Bài 1.Giới thiệu về Khoa học dữ liệu

Tìm hiểu xong bài này bạn sẽ:

- Nêu được khái niệm Khoa học dữ liệu
- Trình bày được sơ lược về lịch sử phát triển của Khoa học dữ liệu
- Kể tên được một số ứng dụng khoa học dữ liệu
- Liệt kê được các công việc trong khoa học dữ liệu cũng như các kỹ năng cần có của nhà khoa học dữ liệu
- Trình bày và giải thích được các bước của vòng đời dự án khoa học dữ liệu

Nội dung

1. Khoa học dữ liệu?
2. Ra quyết định dựa trên dữ liệu
3. Sơ lược về lịch sử
4. Nhà Khoa học dữ liệu
5. Vòng đời dự án khoa học dữ liệu

1. Khoa học dữ liệu?

Khoa học dữ liệu?

- Giải quyết vấn đề dựa trên lập luận và tri thức
  - Ngành toán: dựa trên các mệnh đề, công thức, lập luận… để chứng minh bài toán
  - Ngành vật lý: dựa trên các quan sát, thực nghiệm, tính toán,… kiểm chứng các giả thiết
  - Ngành hóa học:…
  - …

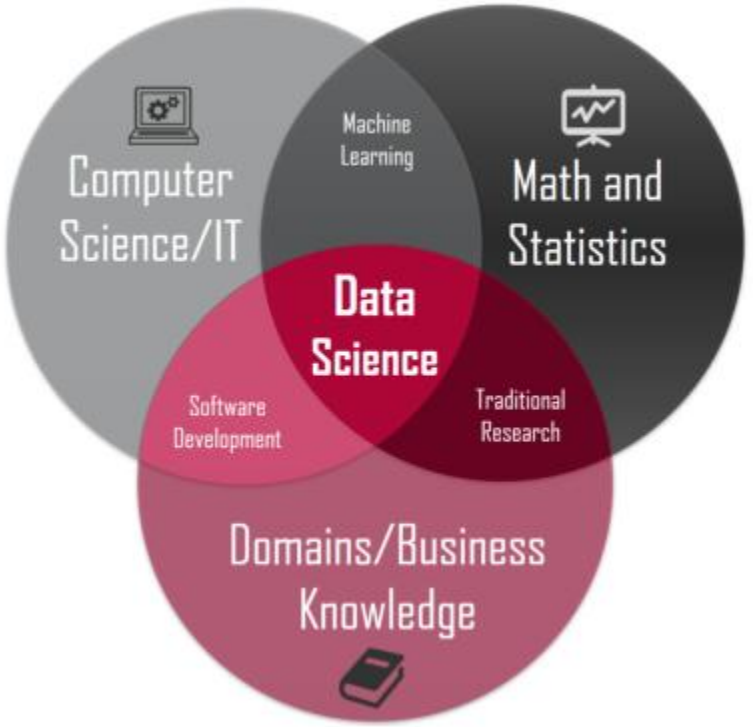
"knowledge-driven" (dẫn dắt bởi tri thức)

Khoa học dữ liệu?

- Khoa học dữ liệu ≠ Khoa học thông thường ở quan điểm: tìm tri thức từ dữ liệu
  - Rút ra tri thức bằng việc tìm tòi từ dữ liệu (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?

- "At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" *Provost and Fawcett, 2013*
- **Khoa học dữ liệu là lĩnh vực liên ngành sử dụng các phương pháp khoa học, quy trình, thuật toán, kết hợp với kiến thức chuyên môn trong lĩnh vực ứng dụng, nhằm rút ra được những hiểu biết sâu sắc từ dữ liệu**



Các nhiệm vụ cụ thể

- **Phân tích và trực quan hoá dữ liệu**: xem xét các mẫu, xu hướng trong tập dữ liệu để hiểu dữ liệu, biểu diễn dữ liệu một cách trực quan và dễ hiểu, giúp người dùng có cái nhìn tổng quan và nhận biết được những yếu tố quan trọng, từ đó phát hiện vấn đề cần giải quyết.
- **Xây dựng mô hình dự đoán, dự báo**: sử dụng dữ liệu để xây dựng mô hình có khả năng dự đoán sự kiện tương lai như sự thay đổi doanh số, xuất hiện rủi ro, biến động về khách hàng,...
- **Tối ưu hoá quyết định**: điều chỉnh quyết định dựa trên dữ liệu, bao gồm việc sử dụng các thuật toán tối ưu hoá để đưa ra quyết định tốt nhất dựa trên các ràng buộc và mục tiêu.
- **Phát hiện tri thức:** tìm ra các mối quan hệ, quy luật ẩn trong dữ liệu, xác định rõ nguyên nhân và kết quả, phát triển kiến thức mới.

Dữ liệu đang trở thành nguyên liệu sản xuất mới

Data are becoming new raw material of business

Craig Mundie, Microsoft

2. Ra quyết định dựa trên dữ liệu

Ra quyết định dựa trên dữ liệu

- Các bài toán dự báo:
  - Dự báo thị trường việc làm: Đến khi tốt nghiệp có xin được việc làm đúng chuyên ngành hay không?
  - Dự báo thời tiết: Tết này đi du xuân chùa Hương có cần mang áo mưa hay không?
  - Dự báo hành vi mua hàng: có thích món hàng này hay không? Mức độ thích như thế nào?
  - ....

Ra quyết định dựa trên dữ liệu

- Các bài toán ra quyết định:
  - Điều chỉnh nhiệt độ điều hòa tối ưu cho hoạt động của người trong phòng?
  - Lái xe tự động?
  - ....
- Các hệ thống gợi ý:
  - Netflix: 2/3 số phim được xem dựa trên khuyến nghị
  - Google News: hệ thống khuyến nghị làm tăng 38% nháy chuột
  - Amazon: 35% lượng hàng bán dựa trên khuyến nghị

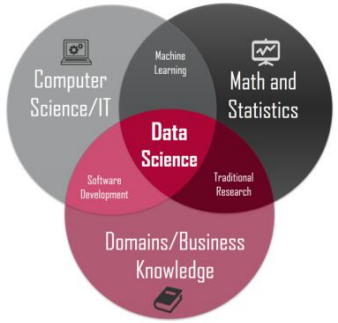
(Xavier, 2014)

Ra quyết định dựa trên dữ liệu

- Các hệ thống phân tích thời gian thực:
  - Cảnh báo cháy thông qua camera
  - Cảnh báo nguy hiểm
  - ...

3.Sơ lược về lịch sử

Vài nét về lịch sử - Statistics



$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Thomas Bayes,
Bayes theorem
(1763)

John W. Tukey
"The Future of data analytics" - "I have come to feel that my central interest is in **data analysis**... **Data analysis**, and the parts of statistics ..." (1962)

Business week "Databased Marketing" (1994)
Companies are **collecting mountains of information about you**, crunching it to **predict how likely you are to buy a product**, and using that knowledge to **craft a marketing message precisely calibrated** to get you to do so...

Prof. Moshe Zviran
" Conventional statistical methods work well with small data sets. Today's databases, however, can **involve millions of rows and scores of columns of data** " (1999)

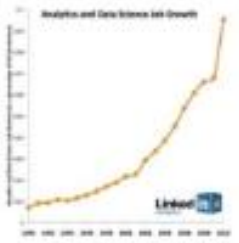
BC

Carl Friedrich Gauss
Linear Regression, Method of least squares
Predict the movement of planet (1809)

Peter Naur, Danish Computer Scientist
"Data Science - The science of dealing with data..." (1976)

