DS-GA 3001.007 Introduction to Machine Learning

Lecture 1

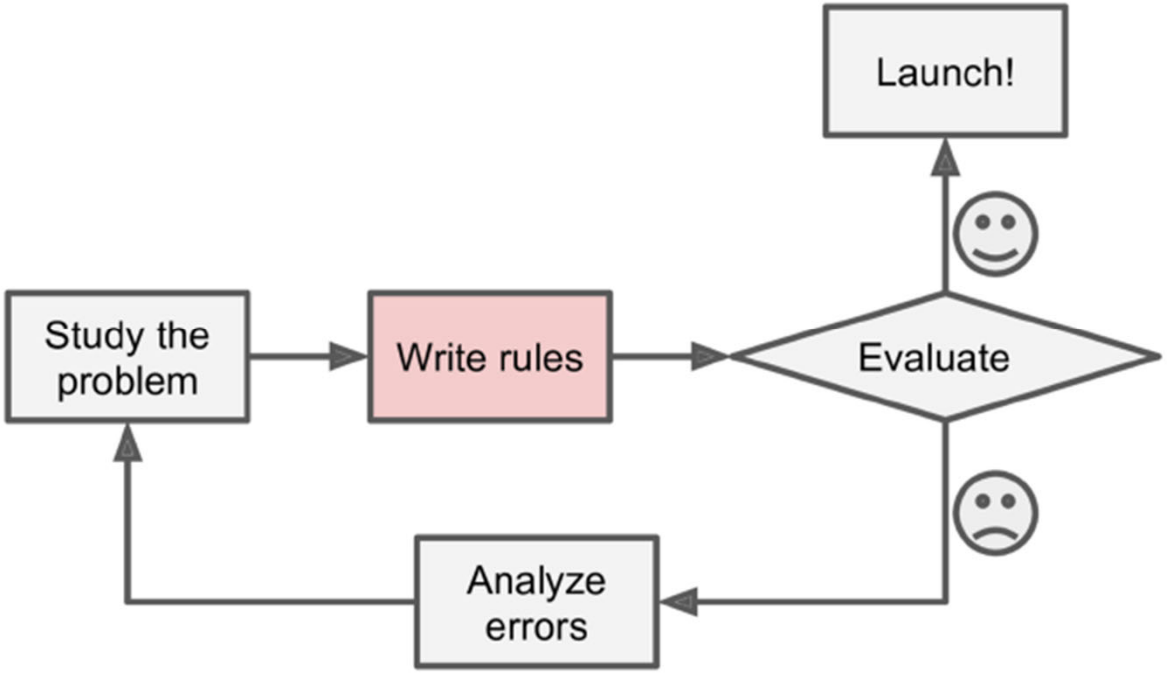
Agenda

- Overview
  - What is ML?
  - Who uses ML?
  - Why study ML?
- Lesson
  - What are the different types?
  - What are the components of an application?
  - How to use ML for data science?
- Demo
  - Churn analysis

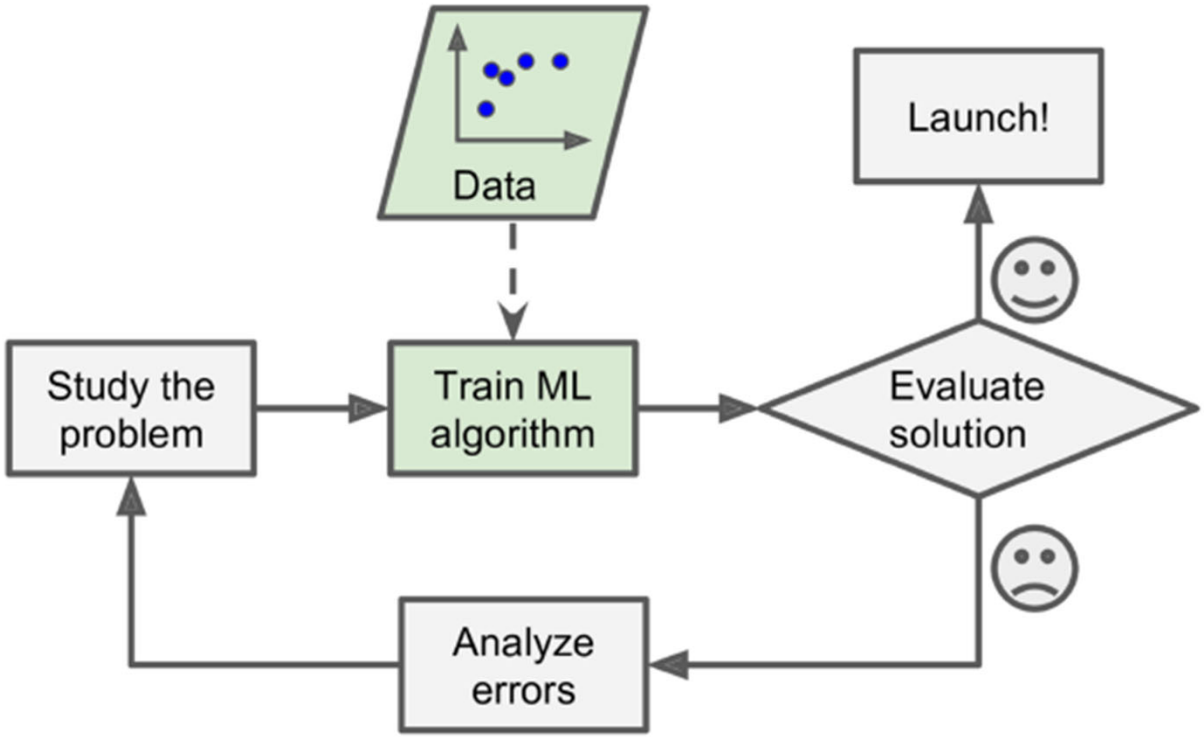
What is machine learning?

A computer program learns from experience ***E*** with respect to some task ***T*** and some performance measure ***P***, if its performance on task ***T***, as measured by ***P***, improves with experience ***E***.

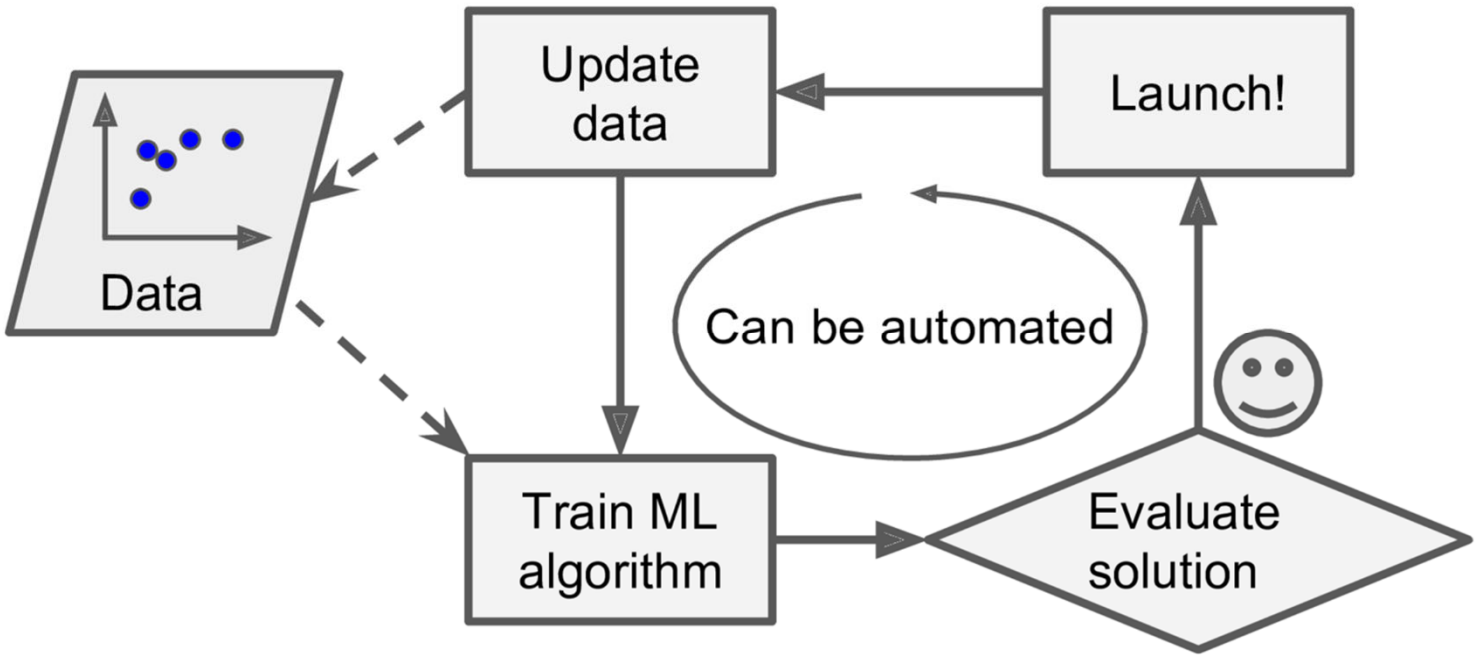
What is machine learning?



