NATIONAL RESEARCH
UNIVERSITY

School of Data Analysis and Artificial
Intelligence Department of Computer Science

DATA SCIENCE FOR BUSINESS

Lecture 3. Introduction to Machine Learning

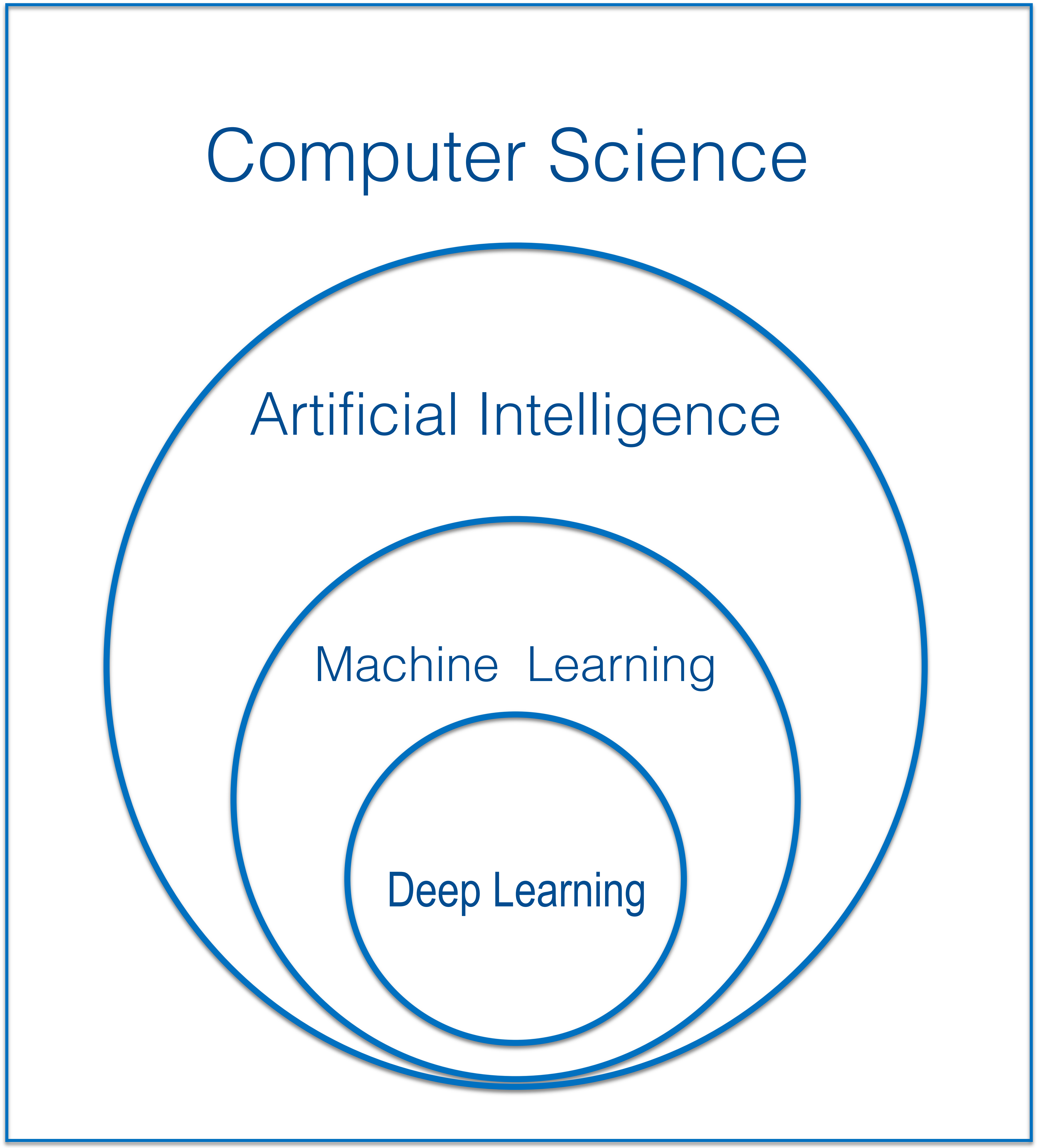
Moscow, April 24th, 2020.

lzhukov@hse.ru

WHAT IS MACHINE LEARNING?

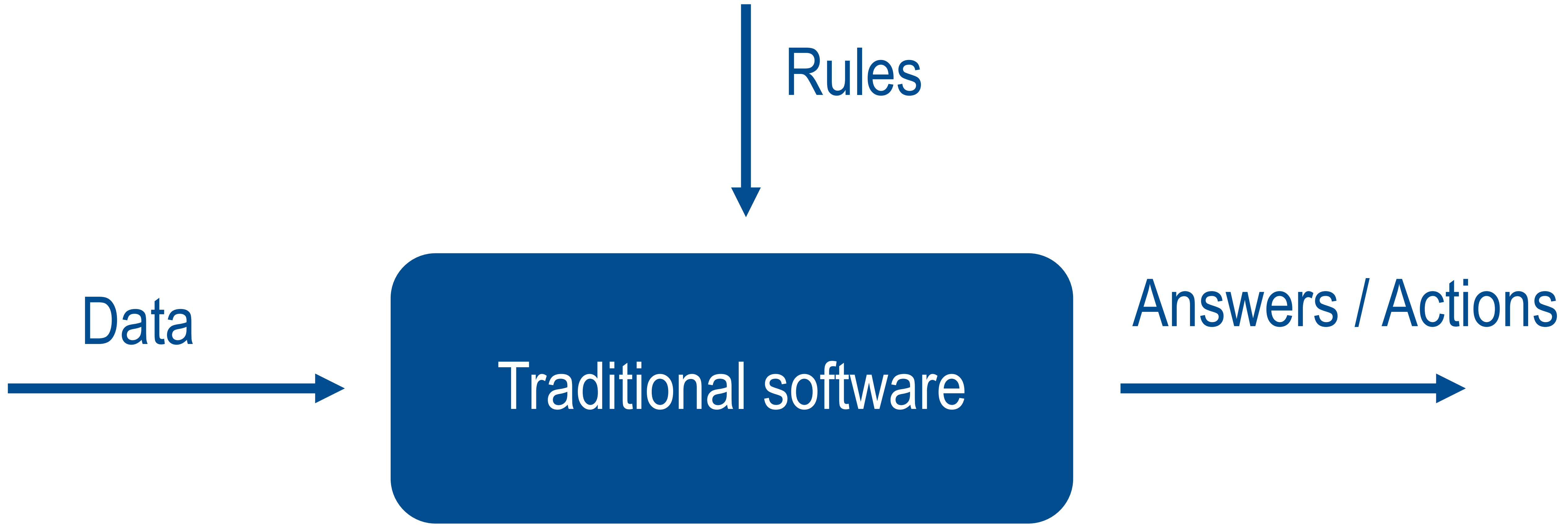
Machine learning is a subfield of computer science that studies and develops algorithms that can learn from data without being explicitly programmed

Machine learning algorithms can detect patterns in data and use them to predict future data

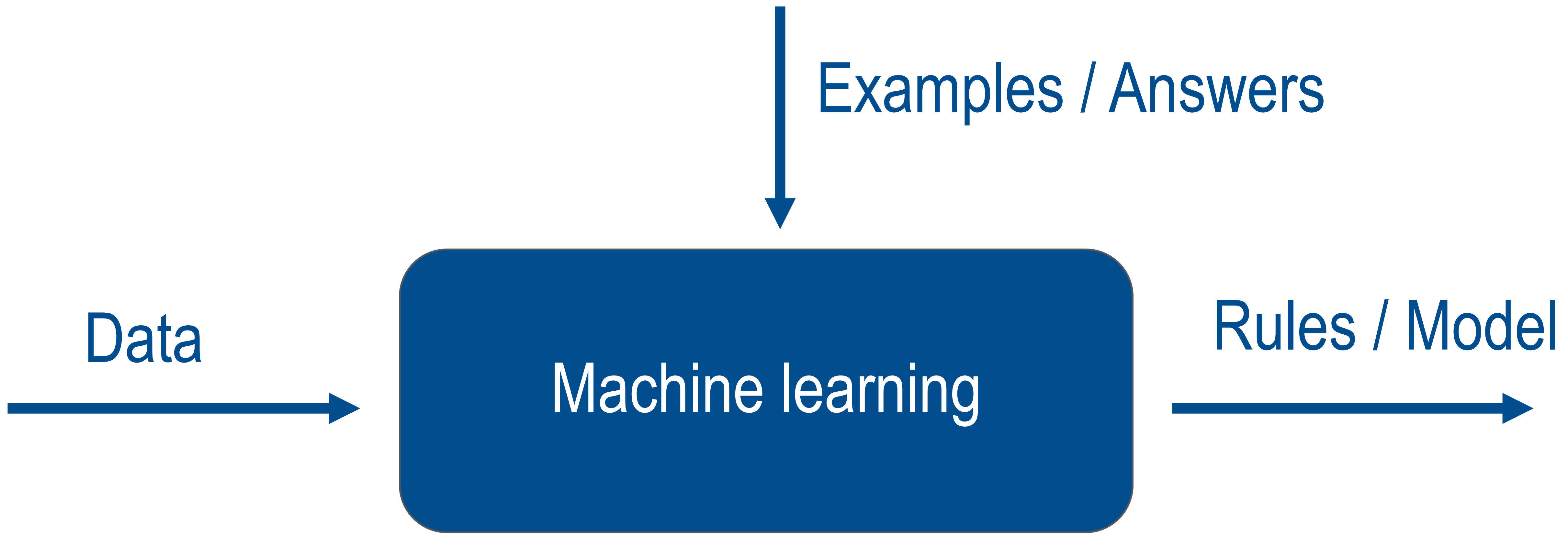


Machine learning – how is it different?

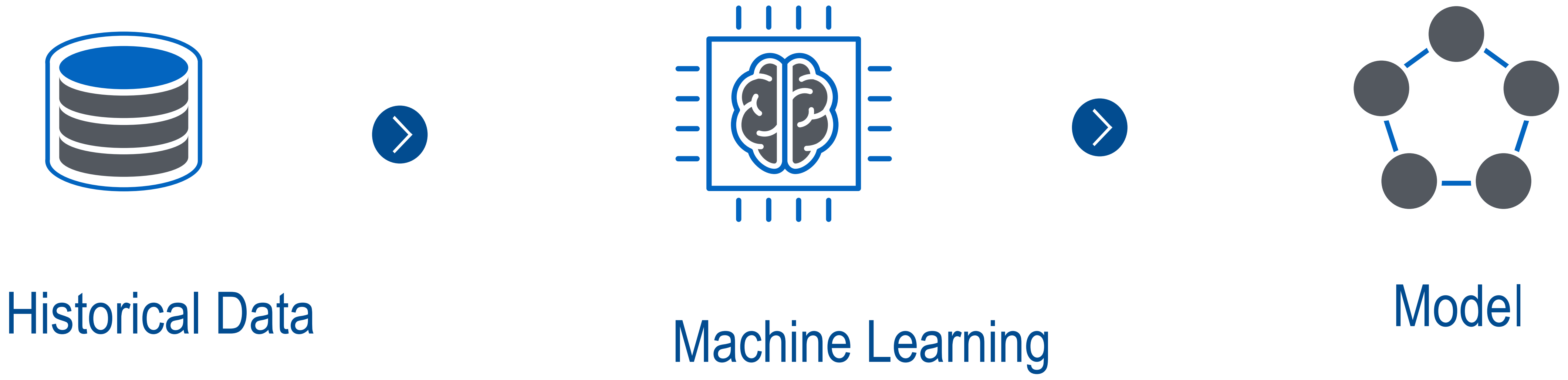
Traditional software: applying given rules to data



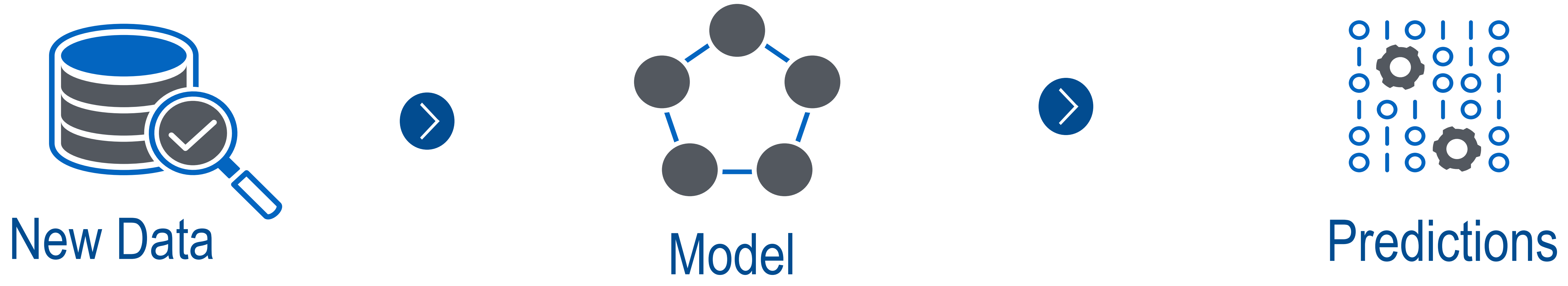
Machine learning: creating rule from data



