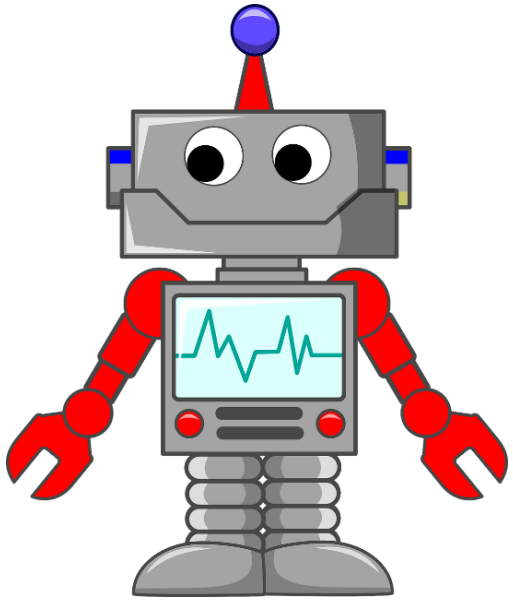
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

BÀI 06

HỌC KHÔNG GIÁM SÁT - UNSUPERVISED LEARNING



TS. Nguyễn Vinh Tiệp

NỘI DUNG

1. GIỚI THIỆU HỌC KHÔNG GIÁM SÁT

2. CÁC MÔ HÌNH GOM NHÓM - CLUSTERING

3. CÁC MÔ HÌNH GIẢM CHIỀU DỮ LIỆU

Giới thiệu

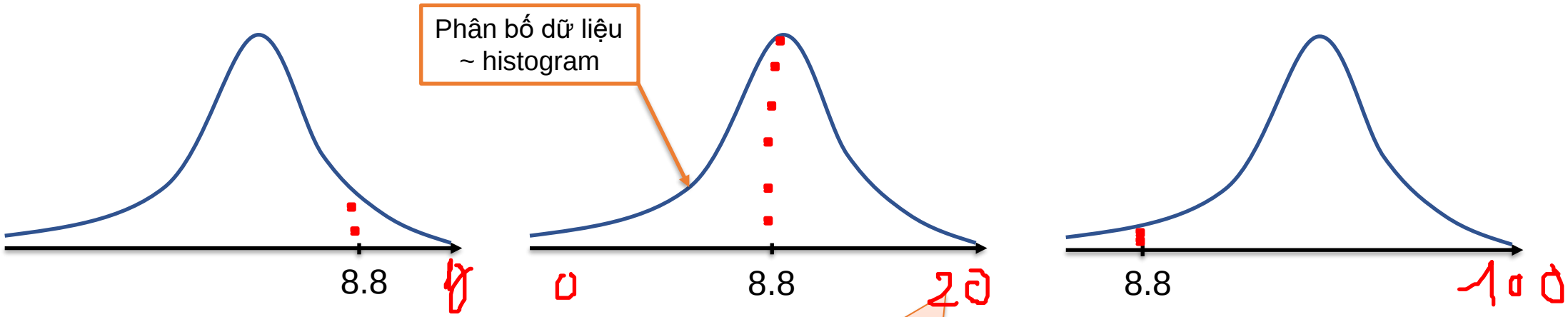
- **Học không giám sát (unsupervised learning)**: là một nhánh của ML, có nhiệm vụ học phân bố của dữ liệu, từ đó có thể biểu diễn dữ liệu hiệu quả hơn
- Dữ liệu cho thuật toán học không giám sát không cần gán nhãn (label)
  - Chỉ cần dữ liệu đầu vào x
  - Không cần nhãn đầu ra tương ứng

không cần y

- Một số chủ đề chính:
  - Gom nhóm dữ liệu
  - Giảm chiều dữ liệu

Phân bố của dữ liệu

- Ví dụ: An có điểm TB là 8.8. Hỏi An xếp loại giỏi, khá hay trung bình?



Phải đưa điểm vào phân bố tổng thể các bạn trong

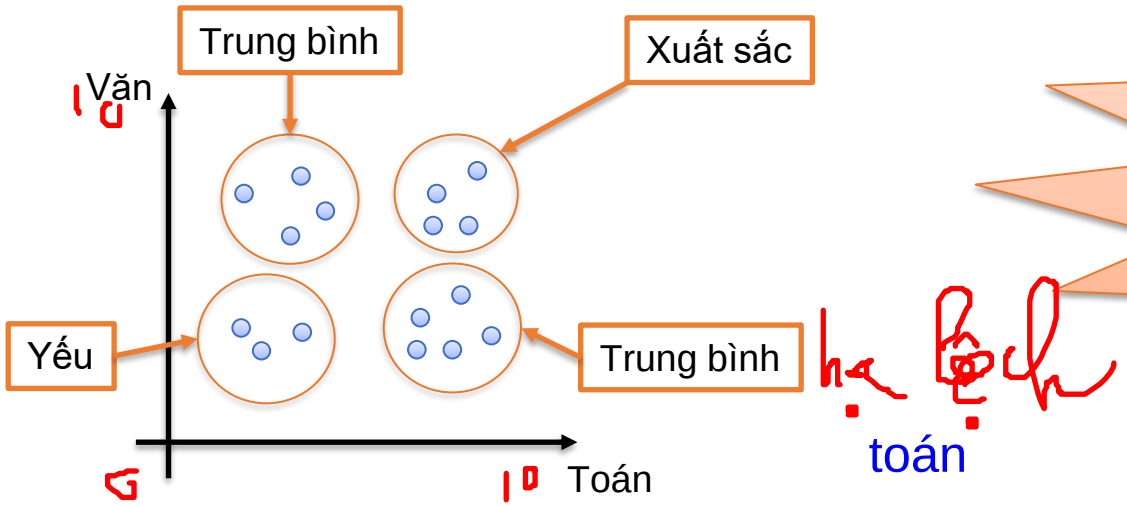
Không thể biết xếp loại nếu không biết phân bố dữ liệu

hoặc 10 thì ta hiểu lớp có nhiều người trên 8.8

Biểu diễn dữ liệu

- Ví dụ: Cho điểm hai môn toán, văn của các bạn trong lớp. Phân loại học lực từng bạn?

học lệch văn



raw gồm toán và văn
chỉ cần lưu học lực thôi

Thay vì lưu dữ liệu "thô"
➔ chỉ cần lưu "đặc trưng" học lực

toán

thuật toán không quan tâm xuất sắc hay gì mà đặt là cluster ID, giá trị gán này là ngẫu nhiên, mình phải quan sát và gọi tên lại nếu cần

- Giảm chiều dữ liệu (VD: giảm 50% số thuộc tính)
- Thể hiện được đặc trưng theo nhóm của mẫu dữ liệu

Lưu ý: ở đây chỉ mượn các khái niệm trong cuộc sống ("yếu", "xuất sắc", "trung bình"). Thực tế, thuật toán UL sẽ mã hóa bằng các cluster ID nào đó!

