

Giới thiệu ngành Data Analytics

23 – 07 – 2016

Hoa M. Le¹ & Linh Nghiem²

¹Research Assistant – Knowledge and Data Engineering Lab, Hanoi University of Science and Technology

²PhD Candidate, Statistical Science, Southern Methodist University, USA

Nội dung

Tổng quan về Data Analytics

- Định nghĩa
- Thành phần

Sự hiện diện của Data Analytics

- ... Trong cuộc sống hàng ngày
- ... Trong kinh tế/kinh doanh
- ... Trong công nghệ
- ... Trong khoa học, nghệ thuật

DATA ANALYTICS LÀ GÌ?

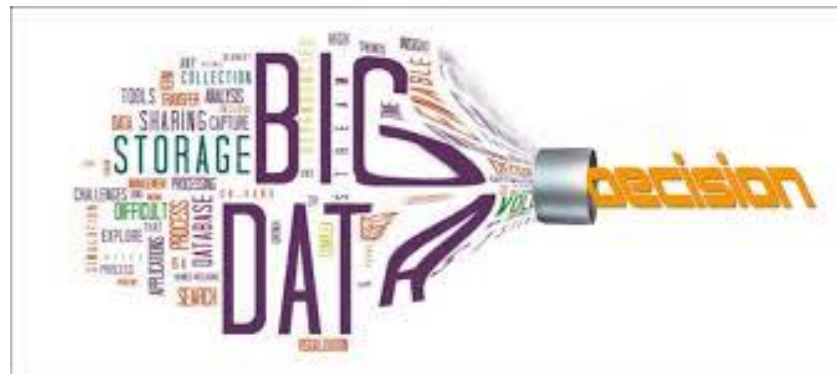
Data Analytics (DA) là gì?

- ... là Khoa học về phân tích dữ liệu nhằm rút ra kết luận.

<http://searchdatamanagement.techtarget.com/definition/data-analytics>

- ... là Khoa học về quá trình phân tích dữ liệu để phát hiện các **mẫu dạng** (pattern), các mối **tương quan** và **thông tin** có giá trị để ra quyết định tốt hơn.

http://www.sas.com/en_us/insights/analytics/big-data-analytics.html



DATA ANALYTICS

```
graph TD; A([DATA ANALYTICS]) --> B[KỸ THUẬT/ PHƯƠNG PHÁP]; A --> C[TRIỂN KHAI]; A --> D[CHUYÊN MÔN TRONG LĨNH VỰC ỨNG DỤNG];
```

KỸ THUẬT/ PHƯƠNG PHÁP

- Thống kê (Statistics)
- Machine Learning
- Khai phá dữ liệu (Data Mining)

TRIỂN KHAI

Computing power:

- Sử dụng Excel, SPSS, ...
- Lập trình (Programming, Engineering)
- Hạ tầng (Infrastructure)

CHUYÊN MÔN TRONG LĨNH VỰC ỨNG DỤNG

- Hiểu - Diễn giải kết quả
- Báo cáo, trình bày
- Đưa ra quyết định
- Đặt câu hỏi cho vấn đề mới

Kỹ thuật/Phương pháp

Thống kê

- Các phương pháp thu thập, mô tả, phân tích và quyết định dữ liệu.
- Nguồn gốc từ các lý thuyết toán học.
- Nhấn mạnh các suy diễn ước lượng và kiểm định giả thuyết.

Machine Learning

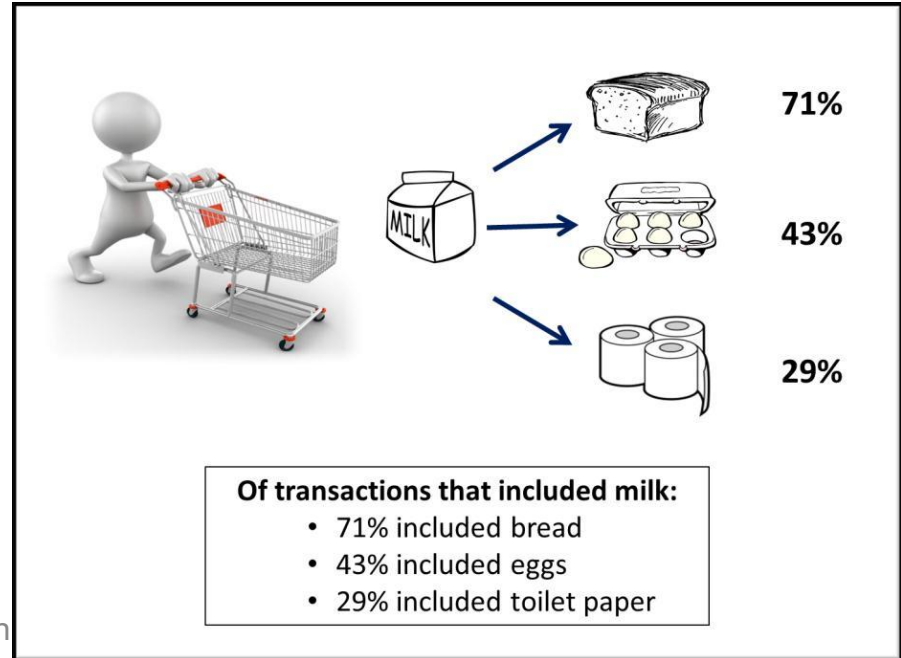
- Xây dựng các hệ máy tính có khả năng thích ứng và học từ kinh nghiệm
- Xây dựng và sử dụng các thuật toán trực cảm (heuristic algorithms)
- Nhấn mạnh các bài toán dự đoán

Khai phá dữ liệu

- Dữ liệu thường không có cấu trúc xác định và không rõ mục tiêu
- Phân tích khám phá
- Phối hợp lý thuyết toán và trực cảm

Ví dụ: Nghiên cứu thị trường (Market Research)

Phân tích giỏ hàng hoá (Market basket analysis)



Ví dụ: Vận chuyển và hậu cần

UPS: ORION (On-Road Integrated Optimization and Navigation)

