

HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

Chương 1

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

ONE LOVE. ONE FUTURE.

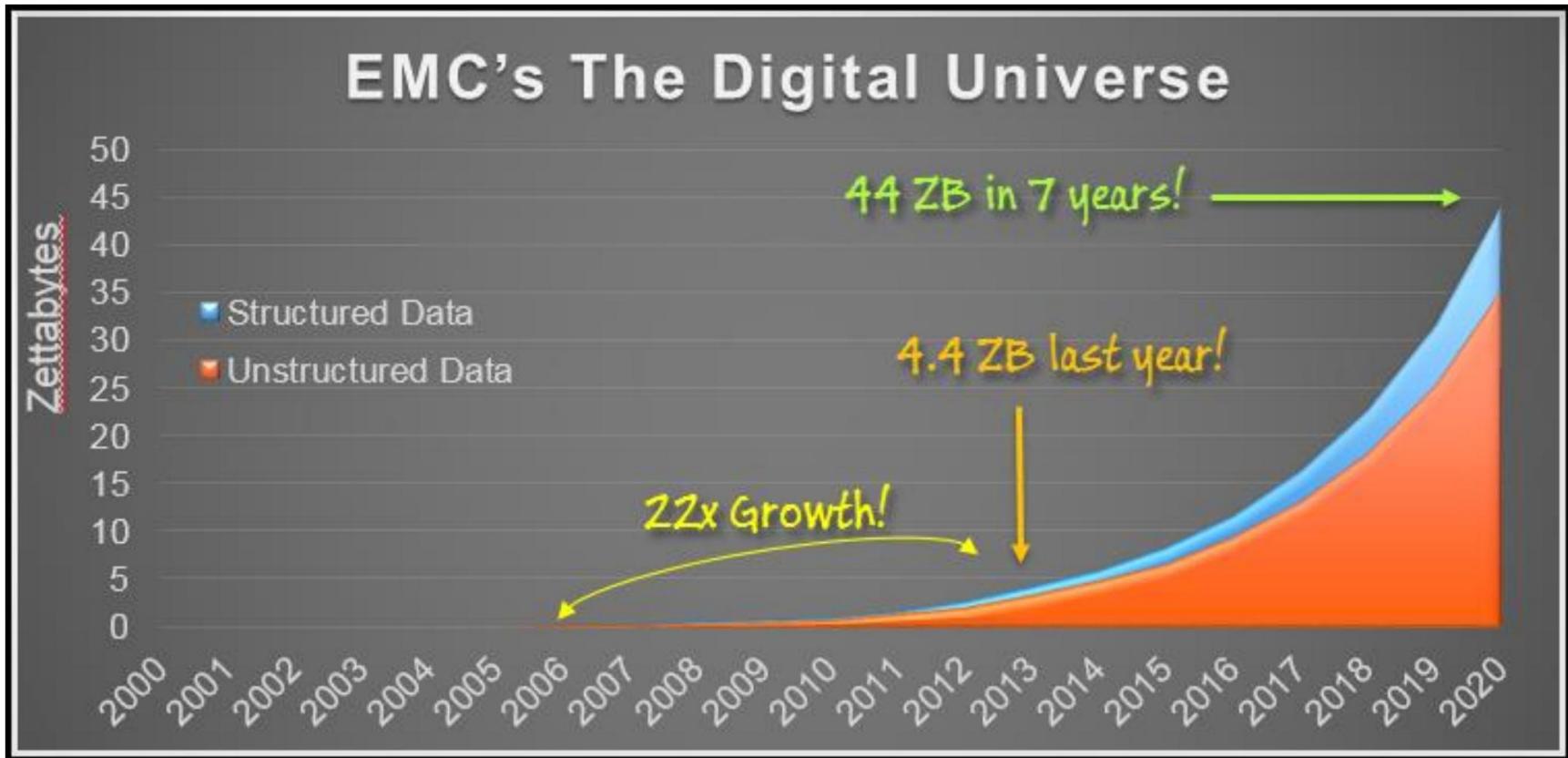
Thông tin chung về môn học

Tên phần học:	Lưu trữ và xử lý dữ liệu lớn (Lưu trữ và xử lý dữ liệu lớn)
Mã số học phần:	IT4931
Khối lượng:	3(3-1-0-6) Lý thuyết: 45 tiết BTL: 15 tiết Thí nghiệm: 0 tiết

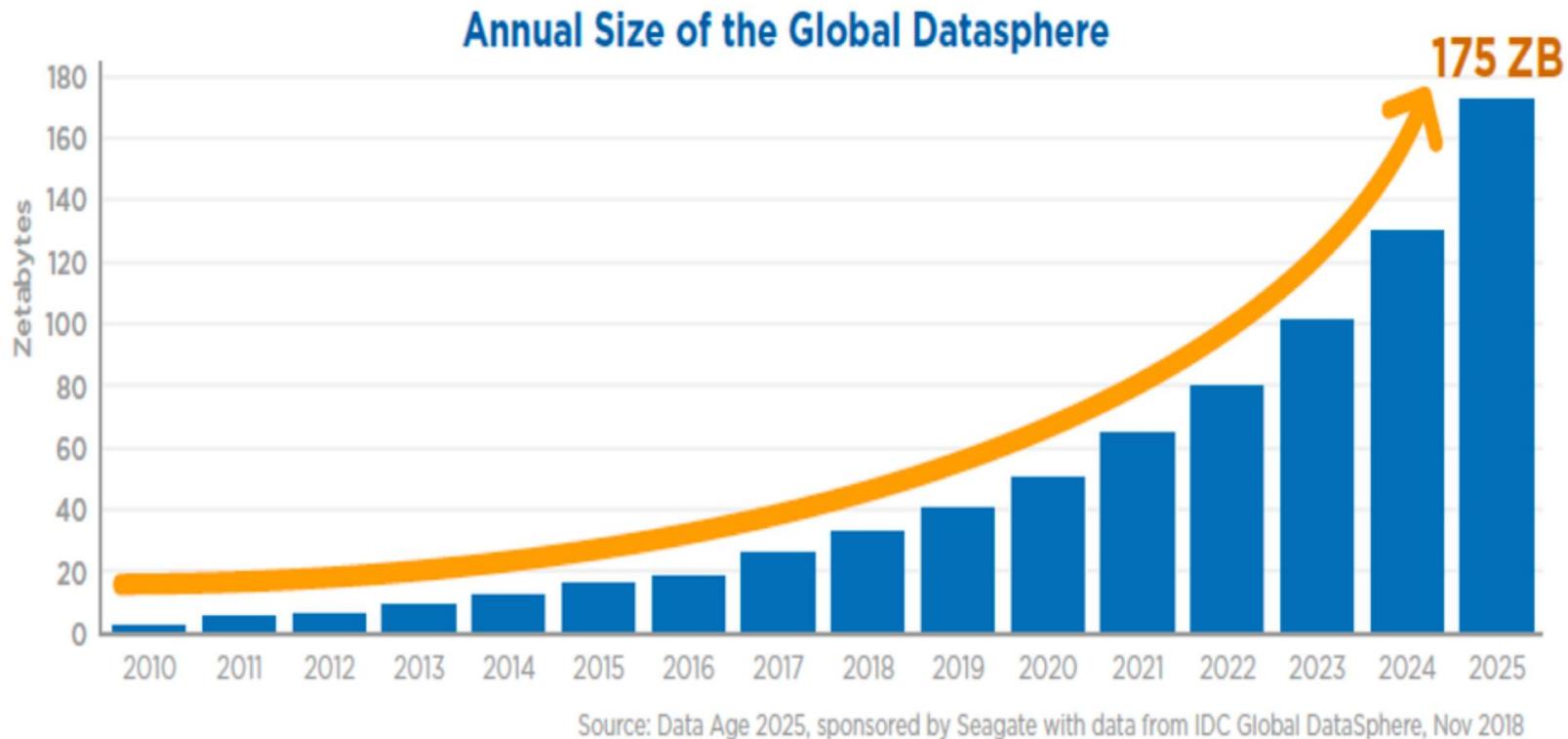
Đề cương học tập

STT	Bài giảng
1	Tổng quan về kho lưu trữ và xử lý dữ liệu lớn
2	Hệ sinh thái Hadoop (Hệ sinh thái Hadoop)
3	Hệ thống phân tích tập tin Hadoop HDFS
4	Cơ sở dữ liệu NoSQL cơ sở dữ liệu - phần 1 Tổng quan
5	Cơ sở dữ liệu NoSQL cơ sở dữ liệu - phần 2 Kiến trúc phổ rộng phân tán
6	Cơ sở dữ liệu NoSQL cơ sở dữ liệu - phần 3 Truy vấn SQL trên NoSQL
7	Phân phối thông điệp hệ thống
8	Các kỹ thuật xử lý dữ liệu lớn theo khối - phần 1 Bản đồ thu nhỏ
9	Các kỹ thuật xử lý dữ liệu lớn theo khối - phần 2 Tia lửa Apache
10	Luồng dữ liệu xử lý kỹ thuật kỹ thuật Spark phát trực tuyến
11	Big data architecture Kiến trúc Lambda
12	Phân tích dữ liệu lớn Tia lửa ML

