# Data Mining & Big data

**Violaine Antoine**

ISIMA / LIMOS

January, 2020

# Outline

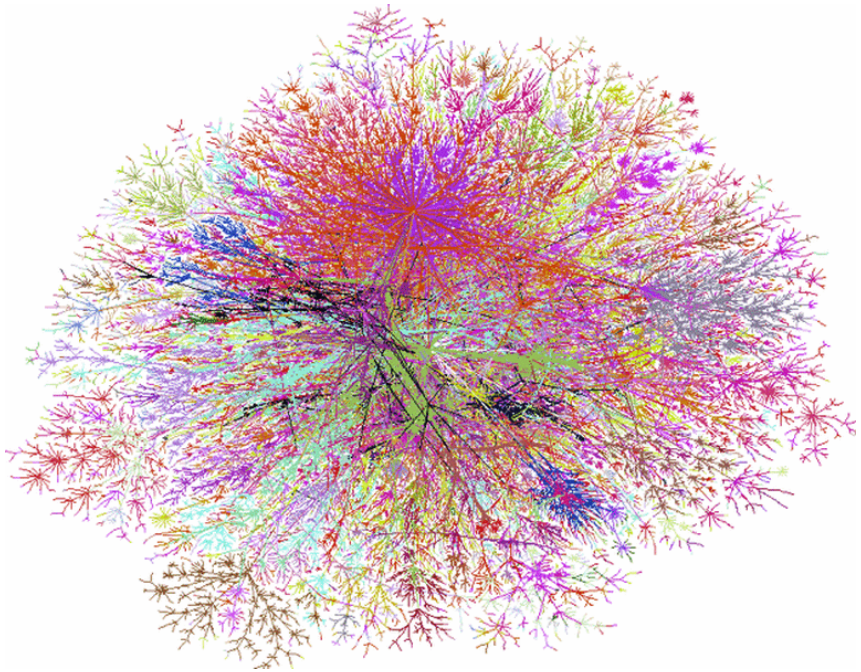
- Introduction to Data Mining
- Data preprocessing
- Clustering
- Feature reduction
- Time Series
- Association rules

# Data definition

- Collection of $n$ objects
- Each object has $p$ dimensions
- The collection is divided in $c$ groups (optional)

# Big Data definition

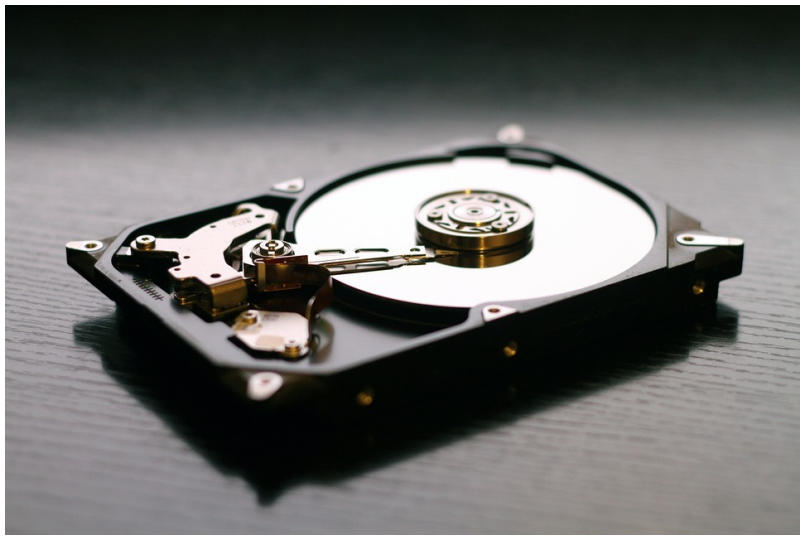
- *n* large, *p* large, *c* large
- Big data corresponds to a dataset too large to be handle with usual data mining techniques.