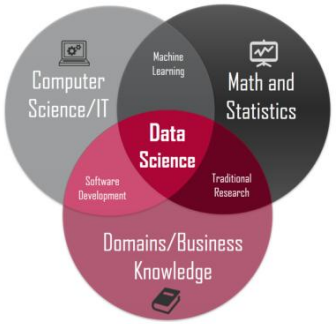
iasc

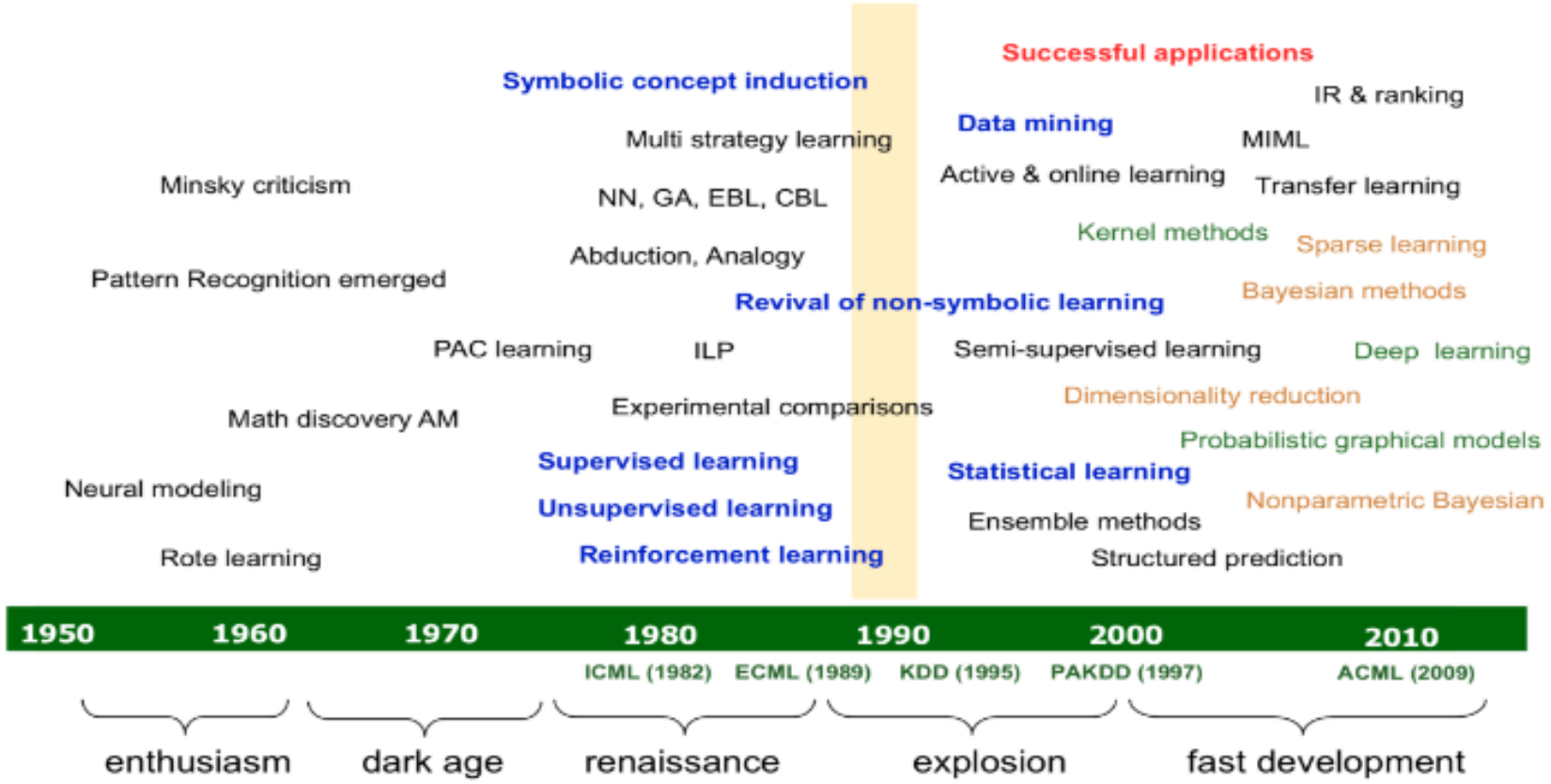
Professor C. F. Jeff Wu - University of Michigan
"Calls for **statistics** to be renamed **data science** and **statisticians** to be renamed **data scientists**." (1997)



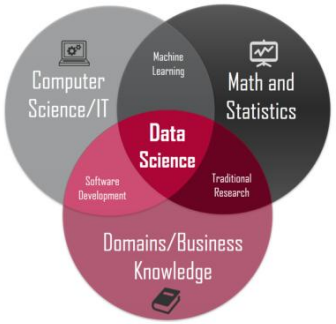
Linked (2010)

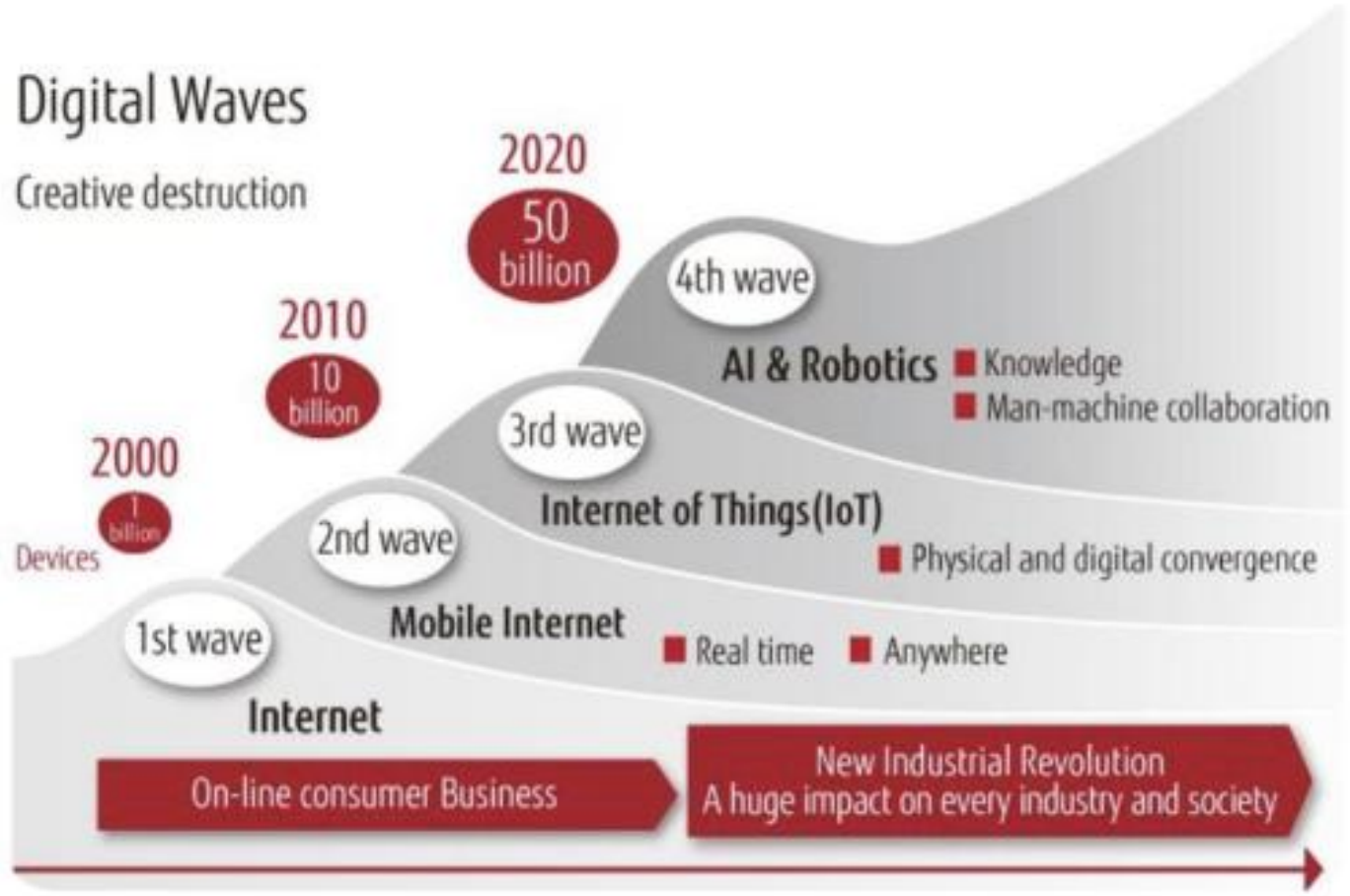
Vài nét về lịch sử - Data Mining -Algorithm



