

Big Data Project

Introduction to Big Data

Jonathan Fürst



2013: What is Big Data?

Big Data

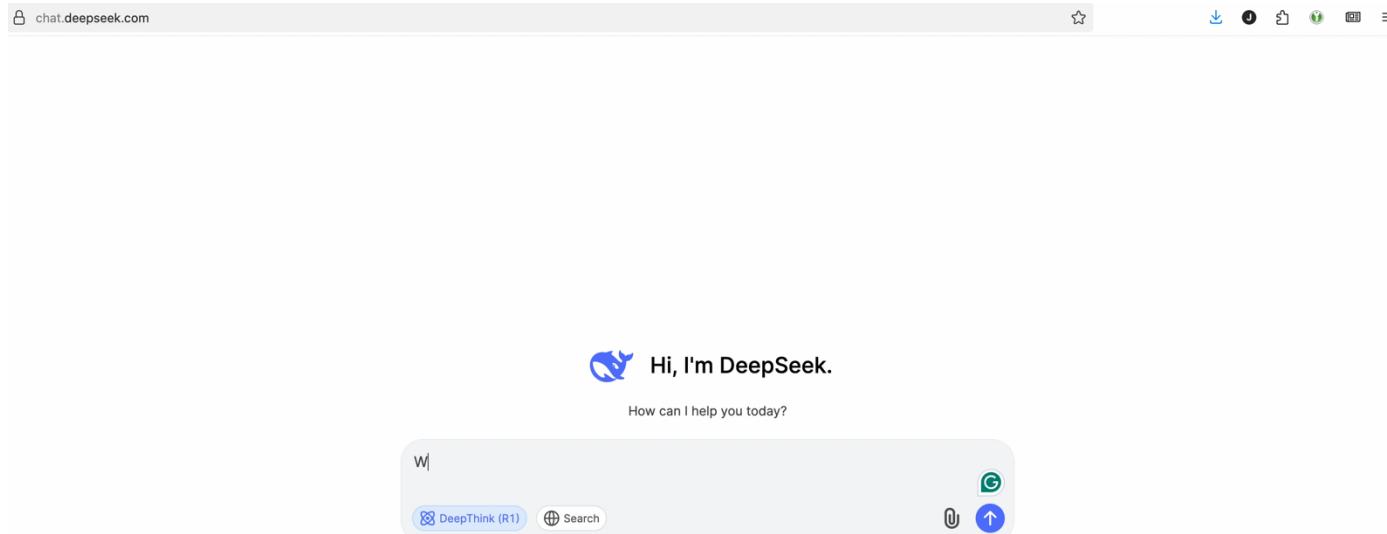


“Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...”

