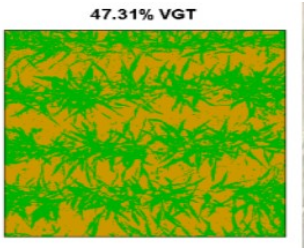
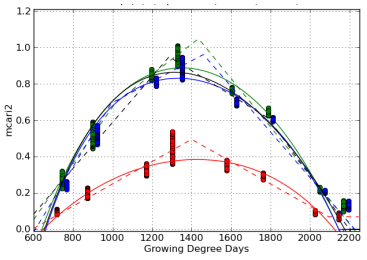
Big Data in Agriculture Challenges

Pascal Neveu INRA Montpellier

The rise of Big Data in agriculture

- **More data production from heterogeneous sources**

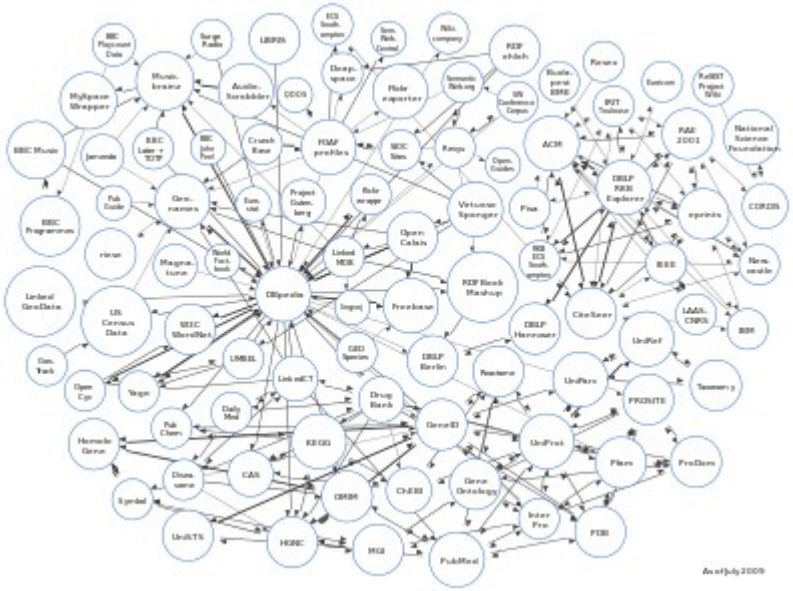




The rise of Big Data in agriculture

More and more data services and datasets on Web

- **Breeding data**
- **Crop data**
- **Weather data**
- **Soil data**
- **Environmental data**
- **Genomic data**
- **Economic, health etc.**