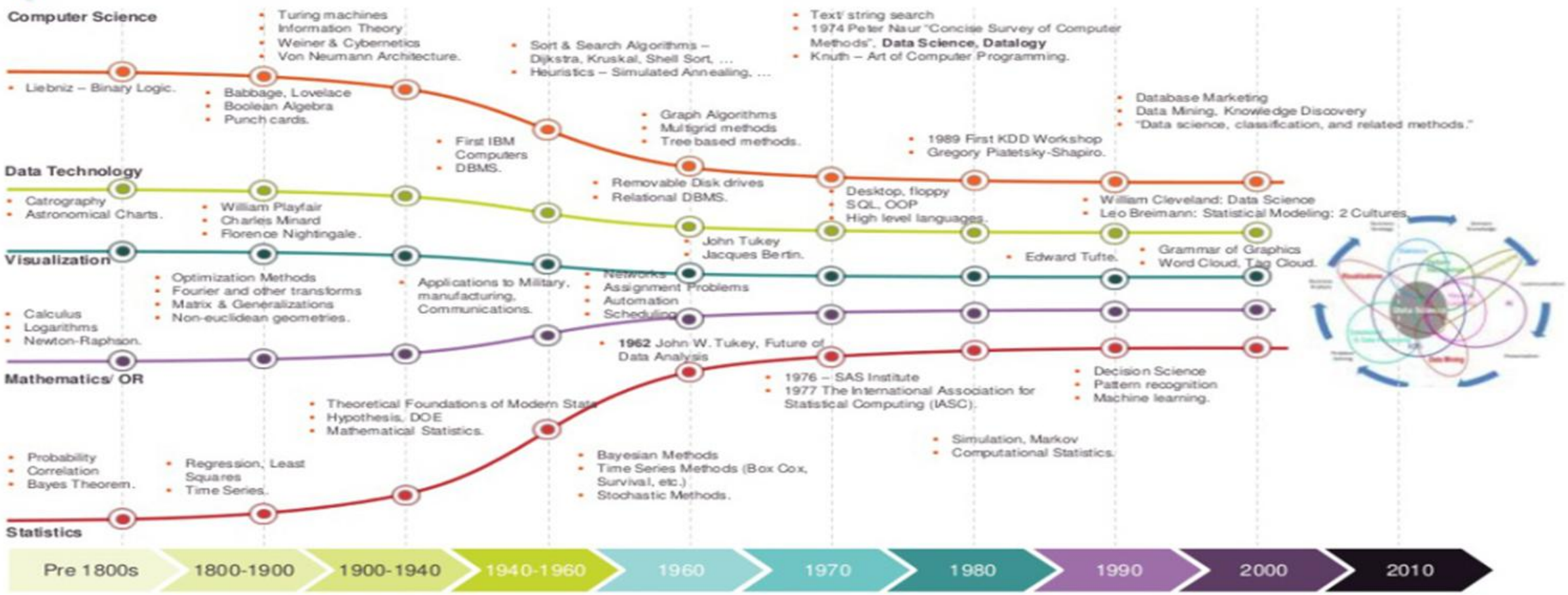


Vài nét về lịch sử - Digital waves

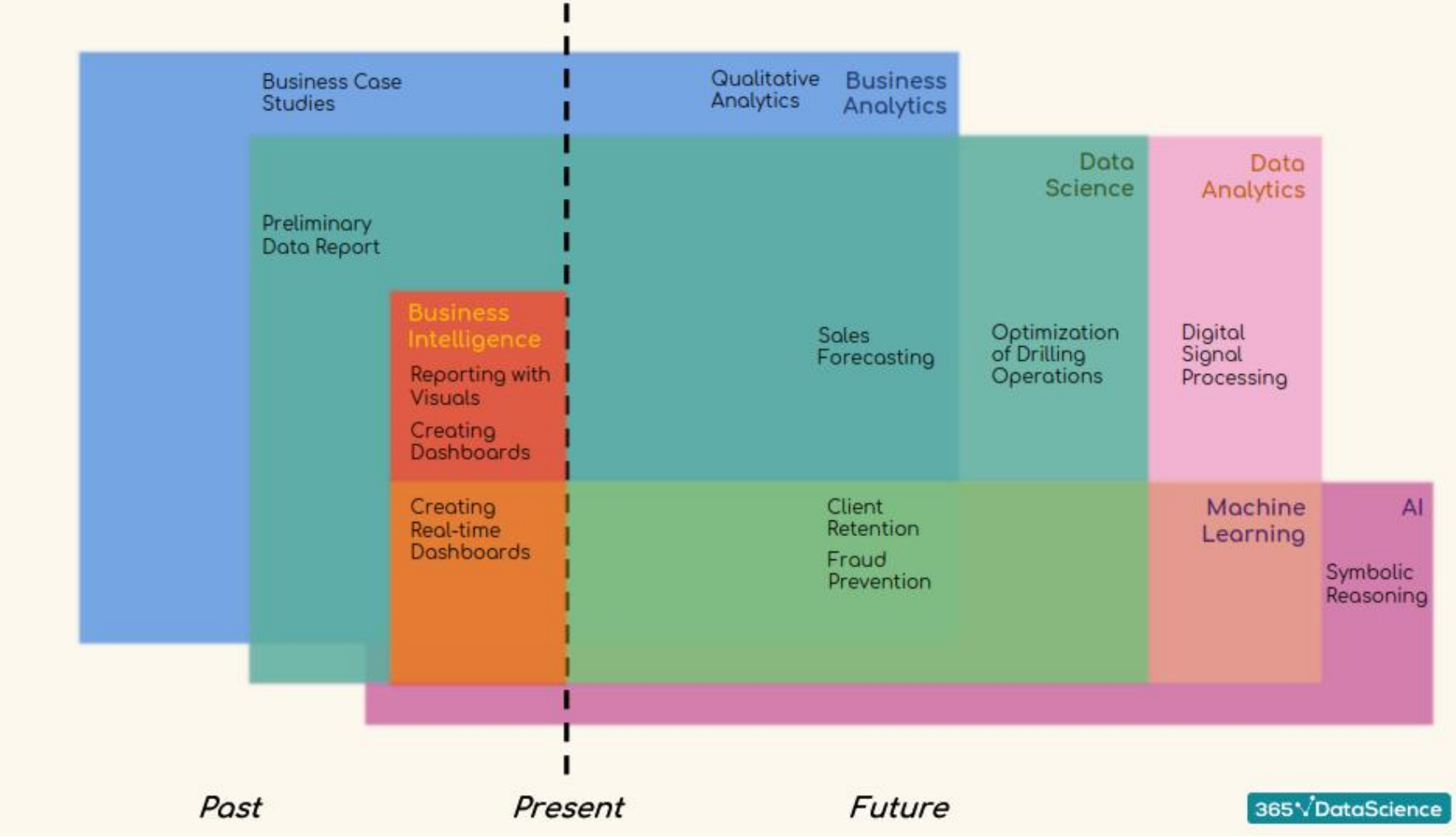




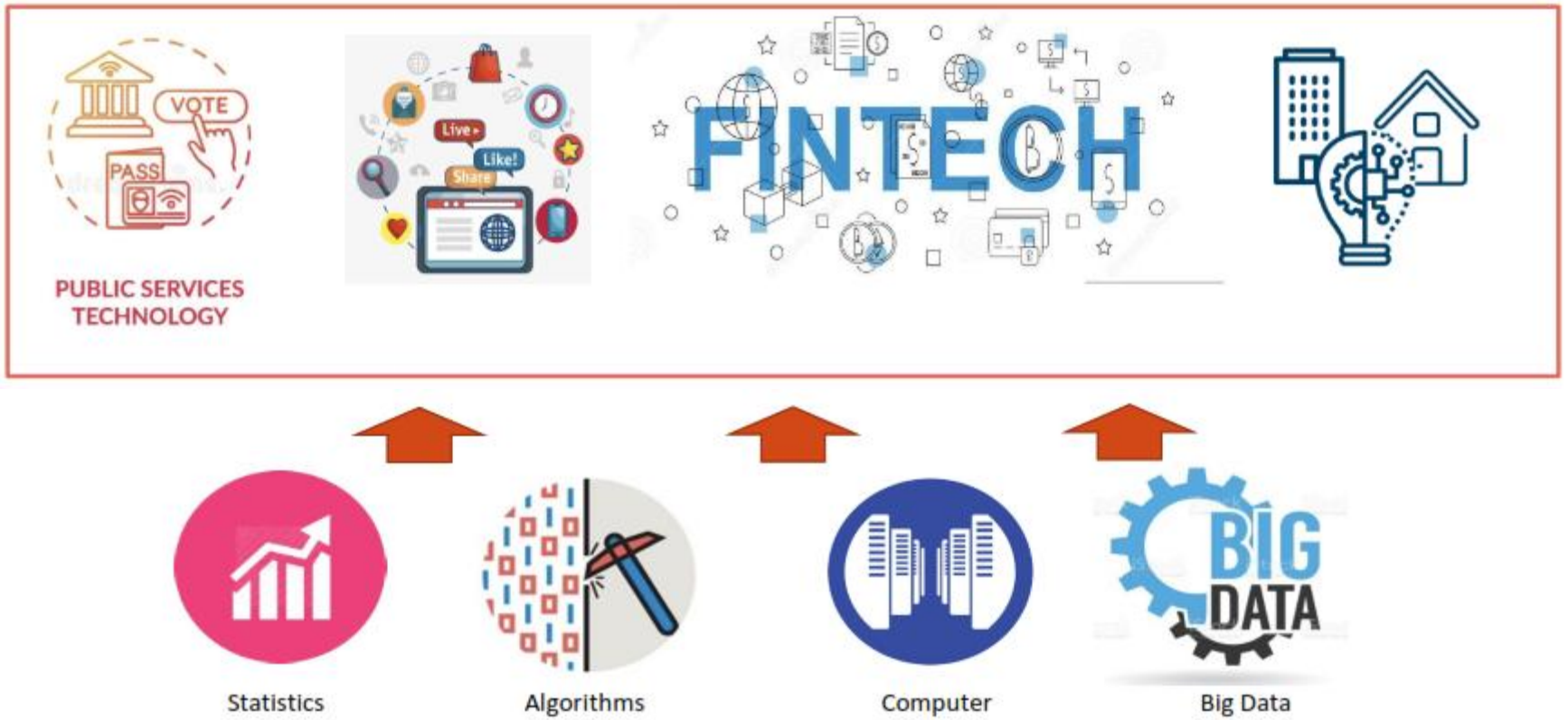
Vài nét về lịch sử - Data Science



Vài nét về lịch sử - Data Science



Vài nét về lịch sử - Data Science



Hoạt động?

