

Plant Digital Phenomics

Designing and developing explainable artificial intelligence and extended reality systems for computational phenomic data analytics



UNIVERSITÀ
DEGLI STUDI DELLA
TUSCIA

Farid Nakhle

Supervisor: Prof. Antoine L. Harfouche

Department for Innovation in Biological, Agro-food, and
Forest Systems

University of Tuscia

June 2023

Plant Digital Phenomics

Designing and developing explainable artificial intelligence and extended reality systems for computational phenomic data analytics



Farid Nakhle

Supervisor: Prof. Antoine L. Harfouche

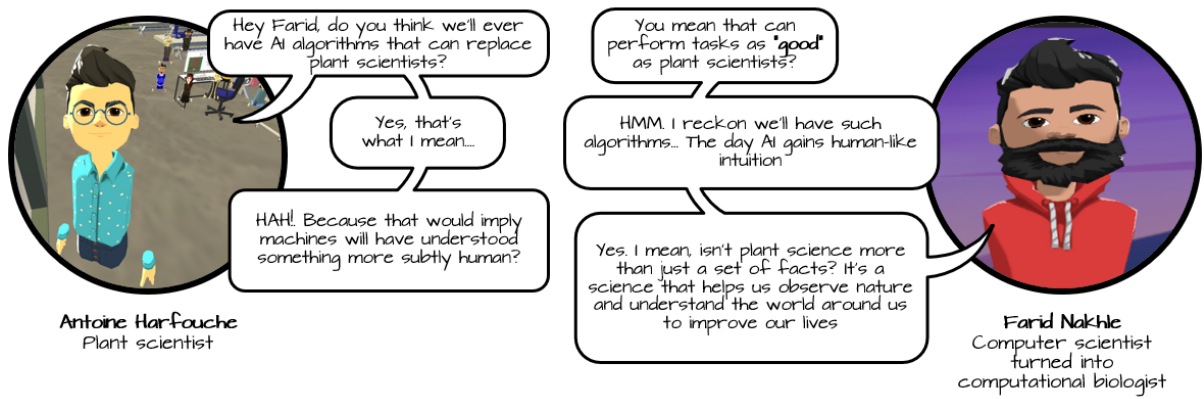
Department for Innovation in Biological, Agro-food, and
Forest Systems

University of Tuscia

June 2023

This dissertation is submitted for the degree of

Doctor of Philosophy



Internal committee:

- Dr. Anna Barbati, Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Viterbo, Italy.
- Dr. Gabriele Chilosi, Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Viterbo, Italy.
- Dr. Silvano Fares, National Research Council of Italy, Institute of Bioeconomy, Rome, Italy.

External committee:

- Dr. Erik Alexandersson, Department of Plant Protection Biology, Swedish University of Agricultural Sciences (SLU), Alnarp, Sweden.
- Dr. Antoine H. Harfouche, Unité de Formation et de Recherche en Sciences Économiques, Gestion, Mathématiques, et Informatique, Université Paris Nanterre, Nanterre, France.

© 2023 Farid Nakhle.
All rights reserved.

CIRCULATION RESTRICTION

To whom it may concern,

I would like to request that a circulation restriction be placed on my dissertation "Plant Digital Phenomic: designing and developing explainable artificial intelligence and extended reality systems for computational phenomic data analytics" for a period of 1 year, because an article on this subject has been submitted for publication and others are being written. I have discussed this with my supervisor and we both agree. I understand that one electronic copy of the dissertation will be available only at the Library of the University of Tuscia.

Sincerely,

Farid Nakhle



Antoine L. Harfouche



Antoine
HARFOUCHE
06.06.2023
12:37:42
GMT+01:00

Preferred citation

Nakhle, F. (2023) Plant digital phenomics: designing and developing explainable artificial intelligence and extended reality systems for computational phenomic data analytics. PhD thesis. University of Tuscia, Italy. 261 pp.

Summary

Plant Digital Phenomics

Designing and developing explainable artificial intelligence and extended reality systems for computational phenomic data analytics

Farid Nakhle

Looking at the formidable challenges of the 21st century that are addressed in the United Nations' Sustainable Development Goals (SDGs), plant digital phenomics is set to advance the achievement of key SDGs aimed at maintaining global food and energy security and combatting climate change. In the midst of a new technological revolution in phenomics-driven data, the plant science community has both the opportunity and responsibility to embrace digital technologies, embedding interdisciplinarity into its research and innovation. In this spirit, this dissertation concerns the outstanding computational bottlenecks that constrain plant phenomics and breeding in the service of addressing a major sustainable agriculture goal – the rapid development of new food and bioenergy crop varieties with increased resilience to climate change. Indisputably, effective explainable artificial intelligence (X-AI) and extended reality (XR) systems are vital to scale up and sustain plant phenomics programmes and biotechnological inventions. This thesis contributions span the design and development of scalable X-AI systems and the implementation of X-AI algorithms to process phenomic data and produce novel biological insight. To move from data to insights, I have developed two X-AI-based phenomic data analysis workflows and implemented various X-AI algorithms to analyze and predict plant diseases, with special emphasis on food and bioenergy crops. I demonstrate the potential of XR and its applications in plant and wood science by showcasing the eXtended Reality and eXplainable Artificial Intelligence (X-RAI) park and the wood science and technology metaverse campus (WOODMeta) that were designed and built for education and training in a creative and innovative manner that advances knowledge and research, and closes the gap between research and practice. If we combine X-AI and XR, plant phenomicists will be far more effective at scaling up and

sustaining human-centric interventions to achieve significant resilient agriculture gains.

Keywords

Digital technologies, computational phenomics, climate change, climate resilience, explainable artificial intelligence, extended reality, metaverse, plant phenomics.

“Advancing technology is the *permissive* source of economic growth, but it is only a potential, a necessary condition, in itself not sufficient.”

Kuznets, 1973, p.247; emphasis in the original
Receiver of the 1971 Nobel Memorial Prize in Economic Sciences

Acknowledgements

My most profound appreciation goes to Professor Antoine L. Harfouche, my PhD advisor and mentor, for his tireless guidance, time, and effort in helping me succeed in this journey. His vast wisdom and wealth of experience have inspired me throughout my studies as he believed in me, pushed me, and never gave up on me even when I wanted to give up on myself.

I am honored that two well-known researchers, Prof. E. Alexandersson of SLU and Professor A.H. Harfouche of Université Paris Nanterre, agreed to read through my dissertation and to serve on my external committee board. I am grateful for their constructive and insightful comments which greatly helped to improve the thesis. Many thanks go to the internal committee board, Prof. A. Barbati, Prof. G. Chilosi, and Prof. S. Fares, who provided consistent support and valuable insights throughout this journey.

I would also like to express my gratitude to my friends who, thanks to their support and encouragement, my time spent studying and living in Italy has been truly rewarding.

To conclude, I would like to extend a special thank you to my parents. It would have been impossible to finish my studies without their unwavering love and support.

Publications and collaborators

The chapters presented in this thesis have been formatted as scientific papers to facilitate publication in academic journals. Co-authorship is attributed, reflecting contributions of collaborators to my research, in the form of supervisory support, logistic support, or contributions to data collection and/or analysis. Therefore, the collective pronoun "we" has been used throughout the text instead of the singular "I". Author contributions are as follows:

- Chapter 1 is entirely my own work.
- Chapter 2 is published: Maesano, M.*, Khoury, S.*, **Nakhle, F.***, Firrincieli, A.*, Gay, A., Tauro, F. and Harfouche, A.L., 2020. UAV-based LiDAR for high-throughput determination of plant height and above-ground biomass of the bioenergy grass *Arundo donax*. *Remote Sensing*, 12(20), 3464. **Impact Factor: 5.349.**

*** Equal contributor as first author.**

Assistance in planning and organization of the manuscript, data collection and analyses, and discussion of analyses was provided by Mauro Maesano, Sacha Khoury, Andrea Firrincieli, Alan Gay, Flavia Tauro, and Antoine L. Harfouche. Supervision of the writing process was provided by Antoine L. Harfouche. I assisted in data analysis and in drafting, writing, and revising the manuscript.

- Chapter 3 is published: **Nakhle, F.** and Harfouche, A.L., 2021. Ready, steady, go AI: a practical tutorial on fundamentals of artificial intelligence and its applications in phenomics image analysis. *Patterns*, 2(9), 100323 (**article chosen for journal cover image, September issue**). *Patterns* is a Cell Press journal that was launched in April 2020 and is expected to receive a strong first Impact Factor in 2023 (<https://www.cell.com/pb-assets/journals/research/patterns/CP22462-Patterns-Authors-Benefits-1663248714727.html>); it publishes ground-breaking original research across the full breadth of data science.

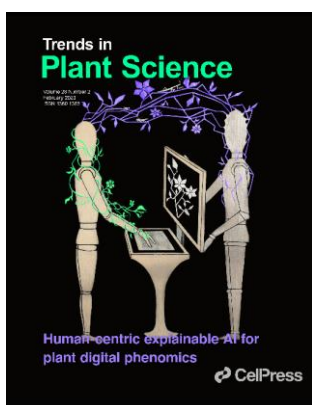


Cover 1. A practical tutorial on fundamentals of AI and its applications in plant digital phenomics

On the cover: Artificial intelligence (AI) can impact and improve our ability to turn data into insights. As these advances have begun to percolate into plant phenomics, it is time to develop educational initiatives for the plant science community so the future workforce will be equipped to leverage AI algorithms and take the lead in this AI revolution that will ultimately bring countless benefits for our breeders and farmers. Nakhle and Harfouche (100323) created “Ready, Steady, Go AI,” an interactive online tutorial to help students and researchers at all levels gain an understanding of how AI can be applied to extract biologically meaningful information from imaging data and to drive innovation within their own research. Image credit: Antoine L. Harfouche and Farid Nakhle.

Assistance in conceiving the idea, designing the X-AI-based analysis workflow, and planning and organization of the manuscript was provided by Antoine L. Harfouche. I developed the tutorial, wrote the codes and notebooks, implemented the algorithms, and conducted the computational research analysis.

- Chapter 4 is published: Harfouche, A.L., **Nakhle, F.**, Harfouche, A.H., Sardella, O.G., Dart, E., and Jacobson, D., 2023. A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey. *Trends in Plant Science*, 28(2), 154–184 (**featured article, article chosen for journal cover image, February 2023 issue**). **Impact Factor: 22.012.**



Cover 2. Human-centric explainable AI for plant digital phenomics

On the cover: Artificial intelligence (AI) predictive models are predominantly black box models. But are black boxes easy to break into? In real-world applications, models must be accurate and correctly and verifiably explainable. Harfouche and colleagues propose a novel human-centric explainable AI (X-AI) system architecture and discuss its potential to improve AI applications in plant phenomics. Pictured is an artist's allegorical drawing of human-AI complementarity. A feedback loop that enables interaction between human (left) and AI (alter-ego; right) to collectively learn and achieve superior results is exemplified by plant branches; mantid flies represent code debuggers. Green curvilinear and purple straight of the hybrid system in which human

and X-AI algorithm (displayed on the screen) can work together emphasize the duality of the nature of the subjects, real and virtual, and human and artificial. Image credit: Ugo Bongarzone (artist, Rome, Italy).

Assistance in conceiving the idea and in planning and organization of the manuscript was provided by Antoine L. Harfouche, Antoine H. Harfouche, Orlando G. Sardella, and Daniel Jacobson. Assistance in designing the network flow was provided by Eli Dart. Assistance in designing the X-AI-based analysis workflow was provided by Antoine L. Harfouche. I implemented, to the best of my knowledge, for the first time, an interpretable by design X-AI algorithm that classifies plant diseases and that explains its own reasoning for each prediction. I developed the tutorial and conducted the computational research analysis.

- Chapter 5 is published: **Nakhle, F.** and Harfouche, A.L., 2023. Extended reality gives digital agricultural biotechnology a new dimension. *Trends in Biotechnology*, 41(1), 1–5. **Impact Factor: 21.942.**

Assistance in conceiving the idea and in planning and organization of the manuscript was provided by Antoine L. Harfouche. I built the virtual world which includes 12 virtual environments (BioInnovation Labs, X-RAI Biotech Enterprise, Biotech Discovery Labs, PCR & Imaging Room, Plant Transformation Facility, Growth Chambers & Growth Rooms, Biotech Data Science & AI Lab, Data Center, RootEye Rhizotron greenhouse, ShootEye greenhouse, X-ray CT 3D RootEye greenhouse, and the Operation & Management Center) and the 3D models (e.g., poplar, tomato, and wheat plant shoots and roots, phenotyping platform and sensors, laboratory equipment such as next-generation sequencing instruments, supercomputers, among others).

- Chapter 6 is in preparation for submission: **Nakhle, F.**, Corona, P., Romagnoli, M., Harfouche, A.L. (2023). Wood science and technology metaverse campus (WOODMeta).

Assistance in conceiving the idea and in planning and organization of the manuscript was provided by Antoine L. Harfouche, Piermaria Corona, and Manuela Romagnoli. I built the virtual world which includes 10 virtual environments (Woodworking Workshop, Wood Art, Architecture, and Design Workshop, WOODMeta Case Classroom, School 42, Wood Collaboration Space, Innovation for Impact Labs, Wood Entrepreneurship Center, Wood Forensic DNA Lab, Wood Forensic Phenomics Lab, and AI, Data Science, and

Blockchain Labs) and the 3D models (e.g., dissecting microscopes, X-ray fluorescence spectrometers, and XyloTrons, machine vision-based automated systems, among others).

During my PhD, I was involved in the following manuscripts which were not included in this thesis:

- Harfouche, A.L. and **Nakhle, F.**, 2020. Creating bioethics distance learning through virtual reality. *Trends in Biotechnology*, 38(11), 1187–1192. **Impact Factor: 21.942.**

Assistance in conceiving the idea and in planning and organization of the manuscript was provided by Antoine L. Harfouche. I built the virtual world and its two environments (Bioethics VR Classroom and Bioethics VR Breakout Room) and the 3D models (e.g., web projectors, whiteboards, pin boards, among others).

- Harfouche, A.L. and **Nakhle, F.**, 2023. Artificial intelligence, blockchain, and extended reality: emerging digital technologies to turn the tide on illegal logging and illegal wood trade. *Annals of Silvicultural Research* <https://doi.org/10.12899/asr-2435> (in press).

I co-conceived the idea and assisted in drafting, writing, and revising the manuscript.

Contents

Summary	XII
Acknowledgements	XVII
Publications and collaborators	XVIII
Contents	XXII
List of Covers	XXVII
List of Graphical Abstracts	XXVIII
List of Figures	XXIX
List of Tables	XXXII
List of Figures 360	XXXIII
List of Glossary Terms	XXXIV
Chapter 1 General introduction	1
1.1 The problem	2
1.2 The solution	6
1.3 Motivation, objectives, and structure	11
Chapter 2 UAV-based LiDAR for high-throughput determination of plant height and above-ground biomass of the bioenergy grass <i>Arundo donax</i>	15
Graphical Abstract	15
Abstract	16
2.1 Introduction	17
2.2 Materials and methods	19
2.2.1 Experimental setup	19
2.2.2 UAV phenotyping platform	22
2.2.3 Data acquisition	23
2.2.4 Data processing and CHM	24
2.2.5 CHM validation and AGB estimation	27

2.3	Results	29
2.3.1	LiDAR CHM	29
2.3.2	AGB estimates derived from 3D LiDAR point clouds.....	33
2.3.3	Heights and biomass of three <i>A. donax</i> crop ecotypes under natural moderate drought.....	35
2.4	Discussion.....	37
2.4.1	Crop height estimation.....	38
2.4.2	AGB estimation.....	38
2.4.3	Comparison of Ecotypes.....	40
2.4.4	Implications.....	41
2.5	Conclusion	42
Chapter 3	Ready, steady, go AI: a practical tutorial on fundamentals of artificial intelligence and its applications in phenomics image analysis	43
	Graphical Abstract	43
	Abstract	44
3.1	Introduction	45
3.2	Getting ready.....	49
3.3	Some basics.....	51
3.3.1.	<i>ML</i>	51
3.3.2.	<i>DL</i>	52
3.3.3.	<i>EL</i>	53
3.4	Getting set.....	53
3.5	Programming languages and IDEs	54
3.6	Software frameworks	61
3.7	Software libraries	61
3.8	Going AI – Materials and methods.....	61
3.8.1.	<i>Step 1: Image dataset selection</i>	63
3.8.2.	<i>Step 2: Data preprocessing</i>	65

3.8.3.	<i>Step 3: Data analysis</i>	71
3.8.4.	<i>Step 4: Performance analysis and explanation</i>	73
3.9	Moving to the cloud	75
3.10	Discussions and the way forward.....	76
3.11	Experimental procedures	77
3.11.1	<i>Data and code availability</i>	77
Chapter 4	A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey	79
	Graphical Abstract	79
	Abstract	80
4.1	Approaching plant phenomics from different angles	81
4.2	How to use AI effectively: ménage-à-trois between plant science, data science, and IS	106
4.2.1	<i>AI and the bias cascade</i>	106
4.2.2	<i>Creating a human-centric X-AI</i>	108
4.2.3	<i>Mitigating bias in human-centric X-AI</i>	111
4.3	How to move from data inputs to outcomes: opening the black box or designing a transparent glass box for explainability.....	113
4.4	Methods and results – how to devise X-AI-driven analytics for phenomics questions	118
4.4.1	<i>Interpretable by design models</i>	118
4.4.2	<i>Dealing with small datasets</i>	121
4.5	How to ensure that human-centric X-AI benefits all: team science, open science, open education, and embedded ethics	122
4.6	Concluding remarks and future perspectives	124
Chapter 5	Extended reality gives digital agricultural biotechnology a new dimension	125
	Graphical Abstract	125

References.....	181
<i>Vitae</i>	224

List of Covers

Cover 1. A practical tutorial on fundamentals of AI and its applications in plant digital phenomics.....	XIX
Cover 2. Human-centric explainable AI for plant digital phenomics	XIX

List of Graphical Abstracts

Graphical Abstract 1. High-throughput field-based phenomics is key to accelerating climate resilient plant breeding.....	15
Graphical Abstract 2. Democratizing AI and AI research is a potential solution to the challenge of turning phenomic big data into insights.....	43
Graphical Abstract 3. Designing a human-centric AI that is explainable, bias-free, embeds human knowledge into its architecture, and keeps humans in the loop presents unprecedented opportunities for computational data analytics in plant digital phenomics and will have an extraordinary impact on plant science in the coming decades.....	79
Graphical Abstract 4. Extended reality is coming of age in digital plant biotechnology.	125
Graphical Abstract 5. WOODMeta acts as a fulcrum of wood frontier research, supporting the development of the wood science and innovation ecosystem.	143

List of Figures

Figure 2.1. The experimental design.	22
Figure 2.2. The unmanned aerial vehicle platform.	22
Figure 2.3. Aerial drone photograph and three-dimensional (3D) light detection and ranging (LiDAR) point cloud for EcoA under moderate drought stress (mDr) conditions.	24
Figure 2.4. Workflow of data processing for field phenotyping of <i>Arundo donax</i>	26
Figure 2.5. Correlation matrix plot with significance levels between the different light detection and ranging (LiDAR) metrics and the field measured above-ground biomass (AGB) production in <i>Arundo donax</i>	28
Figure 2.6. Rasterized three-dimensional (3D) light detection and ranging (LiDAR) point clouds.	30
Figure 2.7. Established crop height model (CHM).	31
Figure 2.8. The relationship between manually measured crop height and the height derived from the light detection and ranging (LiDAR).	33
Figure 2.9. The relationship between the measured above-ground biomass (AGB), and the predicted AGB using unmanned aerial vehicle (UAV)-derived light detection and ranging (LiDAR) data.	34
Figure 2.10. Crop height and above-ground biomass (AGB) of <i>Arundo donax</i> as estimated from manual ground measurements and light detection and ranging (LiDAR).	36
Figure 3.1. Ready, Steady, Go AI: A practical AI tutorial for the plant data science community.	48
Figure 3.2. Schematic overview of the proposed X-AI-based image analysis workflow for a typical digital phenomics experiment from selection through preprocessing, analysis, and explanation.	49
Figure 3.3. Visualization of the tutorial outputs at every step of the proposed X-AI-based image analysis workflow.	65
Figure 4.1. Phenomics platforms and sensors for high-throughput plant phenotyping in controlled environments and field conditions: collecting relevant data from a wide range of sources.	81
Figure 4.2. Data architecture blueprint to drive human-centric explainable artificial intelligence (X-AI) innovation.	83

Figure 4.3. Technology infrastructure to support human-centric explainable artificial intelligence (X-AI).....	84
Figure 4.4. Technology infrastructure to support human-centric explainable artificial intelligence (X-AI).....	85
Figure 4.5. Cultivating the conditions for explainable artificial intelligence (X-AI) to flourish in plant digital phenomics.	116
Figure 4.6. Planning, training, and interpreting an explainable artificial intelligence (X-AI)-based analysis in plant digital phenomics requires careful consideration at each stage of the analysis.....	119
Figure 5.1. Screenshot from the X-RAI park.	128
Figure 5.2. Screenshots from the X-RAI biotech laboratories.	131
Figure 5.3. Screenshots from the X-RAI phenomics greenhouses.....	133
Figure 5.4. Screenshots from the X-RAI 'BioInnovation Labs' and 'Biotech Enterprise'.	134
Figure 5.5. Screenshot from the X-RAI 'Operation & Management Center'.	136
Figure 5.6. Application scenarios of XR in digital agricultural biotechnology.....	139
Figure 5.7. Screenshots from the X-RAI 'Biotech Data Science & AI Lab' and 'Data Center'.....	141
Figure 6.1. 'WOODMeta' in the spotlight: into the wood science and technology metaverse campus.....	147
Figure 6.2. Expressive avatars that make the WOODMeta a bias-free world with diversity and inclusion at its core.....	152
Figure 6.3. Immersive, interactive, and realistic 3D virtual workshops to pick up the ropes of woodworking.	158
Figure 6.4. 'Wood Learning Center' to assimilate wood science and technology... ..	161
Figure 6.5. 'Wood Collaboration Space' to stimulate confabs, networking, and collaborations.	162
Figure 6.6. Wood technologists, wood scientists, designers, and trainers work alongside the next generation of woodpreneurs to hone their teamwork, communication, and entrepreneurial skills, competencies, and attitudes.	164
Figure 6.7. Wood technologists, wood scientists, designers, and trainers work alongside the next generation of woodpreneurs to hone their teamwork, communication, and entrepreneurial skills, competencies, and attitudes.....	167

Figure 6.8. ‘AI, Data Science, and Blockchain Labs’ to drive wood science and technology innovation.....	168
Figure 6.9. Prioritizing sustainability and circularity in wood science and technology.	172

List of Tables

Table 2.1. Light detection and ranging (LiDAR) crop height model (CHM) metrics. .	28
Table 2.2. Measuring the ability of light detection and ranging (LiDAR) metrics in estimating above-ground biomass (AGB) using simple linear regression.	33
Table 3.1. Representative list of available open-source software frameworks and programming languages for the implementation of AI algorithms.	55
Table 3.2. Representative list of available open-source software libraries and IDEs to aid the implementation of AI algorithms.	58
Table 4.1. Publicly available global datasets and their characteristics as valuable resources for plant digital phenomics research ^{a,b}	90

List of Figures360

Figure360 3.1. An author presentation of Figure 3.1.....	78
Figure360 5.1. An author presentation of Figure 5.1.....	142

List of Glossary Terms

Glossary terms are an alphabetical list of words, with accompanying definitions, pertaining to a specific topic that may not immediately be obvious to readers.

3D point cloud: a set of georeferenced data points plotted in 3D space that can be obtained from a variety of sensors such as light detection and ranging (LiDAR) or 3D scanners to express the characteristics of the surface of a plant and to measure a wide range of its morphological parameters.

Algorithm: a set of well-defined computational instructions that extract, process, calculate, and estimate data to solve a problem.

Augmented reality (AR): an XR technology that superimposes valuable visual content and information over a live view of the real world to provide users an enriched version of reality, allowing them to interact with digital objects in real-time.

Benefit: a desirable outcome, such as biotechnologically improved plants for agricultural purposes, or technologically usable knowledge.

Bias: systematic errors in the ability of AI models to make correct predictions.

Compute node: a backend node used for computing in a cluster and reached via a frontend node.

Computer cluster: a group of interconnected computers working together as a single, integrated computing resource.

Confusion matrix: a visual representation that describes the complete performance of an AI model, summarizing its predictions in four categories: true-positives, true-negatives, false-positives, and false-negatives.

Crowdsourced: the act of collecting data by soliciting contributions from a large group of people rather than from traditional experiments.

Edge case: an unlikely but possible scenario such as two overlapping leaves or two alleles sharing the same signature.

Explicit knowledge: the human knowledge that can be readily assembled and passed on by written or verbal instruction. Metadata is explicit knowledge about data.

Extended reality (XR): a combination of interactive and immersive technologies that add to or change our experience of reality, including VR, AR, and MR.

Federated learning (FL): a collaborative AI training paradigm in which copies of a model are distributed to devices, where data is stored, for local training, and the resulting model weights, rather than the data, are sent back to a central server to update the main model.

Field phenomics: a field-based high-throughput phenotypic evaluation under the real conditions experienced by the plant.

GPU-accelerated: a backend node used mainly for accelerating computing, connected in a heterogeneous manner in a computer cluster.

Human-in-the-loop (HITL): an approach that aims to achieve what neither a human nor a machine can do on their own; it leverages a continuous feedback loop between them to train, evaluate, and deploy AI models that continuously learn and improve their prediction accuracy.

Hyperparameters: a group of variables whose values cannot be estimated from data and are manually tweaked to determine the optimal configuration to train a specific model (e.g., learning rate, batch size, number of training epochs).

Metaverse: a persistent, immersive, and interactive simulated world that can be experienced simultaneously by plurality of users where they share a strong sense of mutual presence and can interact with one another, with the virtual world and its environments, and with the objects within them.

Mixed reality (MR): an XR technology that intertwines virtual with real environments by overlaying and anchoring digital content to the real world, and by allowing users to interact with both real and digital objects in real-time.

Model: an equation that represents relationships and identifies patterns among features of a dataset.

Notebook: a web browser-based interactive computing environment that can be used to combine software code, computational output, explanatory text, and multimedia resources in a single document.

Parallel computing: a form of computation in which multiple compute nodes operating simultaneously are used to solve a large problem broken into independent smaller parts that can be processed concurrently.

Phenomics: the comprehensive large-scale study of high-dimensional phenotypes.

Quantum bit (qubit): the quantum analogue of a classical bit; it may adopt the states 0, 1, or any possible combination of both states.

Quantum processing unit (QPU): a computational unit that leverages quantum mechanical phenomena to manipulate information, relying on qubits.

Quantum volume (QV): a metric that measures the performance of a quantum computer taking into account its number of qubits and error rates.

Remote sensing: the science and technology of observing, assessing, and monitoring an environment at local and global scales without coming into physical contact with it.

Synthetic data: data generated artificially using AI algorithms when real data cannot be collected in sufficient amounts.

Tacit knowledge: the know-how, skills, and intuition that live in the individual's experiences and are hard to impart or transfer to others. It can be shared through advances in information and communications technology, and thus becomes explicit.

Transfer learning (TL): a technique in which an AI algorithm reuses parts of a previously trained model on a new model to perform a different but similar task.

Trustworthiness: a quality of an AI model working reliably in ways that anyone can trust; it should be (i) lawful, ensuring compliance with all applicable laws and regulations; (ii) ethical, demonstrating adherence to ethical principles and values, (iii) robust, able to deal with bias during all of its lifecycle; and (iv) explainable.

Unbalanced dataset: a dataset having certain classes contain substantially more training examples than other classes, misleading the classifier algorithm to overlearn the majority classes and to perform poorly in the prediction of the minority classes.

Values: general evaluative standards on which desirability is assessed; they serve as broad guidelines for society.

Virtual reality (VR): an XR technology that immerses users into interactive and explorative virtual environments based on 3D representation and simulation of the real world and objects within it.

Chapter 1 General introduction

1.1 The problem

Future-proofing our food and energy security is a persistent challenge, especially in a world with an expanding population that, according to a report by the United Nations (United Nations, 2022), is projected to reach nearly 10 billion by 2050. Furthermore, rising global temperatures due to climate change – the increasing changes in the measures of climate that prevail over decades – are altering climatic zones around the planet, thus, shifting growing conditions and affecting global crop production (Lobell and Gourdji, 2012; Xu et al., 2022; Zhao et al., 2017), and ultimately threatening food security. The temperature dependence of grain yields, in addition to the unpleasant surprises that climate change can beset, call for rapid actions to avoid the tipping points (Wagner and Schlenker, 2022). In parallel, the increasing population and economic growth are driving the demand for energy globally, which currently relies heavily on fossil fuels (Erdiwansyah et al., 2019; Nepal and Paija, 2019), resulting in an increase in greenhouse gas (GHG) emissions that have adverse effects on environmental degradation (Vo and Vo, 2021) and are a primary driver of climate change. Bioenergy crops play a vital role in future energy sustainability (Robertson et al., 2017) as they provide renewable energy that can substitute the use of fossil fuels (Sallustio et al., 2022; Smith et al., 2016). Today, efforts to curb climate change rely increasingly on the use of plant biomass to reduce GHG emissions (Wagner and Schlenker, 2022) since bioenergy crops absorb carbon dioxide (CO₂) from the atmosphere and, the biofuels, made from plant biomass, can be burnt without releasing the CO₂ back (Wagner and Schlenker, 2022). As a result, large-scale bioenergy crop cultivation has been identified as a major negative emission technology for reducing GHG from the atmosphere (Rogelj et al., 2018; Rosen, 2018). Thus, with policies pushing to promote clean energy, sustainable cellulosic biofuels have become viable and desirable (Sallustio et al., 2022). Importantly, to meet the global demand for sustainable food and energy from rising population and increasing biofuel consumption, agricultural production must double (Ray et al., 2013; Tilman et al., 2011), and larger crop diversity that are highly resistant to extreme weather conditions and that can rapidly adapt to climate change needs to be unlocked (Nature Plants 2021 p.241; Yu and Li, 2022). Importantly, if bioenergy crops are to make a real contribution to sustainable development, they should be grown on agricultural lands that are less suitable for food crops (Lewandowski, 2015). Plants will continuously get

exposed to various stresses from their environment; phenotyping and understanding their stress responses will provide new opportunities to increase their resilience in the face of climate change. Resilience, however, has multiple levels of meaning ranging from sustainability-related characteristic features to measurable indicators of perturbations that can be assessed (Carpenter et al., 2001). In this thesis, resilience refers to the capacity of plants to cope with stress (e.g., extreme weather conditions), responding in ways that maintain their essential function and capacity for adaptation.

Plant breeding, which can be defined as the science of crop genetic improvements to produce new varieties characterized by climate resilience, improved yields, and nutritional qualities, among other valuable traits (Moose and Mumm, 2008; Zamir, 2001), is one of the most important weapons in the race against the increasing food and energy demand, and climate change. Advances in plant breeding were largely responsible for doubling crop yields during the first green revolution (Harfouche et al., 2019). For further improvement of agricultural productivity over the next 50 years, breeding must increase yield and resource-use efficiencies in an unprecedented way (Harfouche et al., 2019). It is unlikely that cropland will be expanded appreciably in the future due to continuously increasing losses from urbanization, salinization, and desertification. Further, global changing climate will make it harder to grow food and bioenergy crops in many parts of the world. Water for irrigation is also under threat from overuse, and water scarcity induced by climate change afflicts agricultural production worldwide. This necessitates agricultural improvements using idealized plants that can grow in such environments. Diverse, interdisciplinary research teams must come together and create opportunities to identify and fill the gaps needed to achieve sustainable agriculture (Streich et al., 2020).

The basic principle behind crop breeding is to first discover and then select for variants with desired traits (Nature Plants, 2021 p. 241). It thus relies heavily on plant phenotyping. This is expected because a plant phenotype is the result of the interaction between its genotype and the environment ($G \times E$) in which it grows (Poorter et al., 2016). It all started around 12 000 years ago, during the era of plant domestication, when farmers started the selection of outstanding plant phenotypes (Altman et al., 2021). Plant phenotyping, which can be defined as the set of methodologies and protocols used to measure a wide range of plant traits at different scales of organization (Fiorani and Schurr, 2013), is essential to obtain detailed data

on each major aspect of the phenotype and to better understand plant biology and improve crops (Furbank and Tester, 2011; Tardieu et al., 2017; Yang et al., 2020). However, traditional phenotypic data acquisition can be time-consuming, labor-intensive, and subjective. This makes the acquisition of large-scale phenotypic data a major bottleneck, restricting plant breeding (Song et al., 2021). Besides that, our lack of understanding of important biological phenomena suggests that we are not yet measuring all of the important variables (Houle et al., 2010).

Over the last two decades, the rapid development of novel imaging sensors and phenotyping platforms that promise to perform phenotyping of a wide range of plant traits, organs, and environmental conditions at scale, have opened up new avenues for automated data acquisition, enabling **phenomics** (see List of Glossary Terms) to evolve and become a thriving research field of its own (Coppens et al., 2017), thereby, countering the so-called phenotyping bottleneck. These sensing technologies are applied to evaluate crop phenotypes as part of improving crop breeding and production efforts in both greenhouses and the field, enabling **field phenomics**. These platforms include, but are not limited to, ground-based platforms such as ground-based robots and aerial platforms such as unmanned aerial vehicles (UAVs). They can mount a variety of cameras and sensors for non-invasive, high-throughput phenotyping. For example, visible light cameras for red–green–blue (RGB) imaging can be used for morphometric analysis of plants and to rapidly and objectively monitor stress response and formulate vegetation indices that provide information on plant health; light detection and ranging (LiDAR) sensors and 3D laser scanners for 3D imaging allow precise and consistent measurement a wide range of morphological parameters such as plant architecture, canopy height, growth rates, and biomass; multispectral cameras and hyperspectral sensors for spectral imaging capture radiation reflected from plants in the visible–near-infrared (VNIR) range, which may contain information about leaf physiological status, water content, and biochemical traits in response to the environment; and thermal IR (TIR) cameras for thermal imaging allow assessing canopy temperature and providing early diagnostics and quantification of plant responses to heat and water stress (Harfouche et al., 2019). In addition, advances in sensing technologies enabled wood phenomics by allowing for direct measurement of wood chemical and physical traits, which is imperative in woody plant phenomics, especially as planted forests are seen as a foundation for renewable energy.

Additionally, sensing technologies allow for quantitative wood anatomy which enables the analysis of wood anatomical traits, giving in-depth insight into plant functioning and their dynamic responses to environmental change (Fonti et al., 2010). For example, hyperspectral imaging and near-infrared spectroscopy can be used for predicting the chemical and physical properties of a of wood samples (Lepoittevin et al., 2011; Mäkelä et al., 2021); terrestrial and airborne laser scanning can employ LiDAR to measure wood quality indicators to describe elements that influence wood formation and the resulting wood properties (Pyörälä et al., 2019). These phenotyping platforms and sensors are generating a massive amount of data, making phenomic big data a reality. To make sense of these data by leveraging explainable artificial intelligence (X-AI), one of the most important factors to understand and take into account is their attributes, which, can be characterized by eight Vs: (i) volume, the size and amount of data (e.g., images); (ii) velocity, the speed at which data is generated, ingested, accessed, and processed; (iii) variety, the diversity of data types (at rest, typically structured or semistructured, or in motion, typically unstructured); (iv) variability, data with rapidly changing structure or meaning; (v) visibility, the ease of readability, understandability, accessibility, and visualization of data; (vi) veracity, the quality and degree of consistency, accuracy, and trustworthiness of data; (vii) value, the ability to extract useful insights from data; and (viii) vexing, the accuracy and effectiveness of AI **models**, trained from the data (Harfouche et al., 2019).

Phenomic big data present great opportunities to gain insights into the complex biological mechanisms that underlie plant functions in response to environmental stresses. However, they are heterogeneous with high variability, and are high dimensional, deriving from a wide range of experiments that yield many types of information that are possibly ambiguous due to missing values, high data redundancy, and underlying **biases**; and many datasets are noisy, sparse, and irregularly sampled or collected under different conditions and at ambiguous time points, resulting in ill-defined prediction targets (Harfouche et al., 2019). In addition, some phenomic data do not include the necessary metadata of their generation methods and the experimental and environmental conditions in which they were acquired. This leaves us with the daunting challenge of analyzing, interpreting, and integrating this plethora of data.

1.2 The solution

Plant science has seen many revolutions, beginning a few thousand years ago with the domestication of plants, to few hundred years ago with the systematic use of crop rotations, and few decades ago with systematic breeding (Walter et al., 2017). This shows how the plant and agricultural research community not only perseveres over time, but continues to thrive (Nature Plants 2021 p.101). The rise of phenomic big data and the emergence of digital technologies, particularly X-AI and XR, are generating enormous interest in digital phenomics for years to come and are ushering a new era of vegetation observation and insight, giving us the ability to discover and model the dynamics between crop performance and environmental factors. These emerging and enabling technologies hold great promise for overcoming the multiple bottlenecks moving from phenomic data to insights. They constitute a technological revolution that will generate disruptive changes in plant digital phenomics and, consequently, in plant breeding.

Data analytics using traditional statistical methods is limited in analyzing high-dimensional data because these methods are unable to capture complex nonlinear relationships between the independent and dependent variables (Bzdok et al., 2018). The same holds true in the case of phenomic big data (Harfouche et al., 2019). In contrast, AI is capable of learning hidden patterns and relationships in large datasets. The term AI was not formally introduced until 1956, when John McCarthy coined it at a conference at Dartmouth College, in Hanover (Moor, 2006). AI continued to flourish until 1973, when a report, known as the Lighthill report (Lighthill, 1973), criticizing AI research for not delivering impactful discoveries caused a loss of confidence and henceforth a decline in resources for AI research and started a period known as the AI winter (Kaul et al., 2020; Zadpoor, 2019). Despite this, many researchers continued working with AI under other discipline-specific names, including cognitive systems and intelligent systems (Delipetrev et al., 2020). Fast forward to 2012, the AI-based image analysis research marked a significant milestone when AlexNet, an AI **algorithm** that dramatically reduced image analysis error rates, was introduced (Krizhevsky et al., 2017). In the last decade, AI started revealing its disruptive potential and has shown impressive results in phenomic data analysis. For example, It has been successfully employed to predict the yield of individual plants of barley and wheat (Nevavuori et al.,

2019), to classify and quantify biotic and abiotic stresses in leaves of various fruit and vegetable crops (Geetharamani and Pandian, 2019), to segment roots of chicory, wheat, and rapeseed (Gaggion et al., 2021; Smith et al., 2020; Yasrab et al., 2019), to count tobacco leaves (Ubbens and Stavness, 2017), to classify biotic and abiotic stresses in soybean leaves (Ghosal et al., 2018), to estimate the nitrogen nutrition index for improving nitrogen use efficiency in rice (Qiu et al., 2021; Shi et al., 2021), to analyze the intrinsic variability of wood anatomical features in conifers, alder, beech, and oak (Resente et al., 2021), to estimate above ground biomass in maize (Shu et al., 2021), to detect and segment root decay in wheat (Bodner et al., 2018), to detect response to drought stress in bell pepper, courgette, sunflower, radish, foxtail millet, sorghum (Burnett et al., 2021), and spinach (Raza et al., 2014), to estimate yield and canopy geometric characterization in apple (Gené-Mola et al., 2020a), to predict plant water status in grapevine (Gutiérrez et al., 2018), and to identify abnormalities in organelle morphology in *Arabidopsis thaliana* (J. Li et al., 2021). AI is also disrupting the field of wood digital phenomics, which Harfouche et al., 2022 defined as the use of digital tools to upgrade the phenotyping of wood anatomy and wood identification, and is raising awareness about its importance for an in-depth understanding of wood as a complex system. For example, AI has been employed to analyze the intrinsic variability of wood anatomical features in conifers, alder, beech, and oak (Resente et al., 2021) and to identify wood (Hwang and Sugiyama, 2021; Ravindran and Wiedenhoef, 2020). Because woody biomass is considered a promising resource for sustainable biorefineries, wood digital phenomics becomes extremely important to guide forest tree breeding to develop highly productive genotypes that are resilient to climate change. Not to mention that wood phenomics is an integral weapon in the fight against illegal logging, which often causes forest degradation and, sometimes, deforestation in the geographies where it is prevalent, contributing to climate change (Grant and Keong Chen, 2021).

Phenomic big data and AI have a synergistic relationship. In fact, AI models require a large amount of data to be trained for prediction. But some phenotyping experiments generate small amounts of data. In such cases, what can be done for models to learn well from small datasets? In Chapter 4, several methods and techniques to learn from small datasets are described in detail. These include using low-complexity models, **transfer learning**, cumulative learning, the cosine loss function, and one-shot or few-

shot learning. The chapter also discusses alternative workarounds in which data can be oversampled to augment small datasets. This can be achieved by performing geometric transformations on existing images, pairing existing images using methods such as cut-and-paste or CutMix, generating new **synthetic data** with generative adversarial networks (GANs), or simulating real-world data by making use of **extended reality** (XR).

AI is creating an unprecedented opportunity for phenomic data analytics and knowledge discovery and will underpin efforts to develop plants with improved climate resilience. However, AI algorithms evolved to an extent that humans do not understand the complex mechanisms by which AI models make predictions. Our inability to see inside the black box of these models can result in AI adoption being hindered. To be trusted, AI must not only predict correctly but must also explain the logic behind its decisions. To achieve this goal, *post hoc* X-AI algorithms have been developed. They aim to make the rationale behind the output of black box models understandable to humans. These algorithms are employed after a black box model is trained. But, because they are not connected to the black box model's internal design, *post hoc* algorithms employ data perturbation strategies by modifying the input data and observing the changes in the outputs of the black box model to generate an explanation of a prediction. However, since this is only an approximation of the inner workings of the black box model, it is possible that the generated explanations do not provide sufficient detail to understand what the black box model is actually computing. This created a need to design new X-AI algorithms that explain their own predictions without the need for another *post hoc* algorithm. These algorithms are interpretable by design and their explanations are faithful to what their models are computing. But scientific, technical, social, ethical, and legal concerns around X-AI regarding the handling of data, biases, and responsibility have prompted the question of how to design, implement, deploy, and monitor post-deployment X-AI technology in an ethical, human-centric manner, where explainable algorithms embed human-knowledge into their models, are bias-free, and keep **human-in-the-loop (HITL)** for harmonious human and X-AI system symbiosis. In this regard, we propose in Chapter 4 a novel human-centric X-AI system architecture for plant phenomic research, consisting of data architecture, technology infrastructure, and AI architecture design.

Today, AI is topping the list of industry trends and innovations¹ and is becoming increasingly ubiquitous in plant phenomics and biotechnology. Second on the list is XR, which encompasses interactive and immersive technologies, including **virtual reality** (VR), **augmented reality** (AR), and **mixed reality** (MR). Jaron Lanier first coined the term VR in 1987 when he founded the first company to commercialize VR headsets (Flores-Arredondo and Assad-Kottner, 2015). Following VR, Neal Stephenson coined the term **metaverse** in his 1992 Snow Crash novel, where he describes the metaverse as a virtual world where characters could go to escape a dreary reality (Stephenson, 1992). Later that year, Thomas Caudell coined AR after designing a head-mounted apparatus that displays planes' wiring instructions (Caudell and Mizell, 1992). Finally, in connection with his investigations into interactions with virtual objects, Paul Milgram coined the term MR in 1994 (Milgram and Kishino, 1994). Now and in the future, XR will undoubtedly play a striking role in making plant phenomics and biotechnology more accessible and tangible to users globally. Beside its potential application in augmenting small phenomic datasets, it serves as a potent tool for effective data visualization, a key part of the knowledge discovery process, bridging between the quantitative content of the data and human intuition (Donalek et al., 2014). The accumulation of vast quantities of phenomic data poses a tremendous challenge for researchers to effectively visualize and mine these data to explain a biological phenomenon of interest. Data visualization is also an important aspect for quality control of the input data, making it possible to flag outliers for further inspection (Coppens et al., 2017). VR allows researchers to intuitively explore and manipulate complex phenomics data where they can observe and gain a better understanding of the relationship among different elements, and consequently identify previously unappreciated links between them. Compared to conventional data visualization methods which condense high-dimensional data into a digestible 2D format, VR does not necessitate any dimensionality reductions and allows researchers to explore high-dimensional data in its natural dimension. For example, interactive VR environments allowed researchers to collaboratively gain an understanding of hyperspectral data of Indian Pines (Kvinge et al., 2019). In wood phenomics, VR allowed researchers to interactively visualize X-ray micro-computed tomography (micro-CT) data to increase their understanding of complex wood anatomy features (Feng et al., 2020). This power of visualization can give researchers a holistic view by combining phenomic datasets

and, with the aid of X-AI, can guide knowledge discovery to turn these data into insights. XR is expected to offer a variety of plant phenomics and biotechnology applications across a wide range of technological innovations. For example, with data coming from various sensors, through a live camera feed on a tablet or a smartphone, an AR application (app) can enable plant phenomicists to enrich what they see with their eyes to visualize the effect of different environmental stresses across an entire field trial for them to select optimal genotypes or varieties. A similar AR app can use the camera to scan wood logs, and based on wood anatomical features, can identify the species and origin of wood. This could be extremely useful in combatting illegal logging and illegal wood trade. Further, VR simulations allow the exploration of unobservable phenomena by, for example, allowing users to closely observe plants' root systems, walk inside the cells of a leaf, or explore the anatomy of a wood sample to address phenomic questions related to plant functioning, growth, and the environment. This could help scientists understand fundamental problems in plant biology as well as understanding the concept of the plant–disease–environment triangle in a changing climate. Importantly, XR enables the digital twins technological innovation, which are virtual environments and simulations driven by data collected from cameras and sensors to mirror physical and biological objects such as plants in near-real-time. Such technology can be used to create digital twins of crops in a greenhouse to predict their growth and development in response to real-time environmental factors, speeding breeding programs and allowing for better informed breeding decisions. Digital twins of forests can also be created to study the effects of different disturbances, such as illegal logging and drought. Finally, XR, which empowers the metaverse revolution, can be used to build virtual worlds, network of virtual environments, that demonstrate how XR applications in plant science and wood science are far-reaching. These virtual worlds can simulate greenhouses, fields, and forests, not to mention phenomics platforms, biotech laboratories, and wood workshops, among others. The metaverse brings real-life environments that are not yet available for access to the living room of phenomicists around the world, in a fully interactive, immersive, and risk- and bias-free environment. The possibilities are endless. Using custom-built metaverses, phenomicists can experience these virtual worlds at their own pace, or synchronously alongside other researchers from various disciplines including, but not limited to, plant scientists, computer scientists, data

scientists, and information system scientists, enabling them to network, collaborate, and work together to usher new discoveries. Educating in a creative and innovative manner is a powerful resource to close the gap between research and practice and make digital phenomics more accessible, desirable, and useful.

While X-AI and XR technologies differ in their primary objectives, their combination has three-fold added value: (i) phenomic data visualization in XR where data can be immersively and collaboratively interacted with while it spreads across the surrounding virtual environment, allowing for a holistic view of the data that is impossible without XR. In addition, X-AI models can get trained in XR, and can be directly deployed to glean insights from data; (ii) training X-AI models using XR-generated synthetic data where XR can complement collecting real-world training data, which can be a costly and time-consuming process, by generating endless streams of accurate synthetic data to increase diversity in datasets and to increase the robustness of AI models. Notably, XR-generated data can effectively cover **edge cases** that can occur in the real world, ensuring that synthetic data are consistent and bias-free (Nakhle and Harfouche, 2023); and (iii) X-AI conferring intelligence on XR where X-AI models can make simulated environments dynamic and more realistic by generating 3D models that are as close as virtually possible to real objects. The combination of X-AI and XR is emerging as a powerful tool for addressing prominent challenges and is poised to transform how we research and innovate plant phenomics and biotechnology. Ultimately, XR and X-AI will help us dissect the biological mechanisms that underlie desirable traits and will shed light on the flow of information that underpins plant responses to environmental stresses, which is crucial for designing new crop ideotypes that are optimized for niche environments, safeguarding global climate commitments, advancing sustainable development goal (SDG) targetsⁱⁱ, Ecohealthⁱⁱⁱ, planetary health^{iv}, and One health^v.

1.3 Motivation, objectives, and structure

This thesis explores how field phenomics approaches can be used to unlock valuable phenomic datasets and takes on the computational challenge of turning the vast seas of data into valuable predictions and insights. While several studies have looked at using AI to analyze phenomic data, most are limited as they do not provide explanations for the prediction of their models, which raises many technical and

scientific concerns. Explainability of AI models would not only increase the trust of users in why and how predictions were made, but would also help enabling better diagnostics and consequently enhance the models' predictive performance, and would drive their successful adoption. Ultimately, if designed and implemented in a human-centric manner, X-AI will enable bidirectional interaction such that human intelligence and AI are brought together to collectively achieve superior results and continuously improve by learning from each other. This thesis aims to bridge these gaps by exploring the unprecedented opportunities that emerging technologies, such as X-AI and XR, provide for computational data analytics in plant digital phenomics, and by uncovering their extraordinary impact on plant and wood science. The main objectives are to:

(i) Design and develop a phenomic data collection and processing workflow to assess plant morpho-physiological traits, then apply the workflow using a UAV equipped with LiDAR to determine the height and biomass of the bioenergy crop *Arundo donax*.

(ii) Design and develop an X-AI-based phenomic data analysis workflow using *post hoc* X-AI algorithms to generate insights into plant phenomic data, then apply the workflow to identify and classify eight tomato (*Lycopersicum esculentum*) leaf diseases (bacterial spot, leaf mold, *Septoria* leaf spot, target spot, early blight, late blight, mosaic virus, and yellow leaf curl virus) and spider mites.

(iii) Design and develop an X-AI-based phenomic data analysis workflow using interpretable by design X-AI algorithms, discuss their advantages over *post hoc* algorithms, and then apply the workflow to identify and classify four cassava (*Manihot esculenta* Cranz) leaf diseases (bacterial blight, brown streak, green mottle, and mosaic virus).

(iv) Conceive, design, and implement a metaverse for plant phenomics and biotechnology research and innovation, and determine the potential of its far-reaching applications in these fields, setting the stage for phenomic data visualization, time-resolved monitoring of crops, X-AI powered AR apps for plant disease classification, and synthesizing phenomic data in XR.

(v) Conceive, design, and implement a metaverse for wood science research and innovation and explore how it can bring forest and wood industry into the laboratory, workshop, and field in a fully interactive, immersive, and risk- and bias-free

environment that respects human values and promotes developmental and environmental sustainability.

In **Chapter 2** we propose a novel unmanned aerial vehicle (UAV)-LiDAR phenomic data collection and processing workflow to provide new opportunities in assessing plant morpho-physiological traits and in delivering improved genotypes for biorefining. The workflow was applied by testing, for the first time, to the best of our knowledge, a UAV for LiDAR-based phenotyping in the field to remotely determine the height and biomass of the bioenergy crop *Arundo donax*.

In **Chapter 3** we propose an X-AI-based workflow for plant phenomic data analysis and develop an interactive tutorial that introduces the basic principles for implementing X-AI algorithms with *post hoc* explainability in computational plant phenomics. The workflow was successfully applied to accurately identify and classify tomato leaf diseases and spider mites, as well as to explain which visual symptoms are used to make predictions. We choose tomato because it is widely considered one of the most economically important vegetable crops grown all over the world, but it is highly affected by devastating diseases that hinder its production.

In **Chapter 4** we propose a novel human-centric X-AI system architecture for plant phenomics research, consisting of data architecture, technology infrastructure, and AI architecture design. We clarify the difference between *post hoc* models and 'interpretable by design' models, describing the advantages of the latter over the former. We implemented for the first time, to the best of our knowledge, an interpretable by design X-AI algorithm and trained a model that classifies cassava leaf diseases and explains its own reasoning for each prediction. This choice is motivated by the fact that cassava is a key food security crop, widely cultivated in tropical and subtropical regions of the world. Cassava serves as the main carbohydrate source after maize and rice. However, diseases that plague this crop are a major cause of poor yield.

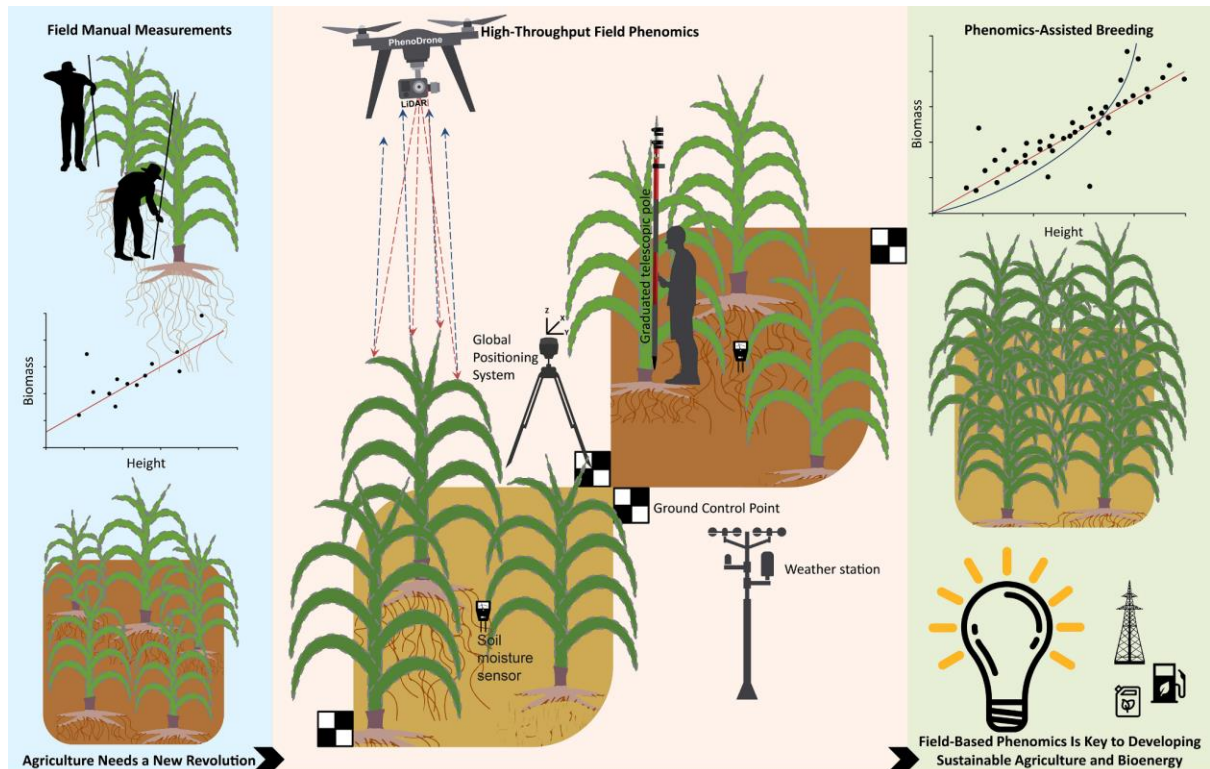
In **Chapter 5** we develop and showcase the X-RAI park to explore the **value** that XR technology brings to research and innovation in plant phenomics and biotechnology. We seek to stimulate technological innovation by empowering future innovators, researchers, academics, and startups to think and partner creatively.

In **Chapter 6** we develop and showcase the WOODMeta that brings the forest and wood industry into the laboratory and field, and consequently into the classroom and workshop in a fully interactive, immersive, and risk- and bias-free environment with diversity and inclusion from the start, while promoting innovation, compatibility, and competitiveness throughout the value chain with environmental and developmental sustainability. Ultimately, WOODMeta will transform how we teach, learn, train, get trained in, research, and innovate wood science and technology.

In the last chapter, **Chapter 7**, I summarize the different findings in this thesis and discuss how they inform each other and can all be combined to turn phenomic big data into knowledge and insight on complex biological phenomena (e.g., biological mechanisms that underlie desirable traits and the flow of information that underpins plant responses to various environmental stressors). I highlight some challenges and key outstanding questions that future plant phenomic and biotechnology research should address, and mention potential avenues to expand on this work.

Chapter 2 UAV-based LiDAR for high-throughput determination of plant height and above-ground biomass of the bioenergy grass *Arundo donax*

Graphical Abstract



Graphical Abstract 1. High-throughput field-based phenomics is key to accelerating climate resilient plant breeding

Abstract

Replacing fossil fuels with cellulosic biofuels is a valuable component of reducing the drivers of climate change. This leads to a requirement to develop more productive bioenergy crops, such as *Arundo donax* with the aim of increasing above-ground biomass (AGB). However, direct measurement of AGB is time consuming, destructive, and labor-intensive. Phenotyping of plant height and biomass production is a bottleneck in genomics- and phenomics-assisted breeding. Here, an unmanned aerial vehicle (UAV) for **remote sensing** equipped with light detection and ranging (LiDAR) was tested for remote plant height and biomass determination in *A. donax*. Experiments were conducted on three *A. donax* ecotypes grown in well-watered and moderate drought stress conditions. A novel UAV-LiDAR data collection and processing workflow produced a dense three-dimensional (3D) point cloud for crop height estimation through a normalized digital surface model (DSM) that acts as a crop height model (CHM). Manual measurements of crop height and biomass were taken in parallel and compared to LiDAR CHM estimates. Stepwise multiple regression was used to estimate biomass. Analysis of variance (ANOVA) tests and pairwise comparisons were used to determine differences between ecotypes and drought stress treatments. We found a significant relationship between the sensor readings and manually measured crop height and biomass, with determination coefficients of 0.73 and 0.71 for height and biomass, respectively. Differences in crop heights were detected more precisely from LiDAR estimates than from manual measurement. Crop biomass differences were also more evident in LiDAR estimates, suggesting differences in ecotypes' productivity and tolerance to drought. Based on these results, application of the presented UAV-LiDAR workflow will provide new opportunities in assessing bioenergy crop morpho-physiological traits and in delivering improved genotypes for biorefining.

2.1 Introduction

Given the rapidly growing world population and pressure placed on availability and productivity of agricultural lands by climate change, it is imperative to explore new technologies and approaches to help increase the rate of crop improvement. These improvements aim to increase biomass of food and feed crops and that of bioenergy crops which reduce the need for fossil fuels. An important step towards this goal has been the transformation in the detail and speed of genetic information brought about by next-generation sequencing. However, there remains a bottleneck in acquiring the information on the detailed morphological and physiological characteristics of large number of crop genotypes, an activity known as phenotyping. In part, this requirement has been fulfilled by the development of phenomics platforms for fast, accurate, high-resolution, and non-destructive high-throughput phenotyping (Yang et al., 2017), which have several advantages, including flexibility, convenient operation (Harfouche et al., 2019; Ludovisi et al., 2017) and the ability to provide controlled experimental treatments. There have also been useful developments in the provision of high-resolution field phenotyping systems, but these are relatively large systems deployed on fixed sites (Virlet et al., 2017). However, neither of these approaches allows the flexibility to operate directly on field crops on any site required, such as a large breeding nursery, an experimental agriculture site or a typical commercial farm where detailed crop phenotyping is required. Thus, the overall aim of this study is to investigate the application of unmanned aerial vehicle (UAV)-based high-resolution remote sensing to phenotyping ecotypes of the bioenergy crop *Arundo donax* under contrasting water stress levels to determine the associations between genomic and phenotypic data that will allow rapid crop improvement (Tardieu et al., 2017; Yang et al., 2017). Key physical parameters that are targets for *A. donax* high-throughput phenotypic evaluation under the real conditions experienced by the plant include crop height and above-ground biomass (AGB) production.

Formally the phenotype is an expression of the genotype and the environment in which it grows, including structural traits (e.g., plant height, leaf area index, lodging, crop canopy cover), canopy spectral texture (spectral features), physiological traits (e.g., chlorophyll, biomass, pigment content, photosynthesis), abiotic/biotic stress indicators (e.g., stomatal conductance, canopy temperature difference, leaf water

potential, and senescence index), nutrients (nitrogen concentration, protein content), and yield (Yang et al., 2017). Many of these traits can be used as indicators of plant fitness and for biomass estimation and have been used in selective plant breeding for decades. An important feature of these measurements is they often require manual measurement or sampling. This is invariably time consuming, can result in considerable crop disturbance and for some measurements requires destructive sampling. Thus, any non-invasive and non-destructive measurements which can be used as an indicator of likely biomass that can readily be applied to a biomass crop would be advantageous. Crop height is recognized as one of the most important components of AGB and yield in biomass crops (Madec et al., 2017; Tilly et al., 2015). Height can be reduced by environmental stress and can therefore be used to separate drought sensitive from drought tolerant crops (Kerby et al., 2013; Su et al., 2019). Traditionally, crop height and diameter are measured manually with graduated poles and measuring tapes; this is difficult to do without canopy disturbance in the central region away from plot edges of tall (3.2 to 4.2 m) *A. donax* plants with their typically high shoot density. In terms of the choice of sensor, light detection and ranging (LiDAR) is a promising technology to measure plant height and predict biomass (Hollaus et al., 2006). LiDAR equipped UAVs have been demonstrated to work well in forestry (Dalla Corte et al., 2020; Picos et al., 2020), maize (Zhu et al., 2019), and shorter annual crops such as wheat (Zhu et al., 2019). There is one study using tractor mounted LiDAR to characterize a bioenergy crop, *Miscanthus giganteus*, with good results (reported accuracy 92–98.2%) (Zhang and Grift, 2012). However, the approach is limiting in terms of field size and shape as well as crop horizontal density, thus demonstrating the advantage of UAV platforms. Prior to our study, there was no survey that has applied UAV-based LiDAR to *A. donax* which contrasts both in terms of crop structure and leaf reflectivity with other crops studied to date.

