

The image shows the HUST logo and course information on a white background with a pattern of blue dots. The logo consists of a small red square with a white star and the text "ĐẠI HỌC BÁCH KHOA HÀ NỘI" and "HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY" in small, black, sans-serif capital letters. Below the logo, the course title "Nhập môn Khoa học dữ liệu (IT4142)" is written in large, bold, red, sans-serif capital letters. Below the title, the text "PGS.TS Thân Quang Khoát & PGS.TS Phạm Văn Hải" is written in red, sans-serif capital letters, followed by "Team lecturers" in smaller, red, sans-serif capital letters. At the bottom, the slogan "ONE LOVE. ONE FUTURE." is written in red, sans-serif capital letters. The background is a solid white color with a pattern of blue dots of varying sizes arranged in a circular, halftone-like pattern.

Nhập môn
Khoa học dữ liệu
(IT4142)

PGS.TS Thân Quang Khoát & PGS.TS Phạm Văn Hải
Team lecturers

ONE LOVE. ONE FUTURE.

Contents

- **Lecture 1: Tổng quan về Khoa học dữ liệu**
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Làm sạch và tích hợp dữ liệu
- Lecture 4: Phân tích và khám phá dữ liệu
- Lecture 5: Trực quan hoá dữ liệu
- Lecture 6: Trực quan hoá dữ liệu đa biến
- Lecture 7: Học máy
- Lecture 8: Phân tích dữ liệu lớn
- Lecture 9: Báo cáo tiến độ bài tập lớn và hướng dẫn
- Lecture 10+11: Phân tích một số kiểu dữ liệu
- Lecture 12: Đánh giá kết quả phân tích



Mở đầu

- Mục tiêu của Khoa học dữ liệu



Câu hỏi

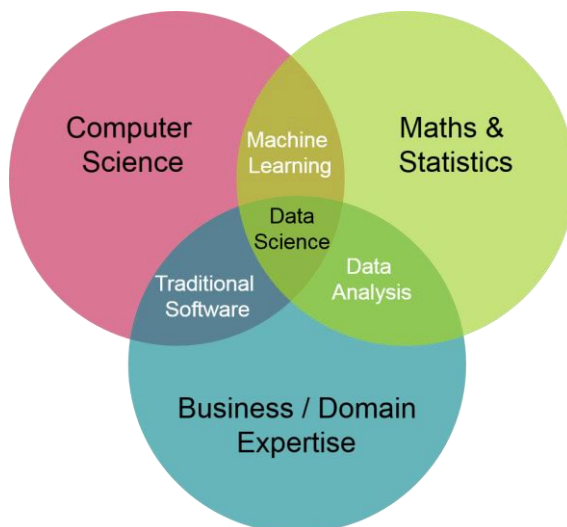
- Vài câu hỏi mà một hệ ra quyết định cần phải trả lời:
 - Sự phát triển về doanh thu của các cửa hàng của tôi theo tháng, theo từng cửa hàng như thế nào?
 - Hồ sơ của các khách hàng chủ yếu của tôi là gì? Dựa trên thông tin đó tôi nên gợi ý sản phẩm nào cho họ?
 - Các sản phẩm nào hay được mua cùng nhau? Làm sao tôi có thể xem xét thông tin đó để tối đa hóa doanh số bán hàng?
 - Liệu có thể dự đoán độ tuổi của khách hàng dựa trên những gì mà họ mua?
 - Nếu tôi chấp nhận cho khách hàng này vay một khoản tiền, liệu họ có khả năng hoàn tiền cho tôi không?
- Các câu hỏi này thường:
 - Cụ thể
 - Đôi khi được thể hiện trong nhiều cách hỏi khác nhau.
 - Không thể biết trước
 → Một bản báo cáo đơn giản là không đủ thông tin!

Thể hiện qua dữ liệu



5

Khoa học dữ liệu (KHDL) là gì?



Khoa học dữ liệu là ngành học (khai phá tri thức) *từ dữ liệu*.

Data science is the science of *learning from data*.

