Predictive Analytics Pt. I

Lab 5

Agenda

- Models and prediction
- Concepts in machine learning

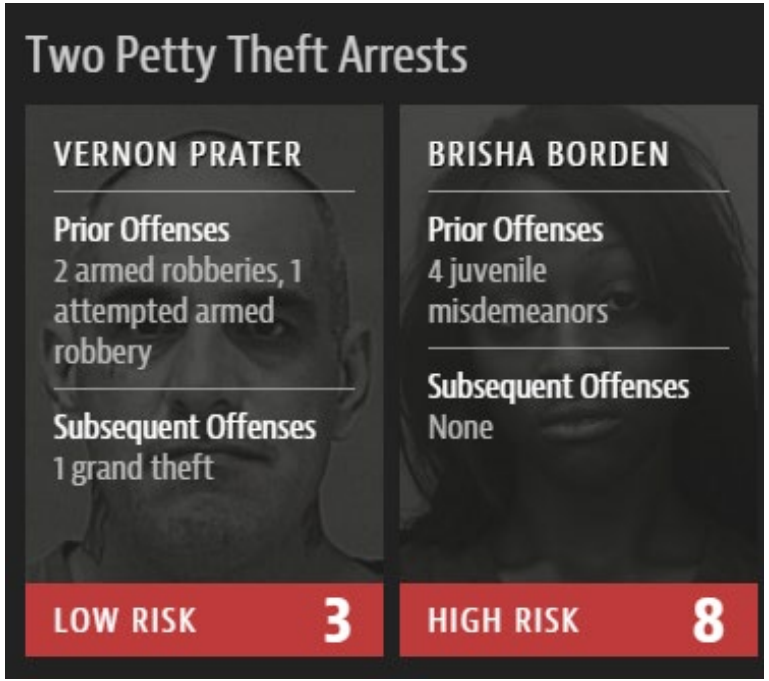
- Random forests, training & predicting

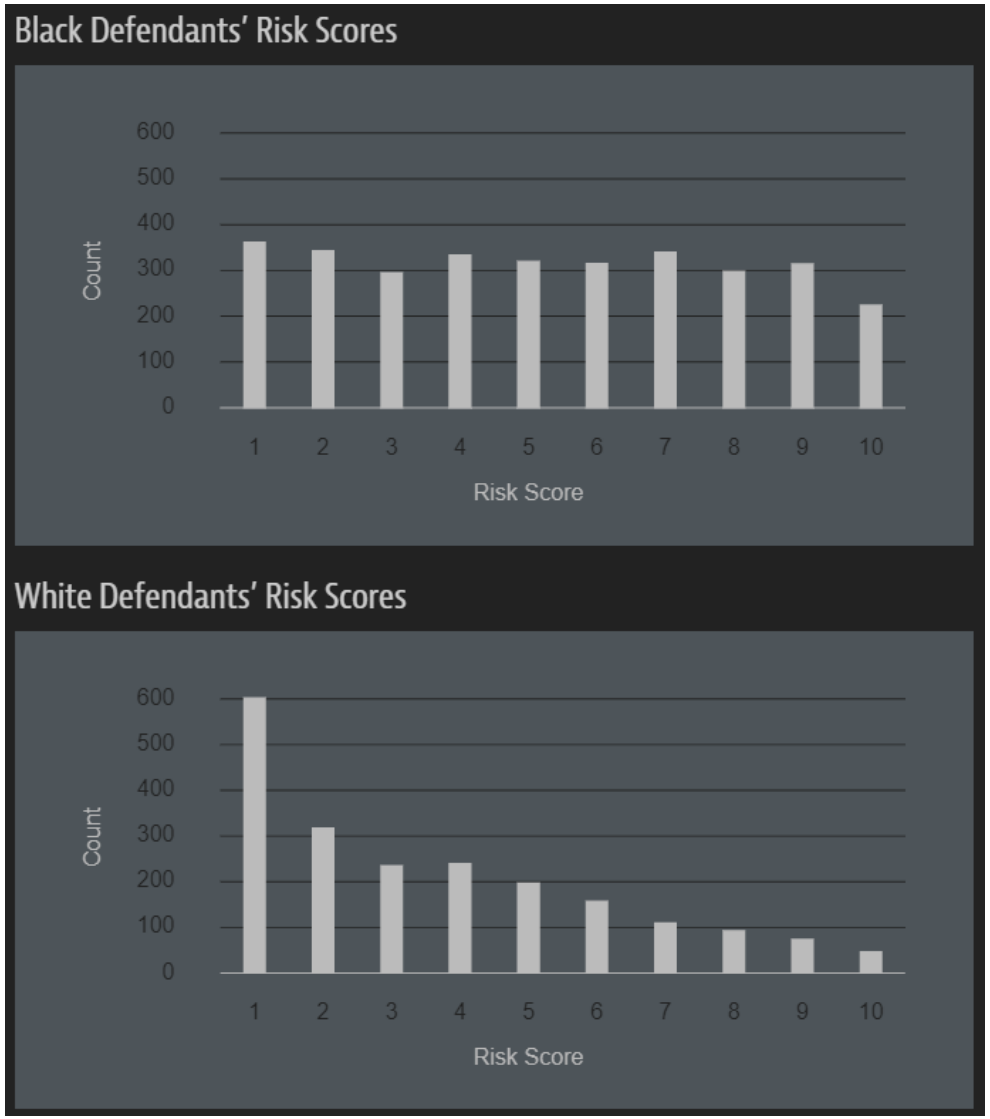
The COMPAS risk assessment tool (model)

137-item questionnaire

- “Have you ever been a gang member?”
- “Were you using drugs or under the influence when arrested?”
- “What was your final grade completed in school?”

New York, Wisconsin, California, Florida





2016 – ProPublica investigation and analysis

- **Risk scores** predicted by COMPAS model, ProPublica analyzed predictions from **risk scores**
- Calculation of risk scores not disclosed
- Correct predictions of recidivism (re-offense) 61% of the time

- Dataset is publicly available

- Analysis notebook with R/Python code

- Much more to the story

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Further reading

Predictive policing in Chicago

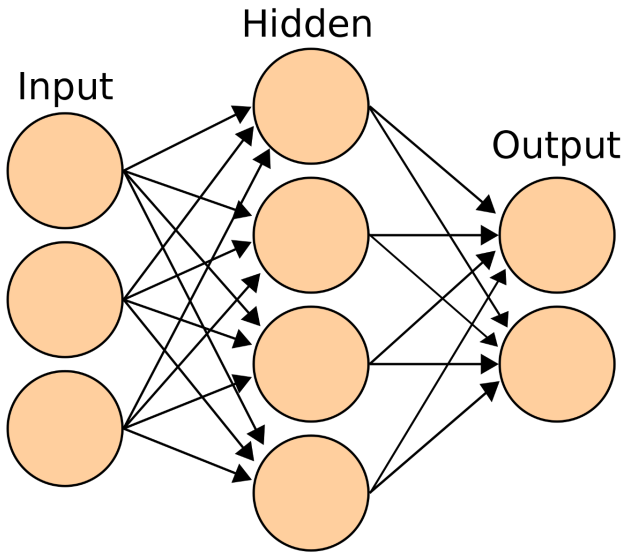
Racial bias in health care decision-making

Bias in predicting child abuse risk

What *is* a model?

