

25 YEARS ANNIVERSARY
SOICT

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



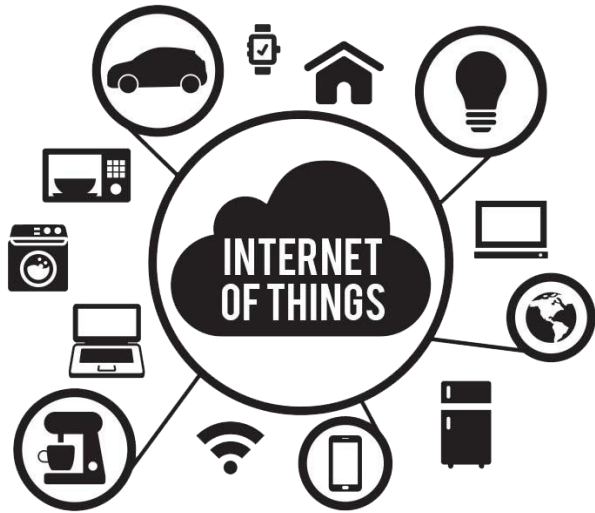
ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nhập môn Học máy và Khai phá dữ liệu (IT3190)

Nội dung môn học

- Lecture 1: Giới thiệu về Học máy và khai phá dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4+5: Phân cụm
- Lecture 6: Phân loại và Đánh giá hiệu năng
- Lecture 7: dựa trên láng giềng gần nhất (KNN)
- Lecture 8: Cây quyết định và Rừng ngẫu nhiên
- Lecture 9: Học dựa trên xác suất
- Lecture 10: Mạng nơron (Neural networks)
- Lecture 11: Máy vector hỗ trợ (SVM)
- Lecture 12: Khai phá tập mục thường xuyên và các luật kết hợp
- Lecture 13: Thảo luận ứng dụng trong thực tế

Nguồn dữ liệu



EACH DAY

50%
of active FB users log in

55 million
status updates are made



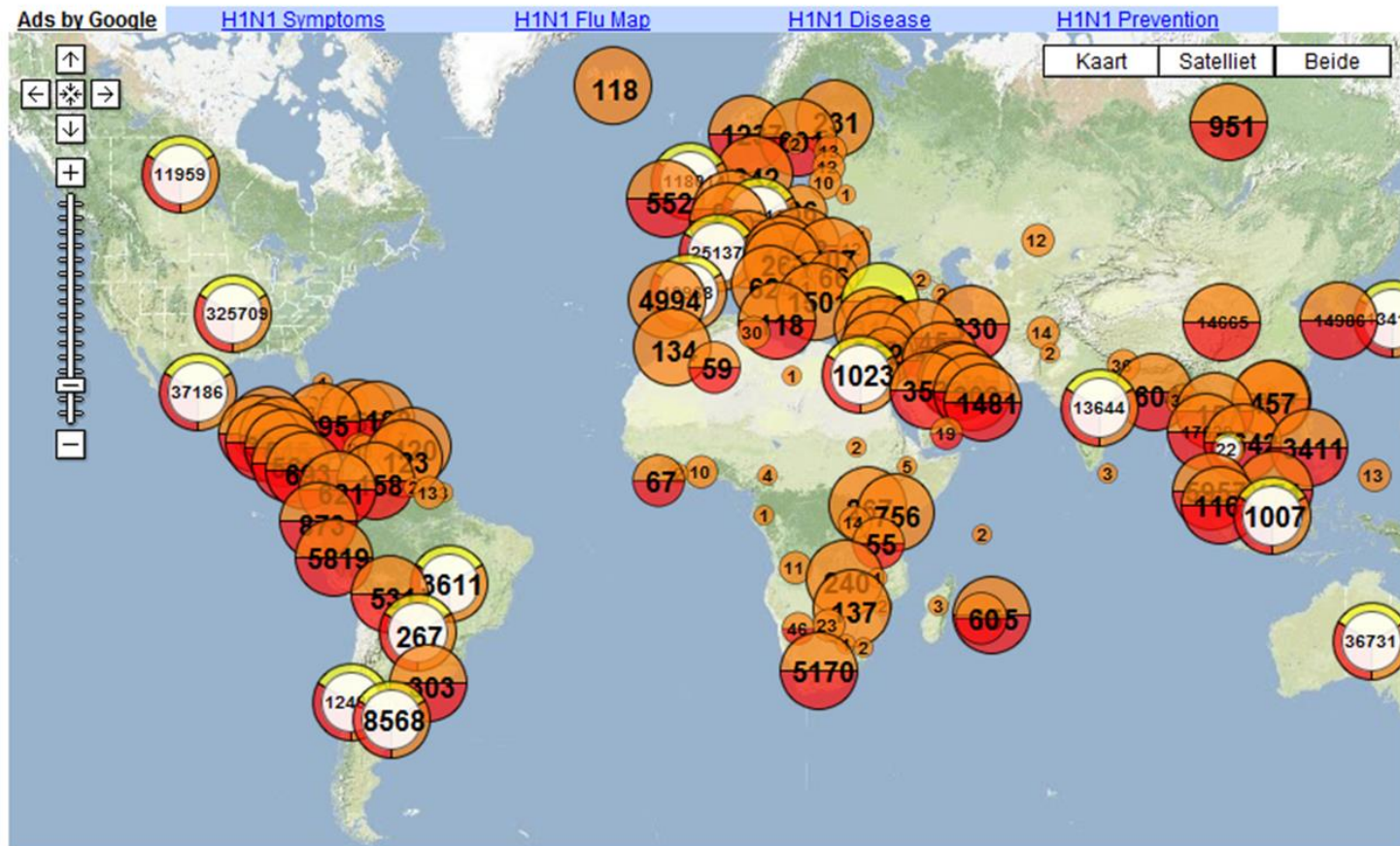
Pages have created
5.30 billion
of fans

35 million
update their status

Khai phá dữ liệu - Dự đoán

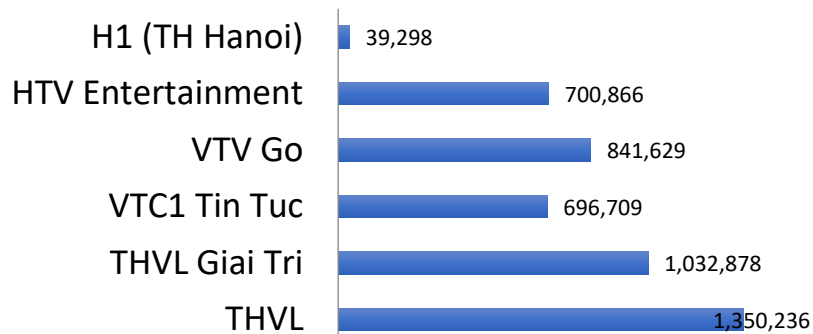
- **Google Flu Trends:** phát hiện các đợt bùng phát trước dữ liệu CDC hai tuần

