Modeling, statistics and optimization for plant science

Estelle Kuhn

Researcher in applied mathematics

French National Research Institute for Agriculture, Food and Environment

Department of Applied Mathematics

3 septembre 2024

French National Research Institute for Agriculture, Food and Environment

"établissement public à caractère scientifique et technologique" (EPST)

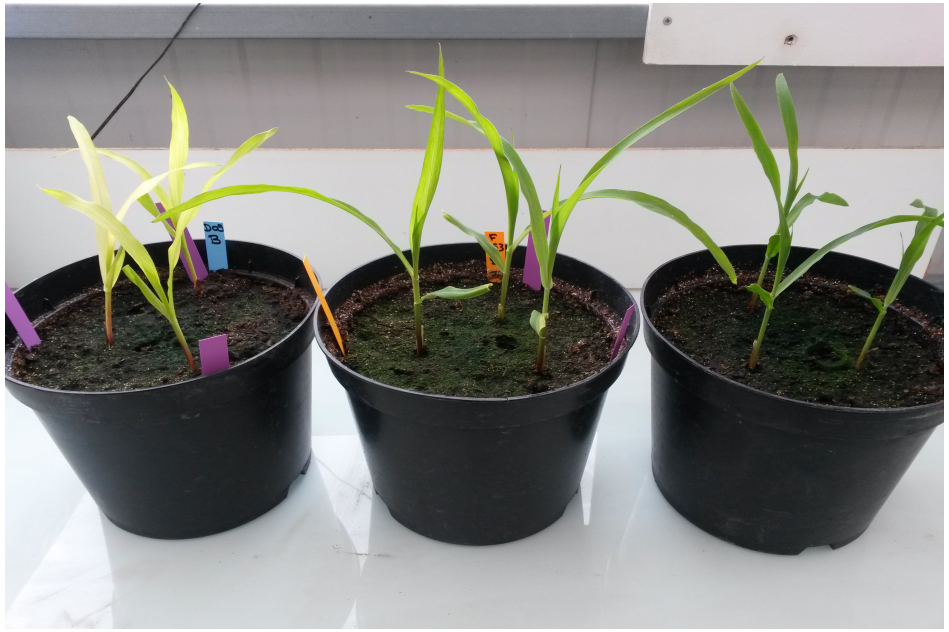
more than 10000 workers :
researchers, engineers, technicians, support staff, etc.

18 research centers located throughout France

14 disciplinary scientific departments :
biology, plant biology and breeding, genetics, animal health, plant health, agronomy and environment, microbiology and food chain, mathematics and digital technology, management sciences and economics...

Plant breeding

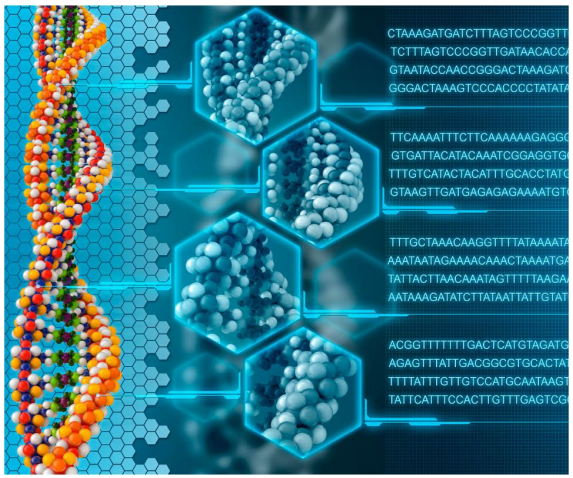
Environmental effect

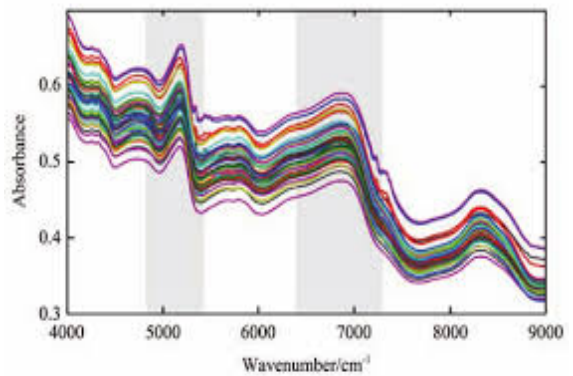


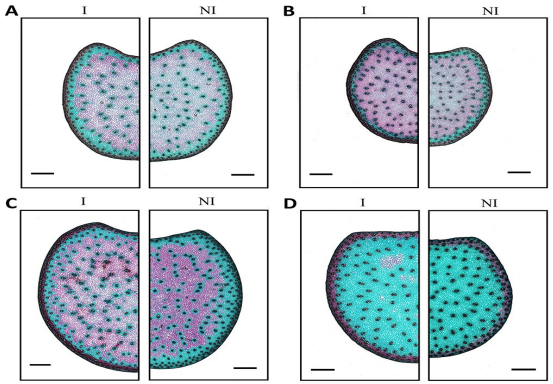
- Strong interaction between varieties (genotypes) and environments (climate, soil, crop management ...)
- Challenge : find *efficient* and *adapted* varieties to a fixed or *random* environment

Data available

- large volume
- heterogeneous sources
- large dimension
- structured
- expert knowledge







Challenge :
make the best use of the information contained in this data

Massive data acquisition

phenotyping platforms in controlled conditions
$\Longrightarrow$ measurement of biomass, height, yield

