UVA HPC & BIG DATA COURSE

Introduction to Big Data

Adam Belloum

Content

- General Introduction
- Definitions
- Data Analytics
- Solutions for Big Data Analytics
- The Network (Internet)
- When to consider BigData solution
- Scientific e-infrastructure – some challenges to overcome

Jim Gray Vision in 2007

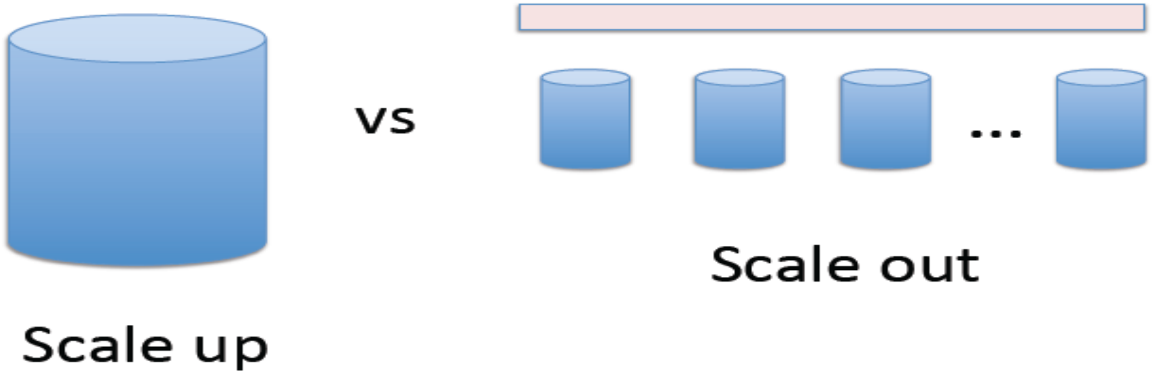
- " We have to **do better at producing tools** to support the whole research cycle—from data capture and data curation to data analysis and data visualization. Today, the **tools** for capturing data both at the mega-scale and at the milli-scale are just **dreadful**. After you have captured the data, you need to curate it before you can start doing any kind of data analysis, and we lack good tools for both data curation and data analysis."

- "Then comes the publication of the results of your research, and the published literature is just the tip of the data iceberg. By this I mean that people collect a lot of data and then reduce this down to some number of column inches in Science or Nature—or 10 pages if it is a computer science person writing. So what I mean by data iceberg is that there is **a lot of data** that is **collected** but not curated or published in any systematic way."

Based on the transcript of a talk given by Jim Gray
to the NRC-CSTB1 in Mountain View, CA, on January 11, 2007

