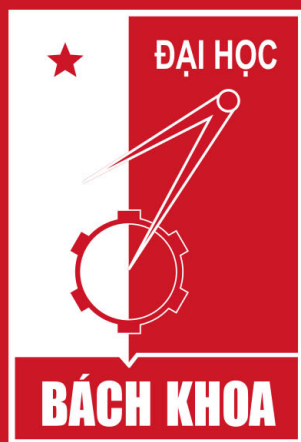




25 YEARS ANNIVERSARY  
SOICT

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

# Nhập môn Khoa học dữ liệu (IT4930)

# 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



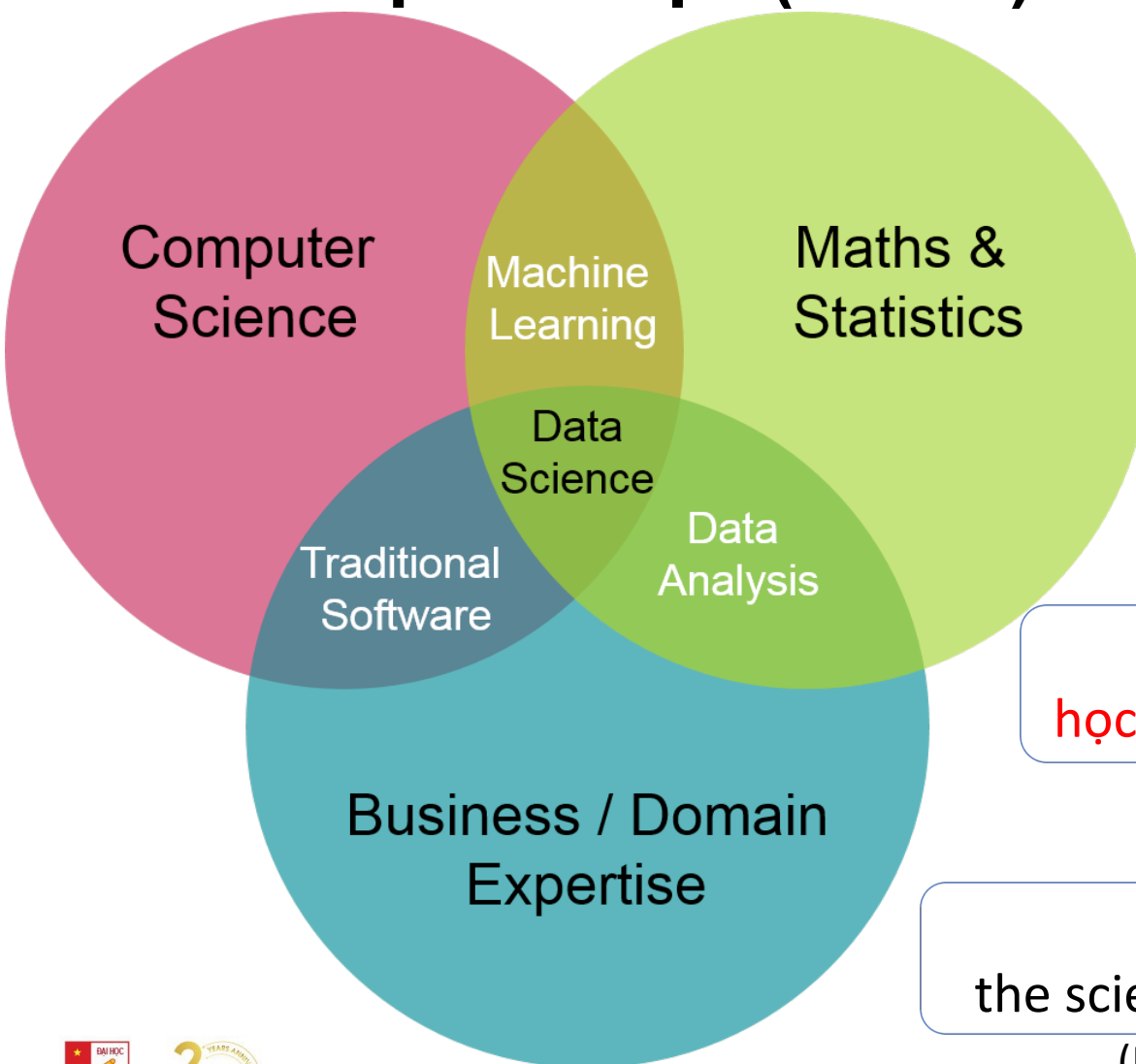
# Câu hỏi

- 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...
  - Các... có thể xem xét t...
  - Liệu... ng gì mà họ mua...
  - Nếu... n, liệu họ có khả n...

## Thể hiện qua dữ liệu

- 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!

# Khoa học dữ liệu (KHDL) là gì?



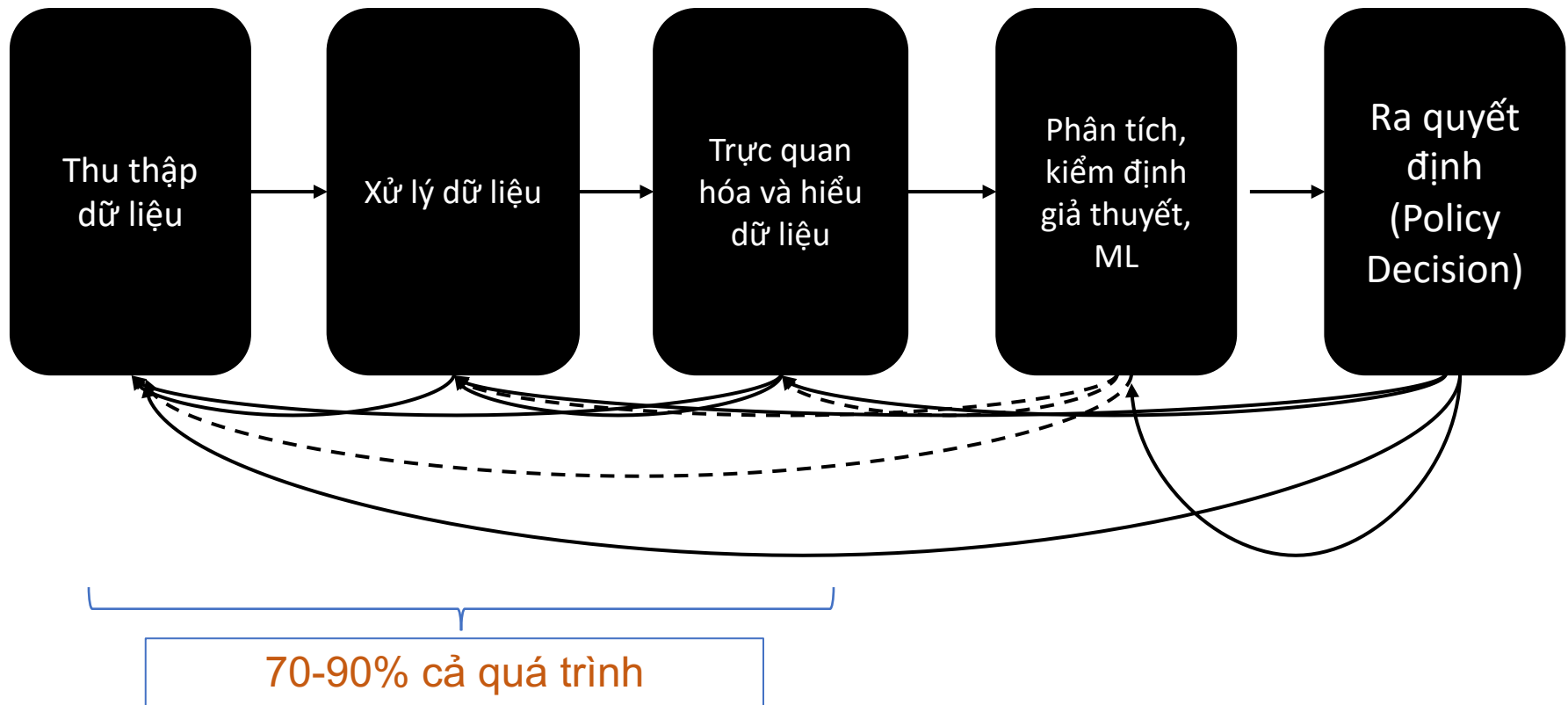
**Khoa học dữ liệu** là ngành học (khai phá tri thức) từ dữ liệu.

