Machine Learning

Instructors: Lei Li, **Yu-Xiang Wang**

About myself

Yu-Xiang Wang 王宇翔

Eugene Aas Chair Assistant Professor

Department of Computer Science,
and Department of Statistics (By Courtesy)

Co-Director, Center for Responsible Machine Learning

UC Santa Barbara

Office: Henley Hall 2013

E-mail: yuxiangw AT cs.ucsb.edu

Curriculum Vita: [link]

Yu-Xiang is pronounced approximately as ['ju:'ʃi:ʌŋ], namely, *y~eu~ee – sh~ih~ah~ng* .

Research area: Statistical Machine Learning. Optimization, reinforcement learning, differential privacy, deep learning.

Short biography:

China => Singapore
⇒ PhD from Carnegie Mellon University
⇒ Scientist at Amazon AI
⇒ Professor at UCSB

Homepage:
https://cs.ucsb.edu/~yuxiangw/

Call me: Yu-Xiang or Professor Wang. **Don't** call me: Wang, Yu, or Professor.

Professor Lei Li

Lei Li

Assistant Professor
Computer Science Department
Univeristy of California Santa Barbara

Research area: natural language processing, machine learning, data mining.

Topics:

- Machine translation, speech translation, multilingual NLP.
- Text generation and summarization.
- Reasoning and question answering.
- Information extraction.
- AI for drug discovery
- Green and Efficient ML
- Time series mining and prediction
- Probabilistic inference, Bayesian sampling methods

Looking for highly motivated students to join my lab.

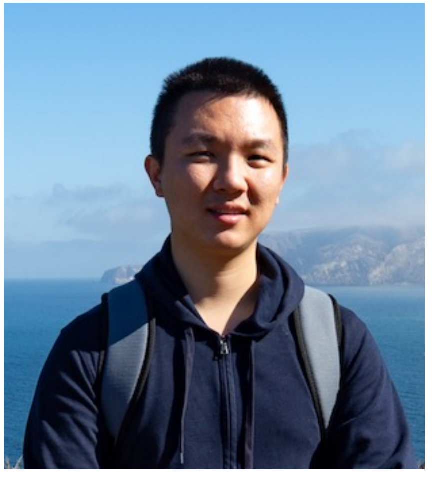
UCSB students: please feel free to contact me if you would like to do intern in my lab.

Website: https://sites.cs.ucsb.edu/~lilei/

Teaching Assistant

- TA: Xuandong Zhao

- PhD student in CS
- Research focus:
  - Machine Learning
  - Natural Language Processing
  - Privacy and confidentiality



Welcome to CS291K Machine Learning!

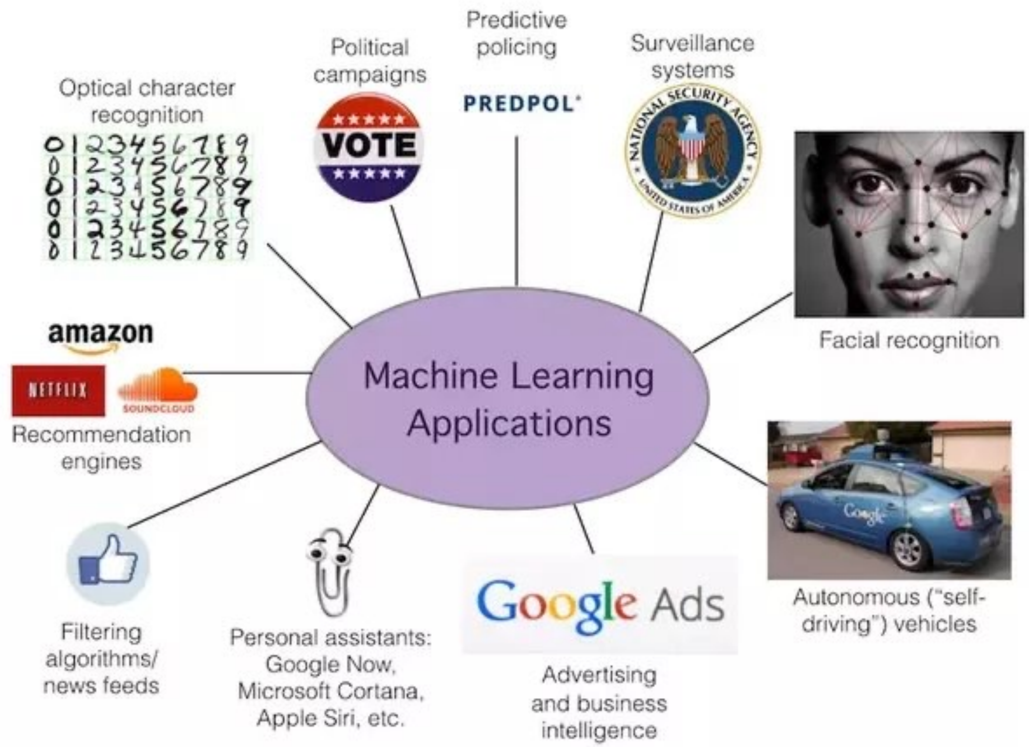
- This is a graduate-level introduction to machine learning.
  - It gives a comprehensive treatment to ML
  - It covers the theory, algorithms and practical implementation of ML all in 12 weeks!

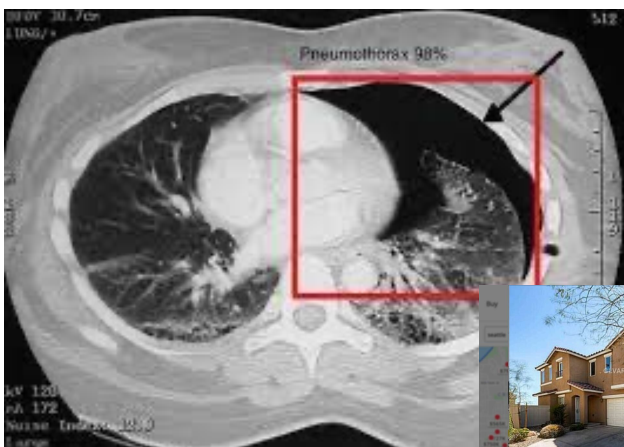
- This will be a tough course
  - Nevertheless, it's a course worth spending time on.
  - Especially if you are just starting your PhD studies.

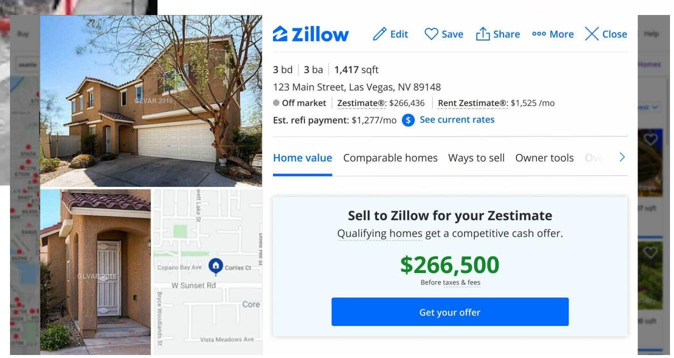
- UCSB Covid policy: https://www.ucsb.edu/COVID-19-information/campus-updates
  - it's your choice on mask-wearing, but highly recommended to keep everybody in the class.

[Discussion] Why are you taking this course?

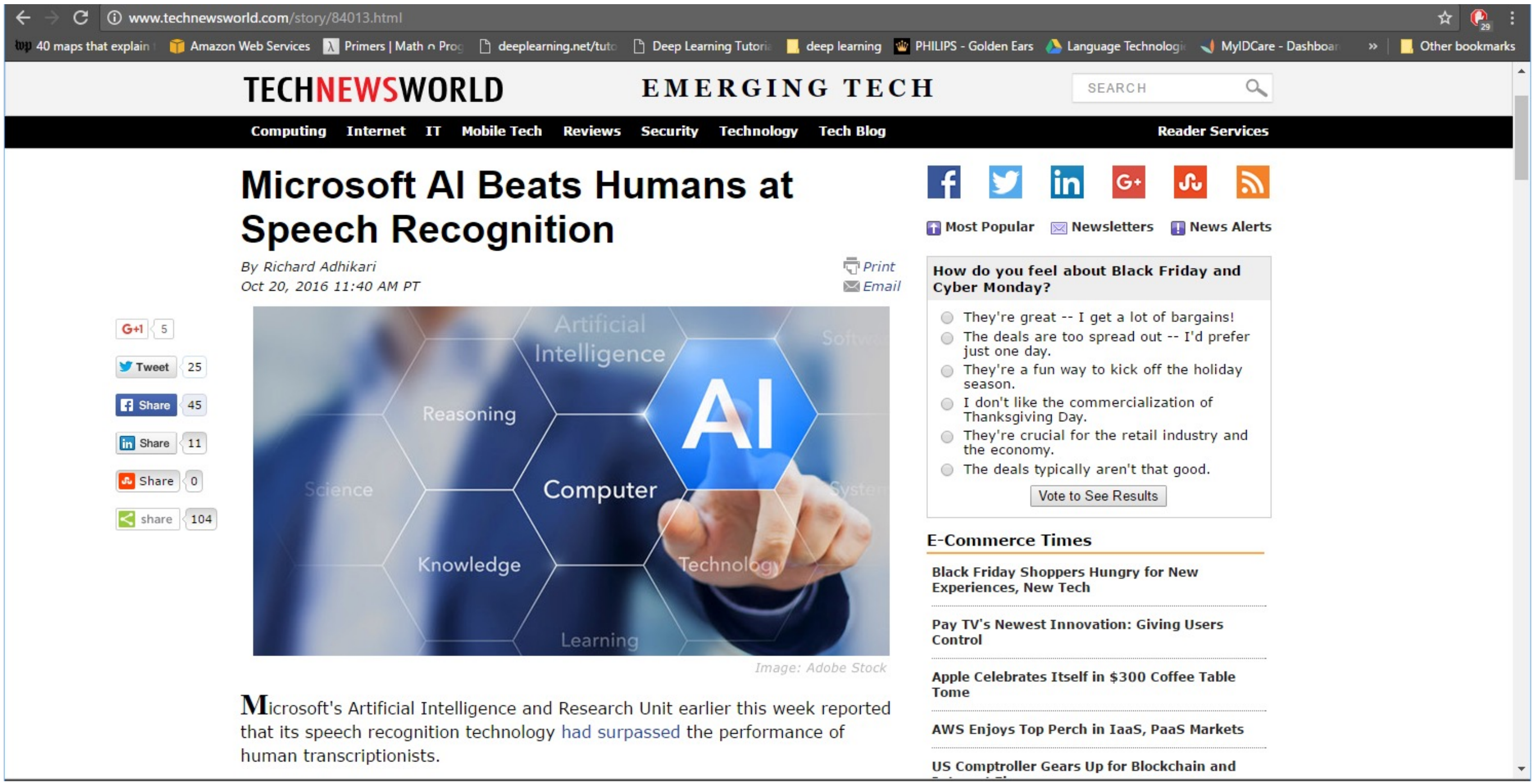
~~AI~~ Machine Learning has revolutionized almost every aspect of our daily life







