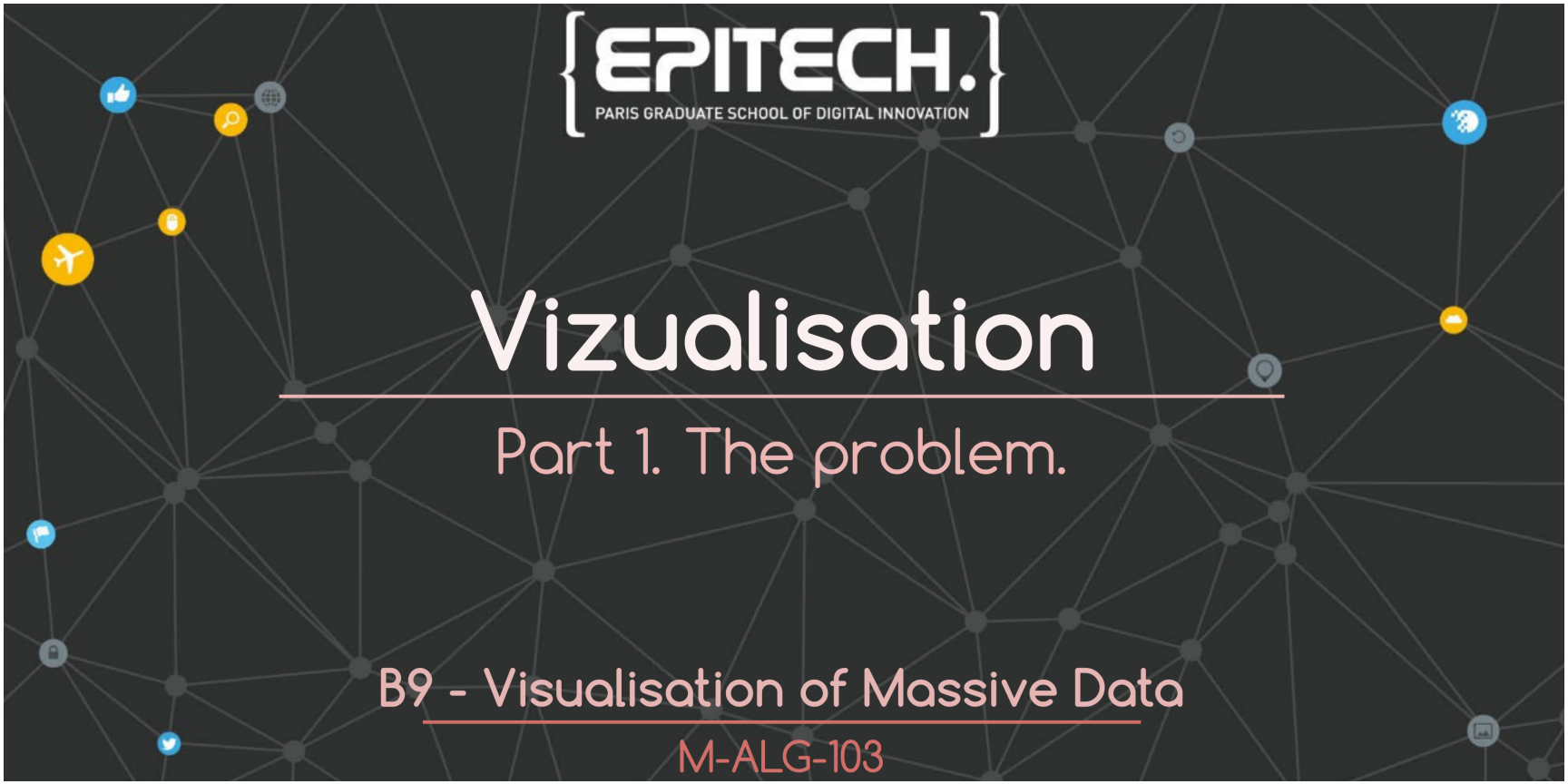

EPITECH.
PARIS GRADUATE SCHOOL OF DIGITAL INNOVATION
Vizualisation
Part 1. The problem.
B9 - Visualisation of Massive Data
M-ALG-103

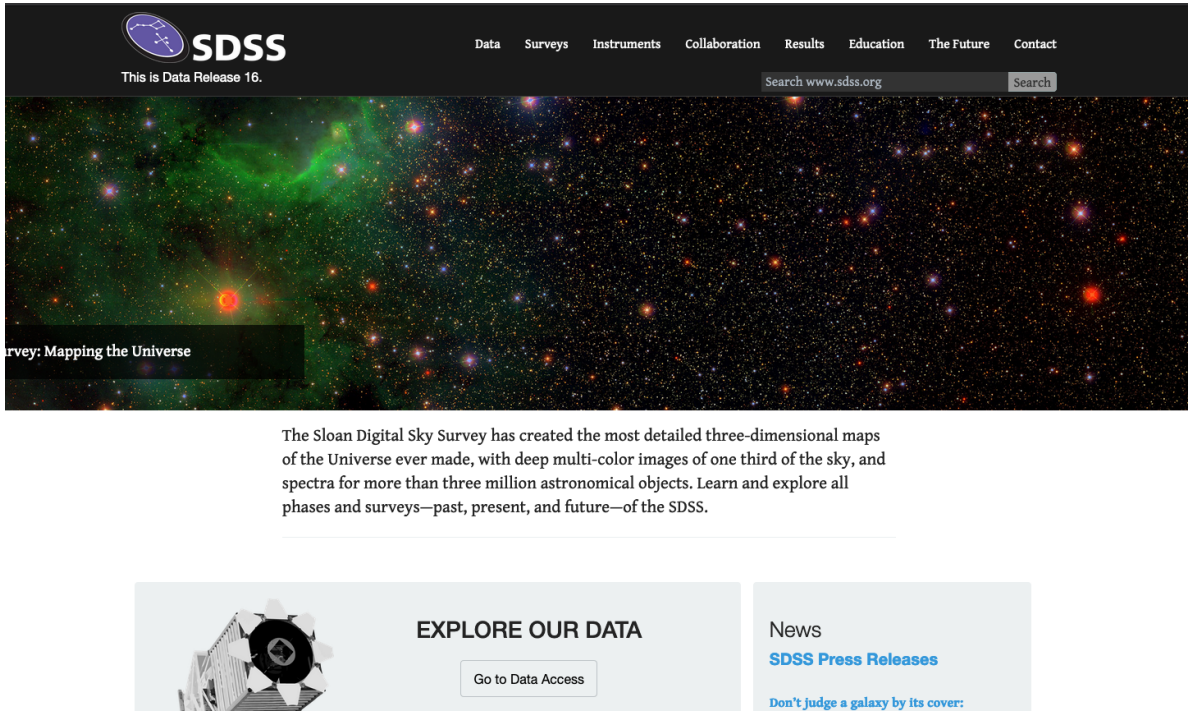


Figure: SDSS `https://www.sdss.org/`

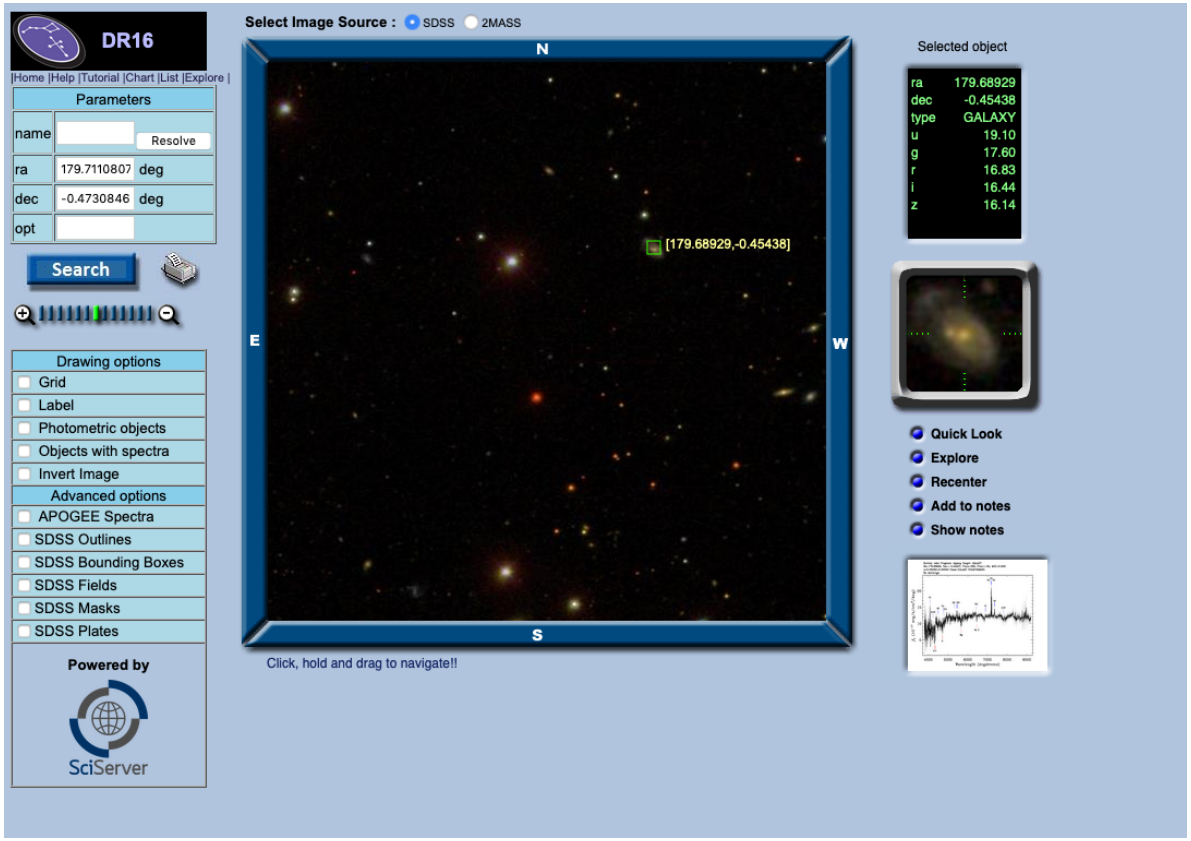


Figure: SDSS `https://www.sdss.org/`

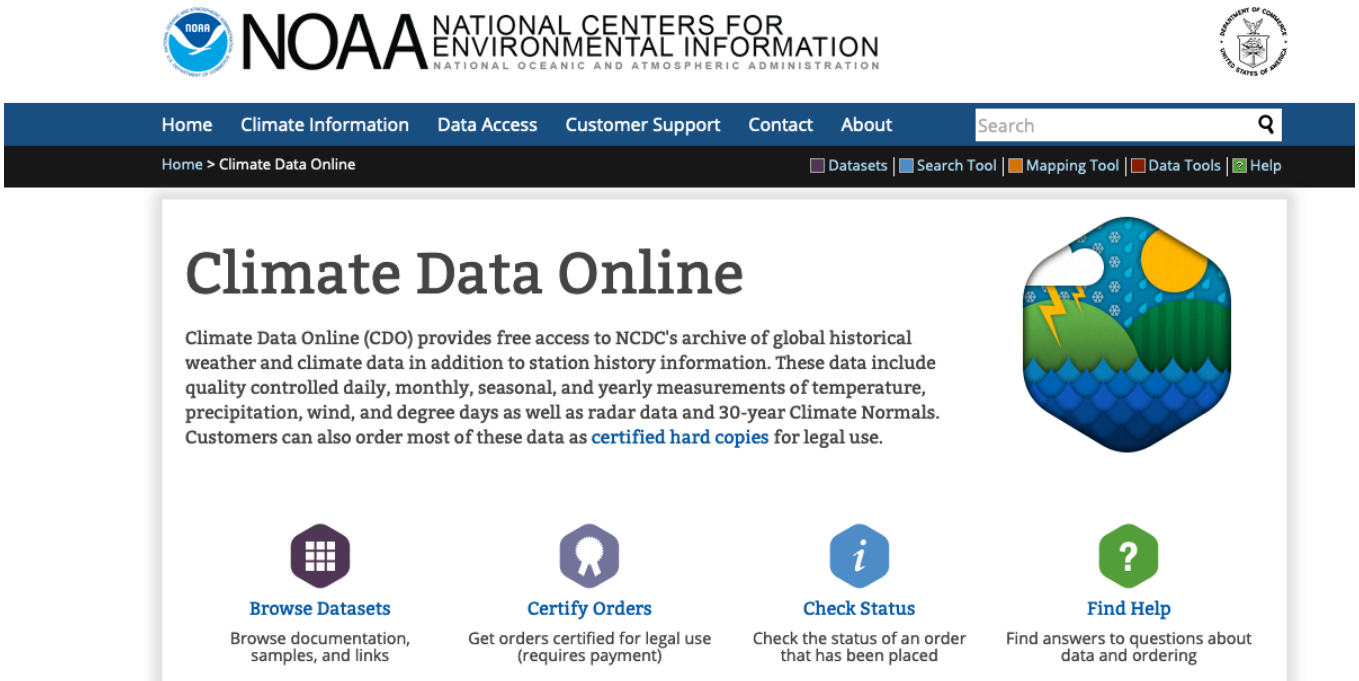


Figure: NOAA dataset `https://www.ncdc.noaa.gov/cdo-web/`

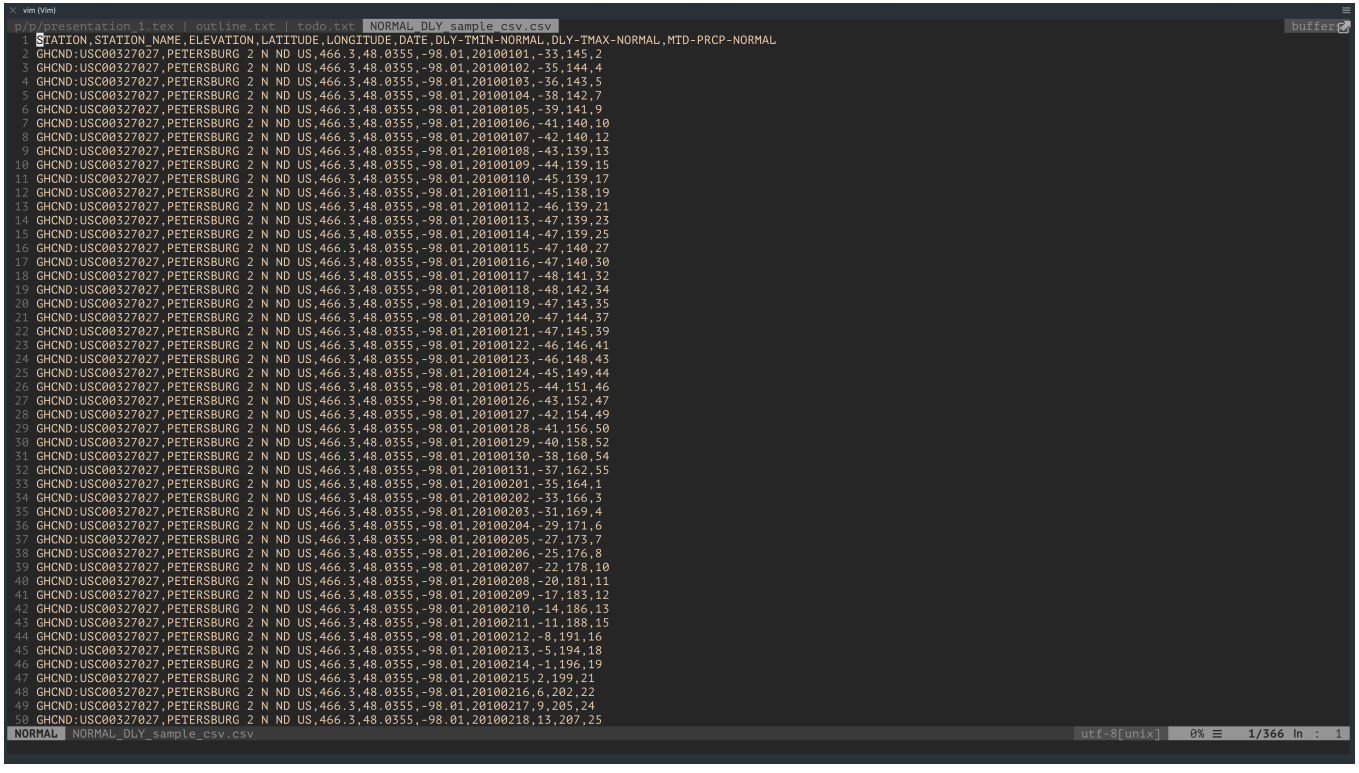


Figure: NOAA data sample. `https://www.ncdc.noaa.gov/cdo-web/`

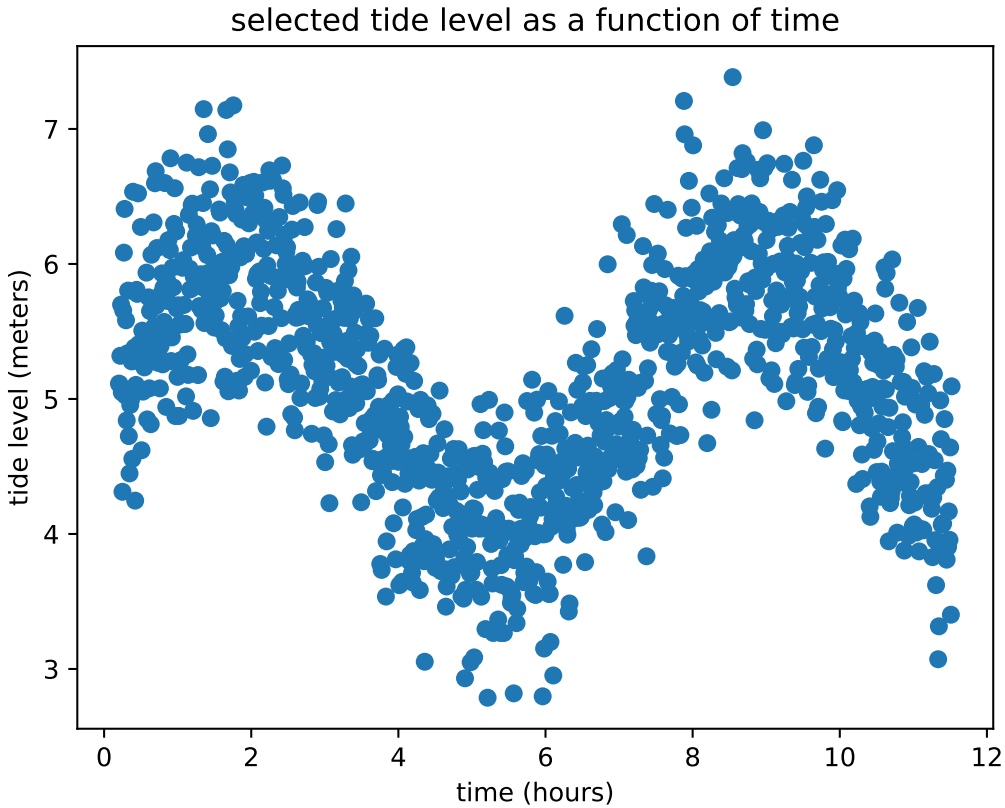


Figure: tide level

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

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)

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

Let us start by defining what big data is.

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 / 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 / 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 / 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]

Scientific areas

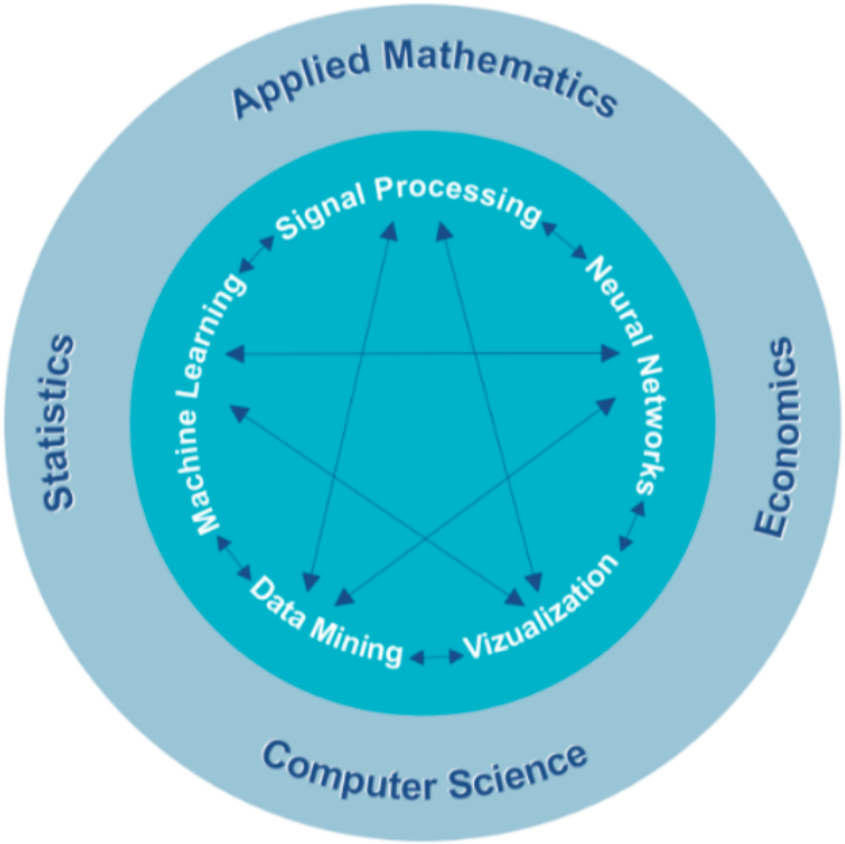


Figure: [Olshannikova et al., 2016]

Big data sources

Where does big data come from ?

Big data sources

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

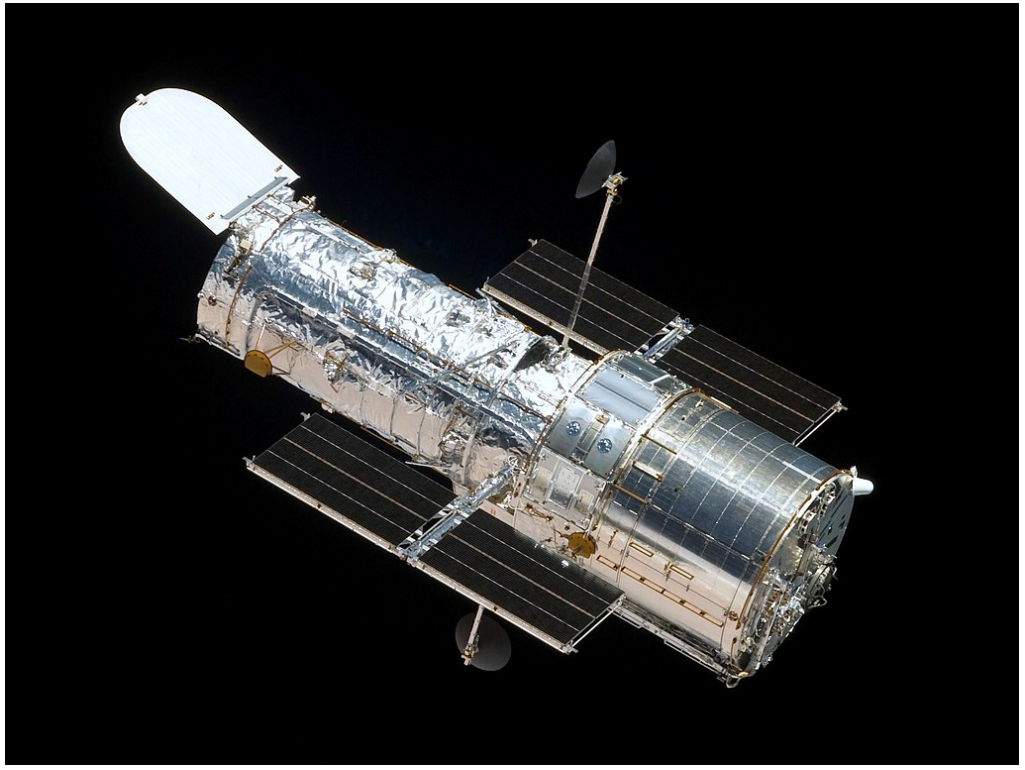


Figure: Hubble telescope (Wikipedia)

