Bài giảng 1: Giới thiệu về Xử lý Số liệu Thống kê

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

Thông tin của môn học

Giảng viên:

TS. Tô Đức Khánh (tôi)
Bộ môn Xác suất-Thống kê, khoa Toán-Tin học
Trường Đại Học Khoa Học Tự Nhiên, Tp. HCM
e-mail: tdkhanh@hcmus.edu.vn
Phòng F.012, Trường Đại Học Khoa Học Tự Nhiên, 227 Nguyễn Văn Cừ, Quận 5, Tp. Hồ Chí Minh.

Thông tin của môn học

Nội dung của môn học trong 12 tuần bao gồm:

- Quy trình xử lý số liệu thống kê;
- A/B testing;
- Hồi quy tuyến tính và tiên đoán;
- Bài toán phân loại;
- Imputation data;
- Imbalance data;
- Phân tích sống sót và ứng dụng trong khoa học dữ liệu.
- Unsupervised learning;

Thực hành trên phần mềm R/Rstudio, với các thư viện: `tidyverse`, `ggplot2`, `xgboost`, `randomForest`, `rpart`, `FNN`, `glm`, splines, `mgcv`, `survival`, `pROC`.

Thông tin của môn học

Số tín chỉ

	Tín chỉ	Số giờ học
Lý thuyết	2	30
Thực hành R	1	30
Tổng cộng	3	60

Lịch học

- Tất cả các bài giảng lý thuyết, thực hành và bài tập được tổ chức tại cở sở 2 của trường Đại Học Khoa Học Tự Nhiên, Linh Trung, Thủ Đức.
- Lớp lý thuyết: thứ 2 (hàng tuần) từ tiết 1 tới 3 (07:30 - 09h50)
- Lớp thực hành R bắt đầu từ tuần thứ 3:

  thứ 2, tiết 4 tới 5 (10:00 - 12:00), phòng C205;

  thứ 3, từ tiết 1 - 2.5 (7h30 - 9h30), phòng C204;

  thứ 3, từ tiết 2.5 tới 5 (10h00 - 12h00), phòng C204;

  thứ 4, từ tiết 1 - 2.5 (7h30 - 9h30), phòng C205.

Xem chi tiết lịch học tại trang quản lý môn học.

Thông tin của môn học

Cấu trúc điểm:

- Điểm quá trình thực hành: 30%
- Điểm thi giữa kỳ: 20%
- Điểm thi cuối kỳ: 50%

Nhóm:
Sinh viên được chia thành các nhóm nhỏ, mỗi nhóm gồm từ 4-6 sinh viên.

Hình thức thi

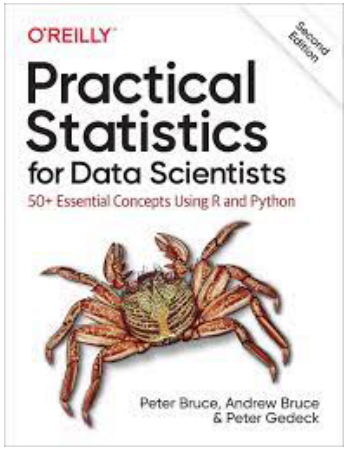
- Thi giữa kỳ: viết tự luận trong thời gian 60 phút.
- Thi cuối kỳ: làm đề tài/bài tập lớn theo nhóm.

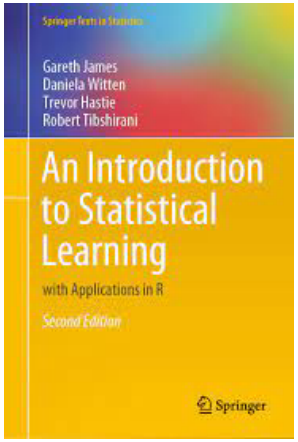
Chú ý: Sinh viên cần phải tham gia ít nhất 80% số tiết học.

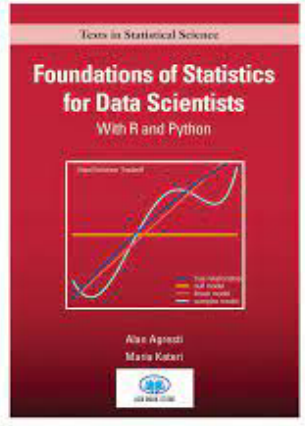
Tài liệu học tập và tài liệu tham khảo

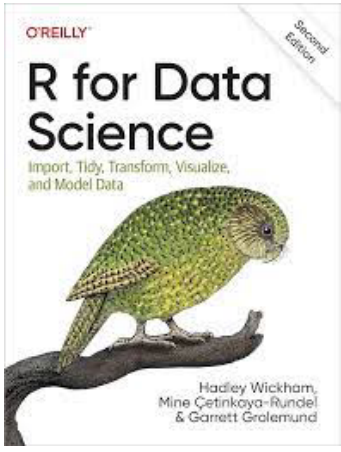
Slide bài giảng được cập nhật tại trang của môn học:
`https://courses.hcmus.edu.vn/course/view.php?id=5579`