[FluTracker map](#) data current as of 09:34 EDT 25 October

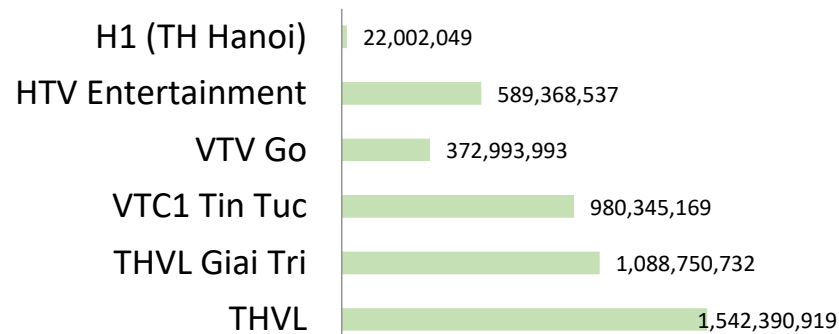


Khai phá dữ liệu - Khám phá

Subscribers in Youtube

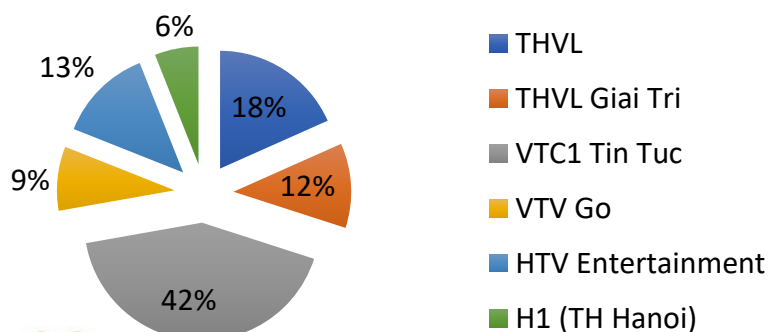


Views in Youtube

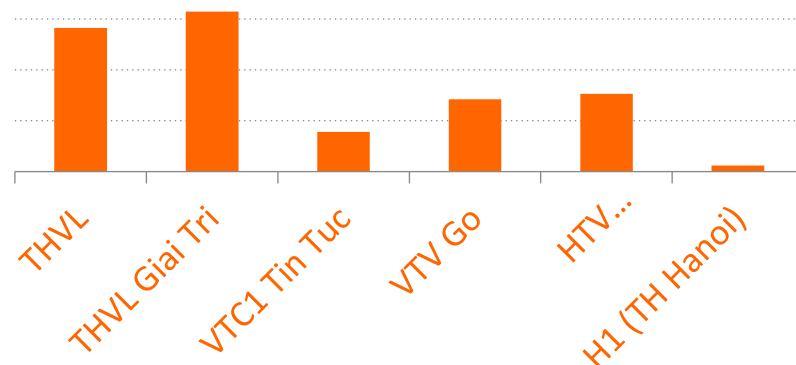


Kênh truyền hình hiệu quả?

Videos in Youtube



Attractiveness



Khai phá dữ liệu

- Dữ liệu giúp mọi thứ rõ ràng hơn

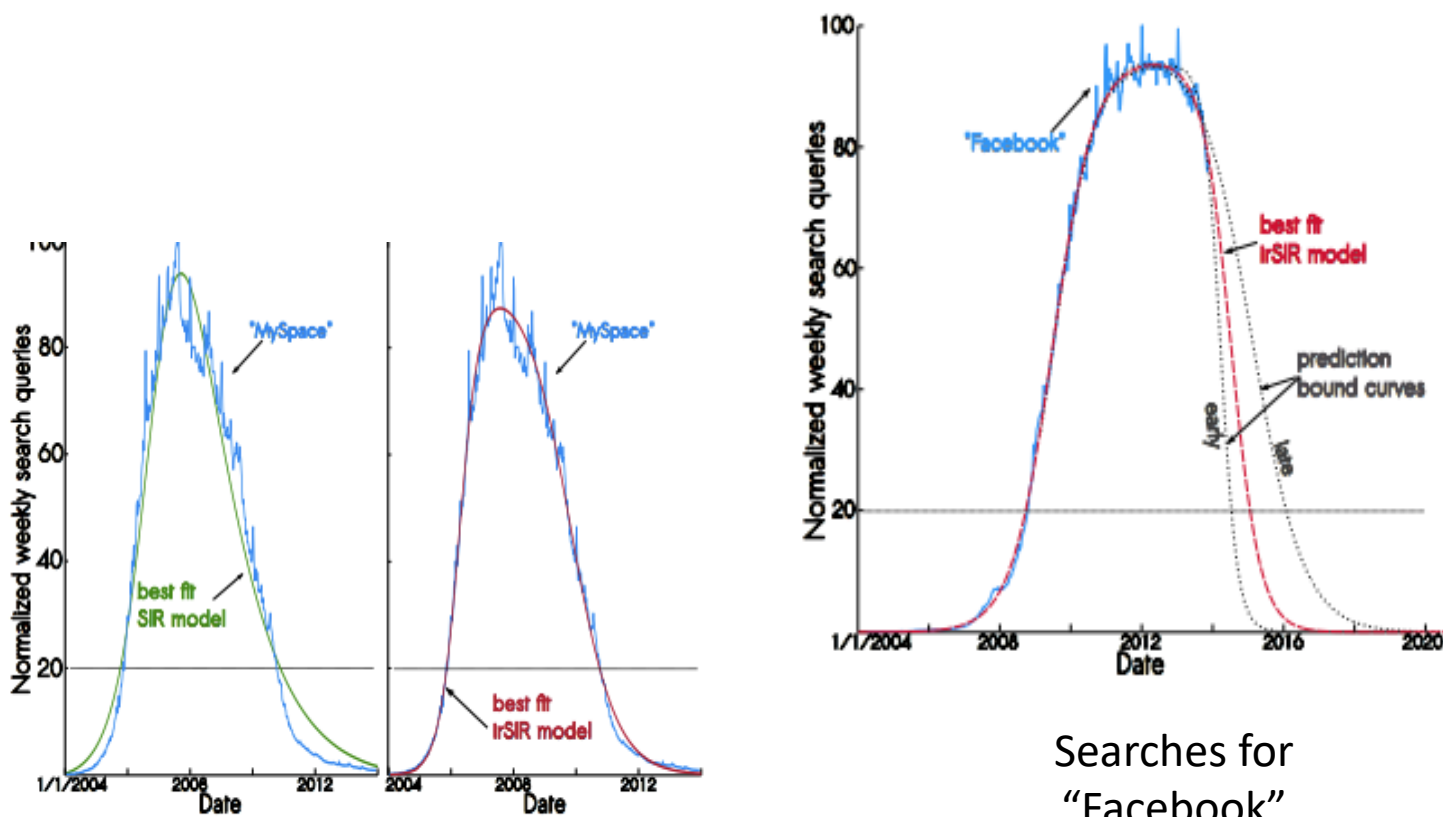


Figure 3: Data for search query “Myspace” with best fit (a) SIR and (b) IrSIR models overlaid. The search query data are normalized such that the maximum data point corresponds to a

Phát hiện tri thức và Khai phá dữ liệu

The **automatic extraction** of non-obvious, **hidden knowledge** from large volumes of data

(tự động trích rút những tri thức ẩn, không tường minh từ dữ liệu lớn)

