



# Introduction to **Machine Learning and Data Mining** (Học máy và Khai phá dữ liệu)

**Khoat Than**

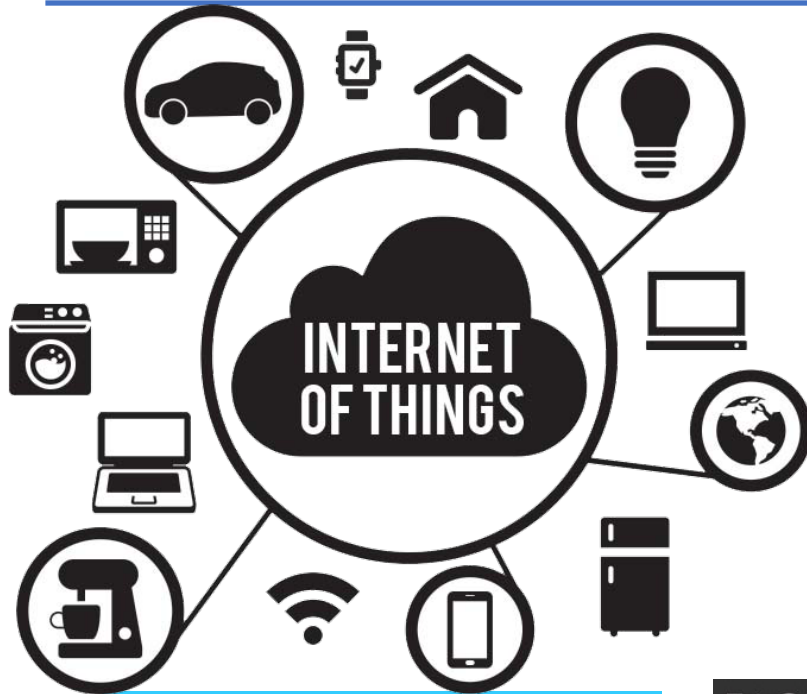
School of Information and Communication Technology  
Hanoi University of Science and Technology

# Contents

---

- Introduction to Machine Learning & Data Mining
- Unsupervised learning
- Supervised learning
- Probabilistic modeling
- **Data mining**
- Practical advice

