OpenAlea

Journées SUCCES
November 23 - 24

InfraPhenoGrid: Une infrastructure orientée workflows scientifiques sur grille de calcul pour le traitement de données de phénotypage de plantes

Christophe Pradal, Sarah Cohen-Boulakia

cirad

INRA
SCIENCE & IMPACT

LABORATOIRE DE RECHERCHE EN INFORMATIQUE

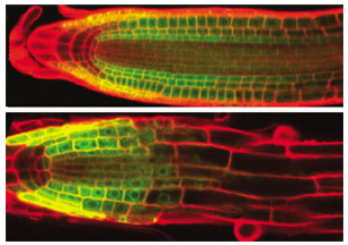
Institut de biologie computationnelle

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Context

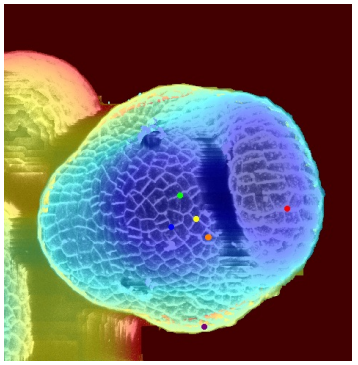
Complex Systems

- **Multiscale**
  From gene to ecosystems
- **Structural complexity**
  Sequences, Graph, Images, …
- **Complex interactions**
  - Gene <-> Shape <-> Environment



Data deluge

- Bioinformatics
- System biology
- Agronomy
- Ecology



Distributed computing

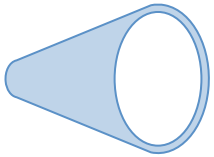
- Multi-core, cluster, grid, cloud

Analyzing and modeling in Phenomics

- The complexity of scientific simulation has increased
  - From simple models to complex simulation pipelines
- The size of data has increased
  - Phenotypic data, Next Generation Sequencing
- Scientific workflow systems provide solutions
  - Visual Programming (drag & drop tools)
  - Provenance modules to keep track of data used/produced during an execution
- How to schedule execution on Grid/Cloud?
- How to reproduce computational experiment?

Phenotyping bottleneck

Phenomics



LD mapping

Genomics

Plant Genotype: Genetic constitution of an individual organism

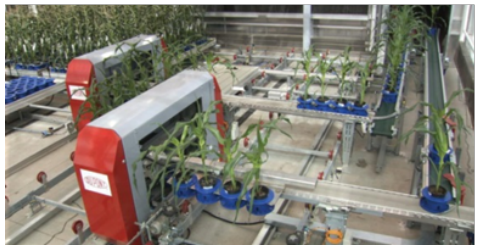
Plant Phenotype: The set of observable characteristics of an individual resulting from its interaction of its **genotype** with the **environment**

High Throughput Phenotyping

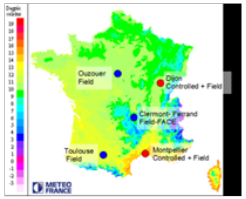
International development

The Plant Accelerator , Adelaide, AUSTRALIA

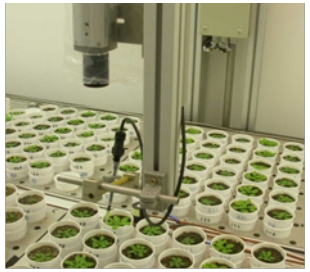
Pioneer platform, Des Moines, USA



Phenome infrastructure



Phenome infrastructure



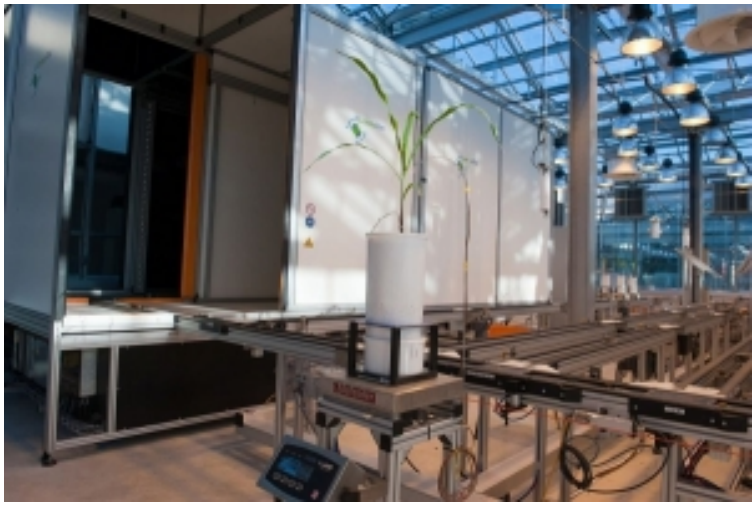
PHENOPSIS, Montpellier

Montpellier Plant Phenotyping Platforms

M3P

PHENOARCH, Montpellier, FRANCE

Overview of PhenoArch / Montpellier



Imaging unit

Conveyor belts

Watering stations

Plant Phenomics / Scientific Challenges

- How plants adapt to different stresses due to **climatic change**?
- Study the impact of different **environmental** conditions for various **genotypes**.
- Quantifying **Topology**, **Geometry** and **Development** of plants by Imaging

ARVALIS Institut du végétal

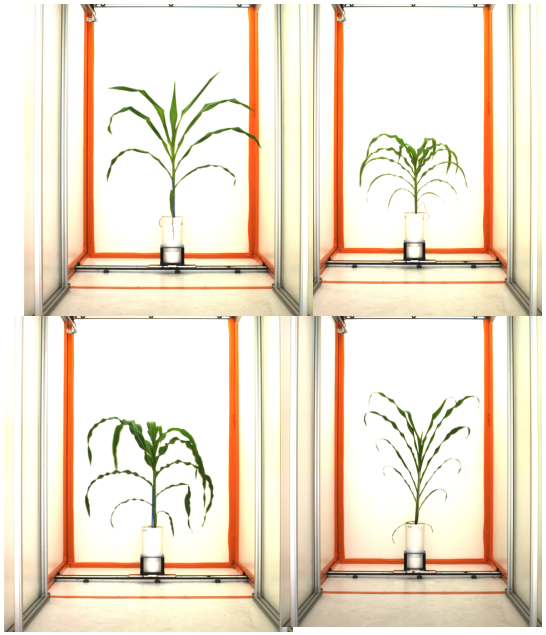
hi-phen

M3P

lepse Montpellier

Plant Phenomics / Experiments

- **Typical experiment**
  - 1680 plants
  - 40 time point per plant
  - Imaging (13 sides & top view)
    - 52 GB/day
    - 2.75 TB/essay
    - 11 TB / year
  - Watering and whole-plant transpiration
    - Temperature + weight measured every 15mn



