

Dữ liệu lớn thời chuyển đổi số

Hồ Tú Bảo

Viện Nghiên cứu Cao cấp về Toán
Viện John von Neumann, ĐHQG HCM
bao.ho@jvn.edu.vn, bao@viasm.edu.vn





1993-2018



2010~



4.2018~



Outline



Ben Chams - Fotolia

- Dữ liệu lớn?
- Dữ liệu lớn và các công nghệ số liên quan



Dữ liệu lớn với
kinh tế số và chính phủ số

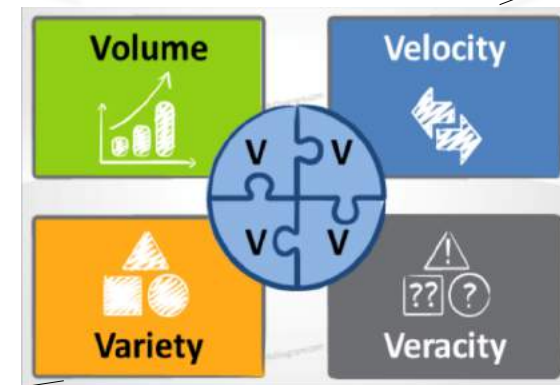
Nhìn lại về dữ liệu lớn

Dữ liệu lớn là dữ liệu **rất lớn** và **rất phức tạp** hiện chưa phân tích hiệu quả với các kỹ thuật CNTT truyền thống.



Từ Terabytes đến Petabytes (10^{15} bytes) hay Zetabytes (10^{18} bytes)

Dòng chuyển động của dữ liệu trực tuyến lớn

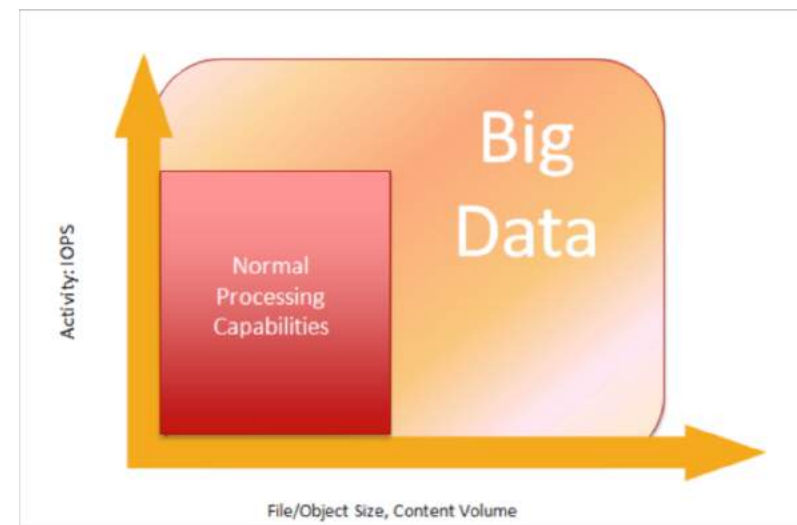


Dữ liệu trong nhiều kiểu cấu trúc đến dữ liệu không cấu trúc.

Độ chính xác và tin cậy, trung thực của dữ liệu.

Nhìn lại về dữ liệu lớn

- Tìm ra được **giá trị** tiềm ẩn trong dữ liệu lớn phụ thuộc vào khả năng phân tích và khai thác được các nguồn dữ liệu này.
- Dữ liệu lớn là một yếu tố ảnh hưởng trong chuyển đổi số.



Dữ liệu lớn đến từ đâu?

online



Nhấp chuột
Mua hàng
Transactions
Networks log
...
Everything online
~ 8 hour / day

Dữ liệu từ mạng xã hội



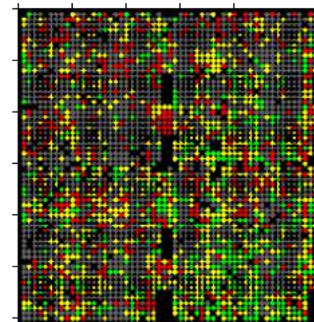
Social sensors

**BIG
DATA**

Kết nối vạn vật và thiết bị thông minh



Dữ liệu từ nghiên cứu khoa học

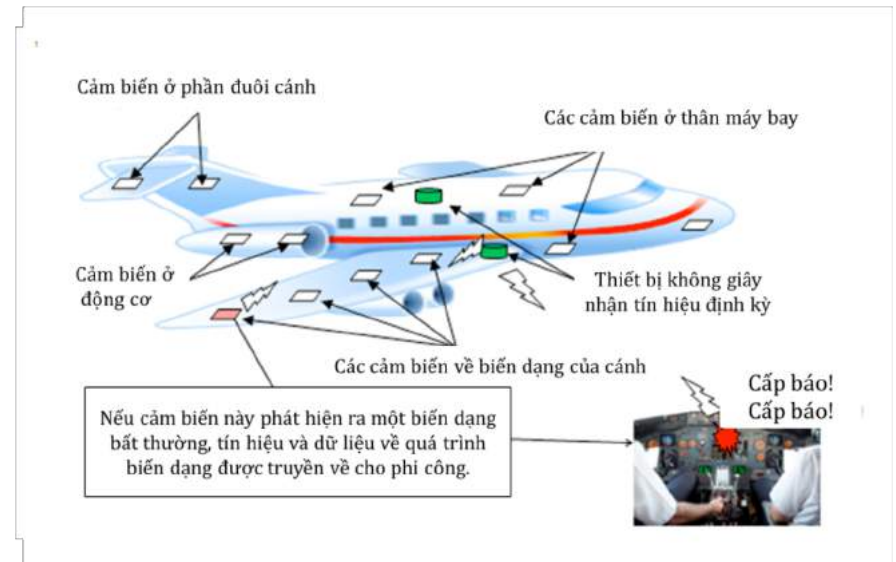


Dữ liệu từ sinh
học (gene
expression)
Nghiên cứu vũ trụ
Nông nghiệp

Dinh's slides

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 → nhỏ hơn 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.

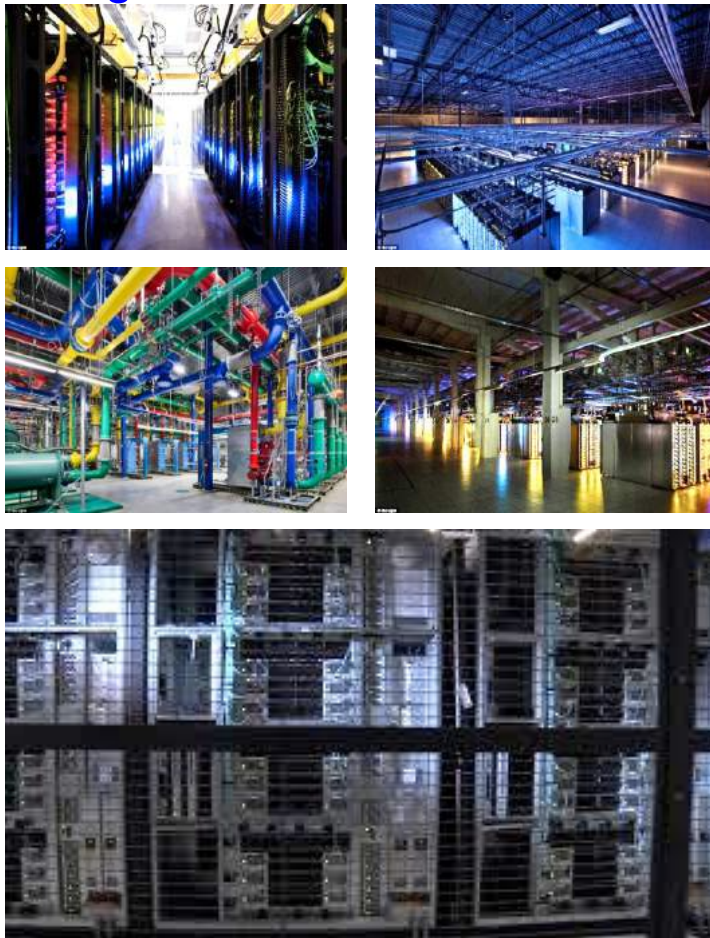


0001010101001001100010101010
0011000101010100100110001010
1001001100010101010010011000
1010100100110001010101001001
0010101010010011000101010100
011000101010100100010010101
0010011000101010100100110001
0101001001100010101010010011