Các tài liệu tham khảo

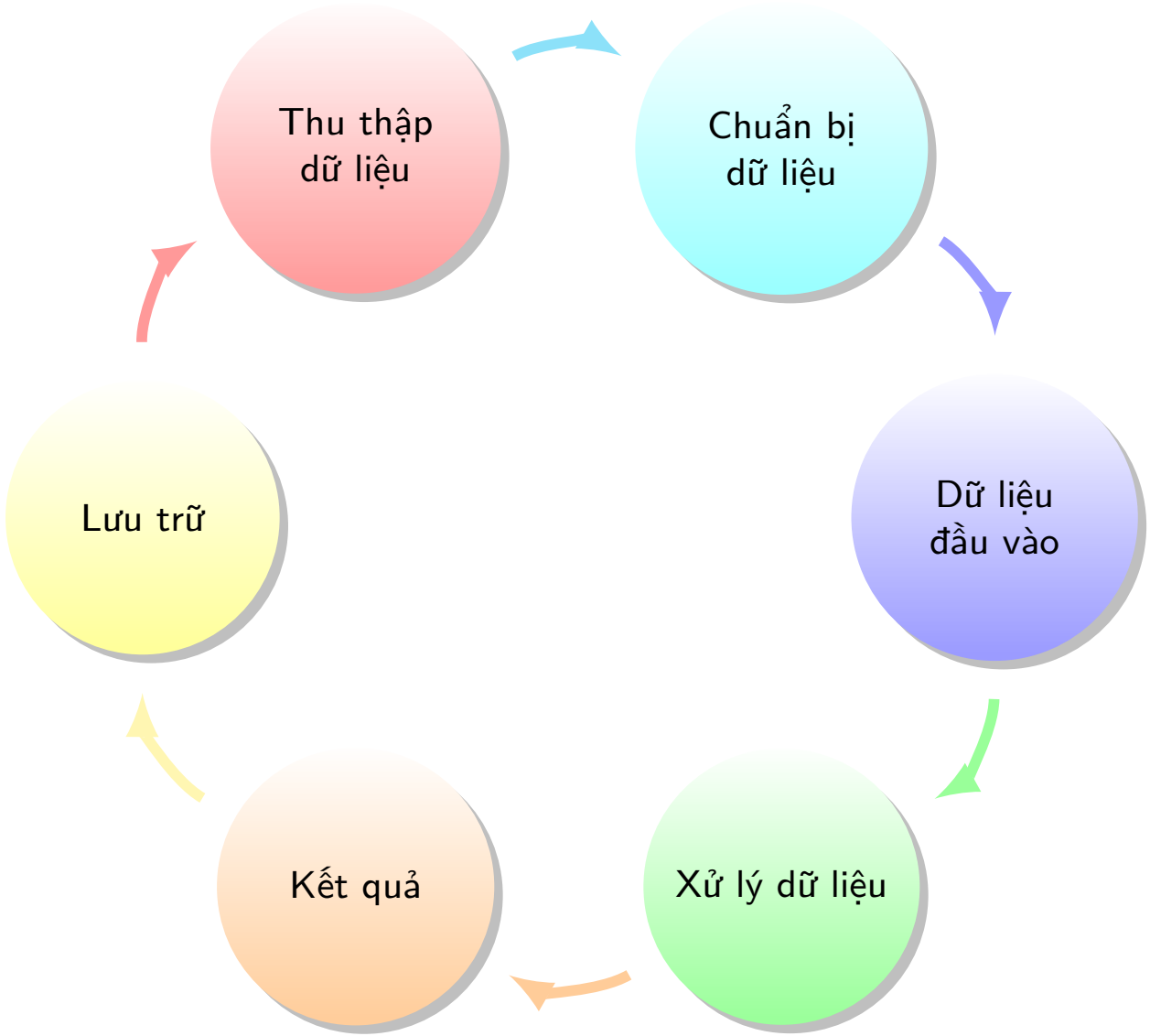








Xử lý số liệu?


Thu thập dữ liệu
Chuẩn bị dữ liệu
Dữ liệu đầu vào
Xử lý dữ liệu
Kết quả
Lưu trữ

Xử lý số liệu?

Thu thập dữ liệu

Dữ liệu thô có thể bao gồm:

- số liệu tiền tệ,
- cookie trang web,
- báo cáo lãi/lỗ của công ty,
- hành vi của người dùng, v.v.

Dữ liệu thô phải được thu thập từ các nguồn được xác định và chính xác để những phát hiện tiếp theo có giá trị và có thể sử dụng được.

Chuẩn bị dữ liệu

Dữ liệu thô được kiểm tra:

- lỗi,
- trùng lặp,
- tính toán sai,
- thiếu dữ liệu,

và chuyển thành dạng phù hợp để phân tích và xử lý tiếp theo.

Xử lý số liệu?

Dữ liệu đầu vào

Dữ liệu thô được chuyển đổi thành dạng máy có thể đọc được và đưa vào bộ xử lý.

Xử lý dữ liệu

Áp dụng

- mô hình thống kê;
- thuật toán máy hoc;
- thuật toán học sâu.

để xử lý dữ liệu nhằm thu được kết quả cho các nhiệm vụ mong muốn.

Xử lý số liệu?

Kết quả

Dữ liệu cuối cùng được truyền và hiển thị cho người dùng ở dạng có thể đọc được:

- biểu đồ;
- bảng;
- video;
- âm thanh;
- tài liệu.

Đầu ra này có thể được lưu trữ và xử lý thêm trong chu trình xử lý dữ liệu tiếp theo.

Lưu trữ

Bước cuối cùng của chu trình xử lý dữ liệu là lưu trữ, nơi dữ liệu và siêu dữ liệu được lưu trữ để sử dụng tiếp:

- excel file;
- sql;
- text file, ...

Xử lý số liệu?

Việc xử lý dữ liệu xảy ra trong cuộc sống hàng ngày của chúng ta cho dù chúng ta có nhận thức được hay không.

- Phần mềm giao dịch chứng khoán chuyển đổi hàng triệu dữ liệu chứng khoán thành biểu đồ đơn giản.
- Một công ty thương mại điện tử sử dụng lịch sử tìm kiếm của khách hàng để giới thiệu các sản phẩm tương tự.
- Một công ty tiếp thị kỹ thuật số sử dụng dữ liệu nhân khẩu học của mọi người để lập chiến lược cho các chiến dịch theo vị trí cụ thể.
- Xe tự lái sử dụng dữ liệu thời gian thực từ các cảm biến để phát hiện xem có người đi bộ và các xe khác trên đường hay không.