# Evolution of the volumes

Over years, hard disk capacity are increasing :

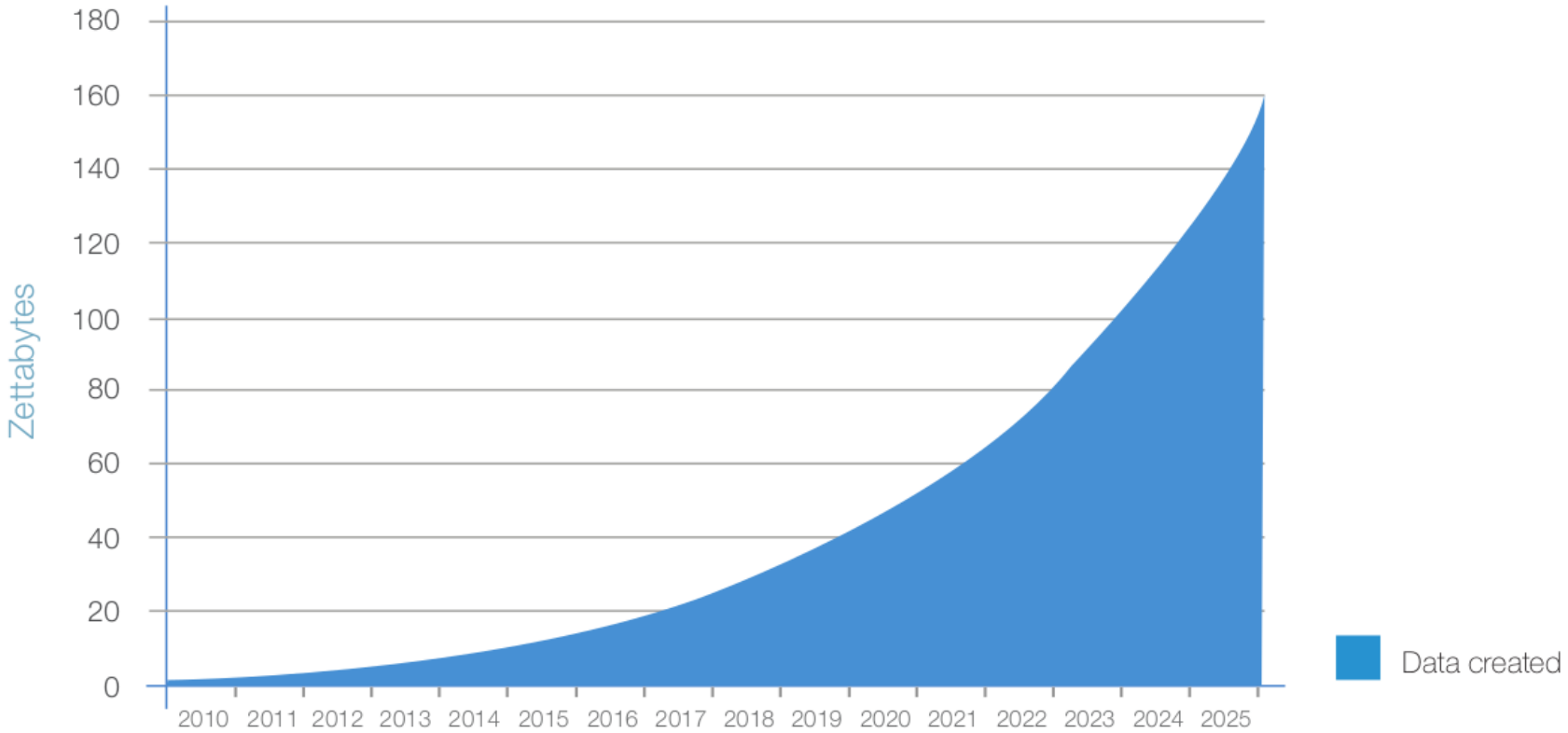
- 1956 : 5 MB
- 1980 : 1.26 GB
- 2007 : 1 TB
- 2017 : 12 TB
- announced for 2018 : 14 TB



# Evolution of the volumes

Data volumes follow the evolution of the hard drives :

- gigabytes → terabytes → petabytes → exabytes → zetabytes




Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

# Examples

Social media and network

- facebook, twitter, web browser, internet of things...