"An informative representation of an object, person or system" - Wikipedia

$$Y = f(X, \beta) + \varepsilon$$

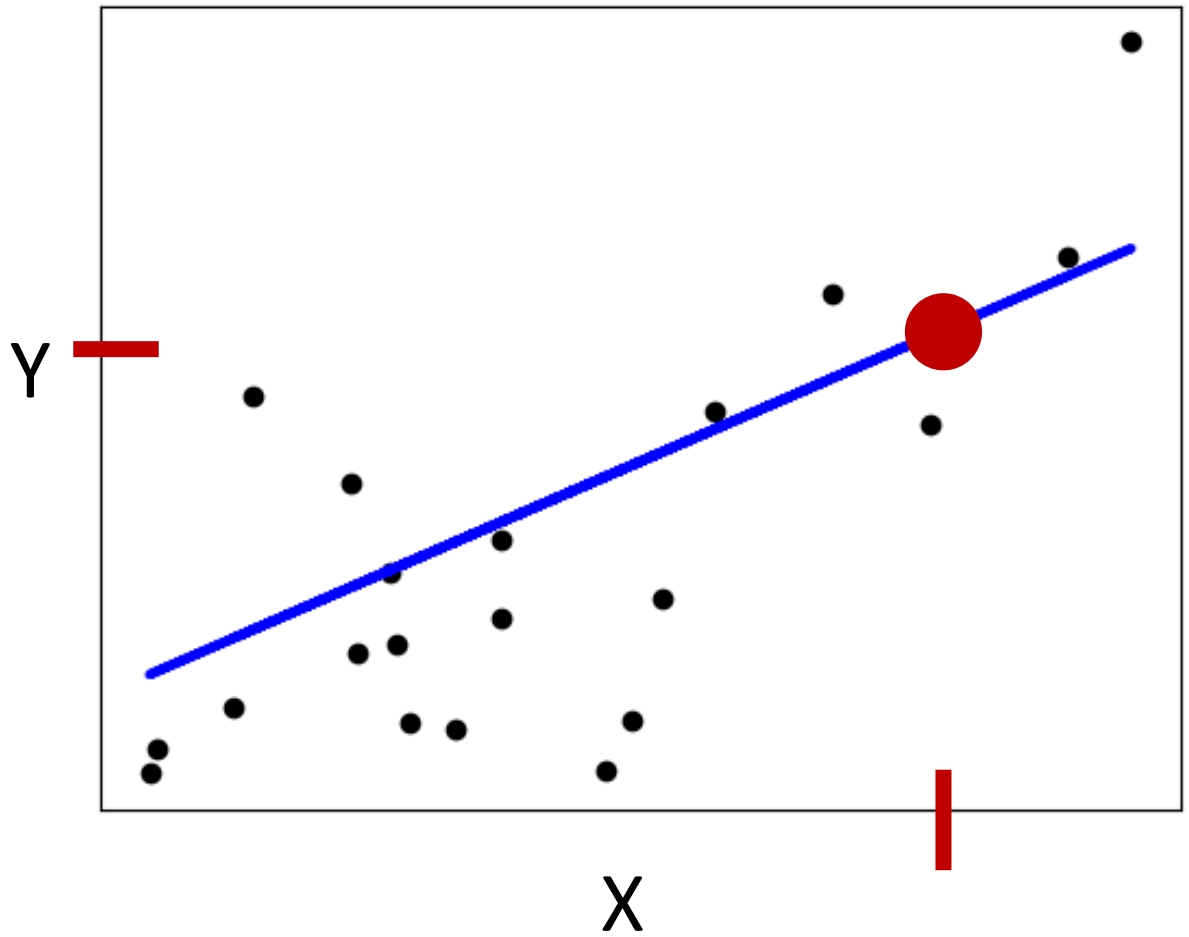


All models require assumptions and decisions

Explanation versus(?) prediction

$$y = \alpha + \beta x + \varepsilon$$

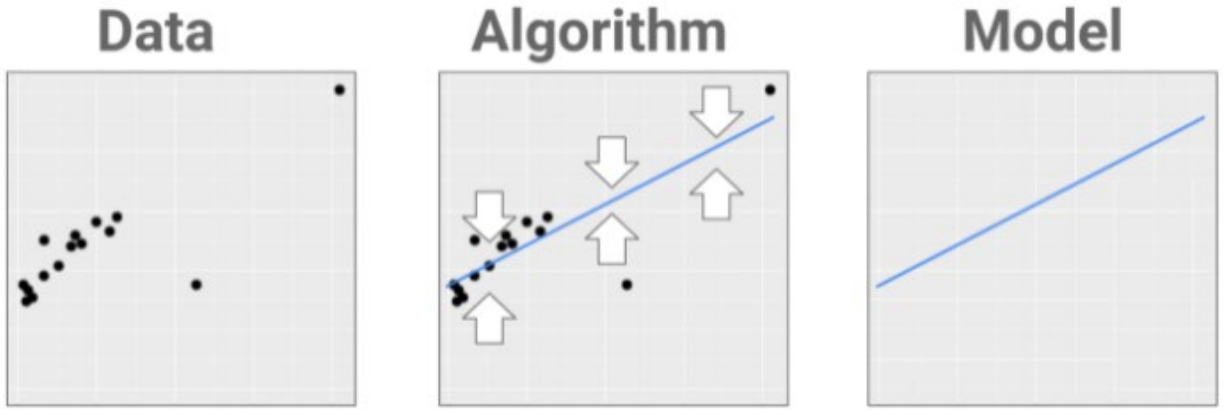
$$y_{new} = \alpha + \beta x_{new}$$



Types of prediction tasks

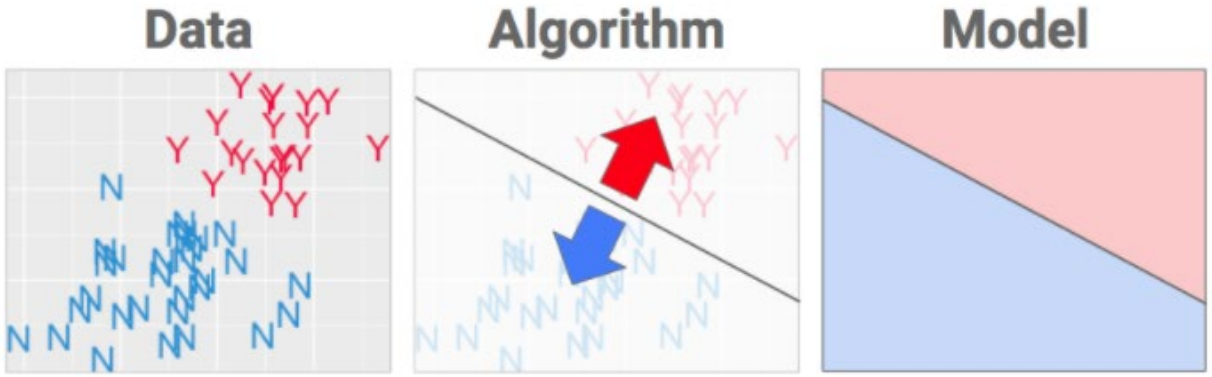
Fitting the data (regression)

- What is this person's risk score?

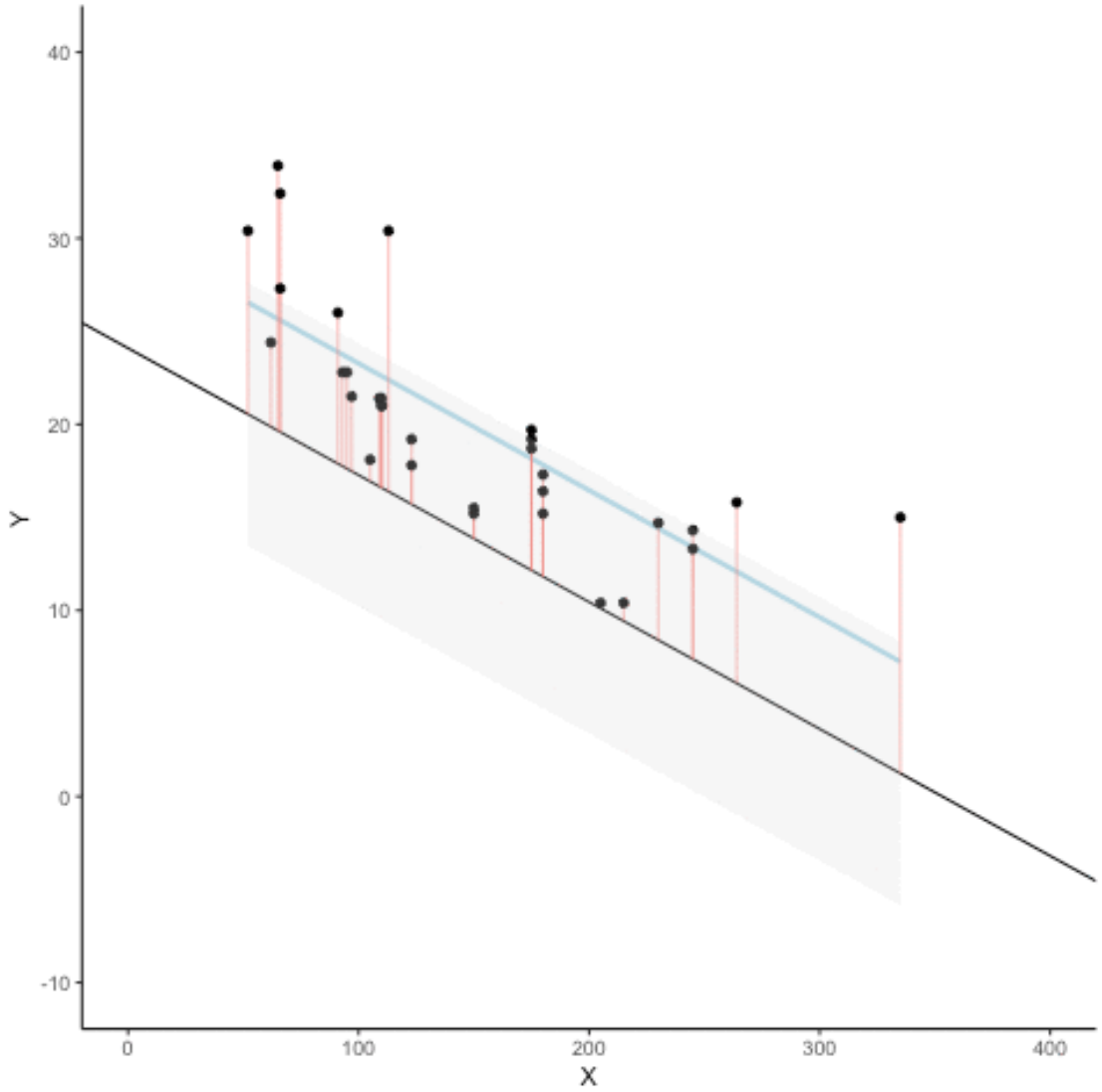


Separating the data (classification)

- What is this person's risk category?