NỘI DUNG

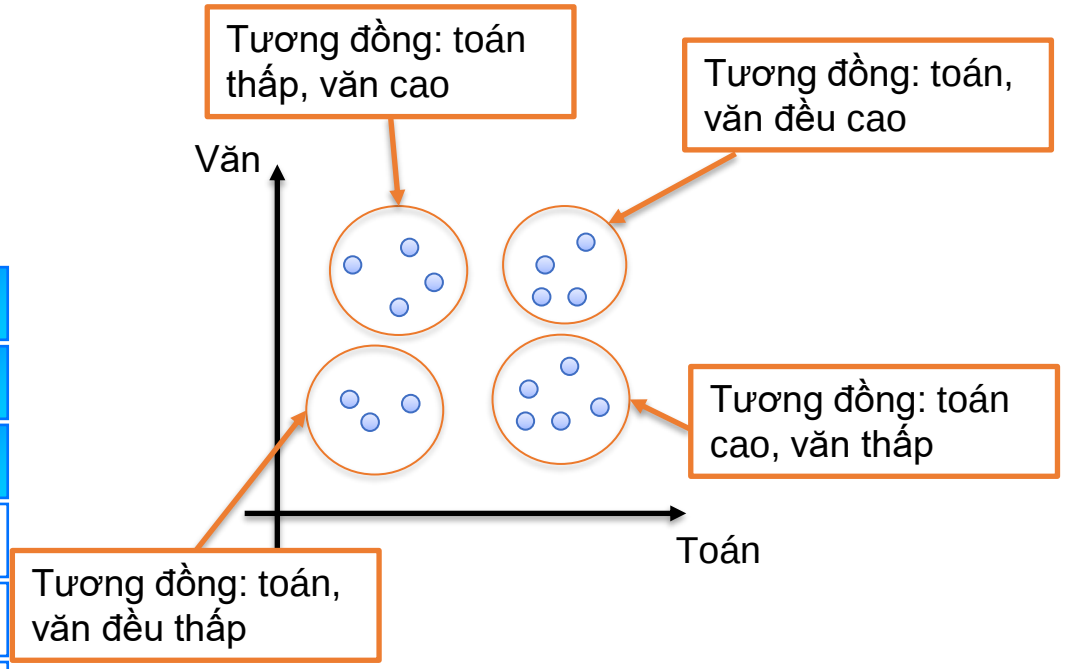
1. GIỚI THIỆU HỌC KHÔNG GIÁM SÁT

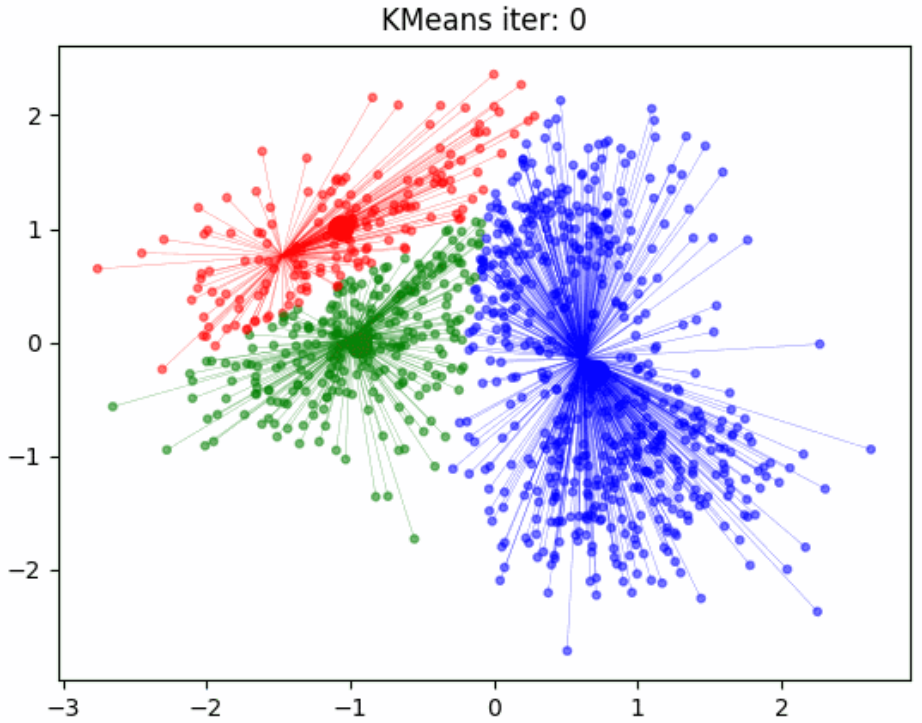
2. CÁC MÔ HÌNH GOM NHÓM - CLUSTERING

3. CÁC MÔ HÌNH GIẢM CHIỀU DỮ LIỆU

Bài toán 1: Gom nhóm dữ liệu

- **Gom nhóm (clustering):** là bài toán gom các đối tượng theo từng cụm sao cho các đối tượng trong cùng một cụm có sự tương đồng với nhau hơn so với những đối tượng thuộc các nhóm khác





Ví dụ gom cụm với thuật toán K-Mean

Bài toán 1: Gom nhóm dữ liệu

- Một số thuật toán gom nhóm dữ liệu:
  - K-Means biết trước k
  - DBSCAN không biết trước k mà chia theo mật độ lớn thì thành 1 cụm
  - Hierarchical clustering
  - Gaussian Mixture Models (GMM)
  - Spectral clustering

