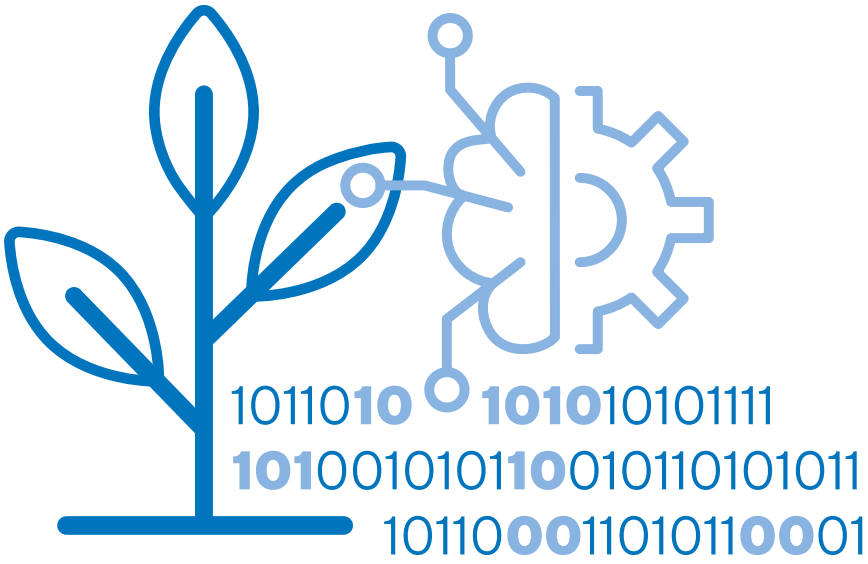
Data Science

für die Pflanzenzüchtung 4.0



Gemeinschaft zur Förderung
von Pflanzeninnovation e. V. (GFPi)

GFPi
Lebensbasis Pflanze

Vorwort

Die Landwirtschaft in Deutschland und Europa steht vor großen Herausforderungen. Der Klimawandel und die in der europäischen Gemeinschaft formulierten gesellschaftlichen Ziele wie der Schutz von Insekten, Biodiversität, Boden und Grundwasser erfordern die Transformation zu nachhaltigeren Wirtschaftsformen.

Pflanzenzüchtung und -forschung werden essenzielle Handlungsfelder bei der Gestaltung zukünftiger Landwirtschafts- und Gartenbausysteme sein. Neue ressourceneffiziente und klimaresiliente Pflanzensorten bilden die Grundlage für diese Systeme. Basierend auf der Auswertung großer Datenmengen tragen prädiktive Züchtungsansätze dazu bei, standortangepasste Sorten effizienter zu entwickeln und somit besser auf geänderte Anforderungen bzw. Anbaubedingungen reagieren zu können.

Wir, die Gemeinschaft zur Förderung von Pflanzeninnovation e. V., sehen in Data Science daher ein zentrales Werkzeug der zukünftigen Pflanzenzüchtung. Mit diesem kann die rasant wachsende Datenmenge resultierend aus Genomanalysen, Phänotypisierung, Gewächshaus- und Feldprüfungen sowie der Züchtungsforschung unter Berücksichtigung von Umweltdaten und Daten aus der Wertschöpfungskette ausgewertet und in Züchtungsfortschritt umgesetzt werden.

Wir Pflanzenzüchter und -forscher werden uns aktiv einbringen und einen Rahmen zur gemeinsamen Nutzung von Daten entwickeln. Die Zielsetzung ist ein von Transparenz und Vertrauen geprägtes Datenökosystem, das allen Unternehmen den Zugang zu digitalen Anwendungen ermöglicht. Um alle notwendigen Datenquellen für die Pflanzenzüchtung nutzbar zu machen, sind darüber hinaus wissenschaftliche Forschungsaktivitäten zur Entwicklung geeigneter Methoden und Werkzeuge erforderlich.

Eine erfolgreiche Forschung im Bereich der Data Science für die Pflanzenzüchtung bedarf aus unserer Sicht eines maßgeblichen Beitrags der öffentlichen Forschungsförderung. Nur so kann Züchtungsfortschritt effektiv gesichert werden, um auch in Zukunft eine produktive, vielfältige und ressourcenschonende Landwirtschaft zu ermöglichen.

Bonn, im August 2021

Wolf von Rhade
(Vorsitzender der GFPi)

Inhalt

Zusammenfassung
Seite 3

Anforderungen an die zukünftige Pflanzenzüchtung
Seite 4

Die Chancen einer Data Science für die Pflanzenzüchtung 4.0
Seite 5

Wo steht die Pflanzenzüchtung heute?
Seite 9

Voraussetzungen – Data Science für die Pflanzenzüchtung 4.0
Seite 14

Glossar
Seite 19

Legende:

Systembiologie

Genotyp × Umwelt × Management-Interaktionen (G×E×M)

Datengrundlage

Tools

Hinweis: Die im Text unterstrichenen Fachbegriffe werden im Glossar erklärt.

Datenschutzhinweis: Die GFPi e. V. nimmt den Datenschutz sehr ernst. Unsere Datenschutzerklärung finden Sie unter https://www.bdp-online.de/de/GFPi/Datenschutz/

Die gewählte männliche Form bezieht gleichermaßen weibliche oder diverse Personen mit ein. Auf eine konsequente Doppelbezeichnung wurde aufgrund besserer Lesbarkeit verzichtet.

Zusammenfassung

Data Science für die Pflanzenzüchtung 4.0

Die Pflanzenzüchtung orientiert sich seit jeher an den Bedürfnissen der Landwirtschaft. Die aktuellen Herausforderungen durch den Klimawandel und gesellschaftliche Anforderungen wie der Schutz von Insekten, Biodiversität, Boden und Grundwasser an die landwirtschaftliche Praxis wirken sich daher direkt auf die Züchtungsziele aus. Die Nutzung großer Datenmengen (Data Science) in der Pflanzenzüchtung birgt immenses Potenzial für die effiziente Umsetzung dieser Züchtungsziele zur Entwicklung standortangepasster und resilienter Sorten.

Auf dem Acker und im Unterglasanbau werden Sorten benötigt, die tolerant gegenüber abiotischen und biotischen Stressfaktoren sind, gute Qualitätseigenschaften aufweisen, eine wirtschaftliche Produktion ermöglichen und gleichzeitig ressourcenschonend angebaut werden können. Die effiziente Entwicklung dieser komplexen Pflanzeneigenschaften in neuen Sorten kann durch Data Science deutlich beschleunigt werden.

Neue Ansätze für die prädiktive Züchtung

Durch das Zusammenführen großer Datenströme aus vielfältigen Quellen kann das virtuelle Abbild einer Pflanze in Form eines Modells geschaffen werden, das alle Informationen über ihre Eigenschaften enthält. Die Interaktionen der Leistungsfaktoren Genotyp, Umwelt und Management könnten mithilfe dieses Modells erstmals detailliert analysiert werden. In prädiktiven Züchtungsansätzen könnten dadurch verbesserte Leistungsvorhersagen getroffen und die optimalen Genotypen ausgewählt werden.

Wo steht die Pflanzenzüchtung heute?

Die Pflanzenzüchtung in Deutschland greift bereits heute auf ein zunehmend vertieftes Verständnis der Funktion einzelner Gene sowie der Struktur und Diversität ganzer Pflanzengenome zurück. Erkenntnisse aus Forschungsprogrammen wie GABI, PLANT 2030 und Agrarsysteme der Zukunft bieten wichtige Anknüpfungspunkte für zukünftige Forschung. An verschiedenen Stellen in der Pflanzenzüchtung und -forschung sowie in der Landwirtschaft werden große Mengen an präzisen Daten zu stetig sinkenden Kosten erhoben. Die Nutzung dieser Daten in innovativen Data-Science-Ansätzen ist für die zukünftige Pflanzenzüchtung von zentraler Bedeutung und stellt eine große Herausforderung dar.

Voraussetzungen für Data Science in der Pflanzenzüchtung

Um Data Science für die Pflanzenzüchtung nutzbar zu machen, sind interdisziplinäre Forschungsansätze in der Grundlagen- und angewandten Forschung notwendig. Das Ziel ist die Entwicklung geeigneter Methoden und Werkzeuge im Bereich Data Science für Pflanzenzüchtungsunternehmen in Deutschland. Flankiert werden muss die digitale Transformation durch eine berufliche Aus- und Weiterbildungsoffensive in der Pflanzenzüchtung. Die Branche ist gefordert, geeignete Strukturen für Datenökosysteme zu entwickeln.

