

Bảo vệ khóa luận: TÌM HIỂU VÀ VẬN DỤNG PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU LỚN



GVHD : Th.S Cao Mạnh Hùng
Sinh Viên : Trịnh Đình Phúc



Nội dung báo cáo

1 Giới thiệu

- Dữ liệu lớn
- Làm thế nào để khai thác dữ liệu lớn?

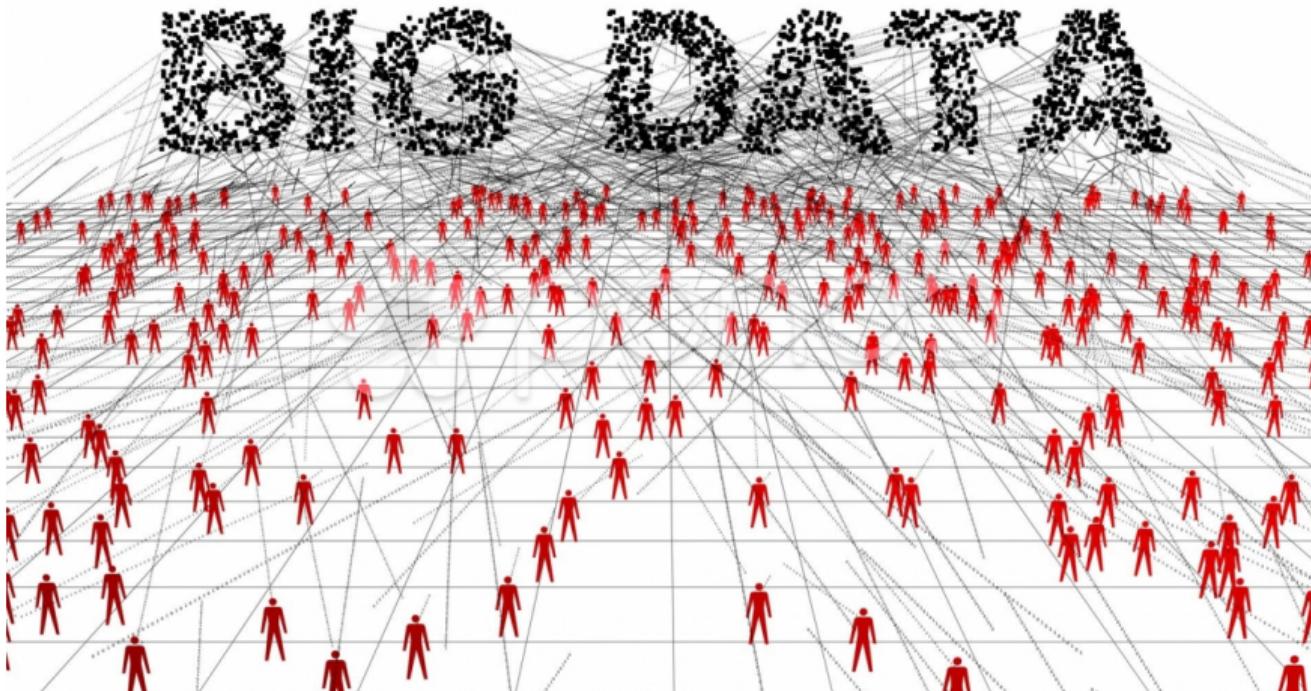
2 Các bước khai thác dữ liệu

- Thu thập dữ liệu (Data collection)
- Chuẩn bị dữ liệu (Data preparation)
- Mô hình hóa (Modeling)
- Đánh giá mô hình (Model Evaluation)

3 Demo áp dụng quy trình xử lý dữ liệu lớn



Dữ liệu lớn là gì?





Dữ liệu lớn

Thống kê:



Dữ liệu lớn

Thống kê:

- 25+ TB dữ liệu được tạo ra mỗi giây trên toàn cầu.



Dữ liệu lớn

Thống kê:

- 25+ TB dữ liệu được tạo ra mỗi giây trên toàn cầu.
- 90+ % dữ liệu của thế giới được tạo ra trong 2 năm vừa qua.



Dữ liệu lớn

Thông kê:

- 25+ TB dữ liệu được tạo ra mỗi giây trên toàn cầu.
- 90+ % dữ liệu của thế giới được tạo ra trong 2 năm vừa qua.
- 90+ % dữ liệu được tạo ra là dữ liệu phi cấu trúc.