Massive data acquisition

open-field phenotyping platforms
$\Longrightarrow$ in semi-controlled conditions

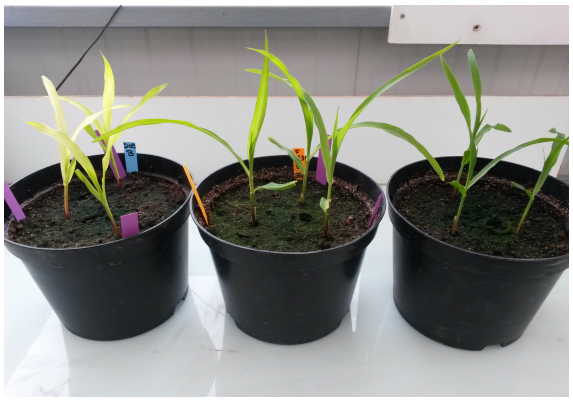
A modeling approach for what purpose ?

better understand complex phenomena
via digital tools

$\Longrightarrow$ complementary to the experiences

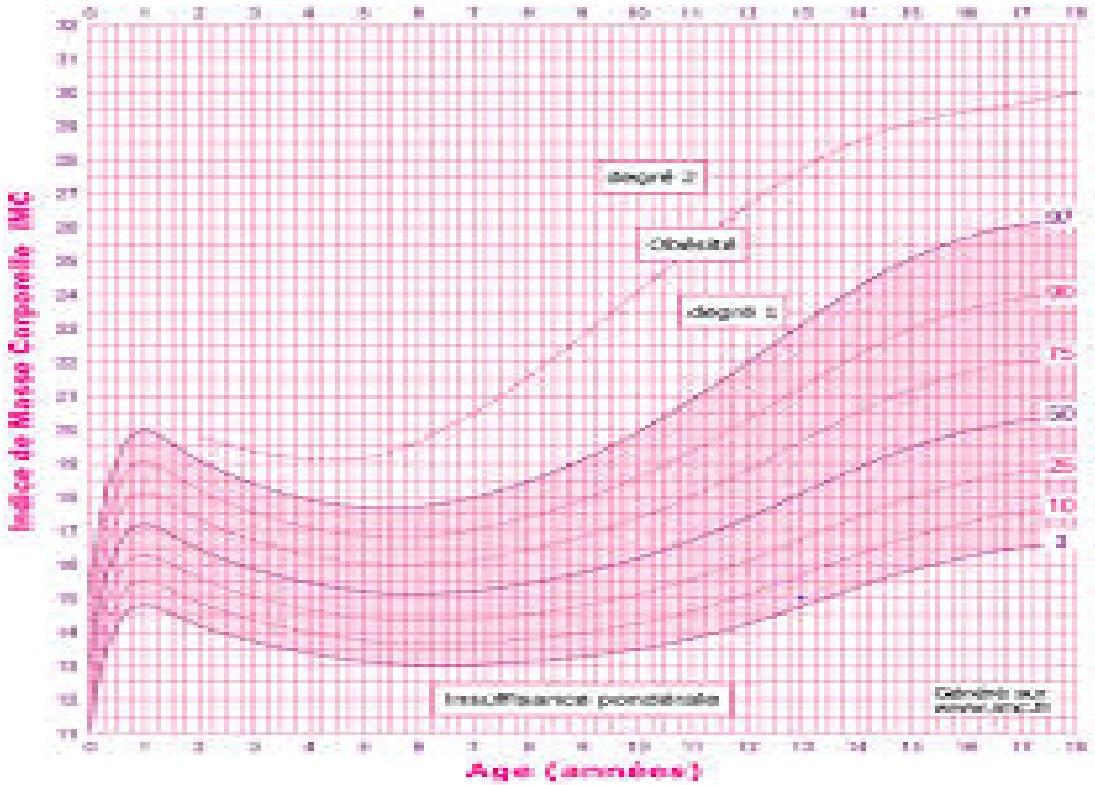
understanding

generalizing



forecasting

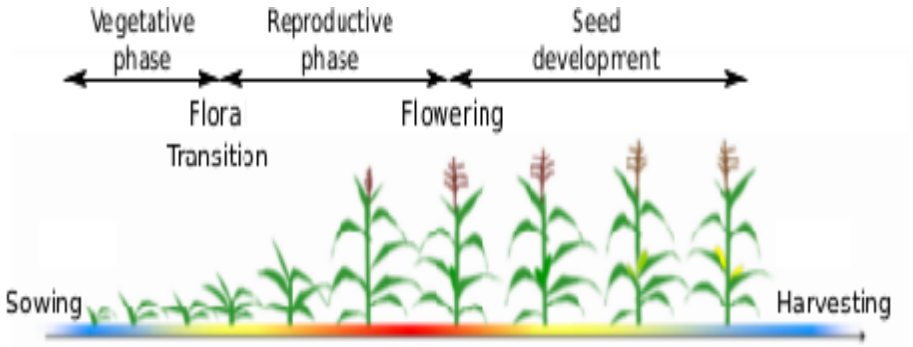
Short reminder of modeling



IMC=function(age,parameter)

- input : age, parameter
- output : IMC
- function : possibly complex system of equations

