



# Vizualisation

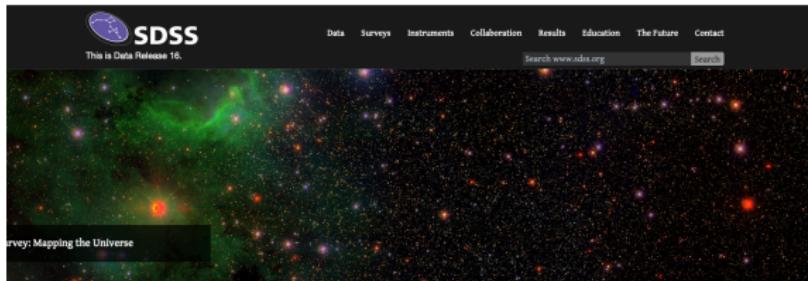
Part 1. The problem.

B9 - Visualisation of Massive Data

M-ALG-103

...

## └ Introduction



**EXPLORE OUR DATA**

[Go to Data Access](#)

**News**

[SDSS Press Releases](#)

[Don't judge a galaxy by its cover:](#)

**Figure:** SDSS <https://www.sdss.org/>

...

## └ Introduction

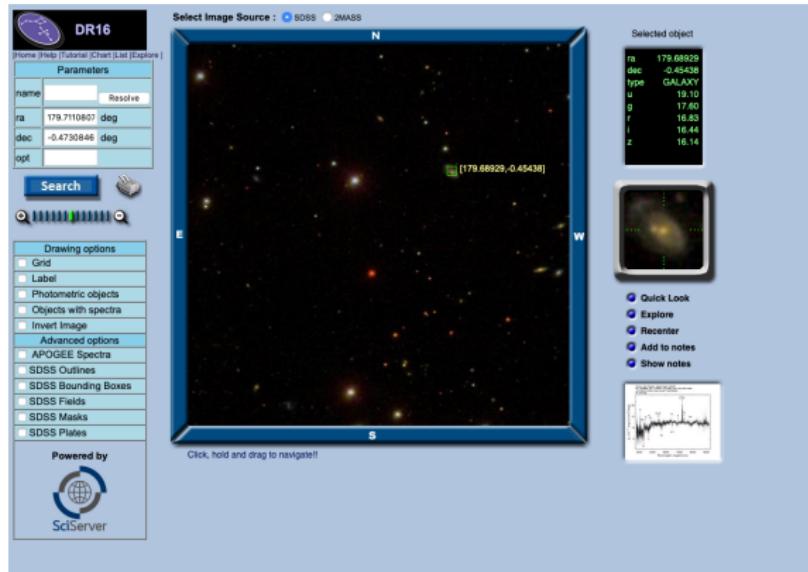


Figure: SDSS <https://www.sdss.org/>

The screenshot shows the official website for the NOAA National Centers for Environmental Information. The header features the NOAA logo and the text "NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION" along with the "NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION". A search bar is at the top right. Below the header, a navigation menu includes "Home", "Climate Information", "Data Access", "Customer Support", "Contact", and "About". A secondary navigation bar below the main menu includes "Datasets", "Search Tool", "Mapping Tool", "Data Tools", and "Help". The main content area has a white background and displays the title "Climate Data Online" in large, bold, black font. To the right of the title is a decorative graphic of a globe showing landmasses, water, and a sun. Below the title, a paragraph describes the service as providing free access to historical weather and climate data, including daily, monthly, seasonal, and yearly measurements. It also mentions radar data and Climate Normals, and the option to order hard copies. At the bottom of the page are four call-to-action boxes with icons: "Browse Datasets" (document icon), "Certify Orders" (keyhole icon), "Check Status" (info icon), and "Find Help" (question mark icon). Each box contains a brief description of its purpose.

Figure: NOAA dataset <https://www.ncdc.noaa.gov/cdo-web/>

## └ Introduction

X: 0m (PMS) C:\Users\...\\Downloads\USC00817.PRECIP.DLY sample.cdo.txt

STATION,STATION\_NAME,ELEVATION,LATITUDE,LONGITUDE,DATE,OVL,TMIN-NORMAL,TMAX-NORMAL,HTD-PROP-NORMAL

1 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180801,-33.145,2

2 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180802,-33.145,2

3 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180803,-33.145,2

4 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180804,-33.145,2

5 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180805,-33.145,2

6 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180806,-33.145,2

7 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180807,-41.148,10

8 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180807,-42.148,12

9 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180808,-43.148,13

10 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180809,-44.159,15

11 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180810,-45.159,17

12 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180811,-46.159,19

13 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180812,-47.159,21

14 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180813,-48.159,23

15 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180814,-49.159,25

16 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180815,-47.159,25

17 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180816,-47.159,25

18 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180817,-48.144,32

19 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180818,-49.144,34

20 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180819,-47.143,35

21 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180820,-47.144,37

22 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180821,-46.144,39

23 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180822,-46.146,41

24 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180823,-46.148,43

25 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180824,-46.148,44

26 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180825,-46.151,46

27 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180826,-43.152,47

28 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180827,-40.156,49

29 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180828,-41.156,49

30 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180829,-40.158,52

31 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180830,-37.162,54

32 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180901,-33.164,51

33 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180902,-33.166,53

34 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180903,-33.167,54

35 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180904,-29.171,6

36 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180905,-27.177,7

37 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180906,-25.178,8

38 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180907,-22.178,10

39 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180908,-20.181,11

40 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180909,-18.183,13

41 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180910,-16.186,13

42 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180911,-11.188,15

43 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180912,-10.188,15

44 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180913,-5.194,18

45 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180914,-1.196,19

46 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180915,-0.197,21

47 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180916,6.282,22

48 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180917,9.285,24

49 GHDNO USCOM817.PETERSBURG 2 M NO US,46.3,48.0555,-98.01,20180918,15.207,25

Normal NOAA\_DLY\_sample.cdo.txt

14% 0% 5/364 In ...

Figure: NOAA data sample. <https://www.ncdc.noaa.gov/cdo-web/>



...

## └ Introduction

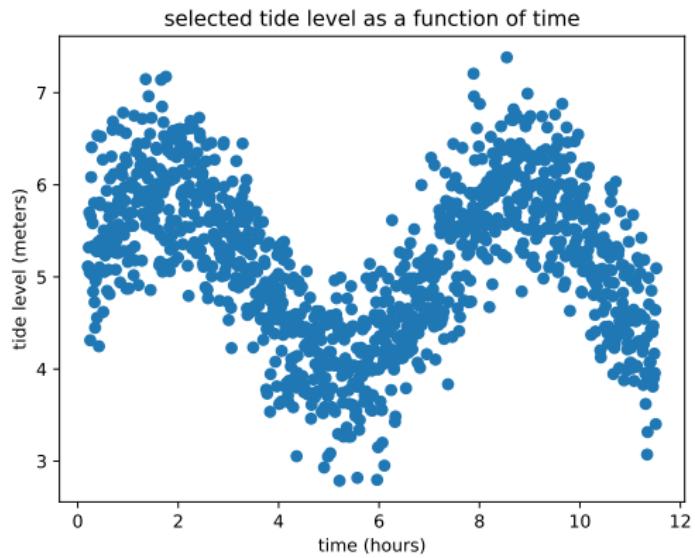


Figure: tide level

# Overview of the module

- Day 1 Position of the problem, formalisation, overfitting, Classic methods, Principal Component Analysis
- Day 2 Visualization platforms, advanced methods.

# Organisation of the module

- ▶ Theoretical course
- ▶ Small coding exercises,
- ▶ Paper + pen exercises

...

└ Introduction

# Organisation

- ▶ The exercices will be in python
- ▶ Clone the following repository :  
<https://github.com/nlehir/Visu.git>
- ▶ We will use **matplotlib**, **numpy**, **pandas**, **seaborn** (we can install them later)

...

└ Introduction

# Overview of day 1

## The problem of big data

Definition

Applications

Data visualization

## Data processing

Data modeling

Algorithm types

Stochastic processes and distributions

Fitting data

...

└ The problem of big data

  └ Definition

# The problem of big data

Let us start by defining what big data is.

...

└ The problem of big data

  └ Definition

## The problem of big data / data mining

**Definition 1:** "Data mining is the set of methods aiming at extracting information from large dataset, in a semi-automatic way. This information can take several forms : rules, tendencies in the data, structure in the data. The objective is that this information is helpful in order to make decisions based on the data."

...

└ The problem of big data

  └ Definition

## The problem of big data / data mining

**Definition 2:** "A useful way of characterizing big data is to understand the three Vs of big data: volume, variety and velocity. These characteristics encapsulate the qualities often associated with big data for example, large amounts of data, different types of data, and streaming or real-time data." [IBM, ]

...

- The problem of big data

- Definition

## The problem of big data / data mining

**Definition 2:** "A useful way of characterizing big data is to understand the three Vs of big data: volume, variety and velocity. These characteristics encapsulate the qualities often associated with big data for example, large amounts of data, different types of data, and streaming or real-time data. [...] But while these characteristics cover the key attributes of big data itself, we believe organizations need to consider an important fourth dimension: veracity. Inclusion of veracity as the fourth big data attribute emphasizes the importance of addressing and managing for the uncertainty inherent within some types of data." [IBM, ]

...

└ The problem of big data

  └ Definition

## The problem of big data / data mining

**Definition 3:** "Big Data is a technology to process high-volume, high-velocity, high-variety data or data-sets to extract intended data value and ensure high veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control."

[Olshannikova et al., 2016]

- ...
  - └ The problem of big data
  - └ Definition

## Scientific areas

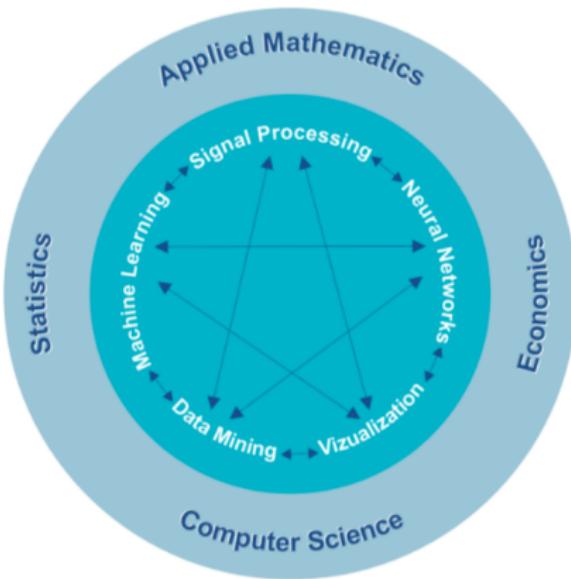


Figure: [Olshannikova et al., 2016]

...

└ The problem of big data

  └ Definition

## Big data sources

Where does big data come from ?

...

- └ The problem of big data

- └ Definition

## Big data sources

In Science : from measurement tools such as satellites, telescopes (astrophysics)

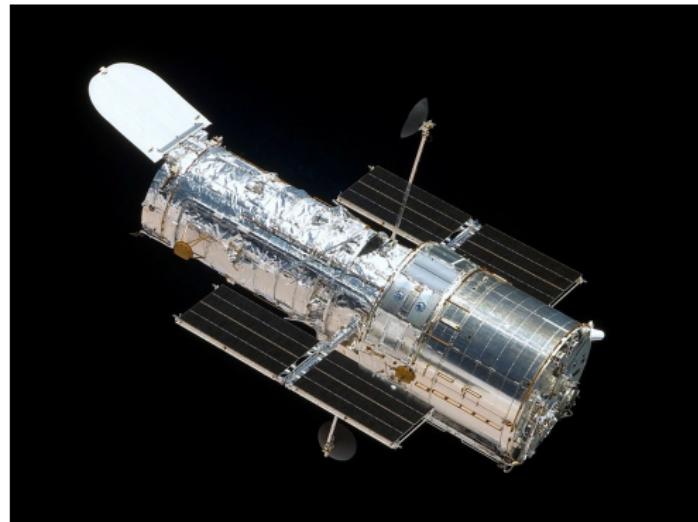


Figure: Hubble telescope (Wikipedia)

- ...
  - └ The problem of big data
  - └ Definition

# Big data sources

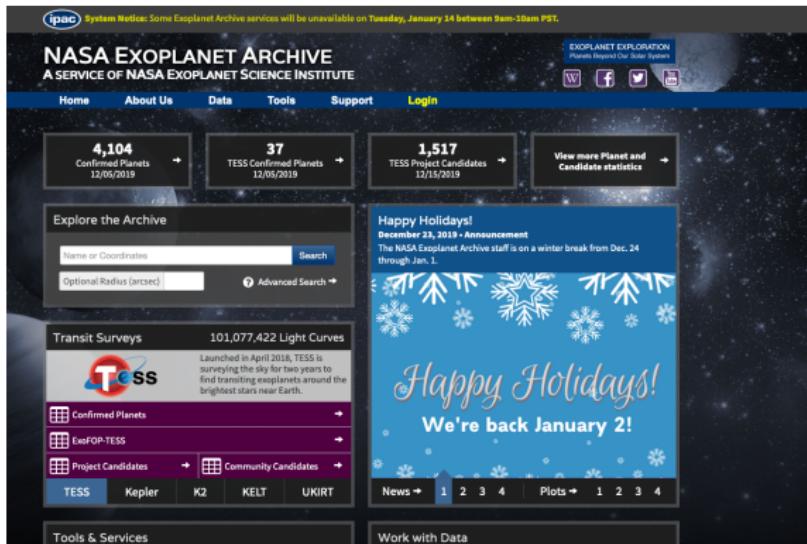
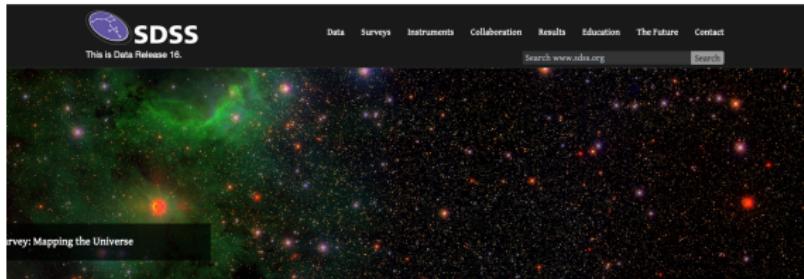


Figure: Nasa exoplanet archive

<https://exoplanetarchive.ipac.caltech.edu/>

- ...
  - └ The problem of big data
    - └ Definition

# Big data sources



A screenshot of the SDSS data access interface. It shows a circular icon representing a telescope or camera, followed by the text "EXPLORE OUR DATA" and a "Go to Data Access" button. To the right, there is a news section with "News" and "SDSS Press Releases" links, and a quote: "Don't judge a galaxy by its cover."

Figure: SDSS <https://www.sdss.org/>

...

└ The problem of big data

└ Definition

# Big data sources

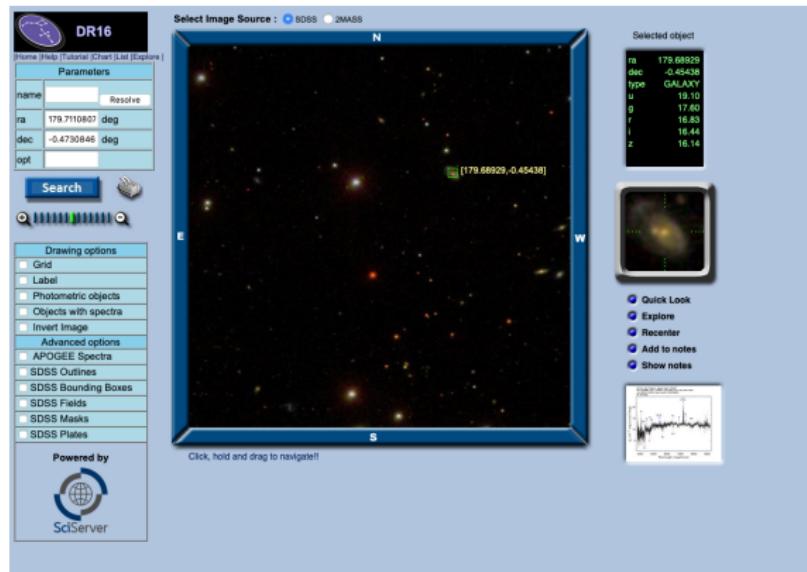


Figure: SDSS <https://www.sdss.org/>

# Big data sources

## Climatic data

The screenshot shows the homepage of the NOAA Climate Data Online (CDO) website. At the top, there is a navigation bar with links for Home, Climate Information, Data Access, Customer Support, Contact, and About. Below the navigation bar, a search bar is present. The main content area features a large heading "Climate Data Online" and a brief description of what the service offers. To the right of the text is a decorative graphic of a globe showing landmasses and clouds. Below the main heading are four buttons labeled "Browse Datasets", "Certify Orders", "Check Status", and "Find Help".

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION  
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Home Climate Information Data Access Customer Support Contact About

Search

Home > Climate Data Online

\_datasets | Search Tool | Mapping Tool | Data Tools | Help

## Climate Data Online

Climate Data Online (CDO) provides free access to NCDC's archive of global historical weather and climate data in addition to station history information. These data include quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, and degree days as well as radar data and 30-year Climate Normals. Customers can also order most of these data as [certified hard copies](#) for legal use.

Browse Datasets  
Browse documentation, samples, and links

Certify Orders  
Get orders certified for legal use (requires payment)

Check Status  
Check the status of an order that has been placed

Find Help  
Find answers to questions about data and ordering

Figure: NOAA dataset <https://www.ncdc.noaa.gov/cdo-web/>

...

## └ The problem of big data

### └ Definition

# Big data sources

## Climatic data

station	station_name	elevation	latitude	longitude	date	oly-min-normal	oly-max-normal	td2-prop-normal
0	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180101	.53-.145-.1
1	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180102	.53-.145-.1
2	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180103	.53-.145-.1
3	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180104	.53-.145-.1
4	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180105	.56-.145-.1
5	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180106	.58-.142-.7
6	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180107	.58-.142-.7
7	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180108	.58-.142-.7
8	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180109	.41-.148-.10
9	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180110	.42-.148-.12
10	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180111	.42-.148-.13
11	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180112	.43-.148-.13
12	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180113	.44-.139-.15
13	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180114	.45-.139-.17
14	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180115	.46-.139-.19
15	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180116	.46-.139-.21
16	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180117	.47-.139-.23
17	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180118	.47-.139-.25
18	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180119	.47-.140-.27
19	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180120	.47-.140-.30
20	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180121	.48-.140-.32
21	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180122	.48-.140-.34
22	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180123	.47-.143-.35
23	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180124	.47-.144-.37
24	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180125	.47-.144-.39
25	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180126	.46-.146-.41
26	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180127	.46-.146-.43
27	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180128	.46-.146-.44
28	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180129	.44-.151-.46
29	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180130	.41-.152-.47
30	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180131	.41-.152-.49
31	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180201	.41-.156-.50
32	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180202	.40-.158-.52
33	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180203	.40-.158-.54
34	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180204	.37-.162-.55
35	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180205	.33-.164-.55
36	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180206	.33-.164-.55
37	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180207	.33-.169-.4
38	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180208	.29-.171-.6
39	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180209	.27-.171-.7
40	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180210	.26-.176-.8
41	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180207	.22-.178-.10
42	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180208	.20-.180-.11
43	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180209	.19-.180-.11
44	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180210	.19-.180-.11
45	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180211	.11-.188-.15
46	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180212	.11-.188-.15
47	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180213	.11-.188-.15
48	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180214	.5-.194-.18
49	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180215	.5-.194-.18
50	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180216	.5-.202-.22
51	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180217	.9-.205-.24
52	GHCND:USC00378027	PETERSBURG 2	N 46.3	W -116.3	48.0555	-98.01	20180218	.13-.207-.25

Figure: NOAA data sample. <https://www.ncdc.noaa.gov/cdo-web/>

...