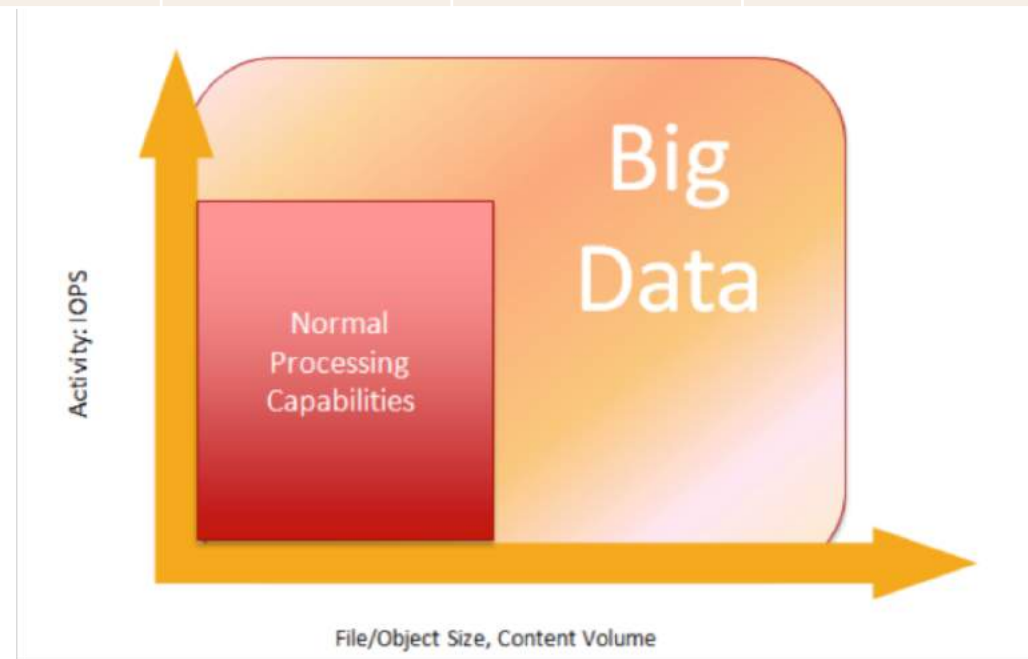
Nguồn: MIKE2.0

Dữ liệu lớn: thách thức

Google Data Center



Bài toán	Số thuộc tính (triệu)	Số thuộc tính lựa chọn (triệu)	Tỷ lệ số thuộc tính lựa chọn
A	868	20	2.3%
B	333	8	2.4%
C	1762	252	14.3%
D	2172	372	17.1%



Singer Yoram, keynote at ACML'14

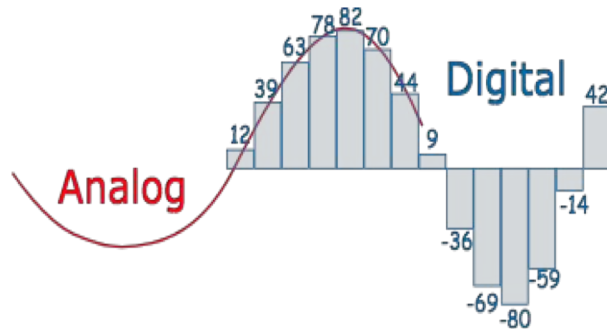
Chuyển đổi số - Digital transformation

- **Chuyển đổi số** là việc chuyển cách giải quyết các vấn đề truyền thống bằng **cách làm mới** khi dùng công nghệ số.
- **Công nghệ số**: Công nghệ **số hoá** và công nghệ **xử lý** các đối tượng được số hoá.
- Hai phạm trù **số hoá**: **Số hoá thông tin** (digitization) và **số hoá công nghiệp và tổ chức** (digitalization).
- Tạo ra **phương thức** sản xuất mới, tăng năng suất lao động.



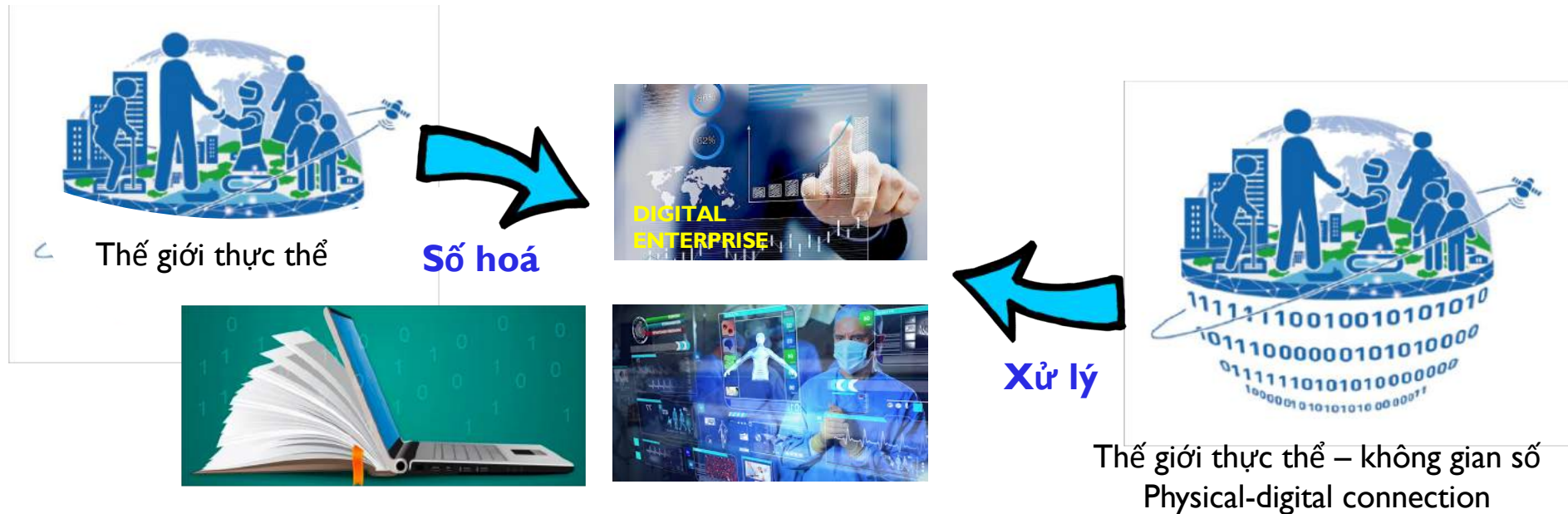
Chuyển đổi số - Digital transformation

Chuyển đổi số là việc chuyển cách giải quyết các vấn đề truyền thống bằng **cách làm mới** khi dùng **công nghệ số**.

