

# Thuật toán phân cụm dữ liệu K-Means áp dụng trong việc phân loại khách hàng

Nhóm 01

📖 Giảng viên : Trịnh Tấn Đạt

# Thành viên trong nhóm



**Nguyễn Minh Thông**

MEMBER

3118410416

lonelynut2k@gmail.com



**Nguyễn Hữu Thắng**

MEMBER

3118410402

jandragon113@gmail.com

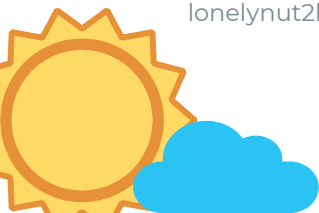


**Trần Huy Khánh**

MEMBER

3118410191

jasontran786@gmail.com





## CÁC PHẦN CHÍNH



**01**

### MACHINE LEARNING

Sơ lược về Machine Learning



**02**

### NGÔN NGỮ LẬP TRÌNH

Python và các ứng dụng



**03**

### PHÂN LOẠI KHÁCH HÀNG

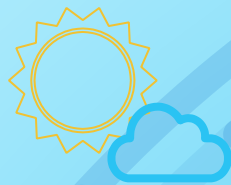
Các phương pháp phân loại



**04**

### THUẬT TOÁN K-MEANS

Và các cải tiến

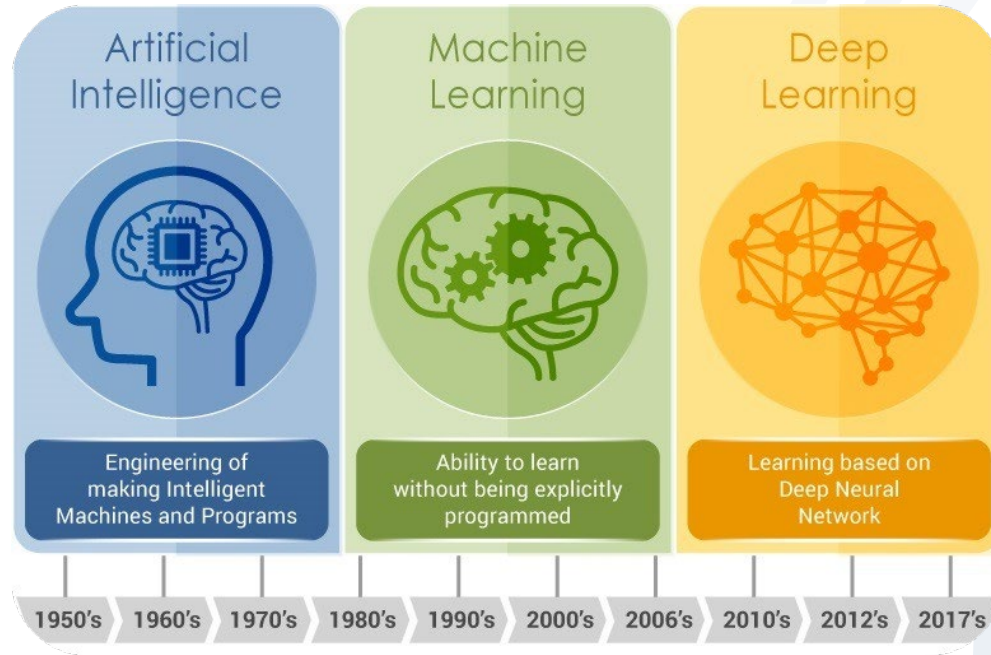


01.

# Machine Learning



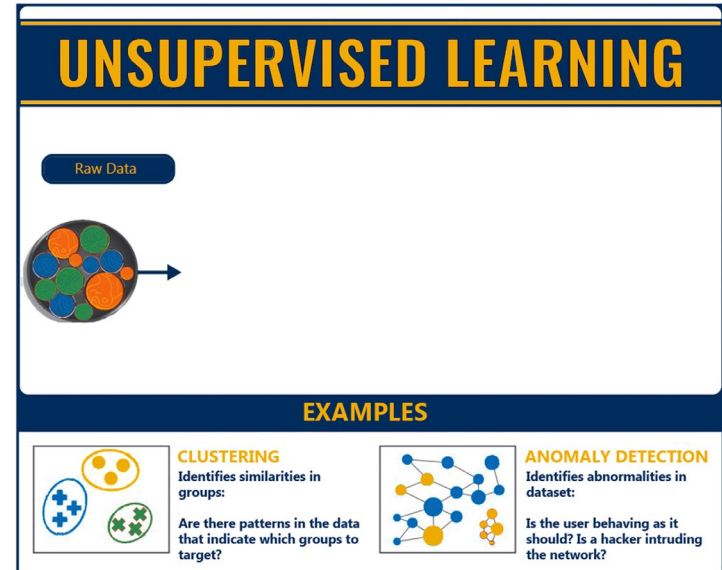
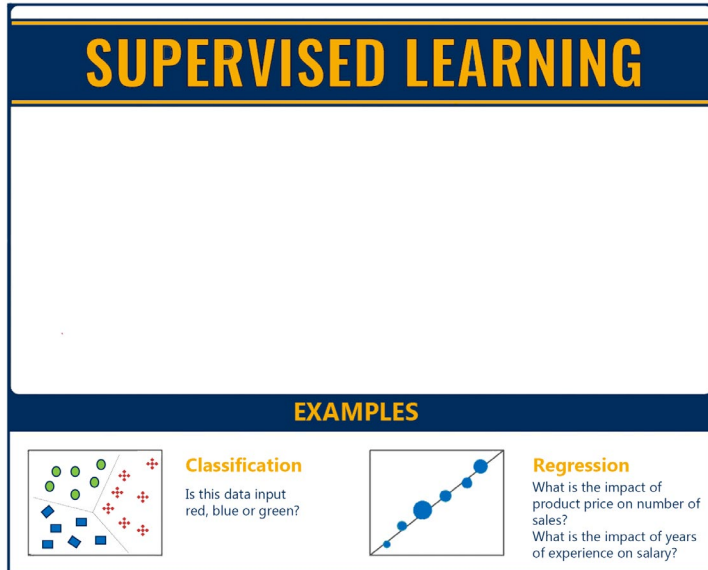
- Là việc tạo ra các cỗ máy có trí thông minh, suy nghĩ tương tự con người để có thể đưa ra những kết quả, quyết định đúng đắn nhất như giải các bài toán, dự báo thời tiết,...