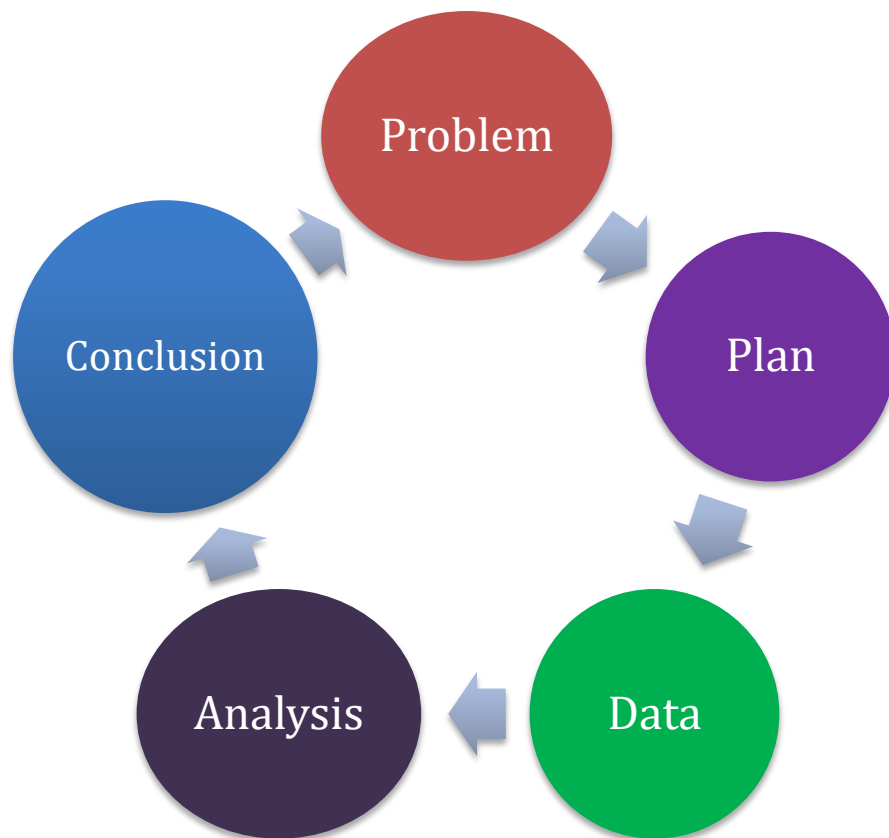
- Giảm quãng đường vận chuyển
- Tăng sự hài lòng của khách hàng
- Tiết kiệm xăng
- Giảm khí thải ra môi trường
- ...



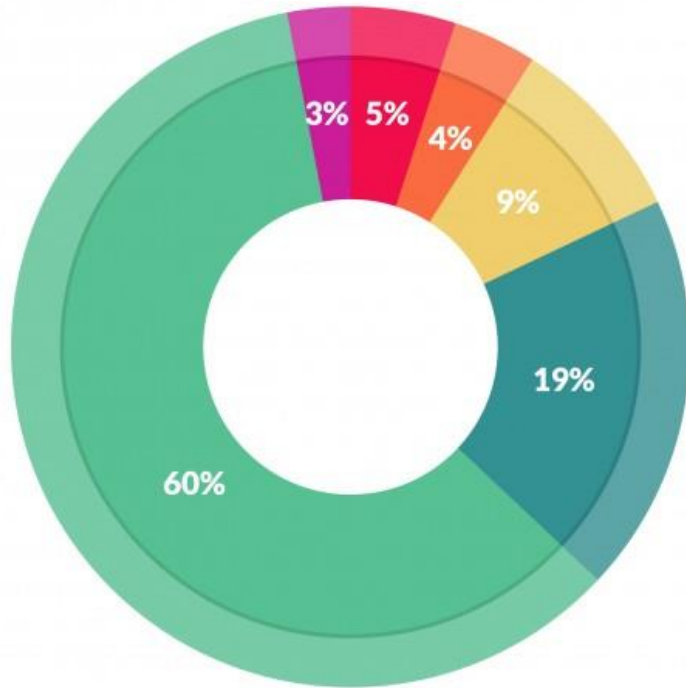
Case-study: Tín hiệu chứng khoán



Chọn những cổ phiếu nào
để đầu tư?



Not really as fun as you expect though

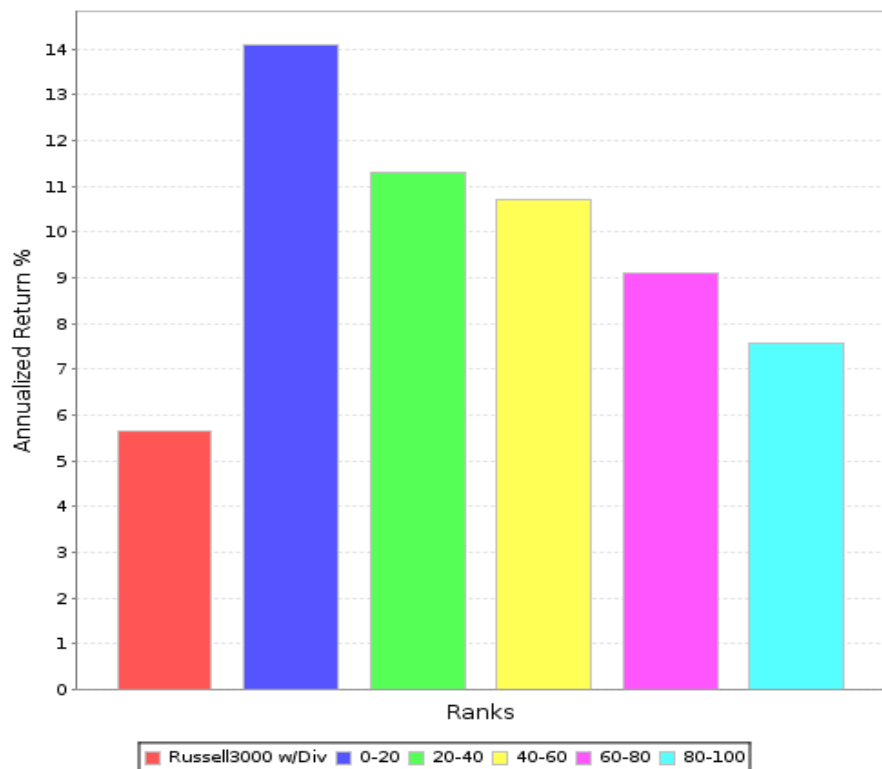


What data scientists spend the most time doing

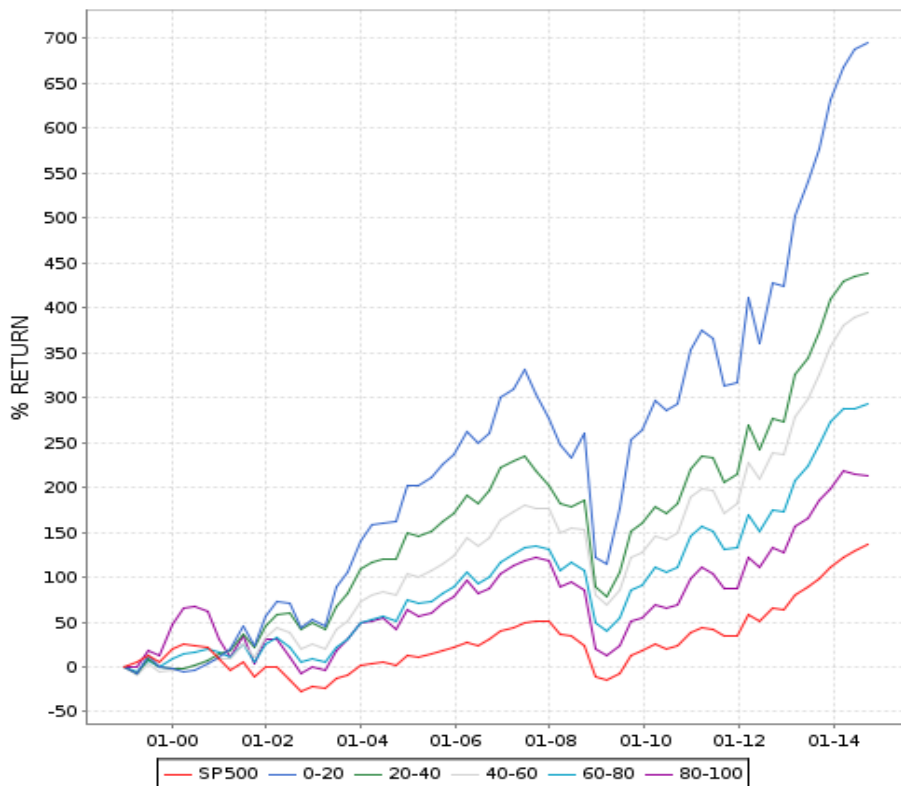
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Signal: P/E ratio