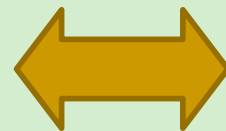


digital transformation of enterprises

Chuyển đổi số - Digital transformation



**Mọi hoạt động trong
thế giới các thực thể**



**Tính toán, điều khiển
trên không gian số**

Phương thức sản xuất mới (kinh tế, xã hội, văn hoá)

Kinh tế số – Digital economy

Nền kinh tế vận hành chủ yếu dựa trên công nghệ số



Hình 11 Các định nghĩa về nền kinh tế số từ hẹp đến rộng

Nguồn từ: Buhkt and Heeks³⁹

Chính phủ điện tử và chính phủ số

■ **Chính phủ điện tử** (E-Governmen):

Sử dụng CNTT-TT, đặc biệt là Internet, như một công cụ để đạt được hiệu quả tốt hơn.

■ **Chính phủ số** (Digital Government):

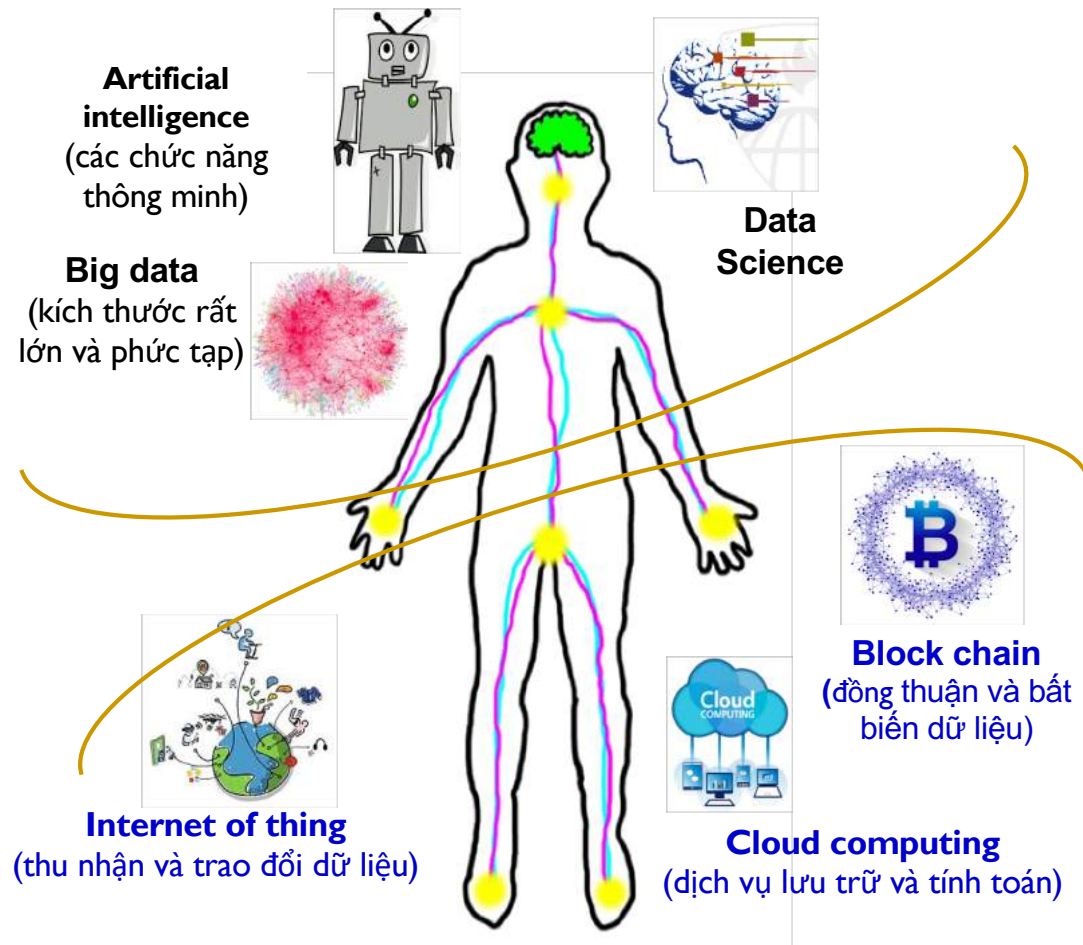
- ❑ Dùng công nghệ số để hiện đại hóa Chính phủ.
- ❑ Tạo hệ sinh thái gồm các Chính phủ, doanh nghiệp, tổ chức xã hội và người dân.
- ❑ Thúc đẩy việc tạo ra và truy cập dữ liệu, dịch vụ và nội dung số thông qua sự tương tác với Chính phủ.



Nguồn: Gabriel Puron-cid

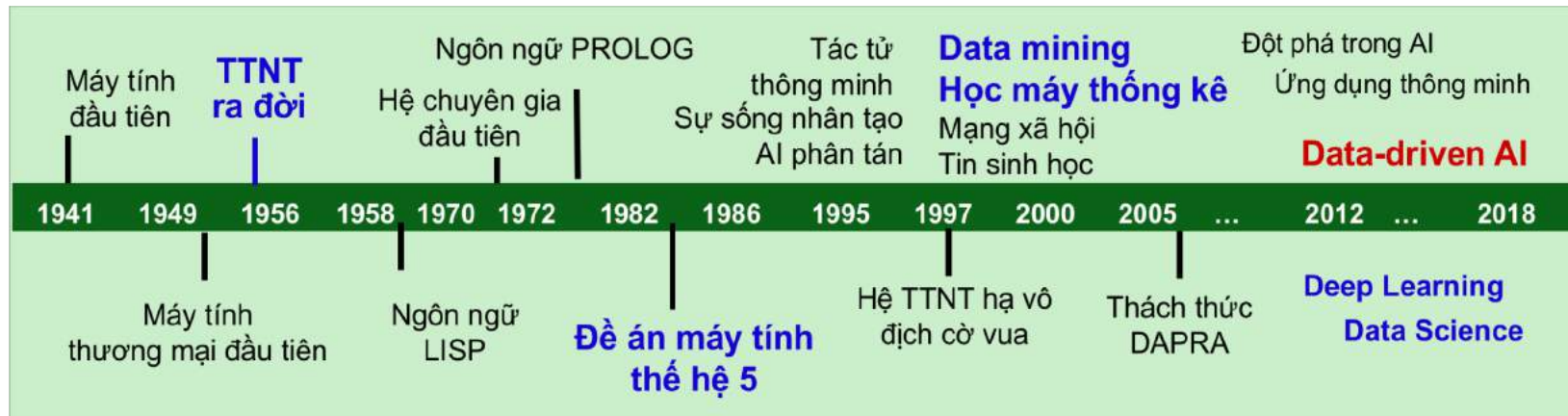
Big data và các công nghệ số đột phá

- **Điện toán đám mây:**
Môi trường
- **Internet vạn vật:**
Huyết mạch
- **Block chain:**
Đồng thuận và bất biến dữ liệu
- **Trí tuệ nhân tạo:**
Chức năng thông minh
- **Dữ liệu lớn:**
Năng lượng
- **Khoa học dữ liệu:**
“Bộ não” phân tích dữ liệu để hỗ trợ đưa ra quyết định.