Linear regression



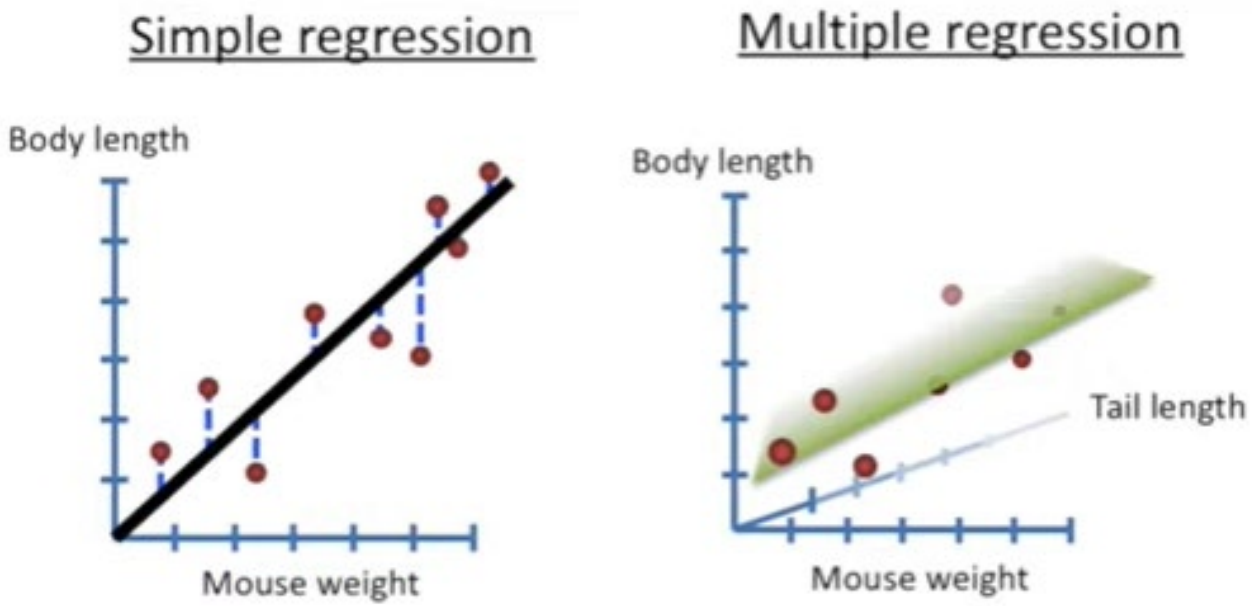
Algorithm:

Find the α and β that minimize residuals

Model:

$$y = \alpha + \beta x + \varepsilon$$

Multiple linear regression



Algorithm:

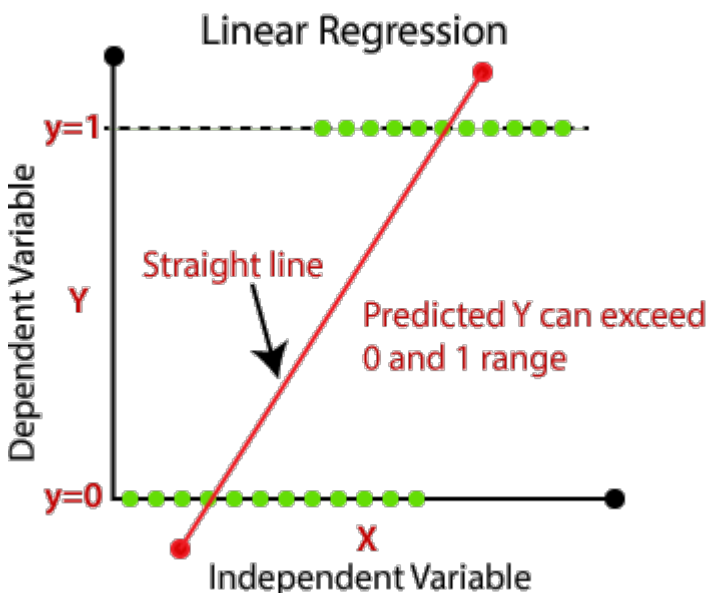
Find the α and $\boldsymbol{\beta}$ that minimize sum of squared residuals

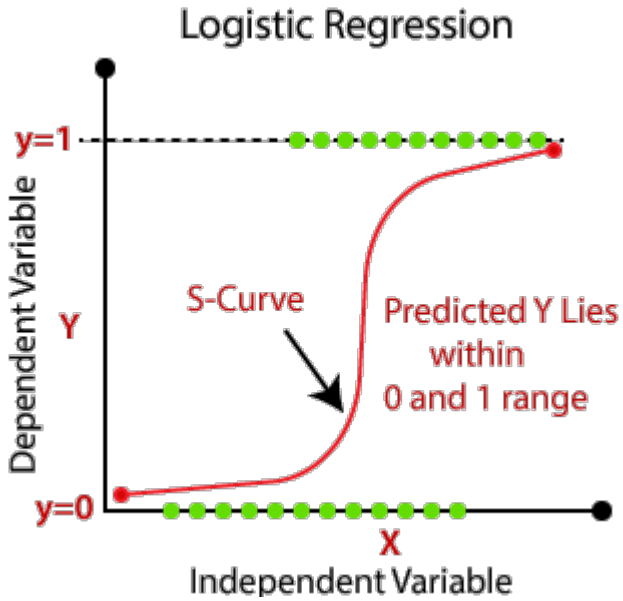
Model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Mouse weight ↑ ↑ Tail length

Logistic regression





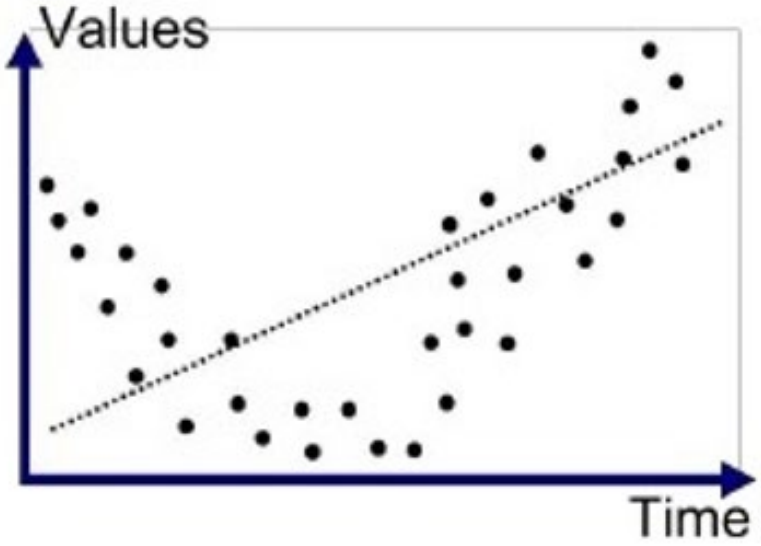
Algorithm:

Find the curve that maximizes the likelihood of observing the data

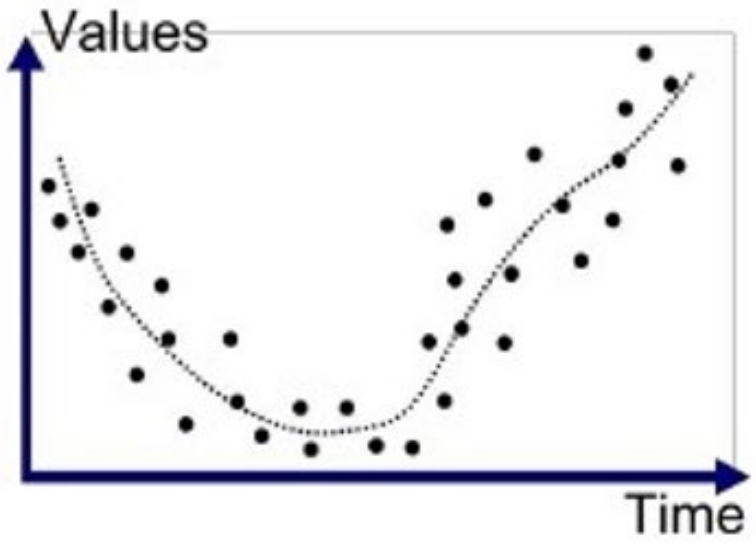
Model:

$$P = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

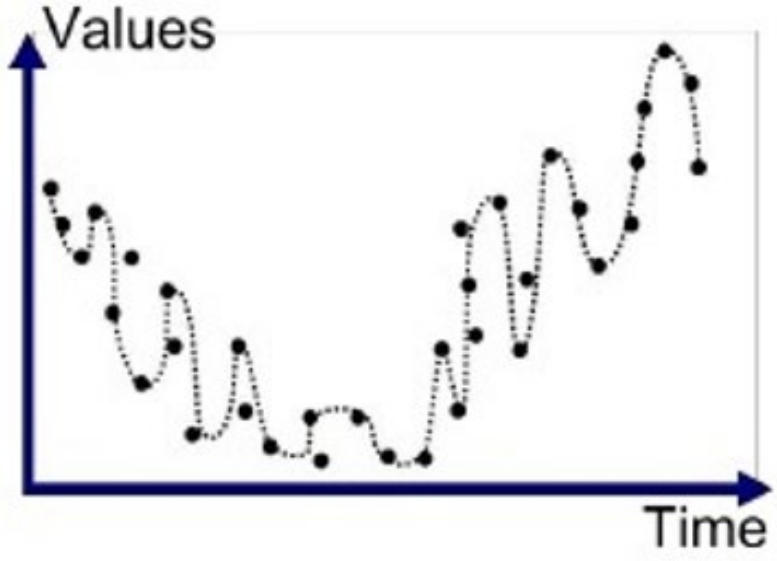
What’s the best fit?