- Mỗi phương pháp có những ưu – khuyết điểm riêng. Sử dụng phương pháp nào tùy vào tính chất dữ liệu và mục tiêu cụ thể

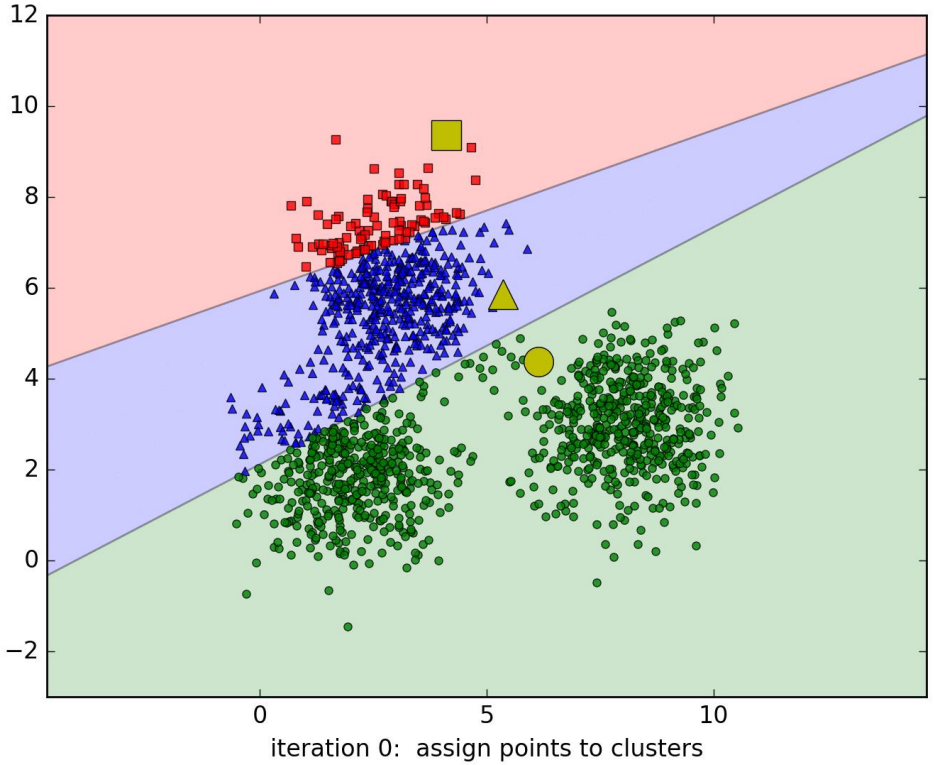
  ví dụ số chiều, số dữ liệu lớn => thuật toán đơn giản: kmeans

  cần biết cụm nào theo phân bố nào, dataset có cụm dày đặc, bỏ các điểm rời rạc => dùng DBSCAN

Thuật toán gom nhóm - KMeans

- **Ý tưởng**: khởi tạo ngẫu nhiên K tâm cụm, sau đó gán các điểm dữ liệu vào trọng tâm gần nhất. Quá trình này được lặp lại cho đến khi các trọng tâm không thay đổi nữa

gán nhãn dữ liệu cho các trọng tâm gần nhất, cập nhật lại trọng tâm cho đến khi trọng tâm không còn thay đổi hoặc rất ít



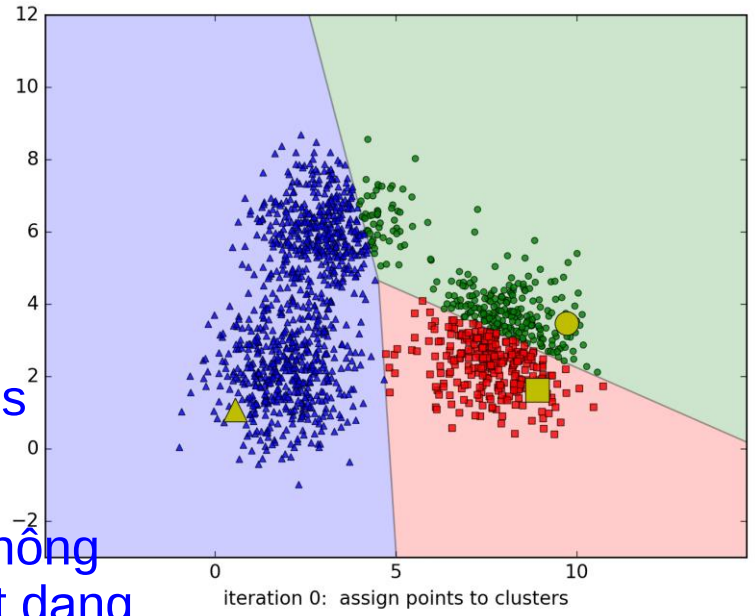
Nguồn: https://machinelearningcoban.com/2017/01/01/kmeans/

Thuật toán gom nhóm - KMeans

- **Ưu điểm**:
  - Đơn giản, dễ cài đặt
  - Hiệu quả với dữ liệu lớn
- **Khuyết điểm**:
  - Cần biết trước số lượng cụm K
  - Dễ bị rơi vào cực tiểu cục bộ
  - Phụ thuộc vào tâm cụm khởi tạo
  - Không hoạt động tốt với dữ liệu có phân bố phức tạp, không phải dạng hình cầu