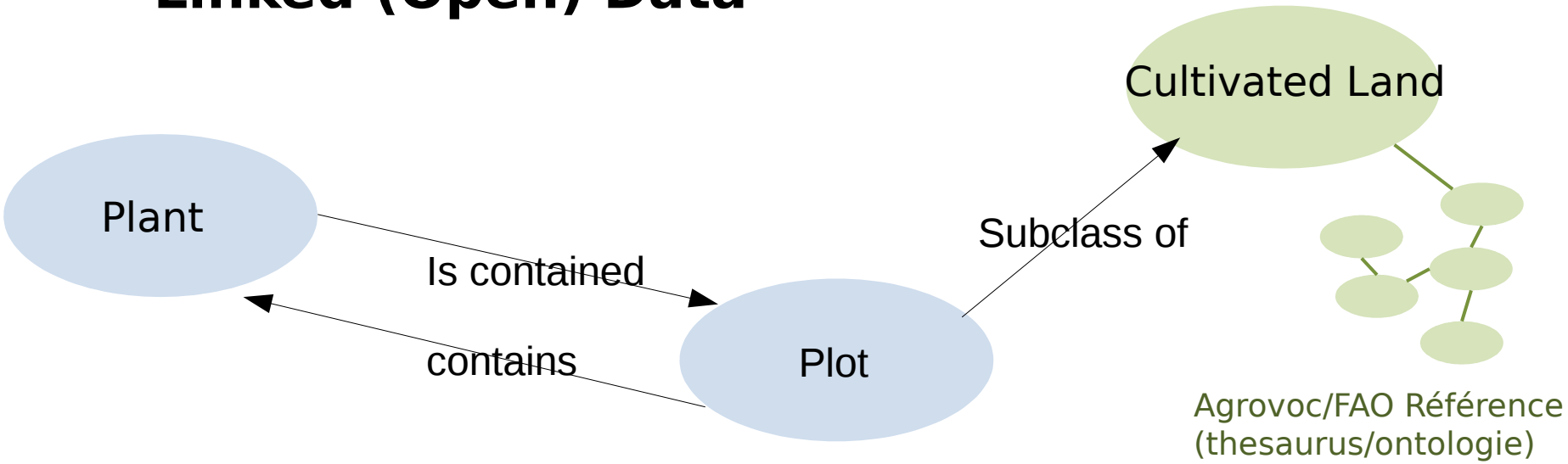


The rise of Big Data in agriculture

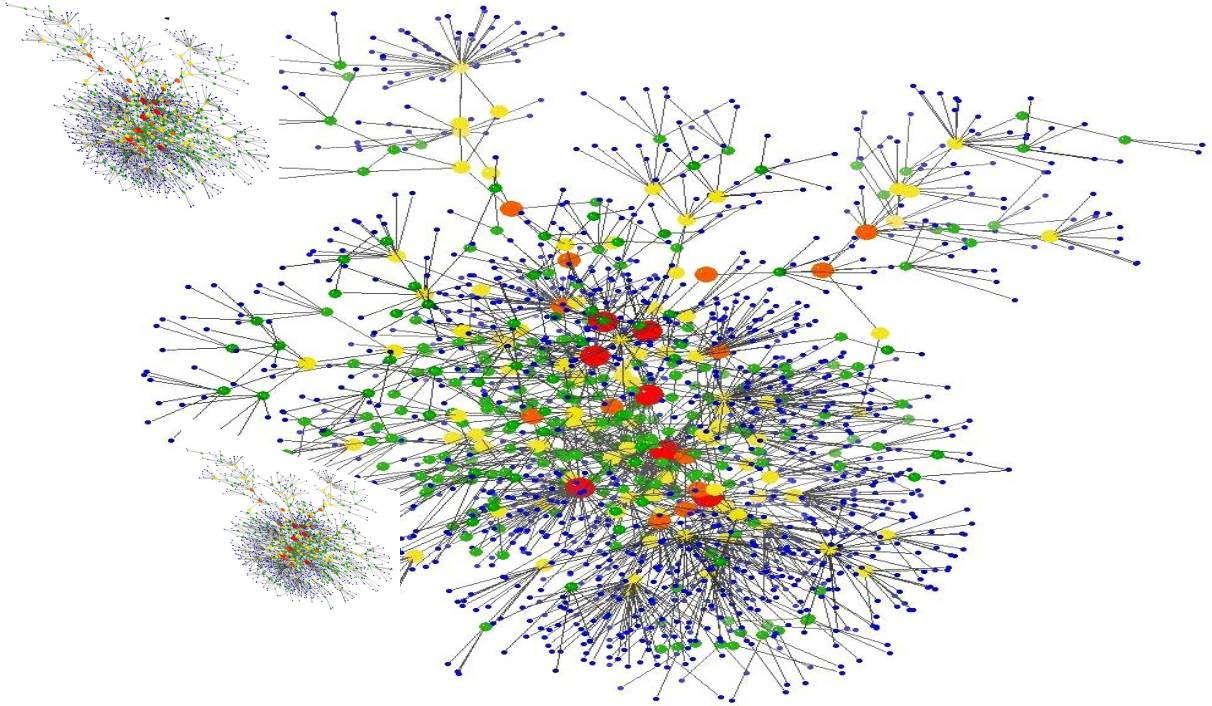
Ontology

Common conceptualisation: to link data we need to define concepts and the relations between concepts

