

25 YEARS ANNIVERSARY  
SOICT

ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

# Chương 1

## Tổng quan về lưu trữ và xử lý dữ liệu lớn

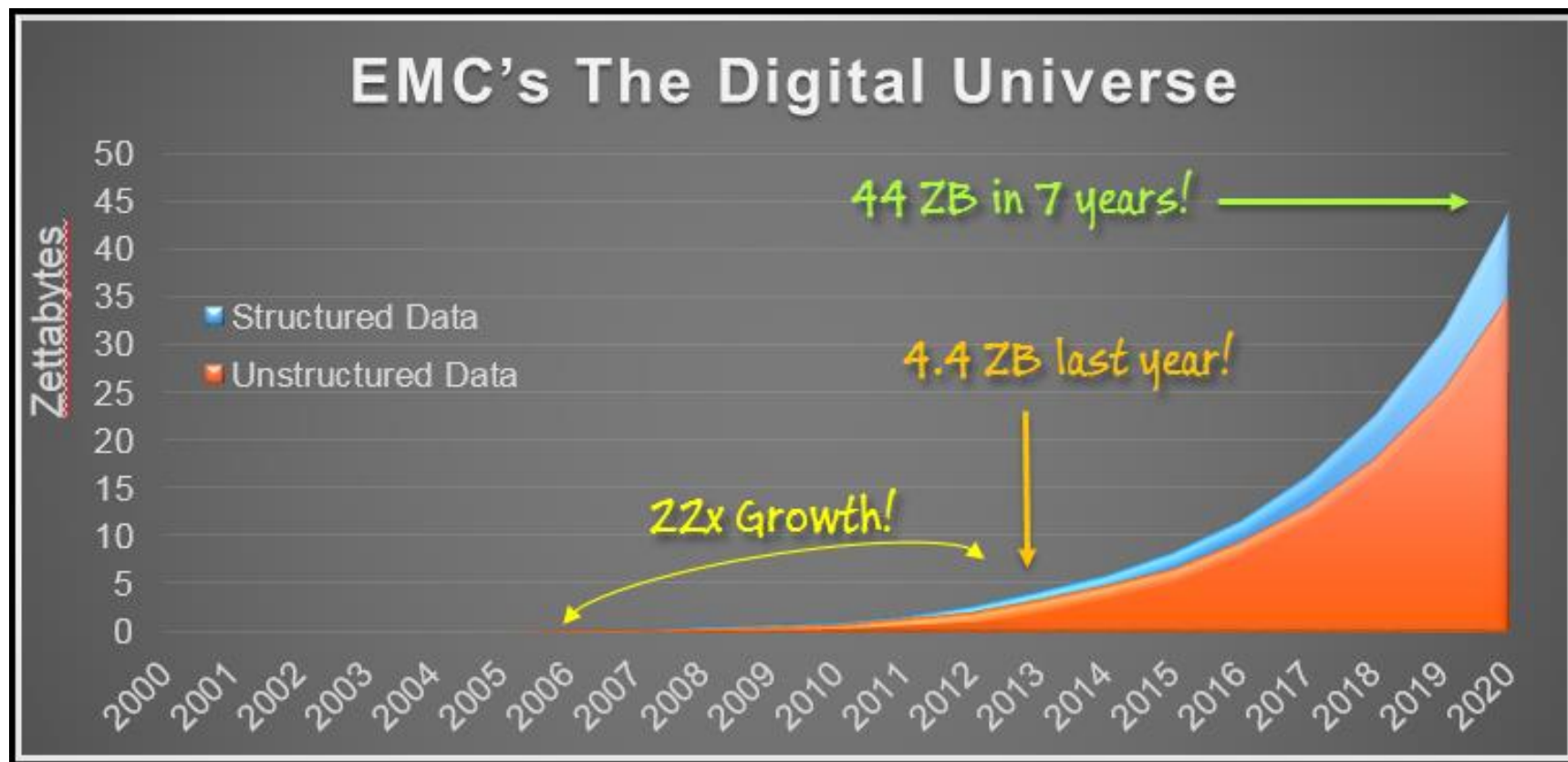
# Thông tin chung về môn học

Tên học phần:	Lưu trữ và xử lý dữ liệu lớn (Big data storage and processing)
Mã số học phần:	IT4931
Khối lượng:	3(3-1-0-6) <ul style="list-style-type: none"><li>– Lý thuyết: 45 tiết</li><li>– BTL: 15 tiết</li><li>– Thí nghiệm: 0 tiết</li></ul>

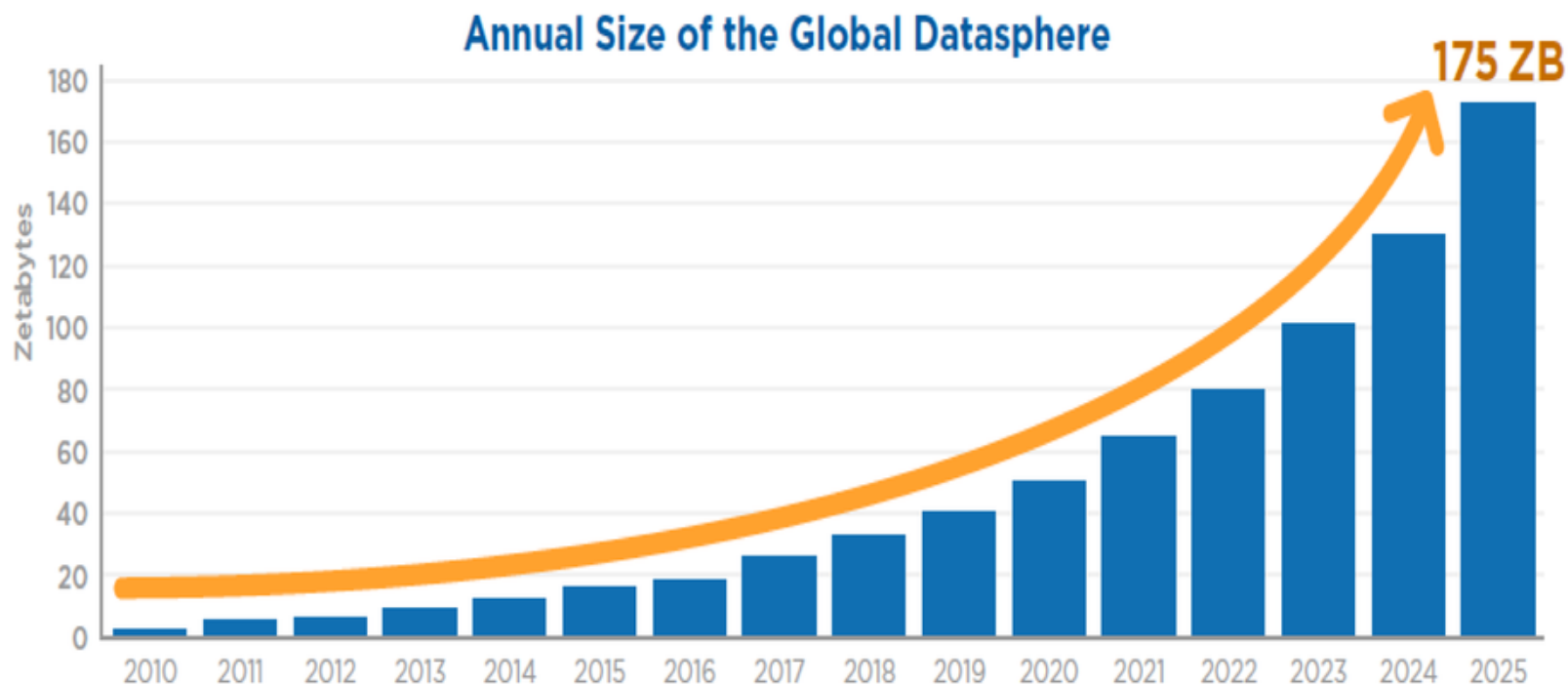
# Đề cương học tập

STT	Bài giảng
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 kĩ thuật xử lý dữ liệu lớn theo khối - phần 1 Map Reduce
9	Các kĩ thuật xử lý dữ liệu lớn theo khối - phần 2 Apache Spark
10	Các kĩ 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