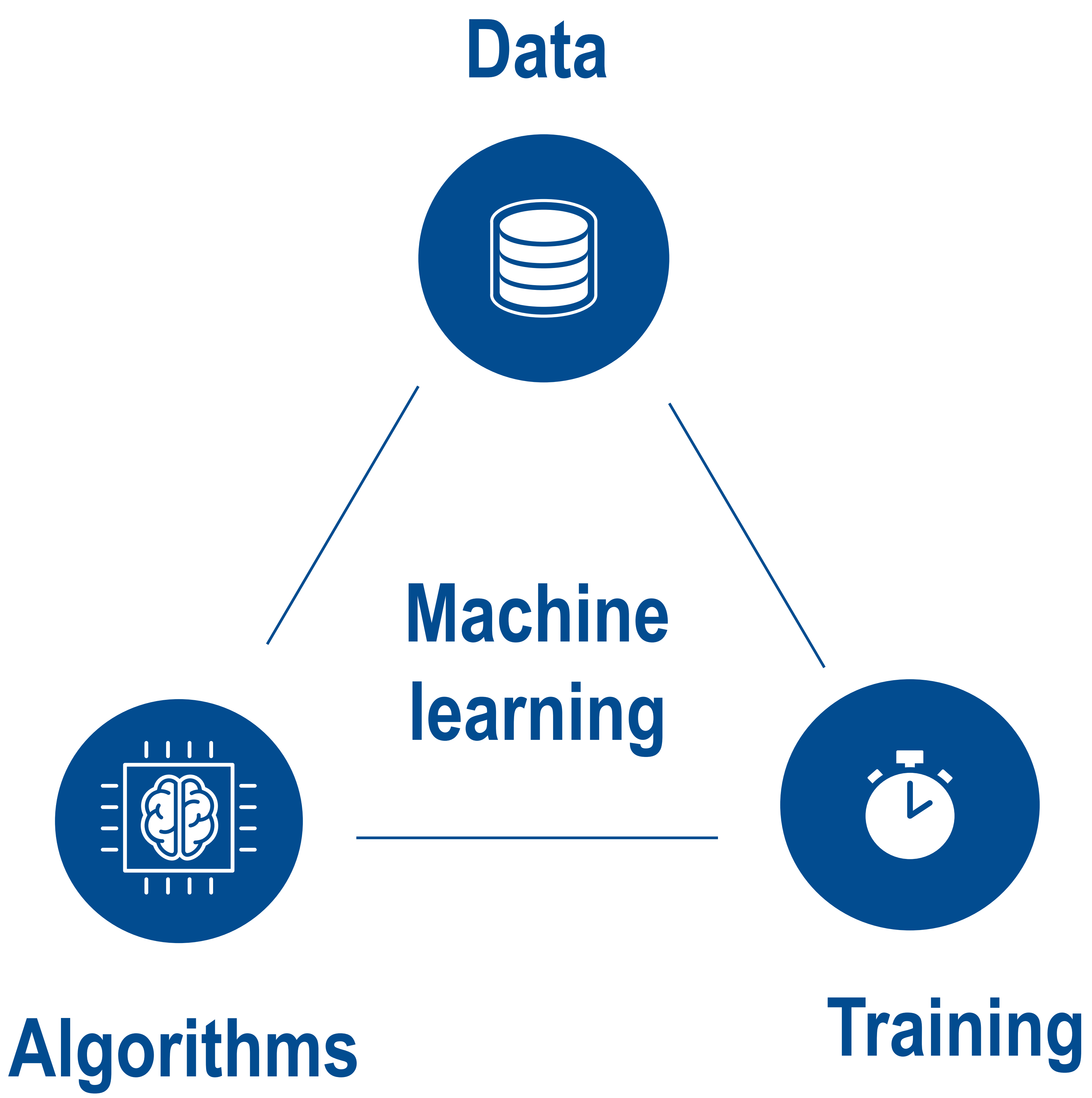
1 Model design, training and testing (model building, feature engineering)



2 Model application (model scoring)



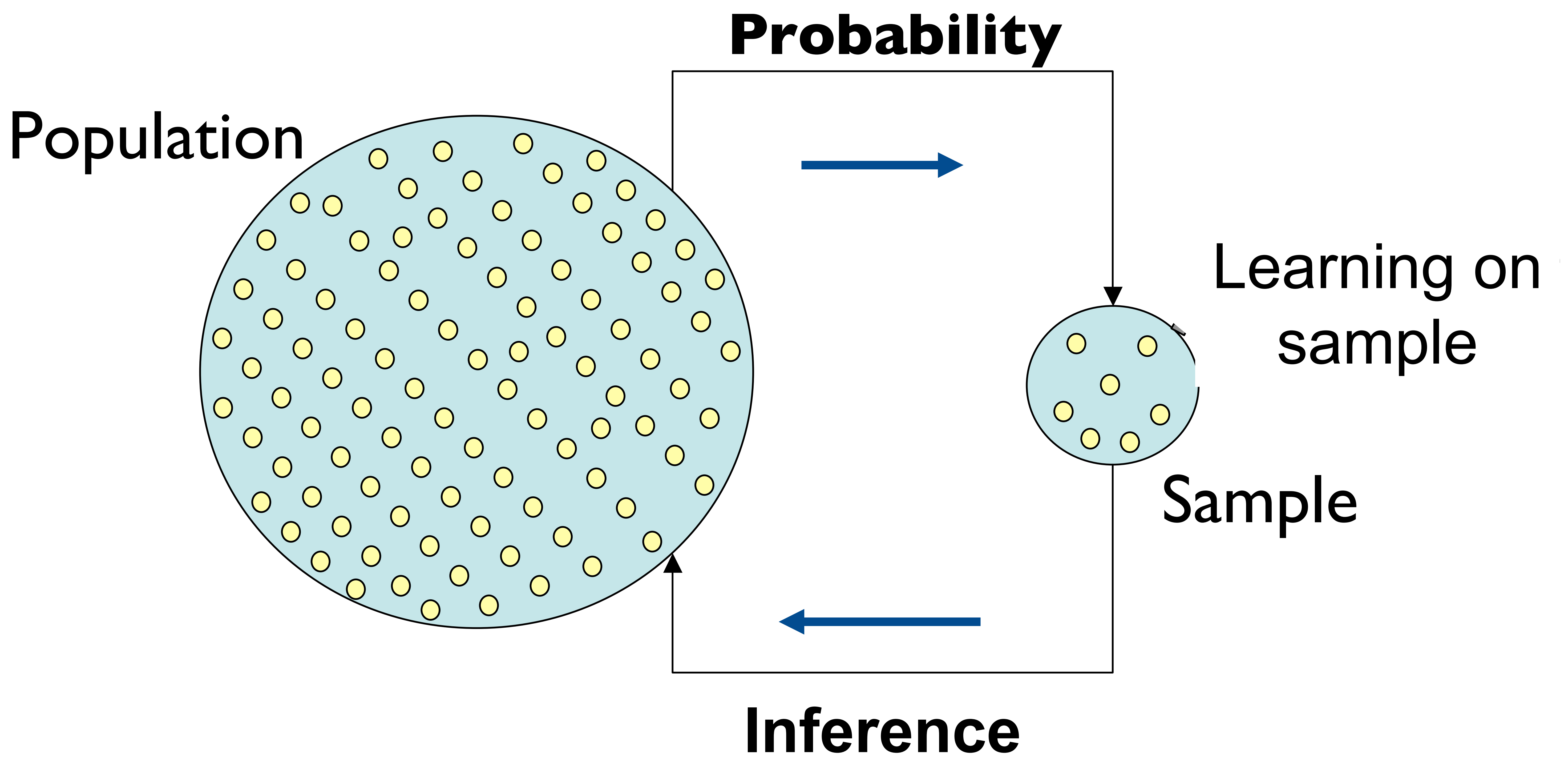
TRIAD OF ALGORITHMS, DATA AND TRAINING



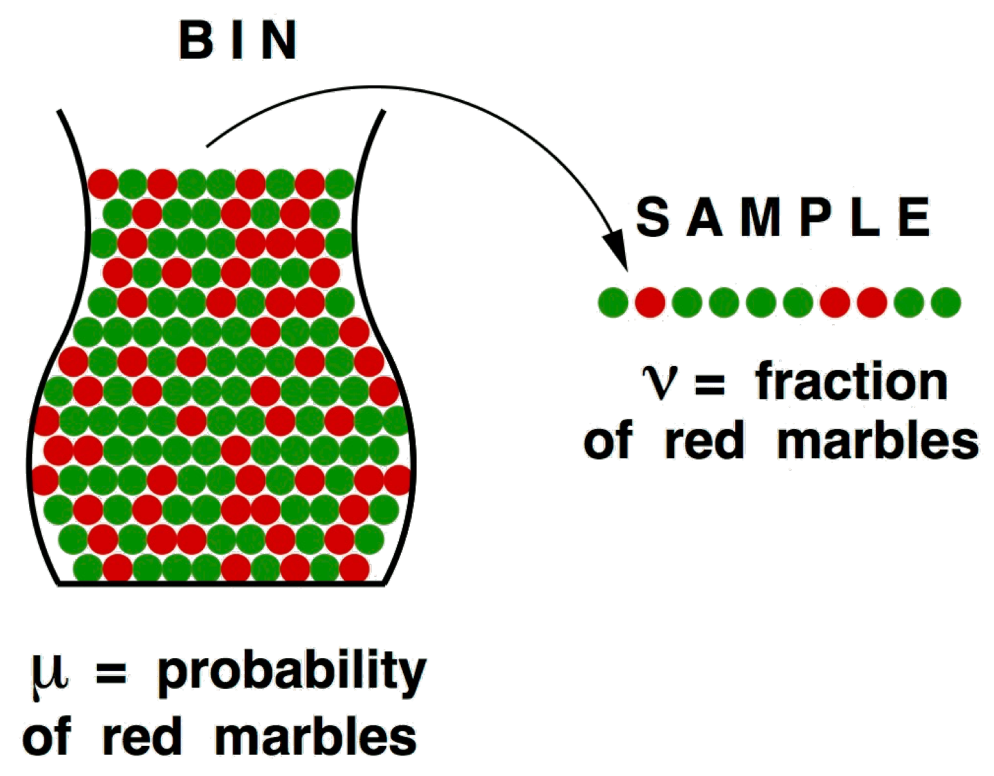
"Learning" is the process of estimating an unknown dependency or structure of a system (building a model) from a limited number of observation (data points) and ability to generalize it onto previously unseen data

THE "CENTRAL DOGMA" OF STATISTICS

Machine learning == statistical learning



Sampling principle



- Sample should be representative of population
- Generalization – extrapolation to entire population
- Watch for population drift!

THREE TYPES OF MACHINE LEARNING

Supervised Learning

The goal is to learn mapping from given inputs X to outputs Y, given a labeled set of input-output (X-Y) pairs
.

X – input data / independent variable
Y – response/ dependent variable

Unsupervised Learning

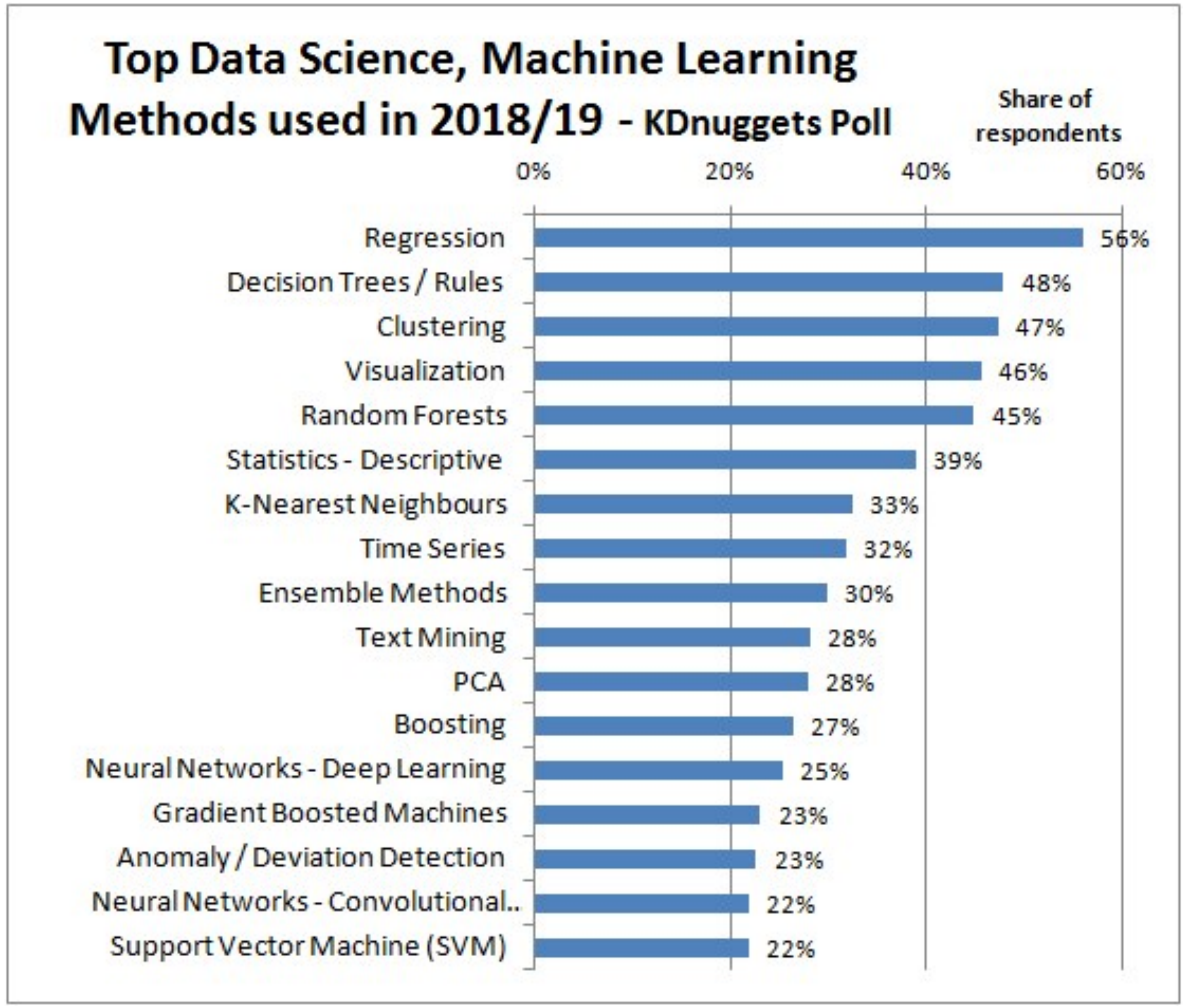
The goal is to learn patterns and structure in data given only inputs X. (no output Y information given at all)

