

25 YEARS ANNIVERSARY
SICT

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY



HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

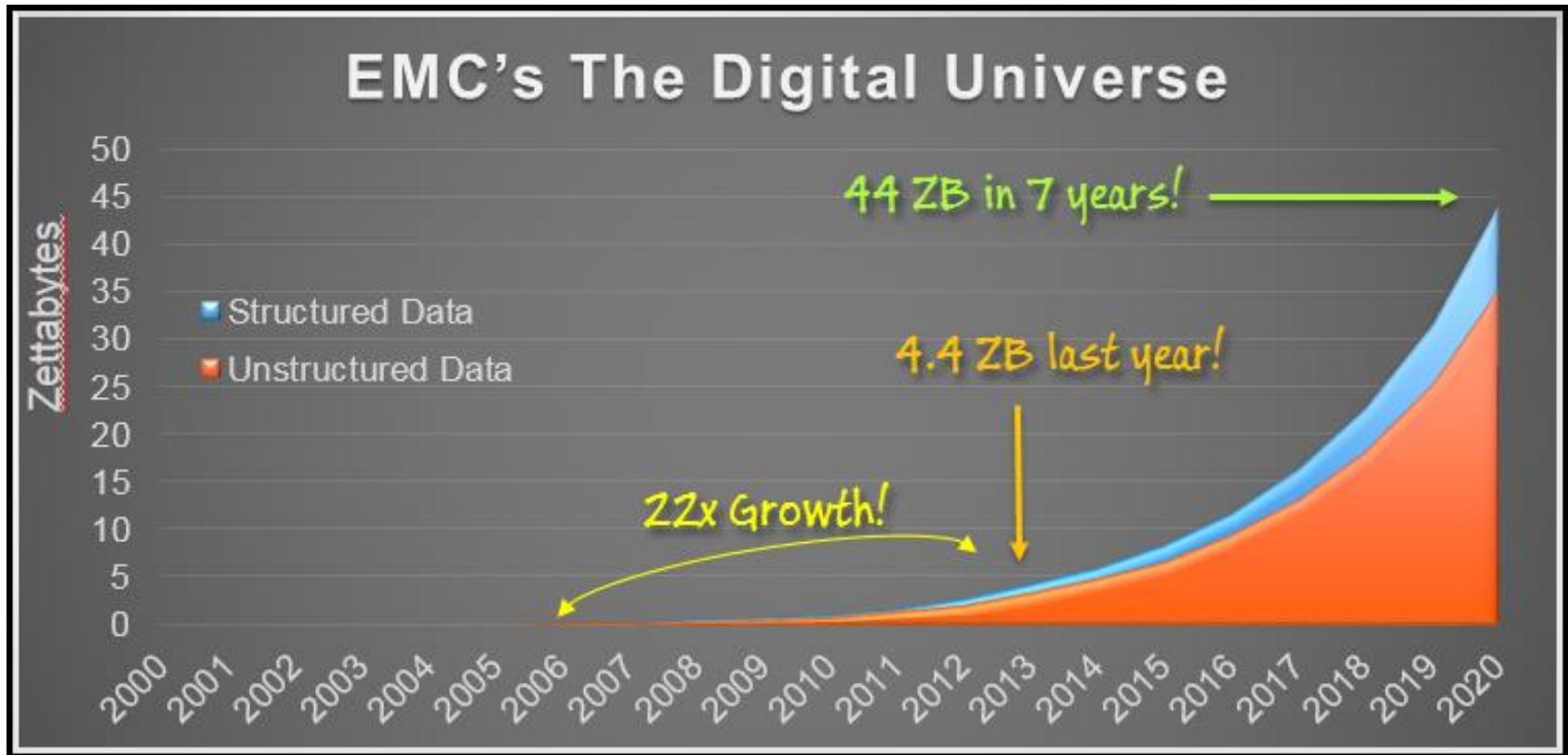
Lecture 1

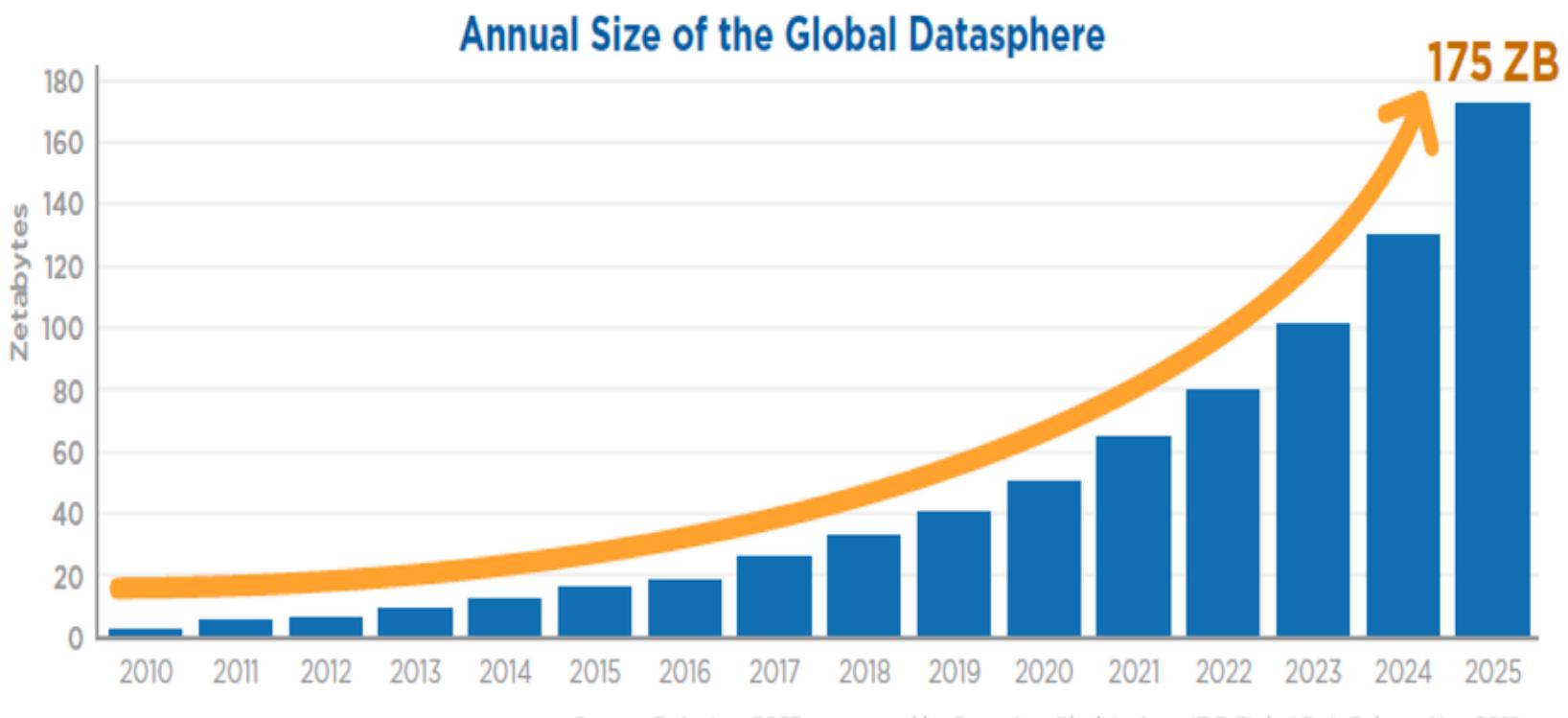
Introduction to big data storage and processing

Syllabus

STT	Lecture
1	Tổng quan về lưu trữ và xử lý dữ liệu lớn
2	Hệ sinh thái Hadoop (Hadoop ecosystem)
3	Hệ thống tập tin phân tán Hadoop HDFS
4	Cơ sở dữ liệu phi quan hệ NoSQL - phần 1 Tổng quan
5	Cơ sở dữ liệu phi quan hệ NoSQL - phần 2 Kiến trúc phân tán phổ biến
6	Cơ sở dữ liệu phi quan hệ NoSQL - phần 3 Truy vấn SQL trên NoSQL
7	Hệ thống truyền thông điệp phân tán
8	Các kỹ thuật xử lý dữ liệu lớn theo khối - phần 1 Map Reduce
9	Các kỹ thuật xử lý dữ liệu lớn theo khối - phần 2 Apache Spark
10	Các kỹ thuật xử lý luồng dữ liệu lớn Spark Streaming
11	Kiến trúc dữ liệu lớn Lambda architecture
12	Phân tích dữ liệu lớn Spark ML

How big is big data?