ARVALIS Institut du végétal

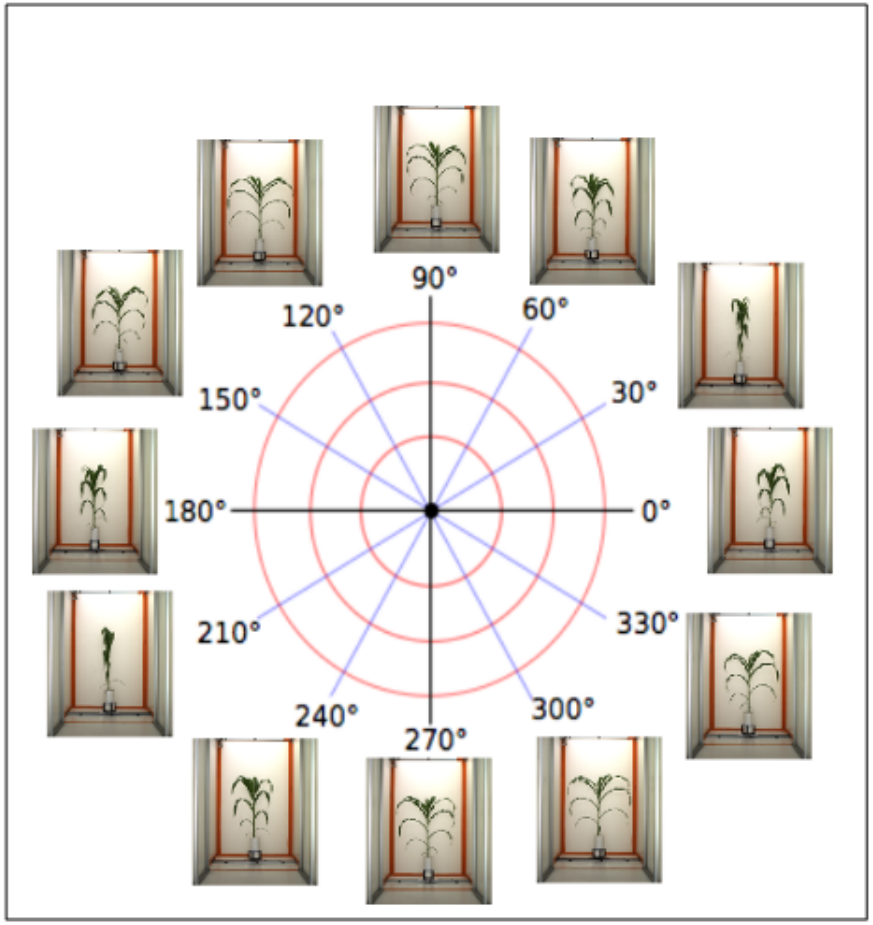
hi-phen

M3P

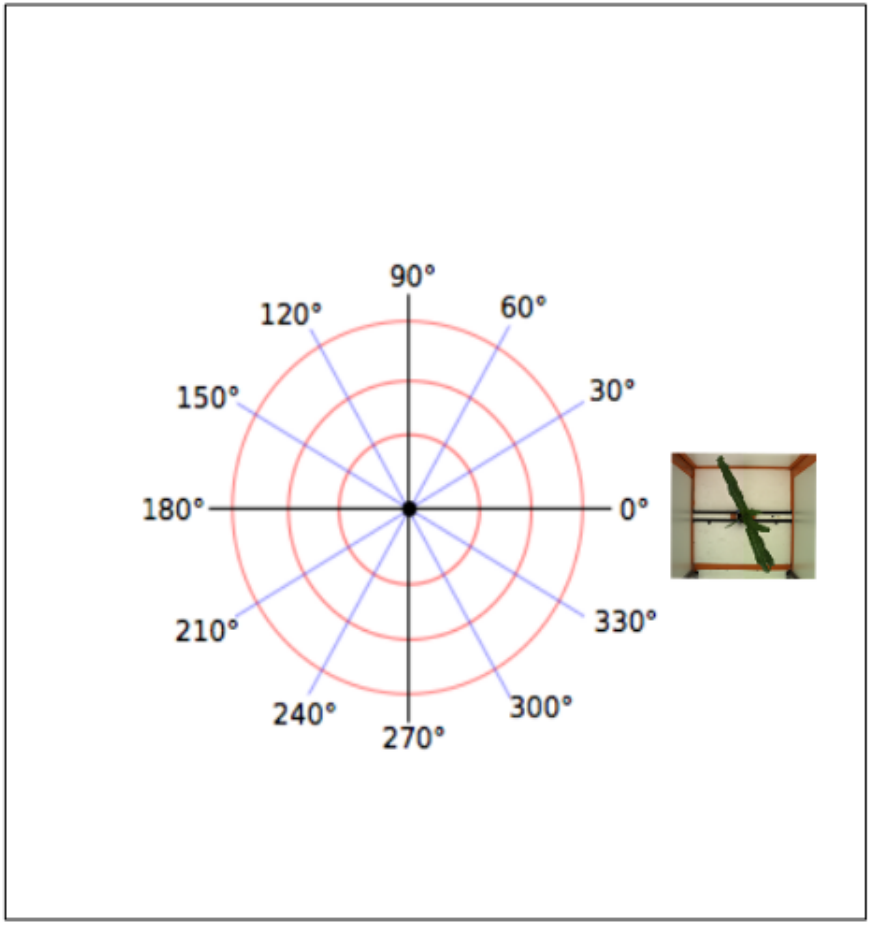
lepse Montpellier

Multiview imaging

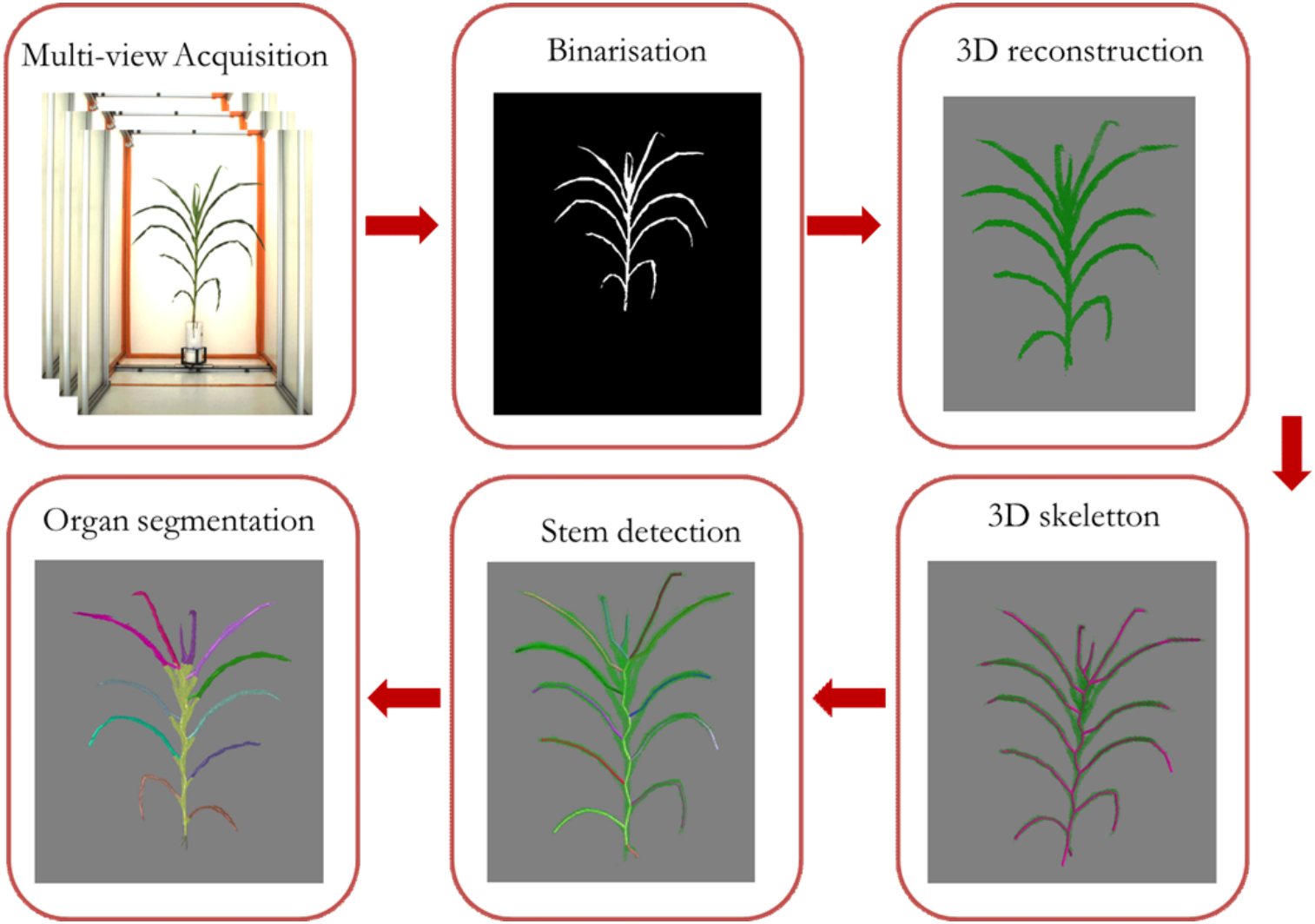
Side Camera



Top Camera



Scientific workflow for Plant Phenotyping

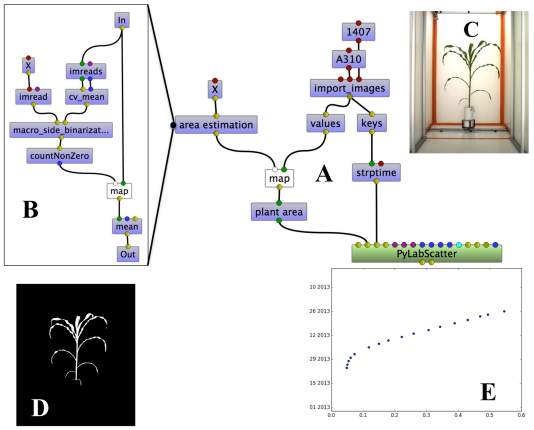


