

Interactions « Human / Data »

Une approche intégrée pour appréhender des données complexes de phénotypage de plantes



Phénotypage de plantes à haut débit

Enjeu : Comprendre la réponse des plantes aux facteurs de l'environnement pour permettre de répondre à la demande d'une production plus importante, de meilleure qualité et plus économe en intrants



Le réseau français de phénomique végétale

- Infrastructure nationale
- Caractériser de grandes séries de génotypes nécessaires pour les études de variabilité génétique, dans des scénarios environnementaux divers





Un défi pour les maths - info

- ❖ De grandes masses de données observées dans le temps, hétérogènes, de nombreuses covariables
- ❖ Gérer / analyser ces données :
 - Organiser données et connaissance
 - Prédire / Expliquer en fonction des cofacteurs
 - ➔ Méthodes génériques
Efficaces en temps de calcul
- ⇒ Développer une nouvelle génération d'outils / méthodes transférables

Approche basée sur les data sciences

Interaction humain - données

Démarche :

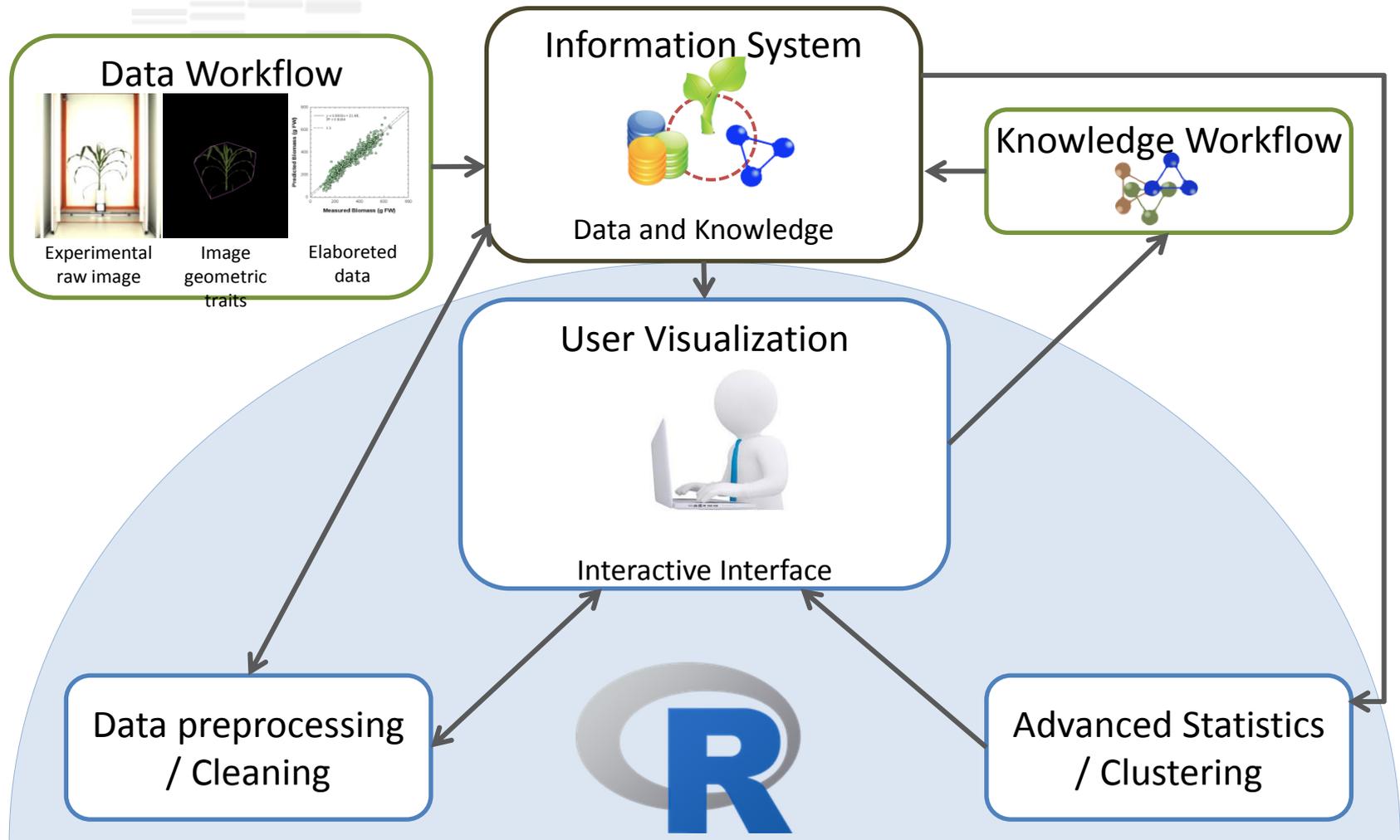
Une approche intégrée qui place **l'utilisateur au cœur du système**

