HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

Chapter 1
Overview of big data storage and processing

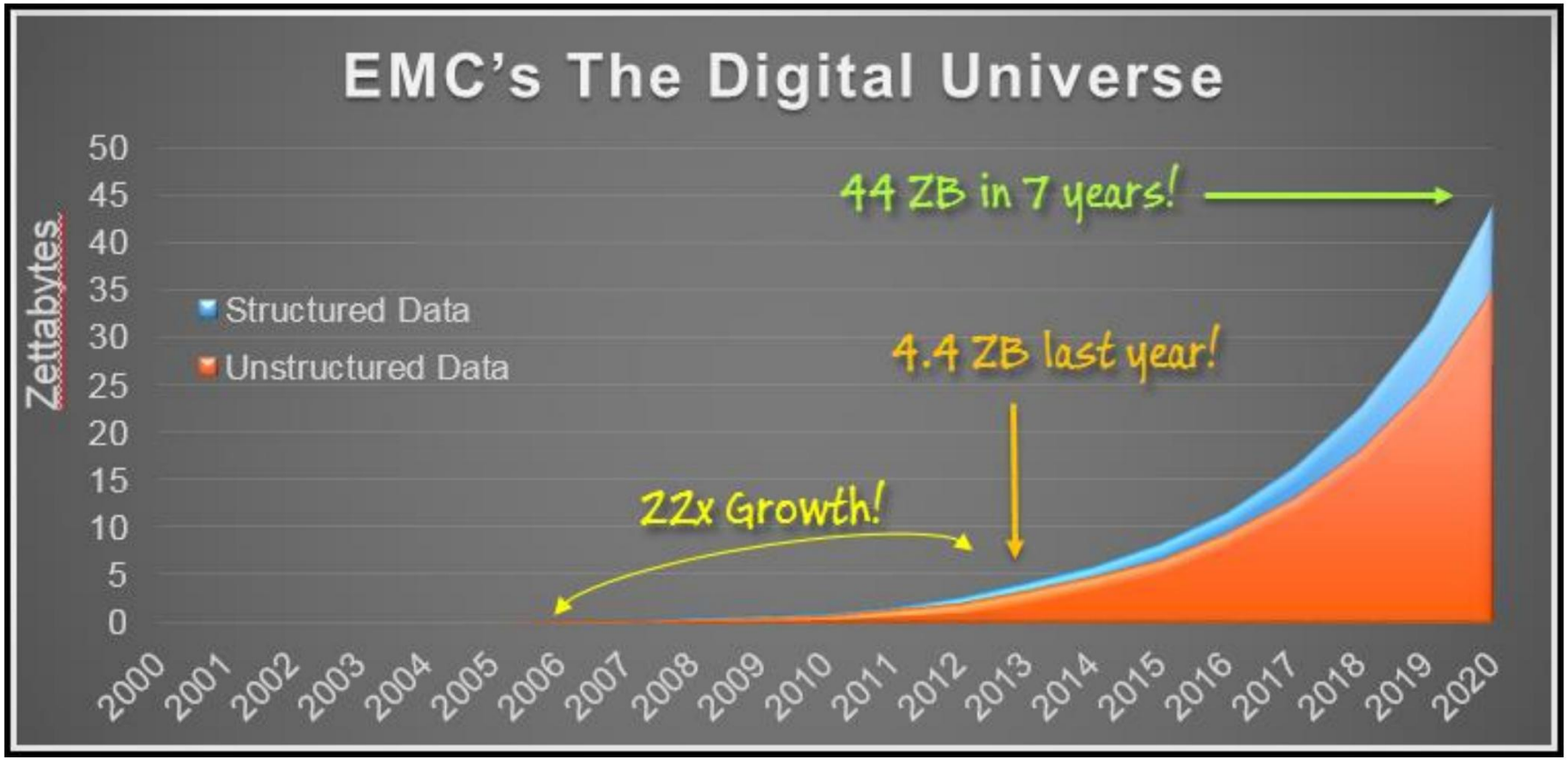
General information about the subject

Course name:	**Big data storage and processing** **(Big data storage and processing)**
Course code: IT4931	
Weight: 3(3-1-0-6)	Theory: 45 lessons ÿ BTL: 15 periods - Experiment: 0 periods

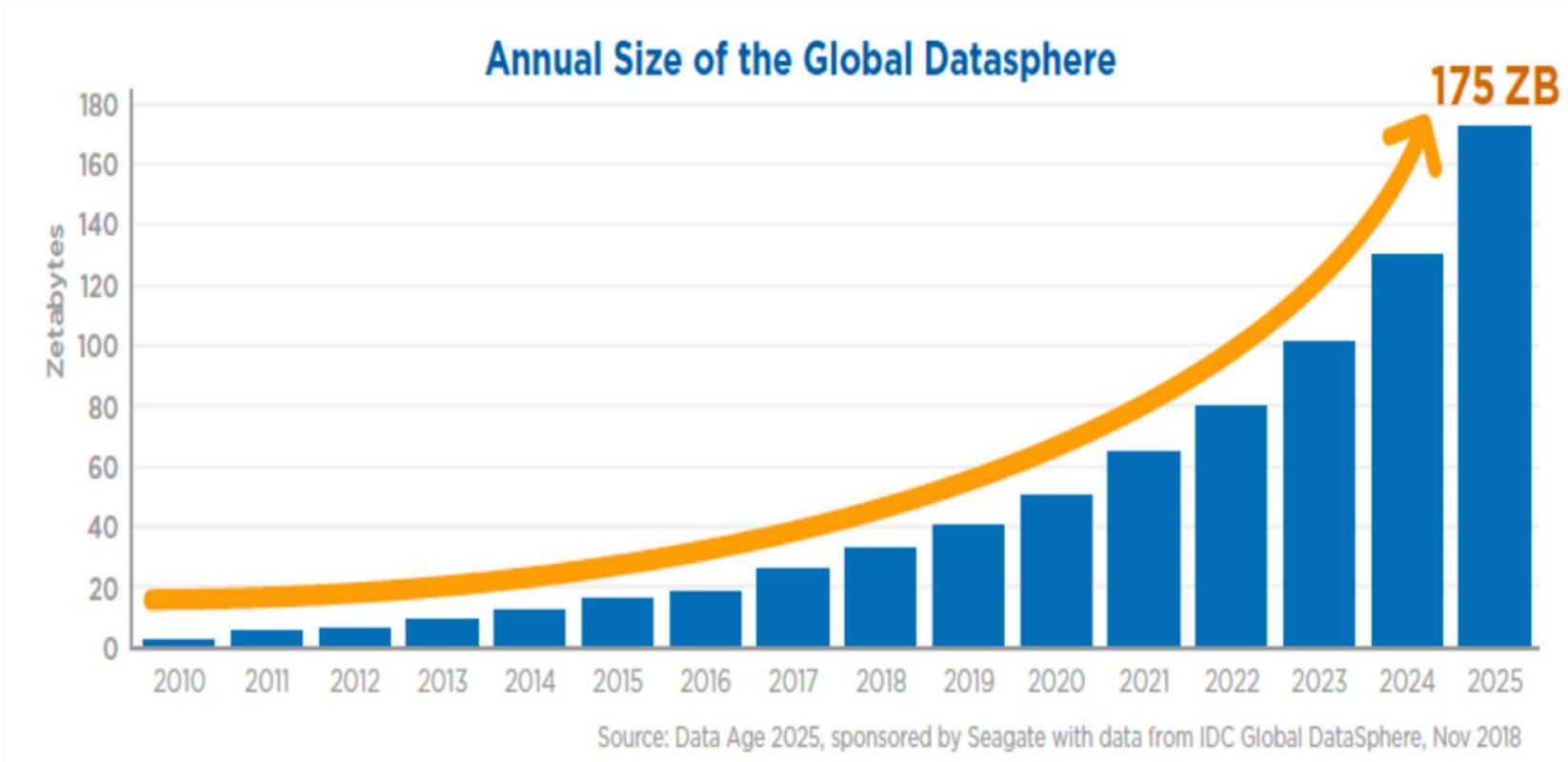
Study outline

STT	Lesson
1	**Overview of big data storage and processing**
2	**Hadoop ecosystem**
3	**Hadoop Distributed File System HDFS**
4	**NoSQL Non-Relational Databases - Part 1** Overview
5	**NoSQL Non-Relational Databases - Part 2** Popular distributed architectures
6	**NoSQL Non-Relational Databases - Part 3** SQL Queries on NoSQL
7	**Distributed messaging system**
8	**Batch Big Data Processing Techniques - Part 1** Map Reduce
9	**Batch Big Data Processing Techniques - Part 2** Apache Spark
10	**Big Data Stream Processing Techniques** Spark Streaming
11	**Big Data Architecture** Lambda architecture
12	**Big Data Analytics** Spark ML