Fournier et al. 2015. IAMPS

Infrastructure for Grid Computing

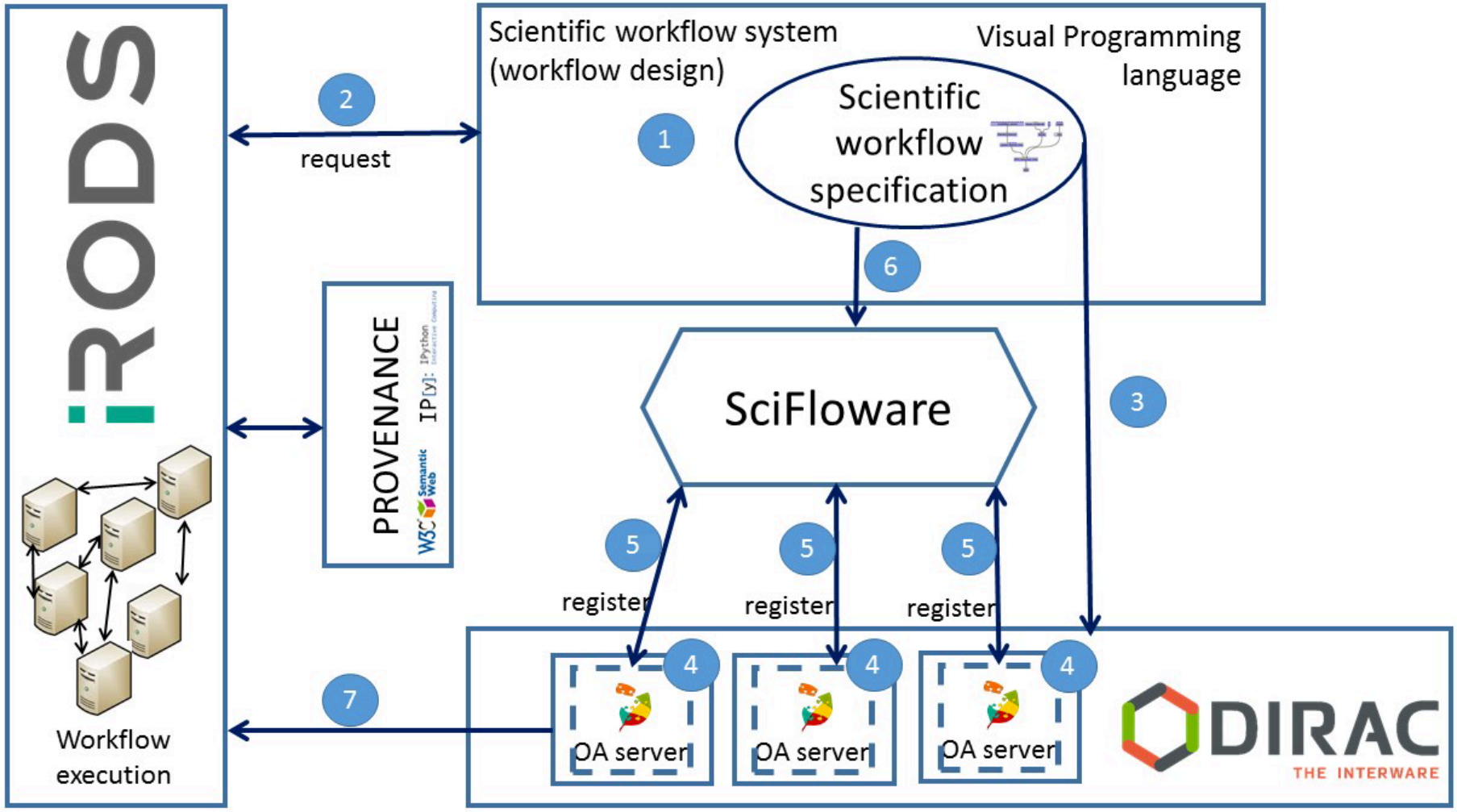
InfraPhenoGrid

- OpenAlea higher-order scientific workflow
- Grid computing using SciFloware middleware
- iRODS : data storage
- DIRAC : jobs management
- Provenance of execution stored as Jupyter Notebooks