Big data sources

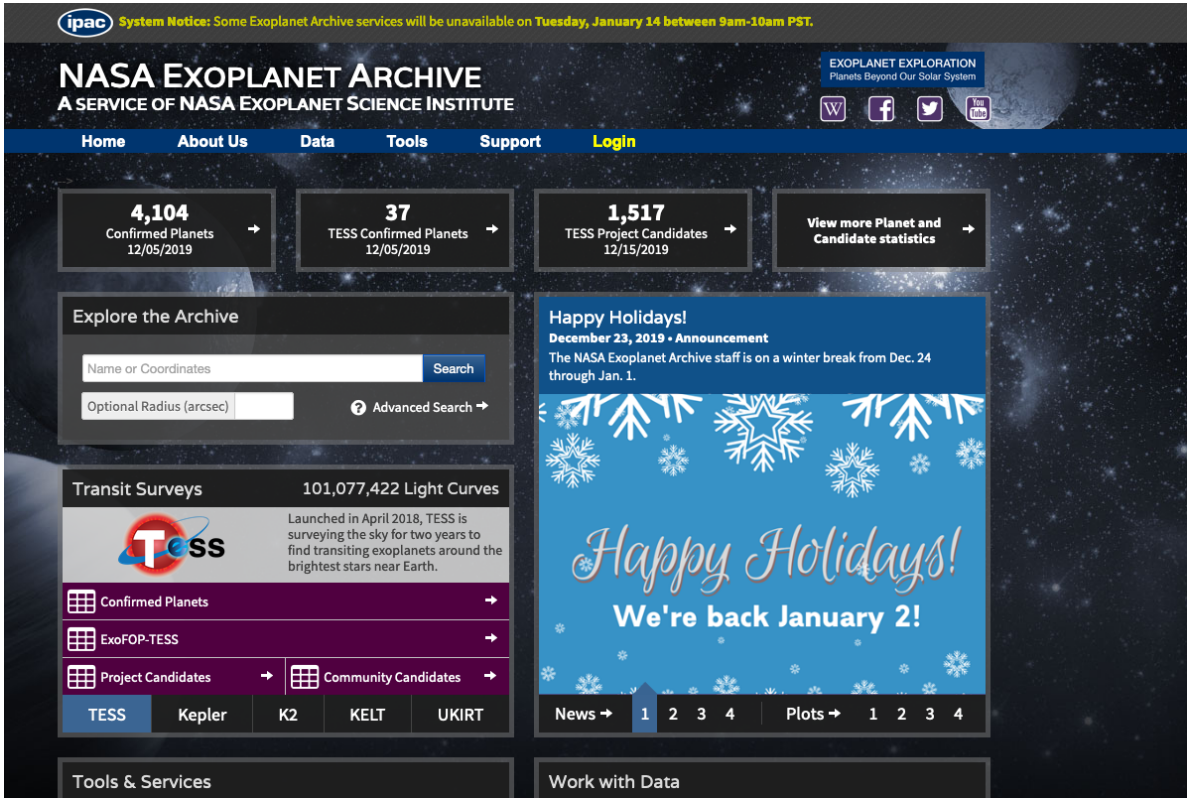


Figure: Nasa exoplanet archive
`https://exoplanetarchive.ipac.caltech.edu/`

Big data sources

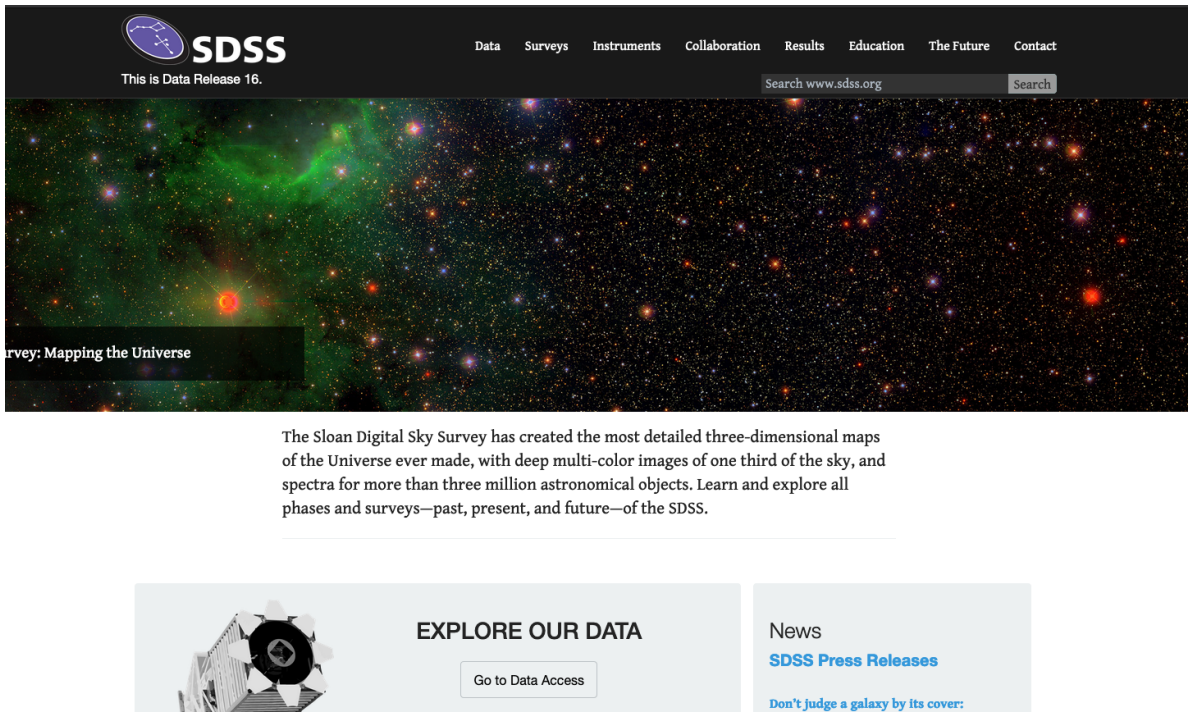


Figure: SDSS `https://www.sdss.org/`

Big data sources

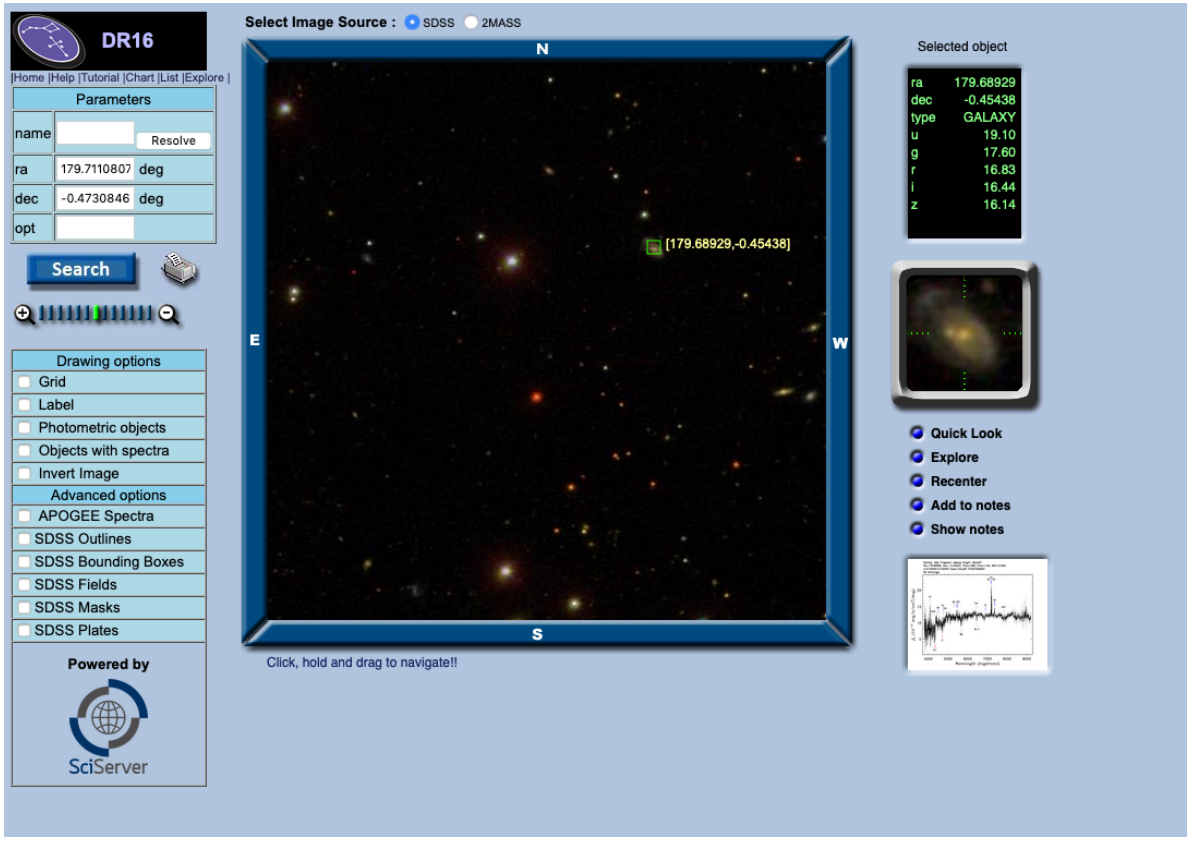


Figure: SDSS `https://www.sdss.org/`

Big data sources

Climatic data

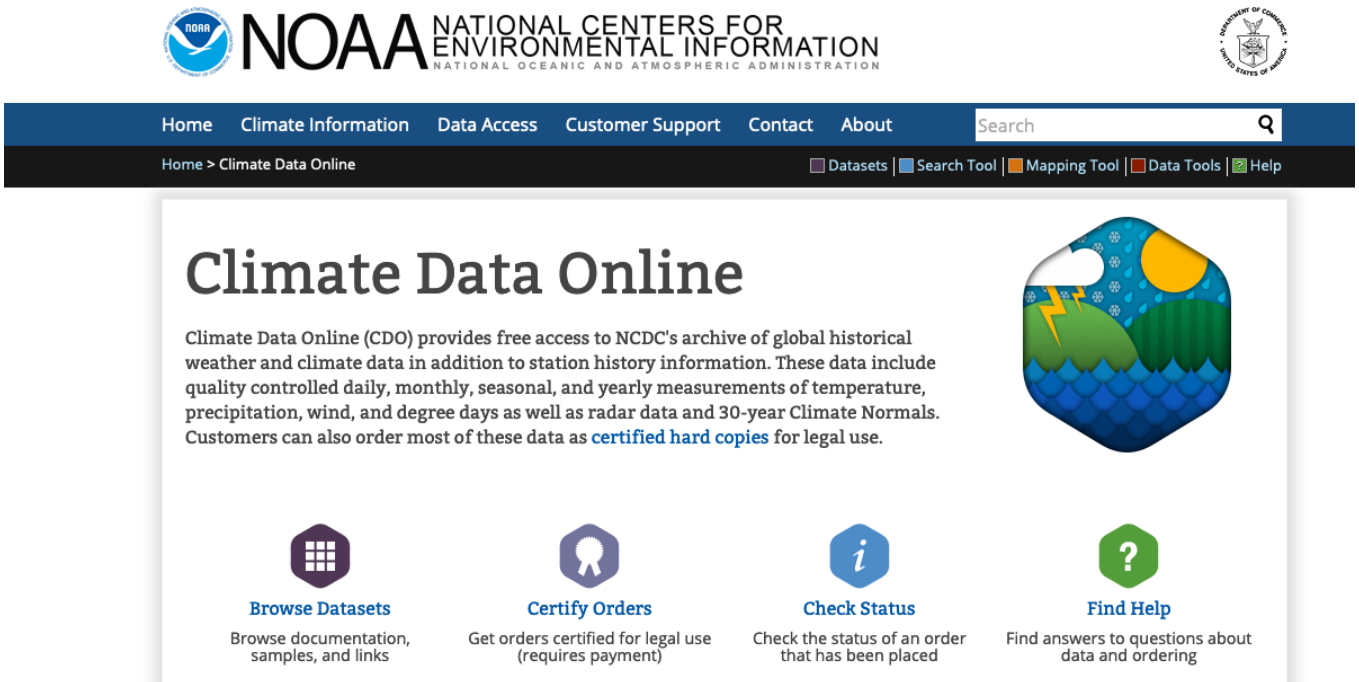


Figure: NOAA dataset `https://www.ncdc.noaa.gov/cdo-web/`

Big data sources

Climatic data

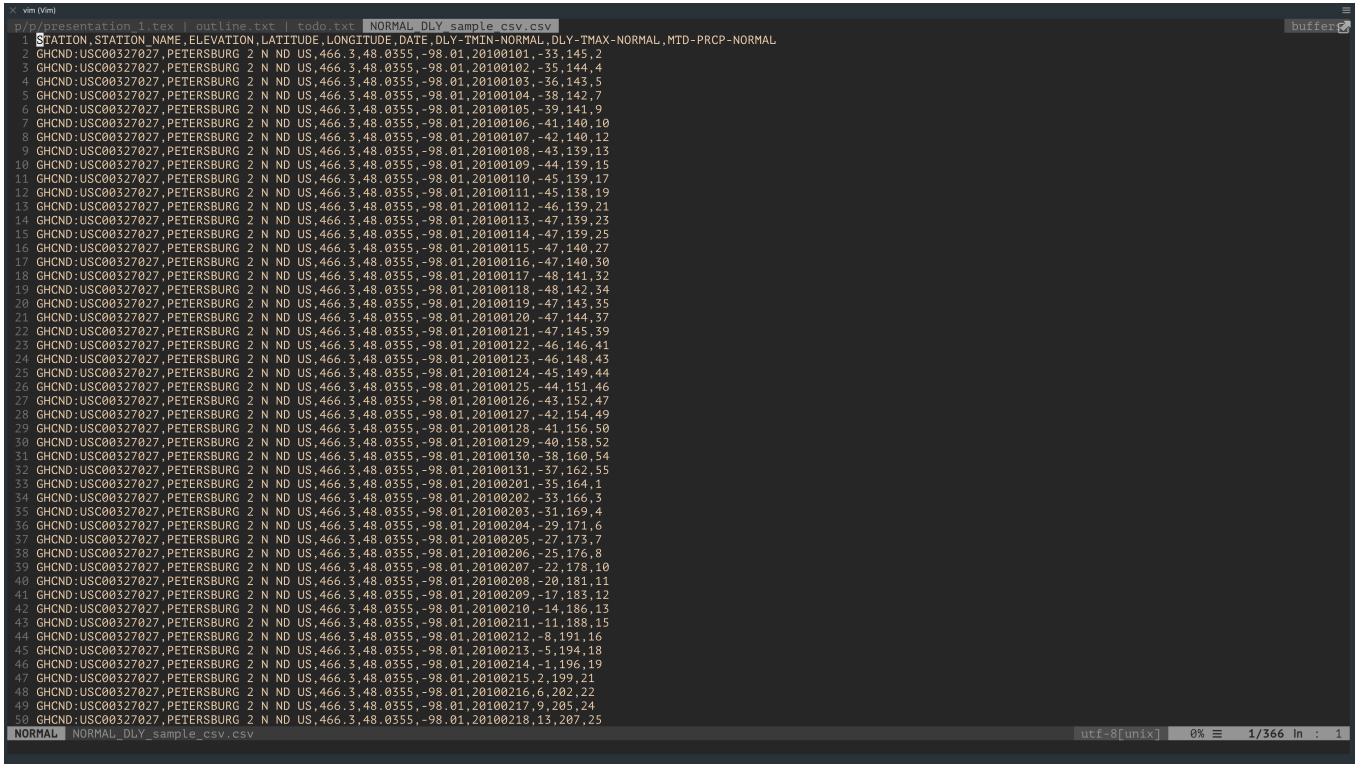


Figure: NOAA data sample. `https://www.ncdc.noaa.gov/cdo-web/`

Big data sources

Bioinformatics

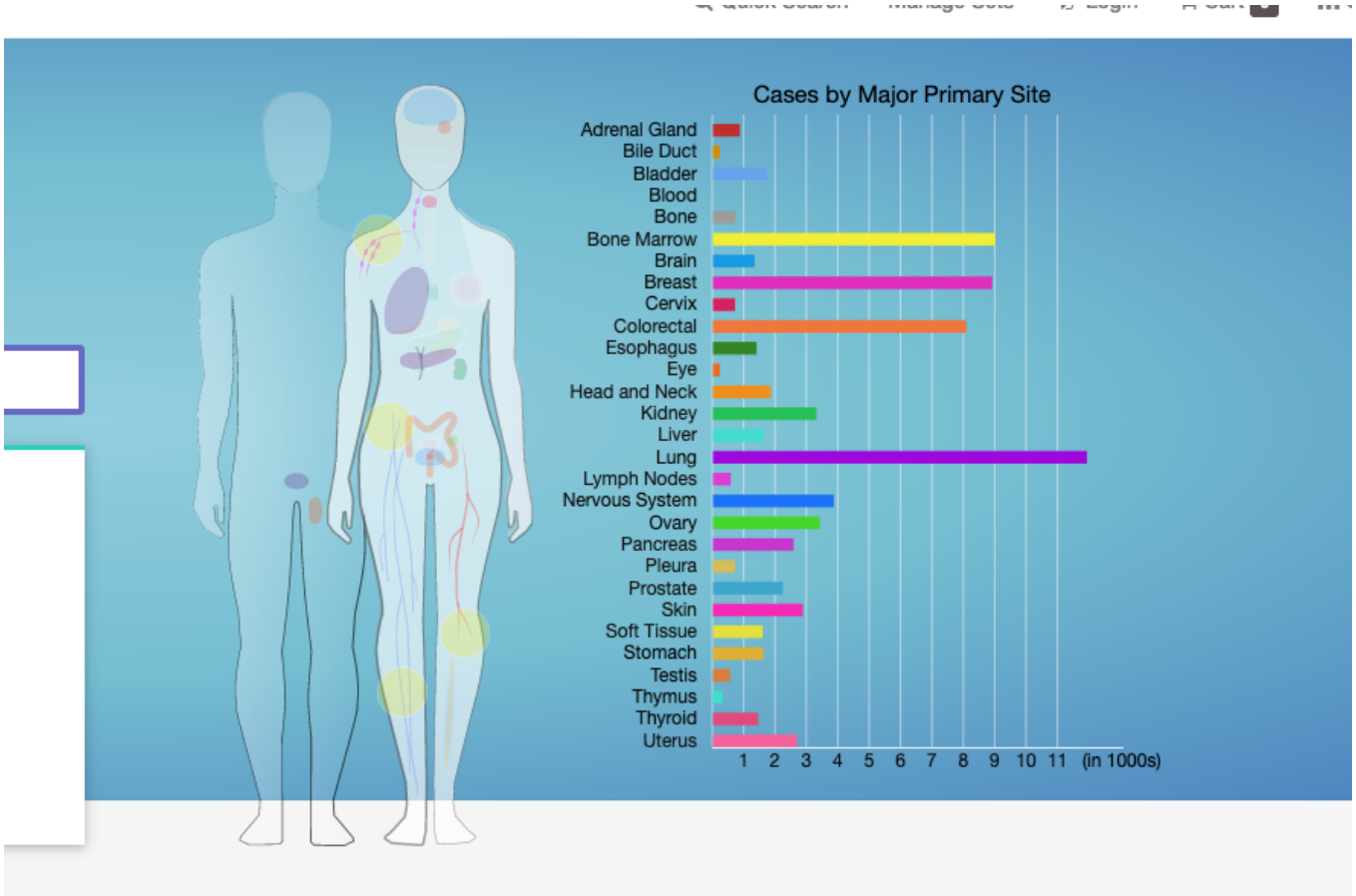


Figure: National Cancer Institute `https://portal.gdc.cancer.gov/`

Big data sources

Bioinformatics

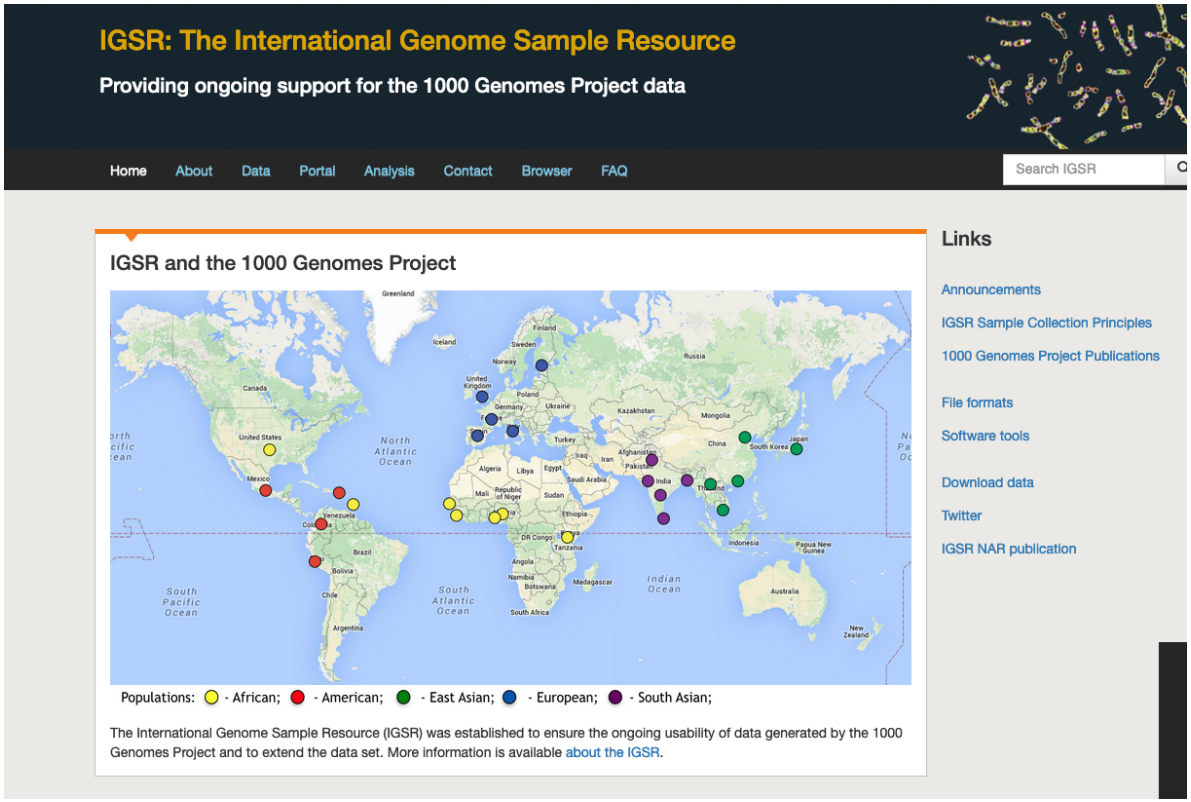


Figure: 1000 Genome project
`https://www.internationalgenome.org/`

Big data sources

The web.

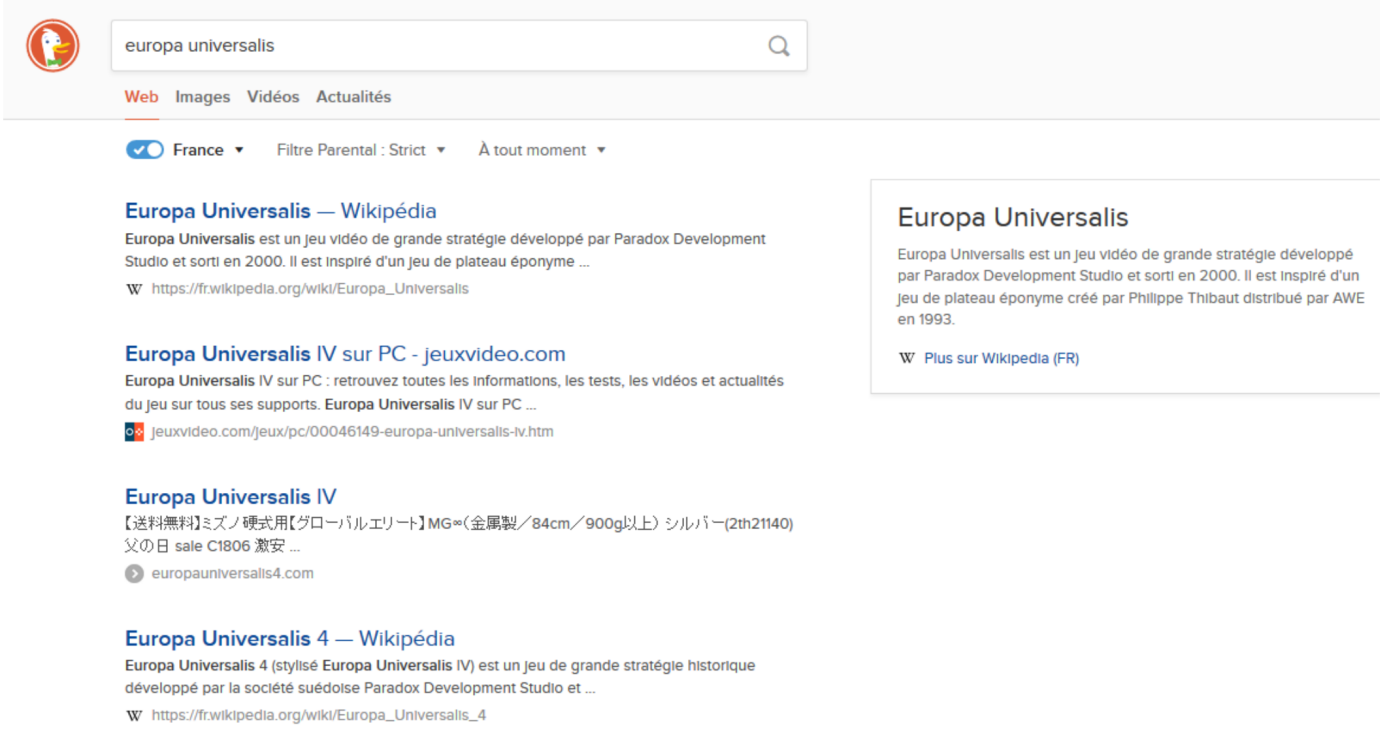
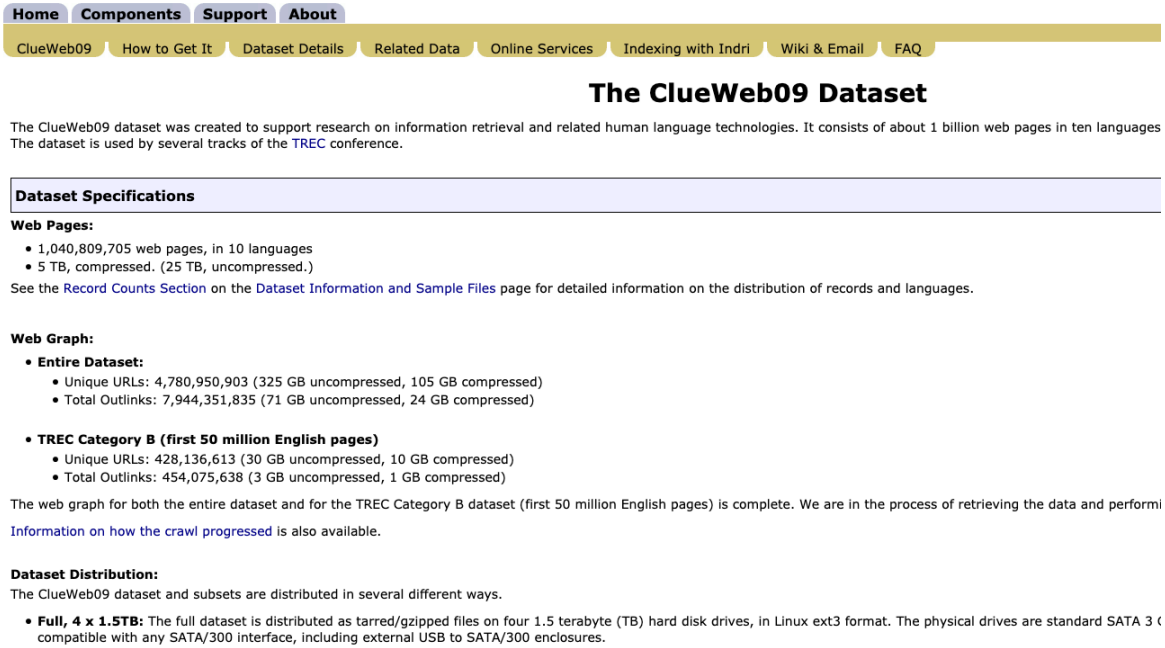


Figure: Duckduckgo web browser

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 performi

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

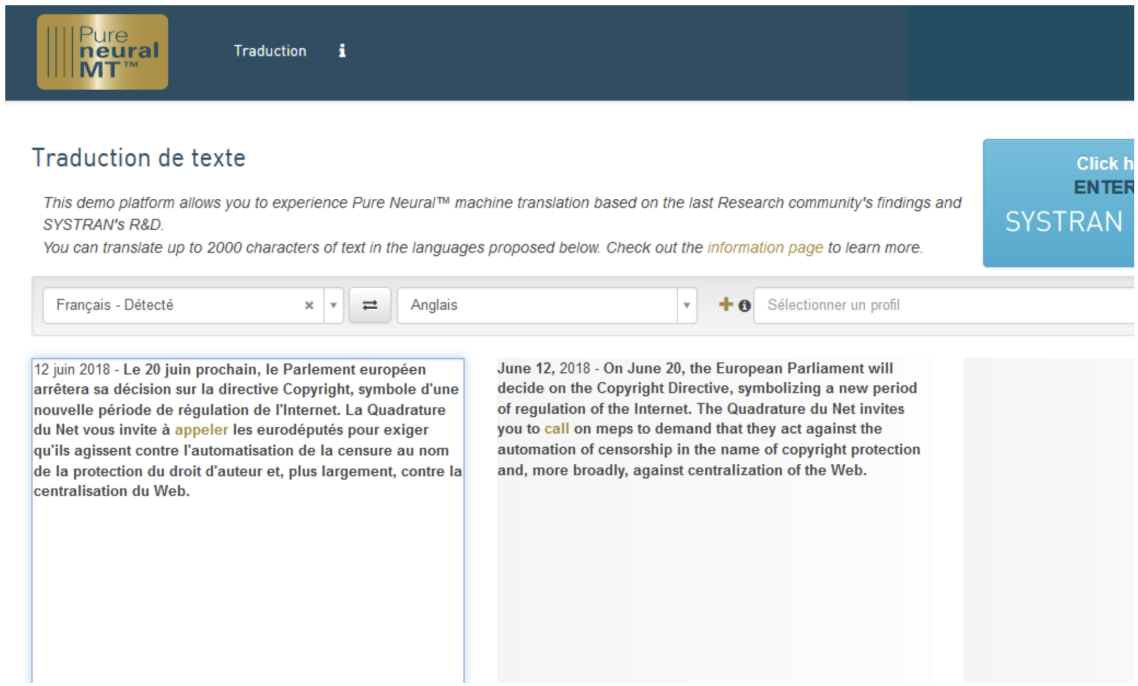


Figure: Text translation (SYSTRAN) `https://www.systransoft.com/fr/systran/technologie/pure-neural-machine-translation/`

Big data sources

Handwrtitten digits.

Figure: MNIST dataset [LeCun and Cortes, 2010]

Big data sources

Lexicometry.

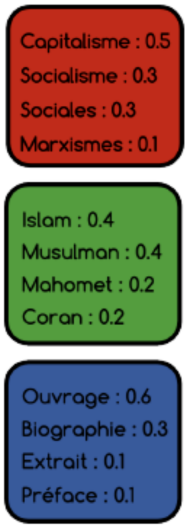


FIGURE : topics

FIGURE : document

Figure: Statistics on texts

Big data sources

Text, audio, video.

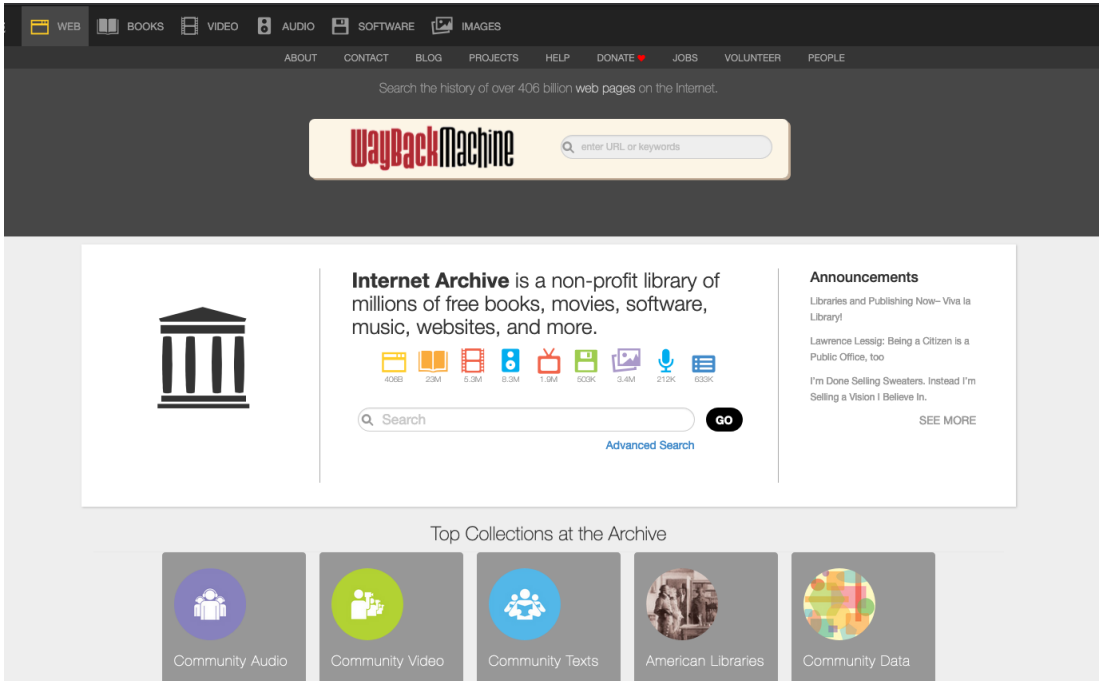


Figure: Archive dataset `https://archive.org/`

Big data sources

Image and video.

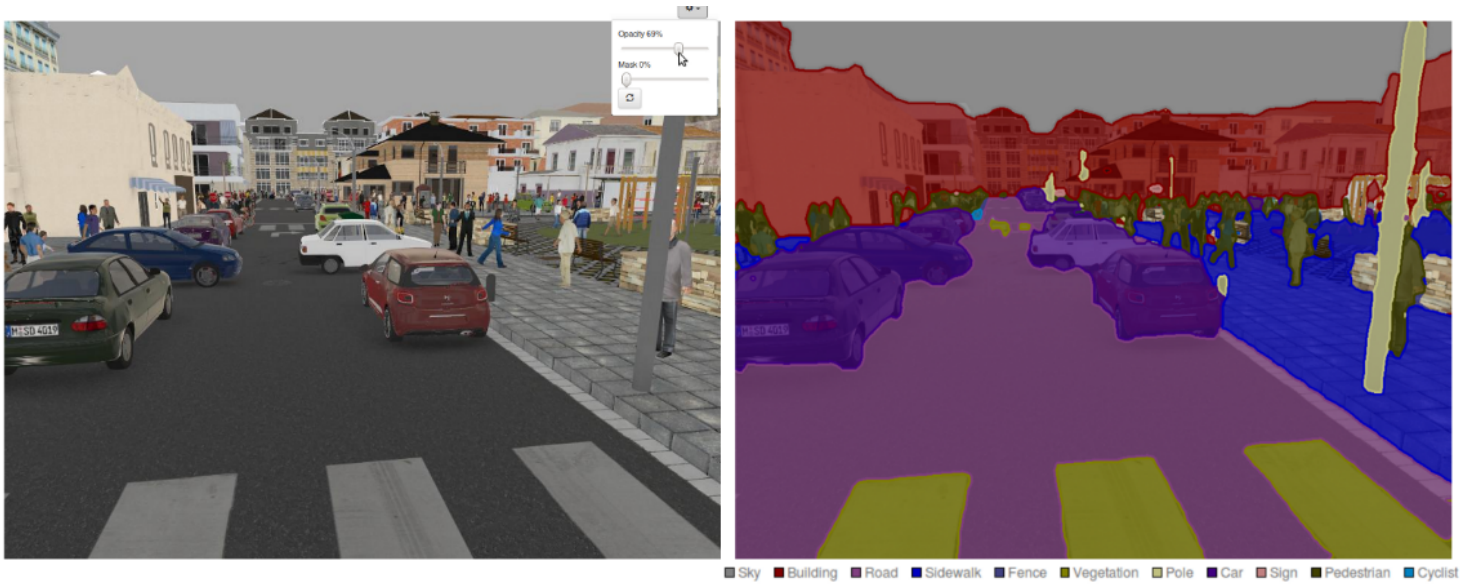


Figure: Source : Nvidia developer blog
`https://devblogs.nvidia.com/`

Big data sources

Financial data (time series)

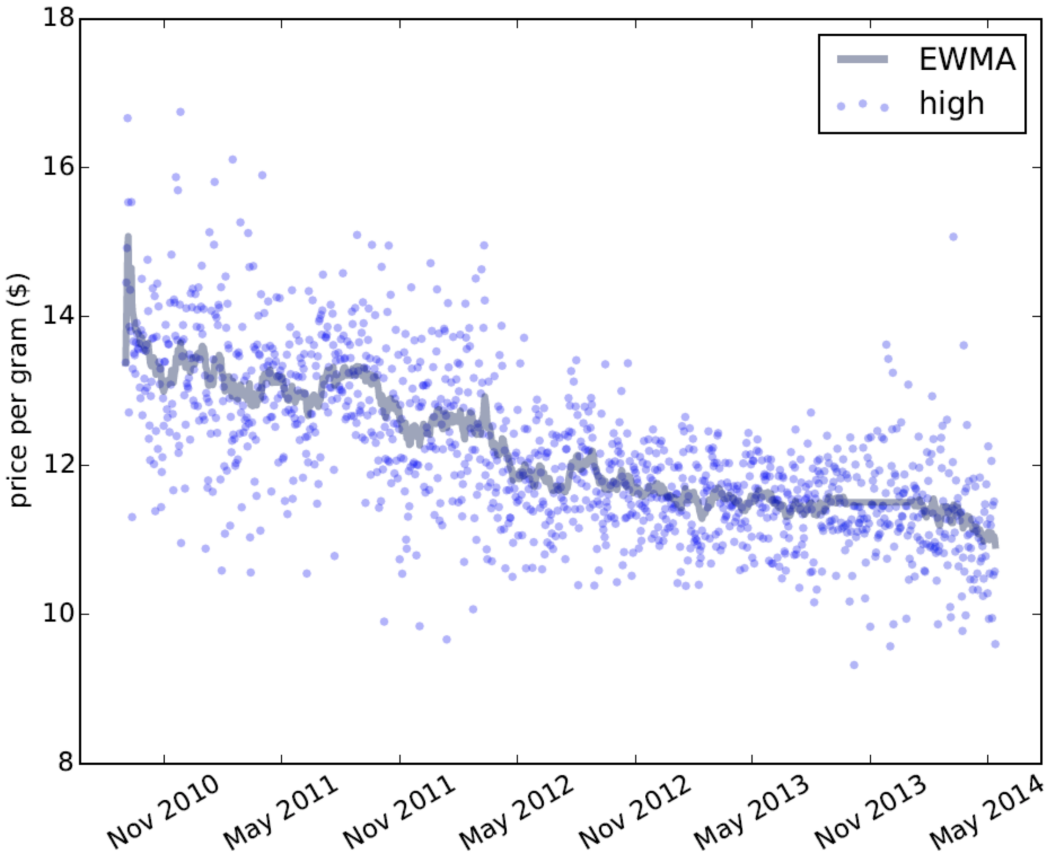


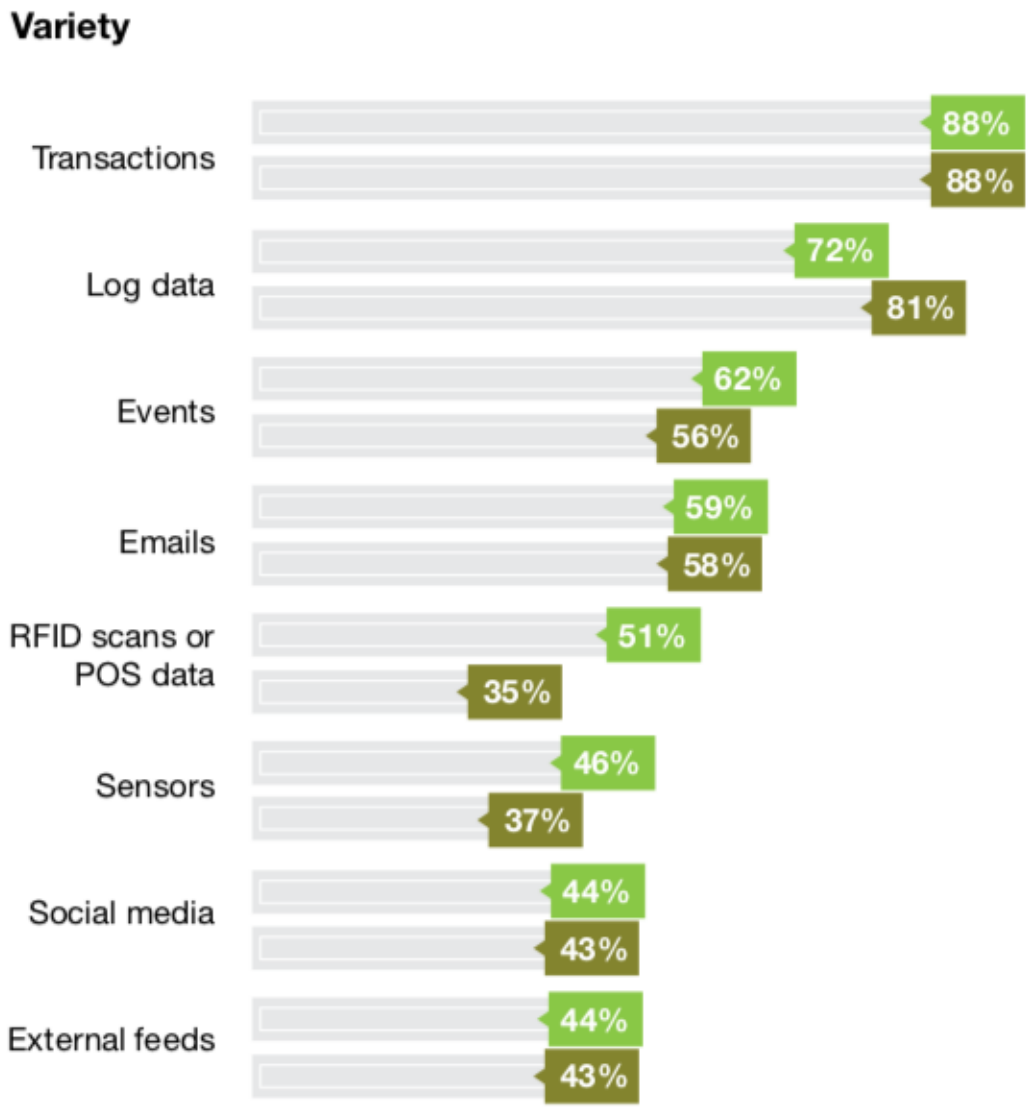
Figure: Price of some plant

Big data sources

Game data

Figure: Europa Universalis
https://fr.wikipedia.org/wiki/Europa_Universalis

Big data sources



Orders of magnitude

Let us discuss the size of data.
Order of magnitude means "Ordre de grandeur".