Tổng dung lượng dữ liệu 2020



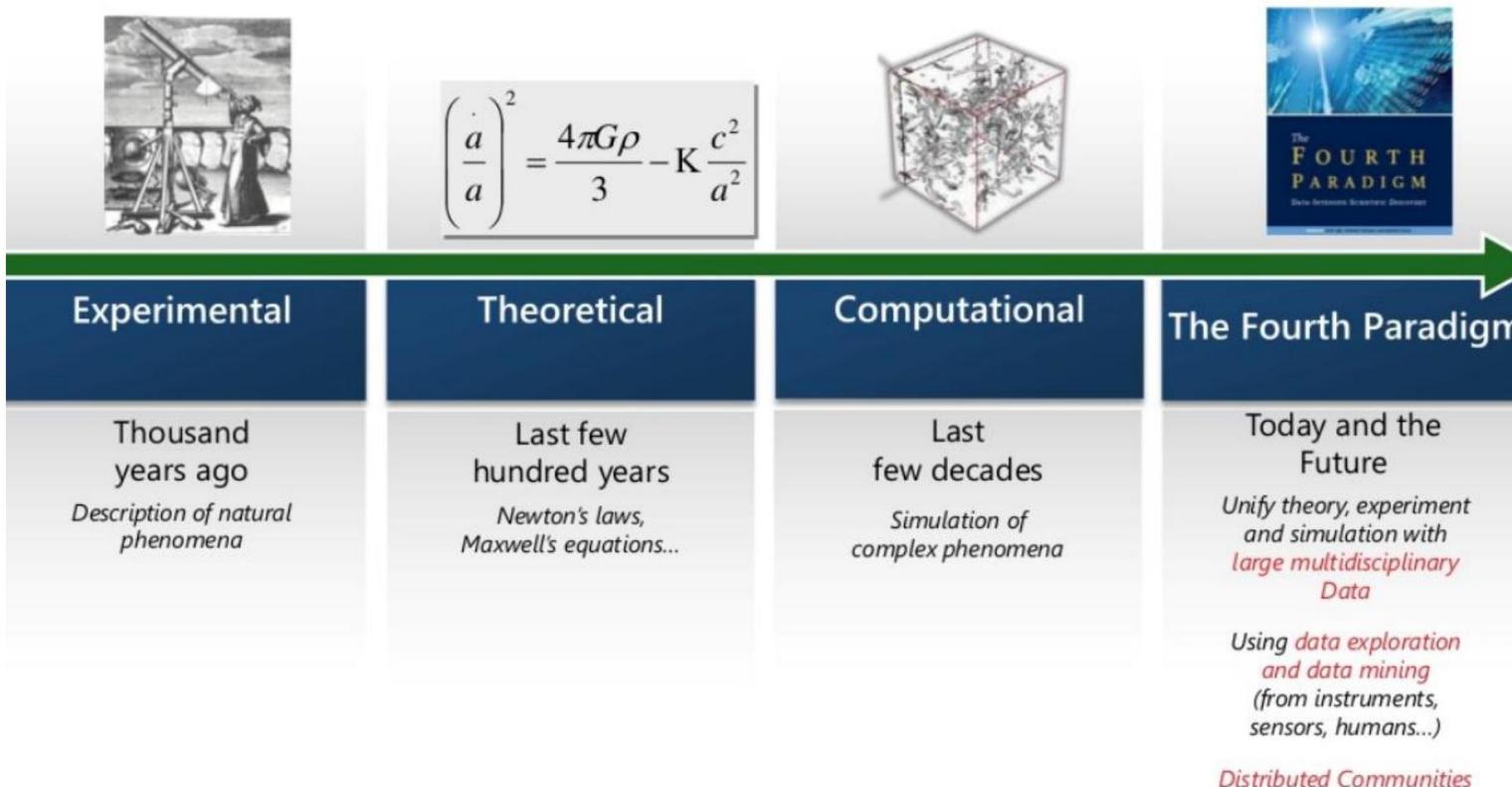
Tổng dung lượng dữ liệu 2025



Data size lớn nhất



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



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

Data 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

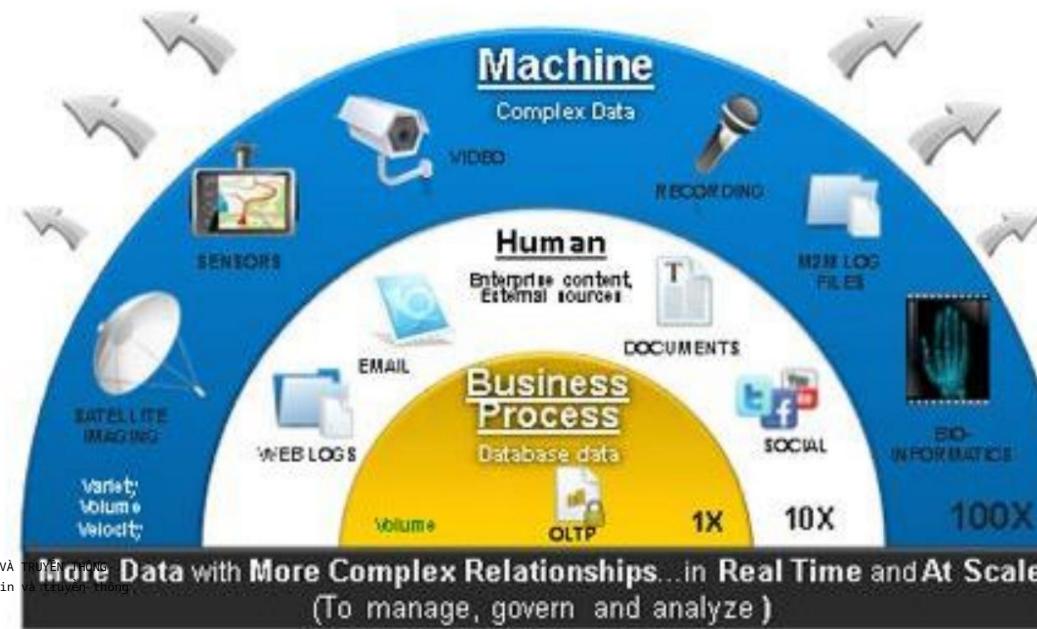
Các thông số về tốc độ sinh dữ liệu

2020 *This Is What Happens In An Internet Minute*

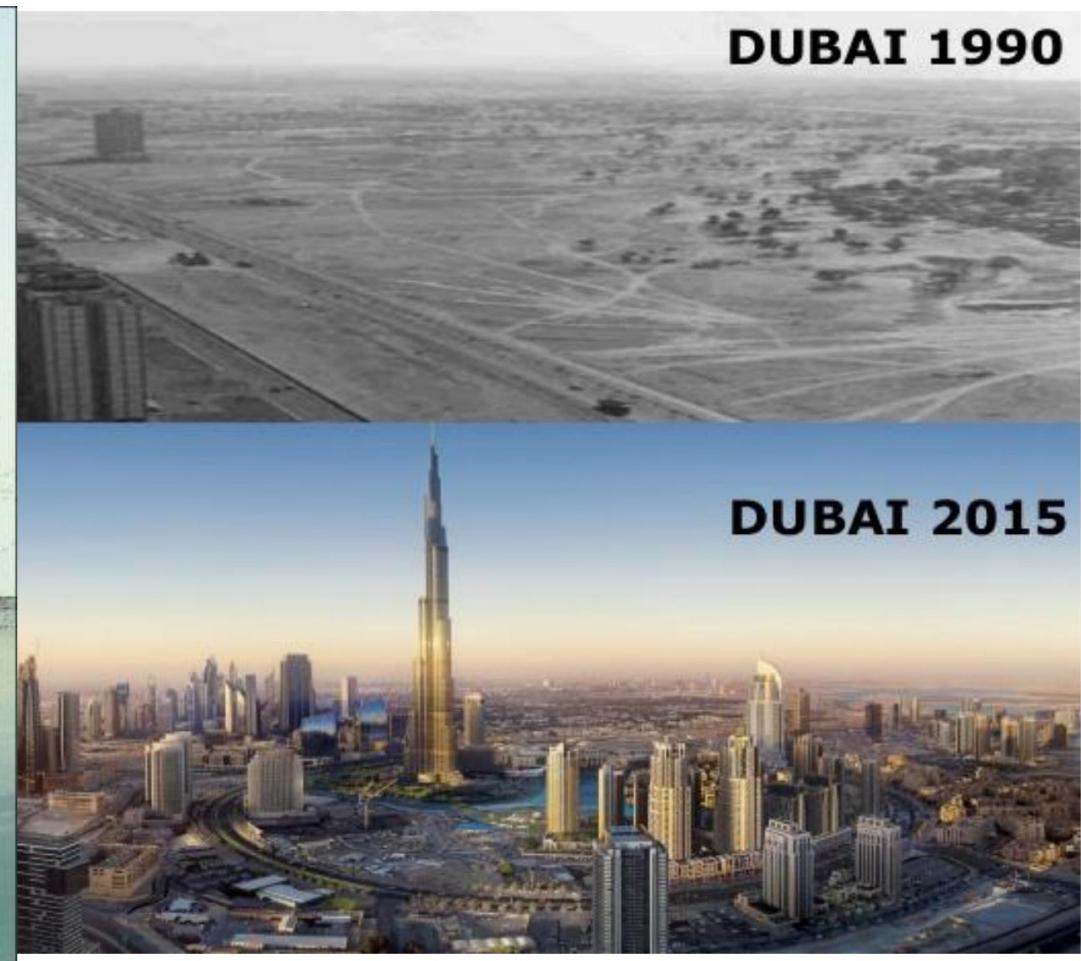


Nguồn tạo dữ liệu lớn

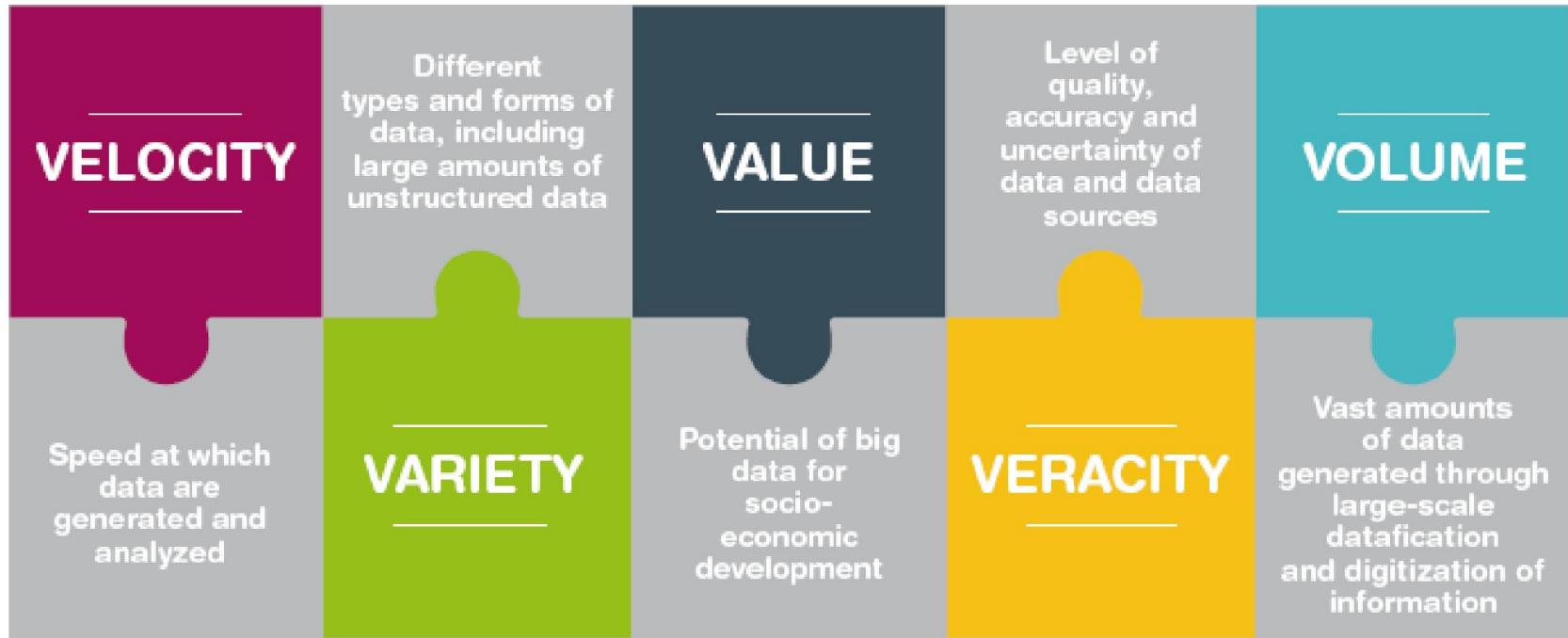
- Điện tử thương mại
- Mạng xã hội
- Internet vạn vật (IoT)
- Các thử nghiệm dữ liệu lớn (tin sinh học, vật chất lượng, vvv)



Dữ liệu được ví như nguồn tài nguyên dầu mỏ mới