How to deal with Big Data
Advice From Jim Gray

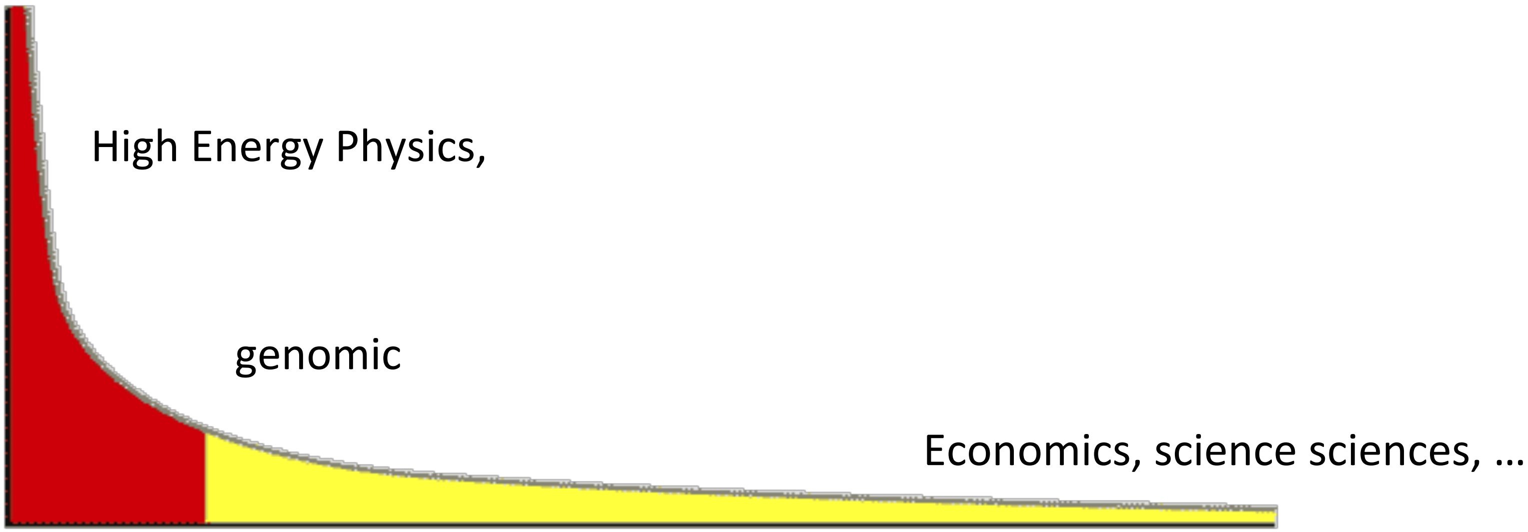
1. Analysing Big data requires scale-out solutions not scale-up solutions
2. **Move** the analysis to the data.
3. Work with scientists to find the most common "**20 queries**" and make them fast.
4. Go from "**working to working.**"



Source: Robert Grossman, Collin Bennec University of Chicago Open Data Group

Data keep on growing

- Google processes 20 PB **a day** (2008)
- Wayback Machine has 3 PB + 100 TB/**month** (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/**day** (4/2009)
- eBay has 6.5 PB of user data + 50 TB/**day** (5/2009)
- CERN's Large Hydron Collider (LHC) generates 15 PB/**year**



Collectively "Long Tail" science is generating **a lot of data**

Data is Big If It is Measured in MW

- A good sweet spot for a data center is 15 MW
- Facebook's leased data centers are typically between 2.5 MW and 6.0 MW.
- Facebook's Pineville data center is 30 MW
- Google's computing infrastructure uses 260 MW

Big data was big news in 2012

- The Harvard Business Review talks about it as "*The Management Revolution*".

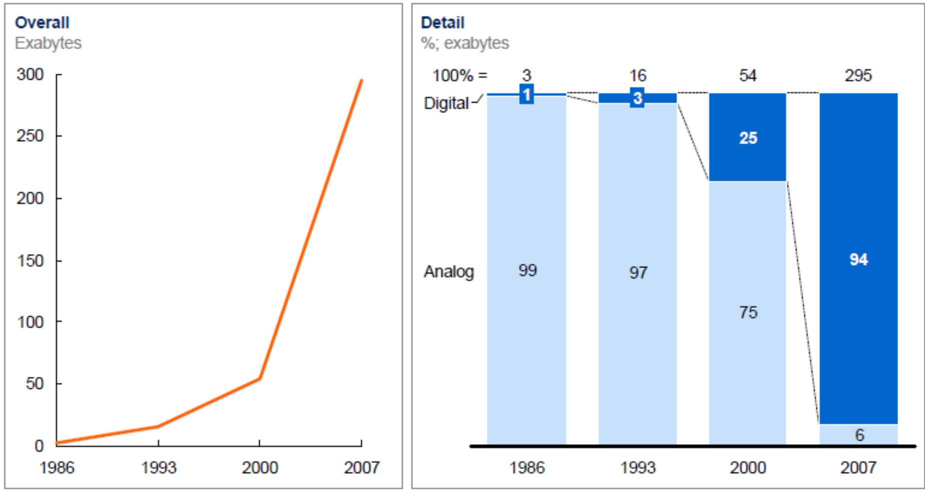
- The Wall Street Journal

  "*Meet the New Big Data*",

  "*Big Data is on the Rise, Bringing Big Questions*".