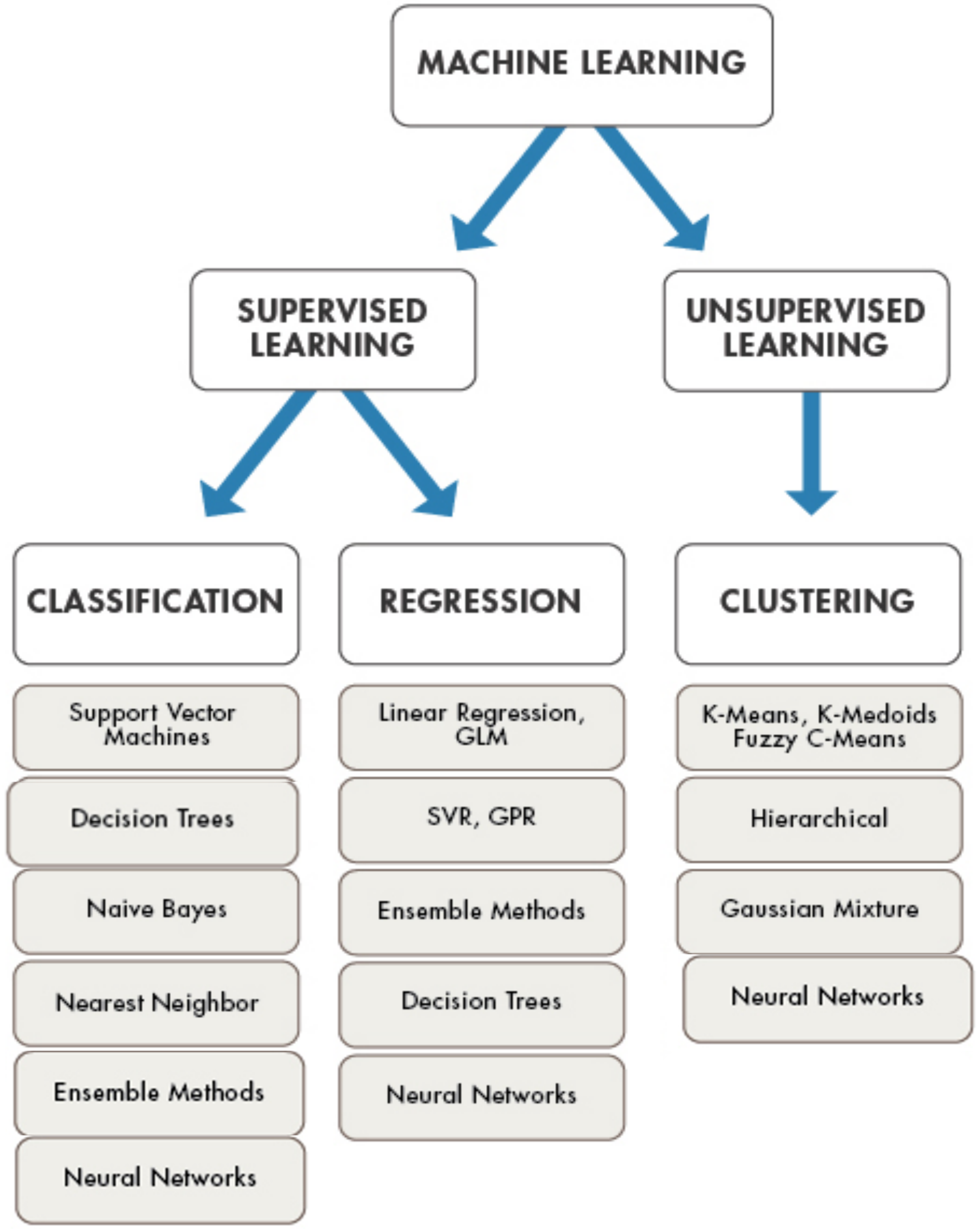
X – input data /independent variable

Reinforcement Learning

The goal is to optimise actions in a way that maximises cumulative reward. no explicitly labeled data is given, but “rewards” and “punishment” signals are provided

