Introduction to Big Data

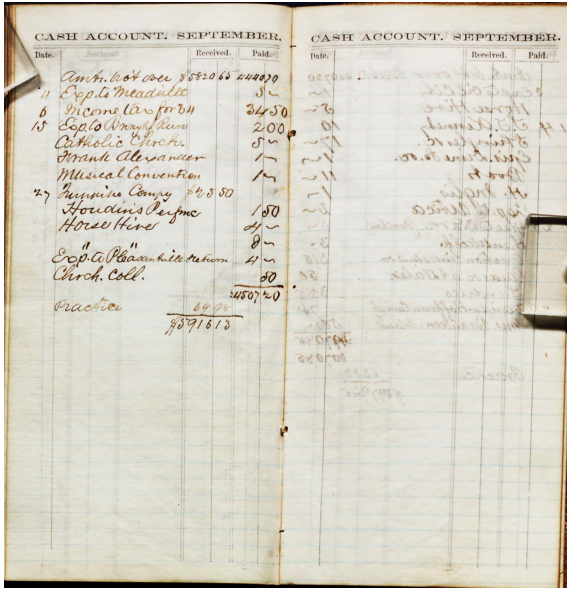
Julián Arenas Guerrero
julian.arenas.guerrero@upm.es

With some slides from Jesús Montes

Nov. 2024

What is Big Data?

Item	Amount
Hop	42
Barley	84

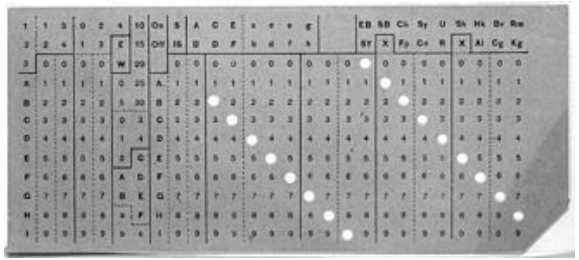


Stone tablet

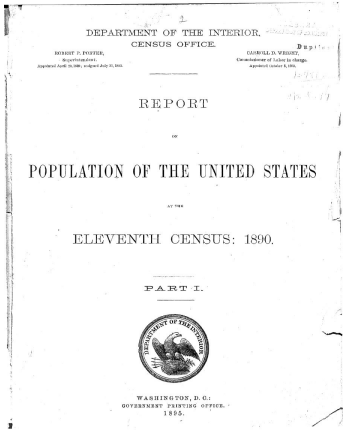
Accounting record

What is Big Data?

Hollerith tabulating machine



Hollerith punched card



DEPARTMENT OF THE INTERIOR.
CENSUS OFFICE.

REPORT

POPULATION OF THE UNITED STATES

ELEVENTH CENSUS: 1890.

PART I.

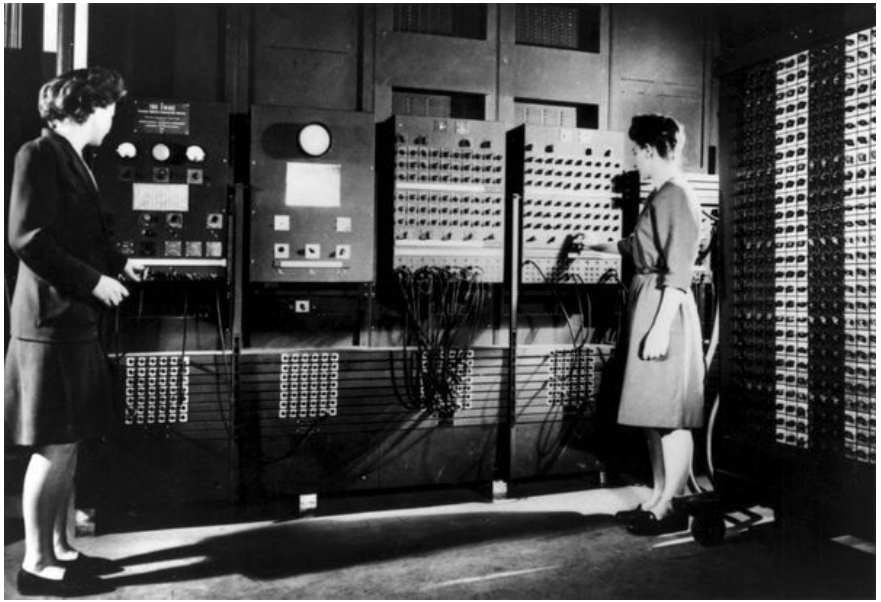
WASHINGTON, D. C.:
GOVERNMENT PRINTING OFFICE.
1895.

What is Big Data?

