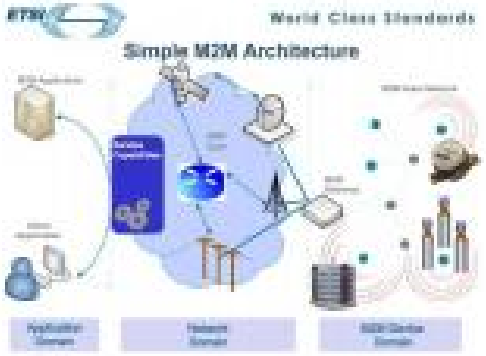
Giới thiệu

- Big Data
- Data Analysis
- Machine learning
- Data mining

Dữ liệu lớn - Big Data

Những xu hướng ảnh hưởng của CNTT

Điện toán đám mây



M2M (Machine to Machine)

Dữ liệu, thông tin và tri thức

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.

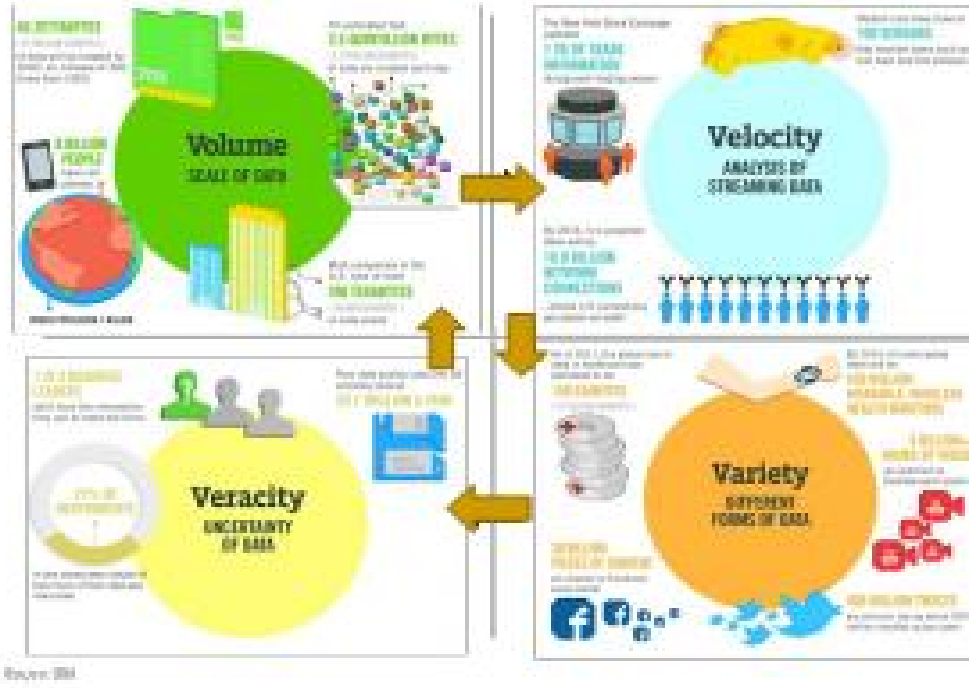
Thông tin là **dữ liệu với ý nghĩa** (data equiped with meaning), thu được khi **xử lý dữ liệu** để lọc bỏ đi các phần dư thừa, tìm ra phần cốt lõi đặc trưng cho dữ liệu.

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…

Dữ liệu ở mức độ trừu tượng thấp nhất và cụ thể nhất, thông tin ở mức trên dữ liệu và tri thức ở mức cao nhất.

Dữ liệu lớn - Big Data

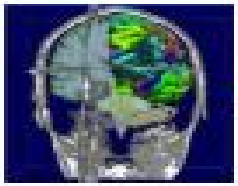
Dữ liệu lớn nói về các **tập dữ liệu rất lớn** và/hoặc **rất phức tạp**, vượt quá khả năng xử lý của các kỹ thuật IT truyền thống (*View 1*).



Dữ liệu lớn - Big Data

Rất lớn là lớn thế nào?