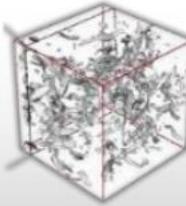
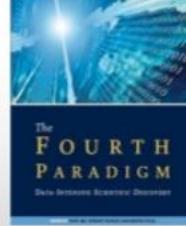




How big is big data?



Data science: The 4th paradigm for scientific discovery

	$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$		
Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago <i>Description of natural phenomena</i>	Last few hundred years <i>Newton's laws, Maxwell's equations...</i>	Last few decades <i>Simulation of complex phenomena</i>	Today and the Future <i>Unify theory, experiment and simulation with large multidisciplinary Data</i> <i>Using data exploration and data mining (from instruments, sensors, humans...)</i> <i>Distributed Communities</i>

Big data in 2008

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

September 2008



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



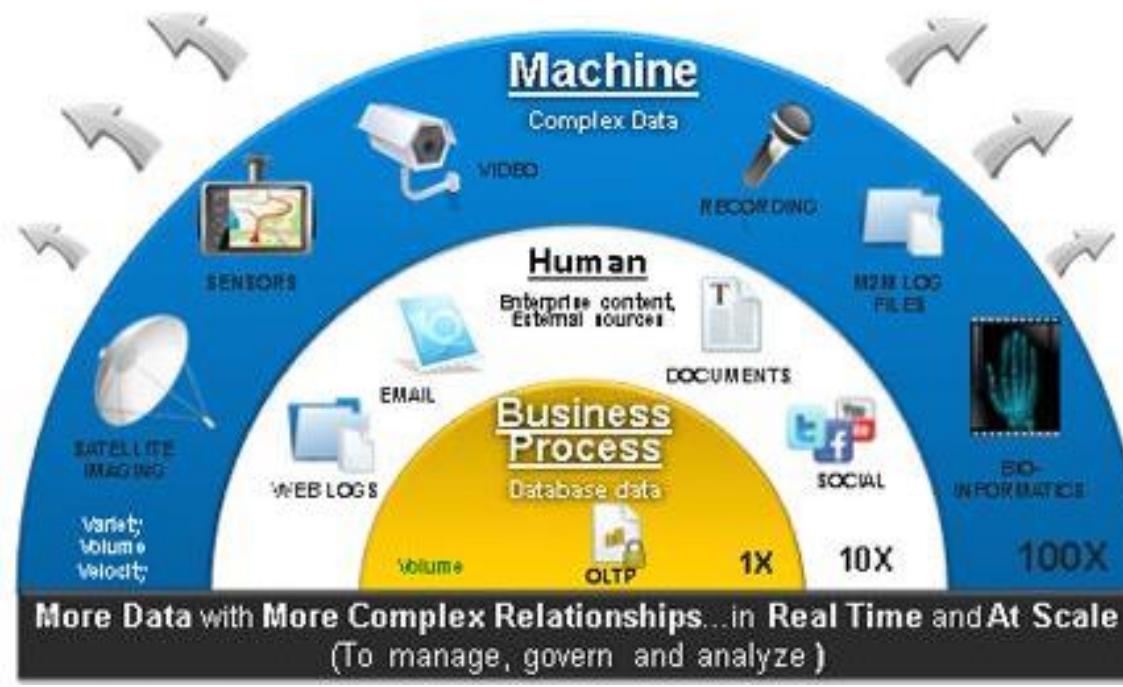
The amount of information generated during the first day of a baby's life today is equivalent to 70 times the information contained in the Library of Congress

Big numbers



Big data sources

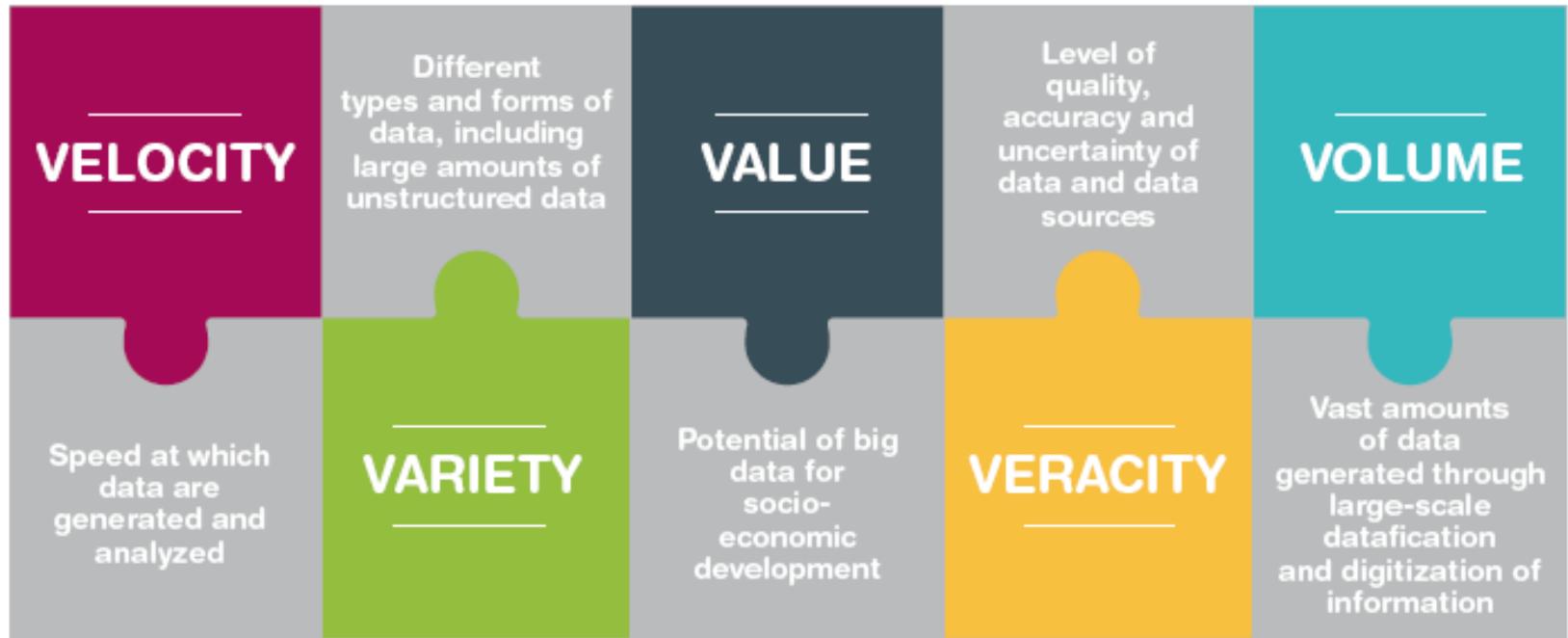
- E-commerce
- Social networks
- Internet of things
- Data-intensive experiments (bioinformatics, quantum physics)