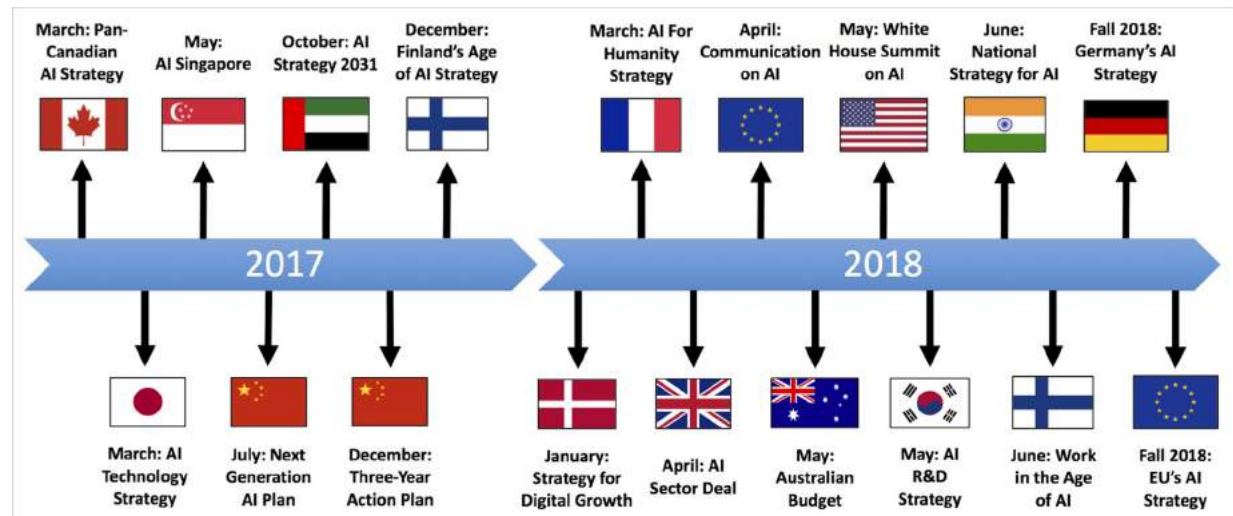


Cần nhìn dữ liệu lớn trong bức tranh về công nghệ số

AI ngày nay là “data-driven AI”

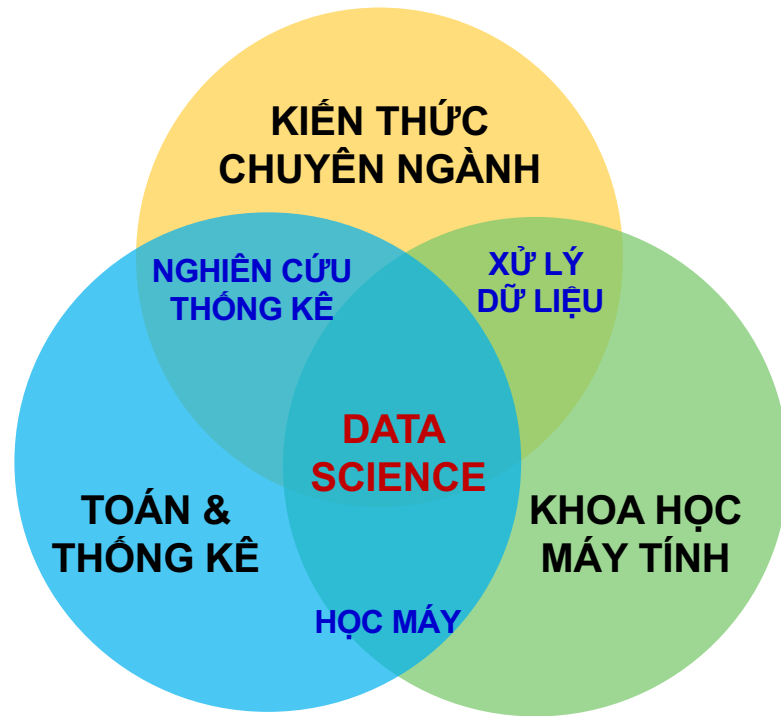


AI = Dùng dữ liệu một cách thông minh



<https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd>

Khoa học Dữ liệu - Data science



Khoa học về phân tích dữ liệu

“In God we trust.
All others bring
data”.

“Ta tin Thượng đế.
Ngoài ra, là dữ
liệu”.

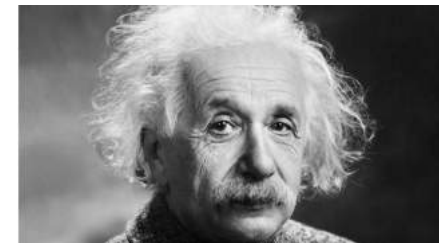
W.E. Deming



“We cannot solve problems
by using the same kind of
thinking we used when we
created them”

Ta không thể giải quyết các
vấn đề với chính cách nghĩ ta
đã dùng khi đặt vấn đề

Albert Einstein



Kết hợp toán học và tin học là công cụ phân tích dữ liệu lớn

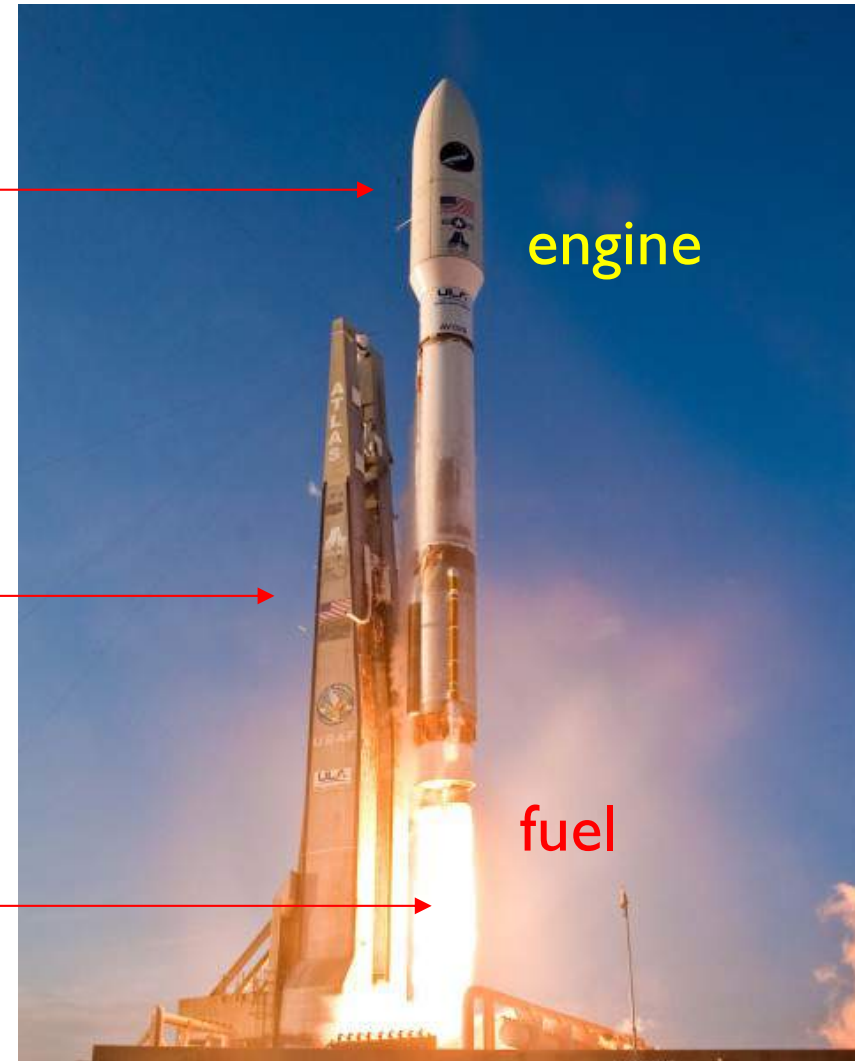
Vai trò của dữ liệu lớn?