# Phân loại

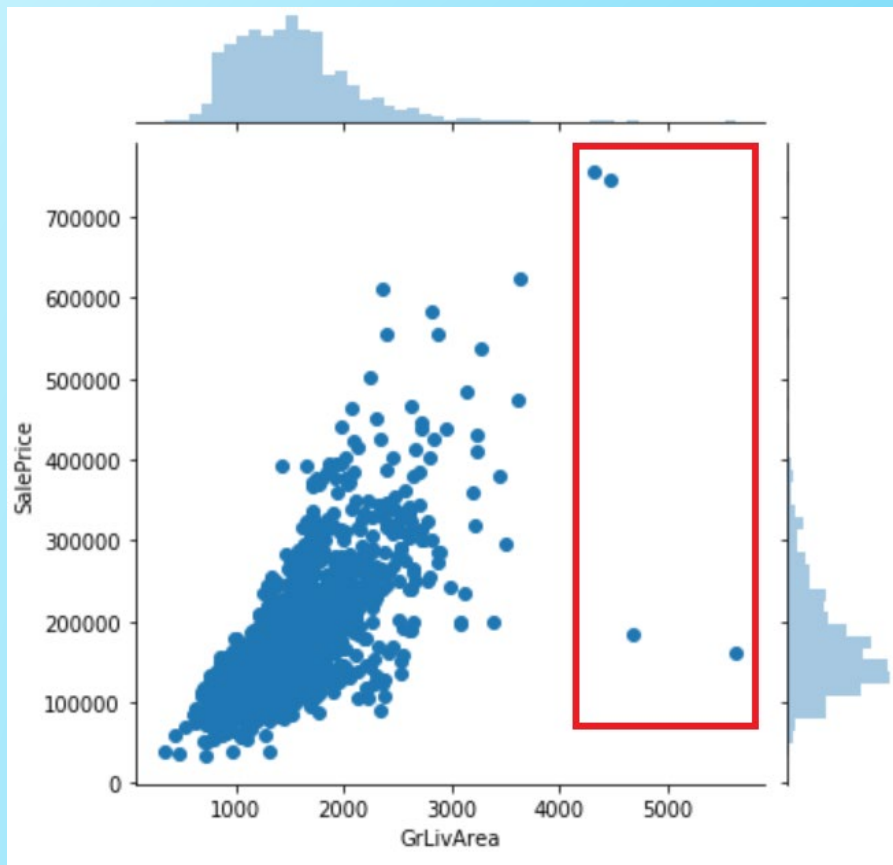
- Gồm 2 loại chính là :
  - Supervised Learning (Học có giám sát)

- Unsupervised Learning (Học không giám sát)



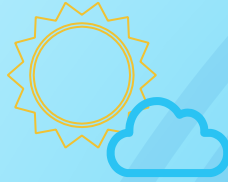


- Học máy vẫn đang còn ở những giai đoạn đầu nên vẫn cần sự can thiệp của con người trong việc tìm hiểu cơ sở dữ liệu và lựa chọn các kỹ thuật phù hợp để phân tích dữ liệu nhằm đem lại kết quả tối ưu nhất.
- Trước khi bắt đầu quá trình học thì dữ liệu phải được làm sạch, không có sai lệch và không có dữ liệu giả (hay còn gọi là loại bỏ các outliers).



Các dữ liệu ngoại lai (hiển thị trong khung màu đỏ)





02.

# Ngôn ngữ lập trình



# Python



- Là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ và phổ biến nhất hiện nay với mục đích lập trình đa năng.
- Python được tạo ra bởi Guido van Rossum vào năm 1980 và lần đầu ra mắt công chúng vào tháng 2 / 1991. Cái tên “Python” cũng được tác giả đặt theo một chương trình hài vào cuối những năm 1970.