(David Donoho, Stanford University)



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

6

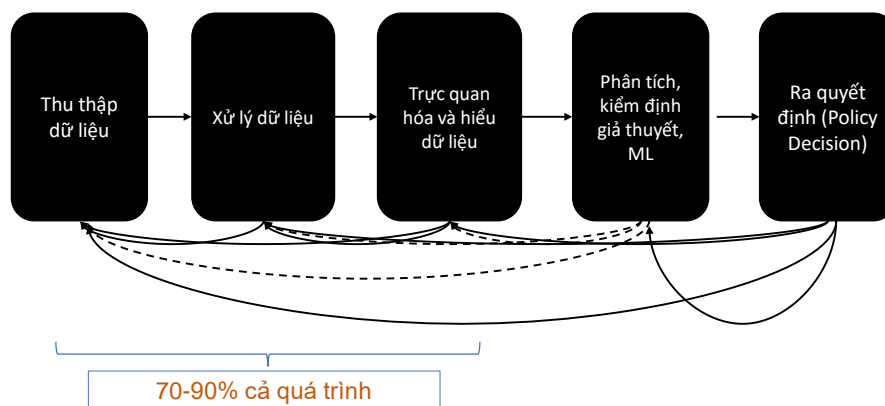
Các mục tiêu của Khoa học dữ liệu

- Các mục tiêu cuối cùng của khoa học dữ liệu có thể được gom thành:
 - Mô tả (Descriptive tasks)
 - Dự đoán (Predictive tasks)
- Để đạt được mục tiêu này đòi hỏi phải thực hiện qua các bước:
 - Thu thập dữ liệu (Data scraping)
 - Tiền xử lý (pre-processing): làm sạch (Data cleaning), biến đổi (transforming), và tích hợp dữ liệu (integration).
 - Học máy
 - Trực quan hóa (Visualization)
- Khoa học dữ liệu có thể áp dụng với nhiều loại (kiểu) dữ liệu khác nhau
 - Dữ liệu thô (số)
 - Dữ liệu văn bản
 - Dữ liệu ảnh, video
 - Dữ liệu đồ thị



7

Quy trình: Hướng tìm kiếm tri thức



(John Dickerson, University of Maryland)



8

Giới thiệu

- Dữ liệu ở đâu?



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

11

Dữ liệu ở đâu? Các mạng xã hội

facebook.



Taylor Swift đã thêm 4 ảnh mới.
4 Tháng 4 lúc 19:52 · 🌐

What an unbelievable run we've had with these memories & all of you. #iHeartAwar

twitter



Basit Alvi @bpk69 · 6m
Swiss banker whistleblower: CIA behind Panama Papers cnb.cx/1WpVigK



Violamagic @TrautCarol · 6m
Why The Panama Papers Scandal Is About Cheating School Children
educationopportunitynetwork.org/why-the-panama...

7,174 Tweets sent in 1 second



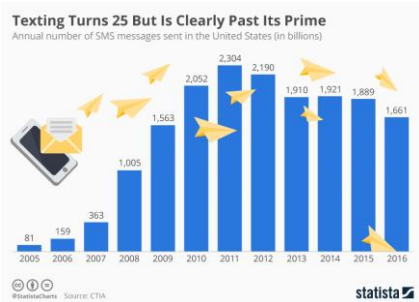
862,696 Tweets since opening this page
0:02:00 seconds ago



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

12

Dữ liệu ở đâu? Tin nhắn di động



Rise and fall of SMS

Rise of messaging apps

WhatsApp Usage Shows No Signs of Slowing Down
Number of WhatsApp messages sent worldwide per day

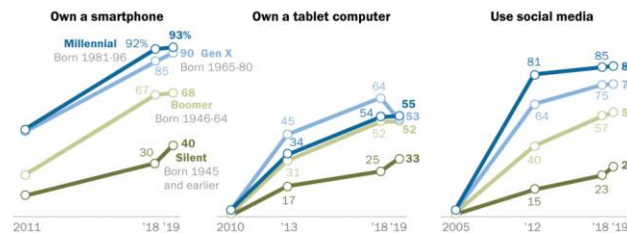


Dữ liệu ở đâu? Internet

- USA:

Millennials lead on some technology adoption measures, but Boomers and Gen Xers are also heavy adopters

% of U.S. adults in each generation who say they ...