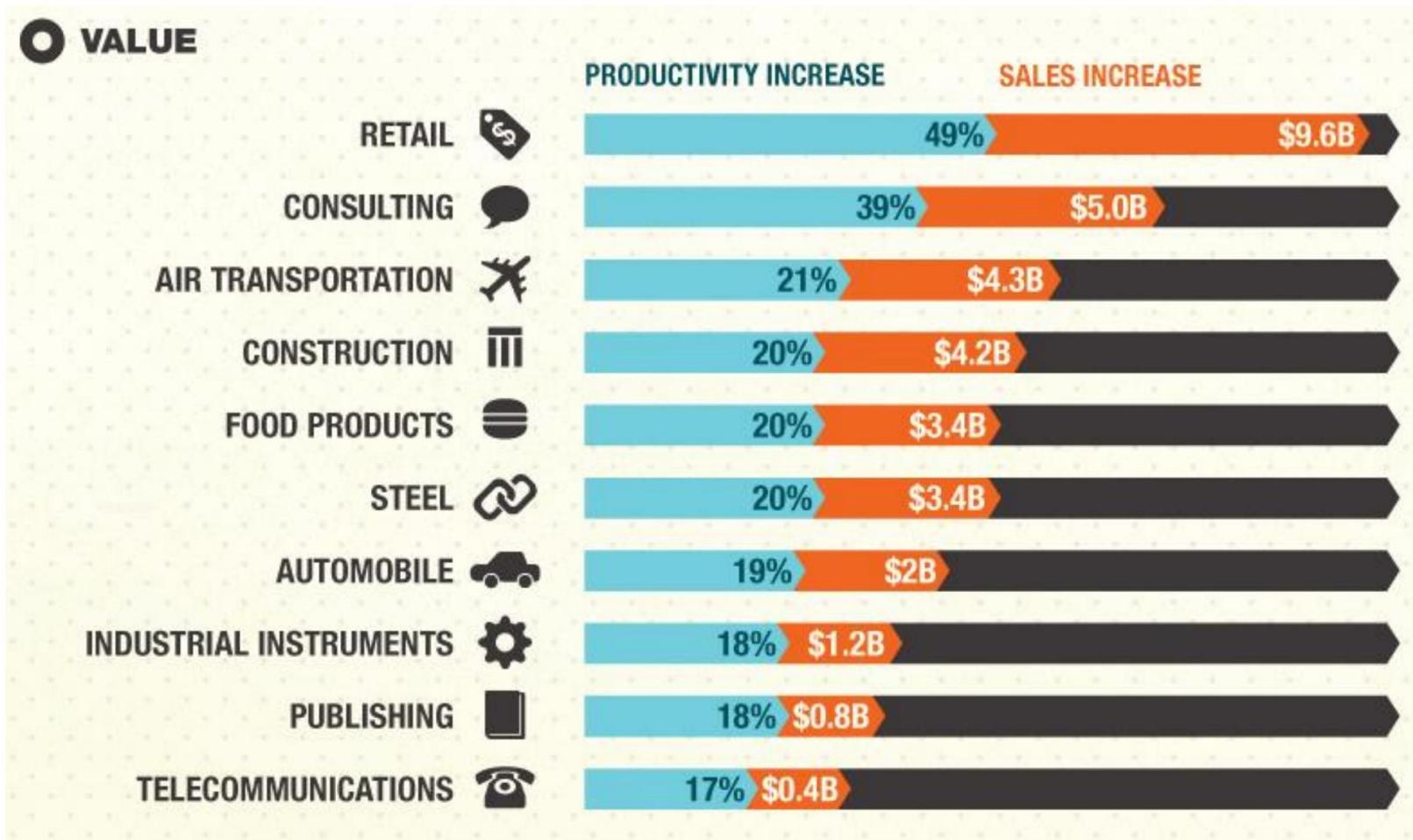


Đặc điểm 5'V của dữ liệu lớn



Dữ liệu lớn là dữ liệu quá lớn hoặc quá phức tạp nên kho lưu trữ nền tảng và xử lý truyền dữ liệu không thể đáp ứng được.

Big data - Mang lại giá trị lớn



Amazon

Let's talk with numbers

How a product recommendation engine can boost your revenue

Amazon's sales



\$280.5B

Amazon's total
2019 revenue



35%

of Amazon.com revenue
is generated by its
recommendation engine



Estimation for 2020 is to touch

\$334.7B

amazon.com

Recommended for You

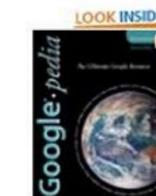
Amazon.com has new recommendations for you based on items you purchased or told us you own.



