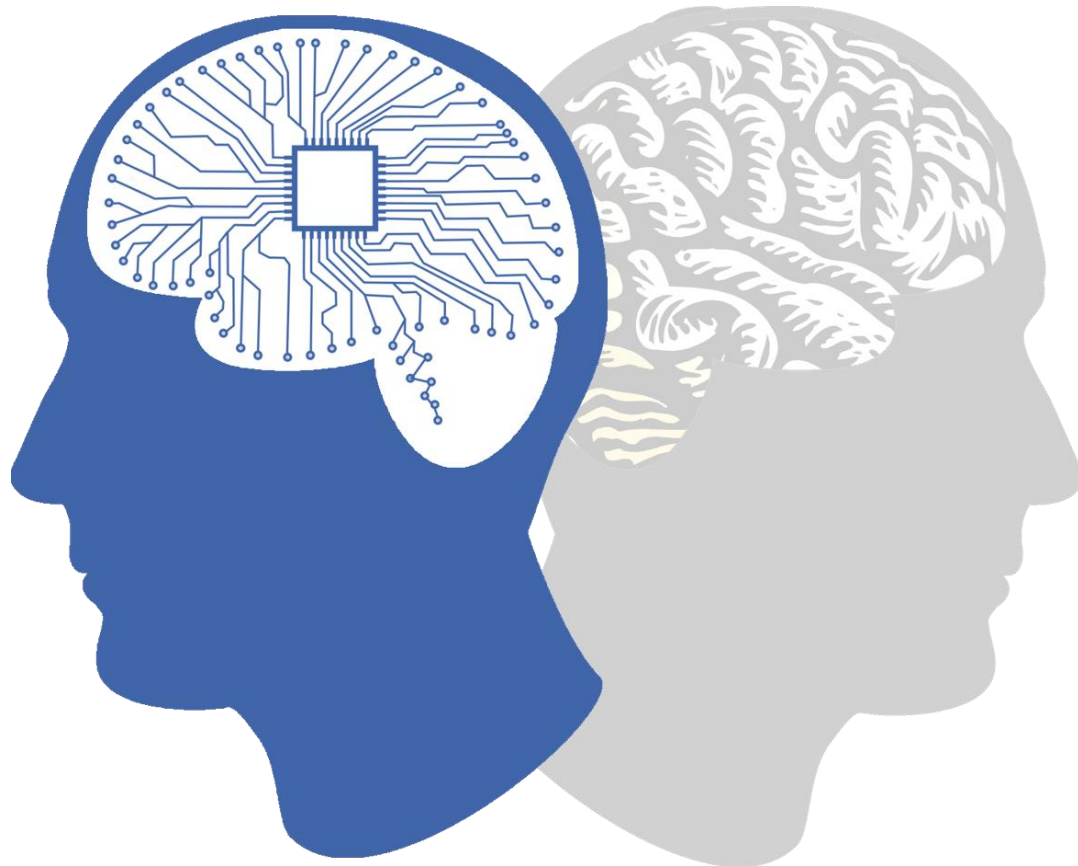
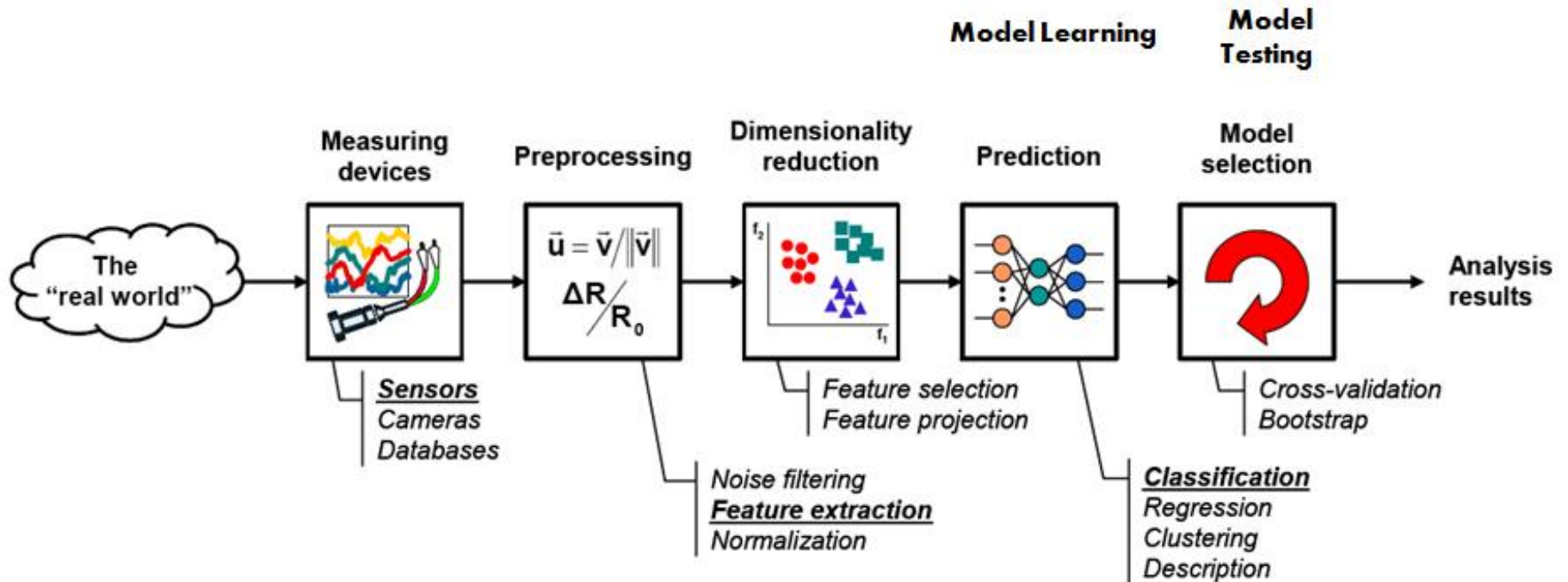