Breakthroughs with deep learning



Breakthroughs with deep learning

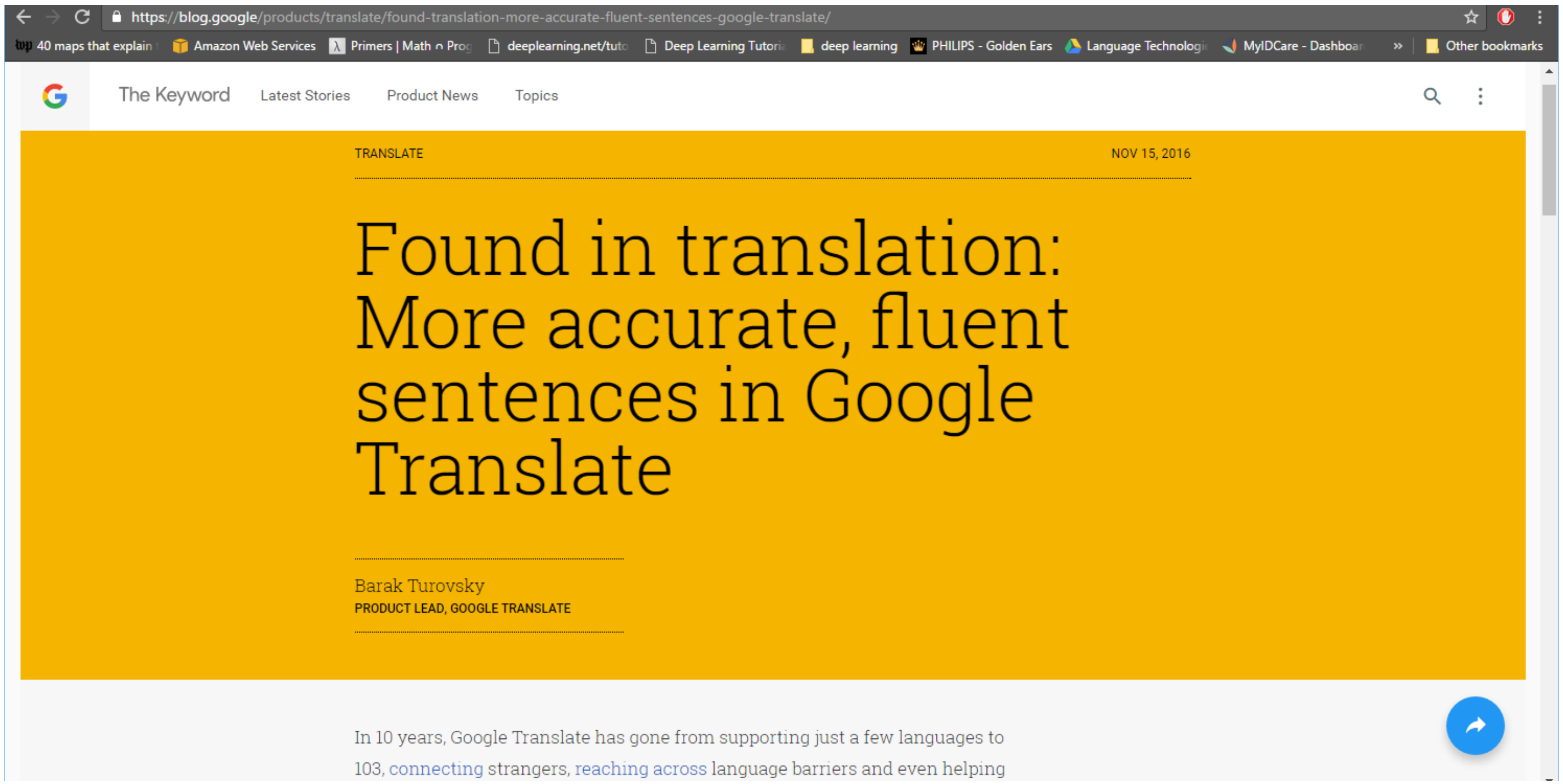
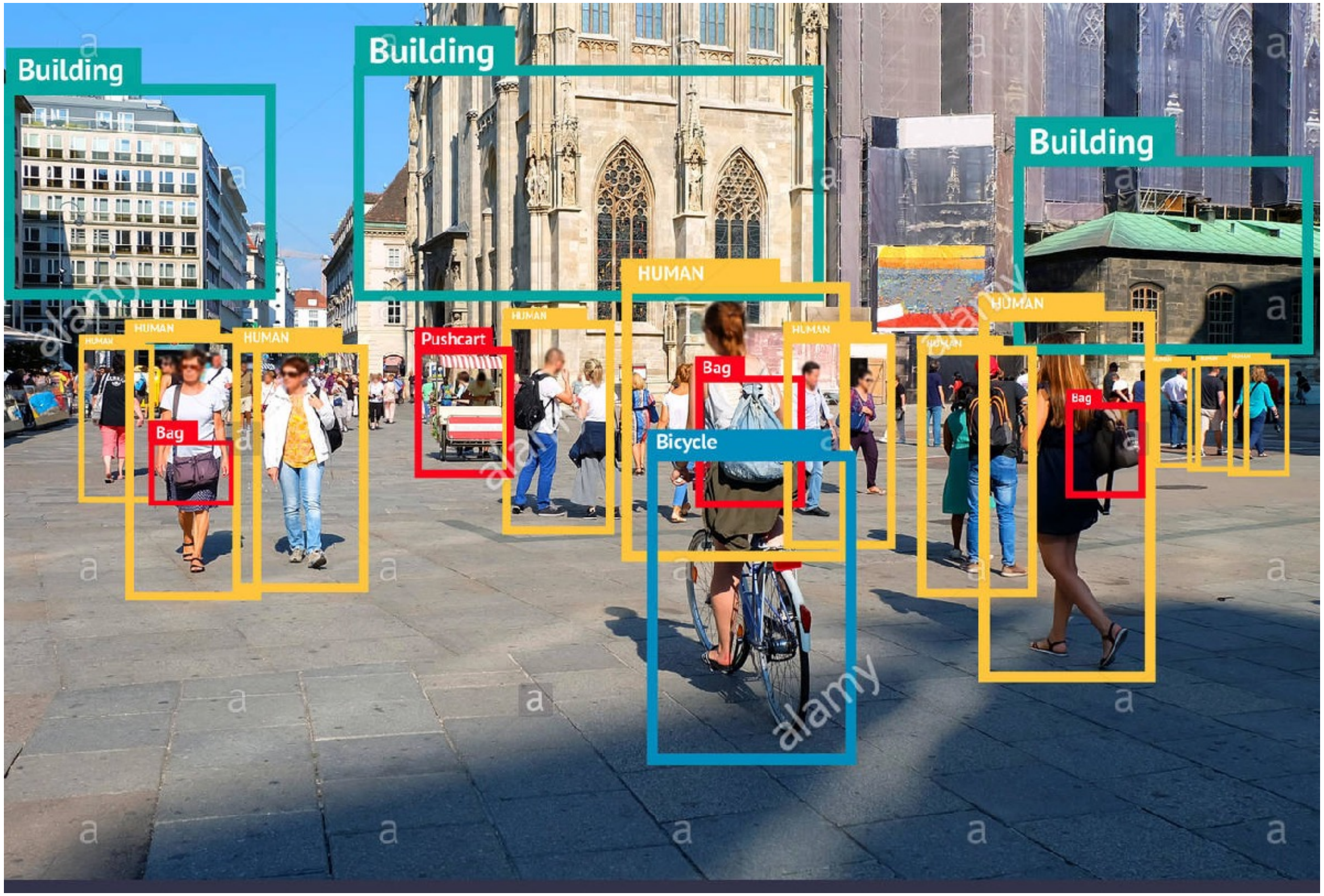
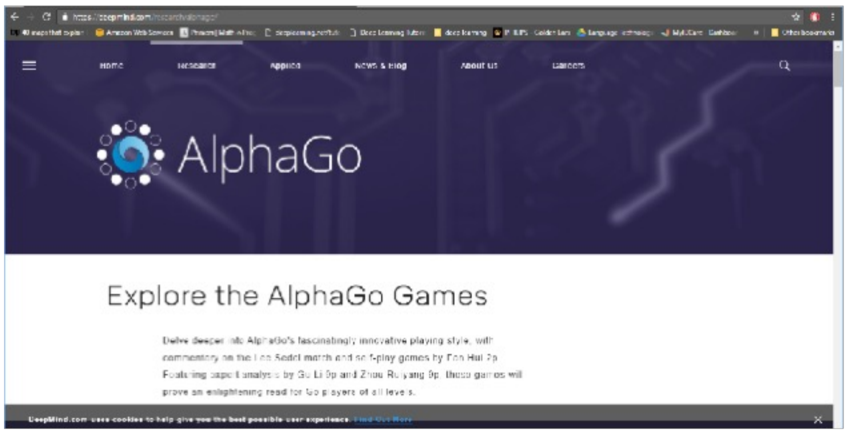