# Ưu điểm

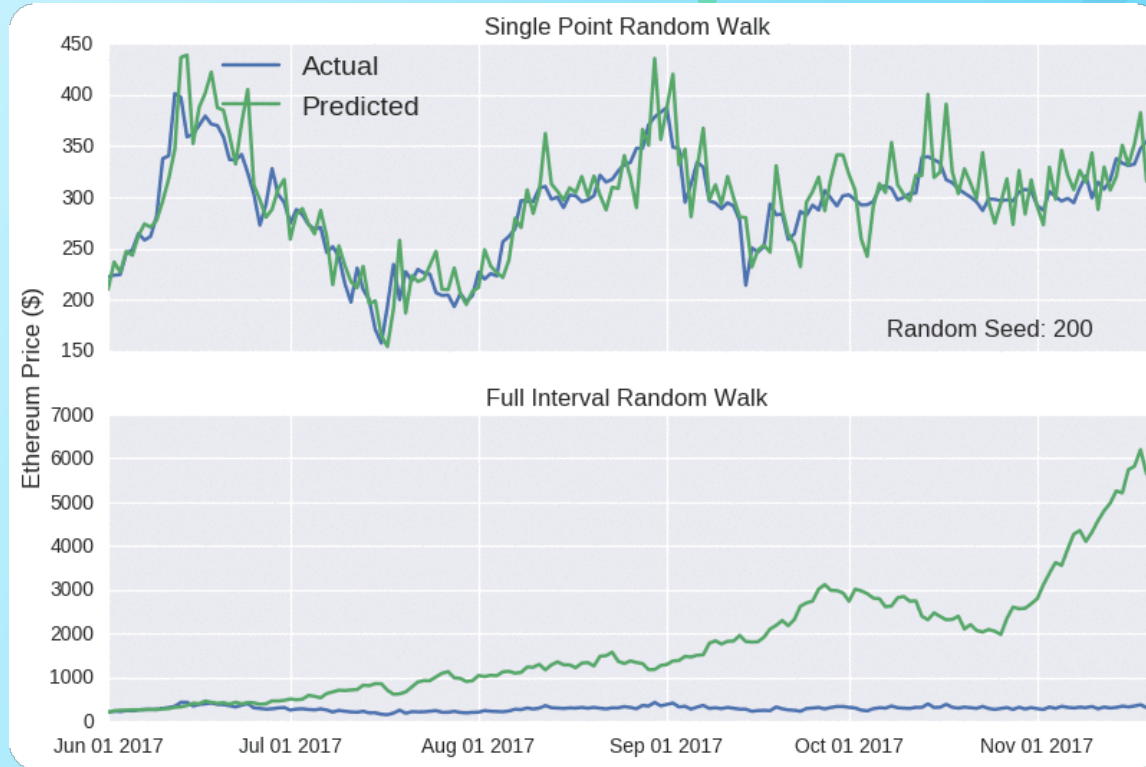


- Không cần khai báo kiểu dữ liệu do sử dụng cơ chế cấp phát bộ nhớ tự động.
- Cú pháp lệnh rõ ràng, dễ đọc, học và nhớ
- Cung cấp hầu như đầy đủ các thư viện cũng như các hàm dựng sẵn.
- Cộng đồng hỗ trợ rộng lớn.

# Ứng dụng

- Hỗ trợ cho lập trình ứng dụng web thông qua các framework như Django, Flask, Pyramid.
- Kết hợp với các nền tảng phần cứng như Raspberry Pi tạo ra các cánh tay Robot được sử dụng trong công nghiệp.
- Thường dùng để phân tích dữ liệu cũng như làm về khoa học và số liệu ứng dụng.

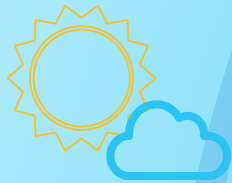




Mô hình dự đoán giá Ethereum sử dụng ngôn ngữ  
Python

03.