AI

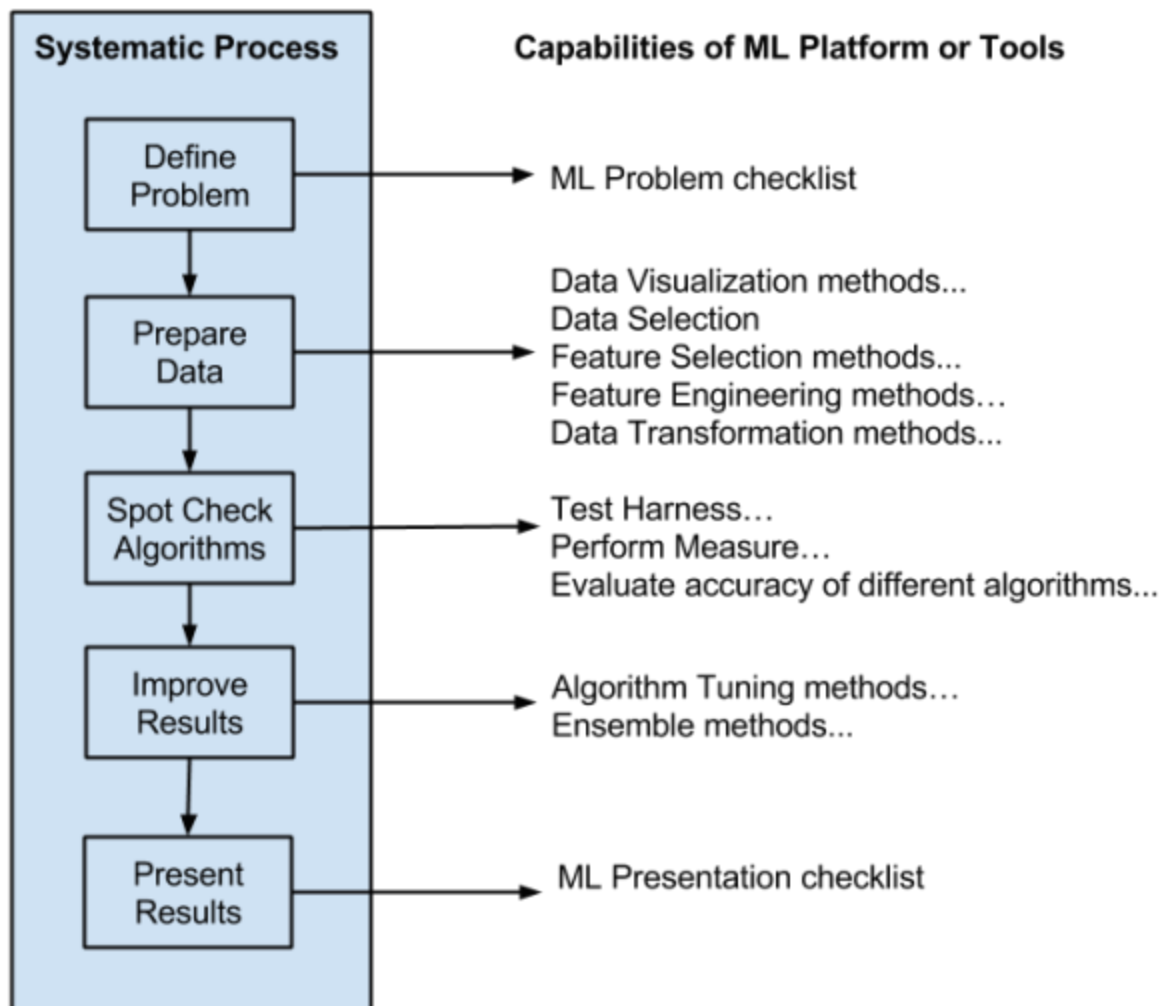
Artificial **I**ntelligence



QUÁ TRÌNH HỌC MÁY CƠ BẢN



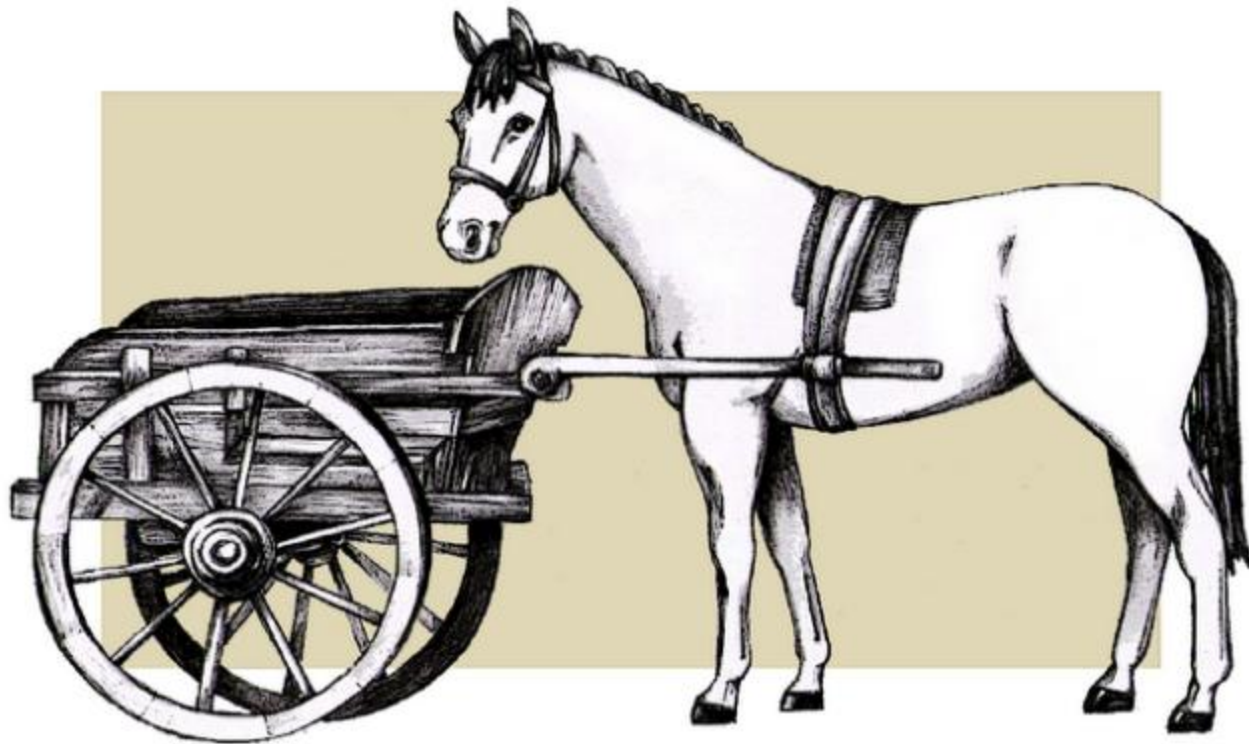
GIẢI QUYẾT MỘT BÀI TOÁN BẰNG ML



ĐỊNH NGHĨA BÀI TOÁN

- **What is the problem?**
 - Define the problem informally and formally.
 - List the assumptions about the problem (e.g. about the data).
 - List known problems similar to your problem.
- **Why does the problem need to be solved?**
 - Describe the motivation for solving the problem.
 - Describe the benefits of the solution (model or the predictions).
 - Describe how the solution will be used.
- **How could the problem be solved manually?**
 - Describe how the problem is currently solved (if at all).
 - Describe how a subject matter expert would make manual predictions.
 - Describe how a programmer might hand code a classifier.

ĐỊNH NGHĨA BÀI TOÁN



What is the problem?

Informal description

- Describe the problem as though you were describing it to a friend or colleague
- For example: *I need a program that will tell me which tweets will get retweets.*

Formalism

- Use this formalism to define the T , P , and E for your problem.
- For example:
 - **Task** (T): Classify a tweet that has not been published as going to get retweets or not.
 - **Experience** (E): A corpus of tweets for an account where some have retweets and some do not.
 - **Performance** (P): Classification accuracy, the number of tweets predicted correctly out of all tweets considered as a percentage.

What is the problem?