Khái niệm dữ liệu

- Dữ liệu chỉ là dữ kiện thô (Long and Long, 1998)
- Dữ liệu... là các luồng dữ kiện thô biểu diễn các sự kiện... trước khi chúng được sắp xếp thành một dạng mà mọi người có thể hiểu và sử dụng (Laudon and Laudon, 1998)
- Dữ liệu bao gồm các dữ kiện (Hayes, 1992), các ký hiệu được ghi lại (McNurlin và Sprague, 1998)

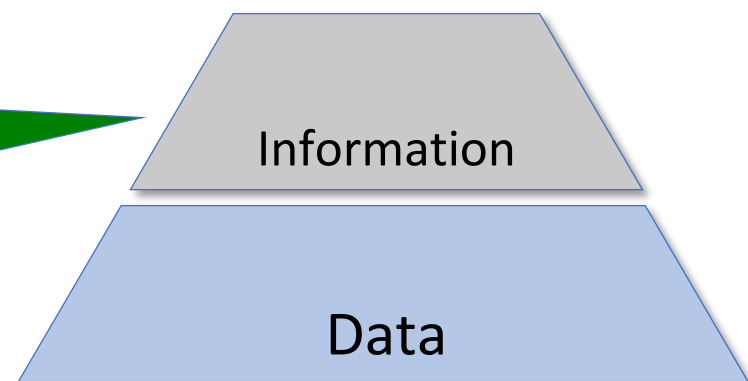
Dữ liệu là tín hiệu (signals) thu được do quan sát, đo đạc, thu thập... từ các đối tượng. Cụ thể, dữ liệu là giá trị (values) của các thuộc tính (features) của các đối tượng, được biểu diễn bằng dãy các bits, các con số hay ký hiệu...

Data

Khái niệm thông tin

- Dữ liệu đã được đưa về một dạng có ý nghĩa và hữu ích đối với con người (Laudon and Laudon, 1998)
- Dữ liệu đã được thu thập và xử lý thành một dạng có ý nghĩa. Đơn giản, thông tin là ý nghĩa mà chúng ta cung cấp cho các dữ kiện tích lũy (Long and Long, 1998)

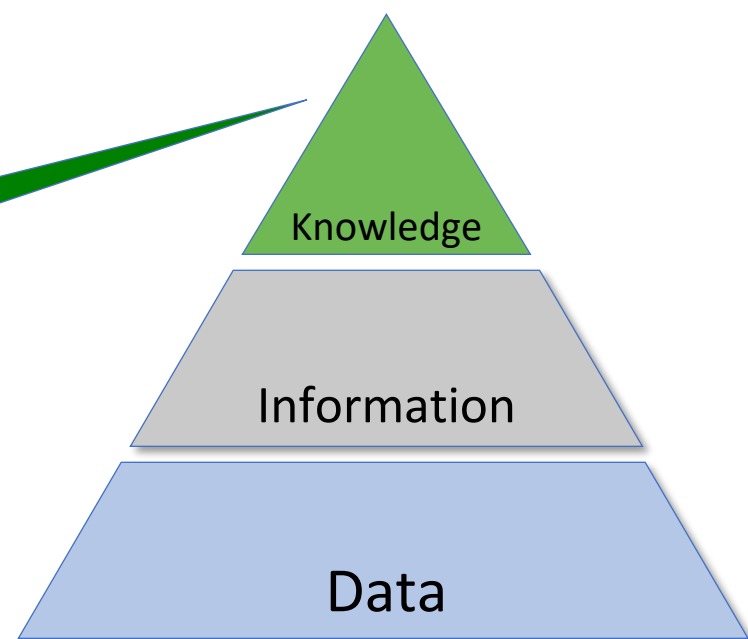
Thông tin là dữ liệu có ý nghĩa (data equipped with meaning), thu được khi xử lý dữ liệu để lọc bỏ đi các phần dư thừa, tìm ra phần cốt lõi đặc trưng cho dữ liệu.



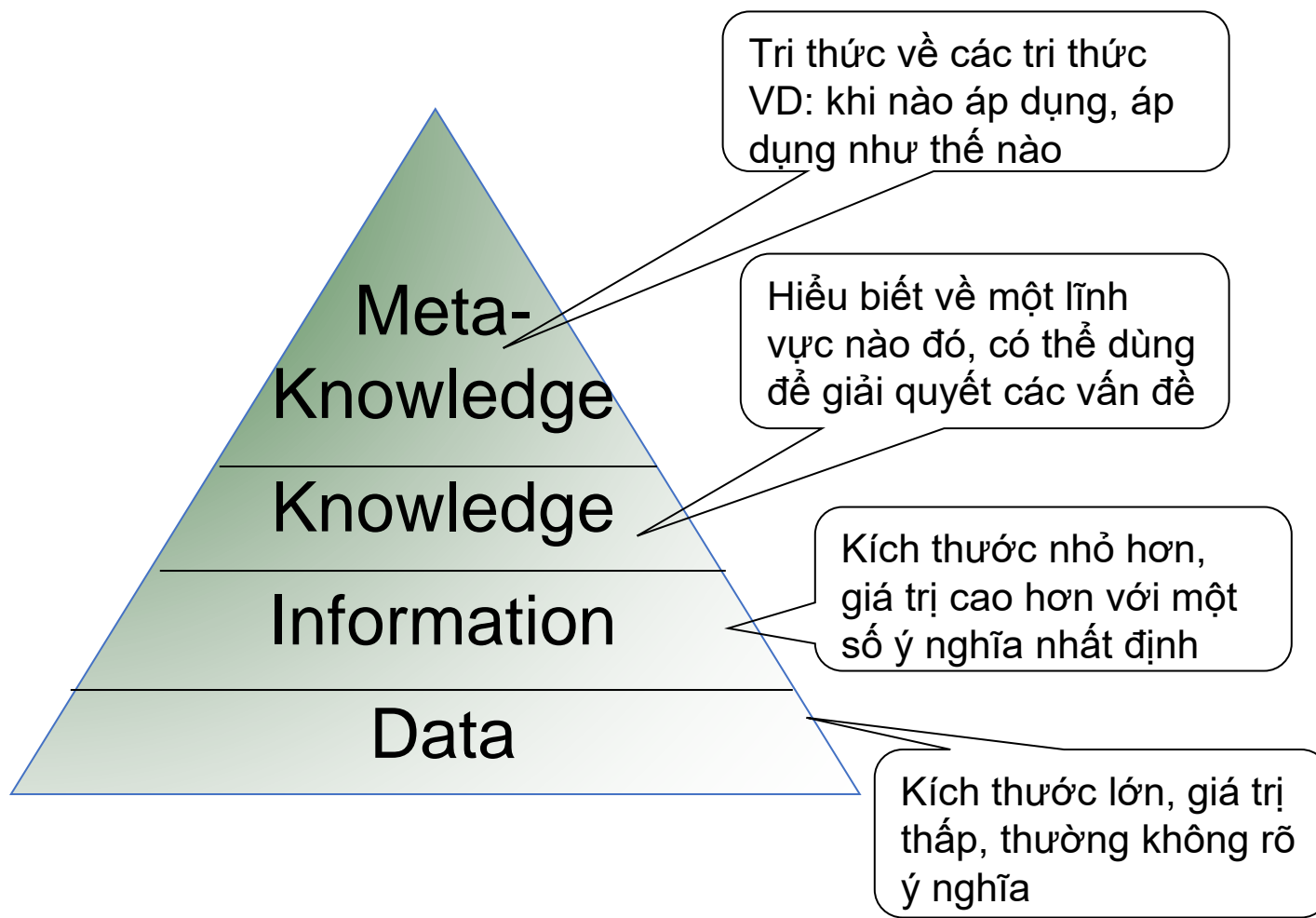
Khái niệm tri thức

- Kết quả của sự hiểu biết thông tin (Hayes, 1992)
- Kết quả của việc ngấm thông tin (Hayes, 1992), Thông tin thu thập về một lĩnh vực quan tâm (Senn, 1990)
- Thông tin có định hướng hoặc ý định, nó giúp hỗ trợ cho một quyết định hoặc một hành động (Zachman, 1987)