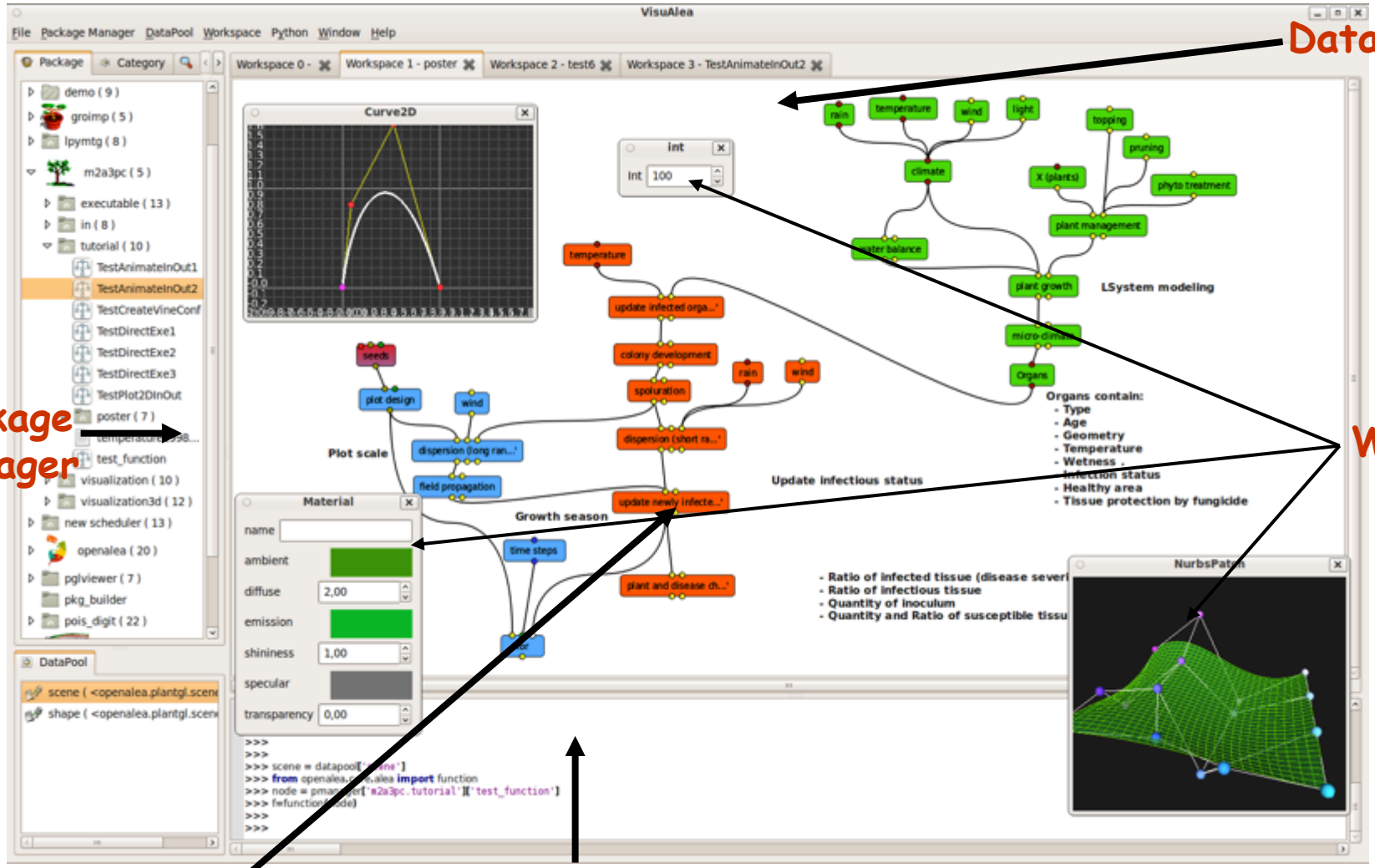


Pradal *et al*. 2016. Future Generation Computer Systems, in press

InfraPhenoGrid - Architecture

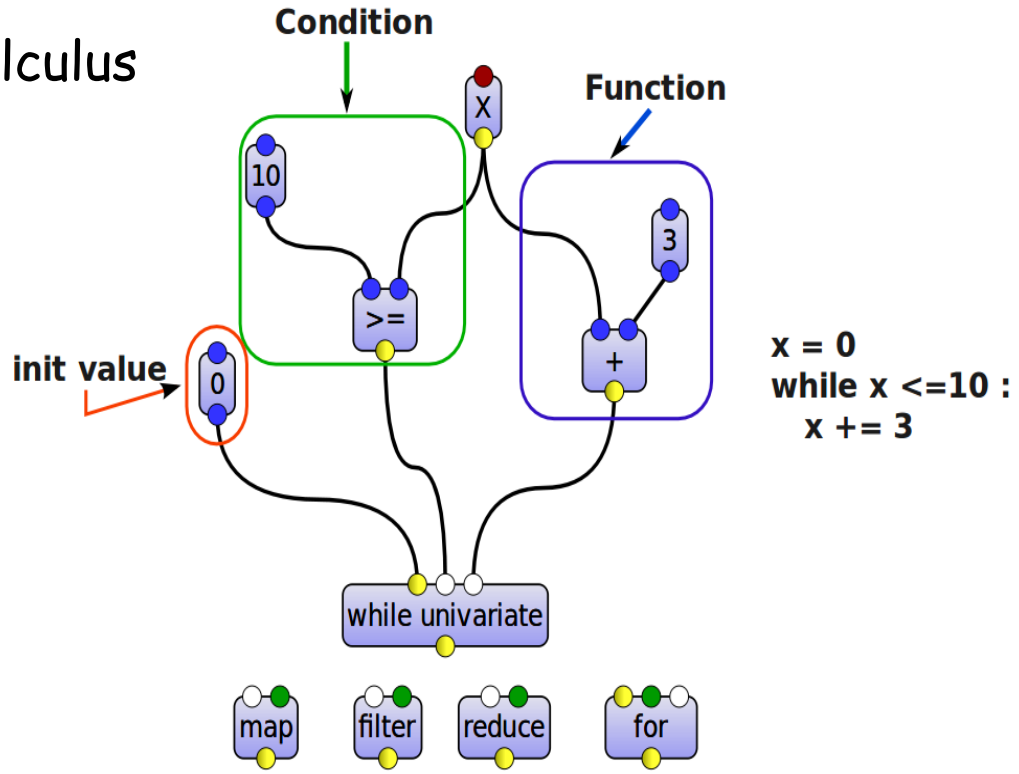


OpenAlea Scientific Workflow



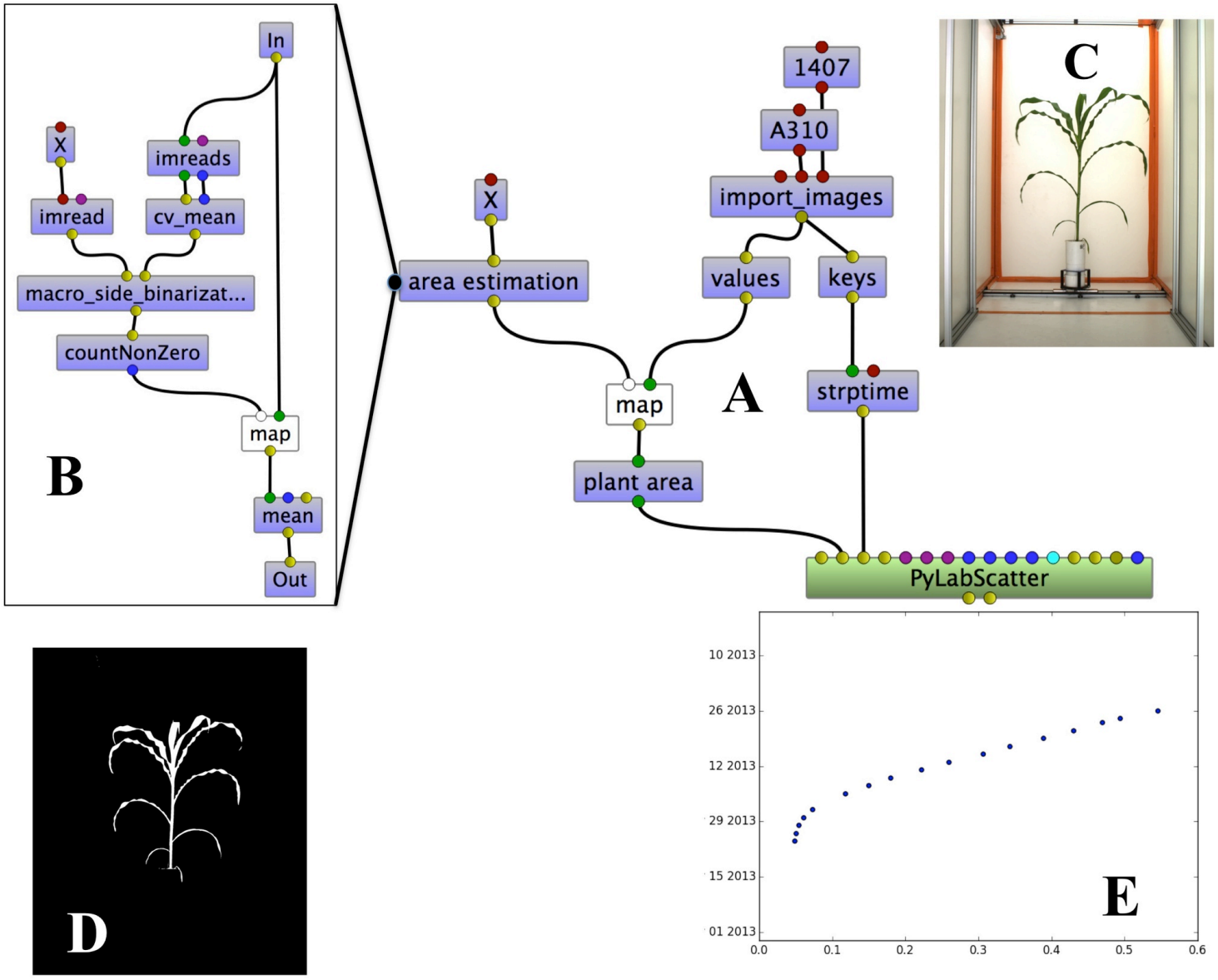
Algebraic Scientific Workflow

- Control-flow using lambda-calculus
- Dataflow Variable (X)
  - Transform a dataflow into a function
- Algebraic Operator
  map, reduce, filter…



