Life Science Workflow Services

– LifeSWS -

motivations and architecture

Patrick Valduriez

Life Science Workflow Services (LifeSWS): motivations and architecture

Reza Akbarinia[1], Christophe Botella[1], Alexis Joly[1], Florent Masseglia[1], Marta Mattoso[2], Eduardo Ogasawara[3], Daniel de Oliveira[4], Esther Pacitti[1], Fabio Porto[5], Christophe Pradal[1,7], Dennis Shasha[6], and Patrick Valduriez[1]

[1] Inria, Univ Montpellier, CNRS, LIRMM, France
[2] Federal University of Rio de Janeiro, Brazil
[3] CEFET/RJ, Brazil
[4] Fluminense Federal University, Brazil
[5] LNCC, Brazil
[6] New York University, USA
[7] CIRAD, AGAP Institute, Univ Montpellier, INRAE, Institut Agro Montpellier, France

Transactions on Large-Scale Data and Knowledge-Centered Systems (to appear)

Outline

- Context and motivations
- Use cases
- Related work
- Objectives
- Architecture
- Platforms
- Research directions

Context: life science

- The study of living organisms (plants, humans, micro-organisms, . . .) and their association with internal or external conditions
- Interdisciplinary research domain spanning agronomy, biology, botany, etc.
  - Very strong in the city of Montpellier
    - CIRAD, INRAE, Inria, IRD, Univ Montpellier, SupAgro
- Many challenges due to climate change
  - Adaptation, resilience, epidemics, land-use conflicts, biodiversity conservation
  - Example of practical question: how to select or breed better plant varieties?

Life Science Data and Workflows

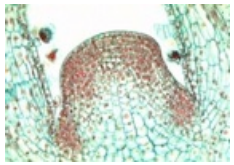
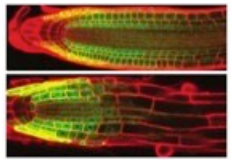
- The data comes from many different data sources
  - Modern platforms, e.g., high-throughput phenotyping, next-generation sequencing, remote sensing, etc.
  - International databases, e.g., Data.World, GenomeHub, AgMIP, EMPHASIS, etc.
- Such data can be used to
  - Produce and train models, e.g., ML models
  - Derive information and knowledge or make predictions using complex workflows
- Example: plant modeling to predict impact
  - Heterogeneous data: plant phenotype, plant genotype, environment data (meteo, soil)
  - Analyze through multiple workflows

Models in Life Science

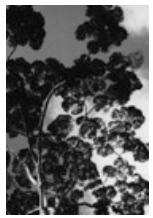
- Statistical machine learning models
  - Find patterns in existing data
  - Simple equations, derived from statistics and regression analysis
  - A lot of data is needed (the more data, the better)
- Mechanistic models
  - Derived from the mathematical modeling of a phenomenon capturing the fundamental laws of natural sciences
  - Less data needed to calibrate the model
  - Running a mechanistic model for a certain number of time steps allows simulating the phenomenon
    - Example: crop simulation models reproduce the main functions of plants such as the evolution of plant architecture, light interception, photosynthesis, and water/nitrogen balance in the crop and soil

Multiscale Plant Modelling

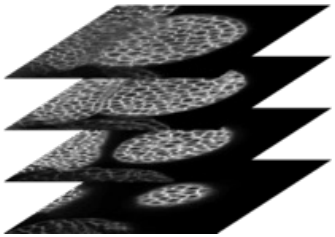
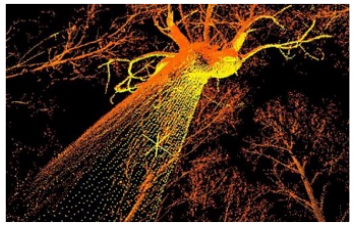
Different scales



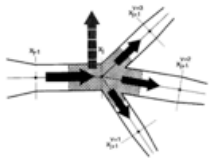



Data acquisition

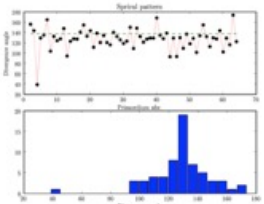

Modeling

Mechanistic

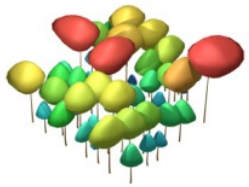
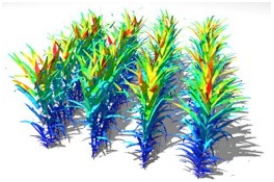


$$\frac{\partial}{\partial x}\left[K(x,P)\frac{\partial \Psi}{\partial x}\right] = c(x,P)\frac{\partial \Psi}{\partial t} + E(t)l(x)$$