Data is the new oil

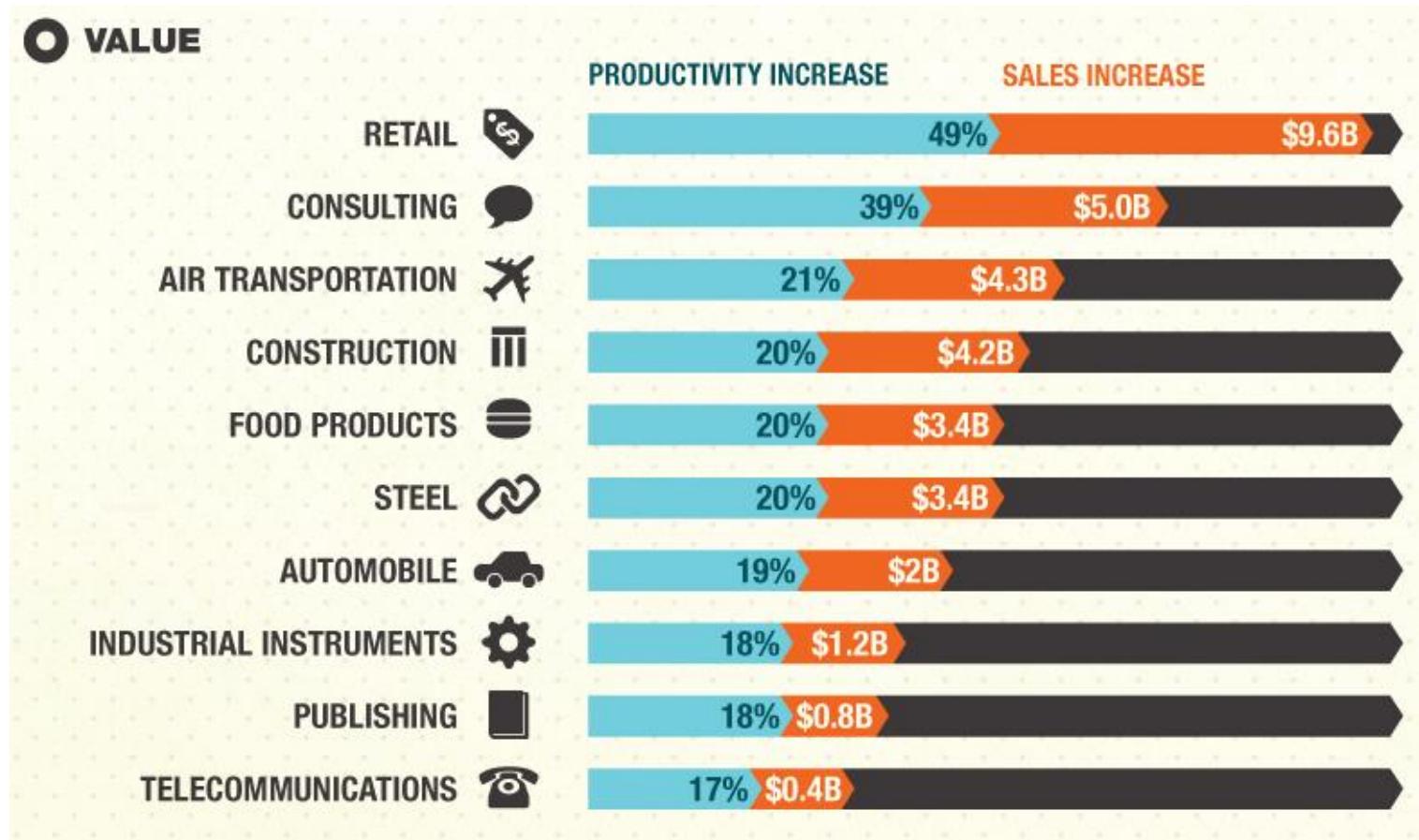


Big data 5'V



Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them (wikipedia)

Big data – big value



Big Data in education industry

- Customized and Dynamic Learning Programs
- Reframing Course Material
- Grading Systems
- Career Prediction



Edtech

- Coursera
- VioEdu
- <https://byjus.com/>
 - Engaging Video Lessons
 - Personalized Learning Journeys
 - Mapped to the Syllabus
 - In-depth Analysis
 - Engaging Interactive Questions



Big Data in healthcare industry

- Reduce costs of treatments, unnecessary diagnosis.
- Predict outbreaks of epidemics and preventive measures.
- Avoid preventable diseases



Big Data in government sector

- Welfare Schemes

- Make faster and informed decisions
- Identify areas that are in immediate need of attention
- Overcome national challenges such as unemployment, terrorism,..

- Cyber Security

- Deceit recognition.
- Catching tax evaders.