In einem ersten Schritt hat die Gemeinschaft zur Förderung von Pflanzeninnovation e. V. (GFPi) ihre Vision, notwendigen Forschungsbedarf und Ansatzpunkte zur Etablierung von Data Science für die Pflanzenzüchtung in dieser Publikation festgehalten. Ein öffentlich gefördertes Forschungsprogramm sollte langfristig, mehrstufig und an die dynamische Entwicklung der Digitalisierung angepasst sein. Der Dialog mit den Stakeholdern entlang der pflanzenbaulichen Wertschöpfungskette sollte die Umsetzung begleiten.

Anforderungen an die zukünftige Pflanzenzüchtung

Die Pflanzenzüchtung hat das Ziel, der Landwirtschaft und dem Gartenbau stetig verbesserte Sorten aus einem breiten Kulturartenspektrum zur Verfügung zu stellen, um die Ernährungssicherung einer wachsenden Weltbevölkerung, die Versorgung mit Futtermitteln sowie die Bereitstellung von Biomasse für die Bioökonomie und als energetischer Rohstoff zu gewährleisten.

Die pflanzliche Erzeugung steht aktuell vor einer Vielzahl von Herausforderungen. Der Green Deal der EU und nationale Strategien zum Schutz von Klima, Insekten, Biodiversität, Boden und Grundwasser unterstreichen die gesellschaftlichen Ansprüche an eine nachhaltige Landwirtschaft und stellen konkrete Anforderungen an die Pflanzenzüchtung. In den nächsten Dekaden wird daher der Bedarf an neuen, innovativen Sorten weiter steigen.

Die von der Pflanzenzüchtungsforschung entwickelte Strategie „Pflanzenzüchtung 4.0“ zeigt die benötigten Forschungs- und Entwicklungsansätze auf. Dazu gehören die prädiktive Pflanzenzüchtung, Genotyp × Umwelt × Management-Interaktionen sowie die funktionelle Genomanalyse.

Data Science stellt die Schlüsseltechnologie dar, um moderne Forschungs- und Entwicklungsansätze in der Pflanzenzüchtung zu realisieren, miteinander zu verknüpfen und in einen beschleunigten Zuchtfortschritt zu übersetzen.

Die zukünftigen Forschungsschwerpunkte hat die GFPi in der Forschungsstrategie „Pflanzenzüchtung 4.0“ festgehalten.

Die Chancen einer Data Science für die Pflanzenzüchtung 4.0

Data Science soll ein neues, zentrales Werkzeug für die Pflanzenzüchtung werden. Als solches soll Data Science zukünftig prädiktive Ansätze ermöglichen, mit deren Hilfe pflanzliche Eigenschaften auf Grundlage ihrer genetischen Ausstattung vorhergesagt werden können. Dies ermöglicht den Züchtern, mit neuen, standortangepassten Sorten noch schneller auf geänderte Anforderungen reagieren zu können. In der Pflanzenzüchtung wachsen die Datenmengen rasant an. Die Berücksichtigung von Umweltdaten sowie weiterer Daten aus der Wertschöpfungskette und der Wissenschaft wird ein entscheidender Baustein sein, um die Zusammenhänge zwischen Genotyp, Umwelt und Bewirtschaftungsverfahren analysieren und prädiktive Züchtungsverfahren auf der Grundlage von Modellierungen und Vorhersagen realisieren zu können.

Leistungsvorhersagen auf Grundlage der „virtuellen Pflanze“

Zukünftig sollen Ergebnisse aus der funktionellen Genomanalyse sowie der systembiologischen Forschung es zusammen mit Data-Science-basierten Modellierungsansätzen ermöglichen, pflanzliche Merkmalsausprägungen auf Grundlage der genetischen Ausstattung mit deutlich verbesserter Präzision vorhersagen zu können. Das Ziel dieser Arbeiten ist die Vorhersage der Leistungsfähigkeit eines bestimmten Genotyps in einer bestimmten Umwelt auf der Basis seiner Genomsequenz. So kann Data Science der Pflanzenzüchtung helfen, standortangepasste Sorten, die den aktuellen Anforderungen an die moderne Landwirtschaft entsprechen, schneller bereitzustellen.

Die dazu notwendigen, vielfältigen Modellierungsansätze sollen in einer „virtuellen Pflanze“ integriert werden. Das Fernziel „virtuelle Pflanze“ soll es ermöglichen, den Effekt aller bekannten Sequenzvarianten einer Kulturart und deren Kombinationen auf die biologischen Prozesse der Pflanze und letztlich ihre Merkmalsausprägung (Phänotyp) zu modellieren. Dadurch soll es möglich werden, bereits vor der Kreuzung und Selektion vorauszusagen, wie sich die Pflanze im Feld verhält.

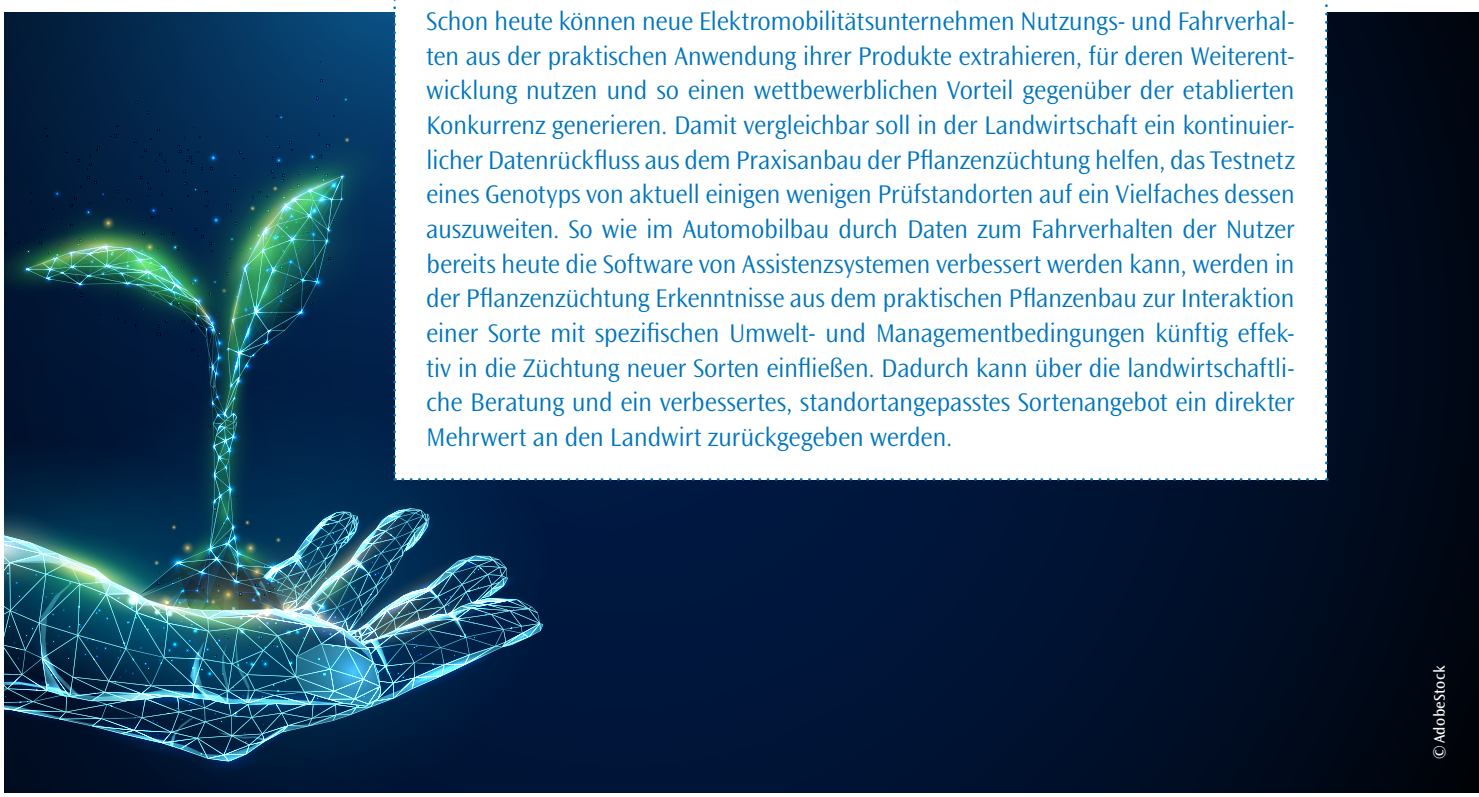
In Analogie zur Humanmedizin, in der personalisierte Humantherapeutika bereits Realität sind, werden die Züchter durch diese Ansätze zukünftig standortangepasste Managementempfehlungen für ihre Nutzpflanzensorten anbieten können. Solche Empfehlungen für vorteilhafte, standortbezogene Managementmaßnahmen können die Sequenz-basierten Vorhersagen zur Leistungsfähigkeit eines bestimmten Genotyps wirkungsvoll ergänzen.

