Buổi 1: Tổng quan về máy học

(Tài liệu nội bộ)

Tháng 3 năm 2020

Nội dung trình bày

1. Giới thiệu về máy học
2. Kiểm thử & Thẩm định
3. Phân loại máy học
4. Thách thức chính của máy học

Nội dung trình bày

1. Giới thiệu về máy học

Máy học là gì?

- Arthur Samuel, 1959:
  - "Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."
- Tom Mitchell, 1997:
  - "A computer program is said to learn from experience E, with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Bài tập

- Bạn hãy cho ví dụ về một tác vụ/bài toán có thể giải quyết bằng máy học.

Các dạng bài toán máy học

- **Phân loại** (Classification)
  - Dự đoán có phải thư rác hay không.
- **Hồi quy** (Regression)
  - Dự đoán giá nhà.
- **Xếp hạng** (Ranking)
  - Xếp hạng các link kết quả tìm kiếm Google search.
- **Phát hiện bất thường** (Anomaly/Fraud Detection)
  - Tình hình tiêu thụ điện có bất thường gì?
- **Tìm kiểu mẫu** (Finding Patterns)
  - Hầu như 80% khách hàng mua "khẩu trang y tế" và "nước rửa tay sát khuẩn" chung một đơn hàng trong mùa dịch cúm.

Tại sao cần sử dụng máy học?

Làm thế nào để phân loại thư rác.
Input: Email.
Output: Spam/Not spam.

Tại sao cần sử dụng máy học?

Hãy xem chương trình lọc **thư rác** sử dụng kỹ thuật lập trình truyền thống sau:

- **Bước 1**: Chúng ta tìm những điểm nổi bật để nhận diện thư rác. Chúng ta quan sát được một số từ, cụm từ, câu văn phổ biến trong thư rác như "*free*", "*4U*", "*amazing*", "*credit card*", v.v....
- **Bước 2**: Chúng ta viết một thuật toán dựa trên các kiểu mẫu "pattern" ở Bước 1: NẾU chứa 4U hoặc free THÌ gán là thư rác.
- **Bước 3**: Kiểm thử chương trình và lặp lại **Bước 1** và **Bước 2** đến khi chương trình đủ tốt để ứng dụng được.

Tại sao cần sử dụng máy học?

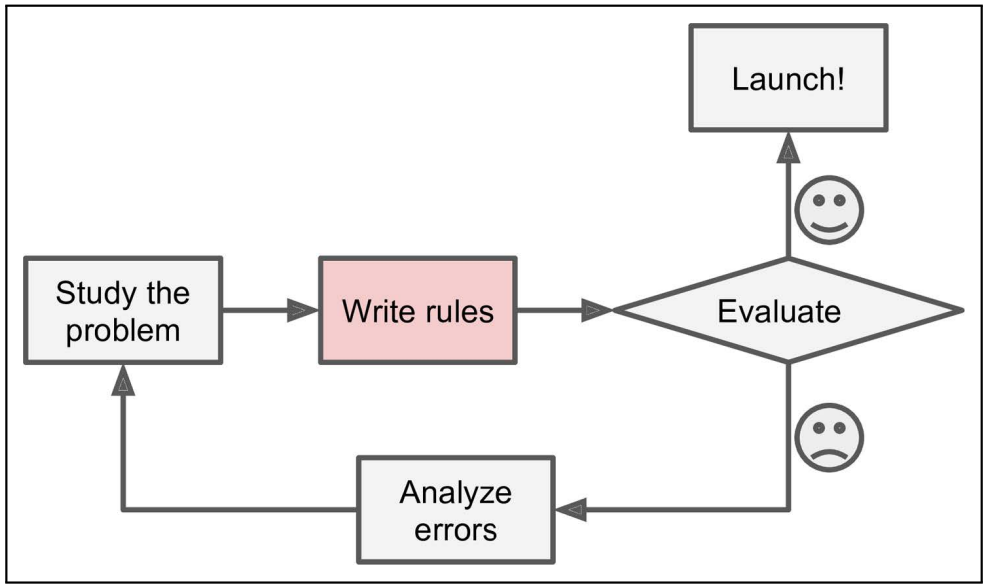
Hãy xem chương trình lọc **thư rác** sử dụng kỹ thuật lập trình truyền thống sau:

- **Bước 1**: Chúng ta tìm những điểm nổi bật để nhận diện thư rác. Chúng ta quan sát được một số từ, cụm từ, câu văn phổ biến trong thư rác như "*free*", "*4U*", "*amazing*", "*credit card*", v.v....
- **Bước 2**: Chúng ta viết một thuật toán dựa trên các kiểu mẫu "pattern" ở Bước 1: NẾU chứa 4U hoặc free THÌ gán là thư rác.
- **Bước 3**: Kiểm thử chương trình và lặp lại **Bước 1** và **Bước 2** đến khi chương trình đủ tốt để ứng dụng được.

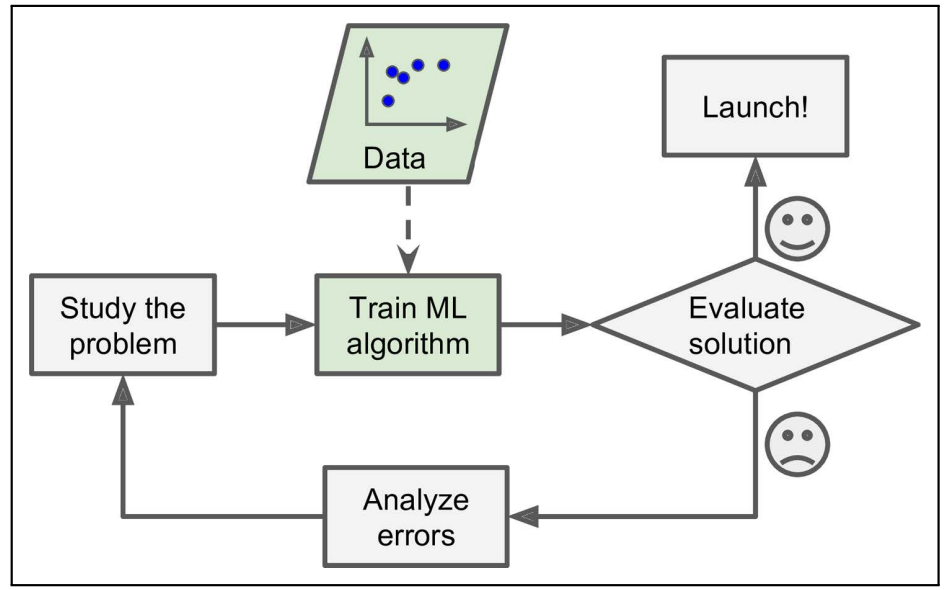
Chuyện gì sẽ xảy ra nếu như người spam thay đổi cách viết để tránh phát hiện? ("for you" thay vì "4U"?)

Tại sao cần sử dụng máy học?

→Máy học giúp cho máy tính có thể tự động rút ra các “pattern/rule”.