Andrew Ng's Analogy

Mô hình tính toán lớn,
e.g., deep learning/AI

Cơ sở hạ tầng tính
toán, nhà nghiên cứu,
môi trường và chính
sách như bộ phóng

Dữ liệu là năng lượng.
Dữ liệu lớn là nguồn
năng lượng khổng lồ



engine

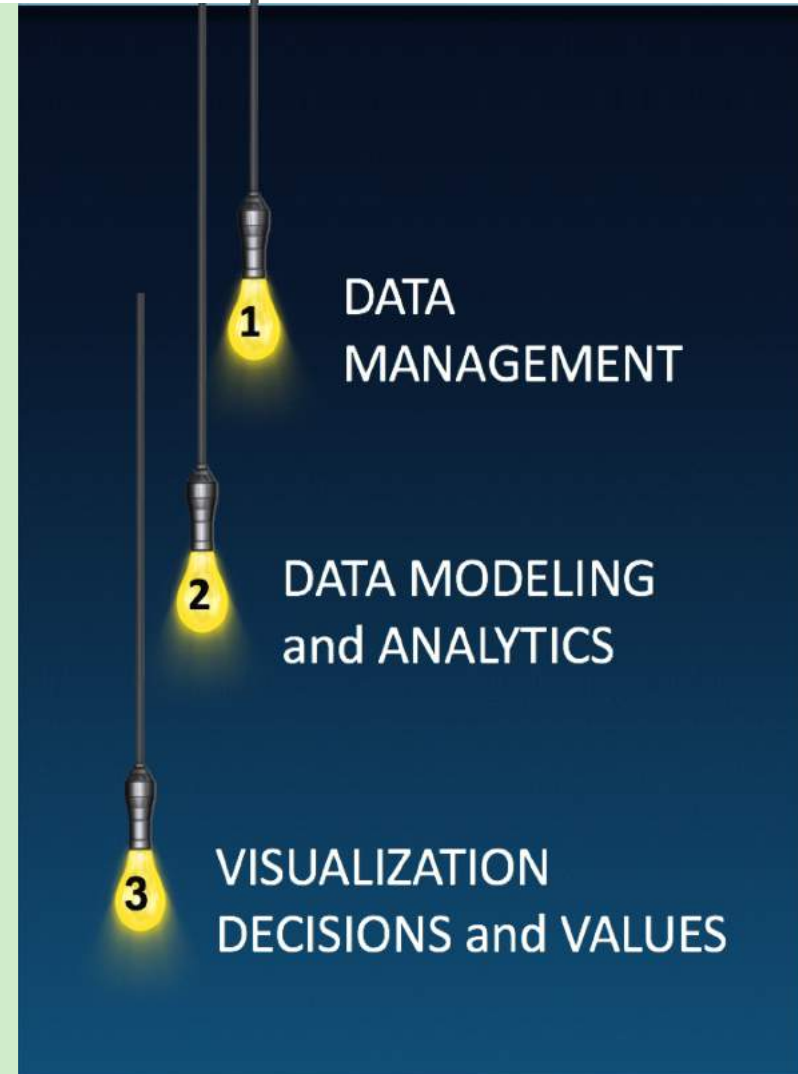
fuel

Dinh's slides

Chìa khoá của dữ liệu lớn

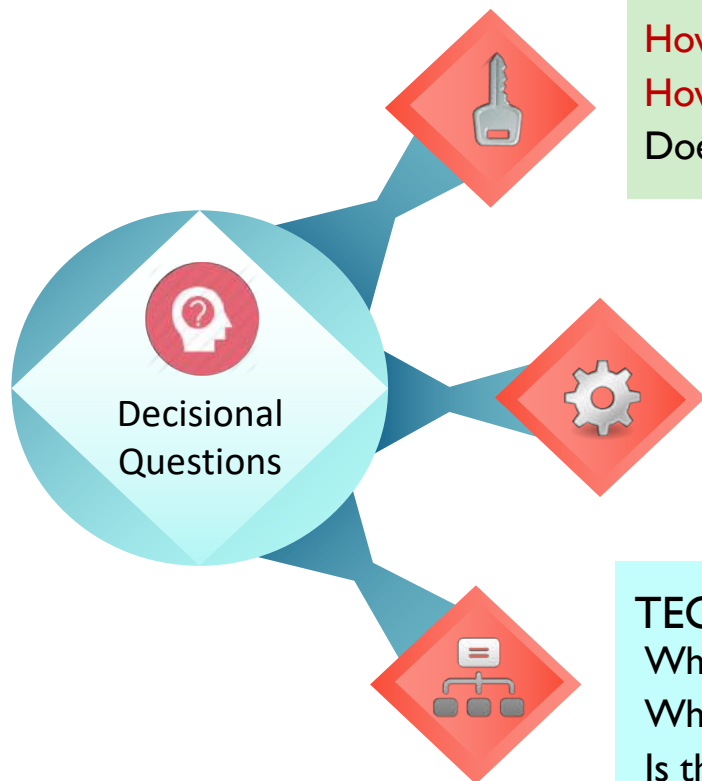
Đâu là chìa khoá khoa học và công nghệ của dữ liệu lớn?

- ❑ **Quản trị dữ liệu**, tức lưu trữ, bảo trì và truy nhập các nguồn dữ liệu lớn.
- ❑ **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.
- ❑ **Trao đổi, hiển thị dữ liệu** và kết quả phân tích dữ liệu để tạo ra sản phẩm hay giá trị.



Chìa khoá của dữ liệu lớn

■ Cần hỏi gì khi tiếp cận dữ liệu lớn?



FUNDAMENTAL CONCERNS

How **quickly** do we need to get the results?

How **big** is the data to be processed?

Does the model building require **several** iterations or a **single** iteration?

SYSTEM CONCERNS

Will there be a need for **more data** processing capability **in the future**?

Is **the rate of data transfer** critical for this application?

Is there a need for **handling hardware failures** within the application?

TECHNOLOGY CONCERNS

What are **the infrastructures** (cloud/physical systems) to be used?

What are **the technologies** to be used for distributed/parallel processing?

Is there a need to invest into **researching a new model**?