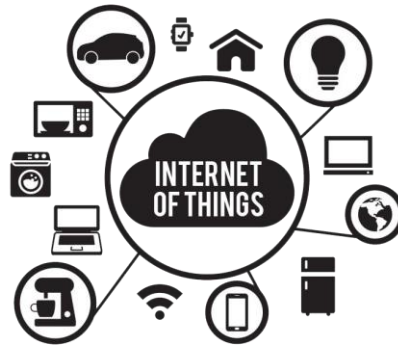
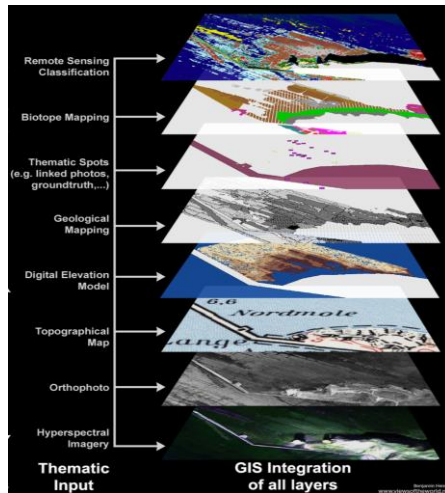


PEW RESEARCH CENTER

- <https://www.internetlivestats.com>



Dữ liệu ở đâu? Nguồn dữ liệu khác

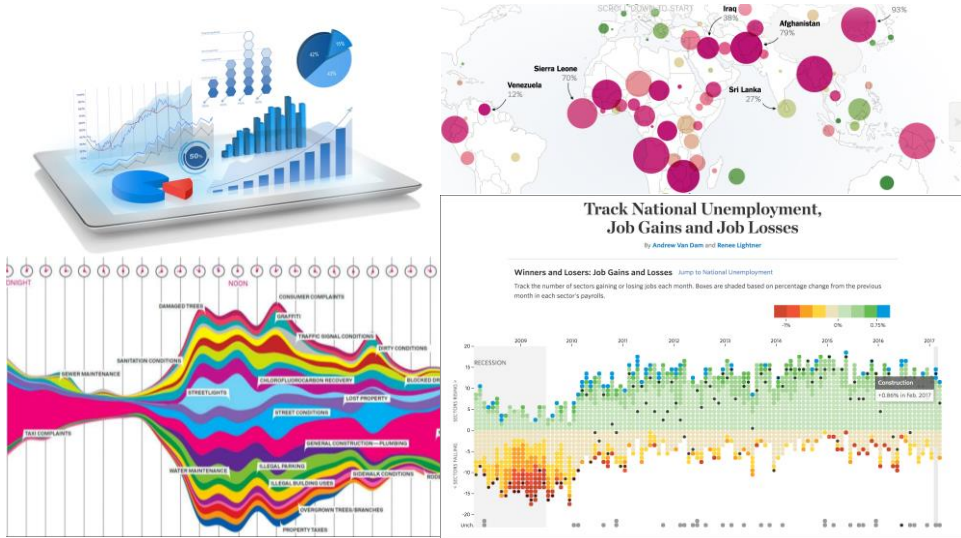


Giới thiệu

- Chúng ta có thể làm gì với dữ liệu?



Có thể làm gì với dữ liệu? Dữ liệu có thể được mô tả thông qua **trực quan hóa**



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

17

Mô tả dữ liệu (Data description)

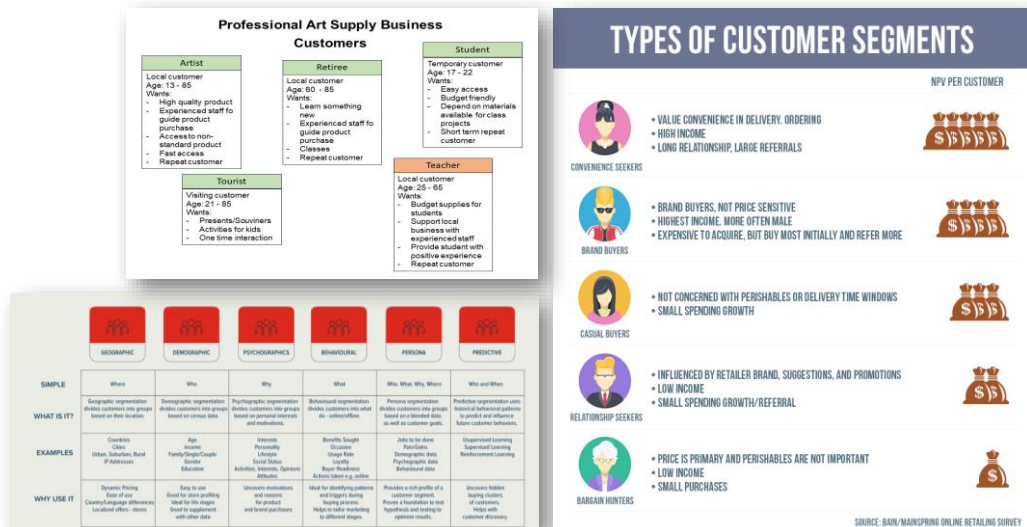
- Mô tả dữ liệu: tóm tắt dữ liệu theo cách “có thể hiểu được”:
 - Thông qua **Phân tích dữ liệu thăm dò (EDA - Exploratory Data Analysis)**
 - Hầu hết các mô tả thống kê đơn giản: average, standard deviation, median, mode, ...
 - Thông qua **Trực quan hóa dữ liệu (data visualization)**