[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)



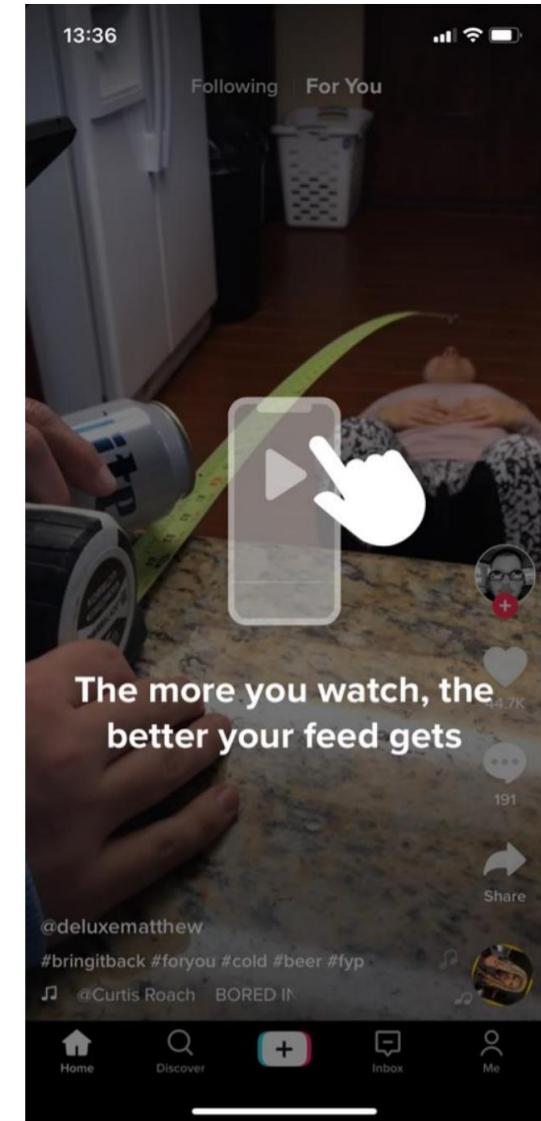
[Administrator Guide: A Private-Label Web Workspace](#)



[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)

Tiktok

- Tính đến năm 2024, TikTok có 1,04 tỷ người dùng hoạt động hàng tháng trên toàn cầu.
- Người dùng TikTok dành 58 phút 24 giây trên ứng dụng này mỗi ngày tính đến năm 2024.
- Phần lớn người dùng TikTok là trong độ tuổi từ 18 đến 34, chiếm 69,3%.



Khai thác dữ liệu lớn trong giáo dục



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

- Giảm chi phí điều trị, các thử nghiệm mô
- Dự kiến quy đại dịch, đưa ra các giải pháp bảo vệ
- Điểm báo 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 • Bắt nhanh các vấn đề xã hội (vi phạm, môi trường, vvv) • Khuyến nghị các biện pháp đổi mới
- An ninh thông tin • Trốn thuế • Lừa đảo



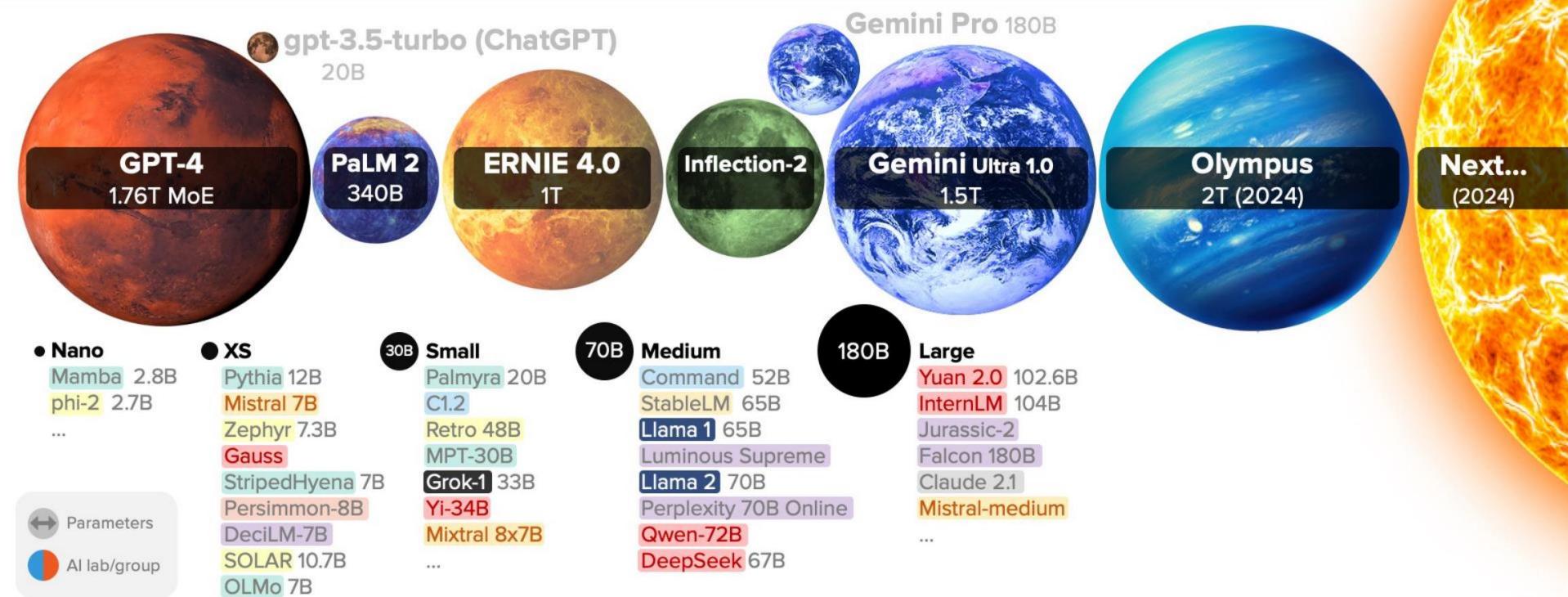
Khai thác dữ liệu lớn trong khoa học khám phá



Máy va chạm Hydron lớn (LHC) của CERN tạo ra 15 PB một năm

Thế giới đang cạn kiệt dữ liệu để đào tạo AI

LARGE LANGUAGE MODEL HIGHLIGHTS (FEB/2024)

