Lecture 01. Course Overview

Nakul Gopalan

Post-doctoral Fellow,

College of Computing, Georgia Tech

This course is made with slides developed from multiple sources including Mahdi Roozbahani, Rodrigo Borela Valente, Eric Eaton, Michael Littman, Byron Boots. I will try to attribute the sources as often as I can.

Meet the Instructor

Nakul Gopalan – Nakul would do

He / him / his

Email: ngopalan3@gatech.edu

Research areas: Language grounding, planning, reinforcement learning, language understanding

Meet the TAs

Head TA:
Vidisha Goyal

TAs:
Zheng Zhang
Christopher J Banks
Aryan J Pariani
Kevin Y Li

How to succeed in this course?

Ask questions
It is your fundamental right as a student to ask questions.
Be inquisitive. I am not known for delivering monologues.

Staying motivated in the lectures over time. You have the opportunity to ask questions directly to me, and you should use it.

I need a response **full of energy** from **all** students.
It gives me a good feeling and shows me you are here to learn new exciting things. That way I stay motivated.

But wait, how are we going to do this?

Refer to:

Class Website

For anything (updates, lectures, logistics, and so on) related to this class

I verbally ask questions in the class

Sometimes I ask questions about previous lectures or something that you have already learnt.

Answer the question even if you think you are wrong (nobody loses point answering the question wrong). It will help you and other students to understand concepts much better.

piazza (the best source for Q/A)

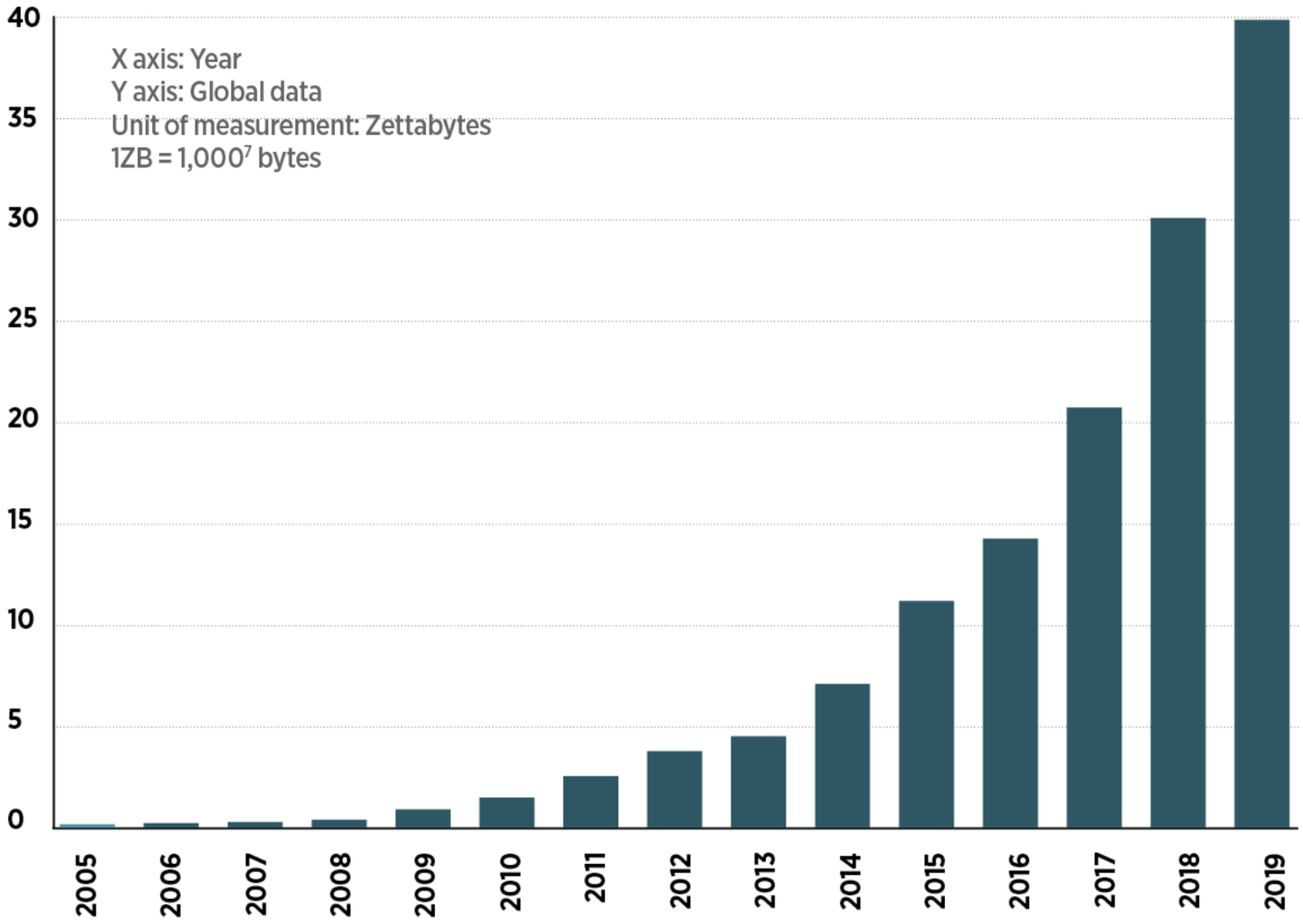
Ask your questions in piazza (make it public to other students), and also please see other questions in piazza, it might answer your question. (Please do not send me or TAs Emails regarding hw questions, exams, and other logistics – you can also ask “private question” on piazza) =>Class participation