Image segmentation and object recognition



Achieving Master Level in GO



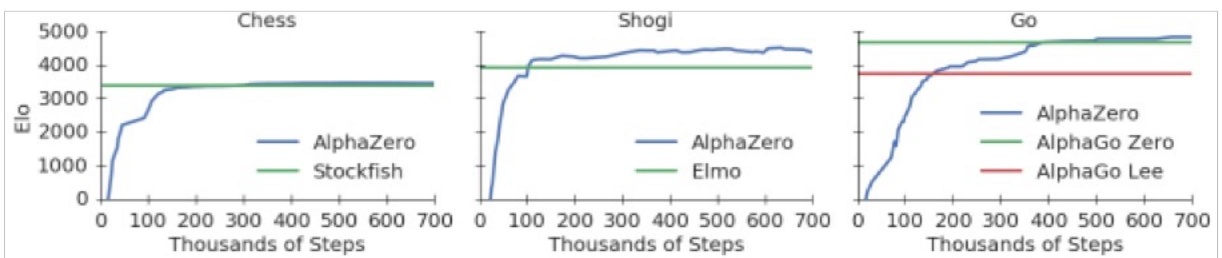
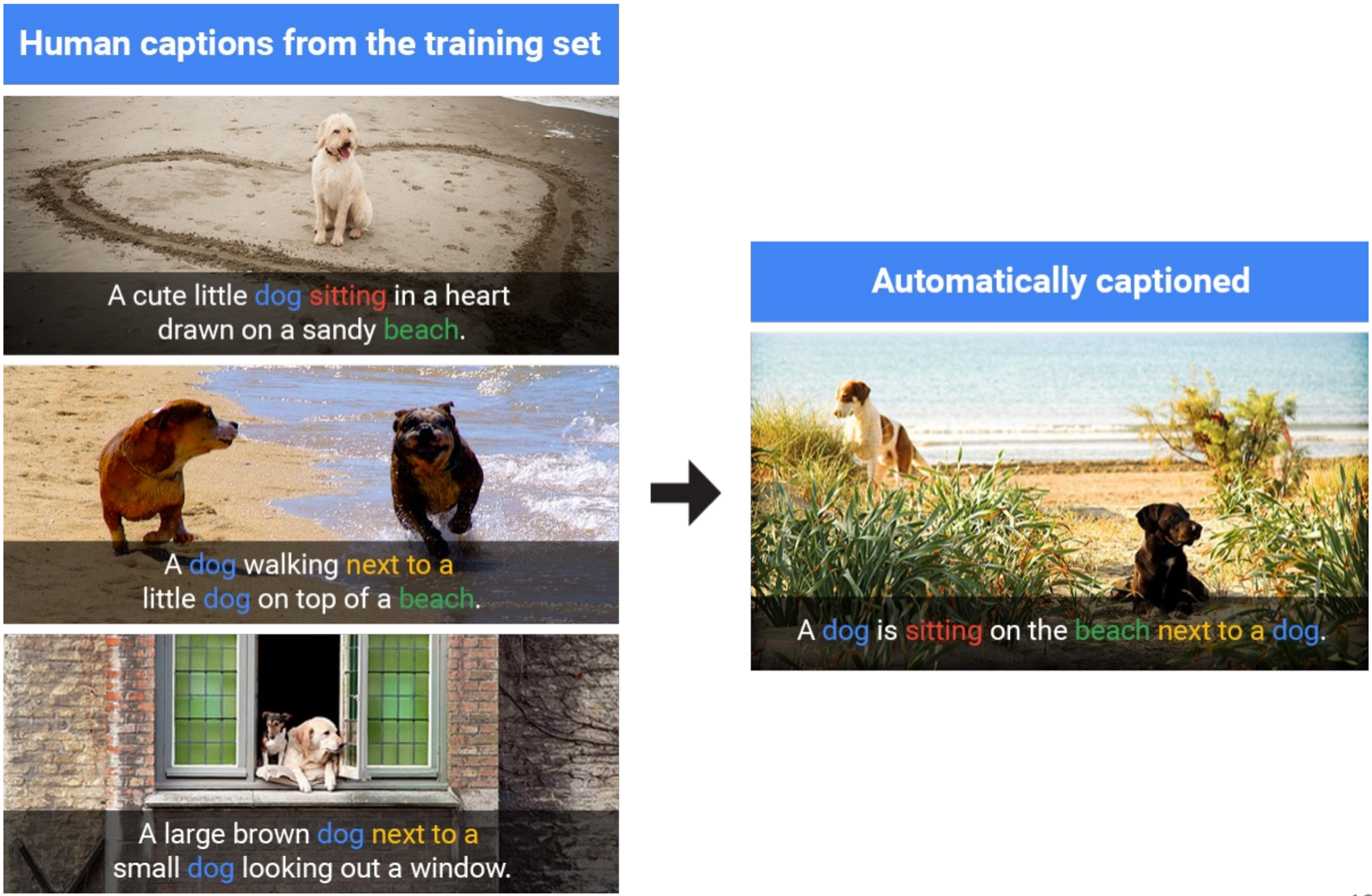


Figure 1: Training *AlphaZero* for 700,000 steps. Elo ratings were computed from evaluation games between different players when given one second per move. **a** Performance of *AlphaZero* in chess, compared to 2016 TCEC world-champion program *Stockfish*. **b** Performance of *AlphaZero* in shogi, compared to 2017 CSA world-champion program *Elmo*. **c** Performance of *AlphaZero* in Go, compared to *AlphaGo Lee* and *AlphaGo Zero* (20 block / 3 day) (29).

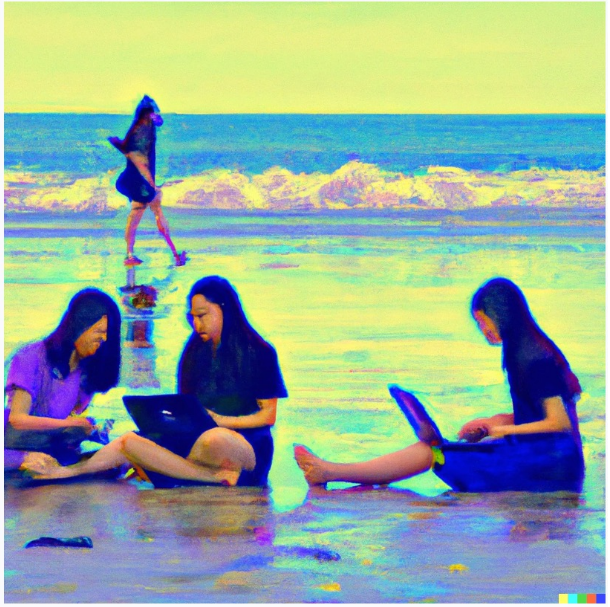
Image Captioning



Art generation

Tell DALL E 2:

"Students eager to learn machine learning on a Californian beach in Monet style"



ML Is Transforming The Industries

- ML has transformed the IT industry
  - Search Engine
  - Speech Recognition
  - Machine Translation
  - Recommendation

- ML is transforming other industries
  - Transportation
  - Healthcare
  - Finance
  - Insurance, Law, HR, Travel, Media, …
  - Semiconductor / Microprocessors

ML is transforming the academia

- SOTA methods in computer vision / speech recognition / natural language processing

- Recent candidates who interviewed with UCSB for faculty positions
  - Top security person: doing security + ML
  - Top database person: doing ML for databases
  - Top software engineering person: doing software engineering for ML systems debugging.
  - Of course, we also interviewed quite a few core ML/ AI candidates too

Why should you learn ML?

- Career opportunities
  - AI / ML jobs are highly paid
  - AI / ML are becoming the standard tools all software engineers are expected to know.

- Research opportunities and potential impact
  - “The golden age of physics is Newton’s time and Einstein’s time, the golden age of AI is right now!”

- Personal development
  - Consolidate your knowledge, connect the dots
  - Becoming better in solving problems

- “It is just my passion! "

How can this course help you?

- At the end of the course you will be able to
  1. Understand jargons of ML
  2. Assess whether ML will be a good solution to your problem; if so, choose appropriate ML models for it
  3. Correctly implement and debug each step of ML workflow
  4. Develop theoretical understanding on how / why / when ML works.
  5. Read research papers in ML (possibly ready to write your own)

Topics we will cover

#	Date	Topic	
1	9/22	Introduction, Supervised Learning, Linear regression	
2	9/27	Classification, Linear Classifier, Decision Tree	Machine Learning basics
3	9/29	Unsupervised Learning, dimensionality reduction	
4	10/4	Optimization basic: Gradient Descent and SGD	
5	10/6	Feedforward Neural Networks	
6	10/11	Convolutional Neural Networks	Modern Deep Learning models
7	10/13	Sequence Modeling and Recurrent Neural Networks	
8	10/18	Attention Mechanism and Transformers	
9	10/20	Graphical Models and MLE	
10	10/25	Gaussian Mixture Models, EM,LDS	
11	10/27	Undirected Graphical Models, Conditional Random Fields	Probabilistic ML models
12	11/1	Deep Latent Model and Approximate Inference	
13	11/3	Sampling Methods	
14	11/8	Convex Optimization	
15	11/10	Support Vector Machine, Kernel Methods	Optimization-based methods
16	11/15	Online Learning	
17	11/17	Statistical Learning Theory I	
18	11/22	Statistical Learning Theory II	ML Foundations
19	11/29	Theory of Deep Learning and Overparameterization	
20	12/1	Reinforcement Learning	
		Final project poster presentation	

Course information

- Websites
  - https://sites.cs.ucsb.edu/~lilei/course/ml22fa/index.html
  - Schedule, lecture notes, assignments, related links
- Discussion on Ed:
  - https://edstem.org/us/courses/22801/discussion/
- Reference books
  - [PRML] Pattern Recognition and Machine Learning, Bishop. available online
  - [FML] Foundations of Machine Learning. Mohri, Rostamizadeh, and Talwalkar. available online.
  - [D2L] Dive into Deep Learning. Zhang, Lipton, Li, Smola. available online.
  - [MML] Mathematics for Machine Learning, Deisenroth, Faisal, and Ong. Free online.

Workload and grades

- 50% Homework assignments
  - HW0: 5%
  - HW1-3: 15% each
  - A mix of theory questions and coding questions
  - Each student needs to write their own homework (discussion / collaboration allowed if declared)
- 40% Project
  - Proposal 5%
  - Midterm milestone 5%
  - Presentation 10% + Final report 20%
- 10% In-class or short after-class quiz
  - You only need to complete them to get the credits

Project timeline

- Group size: 1 – 4
- Milestones:
  - Proposal due 10/6: one pager on what you will do
  - Midterm milestone 11/3: one pager update
  - Poster presentation: early December
  - Final report: Shortly after the presentation

- Basic project:
  - Read a paper, replicate results
- Advanced project:
  - Apply ML to your problem / data from your research
  - Develop new ML methods / theory

There will be some steep learning curves for some

- You need to use math:
  - Calculus, Linear Algebra, Probabilities.
  - Follow mathematical proofs.

