Data Science and Engineering to Support Molecular Breeding Pipelines and Summary

Bayer Russia Molecular Marker Training

26 July 2023

Bayer Russia Molecular Marker Training:
Data Engineering and Training Summary

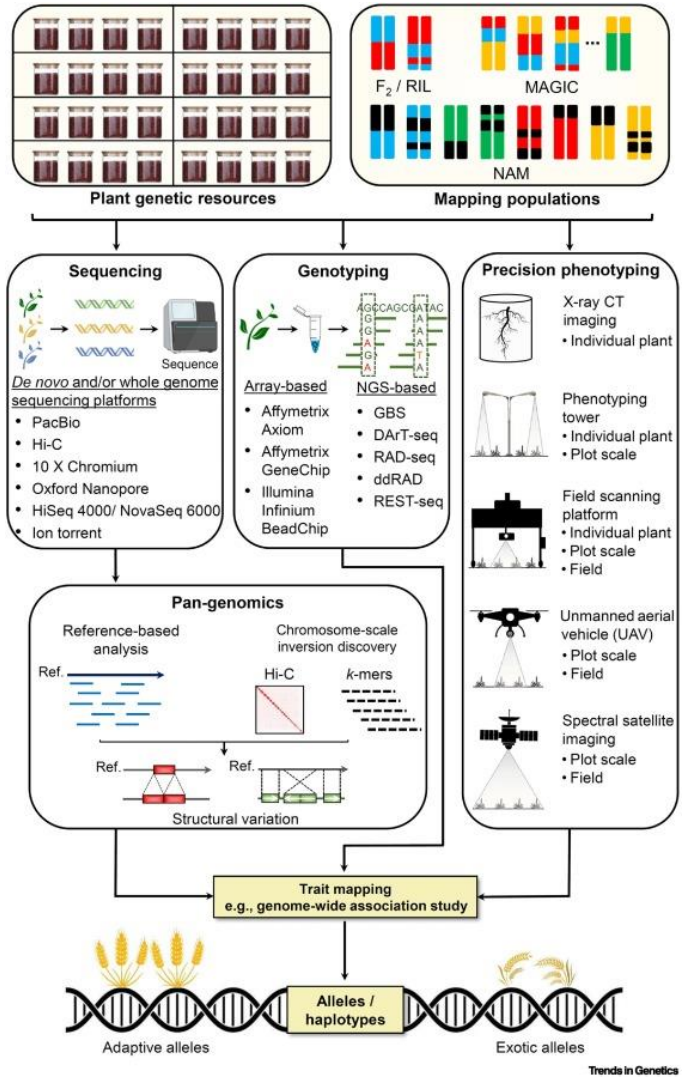
1. **Laboratory Data Infrastructure**
2. **Data Quality**
3. **Automated Analytics**
4. **Modernized methods for computation**
5. **Take-aways and Outlook for Molecular Breeding**

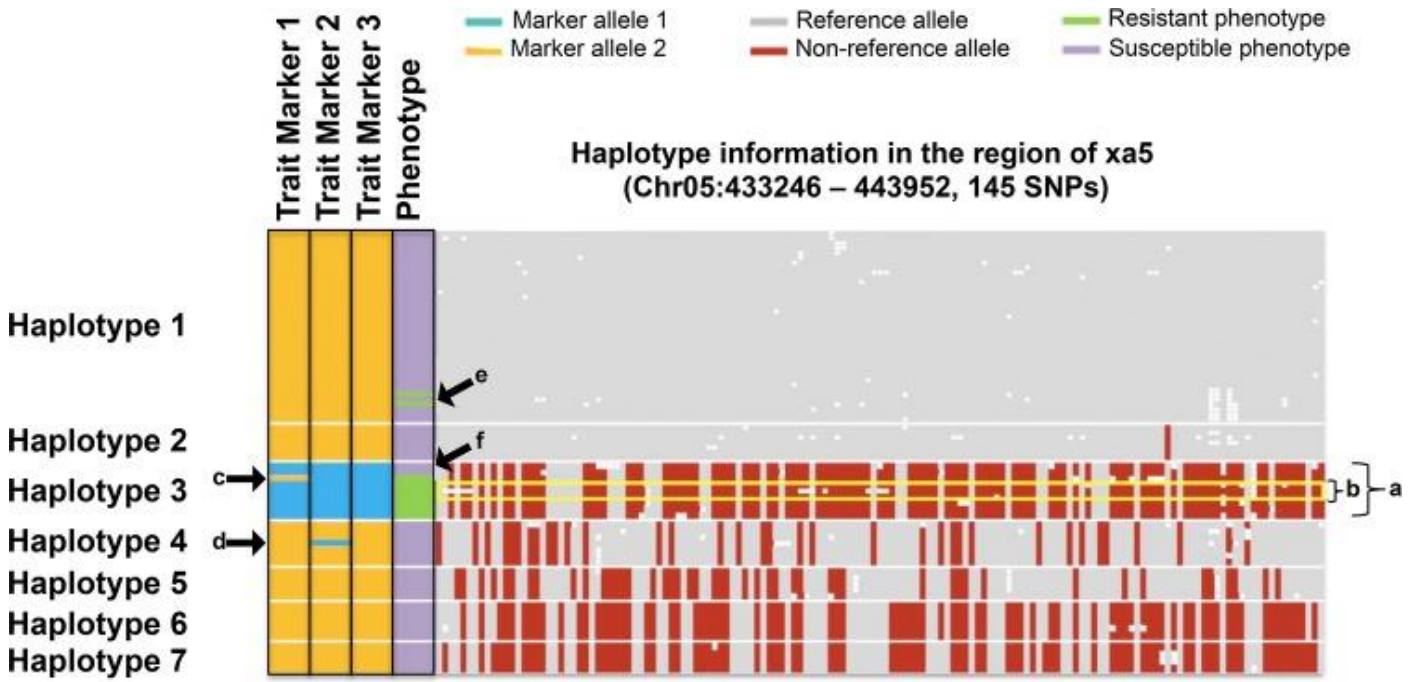
The breeder's equation:
Molecular breeding has positive impact to V and T

$$\text{Genetic gain} \; G = \frac{\text{Genetic variance} \; V \times \text{Phenotyping accuracy} \; A \times \text{Phenotyping precision} \; P \times \text{Selection intensity} \; I}{\text{Cycle time} \; T}$$

Breeders have been successful whenever they had access to useful **genetic variation** and selection has focused on the **right traits** measured with the **right protocol** in the **right environment**

Successful Plant Breeding Relies on Integrated Genotyping for Driving Populations





Cobb, Joshua N et al. "Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation." *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* vol. 132,3 (2019): 627-645. doi:10.1007/s00122-019-03317-0

Varshney, Rajeev K et al. "Fast-forward breeding for a food-secure world." *Trends in Genetics vol. 37,12 (*2021) https://doi.org/10.1016/j.tig.2021.08.002

Genetic Data Pipeline

Laboratory Data Infrastructure

Stable and Flexible Architecture

- Database / Data Asset
- Event Streaming
- API
- Analytics
- Platform

Project Preparation

- Create Cohorts, Entities, Roadmaps
- Sample Submission, Project Creation
- Chipping