Bonus points: Undergrad and grad

Machine Learning

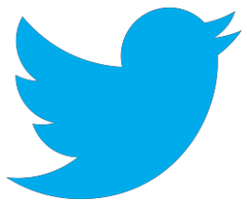
"***We are drowning in information but starved for knowledge.***"

— John Naisbitt



The Booming Age of Data

30 trillion Web pages



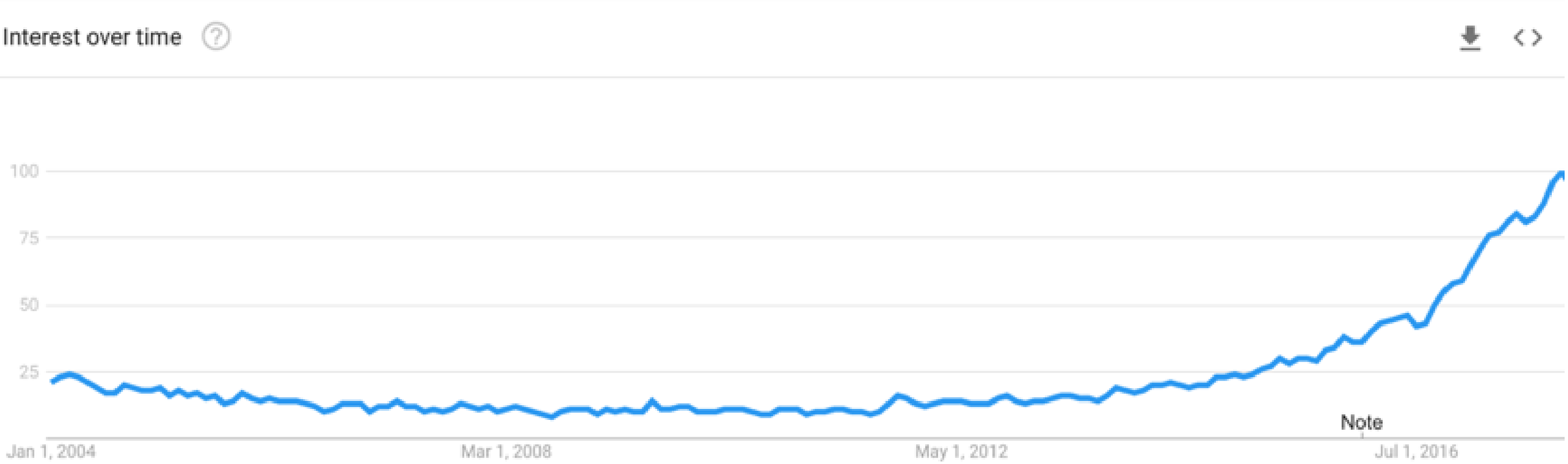
500 million tweets per day

2.27 billion monthly active users

1.8 billion images uploaded to Internet per day

2.9 billion base pairs in human genome

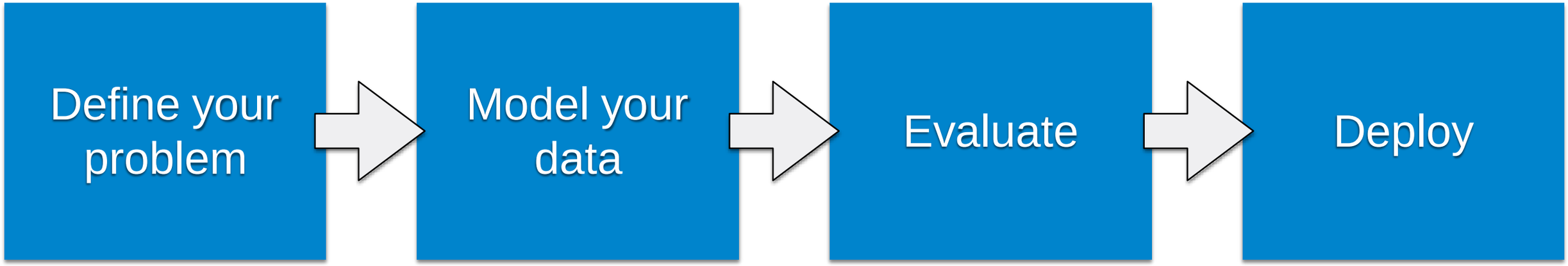
Interest in machine learning



Google trends, “Machine Learning (field of study)”.