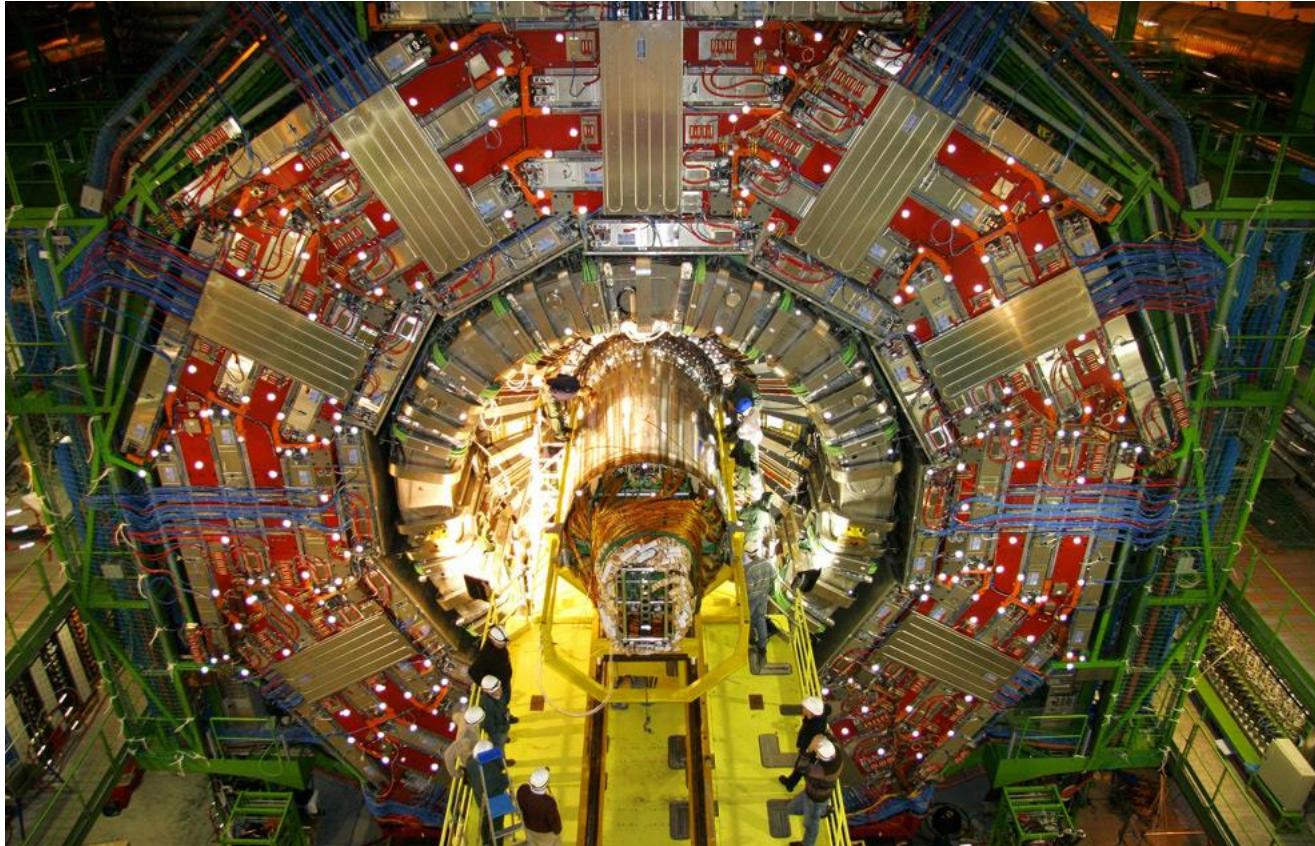


Big Data in media and entertainment industry

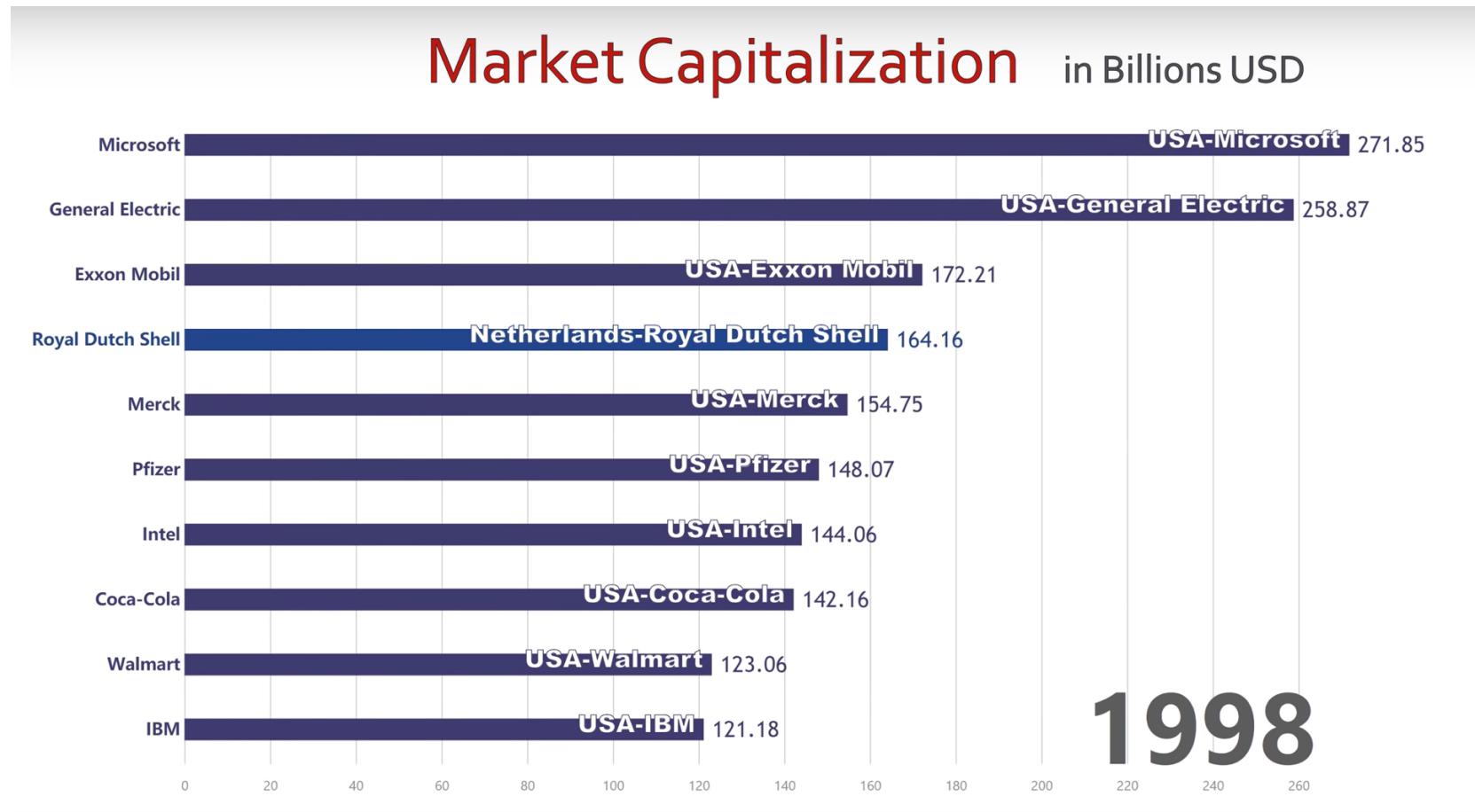
- Predicting the interests of audiences
- Optimized or on-demand scheduling of media streams in digital media distribution platforms
- Getting insights from customer reviews
- Effective targeting of the advertisements
- Example
 - Spotify, Amazon Prime



