Introduction to Big Data

Lesson 1.1 What is big data?

Arthur Katossky & Rémi Pépin

Wednesday, March 30, 2022

From "Big Data" to "Large Scale"

What is "Big Data" ?

What is "Big Data" ?

- started to be used in the 1990's
- is a vague, ill-defined notion
- refers to data that **cannot be managed by commonly-used software**
- is inherently relative to ***who*** is using it

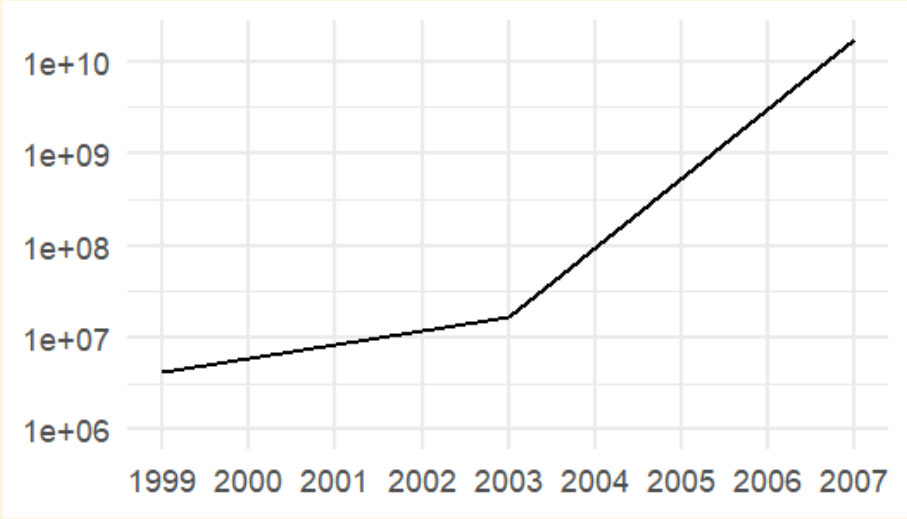
What is "Big Data" ?

What can be considered "big" **evolves over time** since **software constantly improves capacity**.

*Line and row limits of **Microsoft Excel** tablesheets*

	Version	Lines	Columns
until 1995	7.0	16 384	256
until 2003	11.0	65 536	256
from 2007	12.0	1 048 576	16 384

Max. number of items stored in one tablesheet



What is "Big Data" ?

Besides software, **hardware is also evolving** :

- faster and more massive storage
- faster and more massive memory
- more and faster processors
- more specialized processors
- better connectivity
- improved architecture

For instance, it now possible to use graphic cards to perform computation.

What is "Big Data" ?

Size is **not** the only thing that matters.

In a tablesheet program, what kind of information can't you store properly?

- relationnal data
- images, long texts
- unstructured data (ex: web page)
- rapidly varying data (ex: tweeter feed)

What is "Big Data" ?

