TỔNG QUAN
DATA SCIENCE

Lê Văn Hạnh
levanhanhvn@gmail.com

TÀI LIỆU THAM KHẢO

1. Jiawei Han, Micheline Kamber, Jian Pei – Data mining Concept and Techniques - Morgan Kaufmann, 2012, Third edition
2. Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition, Cambridge University Press, March 2020. ISBN: 978-1108473989. (https://dataminingbook.info/book_html/)
3. Mohammed J. Zaki and Wagner Meira, Jr - Data Mining and Machine Learning: Fundamental Concepts and Algorithms - Cambridge University Press, March 2020, ISBN: 978-1108473989 - Second Edition (https://dataminingbook.info/book_html/)
4. PGS.TS. Đỗ Phúc – Giáo trình Khai thác dữ liệu – NXB Đ5ai học Quốc Gia TP.Hồ Chí Minh, 2005.

CÁCH TÍNH ĐIỂM TRONG MÔN HỌC

1. Điểm thường kỳ: **40%**
    - Chuyên cần: 5%
    - Bài tập trên lớp: 5%
    - Đề tài nhóm: 30%
2. Điểm cuối kỳ: **60%**
    - Đề tài cá nhân: 60%

ĐỀ TÀI NHÓM TRONG MÔN HỌC

Mỗi nhóm chọn 1 trong các phần mềm sau:

1. Tableau	6. Looker Studio	11. Keras
2. Power BI	7. OpenRefine	12. Rapid Miner
3. Orange	8. Qlik	13. Talend
4. SPSS	9. SAS	14. KNIME
5. Weka	10. Apache Spark	

YÊU CẦU CHUNG:

- Lớp trưởng tạo và share file excel để SV đăng ký đề tài nhóm và đề tài cá nhân
- Nếu phần mềm không hỗ trợ đủ tính năng chính là tiền xử lý dữ liệu và khai thác dữ liệu => nhóm phải chọn thêm phần mềm thứ 2 để hoàn thiện đề tài nhóm.
- Viết báo cáo nhóm, trong đó cần nêu rõ:
  - Cách cài đặt
  - Giới thiệu các chức năng mà công cụ hỗ trợ.
  - Hướng dẫn cách sử dụng công cụ để tiền xử lý dữ liệu và khai thác dữ liệu
  - Từ buổi thứ 2 trở đi, hàng tuần mỗi nhóm sẽ thực hiện báo cáo nhanh về các công việc đã thực hiện.
  - Ghi rõ việc phân công công việc cho từng cá nhân trong đề tài

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***

- Dữ liệu có thể thuộc tất cả các lĩnh vực (logistic, kinh doanh, kinh tế, xã hội, y tế, khí hậu, môi trường, …)
- SV tự thu thập dữ liệu (tối thiểu 500 records), có thể download từ một số website (Kaggle, WHO, UNDP, …)

– ***Về đề tài cuối môn học***

- Dựa trên lý thuyết được giới thiệu trong môn học, đề tài sẽ thực hiện phân tích trên dữ liệu đã thu thập.
- Công cụ sử dụng:
  - SV sử dụng Python để lập trình thực hiện đề tài sẽ có thể đạt được điểm tuyệt đối (10 điểm).
  - Nếu không sử dụng Python, SV phải sử dụng công cụ mà đề tài nhóm đã nghiên cứu (diểm tối đa là 8 điểm)
- Dàn ý mẫu báo cáo đề tài cá nhân sẽ được GV cung cấp.

Quy ước: Hai SV bất kỳ không được sử dụng cùng bộ dữ liệu và cùng công cụ khai thác dữ liệu

NỘI DUNG MÔN HỌC

1. Tổng quan về Data Science
2. Tìm hiểu dữ liệu
3. Tiền xử lý dữ liệu
4. Khai thác các mẫu phổ biến, mối kết hợp và mối tương quan
5. Phân loại (*Classification*)
6. Phân tích cụm (*Cluster analysis*)

Chương 1

TỔNG QUAN
DATA SCIENCE

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 Văn Tuấn – Phân tích dữ liệu với R – NXB Tổng hợp TP.HCM 2022
4. The Art of Data Science.
5. The Art of Statistics.
6. Storytelling with Data.
7. Good Charts.
8. Introduction to Machine Learning with Python.
9. The Hundred Page Machine Learning Book.
10. AI and Machine Learning for Coders.
11. Deep Learning with Python.
12. Foundations of Deep Reinforcement Learning
13. Deep Learning Illustrated

NỘI DUNG

1. Giới thiệu
2. Lịch sử
3. Khoa học dữ liệu
4. Các phương thức phân tích dữ liệu
5. Các phương pháp phân tích dữ liệu phổ biến
6. Một số kỹ thuật phân tích dữ liệu
7. Các kỹ thuật khai thác dữ liệu
8. Các công cụ phân tích dữ liệu phổ biến
9. Các kỹ năng cần có đối với phân tích viên

1. GIỚI THIỆU

- Dữ liệu đóng một vai trò rất quan trọng trong việc vận hành, ứng dụng cũng như lưu trữ thông tin của người dùng. Ngày nay, với lượng dữ liệu được phát sinh là vô cùng lớn, nếu các Công ty, Tập đoàn rút ra được những tri thức có được từ việc rút trích và phân tích trên lượng dữ liệu hiện có sẽ cực kỳ hữu ích cho các hệ thống ra quyết định và hỗ trợ cuộc sống.
- Ngày nay, hầu hết các công ty và tập đoàn lớn đều đã có những đội ngũ, chuyên gia phân tích dữ liệu của riêng họ. Có thể kể đến là Facebook, Google, ... Sự thành công của các công ty và tổ chức ngày nay đều ít nhiều có liên quan với ngành Khoa học dữ liệu (KHDL). KHDL đang lan rộng ảnh hưởng của nó và mang lại ý nghĩa ngày càng quan trọng hơn đối với đời sống con người.
- *KHDL là khoa học liên quan đến việc quản trị và phân tích dữ liệu, trích xuất các giá trị từ dữ liệu để tìm ra những hiểu biết, các tri thức hành động và đưa ra các quyết định dẫn dắt hành động.*

2. LỊCH SỬ

- ***1960-1996***: thuật ngữ "Khoa học dữ liệu" (data science) đã được sử dụng trong nhiều tài liệu nói về các phương pháp tính toán.
- ***11/1997***: thuật ngữ KHDL mới được dùng chính thức bởi một nhà nghiên cứu tên là Chien-Fu Jeff Wu. Trong bài thuyết trình mang tên "*Statistics = Data Science?*" tại Đại học Michigan, Chien-Fu Jeff Wu đã phổ biến thuật ngữ "Khoa học dữ liệu" và nói rằng thống kê nên được đổi tên thành KHDL và nhà thống kê thành nhà KHDL vì họ đã dành phần lớn thời gian của mình để thao tác và thử nghiệm với dữ liệu.
- ***Năm 2001***: William S. Cleveland đã giới thiệu KHDL như là một ngành độc lập.

2. Lịch sử

- ***4/2002***: International Council for Science cho ra đời Tạp chí KHDL, một ấn phẩm tập trung vào các vấn đề như mô tả hệ thống dữ liệu, ấn phẩm của họ trên internet, các ứng dụng và các vấn đề pháp lý.
- ***01/2003***: Đại học Columbia bắt đầu xuất bản tạp chí KHDL, nhằm cung cấp một công cụ cho tất cả nhân viên dữ liệu trình bày quan điểm của mình và trao đổi ý kiến.
- ***2008***: DJ Patil và Jeff Hammerbacher mới sử dụng thuật ngữ "nhà KHDL" để xác định công việc của họ tại LinkedIn và Facebook.

2. Lịch sử

- ***2013***: Nhóm công tác của IEEE về KHDL và Phân tích nâng cao đã được đưa ra
- ***2014***: tổ chức hội nghị quốc tế đầu tiên về KHDL và Phân tích nâng cao của IEEE (*Institute of Electrical and Electronics Engineers*).
- ***2015***: Tạp chí Quốc tế về KHDL và Phân tích đã được lập bởi Springer để xuất bản tác phẩm ban đầu về KHDL và phân tích dữ liệu lớn.
- ***Hiện nay***: KHDL trở thành 1 ngành khoa học không thể thiếu của hầu hết các Công ty, Tập đoàn, ...

3. KHOA HỌC DỮ LIỆU

1. Data Science là gì?

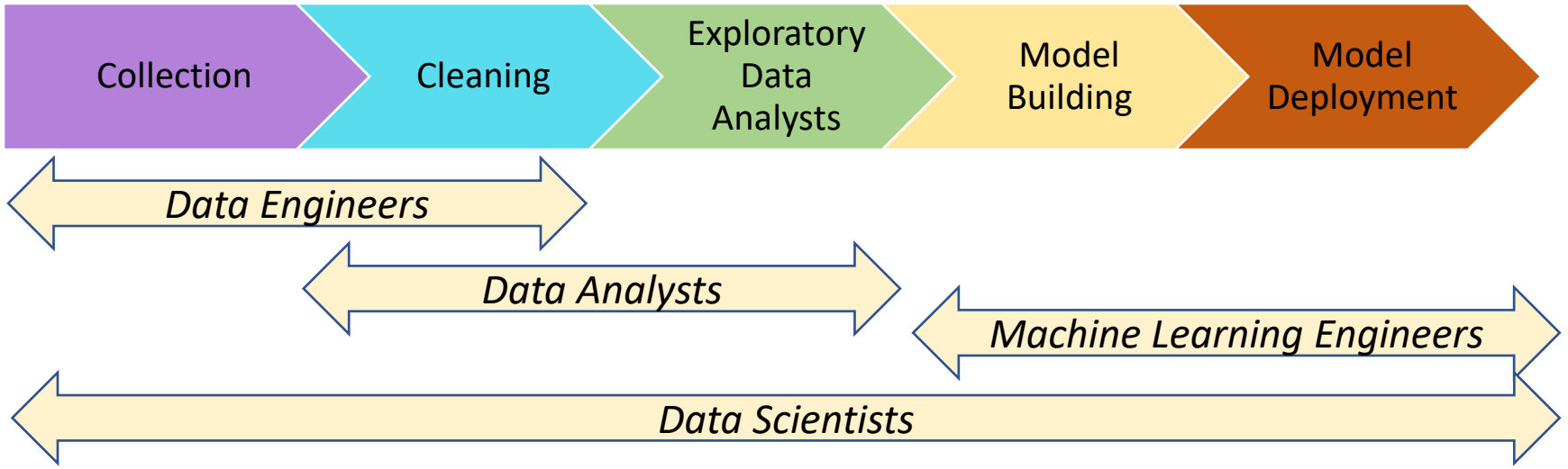
- Data science là khái niệm chỉ tất cả công việc thu thập, khai thác và phân tích dữ liệu để tìm ra được những giá trị hữu ích (insight). Từ đó, nhà quản trị sẽ trực quan hoá các insight này cho các bên liên quan, chuyển hóa nó thành hành động cụ thể.
- Hiện dữ liệu tăng lên theo cấp số nhân. Nhờ đó, khả năng mở ra cơ hội mới cho việc phân tích và chuyển hóa insight ý nghĩa từ Data là rất lớn. Theo đó, yêu cầu về nguồn nhân lực làm trong ngành Data science đang trở nên cấp thiết hơn bao giờ hết.
- Nhân lực ngành này phải có kỹ năng sử dụng các công cụ thống kê, machine learning và khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần lập trình cụ thể. Có thể hiểu một cách đơn giản, Data Science là một lĩnh vực liên quan tới xử lý dữ liệu, phân tích và trích xuất các thông tin chi tiết từ dữ liệu.

2. Một số ứng dụng của KHDL

- Dựa trên phân tích dữ liệu về nhu cầu thị trường để quyết định cần nuôi bao nhiêu lợn trong từng khoảng thời gian trong năm
- Dựa trên phân tích được dữ liệu mô phỏng các phương án xả lũ vào mùa mưa ta có thể chọn được cách xả lũ ít thiệt hại nhất.
- Dựa trên bệnh án điện tử của người bệnh, có thể tìm ra được phác đồ thích hợp để điều trị cho người bệnh.
- Dựa trên phân tích các lần mua hàng trước của khách hàng để dự đoán những món đồ mà khách hàng có thể sẽ thích mua và gửi quảng cáo tới.
- Thông qua khảo sát dữ liệu có được từ các phim sắp chiếu, sở thích xem phim của khách hàng để đưa ra các dự đoán về thời gian phát hành, giờ công chiếu, doanh thu, …
- Dựa trên các dữ liệu truy vấn tìm kiếm của khách hàng để cảnh báo bệnh cúm trong một quần thể.
- …