What is machine learning?



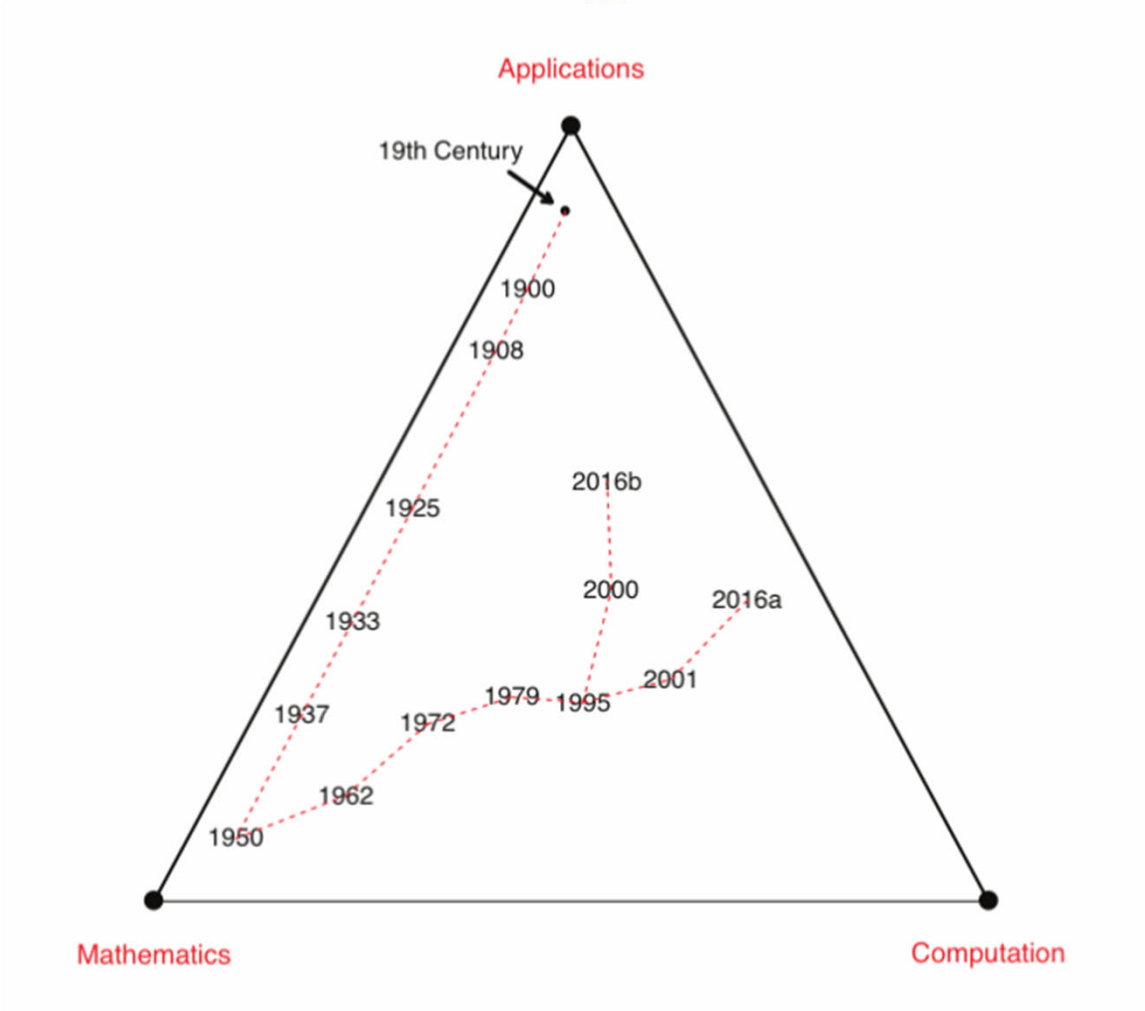
What is machine learning?



Who uses machine learning?

- Natural Language Processing
  - Text classification
  - Part of speech recognition
- Computer Vision
  - Character recognition
  - Image retrieval
- Speech Recognition
  - Source separation
  - Speaker Identification

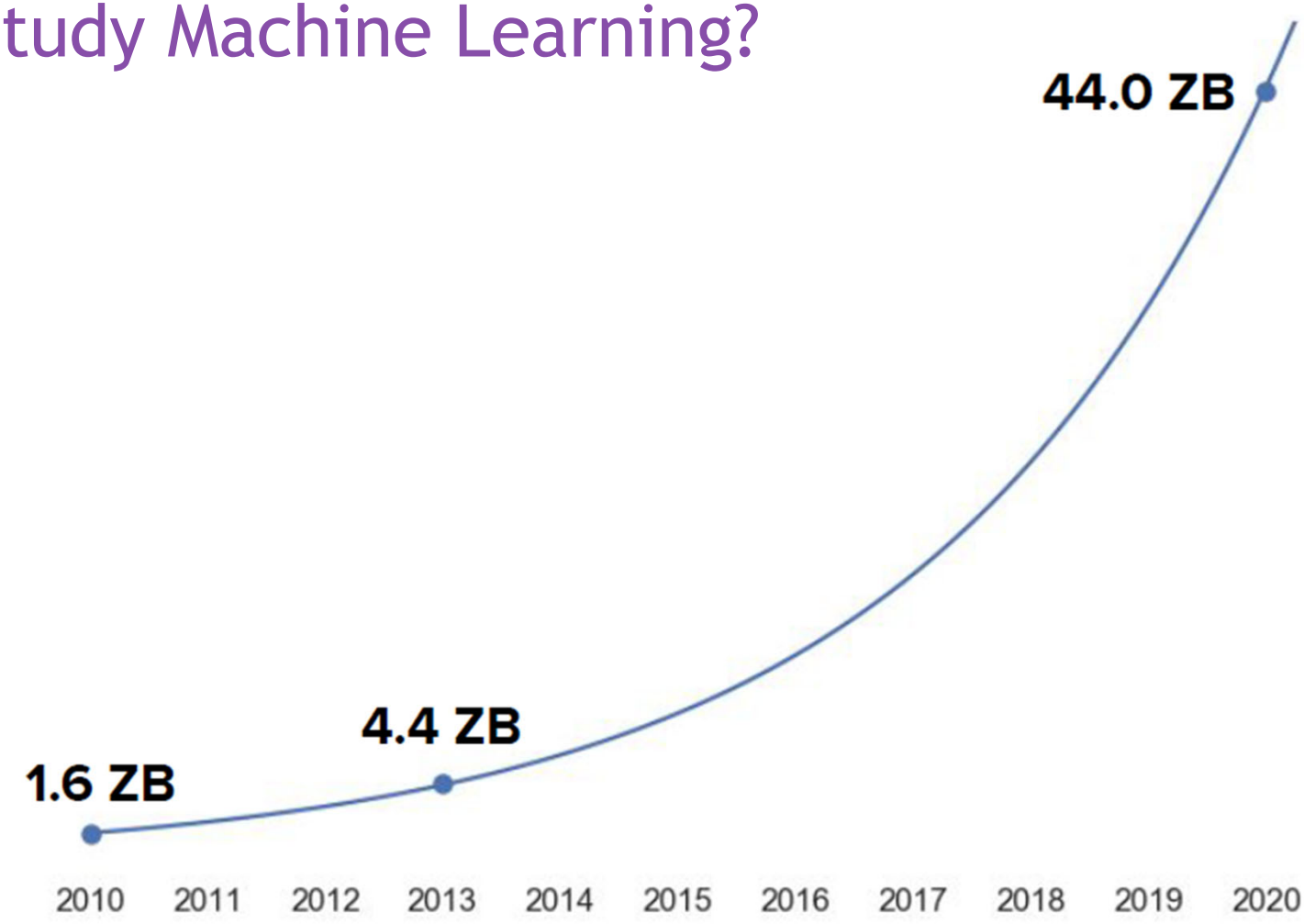
Who uses Machine Learning?