Observe Genotypes

- Genotyping
- Observed Genotype Storage

Establish Entity Connections

- Project Status
- Entity Connectivity

Ensure Quality

- QC

Genotype & Phenotype

- Imputation
- Imputed Genotype Storage
- Phenotype Storage

Select the Best

- Genetic Evaluation
- Selection
- Offloading

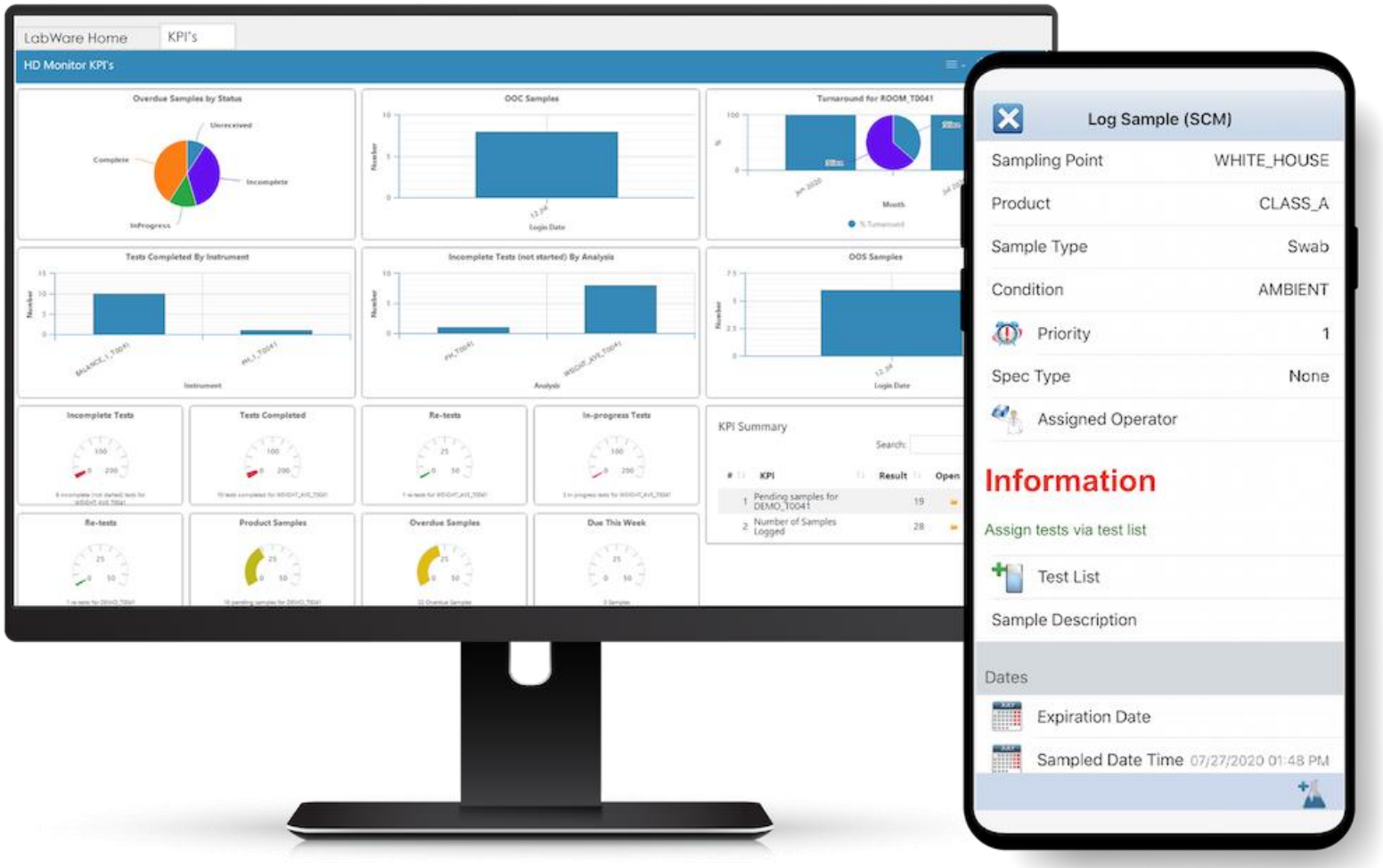
Laboratory Information Management Systems (LIMS)

Laboratory Data Infrastructure

- What do LIMS do?
  - Streamline complex workflows of multiple laboratory instruments
  - Manage, track, and document flow of samples and metadata
  - Automating routine lab tasks

- Benefits:
  - Improve efficiency and accuracy of sample processing
  - Ensure regulatory compliance
  - Reduce errors in data handling
  - Improve data traceability

- Challenges:
  - Connecting both the nursery and the laboratory require shared IT infrastructure



https://www.labware.com/blog/labware-8

Genomics Data Assets

Laboratory Data Infrastructure

// Marker Resource Collections

Markers/Marker Sets
Named genomic features

Marker Call Sets
Observations of genotypes at markers

Genetic Maps
Ordering and distance between markers relative to a reference

Genotypes
Probabilistic genotypes, synthesized from observations and inferences

Identity By Descent
A measure of genetic similarity

Genomics

// Sequence Resource Collections

Reference Genomes
Assembled sequence for a pipeline entity

Sequence Reads
Sequencer reads and aligned reads in standard bifx file formats

Variant Sets
A set of locations where the individual differs from the reference

Interval Sets
Genomic/Genetic intervals with annotations. QTLs, Haplotypes, etc.

Provenances
Source of a dataset. Maps to observation method, algorithms, etc.

// Formats

FASTA

FASTQ / BAM

VCF / GVCF

GFF3

Laboratory Data Quality Control

Data Quality

Primary QC
Aggregate QC

- Missing percent
- Het percent
- PO Percent
- Informative PO

- Number of Missing Markers
- Het Percent Monomorphic Markers
- Het Percent
- Missing percent
- PO percent Monomorphic Markers
- PO Percent
- Relative GP Similarity

Plate QC

Remediate?

Marker QC

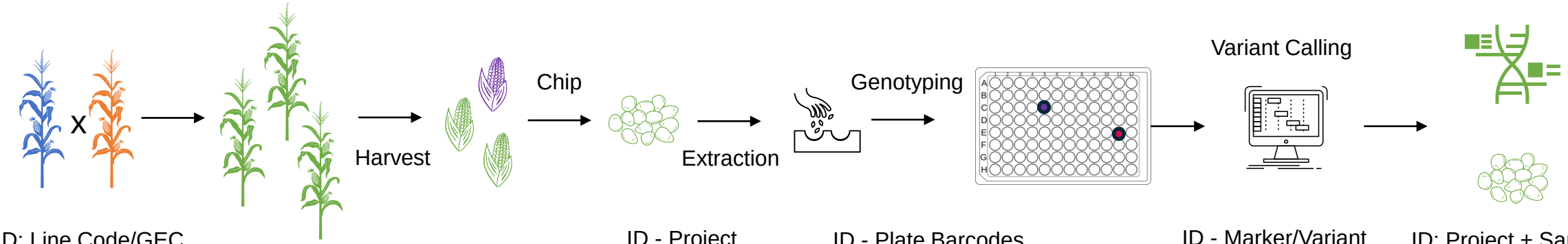
Inactivate Marker?

Sample QC

Redo? Remediate? Drop?

Breeding Origin

Lab Project



ID: Line Code/GEC

ID - Project

ID - Plate Barcodes

ID - Marker/Variant

ID: Project + Sample

- Passed Markers
- Passed Samples
- PO percent
- Non-missing samples percent
- Marker Consistency
- Requested Samples

Project QC

Remediate?

Importance of Laboratory Quality Control

- Accuracy – decrease errors in complex lab workflows for accurate selections
- Consistency – ensure reliability of genotypes to compare across labs and instruments
- Compliance – ensure high regulatory standards to avoid penalties
- Trust – ensure reliability of genotyping
- Data Interpretation – decrease time spent from data scientists interpreting genotypes

Sequencing QC and Remediation

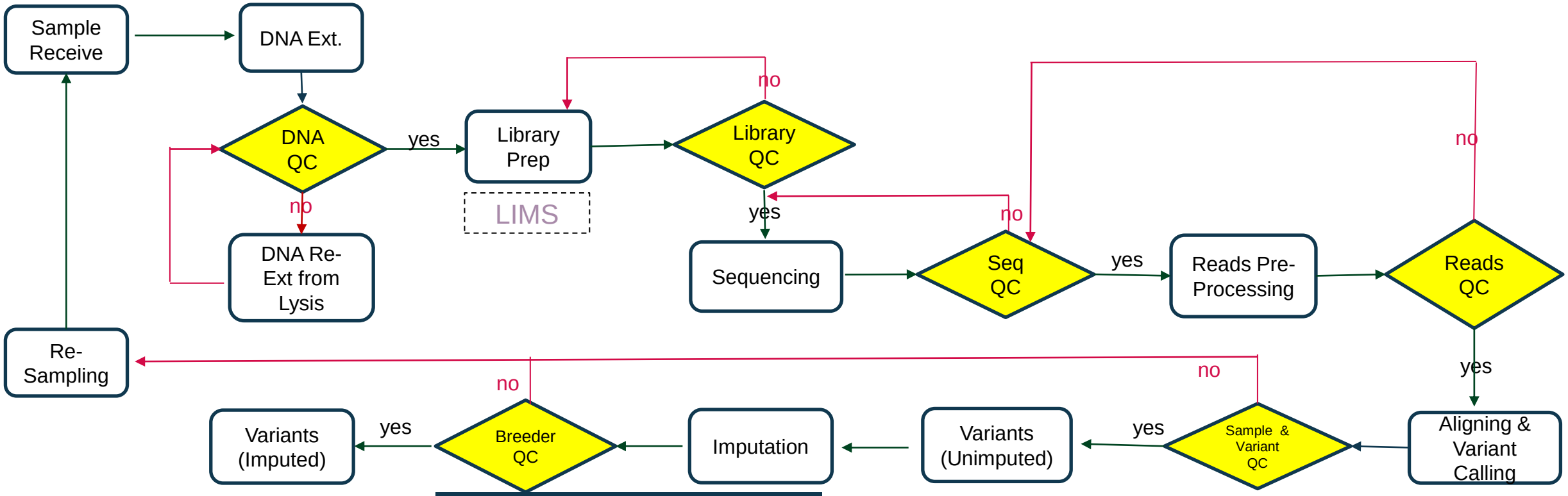
Data Quality

DNA QC: Picogreen
- DNA conc threshold
- 90% samples > threshold

Library QC:
- Bioanalyzer for con. and size
- Pico again for con. before pooling

Seq QC in seq run:
- Total # of reads
- Quality of reads
- Read length

Reads QC:
- # of reads in each sample



Breeder QC:
- Imputed PO Check
- How imputed samples are related to phased panel
- How imputed haplotypes are related to ancestral haplotypes