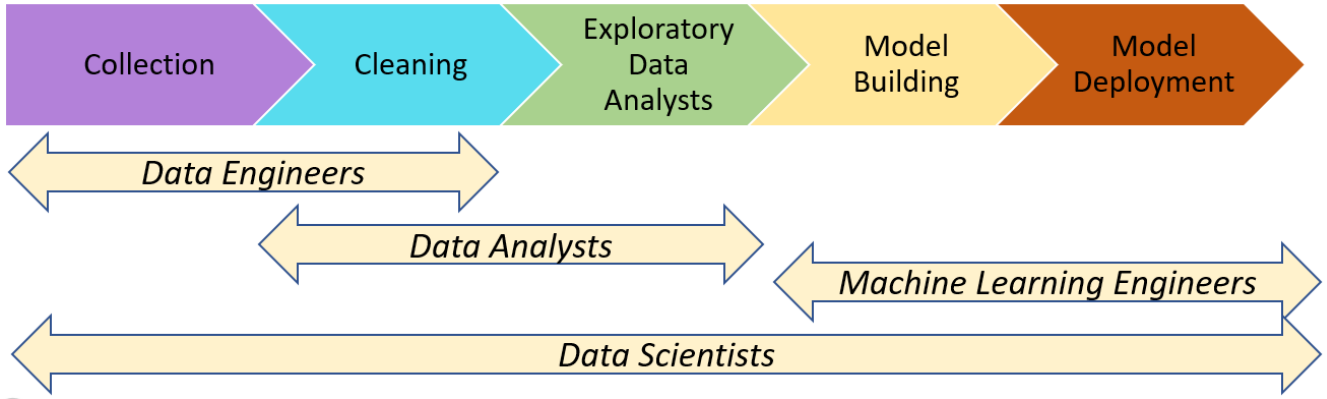
3. Các thành phần của KHDL

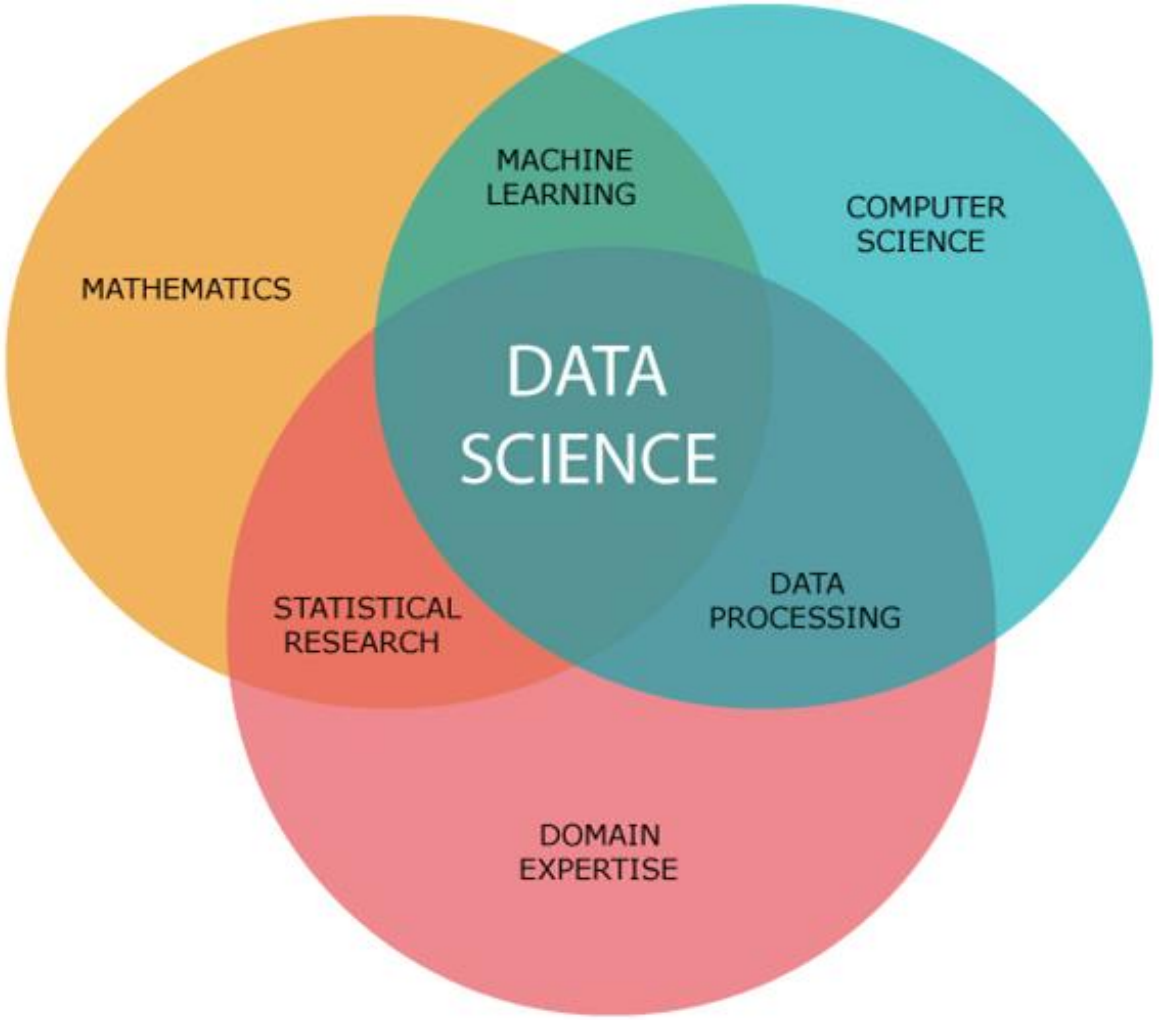
- KHDL là khoa học về việc quản trị và phân tích dữ liệu để tìm ra các hiểu biết, các tri thức hành động, các quyết định dẫn dắt hành động.
- KHDL gồm ba phần chính:
  - Tạo ra và quản trị dữ liệu (số hóa dữ liệu)
  - Phân tích dữ liệu (phân tích, dự đoán, dự báo, …)
  - Chuyển kết quả phân tích thành giá trị của hành động.



3. Các thành phần của KHDL

- Việc phân tích và dùng dữ liệu lại dựa vào ba nguồn tri thức: toán học (thống kê toán học), công nghệ thông tin (máy học) và tri thức của lĩnh vực ứng dụng cụ thể.



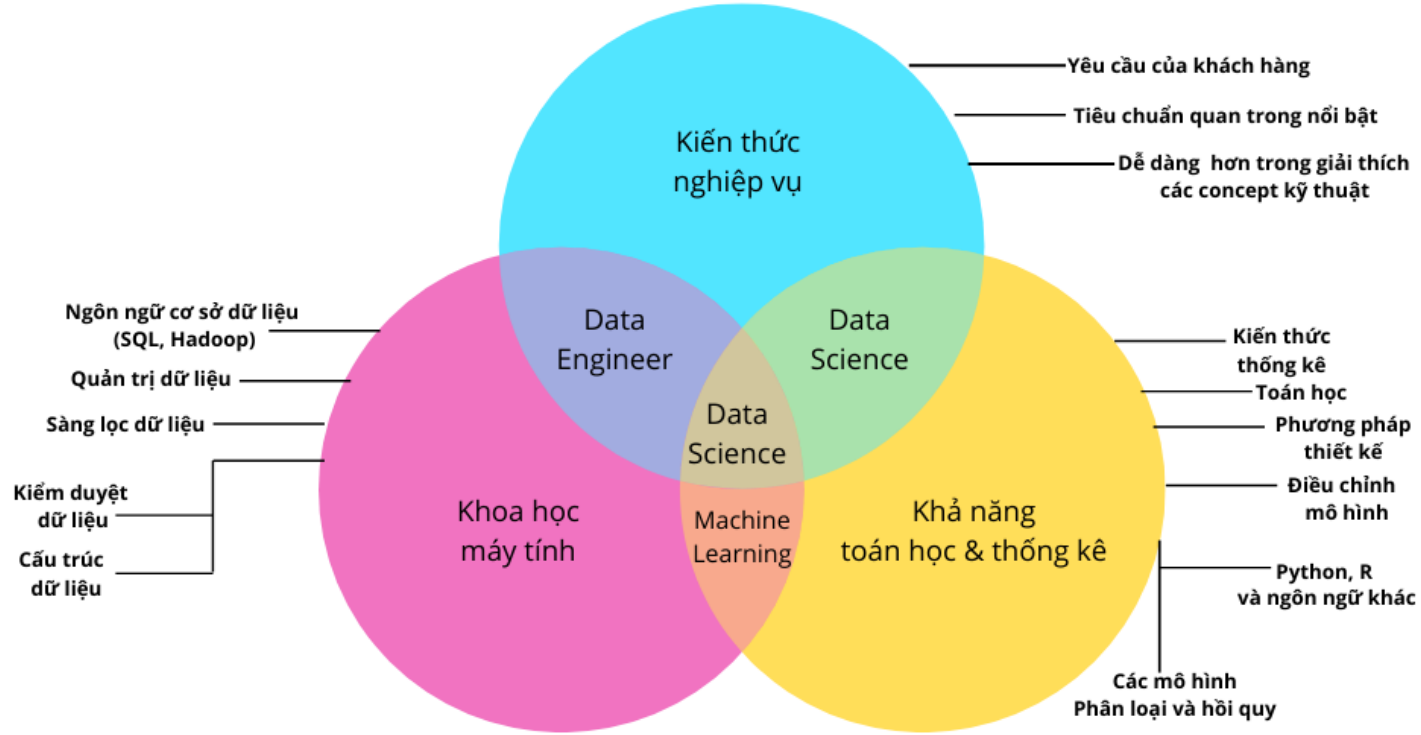


4. Lĩnh vực liên ngành của KHDL

- Như tên gọi, KHDL là một ngành khoa học nghiên cứu về dữ liệu, tức là đối tượng nghiên cứu chính của ngành là dữ liệu.
- Do nhu cầu phát triển của các ứng dụng và cuộc sống con người, dữ liệu:
  - Rất đa dạng có thể đến từ mọi nơi, mọi lĩnh vực.
  - Có khối lượng, tốc độ phát sinh rất lớn.
- Do vậy, ở góc độ chuyên ngành, KHDL là một lĩnh vực nghiên cứu liên ngành vì:
  - KHDL khảo sát rất nhiều loại dữ liệu đến từ các lĩnh vực chuyên ngành khác nhau, về các quá trình và các hệ thống rút trích tri thức hoặc hiểu biết từ dữ liệu ở các dạng khác nhau (có cấu trúc hay phi cấu trúc)
  - KHDL là sự tiếp nối của một số lĩnh vực phân tích dữ liệu như:
    - Khoa học thống kê
    - Khai thác dữ liệu.

5. Mục tiêu chính của KHDL

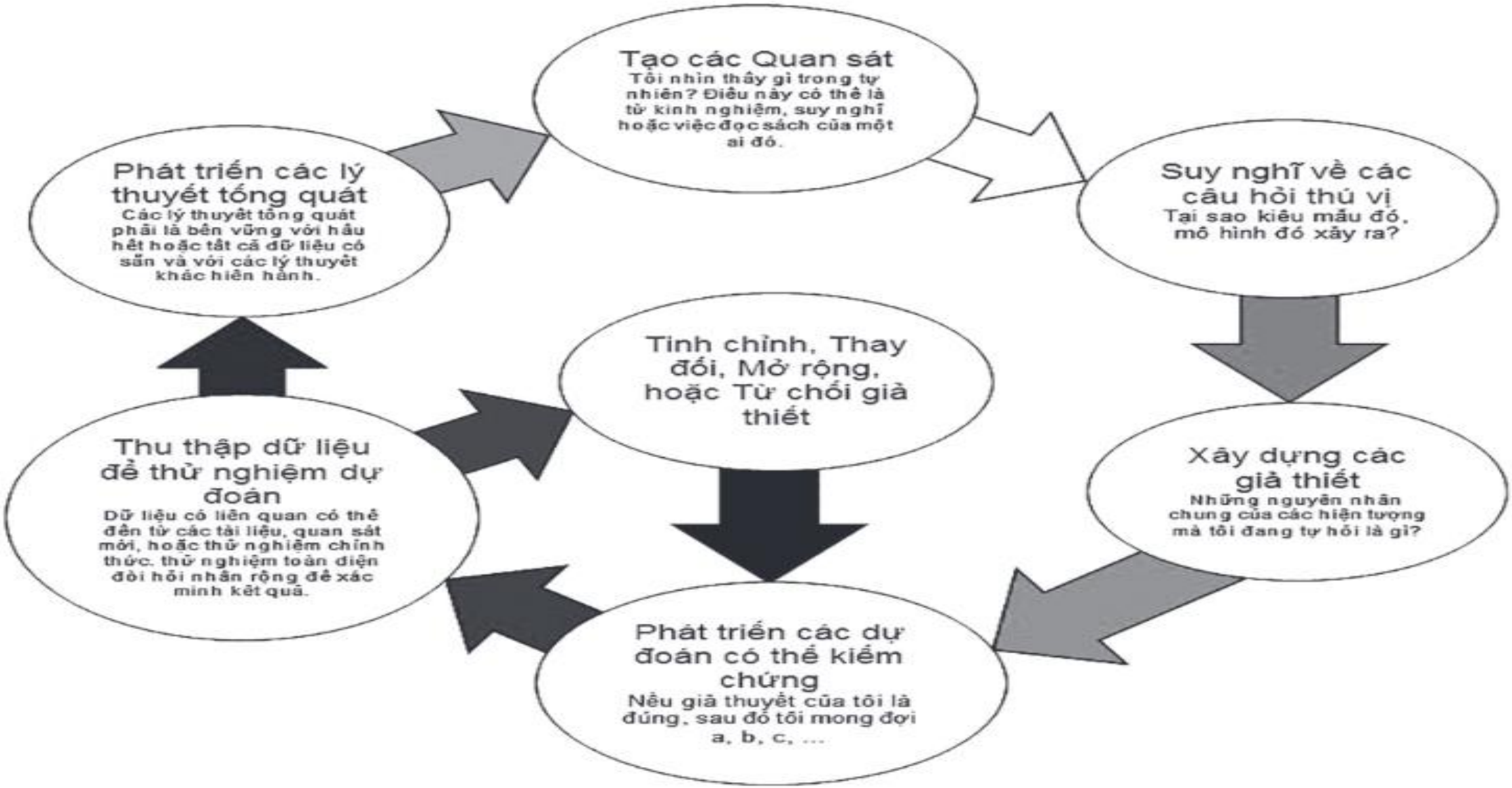
- Mục tiêu chính của ngành KHDL là để có được cái nhìn sâu hơn vào dữ liệu và tạo ra những điều hữu ích cho cuộc sống con người. Thông qua quá trình tiếp nhận, phân tích các đặc tính và rút được các kết quả từ dữ liệu sẽ hỗ trợ chúng ta trong việc đưa ra các quyết định, các dự đoán tốt hơn cho các hệ thống.



