2022

A brief introduction to machine learning

David Makowski

Université Paris-Saclay

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)

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

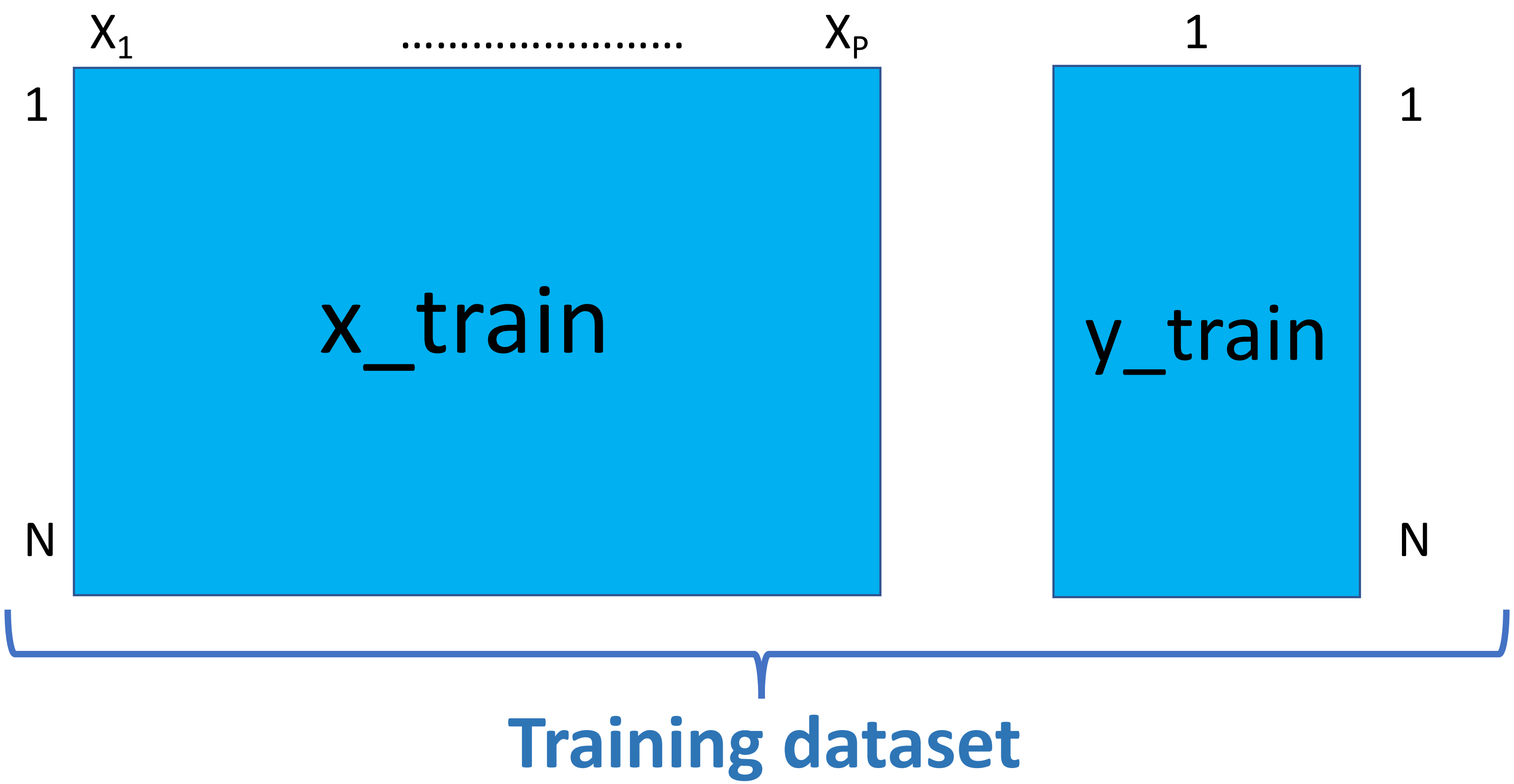
Modelling approach 1: Try to find the true $f(x)$

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

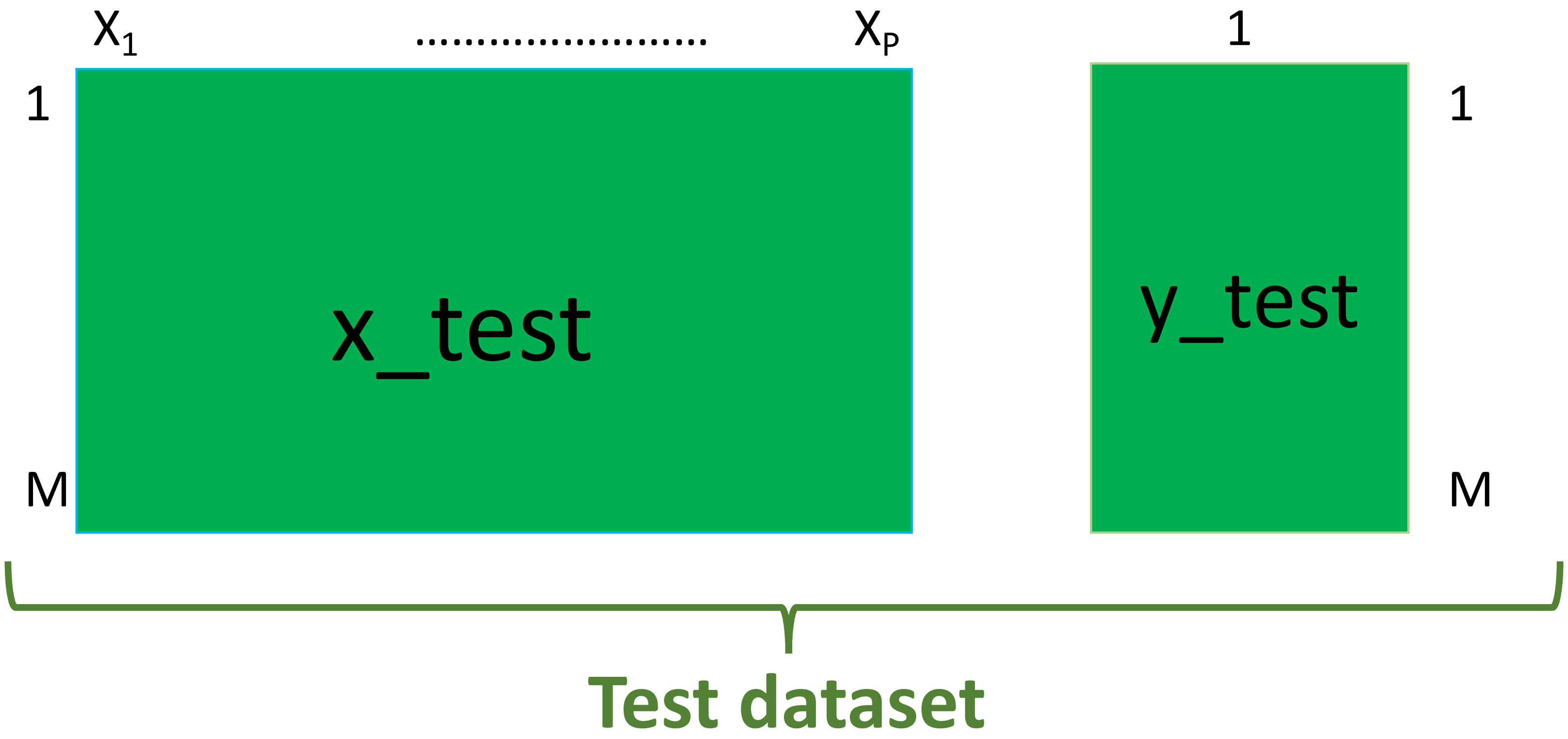
Two important steps

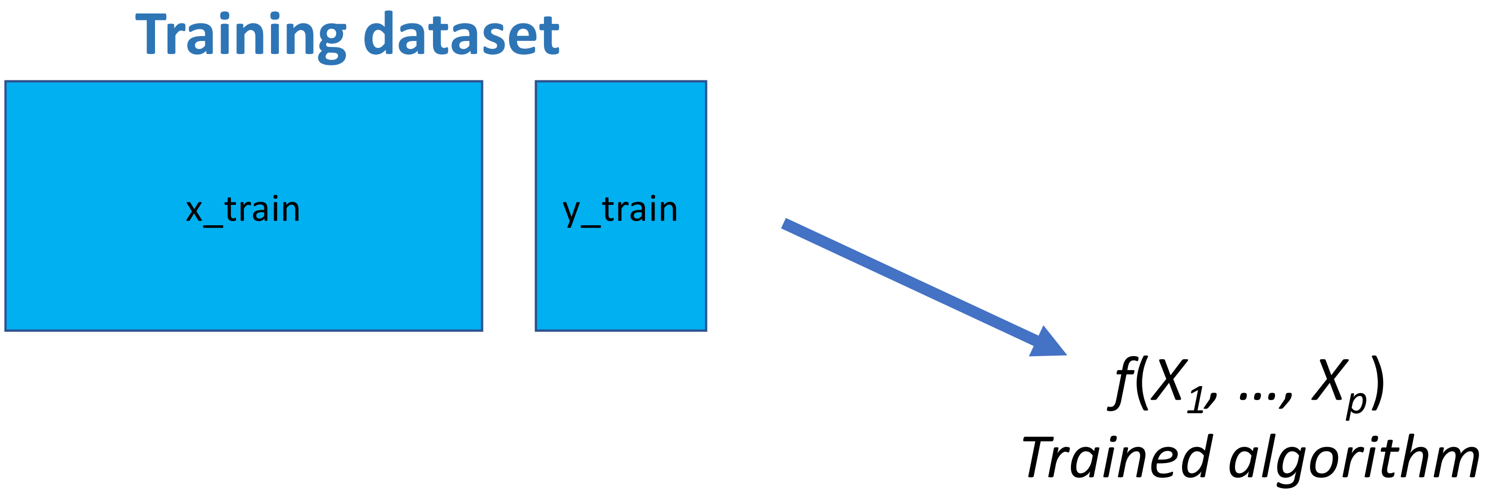
- Training
- Test

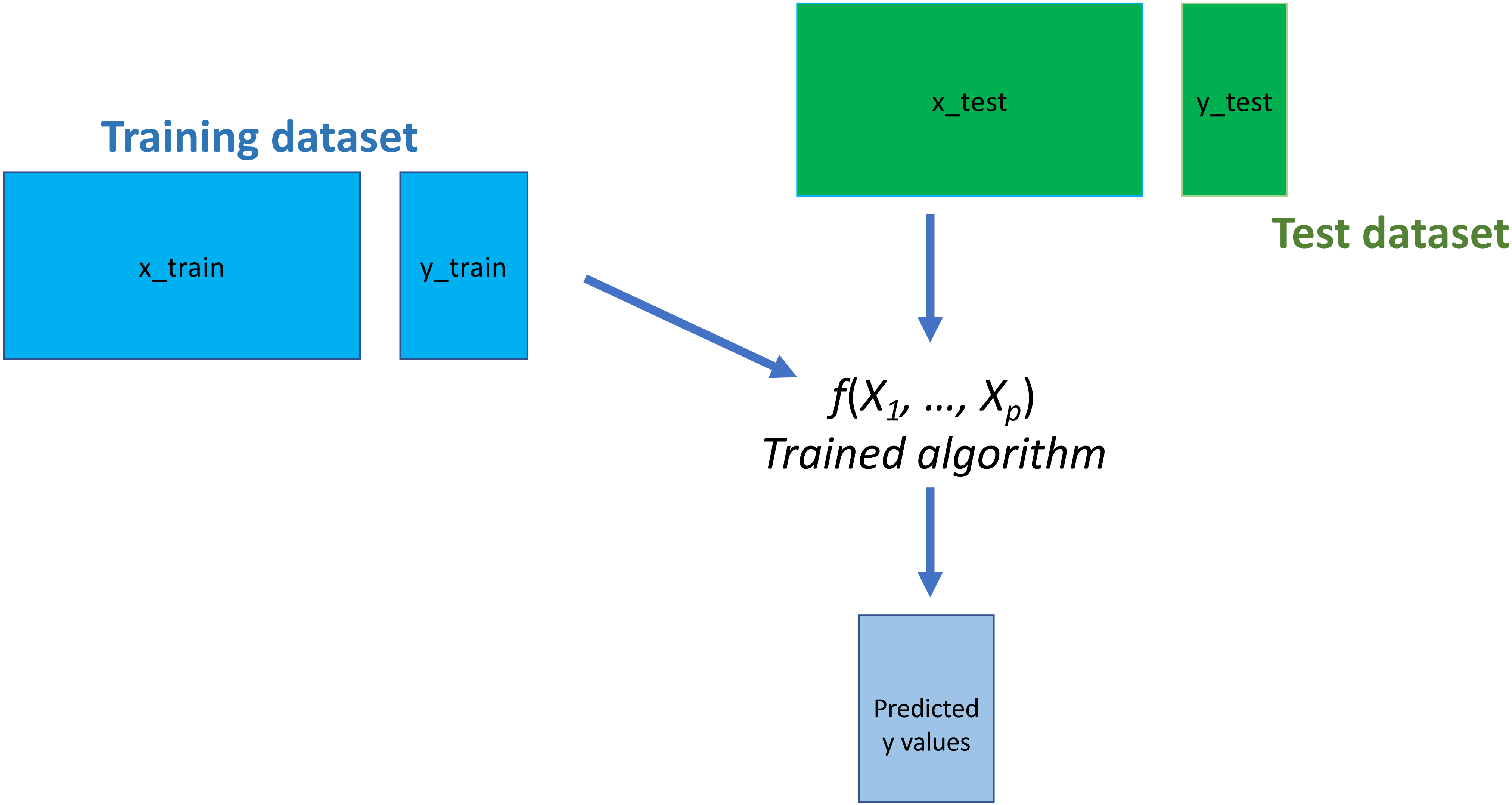
Training: Train an algorithm predicting Y as a function of X_1, …, X_P using a **training dataset**