Computing-Tabulating-Recording Company

International Business Machine Corporation

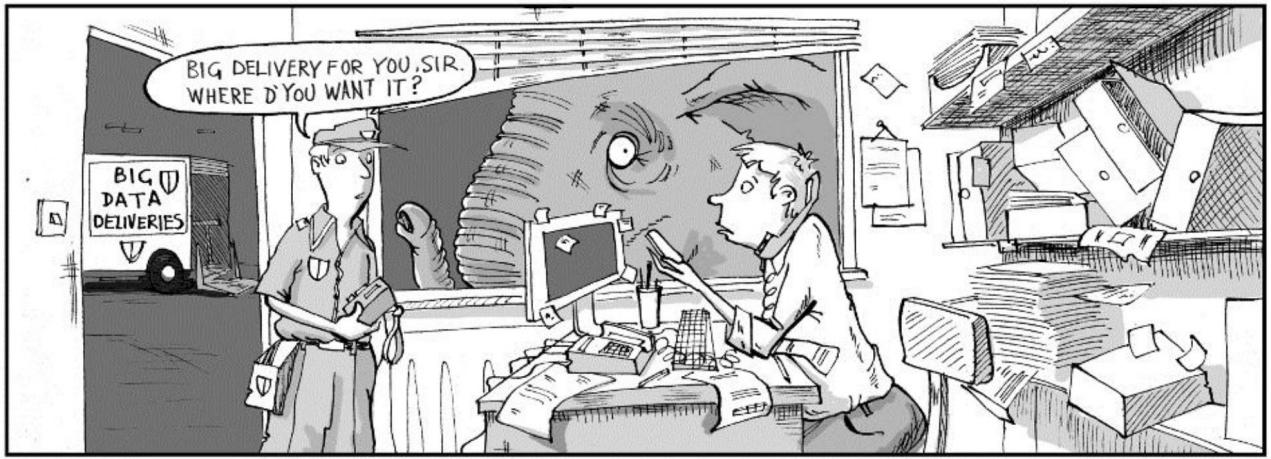
What is Big Data?



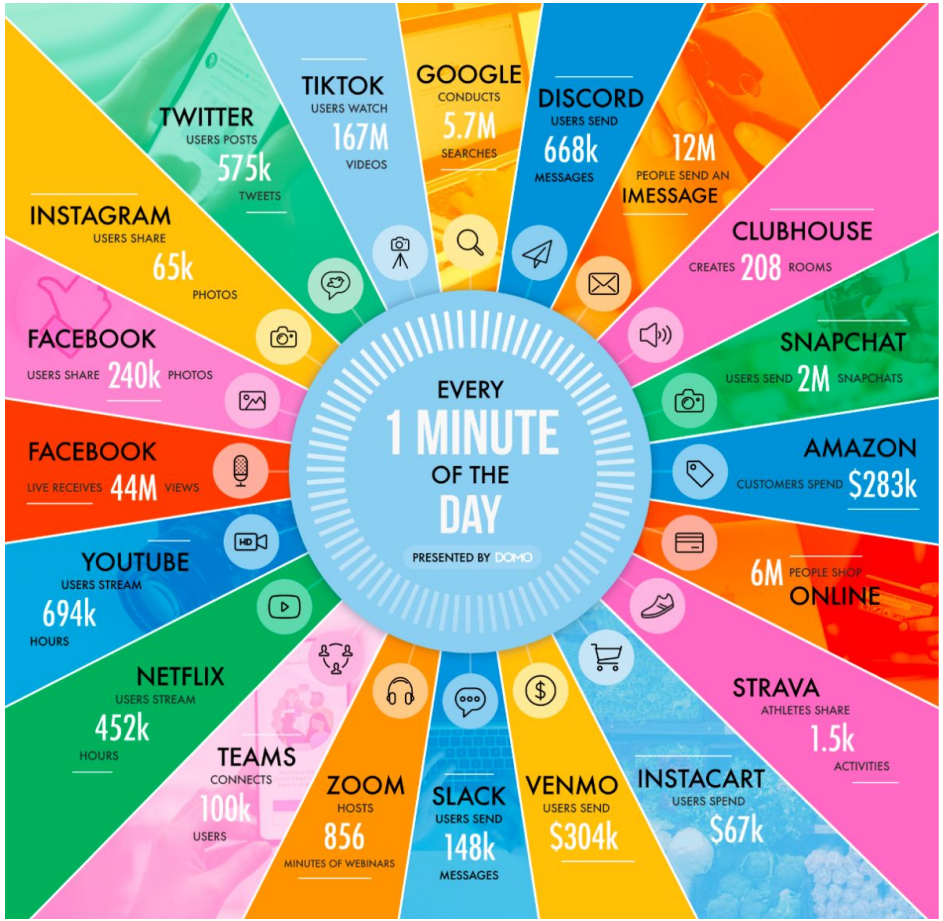
ENIAC computer

What is Big Data?