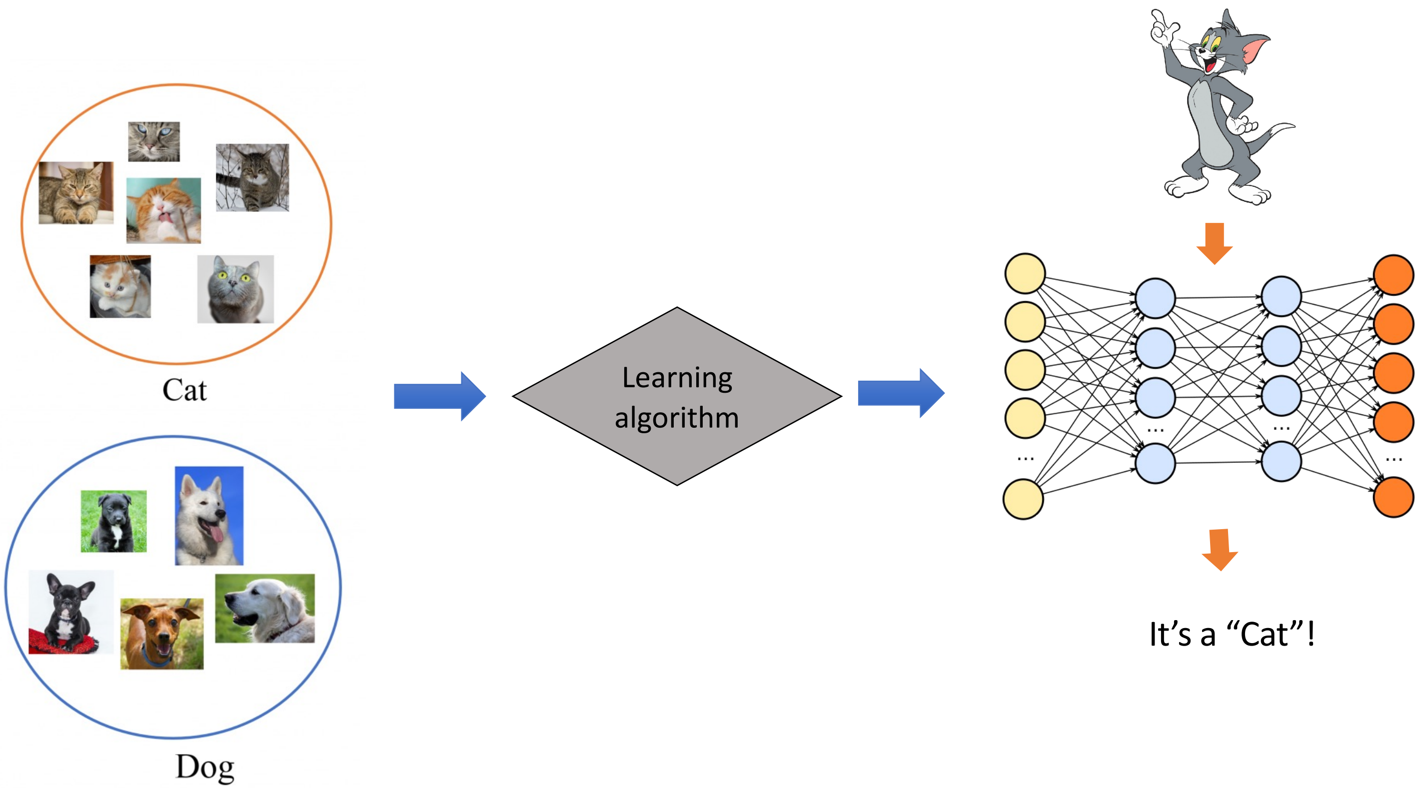
- You need to be able to code:
  - Mainly in Python (numpy, scipy)
  - Learn to use pyTorch for coding assignments

- If you are/were a UCSB undergraduate:
  - This course overlaps with CS 165B but materials are covered with more breath and depth.
  - Having completed the following courses may help you: Linear algebra (MATH 3B), Vector Calculus (6A), Probability and Statistics (PSTAT 120A, 120B), data-structures and algorithms (CS 130A & 130B)

Remaining of today's lecture

- Machine learning overview
- Supervised learning by an example

Machine learning studies "*computer programs that automatically improve (its* ***performance*** *on a task)with* ***experience****.*"



Different tasks / problems in Machine Learning

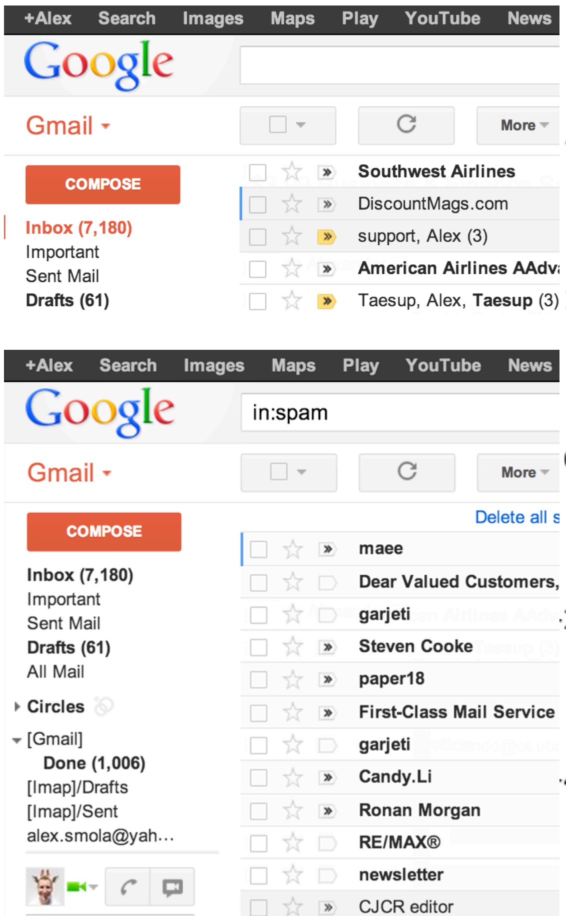
- Supervised Learning — Spam Filter.
- Unsupervised Learning — Topics of a text corpus
- Reinforcement Learning — Atari Games. Serve Ads.
- Structured Prediction — Machine translation.

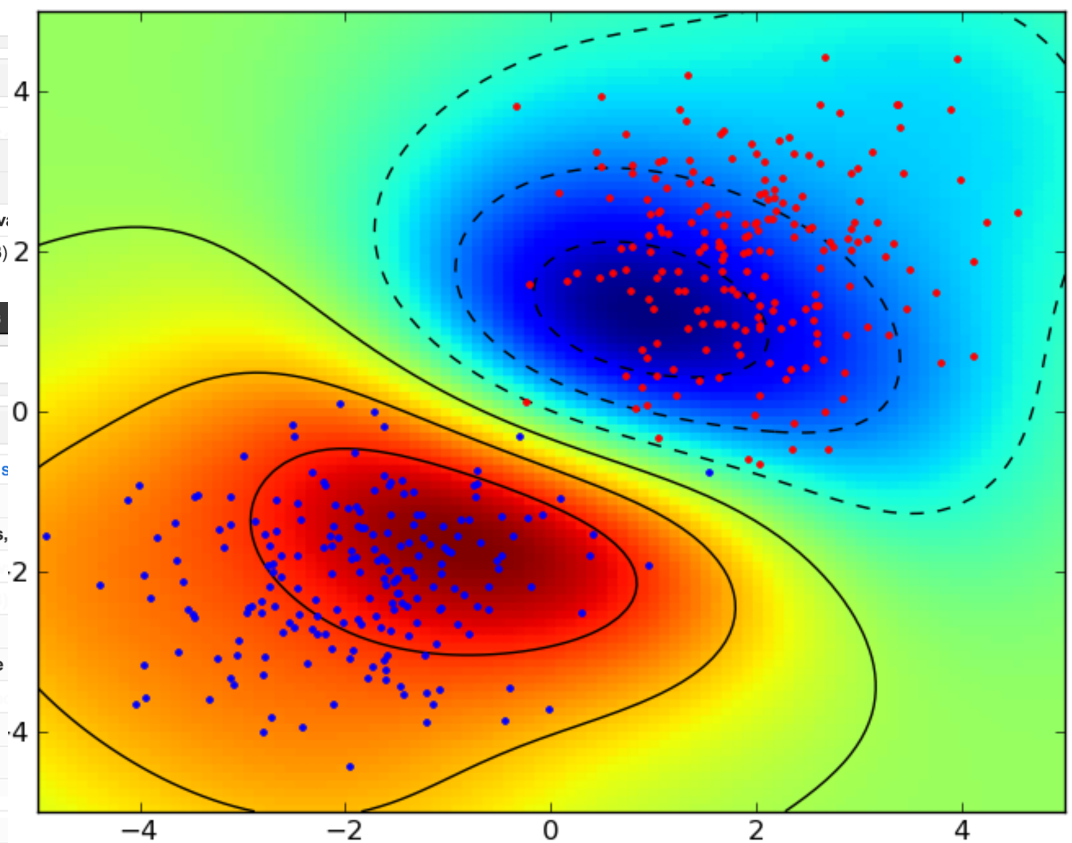
Semi-supervised learning, active learning,

ranking /search / recommendation

self-supervised learning and many more!

Supervised learning is about predicting label y using feature x by learning from labeled examples.