Pradal, Fournier, Valduriez, Cohen-Boulakia. SSDBM 2015

Behind the Scene



Distributed Data-oriented Workflow

D. Parigot, P. Valduriez (Inria)

Approach

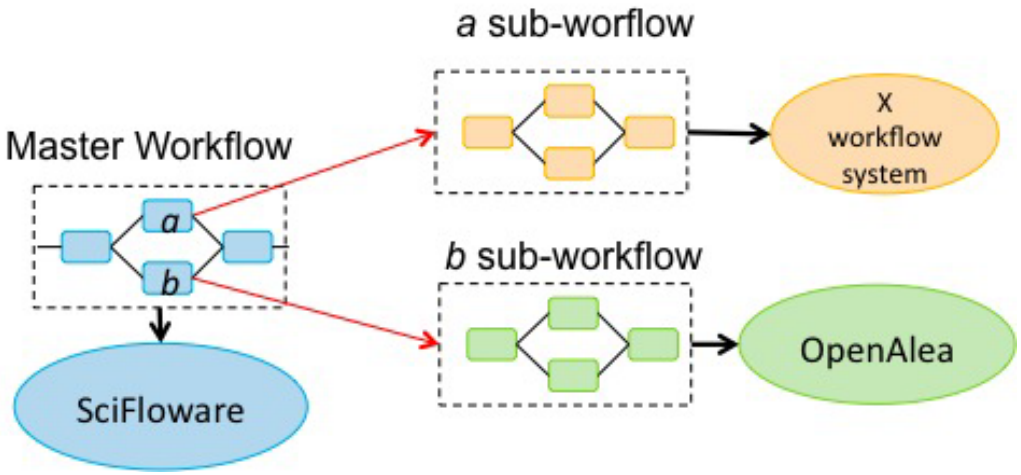
- Exploit distributed data management techniques, e.g. algebraic languages with the definition of a new **data-centric scientific workflow specification**
- Parallel workflow execution in multiple clouds to scale up to big data
- Heterogeneous workflow composition (with workflows in different languages)
- Interoperate with different data management systems (supported by different SWfMS)
  (iRODS, HDFS, Key Value stores, Hbase, etc.)

Software solution:

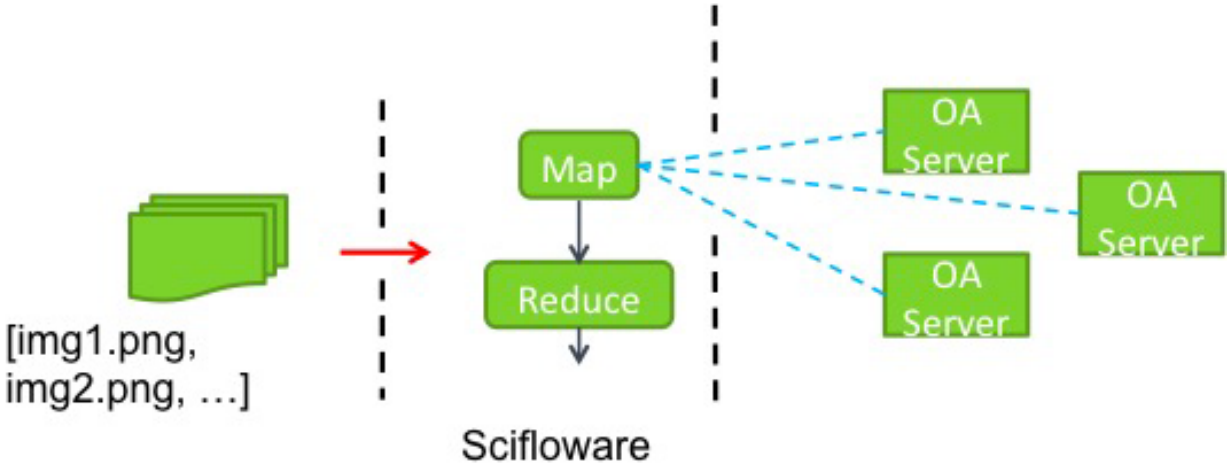
- a Scientific Workflow middleware (**SciFloware**)

SciFloware – Scheduling Workflows

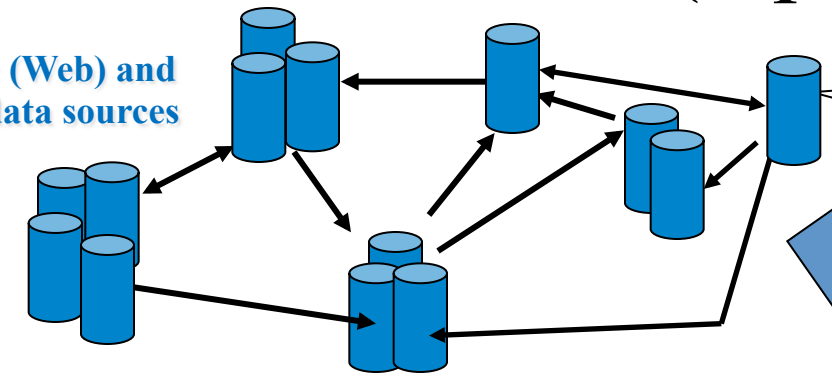
Heterogeneous workflow composition

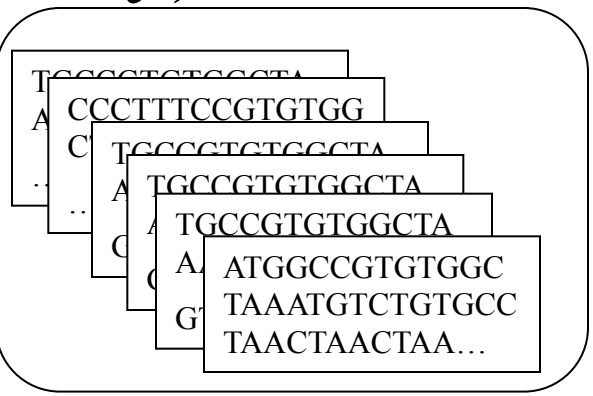


Algebraic workflow operators (Map)



Provenance (reproducibility)



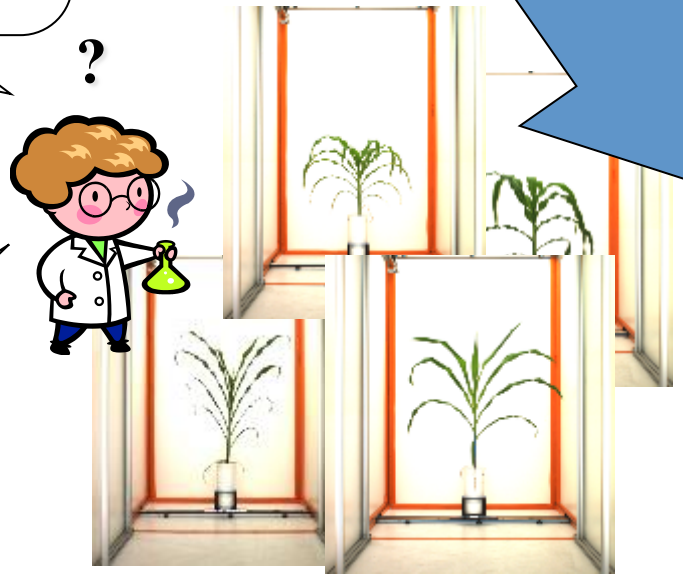


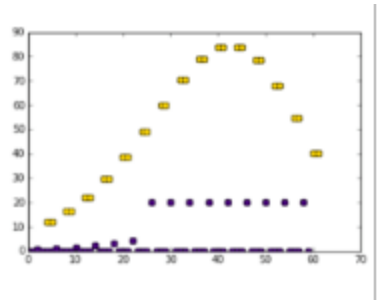
Binarization Water Use Efficiency
Segmentation RUE …

How has this plot been generated?
With which images?
With which binarization algorithms
→ **Provenance**

Which data are really important to inspect? (levels of granularity)

What is the difference between these two workflow executions?





