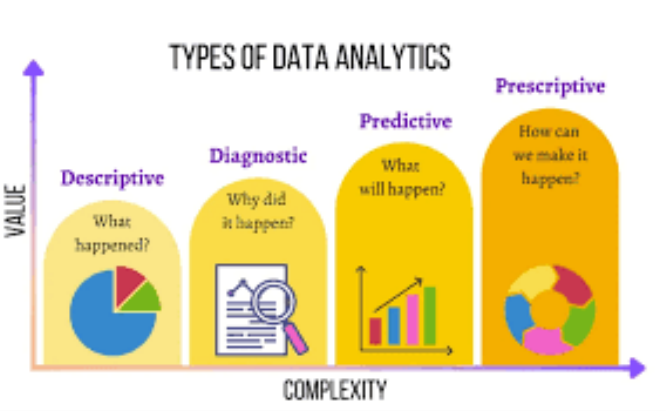


PHÂN TÍCH DỮ LIỆU



(Data Analysis)

THE DATA ANALYSIS PROCESS



Lê Văn Hạnh
levanhanhvn@gmail.com

TÀI LIỆU THAM KHẢO

1. Wes McKinney – Python for Data Analysis - Data Wrangling with Pandas, NumPy and Ipython – O'Reilly - 2nd Edition - 2017
2. Jake VanderPlas – Python Data Science Handbook – Essential tools for working with data – O'Reilly - 2017
3. Nguyễn Đình Thúc, Đặng Hải Vân, Lê Phong - Giáo trình Thống kê máy tính – NXB Khoa Học và Kỹ Thuật - 2010
4. Nguyễn Văn Tuấn – Phân tích dữ liệu với R – NXB Tổng hợp TP.HCM 2022

CÁCH TÍNH ĐIỂM TRONG MÔN HỌC

1. Điểm thường kỳ: **40%**

- Chuyên cần: 10%
- Bài tập trên lớp: 30%

2. Điểm cuối kỳ: **60%**

- Đề tài cá nhân: 60%

Hệ quản trị CSDL – Ngôn ngữ lập trình – Môi trường

1. Hệ quản trị CSDL: SV tùy chọn (SQL Server, mySQL, PostgreSQL, ...)
2. Ngôn ngữ lập trình Python. Khuyến khích sử dụng thêm ngôn ngữ R
3. IDE: Pycharm, Jupyter Notebook, Visual Studio .NET, ...

ĐỀ TÀI CÁ NHÂN TRONG MÔN HỌC

Mỗi SV sẽ thực hiện riêng 1 đề tài, với các yêu cầu :

– *Về dữ liệu sử dụng cho đề tài*

- SV tự thu thập dữ liệu (có thể download từ một số website (Kaggle, WHO, UNDP, UNICEF, ...))
- Dữ liệu có thể thuộc tất cả các lĩnh vực (logistic, kinh doanh, kinh tế, xã hội, y tế, khí hậu, môi trường, ...)
- Tối thiểu 500 records và tối thiểu 12 fields
- Nên gồm đầy đủ các kiểu dữ liệu: số nguyên, số thực, danh nghĩa (category), datetime, boolean

Quy ước: Hai SV bất kỳ không được sử dụng cùng bộ dữ liệu

DÀN Ý ĐỀ TÀI CÁ NHÂN TRONG MÔN HỌC

1. Giới thiệu về bộ dữ liệu cần dùng (liệt kê các thuộc tính cần cho phân tích, số lượng records tối thiểu cần có, ...)
2. Xây dựng phiếu khảo sát để thu thập dữ liệu
3. Quá trình tiền xử lý dữ liệu
 - Làm sạch dữ liệu (*Data Cleaning*)
 - Tích hợp dữ liệu (*Data Integration*)
 - Giảm thiểu dữ liệu (*Data Reduction*)
 - Chuyển đổi dữ liệu và phân tách dữ liệu (*Data Transformation and Data Discretization*)
4. Trừu tượng hóa dữ liệu
 - Nêu các số liệu thống kê cơ bản: mean, median, Q1, Q3, outliers, ...
 - Trình bày tối thiểu 15 đồ thị, trong đó có tối thiểu 10 loại đồ thị khác nhau và có ít nhất 5 đồ thị được vẽ dựa trên dữ liệu được tổng hợp.
 - Khuyến khích sử dụng thêm ngôn ngữ R để trừu tượng hóa dữ liệu.
5. Kết luận và hướng phát triển

NỘI DUNG MÔN HỌC

PHẦN 1 TỔNG QUAN & THU THẬP DỮ LIỆU CHO VIỆC PHÂN TÍCH

1. Khoa học dữ liệu
2. Thu thập dữ liệu
3. Tìm hiểu dữ liệu

PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU (*Data Preprocessing*)

4. Nhiệm vụ chính trong tiền xử lý dữ liệu
5. PANDAS
6. Thao tác với các định dạng khác nhau của tập tin dữ liệu
7. Làm sạch và Chuẩn bị dữ liệu
8. Sắp xếp dữ liệu: nối, kết hợp và định hình lại
9. Tổng hợp dữ liệu và các tác vụ trên nhóm

PHẦN 3 TRỰC QUAN HÓA DỮ LIỆU (*Data Visualization*)

10. Đồ thị và Biểu đồ
11. Vẽ đồ thị và Trực quan hóa



PHẦN 1 TỔNG QUAN

&

THU THẬP DỮ LIỆU CHO VIỆC PHÂN TÍCH

Chương 1

KHOA HỌC DỮ LIỆU (*Data Science*)



Lê Văn Hạnh

levanhanhvn@gmail.com

NỘI DUNG CHƯƠNG 1

1. Giới thiệu
2. Các phương pháp nghiên cứu chính của Khoa học dữ liệu
3. Các kỹ thuật sử dụng trong khoa học dữ liệu
4. Những công nghệ được dùng trong khoa học dữ liệu

1. GIỚI THIỆU

1.1. Khoa học dữ liệu là gì?

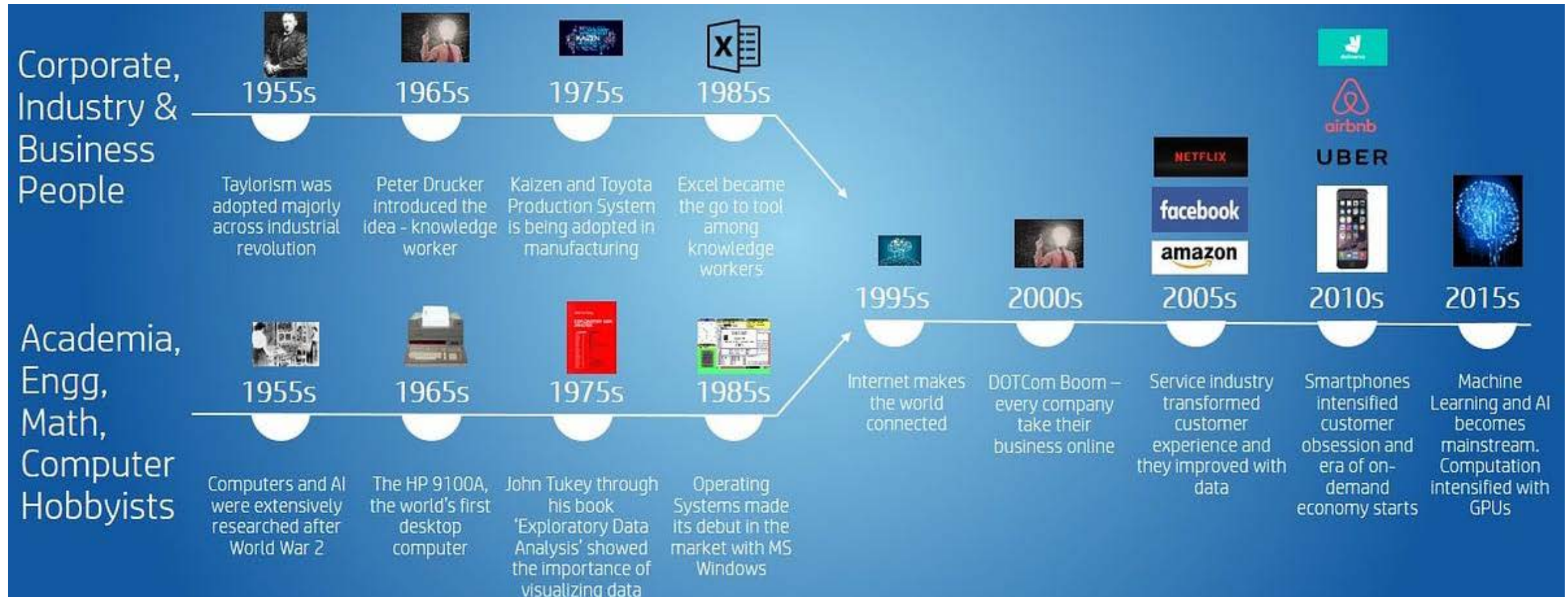
- Khoa học dữ liệu là lĩnh vực nghiên cứu dữ liệu nhằm khai thác những thông tin chuyên sâu có ý nghĩa đối với hoạt động kinh doanh.
- Đây là một phương thức tiếp cận đa ngành, kết hợp những nguyên tắc và phương pháp thực hành của các lĩnh vực toán học, thống kê, trí tuệ nhân tạo, kỹ thuật máy tính để phân tích khối lượng lớn dữ liệu.
- Nội dung phân tích này sẽ giúp các nhà khoa học dữ liệu đặt ra và trả lời những câu hỏi như:
 - Sự kiện gì đã xảy ra?
 - Tại sao nó xảy ra
 - Sự kiện gì sẽ xảy ra?
 - Có thể sử dụng kết quả thu được cho mục đích gì?