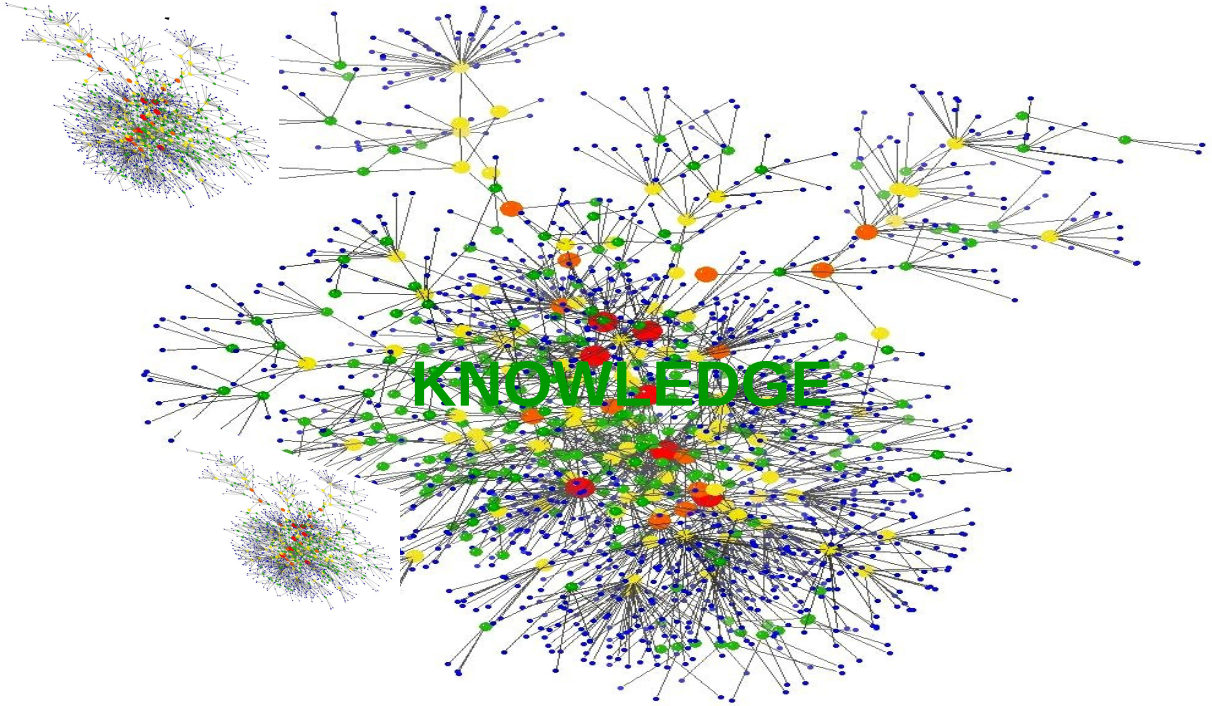
Linked (Open) Data



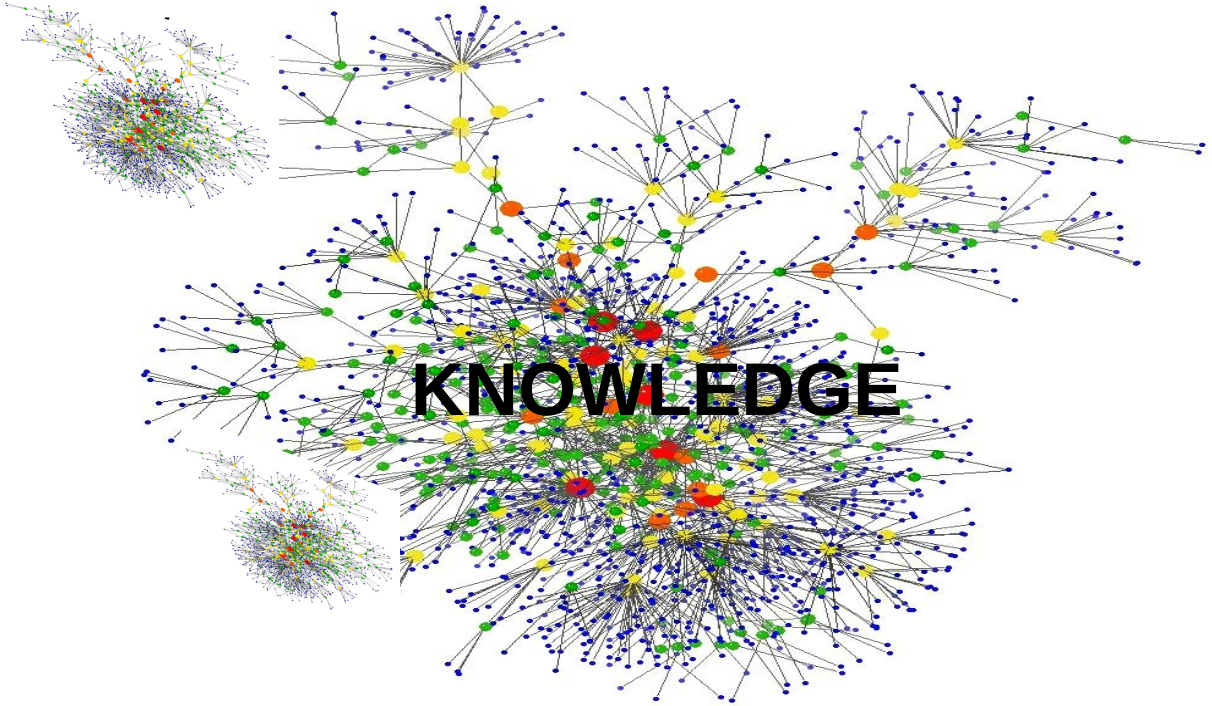
The rise of Big Data in agriculture



The rise of Big Data in agriculture



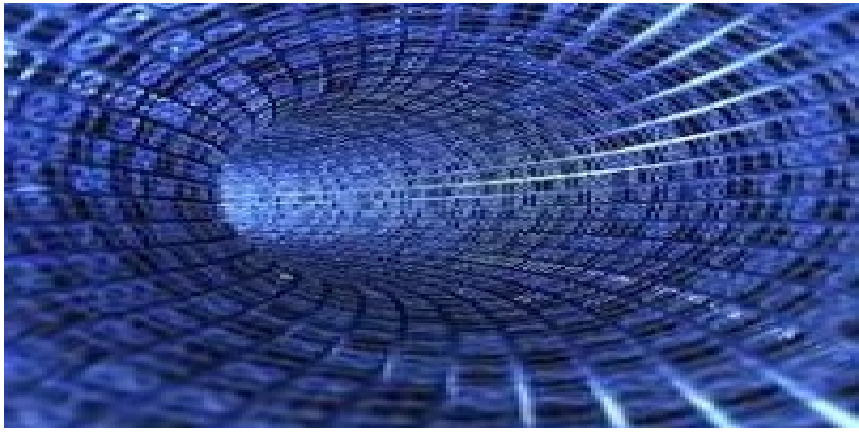
The rise of Big Data in agriculture



Big Data

A definition: *data sets grow so **large** and **complex** that it becomes **impossible** to process using traditional data processing methods (**Management** and **Analysis**)*

Make data **valuable** (information, knowledge, decision support)



Agronomic Big Data
V characteristics

- **Volume**: massive data and **growing size**
  → *hard to store, manage and analyze*