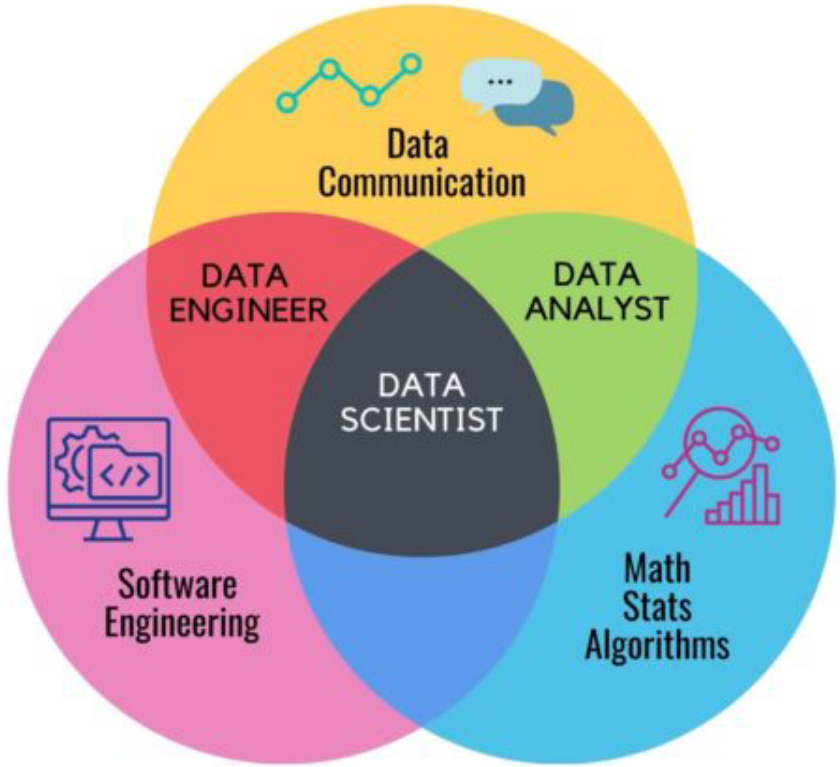
Xử lý số liệu?

	symbol	date	open	high	low	close	volume
1	AMZN	2024-02-15	170.58	171.17	167.59	169.80	49815300
2	AMZN	2024-02-14	169.21	171.21	168.28	170.98	42815500
3	AMZN	2024-02-13	167.73	170.95	165.75	168.64	56345100
4	AMZN	2024-02-12	174.80	175.39	171.54	172.34	51050400
5	AMZN	2024-02-09	170.90	175.00	170.58	174.45	56986000
6	AMZN	2024-02-08	169.65	171.43	168.88	169.84	42316500
7	AMZN	2024-02-07	169.48	170.88	168.94	170.53	47174100
8	AMZN	2024-02-06	169.39	170.71	167.65	169.15	42505500
9	AMZN	2024-02-05	170.20	170.55	167.70	170.31	55081300
10	AMZN	2024-02-02	169.19	172.50	167.33	171.81	117154900
11	AMZN	2024-02-01	155.87	159.76	155.62	159.28	76542400

Xử lý số liệu?

	symbol	date	open	high	low	close	volume
1	AMZN	2024-02-15	170.58	171.17	167.59	169.80	49815300
2	AMZN	2024-02-14	169.21	171.21	168.28	170.98	42815500
3	AMZN	2024-02-13	167.73	170.95	165.75	168.64	56345100
4	AMZN	2024-02-12	174.80				
5	AMZN	2024-02-09	170.90				
6	AMZN	2024-02-08	169.65				
7	AMZN	2024-02-07	169.48				
8	AMZN	2024-02-06	169.39				
9	AMZN	2024-02-05	170.20				
10	AMZN	2024-02-02	169.19				
11	AMZN	2024-02-01	155.87				

Data Analysts - Data Engineer - Data scientists



Data Analysts - Data Engineer - Data scientists

Data analysts - hay **nhà phân tích dữ liệu** là người có trách nhiệm phân tích dữ liệu lịch sử, thiết kế hệ thống lưu trữ dữ liệu và sử dụng các công cụ khác nhau để phân tích dữ liệu.

Data engineer - hay **kỹ sư dữ liệu** là người được giao nhiệm vụ phát triển, xây dựng và quản lý hệ thống dữ liệu. Họ đảm bảo rằng dữ liệu có sẵn đã sẵn sàng để xử lý và phân tích ở định dạng có thể dễ dàng phân tích.

Data scientists - hay **nhà khoa học dữ liệu** là người sử dụng dữ liệu để trả lời các câu hỏi về công ty. Họ sử dụng dữ liệu để phát triển các tính năng sản phẩm mới. Họ sẽ dành một lượng thời gian đáng kể để làm sạch dữ liệu để đảm bảo rằng dữ liệu đó phù hợp với mô hình và kỹ thuật phân tích, xử lý số liệu của họ.

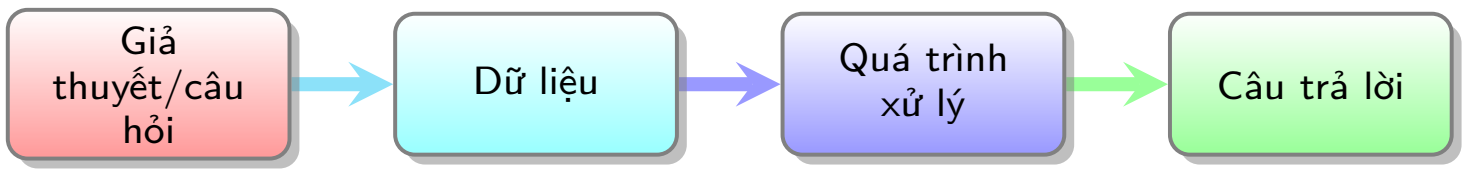
Công việc của Data scientists

Statistical learning vs Machine learning

Thống kê (**Statistics**) là một ngành khoa học của việc:

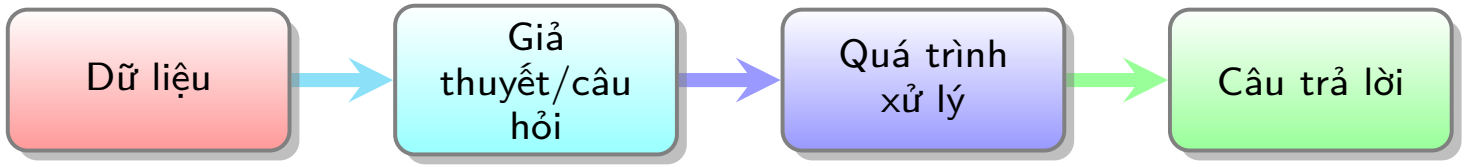
- **thu thập** dữ liệu
- **biểu diễn** dữ liệu
- **phân tích** dữ liệu
- **giải thích** kết quả

nhằm tìm ra các bằng chứng để cung cố hoặc bác bỏ các giả thuyết khoa học hoặc câu hỏi nghiên cứu.