Orders of magnitude

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

Orders of magnitude

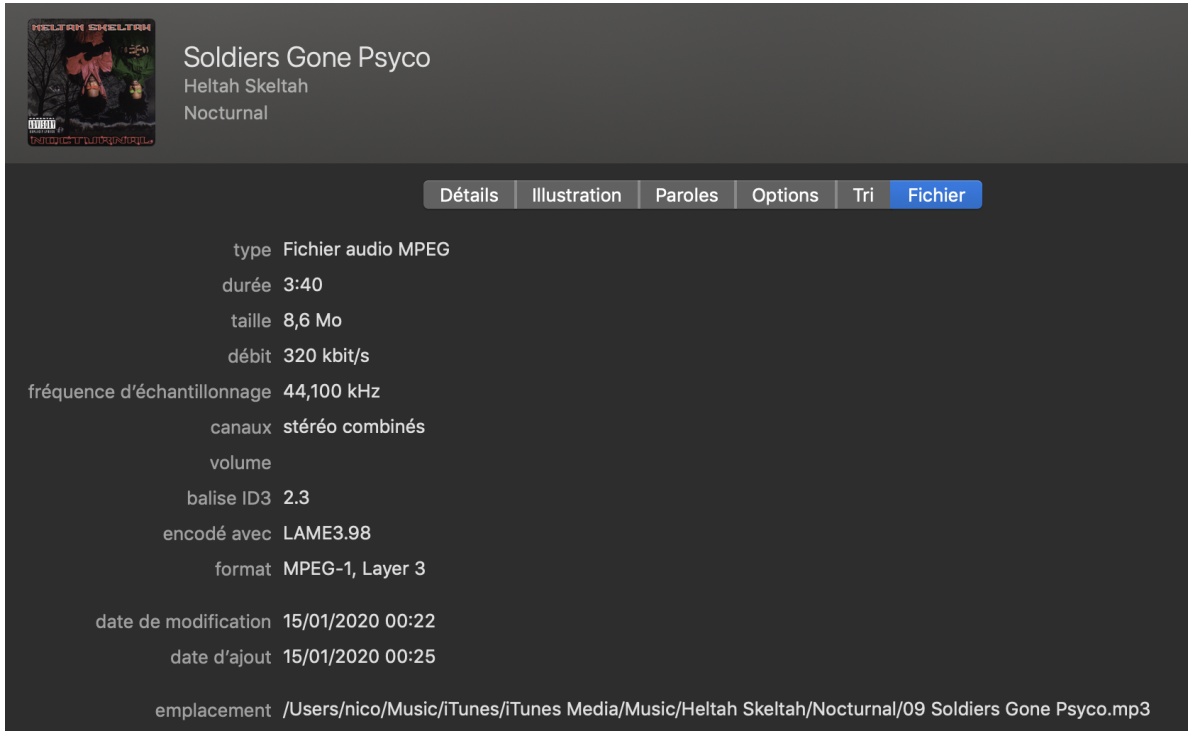


Figure: Approximately 8 Mo

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 Glober 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
```

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
```

Orders of magnitude

PDF slides for this presentation : Approximately 20 Mo

Orders of magnitude

$2,5$ hours MKV movie ?

Orders of magnitude

$2,5$ hours MKV movie ? Around 3 Go.

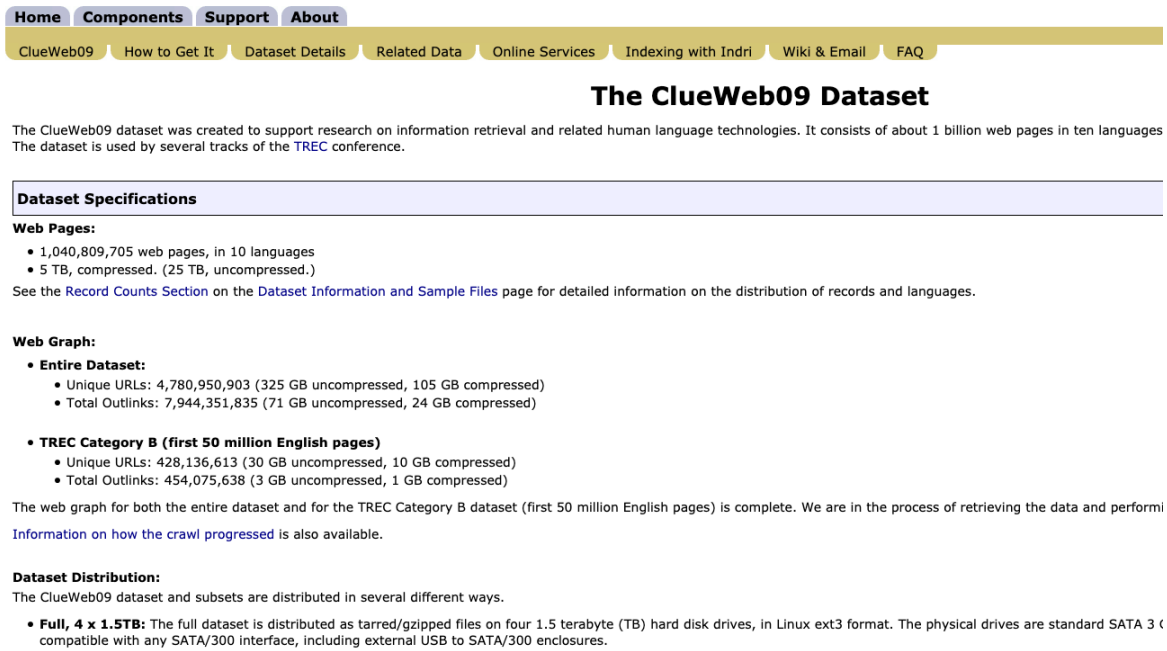
Orders of magnitude

Desktop computer harddrive : 1To.

Orders of magnitude

Big datasets ?

Orders of magnitude



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 performi

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

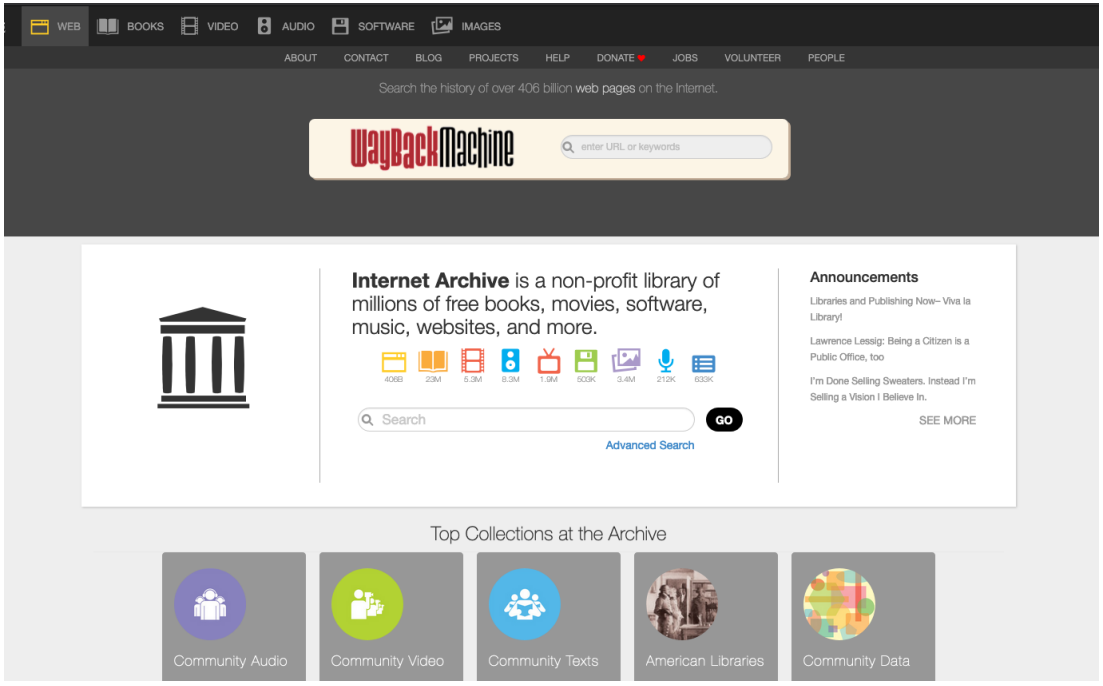


Figure: **20 To** `https://archive.org/`

Orders of magnitude

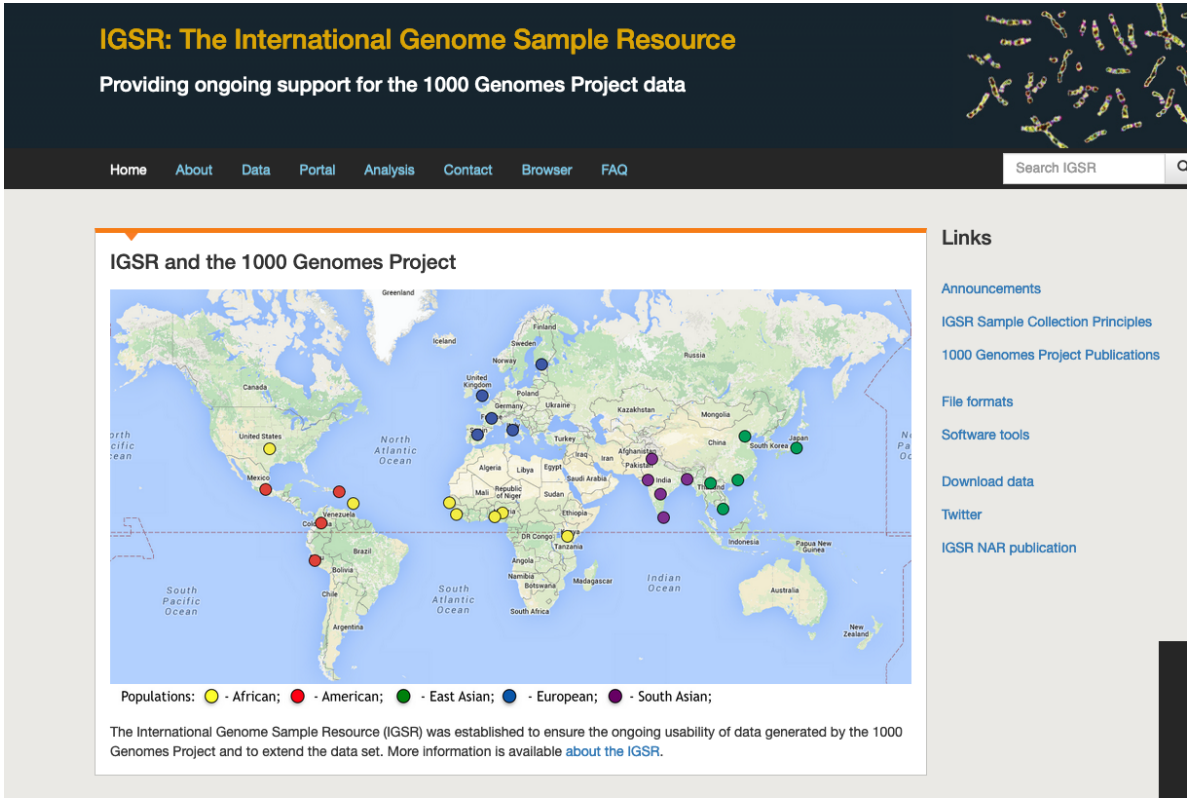


Figure: **260 To** `https://www.internationalgenome.org/`

Orders of magnitude

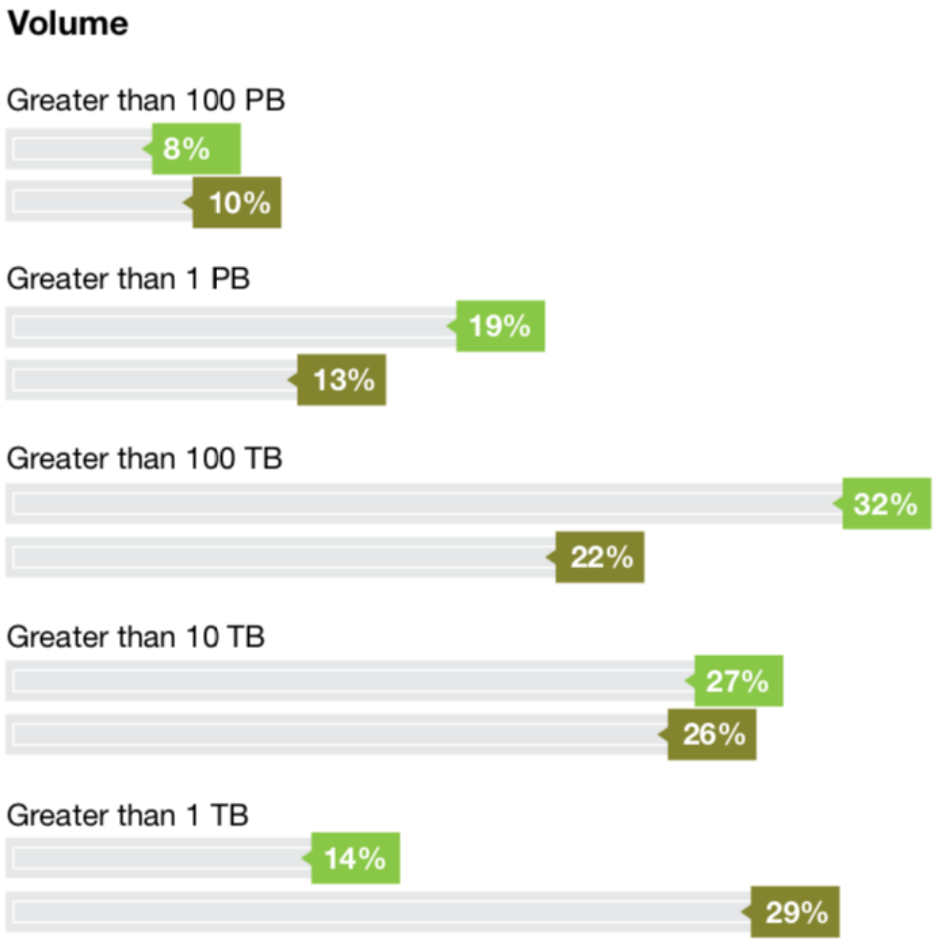


Figure: [IBM,]

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

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	?

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

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]

Who uses big data ?

Who uses big data ?

Mostly scientists and companies.

Who uses big data ?

Mostly scientists and companies.
There is now a market of trading data.

Usage of big data

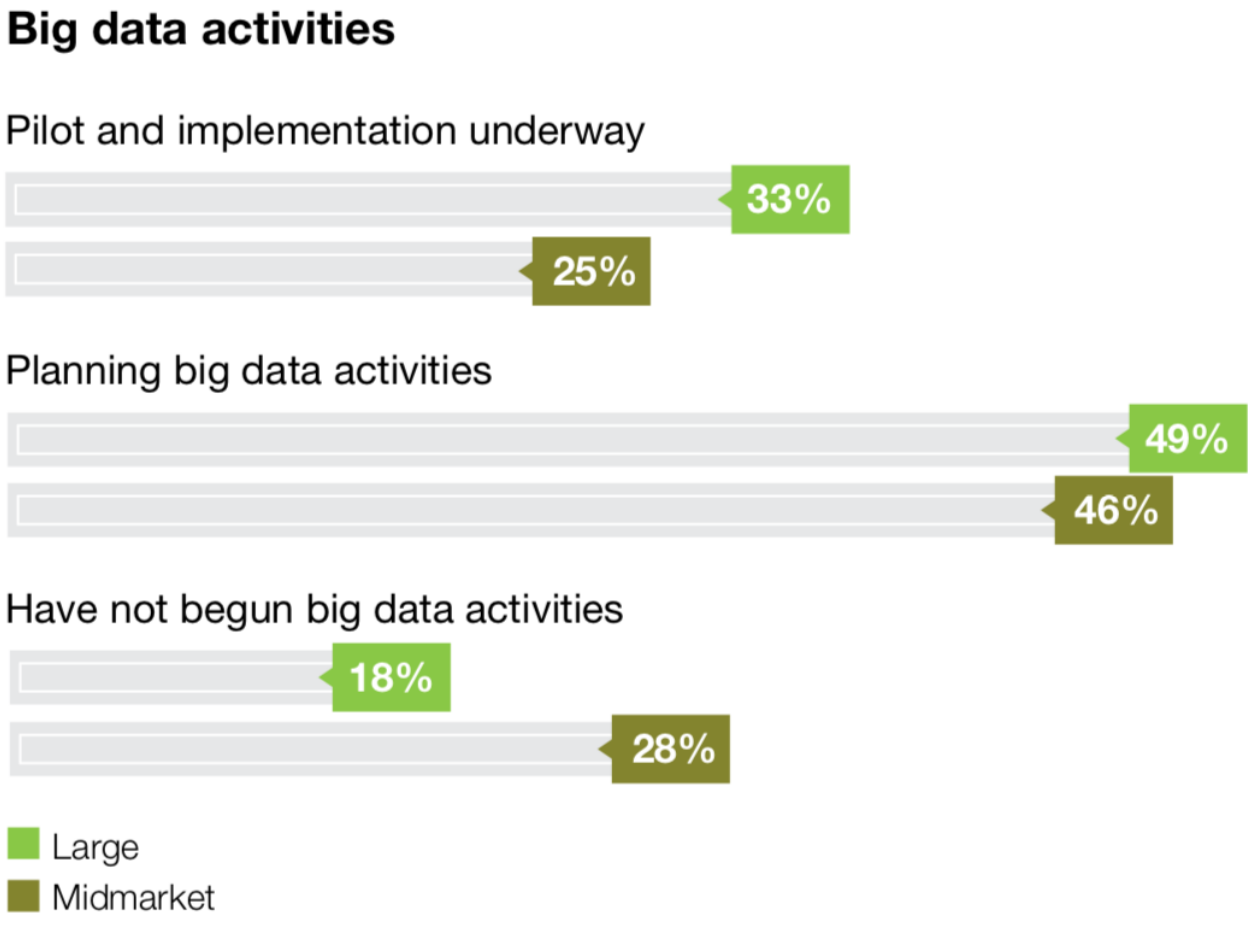


Figure: Big data activites in organizations surveyed by IBM [IBM,]

Applications of big data

Let us discuss some applications of big data.

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.

Example 2 : banking

A bank looses 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 it's offer to those.

Example 3 : web marketing

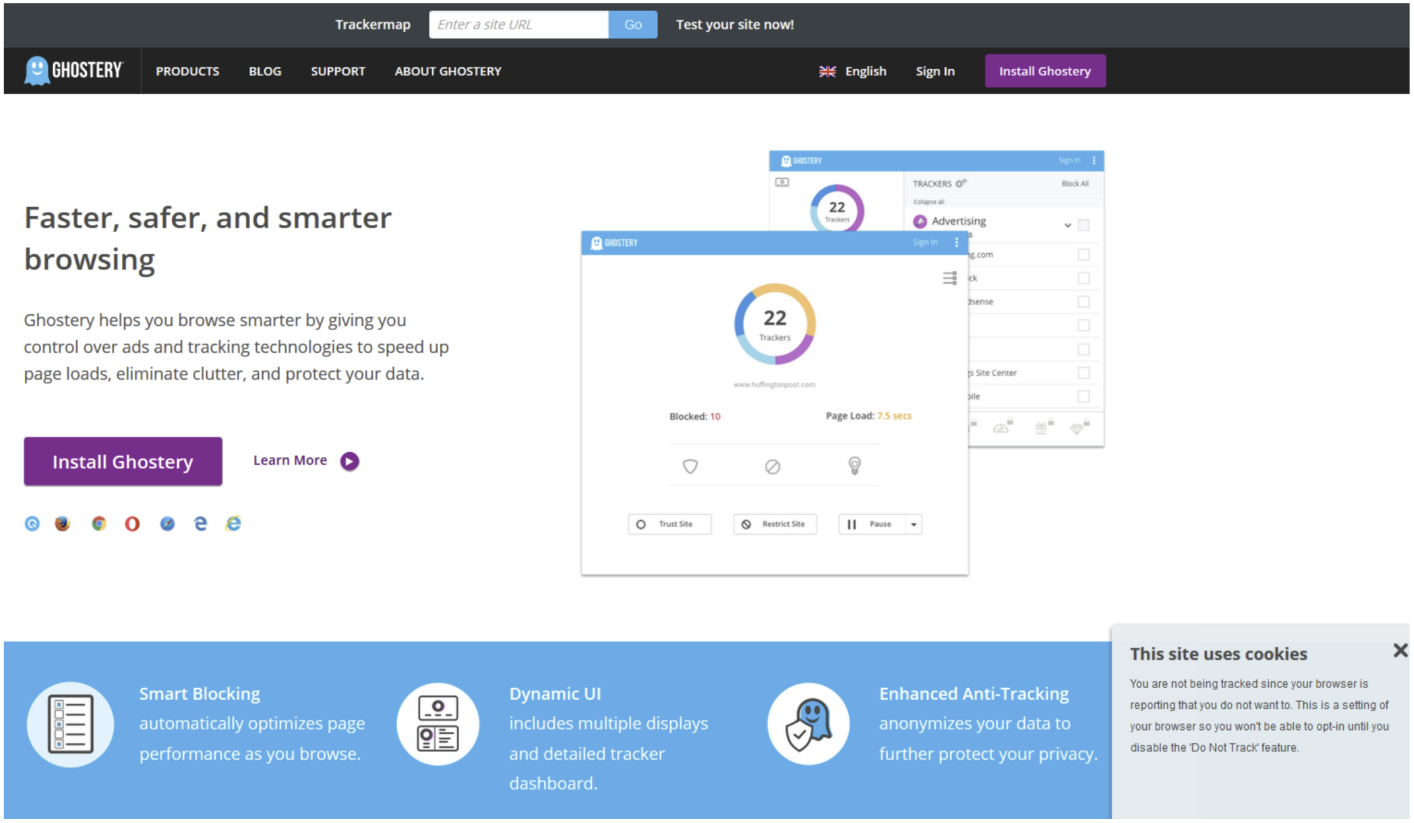


Figure: Ghostery

Other applications

In the future there might be more, unexpected applications.

Data visualization

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

Data visualization

Those datasets are often so large that it is extremely hard to analyze them.
Depending on the information seeked in the dataset, it is sometimes helpful to make a **visualization** of the dataset.

Data visualization

Those datasets are often so large that it is extremely hard to analyze them.
Depending on the information seeked 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.

Big data visualization constraints

The data are often :

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

Big data visualization constraints

The data are often :

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

In that respect, a visualization system should:

- offer the possibilty 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.

Big data visualization constraints

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

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 ojectives of the processing.

Example 1

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

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

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

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

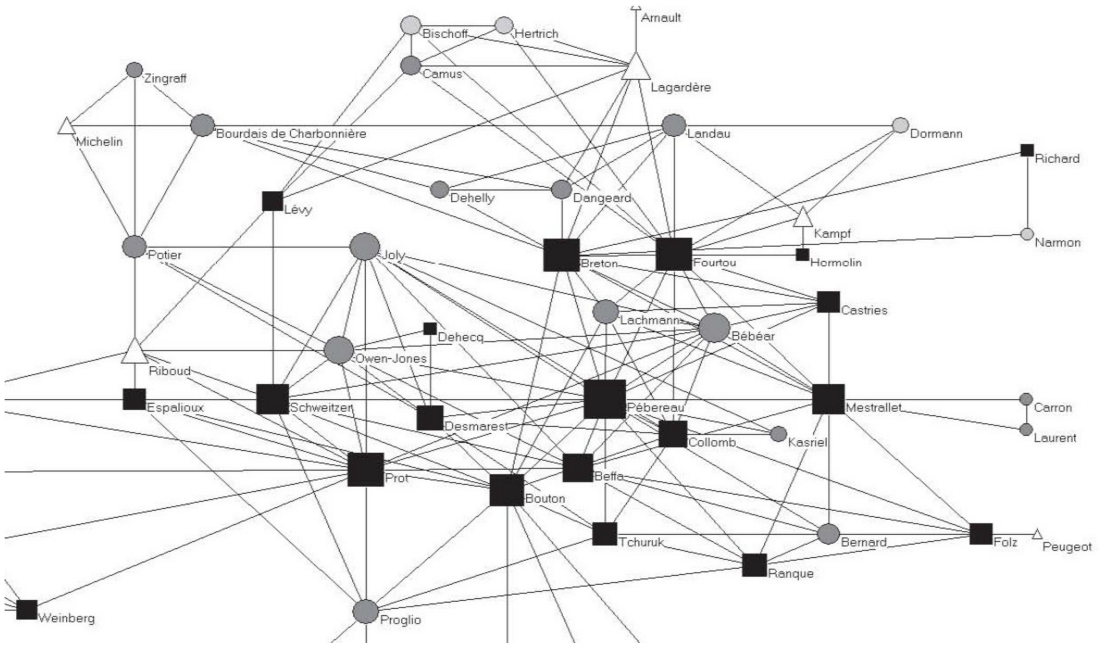


Figure: Graph

Example 3

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

Figure: High dimensional data

Example 4

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

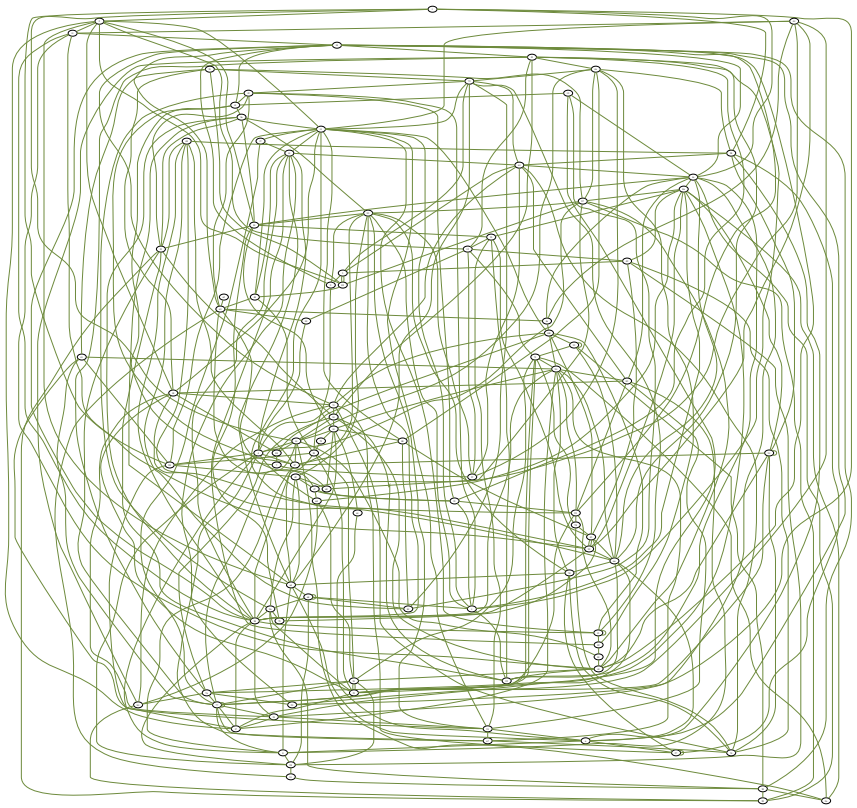


Figure: Dense graph

Conclusion

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

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 modeling

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

Data modeling

Given a unknown dataset, we must **process** it in order to extract information from it.
Before studying visualization, we must discuss the general processing of data.

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

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

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)

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

Several paradigms exist

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

Several paradigms exist

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

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 as **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 as **robot**)
- The paradigms can be mixed : a Reinforcement Learning system can have a supervised learning part.