- Tìm hiểu một số thành tựu khoa học dữ liệu:
  - Dự án Bộ gen người HGP (Human Genome Project)
  - Hệ thống Giám sát đánh bắt cá toàn cầu (Global Fishing Watch)
  - Hệ thống Giám sát đánh bắt cá toàn cầu (Global Fishing Watch): Chat GPT
  - Mô hình phát hiện gian lận của American Express

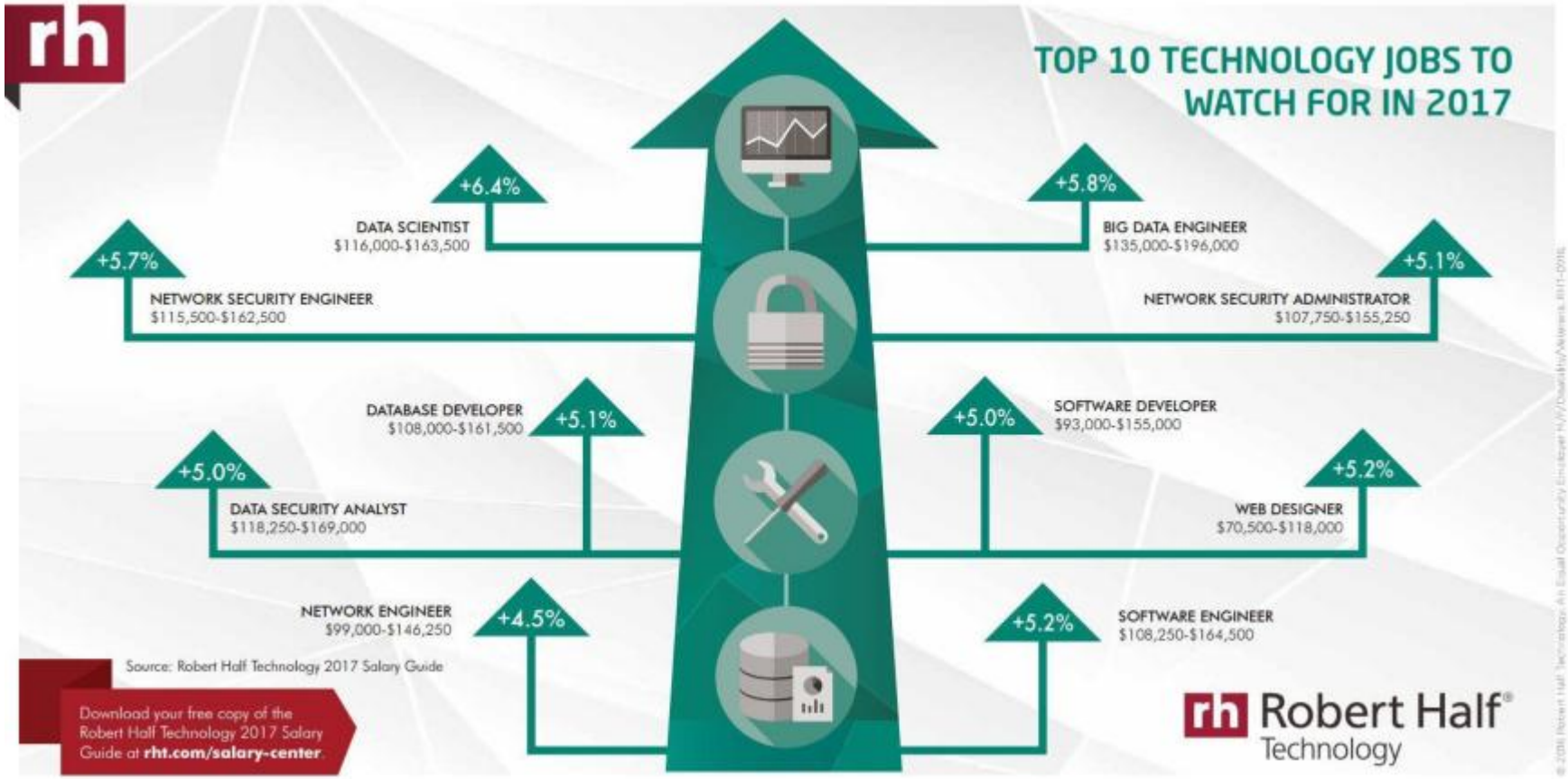
4.Nhà khoa học dữ liệu

Nghề hấp dẫn của thế kỷ 21

"Data scientist is
the sexiest job
of the 21st century."

