Data Graduate Program N°2
Cdiscount Academy

Lundi 20/09 – Mercredi 23/10

SOMMAIRE

- ML

ML

MACHINE LEARNING : WHAT IS IT?

- Deals with building predictive models

Why and when?

Some examples of machine learning

WHICH IRIS IS THAT?

- Setosa
- Versicolor
- Virginica

Sepal length	Sepal width	Petal length	Petal width	Iris type
6cm	3.4cm	4.5cm	1.6cm	versicolor
5.7cm	3.8cm	1.7cm	0.3cm	setosa
6.5cm	3.2cm	5.1cm	2cm	virginica
5cm	3.cm	1.6cm	0.2cm	setosa

IS THIS PERSON RICH?

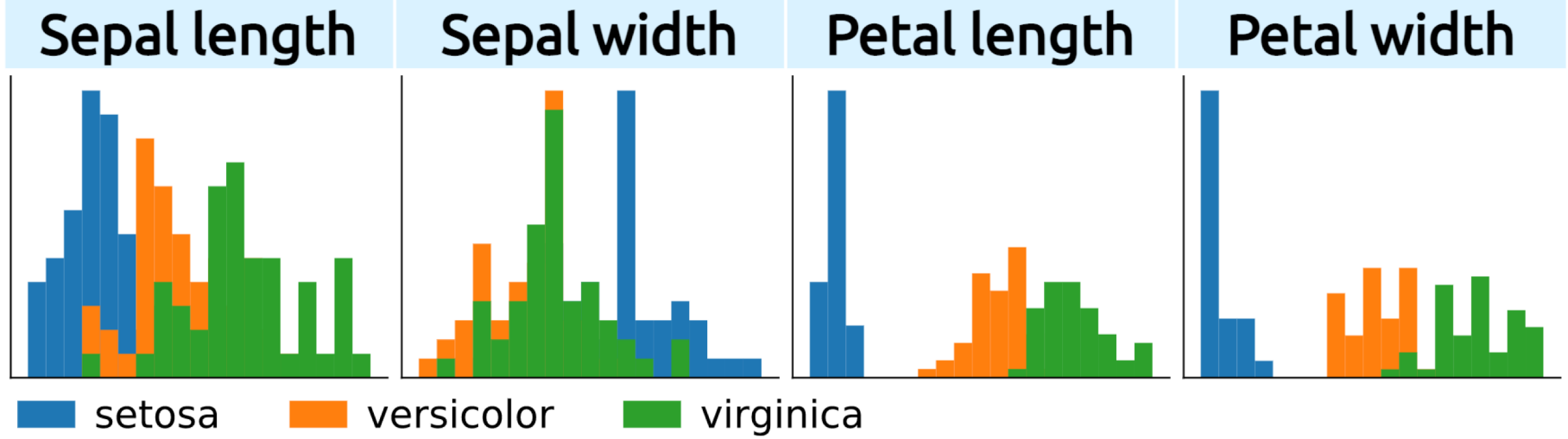
- US Census data:

Age	Workclass	Education	Marital-status	Occupation	Relationship	Race	Sex	Capital-gain	Hours-per-week	Native-country	Class
25	Private	11th	Never-married	Machine-op-inspct	Own-child	Black	Male	0	40	United-States	<=50K
38	Private	HS-grad	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	50	United-States	<=50K
28	Local-gov	Assoc-acdm	Married-civ-spouse	Protective-serv	Husband	White	Male	0	40	United-States	>50K
44	Private	Some-college	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	40	United-States	>50K

ENGINEERING RULES: DATA VERSUS EXPERTS

Cdiscount

Expert knowledge: setosa irises have small petals



This rule can be inferred from the data

KEYRUS data

Cdiscount

Predictive analysis?

Beyond classical statistical tools

KEYRUS data

GENERALIZING: CONCLUDING ON NEW INSTANCES

Cdiscount

- Many sources of variability

- age
- marital-status
- race
- hours-per-week
- workclass
- occupation
- sex
- native-country
- education
- relationship
- capital-gain

- + Noise: unexplainable variance

KEYRUS data

MEMORIZING

- Many sources of variability

- store all known individuals (the census)
- given a new individual, predict the income of its closest match in our database

Trying out this strategy on individuals picked from the data we have (the census) **what error rate do we expect?**

0 errors

Yet, we will make errors on new data

GENERALIZING IS NOT MEMORIZING

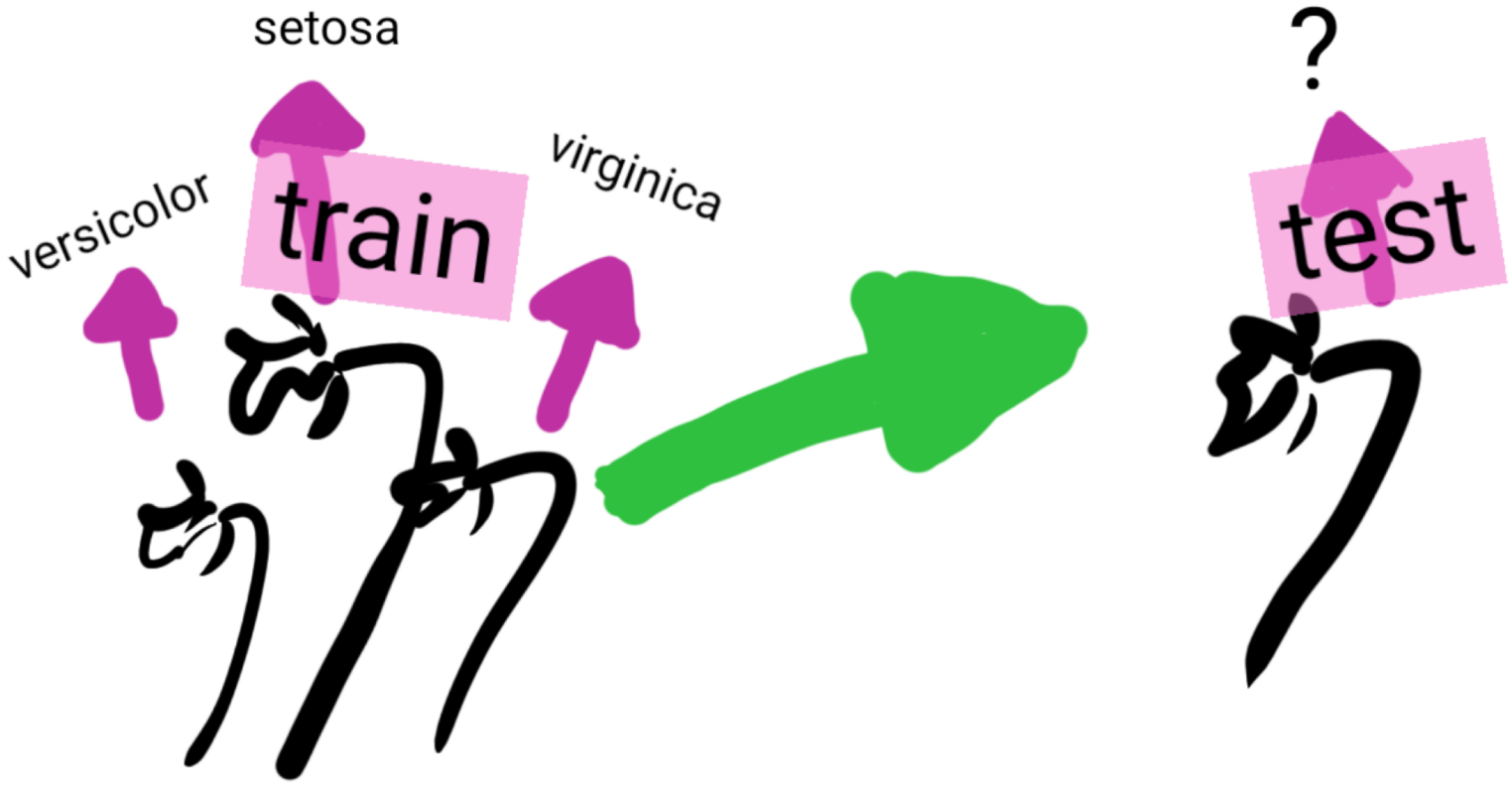
"test" data ≠ "train" data

Data on which the predictive model is applied

Data used to learn the predictive model

- Different sampling of noise
- Unobserved combination of features

THE MACHINE LEARNING WORKFLOW



Some vocabulary

DATA MATRIX

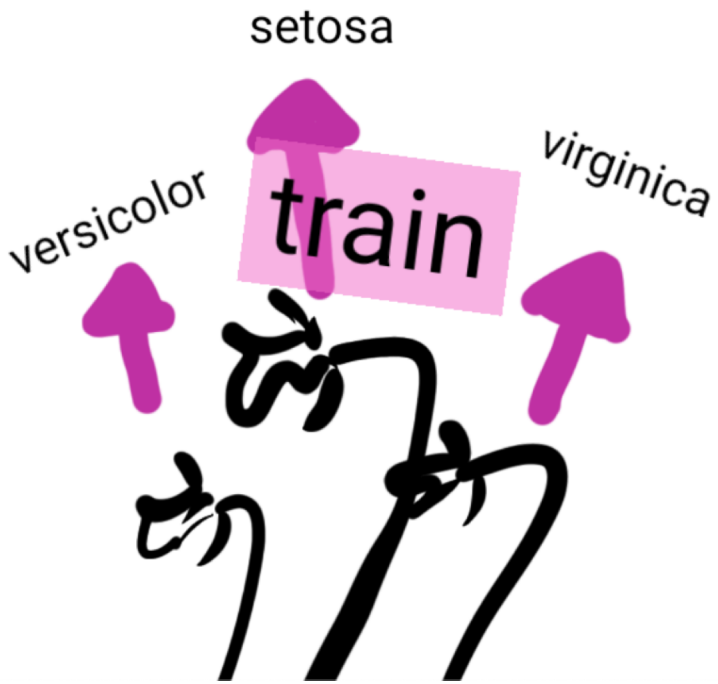
We deal with a table of data (figuratively, an Excel sheet):

- Rows are different observations, or **samples**
- Columns are different descriptors, or **features**

Sepal length	Sepal width	Petal length	Petal width	Iris type
6cm	3.4cm	4.5cm	1.6cm	versicolor
5.7cm	3.8cm	1.7cm	0.3cm	setosa
6.5cm	3.2cm	5.1cm	2cm	virginica
5cm	3.cm	1.6cm	0.2cm	setosa

SUPERVISED MACHINE LEARNING

- A data matrix *X* with *n* observations
- A target *y*: a property of each observation



The goal is to **predict** *y*

UNSUPERVISED MACHINE LEARNING

- A data matrix X with n observations

The goal is to extract from X a structure that generalizes.

Very wide variety of different problems.

REGRESSION AND CLASSIFICATION

Supervised learning: predicting a target *y*

- **Classification**: *y* is discrete (qualitative), made of different classes

  eg types of irises: Setosa, Versicolor, Virginica

- **Regression**: *y* is continuous (quantitative), a numerical quantity

  eg wage prediction

Shriman TIWARI

Tech Lead/Manager Data Science

Mobile: +33 (0)6 49 71 80 68
shriman.tiwari@keyrus.com

www.keyrus.com

KEYRUS