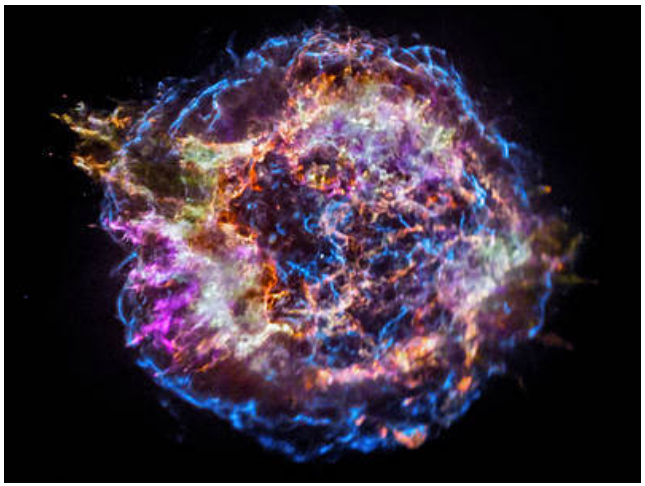
Trades, transactions

Healthcare

- electronic health record, images, questionnaires, sensors...

Science

- astronomy, climate, genetic, hydrology



Cassiopeia A, NASA Image, 12-13-2017

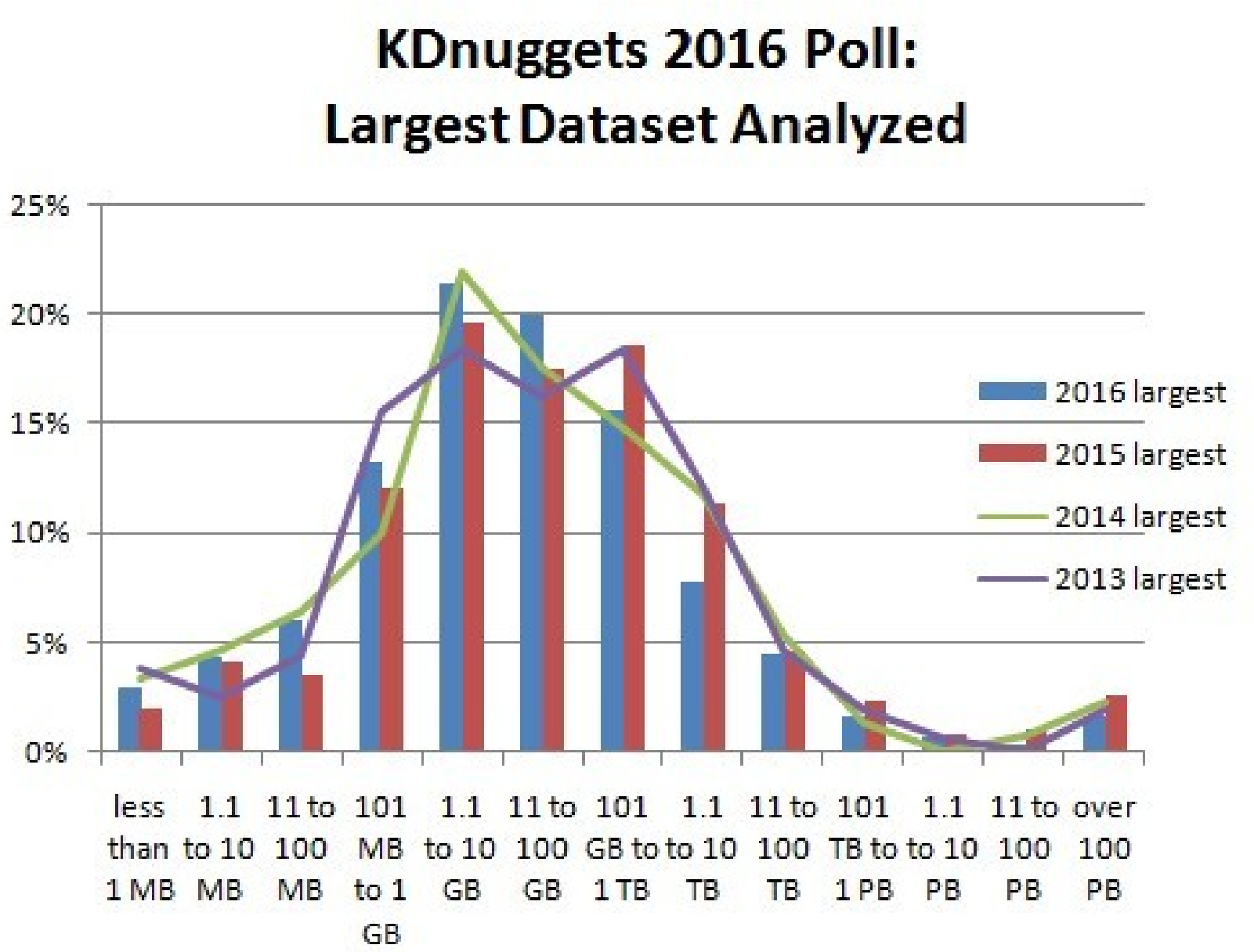
# Examples

Large Synoptic Survey Telescope (LSST) :

