# CHAPTER V: BASIC CONCPETS

## I. Learn

A computer program is said to learn from experience E with respect to some tasks T and performance measure P, if its performance (hiệu suất) at tasks in T, as measured (đánh giá) by P, improves with experience E

## II. Tasks

* Tasks is described (mô tả) by the way how Machine Learning do data point
* Each data point has multiple features
  + Each feature (đặc trưng) is rendered (biểu diễn) by number
  + Aggregate (tập hợp) number be a feature vector/matrix
  + Example: vector **x**

### 1. Classification (phân loại/phân lớp)

* Task requested: pointed out (chỉ ra) label of a data point
  + Label is an element of assemblage has C elements (named )
  + Solve task: Build a function
  + The assigned model (mô hình gán) for a data point is described by feature vector x a label identified (xác định) by number y
  + There is a small variation (biến thể) on output of function when output isn’t number but vector where show probability (xác suất) that data point **x** falls in class Cth
  + The last class selected is which has highest probability show confidence of model
  + higher difference between probabilities higher confidence

### 2. Regression (tiên lượng)

* If label isn’t subgroups and is real number (can infinity) Regression
  + Build a function
* Widen: Predict (dự đoán) multiple outputs at the same time
  + Build a function

### 3. Machine translation

* Translate word, text, … use deep learning algorithm (thuật toán)

### 4. Clustering

* Grouping (phân nhóm) all X data into small groups based on (dựa trên) relationship between datas in each group

### 5. Completion

* Filling in missing values ​​of a data point
* Based on relationship between datas to predict missing values
  + Example: recommendation system (hệ thống khuyến nghị)

## III. Performance

* Data is divided be 2 separate (riêng biệt) part: Training set (tập) & Test set
  + Training set: find parameters (tham số) of model
  + Test set: measure ability (khả năng) of model
  + Note:
    - find only on training set but measure can do on both
    - The line between training set and test set is sometimes ambiguous (mơ hồ)/ not clearly (rõ ràng)
* Online Learning & Offline Learning
  + Online: update new to improve
  + Offline: can’t update, new test set

## IV. Experience

* Training the machine learning model can be considered giving them experience above dataset (tập dữ liệu) \_ training set
* Supervised learning (học có giảm sát/chỉ bảo) & Unsupervised learning
  + Supervised: predict output of 1 or more new inputs based on known pair (input, output)
  + Unsupervised: only known input then use grouping or dimentionality reduction and experience lots of to predict output
* The line between supervised and unsupervised is ambiguous/not clearly
  + Usually, people often consider classification and regression as supervised learning
  + clustering or density estimation (ước lượng 1 phân phối) as unsupervised learning.
* Semi supervised learning (học bán dám sát): use reinforcement learning (học củng cố)