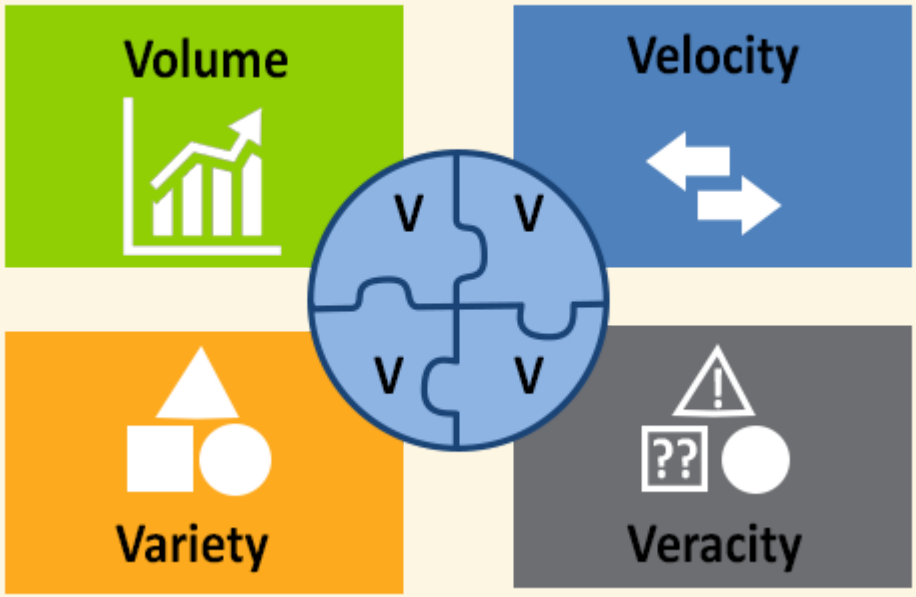
The **"3 V's of Big Data"**

- **V**olume (massive, taking place)
- **V**elocity (fast, updated constantly)
- **V**ariety (tabular, structured, unstructured, of unknown nature, mixed)

Specific challenges to each **V**

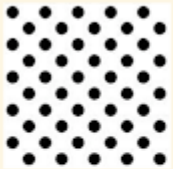
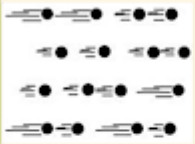
What is "Big Data" ?

Marketers constantly **invent new V's**



What is "Big Data" ?

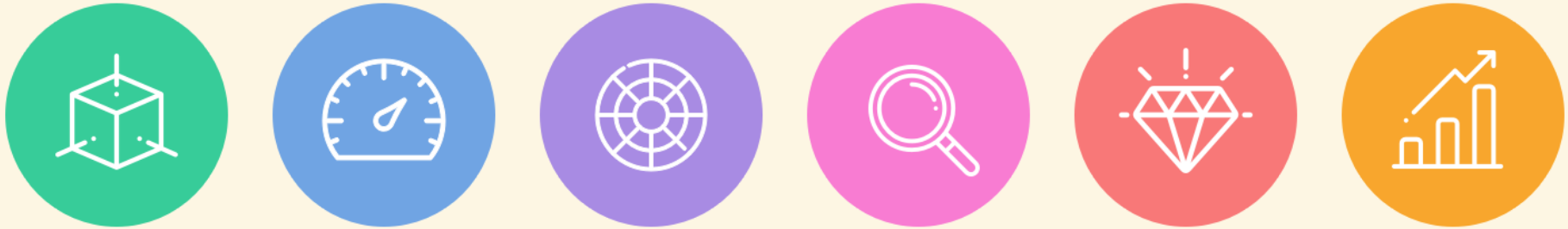
Marketers constantly **invent new V's**

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,...	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

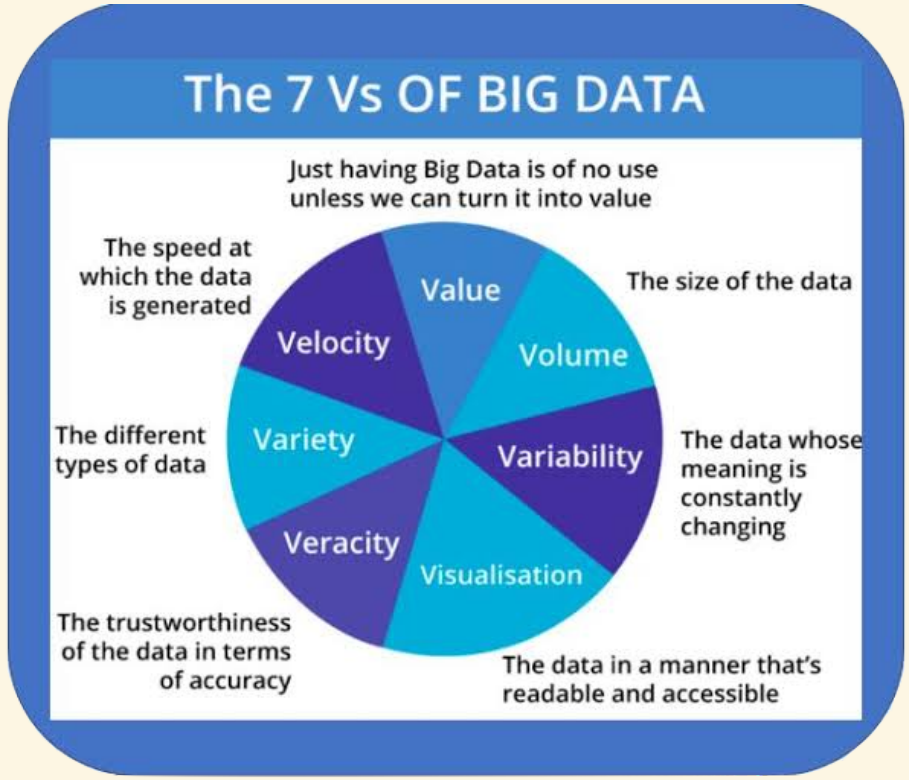
What is "Big Data" ?