- **Variety and Complexity**: different sources, scales, disciplines different semantics, schemas and formats etc.
  → hard to understand, combine, integrate

- **Velocity**: speed of data generation
  → have to be processed on line

- **Veracity, Validity, Vocabulary**, Vulnerability, Volatility, Visibility, Visualisation, Vagueness, etc.

Why Big Data is important for Agriculture?

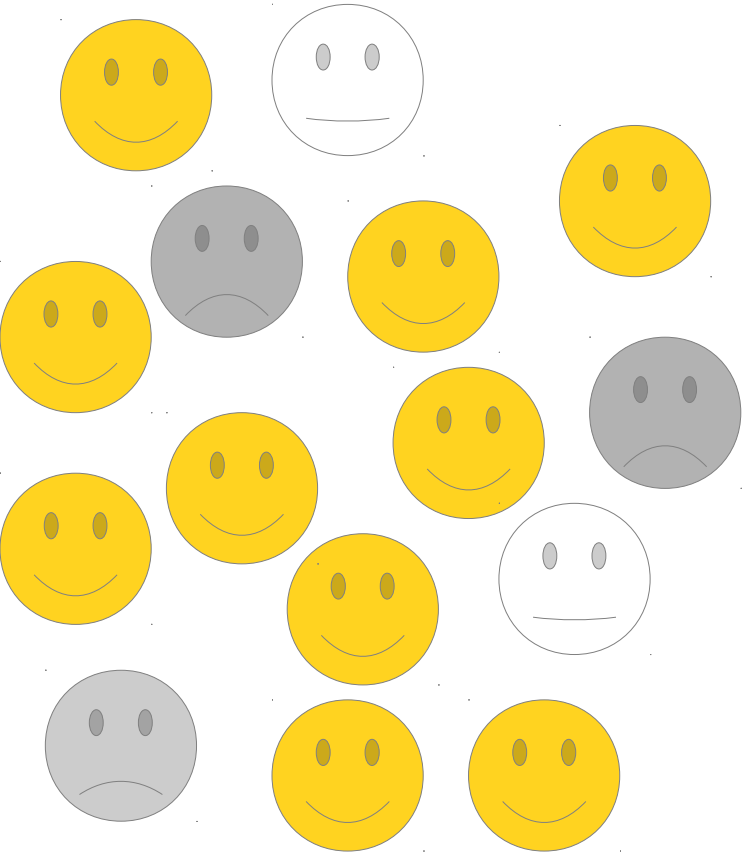
Production of a lot of heterogeneous data for understanding

- **Open new insights**
- **Allow to know:**
  - **Which theories are consistent and which ones are not!**
  - **When data did not quite match what we expect…**
- **Discover patterns, Discover frequent associations**
- **"Big Data" vs "Sample and Survey"**