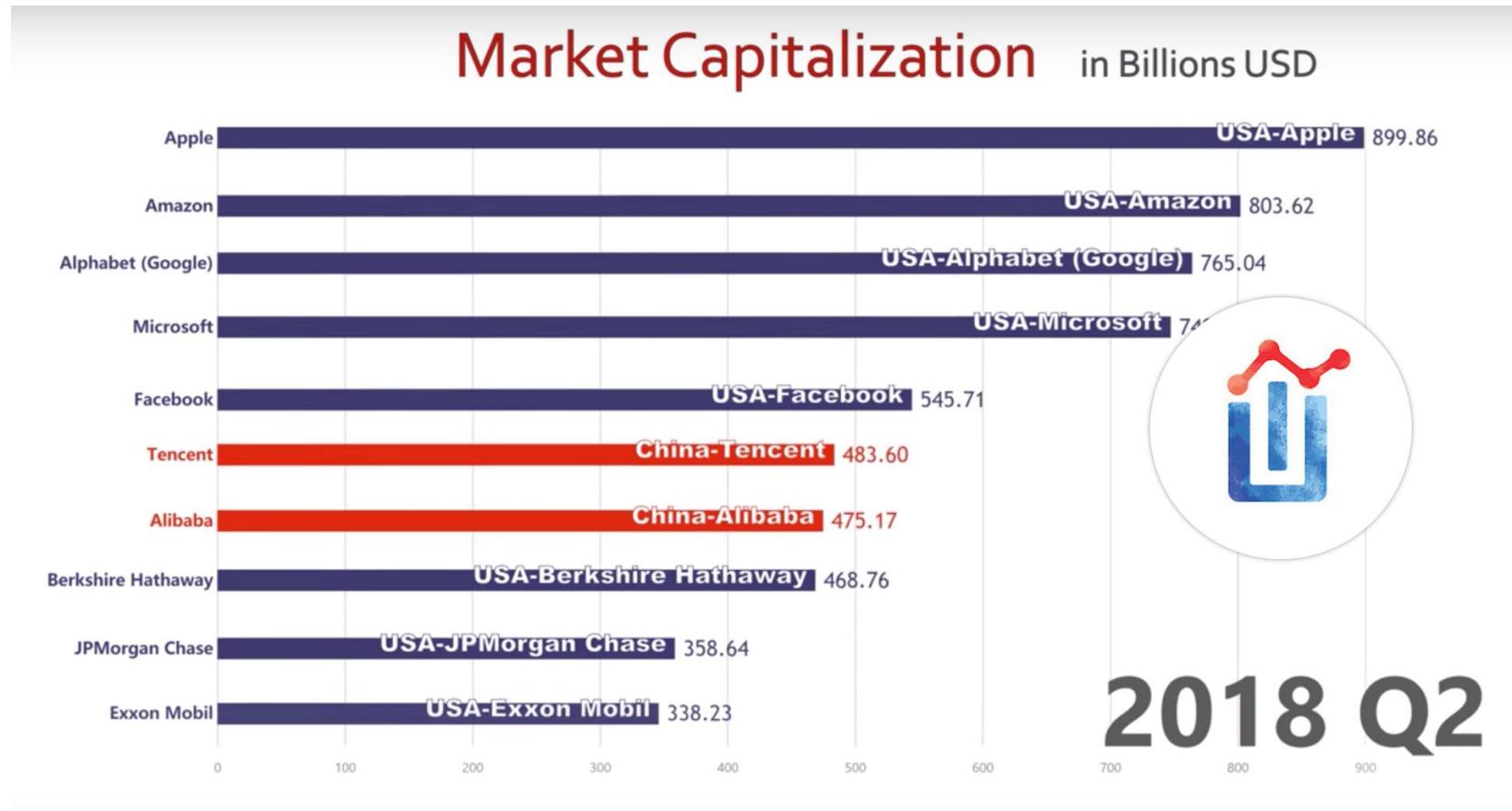
Big data in scientific discovery



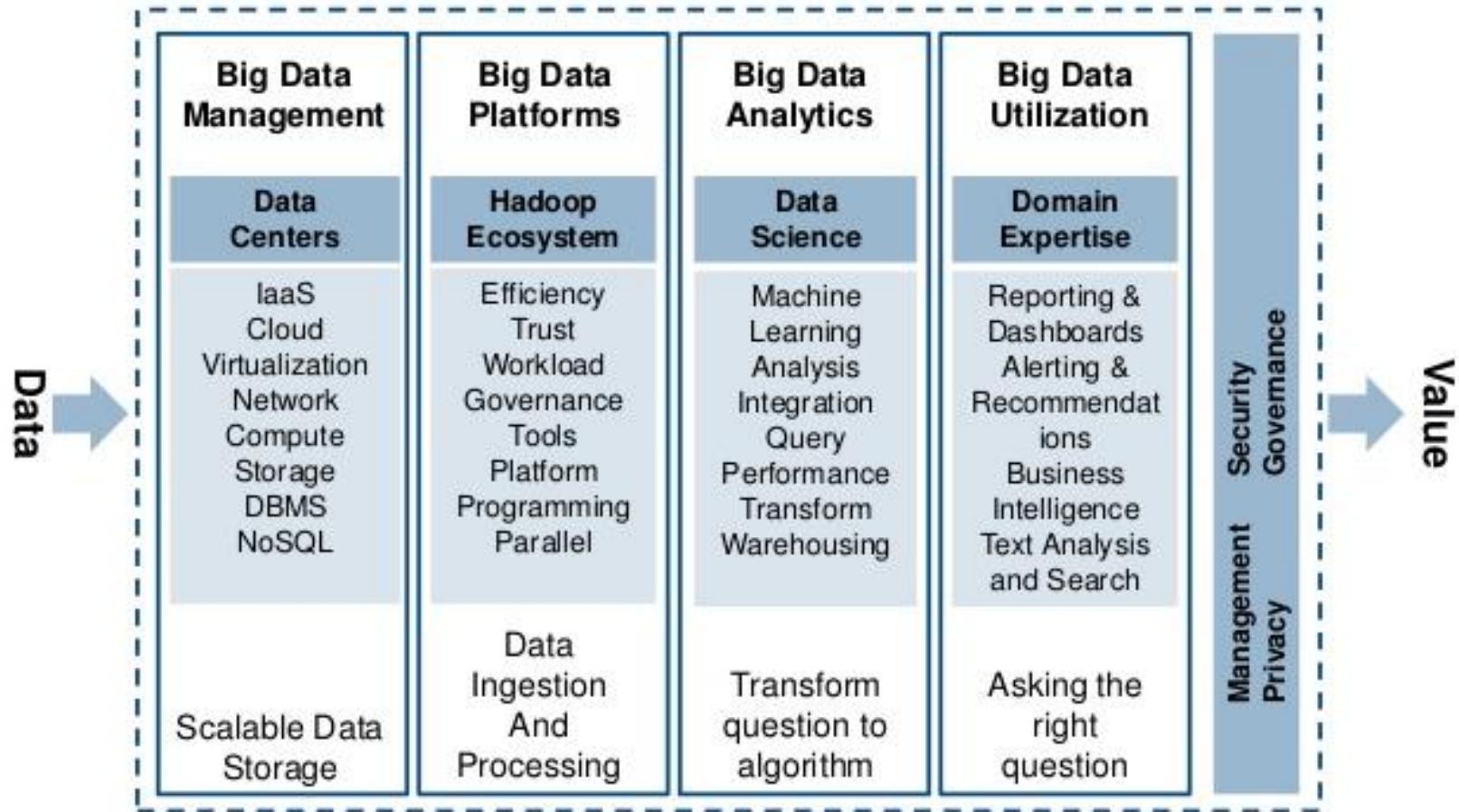
Top 10 Company Market Cap Ranking History (1998-2018)



Top 10 Company Market Cap Ranking History (1998-2018)



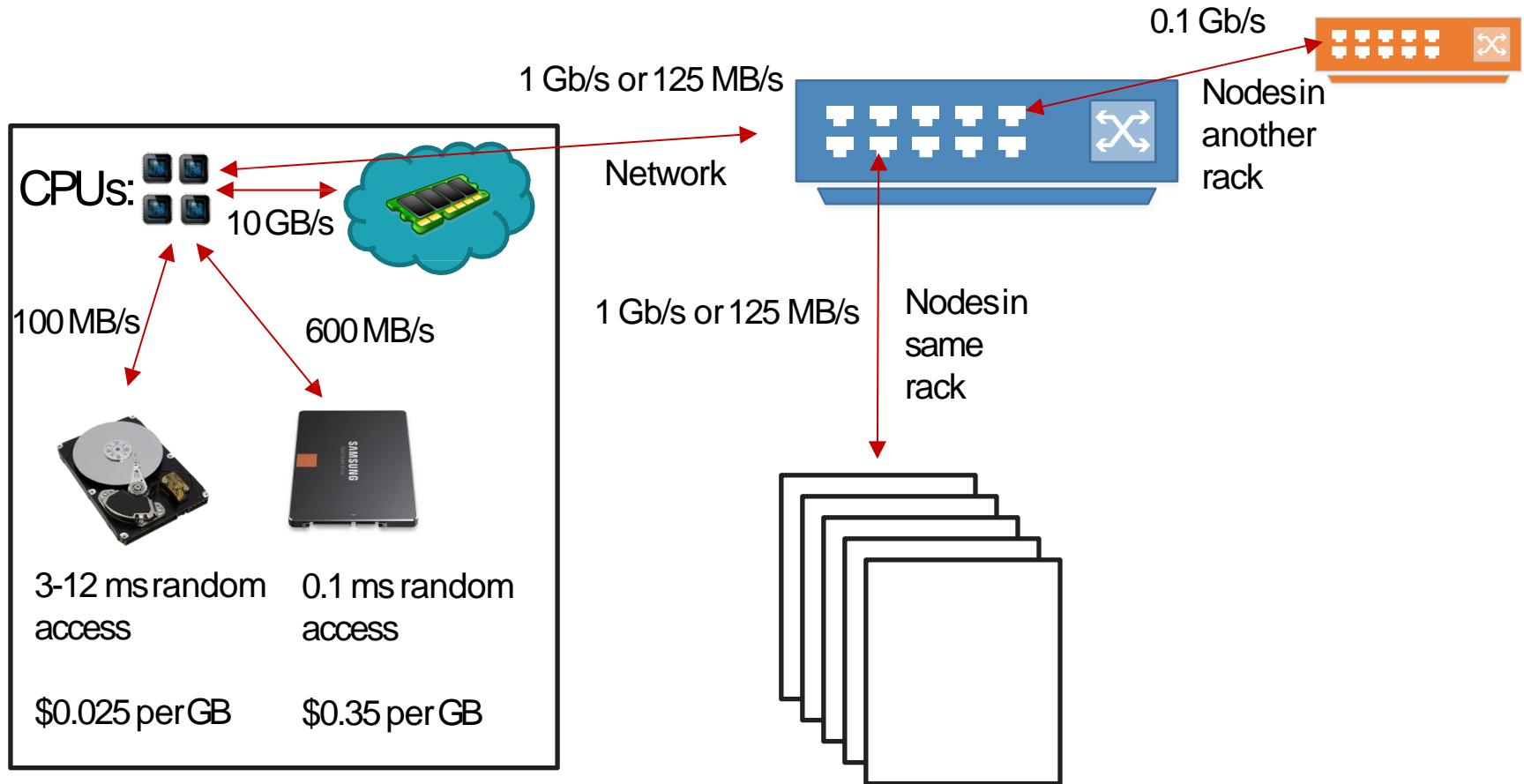
Big data technology stack



Scalable data management

- Scalability
 - Able to manage increasingly big volume of data
- Accessibility
 - Able to maintain efficiency in reading and writing data (I/O) into data storage systems
- Transparency
 - In distributed environment, users should be able to access data over the network as easily as if the data were stored locally.
 - Users should not have to know the physical location of data to access it.
- Availability
 - Fault tolerance
 - The number of users, system failures, or other consequences of distribution shouldn't compromise the availability.