©Portfolio123.com



©Portfolio123.com

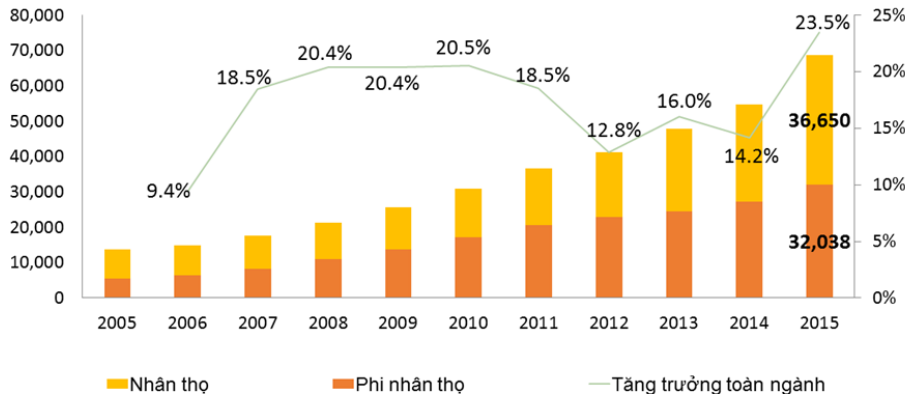


SẢN PHẨM CỦA DATA ANALYTICS

Hiện diện ở nhiều nơi hơn bạn nghĩ

Từ kinh doanh...

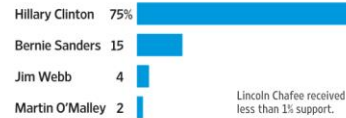
Doanh thu phí toàn ngành bảo hiểm Việt Nam (tỷ VNĐ)



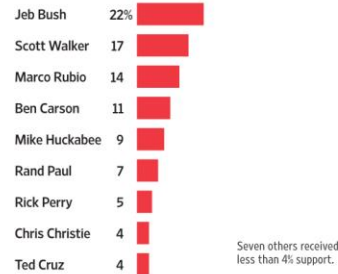
Vietdata

Findings From the Latest WSJ/NBC News Poll

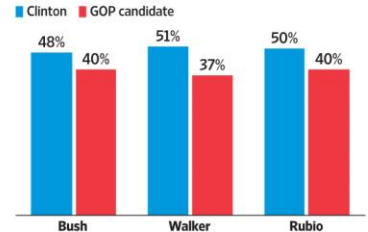
Presidential preference among **Democratic** primary voters



Presidential preference among **Republican** primary voters



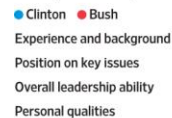
General-election preferences among registered voters



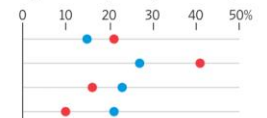
How voters say they would feel if each candidate were elected president



Among those who would feel **satisfied or optimistic** about each, the top reason why:



Among those who would feel **uncertain or pessimistic**, the top reason why:



*Results for 1996-2004 among likely voters; other years among registered voters.
Source: WSJ/NBC News polls

THE WALL STREET JOURNAL.

... đến giải trí

Bundled in the inbox



Housing



Trips



Saved



Purchases



Finance



Social



Updates



Promos

Up next

Autoplay



GIAI ĐIỆU TỰ HÀO THÁNG 6 BẮN FULL

GIAI ĐIỆU TỰ HÀO VTV
5,025 views **NEW**



TÌNH CA TÂY BẮC | GIAI ĐIỆU TỰ HÀO 2016 | 06/2016

GIAI ĐIỆU TỰ HÀO VTV
12,111 views **NEW**



CUỘC ĐỜI VẺ ĐẸP SAO | GIAI ĐIỆU TỰ HÀO 2016 | 06/2016

GIAI ĐIỆU TỰ HÀO VTV
5,533 views **NEW**



BƯỚC CHÂN TRÊN DẢI TRƯỜNG SƠN | GIAI ĐIỆU TỰ HÀO 2016 |

GIAI ĐIỆU TỰ HÀO VTV
4,128 views **NEW**

SUGGESTED PAGES

[See All](#)

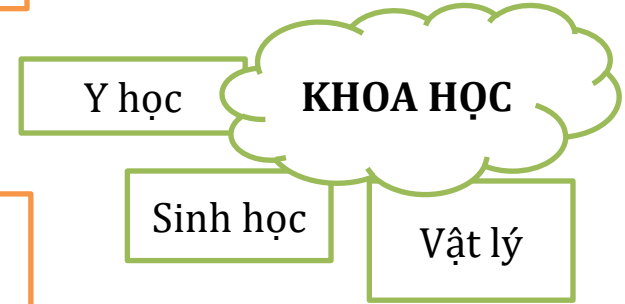
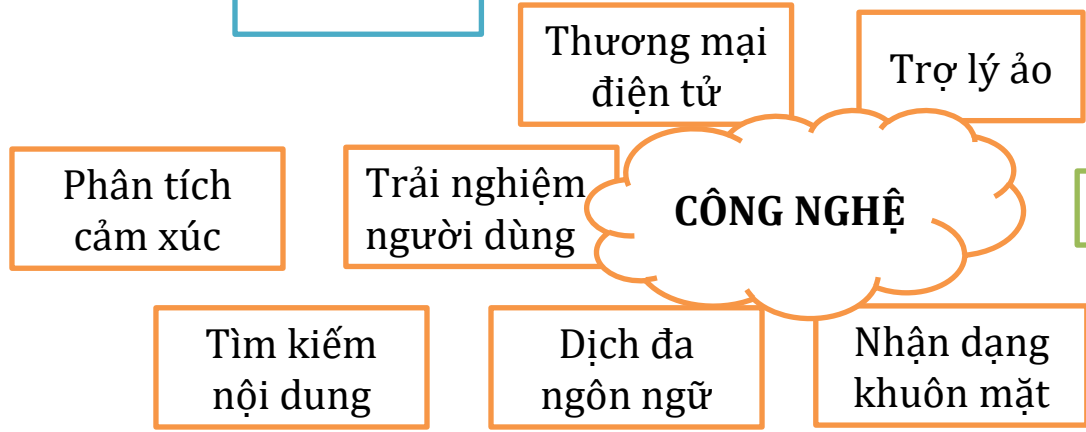
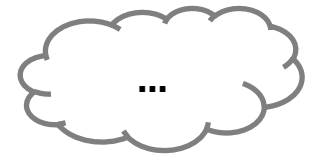
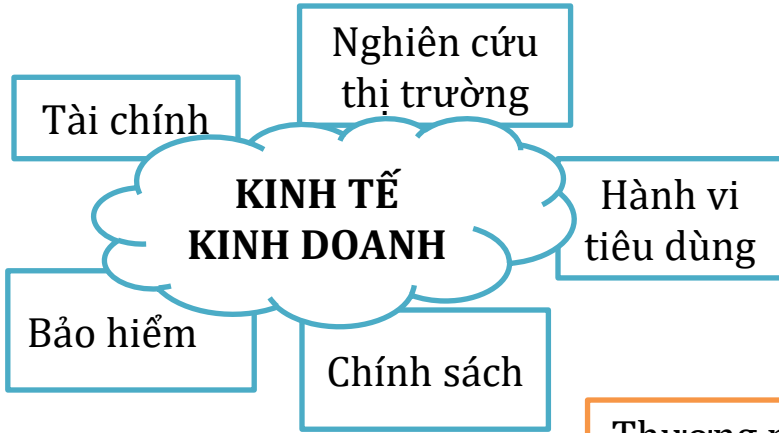