- The Information Age: economy and society built around information technology
- We constantly generate and consume data: manage bank accounts, streaming services, social networks, Wikipedia, etc



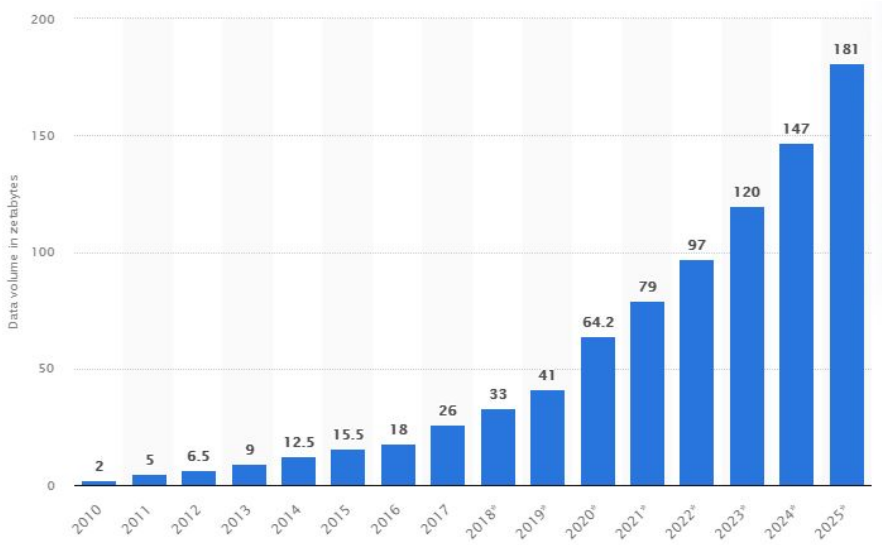
But, how big?



Source:
https://www.visualcapitalist.com/from-amazon-to-zoom-what-happens-in-an-internet-minute-in-2021/

But, how big?

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025



Estimated, in ZB (source)

Wait a minute, ZB (zettabyte)?

$1\ ZB = 1000\ EB = 1000^2\ PB = 1000^3\ TB = 1000^4\ GB$

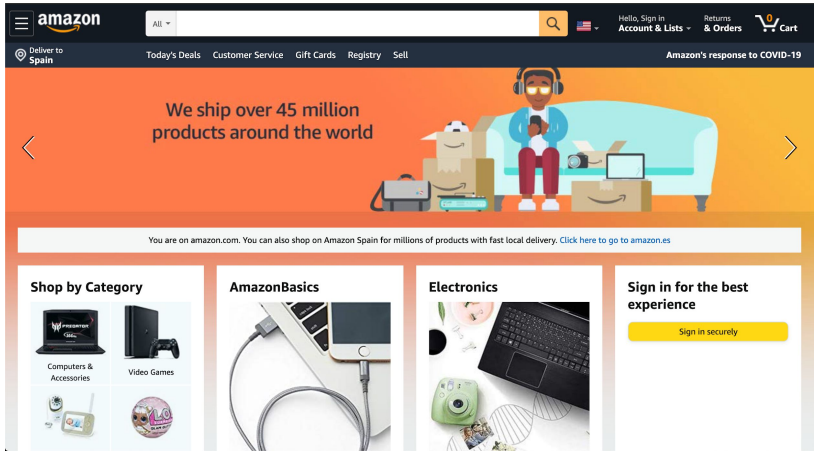
- If we stored 60 ZB in regular blu-ray discs, they would weigh as much as 838 Nimitz-class aircraft carriers.

Some examples of real Big Data applications

- Nowadays, Big Data techniques are being used by many:
  - Large corporations
  - Public services
  - Research institutions
  - Innovative start-ups
- Most people acknowledge the benefits of Big Data for customer management and marketing, but there are many more successful applications.

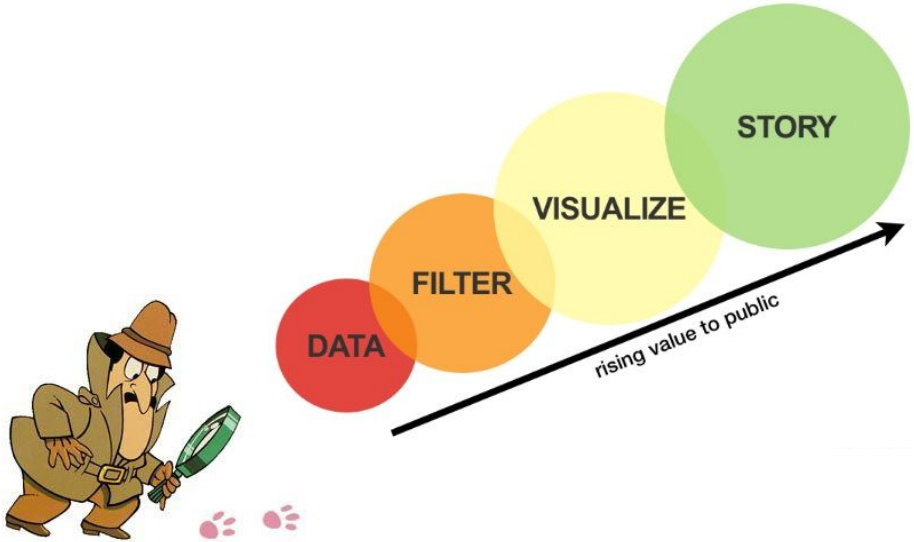
Some examples of real Big Data applications