Sample & Variant QC:
- # of uniquely mapped reads
- Number and coverage of variants called
- Unimputed PO check
- How samples are related to ref panel

Laboratory Data Quality Control Best Practices

Data Quality

Challenges of Maintaining Data Integrity:

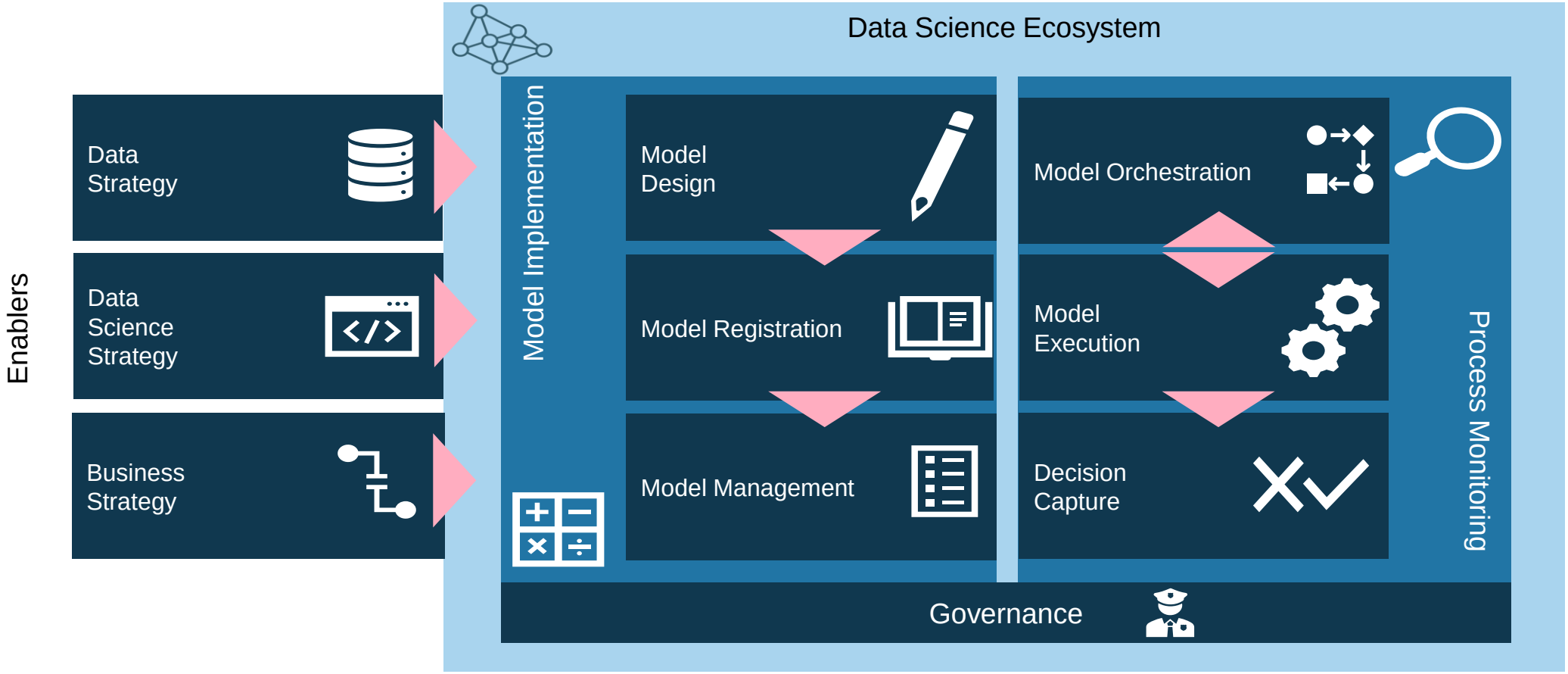
- Contamination/Swaps - Occurs in both the lab (nearby wells) and the field (nearby plots)
- Standardization – Variations in labs, personnel, result in inconsistent results or errors
- Technological Advancement – Continuously update procedures and equipment
- Scalability – Large volumes of sample data often result in unexpected data failures

Best Practices:

- SOPs (Standard Operating Procedures) – Comprehensive SOPs for every lab data workflow to be reviewed frequently
- Quality Assurance – Internal Audits, Proficiency testing, and method validation ensure reliability
- Documentation – Data, observations, and results should be recorded and organized.
- Controls – Using positive and negative controls help determine data expectation
- Error Management – Clear Procedures for identifying and reporting errors and recommend correction to prevent reoccurrence
- Review - Checks should happen as early as possible to prevent erroneous data from ingested to affect downstream decisions
- Data Management – Data collection, security, organized storage, retrieval, and backup systems

