



TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Ứng dụng học máy trong bài toán dự đoán

Viet-Trung Tran

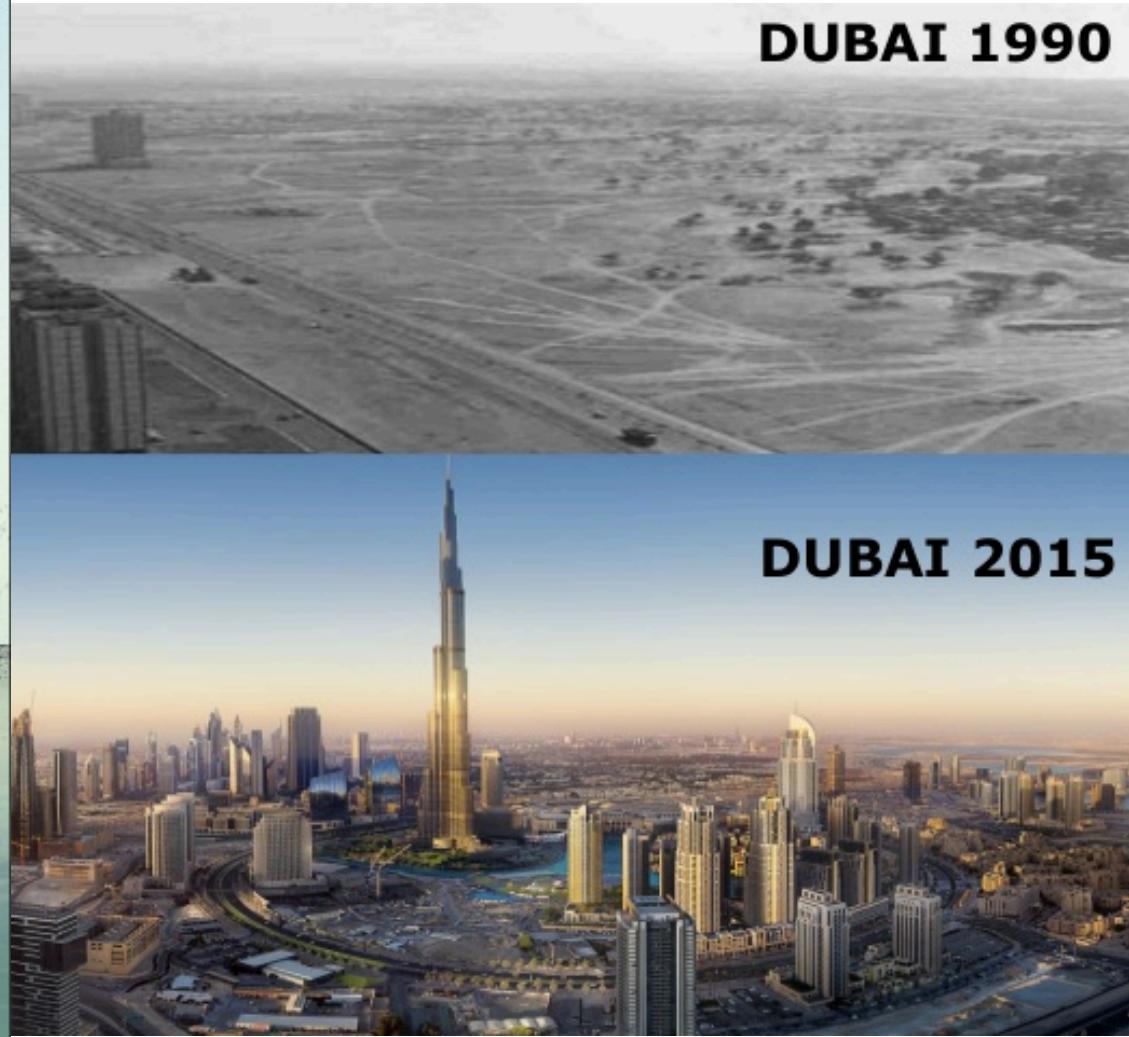
trungtv.github.io

School of Information and Communication Technology

Outline

- Tổng quan về phân tích dữ liệu
- Quy trình làm khoa học dữ liệu
- Ngoại lai và dự đoán ngoại lai

Dữ liệu được ví như nguồn tài nguyên dầu mỏ mới



Đặc điểm 5'V của dữ liệu lớn



Dữ liệu lớn là tập dữ liệu quá lớn hoặc là quá phức tạp mà các nền tảng lưu trữ và xử lý dữ liệu truyền thống không đáp ứng được.

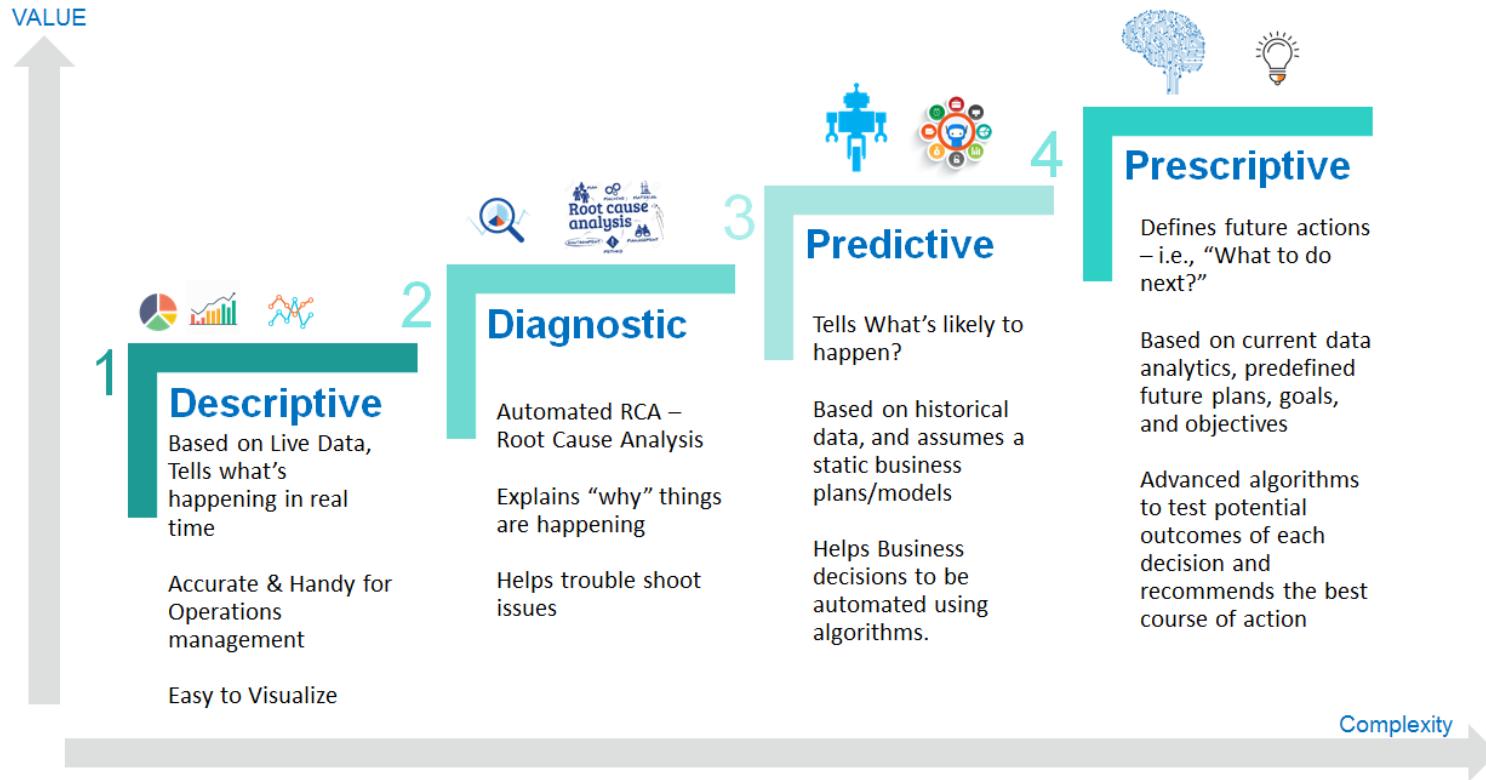
Định nghĩa về phân tích dữ liệu

- Phân tích dữ liệu là một quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định
- Cơ hội và ứng dụng trong điều hành lưới điện thông minh
 - Cải thiện Demand Response: Tìm kiếm tập mẫu khách hàng phù hợp, đấu giá điện theo thời gian, vvv.
 - Phân tích dự báo về giá, phân loại hồ sơ khách hàng
 - Tìm kiếm ngoại lai, bất thường
 - Tối ưu lập kế hoạch – vận hành
 - Giám sát và dự báo các vấn đề hỏng hóc, sự cố, bảo trì phát sinh

4 kiểu phân tích dữ liệu

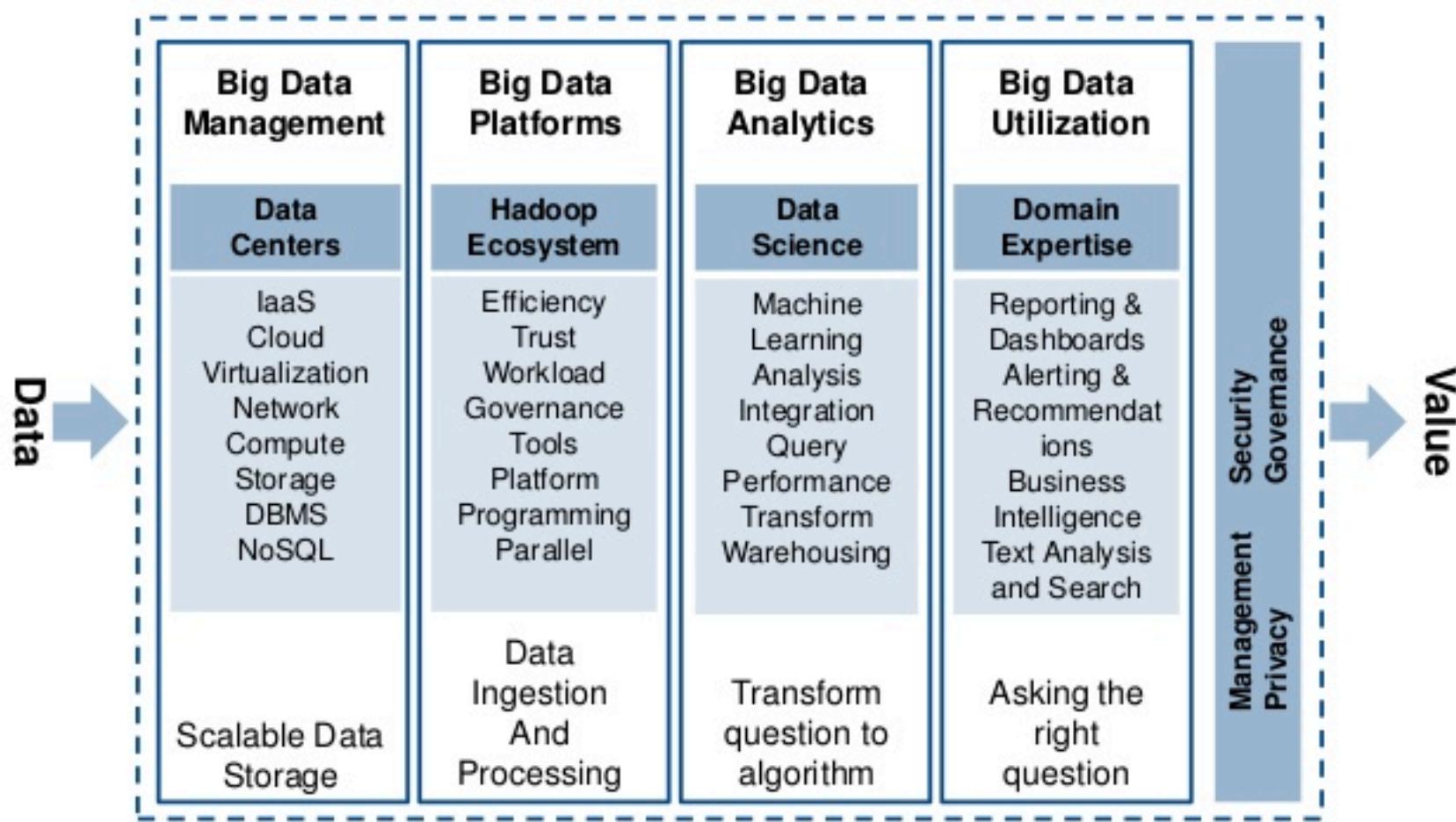


4 Types of Data Analytics



© Arun Kottoli

Các tầng công nghệ cho dữ liệu lớn



Quản lý dữ liệu phải khả mở

- Scalability
 - Khả năng quản lý lượng dữ liệu lớn không ngừng tăng lên theo thời gian.
- Accessibility
 - Cho phép đọc ghi I/O dữ liệu hiệu quả.
- Transparency
 - Truy cập dữ liệu dễ dàng, vị trí lưu trữ dữ liệu trên hệ thống là trong suốt với người dùng cuối.
- Availability
 - Khả năng chống chịu lỗi, khi tăng số lượng người dùng, khi hỏng hóc.