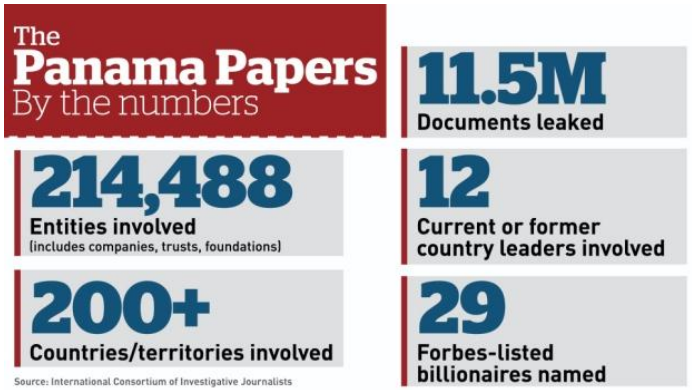
Recommendation systems



Some examples of real Big Data applications

Data driven journalism





The impact of Big Data

Penny Pritzker, US secretary of commerce in a conference at the MIT (march 2014):

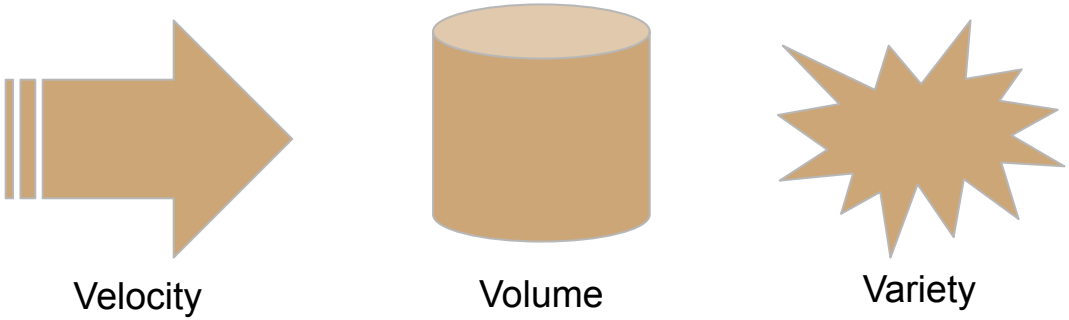
- “Data analysis is the new fuel for American economy”
- Citing a report by McKinsey & Co.: “If open data were available for these main seven sectors: electricity, petroleum, gas, education, transportation, health-care and finances, that could help to unlock up to three trillions dollars”.

A definition

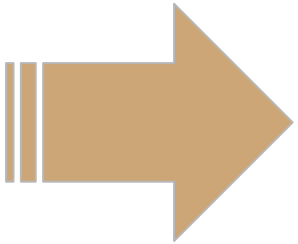
Edd Dumbill (O'Reilly Media):

- 'Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.'

The three V's of Big Data



The three V's of Big Data



Velocity

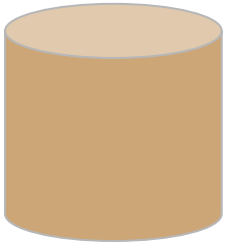
Information is generated faster than it can be analyzed:

- Speed of networks resources do not grow as fast as data volume

What we need:

- Faster stream processing and/or selective storing techniques

The three V's of Big Data



Volume

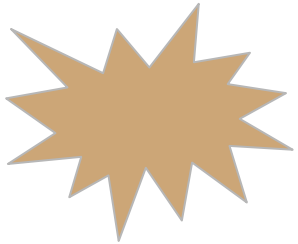
Data volume grows faster than computational resources:

- Volume x10 every 5 years
- Raw CPU power is doubled every 18 months (Moore's Law)

What we need:

- New technologies that store and manage data more efficiently

The three V's of Big Data



Data sources are increasingly heterogeneous:

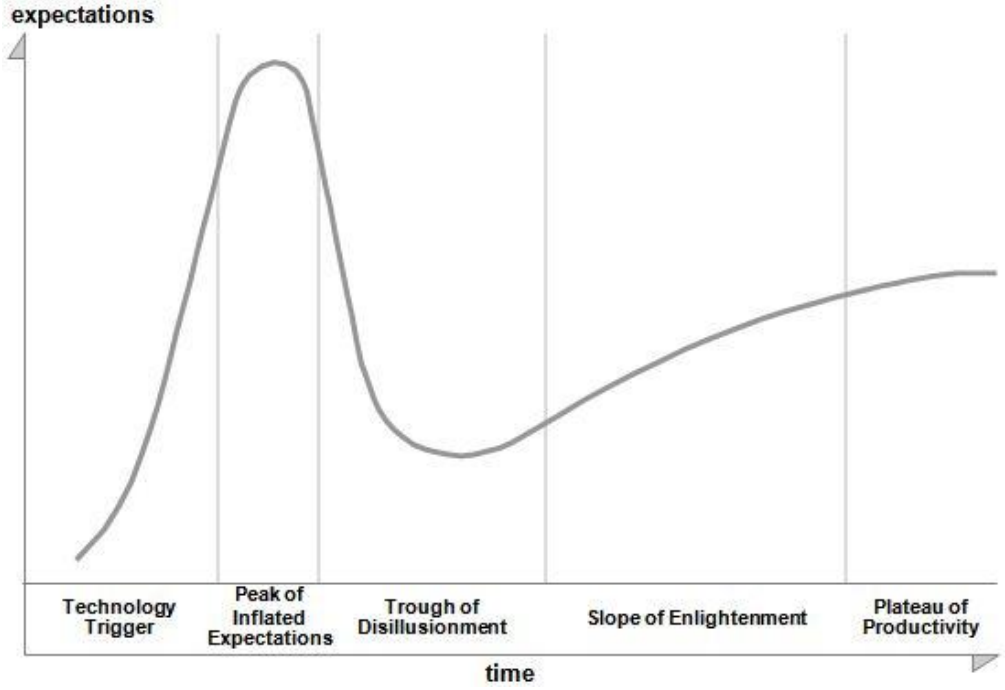
- Multiple-structured or semi-structured data
- Complicated to fit into a classic relational model