Kích thước lớn và rất nhiều chiều



1 human brain at the micron level = 1 PetaByte

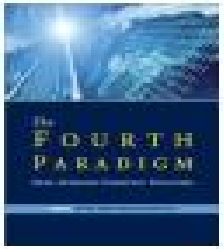
Large Hadron Collider, (PetaBytes/day)

Human Genomics = 7000 PetaBytes 1GB / person

Printed materials in the Library of Congress = 10 TeraBytes

200 of London's Traffic Cams (8TB/day)

1 book = 1 MegaByte



Family photo = 586 KiloBytes

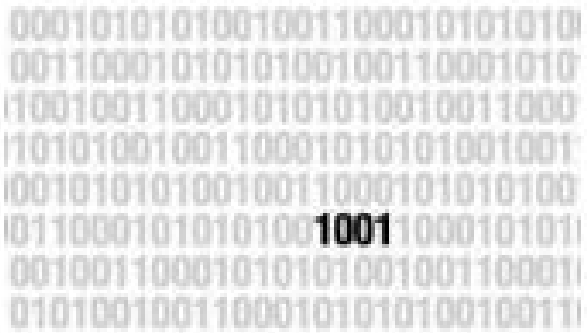
Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}

All worldwide information in one year = 2 ExaBytes

Dữ liệu lớn - Big Data

Dữ liệu lớn có thể rất nhỏ.
Không phải mọi tập dữ liệu to đều lớn
Big data can be very small. Not all large datasets are big

- **Big** liên quan tới sự **phức tạp** nhiều hơn tới **kích thước lớn.**
- **Dữ liệu lớn** nhưng lại nhỏ
  - Lò hạt nhân, máy bay... có hàng trăm nghìn sensors → sự phức tạp của việc tổ hợp dữ liệu các sensors này tạo ra?
  - Dòng dữ liệu của tất cả các sensors là lớn mặc dù kích thước của tập dữ liệu là không lớn (một giờ bay:
    100,000 sensors x 60 minutes
    x 60 seconds x 8 bytes < 3GB).
- Tập dữ liệu **to nhưng không lớn**
  - Số hệ thống dù tăng lên và tạo ra những lượng khổng lồ dữ liệu nhưng đơn giản.



Dữ liệu lớn - Big Data

- Ba nguồn chính là:

  (1) Các phương tiện truyền thông xã hội

  (2) Các máy móc thu nhận dữ liệu, các thiết bị công nghiệp, các cảm biến, các dụng cụ giám sát...

  (3) Giao dịch kinh doanh, từ số liệu giá cả sản phẩm, thanh toán, dữ liệu chế tạo và phân bố…