└ The problem of big data

└ Definition

# Big data sources

## Bioinformatics

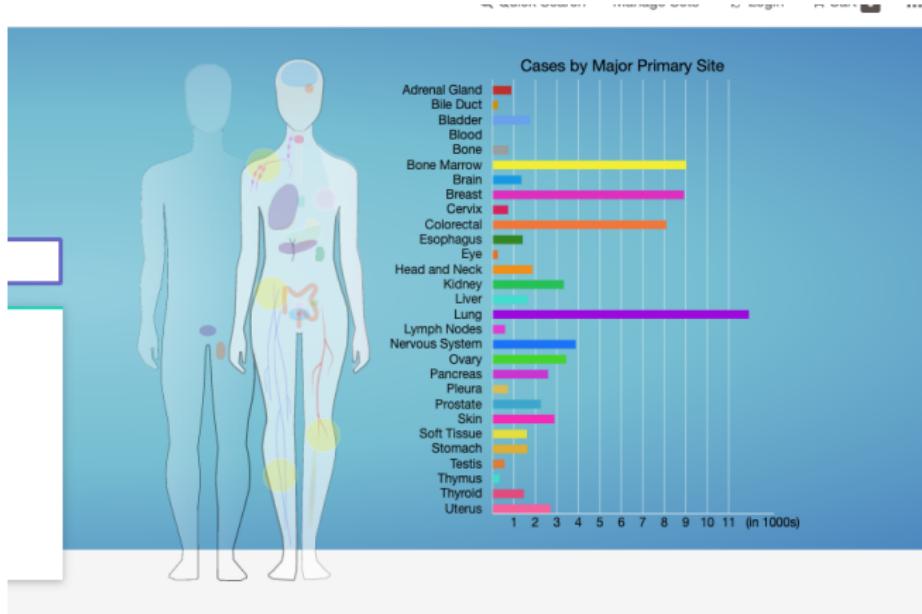


Figure: National Cancer Institute <https://portal.gdc.cancer.gov/>

...

- The problem of big data

- Definition

# Big data sources

## Bioinformatics

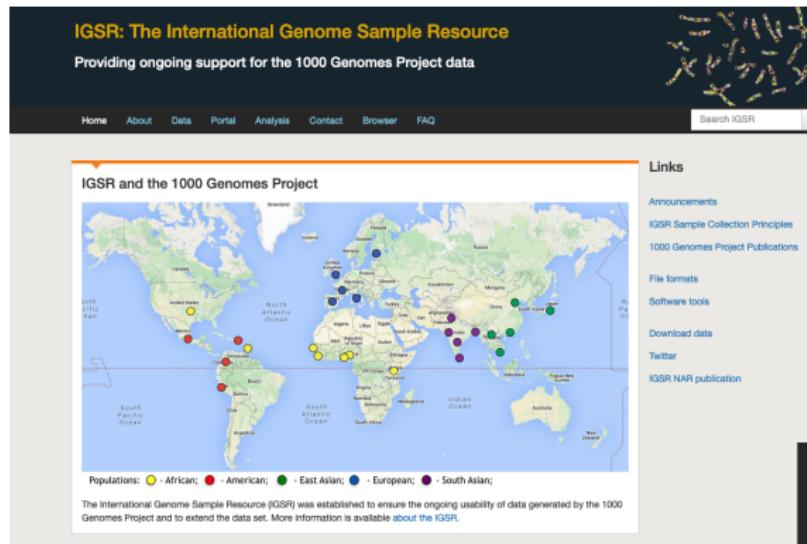


Figure: 1000 Genome project

<https://www.internationalgenome.org/>

# Big data sources

The web.

The screenshot shows a search results page from the DuckDuckGo search engine. The search query 'europa universalis' is entered in the search bar. Below the search bar, there are tabs for 'Web', 'Images', 'Vidéos', and 'Actualités'. A dropdown menu shows 'France' selected under 'Filtre Parental' and 'Strict' under 'À tout moment'. The results are listed in a grid format:

- Europa Universalis — Wikipédia**  
Europe Universalis est un jeu vidéo de grande stratégie développé par Paradox Development Studio et sorti en 2000. Il est inspiré d'un jeu de plateau éponyme ...  
[W https://fr.wikipedia.org/wiki/Europa\\_Universalis](https://fr.wikipedia.org/wiki/Europa_Universalis)
- Europa Universalis IV sur PC - jeuxvideo.com**  
Europe Universalis IV sur PC : retrouvez toutes les informations, les tests, les vidéos et actualités du jeu sur tous ses supports. Europa Universalis IV sur PC ...  
[J jeuxvideo.com/jeux/pc/00046149-europa-universalis-iv.htm](https://www.jeuxvideo.com/jeux/pc/00046149-europa-universalis-iv.htm)
- Europa Universalis IV**  
【送別兼祝】ノゾム 税式用【クローバルエリート】MG=(金属製／84cm／900g)上) シルバーライン(2th21140)  
父の日 sale C1806 対応 ...  
[e europauniversalis4.com](http://europauniversalis4.com)
- Europa Universalis 4 — Wikipédia**  
Europa Universalis 4 (stylisé Europa Universalis IV) est un jeu de grande stratégie historique développé par la société suédoise Paradox Development Studio et ...  
[W https://fr.wikipedia.org/wiki/Europa\\_Universalis\\_4](https://fr.wikipedia.org/wiki/Europa_Universalis_4)

On the right side of the results, there is a sidebar with the title 'Europa Universalis' and a snippet of text about the game, followed by a link 'Plus sur Wikipedia (FR)'.

Figure: Duckduckgo web browser

# Big data sources

## The web.

[Home](#) [Components](#) [Support](#) [About](#)

[ClueWeb09](#) [How to Get It](#) [Dataset Details](#) [Related Data](#) [Online Services](#) [Indexing with Indri](#) [Wiki & Email](#) [FAQ](#)

### The ClueWeb09 Dataset

The ClueWeb09 dataset was created to support research on information retrieval and related human language technologies. It consists of about 1 billion web pages in ten languages. The dataset is used by several tracks of the [TREC](#) conference.

#### Dataset Specifications

**Web Pages:**

- 1,040,809,705 web pages, in 10 languages
- 5 TB, compressed. (25 TB, uncompressed.)

See the Record Counts Section on the [Dataset Information](#) and [Sample Files](#) page for detailed information on the distribution of records and languages.

**Web Graph:**

- **Entire Dataset:**
  - Unique URLs: 4,780,950,903 (325 GB uncompressed, 105 GB compressed)
  - Total Outlinks: 7,944,351,835 (71 GB uncompressed, 24 GB compressed)
- **TREC Category B (first 50 million English pages)**
  - Unique URLs: 428,136,613 (30 GB uncompressed, 10 GB compressed)
  - Total Outlinks: 454,075,638 (3 GB uncompressed, 1 GB compressed)

The web graph for both the entire dataset and for the TREC Category B dataset (first 50 million English pages) is complete. We are in the process of retrieving the data and performing information on how the crawl progressed is also available.

**Dataset Distribution:**

The ClueWeb09 dataset and subsets are distributed in several different ways.

- **Full, 4 x 1.5TB:** The full dataset is distributed as tarred/gzipped files on four 1.5 terabyte (TB) hard disk drives, in Linux ext3 format. The physical drives are standard SATA 3 (compatible with any SATA/300 interface, including external USB to SATA/300 enclosures).

Figure: Clueweb dataset <https://lemurproject.org/clueweb09.php/>

# Big data sources

## Natural Language Processing

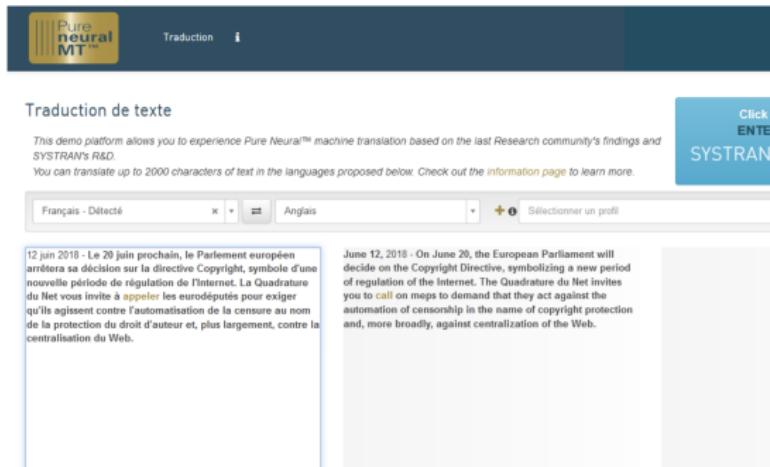


Figure: Text translation (SYSTRAN) [https://www.systransoft.com/  
fr/systran/technologie/pure-neural-machine-translation/](https://www.systransoft.com/fr/systran/technologie/pure-neural-machine-translation/)

...

- └ The problem of big data

- └ Definition

## Big data sources

Handwritten digits.



Figure: MNIST dataset [LeCun and Cortes, 2010]

...

└ The problem of big data

└ Definition

# Big data sources

Lexicometry.

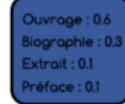


FIGURE : document

FIGURE :  
topics

Figure: Statistics on texts

# Big data sources

Text, audio, video.

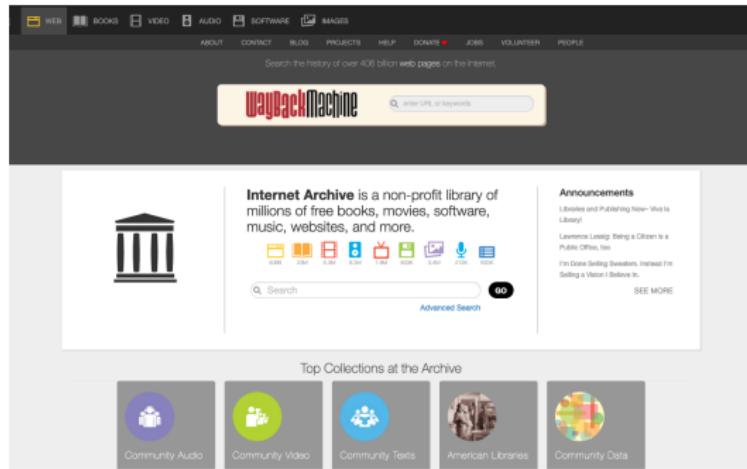


Figure: Archive dataset <https://archive.org/>

...

- └ The problem of big data

- └ Definition

## Big data sources

Image and video.

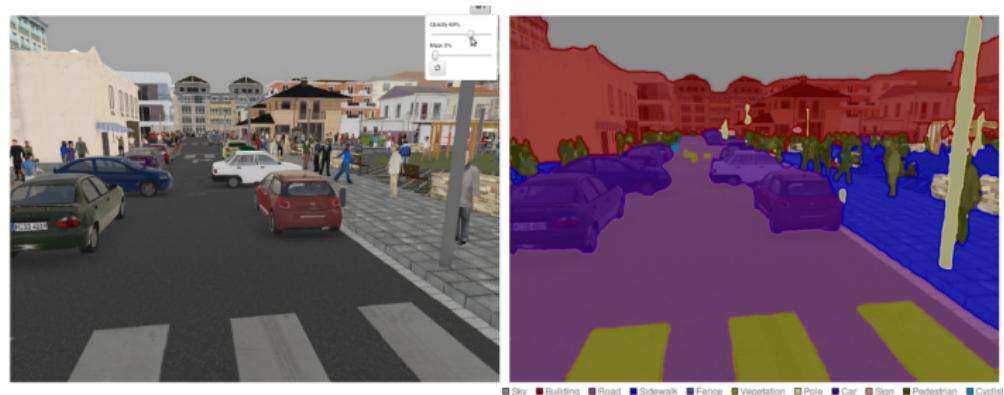


Figure: Source : Nvidia developer blog  
<https://devblogs.nvidia.com/>

...

- └ The problem of big data

- └ Definition

## Big data sources

Financial data (time series)

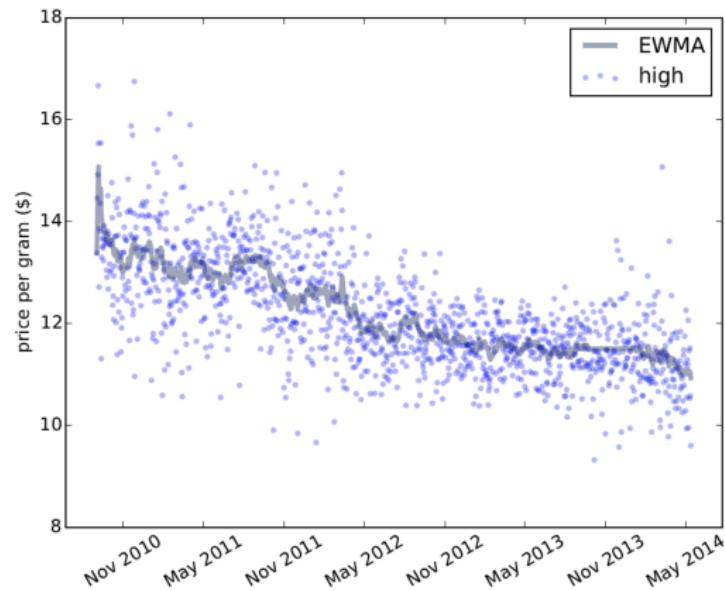


Figure: Price of some plant

...

- └ The problem of big data

- └ Definition

# Big data sources

## Game data



Figure: Europa Universalis

[https://fr.wikipedia.org/wiki/Europa\\_Universalis](https://fr.wikipedia.org/wiki/Europa_Universalis)

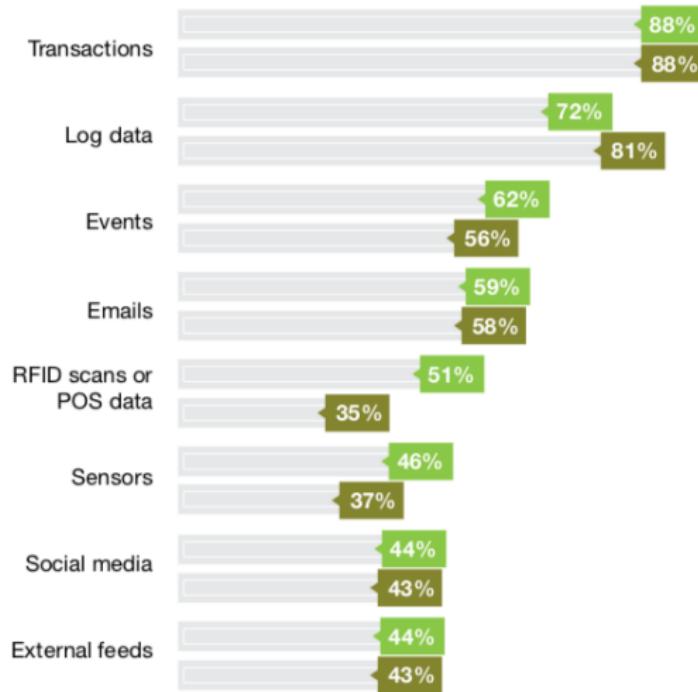
...

- └ The problem of big data

- └ Definition

# Big data sources

Variety



...

└ The problem of big data

  └ Definition

## Orders of magnitude

Let us discuss the size of data.

Order of magnitude means "Ordre de grandeur".

...

└ The problem of big data

  └ Definition

## Orders of magnitude

A 3 minute mp3 file in 320 kbps has a size of approximately ?

# Orders of magnitude

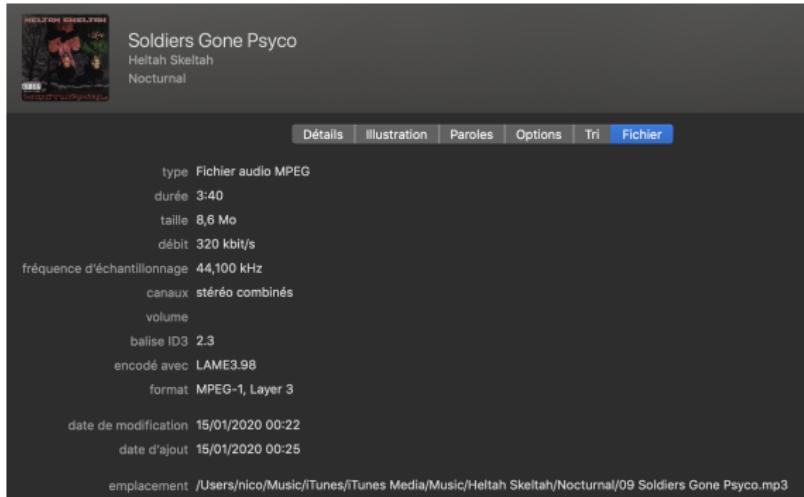


Figure: Approximately 8 Mo

...

└ The problem of big data

  └ Definition

# Orders of magnitude

## MP3 files

```
→ Kan Cassette Vol. 2 du -hs * | gsort -h
1,6M  1-03 Plastic Pine.mp3
3,0M  1-07 Procedure Topic.mp3
3,0M  1-12 Bobby & Marley.mp3
4,0M  1-04 Boom Bop Break.mp3
4,0M  1-05 88 Intellectual.mp3
4,0M  1-10 Early Morning Cruising.mp3
4,0M  1-13 Supreme Motive.mp3
5,0M  1-02 Globet Loner.mp3
5,0M  1-06 DJ Pooh Groove.mp3
5,0M  1-11 Steam.mp3
5,0M  1-14 Heaven.mp3
7,0M  1-01 Rtw (Ill Kid).mp3
7,0M  1-08 Soul Silence Phat (Many Vibrations).mp3
7,0M  1-09 Grand Kan Grooviest.mp3
→ Kan Cassette Vol. 2 └
```

...