## V. Loss function (Hàm mất mát) & Model parameters (Tham số mô hình)

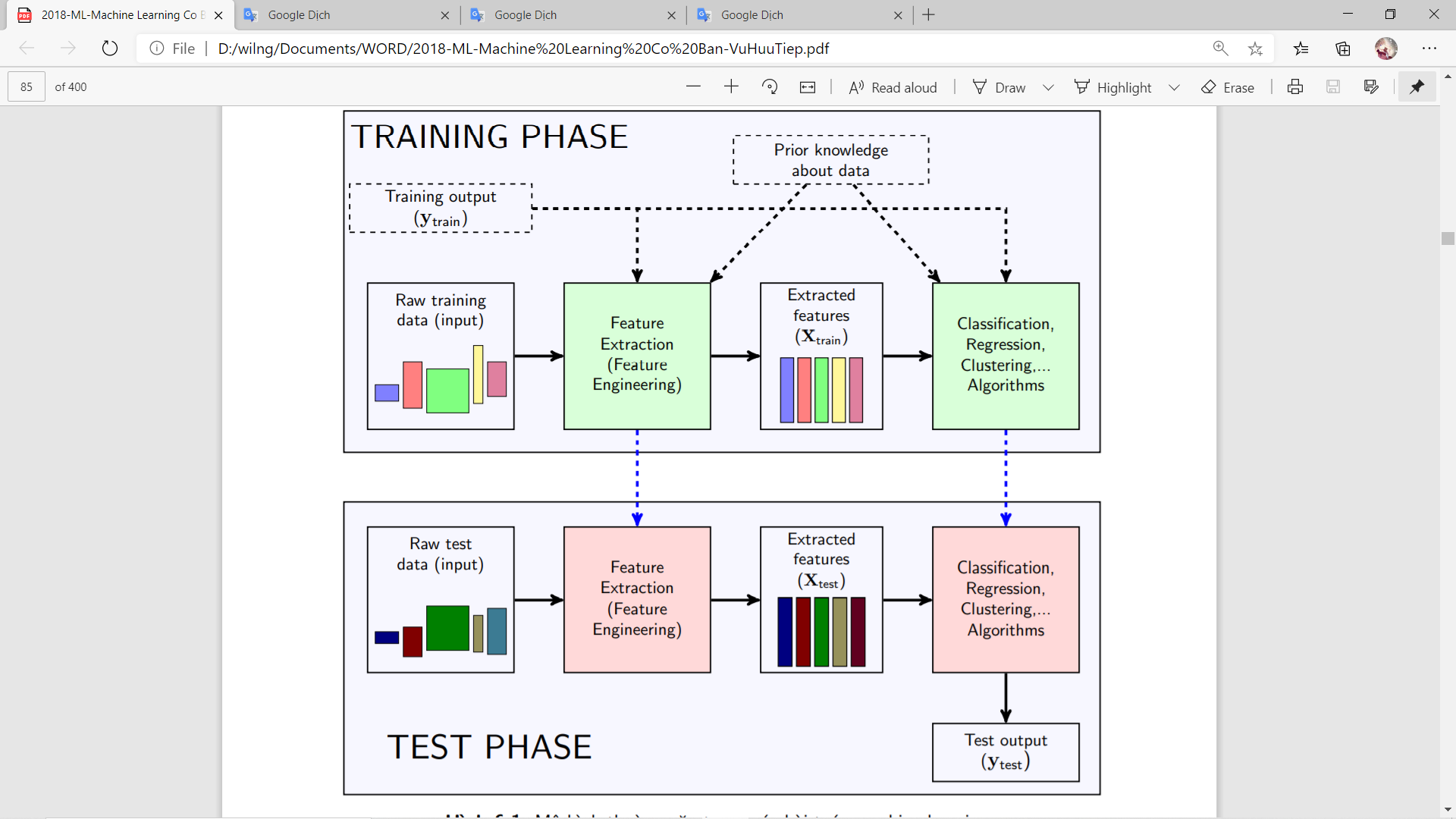
* Each ML model is described by model parameters
* Work of ML algorithm is finding model parameter consistently with problem (phù hợp)
  + Finding model parameters is closely related with performance
  + Aim of finding is lead performance to best result = minimum loss function
  + The relationship between performance & the model parameters is often described through a function called a loss function (or cost function (hàm chi phí))
* Problem: Minimum Loss function
  + Set of model parameters is signed by
  + Loss function is signed by
  + : value of so that function can achieve the smallest value
  + Fact, only necessary to find out a set of parameters θ that minimizes the loss function, or even reaches a minimum value

# CHAPTER VI: FEATURE ENGINEERING

## I. Introduce

* Request: data the same size
* Fact: Raw data (thô) Process
  + expel noisy data & bring raw data which different size or dimensions to the same standard (either vector or matrix)
  + This new standard data must be preserved (đảm bảo) featured information to original (ban đầu) raw data
  + Not only that, depending on (tùy thuộc vào) the article math (bài toán), we need to design transformations which have appropriate (thích hợp) features
* This process is Feature engineering or feature extraction
* Coming up with features is difficult, time-consuming, requires expert knowledge. “Applied machine learning” is basically feature engineering - Andrew Ng

## II. General model of LM problem (mô hình chung)



* Two phase (bước) in ML problem is Training phase & Test phase

### 1. Training phase

Feature Extraction (first green block)

* Mission is creating a feature vector for each input data points
* This feature vector is usually the same size, regardless of (bất kể) how size input data
* Inputs can be:
  + Original raw data (Raw training input): all information about data
    - Example: picture (pixel), document (word), sound …
    - It usually isn’t a vector and different dimensions
  + Output of training set:
    - Unsupervised hasn’t
    - Supervised can no use
  + Prior knowledge about data: other known information about data

Second green block

* Original raw data passed (đưa vào) and generated (tạo ra) the corresponding (tương ứng) feature vectors called ***Extracted feature***
* These extracted features will be put into training for main algorithms such as classification,
* regression, clustering, etc

### 2. Test phase

* When new raw data is available, using the feature extraction to create feature vector corresponding to it
* This feature vector is passed into the main algorithm to make a decision

## III. Example

### 1. Get raw data directly