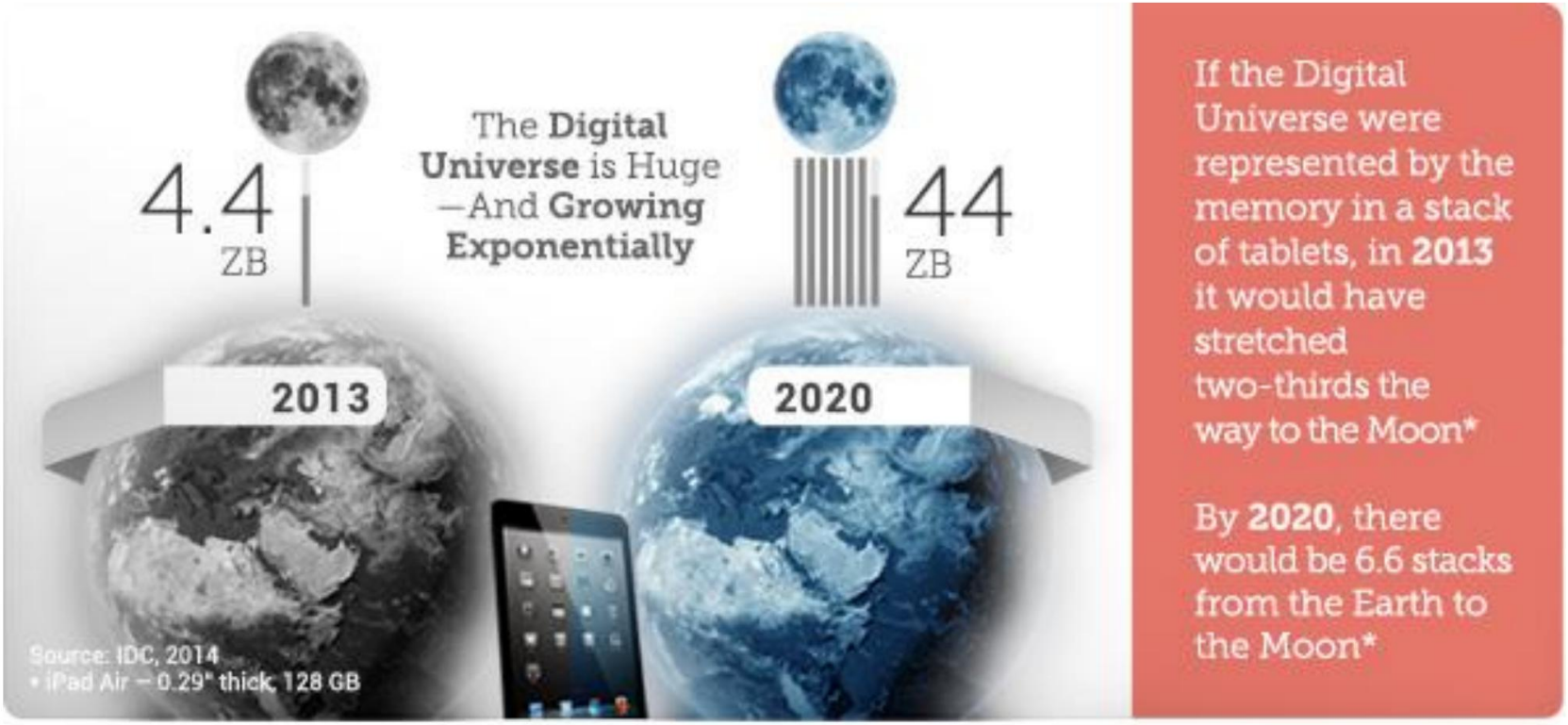
Total data capacity 2020



Total data capacity 2025

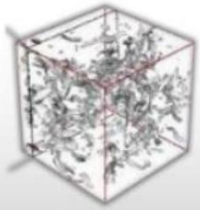


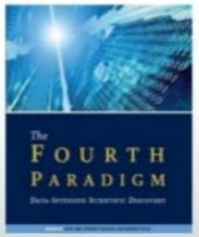
Visualize the size of the data



Data Science: The Fourth Evolution of Discovery Science

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$





Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago	Last few hundred years	Last few decades	Today and the Future
Description of natural phenomena	*Newton's laws, Maxwell's equations...*	*Simulation of complex phenomena*	*Unify theory, experiment and simulation with large multidisciplinary Data* *Using data exploration and data mining (from instruments, sensors, humans...)* *Distributed Communities*

Talking about big data in 2008

http://www.wired.com/wired/issue/16-07

September 2008

Talking about big data in 2014

THE AVERAGE PERSON TODAY PROCESSES MORE DATA IN A SINGLE DAY THAN A PERSON IN THE 1500'S DID IN AN ENTIRE LIFETIME

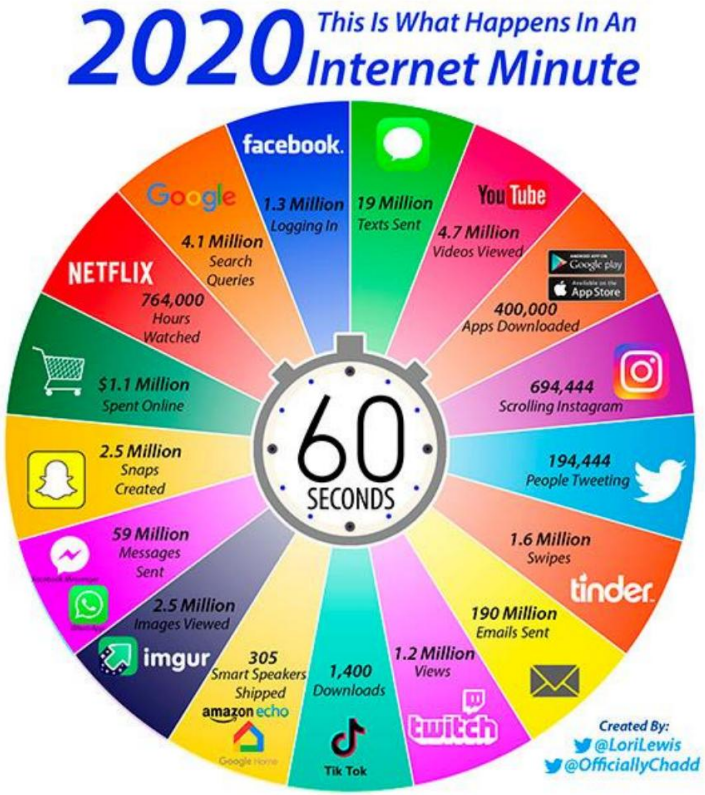
LOOK TO THE LEFT, and you see Times Square at dusk. Look to the right, and you see the same location at midmorning. Internationally acclaimed photographer Stephen Wilkes's time-altering image of New York's Times Square is part of his body of work titled ***Day to Night***.

The image was created by blending more than 1,400 separate photos taken over the course of 15 hours—a meticulous process that took him nearly three months.

PHOTO: STEPHEN WILKES

Big Data Today

Numbers on data generation speed

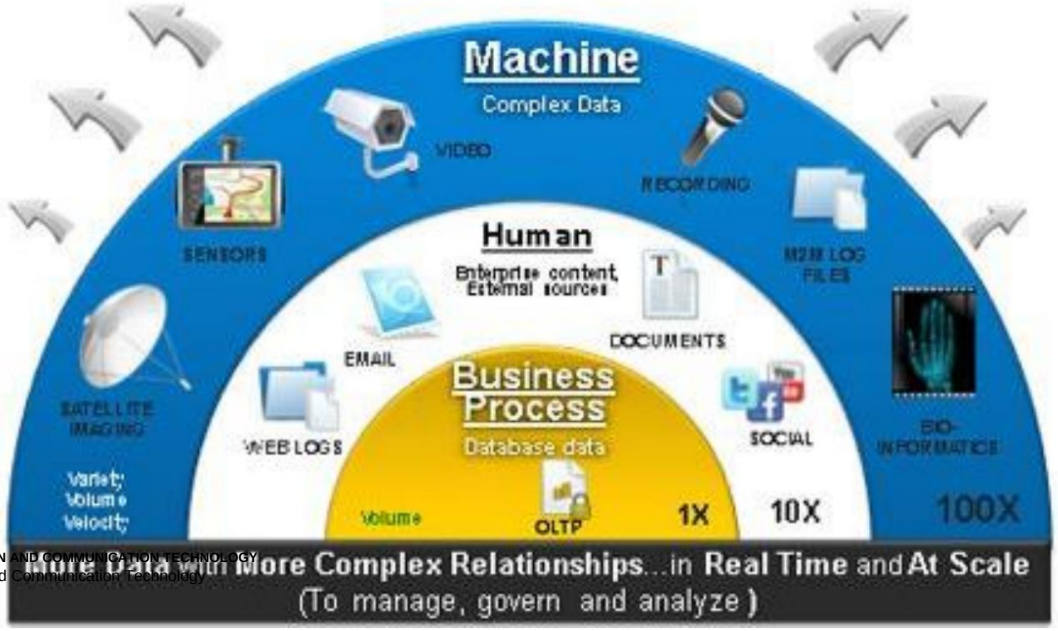


Sources of big data

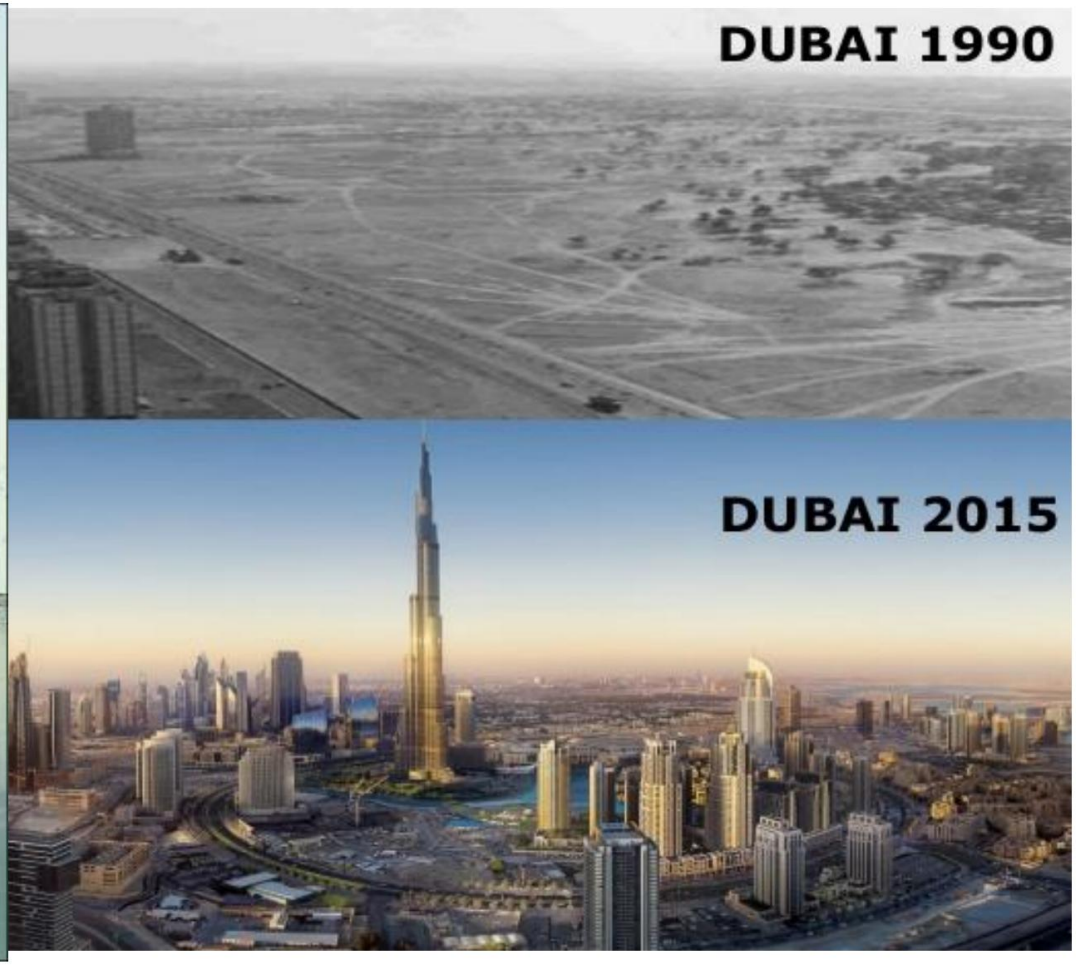
- E-commerce
- Social

Networks • Internet of Things (IoT)