# Tổng dung lượng dữ liệu 2020



# Tổng dung lượng dữ liệu 2025



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

# Hình dung về độ lớn của dữ liệu



If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon\*

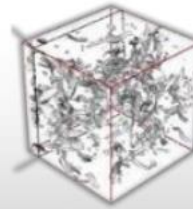
By 2020, there would be 6.6 stacks from the Earth to the Moon\*



# Khoa học dữ liệu: Bước phát triển thứ 4 của khoa học khám phá



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



## Experimental

Thousand  
years ago

*Description of natural  
phenomena*

## Theoretical

Last few  
hundred years

*Newton's laws,  
Maxwell's equations...*

## Computational

Last  
few decades

*Simulation of  
complex phenomena*

## The Fourth Paradigm

Today and the  
Future

*Unify theory, experiment  
and simulation with  
large multidisciplinary  
Data*

*Using data exploration  
and data mining  
(from instruments,  
sensors, humans...)*

*Distributed Communities*



# Nói về dữ liệu lớn năm 2008

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

September 2008



# Nói về dữ liệu lớn năm 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



# Dữ liệu lớn ngày nay



**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**

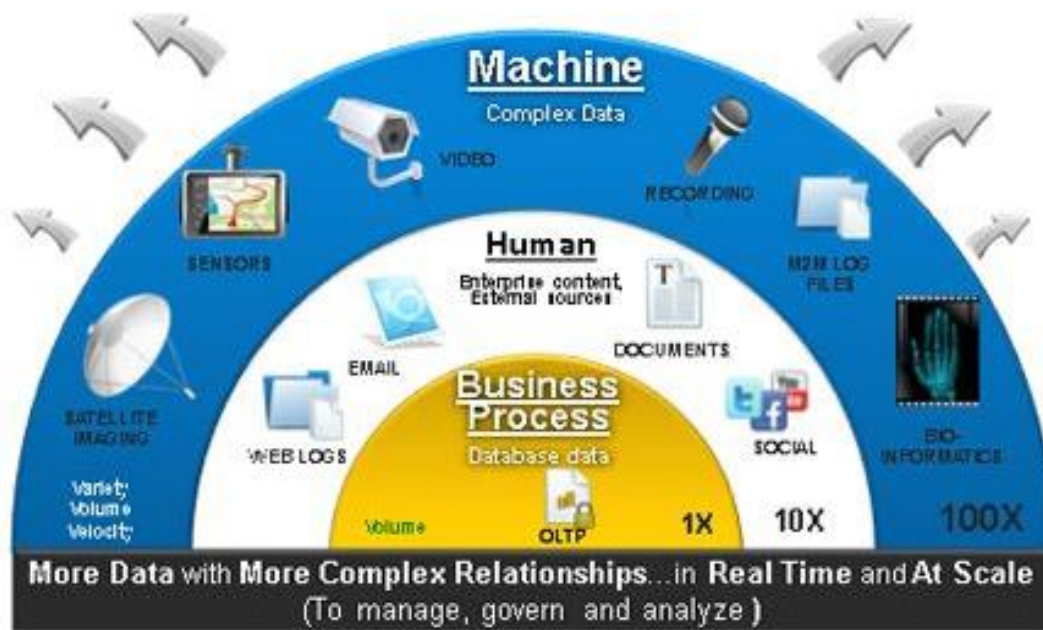
# Những con số về tốc độ sinh dữ liệu

## 2020 *This Is What Happens In An Internet Minute*



# Các nguồn tạo ra dữ liệu lớn

- Thương mại điện tử
- Mạng xã hội
- Internet vạn vật (IoT)
- Các thử nghiệm dữ liệu lớn (tin sinh học, vật lý lượng tử, vvv)