blogradio.vn

Arts/Humanities Website · 361,982 likes

Ha and 26 other friends like this.

Like Page



Câu hỏi: How?

Trả lời: Data come in many forms

	fx 1					
	B	C	D	E	F	G
D	Product	Category	Amount	Date	Country	
1	Carrots	Vegetables	\$4,270	1/6/2012	United States	
2	Broccoli	Vegetables	\$8,239	1/7/2012	United Kingdom	
3	Banana	Fruit	\$617	1/8/2012	United States	
4	Banana	Fruit	\$8,384	1/10/2012	Canada	
5	Beans	Vegetables	\$2,626	1/10/2012	Germany	
6	Orange	Fruit	\$3,610	1/11/2012	United States	
7	Broccoli	Vegetables	\$9,062	1/11/2012	Australia	
8	Banana	Fruit	\$6,906	1/16/2012	New Zealand	
9	Apple	Fruit	\$2,417	1/16/2012	France	
10	Apple	Fruit	\$7,421	1/16/2012	Canada	

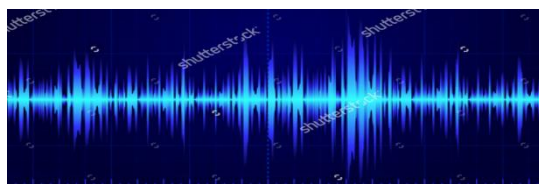
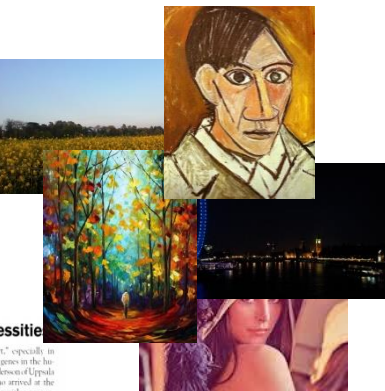
Dữ liệu không chỉ gói gọn dưới dạng bảng số liệu...

Order ID	Product	Category	Amount	Date	Country
1	Carrots	Vegetables	\$4,270	1/6/2012	United States
2	Broccoli	Vegetables	\$8,239	1/7/2012	United Kingdom
3	Banana	Fruit	\$617	1/8/2012	United States
4	Banana	Fruit	\$8,584	1/10/2012	Canada
5	Beans	Vegetables	\$2,626	1/10/2012	Germany
6	Orange	Fruit	\$3,610	1/11/2012	United States
7	Broccoli	Vegetables	\$9,062	1/11/2012	Australia
8	Banana	Fruit	\$6,906	1/16/2012	New Zealand
9	Apple	Fruit	\$2,417	1/16/2012	France

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Dr. Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus number may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing



Sometimes as a father, you ARE the only solution. A real honor making this true story.

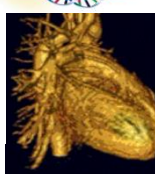
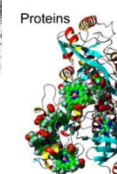
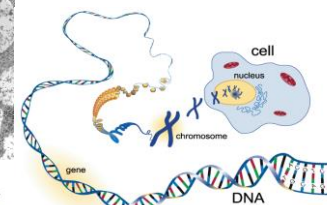
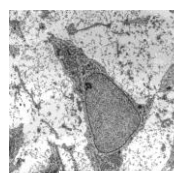
#SNITCH 2/22/13 pic.twitter.com/aJHoF6dt



at Garden is a district in London on the eastern side of the River Thames, between St. Martin's Lane and St. James's Park. It is a historic district, and the site is steeped in history. The district is home to many of the city's most famous landmarks, including the Royal Opera House, the British Museum, and the National Gallery. The district is also home to many of the city's most famous buildings, including the Houses of Parliament and the Palace of Westminster.

at Garden is a district in London on the eastern side of the River Thames, between St. Martin's Lane and St. James's Park. It is a historic district, and the site is steeped in history. The district is home to many of the city's most famous landmarks, including the Royal Opera House, the British Museum, and the National Gallery. The district is also home to many of the city's most famous buildings, including the Houses of Parliament and the Palace of Westminster.

at Garden is a district in London on the eastern side of the River Thames, between St. Martin's Lane and St. James's Park. It is a historic district, and the site is steeped in history. The district is home to many of the city's most famous landmarks, including the Royal Opera House, the British Museum, and the National Gallery. The district is also home to many of the city's most famous buildings, including the Houses of Parliament and the Palace of Westminster.