Data I/O landscape



Scalable data ingestion and processing

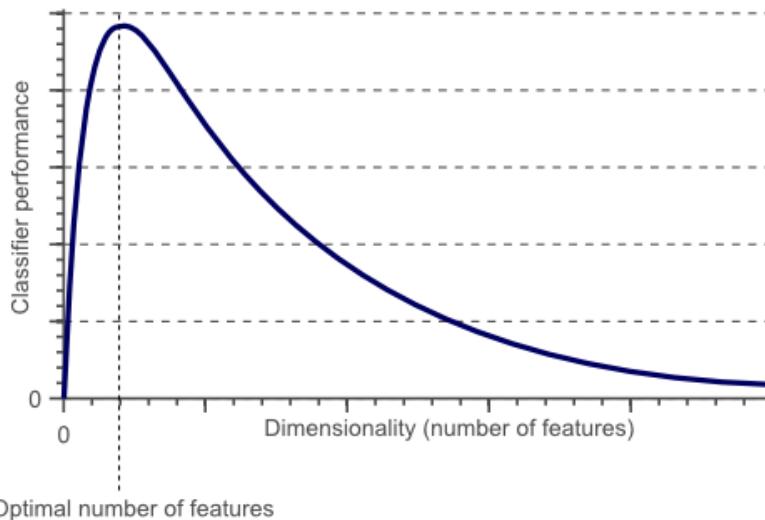
- Data ingestion
 - Data from different complementing information systems is **to be combined to gain a more comprehensive basis** to satisfy the need
 - How to ingest data efficiently from various, distributed heterogeneous sources?
 - Different data formats
 - Different data models and schemas
 - Security and privacy
- Data processing
 - How to process massive volume of data in a timely fashion?
 - How to process massive stream of data in a real-time fashion?
 - Traditional parallel, distributed processing (OpenMP, MPI)
 - Big learning curve
 - Scalability is limited
 - Fault tolerance is hard to achieve
 - Expensive, high performance computing infrastructure
 - Novel realtime processing architecture
 - Eg. Mini-batch in Spark streaming
 - Eg. Complex event processing in Apache Flink

Scalable analytic algorithms

- Challenges
 - Big volume
 - Big dimensionality
 - Realtime processing
- Scaling-up Machine Learning algorithms
 - Adapting the algorithm to handle Big Data in a single machine.
 - Eg. Sub-sampling
 - Eg. Principal component analysis
 - Eg. feature extraction and feature selection
 - Scaling-up algorithms by parallelism
 - Eg. k-nn classification based on MapReduce
 - Eg. scaling-up support vector machines (SVM) by a divide and-conquer approach

Eg. Curse of dimensionality

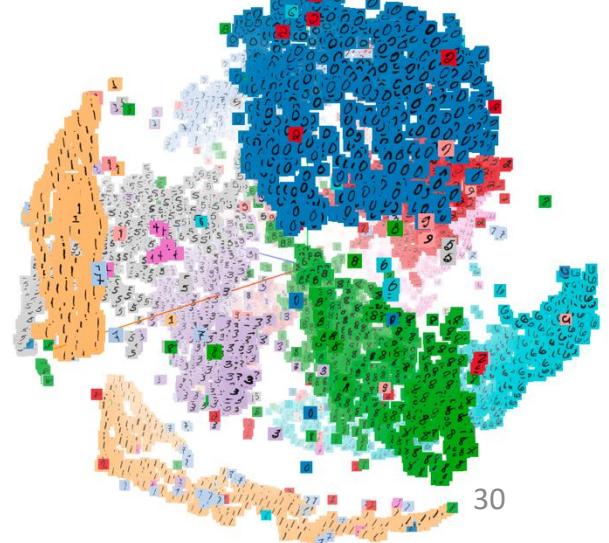
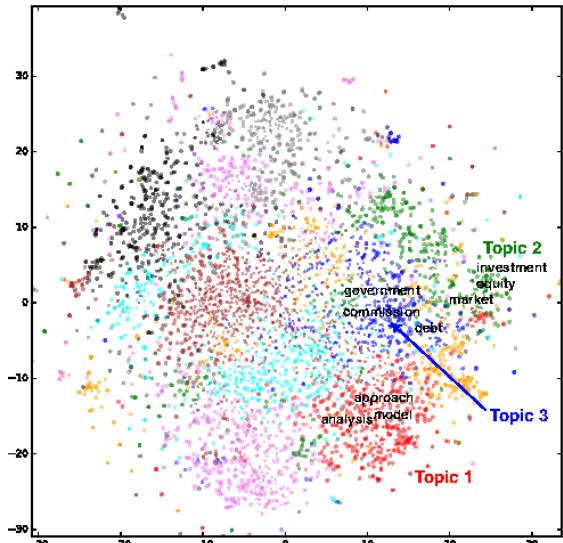
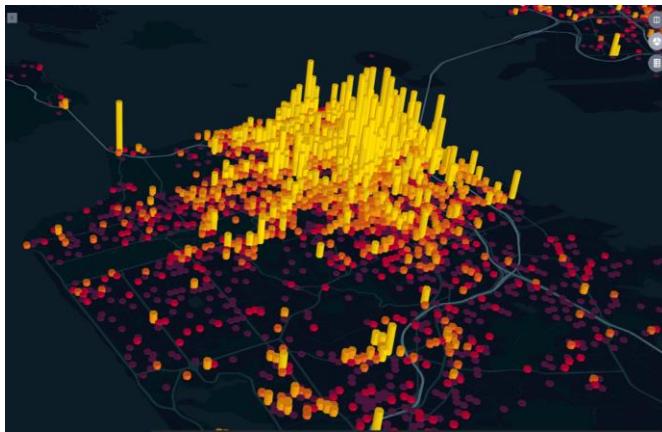
- The required number of samples (to achieve the same accuracy) grows exponentially with the number of variables!
- In practice: number of training examples is fixed!
 - => the classifier's performance usually will degrade for a large number of features!



In fact, after a certain point, increasing the dimensionality of the problem by adding new features would actually degrade the performance of classifier.

Utilization and interpretability of big data

- Domain expertise to findout problems and interprete analytics results
- Scalable visualization and interpretability of million data points
 - to facilitate their interpretability and understanding

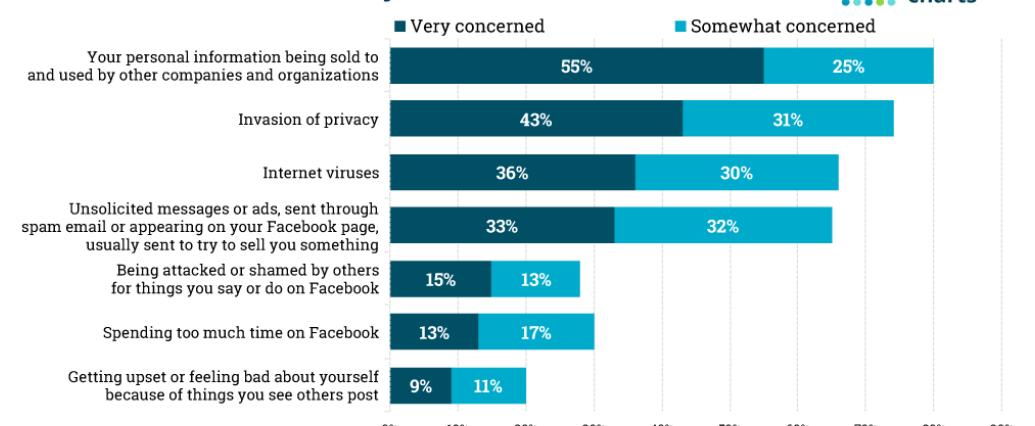


Privacy and security

FTC Settlement with Facebook	
	\$5,000,000,000 Unprecedented penalty
	New privacy structure at Facebook
	New tools for FTC to monitor Facebook

Source: Federal Trade Commission | FTC.gov

Facebook Users' Privacy Concerns



Published on MarketingCharts.com in April 2018 | Data Source: Gallup

Based on telephone interviews conducted April 2-8, 2018 among 1,509 US adults ages 18 and older, of whom 785 are Facebook users.

The remaining respondents answered "Not too concerned" or "Not concerned at all."



SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

How was Facebook users' data misused?

- 1 In 2014 a Facebook quiz invited users to find out their personality type
- 2 The app collected the data of those taking the quiz, but also recorded the public data of their friends
- 3 About 305,000 people installed the app, but it gathered information on up to 87 million people, according to Facebook
- 4 It is claimed at least some of the data was sold to Cambridge Analytica (CA) which used it to psychologically profile voters in the US
- 5 CA denies it broke any laws and says it did not use the data in the US presidential election
- 6 Facebook sends notices to users telling them whether their data was breached

CA denies any wrongdoing. Facebook has apologised to users and says a "breach of trust" has occurred.