Supervised learning

Observations:

- Empirical data points $(\tilde{x_1}, \ldots, \tilde{x_n})$

Supervised learning

Observations:

- Empirical data points $(\tilde{x_1}, \ldots, \tilde{x_n})$
- Each data point $\tilde{x_i}$ is associated with a **label** $\tilde{y_i}$, so we also have a set of labels $(\tilde{y_1}, \ldots, \tilde{y_n})$

Supervised learning

Observations:

- Empirical data points $(\tilde{x_1}, \ldots, \tilde{x_n})$
- Each data point $\tilde{x_i}$ is associated with a **label** $\tilde{y_i}$, so we also have a set of labels $(\tilde{y_1}, \ldots, \tilde{y_n})$

Objective: learn to predict the label $\tilde{y_p}$ of a new data point $\tilde{x_p}$.

Supervised learning

Observations:

- Empirical data points $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$
- Each data point $\tilde{x_i}$ is associated with a **label** $\tilde{y_i}$, so we also have a set of labels $\tilde{y} = (\tilde{y_1}, \ldots, \tilde{y_n})$

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 $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$
- Each data point $\tilde{x_i}$ is associated with a **label** $\tilde{y_i}$, so we also have a set of labels $\tilde{y} = (\tilde{y_1}, \ldots, \tilde{y_n})$

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 $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$
- Each data point $\tilde{x_i}$ is associated with a **label** $\tilde{y_i}$, so we also have a set of labels $\tilde{y} = (\tilde{y_1}, \ldots, \tilde{y_n})$

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 $\tilde{x'}$.

Supervised learning

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

Tide level

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

Tide Level

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

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.

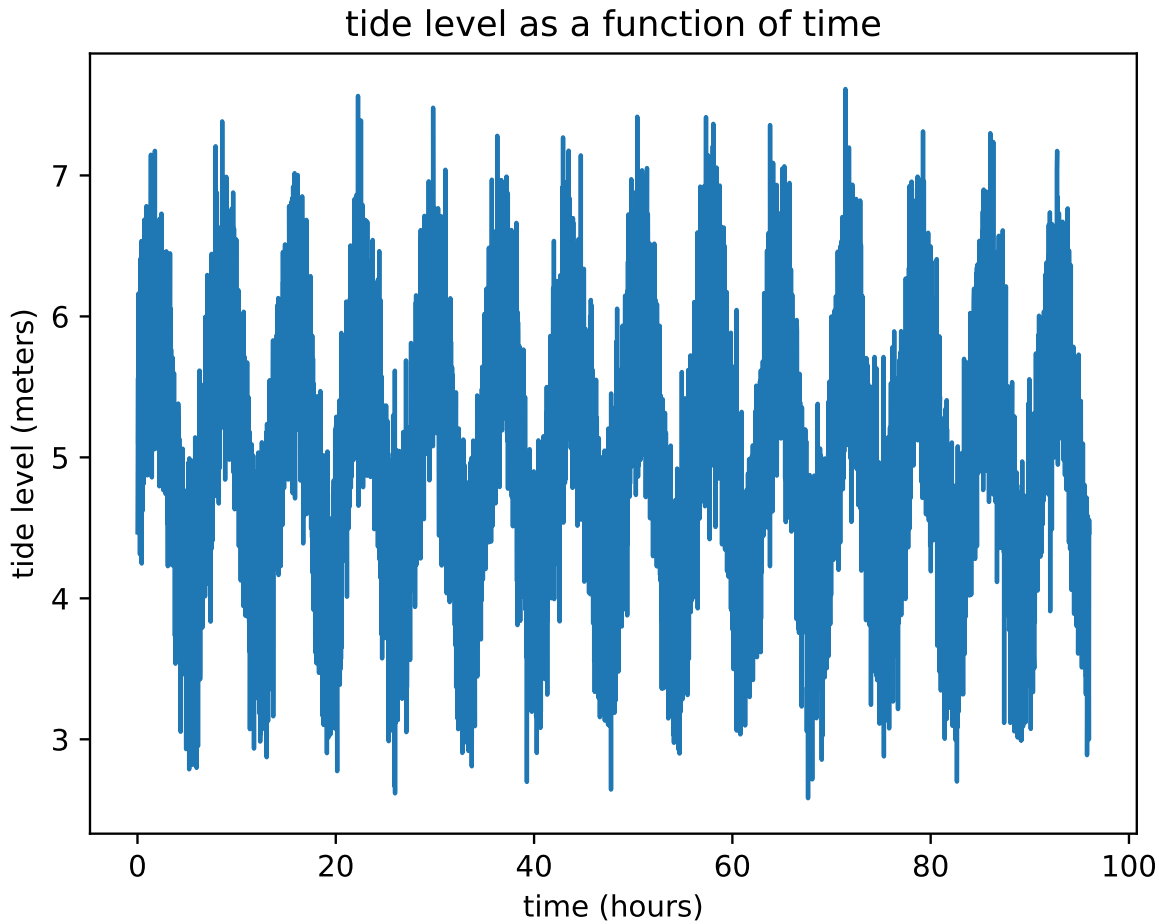
Tide level

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**.

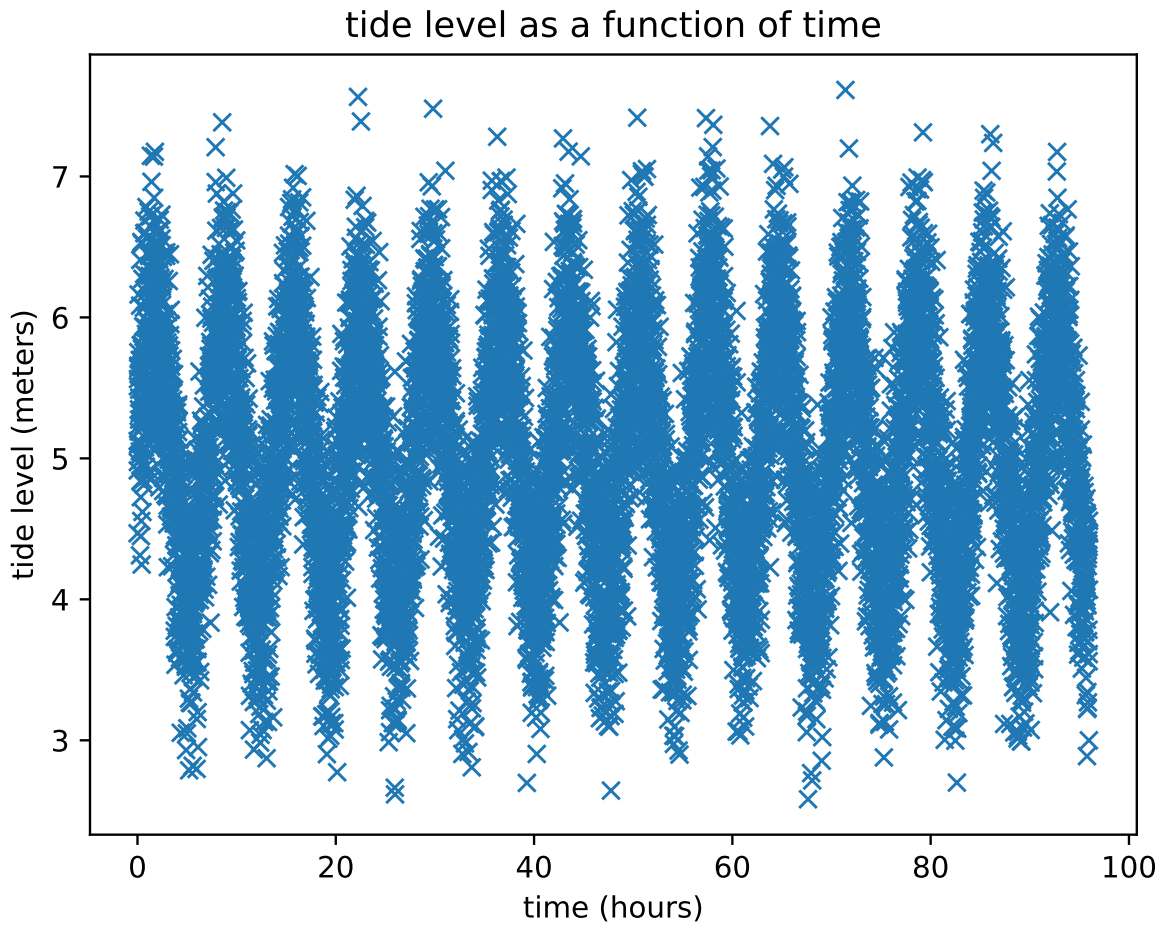
Tide level

Exercice 2 : **Plotting the data**



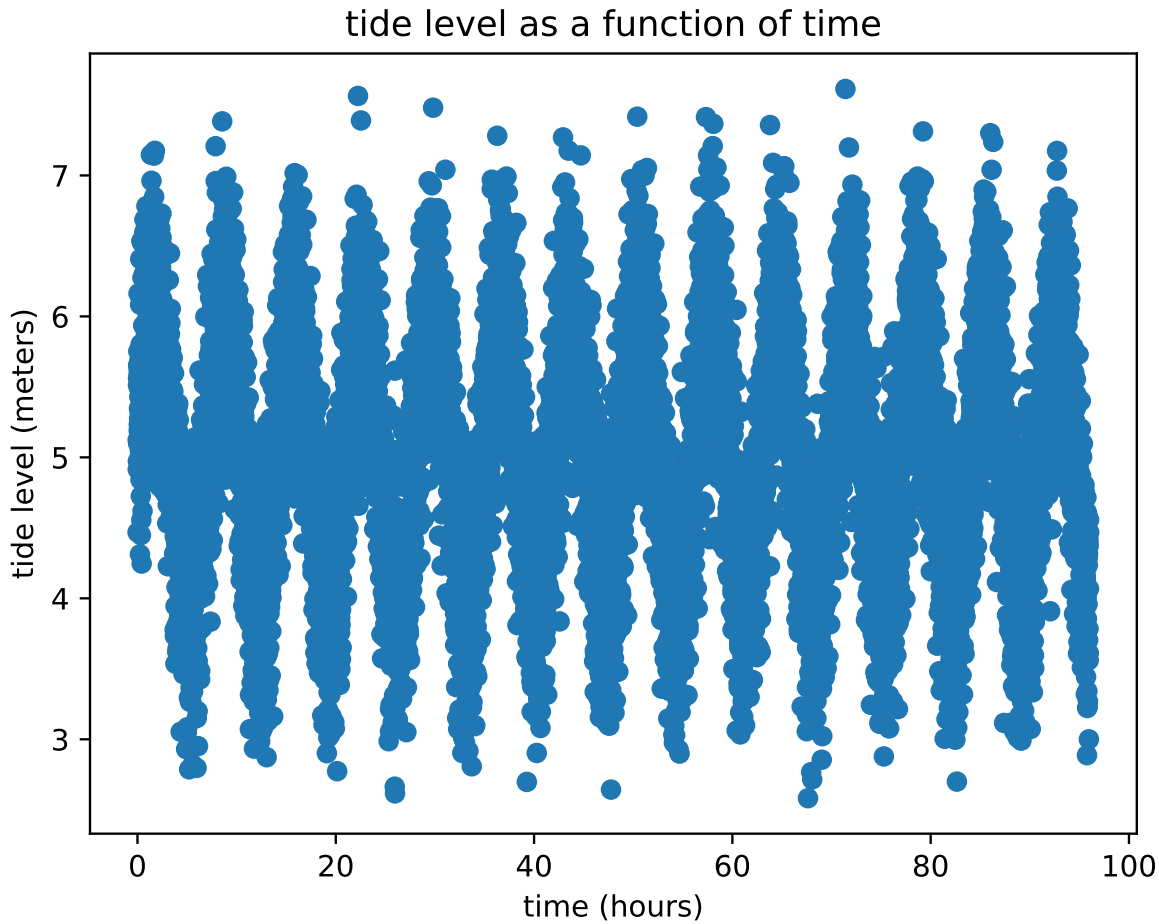
Tide level

Exercice 2 : **Plotting the data**



Tide level

Exercice 2 : **Plotting the data**



Tide level

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

Tide level

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

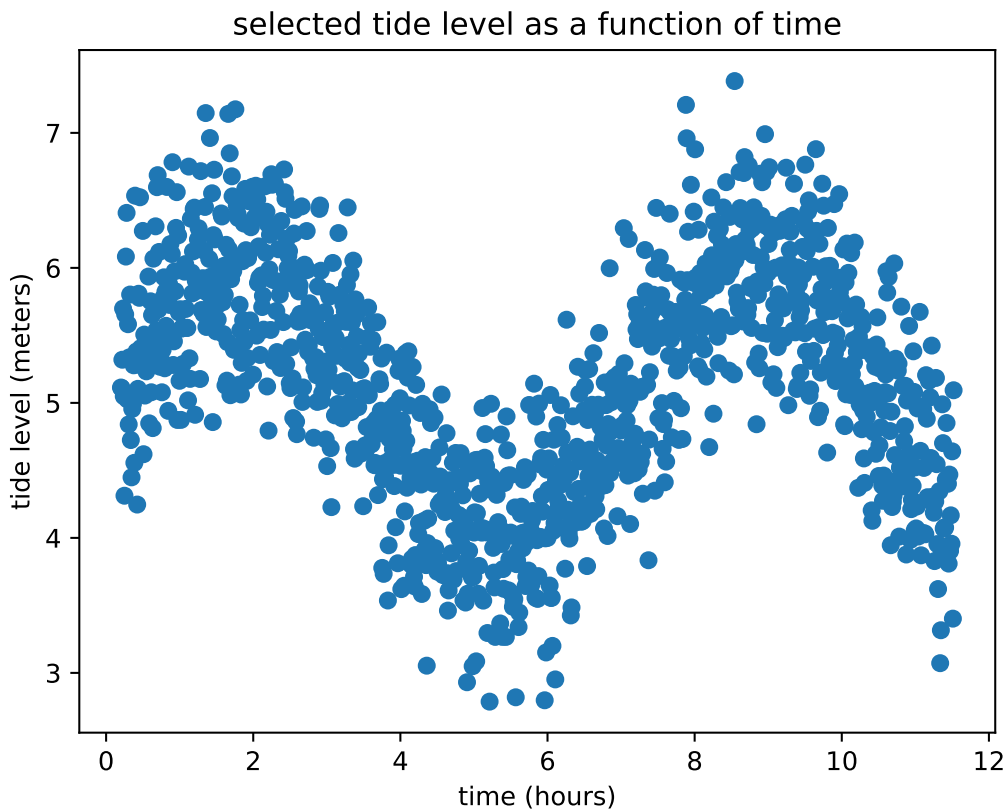


Figure: tide level

Tide level

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

Tide level

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

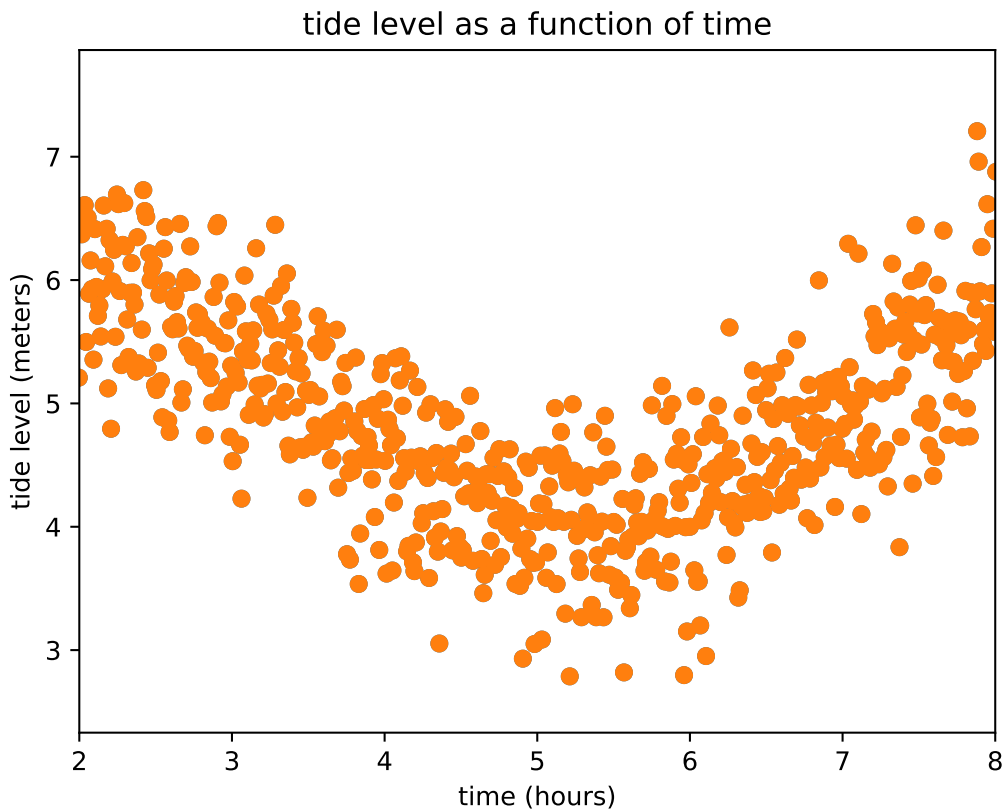


Figure: tide level

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 ?

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 ?

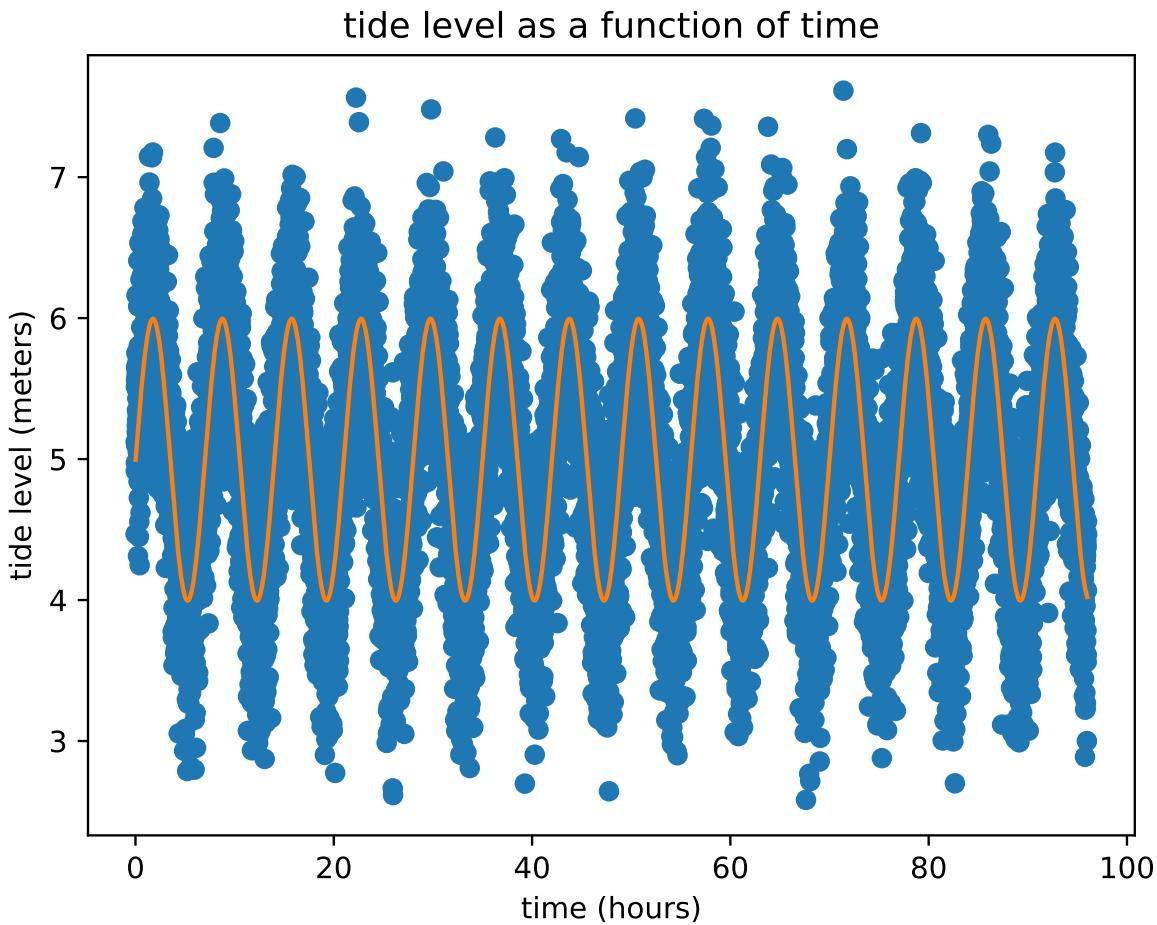
Tide level

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:

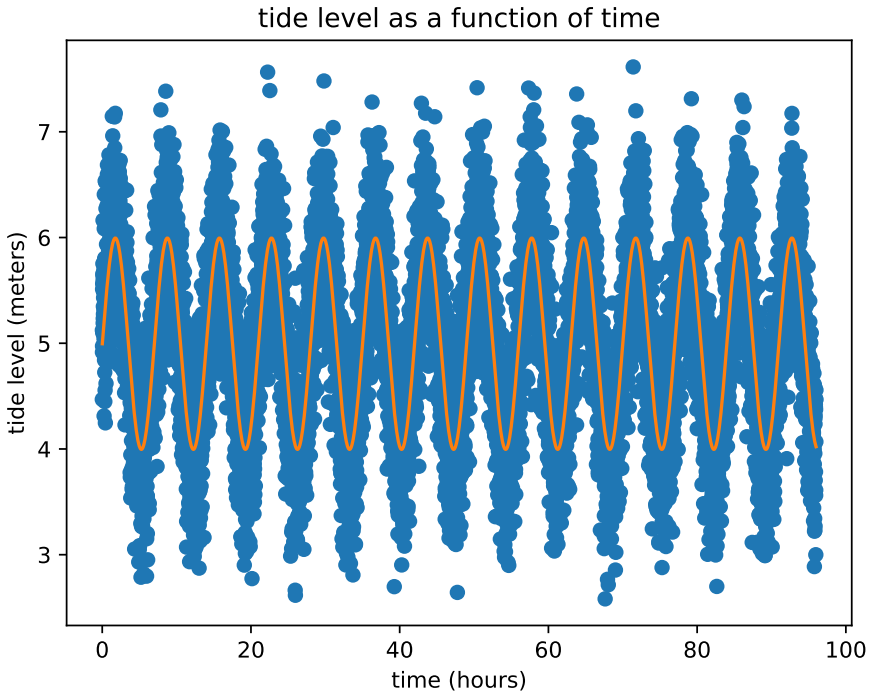
- Amplitude
- pulsation (analog of frequency)
- phase
- offset

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 ?

Tide level


tide level as a function of time
tide level (meters)
7
6
5
4
3
0
20
40
60
80
100
time (hours)

Tide level



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

Interpolation

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

Unsupervised learning

Observations:

- Empirical data points $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$

Unsupervised learning

Observations:

- Empirical data points $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$

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

Unsupervised learning

Observations:

- Empirical data points $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$

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

Hypotheses:

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

Unsupervised learning

Observations:

- Empirical data points $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_n})$

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 $\tilde{x'}$

Unsupersived 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.**

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**.

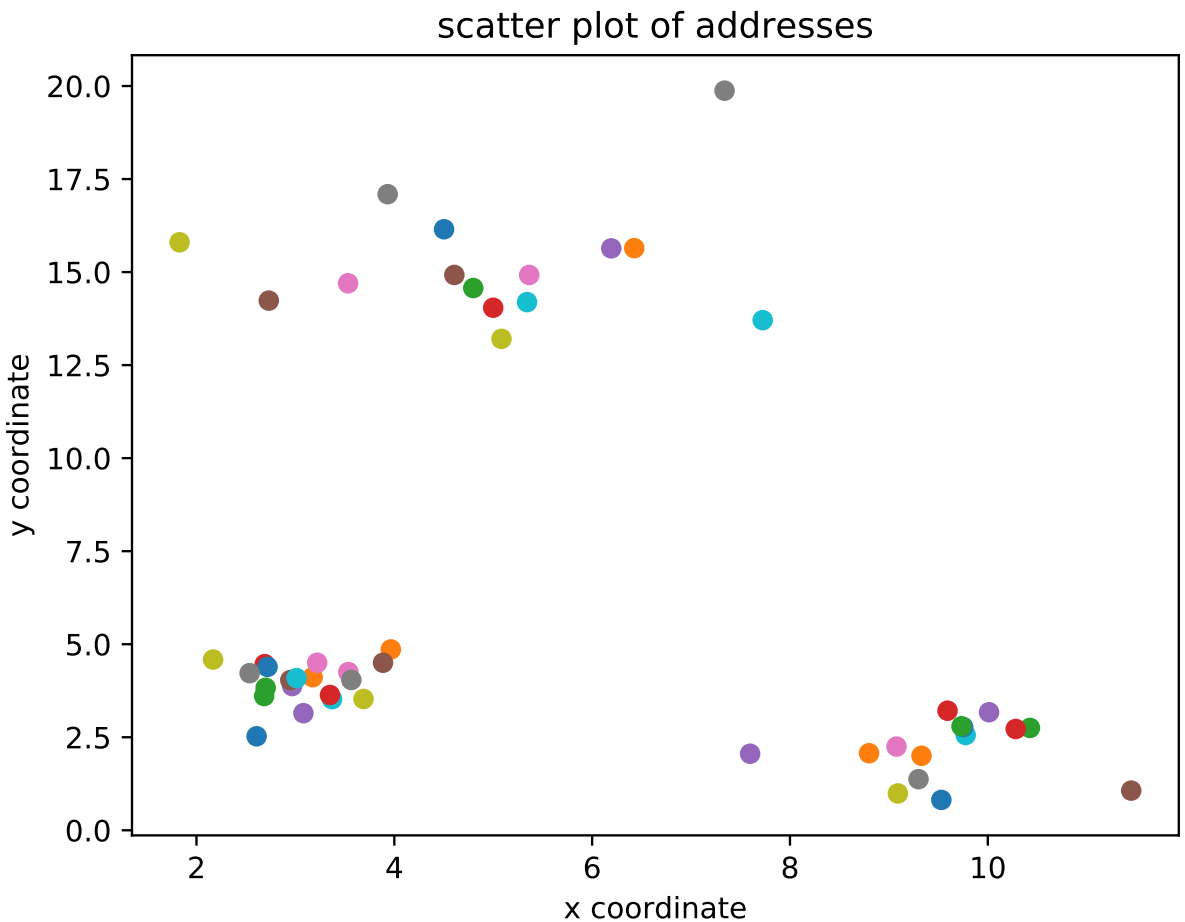
It consists in showing the points at a position corresponding to their coordinates.

Several methods are possible.