-- Dan Arielym, Duke University

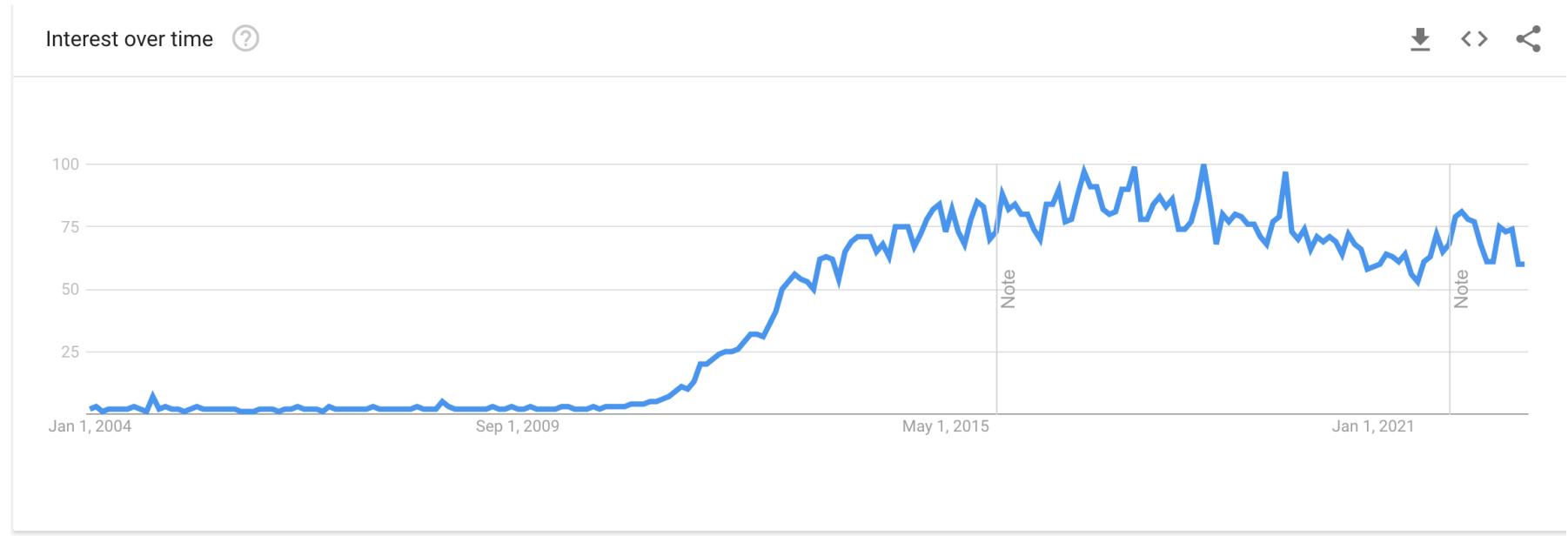
Today: What is Big Data?

Big Data



Buzz

Big Data



<https://trends.google.com/trends/explore?date=all&q=%2Fm%2F0bs2j8q>

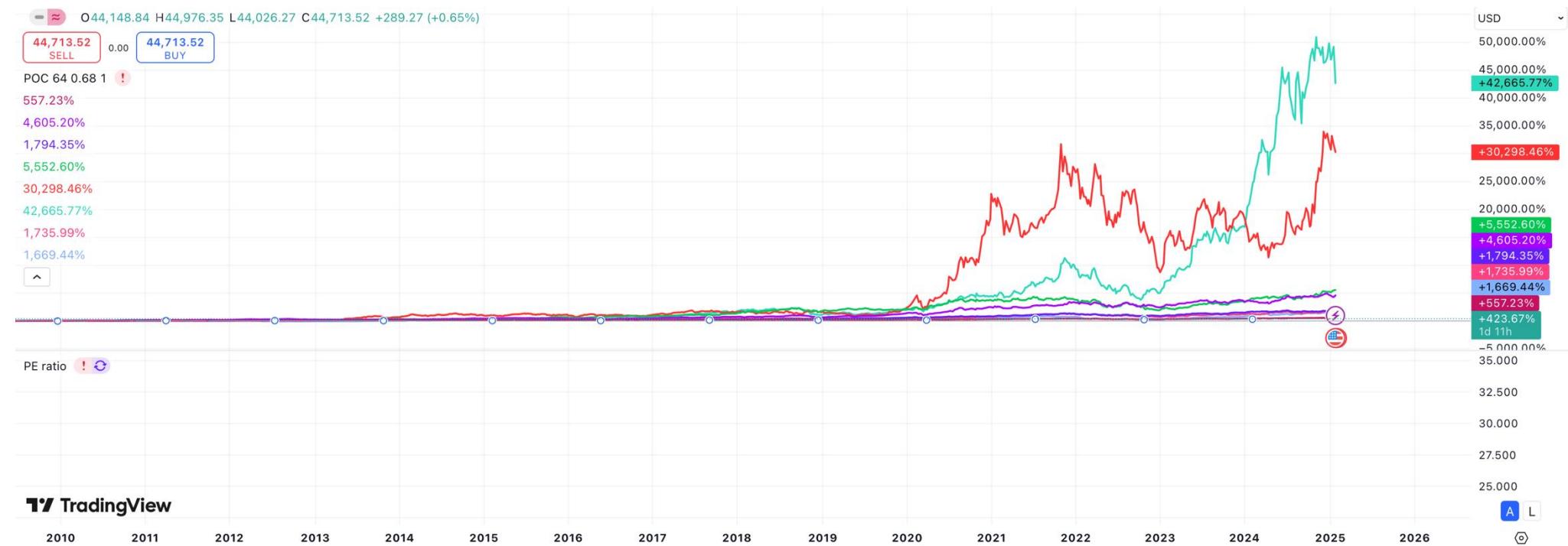
“Global Spending on Big Data and Analytics Solutions Will Reach **\$215.7 Billion** in 2021,
According to a New IDC Spending Guide” IDC, 2021

