INRA's Big Data perspectives and implementation challenges

Pascal Neveu
UMR MISTEA
INRA - Montpellier

Agronomic Sciences

Raises integrated issues and challenges:

- *How to adapt agriculture to climate change?*
- *How agriculture impacts environment?*
- *Agroecology «producing and supplying food in a different way »*
- *Global food security and needs of adaptation*
- *Plant treatment and food safety*
- *…*

Data challenges in Science

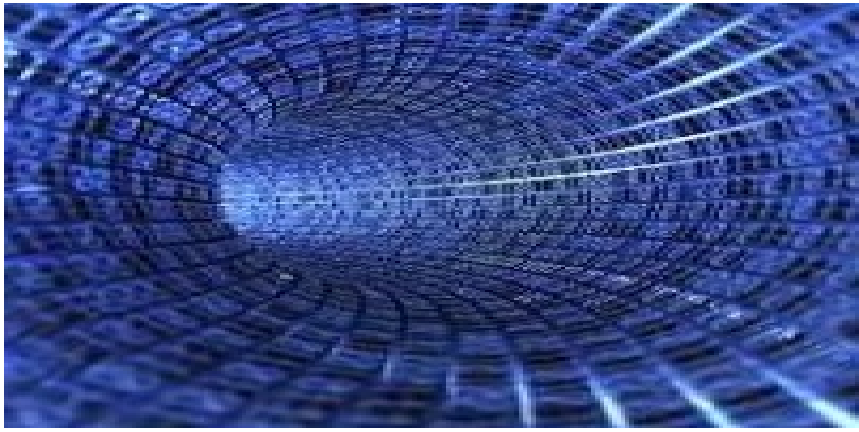
Modern science must deal:

- **More data production**
- **A lot of experimental datasets available on the Web**
- **More collaborative and integrative approaches**

→ Management, sharing and data analysis play an increasing role in research

Discover, combine and analyse these data

→ Big Data Challenges



Big data?

A buzz word

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

Big data

Why data is big?

- Devices, sensors, simulations, etc.
- Collaborative and participative
- Storage capacity, Internet access, etc.

Make data **valuable** (information and knowledge)

But

- Less than 1 % of big data is analyzed
- Less than 20 % of big data is protected
  (New Digital Universe Study)

Big Data vs Survey Sample theory

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 process on line

- **Veracity**

- **Validity,** , Vulnerability, Volatility, Visibility, Visualisation, etc.

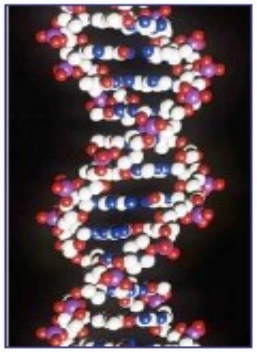
Why Big Data is important in Agronomic Sciences?

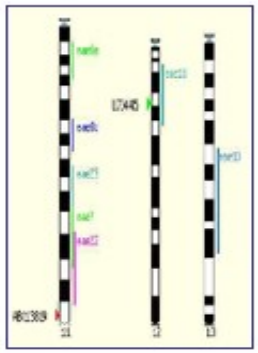
Production of a lot of heterogeneous data for understanding

- **Open new insights**
- **Allow to know:**
  - **Which theories are consistent and which ones are not!**
  - **When data did not quite match what we expect…**
- **Decision support: Combine, transform, analyse, design models, predictive approache needs**

Illustration: High throughpout phenotyping

High throughput?





Many Plant Genotypes

Interactions

Various Environments

High frequency and many trait observations of Phenotypes

Why high throughput phenotyping is important for agriculture?

- **Adaptation to climate change**
- **More efficient use of natural resources (including water and soil) in our farming practices**
- **Sustainable management and equity**
- **Food security**
  Crop performance (yields are globally decreasing)
- **…**

Genotyping and Phenotyping

Plant phenotyping has become a bottleneck for progress in plant science and plant breeding

What to measure?

