



Bài giảng môn học:
Khoa Học Dữ Liệu (7080509)

CHƯƠNG 1: MỞ ĐẦU

Đặng Văn Nam
dangvannam@hmg.edu.vn

Nội dung chương 1

1.1 Vai trò của Khoa học dữ liệu trong thực tế

1.2 Tổng quan về Khoa học dữ liệu

1.3 Khoa học dữ liệu và Dữ liệu lớn

1.4 Quy trình của dự án về Khoa học dữ liệu

1.5 Kiến thức và kỹ năng của nhà khoa học dữ liệu



Nội dung chương 1



“DATA IS THE NEW GOLD”

1. Vai trò của khoa học dữ liệu trong thực tế

WHO USES DATA SCIENCE?



Booking.com



LinkedIn

IBM



NETFLIX

18

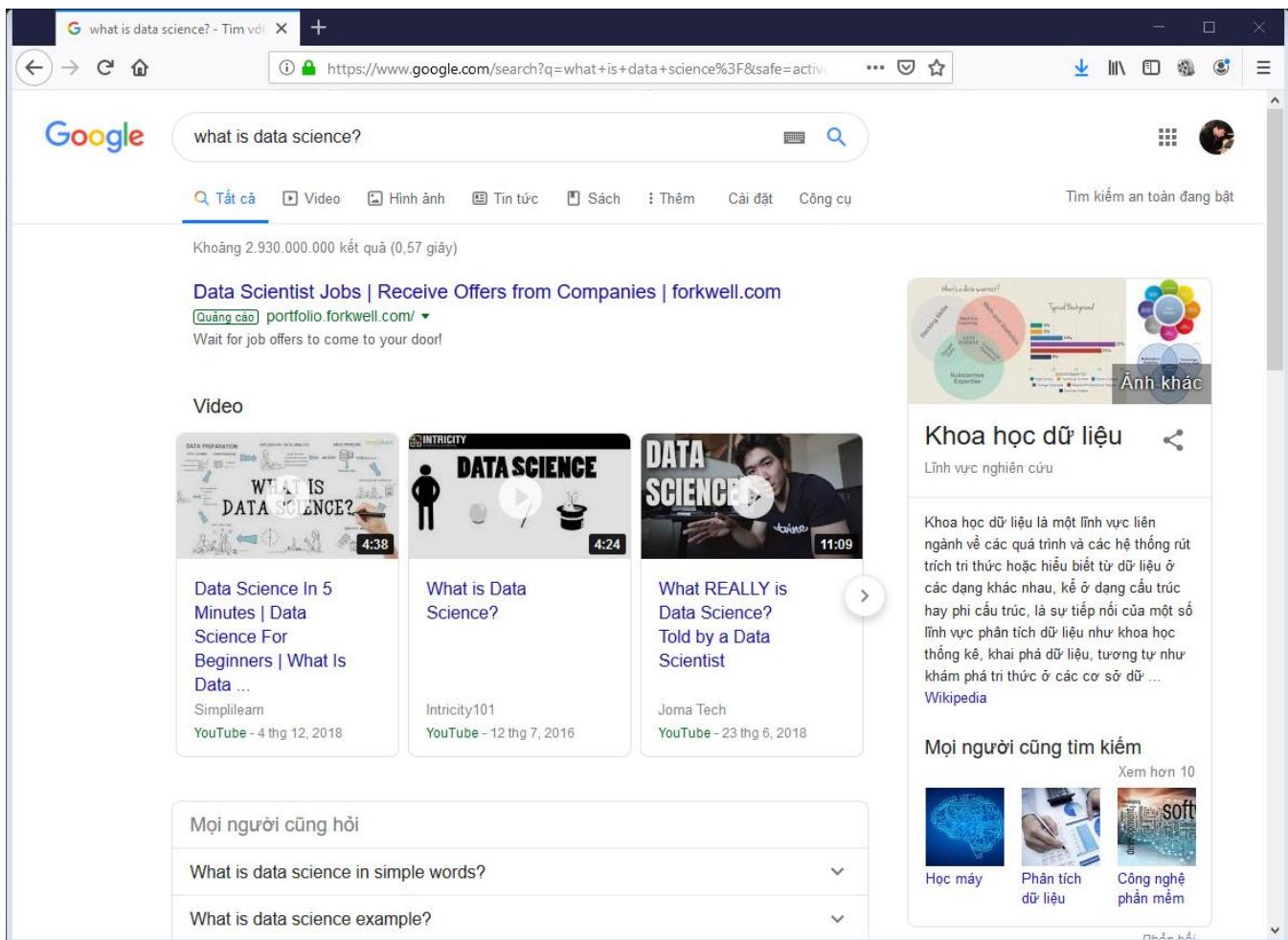
Bank of America



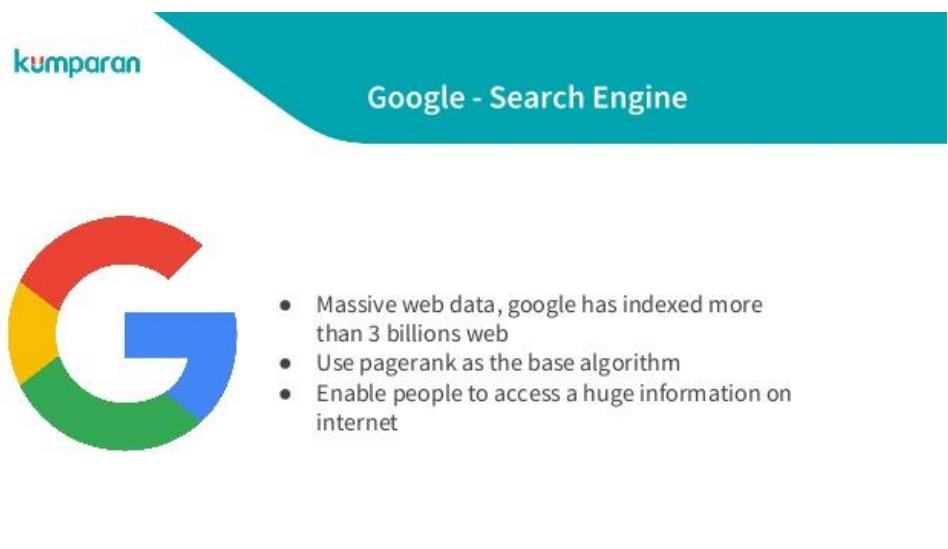
SAMSUNG

Google

Bạn có tự hỏi tại sao google lại có thể trả về hàng triệu kết quả tìm kiếm chỉ trong 1s?



The screenshot shows a Google search results page with the query "what is data science?". The results include a search bar at the top, followed by a summary section stating "Khoảng 2.930.000.000 kết quả (0,57 giây)". Below this, there's a "Data Scientist Jobs" advertisement from forkwell.com, followed by a "Video" section with three video thumbnails: "WHAT IS DATA SCIENCE?", "DATA SCIENCE", and "What REALLY is Data Science? Told by a Data Scientist". To the right of the video section is a "Khoa học dữ liệu" (Data Science) article from Wikipedia, featuring a Venn diagram and some text. At the bottom of the main search results area, there are sections for "Mọi người cũng hỏi" (People also ask) with questions like "What is data science in simple words?" and "What is data science example?", and a "Xem hơn 10" (See more) link.



kumparan

Google - Search Engine

Google logo

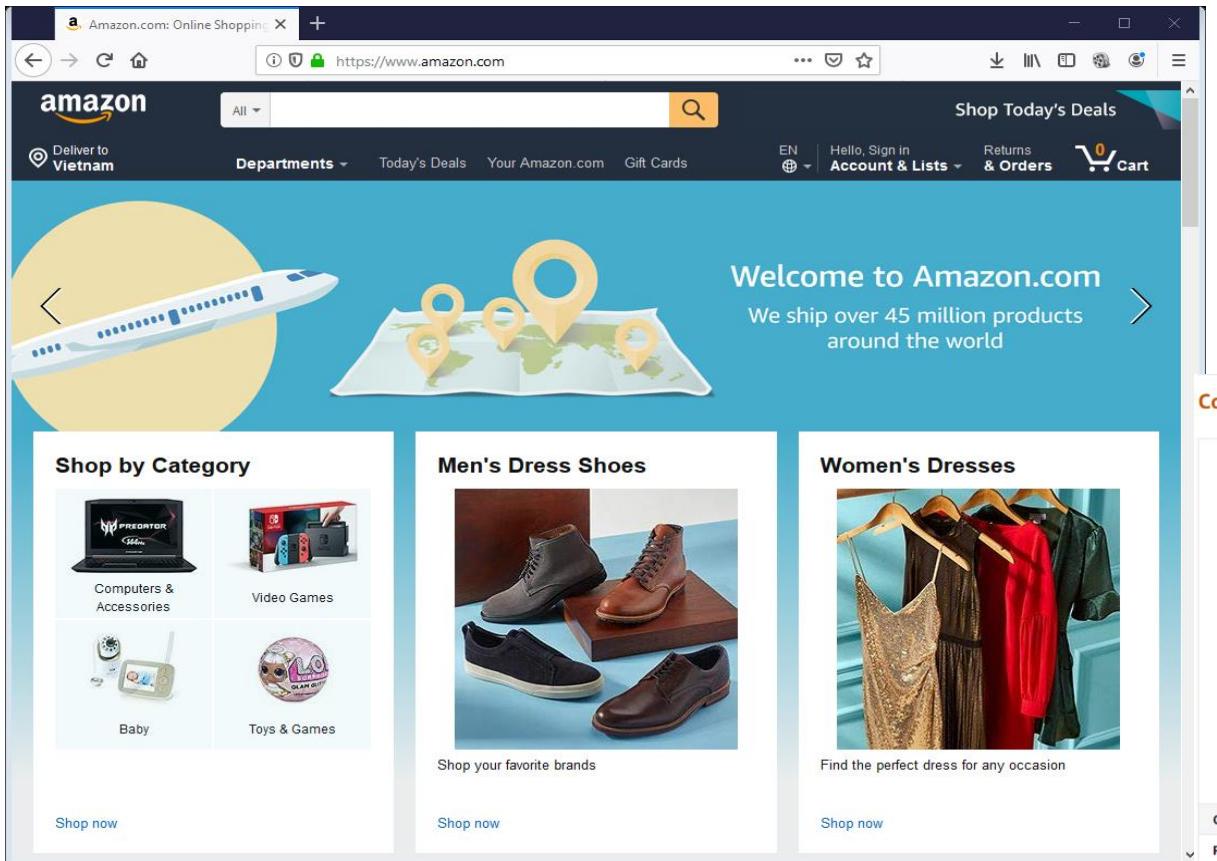
- Massive web data, google has indexed more than 3 billions web
- Use pagerank as the base algorithm
- Enable people to access a huge information on internet

Facebook



Facebook kết nối mọi người thế nào?

E-commerce: Amazon

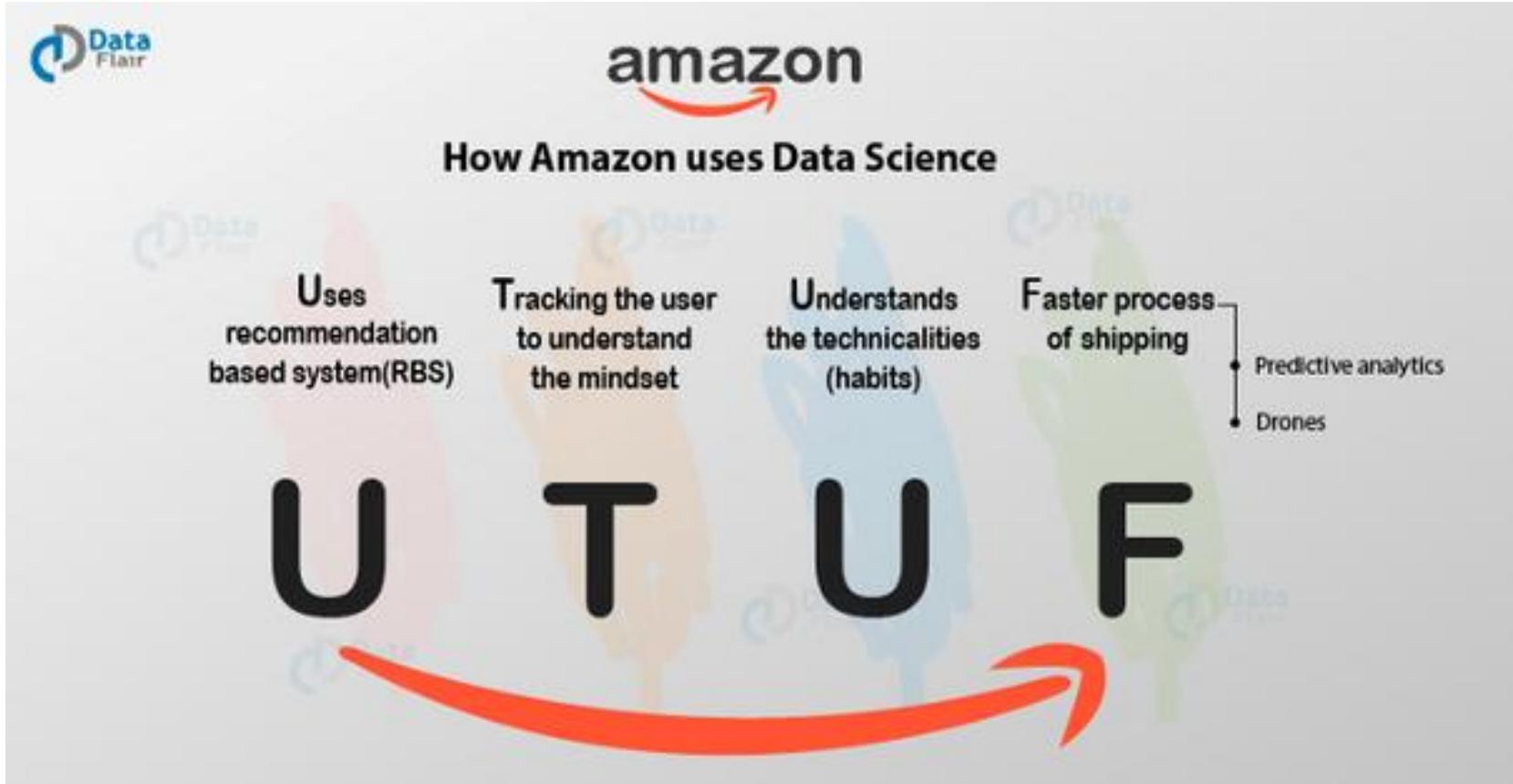


Compare with similar items

			
This item Bose SoundLink Wireless Around-Ear Headphones with Mic (Black)	Sennheiser HD 4.40-BT Bluetooth Headphones (Black)	Bose 741158-0020 SoundLink Wireless Around-Ear Headphones with Mic (White)	Bose 789564-0030 Quiet Comfort 35 Wireless Headphone (Blue)-Special Edition
Add to Cart	Add to Cart	Add to Cart	Add to Cart
Customer Rating  (68)	 (349)	 (22)	 (200)
Price ₹ 19,000.00	₹ 7,490.00	₹ 19,000.00	₹ 29,363.00
Shipping FREE Shipping	FREE Shipping	FREE Shipping	FREE Shipping
Sold By Appario Retail Private Ltd	Appario Retail Private Ltd	Appario Retail Private Ltd	Appario Retail Private Ltd
Colour Black	Black	White	Blue
Connectivity Technology bluetooth wireless	Bluetooth Wireless	Bluetooth Wireless	Bluetooth Wireless

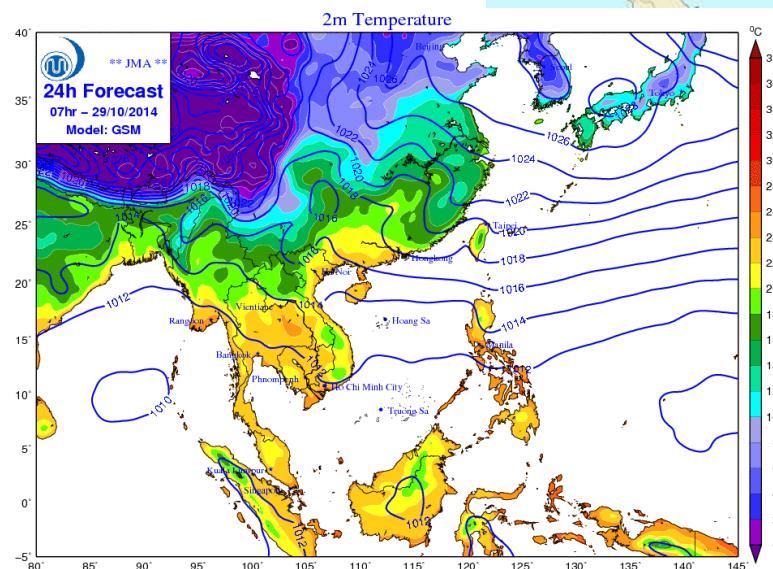
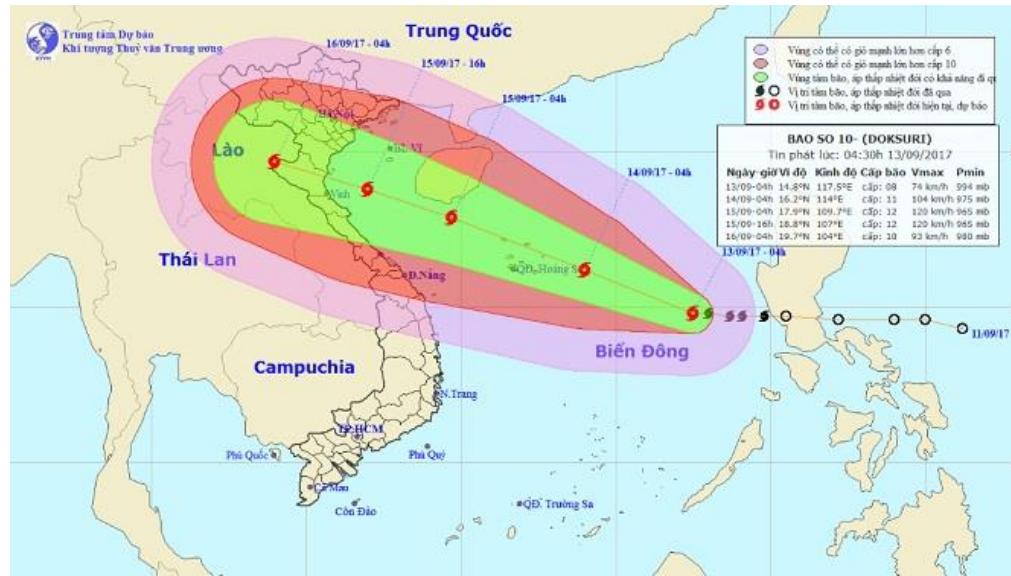
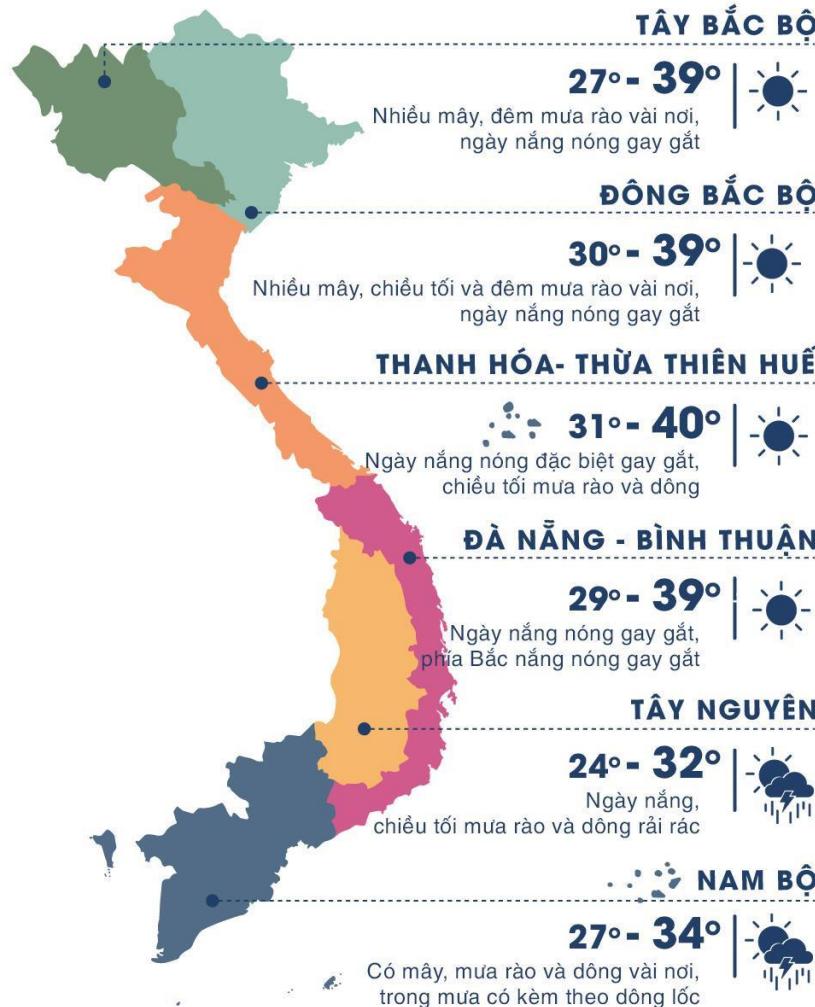
Khi tìm một sản phẩm, Amazon gợi ý cho ta một loạt sản phẩm liên quan như thế nào? So sánh giá? Nhận diện màu sắc....?