Marketers constantly **invent new V's**



What is "Big Data" ?

Marketers constantly **invent new V's**

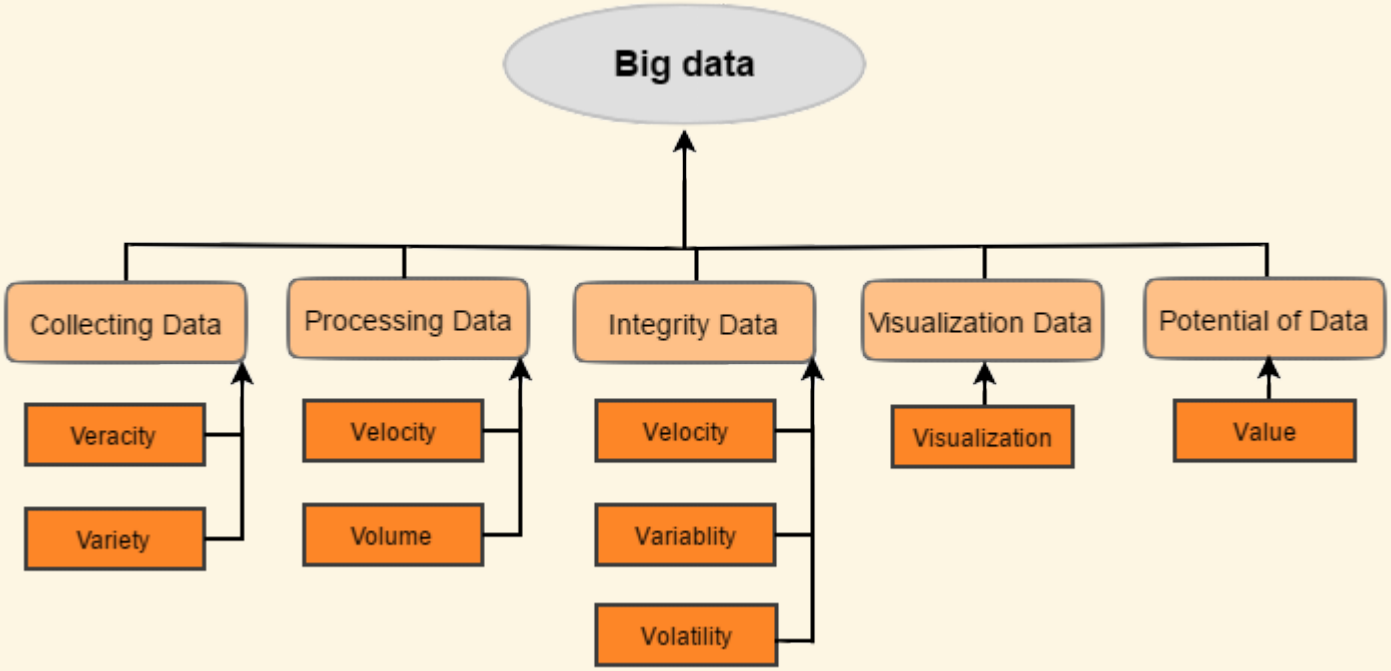


What is "Big Data" ?

Marketers constantly **invent new V's**

What is "Big Data" ?

Marketers constantly **invent new V's**



What is "Big Data" ?