An important aspect of LiDAR operation for agricultural and ecological studies, where timeliness of measurements is important, is the active laser illumination which means it is routinely deployed under a wide range of illumination conditions. The LiDAR data recorded is developed into a three-dimensional (3D) georeferenced point cloud further separated into a digital terrain model (DTM) and a digital surface model (DSM). These are then subtracted to create a normalized DSM, also called crop height model (CHM) when crop positions have been identified. Further analysis of the CHM

allows estimation of heights and biomass (Coomes et al., 2017) to answer ecologically valuable questions related to these two variables and their associated genotypes.

A. donax is a polyploid perennial rhizomatous grass grown for bioenergy production. It was found to be the most productive biomass crop when allowed to grow for more than four years (Laurent et al., 2015). Long considered an invasive plant, *A. donax* is unimproved and understudied as a crop, presenting a great opportunity for breeders and researchers to explore its suitability for growth on marginal lands. Given the interest in its potential for bioenergy applications, the *A. donax* transcriptome was sequenced and de novo assembled (Evangelistella et al., 2017); thus, good genetic information but much less phenotyping data is available. Not only does *A. donax* have high yield compared to other biomass crops, some ecotypes have high tolerance to soil contamination by heavy metals (Cano-Ruiz et al., 2020), tolerance to high salinity or tolerance to water stress (Pompeiano et al., 2017; Sánchez et al., 2015). Thus, it satisfies the need for an improved bioenergy crop that can yield high biomass, grow in harsh environments, and does not compete for higher quality land suitable for food production.

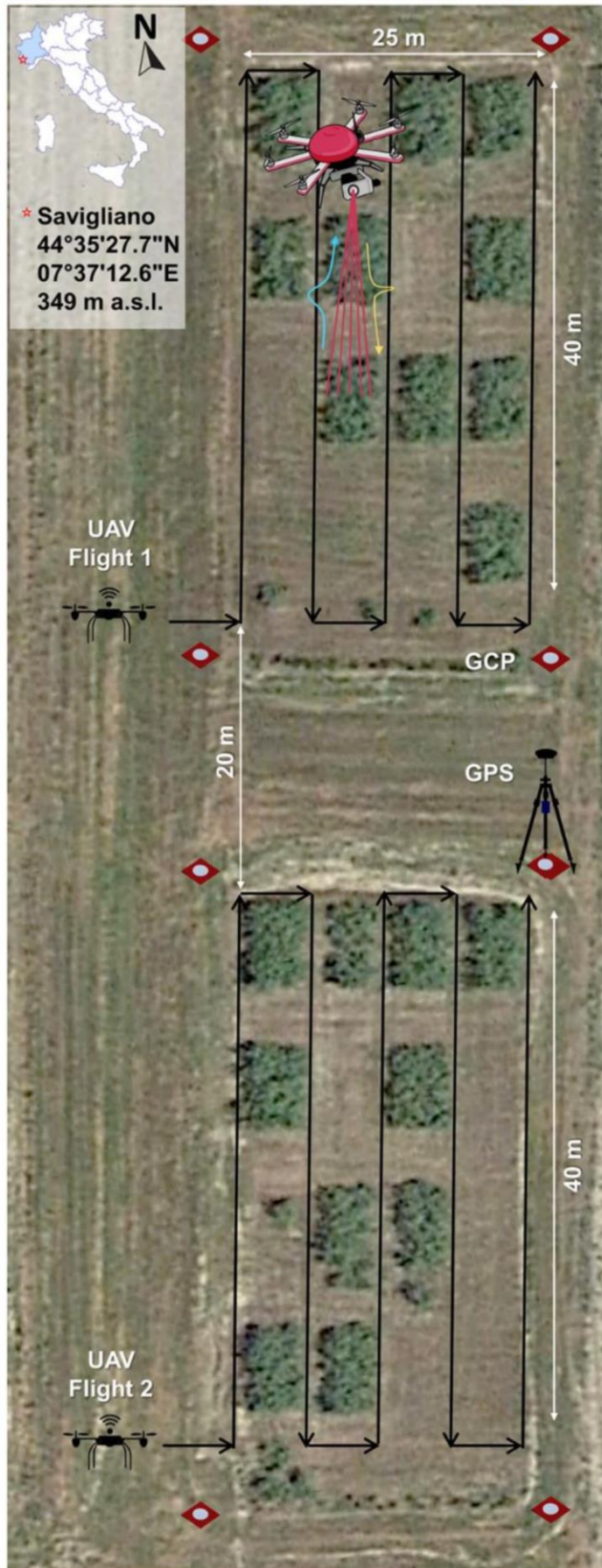
Here, we report the use of drone-mounted LiDAR to determine differences in responses of *A. donax* ecotypes to drought stress under field conditions. We set out to answer three questions: (i) can the proposed method of processing **3D point cloud** data generated from LiDAR determine *A. donax* crop height and estimate biomass production; (ii) can it identify differences between *A. donax* ecotype heights and biomass; and (iii) can it identify differences between drought-stressed and well-watered ecotypes?

2.2 Materials and methods

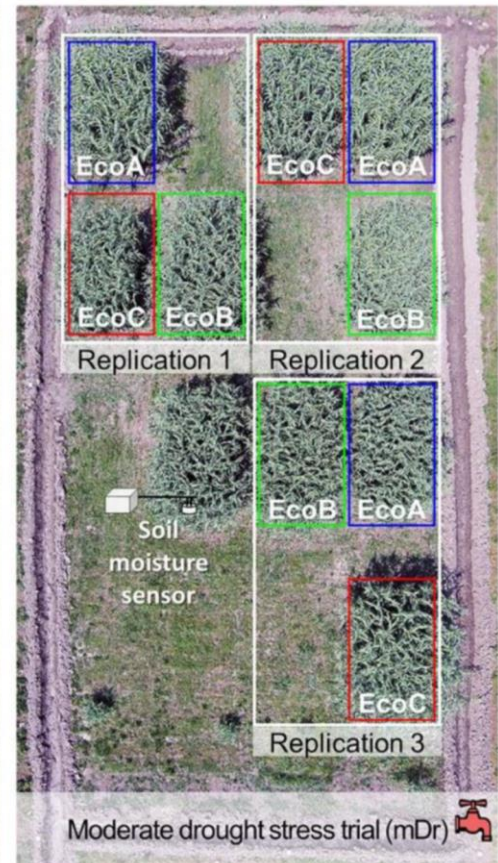
2.2.1 Experimental setup

The study area was located in Savigliano, northern Italy, (44°35' N, 07°37' E, 349 m above sea level) (Figure 2.1a). Three clones of *A. donax* (named hereafter EcoA, EcoB, and EcoC) were selected from a set of 82 EuroMediterranean ecotypes (Fabbrini et al., 2019) and vegetatively propagated by rhizomes. The selected ecotypes originated from three contrasting environments: EcoA was collected from the coastal habitat of Attica region (southern Greece), characterized by typical warm

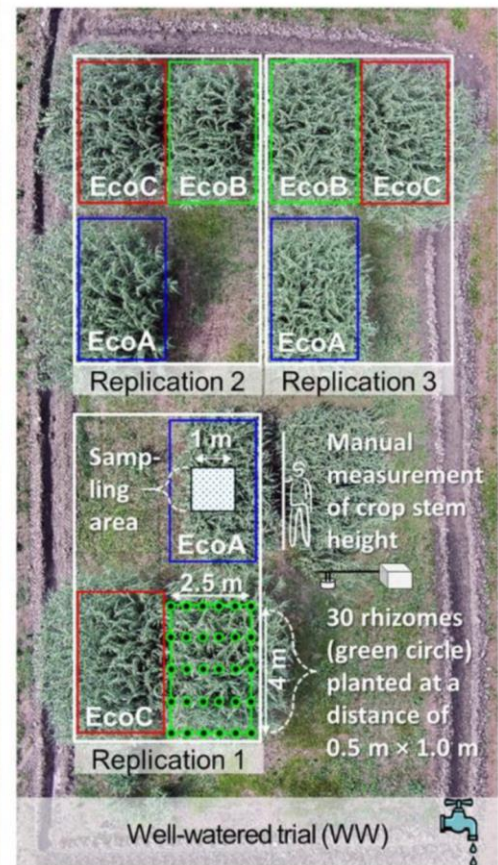
Mediterranean climate; EcoB originated from the coastal habitat of southern Dalmatia region (southern Croatia), characterized by temperate oceanic climate; and EcoC was collected from the hilly area of Viseu district (northern Portugal), characterized by temperate Mediterranean climate. The field experiment began in March 2014. There were three replicate plots of each ecotype grown in each of two watering regimes: the well-watered (WW, kept at 40% volumetric water content) and natural moderate drought stress (mDr, allowed to fall to 15% volumetric water content during June). Each ecotypic replicate consisted of 30 homogeneous cuttings which were planted at distances of 0.5 m × 1.0 m. The experimental design is shown in Figure 2.1.



(a)



(b)



(c)

Figure 0.1. The experimental design.

(a) Aerial view of the experimental site of *Arundo donax* located in northern Italy, Savigliano (see inset), with image showing flight plans over the well-watered (WW) and natural moderate drought stress (mDr) treatments (identified in (b) and (c)), running parallel to plant rows to ensure complete coverage with good overlaps. Four ground control points around each plot are also shown. (b) Layout of mDr treatment with location of soil moisture sensors. (c) Layout of ecotypes and replicates in WW treatment. An example of the central area of one plot used for manual height recording is shown for replication 1 of EcoA (similar positions were used for all plants).

2.2.2 UAV phenotyping platform

In our study, the UAV LiDAR data were acquired by an unmanned G4 SkycraneV2[®] octocopter (Multirotor Service-drone, Germany) equipped with micro LiDAR Velodyne HDL-32E microsensor (Velodyne, San Jose, CA, USA) (Figure 2.2). The hyperspectral sensor was not used in this study. The micro LiDAR sensor is small and lightweight, featuring up to ± 2 cm accuracy, 40° vertical field of view and 360° horizontal field of view with 32 laser beams/scan and dual returns up to 70 m (Velodyne Acoustics, Inc., Morgan Hill, CA, USA). An inertial measurement unit is combined with a real time kinematic global positioning system (GPS) to help georeferencing the captured images, with a precision of 1.0 and 2.0 m in the horizontal and vertical directions, respectively.



Figure 0.2. The unmanned aerial vehicle platform.

The unmanned aerial vehicle (UAV) platform used for field phenotyping of *Arundo donax* on the ground and in flight.

2.2.3 Data acquisition

2.2.3.1 UAV-LiDAR campaign

The airborne campaigns were conducted on 29 June 2016, with UAV LiDAR scanning at an altitude of 25 m above ground with separate flight plans for each treatment (Figure 2.1a). Four ground control points were installed for each treatment (Figure 2.1a), and their positions established with an onboard GPS and an on-ground global navigation satellite system with an accuracy of 3 mm (Leica GS08plus, Leica Geosystems, Switzerland) operating in real-time kinematics.

The flight mission was planned using the autopilot Ground Station v.3.2.304 (Multirotor Service-Drone, Germany). The drone was flown autonomously (GPS-waypoint navigation) at a nominal speed of 3 m/s. Experimental plots were scanned for two 9-min flights under stable cloudless and low-wind conditions. Flights were performed at 14:22 local time above the WW and at 15:29 local time above the mDr field. In order to capture each experimental plot in a single flight pass, the flight plan consisted of transects parallel to the plant rows (Figure 2.1a). A ground station processed the UAV safety manual control and sent telemetry data (position, attitude, and status data) through a radio link at 2.4 GHz to a field laptop. Figure 2.3 shows a UAV RGB image of *A. donax* ecotype in the field and 3D LiDAR point clouds.

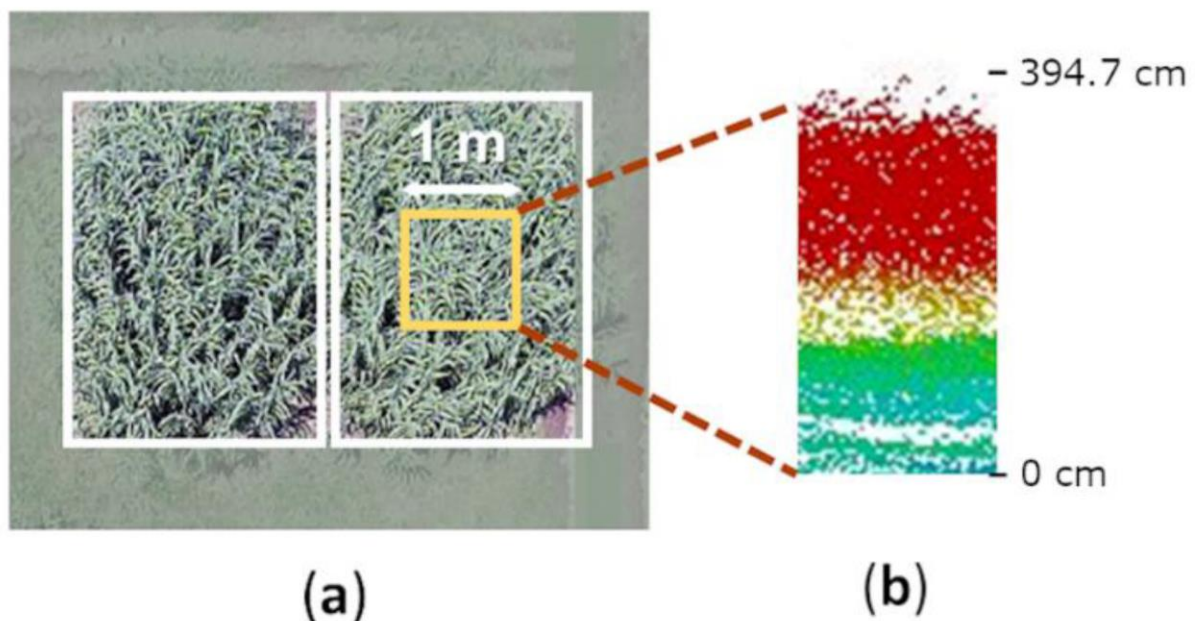


Figure 2.3. Aerial drone photograph and three-dimensional (3D) light detection and ranging (LiDAR) point cloud for EcoA under moderate drought stress (mDr) conditions.

(a) Aerial image of EcoA, from flying the unmanned aerial vehicle (UAV) (Figure 2.2) over the mDr experimental plot. Marked in yellow is the 1 m² area of interest at the center of the plant. (b) 3D LiDAR point cloud color coded by height.

2.2.3.2 Plant height and biomass ground truthing

The heights of four random stems (StH [cm]), in the central 1 m² of each plot, were manually measured to assess the significance of their relationship with the UAV-LiDAR data at the same time the UAV was flying, forming the ground-truth dataset. Heights were measured from the basal node to the top node with a graduated telescopic pole. Nodes are distinct structural components of the stem where the leaves grow out. They are easy to be recognized at a distance by the naked eye as there will be a tiny knob-like on the stem. The lowest node, which is usually just above the ground in *A. donax* is used as the lower reference point for our stem height measurements. For the most part of the *A. donax* growth, the position of the uppermost node is hidden because of overlapping leaf bases and used for the stem height measurement to avoid human bias in the selection of the tallest leaf at the top which is much harder to determine.

Finally, at the end of the growing season (November 2016), the time when *A. donax* is usually harvested as a bioenergy crop, above-ground biomass (total dry mass, AGB, g m⁻²) was estimated for all replicates. Stems were cut at 5 cm above ground level and weighed to determine fresh weight. Then, sub-samples of mixed tissues (stems, leaves, and flowers) were oven-dried at 75 °C for 72 h to determine the conversion to dry weight.

2.2.4 Data processing and CHM

The CHM from the LiDAR data was created according to the workflow in Figure 2.4. The first step was the conversion of each UAV LiDAR scan from .pcap to .las format files using LidarTools (Hyperspec III, version 3.1, Headwall Photonics Inc., Boston, MA, USA). The second step was to clip the LiDAR point clouds to the study areas using LASclip algorithm in LAStools and then eliminate noise (first and last pulse) and outliers inside these point clouds using the LASnoise algorithm in LAStools. LASground, from LAStools, was then used in the third step, to separate ground returns

from non-ground returns. DTMs and DSMs with grid cell size of 20 cm for each field treatment were generated using the “lidR” R package. In the fourth step, DTMs and DSMs were processed in the geographic information system application QGIS (QGIS Development Team, 2018) by normalizing the DSM with the DTM to obtain the CHM (that is the residual distance between the ground and the top of the crop above the ground or the DSM minus the DTM). The CHM was georeferenced by way of ground control points. Finally, an area-based approach (Coomes et al., 2017) was used to generate height and biomass prediction in an area of 1 m² around the center of each ecotypic replicate as described in detail below.

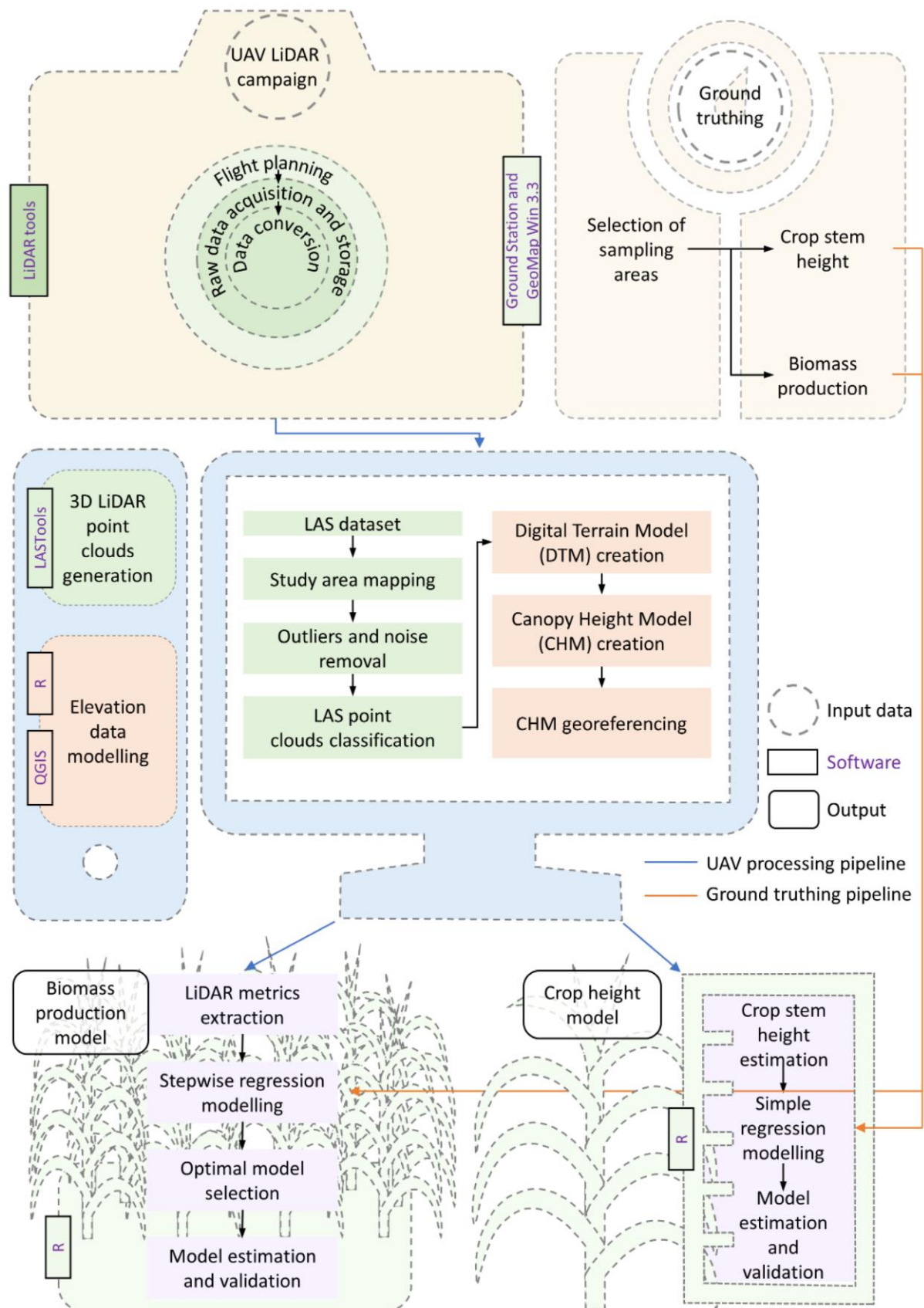


Figure 0.4. Workflow of data processing for field phenotyping of *Arundo donax*.

2.2.5 CHM validation and AGB estimation

LiDAR heights were extracted from a 1×1 m region at the center of each replicate (Figure 2.3a) and averaged for validation against the four randomly measured stems within that central square (Figure 2.1c). The relationship between LiDAR-estimated and manually measured crop heights was evaluated using a linear regression analysis. The coefficient of determination (R^2) was used to define the model's predictive accuracy (considering an F-test p -value < 0.05 as significant and p -value < 0.001 as highly significant), and the root mean square error (RMSE), as a metric for accuracy. Statistical analyses were conducted using R packages such as the “leaps” and the “metrics”. While the leaps performs an exhaustive search for the best subsets of the variables in x for predicting y in linear regression (Lumley, 2020), the metrics implements metrics for regression, time series, binary classification, multiclass classification, and information retrieval problems (Hamner et al., 2018).

For the AGB estimation, eleven LiDAR CHM metrics (defined in Table 2.1), calculated for 1 m^2 area of each plot, were assessed as potential variables for inclusion in the model. Following the area-based approach (Coomes et al., 2017), these metrics were computed for the 1 m^2 area of each ecotypic replicate. We used several linear regressions to study the relationship between each CHM metric variable, extracted from LiDAR data, and the manually measured AGB. Initially, LiDAR CHM metrics with Pearson correlation coefficient (R) values larger than 0.4 were selected to estimate AGB, as is usually done to reduce the number of possible AGB estimators and speed up computations e.g., (Qin et al., 2017; Wan-Mohd-Jaafar et al., 2017; Zhang et al., 2017). The correlation coefficients of all the LiDAR metrics and the field measured AGB are plotted in Figure 2.5.

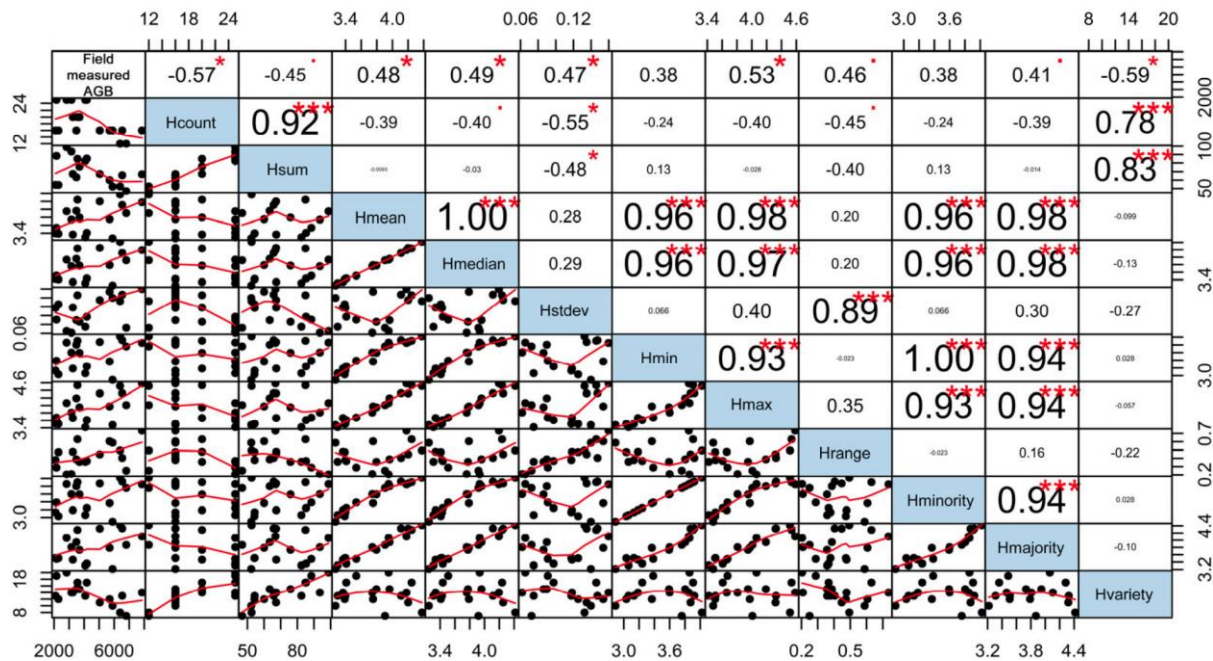


Figure 0.5. Correlation matrix plot with significance levels between the different light detection and ranging (LiDAR) metrics and the field measured above-ground biomass (AGB) production in *Arundo donax*.

The lower triangular matrix is composed of the bivariate scatter plots with fitted smooth red lines. The upper triangular matrix shows the Pearson correlation significance level (as red asterisks or red squares). Each significance level is associated to a symbol: p-values 0.001 (***), 0.05 (*), 0.1 (•).

Table 2.1. Light detection and ranging (LiDAR) crop height model (CHM) metrics.

LiDAR CHM metrics based on the 20 × 20 cm grid cells from the central 1 × 1 m portion of each plot used for estimation of *Arundo donax* crop biomass.

LiDAR CHM metric	Definition	Unit
Hcount	Number of points identified as stem heights	Number
Hsum	Sum of stem height values	cm
Hmean	Mean of stem height values	cm
Hmedian	Median of stem height values	cm
Hstdev	Standard deviation of stem height values	cm
Hmax	Maximum of stem height values	cm
Hmin	Minimum of stem height values	cm
Hrange	Range (Hmax – Hmin) of stem height values	cm
Hmajority	Stem height with most occurrences	cm
Hminority	Stem height with least occurrences	cm

Hvariety	The count of unique stem height values	Number
----------	--	--------

Two way-ANOVA tests and pairwise comparison with Bonferroni correction (to adjust p -values for multiple comparisons) (Wan-Mohd-Jaafar et al., 2017) were performed on the ground data and the LiDAR estimates to determine whether differences between ecotypes and water treatments exist. Linear regressions between AGB and heights were also evaluated for manual measurements and LiDAR estimates.

2.3 Results

2.3.1 LiDAR CHM

The LiDAR system generated high-density 3D data in discrete point clouds (Figure 2.6). The point density of the collected data was 30.1 pt m^{-2} in mDr and 41 pt m^{-2} in WW, while the point spacing was 18 and 16 cm in mDr and WW, respectively.

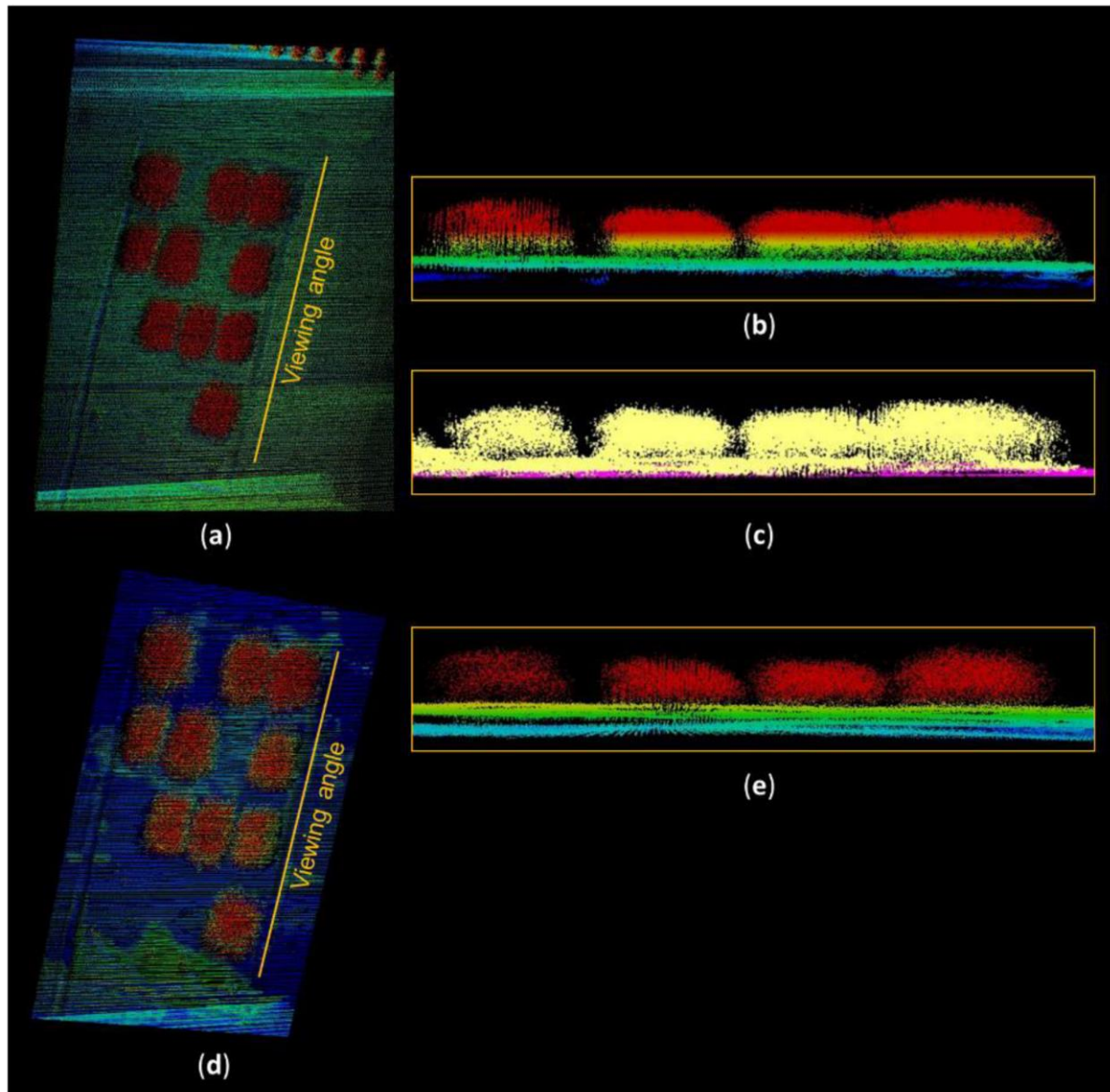


Figure 0.6. Rasterized three-dimensional (3D) light detection and ranging (LiDAR) point clouds. Rasterized 3D LiDAR point clouds before and after they were processed according to the workflow outlined in Figure 2.4. 3D LiDAR point clouds before the removal of outliers from (a) orthographic and (b), frontal view (presentation in the Y-Z plane); (c) visual distinction between the ground surface and plant canopies from the frontal view: the purple indicates ground and yellow indicates plants; (d,e) 3D LiDAR point clouds after the removal of outliers from the orthographic and frontal views, respectively (presentation in the Y-Z plane).

In this study, there was a highly significant positive linear relationship ($R^2 = 0.73$, $p\text{-value} < 0.001$) between the estimated *A. donax* crop heights (LiDAR Hmean) derived from LiDAR CHM (Figure 2.7) and the manually measured crop heights (Figure 2.8a). No significant difference was found between the relationships of the two treatments

(Figure 2.8b) indicating that the relationship between LiDAR estimated crop heights and ground measurement is stable. Although the relationships between Hmean and manual height measurements were significant, the Hmean values were consistently greater. This could be related to the greater number of points identified as stems (12–24; Figure 2.5) compared with the four stems measured manually. However, difference in heights is more likely to result from different detection of the top of the stem between manual and LiDAR measurements. It can be seen in Figure 2.3 and Figure 2.7 that the LiDAR returns are apparently from upper leaves in the canopy, which were not included in manual stem height measurements as mentioned in the methods Section 2.2.3.2.

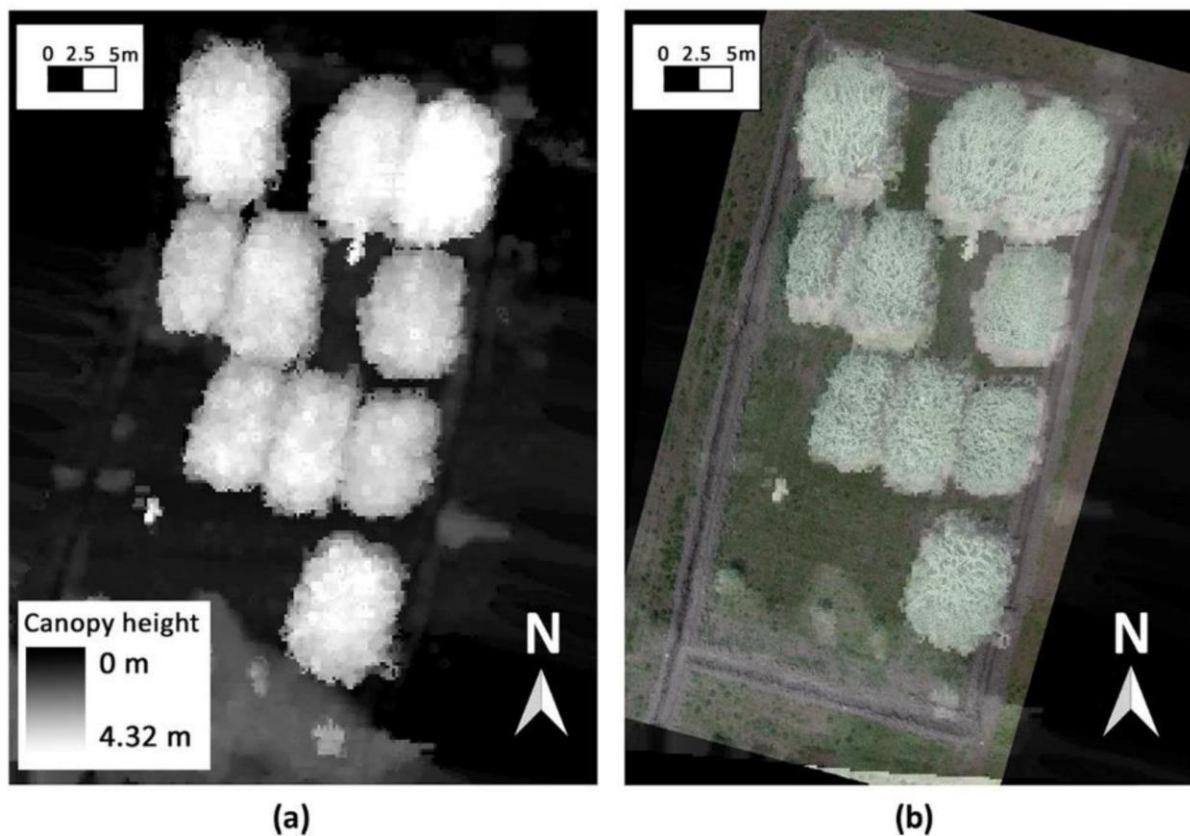
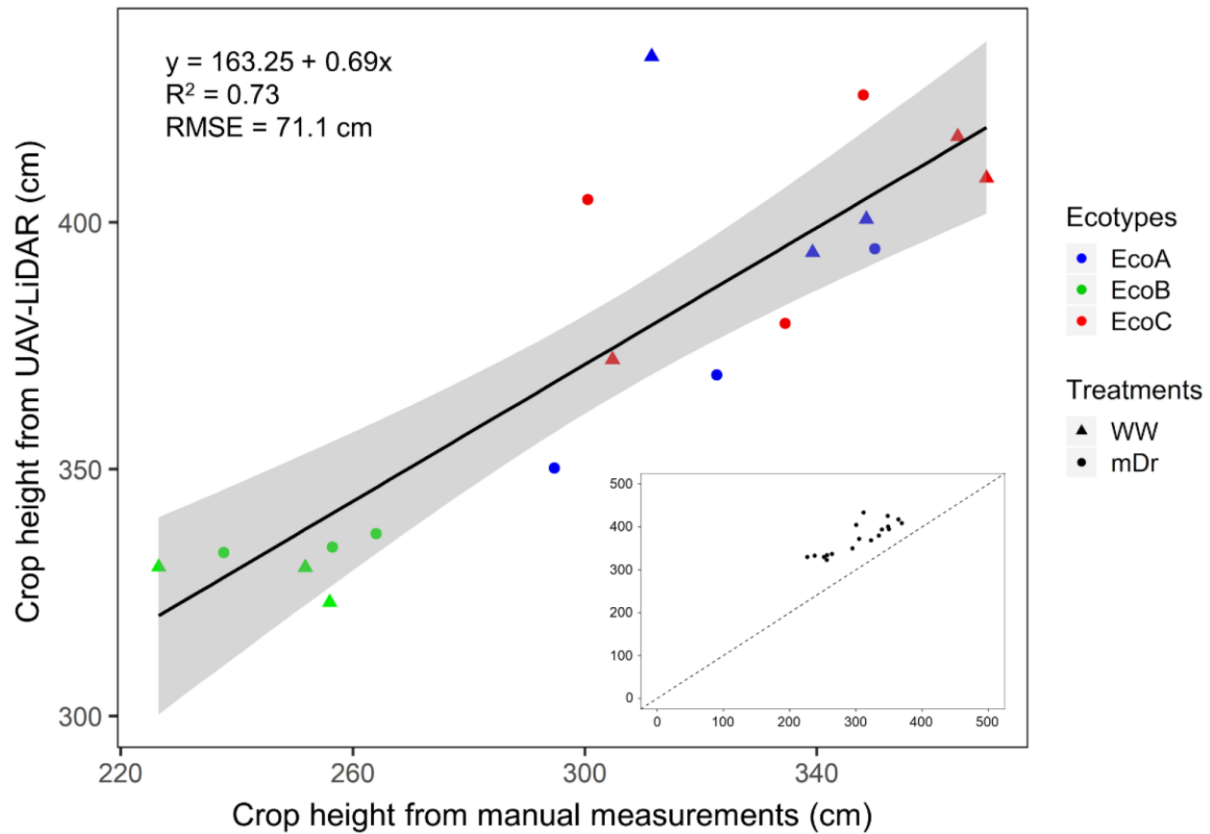
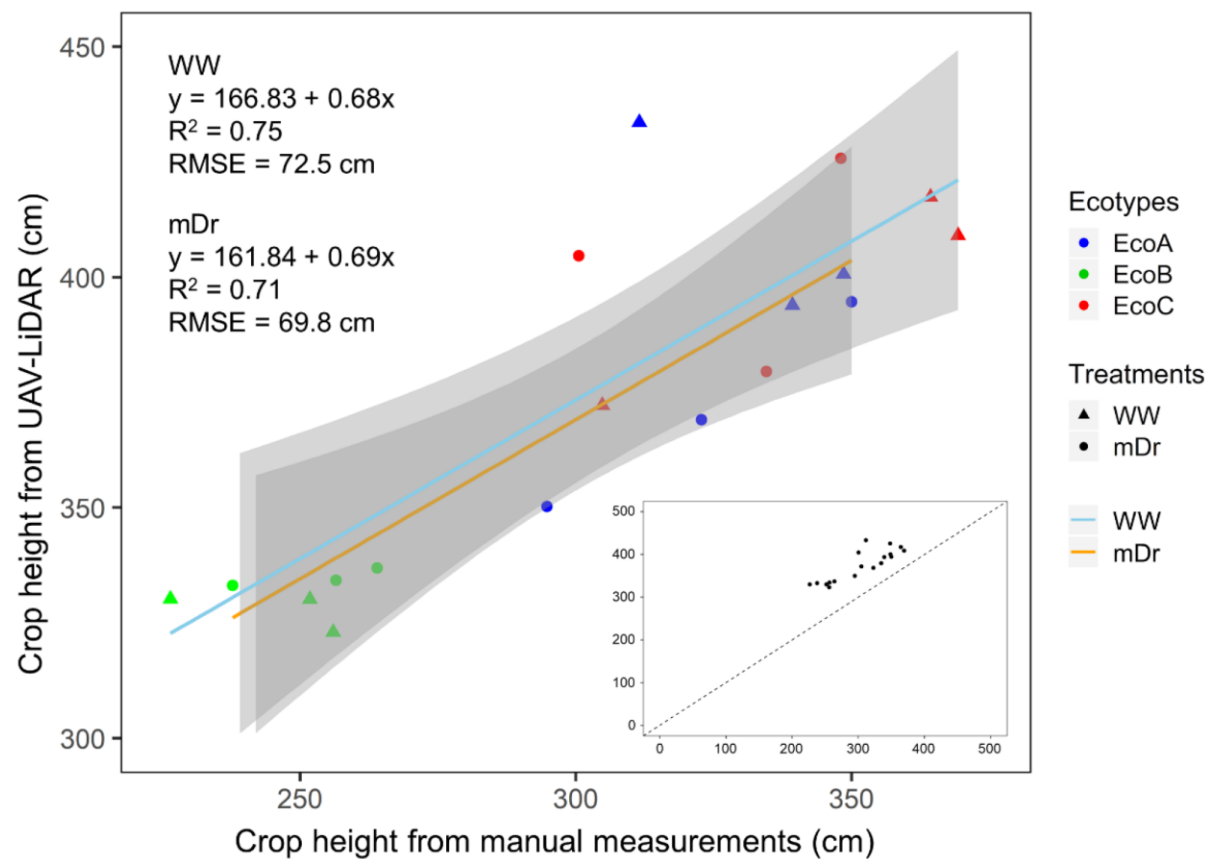


Figure 0.7. Established crop height model (CHM).

(a) CHM of the experimental plot in the moderate drought stress (mDr) treatment; (b) CHM overlapped with an orthophoto of the mDr treatment.



(a)



(b)

Figure 0.8. The relationship between manually measured crop height and the height derived from the light detection and ranging (LiDAR).

The relationship between manually measured crop height and the height derived from LiDAR measurements with fitted lines and grey zone representing confidence intervals (95%). (a) Linear relationship for all crop ecotypes; (b) linear relationship for each of the well-watered (WW) and the natural moderate drought stress (mDr) treatments. The inserts depict all measurements, and the dashed line is the 1:1 line.

2.3.2 AGB estimates derived from 3D LiDAR point clouds

None of the individual LiDAR-derived metrics had a highly significant relationship with AGB (Table 2.2) with the highest R^2 being 0.34. However, six of the eleven metrics showed a small but significant ($R^2 \geq 0.22$, p -value < 0.05) relationship with manually measured AGB (Table 2). Hsum, Hmin, Hrange, Hmajority, and Hvariety did not have a significant relationship (p -value > 0.05) with AGB.

Table 2.2. Measuring the ability of light detection and ranging (LiDAR) metrics in estimating above-ground biomass (AGB) using simple linear regression.

¹ LiDAR metric	Model	R^2	² p -value	³ RMSE (g m ⁻²)
Hcount	⁴ AGB = 8549.6 – 222.2 × Hcount	0.32	0.013 *	1402.8
Hsum	AGB = 7936.2 – 50.6 × Hsum	0.20	0.059	1527.8
Hmean	AGB = –4326.5 + 2310.5 × Hmean	0.22	0.045 *	1504.6
Hmedian	AGB = –4360.8 + 2314.9 × Hmedian	0.23	0.040 *	1494.7
Hstdev	AGB = 1527 + 25569.3 × Hstdev	0.22	0.049 *	1511.8
Hmax	AGB = –5087.7 + 2397.8 × Hmax	0.28	0.023 *	1449.1
Hmin	AGB = –2116 + 1844 × Hmin	0.15	0.116	1581.8
Hrange	AGB = 2032.0 + 5541.4 × Hrange	0.21	0.053	1517.7
Hmajority	AGB = –2797.8 + 1887.9 × Hmajority	0.16	0.090	1561.0
Hminority	AGB = –2116.0 + 1844.0 × Hminority	0.15	0.116	1581.0
Hvariety	AGB = 8491.6 – 302.7 × Hvariety	0.34	0.010 *	1386.6

¹ LiDAR metrics are defined in Table 1. ² p -value of the F-test. Asteriks (*) indicate a significant p -value (p -value < 0.05). ³ RMSE: root mean square error. ⁴ AGB: above-ground biomass, precisely, dry biomass production.

A stepwise multiple linear regression was then used to eliminate multicollinearity issues and to select the optimal regression equation. The LiDAR metrics, Hmax,

Hmajority, and Hvariety, gave the best model to predict total biomass with R^2 of 0.71 and RMSE of 908.5 g m^{-2} (Figure 2.9). The optimal regression equation is defined as:

$$\text{AGB} = 1234.73 + 0.71 \times \text{Hmax} \quad (2.1)$$

where AGB is above-ground biomass (g m^{-2}), Hmax is the maximum of stem height values in cm, Hmajority is the stem height with most occurrences in cm, and Hvariety is the count of unique stem height values.

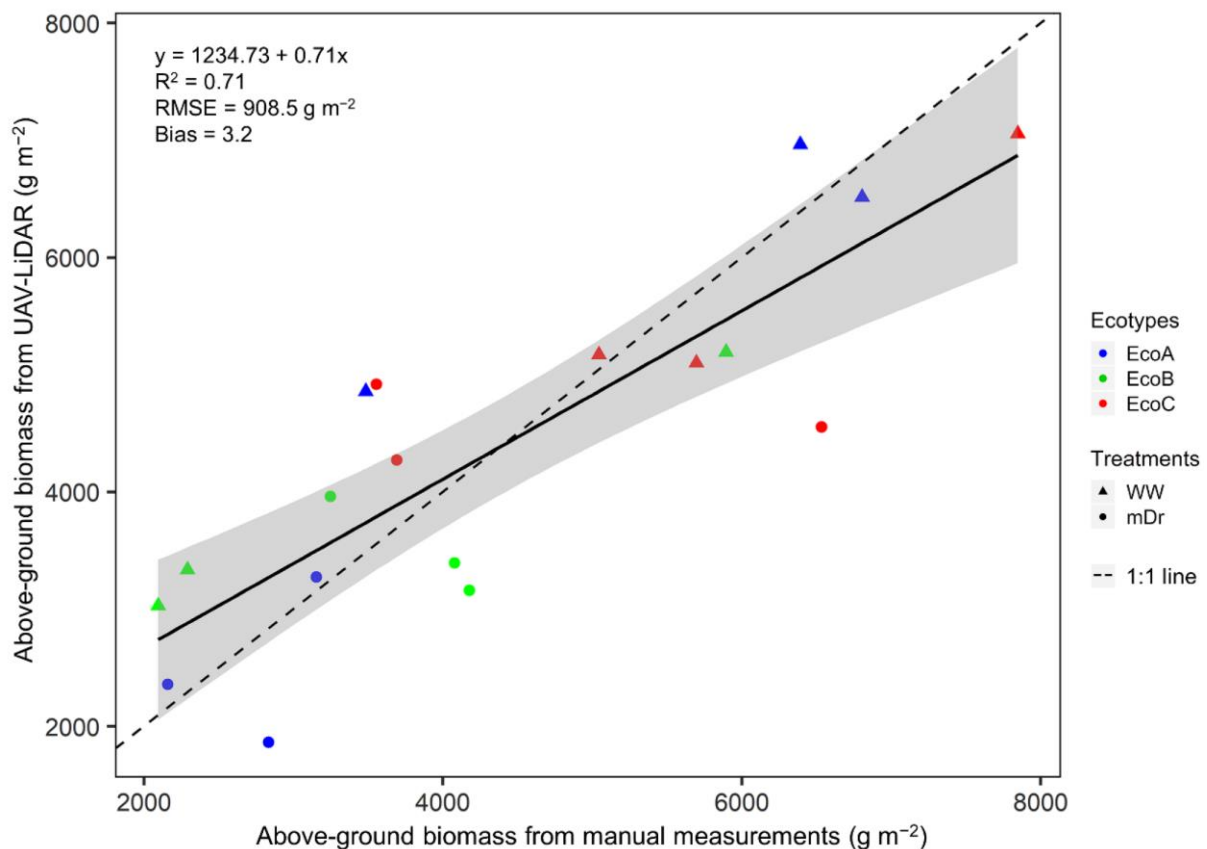


Figure 2.9. The relationship between the measured above-ground biomass (AGB), and the predicted AGB using unmanned aerial vehicle (UAV)-derived light detection and ranging (LiDAR) data.

The relationship between the measured AGB and the predicted AGB using UAV-derived LiDAR data under two water regimes: well-watered (WW) and natural moderate drought (mDr). Grey zone represents confidence interval (95%) of the fitted line. Dashed line is the 1:1 line.

*2.3.3 Heights and biomass of three *A. donax* crop ecotypes under natural moderate drought*

The crop height manual measurement in WW treatment ranged from 226.5 cm for EcoB to 369.3 cm in EcoA with a mean of 307.5 cm, while the mDr treatment varied from 237.8 cm in EcoB to 350 cm in EcoA with a mean of 298.2 cm (Figure 2.10a,b). ANOVA test showed that the differences between the ecotypes were significant, $F(2,12) = 24.47$, $p\text{-value} < 0.001$. No differences were observed across treatments, with $F(1,12) = 0.38$, $p\text{-value} = 0.54$; the interaction of ecotype with water treatment was also not significant, $F(2,12) = 0.48$, $p\text{-value} = 0.63$. Doing a pairwise comparison we found that EcoB is significantly shorter than the other two ecotypes, $p\text{-value} < 0.001$ (Figure 2.10a). For the LiDAR estimated crop height, we also found similar results, but in addition, doing the pairwise comparisons we found a significant effect of drought treatment on the height of EcoA ($p\text{-value} < 0.05$).

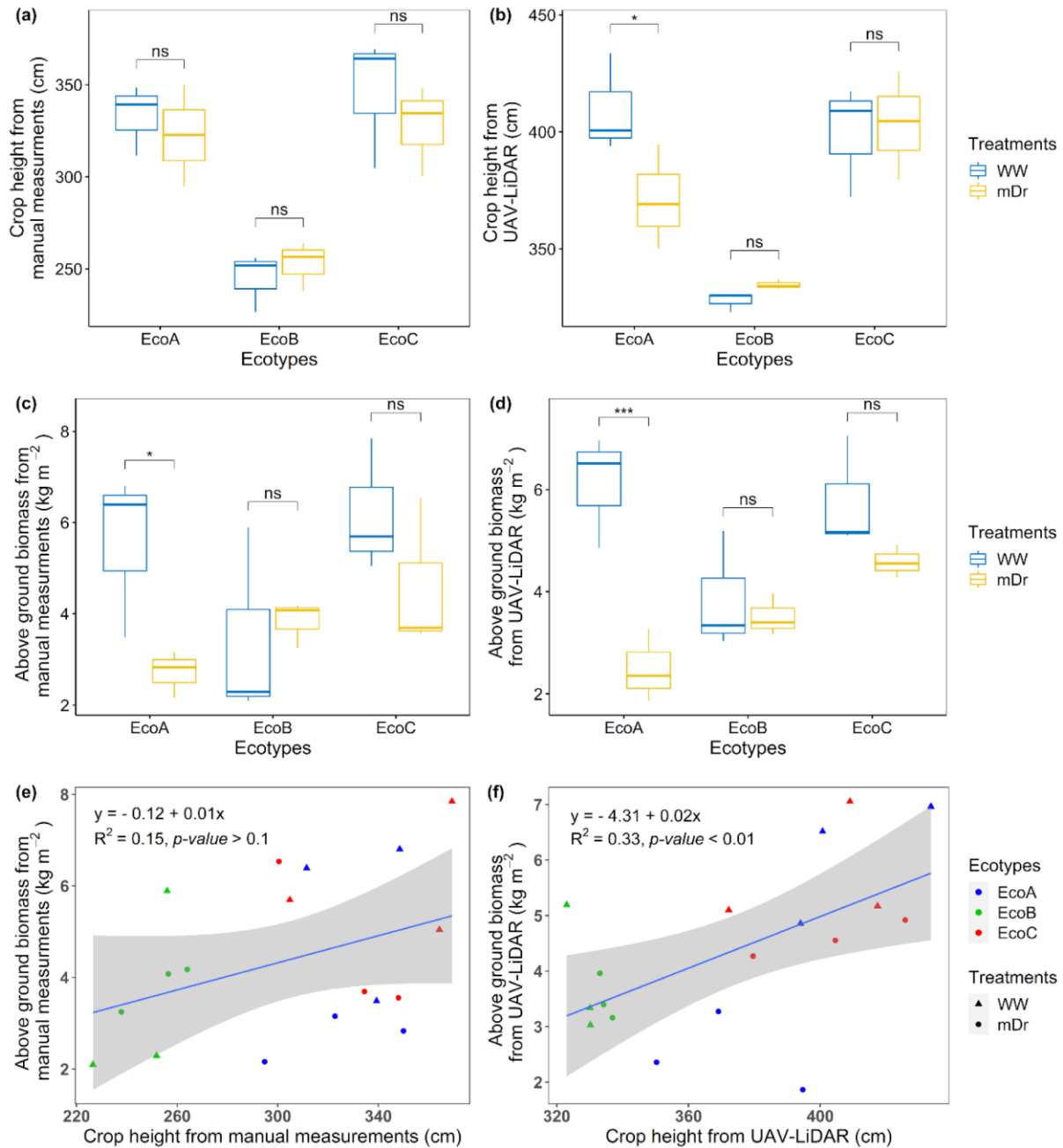


Figure 0.10. Crop height and above-ground biomass (AGB) of *Arundo donax* as estimated from manual ground measurements and light detection and ranging (LiDAR).

Crop height and AGB of *A. donax* as estimated from manual ground measurements and LiDAR under two water regimes: well-watered (WW) and natural moderate drought (mDr). (a) Crop height from manual measurements for EcoA, EcoB, and EcoC. (b) Crop height estimated from LiDAR point cloud for EcoA, EcoB, and EcoC. (c) AGB estimated from manual measurements for EcoA, EcoB, and EcoC. (d) AGB estimated from LiDAR point cloud for EcoA, EcoB, and EcoC. (e) Relationship between manually measured crop heights and AGB. (f) Relationship between LiDAR estimated crop heights and AGB. Bars represent mean $\pm$ standard deviation ($n = 3$). * $p\text{-value} \leq 0.05$, *** $p\text{-value} \leq 0.001$ in analysis of variance (ANOVA). Grey zones represent confidence interval (95%) of the fitted lines.

The AGB, obtained from the harvesting of stems in each treatment, ranged from a minimum of 2096.3 g m⁻² (EcoB in WW) to a maximum of 7845.9 g m⁻² (EcoC in WW) with an average of 4837.6 g m⁻². ANOVA tests showed mainly no significance between ecotypes, water availability treatments and their interaction. We did find, however, a significant difference in EcoA AGB between the treatments p -value < 0.05 (Figure 2.10c). Finally, with regards to LiDAR estimated biomass, we found significant effect of ecotype, drought treatment, and their interaction; $F(2,12) = 4.43$, p -value < 0.05, $F(1,12) = 17.28$, p -value = 0.001, $F(2,12) = 5.61$, p -value < 0.05, respectively (Figure 2.10d). EcoA was again much more productive in the WW plot. Ecotypes situated in WW were on average significantly more productive than ecotypes in mDr. EcoB was found to be less productive than the other two ecotypes; however, because of the interaction between ecotype and treatment, EcoB was not showing any sign of stress in mDr, while EcoC did show some insignificant stress effect (Figure 2.10d).

Looking at crop biomass and crop height, there was a positive but non-significant relationship in the manually measured dataset (Figure 2.10e). This relationship was found to be significant when looking at LiDAR data (Figure 2.10f).

2.4 Discussion

To date, LiDAR-derived metrics have been used to predict attributes such as AGB in forests and crops (Coomes et al., 2017; Knapp et al., 2018; Madec et al., 2017; Nunes et al., 2017; Shao et al., 2018; ten Harkel et al., 2020; Walter et al., 2019; Wang et al., 2016). Given the lack of fast and non-destructive alternatives, the capacity to estimate AGB using UAV-derived LiDAR metrics is a critical outcome for dense bioenergy crops such as *A. donax*. Whilst physical sampling is the most accurate method for estimation of AGB, it is invasive, destructive, costly, and can only be used to study the time course of AGB development in large plots with extensive sampling. The UAV-mounted LiDAR sensor was developed for large field experiments and breeding trials and its deployment has clear advantages over current practice (Harfouche et al., 2019). It is easily transported to any required location overcoming a major limitation of fixed phenotyping platforms. It is significantly more flexible than using airplanes where experiments are constrained by location, cloudiness, and expense (Kirchgessner et al., 2017; Virlet et al., 2017). This study has demonstrated for the first time the practicality of this approach for the bioenergy crop, *A. donax*.

2.4.1 Crop height estimation

Crop heights derived from the LiDAR measurements in *A. donax* had an R^2 of 0.73 with manual height measurements (Figure 2.8). This result is similar to R^2 for other crop height measurements estimated from LiDAR, of sugar beet ($R^2 = 0.70$), wheat ($R^2 = 0.78$), potato ($R^2 = 0.50$) and sorghum ($R^2 = 0.63$) (ten Harkel et al., 2020). Although these results demonstrate that LiDAR-derived height can estimate the differences in crop heights across ecotypes, a systematic difference between the manual stem height measurement and the UAV-measured height was observed (Figure 2.8). One possible explanation for the differing results is that the complex canopy surface makes it difficult to compute a high-quality point cloud (Thapa et al., 2018). Moreover, in practice, most LiDAR returns are likely to come from leaves rather than the denser structures containing more of the biomass such as the stems. As previously mentioned, measuring *A. donax* plant height on the ground by eye is challenging because of the density and height of the stems, and hence to avoid bias in measurement, we standardized the process by measuring each stem from the basal node to the top node. Another explanation is that the LiDAR ground penetration is hindered by the thickness of the canopy (Donoghue et al., 2007; Walter et al., 2019) or by the footprint of the LiDAR itself.

2.4.2 AGB estimation

Linear relations between individual LiDAR-derived-CHM metrics and biomass had low determination coefficients and high mean square errors (Table 2.2). As previously mentioned, these might be related to the poor penetration of the laser within the canopy (Donoghue et al., 2007; Walter et al., 2019). However, an individual metric is unlikely to adequately represent the complex canopy structure. There are also other reports of low correlations between LiDAR and manual measurements (Jimenez-Berni et al., 2018). Such limitation could be reduced by using a LiDAR with a smaller footprint to better penetrate the leafy biomass. Another way to improve biomass prediction is to take into consideration several other crop cover metrics, which may represent other features of canopy structure. Here, from the eleven LiDAR-derived metrics, Hmax, Hmajority, and Hvariety were together the best variables to include in a multiple regression for biomass estimation (Equation (2.1)) with good coefficient of determination ($R^2 = 0.71$) and low bias (bias = 3.2) (Figure 2.9). In the 1 m² evaluated

these variables represent respectively the maximum height captured, the mode of the points and the count of distinct values. As these variables greatly improved the estimate, they may be recovering other features that contribute to canopy volume. While Hmax contributed the most to prediction of biomass, Hmajority and Hvariety individually correlated negatively with the dry biomass estimation, meaning that tall stems contribute more to the yield than the height with the most occurrences in the crop, and that crops which distribute resources more uniformly across stems (smaller Hvariety) yield more biomass. The positive linear relationship obtained is higher than that found for sugar beet ($R^2 = 0.68$), lower than that for the much denser crop of winter wheat ($R^2 = 0.82$), better than that for potato ($R^2 = 0.24$) (ten Harkel et al., 2020), and similar to that obtained for different nitrogen treatments in barley ($R^2 = 0.71$) (Bendig et al., 2015).

In our study, *A. donax* biomass was measured in November, compared with the earlier acquisition of LiDAR data in late June. Since stem height mostly contributes to AGB, several studies have shown that *A. donax* reaches the height plateau around July–August (Acharya et al., 2018; Haworth et al., 2019; Yang et al., 2017). In addition, to maximize biomass yield, *A. donax* harvesting in North Italy is usually done in November just before the crop starts to senesce (Corno et al., 2014; Fabbrini et al., 2019). Waiting this long to collect LiDAR data is however not an option, as the inflorescences produced in August can significantly affect crop height and make it difficult to acquire accurate data manually. Despite the difference in timing between collection of height data and harvesting, multivariate LiDAR-derived biomass estimates related well to the dry weight (Figure 2.9). Biologically this does not come as a surprise, as stems produced in June are the most productive in terms of yield and later stems produced are smaller. If this relationship proves robust and can be established before flowering, it may have useful predictive ability for selecting suitable material for crossing in breeding programs.