6. Một số lỗi thường gặp khi sử dụng KHDL

- Có tư duy sai lầm rằng KHDL có thể giải quyết được bất kỳ vấn đề nào trong thế giới thực.
- Không nên khai thác dữ liệu dưới dạng offline vì có thể đưa ra những giải pháp không phù hợp, lạc hậu, … do dữ liệu không ngừng lớn lên và do nhu cầu, xu hướng, môi trường, … thường xuyên thay đổi.
- Một số lỗi khác:
  - Bắt đầu phân tích mà không đặt câu hỏi.
  - Sử dụng dữ liệu chất lượng kém.
  - Chỉ tập trung vào công nghệ mà không quan tâm đến cơ sở lý thuyết và kiến thức chuyên môn.
  - Nhầm lẫn sự tương quan (*correlation*) và quan hệ nhân quả (*causation*).
  - Thất bại trong việc truyền đạt (*communicate*) các kết quả
  - Làm phức tạp việc phân tích dữ liệu.
  - …

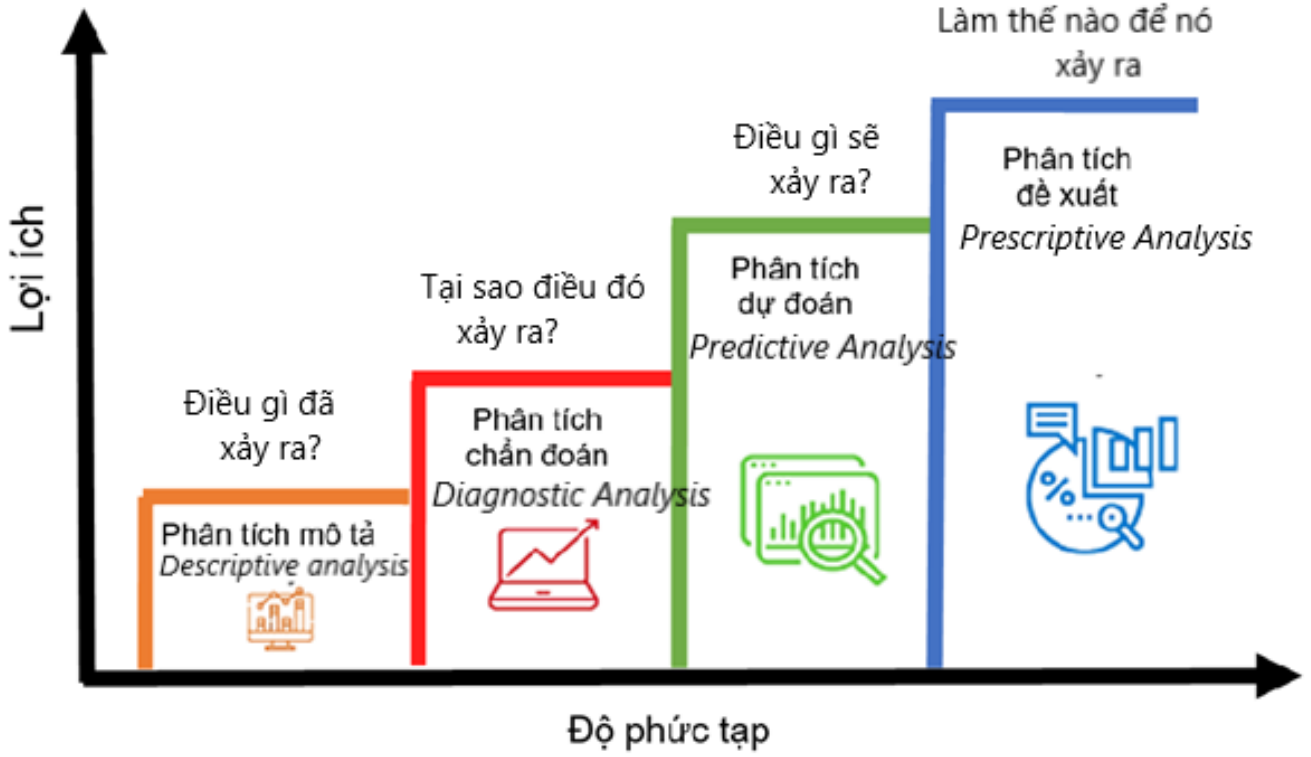
7. Đề xuất quy trình khi sử dụng KHDL

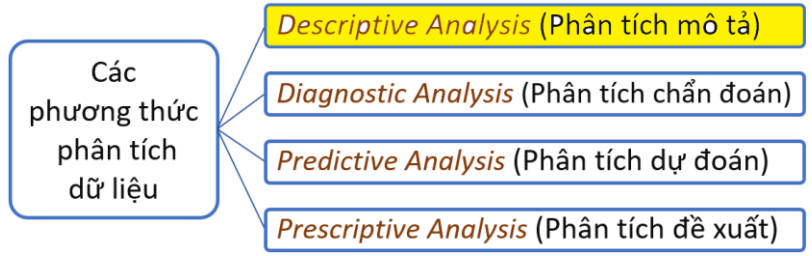


4. CÁC PHƯƠNG THỨC PHÂN TÍCH DỮ LIỆU

Các phương thức phân tích dữ liệu

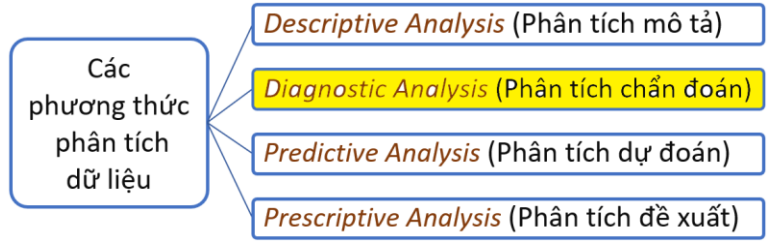
- *Descriptive Analysis* (Phân tích mô tả)
- *Diagnostic Analysis* (Phân tích chẩn đoán)
- *Predictive Analysis* (Phân tích dự đoán)
- *Prescriptive Analysis* (Phân tích đề xuất)





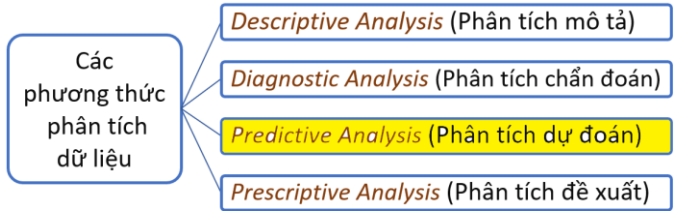
4.1. Descriptive analysis (Phân tích mô tả)

- Là điểm khởi đầu cho bất kỳ quá trình phân tích nào. Là một quy trình tóm tắt dữ liệu quá khứ thành một biểu mẫu mà mọi người có thể dễ dàng đọc được để biết được "*điều gì đã xảy ra?*".
- Nguồn dữ liệu của phân tích mô tả là dữ liệu thu thập được trong một khoảng thời gian nhất định nhằm mô tả những gì đã xảy ra trong vấn đề/lĩnh vực cần nghiên cứu:
  - Trong kinh doanh là các loại báo cáo tài chính, báo cáo doanh thu, tình hình kinh doanh của công ty, ….
  - Trong khí tượng thủy văn là số liệu về tình hình thời tiết, mưa bão, hạn hán.
  - …
- Phân tích mô tả cũng giúp dữ liệu được sắp xếp và sẵn sàng tiến hành phân tích thêm.



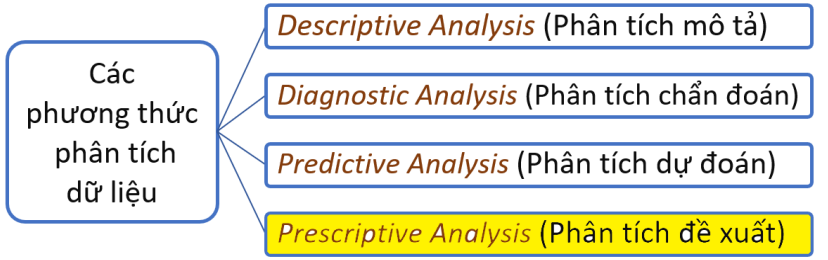
4.2. Diagnostic Analysis (Phân tích chẩn đoán)

- Là phương pháp dựa trên dữ liệu để tìm hiểu "*nguyên nhân để vấn đề đó xảy ra*".
- Phân tích chẩn đoán sẽ tiến thêm một bước nữa để khám phá ra lý do đằng sau 1 kết quả hoặc kết luận.
- Phân tích chẩn đoán thường được thực hiện bằng cách sử dụng các kỹ thuật như:
  - *Exploratory analysis* (khám phá dữ liệu)
  - *Drill-down* (xem chi tiết)
  - *Correlations* (các mối tương quan)
  - *Data recovery* (khôi phục dữ liệu)



4.3. Predictive Analysis (Phân tích dự đoán)

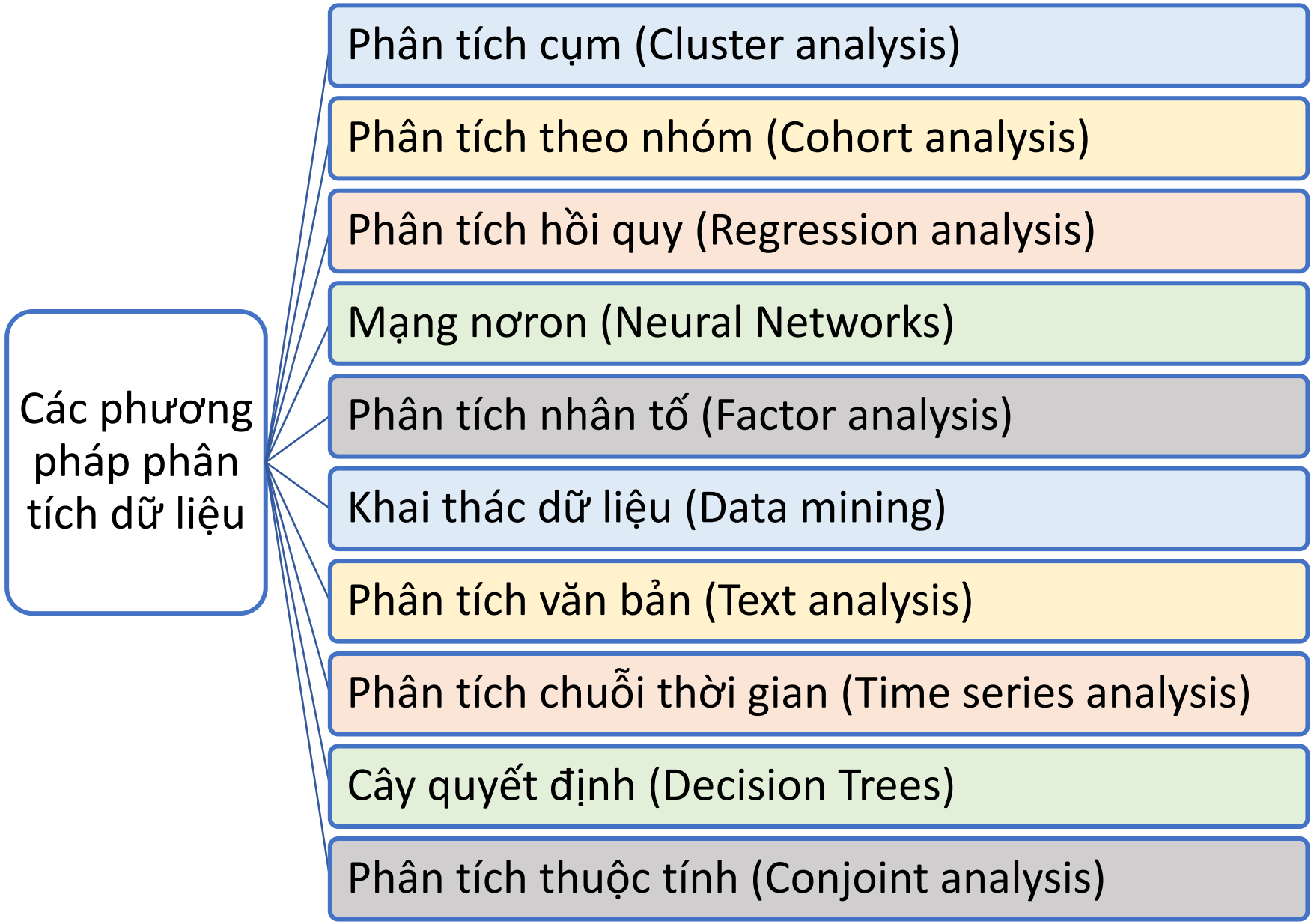
- Phương pháp dự đoán cho phép dự đoán được "điều gì sẽ xảy ra?" dựa trên:
  - Kết quả của phân tích mô tả, phân tích chẩn đoán.
  - *Machine Learning* (ML - học máy)
  - *Artificial Intelligence* (AI - trí tuệ nhân tạo).
  - *Data mining* (khai thác dữ liệu)
- Thông qua Predictive Analysis giúp phát hiện ra các xu hướng trong tương lai. Ví dụ:
  - *Lĩnh vực kinh doanh*: PA giúp giảm rủi ro, tối ưu hóa hoạt động và tăng doanh thu.
  - *Ngành tài chính*: giúp phát hiện và giảm gian lận, đo lường rủi ro tín dụng, tối đa hóa cơ hội bán kèm/bán thêm và giữ chân khách hàng có giá trị. VD: Ngân hàng Commonwealth sử dụng phân tích để dự đoán khả năng xảy ra hoạt động gian lận đối với bất kỳ giao dịch nhất định nào trước khi được phép – trong vòng 40 mili giây kể từ khi bắt đầu giao dịch.



6.4. Prescriptive Analysis (Phân tích đề xuất)

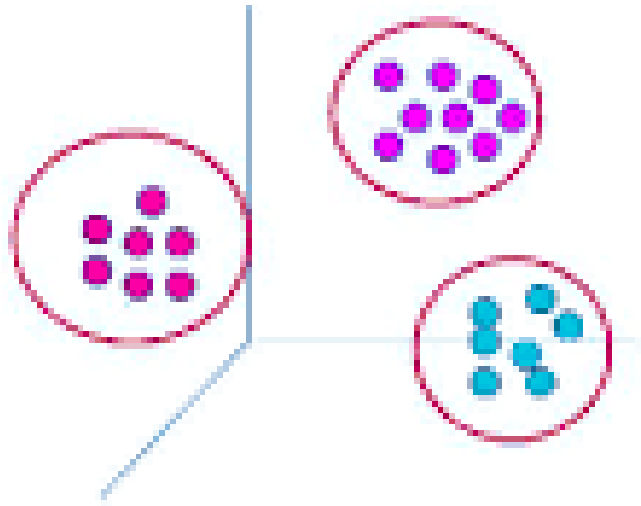
- Phân tích đề xuất là một quy trình phân tích dữ liệu và đưa ra các đề xuất tức thì về cách tối ưu hóa các phương thức kinh doanh để phù hợp với nhiều kết quả dự đoán.
- Phân tích đề xuất nhằm trả lời cho câu hỏi "*Nó sẽ diễn ra như thế nào?*" và "*Nên làm gì tiếp theo?*".
- Về bản chất, phân tích đề xuất lấy "những gì chúng ta biết" (dữ liệu), hiểu một cách toàn diện dữ liệu đó để dự đoán những gì có thể xảy ra và đề xuất các phương án tốt nhất dựa trên các kết quả phân tích mô phỏng.
- Bằng cách phân tích dữ liệu từ các nguồn dựa trên từ ngữ khác nhau, bao gồm đánh giá sản phẩm, bài báo, thông tin liên lạc trên mạng xã hội và câu trả lời khảo sát, sẽ giúp có được những hiểu biết về đối tượng cần nghiên cứu, cũng như nhu cầu, sở thích và khó khan của họ.