└ The problem of big data

  └ Definition

# Orders of magnitude

## Lossless compressed files

```
→ The Infamous du -hs * | gsort -h
6,0M 05 [Just Step Prelude].m4a
6,0M 12 [The Grave Prelude].m4a
12M 02 [The Infamous Prelude].m4a
26M 03 Survival of the Fittest.m4a
28M 06 Give Up the Goods (Just Step).m4a
29M 01 The Start of Your Ending (41st Side).m4a
32M 13 Cradle to the Grave.m4a
33M 04 Eye for an Eye (Your Beef Is Mines).m4a
33M 10 Q.U. - Hectic.m4a
33M 11 Right Back at You.m4a
34M 08 Up North Trip.m4a
35M 07 Temperature's Rising.m4a
35M 14 Drink Away the Pain (Situations).m4a
37M 09 Trife Life.m4a
38M 15 Shook Ones, Part II.m4a
39M 16 Party Over.m4a
→ The Infamous █
```

...

└ The problem of big data

  └ Definition

# Orders of magnitude

PDF slides for this presentation : Approximately 20 Mo

...

└ The problem of big data

  └ Definition

## Orders of magnitude

2,5 hours MKV movie ?

...

└ The problem of big data

  └ Definition

## Orders of magnitude

2,5 hours MKV movie ? Around 3 Go.

...

└ The problem of big data

  └ Definition

## Orders of magnitude

Desktop computer harddrive : 1To.

...

└ The problem of big data

  └ Definition

# Orders of magnitude

Big datasets ?

# Orders of magnitude

The screenshot shows the homepage of the ClueWeb09 dataset. At the top, there is a navigation bar with links for Home, Components, Support, About, ClueWeb09, How to Get It, Dataset Details, Related Data, Online Services, Indexing with Indri, Wiki & Email, and FAQ. Below the navigation bar, the title "The ClueWeb09 Dataset" is displayed in bold. A sub-section titled "Dataset Specifications" contains information about the dataset's size and language distribution.

The ClueWeb09 dataset was created to support research on information retrieval and related human language technologies. It consists of about 1 billion web pages in ten languages. The dataset is used by several tracks of the TREC conference.

## Dataset Specifications

### Web Pages:

- 1,040,809,705 web pages, in 10 languages
- 5 TB, compressed, (25 TB, uncompressed.)

See the Record Counts Section on the [Dataset Information](#) and [Sample Files](#) page for detailed information on the distribution of records and languages.

### Web Graph:

- **Entire Dataset:**
  - Unique URLs: 4,780,950,903 (325 GB uncompressed, 105 GB compressed)
  - Total Outlinks: 7,944,351,835 (71 GB uncompressed, 24 GB compressed)
- **TREC Category B (first 50 million English pages)**
  - Unique URLs: 428,136,613 (30 GB uncompressed, 10 GB compressed)
  - Total Outlinks: 454,075,638 (3 GB uncompressed, 1 GB compressed)

The web graph for both the entire dataset and for the TREC Category B dataset (first 50 million English pages) is complete. We are in the process of retrieving the data and performing information on how the crawl progressed is also available.

### Dataset Distribution:

The ClueWeb09 dataset and subsets are distributed in several different ways.

- **Full, 4 x 1.5TB:** The full dataset is distributed as tarred/gzipped files on four 1.5 terabyte (TB) hard disk drives, in Linux ext3 format. The physical drives are standard SATA 3 (compatible with any SATA/300 interface, including external USB to SATA/300 enclosures).

**Figure:** a **105 Go** <https://lemurproject.org/clueweb09.php/>

## Big data sources

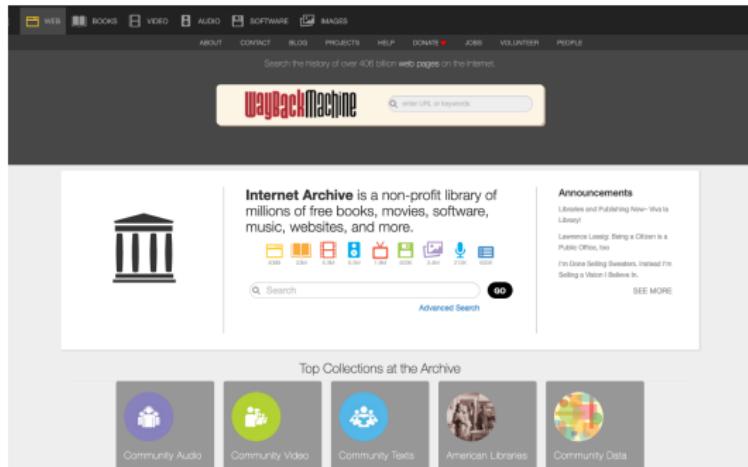


Figure: 20 To <https://archive.org/>

...

└ The problem of big data

└ Definition

# Orders of magnitude

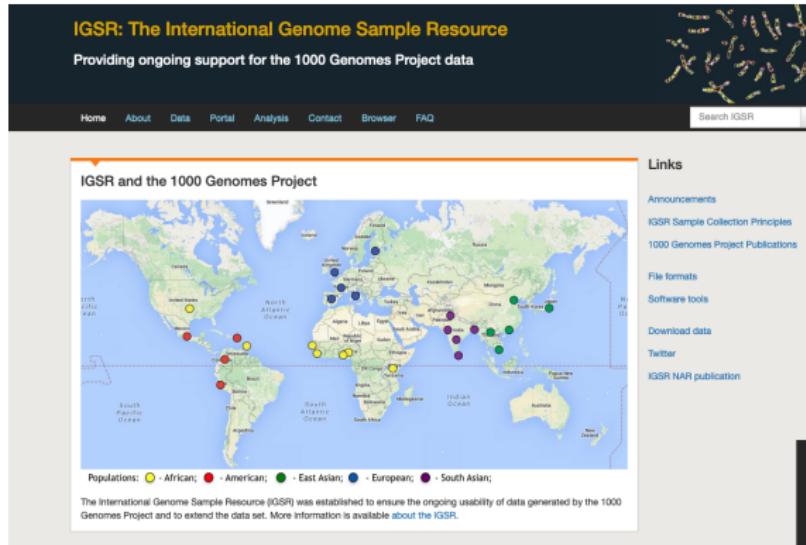


Figure: 260 To <https://www.internationalgenome.org/>

# Orders of magnitude

## Volume

Greater than 100 PB



Greater than 1 PB



Greater than 100 TB



Greater than 10 TB



Greater than 1 TB



Figure: [IBM, ]

...

└ The problem of big data

  └ Definition

## Orders of magnitude

**Exercice 1: Computing orders of magnitude** How many 8 Mo MP3 tracks would a music library of the size of the 1000 Genome project contain ?

And for the Archive project ?

1 Byte = 1 Octet

...

- The problem of big data

- Definition

## Orders of magnitude

Exercice 1: **Sizes in bytes** Fill the following array:

File	Size in Bytes	Number of MP3 tracks
MP3 track	8M	1
Big PDF file	15M	?
2.5 hours movie	3G	?
ClueWeb dataset	105G	?
1 TB hard drive	1T	?
Archive dataset	20T	?
1000 Genome dataset	260T	?

...

└ The problem of big data

└ Definition

## Orders of magnitude

File	Size in Bytes	Number of MP3 tracks
MP3 track	8M	1
Big PDF file	15M	1.8
2.5 hours movie	3G	375
ClueWeb dataset	105G	$\simeq 13000$
1 TB hard drive	1T	125000
Archive dataset	20T	2.5M
1000 Genome dataset	260T	32.5 M

...

└ The problem of big data

  └ Definition

## Amount of data

"According to IDC, data volumes have grown exponentially, and by 2020 the number of digital bits will be comparable to the number of stars in the universe. As the size of bits geminates every two years, for the period from 2013 to 2020 worldwide data will increase from 4.4 to 44 zettabytes." [Olshannikova et al., 2016]

...

└ The problem of big data

└ Applications

# Who uses big data ?

...

└ The problem of big data

└ Applications

## Who uses big data ?

Mostly scientists and companies.

...

└ The problem of big data

└ Applications

## Who uses big data ?

Mostly scientists and companies.

There is now a market of trading data.

# Usage of big data

## Big data activities

Pilot and implementation underway



Planning big data activities



Have not begun big data activities



■ Large

■ Midmarket

Figure: Big data activites in organizations surveyed by IBM [IBM, ]

...

└ The problem of big data

└ Applications

## Applications of big data

Let us discuss some applications of big data.

...

└ The problem of big data

└ Applications

## Example 1 : bioinformatics

In bioinformatics, one can study data coming from genes in order to try to predict the probability of developing a given disease.

...

└ The problem of big data

  └ Applications

## Example 2 : banking

A bank loses an unusual number of clients in a given period. It can then analyze whether there is some characteristic that is common between those clients, and then may adapt its offer to those.

### Example 3 : web marketing



**Faster, safer, and smarter browsing**

Ghostery helps you browse smarter by giving you control over ads and tracking technologies to speed up page loads, eliminate clutter, and protect your data.

[Install Ghostery](#) [Learn More](#)

Smart Blocking automatically optimizes page performance as you browse.

Dynamic UI includes multiple displays and detailed tracker dashboard.

Enhanced Anti-Tracking anonymizes your data to further protect your privacy.

This site uses cookies  
You are not being tracked since your browser is reporting that you do not want to. This is a setting of your browser as you won't be able to opt-in until you disable the Do Not Track feature.

## Figure: Ghostery

...

└ The problem of big data

└ Applications

## Other applications

In the future there might be more, unexpected applications.

...

└ The problem of big data

  └ Data visualization

## Data visualization

Those datasets are often so large that it is extremely hard to analyze them.

...

└ The problem of big data

  └ Data visualization

## Data visualization

Those datasets are often so large that it is extremely hard to analyze them.

Depending on the information sought in the dataset, it is sometimes helpful to make a **visualization** of the dataset.

...

└ The problem of big data

  └ Data visualization

## Data visualization

Those datasets are often so large that it is extremely hard to analyze them.

Depending on the information sought in the dataset, it is sometimes helpful to make a **visualization** of the dataset.

Visualization might be a clue to understand structure or tendencies in the data, or to identify anomalies.

...

- └ The problem of big data
  - └ Data visualization

## Big data visualization constraints

The data are often :

- ▶ dynamic : they evolve a lot and new data is constantly added to the dataset
- ▶ noisy, biased, they contain errors
- ▶ heterogeneous and unstructured

...

- └ The problem of big data
- └ Data visualization

## Big data visualization constraints

The data are often :

- ▶ dynamic : they evolve a lot and new data is constantly added to the dataset
- ▶ noisy, biased, they contain errors
- ▶ heterogeneous and unstructured

In that respect, a visualization system should:

- ▶ offer the possibility to interact with online data (without preprocessing)
- ▶ scale to a large amount of data
- ▶ propose images that are not overloaded with information, taking our cognitive properties into account.

...

└ The problem of big data

  └ Data visualization

## Big data visualization constraints

And all those constraints should be implemented in a system with limited resources.

...

└ The problem of big data

  └ Data visualization

## Extracting information from the data

Most of the time, there are several ways to **visualize** a given dataset. This means that the choice of representation should be justified and will depend on the objectives of the processing.

...

└ The problem of big data

└ Data visualization

## Example 1



Figure: Sky view of Meaux : what information could we extract from it ?

...

└ The problem of big data

└ Data visualization

## Example 1



Figure: Sky view of Meaux : what information could we extract from it ?

- option 1 Display the roads by analyzing lines in the image
- option 2
- option 3

...

└ The problem of big data

└ Data visualization

## Example 1



Figure: Sky view of Meaux : what information could we extract from it ?

option 1 Display the roads by analyzing lines in the image

option 2 Separate water from the land by analyzing the colors

option 3

...

└ The problem of big data

└ Data visualization

## Example 1



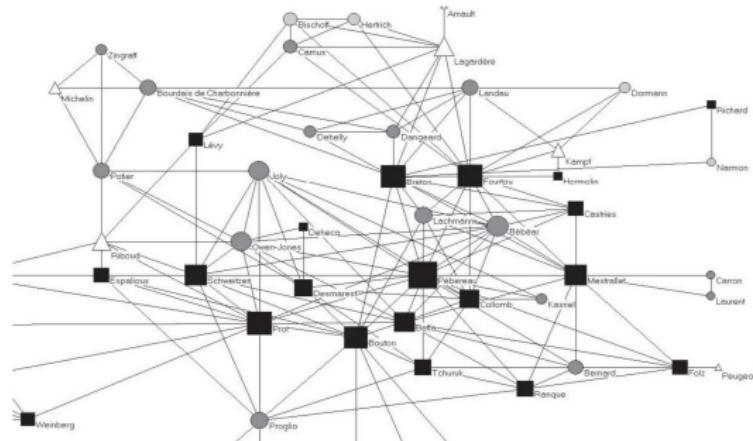
Figure: Sky view of Meaux : what information could we extract from it ?

option 1 Display the roads by analyzing lines in the image

option 2 Separate water from the land by analyzing the colors

option 3 Display the elevation based on the color of the roofs

## Example 2



## Figure: Graph

- └ The problem of big data
  - └ Data visualization

### Example 3

However, most of the time, we start with raw data.

## Figure: High dimensional data

...

- └ The problem of big data

- └ Data visualization

## Example 4

Actually, even with a graph, extracting meaning is not obvious.

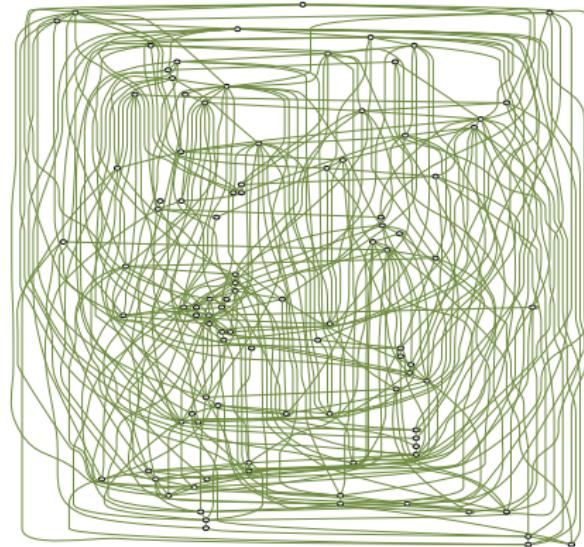


Figure: Dense graph

...

- └ The problem of big data
  - └ Data visualization

## Conclusion

- ▶ In some cases, it is non trivial to propose a useful or meaningful representation of the data.

...

- └ The problem of big data
  - └ Data visualization

## Conclusion

- ▶ In some cases, it is non trivial to propose a useful or meaningful representation of the data.
- ▶ Especially when the dataset is large.

...

└ Data processing

  └ Data modeling

## Data modeling

Given a unknown dataset, we must **process** it in order to extract information from it.

...

└ Data processing

  └ Data modeling

## Data modeling

Given an unknown dataset, we must **process** it in order to extract information from it.

Before studying visualization, we must discuss the general processing of data.

...

└ Data processing

  └ Data modeling

## Position of the problem

- ▶ The examples we encountered seem to be very different
- ▶ In each situation, we can follow these steps :
  - ▶ define and acquire a dataset
  - ▶ fix an objective

...

└ Data processing

  └ Data modeling

## Position of the problem

- ▶ The examples we encountered seem to be very different
- ▶ In each situation, we can follow these steps :
  - ▶ define and acquire a dataset
  - ▶ fix an objective
  - ▶ choose a model

...

└ Data processing

  └ Data modeling

## Position of the problem

- ▶ The examples we encountered seem to be very different
- ▶ In each situation, we can follow these steps :
  - ▶ define and acquire a dataset
  - ▶ fix an objective
  - ▶ choose a model
  - ▶ identify algorithms in order to optimize the model (fitting)

...

└ Data processing

  └ Data modeling

## Position of the problem

- ▶ The examples we encountered seem to be very different
- ▶ In each situation, we can follow these steps :
  - ▶ define and acquire a dataset
  - ▶ fix an objective
  - ▶ choose a model
  - ▶ identify algorithms in order to optimize the model (fitting)
  - ▶ evaluate performances

...

└ Data processing

  └ Algorithm types

## Several paradigms exist

- ▶ During the course we will focus on the two main data processing paradigms :

...

└ Data processing

  └ Algorithm types

## Several paradigms exist

- ▶ During the course we will focus on the two main data processing paradigms : **supervised learning** and **unsupervised learning**

...

└ Data processing

  └ Algorithm types

## Several paradigms exist

- ▶ During the course we will focus on the two main data processing paradigms : **supervised learning** and **unsupervised learning**
- ▶ In AI, there is a third important paradigm, called **Reinforcement Learning**, but it is aimed at designing **agents** that perform **actions** in an environment (such as a **robot**)

## Several paradigms exist

- ▶ During the course we will focus on the two main data processing paradigms : **supervised learning** and **unsupervised learning**
- ▶ In AI, there is a third important paradigm, called **Reinforcement Learning** (RL), but it is aimed at designing **agents** that perform **actions** in an environment (such as a **robot**)
- ▶ The paradigms can be mixed : a Reinforcement Learning system can have a supervised learning part.

...

└ Data processing

  └ Algorithm types

# Supervised learning

## Observations:

- ▶ Empirical data points  $(\tilde{x}_1, \dots, \tilde{x}_n)$

...

└ Data processing

└ Algorithm types

# Supervised learning

## Observations:

- ▶ Empirical data points  $(\tilde{x}_1, \dots, \tilde{x}_n)$
- ▶ Each data point  $\tilde{x}_i$  is associated with a **label**  $\tilde{y}_i$ , so we also have a set of labels  $(\tilde{y}_1, \dots, \tilde{y}_n)$

...

- └ Data processing

- └ Algorithm types

# Supervised learning

## Observations:

- ▶ Empirical data points  $(\tilde{x}_1, \dots, \tilde{x}_n)$
- ▶ Each data point  $\tilde{x}_i$  is associated with a **label**  $\tilde{y}_i$ , so we also have a set of labels  $(\tilde{y}_1, \dots, \tilde{y}_n)$

**Objective:** learn to predict the label  $\tilde{y}_p$  of a new data point  $\tilde{x}_p$ .

# Supervised learning

## Observations:

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$
- ▶ Each data point  $\tilde{x}_i$  is associated with a **label**  $\tilde{y}_i$ , so we also have a set of labels  $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)$

**Objective:** learn to predict the label  $\tilde{y}_p$  of a new data point  $\tilde{x}_p$ .

## Hypotheses:

- ▶ The observations  $\tilde{x}$  stem from a **stochastic process**  $x$
- ▶ The observations  $\tilde{y}$  stem from a **stochastic process**  $y$
- ▶ There is an **underlying relation**  $y = f(x)$

## Supervised learning

### Observations:

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$
- ▶ Each data point  $\tilde{x}_i$  is associated with a **label**  $\tilde{y}_i$ , so we also have a set of labels  $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)$

**Objective:** learn to predict the label  $\tilde{y}_p$  of a new data point  $\tilde{x}_p$ .

### Hypotheses:

- ▶ The observations  $\tilde{x}$  stem from a **stochastic process**  $x$
- ▶ The observations  $\tilde{y}$  stem from a **stochastic process**  $y$
- ▶ There is an **underlying relation**  $y = f(x)$

We then want to build a **function**  $\tilde{f}$  based on  $\tilde{x}$  and  $\tilde{y}$  such that  $\tilde{f}$  is a good approximation of  $f$ .