Amazon

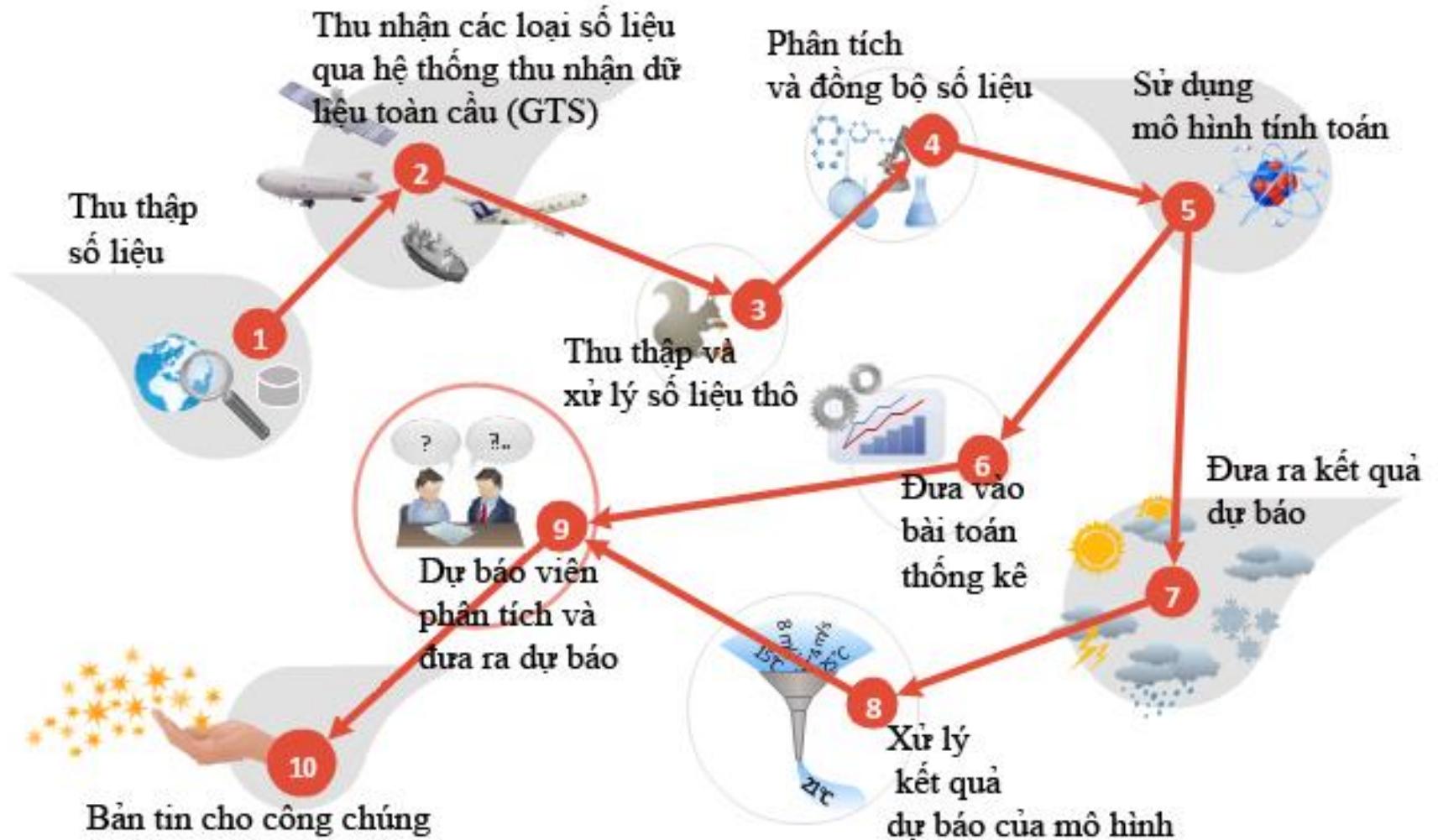


Weather Forecasting

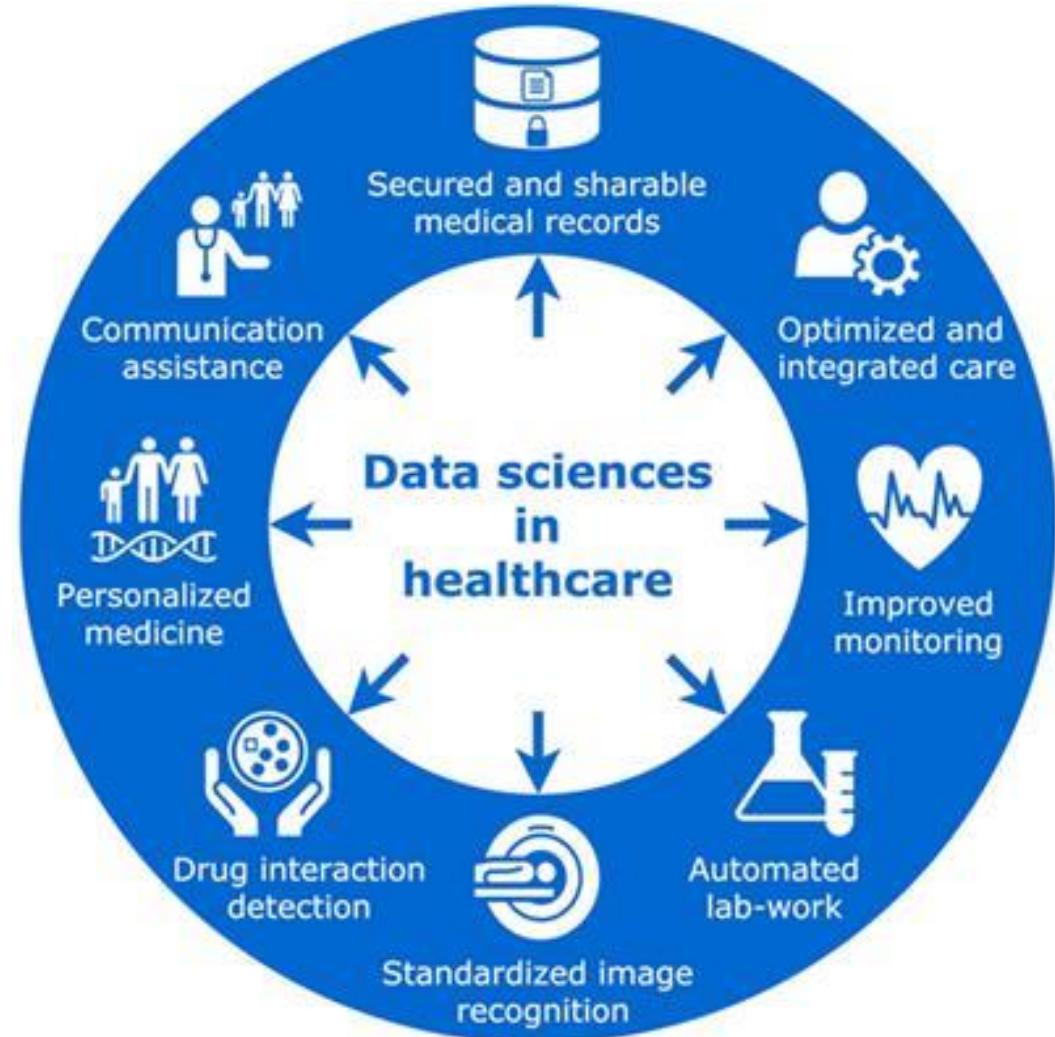
DỰ BÁO THỜI TIẾT NGÀY 28/6



Weather Forecasting

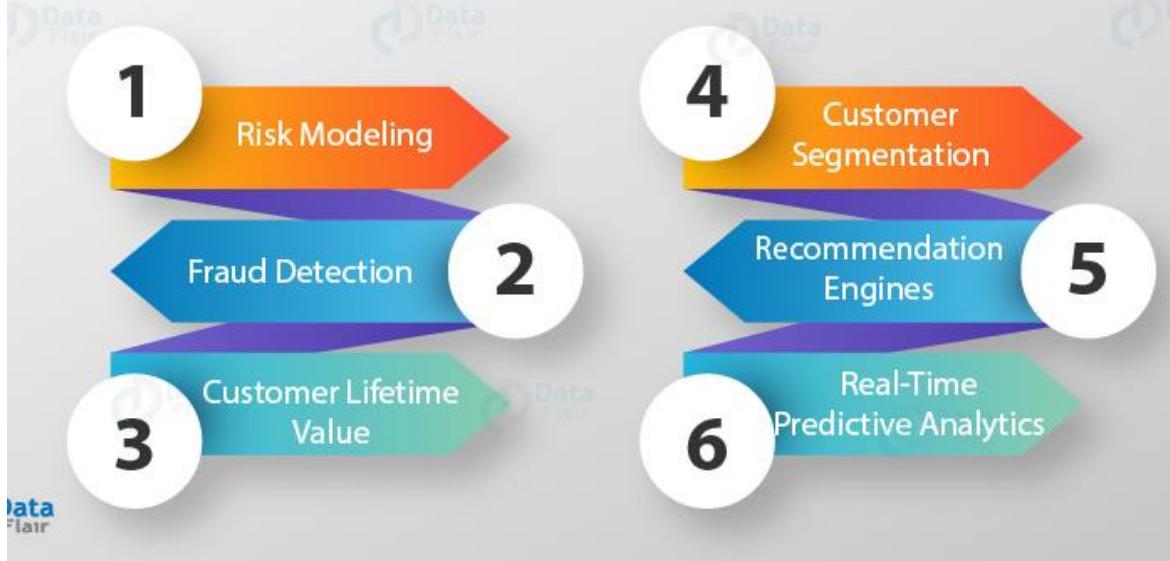


Healthcare



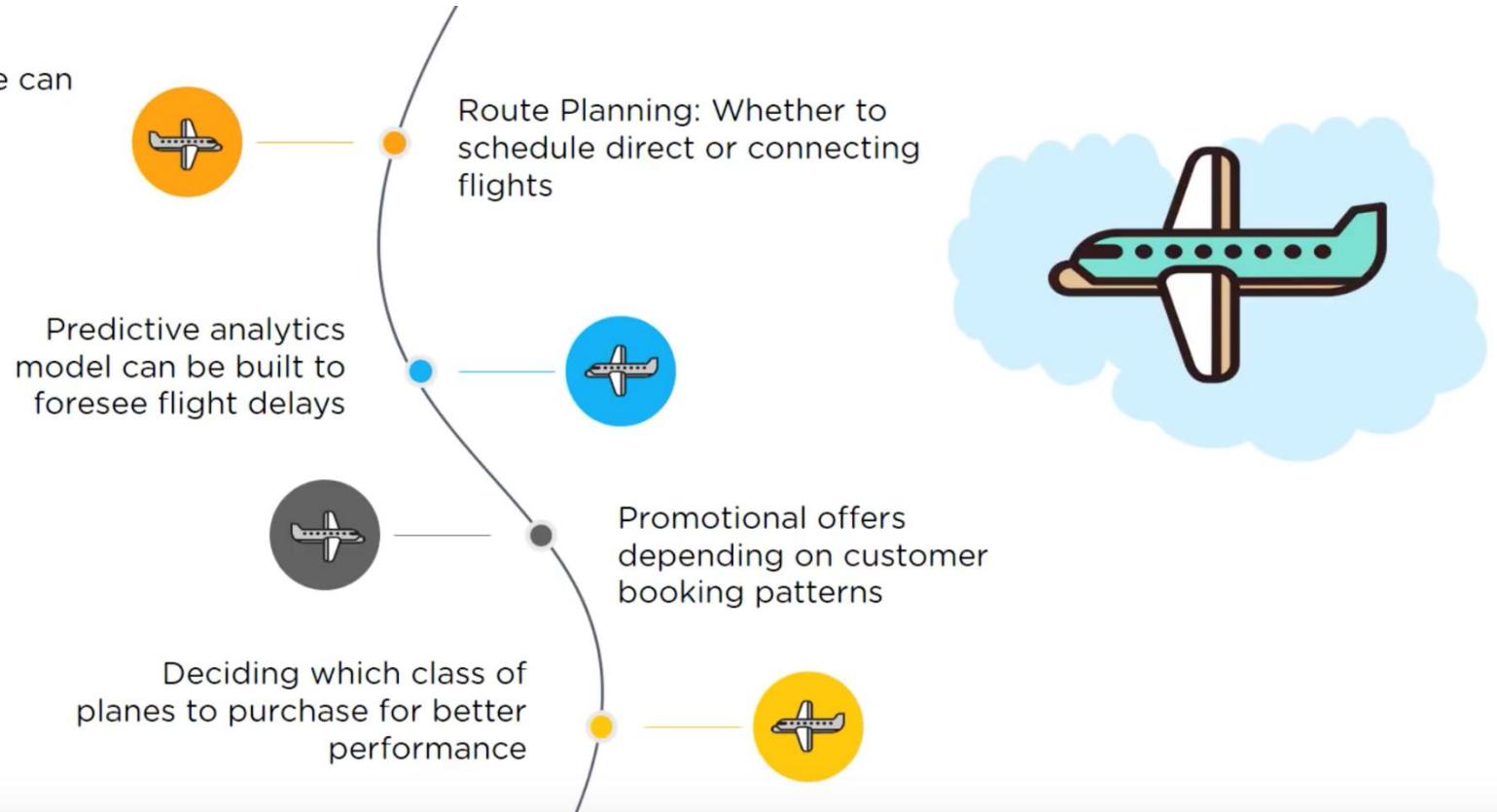
Banking

Applications of Data Science in Banking



Airlines

Using Data Science, we can achieve the following:

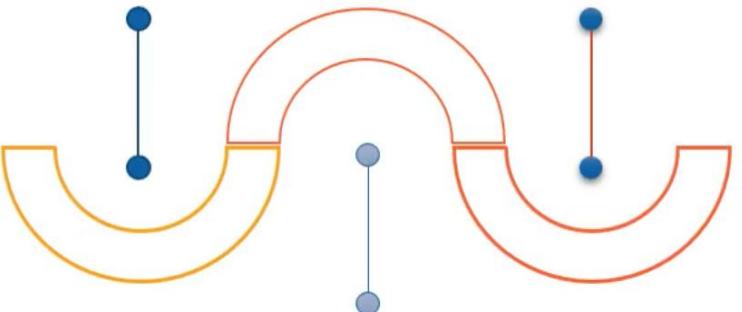


Transport

Logistics companies like FedEx are using Data Science models for operational efficiency

Discover the best routes to ship

The best suited time to deliver



The best mode of transport



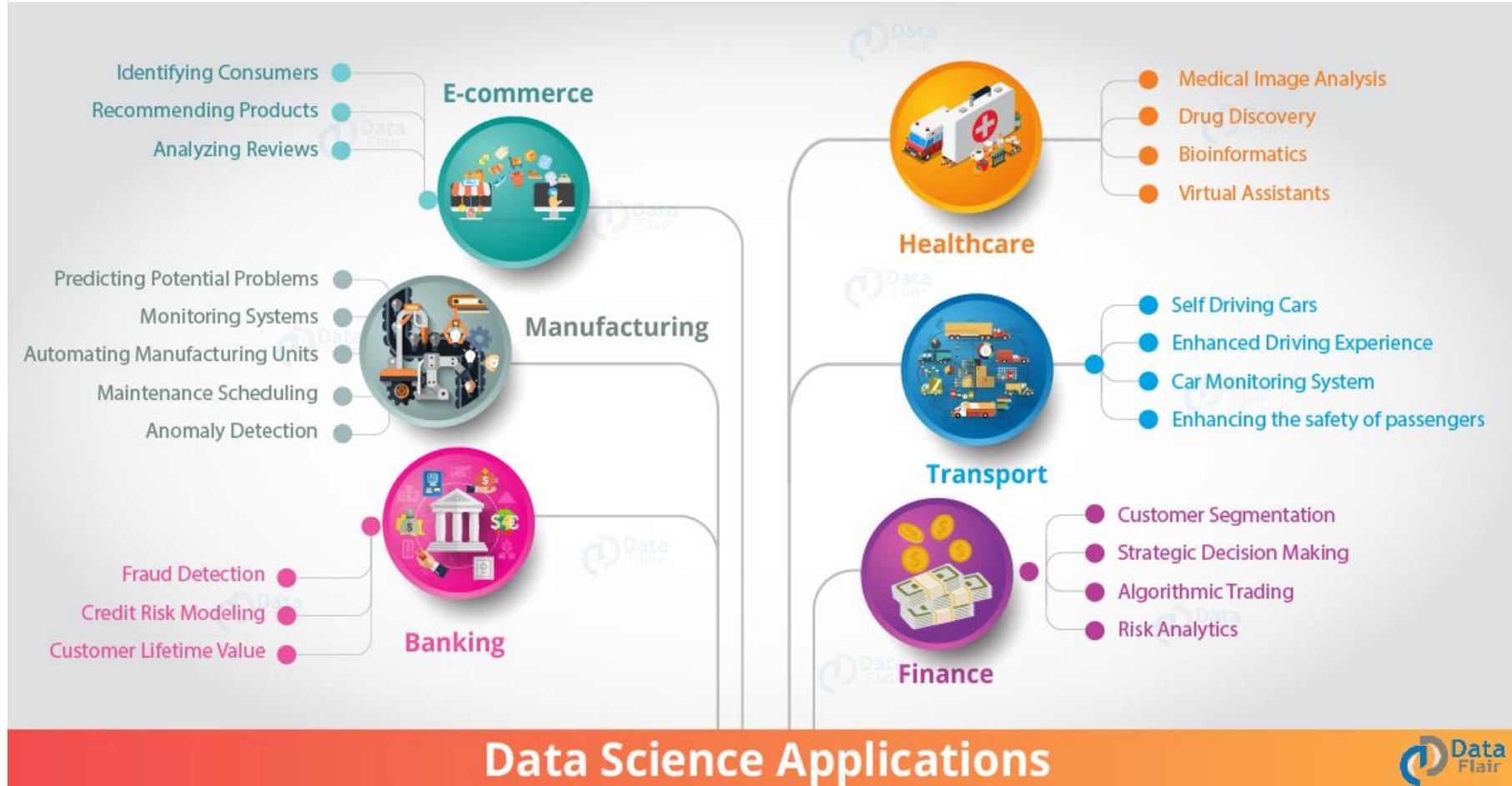
1. Vai trò của Khoa học dữ liệu trong thực tế

- Các công ty thương mại sử dụng khoa học dữ liệu để hiểu rõ hơn về khách hàng, quy trình, nhân viên, hoàn thiện và sản phẩm của họ.