Tri thức là thông tin tích hợp, như quan hệ giữa các sự kiện, giữa các thông tin... thu được qua quá trình nhận thức, phát hiện hoặc học tập.



Dữ liệu – thông tin – tri thức



Ví dụ dữ liệu/thông tin/tri thức

- Dữ liệu
 - Trời nhiệt độ là $5^{\circ}C$
- Thông tin
 - Ngoài trời lạnh quá
- Tri thức
 - Nếu trời lạnh, bạn nên mặc áo ấm khi đi ra ngoài
- Giá trị cảm nhận của dữ liệu tăng lên khi nó được chuyển thành kiến thức.
- Kiến thức giúp đưa ra các quyết định hữu ích

KDD: tác vụ chính

- **Tiên đoán** (predictive task): đưa ra dự đoán về những sự kiện chưa biết trong tương lai và tìm ra lý do đằng sau những sự kiện đó

- Phân loại
- Hồi quy

Tri thức nào giúp ta phân biệt được tế bào ung thư?

- **Mô tả** (descriptive task): phân tích các đặc trưng của dữ liệu để thu được thông tin mới hoặc cho mục đích hữu ích nào đó

- Phân cụm
- Khai phá luật kết hợp

Thói quen nghe nhạc trực tuyến ra sao?



Tiên đoán: **Phân lớp**

- Đoán xem một quan sát x sẽ được cho vào lớp nào
 - “Những người đứng đầu Barcelona có vẻ hài lòng với điều này”
→ **Tích cực** hay **Tiêu cực**?
 - Những người thích nghe



+



-> Có phải người trẻ hay không

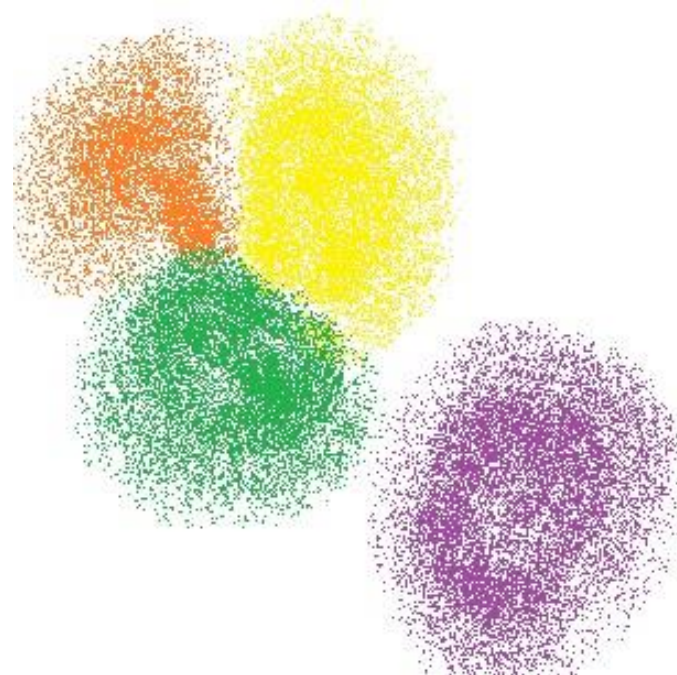
Tiên đoán: **Phát hiện ngoại lai**

- **Ngoại lai:** ngoại lai là một đối tượng mà có khác biệt rất lớn với các đối tượng thông thường, tưởng chừng như nó được sinh ra bởi một cơ chế hoàn toàn khác
 - Một thanh toán tín dụng bất thường
 - Tấn công mạng
 - Giá cổ phiếu bất thường
- Các điểm ngoại lai thường thú vị:
Nó vi phạm các cơ chế sinh dữ liệu thông thường
 - Khác với nhiều
- Nhiệm vụ của chúng ta là phát hiện các ngoại lai này
(outlier detection, anomaly detection)



Khai phá mô tả: Phân cụm

- **Cụm:** Nhóm dữ liệu có cùng đặc trưng nào đó
 - Một nhóm người yêu thích nhảy
- **Phân cụm (Clustering):** tìm tất cả các cụm trong một tập dữ liệu cho trước.



Khai phá mô tả: **Tóm tắt**

- Tìm kiếm mô tả ngắn gọn cho tập dữ liệu
 - VD: Tính toán trung bình và phương sai dữ liệu
 - VD: tổng hợp tin tức

Chúng ta hay viện dẫn câu chuyện thành công của học sinh Việt Nam trong các kì thi toán quốc tế để chứng minh cho năng lực học toán ở đẳng cấp thế giới của người Việt. Đây là do cách truyền thông của ta mà thôi. Đây không chỉ là một định kiến mà còn là một sự huyền hoặc nguy hiểm.

Người Việt giỏi toán: Góc nhìn 'thật' từ người trong cuộc

10/03/2015 01:00 GMT+7

Chúng ta hay viện dẫn câu chuyện thành công của học sinh Việt Nam trong các kì thi toán quốc tế để chứng minh cho năng lực học toán ở đẳng cấp thế giới của người Việt. Đây là do cách truyền thông của ta mà thôi.

Người Việt giỏi toán: có thật vậy không?

Đặt vấn đề có chắc người Việt giỏi toán hay không chắc chắn sẽ gây nhiều tranh cãi vì có thể nó sẽ đi ngược lại quan điểm của đa số chúng ta với một định rằng: người Việt giỏi Toán hay ít nhất là có năng lực và tiềm năng học Toán?

Theo tôi đây không chỉ là một định kiến mà còn là một sự huyền hoặc nguy hiểm.

Chúng ta đều biết trong bảng xếp hạng về các đóng góp của các nước trên thế giới vào khoa học và công nghệ thì Việt Nam luôn xếp ở nhóm cuối.

Trong các cuộc tiếp xúc với các nhà khoa học hàng đầu thế giới chúng tôi đã không ngần ngại hỏi họ nhận định thế nào về vị trí của Việt Nam trên bản đồ khoa học và toán học của thế giới và đây là đánh giá của họ:

Về khoa học: chúng ta là số 0 trên trình.

Về Toán học: chúng ta là một chấm rất nhỏ.