FAANG Stock Performance: 2010 to 2020



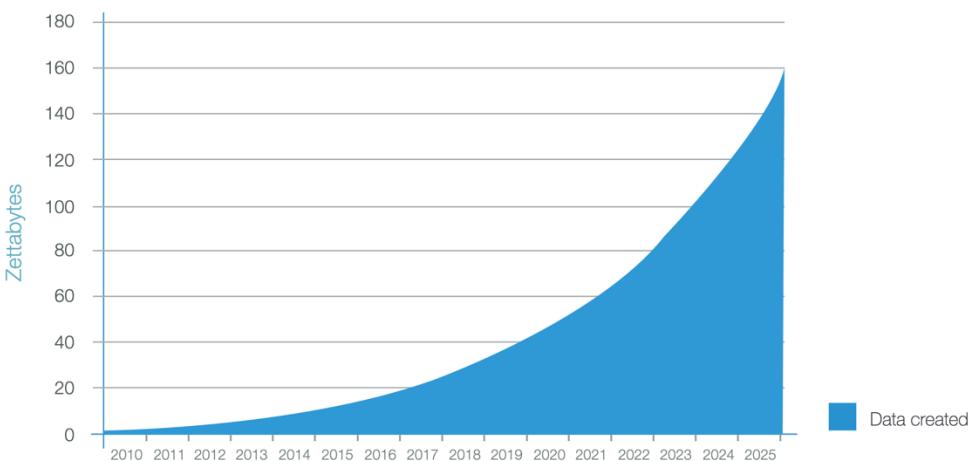
Source: CapIQ, FAANG equal-weight indexed to 1/1/2010

Magnificent Seven (Apple, Microsoft, Google parent Alphabet, Amazon.com, Nvidia, Meta Platforms and Tesla) stock performance vs. S&P 500