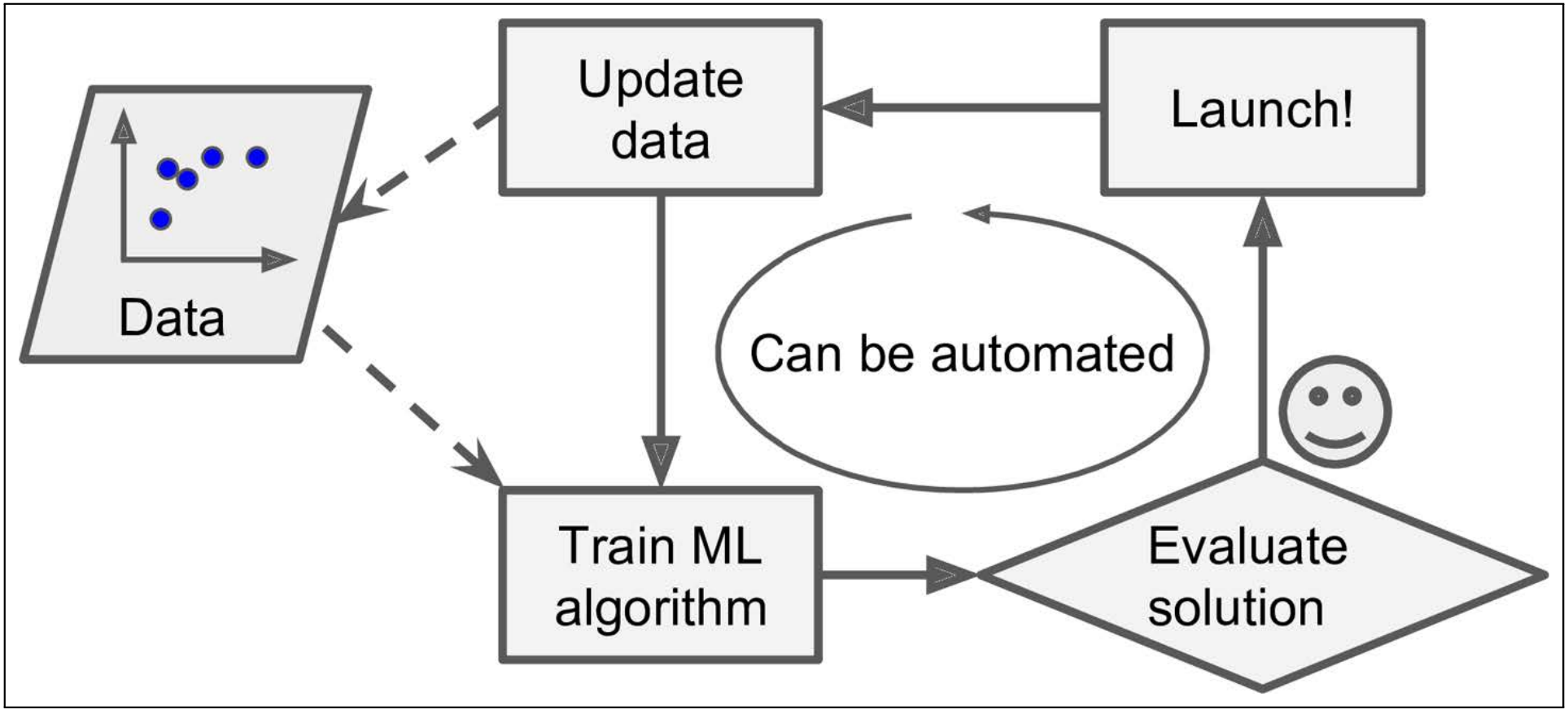


Hình 1: Lập trình truyền thống



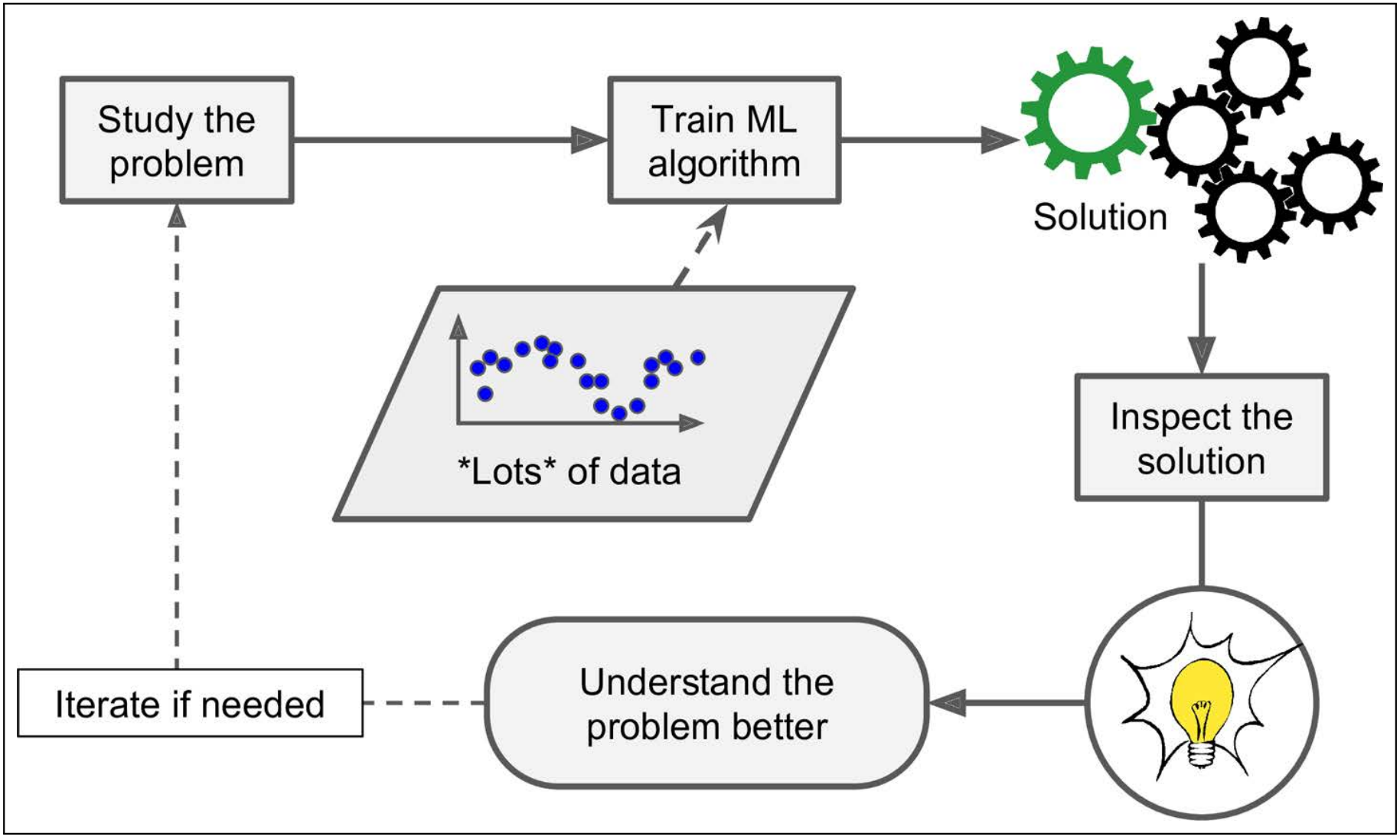
Hình 2: Lập trình máy học

Tại sao cần sử dụng máy học?



Hình 3: Máy học có thể tự động thích ứng với dữ liệu mới

Tại sao cần sử dụng máy học?



Hình 4: Máy học có thể giúp con người rút ra các quy luật từ dữ liệu

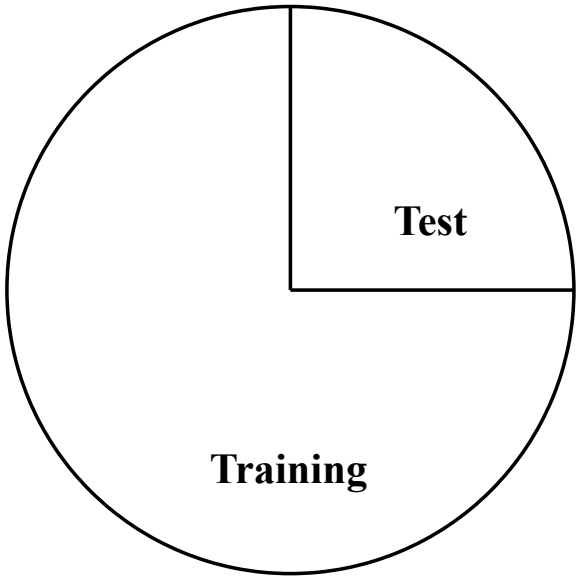
Tóm lại, máy học tốt cho ...

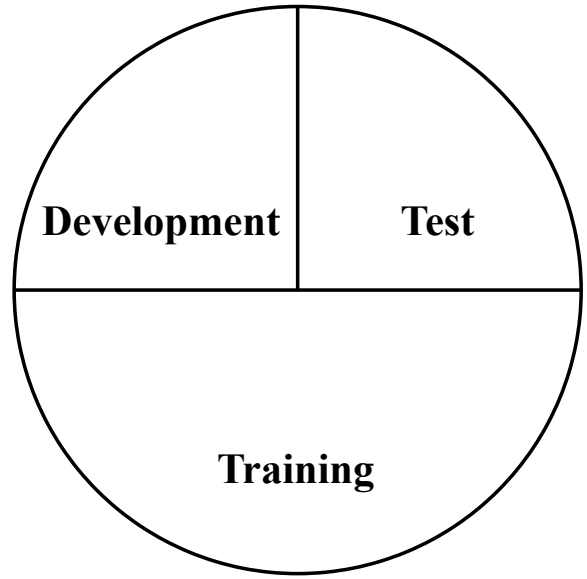
- **Bài toán phải giải bằng một danh sách các luật**: máy học thường đơn giản hóa mã nguồn và cho kết quả tốt hơn.
- **Bài toán phức tạp** không có cách giải tốt bằng những cách truyền thống: máy học có thể giúp tìm lời giải gần đúng.
  - Ví dụ: nhận dạng tiếng nói (nhiều giọng nói khác nhau, môi trường âm thanh nhiễu, nhiều thứ tiếng khác nhau).
- **Ngữ cảnh bài toán biến động:** một hệ thống máy học có thể thích ứng với dữ liệu mới.
- Giúp con người hiểu về dữ liệu lớn
  - Ví dụ: Từ dữ liệu kinh doanh bán lẻ có thể rút ra là người mua hàng mua mặt hàng X hay mua mặt hàng Y.

Nội dung trình bày

❷ Kiểm thử & Thẩm định

Testing (Kiểm thử) & Validation (Thẩm định)





- Các tập dữ liệu cần để xây dựng hệ thống máy học: huấn luyện (training dataset), thẩm định (validation/development dataset), kiểm thử (test dataset).
- Cảnh báo: Lỗi thường gặp!
- Không bao giờ đánh giá một hế thống máy học trện tập dữ liệu dành để phát triển hệ thống (dữ liệu dùng cho huấn luyện, tinh chỉnh tham số)!
- Đánh giá chính thức hệ thống máy học một lần duy nhất trên tập dữ liệu kiểm thử (tập test)!

Nội dung trình bày

❸ Phân loại máy học
- Học có giám sát/không giám sát
- Batch & Online Learning
- Instance-Based & Model-Based Learning

Tiêu chí phân loại hệ thống máy học

Có thể phân loại các hệ thống máy học dựa trên những tiêu chí sau:

- Có sự giám sát của con người hay không (supervised, unsupervised, semi-supervised, và reinforcement learning).
- Có thể học tích lũy một cách nhanh chóng hay không (online và batch learning).
- Học bằng cách so sánh một điểm dữ liệu mới với điểm dữ liệu đã biết, hay phát hiện các kiểu mẫu trong tập dữ liệu huấn luyện và xây dựng mô hình dự đoán như các nhà khoa học làm (instance based và model based learning).

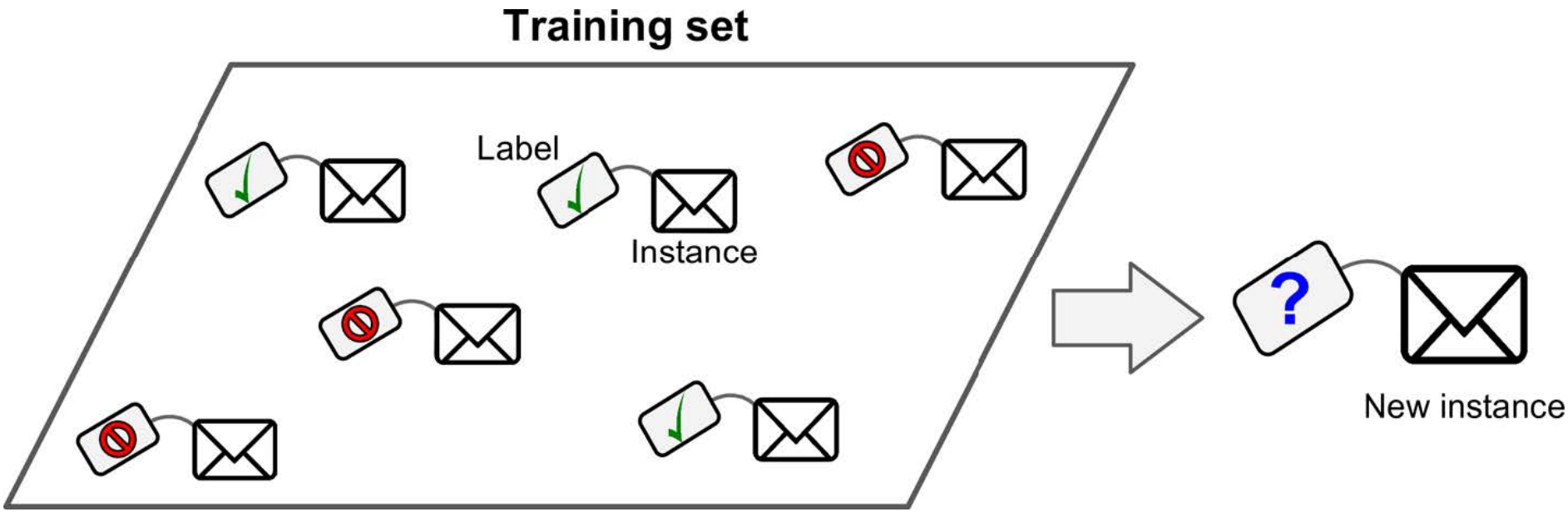
Học có giám sát/không giám sát

Phân loại dựa trên loại hình giám sát trong quá trình huấn luyện.