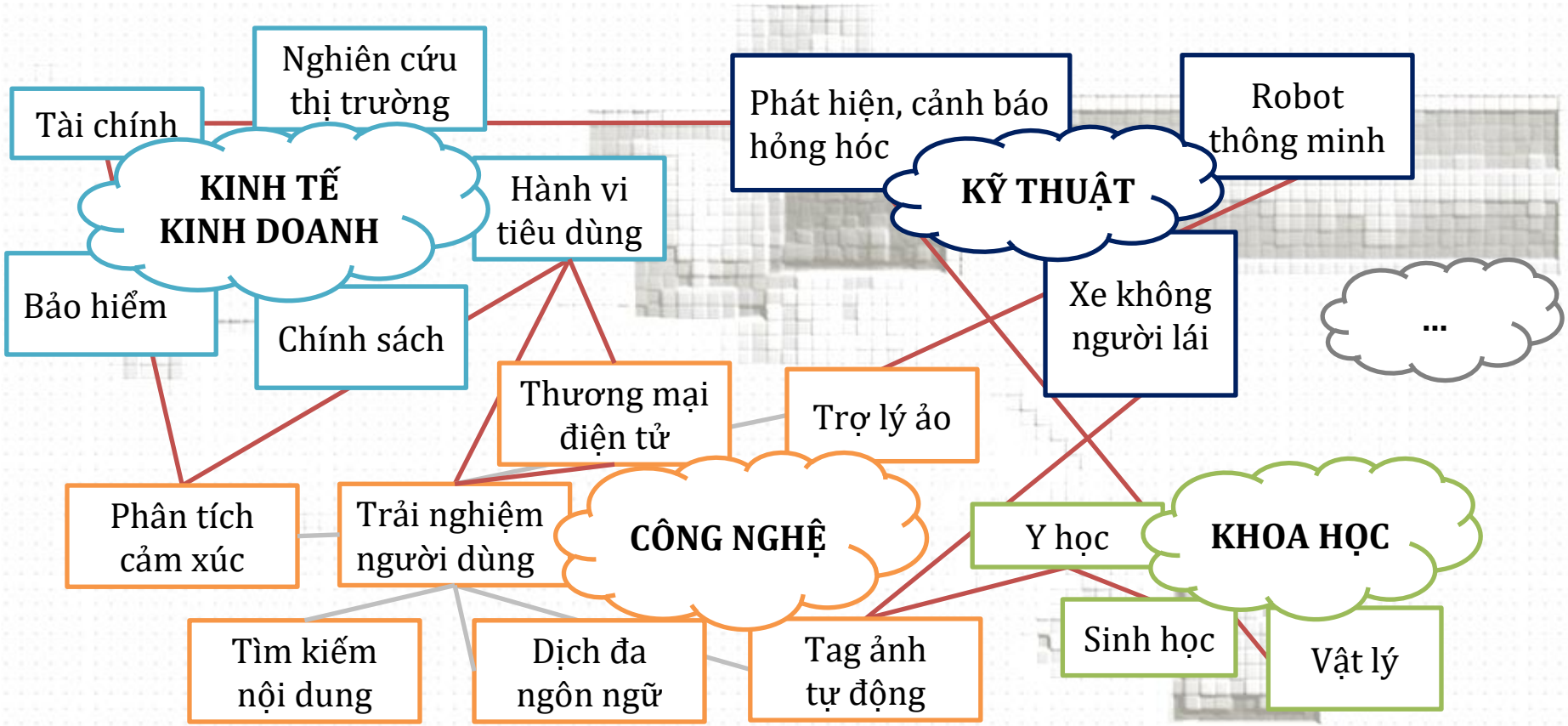


... mà còn tồn tại ở RẤT nhiều hình thái khác

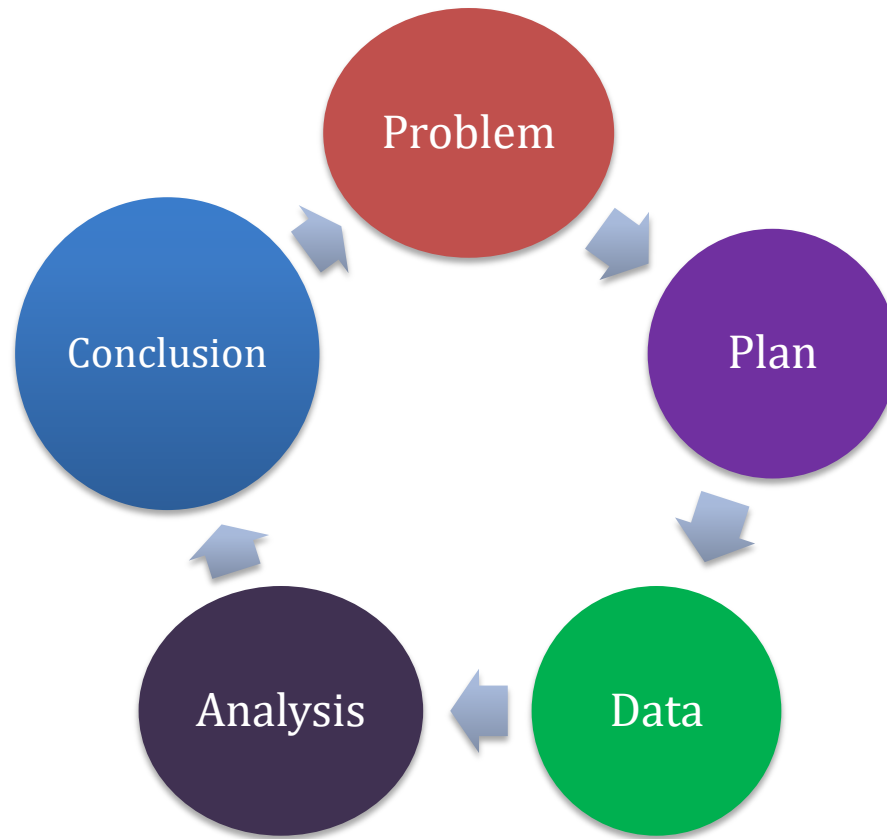
① Số lượng

② Công sức, thời gian thu thập (*)

[illegible]



CASE-STUDY: HỆ THỐNG GỢI Ý VIDEO (YOUTUBE, NETFLIX)



Problem

YouTube VN

Search

Recommended

A man in a black t-shirt with 'techstyle' on it is speaking. The text 'SLI GOT ME LIKE' is overlaid in large, bold letters.

WTF is going on with SLI?
LinusTechTips
515,971 views • 2 days ago

A close-up of a woman's face with her hair styled in two large, round buns.

Actors Who Almost Ruined Their Careers
Looper
619,307 views • 2 days ago

A colorful parrot is shown next to a diagram of a neural network.

Two Minute Papers - Hallucinating Images With Deep Learning
Károly Zsolnai-Fehér
7,746 views • 3 weeks ago

Silhouettes of people standing in a line, looking at a large question mark.

Why no aquarium has a great white shark
Vox
4,231,539 views • 4 days ago

A scene from the game Batman: Arkham Knight showing the Batmobile.

BATMAN: ARKHAM KNIGHT (Honest Game Trailers)
Smosh Games
2,816,385 views • 1 year ago

A hand holding a pen pointing to a line graph showing housing costs over time.

Is Renting Always A Waste Of Money?
Preet Banerjee
1,276,123 views • 1 year ago

A man in a Star Wars costume is sitting in a chair, looking at a screen.

The WAN Show - The Last Show Before We Leave - May 6, 2016
LinusTechTips
175,831 views • 2 months ago

A character from Assassin's Creed 4 is shown in a dynamic pose.

ASSASSIN'S CREED 4 (Honest Game Trailers)
Smosh Games
7,153,343 views • 2 years ago

A Dell XPS 15 laptop is shown from a top-down perspective.

DELL XPS 15 REVIEW
LinusTechTips
697,708 views • 4 months ago

A man in a blue cap and apron is smiling.

Ant-Man "Scott Lang" Funniest Scenes
SuperMarvel
2,309,575 views • 5 months ago

A person is sitting at a desk with a large gaming setup.

Project Backpack: Mobile 200" Gaming Setup
LinusTechTips
399,638 views • 3 days ago

A diagram showing two satellites in space connected by a line, with the text 'Gravitational Waves Explained' and 'Using Stick Figures'.