Chúng tôi không hề ngạc nhiên về đánh giá này. Ở đây chúng tôi thậm chí còn đưa vấn đề đi xa hơn không chỉ với việc đề cập người Việt không giỏi Toán mà còn nói tới việc liệu có phải chúng ta thực sự có đam mê dành cho Toán học hay không?



Cho đến nay, GS Ngô Bảo Châu là người Việt duy nhất theo đuổi nghiệp Toán học và đạt được đỉnh cao. Ảnh AP

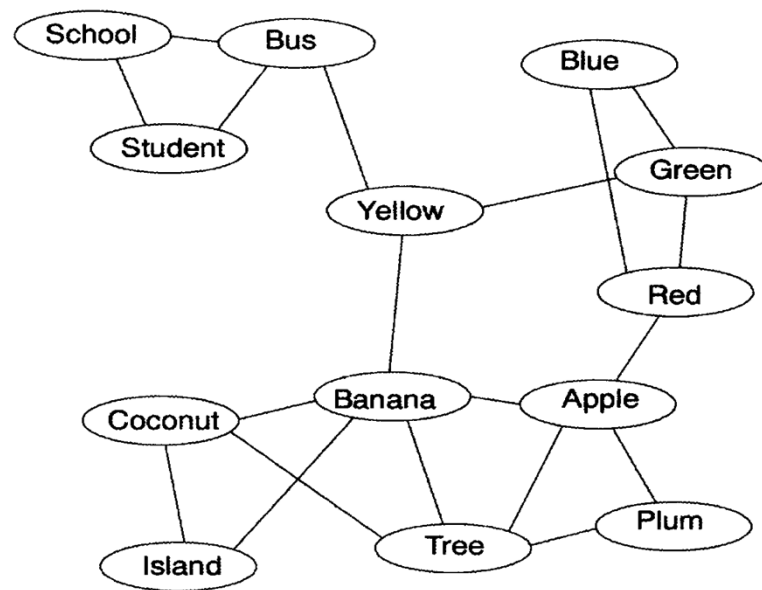
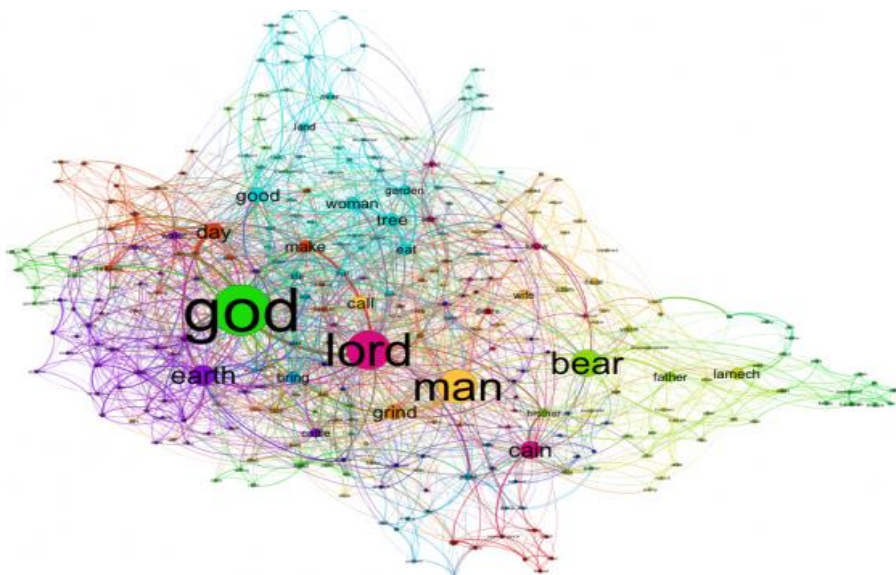
Câu chuyện ở những kỳ thi Toán quốc tế

Chúng ta hay viện dẫn câu chuyện thành công của học sinh Việt Nam trong các kì thi toán quốc tế để chứng minh cho năng lực học toán ở đẳng cấp thế giới của người Việt. Đây là do cách truyền thông của ta mà thôi. Sự thật là:

1. Kỳ thi toán quốc tế IMO chỉ là một cuộc chơi vui vẻ theo đúng nghĩa của nó. Các nước cử đội tuyển tham dự kì thi này theo tiêu chí vui là chính và hoàn toàn không coi đây là sứ mạng mang về vinh dự quốc gia hay giúp nước đó khẳng định vị thế của họ trên bản đồ toán học thế giới. Sẽ thật là sai lầm nếu qua một cái name dành cho học sinh như vậy mà khẳng định VNM Nam là một nước

Khai phá mô tả: **Mô hình phụ thuộc**

- Tìm kiếm mô hình mà nó mô tả những phụ thuộc có ý nghĩa giữa các biến
 - *Mức cấu trúc*: Biến cục bộ phụ thuộc vào nhau như thế nào
 - *Mức định lượng*: độ mạnh của các phụ thuộc vào một số.



KDD: Kiểu dữ liệu

- **Supervised** (có giám sát, có nhãn):
 - Mỗi quan sát x trong tập huấn luyện sẽ có một đầu ra (nhãn)
 - Mục đích là để dự đoán kết quả đầu ra cho một quan sát mới

(x = “Những người đứng đầu Barcelona có vẻ hài lòng với điều này”, y = Positive)



Bát,
Thìa,
ramen

- **Unsupervised** (không giám sát, không nhãn): chúng ta không thể quan sát bất kỳ đầu ra y nào
 - VD: dòng tweets -> xu hướng hiện tại?
- Một số tác vụ có thể có meta-data như tag, likes, links, views,... Những meta-data đó có thể giúp khám phá thêm kiến thức mới.

KDD: Kiểu dữ liệu

Có cấu trúc

	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247	37.37	6.02	3.46	72
8	Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9	Uruguay	Americas	3395	22.05	18.59	2.07	77

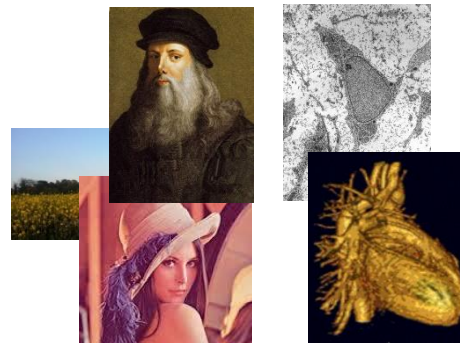
Phi cấu trúc

```
{
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",
  "title": "[Updating] Câu chuyện xuyên mưa về :",
  "url": "http://techtalk.vn/updating-cau-chuyen",
  "labels": "techtalk/Cong nghe",
  "content": "Vào chiều tối ngày 09/12/2016 vừa",
  "image_url": "",
  "date": "2016-12-10T03:51:10Z"
}
```

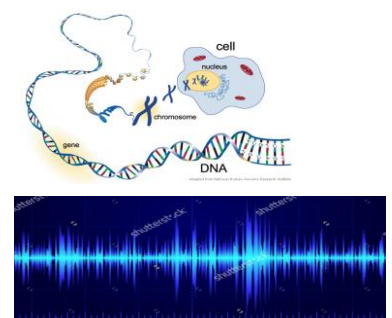
texts in websites, emails, articles, tweets