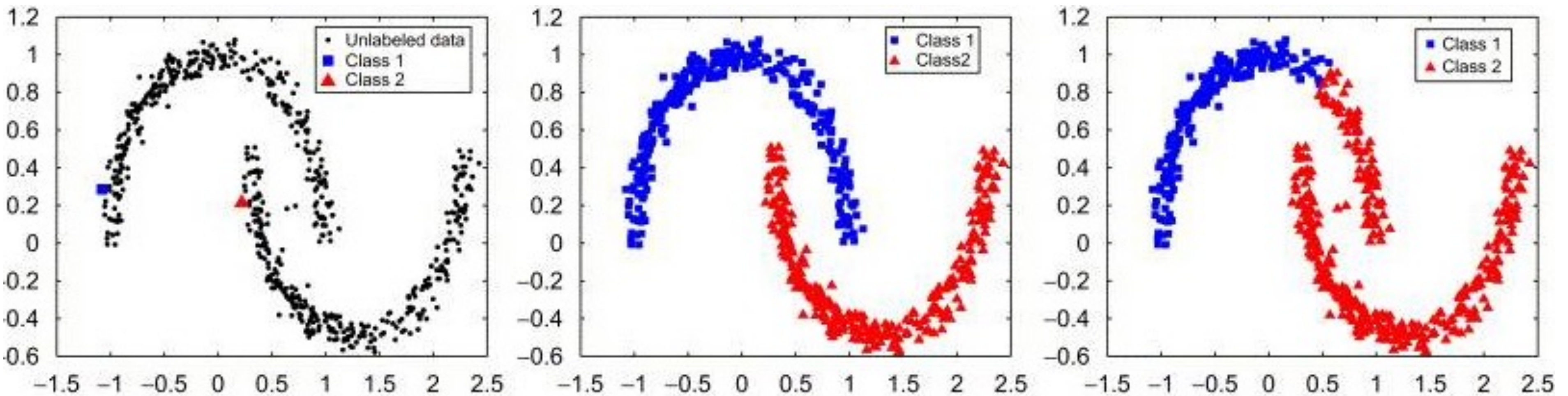


Unsupervised Learning is about finding structures in an unlabeled dataset

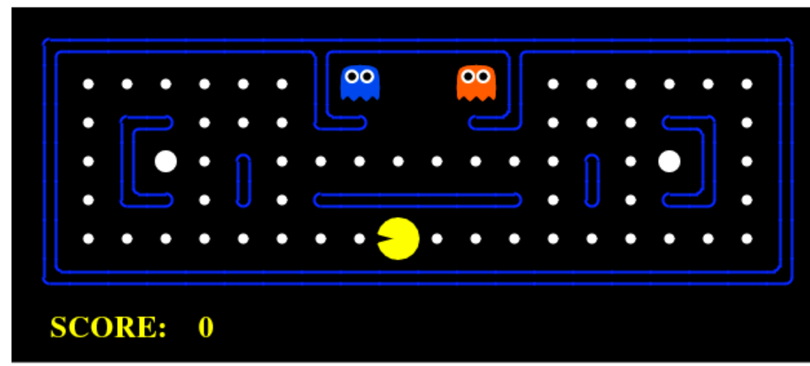
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Semi-supervised Learning using both labeled and unlabeled data.



Reinforcement learning learns to make decisions for long-term rewards by trials-and-errors.

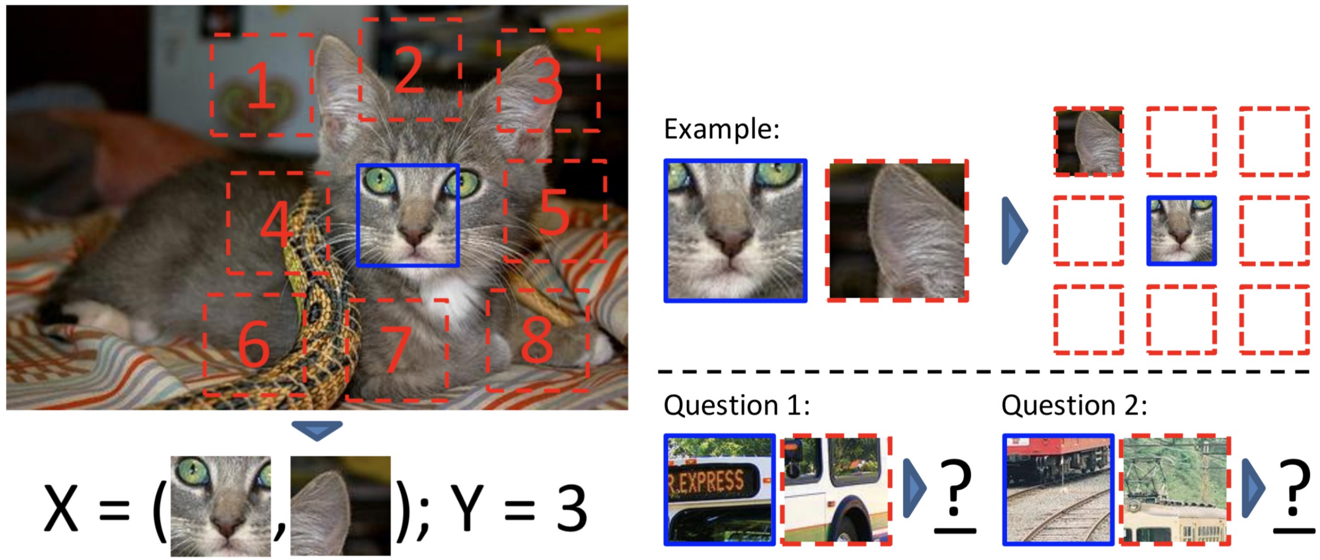


amazon

Recommendations

buy or not buy

Self-supervised learning learns to predict parts of x using other parts of x.



Randomly masked — A quick [MASK] fox jumps over the [MASK] dog

Predict — A quick brown fox jumps over the lazy dog

Image example from [(Doersch et al, 2015)](), text example from [Amit Chaudhary]()

Task, Experience, Performance

	Task	Experience	Performance
Supervised learning	Use X predict Y	(X,Y) – feature-label pairs	error
Unsupervised learning	Organize, index, discover structures in data	Collection of unlabeled data X	"How well it represents actual data distribution"
Semi-supervised learning	Use X predict Y	Unlabeled data X, and labeled data (X,Y)	error
Reinforcement learning	Make savvy decisions	Feedback from Task environment	Expected Long-term reward
Self-supervised learning	Use one part of X to predict another part of X	Collection of unlabeled data X	Reconstruction error

The focus of today’s lecture is “Supervised Learning”

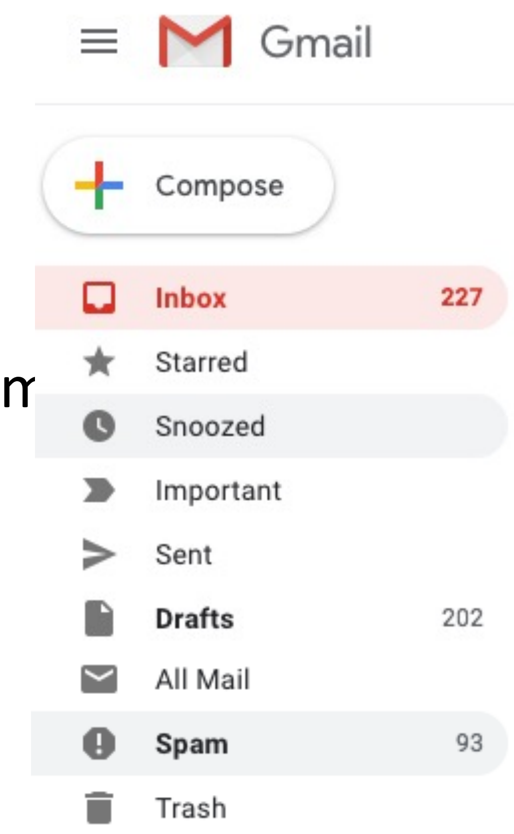
- Actually, just “binary classification”.

- Prototypical Example: Spam filtering

ail

r “Ham

Illustration extracted from [here]



Example of SPAM emails

Mail thinks this message is Junk Mail. Move to Inbox

MICROWORLD CORPORATIO... December 20, 2019 at 2:38 AM MC

CLAIMS.

To: undisclosed-recipients:;,

Reply-To: microworld219@gmail.com

MICROWORLD CORPORATIONS:
CUSTOMER SERVICE:
FRIEDRICHSTRAЯE 10,BERLIN ALEMANHA
REFERENCE NUMBER: MBB-009-D54-DE
BATCH NUMBER: MGC-2019- SM-009
TICKET NUMBERS: 2,6,13,21,26,32

OFFICIAL WINNING NOTIFICATION.
We are pleased to inform you of the released results of Microworld Promotion...
This is a promotional program organized by Microworld Corporations, in conjunction with the Foundation for the promotion of software products, and use of email addresses. Held on Thursday 19th, December 2019, in Berlin, Alemanha.
Your email address won a cash award of Four hundred and eighty eight thousand two hundred and fifty euros (488,250.00 Euros)..
Contact Our Foreign Transfer Manager for claims with your winning details and your contact information.
Mrs. Helena Bosch.
Email: micropromo19@yahoo.com
Congratulations!!
Sincerely,
Rosa Van Beek.

Mail thinks this message is Junk Mail. Move to Inbox

Email ADMIN January 1, 2020 at 10:35 PM EA

(cs.ucsb.edu) APPLICATION -Storage Full Notes- Last -... Details

To: Yu-Xiang Wang,

Reply-To: Email ADMIN

Dear yuxiangw@cs.ucsb.edu,

Your email has used up the storage limit of 99.9 gigabytes as defined by your Administrator. You will be blocked from sending and receiving messages if not re-validated within 48hrs.
Kindly click on your email below for quick re-validation and additional storage will be updated automatically

yuxiangw@cs.ucsb.edu

Regards,
E-mail Support 2020.

Example of another SPAM email

Mail thinks this message is Junk Mail.

Move to Inbox

MARK ZUCKERBERG

Junk - Google August 24, 2018 at 10:48 AM

MZ

WINNING AMOUNT

Reply-To: MARK ZUCKERBERG

WINNING AMOUNT

My name is Mark Zuckerberg,A philanthropist the founder and CEO of the social-networking website Facebook,as well as one of the world's youngest billionaires and Chairman of the Mark Zuckerberg Charitable Foundation, One of the largest private foundations in the world. I believe strongly in'giving while living' I had one idea that never changed in my mind - that you should use your wealth to help people and i have decided to secretly give {$1,500,000.00} to randomly selected individuals worldwide. On receipt of this email, you should count yourself as the lucky individual. Your email address was chosen online while searching at random.Kindly get back to me at your earliest convenience,so I know your email address is valid.(mzuckerberg2444@gmail.com) Email me Visit the web page to know more about me: https://en.wikipedia.org/wiki/ Mark_Zuckerberg/ or you can google me (Mark Zuckerberg)

Regards,
MARK ZUCKERBERG

Example of a HAM (non-spam) email

Dear Professor Foo,

I am a student in your machine learning class.

I have a question about the second term project and I was not able to find the answer on the syllabus. Should our project be only about the topics listed on the second part of the syllabus, or can I incorporate topics from the whole course, as long as it fits with the subject of the class?

I look forward to hearing from you.
Best regards,
Bar

Quoted from [Here].