Finally, the relationship between crop biomass and height was significantly better when looking at LiDAR estimates than at manual measurements (Figure 2.10e,f). The difference between the measured biomass and the estimated biomass can be explained by number of stems weighed (four stems) as opposed to the number of stems present in the 1 m² detected by the LiDAR. The difference between biomass–height relationship, on the other hand, is probably due to the inaccuracy of human

judgment in measuring crop aspects as opposed to LiDAR biomass being derived from the LiDAR crop heights. This is a specific problem in this very tall (3.2 to 4.2 m) dense crop where it is difficult to see the top of the stems, making it really hard to measure their heights without introducing errors. A greater number of plants should be surveyed by LiDAR with accompanying ground truth data in future studies to improve the crop biomass-height relationship and determine whether it is linear or exponential for *A. donax*.

2.4.3 Comparison of Ecotypes

These data confirm that EcoB, originally from a temperate oceanic climate, was significantly shorter than the other two Mediterranean ecotypes of *A. donax* (Figure 2.8a). Significant differences in height and biomass were detected in EcoA between WW and mDr from LiDAR estimates (Figure 2.10b). While ground height measurement showed no effect of drought stress on EcoA (Figure 2.10a), manually measured biomass did (Figure 2.10c). This difference in EcoA response to drought stress was also much more significant when looking at LiDAR estimated biomass (Figure 2.10b,d). This suggests that LiDAR phenotyping was better at determining differences in biomass and height than manual measurements, and LiDAR is currently the only non-destructive alternative. The much higher spatial sampling possible in the LiDAR derived CHM compared with the manual sampling makes LiDAR the preferred approach. Similar height measurement inaccuracy was reported in sorghum (Waliman and Zakhori, 2019), and it suggests that it is difficult without a very large investment in time and care to record accurately the distribution of height in these tall field crops.

The lack of differences in EcoB and EcoC between treatments suggests that these *A. donax* ecotypes are more tolerant than EcoA to moderate drought stress applied during the experiment. There are other reports of stress tolerance in *A. donax* (Ahrar et al., 2017; Cano-Ruiz et al., 2020; Pompeiano et al., 2017; Sánchez et al., 2015), hence its suitability for growth on marginal lands. Future work will aim to increase drought intensity and duration to determine how much is *A. donax* growth reduced at different drought stress levels and which ecotypes are more tolerant to these stress levels.

Our findings demonstrate that UAV-LiDAR is able to capture phenotypic differences across the three *A. donax* ecotypes, and therefore can be applied to larger scale than

currently performed experimental trials monitoring the productivity of this bioenergy crop under different environmental conditions e.g., (Andreu-Rodríguez et al., 2017). The differences in productivity suggest that EcoA is much more productive than EcoB and EcoC. However, given EcoA's low tolerance to drought, it is clear that it might not be the most suitable for bioenergy crop production on marginal lands. EcoB, on the other hand, had the lowest biomass but showed no difference in performance between the two treatments, which suggest that EcoB is more adapted to water scarcity in drought-prone environments, can grow on marginal lands and might yield more biomass than EcoA under similar soil conditions. EcoC showed intermediate performance for biomass production and drought tolerance, indicating a potential for bioenergy crop production on marginal lands. More research on these ecotypes is needed to gain a deeper understanding of the contributions of phenology, physiology, morphology to biomass production in this important bioenergy crop. However, depending on whether we want to maximize biomass yield or drought tolerance, the ecotypes studied demonstrate a huge potential for each scenario.

2.4.4 Implications

Future work will use more ecotypes at several locations and time points to further verify the robustness of the positive linear relationship between LiDAR estimates and biomass or other biophysical traits at various growth stages in response to environmental stresses. Sensors providing hyperspectral and thermal data, for example, will also be tested for their ability to contribute to novel associations between the phenotype and genotype of *A. donax* that are of interest to breeders and growers. Monitoring the productivity during the growing season can also be very important to optimize the biomass yield. Although the results in crop height and biomass estimation have demonstrated the potential of the UAV-LiDAR platform, some limitations are noted and should be worked on to increase utility of the phenotyping platform. For instance, UAV-LiDAR derived height measurements systematically overestimated the height of the *A. donax*. This is acceptable for comparison of crop heights as used in many breeding experiments; however, for absolute height measurements and more accurate biomass estimation, developing robust calibration with ground data should be the eventual aim.

2.5 Conclusion

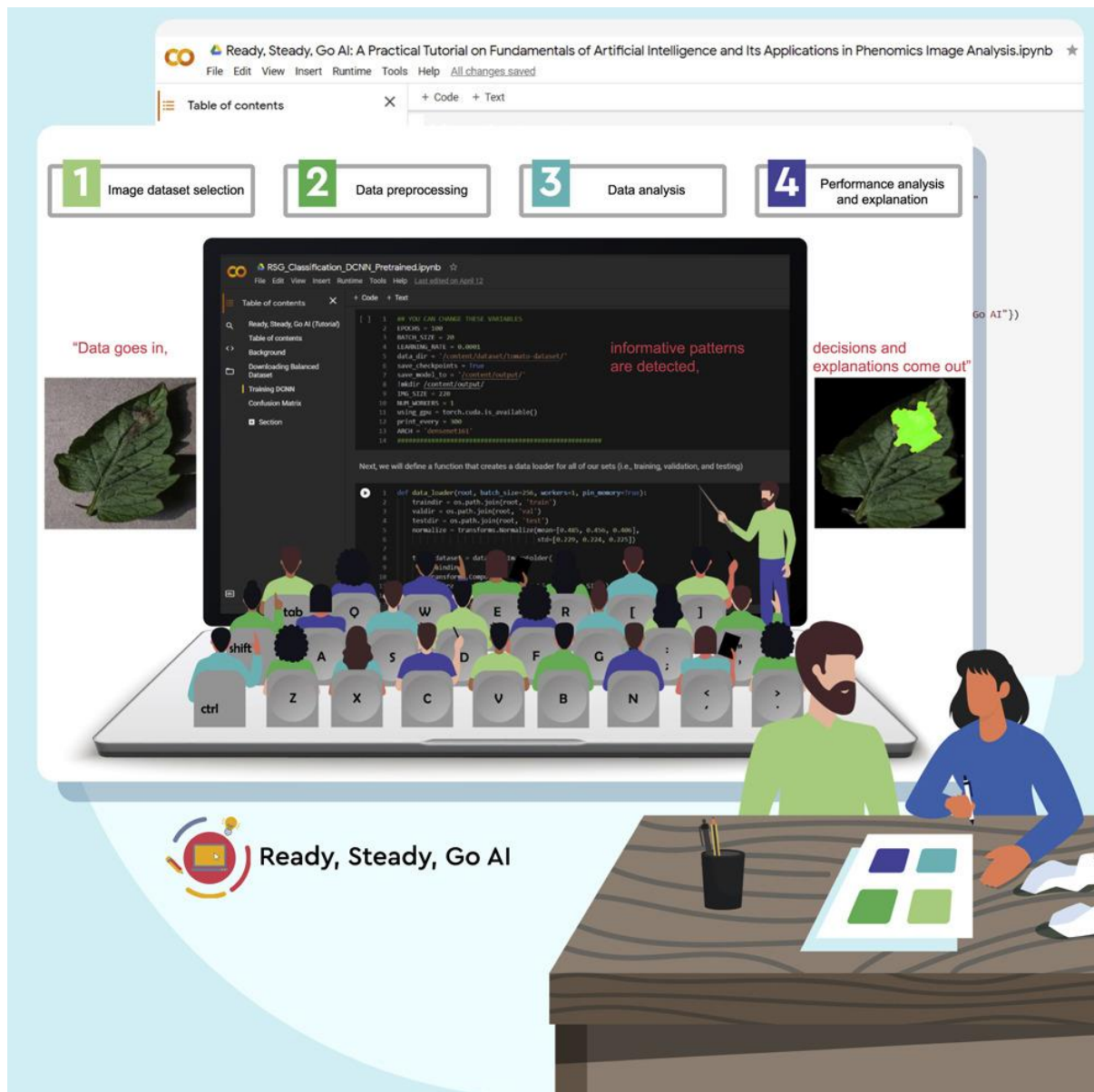
Recent efforts in *A. donax* crop improvement are focused on the measurement of AGB and crop height as determinants of greater biomass yield. We demonstrated that the present methodology with the UAV-mounted LiDAR sensor is particularly useful for high-throughput phenotyping of *A. donax* in the field; it has a high potential for advancing *A. donax* crop phenomics and genomics-assisted breeding.

Our findings could help biomass crop breeders and researchers to reduce the gap between phenomics and genomics by generating accurate *A. donax* information in an efficient, non-destructive, and inexpensive way when compared to traditional laborious approaches. The methodology described also has potential for data mining to select the most promising ecotypes, making it possible to assess specific multiple criteria for ranking ecotypes, one of the main techniques used in crop breeding. This would be a valuable extension of the crop height and AGB successfully estimated from LiDAR in *A. donax*. We were able to capture differences in crop heights between treatments and between ecotypes that manual measurements were unable to detect. This improvement was probably related to the unbiased sampling at 30 or more points per m² of the crop compared with the four stems selected and measured manually. These results are the first demonstration of the suitability of this UAV-based workflow for *A. donax* field phenotyping. Our findings show that UAV-based LiDAR scanning can replace traditional field-based crop surveys of mean height and AGB of different ecotypes and under different environmental conditions. This methodological and technological tool could be adapted to provide technical support to promote commercially viable applications of UAVs in crop phenotyping of other bioenergy crops (e.g., miscanthus and switchgrass) for other research programs and breeding companies.

Future investigations could be focused on the development of comparative analyses of *A. donax* ecotypes to test their growth habits and AGB yields over multiple environments, locations, and years to select the best performing and stable ecotypes. Considering that our research study was on three *A. donax* ecotypes stressed at one time point, next work could focus on time-series measurements of a large ecotypic panel, since breeders may need to study a large pool of plant materials and produce ecotypes that are also suitable for use in multiple cropping systems.

Chapter 3 Ready, steady, go AI: a practical tutorial on fundamentals of artificial intelligence and its applications in phenomics image analysis

Graphical Abstract



Graphical Abstract 2. Democratizing AI and AI research is a potential solution to the challenge of turning phenomic big data into insights.

Abstract

High-throughput image-based technologies are now widely used in the rapidly developing field of digital phenomics and are generating ever-increasing amounts and diversity of data. Artificial intelligence (AI) is becoming a game changer in turning the vast seas of data into valuable predictions and insights. However, this requires specialized programming skills and an in-depth understanding of machine learning, deep learning, and ensemble learning algorithms. Here, we attempt to methodically review the usage of different tools, technologies, and services available to the phenomics data community and show how they can be applied to selected problems in explainable AI-based image analysis. This tutorial provides practical and useful resources for novices and experts to harness the potential of the phenomic data in explainable AI-led breeding programs.

3.1 Introduction

In his tutorial memoranda introducing B and C programming languages in 1972 and 1975, respectively, Brian Kernighan wrote an example code that prints “hello, world!” on the computer terminal (Kernighan, 1975, 1972). In 1978, he reused it in “The C Programming Language” book, a worldwide bestseller, which he co-authored with Dennis Ritchie, who originally designed and implemented the language (Kernighan and Ritchie, 1978). Ever since, “hello, world!” has become the first program taught when learning a programming language with a purpose to illustrate its basic syntax and verify that the programming environment is properly set up. Seeing these two words on the screen means the code can compile and run correctly.

While this example is used as an illustration for almost every programming language introduced after C, it is also being incorporated in tutorials for older programming languages, such as list processor (LISP), the first high-level functional programming language, designed by John McCarthy in 1958 (McCarthy, 1960); LISP became predominant over the information processing language (IPL), the first programming language tailored for artificial intelligence (AI) programming, released in the style of low-level assembly language in 1956 (Newell and Shaw, 1957; Yeung and Yeung, 1985).

Although IPL was introduced in 1956, the roots of AI date back to 1947 when Alan Turing delivered what is perhaps the earliest public lecture on computer intelligence (Copeland and Proudfoot, 2009). In 1950, his paper discussed how to build intelligent machines and test their intelligence (Turing, 1950). Five years later, Allen Newell, Clifford Shaw, and Herbert Simon wrote a theorem-proving program, known as the logic theorist, designed to mimic human problem-solving skills (Newell and Simon, 1956). In 1956, their program, often cited as the first AI program (Simon, 1993), was presented at the Dartmouth summer research project on AI conference hosted by John McCarthy and Marvin Minsky (Moor, 2006) during which McCarthy coined the term AI.

AI continued to flourish until 1973, when the UK science research council published a report, known as the Lighthill report, criticizing AI research for not delivering impactful discoveries (Lighthill, 1973). This caused a loss of confidence and henceforth a decline in resources for AI research (Agar, 2020; Olazaran, 1993), starting a period known as the AI winter (Kaul et al., 2020; Zadpoor, 2019). In the 1980s, with Edward

Feigenbaum introducing expert systems (Delipetrev et al., 2020; Feigenbaum, 1992), and the Japanese government funding the fifth-generation computer project (Feigenbaum and Shrobe, 1993), the AI winter ended. The success, however, did not last long as expert systems proved too expensive to maintain and difficult to update (Delipetrev et al., 2020). Similarly, the Japanese project failed to meet its goals (Feigenbaum and Shrobe, 1993), marking the second AI winter in the 1990s (Delipetrev et al., 2020). Despite this, many researchers continued working with AI under other discipline-specific names, including cognitive systems and intelligent systems (Delipetrev et al., 2020). AI continued to advance with the rapid development of computer architectures. Notably, in 1997, International Business Machines (IBM)'s Deep Blue, a chess-playing computer program, defeated world chess champion and grandmaster Gary Kasparov for the first time (Hassabis, 2017). In 2012, the AI-based image analysis research marked a significant milestone when Alex Krizhevsky and colleagues published a highly influential paper introducing AlexNet, an AI algorithm that dramatically reduced image analysis error rates and won the ImageNet Large Scale Visual Recognition Challenge with 10.8% fewer errors than that of the runner-up (Krizhevsky et al., 2017). Next, in 2017, Google's AlphaGo was able to beat Chinese Go champion, Ke Jie (Chao et al., 2018). The program simulated millions of Go games by playing against itself and progressively learned the game, accumulating thousands of years of human knowledge during a period of just a few days (Silver et al., 2017).

Deep Blue and AlphaGo successfully illustrated how data can contribute to the improvement of AI. As data and AI complement each other, AI-based predictions become better the more quality data the AI is given, but this makes AI programs computationally heavy. In fact, some datasets are too large to fit in computers' memories. Michael Cox and David Ellsworth first coined the term “big data” in 1997 to describe such a scenario (Cox and Ellsworth, 1997). Afterward, big data gets more complex and challenging due not only to its volume but also to field-specific attributes. In this context, Harfouche et al. (Harfouche et al., 2019) expanded the dimensions of omics big data to eight Vs: volume, velocity, variety, variability, visibility, value, veracity, and vexing. Nowadays, graphical processing units (GPUs) are accelerating AI with their ability to process multiple tasks simultaneously due to their large number of processing units and high memory bandwidth. With supercomputers breaking the

exascale barrier (performing a quintillion floating-point operations per second) (Mann, 2020; Streich et al., 2020) and the ability of AI to analyze massive amounts of data, researchers can take advantage of big data, rather than relying on smaller representative samples, to extract knowledge and insights without loss of information.

To this end, this tutorial aims to introduce the basic principles for implementing AI and explainable AI (X-AI) algorithms in image-based data analysis (Figure 3.1) using the PlantVillage dataset as a case study to accurately identify and classify tomato leaf diseases and spider mites, as well as to explain which visual symptoms are used to make predictions. Tomato is the most economically important vegetable crop grown all over the world, but it is highly affected by devastating diseases that hinder its production. This interactive online tutorial provides students and early career researchers who seek an understanding of this rapidly evolving field and want to effectively learn to extract biologically meaningful information from imaging data and the underlying explanation with an easy-to-follow four-step workflow by using existing AI algorithms (Figure 3.2). It is linked to multiple AI resources that are relevant for experts who are interested in applying AI approaches in phenomics and wish to make use of our open computational **notebooks** to design and run their own projects. Topics covered include (1) tools, technologies, and services available to the plant data science community; (2) how to automate leaf detection, cropping, and segmentation; (3) how to balance a dataset by means of oversampling and undersampling; (4) how to classify leaf images as healthy, diseased, or spider mites damaged; (5) how to validate the model; and (6) how to generate explanations. As with any new methodology, it is advisable to seek expert assistance and advice before diving too deep.

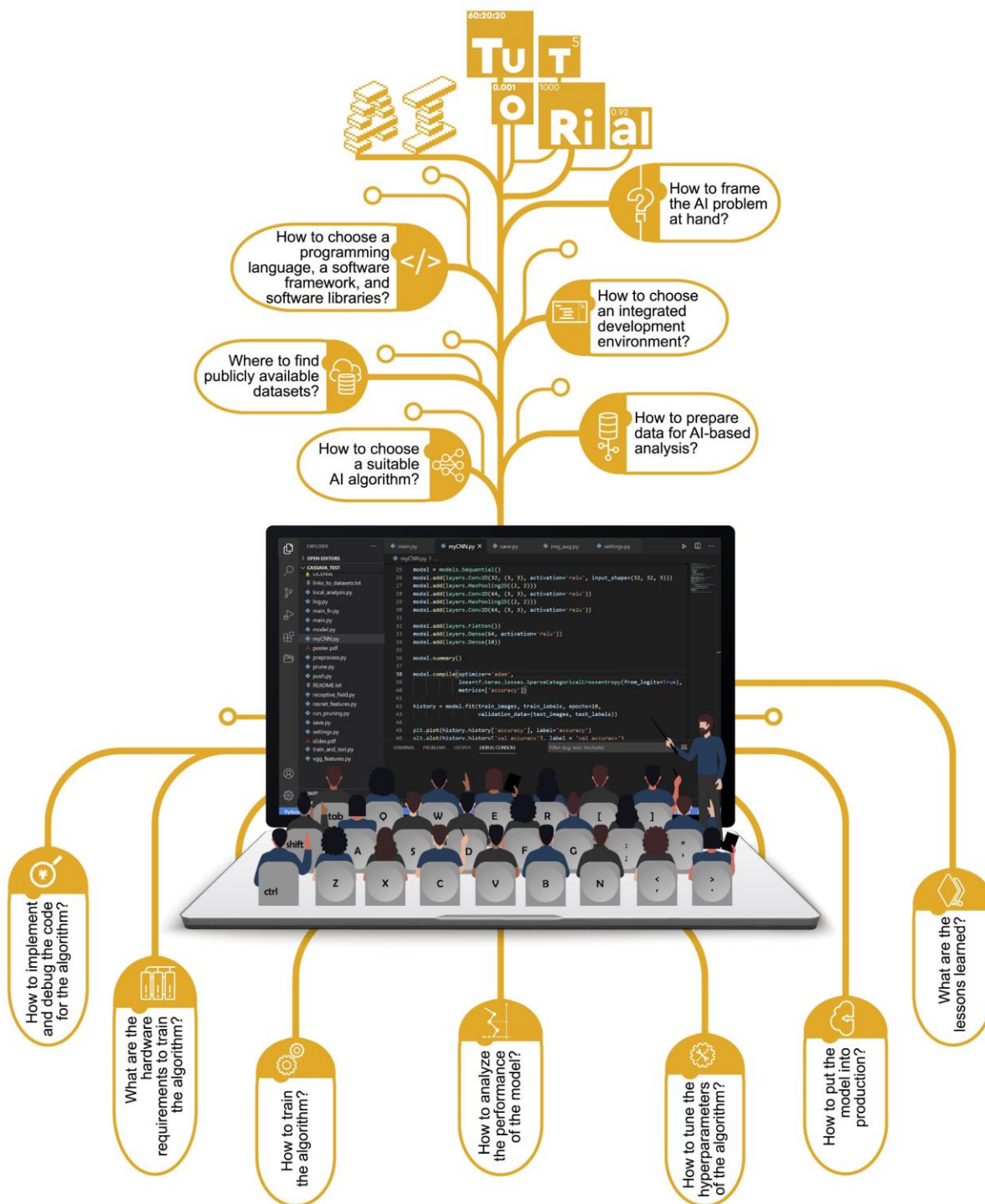


Figure 0.1. Ready, Steady, Go AI: A practical AI tutorial for the plant data science community.

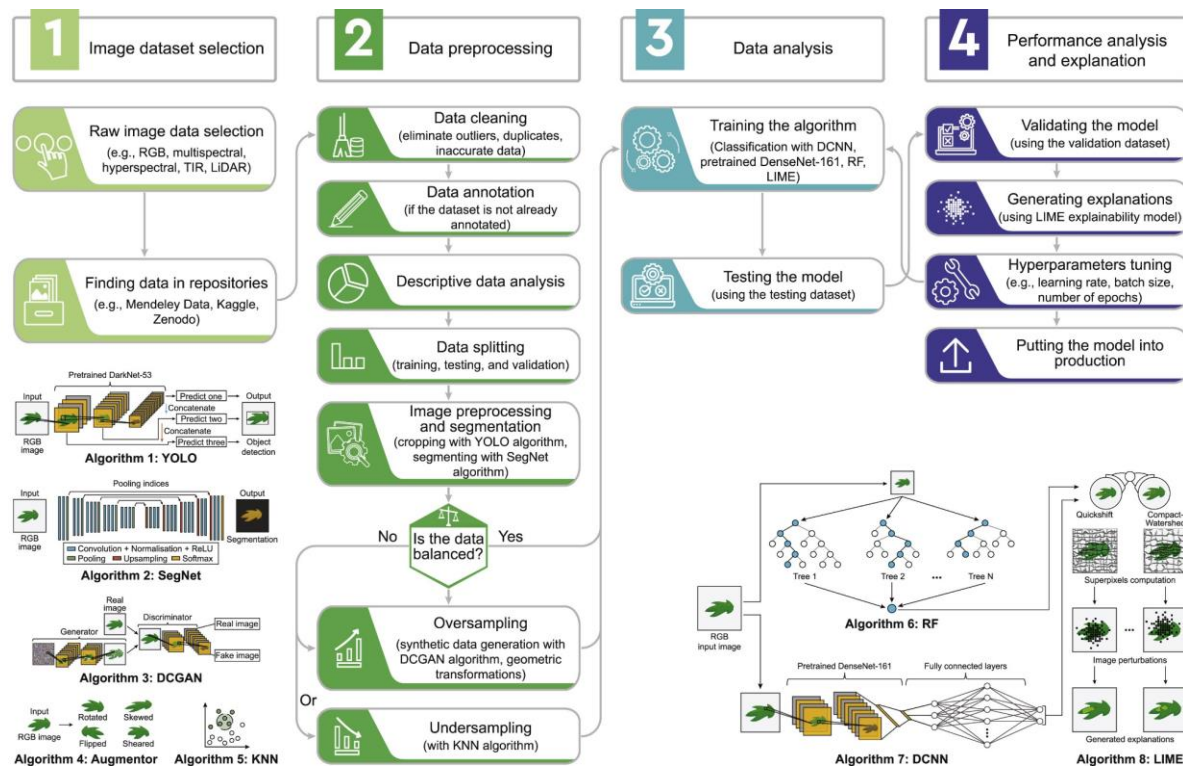


Figure 0.2. Schematic overview of the proposed X-AI-based image analysis workflow for a typical digital phenomics experiment from selection through preprocessing, analysis, and explanation.

3.2 Getting ready

Phenomics, the comprehensive large-scale study of high-dimensional phenotypes, is essential to obtaining detailed data of each major aspect of the phenotype and to better understanding plant biology and improve crops (Furbank and Tester, 2011; Tardieu et al., 2017; Yang et al., 2020). However, learning how to use such data more efficiently, easily, and on a large scale is a key challenge to overcome if the implementation of phenomics to improve on-farm response to climate change is to become a reality (Harfouche et al., 2019; Streich et al., 2020).

Digital phenomics is at the forefront of this tour-de-force effort. Here, we define it as the use of phenome data and metadata to guide decisions along the entire data analytics cycle. More broadly, it cycles between data collection and/or selection, preprocessing, analysis, and interpretation to produce new knowledge and, consequently, societal **benefits**. Importantly, as imaging is a promising source of phenotypic data, digital phenomics is not restricted to image analysis but can span the environment-to-phenotype map, which allowed the exploration of a wide range of

organisms' phenotypic responses to environmental conditions (Xue et al., 2019), as well as the genotype-to-phenotype map, which describes how genetic variation affects phenotypes (Ahnert, 2017; Alberch, 1991; Pigliucci, 2010). These aspects of digital phenomics are crucial to accelerate breeding.

Two of the main priorities for digital phenomics are (1) to develop the tools and resources that facilitate the data analytics cycle, and (2) to deploy digital platform technologies that provide powerful solutions not just to researchers but, more importantly, to breeders and farmers. The need for a bigger-science approach is most apparent in the development of digital phenomics approaches in which digitalization of phenotyping enables, improves, and transforms research design, innovation models, and entrepreneurship activities, by leveraging digital technologies and data.

Data analytics is crucial for working with large-scale phenomics data. Traditional statistical methods are limited in analyzing high-dimensional data and are unable to capture complex relationships between the predictor (independent) and response (dependent) variables (Bzdok et al., 2018). The advancement of computing technologies and power as well as AI-driven data analytics offer the possibility that digitalization of phenomics can provide new solutions to these complex challenges. Crucially, AI, in its various techniques (e.g., machine learning [ML], deep learning [DL], and ensemble learning [EL]), could be an alternative approach. For example, AI has shown impressive results in plant stress phenotyping, as reviewed by Singh and coworkers (Singh et al., 2021, 2018), and in predicting crop yield (Chlingaryan et al., 2018; van Klompenburg et al., 2020) and plant phenotypes (Azodi et al., 2020a; Liu et al., 2019; Orhobor et al., 2020). However, complex AI models typically operate as black boxes and require a leap of faith to believe their predictions (Azodi et al., 2020b; Harfouche et al., 2019). Fortunately, efforts to develop approaches that make their inner workings understandable to humans have improved our ability to explain their predictions (Azodi et al., 2020b) and introduced X-AI. With X-AI, models operating as black boxes become transparent, establishing trust between humans and AI-based predictions. While a wide range of X-AI approaches and algorithms have been developed and covered by Azodi et al. (Azodi et al., 2020b), Rudin et al., (Rudin, 2019) and Barredo Arrieta et al., (Barredo Arrieta et al., 2020) this tutorial focuses on *post hoc* model-agnostic algorithms such as local interpretable model-agnostic explanations (LIME), which is the X-AI algorithm employed in this study. Such

algorithms are used to explain any black box models by extracting the relationships between their input values and predictions to approximate their predictive behavior. *Post hoc* algorithms are commonly categorized by the scope of their explanations, which can be global or local. While the former describe the overall extracted relationships based on the entire model behavior, the latter, which LIME provides, reveal the rationale behind a specific prediction (Azodi et al., 2020b; Barredo Arrieta et al., 2020).

3.3 Some basics

3.3.1. ML

ML traces back to 1943, when Warren McCulloch and Walter Pitts modeled an electrical circuit to describe how neurons in the brain might work (McCulloch and Pitts, 1943). Their research inspired the design of the artificial neural network (ANN) algorithm, in which layers of connected nodes called artificial neurons mimic how human brains process information.

ML is a multidisciplinary approach to data analysis that makes use of probability theory, statistics, decision theory, and mathematical optimization (Singh et al., 2016) to learn from data in order to extract important information, find hidden patterns, and make associations and predictions (Harfouche et al., 2019). It can be classified into either unsupervised, supervised, or semi-supervised learning. Unsupervised learning algorithms find patterns and learn meaningful relationships in unlabeled data to describe its underlying structure. They can be categorized into clustering, association, and dimensionality reduction. On the other hand, supervised learning algorithms learn from labeled data to assign a label for previously unseen data. Supervised learning can be further divided into classification, when predicting a class label, and regression, when predicting a numerical value. For example, if we were to write a classification program to distinguish between healthy and diseased tomato leaves, all we have to do is to use a large number of leaf images, each labeled as healthy or diseased, as input for a classification algorithm. From labeled images, the algorithm being trained accumulates experience and creates its own set of rules to ultimately predict the correct class of an unlabeled image. Semi-supervised learning falls in between the two, where only part of the data needs to be labeled.

Both Turing and McCarthy suggested that the combination of human knowledge with ML is central to the development of AI research (McCarthy, 1968; Turing, 1950). ML algorithms can integrate knowledge in the form of equations and logic rules to enhance learning performance (Deng et al., 2020). Consequently, methods such as logic-based ML make use of inductive logic programming (Muggleton, 1991) to integrate information and patterns extracted from data with human knowledge in order to make predictions (Cropper and Morel, 2021; Muggleton and Marginean, 2000).

Various ML algorithms have been developed and used in plant phenotyping, as reviewed by Singh et al. (Singh et al., 2016) and van Dijk et al. (van Dijk et al., 2021). While unsupervised, supervised, and semi-supervised learning can be useful depending on the goal to be achieved, in this tutorial, we focus on supervised learning, with an aim to provide researchers with ways to identify and classify healthy, diseased, and spider mite-damaged tomato leaf images.

3.3.2. DL

ANN has undergone significant improvements, mainly by increasing its number of layers (LeCun et al., 2015). This multilayered architecture is known as deep neural networks (DNNs). DL is a class of ML that extracts features from annotated data and trains algorithms via DNNs to provide predictions based on discovered patterns (Harfouche et al., 2019; Zou et al., 2019).

Historically, the earliest efforts in developing DL algorithms came from Alexey Ivakhnenko and Valentin Lapa in 1965, when they created the group method of data handling (GMDH) algorithm which gradually increments the number of layers of an ANN based on evaluation of its performance (Schmidhuber, 2015). In 1979, Kunihiko Fukushima designed an ANN with multiple data downsampling (pooling) and filtering (convolutional) layers called Neocognitron. It was capable of analyzing images (Fukushima, 2014, 1980) and led to the birth of deep convolutional neural networks (DCNNs). Slow progress was being made in this field until 2012 when AlexNet, trained to classify 1.2 million high-resolution images, achieved a top-five error rate of 15.3%, making a major jump in the accuracy of image analysis (Krizhevsky et al., 2017). Since then, DCNNs became the default approach for most AI-based image analysis tasks (Levine et al., 2019). Various DL algorithms have been developed and used in plant phenotyping (Singh et al., 2021, 2018).

3.3.3. *EL*

EL is a technique where a set of algorithms are individually trained to solve the same problem, having their output combined to improve predictive performance and reduce error variance (Dietterich, 2000; Harfouche et al., 2019).

Tracing the origin of EL proved difficult as the idea of training multiple algorithms and combining their predictions has been used for a long time (Yang, 2017). However, EL research was highly influenced by two main events in 1990 (Yang, 2017). The first was when Robert Schapire introduced boosting by theoretically proving that the prediction accuracy of learners that are not highly accurate, known as weak learners, can be boosted by their combination (Schapire, 1990). The second was when Lars Hansen and Peter Salamon conducted applied research and found that combining the predictions of a set of algorithms is often more accurate than using a single one (Hansen and Salamon, 1990). Later, in 1996, Leo Breiman introduced bagging, which stands for bootstrap aggregating, where the algorithms of an ensemble are trained on different randomly sampled subsets of data (Breiman, 1996).

The combination of algorithms to create an ensemble can be homogeneous, by applying the same algorithm for all combined estimators (e.g., decision trees), or heterogeneous, by combining a variety of algorithms (e.g., support vector machines with logistic regression). Notably, the combination of multiple DL algorithms is sometimes referred to as deep EL.

The remarkable flexibility and adaptability of EL helped the proliferation of its application in plant phenotyping where it has been used for various tasks, including biotic stress identification and classification (Gomez Selvaraj et al., 2020; Yuan et al., 2020), aboveground biomass estimation (Li et al., 2020; Wang et al., 2017), and performance prediction of crosses in breeding (Ansarifar et al., 2020; Shahhosseini et al., 2021).

3.4 Getting set

At the end of his speech, on February 9, 1941, Winston Churchill said “give us the tools, and we will finish the job” in an attempt to raise the morale of the British public during World War II. While the task of implementing AI algorithms is not to be

compared with war, choosing the right tools, technologies, and services is paramount to its success. However, the number of available programming languages, software frameworks, software libraries, and integrated development environments (IDEs) can be daunting. Here, we guide researchers through the options of making a suitable choice.

3.5 Programming languages and IDEs

Programming languages are tools used to write instructions for computers to follow. They fall into two categories: low level and high level. Low-level languages are close to binary; they have the advantage of being faster to execute and use less computer memory. In contrast, high-level programming languages are closer to how humans communicate; this makes them easier to use compared with low-level languages. However, they are slower to run as, prior to execution, they get translated into machine code through a compiler that scans the entire program and translates it all at once, or an interpreter, which translates just one statement at a time. However, as computing power continues to increase, the runtime difference between low-level and high-level languages becomes negligible. Examples of high-level programming languages include Python, C, and R.

A programming language is said to be Turing complete if it is possible to implement any algorithm with it, regardless of its performance (Scott, 2009). In this sense, any Turing complete programming language can be used to implement any AI algorithm. Unfortunately, a single ideal programming language does not exist. Choosing one requires consideration of all of its aspects, benefits, and shortcomings in respect to the task at hand. It is advisable to consider the ecosystem around it, such as its community, contributors, and available software frameworks and libraries (Table 3.1 and Table 3.2). It is also prudent to choose the language that is adopted in the user's professional environment, which increases code reusability. Well-established, stable languages are often a convenient choice. For example, Julia combines the interactivity of scripting languages, such as Python and R, with the speed of compiled languages such as C (Perkel, 2019). It was originally designed for high-performance scientific computing and data analysis (Gao et al., 2020). However, being in early-stage development, Julia has a limited amount of software frameworks and libraries (Gao et al., 2020). Its influence is far less than that of other popular programming languages



such as Python and R, which are considered the standard tools of the trade in research and industry, with Python being a more popular choice among biologists (Grabowski and Rappsilber, 2019); they both have an easy-to-learn syntax and have no major speed differences when it comes to typical data operations (Grabowski and Rappsilber, 2019).

Table 3.1. Representative list of available open-source software frameworks and programming languages for the implementation of AI algorithms.

Software framework	Programming languages	Operating systems	GPU computing support ^{BS}	Distributed training support	Availability of pretrained models ^b	Software license	Website
			D, Berkeley Software Distribution; BSL, Boost Software License; GPL, GNU General Public License; GPU, graphics processing unit; LGPL, GNU Lesser General Public License; MIT, Massachusetts Institute of Technology				
Deep ML frameworks ^c							
Tensorflow	& Python,	Linux,	Linux,	multiple GPUs, yes		Apache	tensorflow.
Keras ^d	JavaScript, C++, Java, Go, Swift (early release)	Windows, macOS	Windows	multiple machines		2.0	org

PyTorch	Python, Java	C++, Linux, Windows, macOS	Linux, Windows	multiple machines	GPUs, yes	BSD-3	pytorch.org
Apache MXNet	Python, Scala, Clojure, Java, Perl	C++, Julia, R, macOS	Linux, Windows, macOS	multiple machines	GPUs, yes	Apache 2.0	mxnet.apache.org
H2O	Python, Java, R	Scala, Linux, Windows	Linux, Windows	multiple machines	GPUs, no	Apache 2.0	h2o.ai
Deeplearning4j	Java, Scala, Clojure	Groovy, Kotlin, macOS, Android	Linux, Windows	multiple machines	GPUs, yes	Apache 2.0	deeplearning4j.org
Chainer	Python	Linux	Linux	multiple machines	GPUs, yes	MIT	chainer.org
PaddlePaddle	Python	Linux, Windows, macOS	Linux, Windows	multiple machines	GPUs, yes	Apache 2.0	paddlepaddle.org
SINGA	Python, C++	Linux, Windows, macOS	Linux, Windows	multiple machines	GPUs, no ^e	Apache 2.0	singa.apache.org
Flux	Julia	Linux	Linux	multiple	GPUs no	MIT	fluxml.ai
OpenNN	C++	Linux, Windows, macOS	Linux, Windows	no	no	LGPLv3	opennn.net
Dlib	Python, C++	Linux, Windows, macOS	Linux, Windows	multiple	GPUs no	BSL-1.0	dlib.net
MLBox	Python	Linux, Windows, macOS	no	no	no	BSD-3	github.com/AxeldeRochembay/MLBox
ML frameworks							
scikit-learn	Python	Linux, Windows, macOS	no	no	no	BSD-3	scikit-learn.org
Apache Mahout	Java, Scala	Linux, Windows, macOS	Linux, Windows	multiple machines	GPUs, no	Apache 2.0	mahout.apache.org

xLearn	Python, R	Linux, Windows, macOS	no	multiple machines	no	Apache 2.0	github.com/aksznzhy/xlearn
Shogun	Python, Octave, R, Java, Scala, Lua, C#, Ruby	Linux, Windows, macOS	no	no	no	GPLv3	shogun-toolbox.org
SystemDS	Python, Java	Linux, Windows, macOS	Linux, Windows	multiple machines	GPUs, no	Apache 2.0	systemds.apache.org

BSD, Berkeley Software Distribution; BSL, Boost Software License; GPL, GNU General Public License; GPU, graphics processing unit; LGPL, GNU Lesser General Public License; MIT, Massachusetts Institute of Technology

^aWhile all software frameworks support CPU-based analysis, some support GPU-based analysis, which can make key computations very fast, particularly convolution operations used in image analysis.

^bThe availability of pretrained models allows giving an AI model a warm start by applying information learned from another previously trained model using a process known as transfer learning. For example, frameworks such as Tensorflow with Keras, PyTorch, and PaddlePaddle allow the transfer of a pretrained DenseNet-161 model to a DCNN model, and DarkNet-53 pretrained model to a YOLO model (Figure 3.2).

^cDeep ML frameworks can support both ML and DL AI algorithms.

^dKeras is the high-level application programming interface (API) of TensorFlow, allowing users to train, evaluate, and use AI models in just a few lines of code.

^eWhile SINGA does not provide pretrained AI models, it supports transferring models in Open Neural Network Exchange (ONNX) format, an open-representation format for AI models that enables users to use them across different software frameworks.

Table 3.2. Representative list of available open-source software libraries and IDEs to aid the implementation of AI algorithms.

Software library/IDE	Programming languages	Description	Website
Software libraries			
OpenCV	Python, R, C, C++, Java, Julia, Closure	helps applying various image processing operations such as image resizing, object detection, and geometrical transformations	opencv.org
Plotly	Python, R, Julia	enables plotting charts such as histograms, bar charts, line charts, and boxplots	plotly.com
NumPyPNG, portable graphics format; TIFF, tag image file format	Python	provides multidimensional arrays and various mathematical functions and statistical operations to operate on these arrays	numpy.org
pandas	Python	offers data structures and operations that enable the analysis and manipulation of different kinds of data, such as tabular, time series, and arbitrary matrix data	pandas.pydata.org
Matplotlib ^b	Python	helps generating different types of charts such as line charts, scatterplots, heat maps, and bar charts	matplotlib.org
Seaborn	Python	based on Matplotlib, it enables data visualizations by providing a simpler API	seaborn.pydata.org
Bokeh	Python	allows creating web-based data visualizations and charts ranging from simple plots to complex data dashboards	bokeh.org
split-folders ^c	Python	allows splitting a dataset into training, validation, and test sets	github.com/jfilter/split-folders
Augmentor ^d	Python	oversamples an image dataset by applying transformations to image geometries	augmentor.readthedocs.io
DataExplorer	R	automates the descriptive data analysis process by generating basic statistics for a dataset (e.g., types of variables, missing values) and plotting distributions of variables and their correlation matrices	cran.r-project.org/web/packages/DataExplorer/
ggplot2	R	helps plotting charts such as line charts, histograms, scatterplots, density plots, boxplots, and bar charts	ggplot2.tidyverse.org
tidyr	R	provides table data structures and organizes variables in columns and values in rows	tidyr.tidyverse.org
magick	R	helps editing and converting image data from and to a variety of formats, including PNG, JPEG, and TIFF. Enables resizing and applying geometric transformation to images	cran.r-project.org/web/packages/magick/

Magick++	C++	a C++ API for the magick library	imagemagick.org/Magick++
accuracy-evaluation-cpp	C++	helps calculating performance metrics for classification algorithms, including accuracy, error rate, precision, recall, and F score	github.com/ashokpant/accuracy-evaluation-cpp
matplotlib-cpp	C++	a C++ API for the Matplotlib library	github.com/lava/matplotlib-cpp
Armadillo	C++	provides data structures and linear algebra operations for vectors, matrices, and cubes	arma.sourceforge.net
ND4J	Java	provides multidimensional arrays and various linear algebra functions to operate on them	github.com/deeplearning4j/nd4j
matplotlib4j	Java	a Java API for the Matplotlib library	github.com/sh0nk/matplotlib4j
cv4j	Java, Android	allows the application of various image filtering techniques, including the median filter for image denoising, sharp filter for image sharpening, and Laplacian filter for edge detection	github.com/imageprocessor/cv4j
Tables.jl	Julia	offers data structures and operations that enable the analysis and manipulation of tabular data	github.com/JuliaData/Tables.jl
Plots	Julia	brings plotting functionalities to Julia from several plotting libraries, including Pyplot from the Python Matplotlib library	juliaplots.org
Augmentor.jl	Julia	a Julia API for the Augmentor library	github.com/Evizero/Augmentor.jl
Vegas	Scala	helps generating different types of charts, such as line charts, scatter plots, heat maps, and bar charts	github.com/vegas-viz/Vegas
Breeze	Scala	provides data structures for vectors and matrices, and statistical functions to calculate descriptive statistics, such as mean, variance, and standard deviation	scalantlp.org
kixi.stats	Closure	generates descriptive statistics such as mean, variance, and standard deviation, and provides functions for performing statistical tests such as t test, z test, and chi-squared test	clidoc.xyz/d/kixi/stats
Neanderthal	Closure	provides data structures for vectors and matrices in addition to a variety of functions that support their manipulation	neanderthal.uncomplicate.org/
Hanami	Closure	helps plotting charts such as bar charts, contour plots, and tree layouts using the declarative languages Vega and Vega-Lite	github.com/jsa-aerial/hanami
panthera	Closure	an API that imports NumPy and pandas functionalities from Python to Closure	github.com/alanmarazzi/panthera

IDEs

Visual Studio Code	Python, R, C, C++, C#, Julia, Java	is a cross-platform code editor ^e offering code syntax highlighting, smart completion, and debugging. It helps reviewing differences in code versions and integrates with SCM services. It is extensible, providing the possibility to add new languages, themes, and debuggers	code.visualstudio.com
--------------------	------------------------------------	--	---

Eclipse	Python, R, C, C++, Java, C#, Clojure, Julia, Scala	is composed of plugins and designed to be extensible using additional plugins. It is customizable, cross-platform, and offers code syntax highlighting, code debugging, and the ability to review differences in code versions	eclipse.org
NetBeans	C, C++, Java	provides code syntax highlighting, templates, and a range of tools for code debugging and refactoring, and integration with SCM services. It is cross-platform and is able to identify and fix common problems in Java code	netbeans.org
PyCharm Community Edition	Python	provides smart code completion, error highlighting, and code debugging capabilities. It is cross-platform and customizable, providing various code editor color templates	jetbrains.com/pycharm/
IDLE	Python	offers an interactive cross-platform Python interpreter with code syntax highlighting, smart indentations, and code debugging	docs.python.org/3/library/idle.html
IntelliJ IDEA Community Edition	Java, Scala, Android	provides code syntax highlighting, debugging, tools for code refactoring, and a unified interface for integration with SCM services. It is customizable, cross-platform, and allows reviewing differences in code versions	jetbrains.com/idea
Code::Blocks	C, C++, Fortran	is composed of plugins and designed to be extensible using additional plugins. It is cross-platform and provides code syntax highlighting, smart indentations, and code debugging	codeblocks.org
KDevelop	Python, C, C++	provides code syntax highlighting and code completion. It is customizable, cross-platform, and allows code debugging. It supports integration with various SCM services	kdevelop.org
RStudio	R	offers code syntax highlighting and supports direct code execution. It is cross-platform and provides tools for debugging and viewing code execution history	rstudio.com

PNG, portable graphics format; TIFF, tag image file format

^aBy loading image data as NumPy arrays, various image processing tasks can be performed using functions provided by other software frameworks and libraries; in this tutorial, NumPy aided the implementation of the DCNN, YOLO, SegNet, and DCGAN algorithms.

^bAided the visualization of the confusion matrix used to evaluate the performance of the DCNN algorithm.

^cHelped the random splitting of PlantVillage dataset, used in this tutorial, into training, testing, and validation sets.

^dAssisted in oversampling PlantVillage to ensure the balance between different classes and avoid overfitting.

^eWhile it is a code editor, its support for development operations, including debugging and version control, and its high extensibility and customizability level allow it to be configured as an IDE.

The choice of a programming language is often complemented by the choice of an IDE. IDEs allow researchers to write codes quickly and increase their productivity by combining code editing, compiling, and debugging into a single application. The choice of IDEs is personal. The reason to pick one over the others depends on the tasks required to be performed. Common features to look for are code autocompletion, syntax highlighting, code refactoring, and more that could increase productivity (Table

3.2). For example, while PyCharm provides smart code completion, error highlighting, and code debugging capabilities for Python, Visual Studio Code supports various programming languages and can help review differences in code versions and integrate with source code management (SCM) services.

3.6 Software frameworks

Software frameworks are a working environment that helps researchers quickly implement or design algorithms without compromising quality. They provide generic codes that are tested and optimized for others to reuse, extend, and customize to their needs. Additionally, they include functions that can be used to process input, manage devices, and interact with system software. For example, using a software framework can remove the burden of writing a code that loads images from a disk drive. Instead, a pre-written code provided by the framework can be used. Various software frameworks are available to help the implementation of AI algorithms (Table 3.1).

3.7 Software libraries

Similar to software frameworks, software libraries consist of pre-written codes. They can be used to import a specific functionality without writing a code for it. For example, when implementing an AI algorithm with Python, importing NumPy provides multidimensional arrays and various mathematical functions to operate on them, eliminating the need to code such complex functionality.

Unlike software frameworks, libraries typically focus on a narrow scope where they perform a set of specific, well-defined operations. Technically, the key difference between them is known as the inversion of control (IoC), meaning that, with software libraries, researchers are in control of the code as they use library functions whenever they see fit. However, software frameworks require them to follow specific workflow and software design patterns. Table 3.2 provides a list of software libraries that can come in handy for implementing AI algorithms.

3.8 Going AI – Materials and methods

AI can potentially revolutionize many aspects of image analysis in phenomics, but it would have a greater impact when everyone can access it, including not only

computer scientists but also researchers across disciplines. While the emergence of freely available and open-source AI resources is opening the way toward wider AI accessibility, the substantial expertise required for its adoption remains a hurdle. Adopters do not need to have extensive experience in programming, but, in addition to their domain knowledge, they should be open to learning a programming language and writing some codes for their experiments. Going AI is set to avoid exposing tutees to heavy jargon and tedious definitions, but, instead, to tutor them on the concepts behind AI algorithms relevant to the task at hand and equip them with the right skills that make them comfortable using those resources. This would enable them to gain the practical know-how needed to address new problems in phenomics image analysis. When they complete this tutorial experience, they should be in a position to ask biological questions, analyze available resources, and approach several types of plant image datasets using various AI algorithms. Although a single workflow cannot be considered the best choice as different algorithms are to be assessed for different tasks and objectives, our approach is in agreement with common image-based analytical methods, along the lines of those described elsewhere (Goluguri et al., 2021; Singh et al., 2021, 2018), as essential key steps in the workflow of image analysis. The four steps of this workflow are illustrated in Figure 3.2. For each of the eight algorithms used, we created an interactive notebook runnable on Google Colab, or any other platform supporting Jupyter notebooks, and made it available on our public GitHub repository^{vi}. The repository includes the notebooks and corresponding ReadMe files that instruct researchers through the process of running our code either to reproduce the results or to run the implemented algorithms with a new dataset. A continuous integration (CI) framework using the Travis CI service was added so that unit tests, implemented using pytest, can run automatically when code updates are introduced. This provides a safeguard against updates that can break existing functionality by generating a report showing which tests have passed or failed. The coverage of the test is measured using the Coverage.py software library, which reports the amount (percentage) of code covered by existing unit tests. In addition, we have created a set of self-test quizzes and hands-on practices and exercises to provide researchers with opportunities to augment their learning by testing the knowledge they have acquired and by practically applying the concepts explained. For the quizzes, two Google Forms of multiple-choice questions (MCQs) were developed. The first

asks questions about the basics covered in this paper, while the second, intended for the more experienced researchers, includes more advanced questions and is based on real-life scenarios that we have faced while developing this tutorial. This will help researchers identify good practices and develop troubleshooting skills that can be useful when performing AI-based image analysis.

For the practices and exercises, three computational notebooks were developed to get researchers involved in coding. The first, intended for beginners, asks users to complete simple parts of a provided code in order to run the tasks included in our workflow; the second, intended for intermediates, involves users in writing more codes for the same tasks and modifying parts of it to help them practice; the third notebook is intended for the more experienced users where they are asked to write codes that reuse previously learned models to preprocess images of apple leaves included in the PlantVillage dataset and containing four classes of diseases (apple scab, black rot, cedar apple rust, and healthy). Next, the notebook requires them to train a new model that is able to identify and classify apple diseases and explain its decisions. All three notebooks contain the solutions stored in hidden code cells after each task and are available on our public repository along with links to the MCQs.

3.8.1. Step 1: Image dataset selection

Initially, it is important to frame the biological question at hand. This helps researchers identify what task their AI algorithm should perform and guide the choice of the input data. For example, red-green-blue (RGB) imaging data can be used to monitor plant stress (Francesconi et al., 2021), while light detection and ranging (LiDAR) data allow measurement of plant architecture, canopy height, and growth rates (Harfouche et al., 2019; Maesano et al., 2020).

Publicly available plant phenotyping datasets have a 2-fold advantage: on the one hand, the phenomics data science community is given access to valuable data whose collection can be both time consuming and expensive, thus providing the foundation for practical AI-based image analysis applications; on the other hand, freely available benchmarking datasets can serve as a basis to compare the performance of AI algorithms under reproducible settings.

In this tutorial, we performed our analysis using the publicly available PlantVillage tomato dataset (Pennsylvania State University, United States) (Hughes and Salathé, 2015). The dataset includes a total of $n_{\text{tomato}} = 18,160$ individually labeled and non-overlapped RGB images divided into 10 classes (Figure 3.3): class 0 (healthy, $n_h = 1,591$), class 1 (bacterial spot, $n_{bs} = 2,127$), class 2 (leaf mold, $n_{lm} = 952$), class 3 (*Septoria* leaf spot, $n_{sls} = 1,771$), class 4 (target spot, $n_{ts} = 1,404$), class 5 (early blight, $n_{eb} = 1,000$), class 6 (late blight, $n_{lb} = 1,909$), class 7 (mosaic virus, $n_{mv} = 373$), class 8 (yellow leaf curl virus, $n_{ylcv} = 5,357$), and class 9 (spider mites, $n_{sm} = 1,676$). All images were in joint photographic expert group (JPEG) format and had a standard size of 256×256 pixels.

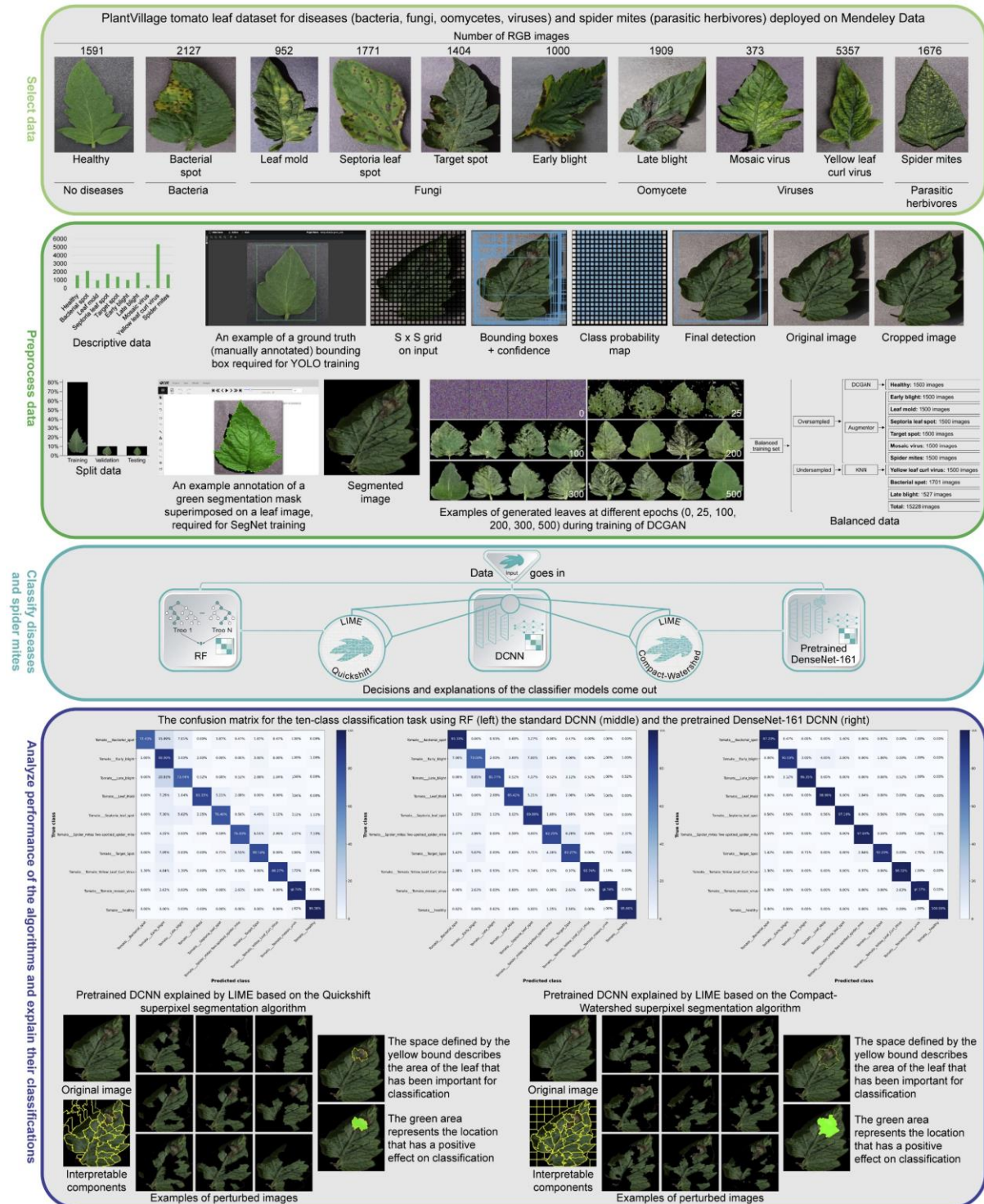


Figure 0.3. Visualization of the tutorial outputs at every step of the proposed X-AI-based image analysis workflow.

3.8.2. Step 2: Data preprocessing

Data preprocessing aims to eliminate noisy (e.g., blurred images), incomplete (e.g., unannotated images), and duplicate data from a dataset to help AI algorithms learn

relevant features that ultimately enhance their accuracy. This step starts with data cleaning, which is the process of inspecting data to remove irrelevant (e.g., images that do not contain a leaf) or wrongly annotated (e.g., diseased images labeled as healthy) images. In case images were not annotated, adding corresponding annotations to images would then be necessary. Both processes are often done manually, preferably with the help of domain experts (e.g., a plant pathologist). Indeed, the tomato dataset has been cleaned and provided with image-level annotations (i.e., assigning a class to each image) by plant pathology experts (Hughes and Salathé, 2015). Next, a descriptive data analysis must be performed to have a better understanding of data characteristics, such as its frequency distribution (i.e., the number of images in each class). A simple Python code was developed to iterate over the directories of the 10 classes, counting the number of images in each. This analysis showed that classes of tomato images are imbalanced (Figure 3.3). This calls for data balancing, done by either oversampling the minority or undersampling the majority classes. However, prior to balancing, the dataset must be split into training, validation, and testing sets to prevent overlap of oversampled images across them. While the training set is used to fit the model, the validation set evaluates it during **hyperparameters** tuning, and the test set serves to provide an unbiased evaluation of the final model. Here, we randomly allocated 80% of images for training ($n_{\text{train}} = 14,523$), 10% for validation ($n_{\text{val}} = 1,812$), and 10% for testing ($n_{\text{test}} = 1,825$), using the `split-folders` Python software library. Then, we applied both oversampling and undersampling to balance the dataset. We used two techniques for oversampling: first, by creating random transformations to image geometries using the `Augmentor` Python software library, and second, by synthesizing leaf images using a deep convolutional generative adversarial network (DCGAN) as suggested by Perez and Wang (Perez and Wang, 2017), concluding that the likely best strategy for oversampling would be to combine the geometrically transformed data with synthesized data. Undersampling was performed using the k-nearest neighbors (KNN) algorithm.

It is worth noting that datasets in public repositories may include images that contain more than one object, typically belonging to background objects (e.g., stem, overlapping leaves) that are not of interest for an AI algorithm to learn. This could negatively affect its training, but it can be solved by image cropping and/or segmentation. However, this process can take a considerable amount of time and

effort. Many types of plant image processing software, such as plant computer vision (PlantCV) (Fahlgren et al., 2015; Gehan et al., 2017), Image Harvest (Knecht et al., 2016), and HTPheno can help with these tasks and extract features from images by automating complex techniques such as histogram thresholding and color analysis (Zhou et al., 2021). In this regard, Lobet et al. (Lobet et al., 2013) developed an online database indexing both commercial and open-source plant image processing software. While such software makes it easy for researchers to get started, more advanced use cases still require them to develop their own. Here, we showed how cropping and segmentation can be automated using the you only look once (YOLO) and SegNet AI algorithms, respectively.

Finally, before diving into image preprocessing and analysis algorithms, it is fundamental to understand how computers process an RGB image. Computers denote every pixel in an image as a 1×3 matrix, representing the three colors R, G, and B with each being an 8-bit integer leading to 256 (28) possible states for every color. These states represent the brightness of each color, with 0 being the darkest and 255 being the brightest. For example, a matrix with values $R = 255$, $G = 255$, and $B = 255$ would lead to a white pixel.

3.8.2.1 Algorithm 1: YOLO

YOLO was first introduced in 2016 to perform object detection and localization in images and videos (Redmon et al., 2016). In 2017 and 2018, improved versions of the algorithm were released, namely YOLOv2 (Redmon and Farhadi, 2017) and YOLOv3 (Redmon and Farhadi, 2018), respectively. YOLO starts by dividing the input image into an $S \times S$ grid, forming S^2 cells (Figure 3.3). For every object in the image, for example a leaf, the grid cell associated with its center is responsible for detecting it. This is done by having the cell predict a number of rectangles that enclose the object, known as bounding boxes. These predictions have five components: x , y , w , h , and confidence. While x and y are the coordinates representing the center of the bounding box, w and h represent the width and height measures that describe its dimensions. Finally, the confidence score is equal to the intersection over union (IOU) between the predicted and the ground truth (manually annotated) bounding box (Figure 3.3).

When tested on the Microsoft common objects in context (COCO) large-scale object detection dataset (Lin et al., 2014), YOLOv3 was able to detect objects faster and more accurately than other relevant object detection algorithms, including faster region-CNN (Faster R-CNN), single shot multibox detector (SSD), and RetinaNet, scoring the highest mean average precision (mAP) in the shortest time (Redmon and Farhadi, 2018).

We used YOLOv3 to automatically detect, localize, and crop the leaves. We randomly selected 1,000 images from all 10 classes and manually annotated them by drawing bounding boxes around the leaves using the MakeSense annotation tool. These images were then used to train YOLOv3, which uses a DCNN architecture (see algorithm 7) named DarkNet-53, pretrained on the ImageNet dataset. An example architecture of YOLO is illustrated in Figure 3.2. After the algorithm has completed 300 training epochs (iterated over the annotated images 300 times), the learned model was used to detect and localize leaves for the rest of the dataset. Once a final detection and a class probability map were generated for each image (Figure 3.3), a custom code was used to crop it, keeping only the leaf.

3.8.2.2 *Algorithm 2: SegNet*

SegNet is a DCNN-based algorithm (see algorithm 7) that uses a modified visual geometry group (VGG)-16 architecture to perform pixel-wise image segmentation (Badrinarayanan et al., 2017), which involves assigning a class to every pixel in an image. SegNet is composed of encoders and decoders where the encoder network consists of 13 convolutional layers followed by their corresponding decoding layers forming the decoder network. The encoder-decoder network is finally followed by a classification layer that uses the softmax activation function. While the encoder network performs convolutions with various filters to produce a set of feature maps, the corresponding decoders upsample them and generate a segmentation mask for the image. Finally, the softmax classifier assigns a class to each pixel. The described SegNet architecture is illustrated in Figure 3.2.