Underfitted

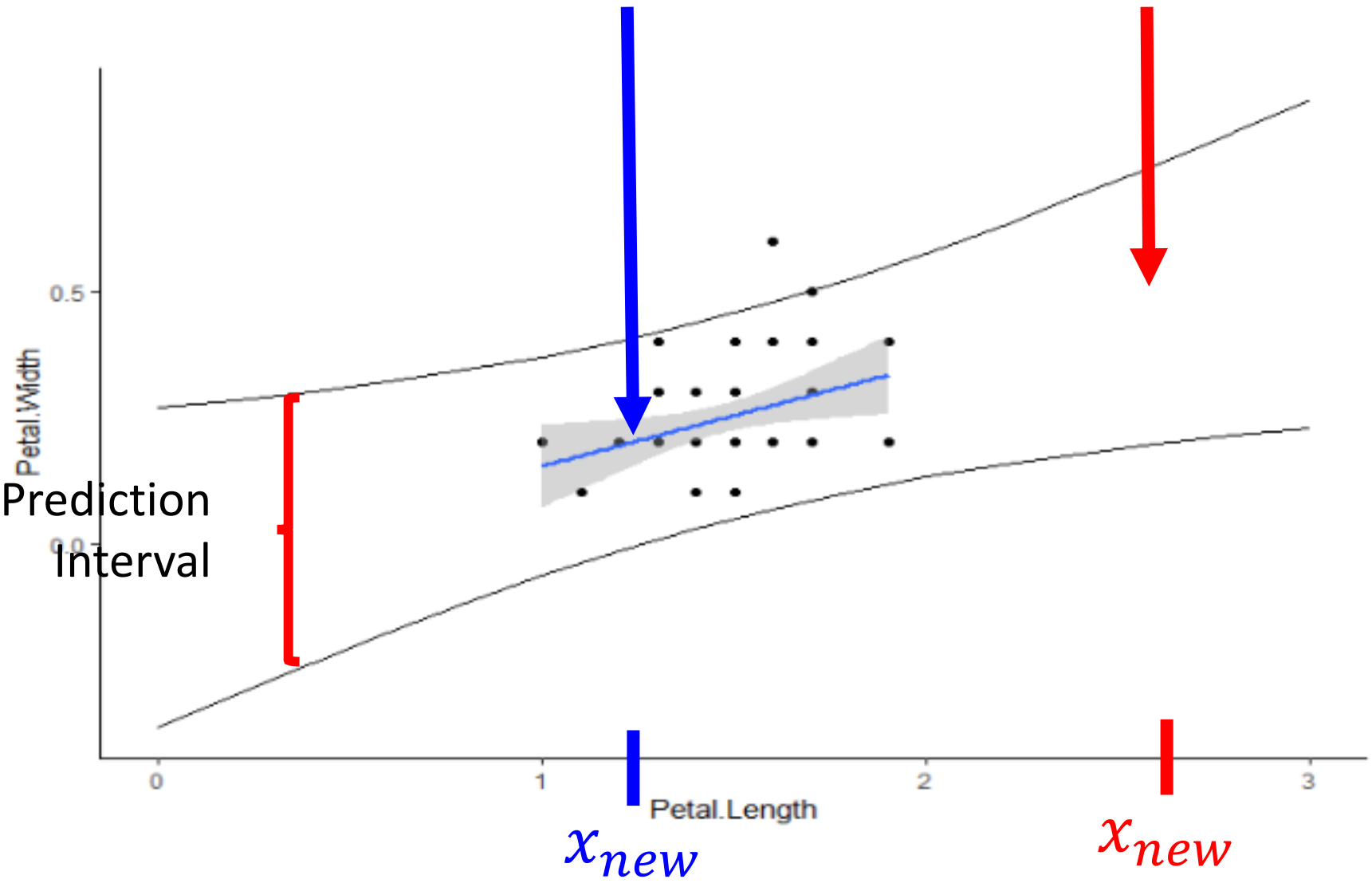


Good Fit/Robust

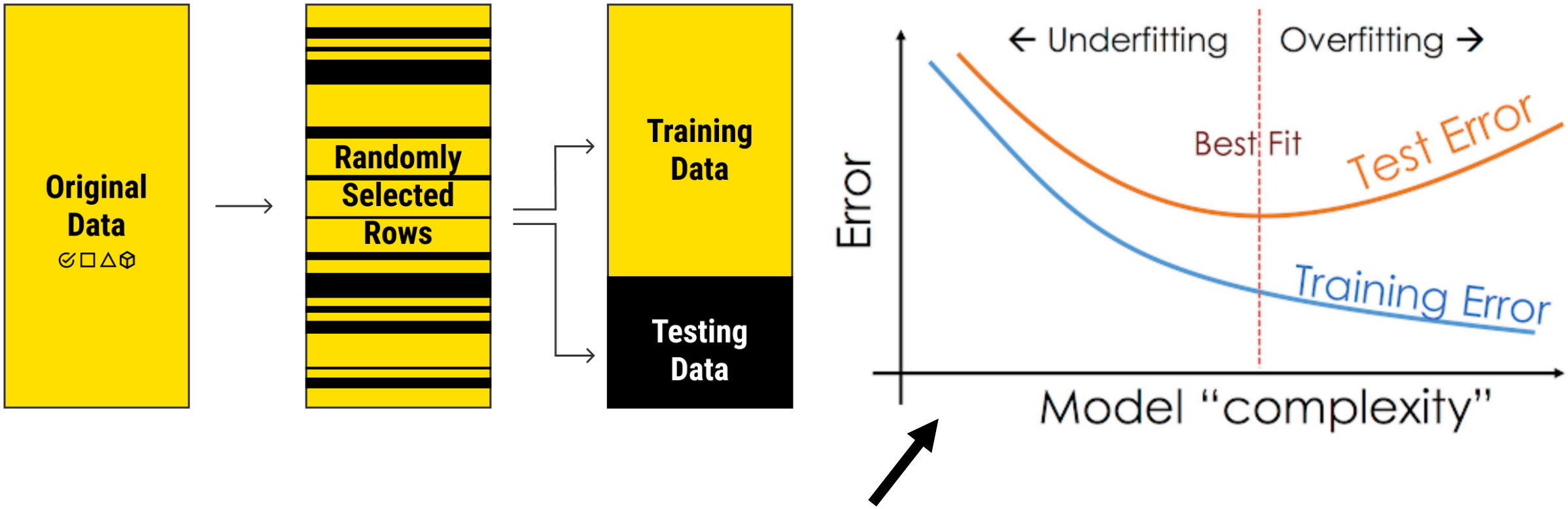


Overfitted

Prediction – interpolation & extrapolation



Training, testing, and overfitting



"bias-variance tradeoff" (related)

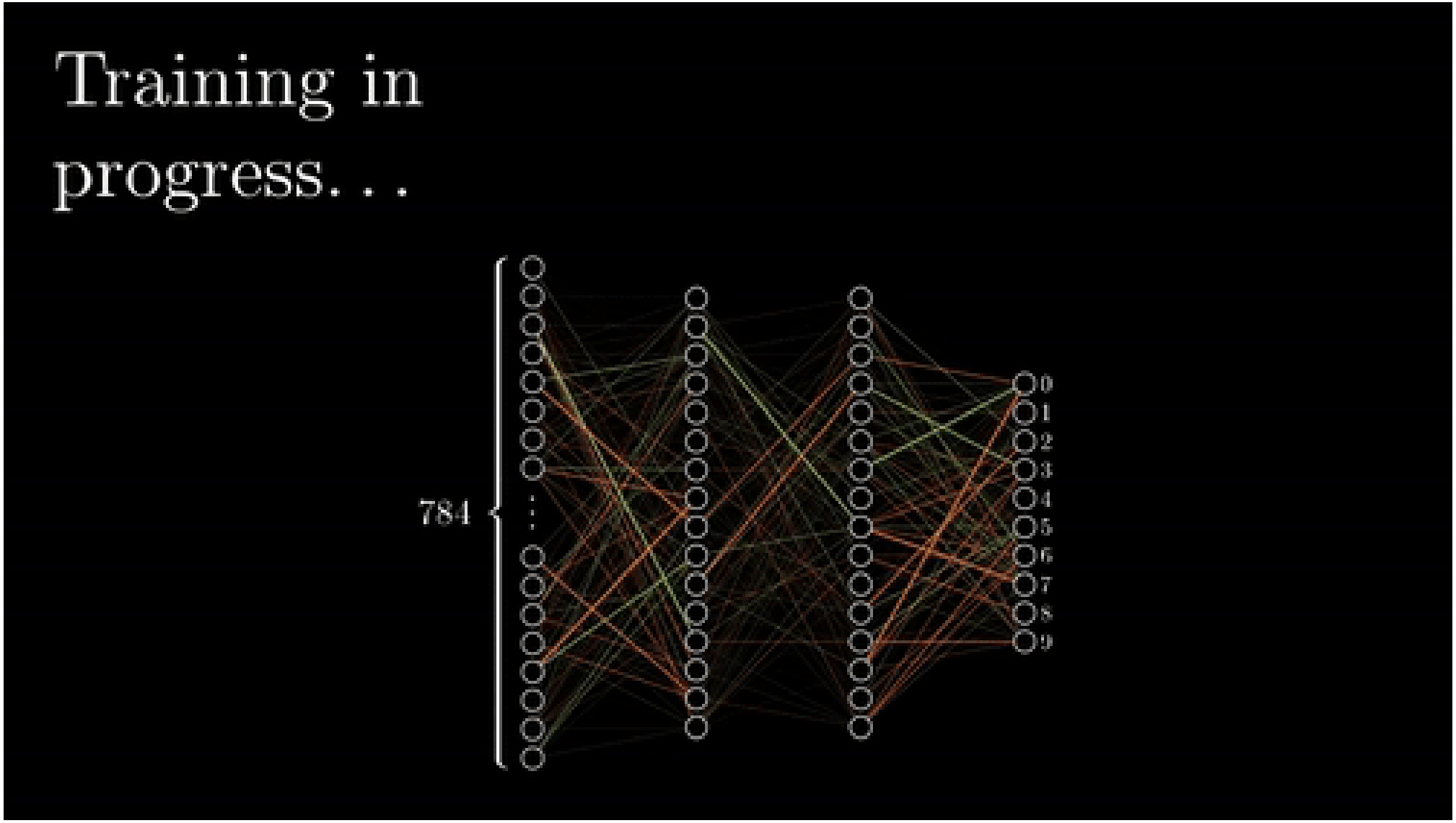
Machine learning

> “Computer algorithms that can improve automatically through **experience** and by the use of **data**”