5. CÁC PHƯƠNG PHÁP PHÂN TÍCH DỮ LIỆU PHỔ BIẾN



5.1. Phân tích cụm (Cluster analysis)

- Thực hiện dựa trên việc nhóm các phần dữ liệu có đặc điểm chung với nhau. Vì không có biến đích khi phân nhóm, phương pháp này thường được sử dụng để tìm các mẫu ẩn trong dữ liệu hoặc cung cấp ngữ cảnh bổ sung cho 1 xu lướng hoặc một tập dữ liệu.
- Trong một thế giới hoàn hảo, các nhà tiếp thị có thể phân tích từng khách hàng riêng biệt và cung cấp cho họ dịch vụ được cá nhân hóa tốt nhất
- Nhưng với một lượng lớn khách hàng, không thể làm được điều đó kịp thời, đúng lúc. Vì vậy xuất hiện tính năng phân nhóm bằng cách nhóm khách hàng thành các cụm dựa trên:
  - Nhân khẩu học
  - Hành vi mua hàng
  - Khả năng tài chính,
  - …



5.2. Phân tích theo nhóm (Cohort analysis)

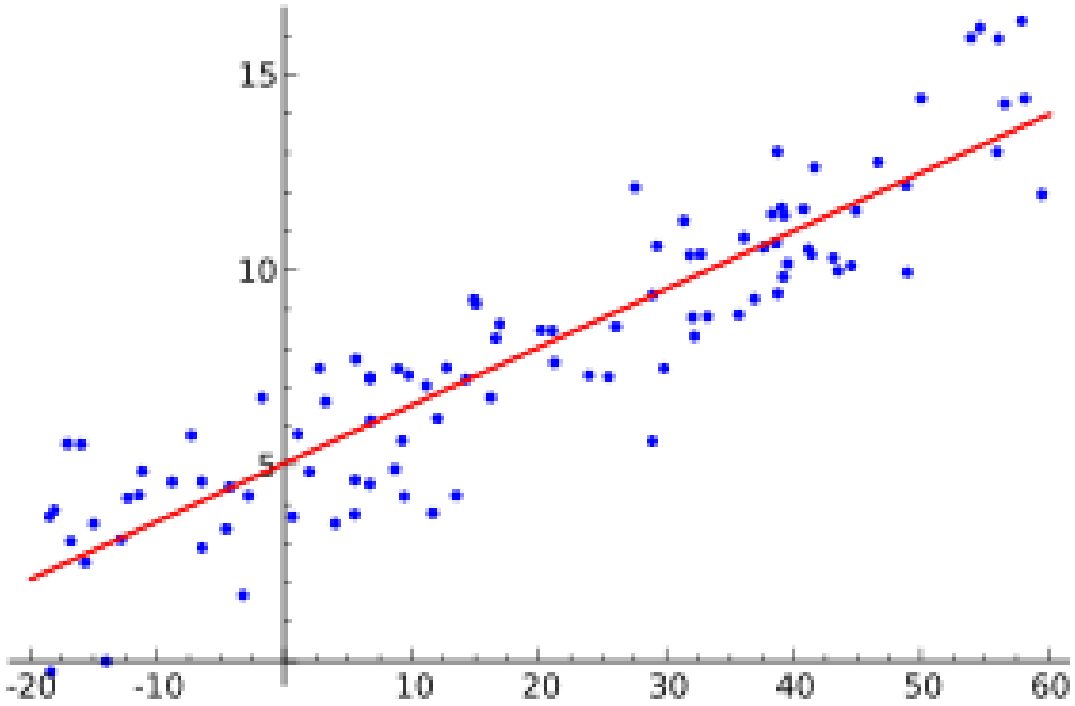
- Phương pháp này sử dụng dữ liệu lịch sử để kiểm tra và đối chiếu một phân khúc xác định về hành vi của người dùng, sau đó nhóm chúng lại với những phân khúc khác có đặc điểm tương tự. Bằng phương pháp này, có thể hiểu được nhu cầu của người tiêu dùng, thậm chí là một nhóm đối tượng mục tiêu cụ thể với số lượng lớn hơn.
- Ví dụ: cần quảng bá 1 sản phẩm qua internet. Bạn tạo hai phiên bản quảng cáo với các thiết kế, nội dung và dịch vụ đáp ứng đến khách hàng khác nhau.

Sau đó, sử dụng phân tích theo nhóm để theo dõi hiệu suất của chiến dịch trong một khoảng thời gian để hiểu loại nội dung, hình thức nào tác động tốt đến sức tiêu thụ của khách hàng hoặc khách hàng đang cần 1 cách tương tác khác.

Cohort Month	Cohort Size	Month 00	Month 01	Month 02	Month 03	Month 04	Month 05	Month 06
Dec 2016	3,949	4.280%	43.277%	70.777%	90.504%	97.316%	99.848%	100%
Jan 2017	6,152	11.297%	50.179%	83.485%	95.936%	99.610%	99.984%	
Feb 2017	5,468	11.595%	64.557%	90.179%	99.305%	99.945%		
Mar 2017	7,257	16.660%	73.033%	97.795%	99.917%			
Apr 2017	7,369	22.405%	91.193%	99.756%				
May 2017	7,586	37.279%	98.800%					
Jun 2017	7,166	64.834%						

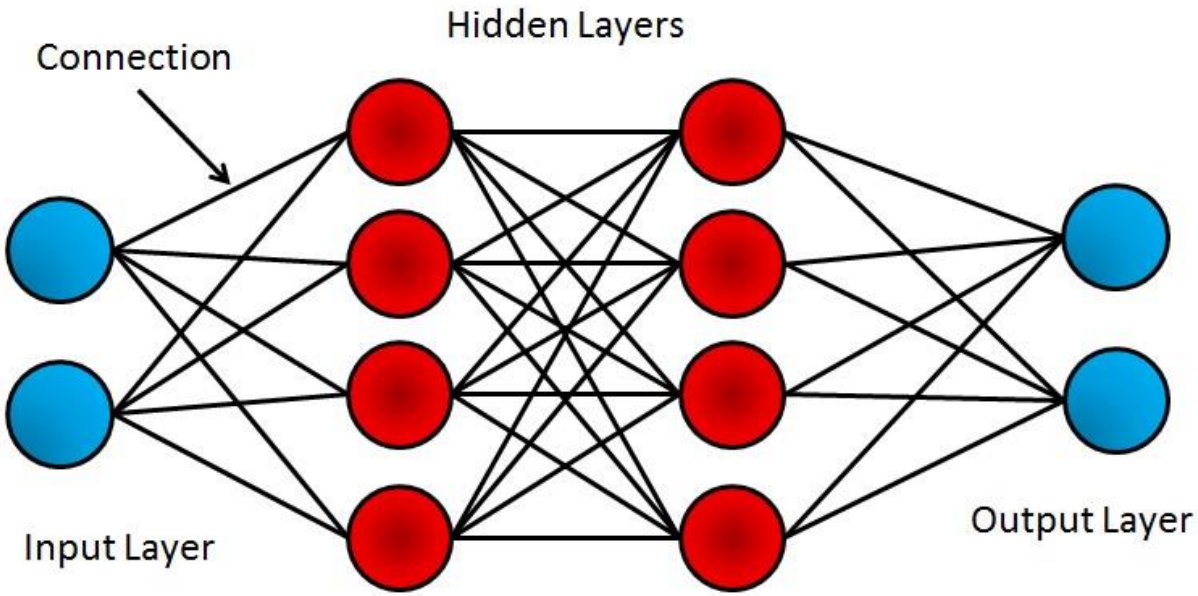
5.3. Phân tích hồi quy (Regression analysis)

- Hồi quy sử dụng dữ liệu lịch sử để hiểu tác động đến giá trị của biến phụ thuộc khi một (*hồi quy tuyến tính*) hoặc nhiều biến độc lập (*hồi quy bội*) thay đổi hoặc giữ nguyên. Bằng cách hiểu mối quan hệ của từng biến (gồm cả biến độc lập và biến phụ thuộc) và cách chúng phát triển trong quá khứ, bạn có thể dự đoán các kết quả có thể xảy ra và đưa ra quyết định tốt hơn trong tương lai.
- Giả sử cần tìm và đánh giá sự ảnh hưởng đến doanh số bán hàng của các biến như:
  - Chất lượng sản phẩm
  - Thiết kế của cửa hàng
  - Vị trí của cửa hàng
  - Dịch vụ khách hàng
  - Chiến dịch tiếp thị.
  - Kênh bán hàng.
  - Hình thức thanh toán
  - …



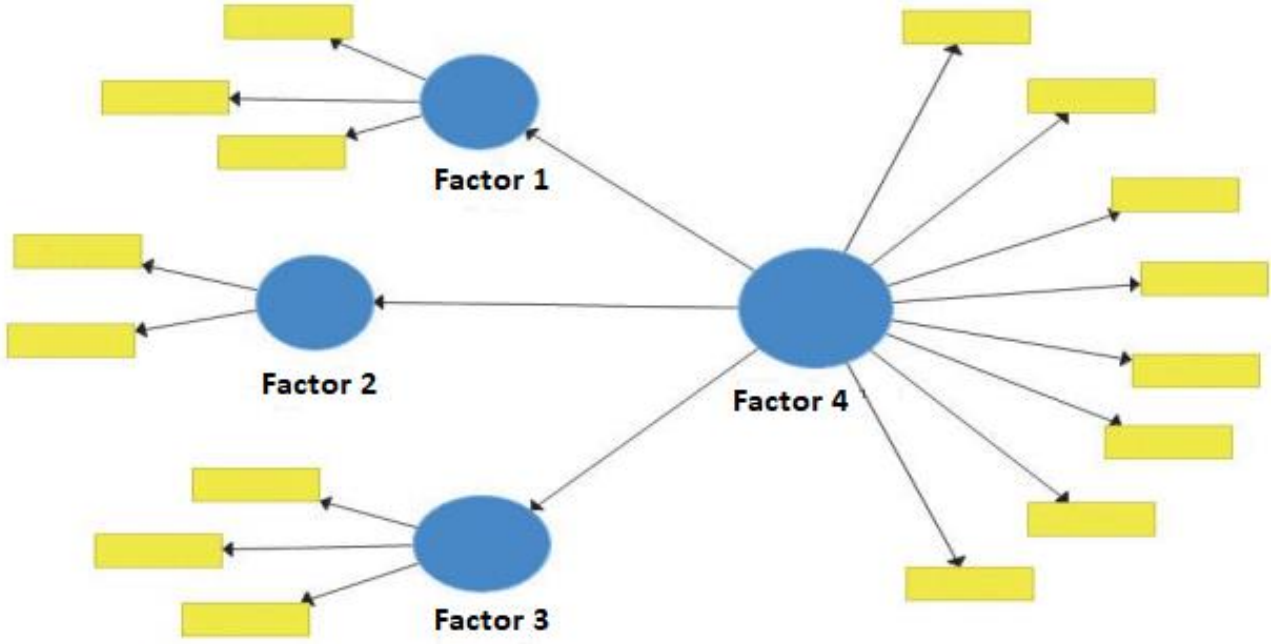
5.4. Mạng nơron (Neural Networks)

- *Neural Network* là nền tảng cho các thuật toán thông minh của *Machine Learning*. Nó là một dạng phân tích có sự can thiệp tối thiểu, để hiểu cách bộ não con người tạo ra những hiểu biết sâu sắc và dự đoán các giá trị. *Neural Network* học hỏi từ mọi dữ liệu, nghĩa là chúng phát triển và tiến bộ theo thời gian.
- Một lĩnh vực ứng dụng điển hình của *Neural Network* là phân tích dữ liệu dự đoán dựa trên dữ liệu lịch sử và hiện tại do người dùng cung cấp.



5.5. Phân tích nhân tố (Factor analysis)

- Phân tích nhân tố còn được gọi là “giảm chiều dữ liệu” (dimension reduction) hay "giảm thứ nguyên“, là một loại phân tích dữ liệu được sử dụng để mô tả sự thay đổi giữa các biến quan sát, tương quan về số lượng các biến không được quan sát có khả năng thấp hơn được gọi là nhân tố với mục đích là phát hiện ra các biến tiềm ẩn độc lập.
- Ví dụ có thể thực hiện đánh giá ban đầu dựa trên các biến số khác nhau như màu sắc, hình dạng, chất liệu, sự thoải mái, cửa hàng, tần suất sử dụng. Trong trường hợp này, danh sách các biến số có thể rất dài, tùy thuộc vào những gì bạn muốn theo dõi. Do đó, phân tích nhân tố đưa ra bức tranh tổng quát bằng cách tóm tắt tất cả các biến này thành các nhóm đồng nhất, ví dụ, bằng cách nhóm các biến màu sắc, vật liệu, chất lượng và xu hướng thành một biến tiềm ẩn của thiết kế.



5.6. Khai thác dữ liệu (Data mining)

- Khai thác dữ liệu là phương pháp phân tích nhằm xác định các yếu tố phụ thuộc, quan hệ, mẫu dữ liệu và xu hướng nhằm xác định xu hướng, mẫu và dữ liệu hữu ích.
- Cùng với phân tích dự đoán, khai thác dữ liệu là một nhánh của khoa học thống kê sử dụng các thuật toán phức tạp, không chỉ bao hàm bước phân tích thô, mà còn liên quan tới cơ sở dữ liệu, quản lý dữ liệu, tiền xử lý dữ liệu, suy luận thống kê, …