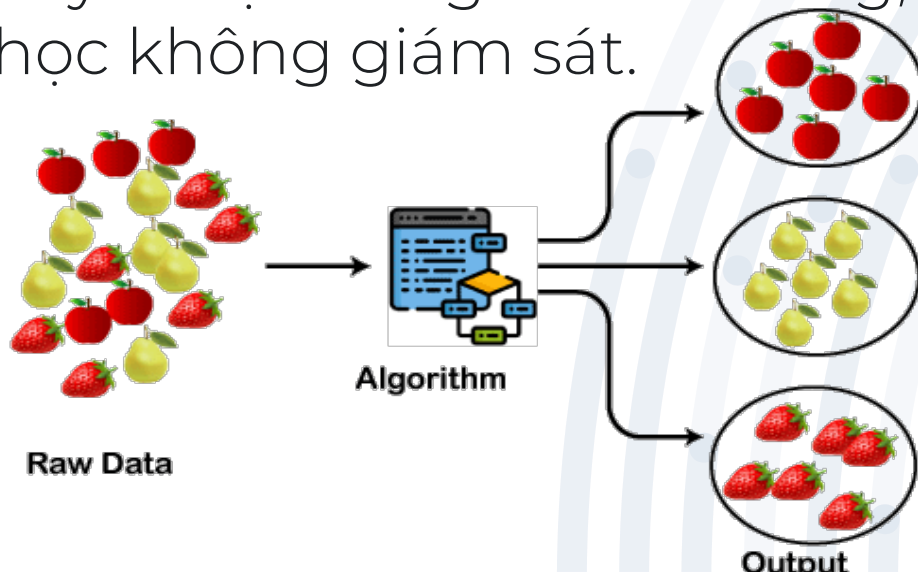
# Phân loại khách hàng



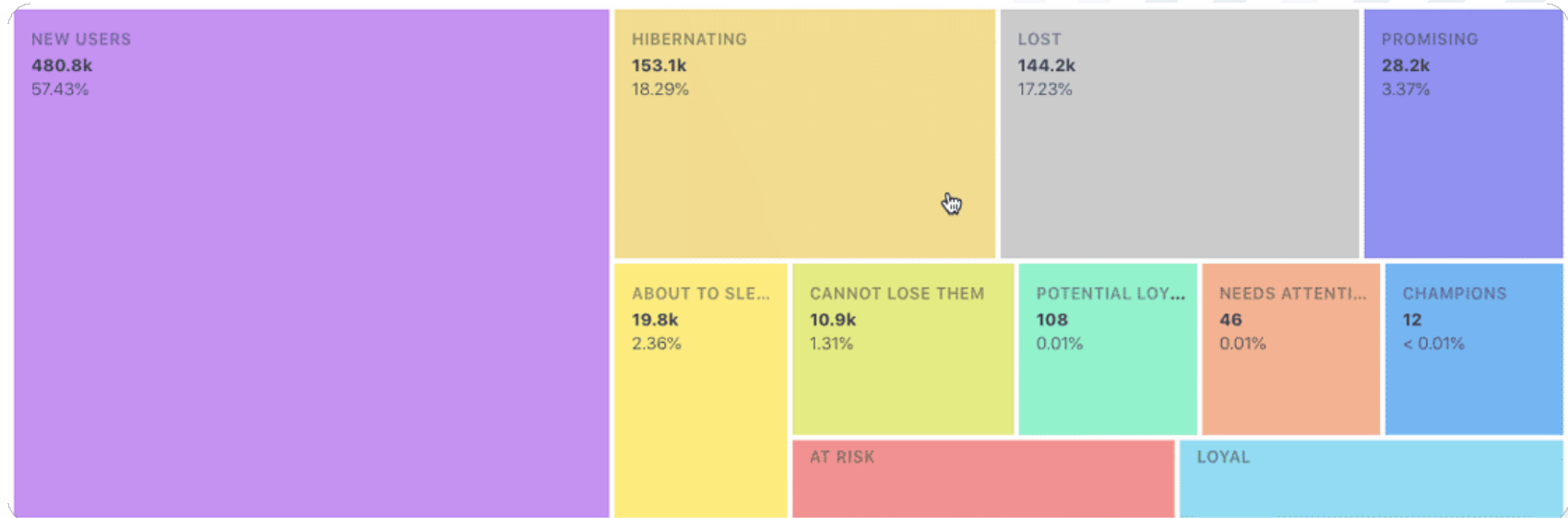


# Phân loại khách hàng

- Là tách tập dữ liệu đầu vào thành các cụm dữ liệu mà những khách hàng trong đó sẽ giống nhau ở một số tiêu chí nào đó so với những khách hàng nằm ở cụm khác.
- Đây là 1 kỹ thuật trong Data Mining, thuộc nhóm học không giám sát.