Statistical

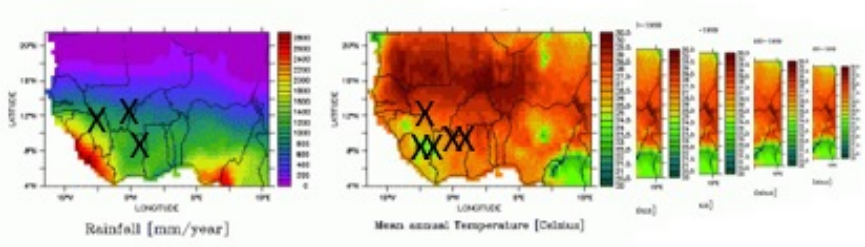


Simulation

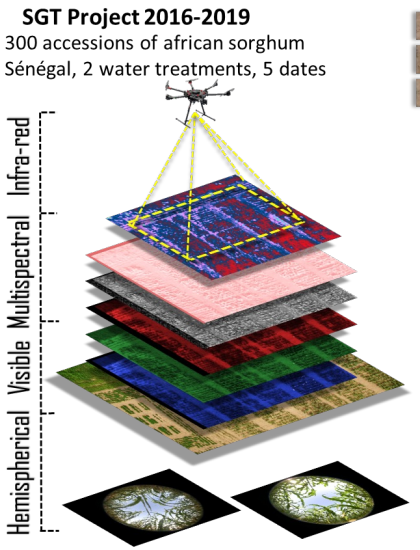



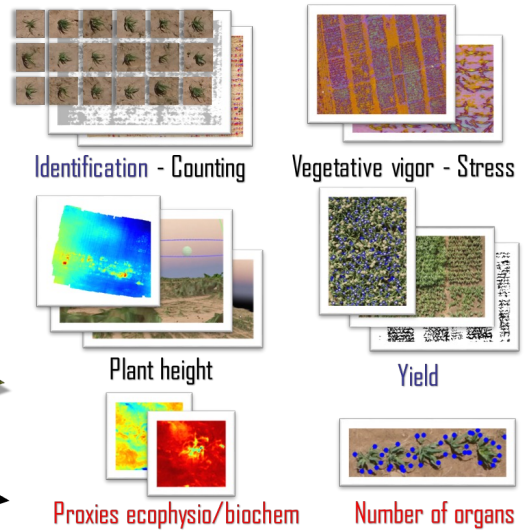
Big Data Sources

- International databases
  - Genomics (Elixir, GenomeHub)
  - Soil, climate, meteo
  - Phenomics
    - Image
    - 3D point cloud (T-Lidar)
  - …