- Học có giám sát
  - Dữ liệu huấn luyện bao gồm nhãn (label)
- Học không giám sát
  - Dữ liệu huấn luyện không bao gồm nhãn
  - Tìm kiếm các cấu trúc ẩn/thú vị trong dữ liệu
- Học bán giám sát (semi-supervised)
  - Dữ liệu huấn luyện gồm một ít nhãn
- Học củng cố (Reinforcement learning)
  - Máy tính học **chiến lược hành động** bằng cách lựa chọn hành động có thể tối ưu phần thưởng nhận được.

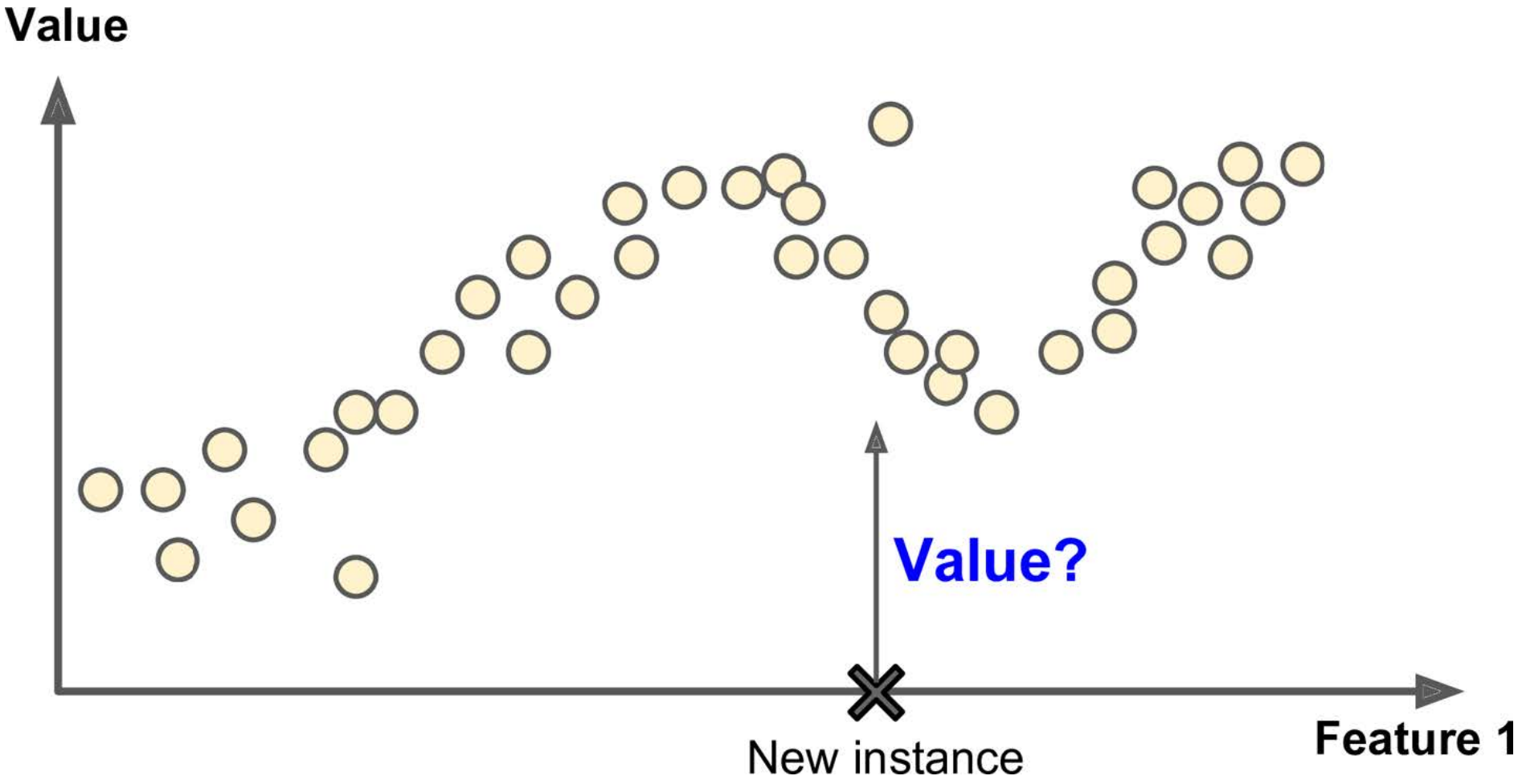
Học có giám sát

Bài toán phân loại thư rác (với hai nhãn: **Phải thư rác** và **Không phải thử rác**).
Mỗi mẫu trong dữ liệu huấn luyện đã được gán nhãn bởi con người.



Học có giám sát: Hồi quy

Ví dụ: Bài toán dự đoán giá xe, giá nhà đất.



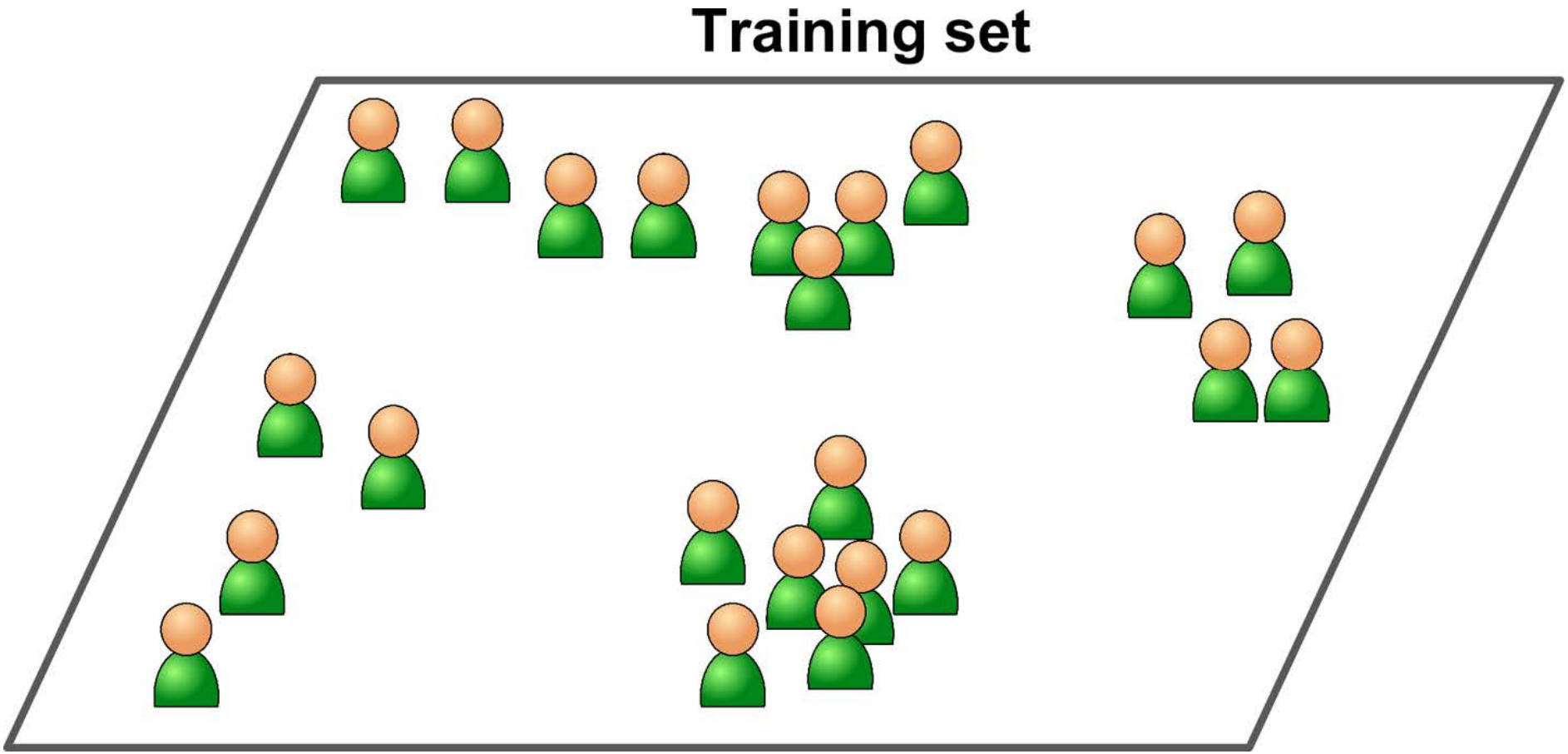
Lưu ý: Một thuật toán máy học cho bài toán hồi quy có thể áp dụng được cho bài toán phân loại và ngược lại.

Học có giám sát

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees & Random Forests
- Neural networks

Học không giám sát

Dữ liệu huấn luyện không được gán nhãn bởi con người.