Why study Machine Learning?

- Adaptive Systems
  - Deploy without hard-coded routines
- Scalable Systems
  - Determine patterns in large and complex sets of observations

Why study Machine Learning?



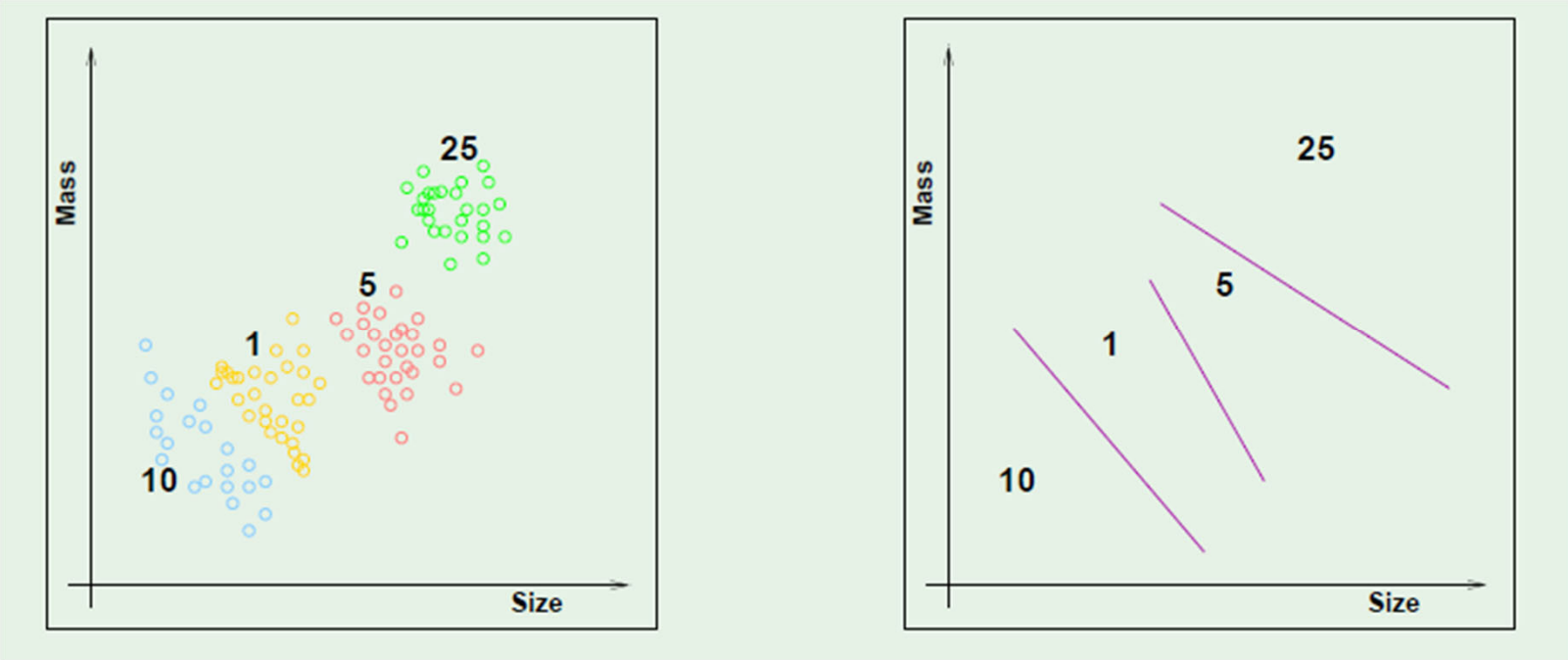
DS-GA 3001.007 Introduction to Machine Learning

- Interdisciplinary course for students in the sciences, engineering and humanities
- Comparable to DS-GA 1003
- Goals
  - Prepare
  - Empower
  - Enable

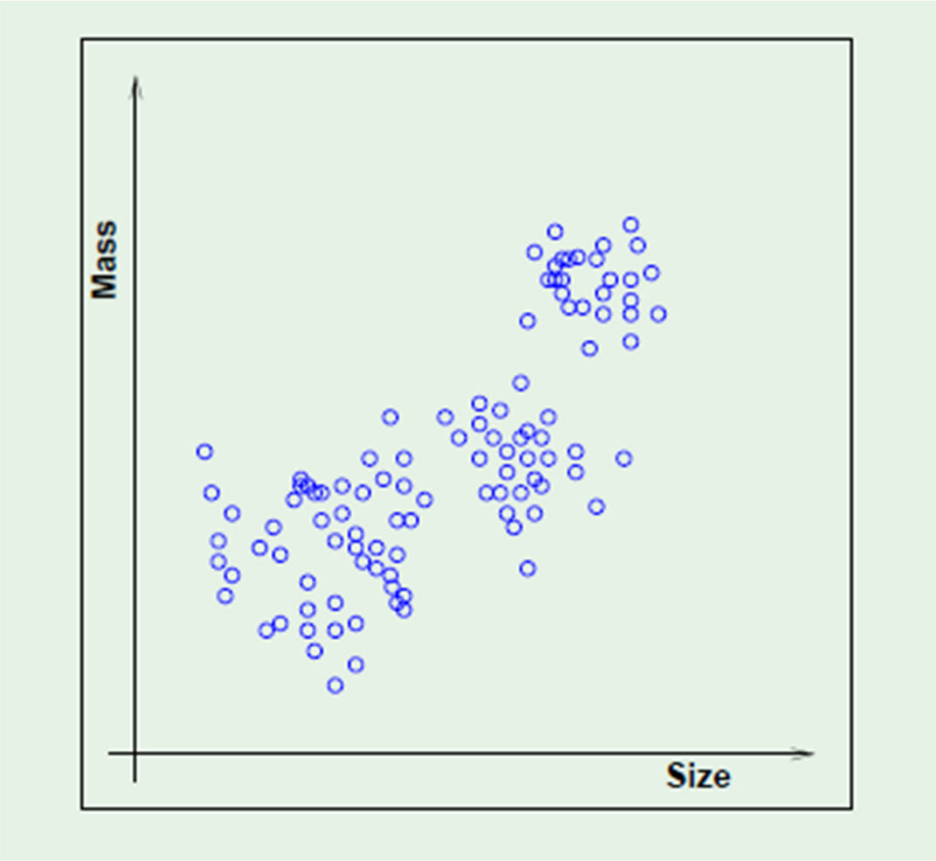
Types of Machine Learning

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Dimension Reduction
- Reinforcement Learning