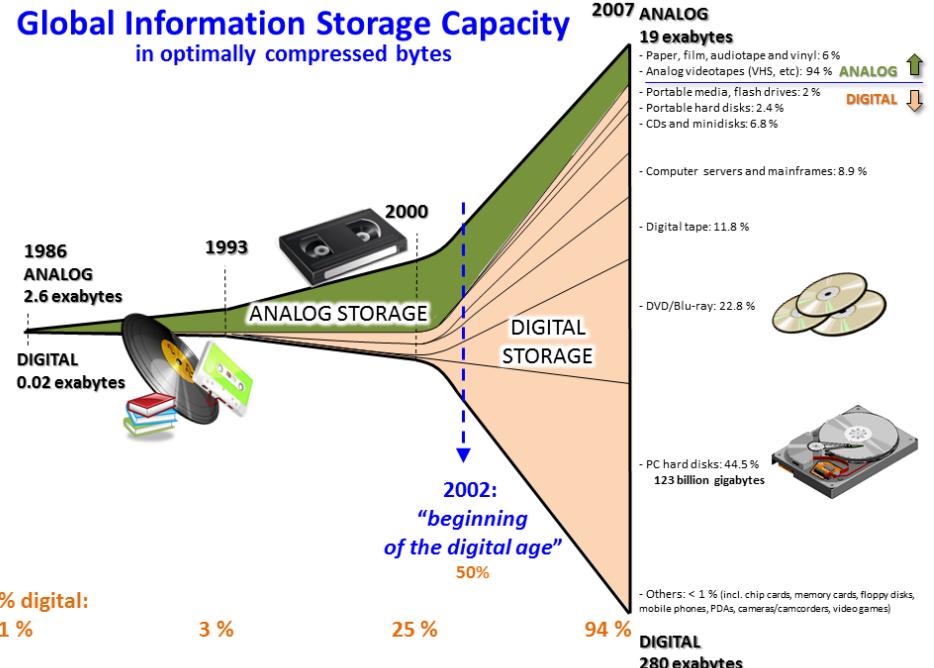


Volume

Big Data



<https://www.idc.com/getdoc.jsp?containerId=prUS48165721>



<https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>

Volume

Big Data

- Library of Congress in Washington
- Contains more than ~~36 38 39 40~~ 51 million books
- ca. 20 TB



<https://www.loc.gov/about/general-information>

Quiz:

- How long do you need to count, if you count one book per second?

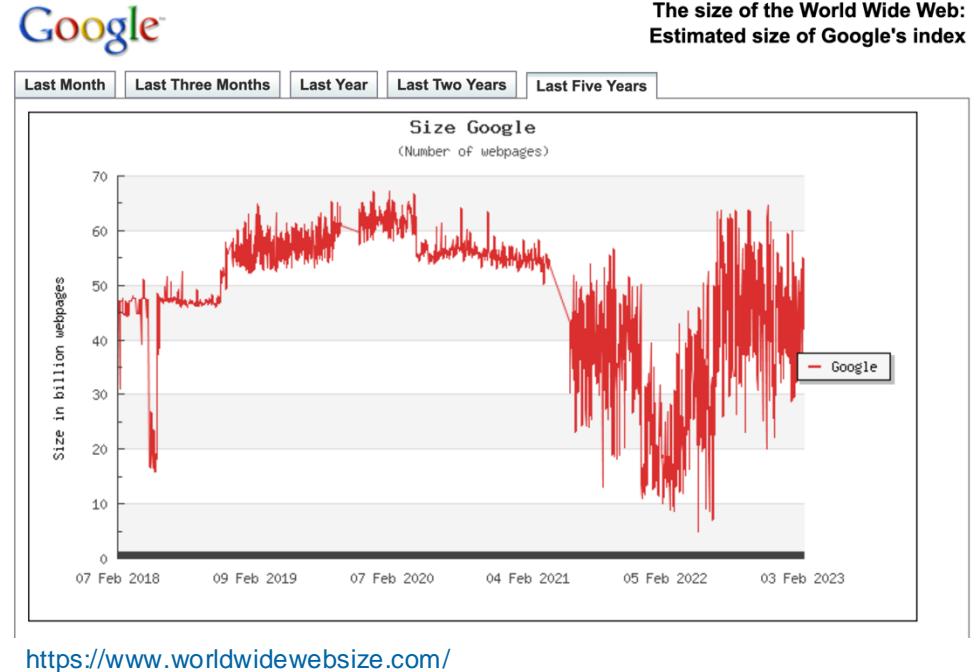
Volume

Big Data

- Google Web Index

Quiz:

- How long do you need to count now?



“Over 500 hours of video are uploaded to YouTube every minute. That equates to an astonishing 30,000 hours of freshly uploaded content by the hour.”