Plant Growth Process

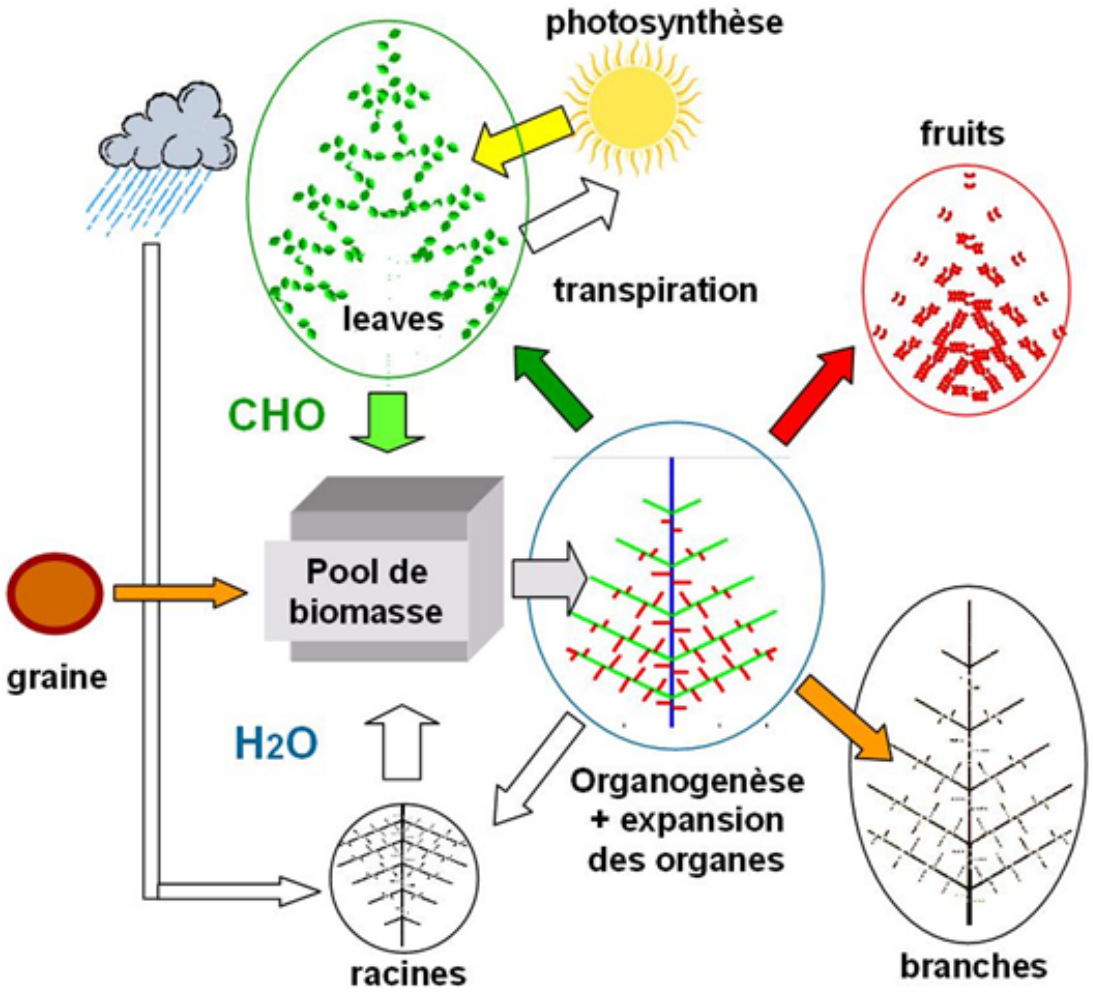


- variable of interest : yield, biomass, flowering date, leaf emergence dates, harvest date
- descriptive variables of the environment : temperature, precipitation, soil composition, technical route

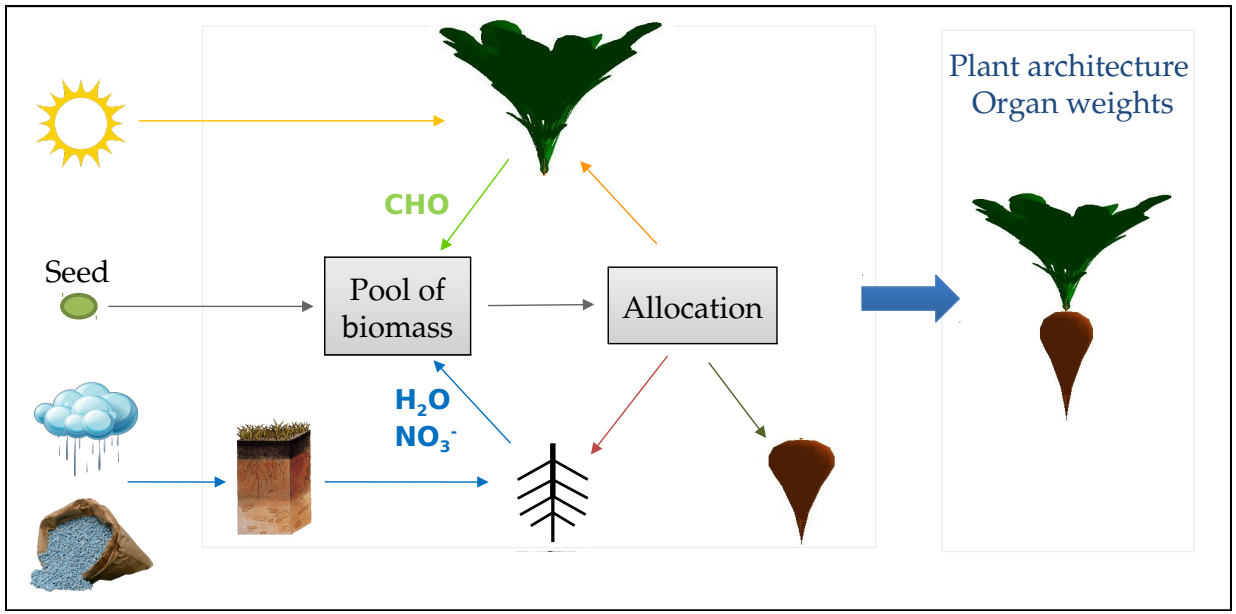
$\Longrightarrow$ describe growth processes precisely

$\Longrightarrow$ plant ecophysiology

Representation of a crop model



Example of models with physical constraints



What level of complexity ?

