2021

A brief introduction to machine learning and deep learning

David Makowski

INRAE

https://www6.inrae.fr/mia-paris/Equipes/Membres/David-Makowski

Outline

- Definition & main principles
- Several extensions of linear regression
- Trees and forests
- Deep learning

Outline

- Definition & main principles
- Several extensions of linear regression
- Trees and forests
- Deep learning

Artificial intelligence

Machine learning

Artificial intelligence

Machine learning

Supervised learning

Objective: « Learning a function that maps an input to an output based on examples of input-output pairs »

Statistical Modeling: The Two Cultures (Breiman, 2001)

$$y = f(x) + e$$

Modelling approach 1: Try to find the true $f(x)$

Modelling approach 2: Predict y from x as accurately as possible

Statistical Modeling: The Two Cultures (Breiman, 2001)

$$y = f(x) + e$$

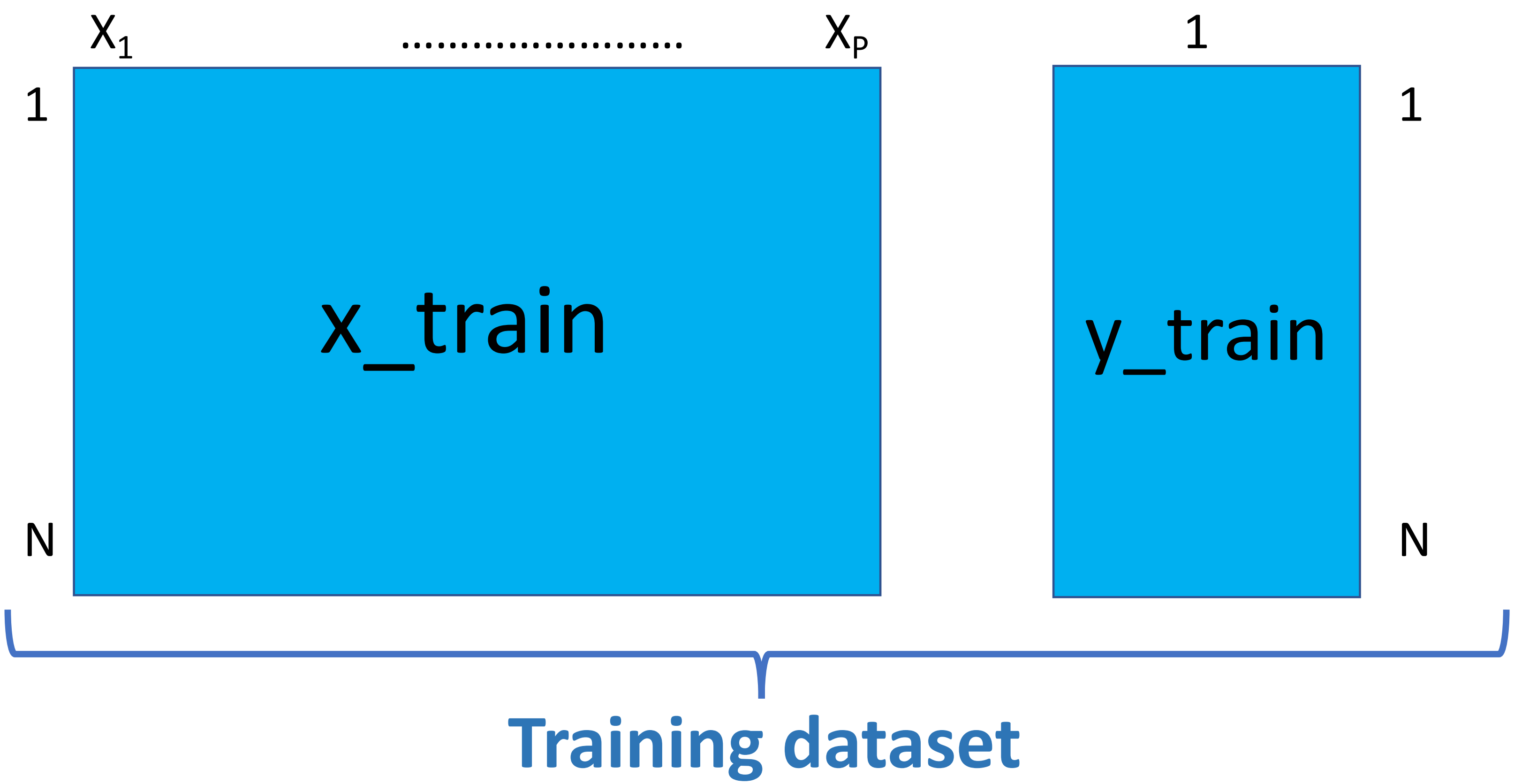
Modelling approach 1: Try to find the true $f(x)$