# Various sources of data



twitter

facebook®

EACH DAY

50%  
of active FB users log in

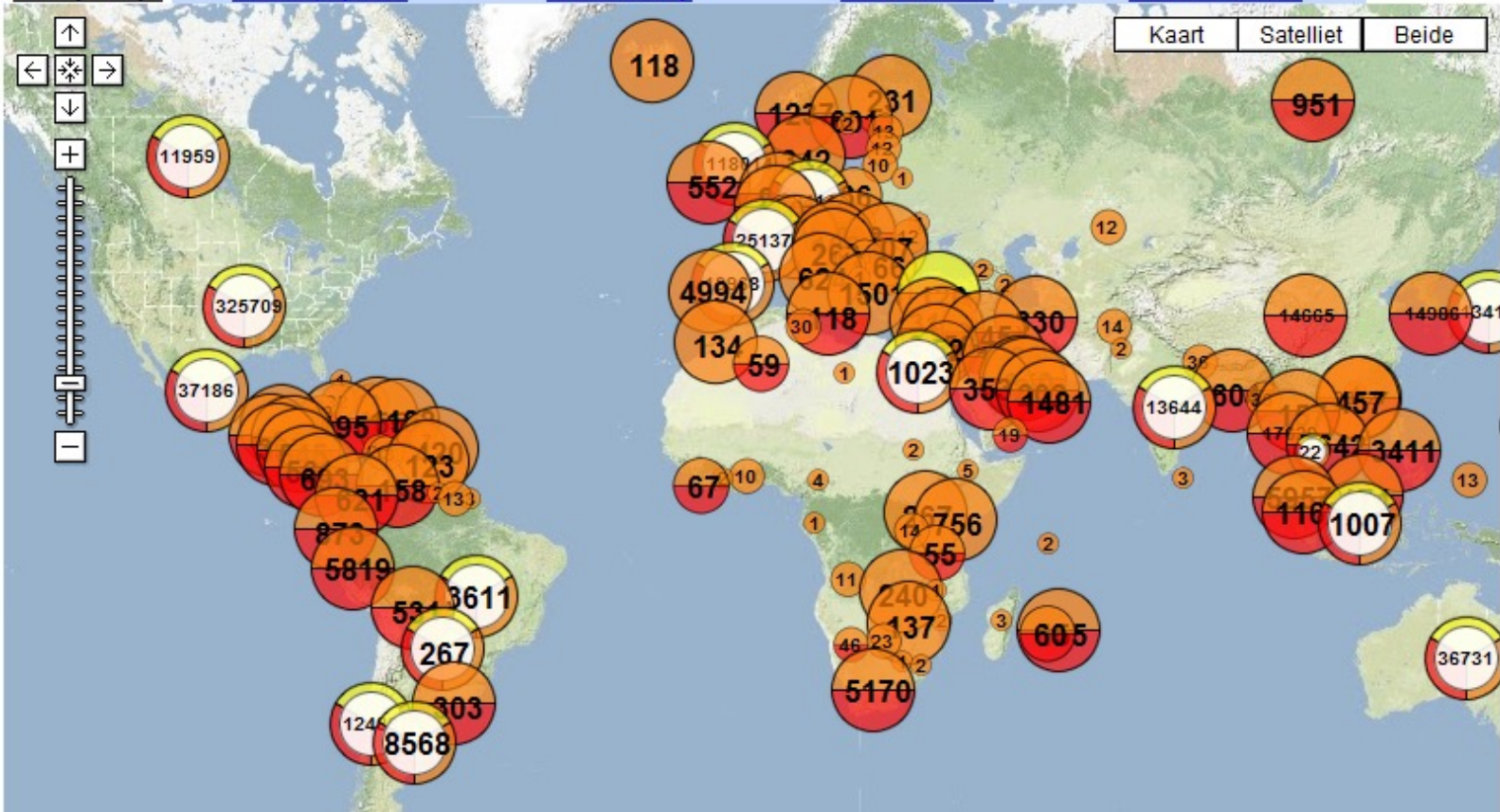
55 million  
status updates are made



Pages have created  
5.30 billion  
of fans

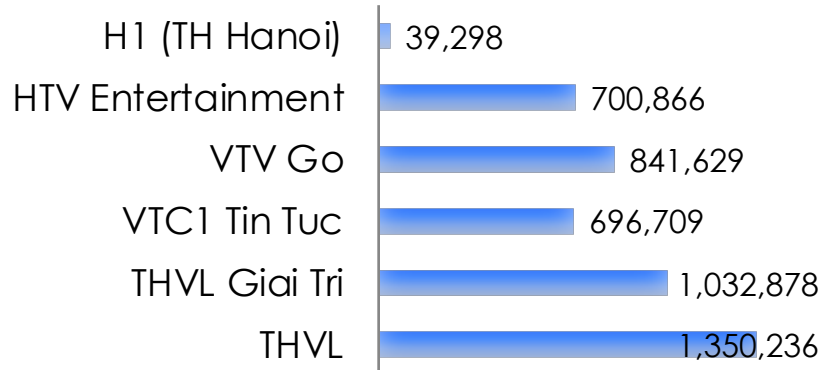
35 million  
update their status

[Ads by Google](#)   [H1N1 Symptoms](#)   [H1N1 Flu Map](#)   [H1N1 Disease](#)   [H1N1 Prevention](#)

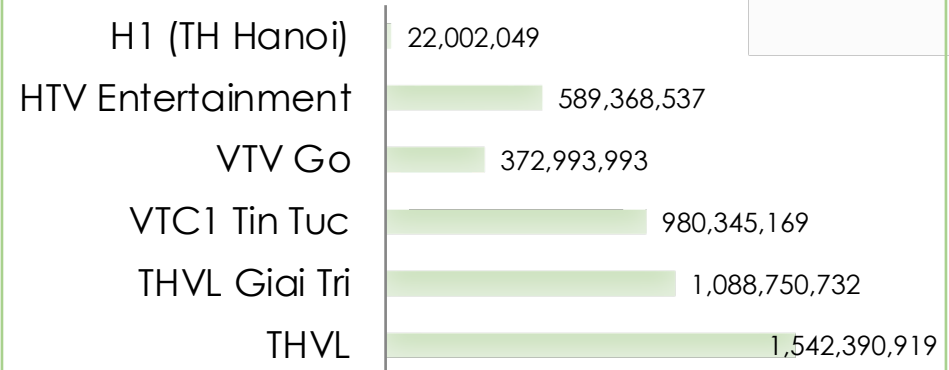


# Why do we need mining? Exploration

## Subscribers in Youtube



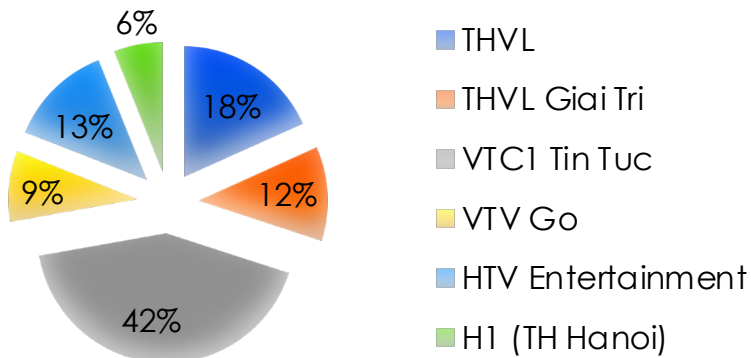
## Views in Youtube



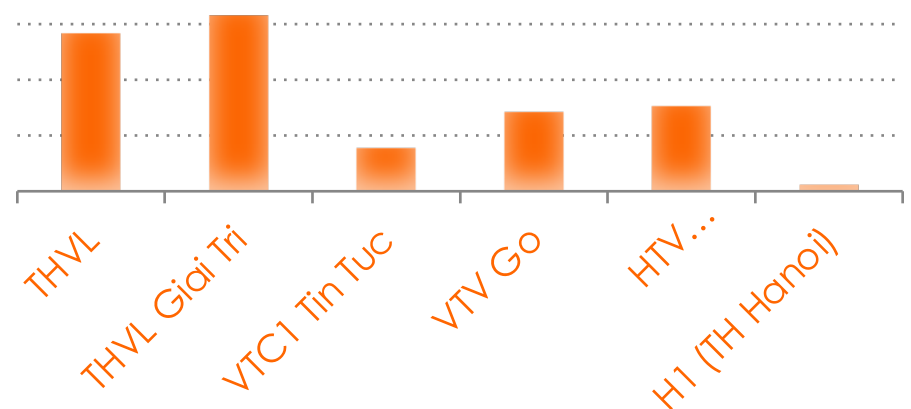
## Effective TV channels?

(July 4, 2017)