1.2. Tại sao khoa học dữ liệu lại quan trọng?

- Các tổ chức hiện nay thuộc hầu hết các ngành nghề (thương mại, điện tử, y tế, tài chính, ...) đều chìm ngập trong dữ liệu và hiện có vô vàn thiết bị có thể tự động thu thập và lưu trữ dữ liệu dưới dạng văn bản, âm thanh, video và hình ảnh.
- Khoa học dữ liệu quan trọng bởi vì lĩnh vực này kết hợp các công cụ, phương pháp và công nghệ để rút ra ý nghĩa từ dữ liệu cho những tổ chức trên.

1.3. Lịch sử lĩnh vực khoa học dữ liệu

- Thuật ngữ khoa học dữ liệu xuất hiện lần đầu vào khoảng thập niên 60, trong vai trò là tên gọi khác của thống kê.
- Đến cuối thập niên 90, các chuyên gia khoa học máy tính đã chính thức hóa thuật ngữ này.



1.4. Những lợi ích chính thu được từ Khoa học dữ liệu

i. Khám phá các mẫu biến đổi tiềm ẩn

Ví dụ: nhờ phân tích dữ liệu giúp Công ty kinh doanh phát hiện

- Có rất nhiều truy vấn của khách hàng được tạo sau giờ làm việc.
- Khách hàng có nhiều khả năng mua hàng hơn nếu họ được phản hồi nhanh chóng thay vì nhận được câu trả lời trong ngày làm việc tiếp theo.

ii. Sáng tạo các sản phẩm và giải pháp mới

Ví dụ: nhờ phân tích dữ liệu giúp Công ty kinh doanh phát hiện Khách hàng quên mật khẩu trong giai đoạn mua sắm cao điểm và không hài lòng với hệ thống khôi phục mật khẩu hiện tại. Công ty có thể sáng tạo ra một giải pháp tốt hơn và nhận thấy mức độ hài lòng của khách hàng tăng lên đáng kể.

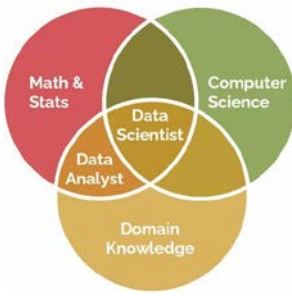
iii. Tối ưu hóa trong thời gian thực

Ví dụ: Công ty vận chuyển bằng xe tải cần giảm thời gian ngừng hoạt động khi xe bị hỏng hóc. Nhờ phân tích dữ liệu giúp xác định được:

- Các mẫu tuyến đường và ca làm việc khiến xe hỏng nhanh hơn
- Loại hỏng hóc thường gặp.

Nhờ vậy họ thực hiện:

- Thay đổi lịch trình vận chuyển.
- Thiết lập một kho phụ tùng thay thế phổ biến cần thay thường xuyên để sửa chữa xe tải nhanh chóng hơn.



1.5. Khoa học dữ liệu với các lĩnh vực dữ liệu khác có liên quan

1.5.1. Khoa học dữ liệu và Phân tích dữ liệu

- Về thuật ngữ:

- Mặc dù hai thuật ngữ này có thể được sử dụng thay thế cho nhau, nhưng phân tích dữ liệu là một nhánh phụ của khoa học dữ liệu.
- Khoa học dữ liệu là một thuật ngữ bao hàm mọi khía cạnh của xử lý dữ liệu—from thu thập dữ liệu đến lập mô hình rồi rút ra thông tin chuyên sâu.
- Phân tích dữ liệu chủ yếu liên quan tới thống kê, toán học và phân tích thống kê. Lĩnh vực này chỉ tập trung vào phân tích dữ liệu.

- Về con người:

- **Nhà phân tích dữ liệu** có thể dành nhiều thời gian hơn cho việc phân tích thông thường, cung cấp các báo cáo thường xuyên. Nói một cách đơn giản, nhà phân tích dữ liệu **diễn giải dữ liệu hiện có**
- **Nhà khoa học dữ liệu** có thể thiết kế phương thức lưu trữ, điều chỉnh và phân tích dữ liệu. Nói một cách đơn giản, nhà khoa học dữ liệu **tạo ra các phương pháp và công cụ mới để xử lý dữ liệu cho các nhà phân tích sử dụng**.
- Tại hầu hết môi trường làm việc, các nhà khoa học dữ liệu và nhà phân tích dữ liệu phối hợp cùng nhau để đạt các mục tiêu chung.

1.5. Khoa học dữ liệu với các lĩnh vực dữ liệu khác có liên quan

1.5.2. Khoa học dữ liệu và Phân tích kinh doanh

- Khác biệt:

- Điểm khác biệt chính giữa hai lĩnh vực này là việc sử dụng công nghệ trong từng lĩnh vực.
- ***Nhà khoa học dữ liệu:***
 - ▢ Làm việc sát với công nghệ dữ liệu hơn các nhà phân tích.
 - ▢ Sử dụng công nghệ để làm việc với dữ liệu kinh doanh. Họ có thể viết ra các chương trình, áp dụng những kỹ thuật máy học để tạo ra mô hình và phát triển thuật toán mới.
 - ▢ Không chỉ nắm rõ vấn đề mà còn có thể xây dựng một công cụ cung cấp giải pháp cho vấn đề đó.

1. Giới thiệu

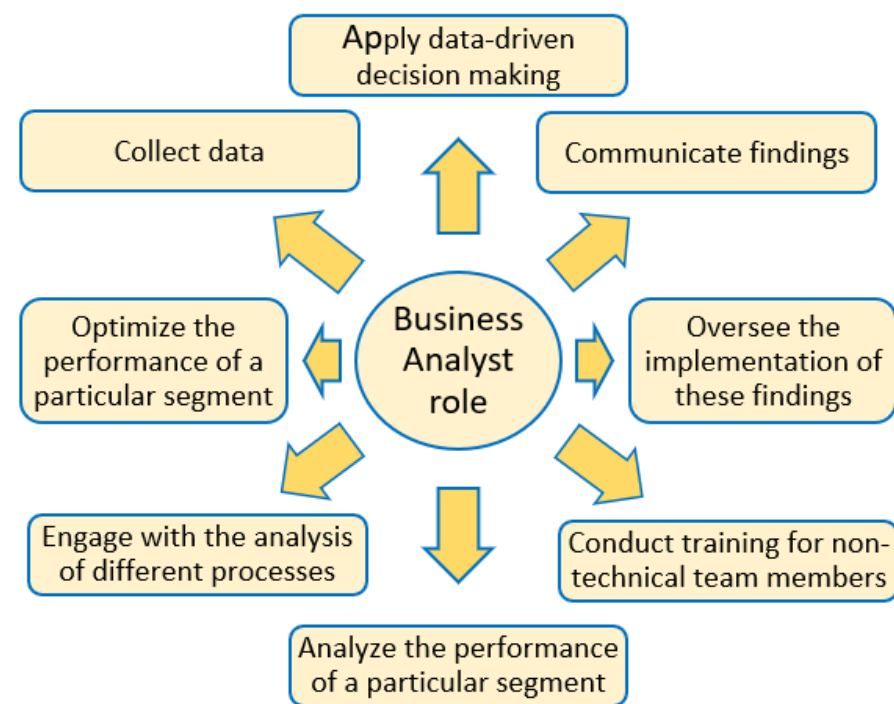
1.5. Khoa học dữ liệu với các lĩnh vực dữ liệu khác có liên quan

1.5.2. Khoa học dữ liệu và Phân tích kinh doanh

- *Khác biệt:*

- **Nhà phân tích kinh doanh:**

- Cố gắng thu hẹp khoảng cách giữa kinh doanh và CNTT.
- Xác định các trường hợp kinh doanh, thu thập thông tin từ những bên liên quan hoặc xác thực các giải pháp.



- *Phối hợp:* Việc các nhà phân tích kinh doanh phối hợp với những nhà khoa học dữ liệu trong cùng nhóm là chuyện không hiếm gặp. Nhà phân tích kinh doanh lấy và sử dụng kết quả từ nhà khoa học dữ liệu để diễn giải theo cách mà toàn thể doanh nghiệp có thể hiểu.



1. Giới thiệu
- 1.5. Khoa học dữ liệu với các lĩnh vực dữ liệu khác có liên quan
- 1.5.3. Khoa học dữ liệu và Kỹ thuật dữ liệu
- **Kỹ thuật dữ liệu**
 - Xây dựng và duy trì các hệ thống cho phép nhà khoa học dữ liệu truy cập và diễn giải dữ liệu.
 - Làm việc chặt chẽ với công nghệ cơ bản hơn là các nhà khoa học dữ liệu. Vai trò này thường liên quan tới việc tạo các mô hình dữ liệu, xây dựng đường ống dữ liệu và giám sát quy trình trích xuất, chuyển đổi, tải (ETL).
 - Tùy thuộc vào quy mô và cơ cấu của tổ chức, kỹ sư dữ liệu cũng có thể quản lý cơ sở hạ tầng liên quan như nền tảng lưu trữ, truyền phát và xử lý dữ liệu lớn.
 - **Nhà khoa học dữ liệu**
 - Sử dụng dữ liệu mà kỹ sư dữ liệu đã xử lý để xây dựng và đào tạo các mô hình dự đoán. Sau đó, các nhà khoa học dữ liệu có thể giao kết quả cho các nhà phân tích để đưa ra quyết định tiếp theo.