# Supervised learning

## Observations:

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$
- ▶ Each data point  $\tilde{x}_i$  is associated with a **label**  $\tilde{y}_i$ , so we also have a set of labels  $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)$

**Objective:** learn to predict the label  $\tilde{y}'$  of a new data point  $\tilde{x}'$ .

## Hypotheses:

- ▶ The observations  $\tilde{x}$  stem from a **stochastic process**  $x$
- ▶ The observations  $\tilde{y}$  stem from a **stochastic process**  $y$
- ▶ There is an **underlying relation**  $y = f(x)$

We then want to build a **function**  $\tilde{f}$  based on  $\tilde{x}$  and  $\tilde{y}$  such that  $\tilde{f}$  is a good approximation of  $f$ .

So we can predict  $\tilde{y}' = \tilde{f}(\tilde{x}')$  for a new sample  $\tilde{x}'$ .

...

└ Data processing

  └ Algorithm types

## Supervised learning

Remark : we can see  $\tilde{x}$  as empirical **inputs** and  $\tilde{y}$  as empirical **outputs**.

...

- └ Data processing

- └ Algorithm types

## Tide level

We will apply this concept to the analysis of the tide level.



...

└ Data processing

  └ Algorithm types

## Tide Level

We have the tide level in meters as a function of time in hours.

...

└ Data processing

  └ Algorithm types

## Tide Level

We have the tide level in meters as a function of time in hours.  
Our goal will be to **predict** the tide level as a function of time.

...

└ Data processing

  └ Algorithm types

## Tide level

### Exercice 2 : Plotting the data

**cd line\_graph/**. We first need to plot the data we have. Please use **analyze\_data.py** in order plot the level as a function of time in a **line graph**. Try several **markers**.

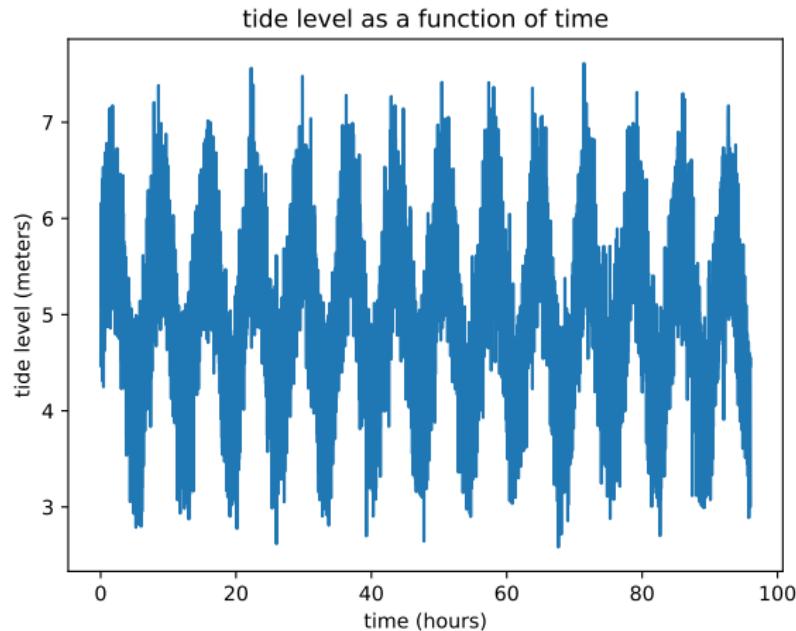
...

└ Data processing

└ Algorithm types

# Tide level

## Exercice 2 : Plotting the data



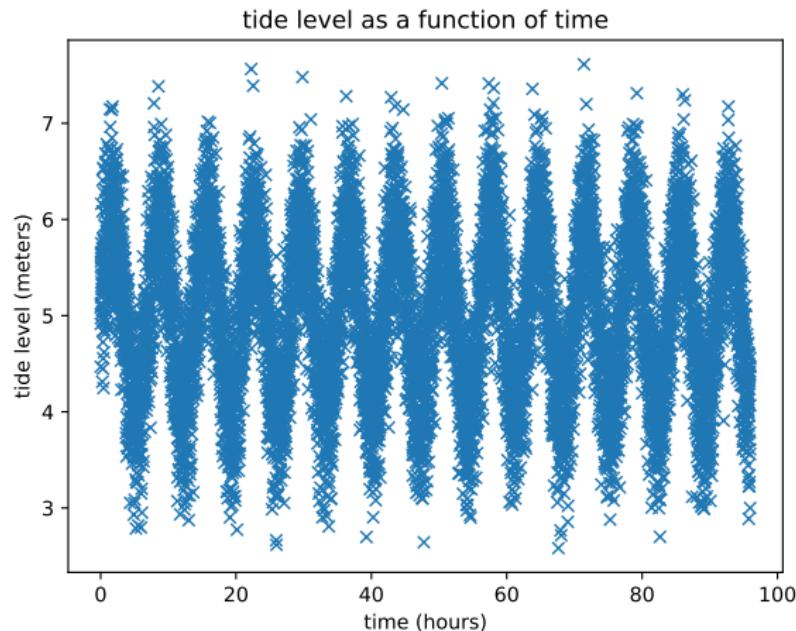
...

└ Data processing

└ Algorithm types

# Tide level

## Exercice 2 : Plotting the data



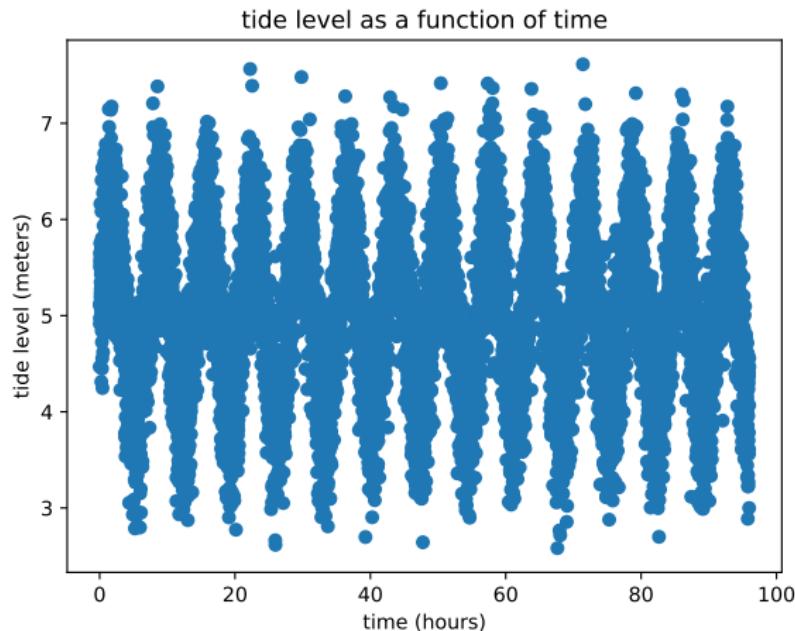
...

└ Data processing

└ Algorithm types

# Tide level

## Exercice 2 : Plotting the data



...

└ Data processing

  └ Algorithm types

## Tide level

Exercice 2 : **Plotting the data** Select indexes in order to graphically see the period of oscillation.

...

└ Data processing

└ Algorithm types

## Tide level

**Exercice 2: Plotting the data** Select indexes in order to graphically see the period of oscillation.

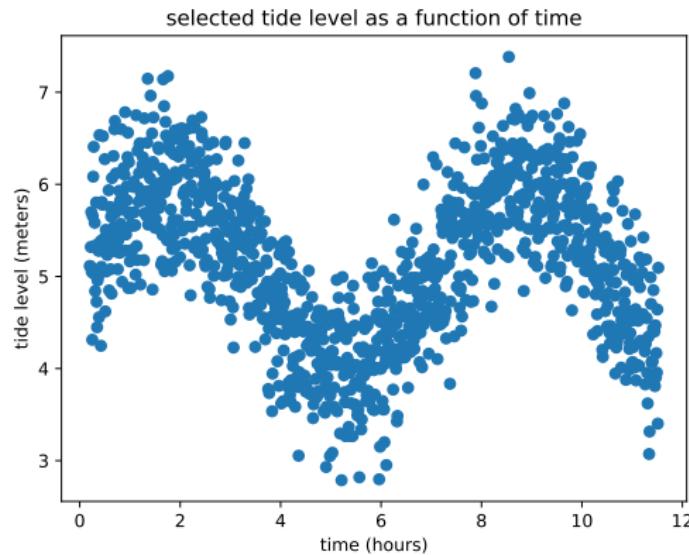


Figure: tide level

...

└ Data processing

  └ Algorithm types

## Tide level

Exercice 2 : **Plotting the data** We could also have manually selected the plot limits.

...

└ Data processing

└ Algorithm types

## Tide level

**Exercice 2: Plotting the data** We could also have manually selected the plot limits.

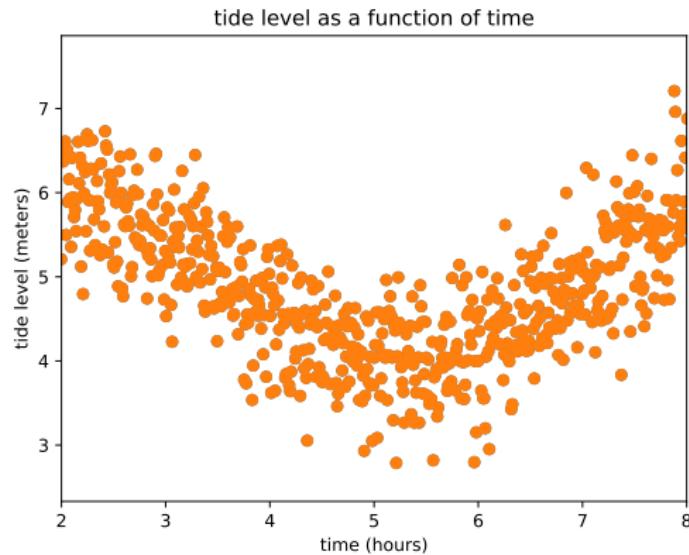


Figure: tide level

...

└ Data processing

  └ Algorithm types

## Tide level

Exercice 2 : **Finding a function** We would like to model the tide level as a function  $f$  of the time. What would you suggest ?

...

└ Data processing

  └ Algorithm types

## Tide level

**Exercice 2 : Finding a function** We would like to model the tide level as a function  $f$  of the time. What would you suggest ?  
We could use a sine function. What would the **parameters** be ?

## Tide level

**Exercice 2: Finding a function** We would like to model the tide level as a function  $f$  of the time. What would you suggest ? We could use a sine function. The parameters are:

- ▶ Amplitude
- ▶ pulsation ( analog of frequency)
- ▶ phase
- ▶ offset

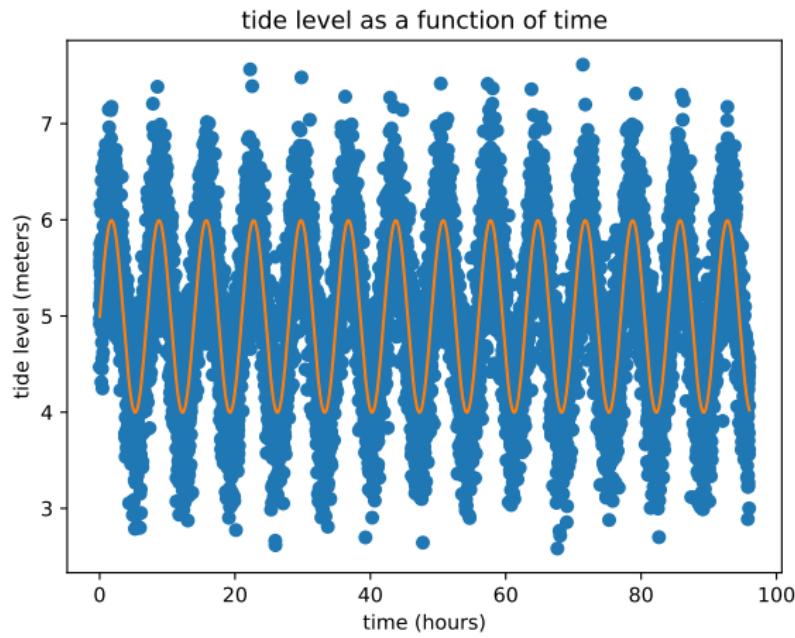
Use the function **fit sinus** in order to find the parameters of the function and to predict future values of the tide. What is the period of oscillation ?

...

└ Data processing

└ Algorithm types

# Tide level

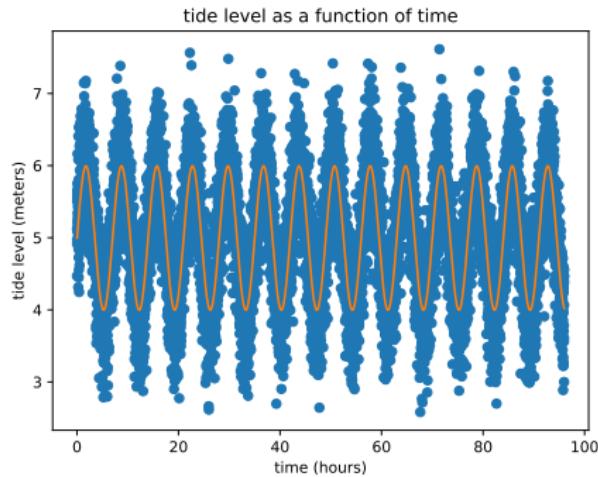


...

└ Data processing

└ Algorithm types

## Tide level



The inaccuracy comes from the **variance** in the data, which comes from **noise**.

...

└ Data processing

  └ Algorithm types

# Interpolation

**Remark :** in order to predict, we could also use **interpolation** between points.

...

└ Data processing

  └ Algorithm types

# Unsupervised learning

## Observations:

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$

...

- └ Data processing

- └ Algorithm types

# Unsupervised learning

**Observations:**

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$

**Objective:** understand the process underlying  $\tilde{x}$ .

# Unsupervised learning

## Observations:

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$

**Objective:** understand the process underlying  $\tilde{x}$ .

## Hypotheses:

- ▶ The observations  $\tilde{x}$  stem from a **stochastic process**  $x$

# Unsupervised learning

**Observations:**

- ▶ Empirical data points  $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$

**Objective:** understand the process underlying  $\tilde{x}$ .

**Hypotheses:**

- ▶ The observations  $\tilde{x}$  stem from a **stochastic process** x

**Then, we can for instance :**

- ▶ Find a smart visualization of the data
- ▶ Sort information in  $\tilde{x}$
- ▶ We want these to be robust so they can be extended to a new sample  $\tilde{x}'$

...

└ Data processing

  └ Algorithm types

## Unsupervised learning

**Remark :** when doing unsupervised learning, only **inputs** are given, they are *a priori* not associated with **outputs**

## Application : hierarchical clustering

We will study one example of unsupervised learning.  
It consists in building a **hierarchy of clusters**.

...

└ Data processing

  └ Algorithm types

## Application : hierarchical clustering

We will study one example of unsupervised learning.

It consists in building a **hierarchy of clusters**.

We will apply to a small example dataset containing addresses.

## Hierarchical clustering

### Exercice 3 : Plotting data

cd **hierarchical\_clustering/** and use **hierarchical\_clustering.py** in order to show the scatter plot of the data (nuage de points) loaded from **addresses.csv**.

It consists in showing the points at a position corresponding to their coordinates.

## Hierarchical clustering

### Exercice 3 : Plotting data

cd **hierarchical\_clustering/** and use **hierarchical\_clustering.py** in order to show the scatter plot of the data (nuage de points) loaded from **addresses.csv**.

It consists in showing the points at a position corresponding to their coordinates.

Several methods are possible.

...

└ Data processing

  └ Algorithm types

## Scatter plots

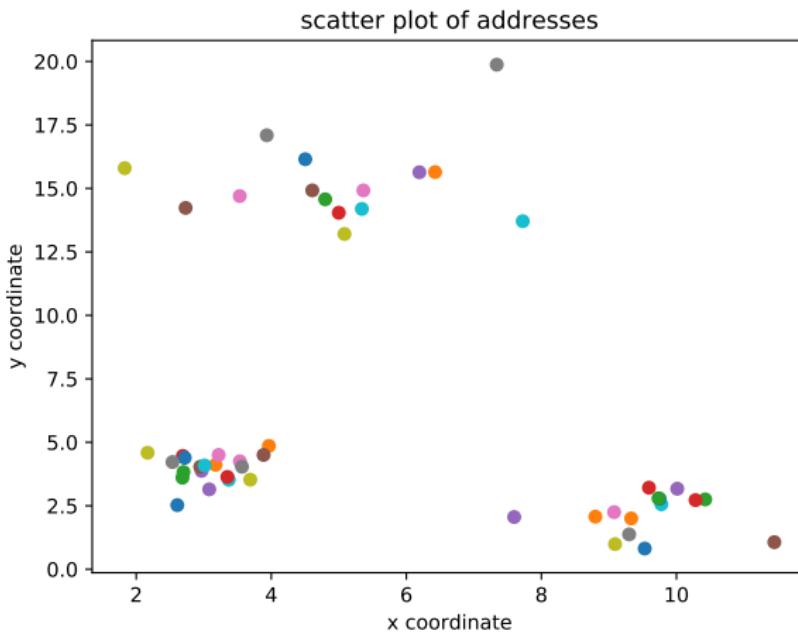
Seaborn lib: <https://seaborn.pydata.org/>

...

└ Data processing

└ Algorithm types

# Hierarchical clustering

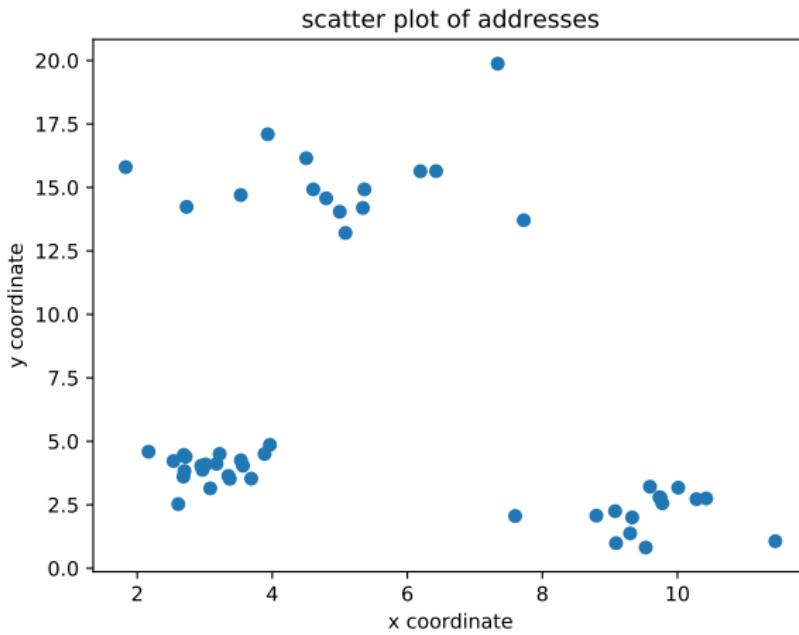


...