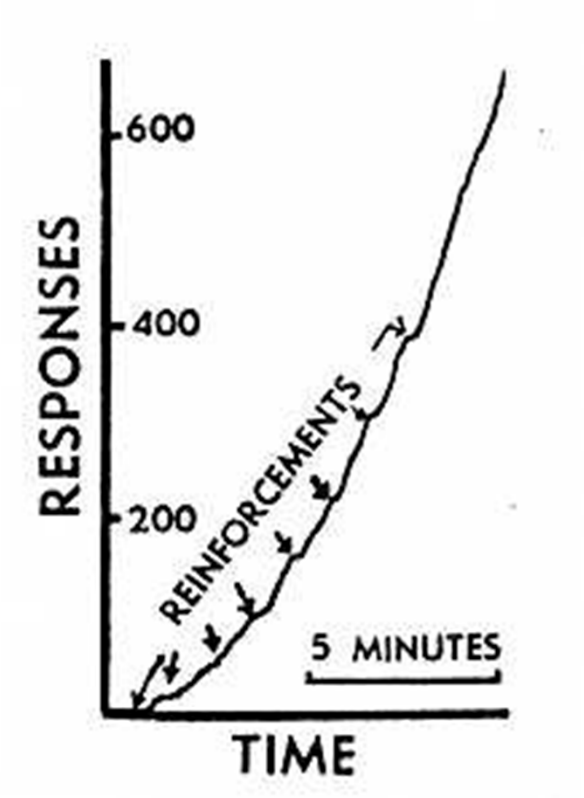
Supervised Machine Learning



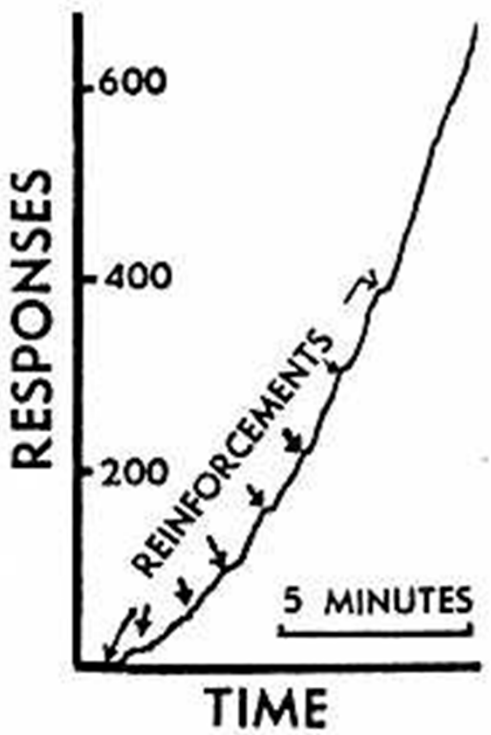
Unsupervised Machine Learning



Reinforcement Learning



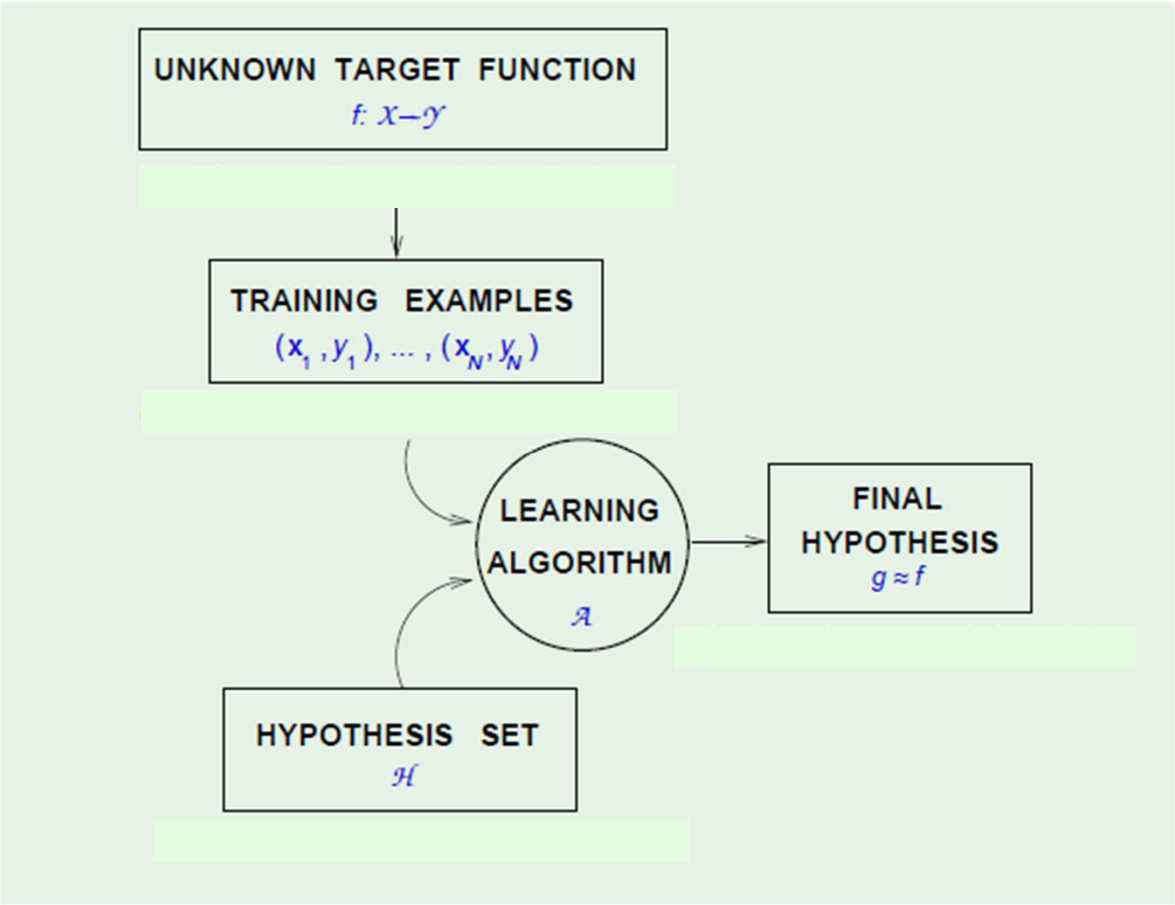
No Free Lunch



Question

- For the spam detection, can you propose examples of…
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning

Components of Machine Learning Application



Components of Machine Learning Application

- Sampling data
  - Train Sample
  - Validation Sample
  - Testing Sample

Components of Machine Learning Application

- Sampling data
  - Train Sample
  - Validation Sample
  - Testing Sample
- Input/Output
  - Features
  - Labels