1.5. Khoa học dữ liệu với các lĩnh vực dữ liệu khác có liên quan

1.5.4. Khoa học dữ liệu và Máy học



- *Kỹ sư máy học*

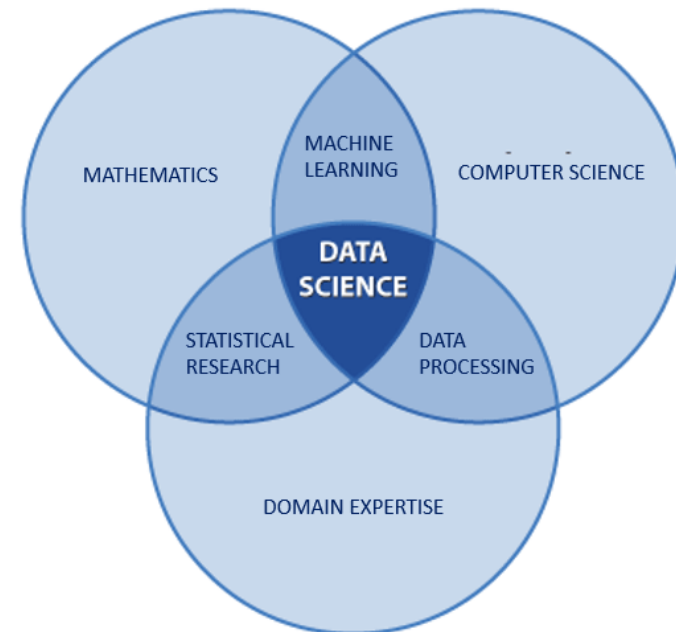
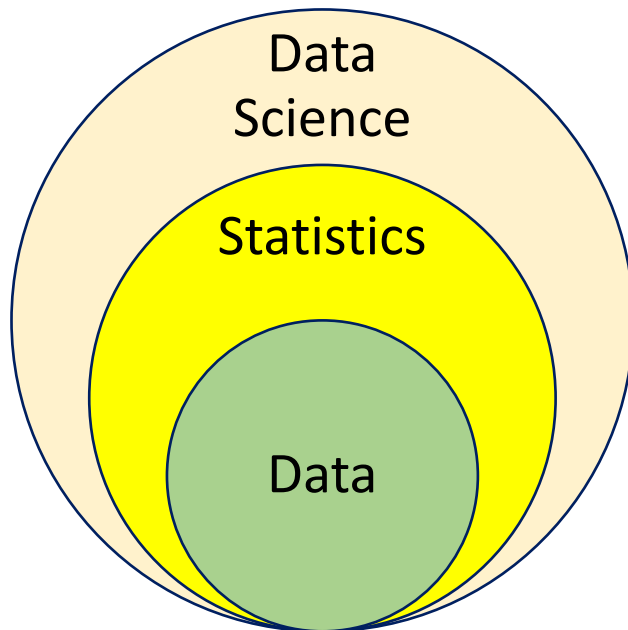
- Máy học là lĩnh vực khoa học về đào tạo máy móc phân tích và học hỏi từ dữ liệu giống như con người. Đây là một trong những phương pháp được sử dụng trong các dự án khoa học dữ liệu nhằm thu thập thông tin chuyên sâu tự động từ dữ liệu.
- Các kỹ sư máy học chuyên về kỹ năng tính toán, thuật toán và viết mã cụ thể cho các phương pháp máy học.

- *Nhà khoa học dữ liệu* có thể sử dụng các phương pháp máy học như một công cụ hoặc hợp tác chặt chẽ với các kỹ sư máy học khác để xử lý dữ liệu.

1.5. Khoa học dữ liệu với các lĩnh vực dữ liệu khác có liên quan




1.5.5. Khoa học dữ liệu và Thống kê

- Thống kê là một lĩnh vực dựa trên toán học nhằm thu thập và diễn giải dữ liệu định lượng.
- Khoa học dữ liệu là một lĩnh vực đa ngành sử dụng các phương pháp, quy trình và hệ thống khoa học để trích xuất tri thức từ dữ liệu dưới nhiều hình thức khác nhau.
- Các nhà khoa học dữ liệu sử dụng các phương pháp từ nhiều lĩnh vực, bao gồm cả thống kê. Tuy nhiên, các lĩnh vực này khác nhau về quy trình và những vấn đề mà chúng nghiên cứu.



1.6. Data Analytics và Data Analysis

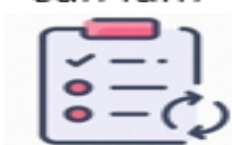

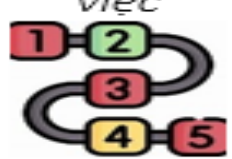

- Các khác biệt chính

	Data Analytics	Data Analysis
<div>Dạng thức</div> <div></div>	Là khoa học phân tích dữ liệu thô để trích xuất kiến thức hữu ích (patterns: các hành vi lặp đi lặp lại) từ chúng	Là một quá trình kiểm tra, làm sạch, chuyển đổi, và mô hình hóa dữ liệu với mục đích tìm ra các thông tin hữu ích, đưa ra kết luận và hỗ trợ ra quyết định của doanh nghiệp.
<div>Mục tiêu</div> <div></div>	Tập trung vào việc sử dụng các kỹ thuật và công cụ để phân tích dữ liệu và tạo ra thông tin hữu ích	Tập trung vào việc khám phá, tìm hiểu và hiểu rõ dữ liệu
<div>Dữ liệu</div> <div></div>	<p>Không có sẵn. Do đó phải tổ chức việc thu thập bao gồm các việc như:</p> <ul style="list-style-type: none">- Hiểu rõ kiến thức về lĩnh vực mà bạn đang làm (domain knowledge)- Xác định được bài toán bạn cần phải giải quyết- Xác định cấu trúc dữ liệu: bạn cần các thông tin xyz của khách hàng, cần thêm lịch sử mua hàng, track thêm hành vi sử dụng ứng dụng của họ, ...- Làm việc với các Team liên quan để triển khai : Data Engineer, Backend, Frontend- Kiểm tra chất lượng dữ liệu	<p>Đã có sẵn, do đó chỉ nghiên cứu để rút ra những thông tin hữu ích</p>