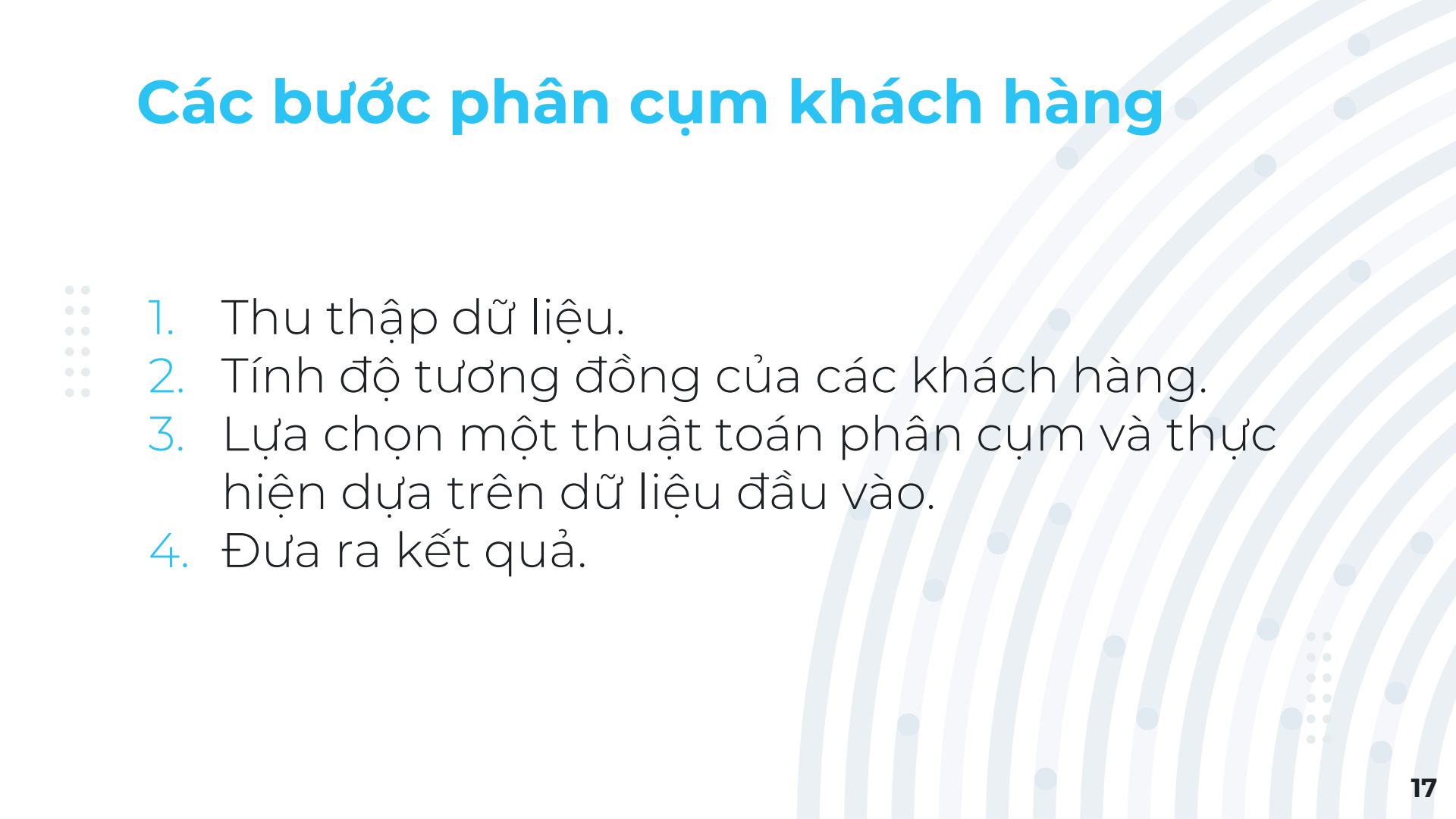



- Việc phân loại và xác định được các tệp khách hàng là yếu tố hàng đầu ảnh hưởng đến việc hoạt động của doanh nghiệp.
- Giúp đưa ra các ưu đãi giảm giá, chương trình khuyến mãi phù hợp đến với khách hàng.





# Các bước phân cụm khách hàng

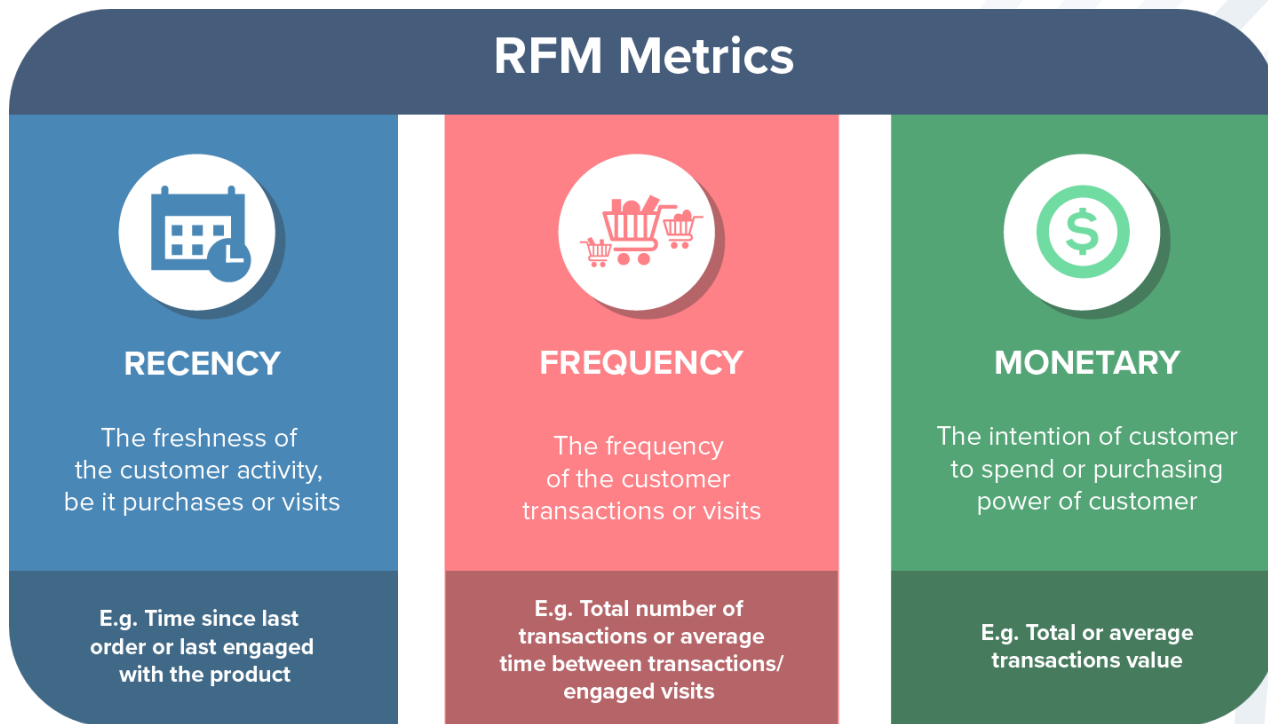
- 
- 
1. Thu thập dữ liệu.
  2. Tính độ tương đồng của các khách hàng.
  3. Lựa chọn một thuật toán phân cụm và thực hiện dựa trên dữ liệu đầu vào.
  4. Đưa ra kết quả.

# Tại sao cần sự hỗ trợ của Machine Learning ?

- Dữ liệu khách hàng của các trung tâm mua sắm, các sàn giao dịch thương mại điện tử,... thường có dung lượng rất lớn vượt ngoài khả năng tính toán của con người.
- Machine Learning sẽ giúp chia khách hàng thành nhiều nhóm khác nhau từ tập dữ liệu ban đầu.
- Giúp tiết kiệm thời gian, công sức, chi phí,...

# Các phương pháp phân loại phổ biến

- Recency, Frequency, Monetary Value (RFM)



# Các phương pháp phân loại phổ biến

- Customer Relationship Management (CRM)

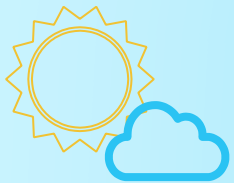


# Các phương pháp phân loại phổ biến

- Finance Resource Management (FRM)



# 04. Thuật toán K-Means



# K-Means



- Là 1 thuật toán phân cụm dữ liệu.
- Ý tưởng chung của K-means là xác định vị trí tâm cụm và xếp đối tượng vào cụm mà có tâm cụm gần nó nhất dựa trên khoảng cách Euclidean hay khoảng cách Manhattan.