Harvard Business Review

Nhu cầu tăng cao

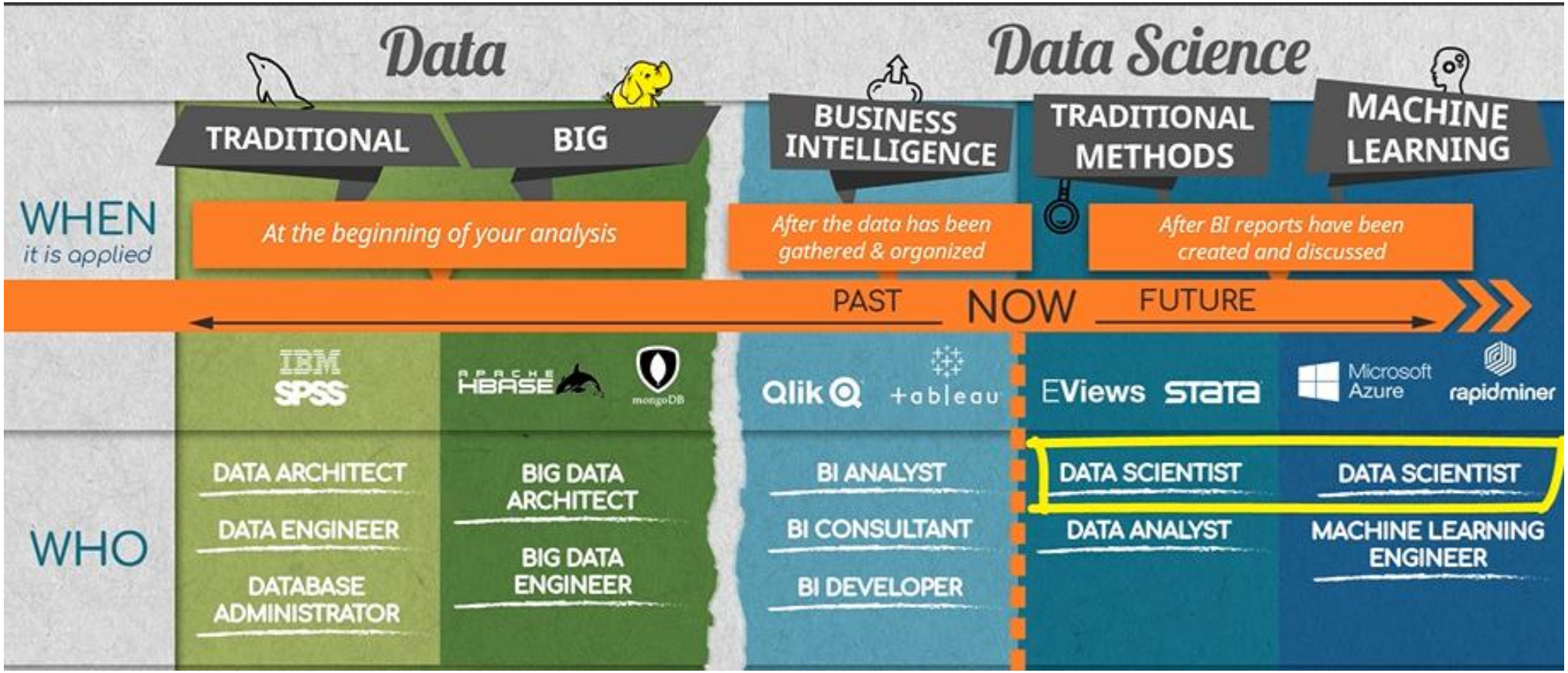


Nhu cầu tăng cao

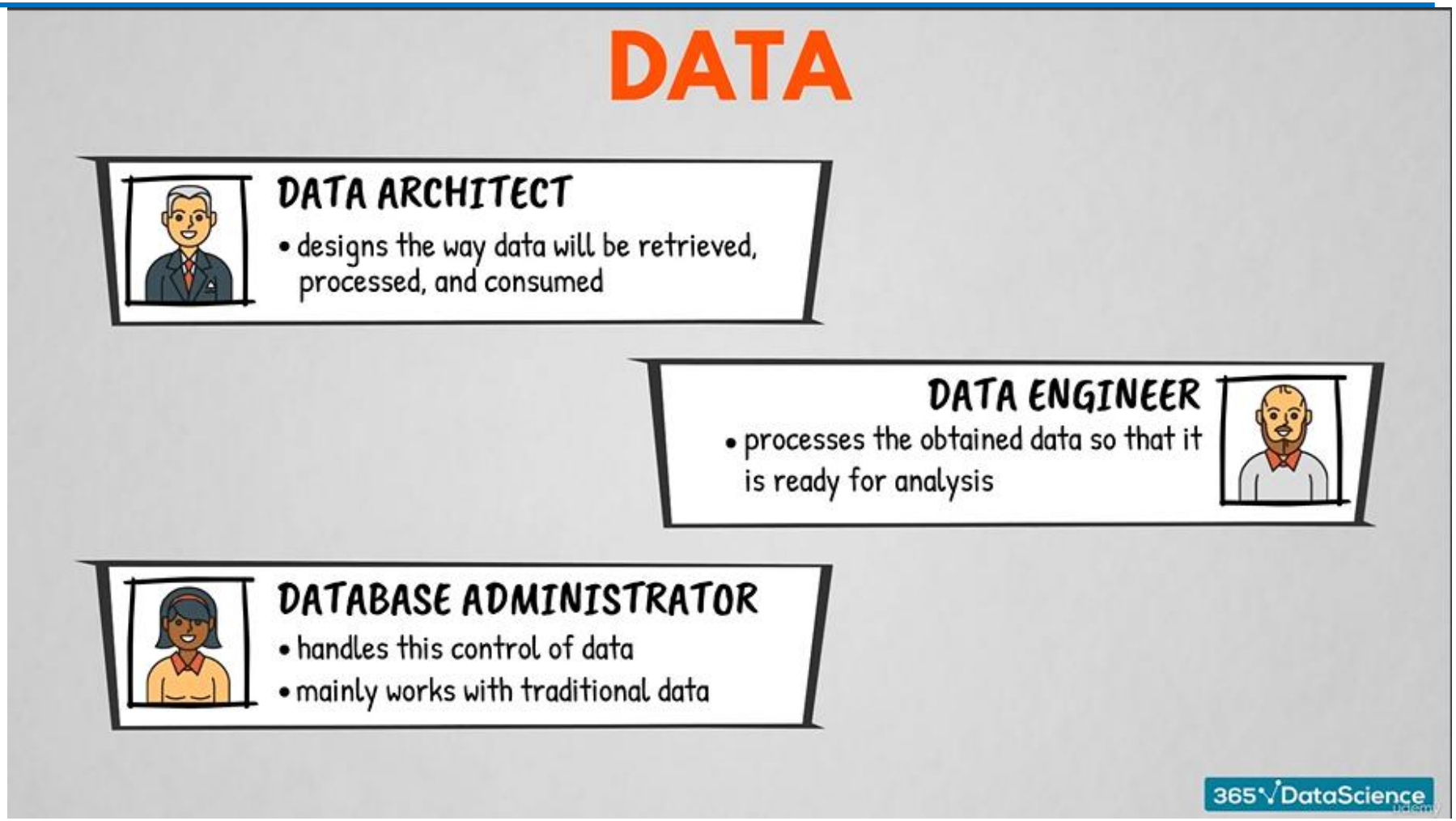
Thu nhập tốt


Big Data, Big Paycheck
Median salary for analytics professionals and those specifically within data science, by level of experience.
Up to 3 years
Analytics professionals
$65,000
Data scientists
$80,000
4 to 8 years
$85,000
$120,000
9+ years
$115,000
$150,000
Note: Data do not include managers
Source: Burtch Works
The Wall Street Journal