## Videos in Youtube

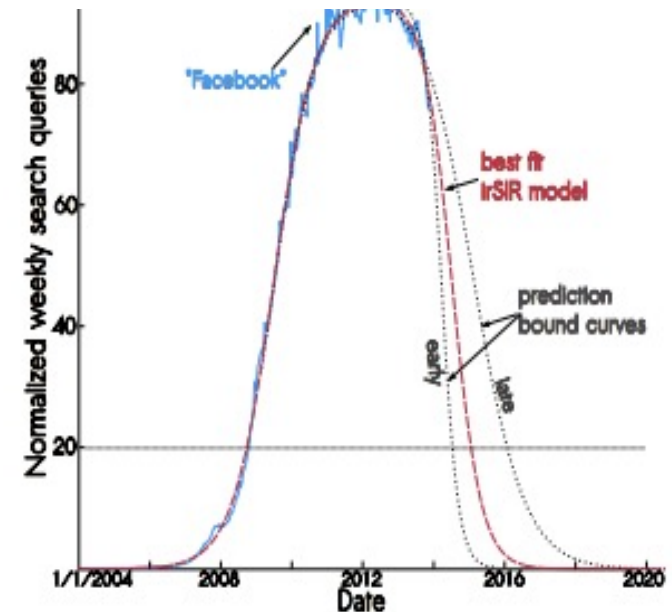
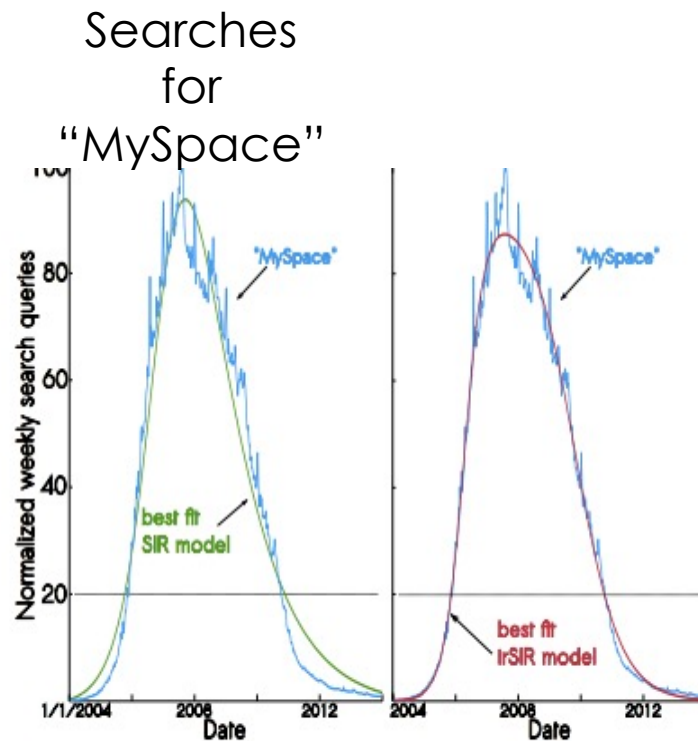


## Attractiveness



# Why do we need mining?

Data make everything clearer



Searches for "Facebook"

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

(John Canny, UC Berkeley)

# Knowledge discovery and data mining

---

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)



# Definition of Data

---

- *Data* are just raw facts (Long and Long, 1998)
- *Data* . . . are streams of raw facts representing events . . . before they have been arranged into a form that people can understand and use (Laudon and Laudon, 1998)
- *Data* is comprised of facts (Hayes, 1992)  
Recorded symbols (McNurlin and 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...



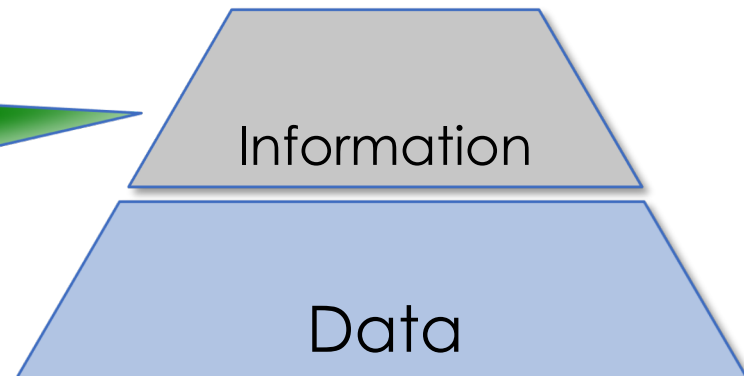
Data



# Definition of Information

- Data that have been shaped into a form that is meaningful and useful to human beings (Laudon and Laudon, 1998)
- Data that have been collected and processed into a meaningful form. Simply, information is the meaning we give to accumulated facts (Long and Long, 1998)
- The property of data which represents and measures effects of processing them (Hayes, 1992)

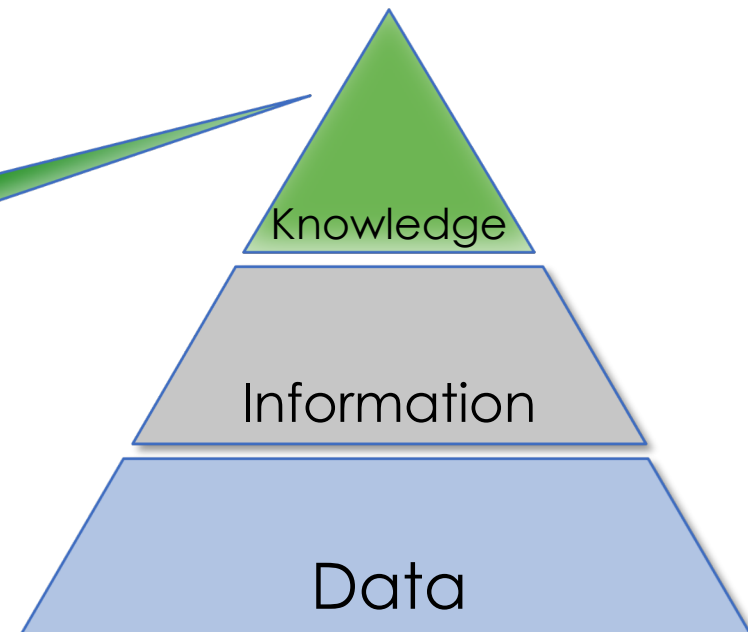
Thông tin là dữ liệu có ý nghĩa (data equipped with meaning), thu được khi xử lý dữ liệu để lọc bỏ đi các phần dư thừa, tìm ra phần cốt lõi đặc trưng cho dữ liệu.