Assumptions

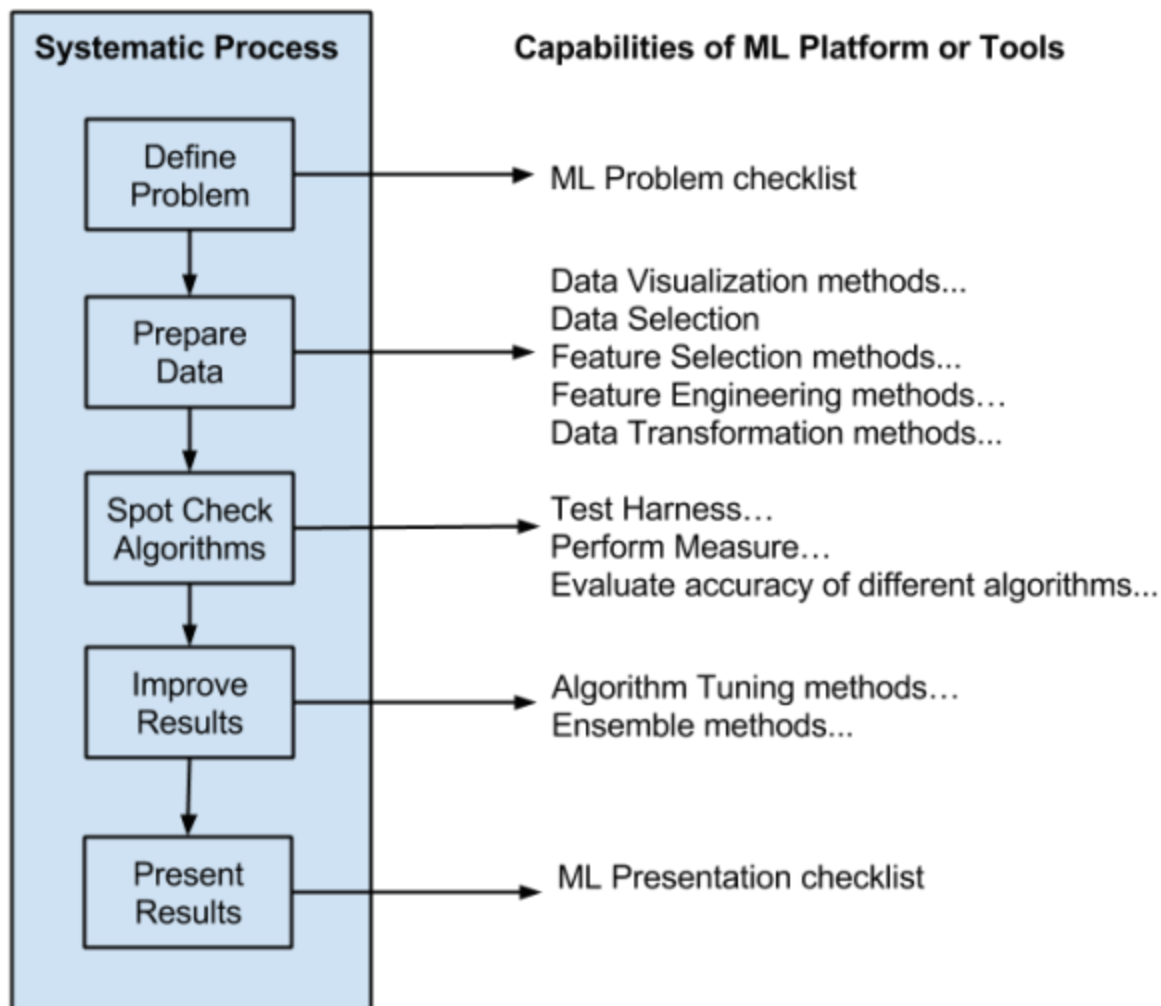
Create a list of assumptions about the problem and its phrasing. These may be rules of thumb and domain specific information that you think will get you to a viable solution faster.

- For example:
 - The specific words used in the tweet matter to the model.
 - The specific user that retweets does not matter to the model.
 - The number of retweets may matter to the model.
 - Older tweets are less predictive than more recent tweets

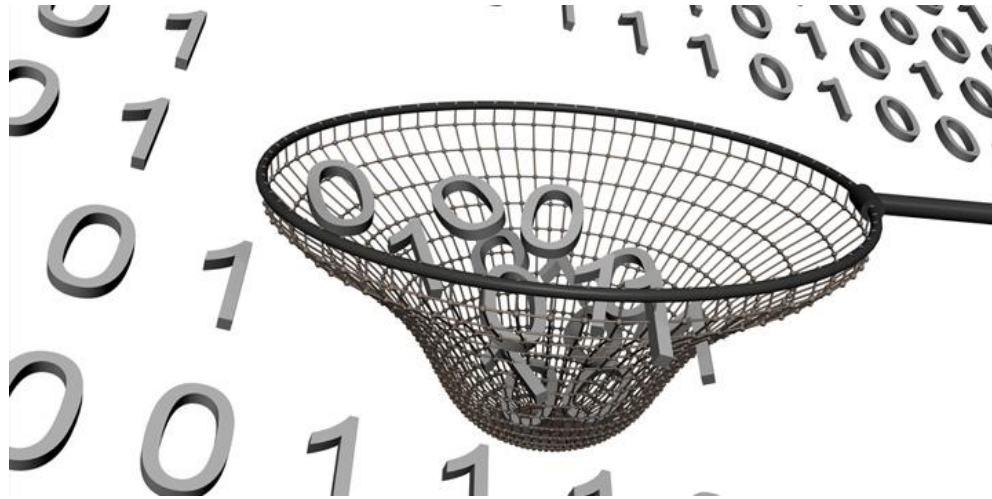
Similar problems

- Other problems can inform the problem you are trying to solve by highlighting limitations in your phrasing of the problem
- Point to algorithms and data transformations that could be adopted to spot check performance.
- **For example:** A related problem would be email spam discrimination that uses text messages as input data and needs binary classification decision.

GIẢI QUYẾT MỘT BÀI TOÁN BẰNG ML



GATHERING DATA



Data for:

- Model training.
- Model evaluation.
- Model tuning.
- Model validation.