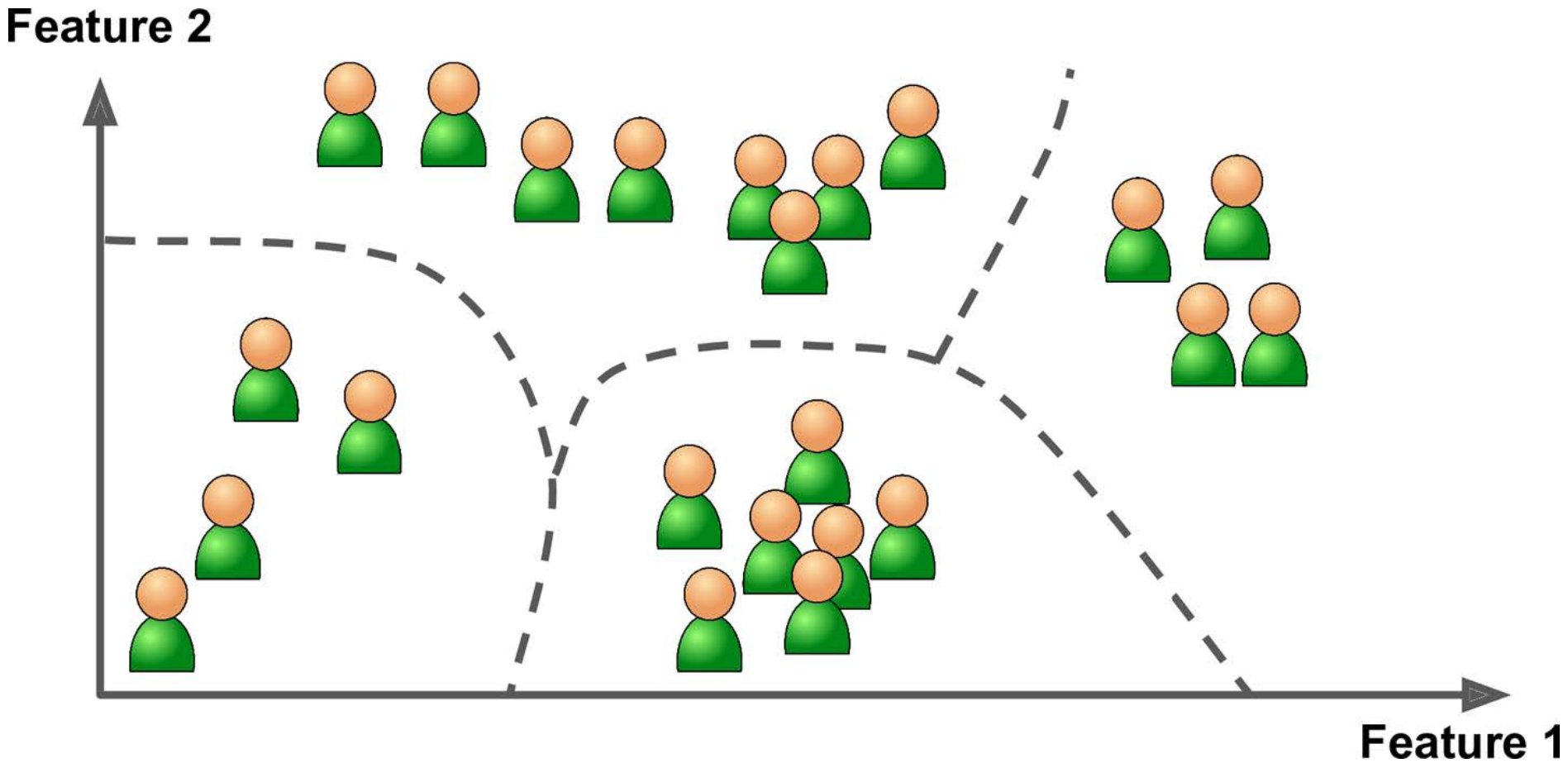


Một số thuật toán quan trọng

- Gom cụm (clustering)
  - k-Means
  - Hierarchical Cluster Analysis (HCA)
  - Expectation Maximization
- Trực quan hóa và Giảm số chiều của dữ liệu (Visualization & Dimensionality Reduction)
  - Principal Component Analysis (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Học luật kết hợp (Association rule learning)
  - Apriori
  - Eclat

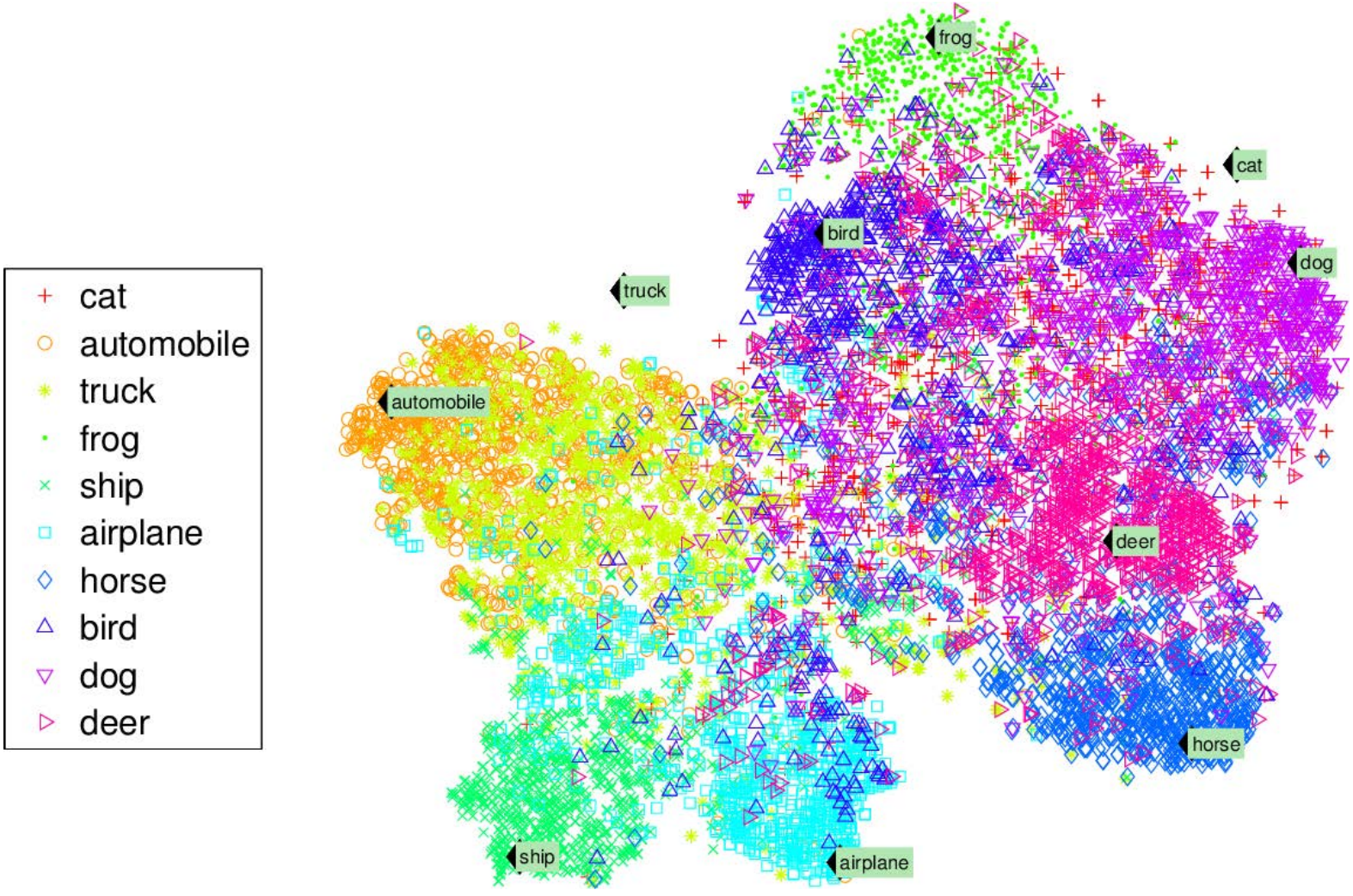
Học không giám sát: Gom cụm (Unsupervised Learning: Clustering)

Ví dụ: Bài toán gom nhóm khách hàng.



Học không giám sát: t-SNE Visualization

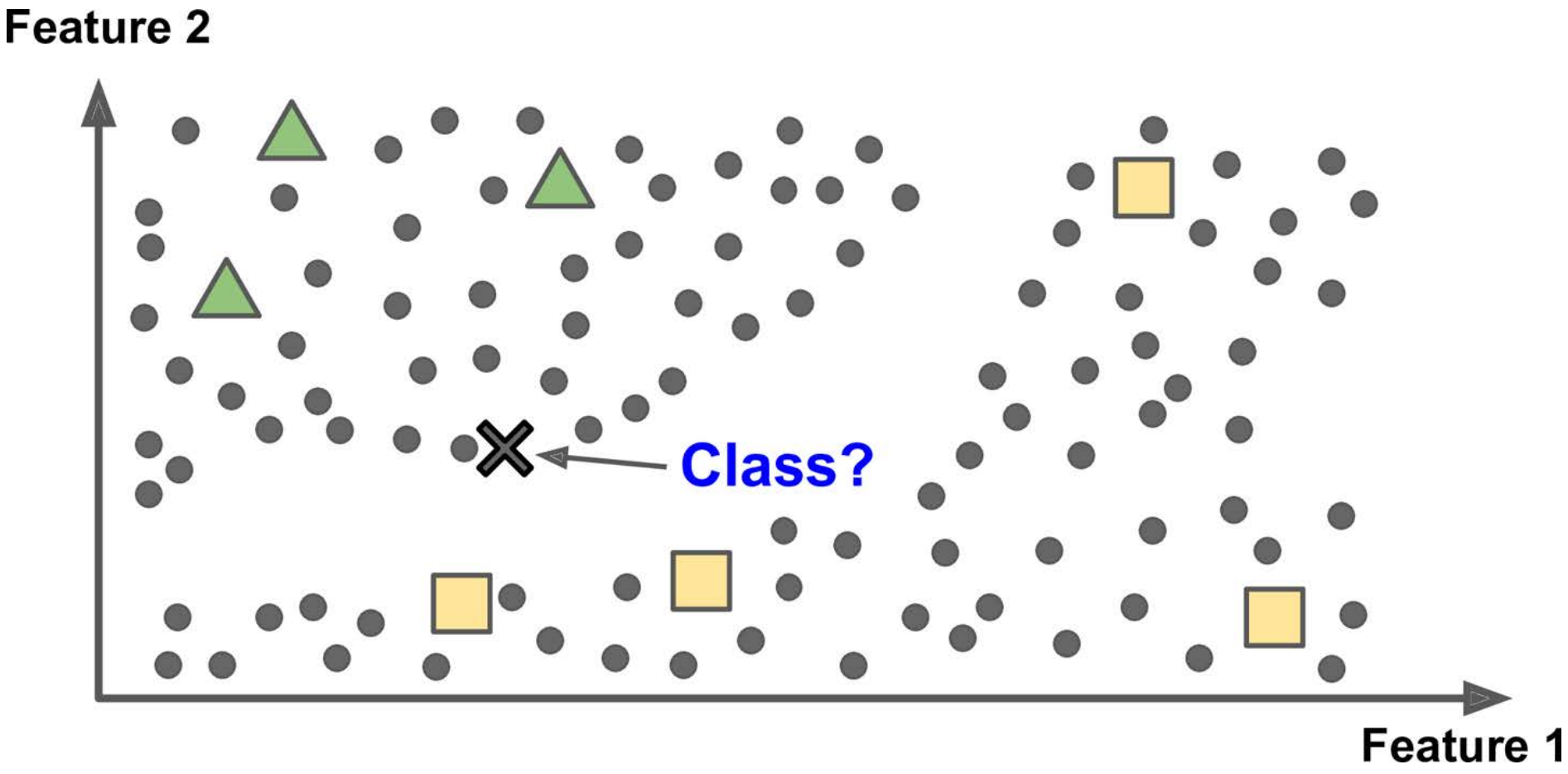
Visualization: Chuyển dữ liệu có cấu trúc phức tạp thành dạng biểu diễn 2D hoặc 3D để có thể vẽ biểu đồ để quan sát trực quan.