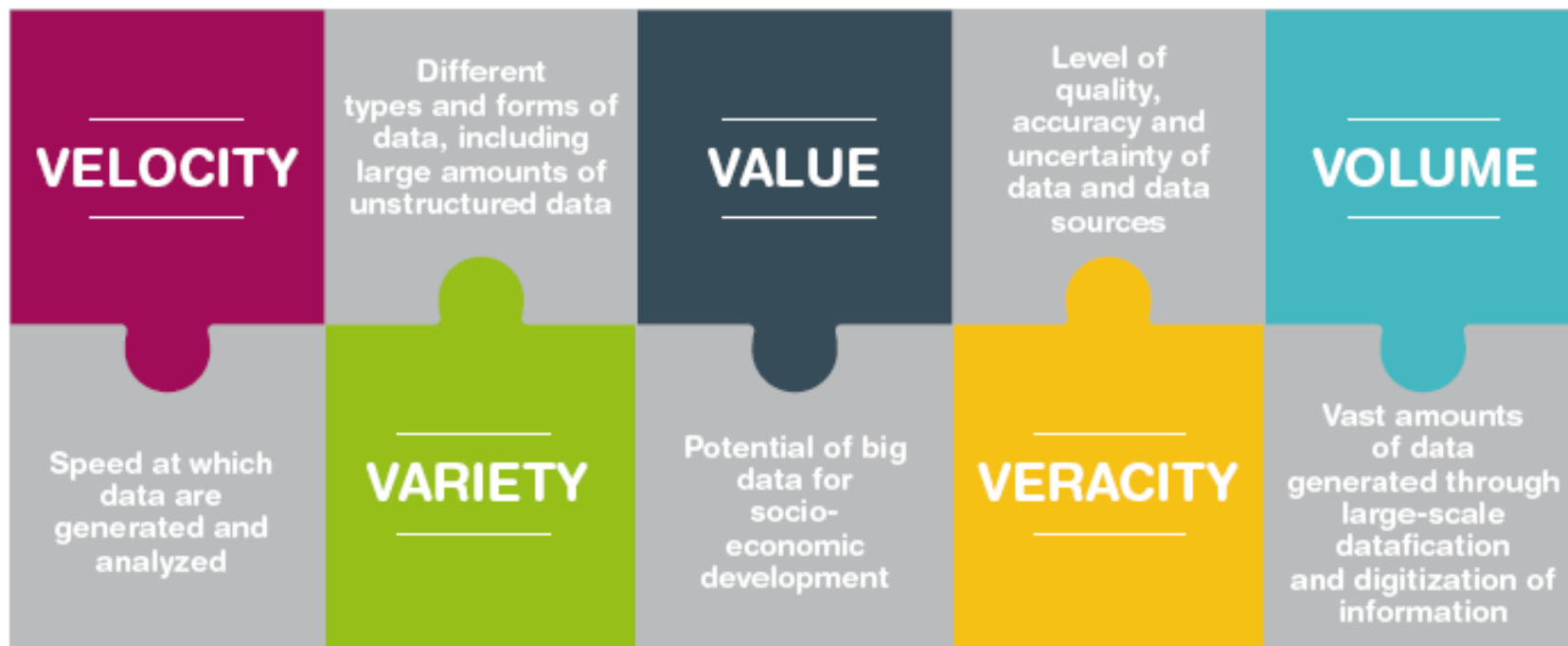




# Dữ liệu được ví như nguồn tài nguyên dầu mỏ mới



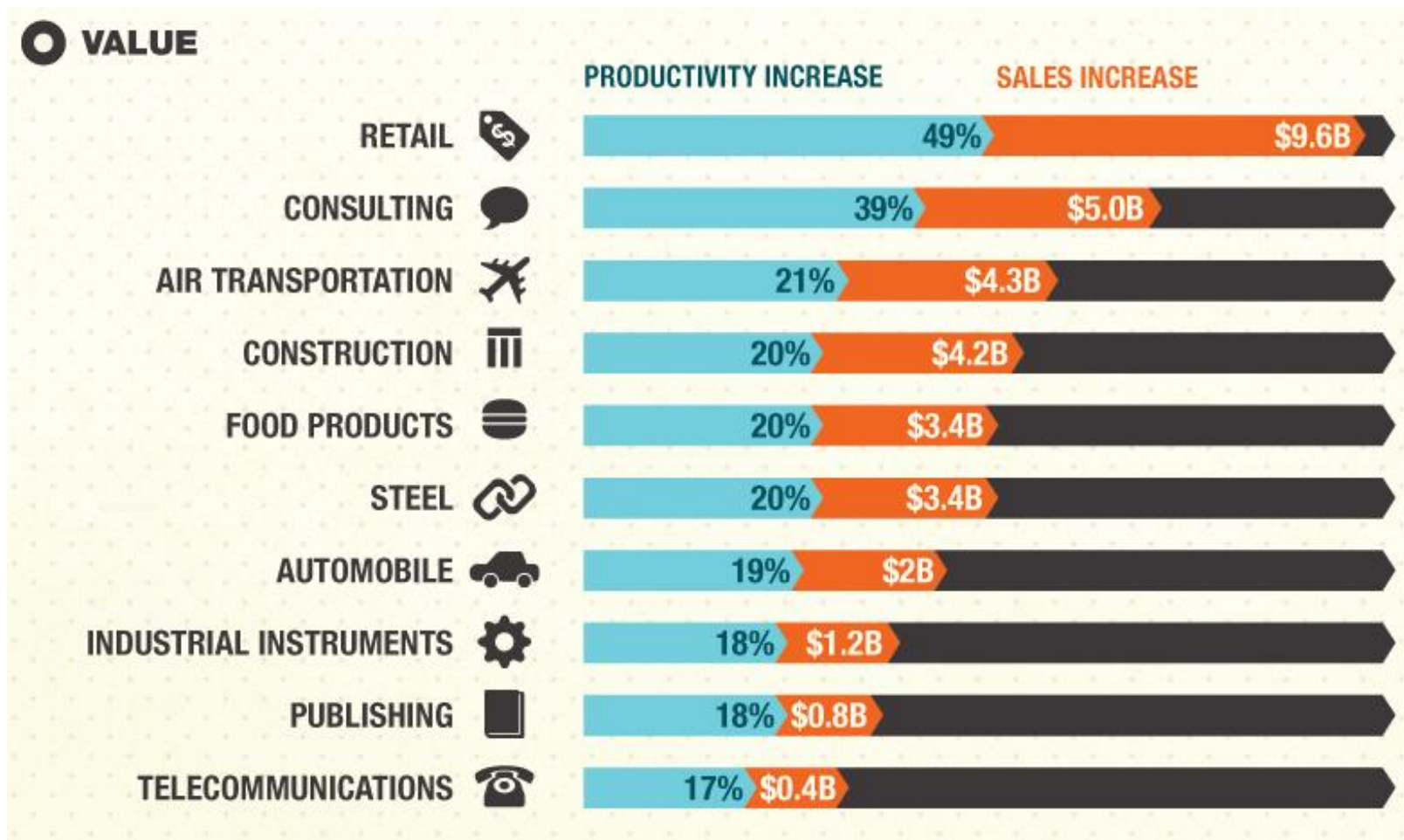
# Đặc điểm 5'V của dữ liệu lớn



Dữ liệu lớn là tập dữ liệu quá lớn hoặc là quá phức tạp mà các nền tảng lưu trữ và xử lý dữ liệu truyền thống không đáp ứng được.



# Dữ liệu lớn – giá trị mang lại lớn



# Khai thác dữ liệu lớn trong giáo dục

- Chương trình học tối ưu, tùy biến phù hợp cho người học
- Cải tiến tài liệu, giáo trình phù hợp
- Đánh giá học tập
- Khuyến nghị lộ trình học tập, sự nghiệp



# Một vài ví dụ

- Coursera
- VioEdu
- <https://byjus.com/>
  - Bài giảng video cá nhân hoá
  - Phân tích tiến độ học tập
  - Các câu hỏi kiểm tra quá trình cá nhân hoá



# Khai thác dữ liệu lớn trong khoa học chăm sóc sức khỏe

- Giảm chi phí điều trị, các xét nghiệm dư thừa
- Dự đoán quy mô đại dịch, khuyến nghị các biện pháp ứng phó
- Ngăn ngừa sớm các bệnh có thể gặp trong tương lai



# Khai thác dữ liệu lớn trong quản lý nhà nước

- Các chương trình phúc lợi xã hội
  - Nắm bắt nhanh chóng các vấn đề xã hội (việc làm, tội phạm, môi trường, vvv)
  - Khuyến nghị các biện pháp đối phó
- An ninh thông tin
  - Trốn thuế
  - Lừa đảo





# Khai thác dữ liệu lớn trong công nghiệp truyền thông và giải trí

- Dự đoán sở thích của tập khách hàng
- Cá nhân hoá theo sở thích
- Truyền thông bám đuổi mục tiêu hiệu quả
- Ví dụ
  - Spotify, Netflix



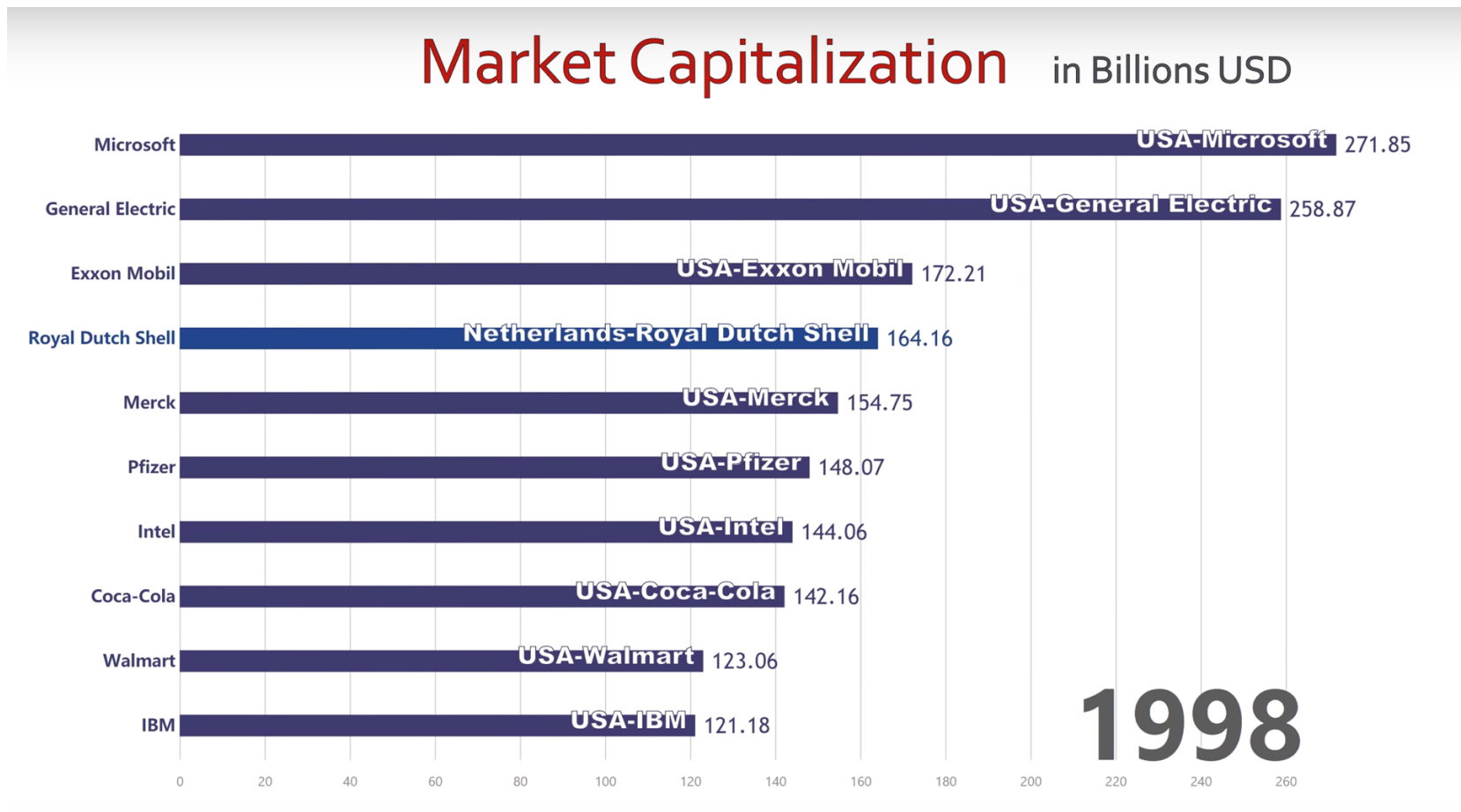
# Khai thác dữ liệu lớn trong khoa học khám phá