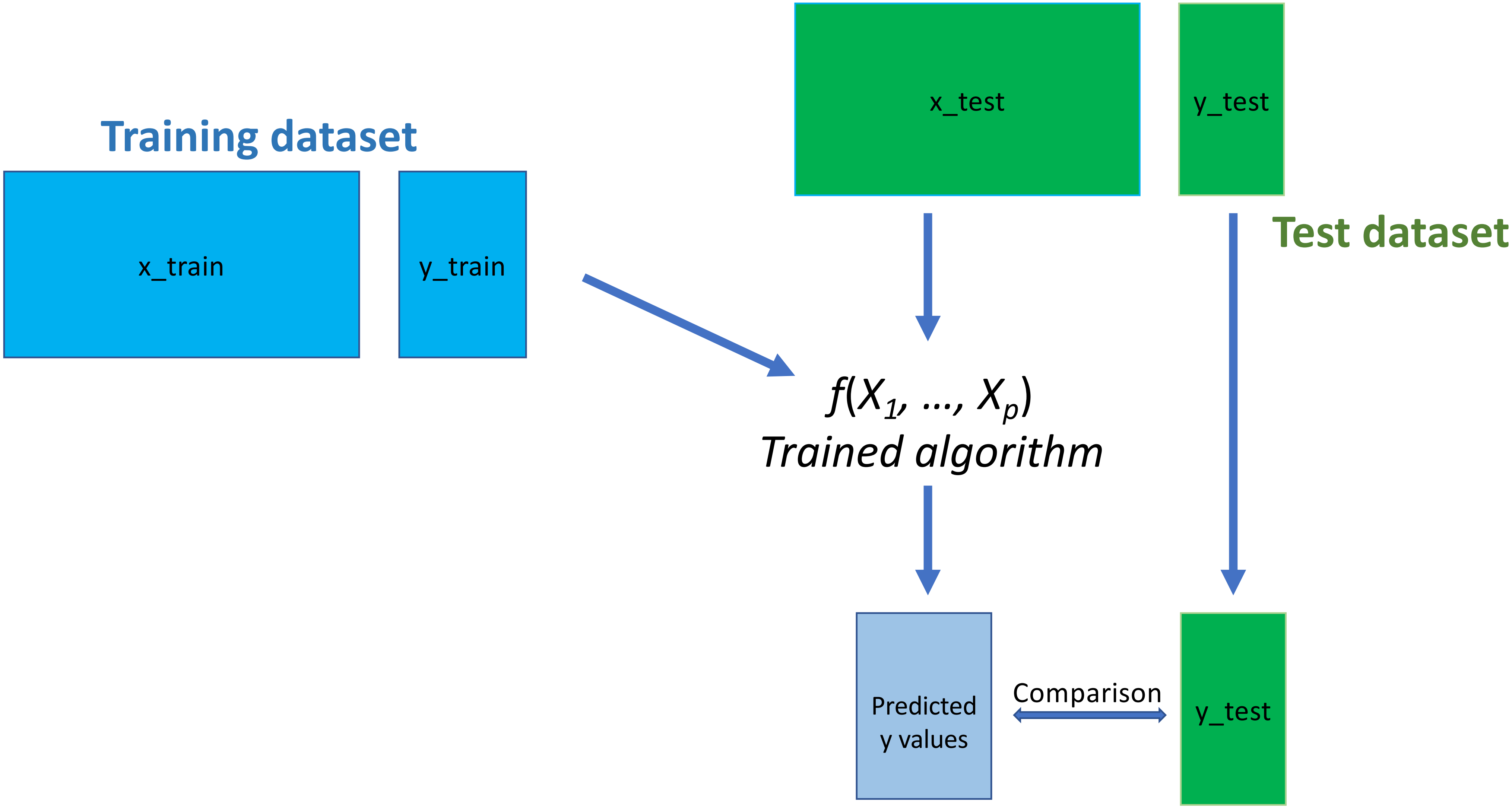


Testing: Assess the predictive capability of the trained algorithm using a **test dataset**




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


Training dataset
x_train
y_train
x_test
y_test
Test dataset
$f(X_1, \ldots, X_p)$
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

RMSE

#	△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

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

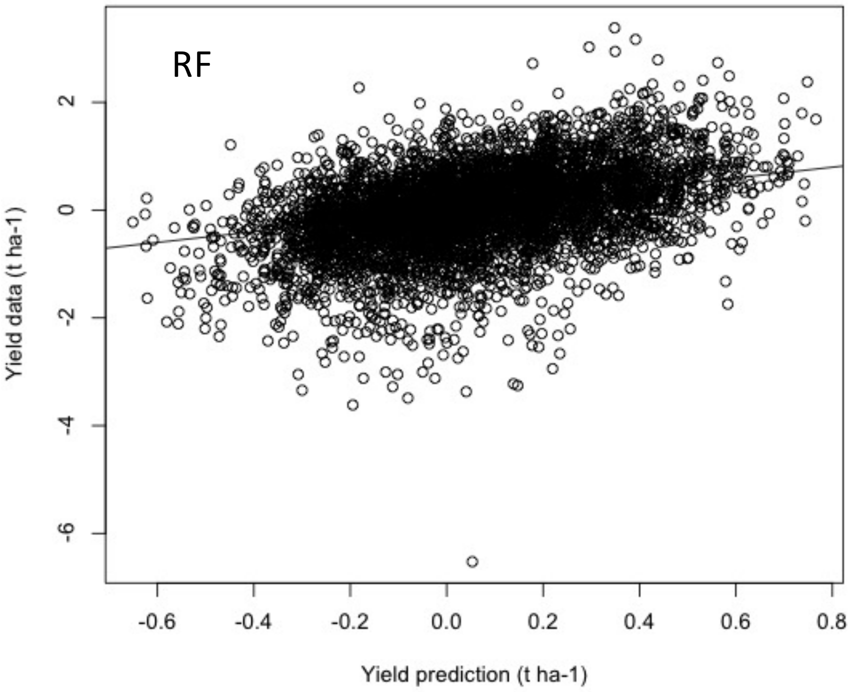
Algorithms developed by the participants

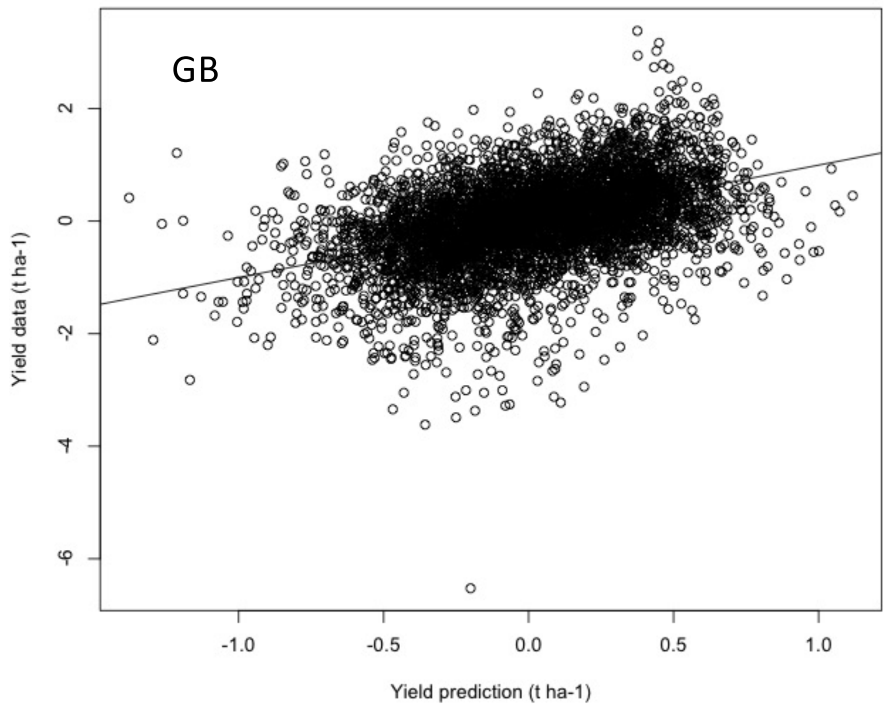
Test dataset

55 inputs
1708 yield data

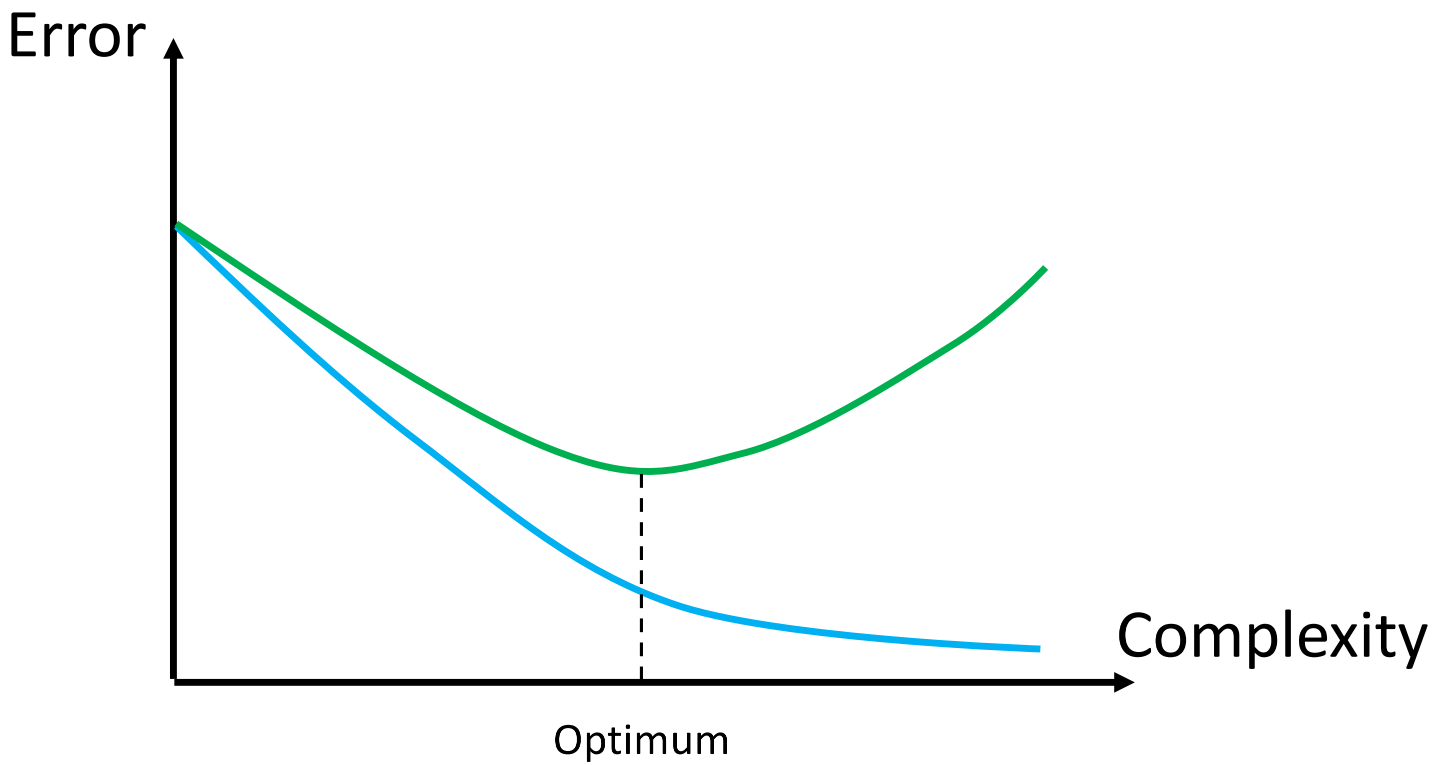
Evaluation of the accuracy of the algorithms by the organizer