# Công thức tính K-Means

$$||x_i - m_k||_2^2$$

- Trong đó :
  - $x_i$  : Điểm dữ liệu được phân vào cụm gần nhất
  - $m_k$  : Phần tử trung tâm của cụm đó



# Độ phức tạp của K-Means

$$O((3nkd)7T^{\text{flop}}))$$

- Trong đó :
  - $n$  : Số đối tượng dữ liệu
  - $k$  : Số cụm dữ liệu
  - $d$  : Số chiều của bài toán
  - $7$  : Số vòng lặp
  - $T^{\text{flop}}$  : Thời gian thực hiện 1 phép tính cơ sở (nhân, chia, cộng,...)

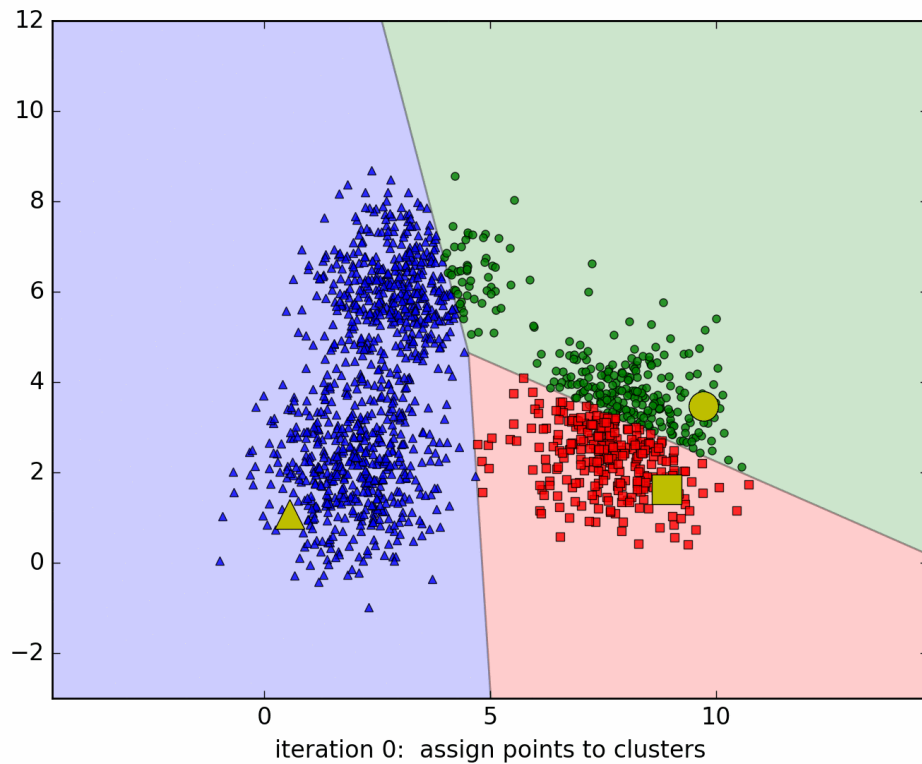
# Công thức tính khoảng cách Euclidean trong K-Means

$$\partial_{ji} = \sqrt{\sum_{s=1}^m (x_{is} - x_{js})^2}$$

- Trong đó :
  - $\partial_{ji}$  : Khoảng cách từ ai tới  $c_j$
  - $x_{is}$  : Thuộc tính thứ  $s$  của đối tượng ai
  - $x_{js}$  : Thuộc tính thứ  $s$  của phần tử trung tâm  $c_j$

# Các bước của thuật toán K-Means

1. Xác định K – số cụm (ngẫu nhiên hoặc nhờ sự giúp đỡ của các thuật toán khác).
2. Chọn K điểm ngẫu nhiên làm tâm cụm.
3. Gán từng điểm vào cụm có tâm gần nó nhất dựa theo khoảng cách hình học Euclidean hay Manhattan.
4. Tính phương sai và đặt lại tâm cụm mới (tâm cụm được tính bằng trung bình cộng tọa độ các thành phần trong cụm).
5. Nếu tâm cụm không đổi (điều kiện hội tụ được thỏa mãn) thì kết thúc thuật toán. Nếu không thì lặp lại bước 3.

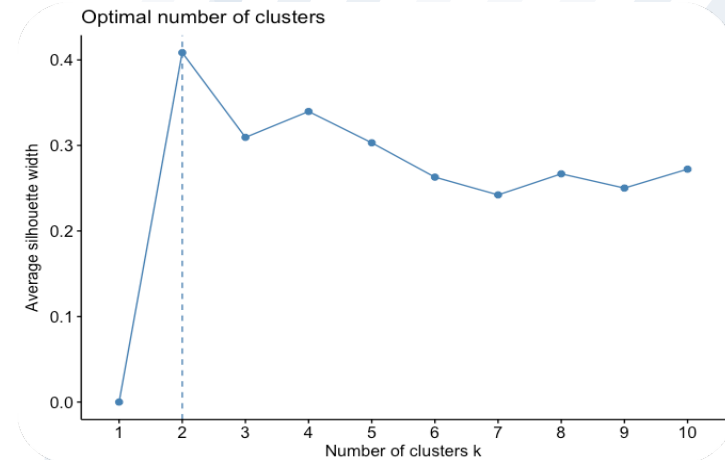
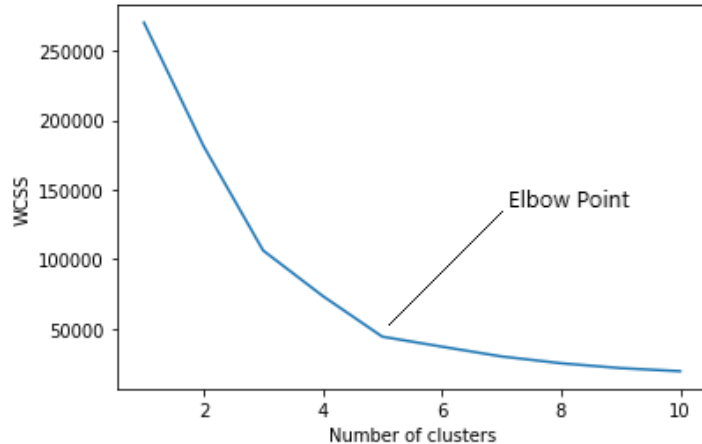