- Các tổ chức tài chính sử dụng khoa học dữ liệu để dự đoán thị trường chứng khoán, xác định rủi ro cho vay tiền và tìm hiểu cách thu hút các nhân viên mới cho dịch vụ của họ.
- Các tổ chức chính phủ sử dụng khoa học dữ liệu để khai phá thông tin, phát hiện các hoạt động tội phạm, tối ưu hóa các hoạt động.
- Các tổ chức NGO, WWF sử dụng Khoa học dữ liệu để tăng hiệu quả các dự án phát triển gây quỹ, bảo vệ tài nguyên.
- Các trường đại học sử dụng khoa học dữ liệu trong nghiên cứu, phát triển và đánh giá các lớp học trực tuyến.

1. Vai trò của Khoa học dữ liệu trong thực tế



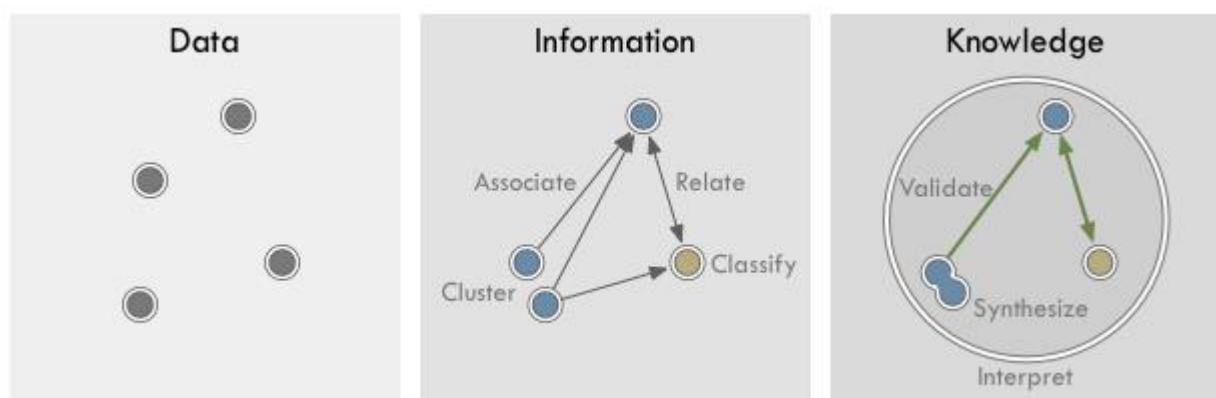
2. Tổng quan về khoa học dữ liệu



Data
Science

Data – Information - Knowledge

- Dữ liệu **Data**: là các yếu tố thô, chưa được xử lý, bao gồm: văn bản, số liệu, ký hiệu, hình ảnh, âm thanh,...
- Thông tin **Information** là dữ liệu đã được xử lý để đáp ứng yêu cầu của người dùng
- Tri thức/khiến thức **Knowledge**: bao gồm những dữ kiện, thông tin, sự mô tả hay kỹ năng có được nhờ trải nghiệm hay thông qua giáo dục.



Khoa học dữ liệu là gì?

Chưa có một định nghĩa về Khoa học dữ liệu (Data Science) được tất cả mọi người chấp nhận.

Một vài định nghĩa:

Viện Tiêu chuẩn và Công nghệ Quốc gia Mỹ (National Institute of Standards and Technology - NIST):

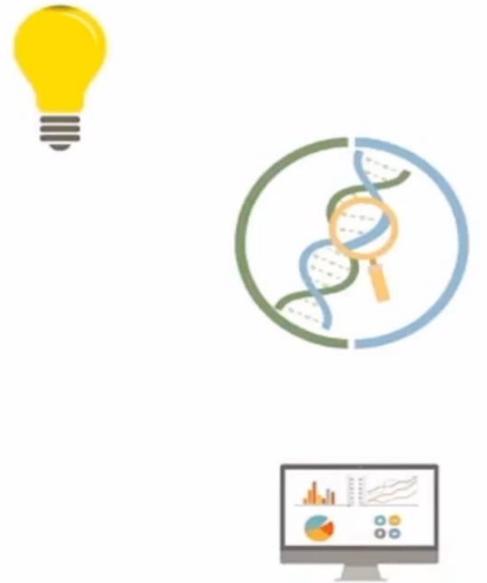
- “Data science is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing”
- Khoa học dữ liệu là trực tiếp trích rút tri thức hành động từ dữ liệu qua quá trình phát hiện, thiết lập và kiểm nghiệm các giả thiết.

Microsoft:

- “Data science is about using data to make decisions that drive actions”
- Khoa học dữ liệu là sử dụng dữ liệu tạo ra các quyết định dẫn dắt hành động.

Khoa học dữ liệu là gì?

Công cụ mới và hiệu quả trong khám phá **thông tin ẩn chứa từ dữ liệu**



Phương pháp tự **động phân tích và trích xuất** thông tin từ **khối lượng dữ liệu lớn**

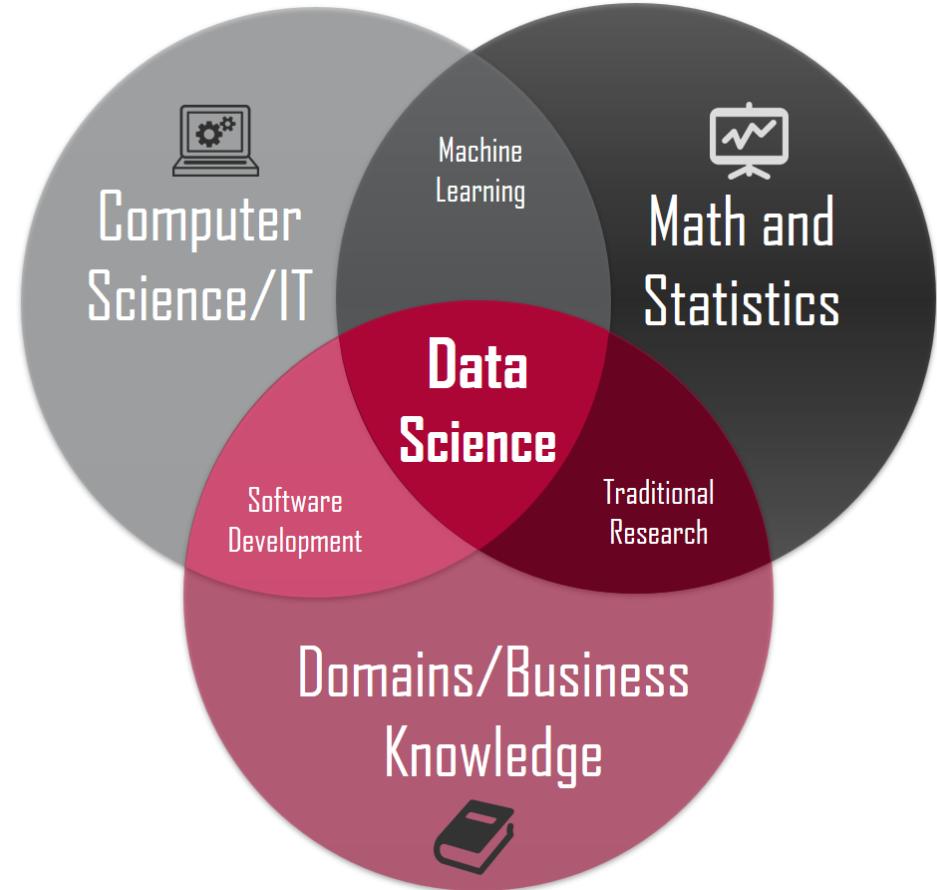
Một lĩnh vực mới kết hợp giữa **thống kê, toán học, kỹ thuật lập trình, mô phỏng** và **biểu diễn** nhằm chuyển đổi **dữ liệu thành thông tin**

Khoa học dữ liệu là gì?

- **Khoa học dữ liệu ≠ Khoa học thông thường** ở quan điểm: **Tìm tri thức (insights) từ dữ liệu (dẫn dắt bởi dữ liệu - “data-driven”)**
 - Rút ra tri thức bằng việc tìm tòi, khám phá từ dữ liệu chứ không nhất thiết phải chứng minh nó.
 - Tri thức tìm ra phải có tính ổn định (luôn có cùng kết quả nếu sử dụng cùng một phương pháp).



Khoa học dữ liệu là gì?



In God we trust, all others bring data.

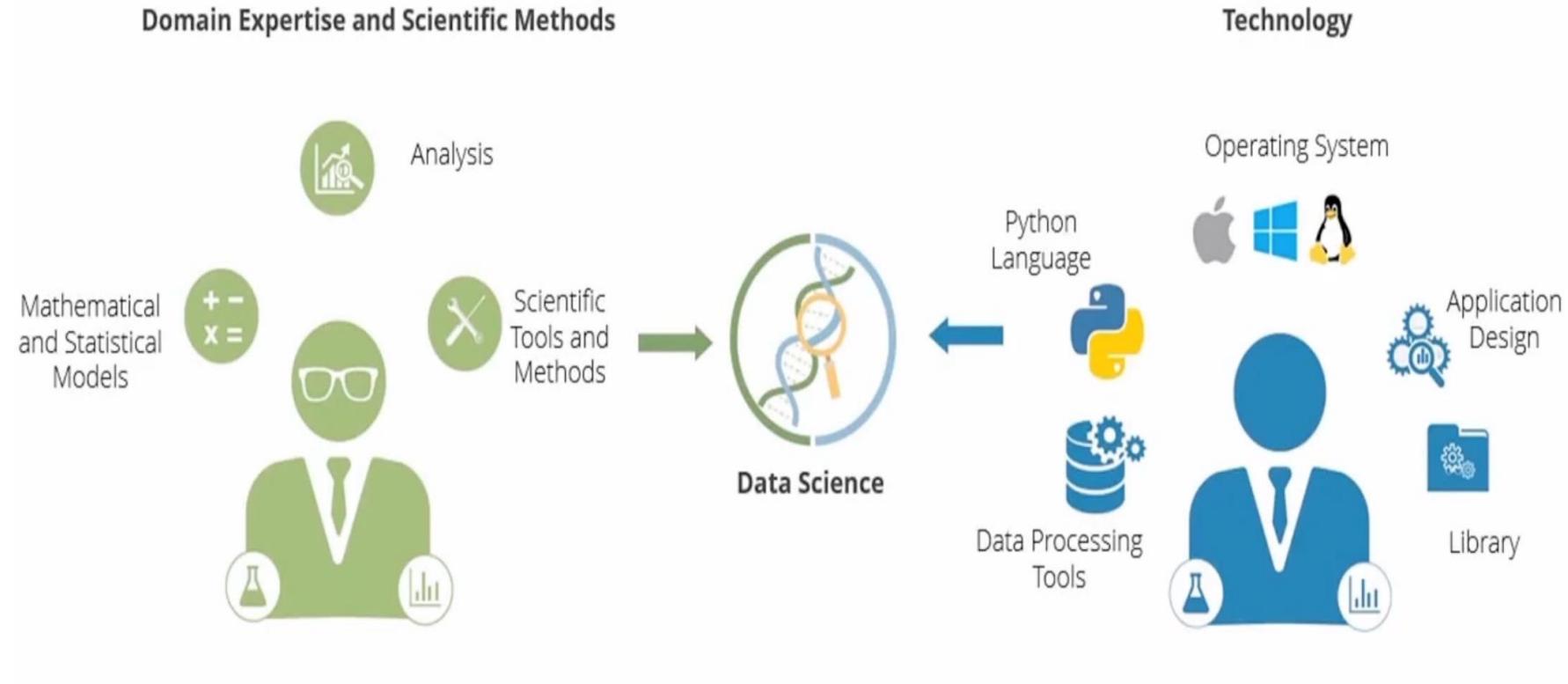
–William E. Deming



**“Ta chỉ tin vào thương đế,
Mọi thứ khác phải dựa vào dữ liệu”**

Các thành phần của khoa học dữ liệu

- Khoa học dữ liệu là sự kết hợp của **Miền chuyên môn** (Domain Expertise) và **Các phương pháp khoa học** (Scientific Methods) với **Công nghệ** (Technology)



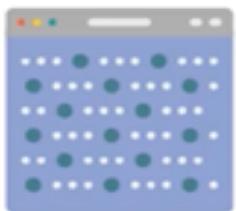
Khoa học dữ liệu cần cho?



Đưa ra các quyết định cái nào tốt hơn: Giữa A và B?

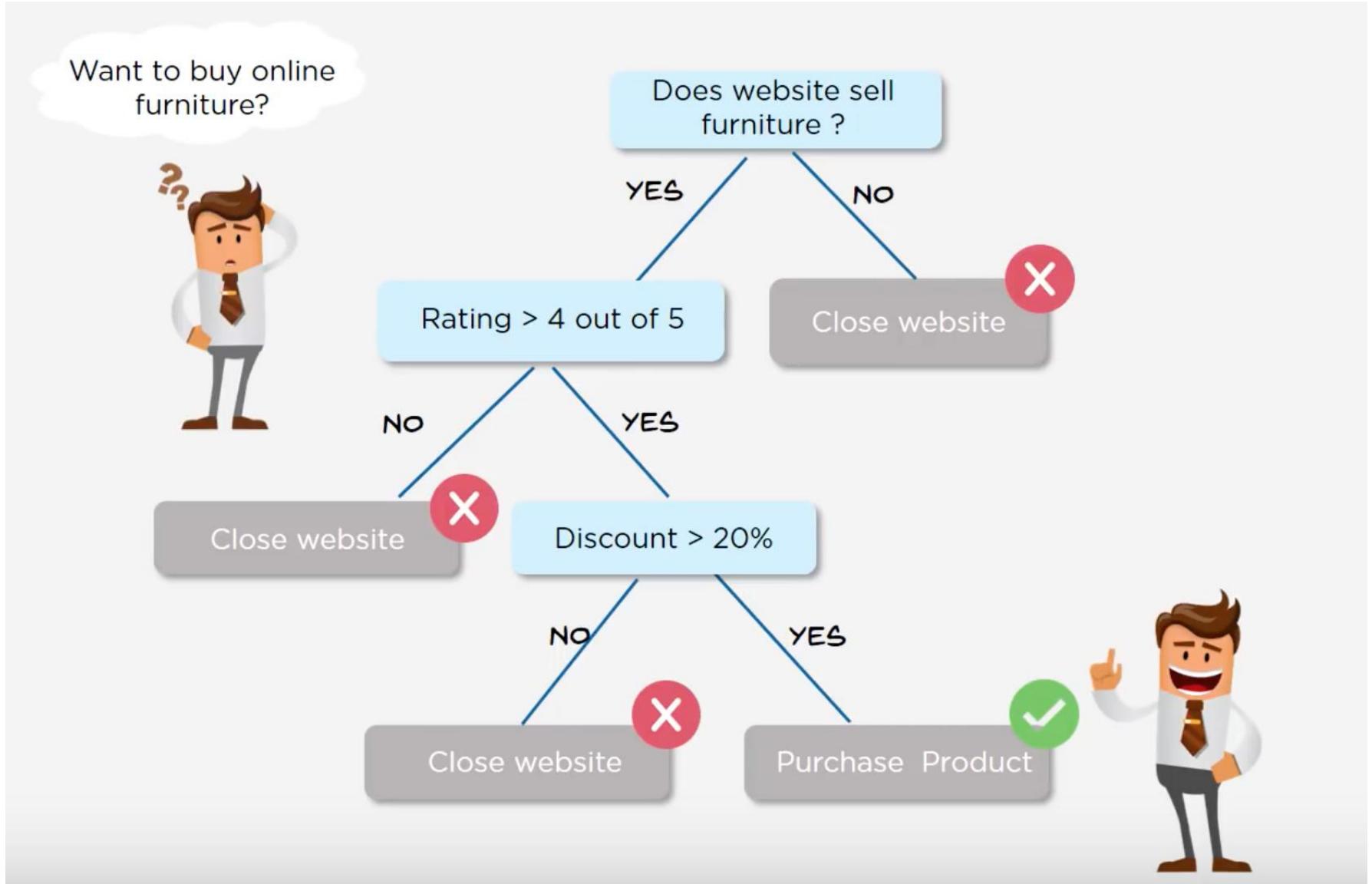


Phân tích đưa ra các dự đoán:
Chuyện gì sẽ xảy ra tiếp theo?



Nhận dạng mẫu: Có bất kỳ thông tin ẩn quan trọng nào trong mẫu không?