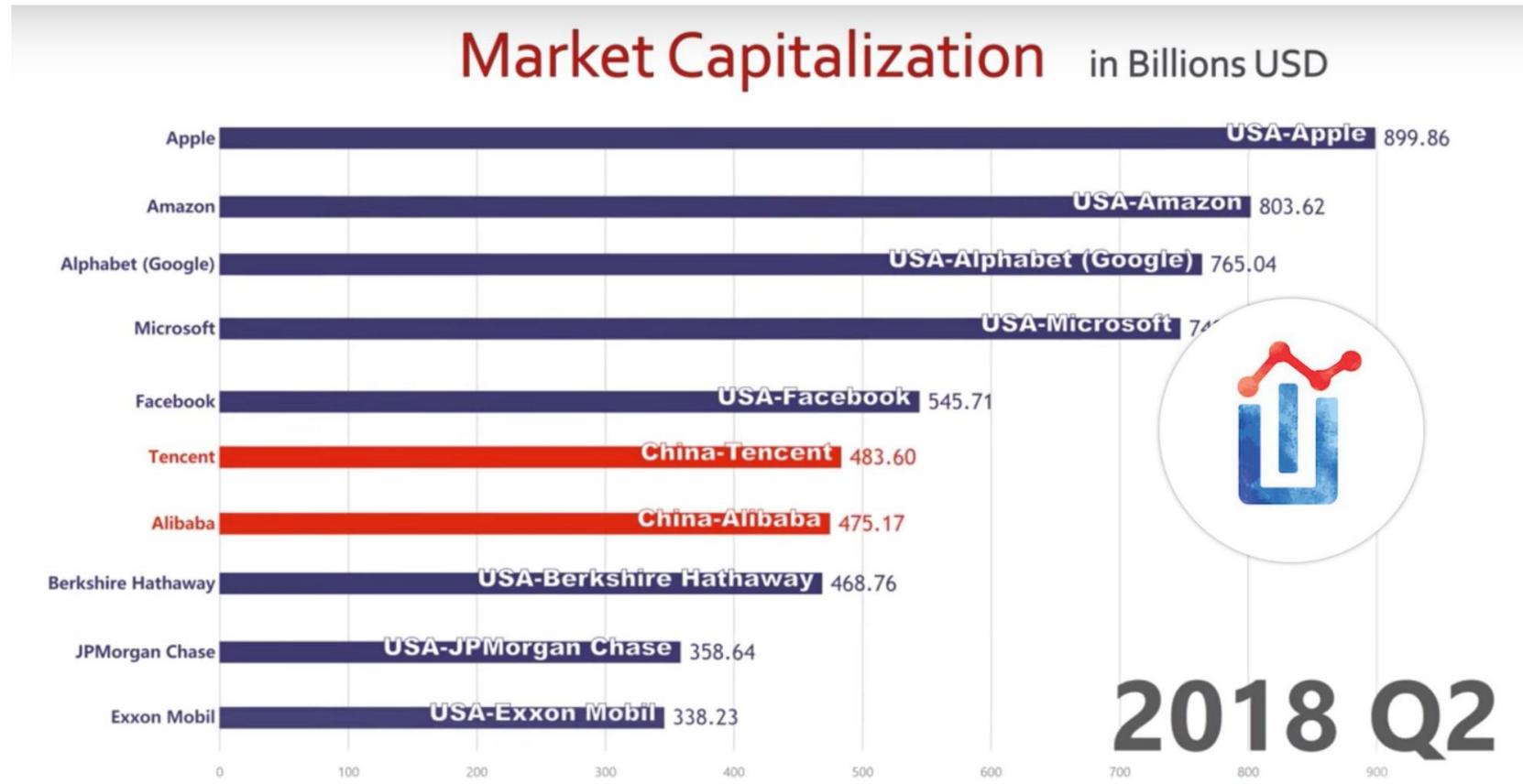


 LifeArchitect.ai/models

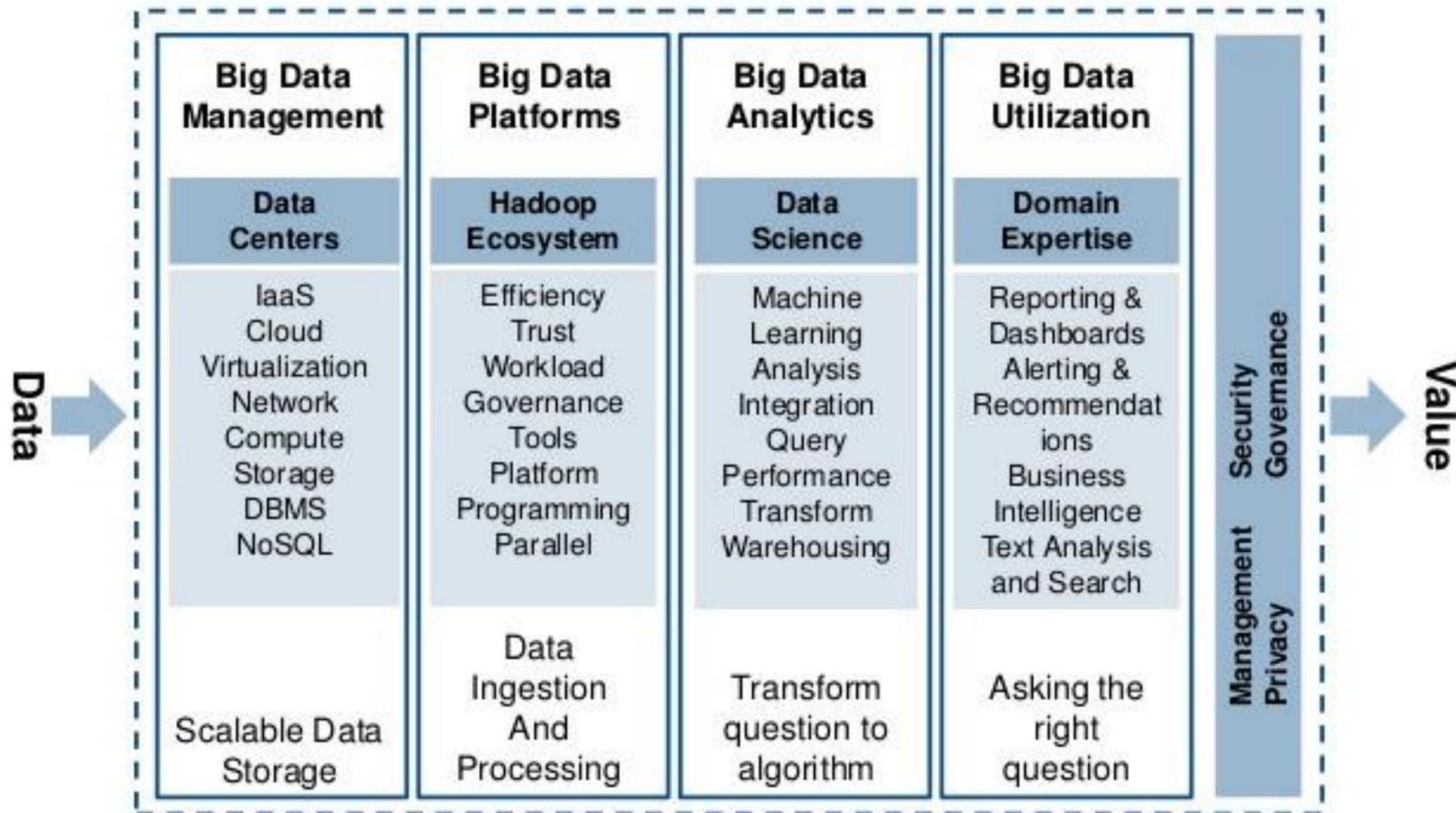
10 công ty lớn nhất (1998-2018)



10 công ty lớn nhất (1998-2018)



Các tầng công nghệ cho dữ liệu lớn



Quản lý dữ liệu phải mở

- Khả năng mở rộng
 - Khả năng quản lý lượng lớn dữ liệu không ngừng tăng lên theo thời gian.
- Khả năng tiếp cận
 - Cho phép đọc kết quả dữ liệu I/O write.
- Minh bạch
 - Truy cập dữ liệu một cách dễ dàng, vị trí lưu trữ dữ liệu trên hệ thống là rõ ràng đối với người dùng cuối cùng.
- Tính khả dụng
 - Có khả năng chịu đựng lỗi, khi tăng số lượng người dùng, khi gặp rắc rối.

Xử lý và tích hợp dữ liệu phải mở

- Tích hợp dữ liệu •

Dữ liệu có các 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 thông tin, quyền riêng tư

- Xử lý dữ liệu

- Xử lý khói lượng dữ liệu rất lớn • Xử lý luồng dữ

- liệu • Xử lý bài hát dữ liệu, phân phối dữ liệu (OpenMP, MPI)

- Phúc tạp, khó học

- Khả năng mở giới hạn • Cơ chế hạn chế lỗi • Chi phí hạ tầng giảm thiểu • Kiến trúc xử lý luồng dữ liệu lớn • Spark mini-batch

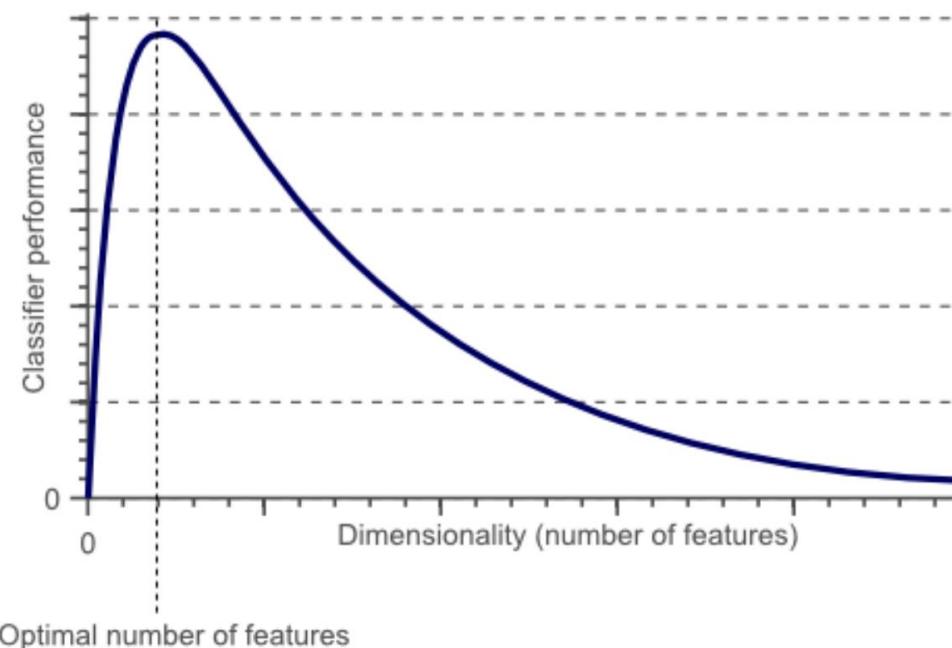
- Apache Flink

Khả năng mở dữ liệu phân tích thuật toán

- Làm lại dữ liệu phù hợp với truyền thông giải thuật • Vd.
Lấy mẫu phụ
 - Ví dụ. Phân tích thành phần chính
 - Ví dụ. Trích xuất tính năng và lựa chọn tính năng
- Song hoá các máy giải thuật
 - Ví dụ: phân loại k-nn dựa trên MapReduce
 - Ví dụ: mở rộng quy mô máy vectơ hỗ trợ (SVM) bằng phương pháp chia để trị

