TS. Nguyễn Thế Vịnh
TS. Quách Xuân Trưởng
TS. Trần Quang Quý
ThS. Nguyễn Văn Việt

# GIÁO TRÌNH

# PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU

2023

Trường Đại học
CNTT &TT, ĐHTN
Quyết Thắng,
Thái nguyên
https://ictu.edu.vn/

# DANH MỤC HÌNH VẼ

# DANH MỤC BẢNG BIỂU

# THUẬT NGỮ VÀ TỪ VIẾT TẮT

| Thuật ngữ | Mô tả |
|---|---|
| vis | Là viết tắt của visualization (trực quan hóa). |
| SPLOM | SPLOM hoặc ma trận biểu đồ phân tán (ScatterPLOt Matrix) là một kỹ thuật được phát minh bởi John Hartigan vào năm 1975, được sử dụng để xác định mối tương quan giữa một chuỗi các biến. |
| HCE | Là viết tắt của từ Hierarchical Clustering Explorer. Đây là công cụ khám phá phân cụm phân cấp được thiết kế để nhập bất kỳ dự án microarray nào có sẵn và kiểm tra tương tác tác động của các định nghĩa do người dùng xác định. |
| Scagnostics | Là viết tắt của cụm từ "scatterplot diagnostics". Nó là một thuật ngữ trong lĩnh vực thống kê và trực quan hóa dữ liệu. |
| Monotonic | Một biểu đồ phân tán được coi là monotonic nếu có một xu hướng tăng dần (monotonic increasing) hoặc giảm dần (monotonic decreasing) giữa các giá trị của hai biến được biểu diễn trên trục x và trục y |
| Stringy | Một biểu đồ phân tán được xem là "stringy" khi có sự tập trung của các điểm dữ liệu thành các dải hoặc chuỗi kéo dài theo một hướng cụ thể |
| Skinny | Một biểu đồ phân tán được xem là "skinny" khi nó có hình dáng hẹp và dài, thường là do sự tập trung của các điểm dữ liệu thành một dải hoặc chuỗi kéo dài theo một hướng cụ thể |
| Convex | Một biểu đồ phân tán được xem là "convex" khi các điểm dữ liệu tạo thành một hình dạng lồi (convex shape) hoặc hình dạng hộp (box shape) |
| Striated | Một biểu đồ phân tán được coi là "striated" khi có sự xuất hiện của các vết sọc, đường kẻ hoặc mẫu vằn trên biểu đồ, tạo thành các hình dạng đặc trưng |
| Sparse | Một biểu đồ phân tán được xem là "sparse" khi có sự thưa thớt và phân tán của các điểm dữ liệu trên biểu đồ. |
| Clumpy | Một biểu đồ phân tán được xem là "clumpy" khi có sự tập trung của các điểm dữ liệu thành các nhóm hoặc cụm. |
| Skewed | Một biểu đồ phân tán được coi là "skewed" khi có sự chênh lệch hoặc méo lệch của phân phối dữ liệu trên trục x hoặc trục y. |
| Outlying | Một biểu đồ phân tán được coi là "outlying" khi có sự xuất hiện của các điểm dữ liệu nằm xa hơn so với phần lớn các điểm dữ liệu khác. |

## Mở đầu

Giáo trình ***Phân tích và trực quan dữ liệu*** được tập thể giảng viên Khoa Công nghệ thông tin biên soạn nhằm phục vụ cho việc giảng dạy theo chương trình đào tạo mới, kết hợp xen kẽ giữa lý thuyết và thực hành.

Nội dung giáo trình cung cấp cho sinh viên các kiến thức cơ bản về phân tích và trực quan dữ liệu; trừu tượng hóa dữ liệu và tác vụ; hiểu các thành phần cơ bản của dữ liệu trực quan; thao tác với các loại dữ liệu khác nhau; trực quan hóa dữ liệu theo không gian, thời gian, và mạng lưới. Sau khi hoàn thành học phần này, sinh viên có thể thành thạo việc trực quan hóa và phân tích các loại dữ liệu khác nhau, qua đó làm tăng khả năng tư duy logic và nâng cao khả năng cơ hội việc làm sau khi tốt nghiệp. Nội dung giáo trình gồm 9 chương:

Chương 1. Tổng quan về phân tích và trực quan dữ liệu.

Chương 2. Trừu tượng hóa dữ liệu.

Chương 3. Trừu tượng hóa tác vụ.

Chương 4. Bốn cấp độ về hợp lệ dữ liệu.

Chương 5. Các thành phần dữ liệu trực quan.

Chương 6. Dữ liệu trực quan theo không gian.

Chương 7. Dữ liệu trực quan theo thời gian.