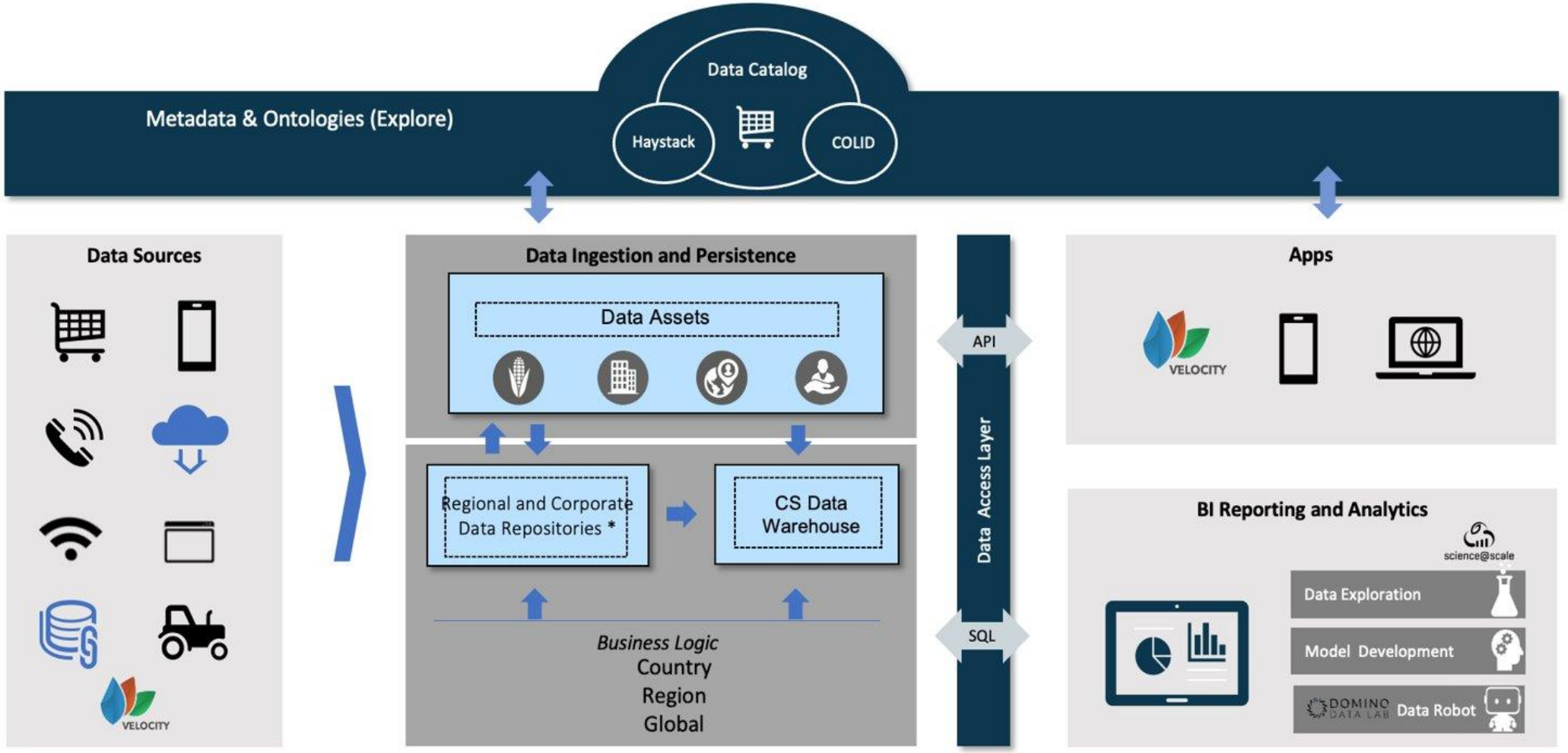
Data Science Ecosystem

Automated Analytics



Data Strategy

Automated Analytics



Data Management and Governance

Automated Analytics

- Data Ownership: Define data ownership and establish guidelines regarding who has rights and responsibilities for different data types (from data collection to analysis)
- Data Security and Access Control: Protecting sensitive data from unauthorized access, data breaches, or data loss. Practices: Authentication, role-based access control, encryption, secure storage
- Data Privacy, Confidentiality, and Compliance: Comply with regulatory standards by country to protect the confidentiality of propriety data. Practice: Anonymization techniques, restriction of data across international boundaries
- Data Integration and Interoperability – Practice: Building standard data formats and data sharing agreements
- Data Lifecycle Management – Define timelines for data collection, storage, retention and disposal. Practice: backups and disaster recovery
- Metadata and Documentation - Standardize documentation on both code and data to provide descriptions for datasets, features, and workflows. Should happen at both the code level as well as the project level. Facilitates easy discovery

Data Science Lifecycle

Automated Analytics

IT Steering Teams

Technical teams, project owners

End users, deployment teams

IT partners, developers, engineers

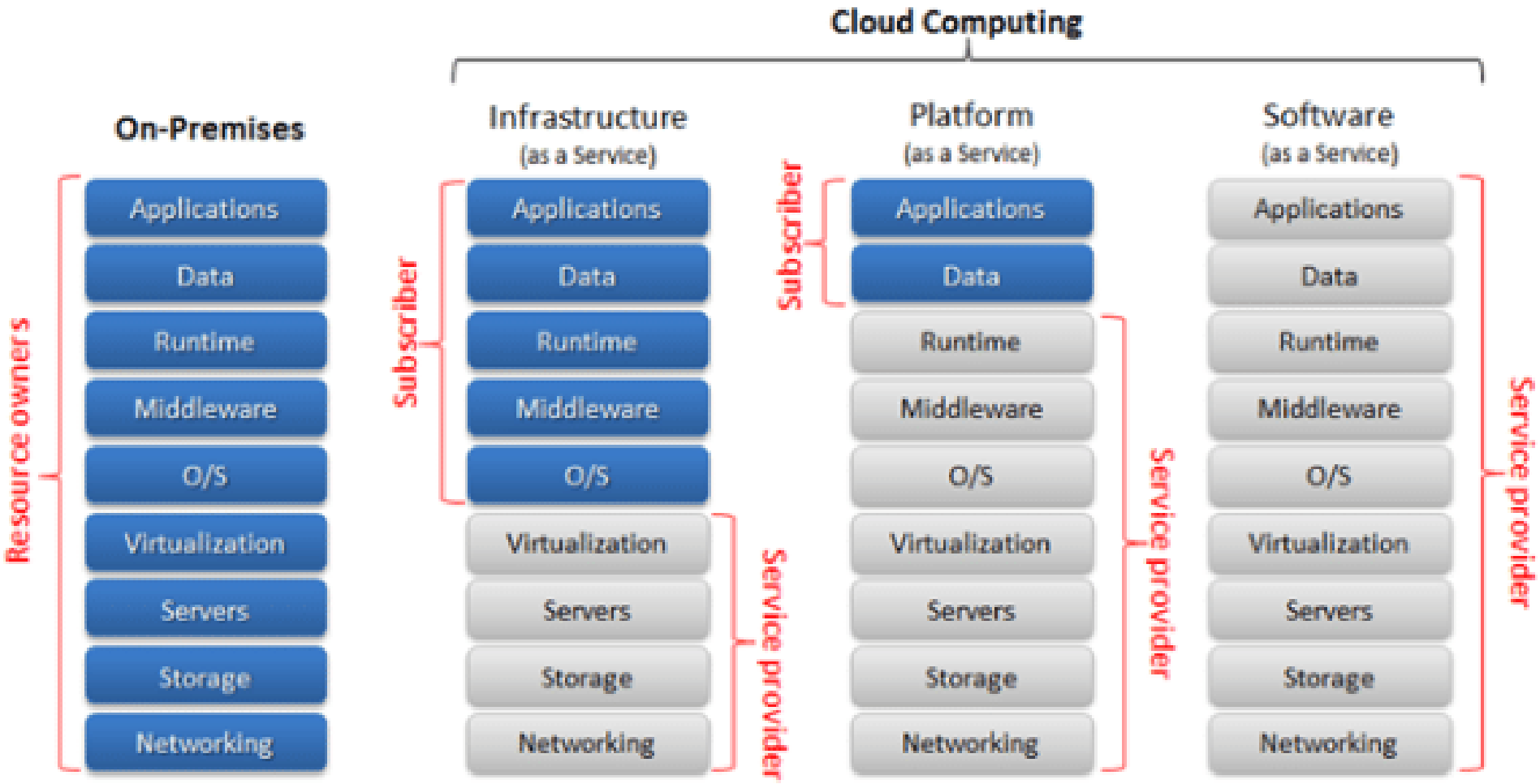
Discovery: Planning	Phase 1: Testing/POC	Phase 2: Validation	Phase 3: Deploy	Phase 4: Expansion

// Business value defined
// Centralization roadmap, resourcing estimated
// Local/Enterprise IT stakeholders informed
// Metadata requirements defined and documented
// Testing planned, sandbox users informed
// Code in repo, versioning/semantics in place
	// Build, execute and assess prototype
	// Execute alpha testing
	// Re-assess business value, resource needs
	// Notebook with tagged code, data versions.
	// Data passes standard format checks
	// Formalize pathway to democratization
	// Sandbox -> Data Lake
		// Refactor resourcing for increases in volume, edge cases and new features
		// Execute Beta testing
		// Process/code review
		// Key stakeholder review
		// Develop change management plan
			// User training/knowledge transfer
			// Launch and deploy
			// Success metrics reported
			// Technical Support and maintenance
			// Documentation complete and made accessible to community
			// Automated QC implemented
				// Expand to different regions, crops, teams
				// Automated QC improved
				// Share progress with stakeholders
				// Monitor and optimize process changes for continuous improvement