Verbesserte komplexe Pflanzeneigenschaften mithilfe von Big Data aus Umwelterhebungen und der landwirtschaftlichen Praxis

Im Rahmen einer Data Science für die Pflanzenzüchtung 4.0 sollen zukünftig auch komplexe pflanzliche Eigenschaften, wie beispielsweise der Ertrag, unter Zuhilfenahme von neu entwickelten Methoden vorhergesagt und damit züchterisch bearbeitet werden können. So lassen sich die Faktoren Umwelt und

Nutzerfeedback als Motor für Innovation

Schon heute können neue Elektromobilitätsunternehmen Nutzungs- und Fahrverhalten aus der praktischen Anwendung ihrer Produkte extrahieren, für deren Weiterentwicklung nutzen und so einen wettbewerblichen Vorteil gegenüber der etablierten Konkurrenz generieren. Damit vergleichbar soll in der Landwirtschaft ein kontinuierlicher Datenrückfluss aus dem Praxisanbau der Pflanzenzüchtung helfen, das Testnetz eines Genotyps von aktuell einigen wenigen Prüfstandorten auf ein Vielfaches dessen auszuweiten. So wie im Automobilbau durch Daten zum Fahrverhalten der Nutzer bereits heute die Software von Assistenzsystemen verbessert werden kann, werden in der Pflanzenzüchtung Erkenntnisse aus dem praktischen Pflanzenbau zur Interaktion einer Sorte mit spezifischen Umwelt- und Managementbedingungen künftig effektiv in die Züchtung neuer Sorten einfließen. Dadurch kann über die landwirtschaftliche Beratung und ein verbessertes, standortangepasstes Sortenangebot ein direkter Mehrwert an den Landwirt zurückgegeben werden.



Daten aus der landwirtschaftlichen Praxis werden im Züchtungsprozess berücksichtigt.

Management sowie deren Interaktionen mit dem pflanzlichen Genotyp sehr viel detaillierter bewerten und im Züchtungsprozess berücksichtigen. Die Grundlage dafür sind große Datenmengen, die beispielsweise auch durch ein kontinuierliches Nutzerfeedback aus der praktischen Landwirtschaft zur Leistung einer Sorte zur Verfügung gestellt werden sollen. Die praktische Landwirtschaft unterstützt die Realisierung komplexer Zuchtziele in der Zukunft damit zusätzlich.

Höhere Diversität und schnelle Adaption von Kulturarten mittels Data Science

Zwei aktuelle Erwartungen an die Pflanzenzüchtung, denen zukünftig mithilfe von Data Science besser begegnet werden können soll, sind die Erhöhung der Diversität auf dem Acker durch zusätzliche Kulturarten sowie eine möglichst schnelle und effiziente Adaptation von Elitesorten an neue klimatische Bedingungen im Zuge des Klimawandels. Anhand der weltweiten Adaptation und Ausbreitung des Maisanbaus im vergangenen Jahrhundert lässt sich zwar eindrucksvoll demonstrieren, dass durch die Kombination klassischer und moderner Zuchtmethoden die Anpassung einer Kulturart an völlig neue Anbau- und Klimabedingungen mit globalem Erfolg erreichbar ist. Vergleichbare Entwicklungen für die Adaptation zusätzlicher Kulturpflanzen im Zuge des aktuellen Klimawandels müssen allerdings wesentlich beschleunigt werden (von vielen Jahrzehnten auf möglichst wenige Jahre).

Projekt BreedPatH: KI in der Rapszüchtung

Ein gutes erstes Beispiel für das Potenzial von Data Science bei der Anpassung von Kulturarten ist das aktuelle BMBF-Verbundvorhaben **BreedPatH**. Hier wird das Ziel verfolgt, die praktische Rapszüchtung mit Vorhersagemethoden, die auf umfangreichen Genomdatensätzen und Künstlicher Intelligenz beruhen, zu beschleunigen. So gelang es beispielsweise bereits, anhand von Computersimulationen neuartige Kreuzungsschemata umzusetzen, mit denen sehr schnell sehr große sogenannte „heterotische Pools“ für die Hybridzüchtung zur Verfügung gestellt werden konnten. Ein vergleichbarer Erfolg, wie ihn die internationale Maiszüchtung im letzten Jahrhundert ohne Genomdaten über viele Dekaden sehr intensiver Züchtungsarbeit erreichte, kann mit der Unterstützung von Data Science ggf. in wenigen Jahren realisiert werden. In Zukunft werden Ansätze dieser Art die Etablierung neuer Kulturarten in Deutschland erleichtern und die Biodiversität auf dem Acker steigern.

Vollumfängliche Datennutzung

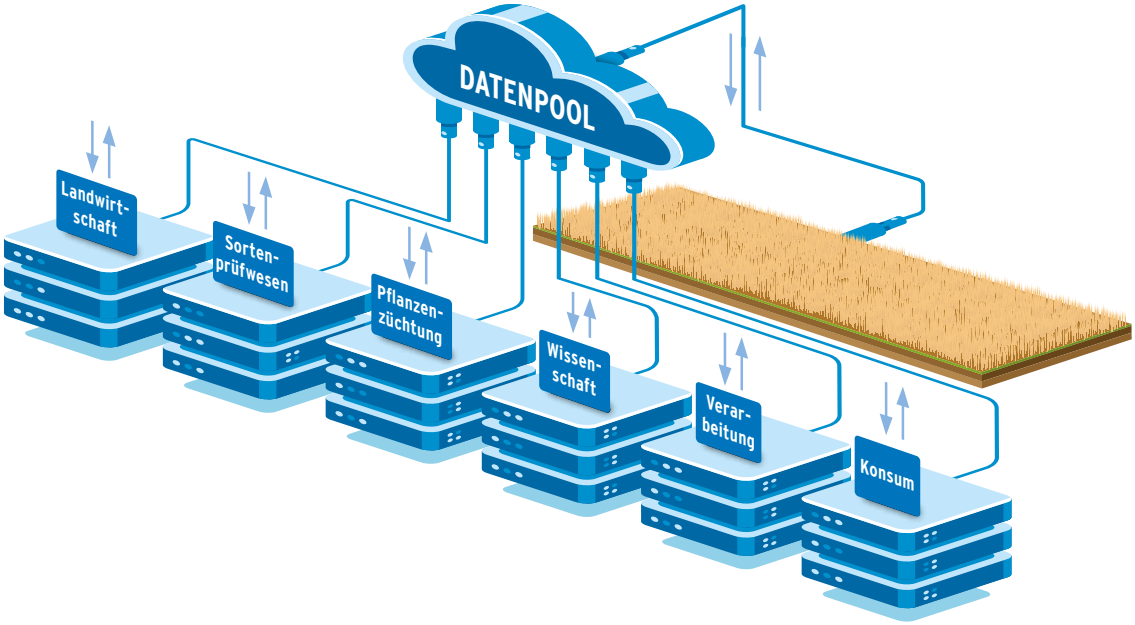
Die Basis der Data Science in der Pflanzenzüchtung 4.0 soll eine breitere und stark wachsende Datengrundlage sein, die aus Datenquellen der gesamten pflanzenbaulichen Wertschöpfungskette einschließlich Klima- und Wetterdaten gespeist wird. Alle verfügbaren Daten und Datenquellen sollen infolge gerichteter Forschungs- und Koordinationsanstrengungen zukünftig miteinander integrier- und nutzbar sein. Entsprechende qualitätsgeprüfte Datensätze werden dazu vor ihrem Eintritt in geeignete Datenräume in standardisierte Formate überführt und detailliert dokumentiert, also durch zusätzliche Metadaten beschrieben und in eine umfängliche Datensemantik eingebettet sein. Die Dokumentation der Daten soll gemäß der FAIR-Prinzipien erfolgen, sodass Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendung der digitalen Daten verbessert werden. Etablierte Datenontologien sollen dabei helfen, die Gesamtheit der potenziell verfügbaren Daten interpretier- und nutzbar zu machen.

Neue Werkzeuge für die Pflanzenzüchtung

Eine weitere Grundvoraussetzung für Data Science für die Pflanzenzüchtung 4.0 sind zukünftige neue Data-Science-Methoden. Diese müssen speziell für die Pflanzenzüchtung entwickelt werden und dazu dienen, die erhobenen Daten mit der zugrunde liegenden genetischen Ausstattung der Pflanzen in Beziehung zu setzen. Dazu sollen alle notwendigen Verfahren der Erzeugung und des Managements von Big Data für die Pflanzenzüchtung adaptiert und anwendbar gemacht werden. Weiterhin soll die traditionelle Lehre der Selektionstheorie für diesen erweiterten Werkzeugkasten grundlegend überarbeitet werden.

Wo steht die Pflanzenzüchtung heute?

