Smart rerun of data intensive scientific workflow in distributed environment

Gaëtan Heidsieck

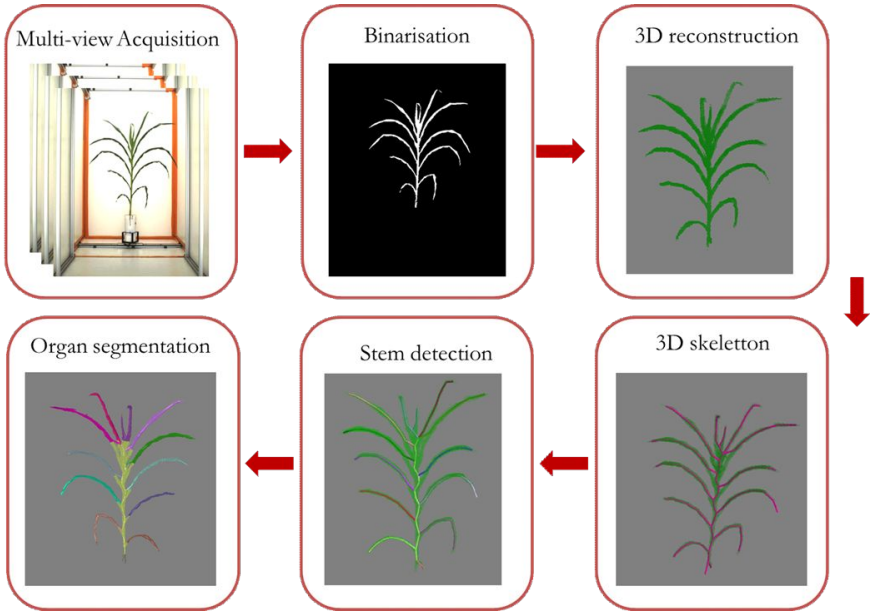
PhD supervisors : E. Pacitti, F. Tardieu, C. Pradal

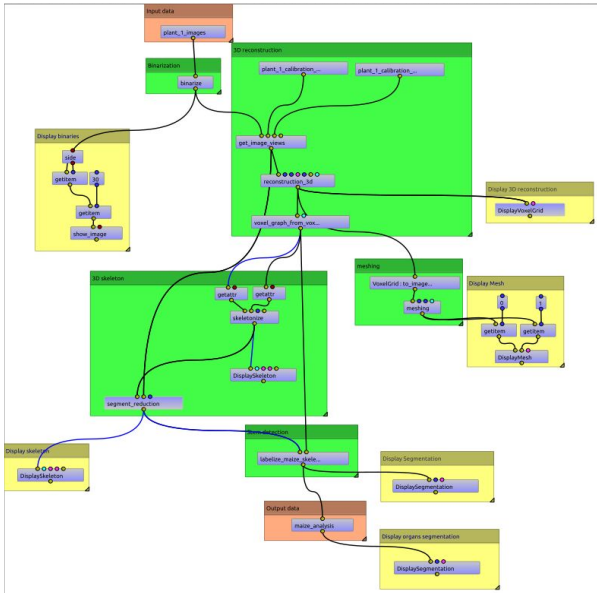
PhenoArch : High-throughput phenotyping platform

- Study the impact of different environmental conditions for various genotypes.
- Quantify plants by Imaging
- Automatic High-throughput system
  - Imaging (12 sides & top view)
    - 52 GB/day
    - 6 TB/essay
    - 11 TB / year
  - Watering and whole-plant transpiration
    - Temperature + weight measured every day

F. Tardieu, L. Cabrera-Bosquet, T. Pridmore T, M. Bennett (2017) Plant Phenomics, From Sensors to Knowledge. Current Biology 27(15):R770-R783

Phenomenal : scientific workflow for HTP image analysis





C. Pradal, S. Artzet, J. Chopard, et al. (2017) InfraPhenoGrid: A scientific workflow infrastructure for plant phenomics on the Grid. Fut. Gen. Comp. Sys. 67: 341-353

Context

Data-intensive scientific workflow execution

- Generally generates provenance data
- Classic execution environment
  - Cluster: Limited resources
  - Cloud: hard to handle the geographically distributed big data

Incremental way of designing workflows

- Several users developing the workflow
- Improvement and update on the workflow after analysing output data

Our goal : How to efficiently execute or re-execute data intensive workflow taking into account provenance data and intermediate data shared among users?