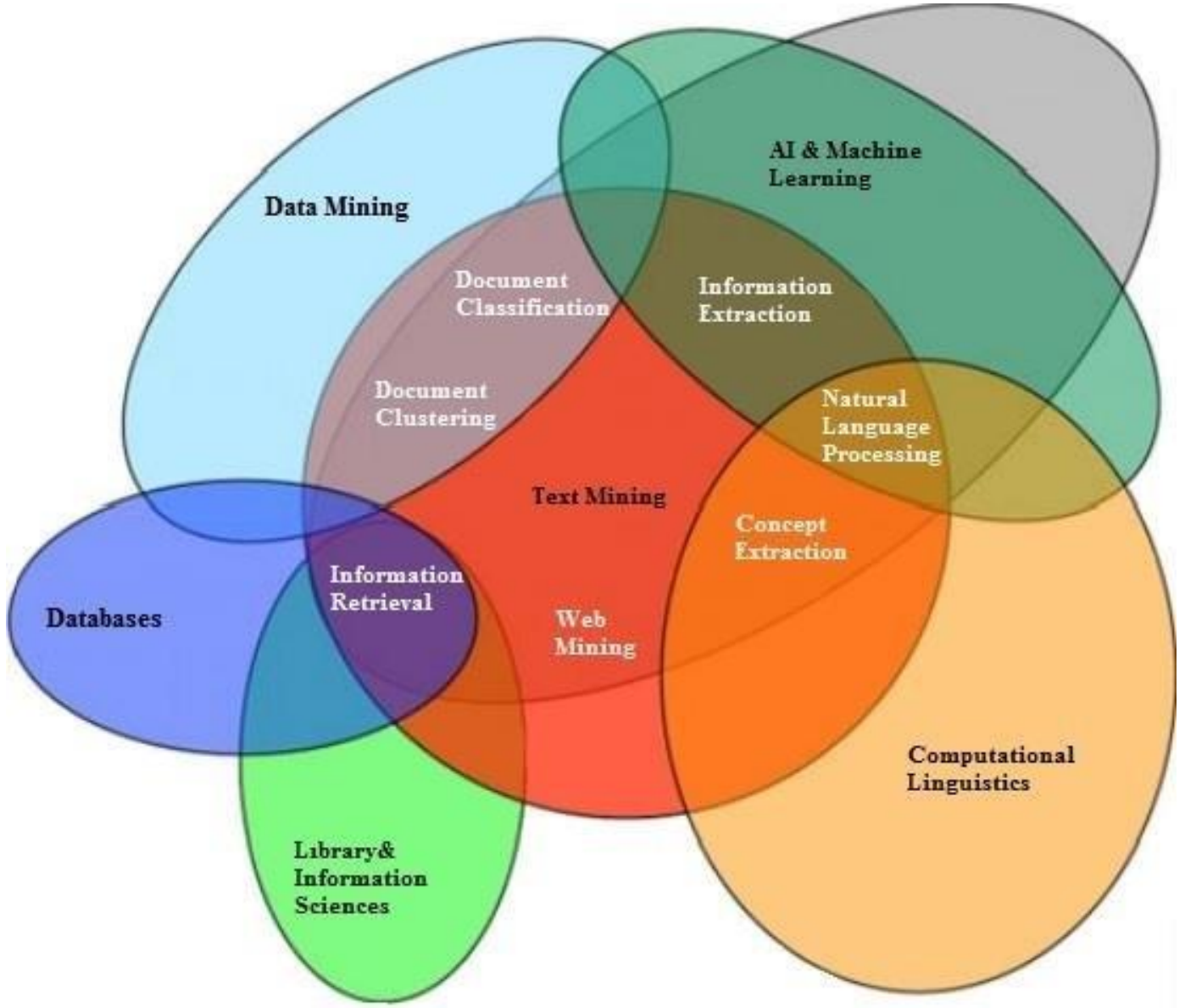
5.7. Phân tích văn bản (Text analysis)

- Phân tích văn bản, còn được gọi là khai thác văn bản (*Text minning*), hoạt động bằng cách lấy các bộ dữ liệu văn bản lớn, thực hiện sắp xếp, trích xuất dữ liệu thực sự liên quan đến vấn đề cần quan tâm nhằm phát triển những thông tin hữu ích phục vụ việc ra quyết định.

Ví dụ: việc phân tích dữ liệu từ nhiều nguồn văn bản khác nhau như bài viết đánh giá sản phẩm trên mạng xã hội hoặc phản hồi khảo sát giúp hiểu sâu sắc hơn về đối tượng mục tiêu, từ đó cho phép xây dựng các chiến dịch, dịch vụ đáp ứng nhu cầu của khách hàng tiềm năng.

5.7. Phân tích văn bản (Text analysis)

- Nhờ sự kết hợp của học máy và các thuật toán thông minh, phân tích văn bản hiện nay còn cho phép thực hiện các quy trình phân tích nâng cao như phân tích cảm xúc. Phân tích cảm xúc thường được sử dụng để theo dõi danh tiếng của thương hiệu và sản phẩm cũng như để hiểu mức độ thành công của trải nghiệm khách hàng.

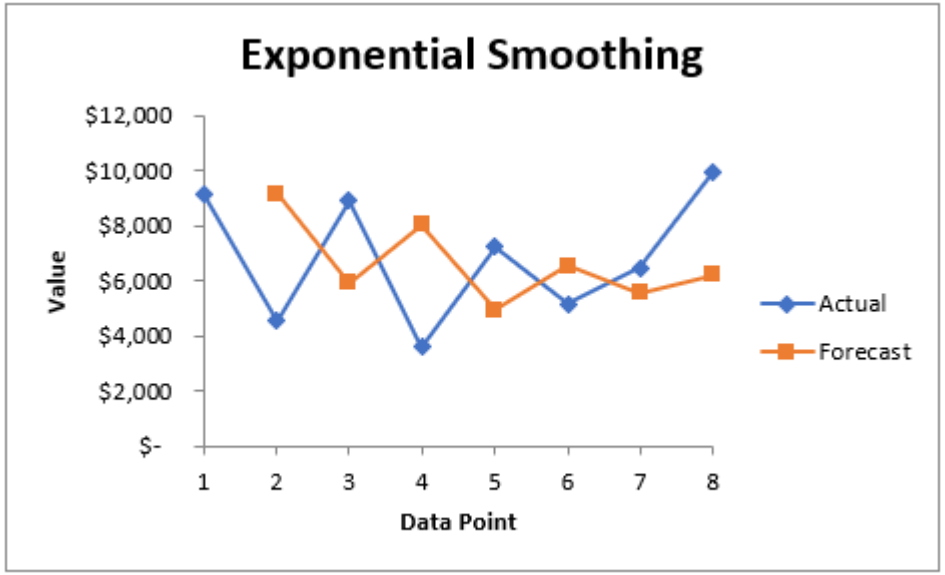


5.8. Phân tích chuỗi thời gian (Time series analysis)

- Phân tích chuỗi thời gian được sử dụng để phân tích một tập hợp dữ liệu thu thập trong một khoảng thời gian xác định.

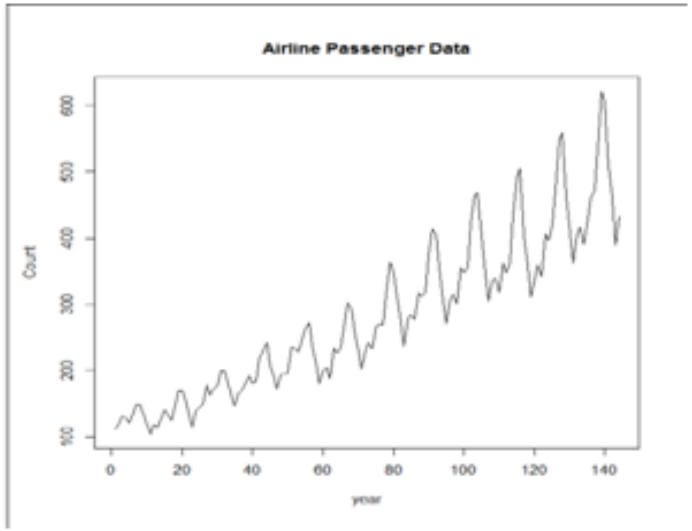
Analysis of Time Series Data

Year	Quarter	Revenue	Smoothed Levels	Standard Errors
2020	1	$ 9,150	#N/A	#N/A
	2	$ 4,560	$ 9,150	#N/A
	3	$ 8,920	5937	#N/A
	4	$ 3,615	8025.1	#N/A
2022	1	$ 7,245	4938.03	4058.545347
	2	$ 5,150	6552.909	3350.09361
	3	$ 6,480	5570.8727	2985.478535
	4	$ 9,950	6207.26181	1644.868454



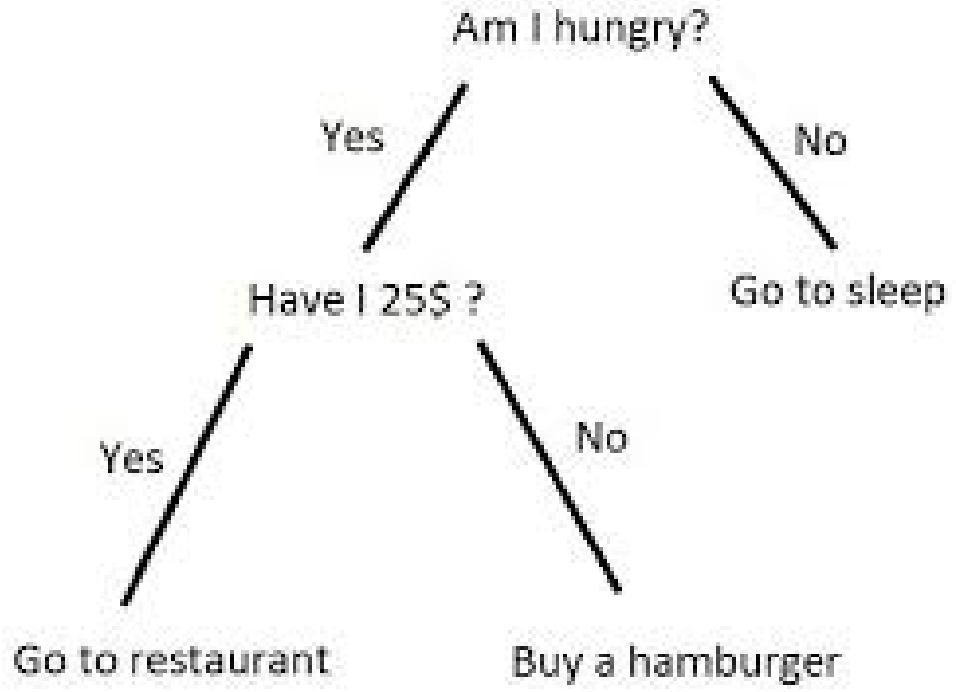
5.8. Phân tích chuỗi thời gian (Time series analysis)

- Ngoài ra, phương pháp này còn cho biết:
  - Giá trị các biến có thay đổi trong suốt thời gian nghiên cứu hay không?
  - Các biến khác nhau phụ thuộc nhau như thế nào?
  - Kết quả cuối cùng ra sao?
- Trong kinh doanh, phương pháp này được sử dụng để hiểu nguyên nhân của các xu hướng và mô hình khác nhau, từ đó rút ra những hiểu biết có giá trị.
- Phương pháp này cũng có thể kết hợp với dự báo chuỗi thời gian nhằm dự báo sự kiện có thể xảy ra trong tương lai.



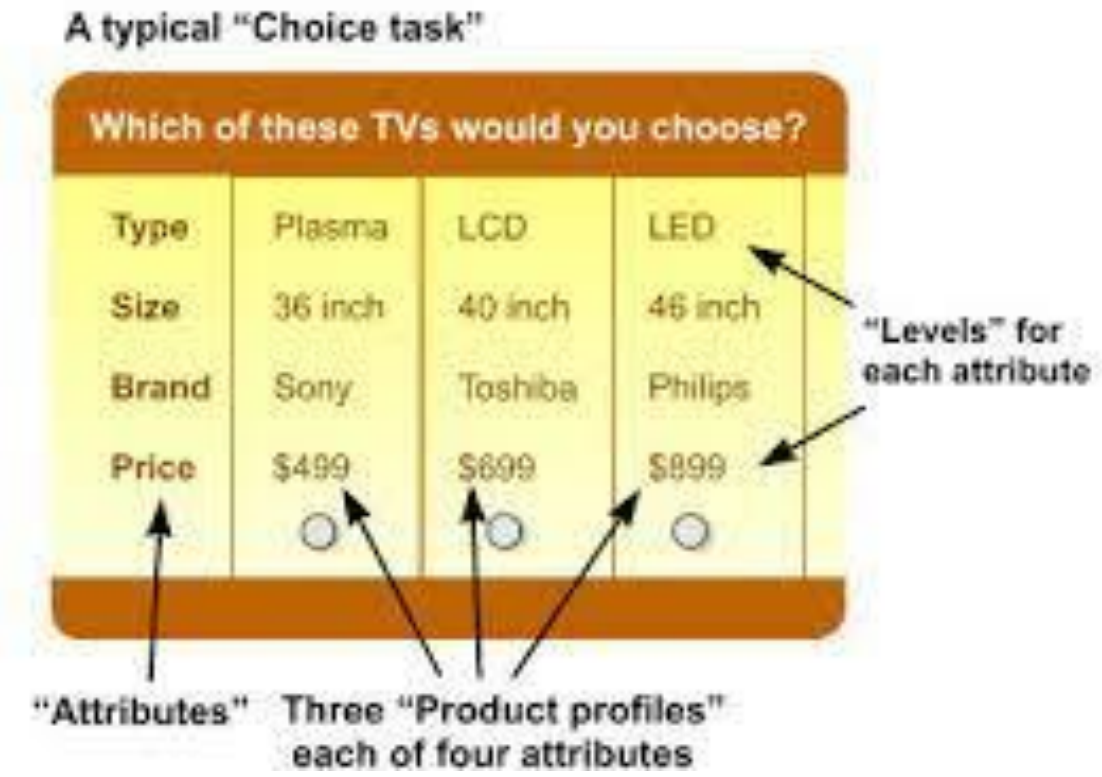
5.9. Cây quyết định (Decision Trees)

- Phân tích dựa trên cây quyết định hoạt động như một công cụ hỗ trợ để đưa ra các quyết định chiến lược và thông minh. Bằng cách hiển thị trực quan các kết quả, hậu quả và chi phí tiềm năng trong mô hình dạng cây, các nhà nghiên cứu và người dùng doanh nghiệp có thể dễ dàng đánh giá tất cả các yếu tố liên quan và chọn cách hành động tốt nhất.
- Cây quyết định thường được dùng để phân tích dữ liệu định lượng, nó cho phép cải thiện quy trình ra quyết định bằng cách giúp bạn xác định các cơ hội cải tiến, giảm chi phí, nâng cao hiệu quả hoạt động và sản xuất.