Biologist's workspace

Provenance

Provenance

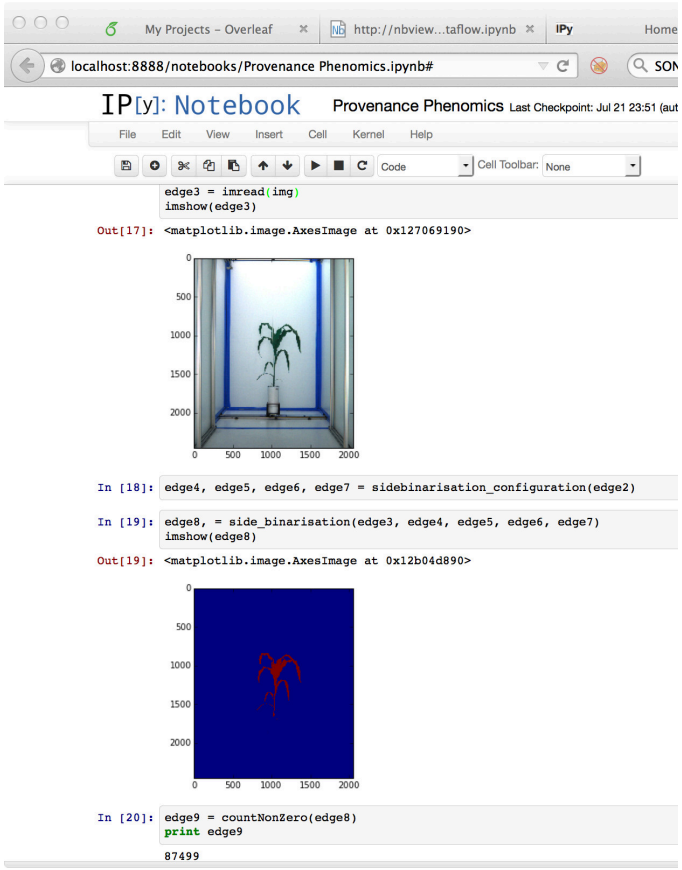
- Log execution information
- Prov-DM (w3c)

Workflow execution saved in Jupyter notebooks

- Actors in the workflow -> cells in the notebook

- Input and Output data used and produced during an execution can be visualized

- Stored in iRODS



Execution in Jupyter Notebook

IP[y]: Notebook

Scientific workflows meet modeling and simulation Last Checkpoint: Mar 18 18:39 (autosaved)

File Edit View Insert Cell Kernel Help

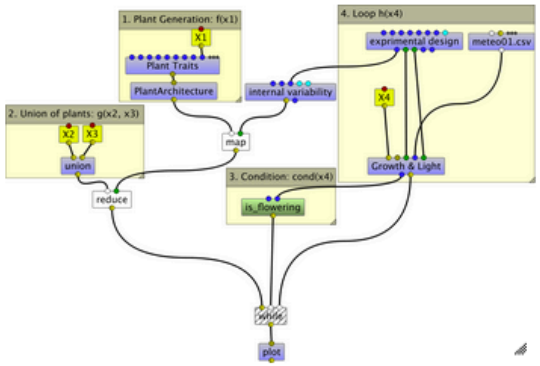
Code Cell Toolbar: None

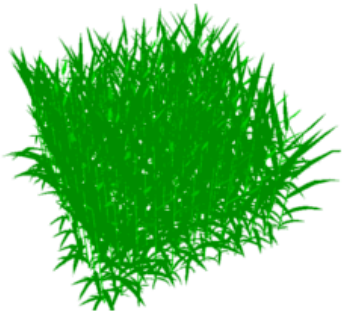
OpenAlea: Scientific workflows meet modeling and simulation

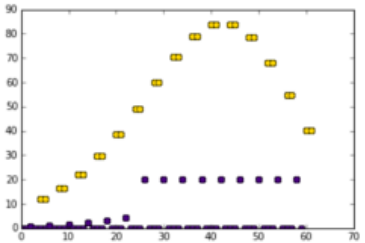
In [2]:
```
from vpltkdisplay import *
from IPython.display import display
from openalea.plantgl.all import *
from openalea.core import *
import numpy as np
```

Simulation of the growth of a crop

In [3]:
```
pkname='alinea.adel.tutorials.ssdbm'
node='5- while'
display(Dataflow(pkname, node))
```







Pradal, Fournier, Valduriez, Cohen-Boulakia. SSDBM 2015

France Grille Support

Infrastructure

- Provides a very large scalable infrastructure for storing and processing
- From Grid to Cloud