On Premise vs Cloud Deployments

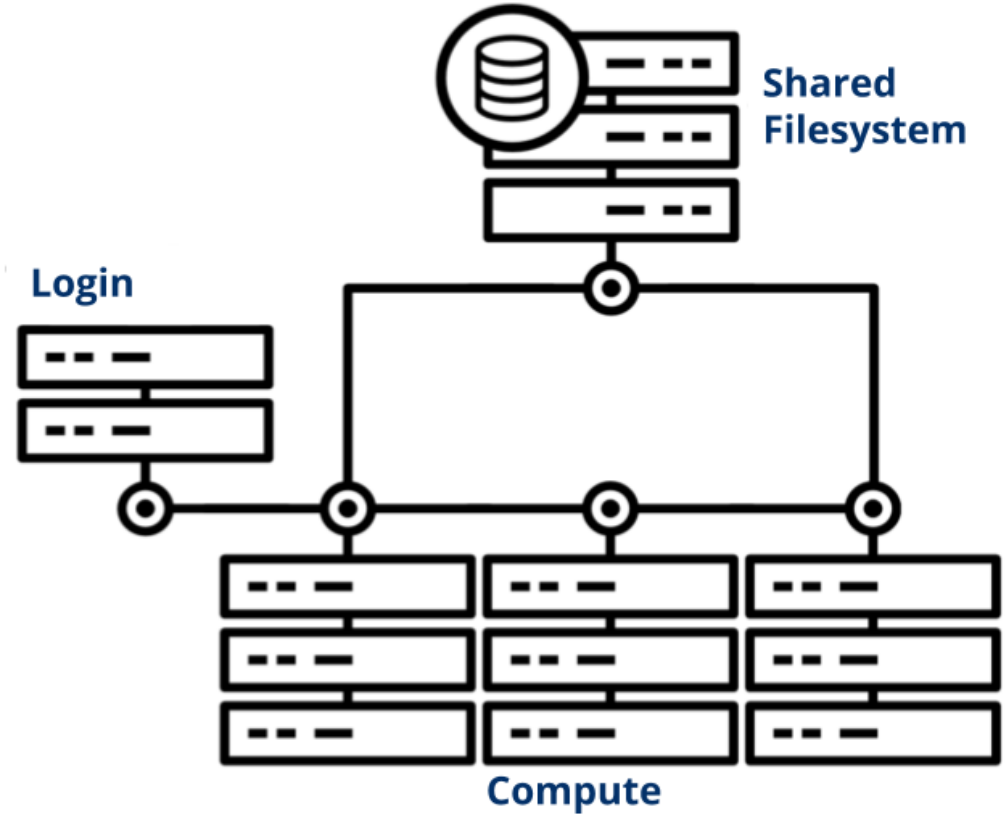
Automated Analytics

Cloud Environments:
- Yandex Cloud
- MTS.cloud
- Selectel



Model Execution – Local Deployment

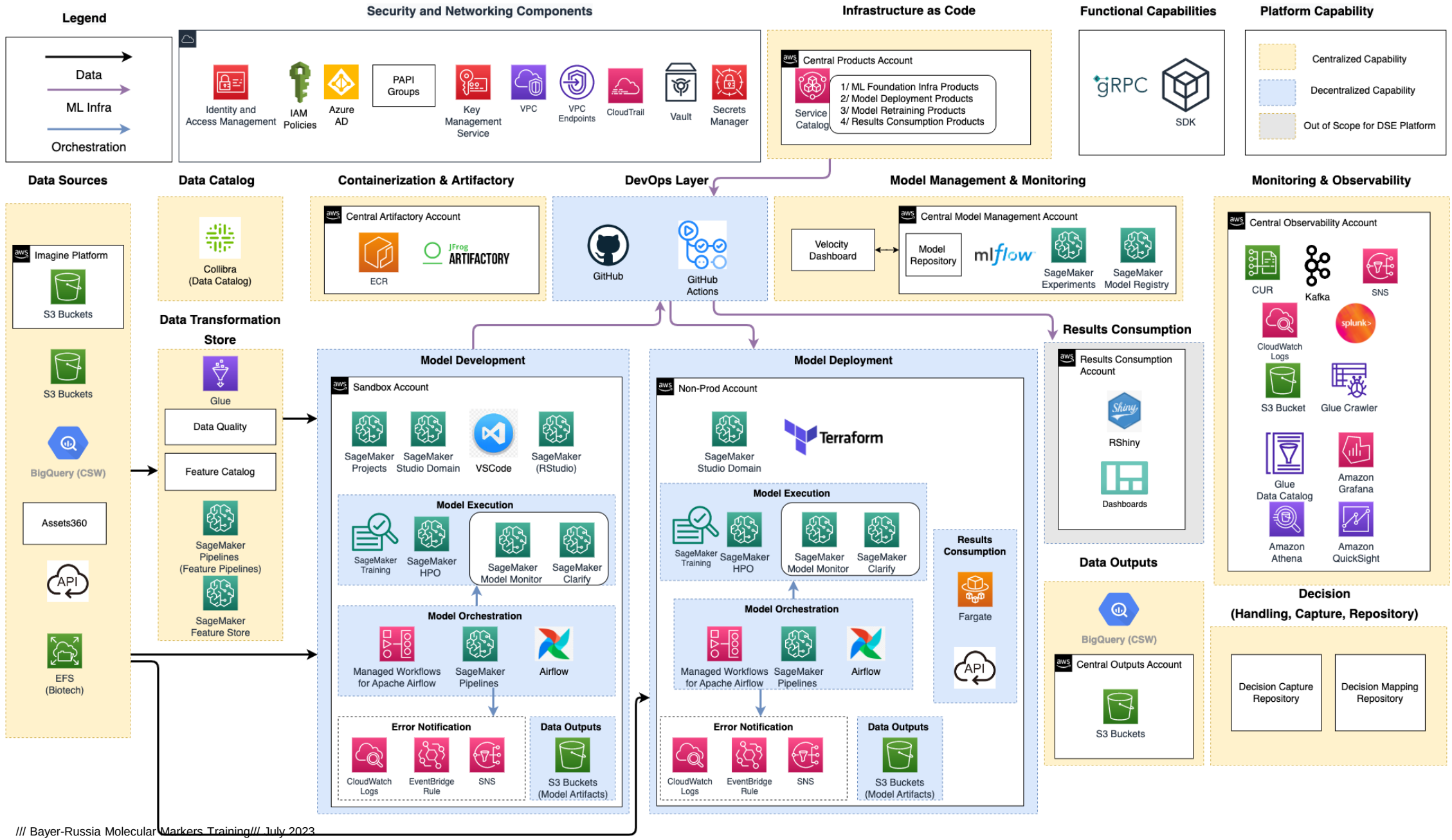
Modernized Methods for Computation



https://docs.ycrc.yale.edu/

- High Performance Computing (HPC) Clusters: linked computers (nodes) that work together to perform complex computation
- Each node operates like an individual computer but interconnected to function as a single system
- Deliver speed for processing large volumes of data by dividing into smaller, independent tasks simultaneously distributed across nodes
- Usually deployed as an on-premise solution, however, concepts extend to cloud

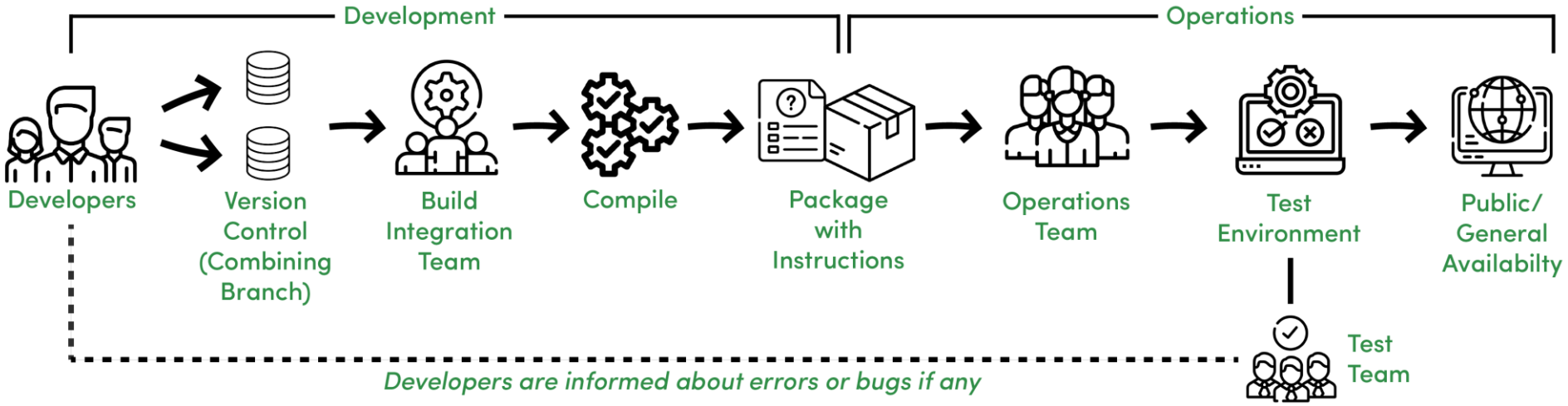
DECISION SCIENCE ECOSYSTEM - HIGH LEVEL ARCHITECTURE



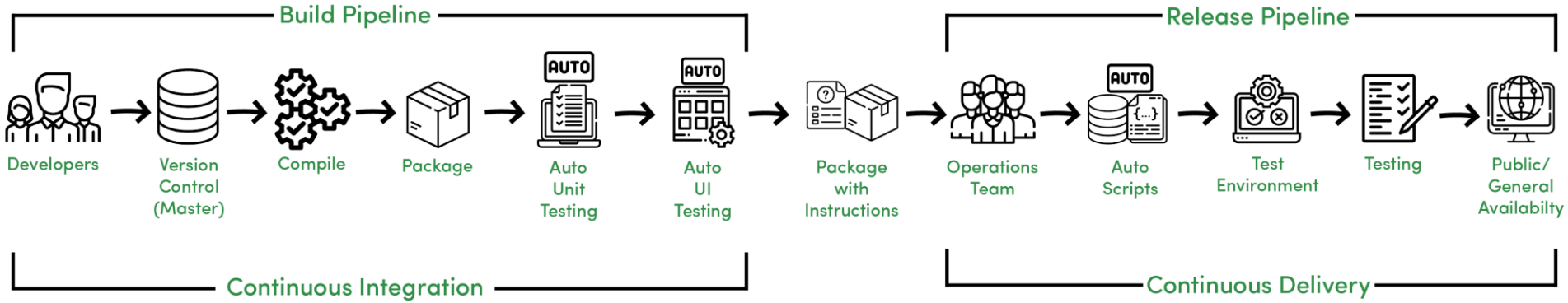
Model Implementation - Software Engineering Best Practices