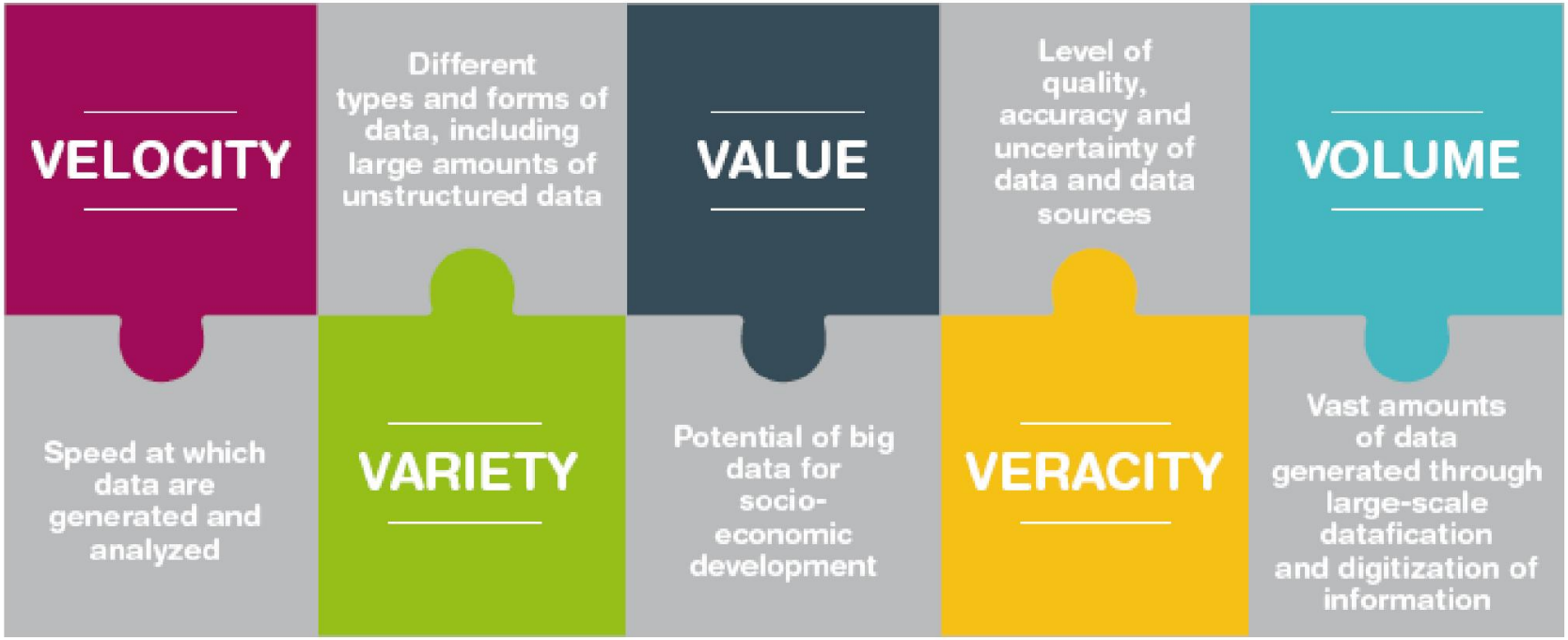
- Big data experiments (bioinformatics, quantum physics, etc.)



Data is the new oil

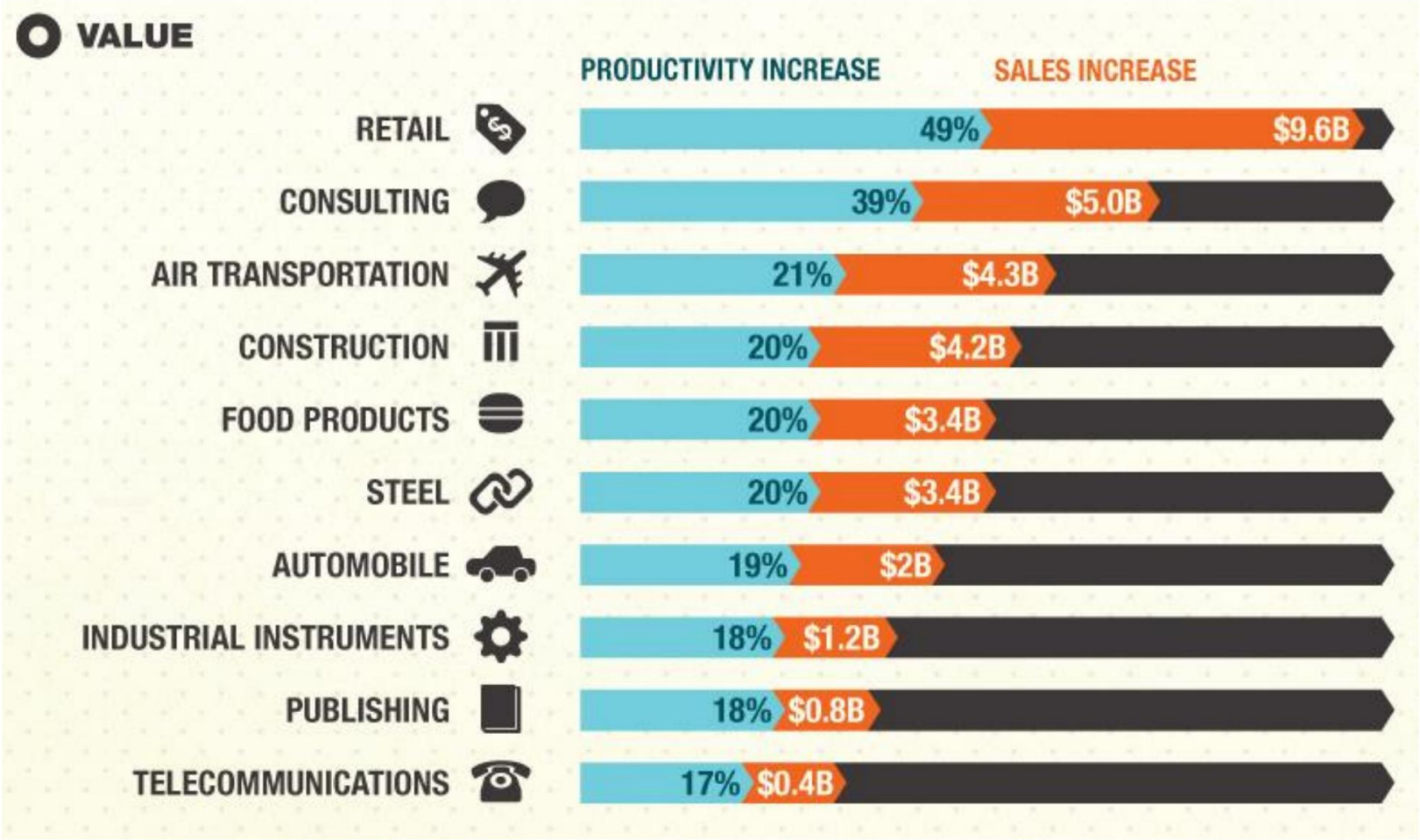


The 5'V's of Big Data



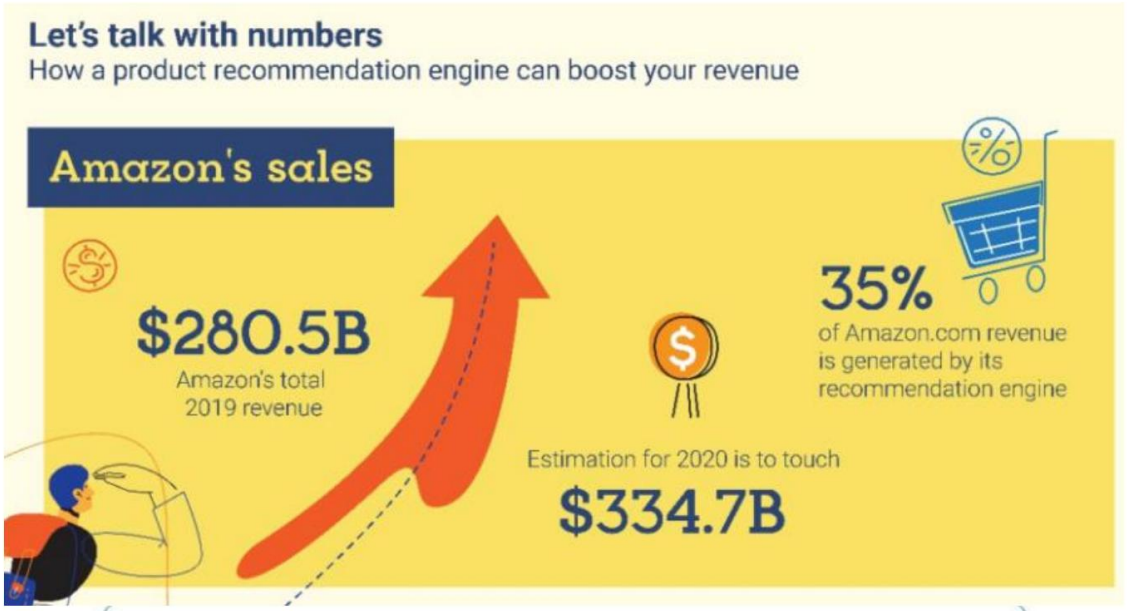
Big data is data sets that are too large or too complex for traditional data storage and processing platforms to handle.

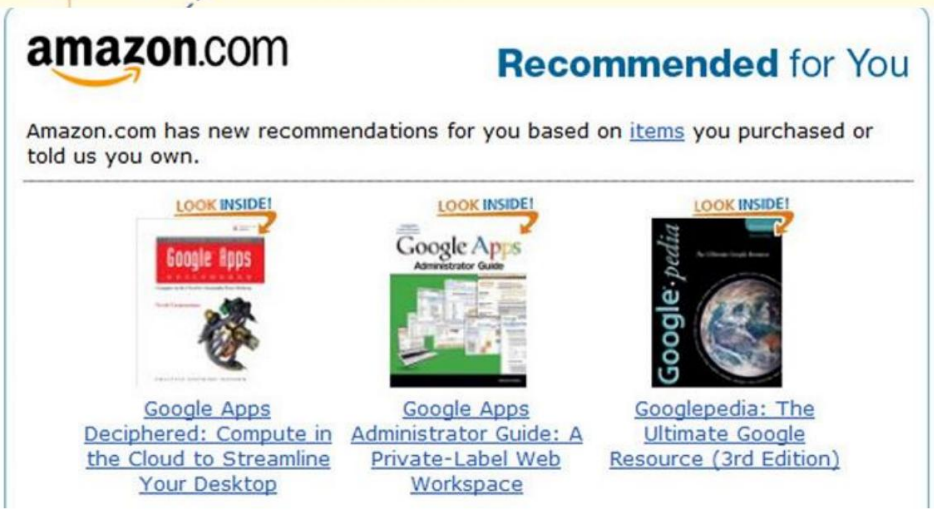
Big data – big value



source: wipro.com

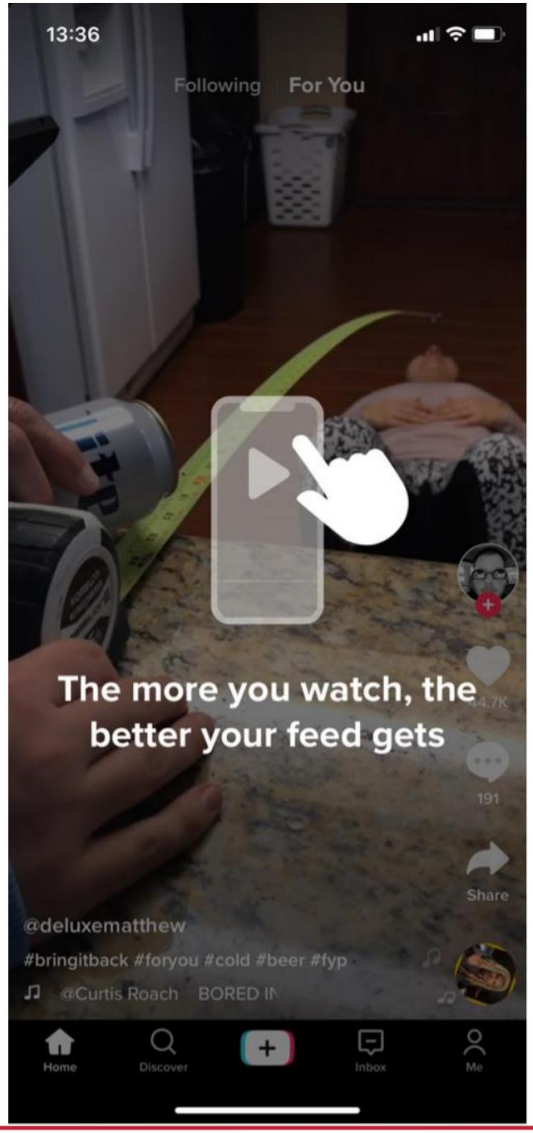
Amazon





Tiktok

- TikTok has 1.04 billion monthly active users globally as of 2024.
- TikTok users spend 58 minutes and 24 seconds on the app daily as of 2024.
- The majority of TikTok users are between the ages of 18 to 34, at 69.3%.