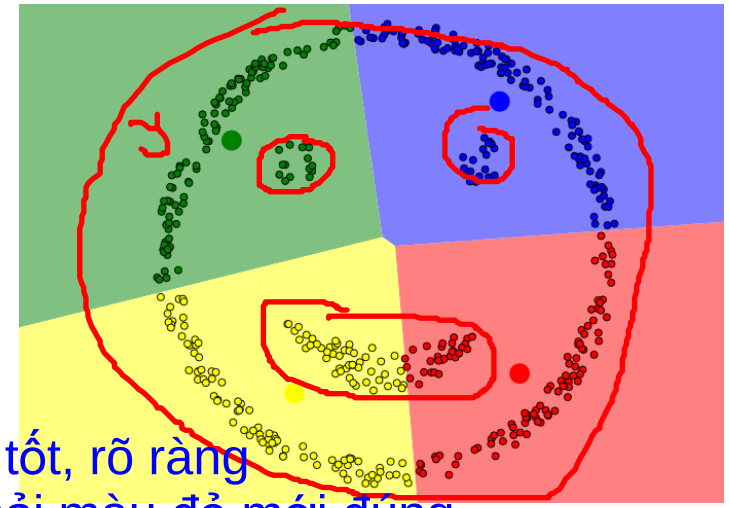
trọng tâm ban đầu

> 1M mẫu thì cũng bắt đầu chậm và không hiệu quả nữa =>dùng AK Means approximate



nhiều tình huống không biết k, vì không biết dạng dữ liệu

phụ thuộc vào khởi tạo k cụm



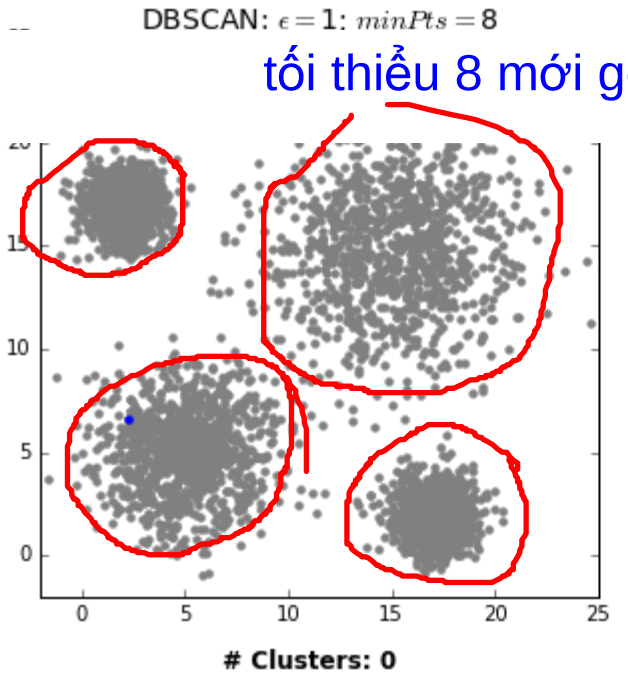
rải rác thì không chạy tốt, rõ ràng đã phân sai 4 cụm, phải màu đỏ mới đúng

Thuật toán gom nhóm - DBSCAN

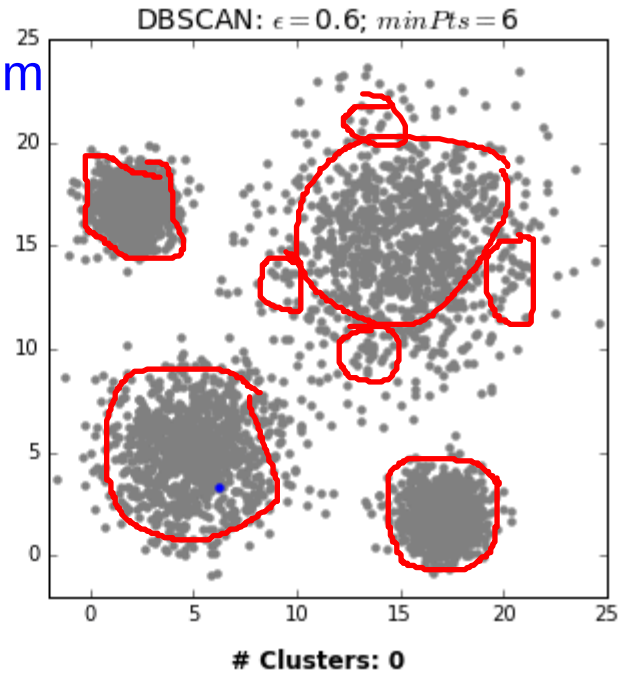
thỏa 2 điều kiện này mới được: điểm gần nhau và mật độ cao

- **Ý tưởng**: phân cụm dựa trên mật độ, gom các điểm gần nhau (có khoảng cách nhỏ hơn ε) và có mật độ cao (số điểm tối thiểu trong cụm là $minPts$)
- Các điểm nằm trong các cụm mật độ thấp được gán là nhiễu (noise)

epsion =1 thì xa tí cũng xem là láng giềng



tối thiểu 8 mới gom



0.6 thì hơi cách tí

tối thiểu 6 thì thành 1 cụm

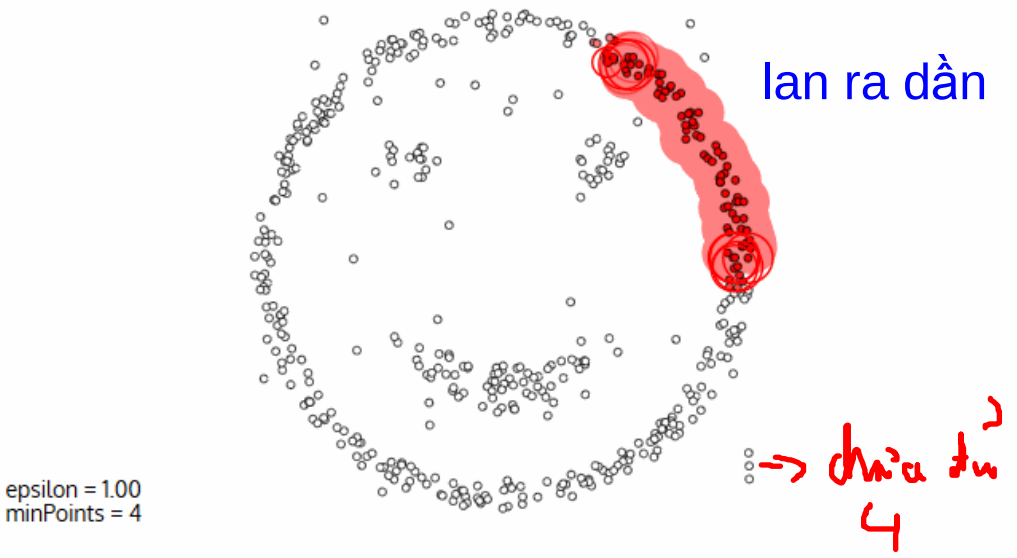
bên đây thì khắc khe hơn về khoảng cách nhưng lại dễ dãi về số phần tử để trở thành cụm, khiến cho nhiều cụm phân bố gần-> tạo ra nhiều cụm hơn