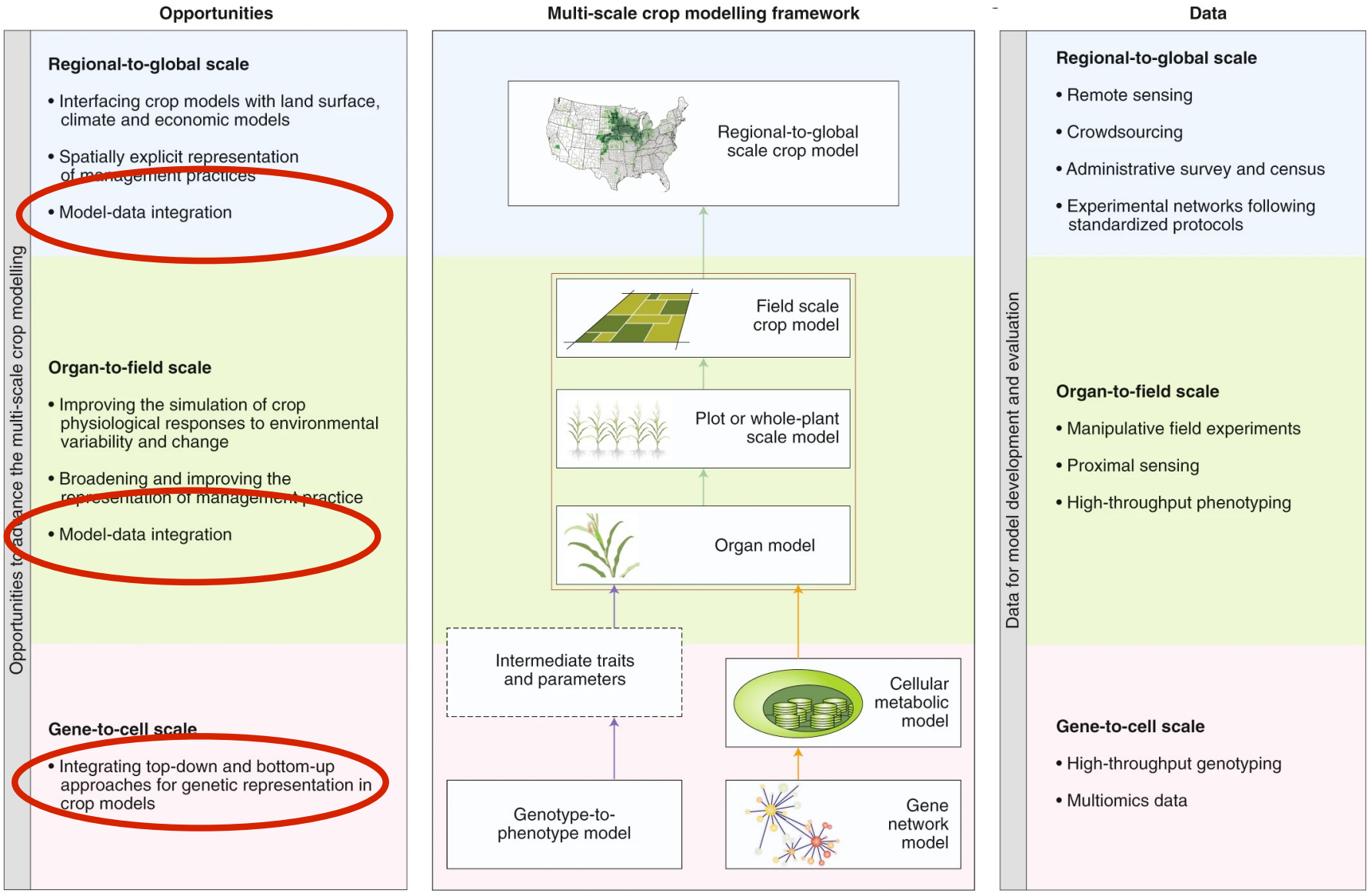


- Data produced by platforms and models
  - Time series
  - 3D data





Multiscale Crop Modeling Framework

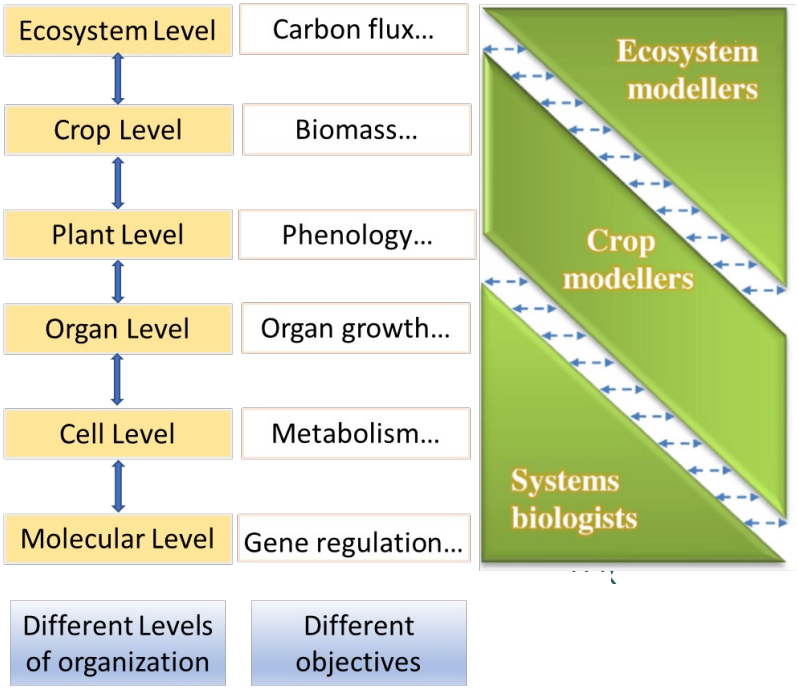


B. Peng et al.: Towards a Multiscale Crop Modelling Framework for Climate Change Adaptation Assessment. Nature Plants, 2020

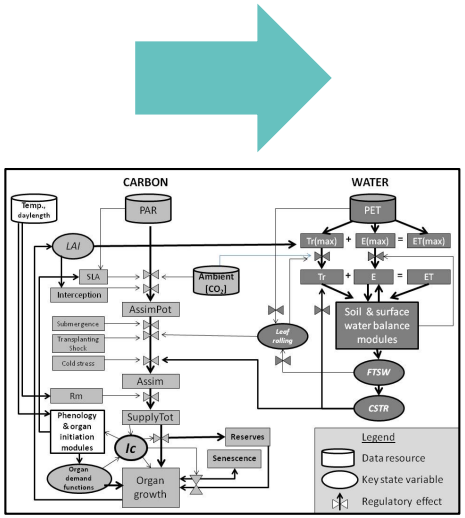
Model Intercomparison and Improvement Project (AgMIP)