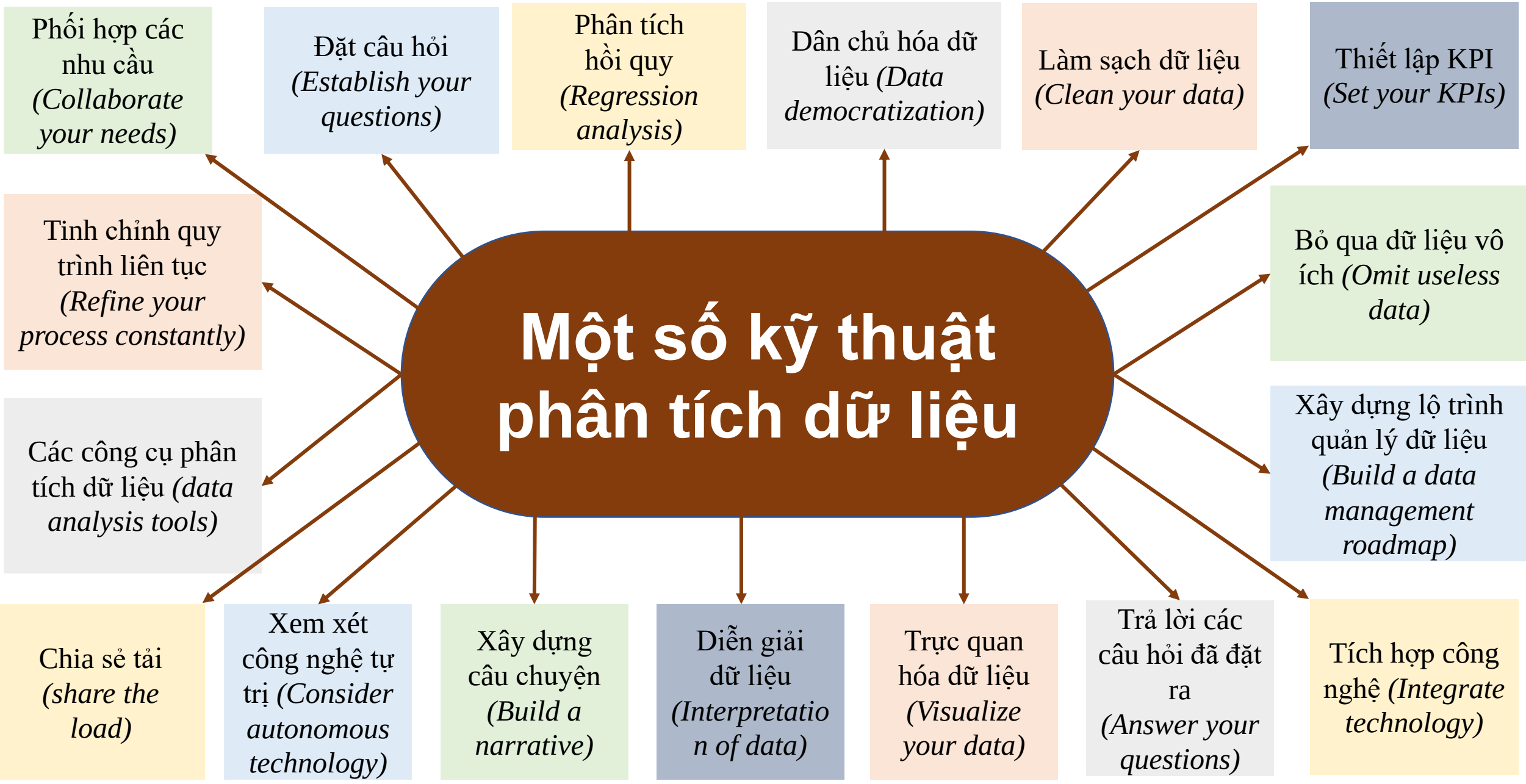


5.10. Phân tích thuộc tính (Conjoint analysis)

- Phân tích thuộc tính thường được sử dụng trong các cuộc khảo sát để hiểu cách người dùng đánh giá các thuộc tính khác nhau của một sản phẩm hoặc dịch vụ.
- Ví dụ, khi nói đến việc mua hàng, một số khách hàng có thể tập trung vào giá, những người khác tập trung vào tính năng, hay tính bền vững của sản phẩm. Có thể tìm thấy các thuộc tính này bằng phân tích kết hợp. Như vậy, các công ty có thể xác định chiến lược giá cả, tùy chọn gói sản phẩm, dịch vụ,….



6. MỘT SỐ KỸ THUẬT PHÂN TÍCH DỮ LIỆU



6.1. Phối hợp các nhu cầu của bạn *(Collaborate your needs)*

Trước khi bắt đầu phân tích dữ liệu hoặc đi sâu vào bất kỳ kỹ thuật phân tích nào, cần:

- Xác định được mục tiêu chiến lược của tổ chức.
- Có được sự hiểu biết cơ bản về các loại thông tin sẽ được dung để phân tích.

⇨ Giúp việc phân tích dữ liệu mang lại lợi ích hoặc cung cấp cho bạn tầm nhìn cần thiết để phát triển tổ chức của mình.

6.2. Đặt câu hỏi của bạn *(Establish your questions)*

- Sau khi đã xác định được các mục tiêu cốt lõi, bạn cần đặt ra được những câu hỏi phân tích dữ liệu phù hợp với dữ liệu đang có để giúp đạt được mục tiêu. Đây là một trong những kỹ thuật phân tích dữ liệu quan trọng nhất vì nó sẽ định hình nền tảng thành công của bạn.

6.3. Dân chủ hóa dữ liệu *(Data democratization)*

- Dân chủ hóa dữ liệu là một quá trình nhằm mục đích kết nối dữ liệu từ nhiều nguồn khác nhau một cách hiệu quả và nhanh chóng để bất kỳ ai trong tổ chức đều có thể truy cập dữ liệu vào bất kỳ thời điểm nào.
- Có thể trích xuất dữ liệu ở dạng văn bản, hình ảnh, video, số hoặc bất kỳ định dạng nào khác. Và sau đó thực hiện phân tích cơ sở dữ liệu chéo để đạt được thông tin chi tiết nâng cao hơn để chia sẻ tương tác với phần còn lại của công ty.

6.3. Dân chủ hóa dữ liệu *(Data democratization)*

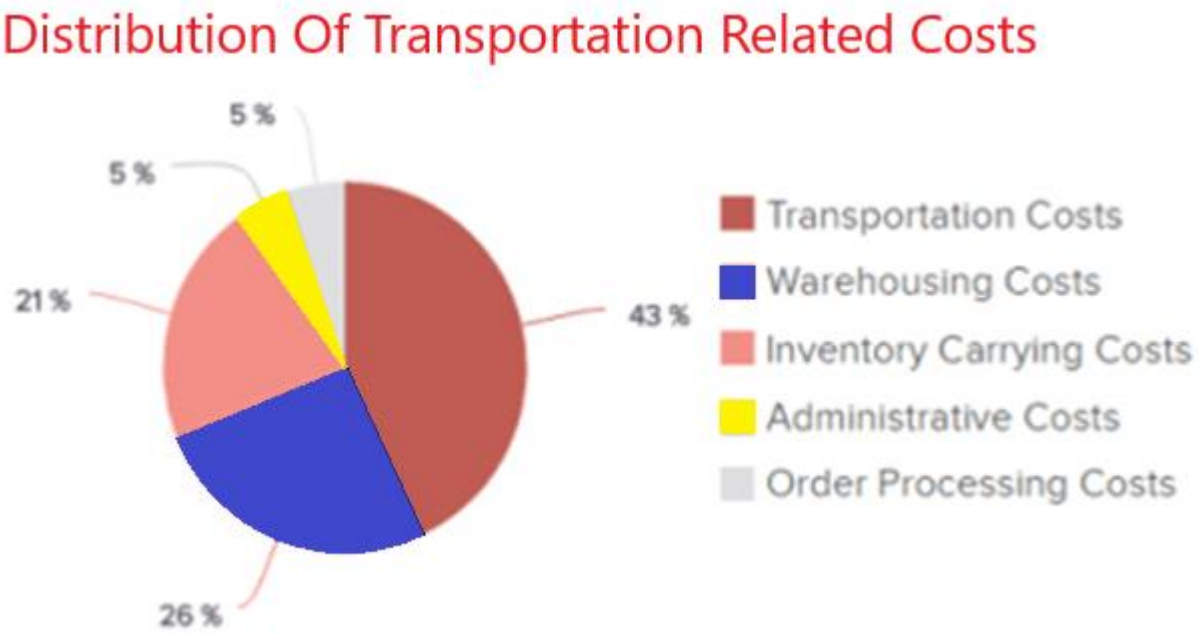
- Khi bạn đã quyết định các nguồn dữ liệu có giá trị nhất của mình, bạn cần đưa tất cả thông tin này vào một định dạng có cấu trúc để bắt đầu thu thập thông tin chi tiết của mình.
- Với mục đích này, datapine cung cấp tính năng kết nối tất cả dữ liệu để tích hợp tất cả các nguồn dữ liệu bên trong và bên ngoài của bạn và quản lý chúng theo ý muốn của bạn. Ngoài ra, giải pháp end-to-end của datapine tự động cập nhật dữ liệu của bạn, cho phép bạn tiết kiệm thời gian và tập trung vào việc thực hiện phân tích phù hợp để phát triển doanh nghiệp của mình.

6.4. Làm sạch dữ liệu của bạn *(Clean your data)*

- Mục đích chính của việc làm sạch dữ liệu là để ngăn bạn đưa ra những kết luận sai lầm.
- Với một lượng lớn dữ liệu thu thập được, có thể bạn phải đối mặt với dữ liệu không chính xác có thể gây hiểu lầm cho phân tích của bạn.
- Điều thông minh nhất bạn có thể làm để tránh phải đối mặt với vấn đề này trong tương lai là làm sạch dữ liệu. Quá trình này là cơ bản vì nó sẽ đảm bảo rằng những thông tin chi tiết bạn trích xuất từ nó là chính xác.
- Có nhiều thứ bạn cần tìm trong khi làm sạch dữ liệu của mình:
  - Loại bỏ bất kỳ quan sát trùng lặp (do sử dụng từ nhiều nguồn dữ liệu bên trong và bên ngoài).
  - Có thể thêm bất kỳ mã nào bị thiếu
  - Sửa các trường trống
  - Loại bỏ dữ liệu được định dạng không chính xác (như ký tự không hợp lệ, lỗi cú pháp hoặc chính tả).

6.5. Đặt KPI của bạn *(Set your KPIs)*

- Cần đặt một loạt các chỉ số hiệu suất chính (KPIs - *Key Performance Indicators*) để giúp bạn theo dõi, đo lường và định hình tiến bộ trên một số lĩnh vực chủ yếu.
- KPI rất quan trọng đối với cả phương pháp phân tích trong nghiên cứu định tính và định lượng. Đây là một trong những phương pháp phân tích dữ liệu chính mà bạn không nên bỏ qua.



6.6. Bỏ qua dữ liệu vô ích *(Omit useless data)*

- Bất kỳ số liệu thống kê, dữ kiện, số liệu hoặc chỉ số nào không phù hợp với mục tiêu hoặc không phù hợp với chiến lược quản lý KPI sẽ bị loại bỏ.

6.7. Xây dựng lộ trình quản lý dữ liệu *(Build a data management roadmap)*

- (Bước tùy chọn) Việc tạo ra một lộ trình quản trị dữ liệu sẽ giúp các phương pháp và kỹ thuật phân tích dữ liệu của bạn trở nên thành công và bền vững hơn.
- Đầu tư nhiều thời gian vào việc phát triển một lộ trình sẽ giúp bạn lưu trữ, quản lý và xử lý dữ liệu của mình sẽ làm cho các kỹ thuật phân tích trở nên linh hoạt hơn.

6.8. Tích hợp công nghệ *(Integrate technology)*

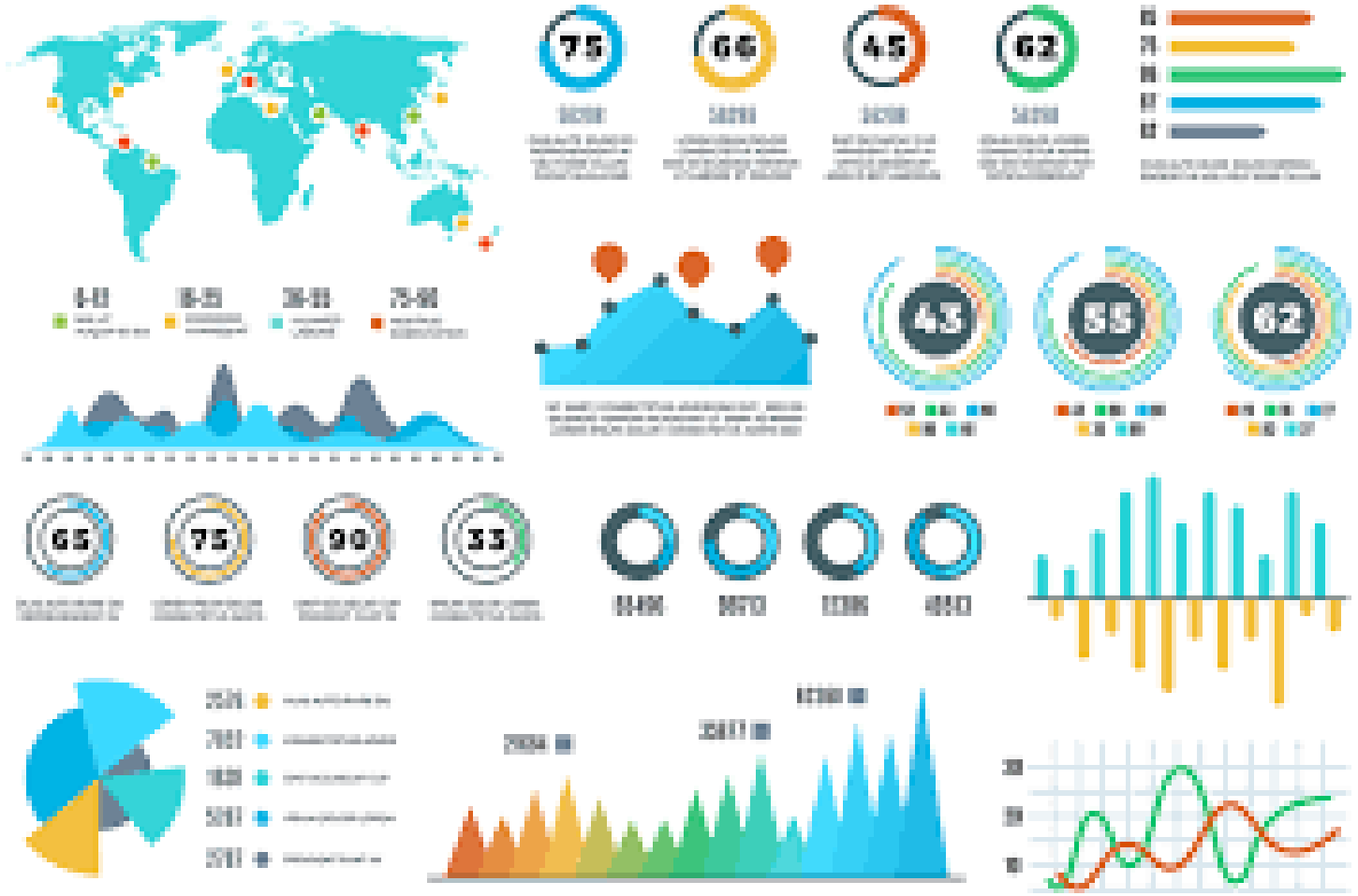
- Bằng cách tích hợp công nghệ phù hợp cho phân tích dữ liệu theo phương pháp thống kê và phương pháp phân tích dữ liệu cốt lõi, bạn sẽ tránh phân mảnh thông tin chi tiết của mình, tiết kiệm thời gian và công sức đồng thời cho phép bạn tận hưởng giá trị tối đa từ những thông tin chi tiết có giá trị nhất của doanh nghiệp.