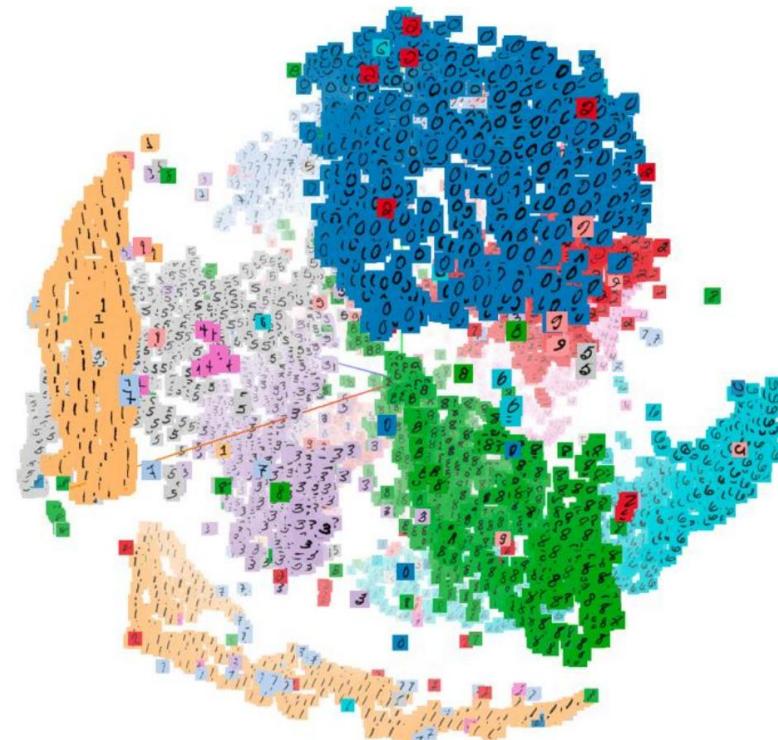
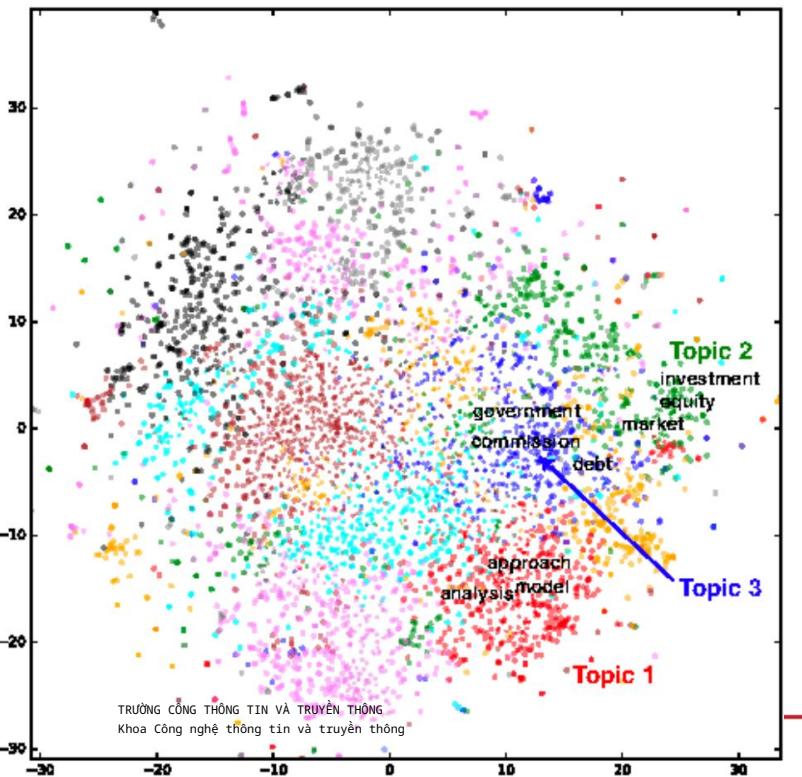
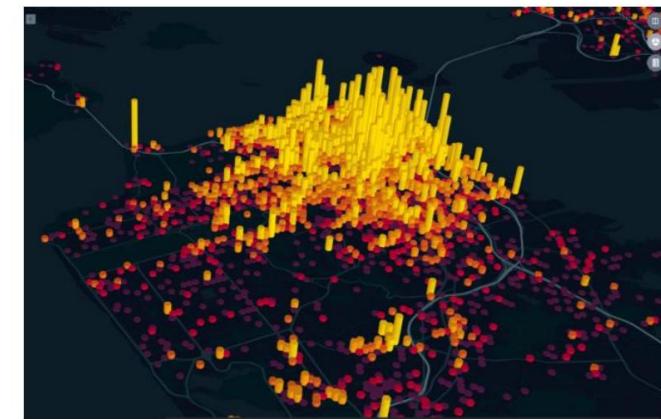
Ví dụ. Sự kiện mong đợi số chiều trong dữ liệu (Lời nguyền của chiều)

- Số lượng mẫu cần tăng mô hình học khi tăng số chiều dữ liệu • Trong quá trình triển khai: Số lượng mẫu để cố định học thường xuyên
 - => Độ chính xác của mô hình giảm khi tăng số chiều trong dữ liệu tài liệu học



Sử dụng và trực tiếp hóa dữ liệu lớn

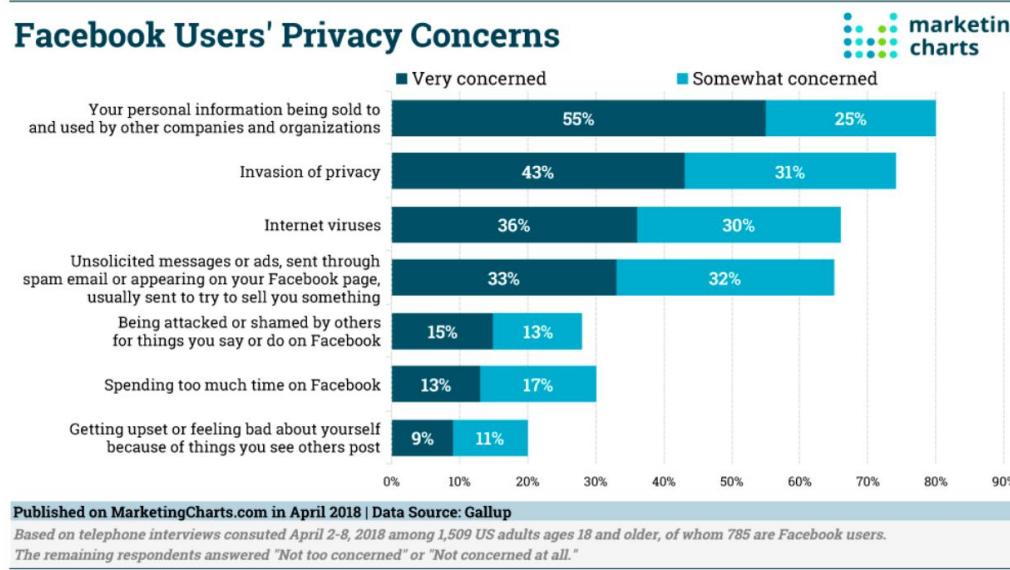
- Cần kiến thức chuyên môn
- Cần kỹ thuật và công cụ để hỗ trợ hiệu quả trình diễn và hiểu về dữ liệu lớn



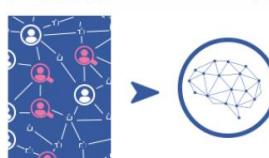
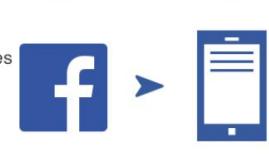
Bảo mật và quyền riêng tư



Facebook Users' Privacy Concerns



How was Facebook users' data misused?

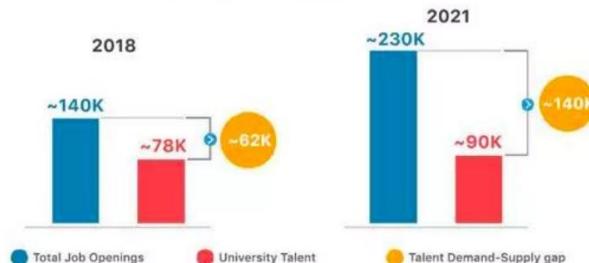
- 1 In 2014 a Facebook quiz invited users to find out their personality type 
 - 2 The app collected the data of those taking the quiz, but also recorded the public data of their friends 
 - 3 About 305,000 people installed the app, but it gathered information on up to 87 million people, according to 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.