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Develops simulation crop models to analyze and predict plant and crop development from molecular to ecosystem levels



40+ AgMIP wheat models in use



Use Case: plant phenotyping

- High-throughput phenotyping platforms
  - Enable the collection of quantitative data on thousands of plants under controlled environmental conditions
    - Temperature, humidity, drought, light, etc.
- Scientific workflows
  - To analyze, reconstruct, and visualize the spatial and temporal development of the geometry and topology of plants in various environmental conditions

PhenoArch platform
INRAE, Montpellier

OpenAlea

Workflow system
CIRAD & Inria, Montpellier

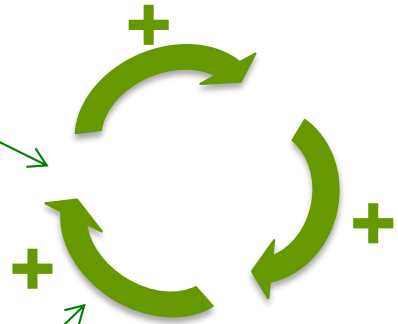
C. Pradal, C. Fournier, P. Valduriez, S. Boulakia
Openalea: scientific workflows combining data analysis and simulation. SSDBM 2015

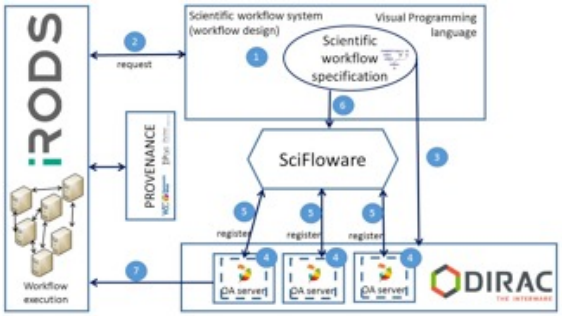
OpenAlea Phenomenal Workflow

- Coupling high-throughput phenotyping analysis with biophysical models

PhenoArch platform, Montpellier

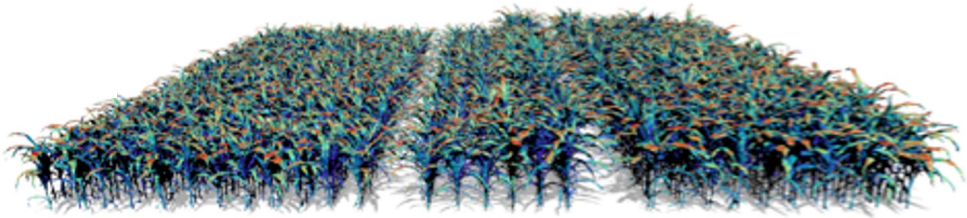
Biophysical data





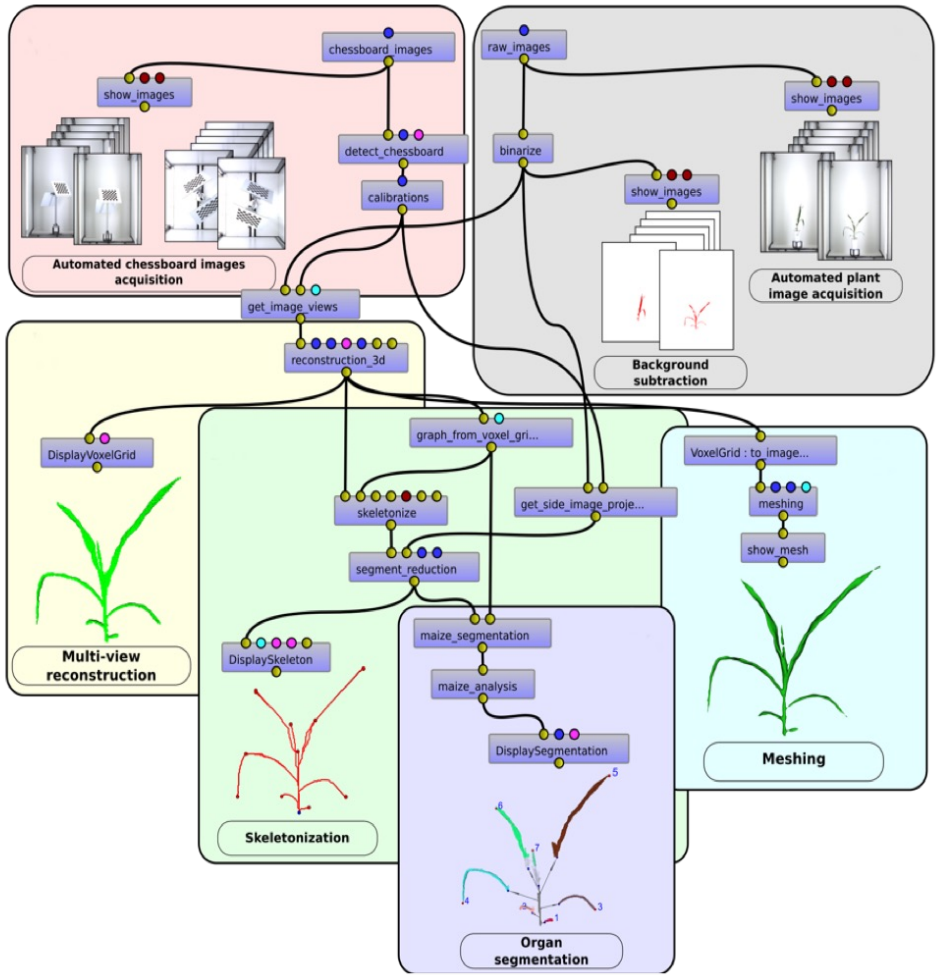
- **InfraPhenoGrid**: a grid infrastructure for phenotyping
- **Phenomenal**: automatic 3D shoot reconstruction