Source: <https://www.zippia.com/answers/how-many-hours-of-video-are-uploaded-to-youtube-every-minute/>



Shots

TREATMENTS

Big Data Coming In Faster Than Biomedical Researchers Can Process It

November 28, 2016 · 2:04 PM ET

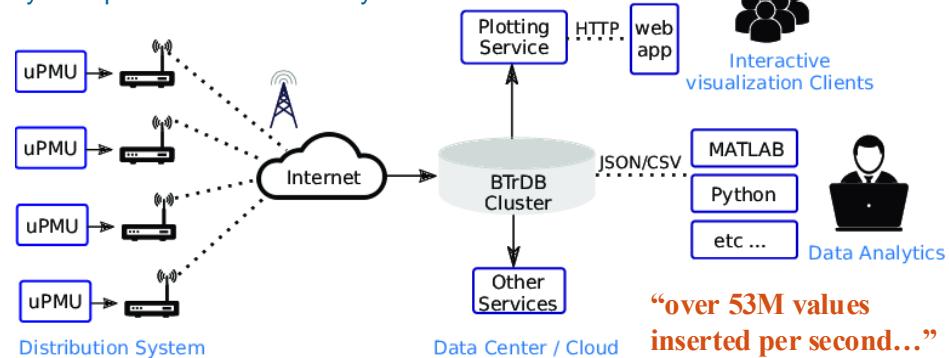
Heard on [All Things Considered](#)

RICHARD HARRIS

<https://www.npr.org/sections/health-shots/2016/11/28/503035862/big-data-coming-in-faster-than-biomedical-researchers-can-process-it>



<https://citrис-uc.org/research/project/micro-synchrophasors-distribution-systems/>



"over 53M values inserted per second..."

Ethics & Privacy

Big Data

GLOBE IN BEIJING

China using big data to detain people before crime is committed: report

NATHAN VANDERKLIPPE > INTERNATIONAL CORRESPONDENT

BEIJING

PUBLISHED FEBRUARY 27, 2018

MOTHERBOARD
TECH BY VICE

Leaked Document Shows How Big Companies Buy Credit Card Data on Millions of Americans

Yodlee, America's largest financial data broker, says the data it sells is anonymous. A confidential document obtained by Motherboard shows people could be unmasked in the data.



By Joseph Cox



EU data watchdogs ruling sharpens focus on Facebook, big tech

By Foo Yun Chee



≡ WIRED

JUSTIN SHERMAN IDEAS DEC 19, 2021 8:00 AM

Big Data May Not Know Your Name. But It Knows Everything Else

Data brokers claim that deidentified data on millions of Americans is risk-free. Lawmakers need to know that “anonymity” is an abstraction.

The Vs

Big Data

- **Volume (big):**
 - “Large” amounts of data
- **Velocity (fast):**
 - Streams of data need to be processed fast
- **Variety (different data sources/structure):**
 - Text, images, videos, databases, blogs, social network data
- **Veracity (quality):**
 - Data of different quality
- **Value:**
 - Some data are more valuable than others (customer records vs. product description)



Used to characterize Big Data but this does not mean all of them are needed to consider something “Big Data”.

Traditional Database Management

Big Data

Narrow scope

- A database is created to serve a well-defined purpose

Structured data

- Conceptual/Logical/Physical schema
- Relational model dominates since 80s
- Entity Relationship defines conceptual schema

Close world assumption

- Data as an instance of the schema
- The data which is not part of an instance *does not exist*
 - *Any query on the database returns a value based on the current instance*

Data at rest

- Data is loaded and stored in the database, on disk.

Wide scope

- Data is made available for yet-to-be-defined analysis

Data Variety

- Time series are highly structured; Text is not

Open world assumption

- Data sources might be added or removed
- So any analysis is only valid based on the current state of the big data resource

Data in movement, and at rest

- Data streams complements stored data
- Some data streams are stored, others are not

Take Away Points

Conclusion

- "Big Data" still means different things to different people, however the term is clearer than it was 10 years ago.
- There is lots of buzz around it, but also it represents some profound changes from a **closed-world assumption** with well structured data in a relational model to **Big Data**, where a large amount of data (**Volume**) exists in various formats (**Variety**), produced rapidly (**Velocity**), often without an immediate application pull.
- Lately, in context of IoT, streaming data has also become more important.
- Often we want to make sense of large amounts of data with a short latency and in quick iterations.