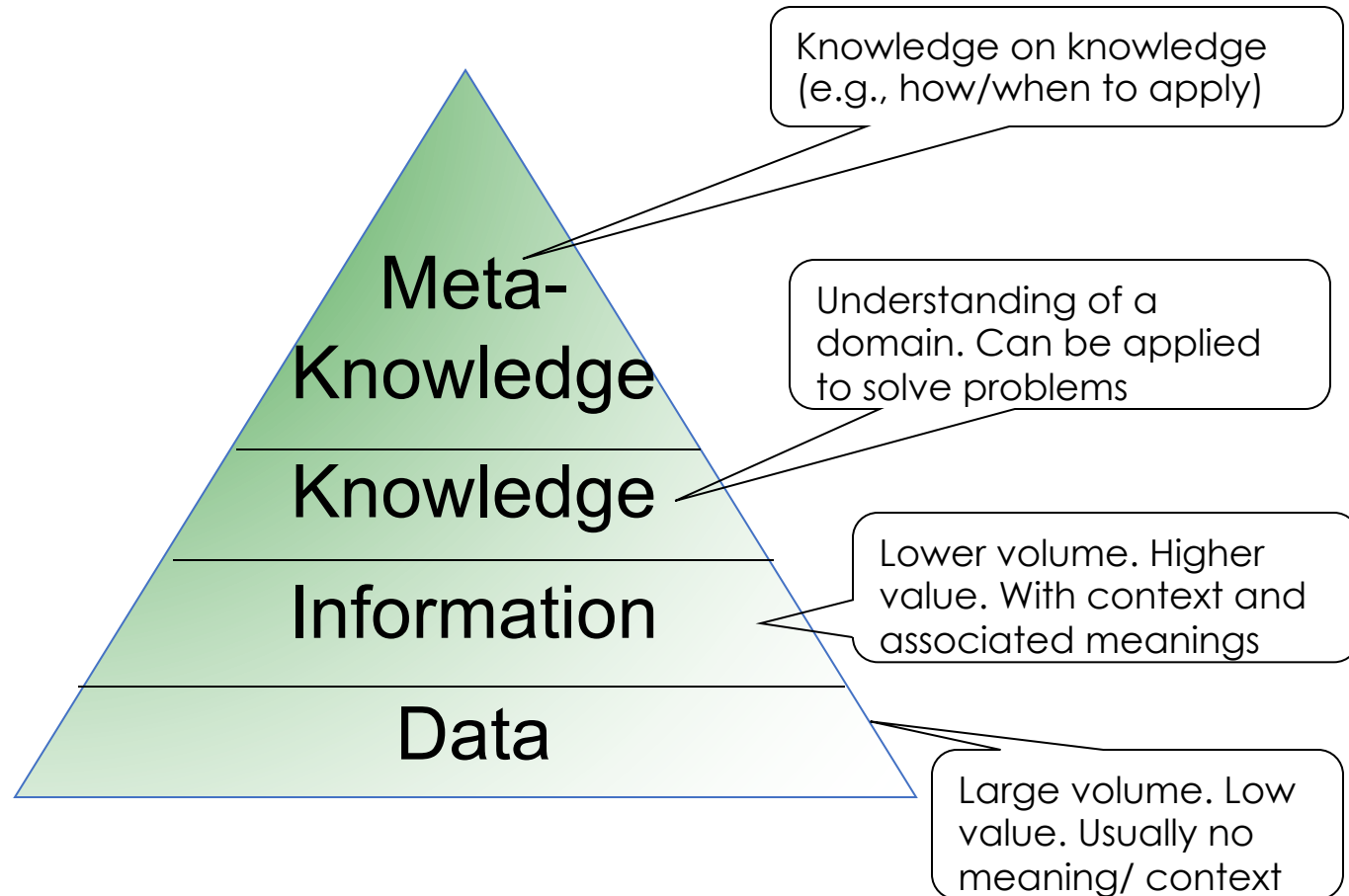
# Definition of Knowledge

- The result of the understanding of information (Hayes, 1992)
- The result of internalizing information (Hayes, 1992)  
Collected information about an area of concern (Senn, 1990)
- Information with direction or intent – it facilitates a decision or an action (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.



# Data-Information-Knowledge



# Example of Data/Information/Knowledge

---

- Data
  - Nhiệt độ ngoài trời là 5°C
- Information
  - Ngoài trời lạnh quá
- Knowledge
  - Nếu trời lạnh thì bạn nên mặc áo ấm khi đi ra ngoài
- The perceived value of data increases as it is transferred into knowledge.
- Knowledge enables useful decisions to be made.

# KDD: main tasks

- **Predictive task:** make predictions about unknown future events and disclose the reasons behind them.  
(tạo các phán đoán về những sự kiện tương lai và vạch ra những lý do đứng sau những sự kiện đó)

- Classification
- Regression

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

- **Descriptive task:** characterize the properties of the data to gain information, or for other useful purposes.  
(phân tích các đặc trưng của dữ liệu để thu được thông tin mới hoặc cho mục đích hữu ích nào đó)

- Clustering
- Association

Thói quen nghe nhạc trực tuyến ra sao?



# Predictive mining: classification

---

- Predict what label should be assigned to observation  $x$ ?  
what class should  $x$  belong to?  
(cần đoán xem  $x$  thuộc lớp nào)
- “Những người đứng đầu Barcelona có vẻ hài lòng với điều này” →  
Positive or negative?