└ Data processing

└ Algorithm types

## Hierarchical clustering

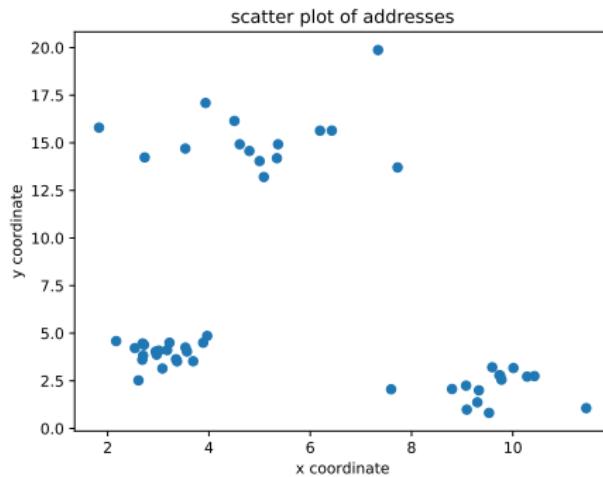


...

└ Data processing

└ Algorithm types

## Hierarchical clustering



Hierarchical clustering consists in progressively grouping points together in **classes**.

## Hierarchical clustering

**Exercice 4 : Hierarchical clustering** Edit the function **distance\_between\_classes.py** in order to compute the distance between two classes of points.

...

└ Data processing

  └ Algorithm types

## Hierarchical clustering

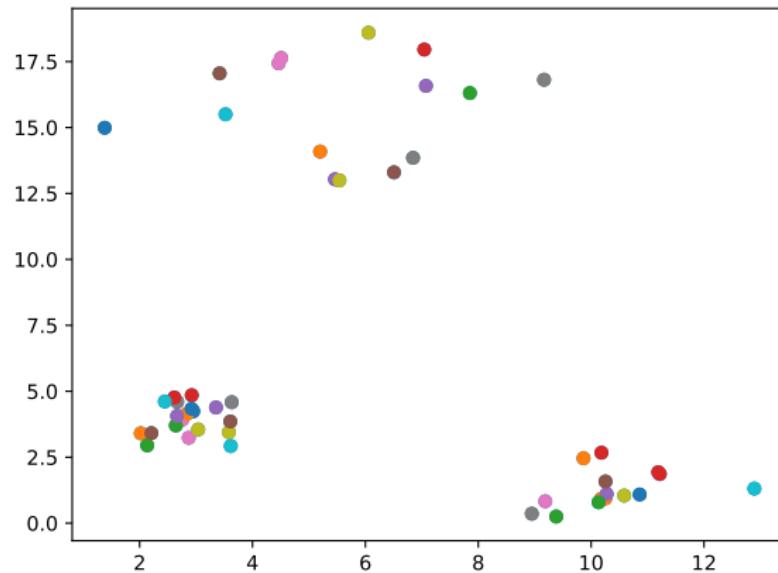
Exercice 4 : **Hierarchical clustering** Edit the function **find\_closest\_classes.py** in order to find the closest classes. Then they can be merged in the while loop.

...

└ Data processing

└ Algorithm types

## Hierarchical clustering

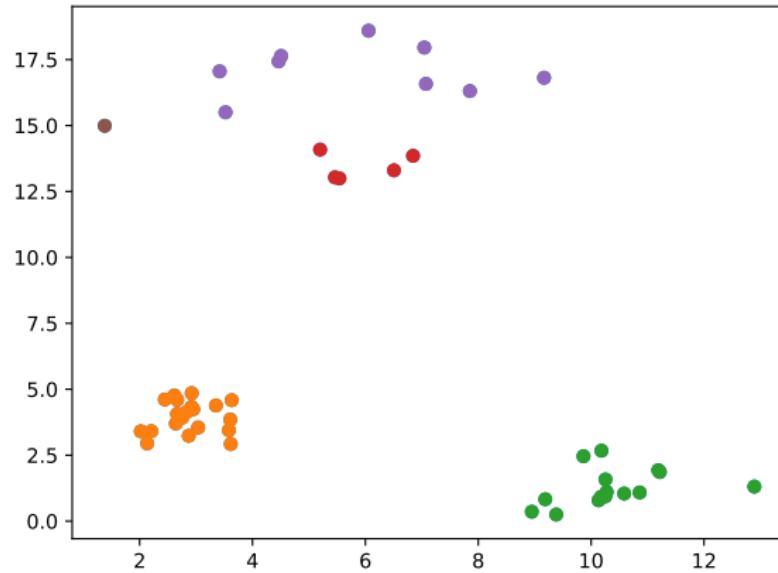


...

└ Data processing

└ Algorithm types

## Hierarchical clustering

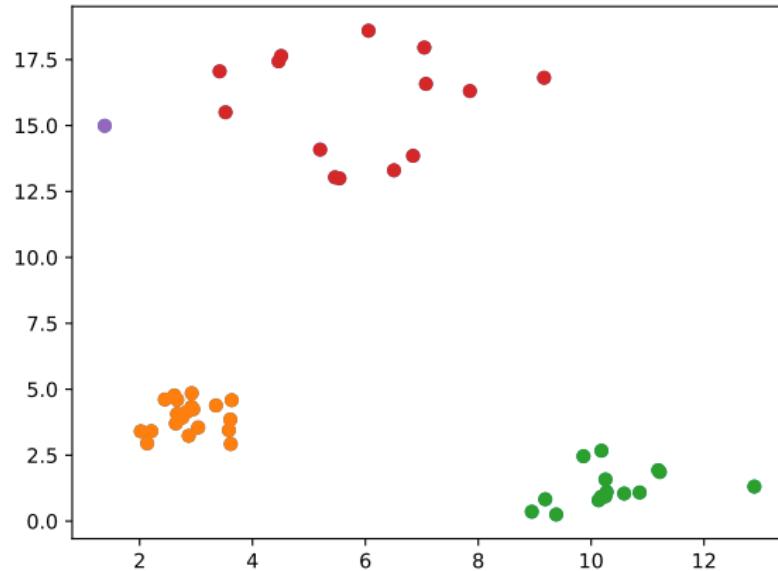


...

└ Data processing

└ Algorithm types

## Hierarchical clustering

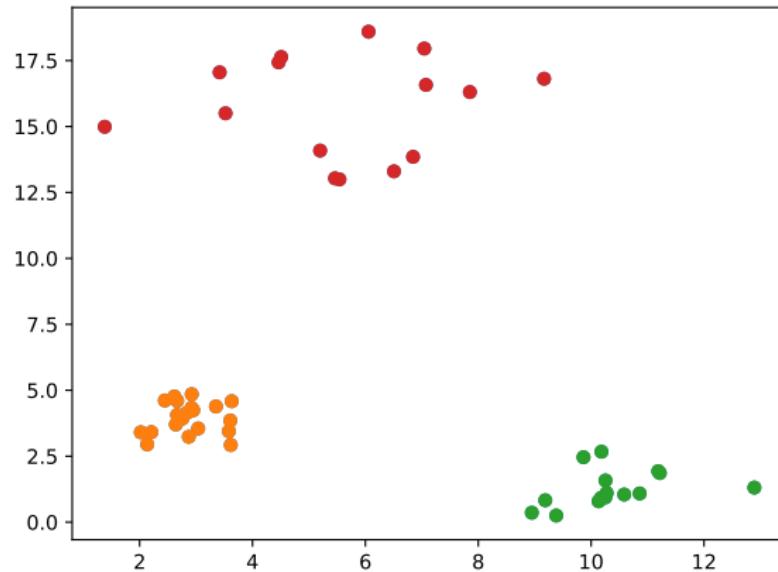


...

└ Data processing

└ Algorithm types

## Hierarchical clustering

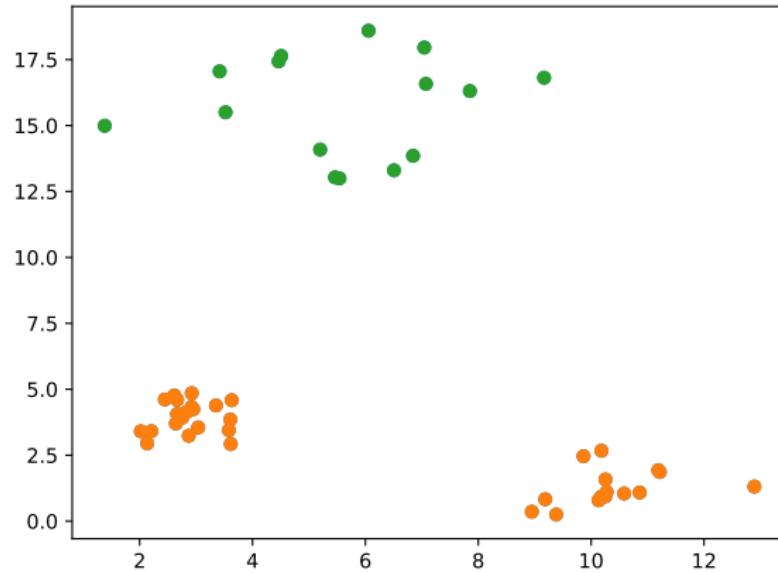


...

└ Data processing

└ Algorithm types

## Hierarchical clustering

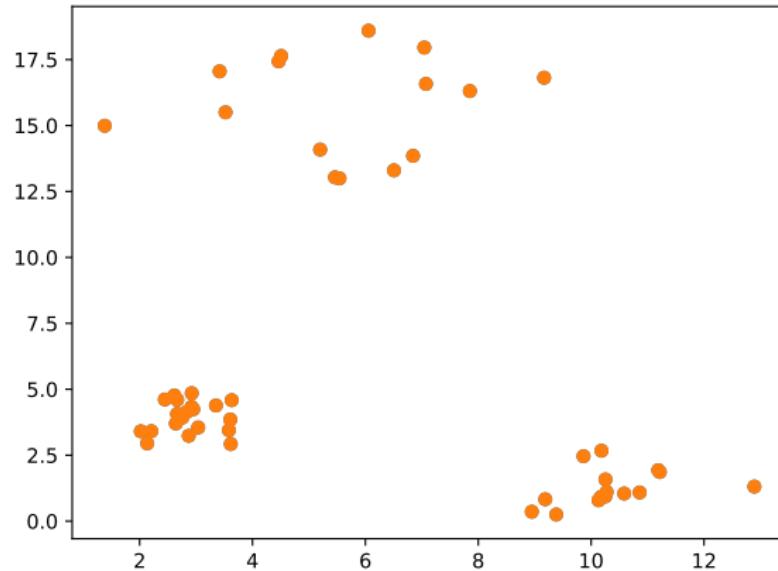


...

└ Data processing

└ Algorithm types

## Hierarchical clustering



...

└ Data processing

  └ Algorithm types

## Hierarchical clustering

An **essential** aspect of hierarchical clustering is that different criteria can be used in order to merge the classes.

## Hierarchical clustering

An **essential** aspect of hierarchical clustering is that different criteria can be used in order to merge the classes. The distance between class 1 and class 2 can for instance be:

- ▶ the minimum distance between one point of class 1 and one point of class 2: **single-linkage clustering**.
- ▶ the average distance between points un class 1 and points un class 2: **unweighted average linkage clustering**

## Hierarchical clustering

An **essential** aspect of hierarchical clustering is that different criteria can be used in order to merge the classes. The distance between class 1 and class 2 can for instance be:

- ▶ the minimum distance between one point of class 1 and one point of class 2: **single-linkage clustering**.
- ▶ the average distance between points un class 1 and points un class 2: **unweighted average linkage clustering**

The two methods can lead to a different hierarchy of clusters.

...

└ Data processing

  └ Algorithm types

## Semi-supervised learning

- ▶ The paradigms can be mixed :

...

└ Data processing

  └ Algorithm types

## Semi-supervised learning

- ▶ The paradigms can be mixed : in semi-supervised learning, only a small amount of data is labelled to help the unsupervised learning step.

...

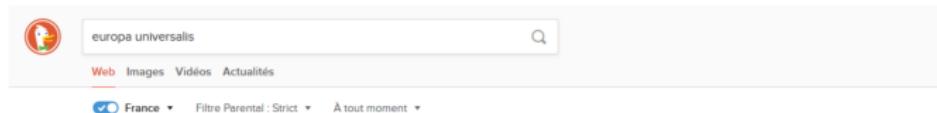
└ Data processing

  └ Algorithm types

## Application of the paradigms

Let's discuss the introductory examples in terms of learning paradigms.

# Example 1 : web browser and page ranking



**Europa Universalis IV**  
【送付無料】ヨーロッパ 硬式用【クローラルエリート】MG=〔金属製／84cm／900g〕上 シルバー(2th21140)  
火の日 sale C1806 葵 安 ...  
S [europauniversalis4.com](http://europauniversalis4.com)

**Europa Universalis 4 — Wikipédia**  
Europe Universalis 4 (stylisé Europa Universalis IV) est un jeu de grande stratégie historique développé par la société suédoise Paradox Development Studio et ...  
W [https://fr.wikipedia.org/wiki/Europa\\_Universalis\\_4](https://fr.wikipedia.org/wiki/Europa_Universalis_4)

Figure: Duckduckgo web browser

## Example 1 : web browser and page ranking

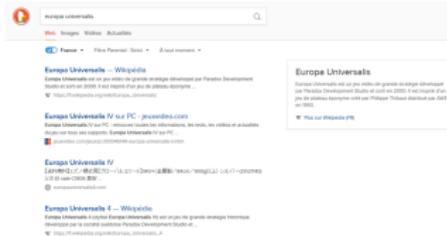


Figure: Duckduckgo web browser

- ▶ User  $i$  types a query  $\tilde{x}_i$ , and finally chooses a web page  $\tilde{y}_i$ .
- ▶ The web browser wants to propose the most relevant output
- ▶ **Supervised learning**

## Example 2 : targetted ads

The screenshot displays the Ghostery website with the following elements:

- Header:** Includes a search bar ("Enter a site URL") and a "Test your site now!" button.
- Navigation:** Links to PRODUCTS, BLOG, SUPPORT, ABOUT GHOSTERY, English, Sign In, and Install Ghostery.
- Callout:** "Faster, safer, and smarter browsing".
- Text:** "Ghostery helps you browse smarter by giving you control over ads and tracking technologies to speed up page loads, eliminate clutter, and protect your data."
- Buttons:** "Install Ghostery" and "Learn More".
- Browsers Supported:** Icons for Chrome, Firefox, Opera, Safari, Microsoft Edge, and Internet Explorer.
- Extension Interface:** Shows a dashboard with 22 trackers found on the current page, 0 tracked, and 0 blocked. It includes sections for Advertising, Cookies, and Scripts, along with a sidebar for Site Center and Help.
- Footer Features:** Smart Blocking, Dynamic UI, Enhanced Anti-Tracking, and a cookie consent message.
- Cookie Consent Message:** "This site uses cookies" with a "X" icon, stating: "You are not being tracked since your browser is reporting that you do not want to. This is a setting of your browser as you won't be able to opt-in until you disable the 'Do Not Track' feature."

Figure: Ghostery

## Example 2 : targetted ads

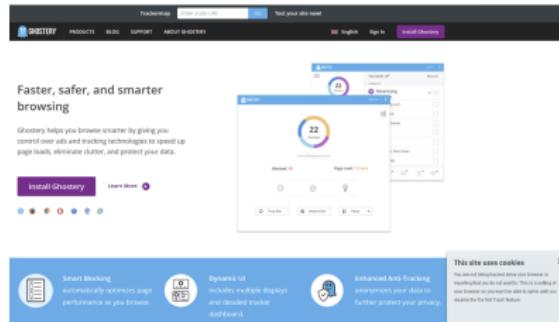


Figure: Ghostery

- ▶ Propose products (output) to a user (input)
- ▶ **Supervised**

...

└ Data processing

└ Algorithm types

## Example 4 : Image segmentation

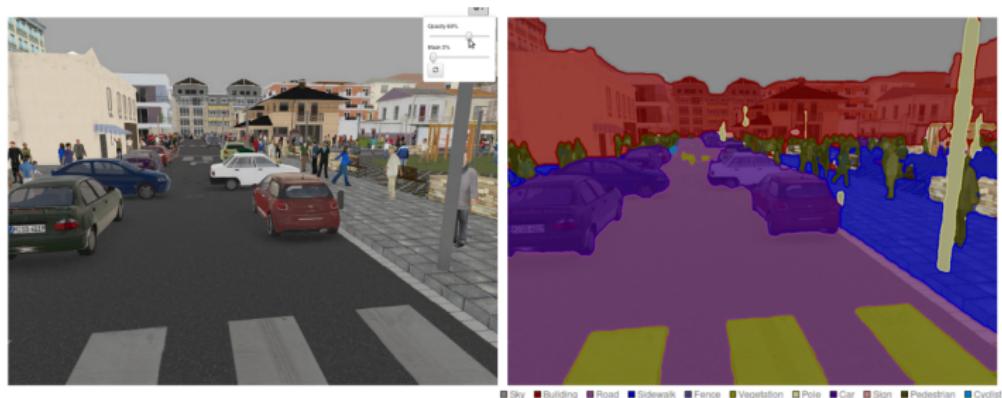


Figure: Source : Nvidia dev blog

...

└ Data processing

└ Algorithm types

## Example 4 : Image segmentation



Figure: Source : Nvidia dev blog

- ▶ Often people give labeled examples of segmented images
- ▶ **Supervised**

...

└ Data processing

└ Algorithm types

## Example 4 : Image segmentation

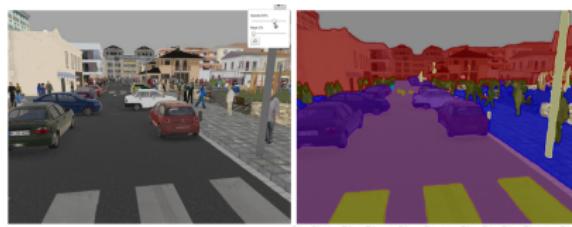


Figure: Source : Nvidia dev blog

- ▶ One could also group pixels together based on their properties
- ▶ **Unsupervised** (but most of the time the supervised method is used)

...

└ Data processing

└ Algorithm types

## Example 5 : Digit classification



Figure: MNIST dataset [LeCun and Cortes, 2010]

...