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

18

Có thể làm gì với dữ liệu? **Phân khúc** khách hàng (*Customer segmentation*)



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

19

Phân nhóm dữ liệu (**Data segmentation**)

- Phân nhóm dữ liệu: nhóm các bản ghi dữ liệu giống nhau thành các nhóm đồng nhất (tạo ra các cụm dữ liệu)
 - Các bản ghi trong một nhóm có các giá trị thuộc tính tương tự nhau
 - Mục tiêu học ra một thuộc tính (nhóm) "mới" từ các thuộc tính ở các bản ghi.
 - Có thể sử dụng các phương pháp học **không giám sát** (Unsupervised learning): chương 7



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

20

Có thể làm gì với dữ liệu? Hệ thống gợi ý của Amazon (*association*)



“The company reported a **29% sales increase** to \$12.83 billion during its second fiscal quarter, up from \$9.9 billion during the same time last year.”
– Fortune, July 30, 2012



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

21

Luật kết hợp (*Association rules*)

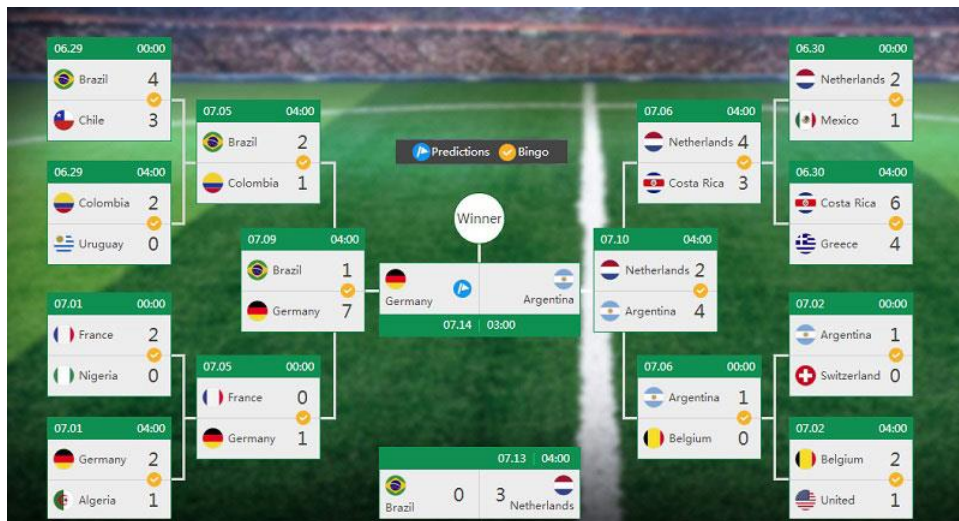
- Khám phá các quy tắc liên kết giữa các bản ghi dữ liệu dựa trên các tiêu chí được xác định trước
 - Vd: các sản phẩm thường được mua cùng nhau trong một lần mua sắm
 - Học ra thông tin “mới” (các quy tắc) dựa trên các thuộc tính của dữ liệu
 - Có thể sử dụng các phương pháp học **không giám sát** (Unsupervised learning): chương 7



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

22

Có thể làm gì với dữ liệu? Dự đoán kết quả FIFA (2014)



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

Accuracy ~93%.