Question

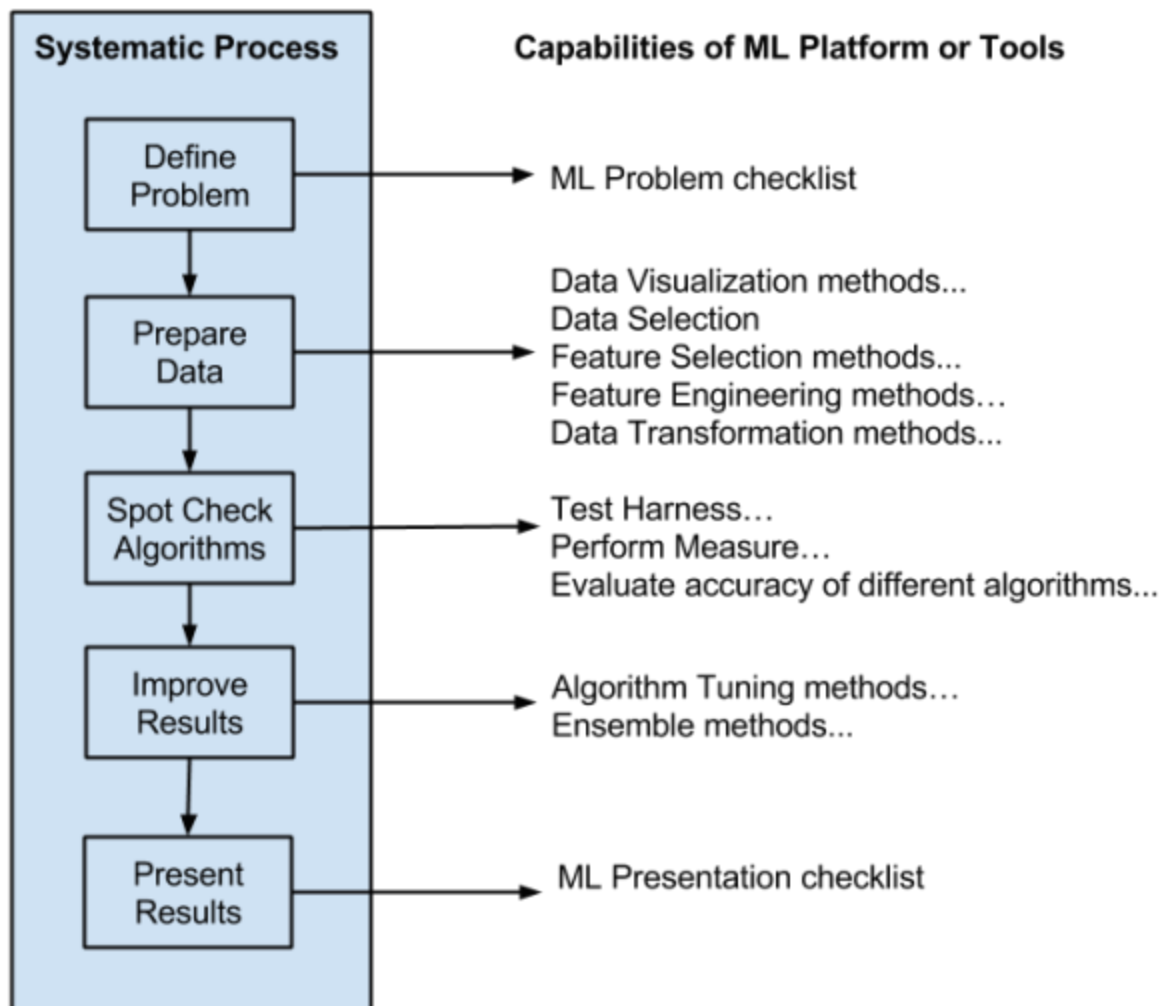
- Has anybody done it before?
- What is the domain of your problem? is it related to Computer Vision, Natural Language Processing, Sensor data, or some XYZ?

How Much Training Data is Required for Machine Learning?

- **It Depends; No One Can Tell You**
 - The complexity of the problem
 - The complexity of the learning algorithm
- **Reason by Analogy** : look at studies on problems similar to yours as an estimate for the amount of data that may be required.
- **Use Domain Expertise** : Use your domain knowledge, or find a domain expert
- **Use a Statistical Heuristic**
 - Factor of the number of classes
 - Factor of the number of input features
 - Factor of the number of model parameters
 - Nonlinear Algorithms Need More Data

Get More Data (No Matter What!?)

GIẢI QUYẾT MỘT BÀI TOÁN BẰNG ML



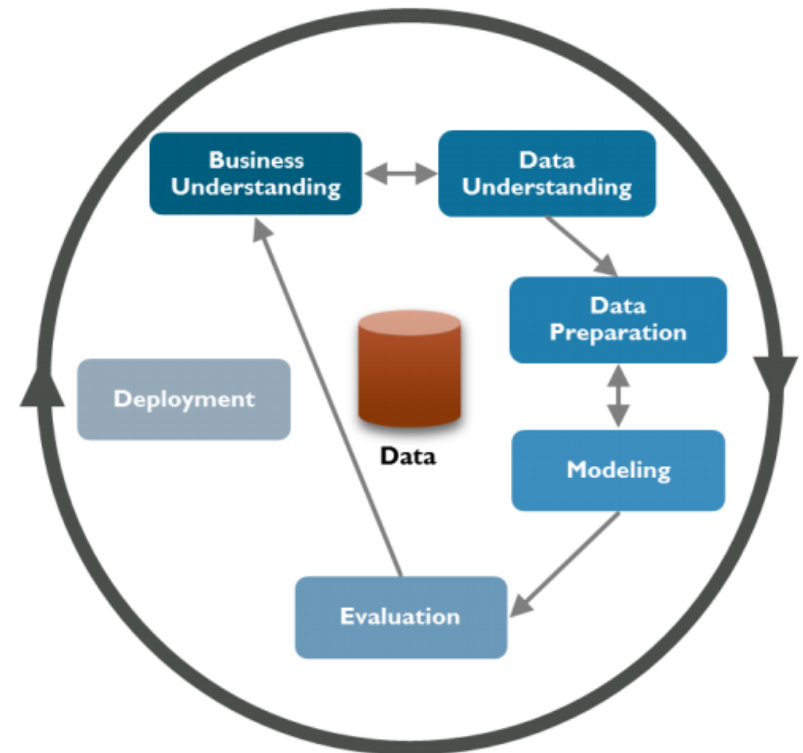
DATA PREPROCESSING

1. Data Engineering – 80%

- Data extraction
- Data cleaning
- Data transformation
- Data normalization
- Feature extraction