└ Data processing

└ Algorithm types

## Example 5 : Digit classification



Figure: MNIST dataset [LeCun and Cortes, 2010]

- ▶ **input** : pixels
- ▶ **output** : digit (class)
- ▶ **Supervised**

## Example 6 : Lexicometry

Capitalisme : 0.5  
Socialisme : 0.3  
Sociales : 0.3  
Marxismes : 0.1

Islam : 0.4  
Musulman : 0.4  
Mahomet : 0.2  
Coran : 0.2

Ouvrage : 0.6  
Biographie : 0.3  
Extrait : 0.1  
Préface : 0.1

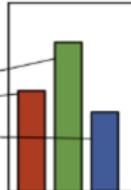
### Le monde musulman, Marx et le socialisme

*Forts en 1900, [Islam] [Capitalisme] de [Ouvrage] [Musulman] [Mahomet] sont les deux derniers termes d'un tableau qui, entre autres, d'après [Biographie] la prophète [Mahomet] (1), n'avait encore jamais été établi (2). Cet arrangement pose des problèmes d'interférence assez basiques : quels rapports existent-ils entre [Islam] et [Capitalisme] ?*

« ISLAM ET CAPITALISME », DE MAXIME ROBINSON

PAR ALAIN GREEN

*appartient dans le monde [Capitalisme]. Pour [Musulman] ces dernières années ont entraîné un rapprochement avec [Islam] et l'Occident. Le rôle de l'islamologue à [Ouvrage] a donc été délaissé au profit de l'islamologue à [Islam].* Robinson préfère adopter pour son analyse le terme [Islam] comme filtre une pensée qui n'est pas dénuée de sens. Non seulement [Islam] ne connaît pas la religion, mais aussi [Islam]



**FIGURE :** document

**FIGURE :**  
topics

Figure: Statistics on texts

## Example 6 : Lexicometry



FIGURE :  
topics



FIGURE : document

Figure: Statistics on texts

- We want to understand the underlying structure of a text, no output.
- **Unsupervised**

## Example 3 : Text Translation

The screenshot shows a web-based machine translation interface. At the top left is the SYSTRAN logo with the text "Pure neural MT™". To its right is a "Traduction" button with a small info icon. Below the logo is a dark blue header bar. On the left side of the main content area, there's a "Traduction de texte" section with a sub-instruction: "This demo platform allows you to experience Pure Neural™ machine translation based on the last Research community's findings and SYSTRAN's R&D." It also says "You can translate up to 2000 characters of text in the languages proposed below. Check out the [information page](#) to learn more." On the right side, there's a blue sidebar with the text "Click h ENTER" and the SYSTRAN logo.

Below the header, there's a toolbar with language selection buttons: "Français - DéTECTé" (selected), "Anglais" (disabled), and a "Selectionner un profil" button. The main content area has two columns. The left column contains the French input text: "12 juin 2018 - Le 20 juin prochain, le Parlement européen arrêtera sa décision sur la directive Copyright, symbole d'une nouvelle période de régulation de l'Internet. La Quadrature du Net vous invite à [appeler](#) les eurodéputés pour exiger qu'ils agissent contre l'automatisation de la censure au nom de la protection du droit d'auteur et, plus largement, contre la centralisation du Web." The right column contains the English output text: "June 12, 2018 - On June 20, the European Parliament will decide on the Copyright Directive, symbolizing a new period of regulation of the Internet. The Quadrature du Net invites you to [call](#) on MEPs to demand that they act against the automation of censorship in the name of copyright protection and, more broadly, against centralization of the Web."

Figure: Text translation

## Example 3 : Text Translation

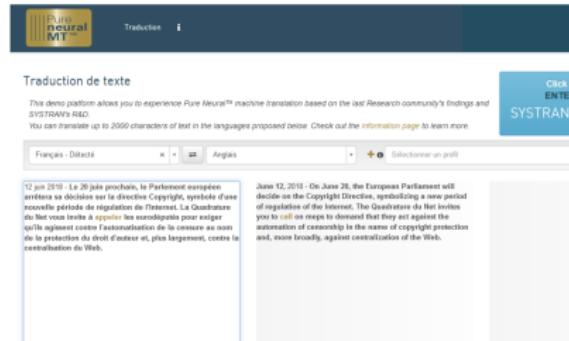


Figure: Text translation

- ▶ Based on translated examples
- ▶ Supervised

...

└ Data processing

└ Algorithm types

## Example 7 : Time Series Analysis

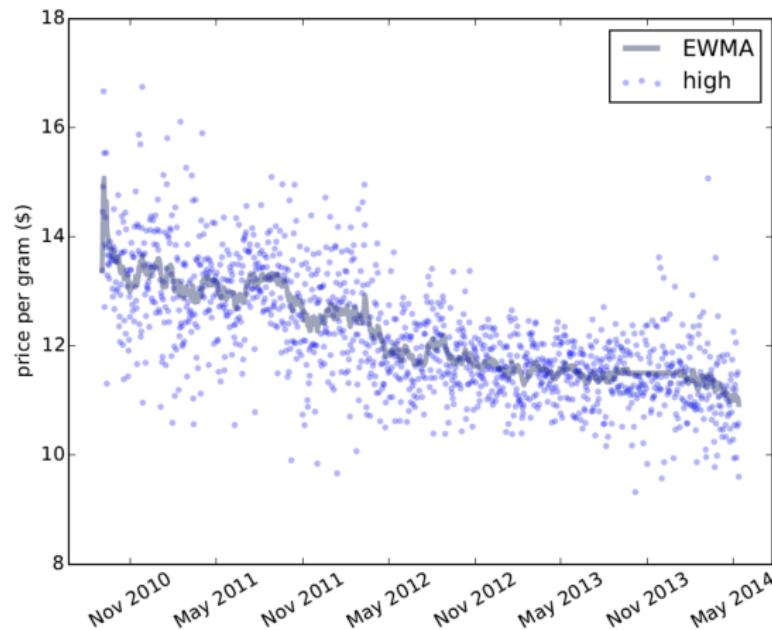


Figure: Price of some plant

...

└ Data processing

└ Algorithm types

## Example 7 : Time Series Analysis

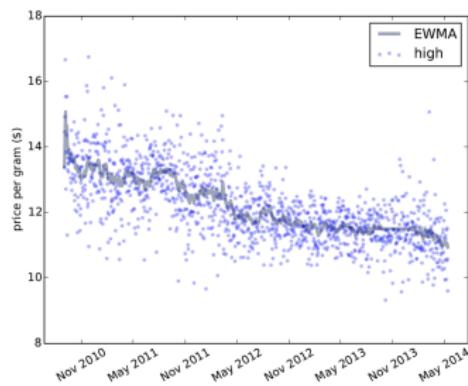


Figure: Price of some plant

### ► Supervised

...

└ Data processing

└ Algorithm types

## Example 8 : Clustering

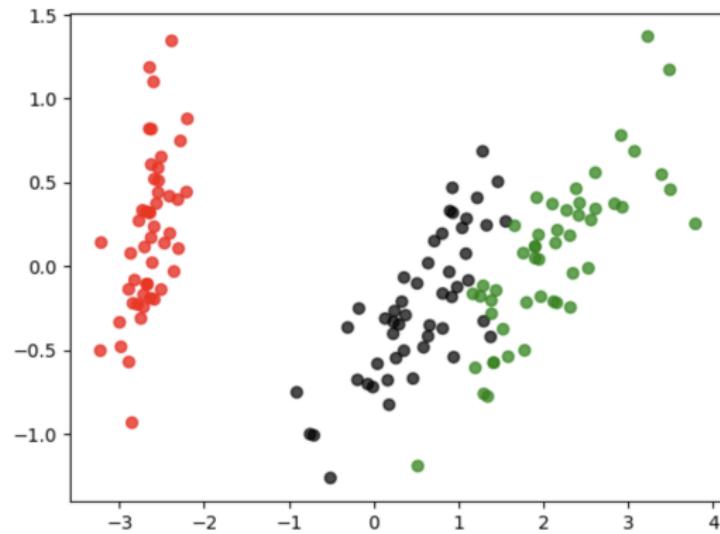


Figure: Iris dataset

...

└ Data processing

└ Algorithm types

## Example 8 : Clustering

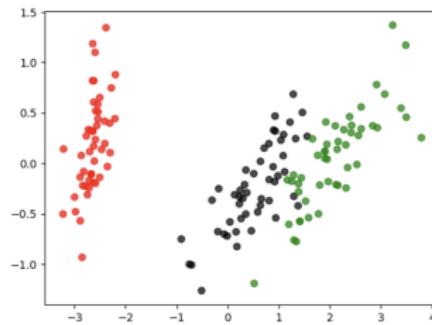


Figure: Iris dataset

- ▶ Grouping together points as a function of their features
- ▶ **Unsupervised learning**

...

└ Data processing

└ Algorithm types

## Example 9 : game AI



Figure: Europa Universalis

...

- └ Data processing

- └ Algorithm types

## Example 9 : game AI



Figure: Europa Universalis

- ▶ Designing an **agent** performing **actions**
- ▶ **Reinforcement learning**

...

└ Data processing

  └ Algorithm types

## Conclusion

Sometimes there is an ambiguity between RL, supervised and unsupervised.

## Stochastic processes

- ▶ For both supervised learning and unsupervised learning, we assumed that the data stemmed from a **stochastic process**.

...

└ Data processing

└ Stochastic processes and distributions

## Stochastic processes

- ▶ For both supervised learning and unsupervised learning, we assumed that the data stemmed from a **stochastic process**.
- ▶ This means that they are **random variables** with a certain **law**

## Random variables

- ▶ A **random variable** is a variable that can take several values.  
These can be **discrete** or **continuous**
- ▶ Examples :

...

 Data processing Stochastic processes and distributions

## Random variables

- ▶ A **random variable** is a variable that can take several values.  
These can be **discrete** or **continuous**
- ▶ Examples : dice game, weather forecast, outcome of a game,  
number of cars on a highway, waiting time at a bus stop.

...

└ Data processing

  └ Stochastic processes and distributions

# Probability distributions

- ▶ A random variable is linked to a **probability distribution**.

## Probability distributions

- ▶ A random variable is linked to a **probability distribution**.
- ▶ It quantifies the probability of observing one outcome.

...

- └ Data processing

- └ Stochastic processes and distributions

## Probability distributions

- ▶ A random variable is linked to a **probability distribution**, which is a function  $P$
- ▶ It quantifies the probability of observing one outcome.
- ▶ For a discrete variable : each possible outcome is associated with a number between 0 and 1

...

└ Data processing

└ Stochastic processes and distributions

# Probability distributions

- ▶ For a dice game, the possible outcomes are in the set  $\{1, 2, 3, 4, 5, 6\}$
- ▶ For a dice game :  $P(1) = ?$   $P(2) = ?$   $P(3) = ?$   $P(4) = ?$   
 $P(5) = ?$   $P(6) = ?$

## Probability distributions

- ▶ For a dice game, the possible outcomes are in the set  $\{1, 2, 3, 4, 5, 6\}$
- ▶ For a dice game :  $P(1) = \frac{1}{6}$ ,  $P(2) = \frac{1}{6}$ ,  $P(3) = \frac{1}{6}$ ,  $P(4) = \frac{1}{6}$ ,  
 $P(5) = \frac{1}{6}$ ,  $P(6) = \frac{1}{6}$
- ▶ This is called a **uniform distribution**

...

└ Data processing

  └ Stochastic processes and distributions

## Continuous variables

- ▶ How would you model a continuous variable ? Can you assign a number to a waiting time or a weather ?

...

└ Data processing

└ Stochastic processes and distributions

## Continuous variables

- ▶ How would you model a continuous variable ? Can you assign a number to a waiting time or a weather ?
- ▶ One needs to use **probability densities**. Formally, the probability of being between  $x$  and  $x + dx$  is  $p(x)dx$ .

...

└ Data processing

└ Stochastic processes and distributions

## Continuous variables

- ▶ How would you model a continuous variable ? Can you assign a number to a waiting time or a weather ?
- ▶ One needs to use **probability densities**. Formally, the probability of being between  $x$  and  $x + dx$  is  $p(x)dx$ .
- ▶ Let's see some examples

## Uniform discrete

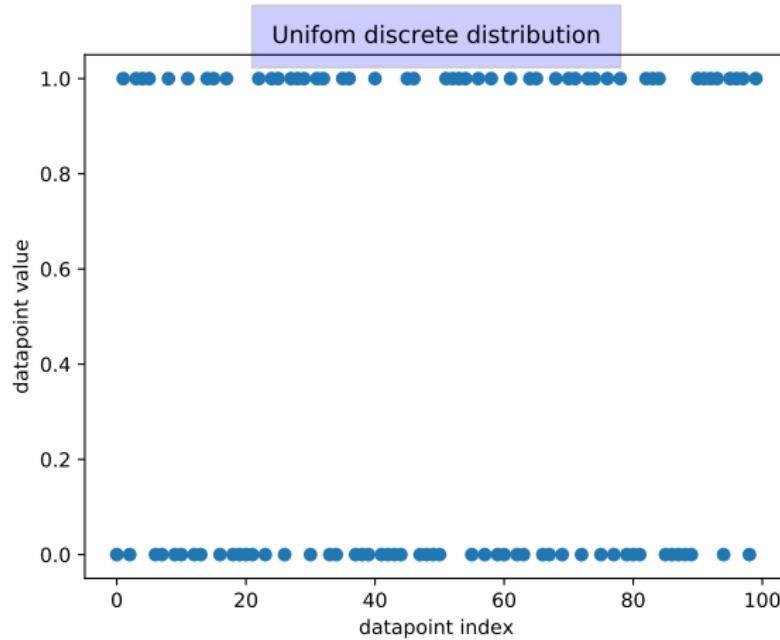


Figure: Uniform discrete distribution

...

└ Data processing

└ Stochastic processes and distributions

# Bernoulli

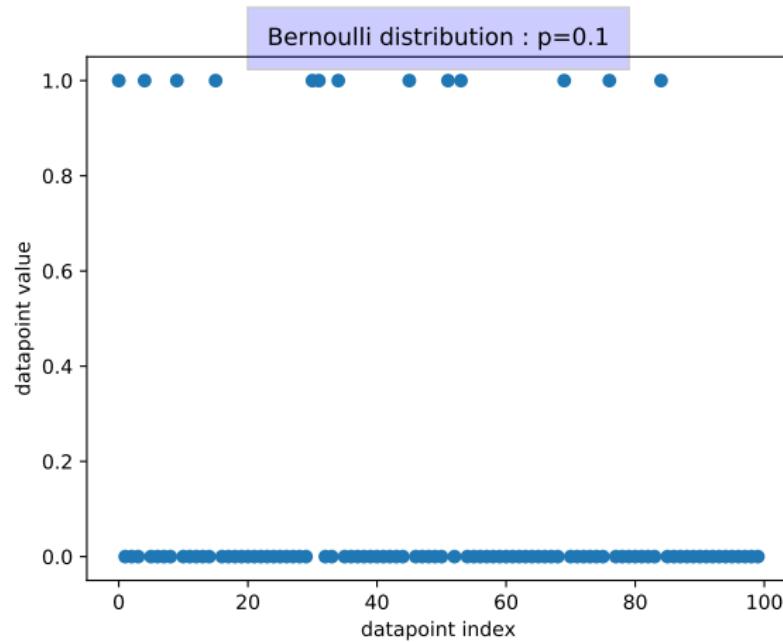
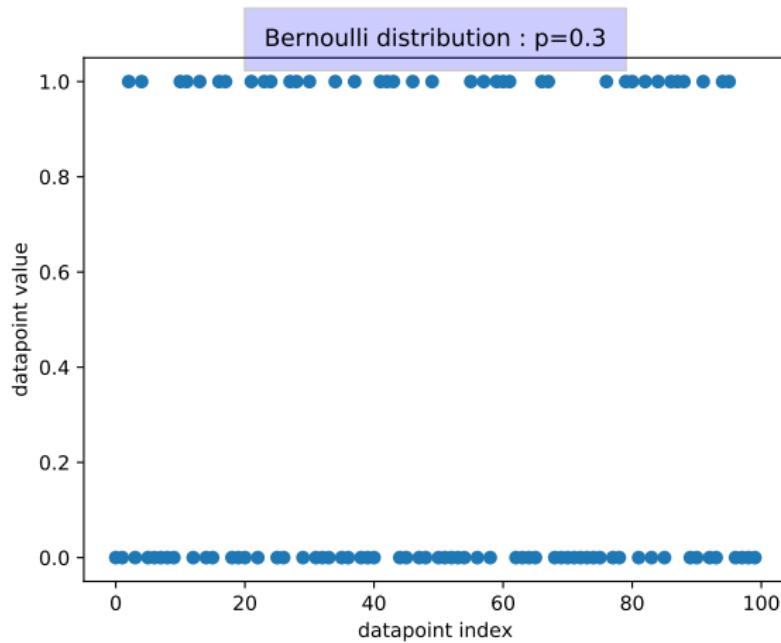