Statistical learning vs Machine learning

Statistical learning/**Machine learning** là sự áp dụng của các phương pháp thống kê (chủ yếu là hồi quy) để tạo ra sự tiên đoán về các dữ liệu chưa nhìn thấy.



Statistical learning vs Machine learning

Statistical learning tập trung vào phát triển và phân tích các mô hình có thể đưa ra dự đoán hoặc suy luận dựa trên dữ liệu:

- so sánh;
- mối tương quan;
- sự ảnh hưởng.

Ví dụ:

- so sánh số lượng click của hai thiết kế tiêu đề trang web;
- sự ảnh hưởng của thời tiết tới số lượng thuê xe đạp;
- nguyên nhân nào tác động tới tỷ lệ từ bỏ dịch vụ của khách hàng.

Statistical learning hoạt động dựa trên các giả định cho dữ liệu:

- phân phối chuẩn;
- độc lập của các quan sát;
- đồng nhất phương sai.

Statistical learning vs Machine learning

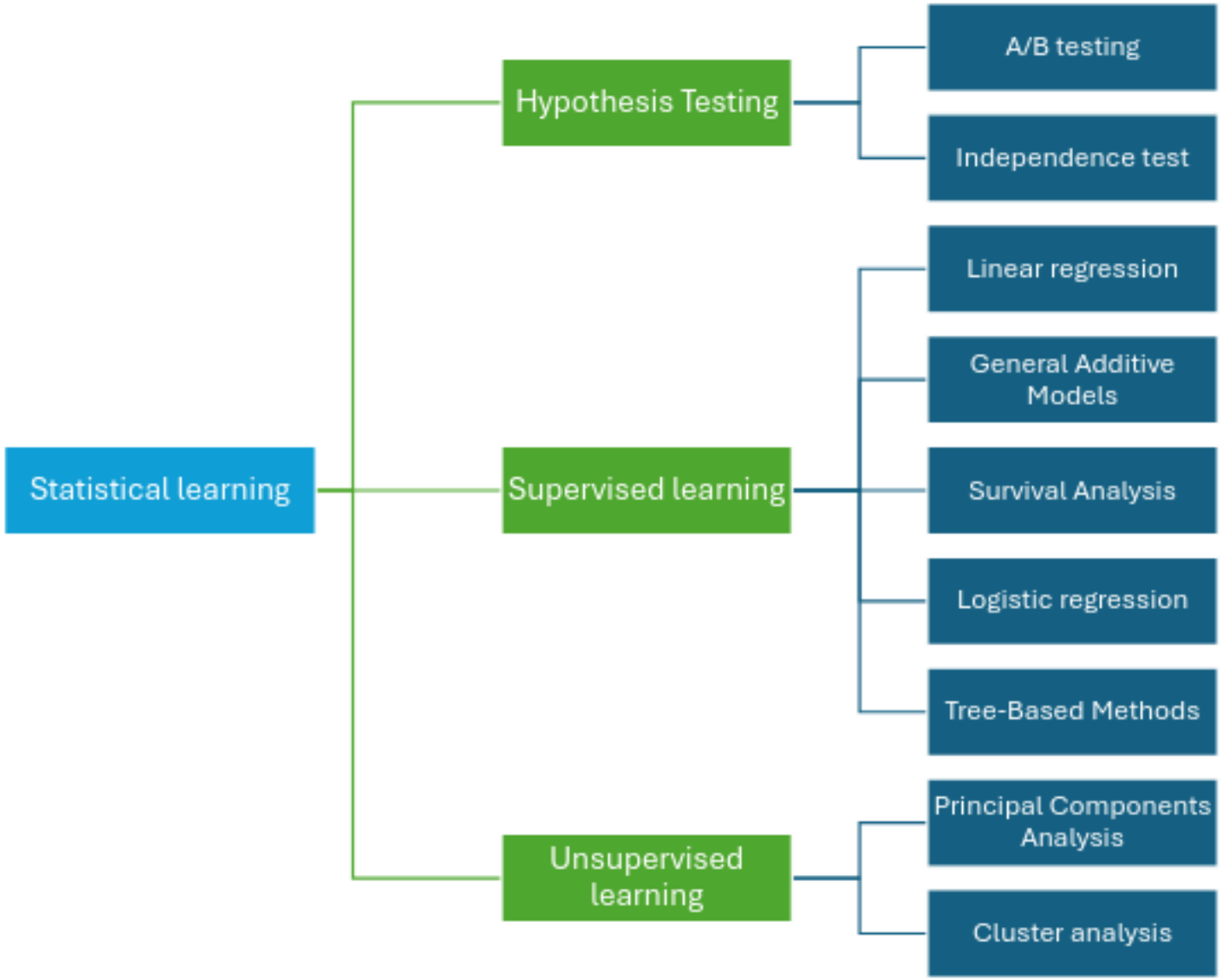
Machine learning tập trung vào việc trích xuất kiến thức từ dữ liệu, bao gồm:

- các kỹ thuật hình thành khái niệm;
- phát hiện mẫu;
- các phương pháp lựa chọn tính năng tự động

Machine learning không phụ thuộc vào các giả định và trong hầu hết các trường hợp đều bỏ qua chúng.