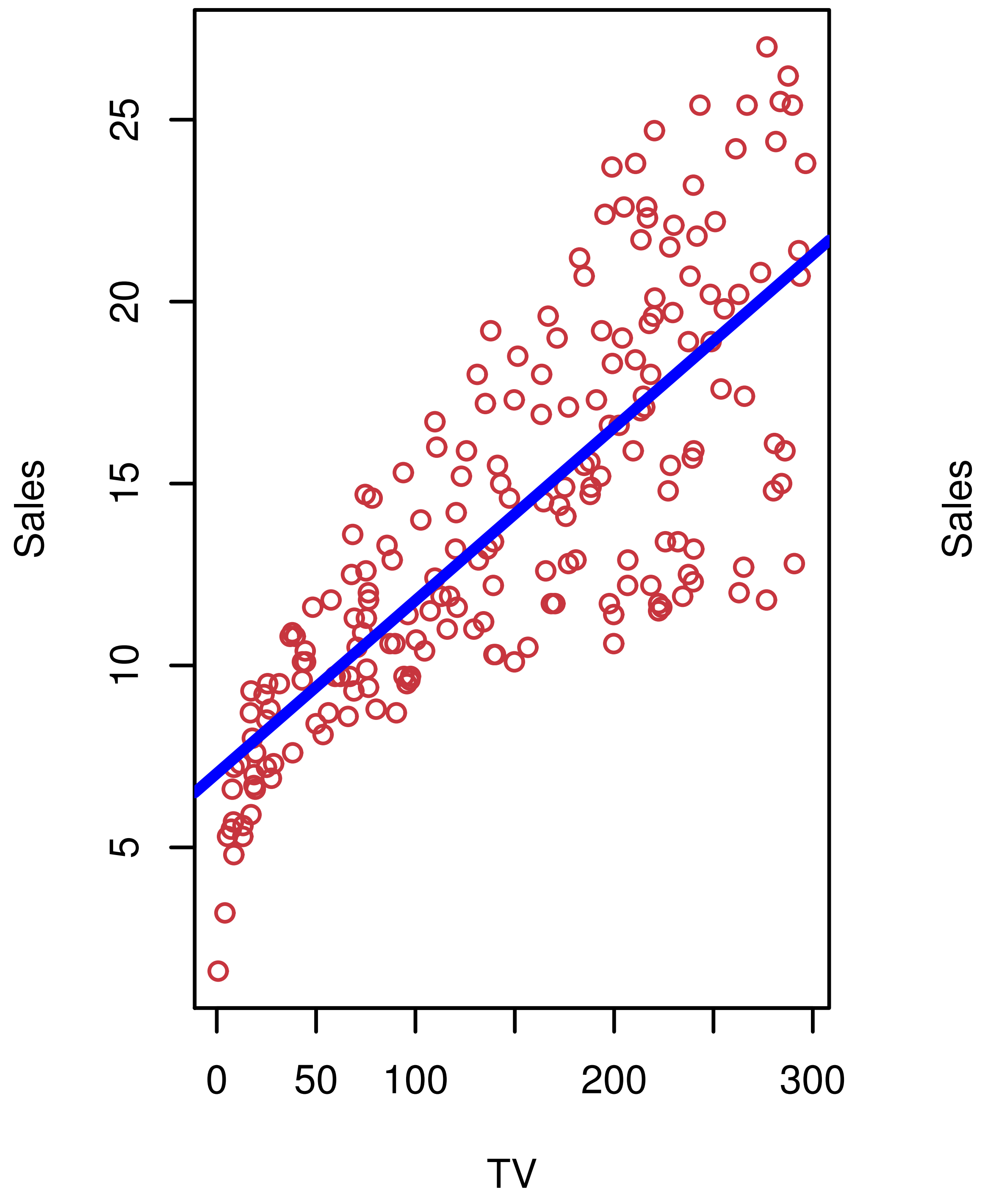
MACHINE LEARNING METHODS



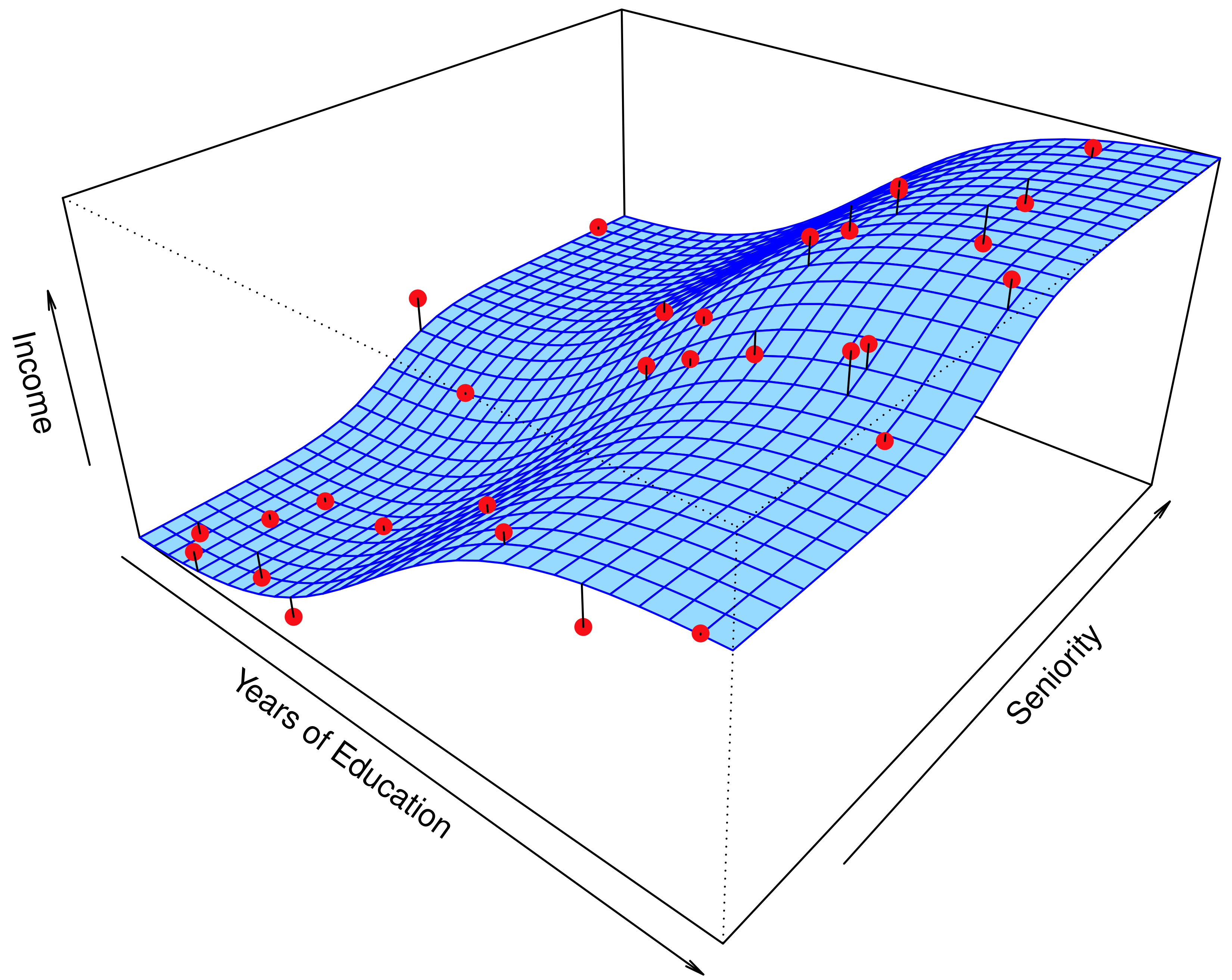


SUPERVISED LEARNING: REGRESSION

Response variable Y – real valued



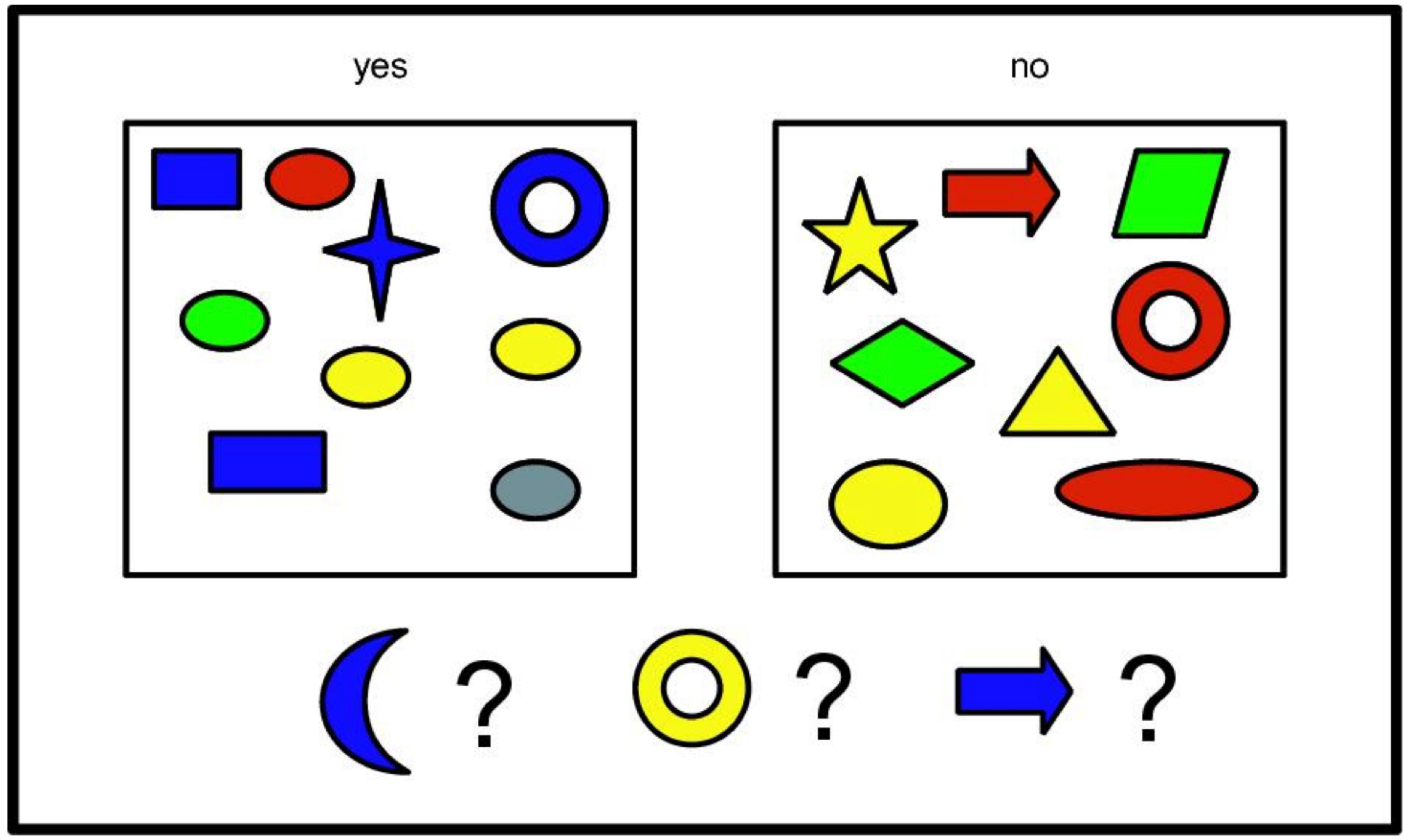
univariate



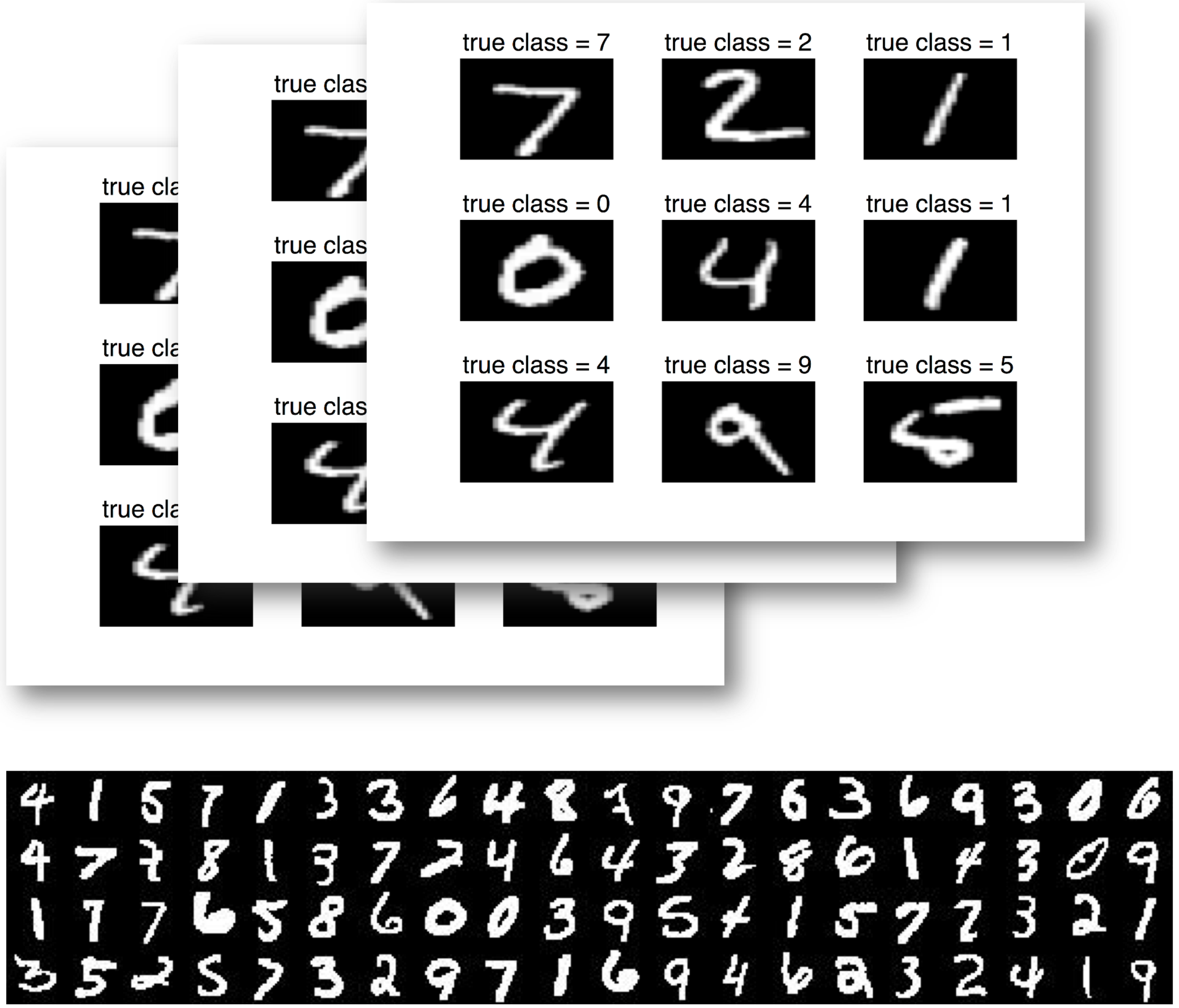
multivariate

SUPERVISED LEARNING: CLASSIFICATION

Response variable Y – categorical

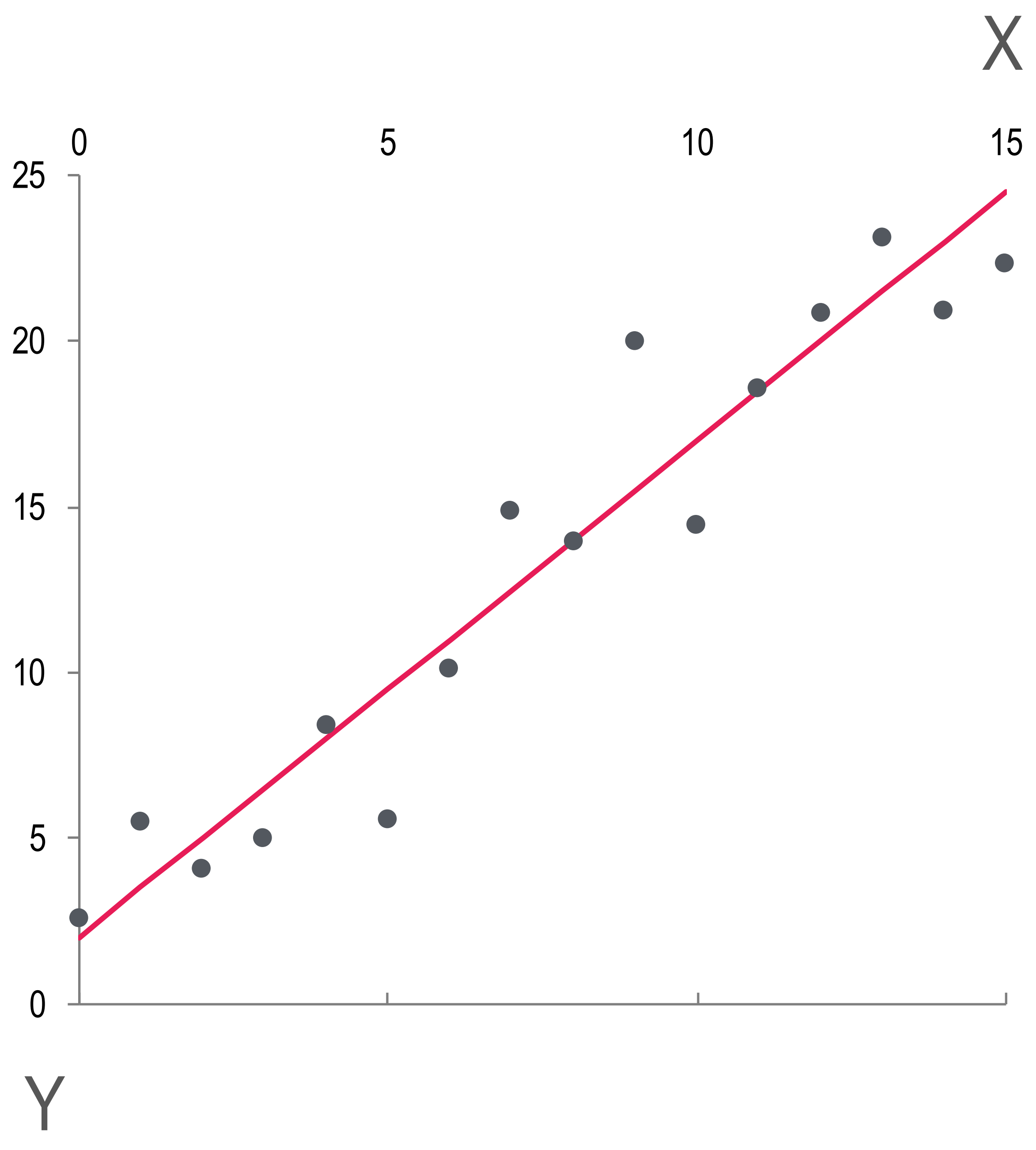


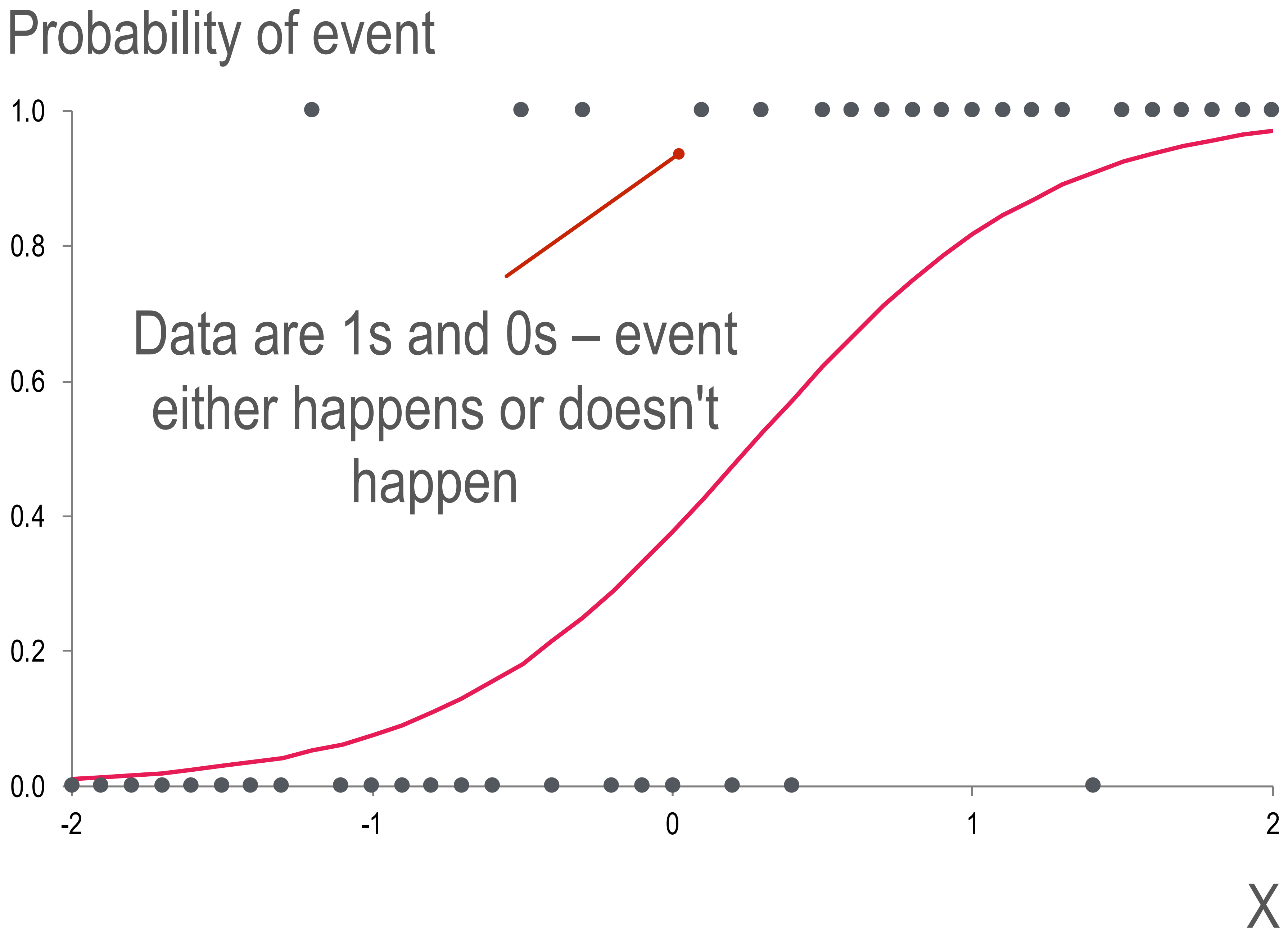
binary



multiclass

REGRESSION AND CLASSIFICATION ARE SIMILAR





Regression

Predict a numeric variable

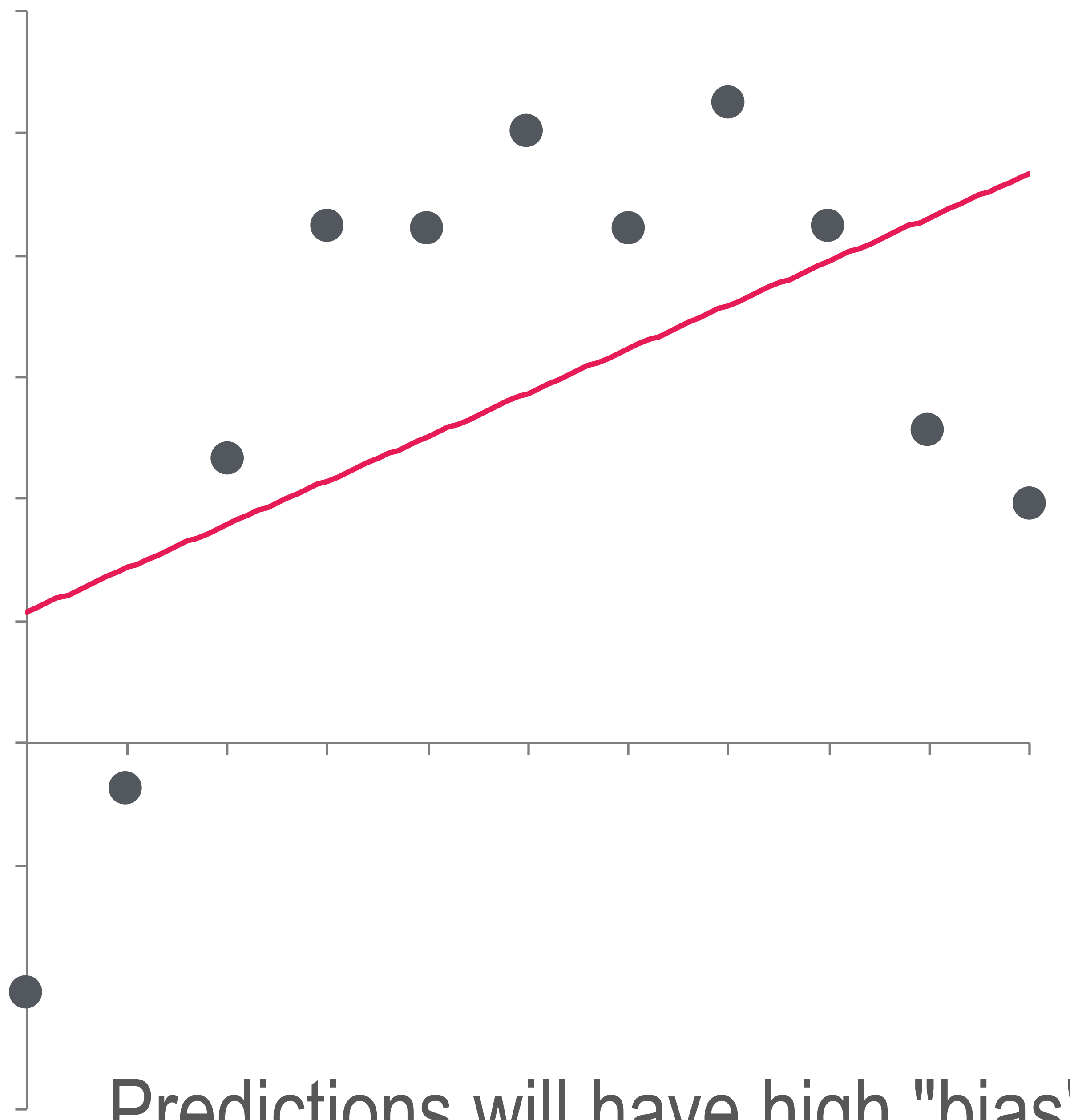
Classification