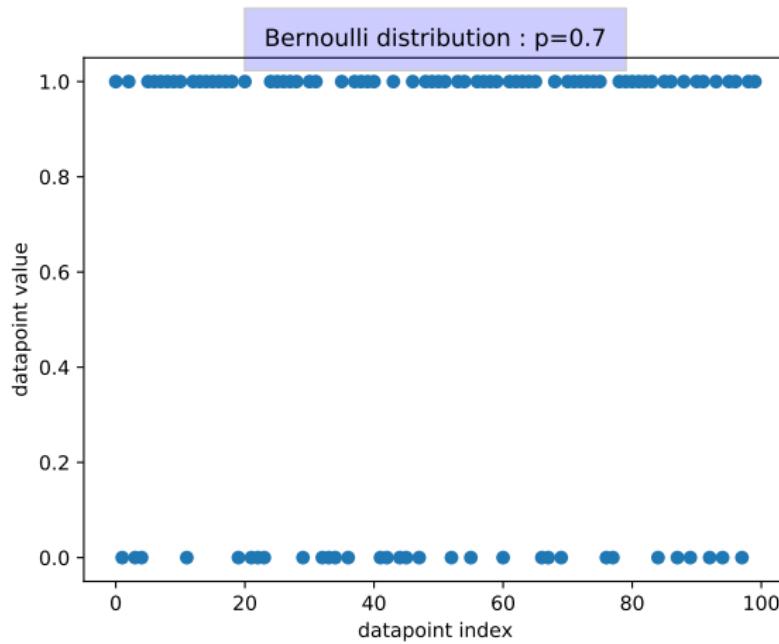
Figure: Bernoulli distribution

# Bernoulli



## Figure: Bernoulli Distribution

# Bernoulli



## Figure: Bernoulli Distribution

## Uniform continuous

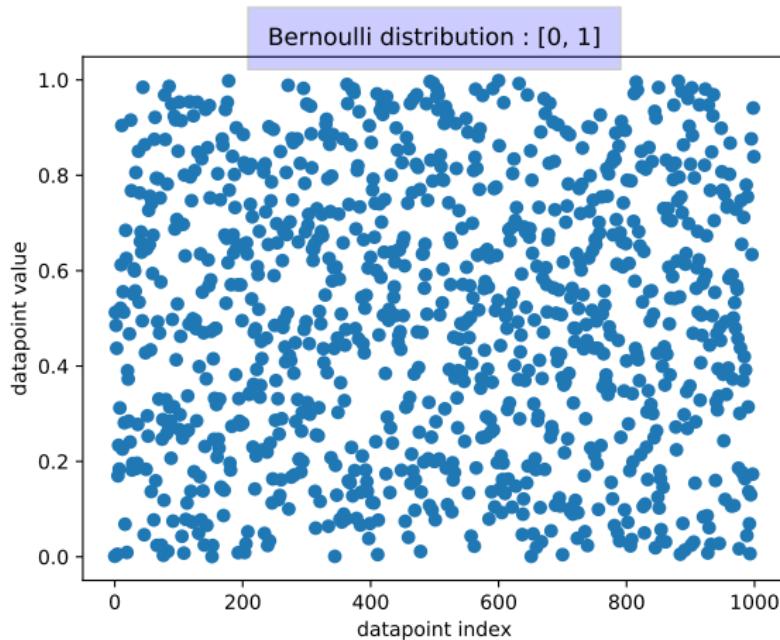


Figure: Uniform continuous distribution

## Uniform continuous

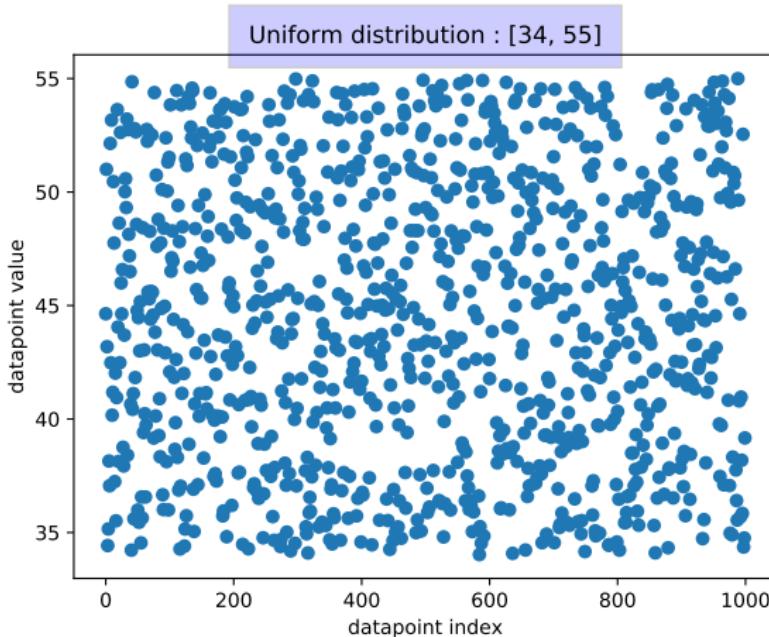
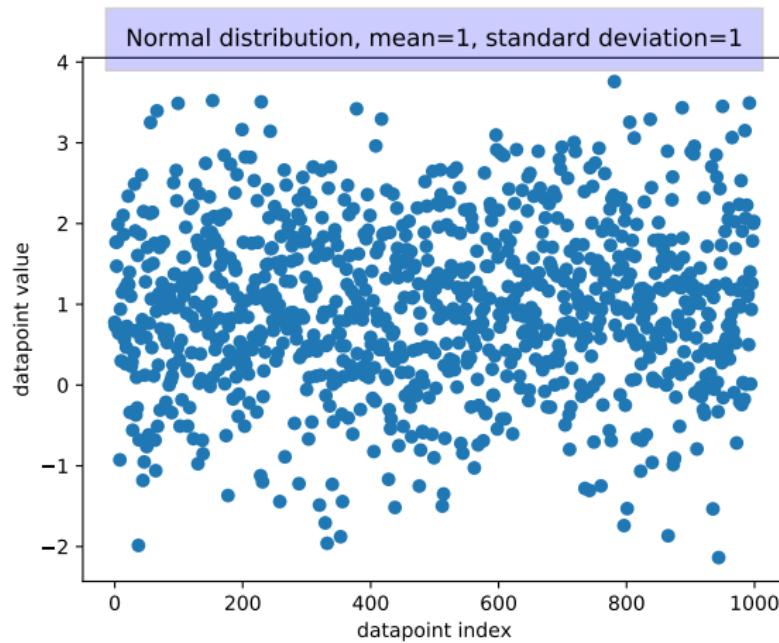


Figure: Uniform continuous distribution

## Normal



## Figure: Normal distribution

## Normal

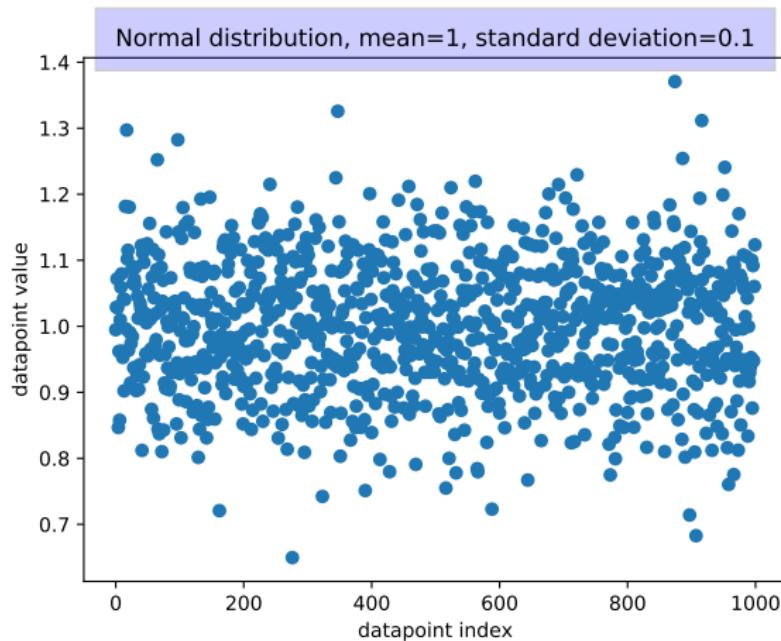


Figure: Normal distribution

## Normal

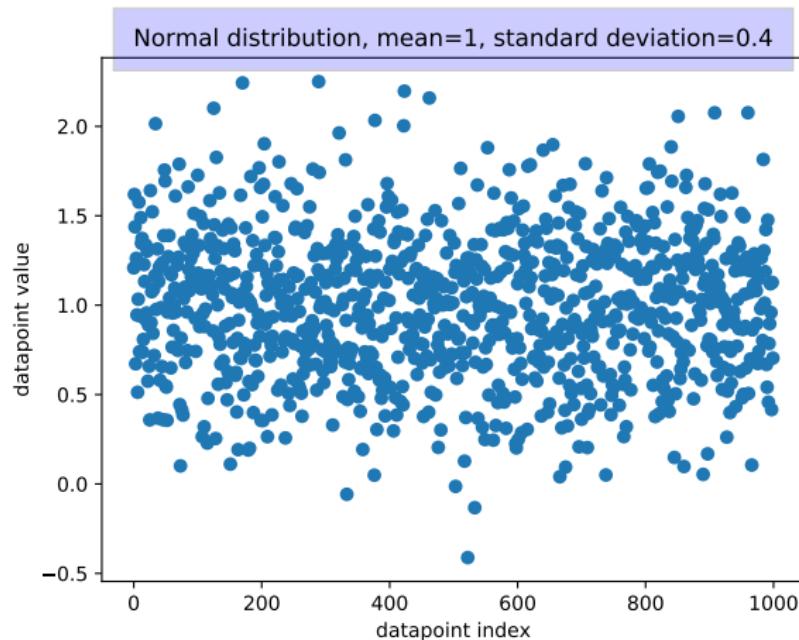


Figure: Normal distribution

## White noise

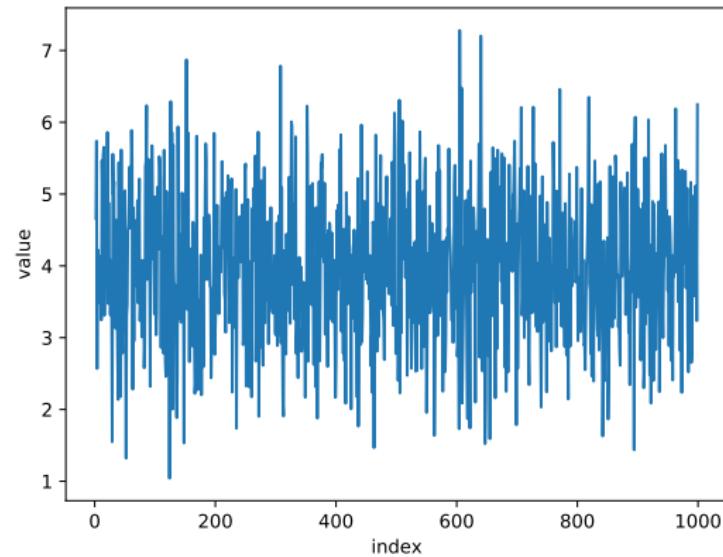


Figure: White noise

...

└ Data processing

  └ Stochastic processes and distributions

# Histograms

Is looking at the raw dataset really **informative** ?

...

└ Data processing

  └ Stochastic processes and distributions

# Histograms

Is looking at the raw dataset really **informative** ?  
It is informative, but often a **histogram** tells more.

## Uniform discrete

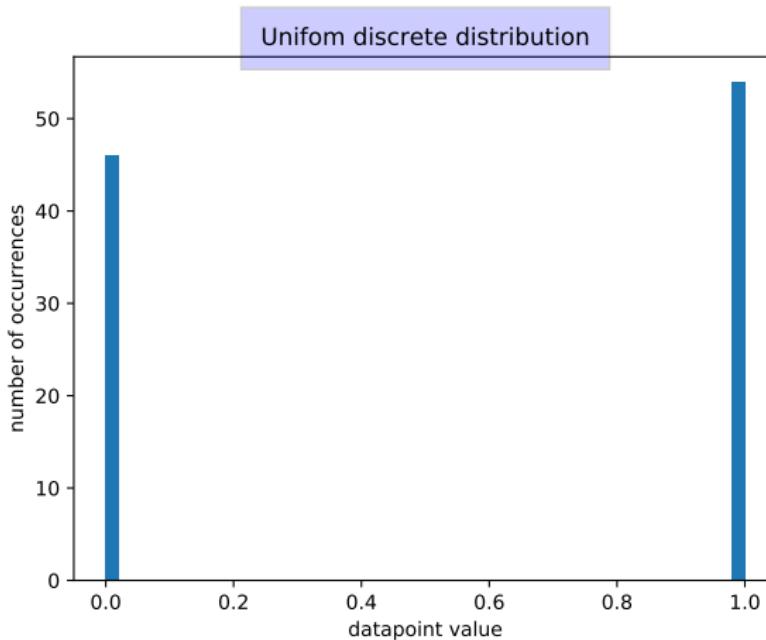


Figure: Histogram 1

## Bernoulli

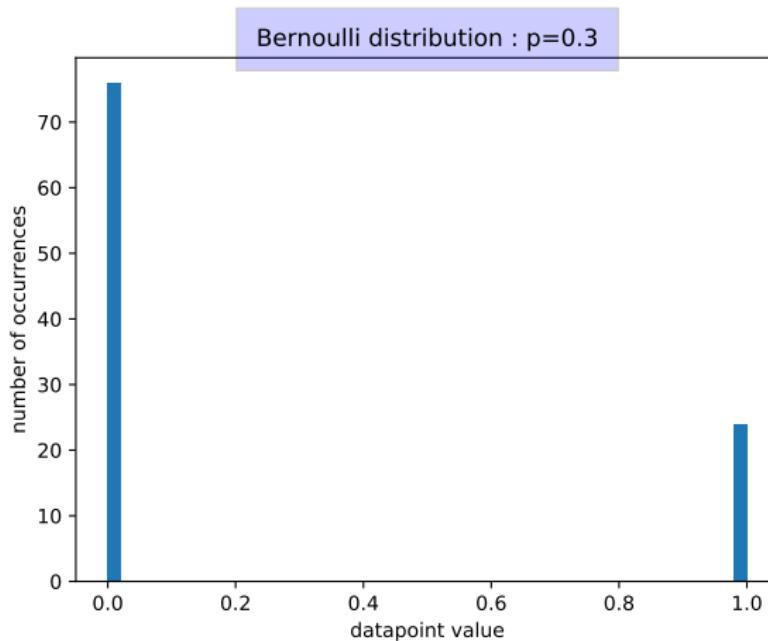


Figure: Histogram 2

## Uniform continuous

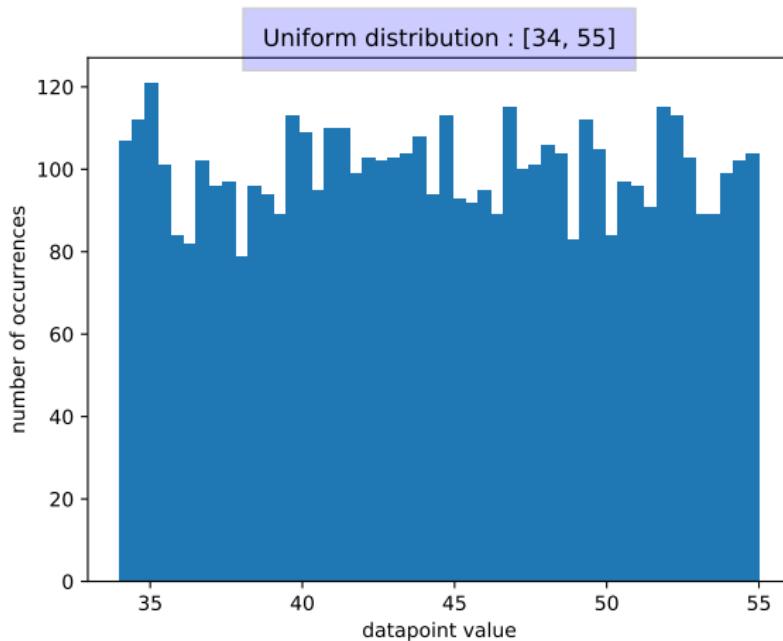


Figure: Histogram 3

## Normal

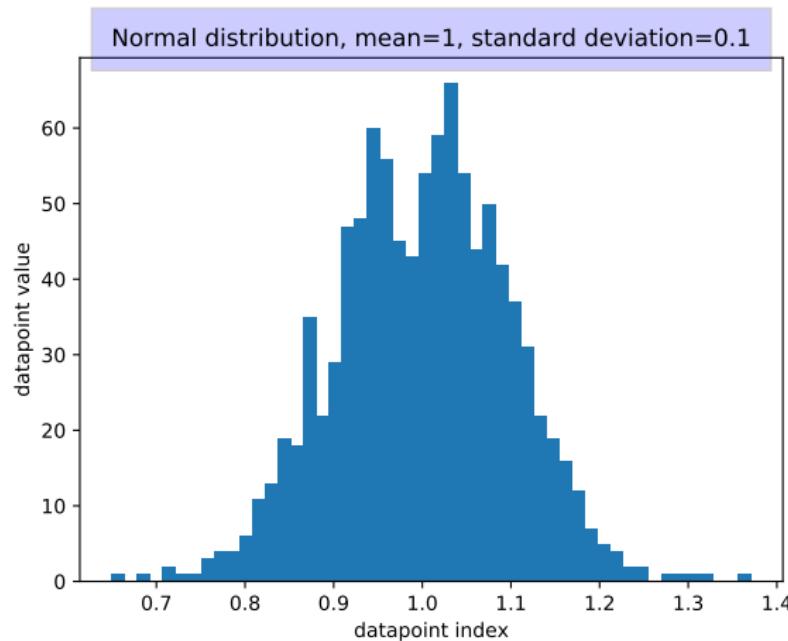


Figure: Histogram 4

...

## The choice of a model and overfitting

- ▶ In the case of supervised learning, we mentioned a **function** that links the input  $\tilde{x}$  to the output  $\tilde{y} = \tilde{f}(\tilde{x})$ .
- ▶ In the case of unsupervised learning, if we study the underlying process of the data, we can look for a **probability distribution**  $p$  to fit them.

...

└ Data processing

  └ Fitting data

## Choice of the model

- ▶ Given a dataset, several models or functions are possible
- ▶ It is necessary to constrain the model to a certain form in order to be able to optimize it

...

## Choice of the model

- ▶ Given a dataset, several models or functions are possible
- ▶ It is necessary to constrain the model to a certain form in order to be able to optimize it
- ▶ For instance
  - ▶ a mixture of normal laws when fitting a probability distribution to data
  - ▶ a neural network when classifying digit
- ▶ The model must be chosen depending on the problem

...

└ Data processing

  └ Fitting data

## Evaluation

How do we evaluate the quality of a model ?

...

└ Data processing

  └ Fitting data

## Evaluation

How do we evaluate the quality of a model ?

- ▶ Accuracy of the prediction,
- ▶ Evaluation on a test sample different from the learning sample
- ▶ Computation time, speed of convergence, robustness of the results

...

└ Data processing

  └ Fitting data

# Overfitting

- ▶ What could be the drawbacks of using a very simple model (very few parameters) ?

...

└ Data processing

  └ Fitting data

# Overfitting

- ▶ What could be the drawbacks of using a very simple model (very few parameters)
  - ▶ Weak expressive power

...

└ Data processing

  └ Fitting data

## Overfitting

- ▶ What could be the drawbacks of using a very simple model (very few parameters)
  - ▶ Weak expressive power
- ▶ What could be the drawbacks of having a very complex model (that contains a very large number of parameters, e.g. millions as in a very deep neural network) ie a very high expressive power ?

...

## Overfitting

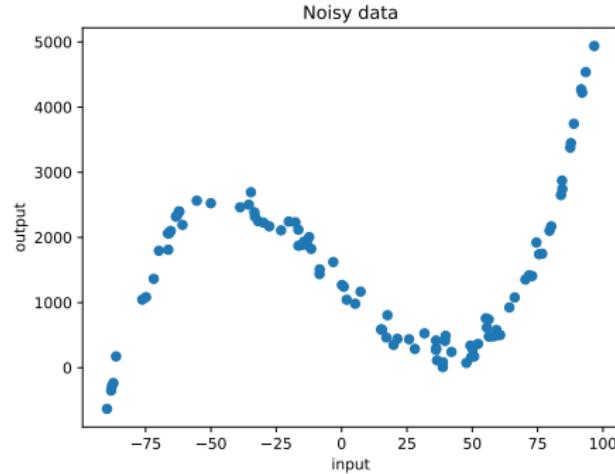
- ▶ What could be the drawbacks of using a very simple model (very few parameters)
  - ▶ Weak expressive power
- ▶ What could be the drawbacks of having a very complex model (that contains a very large number of parameters, e.g. millions as in a very deep neural network) ie a very high expressive power ?
  - ▶ Harder to optimize
  - ▶ Harder to interpret
  - ▶ Can **overfit**

...