1. Giới thiệu

1.6. Data Analytics và Data Analysis

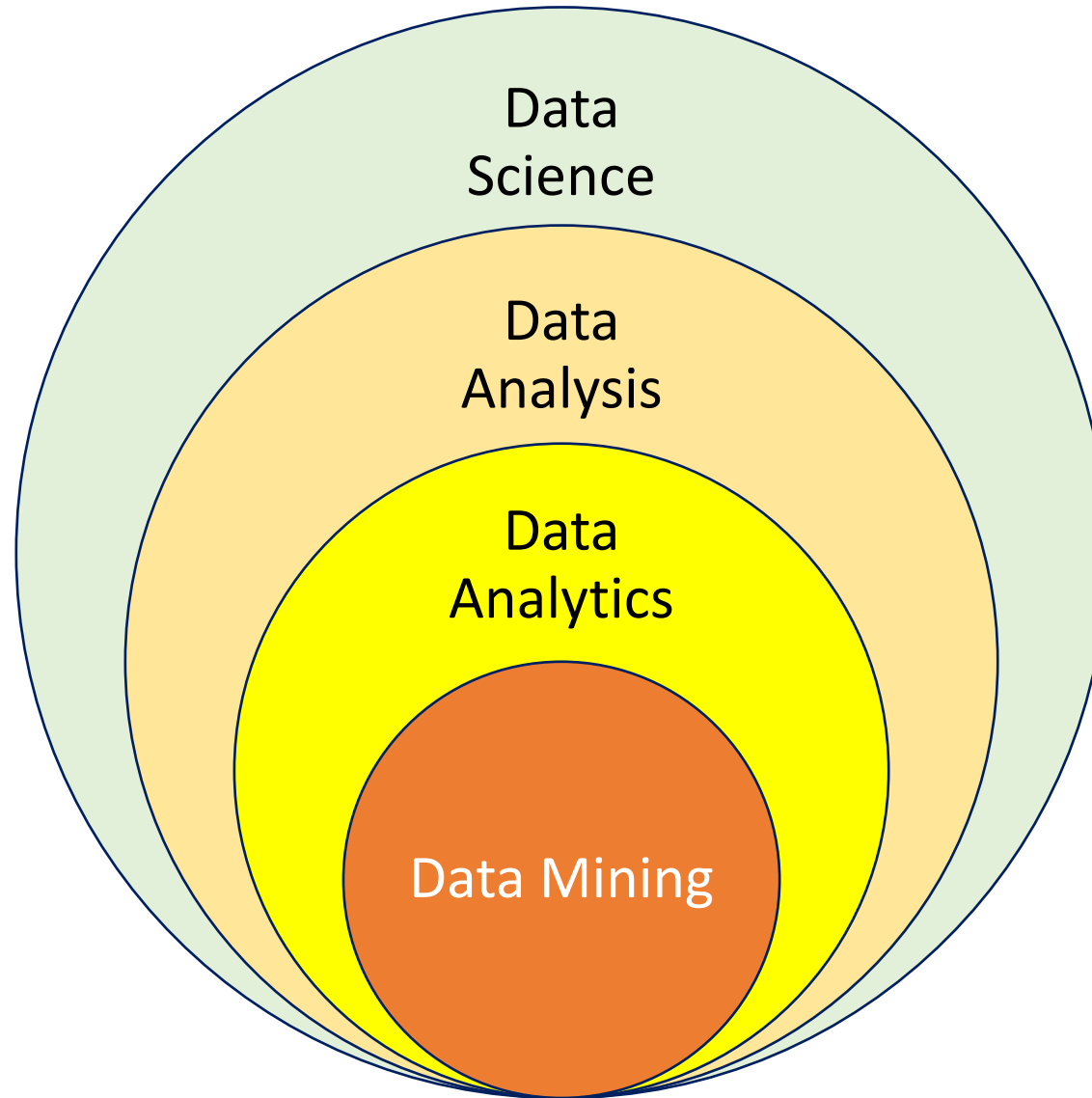
- Các khác biệt chính

	Data Analytics	Data Analysis
<div>Các quy trình cần làm</div> <div></div>	<div>Gồm các quy trình khác nhau như:</div> <div><ul style="list-style-type: none">- Thu thập dữ liệu (data collection)- Lọc dữ liệu (data filtering)- Chuẩn hóa dữ liệu (data standardization)- Diễn giải kết quả (interpreting result)- Thực hiện các thay đổi được đề xuất (making recommended changes)</div>	<div>Gồm ba quy trình chính:</div> <div><ul style="list-style-type: none">- Làm sạch dữ liệu (data cleaning - loại bỏ các giá trị trùng lặp và không đầy đủ)- Trực quan hóa dữ liệu (data visualization – tạo ra các biểu đồ để giúp xem dữ liệu rõ ràng)- Tạo câu chuyện của dữ liệu (data stories) bằng cách sử dụng thông tin chi tiết</div>
<div>Công cụ sử dụng</div> <div></div>	<div><div><ul style="list-style-type: none">- R- Tableau Public- Python- SAS</div><div><ul style="list-style-type: none">- Apache Spark- Excel- Google Sheets- ChartExpo</div></div>	<div><div><ul style="list-style-type: none">- OpenRefine- Tableau Public- Google Fusion Tables</div><div><ul style="list-style-type: none">- KNIME- NodeXL- RapidMiner</div></div>
<div>Trình tự công việc</div> <div></div>	<div>9. Nhận dạng dữ liệu (Data identification)</div> <div>10. Thu thập và lọc dữ liệu (Data acquisition & filtering)</div> <div>11. Trích xuất dữ liệu (Data extraction)</div> <div>12. Tiền xử lý dữ liệu (Data Preprocessing)</div> <div>13. Tổng hợp và biểu diễn dữ liệu (Data aggregation & representation)</div> <div>14. Phân tích dữ liệu (Data analysis)</div> <div>15. Trực quan hóa dữ liệu (Data visualization)</div> <div>16. Tạo câu chuyện dữ liệu (Creating Data stories)</div>	<div>4. Làm sạch dữ liệu (Data cleaning)</div> <div>5. Trực quan hóa dữ liệu (Data visualization)</div> <div>6. Tạo thông tin chi tiết về câu chuyện dữ liệu (Creating Data stories insights)</div>
<div>Sử dụng</div> <div></div>	<div>Để phân tích dự đoán toàn diện</div>	<div>Để phân tích thống kê</div>

1. Giới thiệu

1.6. Data Analytics và Data Analysis

- Mối liên hệ



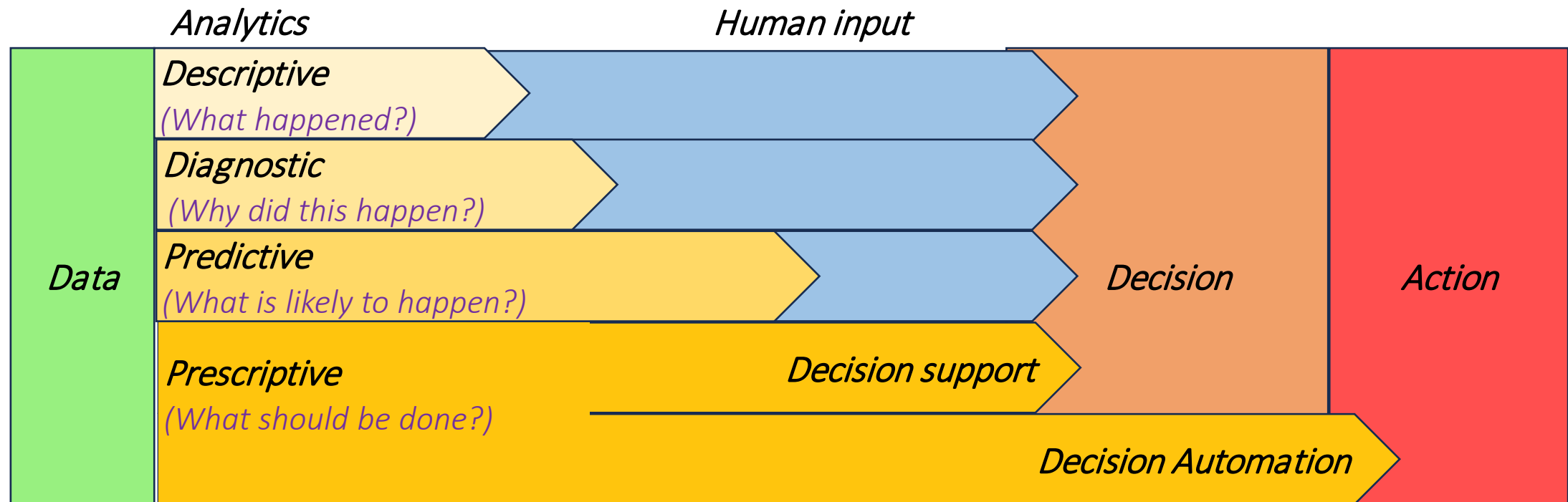
NỘI DUNG CHƯƠNG 1

1. Giới thiệu
2. Các phương pháp nghiên cứu chính của Khoa học dữ liệu
3. Các kỹ thuật sử dụng trong khoa học dữ liệu
4. Những công nghệ được dùng trong khoa học dữ liệu

2. CÁC PHƯƠNG PHÁP NGHIÊN CỨU CHÍNH CỦA KHOA HỌC DỮ LIỆU

Khoa học dữ liệu được sử dụng để nghiên cứu dữ liệu theo 4 phương pháp chính:

- Phân tích mô tả (*Descriptive Analytics*)
- Phân tích chẩn đoán (*Diagnostic Analytics*)
- Phân tích dự đoán (*Predictive Analytics*)
- Phân tích đề xuất (*Prescriptive Analytics*)



2.1. Phân tích mô tả (Descriptive Analytics) - Điều gì đã xảy ra?

- Phân tích mô tả xem xét dữ liệu để thu thập thông tin chuyên sâu về những sự kiện đã hoặc đang xảy ra trong môi trường dữ liệu.

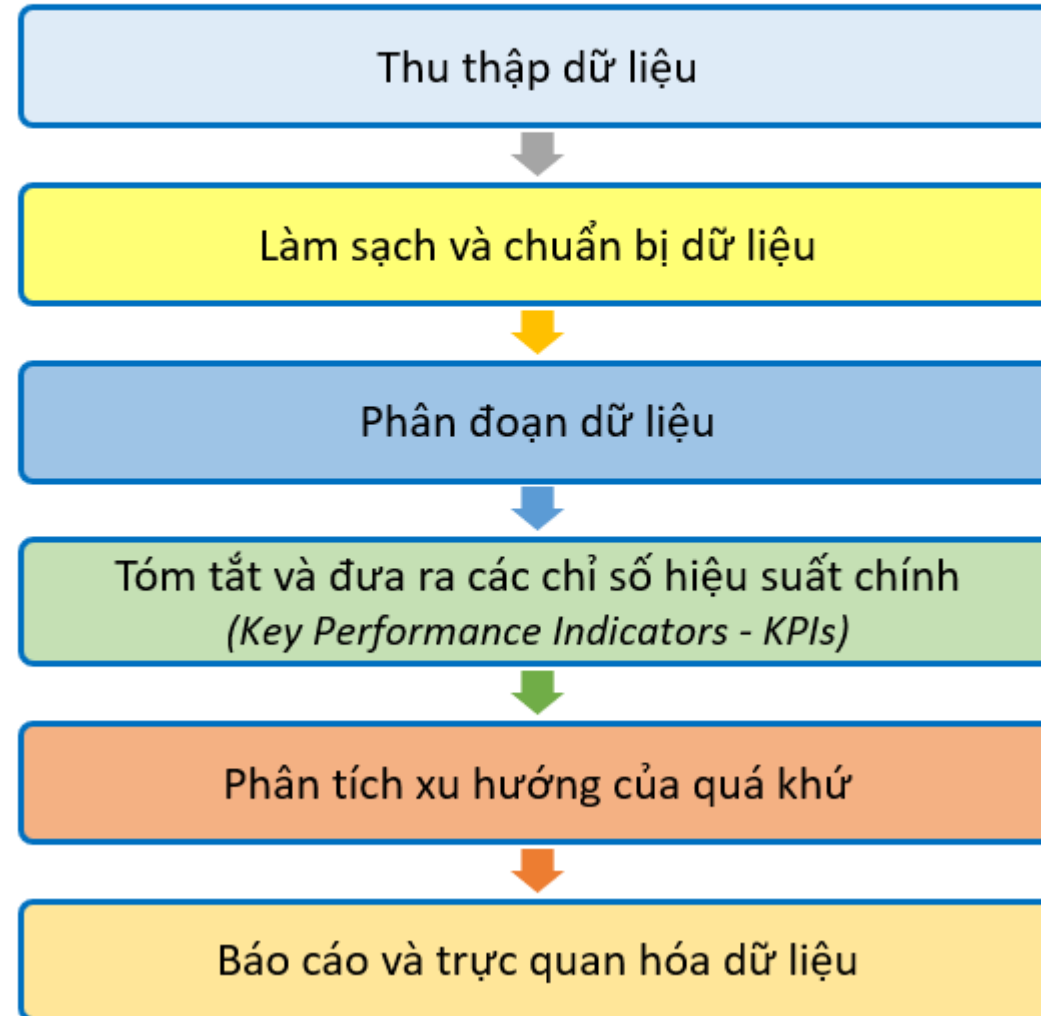
Một số ví dụ:

- Doanh thu trung bình trong tháng 8 là bao nhiêu? Và so sánh với các tháng trước đó.
- Hoặc có bao nhiêu trẻ em từ hai đến mười tuổi được đến trường trên từng quận/huyện của từng tỉnh/thành phố?
- Những triệu chứng đang có của trẻ em trước khi phát hiện bệnh sởi? Những triệu chứng nào là nổi bật?
- Huyện nào có số lượng hộ dân chưa được sử dụng nguồn nước sạch nhiều nhất? Tỷ lệ số hộ dân chưa được sử dụng nguồn nước sạch trên tổng số hộ dân của huyện. Tỷ lệ nam/nữ tham gia khảo sát này là bao nhiêu? Nguồn nước sạch (nếu có) có thường xuyên bị hỏng không? Có phải chủ yếu là phụ nữ đang cho biết rằng khoảng cách mà họ đi lấy nước là quá xa hay không? ...

2. Các phương pháp nghiên cứu chính của khoa học dữ liệu

2.1. Phân tích mô tả (Descriptive Analytics) - Điều gì đã xảy ra?

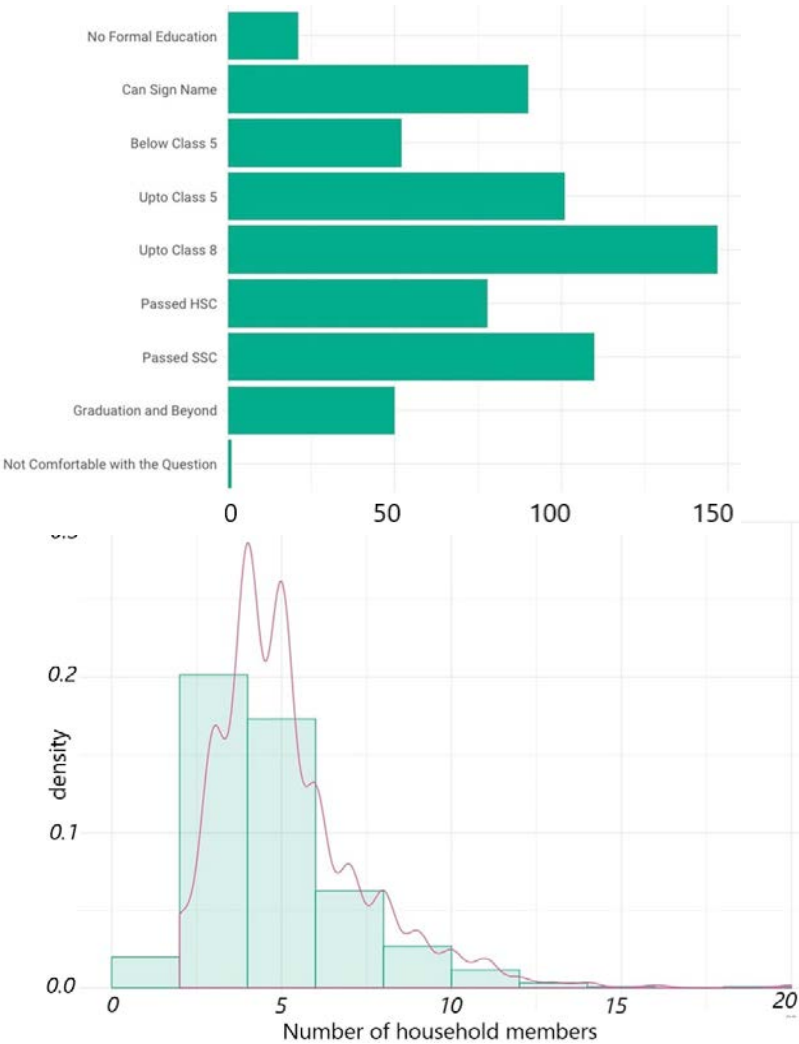
Quy trình Phân tích mô tả (The Descriptive Analytics process)



2. Các phương pháp nghiên cứu chính của khoa học dữ liệu

2.1. Phân tích mô tả (Descriptive Analytics) - Điều gì đã xảy ra?

- Đặc trưng của phương pháp này là sự trực quan hóa dữ liệu thông qua các đồ thị /biểu đồ (Pie, Bar, Line, Histogram, Boxplot, ...) hoặc văn bản thuyết minh. Ví dụ:



Report Objective: Measure the Effectiveness of Overall Marketing Campaigns and Strategies for the Last Quarter	
Outline	
I. Summary	
II. Introduction	
III. Main Point #1 Marketing and Sales Summary	
<input type="checkbox"/> Monthly Highlights	
<input type="checkbox"/> Monthly sales from marketing qualified leads	
<input type="checkbox"/> Monthly revenue generated	
<input type="checkbox"/> Subsection #4	
IV. Main Point #2	
<input type="checkbox"/> Subsection #1	
<input type="checkbox"/> Subsection #2	
<input type="checkbox"/> Subsection #3	
V. Main Point #3	
<input type="checkbox"/> Subsection #1	
<input type="checkbox"/> Subsection #2	
<input type="checkbox"/> Subsection #3	
VI. Conclusion/Recommendations	
VII. Sources	

Sales				
	Sep	Oct	Nov	Total
Apples	250	590		840
John		180		180
Mike		120		120
Pete		290		290
Sally	250			250
Bananas		430	600	1030
John			400	400
Mike			200	200
Pete		180		180
Sally	250			250
Cherries	580	910		1490
John		250		250
Mike	250	330		580
Pete		330		330
Sally	330			330
Oranges		120	720	840
John			120	120
Mike			400	400
Pete		120		120
Sally		200		200
Total	830	2050	1320	4200

Month	(All)				
Sales	Product				
Reseller	Apples	Bananas	Cherries	Oranges	Total
John	\$180	\$400	\$250	\$120	\$950
Mike	\$120	\$200	\$580	\$400	\$1,300
Pete	\$290	\$180	\$330	\$120	\$920
Sally	\$250	\$250	\$330	\$200	\$1,030
Total	\$840	\$1,030	\$1,490	\$840	\$4,200

Product	(All)				
Sales	Month				
Reseller	Sep	Oct	Nov		Total
John			\$430	\$520	\$950
Mike	\$250	\$450	\$600		\$1,300
Pete		\$920			\$920
Sally	\$580	\$250	\$200		\$1,030
Total	\$830	\$2,050	\$1,320		\$4,200

2.1. Phân tích mô tả (Descriptive Analytics) - Điều gì đã xảy ra?

- Về dữ liệu đang có:
 - Các câu hỏi quan trọng cần ghi nhớ khi xem xét số liệu thống kê mô tả là:
 - Dữ liệu đã đầy đủ thông tin (thuộc tính) chưa?
 - Số lượng mẫu thu được đã đủ để phân tích hay chưa?
 - Có bất kỳ lý do nào có thể ảnh hưởng đến dữ liệu đang có không? Ví dụ: điều kiện thời tiết buộc phải sử dụng những người tham gia khác ngoài dự định hoặc do người vợ đi vắng nên những khảo sát về việc nội trợ của phụ nữ lại do người chồng trả lời hộ, ...
 - Nguồn gốc của dữ liệu:
 - *Nếu tập dữ liệu được cung cấp sẵn*: Khi đó nhà phân tích có thể không biết về bối cảnh và chi tiết của các mẫu dữ liệu. Nếu đúng như vậy, điều quan trọng là phải tìm hiểu bối cảnh của mẫu và tổng thể mà mẫu đó trước khi tiến hành các bước phân tích mô tả.
 - *Nếu tập dữ liệu có được do tự thu thập* (hoặc là thành viên trong nhóm thu thập): người phân tích có thể đã biết rõ về thông tin mà dữ liệu và biết điều gì diễn ra tốt đẹp và điều gì không khi thu thập dữ liệu. Nhờ vậy, sẽ có nhiều lợi thế hơn khi phân tích mô tả.

2.2. Phân tích chẩn đoán (Diagnostic Analytics) - Tại sao điều này xảy ra?

- Phân tích chẩn đoán là một phương pháp phân tích chuyên sâu hoặc chi tiết dữ liệu để nắm được nguyên nhân khiến một sự kiện xảy ra.
- Phương pháp chẩn đoán sử dụng các kỹ thuật như
 - Truy vết
 - Khám phá dữ liệu
 - Khai thác dữ liệu
 - Đối sánh.
- Nhiều thao tác vận hành và chuyển đổi dữ liệu có thể được thực hiện trên một tập dữ liệu nhất định để phát hiện ra những mẫu “đặc biệt” trong từng kỹ thuật này.