Modeling-Inference-Learning

Modeling

- Feature engineering
- Specify a family of classifiers

Inference

Apply the classifier to emails

Learning

Learning the best performing classifier

What are the features that we can use to describe an email (3 min discussion)

- What are the information that we can extract from text, and hyper-texts to describe an email?
- What are typical characteristics of a spam email?
- What are typical characteristics of a non-spam email?

Possible features

- Number of special characters: $, %
- Mentioning of: Award, cash, free
- Greetings: generic, or specific
- Bad grammars and misspelled words: e.g. m0ney, c1ick here.
- Excessive excitement: Many “!”, “!!!”, “?!”, words in CAPITAL LETTERS.

- Whether the senders on the contact list
- Length of an email
- Whether the receiver has responded to sender before

Example of a feature vector of dimension 4

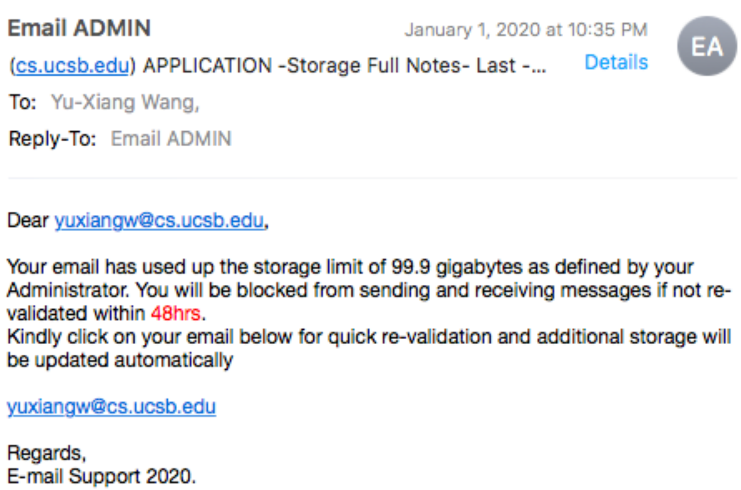
Contains hyperlinks

Proportion of misspelled words

1	**0**	**0.0375**	**80**

Whether the contact list

Length of the message


Email ADMIN
January 1, 2020 at 10:35 PM
EA
(cs.ucsb.edu) APPLICATION -Storage Full Notes- Last -...
Details
To: Yu-Xiang Wang,
Reply-To: Email ADMIN

Dear yuxiangw@cs.ucsb.edu,

Your email has used up the storage limit of 99.9 gigabytes as defined by your Administrator. You will be blocked from sending and receiving messages if not re-validated within 48hrs.
Kindly click on your email below for quick re-validation and additional storage will be updated automatically

yuxiangw@cs.ucsb.edu

Regards,
E-mail Support 2020.

Step 1 in Modelling
Feature extractor:
Converting the object of interest to a vector of numerical values.

Mathematically defining a classifier

- Feature space: $\mathcal{X} = \mathbb{R}^d$
- Label space: $\mathcal{Y} = \{0, 1\} = \{\text{non-spam}, \text{spam}\}$
- A classifier (hypothesis): $h : \mathcal{X} \to \mathcal{Y}$

How do we make use of this feature vector? What is a reasonable “classifier” based on this feature representation?

1	**0**	**0.0375**	**80**

Contains hyperlinks

Whether the contact list

Proportion of misspelled words

Length of the message

- Feature space: $\{0, 1\} \times \{0, 1\} \times \mathbb{R} \times \mathbb{N}$
- Label space: $\mathcal{Y} = \{0, 1\} = \{\text{non-spam}, \text{spam}\}$

- **How are we going to use these features as a human?**
  - (3 min discussion)

Specifying a family of classifiers --- a “hypothesis class”

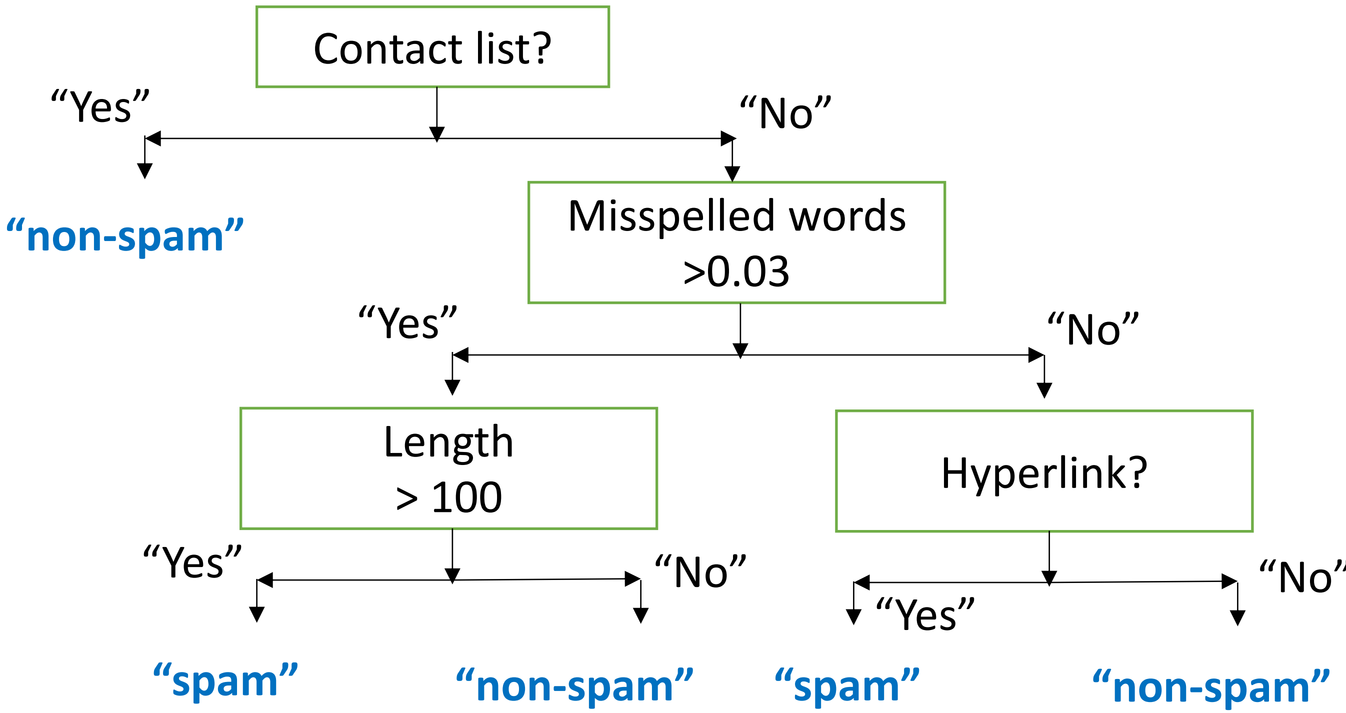
- Hypothesis class
  - A family of classifiers: $\mathcal{H}$
  - Also known as “concept classes”, “models”, “decision rule book”
  - “Neural networks” and “Support Vector Machines” are hypothesis classes.
  - Typically we want this family to be large and flexible.

- The task of machine learning:
  - A **selection problem** to find a

$$h \in \mathcal{H}$$

that “**works well**” on this problem using the **observed data**.

Decision trees



- **Question:** What are the "free parameters" if we are to learn such a decision tree? Using data?

Learning a decision tree

- Free parameters:
  - Which feature(s) to use when branching branch?
  - How to branch? Thresholding? Free threshold?
  - Which label to assign at leaf nodes?
- Hyperparameters:
  - Max height of a decision tree?
  - Number of parameters the tree can use in each
- **Question**: Consider a problem with **4 binary features**.
  - How many decision trees of **3 layers** are there? If each decision uses only one feature? (you may repeat features)
  - How many possible feature vectors are there?
  - How many classifiers are there (without restrictions)?

Example: Linear classifiers

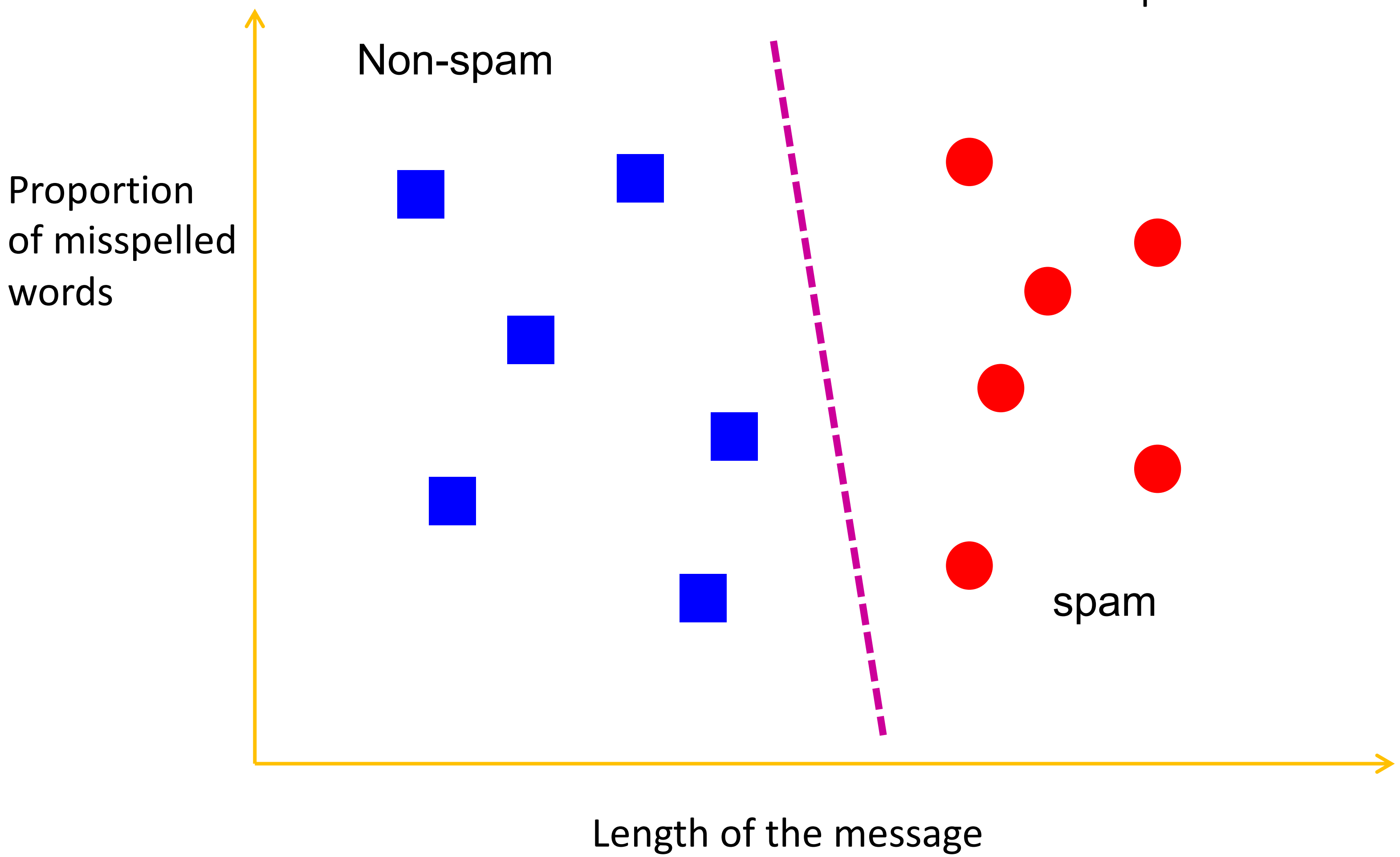
- Score(x) = $w_0 + w_1$ * **1**(hyperlinks) + w_2 * **1**(contact list) + w_3 * misspelling + w_4 * length
- A linear classifier: h(x) = 1 if Score(x) > 0 and 0 otherwise.
- Question: What are the "free-parameters" in a linear classifier?
  - If we redefine $\mathcal{Y} = \{-1, 1\}$
  - A compact representation:

$$h(x) = \text{sign}(w^T [1; x])$$

Geometric view: Linear classifier are “half-spaces” !

$\{ x \mid w_0 + w_1 * x1 + w_2 * x2 + w_3 * x3 + w_4 * x4 > 0\}$

The set of all ”emails” that will be classified as “Spams”.



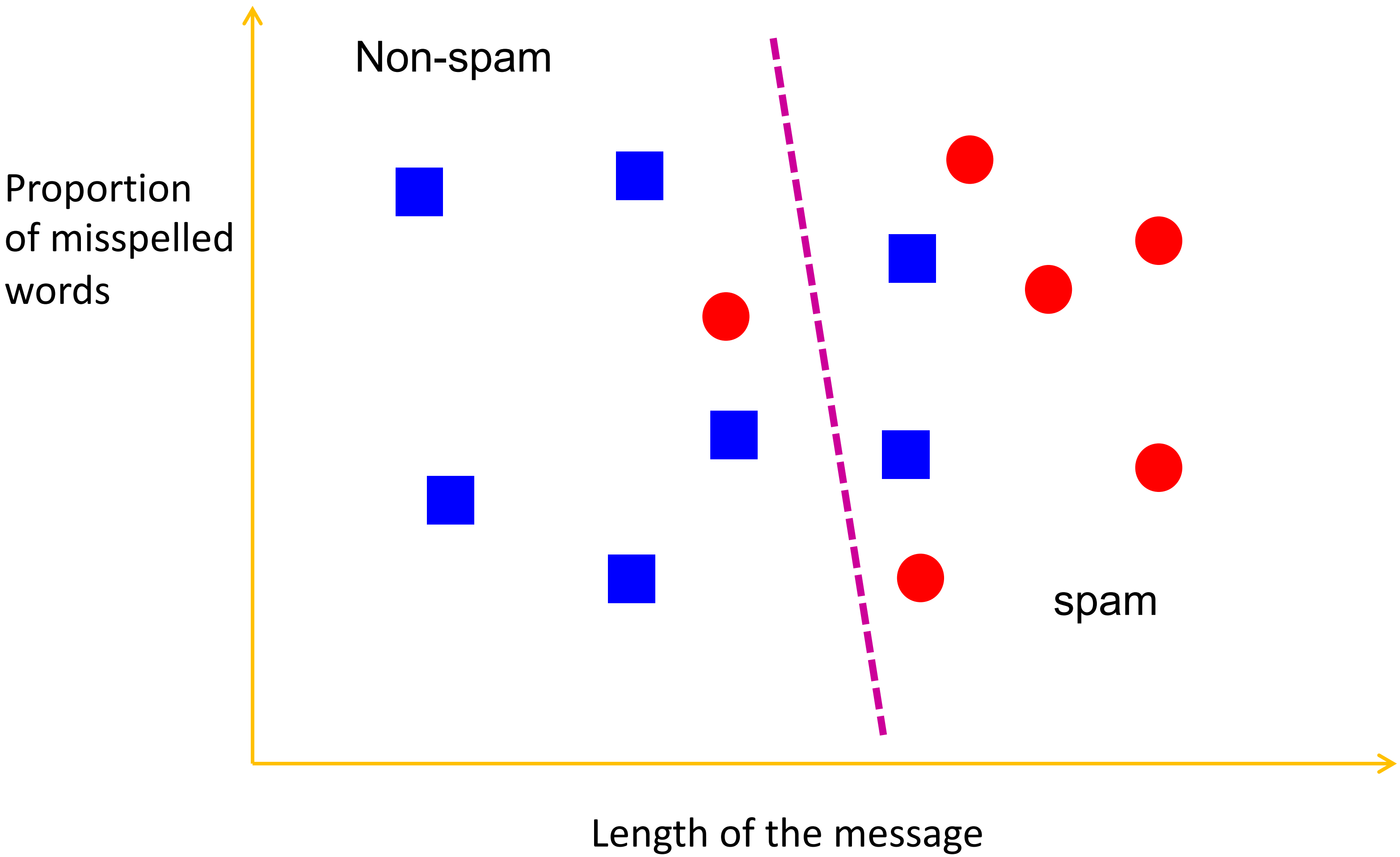
Learning linear classifiers

- Training data:

$$(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$$

- In the above example, there is a clean cut boundary that distinguishes “spams” from “non-spams”.
    - “Linearly separable” problem
    - Learning linear classifier: Finding vector w that is consistent with the observed training data.

Example: Linearly non-separable cases



How do we learn a linear classifier in a non-linearly separable case?

- Training data:

$$(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$$

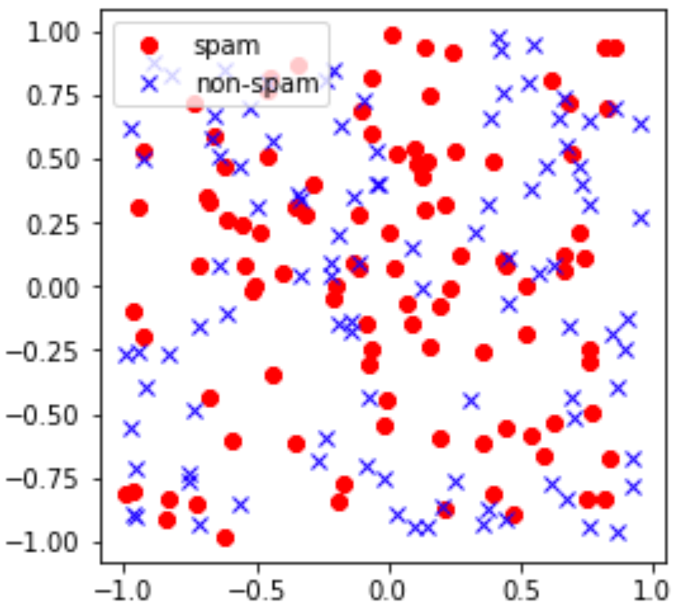
- Solving the following optimization problem:

$$\min_{w \in \mathbb{R}^d} \mathrm{Error}(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(h_w(x_i) \neq y_i)$$

- Learning: Find the linear classifier that makes **the smallest number of mistakes** on the training data.

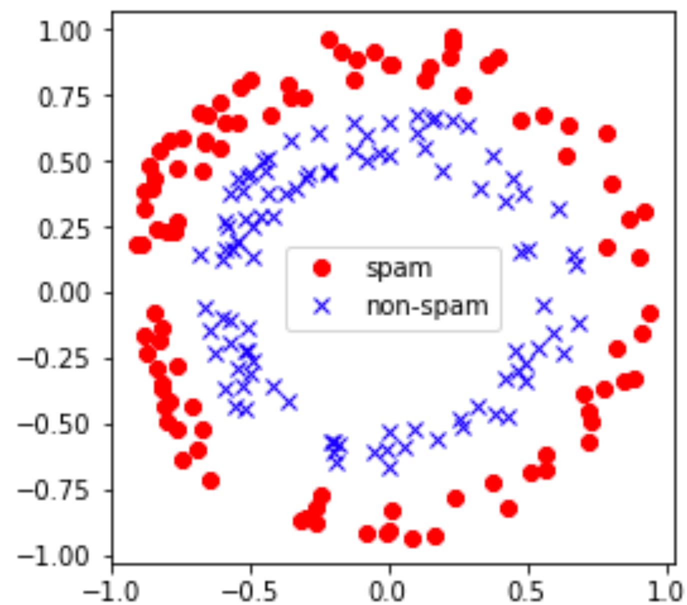
What happens if the linear classifier with the smallest number of mistakes still makes a mistake 49% of the time?