(<http://yourstory.com/2014/07/germany-argentina-fifa-world-cup-2014/>)

23

Dự đoán (Prediction)

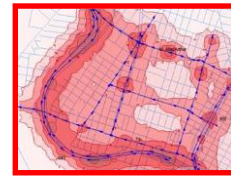
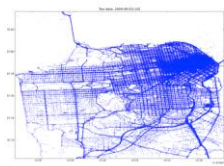
- **Dự đoán** có thể làm:
 - **Dự đoán** hoặc **ước lượng** các giá trị của một thuộc tính cho một tập các bản ghi (điểm) dữ liệu.
 - Thuộc tính được biết bởi các bản ghi khác
 - Trị thức mới sẽ được dùng để dự đoán các giá trị của thuộc tính này trên một tập các bản ghi
 - Phương pháp học có **giám sát** có thể được dùng để giải quyết bài toán này (**Supervised learning**)



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

24

Có thể làm gì với dữ liệu? Hơn thế nữa!!!



Crowdsourcing + physical modeling + sensing + data assimilation
to produce:



(Alex Bayen, UC Berkeley)



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

25

Dữ liệu lớn

- Dữ liệu lớn là gì?



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

26

Dữ liệu lớn – 2008

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

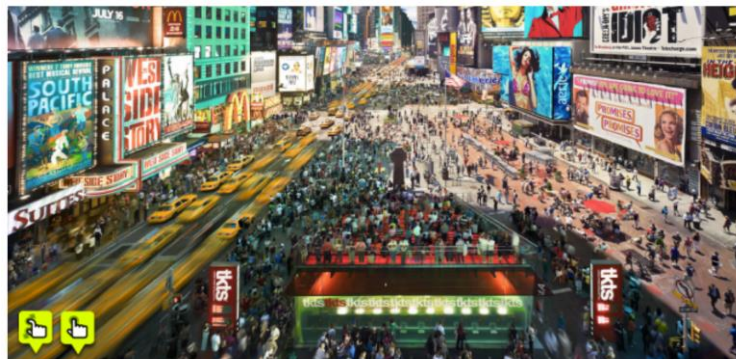
September 2008



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

27

Dữ liệu lớn – 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



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

28

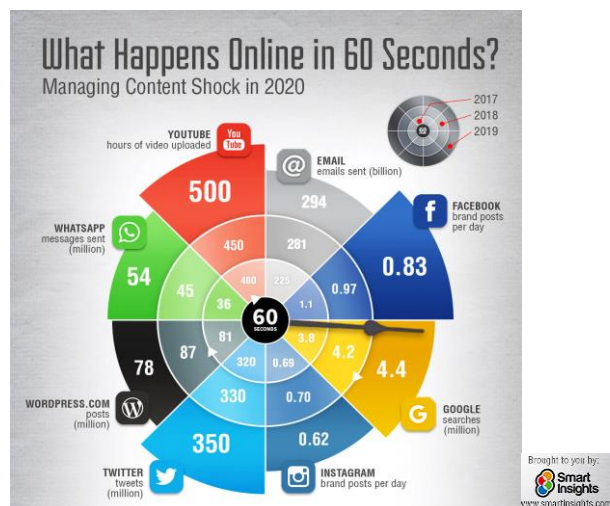
Dữ liệu lớn – ngày nay



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

29

Dữ liệu lớn – ngày nay: Một số thống kê



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

30

Dữ liệu lớn

- Các thách thức



10 thuộc tính của dữ liệu lớn (The 10 Vs of Big data)



[Source: houseofbots.com]



10 Vs of Big data: **Volume**

- **Volume**: là đặc điểm được quan tâm nhiều nhất của dữ liệu lớn
- Hơn 90% dữ liệu từ trước đến nay được tạo từ 2 năm gần đây
- Đặt ra các thách thức:
 - Phân tích dữ liệu khai phá (**Exploratory Data Analysis**): chương 4
 - Trực quan dữ liệu (**Data visualization**): chương 5