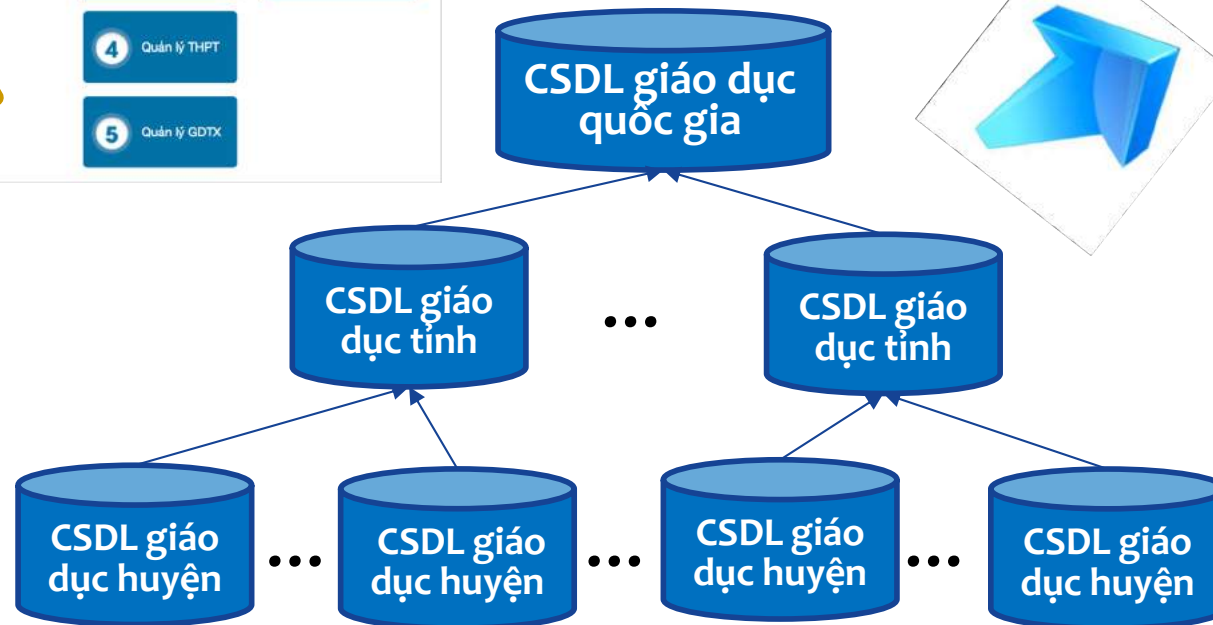
[Reddy and Singh, A survey on platforms for big data analytics, *Journal of Big Data*, 2014]

Dữ liệu lớn: CSDL giáo dục quốc gia

www.csdل.gov.vn



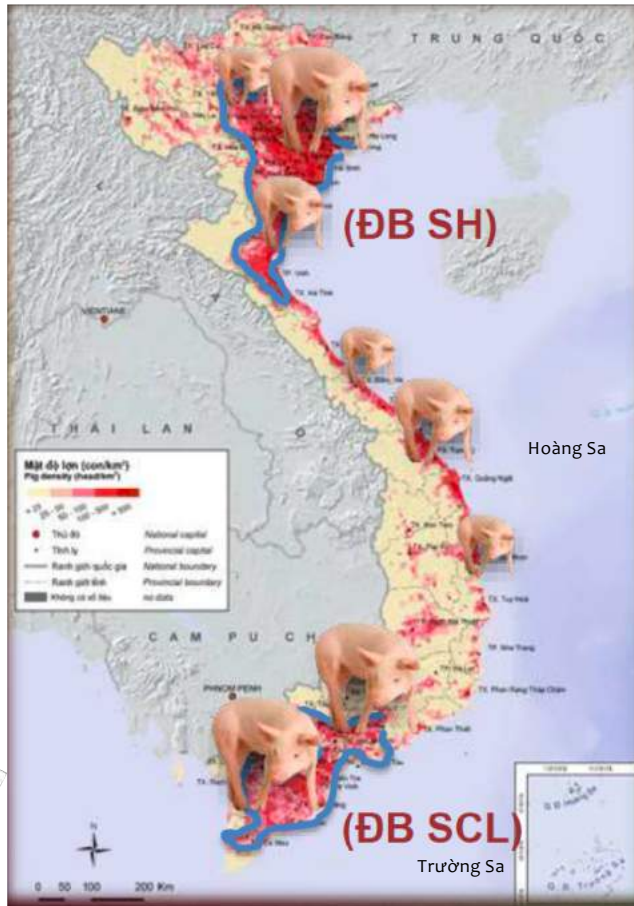
Số định danh cá nhân ngành giáo dục gồm 10 số, thống nhất và duy nhất.



CSDL lõi, gồm

- Trường (44 nghìn)
- Lớp (760 nghìn)
- Giáo viên (>1 triệu)
- Học sinh (23 triệu)

Dữ liệu lớn: CSDL nông nghiệp?



Nông nghiệp thông minh không chỉ làm trang trại thông minh, mà phải tính được cung-cầu, chọn sản phẩm “át chủ bài”, dự đoán thị trường... tất cả dựa vào **cơ sở dữ liệu nông nghiệp** và phân tích dữ liệu.

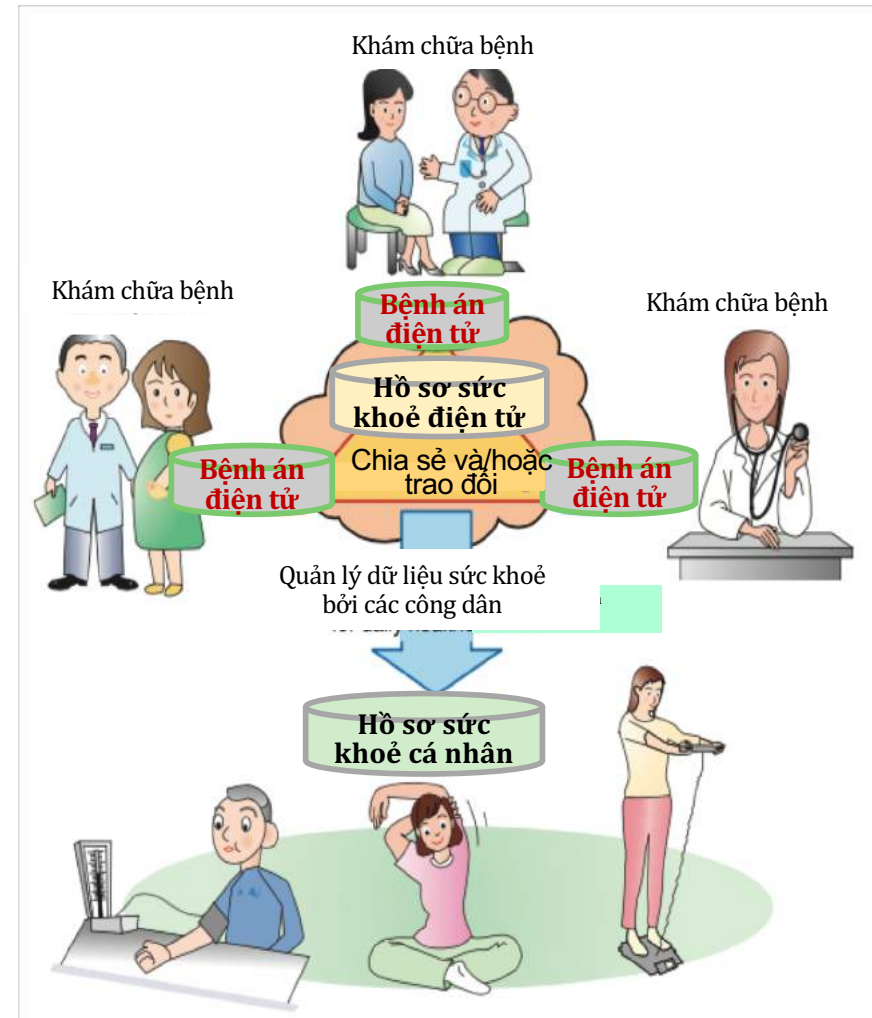


Dữ liệu lớn: Y tế thông minh

Problem: Xây hạ tầng số y tế

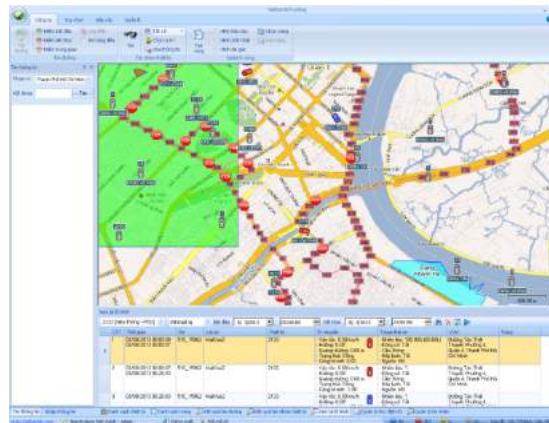
→ Khai thác nâng chất lượng y tế

1. Triển khai **bệnh án điện tử (BAĐT)** tại bệnh viện từ 1/3/2019 (thông tư 46/2018/TT-BYT).
2. Kết nối BAĐT giữa các bệnh viện để tạo dựng CSDL **hồ sơ sức khỏe điện tử** của công dân.
3. Mọi công dân có và quản lý **hồ sơ sức khỏe cá nhân** như tài sản riêng.
4. Khai thác **dữ liệu lớn y tế** hỗ trợ chăm sóc sức khỏe cho công dân.