Missing link kernel to data big

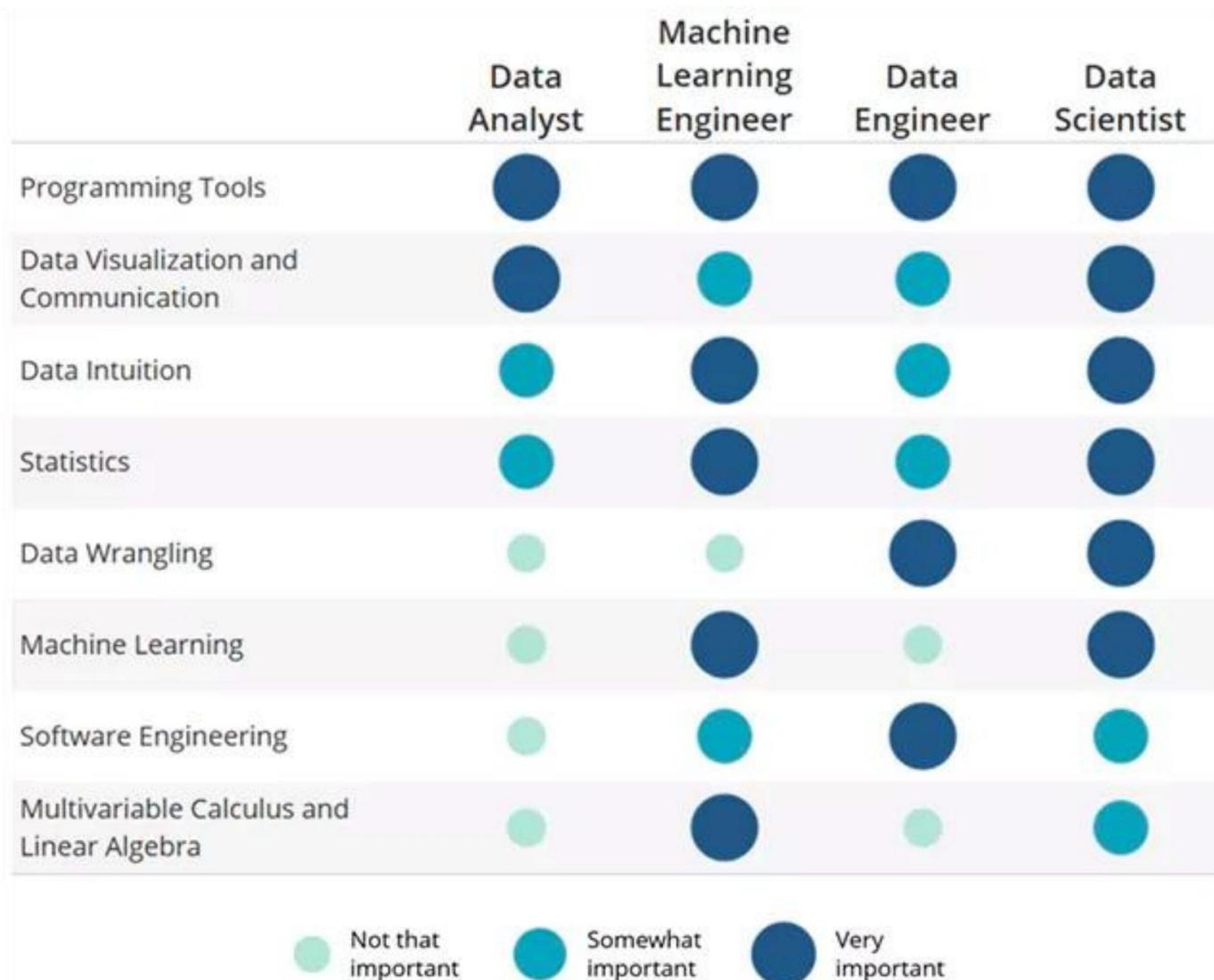
Table 2. Summary Demand Statistics

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 kỹ năng cần thiết theo vị trí



Ý nghĩa của việc tìm hiểu về dữ liệu lớn

- Học lập trình

 - Coursera •

 - Udacity •

 - Freecodecamp •

 - Codecademy • Học

máy, toán, toán thống kê • Kaggle • Hadoop, NoSQL,

Spark • Các

công cụ báo cáo và trực quan hóa

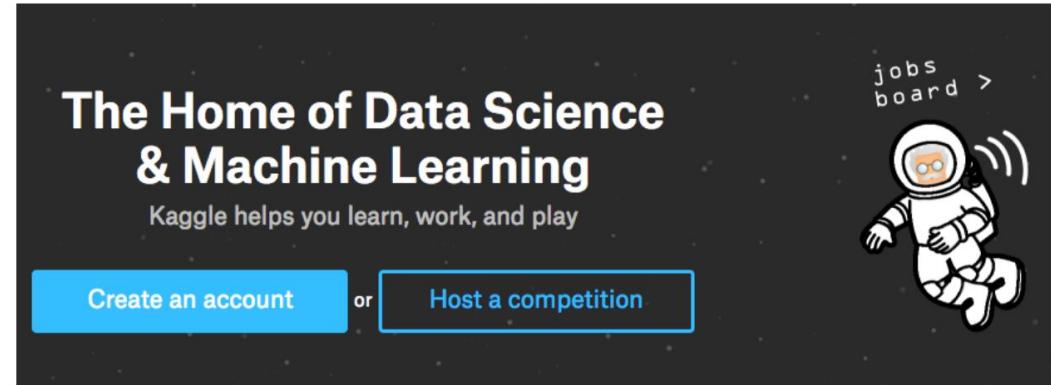
- Bảng •

 - Pentaho

- Gỡ lỗi và chia sẻ

- Tìm sự cố •

Thực tập, dự án



The Home of Data Science & Machine Learning

Kaggle helps you learn, work, and play

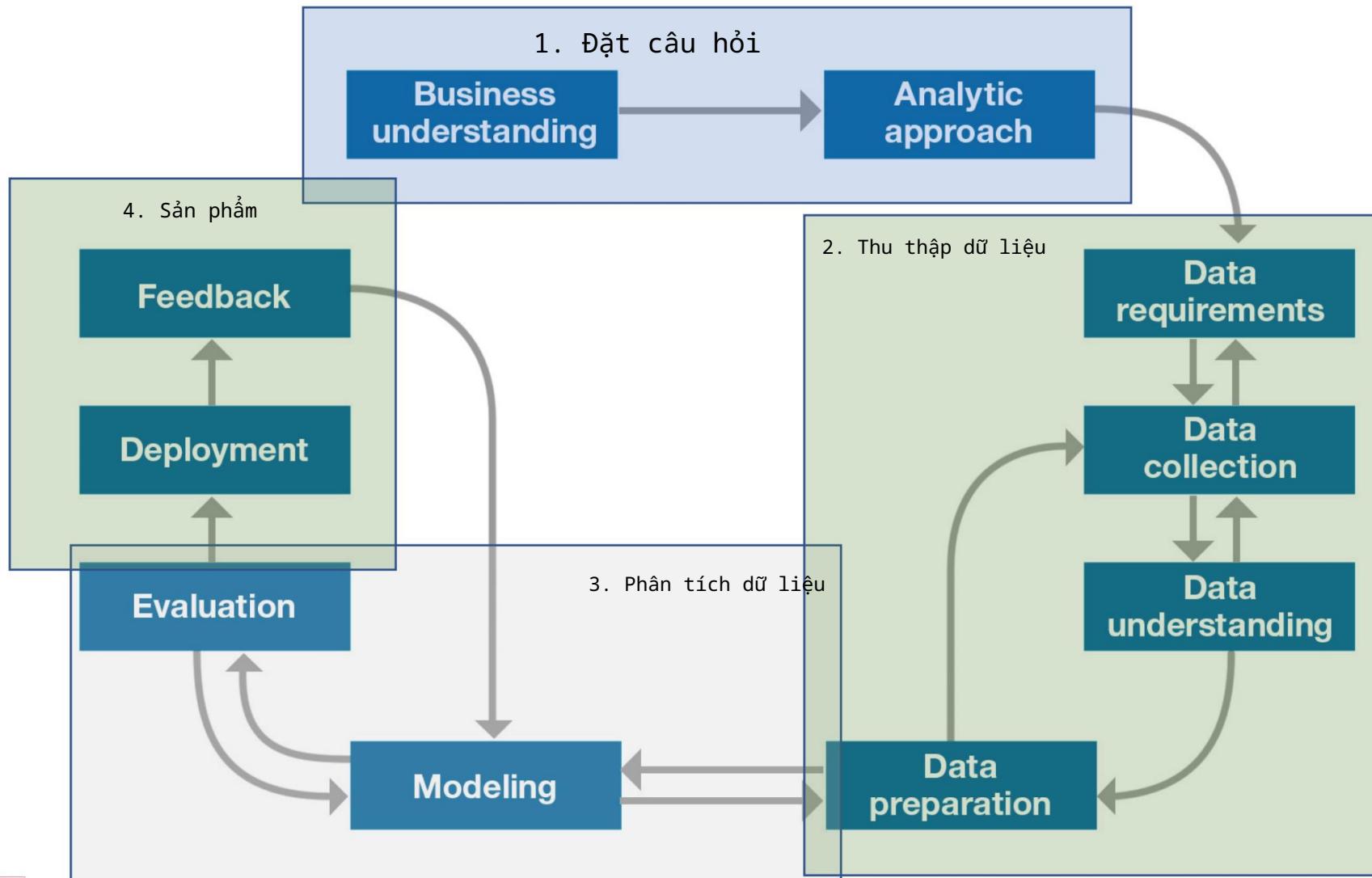
[Create an account](#)