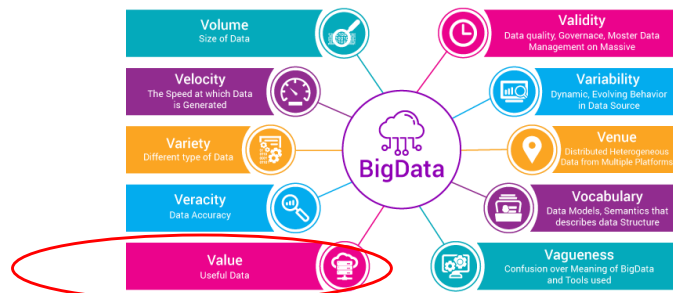
10 Vs of Big data: **Velocity**

- **Velocity**: đề cập đến tốc độ mà dữ liệu được sinh ra, được làm mới
 - Sự gia tăng tốc độ (**velocity**) dẫn đến sự tăng trưởng theo cấp số mũ về khối lượng dữ liệu (**volume**)
 - Đặt ra nhiều thách thức trong việc tích hợp dữ liệu (**data integration**): chương 3



10 Vs of Big data: **Value**

- Khi có quá nhiều dữ liệu vấn đề được đặt ra là giá trị (**Value**) của dữ liệu
 - Do đó phải chọn/ tiền xử lý/ tích hợp (**select / pre-process / integrate**) các dữ liệu có liên quan: chương 2



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

35

10 Vs of Big data: **Validity**

- Tính hợp lệ của dữ liệu (**data validity**)
 - Kiểm tra chất lượng của dữ liệu
 - Kiểm tra tính liên kết với các nguồn dữ liệu khác
 - Loại bỏ nhiễu (ngoại lai)
 - Quá trình này được gọi là **tiền xử lý** dữ liệu, được thực hiện trước khi tiến hành tích hợp dữ liệu (integrating)



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

36

10 Vs of Big data: **Venue**

- **Venue** đề cập đến sự đa dạng của các nguồn dữ liệu (e.g. Excel files, OLTP databases, websites,...)
 - Cần tích hợp dữ liệu (**data integration**): chương 3



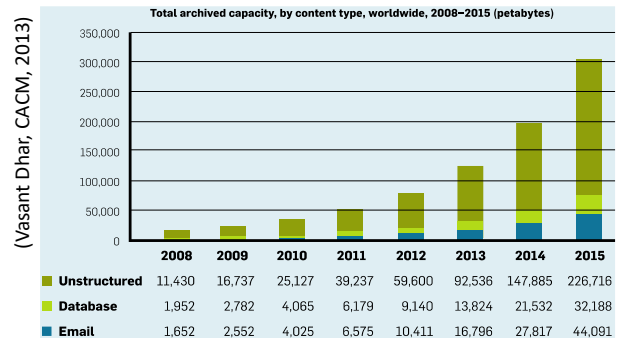
10 Vs of Big data: **Variability**

- Sự thay đổi (**Variability**) đề cập đến
 - Những thay đổi có thể xảy ra trong cấu trúc của nguồn dữ liệu
 - Các tốc độ khác nhau mà khi đó các nguồn dữ liệu được làm mới
 - Đặt ra nhiều khó khăn cho bước tích hợp dữ liệu (**data integration**)



10 Vs of Big data: **Variety**

- Sự đa dạng (**Variety**) đề cập đến các loại dữ liệu khác nhau cần phải được xử lý:
 - Dữ liệu có cấu trúc (**Structured data**): OLTP datasets,...
 - Dữ liệu phi cấu trúc (**Unstructured data**) tăng nhanh: texts, images, tags, links, likes, emotions, ...



10 Vs of Big data: **Vocabulary**

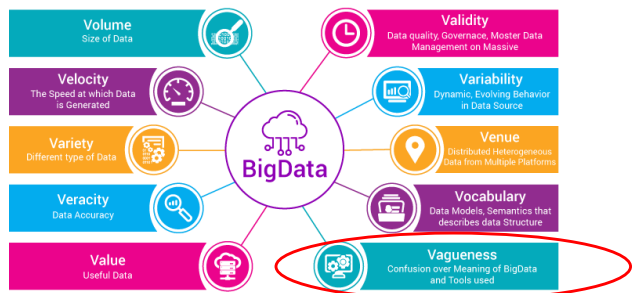
- Vocabulary** đề cập đến việc đem các mô hình dữ liệu / mô hình ngữ nghĩa vào dữ liệu để hiểu cấu trúc / giải thích dữ liệu.
 - Khóa học AI