Data storage has grown significantly, shifting markedly from analog to digital after 2000

Global installed, optimally compressed, storage

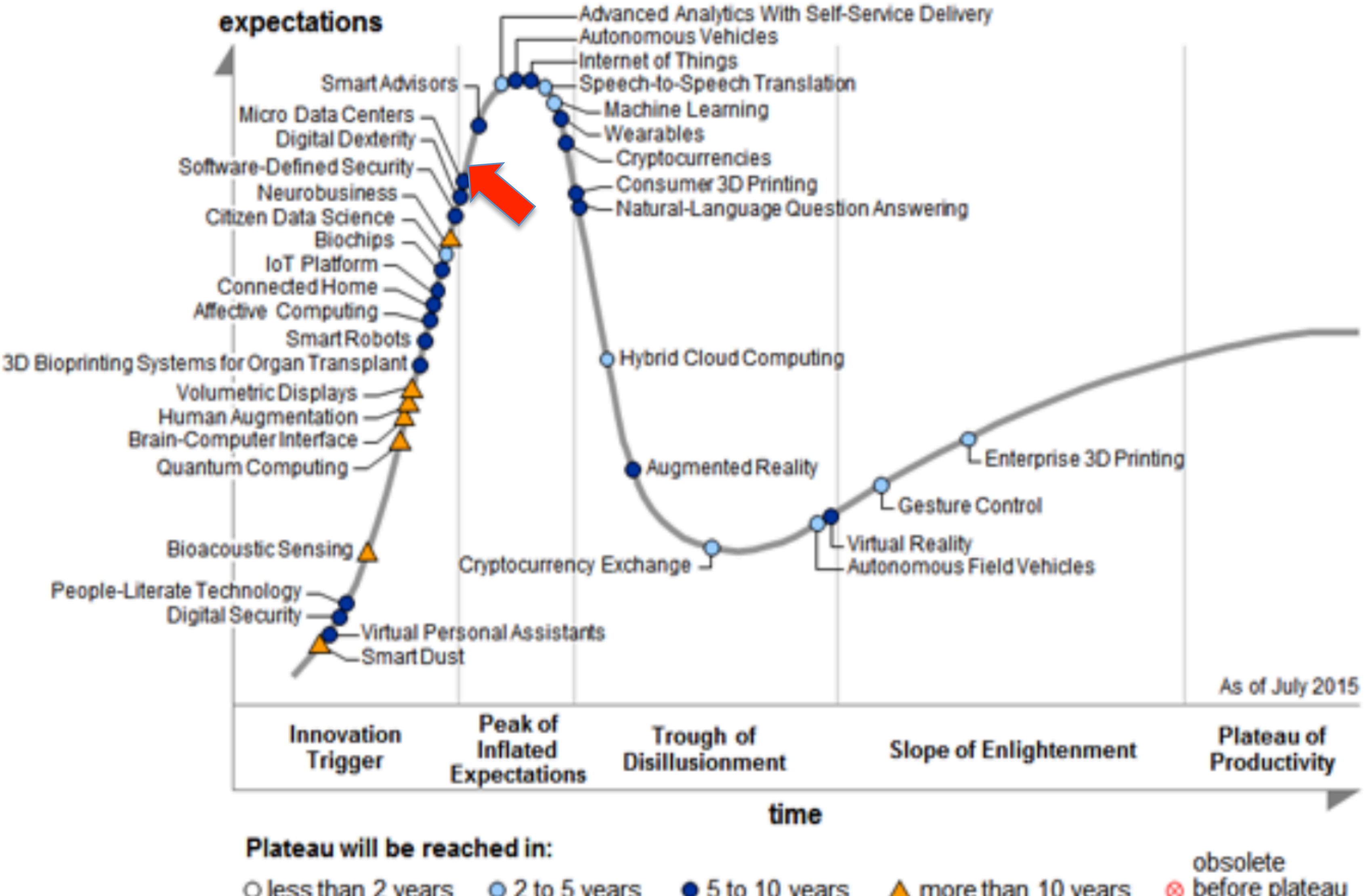


NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