- **Climate**
- **Pathogen Pressure**
- **Soil**
  - Moisture/ Tension
  - Root Biomass and distribution

Environnement

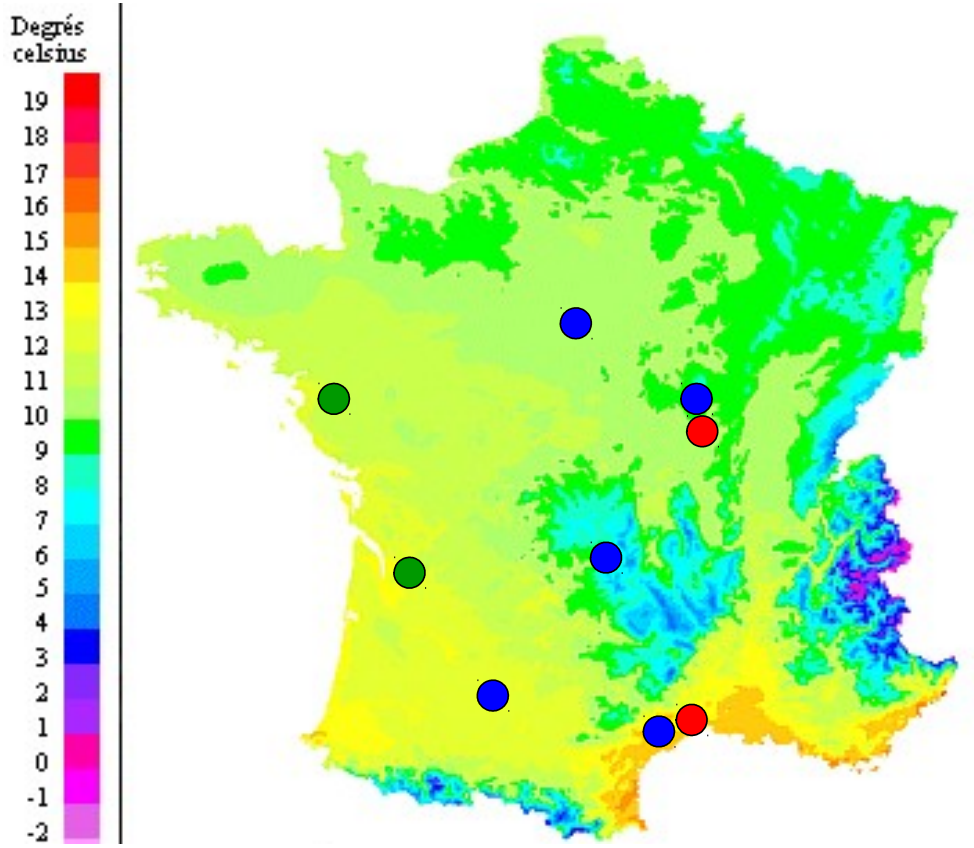
- **Structure**
  - Leaf area (GAI per layer)
  - Clumping
  - Inclination/orientation of organs
  - Density of plants/stems/ears
  - Height (per layers)
  - FIPAR
  - Leaf rolling ….
- **Biochemical content**
  - Chlorophyll, water, dry matter, Nitrogen…
- **State**
  - Fluorescence
  - Skin temperature

µ-plot

Phenome

High throughput plant phenotyping
French Infrastructure
9 multi-species plateforms