Thuật toán gom nhóm - DBSCAN

những dạng không chỉ hình cầu vẫn xử lý tốt

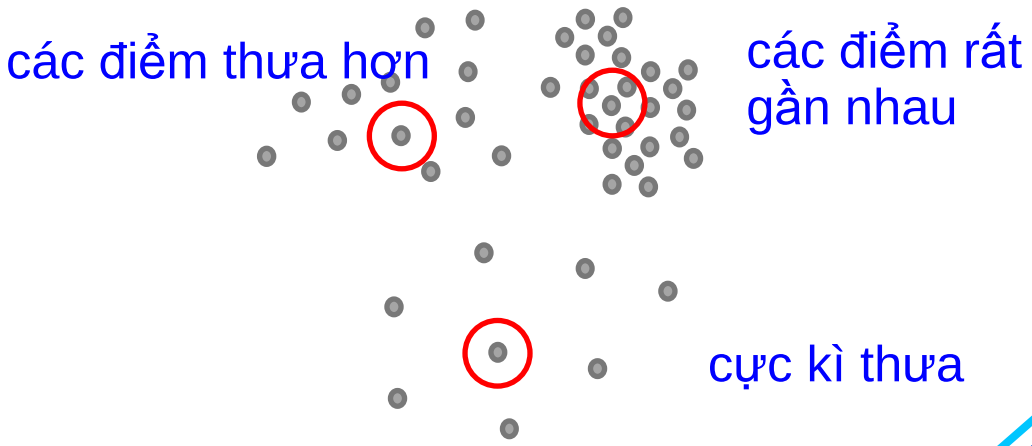
- **Ưu điểm**:
  - Không cần biết trước số cụm
  - Hiệu quả với dữ liệu mật độ cao
- **Khuyết điểm**:
  - Không hiệu quả khi dữ liệu có mật độ biến động
  - Phải chọn tham số ε và $minPts$

mật độ không đồng đều dù có thể là 1 cluster



Nguồn: https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/

do epsilon là con số cố định nên khi mật độ có biến động epsilon không cập nhật theo vậy có thể cập nhật epsilon và minPts để hoàn thiện hơn thuật toán

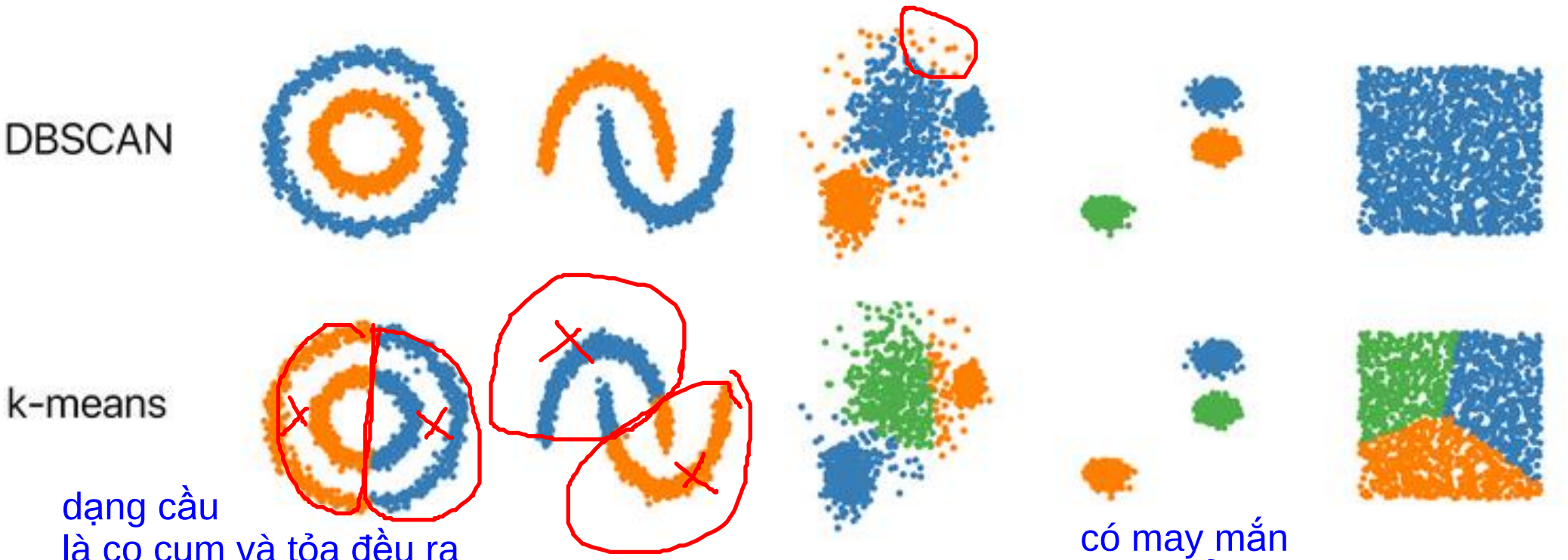


Tham số ε cố định không hiệu quả khi số điểm giảm

So sánh K-Means và DBSCAN

- So sánh K-Means với DBSCAN khi gom nhóm trên các “toy example”:

min point nếu đủ lớn thì đây thành nhiễu



phải để chung thành 1 cụm

dạng cầu
là co cụm và tỏa đều ra

có may mắn
khi 3 điểm tâm
được tách ra hơi xa
để gom nhóm đúng

Nguồn: scikit-learn.org

NỘI DUNG

1. GIỚI THIỆU HỌC KHÔNG GIÁM SÁT

2. CÁC MÔ HÌNH GOM NHÓM - CLUSTERING

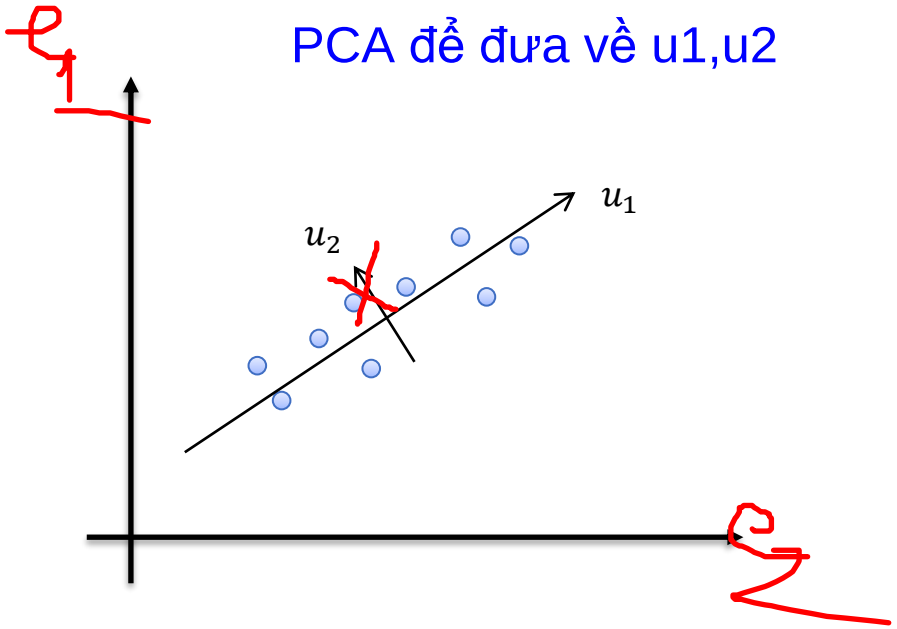
3. CÁC MÔ HÌNH GIẢM CHIỀU DỮ LIỆU

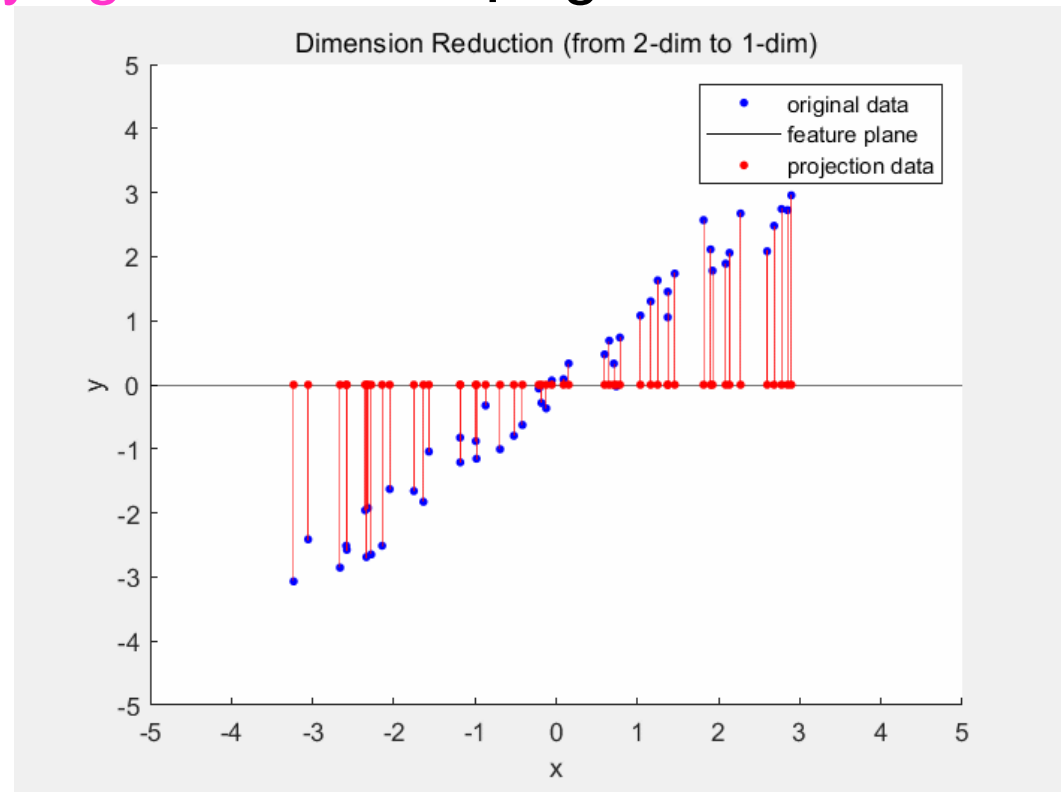
Bài toán 2: Giảm chiều dữ liệu

không được là vector ít chiều hơn nhưng ngẫu nhiên thì vô nghĩa

- **Giảm chiều dữ liệu**: là quá trình chuyển đổi dữ liệu từ không gian đa chiều sang không gian ít chiều sao cho biểu diễn không gian ít chiều vẫn giữ được một số tính chất có ý nghĩa của dữ liệu gốc

PCA để đưa về u1,u2





Trực quan hóa thuật toán PCA

Nguồn: https://www.programmersought.com/article/63044502718/

Bài toán 2: Giảm chiều dữ liệu

- Một số thuật toán giảm chiều dữ liệu:
  - Phân tích thành phần chính – PCA
  - Nhúng t-SNE trực quan hóa dữ liệu

không gian N chiều đưa về 2D hoặc 3D thì phải vẫn giữ tính chất nhất định của không gian gốc

- Mỗi phương pháp có những đặc điểm và ứng dụng riêng, tùy thuộc vào bộ dữ liệu và mục tiêu cụ thể

Thuật toán Principal Component Analysis - PCA

- **Ý tưởng**: Cho bảng dữ liệu điểm của một lớp học như sau, bạn có nhận xét gì?

đều trải đòng đều, điểm văn lại giống nhau

xấp xỉ vì thêm 1 vùng dữ liệu cho số 7

Giảm được $\approx \frac{1}{4} = 25\%$ dữ liệu

Họ tên	Giới tính	Toán	Văn
Họ tên 1	Nam	6.5	7
Họ tên 2	Nữ	7.5	7
Họ tên 3	Nam	9.0	7
Họ tên 4	Nữ	6.0	7
Họ tên 5	Nữ	9.5	7

Độ lệch bằng 0 nên có thể xóa cột này, **chỉ cần lưu điểm chung là 7**

Thuật toán Principal Component Analysis - PCA

- **Ý tưởng**: Cho bảng dữ liệu điểm của một lớp học như sau, bạn có nhận xét gì? Ví dụ khác.