Các dạng dữ liệu lớn

Dữ liệu có ở các dạng sau:

<i>Kiểu dữ liệu</i>	<i>Ứng dụng trong việc khai thác</i>
Văn bản	Xử lý ngôn ngữ tự nhiên
Ảnh và Video	Thị giác máy tính
Âm thanh	Xử lý tín hiệu số
Social Network	Phân tích đồ thị
Business	Khai thác dữ liệu
DNA	Tin sinh học
...	...



Làm thế nào để khai thác dữ liệu lớn?





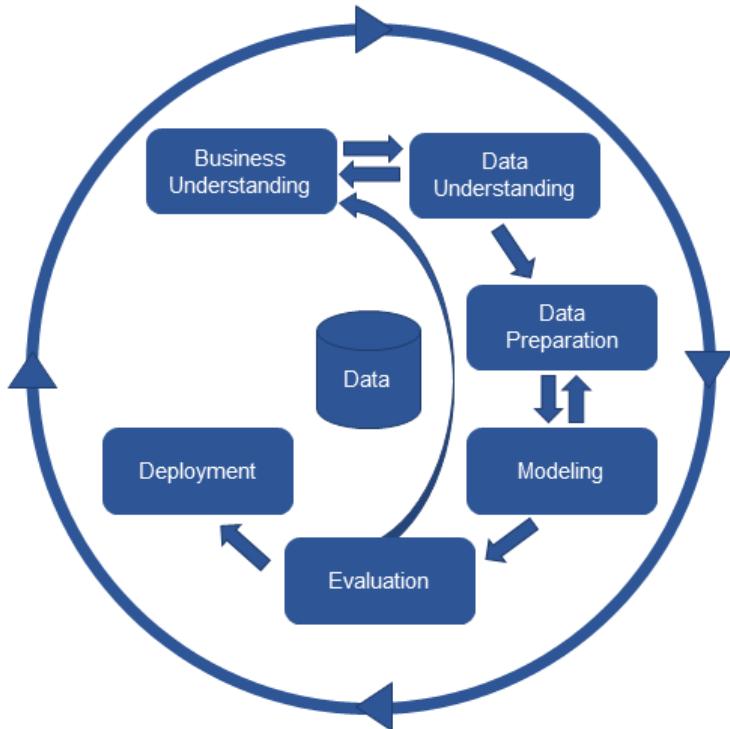
Làm thế nào để khai thác Dữ liệu lớn?

Machine Learning
(Máy học)





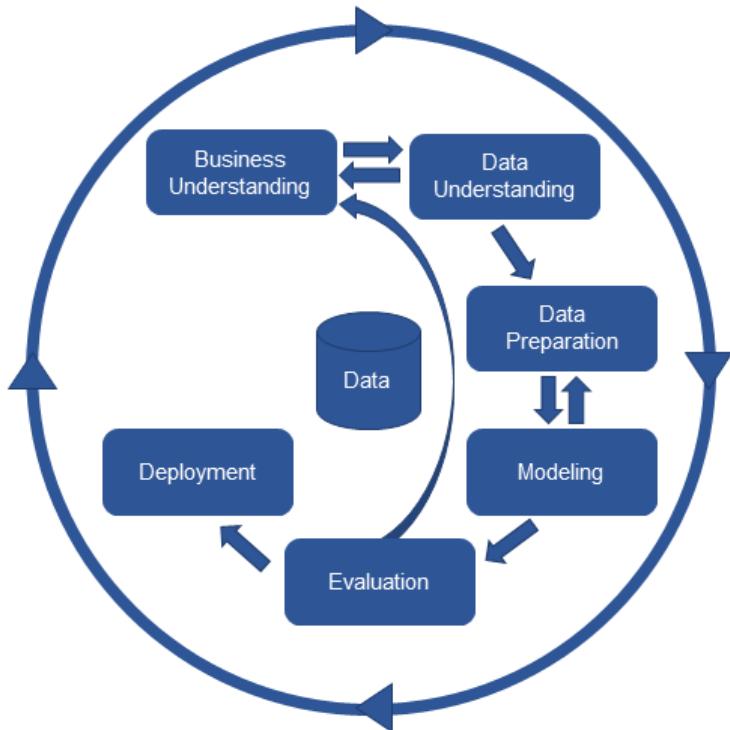
Các bước khai thác dữ liệu



Quy trình khai thác dữ liệu có 6 bước chính:



Các bước khai thác dữ liệu

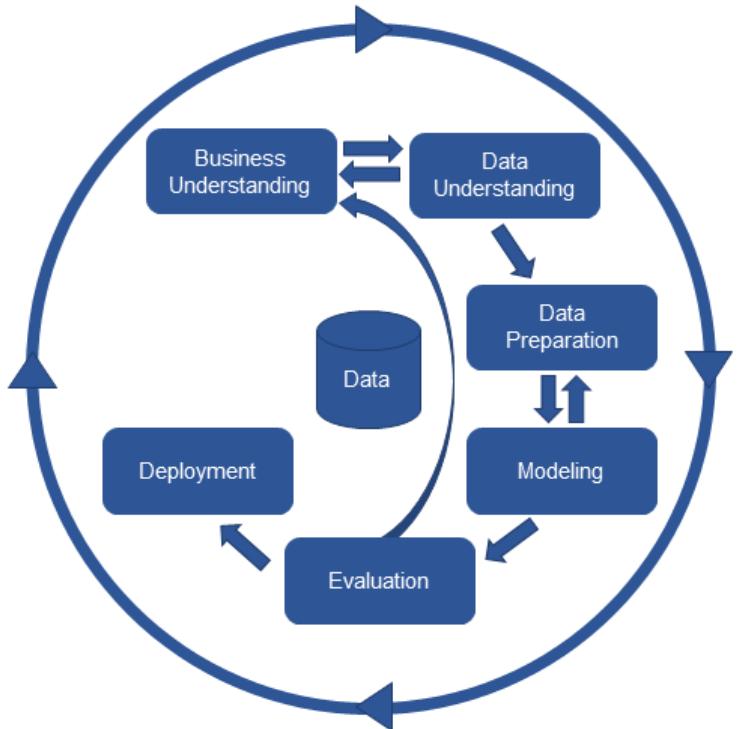


Quy trình khai thác dữ liệu có 6 bước chính:

- ① Tìm hiểu nghiệp vụ (Business understanding)



Các bước khai thác dữ liệu

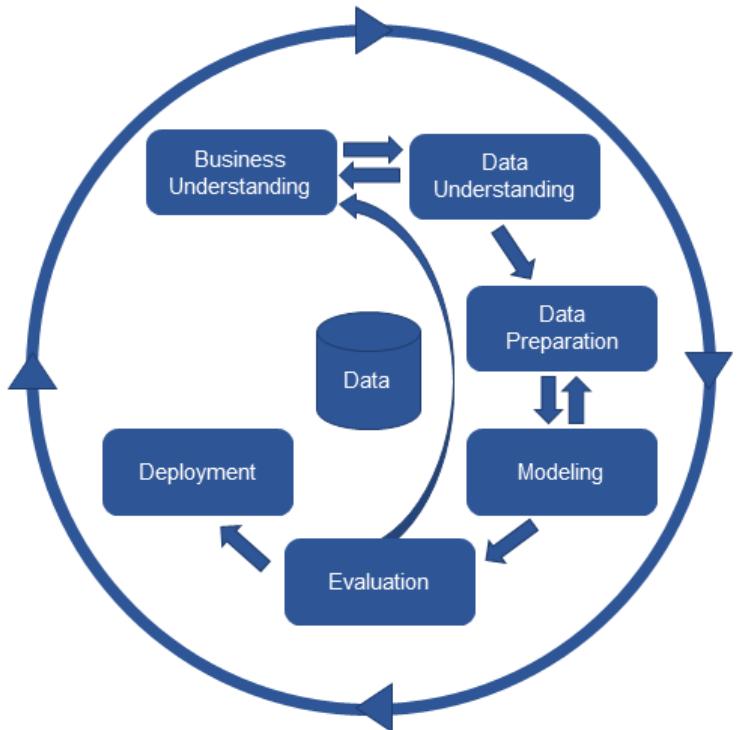


Quy trình khai thác dữ liệu có 6 bước chính:

- ① Tìm hiểu nghiệp vụ (Business understanding)
- ② Tìm hiểu dữ liệu (Data understanding)