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

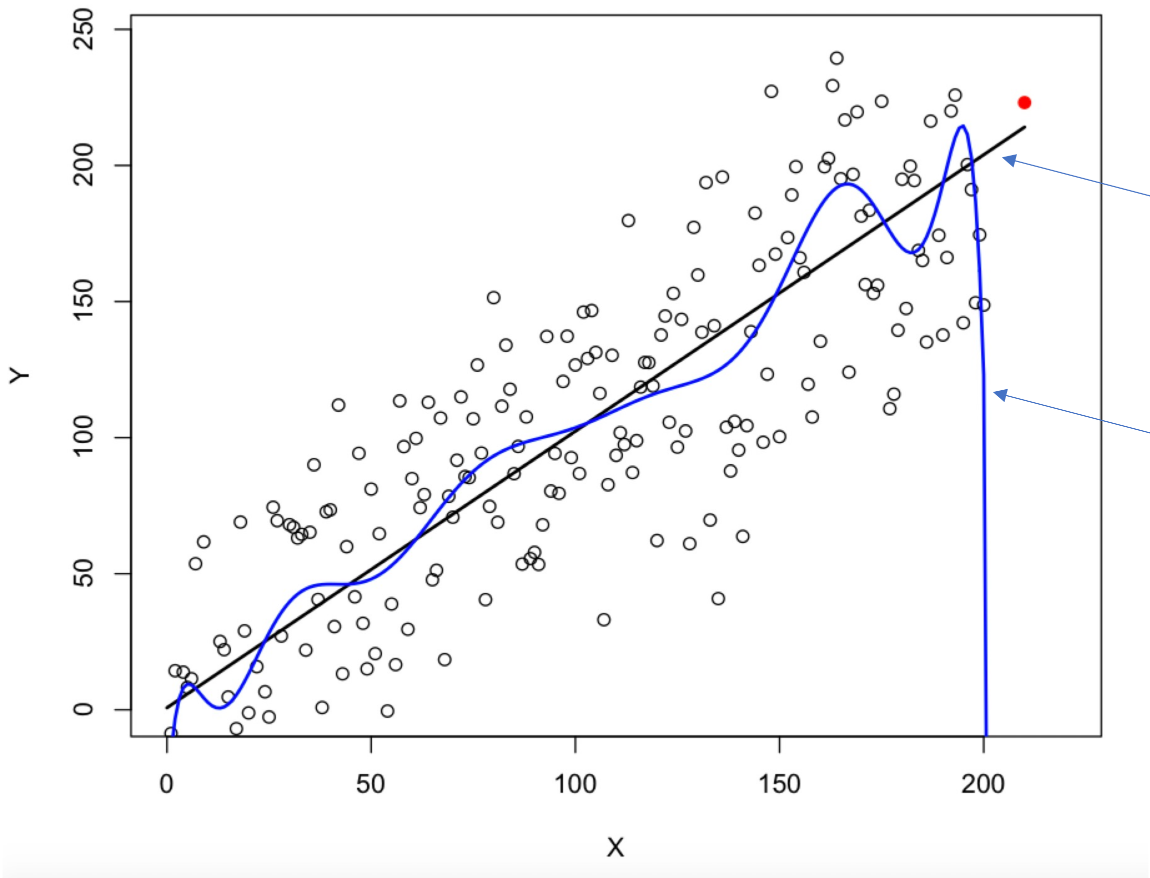




Model testing should be taken seriously to avoid risk of overfitting



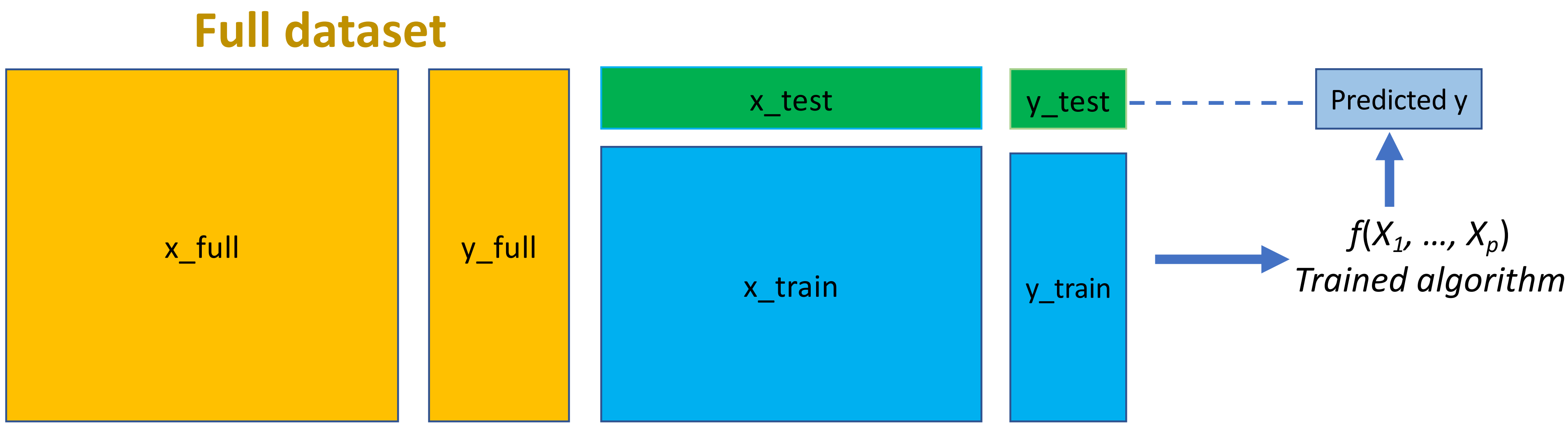
Model testing should be taken seriously to avoid risk of overfitting

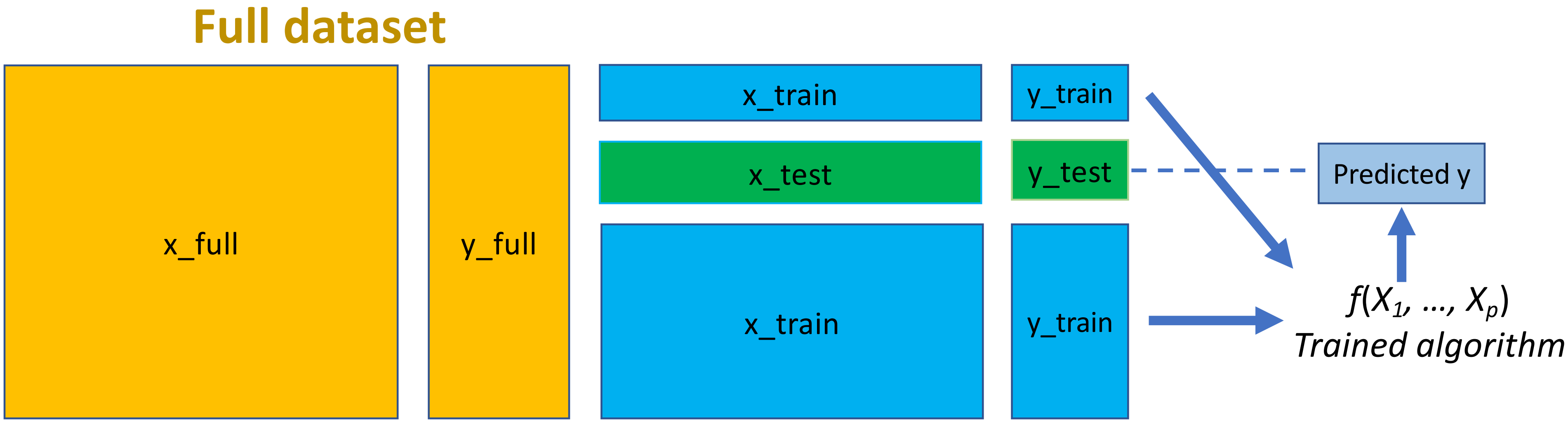


Linear regression:
R2=0.69

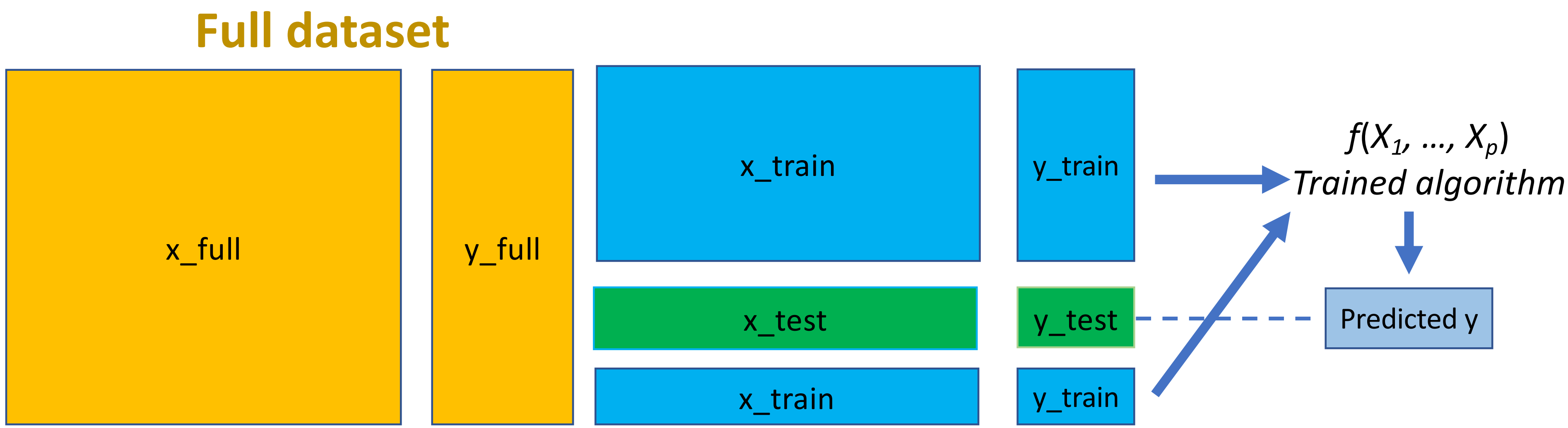
Polynomial regression:
R2=0.72

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

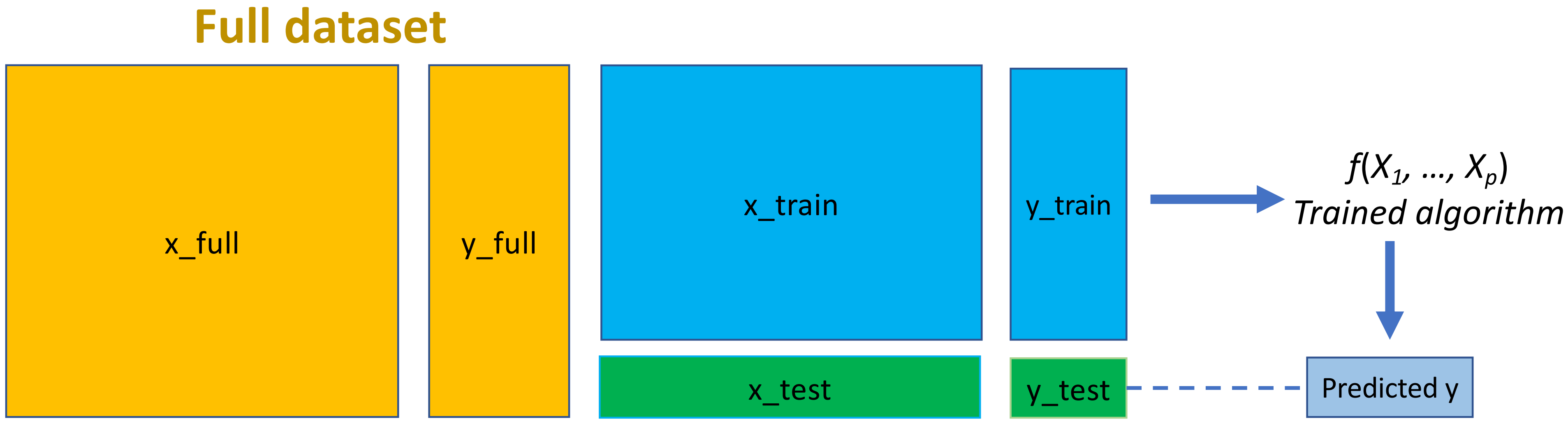

Full dataset
x_full
y_full
x_test
y_test
Predicted y
x_train
y_train
$f(X_1, \ldots, X_p)$
Trained algorithm