Exploiting Big Data in Education

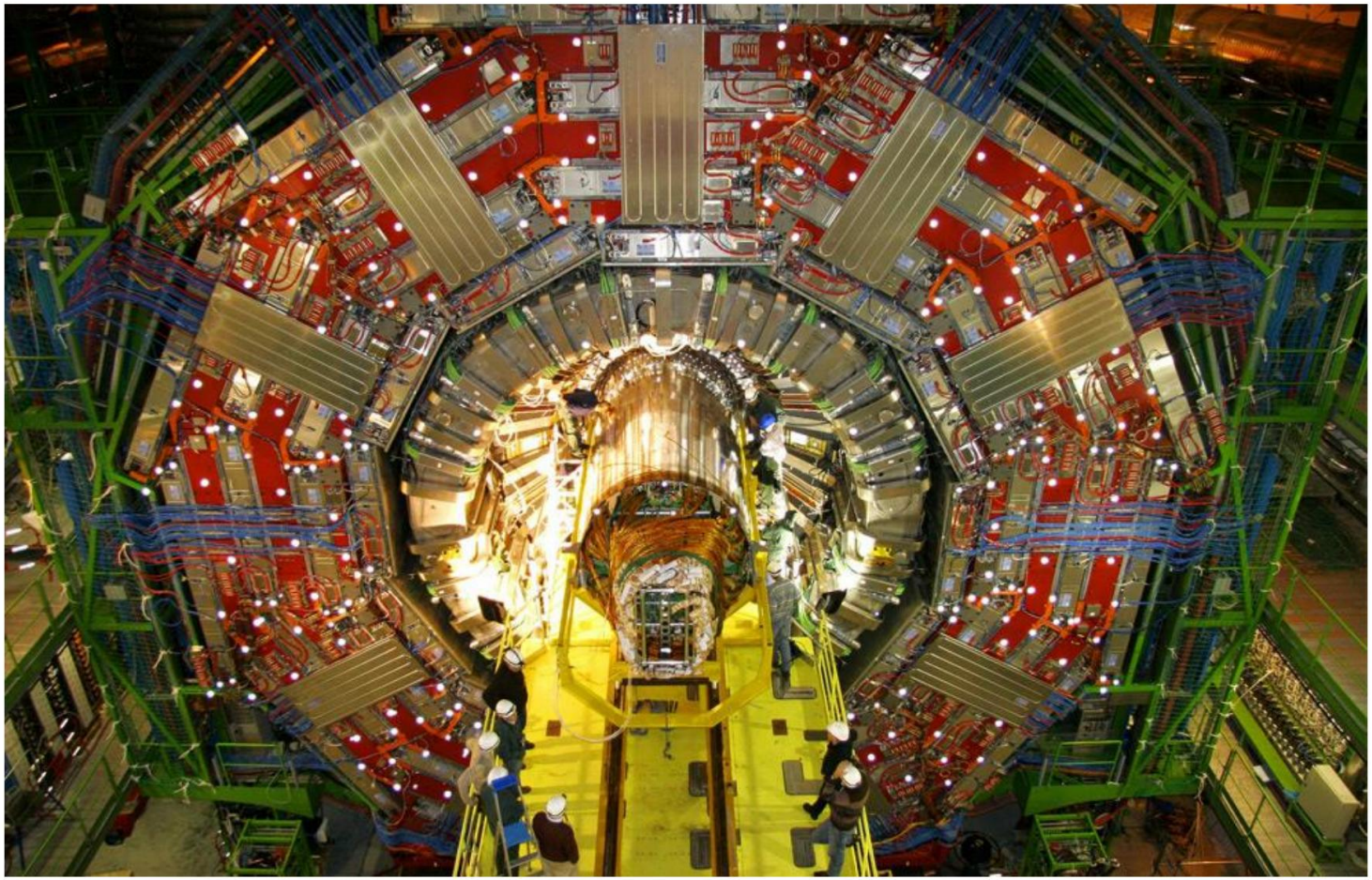
Exploiting Big Data in Healthcare Science

- Reduce treatment costs, redundant testing
- Predict the scale of the pandemic, recommended countermeasures
- Early prevention of diseases that may occur in the future

Exploiting big data in state management

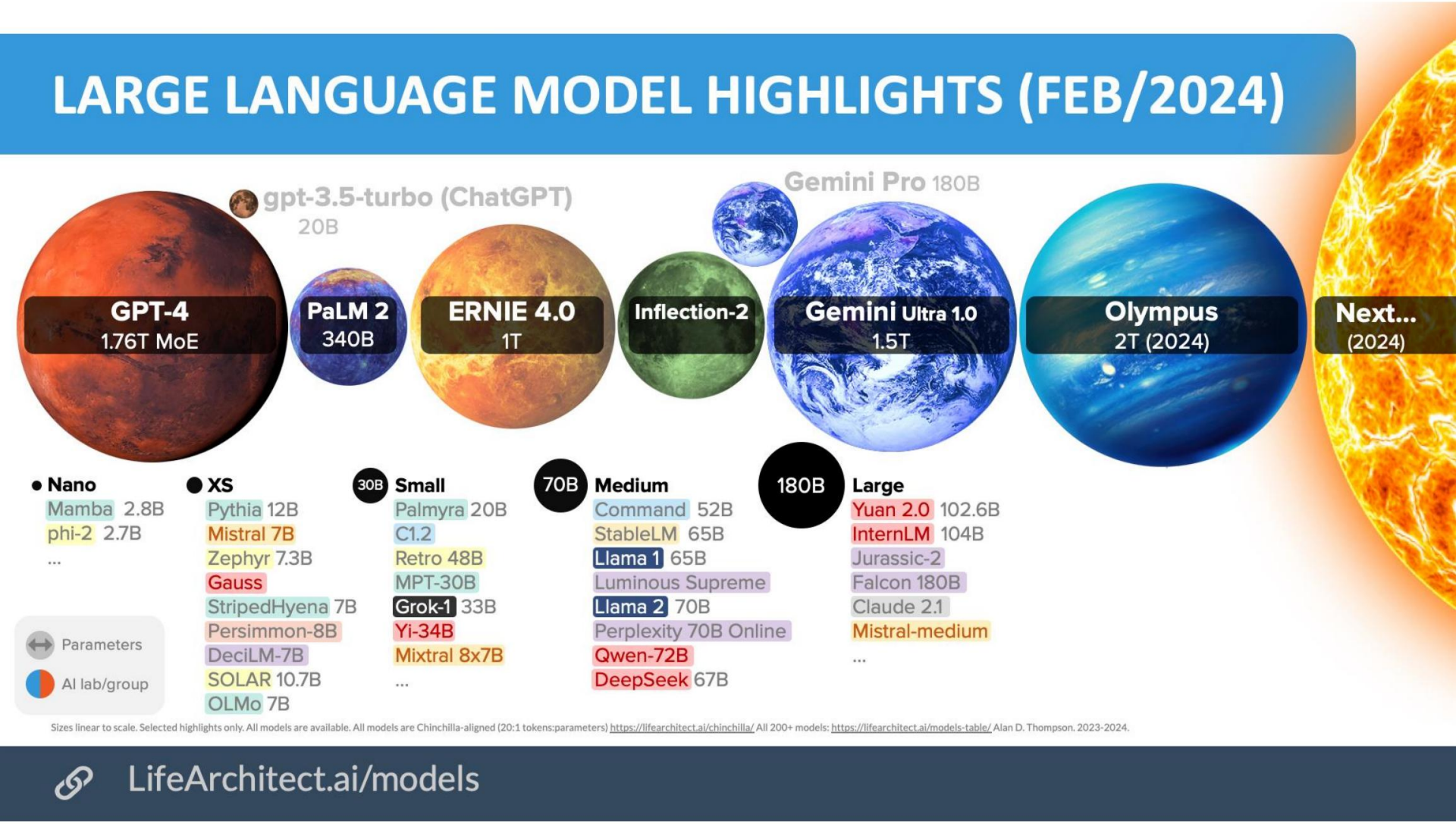
- Social welfare programs • Quickly grasp social issues (employment, crime, environment, etc.) • Recommend countermeasures
- Information security • Tax evasion • Fraud