Modelling approach 2: Predict y from x as accurately as possible

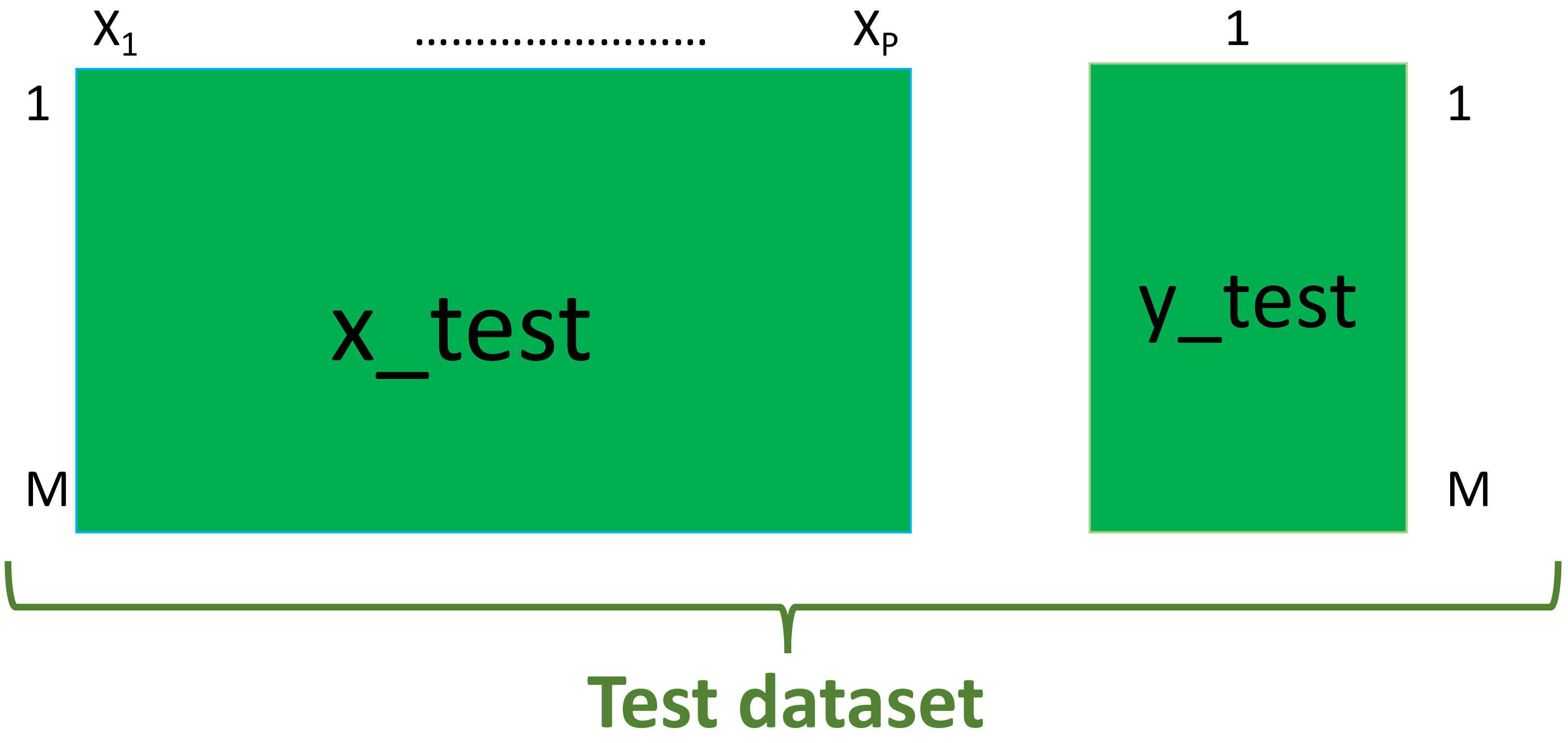
Two main steps

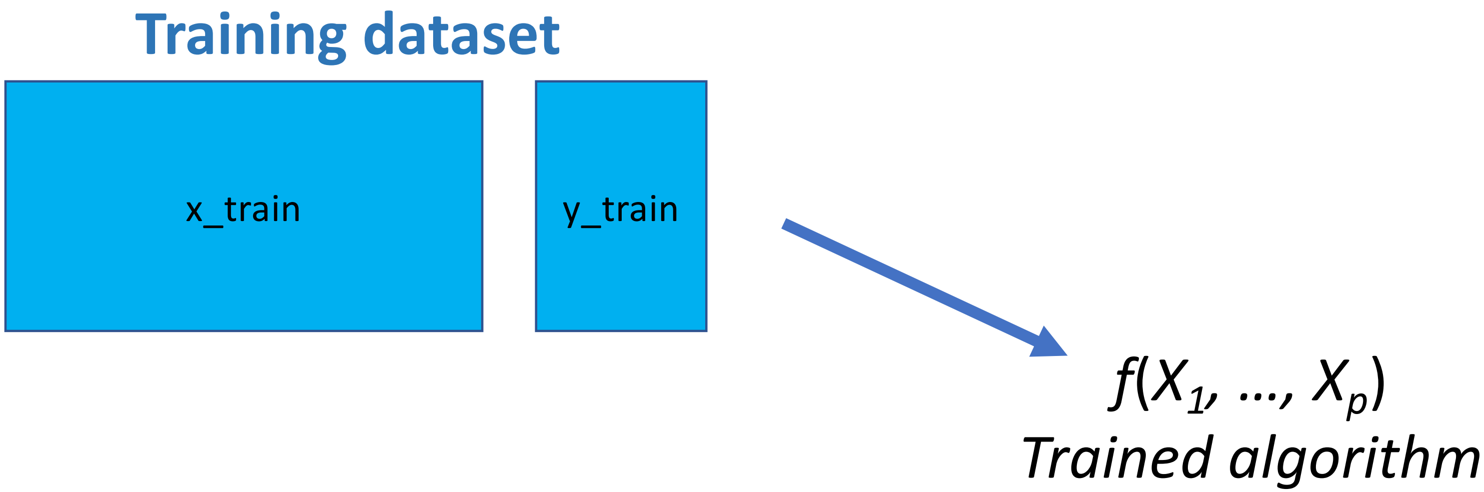
- Step 1: Training
- Step 2: Test

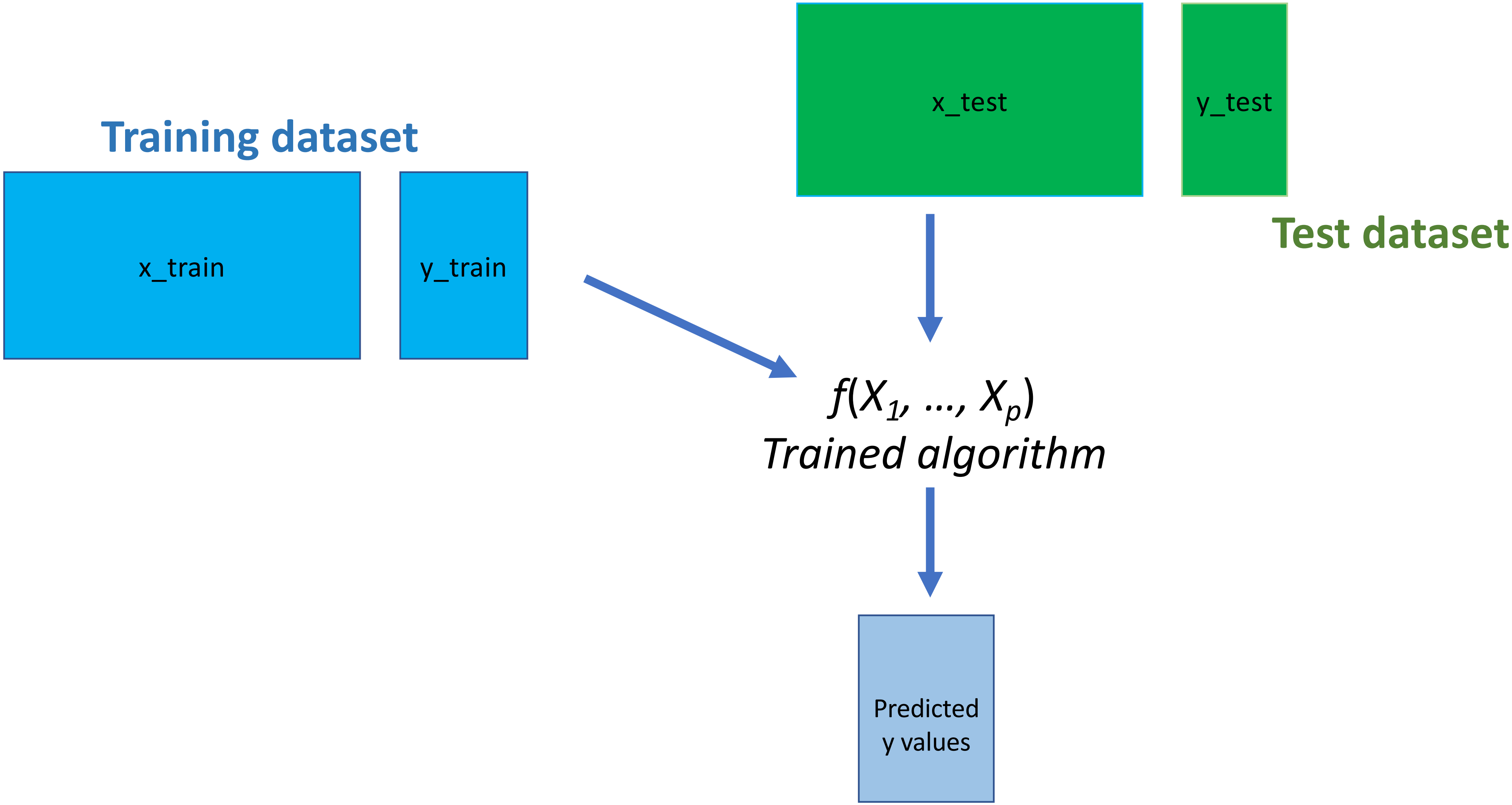
Step 1: Train an algorithm predicting Y as a function of X_1, …, X_P using a **training dataset**



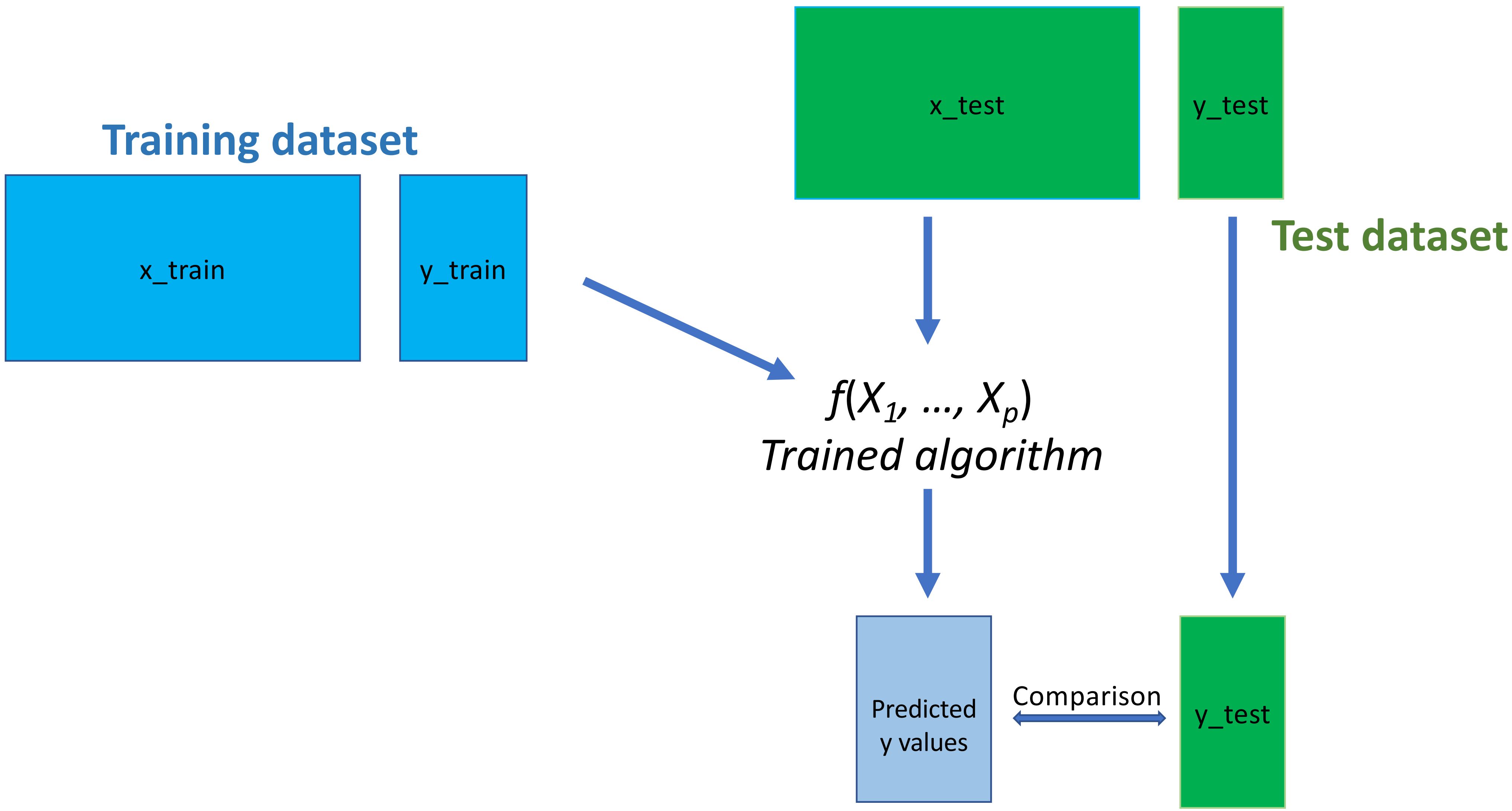
Step 2: Assess the predictive capability of the trained algorithm using a **test dataset**




Training dataset
x_train
y_train
$f(X_1, \ldots, X_p)$
Trained algorithm



Training dataset
x_train
y_train
x_test
y_test
Test dataset
f(X1, ..., Xp)
Trained algorithm
Predicted
y values



Training dataset
x_train
y_train
x_test
y_test
Test dataset
$f(X_1, \ldots, X_p)$
Trained algorithm
Predicted
y values
Comparison
y_test


kaggle
Search
Competitions
Datasets
Notebooks
Competitions

kaggle

Search

- Home
- Compete
- Data
- Notebooks
- Discuss
- Courses
- More

Genentech
A Member of the Roche Group

Flu Forecasting

Predict when, where and how strong the flu will be

$125,000 · 50 teams · 6 years ago

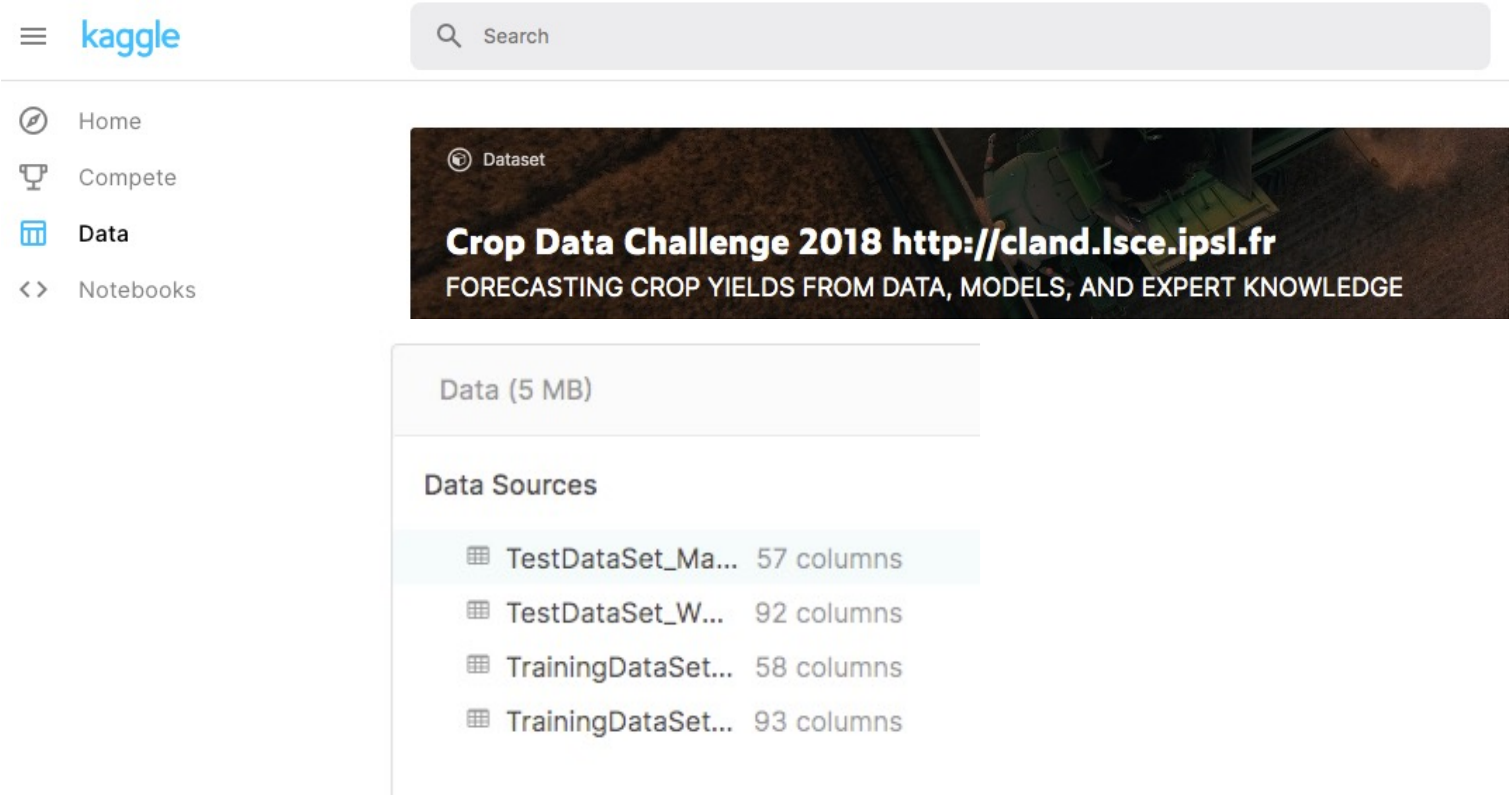
Overview Data Discussion Leaderboard Rules

« The objective of this competition is to build an algorithm that helps predict the occurrence, peak and severity of influenza in a given season ».