Machine Learning

Machine Learning is the process of **turning data into actionable knowledge** for **task support** and **decision making**.



Course Objectives

- Introduce to you the **pipeline of Machine Learning**
- Help you understand **major machine learning algorithms**
- Help you learn to **apply tools** for **real data analysis problems**
- Encourage you to **do research** in data science and machine learning

Brief History of Machine Learning

1950s

Samuel's checker player

Selfridge's Pandemonium

1960s:

Neural networks: Perceptron

Pattern recognition

Learning in the limit theory

Minsky and Papert prove limitations of Perceptron

1970s:

Symbolic concept induction

Winston's arch learner

Expert systems and the knowledge acquisition bottleneck

Quinlan's ID3

Michalski's AQ and soybean diagnosis

Scientific discovery with BACON

Mathematical discovery with AM (Automated Mathematician)

Slide Credit: Ray Mooney

Brief History of Machine Learning

1980s:

Advanced decision tree and rule learning
Explanation-based Learning (EBL)
Learning and planning and problem solving
Utility problem
Analogy
Cognitive architectures
Resurgence of neural networks (connectionism, backpropagation)
Valiant's PAC Learning Theory
Focus on experimental methodology

1990s

Data mining
Adaptive software agents and web applications
Text learning
Reinforcement learning (RL)
Inductive Logic Programming (ILP)
Ensembles: Bagging, Boosting, and Stacking
Bayes Net learning

Slide Credit: Ray Mooney

Brief History of Machine Learning

2000s:

Support vector machines
Kernel methods
Graphical models
Statistical relational learning
Transfer learning
Sequence labeling
Collective classification and structured outputs
Computer Systems Applications
Learning in robotics and vision

2010s:

Deep learning
Reinforcement learning
Generative models
Adversarial learning
Muti-task learning
Learning in NLP, CV, Robotics, …

Based on Ray Mooney's slide

Syllabus

Part I: Basic math for computational data analysis

- Probability, statistics, linear algebra

Part II: Supervised learning for predictive analysis

- Tree-based models, linear classification/regression, neural networks

Part III: Unsupervised learning for data exploration

- Clustering analysis, dimensionality reduction, kernel density estimation

Part IV: Advanced topics for learning behaviors

- Reinforcement learning, Hidden Markov Models

The classic question Cat or Dog

Unsupervised and Supervised learning

	Weight(lb)	Height(cm)	Fur color	Eye color		Label
Point 1	10	20	w	g		cat
Point 2	50	100	br	bl		dog
Point 3	8	15	bl	bl	=	dog
Point 4	12	25	w	bl		cat
Point 5	14	10	bl	g		dog
				$X_{n\times d}$		$Y_{n\times 1}$

Unsupervised just focuses on $X_{n\times d}$

Supervised focus on $X_{n\times d}$ and $Y_{n\times 1}$

Syllabus: Supervised Learning

Tree-based models

- Decision tree
- Ensemble learning/Random forest

Linear classification/regression models

- Linear regression
- Naive Bayes
- Logistic regression
- Support vector machine

Neural networks

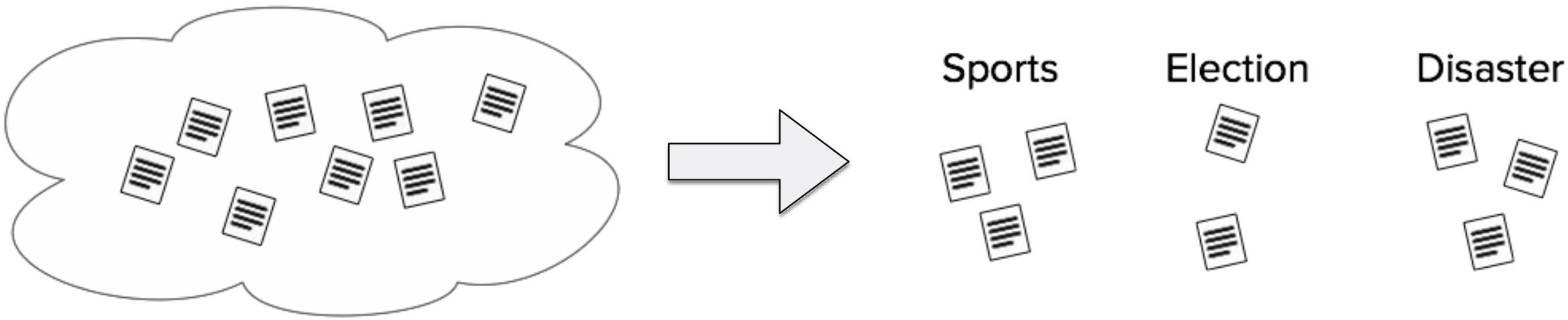
- Feedforward neural networks and backpropagation analysis