Xử lý và tích hợp dữ liệu phải khả mở

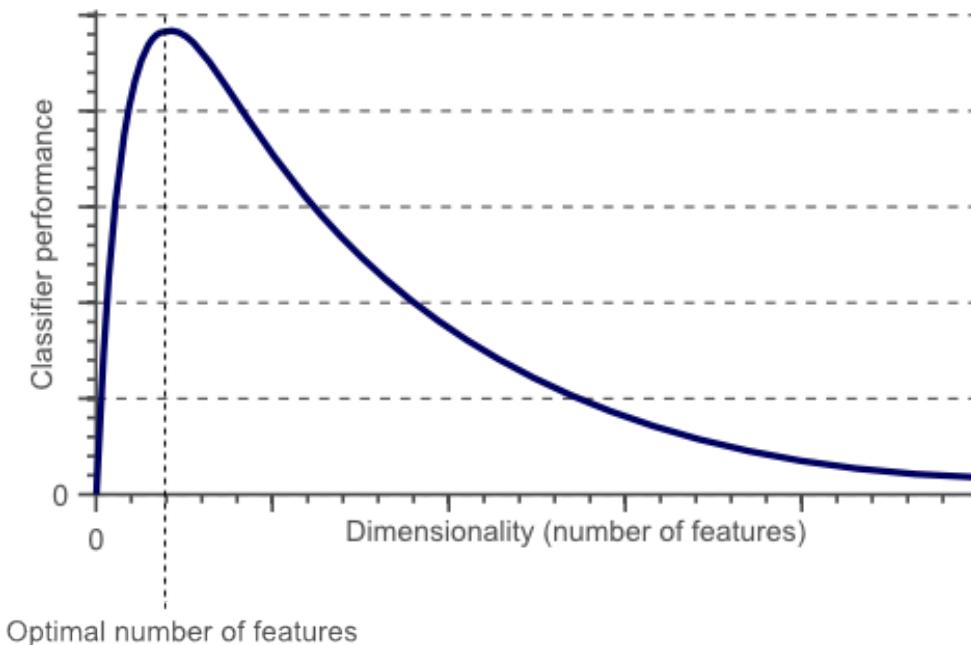
- Tích hợp dữ liệu
 - Dữ liệu có định dạng khác nhau
 - Dữ liệu tồn tại ở các mô hình và lược đồ dữ liệu khác nhau
 - Các vấn đề liên quan đến an toàn an ninh thông tin, quyền riêng tư
- Xử lý dữ liệu
 - Xử lý khối lượng dữ liệu rất lớn
 - Xử lý luồng dữ liệu lớn
 - Xử lý dữ liệu song song, phân tán truyền thống (OpenMP, MPI)
 - Phức tạp, khó học
 - Khả năng khả mở có giới hạn
 - Cơ chế chịu lỗi kém
 - Chi phí hạ tầng đắt đỏ
 - Kiến trúc xử lý dữ liệu luồng dữ liệu lớn
 - Spark mini-batch
 - Apache Flink

Các giải thuật phân tích dữ liệu khả mở

- Làm nhỏ lại dữ liệu cho phù hợp với các giải thuật truyền thống
 - Eg. Sub-sampling
 - Eg. Principal component analysis
 - Eg. Feature extraction and feature selection
- Song song hoá các giải thuật học máy
 - Eg. k-nn classification based on MapReduce
 - Eg. scaling-up support vector machines (SVM) by a divide and-conquer approach

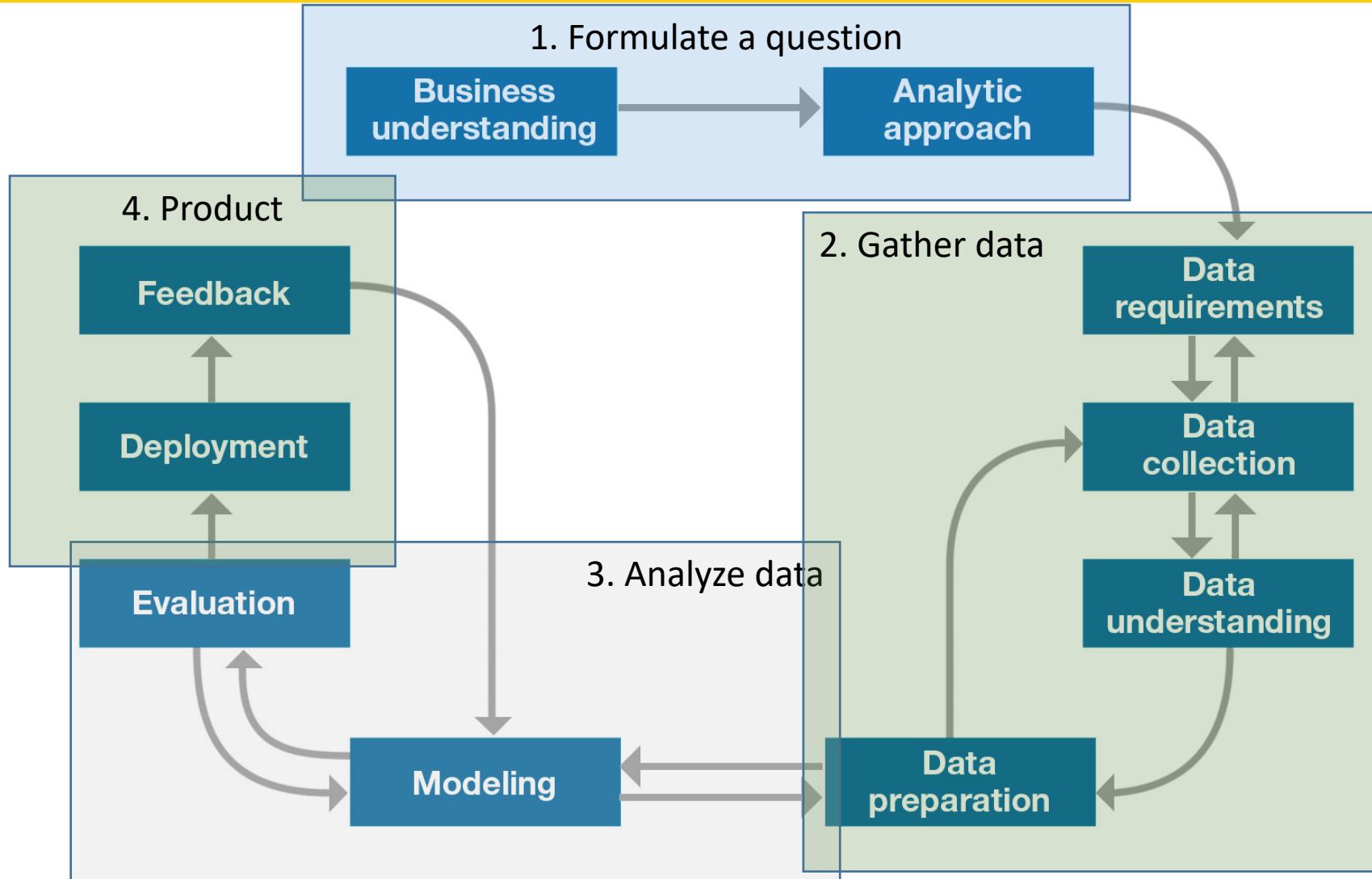
Eg. Sự bùng nổ số chiều trong dữ liệu (Curse of dimensionality)

- Số lượng mẫu cần cho mô hình học tăng lên khi số chiều dữ liệu tăng
- Trong thực tiễn: Số lượng mẫu để học thường cố định
=> Độ chính xác của mô hình giảm khi tăng số chiều trong dữ liệu học



Quy trình làm khoa học dữ liệu

Quy trình làm khoa học dữ liệu



1. Hình thành câu hỏi

Đặt câu hỏi đúng

Hiểu nghiệp vụ, bài toán cần giải quyết

- Cần sự hợp tác chặt chẽ giữa các bên liên quan
 - Chuyên gia về lĩnh vực, nghiệp vụ của bài toán: để đặt các mục tiêu đúng
 - Chuyên gia quản trị dữ liệu: để xây dựng và tổ chức dữ liệu phù hợp
 - Chuyên gia mô hình hóa: để thiết kế và đánh giá các mô hình dự báo

Xác định bài toán mục tiêu

- Bài toán thực tiễn cần giải quyết bởi mô hình dự báo là gì?
- Bài toán thực tiễn này định lượng thế nào, bằng các con số, dữ liệu nào
- Các phương pháp mô hình hóa nào có thể dùng được
- Đánh giá chất lượng mô hình như thế nào để phản ánh đúng bài toán mục tiêu
- Các bước đưa mô hình dự báo vào vận hành trong thực tiễn

2. Thu thập dữ liệu

Thu thập dữ liệu

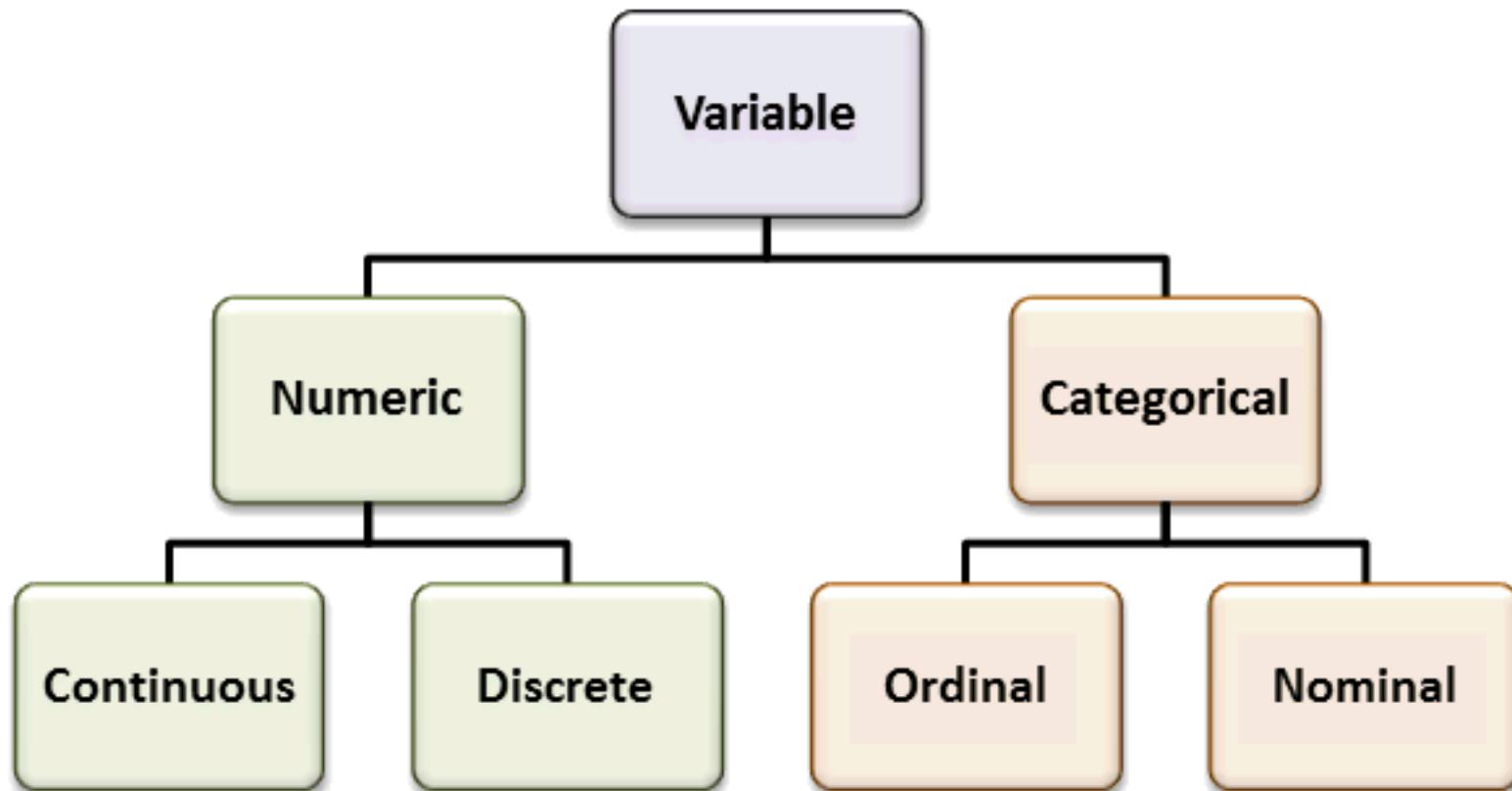
- Xác định dữ liệu, các thuộc tính của dữ liệu (thường được biết dưới các tên như tên biến, trường thuộc tính, đặc trưng của dữ liệu)
- Xác định biến mục tiêu
- Thu thập tập dữ liệu (tập hợp các mẫu dữ liệu sử dụng trong bài toán)

