

Nhập Môn Trí Tuệ Nhân Tạo

Khoa Học Dữ Liệu

Anh-Tuan Nguyen
tuan.nguyenanh2@phenikaa-uni.edu.vn

Ví dụ

Xác định khách hàng tiềm năng?

Gợi ý sản phẩm cho khách hàng?

Tăng cường hiệu quả của quảng cáo?

Hệ thống mở cửa tự động?

Hệ thống tự động trả lời khách hàng?

Tối ưu hóa quá trình vận tải hàng hoá?

Dự đoán giá chứng khoán?

Tự động dịch văn bản?

....

What is data?

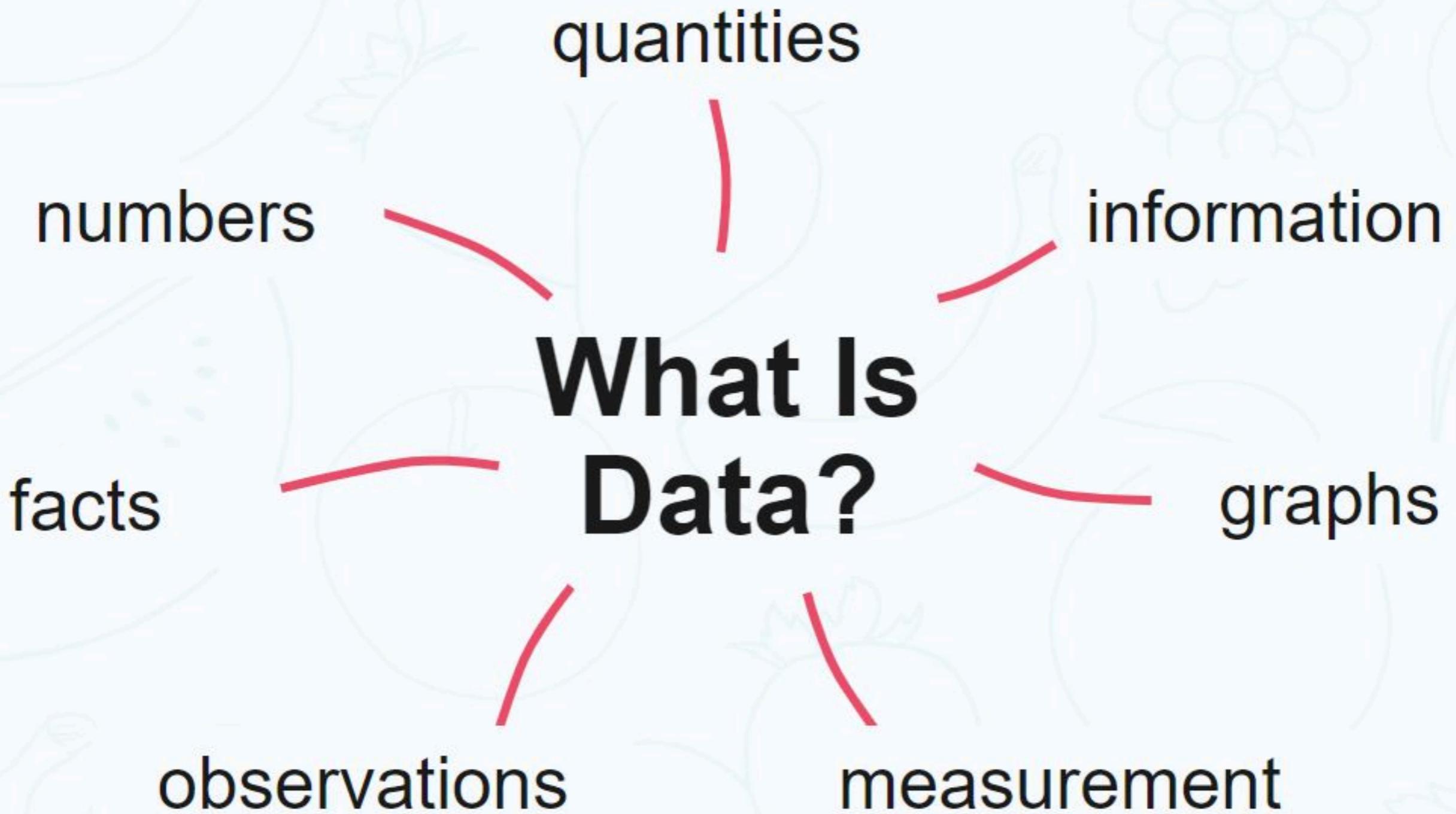


What you created while using your phone

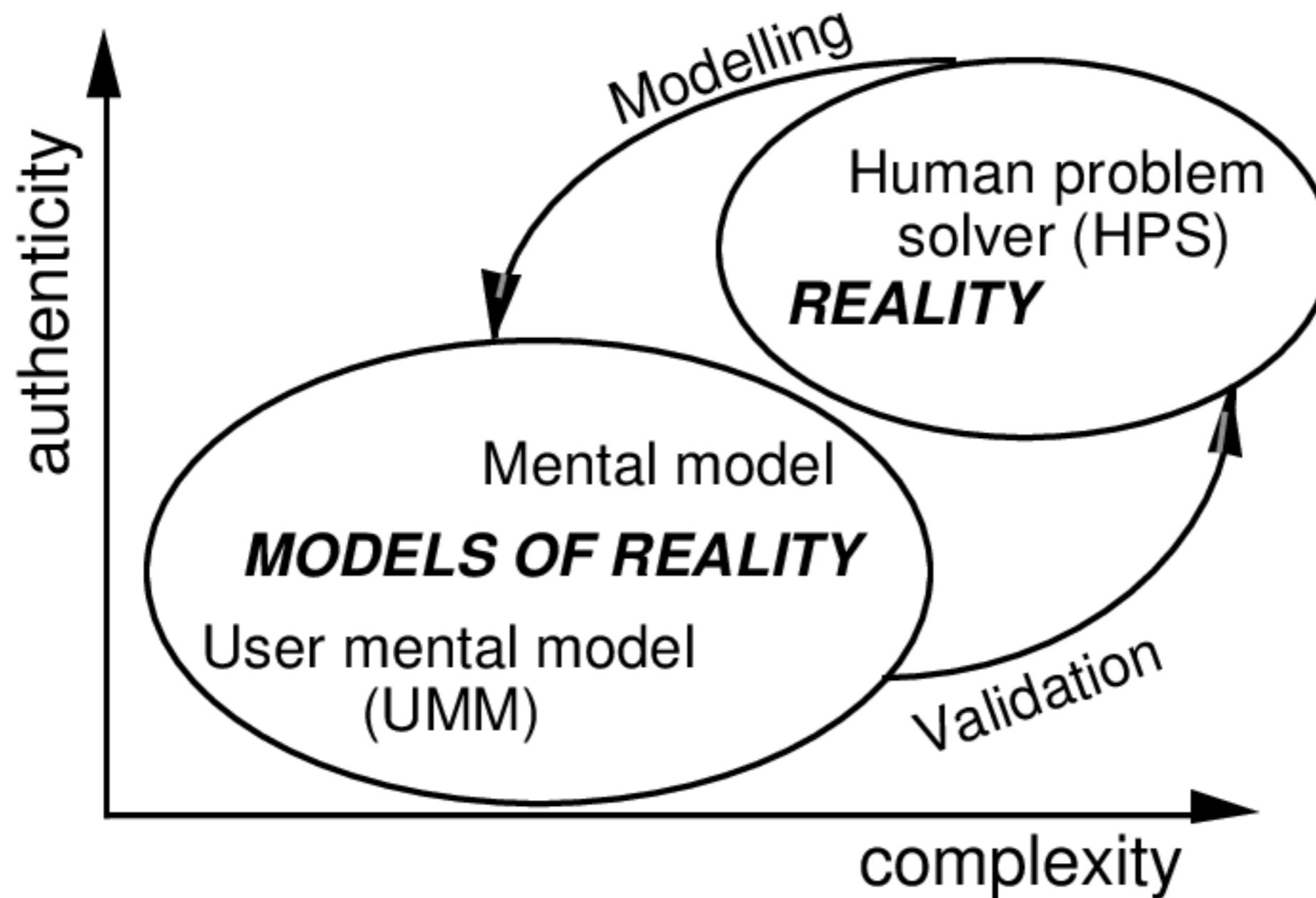
- Make a calls
- Capture photos
- Text message
- Browse internet
- Reading books
- Learning
-

All of your data are collected

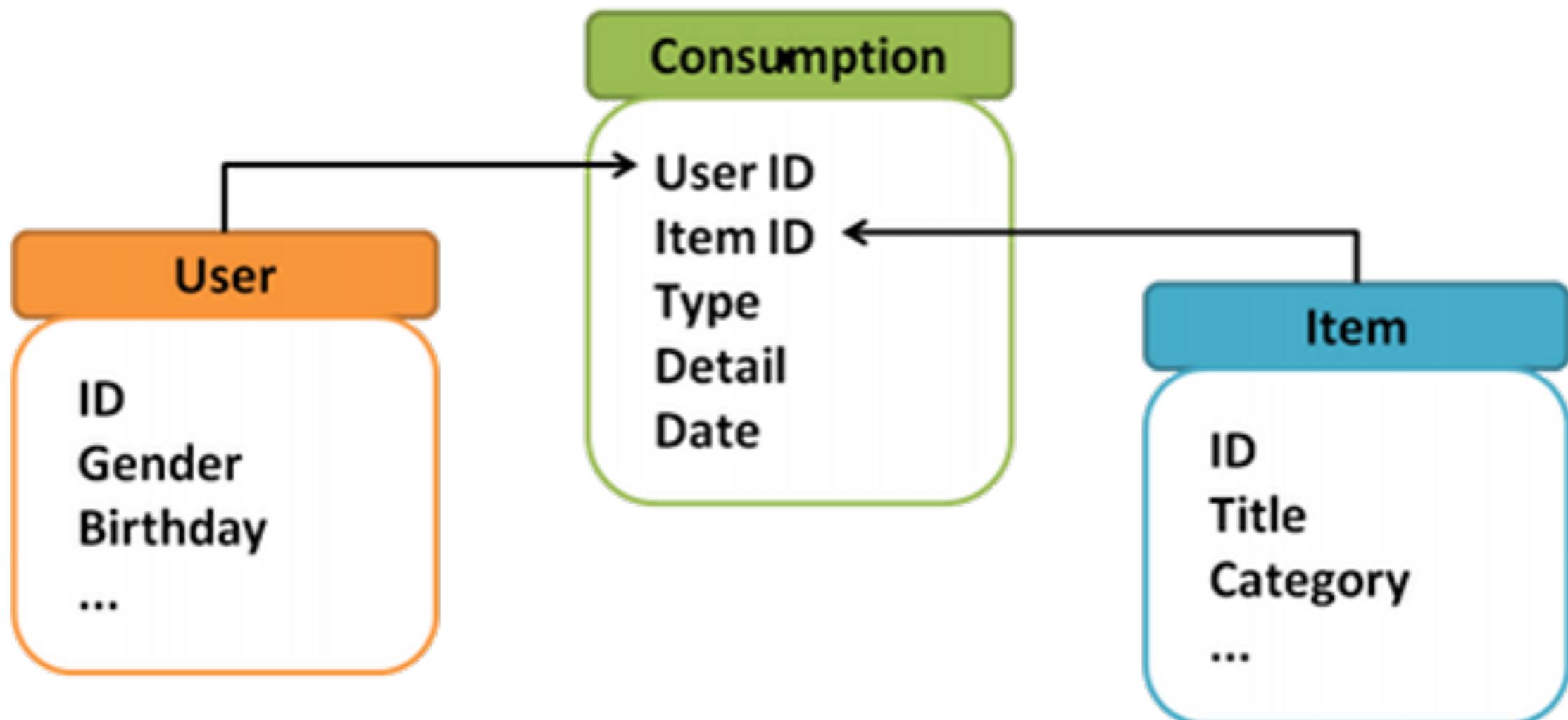
What is data?