Quantitative comparison of deep networks for semantic pixel-wise image segmentation using the Cambridge-driving labeled video database (CamVid) (Brostow et al., 2009) showed that SegNet learned to perform better in a shorter time than

competing algorithms such as DeepLab, fully convolutional network (FCN), and deconvolution network (DeconvNet) (Badrinarayanan et al., 2017).

In this tutorial, 150 images, chosen randomly from the 10 classes, were annotated with segmentation masks using the computer vision annotation tool to train SegNet on leaf segmentation (Figure 3.3). The trained model was then used to segment the remaining images by deleting the background around the leaves (Figure 3.3).

3.8.2.3 *Algorithm 3: DCGAN*

DCGAN is an extension of the generative adversarial network (GAN) that was designed in 2014 to generate new data instances that resemble training data (Goodfellow et al., 2014). The idea of GAN was to pair two learning models, typically two ANNs, named generator and discriminator, where the former learns to produce synthetic data and the latter learns to distinguish true data from the output of the generator. During training, the generator tries to deceive the discriminator by synthesizing better data, while the discriminator becomes a better classifier. The equilibrium of this zero-sum game is reached when the discriminator can no longer distinguish real images from synthetic ones.

DCGAN, released in 2016, is very similar to GAN, except that it uses convolutional and convolutional-transpose layers in the discriminator and generator, respectively, making it more suitable for synthesizing imaging data (Radford et al., 2016). While there exist other algorithms to synthesize data, such as the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) and adaptive synthetic sampling (ADASYN) (Haibo He et al., 2008), they are mostly suitable for tabular data, which restricts their application in high-dimensional complex image data (Sampath et al., 2021). DCGAN instead can generate synthetic images with high visual fidelity. When it was first introduced, Radford et al. (Radford et al., 2016) identified an architecture that allows training the algorithm to generate higher-resolution images. We adopted this architecture by (1) using stride convolutions instead of pooling layers (see algorithm 7); (2) using batch normalization in both the generator and discriminator; (3) refraining from using hidden layers in the fully connected network in both the generator and discriminator; (4) using the rectified linear unit (ReLU) activation function in the generator for all layers except for the output, which uses the

hyperbolic tangent (Tanh) activation function; and (5) using LeakyReLU activation function in the discriminator for all layers. Finally, the binary cross entropy was employed as a loss function.

Figure 3.2 shows an example architecture of DCGAN. Here, we provided DCGAN with the healthy class images of the training set ($n_h \subset \text{train} = 1,272$) to train it on generating synthetic leaves, and, thus, oversampling the set to 1,500 images after 500 training epochs (Figure 3.3), on a learning rate of 0.0002 as suggested by Radford et al. (Radford et al., 2016)

3.8.2.4 *Algorithm 4: Augmentor*

Another approach to oversampling is using geometric transformations to generate new images. This helps models learn patterns invariant to different geometric perturbations. Augmentor automates this task by building an oversampling pipeline defined by a series of operations, such as rotating and flipping, to perform on a set of images. This approach allows chaining transformation operations together to create new images (Bloice et al., 2017) and makes Augmentor a popular choice (Abdelhalim et al., 2021; Bloice et al., 2019; Jager et al., 2021).

Our Augmentor pipeline included rotations (-15° – 15°), shearing (-15° – 15°), skewing (0%–20%), and horizontal flips. Figure 2 displays an illustration of the output of each operation. The pipeline was used to oversample class 2 ($n_{lm} \subset \text{train} = 761$), class 3 ($n_{sls} \subset \text{train} = 1,416$), class 4 ($n_{ts} \subset \text{train} = 1,123$), class 5 ($n_{eb} \subset \text{train} = 800$), class 7 ($n_{mv} \subset \text{train} = 298$), and class 9 ($n_{sm} \subset \text{train} = 1,340$) of the training set to 1,500 images each (Figure 3.3).

3.8.2.5 *Algorithm 5: KNN*

In contrast to oversampling, undersampling techniques remove data from the majority class to balance data distribution. While the easiest technique involves randomly undersampling the majority class, it does not show any concern for the importance of the data being removed, so that useful information can be lost. An alternative approach is to be more selective with the instances being removed by involving learning models that can identify redundant data such as the KNN ML algorithm. KNN calculates feature similarity between its training data to predict values

for new, previously unseen data. When given a new input, KNN finds k (a user-predefined number) most resembling data (nearest neighbors) using similarity metrics, such as the Euclidean distance. Based on the majority class of those similar data, the algorithm classifies the input. Despite its simplicity, KNN is considered one of the top 10 data-mining algorithms (Wu et al., 2008). When used for undersampling on 33 different datasets, KNN outperformed the results of the edited NN (ENN), neighbor cleaning rule (NCL), and random undersampling algorithms (Beckmann et al., 2015).

Figure 3.2 shows an example of KNN circling three detected neighbors with similar features belonging to the same class. Here, KNN was used to discover similarities of leaves in class 8 of the training set ($n_{ylcv} \subset \text{train} = 4,285$), removing images with redundant features, ultimately undersampling this majority class to 1,500 images (Figure 3.3). The input images were first converted to three dimensional arrays with values ranging from 0 to 255 (representing the brightness of each color in the RGB channels). Following Beckmann et al. (Beckmann et al., 2015), for each image, the algorithm deleted the nearest k neighbors (here $k = 3$). The process repeats until the number of images reaches the undersampling goal (1,500 images).

3.8.3. *Step 3: Data analysis*

3.8.3.1 *Algorithm 6: Random forest*

Random forest (RF) is an EL algorithm introduced in 2001 (Breiman, 2001). RF employs multiple decision trees (a forest) to create an ensemble (Figure 3.2). A decision tree is an ML algorithm used for both classification and regression tasks. It has a tree-like structure with two types of nodes: decision and leaf nodes. Decision nodes represent a test that is applied on the input data, where a decision will be made based on its result. Leaf nodes are the final output of those decisions. They represent the final prediction and do not contain any further branches.

In RF, each decision tree in the ensemble is generated from a random subset of the training set, with replacement (i.e., the same data can be sampled more than once), called the bootstrap sample. Lastly, RF determines the prediction by averaging the outputs of all decision trees for regression tasks, or by using majority voting for classification tasks. When evaluated on 121 different datasets against 17 algorithms, including discriminant analysis, support vector machines, and decision trees, RF

proved to be the best classifier among them (Fernández-Delgado et al., 2014). Although Wainberg et al. (Wainberg et al., 2016) questioned this conclusion, arguing that the results are biased by the lack of a held-out test set, RF is still possibly one of the most famous EL algorithms because of its effectiveness and simplicity (González et al., 2020).

We used the preprocessed training set to train an RF classifier with 1,000 decision trees to classify tomato leaf diseases and spider mites. We extracted three features from the training dataset to be used as input: (1) Haralick texture (Haralick et al., 1973) to identify texture characteristics of leaves; (2) Hu moments, also known as image moments, (Ming-Kuei Hu, 1962) to characterize the shape of leaves, including information such as their area, centroid, and orientation; and (3) color histograms (Chakravarti and Meng, 2009) to calculate the distribution of pixel color intensities, which improves classification on the assumption that images from the same class will have similar histograms. Figure 3.3 shows the resulting **confusion matrix** for the 10-class classification task.

3.8.3.2 *Algorithm 7: DCNN*

Similar to DNNs, DCNNs are made up of layers of artificial neurons that have learnable weights and biases. However, they use convolution operations instead of matrix multiplications between layers, making them more effective in processing imaging data. A DCNN is commonly composed of three main types of layers. The network starts with a convolutional layer, which can be followed by additional convolutional or pooling layers and ends with the fully connected layer. The convolutional layer employs a feature detector, known as a filter or kernel, that moves across an input image and checks for the presence of important features. This operation is known as convolution. A dot product is calculated between the input pixels and the filter, storing the output in a new matrix. Afterward, the filter shifts by a predefined number of pixels, known as stride, repeating the process until it has swept the entire image. The final output is known as feature map or activation map. After each convolution, a nonlinear activation function is applied to the feature map, such as ReLU transformation, which simply outputs the input value directly if it is positive, and outputs zero for all negative inputs. The pooling layer reduces the number of features in representations captured by convolutional layers. Similarly, it sweeps a

filter across the entire input, applying an aggregation function that chooses the maximum (maximum pooling) or the average (average pooling) value of the input to populate an output matrix. Finally, the fully connected layer, typically a DNN, performs classification on the features extracted by the series of convolutional and pooling layers. An example architecture of DCNN is illustrated in Figure 3.2. Features learned at different layers of DCNN correspond to different levels of abstractions. For example, initial layers detect edges and extract color information, and the following layers encode shapes and textures (Gopalakrishnan et al., 2017). Such learned features can be transferred from one model to another in a process called transfer learning. Transferring knowledge from previously trained models can be more efficient than training a DCNN from scratch. Among various DL algorithms, DCNNs are the most extensively studied (Gu et al., 2018). They have made a series of breakthroughs in image-based analysis (Gu et al., 2018; Krizhevsky et al., 2017; LeCun et al., 2015) and became the default choice for image classification tasks (LeCun et al., 2015; Russakovsky et al., 2015). We trained a standard DCNN using the preprocessed training set. Then, we repeated the process with another DCNN that transfers a densely connected convolutional network (DenseNet)-161 model pretrained on the ImageNet dataset. Both networks followed the architecture of DenseNet-161 described by Huang et al. (Huang et al., 2018), used Adam optimizer, and had a learning rate of 0.0001 and a batch size of 20. These hyperparameters can be tuned based on the performance of the validation set. After 100 epochs of training, the best-performing model on the validation set for the standard DCNN was generated at epoch 93, compared with 83 for the pretrained model. Figure 3.3 shows the confusion matrix for both models generated using the test set, suggesting that the pretrained model not only trained faster but is also a better classifier.

3.8.4. *Step 4: Performance analysis and explanation*

Once an algorithm is trained, it generates a model that can be used to make new predictions on data that it has not seen before. The test set is therefore used to evaluate the model predictions; they can either be correct or wrong. The ratio of correct predictions to total number of predictions shows how accurate is the model. Intuitively, a good classifier should generate a higher number of correct predictions and thus is expected to have a higher accuracy. However, the accuracy measure does not inform

whether classes are being predicted equally well or some classes are performing better than others. Moreover, in the case of imbalanced data, a classifier may achieve an extremely high accuracy by correctly predicting the majority class only.

Alternatively, a confusion matrix describes the complete performance of the model. It summarizes predictions in four categories: true-positive, true-negative, false-positive, and false-negative. It also forms the basis for the other types of metrics, including precision, recall, and F1 score. Figure 3.3 depicts the confusion matrix for each of the three trained models, cross-tabulating the true labels against the predicted ones. The matrices revealed that, out of the three models, the pretrained DenseNet-161 DCNN had the highest prediction accuracy for all 10 classes.

3.8.4.1 *Algorithm 8: LIME*

LIME was released in 2016 with the purpose of opening any ML or DL black box models, making their predictions explainable to humans (Ribeiro et al., 2016). To explain imaging data, LIME starts by partitioning an image into multiple segments that share common characteristics such as pixel intensity, known as superpixels (Figure 3.3). It then generates perturbations by deleting some superpixels in the image and monitors how they affect the prediction of the black box model. Finally, it identifies which areas of the image have been important for classification and, thus, generates the explanation by highlighting them (Figure 3.3). This shows how crucial the superpixels segmentation is for LIME to generate explanations. By default, LIME uses the quickshift algorithm to segment images. However, there are other approaches that have been tested to improve the quality of explanations. For example, Schallner et al. (Schallner et al., 2019) studied the effects of using different superpixel segmentation algorithms on LIME explanations. Their results suggested that the Compact-Watershed algorithm yielded the best results, introducing an improvement of 45.58% over the default quickshift algorithm. LIME is one of the most popular X-AI *post hoc* algorithms that can generate explanations for predictions produced by any AI black box models (Linardatos et al., 2020). Its performance turned out to be the best one when compared with Shapley additive explanations (SHAP) (Lundberg and Lee, 2017) and model agnostic supervised local explanations (MAPLE) (Plumb et al., 2018) (Plumb et al., 2018) to explain image-based classifications (Guidotti, 2021). We used LIME to explain our three trained models, RF, the standard DCNN, and the pretrained

DCNN. Both quickshift and Compact-Watershed algorithms were implemented. Figure 3 shows the explanation of the results of the pretrained DCNN, comparing the two algorithms. Our results are in agreement with findings in Schallner et al. (Schallner et al., 2019), highlighting more precise coverage of the late blight-diseased area on the tomato leaf using Compact-Watershed. Late blight disease remains one of the most significant threats to tomato agriculture, raising global food security concerns and causing economic hardship to farmers. While it is often misdiagnosed at its first encounter or not managed timely, its real-time, AI-driven detection and interpretation by researchers and farmers become of paramount importance.

3.9 Moving to the cloud

Large-scale high-throughput plant phenotyping experiments can produce images in the order of gigabytes (Roitsch et al., 2019), making it crucial to have access to greater computing power. GPUs can dramatically increase training speed thanks to their processing cores initially designed to process visual data such as videos and images. GPUs have thousands of these cores that process data in parallel (Bhattacharya and Cui, 2017). Some academic supercomputers offer access to GPUs, but, because of the rapid progress in their technology, they are at risk of becoming outdated (Zou et al., 2019).

Cloud computing services offer up-to-date hardware infrastructure that can be accessed over the Internet, eliminating the need for researchers or organizations to set up their own. Commercial services such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure enable the quick setup of a single or a group of servers connected to each other, known as a cluster. Importantly, these servers can be easily upgraded or downgraded based on the need.

Further, emerging services such as Google Colab, originally developed to help disseminate AI education and research, enable code development in a web browser, requiring almost no software to be set up. Colab gives its users the ability to write and execute codes (on central processing units [CPUs] or GPUs), take notes, and upload media such as images all in one place, called notebook, for free, but with varying usage limits. These notebooks are ideal to share well-documented, reproducible data analysis pipelines, promoting transparency and reusability (Clarke et al., 2021).

The increasing availability of cloud computing has led to an overuse of large amounts of data in a compute-centric paradigm, where data are analyzed by clusters in data centers. This approach requires high amount of data transfer, especially with the increased usage of Internet of Things (IoT) devices in agriculture (Tzounis et al., 2017), which can introduce high latency. Although next-generation wireless network technologies such as 5G aspire to provide higher bandwidth and lower latency (Harfouche et al., 2019; Tzounis et al., 2017), further concerns arise from the compute-centric approach, including (1) energy efficiency where data transmission is energy intensive (Aslan et al., 2018), (2) potential privacy violations caused by insecure data transmission and storage on cloud computers (vurukonda and Rao, 2016), and (3) inefficient use of storage due to unnecessary data transmission (Brous et al., 2020). This is leading to the adoption of a data-centric paradigm, where data processing takes place at the origin of data. However, because data generators such as IoT devices are resource constrained in terms of computational power, and because AI algorithms require large computational resources, a new field known as tiny AI or tiny ML (TinyML) has emerged. TinyML aims to compress AI models to fit on hardware with low computational power so that they can process data locally, reducing the need for data transmission. In a typical TinyML workflow, AI algorithms are trained on normal computers or in the cloud. Then, generated models are compressed using compression techniques such as network pruning, low-rank approximation, weight quantization, and network architecture transform (B. Li et al., 2021) to allow for their deployment in devices with low computational resources, without significant decrease in their performance (Cheng et al., 2020). This ability empowers IoT and resource-constrained devices with AI capabilities, opening up doors to many new possibilities.

3.10 Discussions and the way forward

As AI becomes more and more integrated into agriculture, we envisage a new green revolution with digital phenomics at the helm. Digital phenomics is enormously important for developing novel disease control strategies and accelerating breeding of tomato, widely considered one of our most important vegetable crops. We used an existing expert-curated open dataset and publicly available image processing resources to identify and classify nine biologically and economically important diseases and spider mites that can infect leaves of tomato fruit-bearing crops, resulting

in yield losses worldwide, and threatening our food supply and safety. In efforts toward embracing open science, the contextualized programming codes and computational notebooks are open source and ready for use immediately in image-based analyses. We believe that code, notebooks, and data sharing along with the availability of open-source programming languages, software frameworks, software libraries, and IDEs are bringing us closer to democratized AI research. Crucially, though, to support this, it is essential to promote an open research culture and to nurture talents for long-term progress of digital phenomics. Thinking across disciplinary boundaries in a time of health and climate crisis, this could better support the many ways in which our research community can contribute to the UN's Sustainable Development Goals (SDGs) and promote ethically responsible research and innovation (Harfouche et al., 2021). However, to make global agriculture more sustainable, smaller farms in developing countries must have free access to these resources and tools as well. Ready, Steady, Go AI is more than just a training and learning tool for driving ongoing and likely future biotechnology innovations; it is also useful for motivating graduate students and early-career researchers to engage in co-produced research studies that create space to work together with people with a shared interest. We advocate for creating perforations in the funnel of learning so that exposure to such skills and insights can be integrated into a full mainstream educational program to increase students' and early career researchers' own confidence and ability to undertake complex phenomics procedures using AI. We have a responsibility to our students to ensure they develop the awareness and skills to engage with experts outside plant sciences, recognizing complementary transferable skills and ways of creative and critical thinking.

3.11 Experimental procedures

3.11.1 *Data and code availability*

The code of this tutorial is available in interactive notebooks runnable on Google Colab and hosted on GitHub^{vi}.

The original PlantVillage dataset is available at <https://doi.org/10.17632/tywbtsjrjv>.

Our processed data, including cropped and segmented images, and the trained AI models are available on Mendeley Data at <https://doi.org/10.17632/4g7k9wptyd>.

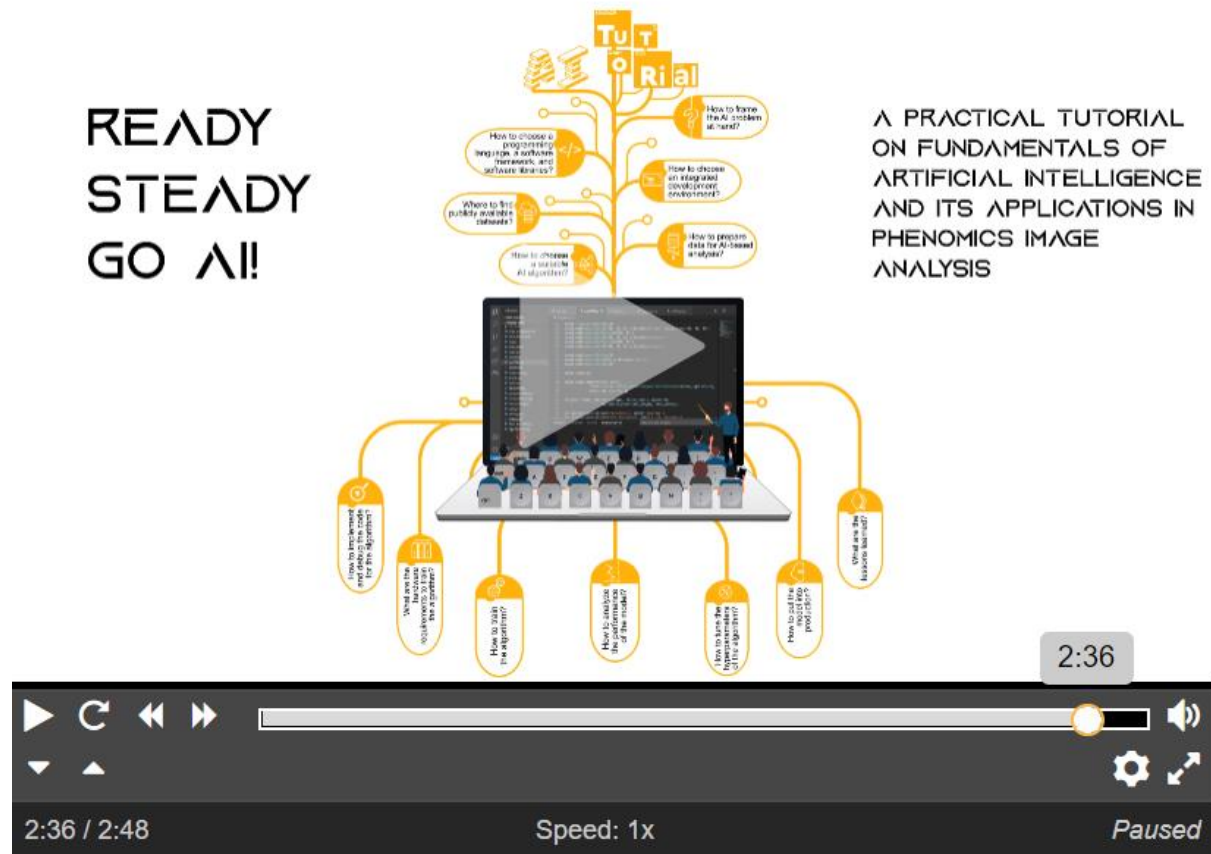
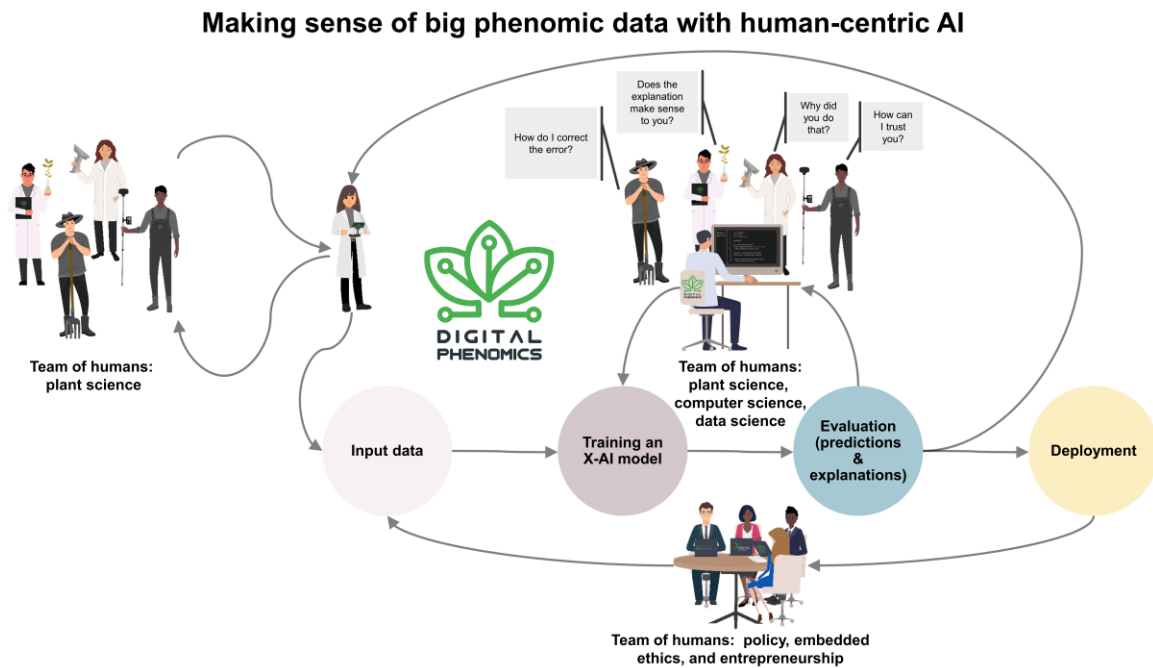


Figure360 3.1. An author presentation of Figure 3.1.

<https://ars.els-cdn.com/content/image/1-s2.0-S2666389921001719-mmc1.mp4>

Chapter 4 A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey

Graphical Abstract



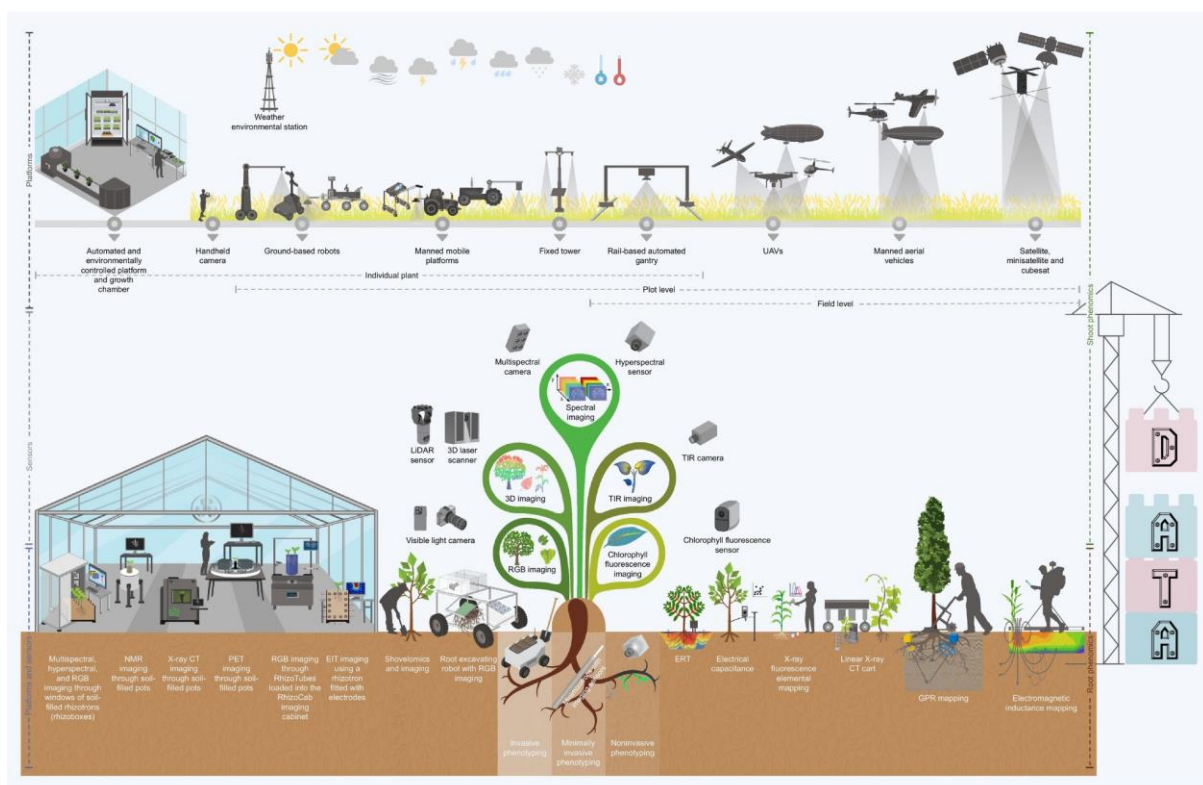
Graphical Abstract 3. Designing a human-centric AI that is explainable, bias-free, embeds human knowledge into its architecture, and keeps humans in the loop presents unprecedented opportunities for computational data analytics in plant digital phenomics and will have an extraordinary impact on plant science in the coming decades.

Abstract

Artificial intelligence (AI) has emerged as a fundamental component of global agricultural research that is poised to impact on many aspects of plant science. In digital phenomics, AI is capable of learning intricate structure and patterns in large datasets. We provide a perspective and primer on AI applications to phenome research. We propose a novel human-centric explainable AI (X-AI) system architecture consisting of data architecture, technology infrastructure, and AI architecture design. We clarify the difference between *post hoc* models and 'interpretable by design' models. We include guidance for effectively using an interpretable by design model in phenomic analysis. We also provide directions to sources of tools and resources for making data analytics increasingly accessible. This primer is accompanied by an interactive online tutorial.

4.1 Approaching plant phenomics from different angles

Crop breeding relies heavily on phenotypic information, which remains a bottleneck for realizing its full potential. The advent of plant phenomics (reviewed in Dhondt et al., 2013; Feng et al., 2021; Fiorani and Schurr, 2013; Furbank and Tester, 2011; Houle et al., 2010; Huang et al., 2020; Song et al., 2021; Tardieu et al., 2017c; Yang et al., 2020), which can broadly be considered to be the systematic study of phenotypes, however, marks a turning point. Phenomic platforms equipped with novel imaging sensors promise to make it possible to perform phenotyping of a wide range of plant traits, organs, and environmental situations at scale (Figure 4.1). These new technological developments have opened up avenues for automated data acquisition, enabling phenomics to evolve to become a thriving research field of its own (Coppens et al., 2017).



Trends in Plant Science

Figure 0.1. Phenomics platforms and sensors for high-throughput plant phenotyping in controlled environments and field conditions: collecting relevant data from a wide range of sources.

A network of comprehensive automated weather stations collects hourly weather and soil data including, among others, rainfall, air temperature, solar radiation, relative humidity, and soil moisture and temperature. Varying phenotyping scales allow precise and consistent monitoring of individual

plants, plots, and fields. Ground-based and aerial platforms can mount a variety of cameras and sensors for non-invasive, high-throughput phenotyping (HTP): visible light cameras for RGB imaging, LiDAR sensors and 3D laser scanners for 3D imaging, multispectral cameras and hyperspectral sensors for spectral imaging, TIR cameras for thermal imaging, and chlorophyll fluorescence sensors for chlorophyll fluorescence imaging. Automated and environmentally controlled platforms, growth chambers, and multifunction printers can be used for HTP in controlled environments. Root phenotyping in the field can be invasive (e.g., shovelomics and its automation with root excavating robots), minimally invasive (e.g., minirhizotrons), or non-invasive (e.g., ERT, electrical capacitance, GPR mapping, and electromagnetic inductance mapping). Field-deployable linear X-ray CT cart, and handheld X-ray fluorescence elemental mapping, are being explored for non-invasive field root phenotyping. Multispectral, hyperspectral, RGB, and EIT imaging can be used to phenotype roots in soil-filled rhizotrons (rhizoboxes) in controlled environments. Similarly, NMR, X-ray CT, and PET imaging can be used to phenotype roots in soil-filled pots. RhizoTubes, which are cylindrical rhizotrons, allow full visualization of the root system of a single or up to six plants simultaneously. The RhizoCab is designed to take images of the entire root systems of plants growing in RhizoTubes. These platforms and sensing technologies are generating a massive amount of data, and this creates a need for proper data management and processing – the first step of the data lifecycle in digital phenomics (Figure 4.2). Abbreviations: EIT, electrical impedance tomography; ERT, electrical resistance tomography; GPR, ground-penetrating radar; LiDAR, light detection and ranging; NMR, nuclear magnetic resonance; PET, positron emission tomography; RGB, red–green–blue; TIR, thermal infrared; UAV, unmanned aerial vehicle; X-ray CT, X-ray computed tomography.

Embracing digital technology holds tremendous potential for driving transformative changes in plant phenomics by improving the collection of, access to, and analysis of phenomic big data. As such, digital phenomics furnishes tools and resources that aid in the digitization of plant phenomics. It uses phenome data and metadata to guide decision making along the entire data analytics cycle (Nakhle and Harfouche, 2021). Because data are growing in volume and vary in their sources, effective management strategies must be put into place (Figure 4.2). Digital phenomics gives rise to computational phenomics, which allows the assembly of a broad array of methods that aid in the discovery of intricate structure and patterns from phenotypic data using technology infrastructure (Figure 4.3) and AI architecture design (Figure 4.4).

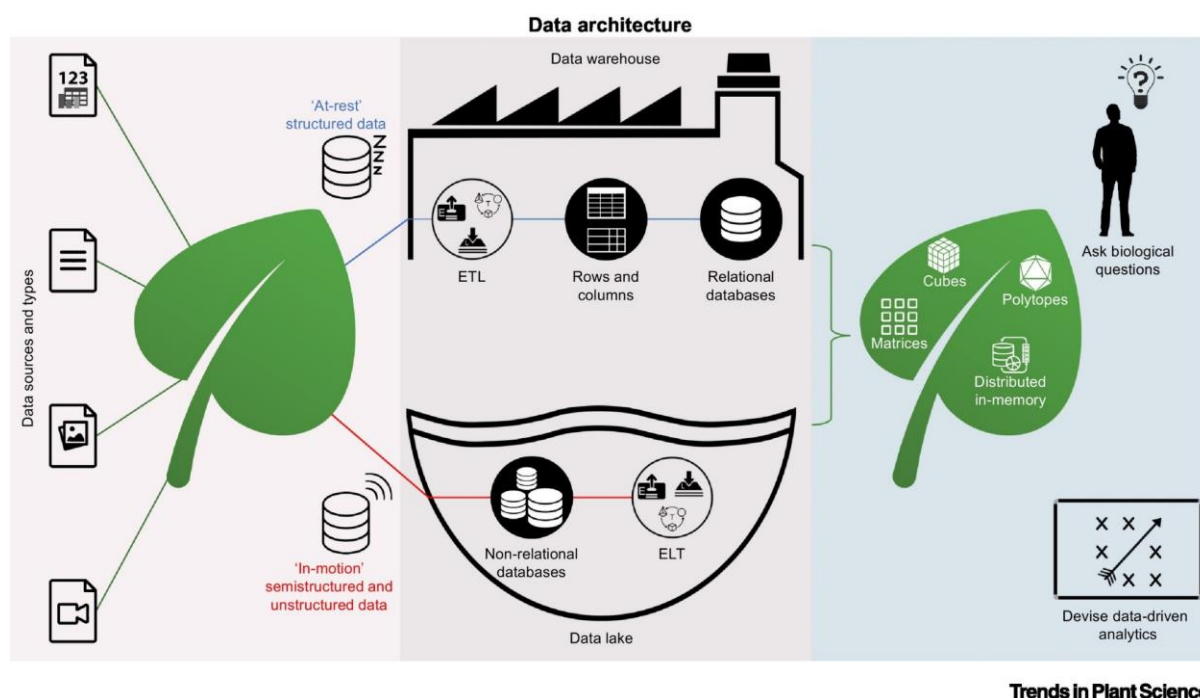


Figure 4.2. Data architecture blueprint to drive human-centric explainable artificial intelligence (X-AI) innovation.

Phenomic data can be structured, semistructured, or unstructured. Structured (e.g., spreadsheet files) data (blue line) are typically 'at-rest', transformed into rows and columns, and loaded into relational databases in data warehouses using a process known as ETL. Semistructured (e.g., extensible markup language files) and unstructured (e.g., flat files) data (red line) are streamed 'in-motion', loaded into non-relational databases, and stored in data lakes in their raw form; their transformation occurs on-demand using a process known as ELT. Because ELT loads data immediately, it prevents any slowdown that often occurs at the transformation step, and thus enables near real-time analytics for fast and practical decision making. When either ETL or ELT is used, data warehouses and data lakes store data as matrices, cubes, polytopes, or distributed in memory. A well-designed data architecture results in higher-quality phenomic datasets that allow plant scientists to ask biological questions and to devise data-driven analytics in the search for answers. Abbreviations: ELT, extract, load, transform; ETL, extract, transform, load.

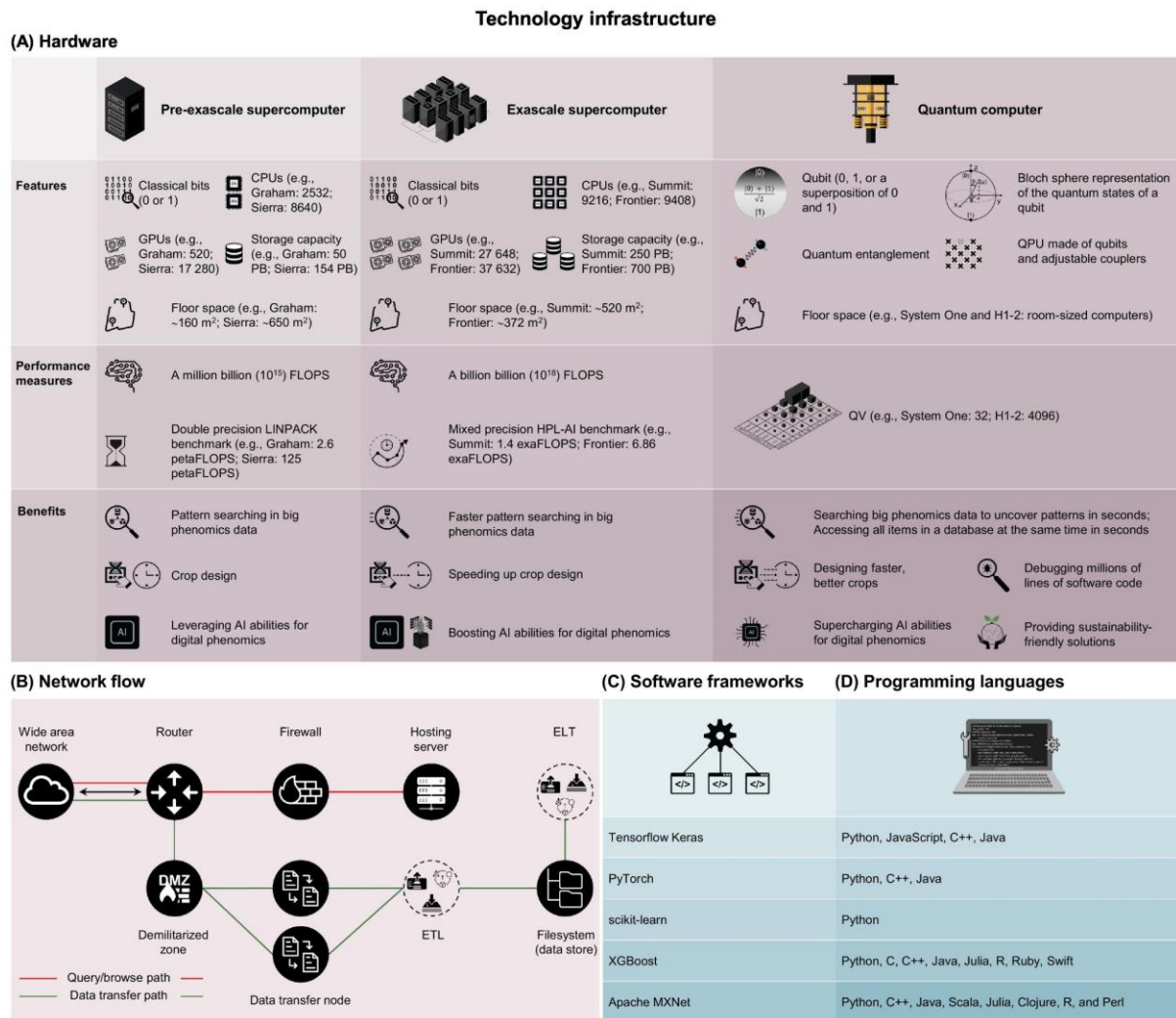


Figure 4.3. Technology infrastructure to support human-centric explainable artificial intelligence (X-AI).

The technology infrastructure consists of the hardware, network flow, software frameworks, and programming languages that enable data transmission, transformation, storage, access, and analysis. (A) Computing hardware supporting data analysis. Pre-exascale supercomputers (e.g., Graham at the University of Waterloo, and Sierra at the Lawrence Livermore National Laboratory) reach a performance of a million billion (10¹⁵) FLOPS. With a similar hardware architecture but an increased number of CPUs and GPUs, exascale supercomputers (e.g., Summit and Frontier at Oak Ridge National Laboratory) reach a billion billion (10¹⁸) FLOPS and can deliver higher performance in pattern searching in phenomic big data, thus speeding up crop design. Quantum computers (e.g., System One of International Business Machines and H1-2 of Quantinuum) represent a new paradigm in computation that leverages the fundamental principles of quantum mechanics to perform calculations. They employ **quantum bits** (qubits) that can be entangled, giving them the ability to manipulate vast amounts of data with few operations, and thus can solve problems polynomially faster than classical computers (i.e., pre-exascale and exascale supercomputers) ultimately to more rapidly design better crops. Researchers can simulate quantum circuits on classical computers using free and open-source

software development kits such as Cirq or Qiskit, and the cuQuantum software library, to leverage the power of GPUs and **parallel computing** to perform faster calculations. Examples of classical and quantum computers are compared based on their peak performance that is the theoretical highest processing power they can reach. For classical computers, the LINPACK benchmark tests the performance in double precision (64-bit) compute capabilities, whereas HPL-AI scores performance based on mixed precision (16- and 32-bit). Because quantum computers use **quantum processing units** (QPUs) to manipulate the quantum states of qubits to perform computations, their performance is measured using **quantum volume** (QV). (B) Network flow to enable high-throughput access to and sharing of phenomic datasets. Requests coming from the wide area network are forwarded through a router to one of two paths: (i) the data query and browse path (red line) where requests to browse or search phenomic datasets are filtered through a firewall and processed by the hosting server, and (ii) the data transfer path (green line) where requests to download or upload phenomic datasets are inspected in the DMZ for access control, and are forwarded to the transfer nodes (typically Linux servers) to reach the filesystem where data can be transformed before or after storage using ETL or ELT, respectively (Figure 4.2). (C,D) Representative free and open-source software frameworks and their supported programming languages used to implement AI algorithms. Abbreviations: CPU, central processing unit; ELT, extract, load, transform; ETL, extract, transform, load; FLOPS, floating-point operations per second; GPU, graphics processing unit; HPL-AI, high-performance LINPACK for accelerator introspection; LINPACK, linear equations software package; PB, petabyte.

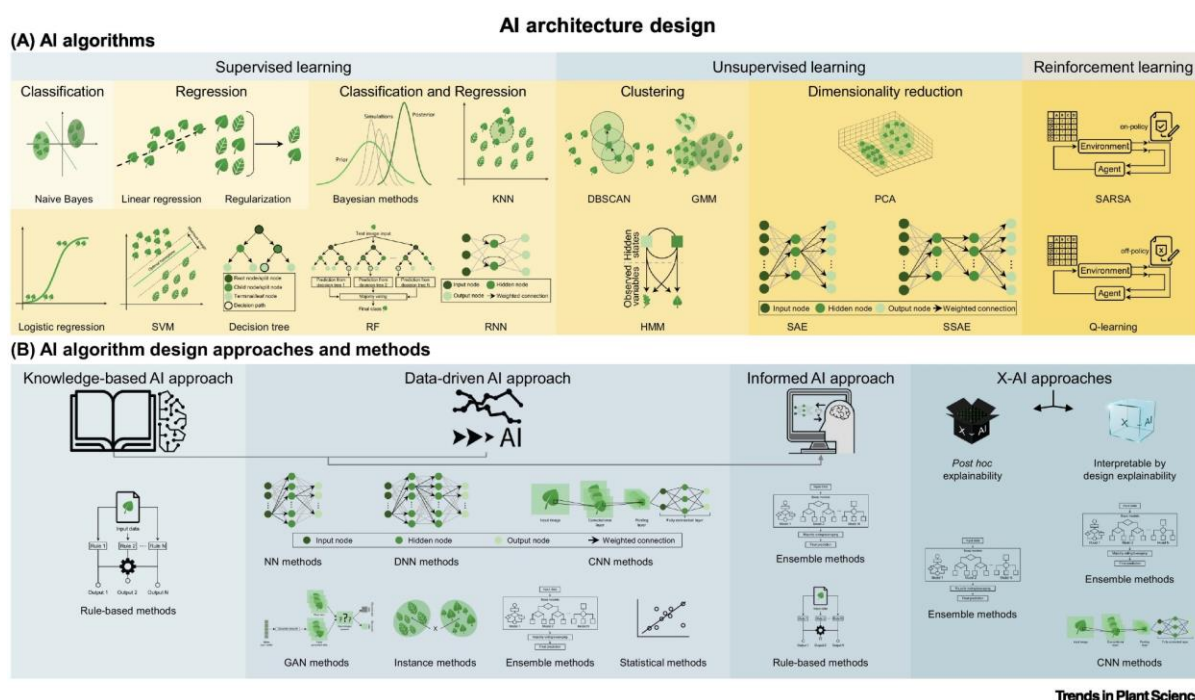


Figure 0.4. Technology infrastructure to support human-centric explainable artificial intelligence (X-AI).

(A) Representative AI algorithms that are used for AI tasks in digital phenomics, including classification and regression (supervised learning) and clustering and dimensionality reduction (unsupervised learning). Reinforcement learning algorithms can be applied to search optimal architecture designs and

improve their performance. (B) Representative AI algorithm design approaches and methods, a higher level of abstraction that helps scientists in their efforts to design and implement novel AI algorithms to answer complex biological questions. The knowledge-based AI approach represents human expert knowledge as a collection of rules to form a knowledge base that is applied to solve a specific problem. It offers a consistent answer to a repetitive problem and its decisions are explainable. It can be implemented using rule-based methods. The data-driven AI approach discovers connections and correlations automatically in a large amount of data and learns a black box model. It can be implemented using various methods including CNN, ensemble, or statistical methods, among others. The informed AI approach combines knowledge-based AI with data-driven AI by leveraging human knowledge with knowledge acquired from data to make faster and more accurate decisions. It can be implemented using ensemble or rule-based methods. Finally, X-AI approaches provide meaningful explanations of decisions made by X-AI models to humans through a decipherable decision-making process. They allow the monitoring of inputs and outputs with the purpose of verifying the adherence of X-AI models to ethical and sociolegal **values** by (i) opening the black box of data-driven or informed AI models using ensemble methods, or by (ii) designing new, transparent glass box algorithms that are interpretable by design using ensemble or CNN methods. Abbreviations: CNN, convolutional neural network; DBSCAN, density-based spatial clustering of applications with noise; DNN, deep neural network; GAN, generative adversarial network; GMM, Gaussian mixture model; HMM, hidden Markov model; KNN, k-nearest neighbors; NN, neural network; PCA, principal component analysis; RNN, recurrent neural network; RF, random forest; SAE, sparse autoencoder; SARSA, state–action–reward–state–action; SSAE, stacked SAE; SVM, support vector machine.

In the past decade AI – the science of studying, designing, and developing intelligent computer systems that can perform tasks that normally require human intelligence – finally began to reveal its remarkable power and disruptive potential. Driven mainly by the advent of machine learning (ML) – a particular approach to AI in which intelligent systems learn and derive models from training datasets – and deep learning (DL) – a specialized branch of ML that leverages neural networks to identify patterns in complex data – AI flexed its muscles by achieving predictive successes in phenomics. For example, in red–green–blue (RGB) image analysis, convolutional neural networks (CNNs) were used to predict the yield of individual plants of barley and wheat (Nevavuori et al., 2019), to classify and quantify biotic and abiotic stresses in leaves of various fruit and vegetable crops (Geetharamani and Pandian, 2019), to segment roots of chicory, wheat, and rapeseed (Gaggion et al., 2021; Smith et al., 2020; Yasrab et al., 2019), and to count tobacco leaves (Ubbens and Stavness, 2017). A CNN was also used to provide explainable classifications of biotic and abiotic stresses in soybean leaves by isolating the top-k feature maps learned by the model

(Ghosal et al., 2018). A random forest (RF), neural network (NN), k-nearest neighbor (KNN), partial least squares (PLS), and support vector machine (SVM) were employed to estimate the nitrogen nutrition index for improving nitrogen use efficiency in rice (Qiu et al., 2021; Shi et al., 2021). In wood anatomical images at a microscopic level, a mask region-based CNN (Mask R-CNN) was used to analyze the intrinsic variability of wood anatomical features in conifers, alder, beech, and oak (Resente et al., 2021). In analyzing multispectral and hyperspectral data, a CNN was used to provide explainable identification of biotic stress in individual soybean plants by incorporating saliency maps (Nagasubramanian et al., 2019); a PLS and a RF to estimate above ground biomass in maize (Shu et al., 2021); an SVM, a KNN, and a linear discriminant analysis (LDA) to detect and segment root decay in wheat (Bodner et al., 2018); an LDA and a PLS to detect response to drought stress in bell pepper, courgette, sunflower, radish, foxtail millet, and sorghum (Burnett et al., 2021); and a data-mining sharpener to guarantee consistent spatial resolution among heterogeneous remote sensing image datasets to dissect the latent heat flux signature of poplar in response to drought (Tauro et al., 2022). In 3D point cloud analysis, an SVM was applied to estimate yield and canopy geometric characterization in apple (Gené-Mola et al., 2020a). In thermal infrared (TIR) image analysis, SVMs and Gaussian processes were used to identify drought stress in spinach (Raza et al., 2014), and rotation forests were used to predict plant water status in grapevine (Gutiérrez et al., 2018). In chlorophyll fluorescence image analysis, a CNN was used to identify abnormalities in organelle morphology in *Arabidopsis thaliana* (J. Li et al., 2021). In X-ray computed tomography (X-ray CT) analysis, an encoder-decoder network was used to segment wheat roots (Soltaninejad et al., 2020).

Simultaneous analysis of data generated by different sensors and imaging techniques of the same biological sample (i.e., plant) can further improve phenotypic trait predictions. Recent studies demonstrated that fusion of multiple data sources originating from the same plants (i.e., paired data) perform better than a single source. For example, using a deep NN (DNN), the fusion of RGB, TIR, and multispectral data delivered superior performance over single-sensor data analytics for yield prediction in soybean (Maimaitijiang et al., 2020); similar results were reported for the estimation of soybean chlorophyll content, nitrogen concentration, leaf area index, and aboveground biomass by employing an extreme learning machine (Maimaitijiang et

al., 2017). Now and in the future, rigorous data integration of phenomics and other different omic datasets that were not originally set out to be integrated and are from distinct biological samples (i.e., unpaired data) may help to dissect the biological mechanisms that underlie desirable traits and shed light on the flow of information that underpins plant responses to environmental stresses (Chhetri et al., 2020, 2019; Coppens et al., 2017; Harfouche et al., 2019; Streich et al., 2020). Because the phenotype of a plant is the result of interaction between its genotype and the environment ($G \times E$) in which it grows (Poorter et al., 2016), integration efforts should also include environmental data such as climatypes; this will be crucial for designing new crop ideotypes that are optimized for niche environments in a world with a rapidly changing climate (Streich et al., 2020).

Importantly, data management strategies should incorporate the findable, accessible, interoperable, and reusable (FAIR) guiding principles (Wilkinson et al., 2016) to put those phenome and envirome data to their most effective use. This requires standards to ensure that the necessary metadata are recorded about data generation methods and the experimental and environmental conditions in which they were acquired (Williamson et al., 2021). In this regard, the minimum information about a plant phenotyping experiment (MIAPPE) standard has been a great step forward to harmonize data from phenotyping experiments with controlled vocabulary and ontologies (Ćwiek-Kupczyńska et al., 2016; Papoutsoglou et al., 2020). Accordingly, the development of tools for capturing the complete set of metadata is poised to have high impact on the support of FAIR data. As standards and tools become more widely disseminated and explored, we envisage that metadata will become commonly annotated by users, expected by referees, and required by journals and data repositories. Furthermore, the workflows that are used to analyze data should themselves be FAIR (Wolf et al., 2021). When data sharing is not viable because of possible privacy or security concerns, **federated learning** (FL) and communication-efficient FL offer an unprecedented opportunity to train AI models without sharing data (Chen et al., 2021; McMahan et al., 2017). FL has gained traction in medical imaging applications (Dayan et al., 2021; Kaissis et al., 2021, 2020) and holds great promise for overcoming data sharing challenges in plant phenomics.

This primer provides suggestions on how to use AI effectively in plant phenomics, on how to ensure that human-centric X-AI can benefit all, and discusses various X-AI

approaches and techniques. We have created a central directory of all publicly available plant imaging datasets, and report their sources, accessibility, and a summary of the species and organ systems represented (Table 4.1). This review is accompanied by an interactive tutorial to train an interpretable by design model to deliver predictive and prescriptive analytics to users. Our primer is intended as an educational resource for phenomicists who are interested in applying X-AI approaches and techniques, and for plant scientists who seek a high-level understanding of this rapidly evolving field. Data scientists and information systems (IS) scientists may also use this primer as an introduction to the promising applications of X-AI in phenomics.

Table 4.1. Publicly available global datasets and their characteristics as valuable resources for plant digital phenomics research^{a,b}.

Dataset	Country of origin	Plant species ^c	Plant organ systems	No. of images	Platform	Sensors	Image annotation types ^d	Potential applications	Access type ^e	Data access details	Data PID ^f	Link to dataset	File format ^g	Refs
Supporting data for "deep machine learning provides state-of-the-art performance in image-based plant phenotyping"	UK	<i>Triticum aestivum</i>	Root	2697	Controlled environment stationary platform	RGB	Point annotations of root tips, leaf tips, leaf bases, ear tips, ear bases	Identification and localization of root tips, leaf, and wheat ear tips	OA	Downloadable tar file	http://doi.org/10.5524/100343	http://gigadb.org/dataset/100343	JPG	(Pound et al., 2017a)
			Shoot	1664										
Wheat 2017	UK	<i>T. aestivum</i>	Shoot, spikes	520	Controlled environment stationary platform	RGB	Point annotations of spikelet, base and tip of each ear	Localization and counting of wheat spikes and spikelets	OA	Register for link to download zip file	–	https://plantimages.nottingham.ac.uk/	JPG	(Pound et al., 2017b)
Global wheat head detection (GWHD)	Japan, France, Canada, UK, Switzerland, China, Australia	Wheat	Shoot	1094	Manned mobile platform	RGB	Bounding boxes	Detection and localization of wheat heads	OA	Downloadable zip file	http://doi.org/10.5281/zenodo.4298502	https://zenodo.org/record/4298502	PNG	(David et al., 2020)
				678	Handheld visible light camera in the field									
				447	Rail-based field automated gantry									

Global wheat head detection (GWHD) 2021	Japan, France, Canada, UK, Switzerland, China, Australia, USA, Mexico, Republic of Sudan, Norway, Belgium	Wheat	Shoot	2307	Manned mobile platform	RGB	Bounding boxes	Detection and localization of wheat heads	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.5092309	https://zenodo.org/record/5092309	PNG	(David et al., 2021)
				2684	Handheld visible light camera in the field									
				1429	Rail-based field automated gantry									
Supporting data for "high throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat"	KS - USA	Wheat	Shoot	>400000	Ground-based field robot	RGB	Image-level annotations	Estimation of plant morphology and developmental stages	OA	Downloadable zip file	http://dx.doi.org/10.5524/100566	http://gigadatabase.org/dataset/view/id/100566	JPEG	(Wang et al., 2019)
RootNav 2.0	UK	<i>T. aestivum</i>	Root	3630	Controlled environment stationary platform	RGB	Segmentation masks, image-level annotations	Root segmentation and species classification	OA	Download each image separately	http://doi.org/10.5524/100651	http://gigadatabase.org/dataset/100651	JPG	(Yasrab et al., 2019)
		<i>Brassica napus</i>		120		NIR							PNG	
		<i>Arabidopsis thaliana</i>		277										
	USA	Buckwheat	Shoot	56		RGB			OA				PNG	

Cosegmentation for plant phenotyping (CosegPP)		Sunflower		56	Controlled environment stationary platform	IR	Segmentation masks	Plant segmentation		Downloadable zip file	https://doi.org/10.5281/zenodo.5117176	https://zenodo.org/record/5117176		(Quiñone s et al., 2021)
				56		CF								
				104		RGB								
				112		IR								
				112		CF								
Aberystwyth leaf evaluation dataset	UK	<i>A. thaliana</i>	Shoot	56	Controlled environment stationary platform	RGB	Semantic segmentation	Plant and leaf segmentation	OA	Downloadable zip file	http://doi.org/10.5281/zenodo.168158	https://zenodo.org/record/168158	PNG	
Deep phenotyping dataset	Australia	<i>A. thaliana</i>	Shoot	2134	Controlled environment stationary platform	RGB	Image-level annotations	Genotype classification	OA	Downloadable zip file	–	https://figshare.com/s/e18a978267675059578f	JPG	(Taghavi Namin et al., 2018)
Supporting data for "ChronoRoot: high-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture"	France	<i>A. thaliana</i>	Root	331	Controlled environment stationary platform	RGB	Segmentation masks	Root segmentation	OA	Downloadable tar file	http://dx.doi.org/10.5524/100911	http://dx.doi.org/10.5524/100911	PNG	(Gaggion et al., 2021)
	Italy	<i>A. thaliana</i>	Shoot	6287	GARNICS controlled	RGB	Segmentation masks,	Plant and leaf segmentation, leaf counting,	OA	Fill in a form and get access and	–	https://www.plant-phenotyping.	HDF5	(Minervini et al., 2016)

Plant phenotyping datasets	Germany	<i>Nicotiana tabacum</i>		165120	environment robot gardener		bounding boxes, point annotations of leaf centers	species classification		download a zip file		org/datasets-home		
Multi-modality plant imagery database (MSU-PID)	MI - USA	<i>A. thaliana</i>	Shoot	576	Controlled environment stationary platform	RGB, RGB-depth, CF, NIR ^h	Polygon and point annotations of leaves and leaf tips	Leaf segmentation, counting, alignment, and tracking	OA	Downloadable zip file	–	http://cvlab.cse.msu.edu/multi-modality-imagery-database-msu-pid.html	PNG	(Cruz et al., 2016)
		<i>Phaseolus vulgaris</i>		175										
Plant segmentation	Hungary	<i>A. thaliana</i>	Shoot	16 ⁱ	Computer scanner	RGB	Segmentation masks	Segmentation and length determination of hypocotyl	OA	Downloadable zip file	–	https://www.kaggle.com/tivadardanka/plant-segmentation	PNG	(Dobos et al., 2019)
		<i>Brachypodium distachyon</i>		8 ⁱ	Handheld visible light camera in a controlled environment									
		<i>Sinapis alba</i>	Shoot and root	15 ⁱ										
Eschikon plant stress phenotyping dataset	Switzerland	<i>Beta vulgaris</i>	Shoot	496	Controlled environment stationary platform	HS (NIR)	Image-level annotations, bounding boxes	Classification of biotic and abiotic stress	OA	Downloadable zip file	–	https://projects.asl.ethz.ch/datasets/doku.php?id=2018plantstressphenotyping	PNG	(Khanna et al., 2019)
				992		RGB-depth								
				496		RGB								
Remote sensing 2018 weed map dataset	Switzerland, Germany	<i>B. vulgaris</i> , weed	Shoot	18746	UAV	MS (NIR), 4 and 5 bands ^j	Semantic segmentation	Identification and segmentation	OA	Downloadable zip file	–	https://projects.asl.ethz.ch/datasets/doku.php?id=	TIF	(Sa et al., 2018)

[illegible]

Crop/weed field image dataset (CWFID)	Germany	Carrot, weed	Shoot	60	Ground-based field robot	MS (NIR), 2 bands	Segmentation masks, semantic segmentation	Identification and segmentation of crops and weeds	OA	Downloadable zip file	–	https://github.com/cwfid/dataset	PNG	(Haug and Ostermann, 2015)
Cassava leaf disease classification	Uganda	<i>Manihot esculenta</i>	Shoot	9436	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Create Kaggle account to download zip file	–	https://www.kaggle.com/c/cassava-disease/data	JPG	(Mwebaze et al., 2019)
Cassava disease classification		Cassava		21397								https://www.kaggle.com/c/cassava-leaf-disease-classification/data		
Cassava root cross-section images	Uganda, Tanzania	Cassava	Root	10052	Handheld visible light camera in the field	RGB	Semantic segmentation	Quantification of root damage	OA	Downloadable zip file	https://doi.org/10.17632/qvp7vshvnh.3	https://data.mendeley.com/datasets/qvp7vshvnh/3	JPG	(Nakatumba-Nabende et al., 2020)
Plant pathology 2020 - FGVC7	NY - USA	Apple	An intact leaf	3651	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Create Kaggle account to download zip file	–	https://www.kaggle.com/c/plant-pathology-2020-fgvc7/data	JPG	(Thapa et al., 2020)
Plant pathology 2021 – FGVC8				23000								https://www.kaggle.com/c/plant-pathology-2021-fgvc8/data		

											pathology-2021-fgvc8			
LFuji-air dataset	Spain	<i>Malus domestica</i>	Shoot	88	Ground-based field robot	LiDAR	Bounding boxes	Identification and localization of fruits, estimation of yield, canopy geometric characterization	OA	Downloadable zip file	–	https://repositori.udl.cat/handle/10459.1/68782	MAT	(Gené-Mola et al., 2020a, 2020b)
MinneApple	MN - USA	Apple	Shoot	1000	Handheld visible light camera in the field	RGB	Polygon annotations	Identification, segmentation, and counting of fruits	OA	Downloadable tar file	http://doi.org/10.13020/8ecp-3r13	https://conservancy.umn.edu/handle/11299/206575	JPG	(Hani et al., 2020)
Data from: multi-species fruit flower detection using a refined semantic segmentation network	WV - USA	Apple	Shoot	18 ⁱ	Manned mobile platform	RGB	Segmentation masks, image-level annotations	Classification of species from fruit flowers	OA	Downloadable zip file	http://doi.org/10.15482/USDA.ADC/1423466	https://data.nal.usda.gov/dataset/data-multi-species-fruit-flower-detection-using-refined-semantic-segmentation-network	JPG	(Dias et al., 2018)
				130	Handheld visible light camera in the field									
		Peach		24 ⁱ										
		Pear		18 ⁱ										
DiaMOS plant dataset: a dataset for diagnosis and	Italy	Pear	Shoot	3505	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.5557313#v5JrXZByMo	https://zenodo.org/record/5557313#v5JrXZByMo	JPG	(Fenu and Mallocci, 2021)

monitoring plant disease											nodo.5557313			
PlantaeK: A leaf database of native plants of Jammu and Kashmir	Jammu, Kashmir	Apple	A single detached leaf	351	Handheld visible light camera in a controlled environment	RGB	Image-level annotations	Identification of biotic stress, classification of plant species	OA	Downloadable zip file	http://doi.org/10.17632/t6j2h22jpx.1	https://data.mendeley.com/datasets/t6j2h22jpx/1	JPG	
		Apricot		270										
		Cherry		212										
		Cranberry		212										
		Grapevine		171										
		Peach		331										
		Pear		228										
		Walnut		378										
Data for: identification of plant leaf diseases using a 9-layer deep convolutional neural network	FL - PA - NY - USA	Multiple crops ¹	A single detached leaf	54305	Handheld visible light camera in a controlled environment	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/tywbtsirjv.1	https://data.mendeley.com/datasets/tywbtsirjv/1	JPG	(Gopal et al., 2019)
ACFR orchard fruit dataset	Australia	Apple	Shoot	1120	Ground-based field robot	RGB	Bounding boxes, circle annotations	Identification and classification of fruits	OA	Downloadable zip file	–	http://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/	PNG	(Bargoti and Underwood, 2017)
		Almond		620										
		Mango		1964										
Dataset for pest classification in Mango farms	Indonesia	Mango	An intact leaf	510	Handheld visible light camera in the field	RGB	Image-level annotations	Classification of pest	OA	Downloadable zip file	https://doi.org/10.17632/94if97jzc8.1	https://data.mendeley.com/datasets/94if97jzc8/1	JPG	(Kusrini et al., 2020)