Thuật toán K-Means tính khoảng cách theo  
Euclidean



# Các cải tiến của thuật toán K-Means

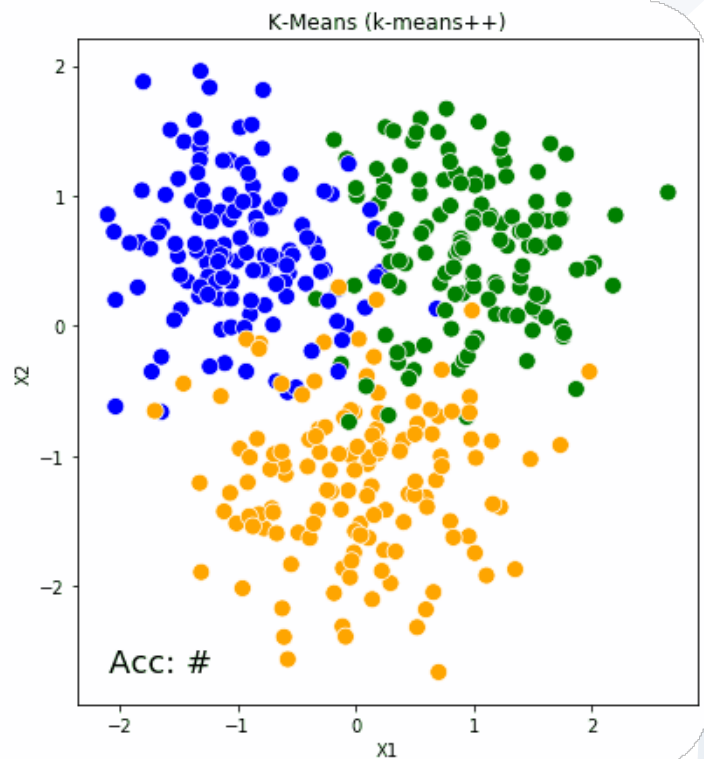
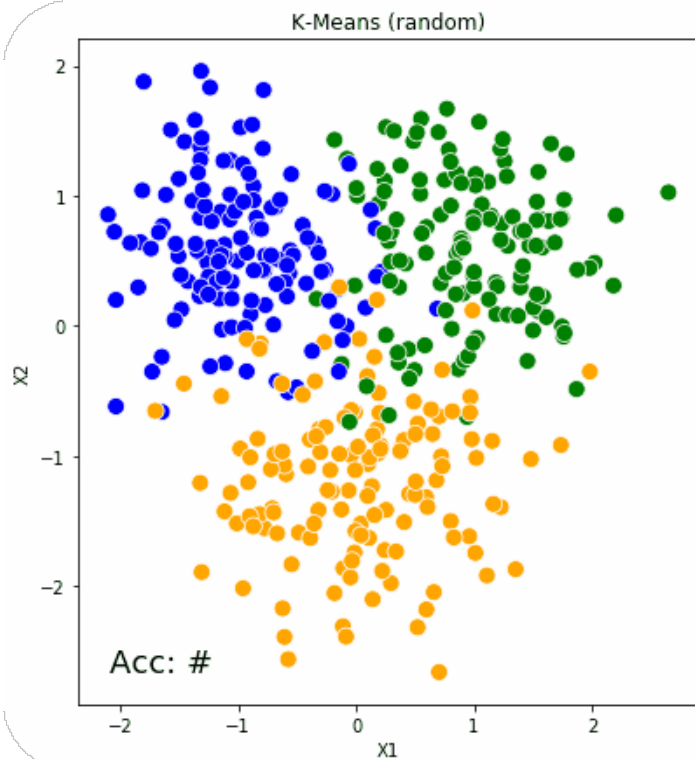
- Tìm số cụm
- Thuật toán Elbow :
  - Dựa vào sự suy giảm của hàm biến dạng .
- Thuật toán Silhouette :
  - Đo lường mức độ tương tự của một điểm với các điểm lân cận gần nhất, giá trị  $\in [-1, 1]$ .