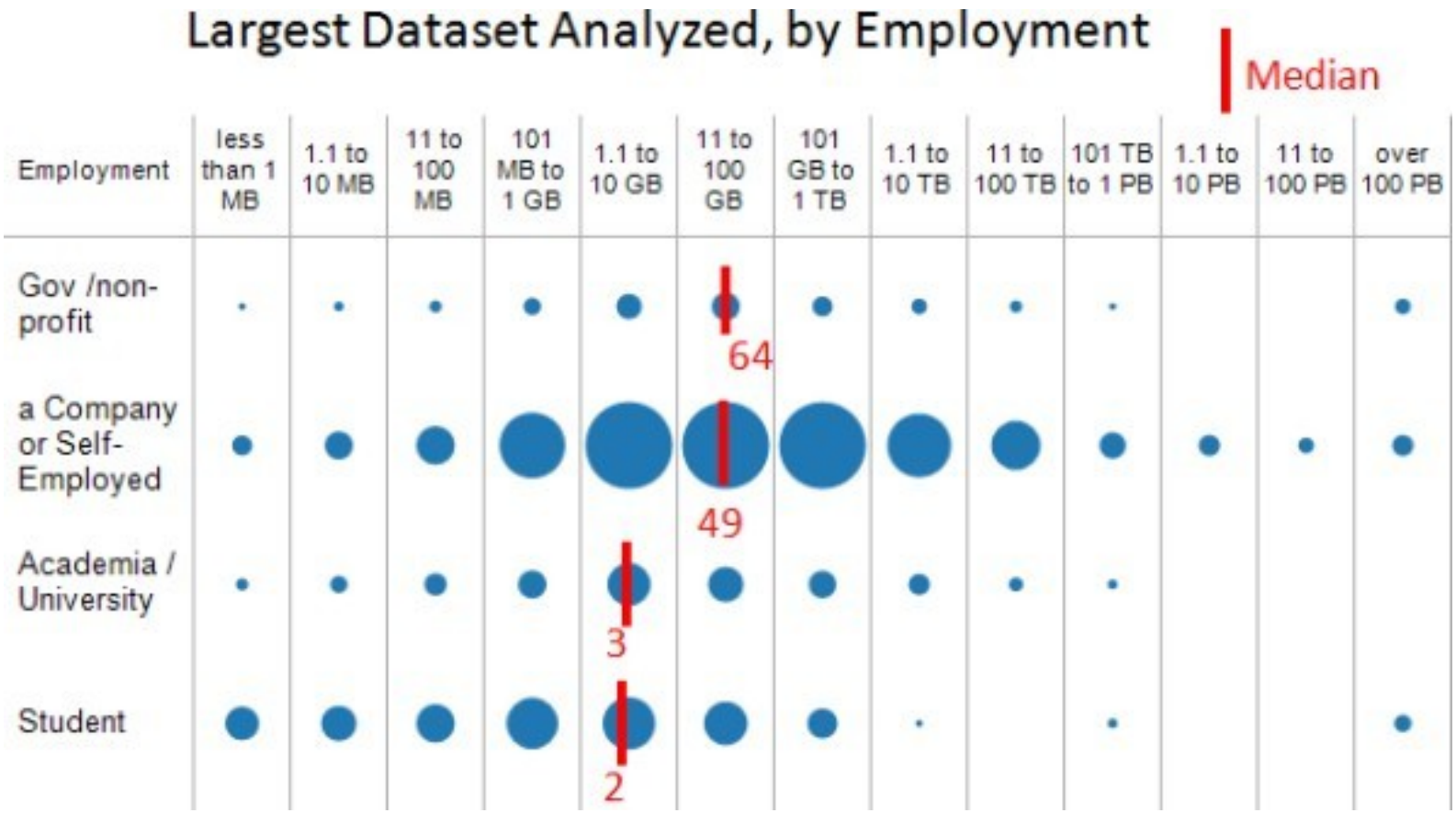
- each image the size of 40 full moons,
- 37 billion stars and galaxies,
- 10 year survey of the sky,
- 10 million alerts, 1000 pairs of exposures,
- 15 Terabytes of data every night.