Big data job trends



Talent shortage in big data

Talent Demand-Supply gap analysis

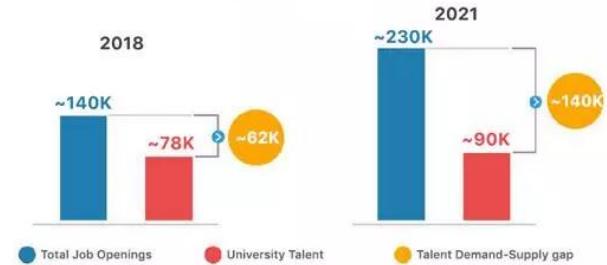
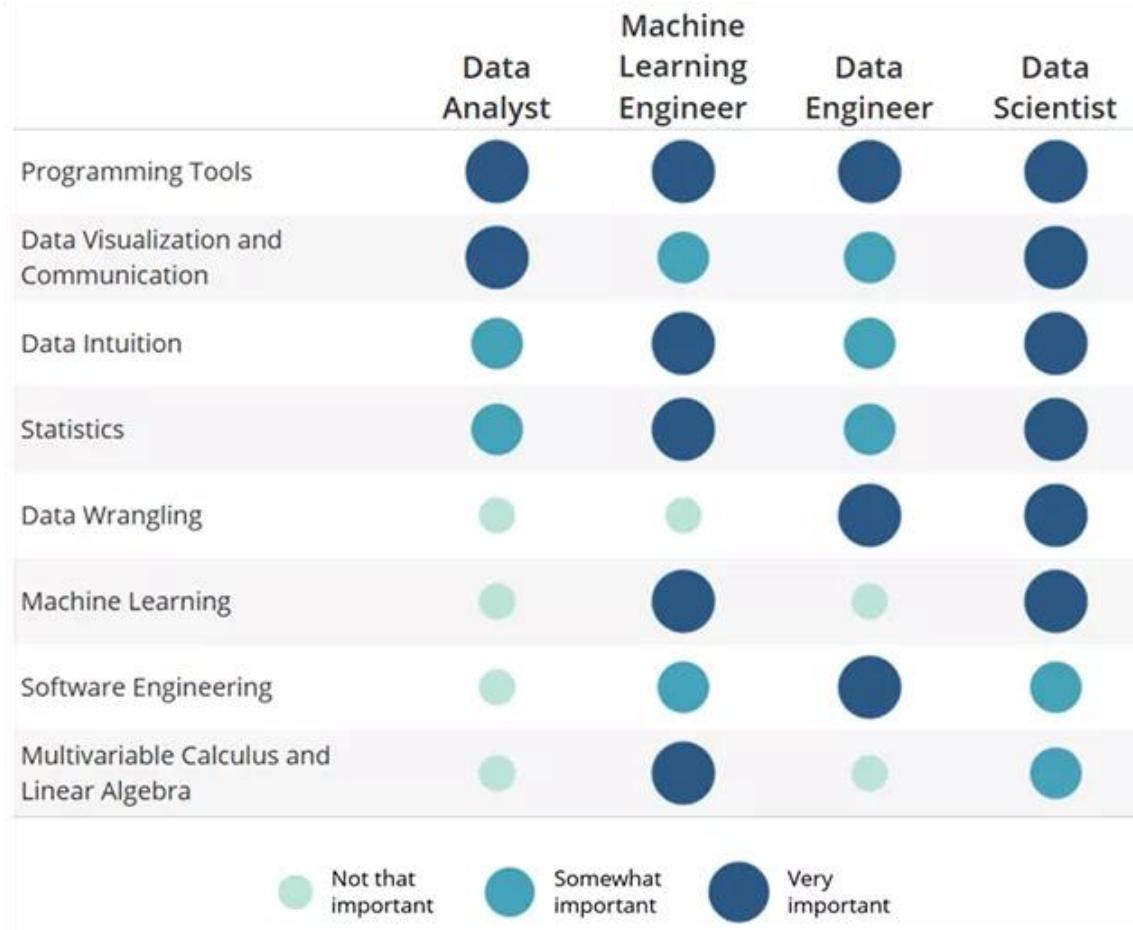


Table 2. Summary Demand Statistics

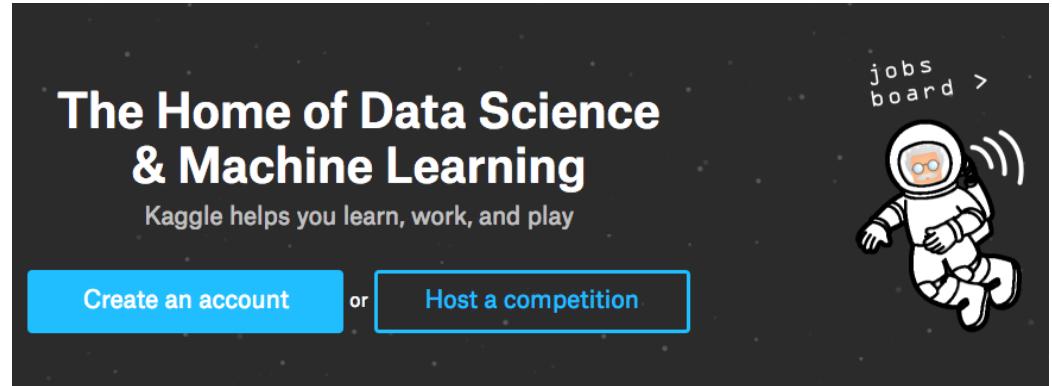
DSA Framework Category	Number of Postings in 2015	Projected 5-Year Growth	Estimated Postings for 2020	Average Time to Fill (Days)	Average Annual Salary
All	2,352,681	15%	2,716,425	45	\$80,265
Data-Driven Decision Makers	812,099	14%	922,428	48	\$91,467
Functional Analysts	770,441	17%	901,743	40	\$69,162
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