Học bán giám sát

Dữ liệu huấn luyện được gán nhãn một phần (thường là nhiều dữ liệu không nhãn và một ít dữ liệu có nhãn)

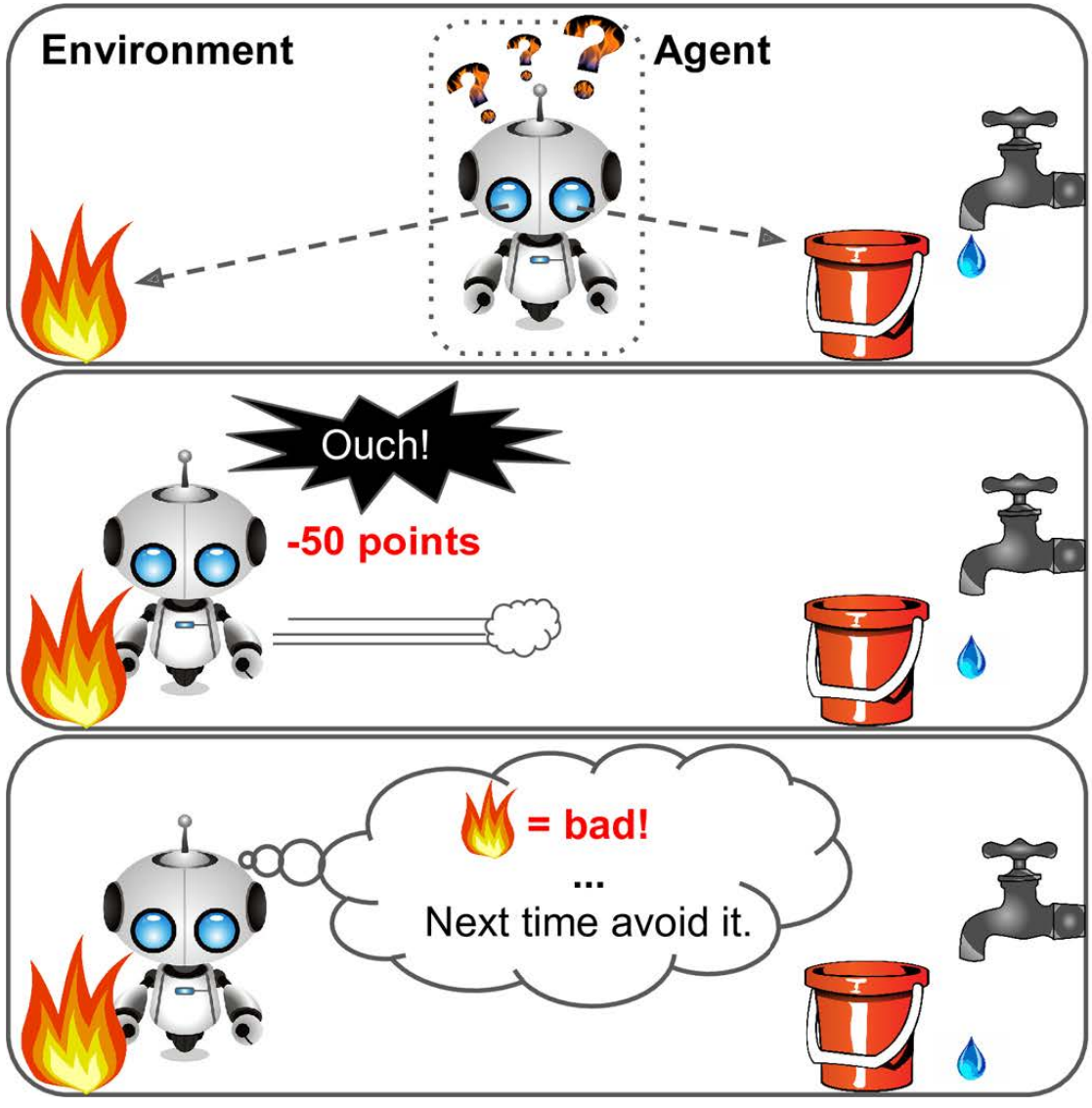
Hình: Bài toán phân 2 lớp: hình vuông và hình tam giác. Những điểm hình tròn là dữ liệu không gán nhãn.



Học củng cố (Reinforcement Learning)

- Hệ thống máy học gọi là "tác tử" (agent) học bằng cách quan sát môi trường, chọn và thực hiện "các hành động" và nhận được "phần thưởng" hoặc "hình phạt".
- Hệ thống này phải học **chiến lược** tốt nhất để nhận được nhiều phần thưởng nhất sau một thời gian.
- Chiến lược xác định hành động gì sẽ được chọn trong một hoàn cảnh nhất định.

Ví dụ minh họa: học củng cố



1. Observe
2. Select action using policy
3. Action!
4. Get reward or penalty
5. Update policy (learning step)
6. Iterate until an optimal policy is found

Kết phần

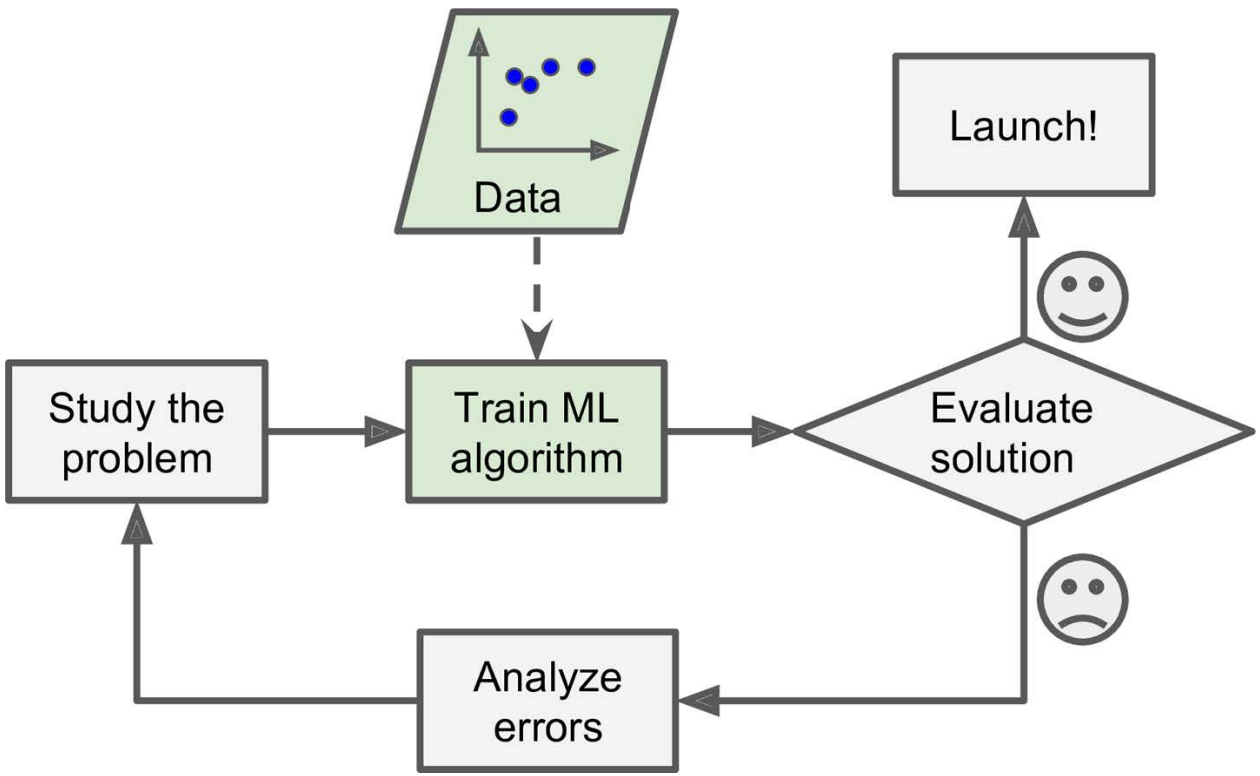
Phân loại các phương pháp máy học theo yếu tố có giám sát của con người hay không.

Batch & Online Learning

Phân loại các hệ thống máy học theo khả năng học tích lũy từ dữ liệu vào liên tục.

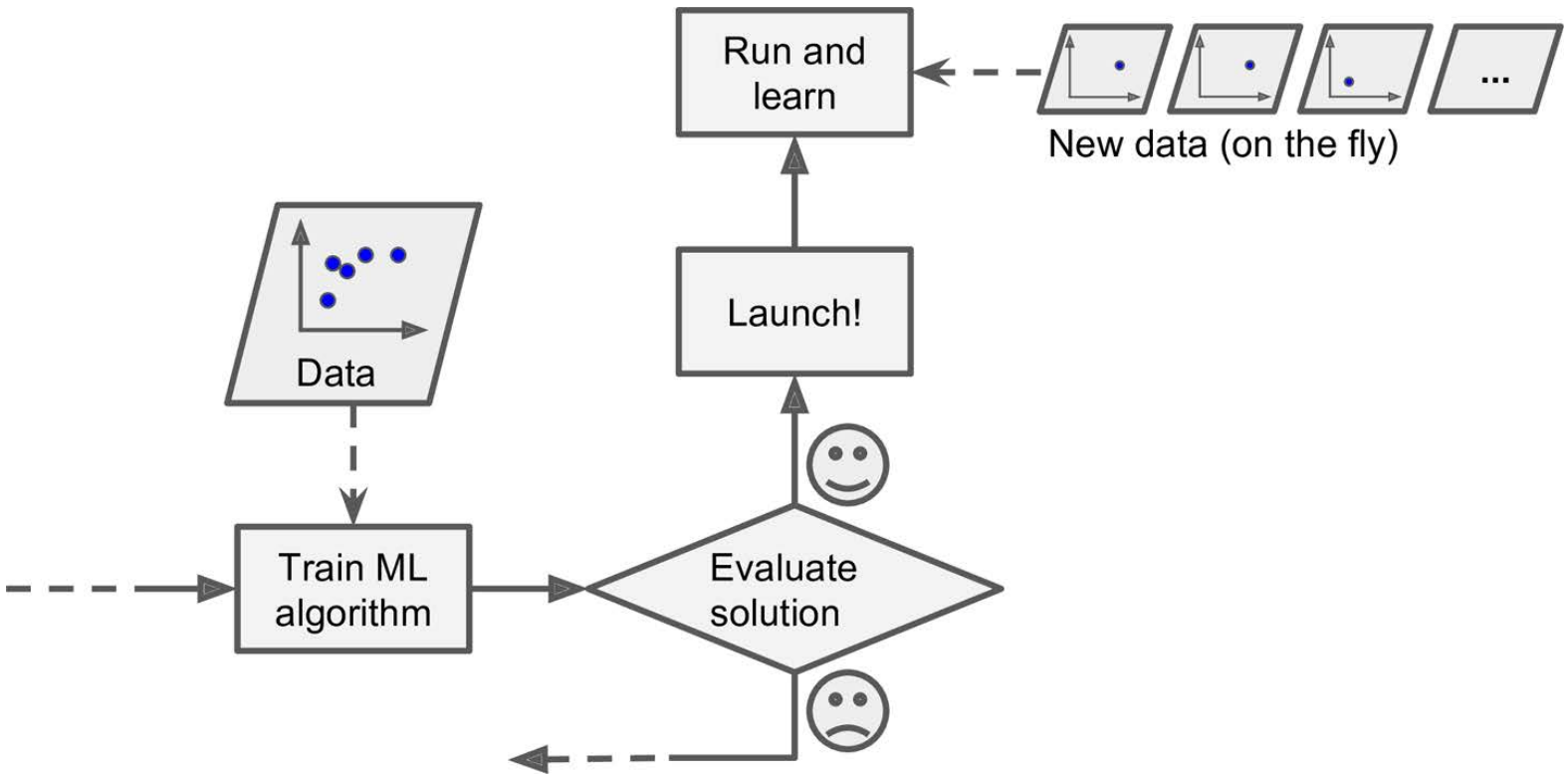
Batch Learning

- Hệ thống được **huấn luyện với toàn bộ dữ liệu** huấn luyện có sẵn, khi vận hành không cần huấn luyện nữa (offline learning).
- Tốn thời gian và **tài nguyên tính toán**.
- Khi có dữ liệu huấn luyện mới thì **phải huấn luyện lại** với dữ liệu cũ + dữ liệu mới.
- **Dữ liệu huấn luyện cực lớn** hoặc **dữ liệu mới phát sinh thường xuyên** thì không thể/không nên dùng cách này.