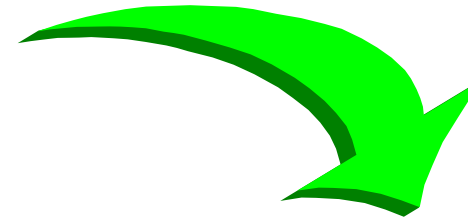
Dữ liệu lớn: Giám sát hành trình

Dữ liệu giám sát hành trình GPS của ~1.000.000 xe tải, xe khách, taxi (~1,5 tỷ bản ghi (~500Gb)/ngày → một phần lớn ở HCMC.

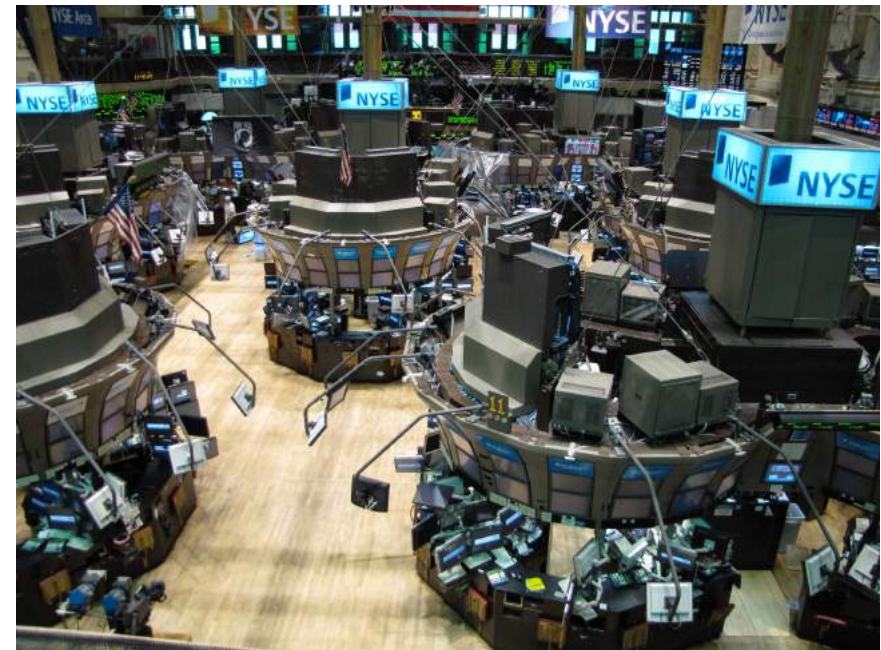


```
{
  "_id" : ObjectId("59e242f65471713a48d46efd"), // ID của bản tin, tự sinh bởi hệ thống
  "vehicle" : "90A01049", // Biển KS
  "ProviderId" : NumberInt(1014), // Mã đơn vị truyền dữ liệu
  "datetime" : NumberInt(1508025744), // Thời gian truyền dữ liệu
  "driver" : "AT094147", // Mã tài xế
  "speed" : NumberInt(78), // Tốc độ tại thời điểm truyền dữ liệu
  "x" : 106.0902, // Tọa độ LONG
  "y" : 20.45321, // Tọa độ LAT
  "heading" : NumberInt(0), // Hướng của xe. Từ 0 - 359. 0 là hướng chính Bắc
  "isOverSpeed" : false, // TRUE: Vss > MaxValidSpeed; FALSE: Vss > MaxValidSpeed
  "Vss" : NumberInt(70), // Vận tốc được tính toán theo hướng dẫn trong phụ lục 02 của TT 09/BGTVT
  "MaxValidSpeed" : NumberInt(90) // Vận tốc tối đa cho phép di chuyển trên đoạn đường đó
}
```

