Introduction to Machine Learning

Software Innovation Institute, ANU

What do we hope to cover?

Modern machine learning (ML) methods for natural language processing (NLP).

What do we hope to cover?

- Representation (for documents/sentences)
- Supervised Machine Learning
  - Linear Classifier
  - Deep Neural Network
- Unsupervised Machine Learning
  - Clustering
  - Self-supervised Learning

What is machine learning?

Machine Learning is about prediction

Examples / features	$x_1, \dots, x_n \sim X$
Labels / annotations	$y_1, \dots, y_n \sim Y$
Predicator	$f_W(x)\colon X \longrightarrow Y$

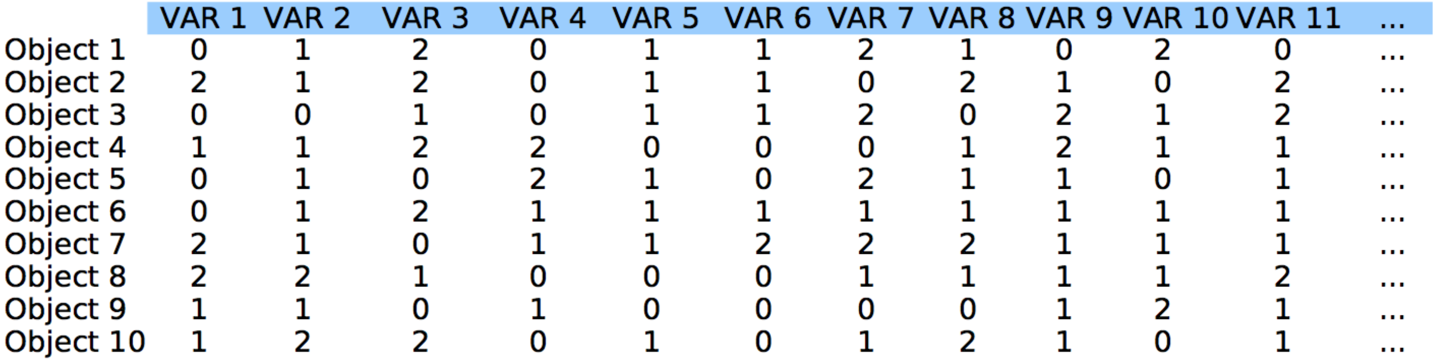
Estimate best predicator = training

Given data (x_1, y_1), … , (x_n, y_n), find a predicator $f_W(x)$

Prediction $\neq$ Understanding

Glossary

Data=a table (dataset, database, sample)



	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5	VAR 6	VAR 7	VAR 8	VAR 9	VAR 10	VAR 11	...
Object 1	0	1	2	0	1	1	2	1	0	2	0	...
Object 2	2	1	2	0	1	1	0	2	1	0	2	...
Object 3	0	0	1	0	1	1	2	0	2	1	2	...
Object 4	1	1	2	2	0	0	0	1	2	1	1	...
Object 5	0	1	0	2	1	0	2	1	1	0	1	...
Object 6	0	1	2	1	1	1	1	1	1	1	1	...
Object 7	2	1	0	1	1	2	2	2	1	1	1	...
Object 8	2	2	1	0	0	0	1	1	1	1	2	...
Object 9	1	1	0	1	0	0	0	0	1	2	1	...
Object 10	1	2	2	0	1	0	1	2	1	0	1	...

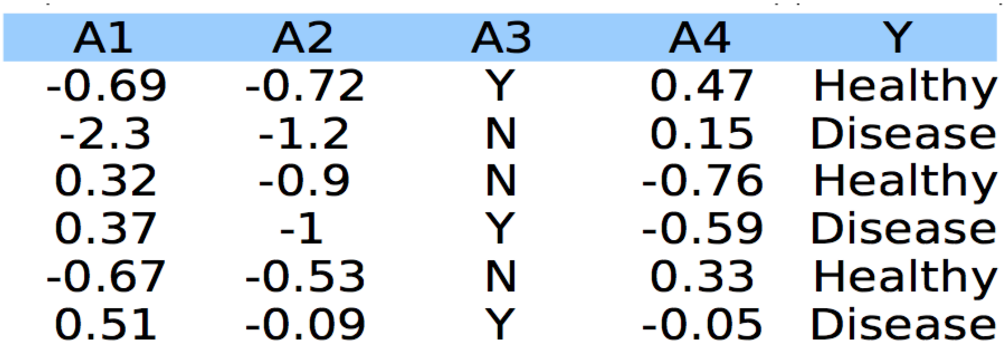
- **Variables** (attributes, features) = measurements made on objects
- **Objects** (samples, observations, individuals, examples, patterns)
- **Dimension** = number of variables
- **Size** = number of objects

For example:

- Objects: samples, patients, documents, images...
- Variables: genes, proteins, words, pixels...