Quan sát và các biến

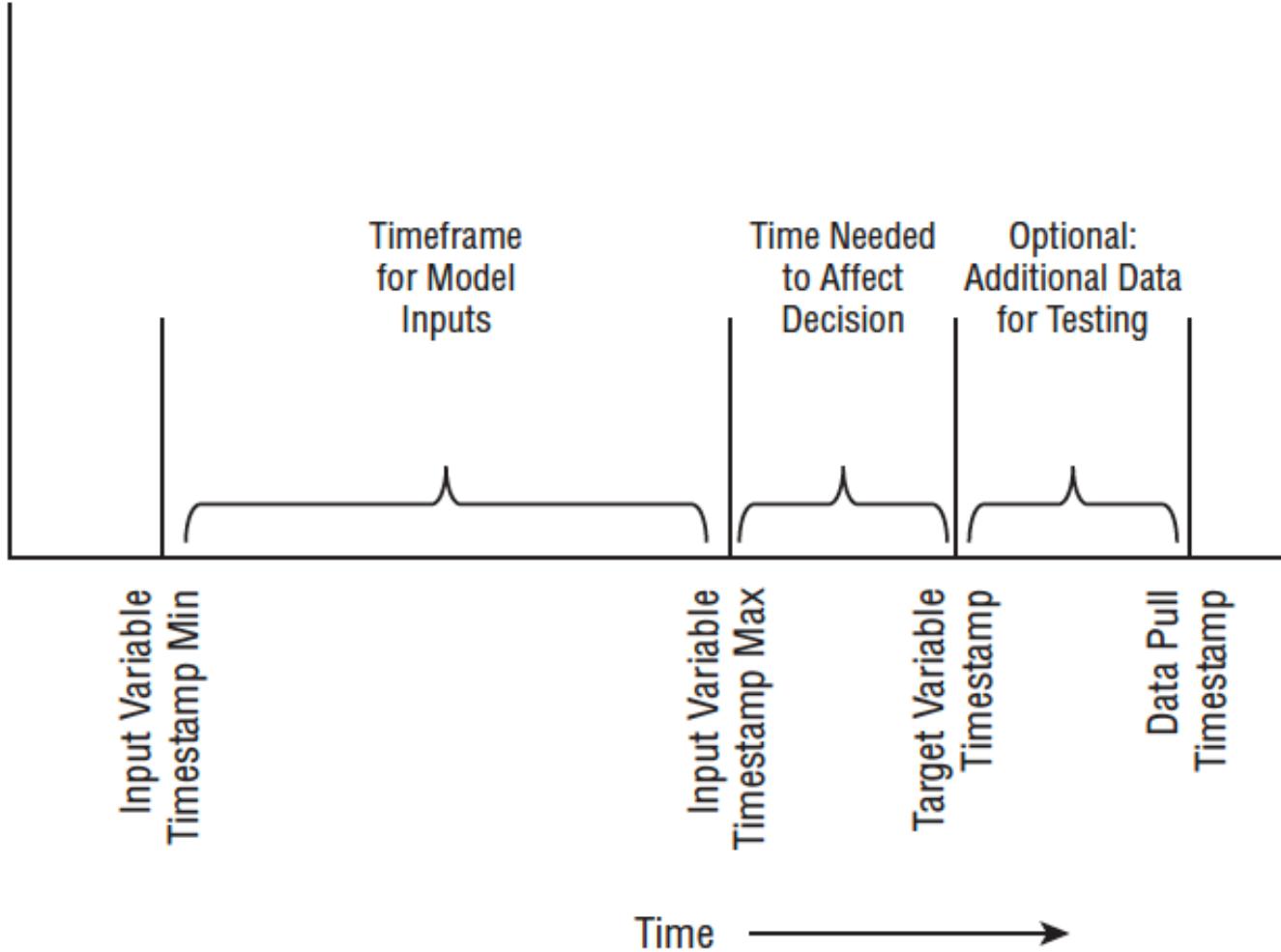
- Dữ liệu là một tập hợp của các quan sát
 - Một thuộc tính, hay 1 biến là 1 tập hợp các giá trị mô tả một khía cạnh nào đó trên tất cả các quan sát

HR Information		Contact	
Position	Salary	Office	Extn.
Accountant	\$162,700	Tokyo	5407
Chief Executive Officer (CEO)	\$1,200,000	London	5797
Junior Technical Author	\$86,000	San Francisco	1562
Software Engineer	\$132,000	London	2558

Các kiểu biến



Yếu tố thời gian trong mối tương quan với biến mục tiêu



2.1 Hiểu dữ liệu

Hiểu dữ liệu

- Là bước quan trọng trước tiên, trước khi xây dựng các mô hình dự báo
- Lý do
 - Xem xét các tính chất, khuôn mẫu có thể có trong dữ liệu
 - Xem xét các lỗi, ngoại lai trong dữ liệu
 - Xem xét các giả định thống kê có thể có trong dữ liệu
 - Đưa ra các giả định phù hợp về dữ liệu
- Nếu không hiểu dữ liệu, sẽ dẫn tới mất thời gian để xây dựng mô hình học, chất lượng mô hình không tốt



Phân tích thăm dò dữ liệu - Exploratory data analysis (EDA)

- EDA không phải là một tập hợp các công cụ, mà là một triết lý, phương pháp cần thiết phải tuân theo
 - Cho phép lựa chọn công cụ phù hợp trong tiền xử lý dữ liệu
 - Cho phép sử dụng kinh nghiệm trong nhìn nhận các khuôn mẫu trong dữ liệu
- Mỗi quan tâm của EDA là dữ liệu, cấu trúc của dữ liệu, các ngoại lai, và lựa chọn mô hình hóa phù hợp với dữ liệu
- EDA sử dụng toàn bộ dữ liệu để xem xét, không bỏ qua bất kỳ thành phần dữ liệu nào
 - Các thông tin thống kê Summary statistics
 - Các biểu đồ trực quan hóa Visualization
 - Gom nhóm và xác định ngoại lai
 - Giảm chiều dữ liệu

Các câu hỏi EDA phổ biến

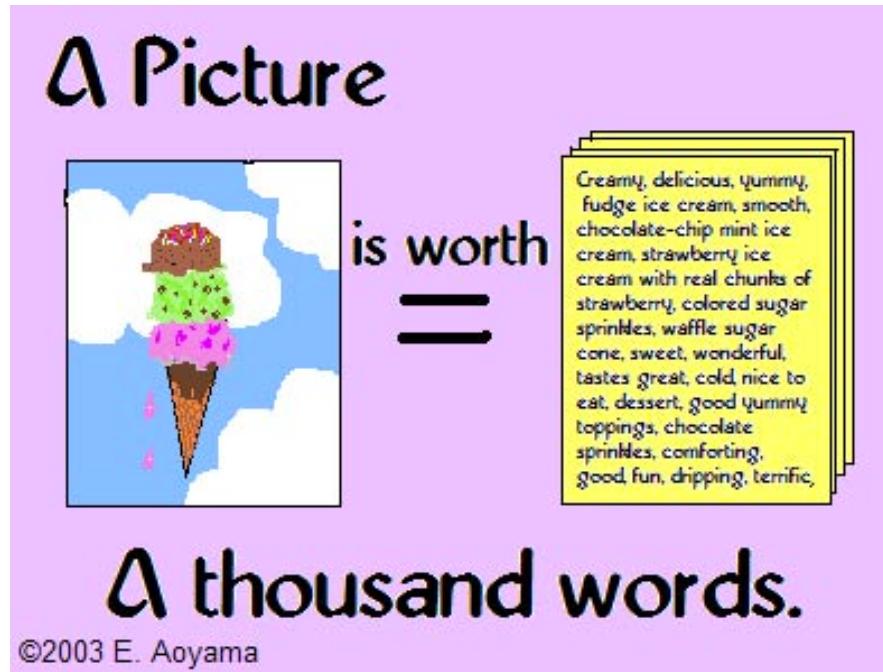
- Giá trị tiêu chuẩn của bộ dữ liệu là bao nhiêu?
- Độ không chắc chắn của giá trị tiêu chuẩn trong bộ dữ liệu?
- Dữ liệu có tuân theo phân bố xác suất nào không?
- Thuộc tính này của dữ liệu có quan trọng không?
- Thuộc tính nào của dữ liệu là quan trọng nhất?
- Các độ đo trong dữ liệu có nhất quán, nếu dữ liệu được tích hợp từ nhiều nguồn khác nhau?
- Biến mục tiêu có mối quan hệ như thế nào với các đặc trưng của dữ liệu?
- Có thể tách nhiễu khỏi dữ liệu hay không?
- Có cấu trúc nào trong dữ liệu đa biến đang xem xét hay không?
- Dữ liệu có ngoại lai không?

EDA là một tiến trình lắp

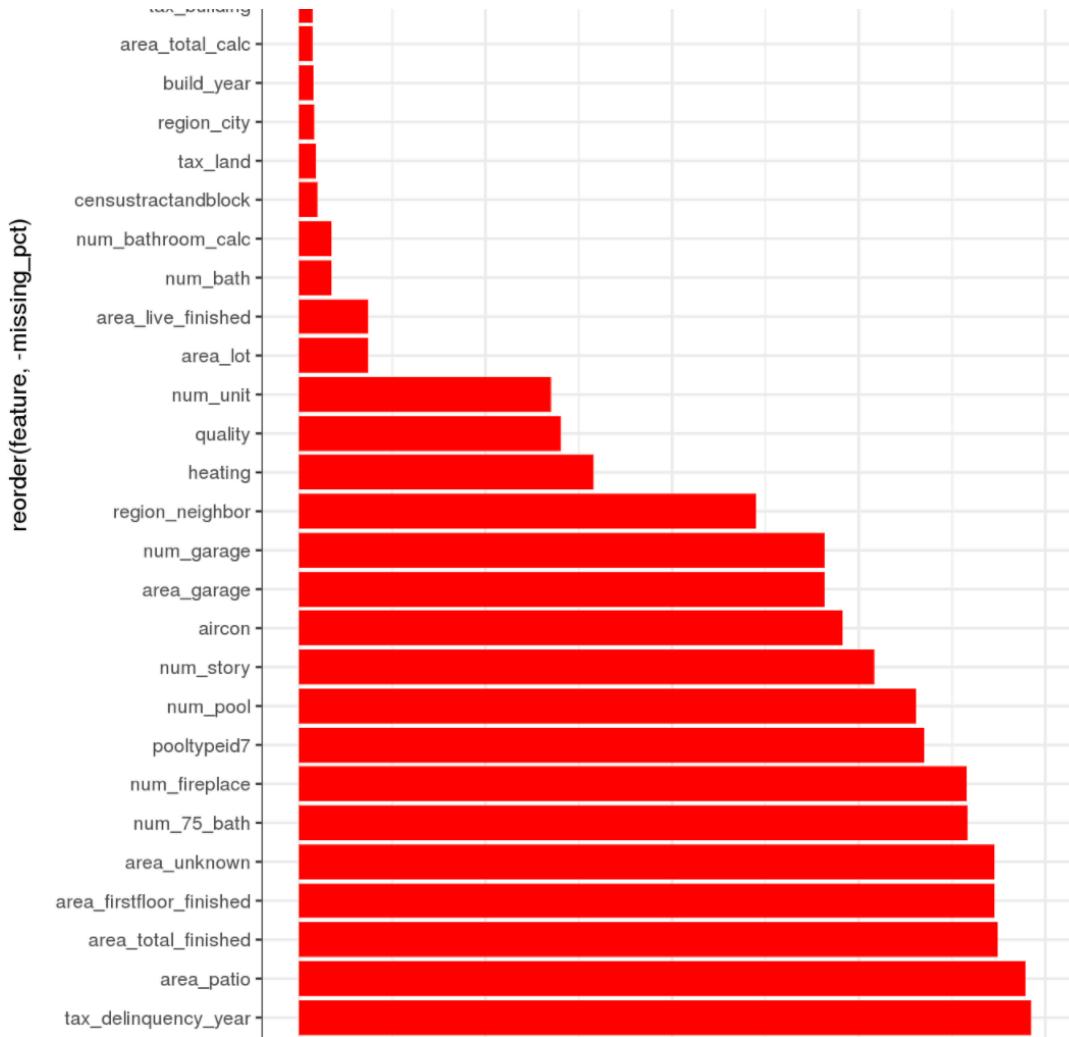
- Vòng lắp
 - Xác định và xét thứ tự ưu tiên cho các câu hỏi theo mức độ giảm dần của độ quan trọng
 - Đặt câu hỏi
 - Xây dựng các biểu đồ để trả lời câu hỏi
 - Xem xét các trả lời rút ra được và đặt câu hỏi mới
- Chiến lược EDA
 - Xem xét lần lượt từng biến, sau đó xem xét mối quan hệ giữa các biến
 - Bắt đầu bằng đồ thị, sau đó thêm các chỉ số thống kê
 - Lưu ý tới kiểu thuộc tính. Ví dụ kiểu catelogy với kiểu số