→ data-driven decision support
(dynamic, integrative and predictive approaches)

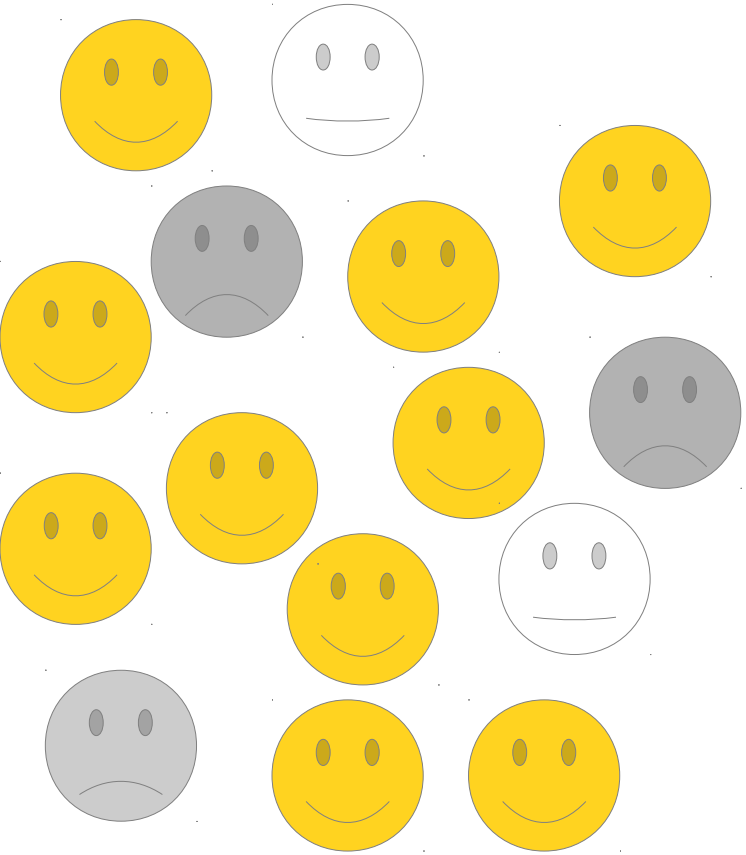
Why Big Data is important for Agriculture?

Population treatment
Some respond and some don’t

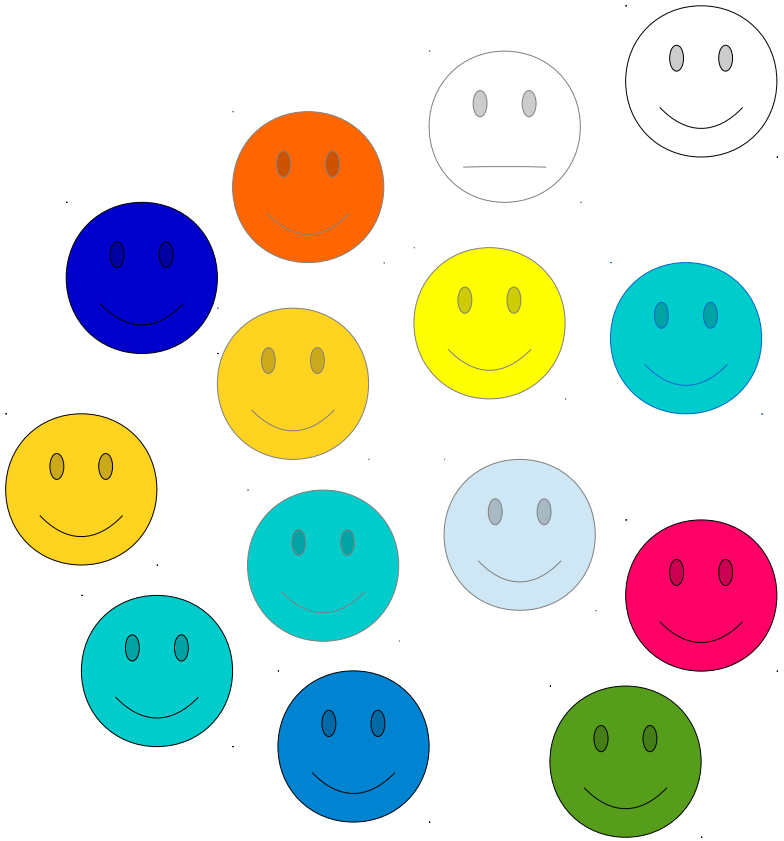


Why Big Data is important for Agriculture?

Population treatment
Some respond and some don't



Individualized treatment
All targeted population respond



Why Big Data is important for Agriculture?