Ví dụ



Tiến trình khoa học dữ liệu



3. Dữ liệu lớn và Khoa học dữ liệu

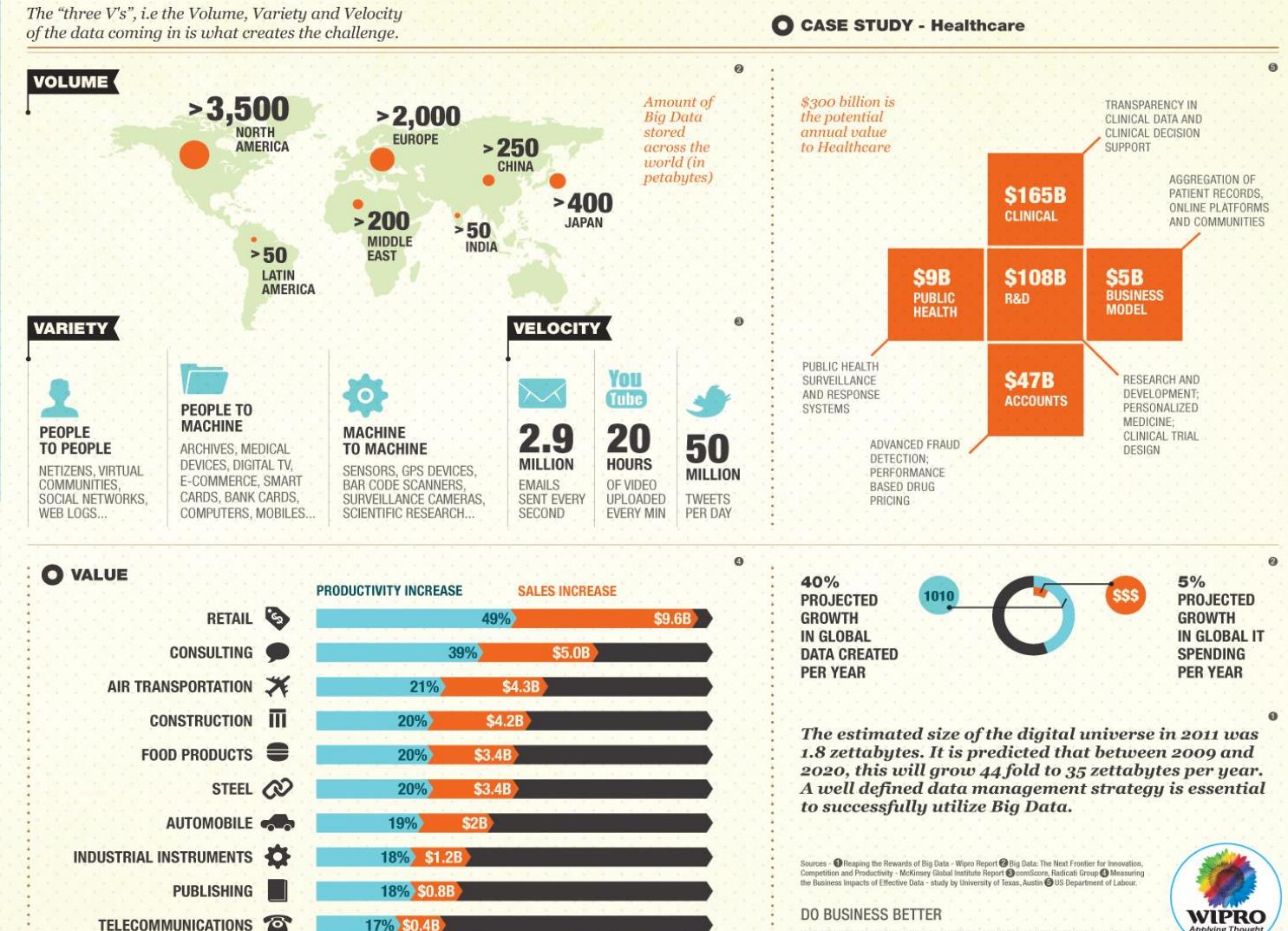


Practically

VS



Dữ liệu lớn (Big Data) là gì?

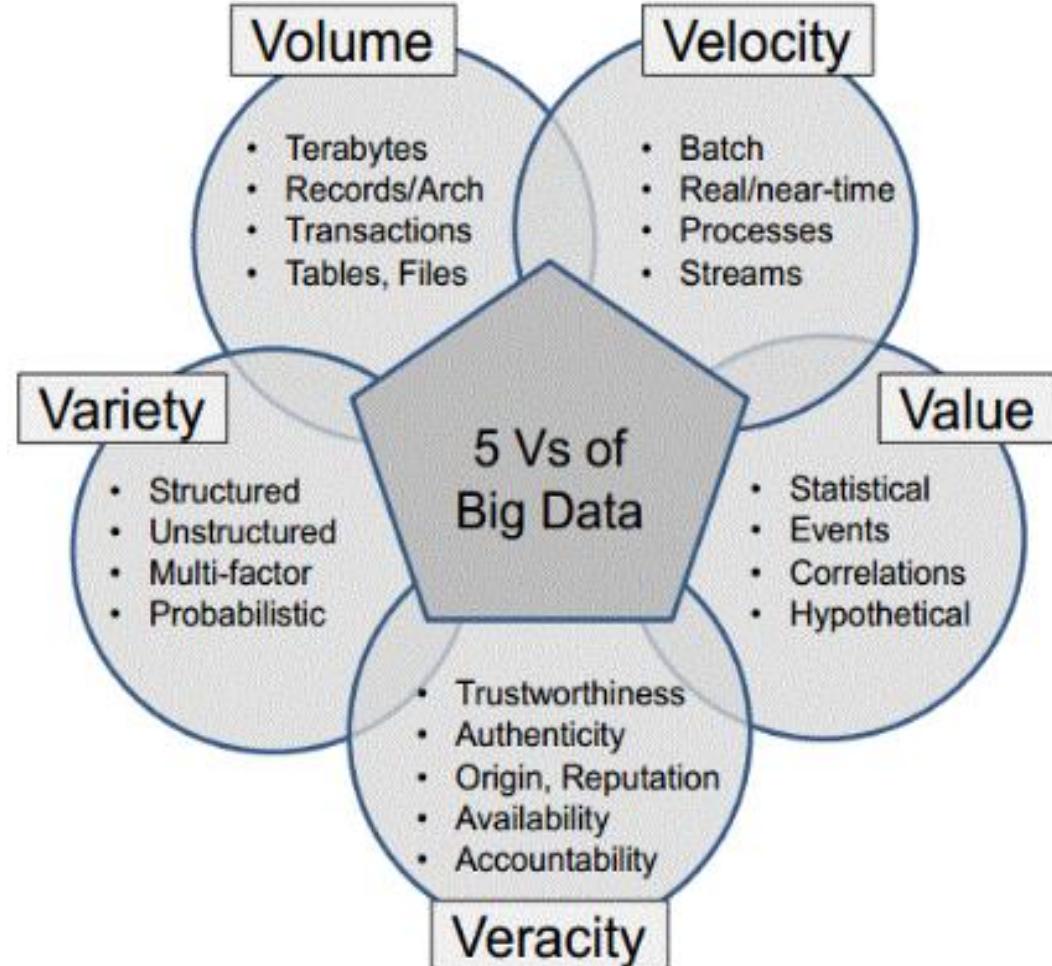


- Dữ liệu lớn (Big Data) là thuật ngữ sử dụng cho tập hợp các dữ liệu quá lớn và phức tạp khiến cho việc xử lý các dữ liệu này trở nên khó khăn khi sử dụng các kỹ thuật quản lý dữ liệu truyền thống.

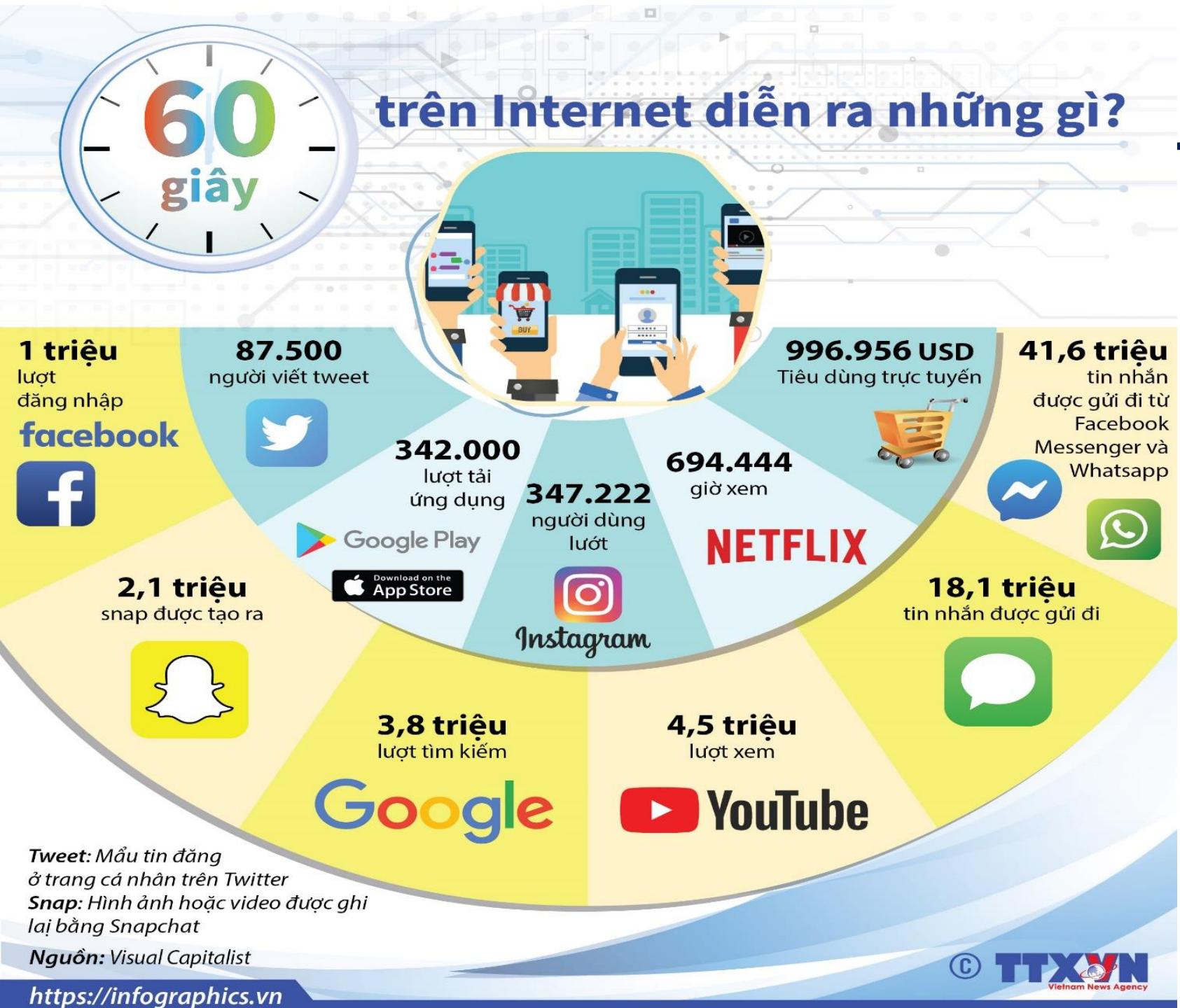
Đặc trưng của dữ liệu lớn:

- 5 V

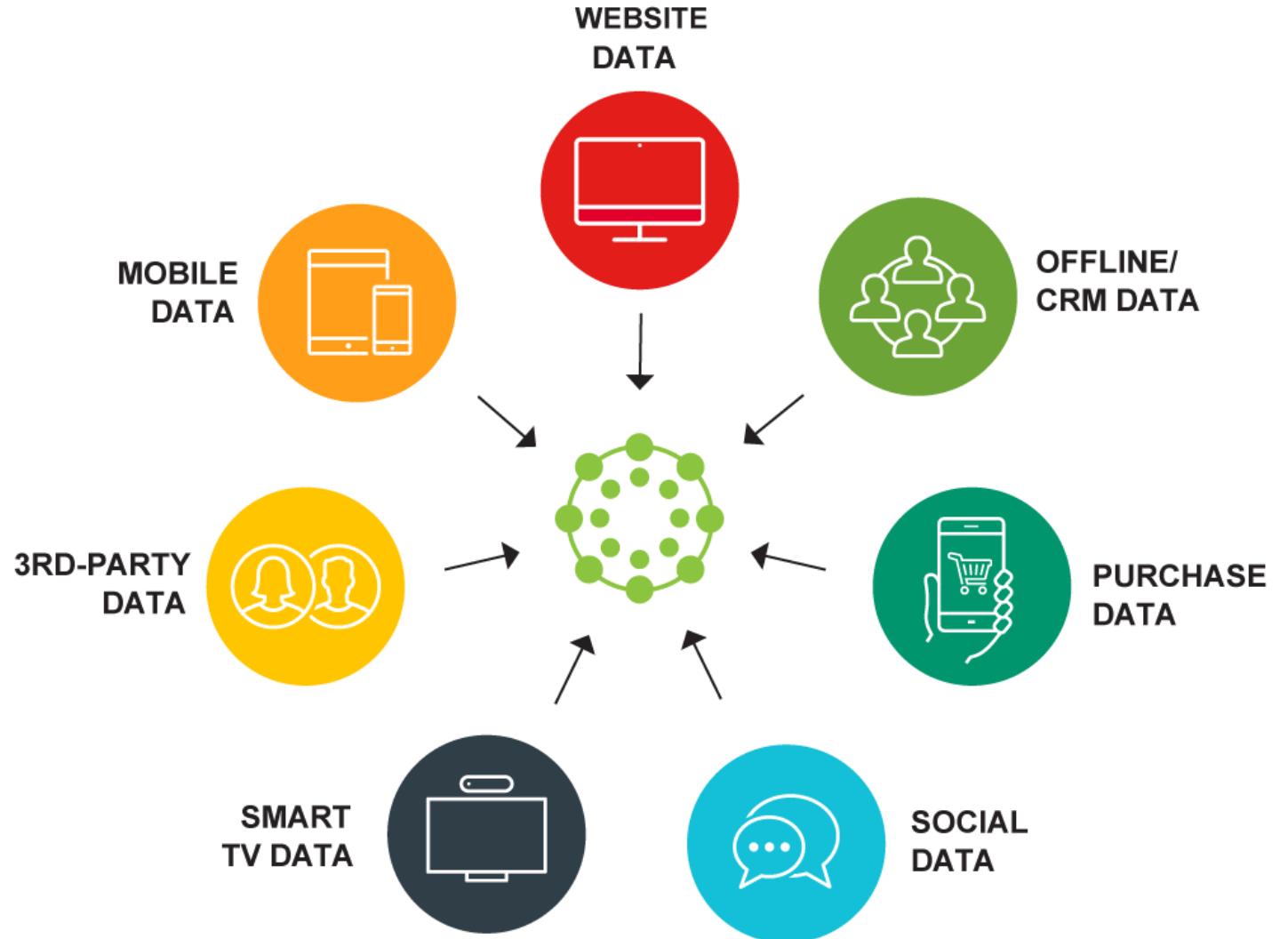
- Khối lượng (**V**olume): lượng dữ liệu được tạo ra
- Tốc độ (**V**elocity): Tốc độ dữ liệu được tạo ra và tốc độ chuyển đổi dữ liệu.
- Đa dạng (**V**ariety): Các kiểu dữ liệu được sử dụng
- Độ chính xác (**V**eracity): Độ tin cậy của dữ liệu
- Giá trị (**V**alue): Giá trị của dữ liệu



trên Internet diễn ra những gì?



Các nguồn dữ liệu tới từ đâu?



Các kiểu dữ liệu (Data type)

- Các kiểu dữ liệu trong khoa học dữ liệu
 - Dữ liệu có cấu trúc (Structured)
 - Dữ liệu phụ thuộc vào mô hình dữ liệu và nằm trong một trường cố định trong một bản ghi.
 - Lưu trữ trong cơ sở dữ liệu.

Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Int
214390830	Total (Age-adjusted)	2008	74.6%		73.8%
214390833	Aged 18-44 years	2008	59.4%		58.0%
214390831	Aged 18-24 years	2008	37.4%		34.6%
214390832	Aged 25-44 years	2008	66.9%		65.5%
214390836	Aged 45-64 years	2008	88.6%		87.7%
214390834	Aged 45-54 years	2008	86.3%		85.1%
214390835	Aged 55-64 years	2008	91.5%		90.4%
214390840	Aged 65 years and over	2008	94.6%		93.8%
214390837	Aged 65-74 years	2008	93.6%		92.4%
214390838	Aged 75-84 years	2008	95.6%		94.4%
214390839	Aged 85 years and over	2008	96.0%		94.0%
214390841	Male (Age-adjusted)	2008	72.2%		71.1%
214390842	Female (Age-adjusted)	2008	76.8%		75.9%
214390843	White only (Age-adjusted)	2008	73.8%		72.9%
214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Các kiểu dữ liệu (Data type)

- Các kiểu dữ liệu trong khoa học dữ liệu

- Dữ liệu phi cấu trúc (Unstructured)

- Dữ liệu không phụ thuộc vào mô hình dữ liệu vì nội dung theo các ngữ cảnh, cách thức và ngôn ngữ khác nhau.
 - Email.



The screenshot shows an email inbox interface with a green header bar containing standard controls like back, forward, delete, move, and spam. Below the header, there are two email messages listed:

- New team of UI engineers**
- CDA@engineer.com** To xyz@program.com Today 10:21 ★

The body of the second message contains the following text:

An investment banking client of mine has had the go ahead to build a new team of UI engineers to work on various areas of a cutting-edge single-dealer trading platform.

They will be recruiting at all levels and paying between 40k & 85k (+ all the usual benefits of the banking world). I understand you may not be looking. I also understand you may be a contractor. Of the last 3 hires they brought into the team, two were contractors of 10 years who I honestly thought would never turn to what they considered "the dark side."

This is a genuine opportunity to work in an environment that's built up for best in industry and allows you to gain commercial experience with all the latest tools, tech, and processes.

There is more information below. I appreciate the spec is rather loose – They are not looking for specialists in Angular / Node / Backbone or any of the other buzz words in particular, rather an "engineer" who can wear many hats and is in touch with current tech & tinkers in their own time.

For more information and a confidential chat, please drop me a reply email. Appreciate you may not have an updated CV, but if you do that would be handy to have a look through if you don't mind sending.

At the bottom of the email view, there are standard reply, reply to all, and forward buttons.