Các kỹ thuật cho EDA

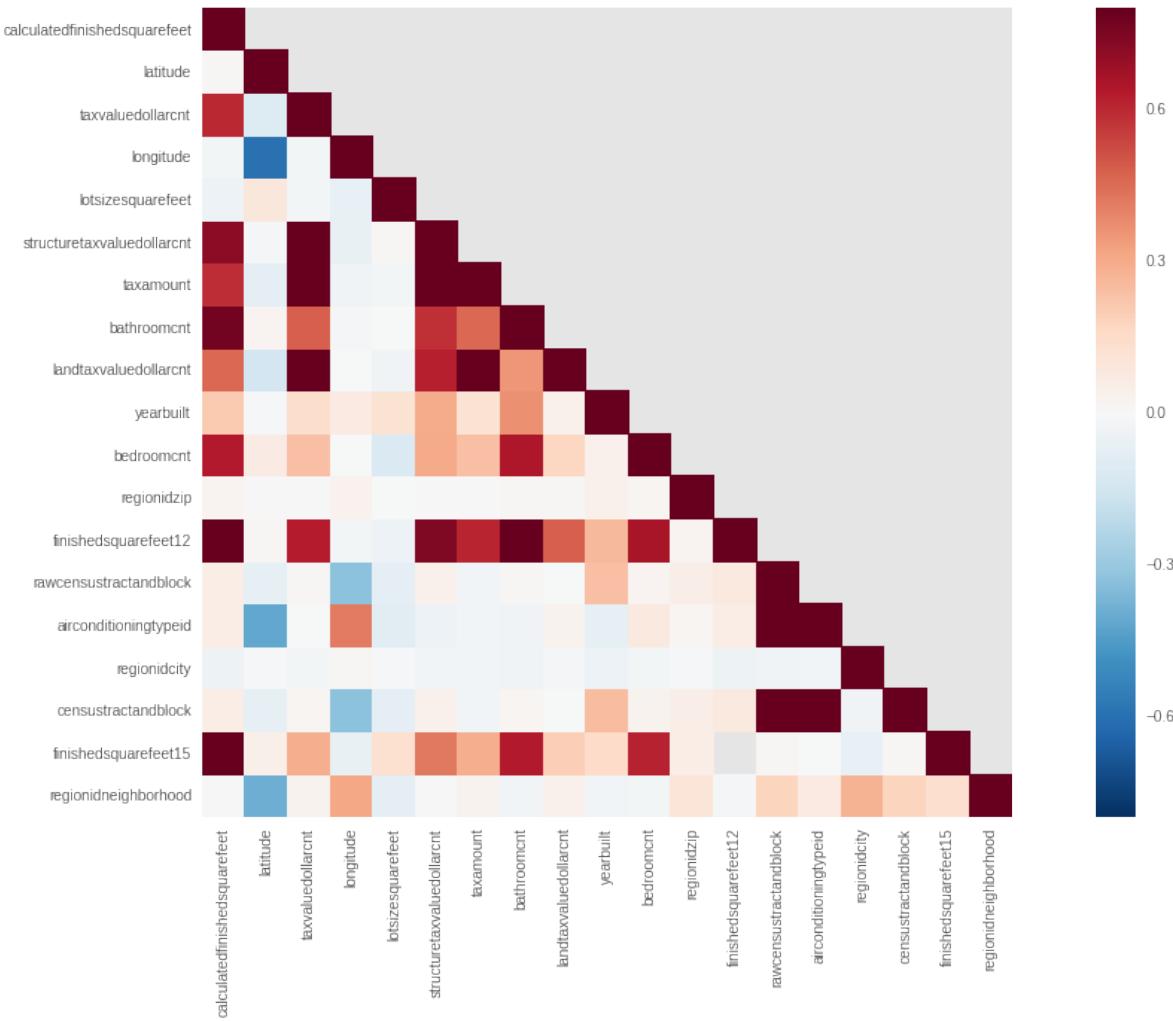
- Kỹ thuật định lượng sử dụng các thông số thống kê
- Kỹ thuật đồ thị
 - Scatter plots, character plots, box plots, histograms, probability plots, residual plots, and mean plots.



Ví dụ: Các thuộc tính thiếu giá trị



Ví dụ: Phân tích tương quan giữa các thuộc tính



2.2 Tiề̂n xū lý dữ liệ̂u

Tại sao cần tiền xử lý dữ liệu?

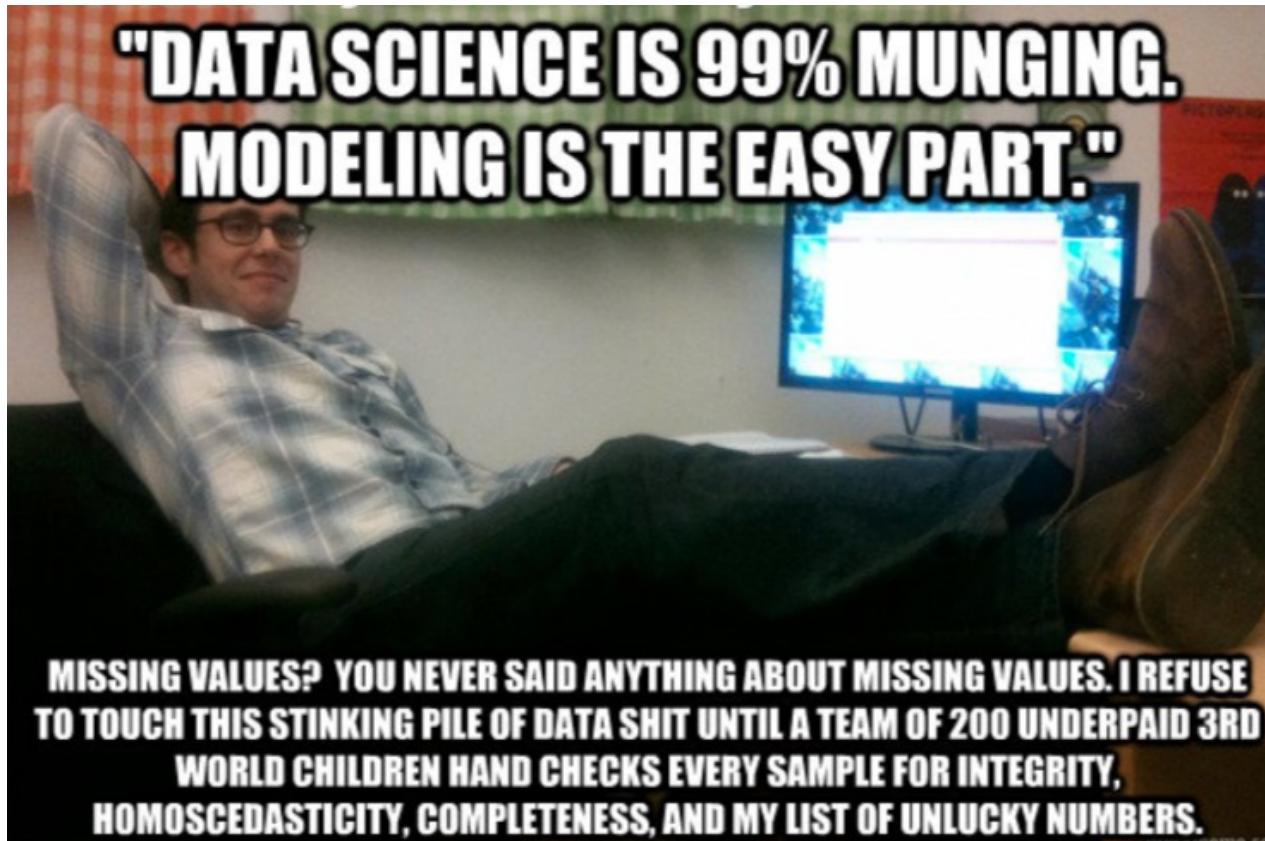
- Dữ liệu trong thế giới thực rất hỗn tạp về chất lượng
- Không đầy đủ (vd. name = "")
 - Thuộc tính thiếu giá trị, Các quan sát thiếu thuộc tính, hoặc chỉ có các giá trị kết tập của dữ liệu
 - Do sự khác nhau về lựa chọn giữa thời gian thu thập dữ liệu và thời gian phân tích dữ liệu
 - Lỗi phần cứng, phần mềm, do con người
- Có nhiễu (vd. salary ='-10k')
 - Chứa lỗi và các ngoại lai
 - Do thiết bị thu thập dữ liệu có nhiễu
 - Do con người nhập vào hệ thống
 - Do lỗi truyền dữ liệu
- Không nhất quán (vd. Age="20" Birthday="02/02/2000")
 - Tích hợp dữ liệu từ nhiều nguồn khác nhau
 - Vi phạm các ràng buộc về dữ liệu
- Các bản ghi (quan sát) trùng lặp cần phải loại bỏ

Ví dụ các vấn đề về chất lượng dữ liệu

	Representation	Contradictions	Ref. integrity																												
CUST	<table border="1"><thead><tr><th>CNr</th><th>Name</th><th>Birthday</th><th>Age</th><th>Sex</th><th>Phone</th><th>ZIP</th></tr></thead><tbody><tr><td>1234</td><td>Costa, Rui</td><td>18.2.80</td><td>37</td><td>m</td><td>9999999999</td><td>1000</td></tr><tr><td>1234</td><td>Ana Costa</td><td>32.2.70</td><td>37</td><td>m</td><td>965432123</td><td>55555</td></tr><tr><td>1235</td><td>Rui Costa</td><td>18.2.80</td><td>27</td><td>m</td><td>963124568</td><td>1000</td></tr></tbody></table>	CNr	Name	Birthday	Age	Sex	Phone	ZIP	1234	Costa, Rui	18.2.80	37	m	9999999999	1000	1234	Ana Costa	32.2.70	37	m	965432123	55555	1235	Rui Costa	18.2.80	27	m	963124568	1000		
CNr	Name	Birthday	Age	Sex	Phone	ZIP																									
1234	Costa, Rui	18.2.80	37	m	9999999999	1000																									
1234	Ana Costa	32.2.70	37	m	965432123	55555																									
1235	Rui Costa	18.2.80	27	m	963124568	1000																									
Uniqueness																															
ADDRESS	<table border="1"><thead><tr><th>ZIP</th><th>Place</th></tr></thead><tbody><tr><td>1000</td><td>Lisboa</td></tr><tr><td>1000</td><td>Lsboa</td></tr><tr><td>1024</td><td>Portugal</td></tr></tbody></table>	ZIP	Place	1000	Lisboa	1000	Lsboa	1024	Portugal	<p>Missing values</p> <p>Incorrect values</p>	<p>Duplicates</p>																				
ZIP	Place																														
1000	Lisboa																														
1000	Lsboa																														
1024	Portugal																														

Tiền xử lý dữ liệu có chi phí cao

- Trích xuất, làm sạch, chuyển đổi dữ liệu chiếm chi phí cao trong xây dựng các kho dữ liệu



Dữ liệu không có chất lượng, đầu ra không thể tốt

- Dữ liệu thiếu, trùng lặp, sai dẫn tới mô hình học sai, đưa đầu ra không đúng



Các đặc trưng của dữ liệu có chất lượng

- “Even though quality cannot be defined, you know what it is.”
Robert Pirsig



Chất lượng dữ liệu: mức giá trị đơn

- Thiếu giá trị
 - Ex:birthdate=""
- Sai cú pháp
 - Ex:zipcode=27655-175;syntactical rule:xxxx-xxx
- Sai chính tả
 - Ex:city='Lsboa', instead of 'Lisbon'
- Sai tập xác định
 - Ex:age=240;age:{0,120}

Mức tập giá trị và bản ghi

- Mức tập giá trị
 - Từ đồng nghĩa
 - Ex: emprego = ‘futebolista’; emprego = ‘jogador futebol’
 - Từ đồng âm khác nghĩa
 - Ex: Tác giả cùng tên
 - Vi phạm ràng buộc duy nhất:
 - Ex: khách hàng có cùng định danh ID
 - Vi phạm ràng buộc toàn vẹn
 - Ex: Tổng % > 100 %
- Mức bản ghi
 - Vi phạm ràng buộc toàn vẹn
 - Ex: Tổng giá khác giá bán + thuế VAT

Mức quan hệ

- Biểu diễn dữ liệu khác nhau:
 - Ex: name = ‘John Smith’; name = ‘Smith, John’
- Vi phạm phụ thuộc hàm
 - Ex: (2765-175, ‘Estoril’) and (2765-175, ‘Oeiras’)
- Trùng lặp tương đối
 - Ex: (1, André Fialho, 12634268) và (2, André Pereira Fialho, 12634268)!
- Vi phạm ràng buộc toàn vẹn
 - Ex: tổng lương của nhân viên > quỹ lương

Mức đa quan hệ

- Biểu diễn dữ liệu khác nhau trên các bảng
 - Ex: khác đơn vị đo
- Từ đồng nghĩa
- Từ đồng âm khác nghĩa
- Thang chia khác nhau:
 - Ex: age:{0-30,31-60,>60};age:{0-25,26-40, 40-65, >65}
- Ràng buộc tham chiếu
- Trùng lắp tương đối
- Ràng buộc toàn vẹn