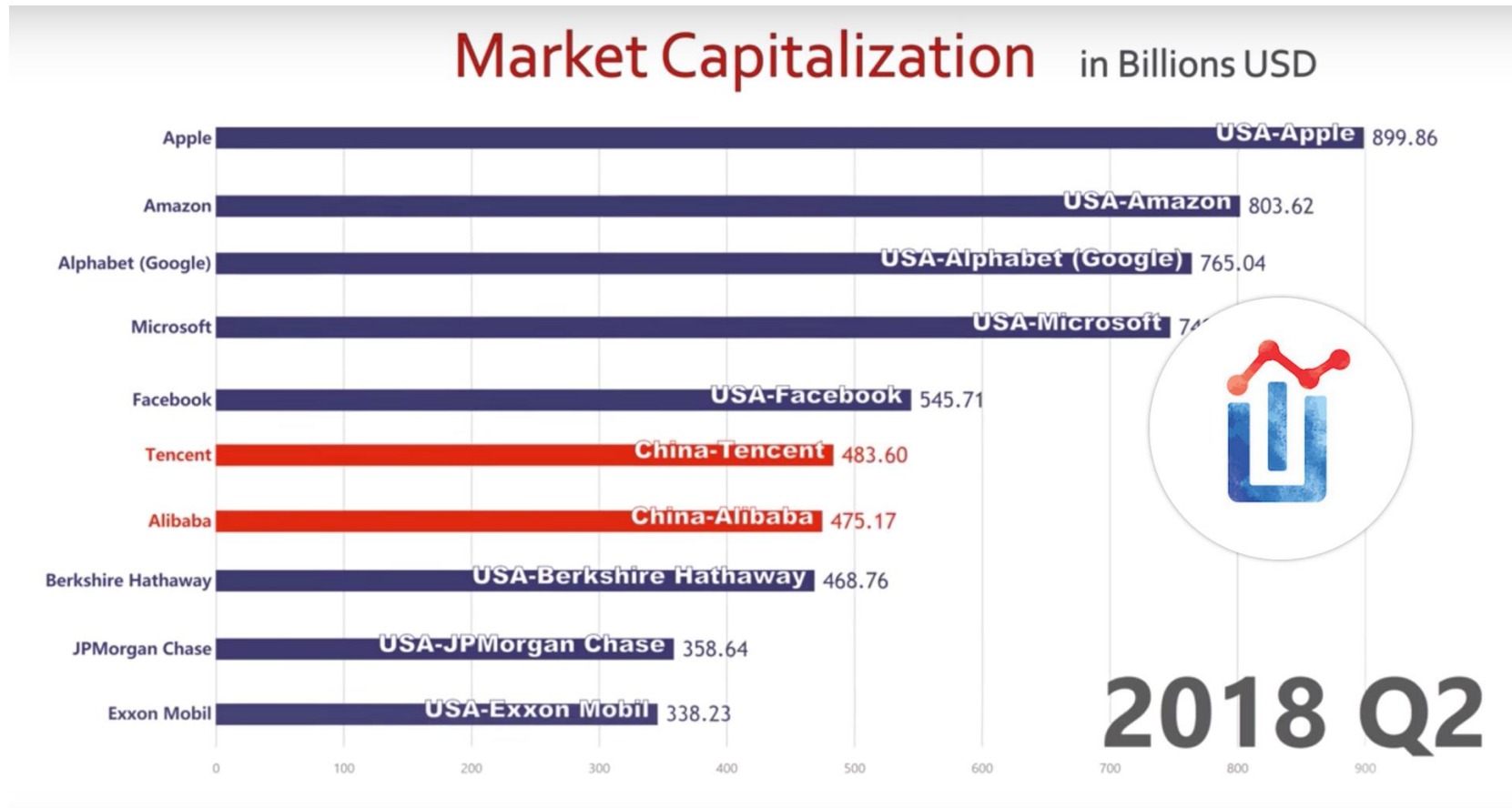
CERN's Large Hydron Collider (LHC) generates 15 PB a year



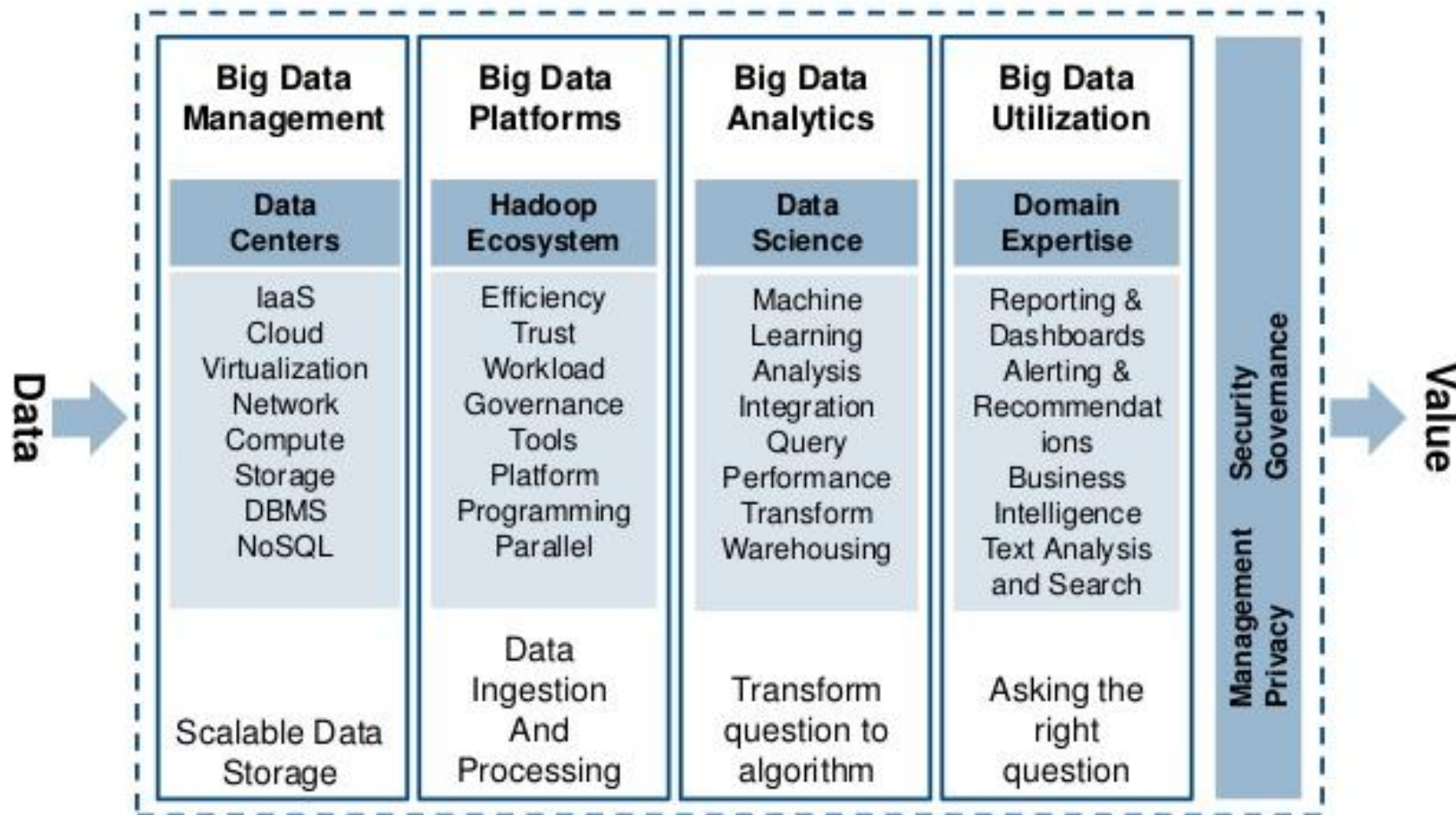
# 10 công ty lớn nhất (1998-2018)