Họ tên	Giới tính	Toán	Văn
Họ tên 1	Nam	6.5	7.0
Họ tên 2	Nữ	7.5	6.5
Họ tên 3	Nam	9.0	7.0
Họ tên 4	Nữ	6.0	7.0
Họ tên 5	Nữ	9.5	7.5

Độ lệch bằng ~0.32 nên vẫn có thể xóa cột này, **chỉ lưu điểm TB là 7**

đánh đổi là không khôi phục dữ liệu gốc, do vậy độ biến động thấp mới bỏ

vẫn bé hơn 0.5 là điểm lệch trong nhập điểm phổ thông nên vẫn xóa cột

Thuật toán Principal Component Analysis - PCA

- **Ý tưởng**: Cho bảng dữ liệu điểm của một lớp học như sau, bạn có nhận xét gì? Ví dụ khác.

Vẫn giảm được $\approx \frac{1}{4} = 25\%$ dữ liệu

Nhưng không khôi phục được dữ liệu gốc

Họ tên	Giới tính	Toán	Văn
Họ tên 1	Nam	6.5	7.0. 7
Họ tên 2	Nữ	7.5	6.5
Họ tên 3	Nam	9.0	
Họ tên 4	Nữ	6.0	
Họ tên 5	Nữ	9.5	

Chiều dữ liệu nào ít biến động →có thể loại bỏ

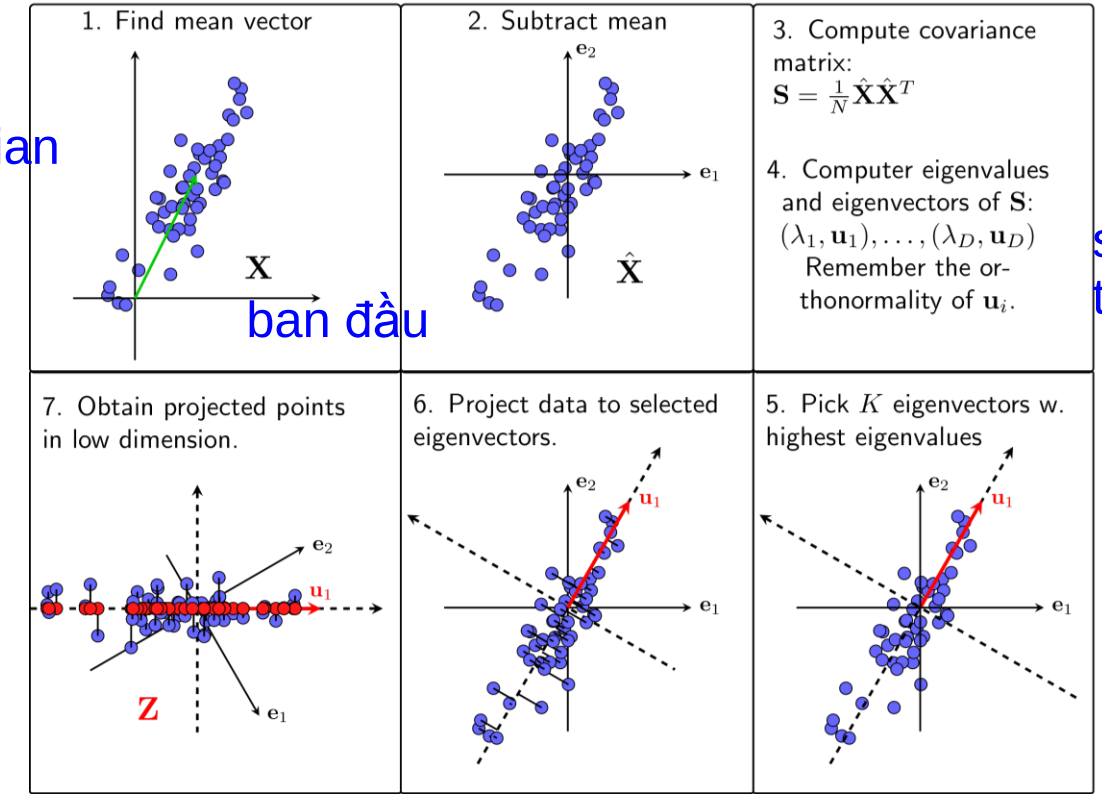
biến đông cao -> có nhiều thông tin

Độ lệch bằng ~0.32 nên vẫn có thể xóa cột này, **chỉ lưu điểm TB là 7**

Thuật toán Principal Component Analysis - PCA

- PCA tìm một không gian con tuyến tính mới mà dữ liệu được biểu diễn một cách hiệu quả nhất
- Trong không gian mới này, các chiều được chọn sao cho tối đa hóa phương sai của dữ liệu

-mean là tịnh tiến không gian



ban đầu

sắp xếp lại theo giá trị luôn

cặp lambda, u là cho biết độ biến động với từng trục u mới nên phải xếp

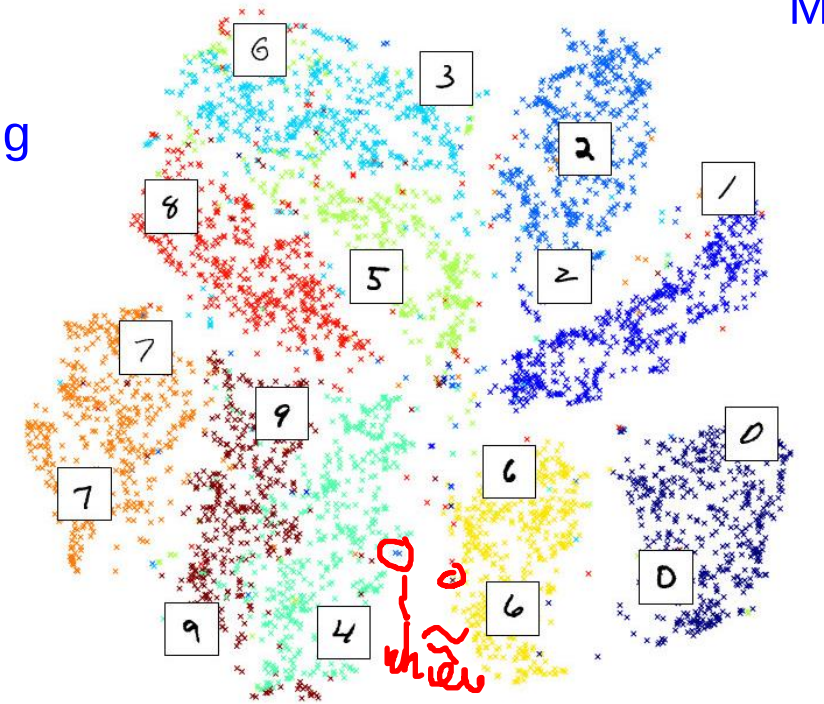
bỏ u nào thì lấy giá trị trung bình nó ra mà ta đã -mean dời gốc nên mean =0