Gather->Organize->Analyse->Understand->Decide

- **Adaptation to climate change**
- **Taking into account human health (farmer, consumer)**
- **More efficient use of natural resources (including water or soil) in our farming practices**
- **Sustainable management, biodiversity and Equity**
- **Food security**
  Crop performance (yields are globally decreasing)
- **...**

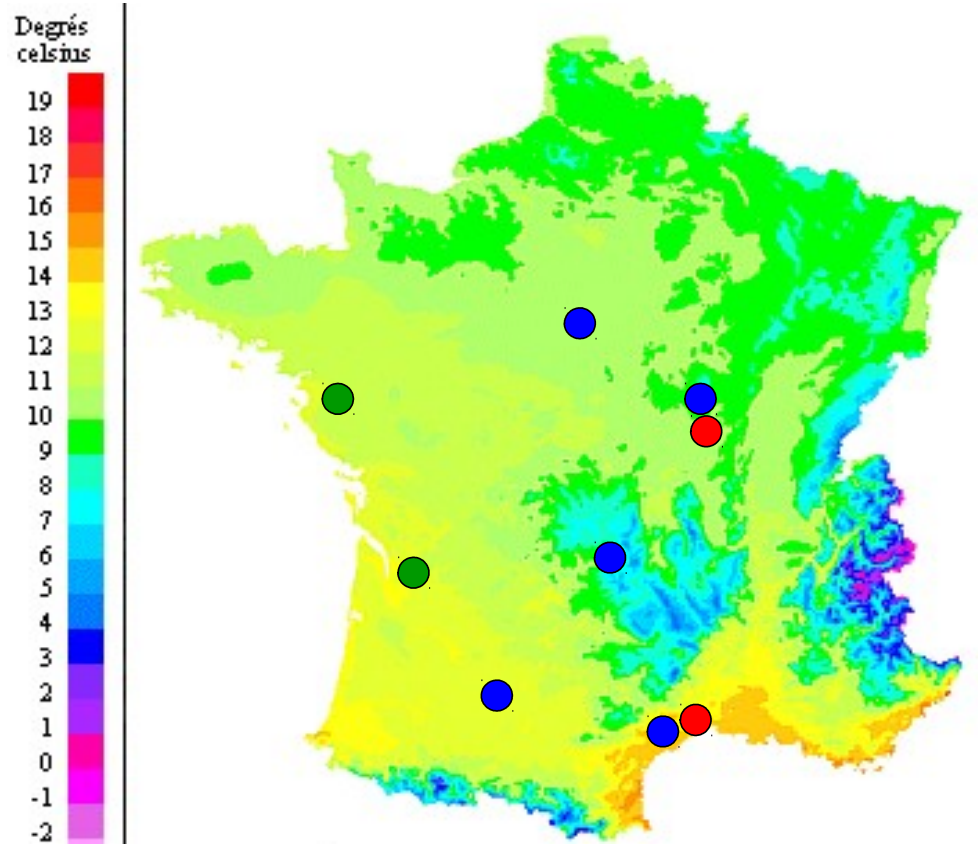
Phenome

High throughput plant phenotyping
French Infrastructure

Phenome

High throughput plant phenotyping
French Infrastructure
9 multi-species platforms

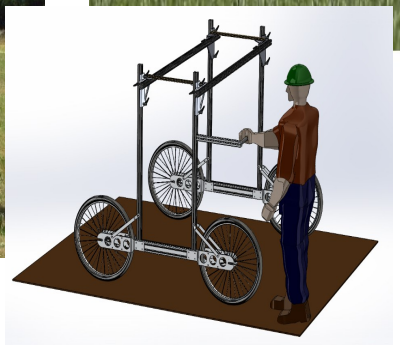
- 2 green house platforms
- 5 field platforms
- 2 omics platforms

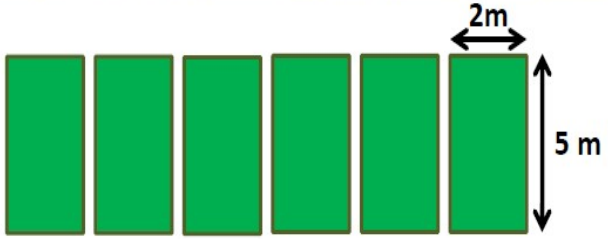


5 Field Platforms

Various scales and data types

- Cell, organ, plant, population
- Images, hyperspectral, spectral, sensors, human readings...