Predict a binary (or categorical) outcome

MODEL OVERFITTING

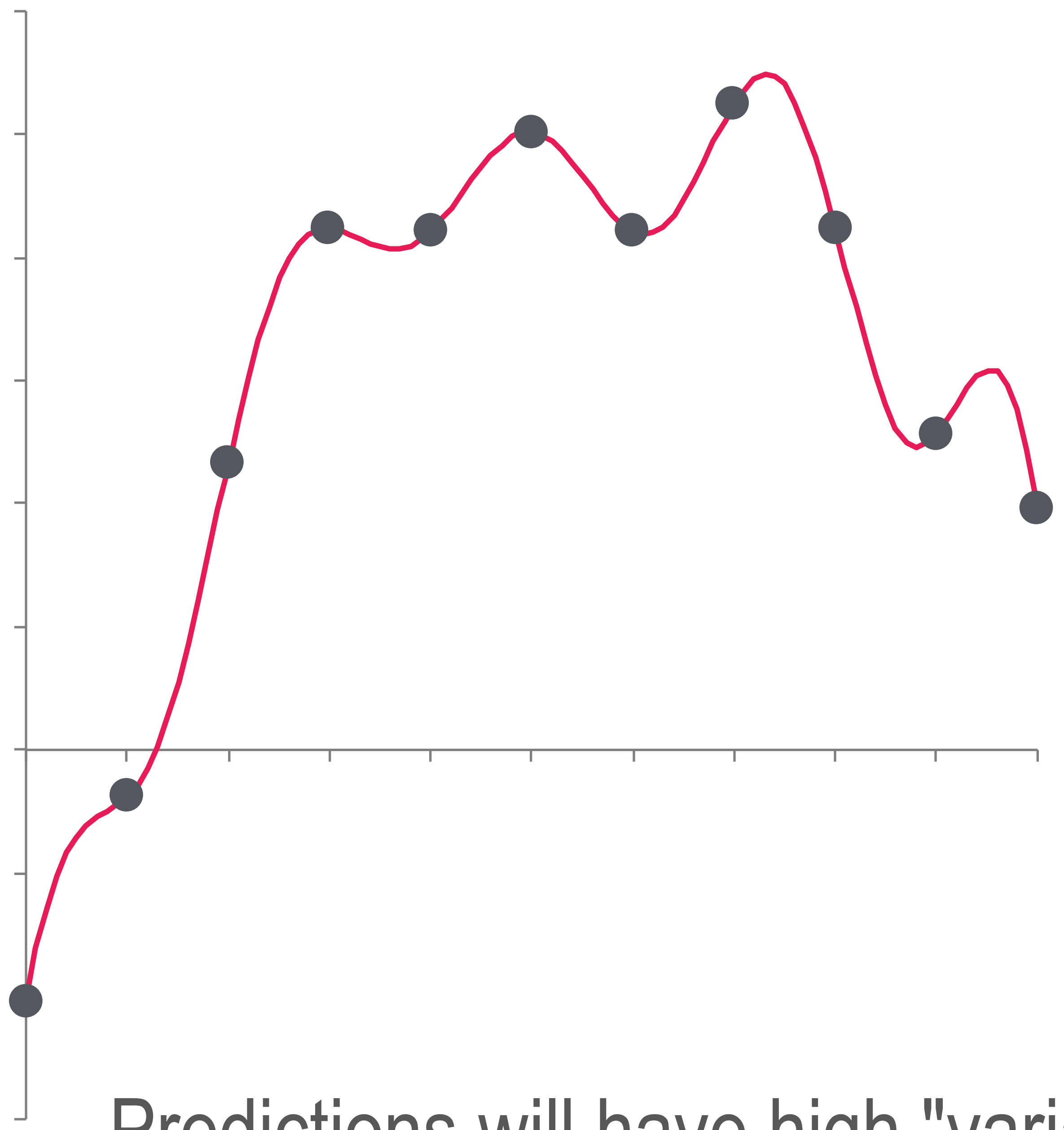
Regression

Too simple



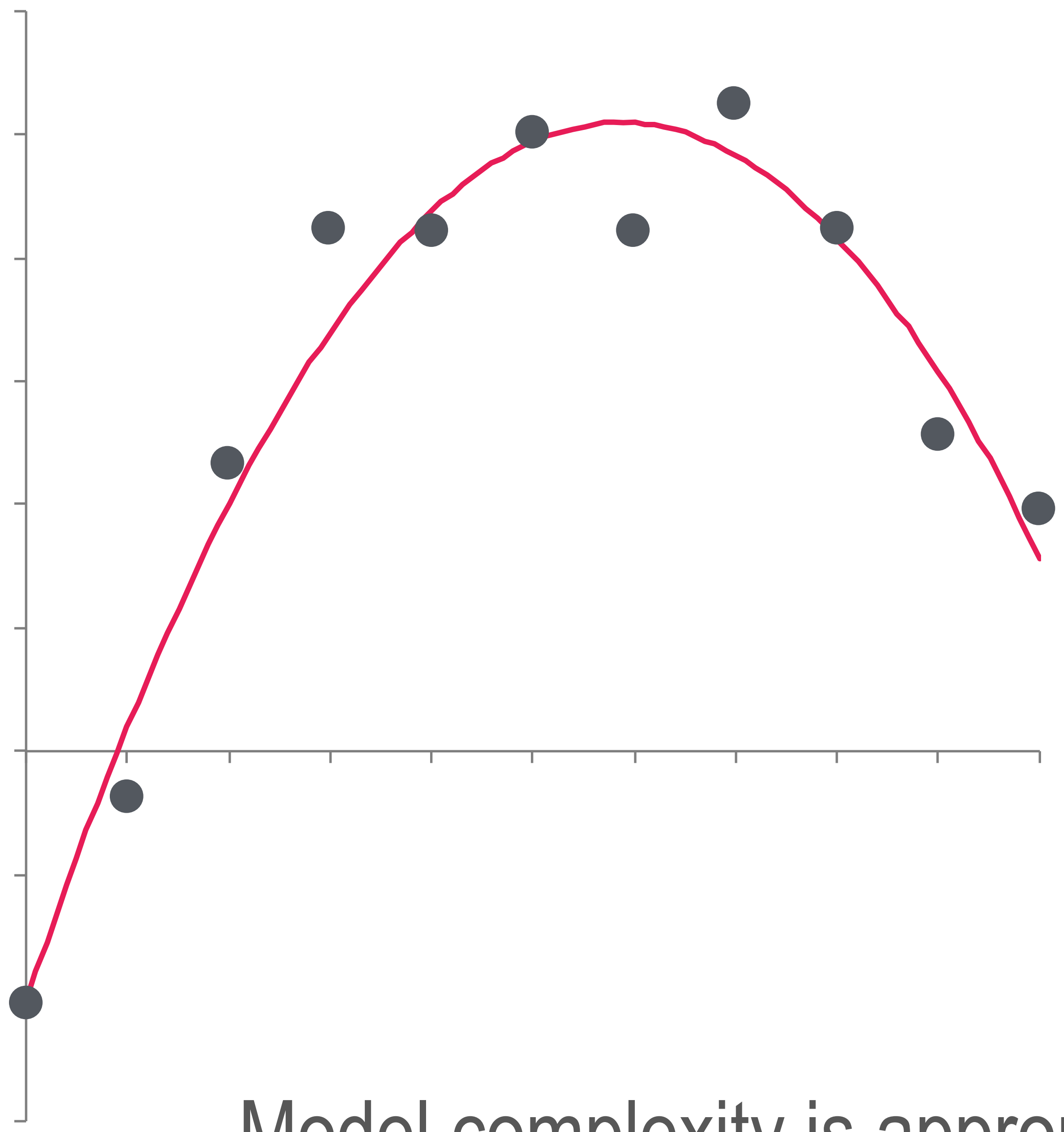
Predictions will have high "bias" – from inadequate assumptions

Too complex



Predictions will have high "variance" – driven by noise in the training data

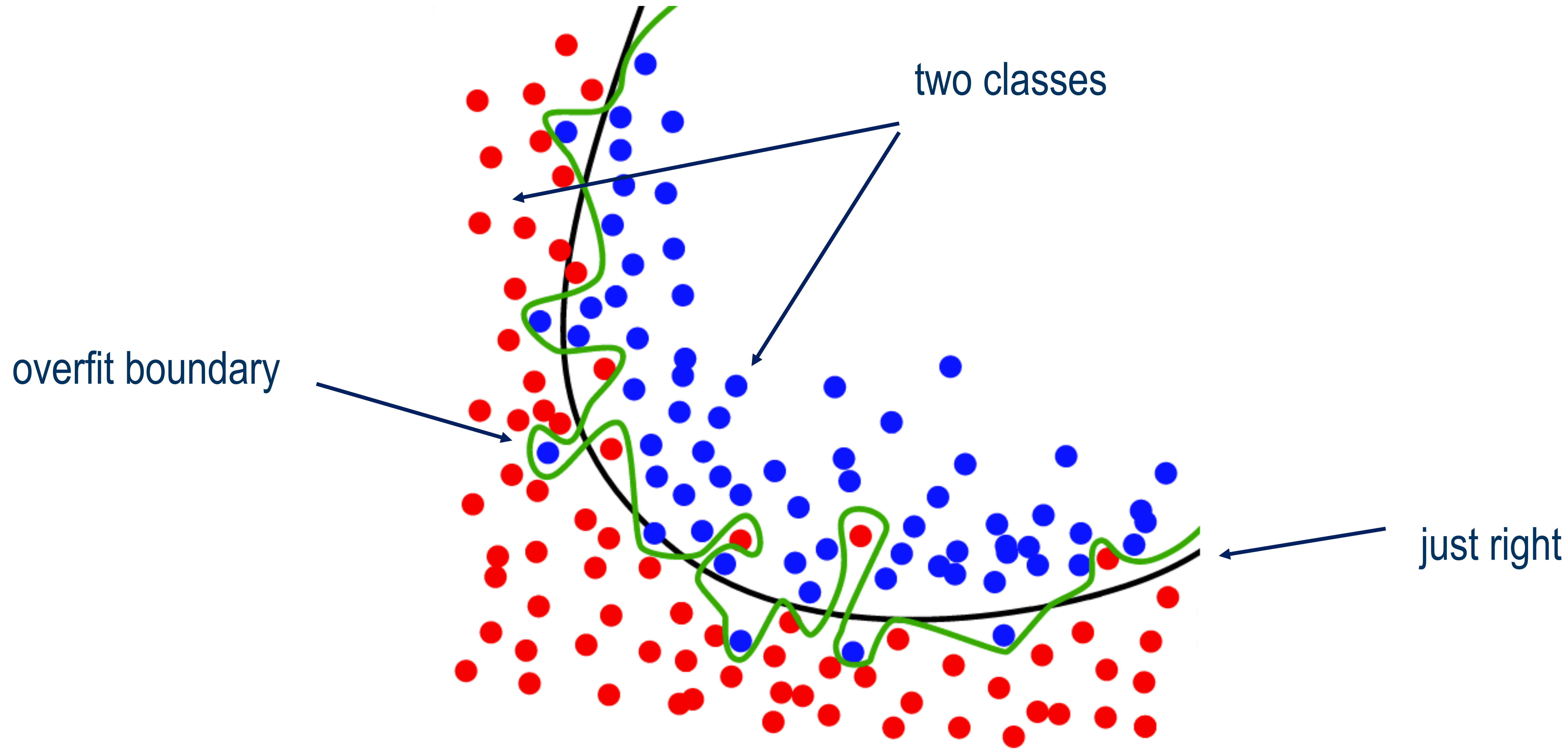
Just right



Model complexity is appropriate given the noise

MODEL OVERFITTING

Classification



PREDICTION ACCURACY VS EXPLAINABILITY

Model explainability

Prediction accuracy

White box models

- Interpretable by design
- Easy to explain
- Quick to run
- Limited tuning needed