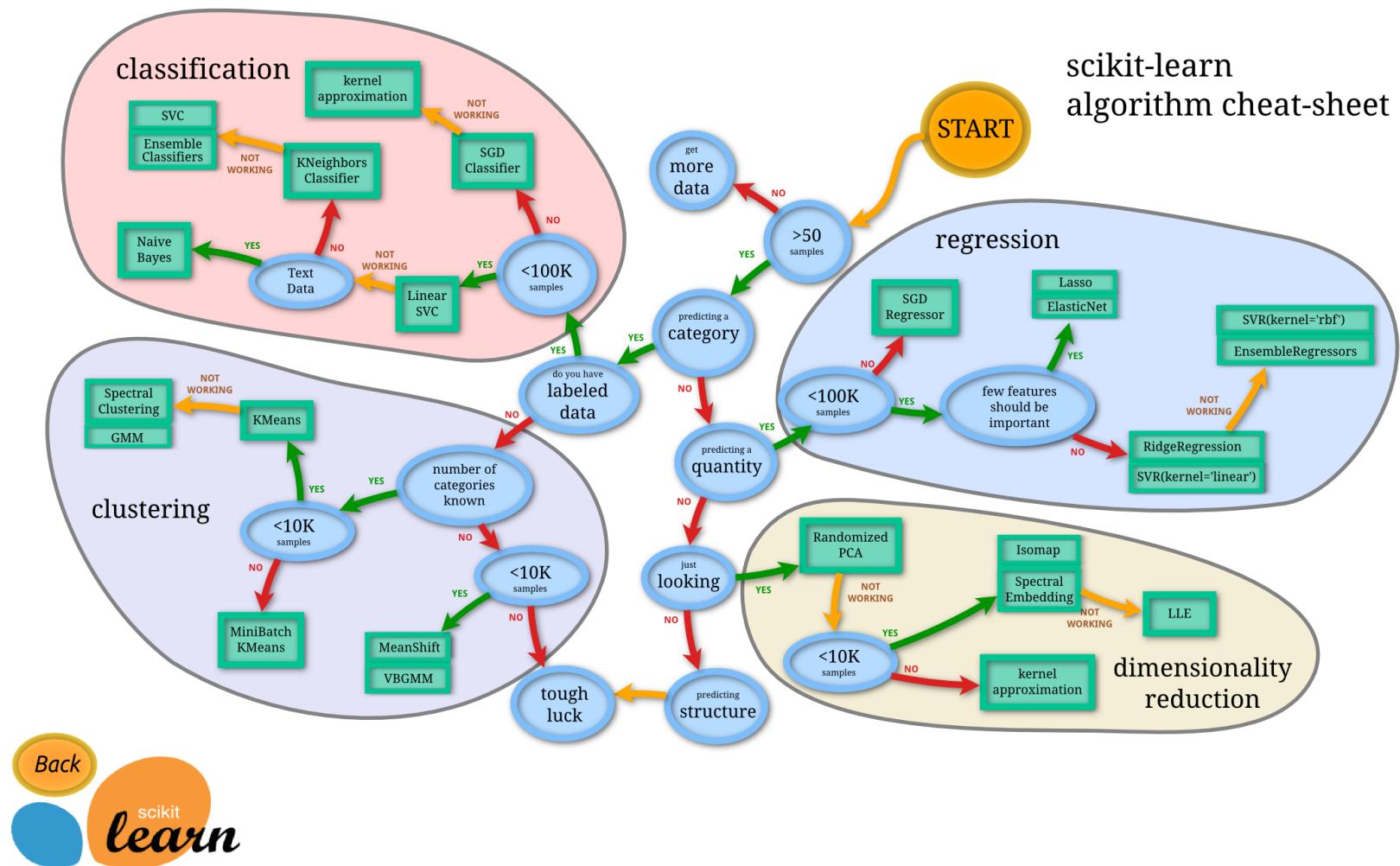
Các tác vụ chính trong tiền xử lý dữ liệu

- Làm sạch dữ liệu
 - Điền các giá trị còn thiếu, làm mịn dữ liệu nhiễu, xác định và loại bỏ các ngoại lai, giải quyết các trường hợp thiếu nhất quán
- Tích hợp dữ liệu từ nhiều nguồn
- Chuyển đổi dữ liệu
 - Chuẩn hoá
 - Kết tập dữ liệu
- Làm giảm dữ liệu
 - Biểu diễn dữ liệu nhỏ hơn về kích thước nhưng vẫn giữ được các đặc trưng của dữ liệu ban đầu
 - Rời rạc hoá dữ liệu
 - Kết tập, giảm chiều, nén, khai quát hoá dữ liệu

3. Phân tích dữ liệu

Predictive modeling

Các giải thuật phân tích dữ liệu



Đánh giá hiệu năng hệ thống học máy

- Đánh giá lý thuyết (theoretical evaluation): nghiên cứu các khía cạnh lý thuyết của một hệ thống mà có thể chứng minh được.
 - Tốc độ học, thời gian học,
 - Bao nhiêu ví dụ học là đủ?
 - Độ chính xác trung bình của hệ thống,
 - Khả năng chống nhiễu,...
- **Đánh giá thực nghiệm (experimental evaluation):** quan sát hệ thống làm việc trong thực tế, sử dụng một hoặc nhiều tập dữ liệu và các tiêu chí đánh giá. Tổng hợp đánh giá từ các quan sát đó.
 - Thường được áp dụng trong thực tiễn

Bài toán đánh giá

- Bài toán đánh giá (model assessment): cần đánh giá hiệu năng của phương pháp (model) học máy A, chỉ dựa trên bộ dữ liệu đã quan sát D.
- Việc đánh giá hiệu năng của hệ thống
 - Thực hiện một cách tự động, sử dụng một tập dữ liệu.
 - Không cần sự tham gia (can thiệp) của người dùng.
- Chiến lược đánh giá (evaluation strategies)
 - Làm sao có được một đánh giá đáng tin cậy về hiệu năng của hệ thống?
- Các tiêu chí đánh giá (evaluation metrics)
 - Làm sao để đo hiệu năng của hệ thống?

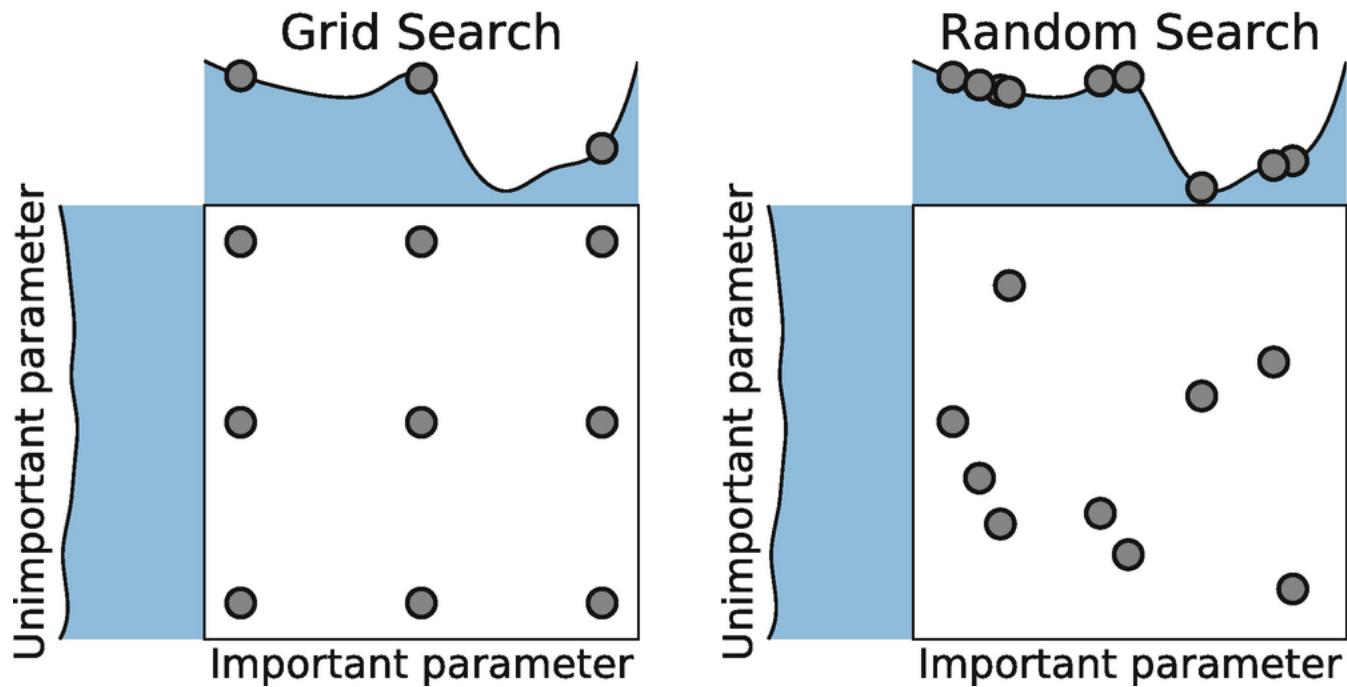
Chiến lược đánh giá

- Hold-out (chia đôi)
- Stratified sampling (lấy mẫu phân tầng)
- Repeated hold-out (chi đôi nhiều lần)
- Cross-validation (đánh giá chéo)
- Bootstrap sampling

Đánh giá và lựa chọn mô hình

- Cho trước tập quan sát D , ta cần lựa chọn tham số λ (model selection) cho phương pháp học A và đánh giá (assessment) chất lượng tổng thể của A .
 - Chọn tập hữu hạn S mà chứa các giá trị tiềm năng cho λ .
 - Chọn độ đo P để đánh giá hiệu năng.
 - Chia tập D thành 3 tập rời nhau: D_{train} , $T_{\text{validation}}$, và T_{test}
 - VỚI MỌI GIÁ TRỊ $\lambda \in S$:
 - Học A từ tập học D_{train} với tham số đầu vào λ . Đo hiệu năng trên tập $T_{\text{validation}}$ và thu được P_λ
 - Chọn λ^* mà có P_λ tốt nhất.
 - Huấn luyện A trên tập $D_{\text{train}} \cup T_{\text{validation}}$, với tham số đầu vào λ^* .
 - Đo hiệu năng của hệ thống trên tập T_{test} .
- Có thể thay Hold-out bằng kỹ thuật khác (cross-validation)