BigData is the new hype

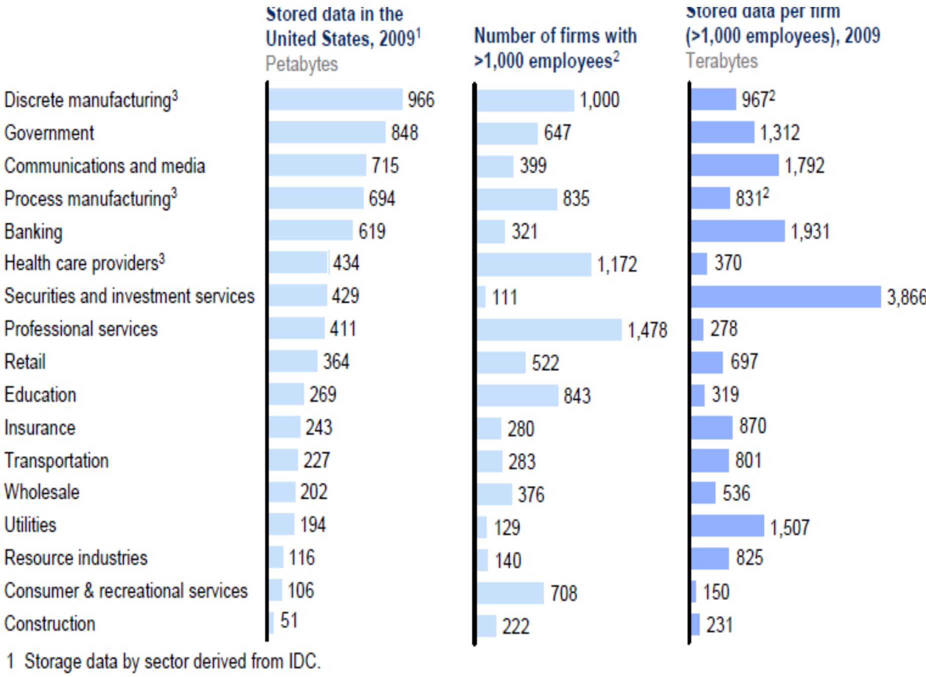
Figure 1. Hype Cycle for Emerging Technologies, 2015



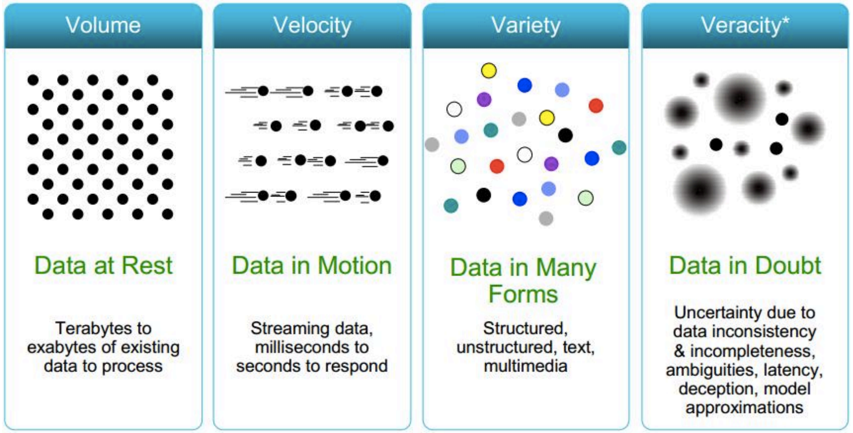
Where Big Data Comes From?