Essentially, all models are wrong, but some are useful.
(George E.P. Box)

gradient of models

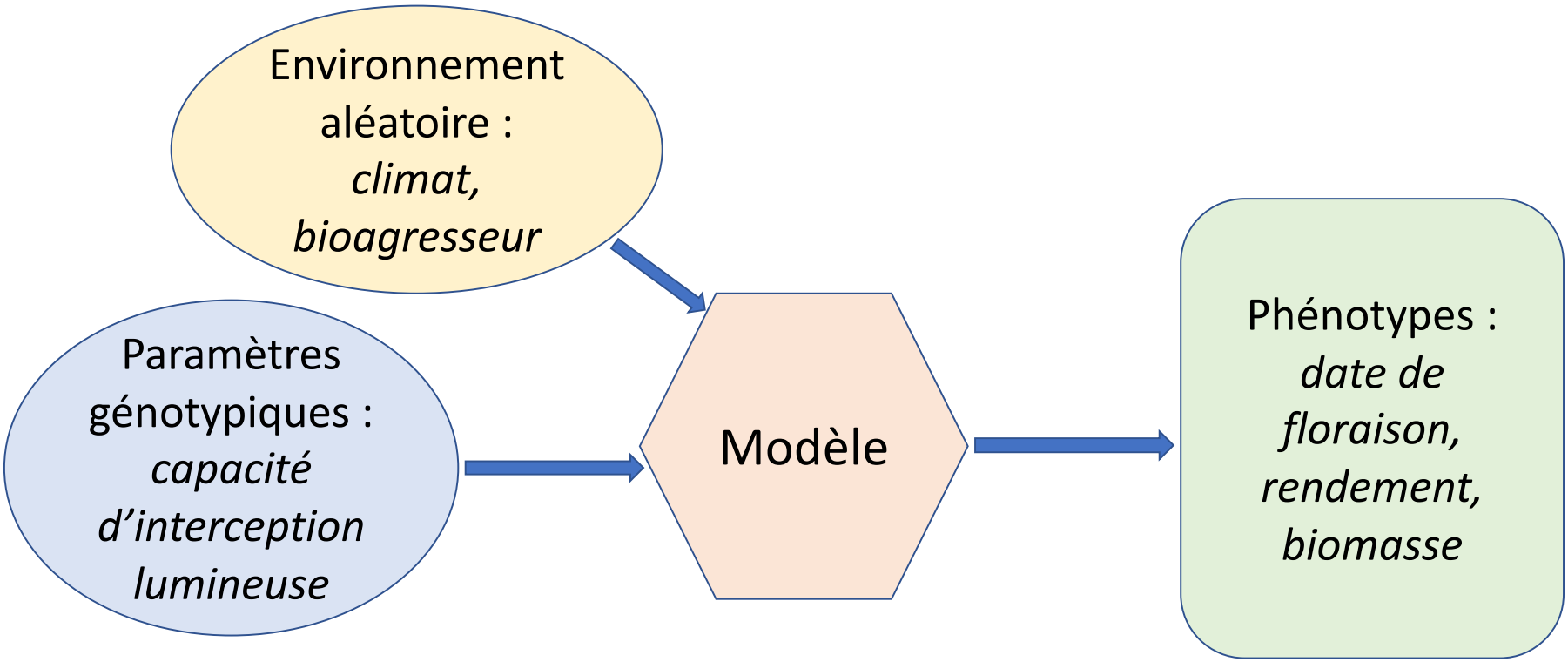
$\Longrightarrow$

simple model	complex model
few parameters	many parameters
few constraints	many constraints
generic	specific
low fit	overfit
....	...

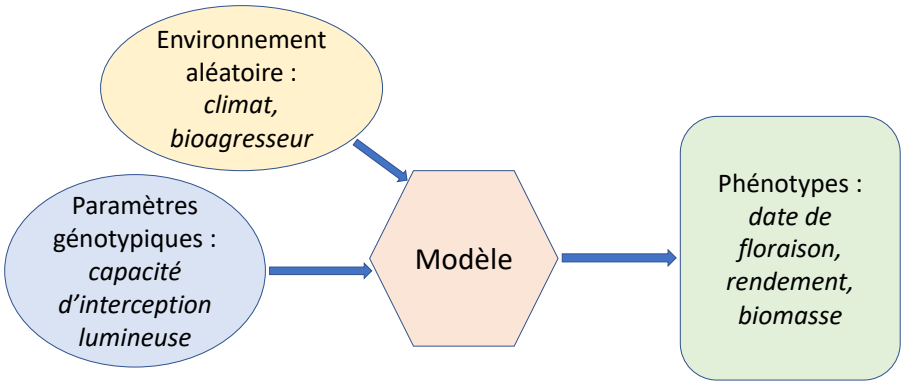
Some other important elements :

- calculation time/effort
- generalization ability
- genericity/specificity of numerical methods
-