# Các cải tiến của thuật toán K-Means

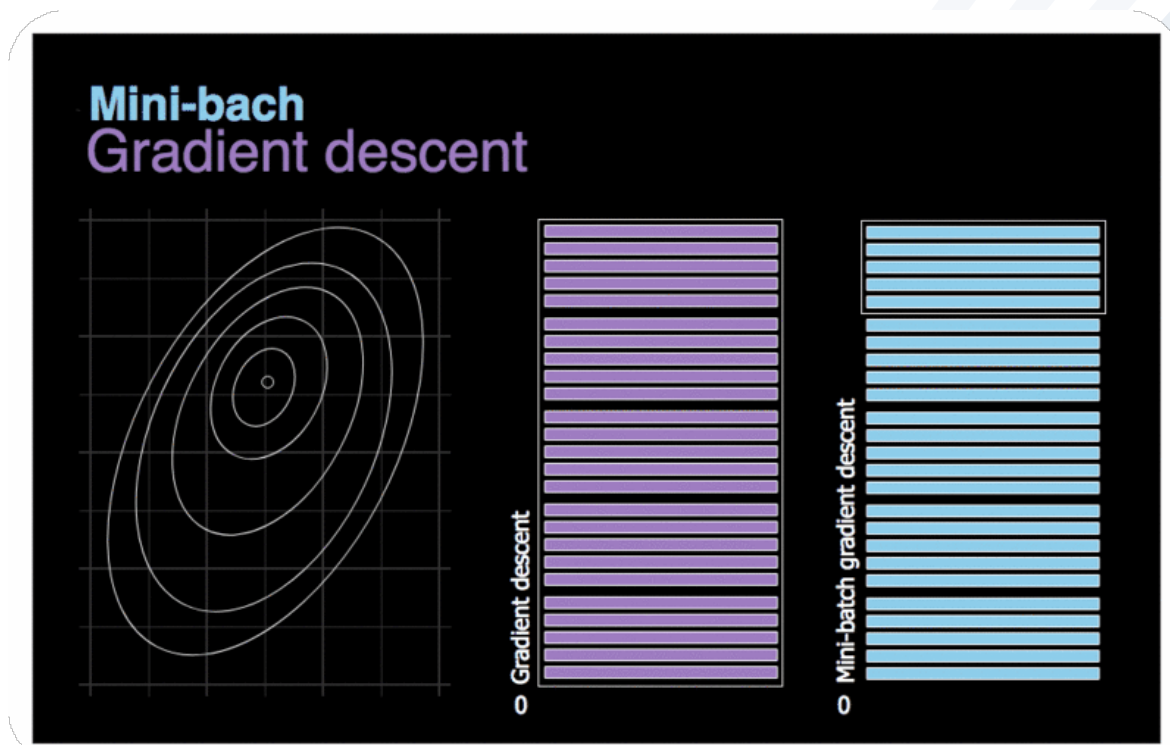
- Chọn K điểm làm tâm cụm :
  - Thuật toán Kmean++





# Các cải tiến của thuật toán K-Means

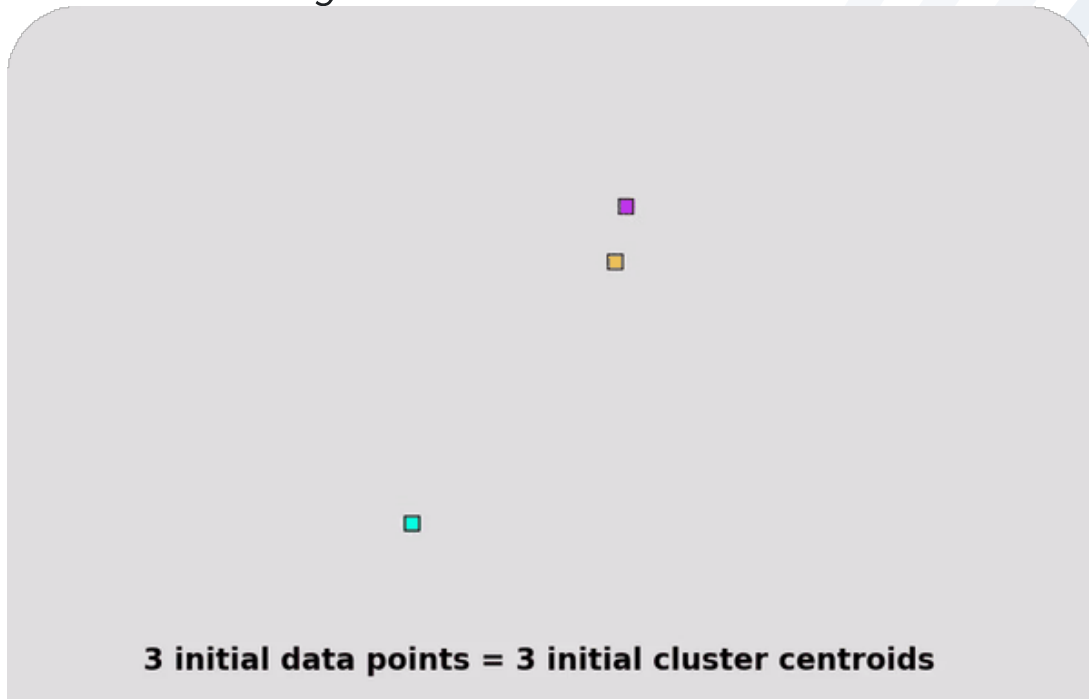
- Chia nhỏ tập dữ liệu đầu vào :
  - Mini Batch K-Means





# Các cải tiến của thuật toán K-Means

- Cập nhật K tâm cụm với mỗi dữ liệu đầu vào :
  - Online Learning K-Means : Phù hợp với bộ dữ liệu rất lớn và yêu cầu tính real-time của dữ liệu.







**Cảm ơn thầy và các bạn  
đã lắng nghe !**