2. Machine Learning – 20%

- Model fitting
- Hyperparameters tuning
- Model evaluation

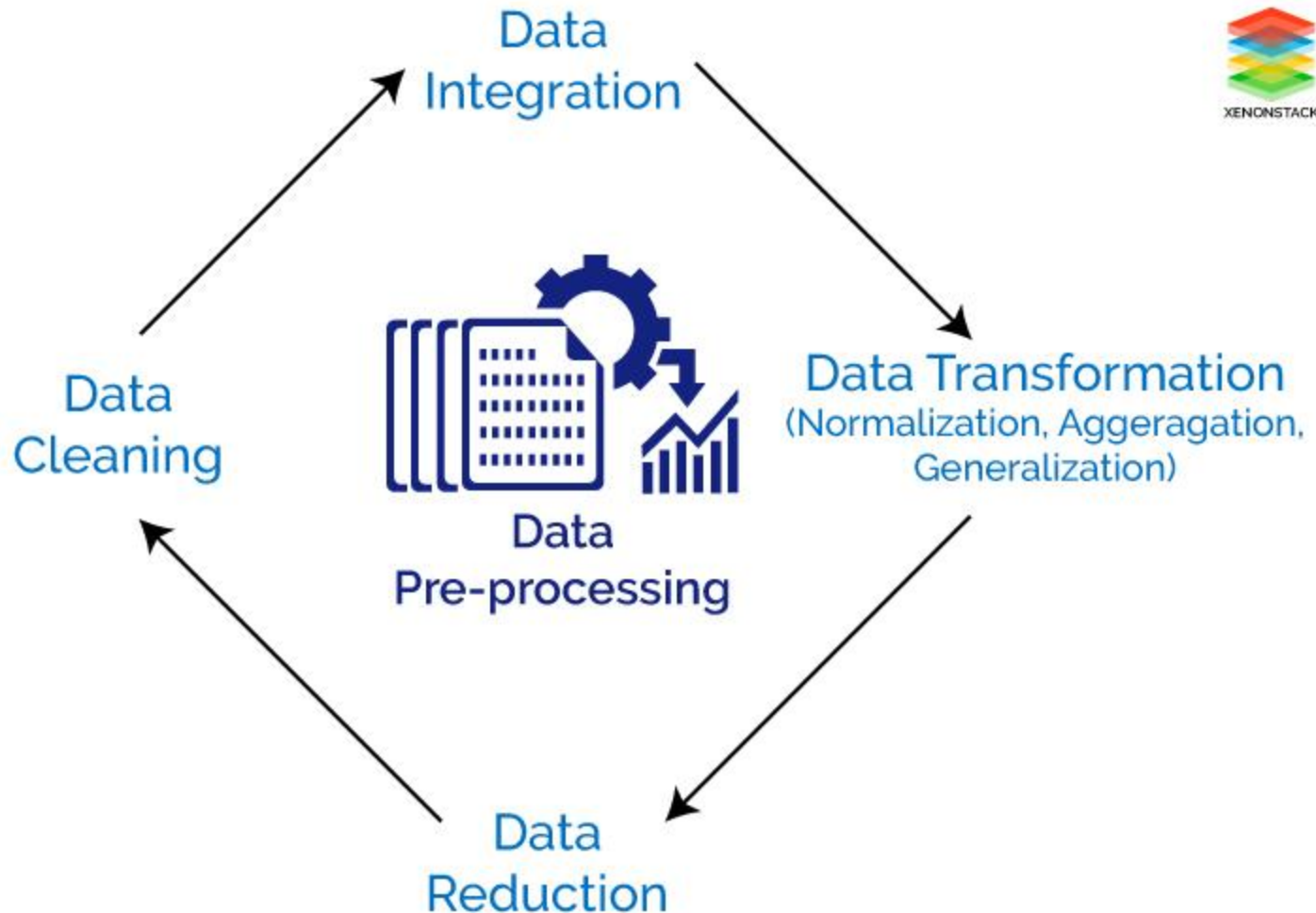


CRISP-DM

DATA PREPROCESSING

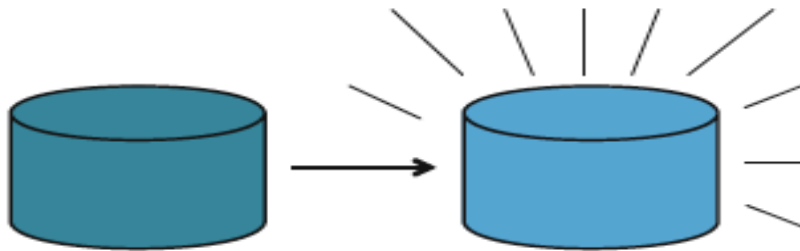
- Data in the real world is dirty
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., occupation=""
 - **noisy**: containing errors or outliers
 - e.g., Salary="-10", Age="222"
 - **inconsistent**: containing discrepancies in codes or names
 - e.g., Age="42" Birthday="03/07/1997"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
 - e.g., discrepancy between duplicate records
 - e.g., *Endereço*: travessa da Igreja de Nevogilde *Freguesia*: Paranhos

DATA PREPROCESSING

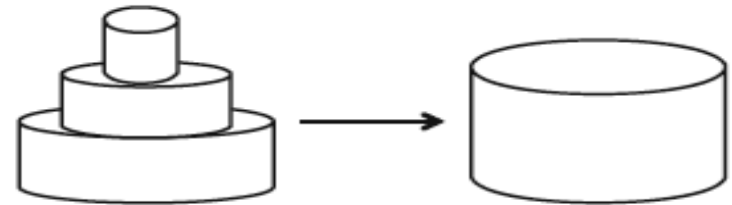


DATA PREPROCESSING

Data Cleaning



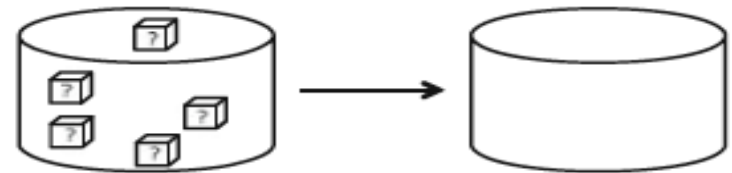
Data Normalization



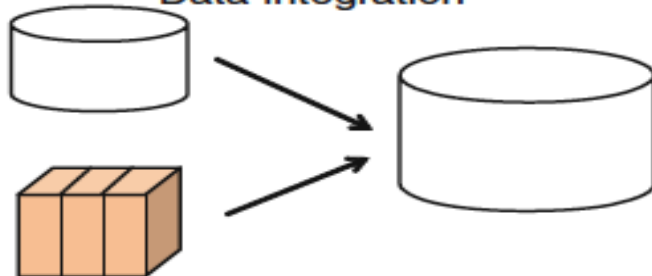
Data Transformation



Missing Values Imputation



Data Integration



Noise Identification



WHY DATA PREPROCESSING

- Preparing data also prepares the miner so that when using prepared data the miner produces better models, faster
- GIGO- good data is a prerequisite for producing effective models of any type

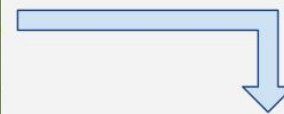
DATA PREPROCESSING

- **Data Cleaning**

- Correct bad data, filter some incorrect data out of the data set and reduce the unnecessary detail of data.

| Name | Date | Duration (s) | Genre | Plays |
|---------------|------------|--------------|---------|-------|
| Highway star | 1984-05-24 | - | Rock | 139 |
| Blues alive | 1990/03/01 | 281 | Blues | 239 |
| Lonely planet | 2002-11-19 | 5:32s | Techno | 42 |
| Dance, dance | 02/23/1983 | 312 | Disco | N/A |
| The wall | 1943-01-20 | 218 | Reagge | 83 |
| Offside down | 1965-02-19 | 4 minutes | Techno | 895 |
| The alchemist | 2001-11-21 | 418 | Bluesss | 178 |
| Bring me down | 18-10-98 | 328 | Classic | 21 |
| The scarecrow | 1994-10-12 | 269 | Rock | 734 |

Original data



Cleaned data

| Name | Date | Duration (s) | Genre | Plays |
|---------------|------------|--------------|---------|-------|
| Highway star | 1984-05-24 | | Rock | 139 |
| Blues alive | 1990-03-01 | 281 | Blues | 239 |
| Lonely planet | 2002-11-19 | 332 | Techno | 42 |
| Dance, dance | 1983-02-23 | 312 | Disco | |
| The wall | 1943-01-20 | 218 | Reagge | 83 |
| Offside down | 1965-02-19 | 240 | Techno | 895 |
| The alchemist | 2001-11-21 | 418 | Blues | 178 |
| Bring me down | 1998-10-18 | 328 | Classic | 21 |
| The scarecrow | 1994-10-12 | 269 | Rock | 734 |

DATA PREPROCESSING

- **Data Transformation (chuyển đổi)**

- The data is consolidated so that the mining process result could be applied or may be more efficient.