Big data skill set

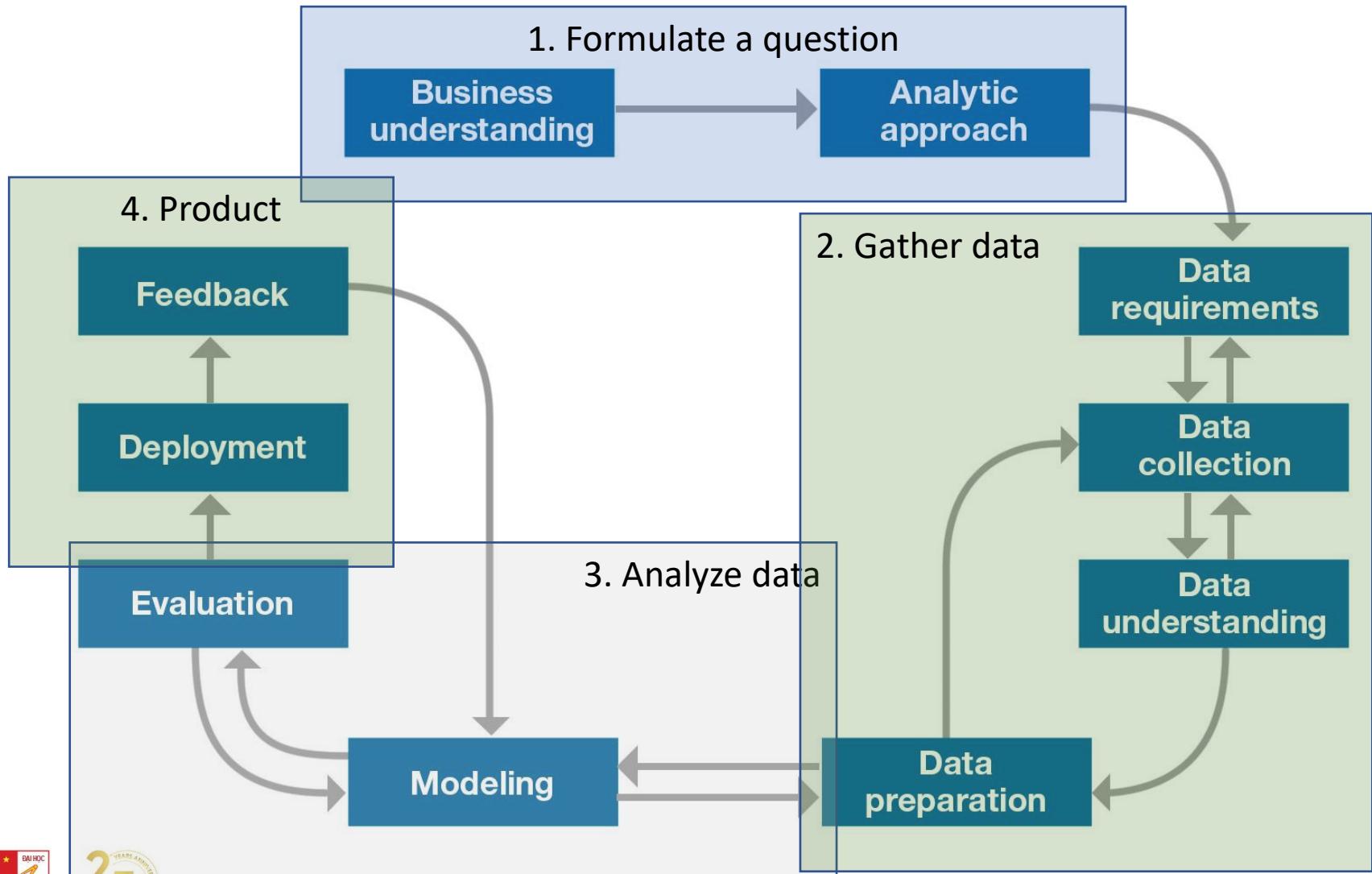


How to land big data related jobs

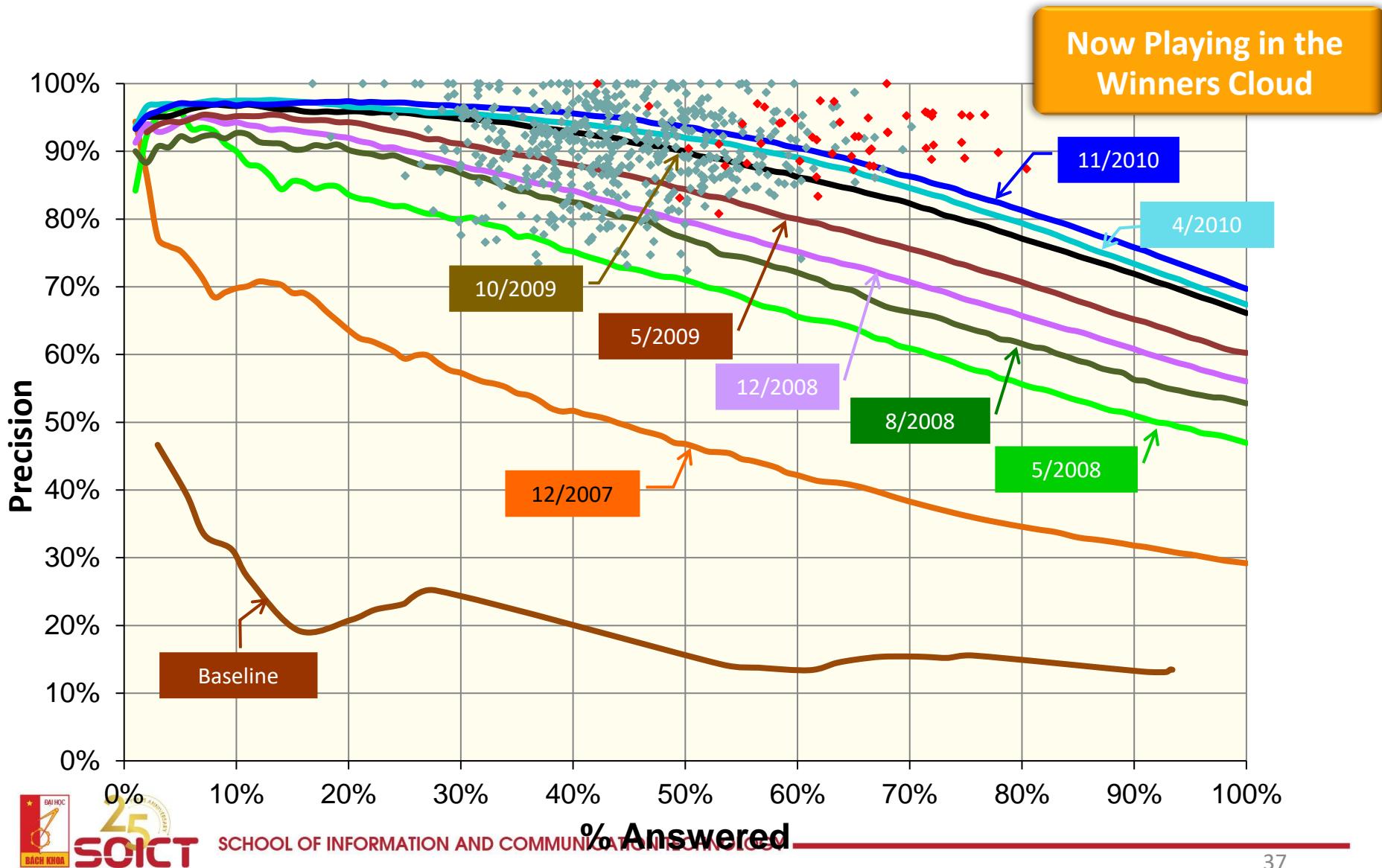
- Learn to code
 - Coursera
 - Udacity
 - Freecodecamp
 - Codecademy
- Math, Stats and machine learning
 - Kaggle
- Hadoop, NoSQL, Spark
- Visualization and Reporting
 - Tableau
 - Pentaho
- Meetup & Share
- Find a mentor
- Internships, projects



Data science method

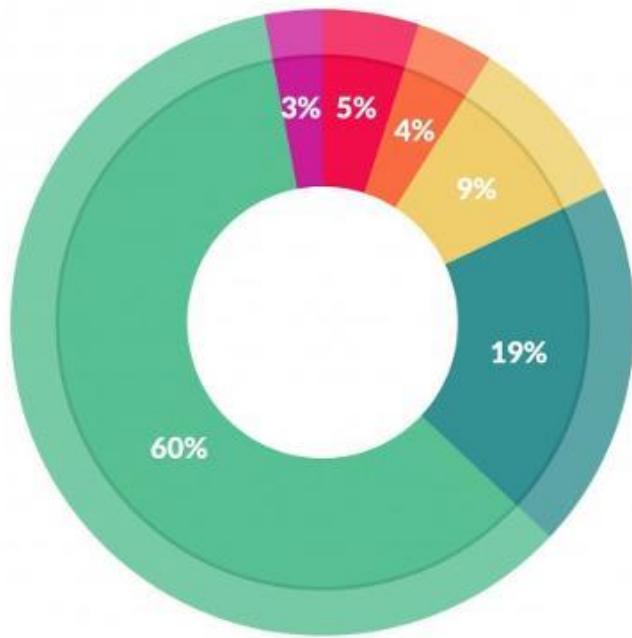


DeepQA: Incremental Progress in Precision and Confidence 6/2007-11/2010



Cleaning big data: most time-consuming, least enjoyable data science task

- Data preparation accounts for about 80% of the work of data scientists



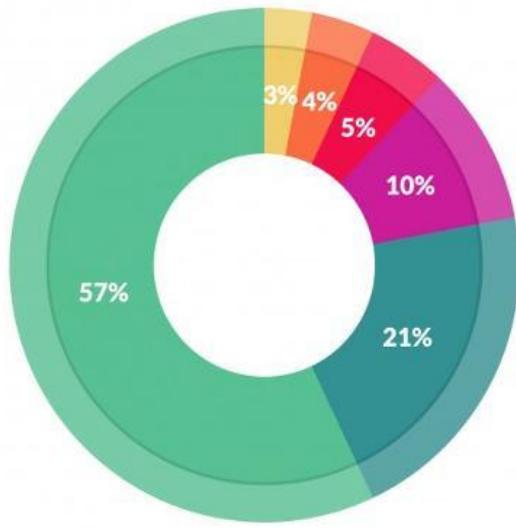
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

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

Cleaning big data: most time-consuming, least enjoyable data science task

- 57% of data scientists regard cleaning and organizing data as the least enjoyable part of their work and 19% say this about collecting data sets.



What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

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, Srinivas. 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 seamlessly scalable 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>



25
YEARS ANNIVERSARY
SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you
for your
attention!!!