In the money Gold Silver Bronze

#	△pub	Team Name	Notebook	Team Members	Score
1	—	Alfonso Nieto-Castanon			0.47415
2	—	J.A. Guerrero (Datrik Intelligen...			0.47567
3	—	Zhanpeng Fang			0.47573
4	—	Tim Salimans			0.47650
5	—	Victor			0.47708
6	—	Nitai Dean			0.48110
7	—	BenPlus			0.48665

RMSE


kaggle
Search
Home
Compete
Data
Notebooks
Dataset
Crop Data Challenge 2018 http://cland.lsce.ipsl.fr
FORECASTING CROP YIELDS FROM DATA, MODELS, AND EXPERT KNOWLEDGE
Data (5 MB)
Data Sources
TestDataSet_Ma... 57 columns
TestDataSet_W... 92 columns
TrainingDataSet... 58 columns
TrainingDataSet... 93 columns

French maize yield prediction (départements)

Training dataset

55 inputs
3394 yield data

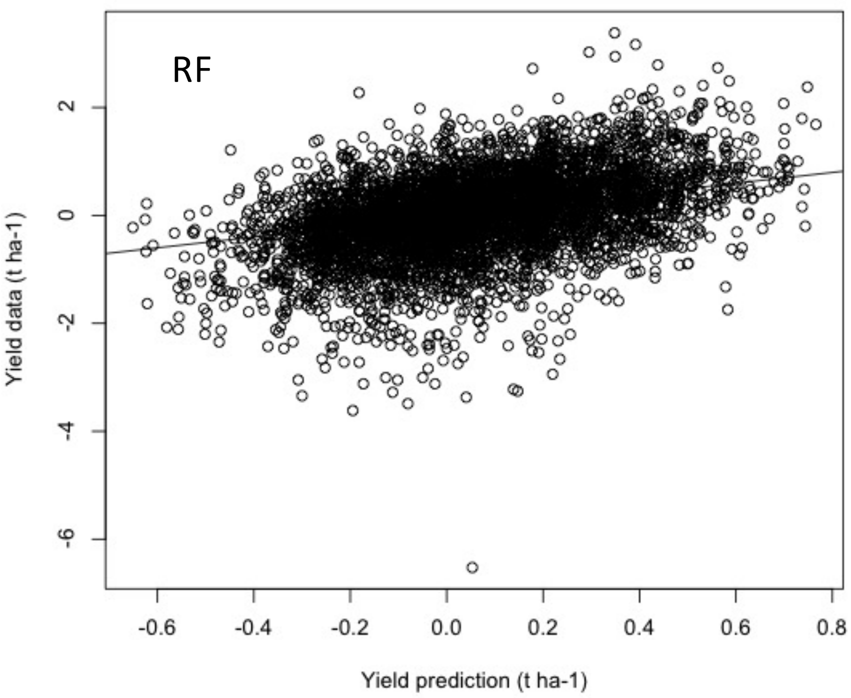
Algorithms developed by the participants

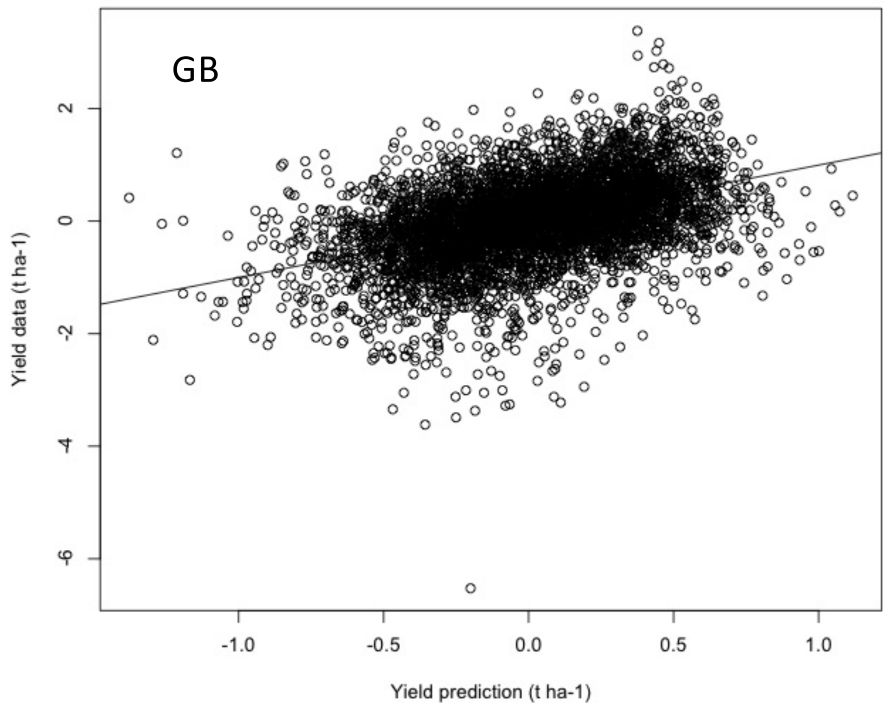
Test dataset

55 inputs
1708 yield data

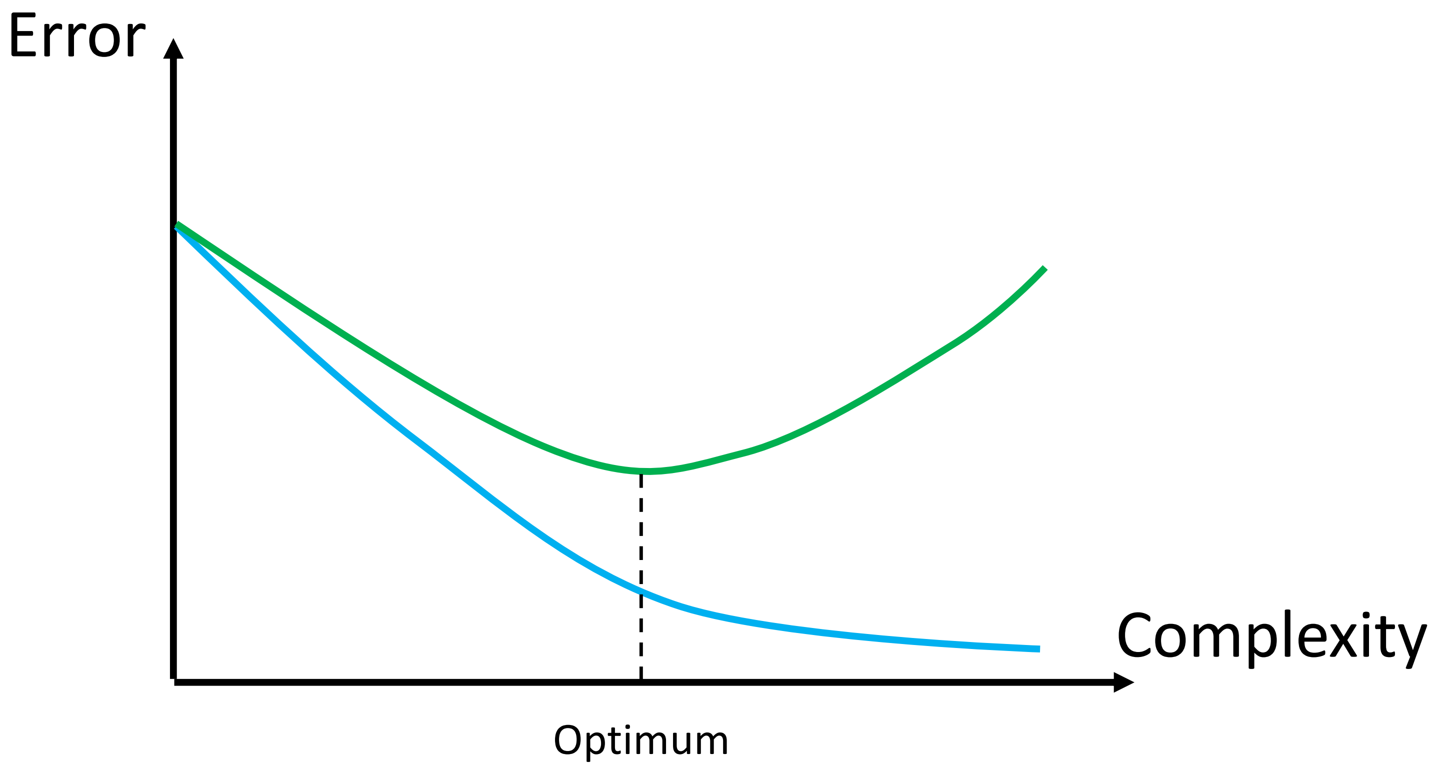
Evaluation of the accuracy of the algorithms by the organizer

Method	RMSEP (maize yield)
Random Forest (RF)	0.71 t/ha
Gradient boosting (GB)	0.70 t/ha