Before transformation

banking.csv

| euribor3m | nr_employed | y |
|-----------|-------------|-----|
| 4.857 | 5191 | no |
| 4.857 | 5191 | no |
| 4.857 | 5191 | yes |
| 4.857 | 5191 | yes |
| 4.857 | 5191 | no |

Target

After transformation

banking.csv

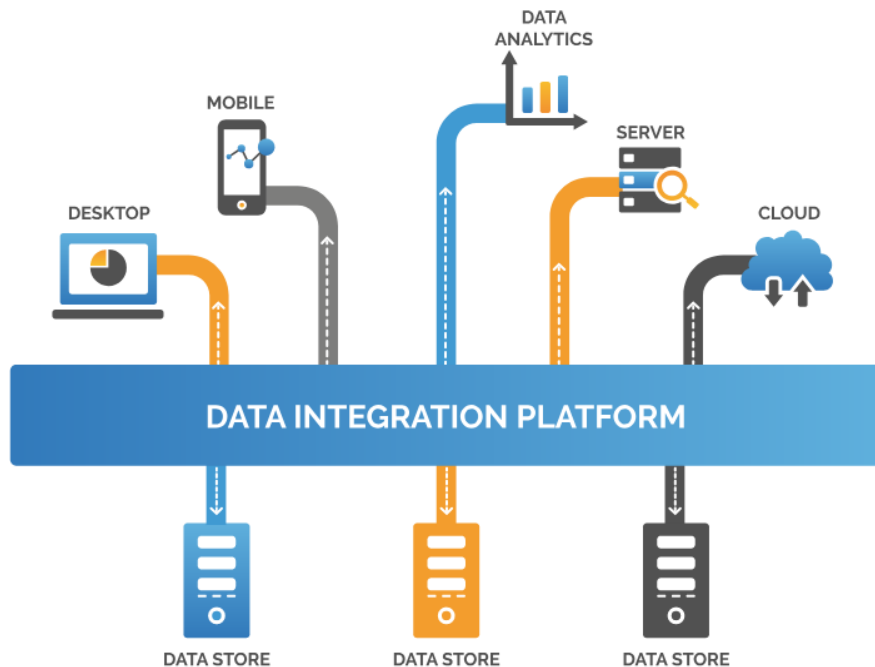
| euribor3m | nr_employed | y |
|-----------|-------------|---|
| 4.857 | 5191 | 0 |
| 4.857 | 5191 | 0 |
| 4.857 | 5191 | 1 |
| 4.857 | 5191 | 1 |
| 4.857 | 5191 | 0 |

Target

DATA PREPROCESSING

- **Data Integration**

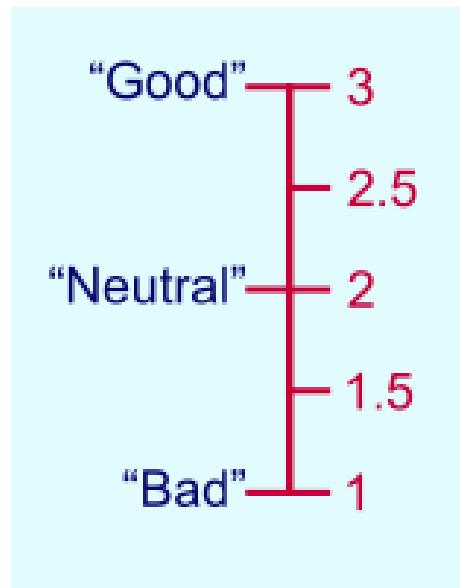
- Merging of data from multiple data stores.



DATA PREPROCESSING

- **Data Normalization**

- To express data in the same measurements units, scale or range.

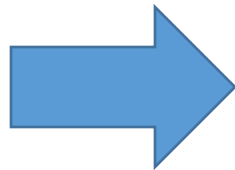


DATA PREPROCESSING

- **Missing Data Imputation**

- To fill the variables that contain missing values with some intuitive data.

| Item | Y | X |
|------|----|----|
| 1 | 9 | 7 |
| 2 | ? | 10 |
| 3 | 11 | 19 |
| 4 | ? | 10 |
| 5 | 15 | 14 |
| 6 | 19 | 18 |
| 7 | 21 | 5 |
| 8 | 8 | 4 |
| 9 | 19 | 21 |
| 10 | 21 | 17 |

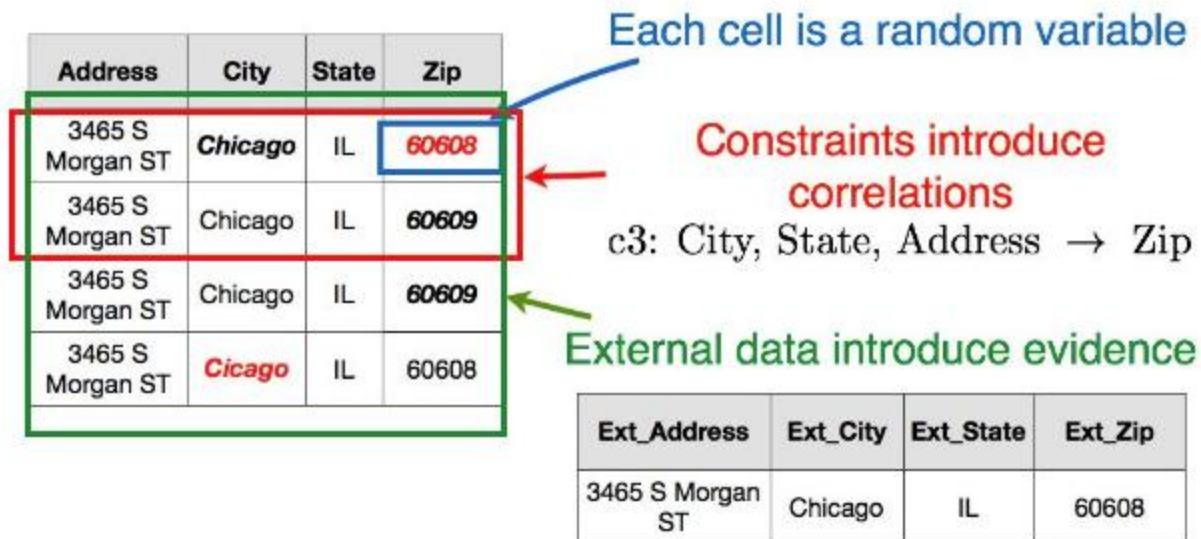


| Item | Y | X |
|------|----|----|
| 1 | 9 | 7 |
| 2 | 9 | 10 |
| 3 | 11 | 19 |
| 4 | 15 | 10 |
| 5 | 15 | 14 |
| 6 | 19 | 18 |
| 7 | 21 | 5 |
| 8 | 8 | 4 |
| 9 | 19 | 21 |
| 10 | 21 | 17 |