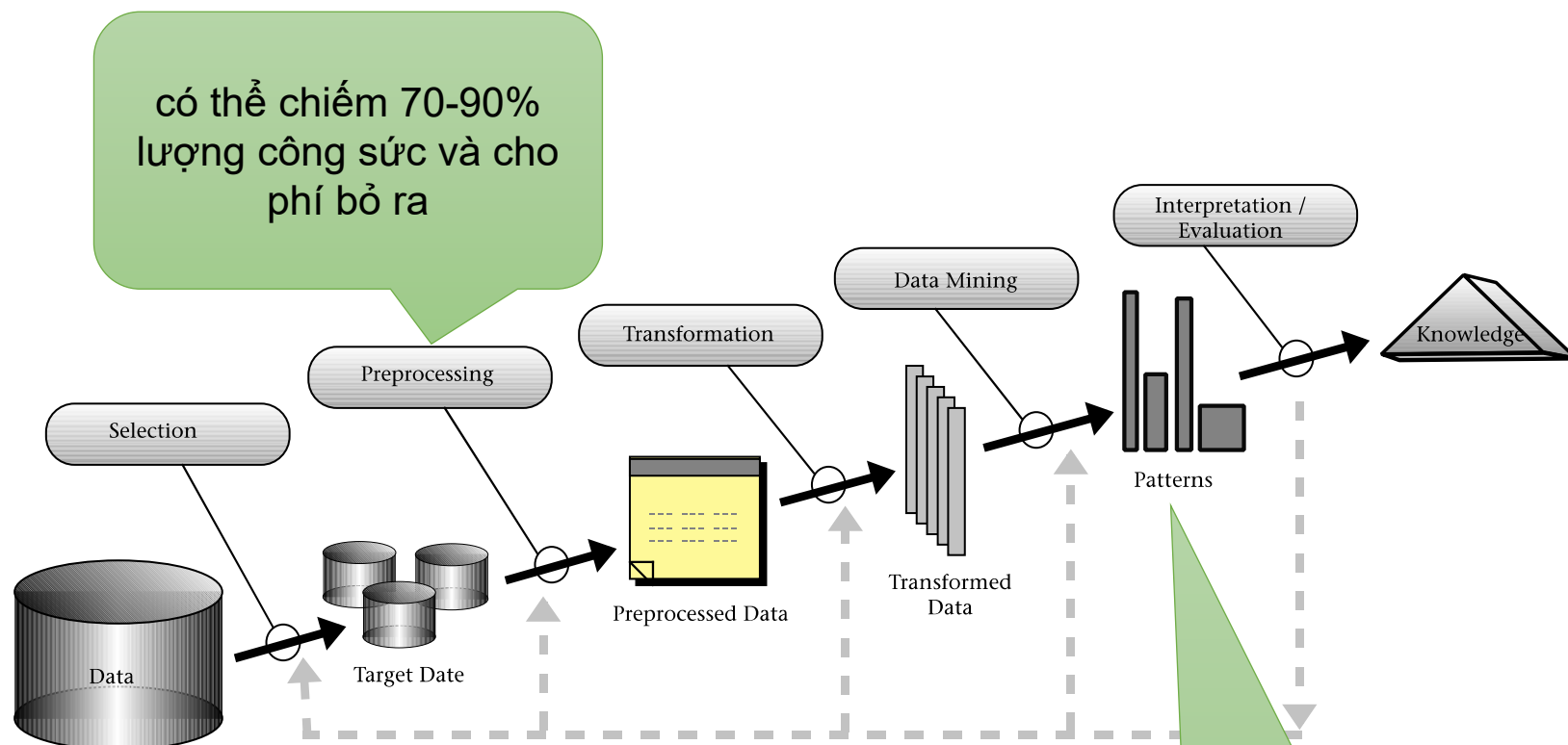
2D/3D images, videos + meta



spectrograms, DNAs, ...