Un système d'information associé à la connaissance formalisée et à des méthodes statistiques et numériques

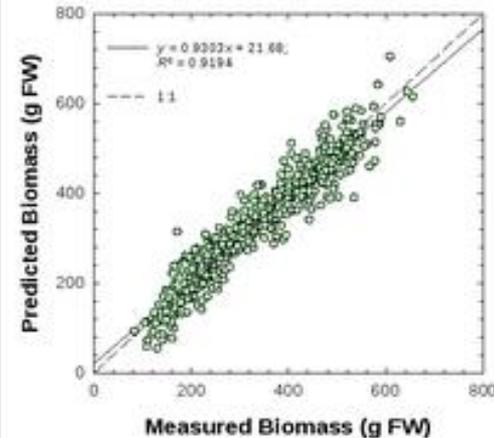
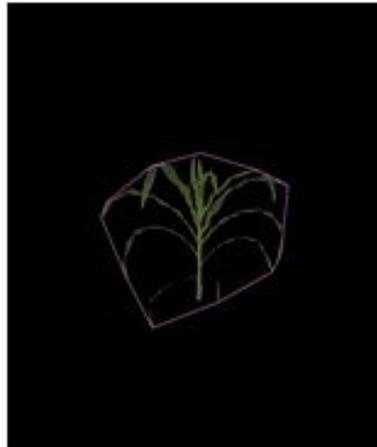
Objectifs :

- ❖ Organiser des données expérimentales complexes
- ❖ Visualiser les données à l'aide d'une interface conviviale
- ❖ Valider les données grâce à des méthodes qui combinent ingénierie de la connaissance et analyses statistiques
- ❖ Explorer les données avec des méthodes avancées (classification..)