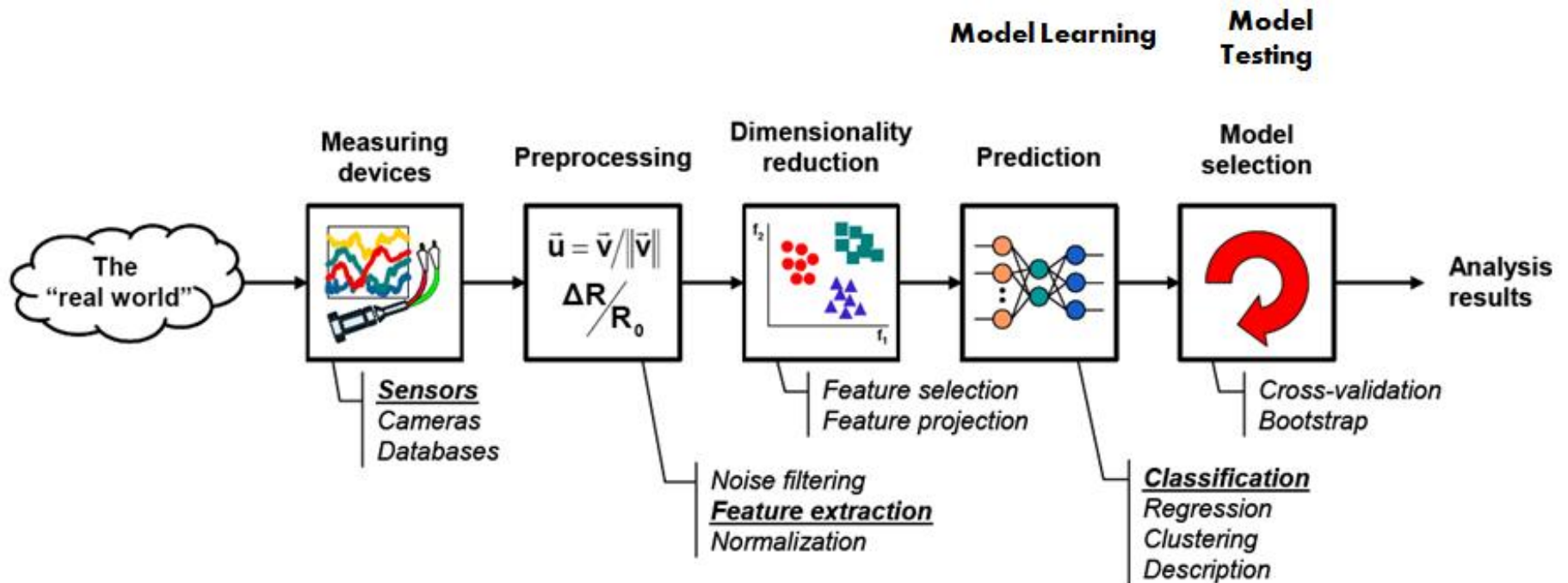
DATA PREPROCESSING

- **Noise Identification**

- To detect random errors or variances in a measured variable.



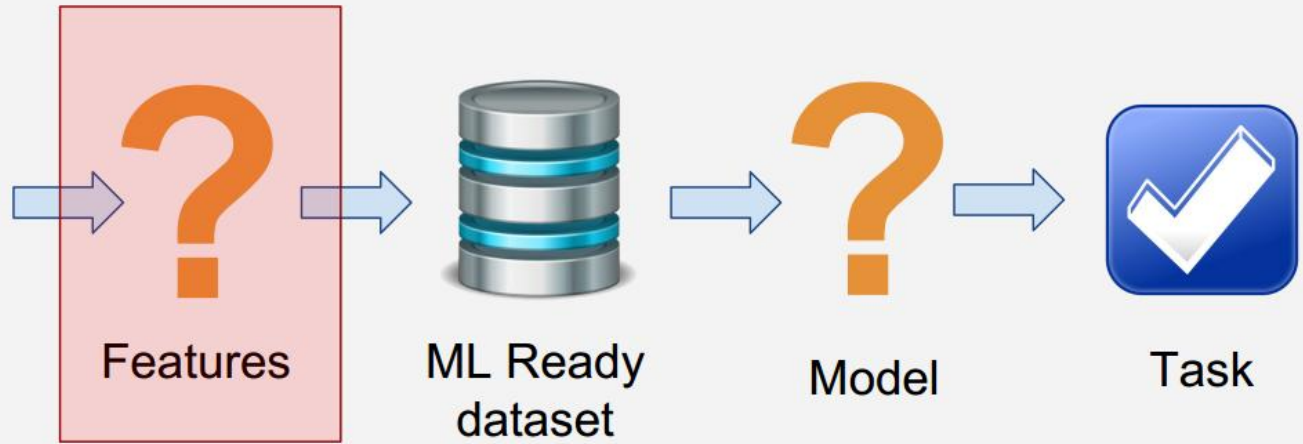
QUÁ TRÌNH HỌC MÁY CƠ BẢN





Raw data

... The Reality







DUCKS



NOT DUCKS

✓ Walks ✓ Quacks ✓ Swims

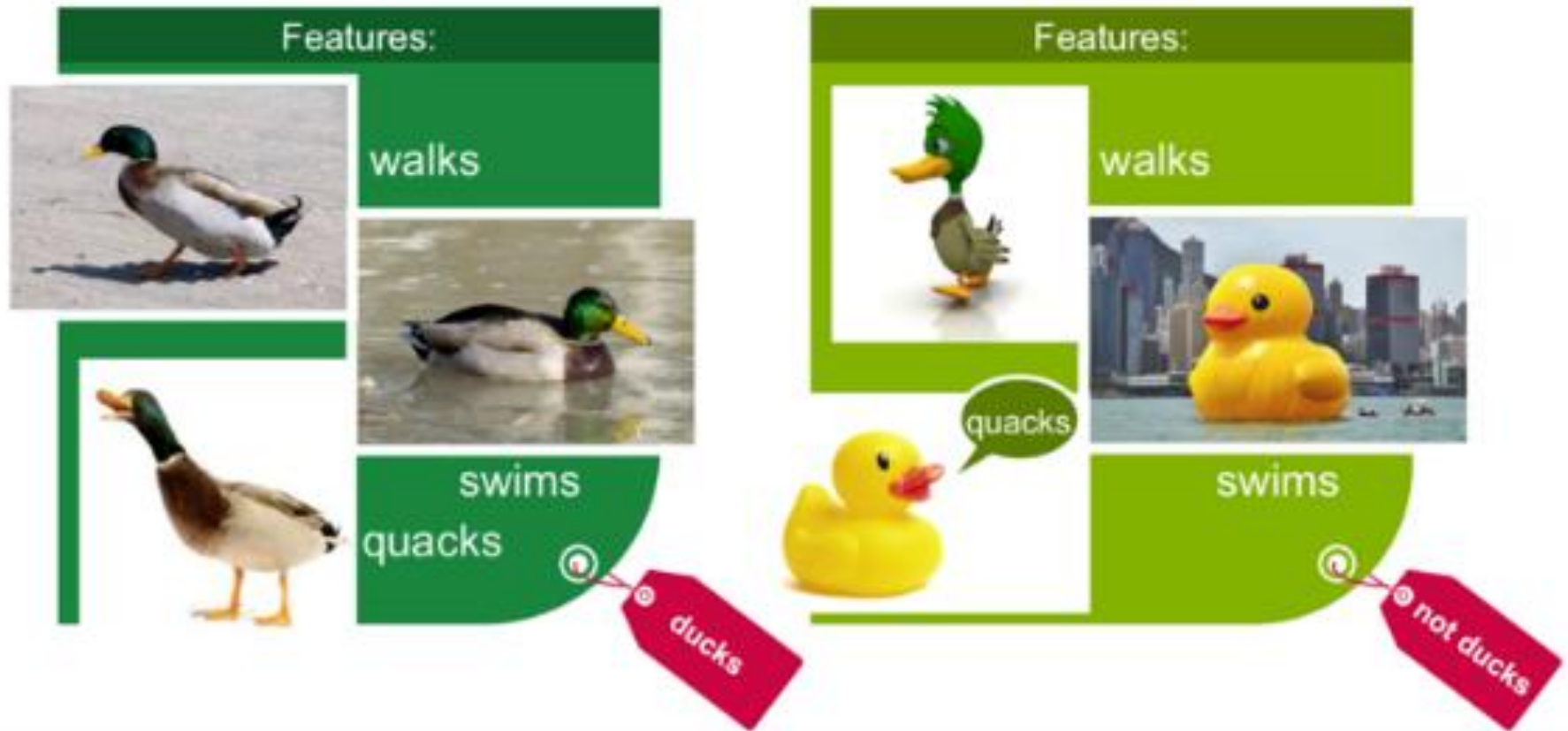


DUCKS



NOT DUCKS

If it Walks/Swims/Quacks Like a Duck Then It Must Be a Duck



Feature ?