# Predictive mining: outlier detection

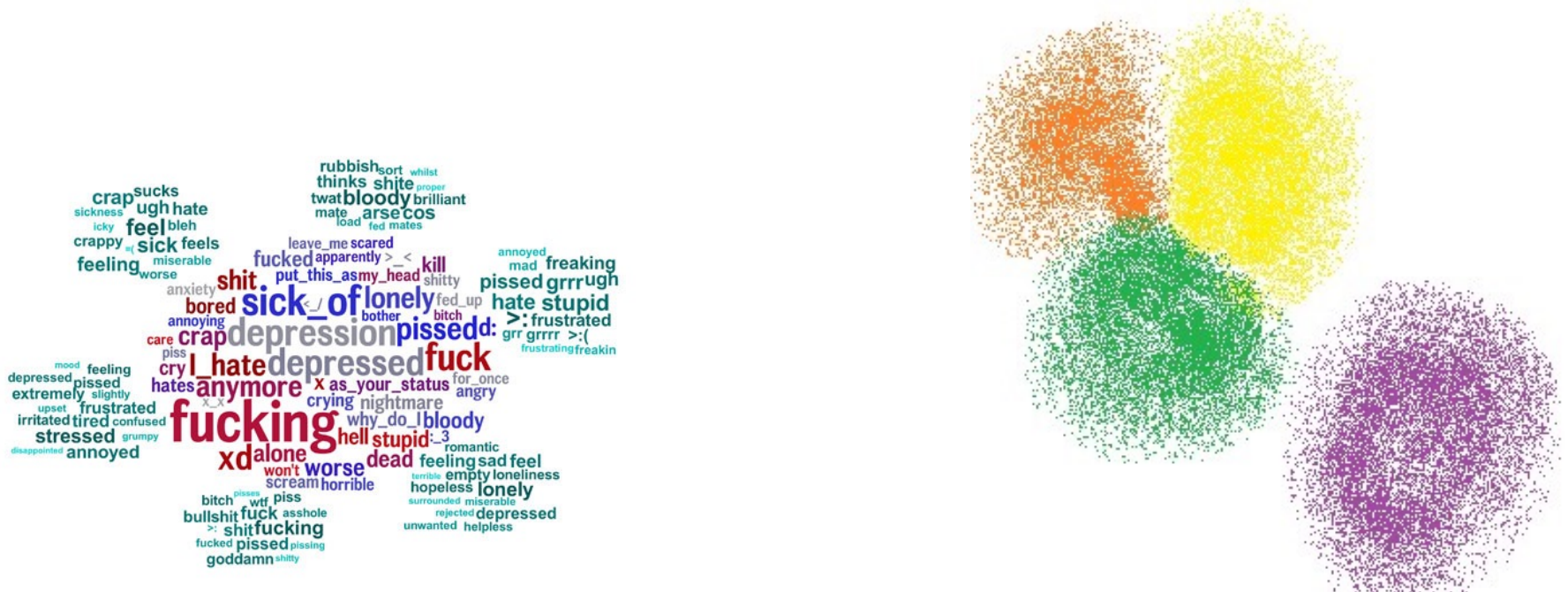
- **Outlier (ngoại lai)**: A data object that **deviates significantly** from the normal objects as if it were generated by a different mechanism  
(ngoại lai là một đối tượng mà có khác biệt rất lớn với các đối tượng thông thường, tưởng chừng như nó được sinh ra bởi một cơ chế hoàn toàn khác)
  - Unusual credit card purchase,
  - Network attacks,
  - Unusual stock price, ...
- Outliers are interesting:  
It violates the mechanism that generates the normal data
  - Different with noises
- Our task is to detect those? (outlier detection, anomaly detection)





# Descriptive mining: clustering

- *Cluster*: a group of data instances that have the same properties (một nhóm dữ liệu mà có cùng đặc trưng nào đó)
  - A group of users who like dancing
- **Clustering**: find all the clusters in a given data set.



# Descriptive mining: summarization

- Finding a compact description for a subset of data  
(tìm kiếm một mô tả ngắn gọn cho tập dữ liệu)
- E.g, compute the mean and deviation of an attribute
- E.g, news summarization

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

**Tuoihoctieu.vn** 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 tròn trĩnh.

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

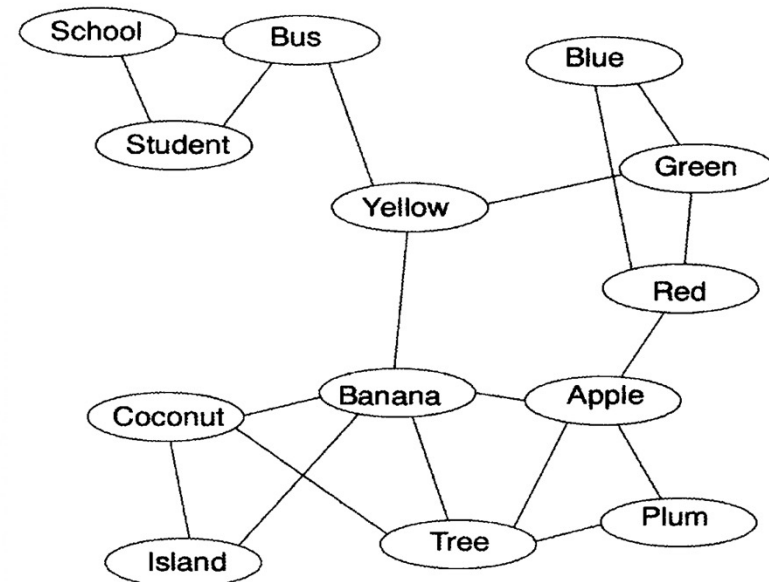
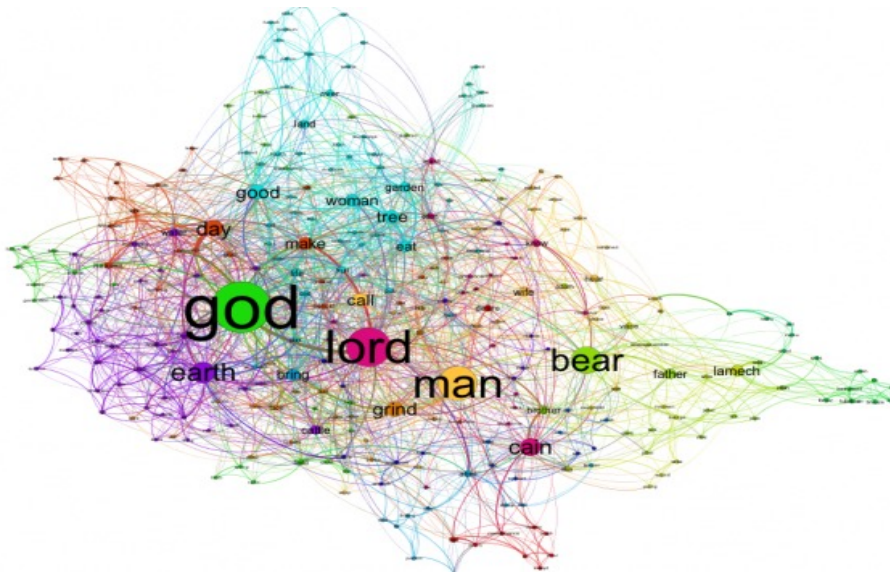
### Câu chuyện ở những kỳ thi Toán quốc tế