Phân tích chẩn đoán (Diagnostic Analytics)

Xác định những bất thường mà hiểu biết hiện tại không thể giải thích đầy đủ

Đi sâu vào dữ liệu để tìm kiếm các mẫu chưa từng được biết đến trước đây

Xác định mối quan hệ nhân quả giữa các mẫu dẫn đến sự bất thường

2.2. Phân tích chẩn đoán (Diagnostic Analytics) - Tại sao điều này xảy ra?

- Một số ví dụ:

- Tại sao doanh thu trung bình trong tháng 8 năm nay lại sụt giảm? So sánh với các tháng từ đầu năm đến tháng 8 và so với tháng 8 của các năm liền trước.
- Tại sao tại huyện X, tỷ lệ trẻ em đến trường ít hơn nhiều so với các huyện lân cận khác?
- Tại sao năm nay lại bùng phát dịch sởi, nhưng nhiều năm liền trước lại không có?
...
- Tại sao huyện Y có số lượng hộ dân chưa được sử dụng nguồn nước sạch nhiều nhất? huyện Z có tỷ lệ số hộ dân chưa được sử dụng nguồn nước sạch trên tổng số hộ dân của huyện là nhiều nhất? Có phải chủ yếu là phụ nữ đang cho biết rằng khoảng cách mà họ đi lấy nước là quá xa hay không? Tại sao nguồn nước sạch (nếu có) lại thường xuyên bị hỏng không (do thiếu kinh phí hay thiếu nhân lực, ...)? ...

2.3. Phân tích dự đoán (Predictive Analytics) - Điều gì có thể xảy ra?

- Vào thời điểm đã biết tại sao điều gì đó lại xảy ra (qua phân tích mô tả và chẩn đoán), ta có thể tiến xa hơn đến việc dự đoán điều gì có thể xảy ra tiếp theo dựa trên kiến thức về các sự kiện trước đó.
- Phân tích dự đoán giúp xem xét các cụm, xu hướng hoặc có thể là các trường hợp ngoại lệ từ đó đưa ra những dự đoán nhất định.
- Phân tích dự đoán sử dụng các kỹ thuật như máy học, dự báo, so khớp mẫu và lập mô hình dự đoán. Trong mỗi kỹ thuật, máy tính được đào tạo để thiết kế ngược các mối quan hệ nguyên nhân - kết quả trong dữ liệu.

Quy trình Phân tích dự đoán (Predictive Analytics)



Pull

Trích xuất
dữ liệu từ
nơi nó tồn
tại



Prepare

Làm sạch
tinh chỉnh
và sửa chữa
dữ liệu



Pick

Xác định
những gì
cần dự đoán



Predict

Tạo
dự
đoán



Plan

Xây dựng kế
hoạch hành
động

2.3. Phân tích dự đoán (Predictive Analytics) - Điều gì có thể xảy ra?

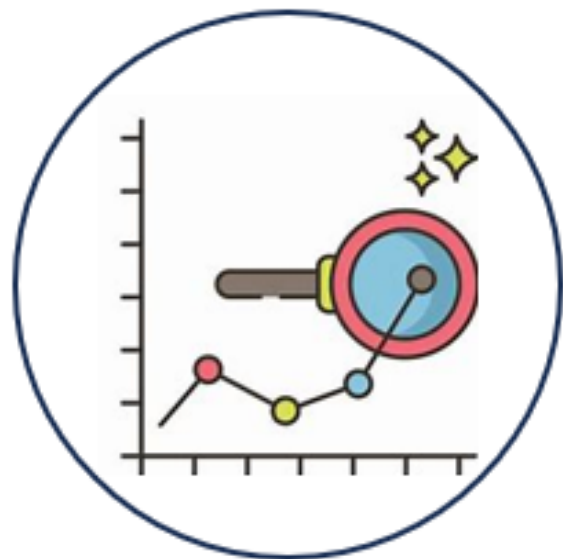
- Một số ví dụ: thông qua phân tích chẩn đoán giúp phát hiện
 - Doanh thu trung bình trong tháng 8 năm nay bị sụt giảm do các hãng đồng loạt ra sản phẩm mới vào đầu tháng 9.
 - Trẻ em huyện X đến trường ít hơn nhiều so với các huyện lân cận khác là do việc giao thông khó khăn, khoảng cách từ nhà đến trường lại quá xa, ...
 - Do tình hình dịch covid cách đây 3 năm dẫn đến số lượng trẻ em tham gia chích ngừa sụt giảm mặc dù đã được chính quyền địa phương thường xuyên nhắc nhở, ...
 - Nguồn nước sạch thường xuyên bị hỏng do đường ống đã xây dựng trên 30 năm và lại là ống kim loại (nên bị gỉ sét) ...
- Những phát hiện này thường được gọi là các mẫu trong dữ liệu. Nhìn chung có hai cách để xem xét các mẫu này: được giám sát (ví dụ: hồi quy) hoặc không được giám sát (ví dụ: phân cụm).

2.4. Phân tích đề xuất (Prescriptive Analytics) - Nên làm gì?

- Sau khi đã có ý tưởng về những gì có thể xảy ra, ta có thể muốn biết cách hành động tốt nhất là gì. Phân tích đề xuất cố gắng trả lời câu hỏi: Nên làm gì? hoặc chúng ta có thể làm gì để những việc đó sẽ (hay không) xảy ra?
- Phương pháp này không chỉ dự đoán sự kiện gì sẽ xảy ra mà còn đề xuất một phản ứng tối ưu cho kết quả đó. Nó có thể phân tích tác động tiềm ẩn của các lựa chọn khác nhau và đề xuất hướng hành động tốt nhất. Nó sử dụng phân tích đồ thị, mô phỏng, xử lý sự kiện phức tạp, mạng nơ-ron và công cụ đề xuất từ máy học.
- Phân tích đề xuất chủ yếu được sử dụng ở các công ty lớn đang tìm kiếm lời khuyên về hàng tồn kho hoặc chuỗi cung ứng của họ. Nó tiến xa hơn một bước so với phân tích mô tả và dự đoán bằng cách đề xuất các kết quả có thể xảy ra. Về cơ bản, có thể dự đoán nhiều tương lai và cho phép các công ty đánh giá một số kết quả có thể xảy ra dựa trên hành động của họ.

2.4. Phân tích đề xuất (Prescriptive Analytics) - Nên làm gì?

Quy trình Phân tích đề xuất (The Prescriptive Analytics Process)



Predictions

- Điều gì sẽ xảy ra
- Khi nào xảy ra
- Tạo sao điều đó lại xảy ra



Decisions

Những dự đoán này
sẽ mang lại lợi ích
như thế nào



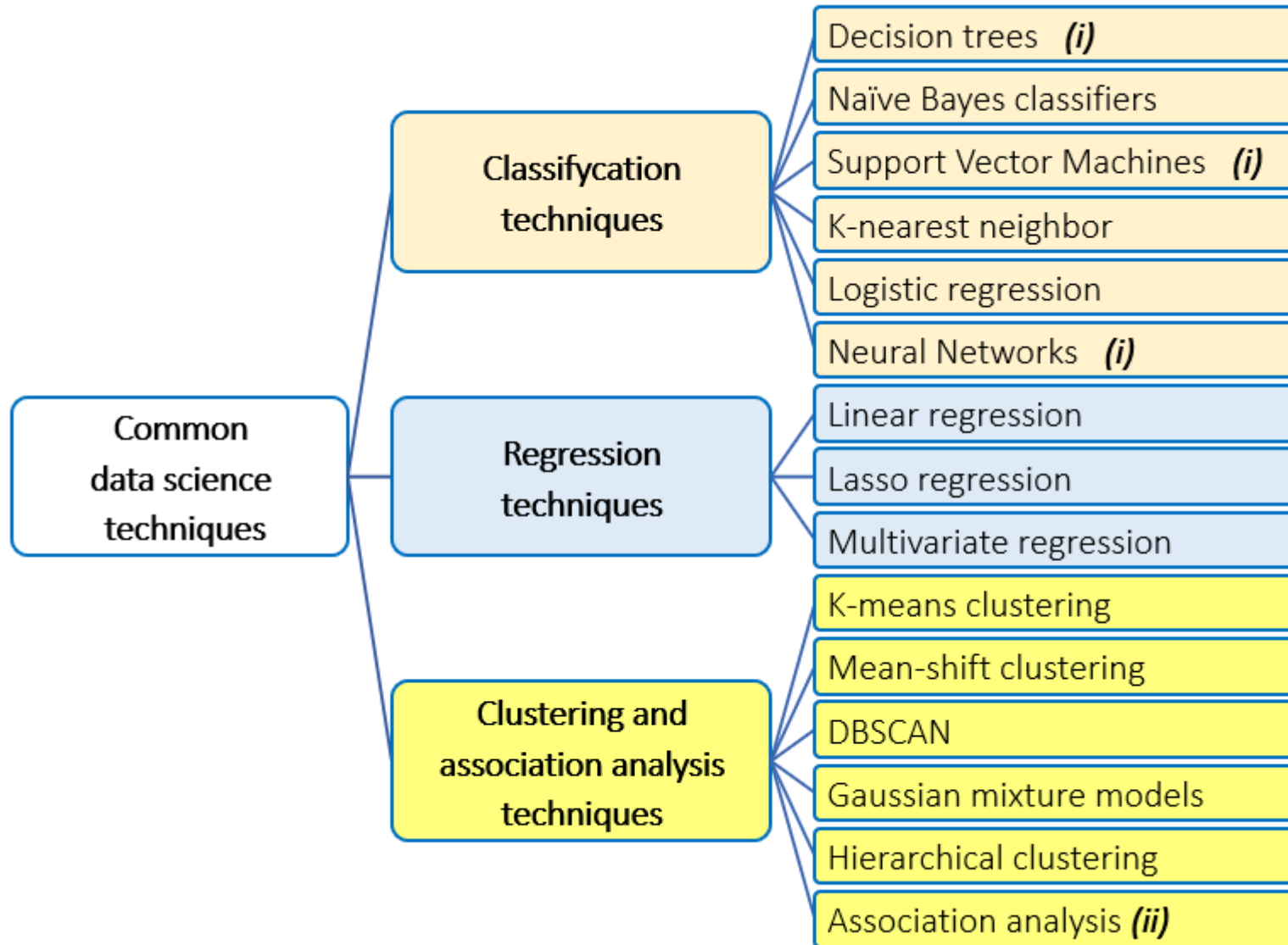
Effects

Các quyết định được
đưa ra sẽ ảnh hưởng
đến những việc còn
lại như thế nào

NỘI DUNG CHƯƠNG 1

1. Giới thiệu
2. Các phương pháp nghiên cứu chính của Khoa học dữ liệu
3. Các kỹ thuật sử dụng trong khoa học dữ liệu
4. Những công nghệ được dùng trong khoa học dữ liệu