Gravitational Waves Explained
minutephysics
343,542 views • 4 days ago

Plan

- Dữ liệu: rất lớn và “thô”
- Kỹ thuật, phương pháp
- Hạ tầng công nghệ

Data

Dữ liệu liên quan trực tiếp:

- video mà người dùng thích/không thích/cho điểm đánh giá

Meta-data:

- Nội dung video:
 - Tags có sẵn/Bio
 - Tags từ quá trình *phân tích tự động video*
- User-logs – ghi chép hoạt động của người dùng:
 - Thông tin cá nhân: độ tuổi, giới tính, ... (*); vị trí địa lý
 - Video đã xem, thích/không thích/điểm đánh giá
 - Thời điểm xem
 - Thiết bị được sử dụng
 - ...
 - *Bản đồ nhiệt của trỏ chuột*

③ Độ “thô”

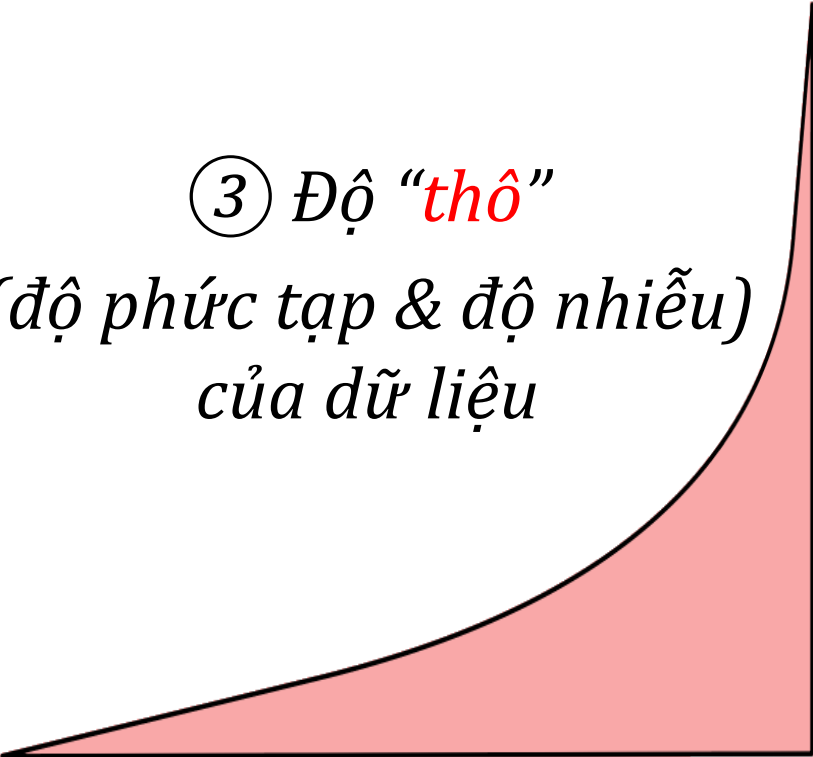
(độ phức tạp & độ nhiều)
của dữ liệu

A2	B	C	D	E	F
1	Order ID	Product	Category	Amount	Date
2	A2	B	C	D	E
3	1	Carrots	Vegetables	\$4,270	1/6/2012
4	2	Broccoli	Vegetables	\$9,129	1/7/2012
5	3	Banana	Fruit	\$6,817	1/8/2012
6	4	Banana	Fruit	\$5,384	1/9/2012
7	5	Beans	Vegetables	\$2,626	1/10/2012
8	6	Orange	Fruit	\$3,610	1/11/2012
9	7	Broccoli	Vegetables	\$9,062	1/12/2012
10	8	Banana	Fruit	\$6,906	1/13/2012
11	9	Apple	Fruit	\$2,417	1/14/2012
12	10	Apple	Fruit	\$2,417	1/15/2012



③ Độ “thô”

(độ phức tạp & độ nhiều)
của dữ liệu



Order ID	Product	Category	Amount	Date	Country
1	Carrots	Vegetables	\$4,270	1/6/2012	United States
2	Broccoli	Vegetables	\$8,229	1/7/2012	United Kingdom
3	Banana	Fruit	\$617	1/8/2012	United States
4	Banana	Fruit	\$8,384	1/10/2012	Canada
5	Beans	Vegetables	\$2,626	1/10/2012	Germany
6	Orange	Fruit	\$5,602	1/11/2012	United States
7	Broccoli	Vegetables	\$9,062	1/11/2012	Australia
8	Banana	Fruit	\$6,900	1/16/2012	New Zealand
9	Apple	Fruit	\$2,417	1/16/2012	France

Số lượng và Độ “thô” lớn đòi hỏi:

- *Kỹ thuật – Phương pháp* phân tích có tính tự động hóa & độ chính xác cao
- *Hạ tầng công nghệ* phù hợp để thu thập, lưu trữ, xử lý dữ liệu cơ bản (đọc/ghi, truy vấn, sắp xếp thứ tự, ...)

Kỹ thuật, Phương pháp

cần **mô hình/thuật toán** xây dựng trên các giả sử phức tạp hơn để:

- (i) tiền xử lý, làm sạch loại dữ liệu cần sử dụng;
- (ii) giải quyết tốt bài toán cụ thể trên tập dữ liệu cụ thể; và
- (iii) áp dụng được trên nhiều tập dữ liệu/bài toán khác nhau

Hạ tầng công nghệ

Hạ tầng công nghệ cổ điển (RDB – SQL) không phù hợp với dữ liệu “thô”

⇒ sự phát triển của công nghệ *(Xử lý) Dữ liệu lớn* – Big Data, *Điện toán đám mây* – Cloud computing

⇒ sự phát triển tương hỗ của công nghệ *Kết nối mọi vật* – Internet of Things

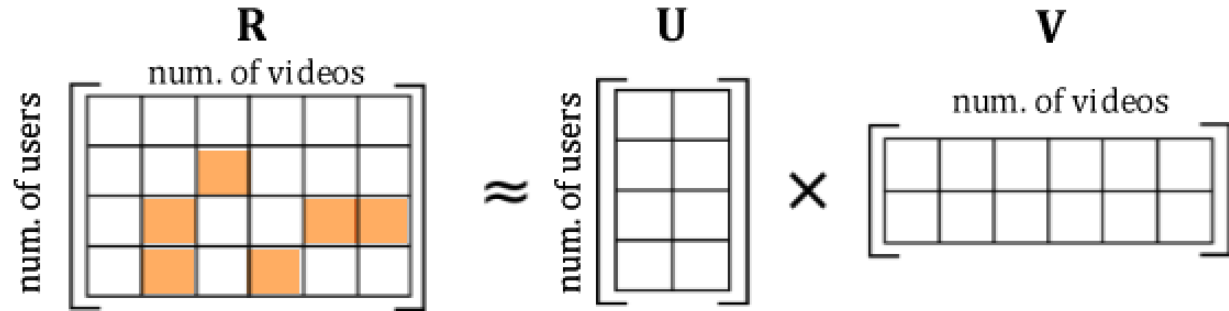
Plan

(cont.)