Các mô hình machine learning được thiết kế để đưa ra dự đoán chính xác nhất có thể.

Statistical learning vs Machine learning



Ví dụ: Dự đoán doanh số bán hàng

Một công ty muốn dự đoán doanh số bán hàng (`sales`) của một mặt hàng tại một thị trường dựa trên chi phí quảng cáo trên các nền tảng:

- truyền hình (`TV`)
- đài phát thanh (`radio`)
- nhật báo in (`newspaper`)

	A	B	C	D	E
1		TV	radio	newspaper	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

Ví dụ: Dự đoán doanh số bán hàng

Machine learning: dự đoán `sales` thông qua `TV`, `radio` và `newspaper`

$$\texttt{sales} = w_0 + w_1 \times \texttt{TV} + w_2 \times \texttt{radio} + w_3 \times \texttt{newspaper}$$

Ta cần:

- xác định các trọng số w_0, w_1, w_2 và w_3 sao cho sự tiên đoán có độ **chính xác cao nhất**:

$$(\widehat{w}_0, \widehat{w}_1, \widehat{w}_2, \widehat{w}_3) = \arg\min \frac{1}{2} \sum_{i=1}^{n_{\text{train}}} \left(\texttt{sales}_i - \widehat{\texttt{sales}}_i\right)^2$$

  trong đó, $\widehat{\texttt{sales}}$ là giá trị tiên đoán;

- đánh giá độ chính xác của tiên đoán trên test set.

Ví dụ: Dự đoán doanh số bán hàng

Statistical learning: xây dựng mô hình ước lượng/giải thích sự biến động của `sales` thông qua `TV`, `radio` và `newspaper`

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \varepsilon,$$

trong đó,

- β_0, β_1, β_2 và β_3 là các hệ số của mô hình;
- ε là sai số của mô hình (sai số đo, sự biến động của `sales` thông qua các yếu tố khác).

Ta cần:

- ước lượng các hệ số β_0, β_1, β_2 và β_3 sao cho trung bình bình phương sai số nhỏ nhất:

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3) = \arg\min \sum_{i=1}^{n_{\text{train}}} \widehat{\varepsilon}_i^2 = \arg\min \sum_{i=1}^{n_{\text{train}}} \left(\texttt{sales}_i - \widehat{\texttt{sales}}_i\right)^2$$

- kiểm tra sự ý nghĩa của biến trong mô hình;
- giải thích sự tác động của hình thức quảng cáo tới sự thay đổi của `sales`;
- đánh giá độ chính xác của tiên đoán trên test set.

Dữ liệu?

Dữ liệu

Dữ liệu là một hoặc nhiều thông tin mô tả **tính chất của các đối tượng nghiên cứu**.

Dữ liệu có thể tới từ nhiều nguồn:

- cảm biến (sensor);
- sự kiện;
- văn bản (text);
- hình ảnh (images);
- âm thanh;
- video.

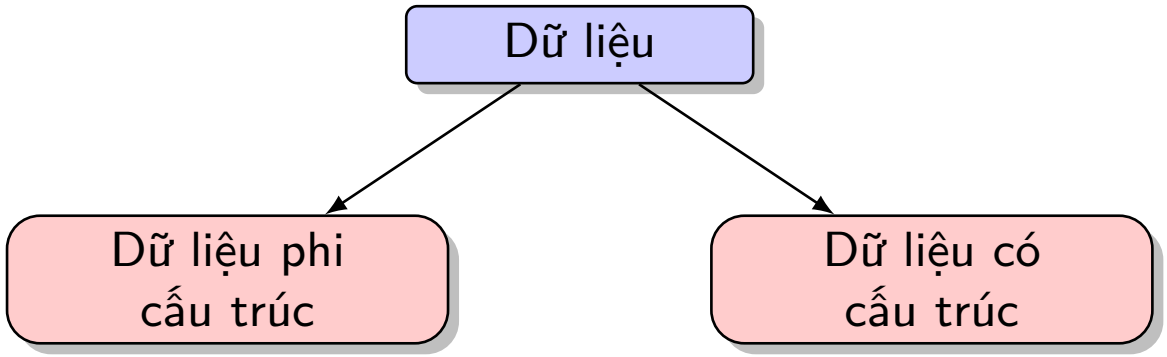
Dữ liệu?