Supervised Learning

Inputs **Output**



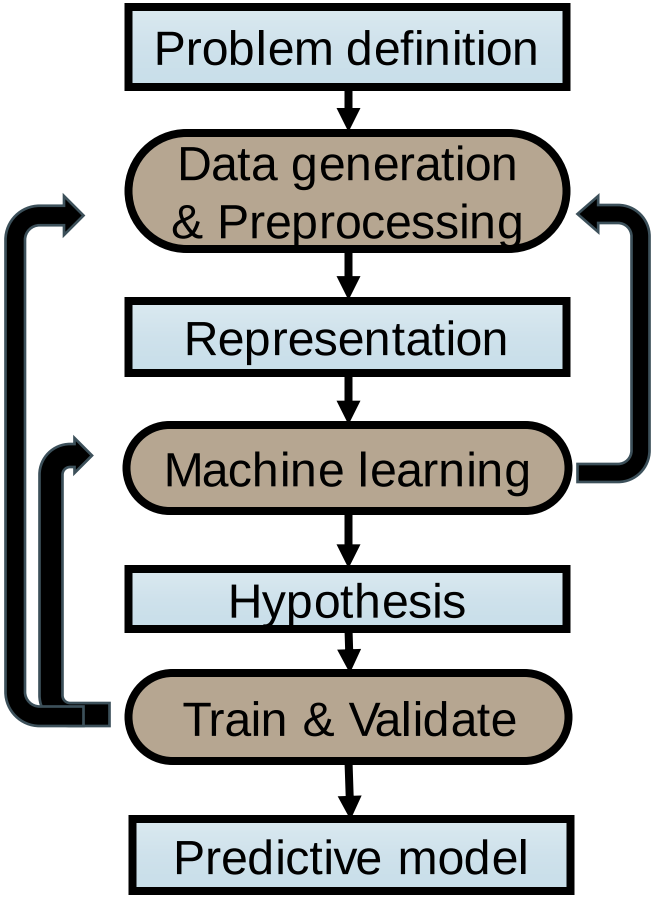
A1	A2	A3	A4	Y
-0.69	-0.72	Y	0.47	Healthy
-2.3	-1.2	N	0.15	Disease
0.32	-0.9	N	-0.76	Healthy
0.37	-1	Y	-0.59	Disease
-0.67	-0.53	N	0.33	Healthy
0.51	-0.09	Y	-0.05	Disease

$$\xrightarrow{\text{Supervised Learning}} \quad \boldsymbol{Y = h(A_1, A_2, A_3 A_4)}$$

Model
Hypothesis

- Goal: from the database (learning sample), find a function *h* of the inputs that approximates ***at best*** the output
- Discrete output ⇒ *classification* problem
- Numerical output ⇒ *regression* problem

How to model NLP as ML problem?

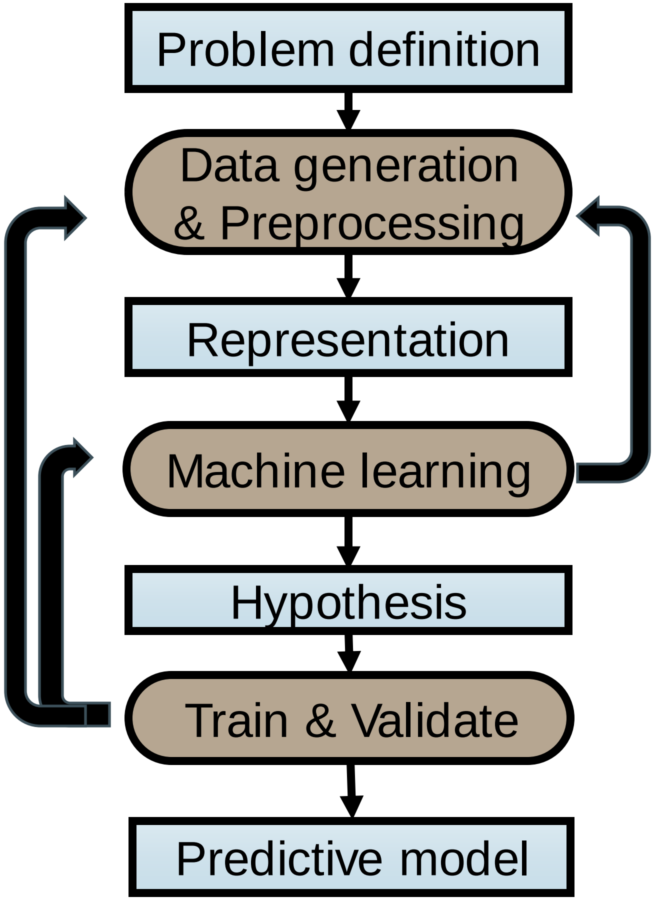


Each step generates many problems:

- Data generation: data types, corpus size, online/offline
- Preprocessing: representation, sampling, noise
- Machine learning: learning paradigm/algorithm
- Train & Validate: evaluation, loss, deployment

Outline

- Problems (NLP in Practice)
- Supervised Learning
  - Linear Classification (Sec 1)
- Representation (Sec 2)
- Supervised Learning
  - Deep Neural Network (Sec 3, 4)
- Unsupervised Learning
  - Clustering (Sec 5)
  - Self-supervised Learning (Sec 6)
- Evaluation (NLP in Practice)


Problem definition
Data generation & Preprocessing
Representation
Machine learning
Hypothesis
Train & Validate
Predictive model