6.9. Trả lời câu hỏi của bạn *(Answer your questions)*

- Với các bước đã qua, từ đây bạn có thể nhanh chóng trả lời các câu hỏi của mình.

6.10. Trực quan hóa dữ liệu của bạn *(Visualize your data)*

- Có thể cho rằng, cách tốt nhất để làm cho các khái niệm dữ liệu của bạn có thể truy cập được trong toàn tổ chức là thông qua trực quan hóa dữ liệu (*data visualization*).



6.11. Diễn giải dữ liệu *(Interpretation of data)*

- Diễn giải dữ liệu là một phần cơ bản của quá trình phân tích dữ liệu. Nó cung cấp ý nghĩa cho thông tin phân tích và nhằm mục đích đưa ra kết luận ngắn gọn từ kết quả phân tích.
- Vì hầu hết thời gian các công ty xử lý dữ liệu từ nhiều nguồn khác nhau, giai đoạn giải thích cần phải được thực hiện cẩn thận và đúng cách để tránh hiểu sai.
- Ba phương pháp phổ biến cần tránh khi xem dữ liệu:

  i. Nhầm lẫn giữa tương quan so với nhân quả

  - Mặc dù hai khía cạnh này có thể tồn tại đồng thời, nhưng sẽ không đúng nếu cho rằng vì hai sự việc xảy ra cùng nhau nên cái này kích động cái kia.
  - Một lời khuyên để tránh rơi vào sai lầm này là đừng bao giờ chỉ tin vào trực giác, hãy tin vào dữ liệu. Nếu không có bằng chứng khách quan về nhân quả, thì hãy luôn gắn bó với mối tương quan.

6.11. Diễn giải dữ liệu *(Interpretation of data)*

- Ba phương pháp phổ biến cần tránh khi xem dữ liệu:

ii. *Sự thiên lệch xác nhận*

- Hiện tượng này mô tả xu hướng chỉ chọn và giải thích dữ liệu cần thiết để chứng minh một giả thuyết, thường bỏ qua các yếu tố có thể bác bỏ nó. Ngay cả khi nó không được thực hiện có chủ đích, sai lệch xác nhận có thể đại diện cho một vấn đề thực sự đối với một doanh nghiệp, vì việc loại trừ thông tin liên quan có thể dẫn đến kết luận sai và do đó, dẫn đến các quyết định kinh doanh tồi.
- Để tránh điều đó, hãy luôn cố gắng bác bỏ giả thuyết của bạn thay vì chứng minh nó, chia sẻ phân tích của bạn với các thành viên khác trong nhóm và tránh đưa ra bất kỳ kết luận nào trước khi toàn bộ quá trình phân tích dữ liệu được hoàn thành.

iii. *Ý nghĩa thống kê*

- Ý nghĩa thống kê giúp các nhà phân tích hiểu được liệu một kết quả có thực sự chính xác hay nó xảy ra do lỗi lấy mẫu hay do cơ hội thuần túy.
- Trong mọi trường hợp, bỏ qua tầm quan trọng của một kết quả khi nó có thể ảnh hưởng đến việc ra quyết định có thể là một sai lầm lớn.

6.12. Xây dựng câu chuyện *(Build a narrative)*

- Sau khi đã thảo luận và khám phá các ứng dụng kỹ thuật của phân tích theo hướng dữ liệu, chúng ta sẽ xem xét cách bạn có thể kết hợp tất cả các yếu tố này lại với nhau theo cách có lợi cho doanh nghiệp của bạn - bắt đầu bằng một thứ gọi là kể chuyện dữ liệu.
- Bộ não con người phản ứng cực kỳ tốt với những câu chuyện hoặc câu chuyện mạnh mẽ. Sau khi đã làm sạch, định hình và hình dung dữ liệu vô giá nhất của mình bằng các công cụ bảng điều khiển BI khác nhau, bạn nên cố gắng kể một câu chuyện - một câu chuyện có mở bài, thân bài và kết luận rõ ràng.
- Bằng cách đó, bạn sẽ làm cho các nỗ lực phân tích của mình trở nên dễ tiếp cận, dễ hiểu và phổ biến hơn, trao quyền cho nhiều người hơn trong tổ chức của bạn sử dụng khám phá của bạn để làm lợi thế cho các hành động của họ.

6.13. Xem xét công nghệ tự trị *(Consider autonomous technology)*

- Các công nghệ tự chủ như trí tuệ nhân tạo (AI) và học máy (ML), đóng một vai trò quan trọng trong việc thúc đẩy sự hiểu biết về cách phân tích dữ liệu hiệu quả hơn.
- Hiện tại, một số công nghệ đang cách mạng hóa ngành phân tích dữ liệu như:
  - Mạng thần kinh (*Neural network*)
  - Cảnh báo thông minh (*Smart alert*)
  - Phân tích cảm xúc (*emotional analysis*)
  - …

6.14. Chia sẻ tải *(share the load)*

- Nếu làm việc với các công cụ và trang tổng quan phù hợp, bạn có thể cho phép hầu hết mọi người trong tổ chức kết nối và sử dụng dữ liệu có liên quan để có lợi cho họ.
- Bảng điều khiển dữ liệu hiện đại hợp nhất dữ liệu từ nhiều nguồn khác nhau, cung cấp quyền truy cập vào vô số thông tin chi tiết tại một vị trí tập trung, bất kể bạn cần theo dõi các chỉ số tuyển dụng hay tạo báo cáo cần được gửi qua nhiều phòng ban. Hơn nữa, những công cụ tiên tiến này cung cấp quyền truy cập vào trang tổng quan từ vô số thiết bị, có nghĩa là mọi người trong doanh nghiệp có thể kết nối từ xa với những thông tin chi tiết thực tế - và chia sẻ tải.
- Một khi tất cả mọi người đều có thể làm việc với tư duy dựa trên dữ liệu, bạn sẽ thúc đẩy sự thành công của doanh nghiệp mình theo những cách mà bạn không bao giờ nghĩ có thể. Và khi biết cách phân tích dữ liệu, thì loại phương pháp hợp tác này là rất cần thiết.

6.15. Các công cụ phân tích dữ liệu *(data analysis tools)*

Bốn loại công cụ phân tích dữ liệu cơ bản

i. Business Intelligence (BI)

- Các công cụ BI cho phép bạn xử lý lượng dữ liệu đáng kể từ một số nguồn ở bất kỳ định dạng nào.
- Nhờ vậy bạn có thể:
  - Phân tích và theo dõi dữ liệu của mình để trích xuất thông tin chi tiết có liên quan.
  - Tạo các báo cáo và trang tổng quan tương tác để trực quan hóa KPI và sử dụng chúng cho lợi ích của mình.
- ***datapine***
  - Là một phần mềm BI trực tuyến tập trung vào việc cung cấp các tính năng phân tích trực tuyến mạnh mẽ có thể truy cập được cho người mới bắt đầu và người dùng nâng cao.
  - Datapine cung cấp một giải pháp đầy đủ dịch vụ bao gồm:
    - Phân tích dữ liệu tiên tiến
    - Trực quan hóa KPI
    - Bảng điều khiển trực tiếp và báo cáo
    - Dự đoán xu hướng và giảm thiểu rủi ro dựa trên các công nghệ trí tuệ nhân tạo

6.15. Các công cụ phân tích dữ liệu *(data analysis tools)*

Bốn loại công cụ phân tích dữ liệu cơ bản

ii. Phân tích thống kê

- Các công cụ này thường được thiết kế cho các nhà khoa học dữ liệu, nhà thống kê, nhà nghiên cứu thị trường và nhà toán học, vì chúng cho phép họ thực hiện các phân tích thống kê phức tạp với các phương pháp như phân tích hồi quy, phân tích dự đoán và mô hình thống kê.
- ***R-Studio***
  - Là một công cụ tốt để thực hiện phân tích thống kê vì nó cung cấp tính năng kiểm tra giả thuyết và mô hình hóa dữ liệu mạnh mẽ có thể bao gồm cả phân tích dữ liệu học thuật và tổng hợp.
  - Công cụ này có khả năng làm sạch dữ liệu, giảm dữ liệu và thực hiện phân tích nâng cao với một số phương pháp thống kê.

6.15. Các công cụ phân tích dữ liệu *(data analysis tools)*

Bốn loại công cụ phân tích dữ liệu cơ bản

ii. Phân tích thống kê

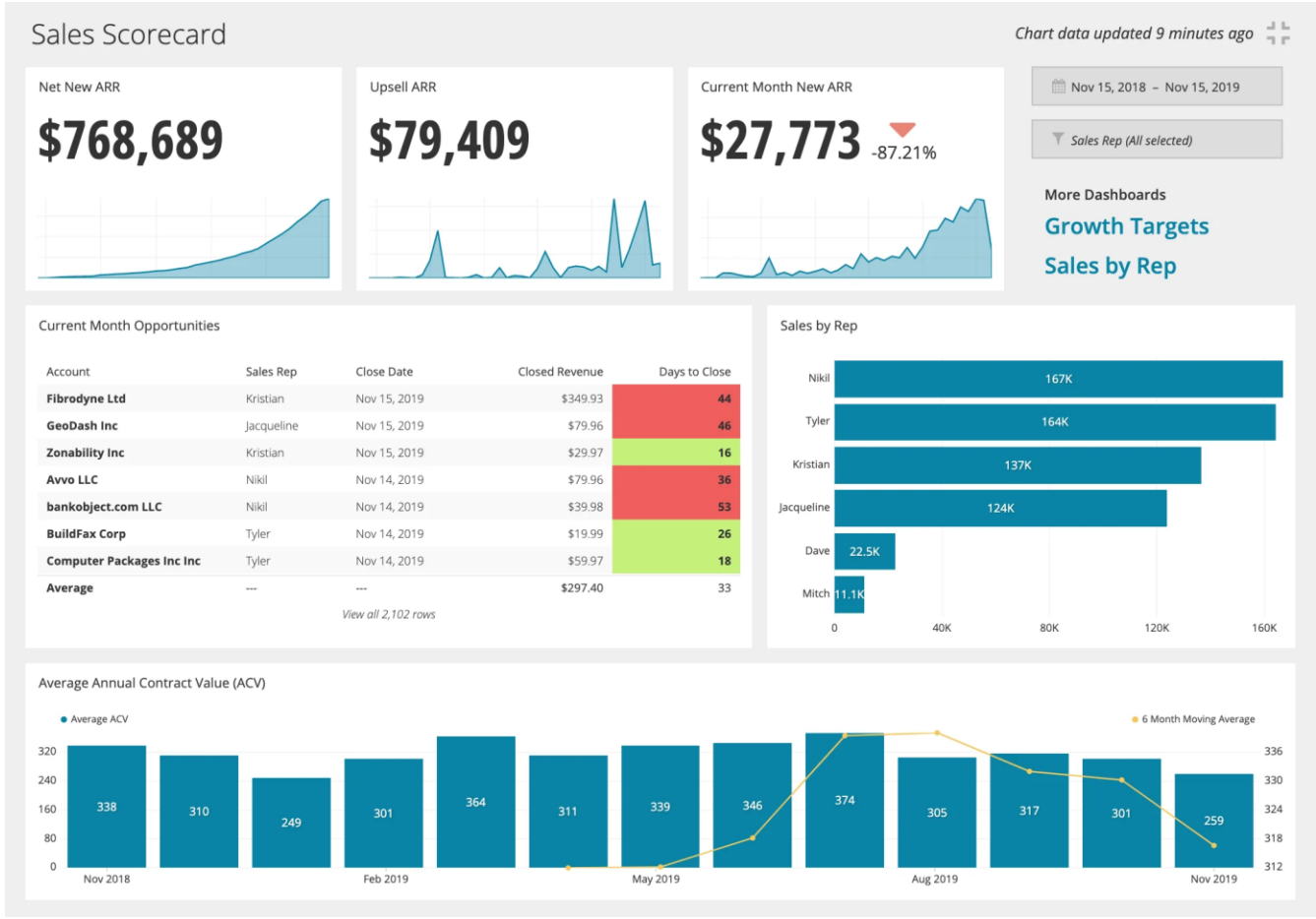
- ***SPSS (của IBM)***
  - Phần mềm cung cấp phân tích thống kê nâng cao cho người dùng ở mọi cấp độ kỹ năng. Nhờ có một thư viện rộng lớn gồm các thuật toán học máy, phân tích văn bản và phương pháp kiểm tra giả thuyết, nó có thể giúp công ty của bạn tìm thấy thông tin chi tiết phù hợp để đưa ra các quyết định kinh doanh tốt hơn.
  - SPSS cũng hoạt động như một dịch vụ đám mây cho phép thực hiện phân tích ở bất cứ đâu.

6.15. Các công cụ phân tích dữ liệu *(data analysis tools)*

Bốn loại công cụ phân tích dữ liệu cơ bản

iii. *Bảng điều khiển SQL*

- SQL là một ngôn ngữ lập trình thường được sử dụng để xử lý dữ liệu có cấu trúc trong cơ sở dữ liệu quan hệ.
- Những công cụ này rất phổ biến đối với các nhà khoa học dữ liệu để mô hình hóa và giám sát cơ sở dữ liệu, tối ưu hóa SQL hoàn chỉnh, cùng với các công cụ quản trị và bảng điều khiển hiệu suất trực quan để theo dõi các KPI.