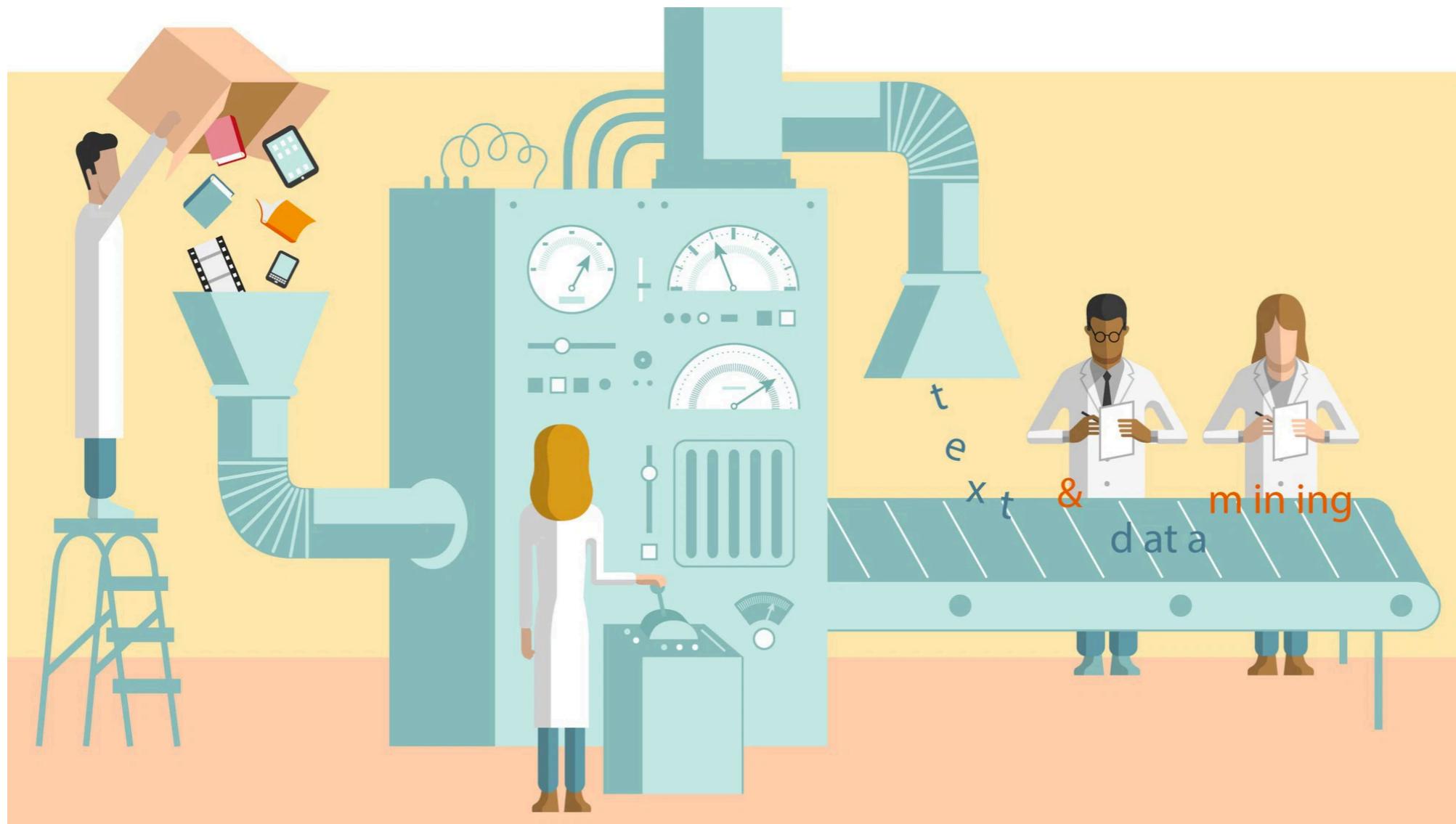
Data as models of reality



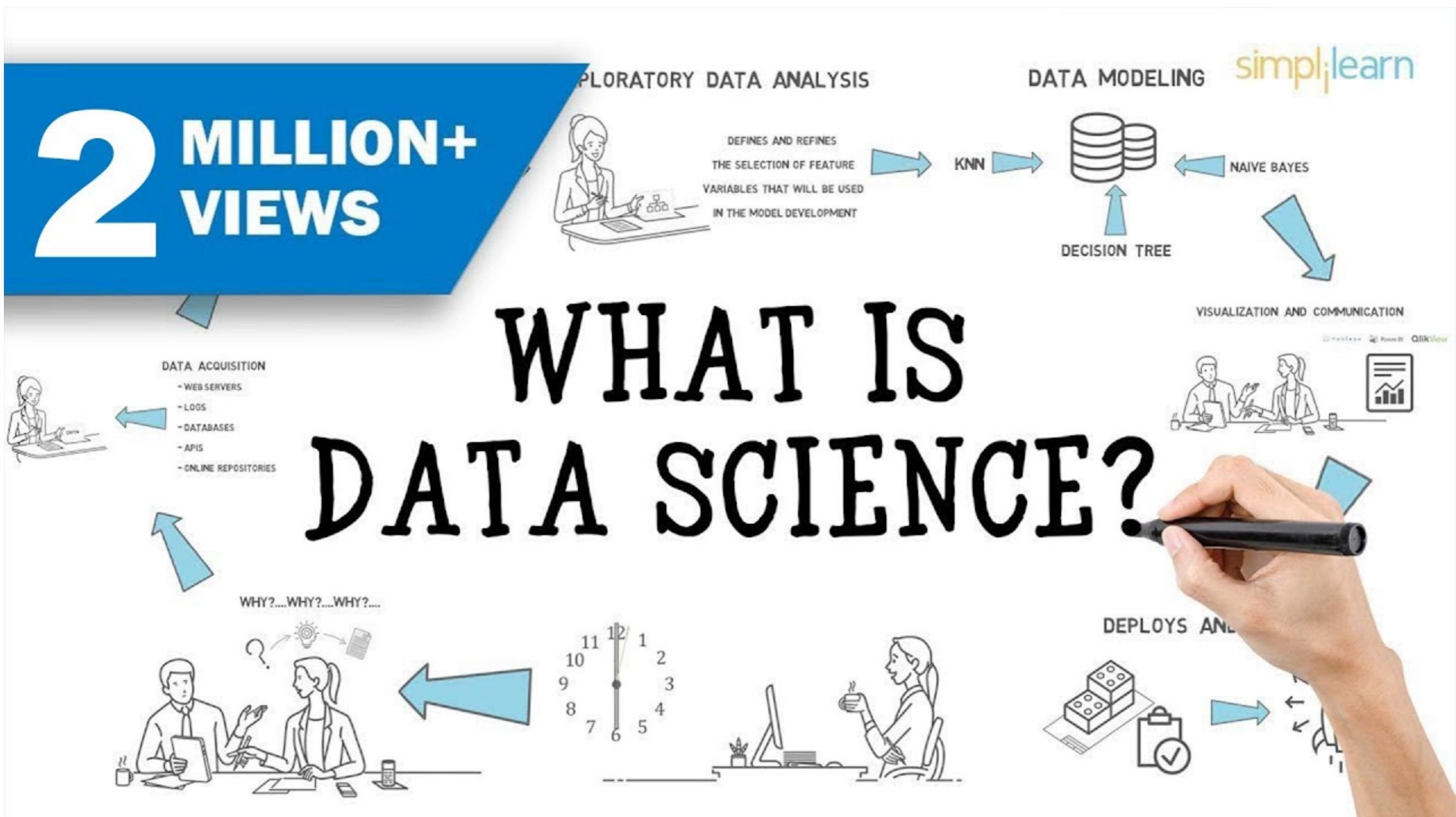
Data models



Data mining



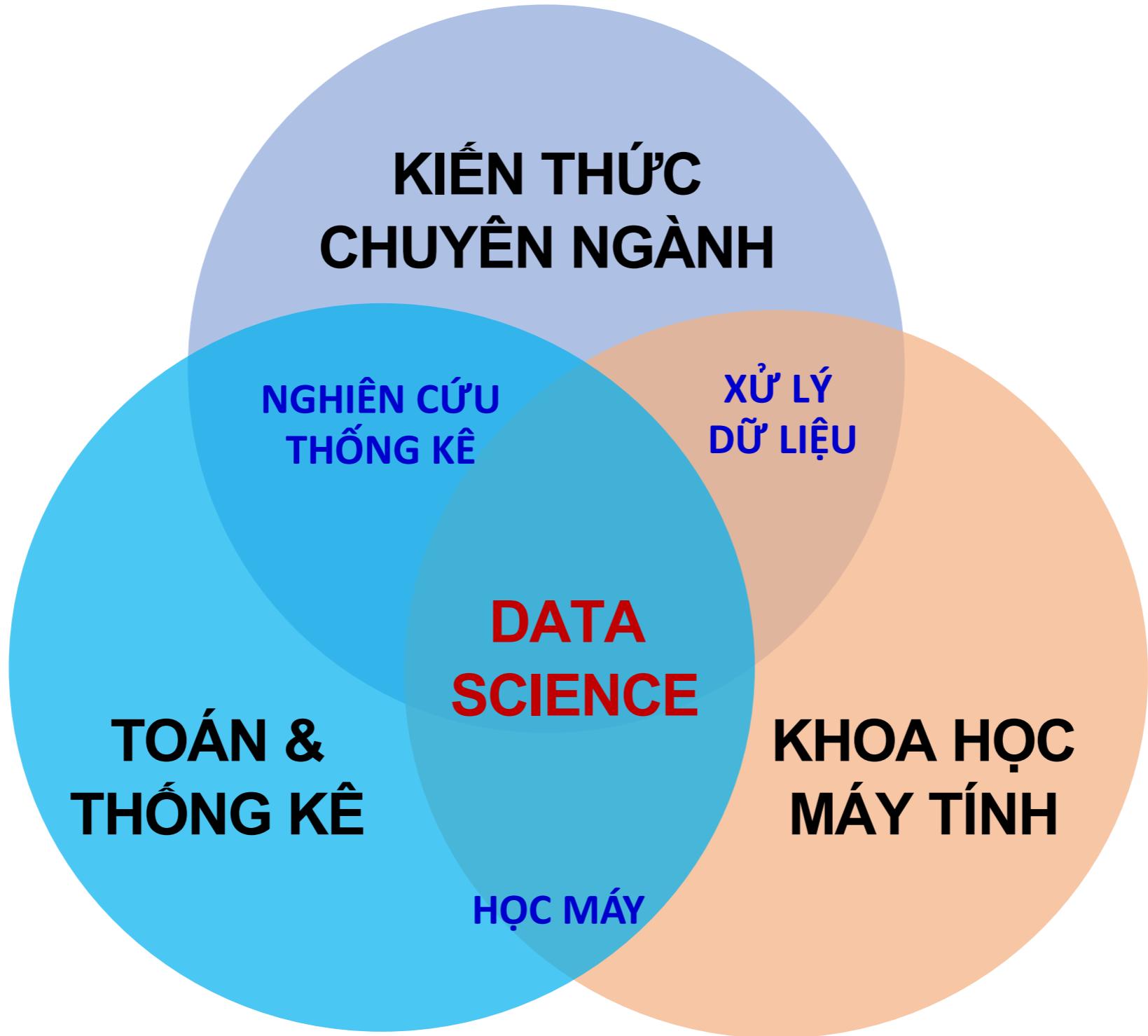
Data science?