Berries in vineyards-color (BIVcolor)	Germany	<i>Vitis vinifera</i>	Shoot	500	Ground-based field robot	RGB	Image-level annotations	Detection of grape size and color	OA	Downloadable zip file	http://doi.org/10.5073/jki-data.2015.1	https://www.openagrar.de/receive/openagrar_mds_00021925	TIF	(Kicherer et al., 2015)
Supporting data for "a novel ground truth multispectral image dataset with weight, anthocyanins and Brix index measures of grape berries tested for its utility in machine learning pipelines"	Spain	<i>V. vinifera</i>	Fruit	1283	Controlled environment stationary platform	MS	Image-level annotations	Prediction of grape variety	OA	Downloadable zip file	http://dx.doi.org/10.5524/102220	http://gigadb.org/dataset/102220	TIF	(Navarro et al., 2022)
ESCA-dataset	Italy	Grapevine	Shoot	1770	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/89cncx58kj1.1	https://data.mendeley.com/datasets/89cncx58kj1	JPG	(Alessandrini et al., 2021)
An annotated image dataset of downy mildew symptoms on	France	Grapevine	Shoot	99	Ground-based field robot	RGB	Semantic segmentation	Identification of downy mildew	OA	Downloadable zip file	–	https://ars.elsa-cdn.com/content/image/1-s2.0-	JPEG	(Abdelghafour et al., 2021)

Merlot grape variety											S2352340921005345-mm1.zip			
The MangoNet semantic dataset	India	<i>Mangifera indica</i>	Shoot	49 ⁱ	Handheld visible light camera in the field	RGB	Semantic segmentation	Identification and counting of fruits	OA	Downloadable zip file	–	https://github.com/avades/h02/MangoNet-Semantic-Dataset	JPG	(Kestur et al., 2019)
Leaves: India's most famous basil plant leaves quality dataset	India	Basil	Shoot	1131	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Download each image separately	https://doi.org/10.2122/7/a4f6-4413	https://iee-dataport.org/open-access/leaves-india%E2%8099s-most-famous-basil-plant-leaves-quality-dataset	JPG	
A database of leaf images: practice towards plant conservation with plant pathology	Jammu, Kashmir	Multiple species ^m	A single detached leaf	4503	Handheld visible light camera in a controlled environment	RGB	Image-level annotations	Identification of biotic stress	OA	Downloadable zip file	http://doi.org/10.17632/hb74ynkicn.4	https://data.mendeley.com/datasets/hb74ynkicn/4	JPG	(Chouhan et al., 2019)
	Pakistan	Citrus	Shoot	150		RGB			OA				JPG	

A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning			A single detached leaf	609	Handheld visible light camera in the field		Image-level annotations	Identification and classification of biotic stress		Downloadable zip file	http://doi.org/10.17632/3f83gxmv57.2	https://data.mendeley.com/datasets/3f83gxmv57/2		(Ali et al., 2017; Sharif et al., 2018)
Supporting data for "morphometric analysis of Passiflora leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade"	Brazil	40 <i>Passiflora</i> species ⁿ	A single detached leaf	5767	Multifunction printer	RGB	Point annotations of leaf edges	Classification of species	OA	Download each image separately	http://doi.org/10.5524/100251	http://qigadb.org/dataset/100251	TIF	(Chitwood and Otoni, 2017a, 2017b)
QuinceSet: dataset of annotated Japanese quince images for object detection	Latvia	<i>Chaenomeles japonica</i>	Shoot	1515	Handheld visible light camera in the field	RGB	Image-level annotations, bounding boxes	Identification and localization of fruits	OA	Downloadable zip file	https://doi.org/10.5281/zenodo.6402251	https://zenodo.org/record/6402251	JPG	(Kaufman et al., 2022)
Thermal images - diseased & healthy leaves	India	<i>Oryza sativa</i>	An intact leaf	636	Handheld TIR camera in the field	TIR	Image-level annotations	Classification of biotic stress	OA	Create Kaggle account to	–	https://www.kaggle.com/sujaradha/th	JPG	

										download zip file	ermal-images-diseased-healthy-leaves-paddy			
Date fruit dataset	Saudi Arabia	<i>Phoenix dactylifera</i>	Shoot	8079	Handheld visible light camera in the field	RGB	Image-level annotations	Detection and maturity classification of fruits	OA	Create IEEE DataPort account to download zip file	http://doi.org/10.21227/x46j-sk98	https://iee-dataport.org/open-access/date-fruit-dataset-automated-harvesting-and-visual-yield-estimation	JPG	(Altaheri et al., 2019b, 2019a)
			Fruit bunch	152										
			Date fruit	256										
Image set for deep learning: field images of maize annotated with disease symptoms	NY - USA	<i>Zea mays</i>	Shoot	7669	UAV	RGB	Line and spline annotation of lesions	Identification of northern leaf blight infected plants	OA	Downloadable zip file	–	https://osf.io/p67rz	JPG	(Wiesner-Hanks et al., 2018)
				10533	Handheld visible light camera in the field									
Vegetable crops dataset for proximal sensing	France	<i>Z. mays</i>	Shoot	1065	Manned mobile platform	RGB	Image-level annotations, bounding boxes, point annotations	Identification and localization of crop stems	OA	Downloadable zip file	https://doi.org/10.17632/d7kbzjr83k.1	https://data.mendeley.com/datasets/d7kbzjr83k/1	JPG	(Lac et al., 2022)
		<i>P. vulgaris</i>		779										
		<i>Allium ampeloprasum</i>		601										
Pheno4D	Germany	Maize	Shoot	84	Controlled environment	3D laser scanner	Semantic segmentation	Estimation of plant traits, growth analysis	OA	Downloadable zip file	–	https://www.pb.uni-	XYZ	(Schunck et al., 2021)
		Tomato		140										

					stationary platform						bonn.de/data/phenod4			
JMuBEN	Kenya	Arabica coffee	A single detached leaf	22591	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/t2r6rszp5c.1	https://data.mendeley.com/datasets/t2r6rszp5c/1	JPG	(Jepkoec h et al., 2021)
JMuBEN2				35964							https://doi.org/10.17632/tqv3zb82nd.1	https://data.mendeley.com/datasets/tqv3zb82nd/1		
BRACOL - A Brazilian Arabica coffee leaf images dataset	Brazil	Arabica coffee	A single detached leaf	1747	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/yv2k5y8mxg.1	https://data.mendeley.com/datasets/yv2k5y8mxg/1	JPG	(Manso et al., 2019)
RoCoLe: a robusta coffee leaf images dataset	Ecuador	Robusta coffee	An intact leaf	1560	Handheld visible light camera in the field	RGB	Image-level annotations	Identification and classification of biotic stress	OA	Downloadable zip file	https://doi.org/10.17632/c5yvn32dzq.2	https://data.mendeley.com/datasets/c5yvn32dzq/2	JPG	(Parraga-Alava et al., 2019)
KOMATSUNA dataset for instance segmentation, tracking and reconstruction	Japan	B. rapa	Shoot	180 60	Controlled environment stationary platform	RGB RGB-depth	Semantic segmentation	Leaf segmentation and plant growth measurement	OA	Downloadable zip file	–	https://limu.a-it.kyushu-u.ac.jp/~agri/komatsuna	PNG	(Uchiyam a et al., 2017)

Single tree point clouds from terrestrial laser scanning	Germany, OR - USA	<i>Quercus petraea</i>	Shoot	22 ^k	Terrestrial laser scanner	3D laser scanner	Image-level annotations	Classification of tree species	OA	Downloadable zip file	https://doi.org/10.25625/FOHJUM	https://data.gettingen-research-online.de/dataset.xhtml?persistentId=doi:10.25625/FOHJUM	XYZ ^o	(Seidel et al., 2021)
		<i>Fraxinus excelsior</i>		39 ^k										
		<i>Picea abies</i>		158										
		<i>Pinus sylvestris</i>		25 ^k										
		<i>Q. rubra</i>		100										
		<i>Fagus sylvatica</i>		163										
		<i>Pseudotsuga menziesii</i>		183										
Early-crop-weed	Greece	<i>S. lycopersicum</i>	Shoot	202	Handheld visible light camera in the field	RGB	Image-level annotations	Identification of crops and weeds	OA	Downloadable zip file	–	https://github.com/AUAgrou/early-crop-weed	JPG	(Espejo-Garcia et al., 2020)
		<i>Gossypium hirsutum</i>		48 ^k										
		<i>S. nigrum</i>		130										
		<i>Abutilon theophrasti</i>		124										
DeepWeeds	Australia	Multiple species ^p	Shoot	8403	Ground-based field robot	RGB	Image-level annotations	Classification of weed species	OA	Downloadable zip file	–	https://github.com/AlexOlsen/DeepWeeds	JPG	(Olsen et al., 2019)

^aAbbreviations: ACFR, Australian center for field robotics; CF, chlorophyll fluorescence; FGVC7/8, the seventh/eight workshop on fine-grained visual categorization; GARNICS, gardening with a cognitive system; HDF5, hierarchical data format version 5; HS, hyperspectral; JMuBEN, Jepkoech, Mugo and Benson; JPG, joint photographic experts group; LiDAR, light detection and ranging; MAT, Matlab; MS, multispectral; MSU-PID, Michigan State University-plant imagery database; NIR, near infrared; OA, open access; PID, persistent identifier; RGB, red–green–blue; TIF, tag image file format; TIR, thermal infrared; UAV, unmanned aerial vehicle.

^bWe identified 56 publicly available global datasets and their characteristics using Google's search engine and Google Dataset Search. The search combined terms describing various plant organ systems, sensors, artificial intelligence (AI) techniques, as well as dataset and database. All pages for each search were systematically collated and screened. Additional datasets are available in

repositories containing large amounts of OA imaging data. Repositories such as the National Ecological Observatory Network (<https://data.neonscience.org/data-products/explore>), Leafsnap (<http://leafsnap.com>), the Institut National de la Recherche Agronomique (<https://data.inrae.fr>), the United States Geological Survey (<https://www.usgs.gov/products/data-and-tools/science-datasets>), the National Aeronautics and Space Administration earth science data (<https://earthdata.nasa.gov>), the plant genomics and phenomics research data repository (<https://edal-pgp.ipk-gatersleben.de>), the computer vision and biosystems signal processing group (<https://vision.eng.au.dk/data-sets>), the Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (<https://terraref.ncsa.illinois.edu/clowder>), figshare (<https://figshare.com>), Dryad (<http://datadryad.org>), the International Maize and Wheat Improvement Center (<https://data.cimmyt.org>) and the *Arabidopsis thaliana* phenotyping database (Phenopsis DB, <http://bioweb.supagro.inra.fr/phenopsis>) provide datasets in downloadable zip files. Similarly, the Oak Ridge National Laboratory Distributed Active Archive (https://daac.ornl.gov/get_data/#themes) provides datasets in downloadable zip files after registering for an account, as well as the University of Nebraska-Lincoln Plant Vision Initiative (<https://plantvision.unl.edu/dataset>) and the X-Plant (<http://www.x-plant.org>) after filling a form. Other sources such as the online database for plant image analysis software tools (<https://www.quantitative-plant.org/dataset>) and the registry of research data repositories (<https://www.re3data.org>) are designed specifically for the discovery of datasets in various repositories.

^cPlant species were reported whenever they were available in the corresponding referenced paper(s); common names were reported otherwise.

^dImages of datasets with semantic segmentation annotations are completely annotated images, where a class is assigned to each pixel.

^eNo datasets were excluded on the basis of access type (i.e., OA, data available on request, or OA with barriers – datasets fulfilling criteria for OA but being inaccessible because of unpredictable reasons such as broken hyperlinks).

^fData PID is a long-lasting digital reference to a dataset, such as a digital object identifier (DOI). A dash (–) indicates that no PIDs are available. DOIs for datasets can be issued automatically by the hosting repositories (e.g., Zenodo, GigaDB, Mendeley Data, and IEEE DataPorts). As datasets should be cited to ensure credit to those who produced and curated them, we recommend that they should include a PID and the minimum metadata suggested by DataCite (a non-profit membership organization that provides DOIs for research data) and FORCE11 (a community of scholars, librarians, archivists, publishers and research funders), i.e., author, year, title, and repository. Data producers can be inferred based on the author contributions of the corresponding referenced paper(s) while data curators can be inferred based on the author(s) that published the dataset to a repository.

^gFile format defines the structure and encoding of the data stored in it and thus guides researchers on how to programmatically input such data to their AI algorithms.

^hThe same number of images was taken with each sensor.

ⁱDatasets containing more than one object per image (e.g., multiple hypocotyls, fruits, flowers). When segmented, each image could become hundreds of samples to train an AI algorithm.

^jTwo multispectral cameras were used: a five-band RedEdge-M camera in Germany and a four-band Sequoia camera in Switzerland.

^kFor datasets with small image number, transfer learning can be applied, giving an AI model a warm start by applying information learned from another previously trained model.

^lCrops and their corresponding number of images: Apple, 3171; blueberry, 1502; cherry, 1906; corn, 3852; grapevine, 4062; orange, 5507; peach, 2657; pepper, 2475; potato, 2152; raspberry, 371; soybean, 6925; strawberry, 1565; tomato, 18160.

^mSpecies and their corresponding number of images: *M. indica*, 435; *Terminalia Arjuna*, 452; *Alstonia Scholaris*, 433; *Psidium guajava*, 419; *Aegle marmelos*, 118; *Syzygium cumini*, 624; *Jatropha curcas*, 257; *Pongamia Pinnata*, 598; *Ocimum basilicum*, 149; *Punica granatum*, 559; *Platanus orientalis*, 223; *C. limon*, 236.

ⁿ*Passiflora* species and their corresponding number of images: *P. coriacea*, 208; *P. misera*, 215; *P. biflora*, 105; *P. capsularis*, 118; *P. micropetala*, 68; *P. organensis*, 84; *P. pohlii*, 16; *P. rubra*, 87; *P. tricuspis*, 257; *P. caerulea*, 99; *P. cincinnata*, 84; *P. edmundoi*, 111; *P. gibertii*, 192; *P. hatschbachii*, 132; *P. kermesina*, 113; *P. mollissima*, 69; *P. setacea*, 189; *P. suberosa*, 352; *P. tenuifolia*, 113; *P. amethystina*, 119; *P. foetida*, 304; *P. gracilis*, 81; *P. morifolia*, 57; *P. actinia*, 95; *P. miersii*, 133; *P. sidifolia*, 145; *P. triloba*, 295; *P. alata*, 235; *P. edulis*, 119; *P. ligularis*, 139; *P. nitida*, 62; *P. racemosa*, 194; *P. villosa*, 58; *P. coccinea*, 169; *P. cristalina*, 220; *P. galbana*, 161; *P. malacophylla*, 168; *P. maliformis*, 156; *P. miniata*, 129; *P. mucronata*, 116.

^oA point cloud data file in XYZ format contains rows of data, each consisting of x, y, and z coordinates of a point.

^pSpecies and their corresponding number of images: *Ziziphus mauritiana*, 1125; *Lantana camara*, 1064; *Parkinsonia aculeata*, 1031; *Parthenium hysterophorus*, 1022; *Vachellia nilotica*, 1062;

Cryptostegia grandiflora, 1009; *Chromolaena odorata*, 1074; *Stachytarpheta* spp., 1016.

4.2 How to use AI effectively: ménage-à-trois between plant science, data science, and IS

Plant science has the potential to provide innovative solutions to the world's most pressing challenges; however, recent advances in discovery methods have greatly accelerated our ability to collect data, leaving us with the challenge of analyzing, interpreting, and integrating the plethora of data (Brink, 2021). To handle such data, data science has attracted considerable attention, and promises to turn data into useful predictions and insights (Cao, 2017). To achieve that, data science needs supporting resources including algorithms, software, and hardware infrastructure. IS combines those resources to create AI architecture designs and to transform, store, and distribute data for analysis. Although the relationship between these disciplines has not been repeatedly reinforced in history, given the current depth of data analysis, the scale and dimensions of the data, and the nature of the scientific questions, an interaction in a ménage-à-trois fashion is highly needed.

4.2.1 *AI and the bias cascade*

Multiple sources of bias can affect the performance of AI systems used in phenomics and can occur across the different development steps of AI applications: data collection or selection, data preprocessing, model development, model evaluation, and deployment. Introduced bias can have a domino effect as it propagates from its entry point to the succeeding development steps, creating a bias cascade.

The bias cascade starts with the data collection or selection step, where experimental data are collected or selected from publicly available datasets (Table 4.1). Bias here can occur for several reasons: (i) 'measurement' bias, when the data contain faulty measurements originating from instrumentation malfunctions, wrong values from miscalibrated sensors, or errors of precision that result in data distortion (Xu et al., 2015); (ii) 'label' bias, when the data are laden with subjective judgments of human experts and are thus inconsistently or wrongly labeled (Miceli et al., 2020); (iii) 'sample selection' bias, when the training data do not represent a random sample from the entire dataset (Akter et al., 2021), causing a model to ignore data belonging to classes that were not represented during the data selection process; and (iv) 'group

attribution' bias, when a data sample is selected from an incorrect target population (Akter et al., 2021), and where a model can fail to distinguish between some classes and consider them to be the same.

Data preprocessing is performed to eliminate noisy (e.g., blurred images, images with unfavorable lighting conditions, images that do not represent the object of interest), incomplete (e.g., unannotated images), and duplicate data, and to normalize datasets as needed to account for batch effects (e.g., groups of images taken under different lighting conditions or with different camera settings) or systematic experimental artifacts (e.g., reflections in images). In this step, even if the training set is representative of the entire dataset, this can generate an intrinsically **unbalanced dataset** in which some plant species, genotypes, or even stresses are under-represented. Such cases can introduce the 'class imbalance' bias.

Bias may also arise during the model development and evaluation steps, where a model is trained and its ability to generalize beyond the training set on new and previously unseen data are evaluated. Because most AI algorithms identify correlations between variables in the underlying data but without being able to detect causal relations, two biases are likely to arise: (i) the 'correlation fallacy' that confuses correlation with causation (Akter et al., 2021) where a model wrongly deduces a cause-and-effect relationship between correlated variables, and (ii) the 'apophenia' when a model sees patterns whereas none actually exists (Boyd and Crawford, 2012). These two biases can be amplified when a massive quantity of training data are used, mistakenly offering connections that radiate in all directions (Boyd and Crawford, 2012), and producing probable but uncertain predictions. Further, training complex models (i.e., models with many trainable parameters) can capture noise-generated patterns, tricking them into thinking that the noise encodes real information (Mehta et al., 2019). This problem introduces an 'overfitting' bias and causes a steep drop-off in predictive performance at the evaluation step. Similarly, such performance drop-offs also occur when models are unable to accurately capture relationships between variables and thus introduce an 'underfitting' bias (Greener et al., 2022).

Finally, at the deployment step, bias can occur in situations where the data used in practice differ from the training data (e.g., different weed or crop species), which is known as the 'domain shift' bias.

4.2.2 *Creating a human-centric X-AI*

It is therefore crucial to mitigate bias to increase the success probability of the AI algorithm for the task at hand. Bias mitigation serves as a building block toward AI **trustworthiness** (Schwartz et al., 2021). What can be done to mitigate detrimental biases in AI in plant phenomics? There is a consensus on the need to develop a human-centric X-AI system that will not only aspire to meet human requirements regarding explainability and trustworthiness but, more importantly, will actively aim to keep a HITL for harmonious human and AI system symbiosis. We believe that such a system should not only put humans at its center but also integrate their knowledge into its predictive process. A close-knit team of specialists with knowledge and experience across a broad range of design science research needs, and spanning various disciplines of science, as well as of relevant stakeholders is thus essential to building a precise and impactful HITL.

Designing human-centric X-AI for plant phenomics is not without challenges; it requires a dedicated and multidisciplinary team effort involving plant scientists, data scientists, and IS scientists to bring AI to its most feasible, desirable, viable, and responsible state. This novel multidisciplinary knowledge is clearly imperative to identify and reduce AI biases, and to facilitate explainability and accountability. We advocate that such a system architecture should constantly realign data architecture and technology infrastructure to serve novel AI architecture designs. Phenotyping complex traits demands the integration of data on different morphological, physiological, temporal, geospatial, and environmental variables (Großkinsky et al., 2015; Harfouche et al., 2019). Although large datasets are vital for creating accurate AI models and validating their results, storing them in a FAIR manner can be challenging. Data architecture plays a fundamental role in meeting these requirements. It consists of a set of standards that govern which data are collected, whether they should be transformed (e.g., data cleaning, deduplication, format conversion, structuring, validation, etc.) before or after storage using extract, transform, load (ETL) or extract, load, transform (ELT) processes, and where (data warehouses or data lakes) and how (matrices, cubes, polytopes, or distributed in-memory) the data are stored (Figure 4.2). Without AI, these datastreams would be overwhelming and chaotic (Harfouche et al., 2019), but reaching the full potential of AI-based analysis of large phenomic datasets depends upon the right technology

infrastructure which defines the components that serve as a foundation for the data lifecycle, including hardware infrastructure, network flow, software frameworks, and programming languages (Figure 4.3). High-performance computing (HPC), including pre-exascale supercomputers, is boosting both the accuracy and predictive power of these approaches. Although central processing units (CPUs) maximize the performance of an algorithm, graphics processing units (GPUs) can dramatically increase AI training speed because their processing cores were initially designed to process visual data such as images and videos (Nakhle and Harfouche, 2021). For example, to take advantage of GPUs, the compute unified device architecture (CUDA) software framework provides a development environment for creating and optimizing AI applications on **GPU-accelerated** local computers or supercomputers. However, CUDA works exclusively on Nvidia GPUs; alternatively, the open computing language (OpenCL) and openACC frameworks work on multiple types of GPUs (Matthews, 2018). Another option is to translate automatically CUDA source code into a portable heterogeneous computing interface for portability (HIP) using source-to-source translators such as HIPify, so that non-Nvidia GPUs can benefit from the rapid development of CUDA applications. In addition, software libraries such as kokkos, RAJA, open multi-processing (OpenMP), and one application programming interface (oneAPI) can be leveraged to unlock the promise of heterogeneous computing where **compute nodes** employ more than one type of processor, including CPUs, GPUs, and tensor processing units (TPUs), among others. This enables the development of scalable AI-based applications in a hardware-agnostic way. With the advent of exascale computing, supercomputers will deliver higher performance in pattern searching in phenomic big data, and will thus boost AI abilities in digital phenomics, speeding up crop design (Figure 4.3A). However, building powerful supercomputers is a never-ending race, and as new ones are launched the number of compute nodes they comprise increases. For example, the first supercomputer to break the exascale barrier, Summit, comprises 4608 compute nodes, whereas the most powerful exascale supercomputer that tops the latest TOP500 list^{vii}, Frontier, contains 9472. This makes it more difficult to exploit supercomputers efficiently because of their need to transmit data back and forth between their nodes while running huge numbers of computations at the same time (Mann, 2020). Implementing AI algorithms (Figure 4.4A) for such parallel computing is not easy. Luckily, emerging free and open-source software frameworks such as Tensorflow Keras, PyTorch, scikit-learn, and XGBoost, among

others, as well as software libraries such as cuNumeric, are enabling scalability in parallel computing. More powerful exascale supercomputers are being anticipated (Skibba, 2021), and researchers may start to utilize quantum computers at some point in the future (Marx, 2021). This will ultimately drive digital phenomics toward designing faster, better crops and providing sustainability-friendly solutions (Figure 4.3A). In addition to hardware infrastructure, properly designed network flows (Figure 4.3B), such as the 'science demilitarized zone (DMZ)' that includes network architecture and performance tools (Dart et al., 2014), enable high-throughput access to datasets in a secure and timely manner while conforming with the FAIR data principles (Wilkinson et al., 2016). Software frameworks (Figure 4.3C) provide a working environment that helps researchers to achieve higher productivity in designing AI algorithms; they support more than one programming language (Figure 4.3D), enabling fast and efficient implementation of algorithms without compromising code quality.

Because data are only as good as the tools and algorithms that are available to analyze them, solving complex biological questions requires a creative process during which efficient AI algorithm architectures are designed and developed. Customized algorithms and architectures can leverage currently available AI architecture designs (Figure 4.4) to develop new architecture designs tailor-made to find answers to the questions at hand. Such promising designs should combine knowledge-based AI, to represent human expert knowledge, with data-driven AI to discover connections and correlations automatically in big data. This combination will result in an informed AI that acquires both the **tacit knowledge** and the **explicit knowledge** of its designers (e.g., the interaction between data scientists, IS scientists, and plant scientists) and users (e.g., breeders and farmers), and integrates that tacit and explicit knowledge with knowledge discovered from data and metadata.

It is noteworthy, however, that new architecture designs should also integrate knowledge into X-AI to enable monitoring of the inputs and outputs of the algorithms, provide more human-comprehensible explanations for their decisions, deliver superior performance, mitigate bias, and aid in verifying the adherence of models to ethical and sociolegal values. Ensemble methods can, for example, be leveraged to design new AI algorithms that are both informed and explainable (Figure 4.4B). Ultimately, improvements in informed X-AI would help to develop novel interpretable algorithms and are likely to be crucial to enable human-centric X-AI in phenomics.

4.2.3 Mitigating bias in human-centric X-AI

A human-centric X-AI system is emerging whereby plant scientists, data scientists, and IS scientists must work together to seize this opportunity to help to identify and mitigate bias by using several strategies.

Starting from the top of the bias cascade, at the data collection or selection step, the minority of data that do not conform to the general characteristics of a given dataset, known as outliers, should be removed during data cleaning to mitigate 'measurement' bias. Regarding 'label' bias, data annotators should be supplied with detailed instructions containing visual examples of the correct output for a given input to be able to reduce ambiguities and avoid mistakes that result from incorrect or incomplete knowledge. For example, when labeling weed species, in addition to their morphological descriptions, a visual representation of each species could be helpful for annotators. Next, 'sample selection' and 'group attribution' biases can be mitigated by establishing random sample selection and statistical correction processes (Castro et al., 2020; Phillips et al., 2009).

When preprocessing data, intrinsically unbalanced datasets can be balanced by means of oversampling (i.e., augmenting the number of training examples in the minority class to be equivalent to other classes) and/or undersampling (i.e., reducing the number of training examples in the majority class to be equivalent to other classes) (Nakhle and Harfouche, 2021) to eliminate 'class imbalance' bias (Buda et al., 2018).

Properly sampled and preprocessed data mitigate the risk of 'correlation fallacy' and 'apophenia' biases, which can occur during the model development step. When training, 'overfitting' can be debiased by increasing the size of the training data, decreasing the model complexity, or ignoring the less important features in a process called regularization (Lever et al., 2016). 'Underfitting' bias can be resolved by increasing the complexity of the model to capture nonlinear relationships in data. During the evaluation step, models yielding incorrect predictions such as misclassifying crops as weeds or vice versa should be inspected carefully; X-AI can be leveraged to better understand how the model reached its predictions which helps to identify previously unknown biases. However, *post hoc* approaches to explainability are not necessarily transparent (i.e., because they only approximate the prediction

procedure of a model), and thus it might be better to employ interpretable by design models (see next section).

Notably, to avoid 'domain shift' bias and identify unwanted biases in deployment, it is crucial that the model is carefully monitored to assess whether the data being used in practice are representative of those used during training. It is important to note that risk-based regulations of AI are on the horizon in the USA^{viii} and Europe^{ix,x}. When new regulations enter into force, post-authorization monitoring of AI applications becomes crucial to ensure that the performance of models does not degrade in practice. Once a model has passed regulatory authorization and is implemented in phenomics, it needs to be retrained periodically using new datasets to prevent it from becoming outdated, thus ensuring 'domain shift' bias mitigation.

Reducing the risk of bias in AI models requires continuous human attention across the five development steps, keeping HITL. Studies have shown that human–computer interaction in HITL AI has improved the predictive performance of AI-based image analysis and reduced biases (Budd et al., 2021; Jin et al., 2020; Sreeram and Nof, 2021). HITL can make a significant impact on phenomic data collection, data preprocessing, model development, evaluation, and deployment. It plays a crucial role in the collection and preparation of data to be used for training an AI model. As such, model training is often a HITL iterative process that identifies biases or weaknesses of the model (e.g., images on which the model fails due to incomplete training sets or inappropriate parameterization), and adjusts the training set and parameters to reduce any biases and ensure the best model performance. It is recommended to start each training step with small iterations and plan on how the feedback of the team of humans can be collected and propagated to other steps, relying on their intelligence to perform complex tasks. This paradigm allows leveraging the advantages of AI while humans at various checkpoints fill the gaps where models are not confident in their predictions or where they may fall short because of underlying biases (Patel et al., 2019). HITL may also offer advantages to evaluating the accuracy of AI predictions and interpreting their decisions by interacting with explainable models. The benefits of HITL extend to deployment by monitoring the model for possible biases and ensuring the reliability of the AI system. HITL can feed back into itself to respond to changes in the real-world environment. For example, after data collection and preprocessing, in each training iteration, plant scientists are shown a list of misclassified images with the outputs of

the AI algorithm to hand-verify predictions and assess false positives and false negatives. For instance, the model might misclassify crops as weeds; this could be due to an algorithmic or learning error, or to mislabeled images. They then correct the wrong labels, if any, to ensure high-quality data. Data scientists evaluate the model, tune its hyperparameters, and retrain it. Such iterations between humans and AI are effective for generating training data based on human judgment to increase learning efficiency and enhance model performance (Patel et al., 2019). Data scientists can provide the expertise necessary to help IS scientists to design AI architectures with explanatory capacity supported by theoretical underpinnings. Finally, HITL monitors the model outcomes after deployment to ensure that all biases are identified and mitigated. Furthermore, ensembles of models can be used in the HITL process. An intrinsically interpretable model such as an iterative RF (Basu et al., 2018; Cliff et al., 2019) can be used initially for feature engineering to determine the variables (e.g., wavelet decomposition in RGB and hyperspectral images) that will then be used in a DL predictive model.

This three-way collaboration can amplify knowledge about domain-specific feature engineering and selection to reach a level of augmented intelligence that can help in discovering new ways to make AI more efficient, less biased, and explainable. It also creates new opportunities for human-centric X-AI to predict desirable phenotypic traits and aid efforts to breed climate-proof crops fast enough.

4.3 How to move from data inputs to outcomes: opening the black box or designing a transparent glass box for explainability

AI continues to permeate plant phenomics, as recently reviewed (Danilevicz et al., 2021; Jiang and Li, 2020; Mochida et al., 2019). However, complex AI models are difficult to explain even among data scientists; they operate as black boxes and require a leap of faith to believe their predictions (Harfouche et al., 2019). Explainability of AI models would not only increase the trust of users in why and how predictions were made but also help data scientists to enable better diagnostics and enhance their performance. Although these desirable properties of explainability have led to a recent growing interest in X-AI research (Vilone and Longo, 2020), its origin traces back to

the early 1970s when Edward Shortliffe introduced the AI-based antimicrobial therapy consultation system for assisting physicians who need advice about appropriate therapy. The system made use of a set of decision rules that were coded, categorized, and hand-entered into it to give advice and explain the reasons behind its predictions (Shortliffe et al., 1973). In 1979, Jon Doyle introduced the truth maintenance system (TMS), an independent module that constructs explanations of predictions by recording and maintaining a representation of the knowledge acquired by an expert system (Doyle, 1979). TMS research and development continued until the 1990s, when researchers began to study the possibility of extracting meaningful explanations from non-hand-coded rules that are generated by trained models such as NNs (Tickle et al., 1998).

The rise of DL in the 2010s (Nakhle and Harfouche, 2021) increased the complexity of AI models and consequently the demand for X-AI algorithms. To address this issue, researchers have been developing new approaches and techniques to make these models explainable. Unfortunately, the rush in X-AI development has caused confusion about its various approaches in the literature, where they are not accurately described and are often confused with one another (Lipton, 2018). Although all these approaches revolve around allowing humans to observe how the predictions of an AI model came to be, we can technically distinguish between research involving *post hoc* models and interpretable by design models.

Because current AI models are often developed with only predictive performance in mind, *post hoc* algorithms can be used to explain them. They are employed after a black box model is trained and are not connected to its internal design; they can either be model-specific or model-agnostic (Molnar, 2019). In principle, model-specific algorithms are limited to particular black box models. For example, DL important features (DeepLift) is a model-specific algorithm that can explain DNNs and does not work for any other algorithm. By contrast, model-agnostic algorithms such as the local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) and Shapley additive explanation (SHAP) (Lundberg and Lee, 2017) are more general and can be applied to any black box model. Commonly, *post hoc* algorithms work by (i) probing or inspecting the trained parameters to understand what the black box model has learned, (ii) employing data perturbation strategies which involve modifying the input data and observing the changes in the black box model predictions, or (iii) using a

more interpretable model (e.g., a decision tree) referred to as a surrogate model to approximate and provide explanations of predictions made by the black box model. Recently, researchers have started to apply *post hoc* algorithms in plant phenomics to identify, classify, and quantify plant stresses (Ghosal et al., 2018; Mostafa et al., 2021; Nagasubramanian et al., 2020, 2019; Nakhle and Harfouche, 2021; Toda and Okura, 2019), as well as to count leaves (Dobrescu et al., 2019). However, because *post hoc* algorithms approximate the inner workings of black box models, it is possible that their generated explanations do not provide sufficient detail to understand what the black box model is actually doing (Rudin, 2019). By contrast, interpretable by design algorithms do not need an additional (*post hoc*) algorithm to be explainable; they provide their own explanations which are faithful to what the model actually computes (Rudin, 2019).

These algorithms have existed since the development of expert systems in the 1970s. They have, however, been labeled as less accurate because scientists argue that there is a tradeoff between accuracy and explainability in a way that the highest performing algorithms are the least explainable and the most explainable ones are less accurate^{xi}. This belief proved to be imprecise, especially when analyzing structured data with meaningful features (Rudin, 2019). This also depends on the algorithms being compared. For example, according to Rudin (Rudin, 2019) it would not be fair to compare the 1984 decision tree algorithm to a more recent DL algorithm, and conclude that interpretable by design models are not as accurate. Indeed, the recently developed interpretable by design 'this looks like that' algorithm, derived from a CNN, proved to be as accurate as the non-explainable CNN (Chen et al., 2019). Figure 4.5 highlights the two categories of X-AI, their corresponding representative algorithms, and the explainable outcomes associated with their implementation.

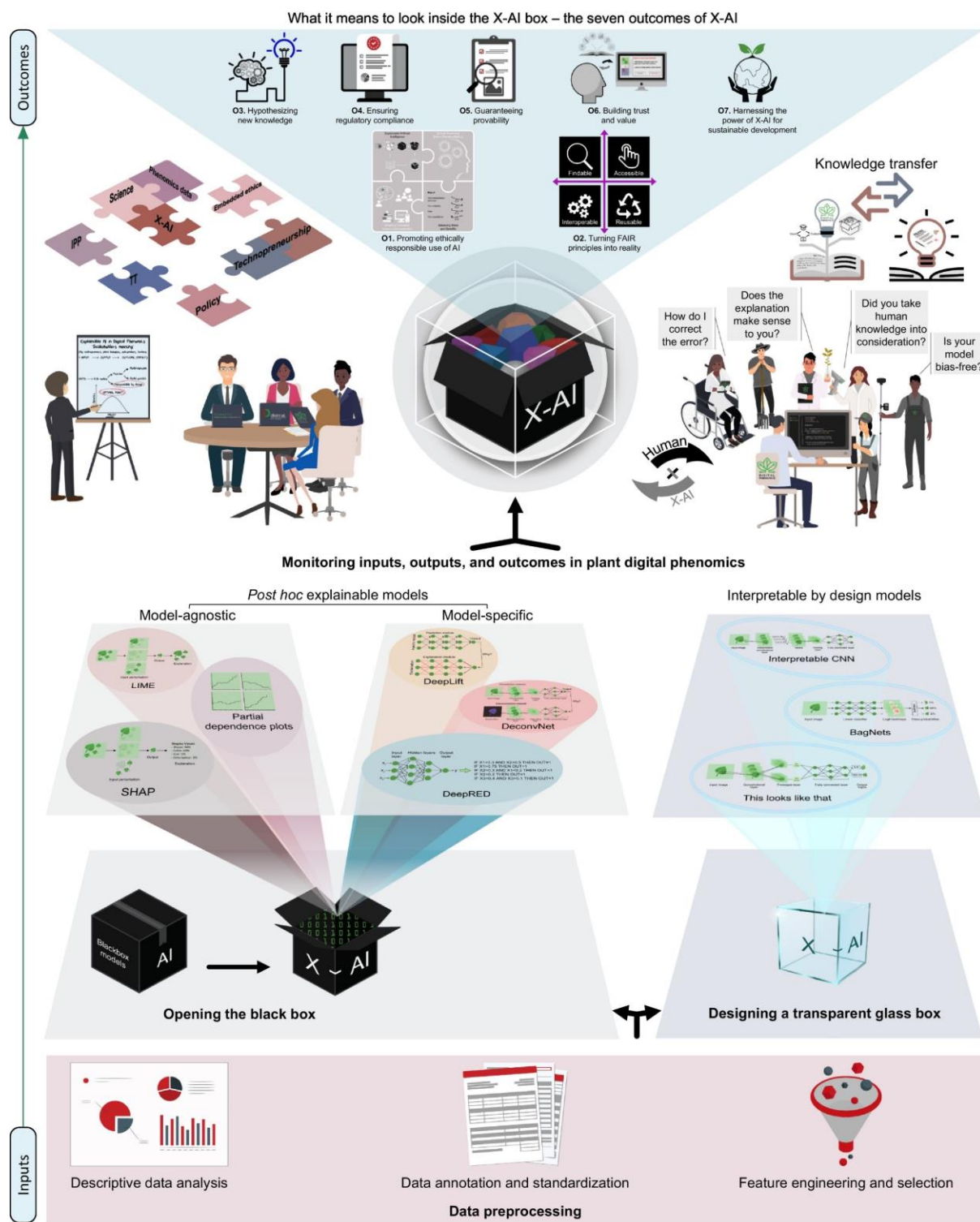


Figure 0.5. Cultivating the conditions for explainable artificial intelligence (X-AI) to flourish in plant digital phenomics.

Data preprocessing prepares input data for X-AI algorithms: descriptive data analysis provides statistical summaries about a dataset to detect anomalies; data annotation and standardization are performed by labeling and adding relevant, structured information about the data such as their source and other details known as metadata; and feature engineering uses existing features to create new

ones whereas feature selection extracts relevant features from the complete set of features in a dataset, thus increasing the predictive precision of learning algorithms. X-AI can be achieved by either opening the black box or designing a transparent glass box. X-AI can be interrogated to understand why a decision has been made, keeping human-in-the-loop (HITL) of such decision making, and allowing two-way transfer of knowledge where, on the one hand, experts assist in the training of X-AI and, on the other hand, explanations can be used to generate scientific hypotheses that can result in new discoveries. An X-AI that takes into account the requirements of all stakeholders interacting with it will drive successful adoption by agricultural technopreneurs, plant biologists, policymakers, and funders. This will help to bridge the gap between science, policy, embedded ethics, and entrepreneurship, allowing responsible TT and leading to technological, regulatory, and social and ethical outcomes. Abbreviations: AI, artificial intelligence; CNN, convolutional neural network; DeconvNet, deconvolution network; DeepLift, deep learning important features; FAIR, findable, accessible, interoperable, reusable; IPP, intellectual property protection; LIME, local interpretable model-agnostic explanations; SHAP, Shapley additive explanation; TT, technology transfer.

Regardless of whether a *post hoc* or an interpretable by design algorithm is used, model explanations can occur at a global or local level. Whereas the former describes the overall extracted relationships based on the entire model behavior, the latter reveals the rationale behind a specific prediction (Barredo Arrieta et al., 2020).

Finally, it is worth noting that, in the same way as there are diverse X-AI techniques, a range of approaches to explainability are necessary because different contexts give rise to different explainability needs (Preece et al., 2018). For example, when training and evaluating an AI model, plant scientists might want to understand which data features are being used for prediction and how they correlate with each other, whereas data scientists might require technical details about how the model functions to help in its testing, debugging, bias identification and mitigation, hyperparameter tuning, and evaluation; IS scientists can leverage details about the model training process to help to optimize the architecture of the algorithm by using suitable design approaches and methods (Figure 4.4B). At the model deployment step, regulators might require assurance about how data are being processed to assess its risk level by inspecting its reliability and the impact of its predictions on its users, to ultimately decide whether or not it requires authorization and regulation. Similarly, farmers and breeders might require explanations to understand why and how the model came to a prediction and to ensure its trustworthiness.

Presently, the hope for human-comprehensible explanations for black box algorithms to increase technical confidence, generate trust, and make better-informed

choices remains an open challenge. In light of this challenge, we strongly recommend that a single prediction might therefore need to be explained in various ways, reflecting the requirements of all stakeholders.

4.4 Methods and results – how to devise X-AI-driven analytics for phenomics questions

4.4.1 *Interpretable by design models*

X-AI holds great potential for the analysis and interpretation of phenomic data. In the following we provide an example of an X-AI workflow design and describe for the first time the steps necessary to foster practical applicability of interpretable by design algorithms in phenomics image analysis (Figure 4.6). We have also accompanied this review by an interactive online tutorial that acts as an educational resource intended for readers with little to no knowledge of X-AI algorithms; it also serves as a good starting point for self-learning and raises early awareness that computational phenomics need not be intimidating. In addition, we have created a set of self-test quizzes and hands-on practice exercises to provide users with opportunities to augment their learning by practically applying the concepts explained to assess their acquired knowledge. The code and computational notebooks are open source and freely accessible through our GitHub repository^{xii}. Collectively, this will accelerate the rate of discovery and move toward open science and AI ethics in digital phenomics.

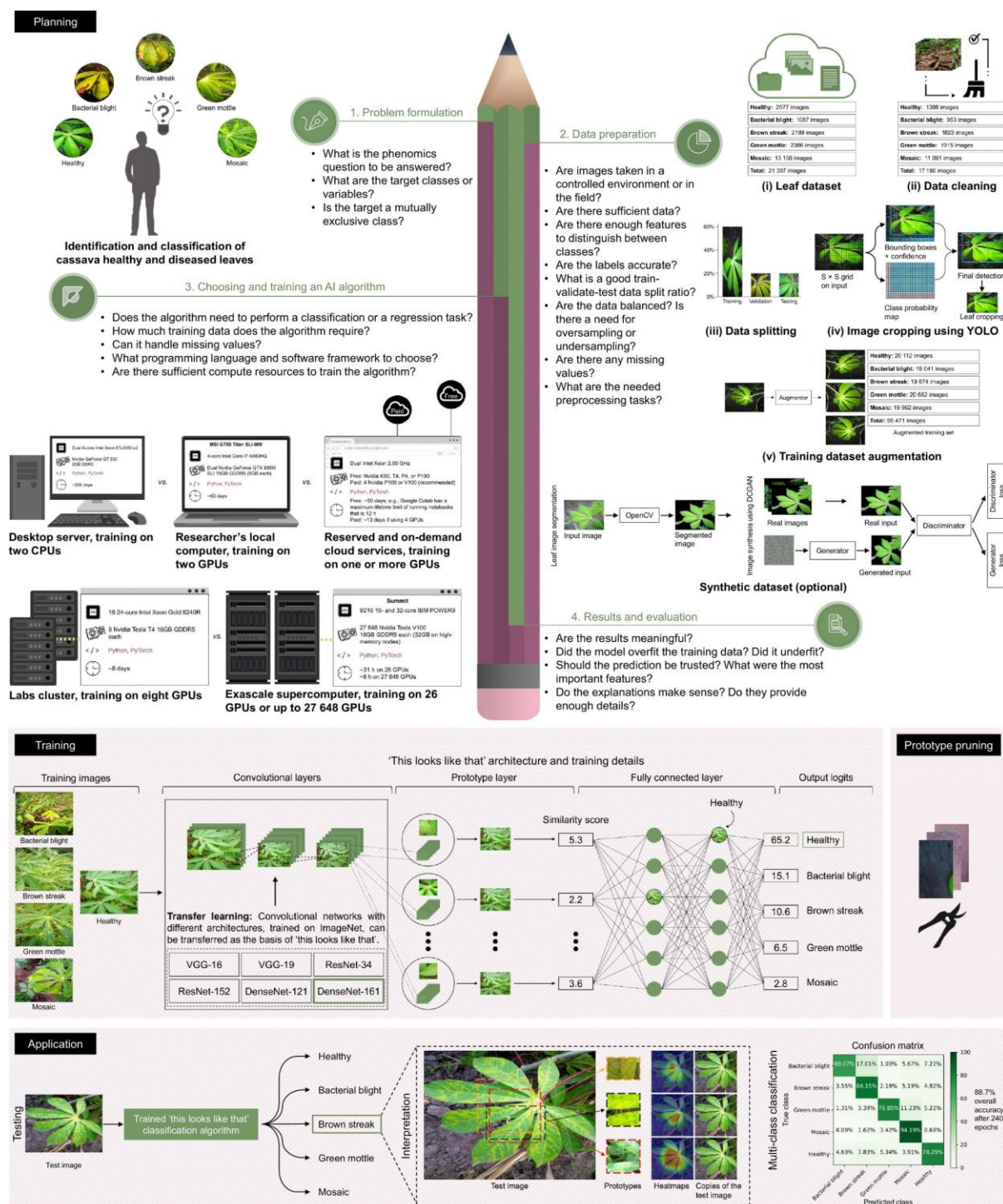


Figure 0.6. Planning, training, and interpreting an explainable artificial intelligence (X-AI)-based analysis in plant digital phenomics requires careful consideration at each stage of the analysis.

This figure sheds light on all the elements of designing such a workflow using a cassava leaf disease classification task as an example. For data preparation (i) a dataset shared on Kaggle by the AI laboratory at Makerere University was used for analysis; (ii) data cleaning was carried out to eliminate outliers and mislabeled images; (iii) the dataset was randomly split for training, validation, and testing; (iv) another shared version of the dataset with images cropped to leaf boundaries using a trained YOLO

model was used to minimize noise in training images; and (v) the training dataset was augmented and balanced by oversampling and creating random transformations to image geometries. An alternative solution to oversampling is to synthesize leaf images; OpenCV can be used to segment leaves to train a deep convolutional generative adversarial network (DCGAN) to generate synthetic data. A 'this looks like that' interpretable by design algorithm, implemented in Python and PyTorch, was carefully chosen for the classification task; its training time was approximated and compared on different hardware, showing the advantages of GPUs over CPUs and exascale over pre-exascale supercomputers. However, increasing the number of GPUs comes at the price of increased network communication and input–output (I/O) operations to synchronize the model over cluster nodes. Such overheads can cause a delay in the training time. For example, although the algorithm is expected to complete 1000 training epochs in 31 h using 26 Nvidia Tesla V100 GPUs, it is still expected to take ~8 h using the full power of the Summit supercomputer (27 649 GPUs). The 'this looks like that' algorithm uses transfer learning to import convolutional layers from pretrained models and, during training, the prototype layer extracts parts of training images (prototypes) and learns a similarity metric between them; the final class prediction is based on the weighted sum of similarities between the input and prototypes. For some prototypes, the nearest image patches come from different classes, and often correspond to a background patch, and thus should be pruned. For interpretation, the model tries to find evidence that a test image belongs to a specific class, and marks activated patches by bounding boxes. Whereas heatmaps show which part of the image is similar to a prototype, the confusion matrix illustrates the percentage of images of a true class classified into the class indicated by the predicted class column, indicating an overall accuracy of 88.7% after 240 training epochs. Abbreviations: CPU, central processing unit; DenseNet, dense convolutional network; GPU, graphics processing unit; OpenCV, open source computer vision library; ResNet, residual network; VGG, visual geometry group; YOLO, you only look once.

In our tutorial we train the 'this looks like that' algorithm to classify diseases using the **crowdsourced** cassava disease classification dataset. This choice is motivated by the importance of cassava – a key food security crop grown by smallholder farmers in Africa, Asia, and South America. However, diseases that plague this crop are a major cause of poor yield (Otim-Napr et al., 2001). Existing methods to identify diseases require governmental agricultural experts to visually inspect and diagnose the plants (Mwebaze et al., 2019). This labor-intensive process makes it difficult to monitor and treat disease progression. With the help of X-AI, we can identify and classify cassava diseases and monitor their progression sufficiently rapidly to address these current limitations in disease surveillance. Our model, initialized with transfer learning (TL), was trained to predict five classes and provide corresponding prototypical explanations by marking activated patches with bounding boxes and generating heatmaps to show which parts of the image are similar to the prototypes.

The resulting confusion matrix illustrates the percentage of correctly classified images in each class. The overall accuracy was 88.7% after cycling through the training set 240 times (Figure 4.6).

A more detailed description of all steps of the analysis, including the computational notebook and code to train, validate, and test/replicate our models, is provided on the tutorial website. This description covers relevant factors such as data preprocessing, image classes and formats, the architecture of the model, model training and evaluation, prototypical explanations, and the **computer cluster** used for training.

4.4.2 Dealing with small datasets

Because some phenotyping experiments generate small amounts of data, X-AI models have fewer training examples to learn from. How can models learn well from small datasets? Low-complexity models that have a small number of trainable parameters can perform better than complex models because they are less prone to overfitting and generalize better (Brigato and Iocchi, 2021). In addition, TL can be employed to transfer knowledge acquired while learning a different but related task from a model trained on a large dataset to fit a new model using a small dataset (Gupta et al., 2021). When TL is not sufficiently powerful because of the absence of large datasets, cumulative learning (CL) can be used to train a model over various small datasets and accumulate knowledge in the resulting network representation (i.e., model weights) (Seddiki et al., 2020). Even when pretraining a model with TL or CL is not possible, the cosine loss function can substantially improve the predictive performance of the model (Barz and Denzler, 2020). Whereas the loss function of the model measures the error between the input and predicted output, the cosine loss function maximizes the cosine similarity between them. One-shot or few-shot learning can also be used to train a model from one or a handful of training image data by basing predictions on a similarity metric (e.g., cosine similarity) that compares training data to new inputs (Wang et al., 2021). Most recently, with the embedding of human knowledge into AI, it will be possible to supplement small training datasets. Representation of such knowledge can be incorporated into AI by means of changes to the input data and loss function (Childs and Washburn, 2019), to the architecture of the algorithm (Hasebe, 2021), or to a combination thereof. Alternatively, oversampling can be another workaround that produces new sample data to augment small

datasets. These new data, however, should be meaningful, sufficient, and realistic, and should contribute to better performance of predictive models (Shamsolmoali et al., 2021). Oversampling can be achieved by (i) performing geometric transformations on existing images using primitive data-manipulation techniques, including flipping, rotation, shearing, cropping, and translation (Khalifa et al., 2022); (ii) generating new synthetic data with generative adversarial networks (GANs), which are powerful models for learning complex distributions to synthesize semantically meaningful samples from an actual training set (Shamsolmoali et al., 2021) – GANs can be employed for image-to-image translation, fusion image generation, label-to-image mapping, and text-to-image translation (Shamsolmoali et al., 2021); (iii) simulating real-world scenarios, by making use of virtual reality (Martinez-Gonzalez et al., 2020) or other extended reality technologies, including augmented and mixed reality, to create immersive 3D virtual environments in which cameras can automatically collect photorealistic synthetic images; and (iv) by pairing existing images using methods such as cut-and-paste (Dwibedi et al., 2017) or CutMix (Yun et al., 2019) which automatically 'cut' objects of interest from training images and 'paste' them onto random backgrounds or other training images, respectively.

4.5 How to ensure that human-centric X-AI benefits all: team science, open science, open education, and embedded ethics

AI in phenomics can potentially impact on many aspects of plant science, from basic research discovery to translational research. It is crucial that these advances in technology broadly benefit society as a whole.

How then can we effectively ensure that human-centric X-AI benefits, and does not harm, individuals and communities? This can be done in several ways.

First, we suggest that pivoting toward multidisciplinary team science is necessary to tackle the most pressing scientific, societal, and ethical problems of plant digital phenomics. Over the past decade funding agencies across the USA and Europe have dedicated resources to facilitating team science. This work is evidenced by interdisciplinary and multidisciplinary team requirements in funding announcements and programs. For example, addressing the problem of bias in phenomics AI requires the integrated knowledge of socially and intellectually diverse researchers who

specialize in plant science, plant phenomics, plant pathology, data science, computer science, IS, social science, and bioethics, to name but a few.

Second, we emphasize the crucial importance of an open science system that aspires to open access not only for research outputs but also for the whole research process, and posit that all phenomics and data centers should participate in these practices. Promotion of open science and team science are synergistic goals, both of which are essential for improving our knowledge and scientific rigor.

Third, we call for mobilization of open educational resources relevant to AI in phenomics and advocate that digitized materials are offered freely and openly to educators, students, and interested learners worldwide, including developing countries, for use and reuse in teaching, learning, training, and research. Open education holds great promise to create knowledge and put it to use, promote content quality through sharing of materials for feedback and continuous improvement, and achieve competencies.

Fourth, we propose the development of socially and ethically responsible AI in phenomics by reforming curricula and embedding bioethicists into the technology development team. Ethical concerns around AI regarding handling of data, data bias, transparency, explainability, and responsibility have prompted us to consider how AI technology can be designed, implemented, deployed, and monitored post-deployment in an ethical manner. Embedding bioethicists into the AI development team can ensure that developers are practically assisted in anticipating, identifying, and addressing ethical issues through critical ethical reasoning and bioethical decision making. Universities across the USA and Europe have recently joined the effort to develop socially responsible AI by reforming curricula. For example, Harvard University initiated an 'Embedded EthiCS' curriculum that integrates ethical issues into the core computer science curriculum. We advocate that universities around the world implement similar approaches to empower students and early-career scientists to think ethically as they develop algorithms and build AI systems in their studies, in their new business ventures, and as they pursue technical work in their careers. These free and open courses should be taught by interdisciplinary teams of computer scientists, social scientists, and bioethicists.

We encourage the scientific community to embrace a growth mindset regarding team science, open science, open education, and embedded ethics which together can be harnessed to create extraordinary phenomic resources that benefit all. The rewards to these efforts come from investments of energy, time, and action.

4.6 Concluding remarks and future perspectives

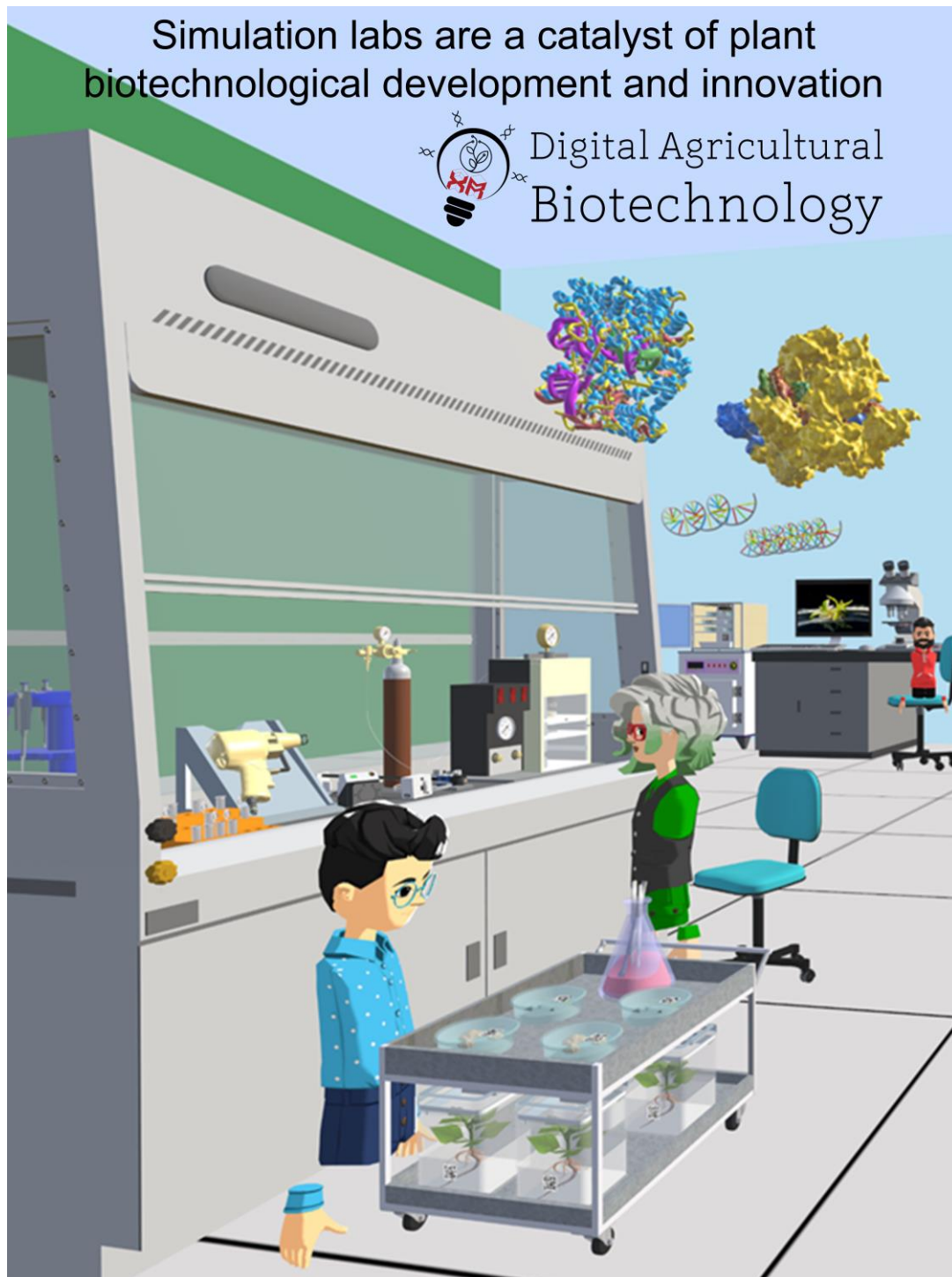
We are experiencing an unprecedented time in which the availability of vast amounts of phenomic data, combined with advances in AI, are providing the opportunity to turbocharge the data to insight journey. This opportunity is an incentive not only to design and implement effective and reliable data management strategies but also to improve the visibility, accessibility, and usability of publicly available datasets that can support research and innovation in plant digital phenomics.

Although AI has demonstrated impressive potential in phenomics, risks due to bias and lack of transparency of models should be considered. Reducing these risks entails multidisciplinary science and technology teams working together. The involvement of plant scientists, data scientists, and IS scientists during the complete lifecycle of AI analysis is integral to ensuring explainability and to identifying bias in the predictive models. Interpretable by design models can potentially be leveraged to mitigate bias and provide transparency to the decision-making process.

In the past, AI research focused on a one-way interaction – from AI to humans; today, human-centric X-AI aims to enable bidirectional interaction such that human intelligence and AI are brought together to collectively achieve superior results and continuously improve by learning from each other. Human-centric X-AI will have an extraordinary impact on phenomics in the near future, and we should do all we can to ensure that it is designed, implemented, deployed, and regulated in a way that maximizes the benefits for breeders, farmers, and consumers. In this regard, the academic and AI communities should ensure that computational phenomics, in addition to social and ethical analysis, is integrated into the plant science curriculum as a step toward this goal.

Chapter 5 Extended reality gives digital agricultural biotechnology a new dimension

Graphical Abstract



Graphical Abstract 4. Extended reality is coming of age in digital plant biotechnology.

Abstract

Facing up to the global challenges of designing climate-resilient biotech crops involves a great deal of out-of-the-box thinking. Extended reality is coming of age in digital agricultural biotechnology. Here, we seek to stimulate technological innovation by empowering future innovators, researchers, academics, and startups to think and partner creatively.