Modernized Methods for Computation

Traditional Software Development Process



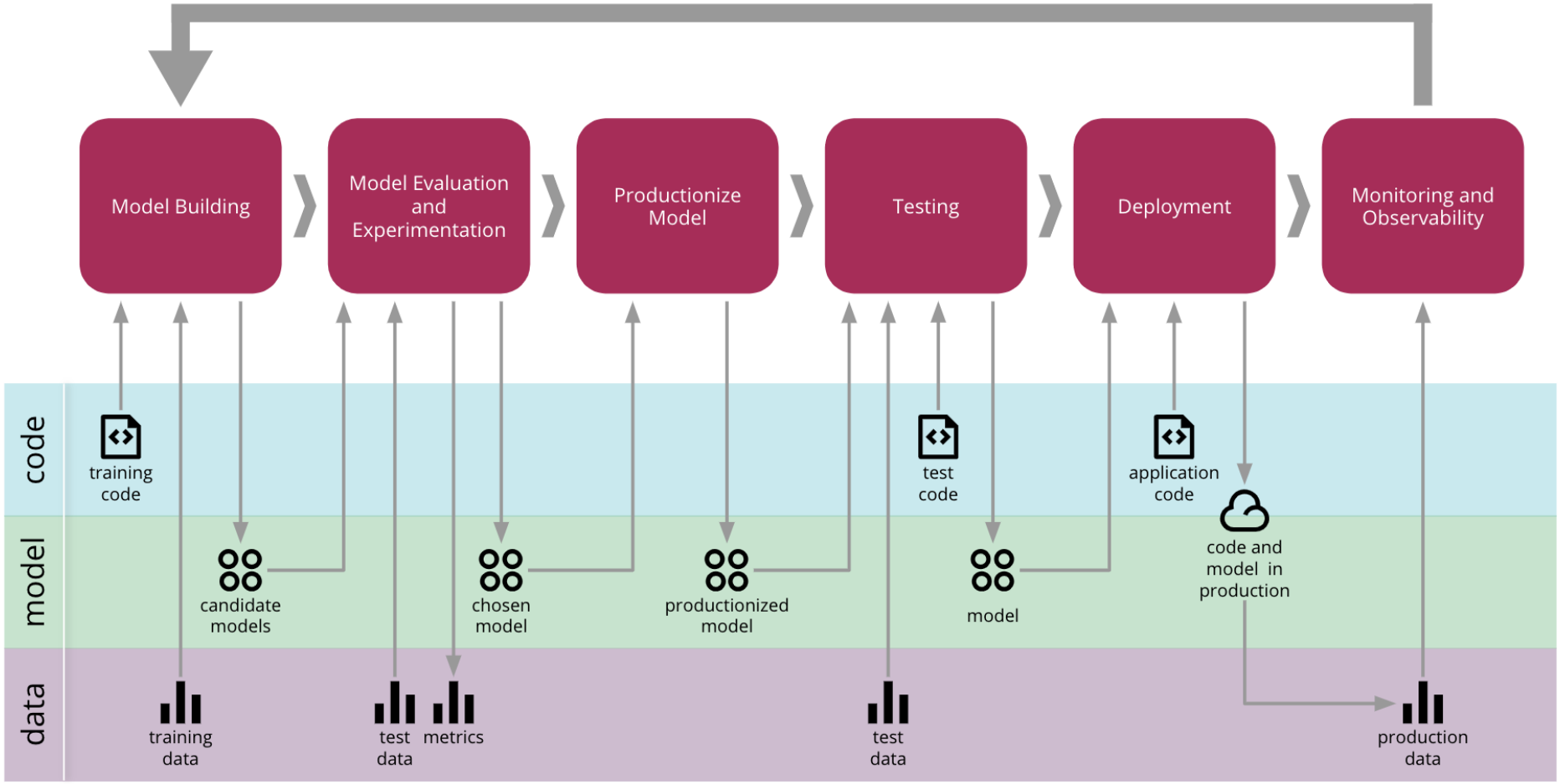
Continuous Integration / Continuous Delivery



https://www.geeksforgeeks.org/ci-cd-continuous-integration-and-continuous-delivery/

Model Implementation – CI/CD for ML and AI

Modernized Methods for Computation



https://martinfowler.com/articles/cd4ml.html

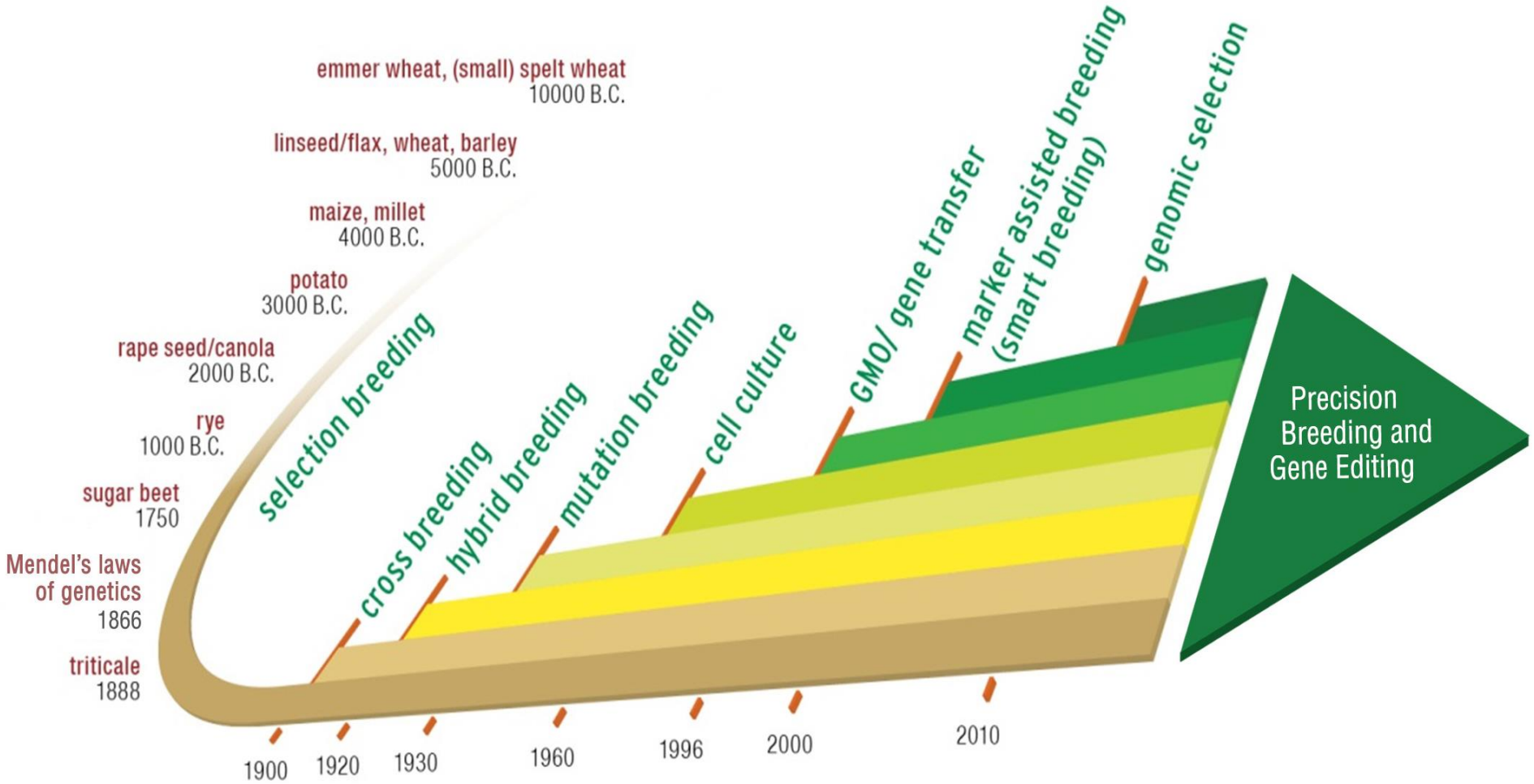
Q&A
Discussion

Bayer Russia Molecular Marker Training:
Data Engineering and Training Summary

1. Laboratory Data Infrastructure
2. Data Quality
3. Automated Analytics
4. Modernized methods for computation
5. **Take-aways and Outlook for Molecular Breeding**