Exploiting big data in discovery science

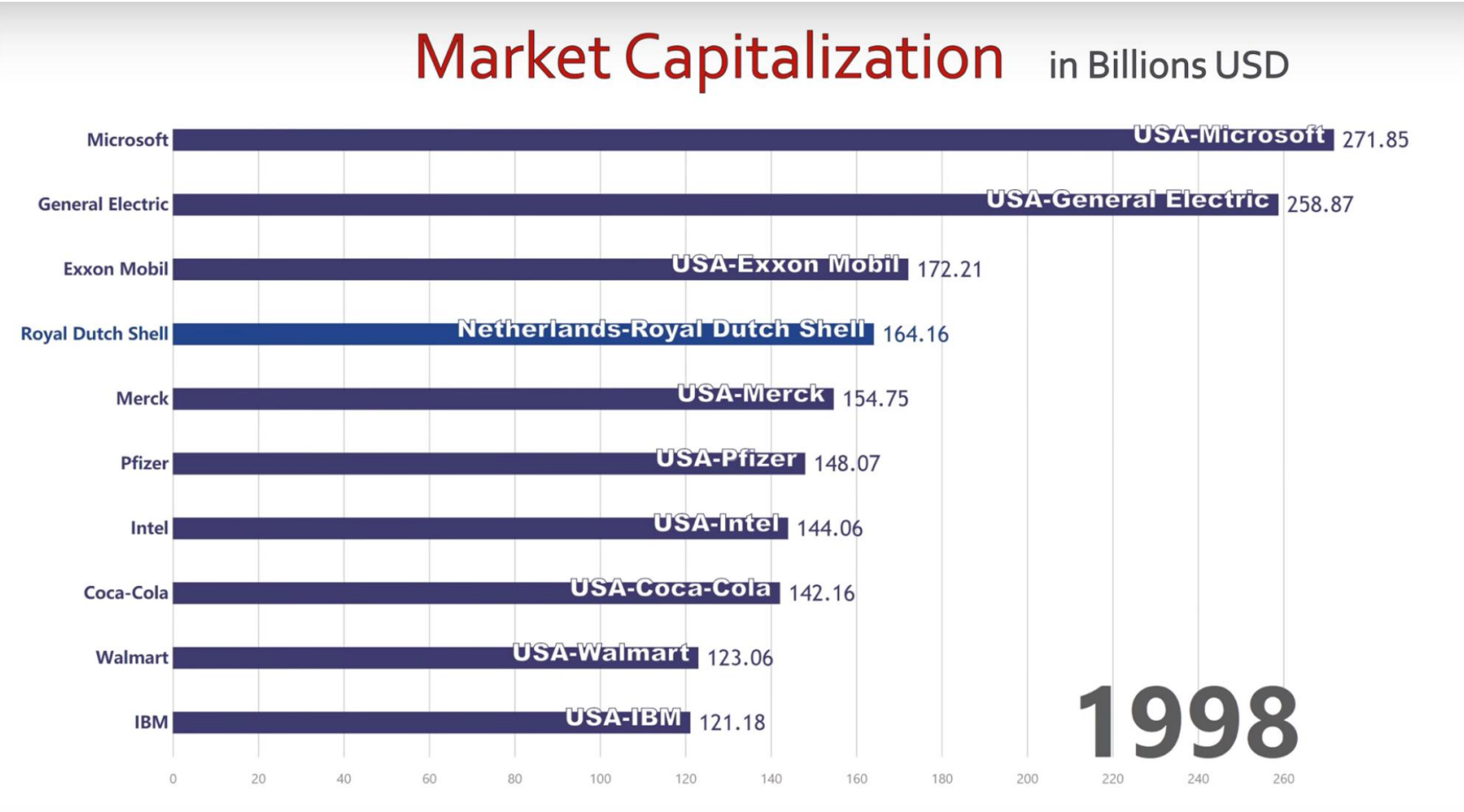


CERN's Large Hydron Collider (LHC) generates 15 PB a year

The world is running out of data to train AI

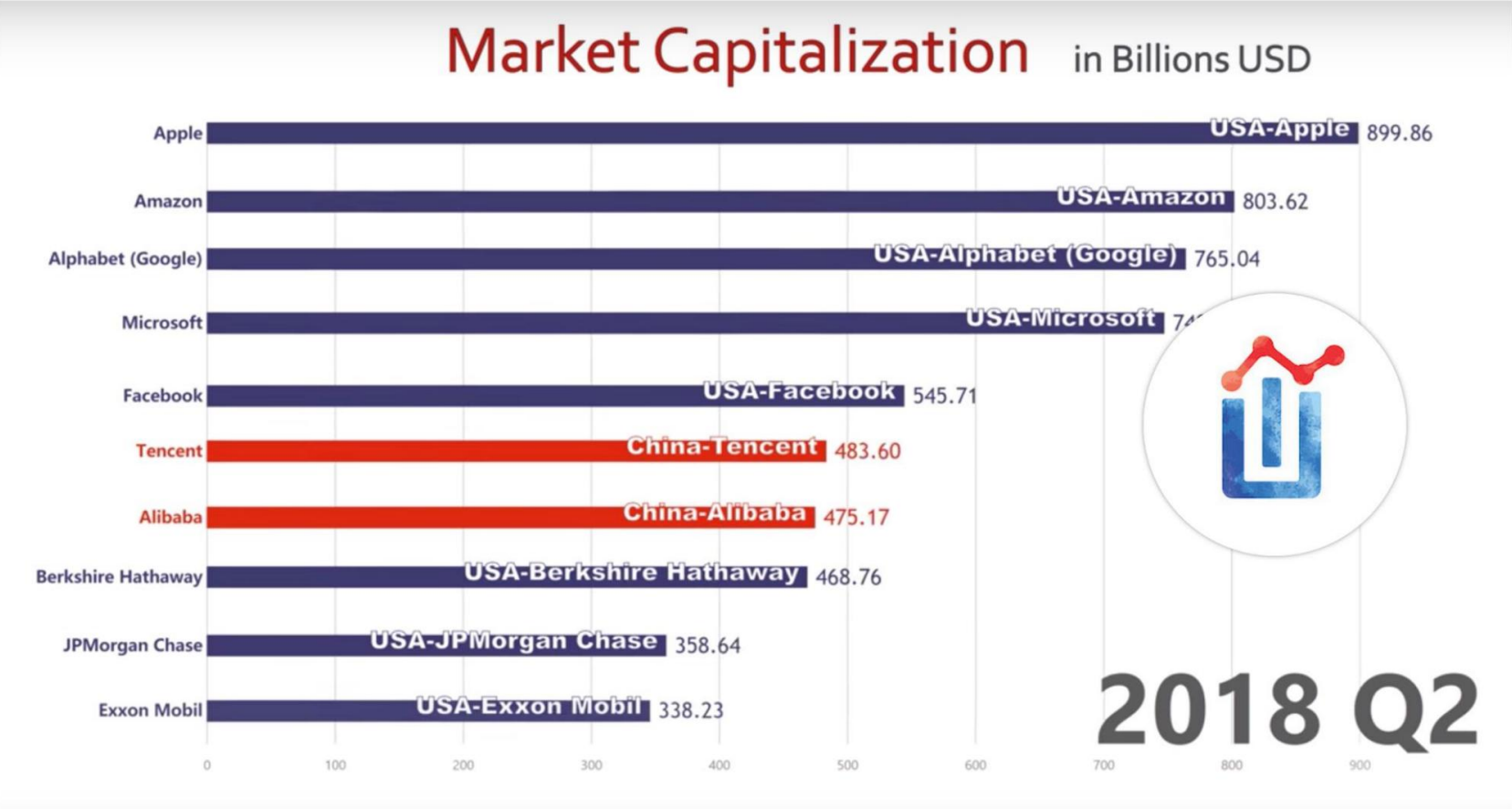


Top 10 Largest Companies (1998-2018)

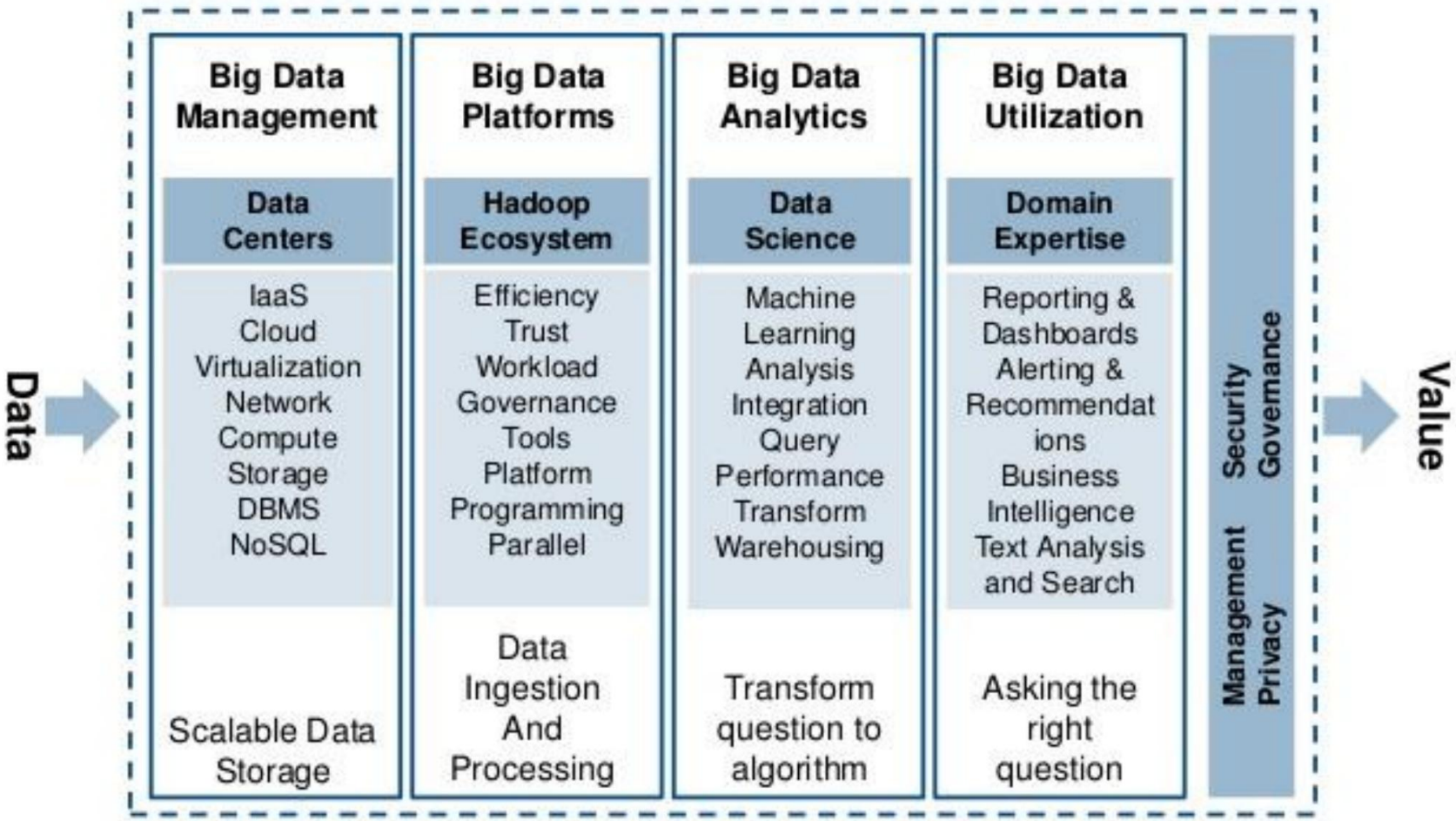


https://www.youtube.com/watch?v=fobx4wIS6W0

Top 10 Largest Companies (1998-2018)



Technology Layers for Big Data



Data management must be open

- Scalability
  - The ability to manage large amounts of data is constantly increasing over time.
- Accessibility
  - Allows efficient data I/O reading and writing.
- Transparency
  - Easy access to data, data storage location on the system is transparent to end users.
- Availability
  - Fault tolerance, when increasing the number of users , when failing.