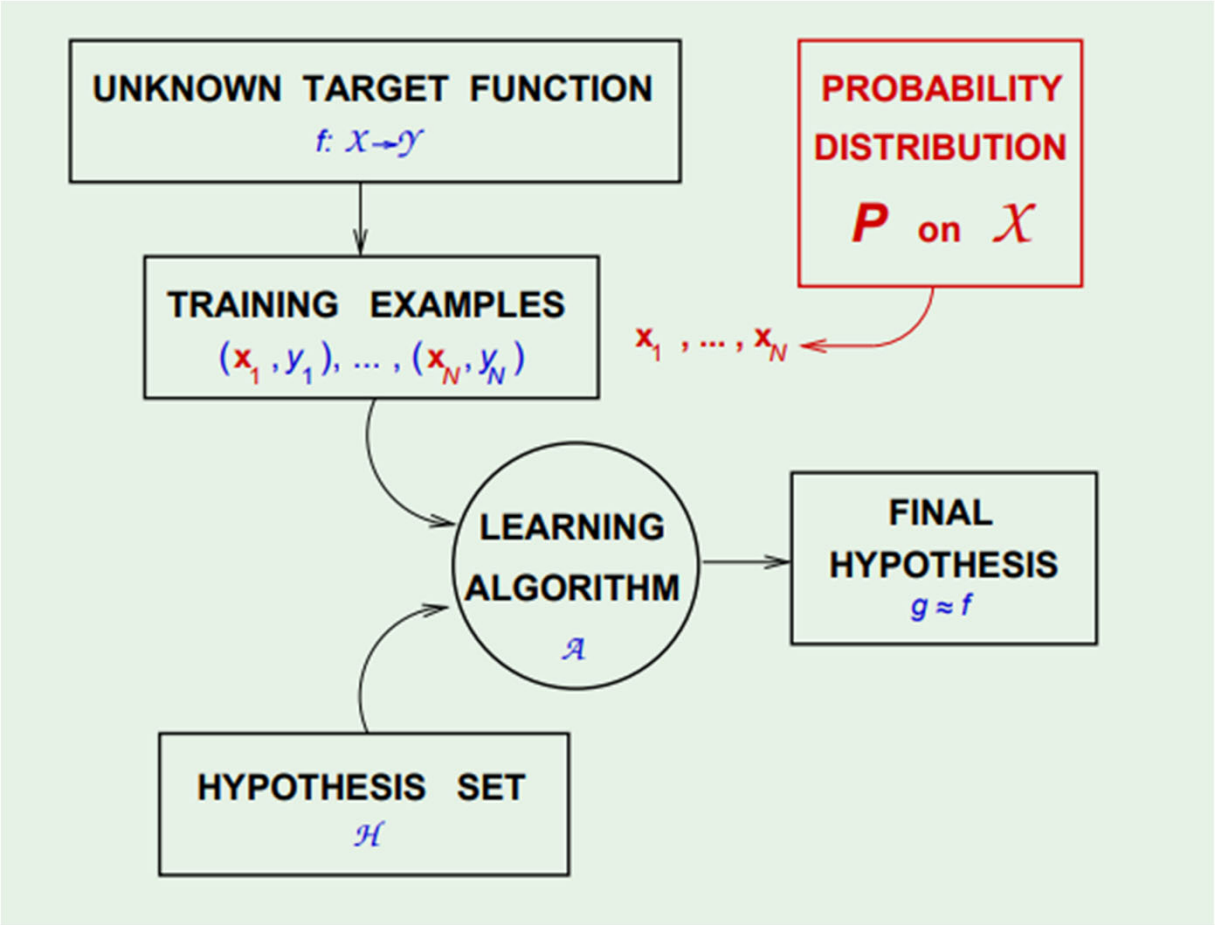
Components of Machine Learning Application

- Sampling data
  - Train Sample
  - Validation Sample
  - Testing Sample
- Input/Output
  - Features
  - Labels
- Fitting a model
  - Hypotheses
  - Hyperparameters

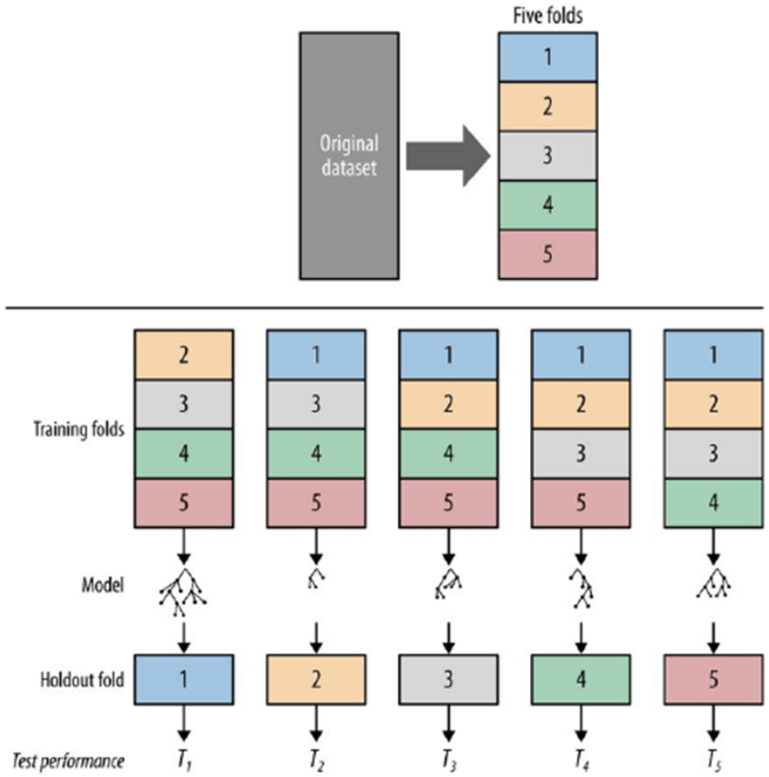
Components of Machine Learning Application

- Sampling data
  - Train Sample
  - Validation Sample
  - Testing Sample
- Input/Output
  - Features
  - Labels
- Fitting a model
  - Hypotheses
  - Hyperparameters
- Error Analysis
  - Loss functions
  - Accuracy Metrics

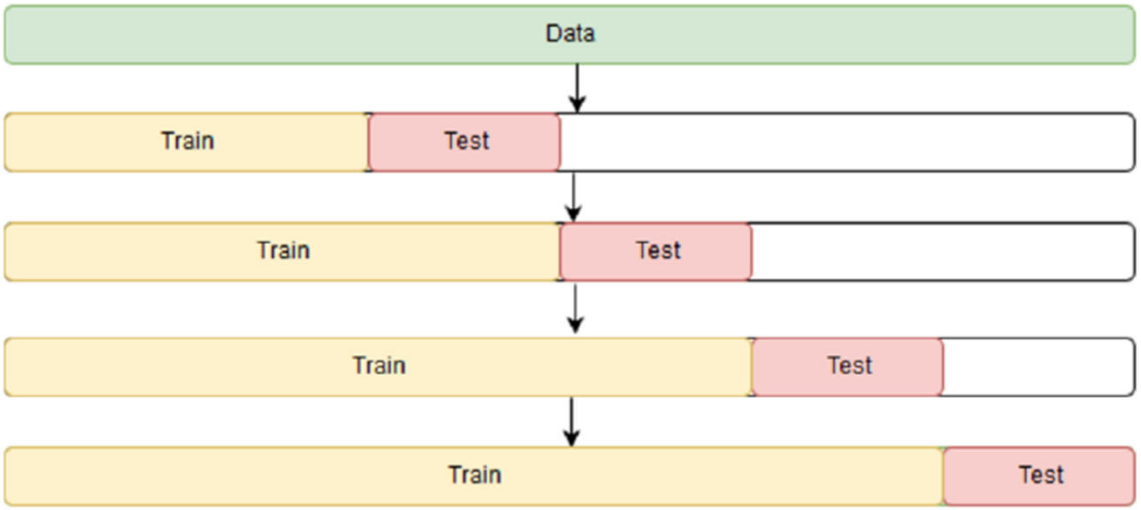
Components of Machine Learning Application