Dữ liệu

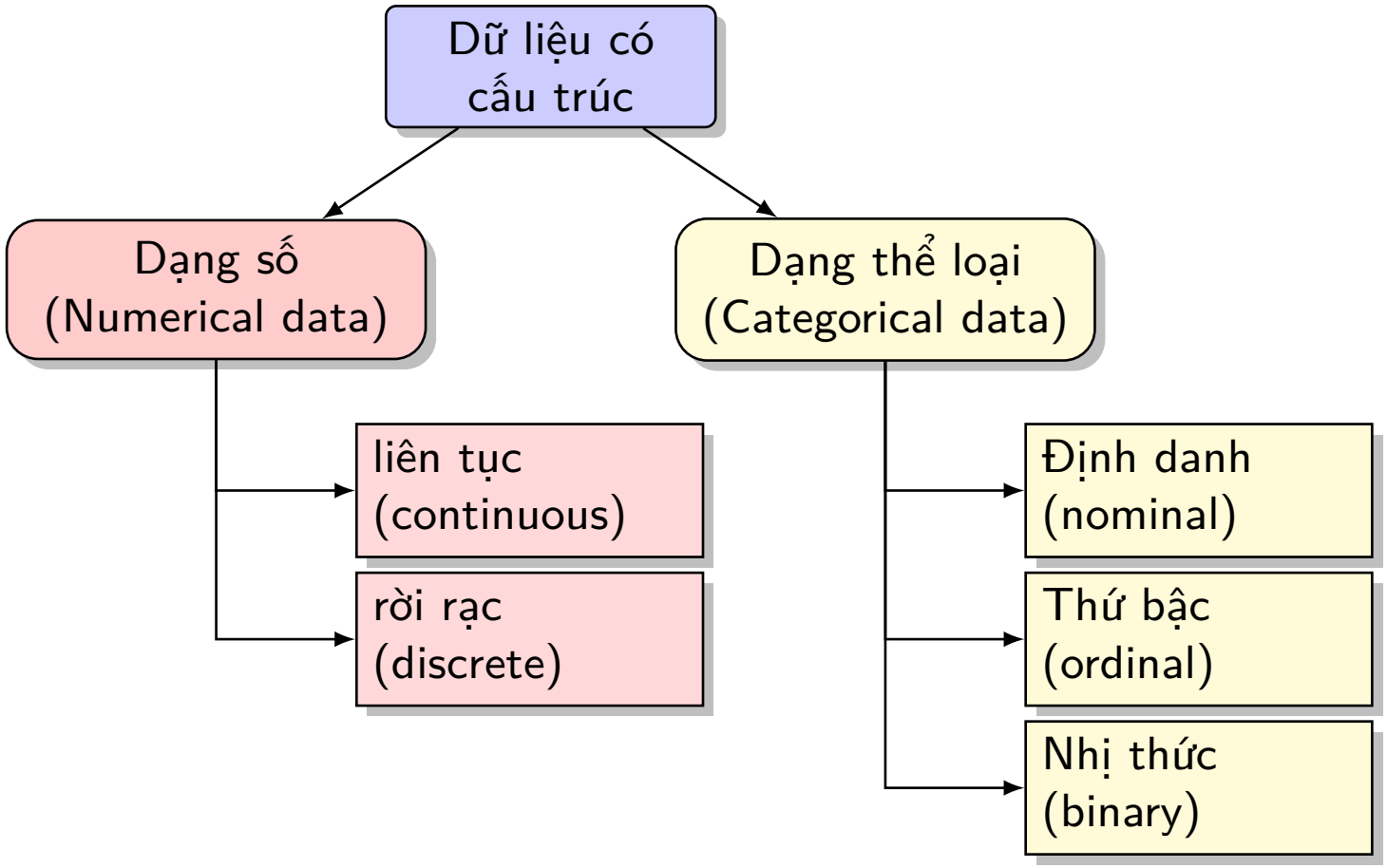
Dữ liệu là một hoặc nhiều thông tin mô tả **tính chất của các đối tượng nghiên cứu**.

Dữ liệu có thể tới từ nhiều nguồn:

- cảm biến (sensor);
- sự kiện;
- văn bản (text);
- hình ảnh (images);
- âm thanh;
- video.



Dữ liệu có cấu trúc - Structured data



Dữ liệu có cấu trúc - Structured data

Dữ liệu dạng số (*numerical data* hay *quantitative data*) là dữ liệu được cấu trúc với giá trị là số:

- liên tục (*continuous*);
- rời rạc hay số đếm (*discrete*),

ghi lại/biểu diễn số lượng của một đối đối tượng nào đó.

Dữ liệu dạng phân loại (*categorical data* hay *qualitative data*) là dữ liệu được cấu trúc với giá trị là các thể loại khác nhau của người hoặc vật. Thông thường ta có hai dạng phổ biến sau:

- biến định danh (*nominal*)
- biến thứ bậc (*ordinal*)

Đặc biệt, đối với dữ liệu định danh, ta có:

- dữ liệu nhị định danh hay dữ liệu nhị phân (*binary data*), khi biến có 2 giá trị;
- dữ liệu đa định danh hay dữ liệu đa thức (*nominal multinomial data* hoặc *multinomial data*).

Dữ liệu có cấu trúc - Structured data

Ví dụ:

- số liên tục: tốc độ gió, doanh thu, khoảng thời gian, chi phí, ...
- số rời rạc: số lượng lượt truy cập web, số lượng khách hàng trong một ngày, ...
- đa định danh: thể loại màn hành TV: plasma, LCD, LED, ...
- thứ bậc: điểm đánh giá sản phẩm, thăng hài lòng về chất lượng dịch vụ, ...
- nhị phân: có/không, đúng/sai, ...

Ví dụ dữ liệu cấu trúc

Chi phí quảng cáo trên các nền tảng:

- truyền hình (`TV`)
- đài phát thanh (`radio`)
- nhật báo in (`newspaper`)

và doanh thu bán sản phẩm của công ty đều là các dữ liệu dạng số liên tục.

	A	B	C	D	E
1		TV	radio	newspaper	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

Biến đại diện - Proxy variable

Biến đại diện (**Proxy variable**) là biến đại diện cho biến được quan tâm thực sự, có thể không có sẵn, quá tốn kém hoặc quá tốn thời gian để đo lường.

- hàm lượng oxy trong lõi băng cổ đại: đại diện cho nhiệt độ;
- body mass index (BMI): đại diện cho % mỡ cơ thể;
- số năm được giáo dục/điểm GPA: đại diện cho khả năng nhận thức;
- ảnh vệ tinh màu sắc bề mặt đại dương: đại diện cho độ sâu mà ánh sáng xuyên qua đại dương trên diện rộng;
- sự thay đổi chiều cao trong một thời gian cố định: đại diện cho nồng độ hormone trong máu.

Biến đại diện - Proxy variable

Ví dụ: *Web stickiness*

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Biến đại diện - Proxy variable

Ví dụ: *Web stickiness*

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Các bước xác định mẫu:

Bán hàng tốt hơn $\Longleftrightarrow$ doanh số bán hàng của dịch vụ từ hai trang web có sự khác biệt lớn.
$\hookrightarrow$ ta cần thu thập tổng doanh số bán hàng của hai trang web.

Biến đại diện - Proxy variable

Ví dụ: *Web stickiness*

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Các bước xác định mẫu:

Bán hàng tốt hơn $\Longleftrightarrow$ doanh số bán hàng của dịch vụ từ hai trang web có sự khác biệt lớn.
$\hookrightarrow$ ta cần thu thập tổng doanh số bán hàng của hai trang web.