Grid search và random search

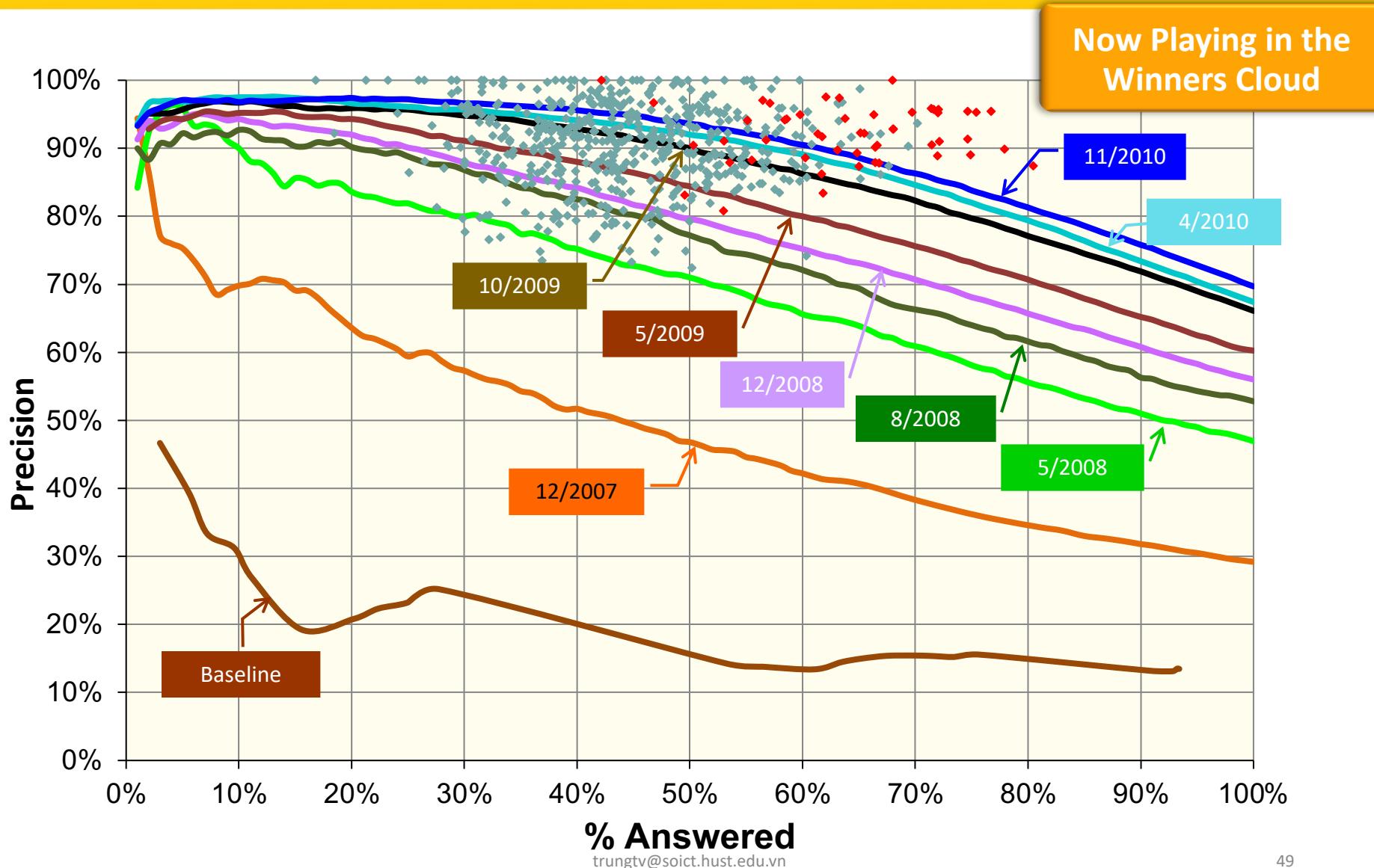


Các tiêu chí đánh giá

- Tính chính xác (Accuracy)
 - Mức độ dự đoán (phân lớp) chính xác của hệ thống (đã được huấn luyện) đối với các ví dụ kiểm chứng (test instances)
- Tính hiệu quả (Efficiency)
 - Chi phí về thời gian và tài nguyên (bộ nhớ) cần thiết cho việc huấn luyện và kiểm thử hệ thống
- Khả năng xử lý nhiễu (Robustness)
 - Khả năng xử lý (chịu được) của hệ thống đối với các ví dụ nhiễu (lỗi) hoặc thiếu giá trị
- Khả năng mở rộng (Scalability)
 - Hiệu năng của hệ thống (vd: tốc độ học/phân loại) thay đổi như thế nào đối với kích thước của tập dữ liệu
- Khả năng diễn giải (Interpretability)
 - Mức độ dễ hiểu (đối với người sử dụng) của các kết quả và hoạt động của hệ thống
- Mức độ phức tạp (Complexity)
 - Mức độ phức tạp của mô hình hệ thống (hàm mục tiêu) học được

DeepQA: Incremental Progress in Precision and Confidence

6/2007-11/2010



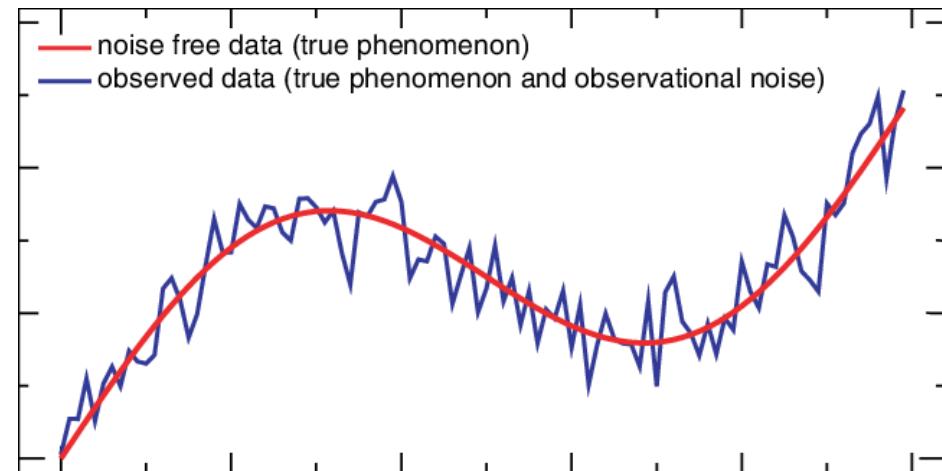
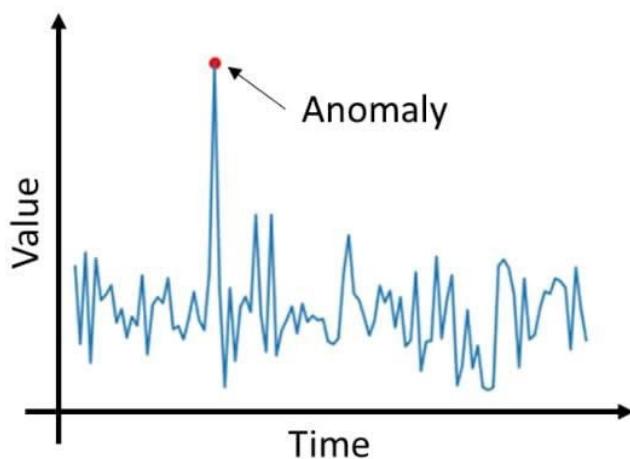
Ngoại lai và phân tích ngoại lai

Bất thường (Ngoại lai) là gì ?

- Bất thường (ngoại lai): khi đối tượng dữ liệu sai lệch một cách rõ rệt khỏi đối tượng bình thường như là được tạo sinh ra bằng một cơ chế khác biệt.
- Ví dụ:
 - Một giao dịch mua sắm bằng thẻ tín dụng không bình thường: bình thường mỗi tuần tiêu hết trong vòng 1tr, đột nhiên có 1 tuần sử dụng hết 50tr, như vậy giá trị này có thể là bất thường ;
 - Dữ liệu bình thường 99 điểm nhận giá trị trong khoảng 300-400, điểm thứ 100 lại nhận giá trị đột biến lên 2000. Đây có thể là bất thường.
 - Một hoạt động thể thao bất thường (doping),
 - ...

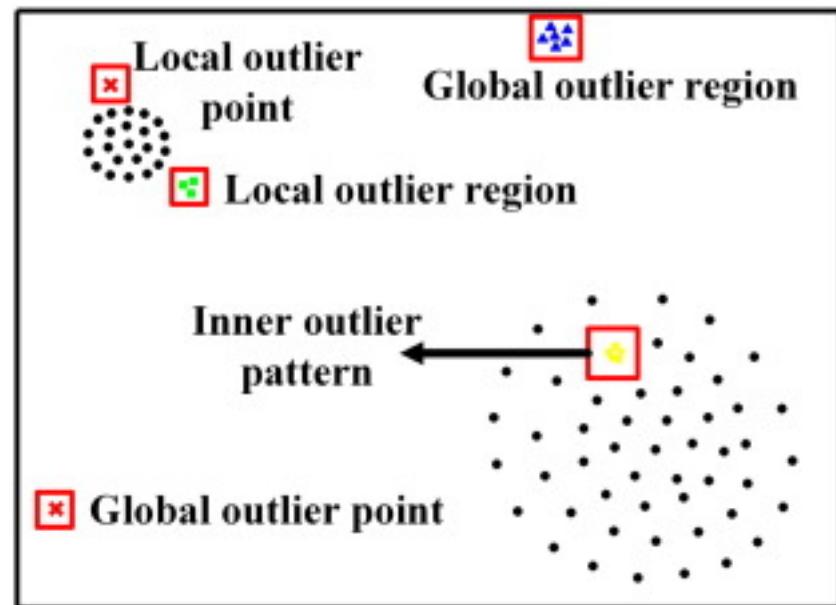
Bất thường và nhiễu

- Các bất thường khác với dữ liệu bị nhiễu.
 - Nhiễu có bản chất là các sai số ngẫu nhiên, hoặc những biến thiên của biến được đo đạc;
 - Nhiễu phải được giảm thiểu trước khi thực hiện phát hiện bất thường.



Các dạng ngoại lai

- Các dạng đối tượng ngoại lai:
 - Ngoại lai toàn cục, ngoại lai cục bộ
 - Ngoại lai theo ngũ cảnh,
 - Ngoại lai tập thể
- Ngoại lai toàn cục (point anomaly)
 - Đối tượng là O_g nếu đối tượng đó lệch đáng kể so với phần còn lại của tập dữ liệu. Khi đó các quan sát không phù hợp với toàn bộ tập dữ liệu
 - Ví dụ: trong phát hiện xâm nhập mạng
 - Vấn đề: Tìm độ đo thích hợp để đo độ sai lệch
- Ngoại lai cục bộ
 - giá trị đo được so với toàn bộ tập dữ liệu thì bình thường, nhưng trong lân cận cục bộ thì các quan sát nhận đột biến.



Các dạng ngoại lai [2]

- Ngoại lai theo ngũ cảnh (ngoại lai có điều kiện)
 - Đối tượng là O_c nếu đối tượng đó sai lệch đáng kể dựa trên một số ngũ cảnh xác định
 - Ví dụ: 41°C ở Hà Nội là bất thường? Điều này phụ thuộc vào mùa đông hay mùa hè !
 - Các thuộc tính của đối tượng dữ liệu được chia thành 2 nhóm:
 - Các thuộc tính ngũ cảnh: xác định các ngũ cảnh, thời gian, địa điểm, tình huống, ...
 - Các thuộc tính hành vi: các đặc trưng của đối tượng, dựa vào đó để xác định các ngoại lai.. Ví dụ – như trong ví dụ trước - nhiệt độ
 - Ngoại lai theo ngũ cảnh có thể coi là tổng quát hóa của ngoại lai cục bộ—tại đó mật độ lệch khỏi vùng cục bộ
- Nhóm ngoại lai (ngoại lai tập thể)
 - Một tập con của một số đối tượng dữ liệu cùng lệch đáng kể khỏi toàn bộ tập dữ liệu, thậm chí cả khi từng đối tượng dữ liệu có thể không phải là ngoại lai.

Các thách thức đối với phát hiện ngoại lai

- Mô hình hóa chính xác đối tượng bình thường và đối tượng ngoại lai:
 - Rất khó xác định hết các khả năng có thể có của hành vi bình thường của đối tượng trong ứng dụng
 - Ranh giới giữa đối tượng bình thường và ngoại lai thường khá mờ
- Phát hiện ngoại lai đặc trưng theo ứng dụng
 - Lựa chọn độ đo khoảng cách giữa các đối tượng và mô hình quan hệ giữa các đối tượng thường phụ thuộc vào ứng dụng
 - Ví dụ:
 - Trong y khoa: sự chênh lệch nhỏ có thể đã là ngoại lai.
 - Đối với phân tích thị trường chứng khoán, sự biến động lớn là ngoại lai.

Các thách thức đối với phát hiện ngoại lai [2]

- Xử lý nhiễu trong phát hiện ngoại lai
 - Nhiễu có thể làm biến dạng các đối tượng bình thường, làm mờ sự khác biệt giữa đối tượng bình thường và ngoại lai.
 - Nhiễu có thể làm ẩn các đối tượng ngoại lai và giảm hiệu quả phát hiện
- Khả năng diễn giải
 - Hiểu được vì sao các đối tượng đó là ngoại lai
 - Xác định được mức độ của đối tượng ngoại lai

Phát hiện ngoại lai

- Phân loại phương pháp phát hiện ngoại lai:
 - Dựa trên các mẫu đối tượng ngoại lai được gán nhãn có thể thu thập được:
 - Có giám sát,
 - Bán giám sát,
 - Không giám sát
 - Dựa trên các giả thiết về dữ liệu bình thường và đối tượng ngoại lai:
 - Phương pháp thống kê,
 - Phương pháp dựa trên độ lân cận,
 - Phương pháp phân cụm

Các phương pháp có giám sát

- Mô hình hóa bài toán phát hiện ngoại lai như bài toán phân loại
 - Các mẫu sẽ được khảo sát bởi các chuyên gia trong lĩnh vực để huấn luyện và kiểm chứng.
- Phương pháp huấn luyện bộ phân loại để phát hiện ngoại lai hiệu quả:
 - Mô hình hóa đối tượng bình thường; xác định những đối tượng không phù hợp với mô hình là ngoại lai. Hoặc
 - Mô hình hóa ngoại lai, và xác định những đối tượng không phù hợp mô hình là đối tượng bình thường.

Những thách thức trong phương pháp học có giám sát

- Sự mất cân bằng giữa các phân lớp: đối tượng ngoại lai thường hiếm gặp:
 - Gia tăng lớp đối tượng ngoại lai, tự tạo sinh ra một số ngoại lai
- Thu nhận nhiều đối tượng ngoại lai nhất có thể.
 - Tham số Recall quan trọng hơn Accuracy – tức là không phân loại nhầm đối tượng bình thường thành ngoại lai

$$Accuracy = \frac{TP + TN}{Pos + Neg}$$

$$Precision = \frac{TP}{TP + FP}$$

$$TPR = Recall = \frac{TP}{Pos}$$

$$TNR = \frac{TN}{Neg} = \frac{TN}{FP + TN}$$

$$FNR = \frac{FN}{Pos} = \frac{FN}{FN + TP}$$

Các phương pháp không giám sát

- Giả thiết
 - Các đối tượng bình thường được phân cụm thành các cụm, mỗi cụm có các đặc trưng riêng biệt.
 - Đối tượng ngoại lai là các đối tượng nằm xa tất cả các nhóm đối tượng bình thường
- Điểm yếu: Khó xác định các nhóm đối tượng ngoại lai một cách hiệu quả.
 - Đối tượng bình thường có thể không có đặc điểm chung rõ rệt,
 - Những ngoại lai tập thể có thể có sự giống nhau trong một vùng hẹp – tính co cụm.
- Ví dụ:
 - Trong phát hiện xâm nhập mạng hoặc phát hiện virus, các hành vi bình thường có thể rất đa dạng.
 - Phương pháp không giám sát có thể có chỉ số cảnh báo sai (false positive) cao, tuy vậy vẫn bị lọt một số đối tượng ngoại lai thực sự.