Các kiểu dữ liệu (Data type)

- **Các kiểu dữ liệu trong khoa học dữ liệu**

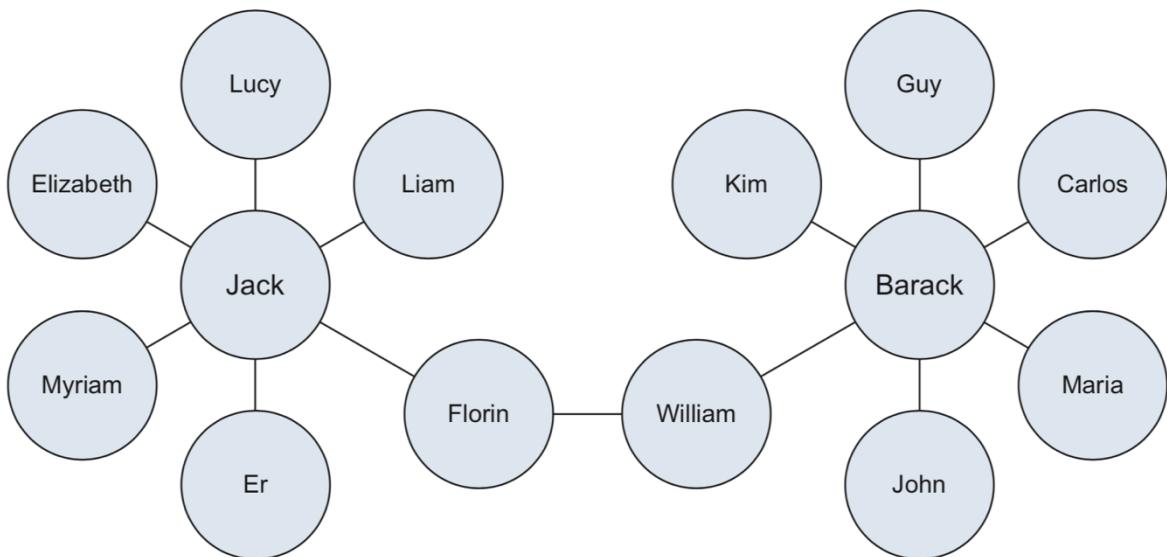
- **Dữ liệu thiết lập từ máy tính (Machine-generated)**

- Thông tin được thiết lập tự động bởi máy tính, các quá trình, ứng dụng hoặc các loại máy móc khác mà không cần sự can thiệp của con người.
- Nguồn thiết lập dữ liệu chính (The Internet of Things).
- Yêu cầu các công cụ và kỹ thuật xử lý mạnh mẽ.
- Nhật ký máy chủ web, bản ghi chi tiết cuộc gọi, nhật ký sự kiện mạng.

CSIPERF:TXCOMMIT;313236 2014-11-28 11:36:13, Info 69), objectname [6]"(null)" 2014-11-28 11:36:13, Info result 0x00000000, handle @0x4e54 2014-11-28 11:36:13, Info Beginning NT transaction commit... 2014-11-28 11:36:13, Info trace: CSIPERF:TXCOMMIT;273983 2014-11-28 11:36:13, Info 70), objectname [6]"(null)" 2014-11-28 11:36:13, Info result 0x00000000, handle @0x4e5c 2014-11-28 11:36:13, Info Beginning NT transaction commit... 2014-11-28 11:36:14, Info trace: CSIPERF:TXCOMMIT;386259 2014-11-28 11:36:14, Info 71), objectname [6]"(null)" 2014-11-28 11:36:14, Info result 0x00000000, handle @0x4e5c 2014-11-28 11:36:14, Info Beginning NT transaction commit... 2014-11-28 11:36:14, Info trace: CSIPERF:TXCOMMIT;375581	CSI	00000153 Creating NT transaction (seq 69) 00000154 Created NT transaction (seq 69) 00000155@2014/11/28:10:36:13.471 00000156@2014/11/28:10:36:13.705 CSI perf 00000157 Creating NT transaction (seq 70) 00000158 Created NT transaction (seq 70) 00000159@2014/11/28:10:36:13.764 0000015a@2014/11/28:10:36:14.094 CSI perf 0000015b Creating NT transaction (seq 71) 0000015c Created NT transaction (seq 71) 0000015d@2014/11/28:10:36:14.106 0000015e@2014/11/28:10:36:14.428 CSI perf
--	-----	---

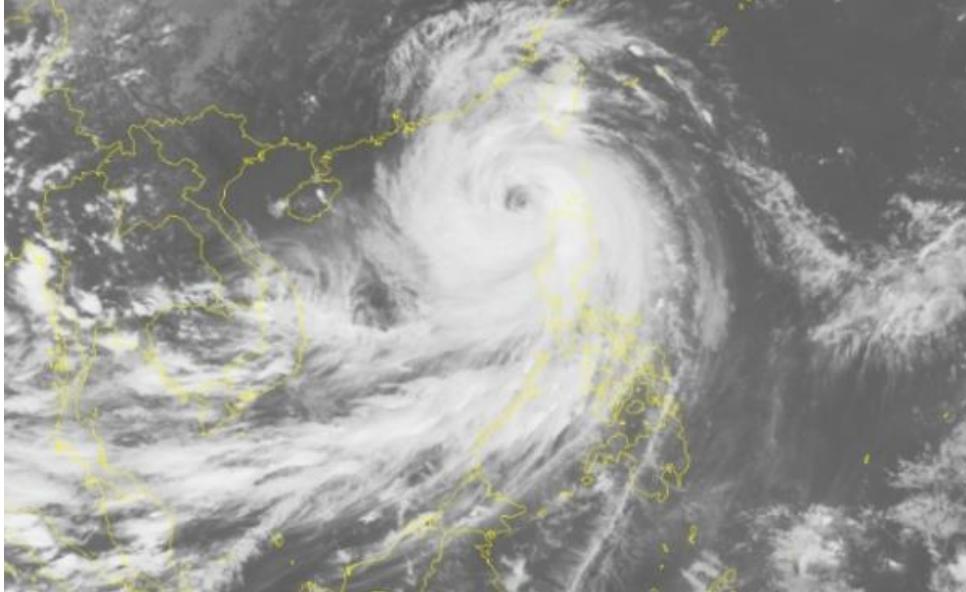
Các kiểu dữ liệu (Data type)

- Các kiểu dữ liệu trong khoa học dữ liệu
 - Dữ liệu dạng đồ thị hoặc mạng lưới (Graph-based or Network data)
 - Dữ liệu phụ tập trung vào mối quan hệ hoặc tính phụ thuộc của các đối tượng.
 - Sử dụng các nút, cạnh và thuộc tính để biểu diễn và lưu trữ dữ liệu đồ thị.
 - Dữ liệu dựa trên đồ thị là một cách tự nhiên để thể hiện các mạng xã hội và cấu trúc của nó. **LinkedIn Twitter Facebook**



Các kiểu dữ liệu (Data type)

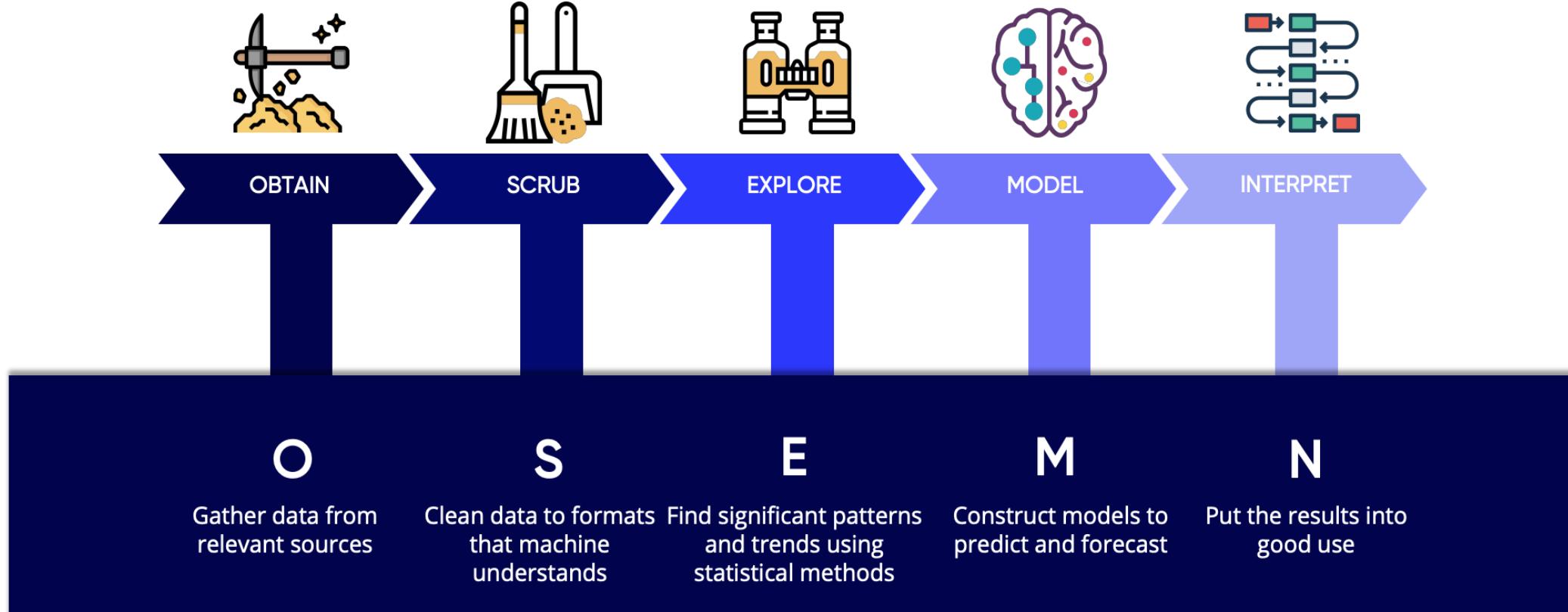
- **Các kiểu dữ liệu trong khoa học dữ liệu**
 - **Dữ liệu âm thanh, hình ảnh và video (Audio, Image, Video)**
 - Thách thức đối với các nhà khoa học dữ liệu.
 - Tự động nhận dạng đối tượng cụ thể qua âm thanh, hình ảnh.
 - Tính toán chuyển động của đối tượng theo thời gian thực.
 - Thu nhận thông tin thông qua các video sử dụng thuật toán Deep Learning.



Mối quan hệ giữa Dữ liệu lớn và KHDL



4. Quy trình của một dự án về Khoa học dữ liệu



Quy trình tổng quát

1. Setting the research goal

Thiết lập các mục tiêu nghiên cứu (Xác định bài toán cần giải quyết)

2. Data Collection

Thu thập tất cả các dữ liệu từ các nguồn khác nhau liên quan đến bài toán

3. Data Preparation

Sử dụng các công cụ kỹ thuật để chuẩn bị dữ liệu (Data cleaning, Scale...)

Trình bày kết quả cuối cùng

6. Presentation & Automation

Mô hình hóa dữ liệu (ML, DL...)

5. Data Modelling

Phân tích và khám phá dữ liệu

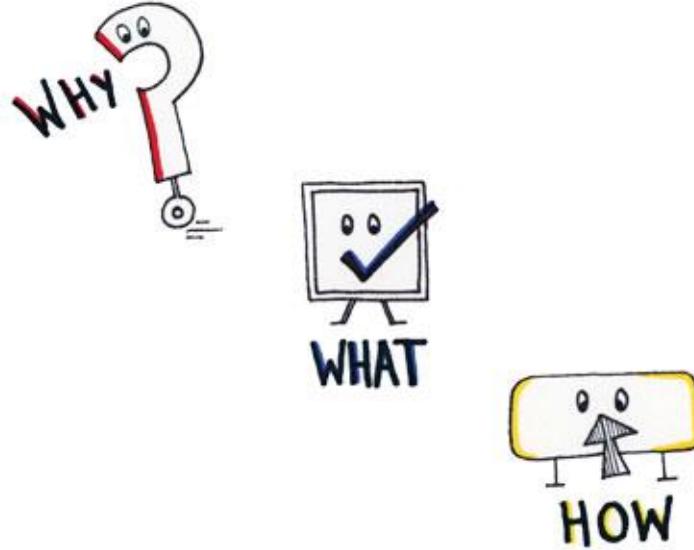
4. Data Exploration & Analysis



B1: Thiết lập mục tiêu nghiên cứu

- **Setting the research goal**

- Một dự án data science bắt đầu bằng việc tìm hiểu xác định mục tiêu (what), lý do cần thực hiện (why) và thực hiện nó như thế nào (how)



- Trả lời ba câu hỏi này là mục tiêu của giai đoạn đầu tiên, để mọi người biết phải làm gì và có thể đồng ý về hướng hành động tốt nhất.

B1: Thiết lập mục tiêu nghiên cứu (2)

- Kết quả của quá trình này bao gồm:
 - Mục tiêu nghiên cứu
 - Hiểu nhiệm vụ và bối cảnh dự án
 - Dự án cần dữ liệu gì? Lấy như thế nào? Ở đâu? Số lượng như thế nào?
 - Tài nguyên nào bạn muốn sử dụng (nhân lực, thời gian)
 - Sản phẩm của dự án là gì, sẽ được sử dụng ở đâu?
 - Đâu là thước đo dự án có thành công hay không?
 - Một kế hoạch hành động với thời gian biểu (timetable)

B1: Thiết lập mục tiêu nghiên cứu (3)

- **Mục tiêu nghiên cứu**
 - Mục tiêu nghiên cứu có thể được bắt nguồn từ nhu cầu hoặc nhiệm vụ từ công ty, xã hội hoặc phát sinh trong quá trình thực hiện một dự án
 - Tìm danh sách khách hàng tiềm năng đang cần vay vốn
 - Từ tag người vào ảnh có trước (facebook)
 - Dự đoán khả năng bị bệnh của một người
 - Hệ thống recommend mua hàng cho khách hàng (Amazon)

B1: Thiết lập mục tiêu nghiên cứu (4)

- **Ví dụ**
 - Mục tiêu: dự đoán giá viên kim cương 1.35 carats
 - Tìm hiểu về ngành công nghiệp kim cương



B2: Thu thập dữ liệu

- **Thu thập dữ liệu (Data collection)**
 - Mục tiêu: thu thập tất cả các dữ liệu cần cho dự án
 - Dữ liệu có thể được lưu trữ ở nhiều dạng, từ các tệp văn bản đơn giản đến các bảng trong cơ sở dữ liệu
 - **Thu thập giá kim cương tại các cửa hàng bán lẻ**



B3. Chuẩn bị dữ liệu

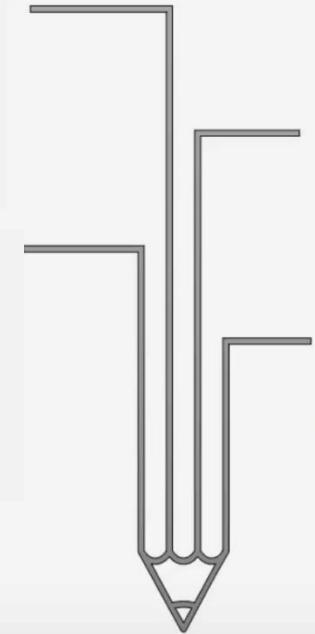
Là bước **quan trọng**, chiếm **nhiều thời gian** và **nguồn lực** nhất trong bất kỳ một dự án khoa học dữ liệu nào (80%)

Làm sạch dữ liệu

Chỉnh sửa dữ liệu bằng cách bổ sung các dữ liệu còn thiếu, thay thế và hiệu chỉnh các dữ liệu nhiễu

Giảm kích thước dữ liệu

Đảm bảo chất lượng ban đầu của dữ liệu



Chuyển đổi dữ liệu

Bao gồm chuẩn hóa, chuyển đổi và tổng hợp dữ liệu sử dụng các phương pháp ETL.

Tích hợp dữ liệu

Xử lý sự không tương thích giữa các dữ liệu



80%

of time & resources spent
on any data project is
data preparation*

B3: Chuẩn bị dữ liệu (2)

- Chuẩn bị dữ liệu

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	
0.6	4172
Two	21764
1.1	4682
1.31	6171

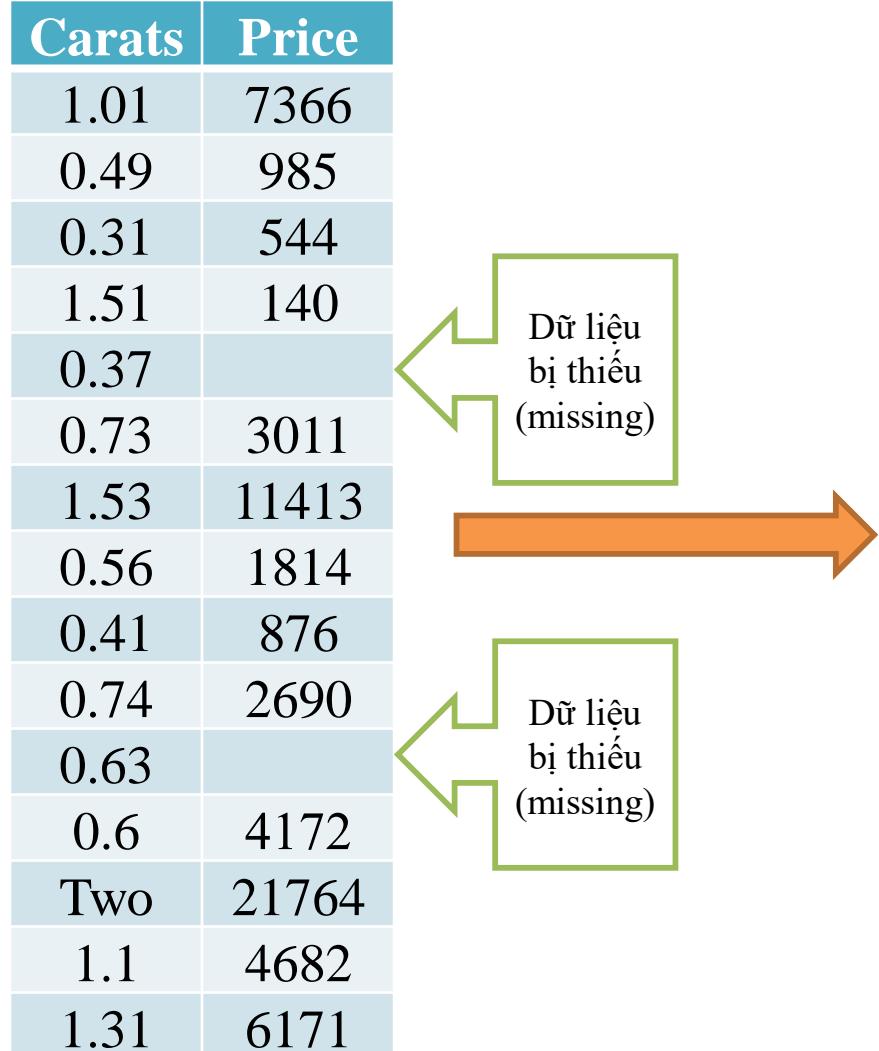
Dữ liệu
không hợp lệ
(Improper)

Dữ liệu
bị thiếu
(missing)

Dữ liệu
bị thiếu
(missing)

B3: Chuẩn bị dữ liệu (4)

- Chuẩn bị dữ liệu

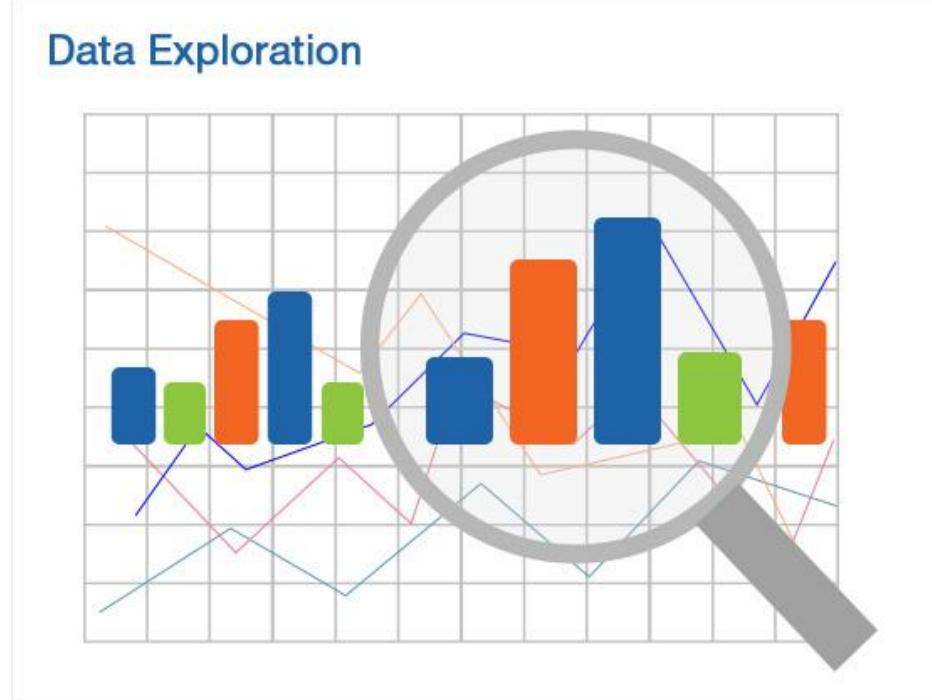


The diagram illustrates the transformation of a dataset from a blue table to an orange table. The blue table contains 15 rows of data with columns 'Carats' and 'Price'. The orange table contains 15 rows of data with columns 'Carats' and 'Price'. Two green arrows point from a box labeled 'Dữ liệu không hợp lệ (Improper)' to the 'Two' entry in the blue table's 'Carats' column. A green box labeled 'Dữ liệu bị thiếu (missing)' points to the empty cell in the blue table's 'Price' column at row 5. An orange arrow points from the blue table to the orange table.