10 Vs of Big data: **Vagueness**

- **Vagueness** đề cập đến:

- Vấn đề giao tiếp giữa nhà cung cấp và người sử dụng
- Khó khăn đối với một người không phải là chuyên gia để diễn giải kết quả phân tích
 - Vd: sự khác biệt giữa mối tương quan và quan hệ nhân quả



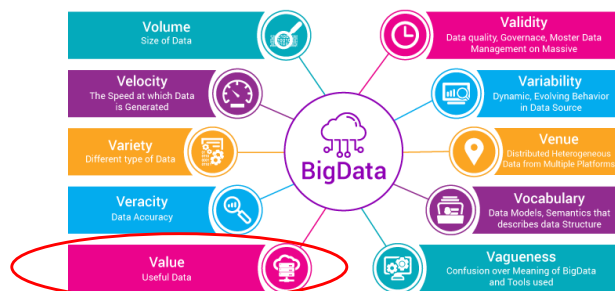
ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

41

10 Vs of Big data: **Veracity**

- **Veracity**: dữ liệu có phản ánh tính thực thế?

- **Không phải tất cả thông tin trên Internet là chính xác!!!**
- Cần kiểm tra chất lượng của các nguồn dữ liệu: xem chương 2
 - Có thể coi là một vấn đề đạo đức (ethical issue)



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

42

Một số thách thức khác

- Các tương tác hoặc mối tương quan ẩn trong dữ liệu có thể thực sự rất lớn
- Các bài toán thực tế thường làm việc trên số chiều lớn (số lượng tham số lớn)
 - Bicycle runs: 2 dimensions (a road)
 - We live in 4 dimensions
 - But an image 1024x1024: ~1 million dimensions
 - Text collections: million dimensions
 - Recommenders' system: billion dimensions (items/products)

→ The **curse of dimensionality**

Dữ liệu dù thu thập được
lớn đến đâu thì cũng là
quá nhỏ so với không
gian của chúng

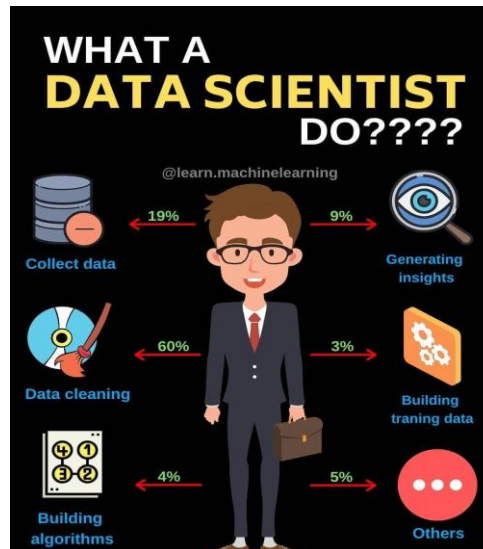


Vấn đề đạo đức

- **Privacy (riêng tư)**
 - Breach of privacy, collection of data without informed consent
- **Security (bảo mật)**
 - The ease of stealing, including identity theft, the stealing of national security information
- **Commercial exploitation (khai thác thương mại)**
 - Commercial mining of information; targeting for commercial gain
- **Issue of Power and politics (quyền lực và chính trị)**
 - The use of data to perpetuate particular views, ideologies, propaganda
- **Issue of Truth (sự thật)**
 - Rumors, hoaxes, fake news
 - Bias introduced by social networks' recommender systems
- **Issue of social justice (công bằng xã hội)**
 - Information is overwhelmingly skewed towards certain groups and leaves others out of the 'digital revolution'