**Data science** is the science of *learning from data*.  
(David Donoho, Stanford University)

# Các mục tiêu của Khoa học dữ liệu

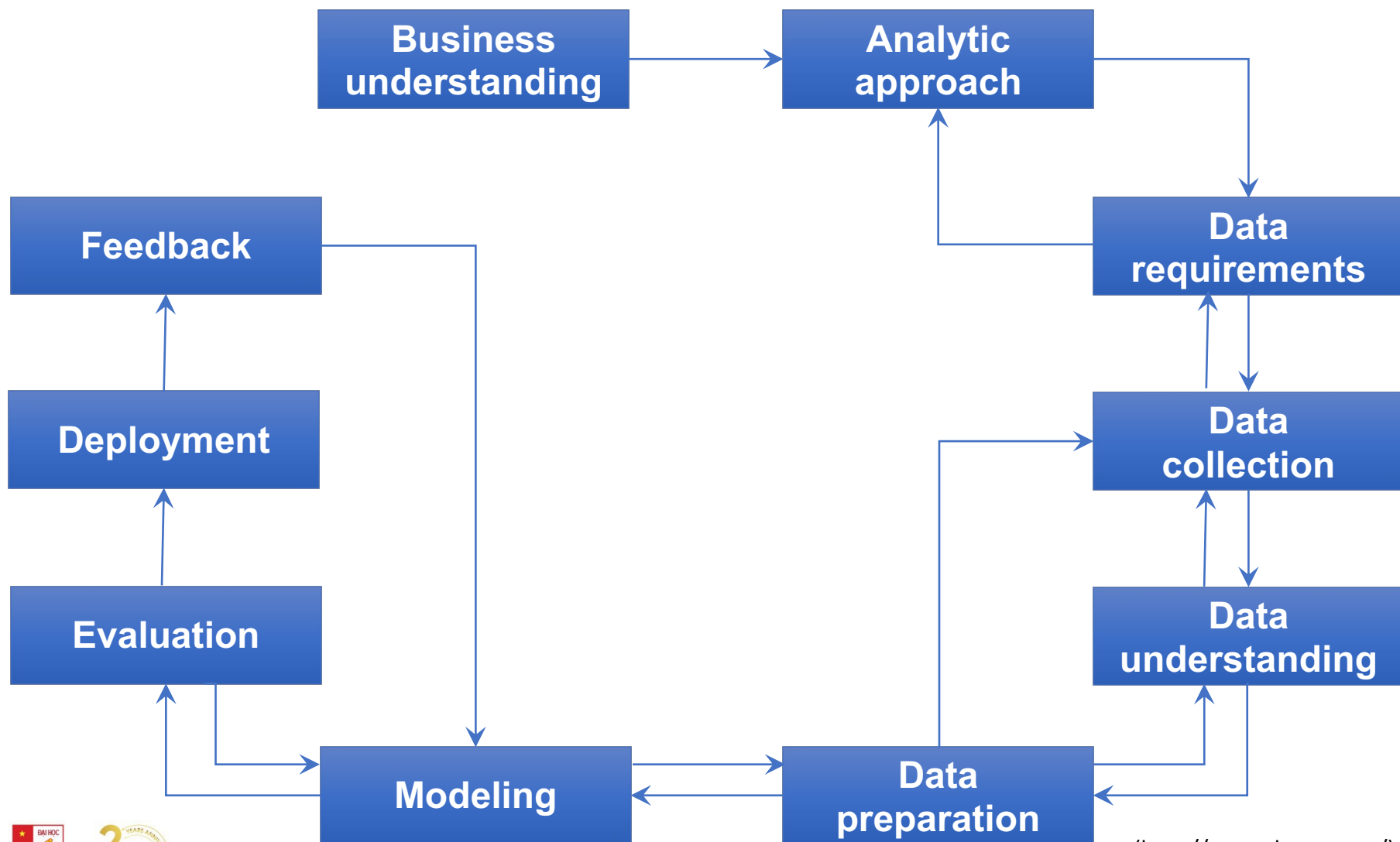
- Các mục tiêu của khoa học dữ liệu có thể được gom thành:
  - Mô tả (Descriptive tasks)
  - Dự đoán (Predictive tasks)
- Để đạt được mục tiêu này thường phải thực hiện qua các bước:
  - Thu thập dữ liệu (Data scraping)
  - Tiền xử lý (pre-processing): làm sạch (Data cleaning), biến đổi (transforming), 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, đồ thị, ...

# Quy trình: **hướng tìm kiếm tri thức**



(John Dickerson, University of Maryland)

# Quy trình: **hướng tạo sản phẩm**



(<http://www.theta.co.nz/>)

# Một số nền tảng trực tuyến cho các cuộc thi KHDL



Analytics, Data Science, Data Mining  
Competitions

kaggle™

## Notable Recent Competitions

- [GE NFL \\$10 Million Head Health Challenge](#), for more accurate diagnoses of mild brain injury and prognosis for recovery following acute and/or repetitive injuries.
- [GE Hospital Quest on Kaggle](#).  
Your challenge: Contribute to the design of the ultimate patient experience. Prize Pool: \$100,000
- [GE Flight Quest on Kaggle](#).  
Your Challenge: Develop a usable and scalable algorithm that determines real-time flight profile to the pilot, helping them make flights more efficient and reliably on time. Prize Pool: \$250,000
- [Heritage Health Data Analysis Prize \(\\$3M\)](#), can administer health care data be used to accurately predict which patients

jobs board >  
**Your Home for Data Science**  
Kaggle helps you learn, work, and play



Create an account

or

Host a competition

Competitions ›

Climb the world's most elite machine learning leaderboards

Datasets ›

Explore and analyze a collection of high quality public datasets

Kernels ›

Run code in the cloud and receive community feedback on your work

# Giới thiệu

Dữ liệu ở đâu?

# 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

EACH DAY

50%  
of active FB users log in

55 million  
status updates are made



Pages have created  
5.30 billion  
of fans

35 million  
update their status



**Basit Alvi** @bpk69 · 6m

Swiss banker whistleblower: CIA behind **Panama Papers** [cnb.cx/1WpVjgK](http://cnb.cx/1WpVjgK)



View summary



**Violamagic** @TrautCarol · 6m

Why The **Panama Papers** Scandal Is About Cheating School Children  
[educationopportunitynetwork.org/why-the-panama...](http://educationopportunitynetwork.org/why-the-panama...)



