CHALLENGES OF MACHINE LEARNING

# Bad data

## Insufficient quantity of data

Trong machine learning dử liệu có thể ảnh hưởng đến độ hiệu quả của thuật toán.

## Nonrepresentative data

Noisy data: dử liệu thu được có một số phần tử bất thường có thể ảnh hưởng đến kết quả thuật toán.

Sampling bias: dử liệu chỉ được thu thập ở một vùng nào đó và không trải đều ra các trường hợp của bài toán

## Irrelevant feature:

Dử liệu có các feature không liên quan tới bài toán.

## Missing data

Không thu được một số các feature củ dử liệu. Có thể loại bỏ các feature có quá nhiều ô trống hoặc bỏ các sample có quá nhiều feature trống. Ngoài ra, có thể thêm dử liệu vào feature trống nếu thấy hợp lý.

# Bad Algorithm

## Overfitting:

Sử dụng model quá phức tạp với bài toán.

## Undefitting:

Sử dụng model quá đơn giản với bài toán.

# Testing and hyperparameter tunning

## Testing and generalization error:

Để test ta chia dữ liệu ta có thành 2 phần, một phần dùng để trainning, một phần dùng để test.

Đối với những tập dử liệu lớn, ta có thể chọn ngẩu nhiên dử liệu để đưa vào tập test (Random Sampling).

Đối với những tập dử liệu không quá lớn, ta có thể dùng Strategy Sampling.

## Hyperparameter tunning:

Hyperprameters là các tham số mà thuật toán cần để chạy (trừ dử liệu). Hyperprameters có thể ảnh hưởng trực tiếp tới độ hiệu quả của thuật toán.

Hyperparameter tunning là điều chỉnh hyperparameter để thuật toán chạy hiệu quả. Để là được việc đó ta cần tạo một tập dử liệu gọi là Validation Set để kiểm tra độ hiệu quả của Hyperparameter.