# 10 công ty lớn nhất (1998-2018)



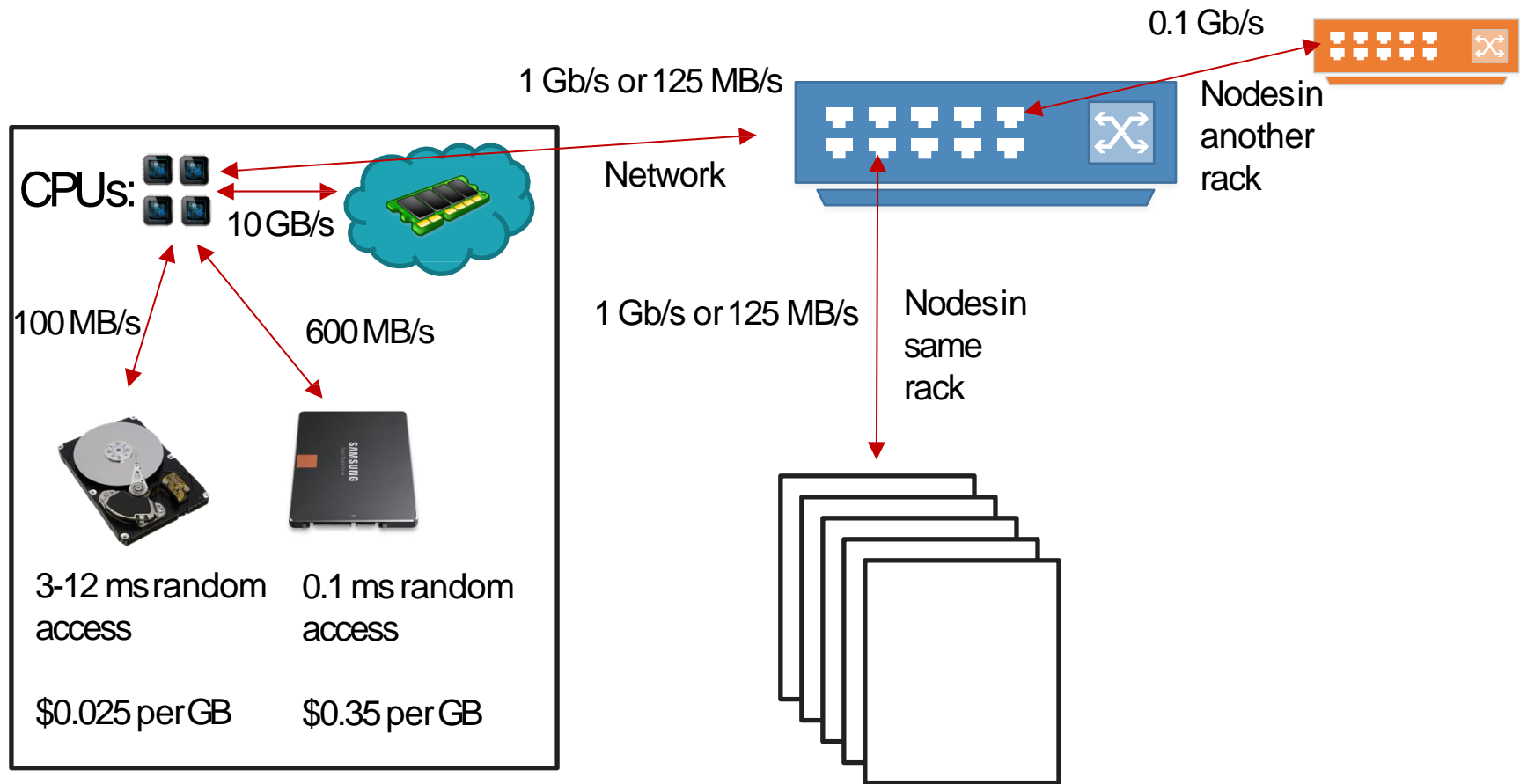
# Các tầng công nghệ cho dữ liệu lớn



# Quản lý dữ liệu phải khả mở

- Scalability
  - Khả năng quản lý lượng dữ liệu lớn không ngừng tăng lên theo thời gian.
- Accessibility
  - Cho phép đọc ghi I/O dữ liệu hiệu quả.
- Transparency
  - Truy cập dữ liệu dễ dàng, vị trí lưu trữ dữ liệu trên hệ thống là trong suốt với người dùng cuối.
- Availability
  - Khả năng chống chịu lỗi, khi tăng số lượng người dùng, khi hỏng hóc.

# Tốc độ I/O phổ biến



# Xử lý và tích hợp dữ liệu phải khả mở

- Tích hợp dữ liệu
  - Dữ liệu có định dạng khác nhau
  - Dữ liệu tồn tại ở các mô hình và lược đồ dữ liệu khác nhau
  - Các vấn đề liên quan đến an toàn an ninh thông tin, quyền riêng tư
- Xử lý dữ liệu
  - Xử lý khối lượng dữ liệu rất lớn
  - Xử lý luồng dữ liệu lớn
  - Xử lý dữ liệu song song, phân tán truyền thống (OpenMP, MPI)
    - Phức tạp, khó học
    - Khả năng khả mở có giới hạn
    - Cơ chế chịu lỗi kém
    - Chi phí hạ tầng đắt đỏ
  - Kiến trúc xử lý dữ liệu luồng dữ liệu lớn
    - Spark mini-batch
    - Apache Flink

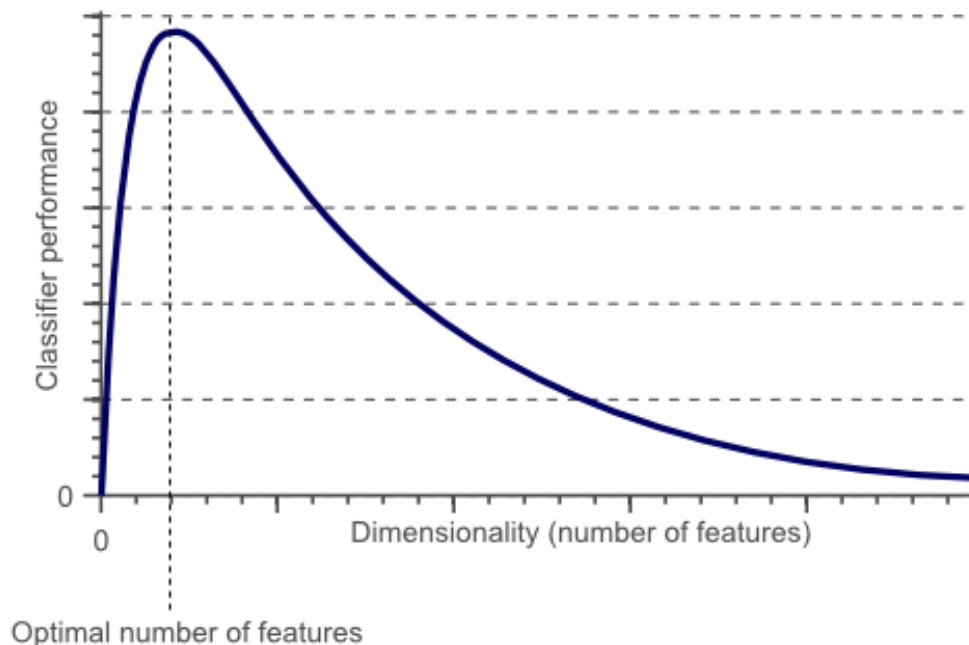
# Các giải thuật phân tích dữ liệu khả mở

- Làm nhỏ lại dữ liệu cho phù hợp với các giải thuật truyền thống
  - Eg. Sub-sampling
  - Eg. Principal component analysis
  - Eg. Feature extraction and feature selection
- Song song hoá các giải thuật học máy
  - Eg. k-nn classification based on MapReduce
  - Eg. scaling-up support vector machines (SVM) by a divide and-conquer approach