- Big Data is not Specific application type, but rather a trend –or even a collection of Trends- napping multiple application types

- Data growing in multiple ways
  - More data (**volume** of data)
  - More Type of data (**variety** of data)
  - Faster Ingest of data (**velocity** of data)
  - More Accessibility of data (internet, instruments , …)
  - Data Growth and availability exceeds organization ability to make intelligent decision based on it



1 Storage data by sector derived from IDC.



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Addison Snell CEO. Intersect360, Research

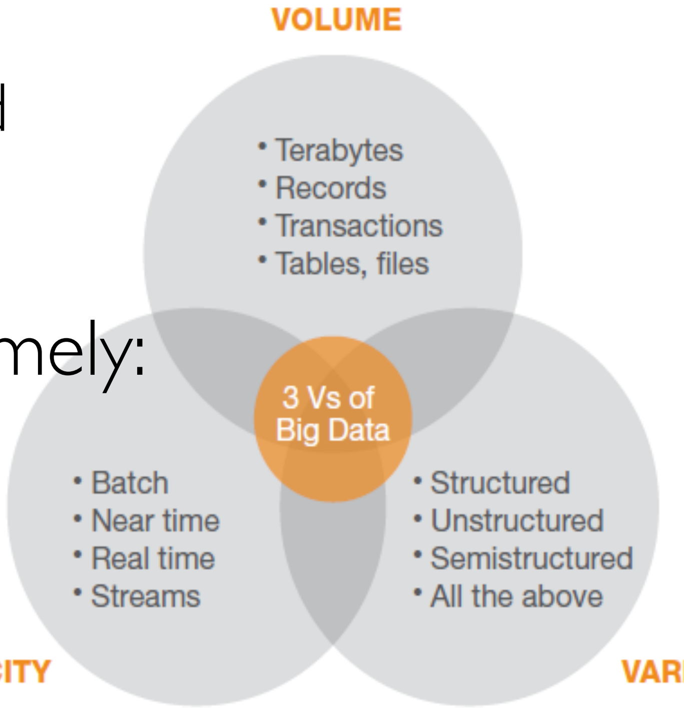
content

- General Introduction
- Definitions
- Data Analytics
- Solutions for Big Data Analytics
- The Network (Internet)
- When to consider BigData solution
- Scientific e-infrastructure – some challenges to overcome

How do We Define Big Data

- Big in Big Data refers to:
  - Big size is the primary definition.
  - Big complexity rather than big volume. it can be small and not all large datasets are big data
  - size matters... but so does accessibility, interoperability and reusability.

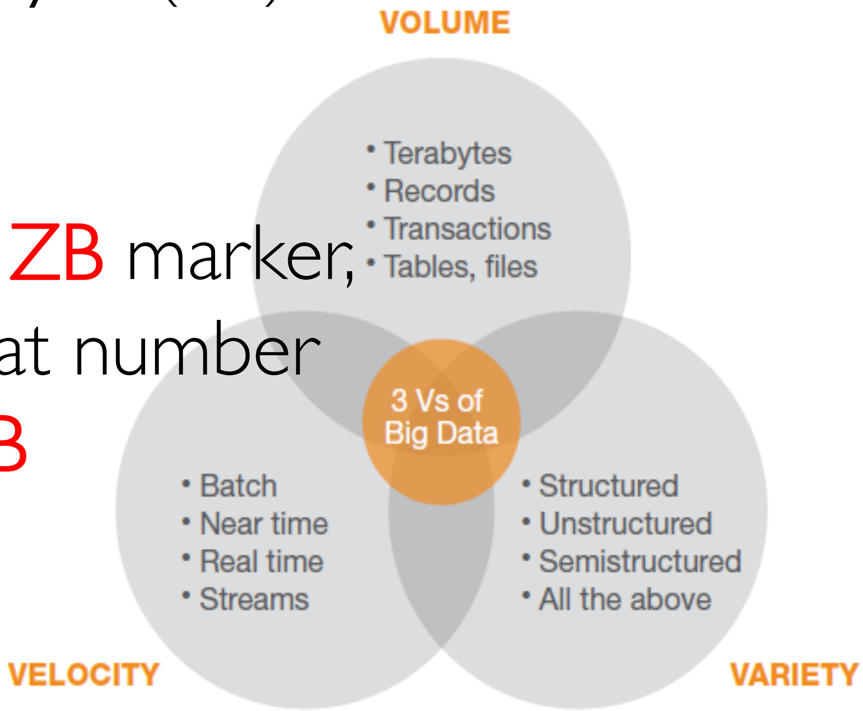
- define Big Data using 3 Vs; namely:
  - volume, variety, velocity