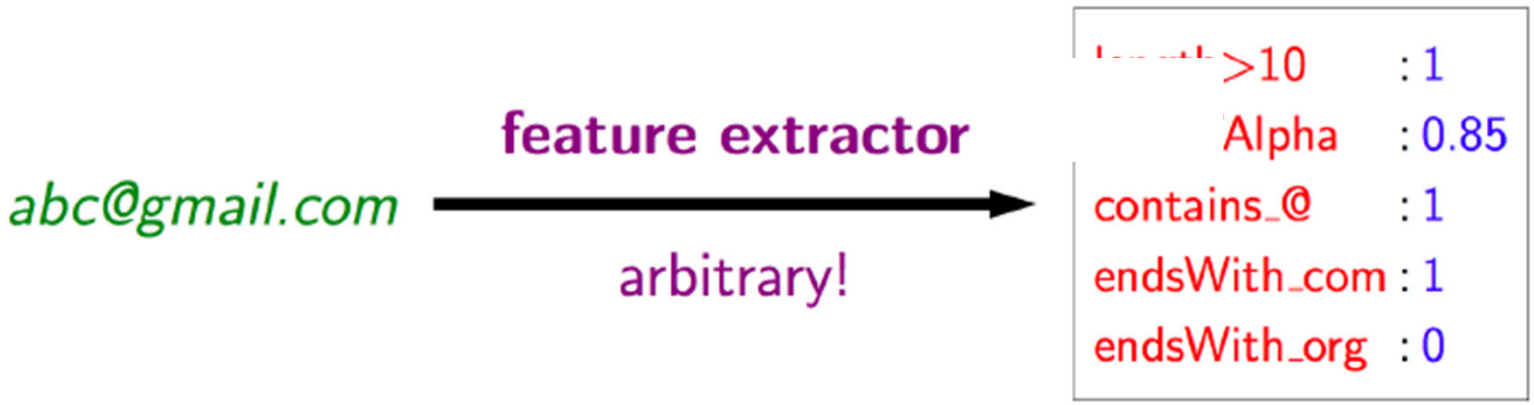
Train, Validate, Test



Train, Validate, Test



Features and Labels

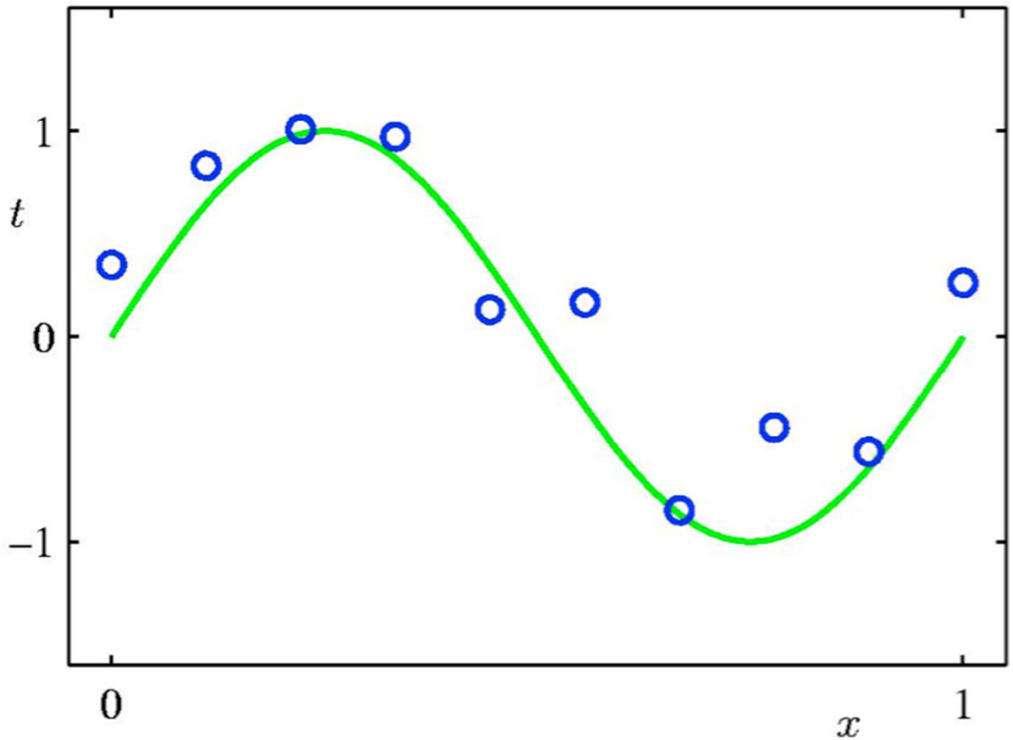


Features and Labels

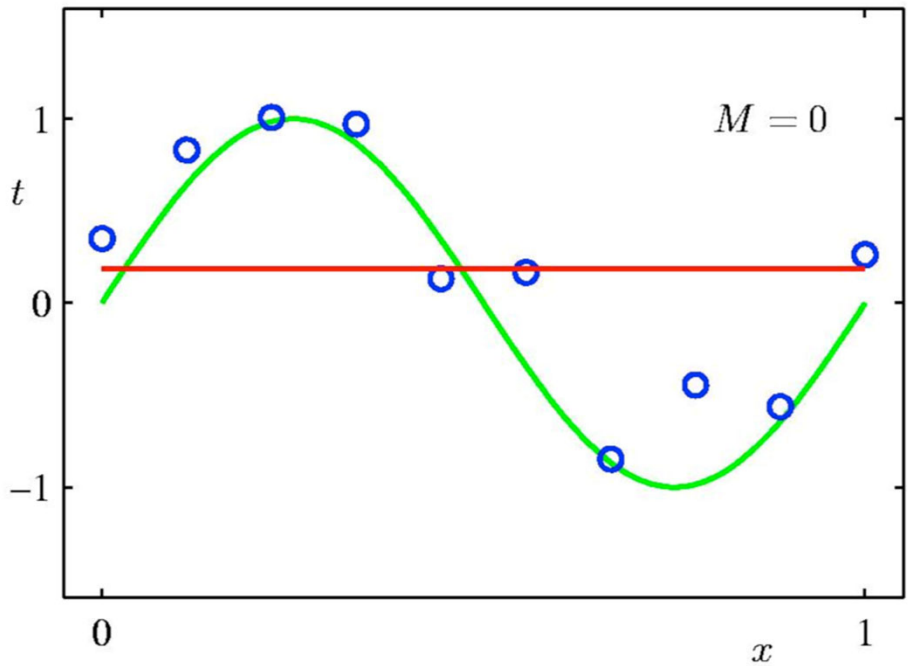
abc@gmail.com →

endsWith_aaa	: 0
endsWith_aab	: 0
endsWith_aac	: 0
...	
endsWith_com	: 1
...	
endsWith_zzz	: 0