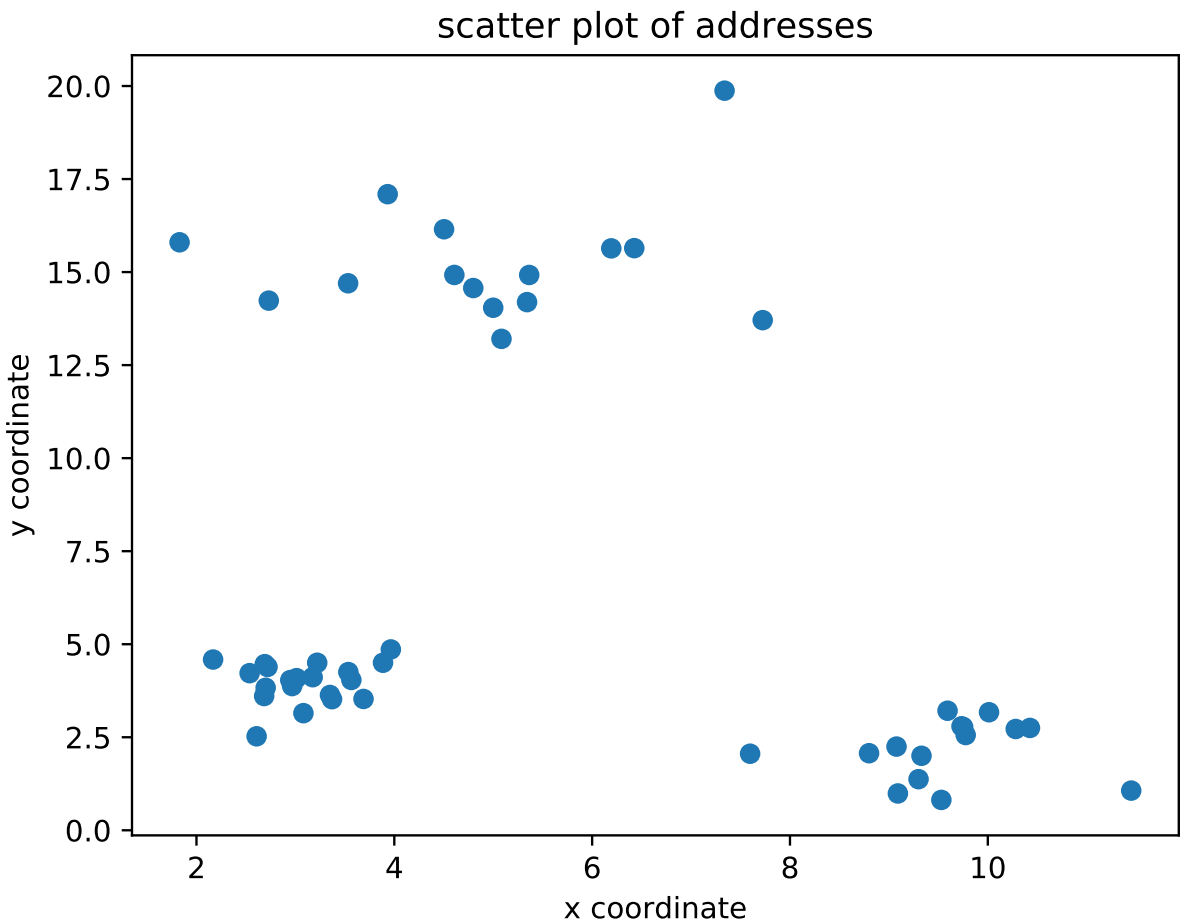
Scatter plots

Seaborn lib: `https://seaborn.pydata.org/`

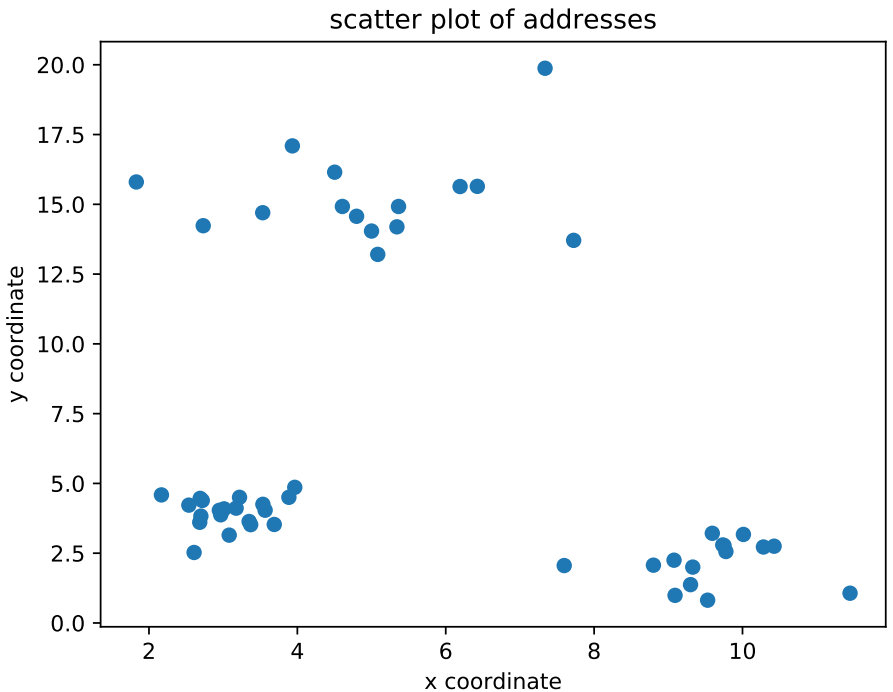
Hierarchical clustering



Hierarchical clustering



Hierarchical clustering



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

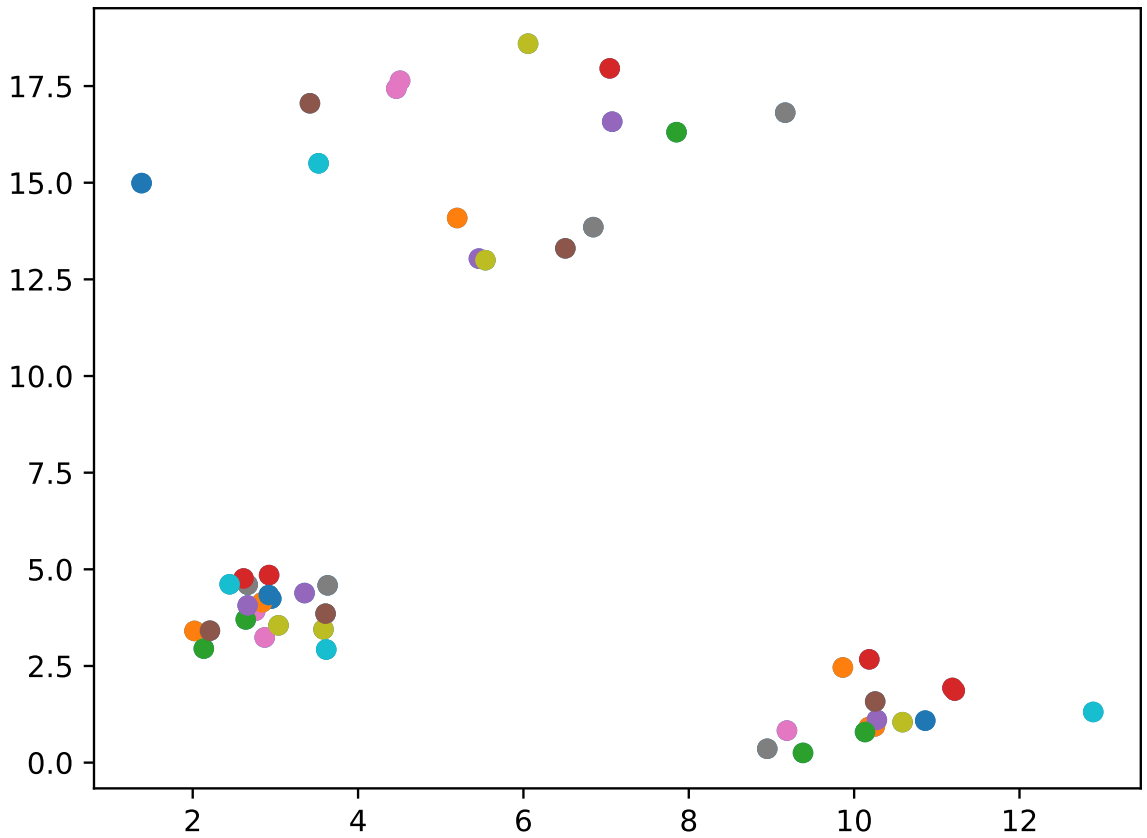
Hierarchical clustering

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

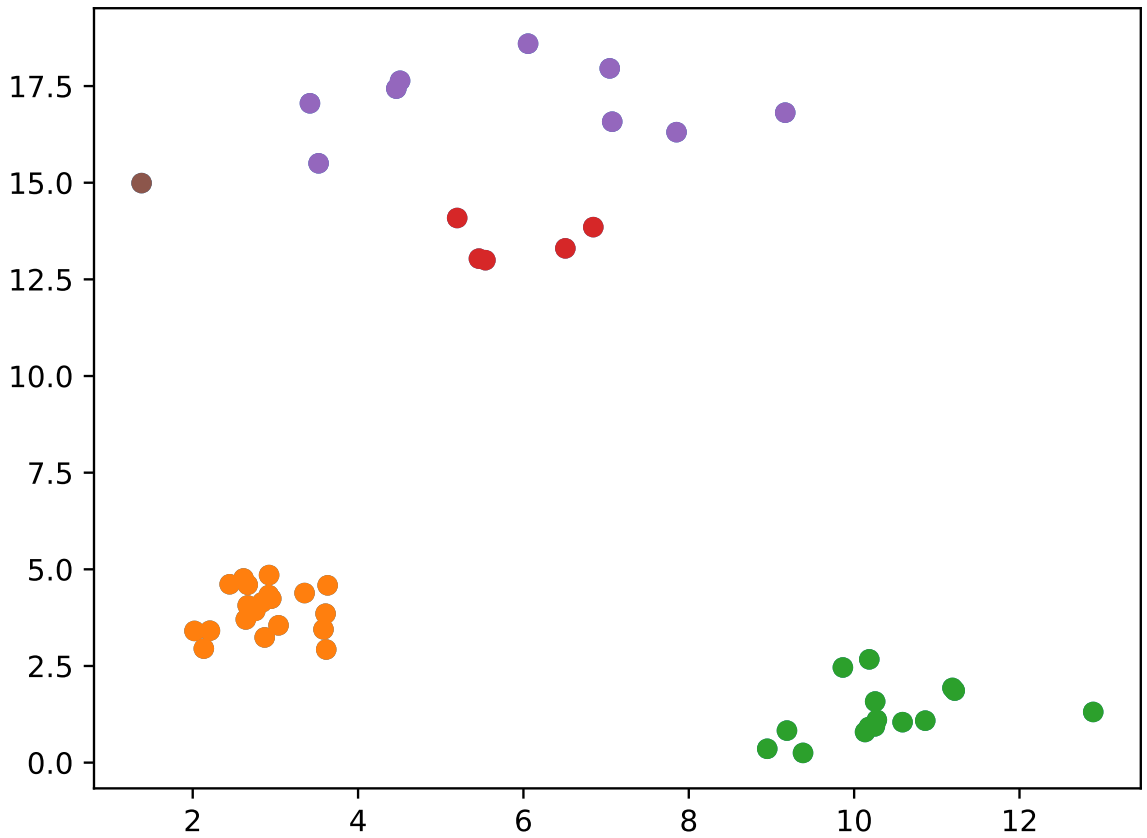
Hierarchical clustering

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.

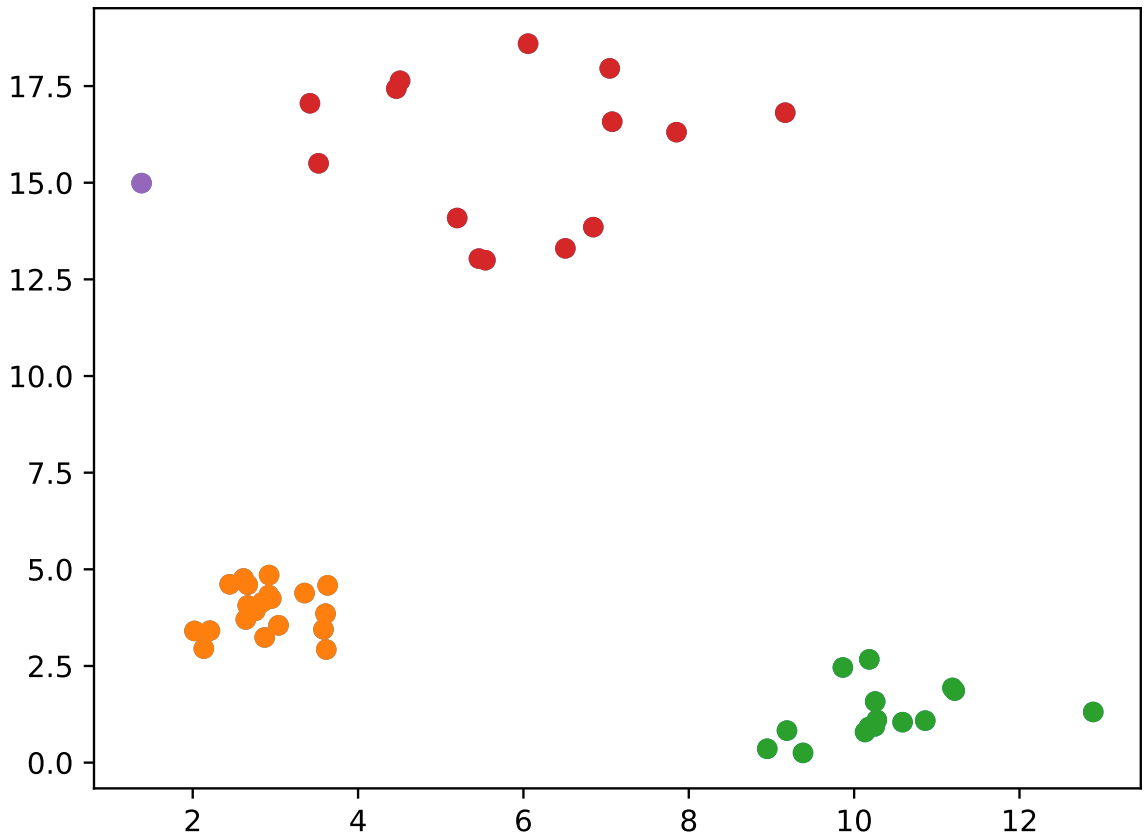
Hierarchical clustering



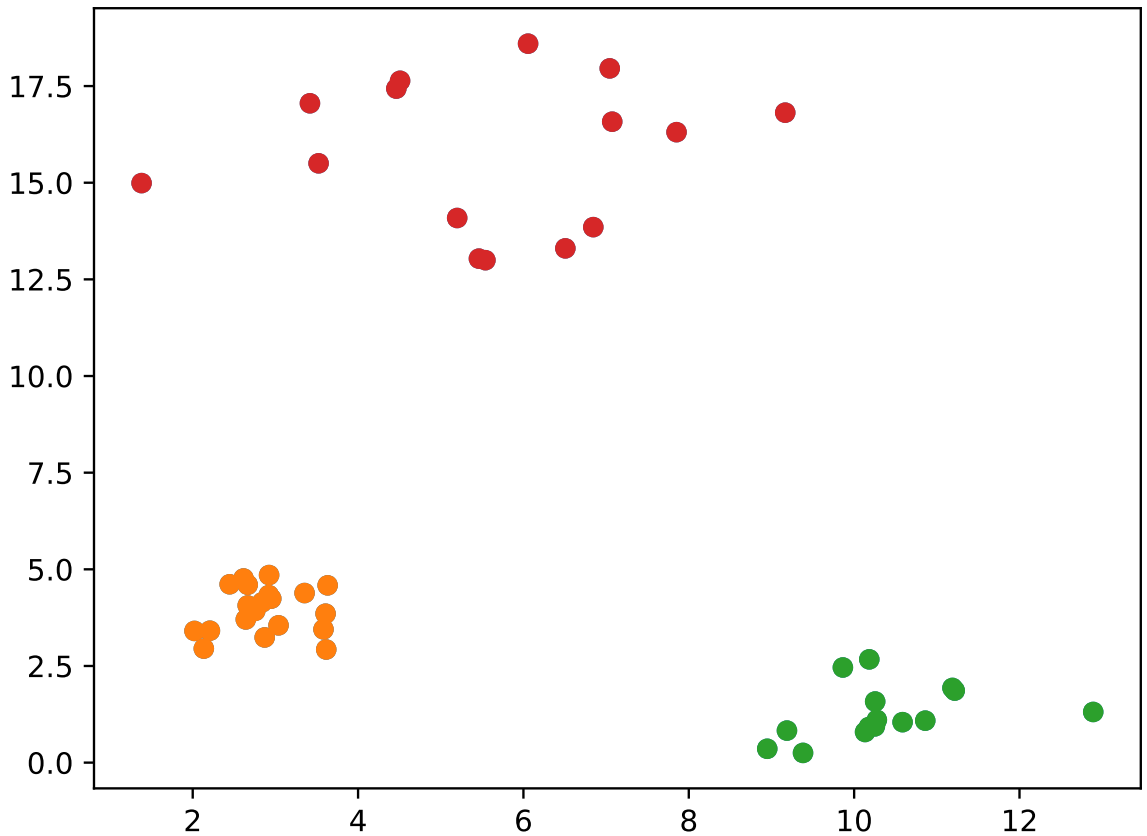
Hierarchical clustering



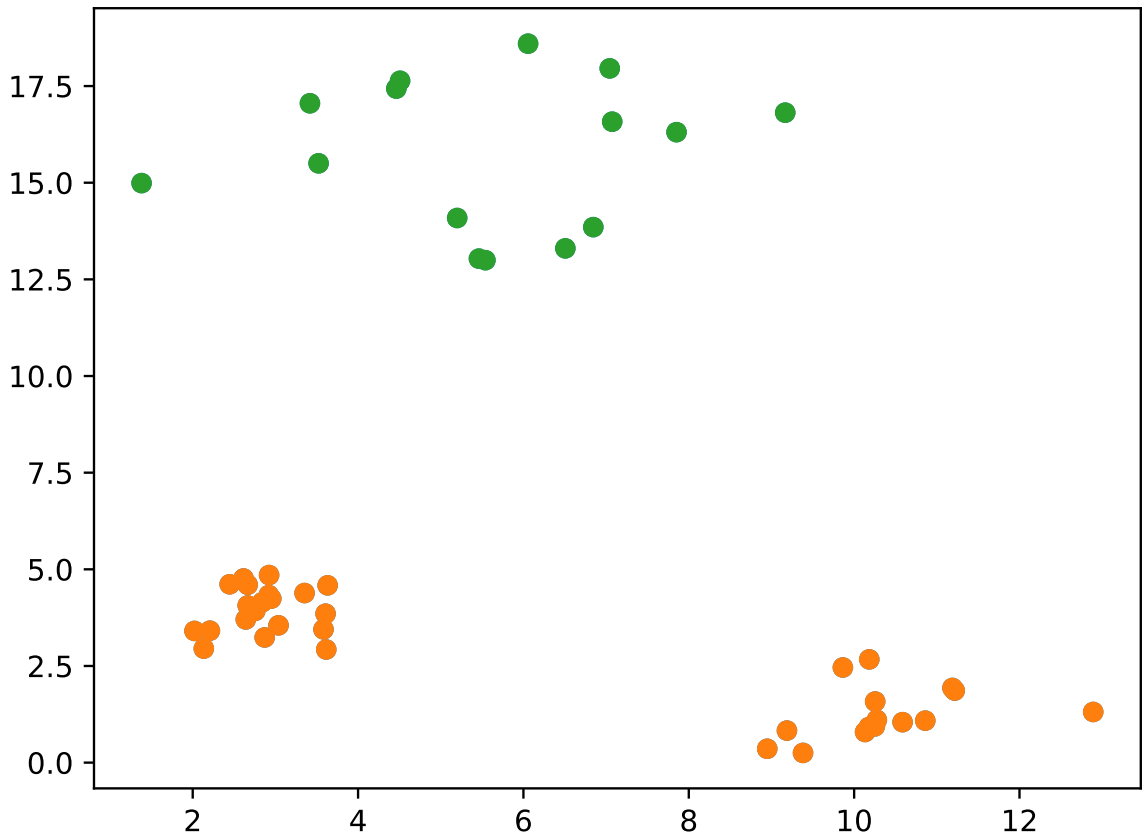
Hierarchical clustering



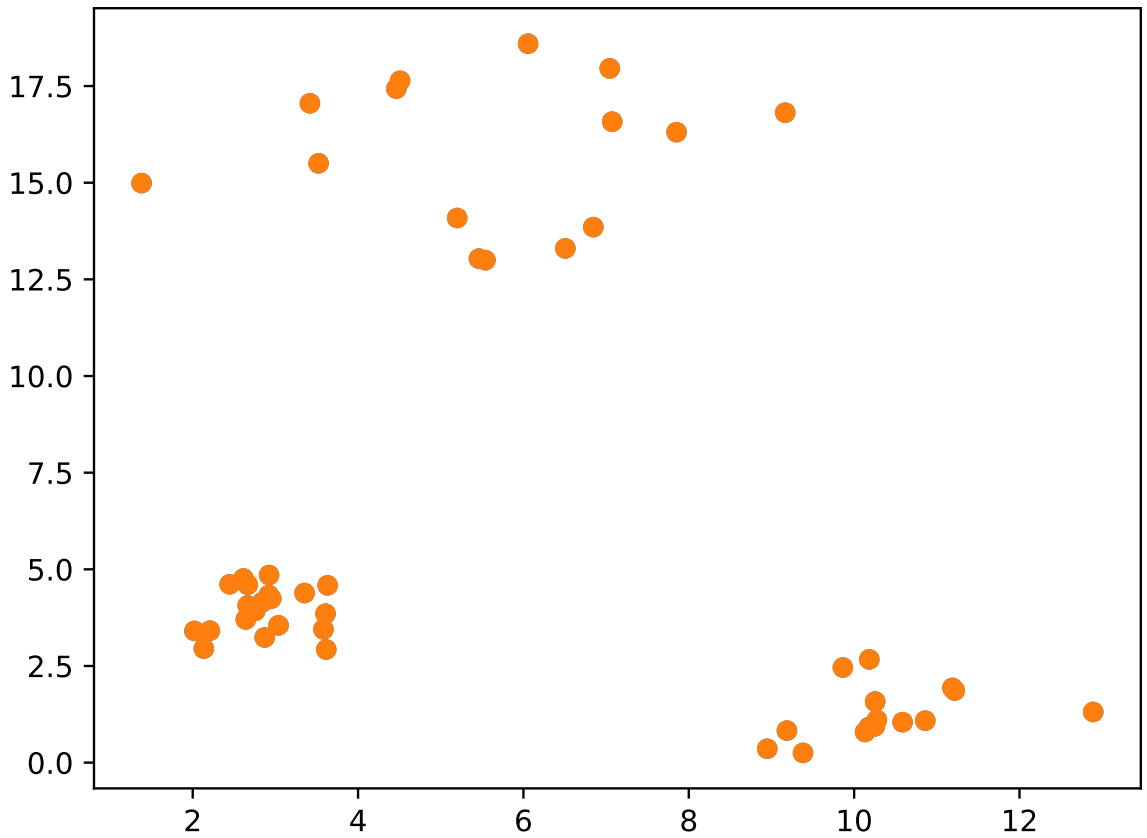
Hierarchical clustering



Hierarchical clustering



Hierarchical clustering



Hierarchical clustering

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

Hierarchical clustering

An **essential** aspect of hierarchical clustering is that different critera 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 critera 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.

Semi-supervised learning

- The paradigms can be mixed :

Semi-supervised learning

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

Application of the paradigms

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

Example 1 : web browser and page ranking

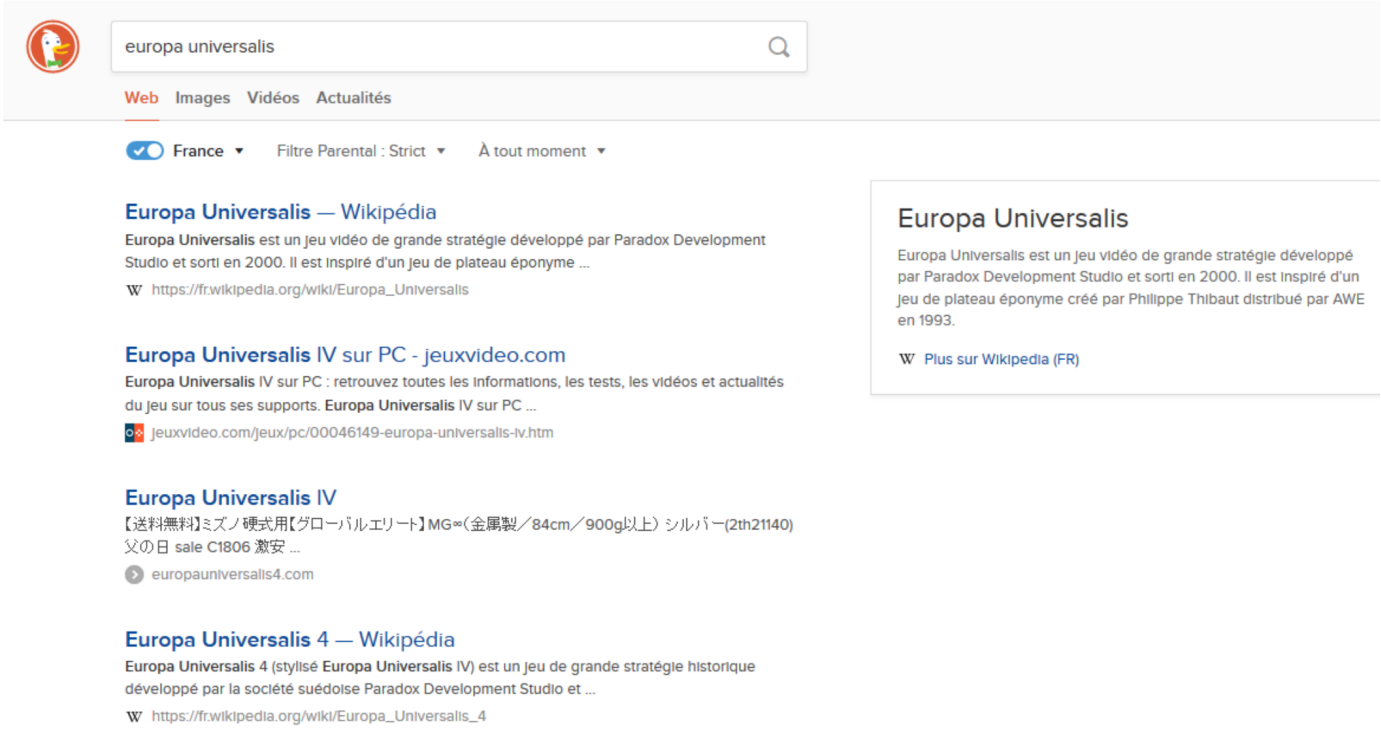


Figure: Duckduckgo web browser

Example 1 : web browser and page ranking

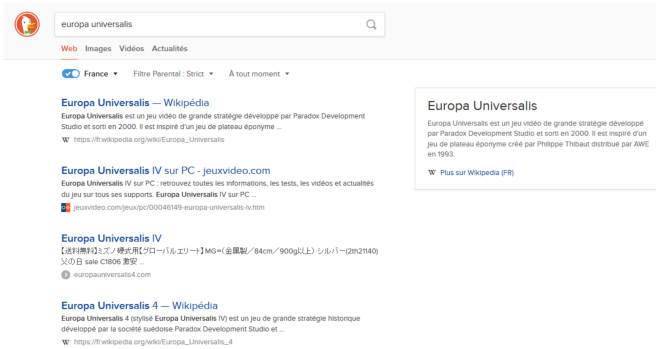


Figure: Duckduckgo web browser

- 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

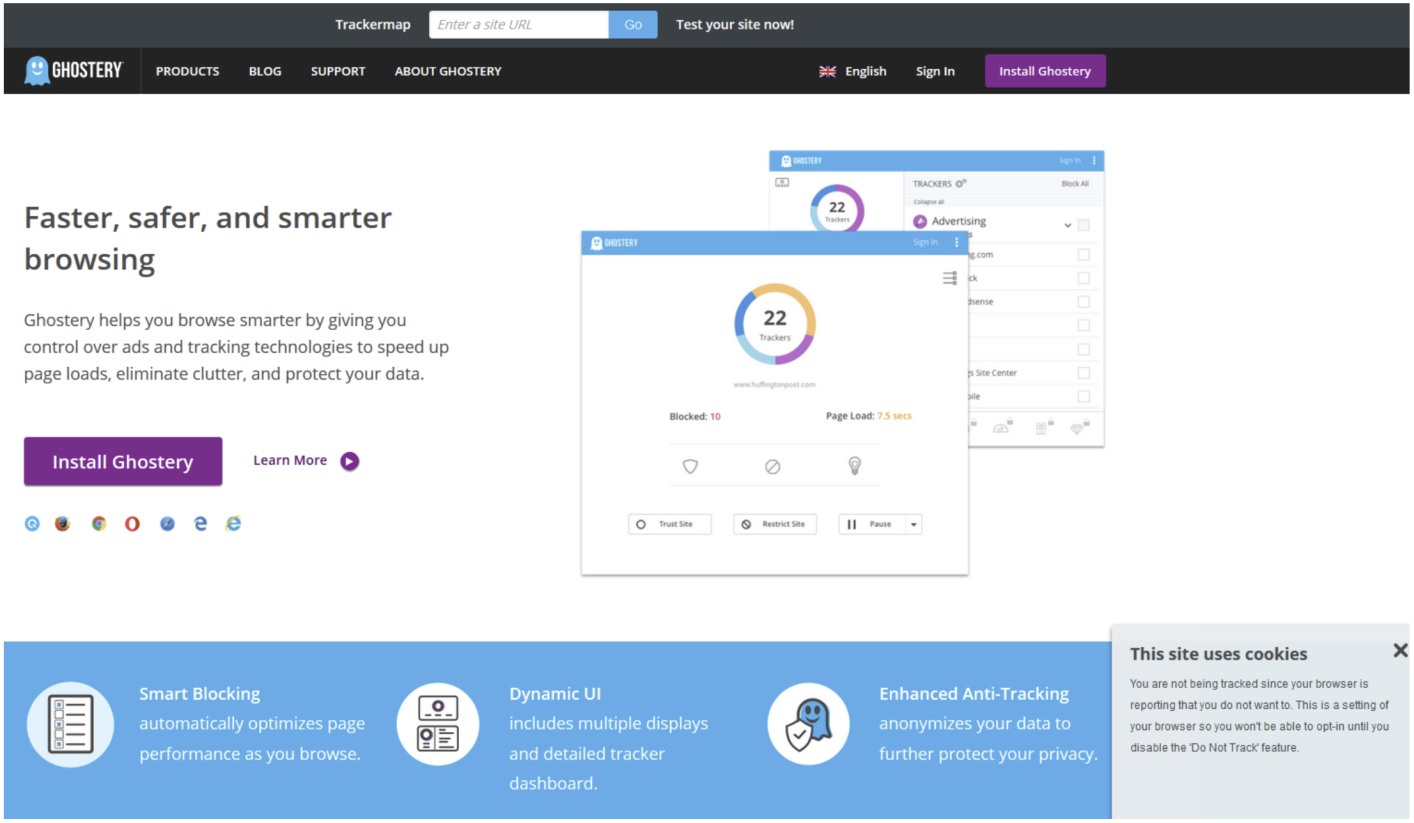


Figure: Ghostery

Example 2 : targetted ads

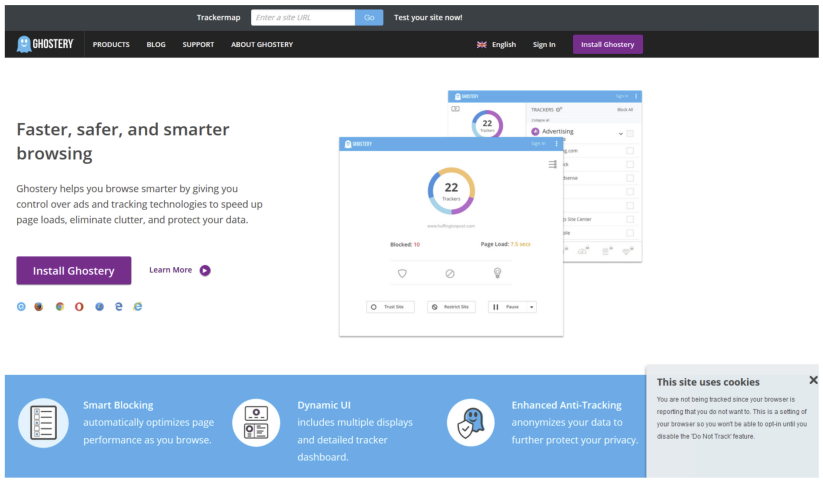


Figure: Ghostery

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

Example 4 : Image segmentation

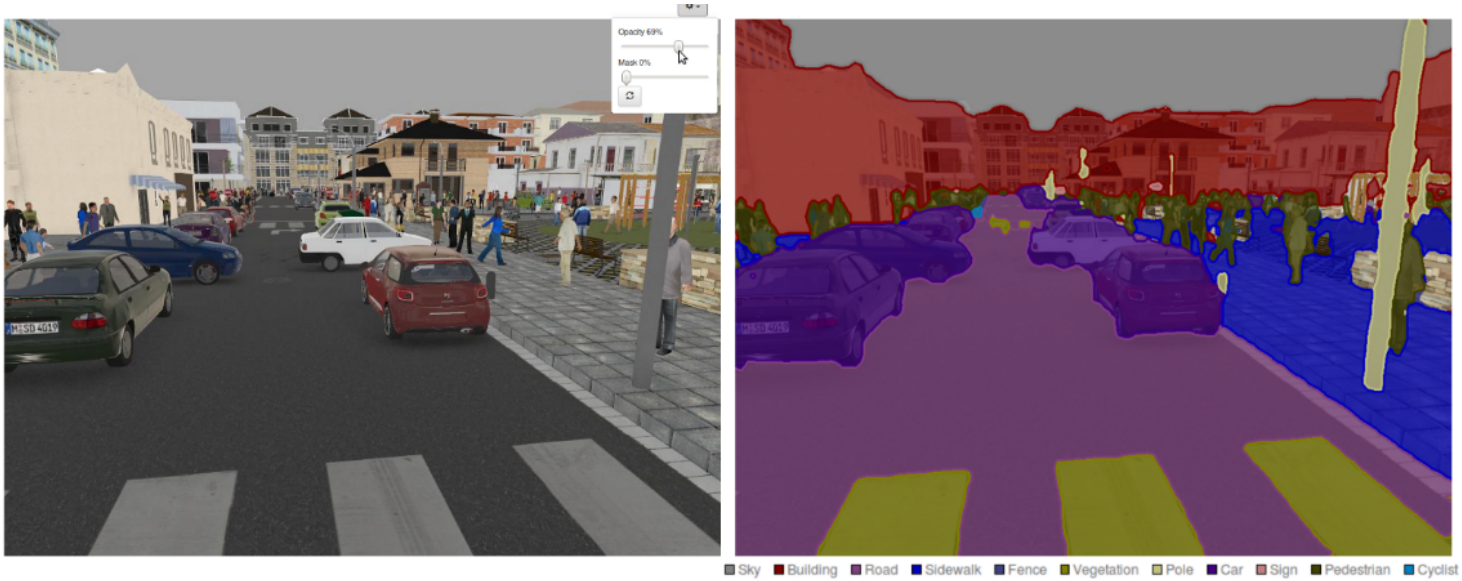


Figure: Source : Nvidia dev blog

Example 4 : Image segmentation

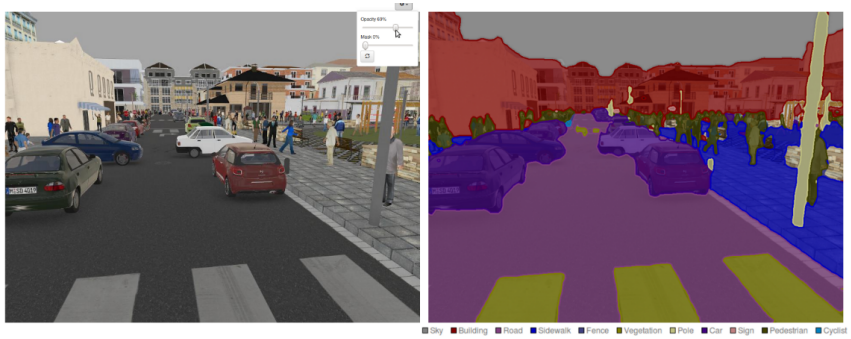


Figure: Source : Nvidia dev blog

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

Example 4 : Image segmentation

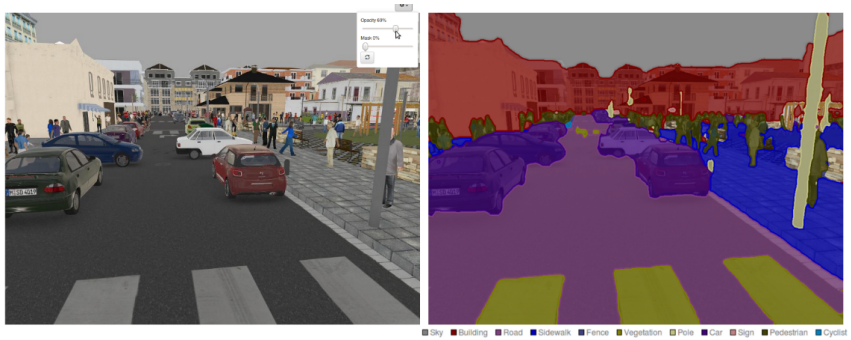


Figure: Source : Nvidia dev blog

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

Example 5 : Digit classification

Figure: MNIST dataset [LeCun and Cortes, 2010]