Model the different sources of variability

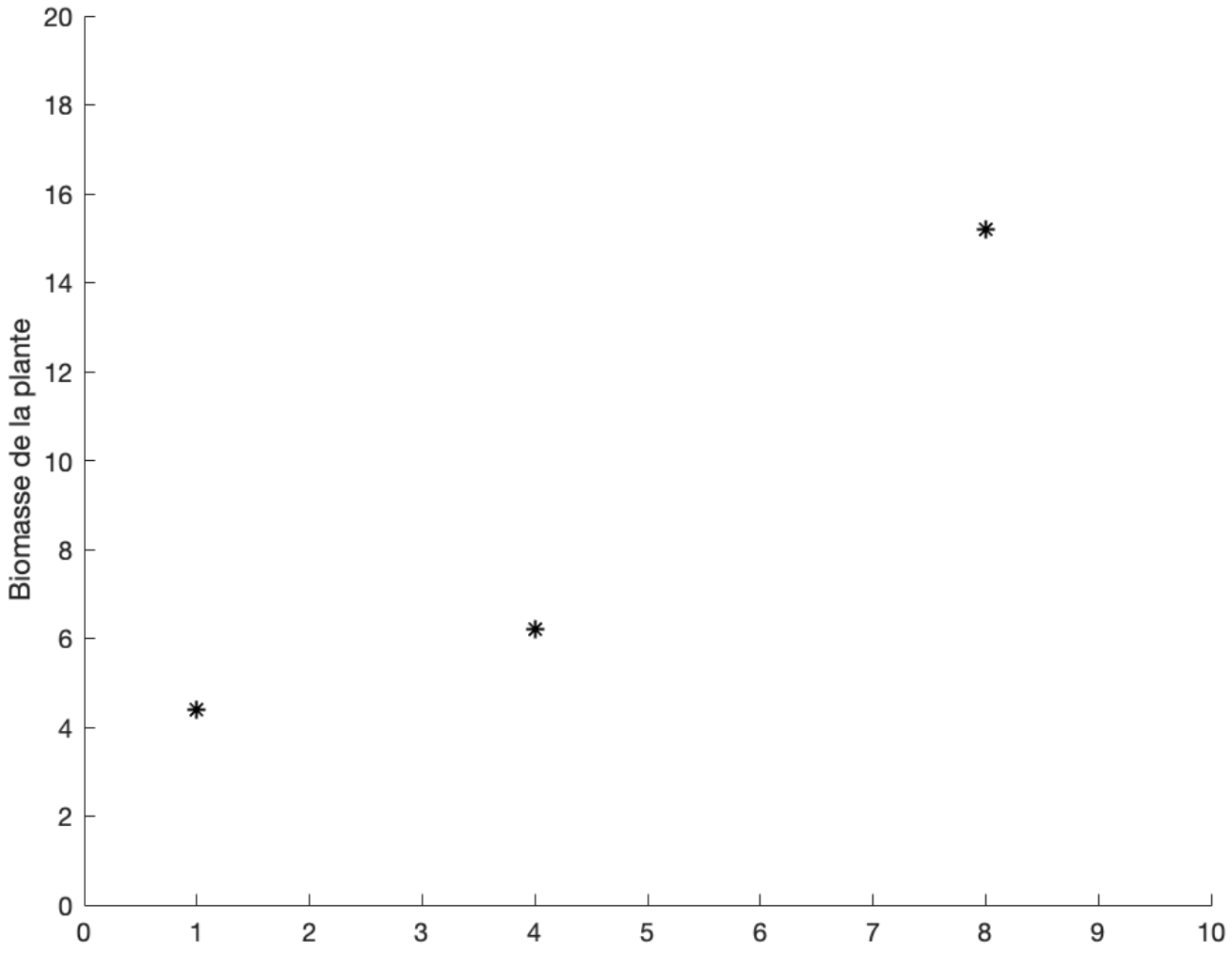


Parameter estimation/calibration

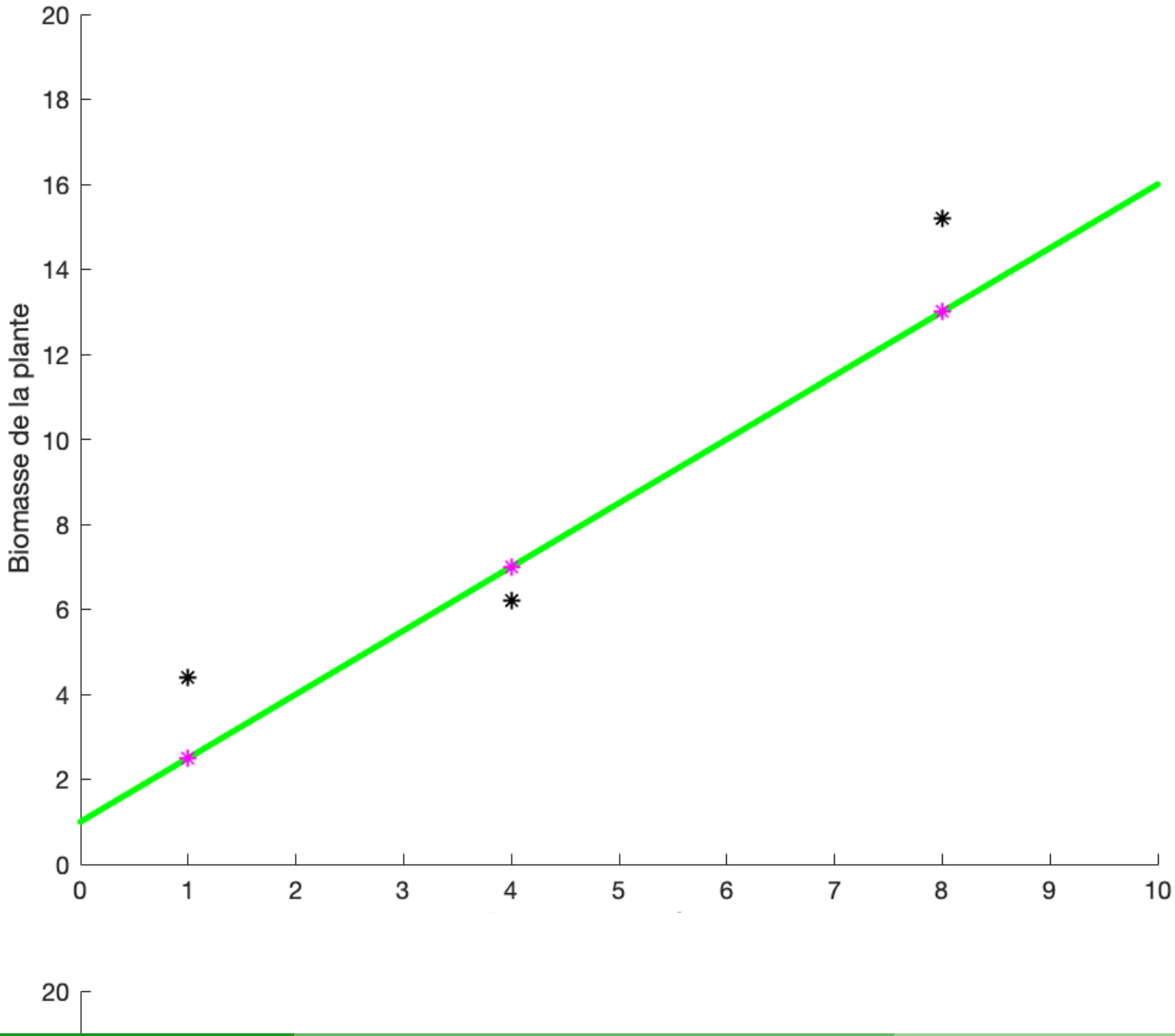


- from experimental data :
  value of the parameter that makes the model outputs most similar/likely to the observations
- from expert knowledge and literature