Data processing and integration must be open

- Data integration • Data
  in different formats • Data exists in
  different data models and schemas • Issues related to information
  security, privacy
  private
- Data processing
  - Processing very large volumes of
    data • Processing large
    data streams • Traditional parallel, distributed data processing (OpenMP, MPI)
    - Complex, hard to
      learn • Limited scalability • Poor
      fault tolerance •
      Expensive infrastructure
    costs • Big data streaming data processing
    architecture • Spark
    mini-batch • Apache Flink

Scalable data analytics algorithms

- Reduce data to fit traditional algorithms • Eg. Sub-sampling
  - Eg. Principal component analysis
  - Eg. Feature extraction and feature selection
- Parallelizing machine learning algorithms
  - Eg. k-nn classification based on MapReduce
  - Eg. scaling-up support vector machines (SVM) by a divide and-conquer approach

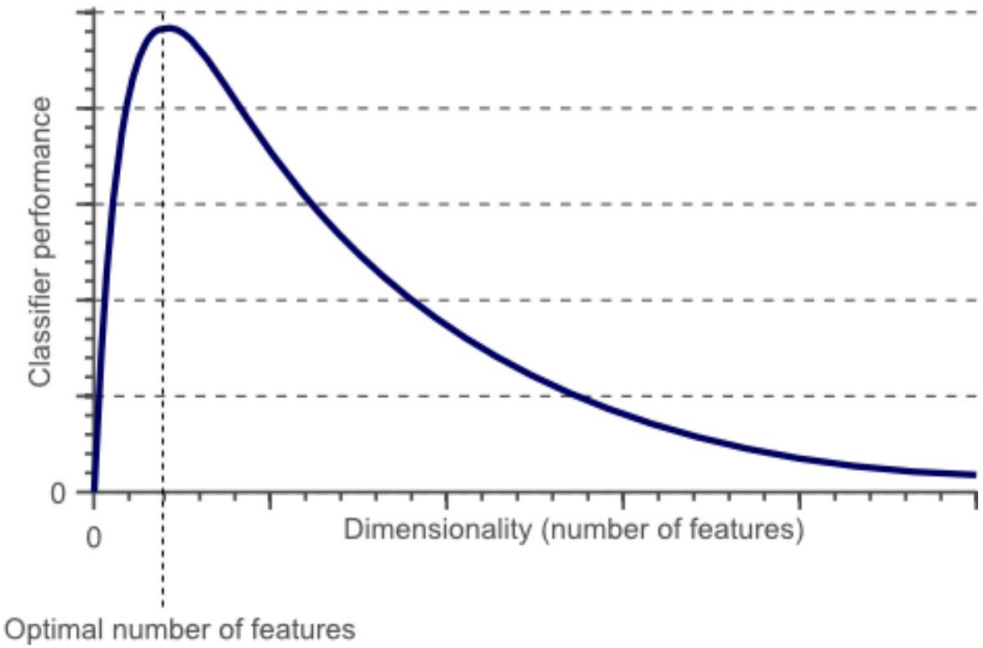
Eg. Curse of dimensionality

- The number of samples needed for the model to learn increases as the data dimension

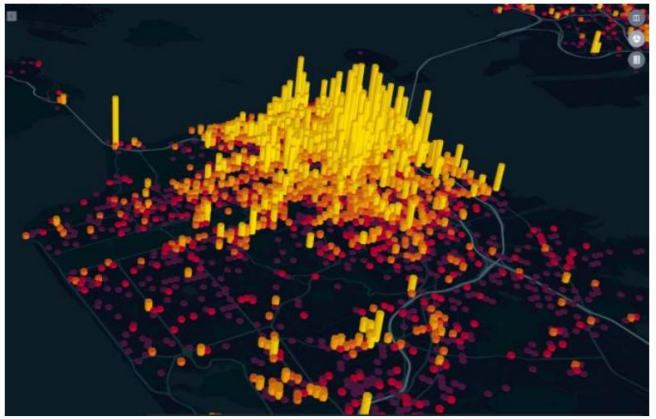
increases • In practice: The number of samples to learn is usually fixed

- => The accuracy of the model decreases when the number of dimensions in the data increases.

study material

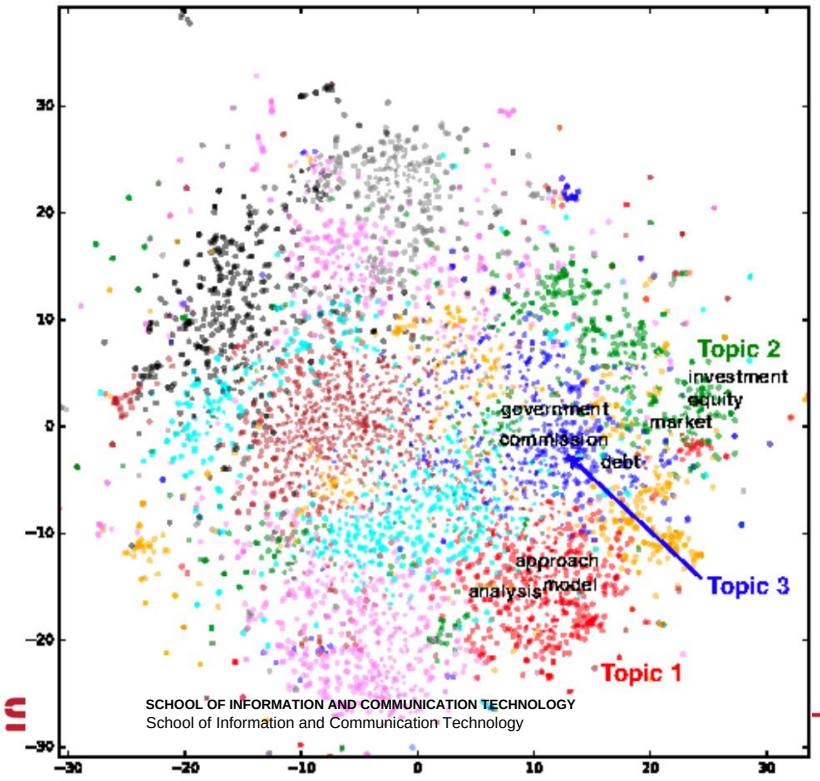


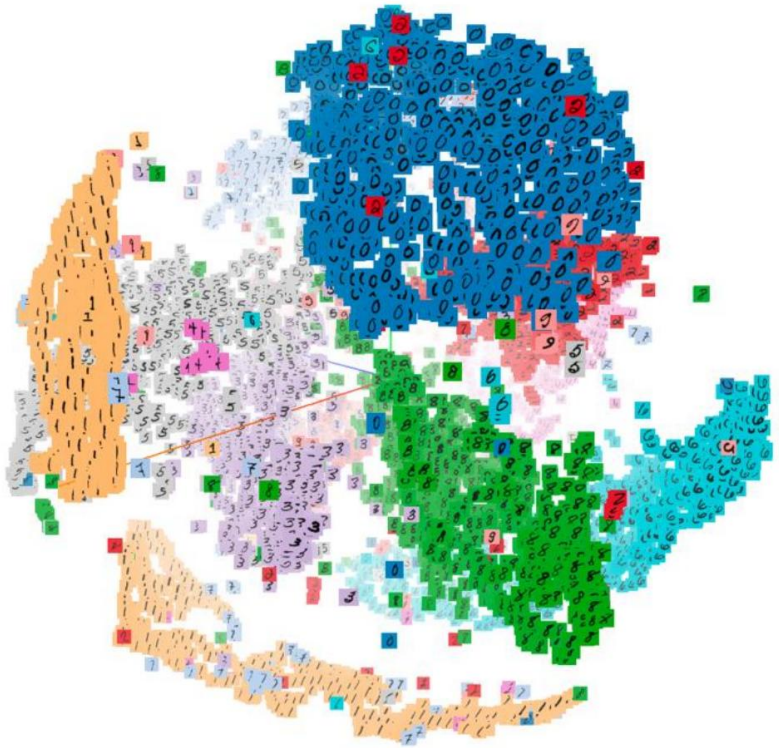
Using and visualizing big data



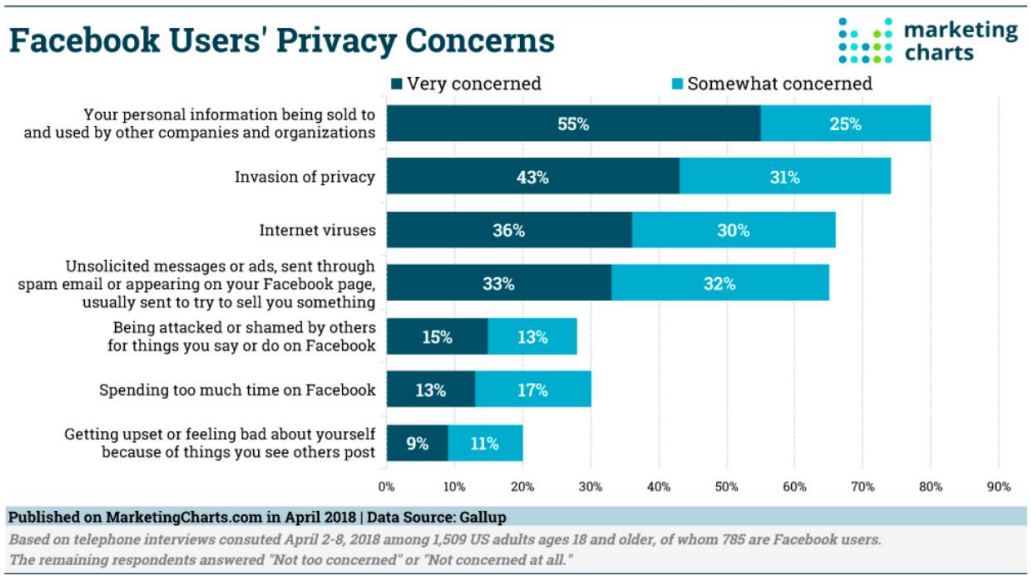
• Requires expert knowledge •

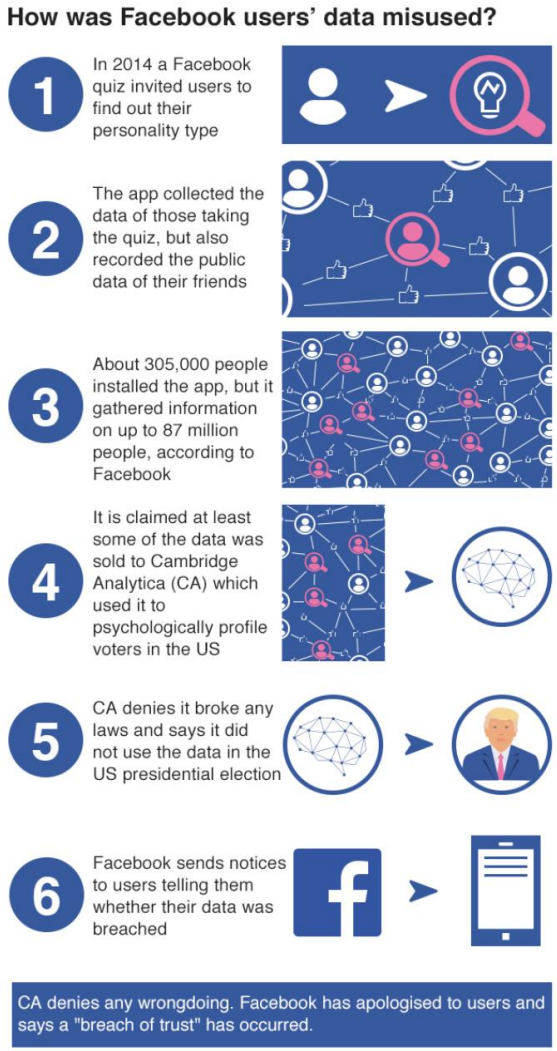
Requires techniques and tools to effectively support the visualization and understanding of big data