Dữ liệu lớn - Big Data

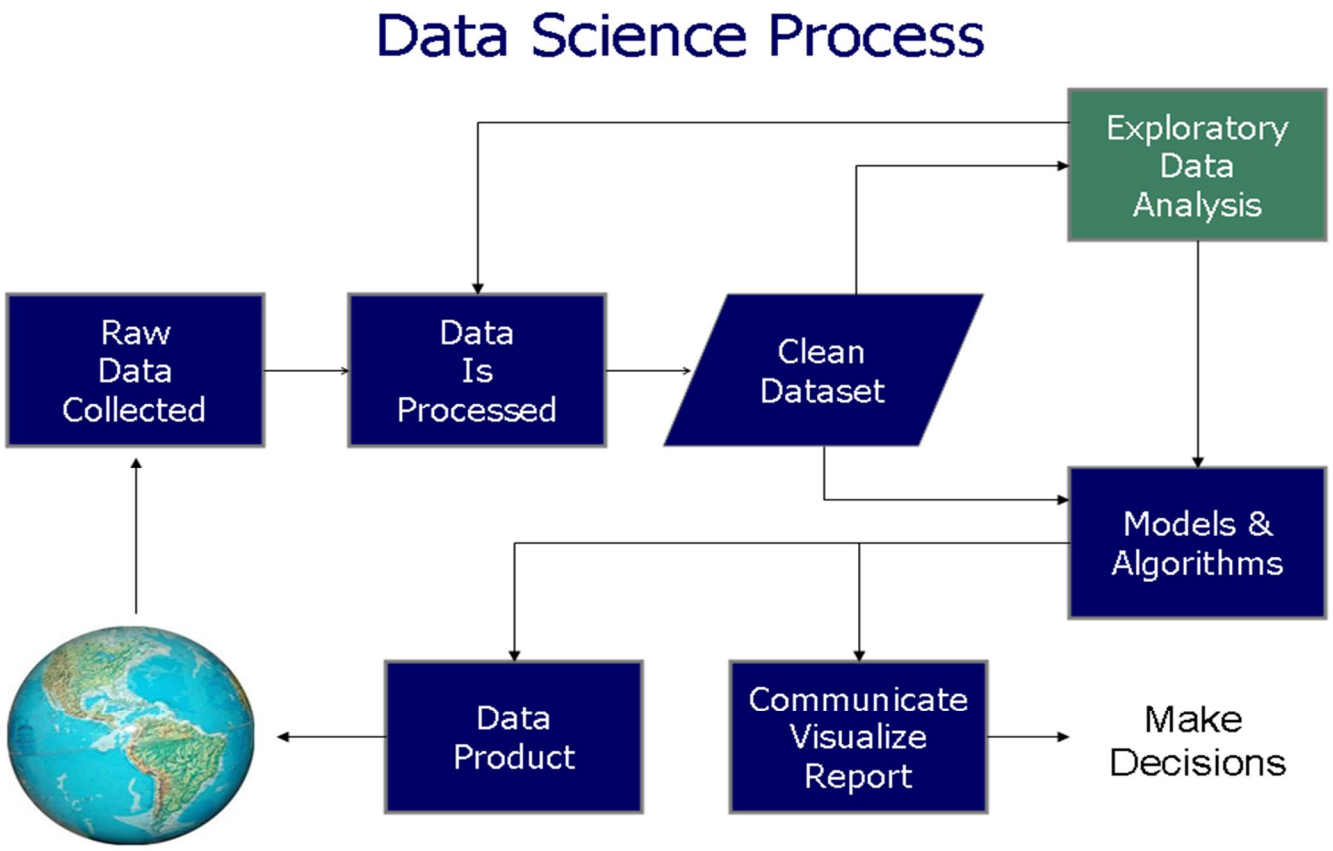
- Ba chìa khóa chính của khai thác dữ liệu lớn:

  (1) Quản trị dữ liệu, tức lưu trữ, bảo trì và truy nhập các nguồn dữ liệu lớn;

  (2) Phân tích dữ liệu, tức tìm cách hiểu được dữ liệu và tìm ra các thông tin hoặc tri thức quý báu từ dữ liệu;

  (3) Hiển thị dữ liệu và kết quả phân tích dữ liệu
- Thách thức chính của dữ liệu lớn là các phương pháp phân tích dữ liệu, trong đó chủ yếu là các phương pháp của hai lĩnh vực học máy và khai phá dữ liệu.

Phân tích dữ liệu-Data Analysis

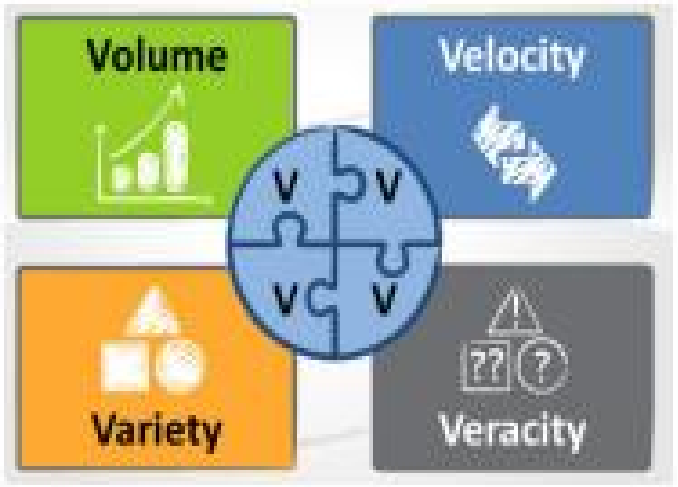


Khoa học dữ liệu là gì?