volume, variety, and velocity

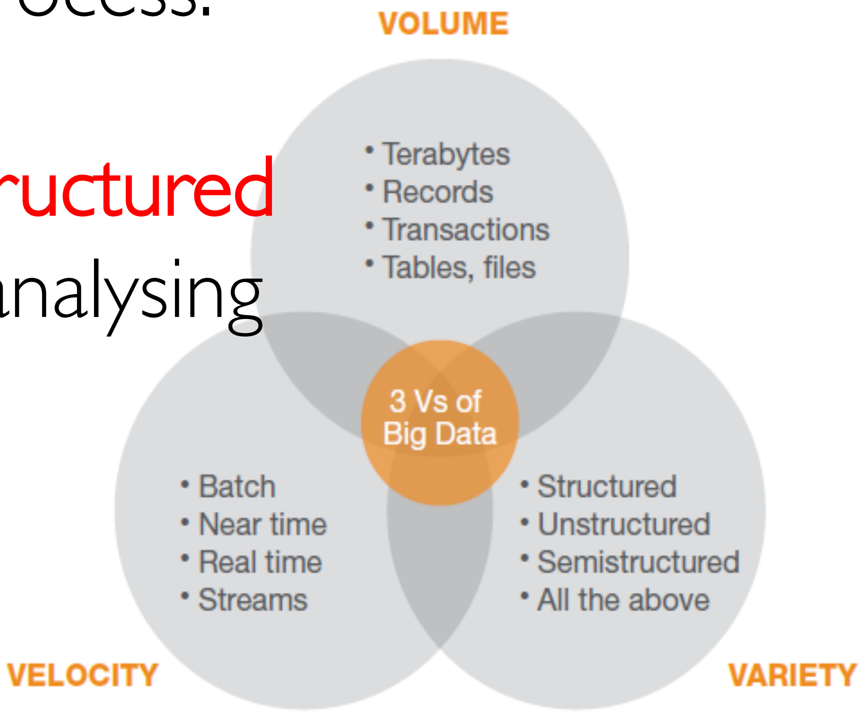
- Aggregation that used to be measured in petabytes (PB) is now referenced by a term: zettabytes (ZB).
  - A **zettabyte** is a **trillion gigabytes** (GB)
  - or a **billion terabytes**

- in 2010, we crossed the 1ZB marker, and at the end of 2011 that number was estimated to be 1.8ZB