OpenAlea



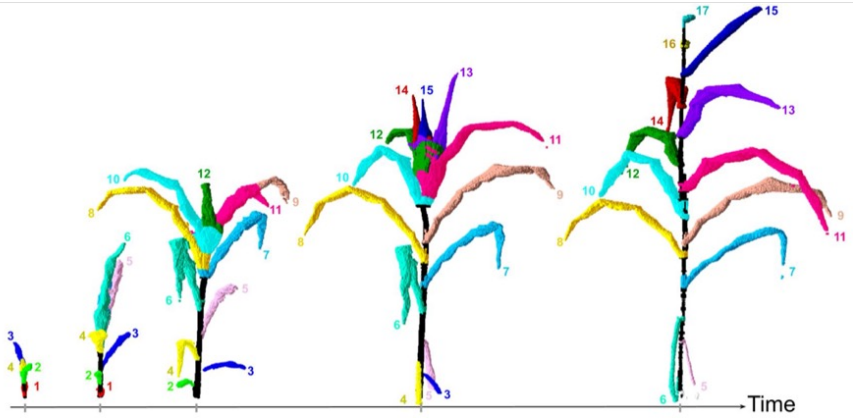
G. Heidsieck, D. de Oliveira, E. Pacitti, C. Pradal, F. Tardieu, P. Valduriez
Cache-aware Scheduling of Scientific Workflows in a Multisite Cloud, FGCS 2021

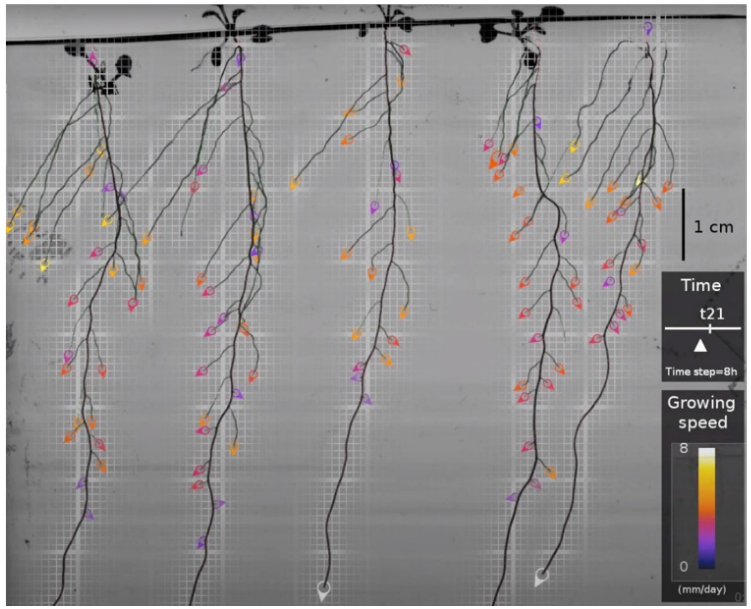
OpenAlea Workflows



Phenomenal workflow

3D organ tracking of a maize plant with PhenoTrack3D workflow





Reconstructed root system architecture using RootSystemTracker workflow

Use Case Requirements

- Efficient execution of OpenAlea workflows
  - Parallel transfer of large image datasets
  - Distributed execution within a cluster
  - Caching of intermediate results for later use
- Ease of use
  - Reproducibility of executions using provenance information
  - Dashboards to ease reruns of workflows with different parameters and display execution results
- Integration with other workflows
  - To understand the genotype-to-phenotype relationships, and relate plant traits with genotyping workflows, e.g., Galaxy

Related Work

- Wide spectrum from generic to specific
- Cloud services
  - Many ready-to-use services within a PaaS
  - Lack of services for scientific applications
  - Vendor lock-in
- Data-based systems
  - Scientific workflows: Galaxy, Kepler, OpenAlea, etc.
  - Data analytics: Spark, Flink, etc.
  - Polystores: BigDAWG, BigIntegrator, CloudMdSQL, etc.
- Model life-cycle frameworks
  - MLaaS by cloud providers, Mlflow, ProvLake, etc.