Tại sao phân tích dữ liệu lớn lại khó?

Bốn tính chất của dữ liệu (4V) và hai việc: dự đoán và phân tích quan hệ.

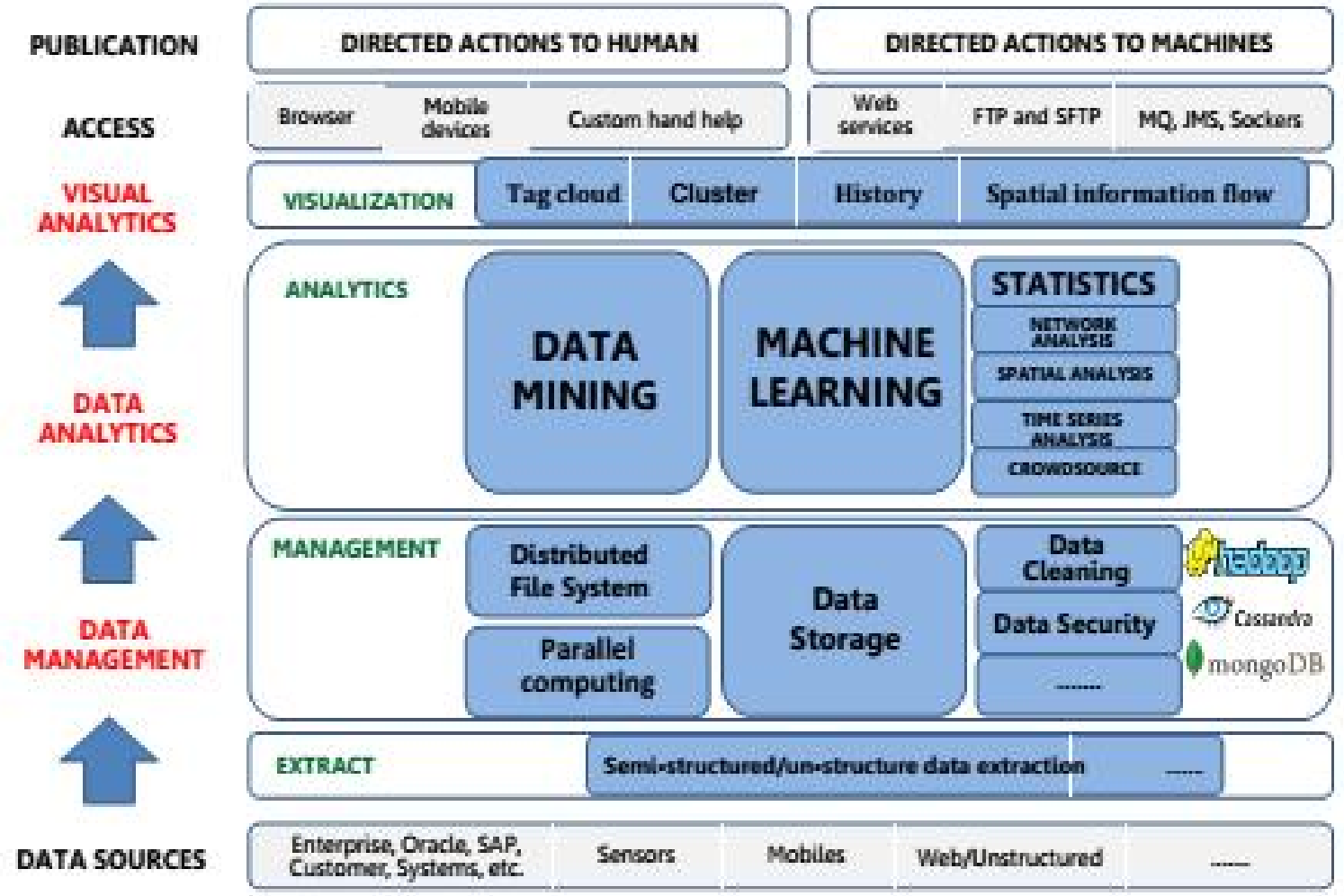
1. Số chiều rất lớn + dữ liệu kiểu khác nhau, chuyển động của dữ liệu, nhiễu trong dữ liệu → kém hiệu quả.
2. Số chiều rất lớn + số đối tượng rất lớn → tính toán nặng nề và thuật toán không khả kích (scalable)
3. Dữ liệu lớn đến từ nhiều nguồn, thu thập ở những thời điểm khác nhau bởi kỹ thuật khác nhau → không thuần nhất, khác biệt và lệch (bias)