Học trực tuyến (Online Learning)

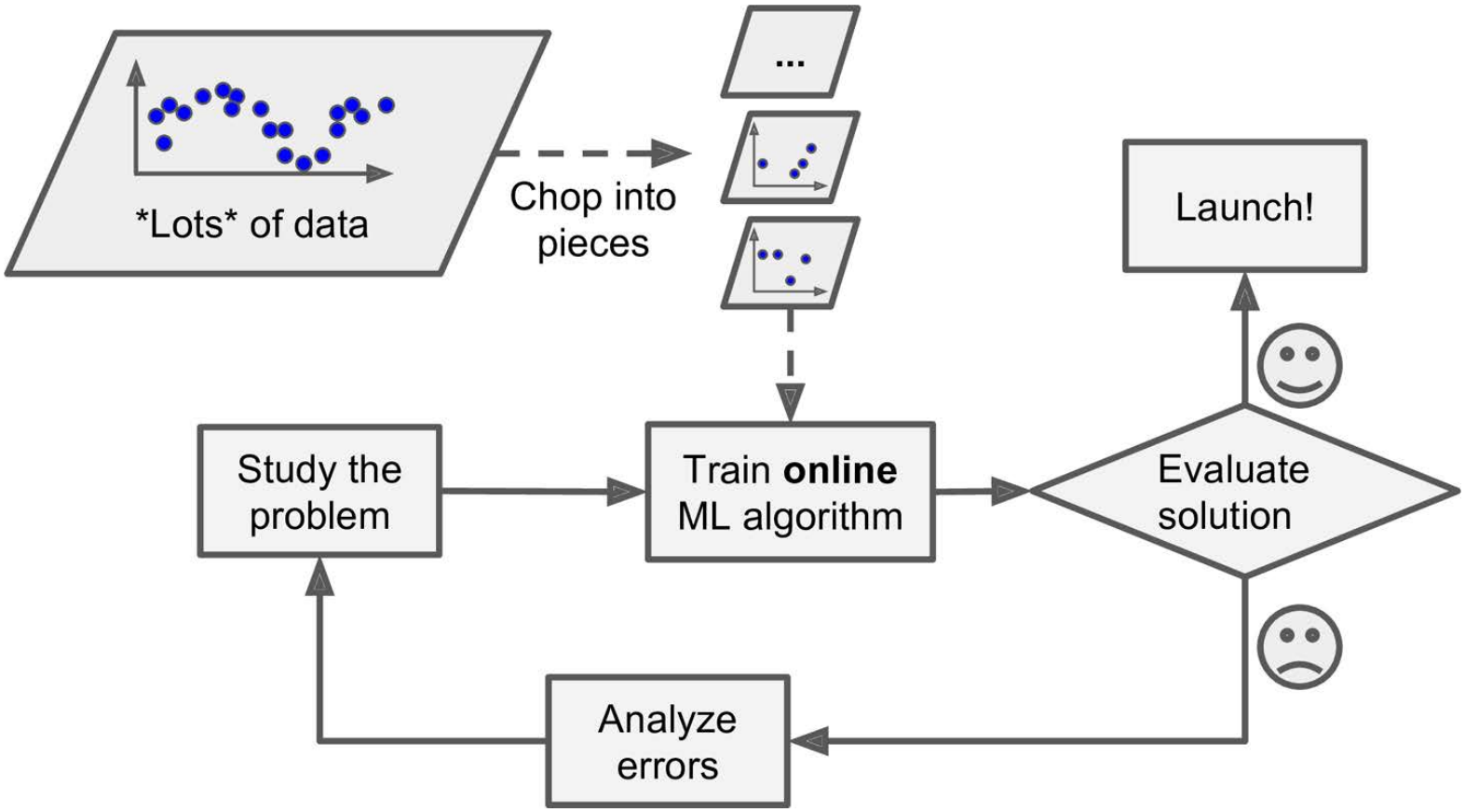
- Hệ thống được huấn luyện tăng cường theo từng mẩu dữ liệu hoặc từng nhóm nhỏ dữ liệu (mini-batch).
- Mỗi bước học nhanh và ít tốn tài nguyên, vì vậy hệ thống có thể học nhanh chóng với dữ liệu mới (on-the-fly).



Học trực tuyến (online learning) (tt)

- Thích hợp khi dữ liệu vào liên tục thay đổi (dữ liệu chứng khoán) và hệ thống cần thay đổi nhanh chóng.
- Thích hợp cho dữ liệu lớn không thể chứa tất cả vào bộ nhớ.
- Learning rate: Là một tham số quan trọng của thuật toán học online, cho biết hệ thống phải thay đổi nhanh chóng cỡ nào với dữ liệu mới.
- Rủi ro: dữ liệu vào chất lượng không tốt sẽ làm giảm chất lượng hệ thống nên phải theo dõi hoạt động của hệ thống.

Online Learning: Xử lý dữ liệu lớn

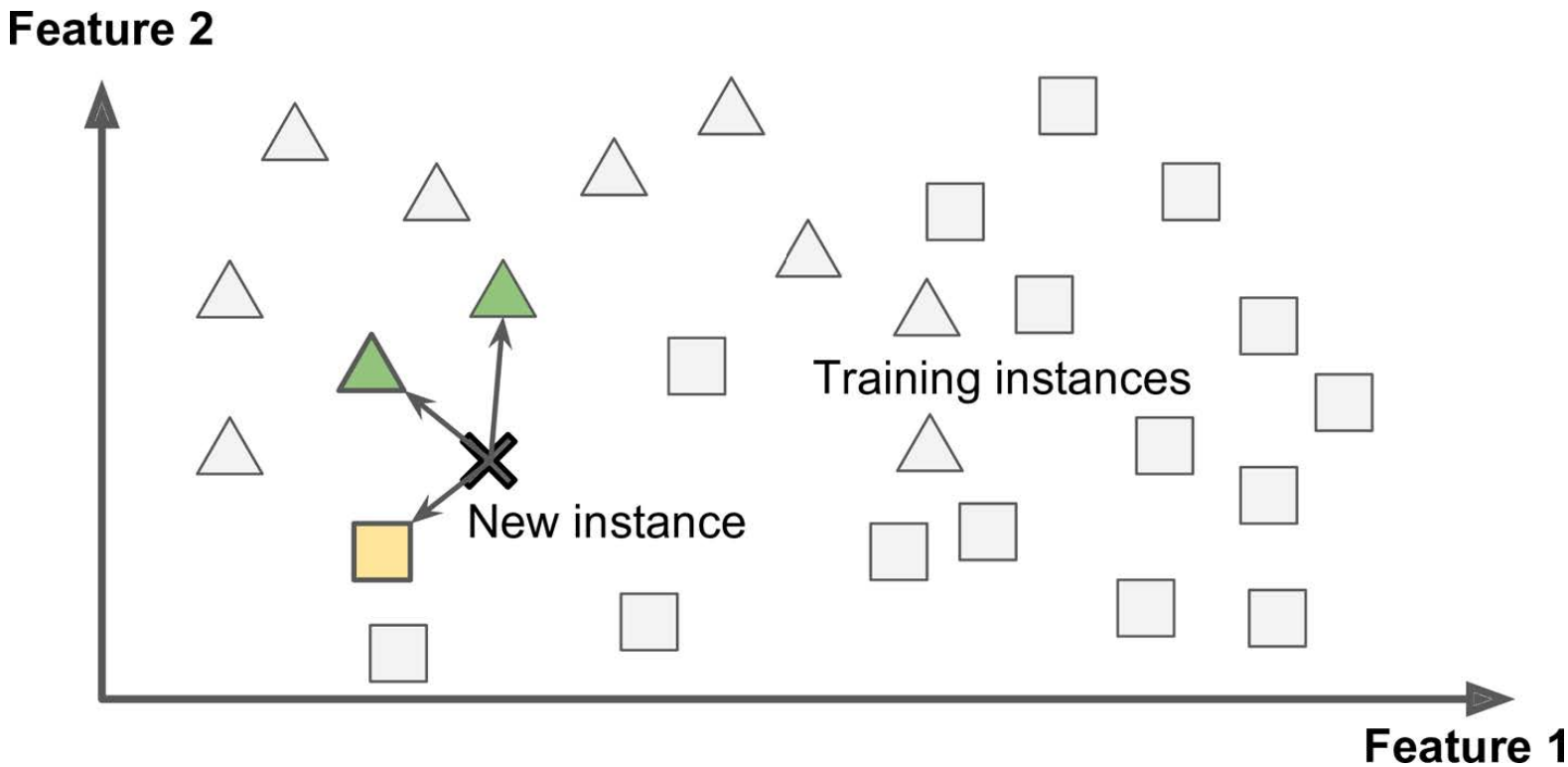


Học dựa trên mẫu và Học dựa trên mô hình

- **Khả năng tổng quát hóa** của hệ thống máy học: Là khả năng dự đoán cho dữ liệu mới chưa từng gặp (trong quá trình huấn luyện).
- Có hai cách tiếp cận để làm cho hệ thống có khả năng tổng quát hóa:
  - Học dựa trên mẫu
  - Học dự trên mô hình