View summary

7,174 Tweets sent in 1 second



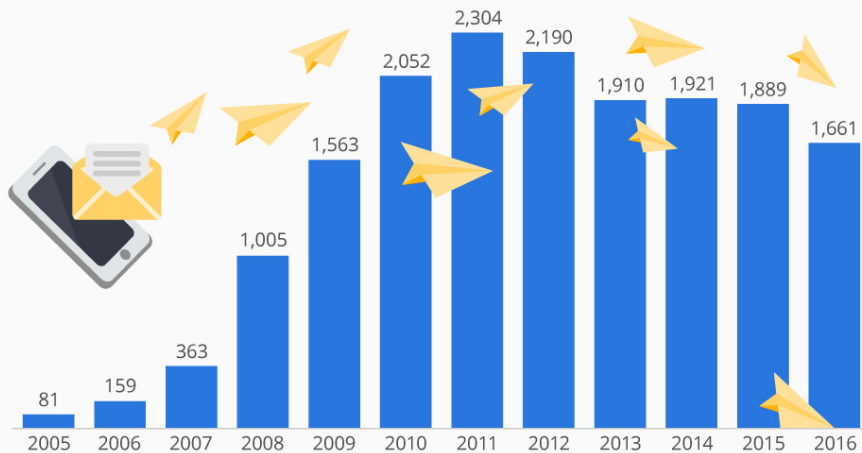
862,696 Tweets since opening this page  
0:02:00 seconds ago



# Dữ liệu ở đâu? Tin nhắn di động

## Texting Turns 25 But Is Clearly Past Its Prime

Annual number of SMS messages sent in the United States (in billions)



@StatistaCharts Source: CTIA

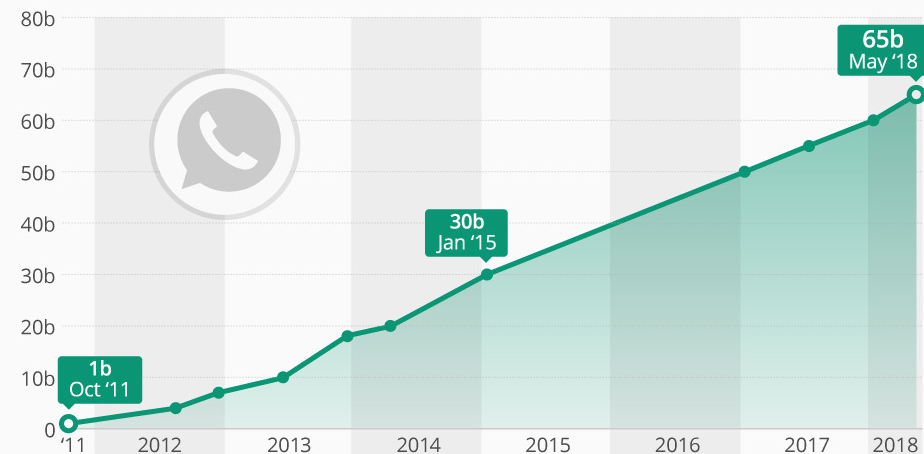
statista

## 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\*



@StatistaCharts

\* a message sent to a WhatsApp group is counted as one sent message  
Source: Company announcements

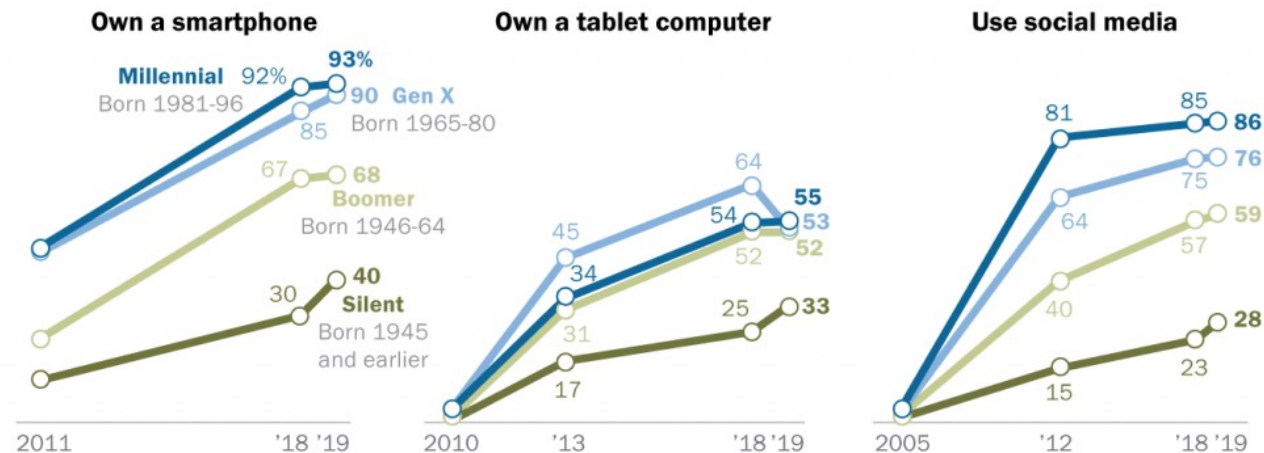
statista

# 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 ...*



Note: Those who did not give an answer are not shown.

Source: Survey conducted Jan. 8 - Feb. 7, 2019.

PEW RESEARCH CENTER

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



Là cái gì

nột hộp sản phẩm "Bioscalin NOVA Genina" với chức năng "Shampoo Fortificante Rivitalizzante," có nghĩa là dầu gội tăng cường và phục hồi, giúp tăng cường sức khỏe của tóc và da đầu, đặc biệt là cho những người có tóc yếu. Sản phẩm này được sản xuất bởi công ty Giuliani.

is sq.

## Dữ liệu ở đâu? GenAI

Yes, the inequality  $\sqrt{a^2 + b^2} \leq a + b$  holds for all non-negative real numbers  $a$  and  $b$ .

Here's a proof:

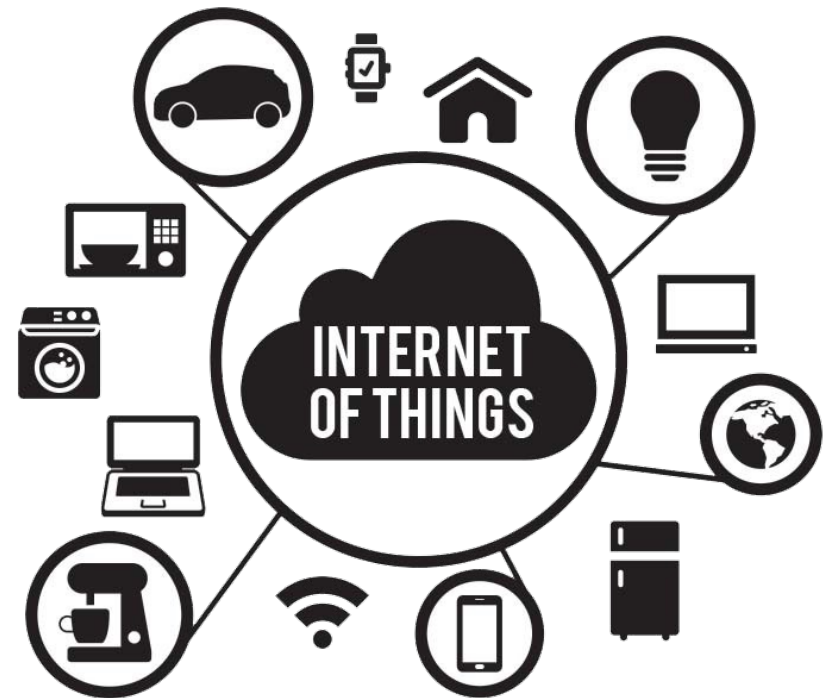
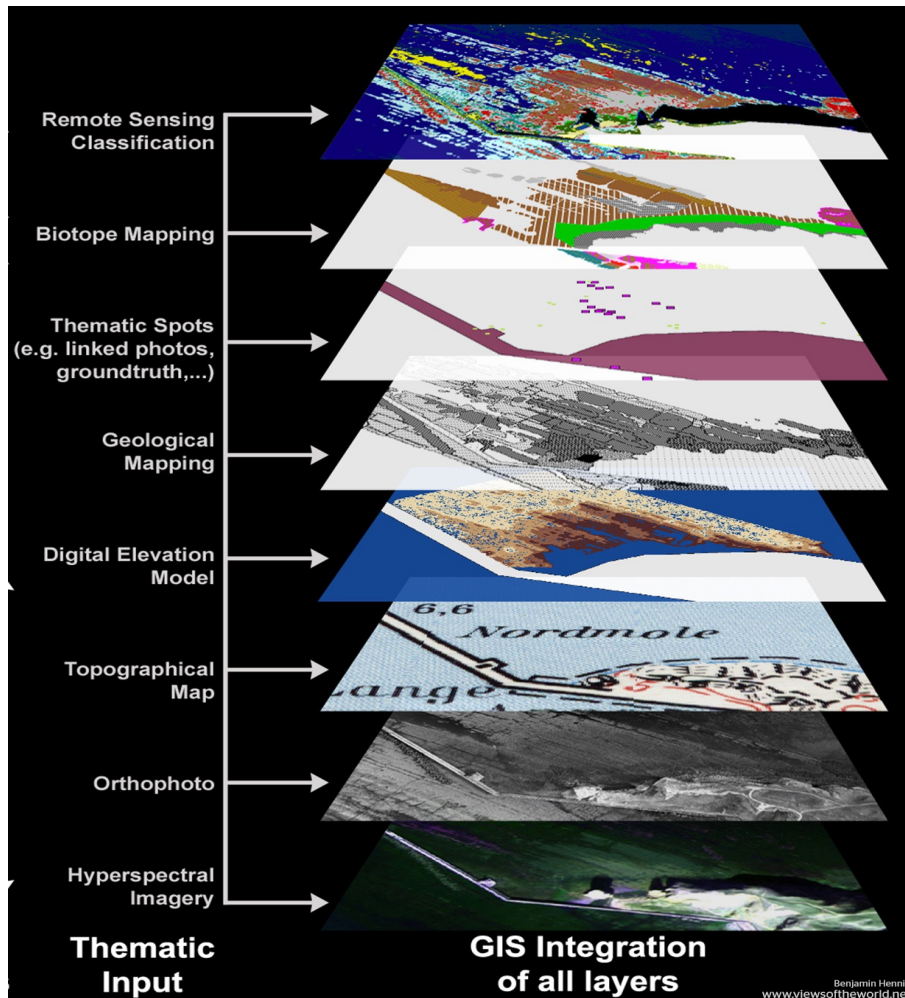
**Proof:**

1. Start with the inequality we want to prove:

$$\sqrt{a^2 + b^2} \leq a + b$$



# Dữ liệu ở đâu? Nguồn dữ liệu khác

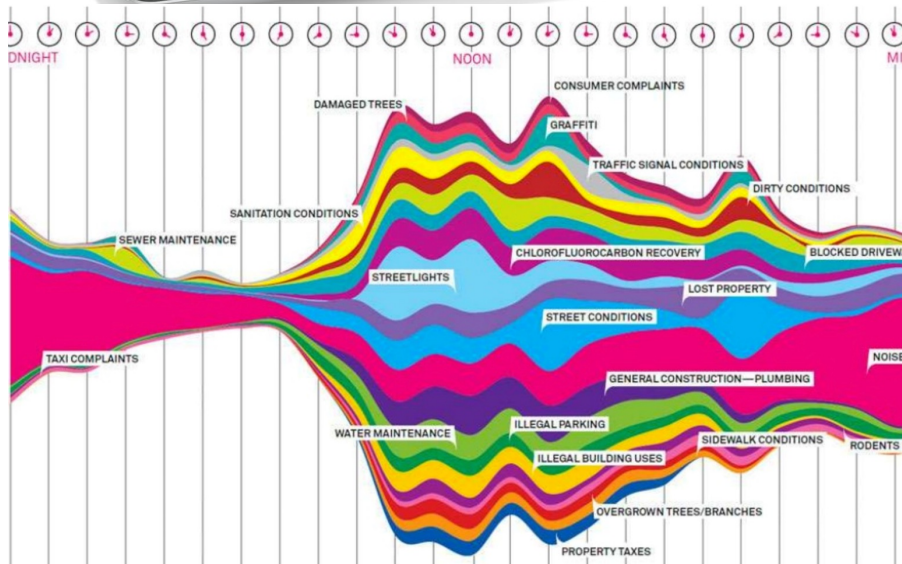
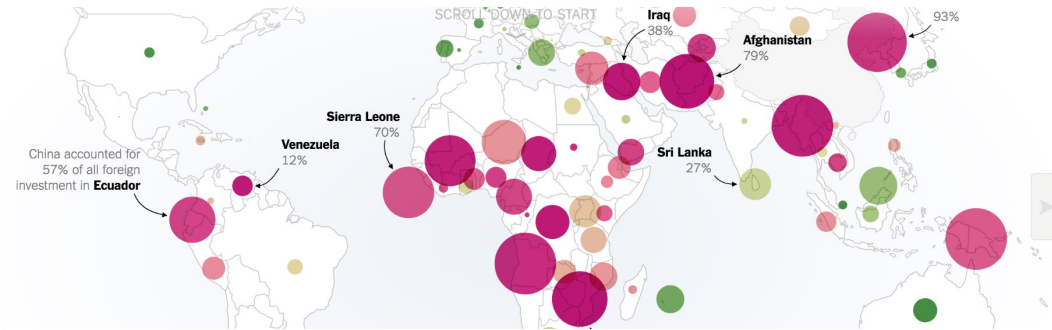


# Giới thiệu

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**

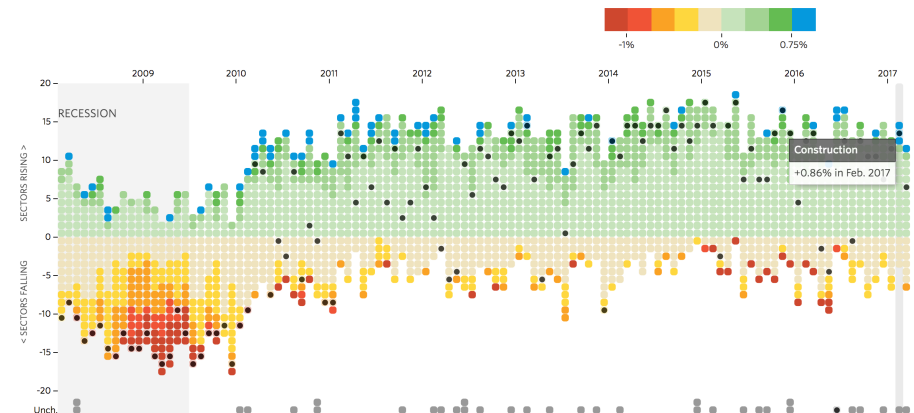


### Track National Unemployment, Job Gains and Job Losses

By Andrew Van Dam and Renee Lightner

#### Winners and Losers: Job Gains and Losses [Jump to National Unemployment](#)

Track the number of sectors gaining or losing jobs each month. Boxes are shaded based on percentage change from the previous month in each sector's payrolls.