Chương 8. Dữ liệu trực quan theo mạng lưới.

Chương 9. Phân tích dữ liệu trực quan qua các ví dụ điển hỉnh.

Mặc dù tập thể tác giả đã dành nhiều thời gian và công sức để biên soạn, song khó tránh khỏi thiếu sót. Vì vậy, chúng tôi kính mong quý thầy cô và các bạn sinh viên đóng góp ý kiến để cuốn giáo trình được hoàn thiện hơn. Xin trân trọng cảm ơn !

# Chương 1 : TỔNG QUAN VỀ PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU

## *Nội dung chính của chương*

Chương này nhằm giới thiệu cho sinh viên vai trò và tầm quan trọng của trực quan hóa dữ liệu, các ví dụ điển hình cũng như các công cụ và tài liệu sẽ được sử dụng trong suốt quyển giáo trình. Bố cục của chương được chia thành 03 phần. Phần 01 giới thiệu một số các ví dụ điển hình khi áp dụng trực quan hóa dữ liệu so với phương pháp phân tích thuần túy, qua đó sinh viên có thể tự suy ngẫm và cân nhắc có cần phải trực quan hóa dữ liệu hay không. Phần 02 tập trung vào một số các khái niệm cơ bản, công cụ điển hình nhằm giúp cho sinh viên làm quen với chủ đề. Phần 03 hướng dẫn cài đặt một số công cụ, thư viện, dữ liệu để giúp cho sinh viên thực hành xuyên suốt quyển giáo trình.

## *Mục tiêu cần đạt được của chương*

- Hiểu được tầm quan trọng của việc phân tích và trực quan dữ liệu
- Nhớ và liên tưởng tới một số các ví dụ điển hình của trực quan dữ liệu
- Nắm được một số công cụ, thư viện, dữ liệu hữu ích cho việc thực hành trực quan dữ liệu

## Một số ví dụ điển hình của phân tích và trực quan dữ liệu

### *1.1 Cuộc xâm lược nước Nga của Napoléon*

Việc học lịch sử có lẽ là vấn đề gây ra nhiều khó khăn đối với các bạn học sinh bởi lẽ nó liên quan đến rất nhiều câu chữ, dấu mốc thời gian, số lượng quân đội, địa điểm,... Chúng ta thường gặp những bài giảng lịch sử kiểu như:

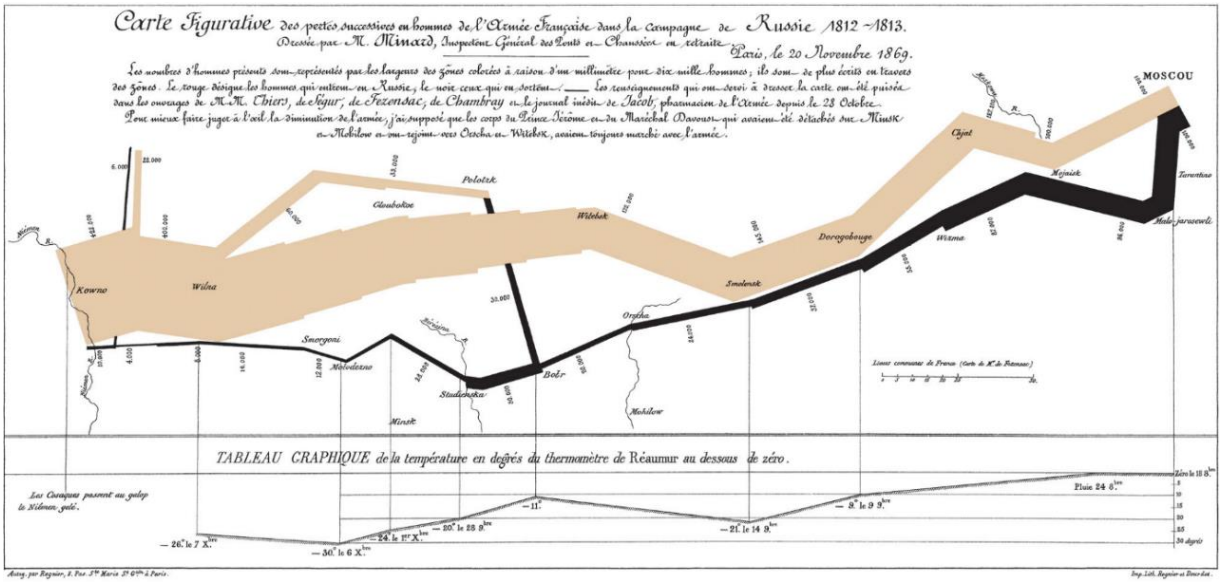
> "...Chiến dịch nước Nga là bước ngoặt trong các cuộc chiến tranh của Napoléon. Trong vòng 10 năm trước đó, quân Pháp thắng trận liên tục và xâm chiếm phần lớn châu Âu, nhưng tại Nga thì họ đã chịu thất bại lớn với hơn 550.000 quân thương vong hoặc đào ngũ. Thất bại này làm sụt giảm nghiêm trọng sức mạnh của Đế chế Pháp và lực lượng đồng minh, gây ra sự thay đổi lớn trong nền chính trị ở châu Âu và làm suy giảm đáng kể quyền bá chủ của Pháp ở châu Âu. Danh tiếng của Napoléon là một thiên tài quân sự bất khả chiến bại đã mất đi sau thất bại này. Các đồng minh của Pháp, đầu tiên là Phổ và sau đó là Áo, đã lần lượt phá vỡ liên minh và chuyển sang chống lại Pháp, gây nên chiến tranh Liên minh thứ sáu.
>
> Chiến dịch bắt đầu ngày 24 tháng 6 năm 1812, khi Napoléon vượt sông Neman. Napoléon buộc Hoàng đế nước Nga là Aleksandr I ở lại trong Hệ thống phong tỏa Lục địa của Vương quốc Anh, mục đích chính là để tránh các đe dọa của Nga tới Ba Lan. Napoléon đặt tên cuộc xâm lăng này là chiến dịch Ba Lan lần thứ hai, người Nga tuyên bố phát động một cuộc chiến tranh Vệ quốc...
>
> Hơn 680.000 tinh binh của đại quân Napoléon hành quân qua phía tây nước Nga, họ giành chiến thắng trong một số trận đánh nhỏ và một trận đánh lớn ở Smolensk vào 16-18 tháng 8. Tuy nhiên, trên cùng một ngày, cánh bắc của quân đội Nga do nguyên soái Pyotr Khristianovich Wittgenstein đã chặn cánh bắc của quân đội Pháp, dẫn đầu bởi Thống chế Nicolas Oudinot trong trận Polotsk. Điều này ngăn chặn cuộc hành quân của Pháp tới kinh thành Sankt-Peterburg; số phận của cuộc chiến