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 game dành cho học sinh như vậy mà khẳng định Việt Nam là một cường

# Descriptive mining: dependency modeling

- Find a model that describes significant dependencies between variables. (tìm kiếm mô hình mà nó mô tả những phụ thuộc có ý nghĩa giữa các biến)
- *Structural level*: which variables are locally dependent on each other.
- *Quantitative level*: the strength of the dependencies in term of a number.



# KDD: data type

- **Supervised** (có giám sát, có nhãn):

- Each observation in the training set will have an output (label)
- The aim is to predict what output for a new observation

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



Bow,  
Spoon,  
ramen

- **Unsupervised** (không giám sát, không nhãn): we could not observe any output of the training data.
  - Ex: data = tweets → what is the current trend?
- Some tasks may have meta-data such as tags, likes, links, views,... Those meta-data can help further discovery of new knowledge.



# KDD: data type

## Structured – relational (table-like)

	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

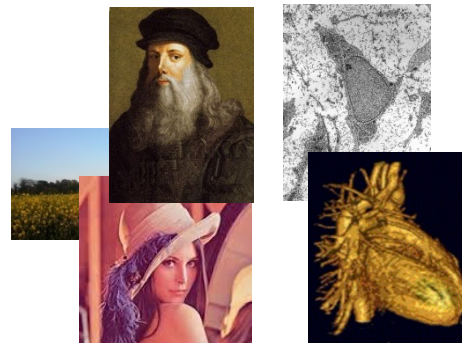
## Un-structured

```
{
  "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"
}
```

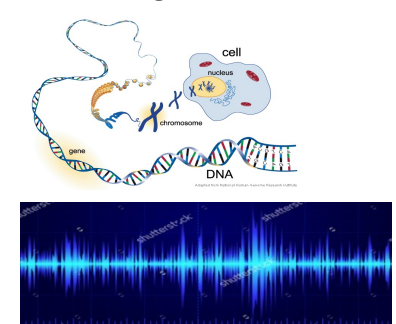
texts in websites, emails, articles, tweets

The collage shows three examples of unstructured text data: a Wikipedia 'Welcome' page, a tweet from Dwayne Johnson (@TheRock) saying 'Sometimes as a father, you ARE the only solution. A real honor making this true story. eSNITCH 9/19/19 nic twitter.com/AJhoF6dt', and a news article titled 'Seeking Life's Bare (Genetic) Necessities' from Cold Spring Harbor, New York, discussing genome mapping and sequencing.