- *An attribute or group of attributes that constitute a characteristic property or set of properties which is **unique, measurable and differentiable** .*
- Feature is something that **makes a class different from another class.**

Outbrain Click Prediction - Kaggle competition

Can you predict which recommended content each user will click?

The screenshot shows a CNN article page with several annotations:

- Publisher:** A green line points to the URL `edition.cnn.com` in the address bar.
- Document:** A green line points to the full URL `2016/08/23/middleeast/iraq-nineveh-mosul-scene/index.html` in the address bar.
- Promoted Content Set:** A red line points to a red-bordered box containing a list of recommended content items.
- Promoted Content Item:** A blue line points to one specific item within the recommended content set.

The article text visible is: "I am so happy for them," the man said. "But I am heartbroken myself. My parents were not able to come with me. I don't know how I am going to get them out."

The recommended content set includes:

- Paid Content:** Mapping the Startup Nation: The 12 most popular Tech Hubs in... (Viola Notes)
- First time in Israel:** Business degrees in Ramat Gan and New... (Israel News)
- The most addictive game of the year!** Play with 15 million Players... (Forge Of Empires)
- How to Avoid Everyday Pain Landmines** (Womens Health)
- How One Brand is Disrupting the \$63 Billion Makeup Industry** (The Huffington Post)
- Find out what special ingredient makes this omelette so tasty** (HomeMadebyYou)

Dataset

- Sample of users page views and clicks during 14 days on June, 2016
- 2 Billion page views
- 17 million click records
- 700 Million unique users
- 560 sites

Feature Engineering

"Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data."

– Jason Brownlee

Feature Engineering

- Là quá trình chuyển **đổi tập dữ liệu thô** đầu vào thành tập **các thuộc tính (features)** có thể giúp biểu diễn tập dữ liệu tốt hơn.
- Feature engineering cố gắng **biểu diễn tốt nhất** tập dữ liệu ban đầu sao cho **tương thích với mô hình dự đoán** đang sử dụng.

Feature Engineering

- Phân loại sinh viên có học tốt hay không
- Sử dụng tất cả các thuộc tính liên quan đến sinh viên đó để áp dụng cho bài toán phân lớp, mà các thuộc tính này thường rất nhiều từ 20-50 cột thuộc tính.
- ➔ Thời gian để máy traing rất lâu, đồng thời kết quả dự đoán có độ chính xác thấp.
- ➔ Sử dụng kĩ thuật feature engineering để chọn ra một số thuộc tính phù hợp hơn như có vay mượn để đóng học phí không, số điểm đầu vào là bao nhiêu, quá trình tiến bộ trong học tập là bao nhiêu,...

Một số bài toán trong Feature Engineering

- **Bài toán rút trích đặc trưng (feature extraction)**
- **Bài toán đánh giá độ hữu dụng của các feature**
- **Bài toán lựa chọn đặc trưng (feature selection)**
- **Bài toán xây dựng đặc trưng mới (feature construction)**
- **Bài toán xác định feature thông qua training dữ liệu (feature learning).**

Feature extraction

- Là tiến trình tự động hoá được dùng để **giảm số chiều dữ liệu** sao cho dữ liệu ban đầu được chuyển đổi sang dạng **đơn giản và nhỏ hơn dữ liệu ban đầu**, trước khi đưa vào mô hình dự đoán



Image ' I_1 '

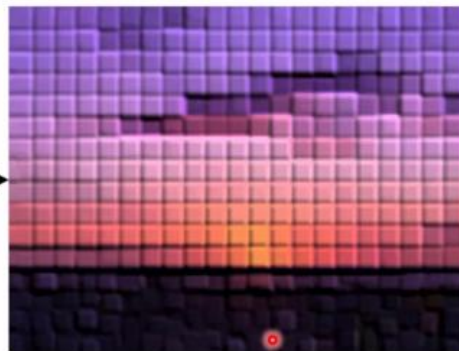
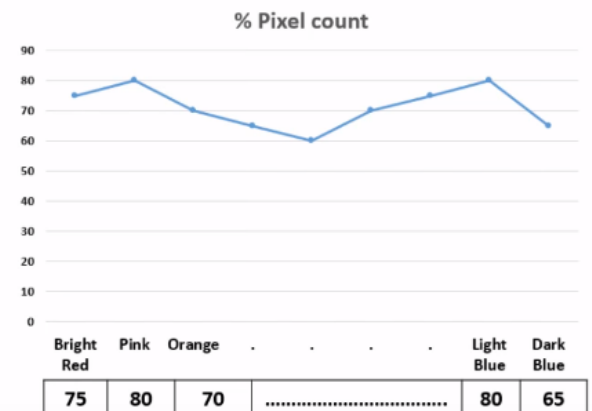
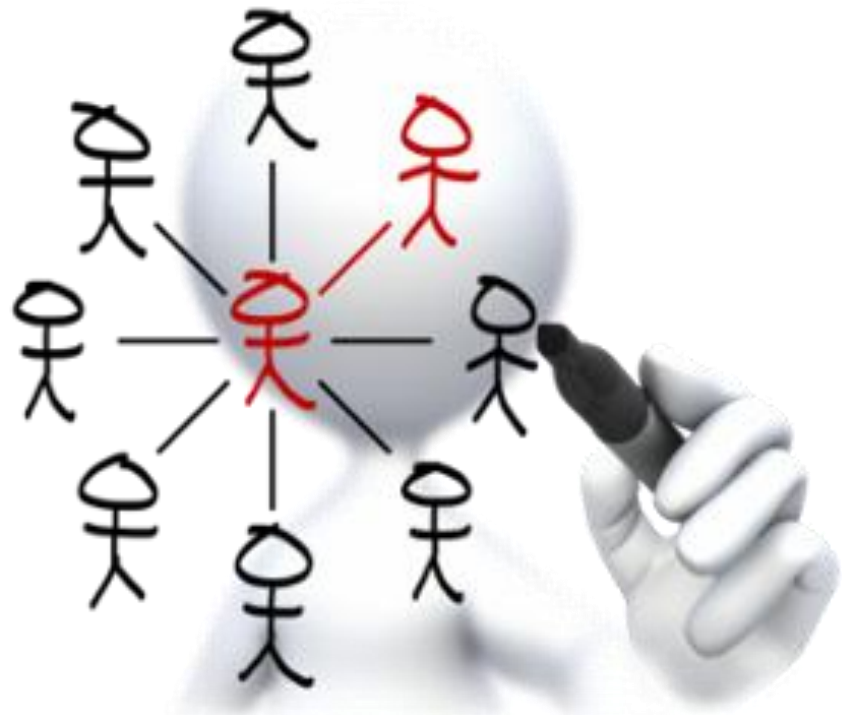


Image ' I_1 ' represented in form of pixels



Feature Selection

- Tự động hoá lựa chọn tập con trong số các feature ban đầu sao cho các feature được lựa chọn này phù hợp với bài toán hiện tại



QUÁ TRÌNH HỌC MÁY CƠ BẢN