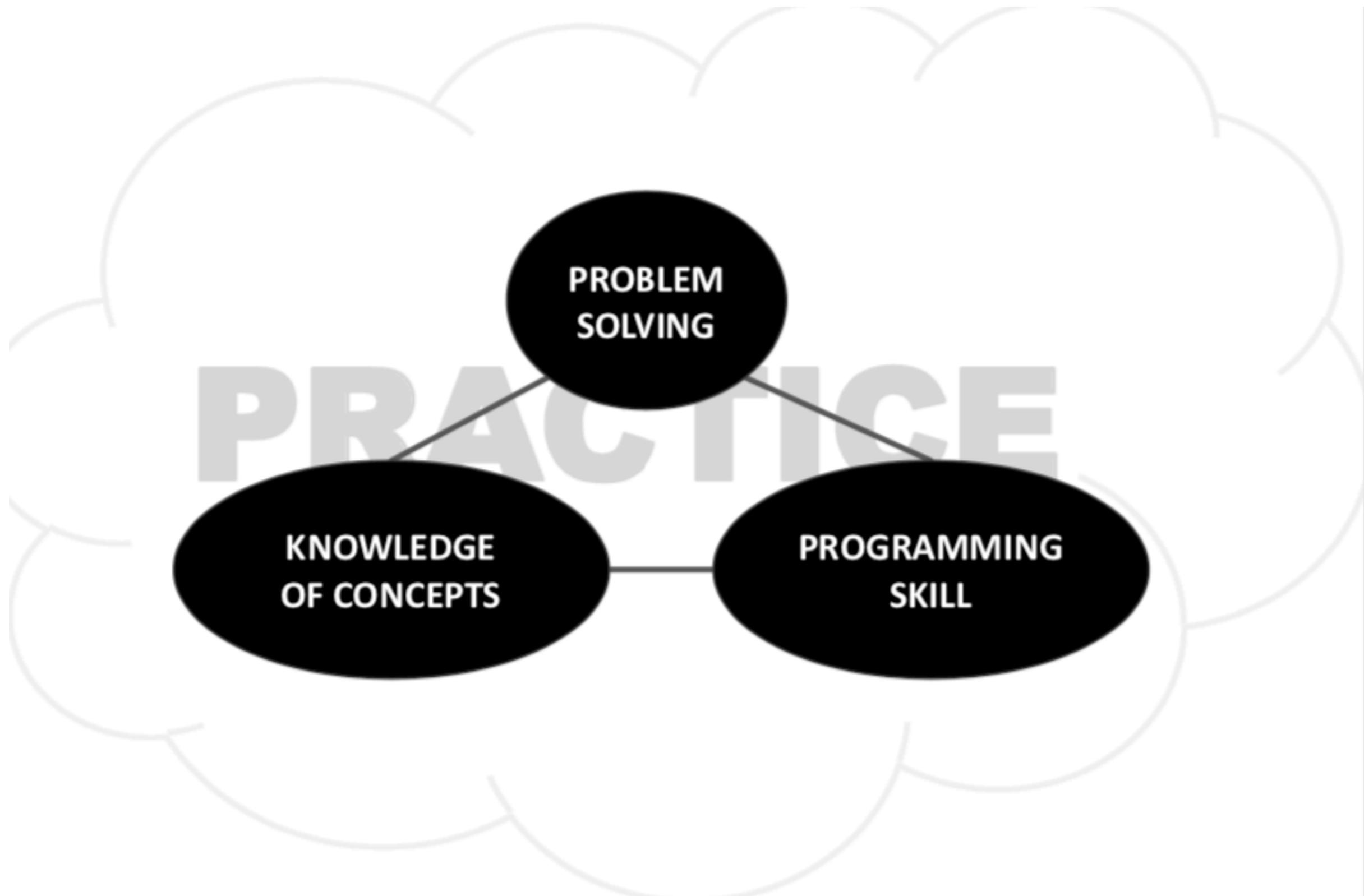
Data science

- Define your problems
- Data collection: web, databases, log, APIs, and others
- Data preparation (data model):
 - ✓ Data cleaning: missing data, inconsistency, duplications
 - ✓ Data transformations
- Exploratory analysis: in sight of data
- Modeling: statistics, machine learning
- Report
- Deploy and maintenance

Data Science



Leaning for your Future



Computer vs. Human

- **Computer:**

- ✓ Computing: billion of calculations / s: +, -, *, /, NOT, AND ..
- ✓ Remember: 100s GB
- ✓ Do what you tell it to do -> Programming -> instructions

- **Human:**

- ✓ Observe nature
- ✓ Imagine
- ✓ Learn
- ✓ Calculation?
- ✓ Remember important features?
- ✓ Interact with environment and action independently

You can be

- Data analyst
- Machine learning engineer
- Deep learning engineer
- Data engineer
- Data scientist