# 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 là các mô tả thống kê đơn giản: average, standard deviation, median, mode, ...
  - Thông qua trực quan hóa dữ liệu (data visualization)

# Có thể làm gì với dữ liệu?

## Phân khúc khách hàng (*Customer segmentation*)

### Professional Art Supply Business

#### Customers

| Artist                                                                                                                                                                                               | Retiree                                                                                                                                                                                                 | Student                                                                                                                                                                         |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Local customer</b><br>Age: 13 - 85<br>Wants:<br>- High quality product<br>- Experienced staff to guide product purchase<br>- Access to non-standard product<br>- Fast access<br>- Repeat customer | <b>Local customer</b><br>Age: 60 - 85<br>Wants:<br>- Learn something new<br>- Experienced staff to guide product purchase<br>- Classes<br>- Repeat customer                                             | <b>Temporary customer</b><br>Age: 17 - 22<br>Wants:<br>- Easy access<br>- Budget friendly<br>- Depend on materials available for class projects<br>- Short term repeat customer |
| Tourist                                                                                                                                                                                              | Teacher                                                                                                                                                                                                 |                                                                                                                                                                                 |
| <b>Visiting customer</b><br>Age: 21 - 85<br>Wants:<br>- Presents/Souvenirs<br>- Activities for kids<br>- One time interaction                                                                        | <b>Local customer</b><br>Age: 25 - 65<br>Wants:<br>- Budget supplies for students<br>- Support local business with experienced staff<br>- Provide student with positive experience<br>- Repeat customer |                                                                                                                                                                                 |

### TYPES OF CUSTOMER SEGMENTS

|                                                                                                             | NPV PER CUSTOMER                                                                                                                                                                                                                                                                            |
|-------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <br>CONVENIENCE SEEKERS  | <ul style="list-style-type: none"> <li>• VALUE CONVENIENCE IN DELIVERY, ORDERING</li> <li>• HIGH INCOME</li> <li>• LONG RELATIONSHIP, LARGE REFERRALS</li> </ul>                                         |
| <br>BRAND BUYERS         | <ul style="list-style-type: none"> <li>• BRAND BUYERS, NOT PRICE SENSITIVE</li> <li>• HIGHEST INCOME. MORE OFTEN MALE</li> <li>• EXPENSIVE TO ACQUIRE, BUT BUY MOST INITIALLY AND REFER MORE</li> </ul>  |
| <br>CASUAL BUYERS        | <ul style="list-style-type: none"> <li>• NOT CONCERNED WITH PERISHABLES OR DELIVERY TIME WINDOWS</li> <li>• SMALL SPENDING GROWTH</li> </ul>                                                             |
| <br>RELATIONSHIP SEEKERS | <ul style="list-style-type: none"> <li>• INFLUENCED BY RETAILER BRAND, SUGGESTIONS, AND PROMOTIONS</li> <li>• LOW INCOME</li> <li>• SMALL SPENDING GROWTH/REFERRAL</li> </ul>                            |
| <br>BARGAIN HUNTERS      | <ul style="list-style-type: none"> <li>• PRICE IS PRIMARY AND PERISHABLES ARE NOT IMPORTANT</li> <li>• LOW INCOME</li> <li>• SMALL PURCHASES</li> </ul>                                                  |

SOURCE: BAIN/MAINSRING ONLINE RETAILING SURVEY

|                    | GEOGRAPHIC                                                                                  | DEMOGRAPHIC                                                                                            | PSYCHOGRAPHICS                                                                                         | BEHAVIOURAL                                                                                                       | PERSONA                                                                                                                | PREDICTIVE                                                                                                      |
|--------------------|---------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| <b>WHAT IS IT?</b> | Where                                                                                       | Who                                                                                                    | Why                                                                                                    | What                                                                                                              | Who, What, Why, Where                                                                                                  | Who and When                                                                                                    |
| <b>EXAMPLES</b>    | Geographic segmentation divides customers into groups based on their location.              | Demographic segmentation divides customers into groups based on census data.                           | Psychographic segmentation divides customers into groups based on personal interests and motivations.  | Behavioural segmentation divides customers into what do - online/offline.                                         | Persona segmentation divides customers into groups based on a blended data, as well as customer goals.                 | Predictive segmentation uses historical behavioral patterns to predict and influence future customer behaviors. |
| <b>WHY USE IT</b>  | Countries<br>Cities<br>Urban, Suburban, Rural<br>IP Addresses                               | Age<br>Income<br>Family/Single/Couple<br>Gender<br>Education                                           | Interests<br>Personality<br>Lifestyle<br>Social Status<br>Activities, Interests, Opinions<br>Attitudes | Benefits Sought<br>Occasion<br>Usage Rate<br>Loyalty<br>Buyer Readiness<br>Actions taken e.g. online              | Jobs to be done<br>Pain/Gains<br>Demographic data<br>Psychographic data<br>Behavioural data                            | Unsupervised Learning<br>Supervised Learning<br>Reinforcement Learning                                          |
| <b>WHY USE IT</b>  | Dynamic Pricing<br>Ease of use<br>Country/Language differences<br>Localized offers - stores | Easy to use<br>Good for store profiling<br>Ideal for life stages<br>Good to supplement with other data | Uncovers motivations and reasons for product and brand purchases                                       | Ideal for identifying patterns and triggers during buying process. Helps to tailor marketing to different stages. | Provides a rich profile of a customer segment. Proves a foundation to test hypothesis and testing to optimize results. | Uncovers hidden buying clusters of customers. Helps with customer discovery.                                    |



# 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

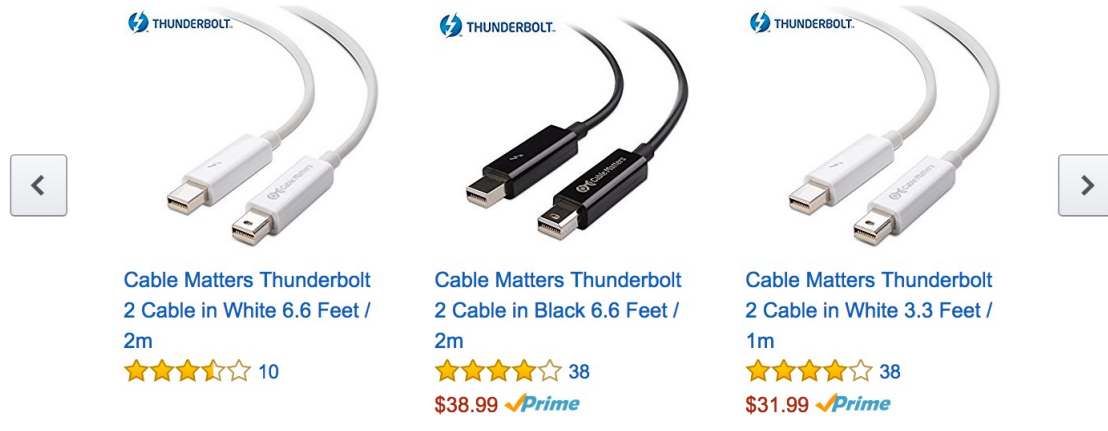
# Có thể làm gì với dữ liệu?

## Hệ thống gợi ý của Amazon (*association*)



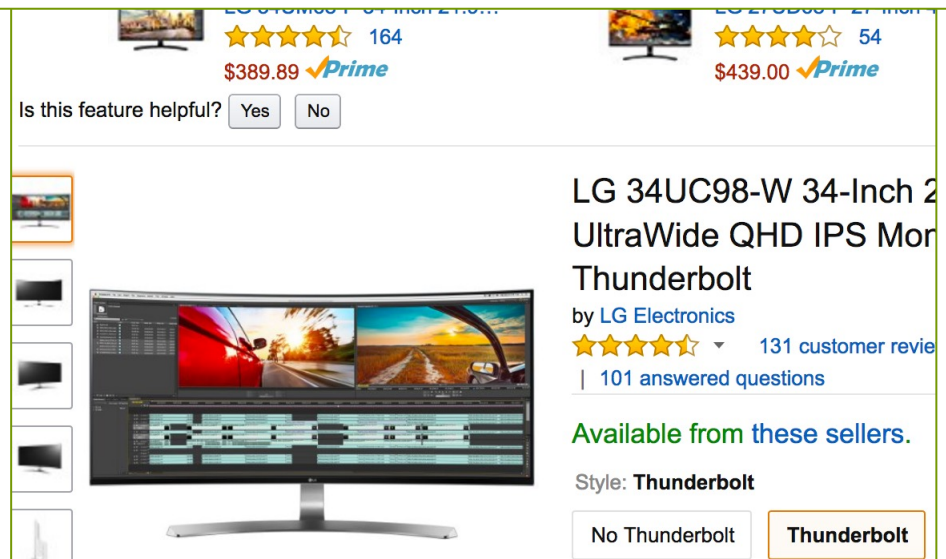
### Customers Who Bought This Item Also Bought