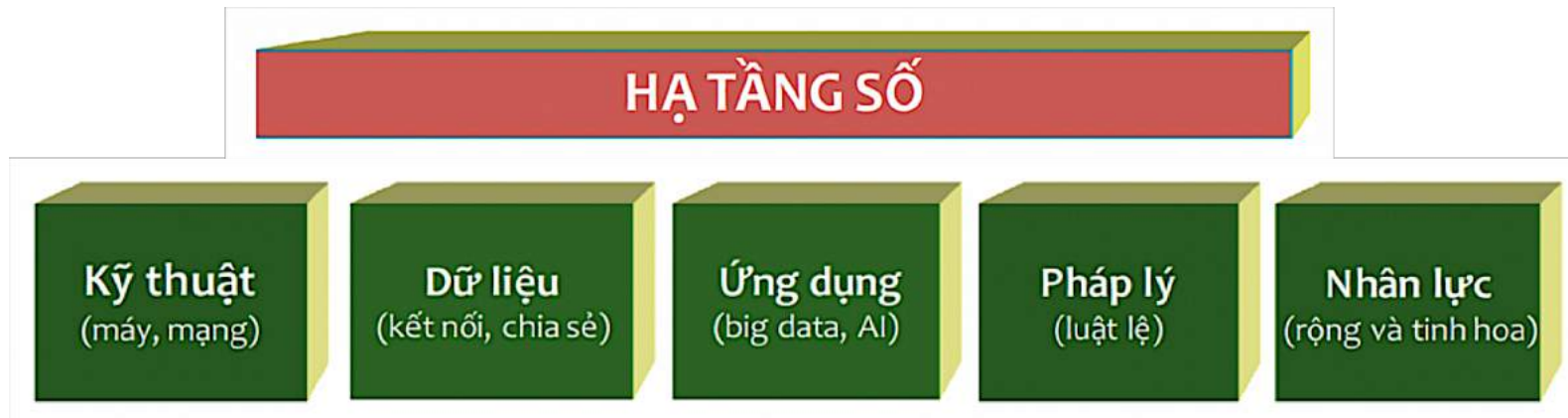
Dữ liệu lớn: Chuyển đổi số tài chính



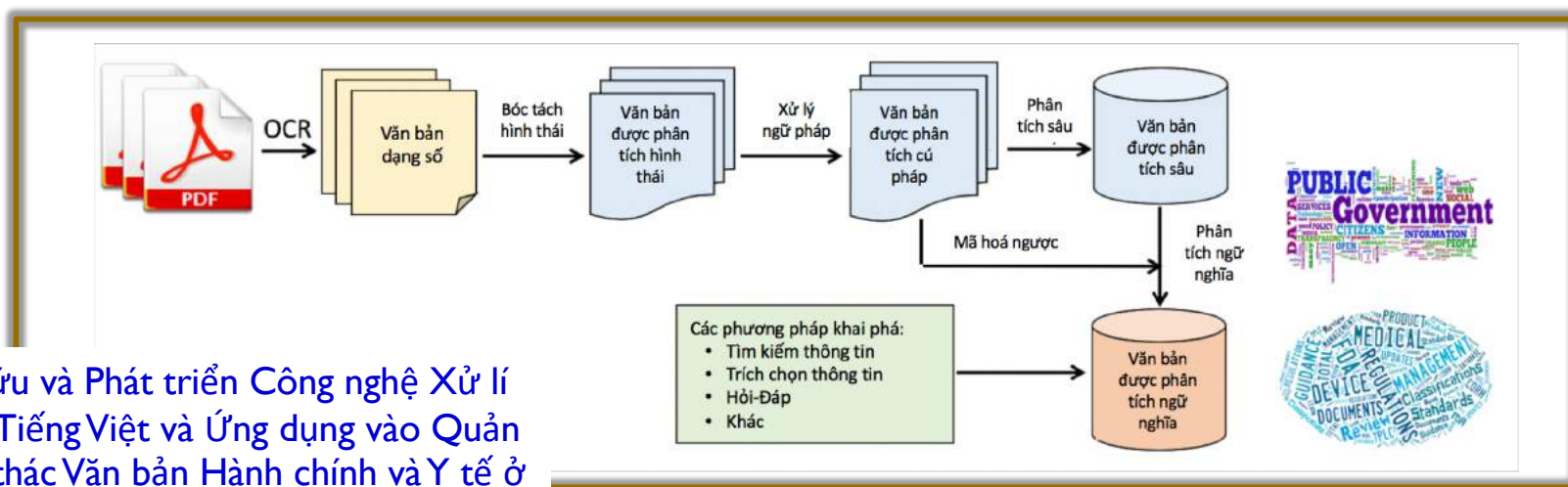
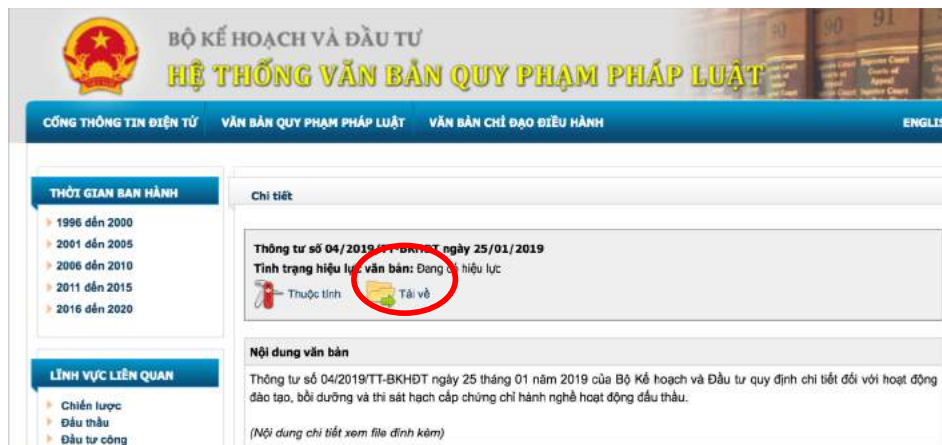
- Tài chính là lĩnh vực được số hoá rất rộng rãi.
- Số hoá quá trình (digitalization), xây dựng các CSDL và phân tích, khai thác các dữ liệu này trong hoạt động tài chính vẫn là thách thức lớn.



Dữ liệu lớn: Hạ tầng số của smart cities

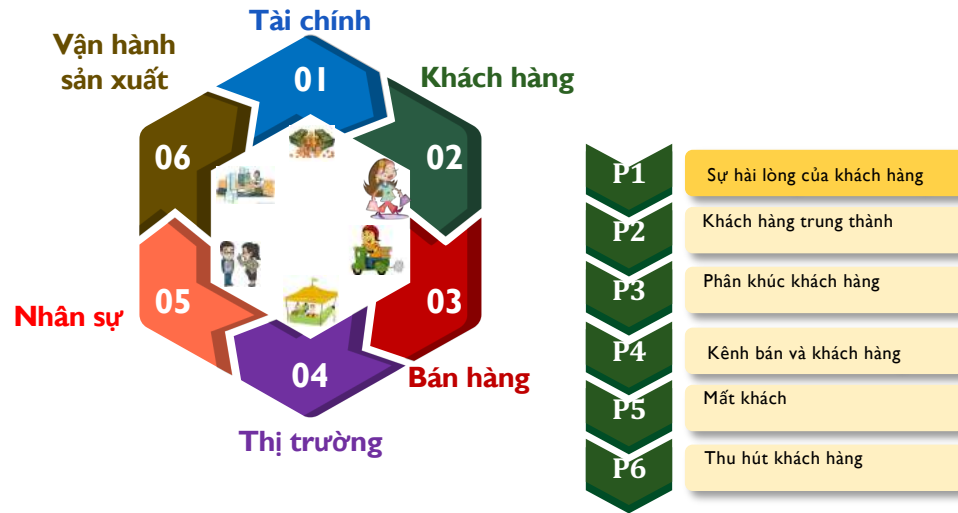


Dữ liệu lớn: Văn bản hành chính



“Nghiên cứu và Phát triển Công nghệ Xử lý Ngôn ngữ Tiếng Việt và Ứng dụng vào Quản lý và Khai thác Văn bản Hành chính và Y tế ở Việt Nam”, DSLab/VIASM project.

Dữ liệu lớn: Phân tích kinh doanh



ID	ISDN	DI	DISTRICT	NAME	PACKAGE	NAME	ACTIVE	DATE	DISCOUNT	CREDIT	DEVICE	TYPE	MODEL	OS	
1	090342002	Thành phố Hồ Chí Minh	Quận 12	Moviscid	07/24/2008	1	0.000000	1	OPTIMUS G (J370)					Google Android 4.2.2	
2	090342002	Thành phố Hồ Chí Minh	Quận 12	Moviscid	10/30/2013	0	0.000000	0							
3	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	8/9/2008	0	0.000000	0						2020 Symbol Series 40	
4	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	7/2/2008	0	0.000000	0						Symbol Series 40	
5	090342002	Thành phố Hồ Chí Minh	Quận 12	Moviscid	10/30/2008	0	0.000000	0			1	GT-T9000-Phone		Android 2.2 v1 (Jelly Bean)	
6	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	6/25/2014	0	0.000000	0					Galaxy S III Mini (GT-S7570)		
7	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
8	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
9	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
10	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
11	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
12	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
13	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
14	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
15	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
16	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
17	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
18	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
19	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
20	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
21	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
22	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
23	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3
24	090342002	Thành phố Hồ Chí Minh	Quận 10	Moviscid	3/20/2013	0	0.000000	0					1	Galaxy Note II Lite	Google Android 4.3



- Doanh nghiệp có đang cung cấp đúng những sản phẩm, dịch vụ mà cái khách hàng mong muốn?
- Khách hàng có hài lòng về trải nghiệm khi mua sản phẩm, dịch vụ của doanh nghiệp?
- Khách hàng có hài lòng về sản phẩm, dịch vụ của doanh nghiệp?
- Họ hài lòng ở mức độ nào và đâu là yếu tố khiến họ hài lòng nhất? Nếu không thì nguyên nhân là vì sao?

Hiding within huge mounds
of data is knowledge that
could change the our world.

Ẩn trong khối dữ liệu khổng
lồ là kiến thức có thể thay
đổi thế giới của chúng ta.

Thank you