Middleware

- iRODS: de facto solution
- Dirac

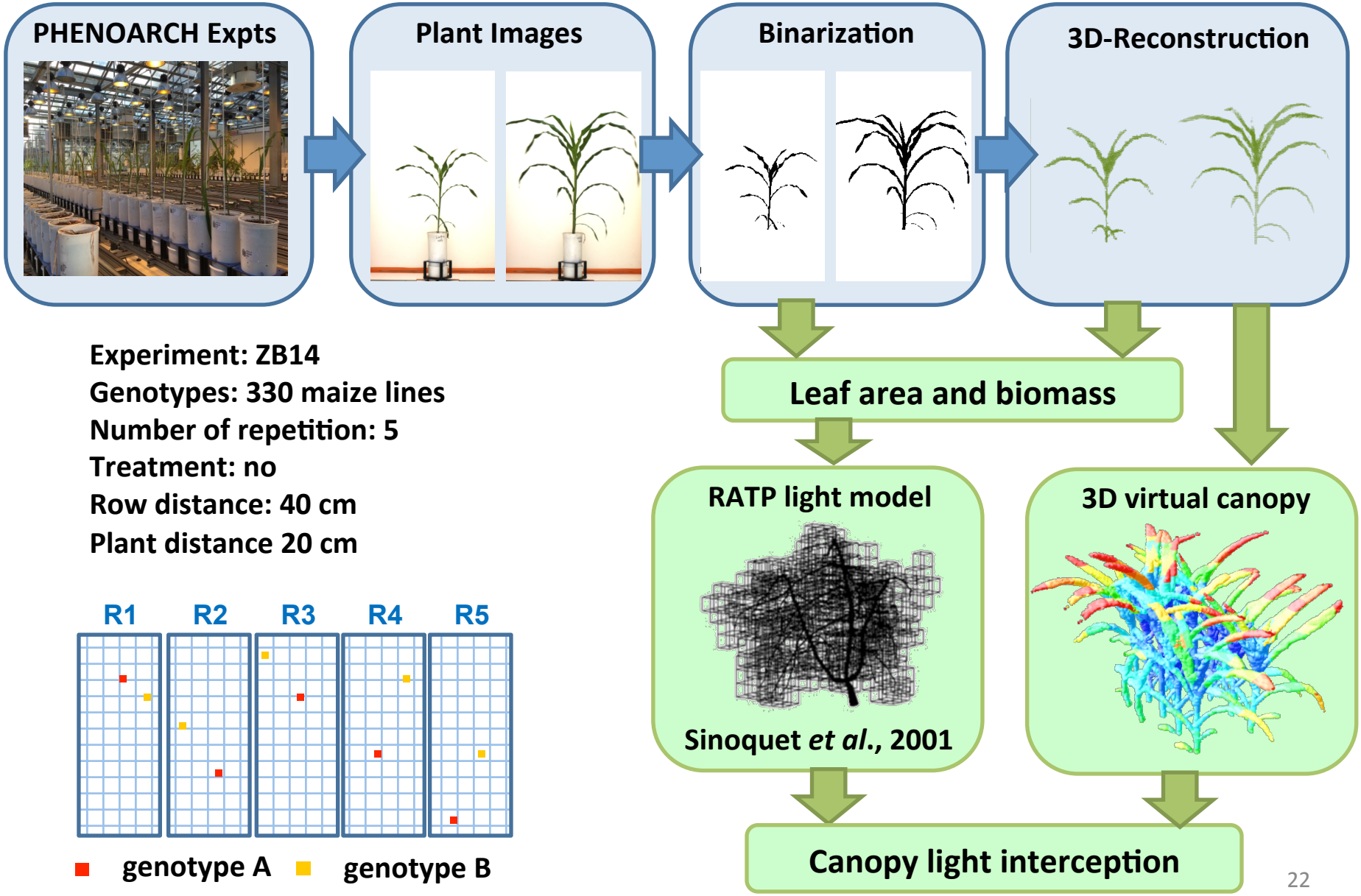
Pro

- Excellent support
- Free

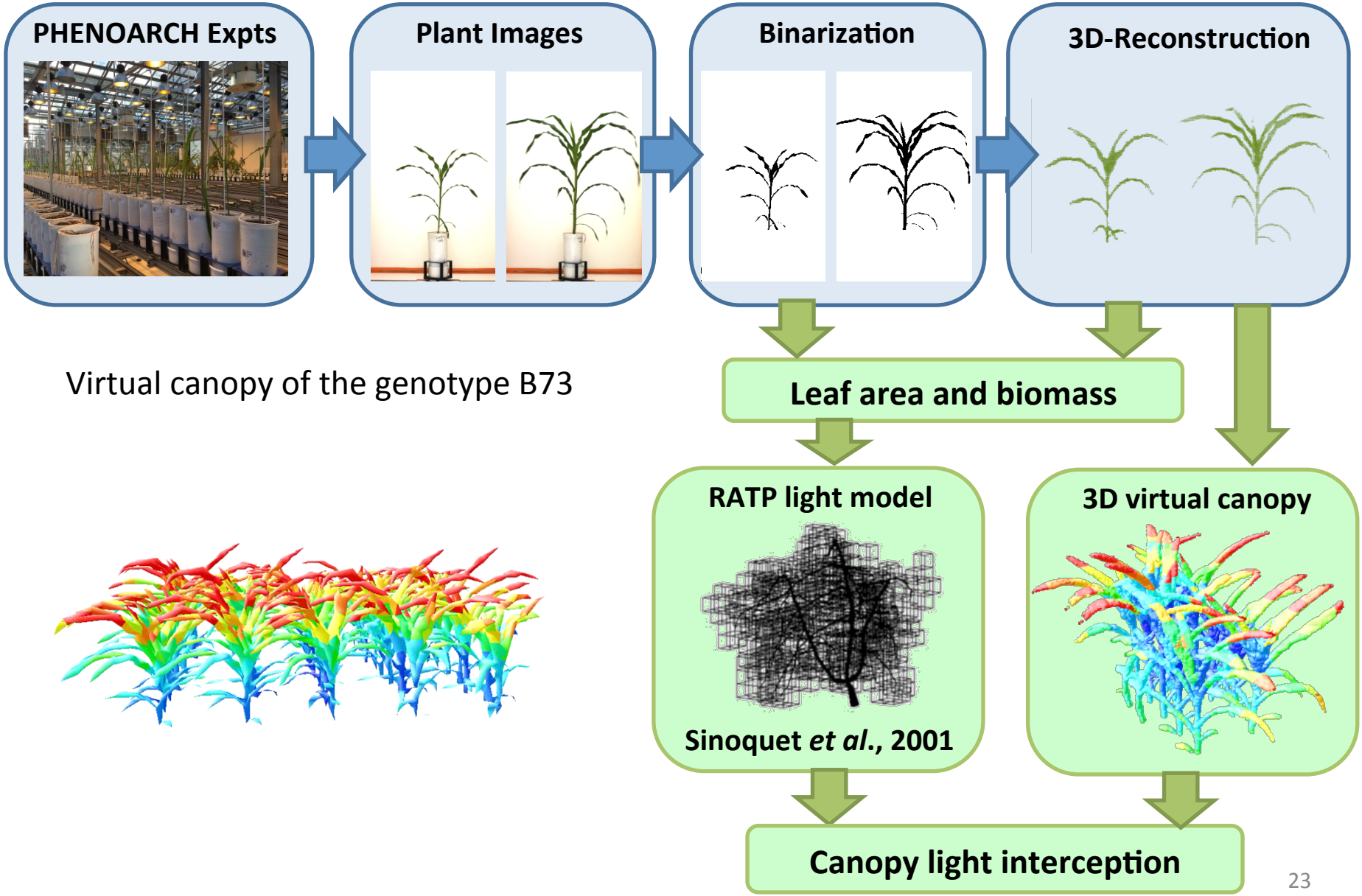
Cons

- Complexity to drive the computation outside the Grid (we implement our communication protocol)