Page 1 of 31



“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

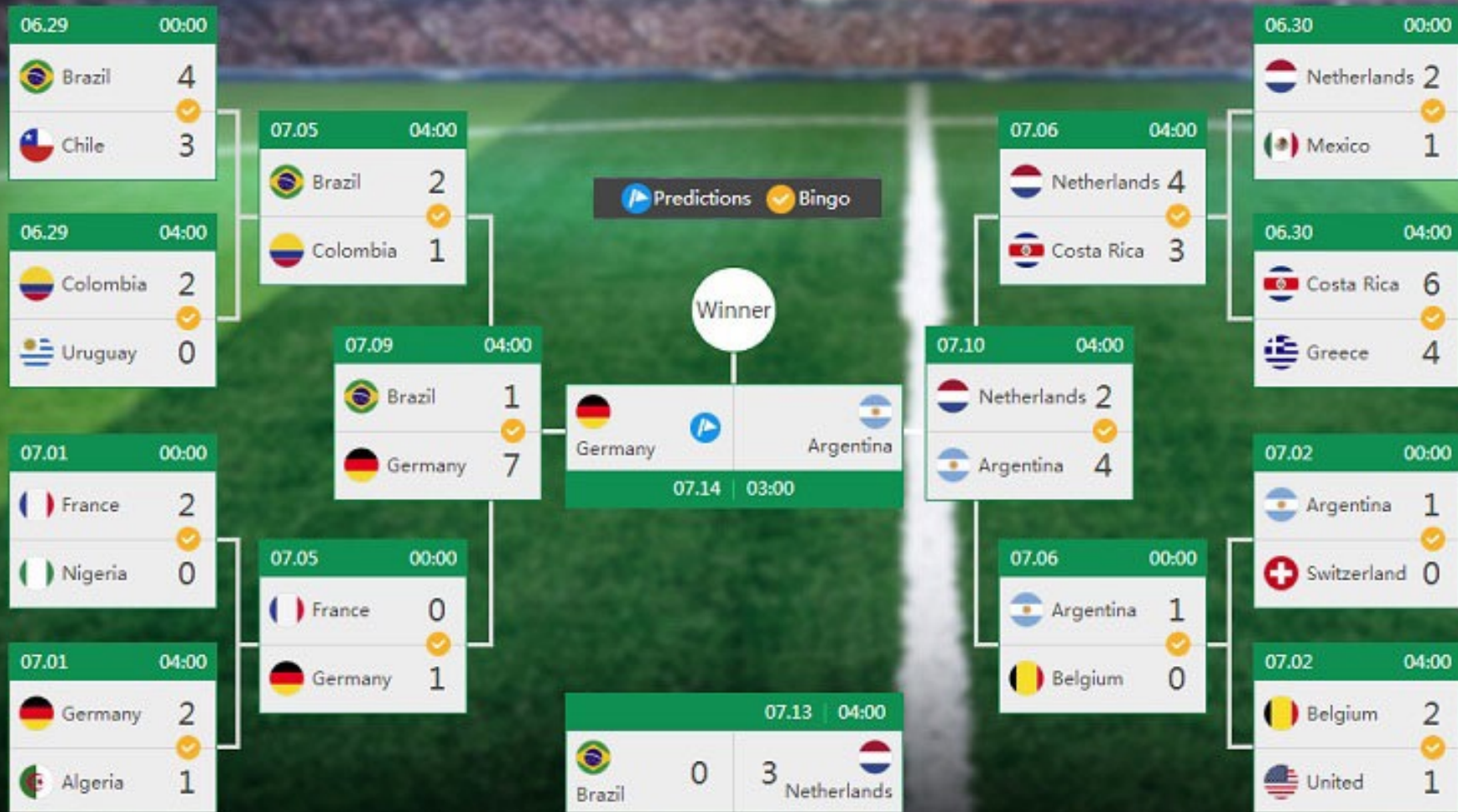


# Luật kết hợp (**Association** rules)

- 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

# Có thể làm gì với dữ liệu?

## *Dự đoán kết quả FIFA (2014)*



Accuracy ~93%.

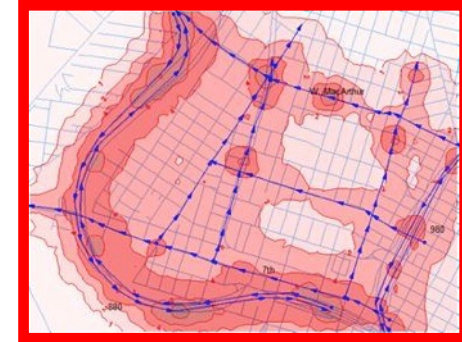
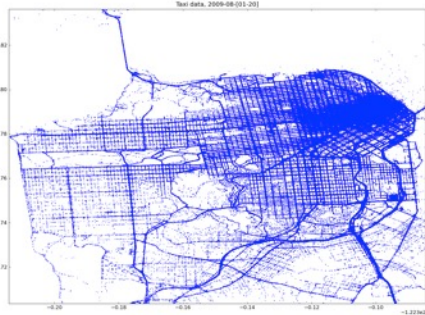
(<http://yourstory.com/2014/07/germany-argentina-fifa-world-cup-2014/>)

# 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
    - Tri 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**)

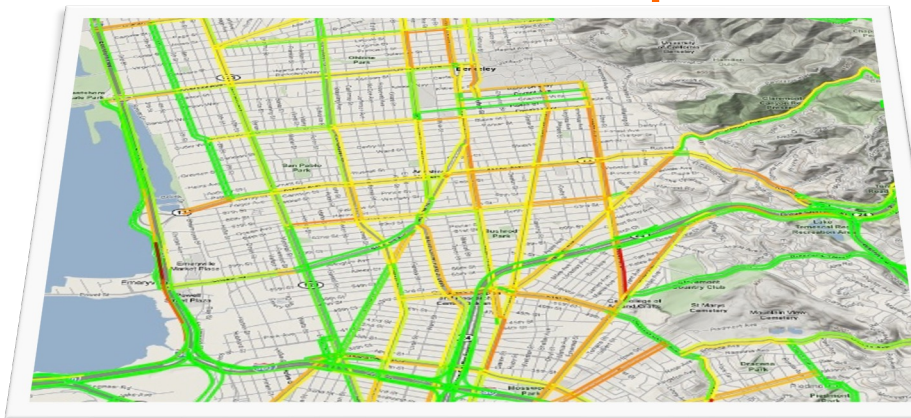


# Có thể làm gì với dữ liệu? *Hơn thế nữa!!!*



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



(Alex Bayen, UC Berkeley)

# Dữ liệu lớn

Dữ liệu lớn là gì?