Data Driven World

3 Important Statistics About How Much Data Is Created Every Day



1 How much data is generated every minute?

 **41,666,667**

messages shared
by WhatsApp users

 **1,388,889**

video / voice calls made
by people worldwide

 **404,444**

hours of video streamed
by Netflix users

 **347,222**

stories posted by Instagram users

 **150,000**

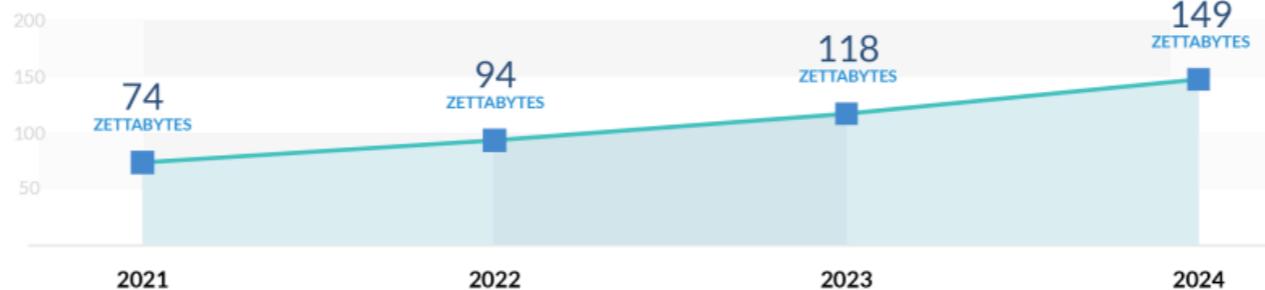
messages shared by Facebook users

 **147,000**

photos shared by Facebook users

Source: Domo

2 Estimated Data Consumption from 2021 to 2024



Source: IDC / Statista

3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

 **2 TRILLION**

searches on Google by the end of 2021

 **1.134 TRILLION MB**

volume of data created every day

 **3,026,626**

emails sent every second, 67% of which are spam

 **278,108 PETABYTES**

global IP data per month by the end of 2021

 **230,000**

new malware versions created every day

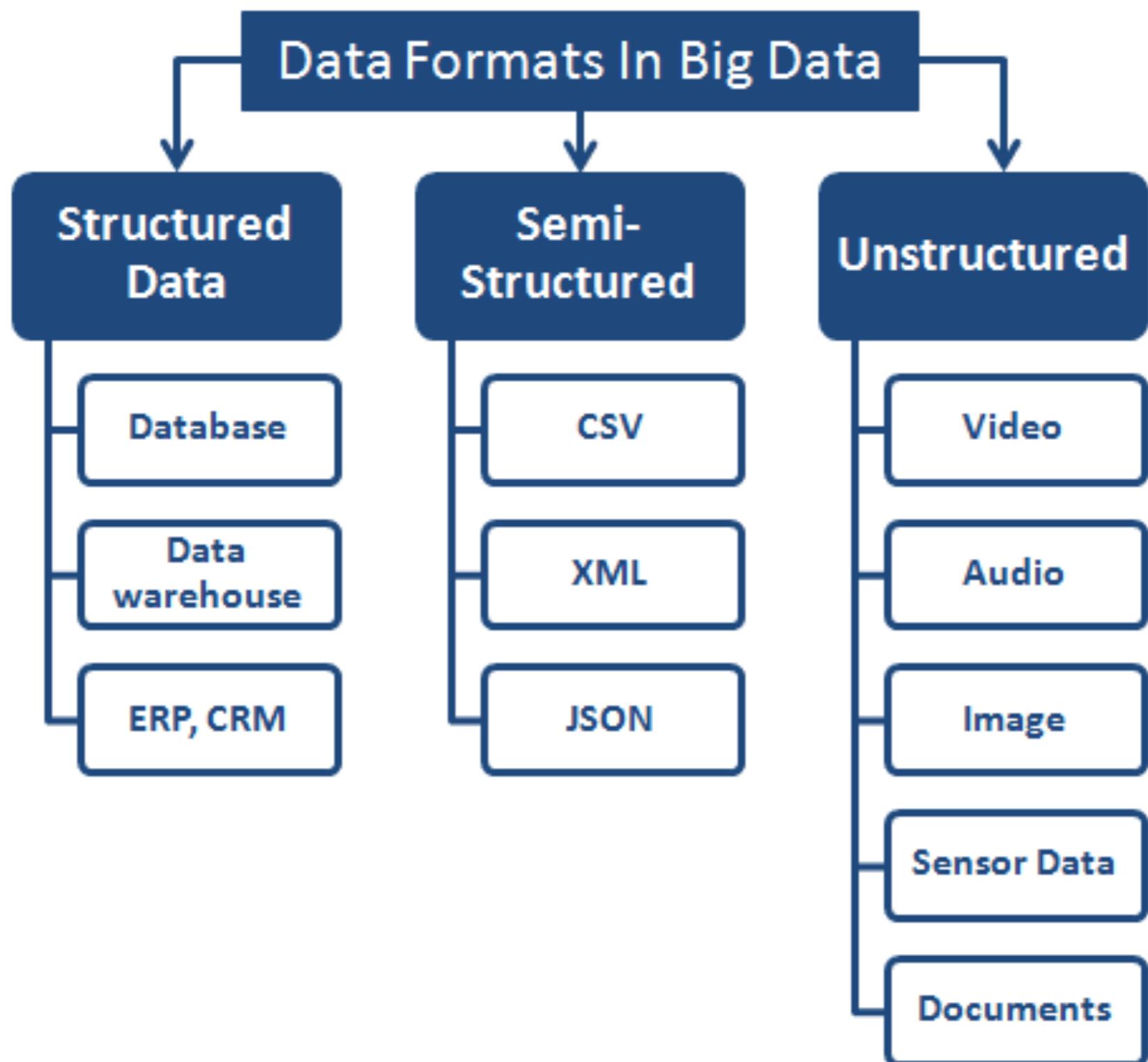
 **82%**

share of video in total global internet
traffic at the end of 2021

Dữ liệu lớn

| Abbreviation | Unit | Value | Size (in bytes) |
|--------------|-----------|--------------------------|---|
| b | bit | 0 or 1 | 1/8 of a byte |
| B | bytes | 8 bits | 1 byte |
| KB | kilobytes | 1,000 bytes | 1,000 bytes |
| MB | megabyte | 1,000 ² bytes | 1,000,000 bytes |
| GB | gigabyte | 1,000 ³ bytes | 1,000,000,000 bytes |
| TB | terabyte | 1,000 ⁴ bytes | 1,000,000,000,000 bytes |
| PB | petabyte | 1,000 ⁵ bytes | 1,000,000,000,000,000 bytes |
| EB | exabyte | 1,000 ⁶ bytes | 1,000,000,000,000,000,000 bytes |
| ZB | zettabyte | 1,000 ⁷ bytes | 1,000,000,000,000,000,000,000 bytes |
| YB | yottabyte | 1,000 ⁸ bytes | 1,000,000,000,000,000,000,000,000 bytes |

Các dạng dữ liệu



Cách mạng công nghiệp lần thứ 4



Chúng ta thu được gì từ dữ liệu

HOW 5 TECH GIANTS MAKE THEIR BILLIONS

Comparing the revenue streams of the five largest tech companies

Technology has taken over.

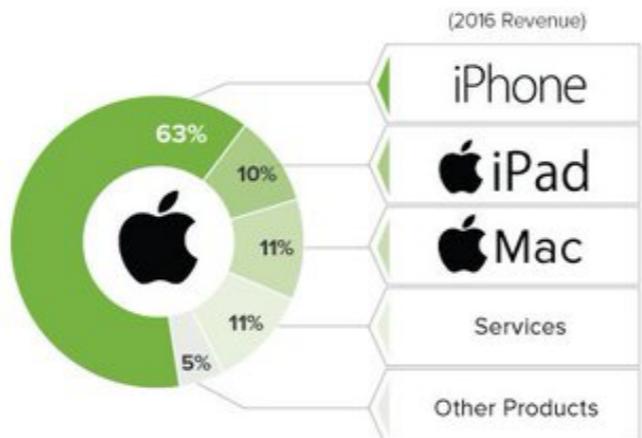
These five tech companies are the most valuable stocks in the U.S. market, worth a collective \$2.9 trillion in market capitalization.

In 2016, these companies combined for \$555 billion in revenue, and a \$94 billion bottom line.



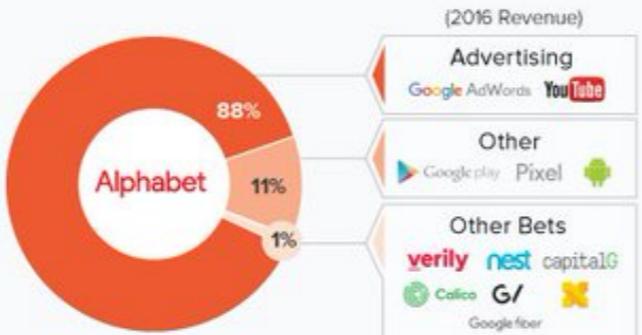
1 Apple

Market Cap \$804B
Revenue \$216B
Earnings \$46B



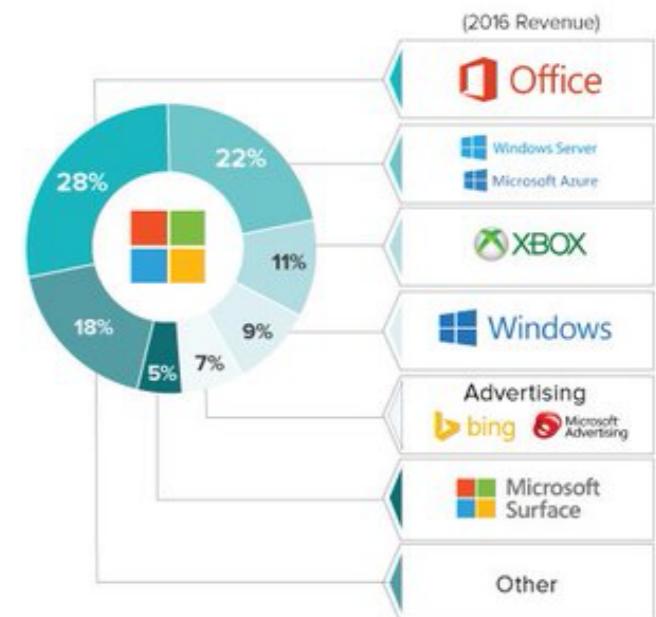
2 Alphabet

Market Cap \$651B
Revenue \$90B
Earnings \$19B



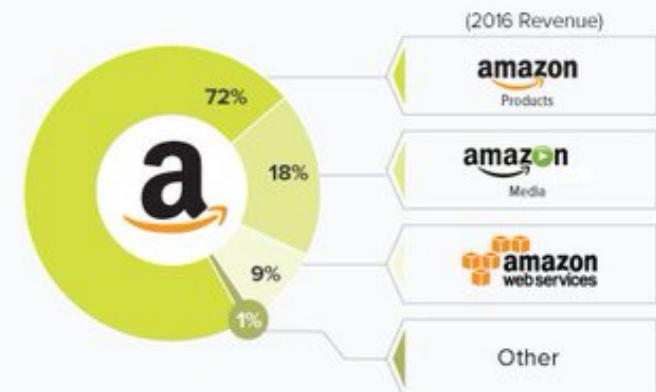
3 Microsoft

Market Cap \$536B
Revenue \$85B
Earnings \$17B



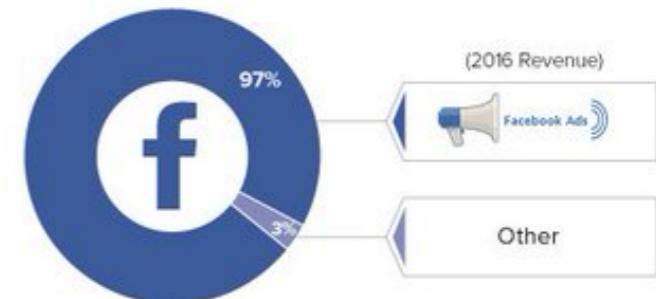
4 Amazon

Market Cap \$455B
Revenue \$136B
Earnings \$2B



5 Facebook

Market Cap \$434B
Revenue \$28B
Earnings \$10B



Việt Nam



- VinID: thu thập dữ liệu
- VinTech: nghiên cứu công nghệ cao và quỹ đầu tư (~500 FTE)
- VinSmart: thiết bị thông minh(mobile, TV, cameras, home???)
 - 51% của BQ ở Tây Ban Nha
 - 29.5% của Archos ở Pháp
 - Factory in Hoa-Lac High Tech park
- Khách hàng dựa trên những giá trị ngành dọc



- Telecoms: 3G/4G/5G, NB-IoT, ADSL/FTTH
- Giải pháp kinh doanh
- Trung tâm nghiên cứu Hòa Lạc
- VTTC: Open AI platform (120 FTE)
- Sản xuất
- Khách hàng toàn cầu
- Quỹ Viettel Startup



- Telecoms: ADSL/FTTH
- Tích hợp hệ thống
- Phát triển phần mềm
- FPT.AI (150 FTE)
- VISSA

Phenikaa?

Ứng dụng giáo dục



Ánh sáng thông minh
(trong nhà/ngoài trời)