Pros: Outstanding tools for prediction (better than humans)

Cons: Difficult to understand for explanation

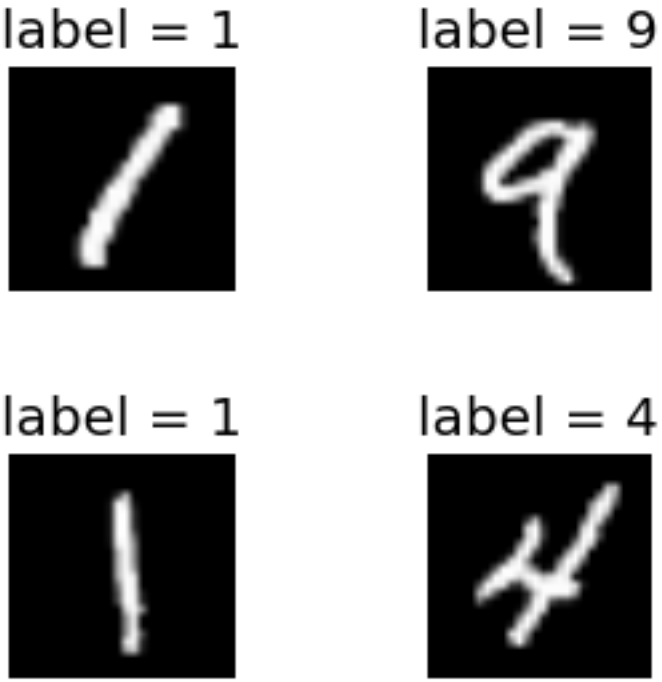
Machine learning



Many kinds of machine learning algorithms

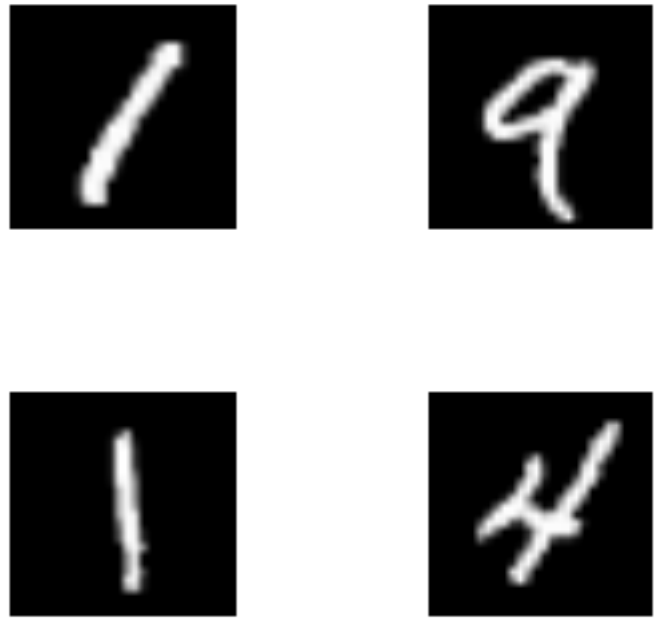
Supervised

Ground truth available



Unsupervised

Ground truth unavailable

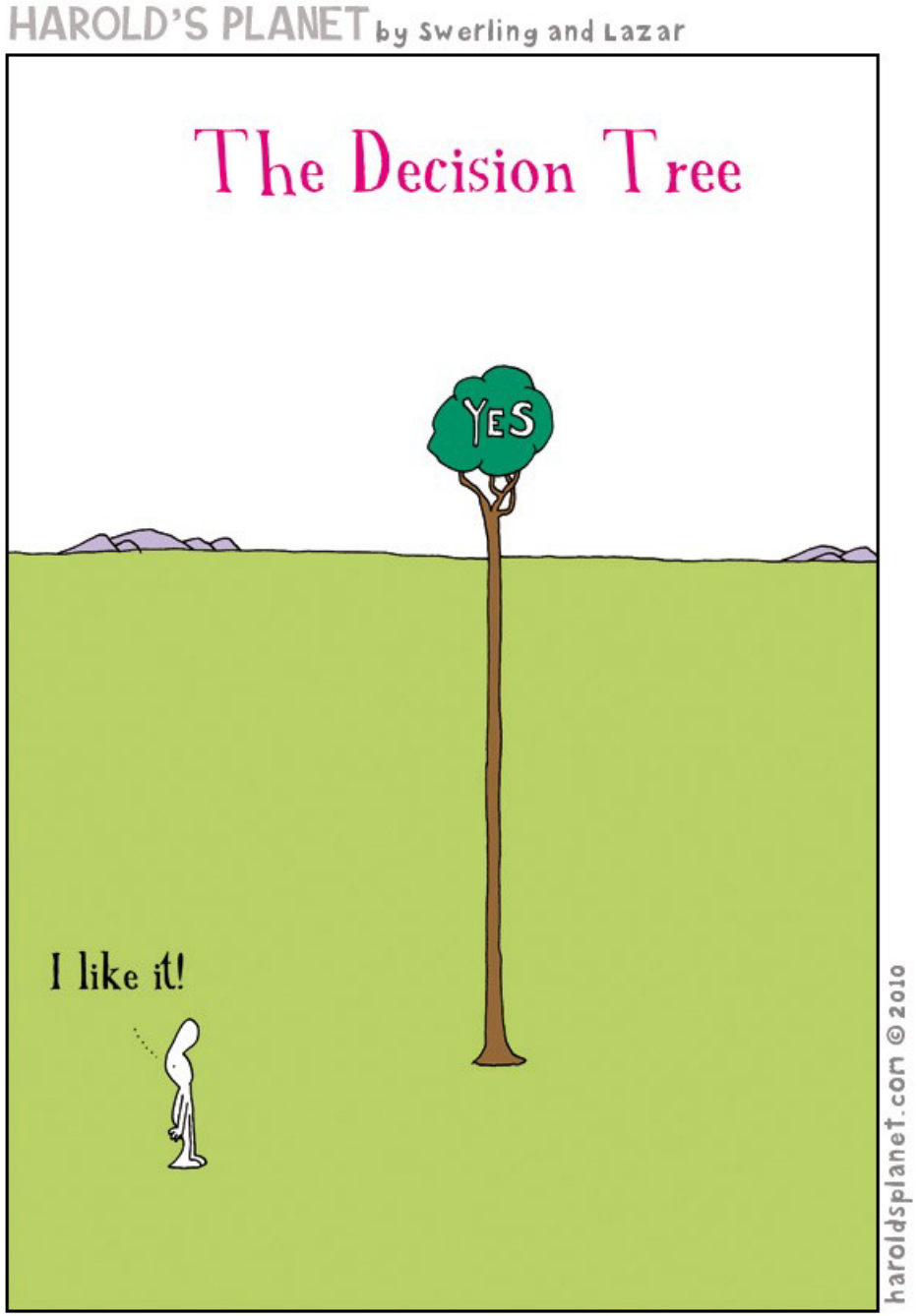


Random forests

[Recommended Reading by Breiman and Cutler]