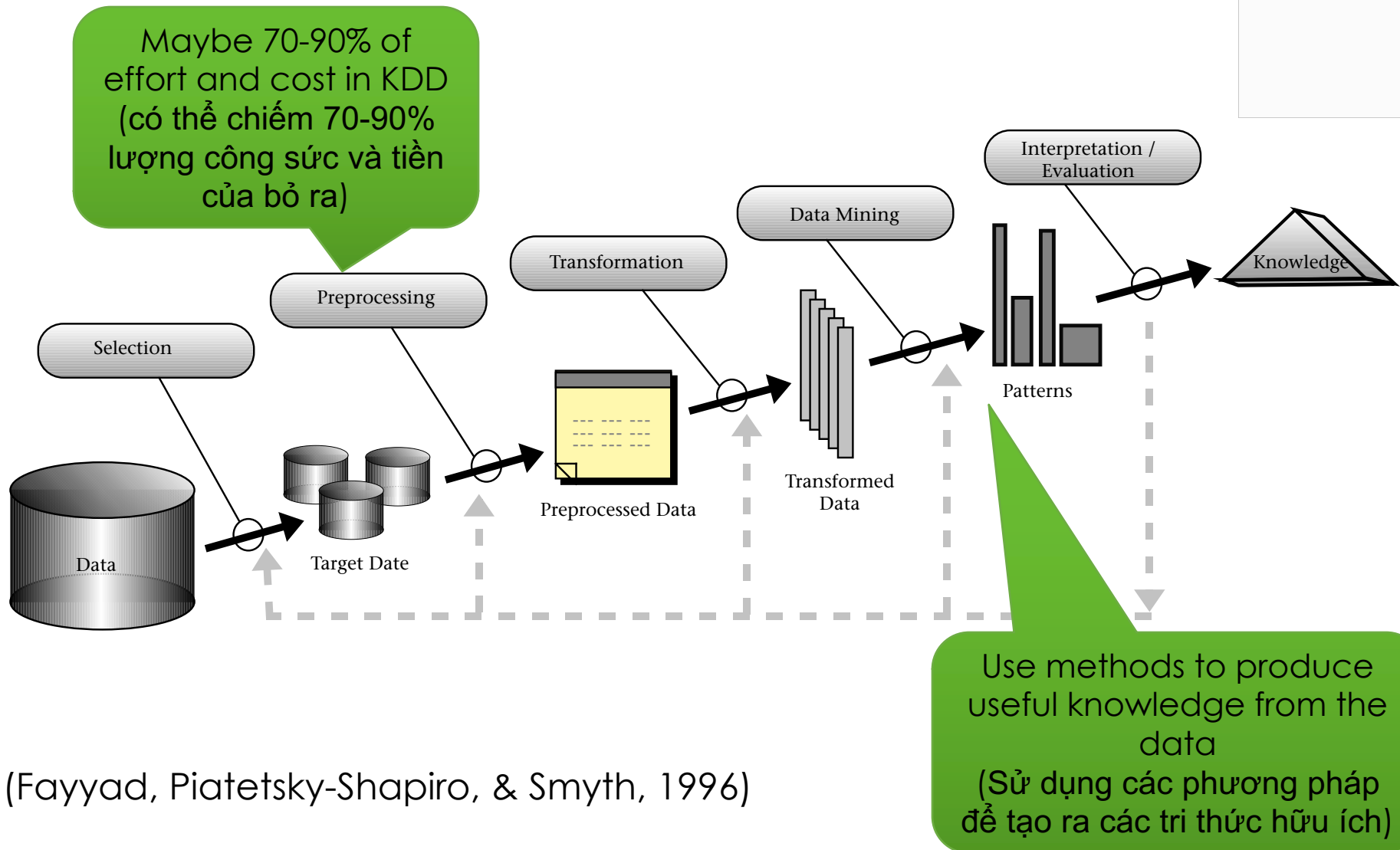
2D/3D images, videos + meta



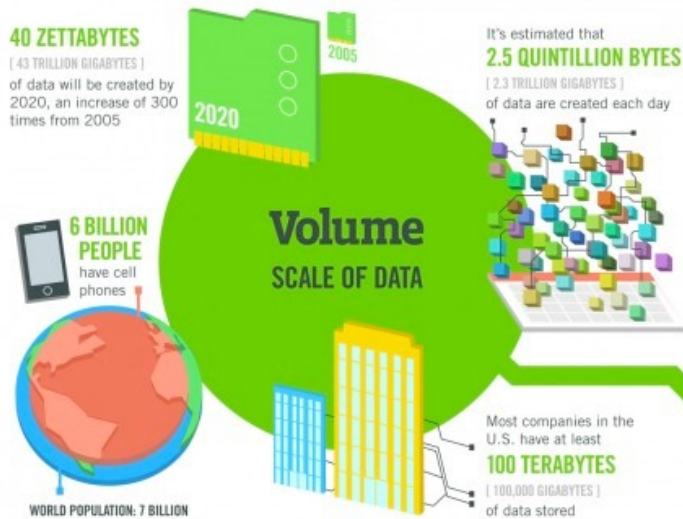
spectrograms, DNAs, ...



# KDD: methodology



# KDD: challenges



## The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015  
**4.4 MILLION IT JOBS**  
will be created globally to support big data,  
with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**  
[ 161 BILLION GIGABYTES ]



**30 BILLION PIECES OF CONTENT**  
are shared on Facebook every month



**Variety**  
DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**  
are watched on YouTube each month



**400 MILLION TWEETS**  
are sent per day by about 200 million monthly active users



The New York Stock Exchange captures  
**1 TB OF TRADE INFORMATION**  
during each trading session



**Velocity**  
ANALYSIS OF STREAMING DATA

Modern cars have close to **100 SENSORS**  
that monitor items such as fuel level and tire pressure



By 2016, it is projected there will be  
**18.9 BILLION NETWORK CONNECTIONS**  
— almost 2.5 connections per person on earth