What we need:

- Flexible data representation models
- Data storing and processing tools optimized for these new models

Big Data and the Gartner hype cycle

- In 2013, a Gartner article claimed that Big Data was entering the “Trough of Disillusionment”
- Nowadays we should be in the “Plateau of Productivity”, but it depends on multiple factors (region, industry, ...)



http://blogs.gartner.com/svetlana-sicular/big-data-is-falling-into-the-trough-of-disillusionment/

Big Data ~~is dead~~ evolved

BIG DATA IS DEAD

MotherDuck

2023/02/07

BY JORDAN TIGANI

For more than a decade now, the fact that people have a hard time gaining actionable insights from their data has been blamed on its size. "Your data is too big for your puny systems," was the diagnosis, and the cure was to buy some new fancy technology that can handle massive scale. Of course, after the Big Data task force purchased all new tooling and migrated from Legacy systems, people found that they still were having trouble making sense of their data. They also may have noticed, if they were really paying attention, that data size wasn't really the problem at all.

The world in 2023 looks different from when the Big Data alarm bells started going off. The data cataclysm that had been predicted hasn't come to pass. Data sizes may have gotten marginally larger, but hardware has gotten bigger at an even faster rate. Vendors are still pushing their ability to scale, but practitioners are starting to wonder how any of that relates to their real world problems.

https://motherduck.com/blog/big-data-is-dead/

Big Data ~~is dead~~ evolved

- The obligatory intro
  - Traditional data management techniques not able to process increasing amounts of data
- Most people don’t have that much data
  - Data sizes follow a power-law distribution: largest dataset is much larger than most datasets
- Workload sizes are smaller than overall data sizes
  - Small tables queried more frequently, giant tables more selectively
  - Some workloads involve representative samples
- Most data is rarely queried
  - Newer data is more likely to be queried than historical data
- The Big Data frontier keeps receding
  - Machines have now more CPUs and RAM than years ago
- Data is a liability

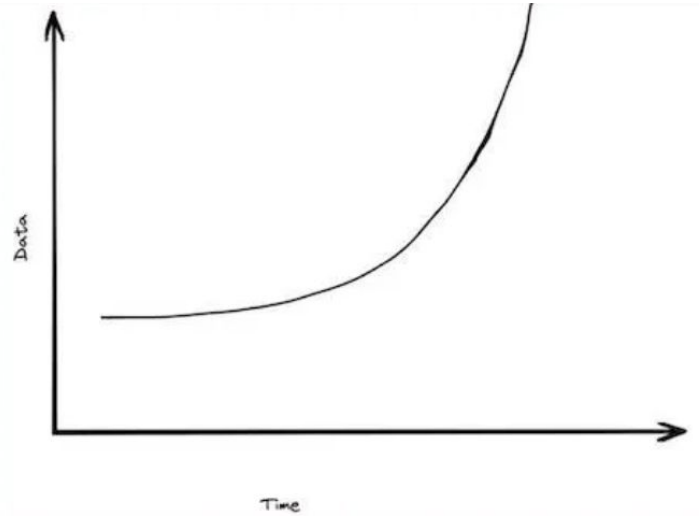
Big Data ~~is dead~~ evolves

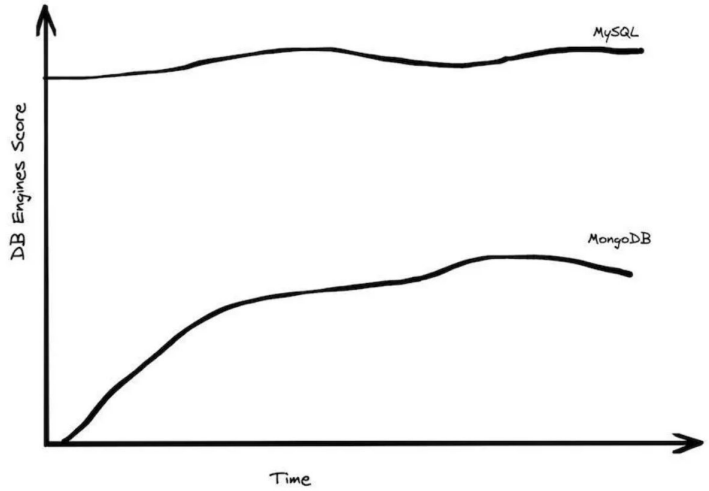
Are you in the Big Data one percent?