- **Dữ liệu**
 - Số người dùng & số video \Rightarrow rất lớn và “thô”
- **Kỹ thuật, phương pháp**
 - [...]
 - Constraint: Thời gian đáp ứng nhanh
 - \Rightarrow pipeline kết hợp nhiều thuật toán/mô hình để phù hợp cho phân tích online/offline
 - Kiểm định, đánh giá \Rightarrow nhiều hơn 1 thước đo: độ “chính xác”, thời gian xử lý, mức độ tiêu tốn tài nguyên, ...
 - Văn hóa, Pháp luật \Rightarrow lọc nội dung không phù hợp
- **Hạ tầng công nghệ: [...]**

Analysis

- Tiền xử lý & Biểu diễn dữ liệu về 1 dạng chung: số
- Xây dựng mô hình: *Matrix factorization*



- Huấn luyện mô hình (giải các tham số chưa biết)
- Kiểm định, đánh giá để cải thiện mô hình dựa vào các chỉ số đo trên các tập dữ liệu khác nhau

- Đưa ra Top-30 video có điểm *dự đoán* cao nhất để gợi ý cho *từng* người dùng
- Thu nhận thêm dữ liệu từ phản hồi của người dùng để cải thiện (tự động) mô hình hiện tại

Conclusion

- Theo dõi hoạt động của hệ thống trong một khoảng thời gian để có đánh giá chân thực cho mục tiêu kinh doanh ban đầu
- Thử nghiệm cải thiện mới? Thông tin mới? Tính năng mới? Hệ thống mới? ... ⇒ **tiếp tục vòng quay**

DATA ANALYTICS

```
graph TD; A([DATA ANALYTICS]) --> B[KỸ THUẬT/ PHƯƠNG PHÁP]; A --> C[TRIỂN KHAI]; A --> D[CHUYÊN MÔN TRONG LĨNH VỰC ỨNG DỤNG];
```

KỸ THUẬT/ PHƯƠNG PHÁP

- Thống kê (Statistics)
- Machine Learning
- Khai phá dữ liệu (Data Mining)

TRIỂN KHAI

Computing power:

- Sử dụng Excel, SPSS, ...
- Lập trình (Programming, Engineering)
- Hạ tầng (Infrastructure)

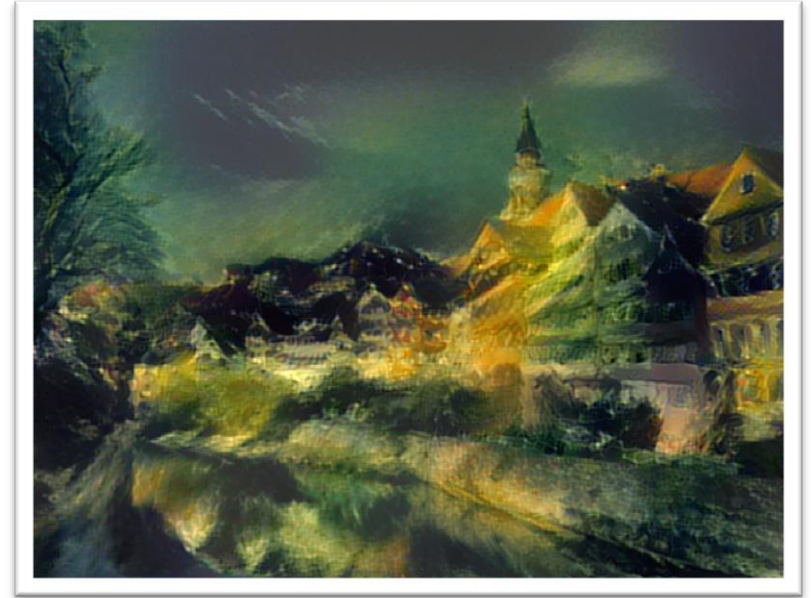
CHUYÊN MÔN TRONG LĨNH VỰC ỨNG DỤNG

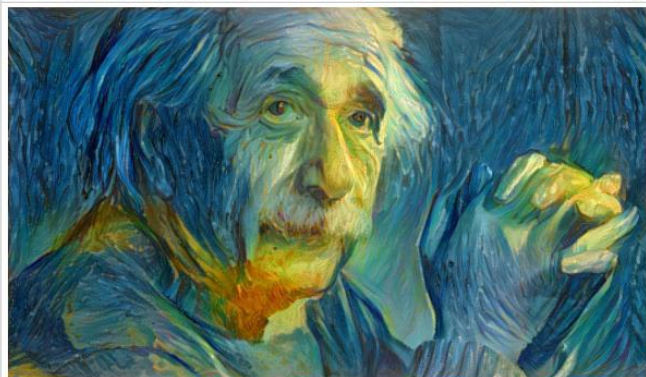
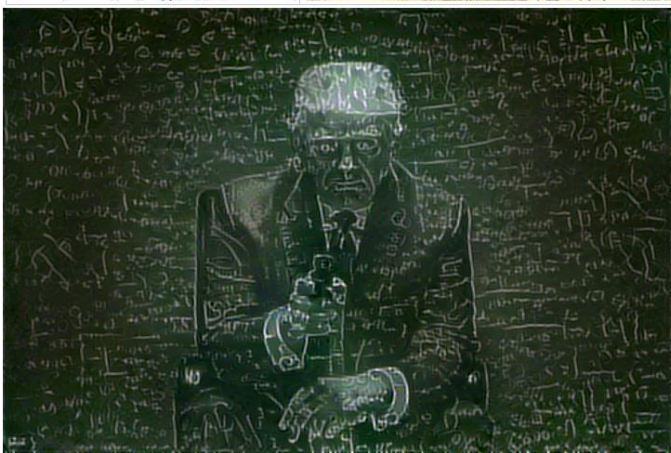
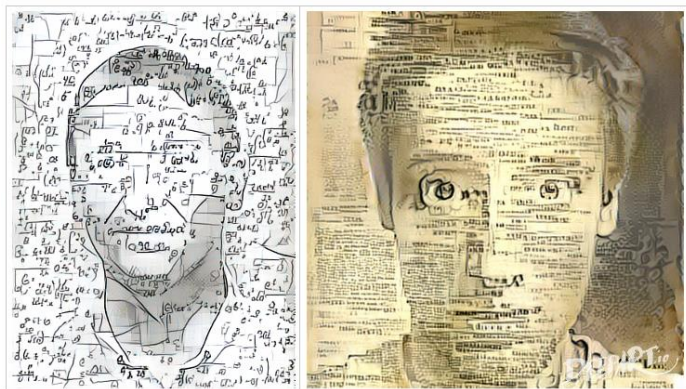
- Hiểu - Diễn giải kết quả
- Báo cáo, trình bày
- Đưa ra quyết định
- Đặt câu hỏi cho vấn đề mới