5.1 Thinking outside the box: eXtended Reality and eXplainable Artificial Intelligence ('X-RAI') park for digital agricultural biotechnology

Agriculture futures are unfolding in an era of accelerating environmental, economic, societal, and technological changes brought about by the short- and long-term effects of food and energy insecurities induced by climate change.

Digital agricultural biotechnology is increasingly gaining traction, and the arsenal of digital tools available to today's biotechnologists is impressive. It carries extraordinary potential to (i) bring both enormous long-term benefits and substantial disruption; (ii) provide individuals and communities with opportunities to acquire new skills, develop critical thinking and analytical abilities, and foster a sense of purpose; and (iii) accelerate the pace and variety of biotech discoveries in the face of climate change and global megatrends that are shaping the future world.

To leverage these opportunities, we must think outside the box and embrace creativity and innovation. As the saying goes, 'Creativity is thinking up new things. Innovation is doing new things.' Extended reality (XR) technology encompassing virtual reality (VR), augmented reality (AR), and mixed reality (MR) is a prominent example. It will play a striking role in making agricultural biotechnology more accessible and tangible to users globally (Section 5.3). It has already been incorporated into biotechnology training and education (Harfouche and Nakhle, 2020; Jones, 2018; Wismer et al., 2021) as well as data simulation and visualization (Kvinge et al., 2019; Legeth et al., 2021; Spark et al., 2020), among others. Likewise, explainable artificial intelligence (X-AI) has advanced at an impressive pace in recent years, creating new opportunities in agricultural biotechnology (Harfouche et al., 2019). Indeed, the merging of XR with X-AI is a compelling combination that promises to bring new, exciting applications that will fundamentally transform how academia and industries try to gain first-mover advantage, compete, and ultimately deliver added-value biotechnology products and services (see Figure 5.1).

Echoing this need, the X-RAI park was conceived and implemented to explore the value that virtual environments bring to research, innovation, and education in the context of other ways to teach and build capacity in digital agricultural biotechnology

(Figure 5.1). The park strives to be a hub of biotech innovation, highlighting how XR applications in this field are far-reaching. If a picture is worth a thousand words, a 3D scene is worth even more. Biotechnologists are quickly discovering the value of creating realistic simulated environments, making the X-RAI park an increasingly worthwhile endeavor.



Trends in Biotechnology

Figure 0.1. Screenshot from the X-RAI park.

The X-RAI park provides a virtual platform for learning, training, research, and innovation and serves as a biotechnology accelerator through its various immersive digital environments. Abbreviations: X-RAI, eXtended Reality and eXplainable Artificial Intelligence.

5.2 Exploring innovative XR applications in digital agricultural biotechnology

XR is beginning to fulfil its potential; the following applications offer a glimpse of what's to come.

5.2.1 Learning and training

When learning agricultural biotechnology, besides theoretical courses, learners often benefit from an experimental counterpart to help them acquire practical skills and knowhow. But running real laboratory experiments can be time-consuming, constrained by safety concerns (Bonde et al., 2014) and limited space availability (Jones, 2018). Additionally, equipment such as next-generation sequencing instruments can be prohibitively expensive and unfamiliar to students without previous training in biology, not to mention that phenomics platforms require a high budget and on-campus infrastructures exist in limited numbers globally. In comparison, virtual

worlds provide a promising solution to these deficiencies because they make environments that are not always available on an in-person tour, such as simulated biotech laboratories (Figure 5.2) and phenomics greenhouses (Figure 5.3), accessible. These realistic 3D learning environments are cost- and time-effective and allow learners to access them at different rates and times (Jones, 2018) from anywhere. Their advantages include encouraging deeper thinking about the biotech experiments before they are carried out and exploring unobservable phenomena, such as gene editing (Figure 5.2C). This can increase motivation and learning outcomes when compared with, and particularly when combined with, traditional learning (Bonde et al., 2014). That said, an emerging viewpoint in cognition suggests that, with respect to learning, the body and the mind are intrinsically linked (Reen et al., 2021). The illusion of embodiment through VR avatars can thus enhance the breadth and depth of learning experience. Embodiment also helps teachers and students develop higher levels of empathy by allowing them to walk in each other's shoes, which can, in turn, reduce human bias, such as gender, race, ethnicity, and age, in imparting knowledge that, in many instances, may be unconscious. Equally important, VR can help learners boost their imagination and creativity to develop innovative ideas. For example, in the BioInnovation Labs, users can learn and hone bioentrepreneurial skills working with LEGO sets, which have emerged as building blocks for designing and assembling scientific systems for biotechnology (Boulter et al., 2022), to build physical representations of their ideas (Figure 5.4A).



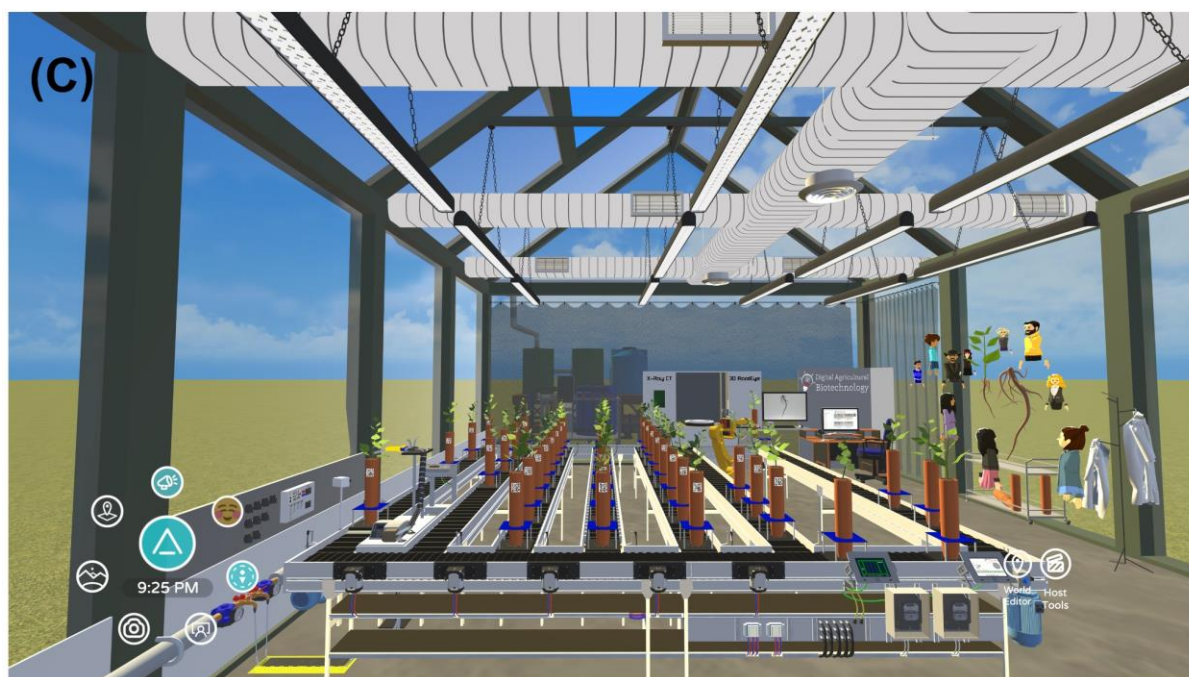


Trends in Biotechnology

Figure 0.2. Screenshots from the X-RAI biotech laboratories.

(A) 'Biotech Discovery Labs' simulate plant improvement by applying cutting-edge biotechnology to key crops such as wheat, tomato, and poplar. They consist of three stations: (B) 'PCR & Imaging Room'; (C) 'Plant Transformation Facility' targeting CRISPR-Cas9 genome editing, biolistics particle delivery system, and *Agrobacterium tumefaciens*; and (C) 'Growth Chambers & Growth Rooms'. Abbreviations: Cas9, associated protein 9; CRISPR, clustered regularly interspaced palindromic repeats; X-RAI, eXtended Reality and eXplainable Artificial Intelligence.

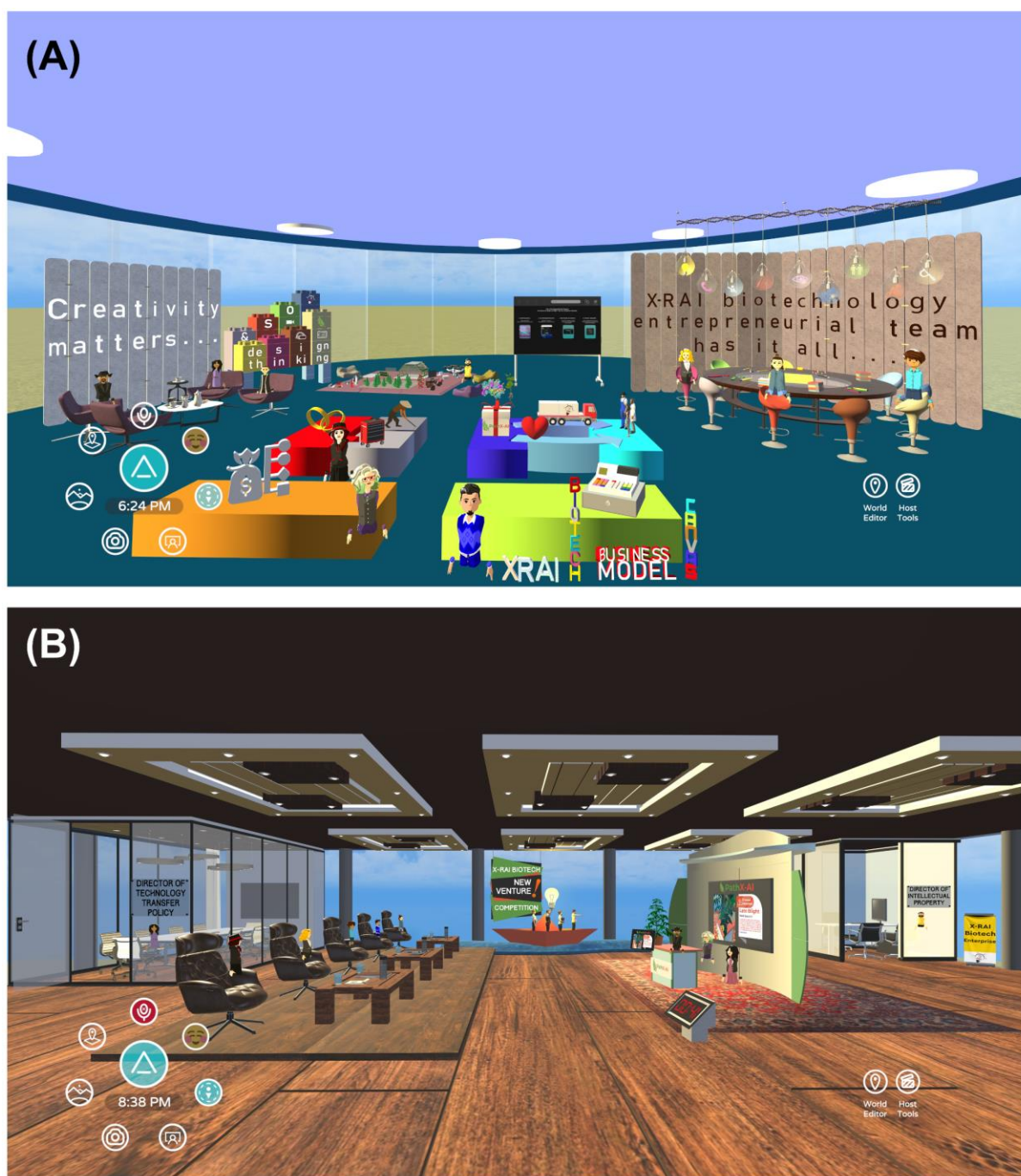




Trends in Biotechnology

Figure 0.3. Screenshots from the X-RAI phenomics greenhouses.

'Phenomics Biotech Greenhouses,' including (A) 'RootEye Rhizotron' simulates phenotyping of wheat root developmental plasticity in response to salinity stress by RGB-IR imaging; (B) 'ShootEye' simulates phenotyping of tomato shoot growth dynamics in response to drought stress using fluorescence, 3D, hyperspectral, RGB, and thermal IR imaging; and (C) 'X-ray CT 3D RootEye' simulates monitoring of poplar root system growth in response to drought stress. Abbreviations: CT, computed tomography; IR, infrared; RGB, red–green–blue; X-RAI, eXtended Reality and eXplainable Artificial Intelligence.



Trends in Biotechnology

Figure 0.4. Screenshots from the X-RAI ‘BiInnovation Labs’ and ‘Biotech Enterprise’.

(A) BiInnovation Labs feature a creativity café and a design thinking workshop to turn ideas into reality. The business model canvas helps Belbin’s biotech teams conduct structured and tangible conversations around their new bioventures. (B) Biotech Enterprise hosts new biotech venture competition. Abbreviations: X-RAI, eXtended Reality and eXplainable Artificial Intelligence.

Virtual environments can also be viable for bias-free skills training where biotech trainees interact with simulated equipment and their trainers. For example, a recent study compared real life to competitive VR training in the biopharma sector and found

that training in VR was as efficient and engaging as in real life and was more standardized, with all trainees exposed to exactly the same content (Wisner et al., 2021).

5.2.2 Omics data visualization and time-resolved monitoring of crop growth

VR serves as a potent tool for immersive 3D omics data visualization that aids in the process of knowledge discovery. For example, an interactive VR environment for single-cell experiments allowed researchers to collaboratively gain an understanding of RNA-sequencing data (Legeth et al., 2021). These advantages extend to phenomics data, including, for example, single-molecule localization microscopy (Spark et al., 2020) and hyperspectral data where they can be explored in their natural dimension (Kvinge et al., 2019).

With their ability to track head movements, VR headsets can help display 360-degree videos and images collected autonomously by omnidirectional cameras mounted on drones or on ground robots. This would increase the frequency and scale at which biotech field and greenhouse trials are inspected and would enable the continuous monitoring and comparison of crop growth over time. Furthermore, detailed 3D reconstructions of these trials can be created in VR (e.g., Figure 5.5). Thus, analyzing data on individual plants and running simulations to predict outcomes of experimental scenarios become possible.

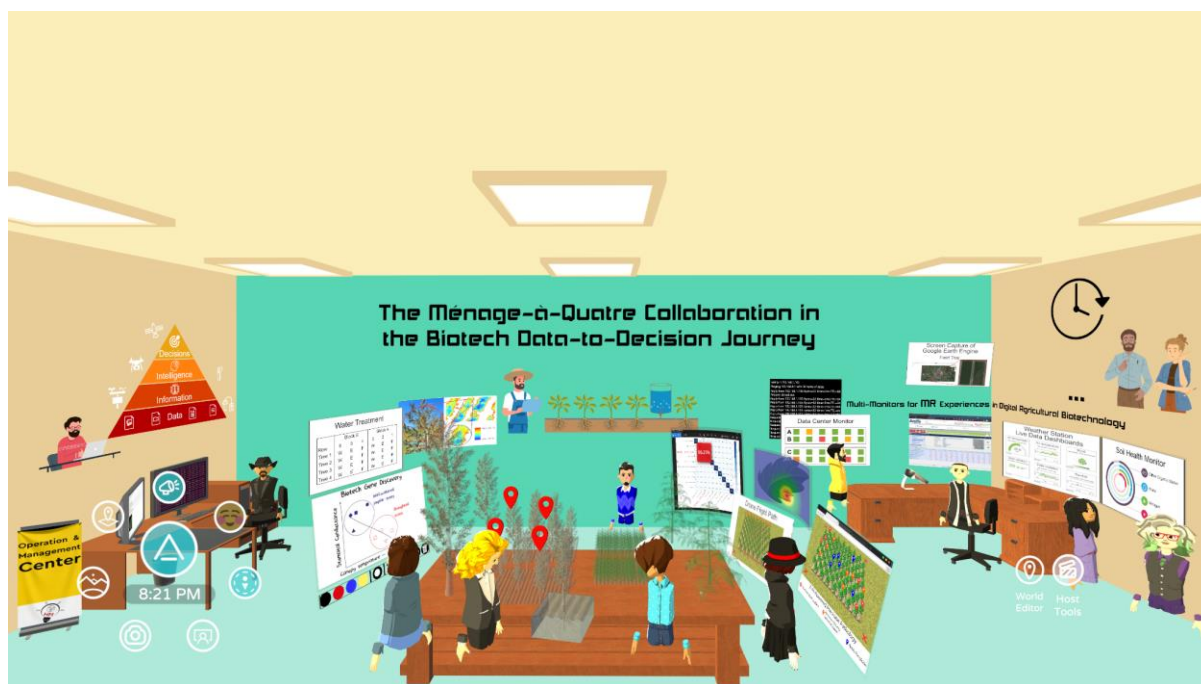


Figure 5.5. Screenshot from the X-RAI 'Operation & Management Center'.

The Operation & Management Center leverages virtual monitors to simulate control of X-RAI park. X-AI algorithms analyze data, and simulated results are displayed on the table using MR. Abbreviations: MR, mixed reality; X-AI, explainable artificial intelligence; X-RAI, eXtended Reality and eXplainable Artificial Intelligence.

5.2.3 Team communication and decision making for data-driven biotech discovery

Although the future of farming is autonomous, farmers often still attempt to detect new outbreaks in biotech crops with their naked eyes. But some diseases present similar symptoms and can be misidentified, which adds confusion to how susceptible crops are to certain diseases. AR applications (apps) can leverage X-AI to classify diseases of biotech crops by pointing a tablet's camera to a diseased leaf (see Figure 5.1 for a video version of Figure 5.1 and Figure 5.4B demonstrating the simulated PathX-AI app). Accordingly, AR can display relevant information on the emerging diseases and suggest actionable prevention and mitigation strategies. When doubt arises, team collaboration and communication can be a cornerstone of success. MR allows interaction with immersive communication where farmers can request assistance from the plant pathologist and biotechnologist, farm manager, and data scientist by sharing their field of view in real time, without requiring them to be in the same physical space (Figure 5.5). They can also benefit from holoportation, an

immersive holographic communication technology that uses depth cameras to capture high-quality 3D models of people and objects and transmit them anywhere in real time (Orts-Escalano et al., 2016). Through this, using XR, they will be able to interact with each other remotely as if they were physically present in the same location. Together, they can reach a decision on the most effective treatment plan.

With the combination of these applications, XR will pave the path towards making the ability to share, manage, use, and reuse omics data almost instantaneously a reality. XR will help experts leverage their individual strengths toward becoming more impactful as a collective. With data becoming a centerpiece of modern biotech discovery, XR will give researchers a holistic view by combining these datasets and, with the aid of X-AI, will guide breeding and gene editing programs to drive biotech discovery forward.

5.2.4 Biotech crop design

Modern breeding simulation in XR can be a powerful tool to devise and employ new approaches to design biotech crop ideotypes that can grow in niche environments. XR developments promise, for example, to increase users' knowledge of genetic architecture by offering characteristic 3D representations of nucleic acid structure, which allow performing base mutations to optimize gene editing and to widen the mutation spectra used for extensive screening for useful varieties (Figure 5.2C). Remarkably, a VR platform was developed to interactively explore genome-scale molecular networks to help identify genes associated with phenotypic traits (Pirch et al., 2021). Accordingly, a VR app can be developed to simulate gene editing and explore the resulting changes in a plant's phenotype.

5.3 Why XR? why now and in the future?

XR is a combination of interactive and immersive technologies that add to or change our experience of reality, including VR, AR, and MR. VR was coined in 1987 by Jaron Lanier (Flores-Arredondo and Assad-Kottner, 2015), who founded the first company to commercialize VR headsets. In 1992, Thomas Caudell coined AR after designing a head-mounted apparatus that displays planes' wiring instructions (Caudell and Mizell, 1992), whereas MR was coined in 1994 by Paul Milgram in connection with his investigations into virtual objects (Milgram and Kishino, 1994). Today, XR has come a

long way from a niche technology to a growing digital technology, topping the 2022 industry trends and innovations, and it is becoming increasingly ubiquitous in digital agricultural biotechnology (Figure 5.6).

As XR development continues to move forward, it is time to push the boundaries of biotechnology to be more immersive, interactive, and 3D. VR, AR, and MR are no longer new technologies that biotechnologists can wait and watch upon. Designing and implementing XR experiences require new skills and expertise relevant to workforce participation for decades to come. Now that XR devices are becoming more affordable and 3D engines and XR platforms are becoming freely available, it is time to take the plunge and start experimenting with these tools, make mistakes, learn, and develop a new competitive advantage.

In addition to the myriad of exciting biotech applications already described elsewhere in this article, we also explored whether XR has further added value for expanding digital agricultural biotechnology impact and scale. Thanks to steadfast investigations, here we include other applications that are particularly being pursued (Figure 5.6). First, ‘demonstration and exploration’ leverage immersion in VR to seemingly allow worldwide visitors to time-travel to the first-ever digital agricultural biotechnology park and transport them into a realistic alternative dimension, turning them into explorers, imaginers, creators, and innovators. It costs nothing to generate innovative ideas, and the positive impacts are priceless. Second, ‘storytelling’ is a powerful exploration-based learning and training tool in VR to help deliver meaningful and relevant biotechnology content and allow users to become part of the story instead of being passive observers. For example, users can walk into a virtual recreation of a biological phenomenon (e.g., DNA replication inside cells, functional traits in wood anatomy) and experience it in an immersive 3D environment. This can be effective in promoting empathy and can significantly improve students’ and trainees’ learning retention and team collaboration. Third, ‘networking’ in VR has the potential to break down barriers and connect an unlimited number of people, startups, and research clusters around the world in an effective way. This will allow affordable solutions for engaging authentically and forming a strategic business network that can provide a launchpad for bioentrepreneurs pursuing collaborative development in agricultural biotechnology. Fourth, ‘omics data simulation’ in VR arises from the fact that accelerating plant breeding through X-AI relies heavily on collecting real-world training

data, a process known to be costly and time-consuming; thus, this application can fill this gap by enabling endless streams of accurate synthetic data to increase diversity in datasets and to increase the robustness of X-AI models. Notably, VR simulations can help data scientists and biotechnologists account for and effectively cover edge cases (i.e., unlikely but possible scenarios such as two alleles sharing the same signature) that can occur in the real world, ensuring that synthetic data are consistent and bias-free. Last, ‘remote operation and management’ focuses on the interactions of genetics, environment, and management by harnessing MR technology and spawning multiple virtual monitors to facilitate real-time monitoring and control of fields and greenhouses while maintaining portability and cost-effectiveness. A simulation of such an application (e.g., Figure 5.5) issues near-real-time alerts for drought stress signs by displaying 3D models of droughted poplar trees and pinpointing gene-edited genotypes with higher tolerance. We are acutely aware of the long-term threats facing agriculture. We view the creation of XR technologies as critical to achieving sustainable biotechnology solutions to these challenges.

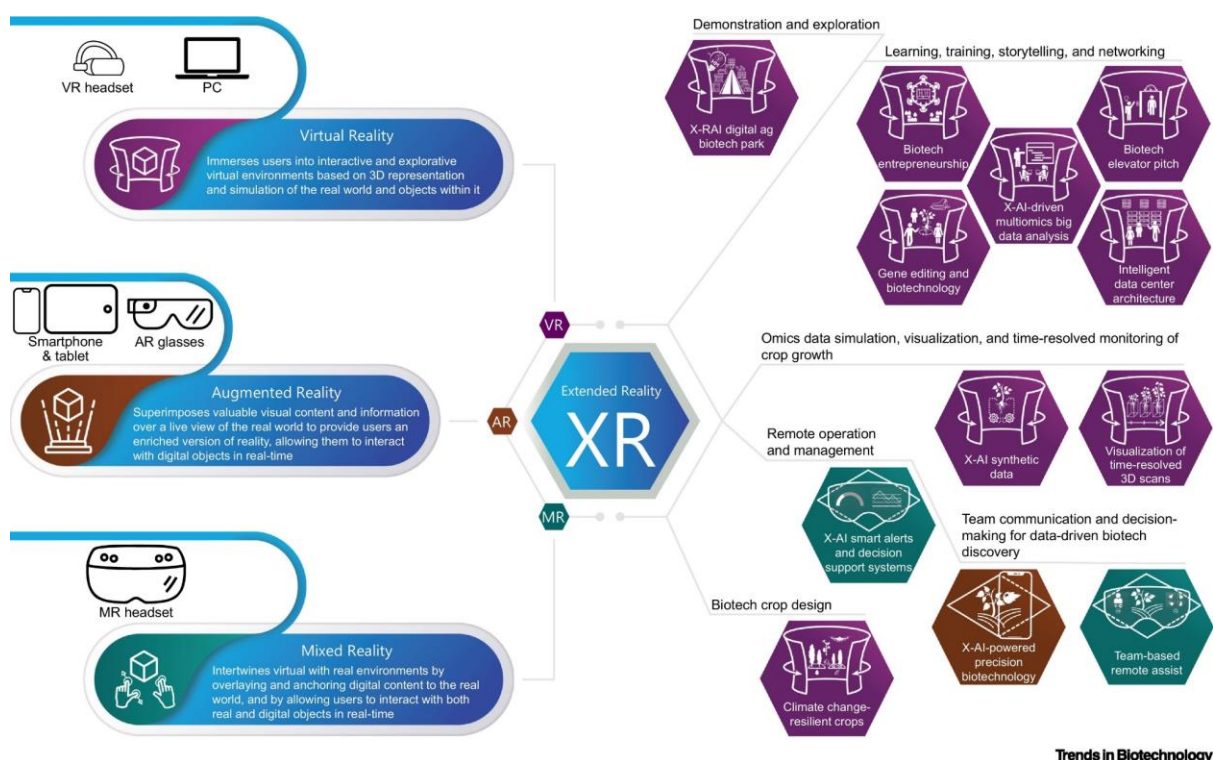
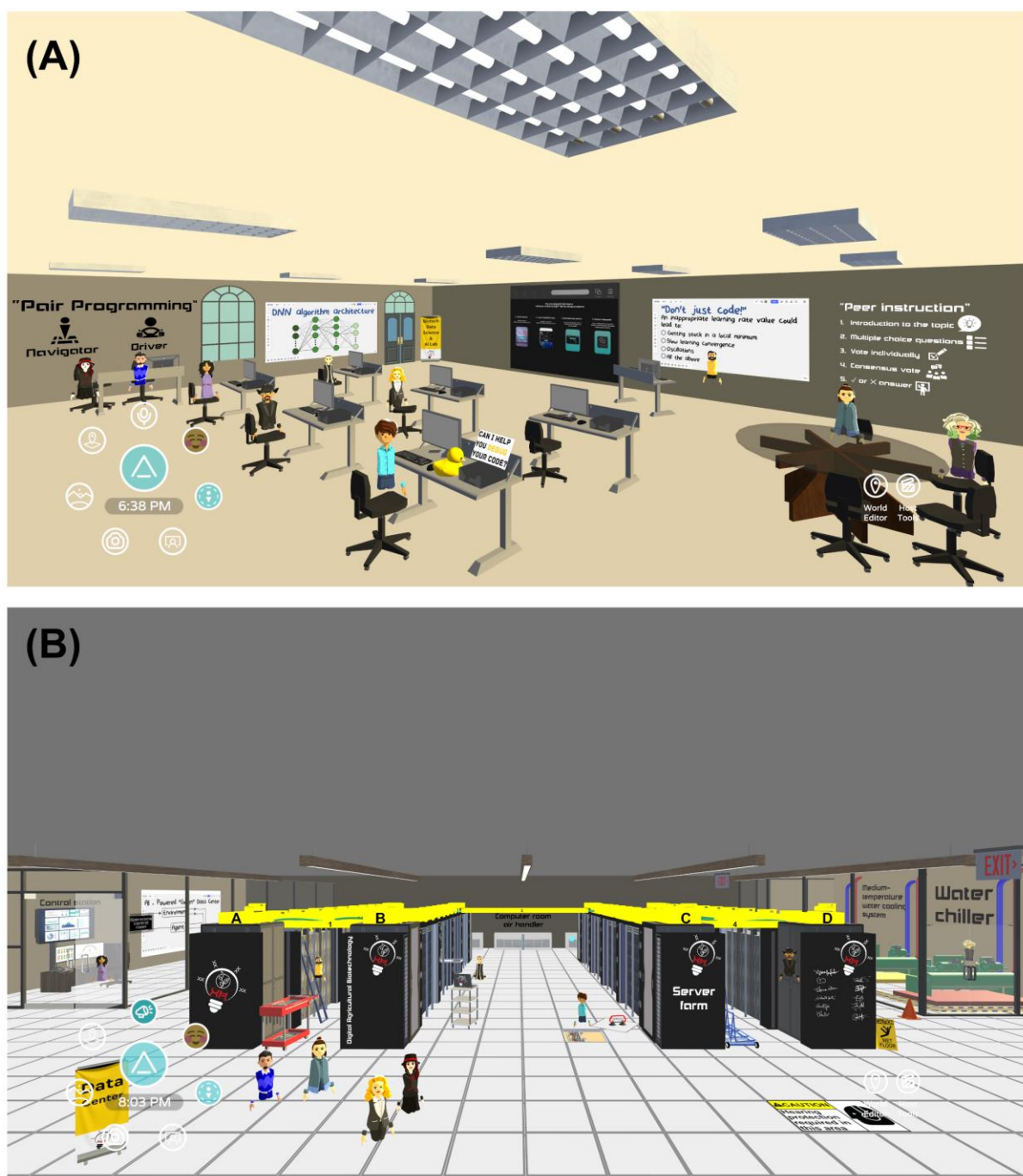


Figure 0.6. Application scenarios of XR in digital agricultural biotechnology.

XR is expected to offer a variety of digital agricultural biotechnology applications across a wide range of technological innovations. Using a VR headset (e.g., Valve Index) or a computer, VR enables the demonstration and exploration of virtual environments such as our ‘X-RAI’ park (Figure 5.1). It facilitates educating and training the next generation of agricultural biotechnologists and bioentrepreneurs, as well

as sharing experiences through storytelling and networking, as demonstrated in our 'BioInnovation Labs' (Figure 5.4A), 'X-RAI Biotech Enterprise' (Figure 5.4B), 'Biotech Discovery Labs' (Figure 5.2), 'Biotech Data Science & AI Lab' (Figure 5.7A), and 'Data Center' (Figure 5.7B). VR also allows the simulation and visualization of omics data and time-resolved monitoring of biotech crop growth, leading to designing climate-resilient ideotypes as simulated in 'Phenomics Biotech Greenhouses' (Figure 5.3) and 'Digital Agricultural Biotechnology Farm' (Figure 5.1, right). In turn, AR, using a smartphone, a tablet, or AR glasses (e.g., Google Glass), enables team communication and decision making for data-driven biotech discovery through, for example, direct-to-farmer precision pathology applications, aided by MR via team-based remote assistance applications. Using MR headsets (e.g., Microsoft HoloLens 2), MR will allow remote operation and management of the X-RAI park, including data center, greenhouses, and experimental fields; it will also strengthen decision-making processes via X-AI smart alerts and decision support systems, as showcased in our 'Operation & Management Center' (Figure 5.5). Abbreviations: AR, augmented reality; MR, mixed reality; VR, virtual reality; X-AI, explainable artificial intelligence; XR, extended reality; X-RAI, eXtended Reality and eXplainable Artificial Intelligence.



Trends in Biotechnology

Figure 0.7. Screenshots from the X-RAI 'Biotech Data Science & AI Lab' and 'Data Center'.

(A) Biotech Data Science & AI Lab conducts live coding sessions, provides one-to-one mentorship through peer instruction, and offers pair programming. (I) Data Center houses a supercomputer. Abbreviations: AI, explainable artificial intelligence; X-RAI, eXtended Reality and eXplainable Artificial Intelligence.

5.4 Concluding remarks and future perspectives

XR is more than just a sophisticated tool for driving biotechnological innovations; it is also rich with possibilities for shaping agricultural biotechnology's digital future.

Although XR can help ease learning, training, and research processes, coevolution with impactful X-AI will lead to significantly better returns on biotech innovation. What will the future hold? Hopefully, soon, XR open-source platforms and tools will become available. The shift to XR will bring forth benefits to the digital agricultural biotechnology innovation community that should be embraced moving forward. For XR to actualize its potential, concerted efforts are needed across pioneers in the field to increase direct funding for its research and innovation and to expand educational and capacity-building initiatives. An open-source approach provides the potential to encourage the greatest uptake by the research and training community, speed development, promote equitable partnerships, and fill the gaps where unequal access to tools and devices in poorer countries prevails. We are at a critical point where we can get it right if we also act in an ethical and human-centered way.



Figure360 5.1. An author presentation of Figure 5.1.

<https://ars.els-cdn.com/content/image/1-s2.0-S0167779922002396-mmc1.mp4>

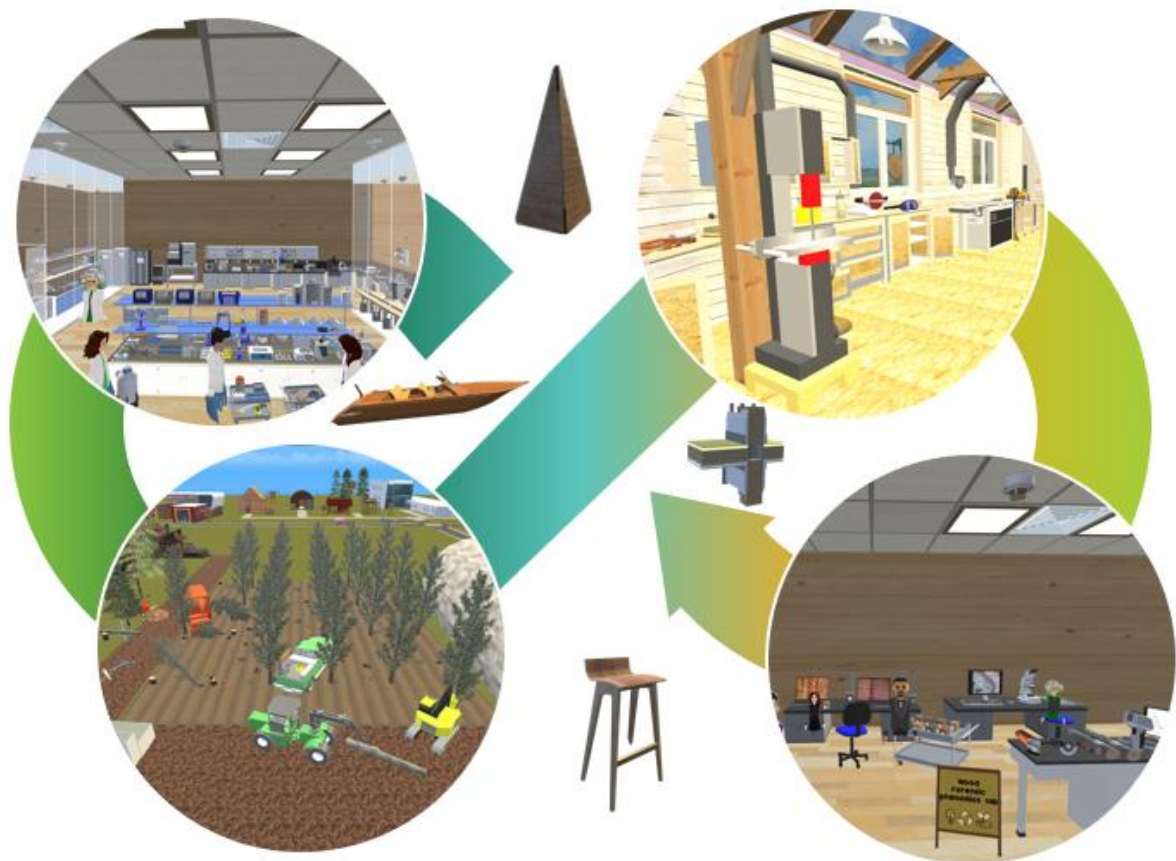
Chapter 6 Development of wood science and technology metaverse campus (WOODMeta)

Graphical Abstract

The WOOD Science and Technology Metaverse Campus

WOODMeta   

The metaverse revolution is upon us; it is here



Graphical Abstract 5. WOODMeta acts as a fulcrum of wood frontier research, supporting the development of the wood science and innovation ecosystem.

Abstract

The metaverse is rapidly expanding the potentials for disruption and innovation. Its key enabling technologies such as virtual reality (VR), augmented reality (AR), and mixed reality (MR) – collectively extended reality (XR) – have gained popularity in recent years in various domains ranging from entertainment to education due to their increasing accessibility and affordability. Harnessing these potentials to leverage wood science and technology is our primary objective. Herein we showcase the wood science and technology metaverse campus (hereafter WOODMeta) that we conceived and created for advancing education, training, research, and innovation. The WOODMeta allows educators and researchers to bring the forest and wood industry into the classroom, laboratory, workshop, and field in a fully interactive, immersive, and risk- and bias-free environment with diversity and inclusion from the start, while promoting innovation, compatibility, and competitiveness throughout the value chain with environmental and developmental sustainability. A key component in our VR designs is the opportunity for users to customize their avatars with an extended range of possibilities and useful features that can be used to represent themselves. The XR revolution is upon us; it is here.

6.1 Introduction

‘The sky and the ground are black, like a computer screen that hasn’t had anything drawn into it yet; it is always nighttime in the metaverse, and the street is always garish and brilliant, like Las Vegas freed from constraints of physics and finance’. With this sentence, Neal Stephenson described what does the metaverse look like when he first coined the term at the beginning of his 1992 novel, *Snow Crash* (Stephenson, 1992). The metaverse can be defined as a persistent, immersive, and interactive simulated world that can be experienced simultaneously by plurality of users where they share a strong sense of mutual presence and can interact with one another, with the virtual world and its environments (i.e., virtual spaces contained in a virtual world), and with the objects within them (Pyun et al., 2022; Rosenberg, 2022; Sun et al., 2022). It is like standing in your bedroom but virtually exploring a forest and searching for felled tree trunks alongside a friend living on the other side of the world; this can be achieved by accessing the metaverse using extended reality (XR) technologies. XR encompasses interactive and immersive technologies, including virtual reality (VR), augmented reality (AR), and mixed reality (MR), that provide the possibility to add to or change our experience of reality (Nakhle and Harfouche, 2023). Wood science and technology are no exception; the metaverse holds novel dimensions to integrate into this fruitful medium.

Through XR, the metaverse is poised to make the jump from forests, sawmills, woodworking workshops (hereafter woodshops), and wood science laboratories, to name a few, to our classrooms, offices, and living rooms. Ultimately, it will transform how we teach, learn, train, get trained in, research, and innovate wood science and technology. As a step in this direction, some applications (apps) have been developed to infuse VR technology into wood experiential learning, woodworking machines and equipment training, and wood imaging data visualization. For example, ForestVR is an app that showcases various forest environments and processing mills to help learners explore the wood processing journey (Australian Forest Education Alliance, 2022). However, virtual environments in ForestVR are based on 360-degree images and videos that are not interactive, which may induce passive viewing; thus, they are not suitable for real-time interaction. Yet, in another example, the app InductVR allows learners to interact with woodworking machinery and guides them through their functionality (Insilico Ltd, 2018). Similarly, vrkshop is a game that challenges learners

to build wood products with hand tools, putting their woodworking skills to the test (Wooden robot, 2021). However, limitations on the experiences that these simulations provide still need to be resolved. For example, each environment in InductVR comprises one machine, which limits users' exploration ability; in addition, the machines are limited to band saws, drill presses, and routers. vrkshop is limited to a single user, and thus, does not allow for interaction between trainers and trainees. Another example comes from a VR system that was developed to interactively visualize X-ray micro-computed tomography (micro-CT) data in VR to increase learners' understanding of complex wood anatomy features (Feng et al., 2020). However, this VR system is limited in scope and it does not support multi-users interaction.

Amidst the incumbent constraints, we conceived, designed, and built the wood science and technology metaverse campus, hereafter WOODMeta, as a solution to the existing limitations (Figure 6.1). We built it from the ground up with diversity and inclusion in mind. The WOODMeta brings learners, trainees, researchers, and woodworkers together to immersively and interactively embark on educational safaris that are as close as virtually possible to the real experience, touring sawmills, wood processing facilities, and woodshops. These woodshops allow users to explore and unravel the functional mechanisms behind various woodworking machines and to learn through the reflection of doing. The WOODMeta's learning centers and collaborative spaces enable users to brainstorm, share ideas, and co-create with ease in the metaverse. Further, the biotechnology and forensic laboratories provide a path to collaborative wood phenomic and genomic data visualization and exploration. Although the WOODMeta exists in a virtual world, it realistically respects human values and promotes developmental and environmental sustainability.



Figure 6.1. ‘WOODMeta’ in the spotlight: into the wood science and technology metaverse campus.

Screenshot of the WOODMeta conceived, designed, and built in this article.

6.2 The WOODMeta – a consolidating environment for broad technological, educational, and entrepreneurial experience equal to reality

With the world facing a race to catch up on declining student’s learning outcomes, and with an economic growth that has been lackluster for more than a decade now, the metaverse has the potential to be a great technological, educational, and entrepreneurial equalizer. The WOODMeta is not just another XR app; it aims to bridge the real world with the digital one through a metaverse of opportunities that empowers everyone in a geographically agnostic, bias-free world, with diversity and inclusion at its core. It is a once in a generation opportunity to seek discoveries without boundaries, to create future wood science leaders, to collaborate creatively to advance wood science and innovation and to engage and inspire the multi- and inter-disciplinary atmosphere that maximizes openness and permeability to ideas and talents from different disciplines. It will act as a fulcrum of wood frontier research, catalyzing new and dynamic interactions between relevant stakeholders and across research disciplines, and supporting the development of the wood science and innovation ecosystem. To the best of our knowledge, the WOODMeta is the first metaverse established for wood science and technology; the following sections serve to exemplify the vast scope of possibilities that the WOODMeta has to offer.

6.2.1 *Creating the WOODMeta experience: how did we do it!*

Our vision from the start was to build an entirely new design of human coexperience – nothing less than the realization of the next phase of human interaction and a newfangled creativity and innovation ecosystem that anyone can be part of. We imagined a metaverse campus, a network of virtual environments where people from anywhere in the world could intuitively interact, just as they would in person. To put our ideas into action, we started coding. Soon we had a prototype, which we called WOODMeta.

The WOODMeta conglomerates various tech elements, and it was developed entirely using the following free software: ‘Blender’, an open-source three-dimensional (3D) creation suite; ‘Unity’, a cross-platform game engine; and ‘AltspaceVR’, a metaverse platform for creating MR experiences. The 3D models it comprises, were created in Blender and then imported into Unity, where they were organized into world-building kits (i.e., small databases of 3D objects). The WOODMeta virtual world and its 11 environments were created in Unity and were uploaded, along with the world-building kits, to AltspaceVR using the ‘AltspaceVR Unity uploader’. The AltspaceVR platform provides a ‘world editor’ that contains the tools required to spawn, manipulate, and place 3D objects from the uploaded kits. As such, the world editor was used to furnish the virtual world and its environments with 3D models.

6.2.2 *Offering an inductive near real world environment*

6.2.2.1 *A personalized experience with expressive, diverse, and inclusive avatars representative of the people behind them*

Personal avatars are a central part of our vision of what the future of virtual communication might be. The WOODMeta strives to create a metaverse that values diversity, promotes inclusion, and establishes a sense of belonging for each user right from the start. It represents users by customizable avatars with distinctive art styles that balance between virtuality and photorealism (Figure 6.2). These avatars achieve a high level of expressiveness to create immersive interactions through facial, head, and upper body movements. They also feature lips-syncing technology and maintain eye contact with other avatars, allowing users to take on a virtual character, which boosts their confidence and facilitates social interactions by creating the sensation of

being face-to-face with others (Harfouche and Nakhle, 2020). In addition, through the AltspaceVR avatar configurator, users can express their identity by configuring skin color, face shape, hairstyle, and clothing. This can help users develop higher levels of empathy by allowing them to walk in each other's shoes, richly understanding and appreciating one another, and to see the world through each other's eyes, sympathetically appreciating a person's specific thoughts and feelings, which can in turn reduce human bias, such as gender, race, ethnicity, and age (Nakhle and Harfouche, 2023).

In the midst of such activity, bespoke avatars were successfully designed using a custom-developed MR extension (MRE) through the AltspaceVR software development kit (SDK), which allows users with programming experience to extend the functionality of AltspaceVR. We leveraged the SDK to blur the line between the real and the virtual world by coding an MRE that makes the WOODMeta more interactive and its avatars more diverse and inclusive, so users can dynamically add or remove clothing and accessories to their avatars. The new MRE, named WOODMeta AVATARS, was developed using the Node.js application programming interface (API) and the TypeScript programming language. It was wrapped into a Docker image to ensure its reliable deployment on a cloud server. WOODMeta AVATARS offers disability-centric accessories to our novel, inclusive avatars to represent users with hearing aids, cochlear implants, oxygen tubes, soft helmets, and wheelchairs (Figure 6.2a), much more than was traditionally possible. This realistic presence is a key to creating a connected experience and will help the WOODMeta ensure equal access to education and training for persons with disabilities. Through WOODMeta AVATARS, they can practice skills and interact with others in a safe environment without feeling overwhelmed or anxious. It can also help those with mobility issues to improving their quality of life. These benefits are compatible with the UN's Sustainable Development Goals (SDGs) agenda, existing goal (SDG 4) and targets (SDG 4.5 and 4.a). Furthermore, the app promotes diversity by offering safety gear, such as goggles, gloves, and hearing protection for the avatars representing, for example, woodworkers (Figure 6.2b,c). We also allow students to outfit their avatars with custom-designed clothing, such as Belbin team-specific shirts honoring the 'Wood Phenomics', the 'Wood Biotechnology and Bioeconomy', the 'Wood Art, Architecture,



and Design', and the 'Wood Ethics' teams (Figure 6.2d). These equalizing benefits would be followed by creativity and imagination enhancement.





Figure 6.2. Expressive avatars that make the WOODMeta a bias-free world with diversity and inclusion at its core.

(a) Avatars with disability-centric accessories to represent users with (from left to right) hearing aids, cochlear implants, oxygen tubes, soft helmets, and wheelchairs, respectively. (b) and (c) Avatars representing woodworkers by offering safety gear including goggles, gloves, and hearing protection. (d) Avatars wearing team-specific shirts to represent the four Belbin teams (from left to right): (i) the 'Wood Phenomics' team, (ii) the 'Wood Biotechnology and Bioeconomy' team, (iii) the 'Wood Art, Architecture, and Design' team, and (iv) the 'Wood Ethics' team. Our avatars will constantly become more sophisticated, letting people be whomever they want to be in the WOODMeta campus.

6.2.2.2 *Virtual environments that look, sound, and feel real*

The time has come to reimagine how people come together to connect, socialize, network, learn, research, and innovate in ways that open up opportunities that wouldn't otherwise be available, and WOODMeta can be the blueprint. Virtual environments in the WOODMeta are conceived and designed to look like real-world environments, where communications, interactions, and explorations, feel like happening in a real place, and where conversing with others feels natural and effortless. Each environment is purpose-built to create the right atmosphere and conditions to get users in the best frame of mind for what they can do at a glance and for what they are trying to accomplish. The environments blend realistic 3D models and visuals with spatial audio technology to increase immersion and reinforce the sense of being physically present alongside other users, creating a foundation for effective and efficient teaching, learning, training, capacity building, researching, and innovating in the metaverse. The WOODMeta environments are non-static – they will change all the time, and the WOODMeta experience is persistent – it's a living experience that exists consistently for everyone and in real-time.

6.2.3 *Enabling the age of immersive education and experiential learning*

6.2.3.1 *The educational safari prototype*

Education in the metaverse is certainly garnering interest, raising hope of a more inclusive and equitable quality education for all in the future. The metaverse can truly democratize education, by allowing people all over the world to connect in a cost-effective and flexible manner. The WOODMeta embarks participants on an immersive and interactive 3D educational safari as an initial phase of their experiential learning. This provides learners with ample exciting opportunities to bridge the theory-practice gap, to increase their knowledge, to clarify values, and to develop skills, competencies, and attitudes in a practical context by authentically engaging them in hands-on experiences. Remarkably, virtual safaris would resolve the challenge of deploying learners to real-life woodworking sites, which are not always available for access, and would augment the benefits associated with experiential learning in general. The VR-based educational safari is designed to allow participants to experientially learn and

transform by exploring relevant issues, transferring knowledge, and gaining firsthand practical experience from the simulated ecosystem inside the WOODMeta touching on the wood value chain, starting from how trees are safely and sustainably felled to how wood arrives to woodshops. As first step, 3D models that mimic naturally broken tree branches and fallen trunks (e.g., due to extreme high temperatures, low rainfall levels, fires, diseases, or compromised root systems) are collected from the sustainably managed natural forest, whereas trees are harvested from the responsibly developed plantation forest (Figure 6.1). WOODMeta's educational safari submerges students in virtual worlds made for learning how wood logs are sorted according to diameter, length, and end-use before sawmilling, and how they are bucked, debarked, and cut into boards in the 'Equipment Shelter', using 3D representations of equipment, such as circular saws and bandsaws. The safari continues with a simulation of how wood undergoes a drying process, which is essential to protect it from microorganisms, increase its strength, and reduce its weight and transportation costs. This is accomplished in yards in open air or through kiln drying (Figure 6.1). Afterwards, wood can be fireproofed and protected from destructive agents such as wood-decaying insects, molds, and fungi by using a simulated preservative treatment in an autoclave, where chemicals are impregnated into the wood using vacuum and pressure (Figure 6.1). Wood is then stored in the 'Wood Shelter' and is finally ready to be transported and used in downstream processing and value adding (Figure 6.1).

Woodshops are extremely important for wood science and technology training as they equip learners and trainees with hands-on experience in using woodworking equipment and machinery. Simulated woodshops can complement physical ones, especially that the latter are expensive, time-consuming, difficult to scale, and might raise safety concerns. Indeed, in the WOODMeta woodshops, users have the opportunity to learn and get trained on simulated materials and machinery until they are well-versed in woodworking foundational techniques and safety practices prior to using the real ones (Figure 6.3a–e); the materials and machinery include, among others, band saws, table saws, drill presses, sanding machines, lathes, jointers, planers, spindle molders, computer numerical control routers, and dust collectors, in addition to various hand tools. Participants can also interact with the woodshop products and learn the future potential use of the processed wood, for example, in 3D



designs of building construction, intelligent furnitures, architectural accents and home goods, among others (Figure 6.3f).

At WOODMeta, we consider an educational safari to be a safari experience with a positive impact on the environment and the users since it enhances learning without loss of natural resources nor jeopardy of safety measures.





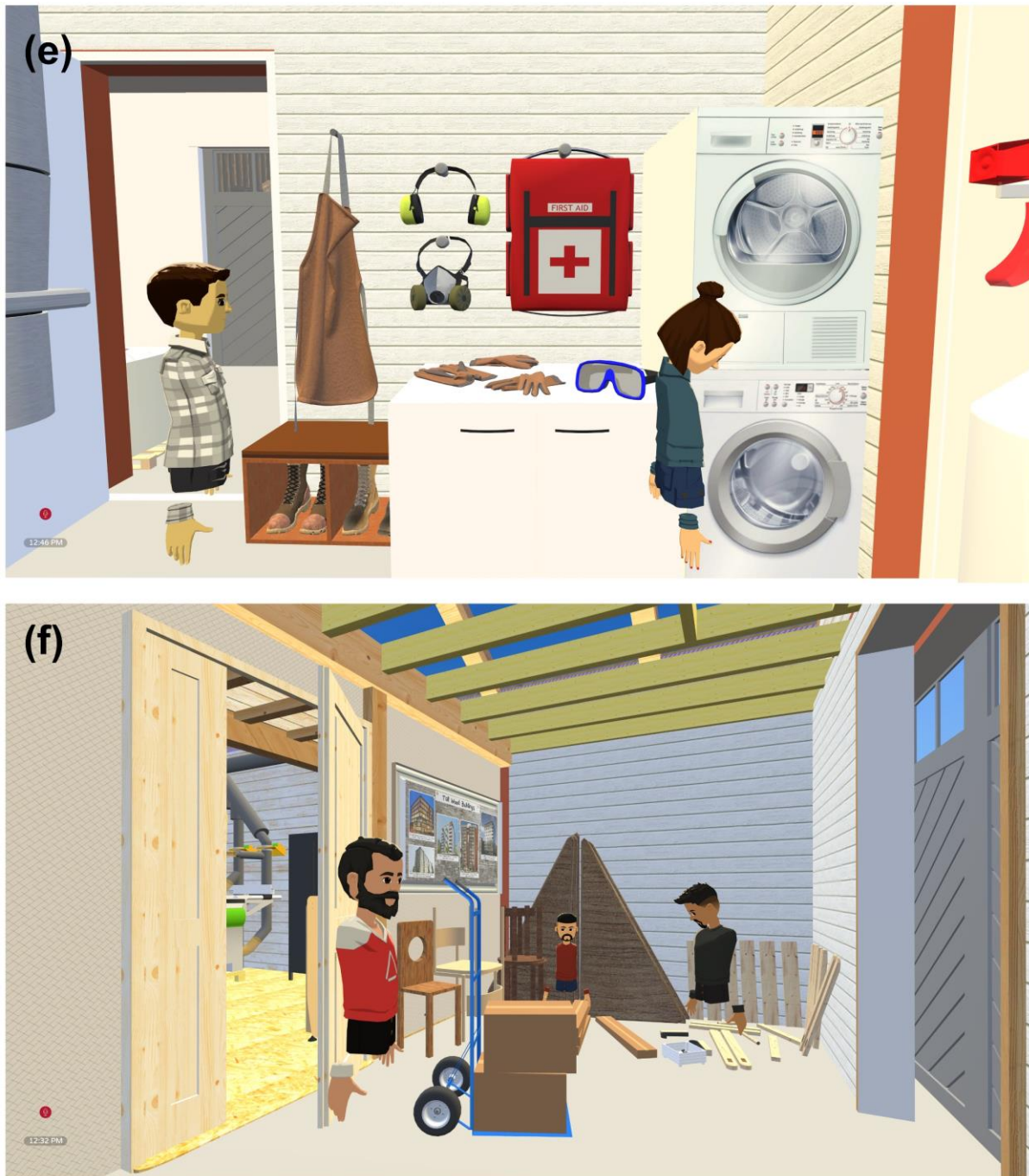


Figure 6.3. Immersive, interactive, and realistic 3D virtual workshops to pick up the ropes of woodworking.

(a) Front view of the 'Woodworking Workshop'. (b) Back view of the Woodworking Workshop. (c) Front view of the 'Wood Art, Architecture, and Design Workshop'. (d) Back view of the Wood Art, Architecture, and Design Workshop. (e) Woodworking safety gear in the Wood Art, Architecture, and Design Workshop. (f) Staging area of the Wood Art, Architecture, and Design Workshop.

In the following phase, students, wood scientists, wood technologists, wood biotechnologists, data scientists, and woodworkers can teleport to the 'WOODMeta Case Classroom' where they learn by doing on multiple levels and gain skills in critical

thinking, communication, and group dynamics (Figure 6.4a). The VR case classroom has a horseshoe-shaped design that allows teachers to move easily through the classroom, to promote dynamic interaction among learners, and to provide a class of seated students an unobstructed line of sight to the digital interactive whiteboards (Figure 6.4). These whiteboards can also be used to display educational multimedia content (Figure 6.4a). By virtue of being virtual, the classroom has no geographical boundaries and thus, can be filled with learners from diverse wood industries, countries, and backgrounds, all in search of solutions to real-world wood science and technology problems and challenges through cooperation with colleagues, and through self-observation and reflection. For example, when discussing the problem of illegal logging and illegal wood trade, participants can have different opinions, different considerations, and each person's perspective might provide a different angle to the problem.

Afterwards, participants move to 'School 42' which focuses on project-based learning and teamwork (Figure 6.4b). School 42 is inspired by the nonprofit computer science school created by Xavier Niel. 42 refers to the novel series 'The Hitchhiker's Guide to the Galaxy' where that number is the answer given by a supercomputer to the ultimate question of life, the Universe, and everything (Association 42, 2021). In the WOODMeta, simulated School 42 can facilitate group discussions in VR by featuring round tables and leveraging spatial audio technology to mimic the pitch, volume, reverberation level, and other audio cues that are expected in real-world experience to reduce distraction and help students, wood scientists, wood technologists, wood biotechnologists, data scientists, and woodworkers stay focused. In the example of illegal logging, users breakout into four teams; team 1, Wood Phenomics, works on how to effectively leverage digital technologies such as explainable artificial intelligence (X-AI), next-generation blockchain, digital forest twins and digital wood twins, and XR in wood identification and geographic origin investigations to turn the tide on illegal logging (Harfouche and Nakhle, 2023). Team 2, Wood Biotechnology and Bioeconomy, framestorms and brainstorm with ease why and how does the future depend on development of wood bio-products, and the effect of illegal logging on its sustainability. Team 3, Wood Art, Architecture, and Design, might discuss the market demands, designer needs, and the end-user preferences. Team 3 would also attempt to implement ideas into prototypes and to promote

sustainability and responsibility by helping woodworkers, designers, and consumers make informed purchasing choices about wood and wood products, ensuring the legality of their sources. Last, team 4, Wood Ethics, reasons critically about the relevant ethical questions raised by this case and how to leverage ethical decision-making models to promote responsible use of the above-mentioned technologies in the fight against illegal logging (Figure 6.4b). Impactfully, users from anywhere in the world can experience these virtual environments at their own pace, or synchronously alongside teachers and trainers, enabling them to upskill and gain new competencies.



Figure 0.4. ‘Wood Learning Center’ to assimilate wood science and technology.

(a) ‘WOODMeta Case Classroom’. (b) ‘School 42’.

6.2.3.2 *The network and the team imperative*

Inspired by a variety of 3D designs of wood prototypes, the ‘Wood Collaboration Space’ is a virtual space where teams and networks between individuals and teams within wood science and technology organizations from anywhere in the world can collaborate and work together on new ideas in a fun-filled manner. More specifically, they do so as efficiently as possible, enhancing their overall productivity. The space is a cozy VR environment with a distinct atmosphere that breaks the four-wall mold and geographical barriers and embraces a more flexible and open way of working and networking, from brainstorming to making new connections and socializing that previously were only possible in the physical world (Figure 6.5). It features 3D models of interior designs with plenty of vibrant colors to encourage social interactions and inspire creative, innovative, and collaborative work, capitalizing on the power of diversity and inclusivity, and this potent combo can further bring about a decrease in conflict and stress. While this kind of prototypes spark rich conversations, the VR space also includes a bar setting that stimulates confabs and allows users to come together in a more relaxed, safe, and natural way rather than working in isolation to quickly turn informal interactions amongst them into co-creative collaboration. In addition, the space contains a digital whiteboard that can enable users to work together more efficiently and effectively by sharing educational multimedia content. In the WOODMeta, people can have rich interactions that help them understand each other better and work toward the same goal faster. They can see each other, talk to each other, and collaborate in ways that wouldn't be possible in the physical world. The stimuli in the Wood Collaboration Space make it effective environment to work; the combination of ambient noise, informal crowds, and visual variety as well as the social-facilitation effect can give participants just the right amount of distraction to help them be the most creative. If you want a collaboration space brimming with creativity and effective teamwork, you'll need WOODMeta.



Figure 0.5. 'Wood Collaboration Space' to stimulate confabs, networking, and collaborations.

6.2.3.3 *The entrepreneurship learning model*

The metaverse might still be in the early stages of development, but it will undoubtedly influence every corner of the education, research, and innovation. In developing the WOODMeta, we decided to make it useful to users, and fruitful to entrepreneurs and start-ups, as the best way to spark creativity, to fuel ideation, and to inspire entrepreneurship across the wood industry. The 'Innovation for Impact Labs' feature a creativity and design thinking café for users in VR to meet and work together informally in a hackathon fashion (Figure 6.6a). In these labs, users can brainstorm their creative ideas in a judgment-free environment following the inventure cycle which defines the pathway from inspiration to implementation (Seelig, 2015) (Figure 6.6a). The digital interactive whiteboards can be used to display educational multimedia content and colored sticky notes. Participants are formed into teams based on their corresponding Belbin team roles. They leverage the 3D model of the business model canvas to conduct structured and tangible conversations about their innovative ideas in a way that may be more convenient and not even possible in the physical world (Figure 6.6a). Teams then learn strategic communication skills following the golden circle of communication (Sinek, 2011). These simulated laboratories have the effect of raising participants' energy levels, partly because they're mobile during discussions, and partly because they can richly interact with a selection of 3D wood processed

materials coming from the WOODMeta woodshops to boost their creativity and allow them to craft realistic woodwork of their choice. Through gamification of wood building blocks, they can prototype and build their own 3D models ranging from intelligent furniture such as music tables and music chairs, to the marvels of prefabricated construction such as ‘Chris Moller Architecture + Urbanism’s shell structure’ design (Figure 6.6a). This wood architecture uses thin triangular billets of pine laminated veneered lumber reinforced and held together at the hinge to provide the ductile strength necessary to resist earthquakes with minimal material. The novel 3D model, a digital replica of the new generation of strong safe lean architecture, is used in the ‘Wood Entrepreneurship Center’ (Figure 6.6b). It was designed specifically to be more accessible by more people across the globe and to raise awareness of the earthquake resistant wood construction.