- 2 controlled platforms
- 5 field platforms
- 2 high throughout omics

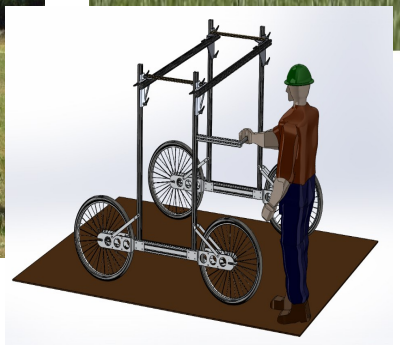


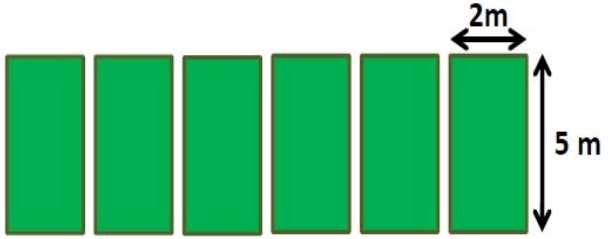
5 Field Platforms

Various scales and data types

time

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

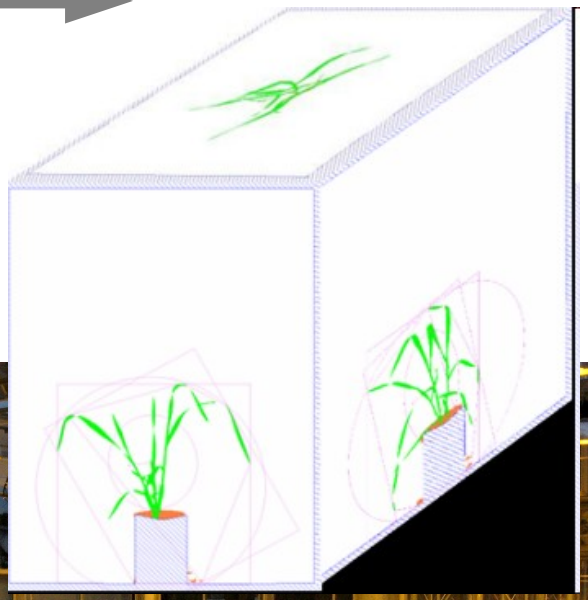


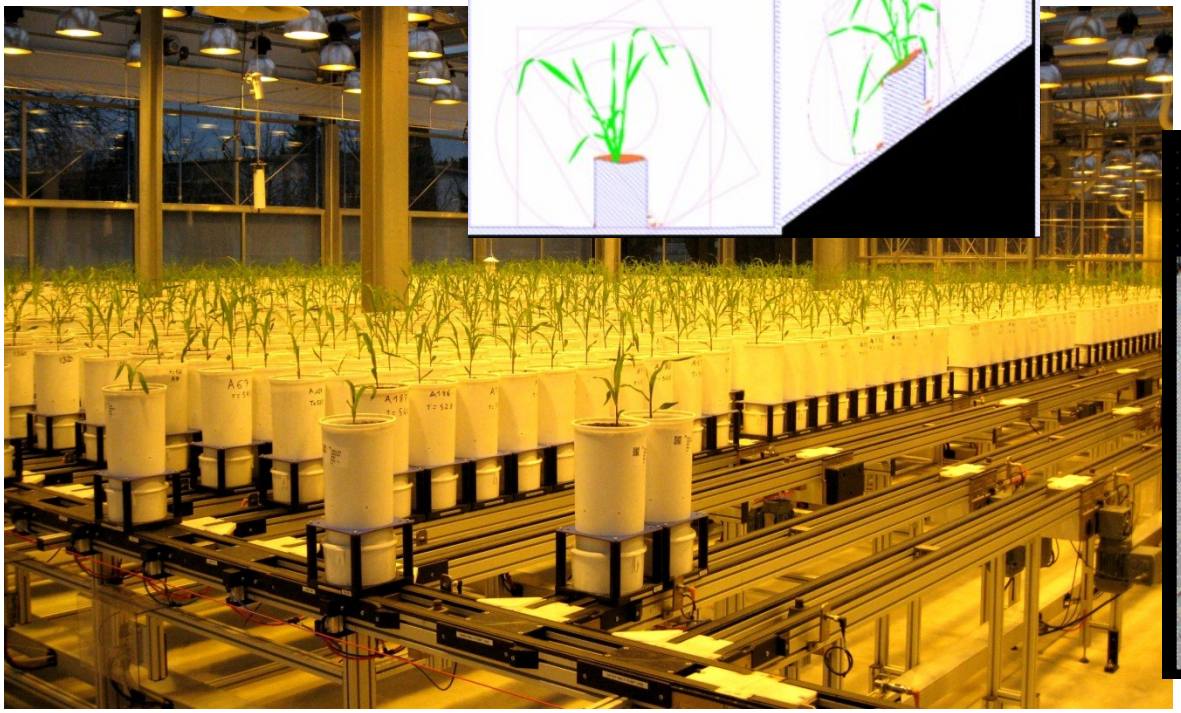


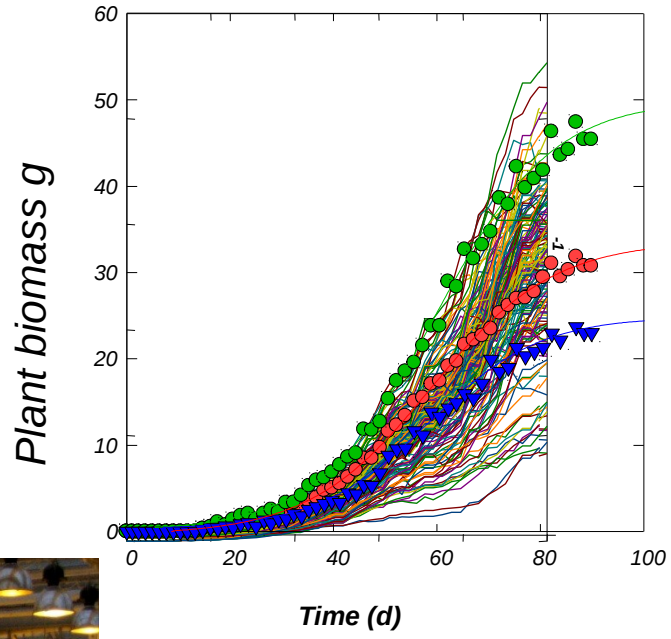
2 Controlled Platforms

Various scales and data types

time





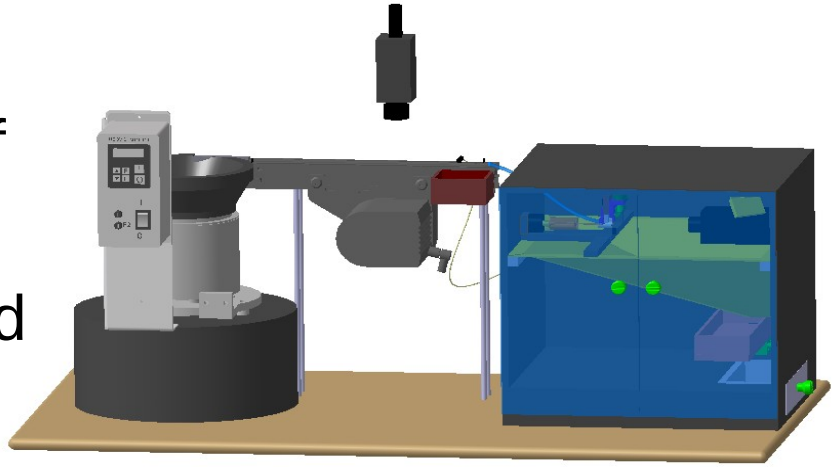


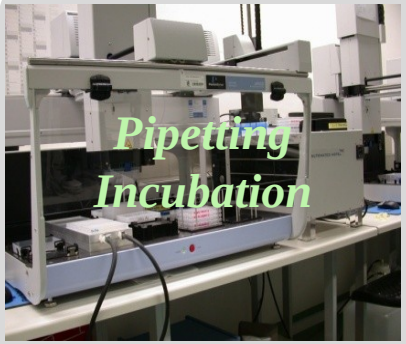
2 « Omics » platforms

Various data complex types

composition and the structure of biopolymers

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 (FranceGrille, EGI)

- **Easy to use with a sort of « unlimited scalability »**
- **On-demand infrastructure and Elasticity (season)**
- **Virtualization technologies**
- **Data-Based parallelism (same operation on different data)**

Data management challenges in Phenome: **Variety**

- **Can be produce by differents communities (geneticians, ecophysiologists, farmers, breeders, etc)**
- **Data integration needs extensive connections to other types of data (genotypes, environments, experimental methods, etc.)**
- **Different semantics, data schemas, …**
- **Can be associated in many ways (environments, individuals, populations, etc.)**

- **Extremely diverse data**

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

Data management in Phenome: **Velocity**

- **Controlled platforms produce tens of thousands images/day (200 days per year)**
- **Field platforms produce tens of thousands images/day (100 days per year)**
- **Omics platforms produce tens of Gbytes/day (300 days per year)**

Scientific Workflow

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

Data management challenges in Phenome: **Validity**

Data cleaning

- **Automatically diagnose and manage:**
  - **Consistency?, duplicate? Wrong?**
  - **annotation consistency?**
  - **Outliers?**
  - **Disguised missing data?**
  - **...**

Some approaches

- Unsupervised Curve clustering (Zenith INRIA team)
- Curve fitting over dynamic constrains
- Clustering of Image histograms

Conclusion

High throughput phenotyping data:

- **Hard to produce**
- **Hard to manage**
- **Also hard to analyse**

Thank you for your attention