- └ Data processing
- └ Fitting data

# Overfitting

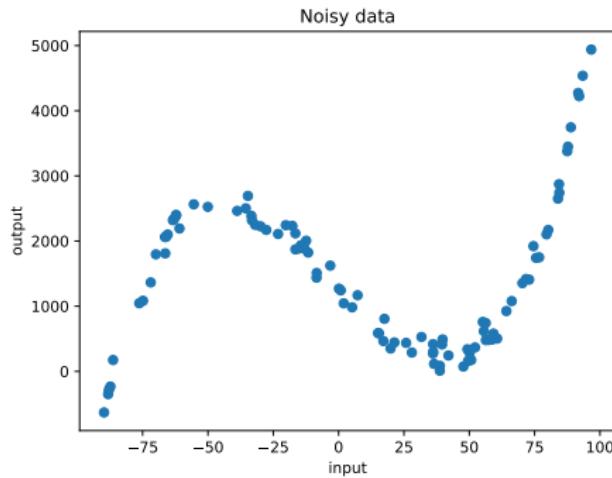
We will learn a **model** of the following data, in a **supervised learning** context.



...

Data processing

Fitting data

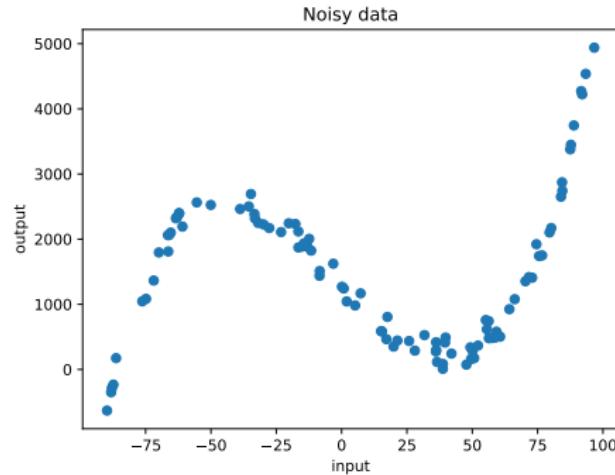


Our **model** should allow us to predict the **output** for new **inputs**.  
For instance what should be predicted for an input of  $-48$  ?

...

└ Data processing

└ Fitting data



We need to choose:

- ▶ A **class** of model.
- ▶ A relevant **complexity** once the class is chosen.

...

└ Data processing

  └ Fitting data

# Overfitting

- ▶ What could be the drawbacks of using a very simple model (very few parameters) ?

...

└ Data processing

  └ Fitting data

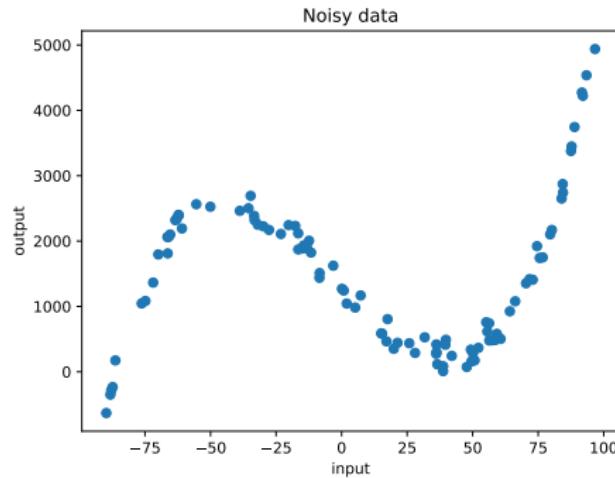
# Overfitting

- ▶ What could be the drawbacks of using a very simple model (very few parameters)
  - ▶ Weak expressive power

...

# Fitting

## Exercice 5: Fitting polynomials to data

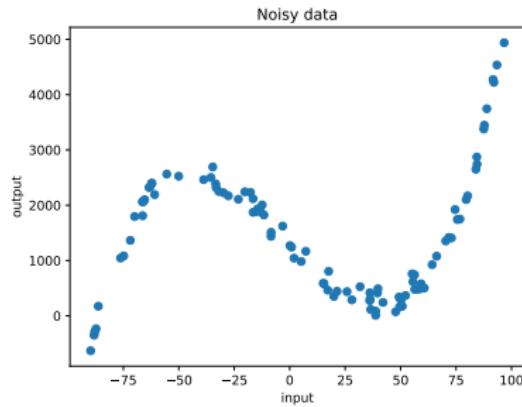


We want to perform supervised learning in order to be able to predict the output  $y$  for a new sample  $x$ .

# Fitting

## Exercice 5 : Fitting polynomials to data

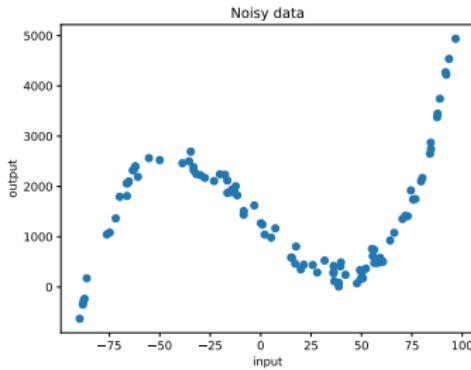
- ▶ We want to perform supervised learning in order to be able to predict the output  $y$  for a new sample  $x$ .



- ▶ To illustrate the problem of overfitting, we will use **polynomials** as models.

# Fitting

## Exercice 5 : Fitting polynomials to data



We will divide the dataset into two subsets :

- ▶ a **training set** : used to learn the most relevant polynom once the degree is chosen
- ▶ a **test set** : used to evaluate overfitting

# Fitting

## Exercice 5: Fitting polynomials to data

- ▶ `cd ./overfitting`. Use the dataset contained in `linear_noisy_data.csv`, load it from `fit_data.py` in order to assess the impact of the `degree` of the polynom on overfitting.
- ▶ You need to edit the loop at the end of the file.

# Fitting

## Exercice 5 : Fitting polynomials to data

- ▶ The higher the degree of the polynom, the more parameters it has and the better it can fit the training points :

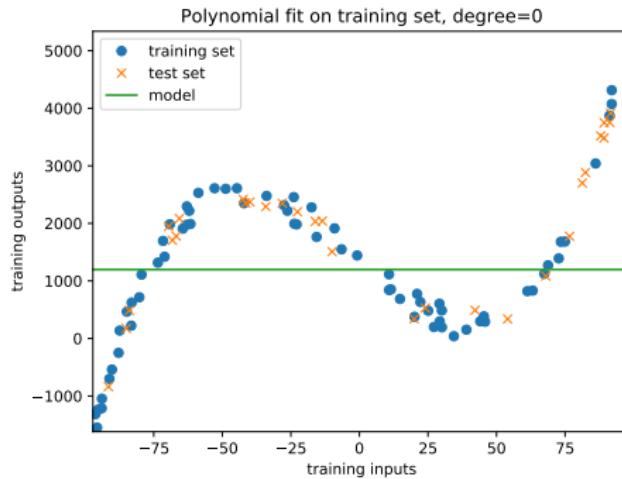


Figure: degree 0

# Fitting

## Exercice 5 : Fitting polynomials to data

- ▶ The higher the degree of the polynom, the more parameters it has and the better it can fit the training points :

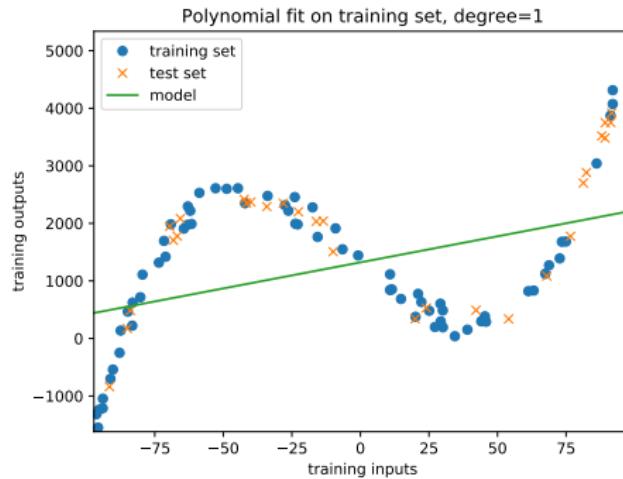


Figure: degree 1

# Fitting

## Exercice 5 : Fitting polynomials to data

- ▶ The higher the degree of the polynom, the more parameters it has and the better it can fit the training points :

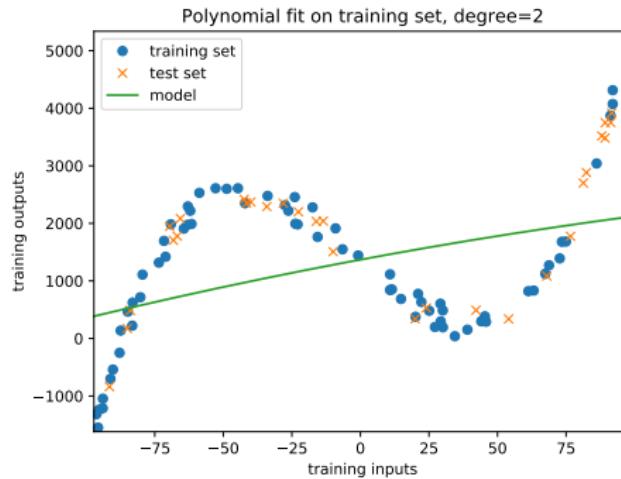


Figure: degree 2

# Fitting

## Exercice 5 : Fitting polynomials to data

- ▶ The higher the degree of the polynom, the more parameters it has and the better it can fit the training points :

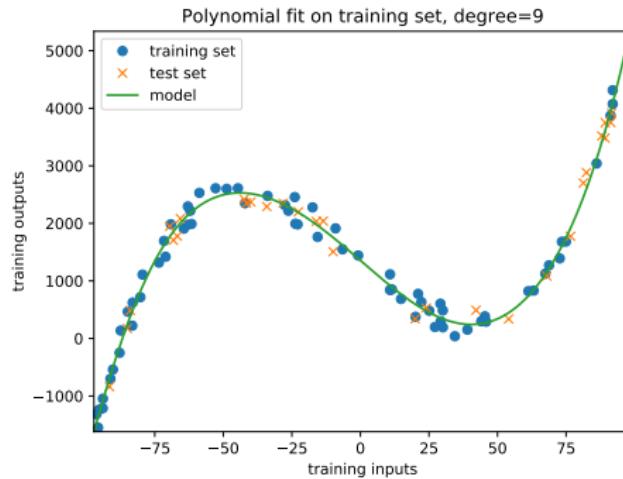


Figure: degree 9

# Fitting

## Exercice 5 : Fitting polynomials to data

- ▶ The higher the degree of the polynom, the more parameters it has and the better it can fit the training points :

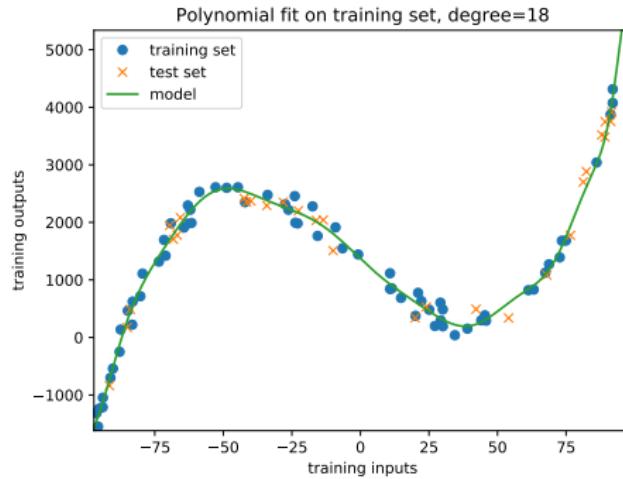


Figure: degree 19

# Fitting

## Exercice 5 : Fitting polynomials to data

- ▶ However, the error on the test set increases and the model loses **signification**

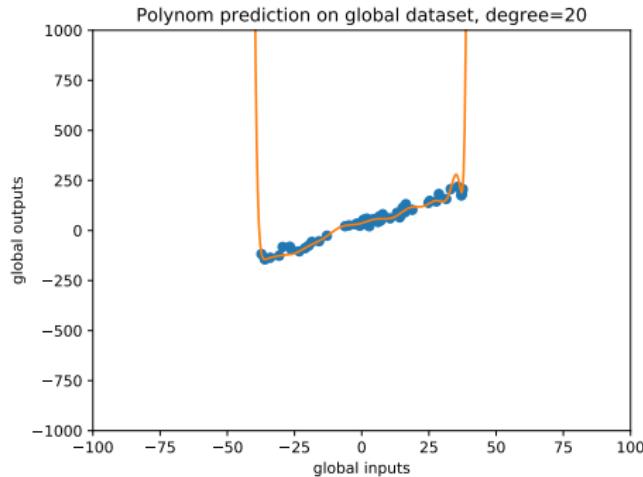


Figure: Useless solution

...

Data processing

Fitting data

# Fitting

## Exercice 5 : Fitting polynomials to data

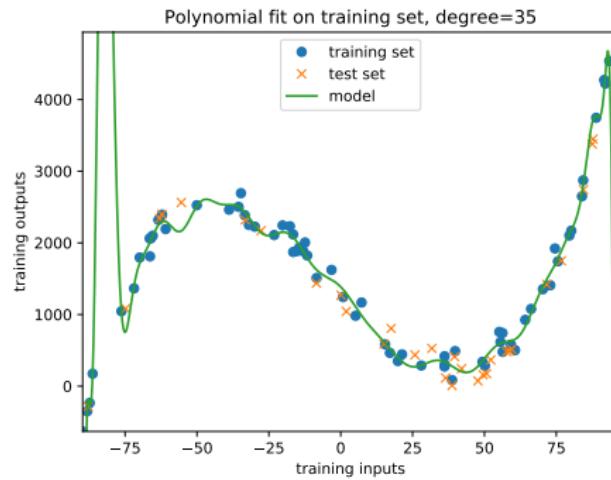


Figure: When the degree is too high.

# Fitting

## Exercice 5 : Fitting polynomials to data

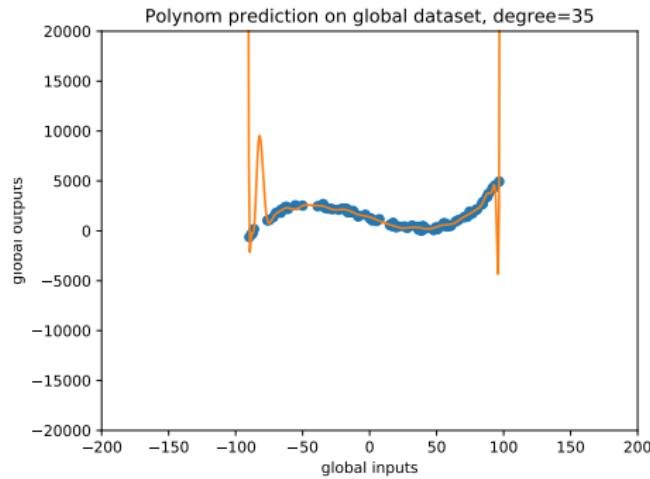
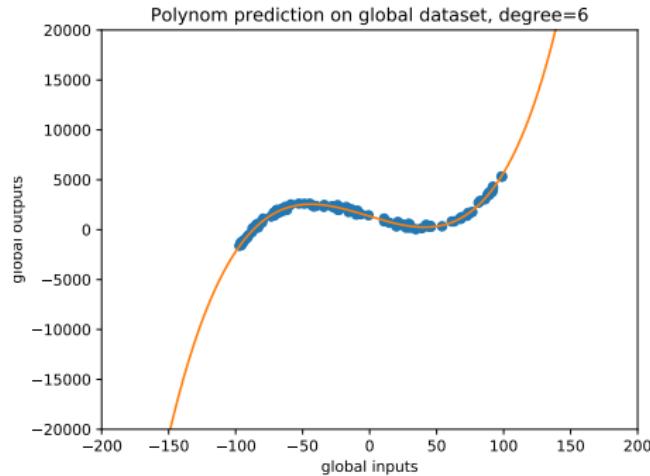


Figure: When the degree is too high.

## Fitting

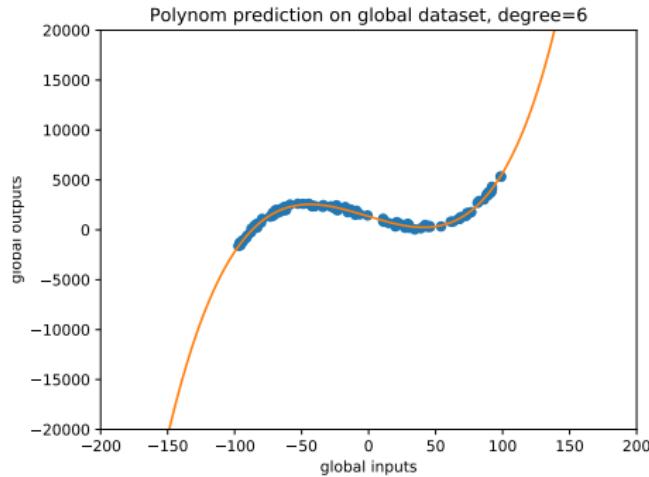
**Exercice 5: Fitting polynomials to data** When does the test error start to increase ?



**Figure:** When the degree is too high.

## Fitting

**Exercice 6 : Fitting polynomials to data** In that situation, what degree should we use ? (use a **quantitative criteriton** )



**Figure:** When the degree is too high.

...

└ Data processing

  └ Fitting data

## Trying to prevent overfitting

- ▶ The problem of overfitting is linked to that of **generalisation** : to what extent are we allowed to extrapolate the knowledge obtained on the training samples to new samples ?
- ▶ To improve generalisation, one can use :
  - ▶ a **validation set**
  - ▶ **regularization**

...

└ Data processing

  └ Fitting data

## Regularization methods

- ▶ Penalize the magnitude of the weight in a neural network
- ▶ Remove neurons in a neural network (pruning)
- ▶ use smooth functions (continuous)

...

- └ Data processing
  - └ Fitting data

## References

-  IBM.  
Analytics : The real-world use of big data in financial services.
-  LeCun, Y. and Cortes, C. (2010).  
{MNIST} handwritten digit database.
-  Olshannikova, E., Ometov, A., Koucheryavy, Y., and Olsson, T. (2016).  
*Chapter 4 VISUALIZING BIG DATA.*  
Number October.