CGIAR
Science for a food-secure future

AgPile

Data Ecosystem for AI Enablement in Ag

Jan 13 2025

CGIAR
INITIATIVE ON
Breeding Resources

Agpile Project Project timeline

- 2024 Gates Foundation funded project for data ecosystem evaluation
- 2024 Accenture – Completed Current State Review, Initial Scoping and Design, Architecture Recommendations
- 2025 Product Team assembly & MVP version of Agpile
- 2026+ TBD

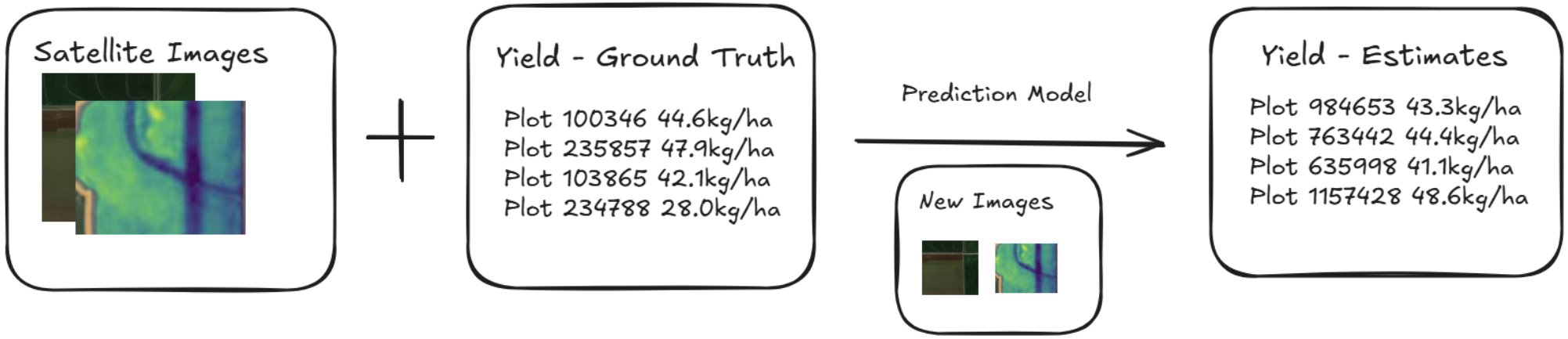
Agpile Project

Mission:

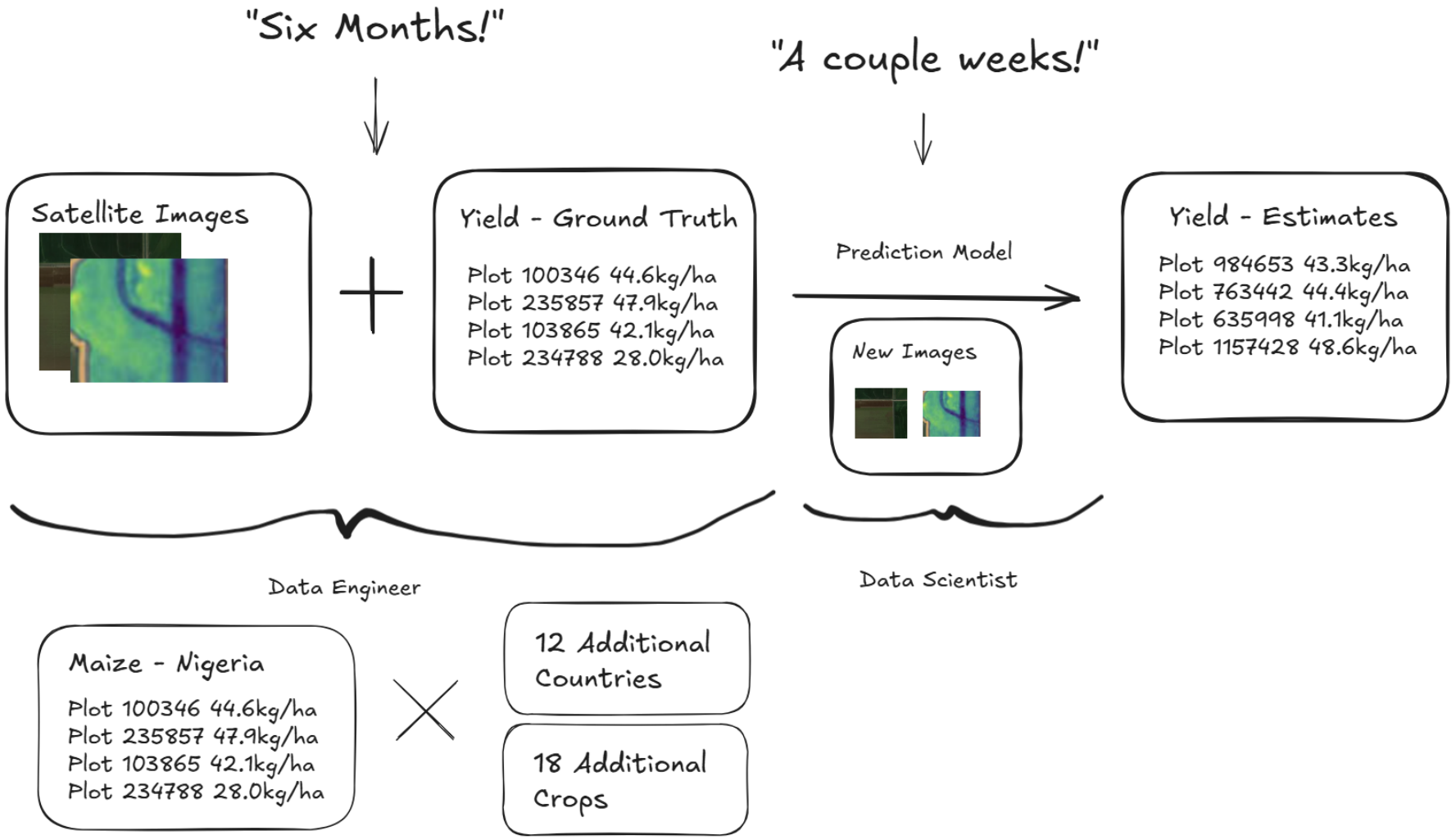
Provide the ability to curate, federate and steward disparate data sources to enable AI and other data products.

Background: Example of Developing a Data Product

Example Project – Use Satellite imagery to Estimate Crop yield



Example of Developing a Data Product



Data Landscape

Why is it challenging to get the data needed?

- Cross organizational boundaries
- Complex data governance rules
- Data silos within silos
- Limited tooling for hosting
- Limited incentives for sharing/reuse

Goal:

F findable

A accessible

I interoperable

R re-usable

Agpile Project

Getting to FAIR

F findable

A accessible

I interoperable

R re-usable

Key Capabilities

- Search
  - Catalog/Registry
  - Metadata Storage
- Secure
  - Federated Identity Providers
  - Access Controls
- Standard APIs
- Storage for Raw and Intermediate data

Why it matters

- Data re-use; especially important for data collection like trials that require a growing season

- AI enablement

- Hidden Cost of untracked data assets

- Governance enablement & Auditing

Approach

- Avoid “Bridge to Nowhere”
- Use cases for willing partners
- Buy/use existing solutions for use cases
- Use cases to serve as templates for additional projects
- MVP – platform enables FAIR and AI ready data; reduces the first three steps in a typical data analysis pipeline [locating, contextualizing, refining]
- Focus on enabling self-service
- Snowball effect of adoption

Use cases

- Artemis –
  - Computer vision applied to field collected images for collecting phenotype data
- Tumaini
  - Computer vision applied for disease identification

- Operational Breeding Data (combine historic breeding trial data with data hosted in different operational systems like Breedbase, BMS, EBS)