# Dữ liệu lớn – 2008

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

September 2008



# 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



## 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**

# Dữ liệu lớn – ngày nay: một số thống kê

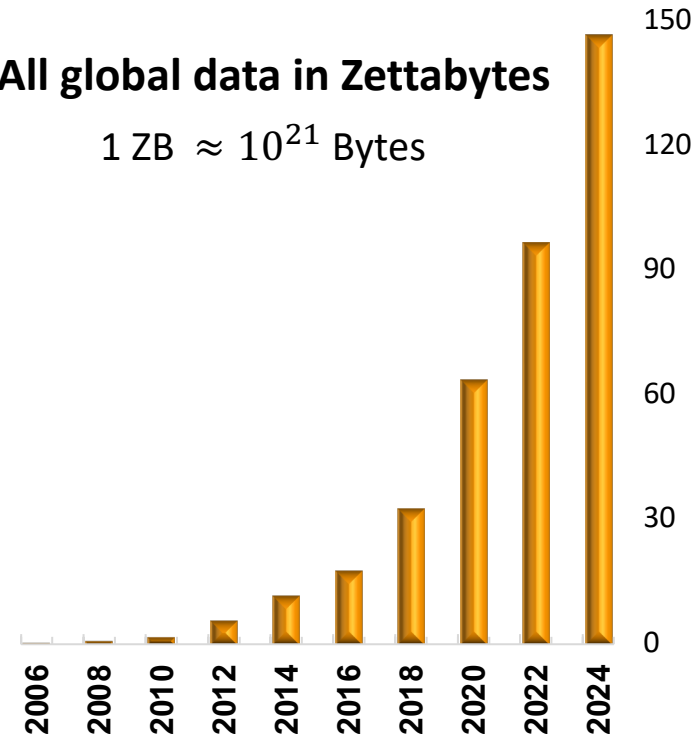
## What Happens Online in 60 Seconds?

Managing Content Shock in 2020



## All global data in Zettabytes

1 ZB  $\approx 10^{21}$  Bytes



Brought to you by:



www.smartinsights.com

# Dữ liệu lớn

Các thách thức

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

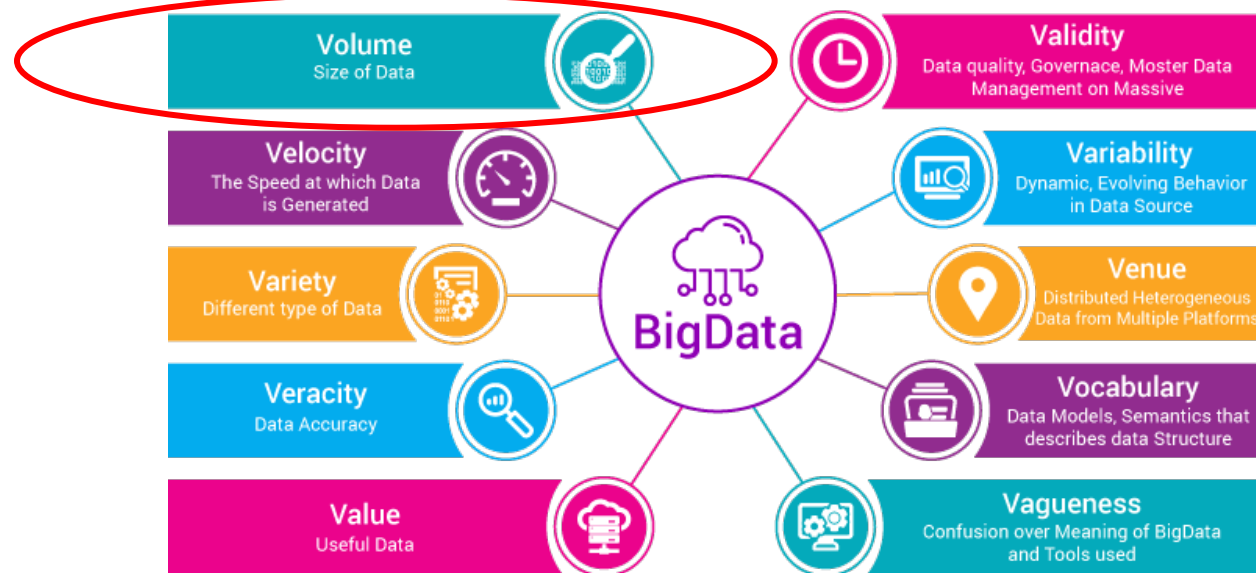


[Source: houseofbots.com]