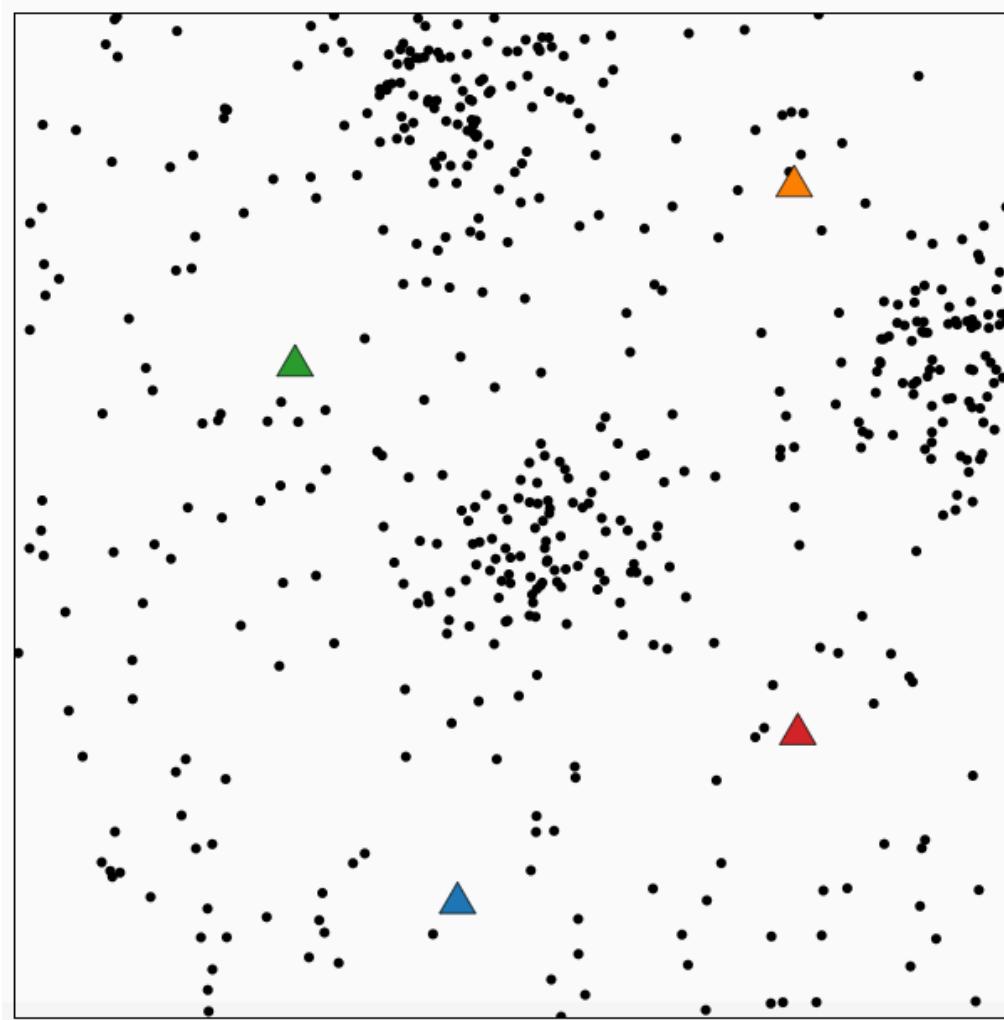
Phát hiện ngoại lai dựa trên phân cụm

- Đối tượng là ngoại lai nếu
 - Đối tượng đó không thuộc một cụm nào
 - Khoảng cách lớn giữa đối tượng và cụm gần nhất
 - Đối tượng đó thuộc một cụm nhỏ và tản mạn
- Các phương pháp phân cụm
 - Phân cụm theo khoảng cách
 - Phân cụm dựa trên mật độ

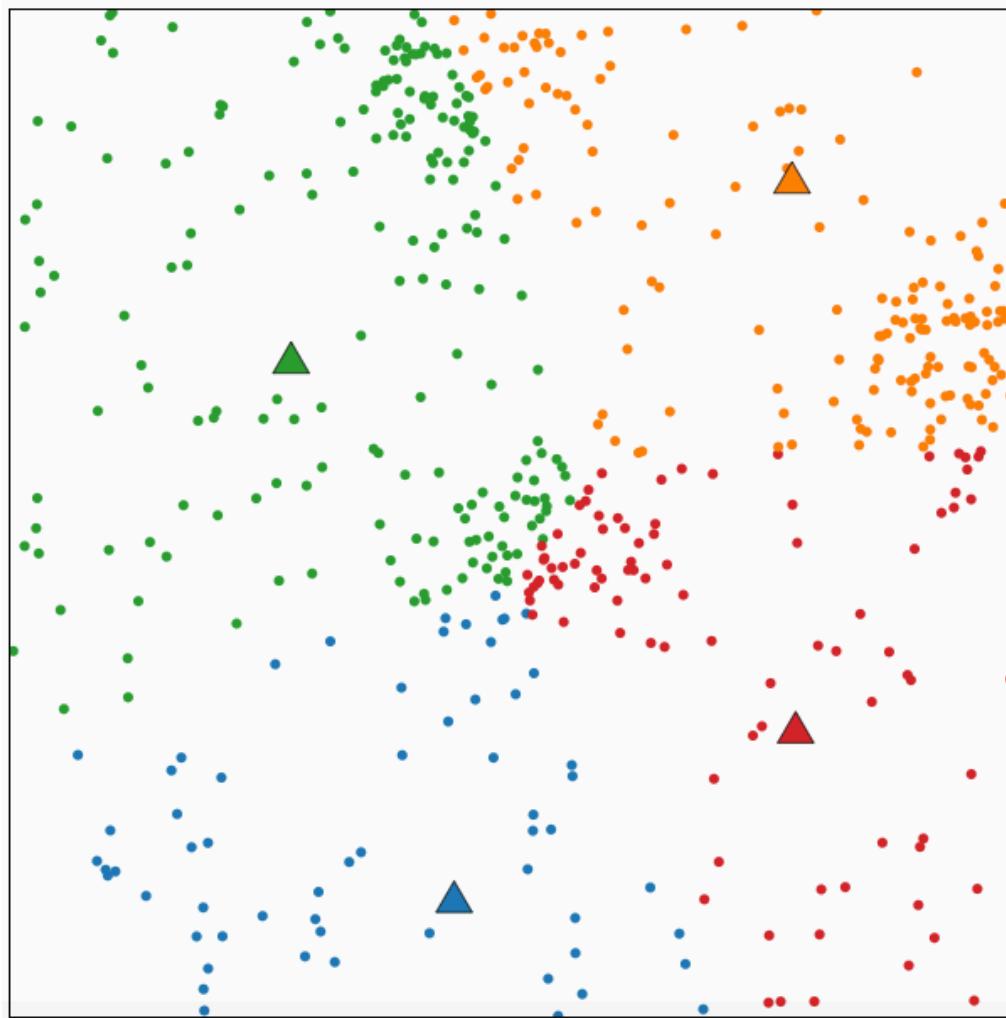
Phân cụm theo giải thuật K-mean

- Hướng tiếp cận
 - Cố định số lượng cụm
 - Tính toán “centroids” của mỗi cụm
 - Gán lại các đối tượng tới cụm theo centroid của cụm gần nhất
 - Tính toán lại các centroids
- Yêu cầu cần có
 - Cần xác định 1 hàm khoảng cách
 - Cần xác định số lượng cụm
 - Khởi tạo các centroids ban đầu