Simulation of the greenhouse and its ligth environment

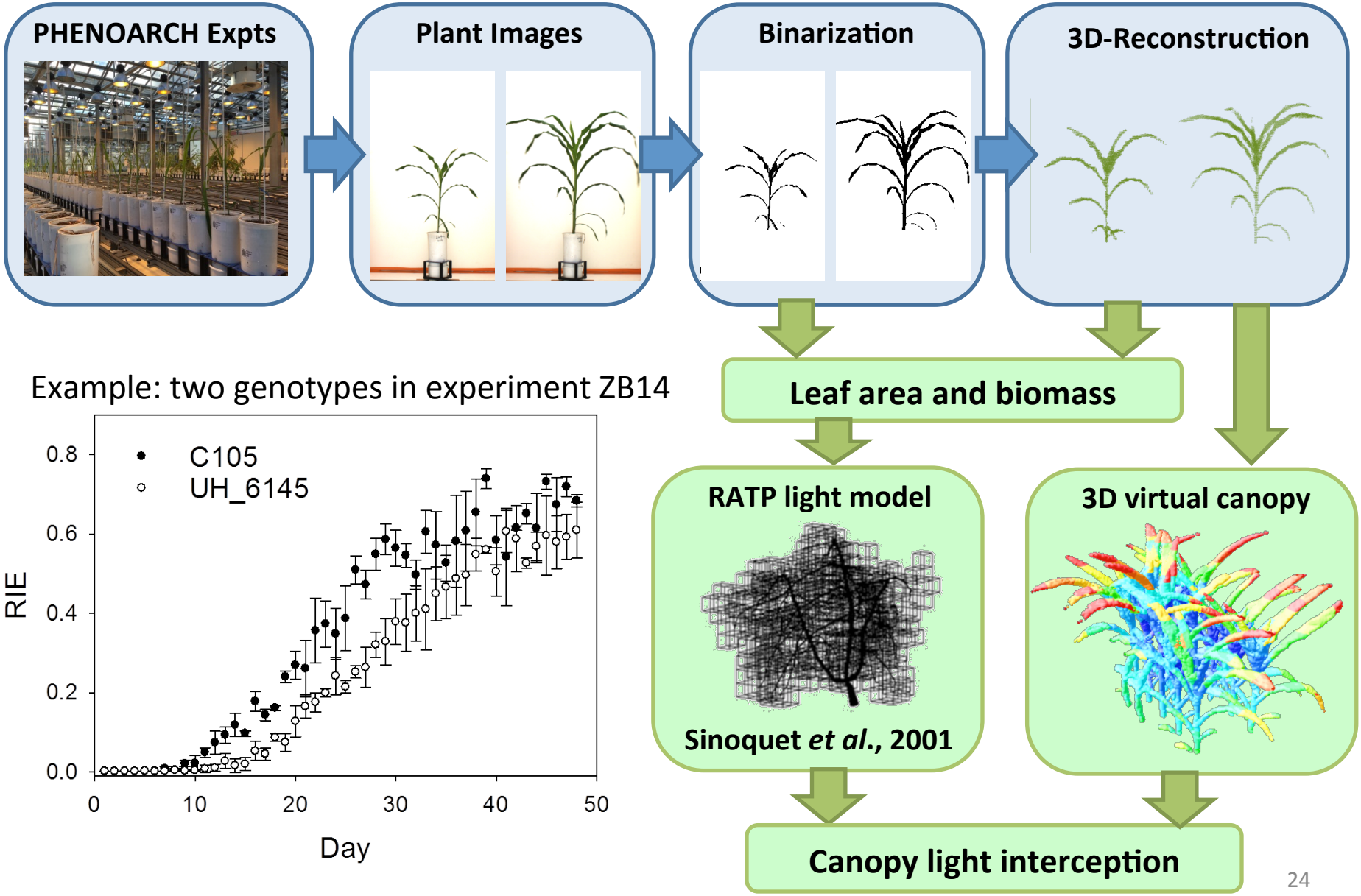


Retrieving light interception for individuals



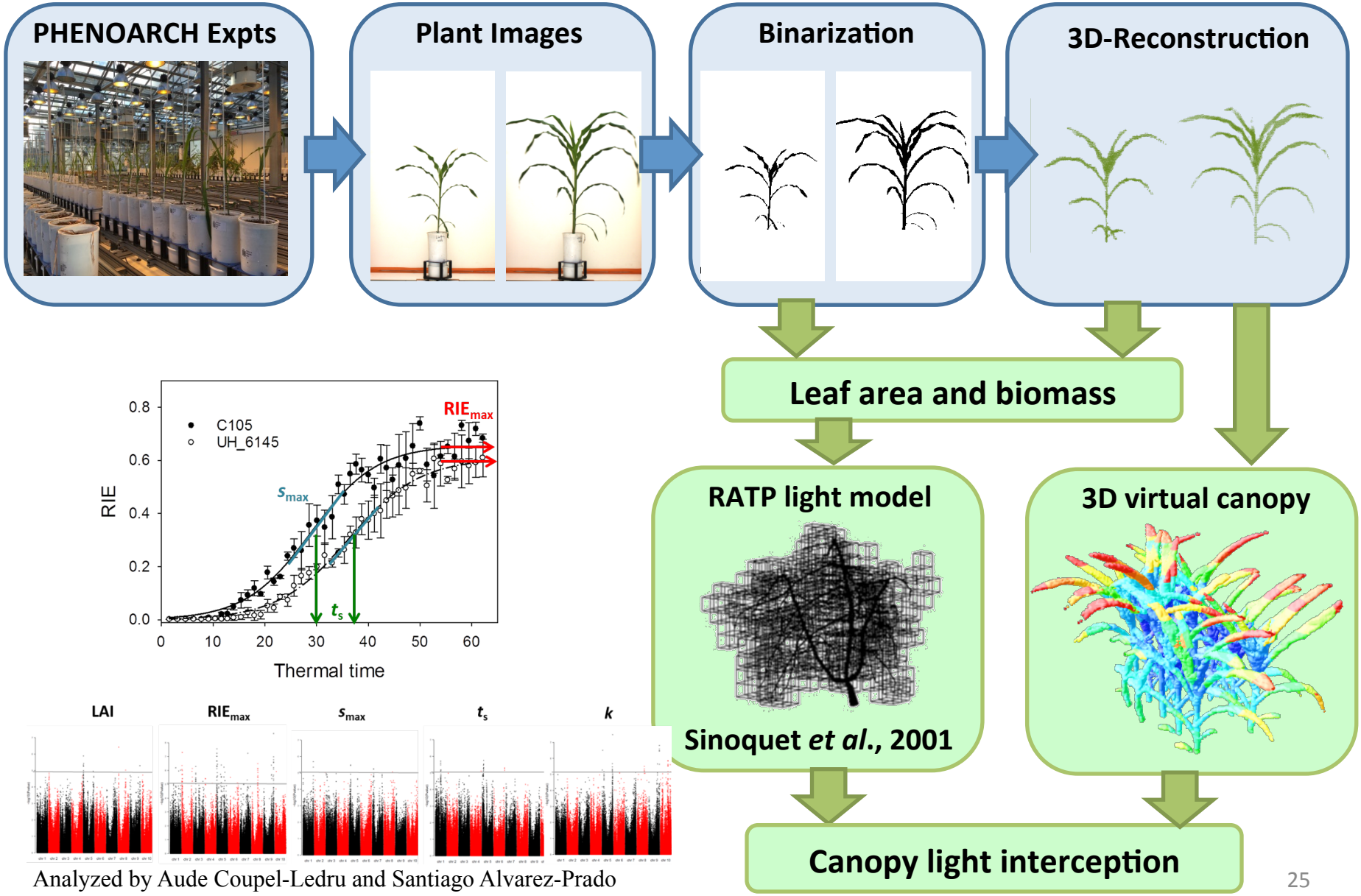
Virtual canopy of the genotype B73

Radiation interception efficiency



Example: two genotypes in experiment ZB14

Traits & QTL detection



Analyzed by Aude Coupel-Ledru and Santiago Alvarez-Prado

Chen et al. 2016 IEEE FSPMA

Radiation use efficiency: biomass + light

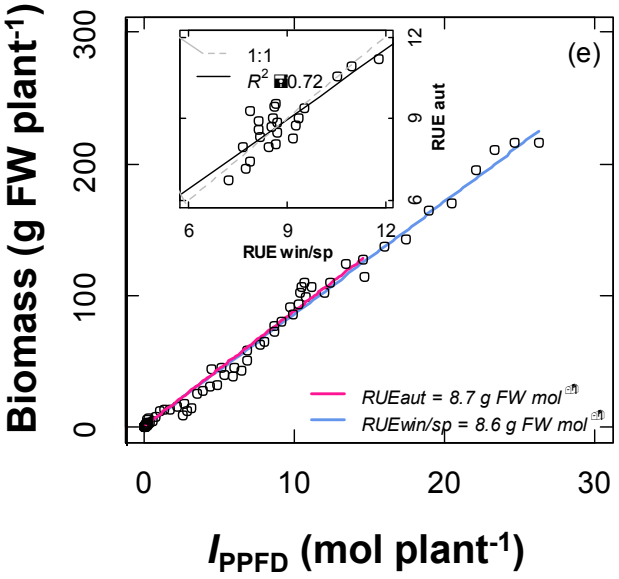
PHENOARCH Expts

Plant Images

Binarization

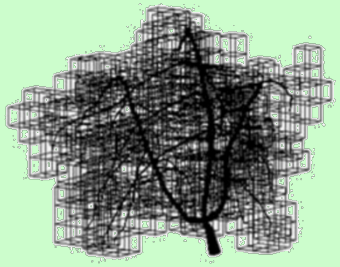
3D-Reconstruction

Leaf area and biomass



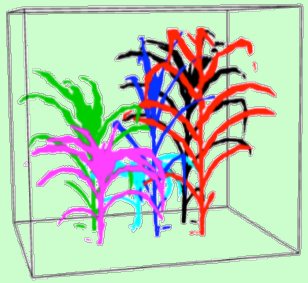
Slope = Biomass/Light Interception
= Radiation Use Efficiency
(*Cabrera-Bosquet et al*. in press)

RATP light model



Sinoquet *et al*., 2001

Reconstructed 3D Canopy



Cabrera-Bosquet et al.

Single plant light Interception

Perspectives

Cloud computing

- Use France-Grid cloud infrastructure

Distribute over several SWFMS

- Distribute computation between OpenAlea and Galaxy (bioinformatics)

Standardisation

- Provenance
- Execution of scientific workflows

Conclusion

OpenAlea is an Open Source platform for plant modeling at different scales

Scientific Workflows provides several abstractions (Composition, Mapping, Provenance)

Provide a systematic way of **describing** the scientific and data methods, and **execute** complex experiment on a variety of distributed resources.

The challenge is to link **Phenotyping data to models** to predict environmental traits.

The infrastructure need to be **transparent** for scientists (end-users)

References

Pradal, C., Artzet, S., Chopard, J., Dupuis, D., Fournier, C., Mielewczik, M., Nègre, V., Neveu, P., Parigot, D., Valduriez, P. & Cohen-Boulakia, S. (2016). InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid. Future Generation Computer Systems.

Pradal, C., Fournier, C., Valduriez, P., & Cohen-Boulakia, S. (2015). OpenAlea: scientific workflows combining data analysis and simulation. In Proceedings of the 27th International Conference on Scientific and Statistical Database Management. ACM.

Acknowledgements

VirtualPlants (Inria)
C. Fournier, S. Artzet,
J. Chopard

Zenith (Inria)
P. Valduriez,
D. Parigot, D. Dupuis

ScanAlea / Lepse
X. Sirault (CSIRO)
J. Guo (CSIRO)
F. Tardieu (INRA)
Tsu Wei Chen (INRA)
P. Neveu (INRA)
M. Mielewczik, V. Negre