The innovative ideas are then pitched live in VR to a panel of experienced judges in a new venture competition hosted in the ‘Wood Entrepreneurship Center’, providing teams with a unique opportunity to put the teamwork, communication, and entrepreneurial skills and competencies gained and the attitudes developed into practice (Figure 6.6b). Teams demonstrate how they are helping to solve customers’ problems and to satisfy customer needs by incorporating 3D prototypes into their elevator pitch. The center cultivates a high-risk, high-gain value-creation mindset in the next-generation of woodpreneurs, aiming to bring feasible, viable, desirable, and ethically responsible innovations to market.



Figure 6.6. Wood technologists, wood scientists, designers, and trainers work alongside the next generation of woodpreneurs to hone their teamwork, communication, and entrepreneurial skills, competencies, and attitudes.

(a) 'Innovation for Impact Labs'. (b) 'Wood Entrepreneurship Center'.

6.2.4 Promoting a greater level of enthusiasm for a human-centric research approach to visualize, analyze, and interpret wood data

The WOODMeta Biotechnology and Forensic Labs harbor the ‘Wood Forensic DNA Lab’ and the ‘Wood Forensic Phenomics Lab’. Simulation laboratories is an essential step in the training of current and future wood science and technology experts, and a catalyst for transforming them into practicing scientists and technologists. Realistic situations are created using equipment similar to what students and trainees will encounter in biotechnology laboratories and other wood science settings. These VR simulated laboratories enable remote and yet seamless learning irrespective of time and location. They bring a wealth of exciting and engaging opportunities; they can serve as a potent VR tool to visualize wood phenomic and genomic data in their natural dimension rather than 2D representations, thus enhancing visualization capabilities and interaction with, for example, microscopic images and anatomical datasets, irrespective of users’ location. The VR laboratories underline how data visualization in VR can be a key part of the discovery process that bridges between the quantitative content of the data and human intuition (Donalek et al., 2014). In this context, they can bring together a team of wood scientists, wood technologists, wood biotechnologists, data scientists, and woodworkers to a realistic, immersive, and interactive VR environment where they can simultaneously and collaboratively visualize and interact with big data (Figure 6.7) that are otherwise logistically challenging and expensive. This can deeply improve their knowledge of, for example, the underlying mechanisms of how wood anatomical traits respond to a changing environment (Nakhle and Harfouche, 2023) and how these traits enable wood identification to fight against illegal wood logging and illegal wood trade (Harfouche and Nakhle, 2023). This kind of advanced training in dedicated VR laboratories can add significant value by building empathy, decision-making and other social skills essential to wood science and technology practitioners and leadership roles.

Some topics cannot be grasped completely solely on an experimental level. Besides visual cues, simulation could facilitate learning in the VR environment. The Wood Forensic DNA Lab simulates DNA extraction and genomic sequencing from wood samples as well as DNA profiling of unprocessed and processed wood materials. With the aid of X-AI algorithms that embed human knowledge into their models, keeping humans in the loop (Harfouche et al., 2023, 2019), DNA profiling serves as an ideal tool to identify wood species and to trace the geographic origin of wood and wood products (Figure 6.7a). The Wood Forensic Phenomics Lab simulates wood

identification by visualizing and analyzing phenomic data coming from, for example, dissecting microscopes, X-ray fluorescence spectrometers, or XyloTrons (Figure 6.7b). Users in VR can grab, rotate, and scale output images to focus on a specific region of interest with greater ease of navigation, diagnostic confidence, and image quality.

Both simulated laboratories can reduce infrastructure costs by eliminating the need for expensive equipment and maintenance costs. They can help protect the environment by minimizing chemical waste and pollution and maximizing the purity of our natural resources for future generations. They are easy to scale and thus, users do not need to wait for their turns to explore the equipment, to experiment freely, and to develop practical skills. They provide users with the opportunity to enhance their practical understanding in a safe environment. For example, they can be leveraged to help users understand how to keep track of wood origins, how to trace wood along its production and trade cycle, and how to help all stakeholders of the wood supply chain, including consumers, validate the legitimacy of the wood and the wood products and the legality of their sources to make informed purchasing choices and ultimately help to put an end to illegal logging and illegal wood trade (Harfouche and Nakhle, 2023). The WOODMeta laboratories can intrigue students and trainees curiosity, increase their confidence level and further pique their interest in scientific endeavors, not to mention the advantages of fast feedback.

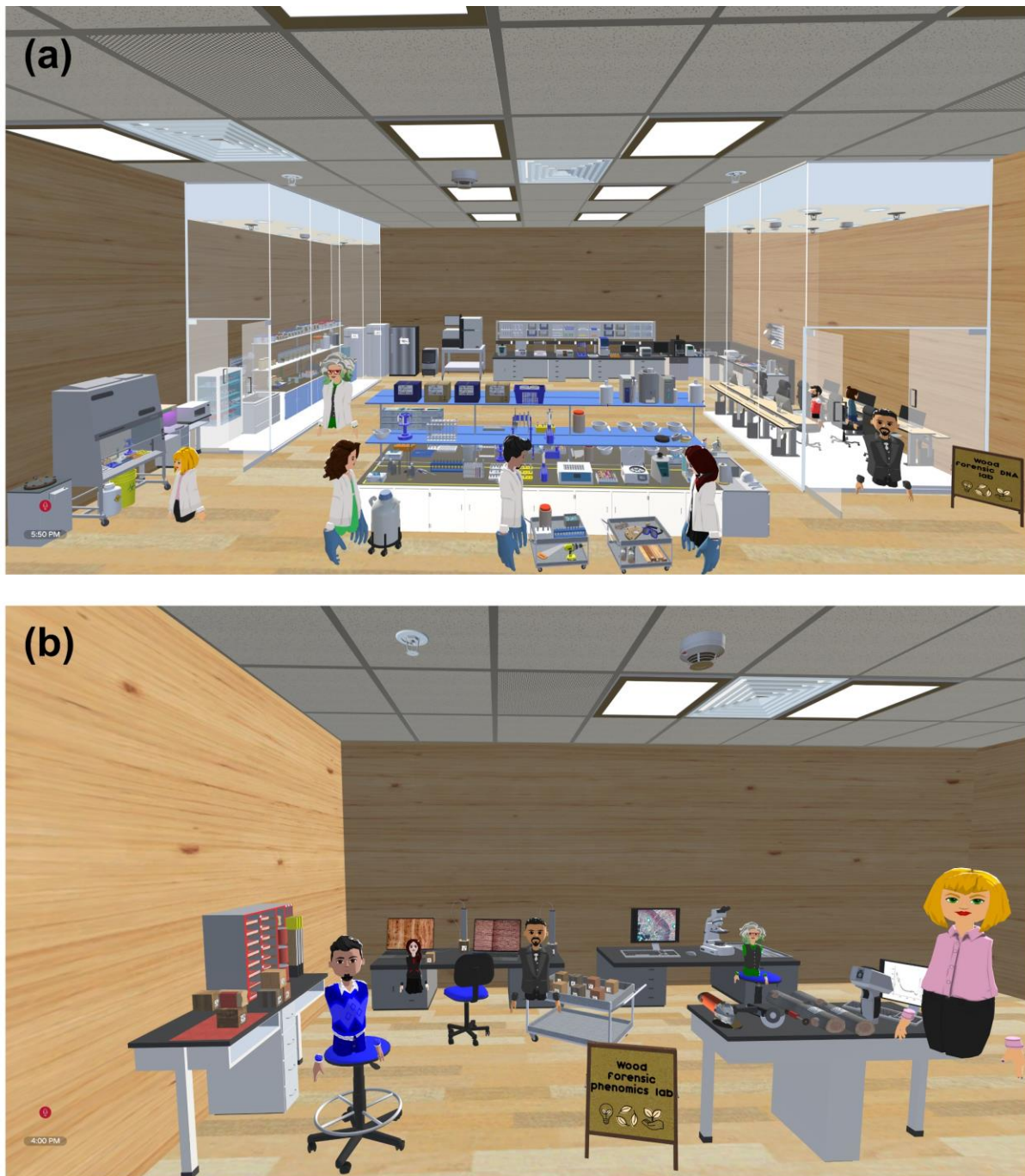


Figure 0.7. Wood technologists, wood scientists, designers, and trainers work alongside the next generation of woodpreneurs to hone their teamwork, communication, and entrepreneurial skills, competencies, and attitudes.

(a) 'Innovation for Impact Labs'. (b) 'Wood Entrepreneurship Center'.

As an effective catalyst to creating innovative education methods, expanding student outreach and enhancing understanding, the 'AI, Data Science, and Blockchain Labs', offer users various tools, technologies, and services for the development of intelligent systems that can be leveraged in the exploration and dissection of wood

phenomic and genomic data in VR (Figure 6.8). The laboratories host live coding sessions and provide one-to-one mentorship through peer instruction. Learners are taught how the efficiency, efficacy, stability, performance, and security of the next-generation blockchain and the traceability, transparency, swiftness, and accountability it promotes could be used to empower X-AI applications in wood science and technology (Harfouche and Nakhle, 2023). They then learn how to design and implement human-centric X-AI algorithms which promise to embed human knowledge into its architecture and to combine it with connections and correlations automatically discovered in big data (Harfouche et al., 2023). During that process, learners in VR benefit from pair programming practices and from rubber duck debugging, a technique that helps them articulate the technical problems they are facing, explaining each step of their code to a 3D model of a rubber duck until they notice and debug the problem (Figure 6.8). This VR environment provides practical and useful resources for novices and experts to harness the potential of big data in X-AI, blockchain, and XR based applications to crunch those data and join the fight against illegal logging (Harfouche and Nakhle, 2023). It can be experienced synchronously by multiple people, anywhere, at any time.

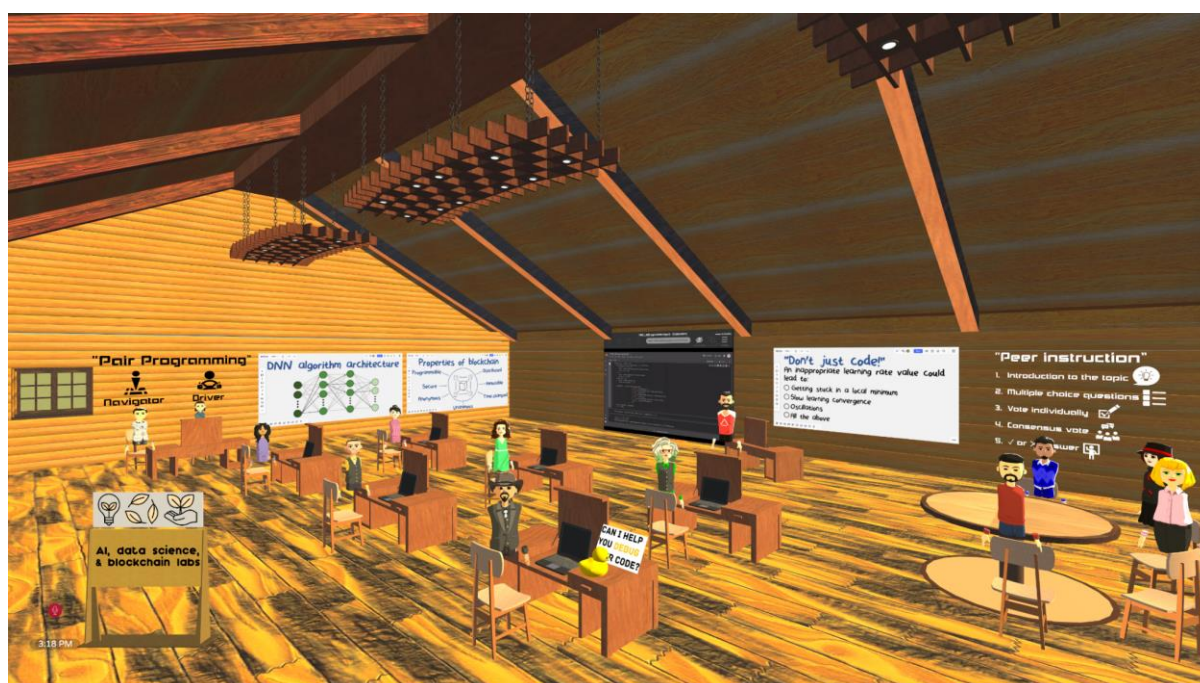


Figure 0.8. 'AI, Data Science, and Blockchain Labs' to drive wood science and technology innovation.

6.3 Sustainability and circularity that you can actually see

WOODMeta grows the educational experience to significant factual contributions to global developmental initiatives and sustainability trends. It makes virtual interactions feel like face-to-face even when participants and users are far across the world; thus, it removes distance barriers in education and promotes learning, training, research, and innovation in the metaverse without having to drive or take long flights. This can be a powerful tool to avoid unnecessary transportation and reduce carbon emissions (SDG 13). WOODMeta can also reduce resource consumption when designing and prototyping wood products by using digital wood instead of real one, enabling significant savings in raw materials (SDG 12). Additionally, the WOODMeta woodshops can help woodworkers test their designs in VR to preempt flaws and troubleshoot errors before the building process begins, again reducing material waste (SDG 12) and promoting sustainable economic growth (SDG 8).

Of equal importance, WOODMeta can help users harness the next circular green innovations in wood products by, for example, simulating how wood processing residue (e.g., bark, sawdust) at the sawmill can be used directly to produce heat and electricity to fuel the mill's primary manufacturing, or indirectly to produce bio-based products such as pellets (Figure 6.1). This promotes the use of affordable, reliable, sustainable, and modern energy (SDG 7), and raises awareness on climate change mitigation (SDG 13). Similarly, a simulation in the woodshops shows how sawdust can be collected (Figure 6.9a) and mixed with an environmentally friendly binder (e.g., lignin, an adhesive material that comes from trees and naturally helps hold together wood) to be used in a 3D printer to churn out wood products such as furniture, medical appliances, fashion accessories, or building blocks for prototyping tall wood buildings, as for example, Brock Commons Tallwood House, an innovative 18 story student residence building at the University of British Columbia (Figure 6.9b). The simulated 3D printer and its 3D printed products bridge the gap between design and reality by enabling woodworkers to visualize how their designs will behave or fit in different real-world scenarios, while reducing the need for costly physical testing. It also presents the 3D printing process in an interactive virtual environment allowing users to experientially learn the operating principles of the state-of-the-art 3D printers which, in real life, are expensive and time-consuming, not to mention that their maintenance require knowledge that users do not necessarily have. This can advance wood

scientific research and inspire the upgrade of the technological capabilities in the wood industry (SDG 9).

In addition, the WOODMeta raises awareness on how wood can be leveraged to create stylish buildings that respect the environment right from the drawing board, using less space and consuming less energy; it does so by showcasing a custom-designed 3D model named 'Zero Energy Passive Casa' which mimics the 'mini zero energy passive house', a small footprint house designed by carpentry students at Seattle Central College that uses Shou Sugi Ban charred cladding and solar generation, airtight construction and insulation, and heat recovery features to minimize the energy usage levels for both heating and cooling by maintaining stable internal temperatures, regardless of climate (Seattle Central College, 2020) (Figure 6.9c). This can encourage making human settlements affordable, accessible, resilient, and sustainable (SDG 11).

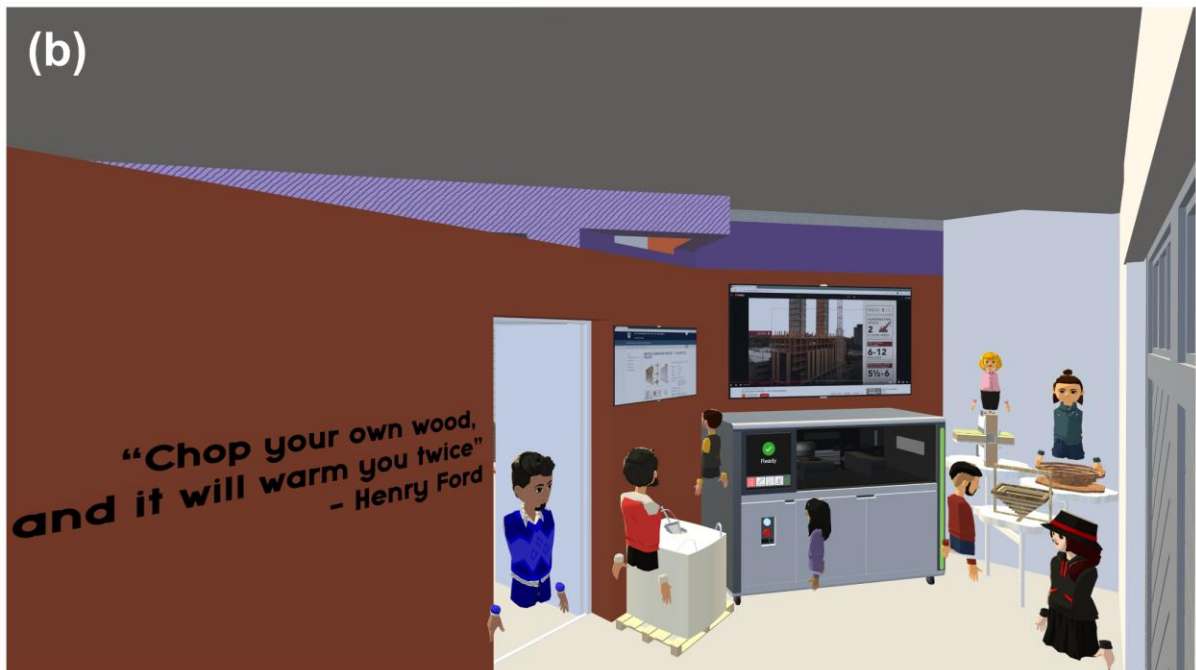




Figure 0.9. Prioritizing sustainability and circularity in wood science and technology.

(a) The dust collector in the 'Wood Art, Architecture, and Design Workshop'. (b) 3D printing area in the Wood Art, Architecture, and Design Workshop. (c) 'Zero Energy Passive Casa'.

6.4 Conclusions

In this article, we have conceived, designed, and built the WOODMeta, a wood science and technology metaverse campus for teaching, learning, training, capacity building, researching, and innovating. It is carefully designed to remove barriers to interaction, communication, collaboration, exploration, and creativity by combining immersive virtual environments that break the traditional bounds of four walls and that feature expressive, diverse, and inclusive avatars and spatial sound technology to generate a sense of true presence and belonging. However, despite the increasing accessibility and affordability of XR technologies and the WOODMeta's ability to create a landscape of diversity and inclusion, there are still communities that might be excluded because the digital divide between high-income and low-income countries remains a structural barrier. In fact, since the virtual worlds, environments, avatars, and 3D objects of the metaverse will reside digitally on a computing server, the delivery of the full range of experiences anticipated by the metaverse will require extensive compute power. While supercomputers and high-speed network infrastructures, such as 4G and 5G, in high-income countries can support the metaverse, a considerable percentage of the world's population, especially in low-income countries, do not have

supercomputers nor internet access. Democratizing computation and enabling inclusive connectivity through investments in sustainable digital public infrastructure can help bridging the digital divide gap and can allow people in those countries to tap into the opportunities that the metaverse has to offer. However, despite advances in the technology supporting the metaverse, there remains a significant divide between potential XR content creators and developers because available resources such as content creation and world building tools may address mainly professional developers and may be kept out of reach for non-computer scientist users who wish to create and innovate their own XR experiences. Thus, to further increase the metaverse and WOODMeta's diversity and inclusion, we should also focus on how to apply digital literacy skills in education and training, and to develop powerful, user-friendly, and open-source tools for creators to ensure that everyone can participate in a meaningful way. On another level, the midpoint of the timeline for the UN 2030 agenda for sustainable development has been reached, and the scientific community is not fully prepared yet. Looking to the future, the growing awareness about the metaverse technology gives the WOODMeta endless opportunities to get closer to the finish line of the UN agenda, categorically, by ensuring quality education for wood science and technology (SDG 4), by supporting decent work and economic growth as well as industry innovation and infrastructure for wood-based society (SDGs 8 and 9), and by promoting affordable and clean energy, sustainable cities and communities, and climate action for all (SDGs 7, 11, 12, and 13). Additionally, the WOODMeta can enhance regional and international cooperation on, and access to, wood science and innovation, and can encourage and promote effective global public-private partnerships (SDG 17). We still, however, need to engage in XR technology design initiatives that strive to secure a human-centric ethical metaverse for all.

Chapter 7 General conclusion and outlook

In the face of a rapidly changing climate, it will be necessary to design climate-resilient food and bioenergy crops that can maintain their productivity and thrive under extreme climate change scenarios.

At present, thanks to the advancements in phenotyping platforms and sensors, large amounts of plant phenomic data are becoming available. AI algorithms are capable of discovering patterns and correlations in the complex structure of these datasets due to their ability of capturing nonlinear relationships. Indeed, AI revealed its remarkable power in plant science by achieving unprecedented predictive successes in phenomic data analysis. Nevertheless, the effectiveness and wider adoption of AI is limited by the models' inability to explain their decisions and actions to human users. Fortunately, efforts in X-AI research have led to the development of several approaches that make the inner workings of AI models understandable to humans and have improved our ability to explain their predictions.

Building on this research, and anticipating such daunting challenges, this thesis develops concrete and innovative computational solutions to challenges associated with data-driven plant phenomics research and proposes novel X-AI systems to achieve the lofty goal of developing better digital technologies for all. As a result, an X-AI phenomic data analysis workflow was successfully designed and applied to accurately identify and classify tomato leaf diseases and spider mites which can impose significant crop losses. This is essential because farmers and plant pathologists often still detect diseases with their eyes and, in many cases, they are not able to accurately diagnose the symptoms, which can turn disease control efforts into a waste of time and money. On the other hand, X-AI achieved an accuracy rate as high as 96.46 percent. The workflow leveraged *post hoc* X-AI algorithms to explain which visual symptoms were used to make predictions. However, it might be possible that the generated explanations do not provide sufficient detail to understand what the black box model is actually computing because *post hoc* algorithms approximate the inner workings of black box models by, for example, data perturbation. In contrast, interpretable by design algorithms provide faithful explanations to what the model actually computes without needing any additional algorithms. That said, to the best of my knowledge, this thesis introduced, for the first time, the implementation of an inherently explainable algorithm to analyze plant phenomic data. The algorithm was

improved to scale and run on a computer cluster, where a model was trained to classify cassava leaf diseases and to explain its own reasoning for each prediction.

But, to better elucidate complex biological processes, it is imperative for future X-AI research to take an integrative approach that analyzes multiomic data (epigenomics, transcriptomics, proteomics, metabolomics, phenomics, and microbiomics) in addition to their metadata, temporal, geospatial, and environmental data. While current X-AI models are designed to solve well-defined tasks of single-omics datasets, a custom X-AI algorithm can be developed to create an ensemble of X-AI models, each performing different well-defined tasks. Such data integration algorithm will help dissect the biological mechanisms that underlie desirable traits and to shed light on the flow of information that underpins plant responses to various environmental stressors, which is crucial for designing new crop ideotypes that are optimized for niche environments. Multiomic data integration can also, for example, advance our understanding of the regulation mechanisms of lignin biosynthesis and its role in wood formation; this will ultimately lead to better genome editing and genome engineering of forest trees, with a focus on wood.

In spite of its successes, for AI to be trustworthy, future research efforts are needed to build on the novel human-centric X-AI system architecture proposed in this thesis, and to design and develop algorithms that reflect the information, goals, and constraints that their models tend to weigh when making a prediction; these human-centric algorithms should analyze input data taking into consideration human expert knowledge and metadata that provides an understanding of the process that generated that input. Human-centric X-AI algorithms should also be inherently interpretable; thus their output should be accompanied by “why” messages that explain the reasoning behind each prediction. Finally, these algorithms should also enable the monitoring of their inputs and outputs, keeping HITL to ensure the sanity of predictions and that unwanted biases have not inadvertently crept into their predictive models. In this regard, the robustness of human-centric X-AI models against biases can be boosted by increasing the diversity of training datasets. Notably, XR technologies such as VR can help generate phenomic datasets that account for and effectively cover edge cases that can occur in the real world, ensuring that training data are consistent and bias-free. To achieve this simulation, future research is needed to leverage AI to generate dynamic 3D models of plants that simulate plant

growth in VR, under various environmental and management conditions. Virtual cameras can be programmed to generate synthetic labelled images periodically to create synthetic training datasets.

But XR is more than just a tool for synthetic data generation; it is also rich with various fascinating applications, providing a wide range of possibilities that will shape the future of plant phenomics and biotechnology. For example, an AR app can leverage X-AI to help farmers classify diseases of crops by pointing a tablet or a phone's camera at a diseased leaf. Accordingly, AR can display relevant information on the emerging diseases and suggest actionable prevention and mitigation strategies. In another example, an AR app can use a camera or an X-ray fluorescence (XRF) spectrometer, and by leveraging X-AI, can identify the species of the trees, logs, and wood sections. Such an app holds great promise for combating illegal logging and illegal log trade. Of equal importance, XR has the capacity to help to crack down on illegal logging through the digital twins technological innovation. Unlike virtual environments and simulations, digital twins are driven by data collected from cameras and sensors in near-real-time, and thus, they mirror physical or biological objects such as forests and trees, allowing for a better understanding of forest disturbances (e.g., logging). In addition, digital twins can, for example, record the complete history of a tree in the forest so that its origin can be traced, even if the physical tree no longer exists in real life. In addition, digital twins will provide immersive 3D digital maps of farms, entire forests, and plantations, bringing unlimited opportunities, from phenotypic features to forest resource inventory measurement capabilities. For example, such technology can be used to create digital twins of crops in a greenhouse to predict their growth and development in response to real-time environmental factors, speeding breeding programs and allowing for better informed management decisions.

XR is poised to transform how we research and innovate plant phenomics and biotechnology, as demonstrated in the X-RAI park and the WOODMeta metaverse campus, introduced in this thesis. The shift to human-centric AI and human-centric XR will bring forth benefits to the digital phenomics community that should be embraced moving forward. Whether for phenomic data visualization, for training X-AI with XR-generated data, or for conferring intelligence on XR experiences, the combination of these emerging and enabling digital technologies can be a powerful tool to devise and

employ new approaches to digitize and monitor forests, fields, and greenhouses in a time-resolved fashion which can lead to the design of biotech crop ideotypes and plants with adequate wood properties that can grow in niche environments to ultimately achieve food and energy security.

Despite the substantial progress, several important outstanding questions remain. For example, what are the challenges to test and validate human-centric X-AI tasks in uncontrolled and unpredictable agricultural environments? How can human-centric X-AI, with the aid of XR, help embed human knowledge into the learning process? How to effectively implement the data architecture blueprint and technology infrastructure to gain the computing power and performance needed to analyze massive amounts of omic data at multiple scales? When will human-centric X-AI infrastructure, analytics, and applications be available for the plant breeding research community, and how can XR help? How can we encourage the breeding, farming, AI, data science, and IS communities to interact, embrace interdisciplinary research, and exchange knowledge for the implementation of longer-term plant breeding X-AI and XR projects? How can XR be leveraged to attract and facilitate public and private investments in human-centric X-AI and XR combination projects? What roles will plant scientists, data scientists, IS scientists, farmers, breeders, and regulators assume in the development, implementation, evaluation, and deployment of human-centric X-AI and XR technologies? And in what ways can we encourage the academic, AI, and XR communities to develop awareness strategies and educational materials to ensure that the current and next generation of researchers will look at X-AI and XR through a human-centric approach?

While plant digital phenomics and biotechnology, powered by human-centric X-AI and XR, offers a promising avenue for meeting the food and energy demands, ethical and sociocultural concerns must be addressed to ensure widespread adoption and public trust. Ethically and socially responsible solutions that are relevant to all, and that can be communicated to the public in a convincing manner need to be developed. As these digital technologies require a supporting hardware infrastructure, rich–poor digital divides should also be addressed to ensure they are implemented responsibly and to support better breeding decisions not only in high-income nations, but also in low-income nations. Integrating the efforts of researchers and relevant stakeholders, including, but not limited to, policymakers, agricultural technopreneurs, funders, and

startups can be an effective way to ensure that these technologies can have a positive impact on society. However, for this collaboration to be successful, it requires careful planning, coordination, and effective science communication.

Finally, human-centric AI, being explainable, informed AI keeping HITL and value-aligned XR would be important to renew an emphasis on creativity and innovative science design thinking, which contribute to a more disruptive plant digital phenomics. Such recommendations will help scientists and scientists-in-training to focus on emphasizing predictability rather than unexpected results and producing results based on meaningful analyses that will be useful for decision making in plant science.

Resources

ⁱ<https://www.startus-insights.com/innovators-guide/top-10-industry-4-0-trends-innovations-in-2021/>

ⁱⁱ<https://sdgs.un.org/goals>

ⁱⁱⁱ<https://www.ecohealthalliance.org/>

^{iv}<https://www.planetaryhealthalliance.org/planetary-health>

^v<https://www.cdc.gov/onehealth/index.html>

^{vi}<https://github.com/HarfoucheLab/Ready-Steady-Go-AI>

^{vii}<https://www.top500.org/lists/top500/2022/06/>

^{viii}<https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>

^{ix}<https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>

^x<https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064>

^{xi}<https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>

^{xii}<https://github.com/HarfoucheLab/a-primer-on-AI-in-plant-digital-phenomics>

References

- Abdelghafour, F., Keresztes, B., Deshayes, A., Germain, C., Da Costa, J.-P., 2021. An annotated image dataset of downy mildew symptoms on Merlot grape variety. *Data Br.* 37, 107250. <https://doi.org/10.1016/j.dib.2021.107250>
- Abdelhalim, I.S.A., Mohamed, M.F., Mahdy, Y.B., 2021. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Syst. Appl.* 165, 113922. <https://doi.org/10.1016/j.eswa.2020.113922>
- Acharya, M., Burner, D.M., Ashworth, A.J., Fritschi, F.B., Adams, T.C., 2018. Growth rates of giant miscanthus (*Miscanthus × giganteus*) and giant reed (*Arundo donax*) in a low-input system in Arkansas, USA. *Am. J. Plant Sci.* 99, 2371–2384. <https://doi.org/10.4236/ajps.2018.912172>
- Agar, J., 2020. What is science for? The Lighthill report on artificial intelligence reinterpreted. *Br. J. Hist. Sci.* 53, 289–310. <https://doi.org/10.1017/S0007087420000230>
- Ahnert, S.E., 2017. Structural properties of genotype–phenotype maps. *J. R. Soc. Interface* 14, 20170275. <https://doi.org/10.1098/rsif.2017.0275>
- Ahrar, M., Doneva, D., Tattini, M., Brunetti, C., Gori, A., Rodeghiero, M., Wohlfahrt, G., Biasiolo, F., Varotto, C., Loreto, F., Velikova, V., 2017. Phenotypic differences determine drought stress responses in ecotypes of *Arundo donax* adapted to different environments. *Exp. Bot.* 68, 2439–2451. <https://doi.org/10.1093/jxb/erx125>
- Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y.K., D'Ambra, J., Shen, K.N., 2021. Algorithmic bias in data-driven innovation in the age of AI. *Int. J. Inf. Manage.* 60, 102387. <https://doi.org/10.1016/j.ijinfomgt.2021.102387>
- Alberch, P., 1991. From genes to phenotype: dynamical systems and evolvability. *Genetica* 84, 5–11. <https://doi.org/10.1007/BF00123979>
- Alessandrini, M., Calero Fuentes Rivera, R., Falaschetti, L., Pau, D., Tomaselli, V., Turchetti, C., 2021. A grapevine leaves dataset for early detection and classification of esca disease in vineyards through machine learning. *Data Br.* 35, 106809. <https://doi.org/10.1016/j.dib.2021.106809>

- Ali, H., Lali, M.I., Nawaz, M.Z., Sharif, M., Saleem, B.A., 2017. Symptom based automated detection of citrus diseases using color histogram and textural descriptors. *Comput. Electron. Agric.* 138, 92–104. <https://doi.org/10.1016/j.compag.2017.04.008>
- Altaheri, H., Alsulaiman, M., Muhammad, G., 2019a. Date fruit classification for robotic harvesting in a natural environment using deep learning. *IEEE Access* 7, 117115–117133. <https://doi.org/10.1109/ACCESS.2019.2936536>
- Altaheri, H., Alsulaiman, M., Muhammad, G., Amin, S.U., Bencherif, M., Mekhtiche, M., 2019b. Date fruit dataset for intelligent harvesting. *Data Br.* 26, 104514. <https://doi.org/10.1016/j.dib.2019.104514>
- Altman, A., Fan, L., Foyer, C., Cowling, W., Mittler, R., Qaim, M., Weber, A., Reynolds, M., Varshney, R., Fernie, A., 2021. Past and future milestones of plant breeding. *Trends Plant Sci.* 26, 530–538. <https://doi.org/10.1016/j.tplants.2021.03.013>
- Andreu-Rodríguez, J., Pérez-Espinosa, A., Pérez-Murcia, M.D., Moral, R., Agulló, E., Ferrández-Villena, M., Ferrández-García, M.T., Bustamante, M.A., 2017. Near infrared reflectance spectroscopy (NIRS) for the assessment of biomass production and C sequestration by *Arundo donax* L. in salt-affected environments. *Agric. Water Manag.* 183, 94–100. <https://doi.org/10.1016/j.agwat.2016.10.005>
- Ansarifar, J., Akhavizadegan, F., Wang, L., 2020. Performance prediction of crosses in plant breeding through genotype by environment interactions. *Sci. Rep.* 10, 11533. <https://doi.org/10.1038/s41598-020-68343-1>
- Aslan, J., Mayers, K., Koomey, J.G., France, C., 2018. Electricity intensity of internet data transmission: untangling the estimates. *J. Ind. Ecol.* 22, 785–798. <https://doi.org/10.1111/jiec.12630>
- Association 42, 2021. What is 42? An Unrivalled Concept to Discover [WWW Document]. URL <https://42.fr/en/what-is-42/42-program-explained/> (accessed 1.7.23).
- Australian Forest Education Alliance, 2022. ForestVR™ [WWW Document]. URL <https://forestlearning.edu.au/forestvr.html> (accessed 11.15.22).
- Azodi, C.B., Pardo, J., VanBuren, R., de los Campos, G., Shiu, S.-H., 2020a.

- Transcriptome-based prediction of complex traits in maize. *Plant Cell* 32, 139–151. <https://doi.org/10.1105/tpc.19.00332>
- Azodi, C.B., Tang, J., Shiu, S.-H., 2020b. Opening the black box: interpretable machine learning for geneticists. *Trends Genet.* 36, 442–455. <https://doi.org/10.1016/j.tig.2020.03.005>
- Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. <https://doi.org/10.1109/TPAMI.2016.2644615>
- Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 3626–3633. <https://doi.org/10.1109/ICRA.2017.7989417>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Barz, B., Denzler, J., 2020. Deep learning on small datasets without pre-training using cosine loss. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1360–1369. <https://doi.org/10.1109/WACV45572.2020.9093286>
- Basu, S., Kumbier, K., Brown, J.B., Yu, B., 2018. Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* 115, 1943–1948. <https://doi.org/10.1073/pnas.1711236115>
- Beckmann, M., Ebecken, N.F.F., Pires de Lima, B.S.L., 2015. A KNN undersampling approach for data balancing. *J. Intell. Learn. Syst. Appl.* 07, 104–116. <https://doi.org/10.4236/jilsa.2015.74010>
- Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M.L., Bareth, G., 2015. Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *Int. J. Appl. Earth Obs. Geoinf.* 39, 79–87. <https://doi.org/10.1016/j.jag.2015.02.012>

- Bhattacharya, A., Cui, Y., 2017. A GPU-accelerated algorithm for biclustering analysis and detection of condition-dependent coexpression network modules. *Sci. Rep.* 7, 4162. <https://doi.org/10.1038/s41598-017-04070-4>
- Bloice, M.D., Roth, P.M., Holzinger, A., 2019. Biomedical image augmentation using Augmentor. *Bioinformatics* 35, 4522–4524. <https://doi.org/10.1093/bioinformatics/btz259>
- Bloice, M.D., Stocker, C., Holzinger, A., 2017. Augmentor: an image augmentation library for machine learning.
- Bodner, G., Nakhforoosh, A., Arnold, T., Leitner, D., 2018. Hyperspectral imaging: a novel approach for plant root phenotyping. *Plant Methods* 14, 84. <https://doi.org/10.1186/s13007-018-0352-1>
- Bonde, M.T., Makransky, G., Wandall, J., Larsen, M. V, Morsing, M., Jarmer, H., Sommer, M.O.A., 2014. Improving biotech education through gamified laboratory simulations. *Nat. Biotechnol.* 32, 694–697. <https://doi.org/10.1038/nbt.2955>
- Bosilj, P., Aptoula, E., Duckett, T., Cielniak, G., 2020. Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. *J. F. Robot.* 37, 7–19. <https://doi.org/10.1002/rob.21869>
- Boulter, E., Colombelli, J., Henriques, R., Féral, C.C., 2022. The LEGO® brick road to open science and biotechnology. *Trends Biotechnol.* 40, 1073–1087. <https://doi.org/10.1016/j.tibtech.2022.02.003>
- Boyd, D., Crawford, K., 2012. Critical questions for big data. *Information, Commun. Soc.* 15, 662–679. <https://doi.org/10.1080/1369118X.2012.678878>
- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24, 123–140. <https://doi.org/10.1007/BF00058655>
- Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brigato, L., Iocchi, L., 2021. A close look at deep learning with small data. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 2490–2497. <https://doi.org/10.1109/ICPR48806.2021.9412492>

- Brink, S.C., 2021. 25 years of trends in plant science: we should all be plant worshippers. *Trends Plant Sci.* 26, 527–529. <https://doi.org/10.1016/j.tplants.2021.04.001>
- Brostow, G.J., Fauqueur, J., Cipolla, R., 2009. Semantic object classes in video: a high-definition ground truth database. *Pattern Recognit. Lett.* 30, 88–97. <https://doi.org/10.1016/j.patrec.2008.04.005>
- Brous, P., Janssen, M., Herder, P., 2020. The dual effects of the Internet of Things (IoT): a systematic review of the benefits and risks of IoT adoption by organizations. *Int. J. Inf. Manage.* 51, 101952. <https://doi.org/10.1016/j.ijinfomgt.2019.05.008>
- Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>
- Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med. Image Anal.* 71, 102062. <https://doi.org/10.1016/j.media.2021.102062>
- Burnett, A.C., Serbin, S.P., Davidson, K.J., Ely, K.S., Rogers, A., 2021. Detection of the metabolic response to drought stress using hyperspectral reflectance. *J. Exp. Bot.* 72, 6474–6489. <https://doi.org/10.1093/jxb/erab255>
- Bzdok, D., Altman, N., Krzywinski, M., 2018. Statistics versus machine learning. *Nat. Methods* 15, 233–234. <https://doi.org/10.1038/nmeth.4642>
- Cano-Ruiz, J., Ruiz Galea, M., Amorós, M.C., Alonso, J., Mauri, P.V., Lobo, M.C., 2020. Assessing *Arundo donax* L. in vitro-tolerance for phytoremediation purposes. *Chemosphere* 252, 126576. <https://doi.org/10.1016/j.chemosphere.2020.126576>
- Cao, L., 2017. Data science: a comprehensive overview. *ACM Comput. Surv.* 50, 1–42. <https://doi.org/10.1145/3076253>
- Carpenter, S., Walker, B., Anderies, J.M., Abel, N., 2001. From metaphor to measurement: resilience of what to what? *Ecosystems* 4, 765–781. <https://doi.org/10.1007/s10021-001-0045-9>

- Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. *Nat. Commun.* 11, 3673. <https://doi.org/10.1038/s41467-020-17478-w>
- Caudell, T.P., Mizell, D.W., 1992. Augmented reality: an application of heads-up display technology to manual manufacturing processes. In: *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*. IEEE, pp. 659–669. <https://doi.org/10.1109/HICSS.1992.183317>
- Chakravarti, R., Meng, X., 2009. A study of color histogram based image retrieval. In: *2009 Sixth International Conference on Information Technology: New Generations*. IEEE, pp. 1323–1328. <https://doi.org/10.1109/ITNG.2009.126>
- Chao, X., Kou, G., Li, T., Peng, Y., 2018. Jie Ke versus AlphaGo: a ranking approach using decision making method for large-scale data with incomplete information. *Eur. J. Oper. Res.* 265, 239–247. <https://doi.org/10.1016/j.ejor.2017.07.030>
- Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., Stachniss, C., 2017. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Rob. Res.* 36, 1045–1052. <https://doi.org/10.1177/0278364917720510>
- Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 8930–8941.
- Chen, M., Shlezinger, N., Poor, H.V., Eldar, Y.C., Cui, S., 2021. Communication-efficient federated learning. *Proc. Natl. Acad. Sci.* 118, e2024789118. <https://doi.org/10.1073/pnas.2024789118>
- Cheng, Y., Wang, D., Zhou, P., Zhang, T., 2020. A survey of model compression and acceleration for deep neural networks.
- Chhetri, H.B., Furches, A., Macaya-Sanz, D., Walker, A.R., Kainer, D., Jones, P.,

- Harman-Ware, A.E., Tschaplinski, T.J., Jacobson, D., Tuskan, G.A., DiFazio, S.P., 2020. Genome-wide association study of wood anatomical and morphological traits in *Populus trichocarpa*. *Front. Plant Sci.* 11, 545748. <https://doi.org/10.3389/fpls.2020.545748>
- Chhetri, H.B., Macaya-Sanz, D., Kainer, D., Biswal, A.K., Evans, L.M., Chen, J., Collins, C., Hunt, K., Mohanty, S.S., Rosenstiel, T., Ryno, D., Winkeler, K., Yang, X., Jacobson, D., Mohnen, D., Muchero, W., Strauss, S.H., Tschaplinski, T.J., Tuskan, G.A., DiFazio, S.P., 2019. Multitrait genome-wide association analysis of *Populus trichocarpa* identifies key polymorphisms controlling morphological and physiological traits. *New Phytol.* 223, 293–309. <https://doi.org/10.1111/nph.15777>
- Childs, C.M., Washburn, N.R., 2019. Embedding domain knowledge for machine learning of complex material systems. *MRS Commun.* 9, 806–820. <https://doi.org/10.1557/mrc.2019.90>
- Chitwood, D.H., Otoni, W.C., 2017a. Morphometric analysis of *Passiflora* leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. *Gigascience* 6, giw008. <https://doi.org/10.1093/gigascience/giw008>
- Chitwood, D.H., Otoni, W.C., 2017b. Divergent leaf shapes among *Passiflora* species arise from a shared juvenile morphology. *Plant Direct* 1, e00028. <https://doi.org/10.1002/pld3.28>
- Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput. Electron. Agric.* 151, 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
- Chouhan, S.S., Singh, U.P., Kaul, A., Jain, S., 2019. A data repository of leaf images: practice towards plant conservation with plant pathology. In: 2019 4th International Conference on Information Systems and Computer Networks (ISCON). IEEE, pp. 700–707. <https://doi.org/10.1109/ISCON47742.2019.9036158>
- Clarke, D.J.B., Jeon, M., Stein, D.J., Moiseyev, N., Kropiwnicki, E., Dai, C., Xie, Z., Wojciechowicz, M.L., Litz, S., Hom, J., Evangelista, J.E., Goldman, L., Zhang, S.,

- Yoon, C., Ahamed, T., Bhuiyan, S., Cheng, M., Karam, J., Jagodnik, K.M., Shu, I., Lachmann, A., Ayling, S., Jenkins, S.L., Ma'ayan, A., 2021. Appyters: turning Jupyter notebooks into data-driven web apps. *Patterns* 2, 100213. <https://doi.org/10.1016/j.patter.2021.100213>
- Cliff, A., Romero, J., Kainer, D., Walker, A., Furches, A., Jacobson, D., 2019. A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. *Genes (Basel)*. 10, 996. <https://doi.org/10.3390/genes10120996>
- Coomes, D.A., Dalponte, M., Jucker, T., Asner, G.P., Banin, L.F., Burslem, D.F.R.P., Lewis, S.L., Nilus, R., Phillips, O.L., Phua, M.H., Qie, L., 2017. Area-based vs tree-centric approaches to mapping forest carbon in Southeast Asian forests from airborne laser scanning data. *Remote Sens. Environ.* 194, 77–88. <https://doi.org/10.1016/j.rse.2017.03.017>
- Copeland, J., Proudfoot, D., 2009. Turing's test. In: *Parsing the Turing Test*. Springer Netherlands, Dordrecht, pp. 119–138. https://doi.org/10.1007/978-1-4020-6710-5_9
- Coppens, F., Wuyts, N., Inzé, D., Dhondt, S., 2017. Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Curr. Opin. Syst. Biol.* 4, 58–63. <https://doi.org/10.1016/j.coisb.2017.07.002>
- Corno, L., Pilu, R., Adani, F., 2014. *Arundo donax* L.: A non-food crop for bioenergy and bio-compound production. *Biotechnol. Adv.* 32, 1535–1549. <https://doi.org/10.1016/j.biotechadv.2014.10.006>
- Cox, M., Ellsworth, D., 1997. Application-controlled demand paging for out-of-core visualization. In: *Proceedings. Visualization '97 (Cat. No. 97CB36155)*. IEEE, pp. 235–244. <https://doi.org/10.1109/VISUAL.1997.663888>
- Cropper, A., Morel, R., 2021. Learning programs by learning from failures. *Mach. Learn.* 110, 801–856. <https://doi.org/10.1007/s10994-020-05934-z>
- Cruz, J.A., Yin, X., Liu, X., Imran, S.M., Morris, D.D., Kramer, D.M., Chen, J., 2016. Multi-modality imagery database for plant phenotyping. *Mach. Vis. Appl.* 27, 735–749. <https://doi.org/10.1007/s00138-015-0734-6>

- Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., van Oeveren, J., Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S.-A., Scholz, U., van Schriek, M., Seren, Ü., Usadel, B., Weise, S., Kersey, P., Krajewski, P., 2016. Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12, 44. <https://doi.org/10.1186/s13007-016-0144-4>
- Dalla Corte, A.P., Rex, F.E., Almeida, D.R.A. de, Sanquetta, C.R., Silva, C.A., Moura, M.M., Wilkinson, B., Zambrano, A.M.A., Cunha Neto, E.M. da, Veras, H.F.P., Moraes, A. de, Klauberg, C., Mohan, M., Cardil, A., Broadbent, E.N., 2020. Measuring individual tree diameter and height using GatorEye high-density UAV-LiDAR in an integrated crop-livestock-forest system. *Remote Sens.* 12, 863. <https://doi.org/10.3390/rs12050863>
- Danilevicz, M.F., Bayer, P.E., Nestor, B.J., Bennamoun, M., Edwards, D., 2021. Resources for image-based high-throughput phenotyping in crops and data sharing challenges. *Plant Physiol.* 187, 699–715. <https://doi.org/10.1093/plphys/kiab301>
- Dart, E., Rotman, L., Tierney, B., Hester, M., Zurawski, J., 2014. The science DMZ: a network design pattern for data-intensive science. *Sci. Program.* 22, 701405. <https://doi.org/10.3233/SPR-140382>
- David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M.A., Pozniak, C., de Solan, B., Hund, A., Chapman, S.C., Baret, F., Stavness, I., Guo, W., 2020. Global wheat head detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics* 2020, 1–12. <https://doi.org/10.34133/2020/3521852>
- David, E., Serouart, M., Smith, D., Madec, S., Velumani, K., Liu, S., Wang, X., Pinto, F., Shafiee, S., Tahir, I.S.A., Tsujimoto, H., Nasuda, S., Zheng, B., Kirchgessner, N., Aasen, H., Hund, A., Sadhegi-Tehran, P., Nagasawa, K., Ishikawa, G., Dandrifosse, S., Carlier, A., Dumont, B., Mercatoris, B., Evers, B., Kuroki, K., Wang, H., Ishii, M., Badhon, M.A., Pozniak, C., LeBauer, D.S., Lillemo, M., Poland, J., Chapman, S., de Solan, B., Baret, F., Stavness, I., Guo, W., 2021.

Global wheat head detection 2021: an improved dataset for benchmarking wheat head detection methods. *Plant Phenomics* 2021, 1–9. <https://doi.org/10.34133/2021/9846158>

Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.-S., Wang, C.-H., Hsu, C.-N., Lee, C.K., Ruan, P., Xu, D., Wu, D., Huang, E., Kitamura, F.C., Lacey, G., de Antônio Corradi, G.C., Nino, G., Shin, H.-H., Obinata, H., Ren, H., Crane, J.C., Tetreault, J., Guan, J., Garrett, J.W., Kaggie, J.D., Park, J.G., Dreyer, K., Juluru, K., Kersten, K., Rockenbach, M.A.B.C., Linguraru, M.G., Haider, M.A., AbdelMaseeh, M., Rieke, N., Damasceno, P.F., e Silva, P.M.C., Wang, P., Xu, S., Kawano, S., Sriswasdi, S., Park, S.Y., Grist, T.M., Buch, V., Jantarabenjakul, W., Wang, W., Tak, W.Y., Li, X., Lin, X., Kwon, Y.J., Quraini, A., Feng, A., Priest, A.N., Turkbey, B., Glicksberg, B., Bizzo, B., Kim, B.S., Tor-Díez, C., Lee, C.-C., Hsu, C.-J., Lin, C., Lai, C.-L., Hess, C.P., Compas, C., Bhatia, D., Oermann, E.K., Leibovitz, E., Sasaki, H., Mori, H., Yang, I., Sohn, J.H., Murthy, K.N.K., Fu, L.-C., de Mendonça, M.R.F., Fralick, M., Kang, M.K., Adil, M., Gangai, N., Vateekul, P., Elnajjar, P., Hickman, S., Majumdar, S., McLeod, S.L., Reed, S., Gräf, S., Harmon, S., Kodama, T., Puthanakit, T., Mazzulli, T., de Lator, V.L., Rakvongthai, Y., Lee, Y.R., Wen, Y., Gilbert, F.J., Flores, M.G., Li, Q., 2021. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* 27, 1735–1743. <https://doi.org/10.1038/s41591-021-01506-3>

Delipetrev, B., Tsinaraki, C., Kostić, U., 2020. AI watch, historical evolution of artificial intelligence: analysis of the three main paradigm shifts in AI., Publications Office of the European Union. <https://doi.org/10.2760/801580>

Deng, C., Ji, X., Rainey, C., Zhang, J., Lu, W., 2020. Integrating machine learning with human knowledge. *iScience* 23, 101656. <https://doi.org/10.1016/j.isci.2020.101656>

Dhondt, S., Wuyts, N., Inzé, D., 2013. Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci.* 18, 428–439. <https://doi.org/10.1016/j.tplants.2013.04.008>

Dias, P.A., Tabb, A., Medeiros, H., 2018. Multispecies fruit flower detection using a refined semantic segmentation network. *IEEE Robot. Autom. Lett.* 3, 3003–3010.

<https://doi.org/10.1109/LRA.2018.2849498>

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: *Multiple Classifier Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–15.

Dobos, O., Horvath, P., Nagy, F., Danko, T., Viczián, A., 2019. A deep learning-based approach for high-throughput hypocotyl phenotyping. *Plant Physiol.* 181, 1415–1424. <https://doi.org/10.1104/pp.19.00728>

Dobrescu, A., Giuffrida, M.V., Tsafaris, S.A., 2019. Understanding deep neural networks for regression in leaf counting. In: Brendel, W., Amer, M. (Eds.), 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 2600–2608. <https://doi.org/10.1109/CVPRW.2019.00316>

Donalek, C., Djorgovski, S.G., Cioc, A., Wang, A., Zhang, J., Lawler, E., Yeh, S., Mahabal, A., Graham, M., Drake, A., Davidoff, S., Norris, J.S., Longo, G., 2014. Immersive and collaborative data visualization using virtual reality platforms. In: Lin, J., Pei, J., Hu, X., Chang, W., Nambiar, R., Aggarwal, C., Cercone, N., Honavar, V., Huan, J., Mobasher, B., Pyne, S. (Eds.), *IEEE International Conference on Big Data*. IEEE, Washington, USA, pp. 609–614. <https://doi.org/10.1109/BigData.2014.7004282>

Donoghue, D.N.M., Watt, P.J., Cox, N.J., Wilson, J., 2007. Remote sensing of species mixtures in conifer plantations using LiDAR height and intensity data. *Remote Sens. Environ.* 110, 509–522. <https://doi.org/10.1016/j.rse.2007.02.032>

dos Santos Ferreira, A., Matte Freitas, D., Gonçalves da Silva, G., Pistori, H., Theophilo Folhes, M., 2017. Weed detection in soybean crops using ConvNets. *Comput. Electron. Agric.* 143, 314–324. <https://doi.org/10.1016/j.compag.2017.10.027>

Doyle, J., 1979. A truth maintenance system. *Artif. Intell.* 12, 231–272. [https://doi.org/10.1016/0004-3702\(79\)90008-0](https://doi.org/10.1016/0004-3702(79)90008-0)

Dwibedi, D., Misra, I., Hebert, M., 2017. Cut, paste and learn: surprisingly easy synthesis for instance detection. In: 2017 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp. 1310–1319. <https://doi.org/10.1109/ICCV.2017.146>

- Erdiwansyah, Mahidin, Mamat, R., Sani, M.S.M., Khoerunnisa, F., Kadarohman, A., 2019. Target and demand for renewable energy across 10 ASEAN countries by 2040. *Electr. J.* 32, 106670. <https://doi.org/10.1016/j.tej.2019.106670>
- Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., Vasilakoglou, I., 2020. Towards weeds identification assistance through transfer learning. *Comput. Electron. Agric.* 171, 105306. <https://doi.org/10.1016/j.compag.2020.105306>
- Evangelistella, C., Valentini, A., Ludovisi, R., Firrincieli, A., Fabbrini, F., Scalabrin, S., Cattonaro, F., Morgante, M., Mugnozza, G.S., Keurentjes, J.J.B., Harfouche, A.L., 2017. De novo assembly, functional annotation, and analysis of the giant reed (*Arundo donax* L.) leaf transcriptome provide tools for the development of a biofuel feedstock. *Biotechnol. Biofuels* 10, 138. <https://doi.org/10.1186/s13068-017-0828-7>
- Fabbrini, F., Ludovisi, R., Alasia, O., Flexas, J., Douthe, C., Ribas Carbó, M., Robson, P., Taylor, G., Scarascia-Mugnozza, G., Keurentjes, J.J.B., Harfouche, A.L., 2019. Characterization of phenology, physiology, morphology and biomass traits across a broad Euro-Mediterranean ecotypic panel of the lignocellulosic feedstock *Arundo donax*. *GCB Bioenergy* 11, 152–170. <https://doi.org/10.1111/gcbb.12555>
- Fahlgren, N., Feldman, M., Gehan, M.A., Wilson, M.S., Shyu, C., Bryant, D.W., Hill, S.T., McEntee, C.J., Warnasooriya, S.N., Kumar, I., Ficor, T., Turnipseed, S., Gilbert, K.B., Brutnell, T.P., Carrington, J.C., Mockler, T.C., Baxter, I., 2015. A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in *Setaria*. *Mol. Plant* 8, 1520–1535. <https://doi.org/10.1016/j.molp.2015.06.005>
- Feigenbaum, E., 1992. A personal view of expert systems: Looking back and looking ahead. *Expert Syst. Appl.* 5, 193–201. [https://doi.org/10.1016/0957-4174\(92\)90004-C](https://doi.org/10.1016/0957-4174(92)90004-C)
- Feigenbaum, E., Shrobe, H., 1993. The Japanese national fifth generation project: introduction, survey, and evaluation. *Futur. Gener. Comput. Syst.* 9, 105–117. [https://doi.org/10.1016/0167-739X\(93\)90003-8](https://doi.org/10.1016/0167-739X(93)90003-8)
- Feng, D., Liu, Y., Mazloomi, M.S., Limaye, A., Turner, M., Evans, P.D., 2020. A virtual

- reality system to augment teaching of wood structure and protection. *Int. Wood Prod. J.* 11, 46–56. <https://doi.org/10.1080/20426445.2020.1737773>
- Feng, L., Chen, S., Zhang, C., Zhang, Y., He, Y., 2021. A comprehensive review on recent applications of unmanned aerial vehicle remote sensing with various sensors for high-throughput plant phenotyping. *Comput. Electron. Agric.* 182, 106033. <https://doi.org/10.1016/j.compag.2021.106033>
- Fenu, G., Mallocci, F.M., 2021. DiaMOS plant: a dataset for diagnosis and monitoring plant disease. *Agronomy* 11, 2107. <https://doi.org/10.3390/agronomy11112107>
- Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181.
- Fiorani, F., Schurr, U., 2013. Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* 64, 267–291. <https://doi.org/10.1146/annurev-arplant-050312-120137>
- Flores-Arredondo, J.H., Assad-Kottner, C., 2015. Virtual reality: a look into the past to fuel the future. *Bull. R. Coll. Surg. Engl.* 97, 424–426. <https://doi.org/10.1308/rcsbull.2015.424>
- Fonti, P., von Arx, G., García-González, I., Eilmann, B., Sass-Klaassen, U., Gärtner, H., Eckstein, D., 2010. Studying global change through investigation of the plastic responses of xylem anatomy in tree rings. *New Phytol.* 185, 42–53. <https://doi.org/10.1111/j.1469-8137.2009.03030.x>
- Francesconi, S., Harfouche, A.L., Maesano, M., Balestra, G.M., 2021. UAV-based thermal, RGB imaging and gene expression analysis allowed detection of *Fusarium* head blight and gave new insights into the physiological responses to the disease in durum wheat. *Front. Plant Sci.* 12, 551. <https://doi.org/10.3389/fpls.2021.628575>
- Fukushima, K., 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. <https://doi.org/10.1007/BF00344251>
- Fukushima, K., 2014. Modeling vision with the Neocognitron. In: *Springer Handbook of Bio-/Neuroinformatics*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 765–

782. https://doi.org/10.1007/978-3-642-30574-0_44

Furbank, R.T., Tester, M., 2011. Phenomics – technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16, 635–644. <https://doi.org/10.1016/j.tplants.2011.09.005>

Gaggion, N., Ariel, F., Daric, V., Lambert, É., Legendre, S., Roulé, T., Camoirano, A., Milone, D.H., Crespi, M., Blein, T., Ferrante, E., 2021. ChronoRoot: high-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture. *Gigascience* 10, giab052. <https://doi.org/10.1093/gigascience/giab052>

Gao, K., Mei, G., Piccialli, F., Cuomo, S., Tu, J., Huo, Z., 2020. Julia language in machine learning: algorithms, applications, and open issues. *Comput. Sci. Rev.* 37, 100254. <https://doi.org/10.1016/j.cosrev.2020.100254>

Geetharamani, G., Pandian, A., 2019. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput. Electr. Eng.* 76, 323–338. <https://doi.org/10.1016/j.compeleceng.2019.04.011>

Gehan, M.A., Fahlgren, N., Abbasi, A., Berry, J.C., Callen, S.T., Chavez, L., Doust, A.N., Feldman, M.J., Gilbert, K.B., Hodge, J.G., Hoyer, J.S., Lin, A., Liu, S., Lizárraga, C., Lorence, A., Miller, M., Platon, E., Tessman, M., Sax, T., 2017. PlantCV v2: image analysis software for high-throughput plant phenotyping. *PeerJ* 5, e4088. <https://doi.org/10.7717/peerj.4088>

Gené-Mola, J., Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiella, R., Escolà, A., Rosell-Polo, J.R., 2020a. Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Comput. Electron. Agric.* 168, 105121. <https://doi.org/10.1016/j.compag.2019.105121>

Gené-Mola, J., Gregorio, E., Auat Cheein, F., Guevara, J., Llorens, J., Sanz-Cortiella, R., Escolà, A., Rosell-Polo, J.R., 2020b. LFuji-air dataset: annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions. *Data Br.* 29, 105248. <https://doi.org/10.1016/j.dib.2020.105248>

Ghosal, S., Blystone, D., Singh, A.A.K., Ganapathysubramanian, B., Singh, A.A.K.,

- Sarkar, S., 2018. An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci.* 115, 4613–4618. <https://doi.org/10.1073/pnas.1716999115>
- Goluguri, N.V.R.R., Suganya Devi, K., Vadaparthi, N., 2021. Image classifiers and image deep learning classifiers evolved in detection of *Oryza sativa* diseases: survey. *Artif. Intell. Rev.* 54, 359–396. <https://doi.org/10.1007/s10462-020-09849-y>
- Gomez Selvaraj, M., Vergara, A., Montenegro, F., Alonso Ruiz, H., Safari, N., Raymaekers, D., Ocimati, W., Ntamwira, J., Tits, L., Omondi, A.B., Blomme, G., 2020. Detection of banana plants and their major diseases through aerial images and machine learning methods: a case study in DR Congo and Republic of Benin. *ISPRS J. Photogramm. Remote Sens.* 169, 110–124. <https://doi.org/10.1016/j.isprsjprs.2020.08.025>
- González, S., García, S., Del Ser, J., Rokach, L., Herrera, F., 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* 64, 205–237. <https://doi.org/10.1016/j.inffus.2020.07.007>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Gopal, G., J., A.P.P., G., G., J., A.P.P., Gopal, G., J., A.P.P., G., G., J., A.P.P., 2019. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput. Electr. Eng.* 76, 323–338. <https://doi.org/10.1016/j.compeleceng.2019.04.011>
- Gopalakrishnan, K., Khaitan, S.K., Choudhary, A., Agrawal, A., 2017. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. *Constr. Build. Mater.* 157, 322–330. <https://doi.org/10.1016/j.conbuildmat.2017.09.110>
- Grabowski, P., Rappsilber, J., 2019. A primer on data analytics in functional genomics: how to move from data to insight? *Trends Biochem. Sci.* 44, 21–32.