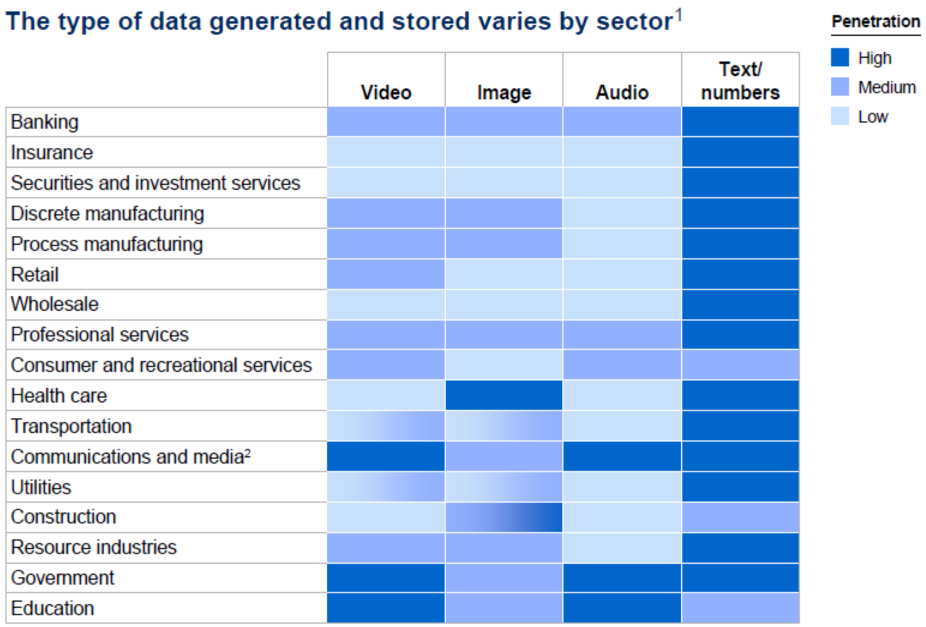
volume, variety, and velocity

- The variety characteristic of Big Data is really about trying to **capture all** of the data that pertains to our decision-making process.
- Making sense out of unstructured data, such as opinion, or analysing images.



volume, variety, and velocity
(Type of Data)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once



The type of data generated and stored varies by sector[1]

Penetration

■ High
▨ Medium
□ Low

	Video	Image	Audio	Text/ numbers
Banking	▨	▨	▨	■
Insurance	□	□	□	■
Securities and investment services	□	□	□	■
Discrete manufacturing	▨	▨	□	■
Process manufacturing	▨	▨	□	■
Retail	▨	□	□	■
Wholesale	□	□	□	■
Professional services	▨	▨	▨	■
Consumer and recreational services	▨	□	▨	▨
Health care	□	■	□	■
Transportation	□ ▨	□ ▨	□	■
Communications and media[2]	■	▨	■	■
Utilities	□ ▨	□ ▨	□	■
Construction	□	▨ ■	□	▨
Resource industries	▨	▨	□	■
Government	■	▨	■	■
Education	■	▨	■	▨

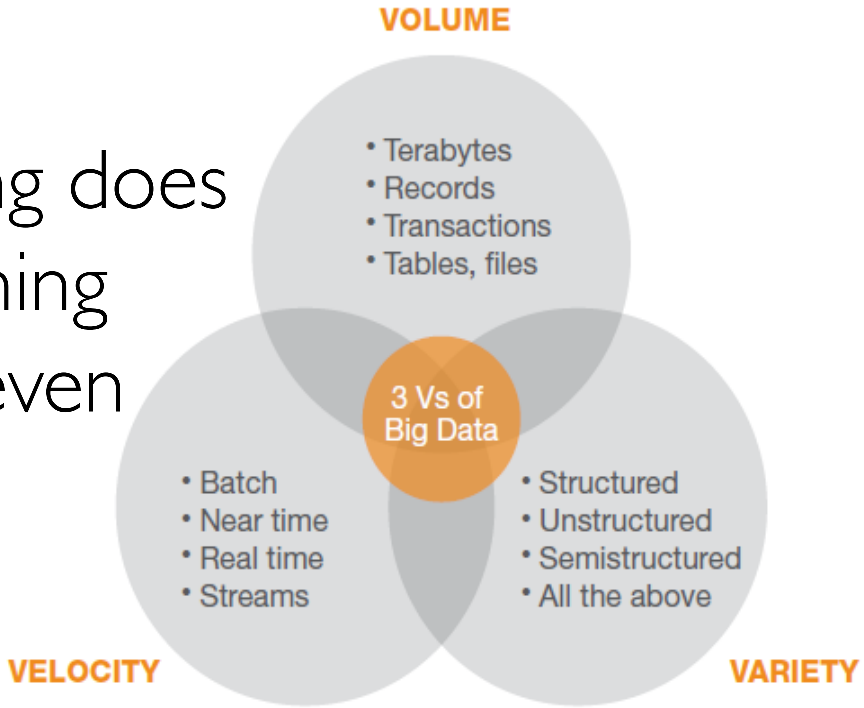
[1] We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
[2] Video and audio are high in some subsectors.