News Classification

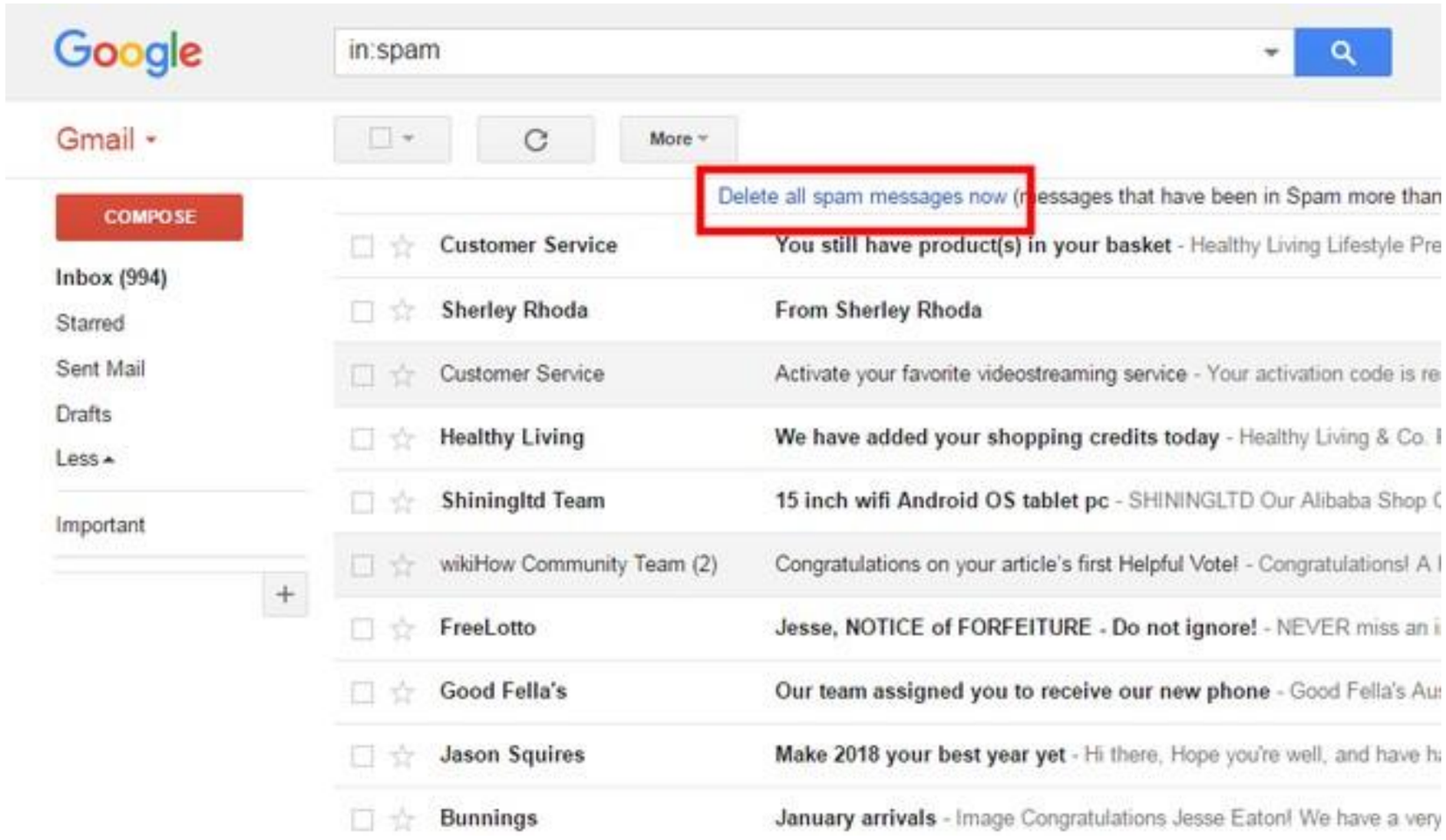
What are the inputs and how to represent them?