<https://doi.org/10.1016/j.tibs.2018.10.010>

Grant, J., Keong Chen, H., 2021. Using wood forensic science to deter corruption and illegality in the timber trade, Targeting Natural Resource Corruption.

Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T., 2022. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55. <https://doi.org/10.1038/s41580-021-00407-0>

Großkinsky, D.K., Svensgaard, J., Christensen, S., Roitsch, T., 2015. Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J. Exp. Bot.* 66, 5429–5440. <https://doi.org/10.1093/jxb/erv345>

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T., 2018. Recent advances in convolutional neural networks. *Pattern Recognit.* 77, 354–377. <https://doi.org/10.1016/j.patcog.2017.10.013>

Guidotti, R., 2021. Evaluating local explanation methods on ground truth. *Artif. Intell.* 291, 103428. <https://doi.org/10.1016/j.artint.2020.103428>

Gupta, V., Choudhary, K., Tavazza, F., Campbell, C., Liao, W., Choudhary, A., Agrawal, A., 2021. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat. Commun.* 12, 6595. <https://doi.org/10.1038/s41467-021-26921-5>

Gutiérrez, S., Diago, M.P., Fernández-Novales, J., Tardaguila, J., 2018. Vineyard water status assessment using on-the-go thermal imaging and machine learning. *PLoS One* 13, e0192037. <https://doi.org/10.1371/journal.pone.0192037>

Haibo He, Yang Bai, Garcia, E.A., Shutao Li, 2008. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. IEEE, pp. 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>

Hamner, B., Frasco, M., LeDell, E., 2018. Metrics: Evaluation metrics for machine learning version 0.1.4 [WWW Document]. URL <https://rdr.io/cran/Metrics/> (accessed 5.17.20).

Hani, N., Roy, P., Isler, V., 2020. MinneApple: a benchmark dataset for apple detection

- and segmentation. *IEEE Robot. Autom. Lett.* 5, 852–858. <https://doi.org/10.1109/LRA.2020.2965061>
- Hansen, L.K., Salamon, P., 1990. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 993–1001. <https://doi.org/10.1109/34.58871>
- Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. *IEEE Trans. Syst. Man. Cybern. SMC-3*, 610–621. <https://doi.org/10.1109/TSMC.1973.4309314>
- Harfouche, A.L., Jacobson, D.A., Kainer, D., Romero, J.C., Harfouche, A.H., Scarascia Mugnozza, G., Moshelion, M., Tuskan, G.A., Keurentjes, J.J.B., Altman, A., 2019. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends Biotechnol.* 37, 1217–1235. <https://doi.org/10.1016/j.tibtech.2019.05.007>
- Harfouche, A.L., Nakhle, F., 2020. Creating bioethics distance learning through virtual reality. *Trends Biotechnol.* 38, 1187–1192. <https://doi.org/10.1016/j.tibtech.2020.05.005>
- Harfouche, A.L., Nakhle, F., 2023. Artificial intelligence, blockchain, and extended reality: emerging digital technologies to turn the tide on illegal logging and illegal wood trade. *Ann. Silv. Res.* <https://doi.org/10.12899/asr-2435> (in press).
- Harfouche, A.L., Nakhle, F., Harfouche, A.H., Sardella, O.G., Dart, E., Jacobson, D., 2023. A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey. *Trends Plant Sci.* 28, 154–184 <https://doi.org/10.1016/j.tplants.2022.08.021>
- Harfouche, A.L., Petousi, V., Meilan, R., Sweet, J., Twardowski, T., Altman, A., 2021. Promoting ethically responsible use of agricultural biotechnology. *Trends Plant Sci.* 26, 546–559 <https://doi.org/10.1016/j.tplants.2020.12.015>
- Harfouche, A.L., Romagnoli, M., Corona, P., 2022. Under the “publish or perish” mantra and the race for grants, insights to catalyze research into wood science. *Ann. Silv. Res.* 47, 116–118. <https://doi.org/10.12899/asr-2407>
- Hasebe, T., 2021. Knowledge-embedded message-passing neural networks: improving molecular property prediction with human knowledge. *ACS Omega* 6,

- 27955–27967. <https://doi.org/10.1021/acsomega.1c03839>
- Hassabis, D., 2017. Artificial intelligence: chess match of the century. *Nature* 544, 413–414. <https://doi.org/10.1038/544413a>
- Haug, S., Ostermann, J., 2015. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In: *Computer Vision - ECCV 2014 Workshops*. pp. 105–116. https://doi.org/10.1007/978-3-319-16220-1_8
- Haworth, M., Marino, G., Riggi, E., Avola, G., Brunetti, C., Scordia, D., Testa, G., Thiago Gaudio Gomes, M., Loreto, F., Luciano Cosentino, S., Centritto, M., 2019. The effect of summer drought on the yield of *Arundo donax* is reduced by the retention of photosynthetic capacity and leaf growth later in the growing season. *Ann. Bot.* 124, 567–579. <https://doi.org/10.1093/aob/mcy223>
- Hollaus, M., Wagner, W., Eberhöfer, C., Karel, W., 2006. Accuracy of large-scale canopy heights derived from LiDAR data under operational constraints in a complex alpine environment. *ISPRS J. Photogramm. Remote Sens.* 60, 323–338. <https://doi.org/10.1016/j.isprsjprs.2006.05.002>
- Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: the next challenge. *Nat. Rev. Genet.* 11, 855–866. <https://doi.org/10.1038/nrg2897>
- Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2018. Densely connected convolutional networks.
- Huang, Y., Ren, Z., Li, D., Liu, X., 2020. Phenotypic techniques and applications in fruit trees: a review. *Plant Methods* 16, 107. <https://doi.org/10.1186/s13007-020-00649-7>
- Hughes, D.P., Salathé, M., 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *CoRR* abs/1511.0.
- Hwang, S.-W., Sugiyama, J., 2021. Computer vision-based wood identification and its expansion and contribution potentials in wood science: a review. *Plant Methods* 17, 47. <https://doi.org/10.1186/s13007-021-00746-1>
- Insilico Ltd, 2018. InductVR [WWW Document]. URL <https://inductvr.github.io/> (accessed 11.15.22).

- Jager, P., Moore, G., Calpin, P., Durmishi, X., Salgarella, I., Menage, L., Kita, Y., Wang, Y., Kim, D.W., Blackshaw, S., Schultz, S.R., Brickley, S., Shimogori, T., Delogu, A., 2021. Dual midbrain and forebrain origins of thalamic inhibitory interneurons. *Elife* 10, e59272. <https://doi.org/10.7554/eLife.59272>
- Jepkoech, J., Mugo, D.M., Kenduiywo, B.K., Too, E.C., 2021. Arabica coffee leaf images dataset for coffee leaf disease detection and classification. *Data Br.* 36, 107142. <https://doi.org/10.1016/j.dib.2021.107142>
- Jiang, Y., Li, C., 2020. Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics* 2020, 1–22. <https://doi.org/10.34133/2020/4152816>
- Jimenez-Berni, J.A., Deery, D.M., Rozas-Larraondo, P., Condon, A. (Tony) G., Rebetzke, G.J., James, R.A., Bovill, W.D., Furbank, R.T., Sirault, X.R.R., 2018. High throughput determination of plant height, ground cover, and above-ground biomass in Wheat with LiDAR. *Front. Plant Sci.* 9, 237. <https://doi.org/10.3389/fpls.2018.00237>
- Jin, L., Yang, J., Kuang, K., Ni, B., Gao, Y., Sun, Y., Gao, P., Ma, W., Tan, M., Kang, H., Chen, J., Li, M., 2020. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. *EBioMedicine* 62, 103106. <https://doi.org/10.1016/j.ebiom.2020.103106>
- Jones, N., 2018. Simulated labs are booming. *Nature* 562, S5–S7. <https://doi.org/10.1038/d41586-018-06831-1>
- Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, I., Mancuso, J., Jungmann, F., Steinborn, M.-M., Saleh, A., Makowski, M., Rueckert, D., Braren, R., 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* 3, 473–484. <https://doi.org/10.1038/s42256-021-00337-8>
- Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2, 305–311. <https://doi.org/10.1038/s42256-020-0186-1>
- Kaufmane, E., Sudars, K., Namatēvs, I., Kalniņa, I., Judvaitis, J., Balašs, R., Strautiņa,

- S., 2022. QuinceSet: dataset of annotated Japanese quince images for object detection. *Data Br.* 42, 108332. <https://doi.org/10.1016/j.dib.2022.108332>
- Kaul, V., Enslin, S., Gross, S.A., 2020. History of artificial intelligence in medicine. *Gastrointest. Endosc.* 92, 807–812. <https://doi.org/10.1016/j.gie.2020.06.040>
- Kerby, T.A., Plant, R.E., Johnson-Hake, S., Horrocks, R.D., 2013. Environmental and cultivar effects on height-to-node ratio and growth rate in Acala cotton. *J. Prod. Agric.* 11, 427. <https://doi.org/10.2134/jpa1998.0420>
- Kernighan, B., 1972. A tutorial introduction to the language B (Bell Laboratories). Murray Hill, New Jersey.
- Kernighan, B., 1975. Programming in C - A tutorial (Bell Laboratories). Murray Hill, New Jersey.
- Kernighan, B., Ritchie, D., 1978. The C programming language. Prentice-Hall, Inc., USA.
- Kestur, R., Meduri, A., Narasipura, O., 2019. MangoNet: a deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. *Eng. Appl. Artif. Intell.* 77, 59–69. <https://doi.org/10.1016/j.engappai.2018.09.011>
- Khalifa, N.E., Loey, M., Mirjalili, S., 2022. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artif. Intell. Rev.* 55, 2351–2377. <https://doi.org/10.1007/s10462-021-10066-4>
- Khanna, R., Schmid, L., Walter, A., Nieto, J., Siegwart, R., Liebisch, F., 2019. A spatio temporal spectral framework for plant stress phenotyping. *Plant Methods* 15, 13. <https://doi.org/10.1186/s13007-019-0398-8>
- Kicherer, A., Herzog, K., Pflanz, M., Wieland, M., Rüger, P., Kecke, S., Kuhlmann, H., Töpfer, R., 2015. An automated field phenotyping pipeline for application in grapevine research. *Sensors* 15, 4823–4836. <https://doi.org/10.3390/s150304823>
- Kirchgessner, N., Liebisch, F., Yu, K., Pfeifer, J., Friedli, M., Hund, A., Walter, A., 2017. The ETH field phenotyping platform FIP: A cable-suspended multi-sensor system. *Funct. Plant Biol.* 44, 168. <https://doi.org/10.1071/FP16165>

- Knapp, N., Fischer, R., Huth, A., 2018. Linking LiDAR and forest modeling to assess biomass estimation across scales and disturbance states. *Remote Sens. Environ.* 205, 199–209. <https://doi.org/10.1016/j.rse.2017.11.018>
- Knecht, A.C., Campbell, M.T., Caprez, A., Swanson, D.R., Walia, H., 2016. Image Harvest: an open-source platform for high-throughput plant image processing and analysis. *J. Exp. Bot.* 67, 3587–3599. <https://doi.org/10.1093/jxb/erw176>
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. <https://doi.org/10.1145/3065386>
- Kusrini, K., Suputa, S., Setyanto, A., Agastya, I.M.A., Priantoro, H., Chandramouli, K., Izquierdo, E., 2020. Data augmentation for automated pest classification in Mango farms. *Comput. Electron. Agric.* 179, 105842. <https://doi.org/10.1016/j.compag.2020.105842>
- Kvinge, H., Kirby, M., Peterson, C., Eitel, C., Clapp, T., 2019. A walk through spectral bands: using virtual reality to better visualize hyperspectral data. *Adv. Intell. Syst. Comput.* 976, 160–165.
- Lac, L., Keresztes, B., Louargant, M., Donias, M., Da Costa, J.-P., 2022. An annotated image dataset of vegetable crops at an early stage of growth for proximal sensing applications. *Data Br.* 42, 108035. <https://doi.org/10.1016/j.dib.2022.108035>
- Laurent, A., Pelzer, E., Loyce, C., Makowski, D., 2015. Ranking yields of energy crops: A meta-analysis using direct and indirect comparisons. *Renew. Sustain. Energy Rev.* 46, 41–50. <https://doi.org/10.1016/j.rser.2015.02.023>
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444. <https://doi.org/10.1038/nature14539>
- Legeth, O., Rodhe, J., Lang, S., Dhapola, P., Wallergård, M., Soneji, S., 2021. CellexalVR: a virtual reality platform to visualize and analyze single-cell omics data. *iScience* 24, 103251. <https://doi.org/10.1016/j.isci.2021.103251>
- Lepoittevin, C., Rousseau, J.-P., Guillemin, A., Gauvrit, C., Besson, F., Hubert, F., da Silva Perez, D., Harvengt, L., Plomion, C., 2011. Genetic parameters of growth, straightness and wood chemistry traits in *Pinus pinaster*. *Ann. For. Sci.* 68, 873–

884. <https://doi.org/10.1007/s13595-011-0084-0>
- Lever, J., Krzywinski, M., Altman, N., 2016. Regularization. *Nat. Methods* 13, 803–804. <https://doi.org/10.1038/nmeth.4014>
- Levine, A.B., Schlosser, C., Grewal, J., Coope, R., Jones, S.J.M., Yip, S., 2019. Rise of the machines: advances in deep learning for cancer diagnosis. *Trends in Cancer* 5, 157–169. <https://doi.org/10.1016/j.trecan.2019.02.002>
- Lewandowski, I., 2015. Securing a sustainable biomass supply in a growing bioeconomy. *Glob. Food Sec.* 6, 34–42. <https://doi.org/10.1016/j.gfs.2015.10.001>
- Li, B., Chen, P., Liu, H., Guo, W., Cao, X., Du, J., Zhao, C., Zhang, J., 2021. Random sketch learning for deep neural networks in edge computing. *Nat. Comput. Sci.* 1, 221–228. <https://doi.org/10.1038/s43588-021-00039-6>
- Li, J., Peng, Jinghao, Jiang, X., Rea, A.C., Peng, Jiajie, Hu, J., 2021. DeepLearnMOR: a deep-learning framework for fluorescence image-based classification of organelle morphology. *Plant Physiol.* 186, 1786–1799. <https://doi.org/10.1093/plphys/kiab223>
- Li, Y., Li, M., Li, C., Liu, Z., 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Sci. Rep.* 10, 9952. <https://doi.org/10.1038/s41598-020-67024-3>
- Lighthill, J., 1973. Artificial intelligence: a general survey. In: *Artificial Intelligence: A Paper Symposium*. Science Research Council, pp. 1–21.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: *European Conference on Computer Vision*. pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
- Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: a review of machine learning interpretability methods. *Entropy* 23, 18. <https://doi.org/10.3390/e23010018>
- Lipton, Z.C., 2018. The mythos of model interpretability. *Queue* 16, 31–57. <https://doi.org/10.1145/3236386.3241340>

- Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., Xu, D., 2019. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* 10, 1091. <https://doi.org/10.3389/fgene.2019.01091>
- Lobell, D.B., Gourdji, S.M., 2012. The influence of climate change on global crop productivity. *Plant Physiol.* 160, 1686–1697. <https://doi.org/10.1104/pp.112.208298>
- Lobet, G., Draye, X., Périlleux, C., 2013. An online database for plant image analysis software tools. *Plant Methods* 9, 38. <https://doi.org/10.1186/1746-4811-9-38>
- Ludovisi, R., Tauro, F., Salvati, R., Khoury, S., Mugnozza Scarascia, G., Harfouche, A.L., 2017. UAV-based thermal imaging for high-throughput field phenotyping of black poplar response to drought. *Front. Plant Sci.* 8, 1681. <https://doi.org/10.3389/fpls.2017.01681>
- Lumley, T., 2020. leaps: Regression Subset Selection version 3.1 [WWW Document]. URL <https://rdr.io/cran/leaps/> (accessed 5.17.20).
- Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U. V, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 4768–4777.
- Madec, S., Baret, F., De Solan, B., Thomas, S., Dutartre, D., Jezequel, S., Hemmerlé, M., Colombeau, G., Comar, A., 2017. High-throughput phenotyping of plant height: Comparing unmanned aerial vehicles and ground LiDAR estimates. *Front. Plant Sci.* 8, 2002. <https://doi.org/10.3389/fpls.2017.02002>
- Maesano, M., Khoury, S., Nakhle, F., Firrincieli, A., Gay, A., Tauro, F., Harfouche, A.L., 2020. UAV-based LiDAR for high-throughput determination of plant height and above-ground biomass of the bioenergy grass *Arundo donax*. *Remote Sens.* 12, 3464. <https://doi.org/10.3390/rs12203464>
- Maimaitijiang, M., Ghulam, A., Sidike, P., Hartling, S., Maimaitiyiming, M., Peterson, K., Shavers, E., Fishman, J., Peterson, J., Kadam, S., Burken, J., Fritschi, F., 2017. Unmanned aerial system (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *ISPRS J. Photogramm.*

- Remote Sens. 134, 43–58. <https://doi.org/10.1016/j.isprsjprs.2017.10.011>
- Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens. Environ. 237, 111599. <https://doi.org/10.1016/j.rse.2019.111599>
- Mäkelä, M., Altgen, M., Belt, T., Rautkari, L., 2021. Hyperspectral imaging and chemometrics reveal wood acetylation on different spatial scales. J. Mater. Sci. 56, 5053–5066. <https://doi.org/10.1007/s10853-020-05597-0>
- Mann, A., 2020. Core concept: nascent exascale supercomputers offer promise, present challenges. Proc. Natl. Acad. Sci. 117, 22623–22625. <https://doi.org/10.1073/pnas.2015968117>
- Manso, G.L., Knidel, H., Krohling, R.A., Ventura, J.A., 2019. A smartphone application to detection and classification of coffee leaf miner and coffee leaf rust. arXiv.
- Martinez-Gonzalez, P., Oprea, S., Garcia-Garcia, A., Jover-Alvarez, A., Orts-Escolano, S., Garcia-Rodriguez, J., 2020. UnrealROX: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. Virtual Real. 24, 271–288. <https://doi.org/10.1007/s10055-019-00399-5>
- Marx, V., 2021. Biology begins to tangle with quantum computing. Nat. Methods 18, 715–719. <https://doi.org/10.1038/s41592-021-01199-z>
- Matthews, D., 2018. Supercharge your data wrangling with a graphics card. Nature 562, 151–152. <https://doi.org/10.1038/d41586-018-06870-8>
- McCarthy, J., 1960. Recursive functions of symbolic expressions and their computation by machine, Part I. Commun. ACM 3, 184–195. <https://doi.org/10.1145/367177.367199>
- McCarthy, J., 1968. Programs with common sense. In: Semantic Information Processing. MIT Press, pp. 403–418.
- McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133. <https://doi.org/10.1007/BF02478259>

- McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A. y, 2017. Communication-efficient learning of deep networks from decentralized data. In: Singh, A., Zhu, J. (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research. PMLR, pp. 1273–1282.
- Mehta, P., Bukov, M., Wang, C.-H., Day, A.G.R., Richardson, C., Fisher, C.K., Schwab, D.J., 2019. A high-bias, low-variance introduction to Machine Learning for physicists. Phys. Rep. 810, 1–124. <https://doi.org/10.1016/j.physrep.2019.03.001>
- Miceli, M., Schuessler, M., Yang, T., 2020. Between subjectivity and imposition: power dynamics in data annotation for computer vision. Proc. ACM Human-Computer Interact. 4, 1–25. <https://doi.org/10.1145/3415186>
- Mignoni, M.E., Honorato, A., Kunst, R., Righi, R., Massuquetti, A., 2022. Soybean images dataset for caterpillar and *Diabrotica speciosa* pest detection and classification. Data Br. 40, 107756. <https://doi.org/10.1016/j.dib.2021.107756>
- Milgram, P., Kishino, F., 1994. Taxonomy of mixed reality visual displays. In: IEICE Transactions on Information and Systems. pp. 1321–1329.
- Minervini, M., Fischbach, A., Scharr, H., Tsaftaris, S.A., 2016. Finely-grained annotated datasets for image-based plant phenotyping. Pattern Recognit. Lett. 81, 80–89. <https://doi.org/10.1016/j.patrec.2015.10.013>
- Ming-Kuei Hu, 1962. Visual pattern recognition by moment invariants. IEEE Trans. Inf. Theory 8, 179–187. <https://doi.org/10.1109/TIT.1962.1057692>
- Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R., Melgani, F., 2019. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. Gigascience 8, giy153. <https://doi.org/10.1093/gigascience/giy153>
- Molnar, C., 2019. Interpretable machine learning. A guide for making black box models explainable. Leanpub.
- Moor, J., 2006. The Dartmouth College artificial intelligence conference: the next fifty years. AI Mag. 27, 87. <https://doi.org/10.1609/aimag.v27i4.1911>

- Moose, S.P., Mumm, R.H., 2008. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.* 147, 969–977. <https://doi.org/10.1104/pp.108.118232>
- Mostafa, S., Mondal, D., Beck, M., Bidinosti, C., Henry, C., Stavness, I., 2021. Visualizing feature maps for model selection in convolutional neural networks. In: Dickinson, S. (Ed.), *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. pp. 1362–1371.
- Muggleton, S., 1991. Inductive logic programming. *New Gener. Comput.* 8, 295–318. <https://doi.org/10.1007/BF03037089>
- Muggleton, S., Marginean, F., 2000. Logic-based machine learning. In: *Logic-Based Artificial Intelligence*. Springer US, Boston, MA, pp. 315–330. https://doi.org/10.1007/978-1-4615-1567-8_14
- Mwebaze, E., Gebru, T., Frome, A., Nsumba, S., Tusubira, J., 2019. iCassava 2019 fine-grained visual categorization challenge. *arXiv*.
- Nagasubramanian, K., Jones, S., Singh, A.A.K.A.A.K., Sarkar, S., Singh, A.A.K.A.A.K., Ganapathysubramanian, B., 2019. Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods* 15, 98. <https://doi.org/10.1186/s13007-019-0479-8>
- Nagasubramanian, K., Singh, A.K., Singh, A., Sarkar, S., Ganapathysubramanian, B., 2020. Usefulness of interpretability methods to explain deep learning based plant stress phenotyping. *arXiv*.
- Nakatumba-Nabende, J., Akera, B., Tusubira, J.F., Nsumba, S., Mwebaze, E., 2020. A dataset of necrotized cassava root cross-section images. *Data Br.* 32, 106170. <https://doi.org/10.1016/j.dib.2020.106170>
- Nakhle, F., Harfouche, A.L., 2021. Ready, steady, go AI: a practical tutorial on fundamentals of artificial intelligence and its applications in phenomics image analysis. *Patterns* 2, 100323. <https://doi.org/10.1016/j.patter.2021.100323>
- Nakhle, F., Harfouche, A.L., 2023. Extended reality gives digital agricultural biotechnology a new dimension. *Trends Biotechnol.* 41, 1–5. <https://doi.org/10.1016/j.tibtech.2022.09.005>

- Nature Plants Editorial, 2021a. Next-generation crop engineering. *Nat. Plants* 7, 241–241. <https://doi.org/10.1038/s41477-021-00890-5>
- Nature Plants Editorial, 2021b. Resilience for sustainability. *Nat. Plants* 7, 101–101. <https://doi.org/10.1038/s41477-021-00871-8>
- Navarro, P.J., Miller, L., Díaz-Galián, M.V., Gila-Navarro, A., Aguila, D.J., Egea-Cortines, M., 2022. A novel ground truth multispectral image dataset with weight, anthocyanins, and Brix index measures of grape berries tested for its utility in machine learning pipelines. *Gigascience* 11, giac052. <https://doi.org/10.1093/gigascience/giac052>
- Nepal, R., Paija, N., 2019. Energy security, electricity, population and economic growth: The case of a developing South Asian resource-rich economy. *Energy Policy* 132, 771–781. <https://doi.org/10.1016/j.enpol.2019.05.054>
- Nevavuori, P., Narra, N., Lipping, T., 2019. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* 163, 104859. <https://doi.org/10.1016/j.compag.2019.104859>
- Newell, A., Shaw, J.C., 1957. Programming the logic theory machine. In: *Western Joint Computer Conference: Techniques for Reliability*. ACM Press, New York, New York, USA, pp. 230–240. <https://doi.org/10.1145/1455567.1455606>
- Newell, A., Simon, H., 1956. The logic theory machine--A complex information processing system. *IEEE Trans. Inf. Theory* 2, 61–79. <https://doi.org/10.1109/TIT.1956.1056797>
- Nunes, M., Ewers, R., Turner, E., Coomes, D., 2017. Mapping aboveground carbon in oil palm plantations using LiDAR: A comparison of tree-centric versus area-based approaches. *Remote Sens.* 9, 816. <https://doi.org/10.3390/rs9080816>
- Olazaran, M., 1993. A sociological history of the neural network controversy. In: Yovits, M.C.B.T.-A. in C. (Ed.), *Advances in Computers*. Elsevier, pp. 335–425. [https://doi.org/https://doi.org/10.1016/S0065-2458\(08\)60408-8](https://doi.org/https://doi.org/10.1016/S0065-2458(08)60408-8)
- Olsen, A., Konovalov, D.A., Philippa, B., Ridd, P., Wood, J.C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., Calvert, B., Azghadi, M.R., White, R.D., 2019. DeepWeeds: a multiclass weed species image dataset for deep learning.

- Sci. Rep. 9, 2058. <https://doi.org/10.1038/s41598-018-38343-3>
- Orhobor, O.I., Alexandrov, N.N., King, R.D., 2020. Predicting rice phenotypes with meta and multi-target learning. *Mach. Learn.* 109, 2195–2212. <https://doi.org/10.1007/s10994-020-05881-9>
- Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P.L., Khamis, S., Dou, M., Tankovich, V., Loop, C., Cai, Q., Chou, P.A., Mennicken, S., Valentin, J., Pradeep, V., Wang, S., Kang, S.B., Kohli, P., Lutchyn, Y., Keskin, C., Izadi, S., 2016. Holoportation. In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, pp. 741–754. <https://doi.org/10.1145/2984511.2984517>
- Otim-Napr, G.W., Alicai, T., Thresh, J.M., 2001. Changes in the incidence and severity of Cassava mosaic virus disease, varietal diversity and cassava production in Uganda. *Ann. Appl. Biol.* 138, 313–327. <https://doi.org/10.1111/j.1744-7348.2001.tb00116.x>
- Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I.N., Chaves, I., Coppens, F., Cornut, G., Costa, B. V., Ćwiek-Kupczyńska, H., Driesbeke, B., Finkers, R., Gruden, K., Junker, A., King, G.J., Krajewski, P., Lange, M., Laporte, M., Michotey, C., Oppermann, M., Ostler, R., Poorter, H., Ramírez-Gonzalez, R., Ramšak, Ž., Reif, J.C., Rocca-Serra, P., Sansone, S., Scholz, U., Tardieu, F., Uauy, C., Usadel, B., Visser, R.G.F., Weise, S., Kersey, P.J., Miguel, C.M., Adam-Blondon, A., Pommier, C., 2020. Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytol.* 227, 260–273. <https://doi.org/10.1111/nph.16544>
- Parraga-Alava, J., Cusme, K., Loo, A., Santander, E., 2019. RoCoLe: a robust coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition. *Data Br.* 25, 104414. <https://doi.org/10.1016/j.dib.2019.104414>
- Patel, B.N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarrappallil, J., Mariano, A.J., Riley, G., Seekins, J., Shen, L., Zucker, E., Lungren, M.P., 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digit. Med.* 2, 111. <https://doi.org/10.1038/s41746-019-0189-7>

- Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning.
- Perkel, J.M., 2019. Julia: come for the syntax, stay for the speed. *Nature* 572, 141–142. <https://doi.org/10.1038/d41586-019-02310-3>
- Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197. <https://doi.org/10.1890/07-2153.1>
- Picos, J., Bastos, G., Míguez, D., Alonso, L., Armesto, J., 2020. Individual tree detection in a *Eucalyptus* plantation using unmanned aerial vehicle (UAV)-LiDAR. *Remote Sens.* 12, 885. <https://doi.org/10.3390/rs12050885>
- Pigliucci, M., 2010. Genotype–phenotype mapping and the end of the “genes as blueprint” metaphor. *Philos. Trans. R. Soc. B Biol. Sci.* 365, 557–566. <https://doi.org/10.1098/rstb.2009.0241>
- Pirch, S., Müller, F., Iofinova, E., Pazmandi, J., Hütter, C.V.R., Chiettoni, M., Sin, C., Boztug, K., Podkosova, I., Kaufmann, H., Menche, J., 2021. The VRNetzer platform enables interactive network analysis in Virtual Reality. *Nat. Commun.* 12, 2432. <https://doi.org/10.1038/s41467-021-22570-w>
- Plumb, G., Molitor, D., Talwalkar, A.S., 2018. Model agnostic supervised local explanations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Pompeiano, A., Remorini, D., Vita, F., Guglielminetti, L., Miele, S., Morini, S., 2017. Growth and physiological response of *Arundo donax* L. to controlled drought stress and recovery. *Plant Biosyst.* 151, 906–914. <https://doi.org/10.1080/11263504.2016.1249427>
- Poorter, H., Fiorani, F., Pieruschka, R., Wojciechowski, T., Putten, W.H., Kleyer, M., Schurr, U., Postma, J., 2016. Pampered inside, pestered outside? differences and similarities between plants growing in controlled conditions and in the field. *New Phytol.* 212, 838–855. <https://doi.org/10.1111/nph.14243>

- Pound, M.P., Atkinson, J.A., Townsend, A.J., Wilson, M.H., Griffiths, M., Jackson, A.S., Bulat, A., Tzimiropoulos, G., Wells, D.M., Murchie, E.H., Pridmore, T.P., French, A.P., 2017a. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience* 6, gix083. <https://doi.org/10.1093/gigascience/gix083>
- Pound, M.P., Atkinson, J.A., Wells, D.M., Pridmore, T.P., French, A.P., 2017b. Deep learning for multi-task plant phenotyping. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, pp. 2055–2063. <https://doi.org/10.1109/ICCVW.2017.241>
- Preece, A., Harborne, D., Braines, D., Tomsett, R., Chakraborty, S., 2018. Stakeholders in explainable AI. *arXiv*.
- Pyörälä, J., Saarinen, N., Kankare, V., Coops, N.C., Liang, X., Wang, Y., Holopainen, M., Hyyppä, J., Vastaranta, M., 2019. Variability of wood properties using airborne and terrestrial laser scanning. *Remote Sens. Environ.* 235, 111474. <https://doi.org/10.1016/j.rse.2019.111474>
- Pyun, K.R., Rogers, J.A., Ko, S.H., 2022. Materials and devices for immersive virtual reality. *Nat. Rev. Mater.* 7, 841–843. <https://doi.org/10.1038/s41578-022-00501-5>
- Qin, H., Wang, C., Xi, X., Tian, J., Zhou, G., 2017. Estimation of coniferous forest aboveground biomass with aggregated airborne small-footprint LiDAR full-waveforms. *Opt. Express* 25, A851. <https://doi.org/10.1364/OE.25.00A851>
- Qiu, Z., Ma, F., Li, Z., Xu, X., Ge, H., Du, C., 2021. Estimation of nitrogen nutrition index in rice from UAV RGB images coupled with machine learning algorithms. *Comput. Electron. Agric.* 189, 106421. <https://doi.org/10.1016/j.compag.2021.106421>
- Quiñones, R., Munoz-Arriola, F., Choudhury, S. Das, Samal, A., 2021. Multi-feature data repository development and analytics for image cosegmentation in high-throughput plant phenotyping. *PLoS One* 16, e0257001. <https://doi.org/10.1371/journal.pone.0257001>
- Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with

deep convolutional generative adversarial networks abs/1511.0.

Ravindran, P., Wiedenhoeft, A.C., 2020. Comparison of two forensic wood identification technologies for ten Meliaceae woods: computer vision versus mass spectrometry. *Wood Sci. Technol.* 54, 1139–1150. <https://doi.org/10.1007/s00226-020-01178-1>

Ray, D.K., Mueller, N.D., West, P.C., Foley, J.A., 2013. Yield trends are insufficient to double global crop production by 2050. *PLoS One* 8, e66428. <https://doi.org/10.1371/journal.pone.0066428>

Raza, S.-A., Smith, H.K., Clarkson, G.J.J., Taylor, G., Thompson, A.J., Clarkson, J., Rajpoot, N.M., 2014. Automatic detection of regions in spinach canopies responding to soil moisture deficit using combined visible and thermal imagery. *PLoS One* 9, e97612. <https://doi.org/10.1371/journal.pone.0097612>

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 6517–6525. <https://doi.org/10.1109/CVPR.2017.690>

Redmon, J., Farhadi, A., 2018. YOLOv3: an incremental improvement. *CoRR* abs/1804.0.

Reen, F.J., Jump, O., McSharry, B.P., Morgan, J., Murphy, D., O’Leary, N., O’Mahony, B., Scallan, M., Supple, B., 2021. The use of virtual reality in the teaching of challenging concepts in virology, cell culture and molecular biology. *Front. Virtual Real.* 2, 670909. <https://doi.org/10.3389/frvir.2021.670909>

Resente, G., Gillert, A., Trouillier, M., Anadon-Rosell, A., Peters, R.L., von Arx, G., von Lukas, U., Wilmking, M., 2021. Mask, train, repeat! Artificial intelligence for quantitative wood anatomy. *Front. Plant Sci.* 12, 767400. <https://doi.org/10.3389/fpls.2021.767400>

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. “Why should I trust you?": explaining the

- predictions of any classifier. In: Krishnapuram, B. (Ed.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Robertson, G.P., Hamilton, S.K., Barham, B.L., Dale, B.E., Izaurrealde, R.C., Jackson, R.D., Landis, D.A., Swinton, S.M., Thelen, K.D., Tiedje, J.M., 2017. Cellulosic biofuel contributions to a sustainable energy future: choices and outcomes. *Science* (80-.). 356, eaal2324. <https://doi.org/10.1126/science.aal2324>
- Rogelj, J., Popp, A., Calvin, K. V., Luderer, G., Emmerling, J., Gernaat, D., Fujimori, S., Strefler, J., Hasegawa, T., Marangoni, G., Krey, V., Kriegler, E., Riahi, K., van Vuuren, D.P., Doelman, J., Drouet, L., Edmonds, J., Fricko, O., Harmsen, M., Havlík, P., Humpenöder, F., Stehfest, E., Tavoni, M., 2018. Scenarios towards limiting global mean temperature increase below 1.5 °C. *Nat. Clim. Chang.* 8, 325–332. <https://doi.org/10.1038/s41558-018-0091-3>
- Roitsch, T., Cabrera-Bosquet, L., Fournier, A., Ghamkhar, K., Jiménez-Berni, J., Pinto, F., Ober, E.S., 2019. Review: new sensors and data-driven approaches—A path to next generation phenomics. *Plant Sci.* 282, 2–10. <https://doi.org/10.1016/j.plantsci.2019.01.011>
- Rosen, J., 2018. The carbon harvest. *Science* (80-.). 359, 733–737. <https://doi.org/10.1126/science.359.6377.733>
- Rosenberg, L., 2022. Regulation of the Metaverse: A Roadmap. In: Sohel, F. (Ed.), The 6th International Conference on Virtual and Augmented Reality Simulations. ACM, New York, NY, USA, pp. 21–26. <https://doi.org/10.1145/3546607.3546611>
- Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. <https://doi.org/10.1007/s11263-015-0816-y>

- Sa, I., Popović, M., Khanna, R., Chen, Z., Lottes, P., Liebisch, F., Nieto, J., Stachniss, C., Walter, A., Siegwart, R., 2018. WeedMap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sens.* 10, 1423. <https://doi.org/10.3390/rs10091423>
- Sallustio, L., Harfouche, A.L., Salvati, L., Marchetti, M., Corona, P., 2022. Evaluating the potential of marginal lands available for sustainable cellulosic biofuel production in Italy. *Socioecon. Plann. Sci.* 82, 101309. <https://doi.org/10.1016/j.seps.2022.101309>
- Sampath, V., Murtua, I., Aguilar Martín, J.J., Gutierrez, A., 2021. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *J. Big Data* 8, 27. <https://doi.org/10.1186/s40537-021-00414-0>
- Sánchez, E., Scordia, D., Lino, G., Arias, C., Cosentino, S.L., Nogués, S., 2015. Salinity and water stress effects on biomass production in different *Arundo donax* L. clones. *Bioenergy Res.* 8, 1461–1479. <https://doi.org/10.1007/s12155-015-9652-8>
- Schallner, L., Rabold, J., Scholz, O., Schmid, U., 2019. Effect of superpixel aggregation on explanations in LIME - A case study with biological Data. In: *Machine Learning and Knowledge Discovery in Databases International Workshops of ECML PKDD*. Springer, Würzburg, Germany, pp. 147–158. https://doi.org/10.1007/978-3-030-43823-4_13
- Schapire, R.E., 1990. The strength of weak learnability. *Mach. Learn.* 5, 197–227. <https://doi.org/10.1007/BF00116037>
- Schmidhuber, J., 2015. Deep learning in neural networks: an overview. *Neural Networks* 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Schunck, D., Magistri, F., Rosu, R.A., Cornelißen, A., Chebrolu, N., Paulus, S., Léon, J., Behnke, S., Stachniss, C., Kuhlmann, H., Klingbeil, L., 2021. Pheno4D: a spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS One* 16, e0256340. <https://doi.org/10.1371/journal.pone.0256340>

- Schwartz, R., Down, L., Jonas, A., Tabassi, E., 2021. A proposal for identifying and managing bias in artificial intelligence. NIST Spec. Publ. <https://doi.org/10.6028/NIST.SP.1270-draft>
- Scott, M.L., 2009. Programming Language Pragmatics, Programming Language Pragmatics. Elsevier. <https://doi.org/10.1016/B978-0-12-374514-9.X0001-8>
- Seattle Central College, 2020. Mini house gives carpentry students experience in green construction [WWW Document]. URL <https://newscenter.seattlecentral.edu/2020/02/18/mini-house-gives-carpentry-students-experience-green-construction> (accessed 11.15.22).
- Seddiki, K., Saudemont, P., Precioso, F., Ogrinc, N., Wisztorski, M., Salzert, M., Fournier, I., Droit, A., 2020. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nat. Commun.* 11, 5595. <https://doi.org/10.1038/s41467-020-19354-z>
- Seelig, T.L., 2015. *Insight out: get ideas out of your head and into the world.* HarperOne, New York, United States.
- Seidel, D., Annighöfer, P., Thielman, A., Seifert, Q.E., Thauer, J.-H., Glatthorn, J., Ehbrecht, M., Kneib, T., Ammer, C., 2021. Predicting tree species from 3D laser scanning point clouds using deep learning. *Front. Plant Sci.* 12, 1–12. <https://doi.org/10.3389/fpls.2021.635440>
- Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S. V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci. Rep.* 11, 1606. <https://doi.org/10.1038/s41598-020-80820-1>
- Shamsolmoali, P., Zareapoor, M., Granger, E., Zhou, H., Wang, R., Celebi, M.E., Yang, J., 2021. Image synthesis with adversarial networks: a comprehensive survey and case studies. *Inf. Fusion* 72, 126–146. <https://doi.org/10.1016/j.inffus.2021.02.014>
- Shao, Gang, Shao, Guofan, Gallion, J., Saunders, M.R., Frankenberger, J.R., Fei, S., 2018. Improving LiDAR-based aboveground biomass estimation of temperate hardwood forests with varying site productivity. *Remote Sens. Environ.* 204, 872–882. <https://doi.org/10.1016/j.rse.2017.09.011>

- Sharif, M., Khan, M.A., Iqbal, Z., Azam, M.F., Lali, M.I.U., Javed, M.Y., 2018. Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. *Comput. Electron. Agric.* 150, 220–234. <https://doi.org/10.1016/j.compag.2018.04.023>
- Shi, P., Wang, Y., Xu, J., Zhao, Y., Yang, B., Yuan, Z., Sun, Q., 2021. Rice nitrogen nutrition estimation with RGB images and machine learning methods. *Comput. Electron. Agric.* 180, 105860. <https://doi.org/10.1016/j.compag.2020.105860>
- Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., Cohen, S.N., 1973. An artificial intelligence program to advise physicians regarding antimicrobial therapy. *Comput. Biomed. Res.* 6, 544–560. [https://doi.org/10.1016/0010-4809\(73\)90029-3](https://doi.org/10.1016/0010-4809(73)90029-3)
- Shu, M., Shen, M., Zuo, J., Yin, P., Wang, M., Xie, Z., Tang, J., Wang, R., Li, B., Yang, X., Ma, Y., 2021. The application of UAV-based hyperspectral imaging to estimate crop traits in maize inbred lines. *Plant Phenomics* 2021, 1–14. <https://doi.org/10.34133/2021/9890745>
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. *Nature* 550, 354–359. <https://doi.org/10.1038/nature24270>
- Simon, H.A., 1993. Anecdotes-A very early expert system. *IEEE Ann. Hist. Comput.* 15, 64–68. <https://doi.org/10.1109/85.222851>
- Sinek, S., 2011. *Start with why: how great leaders inspire everyone to take action*. Portfolio, Brentford, United Kingdom.
- Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S., 2016. Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21, 110–124. <https://doi.org/10.1016/j.tplants.2015.10.015>
- Singh, A., Jones, S., Ganapathysubramanian, B., Sarkar, S., Mueller, D., Sandhu, K., Nagasubramanian, K., 2021. Challenges and opportunities in machine-augmented plant stress phenotyping. *Trends Plant Sci.* 26, 53–69. <https://doi.org/10.1016/j.tplants.2020.07.010>

- Singh, A.K., Ganapathysubramanian, B., Sarkar, S., Singh, A., 2018. Deep learning for plant stress phenotyping: trends and future perspectives. *Trends Plant Sci.* 23, 883–898. <https://doi.org/10.1016/j.tplants.2018.07.004>
- Skibba, R., 2021. Japan's fugaku supercomputer crushes competition, but likely not for long. *Engineering* 7, 6–7. <https://doi.org/10.1016/j.eng.2020.12.003>
- Smith, A.G., Petersen, J., Selvan, R., Rasmussen, C.R., 2020. Segmentation of roots in soil with U-Net. *Plant Methods* 16, 13. <https://doi.org/10.1186/s13007-020-0563-0>
- Smith, P., Davis, S.J., Creutzig, F., Fuss, S., Minx, J., Gabrielle, B., Kato, E., Jackson, R.B., Cowie, A., Kriegler, E., van Vuuren, D.P., Rogelj, J., Ciais, P., Milne, J., Canadell, J.G., McCollum, D., Peters, G., Andrew, R., Krey, V., Shrestha, G., Friedlingstein, P., Gasser, T., Grüber, A., Heidug, W.K., Jonas, M., Jones, C.D., Kraxner, F., Littleton, E., Lowe, J., Moreira, J.R., Nakicenovic, N., Obersteiner, M., Patwardhan, A., Rogner, M., Rubin, E., Sharifi, A., Torvanger, A., Yamagata, Y., Edmonds, J., Yongsung, C., 2016. Biophysical and economic limits to negative CO₂ emissions. *Nat. Clim. Chang.* 6, 42–50. <https://doi.org/10.1038/nclimate2870>
- Soltaninejad, M., Sturrock, C.J., Griffiths, M., Pridmore, T.P., Pound, M.P., 2020. Three dimensional root CT segmentation using multi-resolution encoder-decoder networks. *IEEE Trans. Image Process.* 29, 6667–6679. <https://doi.org/10.1109/TIP.2020.2992893>
- Song, P., Wang, J., Guo, X., Yang, W., Zhao, C., 2021. High-throughput phenotyping: breaking through the bottleneck in future crop breeding. *Crop J.* 9, 633–645. <https://doi.org/10.1016/j.cj.2021.03.015>
- Spark, A., Kitching, A., Esteban-Ferrer, D., Handa, A., Carr, A.R., Needham, L.-M., Ponjavic, A., Santos, A.M., McColl, J., Leterrier, C., Davis, S.J., Henriques, R., Lee, S.F., 2020. vLUME: 3D virtual reality for single-molecule localization microscopy. *Nat. Methods* 17, 1097–1099. <https://doi.org/10.1038/s41592-020-0962-1>
- Sreeram, M., Nof, S.Y., 2021. Human-in-the-loop: role in cyber physical agricultural systems. *Int. J. Comput. Commun. Control* 16, 4166.

<https://doi.org/10.15837/ijccc.2021.2.4166>

Stephenson, N., 1992. *Snow Crash*. Bantam Books, New York, United States.

Streich, J., Romero, J., Gazolla, J.G.F.M., Kainer, D., Cliff, A., Prates, E.T., Brown, J.B., Khoury, S., Tuskan, G.A., Garvin, M., Jacobson, D., Harfouche, A.L., 2020. Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? *Curr. Opin. Biotechnol.* 61, 217–225. <https://doi.org/10.1016/j.copbio.2020.01.010>

Su, Y., Wu, F., Ao, Z., Jin, S., Qin, F., Liu, B., Pang, S., Liu, L., Guo, Q., 2019. Evaluating maize phenotype dynamics under drought stress using terrestrial LiDAR. *Plant Methods* 15, 11. <https://doi.org/10.1186/s13007-019-0396-x>

Sun, Z., Zhu, M., Shan, X., Lee, C., 2022. Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions. *Nat. Commun.* 13, 5224. <https://doi.org/10.1038/s41467-022-32745-8>

Taghavi Namin, S., Esmaeilzadeh, M., Najafi, M., Brown, T.B., Borevitz, J.O., 2018. Deep phenotyping: Deep learning for temporal phenotype/genotype classification. *Plant Methods* 14, 66. <https://doi.org/10.1186/s13007-018-0333-4>

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., Bennett, M., 2017. Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, 770–783. <https://doi.org/10.1016/j.cub.2017.05.055>

Tauro, F., Maltese, A., Giannini, R., Harfouche, A.L., 2022. Latent heat flux variability and response to drought stress of black poplar: A multi-platform multi-sensor remote and proximal sensing approach to relieve the data scarcity bottleneck. *Remote Sens. Environ.* 268, 112771. <https://doi.org/10.1016/j.rse.2021.112771>

ten Harkel, J., Bartholomeus, H., Kooistra, L., 2020. Biomass and crop height estimation of different crops using UAV-based LiDAR. *Remote Sens.* 12, 17. <https://doi.org/10.3390/RS12010017>

Thapa, R., Zhang, K., Snaveley, N., Belongie, S., Khan, A., 2020. The plant pathology challenge 2020 data set to classify foliar disease of apples. *Appl. Plant Sci.* 8, e11390. <https://doi.org/10.1002/aps3.11390>

Thapa, S., Zhu, F., Walia, H., Yu, H., Ge, Y., 2018. A novel LiDAR-based instrument

- for high-throughput, 3D measurement of morphological traits in maize and sorghum. *Sensors* 18, 1187. <https://doi.org/10.3390/s18041187>
- Tickle, A.B., Andrews, R., Golea, M., Diederich, J., 1998. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Trans. Neural Networks* 9, 1057–1068. <https://doi.org/10.1109/72.728352>
- Tilly, N., Aasen, H., Bareth, G., 2015. Fusion of plant height and vegetation indices for the estimation of barley biomass. *Remote Sens.* 7, 11449–11480. <https://doi.org/10.3390/rs70911449>
- Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci.* 108, 20260–20264. <https://doi.org/10.1073/pnas.1116437108>
- Toda, Y., Okura, F., 2019. How convolutional neural networks diagnose plant disease. *Plant Phenomics* 2019, 1–14. <https://doi.org/10.34133/2019/9237136>
- Turing, A.M., 1950. I.—Computing machinery and intelligence. *Mind* LIX, 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Tzounis, A., Katsoulas, N., Bartzanas, T., Kittas, C., 2017. Internet of things in agriculture, recent advances and future challenges. *Biosyst. Eng.* 164, 31–48. <https://doi.org/10.1016/j.biosystemseng.2017.09.007>
- Ubbens, J.R., Stavness, I., 2017. Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* 8, 1190. <https://doi.org/10.3389/fpls.2017.01190>
- Uchiyama, H., Sakurai, S., Mishima, M., Arita, D., Okayasu, T., Shimada, A., Taniguchi, R., 2017. An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, pp. 2038–2045. <https://doi.org/10.1109/ICCVW.2017.239>
- United Nations, 2022. World population prospects 2022: summary of results. UN DESA/POP/2022/TR/NO. 3.
- van Dijk, A.D.J., Kootstra, G., Kruijer, W., de Ridder, D., 2021. Machine learning in plant science and plant breeding. *iScience* 24, 101890.

<https://doi.org/10.1016/j.isci.2020.101890>

- van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. *Comput. Electron. Agric.* 177, 105709. <https://doi.org/10.1016/j.compag.2020.105709>
- Vilone, G., Longo, L., 2020. Explainable artificial intelligence: a systematic review. *arXiv*.
- Virlet, N., Sabermanesh, K., Sadeghi-Tehran, P., Hawkesford, M.J., 2017. Field scanalyzer: An automated robotic field phenotyping platform for detailed crop monitoring. *Funct. Plant Biol.* 44, 153. <https://doi.org/10.1071/FP16163>
- Vo, D.H., Vo, A.T., 2021. Renewable energy and population growth for sustainable development in the Southeast Asian countries. *Energy. Sustain. Soc.* 11, 30. <https://doi.org/10.1186/s13705-021-00304-6>
- vurukonda, N., Rao, B.T., 2016. A study on data storage security issues in cloud computing. *Procedia Comput. Sci.* 92, 128–135. <https://doi.org/10.1016/j.procs.2016.07.335>
- Wagner, G., Schlenker, W., 2022. Declining crop yields limit the potential of bioenergy. *Nature* 609, 250–251. <https://doi.org/10.1038/d41586-022-02344-0>
- Wainberg, M., Alipanahi, B., Frey, B.J., 2016. Are random forests truly the best classifiers? *J. Mach. Learn. Res.* 17, 1–5.
- Waliman, M., Zakhori, A., 2019. Height estimation of biomass sorghum in the field using LiDAR. In: *IS and T International Symposium on Electronic Imaging Science and Technology*. Society for Imaging Science and Technology, pp. 137–138. <https://doi.org/10.2352/ISSN.2470-1173.2019.13.COIMG-137>
- Walter, A., Finger, R., Huber, R., Buchmann, N., 2017. Smart farming is key to developing sustainable agriculture. *Proc. Natl. Acad. Sci. U. S. A.* 114, 6148–6150. <https://doi.org/10.1073/pnas.1707462114>
- Walter, J.D.C., Edwards, J., McDonald, G., Kuchel, H., 2019. Estimating biomass and canopy height with LiDAR for field crop breeding. *Front. Plant Sci.* 10, 1145. <https://doi.org/10.3389/fpls.2019.01145>

- Wan-Mohd-Jaafar, W.S., Woodhouse, I., Silva, C.A., Omar, H., Hudak, A., 2017. Modelling individual tree aboveground biomass using discrete return LiDAR in lowland dipterocarp forest of Malaysia. *J. Trop. For. Sci.* 29, 465–484. <https://doi.org/10.26525/jtfs2017.29.4.465484>
- Wang, C., Nie, S., Xi, X., Luo, S., Sun, X., 2016. Estimating the biomass of Maize with hyperspectral and LiDAR data. *Remote Sens.* 9, 11. <https://doi.org/10.3390/rs9010011>
- Wang, X., Xuan, H., Evers, B., Shrestha, S., Pless, R., Poland, J., 2019. High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. *Gigascience* 8, 1–11. <https://doi.org/10.1093/gigascience/giz120>
- Wang, Y., Wu, G., Deng, L., Tang, Z., Wang, K., Sun, W., Shangguan, Z., 2017. Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm. *Sci. Rep.* 7, 6940. <https://doi.org/10.1038/s41598-017-07197-6>
- Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2021. Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* 53, 1–34. <https://doi.org/10.1145/3386252>
- Wiesner-Hanks, T., Stewart, E.L., Kaczmar, N., DeChant, C., Wu, H., Nelson, R.J., Lipson, H., Gore, M.A., 2018. Image set for deep learning: field images of maize annotated with disease symptoms. *BMC Res. Notes* 11, 440. <https://doi.org/10.1186/s13104-018-3548-6>
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The

- FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Williamson, H.F., Brettschneider, J., Caccamo, M., Davey, R.P., Goble, C., Kersey, P.J., May, S., Morris, R.J., Ostler, R., Pridmore, T., Rawlings, C., Studholme, D., Tsaftaris, S.A., Leonelli, S., 2021. Data management challenges for artificial intelligence in plant and agricultural research. *F1000Research* 10, 324. <https://doi.org/10.12688/f1000research.52204.1>
- Wismer, P., Lopez Cordoba, A., Baceviciute, S., Clauson-Kaas, F., Sommer, M.O.A., 2021. Immersive virtual reality as a competitive training strategy for the biopharma industry. *Nat. Biotechnol.* 39, 116–119. <https://doi.org/10.1038/s41587-020-00784-5>
- Wolf, M., Logan, J., Mehta, K., Jacobson, D., Cashman, M., Walker, A.M., Eisenhauer, G., Widener, P., Cliff, A., 2021. Reusability first: toward FAIR workflows. In: 2021 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, pp. 444–455. <https://doi.org/10.1109/Cluster48925.2021.00053>
- Wooden robot, 2021. vrkshop [WWW Document]. URL <https://www.vrkshop.com/> (accessed 11.15.22).
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* 14, 1–37. <https://doi.org/10.1007/s10115-007-0114-2>
- Xu, L., Cruz, J.A., Savage, L.J., Kramer, D.M., Chen, J., 2015. Plant photosynthesis phenomics data quality control. *Bioinformatics* 31, 1796–1804. <https://doi.org/10.1093/bioinformatics/btu854>
- Xu, S., Wang, R., Gasser, T., Ciais, P., Peñuelas, J., Balkanski, Y., Boucher, O., Janssens, I.A., Sardans, J., Clark, J.H., Cao, J., Xing, X., Chen, J., Wang, L., Tang, X., Zhang, R., 2022. Delayed use of bioenergy crops might threaten climate and food security. *Nature* 609, 299–306. <https://doi.org/10.1038/s41586-022-05055-8>
- Xue, B., Sartori, P., Leibler, S., 2019. Environment-to-phenotype mapping and

- adaptation strategies in varying environments. *Proc. Natl. Acad. Sci.* 116, 13847–13855. <https://doi.org/10.1073/pnas.1903232116>
- Yang, G., Liu, J., Zhao, C., Li, Zhenhong, Huang, Y., Yu, H., Xu, B., Yang, X., Zhu, D., Zhang, X., Zhang, R., Feng, H., Zhao, X., Li, Zhenhai, Li, H., Yang, H., 2017. Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. *Front. Plant Sci.* 8, 1111. <https://doi.org/10.3389/fpls.2017.01111>
- Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J.H., Batchelor, W.D., Xiong, L., Yan, J., 2020. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214. <https://doi.org/10.1016/j.molp.2020.01.008>
- Yang, Y., 2017. Ensemble learning. In: *Temporal Data Mining Via Unsupervised Ensemble Learning*. Elsevier, pp. 35–56. <https://doi.org/10.1016/B978-0-12-811654-8.00004-X>
- Yasrab, R., Atkinson, J.A., Wells, D.M., French, A.P., Pridmore, T.P., Pound, M.P., 2019. RootNav 2.0: deep learning for automatic navigation of complex plant root architectures. *Gigascience* 8, giz123. <https://doi.org/10.1093/gigascience/giz123>
- Yeung, G.C.N., Yeung, H.C., 1985. Principles of programming languages: design, evaluation, and implementation. *Proc. IEEE* 73, 1245–1245. <https://doi.org/10.1109/PROC.1985.13274>
- Yu, H., Li, J., 2022. Breeding future crops to feed the world through de novo domestication. *Nat. Commun.* 13, 1171. <https://doi.org/10.1038/s41467-022-28732-8>
- Yuan, J., Wen, T., Zhang, H., Zhao, M., Penton, C.R., Thomashow, L.S., Shen, Q., 2020. Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt. *ISME J.* 14, 2936–2950. <https://doi.org/10.1038/s41396-020-0720-5>
- Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J., 2019. CutMix: regularization strategy to train strong classifiers with localizable features. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 6022–6031.

<https://doi.org/10.1109/ICCV.2019.00612>

Zadpoor, A., 2019. Fifty years is not a lot of time! Matter 1, 1096–1098.

<https://doi.org/10.1016/j.matt.2019.10.011>

Zamir, D., 2001. Improving plant breeding with exotic genetic libraries. Nat. Rev.

Genet. 2, 983–989. <https://doi.org/10.1038/35103590>

Zhang, L., Grift, T.E., 2012. A LiDAR-based crop height measurement system for

Miscanthus giganteus. Comput. Electron. Agric. 85, 70–76.

<https://doi.org/10.1016/j.compag.2012.04.001>

Zhang, Z., Cao, L., She, G., 2017. Estimating forest structural parameters using

canopy metrics derived from airborne LiDAR data in subtropical forests. Remote

Sens. 9, 940. <https://doi.org/10.3390/rs9090940>

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D.B., Huang, Y., Huang, M., Yao, Y.,

Bassu, S., Ciais, P., Durand, J.-L., Elliott, J., Ewert, F., Janssens, I.A., Li, T., Lin,

E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A.C., Wallach,

D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., Asseng, S., 2017. Temperature

increase reduces global yields of major crops in four independent estimates. Proc.

Natl. Acad. Sci. 114, 9326–9331. <https://doi.org/10.1073/pnas.1701762114>

Zhou, S., Chai, X., Yang, Z., Wang, H., Yang, C., Sun, T., 2021. Maize-IAS: a maize

image analysis software using deep learning for high-throughput plant

phenotyping. Plant Methods 17, 48. <https://doi.org/10.1186/s13007-021-00747-0>

Zhu, W., Sun, Z., Peng, J., Huang, Y., Li, J., Zhang, J., Yang, B., Liao, X., 2019.

Estimating maize above-ground biomass using 3D point clouds of multi-source

unmanned aerial vehicle data at multi-spatial scales. Remote Sens. 11, 2678.

<https://doi.org/10.3390/rs11222678>

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A., 2019. A primer

on deep learning in genomics. Nat. Genet. 51, 12–18.

<https://doi.org/10.1038/s41588-018-0295-5>

Vitae

Farid Nakhle holds a Bachelor's degree in computer science, awarded cum laude, from the American University of Technology (2013) and a Master of Science degree, awarded cum laude, in Computer Science from the American University of Science and Technology (2017). He has an extensive experience in computer networks, programming, AI, and DevOps engineering gained over his 7-year career in software and web development. During his PhD in computational phenomics at Antoine L. Harfouche's laboratory, Farid designed an X-AI workflow for phenomic data analysis. He applied it to successfully implement and train three different AI algorithms to classify leaf diseases and spider mites in tomato plants and used a *post hoc* X-AI algorithm to explain their predictions. Farid then designed another workflow for phenomic data analysis using interpretable by design X-AI algorithms. He successfully applied the workflow to train an interpretable by design algorithm to classify cassava leaf diseases and explain its own reasoning for each prediction. He also successfully implemented many AI algorithms for dimensionality reduction, tree and forest biomass estimation, and image sharpening, among others. Farid also gained extensive knowledge in extended reality (XR). He co-designed and built the eXtended Reality and eXplainable Artificial Intelligence (X-RAI) park for digital agricultural biotechnology as well as the wood science and technology metaverse campus (WOODMeta). Farid also engaged in teaching and training PhD students and early career researchers. He assisted in designing and successfully delivering a practical tutorial on AI and explainable AI in the “Accelerating Climate Resilient Plant Breeding by Applying – Omics and Artificial Intelligence” PhD course organized by the Swedish University of Agricultural Sciences, Alnarp, Sweden, 20–24 April 2020, and in the “Common Data Standards and Integrative Analytic Methods for Plant Phenotyping” PhD course organized by the Nordic Forestry, Veterinary and Agricultural (NOVA) University Network, Alnarp/Lund, Sweden, 28 November–02 December 2022. Farid was appointed as Track Chair for Universal Access (Fintech, Education, Health) in the IEEE International Symposium on Technology and Society (ISTAS23) – Technology and Analytics for Global Development Conference, Swansea University, Wales, United Kingdom, 13–15 September 2023. The research conducted during his time at the University of Tuscia resulted in several scientific writings in reputable journals such as *Trends in Plant Science*, *Trends in Biotechnology*, *Remote Sensing*, and *Patterns*.