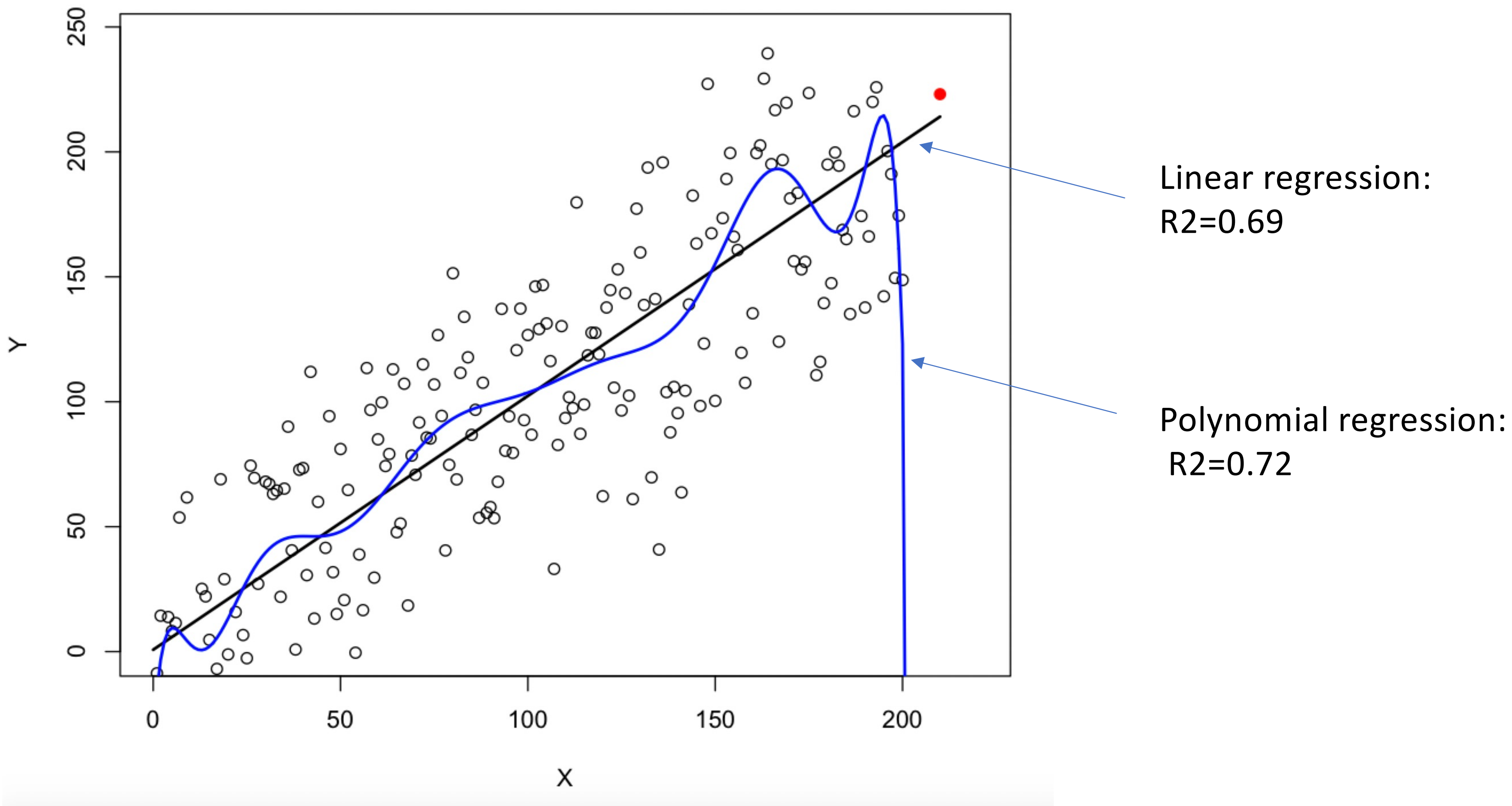




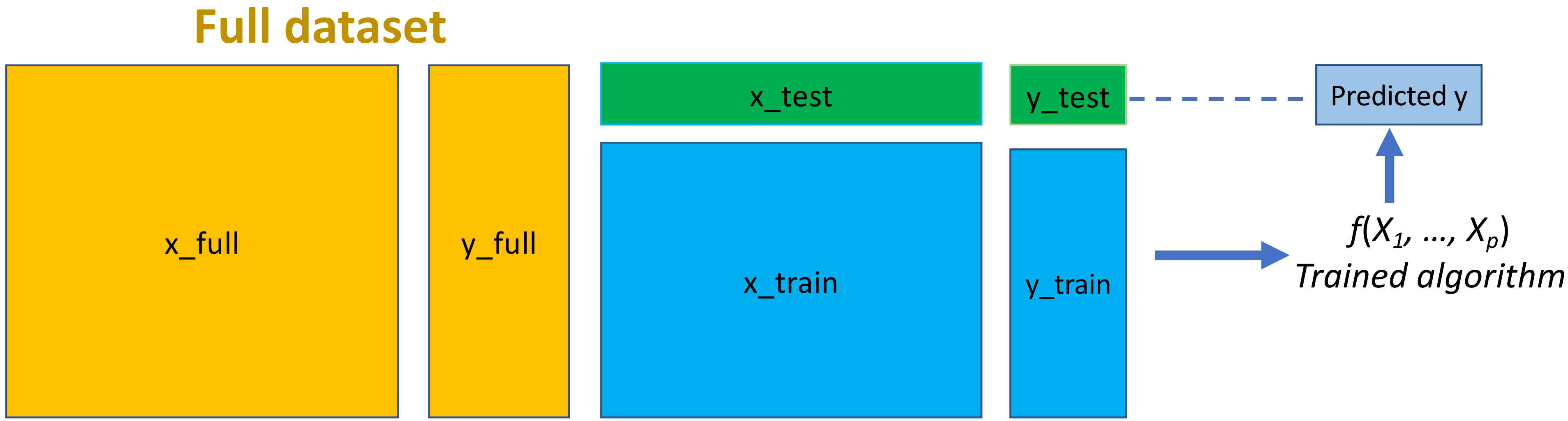
Model testing should be taken seriously to avoid risk of overfitting

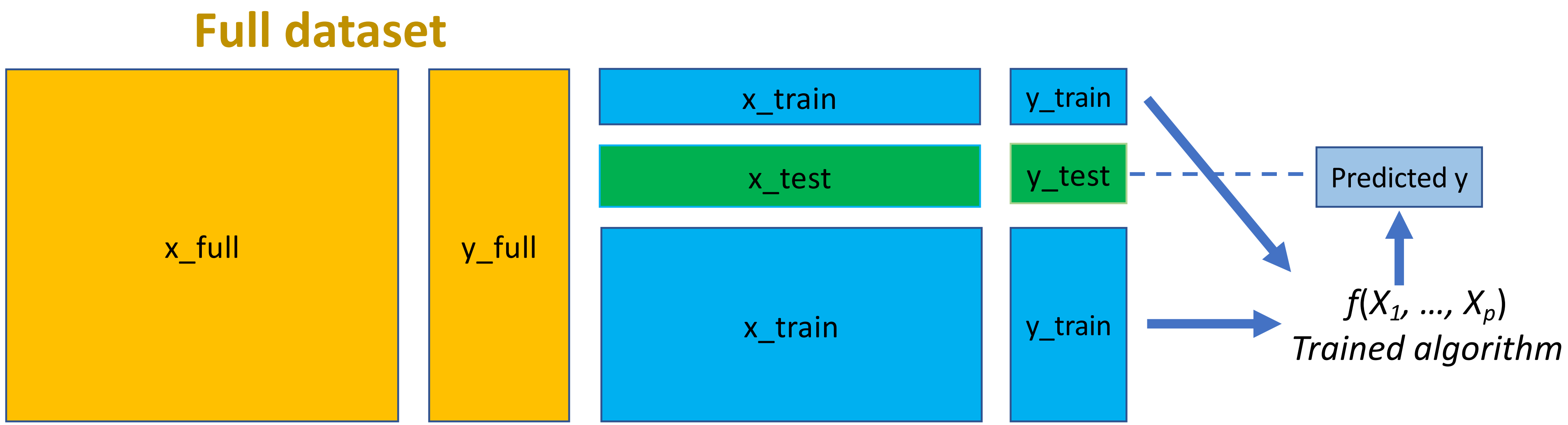


Model testing should be taken seriously to avoid risk of overfitting

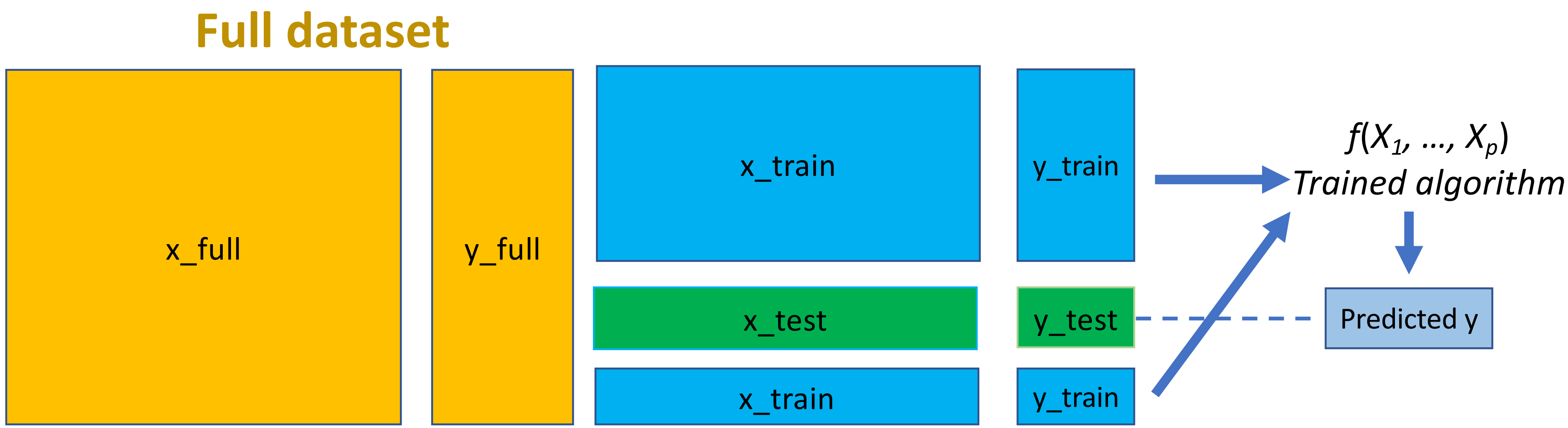


Cross-validation is used when no independent test dataset is available

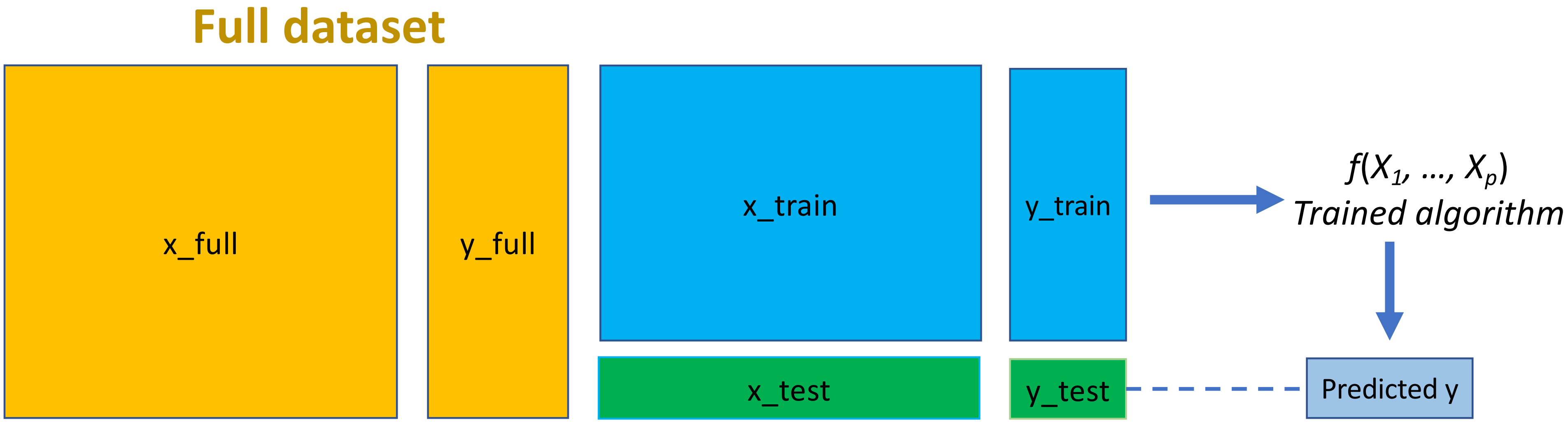

Full dataset
x_full
y_full
x_test
y_test
Predicted y
x_train
y_train
f(X1, …, Xp)
Trained algorithm