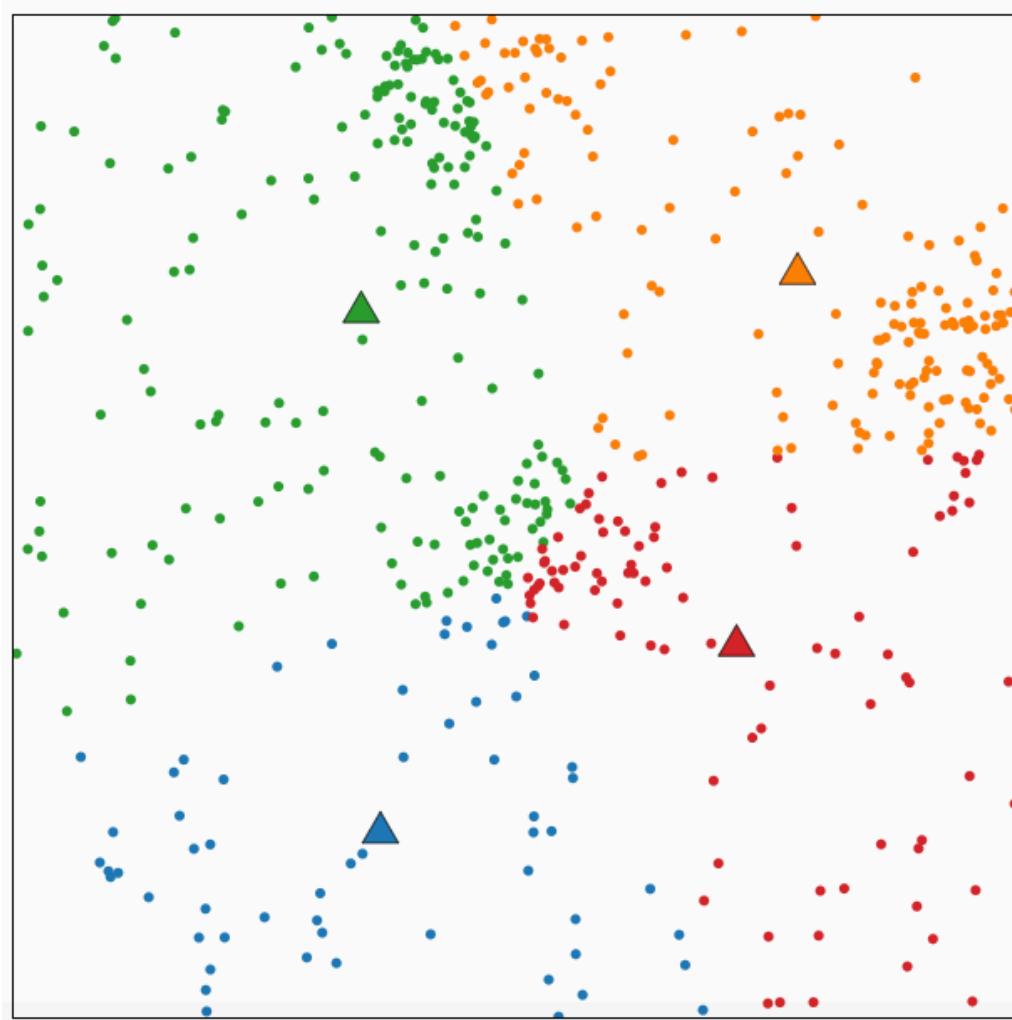
Ví dụ



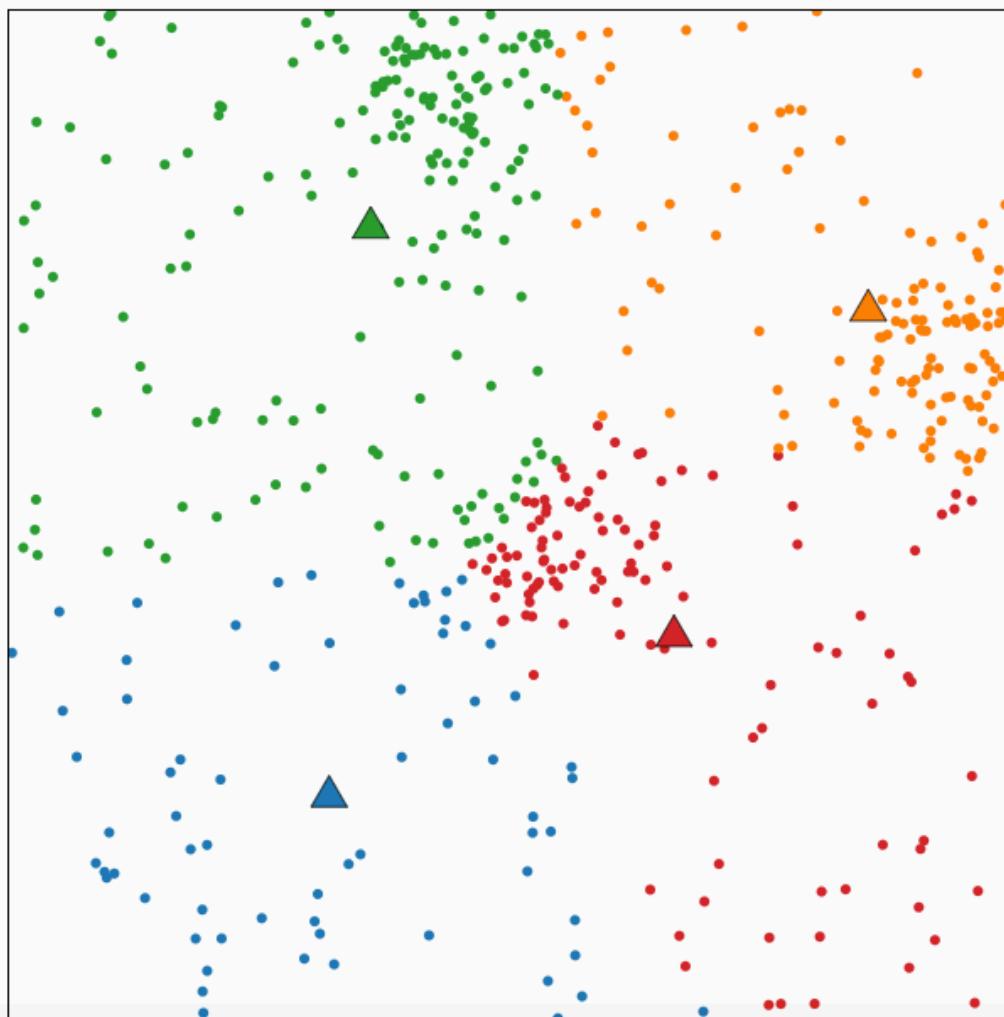
Ví dụ



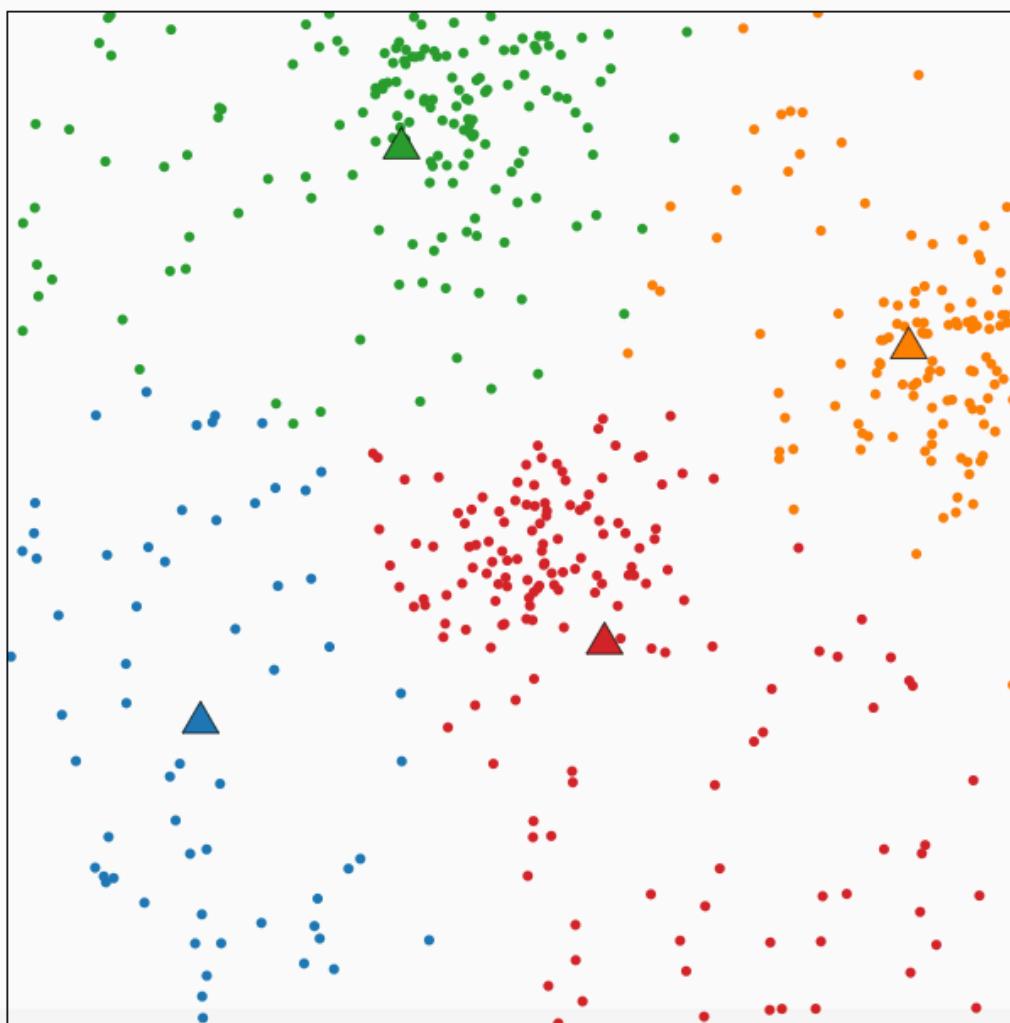
Ví dụ



Ví dụ



Ví dụ

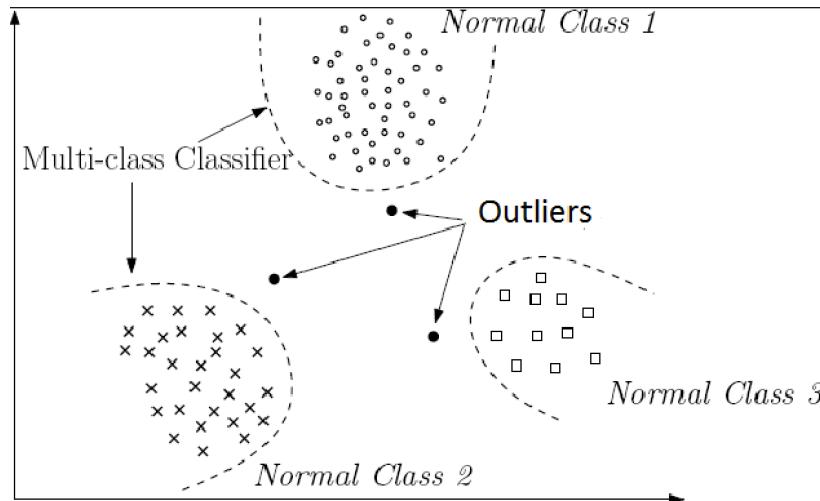


Ví dụ K-mean trong sklearn

```
>>> from sklearn.cluster import KMeans
>>> import numpy as np
>>> X = np.array([[1, 2], [1, 4], [1, 0],
...                 [10, 2], [10, 4], [10, 0]])
>>> kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
>>> kmeans.labels_
array([1, 1, 1, 0, 0, 0], dtype=int32)
>>> kmeans.predict([[0, 0], [12, 3]])
array([1, 0], dtype=int32)
>>> kmeans.cluster_centers_
array([[10.,  2.],
       [1.,  2.]])
```

Phương pháp dựa trên phân loại

- Ý tưởng: Huấn luyện mô hình phân loại có thể phân biệt dữ liệu “bình thường” và dữ liệu ngoại lai.
- Phương pháp vét cạn: Khảo sát tập dữ liệu chứa các mẫu được gán nhãn “bình thường” và các đối tượng khác dán nhãn “ngoại lai”
 - Tuy vậy, tập dữ liệu huấn luyện thường mất cân bằng mạnh: số lượng dữ liệu “bình thường” thông thường vượt xa số lượng các mẫu ngoại lai trong tập dữ liệu
 - Không thể phát hiện các bất thường, ngoại lai chưa biết.



Ví dụ. SVM trong sklearn

```
>>> import numpy as np
>>> from sklearn.pipeline import make_pipeline
>>> from sklearn.preprocessing import StandardScaler
>>> X = np.array([[-1, -1], [-2, -1], [1, 1], [2, 1]])
>>> y = np.array([1, 1, 2, 2])
>>> from sklearn.svm import SVC
>>> clf = make_pipeline(StandardScaler(), SVC(gamma='auto'))
>>> clf.fit(X, y)
Pipeline(steps=[('standardscaler', StandardScaler()),
                 ('svc', SVC(gamma='auto'))])
```

```
>>> print(clf.predict([[0.8, -1]]))
[1]
```

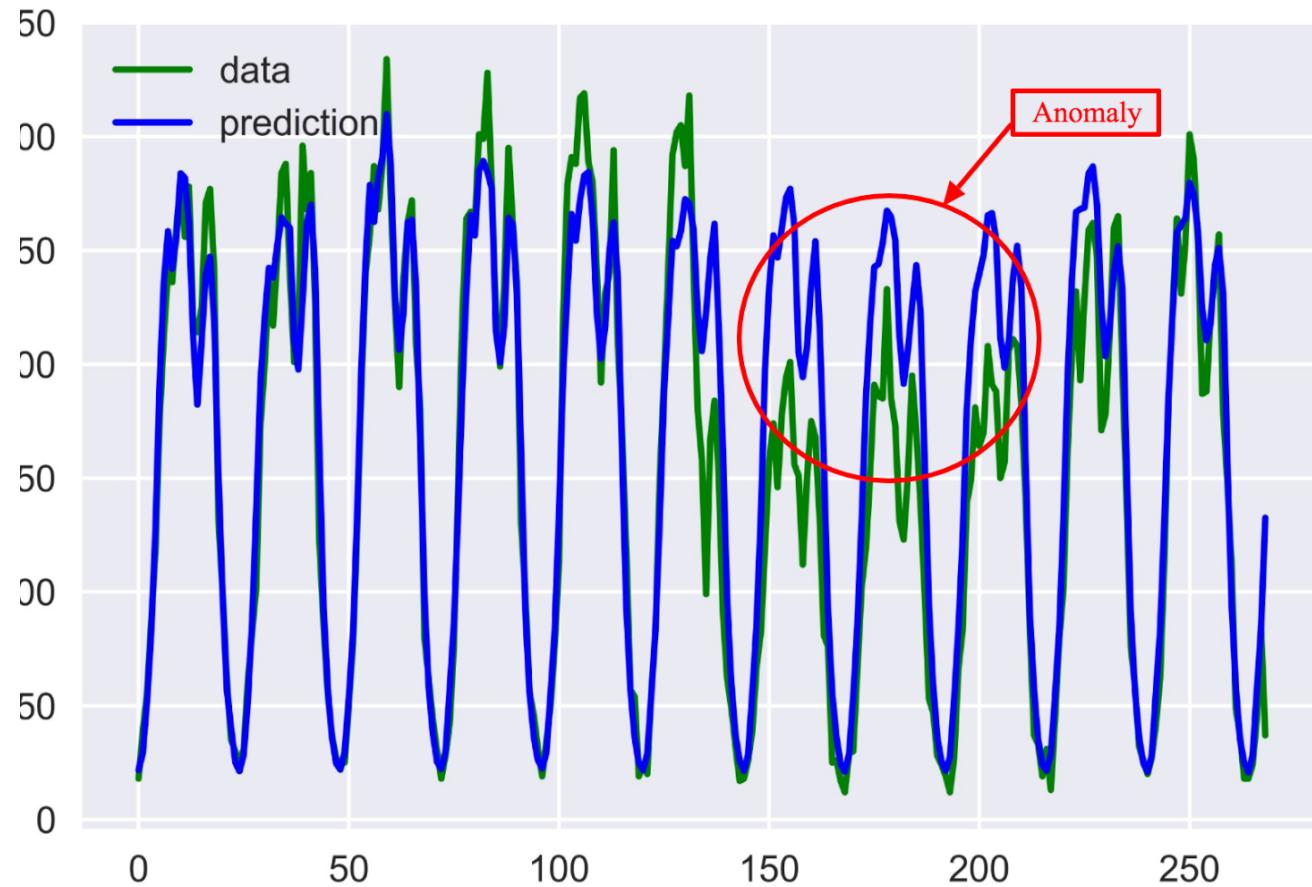
Mô hình một phân lớp

- Bộ phân loại được xây dựng chỉ để mô tả phân lớp bình thường.
 - Huấn luyện và học ranh giới quyết định của phân lớp “bình thường” bằng các bộ phân loại.
 - Bất kỳ đối tượng nào không thuộc phân lớp “bình thường” (không nằm trong vùng quyết định) đều được coi là đối tượng ngoại lai.
- Ưu điểm: có khả năng phát hiện ngoại lai mới có thể chưa xuất hiện gần các ngoại lai đã biết trong tập huấn luyện
- Mở rộng: đối tượng bình thường có thể nằm trong nhiều phân lớp.

```
>>> from sklearn.svm import OneClassSVM
>>> X = [[0], [0.44], [0.45], [0.46], [1]]
>>> clf = OneClassSVM(gamma='auto').fit(X)
>>> clf.predict(X)
array([-1,  1,  1,  1, -1])
>>> clf.score_samples(X)
array([1.7798..., 2.0547..., 2.0556..., 2.0561..., 1.7332...])
```

Mô hình hoá chuỗi thời gian

- <https://www.datacamp.com/community/tutorials/lstm-python-stock-market>





TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thank you for your attention!
Q&A