Vue d'ensemble

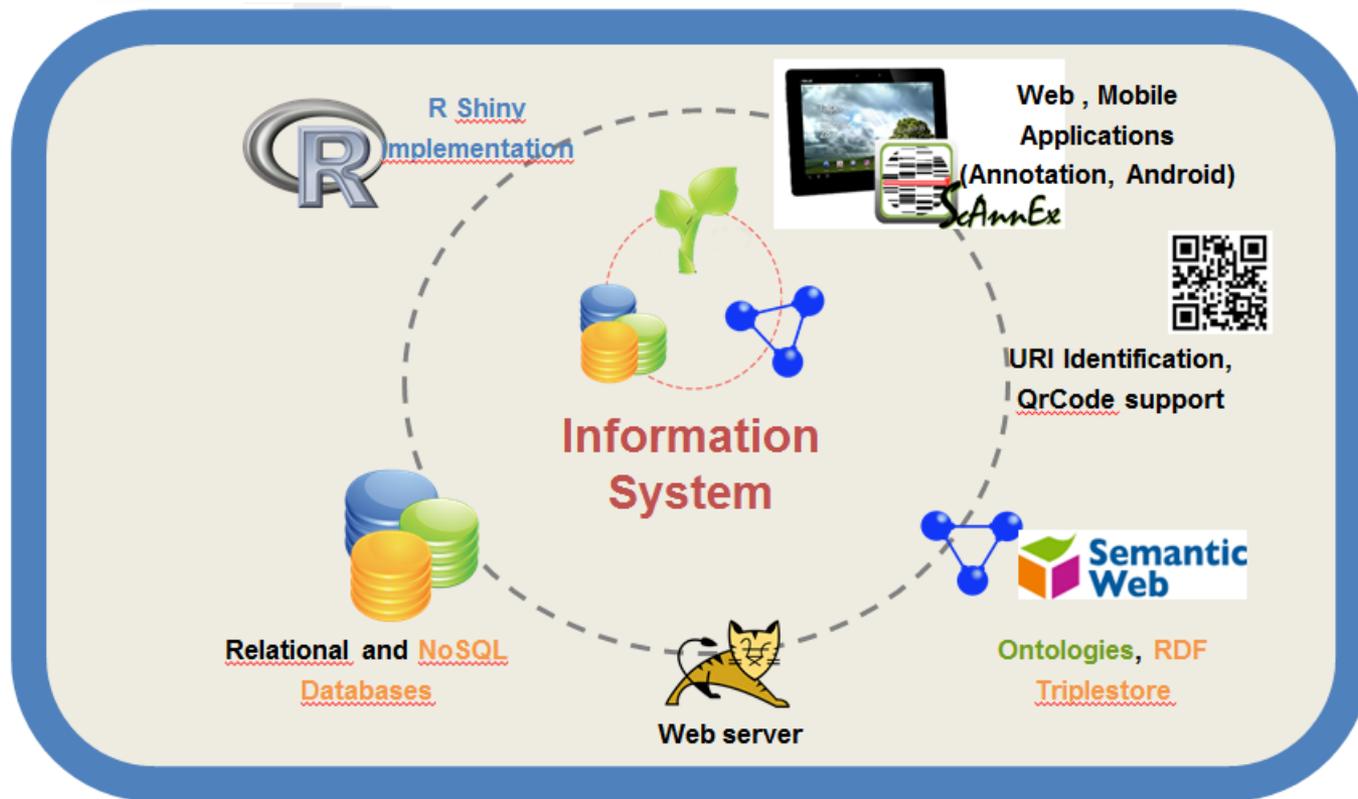


Focus 1 : les données



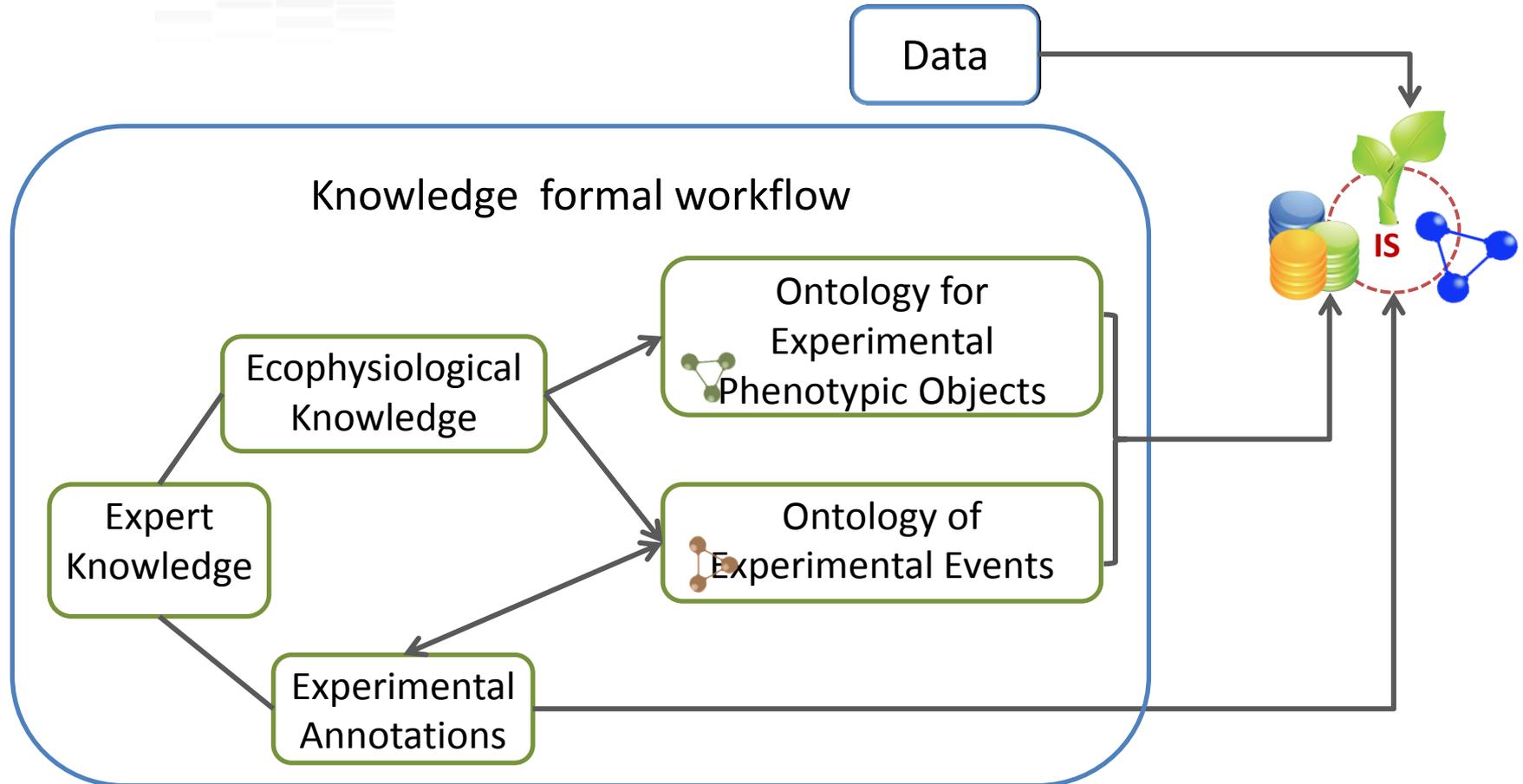
Workflow : 1/ image brute, 2/ traits géométriques intermédiaires obtenus par analyse d'image, 3/ donnée élaborée (trait phénotypique, ici biomasse) obtenue par une étape de calibration (modèle statistique)

