SD701: Big Data Mining

Louis Jachiet

About the course

Calendar

11/09	Amphi Rubis	Intro to Big Data and MapReduce
18/09	Amphi Grenat	Intro to Data Mining
25/09	Amphi Rubis	Classification & Clustering
2/10	Amphi Rubis	Classification & Clustering 2
9/10	Amphi Rubis	Spark
16/10	Amphi Rubis	Spark 2
23/10	Amphi Grenat	Frequent Pattern Mining
6/11	Amphi 3	Links Analysis

https://synapses.telecom-paris.fr/catalogue/occurrence/14437/edt

Objective of the course

Introduce to Big Data platforms

Objective of the course

Introduce to Big Data platforms

Data Mining: glossary, definitions & algorithms

Objective of the course

Introduce to Big Data platforms

Data Mining: glossary, definitions & algorithms

Some practical experience with Hadoop MapReduce and Spark for DataMining

```
FileOutputFormat.setOutputPath(job,
                               new Path(``out1''));
job.setMapperClass(Map.class);
job.setReducerClass(Reduce.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
```

Grades

There will be a project. It will be graded.

Grades

There will be a project. It will be graded.

The grade will be half project and half final exam.

What is Big Data?

From Wikipedia

> *"Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.*
>
> *– Wikipedia @ 09/09/2019*

The three Vs

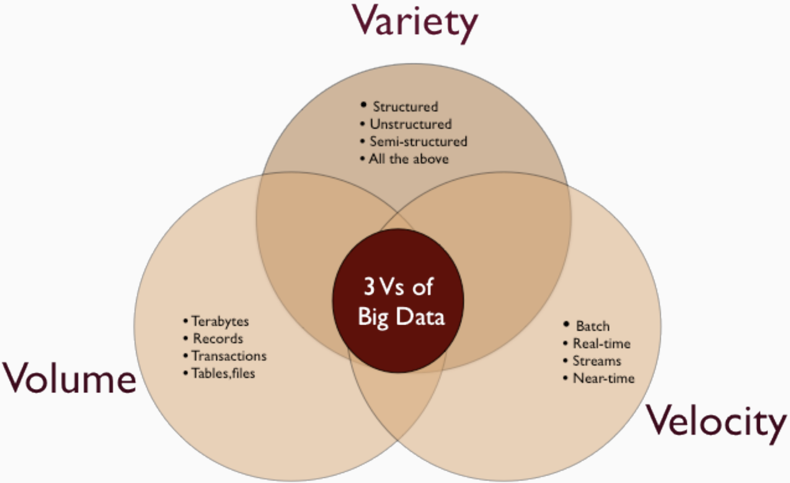


Figure 1: source Wikipedia

Big Data depends on the

CONTEXT.

Controversy of Big Data

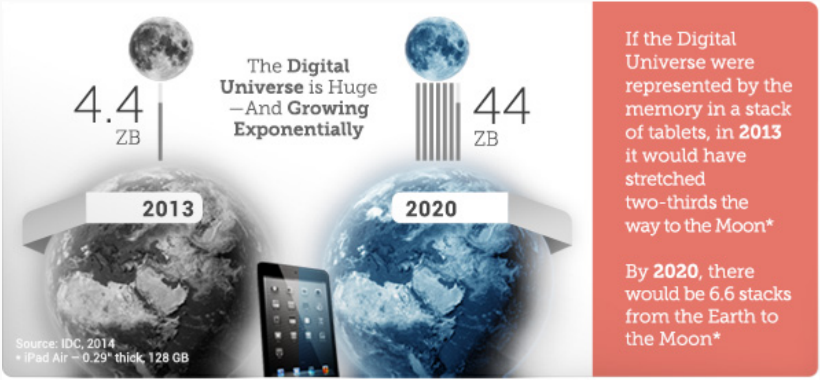
- All data is BIG now
- Hype to sell Hadoop based systems
- Ethical concerns about accessibility
- Limited access to Big Data creates new digital divides
- Statistical Significance:

When the number of variables grows, the number of fake correlations also grows Leinweber: S&P 500 stock index correlated with butter production in Bangladesh

The six Vs?