Tuy nhiên,

- giá trị dịch vụ được bán cao;
- việc bán hàng không thường xuyên;
- chu kỳ bán hàng kéo dài

$\Rightarrow$ cần nhiều thời gian để thu thập đủ doanh số để đánh giá;
$\Rightarrow$ không khả thi, khó thực hiện.

Biến đại diện - Proxy variable

Ví dụ: *Web stickiness*

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Các bước xác định mẫu:

Bán hàng tốt hơn $\Longleftrightarrow$ doanh số bán hàng của dịch vụ từ hai trang web có sự khác biệt lớn.
$\hookrightarrow$ ta cần thu thập tổng doanh số bán hàng của hai trang web.

Tuy nhiên,

- giá trị dịch vụ được bán cao;
- việc bán hàng không thường xuyên;
- chu kỳ bán hàng kéo dài

$\Rightarrow$ cần nhiều thời gian để thu thập đủ doanh số để đánh giá;
$\Rightarrow$ không khả thi, khó thực hiện.

Cần thu thập một dữ liệu khác (biến khác), thay thế tổng doanh số bán hàng.

Biến đại diện - Proxy variable

Một biến đại diện tiềm năng cho tổng doanh số bán hàng trên web là

số lần nhấp chuột vào trang đích chi tiết

Tại sao?

Biến đại diện - Proxy variable

Một biến đại diện tiềm năng cho tổng doanh số bán hàng trên web là

số lần nhấp chuột vào trang đích chi tiết

Tại sao?

Một bản trình bày trên web thu hút sự chú ý của mọi người lâu hơn $\Rightarrow$ doanh số bán hàng cao hơn.

$\hookrightarrow$ dữ liệu là thời gian phiên trung bình (session time), đơn vị giây, so sánh trang web A với trang web B.

	Page	Time
1	Page A	0.21
2	Page B	2.53
3	Page A	0.35
4	Page B	0.71
5	Page A	0.67
6	Page B	0.85

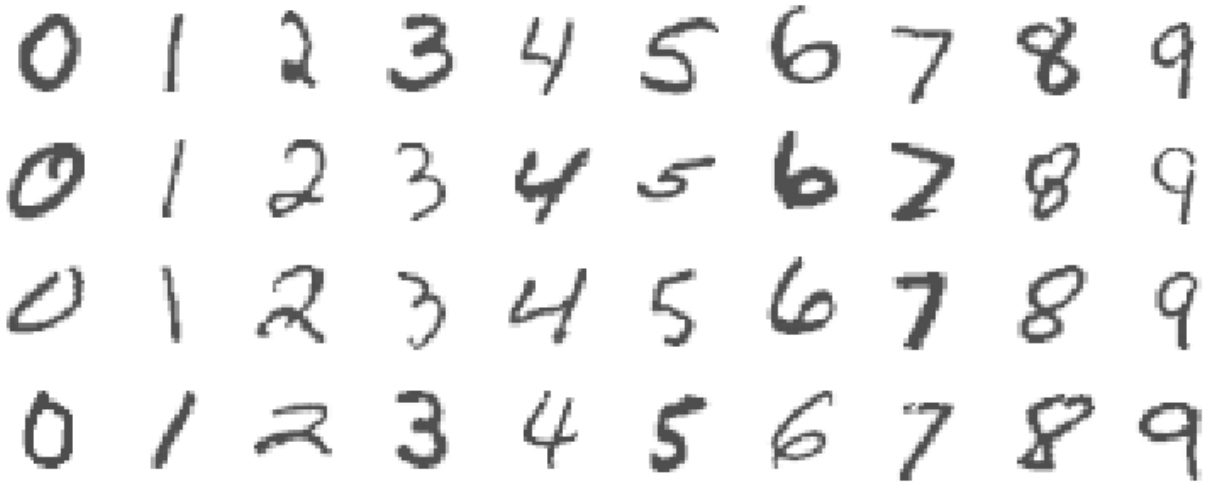
Dữ liệu phi cấu trúc - Unstructured data

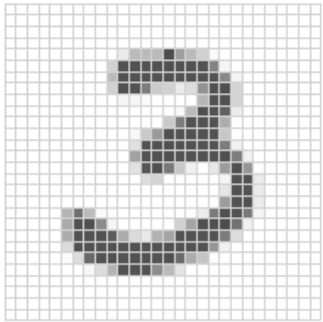
Dữ liệu phi cấu trúc là những dữ liệu được thu thập không có dạng số hoặc dạng thể loại:

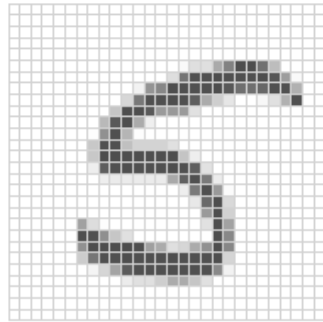
- hình ảnh - images: là tập hợp các pixel với mỗi pixel chứa thông tin màu RGB (đỏ - red, xanh lá - green, xanh dương - blue);
- văn bản - text: là chuỗi các từ và ký tự không phải từ, thường được sắp xếp theo các phần, phần phụ, v.v;
- dòng nhấp chuột - clickstreams: là chuỗi hành động do người dùng tương tác với một ứng dụng hoặc trang web.

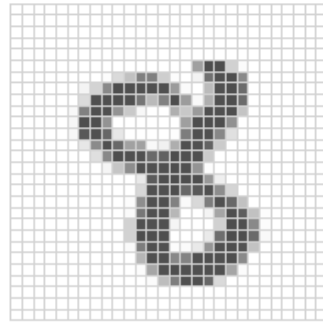
Để áp dụng các mô hình của statistical learning, dữ liệu thô phi cấu trúc phải được xử lý và thao tác thành dạng có cấu trúc.