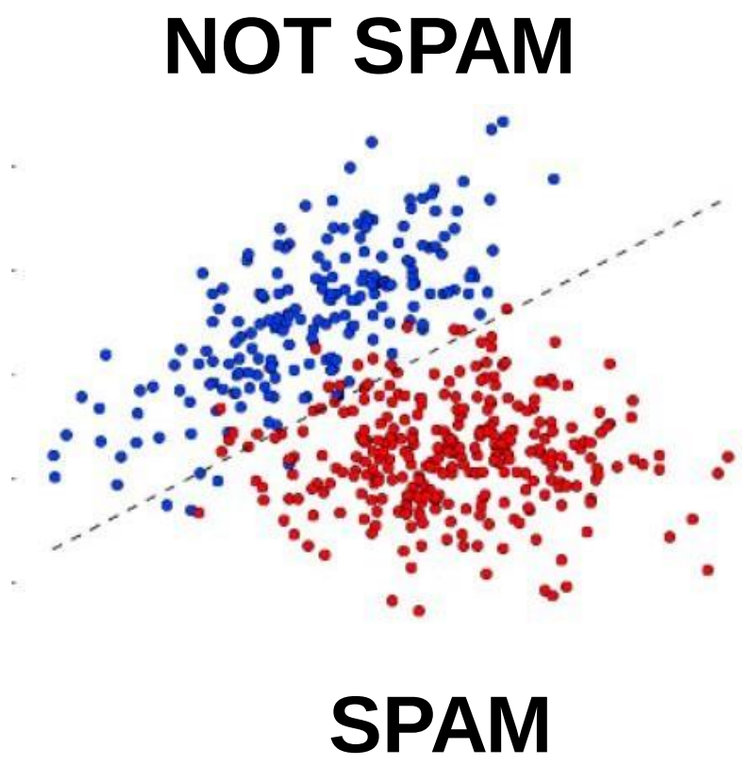
What are the desired outputs?

What learning algorithms to choose?



Spam Detection





What are the inputs and how to represent them?

What are the desired outputs?

What learning algorithms to choose?

Image credit: WikiHow

Examples

FONT TELLER

Root Inspired Anchor Model Project

Music Generation using Machine Learning

Prediction-of-Hard-Drive-Failure

Syllabus: Unsupervised Learning

Clustering Analysis

- K-means
- Gaussian mixture model
- Hierarchical clustering
- Density-based clustering
- Evaluation of clustering algorithms

Dimension Reduction

- Principal component analysis

Kernel Density Estimation

- Parametric density estimation
- Non-parametric density estimation

Community Detection in Social Networks

What are the inputs and how to represent them?

What are the desired outputs?

What learning algorithms to choose?