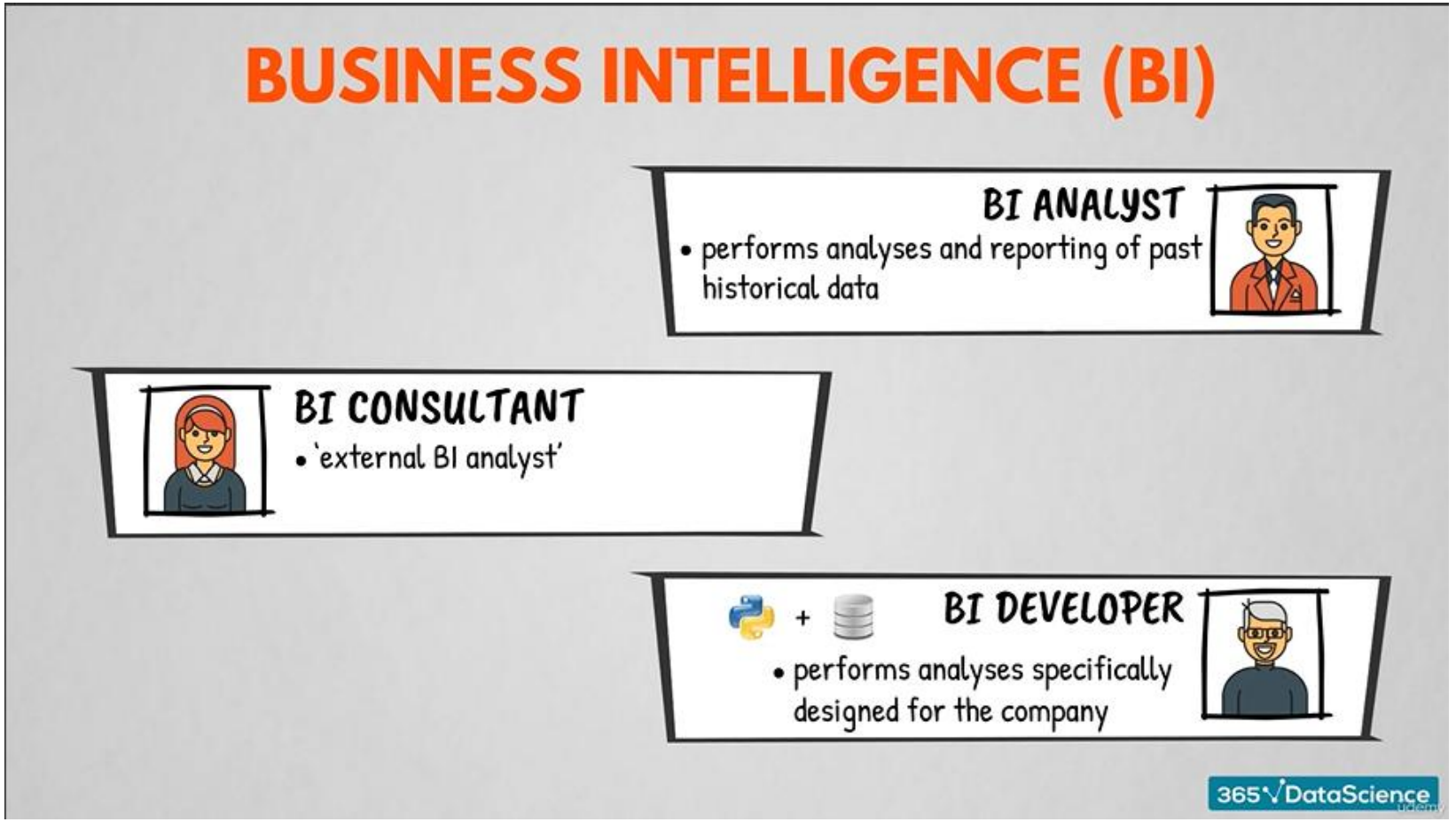
Các vị trí công việc



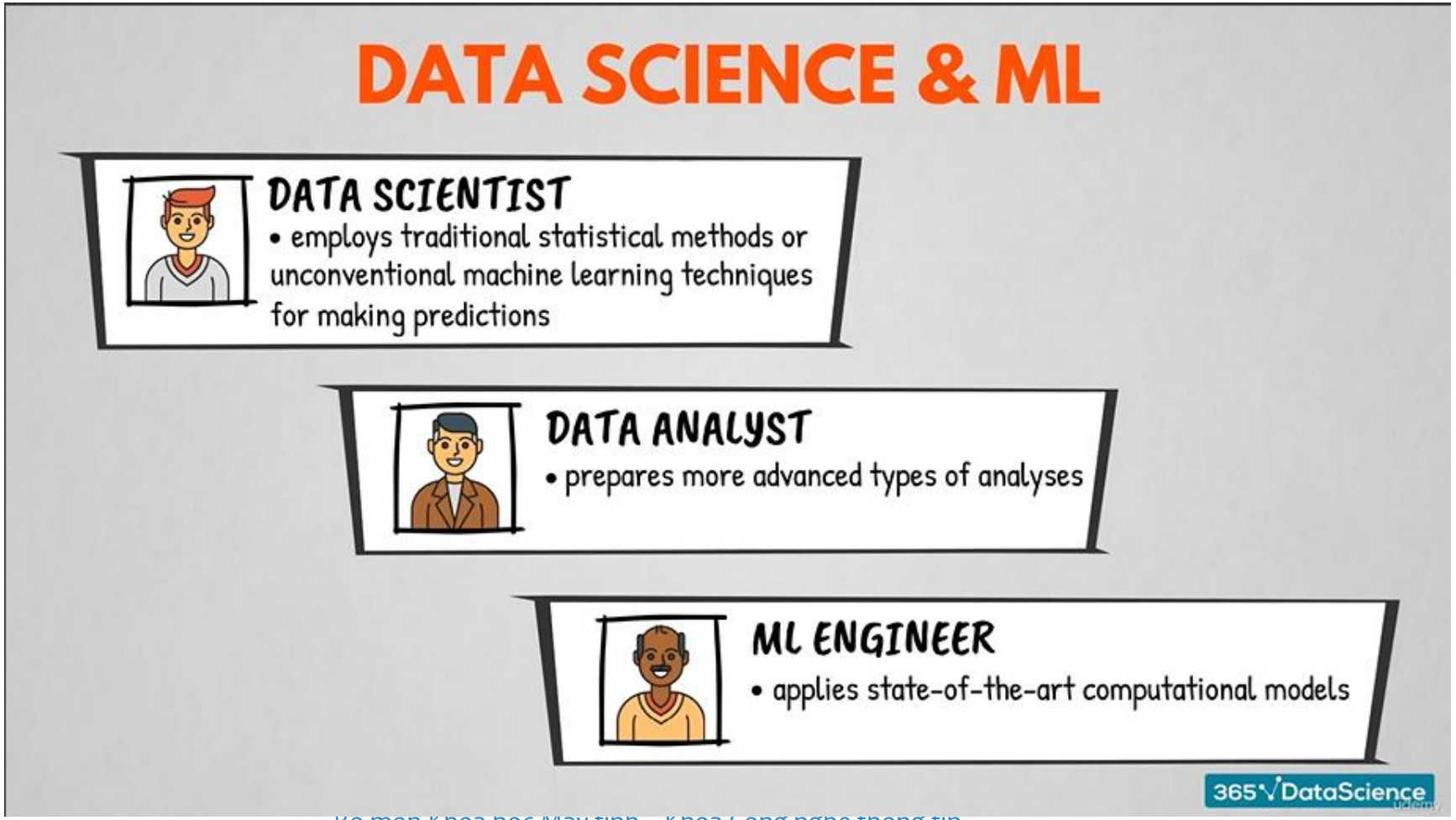
Các vị trí công việc



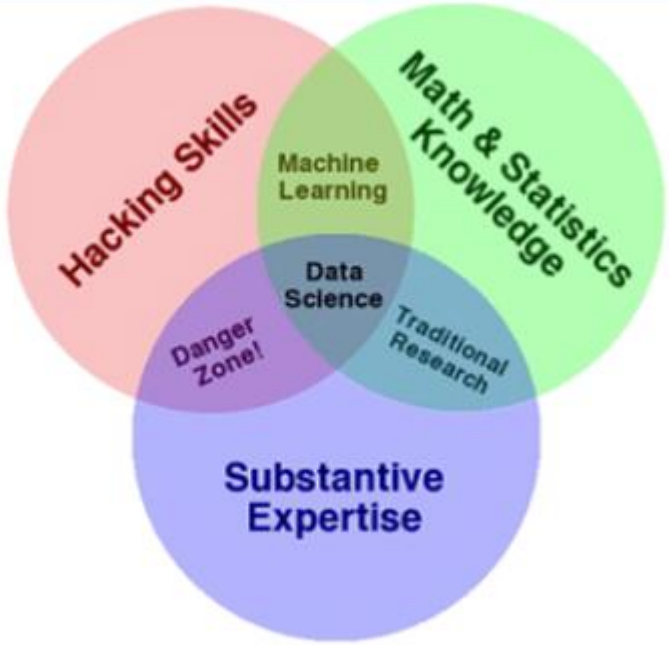
Các vị trí công việc

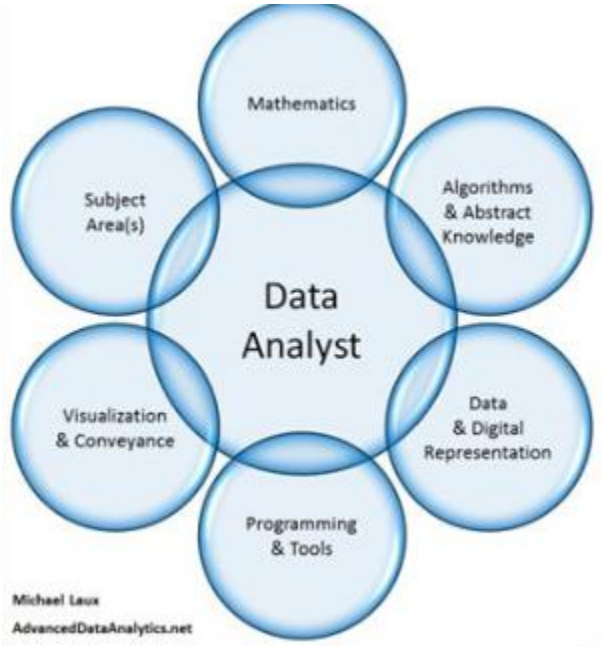


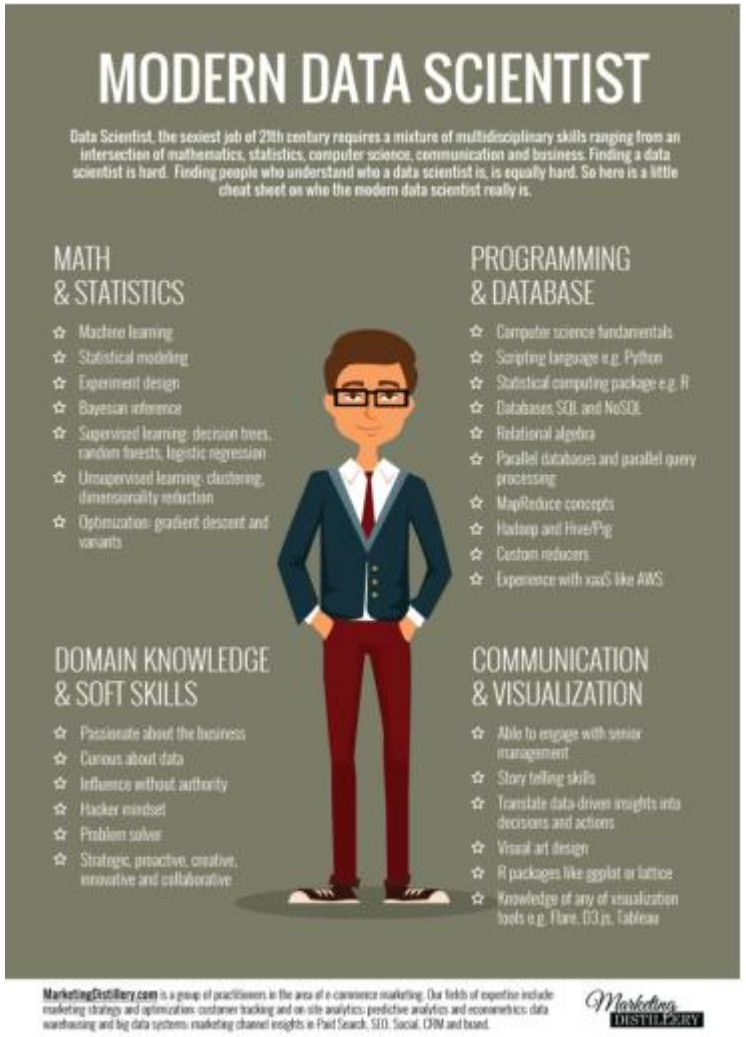
Các vị trí công việc



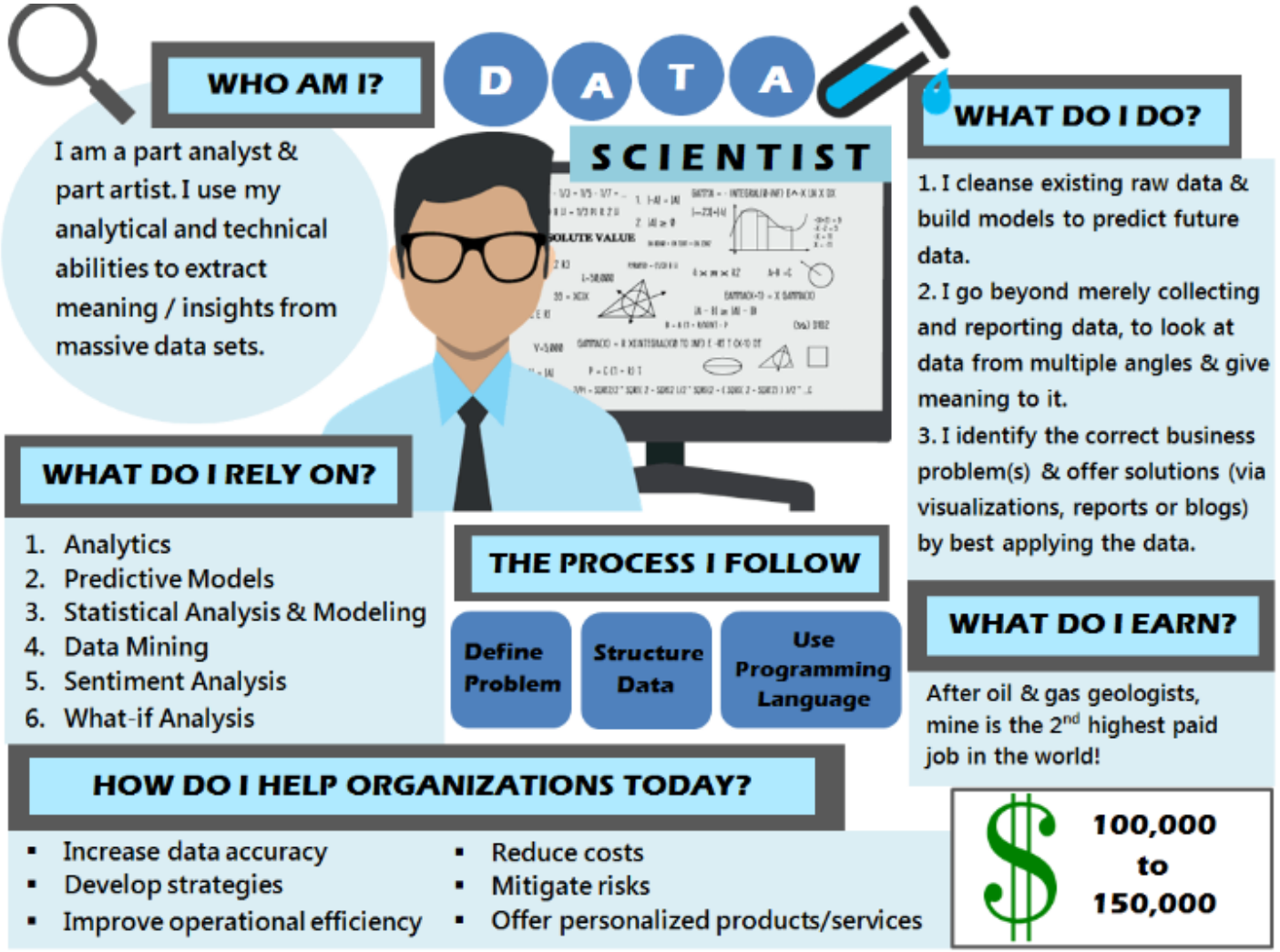
Nhà khoa học dữ liệu cần gì?