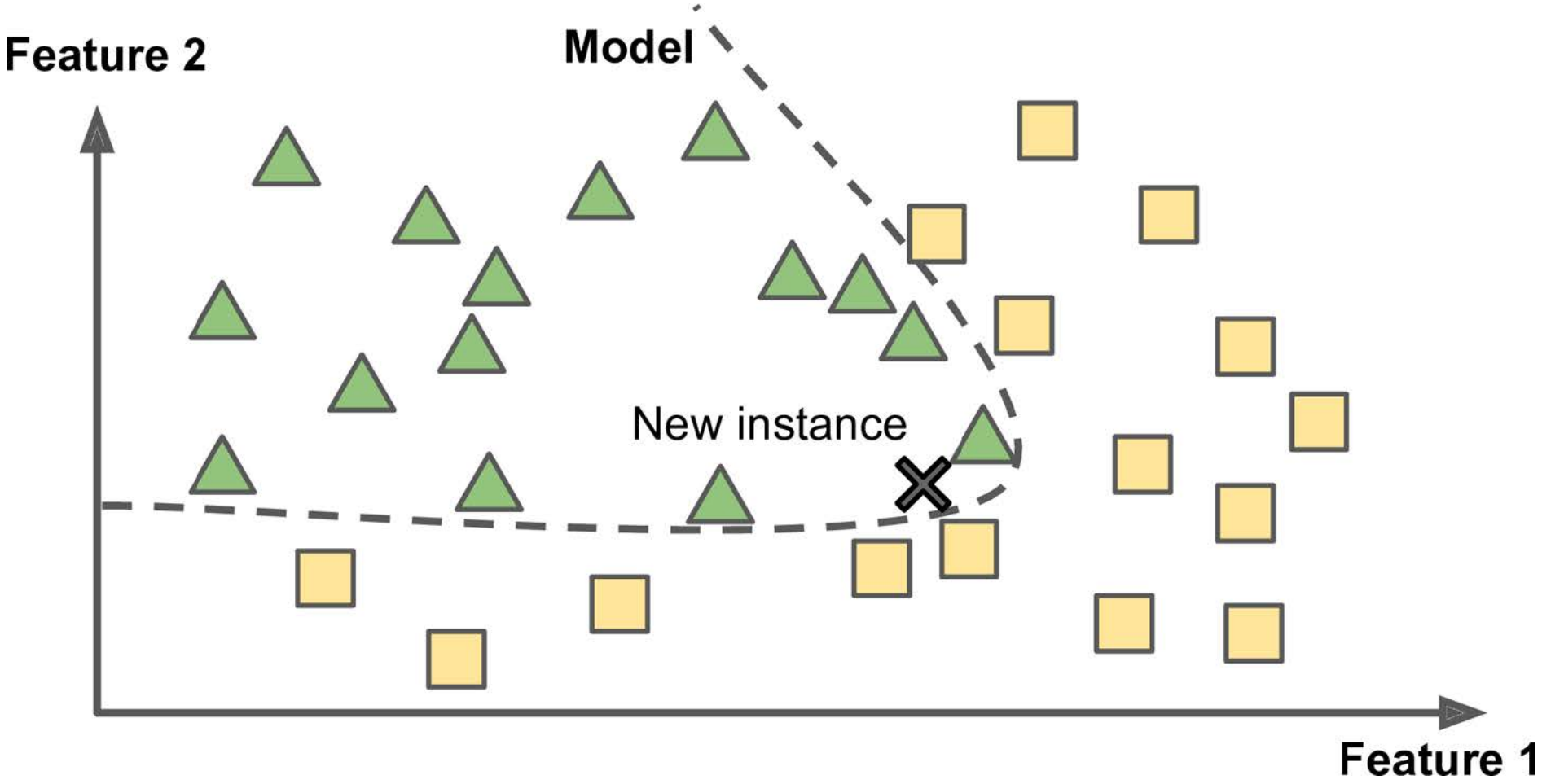
Học dựa trên mẫu

- Hệ thống học thuộc các mẫu trong dữ liệu huấn luyện (mẫu đã học)
- Khi gặp mẫu mới, hệ thống sử dụng một **độ đo tương đồng** (similarity measure) để so sánh khoảng cách giữa mẫu mới và mẫu (hoặc nhóm mẫu) đã học.
- VD bài toán phân loại thư rác: khoảng cách có thể là số từ giống nhau giữa hai thư.



Học dựa trên mô hình

- Hệ thống xây dựng một mô hình từ dữ liệu huấn luyện.
- Khi gặp mẫu mới, hệ thống sử dụng mô hình đó để dự đoán.



Nội dung trình bày

4. Thách thức chính của máy học
   - Thiếu dữ liệu huấn luyện
   - Dữ liệu huấn luyện thiếu tính đại diện
   - Dữ liệu có chất lượng kém
   - Đặc trưng không phù hợp
   - Quá khớp dữ liệu huấn luyện
   - Chưa khớp dữ liệu huấn luyện

Thách thức của máy học

Có 2 loại thách thức bởi "thuật toán không tốt" hoặc "dữ liệu không tốt".

Thách thức do dữ liệu không tốt:

- Thiếu dữ liệu huấn luyện
- Dữ liệu huấn luyện thiếu tính đại diện
- Dữ liệu có chất lượng kém
- Đặc trưng không phù hợp

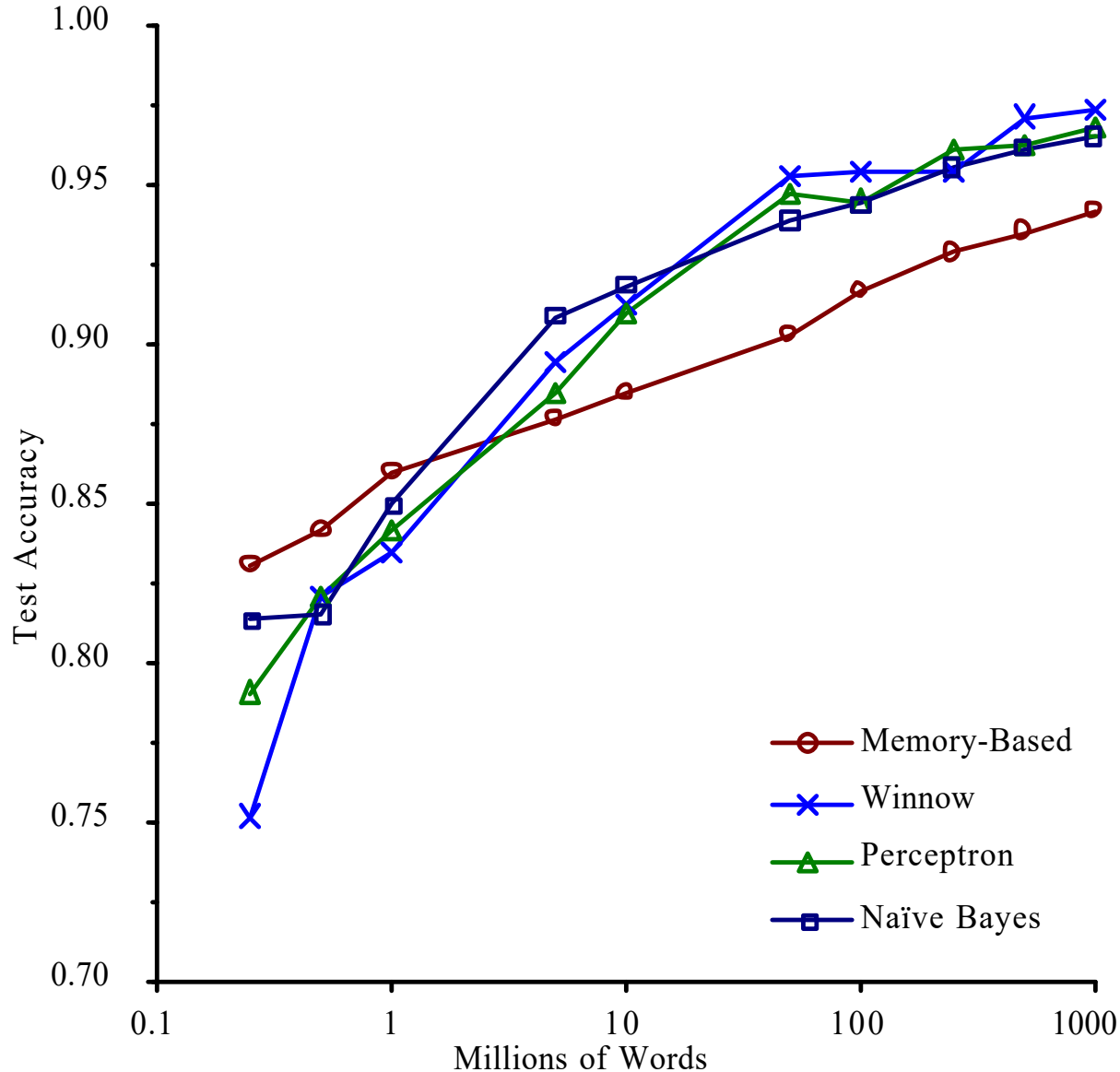
Thách thức do thuật toán không tốt:

- Quá khớp với dữ liệu huấn luyện (over-fitting)
- Chưa khớp với dữ liệu huấn luyện (under-fitting)

Thiếu dữ liệu huấn luyện

- Với một đứa trẻ chập chững học nhận biết quả táo là gì, chúng ta chỉ vào một quả táo và nói rằng "táo" vài lần là trẻ có thể nhận biết quả táo với nhiều màu sắc và hình dạng khác nhau.
- Máy học đến thời điểm hiện tại cần nhiều dữ liệu huấn luyện hơn. Với một bài toán rất đơn giản, thông thường bạn cần **hàng nghìn mẫu** dữ liệu, và đối với các bài toán phức tạp như nhận diện hình ảnh hoặc giọng nói bạn có thể cần **hàng triệu mẫu** dữ liệu.

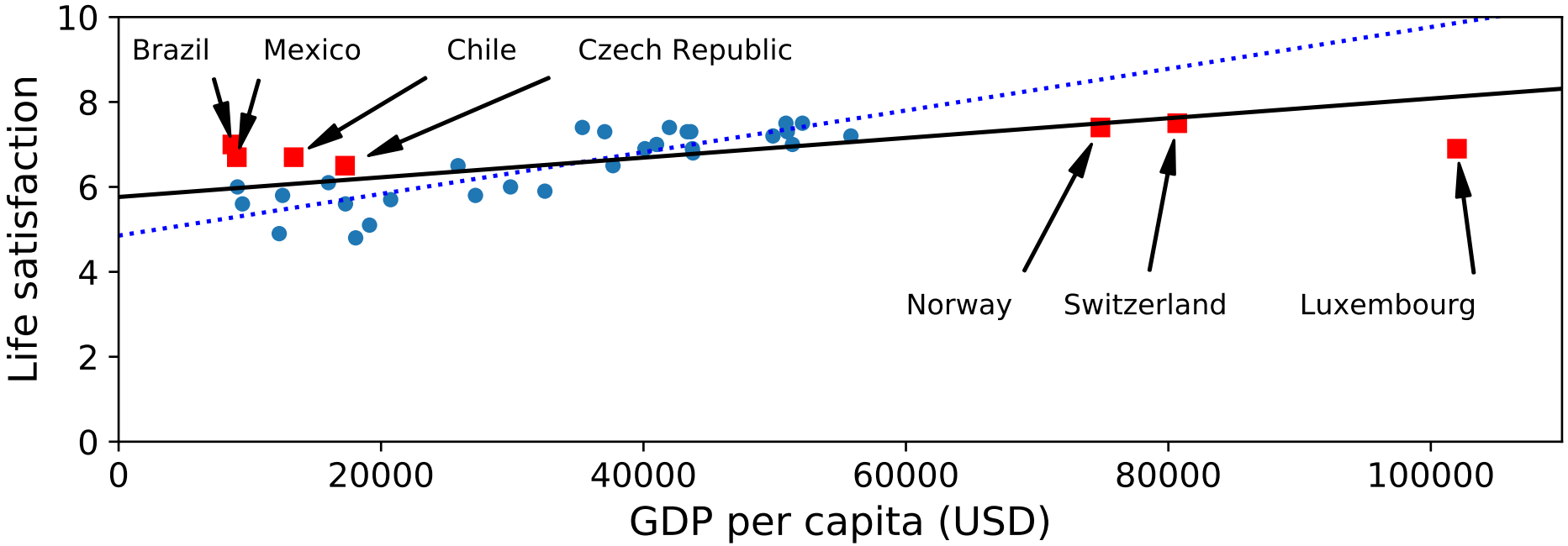
Tầm quan trọng của dữ liệu và thuật toán[1]



[1] [1] Banko and Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation". 2001.