DA trong Nghệ thuật: “Neural” Art

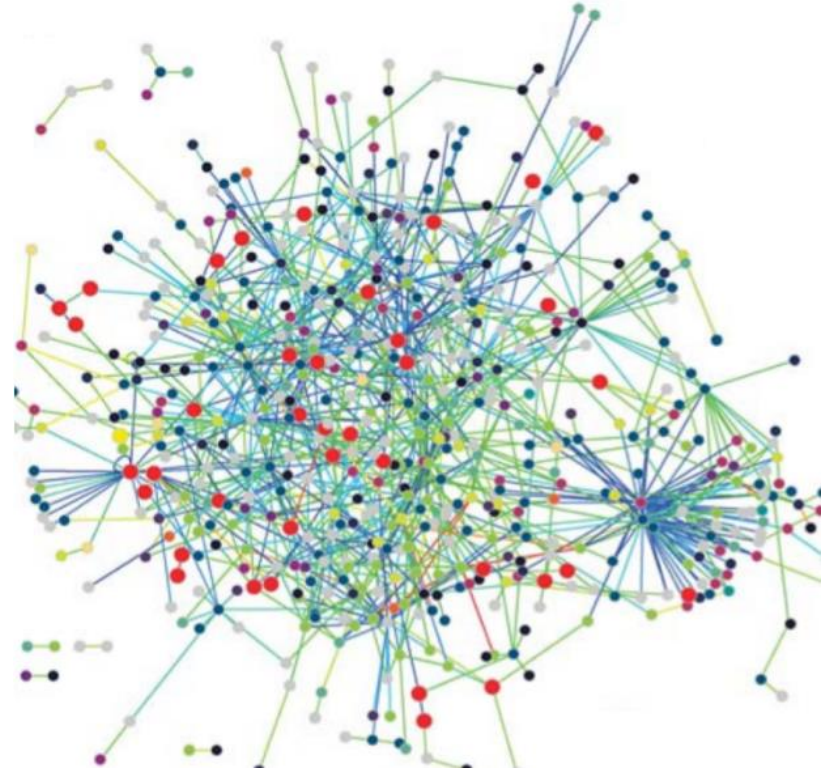
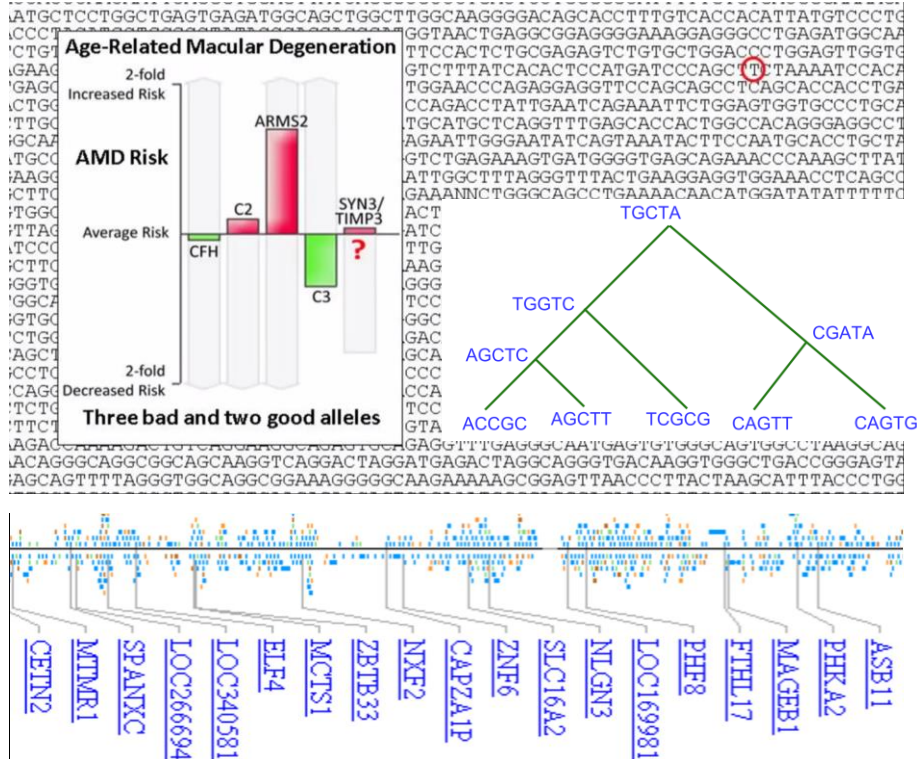


DA trong Nghệ thuật: “Neural” Art





DA trong Y-Sinh: Chẩn trị bệnh tật



Take-home messages

DATA ANALYTICS

*Tạo ra sản phẩm có giá trị
từ dữ liệu*

KỸ THUẬT/PHƯƠNG PHÁP

Thống kê
Machine Learning
Khai phá dữ liệu

TRIỂN KHAI

Computing skills
Kỹ năng đòi hỏi phức thuộc
vào tính chất vị trí công việc

LĨNH VỰC CHUYÊN MÔN

Kiến thức chuyên môn là nhân
tố quan trọng để nắm bắt nhu
cầu & sáng tạo sản phẩm

[Slides phụ]

THE BIG DATA

Ngày càng nhiều công ty sử dụng Dữ liệu lớn (Big data) để phân tích kinh doanh, khiến nhu cầu nhân lực ngành này bùng nổ.

QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU

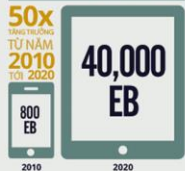


ĐỊNH NGHĨA

BIG DATA

Mô tả các gói dữ liệu quá lớn, quá phức tạp mà không thể xử lý và phân tích theo phương pháp truyền thống

XU THẾ PHÁT TRIỂN



1 exabyte (EB) = 1,000,000 TB
Source: IDC's Digital Intelligence Study, sponsored by EMC, December 2013

CƠ HỘI NGHỀ NGHIỆP

5 KHU VỰC TẠI HOA KỲ



XU HƯỚNG CÁC CÔNG TY LỚN



1.5 TRIỆU

chuyên viên phân tích và quản lý dữ liệu sẽ cần thiết phải bổ sung trong 5 năm tới
Source: "Big Data: The New Frontier for Innovation, Competition, and Productivity" McKinsey Global Institute, May 2011

TĂNG LỢI THẾ CẠNH TRANH

NHỮNG NGÀNH CẦN NHU CẦU NHÂN SỰ BIG DATA CAO NHẤT

- | | |
|----------------------------------|--------------------------------|
| 1. Ngành SQL | 6. Quản lý kho dữ liệu |
| 2. Phát triển KI thông minh (BI) | 7. Quản lý quy trình nghiệp vụ |
| 3. Phân mềm PTHD | 8. Quản lý dữ liệu |
| 4. Phân tích dữ liệu | 9. Người EL trình cuối |
| 5. Phân tích kinh doanh | 10. Thiết kế mô hình dữ liệu |
- Source: Strategy Analytics International report of job postings for business and government big data in the USA analysis field during 2012

"Nghề hấp dẫn nhất trong 10 năm tới là thống kê"

— HAL VARIAN,
Kinh tế trưởng tại Google

CÁC TẬP ĐOÀN LỚN

CÓ NHU CẦU TUYỂN DỤNG CÁN BỘ PHÂN TÍCH DỮ LIỆU

1. Deloitte
2. Capital One
3. IBM
4. Booz Allen Hamilton
5. Northrop Grumman
6. SAIC
7. CGI Group
8. General Dynamics
9. CAI
10. Freddie Mac

Source: Strategy Analytics International report of job postings for business and government big data analysis field during 2012

MỨC LƯƠNG HẤP DẪN

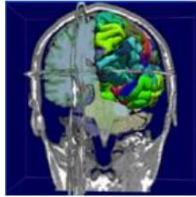


UMUC
University of Maryland
University College

<http://vnexpress.net/infographics/quoc-te/nhu-cau-nhan-luc-khong-lo-cho-big-data-3244918.html>

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

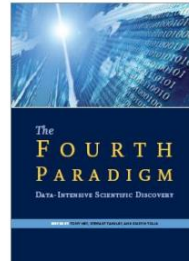


Printed materials in the Library of
Congress = 10 TeraBytes



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

Một lược đồ phân tích dữ liệu lớn