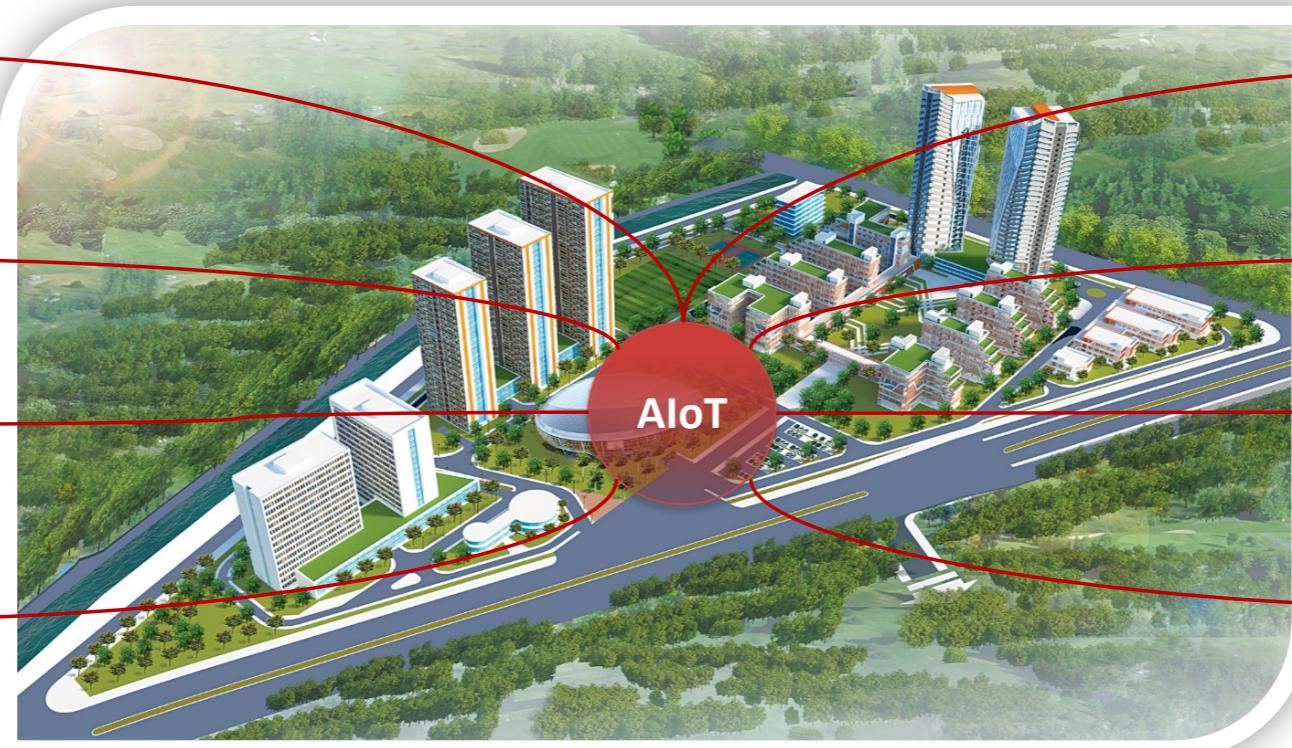
Bãi đỗ xe
thông minh



Nhà ăn
thông minh



Dữ liệu
sinh trắc học



Văn phòng
thông minh



Thư viện
thông minh

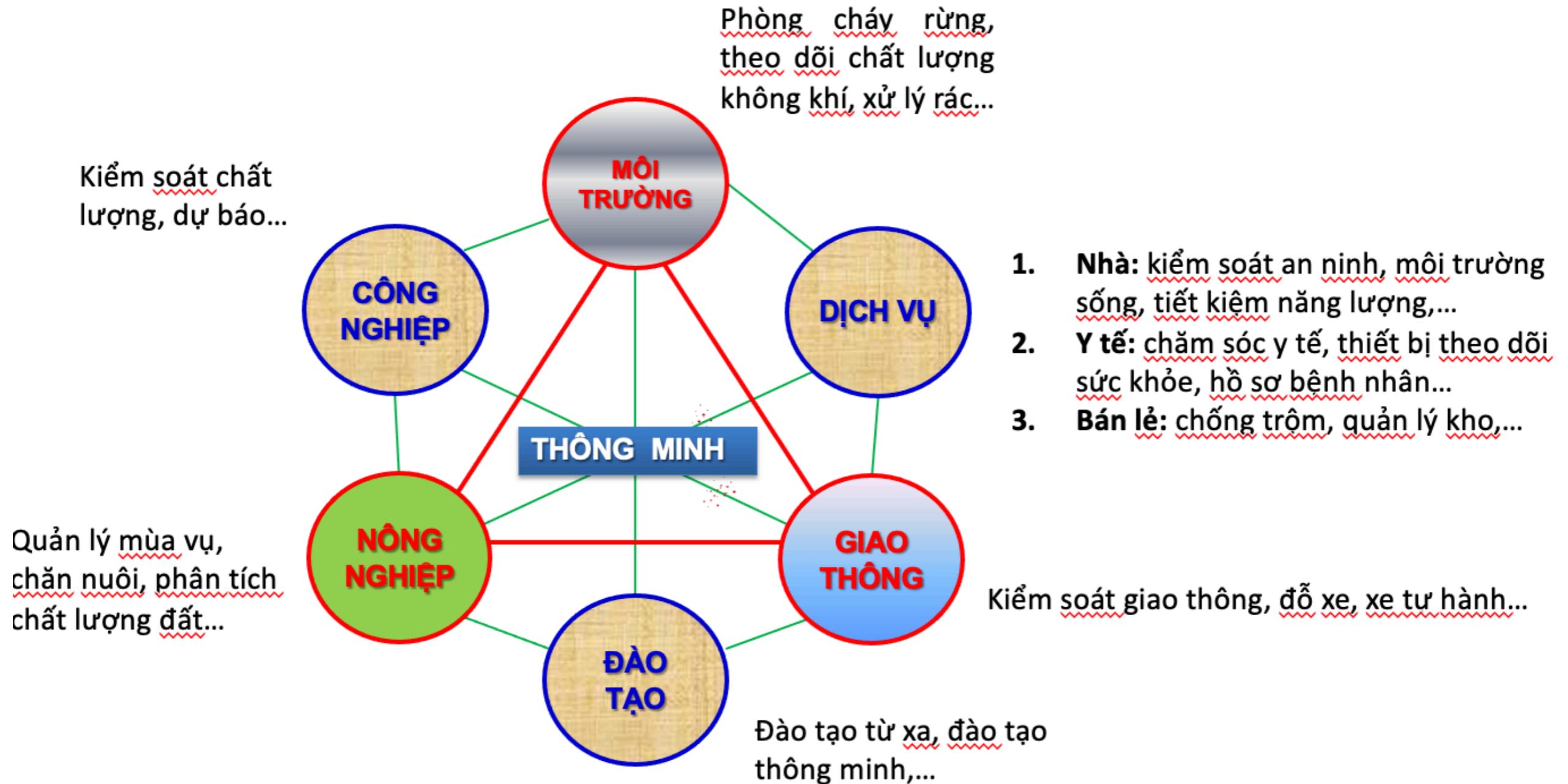


Xe tự hành

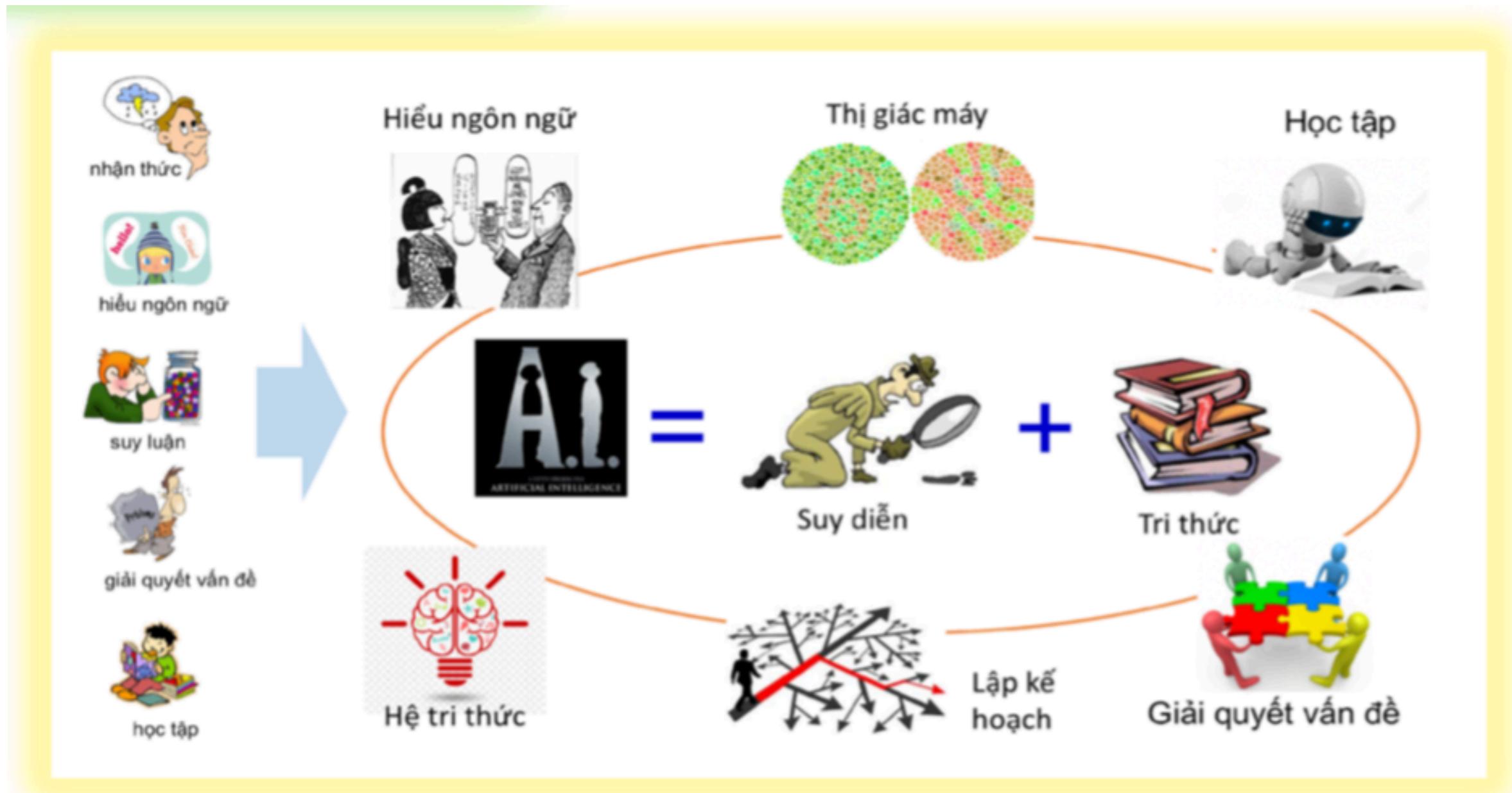
Hệ tầng kết nối



Dữ liệu lớn và nhu cầu của xã hội



Trí Tuệ Nhân Tạo (Artificial Intelligence)



A machine which can do like human?

Trí Tuệ Nhân Tạo (Artificial Intelligence)



Intelligent Machine

Give machine set of rules

Learning from data
(Machine learning)