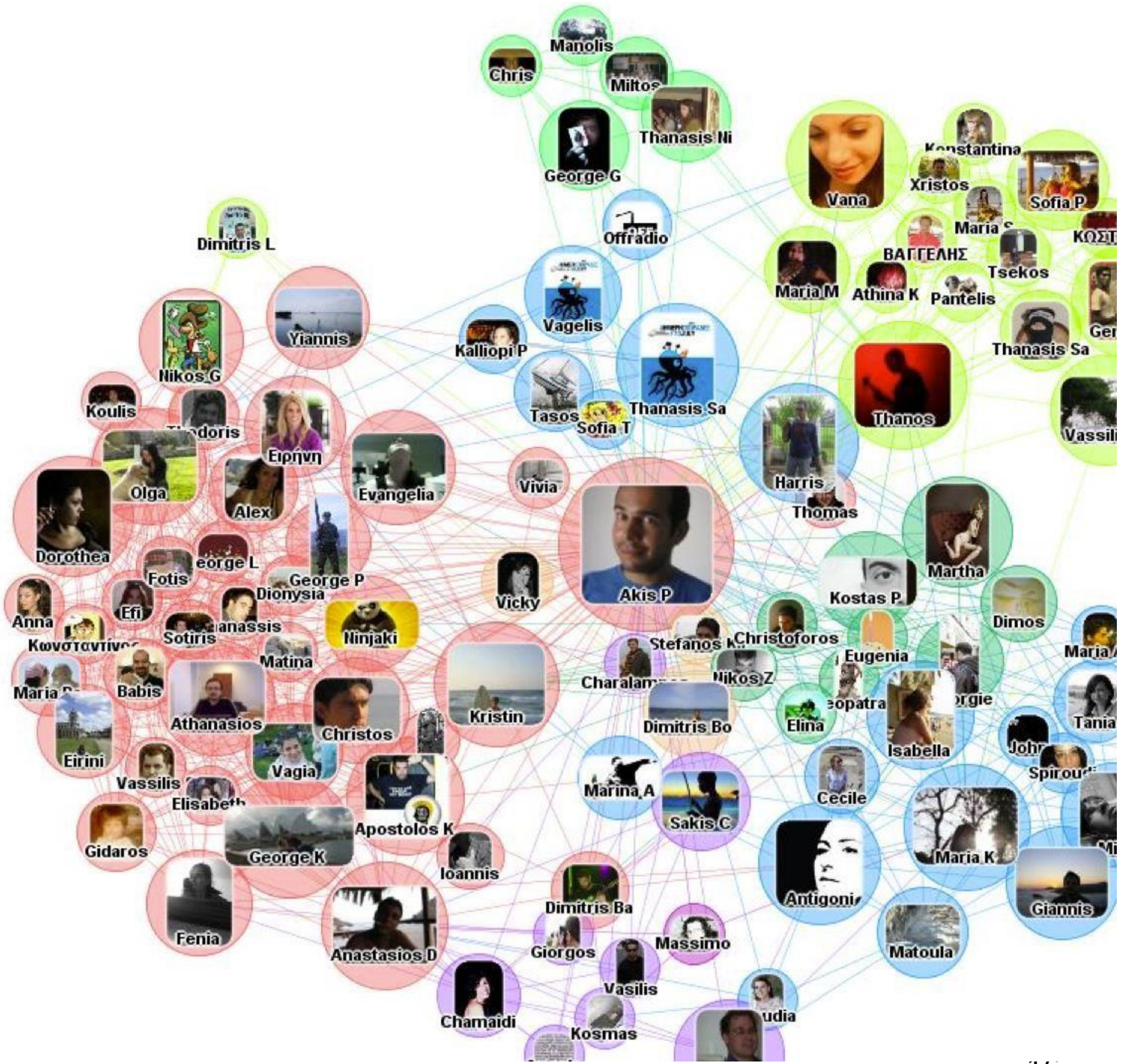
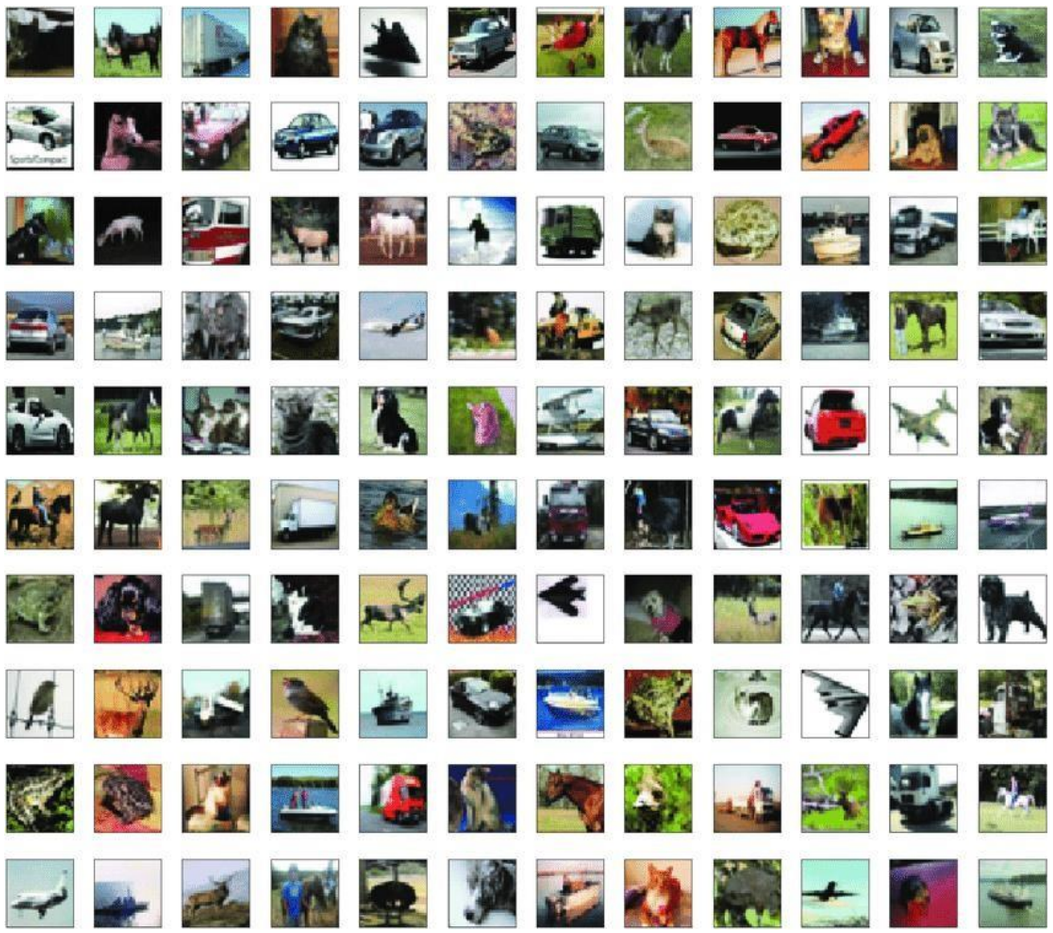