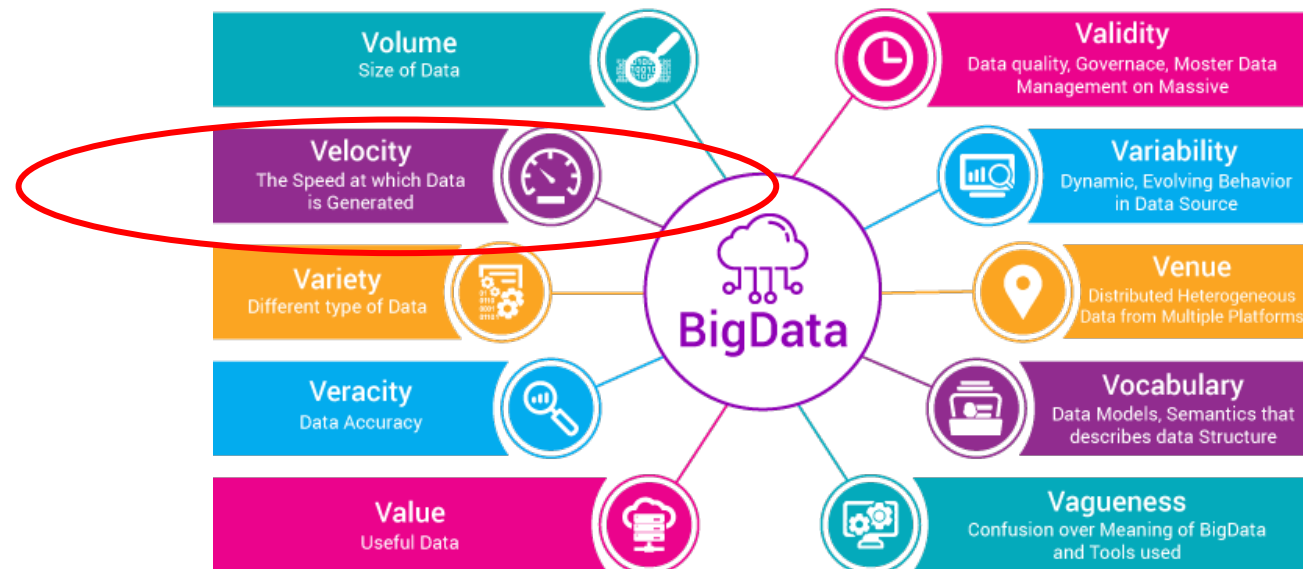
# 10 Vs of Big data: **Volume**

- **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



# The 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



# The 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



# The 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)



# The 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





# The 10 Vs of Big data: **Variability**

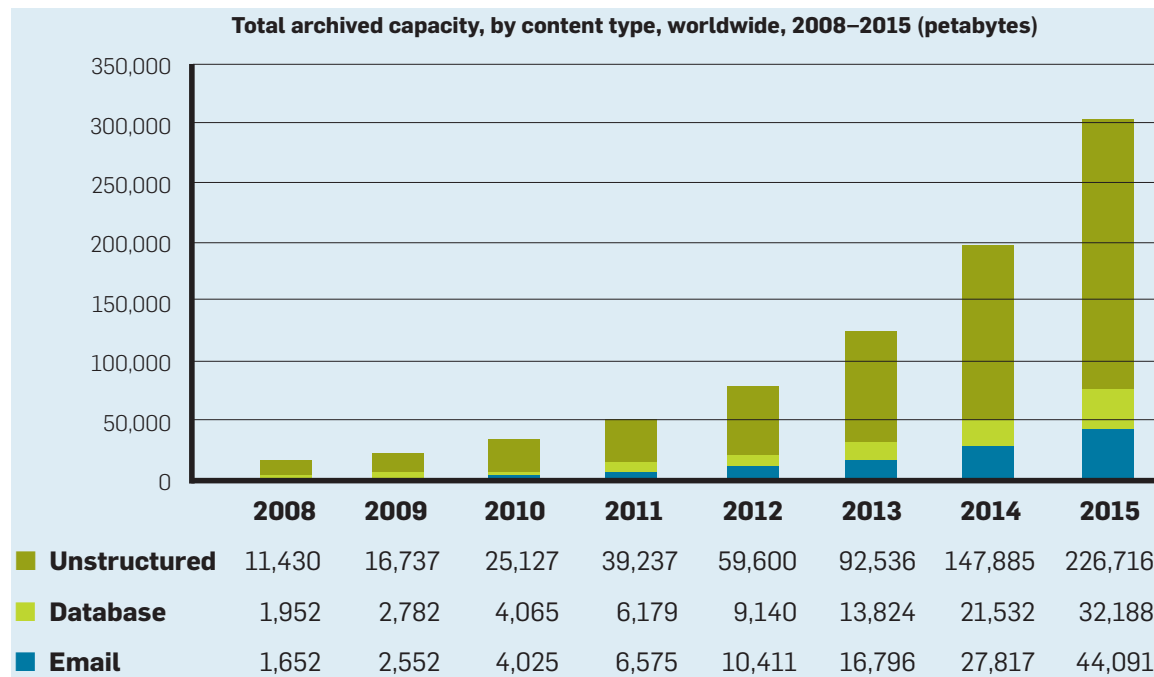
- Sự thay đổi (**Variability**) đề cập đến
  - Những thay đổi có thể xảy ra trong cấu trúc của nguồn dữ liệu
  - Các tốc độ khác nhau mà khi đó các nguồn dữ liệu được làm mới
  - Đặt ra nhiều khó khăn cho bước tích hợp dữ liệu (**data integration**)



# The 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, ...

(Vasant Dhar, CACM, 2013)



# The 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



# The 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*



# The 10 Vs of Big data: **Veracity**

- **Veracity**: dữ liệu có phản ánh đúng thực tế?
  - **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 liên quan đến vấn đề đạo đức (ethical issue)





# 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)
  - Xe đạp: 2 chiều
  - Vũ trụ chúng ta: 4 chiều
  - Bức ảnh 1024x1024:  $\sim 10^6$  chiều
  - Văn bản: triệu chiều
  - Hệ thống gợi ý: tỷ chiều (số sản phẩm)

→ Lời nguyền của số chiều!

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?

# Data science - early days

1935: "The Design of Experiments"

R.A. Fisher



1939: "Quality Control"

W.E. Demming



1958: "A Business Intelligence System"

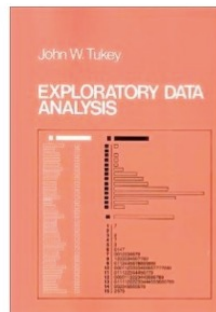


Peter Luhn

1997: "Machine Learning"



1977: "Exploratory Data Analysis"

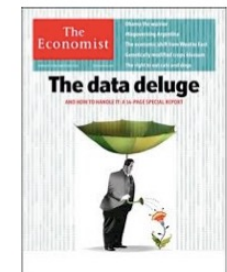


1989: "Business Intelligence"

Howard  
Dresner



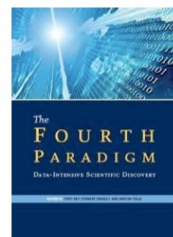
2010: "The Data Deluge"



2009: "The Unreasonable Effectiveness of Data"



2007: "The Fourth Paradigm"



1996: Google



(John Canny, UC Berkeley)

# 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



# Data scientist - nowadays

## Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil



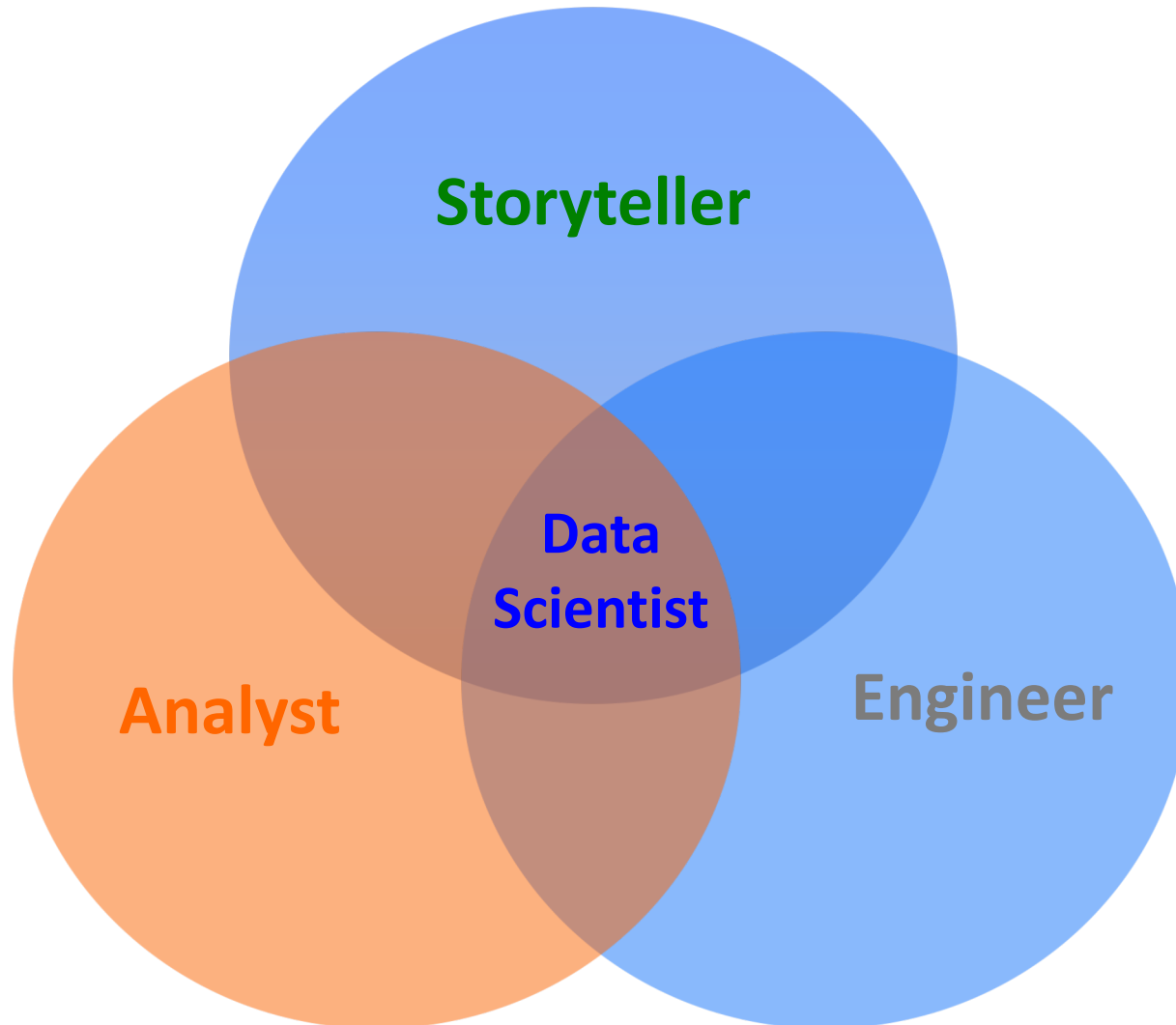


# Skillset



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

# Các vai trò của một nhà khoa học dữ liệu



# 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.



25 YEARS ANNIVERSARY  
**SOICT**

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

**Thank you  
for your  
attentions!**



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



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