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	
0.6	4172
Two	21764
1.1	4682
1.31	6171

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	693
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	4325
0.6	4172
Two	21764
1.1	4682
1.31	6171

B4: Khám phá dữ liệu



- Khám phá dữ liệu để **hiểu rõ** hơn về **mối quan hệ giữa các biến** và nhận biết được các **thông tin** được truyền tải **từ dữ liệu**.
- Lựa chọn các **mô hình phù hợp** và các **biến quan trọng** để đưa vào mô hình.

B4: Khám phá dữ liệu (2)

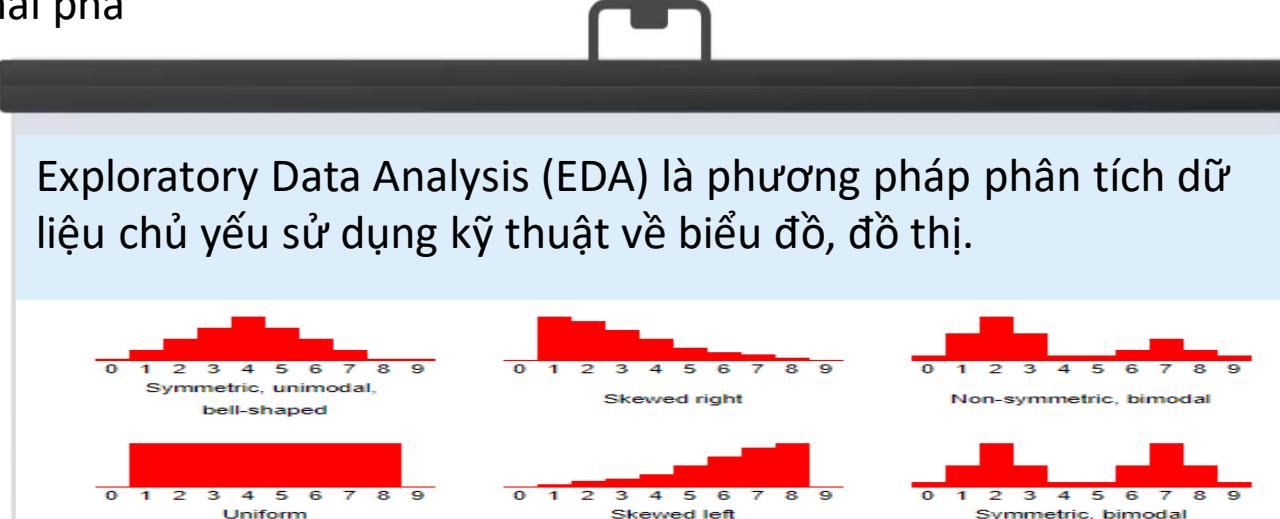
- Thiết kế mô hình

Phân tích dữ liệu khai phá

But what is Exploratory Data Analysis?



Exploratory Data Analysis (EDA) là phương pháp phân tích dữ liệu chủ yếu sử dụng kỹ thuật về biểu đồ, đồ thị.

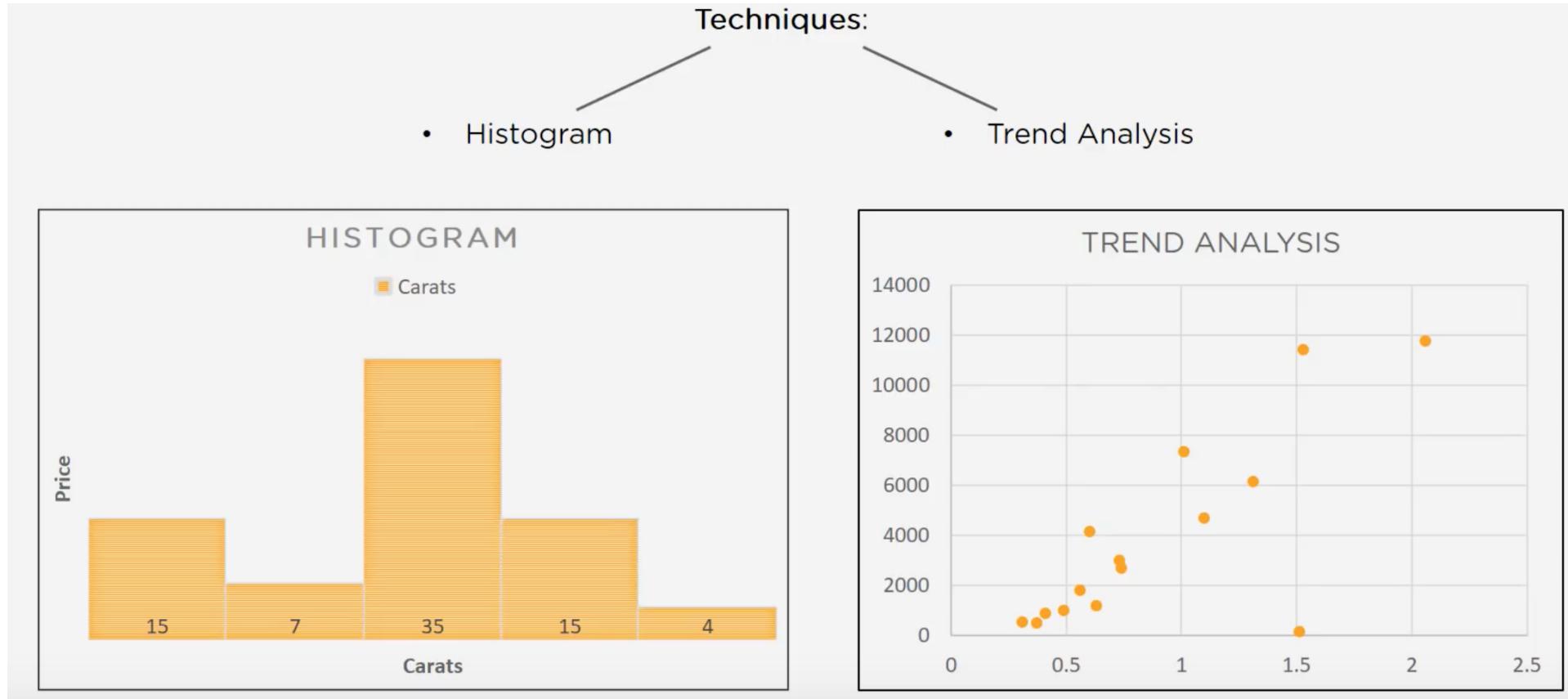


Mục tiêu:

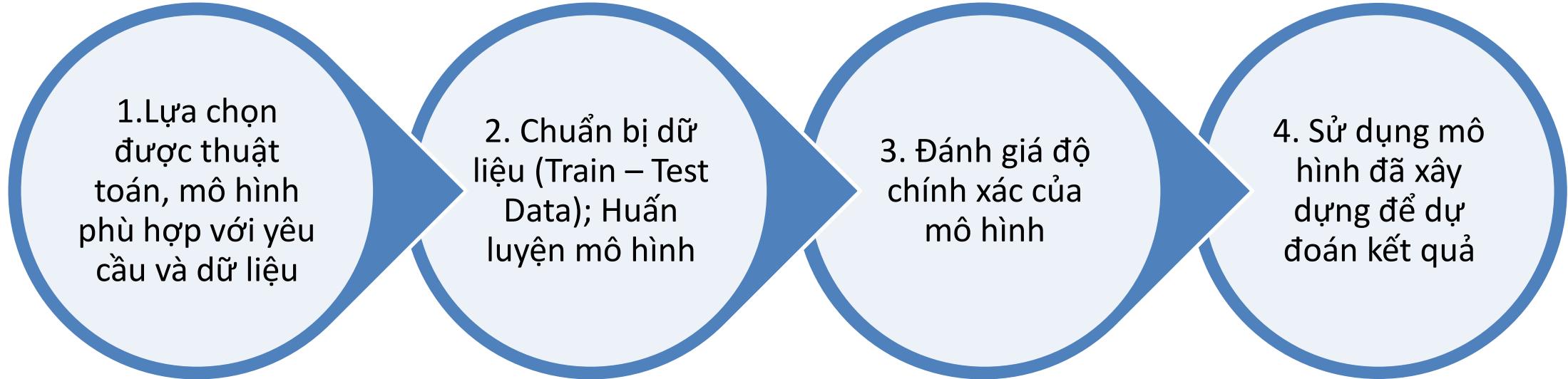
- Hiểu về kiểu dữ liệu và trả lời các câu hỏi về dữ liệu.
- Hiểu rõ cách thức phân phối dữ liệu
- Xác định các trường hợp ngoại lệ (errors)
- Xác định các quy luật có trong dữ liệu (pattern)

B4: Khám phá dữ liệu (3)

- Phân tích khám phá dữ liệu EDA



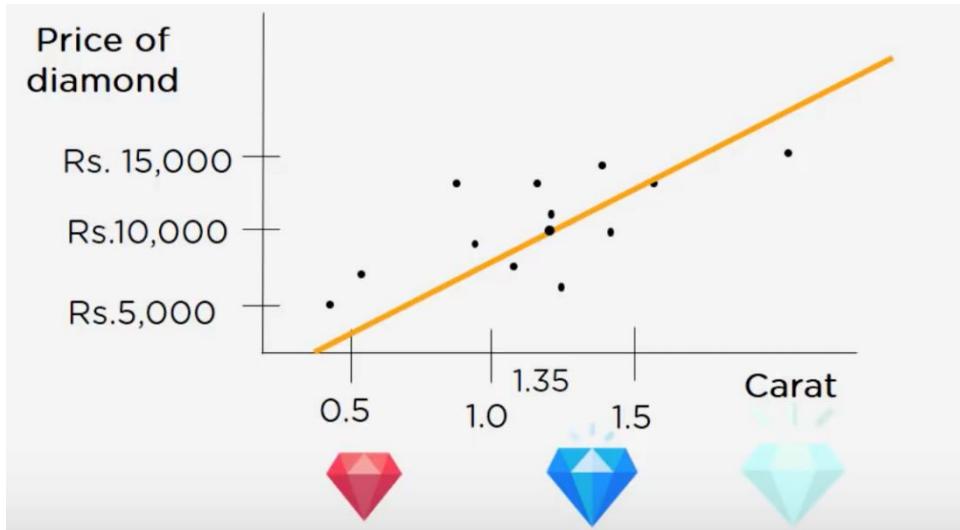
B5: Xây dựng mô hình



B5: Xây dựng mô hình (2)

Lựa chọn thuật toán, mô hình:

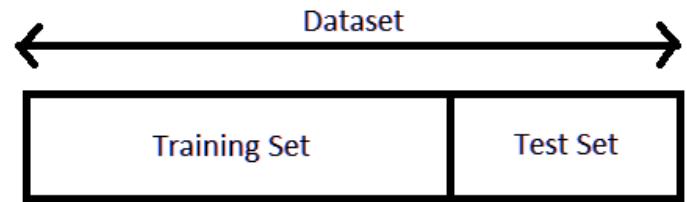
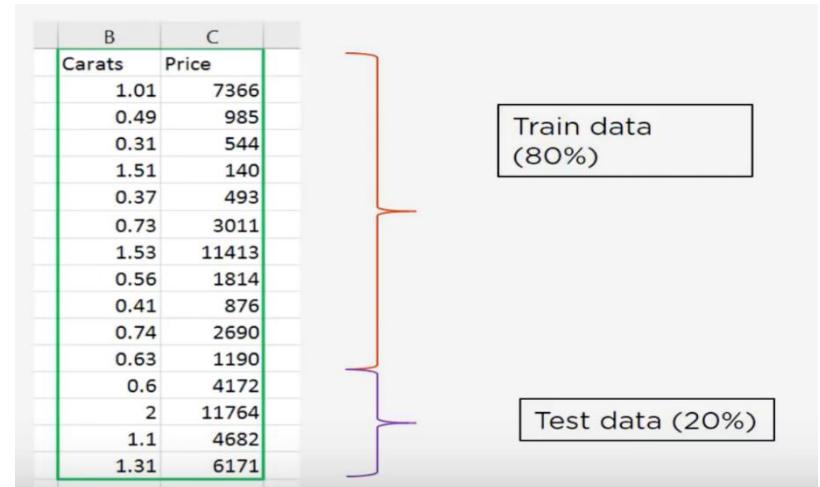
- Dựa vào dữ liệu của bài toán cụ thể, để lựa chọn được thuật toán phù hợp.
 - Với bài toán dự đoán giá Kim cương: kết quả phân tích tiến triển tuyến tính. Do đó thuật toán **Hồi quy tuyến tính** được lựa chọn để xây dựng mô hình trong trường hợp này.
 - Dữ liệu mẫu (train data) được sử dụng để chạy mô hình.



B5: Xây dựng mô hình (2)

Chuẩn bị dữ liệu cho huấn luyện và kiểm thử mô hình:

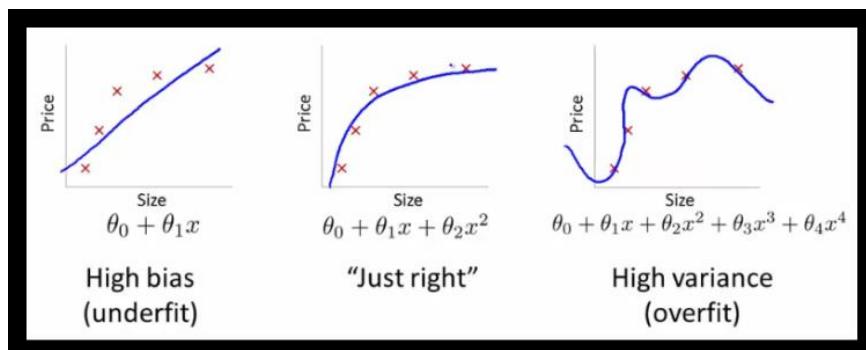
- Phân tách tập dữ liệu cho việc huấn luyện và kiểm thử mô hình. Thông thường tập dữ liệu sẽ được chia theo tỷ lệ **80:20** (80% dữ liệu được sử dụng để huấn luyện mô hình, 20% được sử dụng để kiểm thử mô hình)

The table shows the relationship between Carats (B) and Price (C). A green bracket on the left indicates the "Train data (80%)", and a purple bracket on the right indicates the "Test data (20%)".

B	C
1.01	7366
0.49	985
0.31	544
1.51	140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2	11764
1.1	4682
1.31	6171

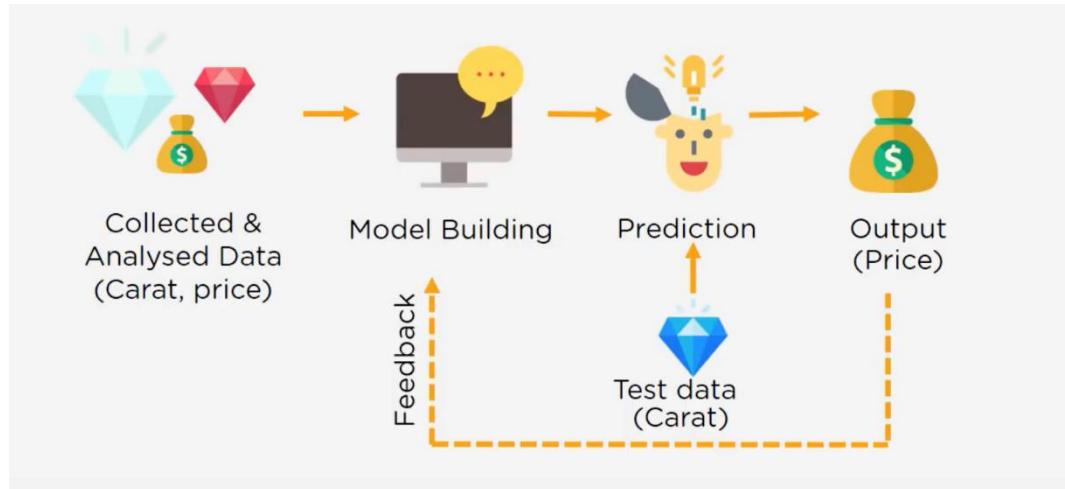
Huấn luyện mô hình:



B5: Xây dựng mô hình (2)

Đánh giá mô hình:

- Sử dụng tập dữ liệu kiểm tra (Test Data), đánh giá độ chính xác của mô hình đã xây dựng



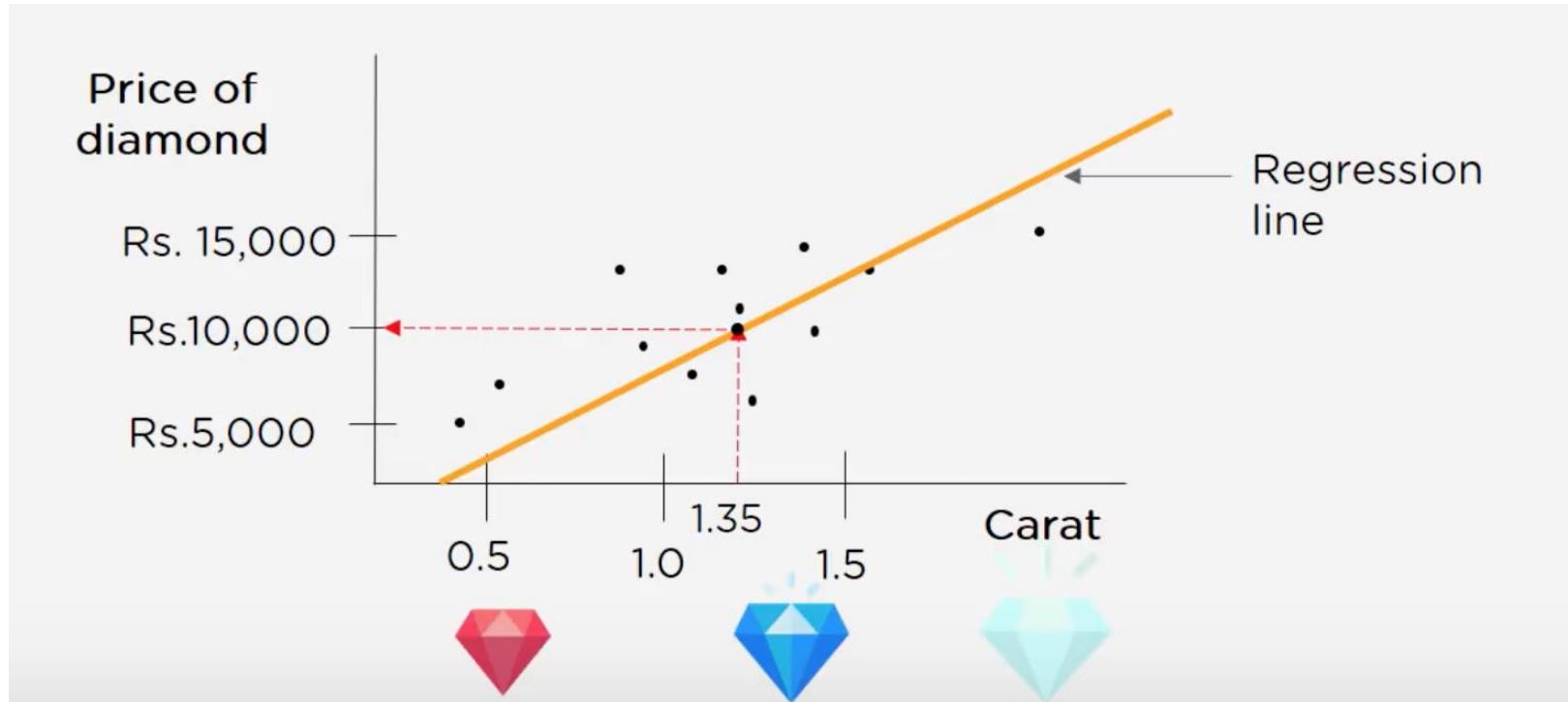
Classification report is as:

	precision	recall	f1-score	support
0	0.95	0.91	0.93	64
1	0.95	0.97	0.96	107
avg / total	0.95	0.95	0.95	171
Accuracy of model is 0.9473684210526315				

B5: Xây dựng mô hình (2)

Chạy mô hình với dữ liệu mới và đưa ra các dự đoán:

- Kết quả dự đoán viên kim cương 1.35 carat có giá 10.000 với mô hình được xây dựng dựa vào thuật toán hồi quy tuyến tính.