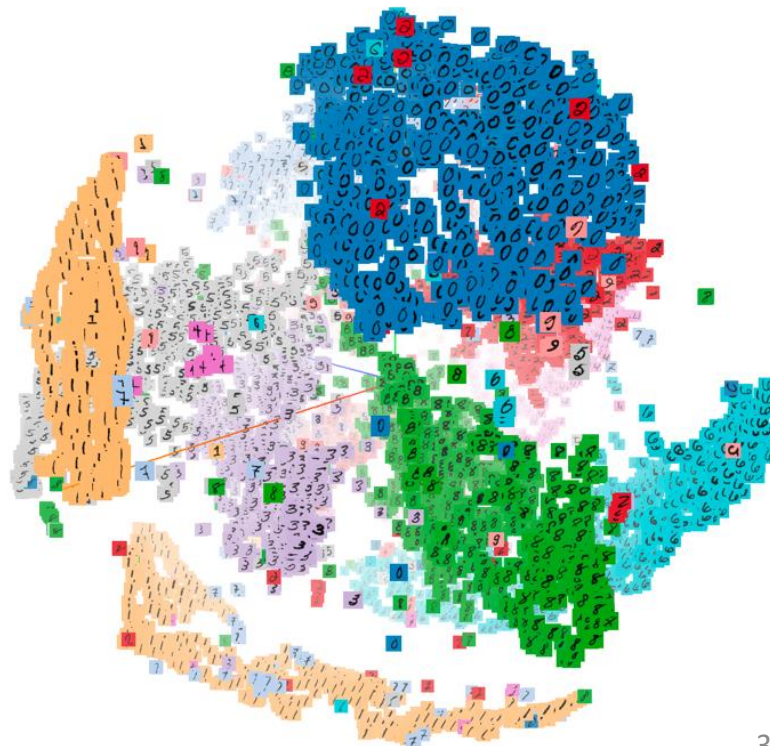
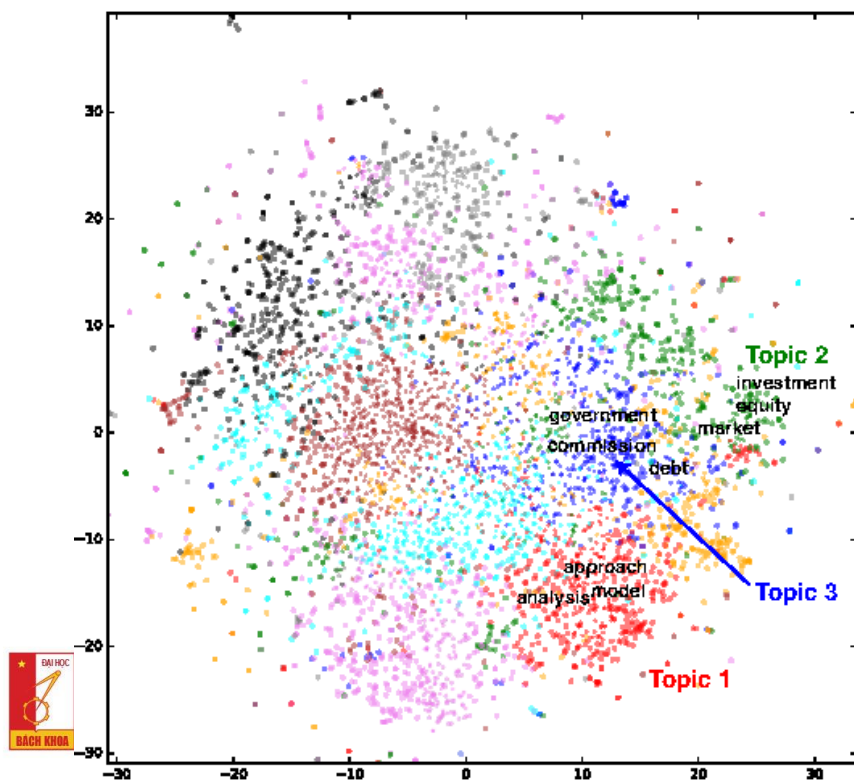
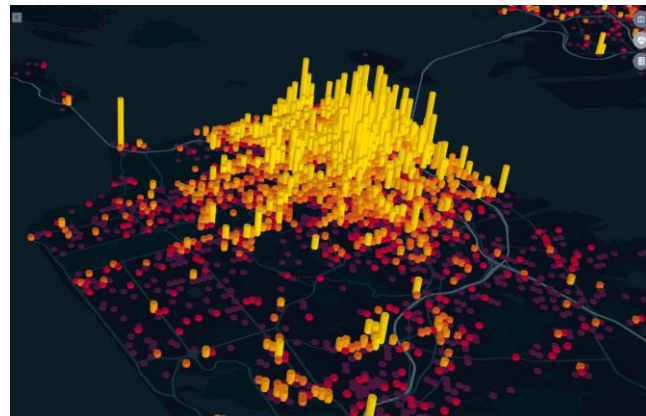
# Eg. Sự bùng nổ số chiều trong dữ liệu (Curse of dimensionality)

- Số lượng mẫu cần cho mô hình học tăng lên khi số chiều dữ liệu tăng
- Trong thực tiễn: Số lượng mẫu để học thường cố định
  - => Độ chính xác của mô hình giảm khi tăng số chiều trong dữ liệu học



# Sử dụng và trực quan hoá dữ liệu lớn

- Cần kiến thức chuyên gia
- Cần kỹ thuật và công cụ để hỗ trợ hiệu quả việc trình diễn và hiểu về dữ liệu lớn