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

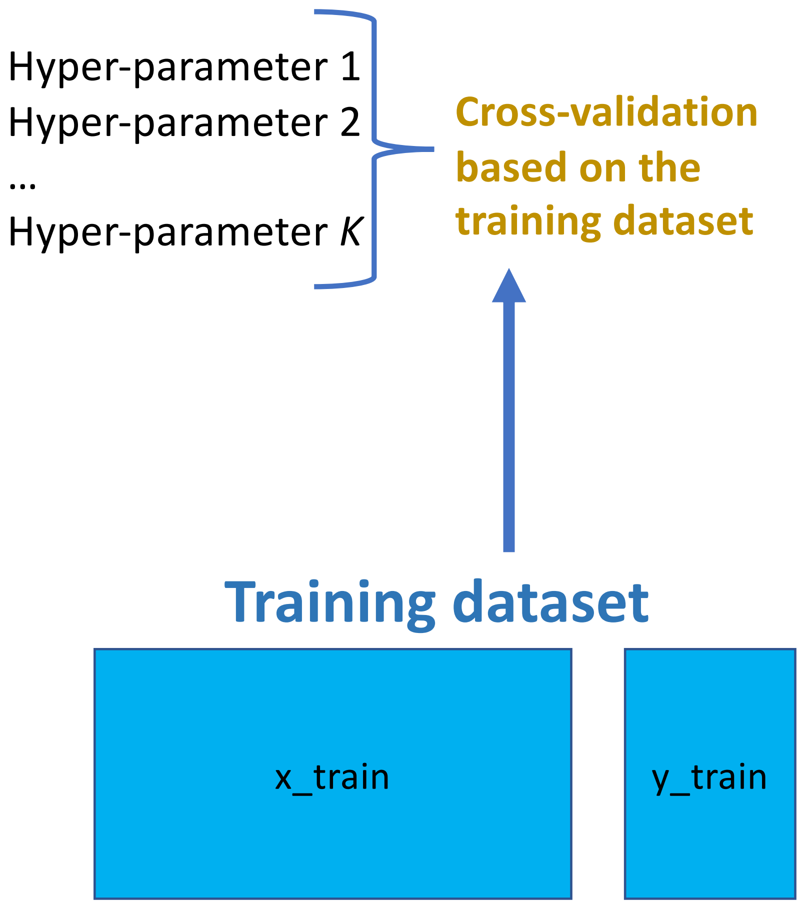


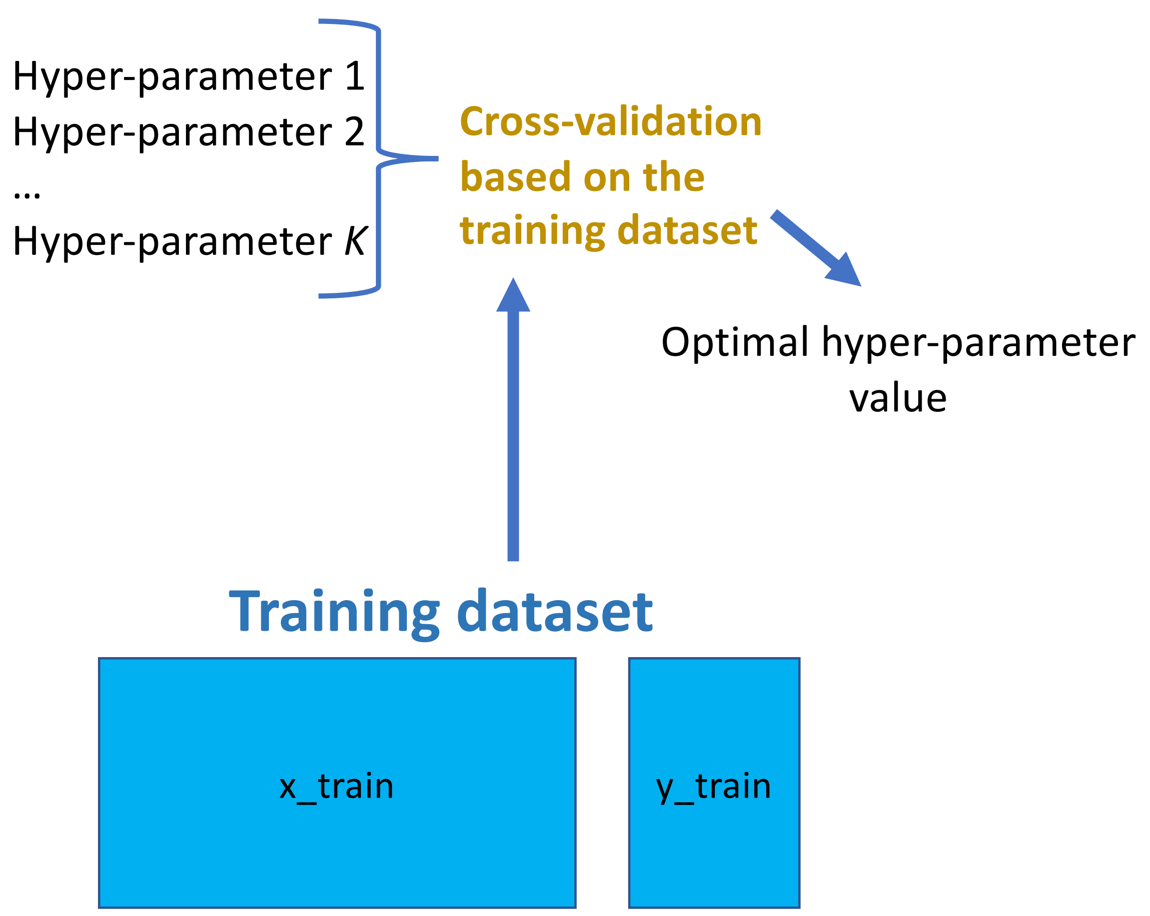
Full dataset
x_full
y_full
x_train
y_train
x_test
y_test
x_train
y_train
f(X1, ..., Xp)
Trained algorithm
Predicted y

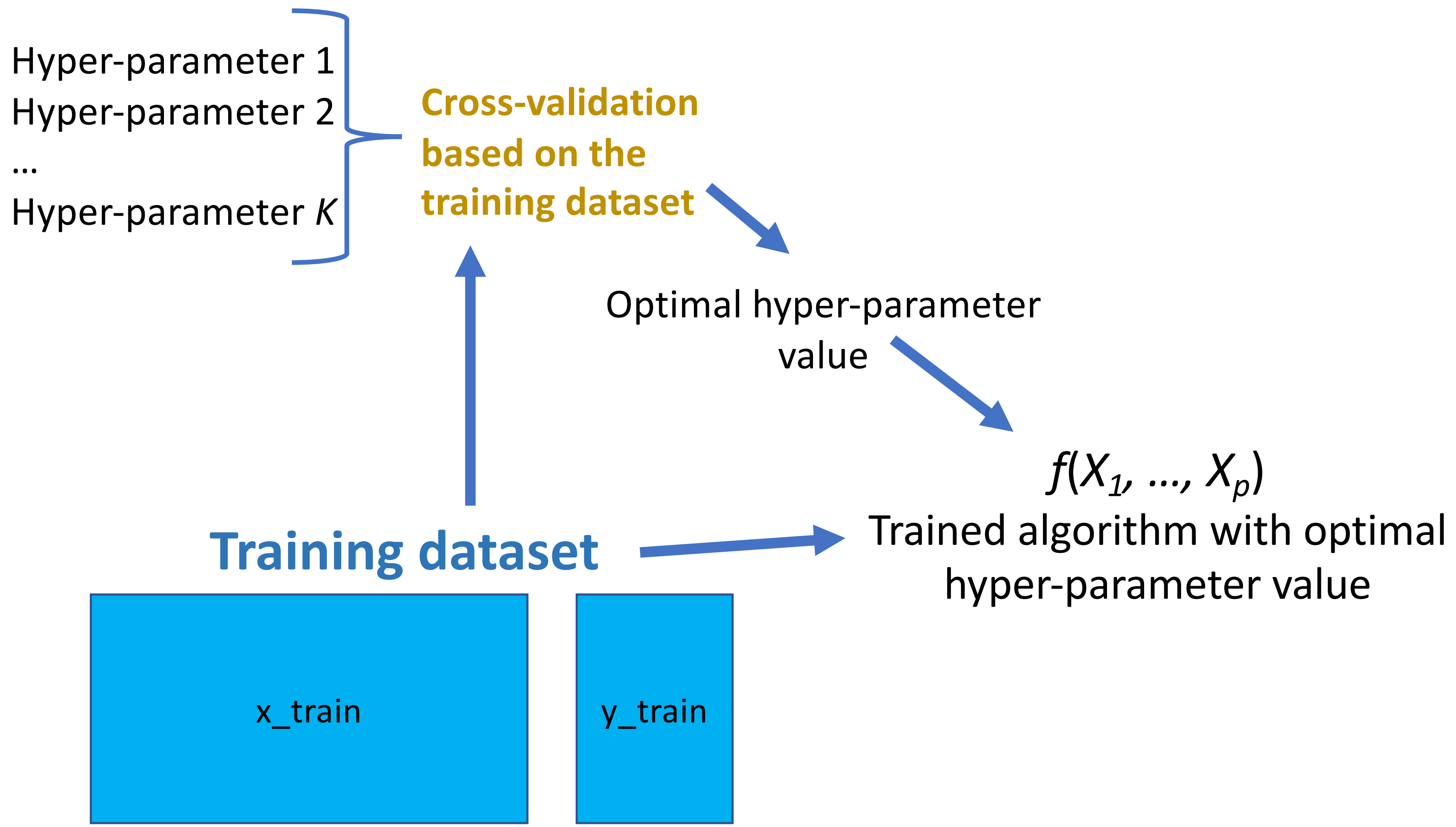


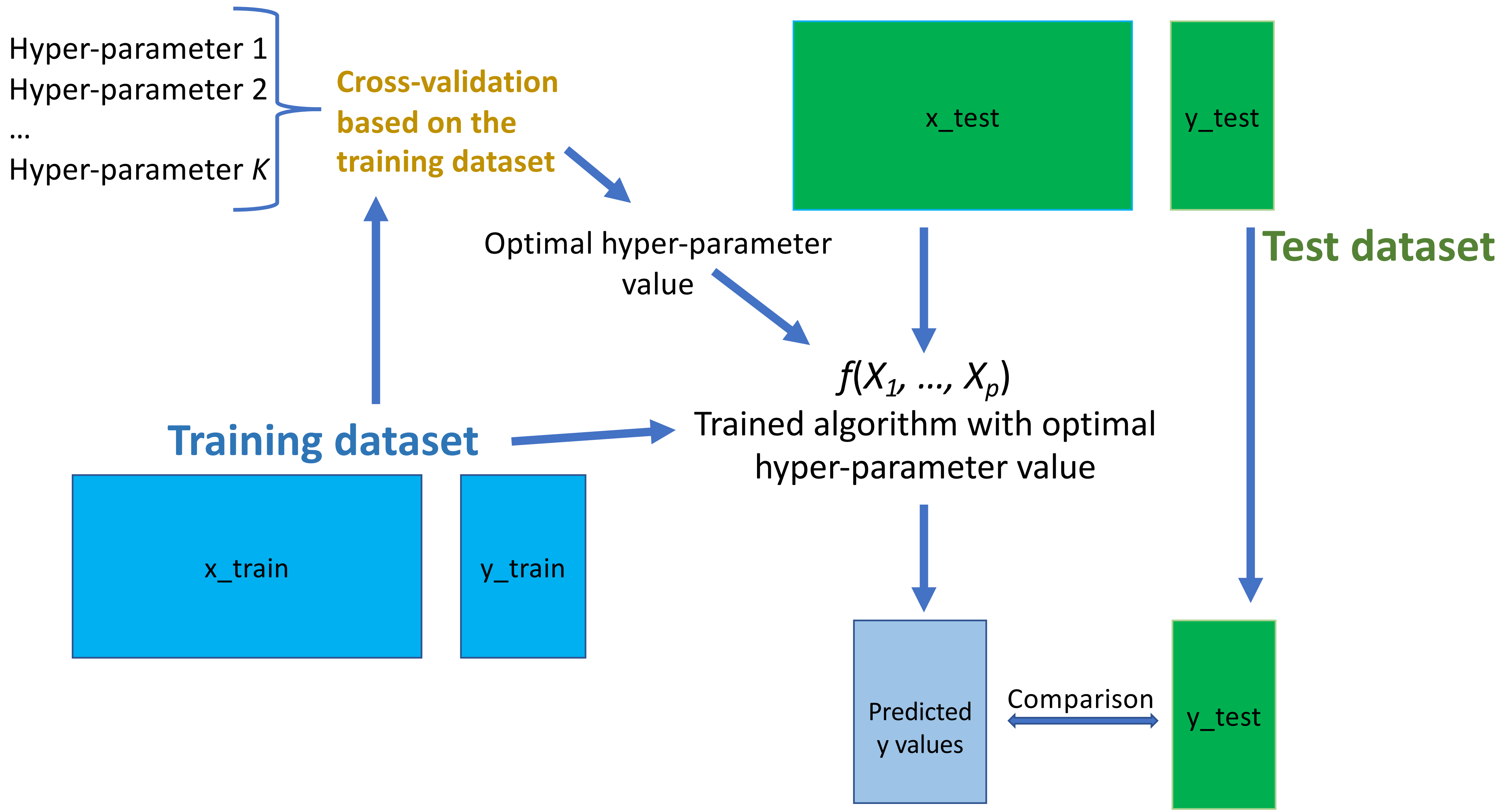
Full dataset
x_full
y_full
x_train
y_train
$f(X_1, \ldots, X_p)$
Trained algorithm
x_test
y_test
Predicted y

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(X1, ..., Xp)
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 »?

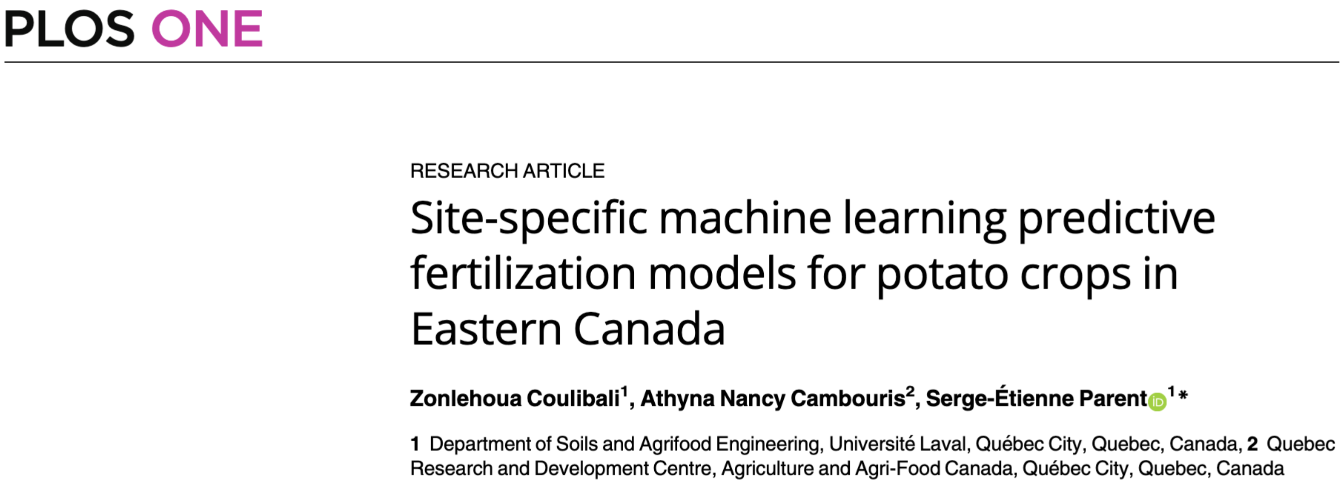
This is less true than before.