or

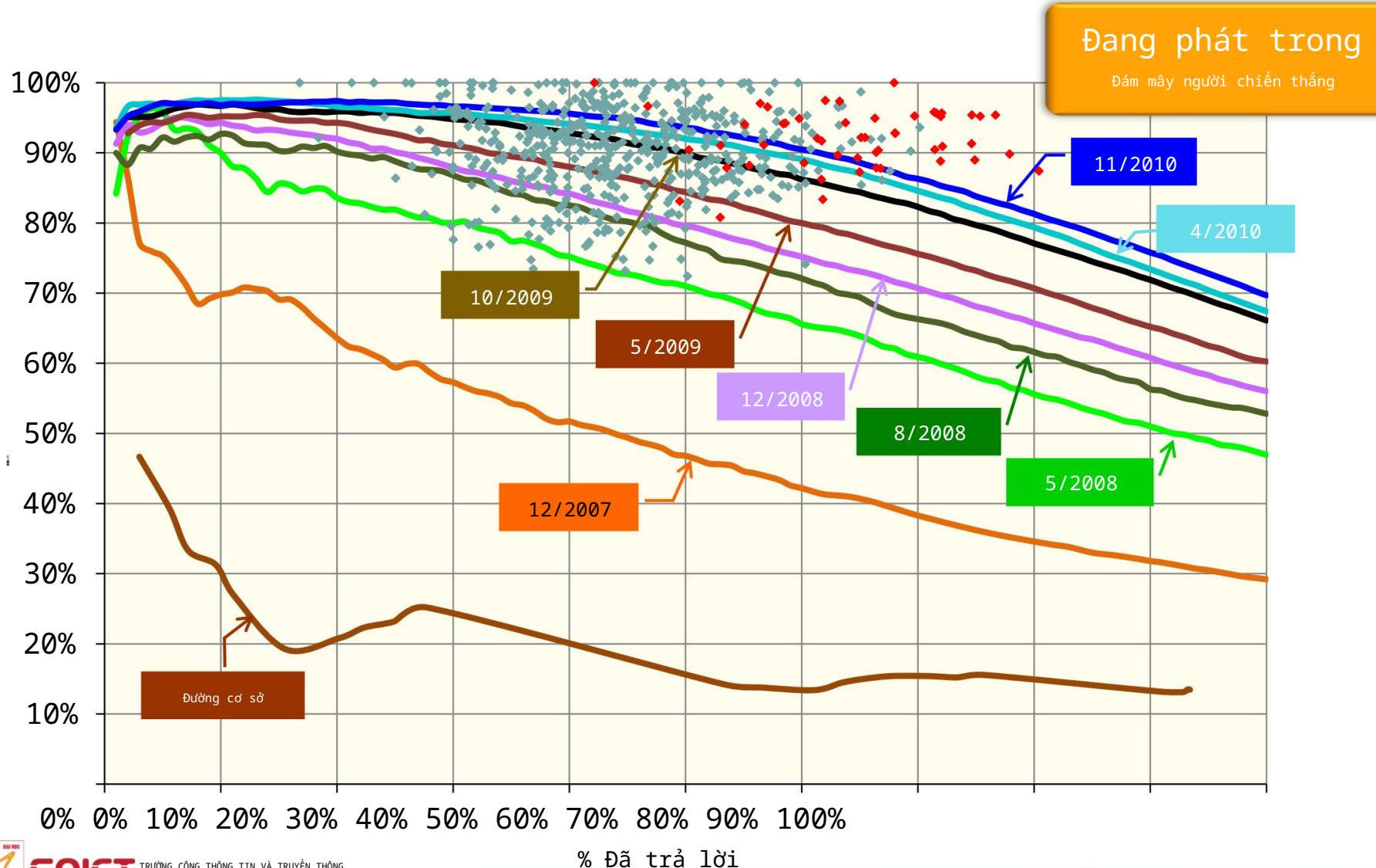
[Host a competition](#)



Quy trình làm khoa học dữ liệu

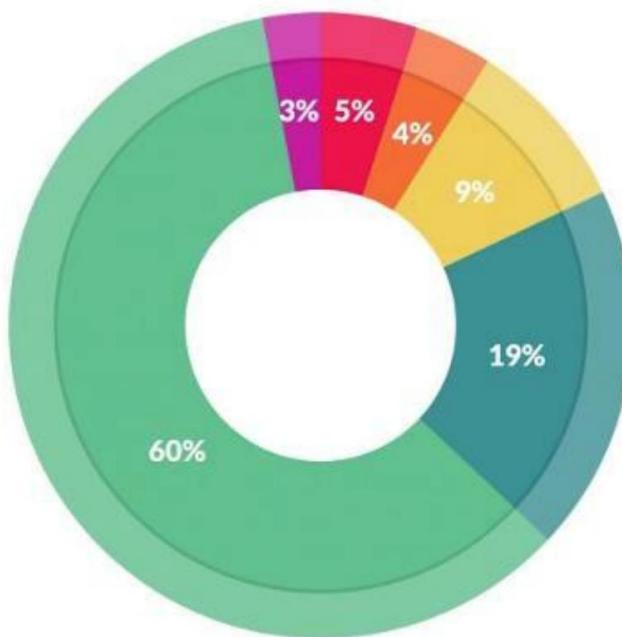


DeepQA: Tiến bộ gia tăng về độ chính xác và độ tin cậy 6/2007-11/2010



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

- Search interval 80% công việc của nhà khoa học dữ liệu



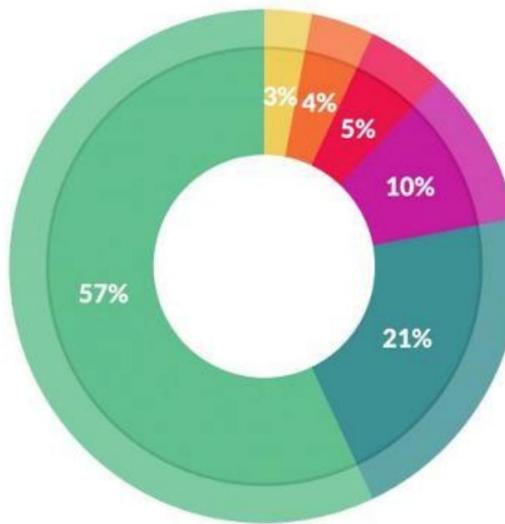
What data scientists spend the most time doing

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

nguồn: <https://www.forbes.com/>

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

- 57% các nhà khoa học dữ liệu cho rằng đây là công việc không 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. NoSQL chuyên nghiệp. John Wiley & Sons, 2011.
- [2] Lam, Chuck. Hadoop trong hành động. Manning Publications Co., 2010.
- [3] Miner, Donald và Adam Shook. Các mẫu thiết kế MapReduce: xây dựng các thuật toán và phân tích hiệu quả cho Hadoop và các hệ thống khác. "Công ty truyền thông O'Reilly", 2012.
- [4] Karau, Holden. Xử lý dữ liệu nhanh với Spark. Packt Publishing Ltd, 2013.
- [5] Penchikala, Srini. Xử lý dữ liệu lớn với tia lửa apache. Lulu. com, 2018.
- [6] White, Tom. Hadoop: Hướng dẫn xác định. "Công ty truyền thông O'Reilly", 2012.
- [7] Gandomi, Amir và Murtaza Haider. "Vượt ra ngoài sự cưỡng điệu: Các khái niệm, phương pháp và phân tích dữ liệu lớn." Tạp chí Quản lý thông tin quốc tế 35.2 (2015): 137-144.
- [8] Cattell, Rick. "Kho dữ liệu SQL và NoSQL có thể mở rộng." Acm Sigmod Record 39.4 (2011): 12-27.
- [9] Gessert, Felix, et al. "Hệ thống cơ sở dữ liệu NoSQL: khảo sát và hướng dẫn quyết định." Khoa học máy tính-Nghiên cứu và phát triển 32.3-4 (2017): 353- 365.
- [10] George, Lars. HBase: hướng dẫn xác đáng: truy cập ngẫu nhiên vào dữ liệu có kích thước hành tinh của bạn. "Công ty truyền thông O'Reilly", 2011.
- [11] Sivasubramanian, Swaminathan. "Amazon dynamoDB: dịch vụ cơ sở dữ liệu phi quan hệ có thể mở rộng liền mạch." Biên bản Hội nghị quốc tế ACM SIGMOD năm 2012 về Quản lý dữ liệu. ACM, 2012.
- [12] Chan, L. "Presto: Tương tác với petabyte dữ liệu tại Facebook." (2013).
- [13] Garg, Nishant. Apache Kafka. Công ty TNHH Xuất bản Packt, 2013.
- [14] Karau, Holden, et al. Tia lửa học tập: phân tích dữ liệu lớn nhanh như chớp. "Công ty truyền thông O'Reilly", 2015.
- [15] Iqbal, Muhammad Hussain và Tariq Rahim Soomro. "Phân tích dữ liệu lớn: Quan điểm về cơn bão Apache." Tạp chí quốc tế về xu hướng máy tính và công nghệ 19.1 (2015): 9-14.
- [16] Toshniwal, Ankit, et al. "Storm@ twitter." Biên bản báo cáo hội nghị quốc tế ACM SIGMOD năm 2014 về Quản lý dữ liệu. ACM, 2014.
- [17] Lin, Jimmy. "Lambda và kappa." IEEE Internet Computing 21.5 (2017): 60-66.

Các khóa 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-
dữ liệu](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-integration-processing?specialization=big-data)
- [https://www.coursera.org/learn/big-data-
management?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>