Attribute	Numerical	Symbolic	
No structure $= \neq$		Places, Color	Nominal (categorical)
Ordinal structure $= \neq \geq$	Age, Temperature, Taste,	Rank, Resemblance	Ordinal
Ring structure $= \neq \geq + \times$	Income, Length		Measurable

→ **Mô hình thưa và giảm chiều (sparse modeling and dimensionality reduction)**

Một lược đồ phân tích dữ liệu lớn



Học máy

- Mục đích của học máy là việc xây dựng các hệ máy tính có khả năng thích ứng và học từ kinh nghiệm (Tom Dieterich).
- Một chương trình máy tính được nói là học từ kinh nghiệm **E** cho một lớp các nhiệm vụ **T** với độ đo hiệu suất **P**, nếu hiệu suất của nó với nhiệm vụ **T**, đánh giá bằng **P**, có thể tăng lên cùng kinh nghiệm
(T. Mitchell Machine Learning book)
- Khoa học về việc làm cho máy có khả năng học và tạo ra tri thức từ dữ liệu.

(from Eric Xing lecture notes)

Học máy

- Học máy là một lĩnh vực của CNTT nhằm làm cho máy tính có một số khả năng học tập của con người, chủ yếu là học để khám phá.
- Cốt lõi của việc tạo ra khả năng tự học này của máy là việc phân tích các tập dữ liệu để phát hiện ra các quy luật, các mẫu dạng, các mô hình.
- Kết hợp ngày càng nhiều hơn với toán học, đặc biệt với hai ngành thống kê và tối ưu, các phương pháp học máy càng mạnh hơn khi phân tích các dữ liệu phức tạp.

Khai phá dữ liệu - Data Mining

Tự động khám phá, phát hiện các tri thức tiềm ẩn từ các tập dữ liệu lớn và đa dạng.

Data mining metaphor: Extracting ore from rock

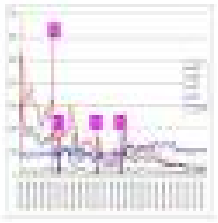
Statistics

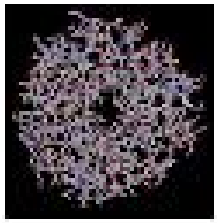
KDD

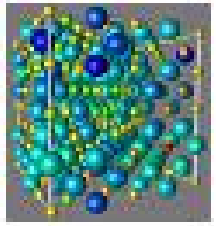
Databases

Machine Learning

Large and unstructured real-life data







Khai phá dữ liệu - Data Mining

- Khai phá dữ liệu tập trung vào việc đưa các phương pháp học máy vào phân tích, khai thác các tập dữ liệu lớn trong các lĩnh vực khác nhau.
- Những hướng nghiên cứu về mô hình làm thưa, giảm số chiều, mô hình đồ thị xác suất... trong hai lĩnh vực học máy và khai phá dữ liệu chính là những hướng đi tới các phương pháp phân tích dữ liệu lớn.

Học máy và Khai phá dữ liệu

Machine learning and data mining

Machine learning

- To build computer systems that learn as well as human does.
- ICML since 1982 (23th ICML in 2006), ECML since 1989.
- ECML/PKDD since 2001.
- ACML starts Nov. 2009.

Data mining

- To find new and useful knowledge from large datasets .
- ACM SIGKDD since 1995, PKDD and PAKDD since 1997 IEEE ICDM and SIAM DM since 2000, etc.