https ://www.lsst.org

# Big Data analysis



# Big Data analysis




KDnuggets 2016 Poll

# Data Mining definition

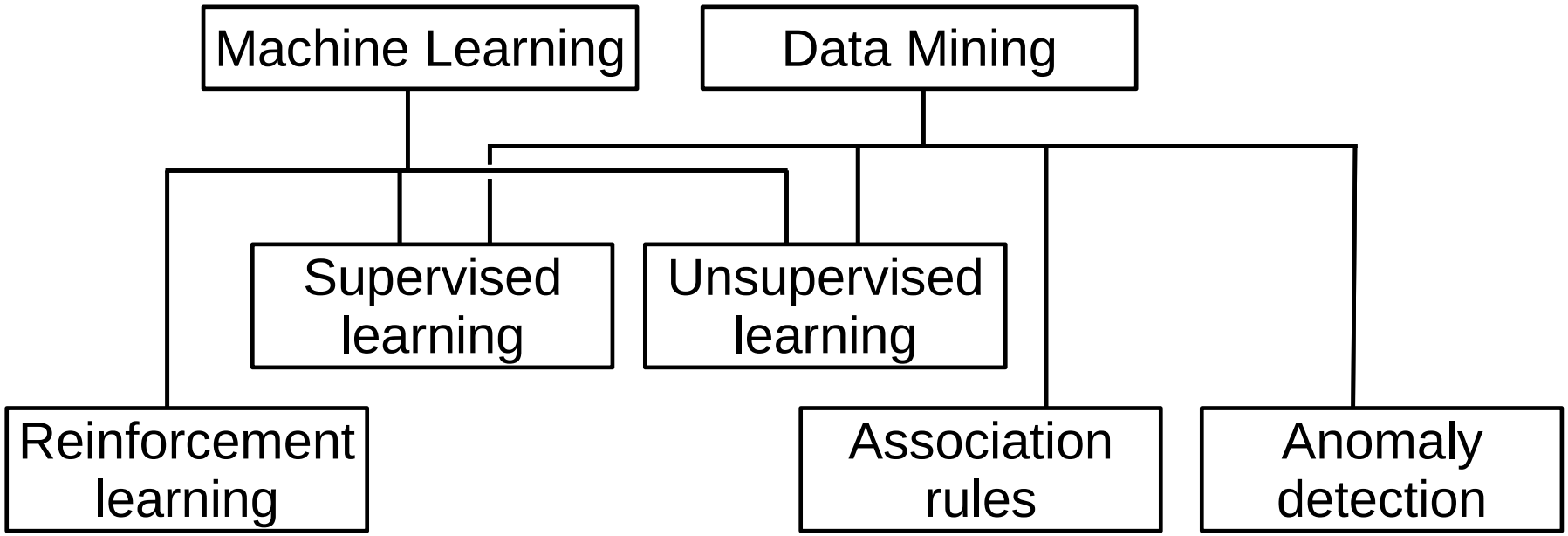
- Discover of unknown knowledge from a specific dataset
- Focus on data, applications and algorithms


Data Mining includes data modeling

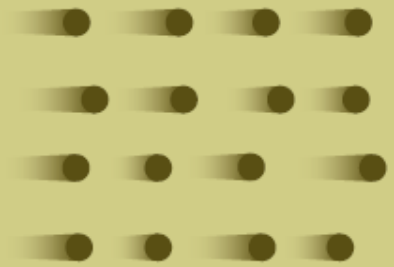
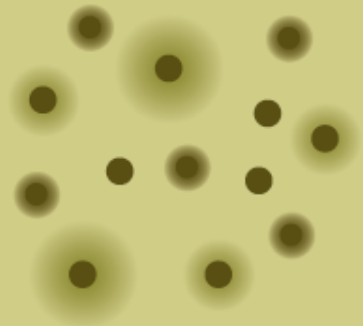
- supervised learning
- unsupervised learning
- association rules
- anomaly detection