SOURCE: McKinsey Global Institute analysis

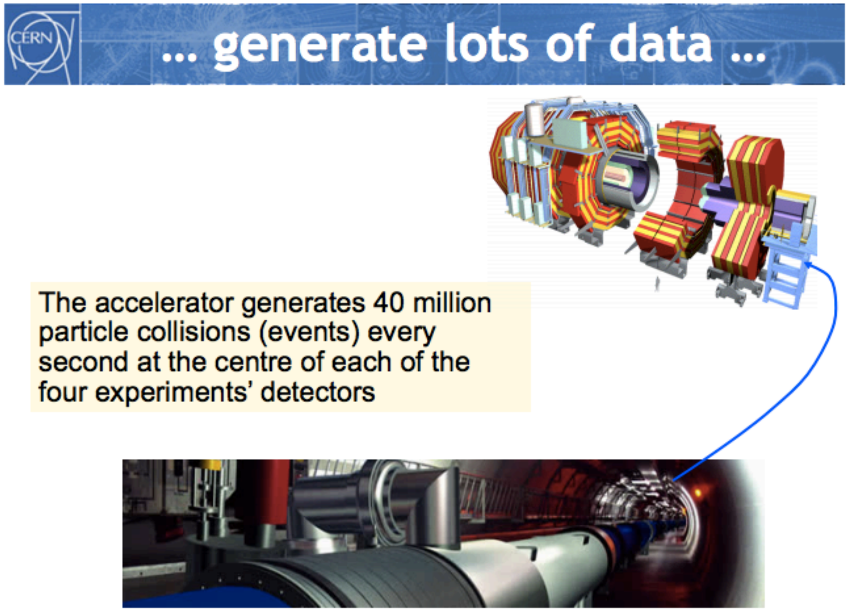
volume, variety, and velocity

- velocity is the rate at which data arrives at the enterprise and is processed or well understood

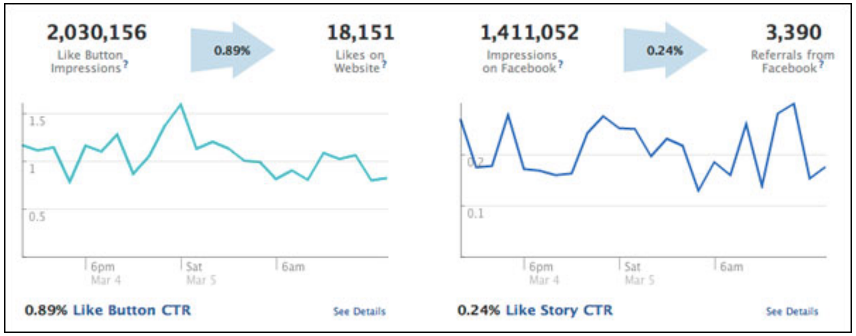
- In other terms "How long does it take you to do something about it or know it has even arrived?"



volume, variety, and velocity



The accelerator generates 40 million particle collisions (events) every second at the centre of each of the four experiments' detectors



Today, it is possible using real-time analytics to optimize Like buttons across both website and on Facebook.

FaceBook use anonymised data to show the number of times people:

- saw Like buttons,
- clicked Like buttons,
- saw Like stories on Facebook,
- and clicked Like stories to visit a given website.