- Are you really generating a huge amount of data?
- If so, do you really need to use a huge amount of data at once?
- If so, is the data really too big to fit on one machine?
- If so, are you sure you are not just a data hoarder?
- If so, are you sure wouldn't be better off summarizing?

Big Data ~~is dead~~ evolves

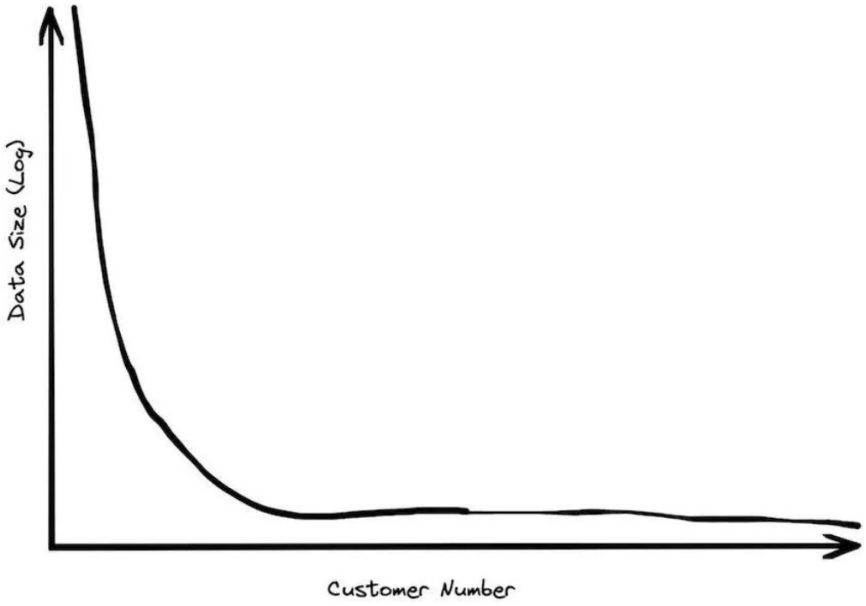
- The obligatory intro
  - Traditional data management techniques not able to process increasing amounts of data





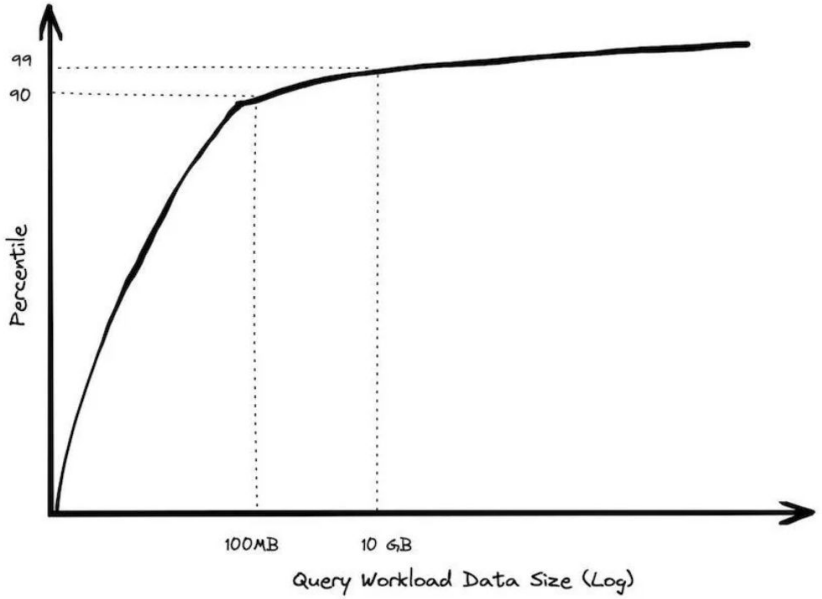
Big Data ~~is dead~~ evolves

- Most people don't have that much data
  - Data sizes follow a power-law distribution: largest dataset is much larger than most datasets