Model properties

Black box models

- Lots of work to get insights Better predictive performance
- Potential for overfitting
- Often lot of tuning required

Algorithm examples

White box models:

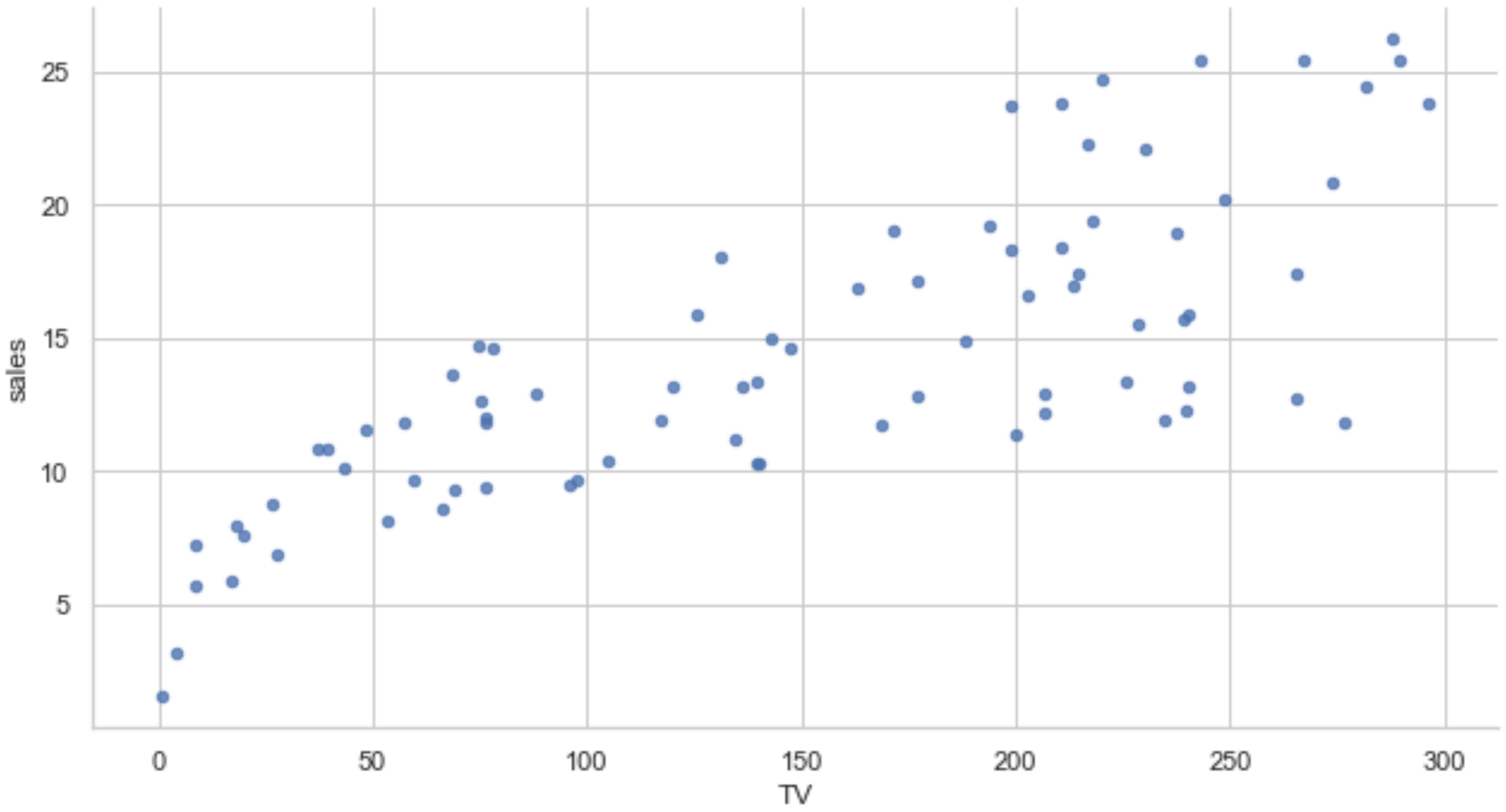
- Linear / logistic regression
- Decision trees

Black box models:

- Random forests
- Gradient boosting
- Neural networks
- Deep learning

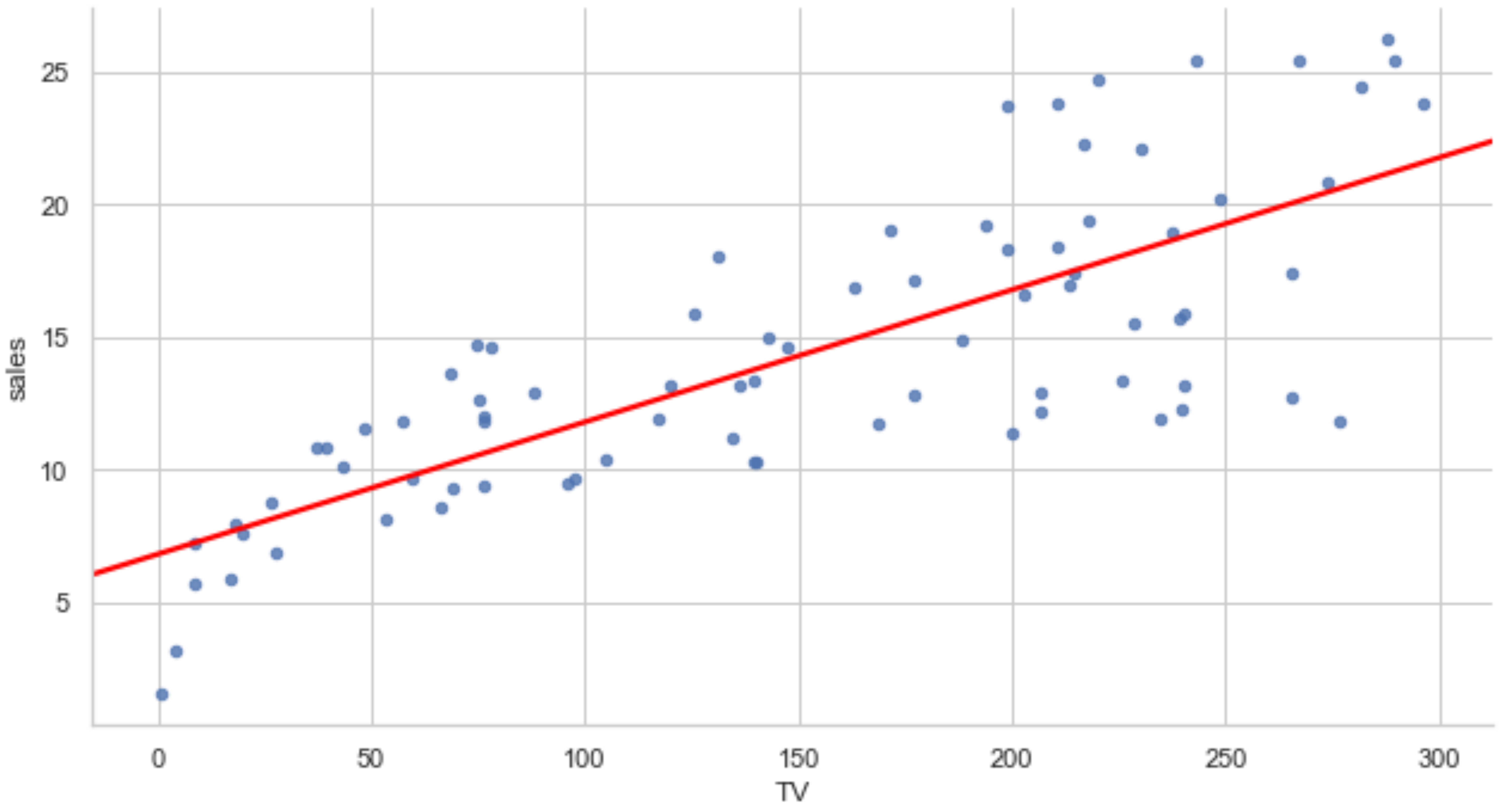
REGRESSION

Modeling



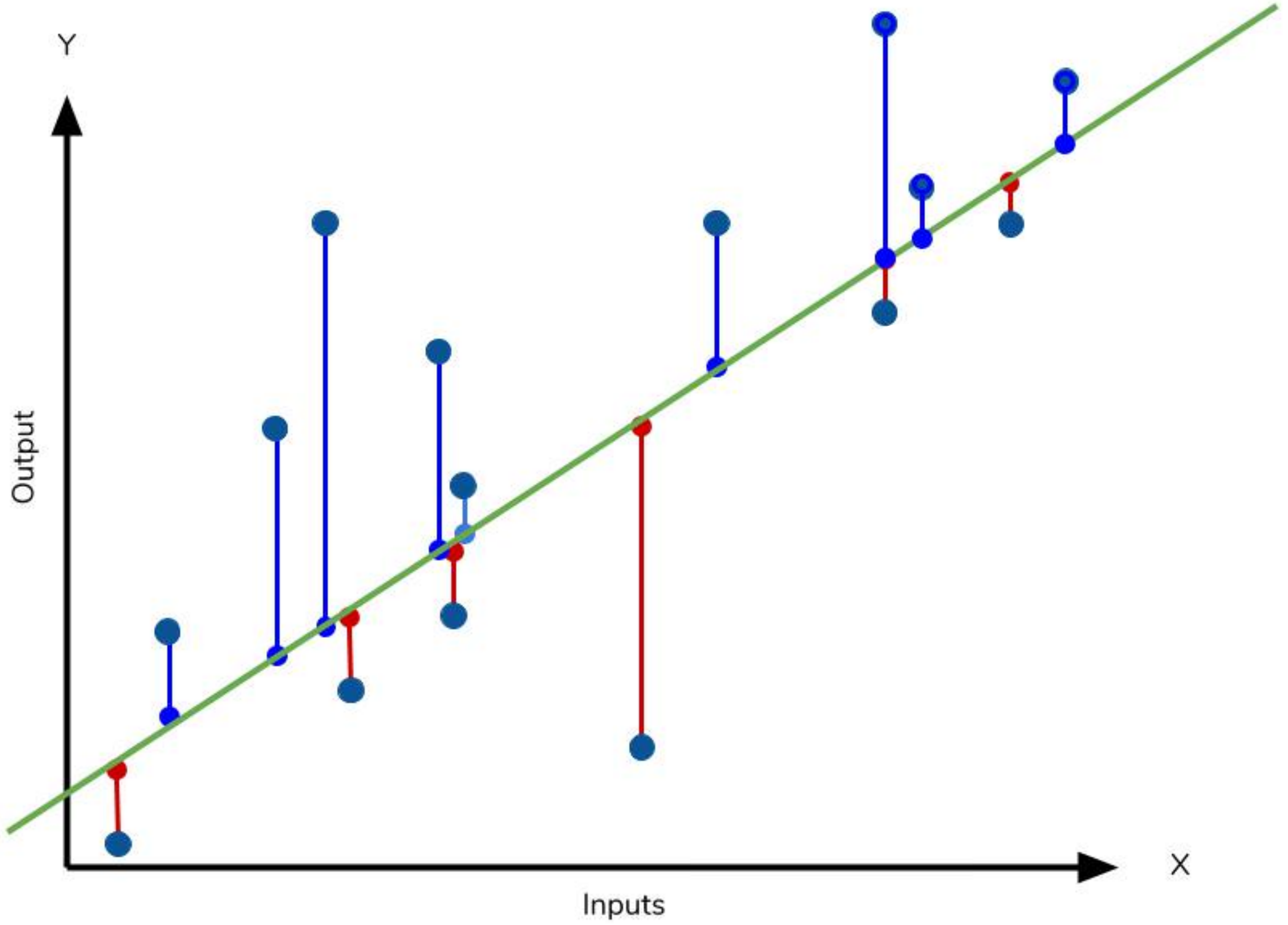
REGRESSION

Quality metrics



REGRESSION EVALUATION

Quality metrics



Standard quality metrics

Mean absolute error: $MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$

Mean squared error: $MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$

Root mean squared error: $RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$