**INPUT
Information
(X)**



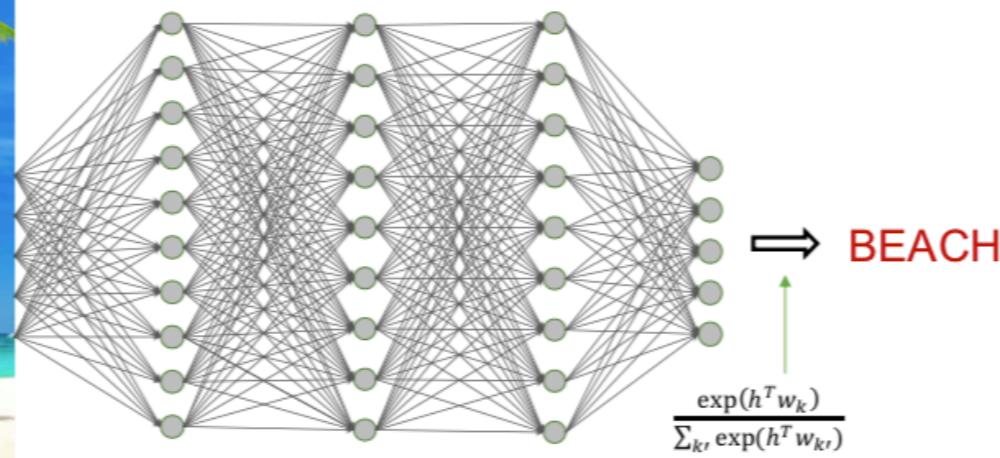
**OUTPUT
Prediction (y)
Action (0, 1)**

Intelligent Machine

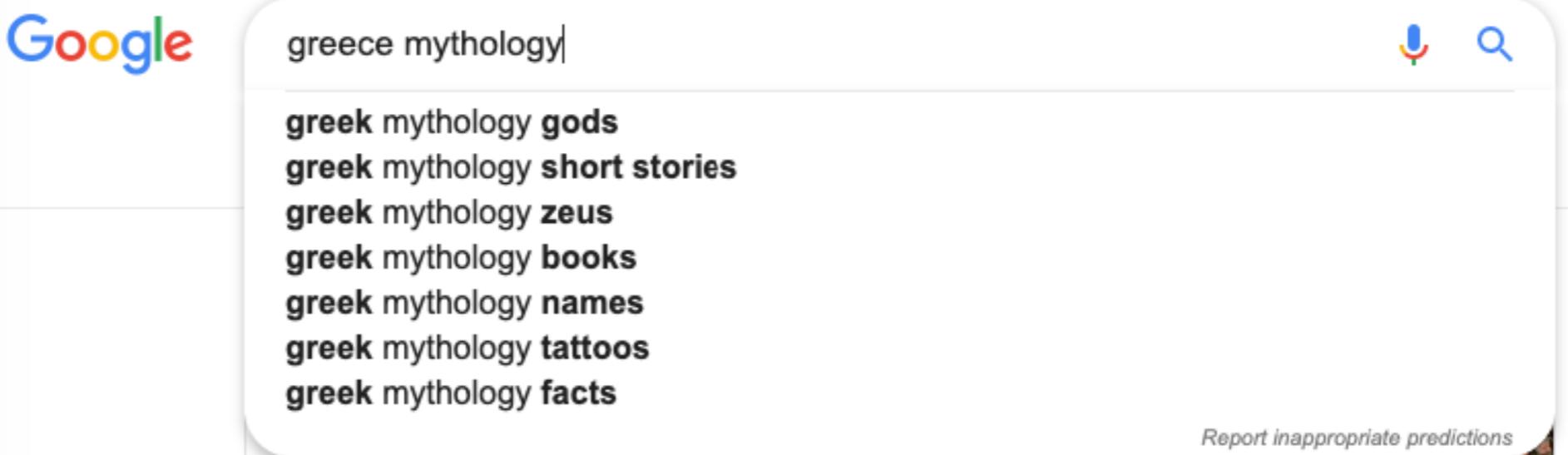
HLV Man Utd cung cấp
thông tin về tiền đạo đang
bị thất sủng ở Old
Trafford, hôm 26/8.

$f(text) \rightarrow reply$
 $\rightarrow sentiment$
 $\rightarrow decision$

...

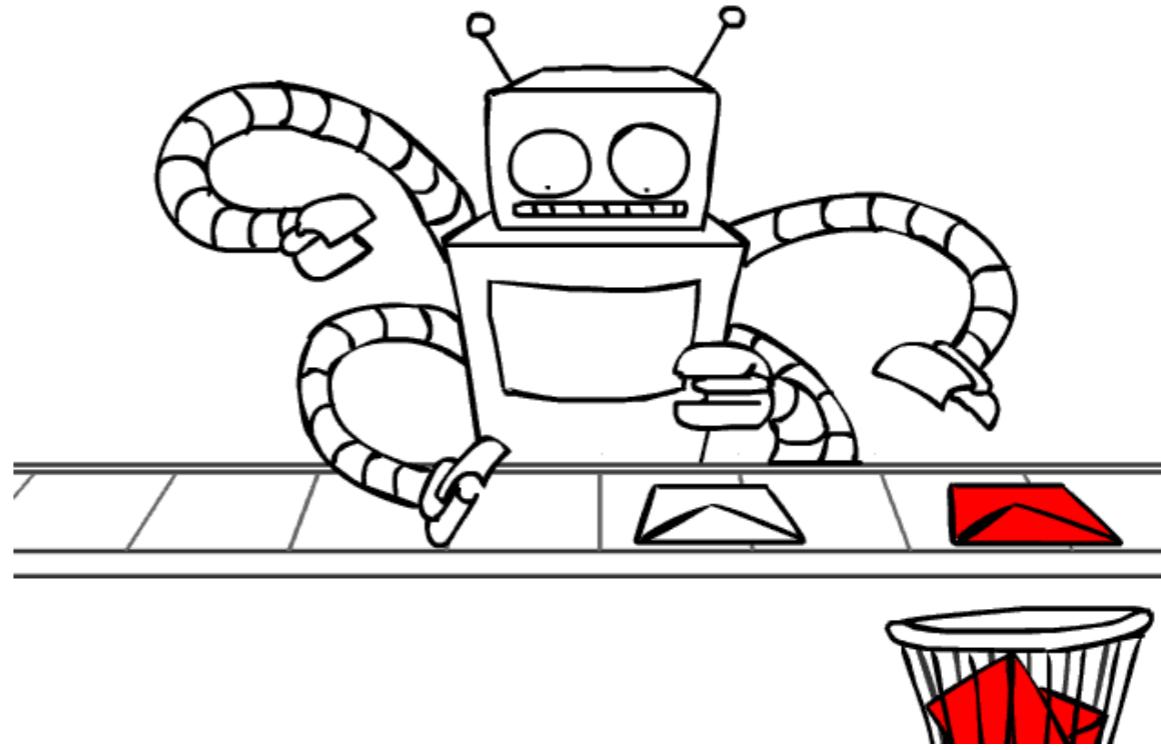
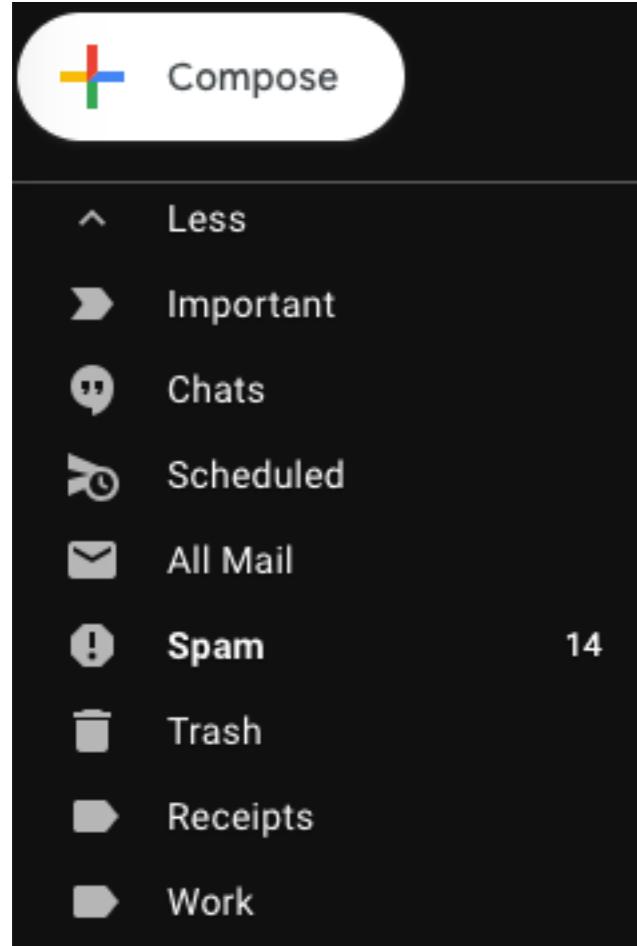


Applications



A screenshot of the Google Translate application. At the top, it shows the logo and a menu icon. Below that, there are tabs for "Text" and "Documents". The language selection bar shows "JAPANESE - DETECTED" on the left, followed by "ENGLISH", "SPANISH", and "ARABIC". The main area displays a Japanese sentence "いい天気です" (Ittenkidesu) with its English translation "Nice weather". There is also a green circular icon with a "G" and a small "X" button between the two translations. At the bottom, there are input fields for microphone and text entry, a character count of "6/5000", a pencil icon, a speaker icon, and additional buttons for document and more options.