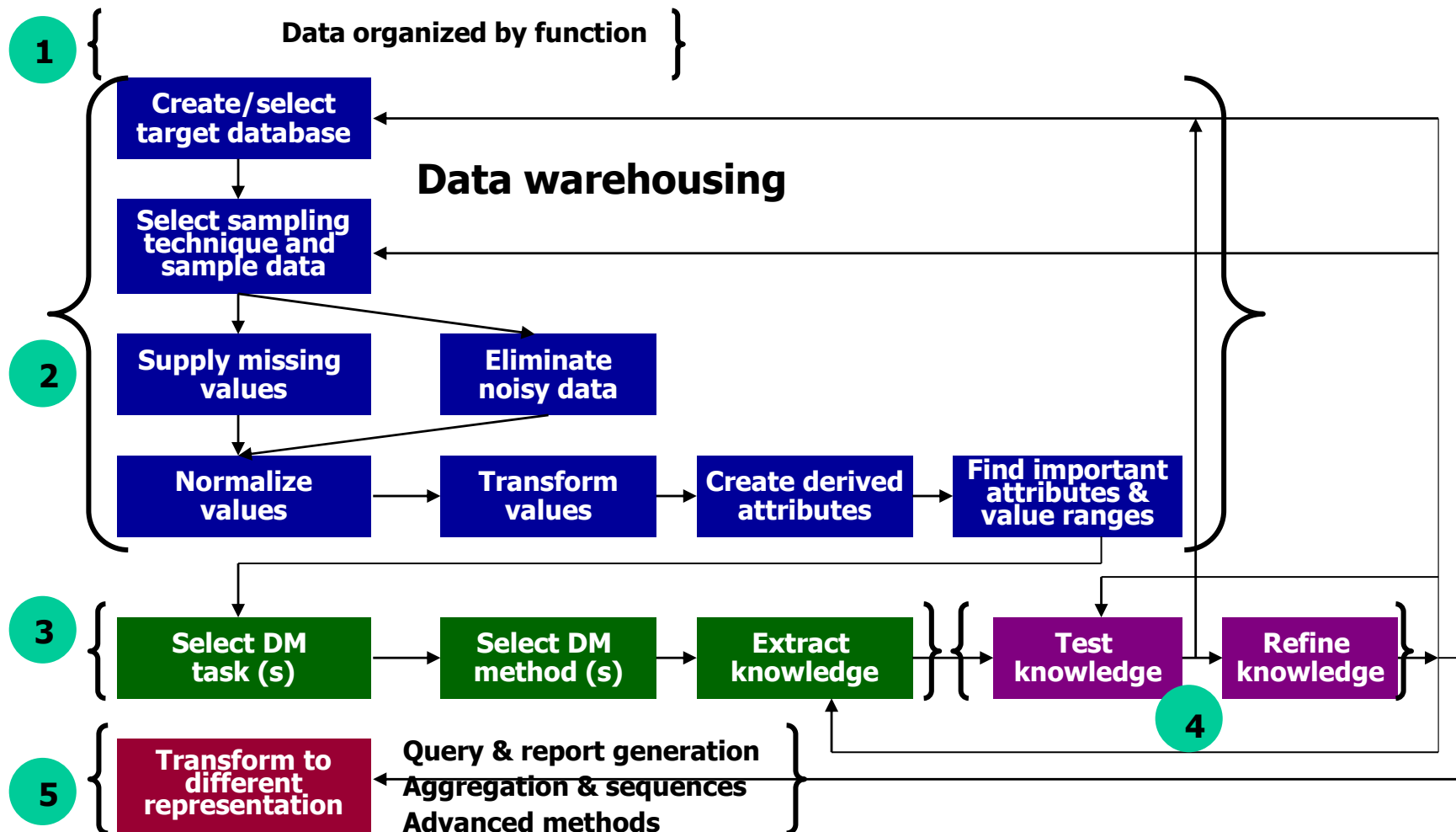


KDD: Phương pháp

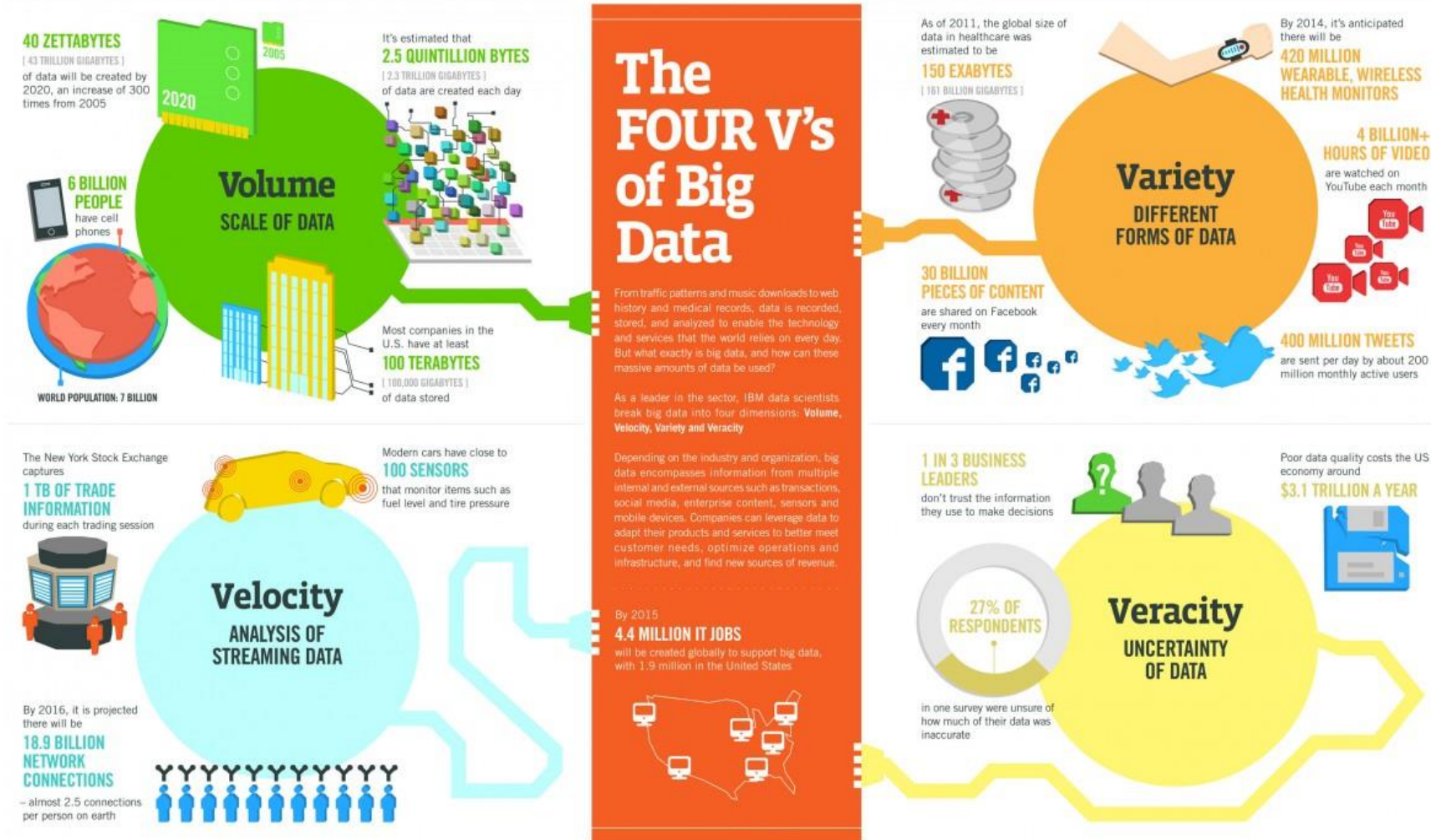


(Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

KDD: Phương pháp



KDD: Thách thức

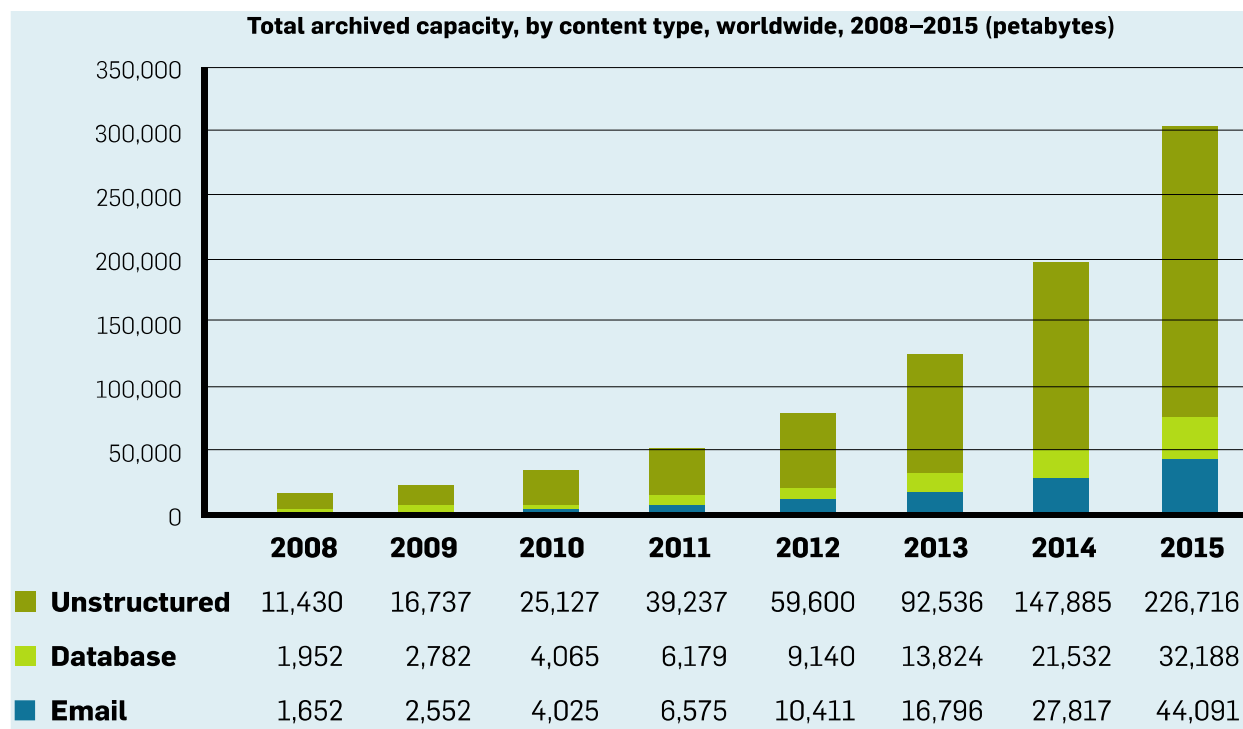


Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTec, QAS

IBM

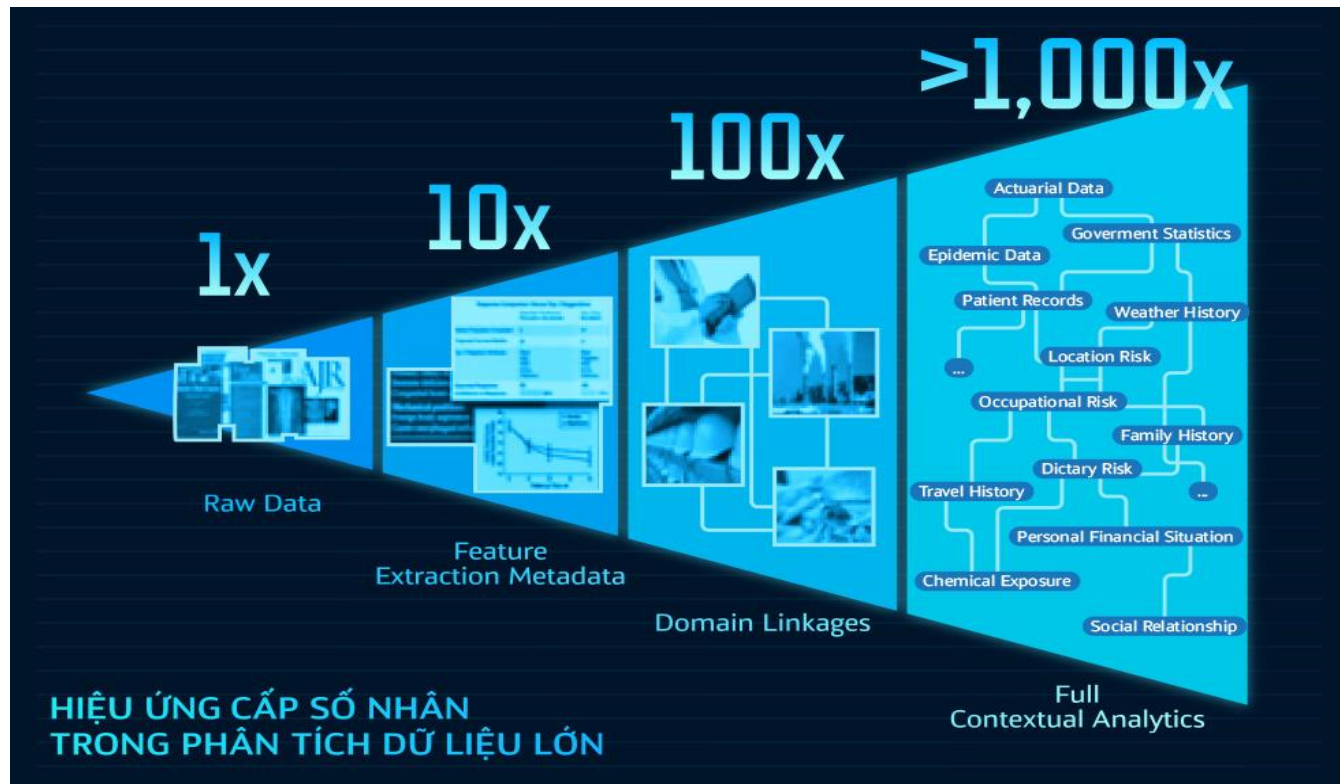
Thách thức: phi cấu trúc

- Các dữ liệu phi cấu trúc phát triển và gia tăng rất nhanh
 - Text, ảnh, tags, links, likes, ...



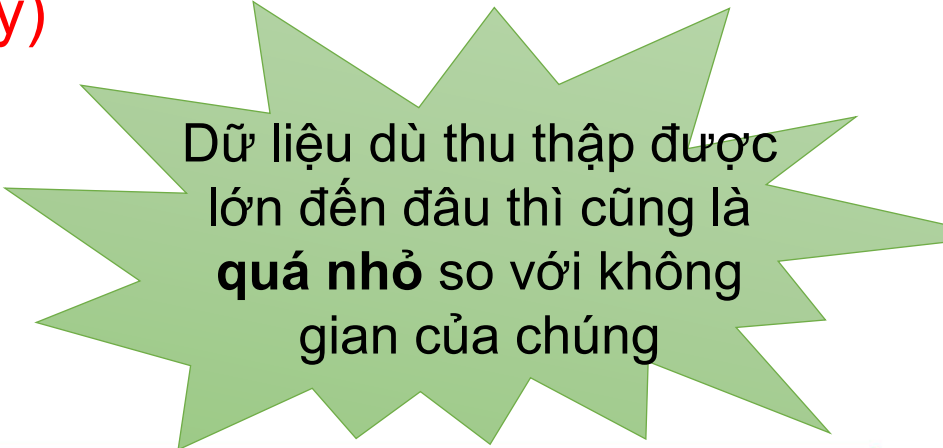
Thách thức: tương tác ẩn

- Những mối tương tác ẩn chứa bên trong dữ liệu có thể rất lớn



Thách thức: số chiều quá lớn

- Các bài toán thực tế thường có số chiều rất lớn
 - Xe đạp chạy: 2 chiều (một con đường)
 - Chúng ta đang sống: 4 chiều
 - Nhưng một hình ảnh 1024×1024 : ~ 1 triệu chiều
 - Bộ sưu tập văn bản: hàng triệu chiều
 - Hệ thống của người đề xuất: hàng tỷ chiều (mặt hàng/sản phẩm)
- Lời nguyền của số chiều không gian
(The **curse of dimensionality**)



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

Tài liệu tham khảo

- 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.
- Vasant Dhar. Data Science and Prediction. *Communication of the ACM*, vol 56 (12), pp 64-73, 2013.
- Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17, no. 3 (1996).
- R. Hayes. The Measurement of Information. In Vakkari, P. and Cronin, B. (editors): *Conceptions of Library and Information Science*, pp. 97–108. Taylor Graham, 1992.
- K. C. Laudon and J. P. Laudon. *Management Information Systems: New Approaches to Organisation and Technology* (5th edition). Prentice-Hall, 1998.
- L. Long and N. Long. *Computers* (5th edition). Prentice-Hall, 1998.
- B. McNurlin and R. H. Sprague. *Information Systems Management in Practice* (4th edition). Prentice-Hall, 1998.
- J. Zachman. A Framework for Information Systems Architecture. *IBM Systems Journal*, 26(3): 276–292, 1987.



25 YEARS ANNIVERSARY
SOICT

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

**Thank
you for
your
attentions
!**