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

FIGURE : topics

FIGURE : document

Figure: Statistics on texts

Example 6 : Lexicometry

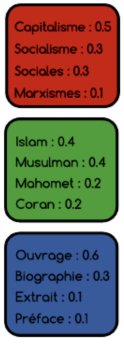


FIGURE : topics

FIGURE : document

Figure: Statistics on texts

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

Example 3 : Text Translation

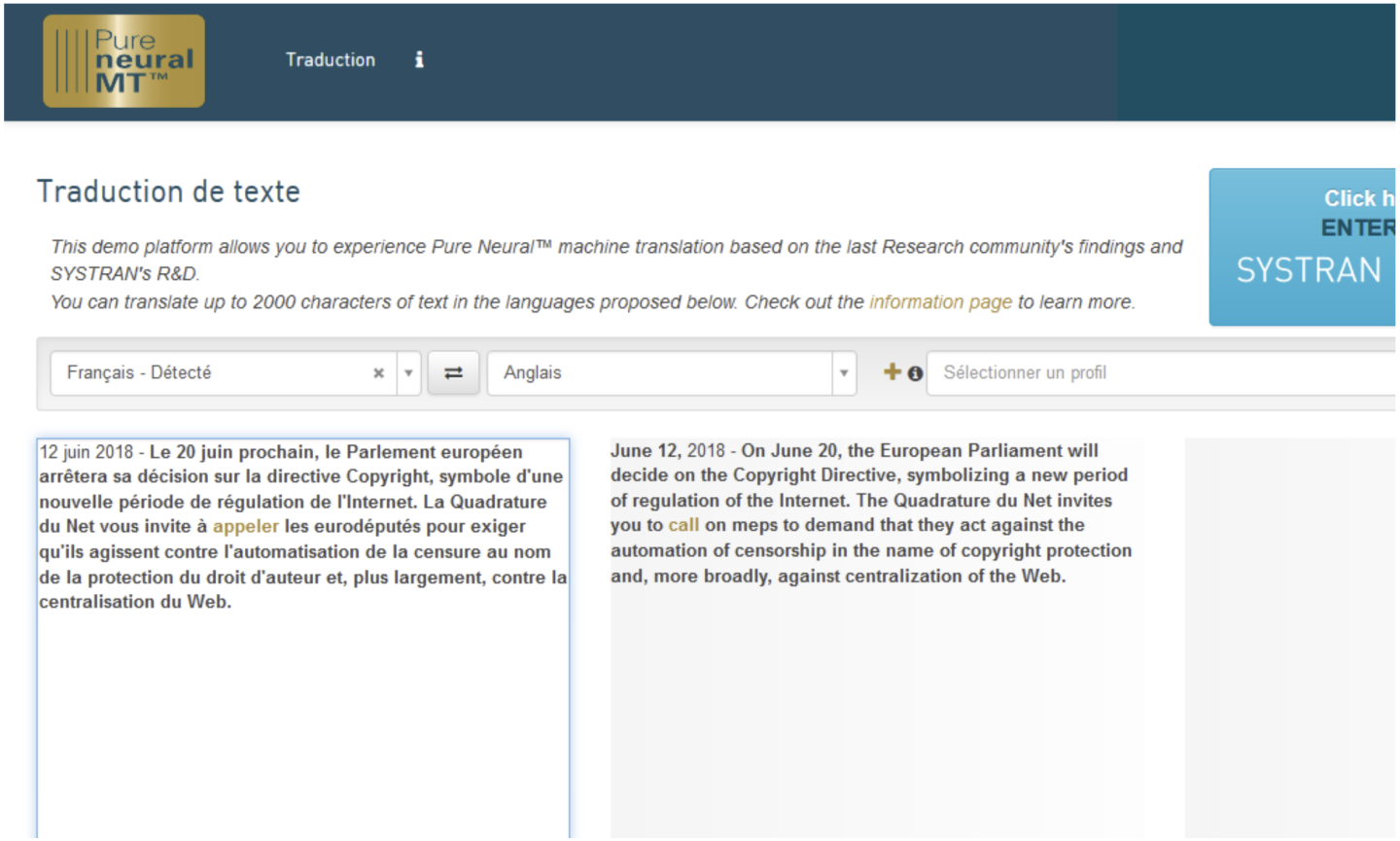


Figure: Text translation

Example 3 : Text Translation

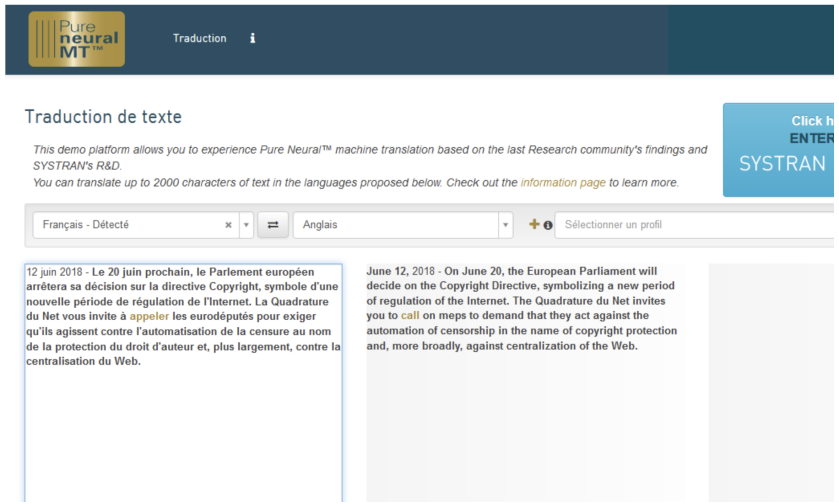


Figure: Text translation

- Based on translated examples
- **Supervised**

Example 7 : Time Series Analysis

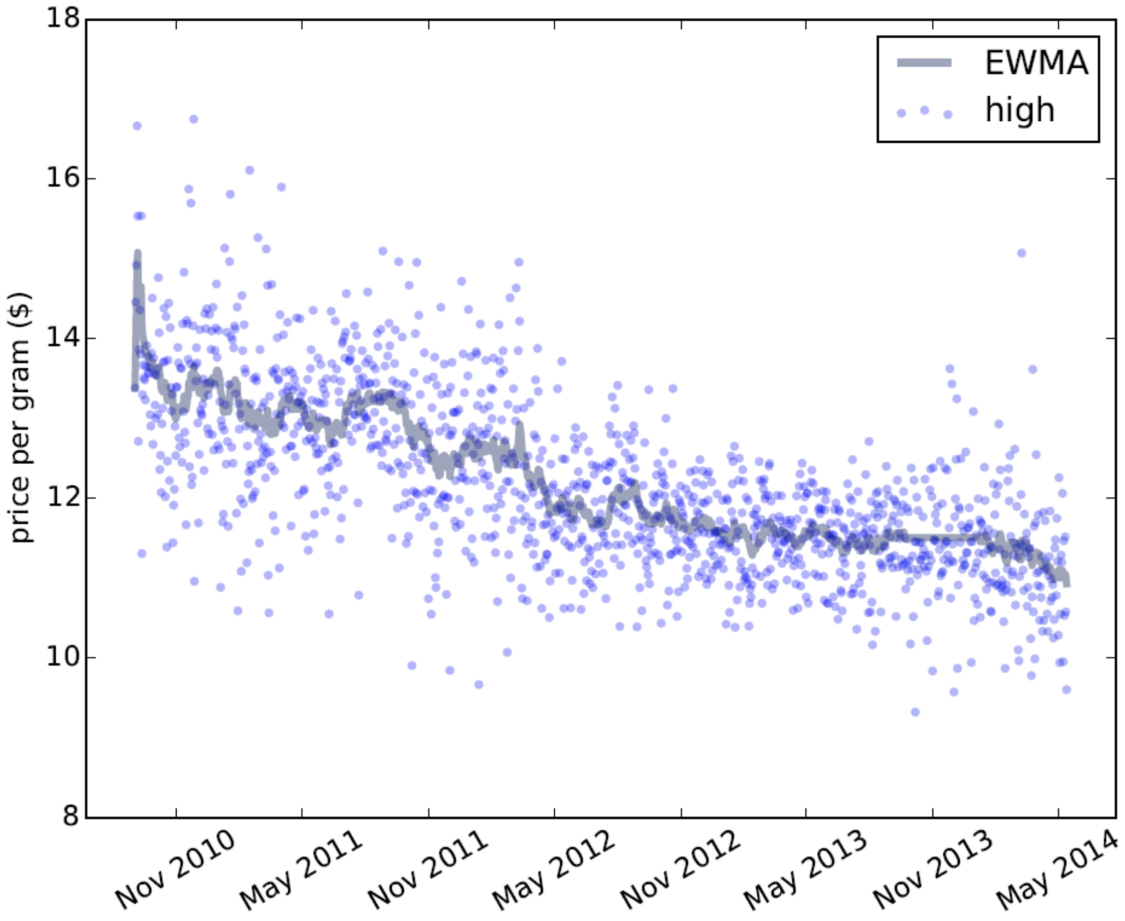


Figure: Price of some plant

Example 7 : Time Series Analysis

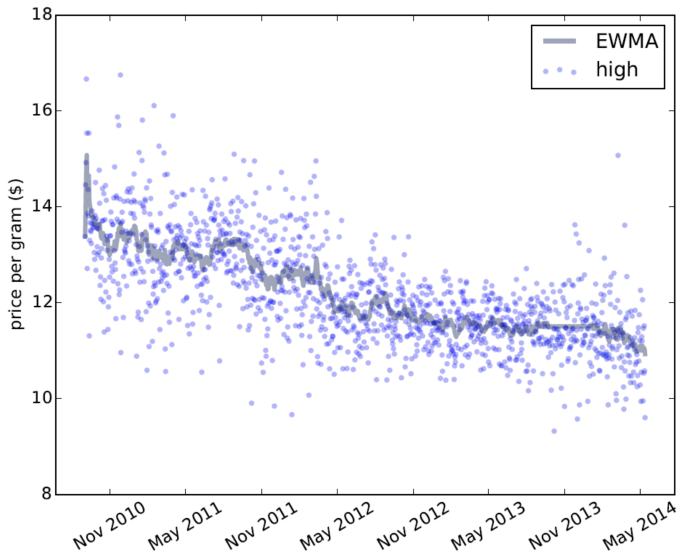


Figure: Price of some plant

- **Supervised**

Example 8 : Clustering

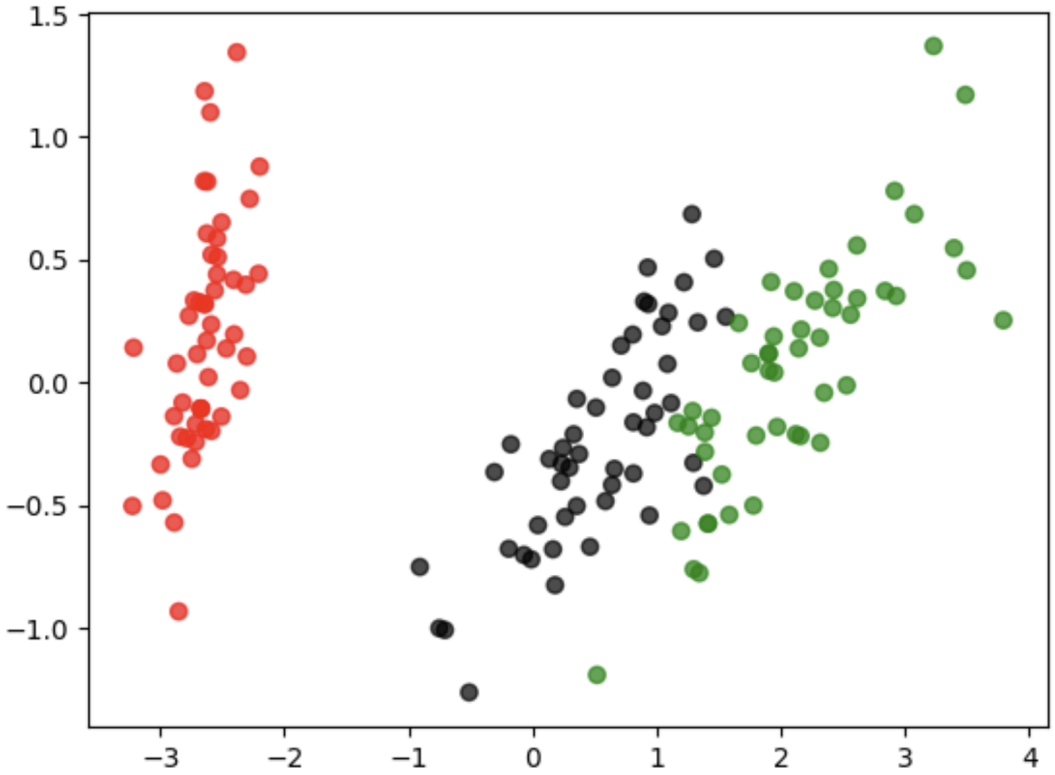


Figure: Iris dataset

Example 8 : Clustering

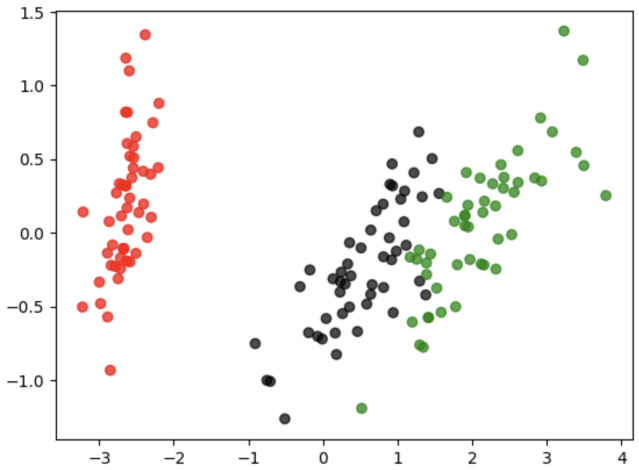


Figure: Iris dataset

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

Example 9 : game AI

Figure: Europa Universalis

Example 9 : game AI

Figure: Europa Universalis

- Desining an **agent** performing **actions**
- **Reinforcement learning**

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**.

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 :

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.

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.

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

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**

Continuous variables

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

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 probably of being between x and $x + dx$ is $p(x)dx$.

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 probably of being between x and $x + dx$ is $p(x)dx$.
- Let's see some examples