R-squared: $R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$

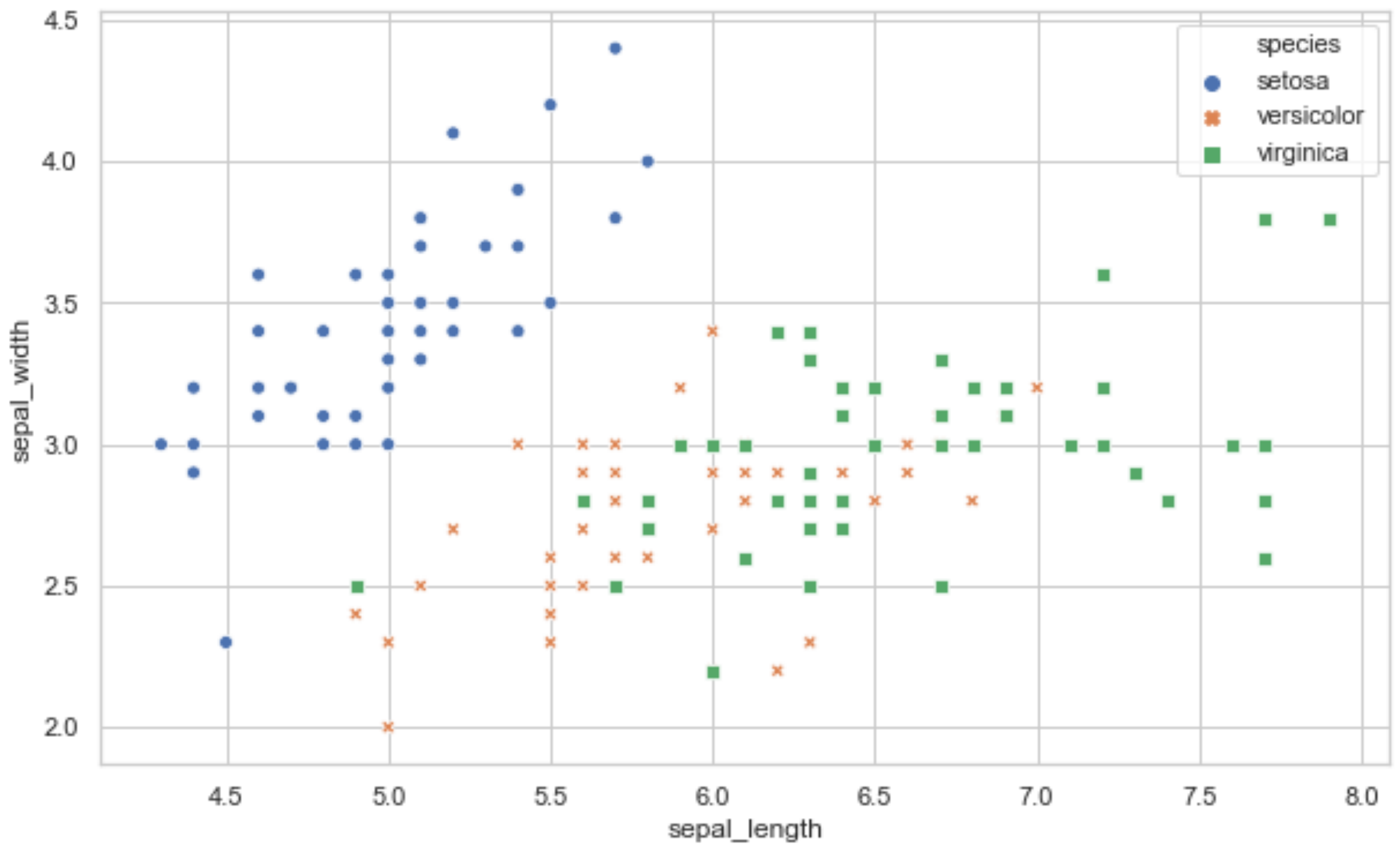
Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

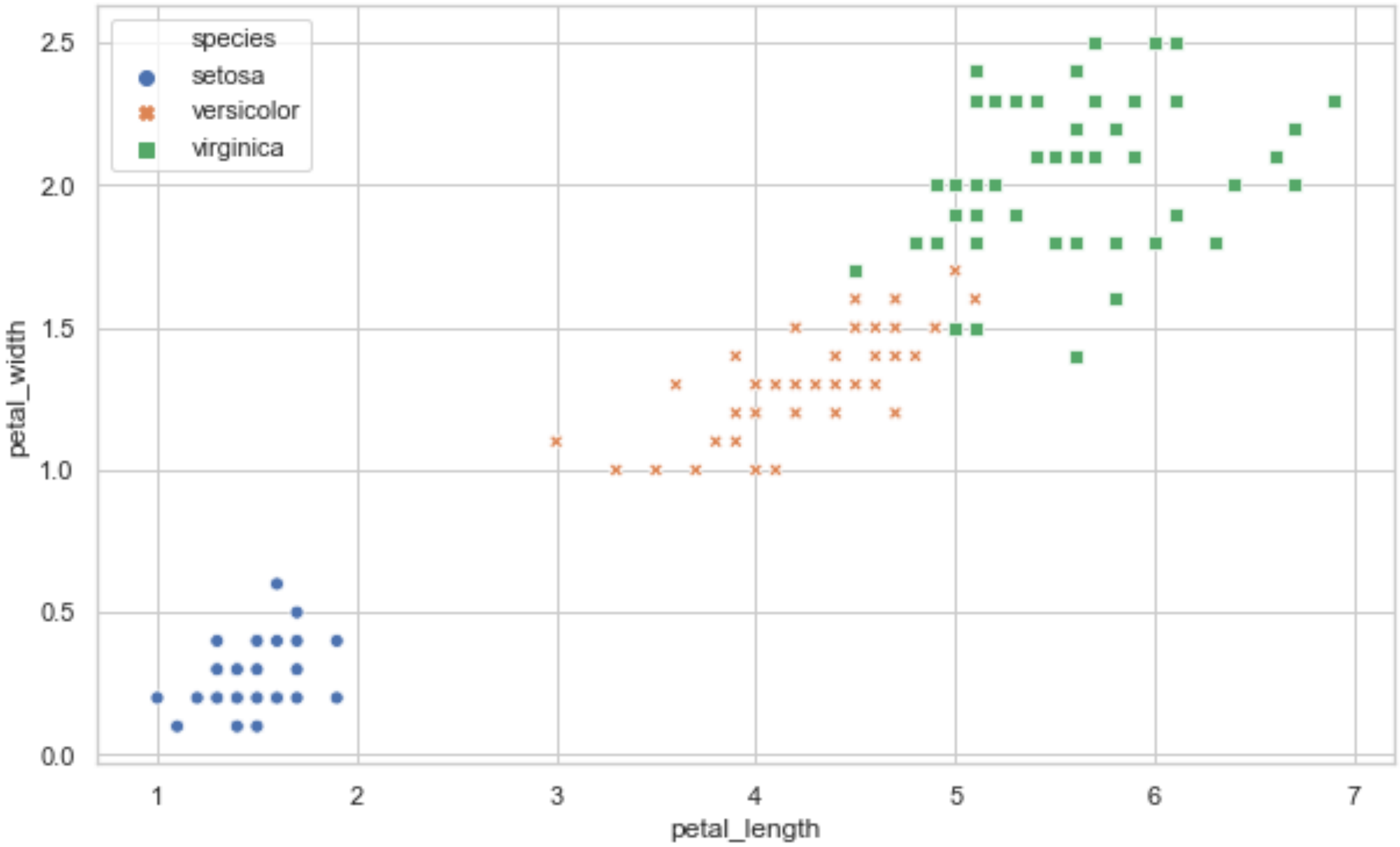
CLASSIFICATION

Classification



CLASSIFICATION

Classification



CLASSIFICATION EVALUATION

Quality metrics

Predicted \ Actual	Yes (or 1)	No (or 0)
Yes (or 1)	True positives TP	False Positives FP
No (or 0)	False Negatives FN	True negatives TN

True positive = Predict event and event happens

True negative = Predict event does not happen, nothing happens

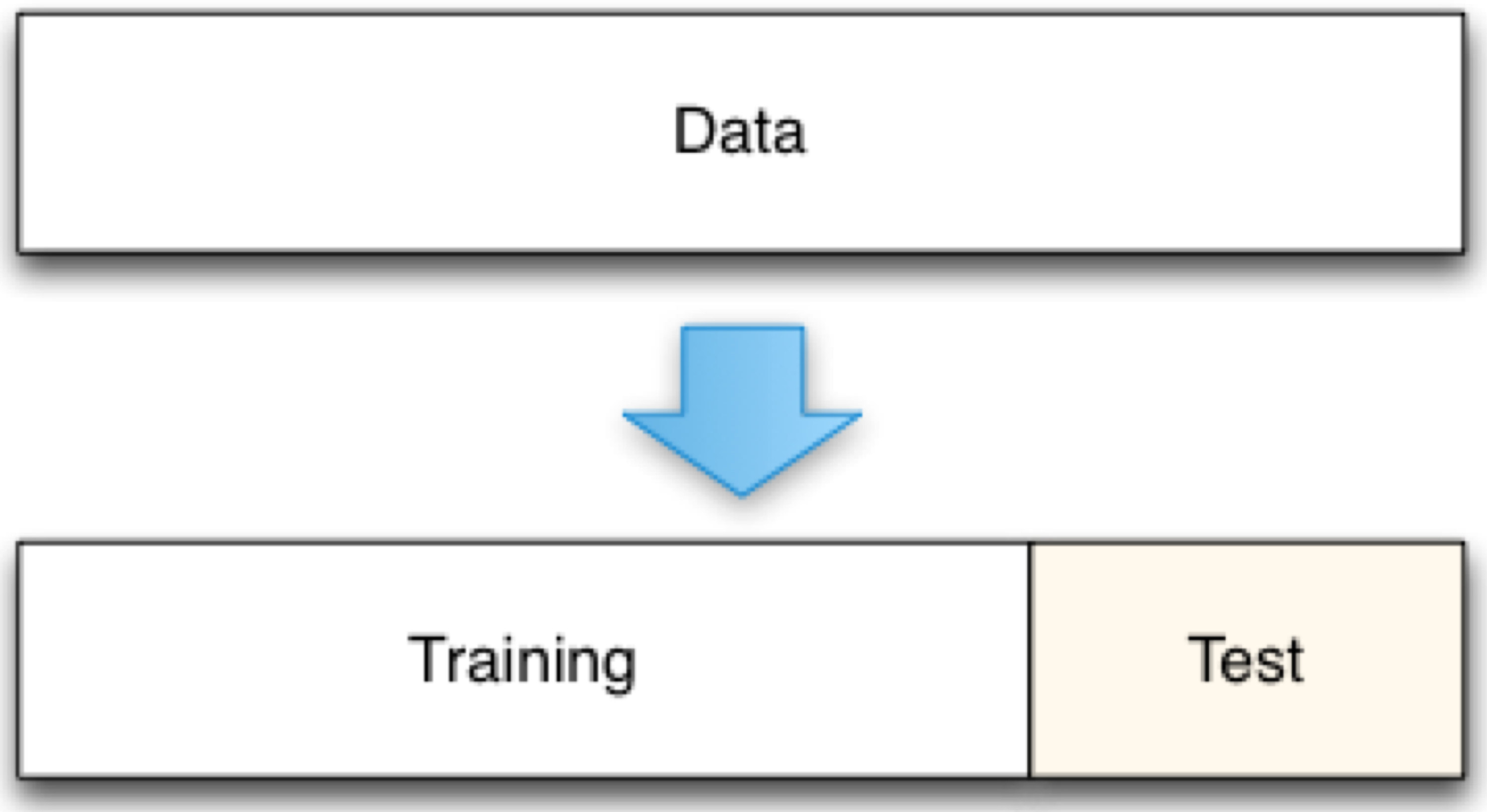
False positive = Predict event and event does not happen (false alarm)

False negative = Fail to predict event that does happen (missed alarm)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

TRAINING AND TESTING

Train-test split



- 70%-90% of the data
- Used to build the model

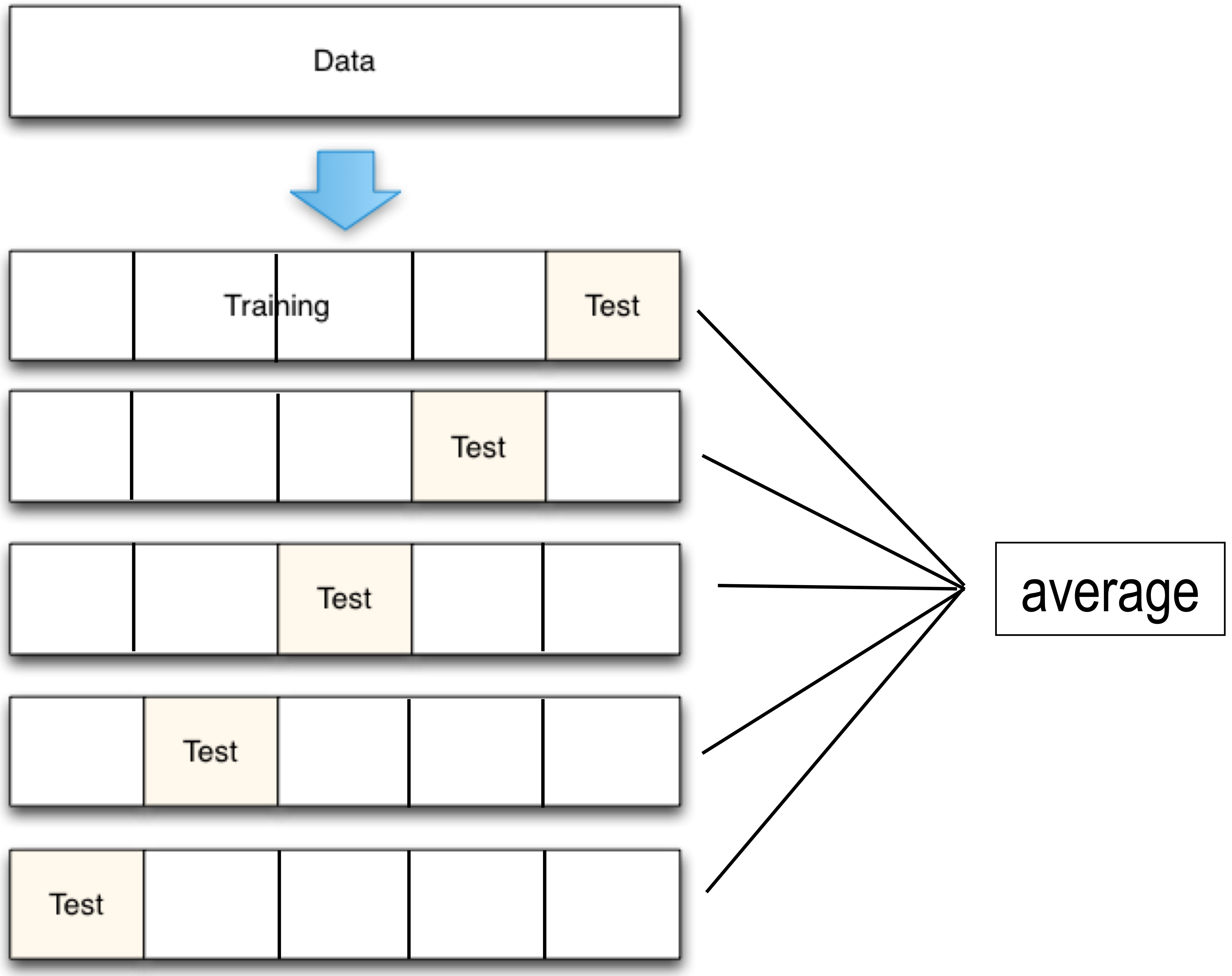
- 10%-30% of the data
- Used to check the performance of the model on unseen data

Train & Test split

- Measure algorithm performance on both train and test sets!
- Performance will be worse on the test set
- Algorithms hyperparameter tuning can be used to improve test set performance
- Avoid overfitting!
- Actual performance of the algorithm in production will not be better than on test set!