# Data Mining vs Machine Learning

| | Data Mining | Machine Learning |
|---|---|---|
| def | Discover of unknown knowledge from a specific dataset | Learn from data to make prediction or decision |
| focus | data, applications and theory | algorithms and theory |



# Challenges

| Volume | Velocity | Variety | Veracity |
|---|---|---|---|
|  |  |  |  |
| **Data at rest** | **Data in motion** | **Data in many forms** | **Data in doubt** |
| Terabytes to exabytes of existing data to process | Streaming data, milliseconds to seconds to respond | Structured, unstructured, text and multimedia | Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations |

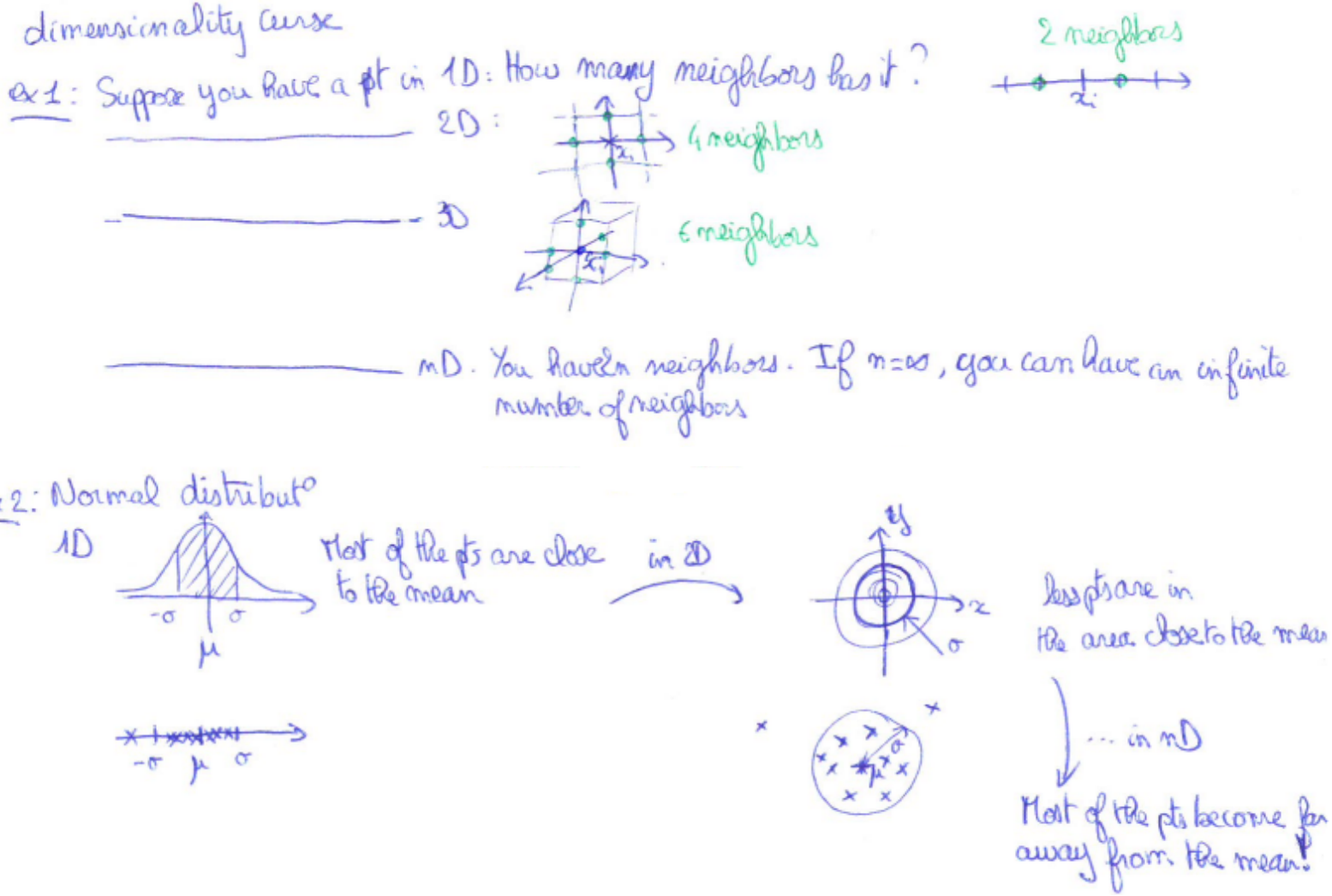
IBM, 2013

# Challenges

## Large number of objects

# Challenges

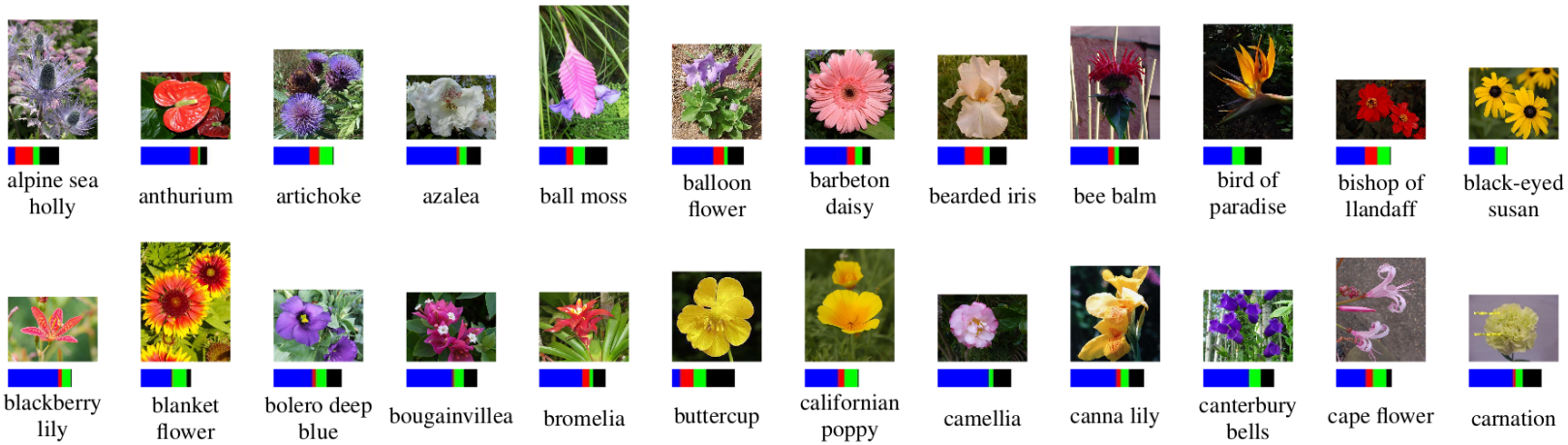
## Dimensionality curse with large number of attributes



# Challenges

Large number of classes :

- plants variety,
- cities in the world,
- whistled tune, ...