Uniform discrete

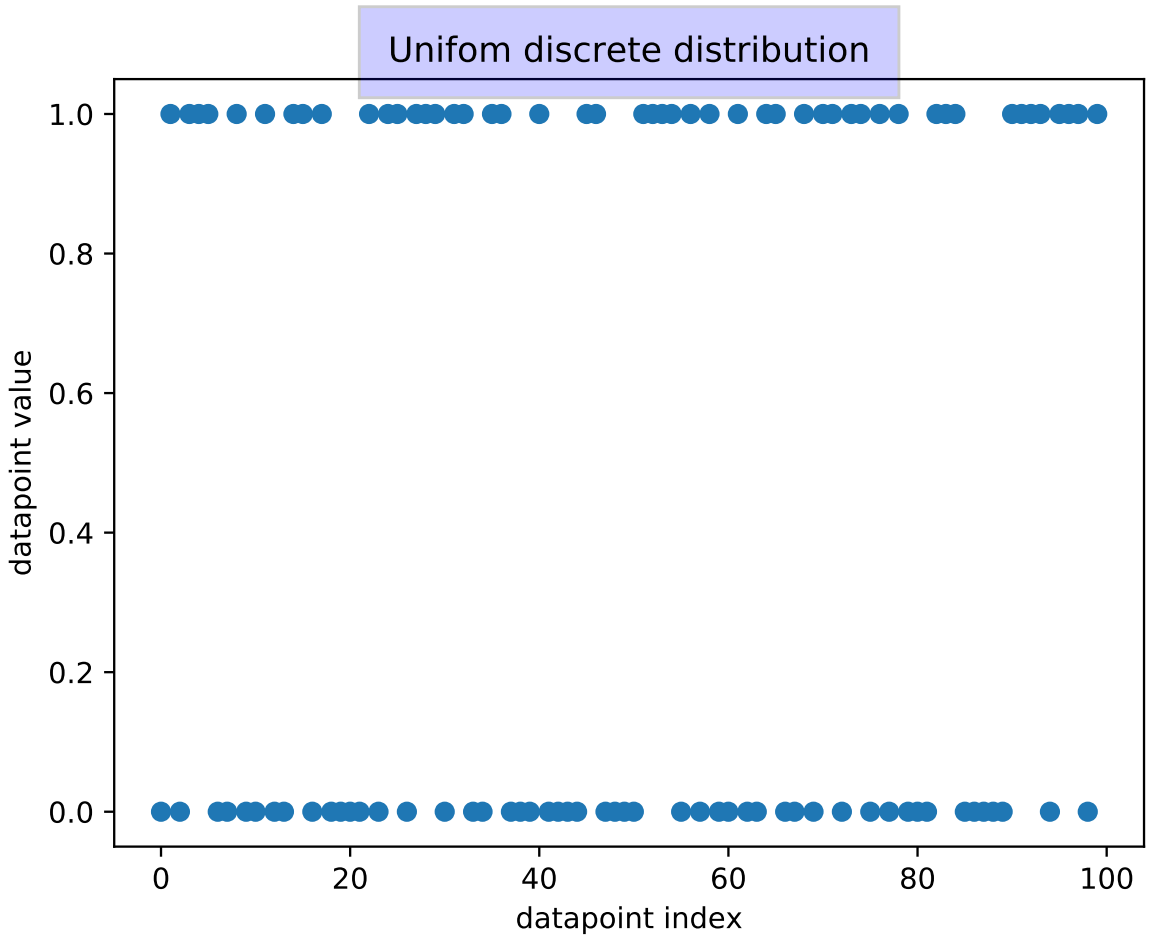


Figure: Uniform discrete distribution

Bernoulli

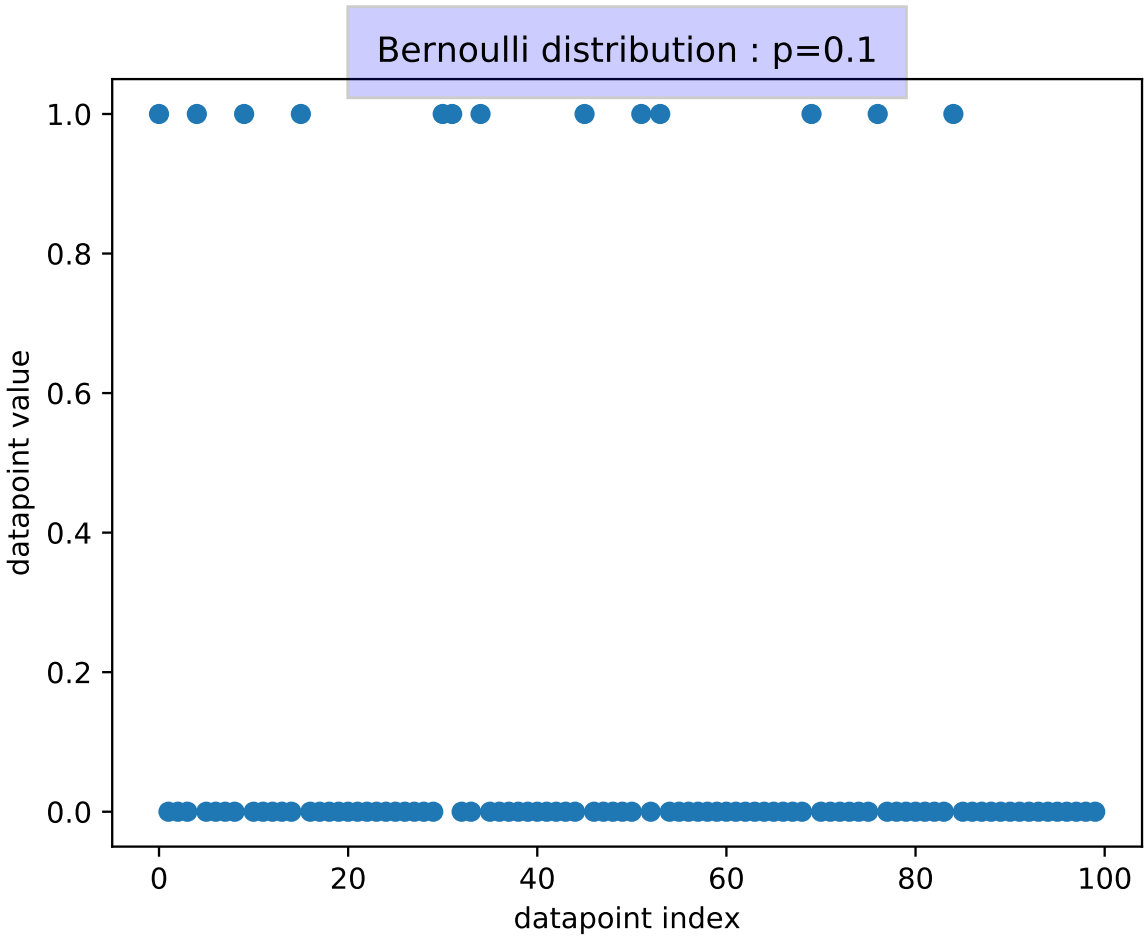


Figure: Bernoulli distribution

Bernoulli

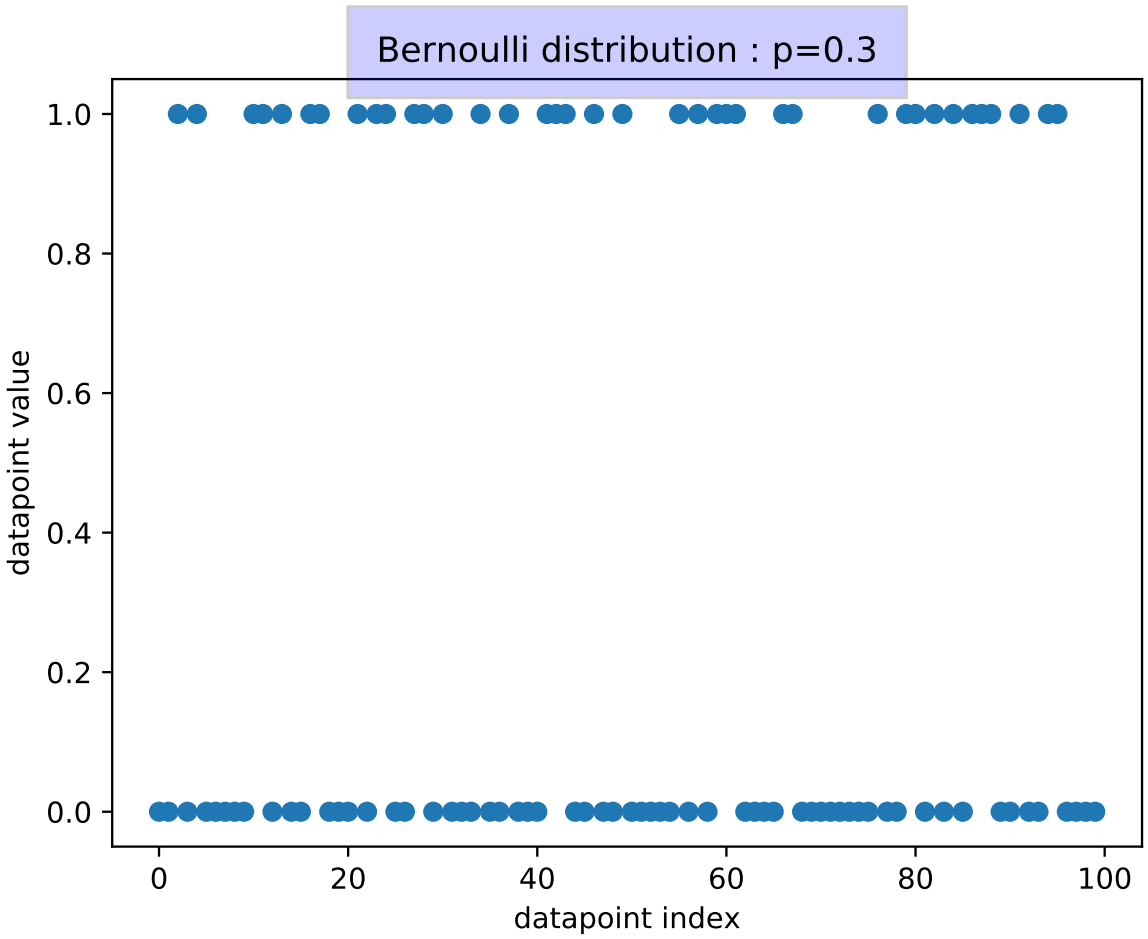


Figure: Bernoulli Distribution

Bernoulli

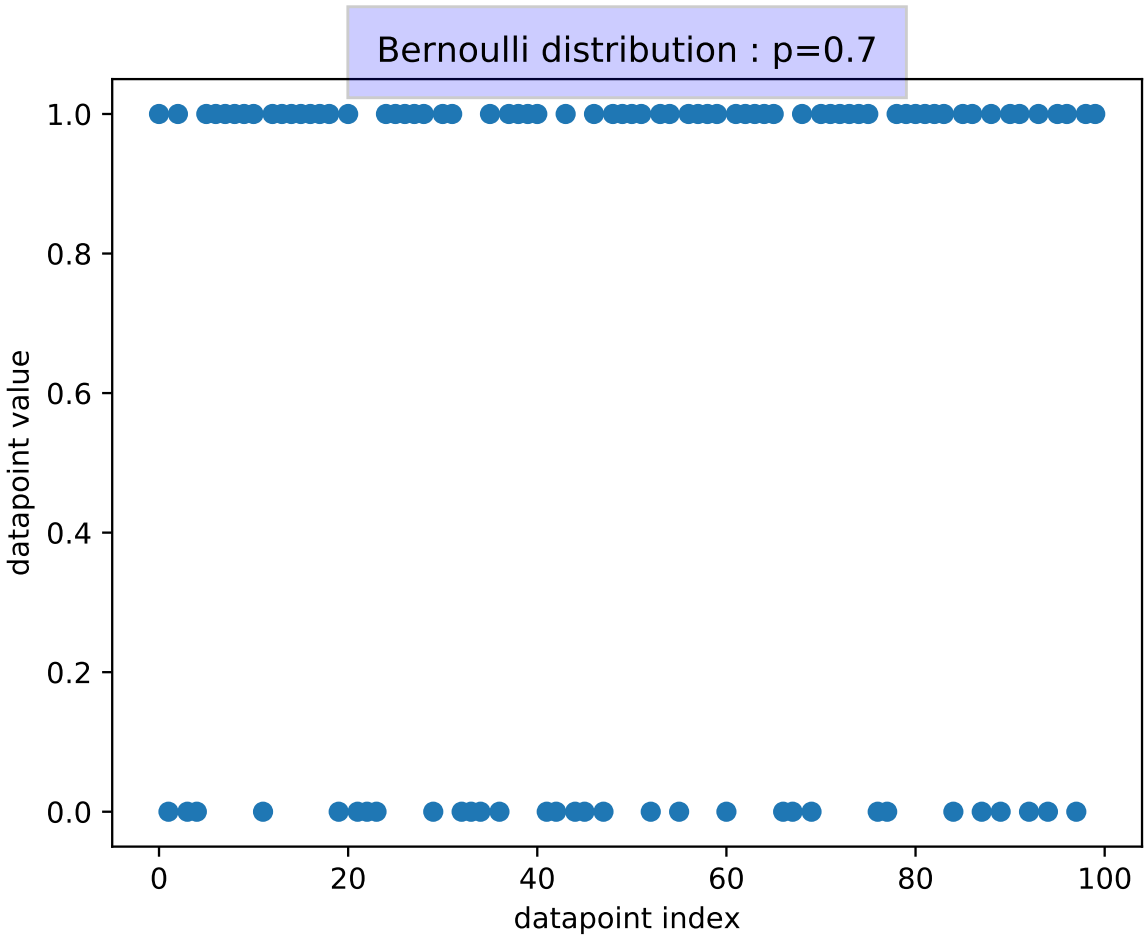


Figure: Bernoulli Distribution

Uniform continuous

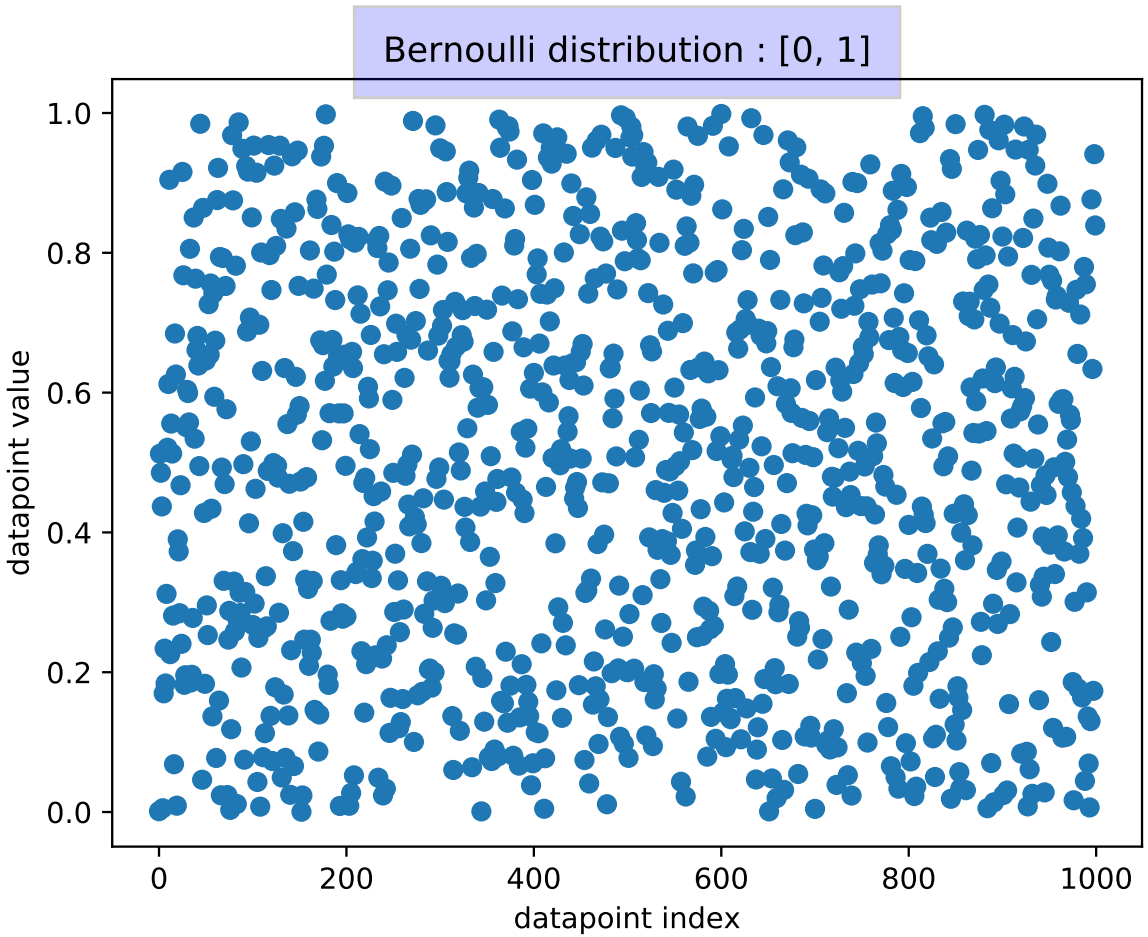


Figure: Uniform continuous distribution

Uniform continuous

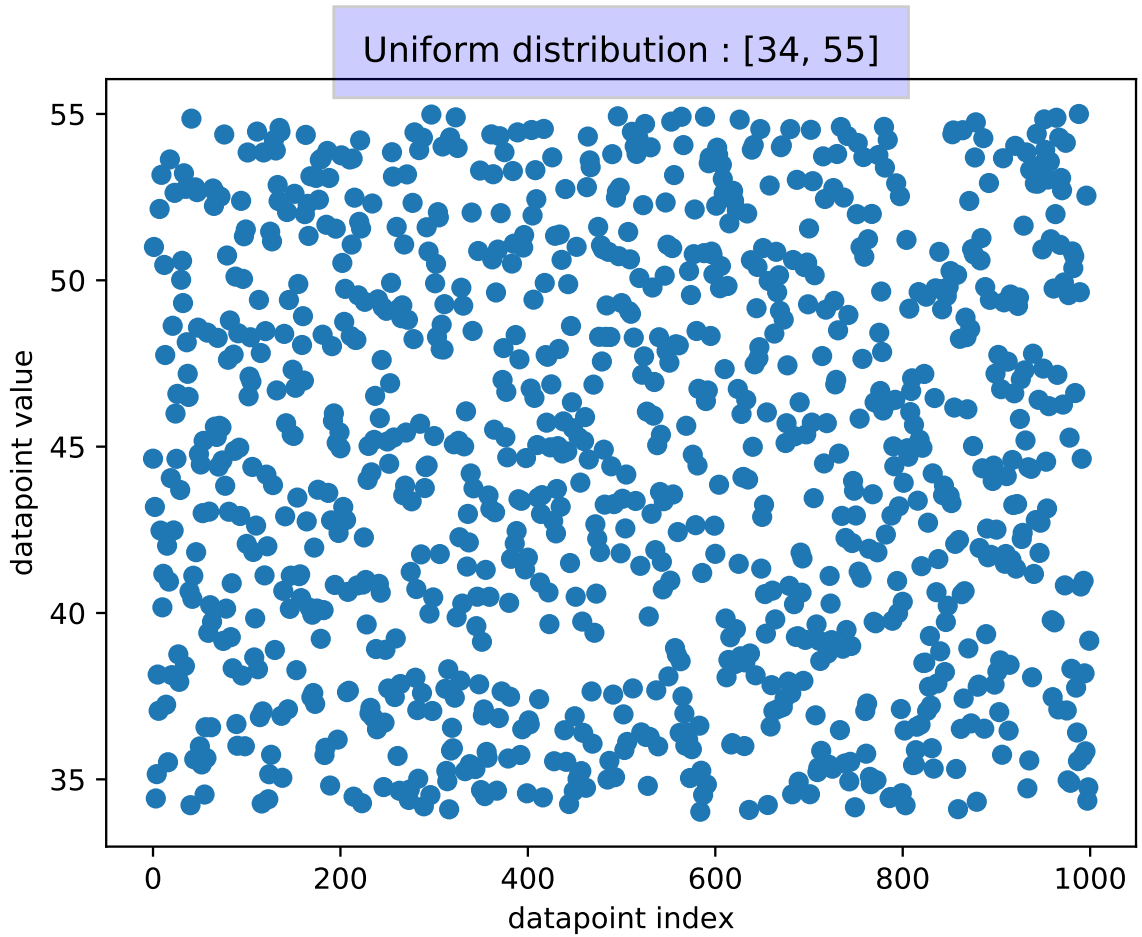


Figure: Uniform continuous distribution

Normal

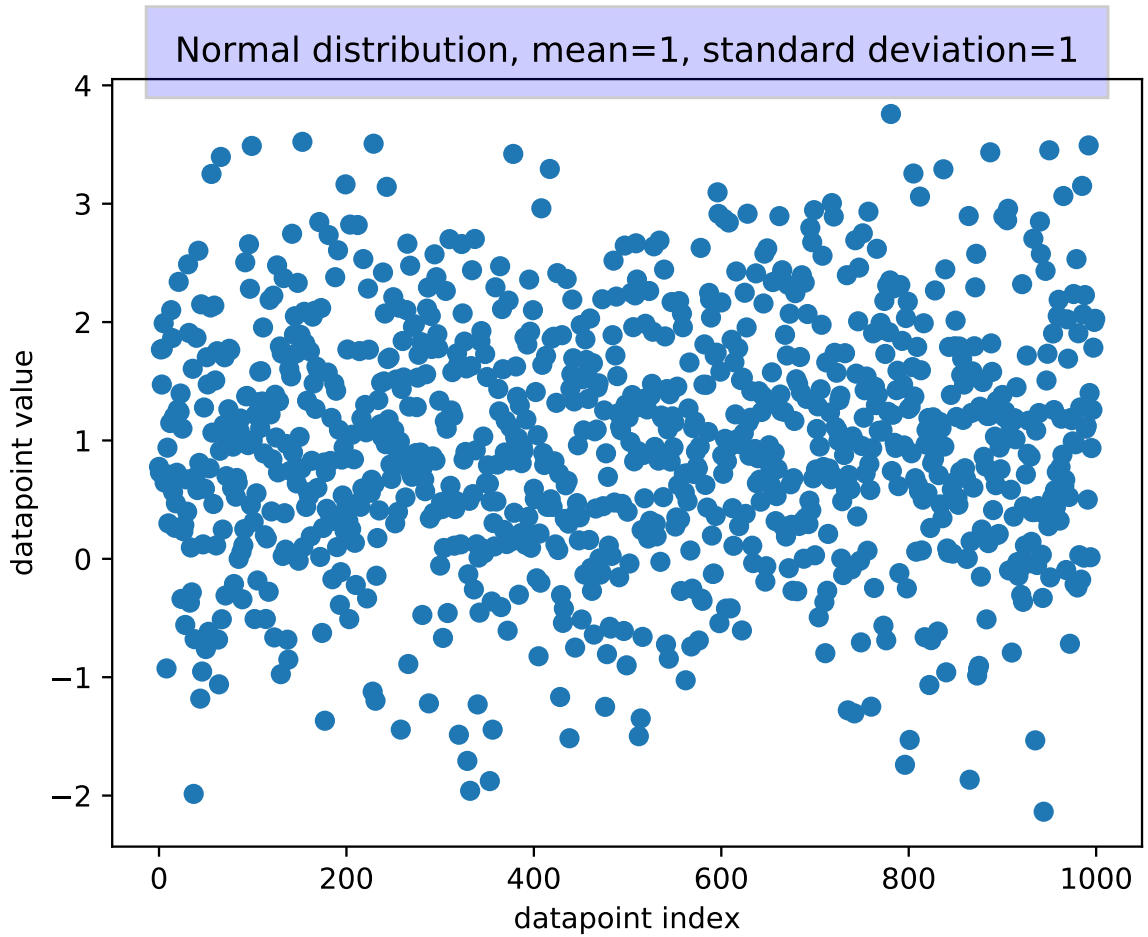


Figure: Normal distribution

Normal

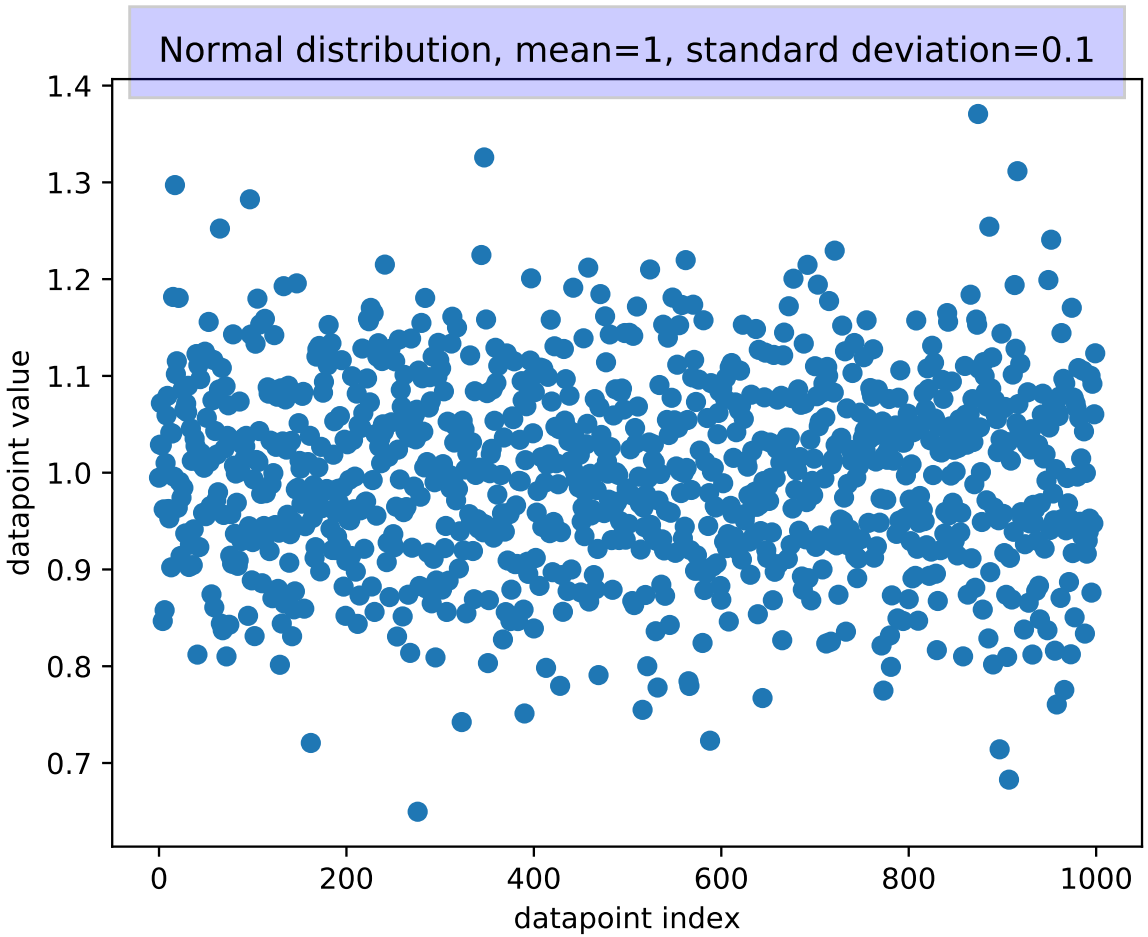


Figure: Normal distribution

Normal

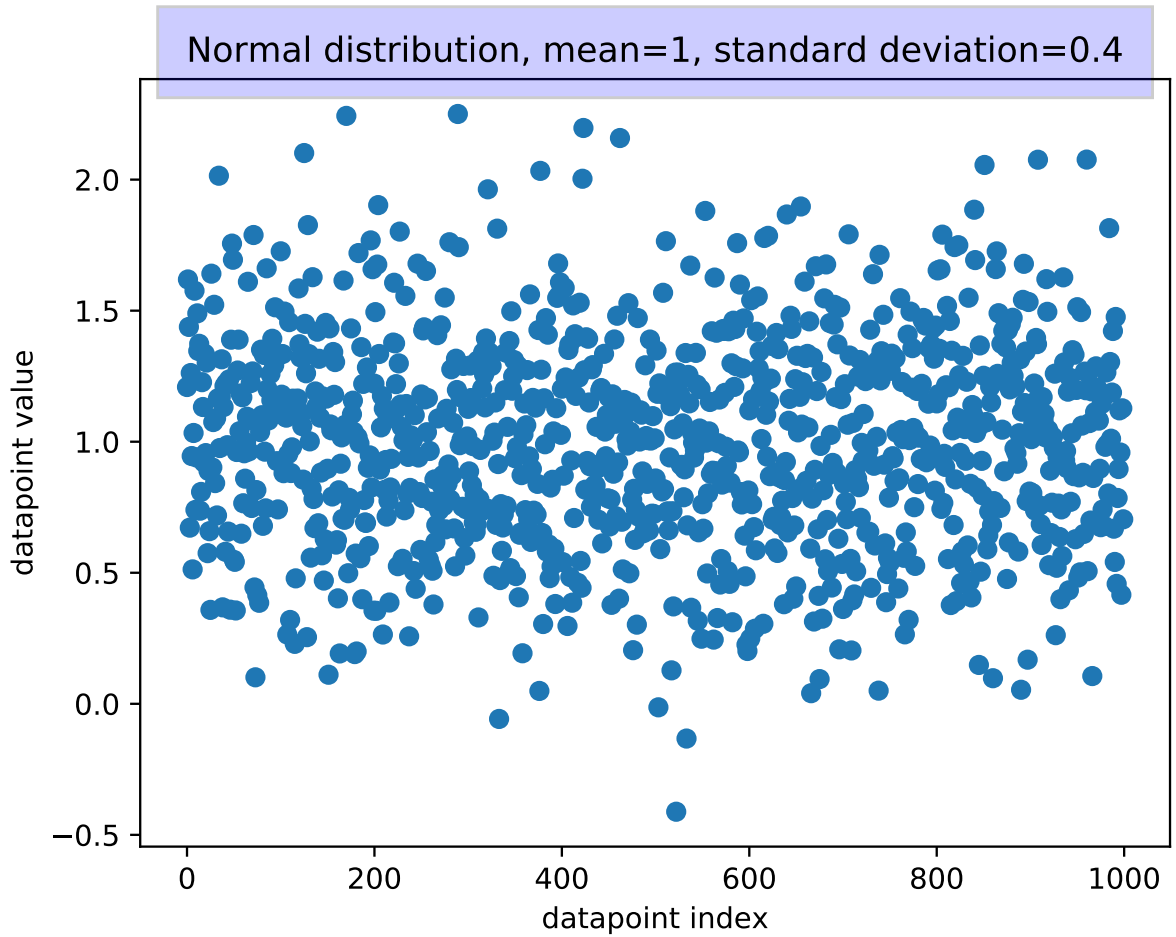


Figure: Normal distribution

White noise

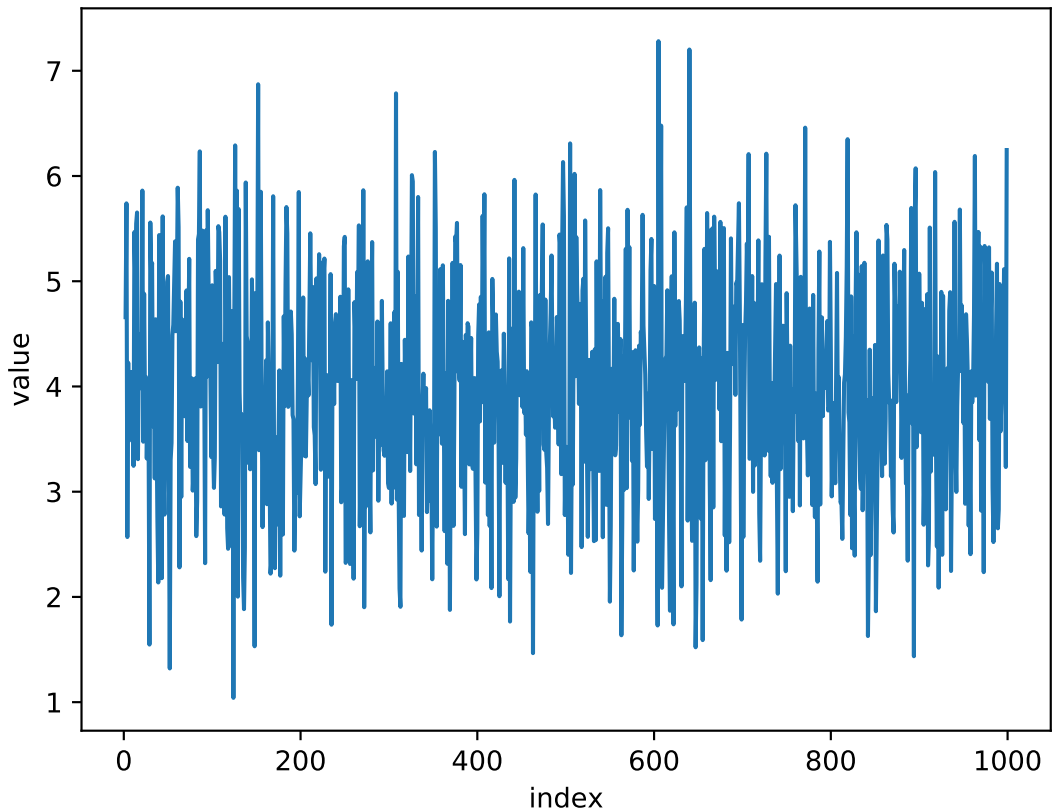


Figure: White noise

Histograms

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

Histograms

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

Uniform discrete

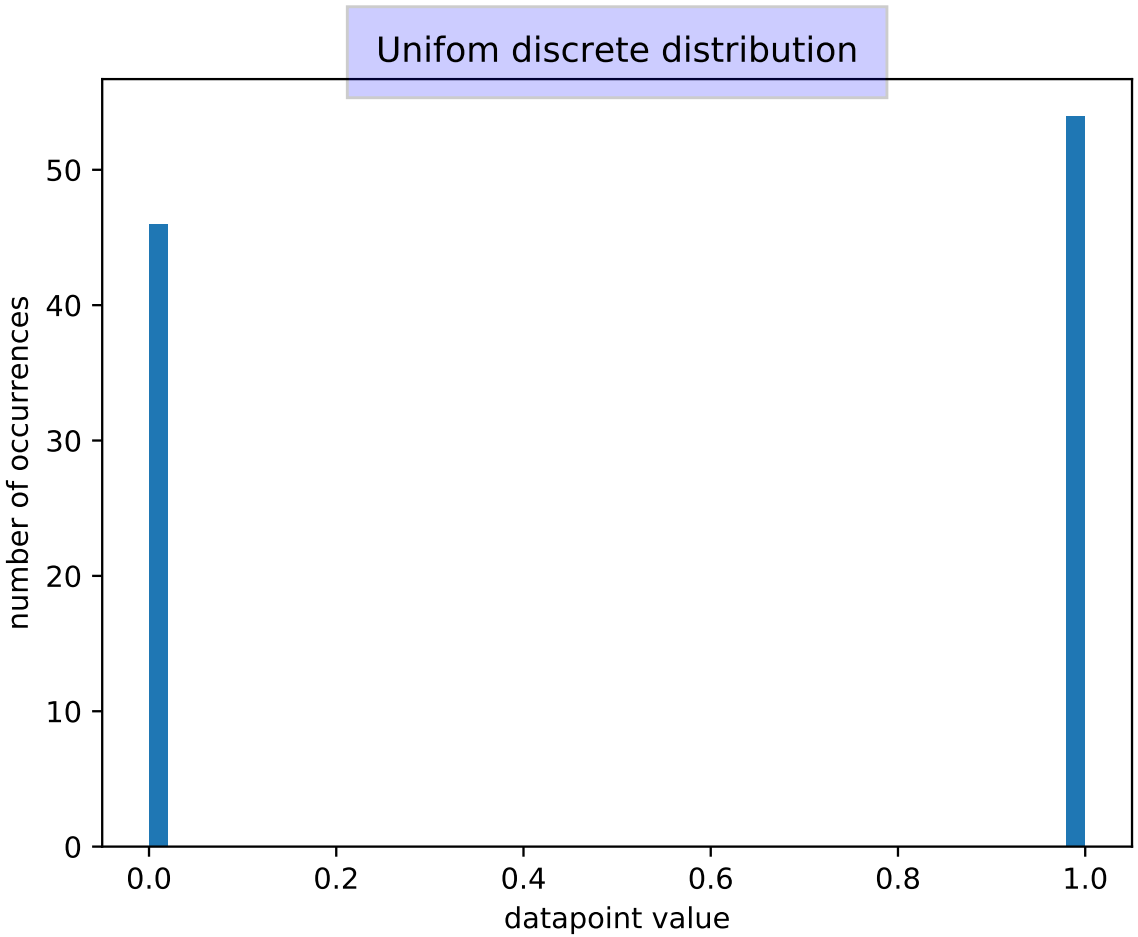


Figure: Historgram 1

Bernoulli

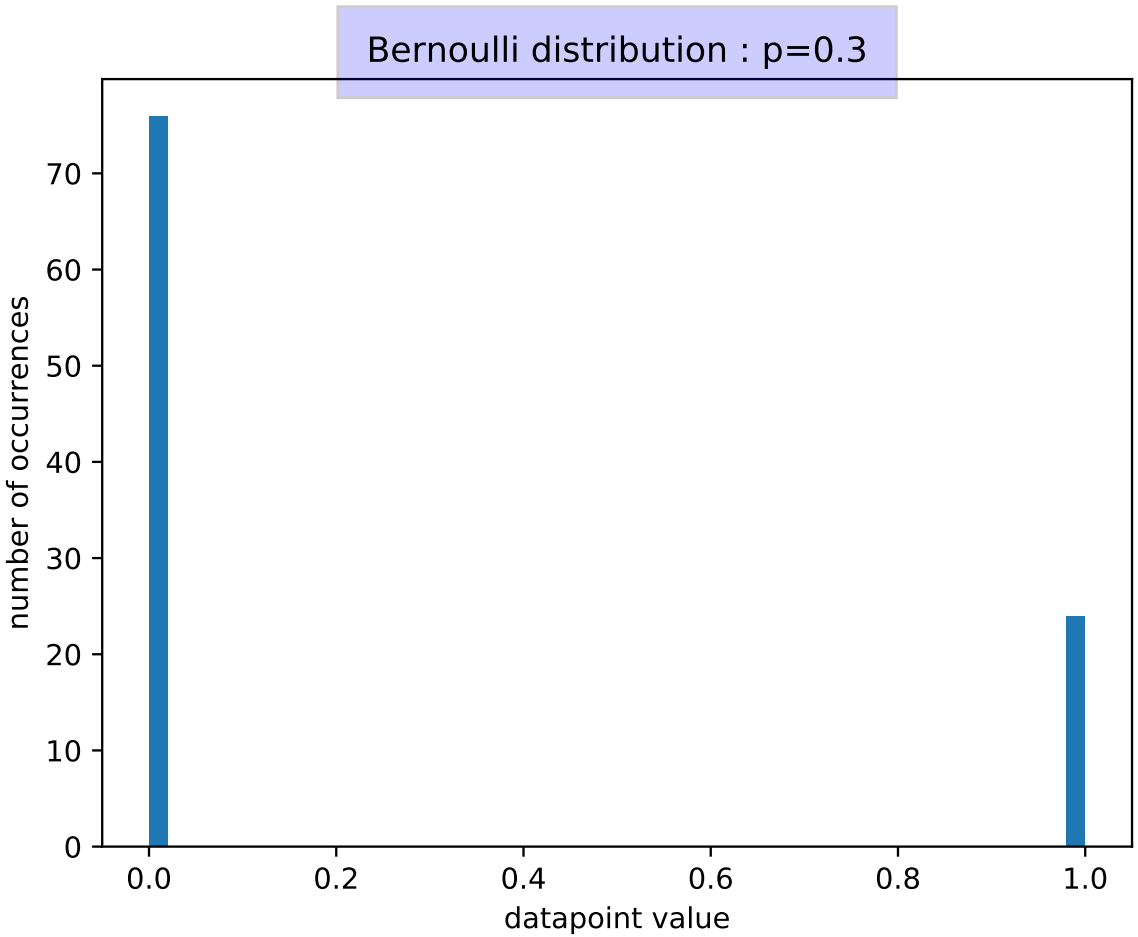


Figure: Historgram 2

Uniform continuous

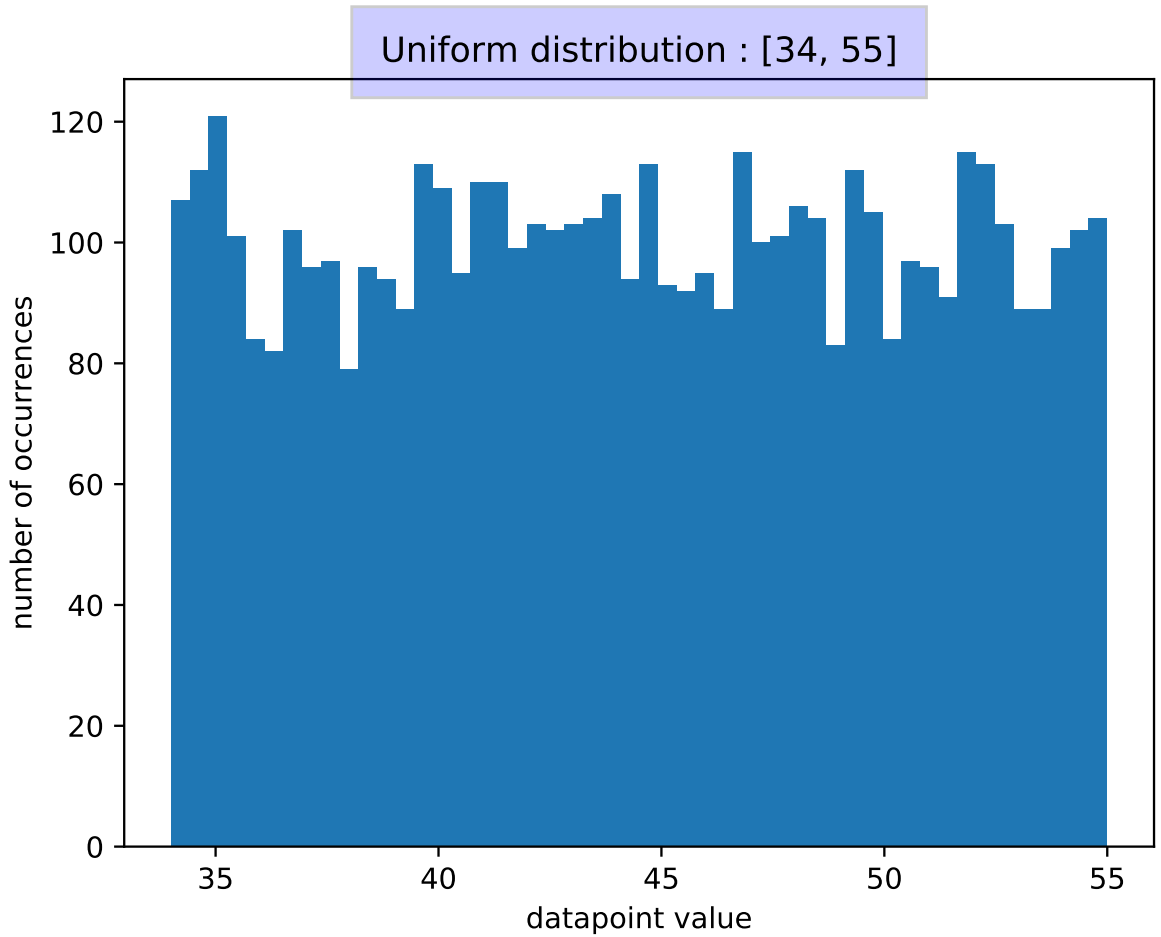


Figure: Historgram 3

Normal

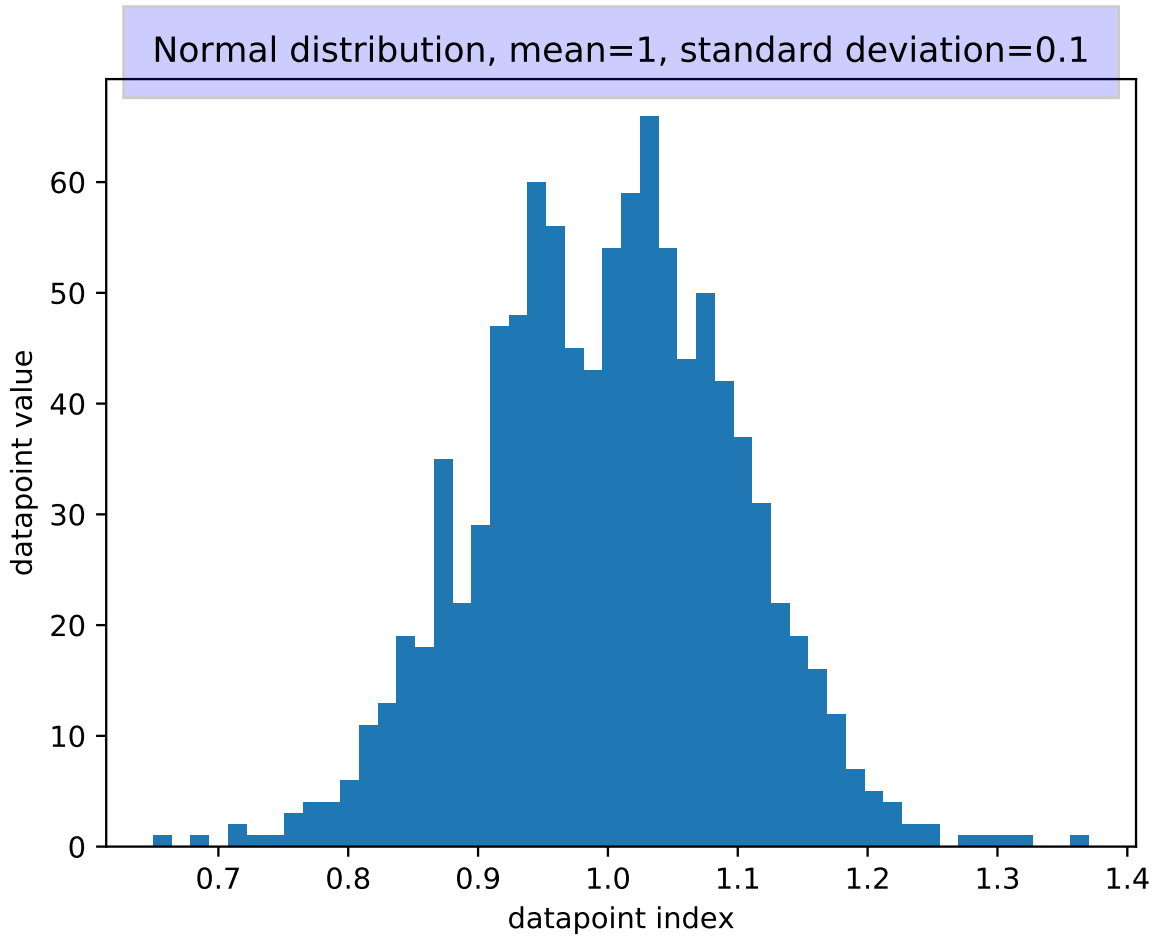


Figure: Historgram 4

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.

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

Evaluation

How do we evaluate the quality of a model ?