- Easy to learn
- Easy to execute
- Easy to interpret
- Difficult to overfit

Supervised: need ground truth data



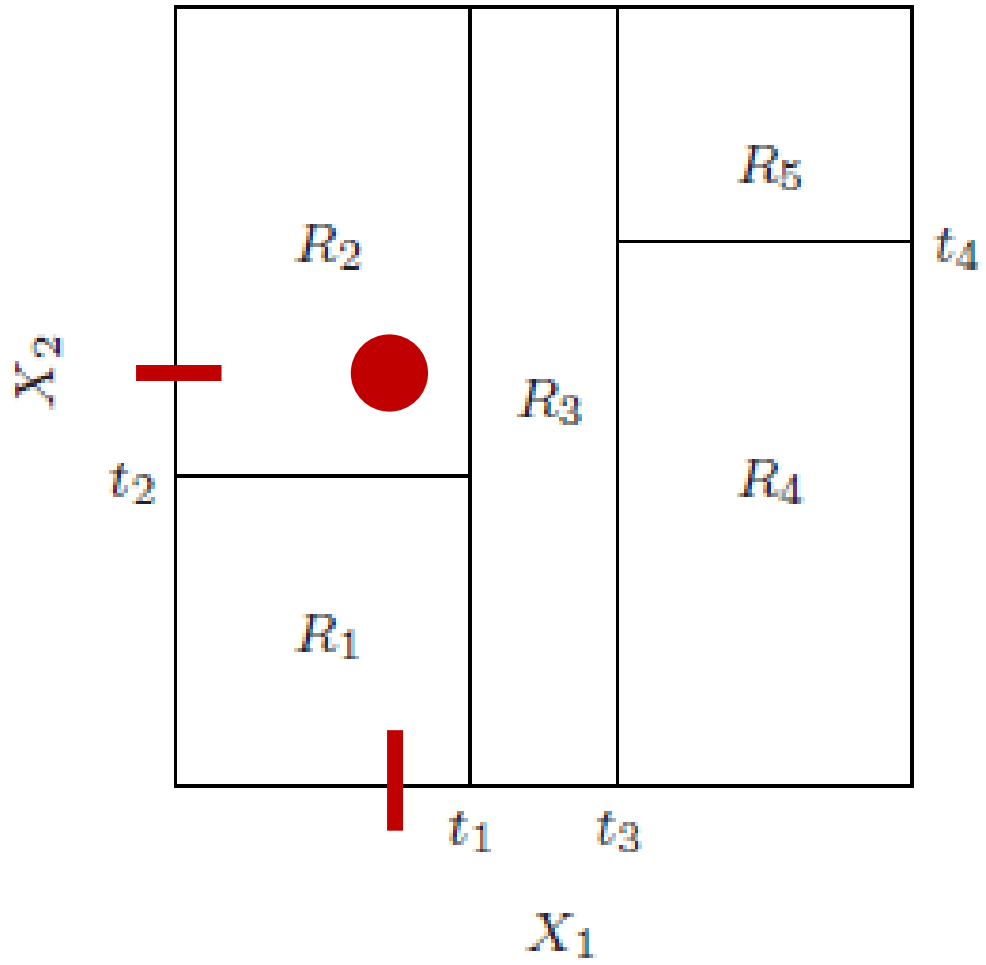
DONUT DECISION MAKER

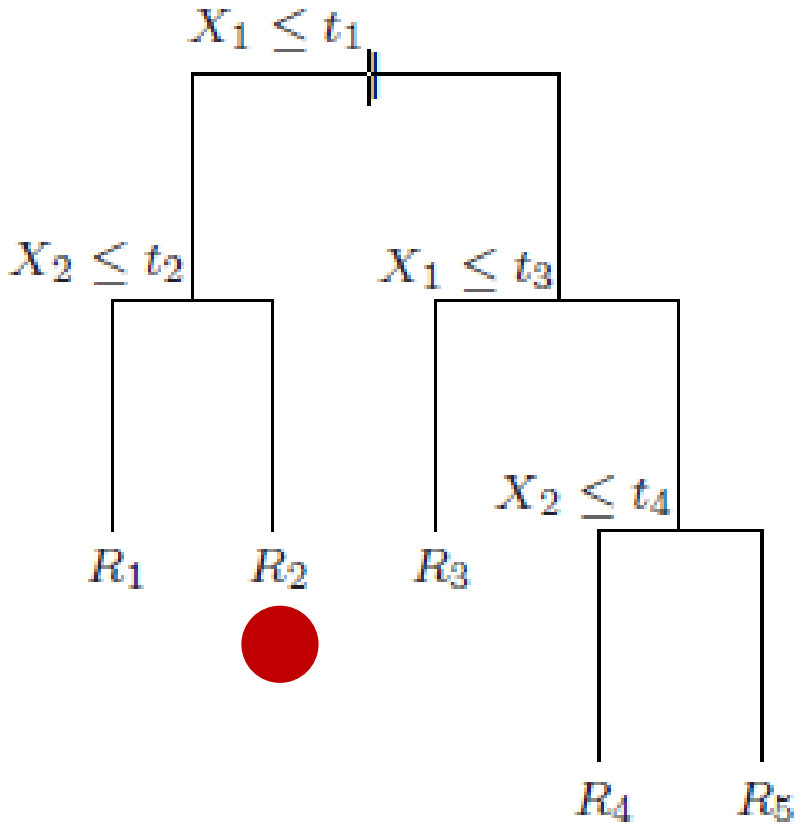
NATIONAL DONUT DAY 2020

DO I WANT A DOUGHNUT?

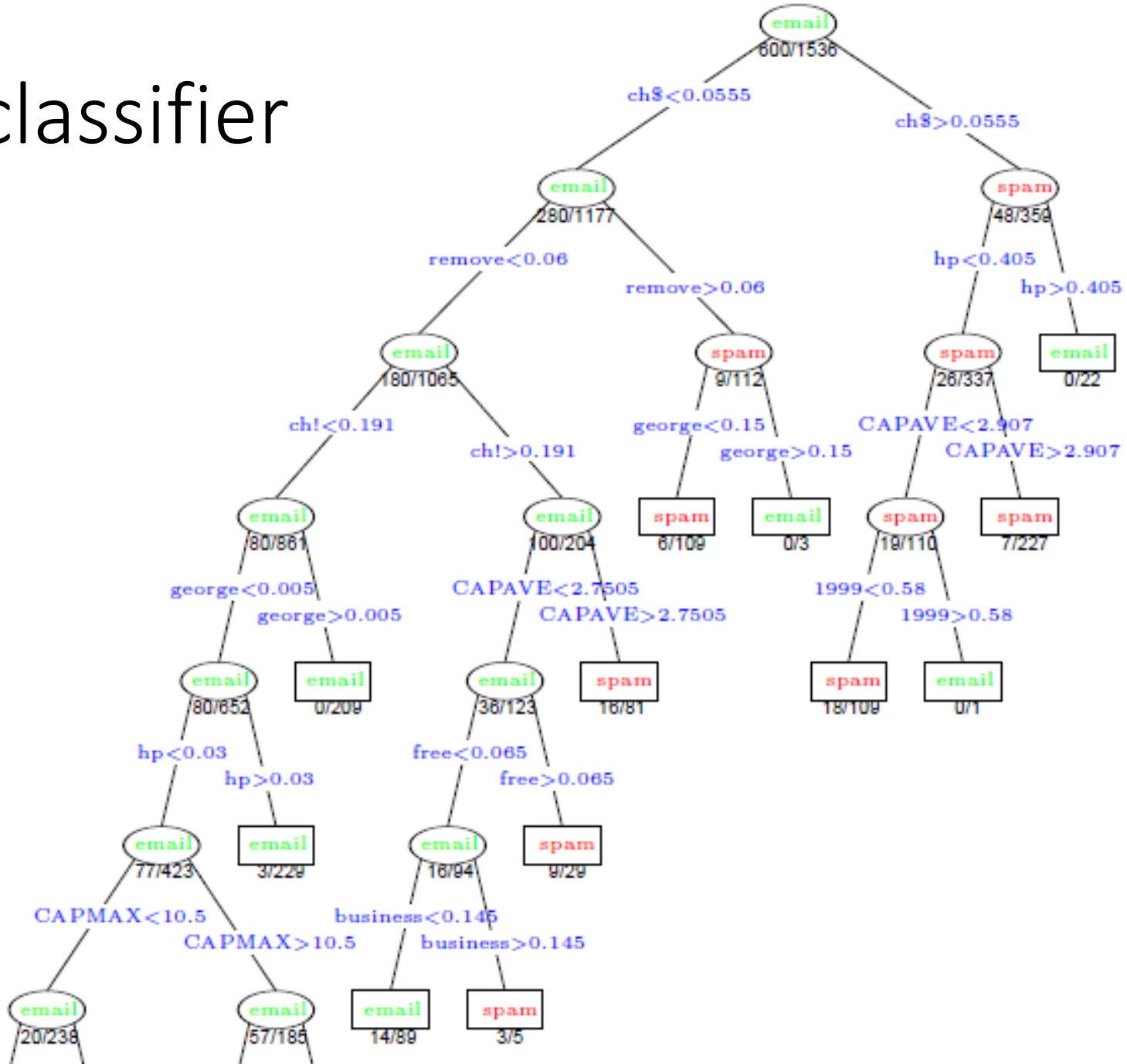
- YES
  - DO I DESERVE IT?
    - YES
      - ARE YOU SURE?
        - YES → GET IT.
        - NO → DO JUMPING JACKS FIRST.
    - NO
      - IS IT A GOOD DOUGHNUT?
        - YES → WHAT ARE YOU WAITING FOR? GRAB IT NOW.
        - NO → WAIT 'TIL YOU FIND A SINFUL, UNFORGET-TABLE DOUGHNUT.
- NO → MAYBE YOU WANT AN APPLE?