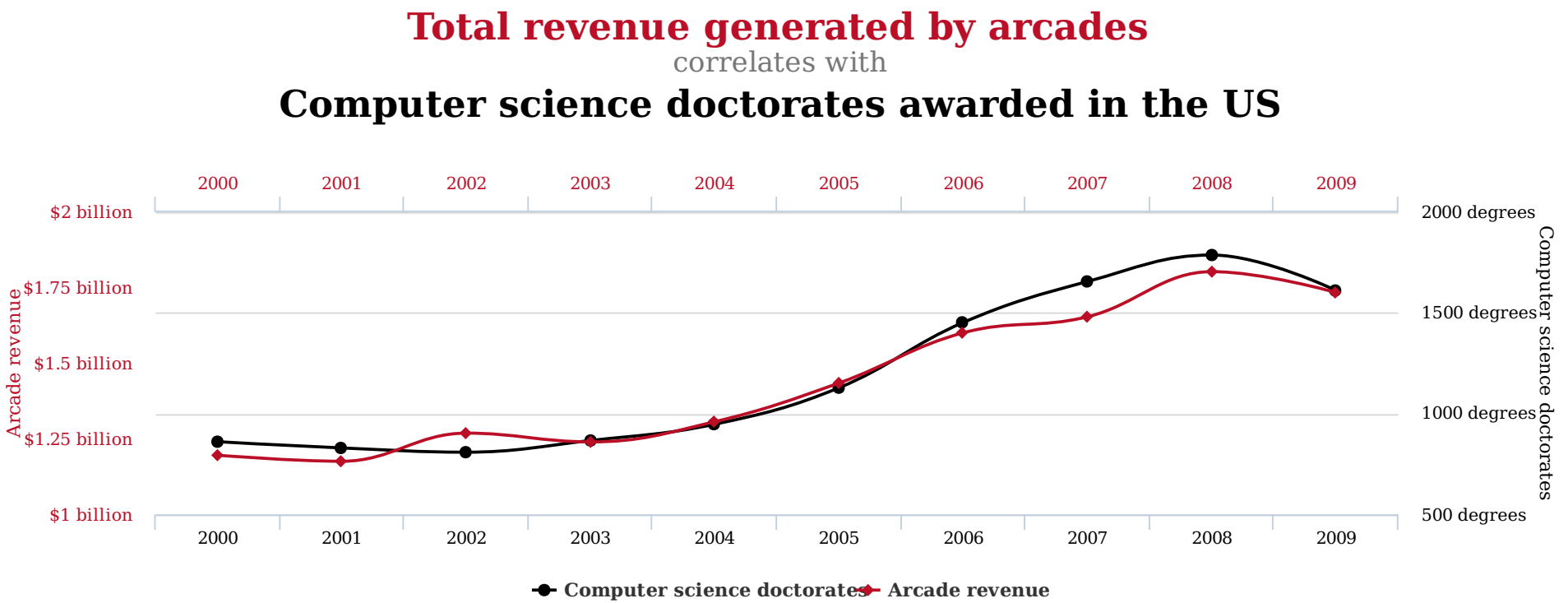

M. Nilsback & al, **Automated flower classification over a large number of classes**, ICVGIP, 2008.

Multiclass classification are usually $O(c)$ !

# Big data criticisms : Result interpretation


xkcd.com

# Big data criticisms : Result interpretation

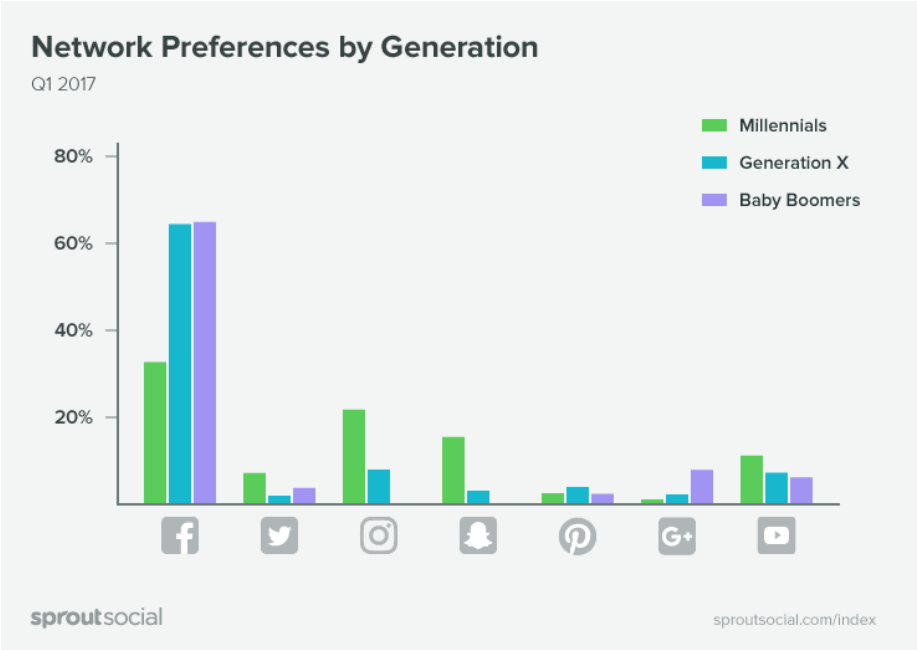



Read more : M. Ebach & al, **Big data and the historical sciences : A critique**. Geoforum, vol 71, pp 1-4, 2016.

# Big data criticisms : Big data bias

Big data might not well represent a community

- Active person on the Internet are in minority, non active are in majority
- Applications or cell phones analysis does not count children and elderly




Millennials : 18-34
Generation X : 35-54
Baby Boomers : 55-$\infty$

$\Rightarrow$ Choose a sample can be better than handling big data.

# Competences needed

- mathematics (statistics, optimization...)
- computer sciences (data science programming and sofwares)
- domain expertise