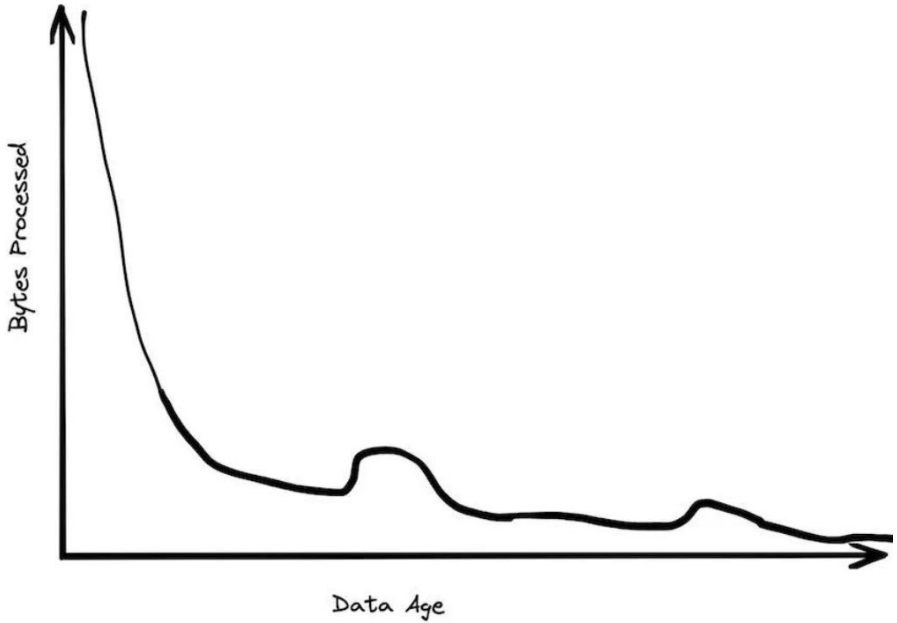
Big Data ~~is dead~~ evolves

- Workload sizes are smaller than overall data sizes
  - Small tables queried more frequently, giant tables more selectively
  - Some workloads involve representative samples



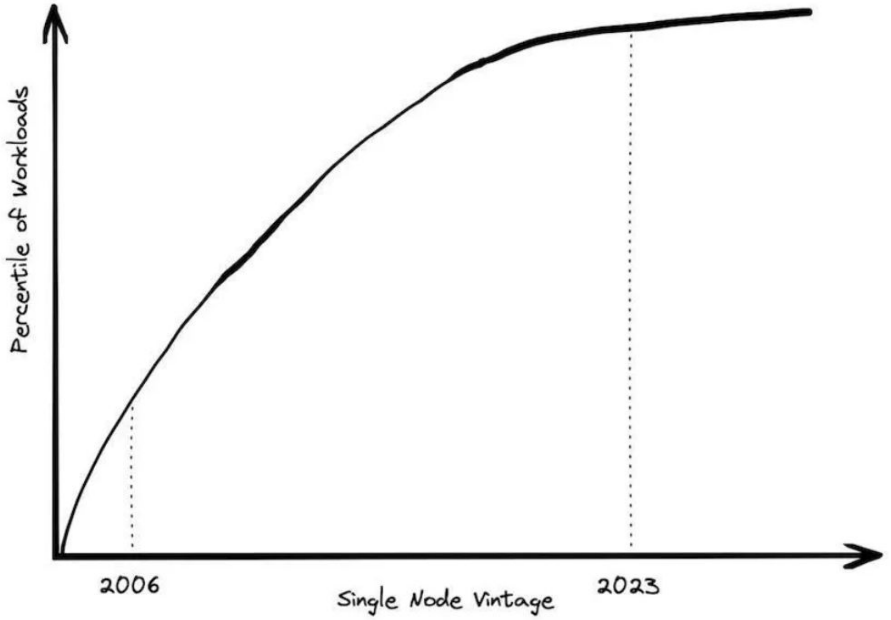
Big Data ~~is dead~~ evolves

- Most data is rarely queried
  - Newer data is more likely to be queried than historical data



Big Data ~~is dead~~ evolves

- The Big Data frontier keeps receding
  - Machines have now more CPUs and RAM than years ago



Big Data ~~is dead~~ evolves

- Data is a liability
  - Cost of keeping data around is higher than cost to store the physical bytes
  - Regulations (GDPR)
  - Data degradation (‘bit rot’)

Demo: OLTP vs OLAP

- LUBM data (modelling the university domain)
- Data scaling factors: 1, 10, 100
- SQLite (OLTP) - DuckDB (OLAP)
- Run SQL query asking for title of a publication (containing ‘magic’) and email of the author (graduate student)

Then, what is Big Data?

Big data is not...

- ... a replacement for traditional databases
- ... a replacement for statistical inference.
- ... a replacement for standard business intelligence procedures.

Big Data tries to address new challenges where these (and other) techniques fall short. Situations where data is being produced...

- ... too fast (velocity).
- ... in an extremely large amount (volume).
- ... from many heterogeneous sources (variety).

Big Data in three “easy” steps…

Data engineering

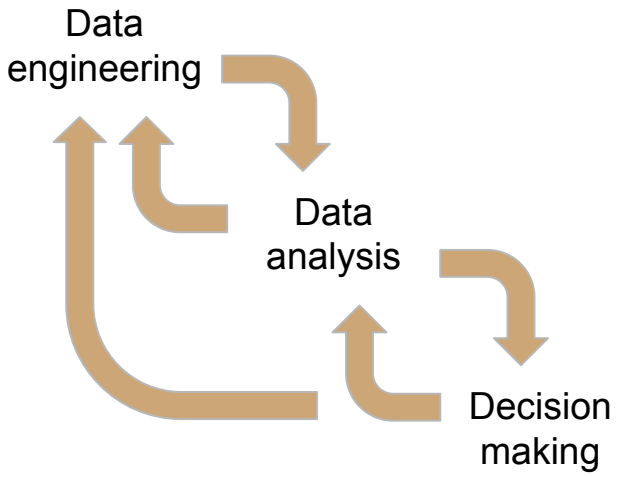
- Storing, managing and operating with data