Applications



Applications

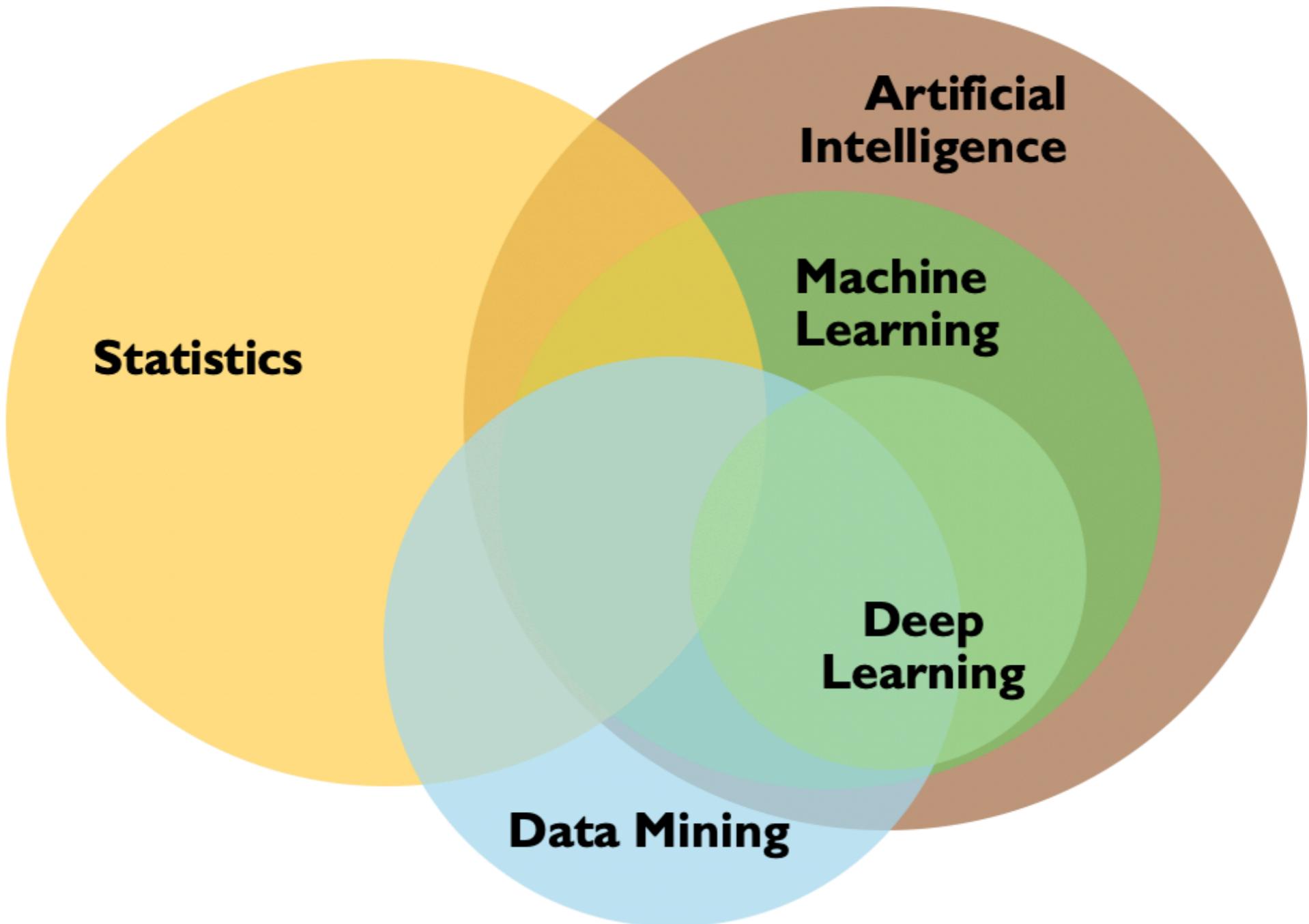


ML applications

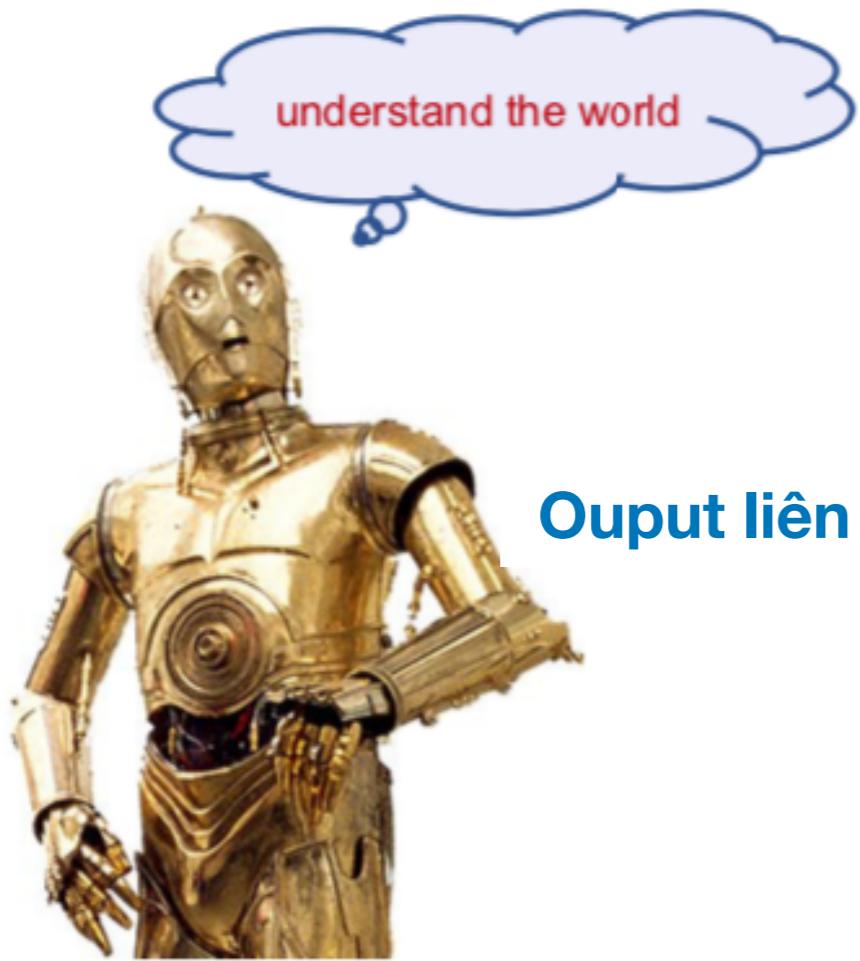


Over 75% of what people watch comes from our recommendations

AI, Data Mining, and Data Science



Classification vs. Regression



Ouput liên tục: y



input: X

Bài toán hồi qui
(regression)



Ouput rời rạc: y

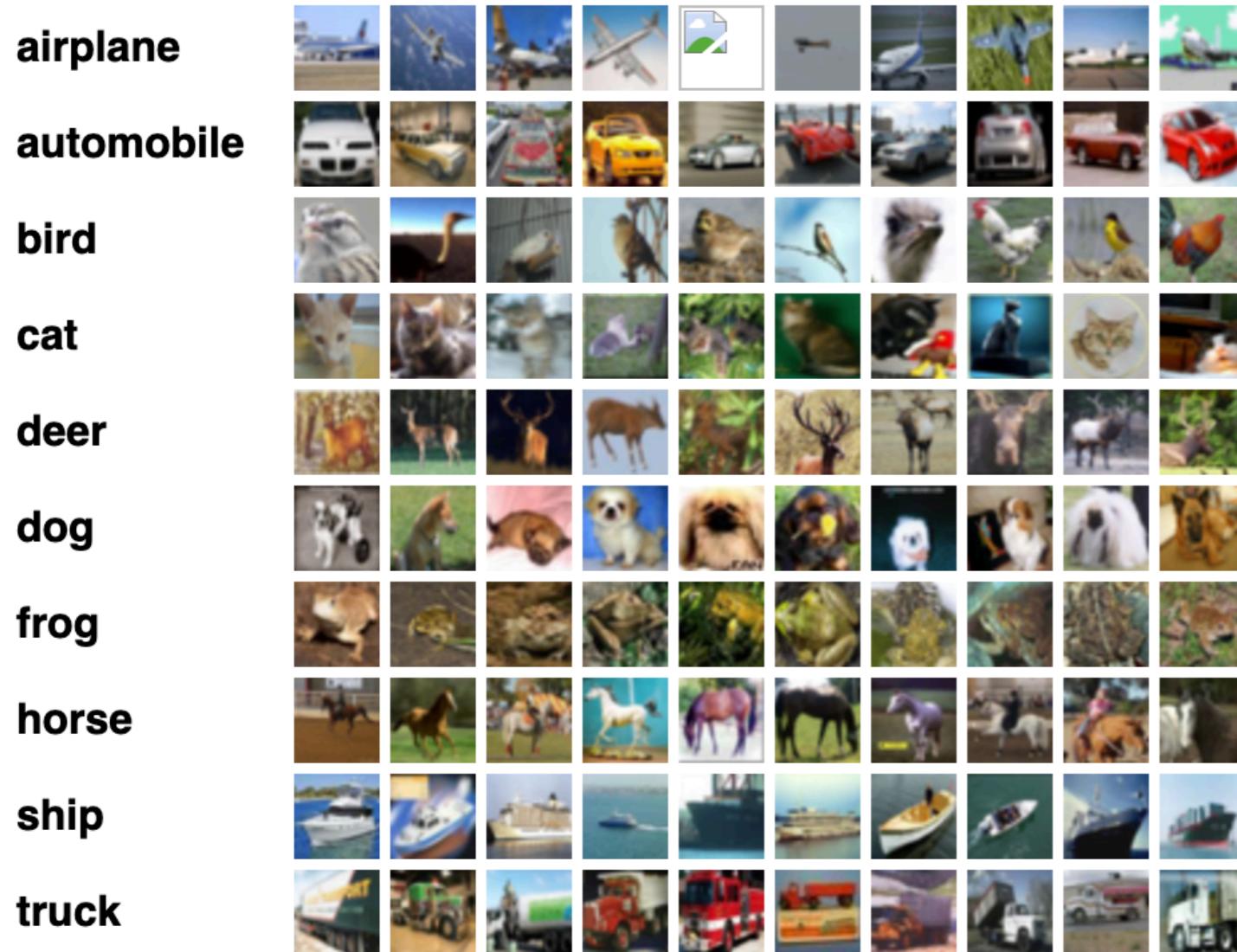


input: X

Bài toán phân loại
(classification)

Data Driven Approach

Nhiệm vụ: gán nhãn cho ảnh $F(imag) = label$



Transitional approach

```
def classify_image(image):  
    # Some magic here?  
    return class_label
```

Data-Driven Approach

(1) Problem setting

(2) Data collection

$$D = \{(x_i, y_i), i = 1, 2, \dots, m\}$$

(3) Modeling and Training Models

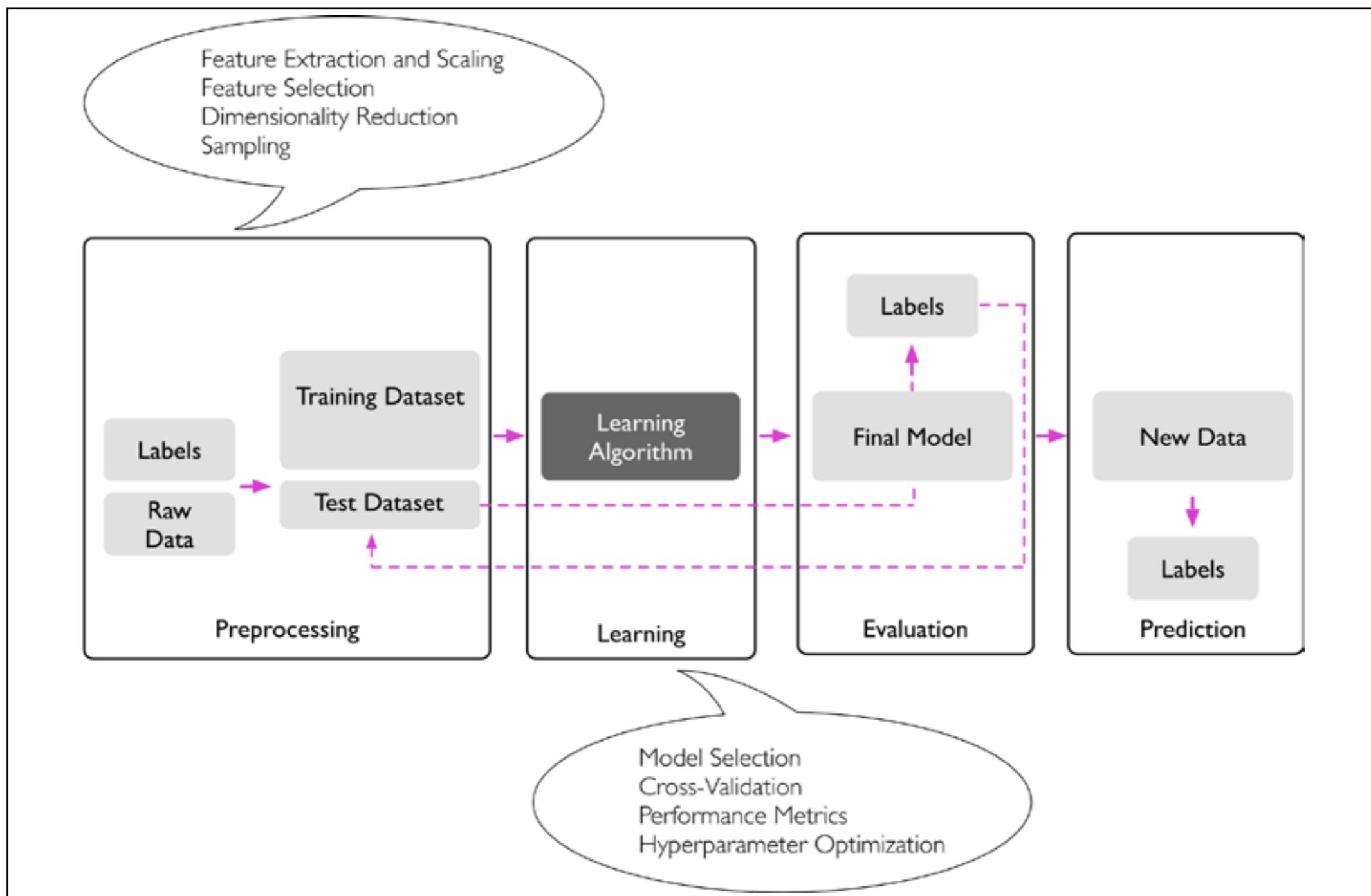
```
def train(images, labels):
    # Machine learning!
    return model
```