Security and Privacy

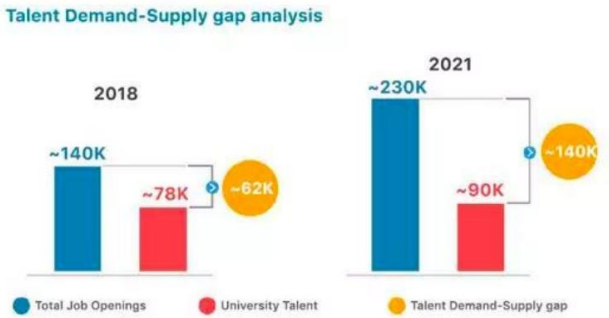




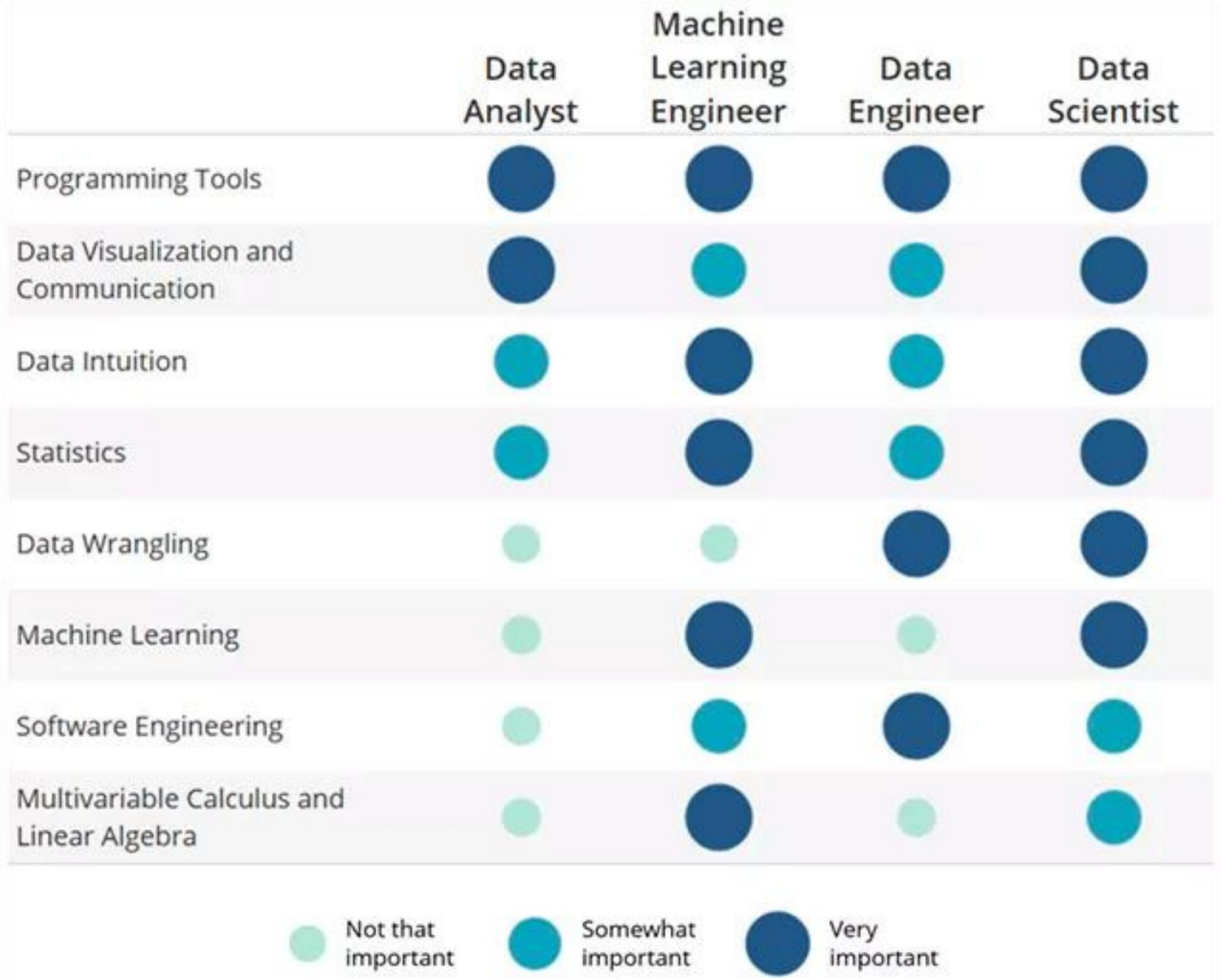
Big Data-Related Talent Shortage

Table 2. Summary Demand Statistics

DSA Framework Category	Number of Postings in 2015	Projected 5-Year Growth	Estimated Postings for 2020	Average Time to Fill (Days)	Average Annual Salary
All	2,352,681	15%	2,716,425	45	$80,265
Data-Driven Decision Makers	812,099	14%	922,428	48	$91,467
Functional Analysts	770,441	17%	901,743	40	$69,162
Data Systems Developers	558,326	15%	641,635	50	$78,553
Data Analysts	124,325	16%	143,926	38	$69,949
Data Scientists & Advanced Analysts	48,347	28%	61,799	46	$94,576
Analytics Managers	39,143	15%	44,894	43	$105,909



Grouping required skills by position



Tips for learning about big data

- Learn programming
  - Coursera •
  - Udacity •
  - Freecodecamp •
  - Codecademy •