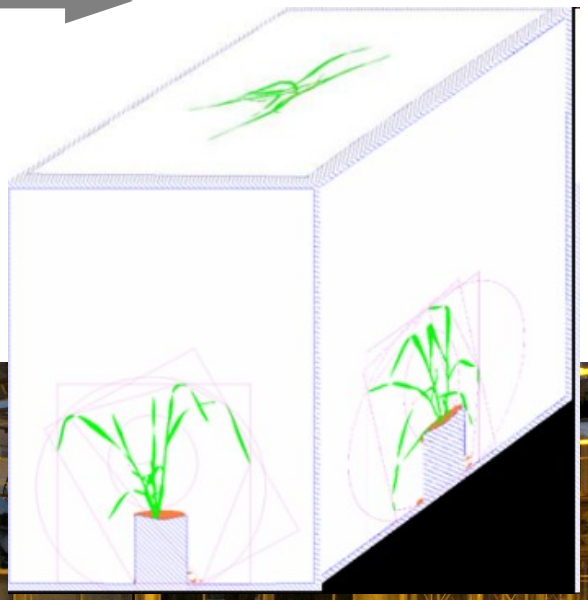


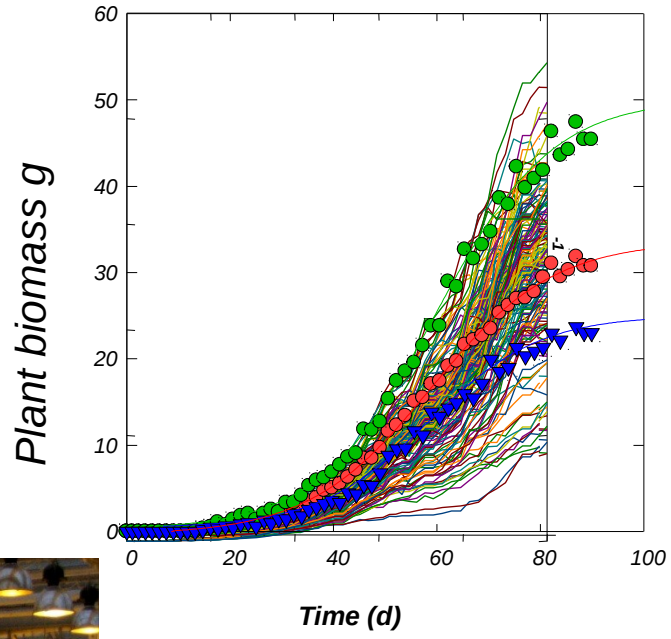


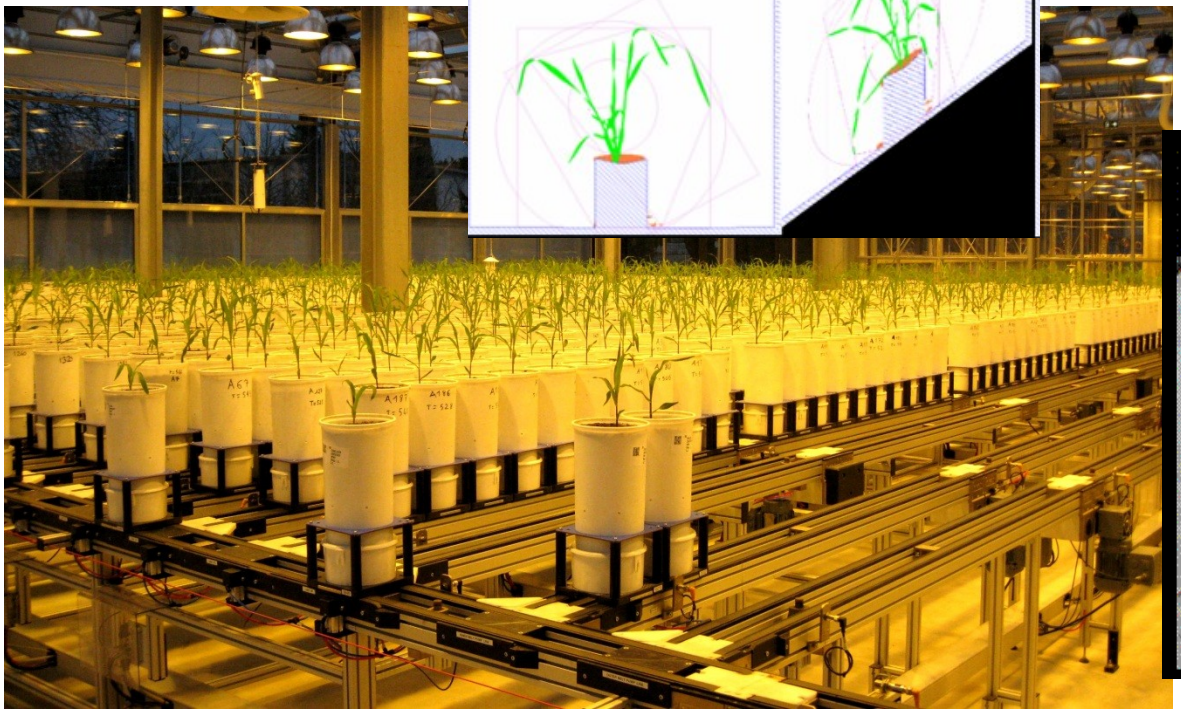
2 Green house Platforms

Various scales and data types

time





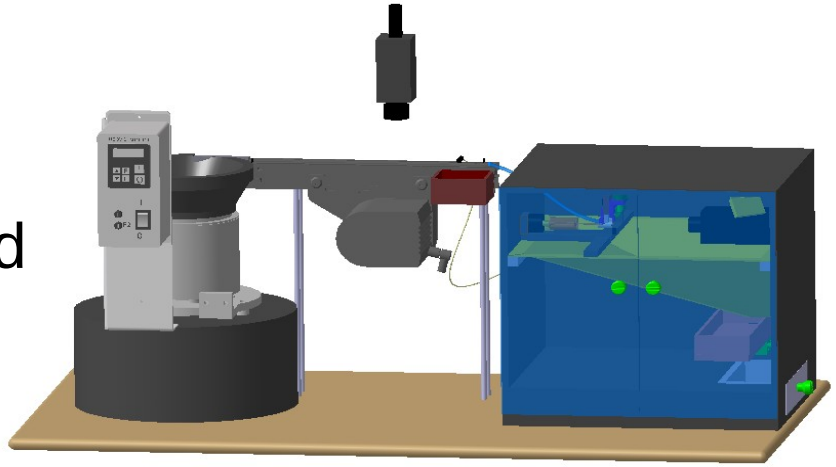


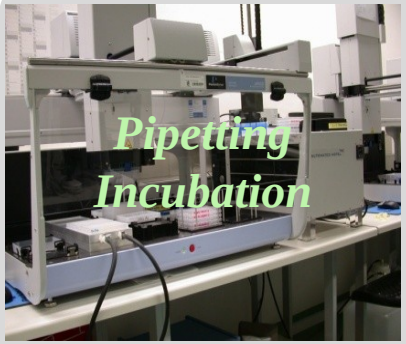
2 « Omics » platforms

Various data complex types

Composition and structure

Quantification of metabolites and enzyme activities





Data management challenges in Phenome: **Volume growth**

40 Tbytes in 2013, 100 Tbytes in 2014, …

- **Volume is a relative concept**
  - **Exponential growth makes hard**
    - **Storage**
    - **Management**
    - **Analysis**

Phenome HPC and Storage → Cloud and Grid computing (FranceGrille, EGI)

- **On-demand infrastructure and Elasticity**
- **Virtualization technologies**
- **Data-Based parallelism (same operation on different data)**

Data management challenges in Phenome: **Variety**

- **Produced by different communities (geneticians, ecophysiologists, farmers, breeders, etc)**
- **Data integration needs extensive connections and associations to other types of data**
- **(environments, individuals, populations, etc.)**
- **Different semantics, data schemas, …**

Extremely diverse data

Approaches:

→ **Web API, Ontology sets, NoSQL and Semantic Web methods**

Data management in Phenome: **Velocity**

- **Green House platforms produce tens of thousands images/day (200 days/year)**
- **Field platforms produce tens of thousands images/day (100 days/year)**
- **Omic platforms produce tens of Gbytes/day (300 days/year)**

Approaches:

- **Scientific Workflow**
  OpenAlea /provenance module (Virtual Plant INRIA team)
  Scifloware (Zenith INRIA team)

Data management challenges in Phenome: **Validity**

Data cleaning

- **Automatically diagnose and manage:**
  - **Consistency? Duplicates? Wrong?**
  - **Annotation consistency?**
  - **Outliers?**
  - **Disguised missing data?**
  - **...**

Approaches (dynamic):

  - **Unsupervised Curve clustering (Zenith INRIA team)**
  - **Curve fitting over dynamic constraints**
  - **Image Clustering**

Conclusion

Make agricultural data:

- Findable
- Accessible
- **Interoperable**
- Reusable (**over discipline**)

Discover knowledge on the world described by data (data mining, integrative methods, etc.)

E-Infrastructure

Big data is Cultural and technical challenges