In den letzten zwei Dekaden hat die Pflanzengenomforschung den pflanzlichen Zuchtfortschritt substanziell beschleunigt. Aufgrund der raschen Entwicklungen auf diesem Gebiet verfügen wir inzwischen über ein vertieftes Verständnis der Funktion einzelner Gene sowie der Struktur und Diversität ganzer Pflanzengenome. Die Verfügbarkeit von Hochdurchsatz-Sequenzierungs- und Genotypisierungstechnologien sowie die Weiterentwicklung der Populationsgenetik, der theoretischen Grundlagen der quantitativen Genetik, der Zuchtmethodik und der automatisierten Phänotypisierung machen es heute möglich, große Mengen an präzisen Daten zu stetig sinkenden Kosten zu erheben. Zusätzlich stellt die fortschreitende Digitalisierung der Landwirtschaft eine Fülle an neuen Daten für die Pflanzenzüchtung zur Verfügung. Die Nutzung dieser Daten in Ansätzen der Data Science ist für die zukünftige Pflanzenzüchtung von zentraler Bedeutung, befindet sich heute aber erst in ihren Anfängen.



Die Ausgangslage für Data Science in der Pflanzenzüchtung

Aktuell stammt ein Großteil der in der kommerziellen Pflanzenzüchtung genutzten Daten von den jeweiligen Unternehmen selbst. Entsprechende genotypische und phänotypische Daten werden in eigenen Zuchtgärten und einem betriebseigenen Prüfnetz an regional verteilten Standorten generiert und fließen direkt in die jeweiligen Zuchtprozesse ein. Allerdings reichen aufgrund der in Deutschland mittelständisch geprägten Pflanzenzüchtungsbranche und entsprechender Unternehmensgrößen die Mengen an proprietären Daten in den meisten Betrieben für Data-Science-getriebene Ansätze bisher nicht aus. Neben proprietären Daten der Züchtungsunternehmen stellen heute auch öffentlich zugängliche Datenbanken für genetische Sequenzinformationen sowie Daten und Kenntnisse aus der Wissenschaft für die Pflanzenzüchtung relevante und genutzte Ressourcen dar, die zur Definition neuer Zuchtziele und zur Auswahl der dafür notwendigen Genotypen und Kreuzungspartner genutzt werden.

Neue Daten- und Informationsquellen

Von Stakeholdern aus der Wertschöpfungskette werden viele für die Pflanzenzüchtung relevante Daten erhoben, die der Züchtung bislang kaum zugänglich sind und daher wenig oder überhaupt nicht genutzt werden können. Dazu gehören beispielsweise Daten aus der praktischen Landwirtschaft. Hier werden die Sorten unter diversen Umwelt- und Managementbedingungen angebaut. In diesem Zusammenhang werden bereits heute von vernetzten und mit Sensoren bestückten Landmaschinen sowie von Drohnen- und Satellitenbetreibern riesige Datensätze generiert.

Entsprechende Daten zu Ertrag, Klima, Umwelt, Management und Qualität könnten die proprietären Datensätze der Pflanzenzüchter deutlich erweitern. Auch vom Erfassungshandel sowie den verarbeitenden

» **GABI** HAT MIT ALL SEINEN ERGEBNISSEN UND DER VERNETZUNG DER FORSCHER DIE WISSENSCHAFTLICHE BASIS DER **PFLANZEN-GENOMFORSCHUNG IN DEUTSCHLAND GESTÄRKT.** «

Zitat aus der Kurzfassung des Evaluationsberichts „Evaluation der BMBF-Förderaktivität – Genomanalyse im biologischen System Pflanze (GABI)“ (2014)

DATA SCIENCE: DATENQUELLEN
Welche Stakeholder könnten welche Datenkategorien zur Verfügung stellen?

Übersicht über verschiedene Arten von Daten mit Relevanz für die Pflanzenzüchtung von Stakeholdern aus der pflanzenbaulichen Wertschöpfungskette

Betrieben werden weitere Rohstoffdaten, beispielsweise zu Qualitätseigenschaften von Erntegut und spezifischen Inhaltsstoffen, erhoben. Ein Zugang zu diesen Daten sowie die wissenschaftlichen Grundlagen zur Verrechnung solcher heterogener Datensätze für die Pflanzenzüchtung sind aktuell nicht gegeben.

Auch in jüngster Vergangenheit entstandene, öffentlich geförderte Strukturen und Netzwerke mit einem Fokus auf Daten und Datenverarbeitung in der Landwirtschaft, wie das Deutsche Netzwerk für Bioinformatik-Infrastruktur – **de.NBI**, die Nationale Forschungsdateninfrastruktur für Agrarwissenschaften **NFDI4Plants** sowie die **Agri-Gaia**-Initiative stellen wertvolle öffentliche Plattformen dar, um Datenressourcen für die Pflanzenzüchtung zugänglich zu machen. Eine entsprechende Vernetzung muss gewährleistet werden, um Daten aus der Pflanzenforschung und der Wertschöpfungskette in der Entwicklung von Anwendungen zu nutzen.

Zusammengefasst sind die Fortschritte in der Digitalisierung der praktischen Landwirtschaft, die entwickelten Methoden zur Genotypisierung und Phänotypisierung sowie die vornehmlich Grundlagen-orientierte funktionelle Genomik und die Systembiologie dazu geeignet, sehr große Datenmengen bereitzustellen, um die Grundlage für eine Data Science in der Pflanzenzüchtung 4.0 bilden zu können.

Automatisierte Feldphänotypisierung in einem GFPi-Gemeinschaftsforschungsprojekt

Big Data: Alle für Einen

Die Möglichkeit, aus großen Datensätzen einen Mehrwert für konkrete Einzelfälle zu generieren, begegnet uns schon heute im Alltag. Algorithmen von Navigationsdienstleistern können bereits heute das Fahrverhalten von Nutzern und das Verkehrsaufkommen analysieren und auf das individuelle Ziel des Nutzers optimierte Wegführungen erstellen.

Mithilfe von Methoden und Werkzeugen aus der Data Science werden in diesem Fall große Datenmengen von Tausenden Individuen gesammelt, um eine optimierte, personifizierte Lösung für das einzelne Individuum vorhersagen und anbieten zu können. In der Pflanzenzüchtung ist das Potenzial zur Generierung großer Datenmengen aus Tausenden Situationen (jeweils eine Sorte (ein Genotyp) in einer bestimmten Umwelt (Anbausituation)) ebenfalls sehr hoch. Das Ziel, das Verhalten einer bestimmten Sorte in einem bestimmten Umwelt-/Anbau-Szenario vorauszusagen, ist deshalb greifbar und kann für die Züchtung verbesserter Sorten und damit zur Erfüllung politischer und gesellschaftlicher Ziele im Bereich der Landwirtschaft eingesetzt werden.

Aktuelle Nutzung von Data Science in der Pflanzenzüchtung

Kuratierung von Daten

Die oben aufgeführten Datenquellen machen die große Diversität der relevanten Daten deutlich. Um diese Diversität für die Pflanzenzüchtung in Ansätzen der Data Science, wie beispielsweise künstlichen neuronalen Netzwerkarchitekturen, nutzen zu können, müssen entsprechende Daten einerseits zunächst verfügbar gemacht werden. Andererseits müssen aber vor allem auch die wissenschaftlichen Grundlagen zur Handhabbarkeit der Daten, d. h. Interpretierbarkeit und Dateninteroperabilität, geschaffen werden. Diese Grundlagen sind heute noch nicht gegeben. Allerdings beschäftigen sich bereits erste Ansätze, wie beispielsweise das Forschungsprojekt „**BigData**“ mit der Kuratierung und Zusammenführung von unterschiedlichen Züchtungsdaten zum Zweck einer gemeinsamen Verrechnung. Erste Erfolge zeigen, dass eine Nutzbarkeit des gesamten in der pflanzlichen Wertschöpfungskette verfügbaren Datenschatzes für die Pflanzenzüchtung durchaus realistisch ist.