Vizualisation tools:

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

Example of machine learning project:
N, P, K fertilization models for potato crops in Eastern Canada

https://doi.org/10.1371/journal.pone.0230888

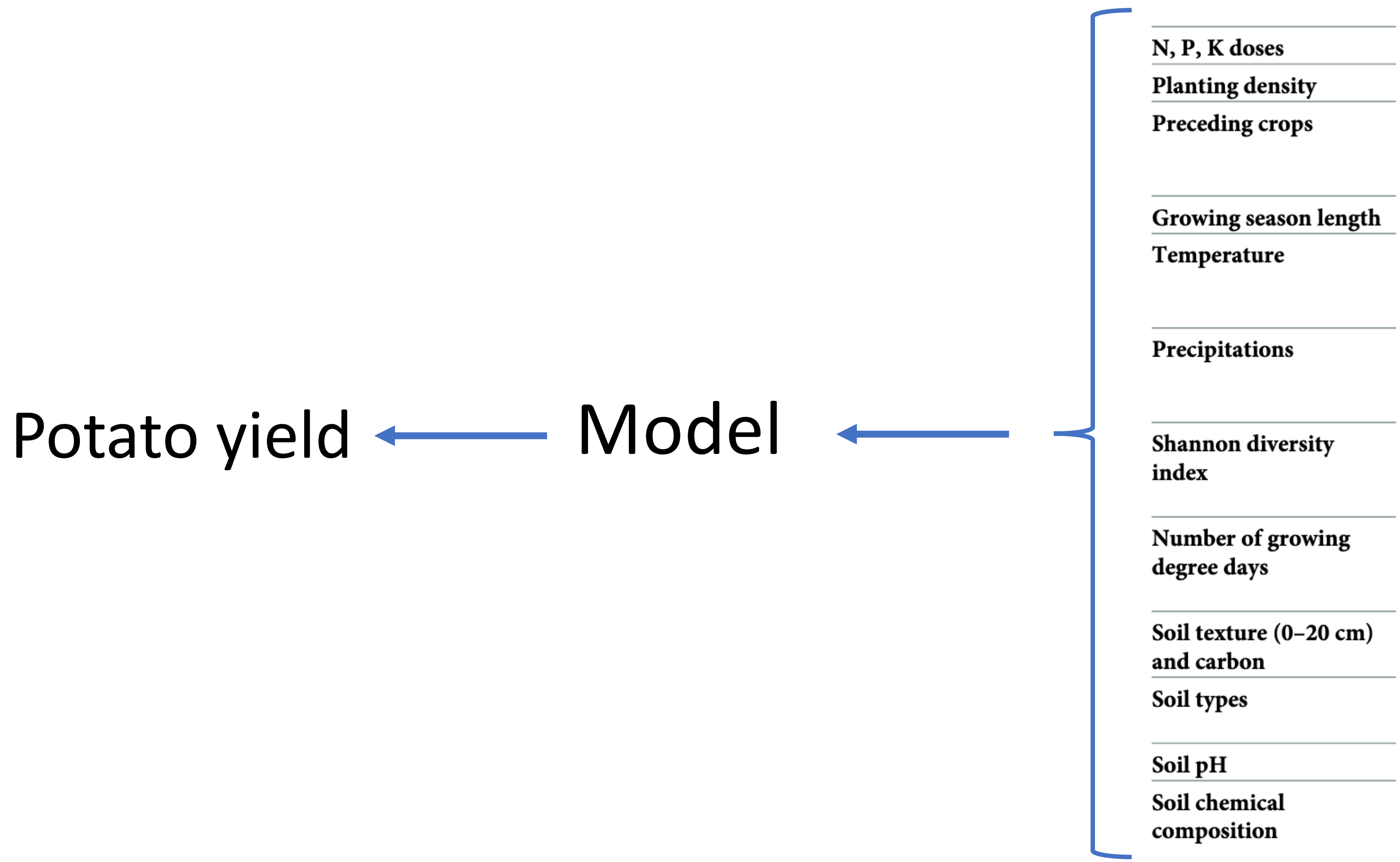


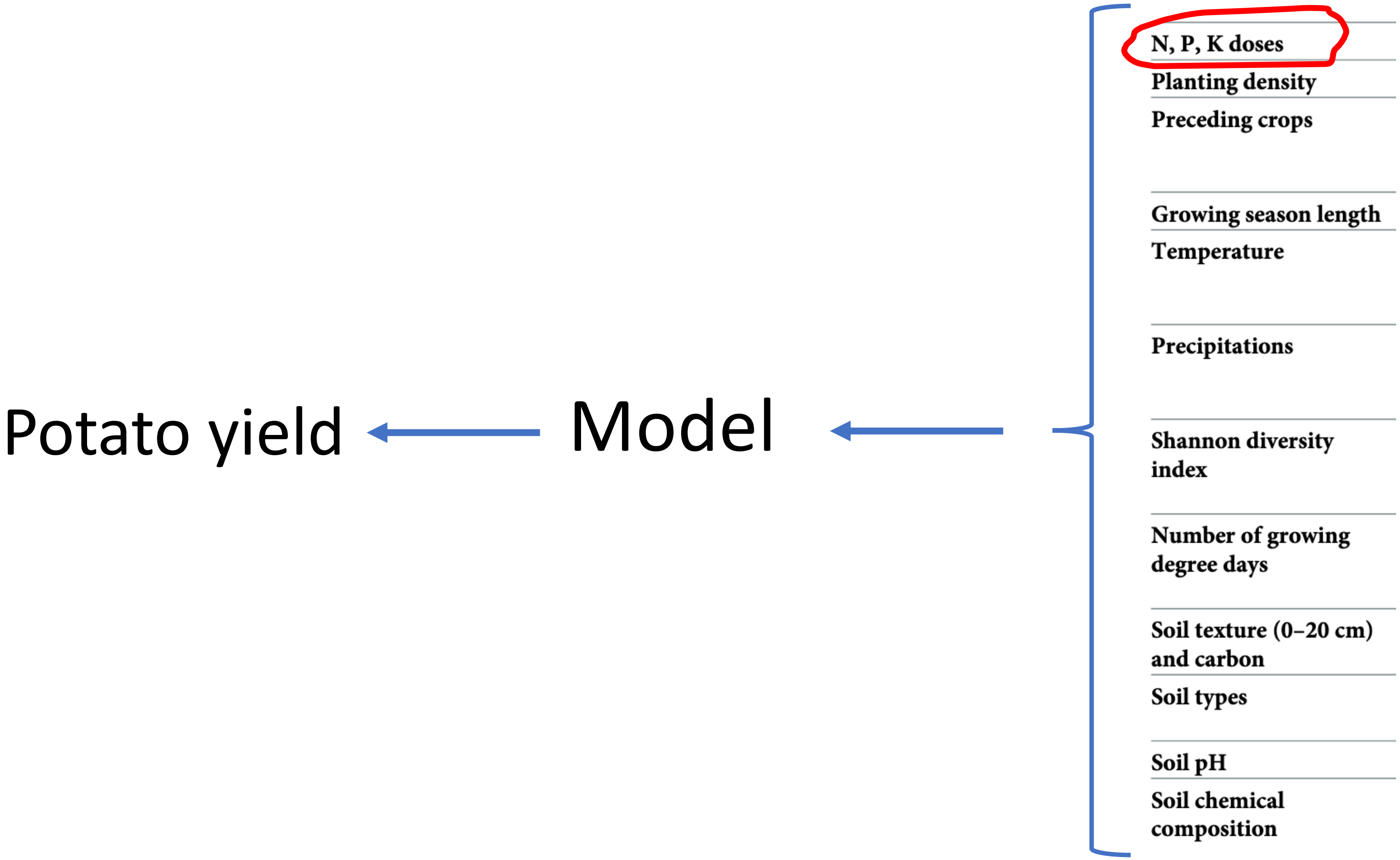
RESEARCH ARTICLE

Site-specific machine learning predictive fertilization models for potato crops in Eastern Canada

Zonlehoua Coulibali[1], Athyna Nancy Cambouris[2], Serge-Étienne Parent[1]*

1 Department of Soils and Agrifood Engineering, Université Laval, Québec City, Quebec, Canada, **2** Quebec Research and Development Centre, Agriculture and Agri-Food Canada, Québec City, Quebec, Canada


Potato yield
Model
N, P, K doses
Planting density
Preceding crops
Growing season length
Temperature
Precipitations
Shannon diversity index
Number of growing degree days
Soil texture (0–20 cm) and carbon
Soil types
Soil pH
Soil chemical composition


Potato yield
Model
N, P, K doses
Planting density
Preceding crops
Growing season length
Temperature
Precipitations
Shannon diversity index
Number of growing degree days
Soil texture (0–20 cm) and carbon
Soil types
Soil pH
Soil chemical composition

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

Step 3: Definition of candidate models

Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

Step 6: Model application

Main steps in a machine learning project

Step 1: Definition of the objective

Develop models to predict yields and calculate optimal N, P, K fertilizer doses for potato crops in Eastern Canada