Related Work (cont.)

- Science platforms
  - Provide services and resources for research communities to perform collaborative research, observation and experimentation
- More or less specialized for some particular science
  - InfraPhenoGrid: grid-based platform for plant phenomics
  - PHIS: phenotyping hybrid information system
  - Pl@ntnet: participatory platform for the production and sharing of botanical data
  - CyVerse: platform for life sciences with services and resources to deal with huge datasets and complex data analyses

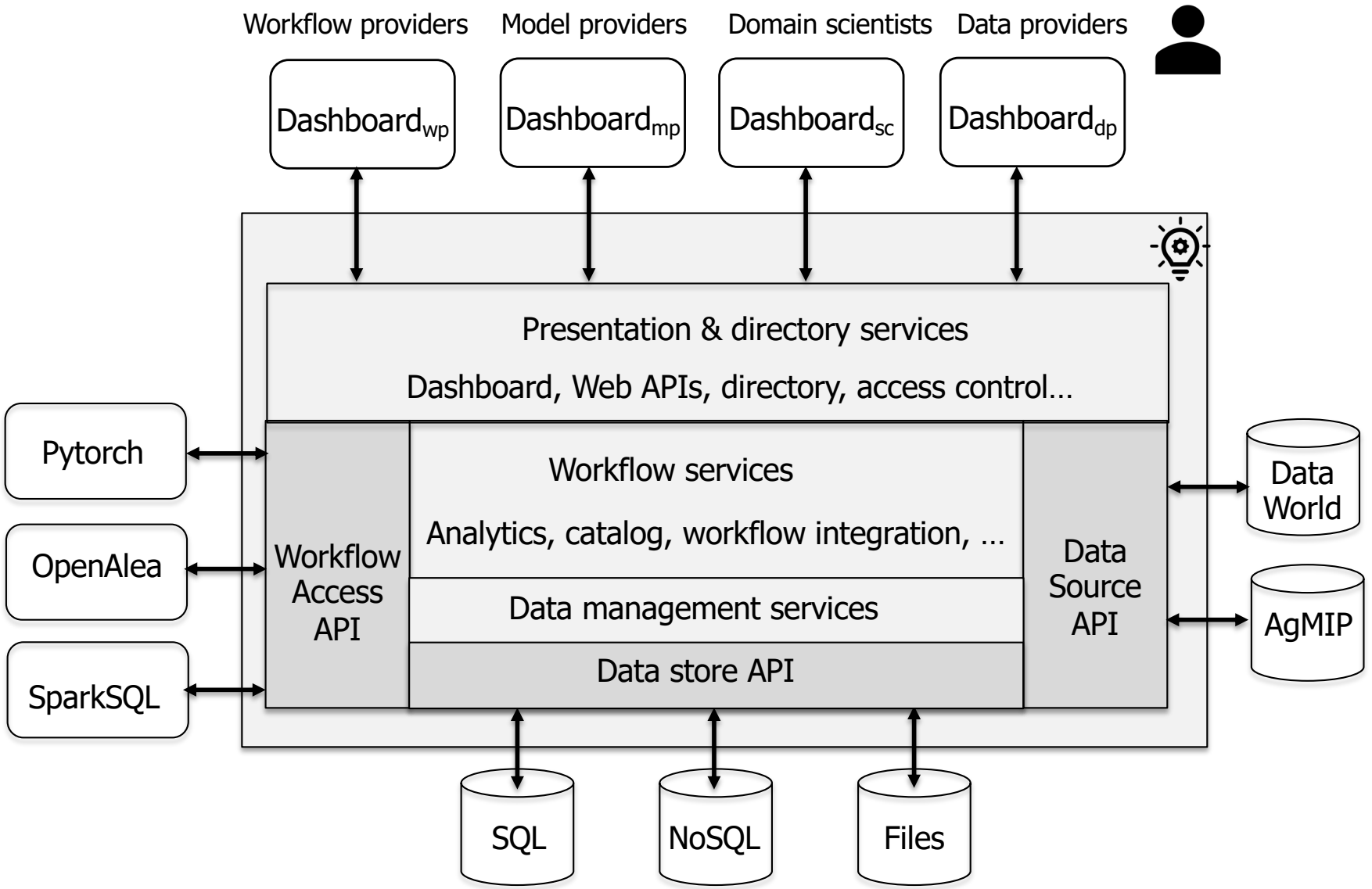
LifeSWS: objectives

- Open service-based architecture
  - Data analysis workflow services for life sciences
- Organize massive and heterogeneous data
  - In connection with models
- Make workflow artifacts easy to search, debug, and parallelize
  - Artifacts = datasets, models, metadata, workflow components, etc.
- Technical goal
  - Make workflows work as seamlessly with data as queries do in business data processing