Những tố chất cần có?



Những tố chất cần có?

- Tính kiên nhẫn
- Giao tiếp tốt
- Thích tìm hiểu và thử cái mới (tò mò)

Các kỹ năng của nhà khoa học dữ liệu

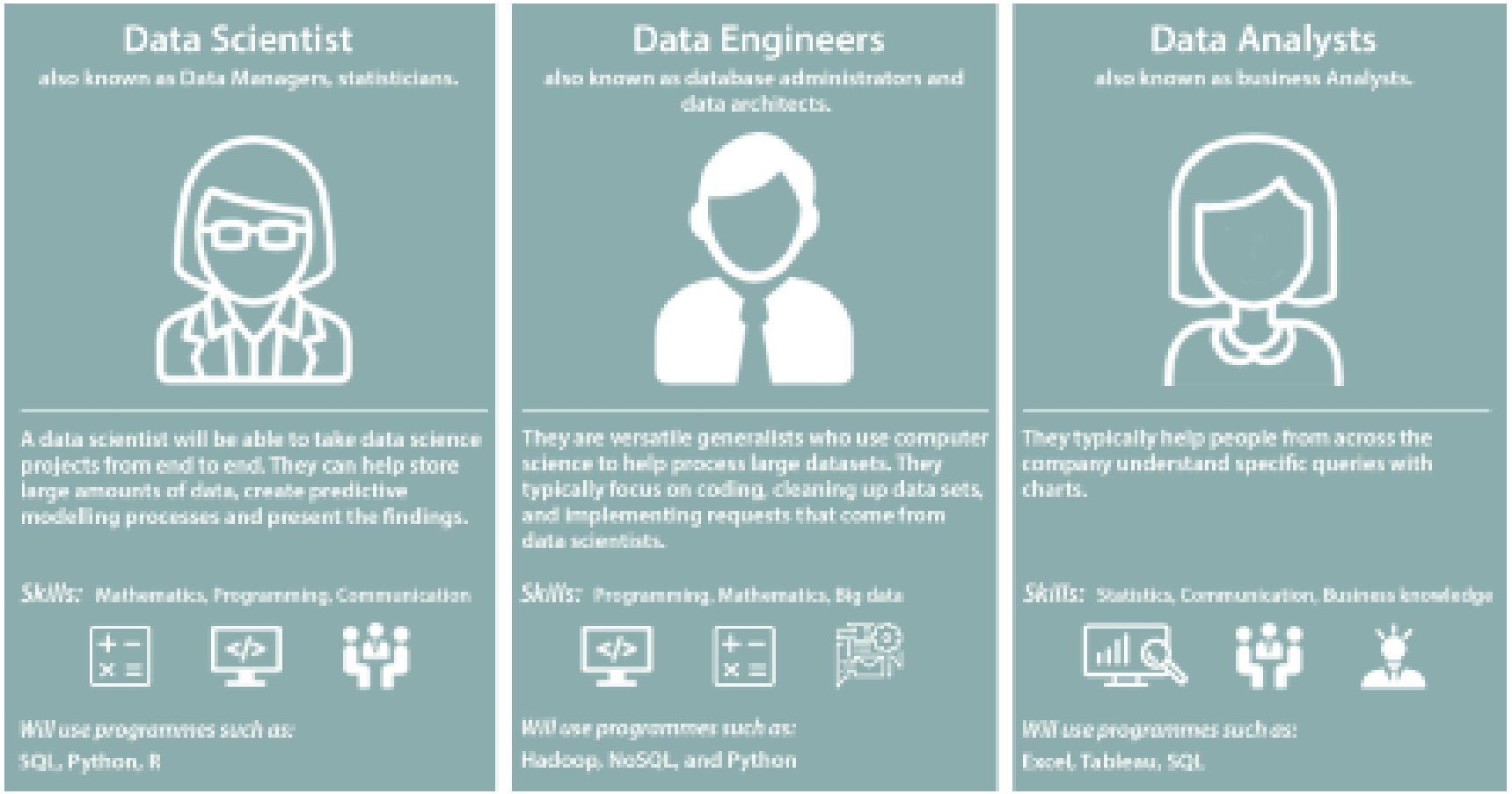
- Hiểu giá trị của dữ liệu
- Hỏi đúng câu hỏi
- Tôn trọng kiến thức ngành
- Hiểu sức mạnh và giới hạn
- Hiểu xác suất thống kê
- Nhạy cảm với các độ đo
- Nhạy cảm với cái quan trọng của dữ liệu
- Chấp nhận thất bại
- Làm việc kiểu AGILE
- Làm việc trong đội có kiến thức nền đa dạng
- Khả năng vừa học vừa làm
- Khả năng kể chuyện
- Khả năng sáng tạo
- Đạo đức và trách nhiệm với dữ liệu

Kỹ năng cần thiết!!!

- Machine Learning
- Database
- Programing Language
- Visualization
- Math
- Acumen (Sự nhạy bén)

Nhà khoa học dữ liệu làm gì?

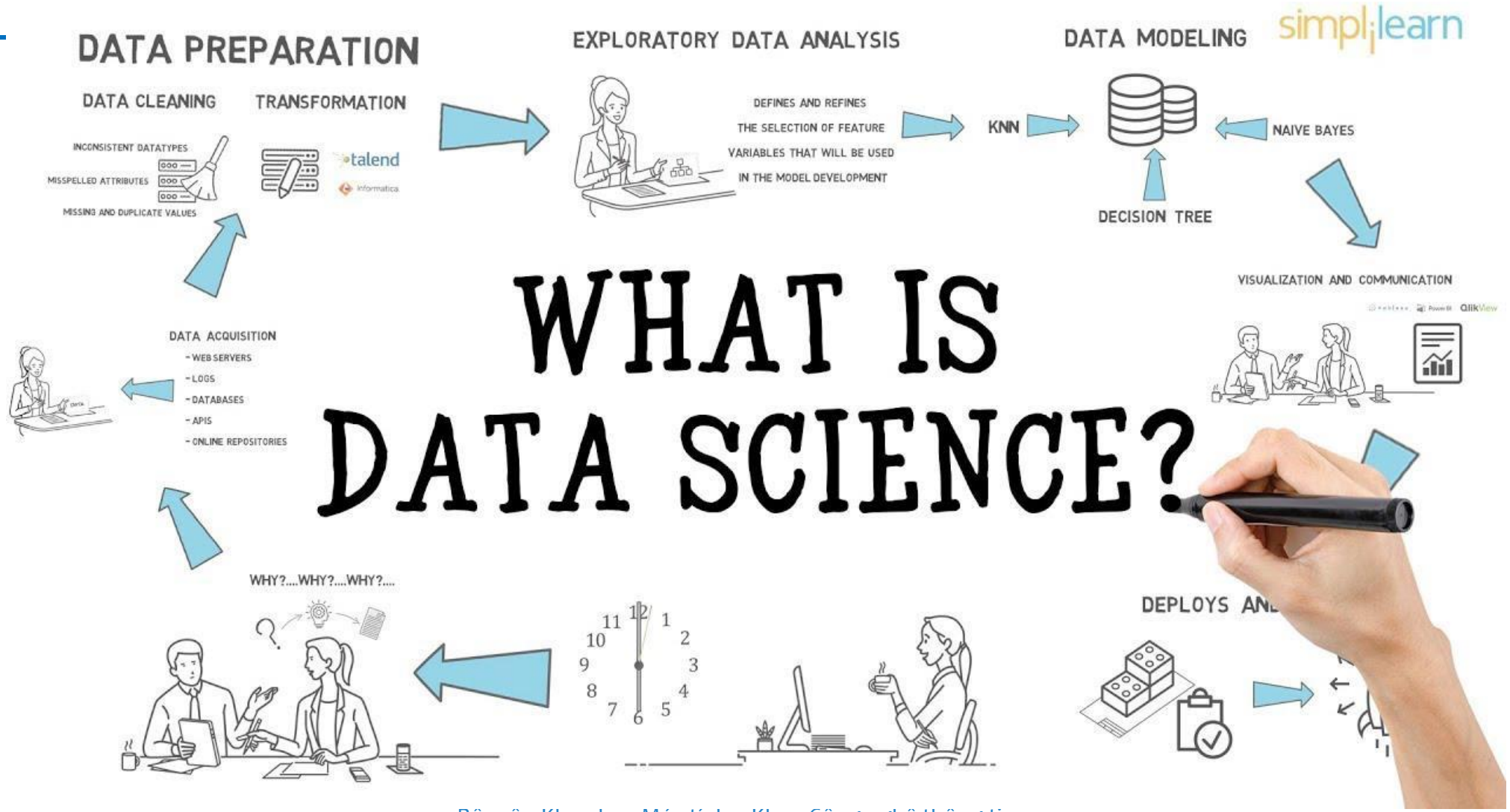
- Thu thập và xử lý dữ liệu phát hiện tri thức (insight) tìm ra các mối quan hệ, quy luật ẩn trong dữ liệu, xác định rõ nguyên nhân và kết quả, phát triển kiến thức mới.
- Giải thích, trình bày những tri thức đó cho các bên liên quan để chuyển hóa insight thành hành động