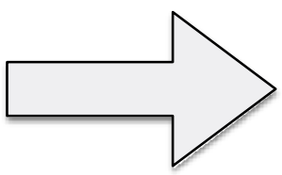
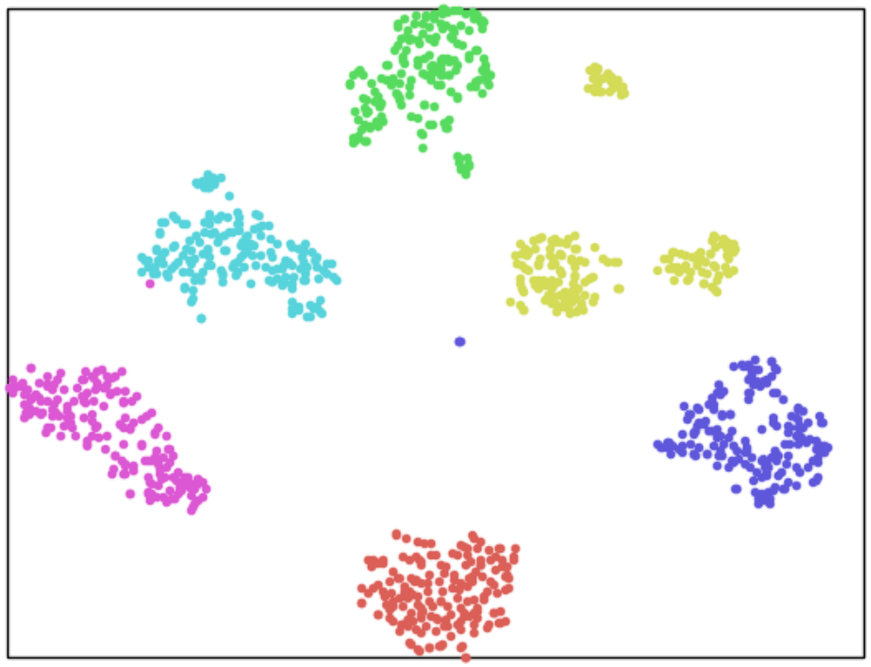