Focus 2 : le système d'information



PHIS Information system architecture

Focus 3 : connaissance



Workflow du traitement de la connaissance

Visualisation : interface web avec graphiques interactifs

PHIS Phenotyping Hybrid Information System Experiment Calibration

Data Validation

Data selection

Experiment: ARCH2013-09-12 Genotypes: 20H2

Treatment: WD

Replicate: 1

Number of bad Extractions: 8 / 1967

[View](#)

Variable

Variables: obj_area_sum View Type and Angle: vis_sv000

Type: imagery

Rendering

Data table: display columns

experiment_id	variable_name	value	date	confidence	genotype	scenario
repetition	view	repScen				

Which plot?
 All replicates All views
 Show smoothed version

Confidence

Hide invalid points

[Save Confidence](#)



Nettoyage des données

Annoter automatiquement des points « déviants » sur une courbe comme peu fiables **dans le Système d'Information**. Les points déviants sont des mesures ponctuelles qui ne suivent pas le comportement attendu (connaissance experte).

Deux types de points déviants :

- ❖ Grossier (valeur négative, etc)
- ❖ Plus subtil : détection grâce à
 - **Ingénierie de la connaissance** : modéliser la connaissance des experts sur le comportement attendu des courbes
 - **Statistique** : utiliser cette connaissance pour évaluer si la courbe observée est compatible

Détection manuelle possible aussi.

Les points détectés déviants sont stockés dans le SI pour de prochaines utilisations ou mises à jour.



Data Cleaning

Geno choice

20H1

Plotting per:

- Scenarios
- Repetitions

Cleaning method

- Growth rate
- Sigmoid regression
- Smooth

Smooth method

Tolerance

1 2 10

1 2 3 4 5 6 7 8 9 10

Display Smooth curves

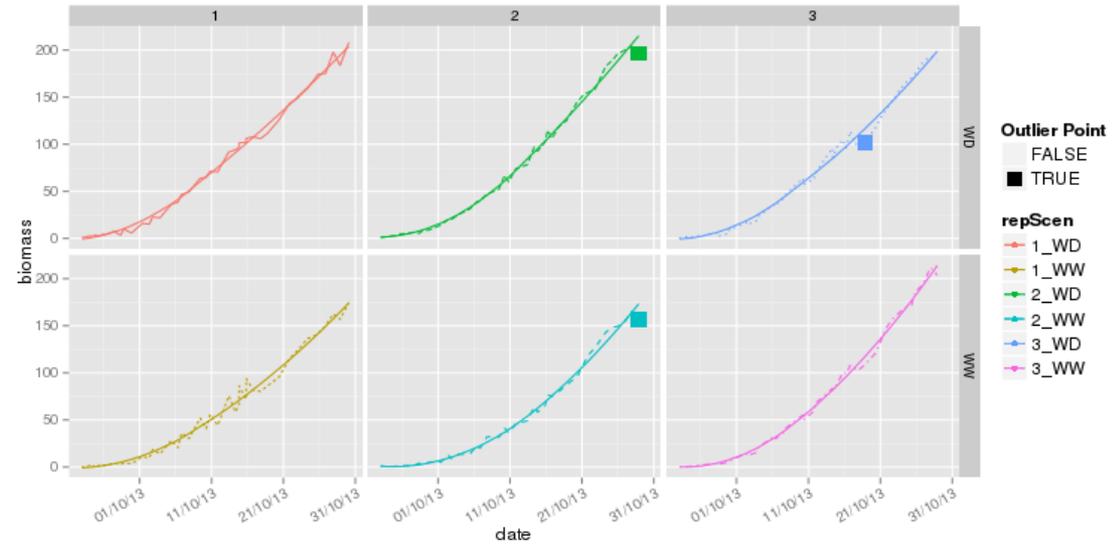
Outlier Point

Atypic plant

Clustering

Export Data

Log



Undo manual mark point

Outlier detection using smoothing



Merci de votre attention !