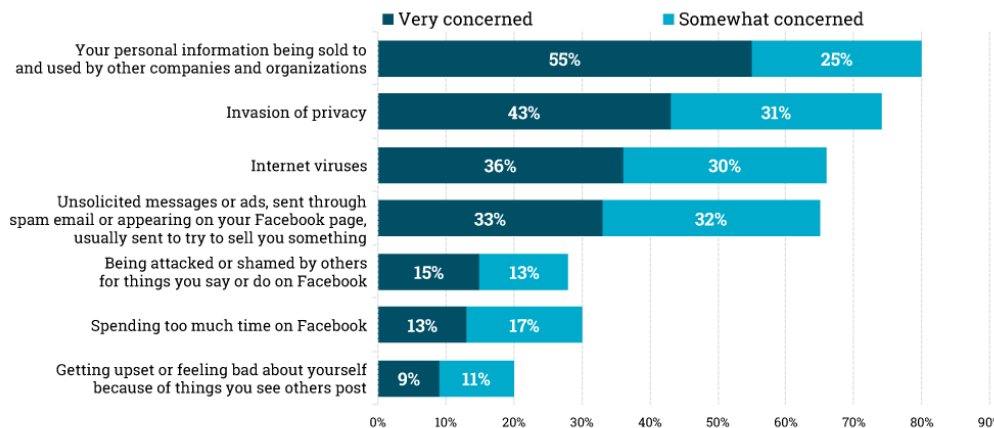


# Bảo mật và quyền riêng tư



## Facebook Users' Privacy Concerns


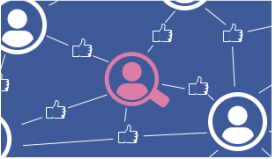


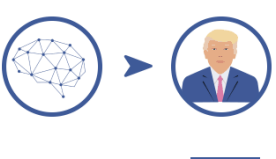

marketing  
charts



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

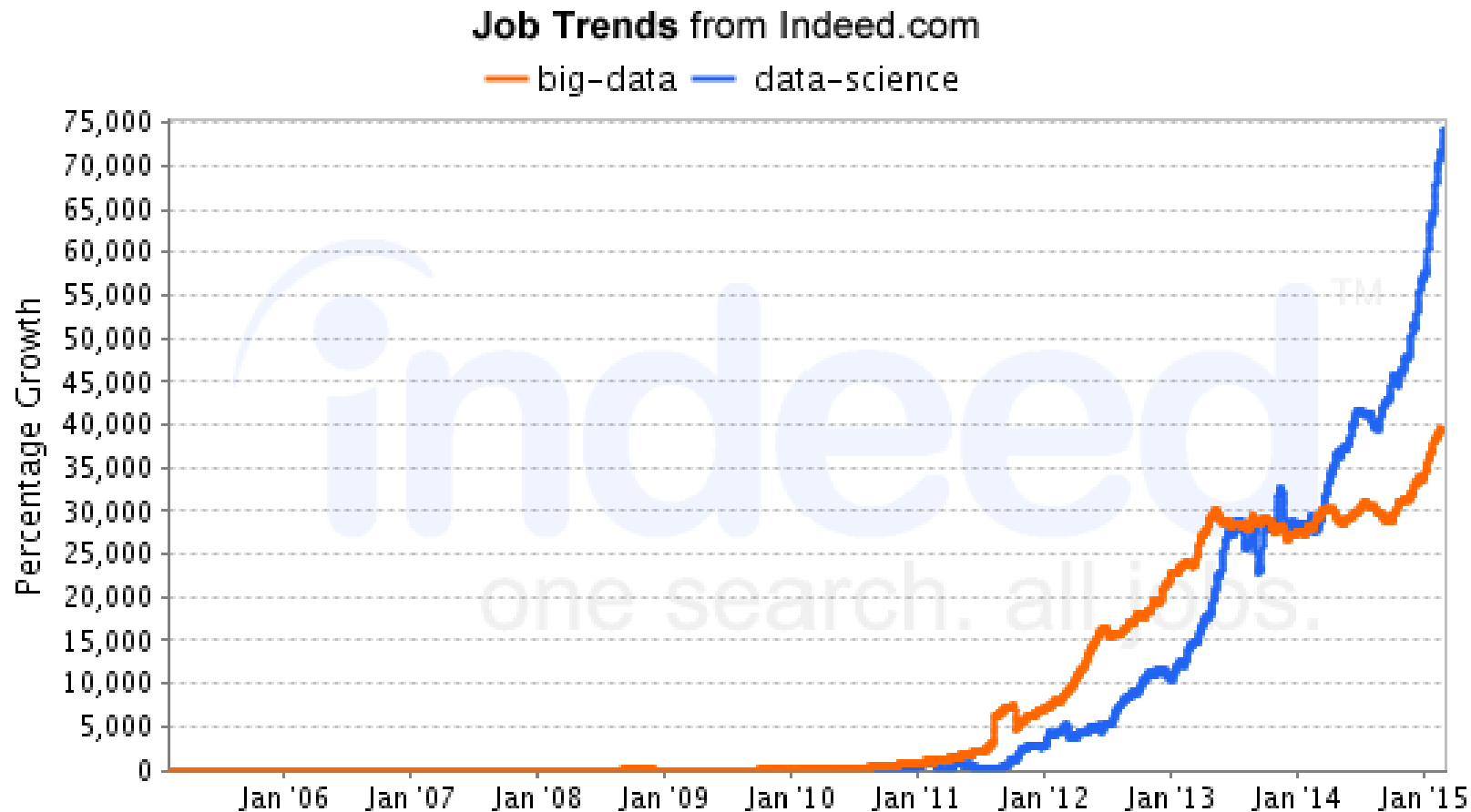
Based on telephone interviews conducted April 2-8, 2018 among 1,509 US adults ages 18 and older, of whom 785 are Facebook users. The remaining respondents answered "Not too concerned" or "Not concerned at all."

## 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.

# Xu hướng nghề nghiệp liên quan tới dữ liệu lớn

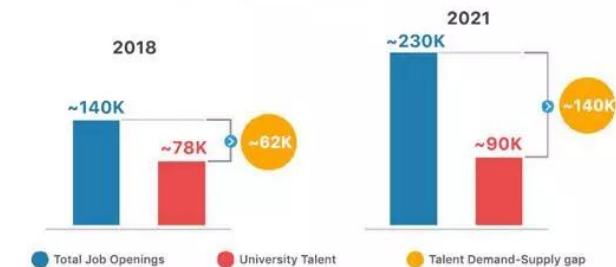


# Thiếu hụt nhân lực liên quan tới dữ liệu lớn

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

Talent Demand-Supply gap analysis





# Nhóm các kỹ năng cần thiết theo vị trí

	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	Very important	Very important	Very important	Very important
Data Visualization and Communication	Very important	Somewhat important	Somewhat important	Very important
Data Intuition	Somewhat important	Very important	Somewhat important	Very important
Statistics	Somewhat important	Very important	Somewhat important	Very important
Data Wrangling	Not that important	Not that important	Very important	Very important
Machine Learning	Not that important	Very important	Not that important	Very important
Software Engineering	Not that important	Somewhat important	Very important	Somewhat important
Multivariable Calculus and Linear Algebra	Not that important	Very important	Not that important	Somewhat important