Die Interaktion von Umwelt und Bewirtschaftungsverfahren mit dem pflanzlichen Genotyp

Erste Ansätze zur Nutzung großer Mengen von Umweltdaten in der Pflanzenzüchtung mittels Wetterstationen und Bodensensoren existieren bereits heute und gewinnen in der kommerziellen Pflanzenzüchtung zunehmend an Bedeutung. Die Ansätze werden beispielsweise genutzt, um eine Klassifizierung von Makroumwelten anhand historischer Klimadaten durchzuführen. Die gewonnenen Daten können verwendet werden, um geeignete Ziel- und Testumwelten für Zuchtprogramme zu identifizieren. Dabei ermöglichen intrasaisonale Umweltdaten es den Züchtern, Ähnlichkeiten von verschiedenen Testumwelten anhand relevanter Umweltparameter zu beschreiben. Die gewonnenen Erkenntnisse können dann im Rahmen der genomischen Leistungsvorhersage für die Selektion genutzt werden. Ein weiteres Anwendungsbeispiel für die Nutzbarkeit von Umweltdaten ist die Verbesserung der Ertragsstabilität. Hier kann unter anderem die Sensitivität potenzieller Sorten gegenüber relevanten Umweltparametern (z. B. Temperatur) berechnet werden. Dies ermöglicht es den Züchtern, jene Sorten zu selektieren, die sich beispielsweise als besonders stabil gegenüber

» [...] ALS **THEMENSCHWERPUNKTE FÜR DIE ZUKUNFT** [WURDEN] INSBESONDERE DIE BEREICHE **PHÄNOTYPISIERUNG, BIOINFORMATIK/BIG DATA**, PFLANZEN(GENOM) FORSCHUNG FÜR DIE **BIOÖKONOMIE** SOWIE **FORSCHUNG AN KULTUR- UND NUTZPFLANZEN** IDENTIFIZIERT. «

Zitat aus der Kurzfassung des Evaluationsberichts „Evaluation der BMBF-Förderaktivität – Genomanalyse im biologischen System Pflanze (GABI)“ (2014)

Digitale Tools unterstützen einen standortangepassten Pflanzenbau.

Temperaturschwankungen bzw. Witterungsextremen auszeichnen. Mithilfe entsprechender Ansätze werden bereits heute trockentolerante Sorten mit verbesserter Ertragsstabilität für Südosteuropa entwickelt. Die standardmäßige Vorhersagbarkeit von Genotyp-Umwelt Interaktionen durch Data Science ist ungleich schwieriger. Zukünftig werden große wissenschaftliche Anstrengungen benötigt, um vor allem komplexe pflanzliche Eigenschaften wie den Ertrag vorhersagen zu können.

Data Science und Systembiologie

Durch systembiologische Ansätze, in denen komplexe Zusammenhänge über alle biologischen Prozessebenen hinweg von der Erbgutvariation bis hin zur Merkmalsausprägung und der Wechselwirkung zwischen Pflanze und Umwelt analysiert werden, sollen die Funktionen der Gene einer Pflanzenart aufgeklärt werden. Erste Ansätze dazu verfolgt beispielsweise das Forschungsprojekt **AVATARS**. Das Ziel von AVATARS ist die Vorhersage und Aufklärung genetischer und umweltbedingter Einflüsse auf Sameneigenschaften bei Raps. Dazu werden einerseits Hochdurchsatzdaten zu Genotypen sowie digitale Phänotyp- und Umweltdaten aus Feldversuchen mittels Künstlicher Intelligenz in Bezug zur Keimfähigkeit von über 350.000 Einzelsamen gesetzt. Andererseits wird ein virtuelles räumlich-zeitliches Samen-Modell erstellt, in dem Organ-spezifische Omics-Daten unterschiedlicher Genotypen aus kontrastierenden Umwelten mittels Verfahren der virtuellen Realität visualisiert werden, um genetische und umweltbedingte Einflüsse auf Regulations- und Stoffwechselprozesse aufzudecken, die die Sameneigenschaften bestimmen.

Wanted: Technologien der Data Science

Trotz einiger erster Beispiele der Nutzung von Data Science in der Pflanzenzüchtung fehlt es heute neben ausreichenden und nutzbaren Daten vor allem auch an geeigneten Methoden und Werkzeugen, um die in großen Datenmengen enthaltenen Informationen und Erkenntnisse für die Pflanzenzüchtung automatisiert extrahieren zu können. Diese müssen durch die Wissenschaft speziell für die Pflanzenzüchtung entwickelt werden.

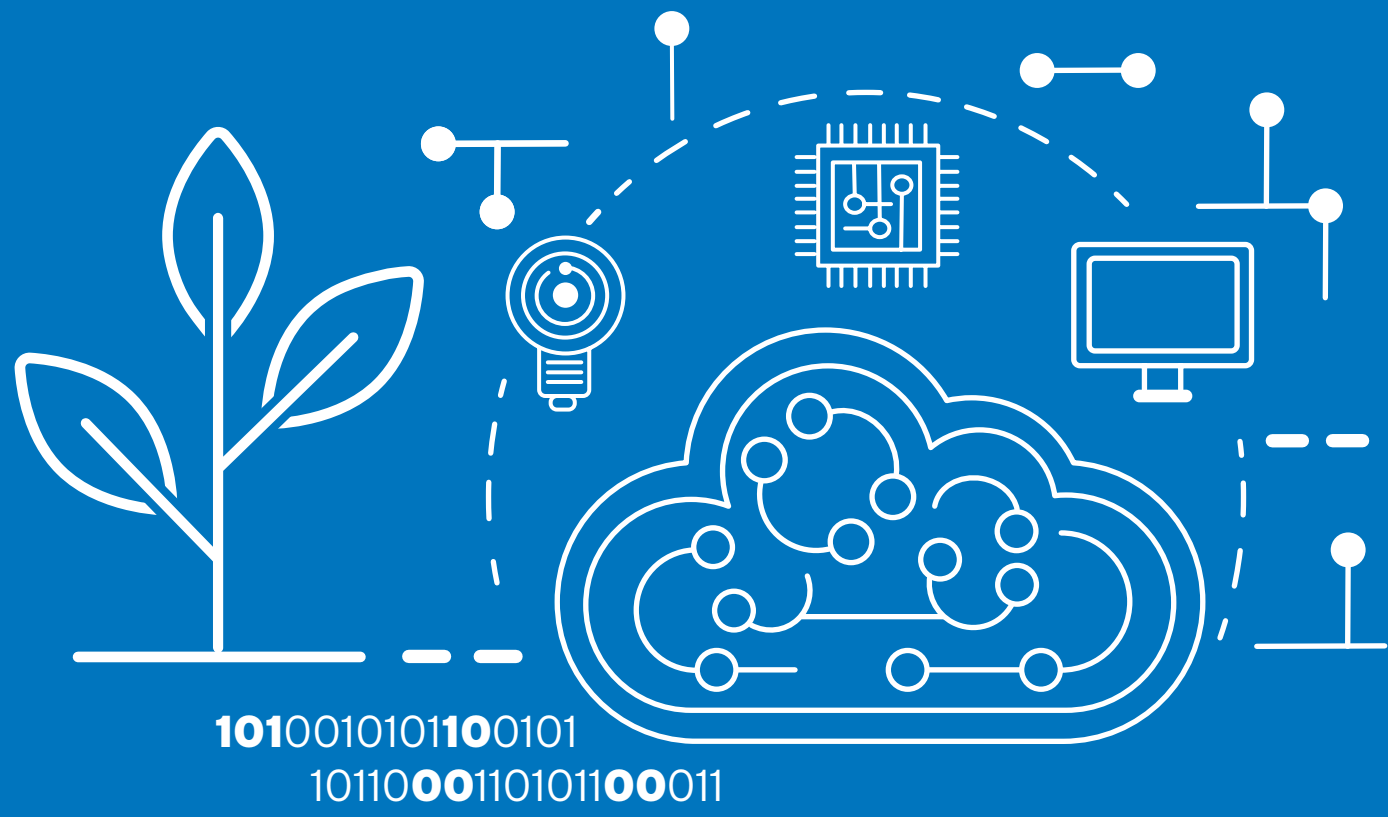
Data Science als möglicher Paradigmenwechsel in der Pflanzenzüchtung

Data Science ist mit Verfahren wie der Künstlichen Intelligenz (KI) ein Innovationstreiber, der sich in vielen Technologiebereichen international in rasantem Tempo entwickelt und die nächste Stufe der digitalen Transformation zur Bearbeitung großer Datenmengen aus verschiedenen Disziplinen darstellt. Die effektive Bewertung und Selektion neuer Zuchtlinien verbessert sich mit der Verfügbarkeit großer Datenmengen und der Fähigkeit ihrer effektiven Nutzung. Eine Zusammenarbeit zwischen Pflanzenzüchtung und Data-Science-Forschung ist deshalb als nächster logischer Schritt und Konsequenz aus der Genomforschung der letzten beiden Dekaden unabdingbar.

Die schnelle, zielgerichtete und effiziente Nutzung von Data Science kann einen Paradigmenwechsel in der Pflanzenzüchtung bewirken, welcher der wissensbasierten Züchtungsforschung neue qualitative und quantitative Dimensionen verleihen wird. Ziel ist die Etablierung von Data Science für eine Pflanzenzüchtung 4.0.

Voraussetzungen – Data Science für die Pflanzenzüchtung 4.0

Zur Etablierung einer Data Science für die Pflanzenzüchtung 4.0 bedarf es einer gesamtheitlichen, gemeinsamen Vorgehensweise der wissenschaftlichen Pflanzen- und Züchtungsforschung sowie der wirtschaftlich tätigen Unternehmen in der Pflanzenzüchtung. Notwendig sind wissenschaftliche Forschungs- und Entwicklungsaktivitäten zur Weiter- und Neuentwicklung geeigneter Methoden und Werkzeuge der Data Science für die Pflanzenzüchtung. Diese müssen durch praktische Konzepte zur Bereitstellung, Verknüpfung und Nutzung der benötigten Datensätze flankiert werden. Bei der Entwicklung und Implementierung der Konzepte müssen alle Stakeholder eingebunden werden.