Data Scientist
also known as Data Managers, statisticians.
A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.
Skills: Mathematics, Programming, Communication
Will use programmes such as:
SQL, Python, R
Data Engineers
also known as database administrators and data architects.
They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.
Skills: Programming, Mathematics, Big data
Will use programmes such as:
Hadoop, NoSQL, and Python
Data Analysts
also known as business Analysts.
They typically help people from across the company understand specific queries with charts.
Skills: Statistics, Communication, Business knowledge
Will use programmes such as:
Excel, Tableau, SQL

Sản phẩm dữ liệu là gì?

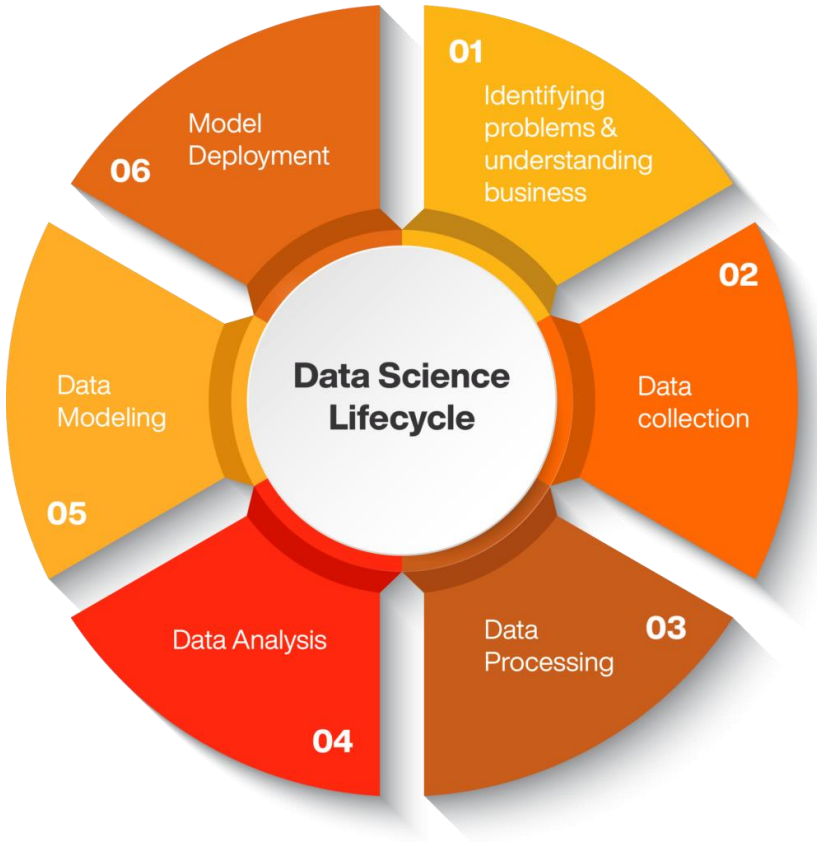
- Sản phẩm được xây dựng dựa trên dựa liệu
- Sản phẩm có thể là một sản phẩm riêng biệt hoặc một phần trong sản phẩm lớn
- Sản phẩm có thể gồm nhiều thành phần. Mô hình dữ liệu là sản phẩm cốt lõi của nó và thường được phát triển bằng các thuật toán học máy.

5.Vòng đời dự án khoa học dự liệu


DATA PREPARATION
DATA CLEANING
TRANSFORMATION
INCONSISTENT DATATYPES
MISSPELLED ATTRIBUTES
MISSING AND DUPLICATE VALUES
talend
Informatica
EXPLORATORY DATA ANALYSIS
DEFINES AND REFINES
THE SELECTION OF FEATURE
VARIABLES THAT WILL BE USED
IN THE MODEL DEVELOPMENT
DATA MODELING
simplilearn
KNN
NAIVE BAYES
DECISION TREE
VISUALIZATION AND COMMUNICATION
DATA ACQUISITION
- WEB SERVERS
- LOGS
- DATABASES
- APIS
- ONLINE REPOSITORIES
WHAT IS
DATA SCIENCE?
WHY?....WHY?....WHY?....
DEPLOYS AN

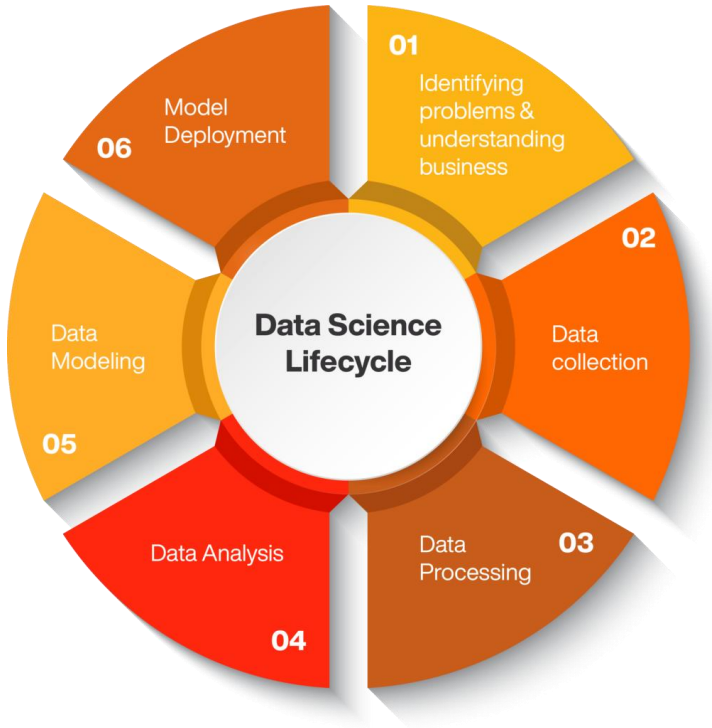
Các bước cơ bản

- **Xác định vấn đề**: Hiểu rõ những vấn đề mà tổ chức, doanh nghiệp cần giải quyết. Từ đó, có thể xác định một số giả thuyết cần kiểm tra, đánh giá và quyết định.
- **Thu thập dữ liệu**: Sau khi hiểu rõ vấn đề, cần thu thập dữ liệu liên quan từ nhiều nguồn.
- **Chuẩn bị dữ liệu**: Lựa chọn dữ liệu; tích hợp dữ liệu từ nhiều nguồn; làm sạch dữ liệu, xử lí các giá trị còn thiếu, không chính xác, loại bỏ ngoại lệ; biểu diễn dữ liệu dưới dạng phù hợp để sử dụng trong các mô hình phân tích.



Các bước cơ bản

- **Phân tích và khai phá dữ liệu**: Áp dụng mô hình cho dữ liệu đã chuẩn bị để chọn lọc một số yếu tố quan trọng nhằm giải quyết vấn đề. Phân tích và khai phá dữ liệu nhằm tìm ra các mối quan hệ, quy luật ẩn trong dữ liệu, xây dựng các mô hình dự báo và phát triển tri thức mới
- **Đánh giá và giải thích**: Sử dụng các tiêu chí cụ thể để đánh giá chất lượng mô hình. Giải thích tác động của mô hình đến hoạt động của tổ chức, doanh nghiệp. Kiểm tra, đánh giá mô hình về mức độ sẵn sàng để triển khai.
- **Ra quyết định và triển khai:** Sau các đánh giá nghiêm ngặt, kết quả phân tích dữ liệu được trình bày cho cấp lãnh đạo quản lí tổ chức, doanh nghiệp để làm cơ sở ra quyết định và triển khai thực tế.



Thảo luận

- Hãy nêu một vài vấn đề liên quan đến địa phương (quê) của bạn, mà bạn cho rằng có thể giải quyết bằng khoa học dữ liệu.
- Theo bạn có những vấn đề nào của trường ta có thể là đối tượng nghiên cứu của khoa học dữ liệu?

Tổng kết

- **Data Science** is about data, models, and evaluation
- **Data Science** can solve a wide variety of problems – once we have the right **data** and **model**
- **Nghề hấp dẫn, nhiều cơ hội việc làm và thu nhập cao**
- **Đòi hỏi nhiều tố chất, kỹ năng, kiến thức: Đặc biệt là Toán học**