LifeSciDS: principles

- Ease of use through web interfaces
  - For different kinds of users
  - Access to various tools and execution environments
- Open, composable architecture with well-defined APIs
  - To foster services interoperability
- Distributed architecture
  - Performance and scalability in the cloud using distributed database principles
- Integrated services
  - Local (in the same data center) or remote (in remote data centers)
  - Support for various databases (SQL, NoSQL, SciDB, etc.)
  - Support for various scientific files (HDF and NetCDF)

Architecture



Presentation and Directory Services

- Web dashboard service to build specific dashboards for different types of users
  - Domain scientists, workflow providers, model providers and data providers
  - These dashboards allow users to analyze and display real-time data as charts and reports
- Directory
  - Manages data about LifeSWS users, access rights, dashboards and services
  - Basis for secured access to services
- Web server-side API
  - Allows LifeSWS developers to access LifeSWS services from more general Web applications
- External data view to ease the development of dashboards and workflows
  - Can be represented by a knowledge graph

Workflow Services

- Make it easy for scientists to develop, debug and optimize their workflows
  - Sharing of data, models and workflow components
  - Model execution using different tools and data sources
- Primary services
  - Catalog (including version management)
  - Model management
  - Workflow integration
  - Data analytics

Model Management

- Support of various types of models
  - Machine learning models
    - Statistical, deep learning, …
  - Mechanistic models
- LifeSWS added value
  - Through a unified view of different model artifacts (produced with different tools), LifeSWS can improve model selection and allow for model integration

Model Integration

- ML model integration
  - Considering an ensemble of models, i.e., a set of models aiming at the prediction of the same target*
    1. Run each individual participant model, possibly across different tools, over the same input
    2. Produce an integrated result, often using a linear combination of the results
- Mechanistic model integration
  - Requires workflow integration (see next)

*R. Zorrilla, E. Ogasawara, P. Valduriez, F. Porto
A Data-Driven Model Selection Approach to Spatio-Temporal Prediction. SBBD 2022

Workflow Integration

- Support for integrating and efficiently executing workflows on different systems using the Workflow Access APIs
  - Some similar goals and functions found in data integration
- Workflow definition
  - Candidate: Common Workflow Language (CWL), an open standard to make workflows portable, reusable and easy to share
  - CWL is explicit about inputs/outputs to form the workflow, data locations and execution models, which can be deployed using software container technologies, such as Docker and Singularity

Workflow Integration (cont.)

- Workflow execution
  - Once an integrated workflow has been defined and its mappings registered, e.g., using CWL, it can be executed using a LifeSWS scheduler that orchestrates execution across different workflow systems, in connection with these systems' schedulers
- Provenance
  - Helps to reproduce, trace, assess, understand, and explain how datasets have been produced
- Caching of intermediate datasets
  - The decision whether to cache can be explicit or made automatic based on workflow fragment analysis

LifeSWS Platforms

- Implement and deploy LifeSWS services to address the specific requirements of vertical applications
  - Using different software and hardware infrastructures
  - Reusing software components that (partially) implement the services
- Examples of deployments
  - Laptop
  - On-premise cluster of servers
  - Cloud (public, private or hybrid)
- Gypscie as a LifeSWS platform

The Gypscie Platform*

- Data Management using SAVIME
  - SAVIME: in-memory array database system
    - Enables simulation real-time monitoring
  - Registration, transformation and metadata description
  - Data locality for transformation and training
  - Querying
- ML Model Management
  - Model building, importing and serving
  - DJEnsemble method for automatic model composition
  - Model Metrics management
- Event detection
  - Integrates the Harbinger component
  - Algorithms for offline event detection
- Multiple Execution Environments
  - Santos Dumont Supercomputer, Spark cluster