B6: Trình diễn kết quả



Trình bày kết quả của dự án với khách hàng, công ty



Tích hợp với các công cụ (ứng dụng) khác

Một số lưu ý:

Không phải mọi dự án khoa học dữ liệu (Data science) sẽ tuân theo một quy trình giống nhau.

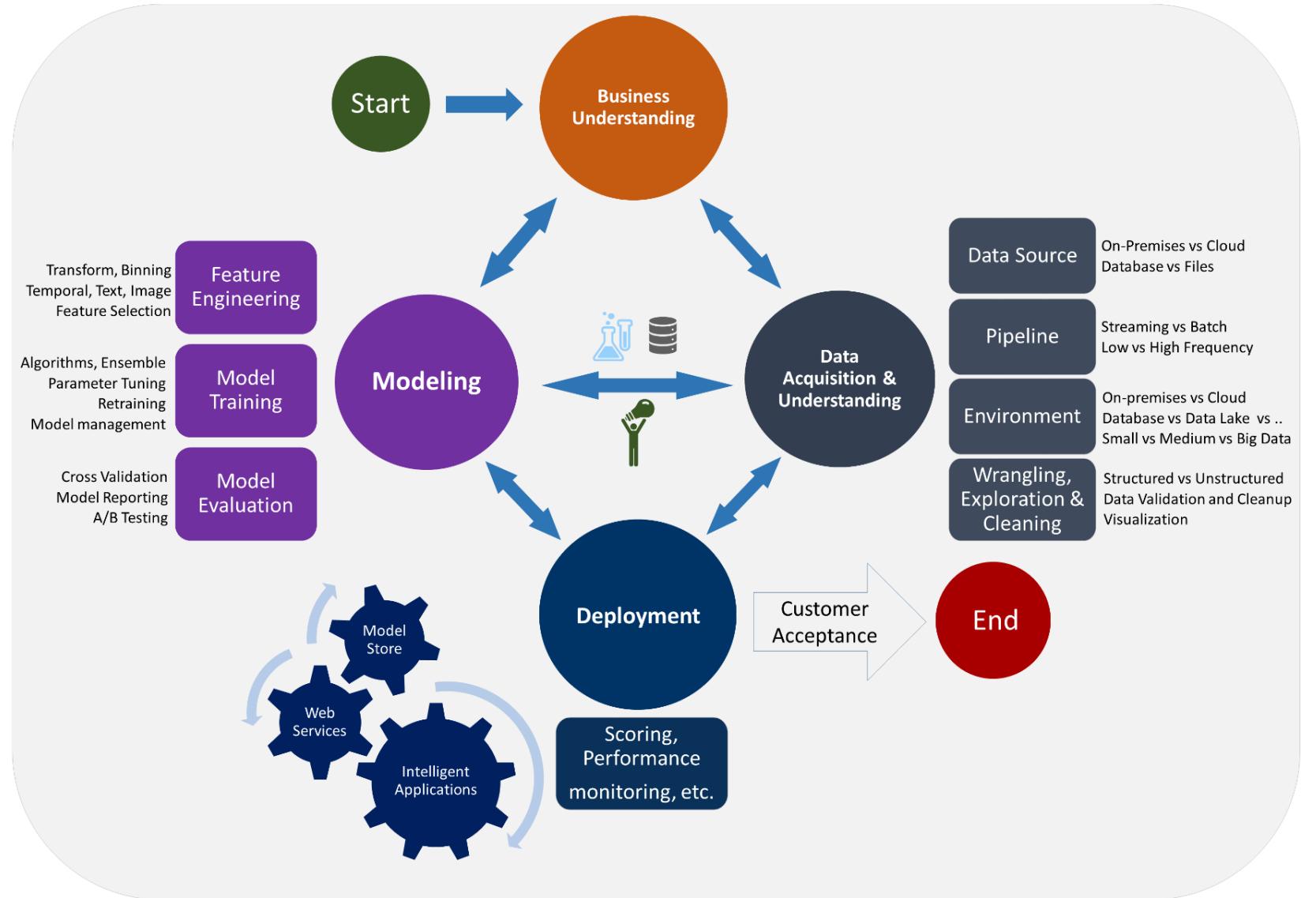


Các bước cụ thể trong các dự án khác nhau có thể khác nhau đôi chút

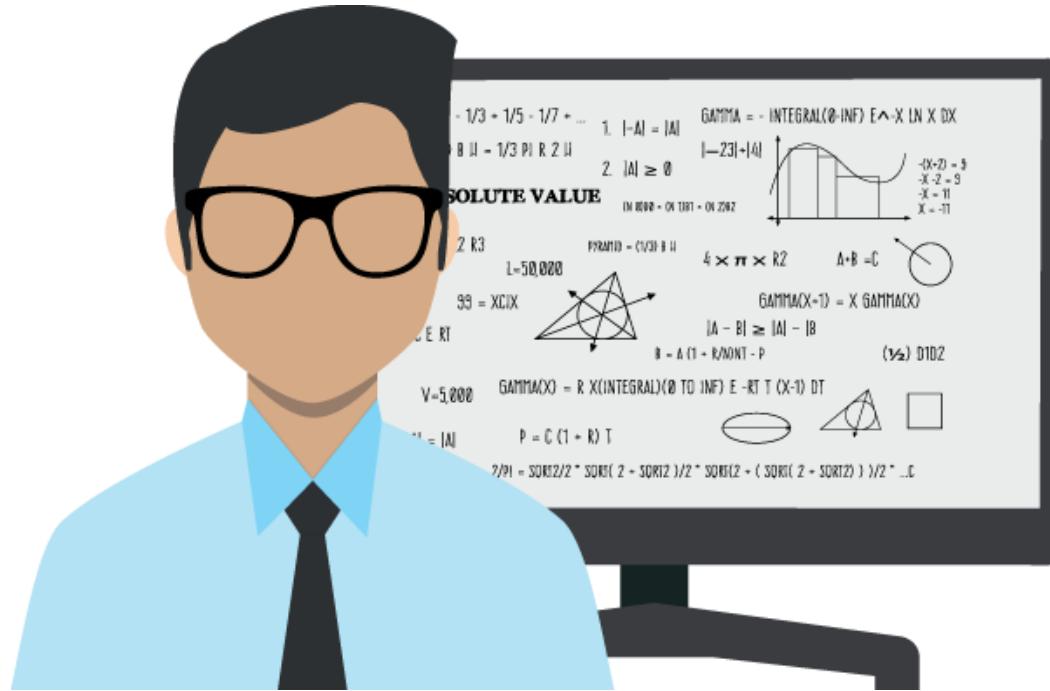


Các bước này phụ thuộc vào nhà khoa học dữ liệu, công ty và các yếu tố khác của dự án.

Vòng đời của Khoa học dữ liệu

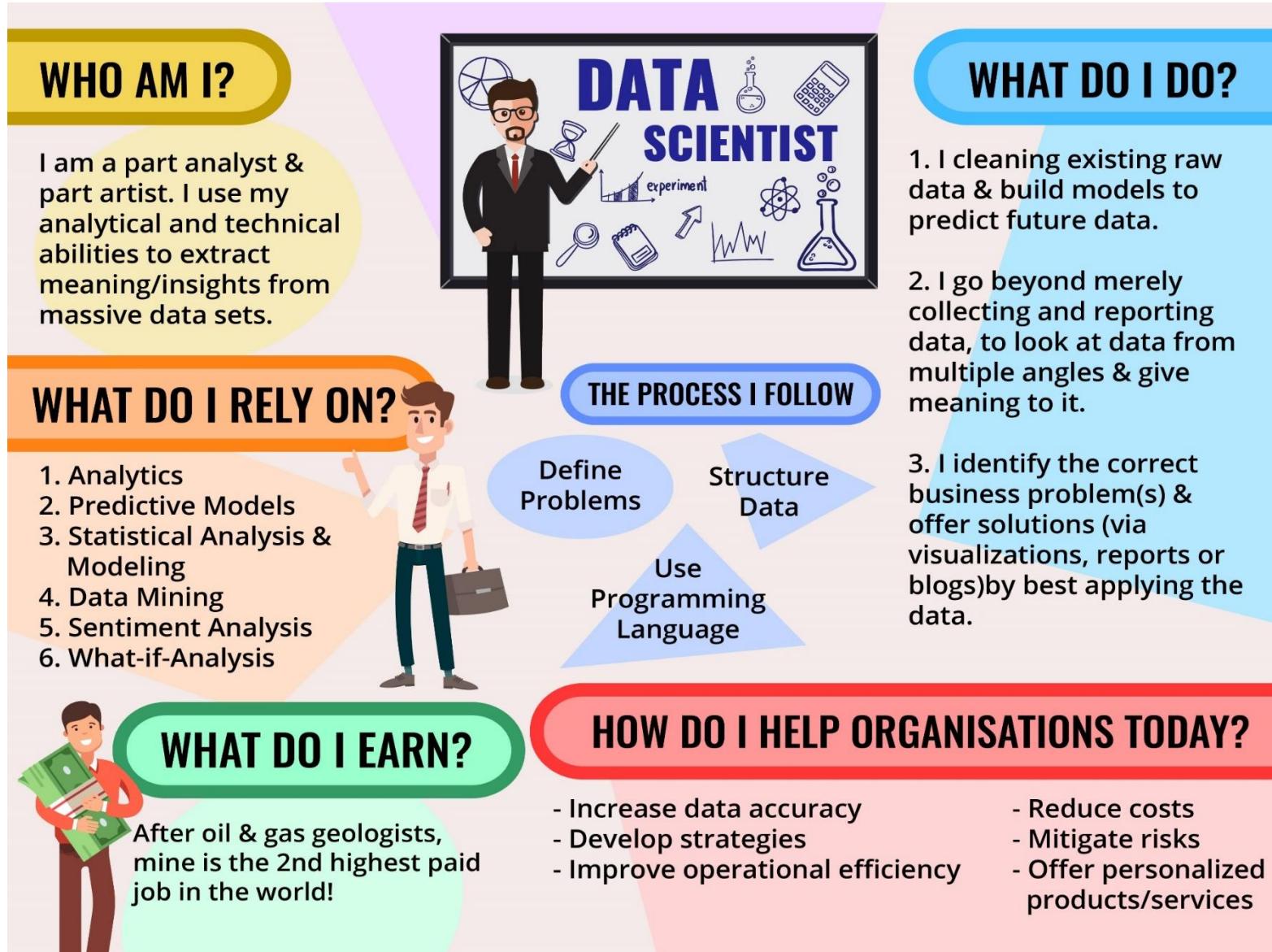


5. Nhà khoa học dữ liệu (Data Scientist)

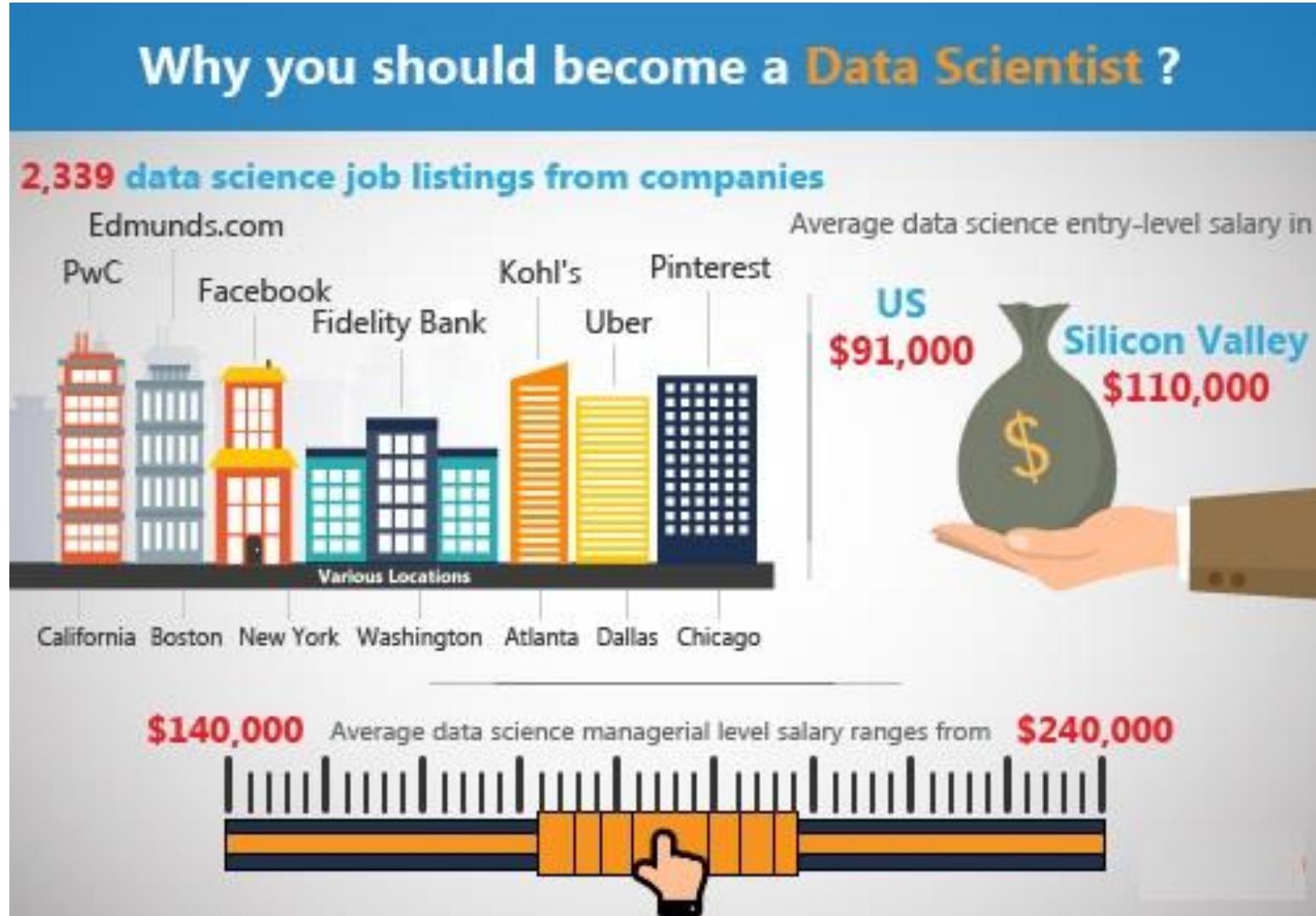


**Kiến thức, kỹ năng và cơ hội việc làm
 của một nhà Khoa học dữ liệu.**

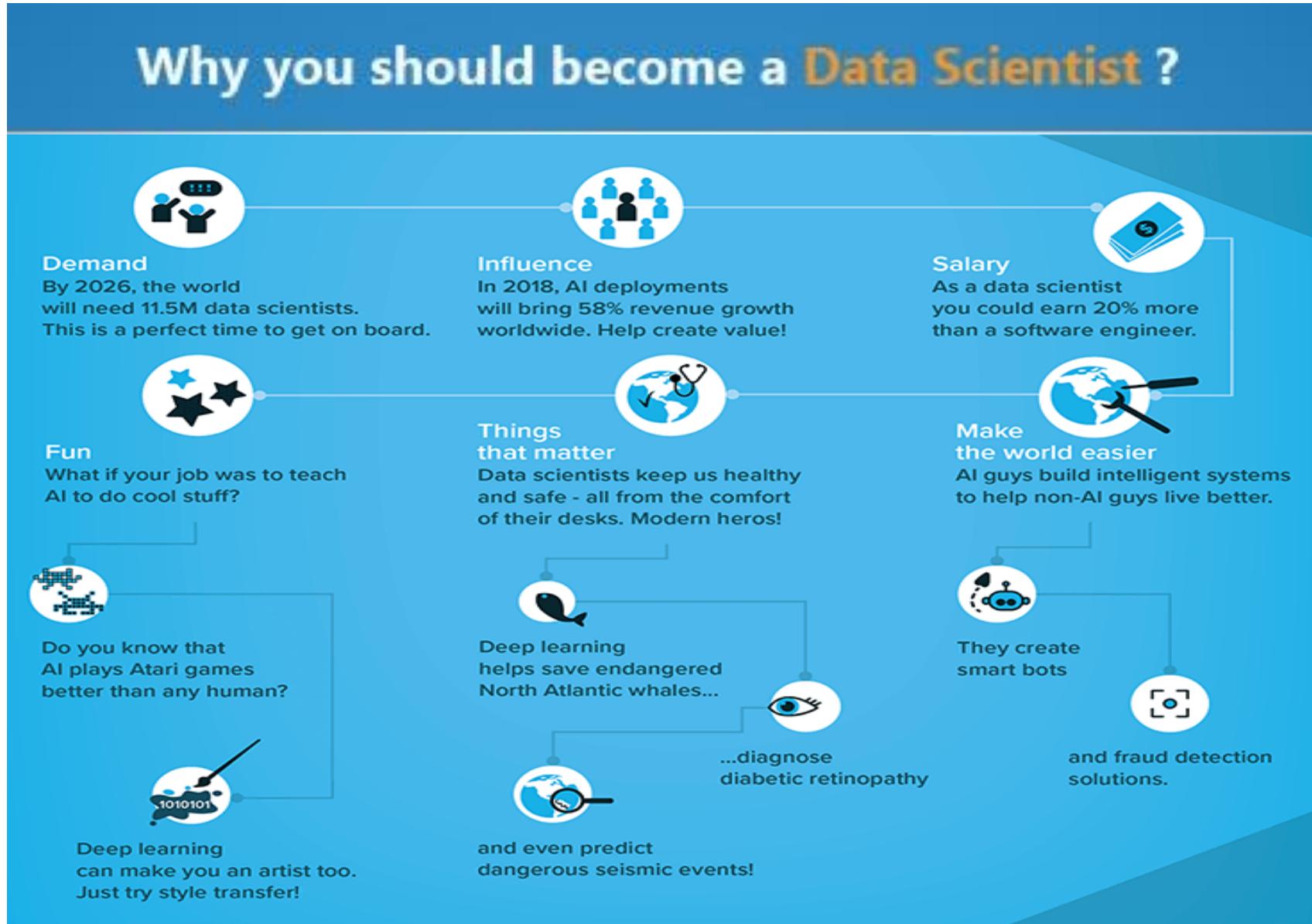
Nhà khoa học dữ liệu?