(4) Model selection

(5) Deploy suitable model (Using the best model to make prediction)

```
def predict(model, test_images):
    # Use model to predict labels
    return test_labels
```

Roadmap for Building Machine Learning Systems



Machine Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell

“Một chương trình máy tính được cho là có khả năng học từ kinh nghiệm (dữ liệu), E, để thực hiện một chức năng T với hiệu năng P, nếu nó thực hiện chức năng T tốt hơn (được đo bằng hiệu năng P) nếu có mặt dữ liệu E”

What You are Learning

- Python Programming
- Introduction to data analysis
- Supervised Learning
 - ✓ Regression
 - ✓ Classification
- Unsupervised Learning
 - ✓ Clustering
 - ✓ Dimensionality reduction
- Q-learning

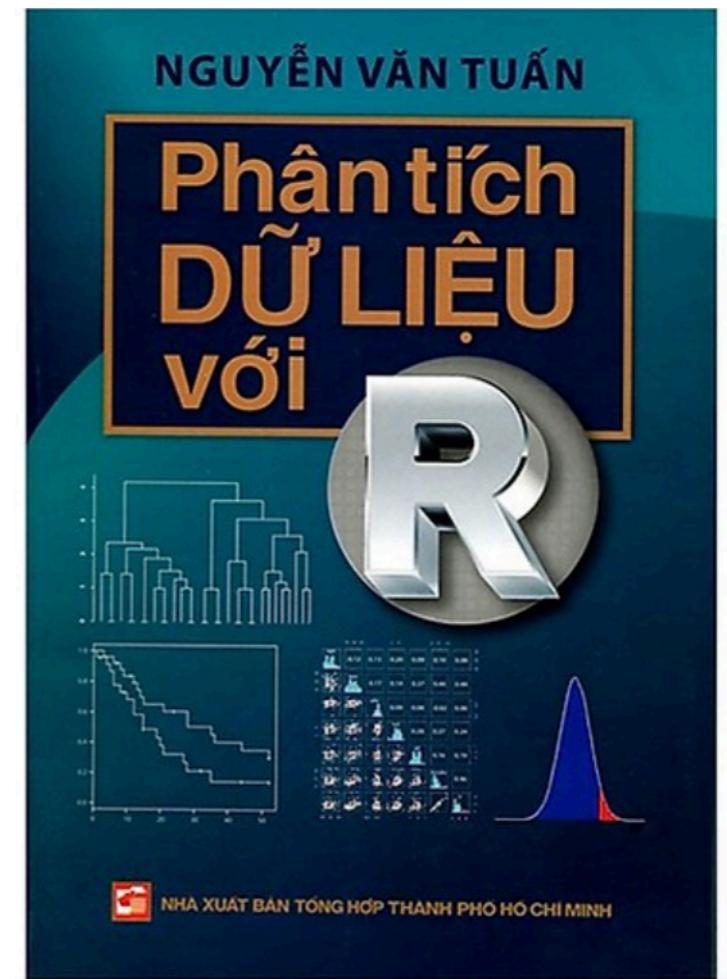
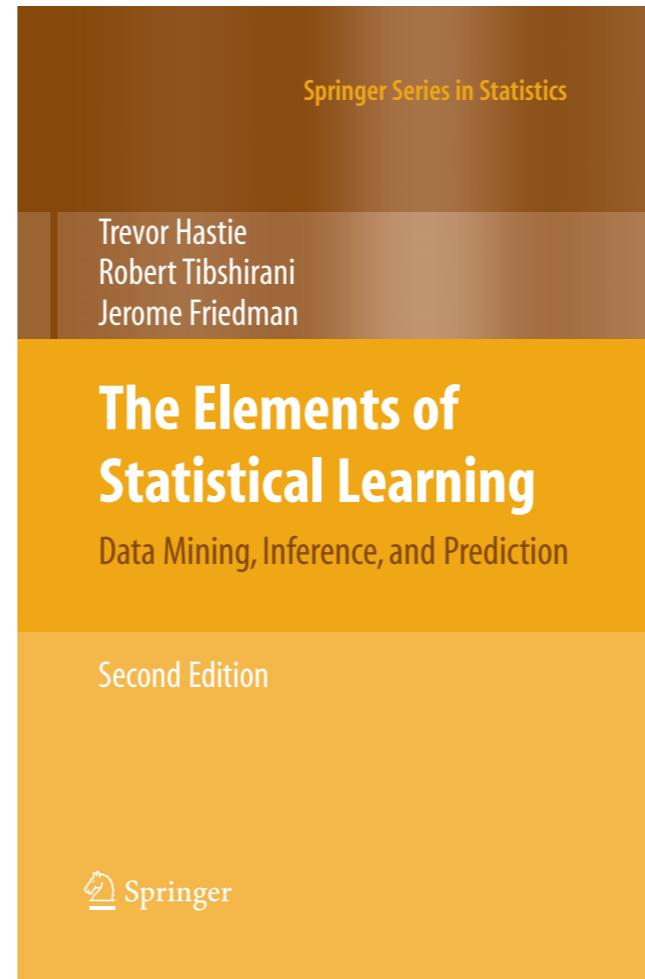
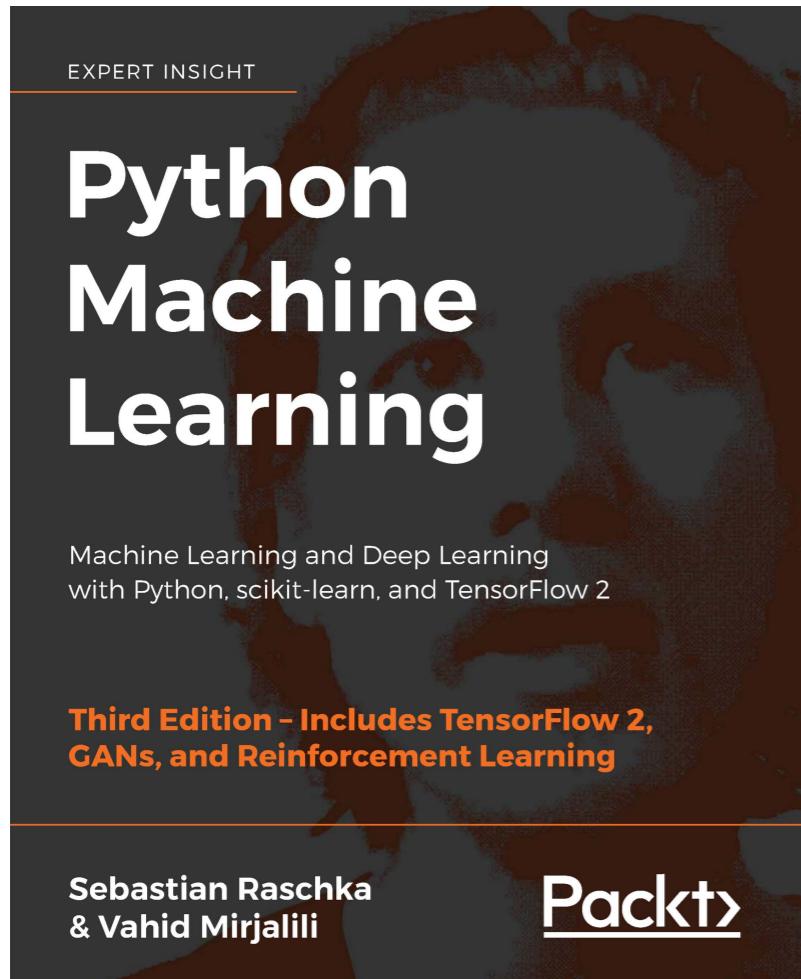
Where to Learn

Machine Learning cơ bản

[About](#) [Index](#) [Tags](#) [Categories](#) [Archive](#) [Math](#) [Copyrights](#) [ebook](#) [Search](#)

[Bài 2: Phân nhóm các thuật toán Machine Learning »](#)

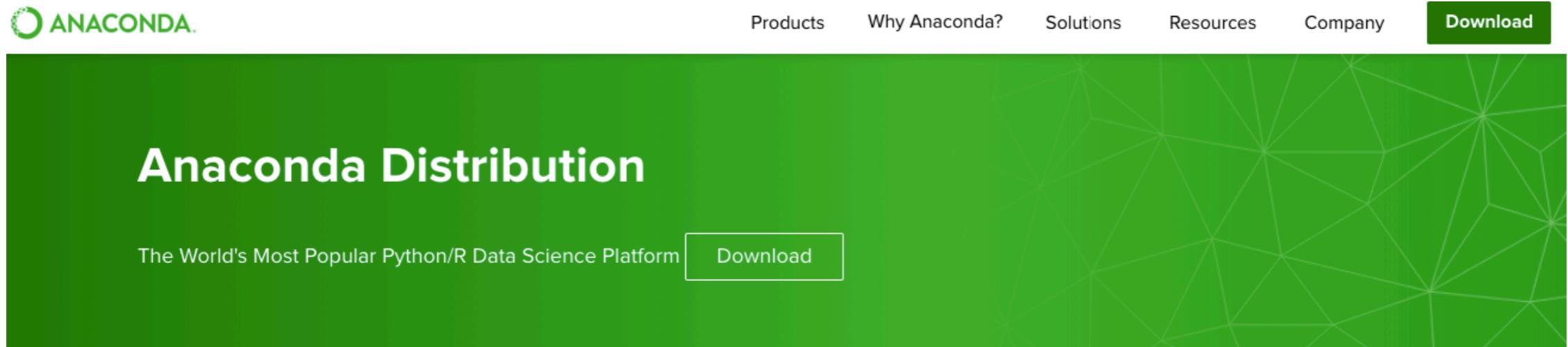
Bài 1: Giới thiệu về Machine Learning



Đánh giá học phần

- Điểm chuyên cần: 10%
- Điểm bài tập: 30 %
- Điểm thi: 60%

Learning by Coding



The image shows the Google Colaboratory interface. At the top, there is a header with back, forward, and refresh buttons, followed by the URL 'colab.research.google.com'. Below the header, the title 'Welcome To Colaboratory' is displayed, along with a 'File' menu. A navigation bar with tabs for 'Examples', 'Recent' (which is underlined), 'Google Drive', 'GitHub', and 'Upload' is shown. On the left side, there is a sidebar with icons for 'Table of contents', 'Getting started', 'Data science', 'Machine learning', and 'More Resources'. The main content area displays a search bar labeled 'Filter notebooks' and a table with columns for 'Title', 'Last opened' (with up and down arrows), and 'First opened' (with up and down arrows). There is also a trash can icon in the bottom right corner of the main content area.