  examples : mechanistic parameters
- by combining the two sources of information

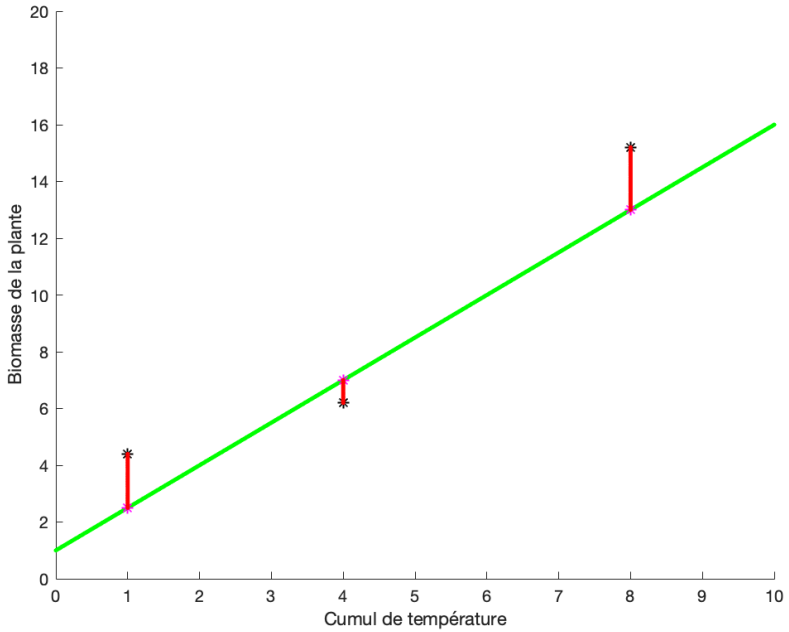
Toy example : model plant biomass as a function of thermal time



Choose a criterion



Least squares criterion



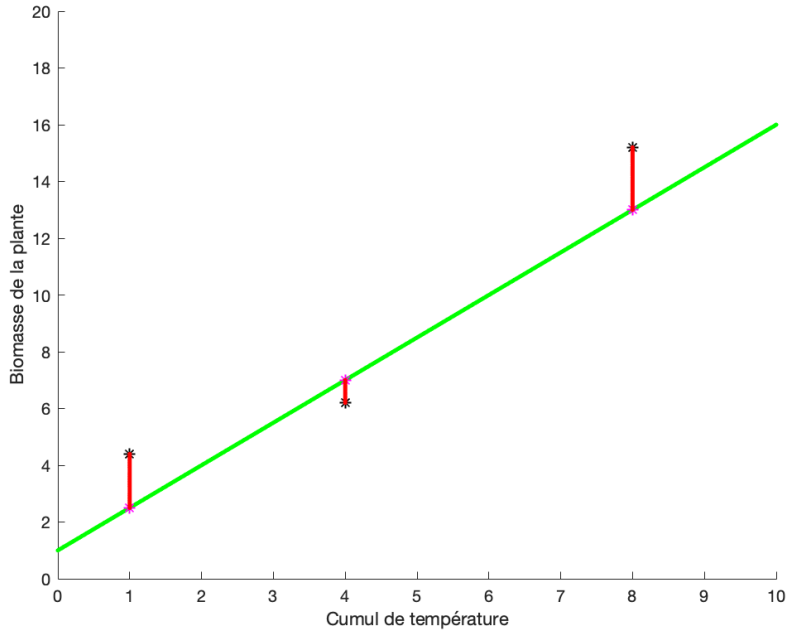
Three points : $M_1(t_1, y_1)$, $M_2(t_2, y_2)$, $M_3(t_3, y_3)$