Learn machine learning, math, statistics • Kaggle •

Hadoop,

NoSQL, Spark • Reporting and

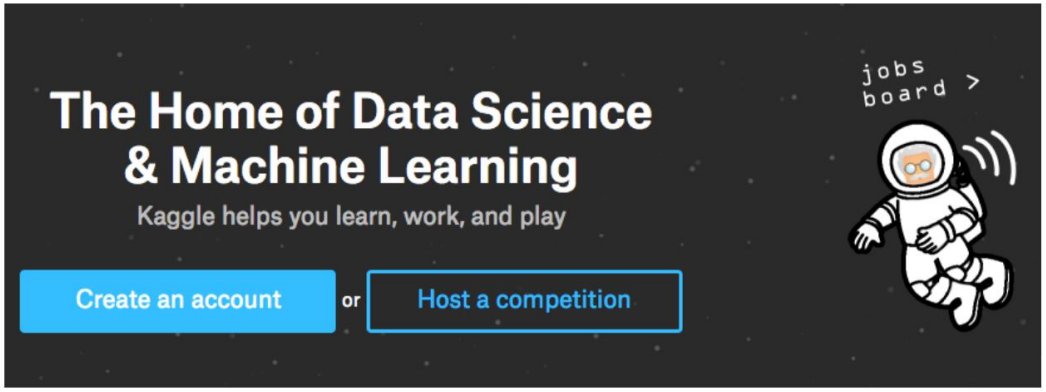
visualization tools

  - Tableau
  - Pentahoo

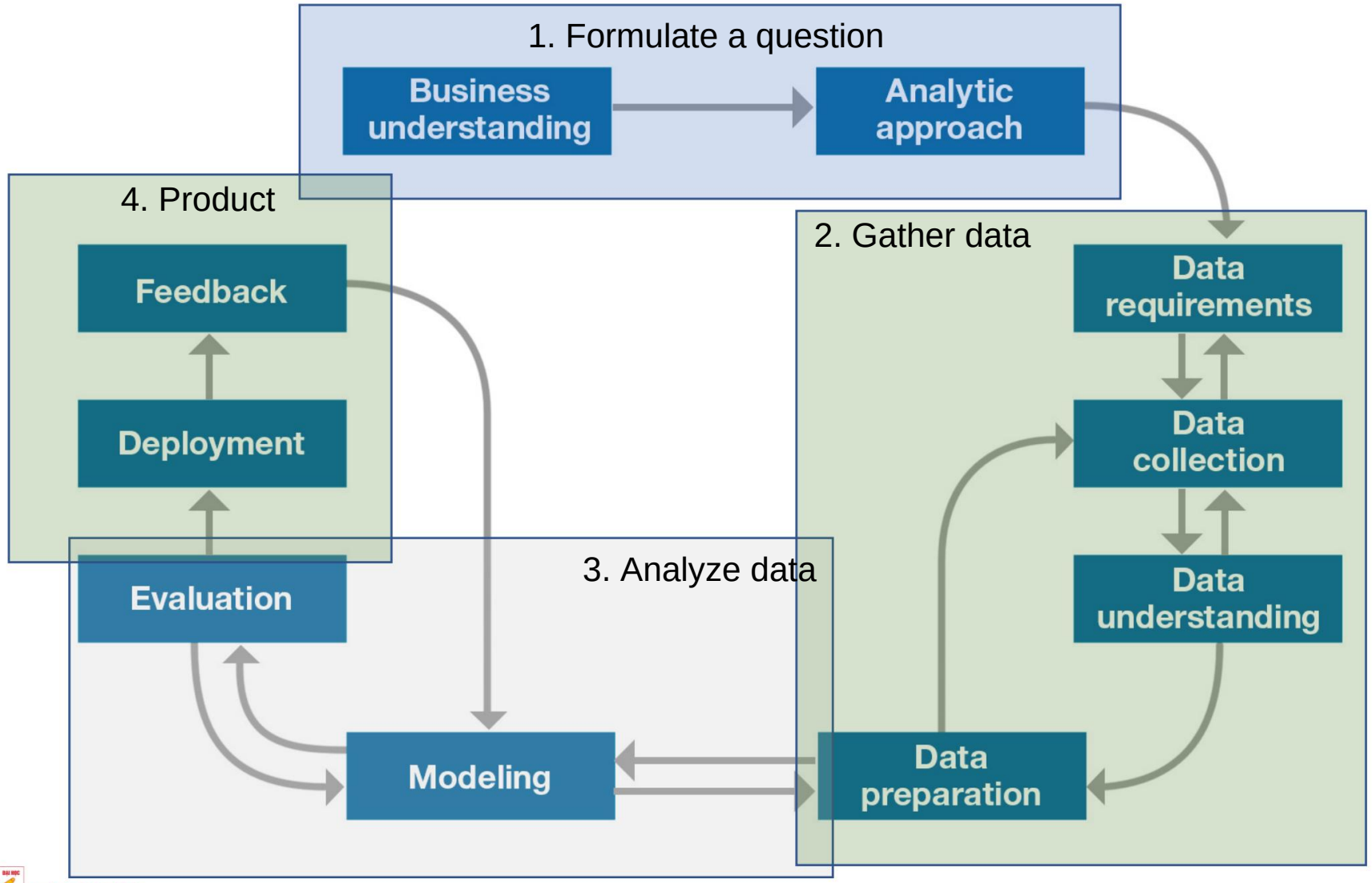
• Meet and share • Find a

mentor • Internship,

project

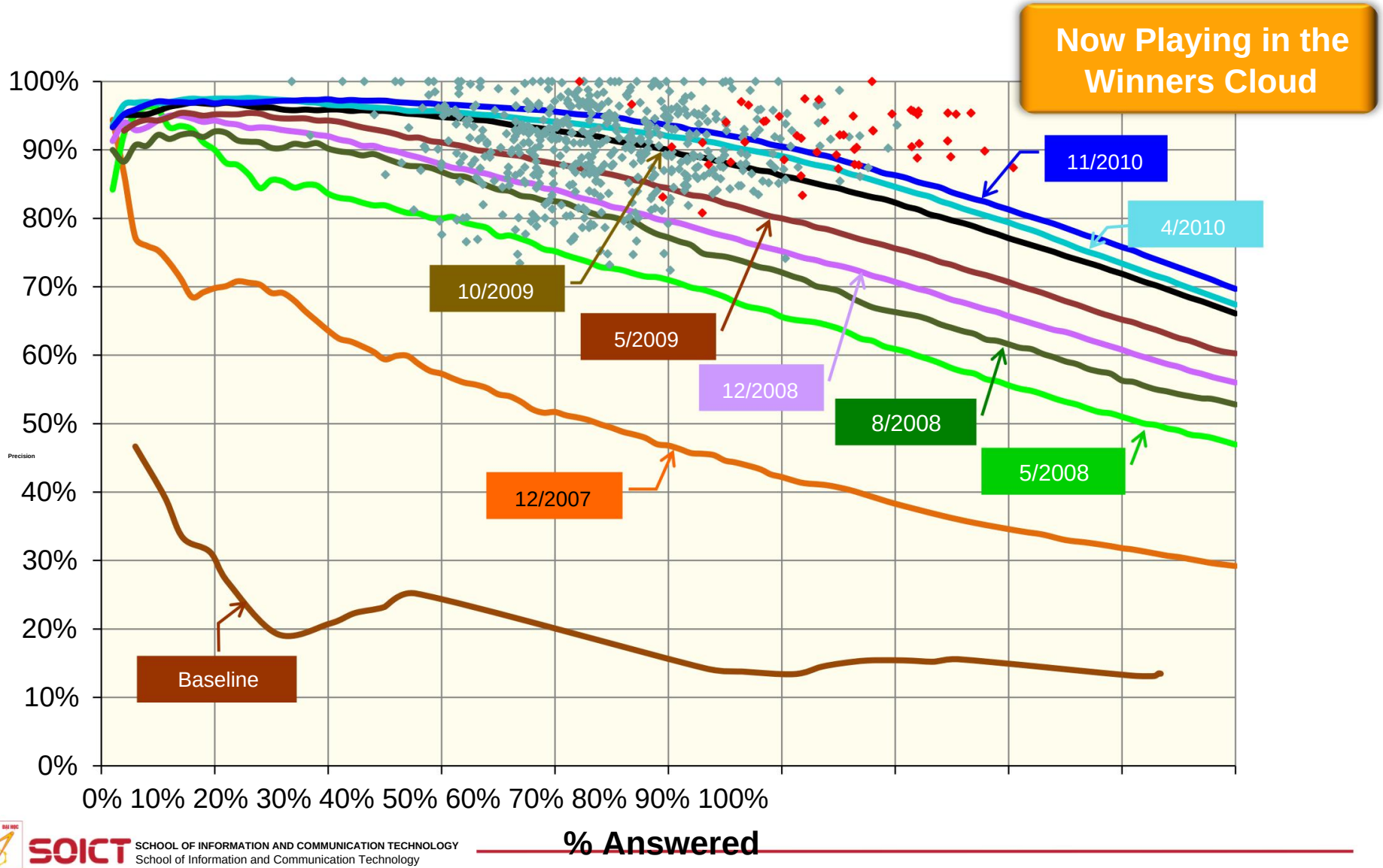


Data Science Process



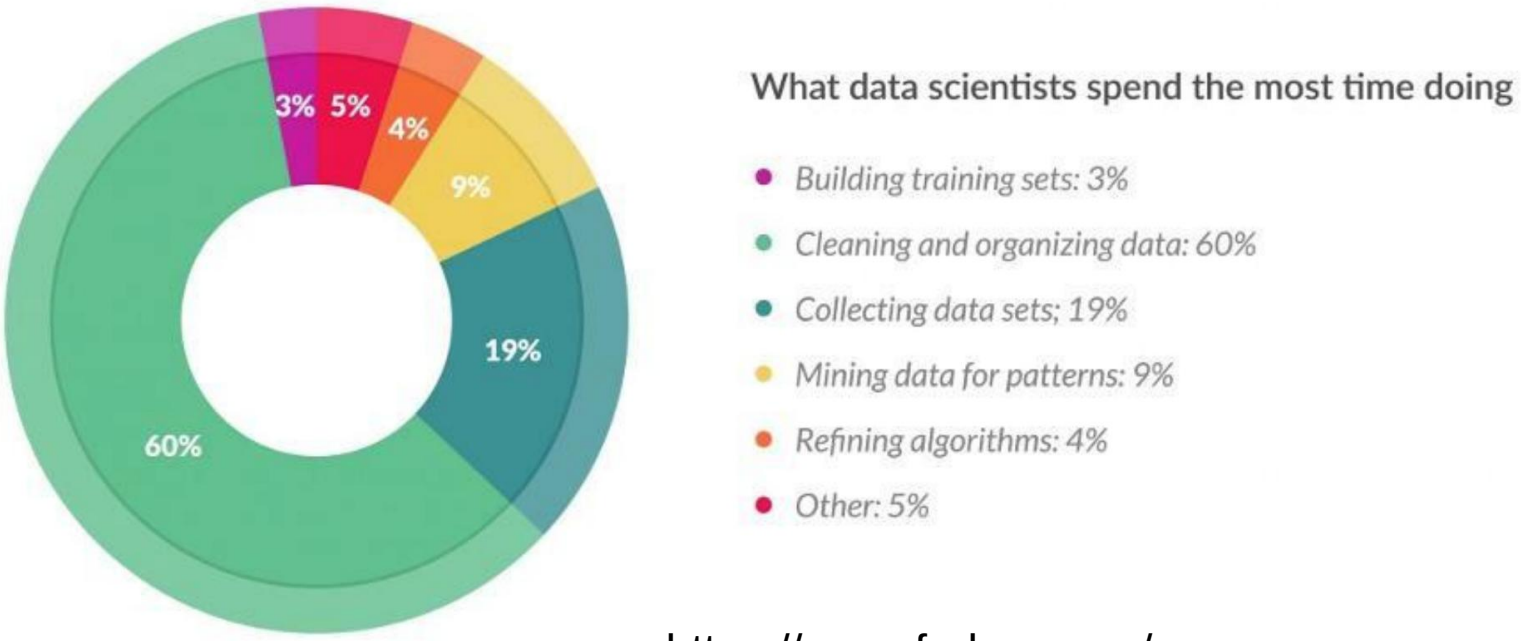
Source: *Foundational Methodology for Data Science,* IBM, 2015

DeepQA: Incremental Progress in Precision and Confidence June 2007-November 2010



Big Data Cleaning: A Time-Consuming and Laborious Task

- Accounts for about 80% of data scientist jobs

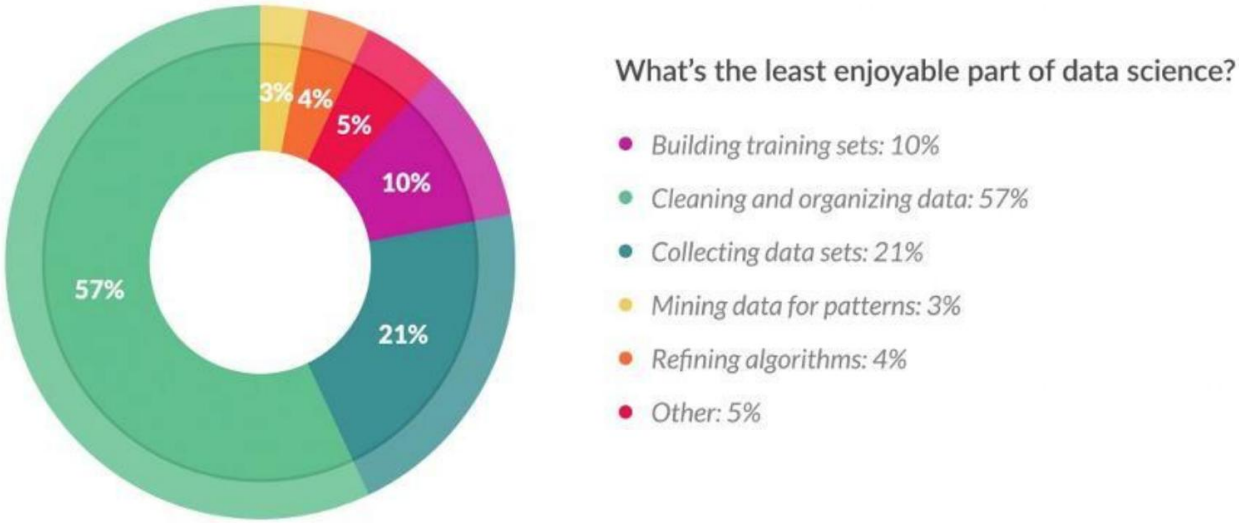


source: https://www.forbes.com/

Big Data Cleaning: A Time-Consuming and Laborious Task

- 57% of data scientists say this is their job

  less interesting



References

[1] Tiwari, Shashank. Professional NoSQL. John Wiley & Sons, 2011.

[2] Lam, Chuck. Hadoop in action. Manning Publications Co., 2010.

[3] Miner, Donald, and Adam Shook. MapReduce design patterns: building effective algorithms and analytics for Hadoop and other systems. "O'Reilly Media, Inc.", 2012.

[4] Karau, Holden. Fast Data Processing with Spark. Packt Publishing Ltd, 2013.

[5] Penchikala, Srini. Big data processing with apache spark. Lulu. com, 2018.

[6] White, Tom. Hadoop: The definitive guide. "O'Reilly Media, Inc.", 2012.

[7] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.

[8] Cattell, Rick. "Scalable SQL and NoSQL data stores." Acm Sigmod Record 39.4 (2011): 12-27.

[9] Gessert, Felix, et al. "NoSQL database systems: a survey and decision guidance." Computer Science-Research and Development 32.3-4 (2017): 353- 365.

[10] George, Lars. HBase: the definitive guide: random access to your planet-size data. "O'Reilly Media, Inc.", 2011.

[11] Sivasubramanian, Swaminathan. "Amazon dynamoDB: a scalable seamless non-relational database service." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.

[12] Chan, L. "Presto: Interacting with petabytes of data at Facebook." (2013).

[13] Garg, Nishant. Apache Kafka. Packt Publishing Ltd, 2013.

[14] Karau, Holden, et al. Learning spark: lightning-fast big data analysis. "O'Reilly Media, Inc.", 2015.

[15] Iqbal, Muhammad Hussain, and Tariq Rahim Soomro. "Big data analysis: Apache storm perspective." International journal of computer trends and technology 19.1 (2015): 9-14.

[16] Toshniwal, Ankit, et al. "Storm@ twitter." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.

[17] Lin, Jimmy. "The lambda and the kappa." IEEE Internet Computing 21.5 (2017): 60-66.

Online courses

- https://www.coursera.org/learn/nosql-database-systems
- https://who.rocq.inria.fr/Vassilis.Christophides/Big/index.htm
- https://www.coursera.org/learn/big-data-introduction?specialization=big-data
- https://www.coursera.org/learn/big-data-integration-processing?specialization=big-data
- https://www.coursera.org/learn/big-data-management?specialization=big-data
- https://www.coursera.org/learn/hadoop
- https://www.coursera.org/learn/scala-spark-big-data