Ví dụ dữ liệu phi cấu trúc









Ví dụ dữ liệu phi cấu trúc

Dữ liệu về hình chụp phiết máu mỏng của hai người:

- nhiễm ký sinh trùng (a);
- không bị nhiễm ký sinh trùng (b).

(a) Parasitized

(b) Uninfected

Rectangular Data

Rectangular Data

Rectangular Data có hình dạng giống như hình chữ nhật, tức là một ma trận hai chiều với các hàng (row) biểu thị các đối tượng (trường hợp) và các cột (column) biểu thị các đặc tính (feature) hay biến (variable).

Category	currency	sellerRating	Duration	endDay	ClosePrice	OpenPrice	Competitive?
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	1
Automotive	US	3115	7	Tue	0.01	0.01	1

Rectangular Data

Các thành phần của rectangular data:

- feature: một cột trong bảng dữ liệu thường ám chỉ tới đặc tính (feature) của đối tượng,

  đồng nghĩa với attribute, input, predictor, variable;
- outcome: là một đặc tính của đối tượng, được sử dụng như là kết quả của sự tiên đoán, hay phân loại,

  đồng nghĩa với dependent variable, response, target, output;
- records: một dòng của bảng dữ liệu thường ám chỉ tới một ghi chép của một đối tượng;

  đồng nghĩa với case, example, instance, observation, pattern, sample.
- data frame: bảng dữ liệu tổng hợp nhiều cột và dòng, biểu diễn ghi chép các đặc tính khác nhau của nhiều đối tượng;

  một bảng dữ liệu có thế chứa hỗn hợp các dạng dữ liệu có cấu trúc.

Rectangular Data

Tuy nhiên, dữ liệu mà ta làm việc, không phải lúc nào cũng ở sẵn dạng rectangular data.

- dữ liệu phi cấu trúc: phải được xử lý và thao tác để có thể biểu diễn dưới dạng tập hợp các tính năng;
- dữ liệu trong cơ sở dữ liệu phải được trích xuất và đưa vào một bảng duy nhất cho hầu hết các nhiệm vụ phân tích và lập mô hình dữ liệu.

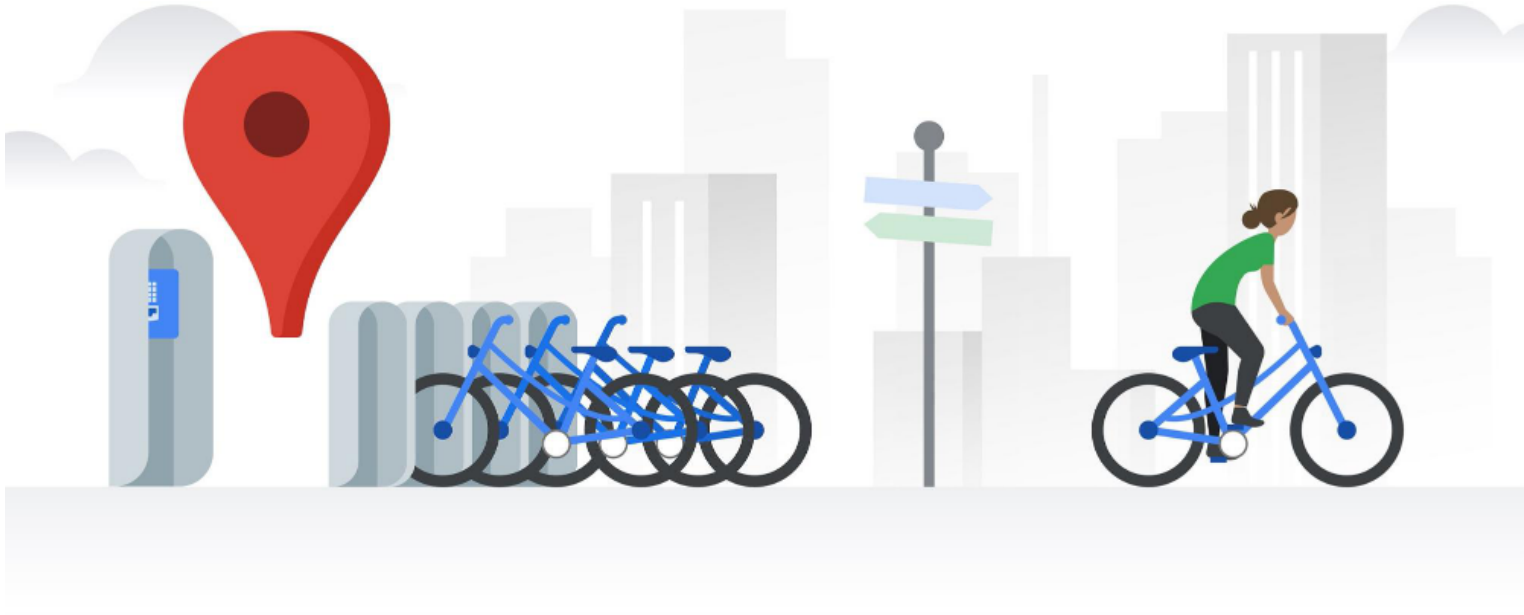
Non-rectangular Data

Một số dạng dữ liệu sau không phải rectangular data:

- dữ liệu dạng danh sách: trong đó, mỗi danh sách là một rectangular data (cross-sectional data, panel data);
- dữ liệu chuỗi thời gian (time series data): ghi lại các phép đo liên tiếp của cùng một biến (feature);
- dữ liệu không gian (spatial data): được sử dụng trong phân tích bản đồ và vị trí;
- dữ liệu đồ thị hoặc mạng lưới (graph, network data): được sử dụng để thể hiện các mối quan hệ vật lý, xã hội và trừu tượng.

Case study 1: So sánh độ hiệu quả của hai dòng tiêu đề của trang web

Case study 2: Dự đoán lượng thuê xe đạp công cộng [1]



[1] @https://www.kaggle.com/datasets/yasserh/bike-sharing-dataset

Case study 3: Phân tích tỷ lệ khách hàng từ bỏ dịch vụ