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

Overfitting

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

Overfitting

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

Overfitting

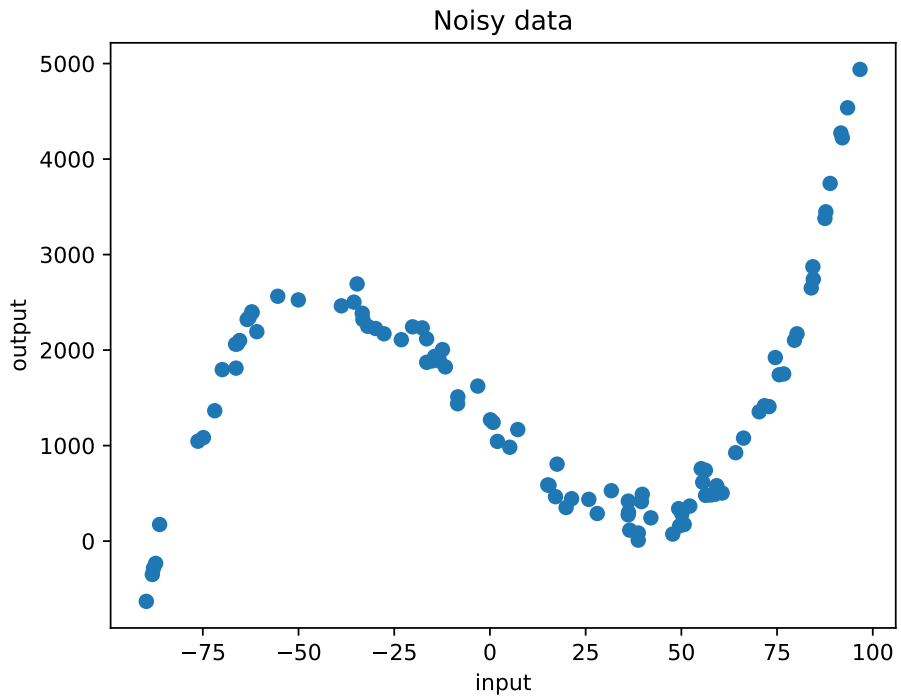
- What could be the drawbacks of using a very simple model (very few parameters)
  - Weak expressive power
- What could be the drawbacks of having a very complex model (that contains a very large number of parameters, e.g. millions as in a very deep neural network) ie a very high expressive power ?

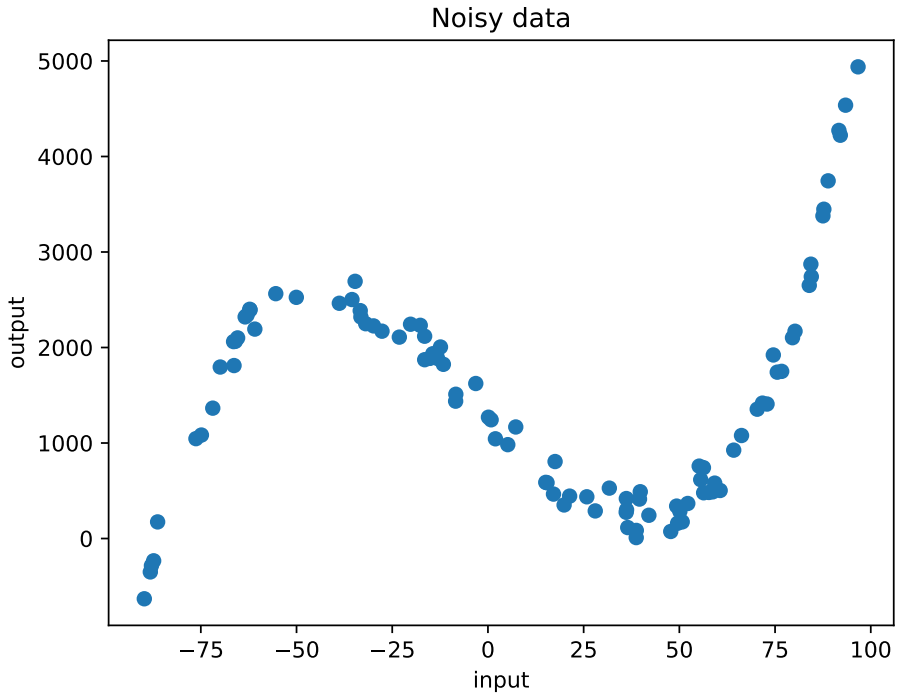
Overfitting

- 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**

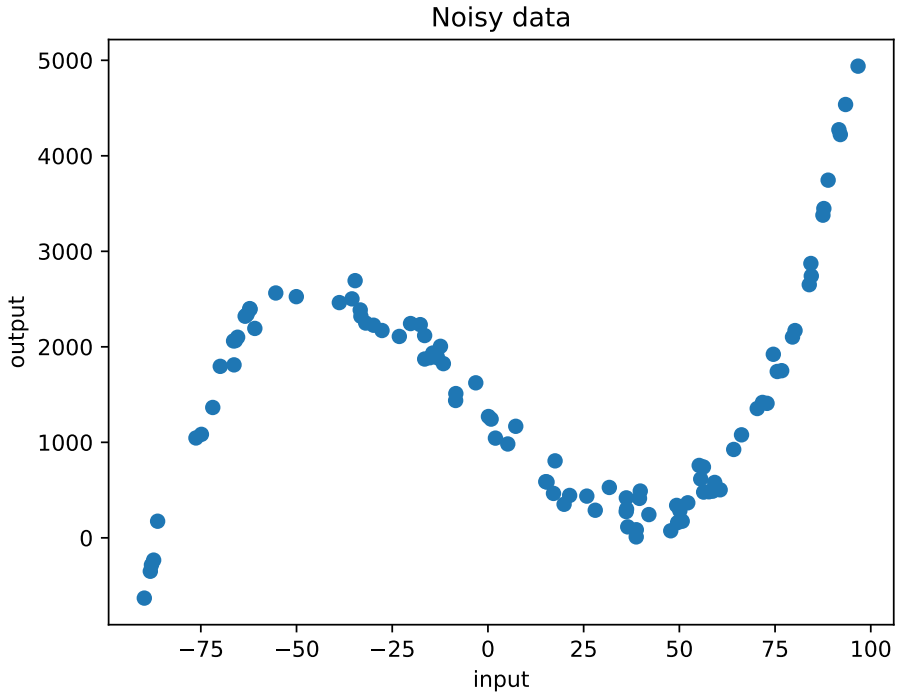
Overfitting

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





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



We need to choose:

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

Overfitting

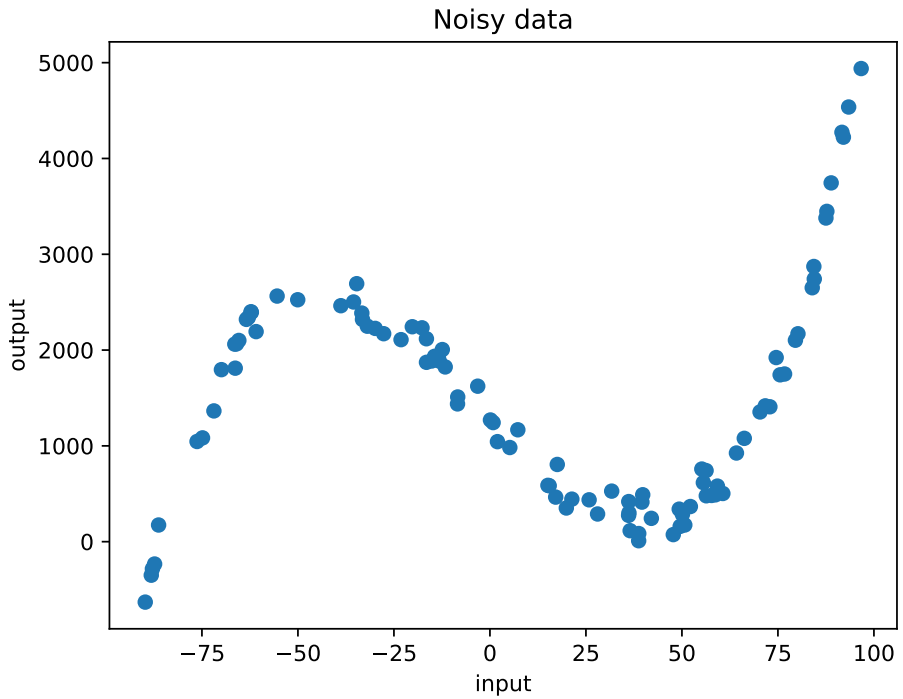
- What could be the drawbacks of using a very simple model (very few parameters) ?

Overfitting

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

Fitting

Exercice 5 : **Fitting polynoms to data**

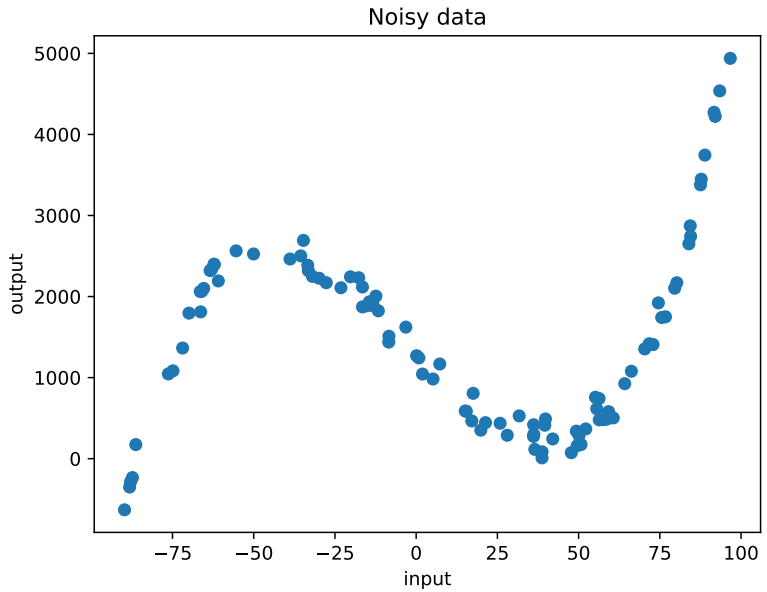


We want to perform supervised learning in order to be able to predict the output y for a new sample x.

Fitting

Exercice 5 : **Fitting polynoms to data**

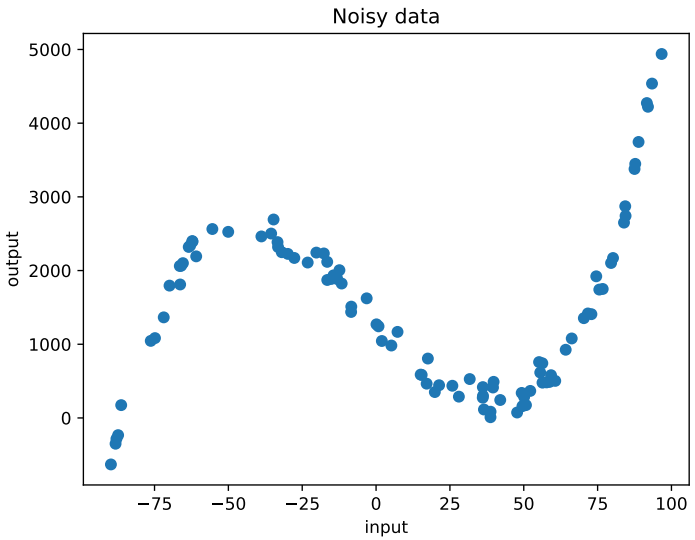
- We want to perform supervised learning in order to be able to predict the output y for a new sample x.



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

Fitting

Exercice 5 : **Fitting polynoms 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 polynoms 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 polynoms to data**

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

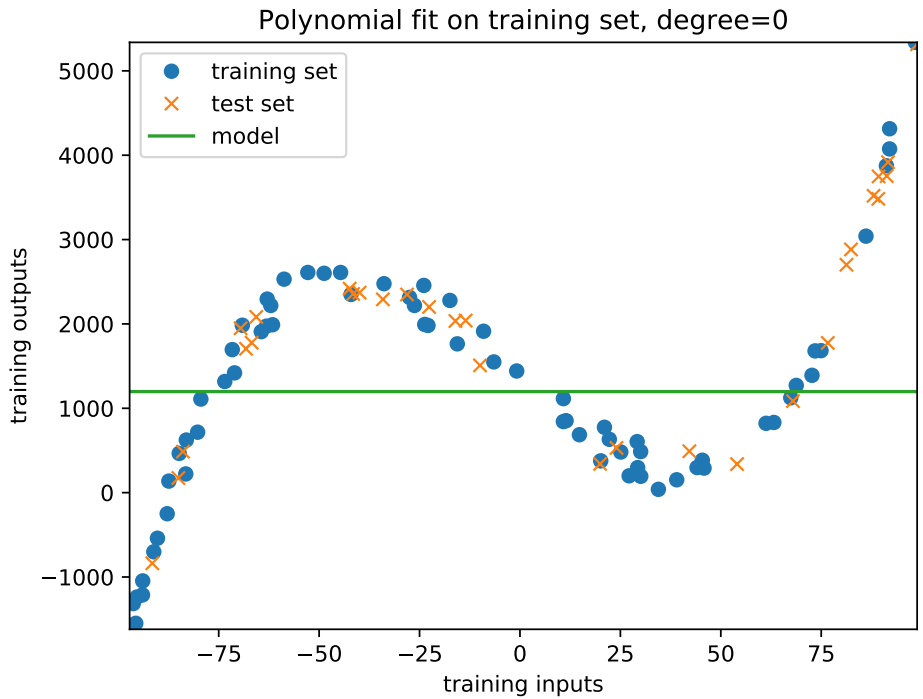


Figure: degree 0

Fitting

Exercice 5 : **Fitting polynoms to data**

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

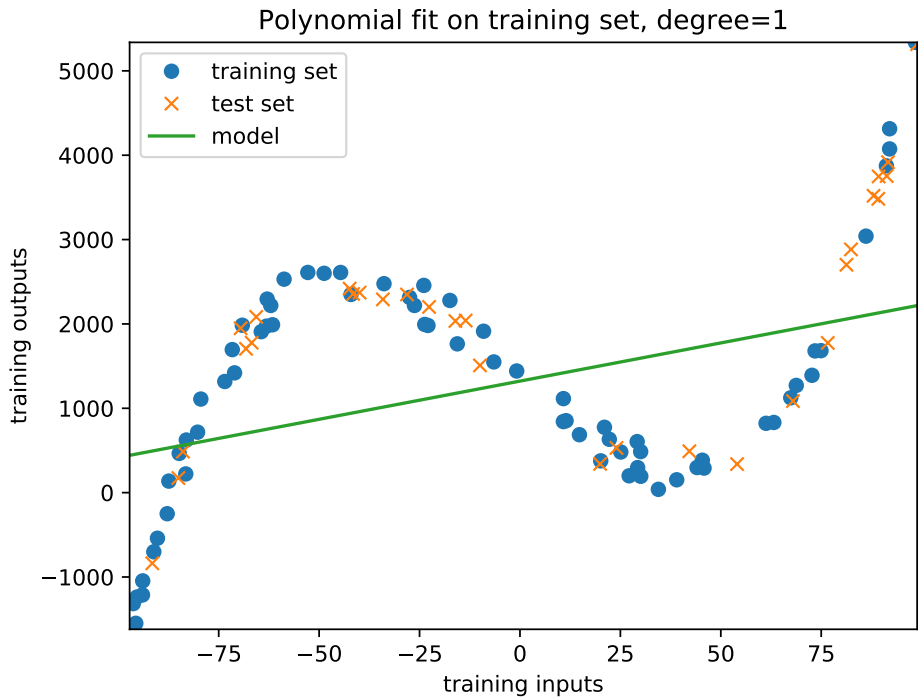


Figure: degree 1

Fitting

Exercice 5 : **Fitting polynoms to data**

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

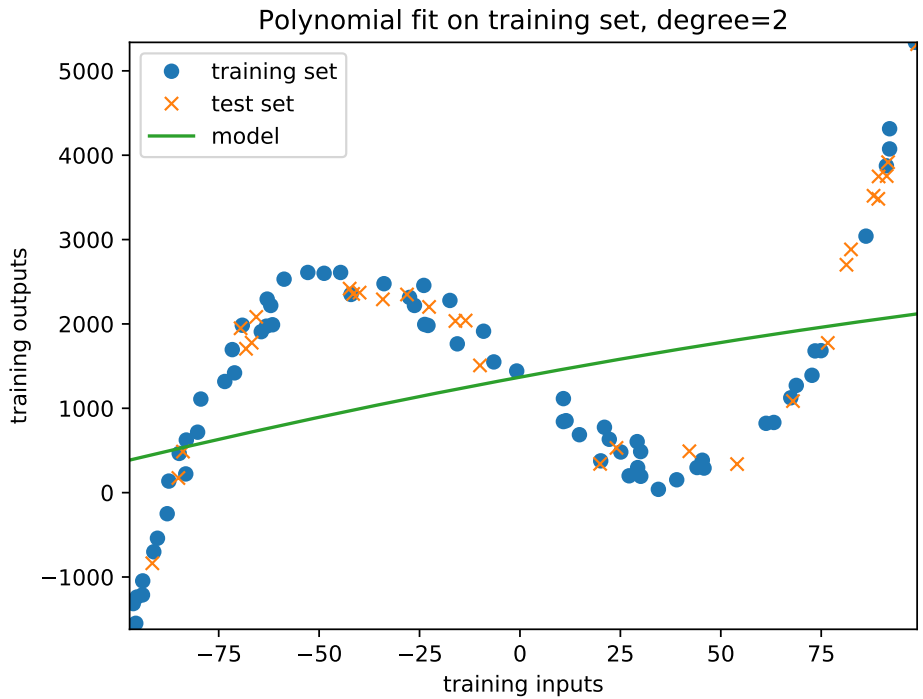


Figure: degree 2

Fitting

Exercice 5 : **Fitting polynoms to data**

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

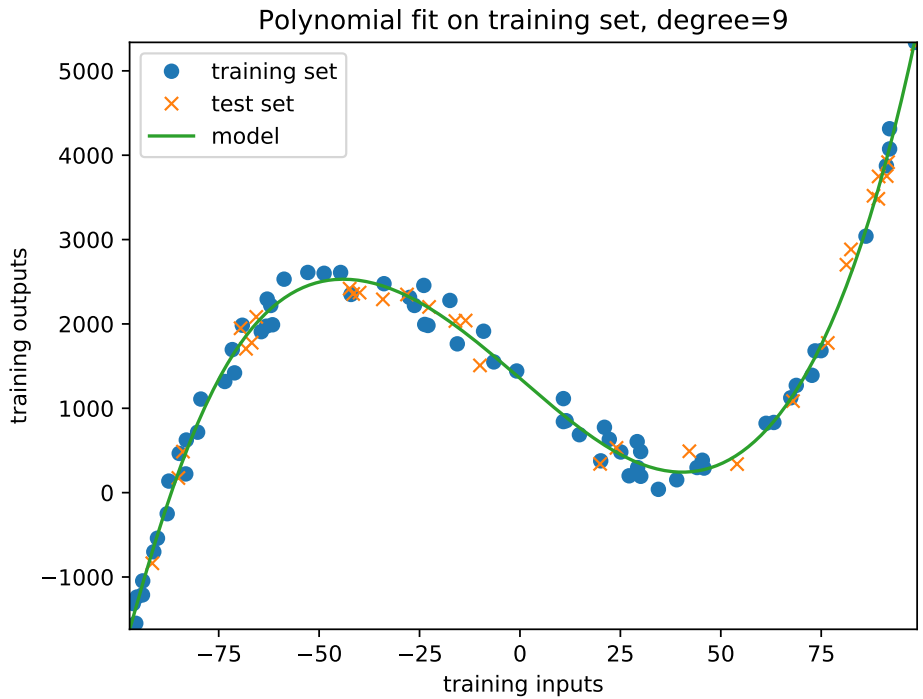


Figure: degree 9

Fitting

Exercice 5 : **Fitting polynoms to data**

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

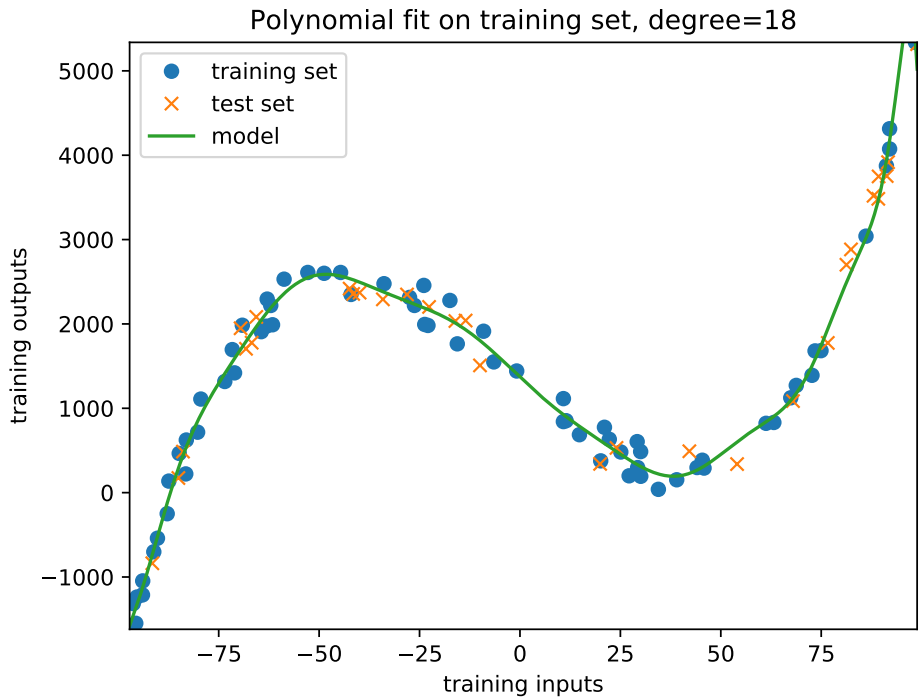


Figure: degree 19

Fitting

Exercice 5 : **Fitting polynoms to data**

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

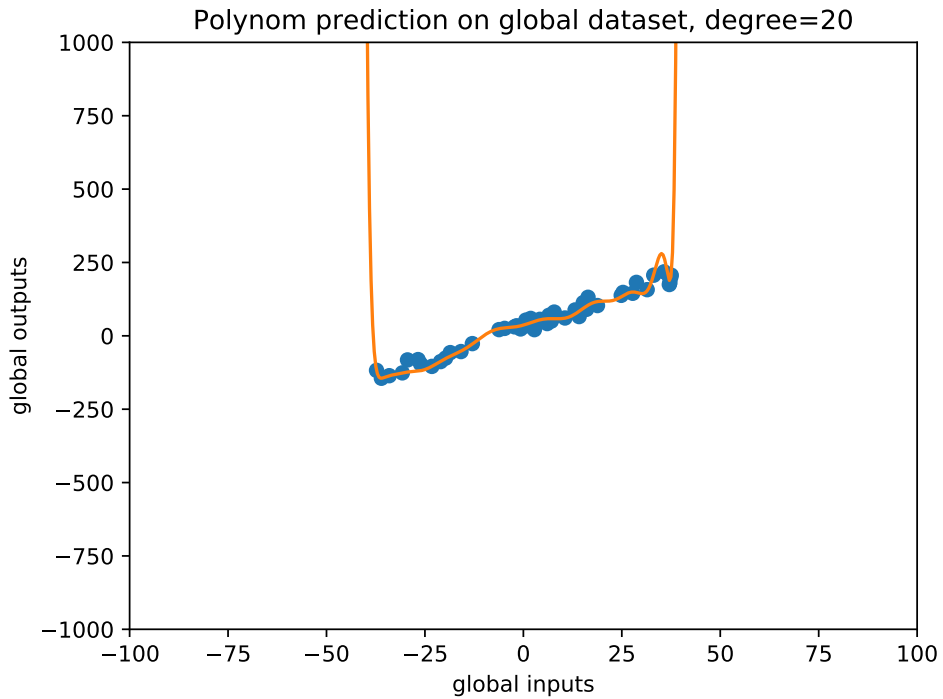


Figure: Useless solution

Fitting

Exercice 5 : **Fitting polynoms to data**

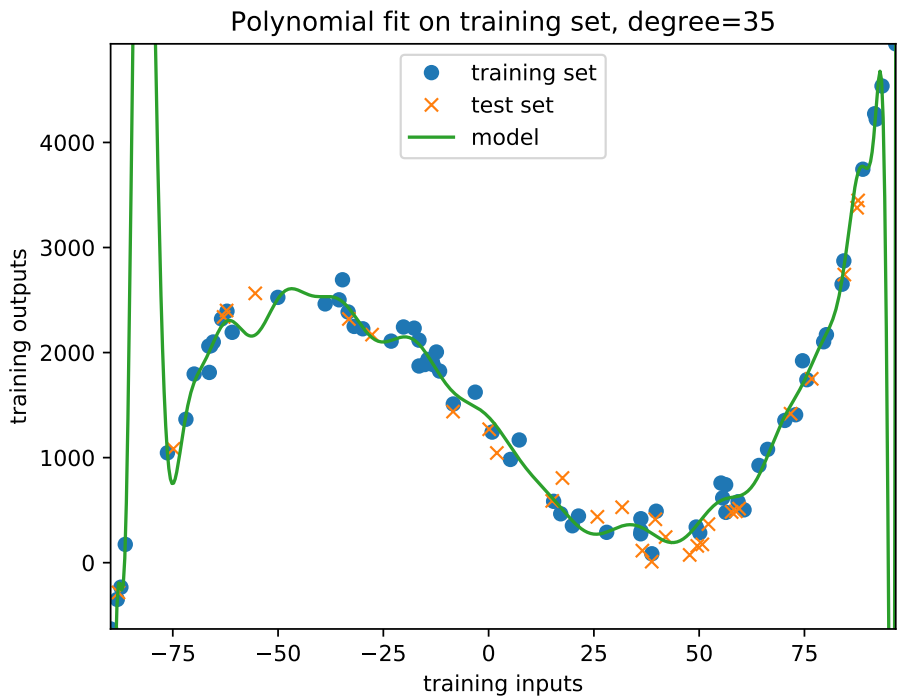


Figure: When the degree is too high.

Fitting

Exercice 5 : **Fitting polynoms to data**

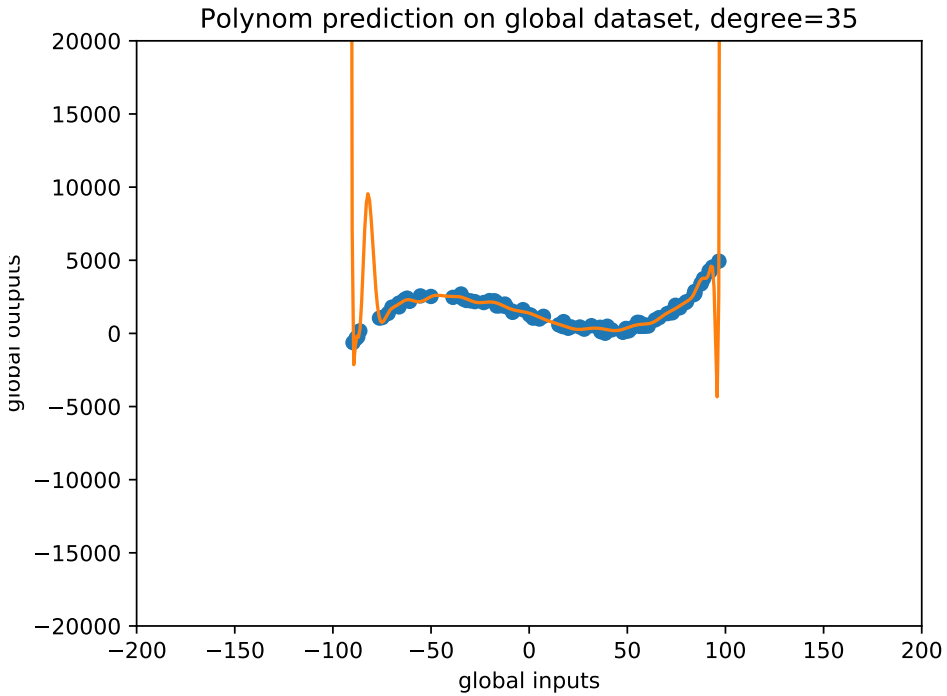


Figure: When the degree is too high.

Fitting

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

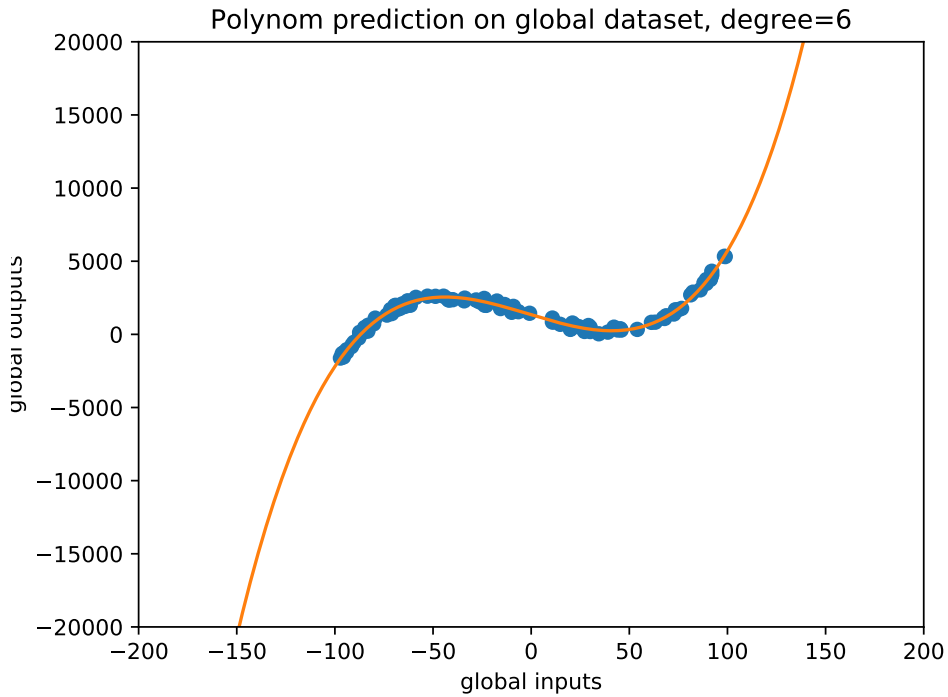


Figure: When the degree is too high.

Fitting

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

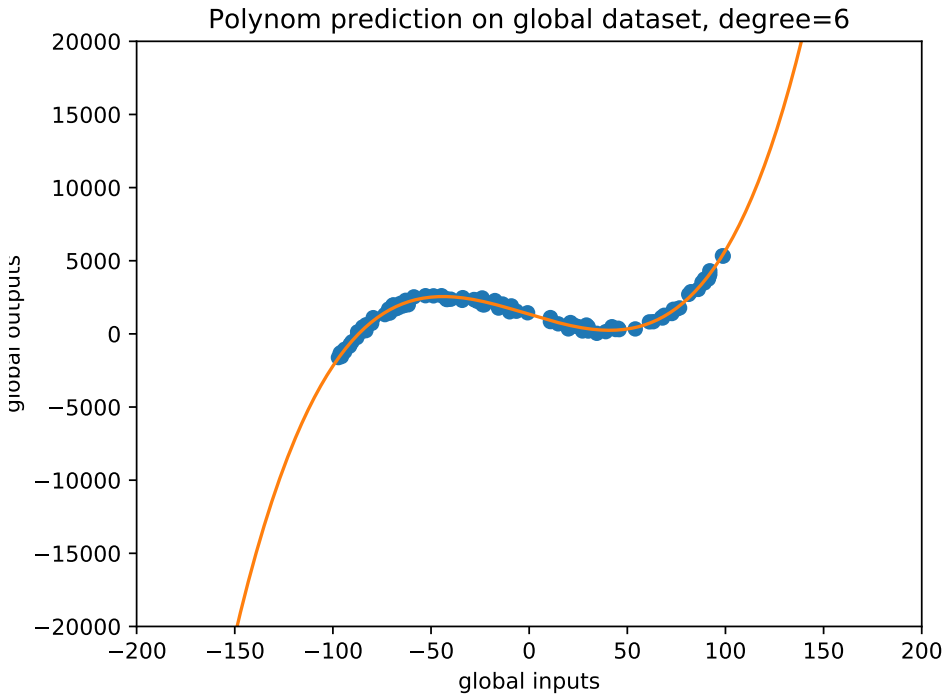


Figure: When the degree is too high.

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**

Regularization methods

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

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.