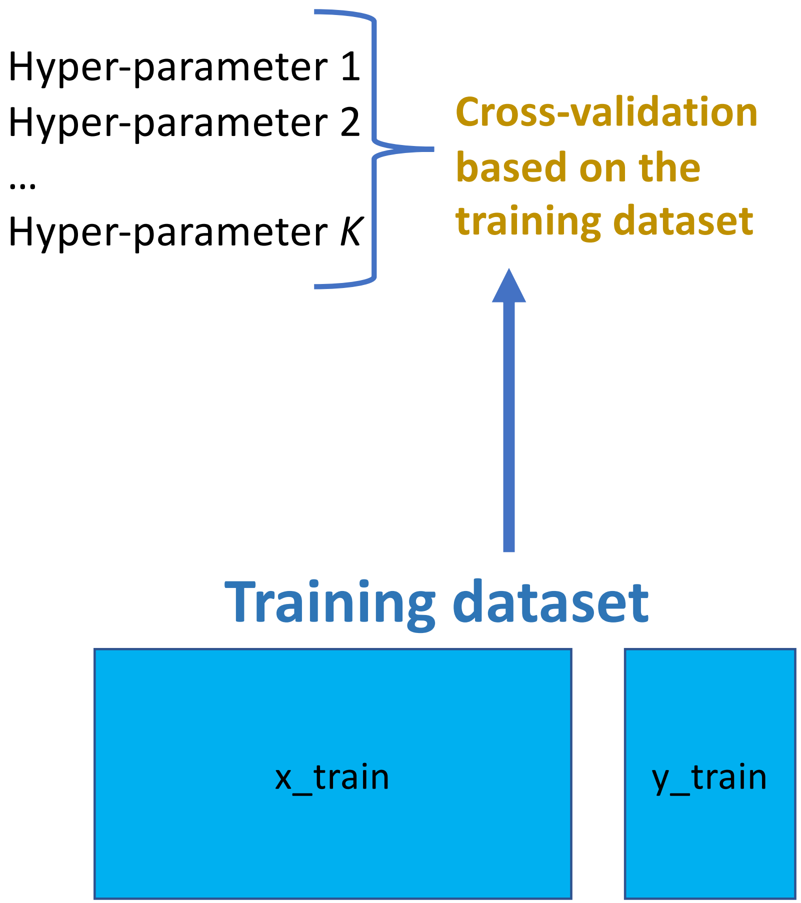
Full dataset
x_full
y_full
x_train
y_train
x_test
y_test
Predicted y
x_train
y_train
$f(X_1, \ldots, X_p)$
Trained algorithm

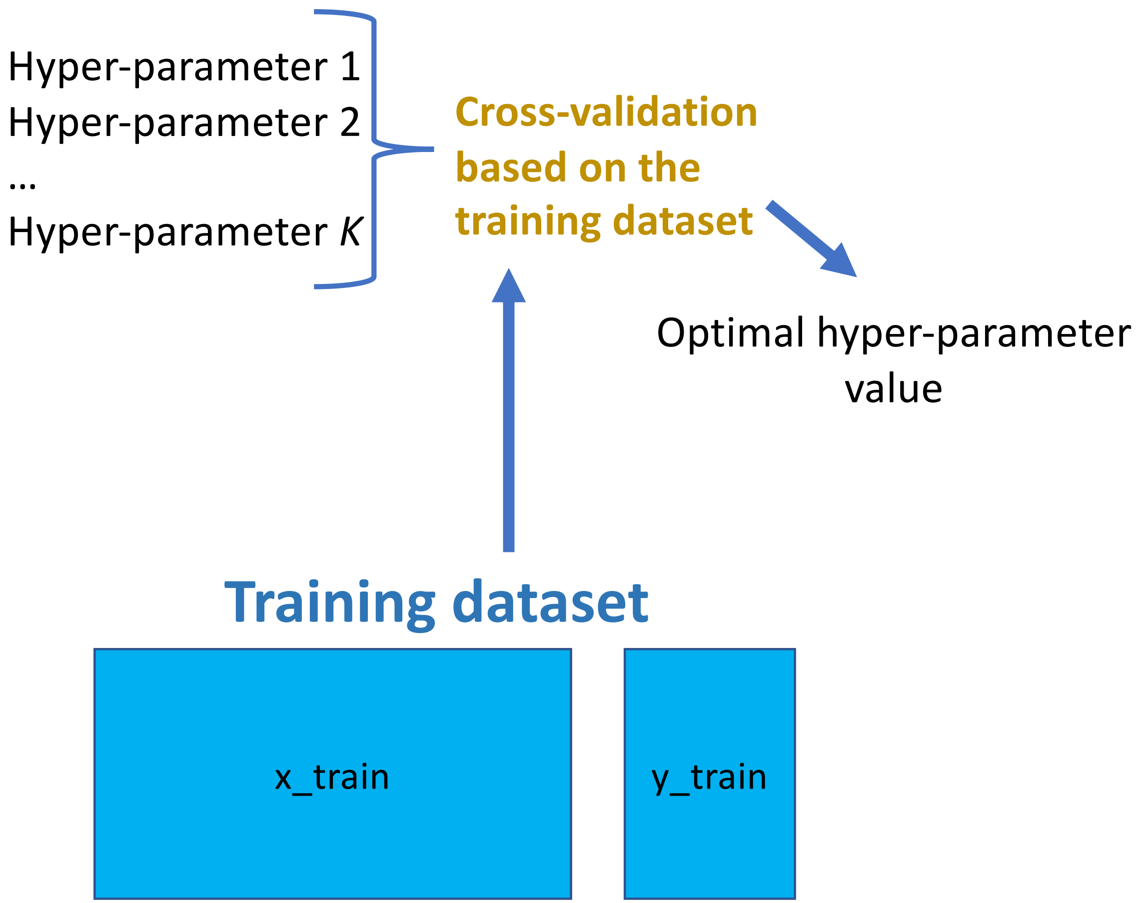


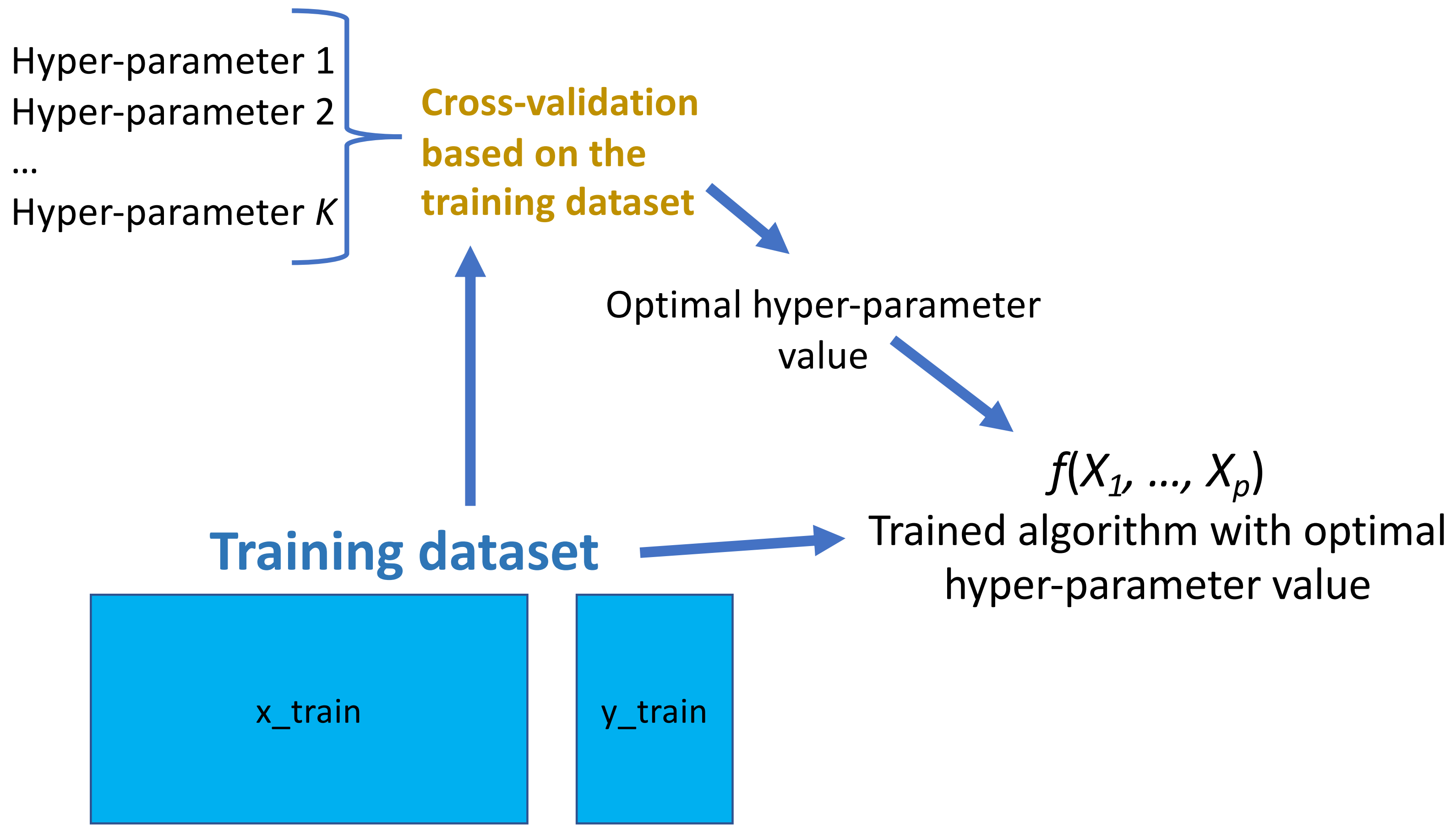
Full dataset
x_full
y_full
x_train
y_train
x_test
y_test
x_train
y_train
$f(X_1, \ldots, X_p)$
Trained algorithm
Predicted y