Các bước khai thác dữ liệu

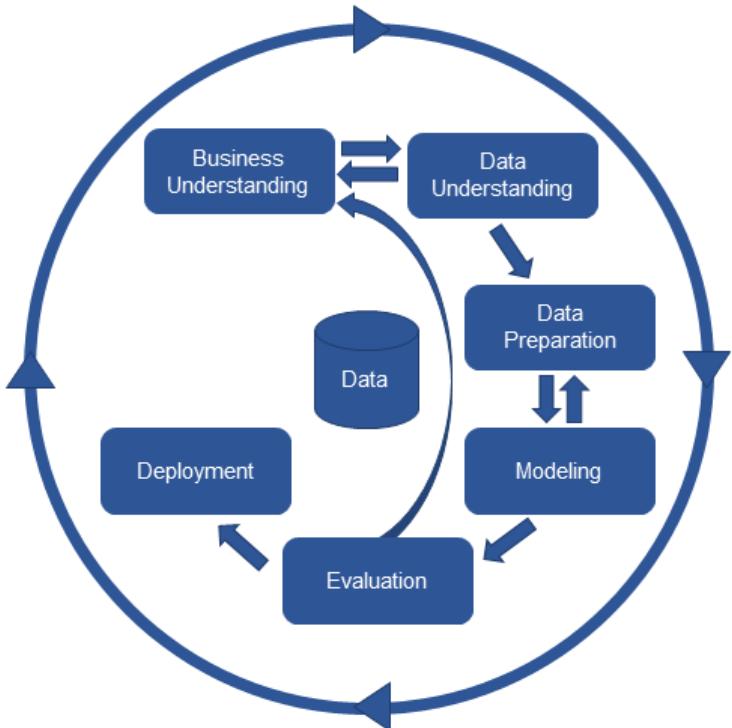


Quy trình khai thác dữ liệu có 6 bước chính:

- ① Tìm hiểu nghiệp vụ (Business understanding)
- ② Tìm hiểu dữ liệu (Data understanding)
- ③ Chuẩn bị dữ liệu (Data preparation)



Các bước khai thác dữ liệu

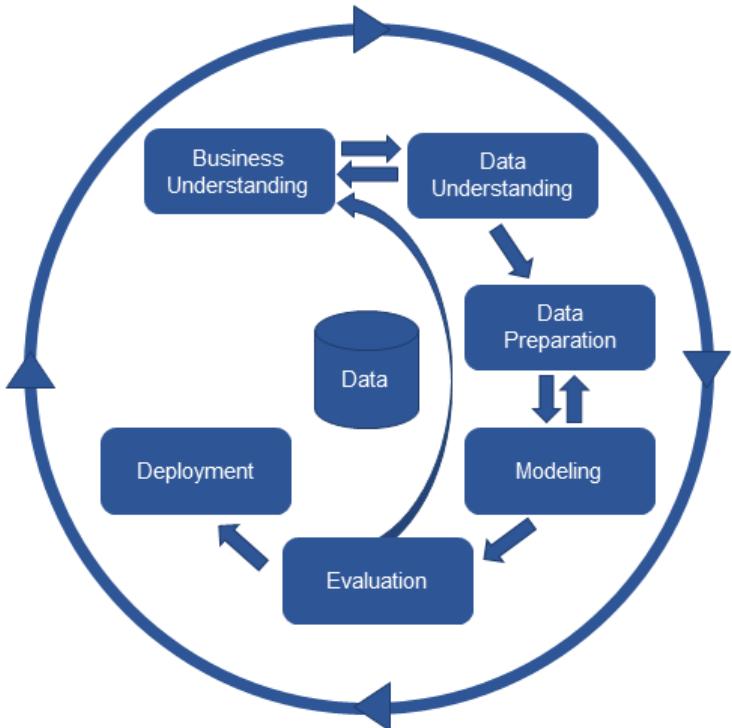


Quy trình khai thác dữ liệu có 6 bước chính:

- ① Tìm hiểu nghiệp vụ (Business understanding)
- ② Tìm hiểu dữ liệu (Data understanding)
- ③ Chuẩn bị dữ liệu (Data preparation)
- ④ Mô hình hóa (Modeling)
- ⑤ Dánh giá mô hình (Model Evaluation)



Các bước khai thác dữ liệu



Quy trình khai thác dữ liệu có 6 bước chính:

- ① Tìm hiểu nghiệp vụ (Business understanding)
- ② Tìm hiểu dữ liệu (Data understanding)
- ③ Chuẩn bị dữ liệu (Data preparation)
- ④ Mô hình hóa (Modeling)
- ⑤ Dánh giá mô hình (Model Evaluation)
- ⑥ Triển khai (Deployment)



Thu thập dữ liệu từ FIFA

		FIFA Index	News	Teams	Players	Squads	Contact	Sign In	English ▾
Club	Men's National	Women's National							
	Name	League	ATT	MID	DEF	OVR	Team Rating		
	Spain	Men's National	84	86	86	86			
	Brazil	Men's National	86	83	85	85			
	Germany	Men's National	81	85	84	85			
	Belgium	Men's National	86	83	85	84			
	France	Men's National	82	85	81	84			
	Italy	Men's National	85	81	85	83			
	Argentina	Men's National	87	81	80	82			



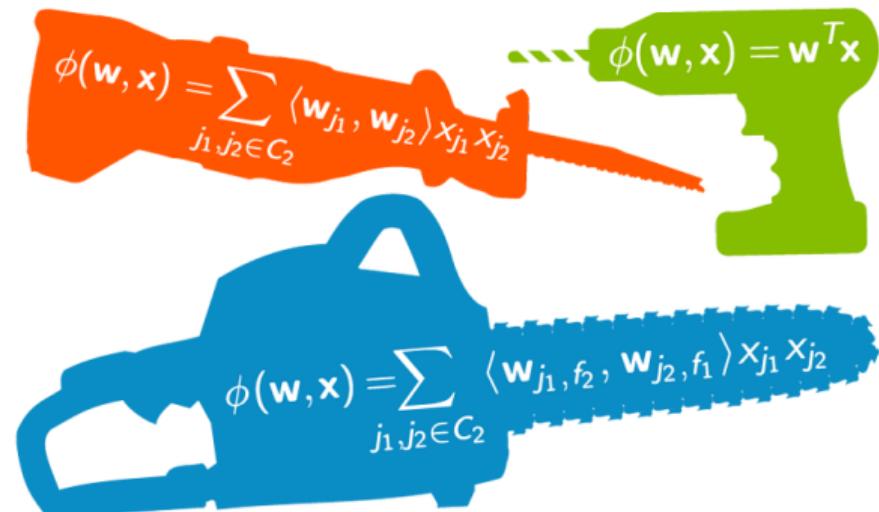
Thu thập dữ liệu từ Mendeley



Chuẩn bị dữ liệu (Data preparation)

Các phương pháp tiền xử lý dữ liệu:

- Bag of word
- TF-IDF
- Normalization
- PCA
- One-Hot-Encoding
- Feature Scaling
- Feature Selection
- Label Encoder





Machine Learning - Học

feat1	feat2	feat3	label
1	1	0	Yes
0	2	1	No
1	3	2	No
0	4	2	Yes
1	1	1	No

Dữ liệu thô

Ta có:

- $feat1 = \{0, 1\}$
- $feat2 = \{1, 2, 3, 4\}$
- $feat3 = \{0, 1, 2\}$
- $label = \{Yes, No\}$



Machine Learning - Dự đoán

feat1	feat2	feat3	label
1	1	0	Yes
0	2	1	No
1	3	2	No
0	4	2	Yes
1	1	1	No
3	9	10	?
0	1	?	?

Dữ liệu thô

Mô hình cần dự đoán được:

- Dữ liệu chưa được học.
- Dữ liệu bị thiếu.
- Dữ liệu bị trùng nhưng nhãn lại khác nhau.
- ...



Machine Learning - Dự đoán

feat1	feat2	feat3	label
1	1	0	Yes
0	2	1	No
1	3	2	No
0	4	2	Yes
1	1	1	No
3	9	10	Yes
0	1	1.2	No

Dữ liệu thô

Mô hình cần dự đoán được:

- Dữ liệu chưa được học.
- Dữ liệu bị thiếu.
- Dữ liệu bị trùng nhưng nhãn lại khác nhau.
- ...