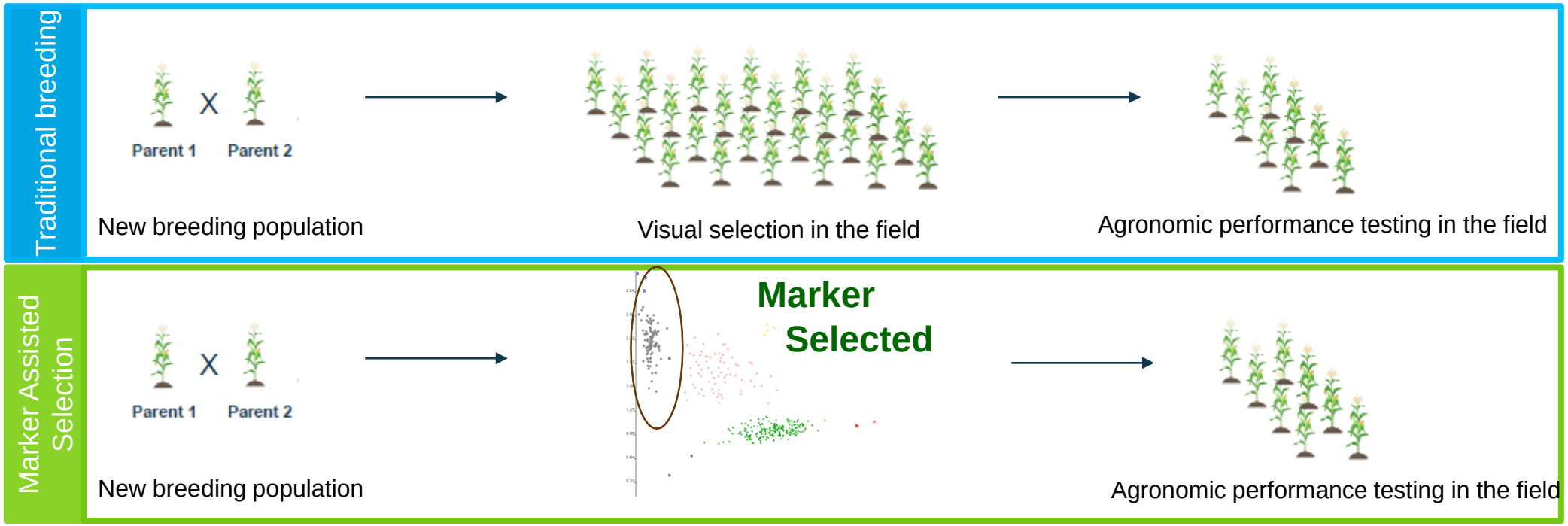
Molecular breeding is an important advancement in product development

// Agriculture has a long history of finding new ways to make the improvements that farmers and society ask of the industry



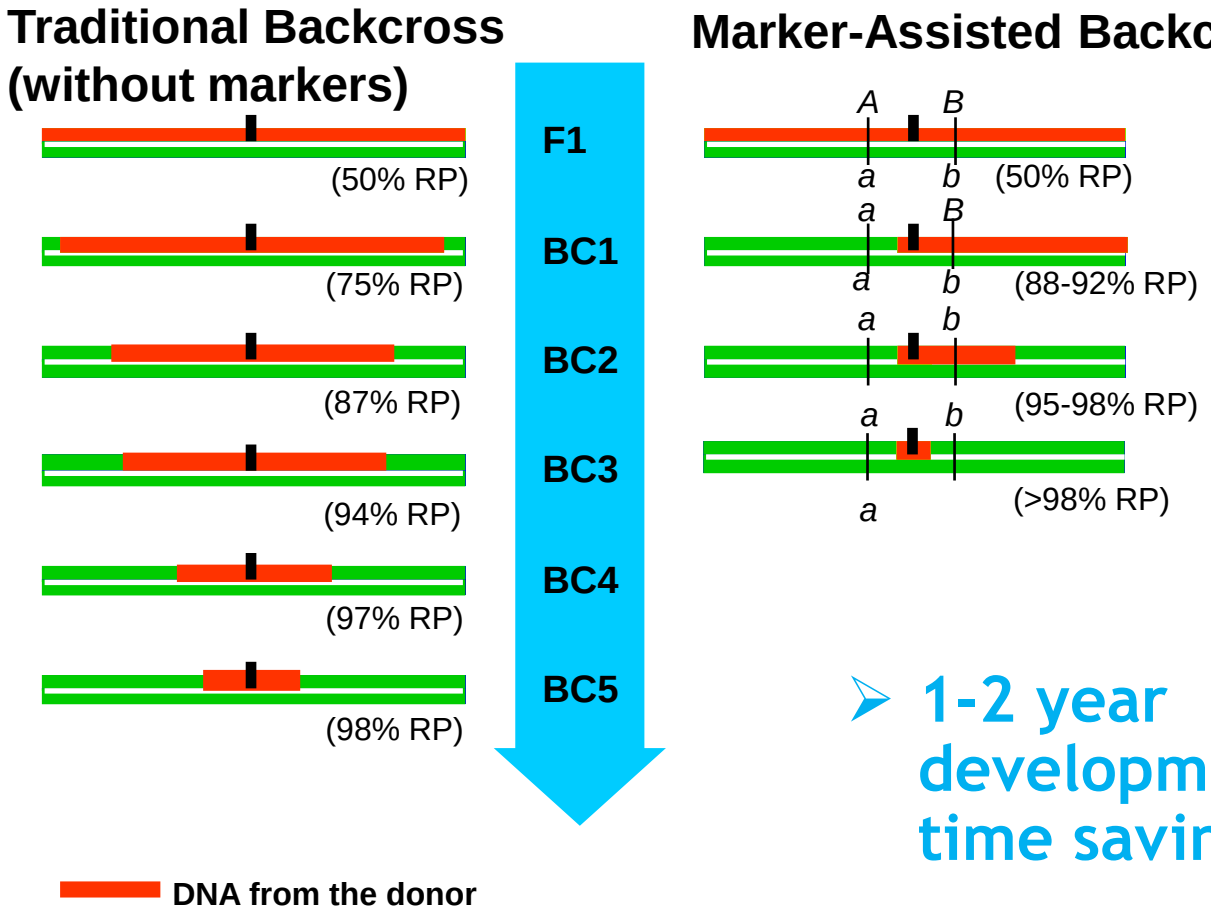
Marker Assisted Selection (MAS)

- MAS is used to improve the accuracy, speed or costs of selecting for agronomic traits which are under simple genetic control (1-3 genes, require associating a trait with a closely-linked molecular marker)
- DNA characterization replaces laborious, costly or inefficient phenotypic screens for the trait in breeding programs



Marker Assisted Backcross (MABC)

Molecular Markers used to identify “elite” vs. “donor” DNA



1. Select for the donor target gene/region
2. Select for progeny that have crossovers near the desired trait gene =>Introgression of reduced donor DNA segment
3. Very fast recovery of recurrent parent

- **1-2 year development time savings**

What is Genome Wide Selection (GWS)?