Tại sao bạn nên trở thành một nhà khoa học dữ liệu?



Tại sao bạn nên trở thành một nhà khoa học dữ liệu?



Phân biệt:

- Phân biệt nhà Khoa học dữ liệu/ Phân tích dữ liệu/ Kỹ sư dữ liệu

Data Scientist

- Nhà khoa học dữ liệu là một người sử dụng trình độ kỹ thuật dữ liệu tiên tiến để đưa ra các quyết định chiến lược.
- Họ là những người có vị trí cao nhất trong nhóm có kiến thức chuyên sâu về thống kê, thao tác dữ liệu và học máy.
- Dựa trên các sản phẩm của các Kỹ sư dữ liệu và các nhà phân tích dữ liệu để đưa những giá trị và phương hướng hành động cụ thể cho doanh nghiệp.

Data Analyst

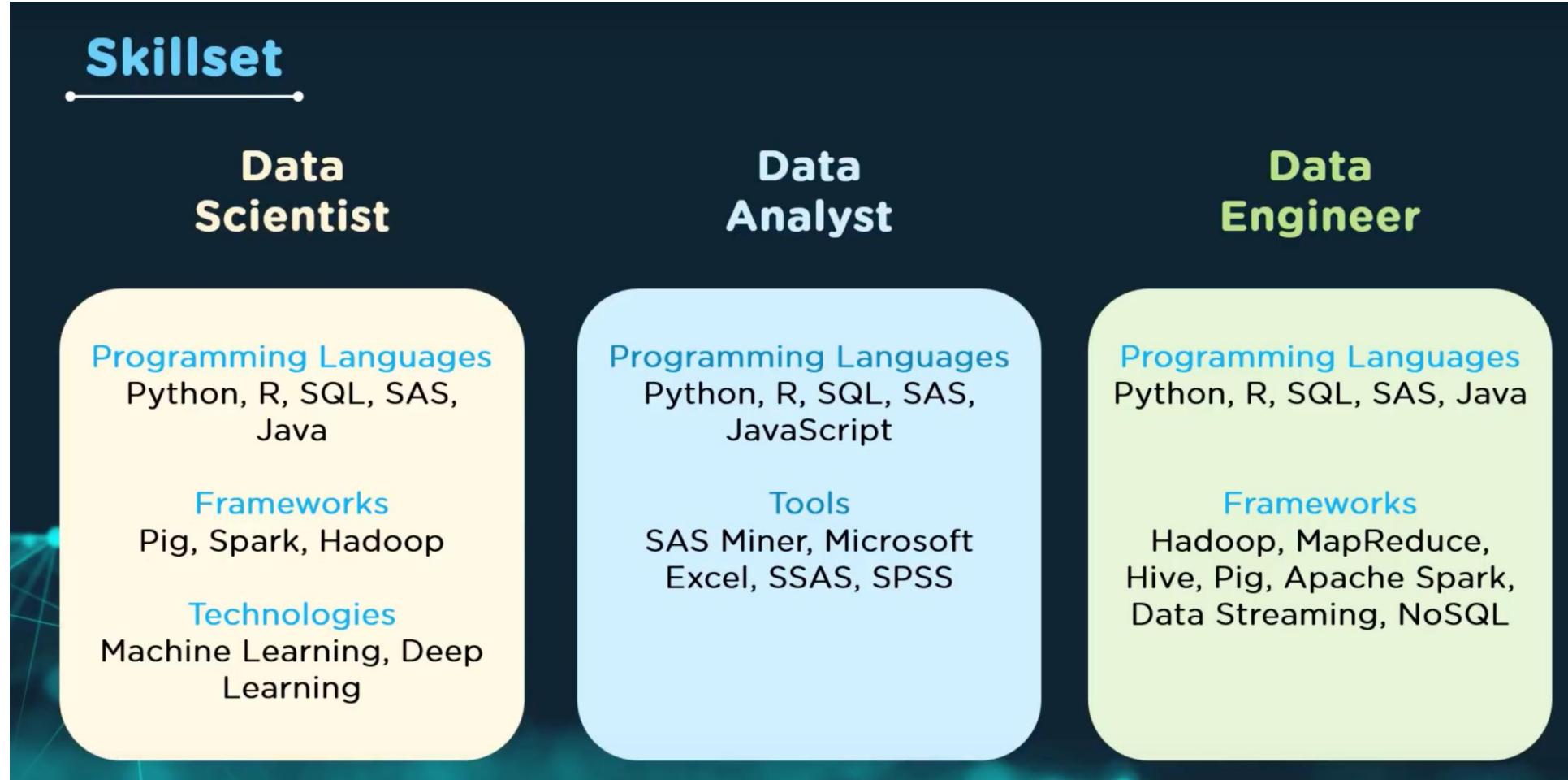
- Nhà phân tích dữ liệu giữ vị trí thấp nhất trong nhóm phân tích dữ liệu.
- Họ cần nắm vững các kỹ thuật về xử lý dữ liệu, mô hình hóa, xây dựng báo cáo.
- Các nhà phân tích dữ liệu có thể trở thành các nhà Khoa học dữ liệu hoặc Kỹ sư dữ liệu khi có nhiều kinh nghiệm trong lĩnh vực này.

Data Engineer

- Một Kỹ sư dữ liệu có vị trí trung gian giữa nhà phân tích dữ liệu và nhà khoa học dữ liệu.
- Họ cần có các kiến thức chuyên môn trong phát triển, xây dựng và bảo trì kiến trúc của hệ thống.
- Họ làm việc với các dữ liệu lớn và gửi báo cáo kết quả cho các nhà khoa học dữ liệu để thực hiện phân tích đánh giá.

Phân biệt

- Phân biệt nhà Khoa học dữ liệu/ Phân tích dữ liệu/ Kỹ sư dữ liệu



Phân biệt:

- Phân biệt nhà Khoa học dữ liệu/ Phân tích dữ liệu/ Kỹ sư dữ liệu

Roles and Responsibilities

Data Scientist

- Làm sạch và khai phá dữ liệu, xử lý các dữ liệu phi cấu trúc.
- Thiết kế mô hình để làm việc với dữ liệu lớn.
- Suy luận và giải thích các phân tích dựa trên dữ liệu lớn.
- Lãnh đạo nhóm phân tích dữ liệu để đạt được các mục tiêu đặt ra.
- Đưa ra các kết luận có ảnh hưởng trực tiếp tới các hoạt động kinh doanh.

Data Analyst

- Thu thập thông tin từ cơ sở dữ liệu thông qua truy vấn
- Xử lý dữ liệu và cung cấp các báo cáo tóm tắt sử dụng các thuật toán cơ bản trong công việc.
- Có kỹ năng cốt lõi về thống kê, khai phá dữ liệu, trộn dữ liệu, trực quan hóa dữ liệu và phân tích dữ liệu khai phá

Data Engineer

- Khai phá dữ liệu phục vụ trích xuất thông tin.
- Chuyển đổi các dữ liệu lõi sang các dạng phù hợp cho các phân tích tiếp theo.
- Xây dựng truy vấn dữ liệu.
- Bảo trì thiết kế và kiến trúc dữ liệu.
- Phát triển các kho dữ liệu lớn

Nhà khoa học dữ liệu

- **Đặc trưng của một nhà Khoa học dữ liệu**



Sự ham hiểu biết

Đặt câu hỏi để hiểu rõ vấn đề. Tò mò muốn khám phá những gì ẩn giấu bên trong.



Khả năng phán đoán

Xác định cách thức mới để giải quyết vấn đề và chỉ rõ các tiêu chí quan trọng.



Kỹ năng giao tiếp

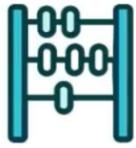
Trao đổi để truyền tải giá trị thu nhận được với người khác

Một nhà khoa học dữ liệu cần gì?

Mô hình toán học giúp tăng hiệu suất tính toán và dự đoán

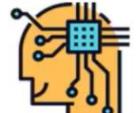
2

MATHEMATICAL MODELLING



1

MACHINE LEARNING



ML là công cụ quan trọng trong KHDL, cung cấp cách thức giải quyết các vấn đề

Thống kê là nền tảng của KHDL, trích xuất kiến thức và thu nhận kết quả phân tích từ dữ liệu



Kiến thức, kỹ năng về lập trình để xây dựng các chương trình trích xuất, mô hình hóa dữ liệu

4

COMPUTER PROGRAMMING



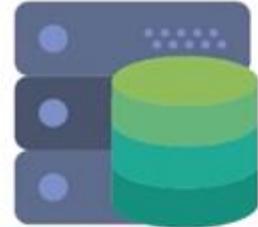
5

DATABASES



Truy vấn trong CSDL giúp đặt ra các câu hỏi chính xác để giải quyết các vấn đề đặt ra trong KHDL

Những kiến thức cần trang bị?



Kiến thức
CSDL



Thống kê



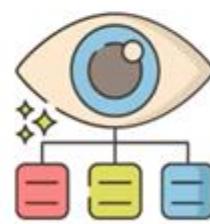
Công cụ
lập trình



Xử lý dữ liệu



Học máy



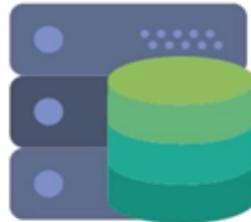
Trực quan hóa
dữ liệu



Dữ liệu lớn

Những kiến thức cần trang bị?

1 – Kiến thức về cơ sở dữ liệu



Kiến thức
CSDL

SQL (Structured Query Language - Ngôn ngữ truy vấn cấu trúc) là một ngôn ngữ cần thiết để trích xuất một lượng lớn thông tin từ tập dữ liệu. Kiến thức về SQL là bắt buộc cho một nhà khoa học dữ liệu

Công cụ cần thiết

ORACLE
DATABASE

MySQL

Microsoft®
SQL Server

TERADATA

Những kiến thức cần trang bị?

2 – Kiến thức về thống kê



Thống kê

Thống kê là một tập con của toán học liên quan đến việc thu thập, phân tích và thể hiện dữ liệu;
Nhà khoa học dữ liệu cần có hiểu biết về thống kê

Thống Kê

Xác Suất



Những kiến thức cần trang bị?

3 – Kiến thức về lập trình



Công cụ
lập trình

Thành thạo ít nhất một trong số những ngôn ngữ lập trình dưới đây là cần thiết cho việc phân tích dữ liệu của bất kỳ nhà khoa học dữ liệu nào



- R is a free software environment for statistical computing and graphics
- Supports most Machine Learning algorithms for Data Analytics like regression, association, clustering, etc.



- Python is an open-source general purpose programming language
- Python libraries like NumPy and SciPy are used in Data Science



- SAS can mine, alter, manage, and retrieve data from a variety of sources
- Can perform statistical analysis on the data

Những kiến thức cần trang bị?

4 – Kiến thức về thu thập và xử lý dữ liệu



Xử lý dữ liệu

Xử lý dữ liệu là quá trình chuyển đổi từ dữ liệu thô (raw data) thành định dạng phù hợp để làm cho nó hữu ích cho phân tích.

Bao gồm:

Làm sạch dữ liệu thô

Phân tích cấu trúc dữ liệu thô

Làm giàu dữ liệu thô

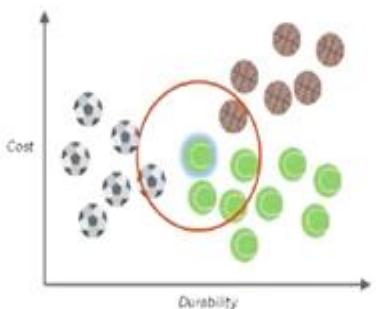
Những kiến thức cần trang bị?

5 – Kiến thức về Học máy

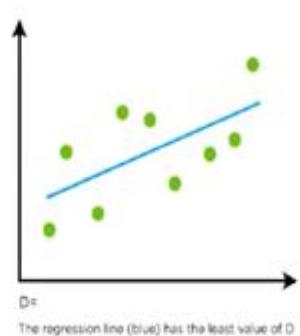


Học máy

Kiến thức về các kỹ thuật học máy như học có giám sát, cây quyết định, hồi quy tuyến tính, KNN... là hữu ích cho công việc của một nhà khoa học dữ liệu



KNN



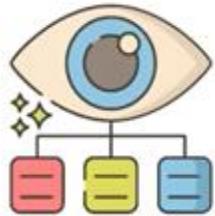
Linear Regression



Decision Tree

Những kiến thức cần trang bị?

6 – Kiến thức về trực quan hóa dữ liệu



Trực quan hóa
dữ liệu

Trực quan hóa dữ liệu là việc nghiên cứu và trình bày dữ liệu trực quan, thông qua các biểu đồ, hình vẽ... Đây là cách truyền tải thông tin một cách rõ ràng và hiệu quả nhất



Tableau



Power BI



Google Data Studio

Những kiến thức cần trang bị?

7 – Kiến thức về Dữ liệu lớn

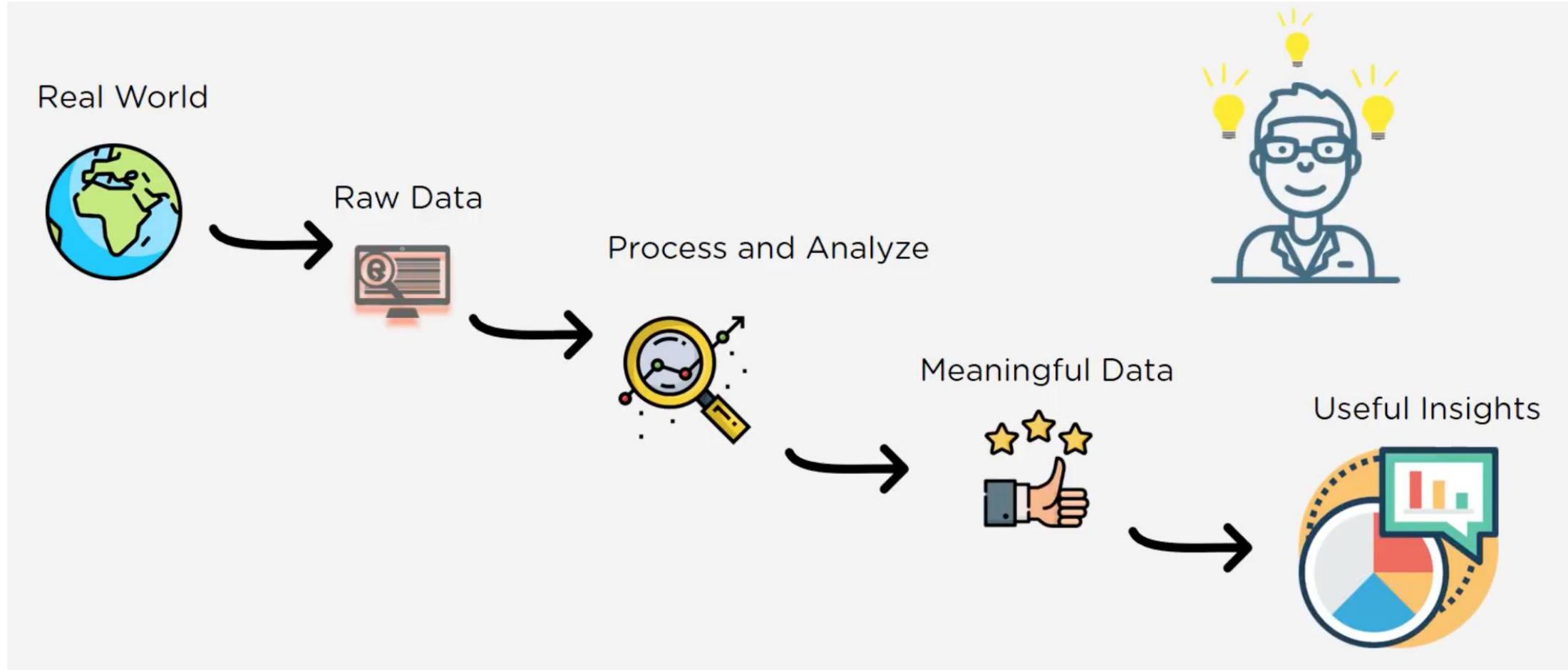


Dữ liệu lớn

Dữ liệu lớn có nhiều lợi ích khác nhau như: Truy cập vào dữ liệu mạng xã hội có thể cho phép điều chỉnh chiến lược kinh doanh, cải thiện trải nghiệm của khách hàng...



Công việc của một nhà Khoa học dữ liệu



Thảo luận



- Khoa học dữ liệu trong hoạt động kinh doanh sách trên Amazon.com
- Khoa học dữ liệu trong hoạt động kinh doanh của GRAB

Hãy nêu một vài bài toán xung quanh bạn có thể áp dụng khoa học dữ liệu để giải quyết nó?

Mục tiêu chung của môn học:

- Hiểu được tầm quan trọng của Khoa học dữ liệu
- Vận dụng được các bước trong quy trình thực hiện một dự án về khoa học dữ liệu.
- Kiến thức, kỹ năng cần thiết để trở thành một nhà khoa học dữ liệu
- Sử dụng được ngôn ngữ lập trình Python.
- Áp dụng được Python và các thư viện phổ biến trong giải quyết một số bài toán cơ bản của Khoa học dữ liệu.