3. CÁC KỸ THUẬT SỬ DỤNG TRONG KHOA HỌC DỮ LIỆU

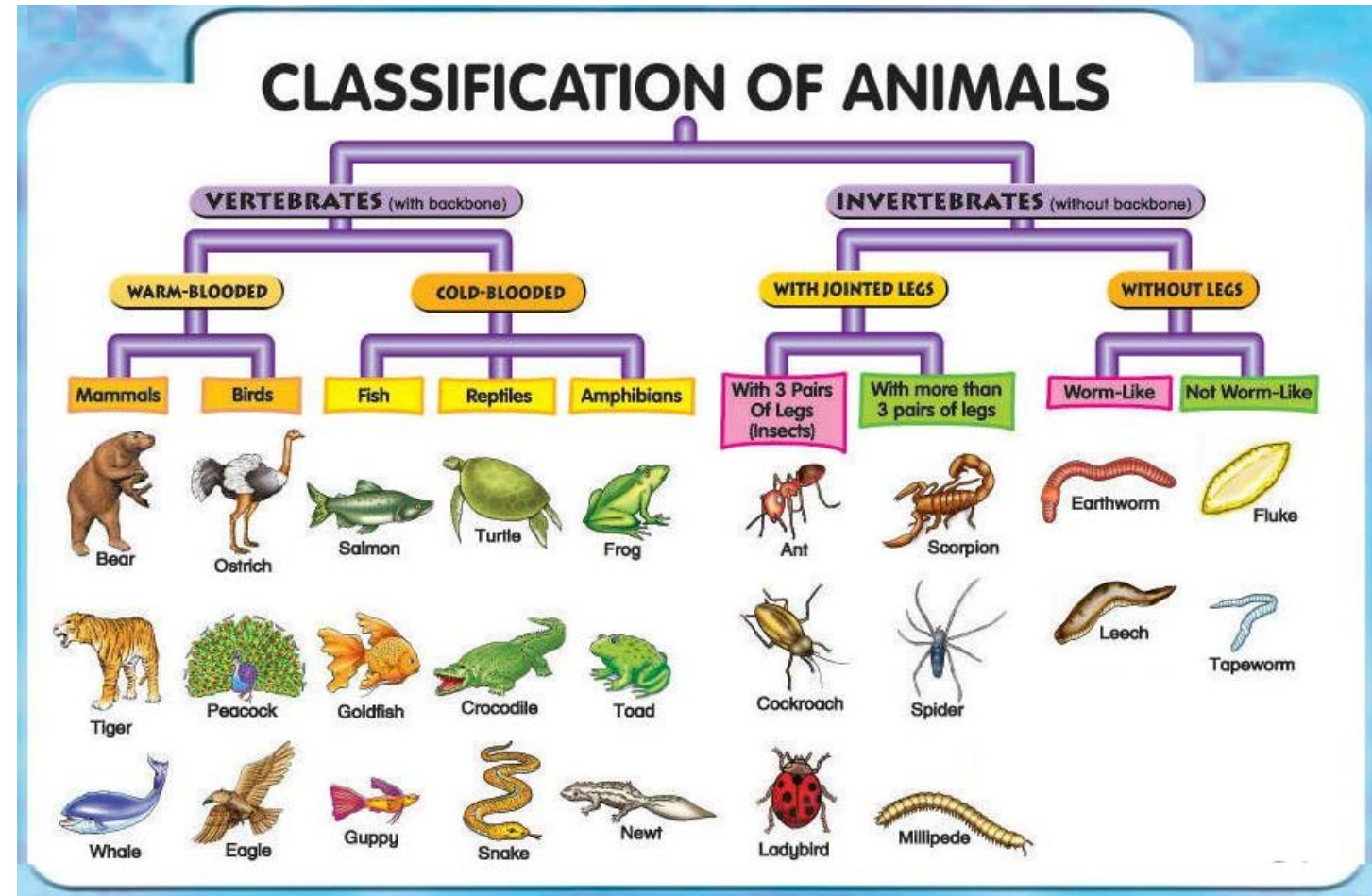


(i): Các kỹ thuật này cũng có thể được sử dụng để thực hiện hồi quy.

(ii): tuy *Association analysis* là một kỹ thuật có tính riêng biệt, nhưng ý tưởng chính trong việc đưa *Association analysis* vào nhóm này là cố gắng xác định khi nào các điểm dữ liệu sẽ xuất hiện cùng nhau.

3.1. Phân lớp (Classification)

- Phân lớp là kỹ thuật sắp xếp dữ liệu thành các nhóm hoặc danh mục cụ thể.
- Các tập dữ liệu đã xác định được sử dụng để xây dựng những thuật toán ra quyết định trong một máy tính có khả năng xử lý và phân loại dữ liệu một cách nhanh chóng.



3.1. Phân lớp (Classification)

- Một số ví dụ:

- Phân loại hình ảnh của chữ viết tay để biết hình ảnh đó đại diện cho chữ cái hoặc chữ số nào?
- Phân loại đơn xin vay tiền để biết liệu nó nên nằm trong danh mục "được phê duyệt" hay "bị từ chối".
- Phân loại để giúp xác định phương pháp điều trị cho bệnh nhân.
- Phân loại để giúp liệu thư email có phải là thư rác hay không?
- Phân loại sản phẩm theo phổ biến hoặc không phổ biến.
- Phân loại đơn bảo hiểm theo rủi ro cao hoặc rủi ro thấp
- Phân loại bình luận trên mạng xã hội thành tích cực, tiêu cực hoặc trung lập.
- ...

3.1. Phân lớp (Classification)

- Các kỹ thuật phân lớp được các nhà khoa học dữ liệu sử dụng phổ biến gồm:
 - i. *Cây quyết định (Decision trees)*: Đây là cấu trúc logic phân nhánh sử dụng các cây tham số và giá trị do máy tạo ra để phân loại dữ liệu thành các danh mục xác định.
 - ii. *Bộ phân loại Naïve Bayes (Naïve Bayes classifiers)*: Sử dụng sức mạnh của xác suất, bộ phân loại Bayes có thể giúp đưa dữ liệu vào các danh mục đơn giản.
 - iii. *Máy vector hỗ trợ (Support Vector Machines - SVM)*: SVM nhằm mục đích vẽ một đường thẳng hoặc mặt phẳng có lề rộng để phân tách dữ liệu thành các danh mục khác nhau.
 - iv. *K-láng giềng gần nhất (K-Nearest Neighbor - KNN)*: Kỹ thuật này sử dụng phương pháp "*quyết định lười biếng*" ("*lazy decision*") đơn giản để xác định danh mục mà điểm dữ liệu nên thuộc về dựa trên danh mục của các điểm lân cận gần nhất trong tập dữ liệu.

3.1. Phân lớp (*Classification*)

- Các kỹ thuật phân lớp được các nhà khoa học dữ liệu sử dụng phổ biến gồm:

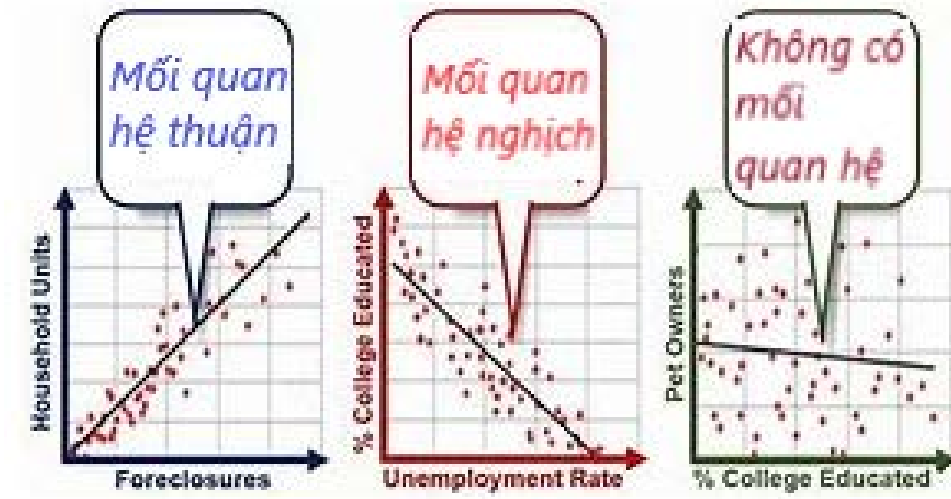
- v. Hồi quy logistic (Logistic regression):* Một kỹ thuật phân loại mặc dù có tên như vậy nhưng sử dụng ý tưởng khớp dữ liệu vào một đường (*line*) để phân biệt giữa các danh mục khác nhau ở mỗi bên. Đường (*line*) này được định hình sao cho dữ liệu được chuyển sang danh mục này hay danh mục khác thay vì cho phép có nhiều mối tương quan linh hoạt hơn.
- vi. Mạng neural (Neural Networks):* Cách tiếp cận này sử dụng mạng thần kinh nhân tạo (*Artificial Neural Networks*) được đào tạo, đặc biệt là các mạng học sâu (*deep learning*) với nhiều lớp ẩn (*multiple hidden layers*). Mạng neural đã thể hiện khả năng phân loại sâu sắc với bộ dữ liệu huấn luyện cực lớn.