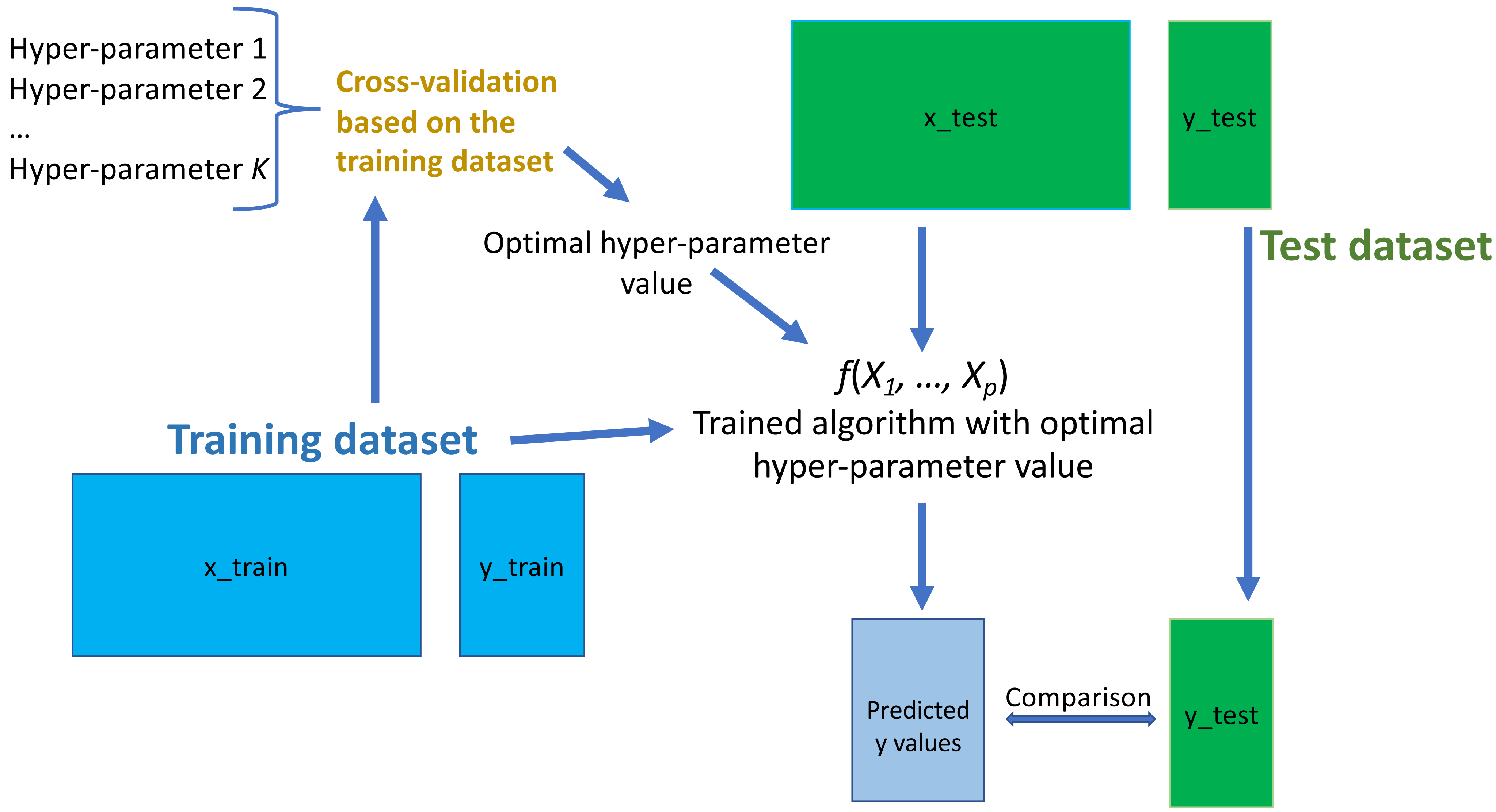

Full dataset
x_full
y_full
x_train
y_train
x_test
y_test
$f(X_1, \ldots, X_p)$
Trained algorithm
Predicted y

Cross-validation and test dataset can be combined together


Hyper-parameter 1
Hyper-parameter 2
…
Hyper-parameter *K*
Cross-validation based on the training dataset
Training dataset
x_train
y_train


Hyper-parameter 1
Hyper-parameter 2
...
Hyper-parameter K
Cross-validation based on the training dataset
Optimal hyper-parameter value
Training dataset
x_train
y_train


Hyper-parameter 1
Hyper-parameter 2
...
Hyper-parameter K
Cross-validation based on the training dataset
Optimal hyper-parameter value
$f(X_1, \ldots, X_p)$
Trained algorithm with optimal hyper-parameter value
Training dataset
x_train
y_train


Hyper-parameter 1
Hyper-parameter 2
…
Hyper-parameter K
Cross-validation based on the training dataset
Optimal hyper-parameter value
x_test
y_test
Test dataset
$f(X_1, \ldots, X_p)$
Trained algorithm with optimal hyper-parameter value
Training dataset
x_train
y_train
Predicted y values
Comparison
y_test

Why machine learning is powerful?

Very flexible methods

+

Computational power

+

Large datasets

→ Increased chance to obtain accurate predictions

Why machine learning is powerful?

Prediction error = g(Bias, Variance)

Why machine learning is powerful?

Prediction error = g(Bias, Variance)

ML is able to fing a good balance between bias and variance

Several « ML tricks »	Principle	Effect
Regularization	Add information to prevent overfitting and simplify the model	Reduce variance at the cost of a small increase of bias
Bagging	Bootstrap aggregation: average together multiple models fitted to resampled dataset	Reduce variance
Boosting	Fit a sequence of weak models to weighted versions of the data (more weight given to poorly predicted data at earlier rounds).	Reduce bias

Numerous methods available

- Regressions (standard, PLS, LASSO, Elastic net…)
- SVM
- Tree and random forest
- Gradient boosting
- Neural network
- Deep neural network
- Deep learning
- Bayesian classification

Numerous methods available

- Regressions (standard, PLS, LASSO, Elastic net…)
- SVM
- Tree and random forest
- Gradient boosting
- Neural network
- Deep neural network
- Deep learning
- Bayesian classification

Relatively easy to run these methods with specialized packages (with R or Python)

Are machine learning models « black boxes »?

This is less true than before.

Vizualisation tools:

- Importance ranking
- Partial dependence plots (PDP)
- Accumulated Local Effects (ALE) Plot

Example 1: Prediction of root biomass

https://doi.org/10.5194/essd-2021-25
Preprint. Discussion started: 29 March 2021

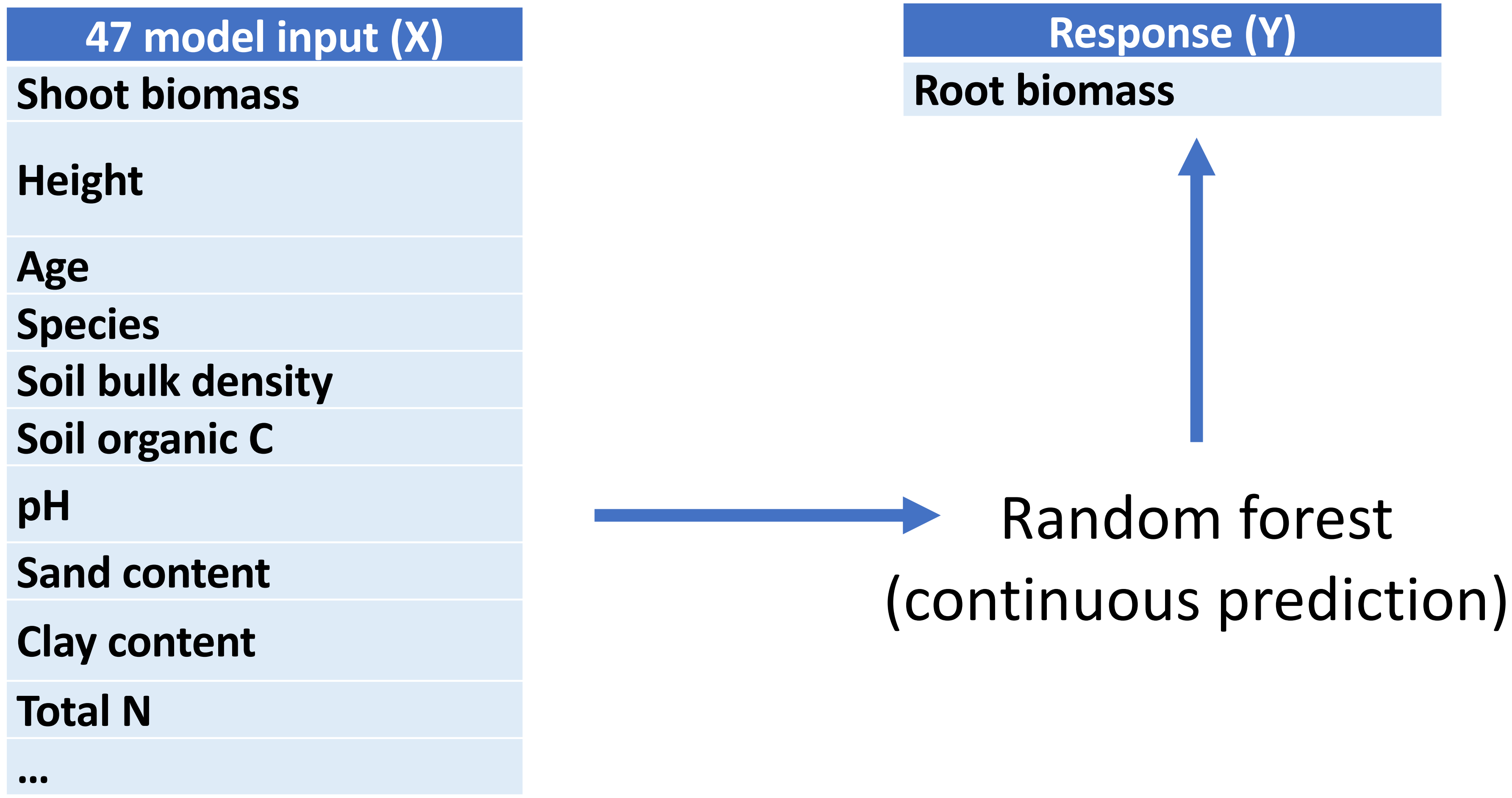
Open Access Earth System Science Data Discussions

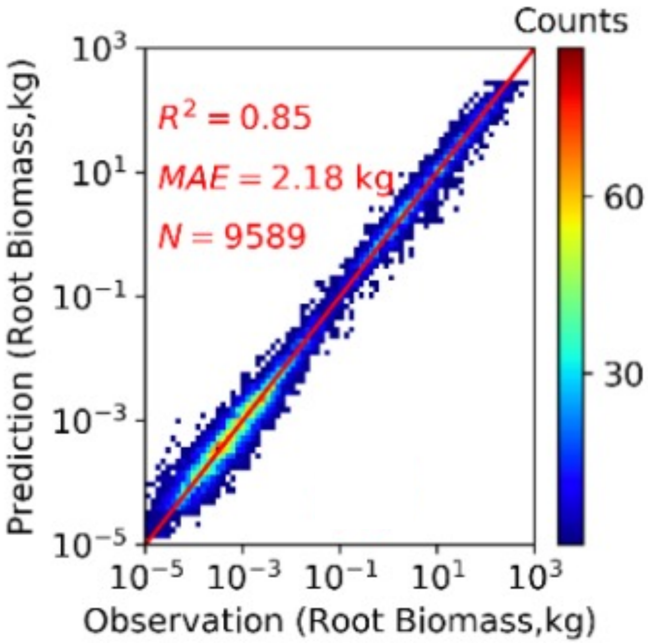
A global map of root biomass across the world's forests