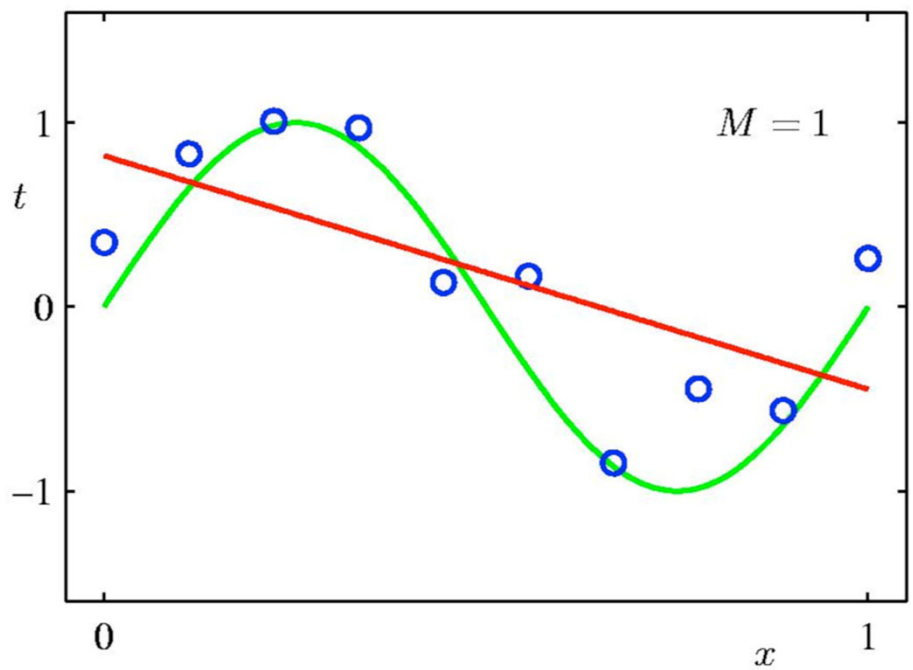
Hypotheses and Hyperparameters



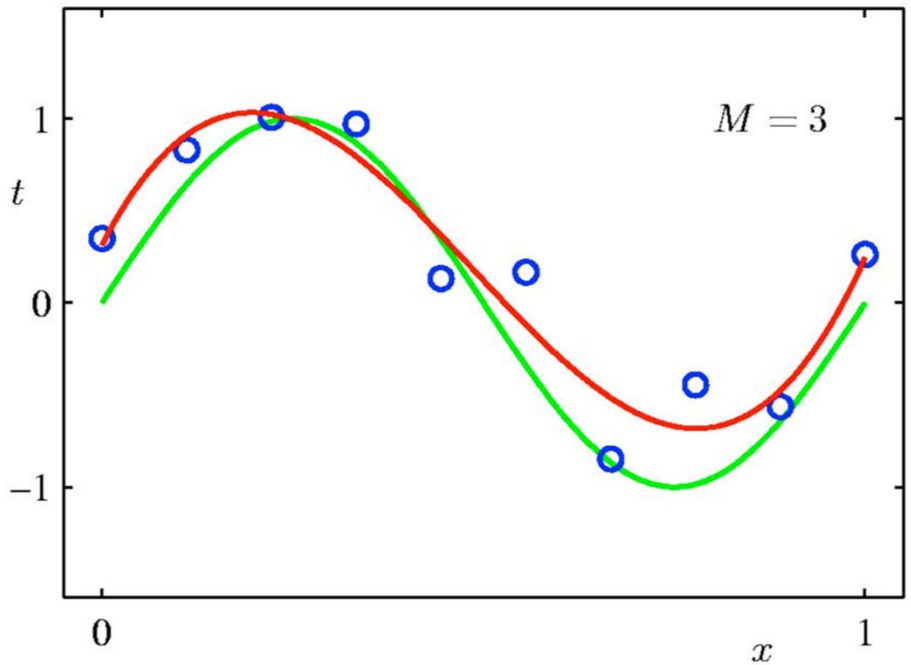
Hypotheses and Hyperparameters



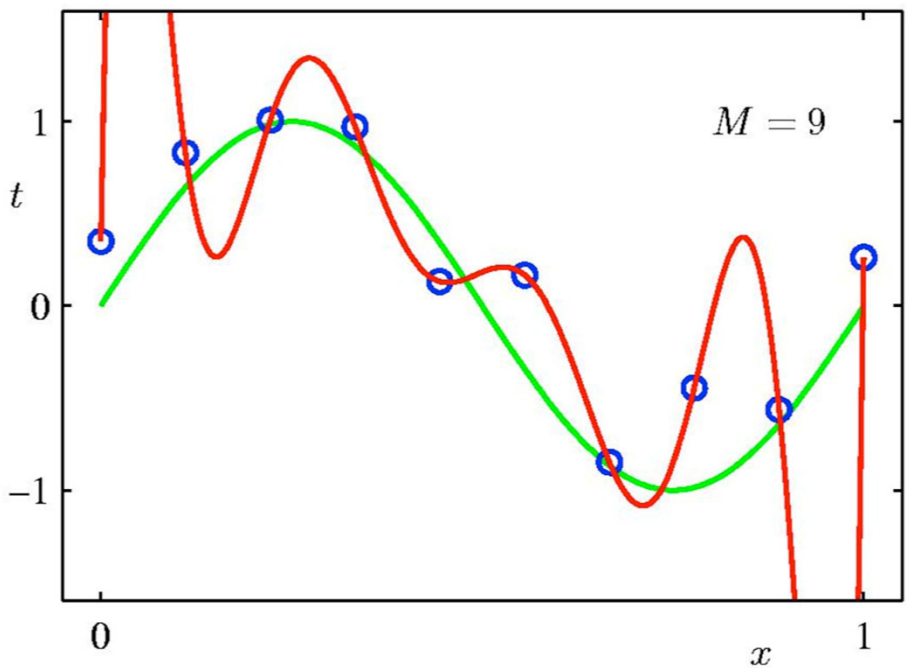
Hypotheses and Hyperparameters



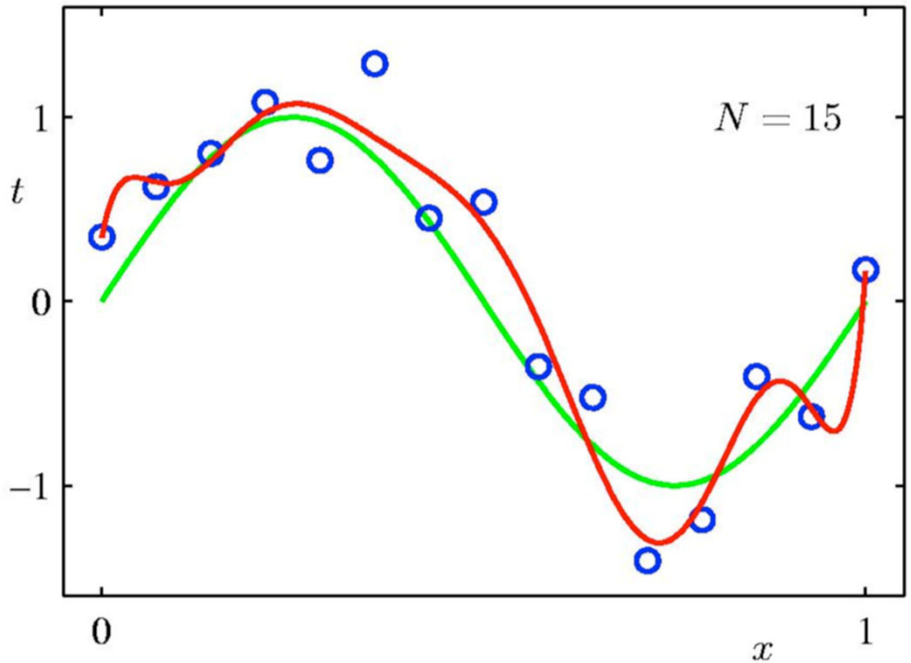
Hypotheses and Hyperparameters



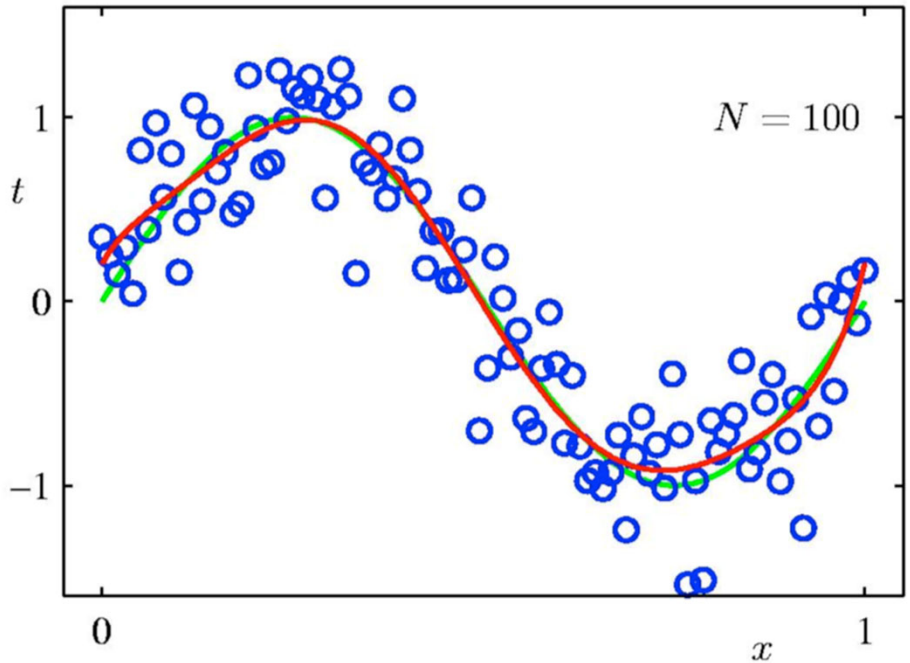
Hypotheses and Hyperparameters



Hypotheses and Hyperparameters



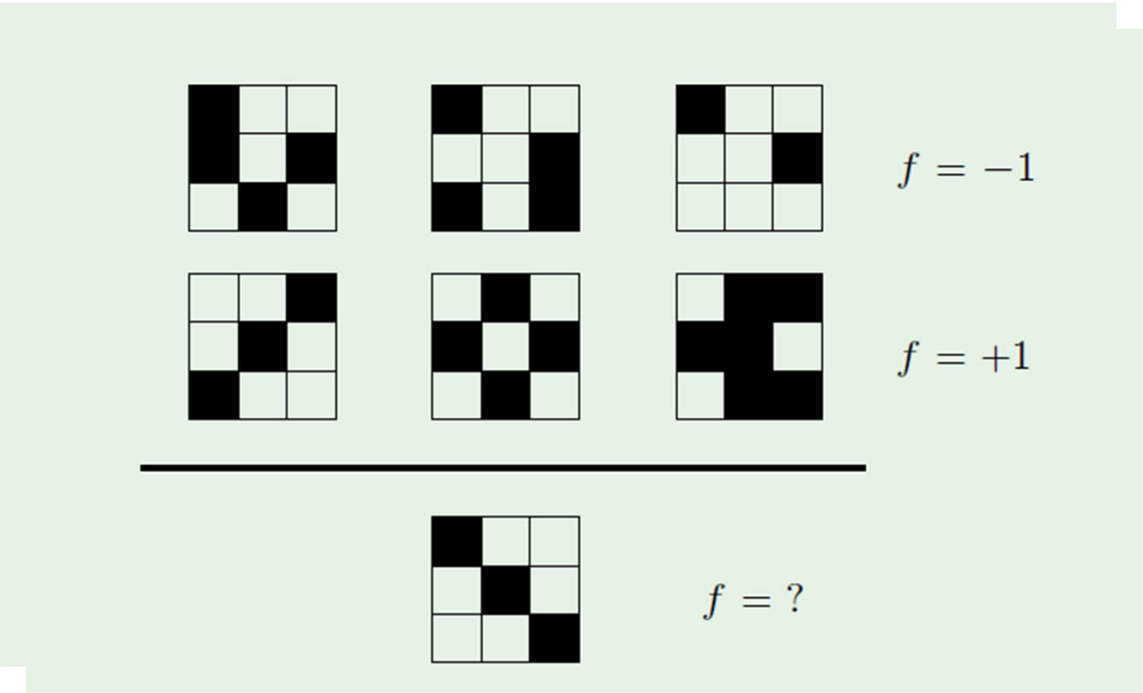
Hypotheses and Hyperparameters



Loss Function and Metrics

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Question



Question

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Question

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Question

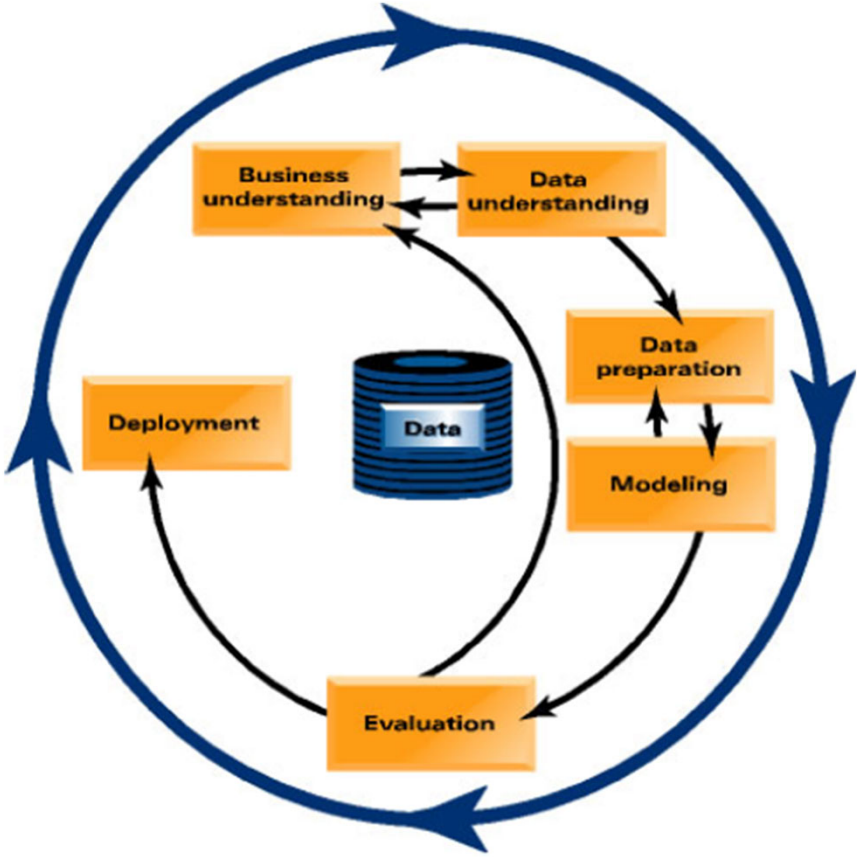
Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

How to use machine learning for data science?

- Components of Data Science
  - Formulate a problem
  - Gather data
  - Explore data
  - Determine a model for prediction and inference
  - Evaluate findings

How to use machine learning for data science?



How to use machine learning for customer churn?

The Data

1. The historical data consist of 39,859 customers. The historical data contain 19,901 customers that churned (*i.e.* left the company) and 19,958 that did not churn (see the "churndep" variable).
3. Here are the data set's 11 possible predictor variables for churning behavior:

Position	Variable Name	Variable Description
1	revenue	Mean monthly revenue in dollars
2	outcalls	Mean number of outbound voice calls
3	incalls	Mean number of inbound voice calls
4	months	Months in Service
5	eqpdays	Number of days the customer has had his/her current equipment
6	webcap	Handset is web capable
7	marryyes	Married (1=Yes; 0=No)
8	travel	Has traveled to non-US country (1=Yes; 0=No)
9	pcown	Owns a personal computer (1=Yes; 0=No)
10	credited	Possesses a credit card (1=Yes; 0=No)
11	retcalls	Number of calls previously made to retention team

The dependent variable, Churndep, = 1 if the customer churned, = 0 otherwise.

Reminders

- Section
  - Section 008 Thursdays 2:25-3:15 pm (60th 5th Avenue, Room 115)
  - Access to https://iml-f19.jupyter.hpc.nyu.edu
- Syllabus
  - Please review the course policies about assignments and grading.
- Surveys
  - Please complete Survey 1