- Volume
- Variety
- Velocity
- Value
- Variability
- Veracity

Digital Universe



The Digital Universe is Huge —And Growing Exponentially

If the Digital Universe were represented by the memory in a stack of tablets, in **2013** it would have stretched two-thirds the way to the Moon*

By **2020**, there would be 6.6 stacks from the Earth to the Moon*

Source: IDC, 2014
• iPad Air – 0.29" thick, 128 GB

Figure 2: source EMC Digital Universe

Digital Universe

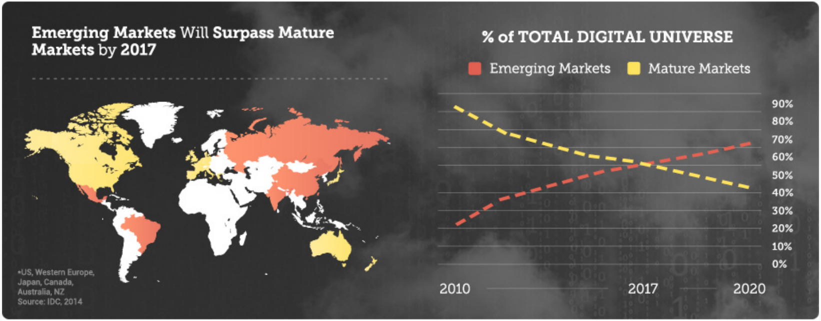


Figure 3: source EMC Digital Universe

Digital Universe

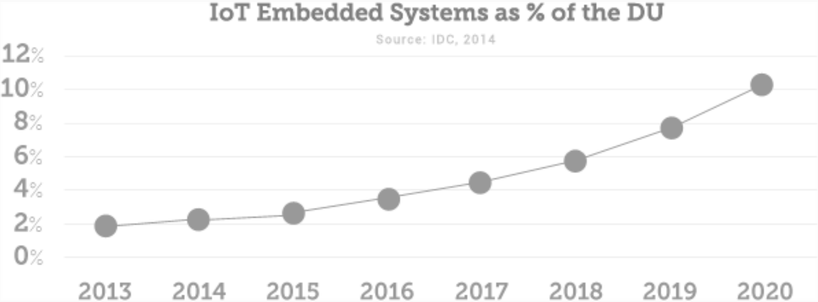
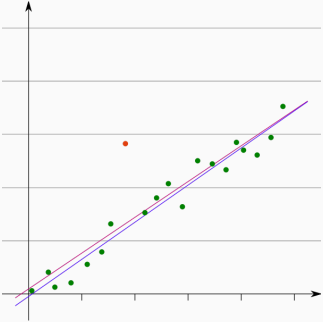


Figure 4: source EMC Digital Universe

Data Mining: the six classes of common tasks (adapted from Wikipedia)

Anomaly detection (outlier/change/deviation detection)

The identification of unusual data records, that might be interesting or data errors that require further investigation.



A website collects which IPs are trying to log in. They can try to detect which IPs have a high failure rate to block them (and prevent brute force guessing).

Association rule learning (dependency modeling)

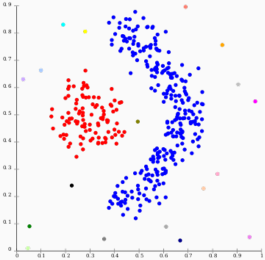
Search for relationships between variables.

$$X \wedge Y \Rightarrow Z$$

A supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering

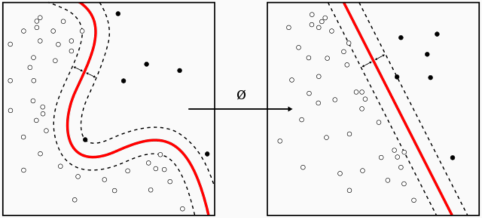
Discover groups and structures in the data that are in some way or another “similar”, without using known structures in the data.



Automatically create categories for collections of items (e.g. music records or movies)

Classification

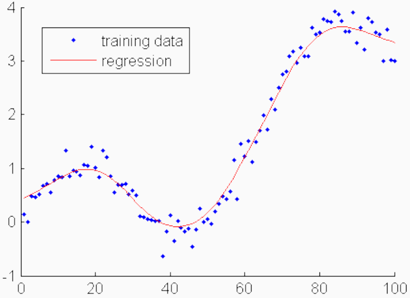
Generalize known structures to apply to new data.



An e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression

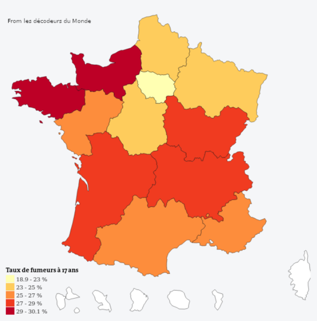
Find a function which models the data with the least error that is, for estimating the relationships among data or datasets.



Predict electricity consumption using: weather forecasts, TV program, day of the week, etc.

Summarization

Provide a more compact representation of the data set, including visualization and report generation.



From les décodeurs du Monde