Marketers constantly **invent new V's**

How big is "Big Data" ?

To this day, can be considered "big" for a statistician :

- more than 1 000 000 rows and/or more than 100 000 columns
- one individual file over 4 Go [1]
- memory needs over 8-10 Go [2]
- storage needs over 0.1 to 1 To [2]
- processing taking over one hour (see later)

[1] Limit for many old computers, external drives or USB sticks (FAT32 file system)

[2] Average desktop

How big is "Big Data" ? (volume)

What is usually **not** an issue:

- file size on modern OS [3]
- DBMS *constraints* [4]

What is usually an issue:

- your physical storage
- computation time

[3] Over 16 To for Window's NTFS, Unix's ext4 and Apple's APFS

[4] You will probably run out of disk space before you can reach MySQL's limit of 65 536 To

How fast is "Big Data" ? (velocity)

To this day, **any data change is challenging** for a statistician :

- **most data analysis is done with data at rest**
- most spreadsheet or statistical software is not well adapted to streams of data

How heterogeneous is "Big Data" ? (variety)

To this day, **any non-textual format** and **any non-standard textual format** is an issue :

- **images**, **sounds**, **videos** are not supported by most spreadsheets or statistical software and rarely supported by non-specialized databases
- any non-standard format is a challenge (ex: annotated texts, vocal message conversations, etc.). How to store to process ?

What is usually **not** an issue:

- storing network data (e.g. social graphs) (neo4J)
- storing formatted text such as XML, JSON, HTML, etc (mongoDB)
- storing raw, unformatted text (Elasticsearch, apache Solr)
- storing spatial data (PostGIS)

Processing those data can be hard and takes time

From big-data to large-scale computing

BUT data do not just sit here:

- we copy, transform, use data
- "big" data are especially **data that take time to process**
- focus shift from "big data" to **"large-scale computing"**

There are **many possible issues** with processing a lot of data...