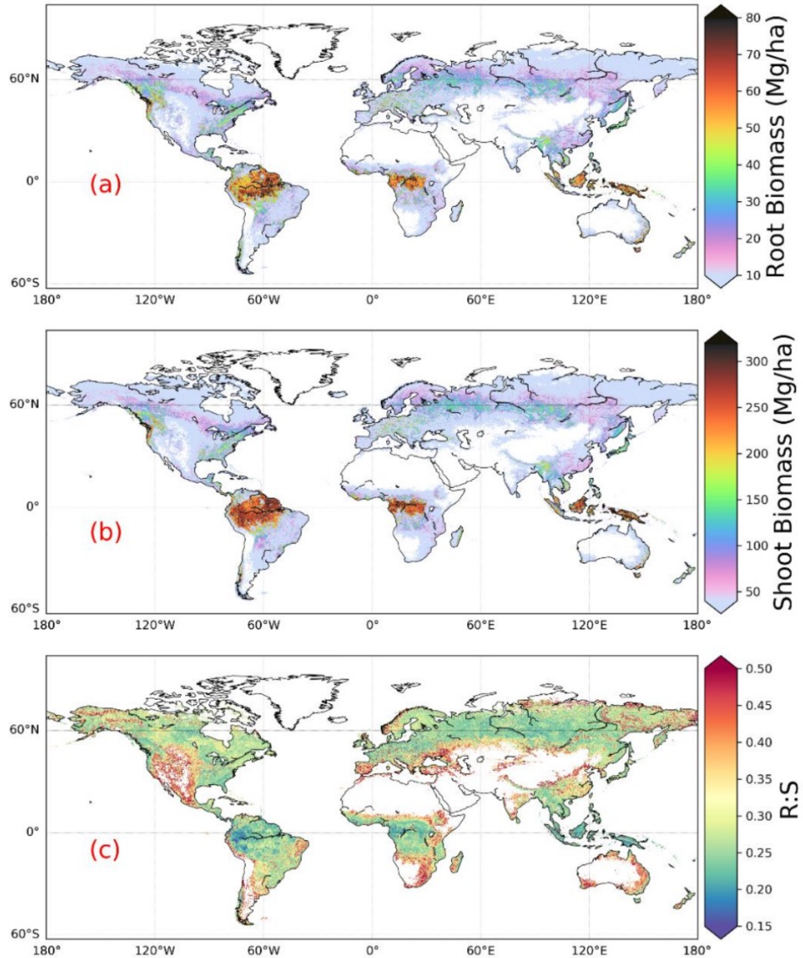
Yuanyuan Huang[1,2], Phillipe Ciais[1], Maurizio Santoro[3], David Makowski[4,5], Jerome Chave[6], Dmitry Schepaschenko[7,8,9], Rose Z. Abramoff[1], Daniel S. Goll[1], Hui Yang[1], Ye Chen[10], Wei Wei[11], Shilong Piao[12,13,14]

Dataset:

10,307 in-situ measurements of the biomass of roots and shoots for individual woody plants, covering 465 species across 10 biomes.


47 model input (X)
Shoot biomass
Height
Age
Species
Soil bulk density
Soil organic C
pH
Sand content
Clay content
Total N
...
Response (Y)
Root biomass
Random forest
(continuous prediction)





Global maps of **forest root biomass** generated through a machine learning model (a), shoot biomass from GlobBiomass-AGB(Santoro, 2018b) (b) and Root:Shoot ratio (c).

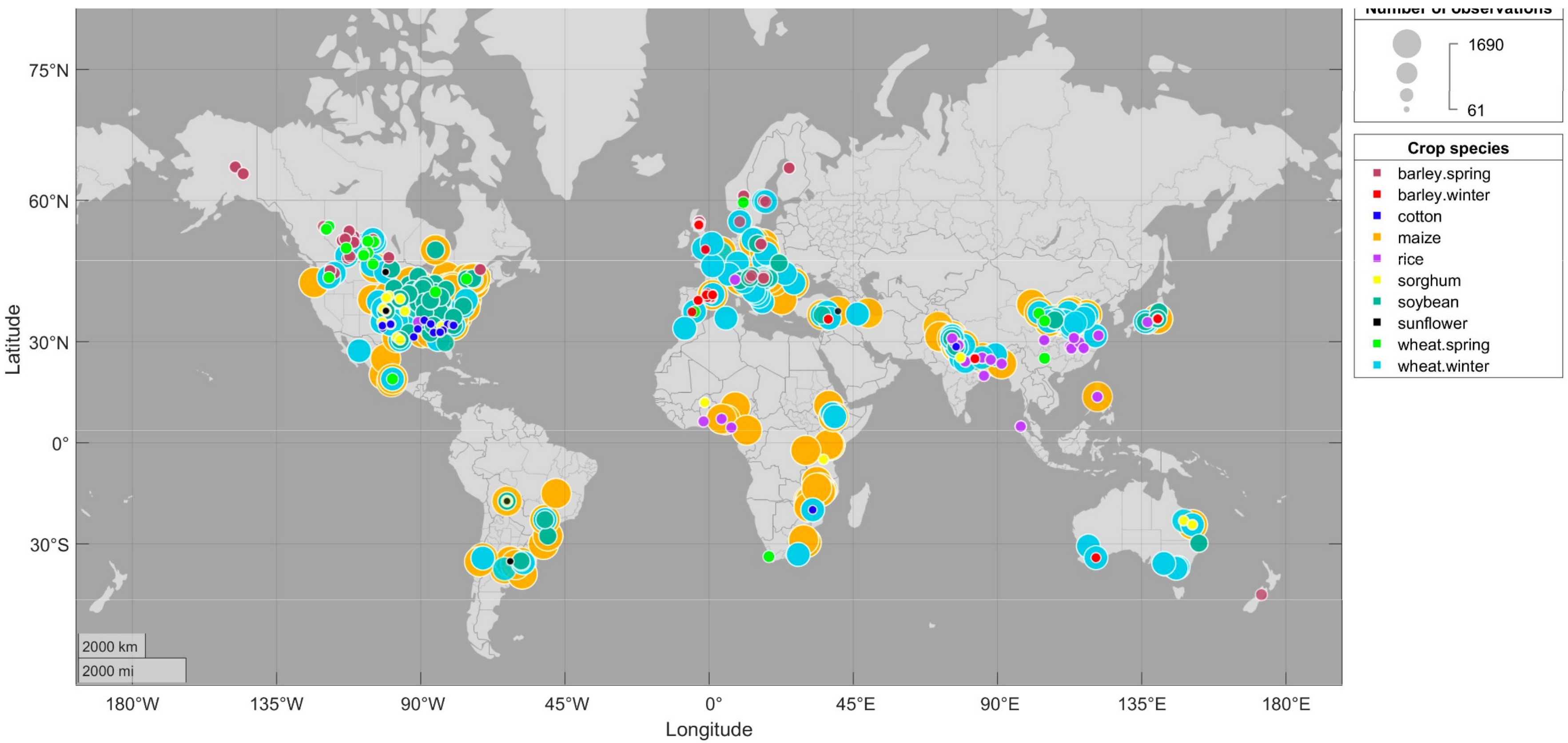
Example 2: Map the probability of yield increase of converting "conventional tillage system" to "conservation agriculture" at the global scale

https://www.nature.com/articles/s41598-021-82375-1.pdf

Locations of the experiments included in the dataset

Each experiment includes yield data for

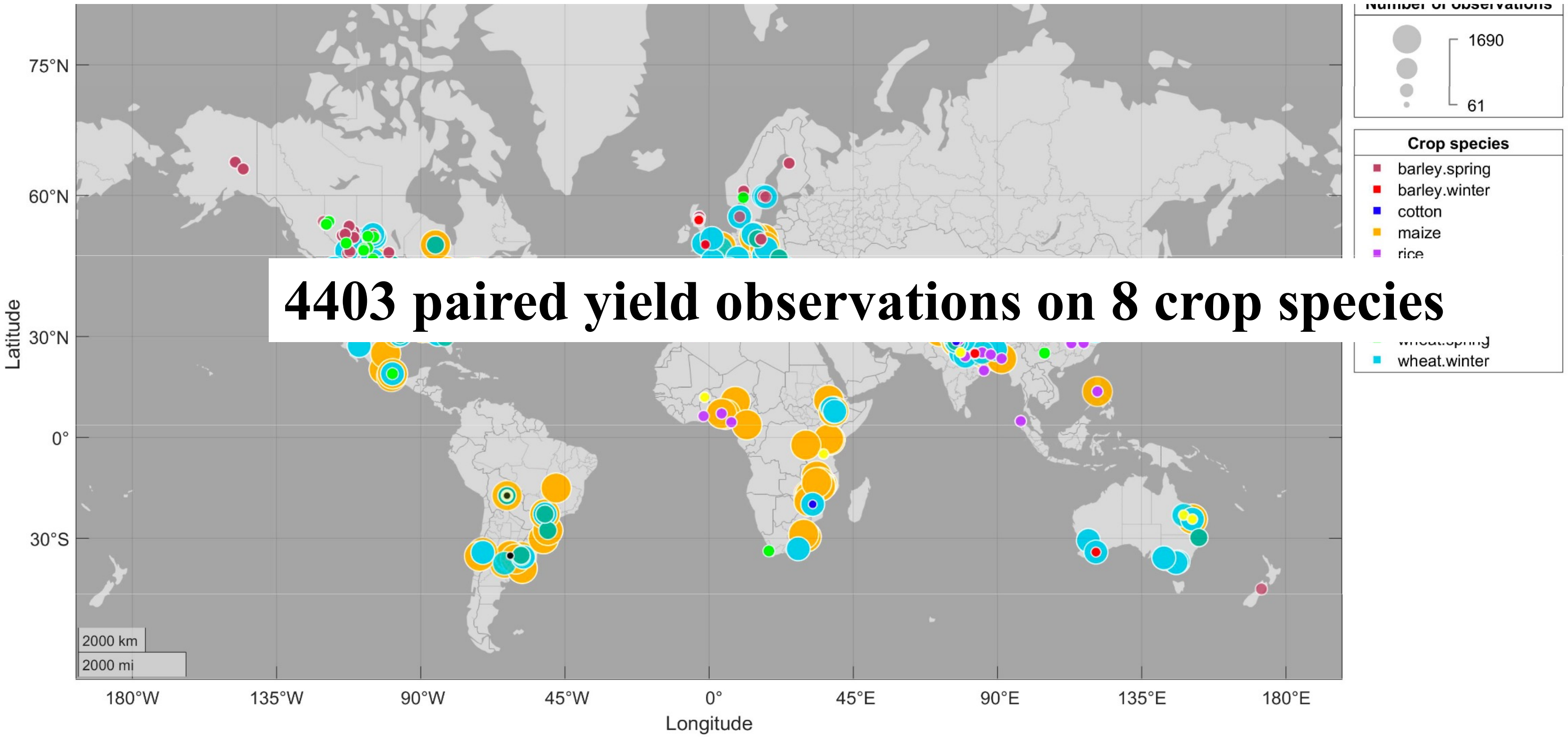
- a conservation agriculture system
- a conventional tillage system



Locations of the experiments included in the dataset

Each experiment includes yield data for

- a conservation agriculture system
- a conventional tillage system



4403 paired yield observations on 8 crop species

Model input (X)
Avg. Precipitation
Avg. Evapotranspiration
Average temperature
Avg. Maximum temperature
Avg. Minimum temperature
Soil texture
Crop type (Barley, maize, soybean, wheat, rice, sorghum, cotton, sunflower)
Fertilizer utilization (Y/N)
Herbicide and pesticide application (Y/N)
Crop rotation (Y/N)
Crop residue management (Y/N)