**1 IN 3 BUSINESS LEADERS**

don't trust the information they use to make decisions



**27% OF RESPONDENTS**

in one survey were unsure of how much of their data was inaccurate

**Veracity**  
UNCERTAINTY OF DATA

Poor data quality costs the US economy around

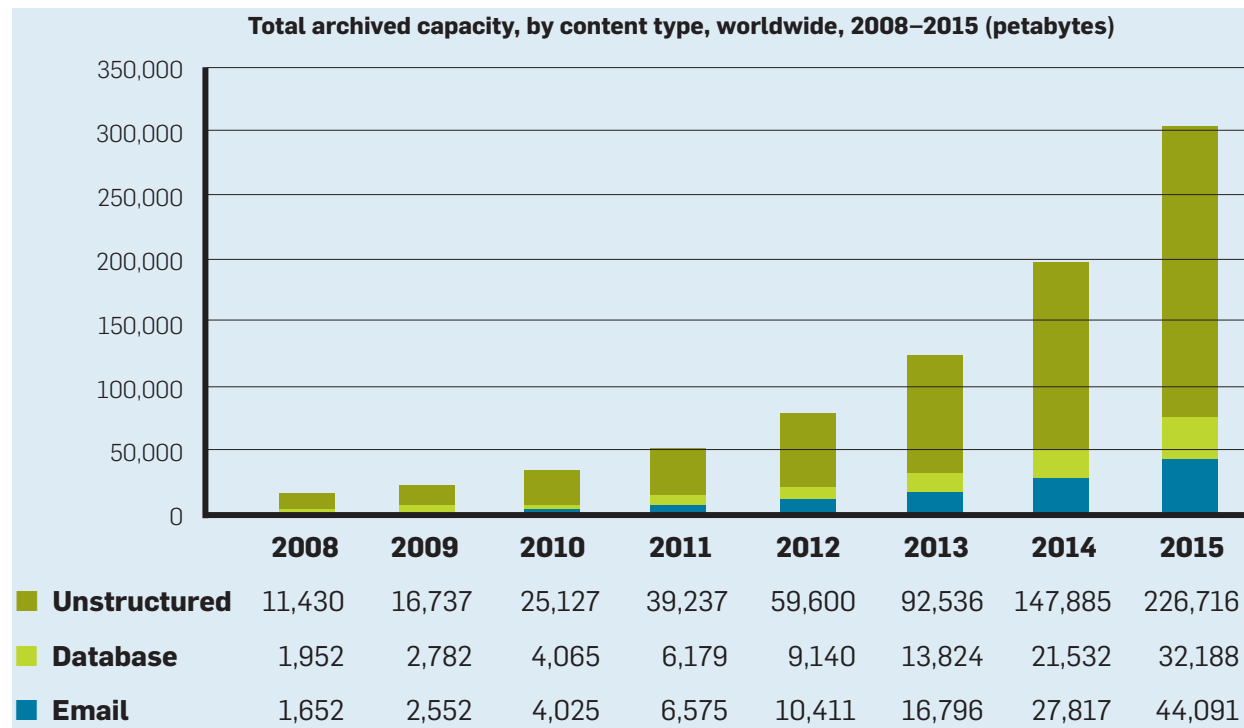
**\$3.1 TRILLION A YEAR**





# Challenges: free structure

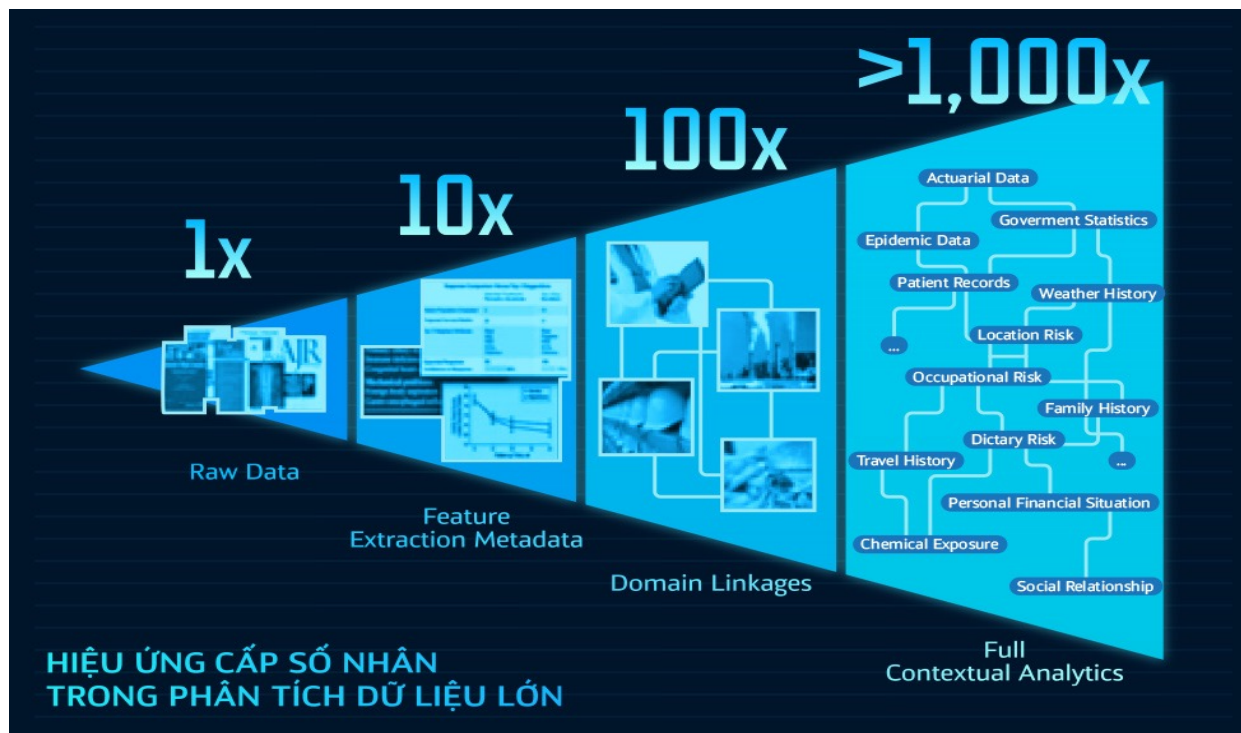
- Unstructured data increases extremely fast
  - Texts, images, tags, links, likes, emotions, ...



(Vasant Dhar, CACM, 2013)

# Challenges: hidden interaction

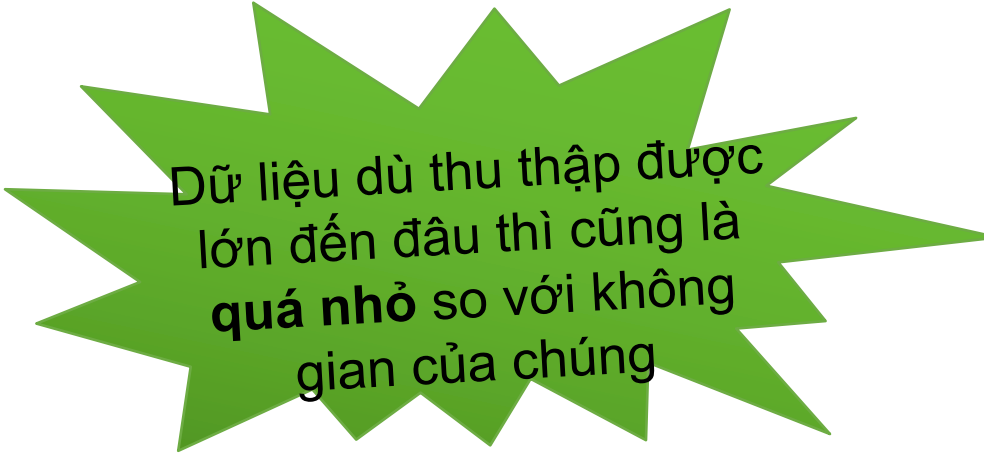
- The interactions or correlations hidden in data might be really huge
  - (Những mối tương tác ẩn chứa bên trong dữ liệu có thể rất lớn)



# Challenges: high dimensionality

---

- Real problems often have extremely **high dimensions** (số chiều của dữ liệu quá lớn)
  - Bicycle runs: 2 dimensions (a road)
  - We live: 4 dimensions
  - But an image 1024x1024: ~1 million dimensions
  - Text collections: million dimensions
  - Recomenders' 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

# References

---

- 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.