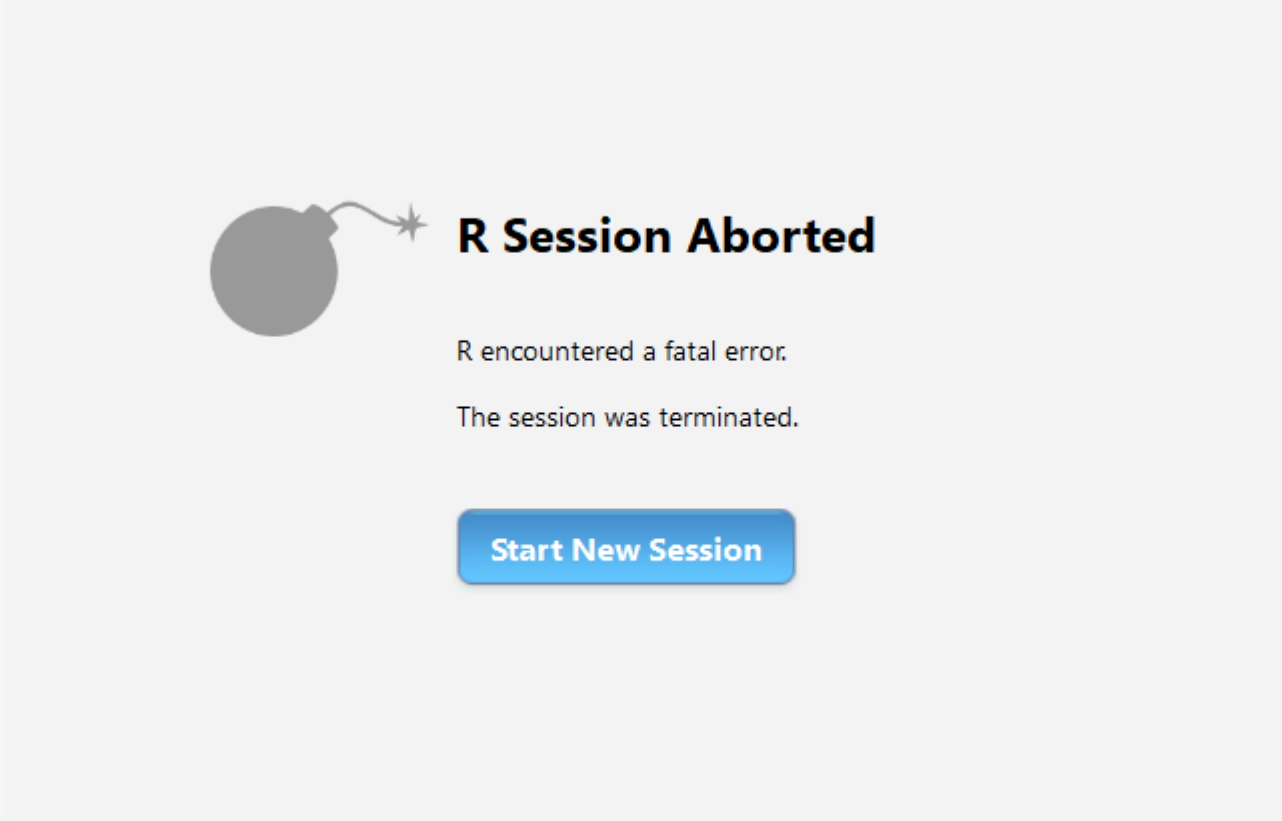
From big-data to large-scale computing

...problems of **memory**...

```
Error: cannot allocate vector of size 3.6 Mb
Error: cannot allocate vector of size 122 Kb
```

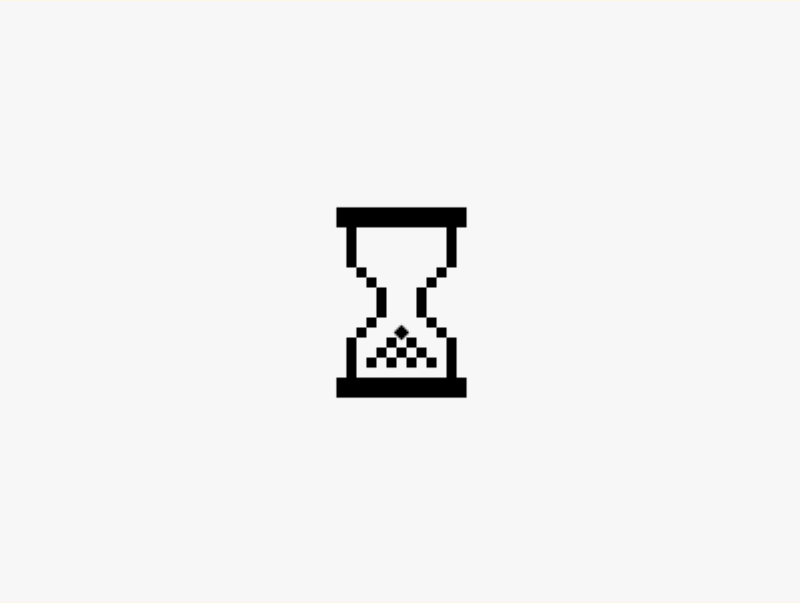
From big-data to large-scale computing

... problems of **unexpected failures**...



From big-data to large-scale computing

... problems of **computation time**...



From big-data to large-scale computing

Shift from:

> How better do my estimations become **for each additional observation** ?

To:

> How better do my estimations become **for each additional second of computation**?

- We often can't use all the data we have at hand
- **Sampling is always possible** and encouraged, but can we do better?

Pragmatic limits to storage and computing

- physical constraints
- financial constraints
- ecological constraints
- ethical constraints
- political consideration

Statistical issues

- the weakest relations become significant asymptotically
- curse of dimensionality: when the number of columns or variables p increases, the p-dimensional space becomes very vast and empty, and the observations become inevitably very sparse
- issue when p increases proportionally to n (the number of rows or observations), giving rise to the field of high-dimmensionnal statistics
- computational issues cumulate (e.g. rounding errors)

IN THE NEXT SECTION

- **goal :** understand why some data are hard to store and use
- **necessary step:** look at how a regular simple computer work