Nguồn: https://machinelearningcoban.com/2017/06/15/pca/

tSNE (t-Distributed Stochastic Neighbor Embedding)

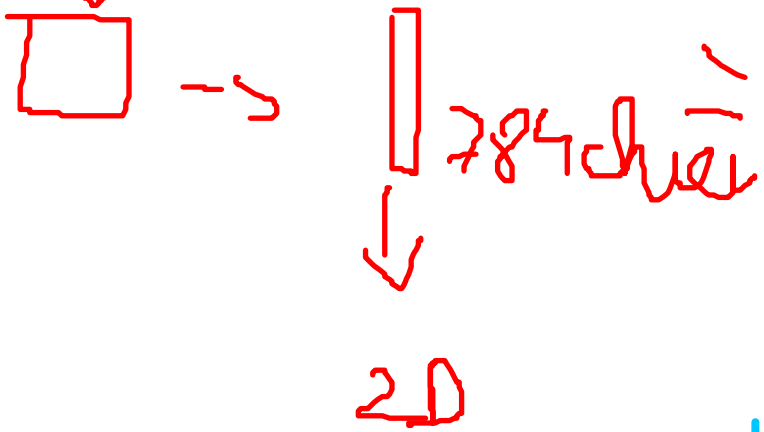
- **t-SNE**: là kỹ thuật giảm chiều phi tuyến tính, để trực quan hóa dữ liệu đa chiều trong không gian có số chiều thấp hơn (thường là 2D hoặc 3D)
- **Ý tưởng**: tạo ra một phân phối xác suất tương tự trong không gian có số chiều thấp hơn

10 chiều chiếu xuống 2 chiều
thì các điểm gần nhau trong 10D cũng gần nhau trong 2 chiều



Trực quan hóa tập MNIST sử dụng tSNE

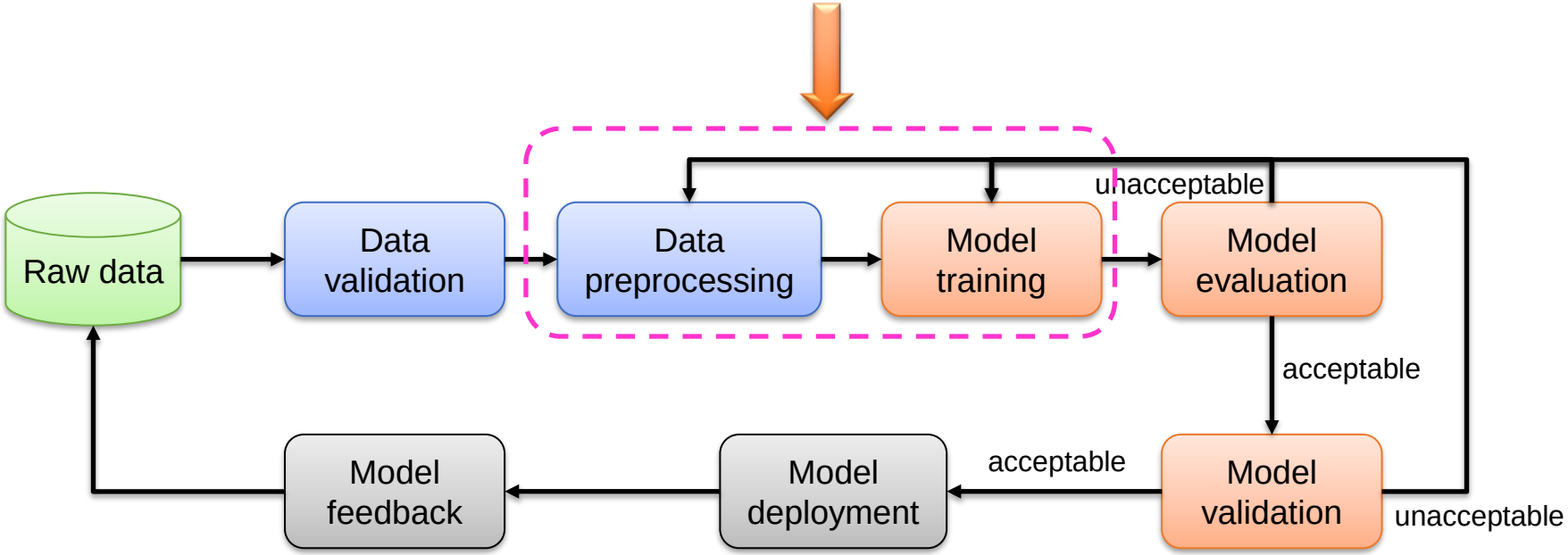
MNIST mỗi ảnh có kích thước 28x28



So sánh PCA và tSNE

	PCA	tSNE
Ưu điểm	- Nhanh, hiệu quả với dữ liệu lớn do thực hiện biến đổi tuyến tính - Giữ lại các thành phần chính với phương sai lớn độ lệch/ biến động lớn	- Có khả năng bắt cấu trúc phi tuyến - Thường được sử dụng để trực quan hóa dữ liệu
Khuyết điểm	- Khả năng bắt cấu trúc phi tuyến kém - Dễ bị ảnh hưởng bởi nhiễu (outlier)	- Không ổn định do yếu tố ngẫu nhiên - Độ phức tạp cao - Khó giải thích hơn so với PCA

Tổng kết – Vị trí của bài học



BÀI QUIZ VÀ HỎI ĐÁP