Dữ liệu huấn luyện thiếu tính đại diện

Để hệ thống máy học có thể tổng quát hóa tốt, tập dữ liệu huấn luyện phải có **tính đại diện** cho tất cả dữ liệu muốn dự đoán.



Nếu kích thước tập dữ liệu quá nhỏ, dữ liệu sẽ có những mẫu không đại diện cho phần nhiều dữ liệu, gọi là **nhiễu do lấy mẫu** "sampling noise".
Thậm chí một tập dữ liệu kích thước rất lớn cũng có thể có tính đại diện không tốt nếu phương pháp lấy mẫu có sai sót, gọi là **lệch do lấy mẫu** "sampling bias".

Dữ liệu có chất lượng kém

Dữ liệu huấn luyện có chất lượng kém là dữ liệu có nhiều lỗi, nhiều giá trị ngoại biên (outlier), nhiễu (noise).
Cần **làm sạch dữ liệu huấn luyện**. Sự thật là, hầu hết các nhà khoa học dữ liệu dành phần lớn thời gian chỉ để làm điều này.

Một số cách làm sạch dữ liệu:

- Bỏ dữ liệu mang giá trị ngoại biên hoặc sửa lại bằng tay.
- Nếu một số mẫu thiếu một vài đặc trưng (VD: thiếu thông tin độ tuổi) thì có 3 cách: bỏ đặc trưng này, bỏ những mẫu này hoặc điền giá trị cho đặc trưng này.

Đặc trưng không phù hợp

Hệ thống của bạn sẽ tệ nếu dữ liệu huấn luyện chứa quá nhiều đặc trưng không hữu ích.
Nhiều phương pháp đòi hỏi phải xác định được tập đặc trưng hữu ích. Công đoạn này gọi là **chế tác đặc trưng** "feature engineering", gồm các bước sau:

- Lựa chọn đặc trưng (Feature selection): lựa chọn các đặc trưng hữu ích nhất từ một tập hợp các đặc trưng.
- Rút trích đặc trưng (Feature extraction): kết hợp các đặc trưng đã có để tạo thêm các đặc trưng mới hữu ích hơn.
- Tạo đặc trưng mới bằng cách thu thập thêm dữ liệu.

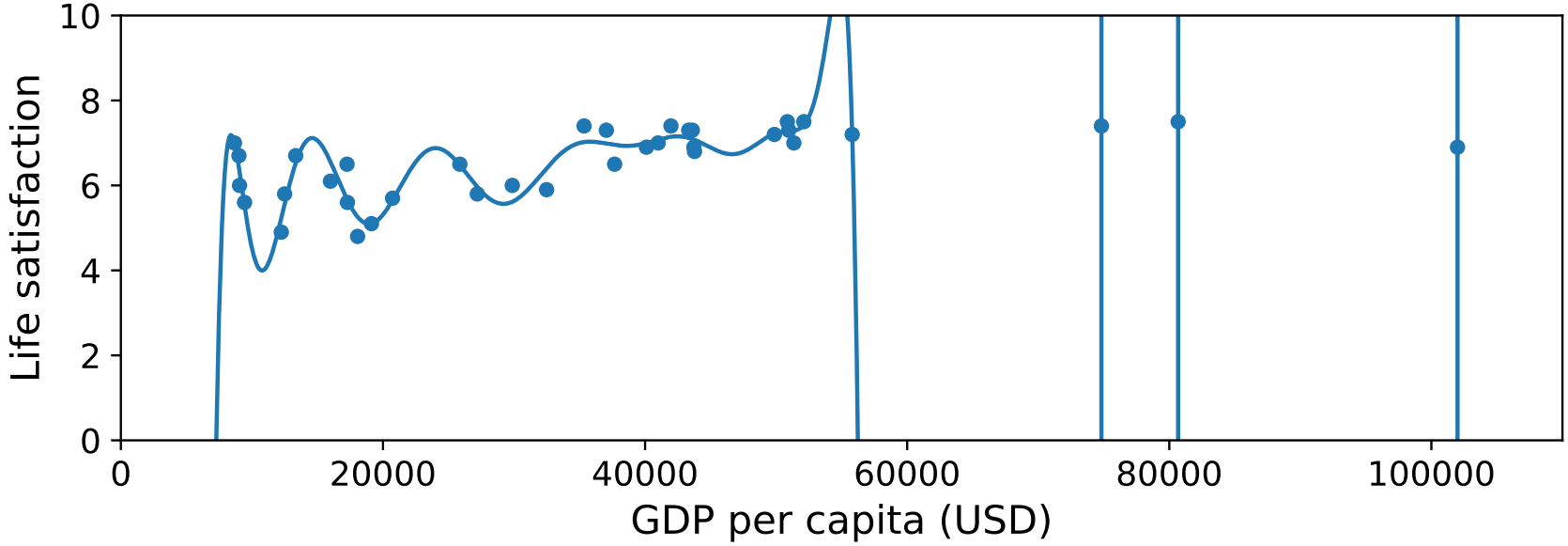
Quá khớp dữ liệu huấn luyện

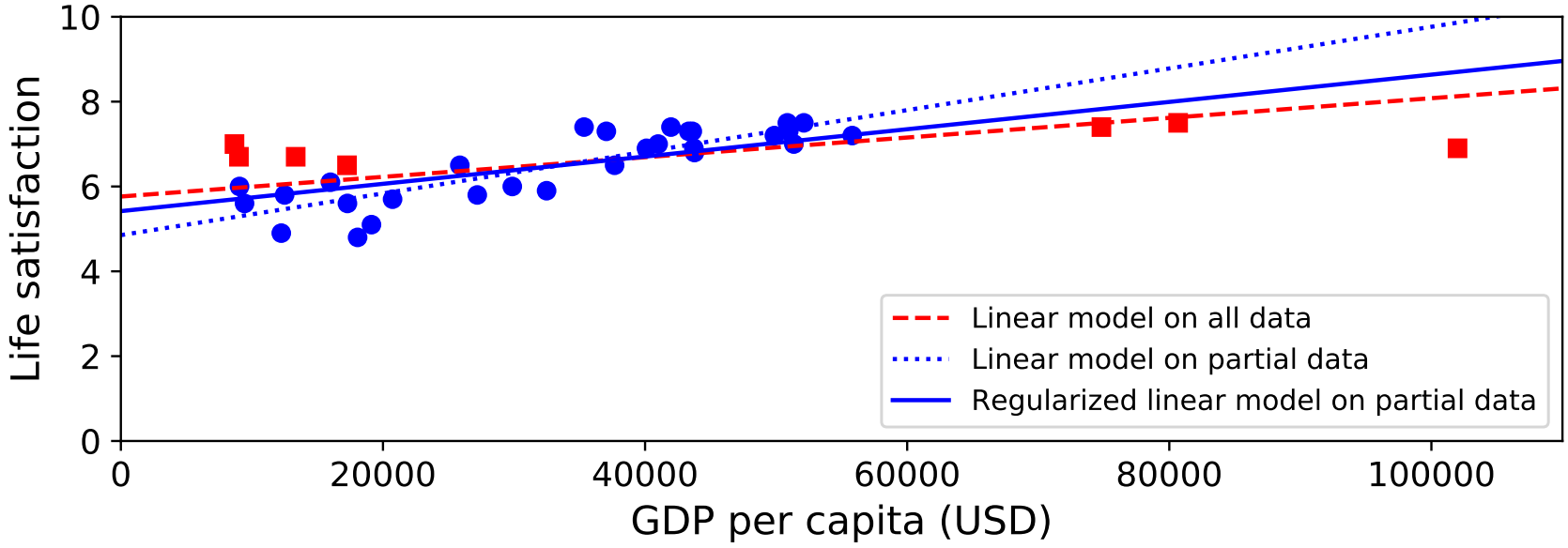
Hiện tượng quá khớp dữ liệu huấn luyện **over-fitting** xảy ra khi mô hình máy học quá chi tiết đến nỗi nó học ra những kiểu mẫu từ mẫu nhiễu.
Một số giải pháp:

- Đơn giản hóa mô hình bằng cách chọn một mô hình có **ít tham số** hơn (ví dụ: một mô hình tuyến tính hơn là một mô hình đa thức bậc cao)
- Đơn giản hóa mô hình bằng cách giảm số lượng đặc trưng của dữ liệu huấn luyện, hoặc bằng cách thêm ràng buộc vào mô hình.
- Thu thập thêm nhiều dữ liệu huấn luyện.
- Giảm nhiễu trong dữ liệu huấn luyện (ví dụ: sửa lỗi dữ liệu bằng tay hoặc xóa giá trị ngoại biên).

2 yếu tố có quan hệ được mất của mô hình (tradeoff): khả năng khớp dữ liệu và khả năng tổng quát hóa.

Quá khớp dữ liệu huấn luyện (tt)





Chưa khớp dữ liệu huấn luyện

Hiện tượng chưa khớp **under-fitting** có tính chất ngược lại với quá khớp **over-fitting**.
Xảy ra khi mô hình quá đơn giản so với cấu trúc cơ bản của dữ liệu.
Một số giải pháp:

- Chọn mô hình máy học mạnh mẽ hơn với nhiều tham số hơn.
- Thêm đặc trưng tốt hơn (chế tác đặc trưng).
- Giảm thiểu ràng buộc cho mô hình (giảm tham số của phương pháp chính quy hóa).

Tóm tắt

- Máy học giúp máy tính giải quyết tốt hơn một số bài toán bằng cách học từ dữ liệu mà không phải viết luật.
- Có nhiều loại hệ thống máy học: giám sát/không giám sát, batch/online, dựa trên mẫu/dựa trên mô hình, ...
- Phân biệt học dựa trên mẫu và học dựa trên mô hình.
- Hệ thống máy học sẽ không hoạt động tốt nếu dữ liệu huấn luyện quá nhỏ, không có tính đại diện, chứa nhiều nhiễu, hoặc nhiều đặc trưng không liên quan.
- Mô hình học không được quá đơn giản mà cũng không được quá phức tạp.
- Phải **đánh giá** hệ thống máy học để biết chất lượng của nó.

Tài liệu tham khảo

- Sách tham khảo “Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd Edition” của tác giả “Aurélien Géron”.