3.2. Hồi quy (Regression)

- Hồi quy là phương pháp tìm ra mối quan hệ giữa 2 điểm dữ liệu dường như không liên quan. Mỗi liên kết này thường được lập mô hình xoay quanh một công thức toán học và được biểu thị dưới dạng đồ thị hoặc đường cong. Khi giá trị của một điểm dữ liệu đã được xác định, hồi quy sẽ được sử dụng để dự đoán điểm dữ liệu còn lại.

- Ví dụ:

- Tốc độ lây nhiễm của các căn bệnh lây qua đường không khí.
- Mối quan hệ giữa mức độ hài lòng của khách hàng và số lượng nhân viên.
- Mối quan hệ giữa số trạm cứu hỏa và số người bị thương do hỏa hoạn tại một địa điểm cụ thể.



Phân tích mối quan hệ hồi quy

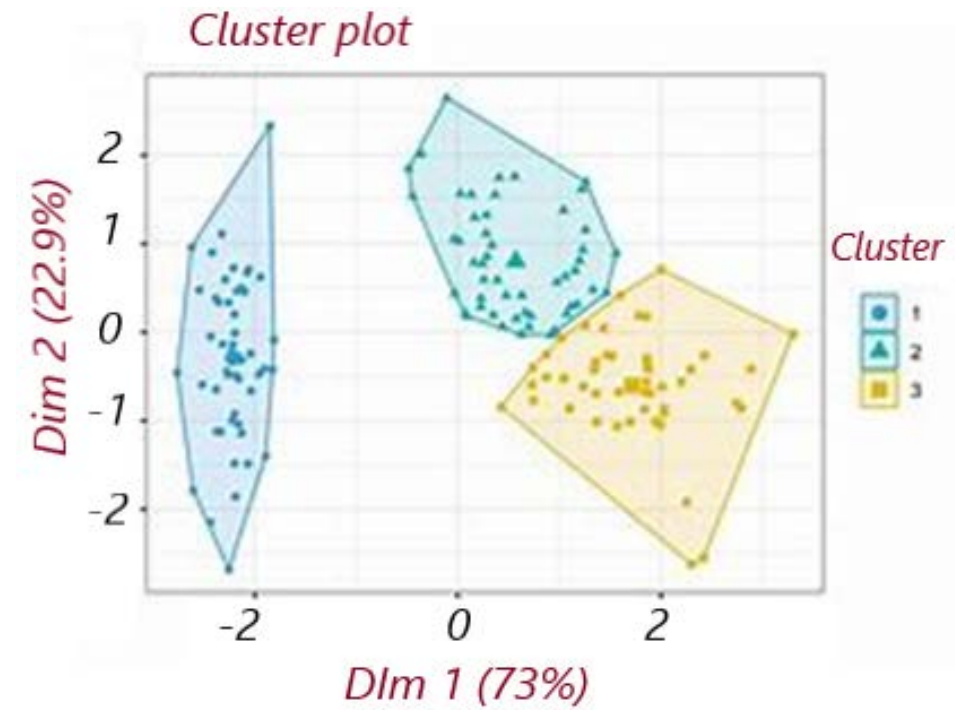
3.2. Hồi quy (Regression)

- Một số kỹ thuật hồi quy thường được các nhà khoa học dữ liệu sử dụng:

- i. *Hồi quy tuyến tính (Linear regression)*: Một trong những phương pháp khoa học dữ liệu được sử dụng rộng rãi nhất, phương pháp này cố gắng tìm ra đường phù hợp nhất với dữ liệu đang được phân tích dựa trên mối tương quan giữa hai biến.
- ii. *Hồi quy LASSO (LASSO regression)*: LASSO (*Least Absolute Shrinkage and Selection Operator* - toán tử lựa chọn và độ co tuyệt đối nhỏ nhất), là một kỹ thuật cải thiện độ chính xác dự đoán của mô hình hồi quy tuyến tính bằng cách sử dụng tập hợp con dữ liệu trong mô hình cuối cùng.
- iii. *Hồi quy đa biến (hay hồi quy bội - Multivariate regression)*: liên quan đến nhiều cách khác nhau để tìm các đường hoặc mặt phẳng phù hợp với nhiều chiều của dữ liệu có khả năng chứa nhiều biến.
- iv. Các kỹ thuật *Decision tree, SVM, Neural Network* cũng có thể được sử dụng để thực hiện hồi quy.

3.3. Phân cụm (Clustering)

- Phân cụm là phương pháp gộp các dữ liệu có liên quan chặt chẽ lại với nhau để tìm kiếm các mẫu và điểm dị thường.
- Phân cụm khác với phân lớp vì dữ liệu không thể được sắp xếp chính xác vào các hạng mục cố định. Do đó, dữ liệu được nhóm thành các mối quan hệ có khả năng xảy ra nhất.
- Thông qua phân cụm, các mẫu và mối quan hệ mới có thể được phát hiện.
- Ví dụ:
 - Nhóm những khách hàng có hành vi mua hàng giống nhau để cải thiện dịch vụ khách hàng.
 - Nhóm lưu lượng mạng để xác định mẫu sử dụng hàng ngày và nhanh chóng phát hiện một cuộc tấn công mạng.
 - Nhóm các bài viết thành nhiều hạng mục tin tức khác nhau và sử dụng thông tin này để tìm kiếm tin giả.



3.3. Phân cụm (Clustering)

- Các Kỹ thuật phân cụm phổ biến:

- i. K-means (K-means clustering)*: xác định một số cụm nhất định trong tập dữ liệu và tìm "trọng tâm" xác định vị trí của các cụm khác nhau, với các điểm dữ liệu được gán cho cụm gần nhất.
- ii. Phân cụm dịch chuyển trung bình (Mean-shift)*: là một kỹ thuật phân cụm dựa trên centroid (trọng tâm) khác, nó có thể được sử dụng riêng biệt hoặc để cải thiện khả năng phân cụm k-mean bằng cách dịch chuyển các trọng tâm được chỉ định.
- iii. DBSCAN (Density-Based Spatial Clustering of Applications with Noise - Phân cụm ứng dụng không gian dựa trên mật độ có nhiễu)*: là một kỹ thuật khác để khám phá các cụm sử dụng phương pháp xác định mật độ cụm nâng cao hơn.

3.3. Phân cụm (Clustering)

- Các Kỹ thuật phân cụm phổ biến:

iv. Mô hình hỗn hợp Gaussian (Gaussian Mixture Models – GMM): giúp tìm các cụm bằng cách sử dụng phân phối *Gaussian* để nhóm dữ liệu lại với nhau thay vì coi dữ liệu là các điểm đơn lẻ.

v. Phân cụm theo thứ bậc (Hierarchical): tương tự như cây quyết định, kỹ thuật này sử dụng cách tiếp cận phân nhánh, phân cấp để tìm các cụm.

vi. Phân tích kết hợp (Association analysis) Phân tích kết hợp (Association analysis) là một kỹ thuật có liên quan nhưng riêng biệt. Ý tưởng chính đằng sau kỹ thuật này là tìm các quy tắc kết hợp mô tả điểm chung giữa các điểm dữ liệu khác nhau. Tuy nhiên, việc đưa *Association analysis* vào nhóm này nhằm cố gắng xác định khi nào các điểm dữ liệu sẽ xuất hiện cùng nhau, mức độ liên kết giữa chúng thay vì chỉ xác định các cụm của chúng.

NỘI DUNG CHƯƠNG 1

1. Giới thiệu
2. Các phương pháp nghiên cứu chính của Khoa học dữ liệu
3. Các kỹ thuật sử dụng trong khoa học dữ liệu
4. Những công nghệ được dùng trong khoa học dữ liệu

4. NHỮNG CÔNG NGHỆ ĐƯỢC DÙNG TRONG KHOA HỌC DỮ LIỆU

- i. Trí tuệ nhân tạo (Artificial intelligence):* Các mô hình máy học và phần mềm liên quan được sử dụng để phân tích dự đoán và phân tích đề xuất.
- ii. Điện toán đám mây (Cloud computing):* Công nghệ đám mây đã trao cho các nhà khoa học dữ liệu sự linh hoạt và sức mạnh xử lý cần thiết để phân tích dữ liệu nâng cao.
- iii. Internet vạn vật (Internet Of Things):* IoT đề cập đến hàng loạt các thiết bị có thể tự động kết nối với Internet. Những thiết bị này thu thập dữ liệu cho các sáng kiến khoa học dữ liệu. Chúng tạo ra khối lượng dữ liệu đồ sộ có thể được sử dụng để khai thác dữ liệu và trích xuất dữ liệu.
- iv. Máy tính lượng tử (Quantum Computing):* Máy tính lượng tử có thể thực hiện các phép tính phức tạp ở tốc độ cao. Các nhà khoa học dữ liệu trình độ cao sử dụng chúng để xây dựng các thuật toán định lượng phức tạp.