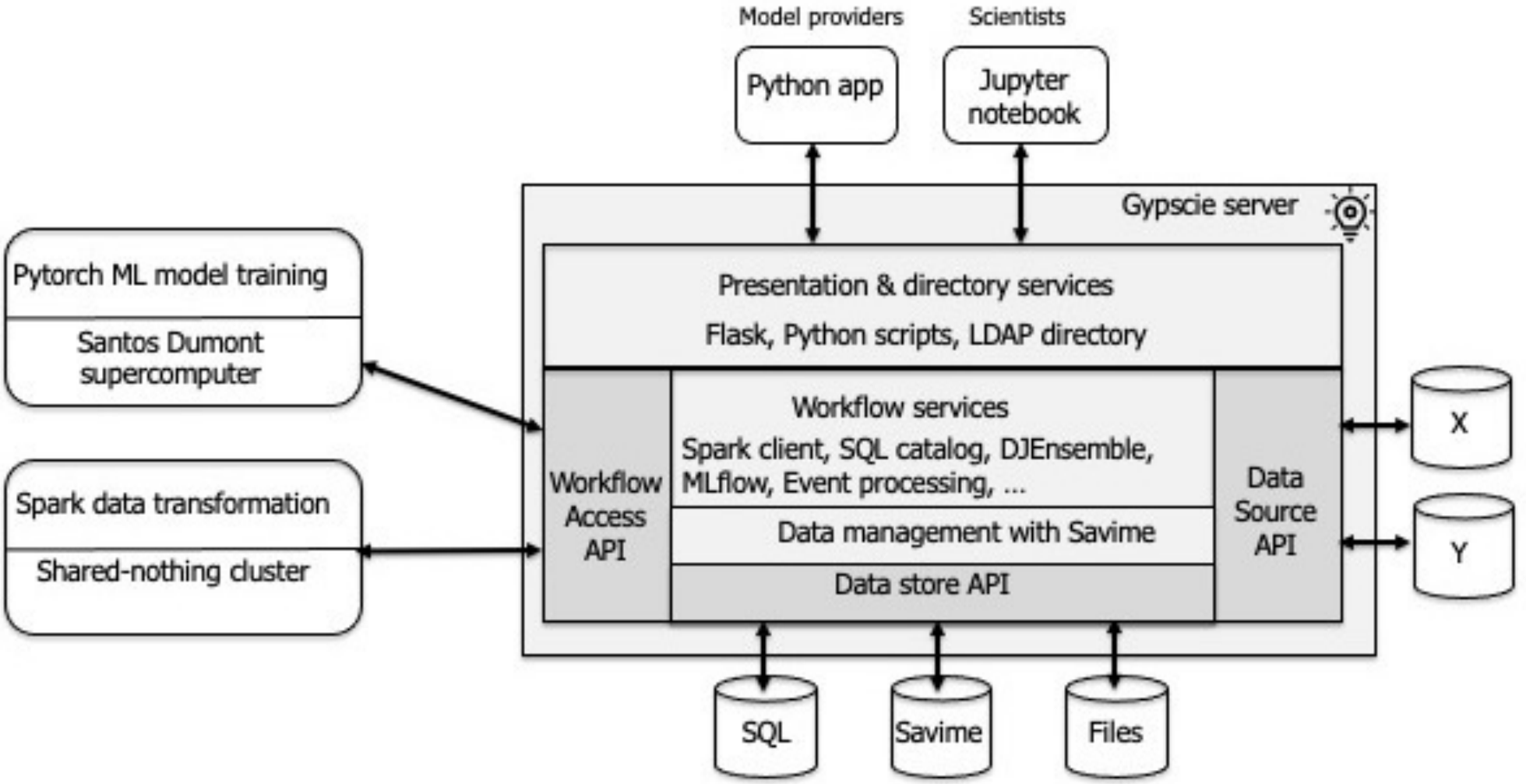
*F. Porto, P. Valduriez. Data and Machine Learning Model Management with Gypscie. CARLA Workshop on HPC and Data Sciences meet Scientific Computing, 2022

DJEnsemble

- Cost-based selection and allocation method of a disjoint ensemble of spatio-temporal models
  - Large data domains partitioned into subdomains according to data similarity properties (e.g., with time series data, we place similar series together)
  - Instead of a single model covering diverse data patterns, opt for specialized models and combine them when needed;
  - Automatic algorithm to combine subdomain models to answer a predictive query
    - Exploits data similarity between query region and models' training regions

R. Zorrilla, F. Porto, E. Ogasawara, P. Valduriez. A Data-Driven Model Selection Approach to Spatio-Temporal Prediction. Nominated for best paper, SBBD 2022 Conf.
R. S. Pereira, et al., DJEnsemble; A Cost-based Selection and Allocation of a Disjoint Ensemble of Spario-temporal Models, SSDBM 2021 Conf.

Gypscie Platform Architecture



Research Directions

- Make it easy to integrate and run heterogeneous workflows
  - Using the Common Workflow Language (CWL)
- Provide efficient execution of heterogeneous workflows
  - By caching intermediate results and performing cache-aware scheduling
- Make it easy for domain scientists to manage the model life cycle
  - Model selection and model integration for different types of models managed using different tools
- Assist scientists in analyzing diverse data types
  - Focus on time series
- Keep track of the provenance of both data sources and software components
  - To aid in debugging and to enhance the reproducibility of computational experiments