TRAINING AND TESTING

Cross-validation



5-fold cross-validation

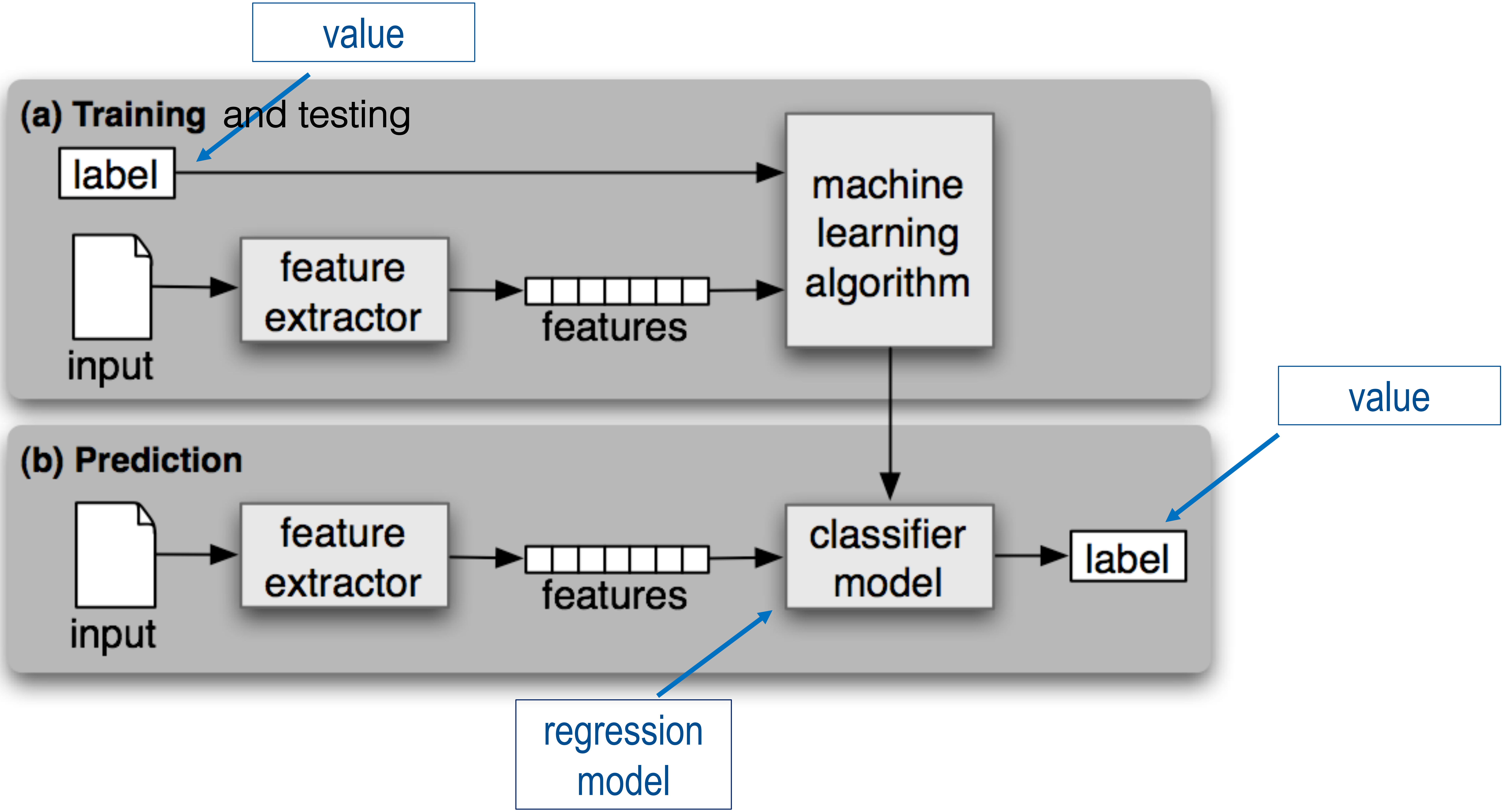
Cross-validation

- Makes best use of the data
- Data split in to N "folds" at random
- N models built. On each model, N-1 folds are used for training and one is used for testing
- Evaluation criteria averaged across folds
- Allows use of eg 90% training data / 10% test data splits for 10-fold cross validation
- More data for training increases predictive power
- Reduces the chance of getting lucky/unlucky just due to the way a single train/test split is done
- More time/computer resources consuming

TYPICAL SUPERVISED LEARNING PIPELINE

Model training

Model application



A SUPERVISED MACHINE LEARNING WORKFLOW

Prepare data

- Define problem and potential solution
- Get the data
- Understand the data
- Clean the data

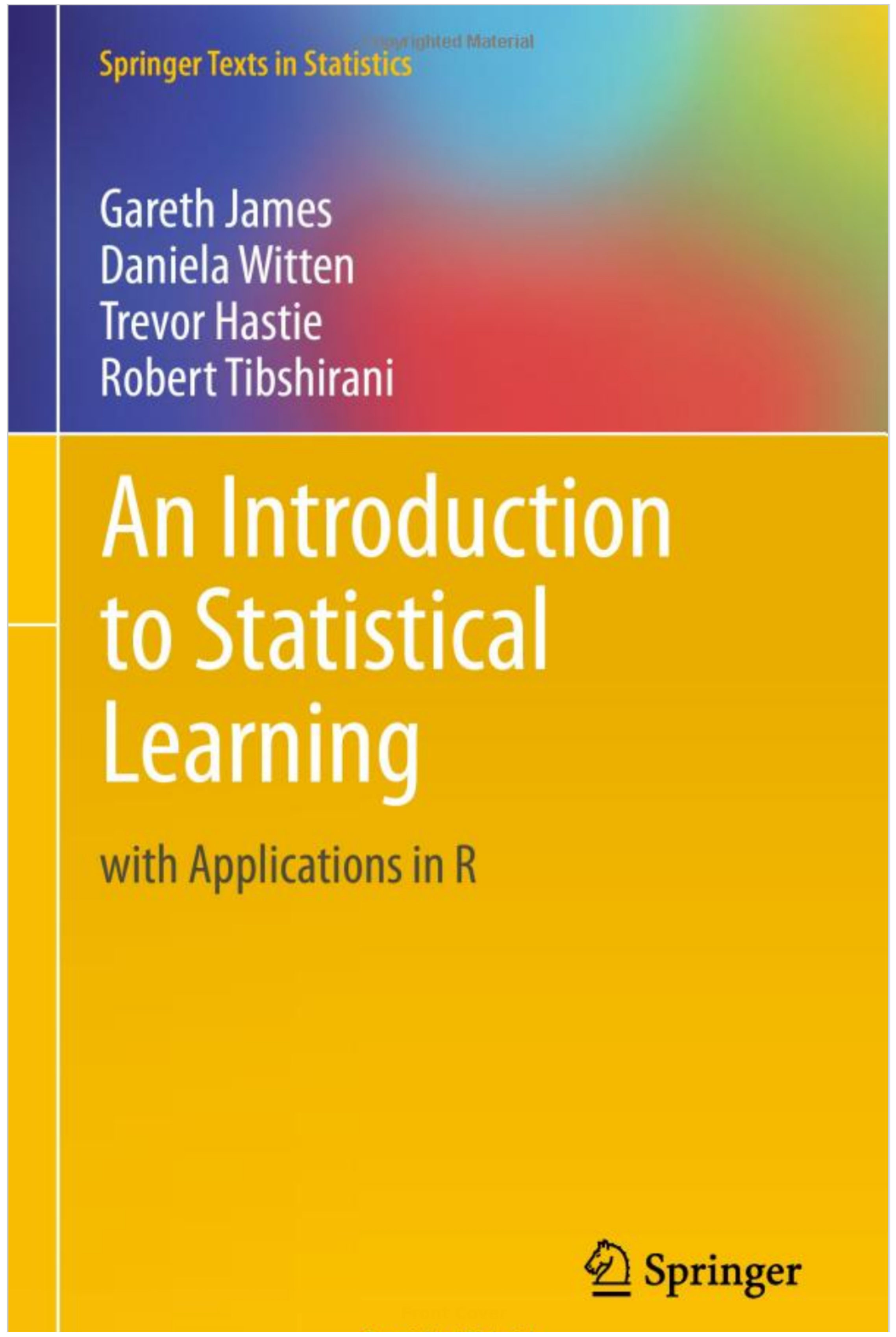
Model and predict

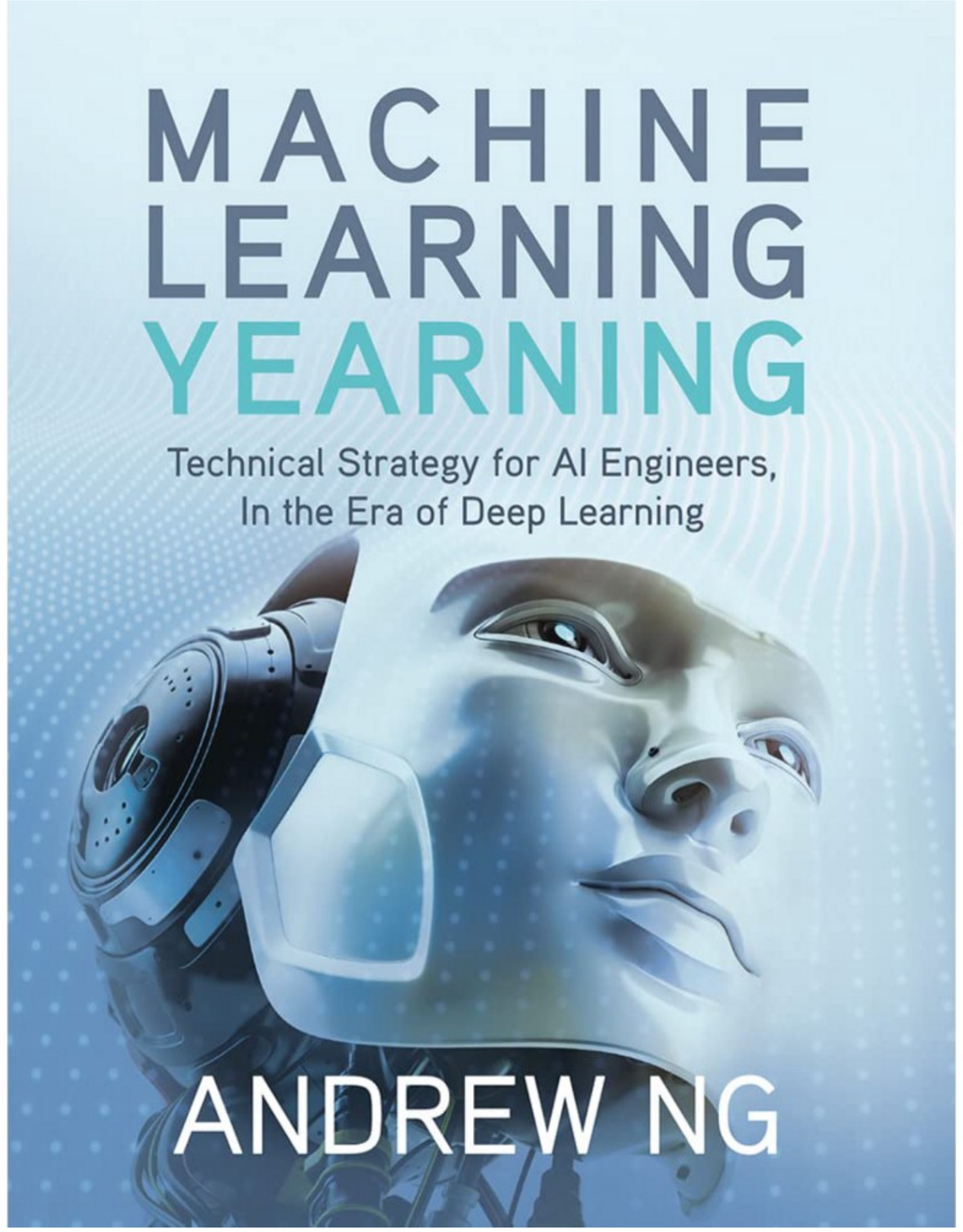
- Feature engineering
- Build and test model
- Understand the model
- Iterate

Impact business

- What does it mean for the business?
- What are we going to change?
- Productionise
- Ongoing monitoring and improvements

A FEW MORE BOOKS





NATIONAL RESEARCH
UNIVERSITY