Step 2: Data collection

Step 3: Definition of candidate models

Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

Step 6: Model application

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

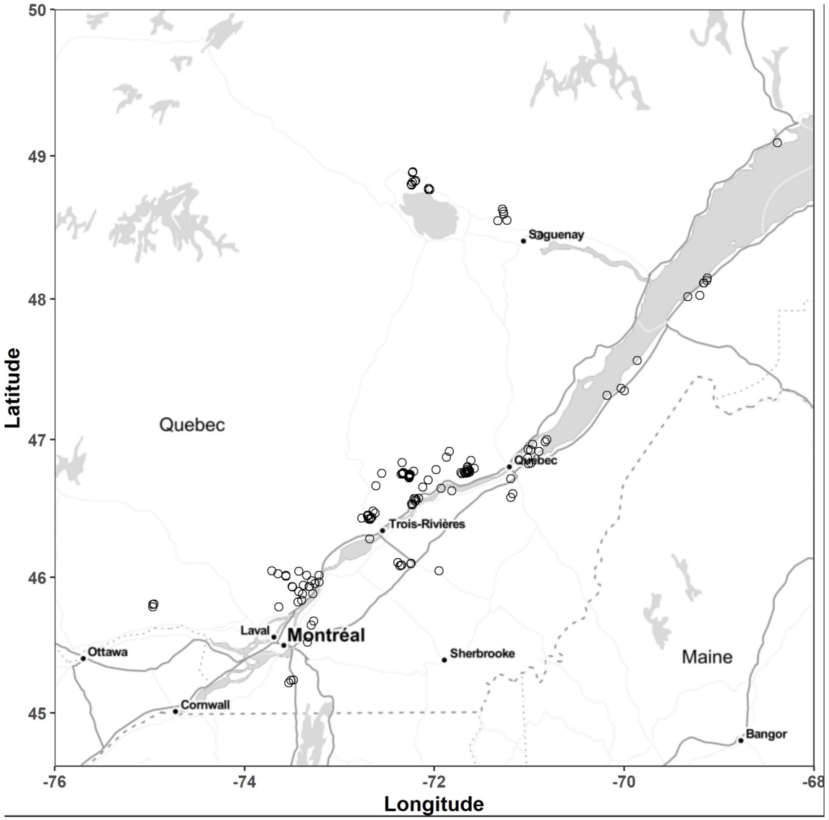
Step 3: Definition of candidate models

Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

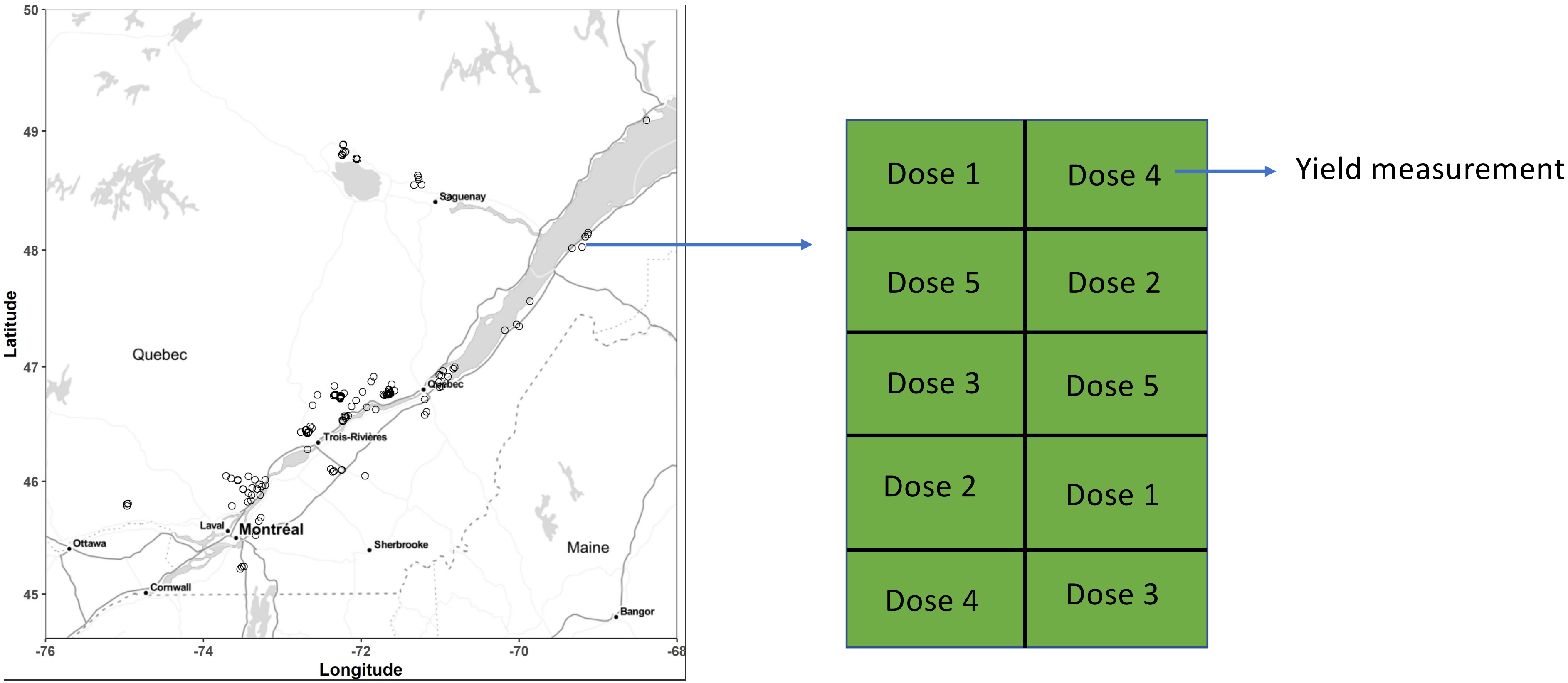
Step 6: Model application

237 field trials



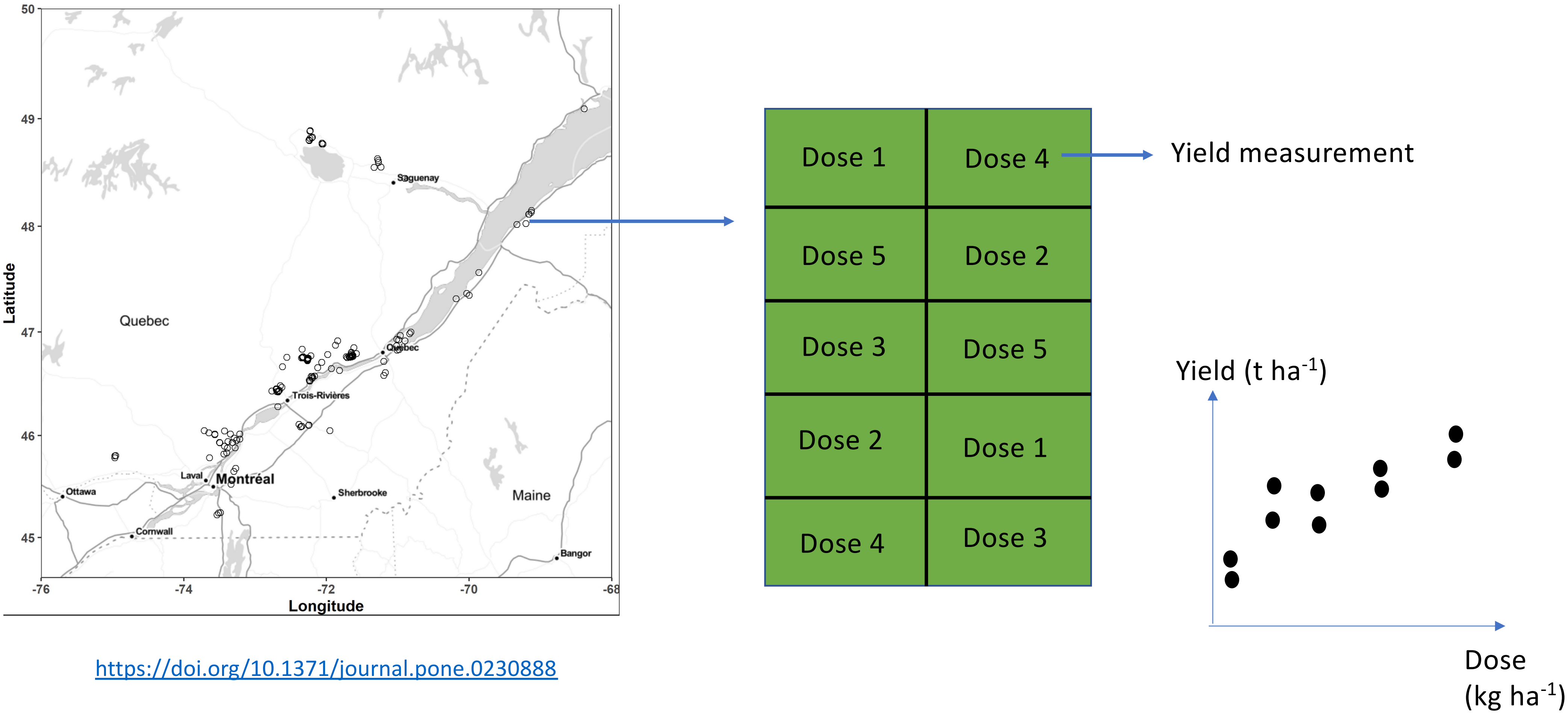
https://doi.org/10.1371/journal.pone.0230888

237 field trials



https://doi.org/10.1371/journal.pone.0230888

237 field trials



https://doi.org/10.1371/journal.pone.0230888

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

Step 3: Definition of candidate models

Step 4: Model training with data (parameter (

Step 5: Model testing with data (model evalu

Step 6: Model application

Five models

1. Mitscherlich
2. KNN
3. Random forest
4. Neural network
5. Gaussian process

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

Step 3: Definition of candidate models

Step 4: Model training with data (parameter

Step 5: Model testing with data (model evalu

Step 6: Model application

Five models

1. Mitscherlich
2. KNN
3. Random forest
4. Neural network
5. Gaussian process

$$Y = A\, x(1 - e^{-R_N \mathrm{x}(E_N + dose_N)})\mathrm{x}(1 - e^{-R_P \mathrm{x}(E_P + dose_P)})\mathrm{x}(1 - e^{-R_K \mathrm{x}(E_K + dose_K)})$$