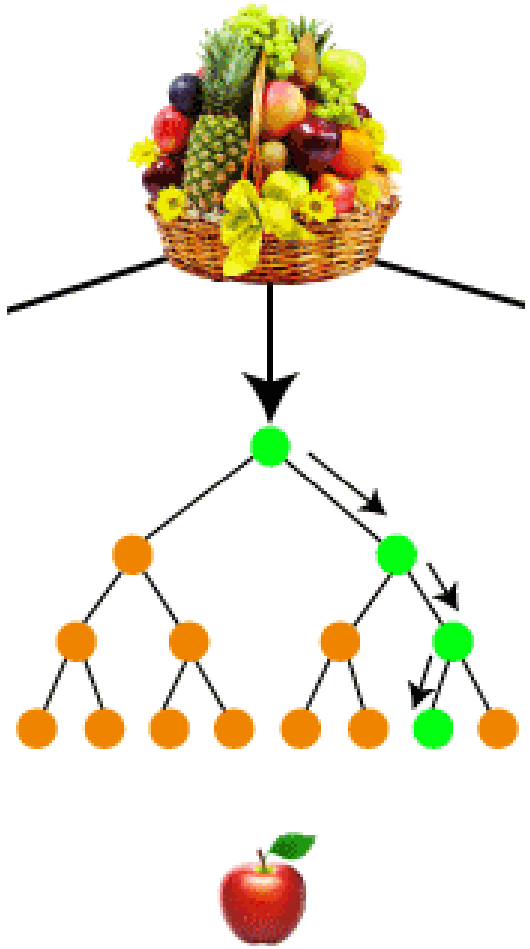
Growing a decision tree – classification task



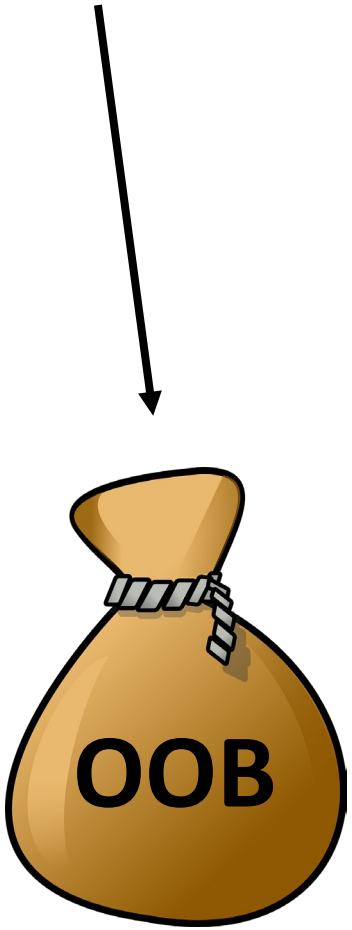


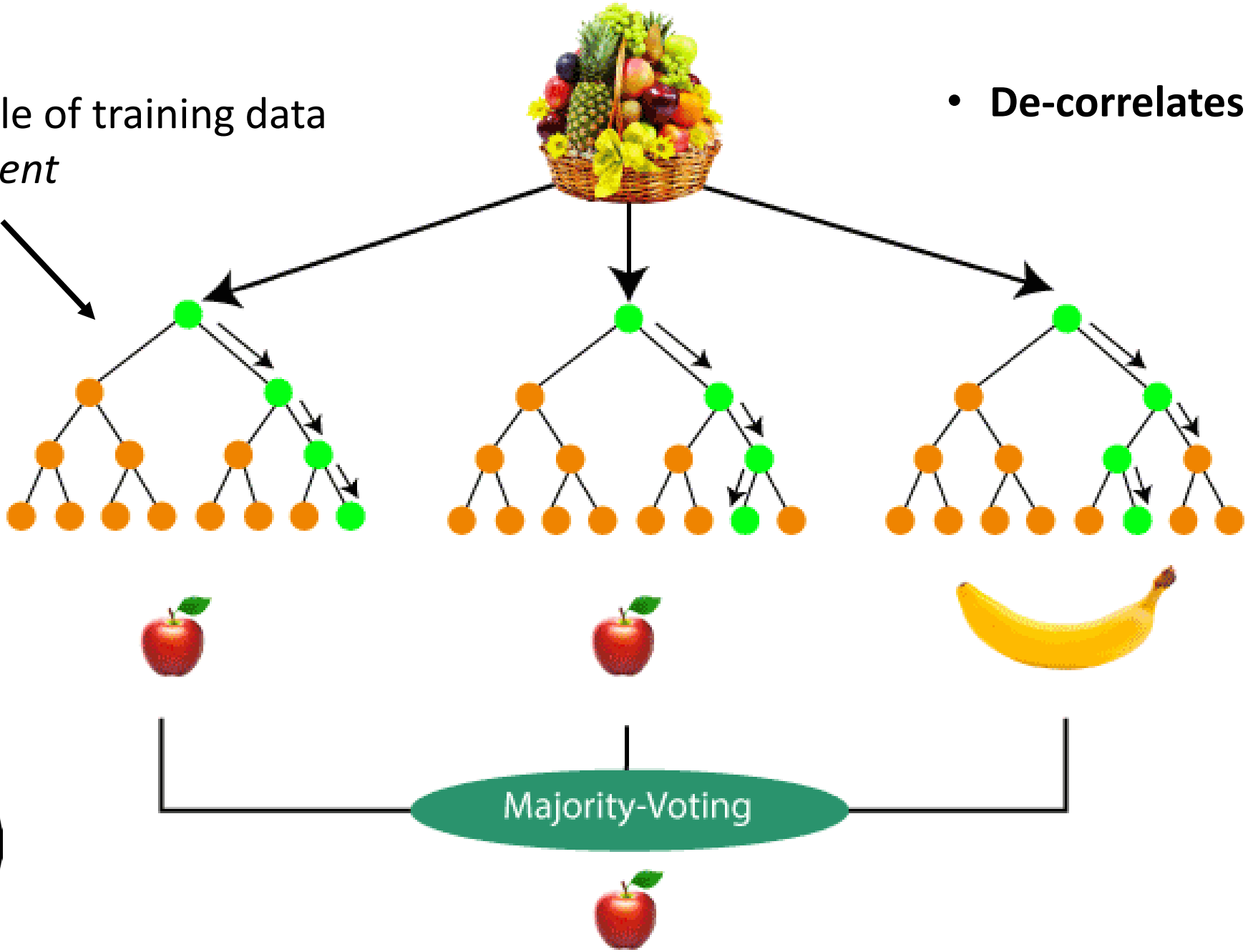
Email classifier





Random sample of training data
with replacement
“bagging”



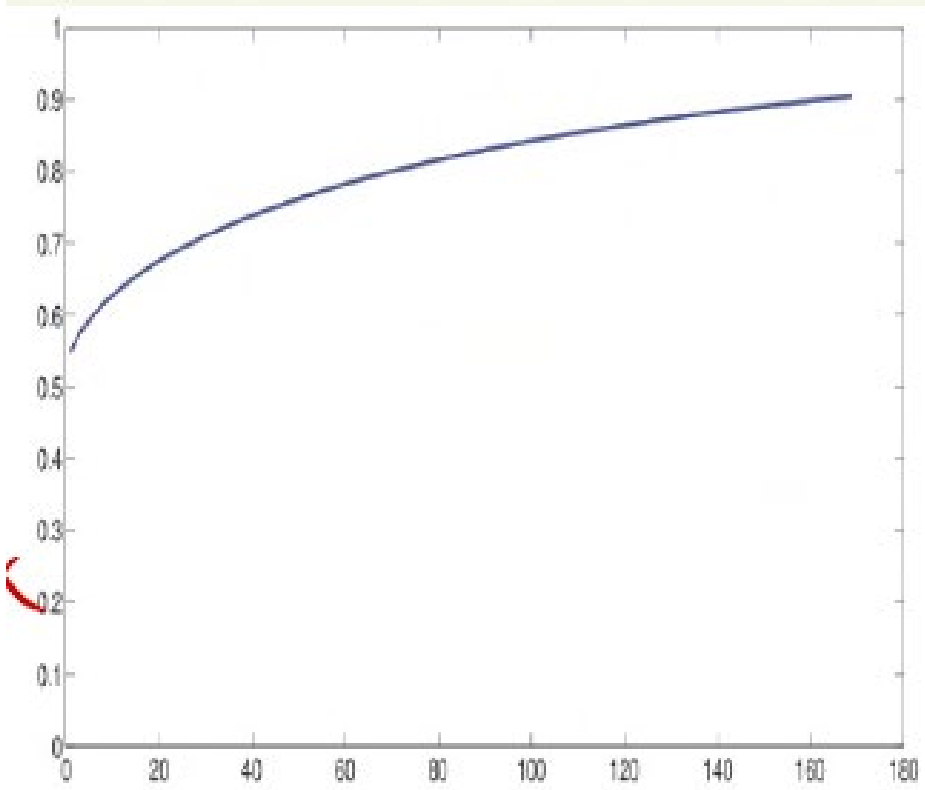


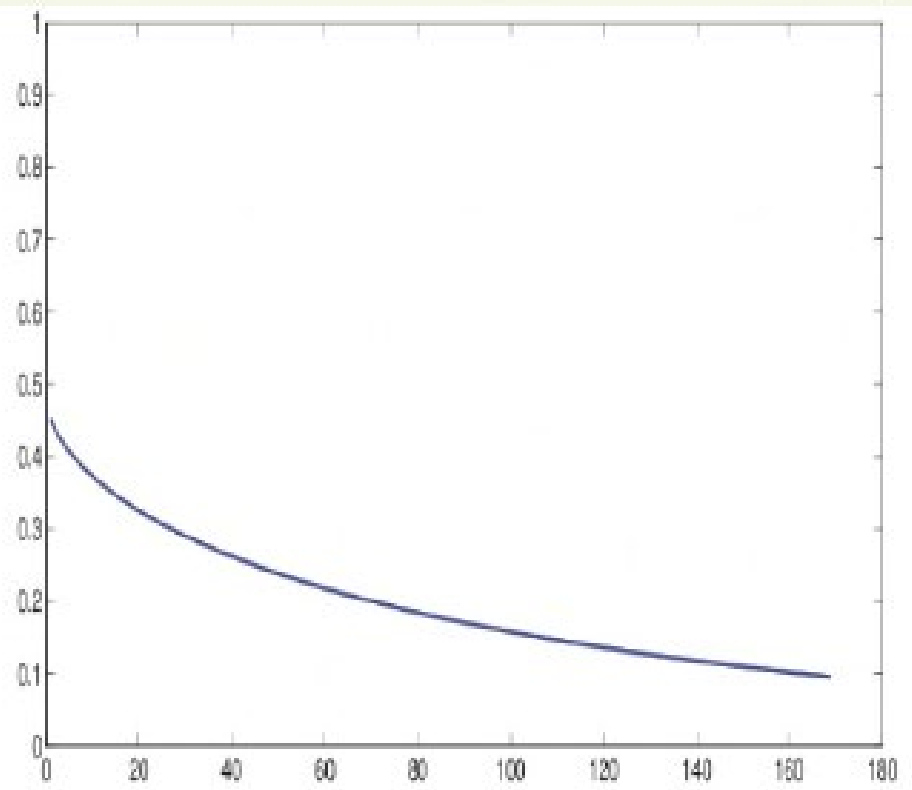
- **De-correlates trees**

Wisdom of crowds – ensemble methods; “weak learners”

Condorcet’s jury theorem

probability the majority is correct





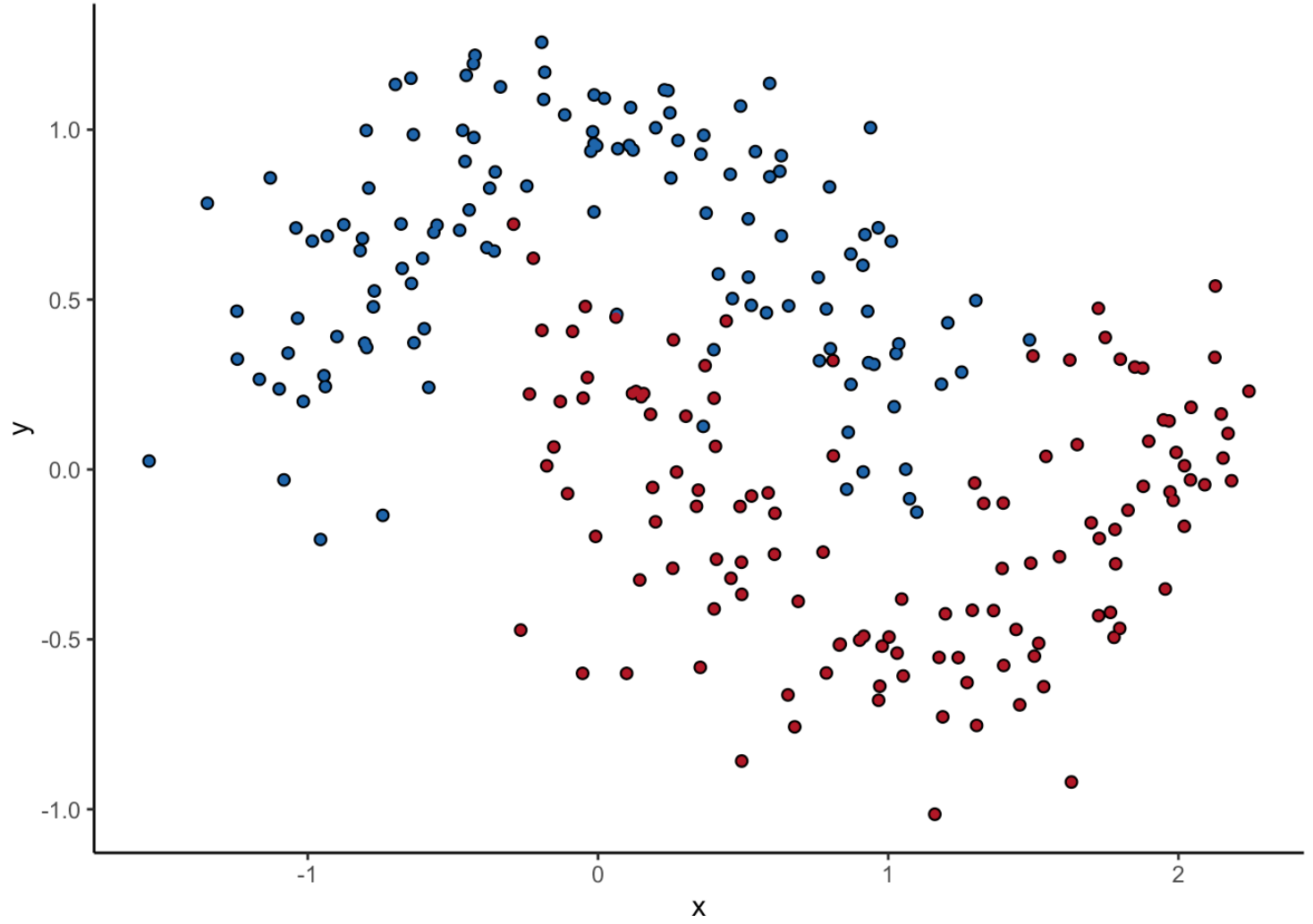
# of individuals

$p = 0.55$

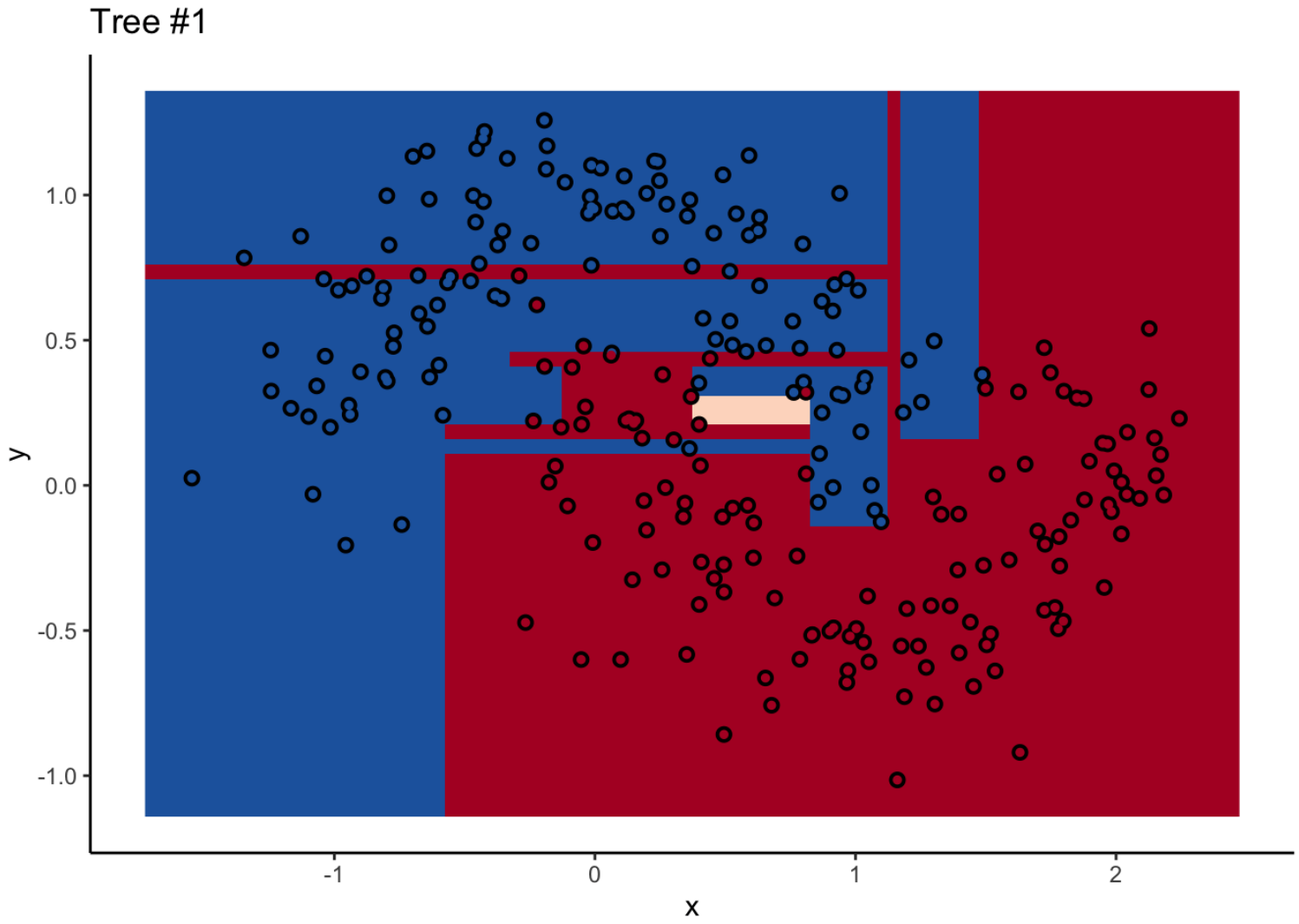
$p = 0.45$

probability an individual is correct

Ensemble methods – “weak learners”



Ensemble methods – “weak learners”



Ensemble methods – “weak learners”