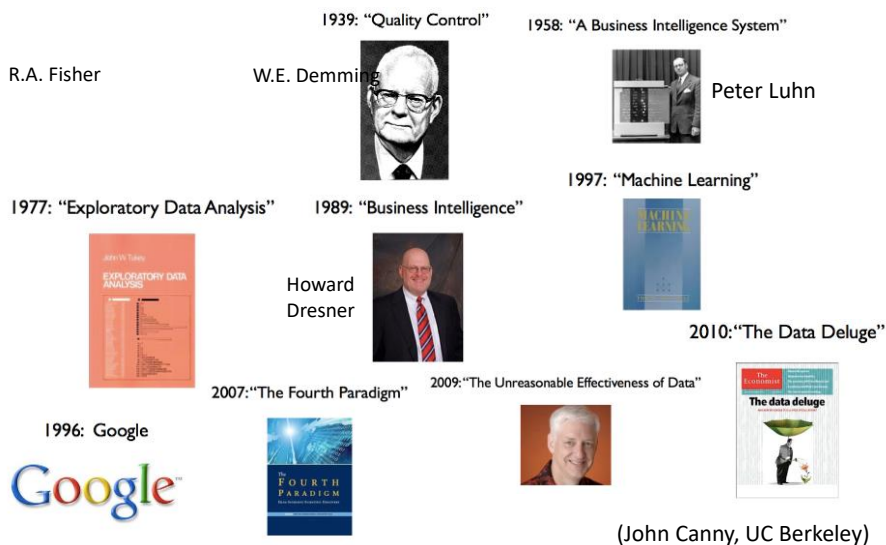
Thế nào là nhà Khoa học dữ liệu?



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

45

Data science - early days



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

46

Sự phát triển của khoa học dữ liệu - 2009

I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

- Hal Varian, Google's Chief Economist, 2009



"The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. **Because now we really do have essentially free and ubiquitous data.**"

- Hal Varian, Google's Chief Economist, 2009



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

47

Data scientist - nowadays

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil



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

48

Skillset

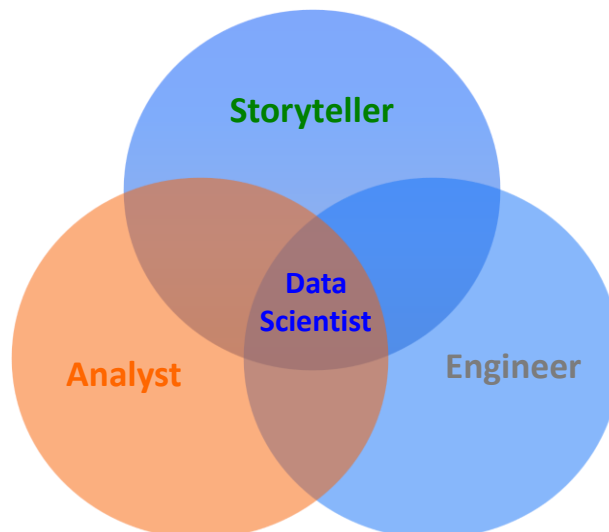


ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

(source: <http://datasciencedojo.com/>)

49

Các vai trò của một nhà khoa học dữ liệu



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

50

Tìm hiểu thêm

- “Job Comparison – Data Scientist vs Data Engineer vs Statistician”
<https://www.analyticsvidhya.com/blog/2015/10/job-comparison-data-scientist-data-engineer-statistician/>
- Big Data Landscape 3.0
<http://mattturck.com/big-data-landscape-2016-v18-final/>
- Ten Lessons Learned from Building (real-life impactful) Machine Learning Systems
<http://technocalifornia.blogspot.com/2014/12/ten-lessons-learned-from-building-real.html>



Tài liệu tham khảo

- John Dickerson. *Lectures on Introduction to Data Science*. University of Maryland, 2017.
- Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 43, 2017.
- Longbing Cao. Data science: nature and pitfalls. *IEEE Intelligent Systems*, 31(5), 66-75, 2016.
- David Donoho. "50 years of Data Science." In *Princeton NJ, Tukey Centennial Workshop*. 2015.
- L. Duan, Y. Xiong. Big data analytics and business analytics. *Journal of Management Analytics*, vol 2 (2), pp 1-21, 2015.
- X. Wu, X. Zhu, G. Wu, W. Ding. Data mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering*, vol 26 (1), pp 97-107, 2014.
- Rafael Irizarry & Verena Kaynig-Fittau. *Lectures on Data Science*. Harvard Univ., 2014.
- John Canny. *Lectures on Introduction to Data Science*. University of California, Berkeley, 2014.
- Vasant Dhar. Data Science and Prediction. *Communication of the ACM*, vol 56 (12), pp 64-73, 2013.
- Michael Perrone. *What is Watson – an overview*. 2011.





HUST

 hust.edu.vn  fb.com/dhbkhn

THANK YOU !

53