Linear model $y = \alpha t + \beta$

Criterion : $\mathcal{C}(\alpha, \beta) = \sum_{k=1}^{3} (y_k - (\alpha t_k + \beta))^2$

Parameter values : $(\hat{\alpha}, \hat{\beta})$ such that $\mathcal{C}(\alpha, \beta)$ is minimum

Statistical point of view



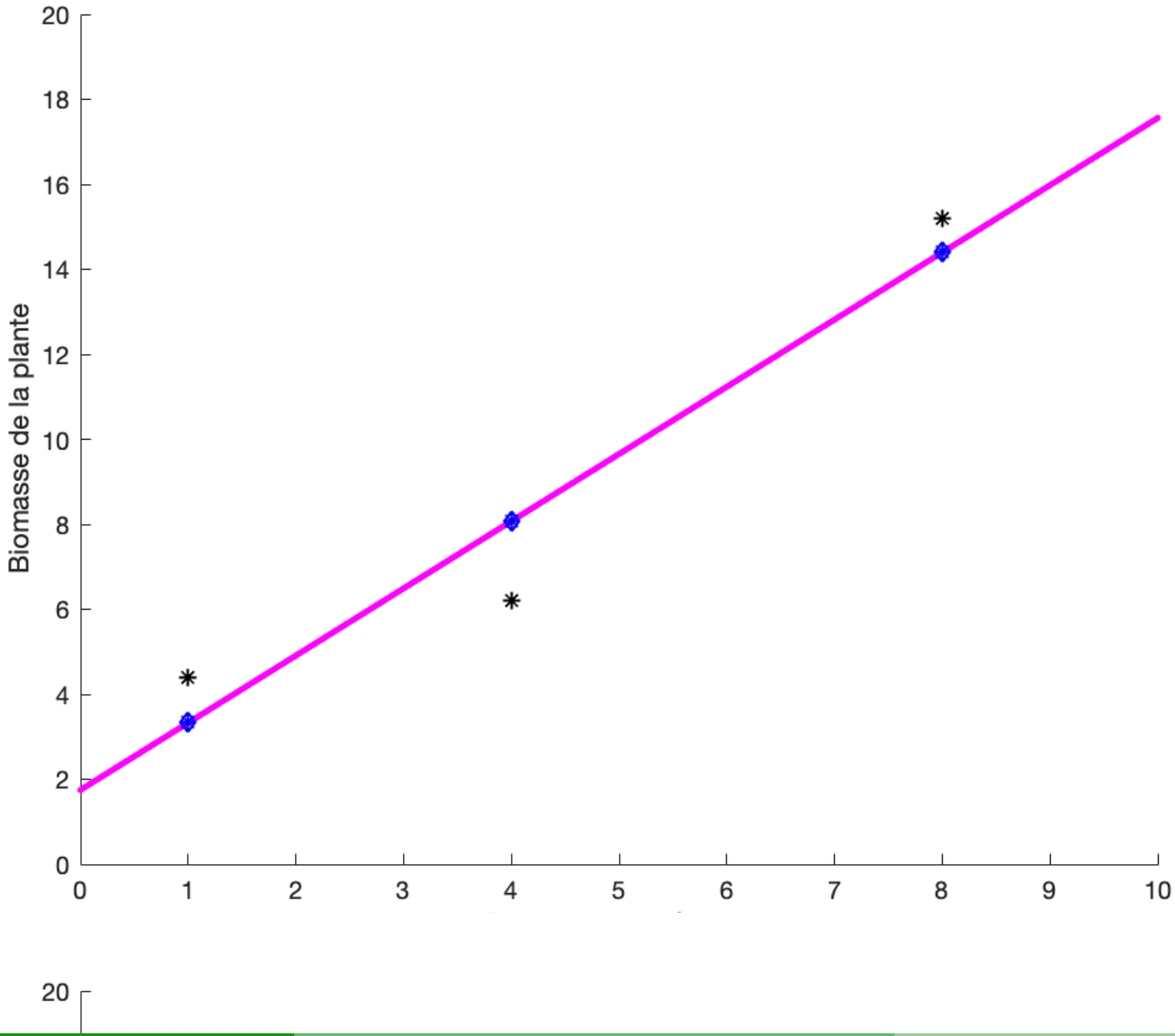
Three observations : $M_1(t_1, y_1)$, $M_2(t_2, y_2)$, $M_3(t_3, y_3)$

Probability distribution for y with expectation $\alpha t + \beta$ and variance 1

Log-likelihood : $\mathcal{L}(\alpha, \beta) = -\sum_{k=1}^{3} (y_k - (\alpha t_k + \beta))^2 / 2 + C$