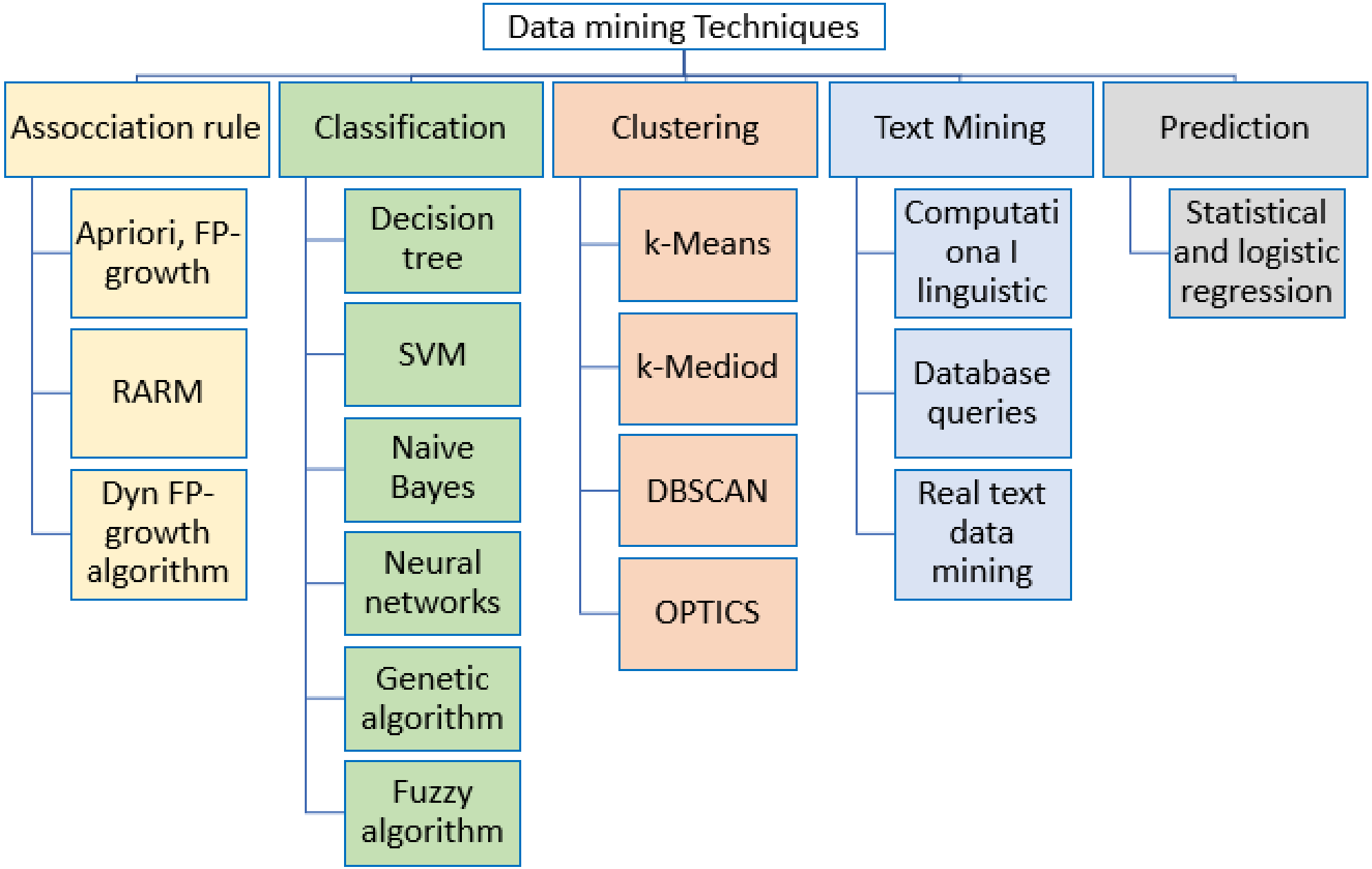
6.15. Các công cụ phân tích dữ liệu *(data analysis tools)*

Bốn loại công cụ phân tích dữ liệu cơ bản

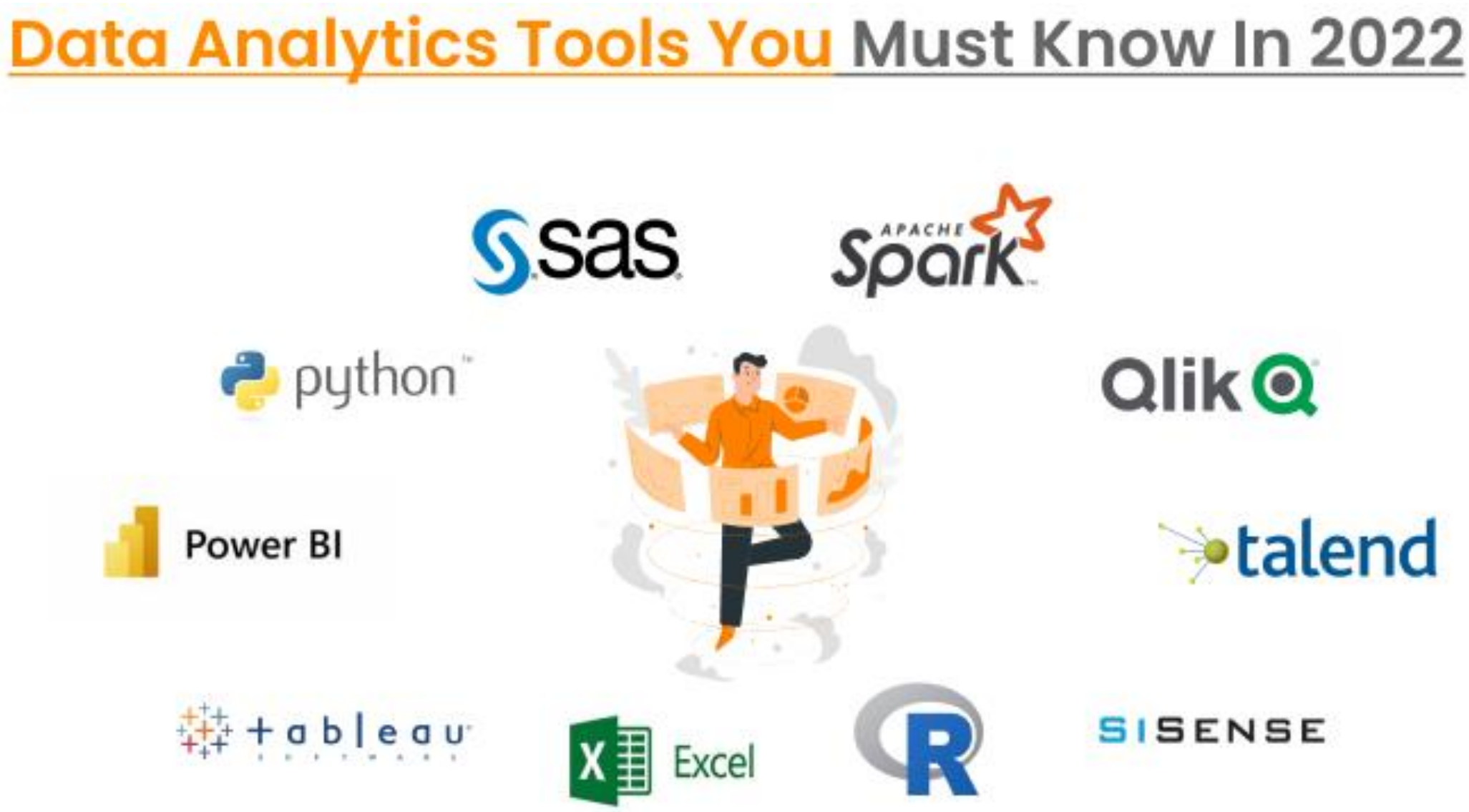
iv. Trực quan hóa dữ liệu

- Các công cụ trực quan hóa dữ liệu được sử dụng để trình bày dữ liệu của bạn thông qua biểu đồ, đồ thị và bản đồ cho phép bạn tìm các mẫu và xu hướng trong dữ liệu.

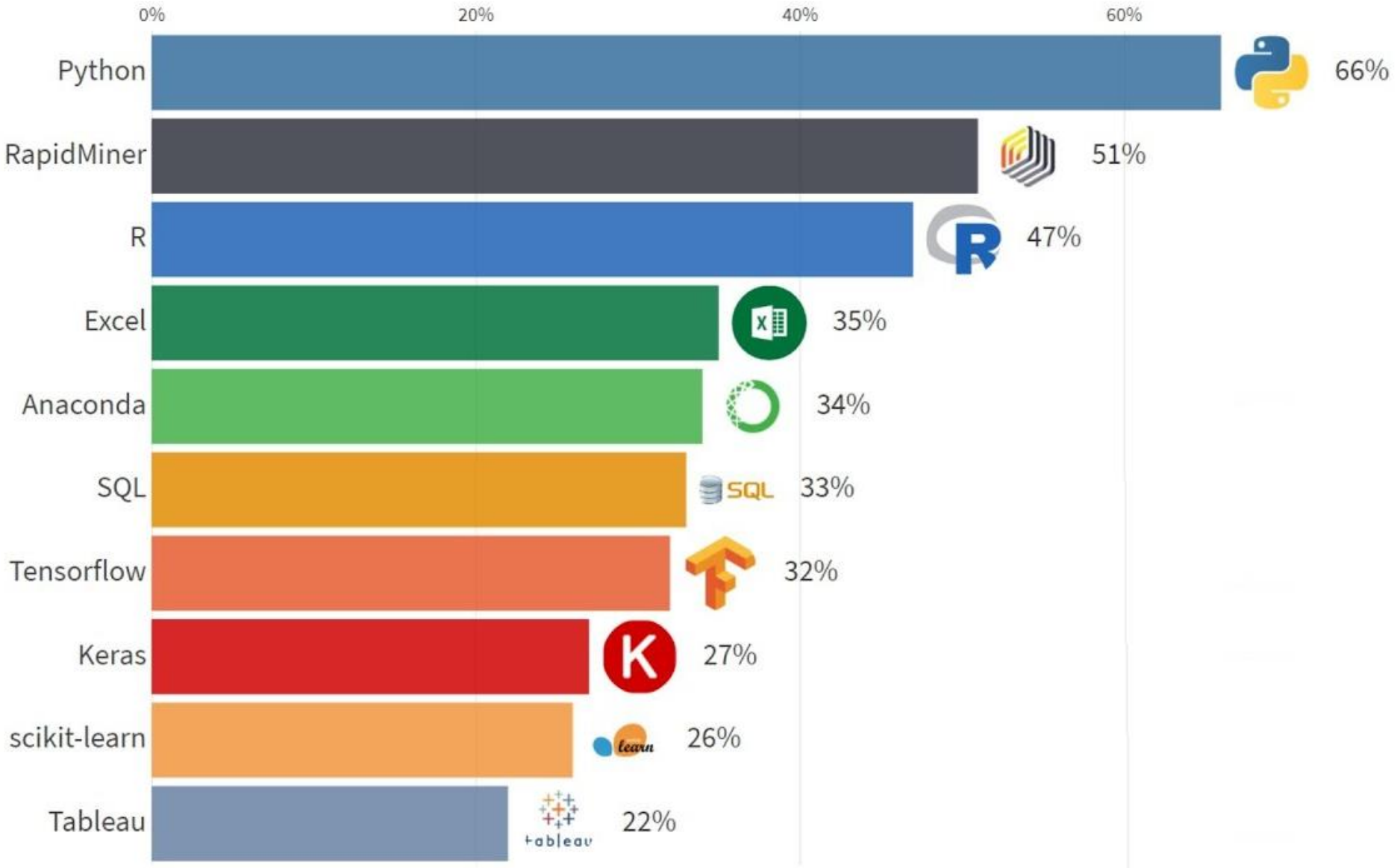
7. Các kỹ thuật khai thác dữ liệu *(Data mining Techniques)*



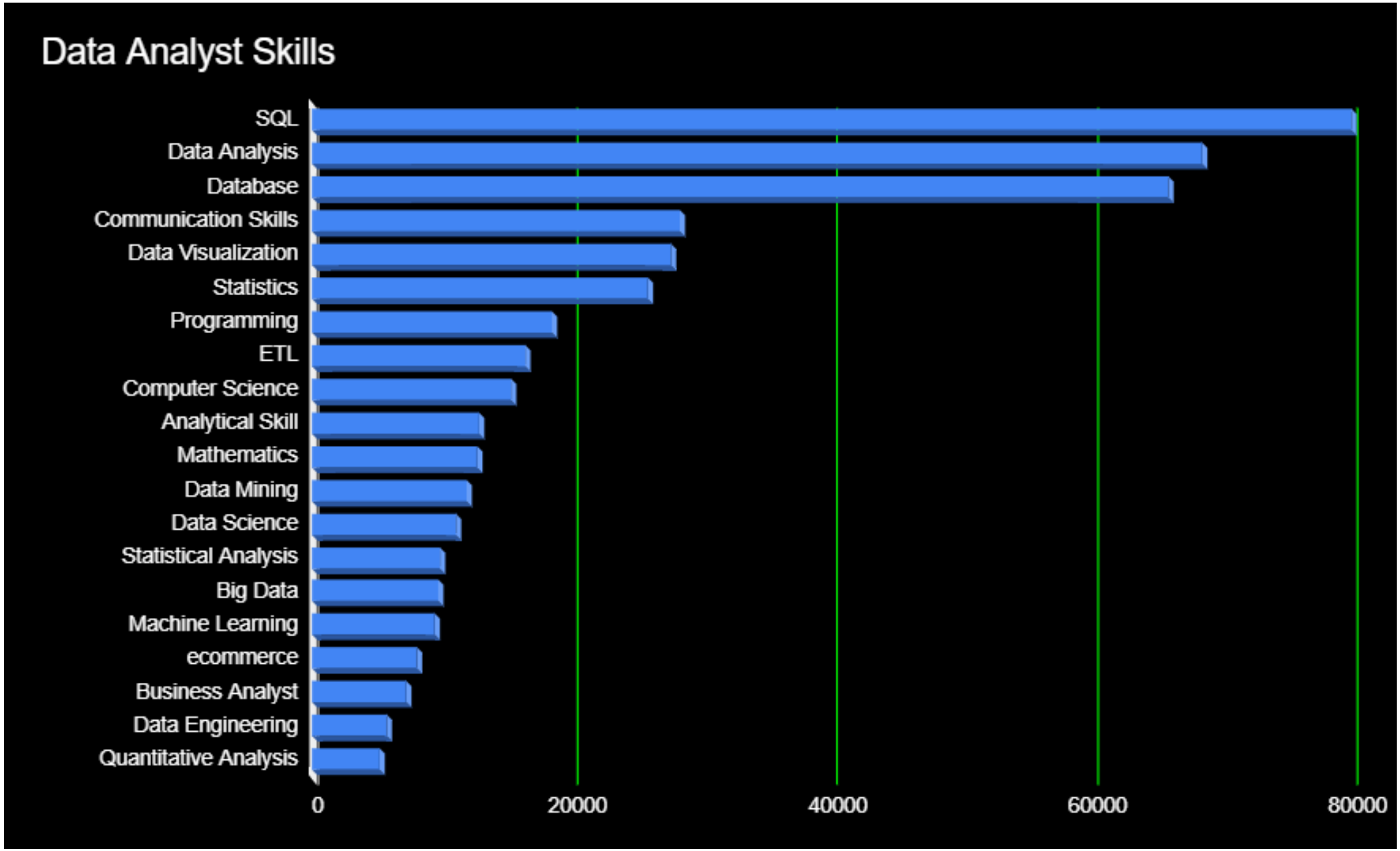
8. Các công cụ phân tích dữ liệu phổ biến



8. Các công cụ phân tích dữ liệu phổ biến



9. Các kỹ năng cần có đối với phân tích viên




Q&A
How
Why
When
Where
What
Who