Machine Learning - Thuật toán

- Linear Regression/Ordinary least squares

$$\mathbf{w} = \mathbf{A}^\dagger \mathbf{b} = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^\dagger \bar{\mathbf{X}}^T \mathbf{y}$$

- Gradient Decent

$$x_{t+1} = x_t - \eta f'(x_t)$$

- Logistic Regression

$$\mathbf{w} = \mathbf{w} + \eta (y_i - z_i) \mathbf{x}_i$$

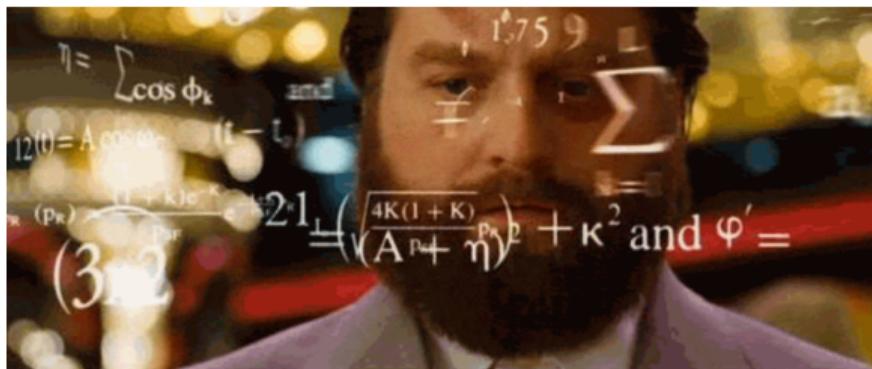
- Support vector machine (SVM)

$$\mathbf{w}^T \mathbf{x} + b = \sum_{m \in \mathcal{S}} \lambda_m y_m \mathbf{x}_m^T \mathbf{x} + \frac{1}{N_S} \sum_{n \in \mathcal{S}} \left(y_n - \sum_{m \in \mathcal{S}} \lambda_m y_m \mathbf{x}_m^T \mathbf{x}_n \right)$$



Dánh giá mô hình

- Độ chính xác (Accuracy)
- Ma trận nhầm lẫn (Confusion Matrix)
- True/False Positive/Negative
- Receiver operating characteristic curve (ROC)
- Area Under the Curve
- Precision và Recall
- Residual sum of squares (RSS)



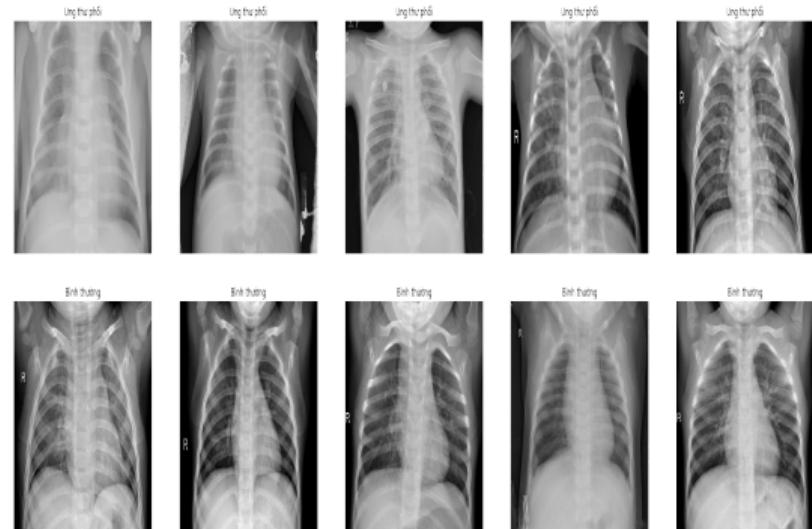


Demo áp dụng quy trình xử lý dữ liệu lớn

Dự đoán kết quả World Cup 2018
(sử dụng máy học)



Chuẩn đoán ung thư phổi qua ảnh
X-quang (sử dụng mạng nơ-ron)





Đã làm được:

- Ứng dụng quy trình khai phá dữ liệu và rút trích các tri thức giá trị của dữ liệu đối với 3 dạng bài toán: Regression, Classification và Clustering.
- Thực hiện Model Tuning để cải thiện độ chính xác mô hình.
- Thực hiện đánh giá mô hình qua nhiều phương pháp.
- Xây dựng Data Story một cách mạch lạc và trực quan.

Chưa làm được:

- Độ chính xác mô hình Machine Learning chưa cao, do dữ liệu vẫn đang còn thiếu.

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