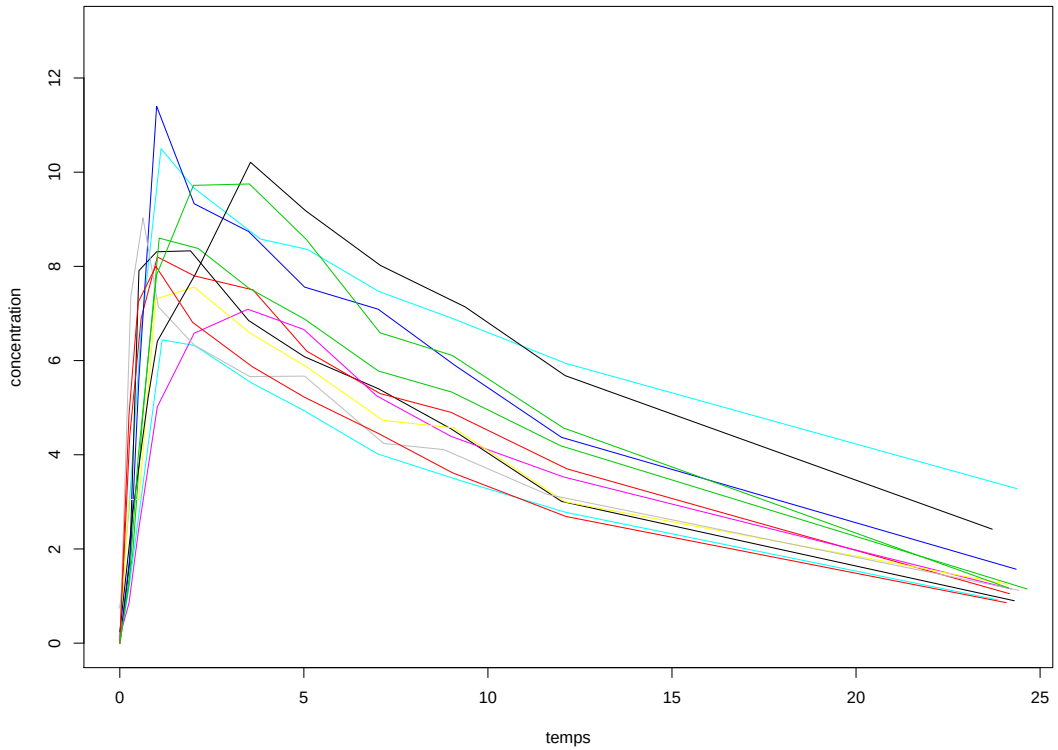
Parameter estimation : $(\hat{\alpha}, \hat{\beta})$ such that $\mathcal{L}(\alpha, \beta)$ is maximum

Forecasting



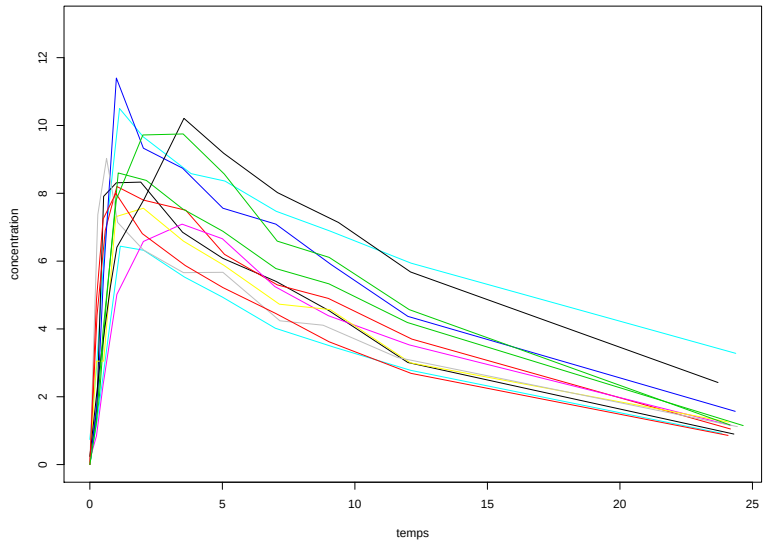
Theophylline Concentration over time

[Davidian and Giltinian (1995)]



12 subjects, same oral dose (mg/kg) times in hours theophylline concentration in mg/L

Theophylline Concentration over time



Models :

- $y_{ij} = G(t_j, x_i, \theta) + \varepsilon_{ij}$
- G solution d'une équation différentielle
- $y_{ij} = G(t_j, x_i, \theta_i) + \varepsilon_{ij}$
- $y_{ij} = \int H(t_j, x_i, \theta, z)dz + \varepsilon_{ij}$
- ....

Interdisciplinary project Stat4Plant (2021-2025)

stat4plant.mathnum.inrae.fr

$\Longrightarrow$ develop new models and new statistical methods to characterize the interactions between the plant and its environment

$\Longrightarrow$ bring together a consortium of researchers in modeling and applied statistics and researchers in biology specializing in phenotype-genotype relationships

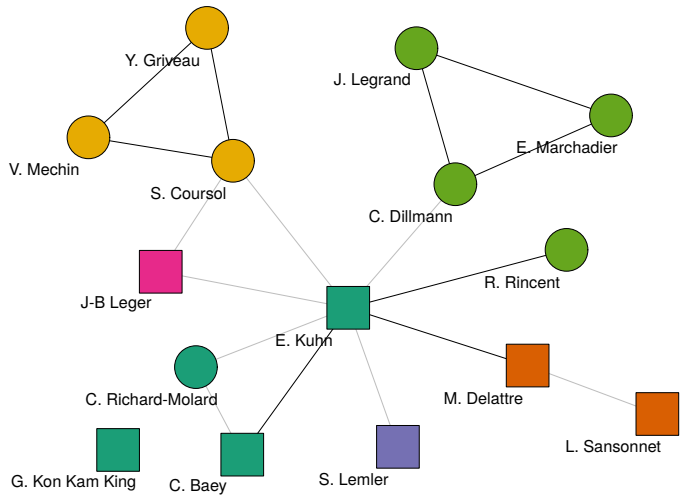


Figure – Collaboration network of the consortium. Each node is a member of the team, colored by its partnership. Squares represent statisticians and circles biologists.

Conclusion and perspectives

- increasingly in-depth knowledge
- modeling increasingly complex
- massive data in number and size
- statistics, optimization and algorithms in "high dimension"
- calculation resources, High-performance computing ...

$\Longrightarrow$ many interdisciplinary challenges to take up

$\Longrightarrow$ environmental impact of research to be controlled