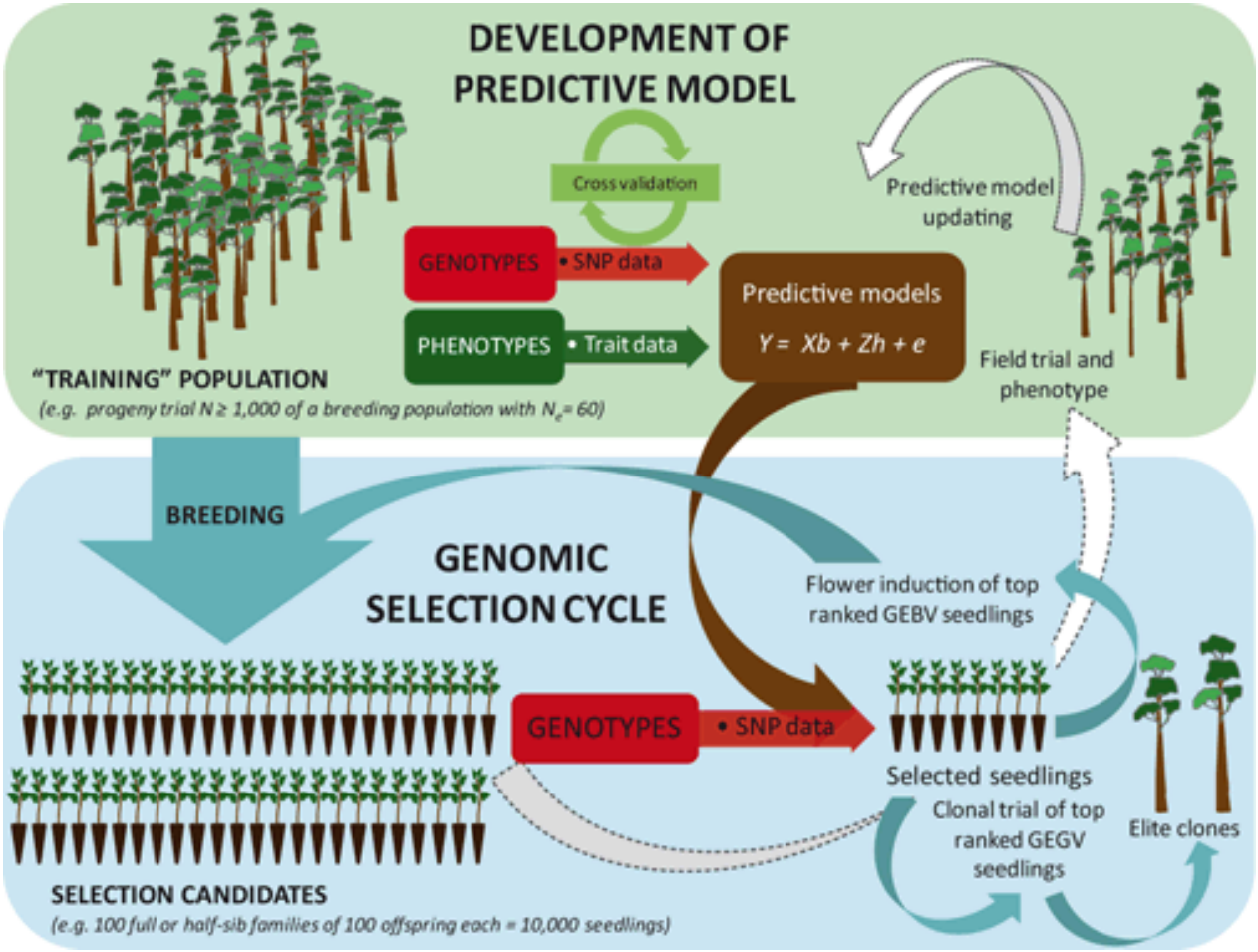
Selection based on genomic estimated breeding value (GEBV)

GWS

Genome wide selection or genomic selection is the selection using genetic evaluation from the whole genome marker-based model.

Pros: improve genetic gain, save resources, works well with complex traits contributed by many small effects

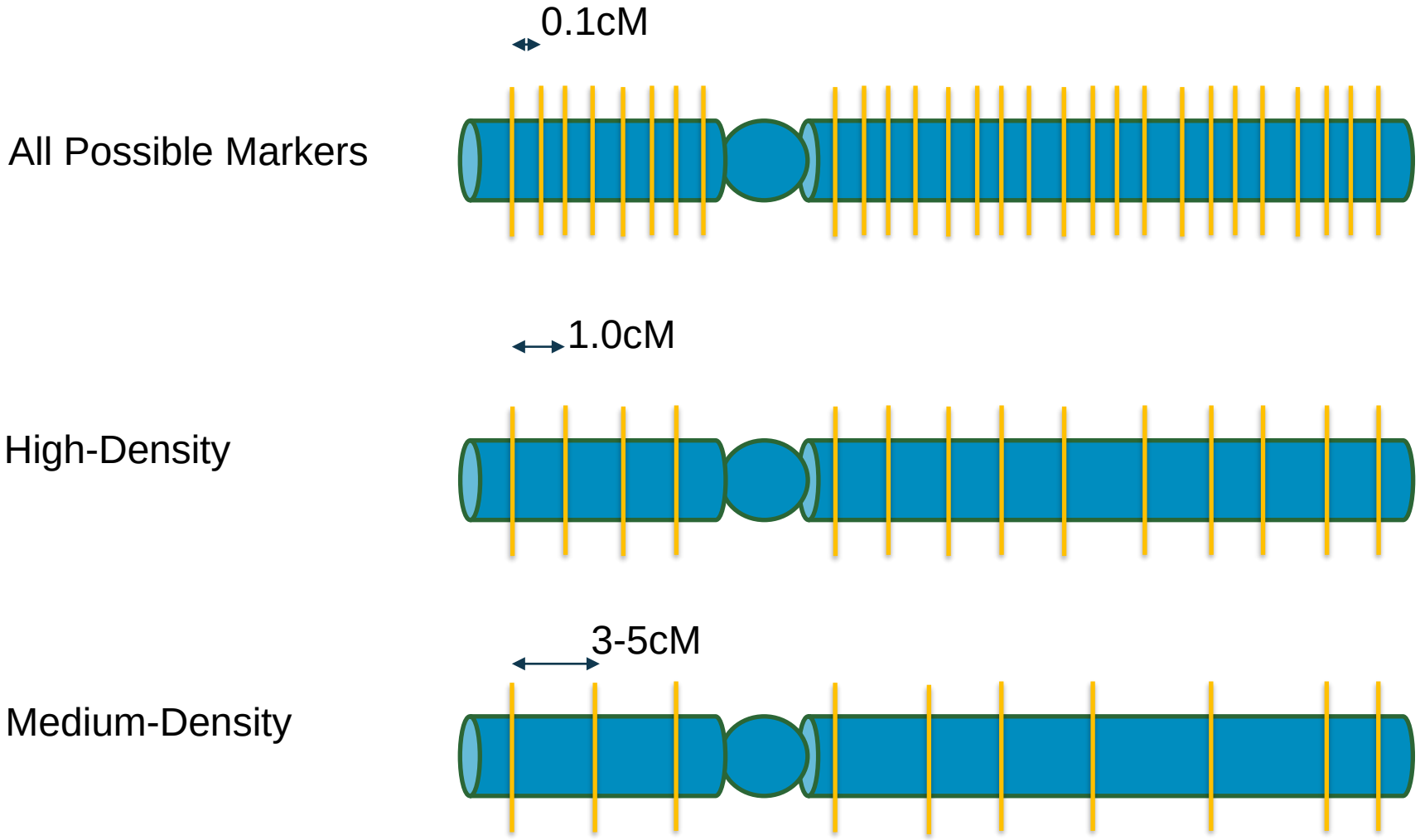
Cons: model accuracy depends on the relationship between training set and prediction set, reduce diversity quickly



Source: https://doi.org/10.1007/978-94-007-7572-5_26

Selecting the minimum number of markers to maximize genetic map coverage

Select Markers



- The exact amount of spacing is decided by breeding stakeholders and optimized by genome size and recombination rate
- If gaps appear, they can be filled by gap-filling algorithms

The breeder's equation:
Molecular breeding has positive impact to V and T

$$\text{Genetic gain} \; G = \frac{\text{Genetic variance} \; V \times \text{Phenotyping accuracy} \; A \times \text{Phenotyping precision} \; P \times \text{Selection intensity} \; I}{\text{Cycle time} \; T}$$

Breeders have been successful whenever they had access to useful **genetic variation** and selection has focused on the **right traits** measured with the **right protocol** in the **right environment**

Portfolio of genomics platforms support Bayer's row crop and vegetable breeding programs

Row Crops

Genomics Platform	Application
Single Marker Taqman	Trait integration, QAQC, Purity, Trait verification
Medium density Genotyping by Sequencing (GBS)	Genetic evaluation, Discovery
High density Fingerprinting (FP)	Origin design, Discovery, Haplotype
Resequencing	Novel polymorphism discovery, Haplotype

Vegetables

Crops may use all or some of the genomic platforms depending on breeding objectives, complexity of genome, and foundational genomics

Q&A
Discussion

DISCLAIMER

THE INFORMATION CONTAINED HEREIN IS EXPERIMENTAL IN NATURE AND IS PROVIDED "AS IS". BAYER MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO THE MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE INFORMATION WILL NOT INFRINGE ANY THIRD-PARTY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS.