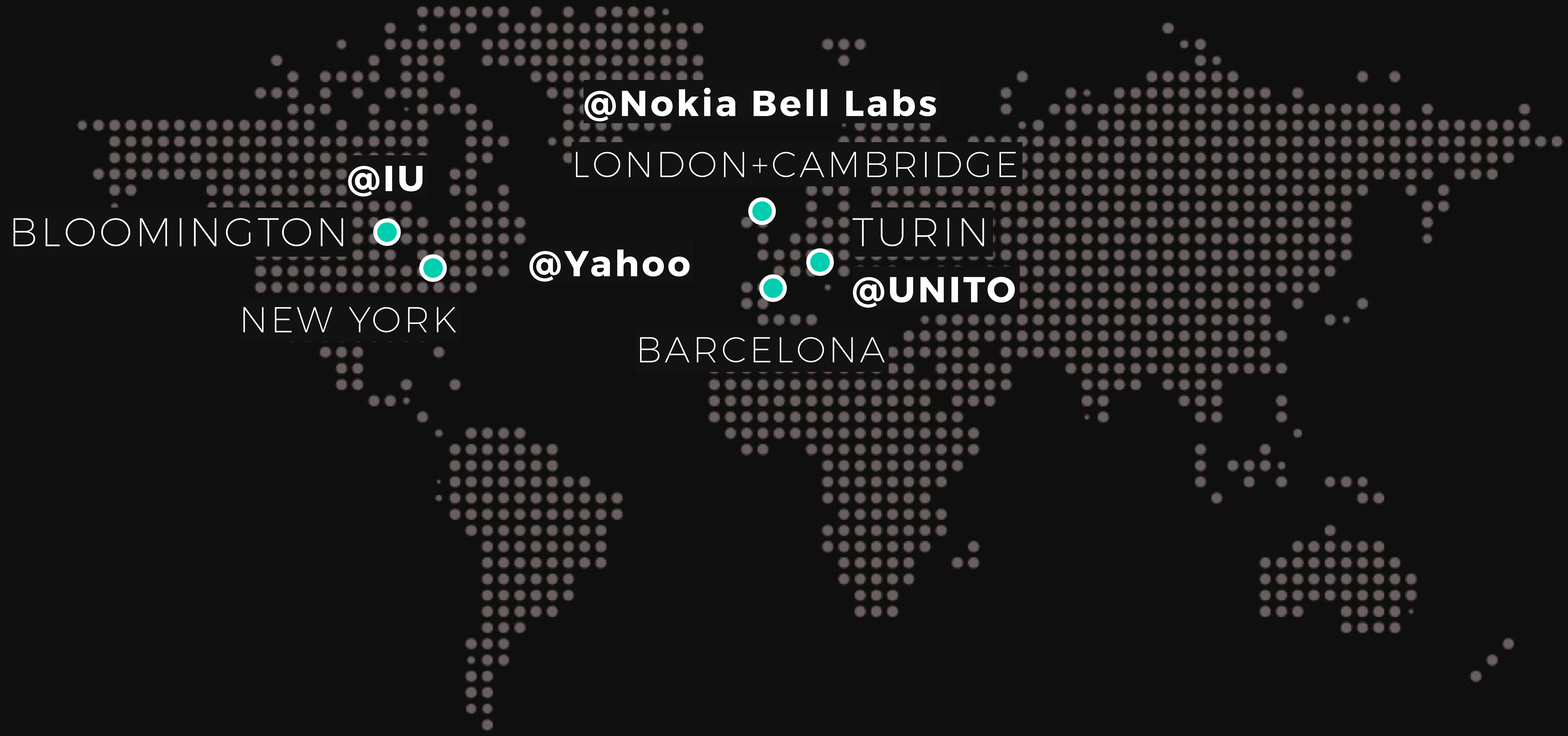
Advanced Machine Learning

Intro

Instructor: Rossano Schifanella

@UDD

Few words about me



Tools and Infrastructure

Python packages for:

scientific computing

numpy, scipy

data visualization

matplotlib, seaborn

data manipulation and analysis

pandas

sharing code+formatted descriptions

Jupyter Notebook

machine learning

scikit-learn

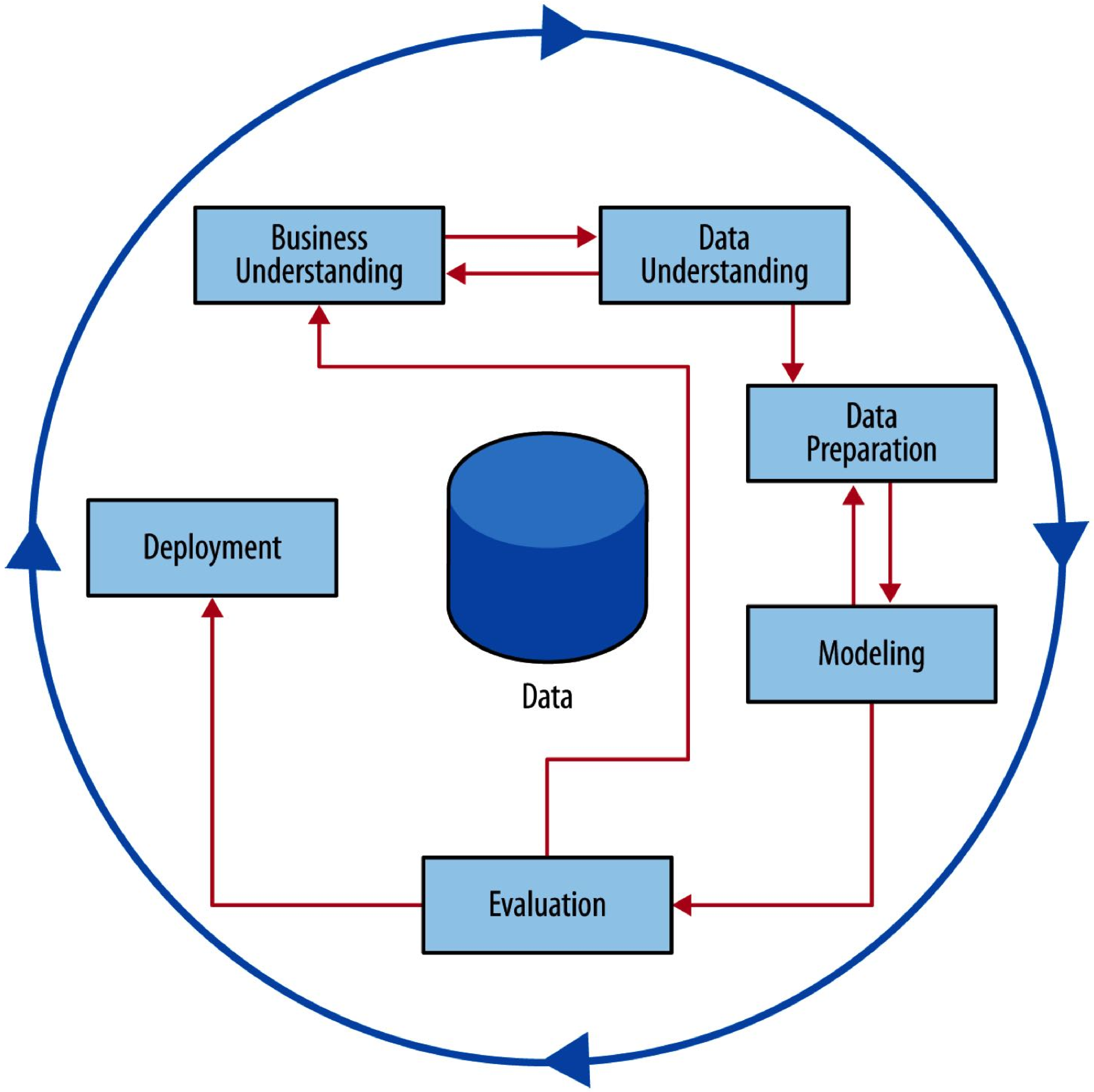
Material

- All the material of the course (slides, exercises) will be uploaded to the **course portal** at
  - **https://github.com/rschifan/MachineLearnig-UDD18**
- Additional resources like
  - links to supplementary material
  - datasets
  - tutorials
  - use cases

will be shared on the same platform contextually with each class

Learning from Data?

- Discover from data patterns and models that are:
  - **Valid**
    - hold on new data with some certainty
  - **Useful**
    - should be possible to act on the item and solve a problem
  - **Unexpected**
    - non-obvious to the system
  - **Understandable**
    - humans should be able to interpret the pattern



Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Business understanding and problem definition

The process starts always with the **definition of the problem** to solve

e.g., predicting customer churn: a telco company you work for has a major problem with customer retention in their wireless business.

What Is Data?

48 k
apps are downloaded by Apple users

3.8 mln
Google searches are performed

149.513
emails are sent

400 hours
of new videos are uploaded to YouTube

3.3 mln
pieces of content are posted on Facebook

Report from IBM states that **90%** of the data in the world today has been created in the last two years alone

---

IDC says that worldwide revenues for big data and business analytics will grow **from $130.1 billion in 2016** to more than $203 billion in 2020

In 2000, only 738 million people used the internet, but by 2017, this number grew to **3,6 billion**

What is Data?

Collection of data objects and their attributes

An **attribute or feature or variable** is a property or characteristic of an **object or sample**

e.g., eye color of a person, temperature, etc.

A **collection of attributes** describe an object

Attributes
Features

Objects
Samples

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Attribute Values

- **Attribute values are numbers or symbols assigned to an attribute**
- Distinction between **attributes** and **attribute values**
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

Quantitative

Temperature

Qualitative

Taste

- **Discrete Attribute**
  - Finite or countably infinite values
  - zip codes, counts, or set of words in a collection of documents
- **Continuous Attribute**
  - Real numbers as values
  - height, weight

Types of data sets

Record

- Data Matrix
- Document Data
- Transaction Data

Graph

- World Wide Web
- Molecular Structures

Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

Characteristics of Structured Data

- **Dimensionality**
  - Curse of dimensionality
- **Sparsity**
  - Only presence counts
- **Resolution**
  - Patterns depend on the scale

Whenever you work with data ask…

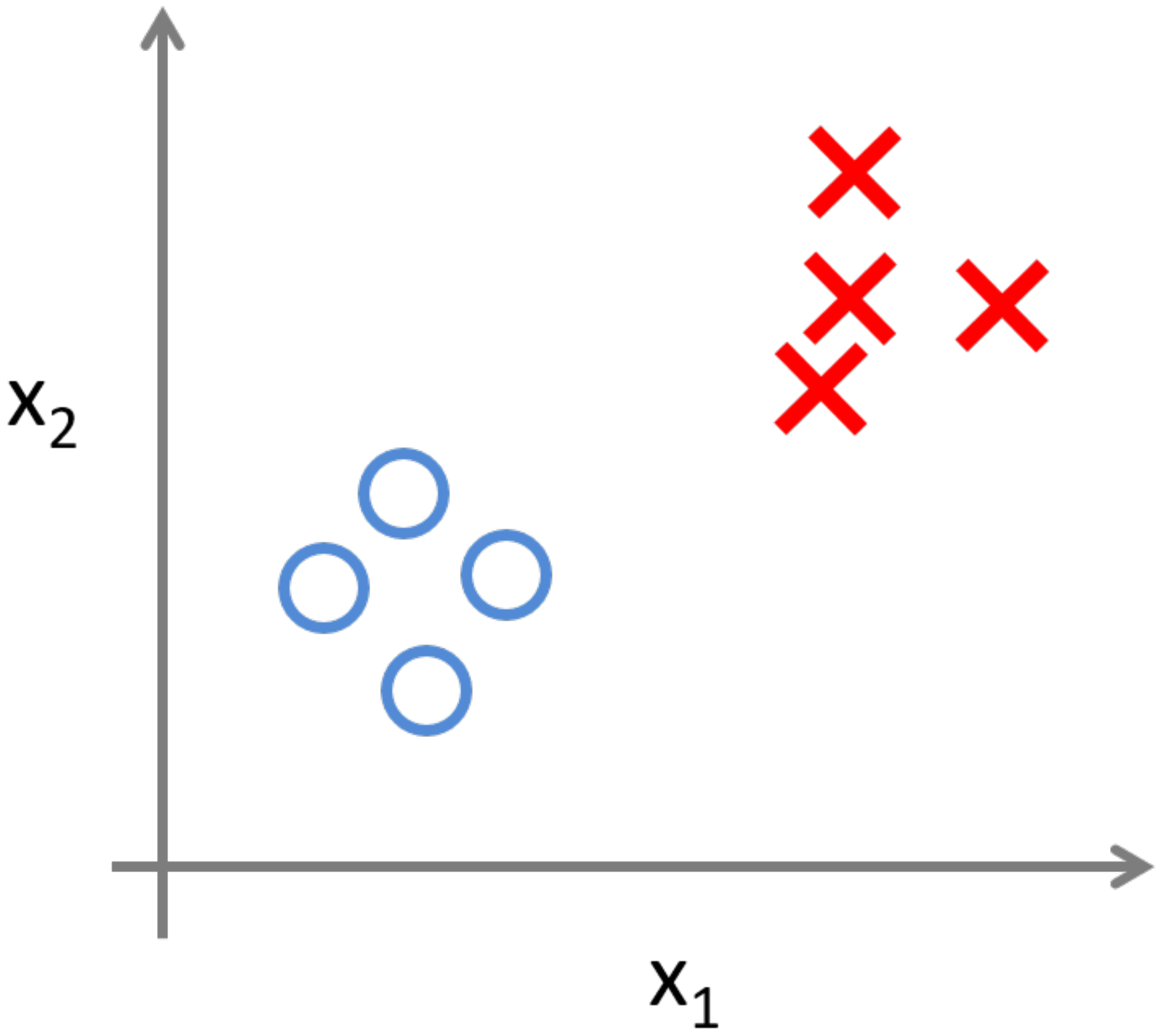
could this occur “randomly”?

are there confounding factors?

Approaches to learning

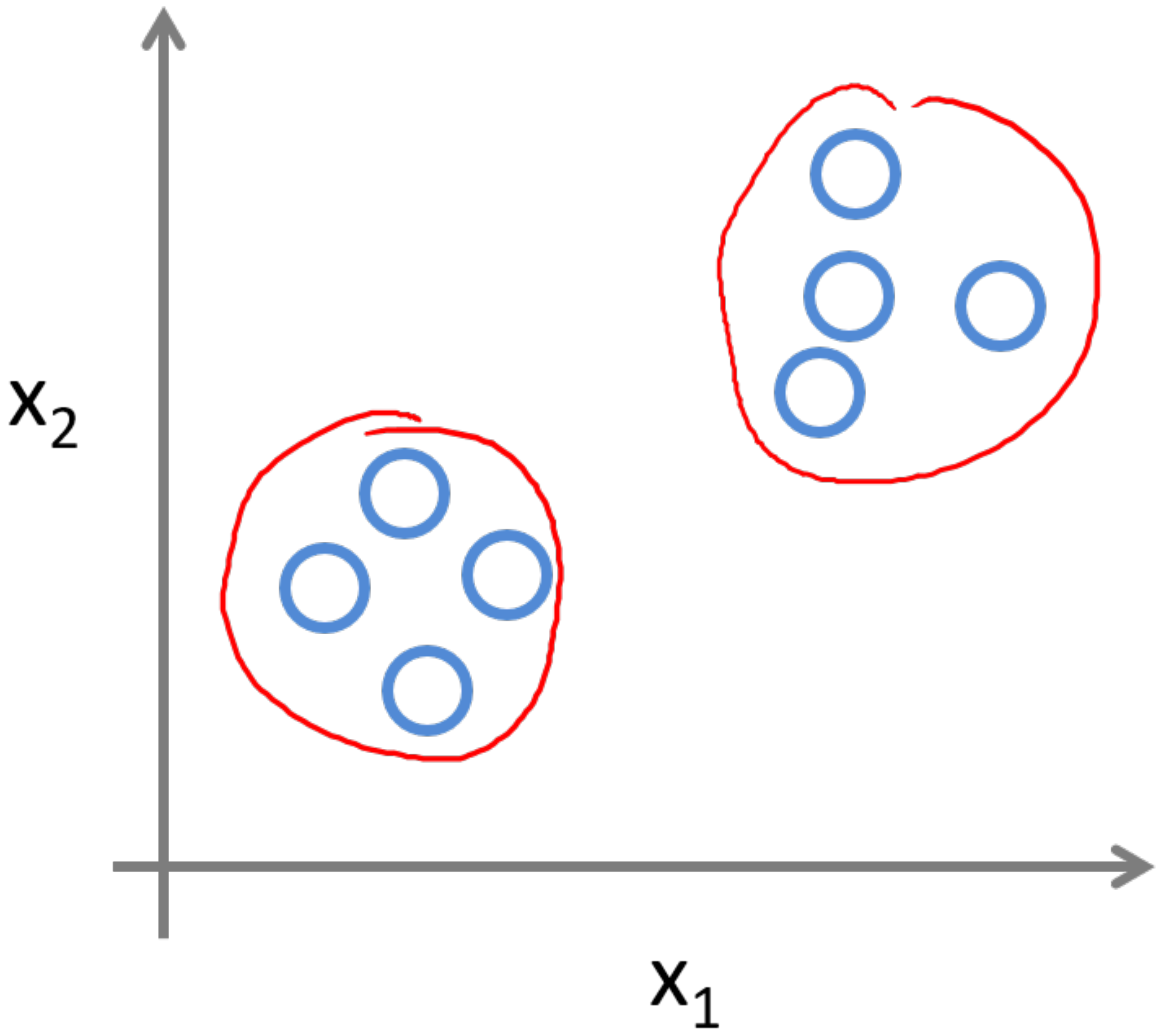
- **supervised**
- **unsupervised**
- **semi-supervised**
- **reinforcement learning**

Supervised Learning



learn from **labeled** data

Unsupervised Learning

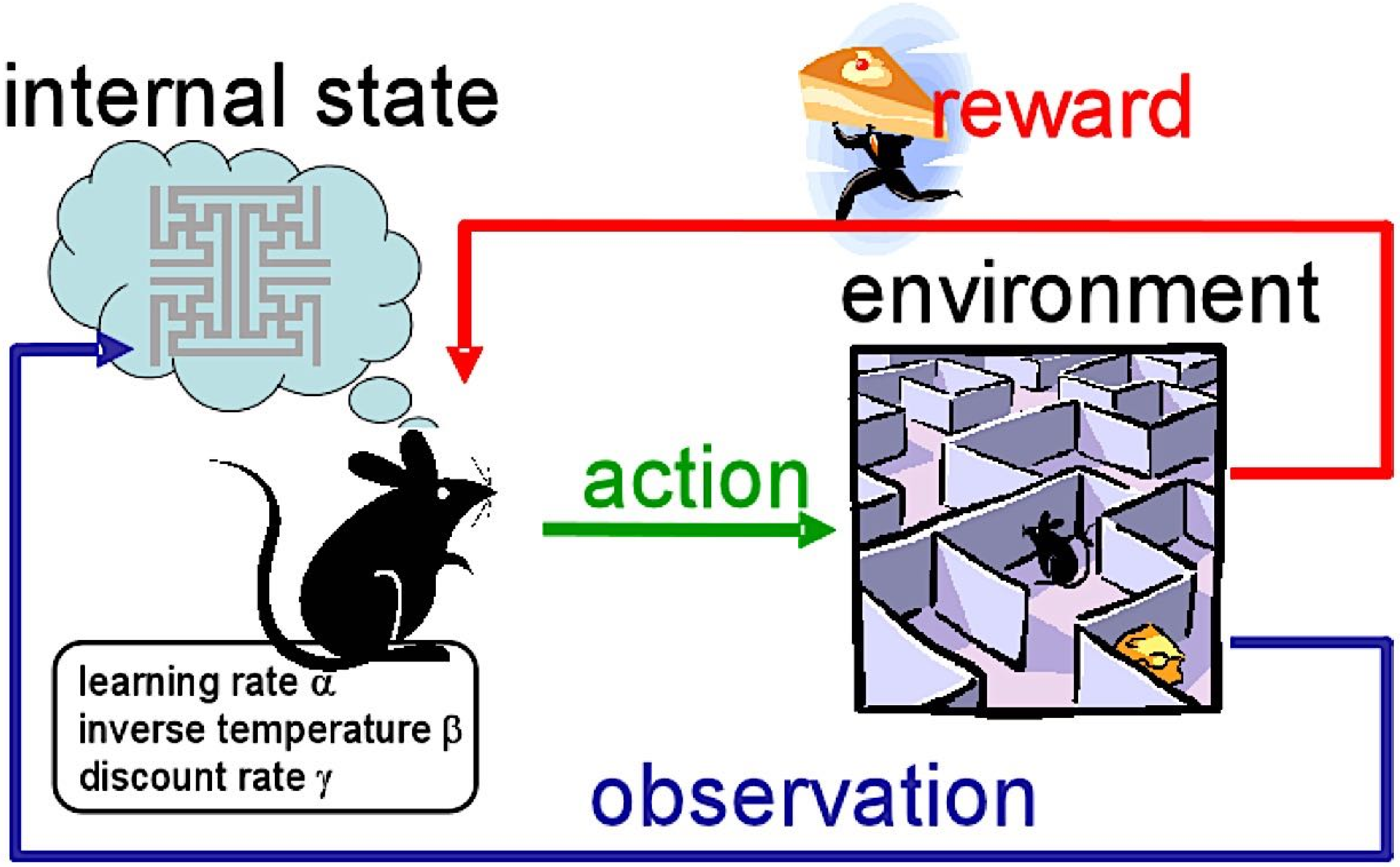


learn from **unlabeled** data

semi-supervised

- **few labeled data**
- **a lot of unlabeled data**
- closer to how **humans** learn
- **current systems are not very good at it**

Reinforcement learning



Tasks

- **Predictive**
  - Use some variables to predict unknown or future values of other variables.
- **Descriptive**
  - Find human-interpretable patterns that describe the data.

Tasks

- **Classification** [Predictive]
- **Regression** [Predictive]
- **Anomaly Detection** [Predictive]
- **Clustering** [Descriptive]
- **Association Rule** [Descriptive]
- **Sequential Pattern Mining** [Descriptive]

Classification

A classifier is a mapping:

$$\hat{c} : \mathscr{X} \to \mathscr{C}$$

where

$$\mathscr{C} = \{C_1, C_2, \ldots, C_k\}$$

the “hat” over the name of the classifier denotes that the classifier is an approximation of the true but unknown function **c**.

An example is a pair:

$$(x, c(x)) \in \mathscr{X} \times \mathscr{C}$$

where **x** is an “instance” and **c(x)** is the true class of the instance (possibly contaminated by noise).

Classification

Learning a classifier involves constructing the function such that it matches **c** as closely as possible.

Example: Email Spam Detection

Class probability estimation

A class probability estimator is a scoring classifier that outputs probability vectors over classes, i.e., a mapping:

$$\hat{\mathbf{p}} : \mathscr{X} \to [0,1]^k$$

We write:

$$\hat{\mathbf{p}}(x) = \big(\hat{p}_1(x), \ldots, \hat{p}_k(x)\big)$$

where the i-th component is the probability assigned to class **C_i** and $\sum_{i=1}^{k} \hat{p}_i(x) = 1$

If we have only two classes, then $\hat{p}(x)$ denotes the estimated probability for the positive class.

Example: Fraud Detection

Regression

A regressor is a mapping

$$\hat{f} : \mathscr{X} \to \mathbb{R}$$

The regression learning problem is to learn a function estimator from examples $(\mathbf{x_i}, \mathbf{f(x_i)})$.

Note that we switched from a low-resolution target variable to one with infinite resolution. It's highly likely that some part of the target values in the examples is due to fluctuations that the model is not able to capture.

Since often the examples are noisy, it is reasonable to assume that the estimator is only intended to capture the general trend of the function.

Example: Stock Price

Association Rules

Let **$I=\{i_1, \ldots, i_n\}$** be a set of binary attributes called **items**.

Let **$D=\{t_1, \ldots, t_m\}$** be a set of **transactions** called the database.

Each transaction in D contains a subset of the items in I.

A rule is defined as an implication of the form:

$$\mathbf{X \Rightarrow Y}$$

where **$X,Y \subseteq I$**.

Example: Supermarket

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

$\{\text{butter, break}\} \Rightarrow \{\text{milk}\}$

Sequential Pattern Mining

Find **statistically relevant patterns** between data examples where the values are delivered in a **sequence**.

Usually with values that are discrete

e.g., strings

Example: DNA Sequences

Acknowledgements

- This course material takes inspiration from many sources that we **thank** sincerely!
- Andreas Müller, Columbia University, NYC
  - http://amueller.github.io/
- Roberto Esposito, Rosa Meo, UNITO, Turin
  - Introduction to data mining http://informatica.i-learn.unito.it/course/view.php?id=1434
- Ciro Cattuto, André Panisson, Laetitia Gauvin ISI, Turin
  - Data Mining, Statistical Modeling and Machine Learning
- We'll cite more sources during the course when used

Questions?

@rschifan

schifane@di.unito.it

http://www.di.unito.it/~schifane