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

Step 3: Definition of candidate models

Step 4: Model training with data (parameter (

Step 5: Model testing with data (model evalu

Step 6: Model application

Five models

1. Mitscherlich
2. KNN
3. Random forest
4. Neural network
5. Gaussian process

Standard machine learning models

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

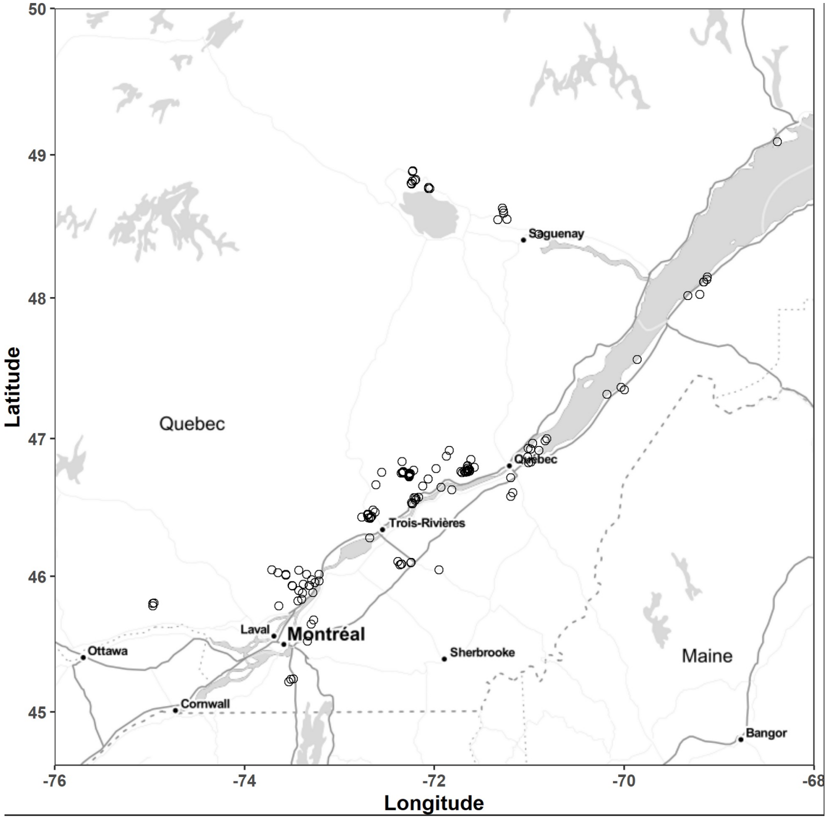
Step 3: Definition of candidate models

Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

Step 6: Model application

237 field trials



https://doi.org/10.1371/journal.pone.0230888

Training dataset
60% of the trials

Parameter estimation
for the five models

Main steps in a machine learning project

Step 1: Definition of the objective

Step 2: Data collection

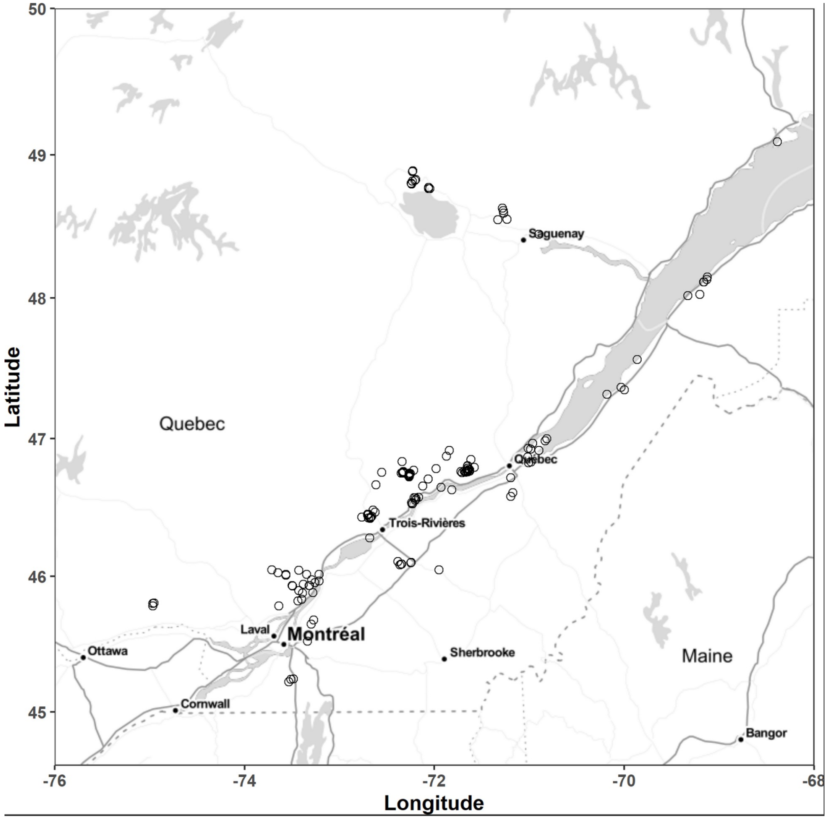
Step 3: Definition of candidate models

Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

Step 6: Model application

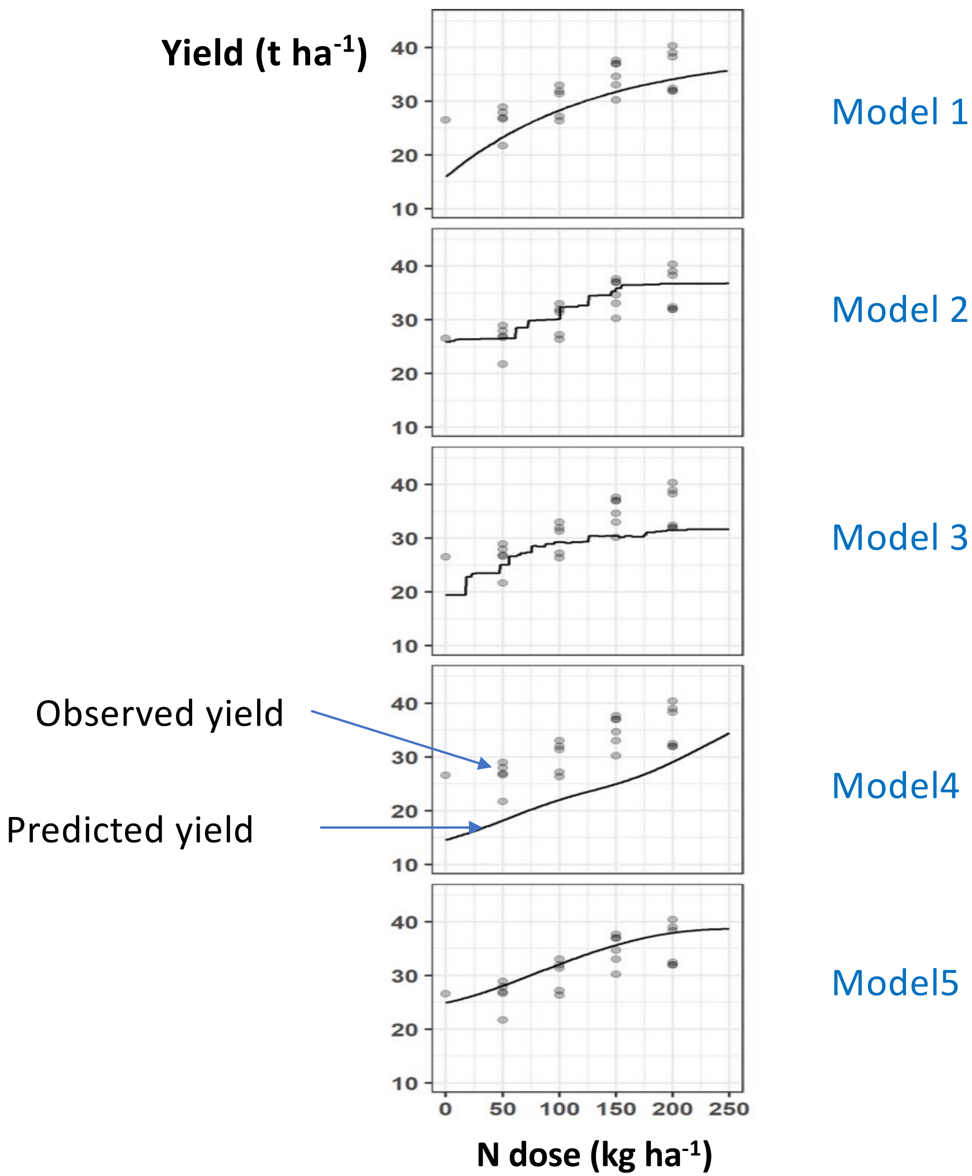
237 field trials



https://doi.org/10.1371/journal.pone.0230888

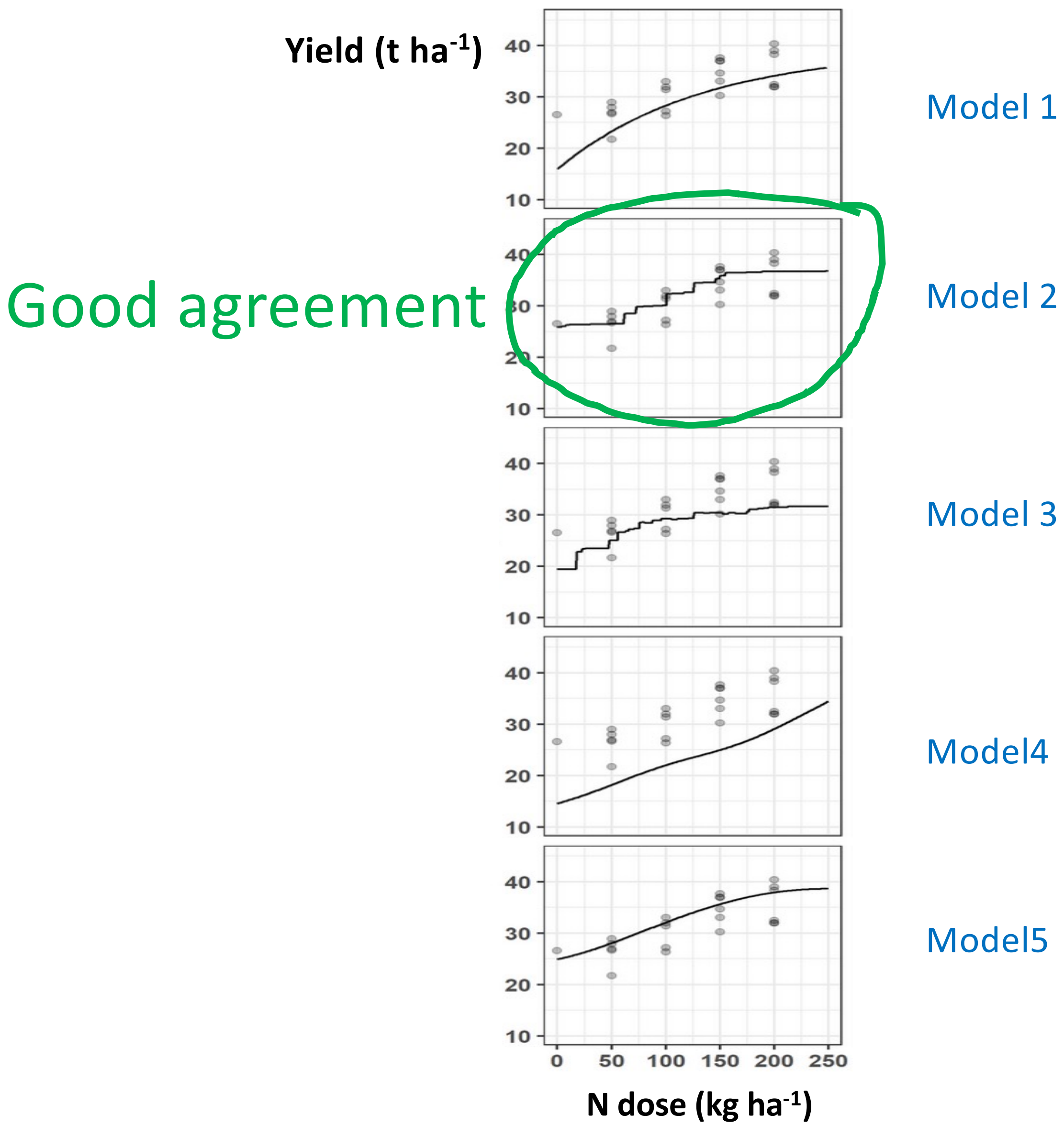
Testing dataset
40% of the trials

Evaluation of the model performances



Five models

1. Mitscherlich
2. KNN
3. Random forest
4. Neural network
5. Gaussian process



Five models

1. Mitscherlich
2. KNN
3. Random forest
4. Neural network
5. Gaussian process

Good agreement

Poor agreement

Model 1

Model 2

Model 3

Model4

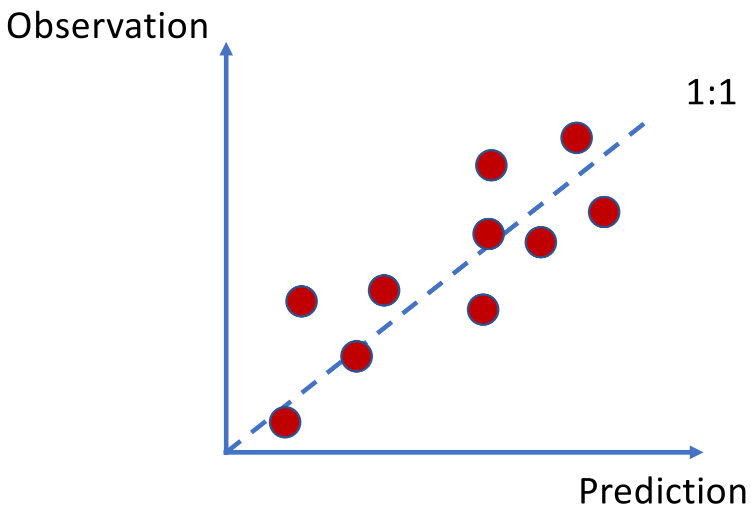
Model5

Five models

1. Mitscherlich
2. KNN