Data analysis/modeling

- Extracting knowledge

Data-driven decision making

- Putting the knowledge to good use



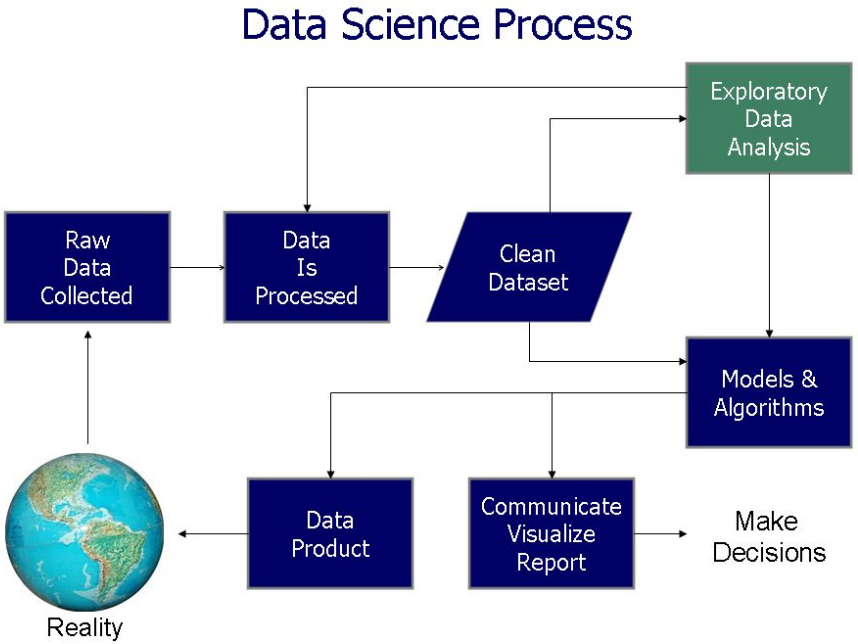
Big Data and Data Science

Data science is the process of extracting knowledge or insights from data in various forms, either structured or unstructured.

- Based on the scientific method
- Can be seen as a continuation/combination of data analysis fields such as statistics, data mining, machine learning, etc.

When a data science problem cannot be addressed using traditional data storing/processing/analyzing techniques, then it also becomes a Big Data problem.

Data Science



A few Data Science use cases

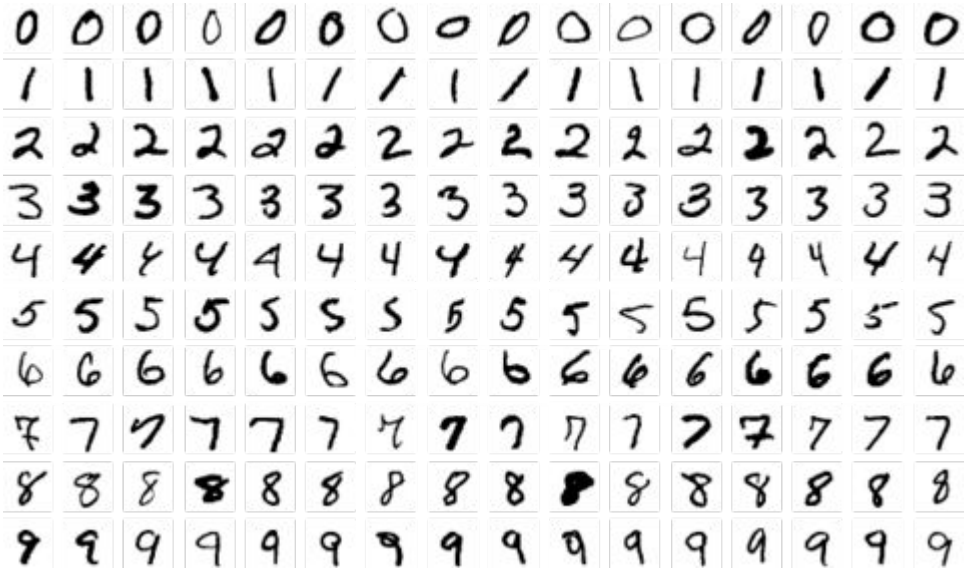
Data science can be applied to many different fields and problems, from decision making in business scenarios to the scientific domain or even more recreational approaches.

Classical examples:

- The Wal-Mart “beer & diapers” case
- Moneyball (The real story behind that 2011 film with Brad Pitt on it)
- The Netflix Prize

A few Data Science use cases

Pattern recognition (MNIST)



A few Data Science use cases

Even create art

Final remarks/reminders

- What is Big Data? → Remember the three Vs
- When are we facing a Big Data problem? → When traditional techniques are not enough
- What are the three “steps” of Big Data?
  - **Data engineering**
  - Data analysis/modeling
  - Decision making

In this course we will cover mostly data engineering.
Data analysis is addressed in depth in other courses.