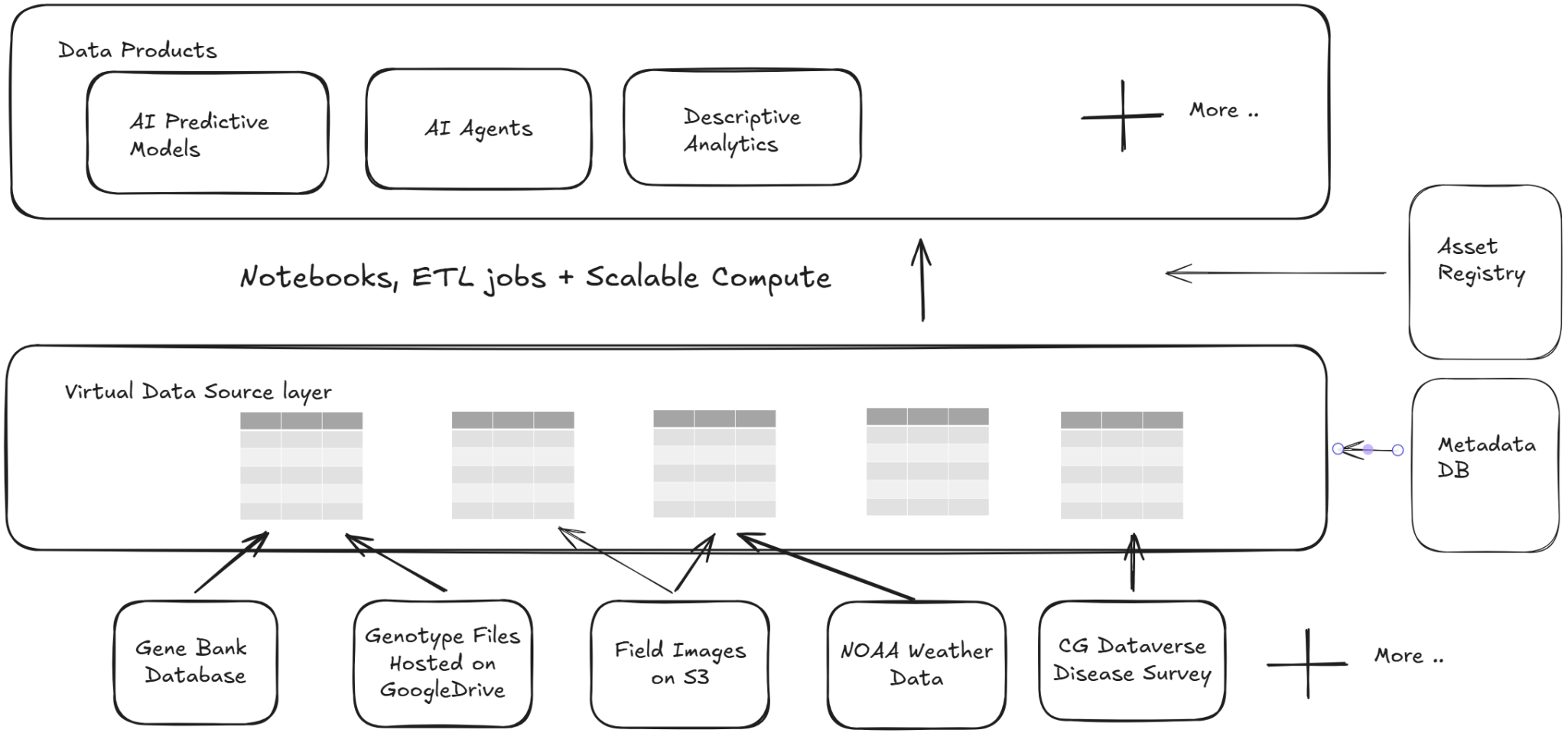
- Will give us us different ‘templates’ for future use cases
  - Artemis and Tumaini – image data
  - Breeding Data – tabular data

Leveraging AI within the platform

- Use existing AI tools to help ingest data into the platform – for example to classify, detect PII, derive metadata

- Store AI/ML workflows for reuse as part of the platform

Concept Diagram



Agpile Project

“Data access and sharing, as well as data interoperability and reuse, have often been approached as a philosophical good rather than an operational need for effective research and translation. This is no longer consistent with the current fast moving AI landscape – large collections of data can be leveraged for insights and tools that go far beyond the goals of the original projects that generated that information.”

- From AgPile grant Proposal

Agpile Project Team

Young Wha Lee, Gates Foundation - Sponsor
Jawoo Koo, IFPRI PI
Andre Embersics, CIMMYT - Project Management
Nsikak Akpakpan, Technical Consultant
Steffen Weber, Consultant Product Management
James Clevenger, Technical Consultant

Thank You!

Visit our website:
www.cgiar.org/initiatives/breeding-resources

LinkedIn:
CGIAR Genetic Innovation

X / Twitter:
@CGIARgenetic

CGIAR
Science for a food-secure future