3. Random forest
4. Neural network
5. Gaussian process

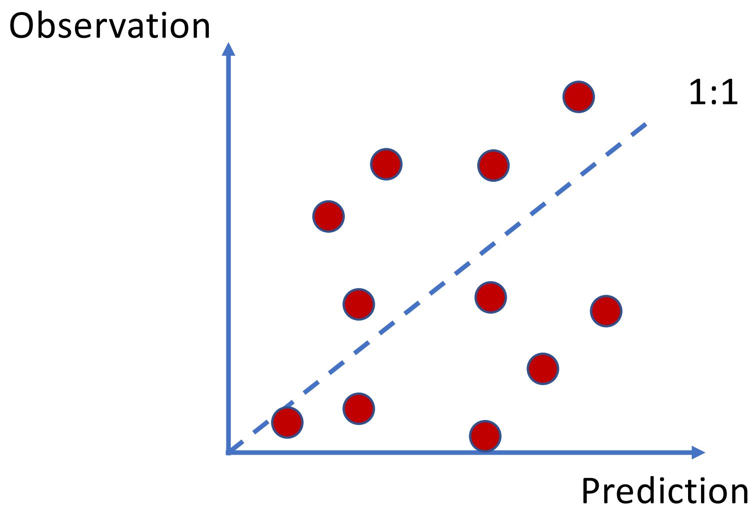
R^2 is a popular evaluation criterion

Good agreement



R^2 close to 1

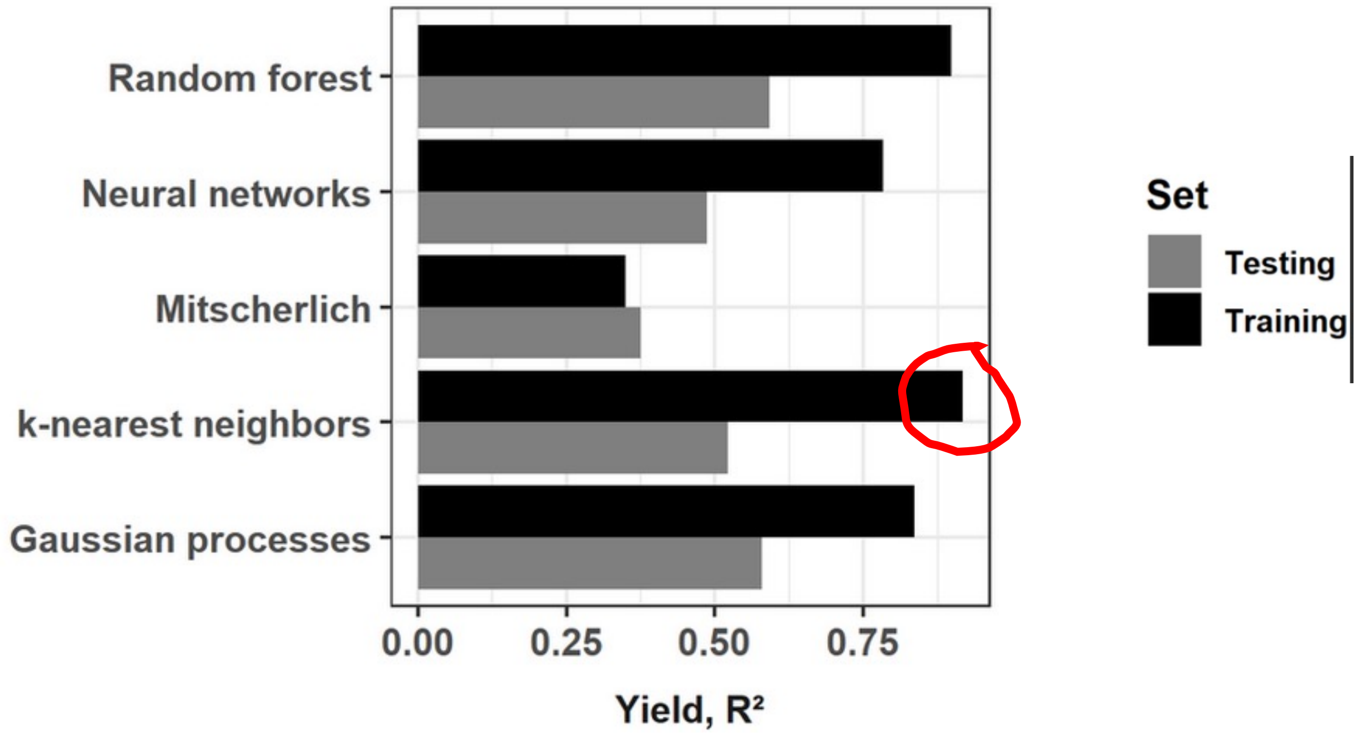
Poor agreement



R^2 close to 0

R^2

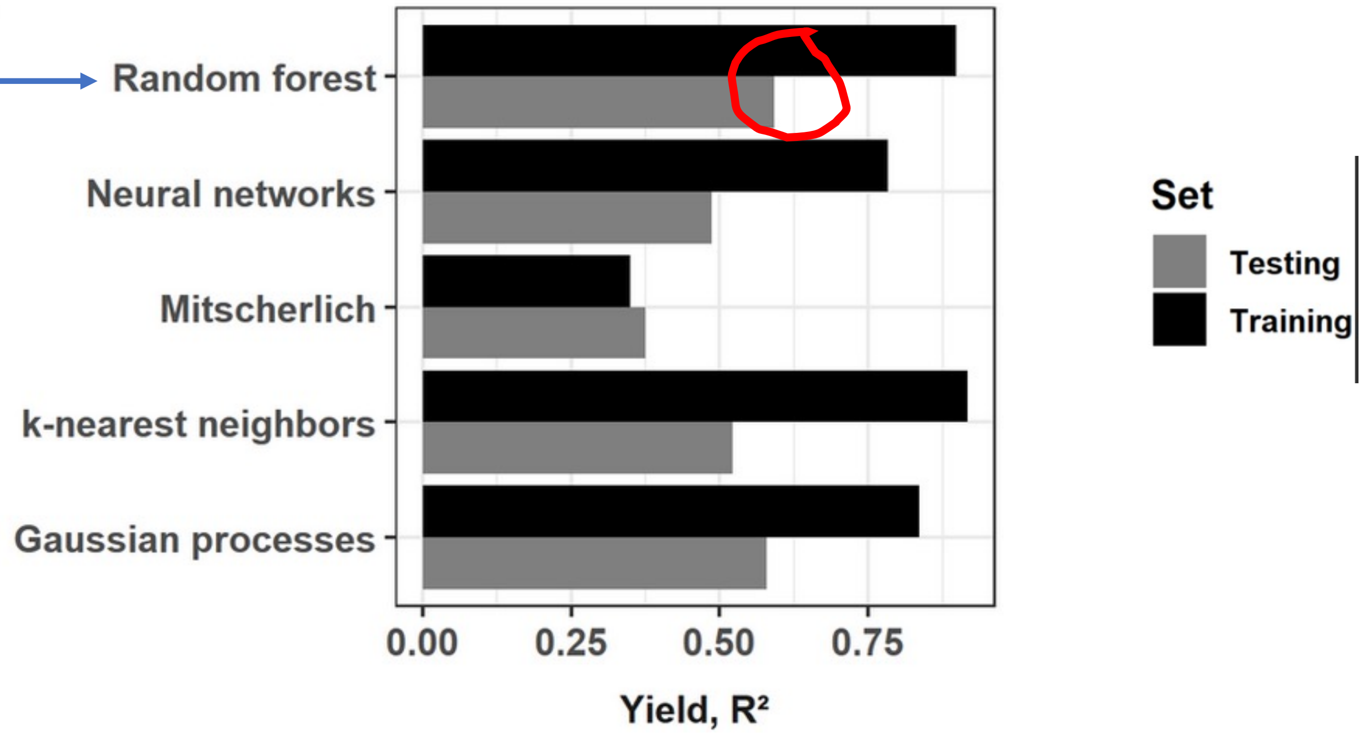
Best model according to the training dataset



https://doi.org/10.1371/journal.pone.0230888

R^2

Best model according to the test dataset

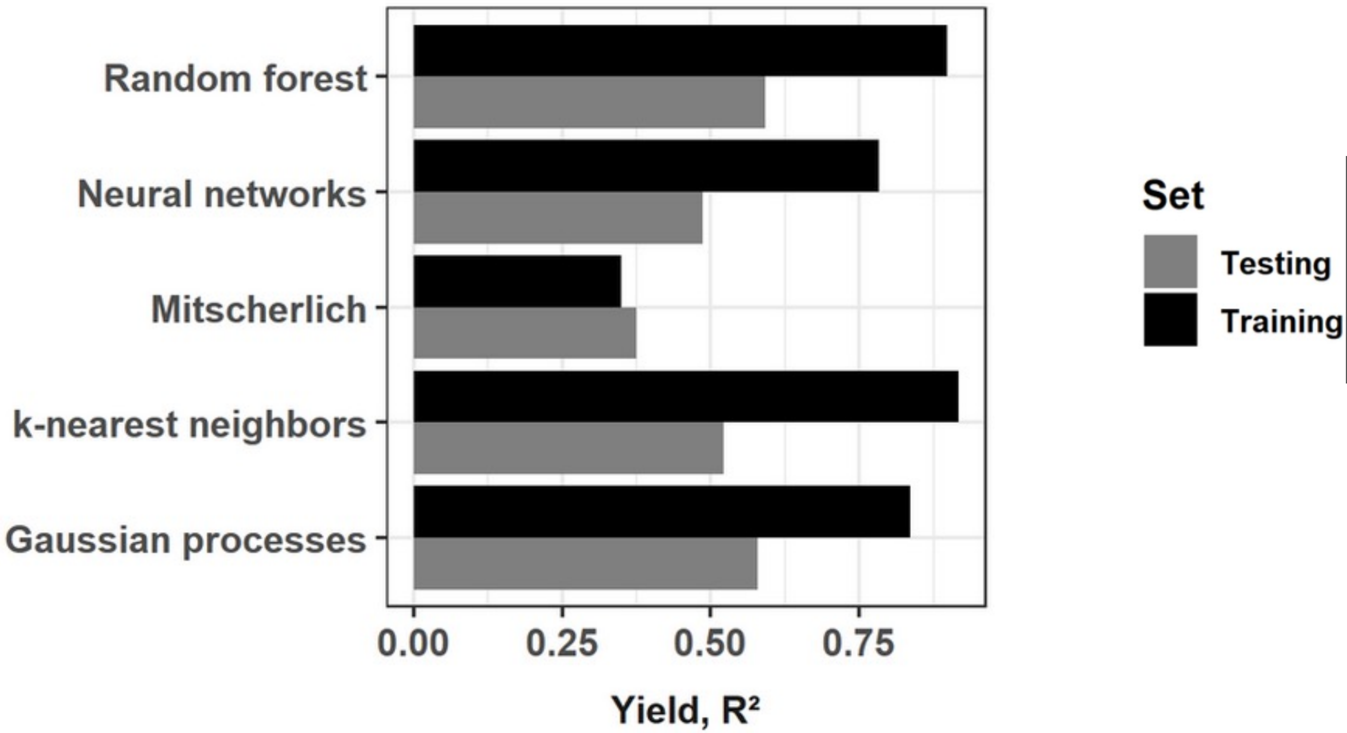


https://doi.org/10.1371/journal.pone.0230888

R^2

Model performances are too optimistic according to the training dataset.

Important to use an independent test dataset !



https://doi.org/10.1371/journal.pone.0230888

Step 1: Definition of the objective

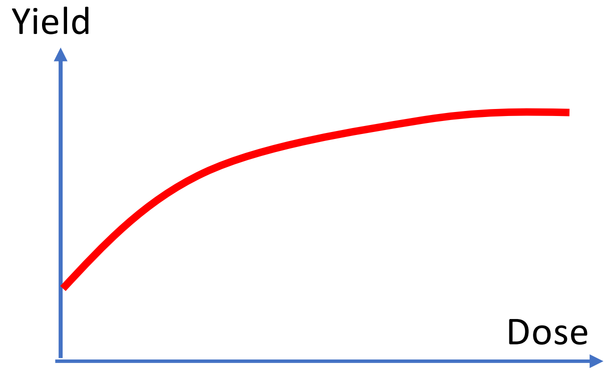
Step 2: Data collection

Step 3: Definition of candidate models

Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

Step 6: Model application



Step 1: Definition of the objective

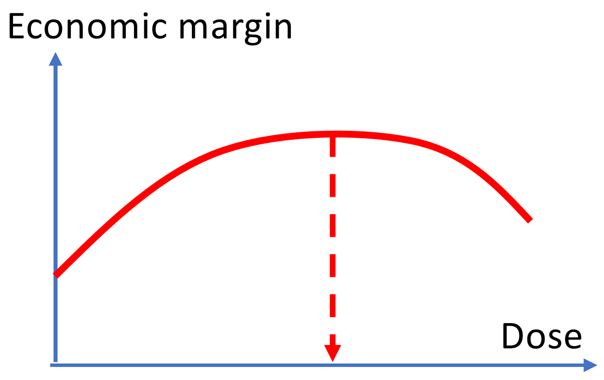
Step 2: Data collection

Step 3: Definition of candidate models

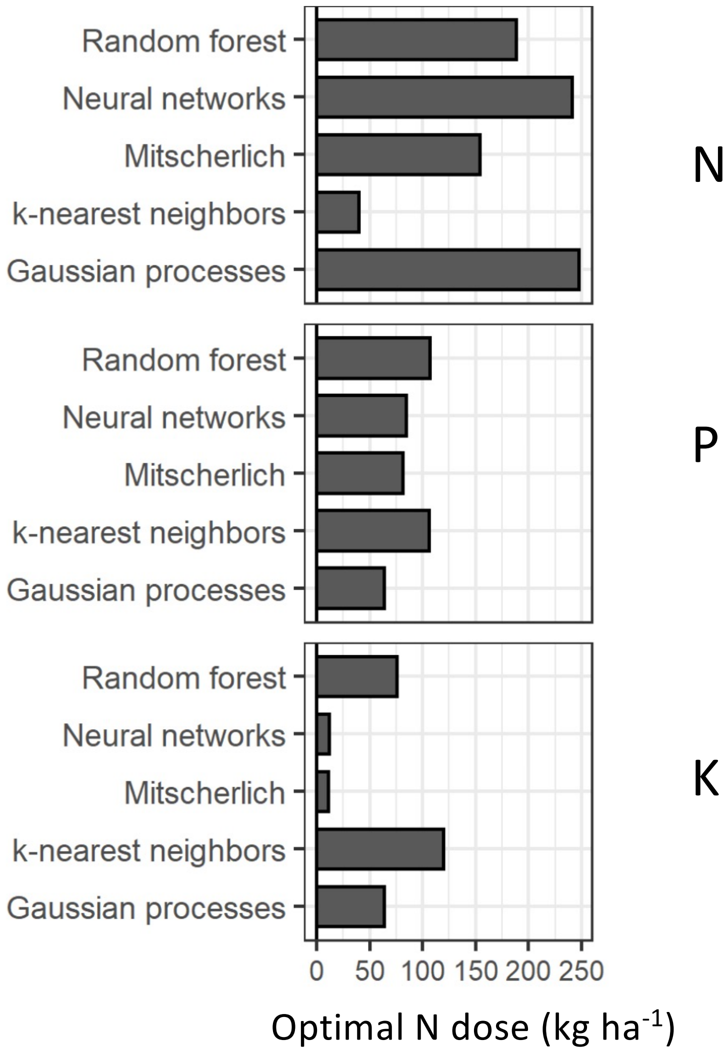
Step 4: Model training with data (parameter estimation)

Step 5: Model testing with data (model evaluation)

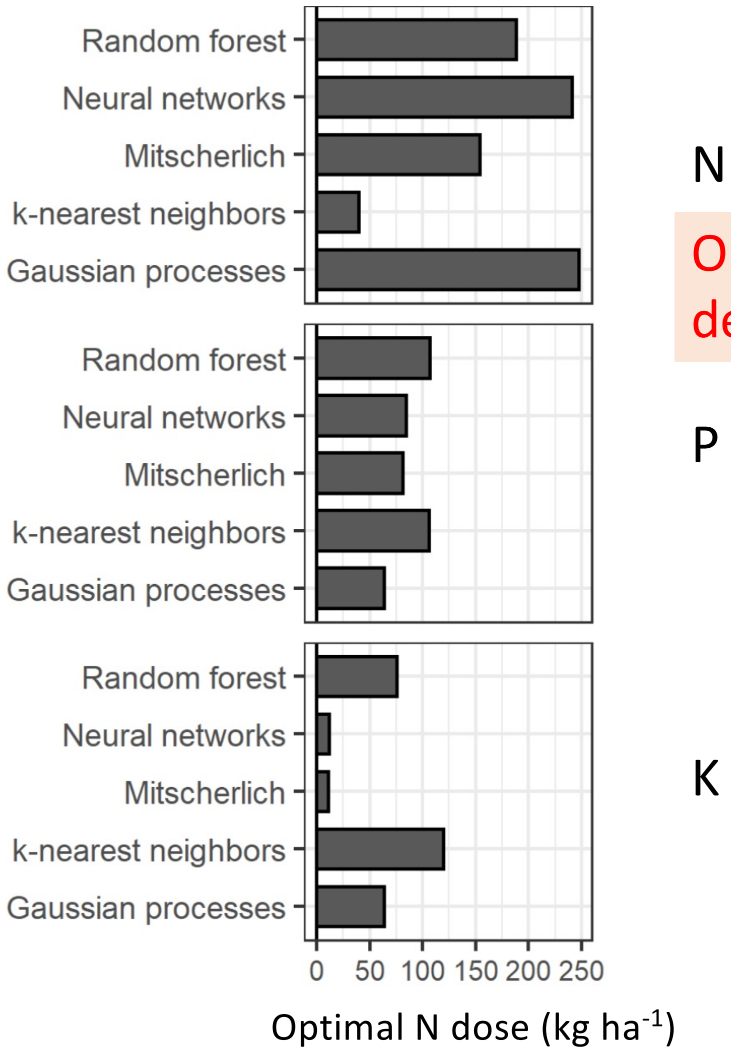
Step 6: Model application



Examples of optimal economic fertilizer doses at one site in Canada



Examples of optimal economic fertilizer doses at one site in Canada



Optimal doses are highly dependent on the selected model!

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