Wissenschaftlicher Forschungsbedarf

11 Standards und Vorgaben zur Datennutzung

Die meisten KI-Algorithmen werden heutzutage mit sehr großen Sammlungen ausgewählter Datenobjekte für eine bestimmte Aufgabe vortrainiert. Die Anwendung solcher Ansätze ist in der Pflanzenzüchtung bisher noch nicht in vollem Umfang möglich: Die Multidimensionalität und die fehlende Orthogonalität von Daten machen es oft schwierig, benötigte Datensätze miteinander zu verknüpfen. Damit eine kritische Masse an Trainingsdaten für das effiziente Training von maschinellem Lernen (ML) und insbesondere KI-Algorithmen für die Pflanzenzüchtung erreicht werden kann, bedarf es neuer Möglichkeiten zur effizienten Standardisierung, Analyse und Kombination sowie zum Austausch von Datensätzen.

Dies kann durch die Etablierung von standardisierten Datenformaten sowie einheitlich strukturierten und für jedermann zugänglichen Metadaten ermöglicht werden. Solche Ansätze zur Metaanalyse, die auf einer definierten Datensemantik basieren, können die Nutzbarkeit heterogener und multidimensionaler Daten mit unterschiedlicher Qualität und Abdeckung verbessern. Im Bereich der Medizin wurde diese Entwicklung bereits durch das vom BMBF geförderte Programm „**i:DSem**“ angestoßen. Analog zu „i:DSem“ wird in der Pflanzenzüchtung eine Datensemantik benötigt, die Datenressourcen von der Gensequenz über Phänotypisierungsdaten bis hin zu Boden-, Klima- und Witterungsparametern miteinander verbindet.

Flankiert werden müssen die Anstrengungen von der Etablierung und Nutzung passender Datenontologien, also Netzwerken aus logischen Relationen verschiedener Informationen und Daten zueinander. Nicht zuletzt muss auf Wunsch auch eine effektive Anonymisierung von Daten zum Schutz der Interessen aller Beteiligten gewährleistet sein. Während solche synthetischen, anonymisierten Analoga von proprietären Daten per Definition nicht direkt interpretiert werden können, ermöglichen sie das Vortraining von KI-Algorithmen mit der gleichen Effizienz wie nicht anonymisierte Datensätze. Forscher und Pflanzenzüchter haben diesen Diskussionsprozess bereits gestartet und streben verbindliche branchenweite Standards auf Grundlage von Transparenz und Vertrauen an.

Über Data Science zur Pflanzensorte

Vor dem Hintergrund der aktuellen Herausforderungen werden standortangepasste Sorten, die Genotyp × Umwelt × Management-Interaktionen berücksichtigen, an Bedeutung gewinnen. Dies ist vor allem dem Umstand geschuldet, dass die meisten wirtschaftlich relevanten Pflanzenmerkmale (z. B. Kornertrag oder Ertragsstabilität) komplexe Merkmale sind, die unter Kontrolle einer Vielzahl interagierender Gene stehen. Aufgrund dessen kommt es zu starken Wechselwirkungen mit der Umwelt und einer verringerten Erblichkeit solcher Merkmale. Entsprechend ist die Ausprägung komplexer pflanzlicher Eigenschaften bisher teilweise weniger auf den Genotyp als auf Umwelteinflüsse und Bewirtschaftungsverfahren (Management) zurückzuführen.

Erste wissenschaftliche Studien zeigen, dass Verfahren der Künstlichen Intelligenz dazu geeignet sind, Interaktionen zwischen Genotyp, Umwelt und Management in den Vorhersagen des Ertrags zu berücksichtigen. Fortschritte in der Sensorik, gekoppelt mit leicht zugänglichen Umweltdaten, erlauben heutzutage eine detaillierte Beschreibung der Umwelten. Solche Daten sind die Basis, um die systematische Modellierung von Genotyp × Umwelt × Management-Interaktionen für die Pflanzenproduktion mittels Deep-Learning-Ansätzen Realität werden zu lassen. Die dazu erforderliche Dateninfrastruktur und für die Pflanzenzüchtung optimierte algorithmische Lösungen müssen jedoch erst noch entwickelt werden.

Funktionelle Genomanalyse und Systembiologie

Um die Leistungsfähigkeit eines Genotyps (G) in einem bestimmten Umweltszenario (E×M) auf biologischen Kenntnissen beruhend vorhersagen zu können, sind zukünftig große Anstrengungen in der systembiologischen Forschung erforderlich. Die systembiologische Innovationsspirale fußt auf bereits bestehenden biologischen Kenntnissen und Daten. Sie beginnt mit der gezielten experimentellen Erhebung umfangreicher molekularer Omics-Daten (z. B. Epi-/Genom, Transkriptom, Proteom, Metabolom, Ionom etc.) und phänotypischer Daten von bestimmten Genotypen in präzise erfassten Umwelten. Es folgt die Identifizierung von Beziehungen zwischen diesen Datensätzen, um regulatorische, metabolische und entwicklungsbiologische Netzwerke aufzudecken.

Aus diesen Erkenntnissen werden Modelle erstellt, die biologische Prozessketten simulieren. Die Modellierungen ermöglichen die Vorhersage der Ausprägung von Eigenschaften von Pflanzen in diversen G×E×M-Szenarien. Diese Vorhersagen werden dann für ausgewählte Genotypkollektionen experimentell geprüft. Auf Grundlage der Prüfergebnisse und weiterer erhobener Daten lassen sich die abgeleiteten Netzwerke und Modelle ergänzen und verfeinern, um die Leistungsvorhersagen im nächsten Zyklus der Innovationsspirale zu optimieren.

Das Fernziel „virtuelle Pflanze“

Die bei jedem Durchlauf der systembiologischen Forschungs- und Innovationsspirale erstellten bzw. optimierten Modelle einer Kulturart ergeben in ihrer Gesamtheit eine virtuelle Pflanze, mit deren Hilfe die Vorhersage der Leistungsfähigkeit eines bestimmten Genotyps (G) auf der Basis seiner individuellen Genomsequenz in einem bestimmten Umweltszenario (G×E×M) ermöglicht werden soll. Die dazu erstellten Modelle können also genutzt werden, um genetische Faktoren (Gene/Allele) für verbesserte Leistungen vorherzusagen und so zielgerichtete Züchtungsansätze zu ermöglichen. Die Erstellung der beschriebenen Crop-Modelle bzw. „virtuellen Pflanzen“ bedarf massiver wissenschaftlicher Forschung. Zukünftig können die Modelle für die prädiktive Selektion in der Pflanzenzüchtung herangezogen werden.

Deep Learning-Technologien als Schlüssel zu Data Science

Um Data Science für die Pflanzenzüchtung anwendbar zu machen und die dargestellten Ziele zu realisieren,

müssen zunächst notwendige bioinformatische Werkzeuge entwickelt werden. Deep-Learning-Verfahren bieten sich zur Entwicklung passender Algorithmen an, die sich mit zunehmenden Mengen an Input-Daten weiter verbessern und es somit ermöglichen,

1. komplexe Muster selbstständig zu erkennen und zu extrahieren,
2. neue Formen und Dimensionen heterogener Daten aufzunehmen,
3. sich im Zusammenspiel mit technologischen Fortschritten wie z. B. verbesserten Genomsequenzierungstechnologien weiterzuentwickeln und
4. sich an genetische Verbesserungen des Zuchtmaterials anzupassen.

Angesichts der rasant wachsenden Datenmengen und der Komplexität in großen Zuchtpopulationen gehören zielgerichtete Deep-Learning-Verfahren für die Vorhersage von merkmalsrelevanten Genommustern und ihre Verbindung mit komplexen Phänotypdaten bzw. mit G×E-Interaktionen zu den aussichtsreichsten Zukunftsansätzen. Bei der Entwicklung und Implementierung von Deep-Learning-Verfahren kann auf bestehende Bibliotheken zurückgegriffen werden, daher ist die Hürde für deren technische Implementierung niedrig. Allerdings ist es sehr herausfordernd, die geeignete Architektur für die Algorithmen auszuwählen. Ein weiterer wichtiger Forschungsbedarf bei der Prädiktion der Interaktion von Umwelt und Bewirtschaftungsverfahren mit dem pflanzlichen Genotyp besteht in der Neuentwicklung der vor mehreren Jahrzehnten entwickelten Selektionstheorie, die an die neuen Möglichkeiten angepasst werden muss.

Koordinativer Bedarf

Forschern, Züchtern, Landwirten, Repository-Managern und Rechtsbeauftragten in der Pflanzenzüchtung fehlt ein rechtliches, organisatorisches und technisches Rahmenwerk zur gemeinsamen Nutzung ihrer Daten in einem Vertrauensraum. Das ist für die Akzeptanz von Data Science in der Pflanzenzüchtung aber essenziell. Entsprechende Konzepte müssen deshalb entwickelt werden. Beispielsweise müssen alle beteiligten Partner immer genau wissen und steuern können, für welche Zwecke, in welcher Art und von wem ihre im System zur Verfügung gestellten Datensätze genutzt werden dürfen. Für die Akzeptanz ist auch eine wirksame Codierung von Datensätzen eine Grundvoraussetzung. So muss beispielsweise sichergestellt werden können, dass Daten einerseits in gemeinsamen Ansätzen zur Erhöhung der Vorhersagefähigkeit von Modellen und Algorithmen beitragen können. Andererseits müssen die Interessen aller beteiligten Partner stets gewahrt bleiben.

Darüber hinaus werden Technologien benötigt, um datengebenden Akteuren die Möglichkeit zu bieten, über potenziell verfügbare Daten zu informieren und/oder diese zu speichern. Die Bereitstellung entsprechender Datensätze von Stakeholdern und der Wertschöpfungskette muss durch Anreizsysteme gefördert werden. Im Bereich der Pflanzenzüchtung werden nicht alle Unternehmen im gleichen Umfang Daten für gemeinsame Ansätze in der Data Science zur Verfügung stellen können, da sich allein schon die Größe der Zuchtprogramme z. T. stark unterscheidet. Damit also kein Partner im Verhältnis zu seinem Beitrag überproportional profitiert oder benachteiligt wird, muss eine faire Gewichtung von unterschiedlichen Beiträgen etabliert werden.

Insgesamt muss der Austausch von Daten in der Wertschöpfungskette weiter verbessert werden, damit Daten leichter in datenbasierte Züchtungsansätze einfließen können. Dazu bedarf es zusätzlicher Anstrengungen zur Vernetzung der jeweiligen Stakeholder. Schnittstellen zur Datenübertragung müssen geschaffen und vorhandene Datenbanken miteinander vernetzt werden. Die Interoperabilität von Datensätzen muss durch die Etablierung von Standards garantiert werden.

Aus diesem Grund müssen flankierend zur Förderung der wissenschaftlichen Forschung auch koordinative Leistungen erbracht werden, um Data Science für die Pflanzenzüchtung zu etablieren. Es gilt, einen transparenten rechtlichen Rahmen, ein langfristig gesichertes und von allen Akteuren getragenes, nachhaltiges Betriebskonzept sowie eine zuverlässige technische Lösung zur gemeinsamen Nutzung von Daten zu Kulturpflanzen aus Forschung und Wirtschaft zu entwickeln. Dabei ist es von hoher Relevanz, durch ein transparentes Vorgehen das nötige Vertrauen für einen branchenweiten Dialog zu schaffen.

Aus- und Fortbildung

Um das Potenzial der Data Science für die praktische Züchtung auszuschöpfen, ist es unerlässlich, spezielle Verfahren des Maschinellen Lernens (ML) zu entwickeln, welche die mathematischen Prinzipien der Data Science aus einer spezifisch züchterischen Perspektive umsetzen. Dies erfordert dringend eine neue Generation von Wissenschaftlern, die ein profundes Wissen über die zugrunde liegende Mathematik der Mustererkennung bzw. über KI-basierte Vorhersagemethoden mit einem breiten Verständnis der Theorie und Praxis der Pflanzenzüchtung in Bezug auf die Züchtungsziele, -methoden und -technologien verbinden.

Einerseits müssen Züchter aus- bzw. weitergebildet werden, die mit Data-Science-basierten Selektionsstrategien im Zusammenhang mit der klassischen und molekularen Zuchtmethodik vertraut sind. Durch eine enge Kooperation der institutionellen Forschung mit der angewandten Pflanzenzüchtung sollten andererseits eine praxisorientierte Anwendung und Optimierung von Data-Science-gestützten Zuchtverfahren in verschiedenen Stufen der praktischen Züchtung umgesetzt werden. Somit sollen an verschiedenen Stellen in der Forschungs- und Entwicklungskette „digitale Züchter“ trainiert werden, die über die notwendigen multidisziplinären Hintergründe und die Expertise verfügen, um Wissen aus der Mathematik, Informatik und praktischen Pflanzenzüchtung zusammenzuführen.

In diesem Dialog mit der pflanzenbaulichen Wertschöpfungskette und durch Einbeziehung aller involvierten Stakeholder sollen folgende Fragen bearbeitet werden:

- Wie sieht das Anforderungsprofil für ein branchenweites Datenökosystem in der Pflanzenzüchtung aus?
- Wie müssen die rechtlichen Rahmenbedingungen, Nutzungsbedingungen und abgestufte Datennutzungsklassen sowie ein Anreizsystem für die Beteiligung an einem solchen Datenökosystem aussehen?
- Wie können effektive Schnittstellen zu weiteren Datenökosystemen entwickelt, implementiert und in ihrer Funktion analysiert werden?
- Wie sehen geeignete Dateninfrastrukturen für das Datenökosystem aus?

Glossar

Algorithmus: Eine eindeutige Handlungsvorschrift zur Lösung eines Problems oder einer Klasse von Problemen, die i. d. R. aus endlich vielen, wohldefinierten Einzelschritten besteht und zur Ausführung in ein Computerprogramm implementiert wird; z. B. ein Verfahren, das eine Schätzung des genotypischen Werts einer Pflanze anhand verfügbarer Daten (z. B. ‚Omics'-Daten oder G×E×M-Informationen) ermöglicht.

Big Data: Datenmengen, die zu groß, zu komplex, zu schnelllebig oder zu schwach strukturiert sind, um mit manuellen oder herkömmlichen Methoden ausgewertet zu werden. Big Data kann in fünf Dimensionen, die 4+1 Big Data V, unterteilt werden: Volume (extreme Datenmenge), Variety (Vielfalt der Dateistrukturen: unstrukturiert, semi-strukturiert, strukturiert), Velocity (Geschwindigkeit, mit der Daten produziert, aber auch verarbeitet werden müssen), Veracity (Unsicherheit der Daten und Datenqualität), Value (Mehrwert durch die (verknüpften) Datenmengen).

Data Science: Datenwissenschaften sind ein interdisziplinäres Wissenschaftsfeld, das sich mit der Extraktion von Wissen aus Daten befasst. Data Science steht für eine zweckorientierte Datenverarbeitung, -aufbereitung und -analyse und die systematische Generierung von Entscheidungshilfen.

Dateninteroperabilität: Interoperabilität ist die Fähigkeit unabhängiger, heterogener Systeme, nahtlos zusammenzuwirken, um Daten auf effiziente und verwertbare Art und Weise auszutauschen bzw. dem Benutzer zur Verfügung zu stellen, ohne dass dazu besondere Adaptierungen notwendig sind.

Datenökosystem: Ein Datenökosystem beschreibt ein Netzwerk, das aus autonom agierenden Akteuren besteht, welche direkt oder indirekt Daten und andere damit zusammenhängende Ressourcen (z. B. Software, Dienste und Infrastruktur) verbrauchen, produzieren oder bereitstellen. Jeder Akteur erfüllt eine oder mehrere Rollen und ist durch Beziehungen mit anderen Akteuren verbunden, sodass die Zusammenarbeit und der Wettbewerb der Akteure die Selbstregulierung des Datenökosystems fördern (vgl. Oliviera & Loscio 2018, S. 4).

Datenontologie: Wissensrepräsentation im Bereich der Künstlichen Intelligenz. Im Unterschied zu einer Taxonomie, die eine hierarchische Untergliederung abbildet, stellt eine Datenontologie ein Netzwerk von Informationen mit logischen Relationen dar. Datenontologien dienen als Mittel der Strukturierung und zum Datenaustausch, um bereits bestehende Wissensbestände zusammenzuführen oder in bestehenden Wissensbeständen zu suchen und diese zu editieren. Während in einer Datenbank Struktur und Inhalt maßgeblich sind, sind in einer Datenontologie die Daten beschrieben sowie Regeln über deren Zusammenhang. Ein Beispiel aus der Bioinformatik ist die Gene Ontology (GO) Initiative: GO ist eine biomedizinische Datenontologie, die drei Bereiche abdeckt: Zelluläre Komponente, Biologischer Prozess und Molekulare Funktion.

Datensemantik: Die Integrative Datensemantik adressiert die inhaltliche Homogenisierung und die automatisierte bzw. teilautomatisierte Analyse der Bedeutung heterogener Daten mit unterschiedlichen Formaten. Datensemantik trägt damit wesentlich zur De-facto-Standardisierung von Datensätzen bzw. komplexen Datenobjekten bei und ist deshalb hoch relevant für jede Form der Datenintegration.

Deep Learning: Spezielle Methode des Maschinellen Lernens. Hierbei werden z. B. künstliche neuronale Netze, die wie das menschliche Gehirn gebaut sind, verwendet, um sehr komplexe/umfangreiche Datensätze zu verarbeiten. Die Fähigkeit des Computers zu lernen wird durch die Generierung einer Hierarchie von Konzepten ermöglicht. Bekannte Beispiele sind Schachcomputer, oder in der Biologie das KI-System AlphaFold zur Vorhersage von 3D Protein-Strukturen.

FAIR data principles („FAIR Guiding Principles for scientific data management and stewardship"): Die FAIR-Prinzipien sollen Leitlinien zur Verbesserung der Auffindbarkeit (F für Findability), Zugänglichkeit (A für Accessibility), Interoperabilität (I für Interoperability) und Wiederverwendung (R für Reusability) digitaler Daten liefern. Hier sollen z. B. globale Identifikationsnummern beim Auffinden helfen, detaillierte Metadaten (wie wurden Daten generiert, wer hat sie erhoben/bearbeitet/publiziert und unter welchen Bedingungen (Lizenz) dürfen sie verwendet werden), die auch maschinenlesbar sein sollten, sind notwendig, um die Daten im richtigen Kontext interpretieren zu können.
Weitere Informationen: https://blogs.tib.eu/wp/tib/2017/09/12/die-fair-data-prinzipien-fuer-forschungsdaten/

G×E×M: Die Leistung einer Pflanzensorte wird – insbesondere bei komplexen Merkmalen wie dem Ertrag – nicht nur durch ihre genetische Zusammensetzung (G: Genotyp) bestimmt, sondern auch durch die Reaktion des Genotyps auf variierende Umweltfaktoren wie z. B. Klima- und Witterungsparameter (E: Environment) sowie auf Bewirtschaftungsverfahren wie z. B. Düngung oder Pflegemaßnahmen (M: Management). Die sehr komplexen Wechselwirkungen zwischen all diesen Faktoren, die sich auf die Leistung von Sorten auswirken, beschreibt man als G×E×M-Interaktion.

Heterogene Daten: Uneinheitlich zusammengesetzte Daten, z. B. aus Feldversuchen, bei denen verschiedene Stichproben von Genotypen in jeweils unterschiedlichen Umwelten (an verschiedenen Standorten oder in unterschiedlichen Jahren) evaluiert wurden.

Künstliche Intelligenz (KI): auch Artifizielle Intelligenz (AI); intelligentes Verhalten wird durch Algorithmen simuliert: Computerprogramme, die intelligentes Verhalten abbilden, sowie Automatisierung von Entscheidungsstrukturen. KI ist ein Teilgebiet der Informatik.

Kuratierung von Daten: Als Datenkuration bezeichnet man die Organisation und Integration von Daten aus verschiedenen Quellen. Um die Verwendung von Big Data zu ermöglichen, ist die Annotation, Veröffentlichung und Präsentation der Daten notwendig, sodass der Wert der Daten über die Zeit erhalten bleibt und die Daten für Wiederverwendung und Aufbewahrung verfügbar bleiben.
Weitere Informationen: https://de.qaz.wiki/wiki/Data_curation

Maschinelles Lernen (ML): Ist eine Methode der KI und beschreibt die künstliche Generierung von Wissen aus Erfahrung. Hier lernt ein künstliches System aus Beispielen und kann nach Abschluss der Lernphase und anhand von Algorithmen dieses Wissen auf unbekannte Daten übertragen (z. B. Klassifizierung von Nukleinsäuren).

Metadaten: Bei Metadaten handelt es sich um begleitende Informationen zu einem Datensatz, welche diesen erklärend beschreiben. Ein Beispiel ist eine Bildunterschrift, die erklärt, was auf dem Bild zu sehen ist. Die Beschreibung durch Metadaten bietet für alle Nutzer die Möglichkeit einer zielgerichteten Nutzung eines Datensatzes, da dessen Aussagekraft besser bewertet und die enthaltenen Informationen so vom Nutzer richtig interpretiert werden können.

Modellierung: Mathematische Modellierung mithilfe mechanistischer Modelle ermöglicht die Vorhersage und Analyse komplexer Prozesse auf Ebene der Gen-, Protein-, Metabolit-Interaktionsnetzwerke bis hin zu virtuellen Ganzpflanzen und Ökosystem-Modellen. Auf letzteren beiden Ebenen können sich genomskalige Stoffwechselmodelle, welche metabolische Stoffflüsse vorhersagen, und funktionell-strukturelle Pflanzenmodelle, welche die 3-dimensionale Struktur der Pflanze repräsentieren, sinnvoll ergänzen.

Multidimensionale Daten: Werte, die für unterschiedliche Eigenschaften/Merkmale von Objekten (z. B. Linien/Genotypen, Individuen, Organe) erhoben werden oder Daten, die nach vielfältigen Kriterien ausgewertet werden können. Für eine Gruppe von Objekten können z. B. genetische Markerdaten, Werte für unterschiedliche äußere Merkmale, oder molekulare Daten (z. B. Transkript-/mRNA- oder Protein-Werte) erhoben werden. Jede Datenreihe (Werte für eine bestimmte Eigenschaft bzw. ein bestimmtes Merkmal) stellt eine Dimension dar und die Objekte nehmen in dem aus den verschiedenen Datenreihen aufgespannten, multidimensionalen Raum unterschiedliche Positionen ein, aus deren Abständen z. B. Ähnlichkeiten zwischen den Objekten abgeleitet werden können. Ebenso können Beziehungen zwischen verschiedenen Merkmalen bzw. Eigenschaften aufgrund von ähnlicher Variation (Korrelation) oder räumlicher Nähe in dem von den Objekten aufgespannten multidimensionalen Raum abgeleitet werden. Für die gleichzeitige Analyse solcher komplexer Datensätze werden multivariate (statistische) Verfahren herangezogen, mit denen Strukturen in den Daten entdeckt oder geprüft werden können.

Neuronale Netze (KNN): Sind ein Zweig der Künstlichen Intelligenz. Nach biologischem Vorbild gebaut, geht es bei KNN aber um eine Modellbildung von Informationsverarbeitung und weniger um das Nachbilden biologischer neuronaler Netze und Neuronen. Die Anwendung von KNN beruht auf Problemen, bei denen kein oder nur ein geringes explizites Wissen vorliegt (z. B. Spracherkennung, Gesichtserkennung).

Omics: Der Begriff ‚Omics' bzw. die Nachsilbe -omik (engl.: -omics) bezeichnet, insbesondere in den Lebenswissenschaften, vollumfängliche Datensätze in einem ganzheitlichen, systemorientierten Ansatz (Systembiologie) und gliedert sich in verschiedene Teilbereiche wie z. B. Genomik, Epigenomik, Transkriptomik, Proteomik, Metabolomik usw. Hier werden Daten des Genoms (Gesamtheit aller Erbinformationen eines Organismus) oder des Transkriptoms (alle von DNA in RNA umgeschriebenen/transkribierten Erbgutbereiche d. h. die Gesamtheit aller vorliegenden RNA-Moleküle) erhoben und analysiert. Entsprechend: Epigenom (Gesamtheit aller epigenetischen Erbgutmodifikationen), Proteom (Gesamtheit aller vorliegenden Proteine), Metabolom (Gesamtheit aller vorliegenden Metabolite = Stoffwechselintermediate) usw.

Orthogonalität von Daten: In der Faktoranalyse sollen die Anzahl der Variablen verringert und ähnliche Variablen gefunden werden. Dabei werden die Daten zur Reduktion der Komplexität in orthogonale Dimensionen eingeteilt. Insgesamt soll durch eine Reduktion der Variablen eine Reduktion der Komplexität des Modells erreicht werden.

Repository-Manager: Ein Repository (englisch für Lager, Depot oder auch Quelle) ist ein verwaltetes Verzeichnis zur Speicherung und Beschreibung digitaler Objekte für ein digitales Archiv.

Gemeinschaft zur Förderung von Pflanzeninnovation e. V. (GFPi)
Büro Bonn
Kaufmannstraße 71
53115 Bonn
Telefon +49 228 98581-40
Telefax +49 228 98581-19
E-Mail gfpi@bdp-online.de
www.gfpi.net

GFPi-EU-Büro
47–51, rue du Luxembourg
B-1050 Brüssel
Mobil +49 172 2643357
E-Mail gfpi-fei@bdp-online.de

Mitglied der

Forschungsnetzwerk Mittelstand

Gemeinschaft zur Förderung von Pflanzeninnovation e. V. (GFPi)

Lebensbasis Pflanze