Image credit: Symeon Papadopoulos

Dimensionality Reduction



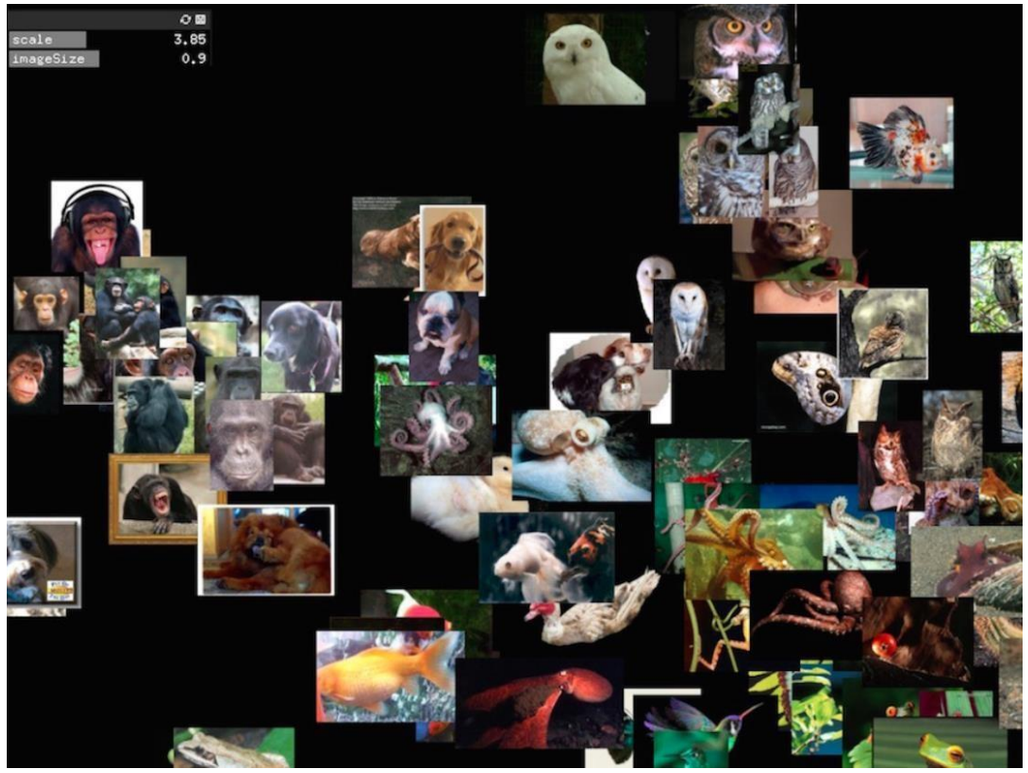




What are the inputs and how to represent them?

What are the desired outputs?

What learning algorithms to choose?



mage credit: Gene Kogan

Advanced topics

Reinforcement Learning

How does “an agent” take actions in the world?

Hidden Markov Model

How to make predictions about the future?
What is the weather tomorrow like?

Repeated theme over the course

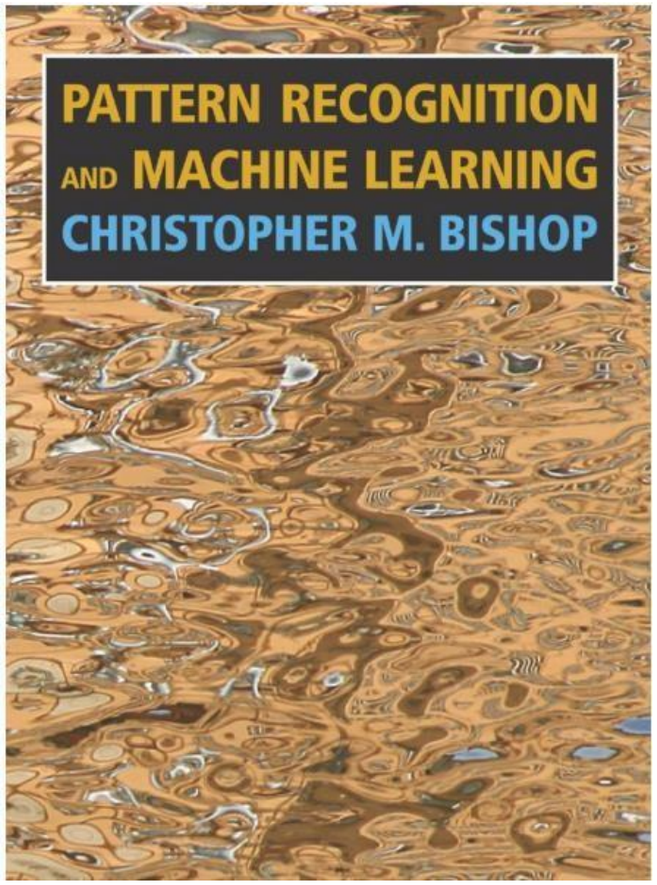
ML algorithm = representation + loss function + optimizer

Prerequisites

Basic knowledge in probability, statistics, and linear algebra

Basic programming skills in Python (Jupyter Notebook)

No background in machine learning is required



Text Books

Pattern Recognition and Machine Learning, by Chris Bishop

Other recommended books:

Learning from data, by Yaser S. Abu-Mostafa

Machine learning, by Tom Mitchell

Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Assignments

Four assignments (GradeScope)

Each can include written analysis or programming

Late policy

Three (3) late days for the entire semester to be used only on assignments.

Don’t copy

Any academic misconduct is subject to F grade as well as reporting to the Dean of students.

Projects

- Work on a real-life Machine Learning problem
  - What is the problem? What is your method? How do you evaluate it?
- Exactly 5 people in a team (Grad and undergrad can’t be mixed in a group)
- GitHub Pages (index.html)
- Start your projects early
- Ask for comments and feedbacks from the teaching staff

Quizzes

- To test knowledge weekly
- About 14 quizzes with a practice quiz
- Practice quiz out next week.
- Top 9 or 10 counted for your final grade.

Grading Policy

Assignments (50%)

Four assignments; programming or written analysis

Project (35%)

5 people in a team (not less and not more); should be done using GitHub Pages)

Quizzes (10%)

About 15 quizzes, we will consider your 10 topmost score

Class participation(5%)