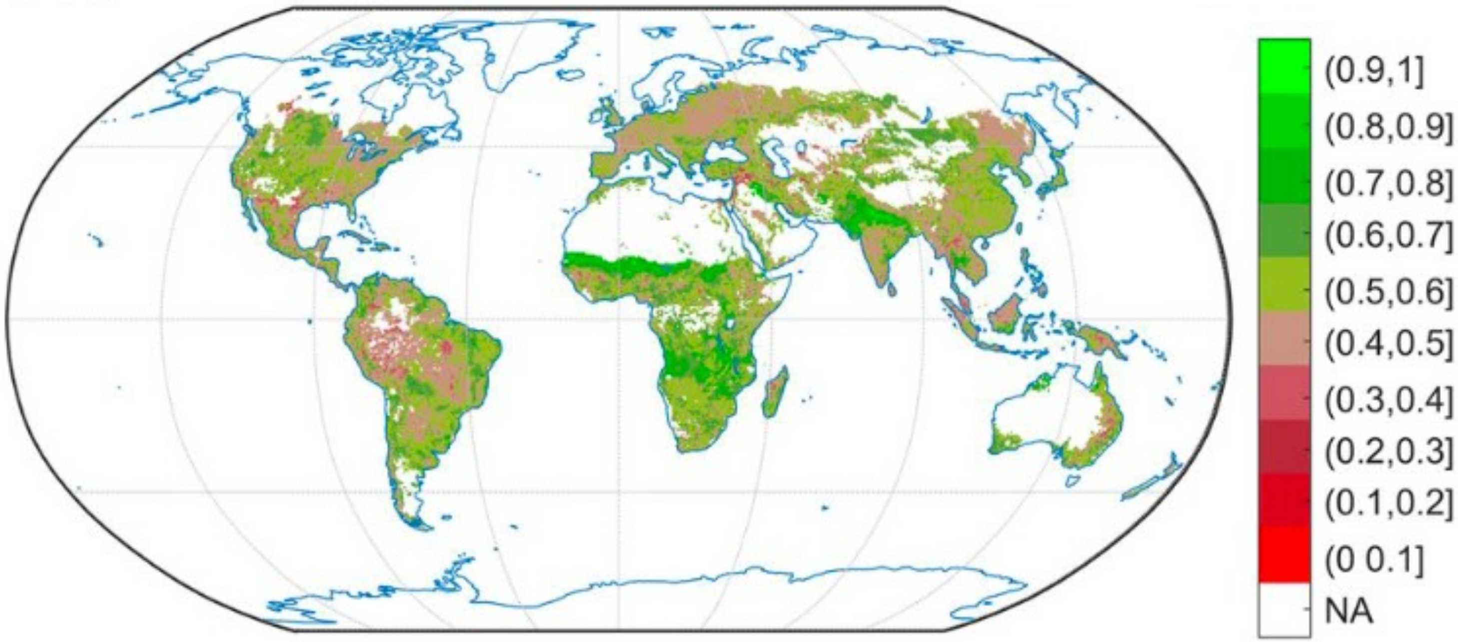
Response (Y)
Yield gain vs. yield loss (1/0) induced by no tillage

Model input (X)
Avg. Precipitation
Avg. Evapotranspiration
Average temperature
Avg. Maximum temperature
Avg. Minimum temperature
Soil texture
Crop type (Barley, maize, soybean, wheat, rice, sorghum, cotton, sunflower)
Fertilizer utilization (Y/N)
Herbicide and pesticide application (Y/N)
Crop rotation (Y/N)
Crop residue management (Y/N)

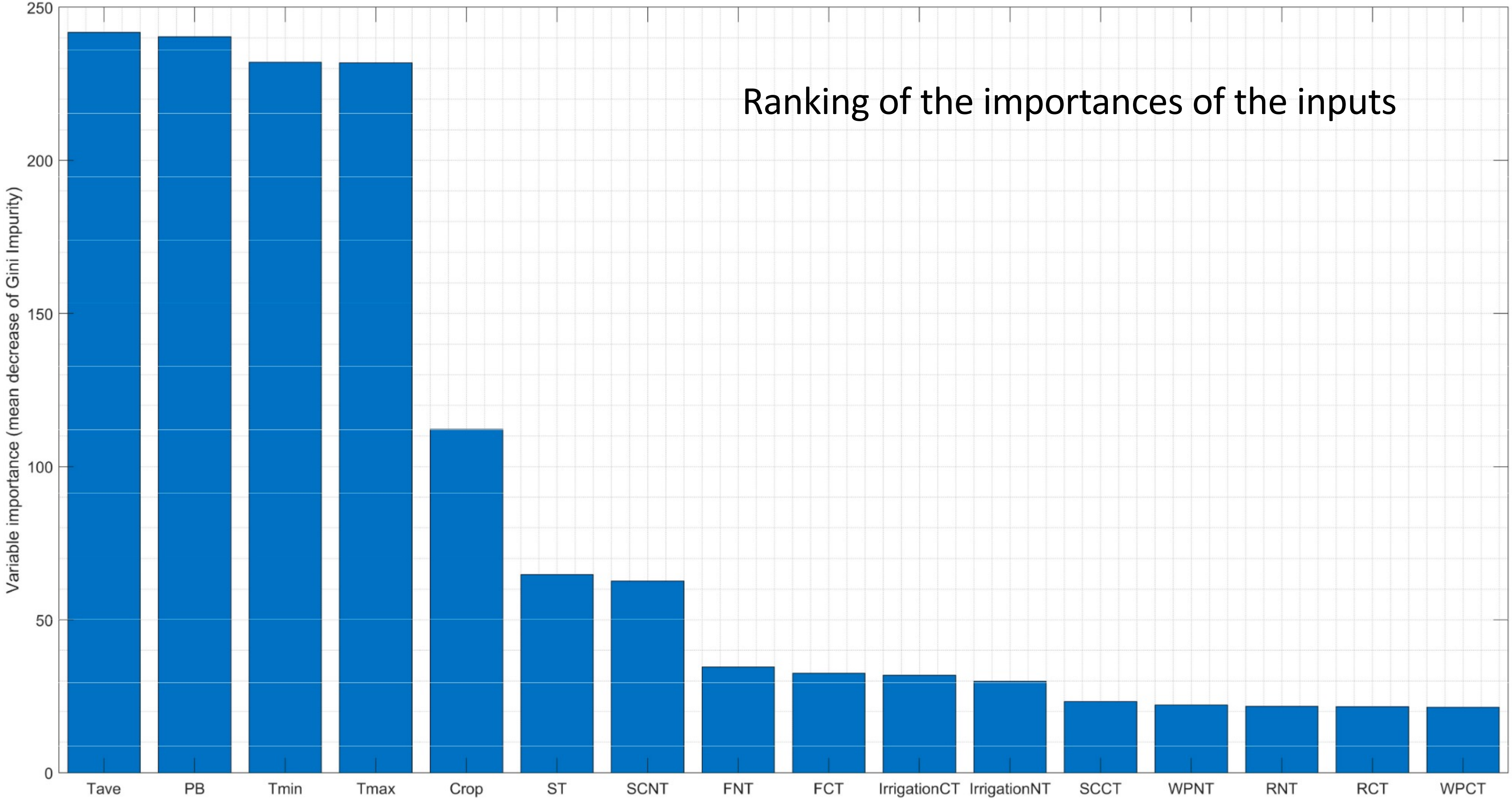
Response (Y)
Yield gain vs. yield loss (1/0) induced by no tillage

Random forest
(binary classification)

probability of yield increase for maize with Conservation agriculture vs. Tillage

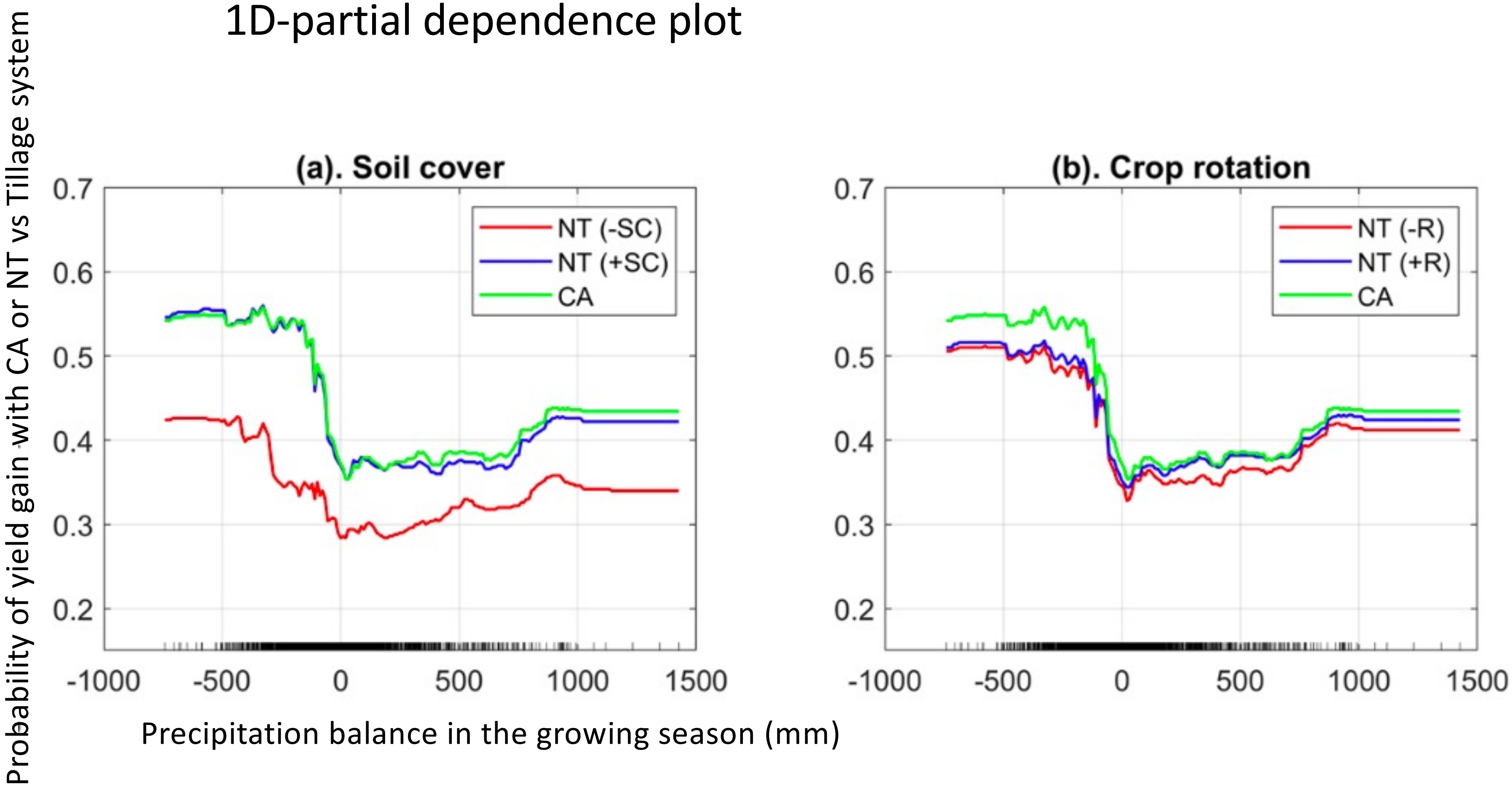


https://www.nature.com/articles/s41598-021-82375-1.pdf



https://www.nature.com/articles/s41598-021-82375-1.pdf

1D-partial dependence plot



https://www.nature.com/articles/s41598-021-82375-1.pdf

NT: No tillage system
R: Rotation
SC: Soil cover
CA: Conservation agriculture (NT+R+SC)

Main challenges in machine learning projects

- Choose a relevant question (Which Y? Which X?)
- Find reliable data
- Calibrate the hyper-parameters
- Assess prediction accuracy without bias
- Optimize computation time
- Vizualisation of output responses

Start simple

Start with two simple methods:

- Penalized linear regression (ex: LASSO)
- Random forest

Some trends

- Visualization tools (to open « the black boxes »)
- Image and text analyses (text mining, deep learning)
- Packages to streamline the development of predictive models (keras, caret, H2O…)
- Including expert knowledge in machine learning