Not that important



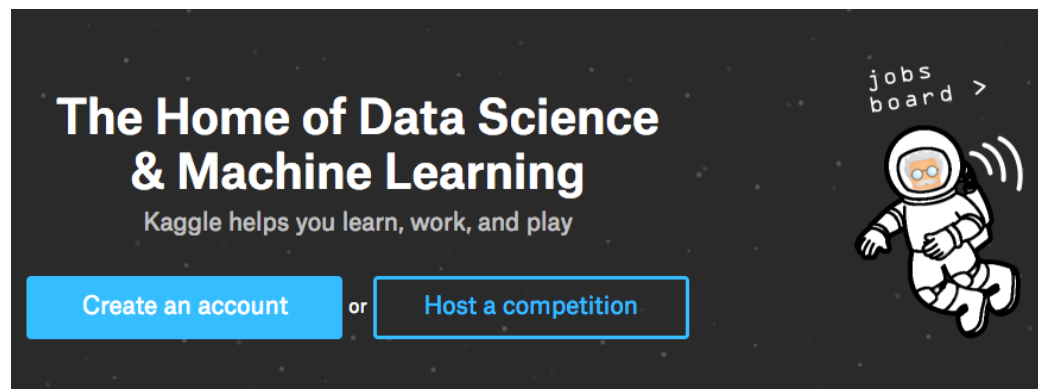
Somewhat important



Very important

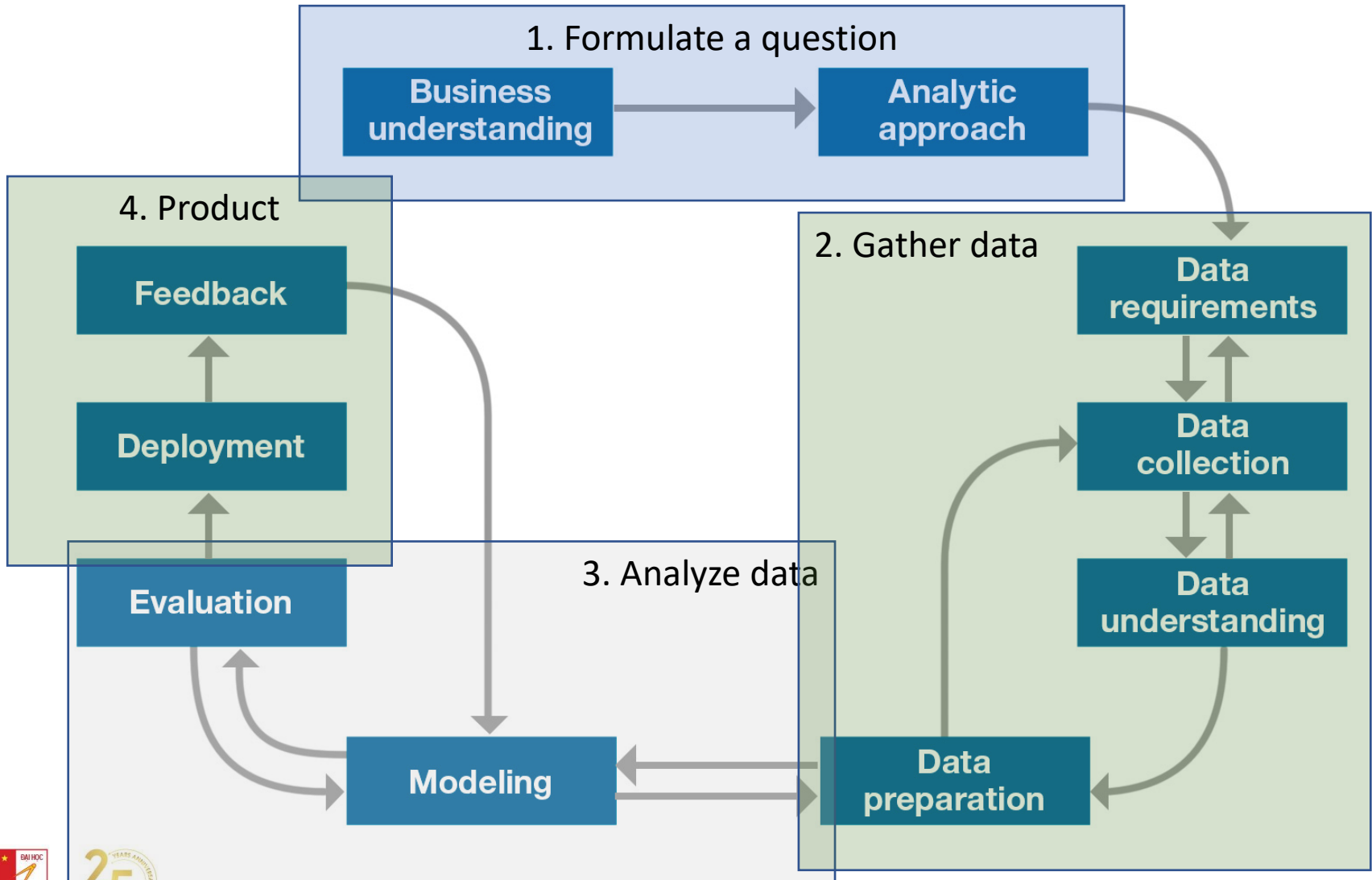
# Gợi ý tìm hiểu về dữ liệu lớn

- Học lập trình
  - Coursera
  - Udacity
  - Freecodecamp
  - Codecademy
- Học học máy, toán, toán thống kê
- Kaggle
- Hadoop, NoSQL, Spark
- Các công cụ báo cáo và trực quan hoá
  - Tableau
  - Pentahoo
- Gặp gỡ và chia sẻ
- Tìm cổ vấn
- Thực tập, dự án

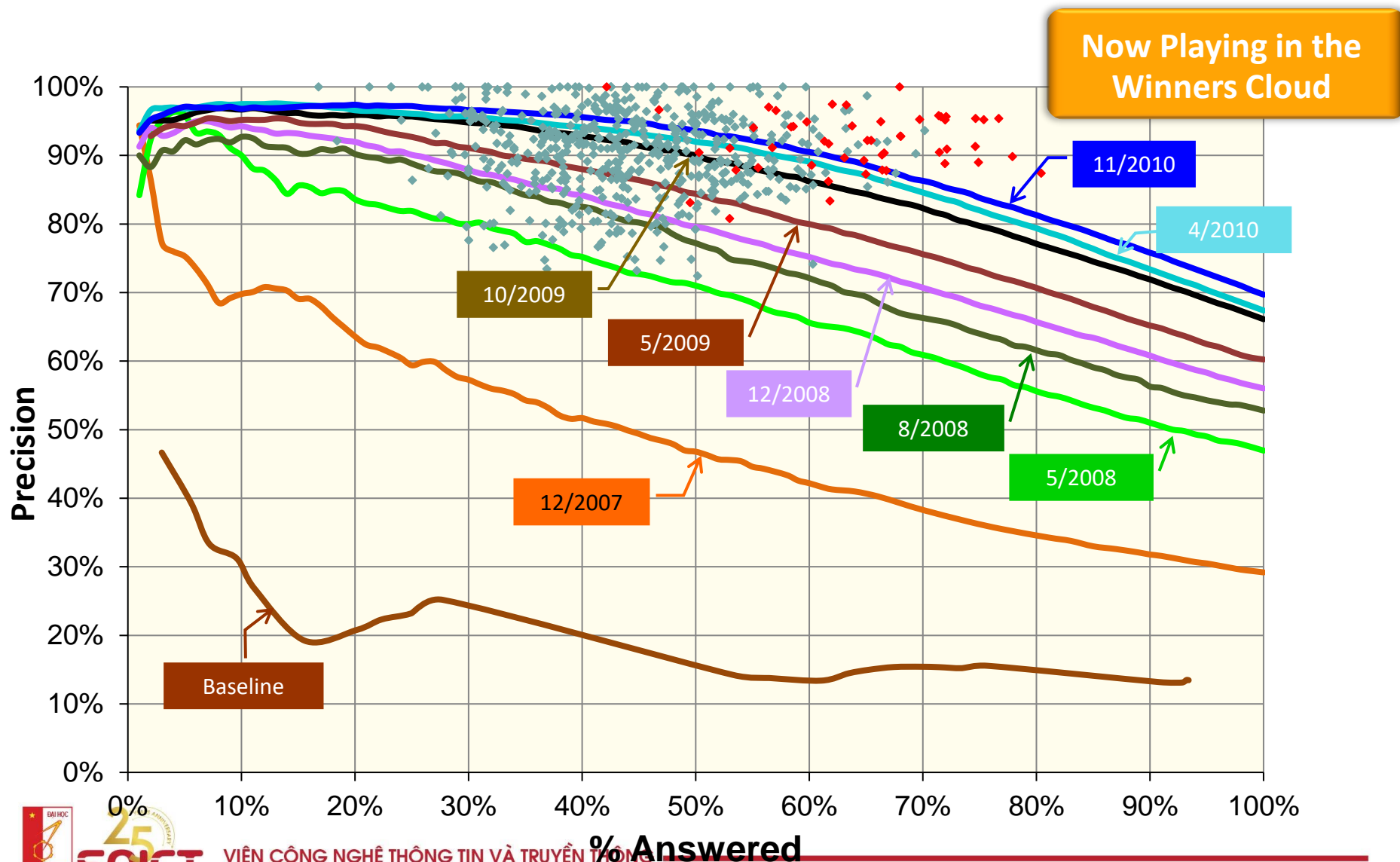




# Quy trình làm khoa học dữ liệu

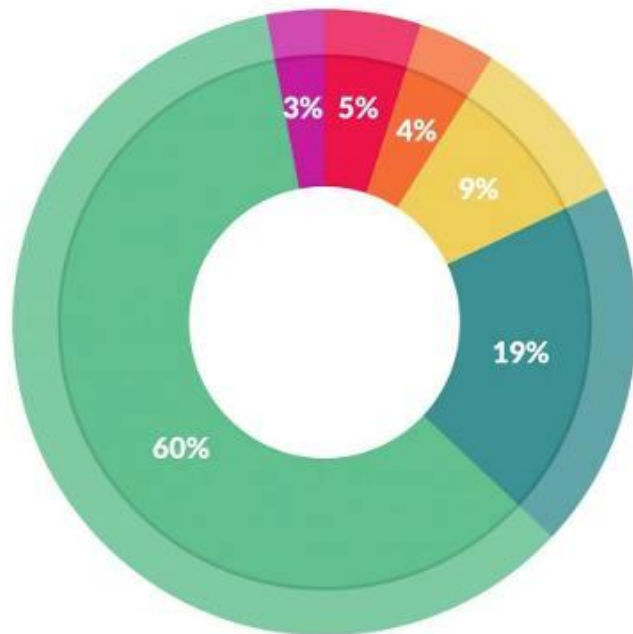


# DeepQA: Incremental Progress in Precision and Confidence 6/2007-11/2010



# Làm sạch dữ liệu lớn: công việc tốn kém thời gian và công sức

- Chiếm khoảng 80% các công việc của nhà khoa học dữ liệu



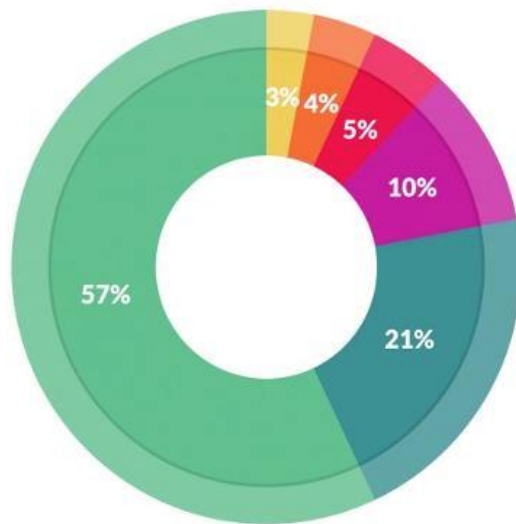
What data scientists spend the most time doing

- 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/>

# Làm sạch dữ liệu lớn: công việc tốn kém thời gian và công sức

- 57% các nhà khoa học dữ liệu cho rằng đây là công việc kém thú vị



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%

# Tài liệu tham khảo

- [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, Srin. 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.

# Các khoá học trực tuyến

- <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

Chân thành  
cảm ơn!



[soict.hust.edu.vn/](http://soict.hust.edu.vn/)



[fb.com/groups/soict](https://fb.com/groups/soict)