Câu hỏi đặt ra là:

***Làm thế nào để hấp thụ những thông tin trên dễ dàng hơn?***

Chúng ta hãy xem một cách tiếp cận khác về vấn đề lịch sử trên từ góc nhìn của trực quan dữ liệu. Biểu đồ của Minard [1] hiển thị sáu loại thông tin: địa lý, thời gian, nhiệt độ, hướng di chuyển của quân đội và số lượng quân còn lại (xem Hình 1-1). Chiều rộng của các con đường vàng (tiến quân) và đen (rút lui) thể hiện quy mô của lực lượng, một milimét cho 10.000 người. Các đặc điểm địa lý và các trận đánh lớn được

đánh dấu và đặt tên, đồng thời nhiệt độ giảm mạnh trên hành trình trở về được hiển thị dọc phía dưới.




*Hình 1-1 Cuộc xâm lược nước Nga của Napoléon*

Rõ ràng, chỉ với một bức hình nhưng nội dung mà nó có thể truyền đạt là **"đáng giá một nghìn từ"**

## *1.2 Phân tích dữ liệu thống kê*

Chúng ta hãy cùng xem xét bốn nhóm dữ liệu sau [2], mỗi nhóm dữ liệu sẽ có 02 thuộc tính (X, Y). Ở đây, điểm trung bình là điểm nằm giữa của dữ liệu, phương sai là mức độ phân tán của dữ liệu, và tương quan là mối liên hệ giữa các thuộc tính (xem Bảng 1-1).

*Bảng 1-1 Trung bình, phương sai và tương quan của 4 nhóm dữ liệu*

| | **Nhóm 1** | | **Nhóm 2** | | **Nhóm 3** | | **Nhóm 4** | |
|---|---|---|---|---|---|---|---|---|
| | X | Y | X | Y | X | Y | X | Y |
| | 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| | 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| | 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| | 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| | 11.0 | 8.33 | 11.0 | 9.23 | 11.0 | 7.81 | 8.0 | 8.47 |
| | 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| | 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| | 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| | 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| | 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| | 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |
| Trung bình | **9.0** | *7.5* | **9.0** | *7.5* | **9.0** | *7.5* | **9.0** | *7.5* |
| Phương sai | 10.0 | 3.75 | 10.0 | 3.75 | 10.0 | 3.75 | 10.0 | 3.75 |
| Tương quan | 0.816 | | 0.816 | | 0.816 | | 0.816 | |