Case 1:



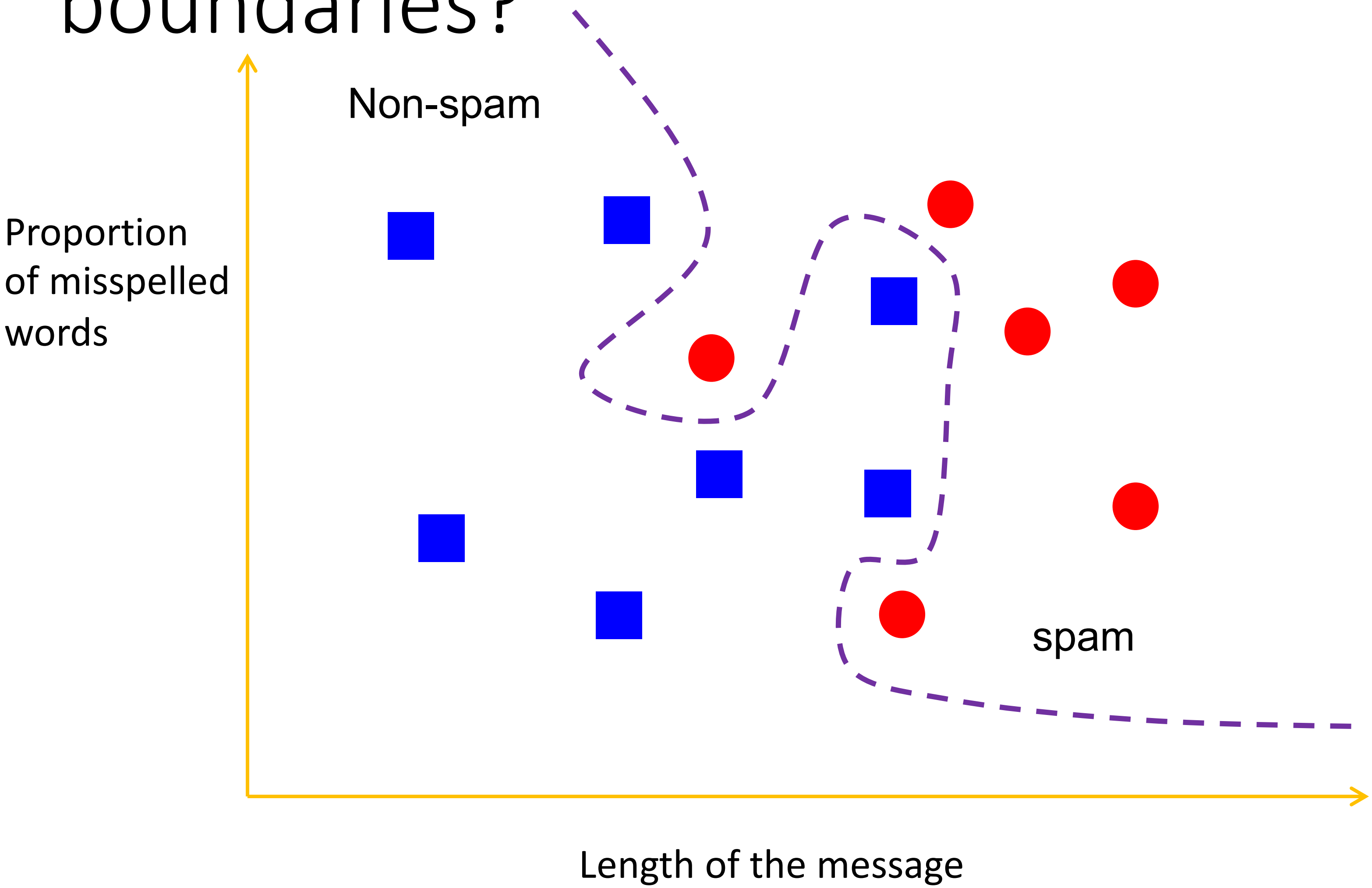
There is no information about the label in the features.
No classifiers are able to do well.

Case 2:

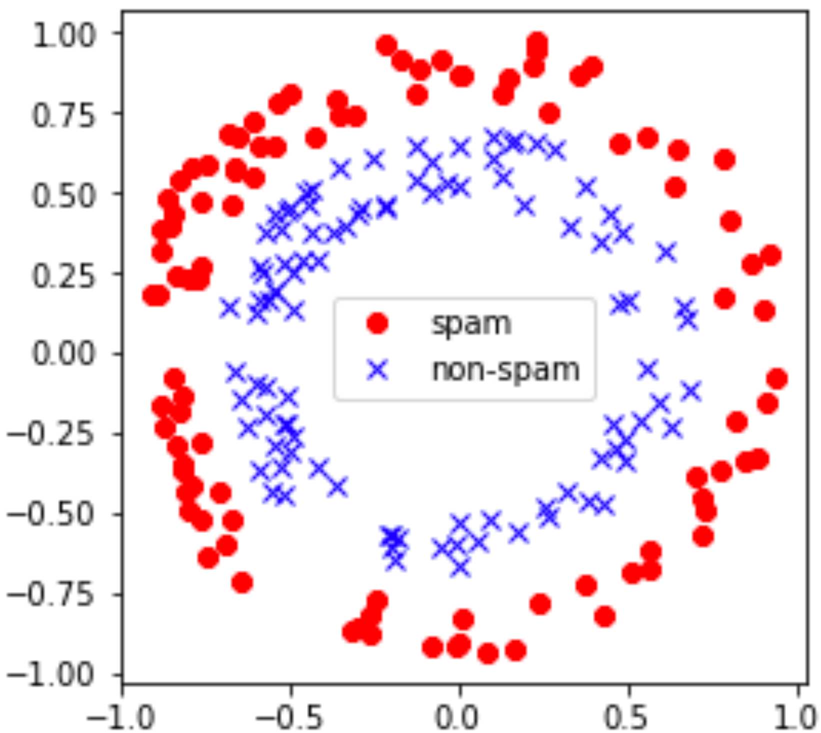


There are some nonlinear classifier that works. But no linear classifiers will do better than chance.

Going to higher dimensions? Maybe we can also allow non-linear decision boundaries?



Example: Feature transformation.



What we can do:

$$(\tilde{x_1}, \tilde{x_2}) = \left(\sqrt{x_1^2 + x_2^2}, \arctan(x_2/x_1)\right)$$

In the redefined space, the two classes are now linearly separable.

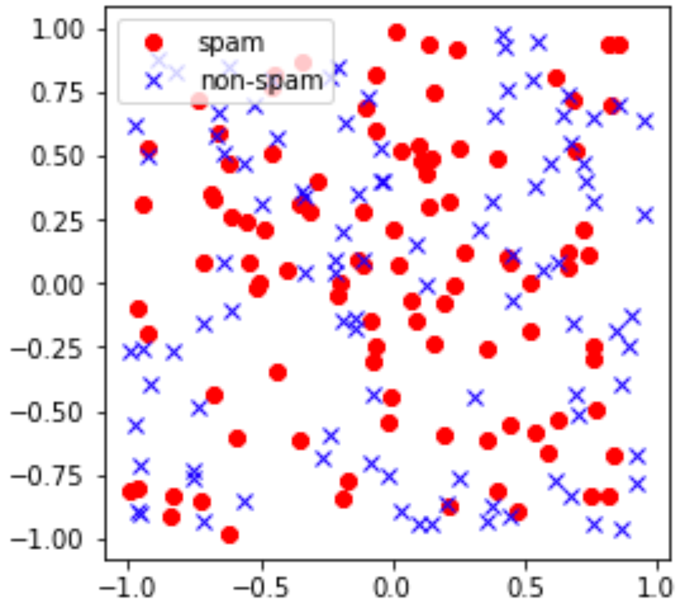
Nonparametric classifiers

- Increasing the complexity of the classifier as we get more data
- For example:
  - We can use the entire training dataset as “free parameters” of the classifier.
  - k-Nearest Neighbor
  - Kernel methods (lifting to infinite dimensional space)
  - Neural networks (design a model for a fixed data size)

Question: What is the classification error of 1-NN classifiers?

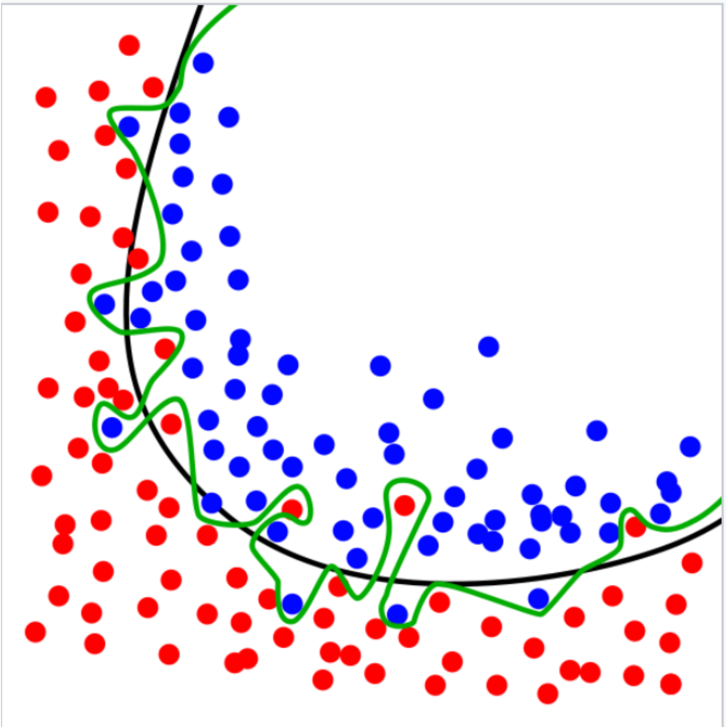
We can make the classifiers arbitrarily accurate… with 1-NN classifier; or with bigger and bigger neural networks.

- Even if the data look like:



- **What went wrong?**

The problem of Overfitting



The green line represents an overfitted model.
While the green line best follows the training data,
it is too dependent on that data and
it is likely to have a higher error rate on new unseen data.

The goal of machine learning is not to obtain 0-training error, but rather to achieve small error rates on **new data points** (that are **not** used for training.)

- Test Error < Training Error + Generalization Error

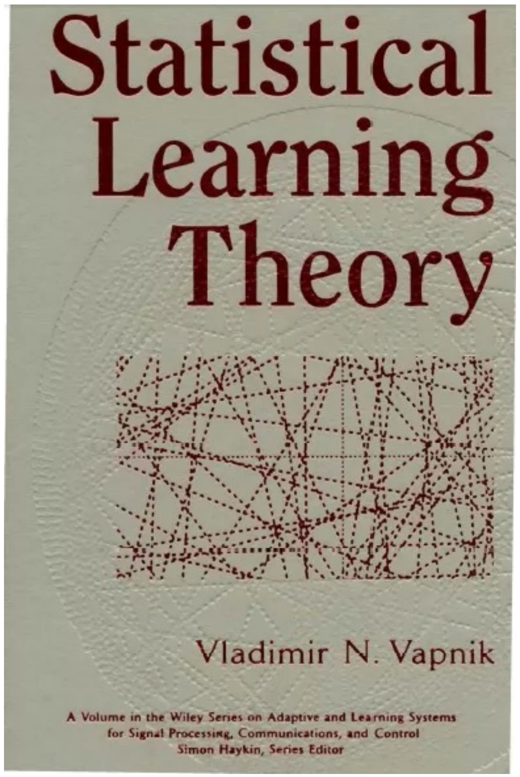
$$\mathrm{Err}(h) := \mathbb{E}[\mathbf{1}(h(x) \neq y)] \qquad \widehat{\mathrm{Err}}(h) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(h(x_i) \neq y_i)$$

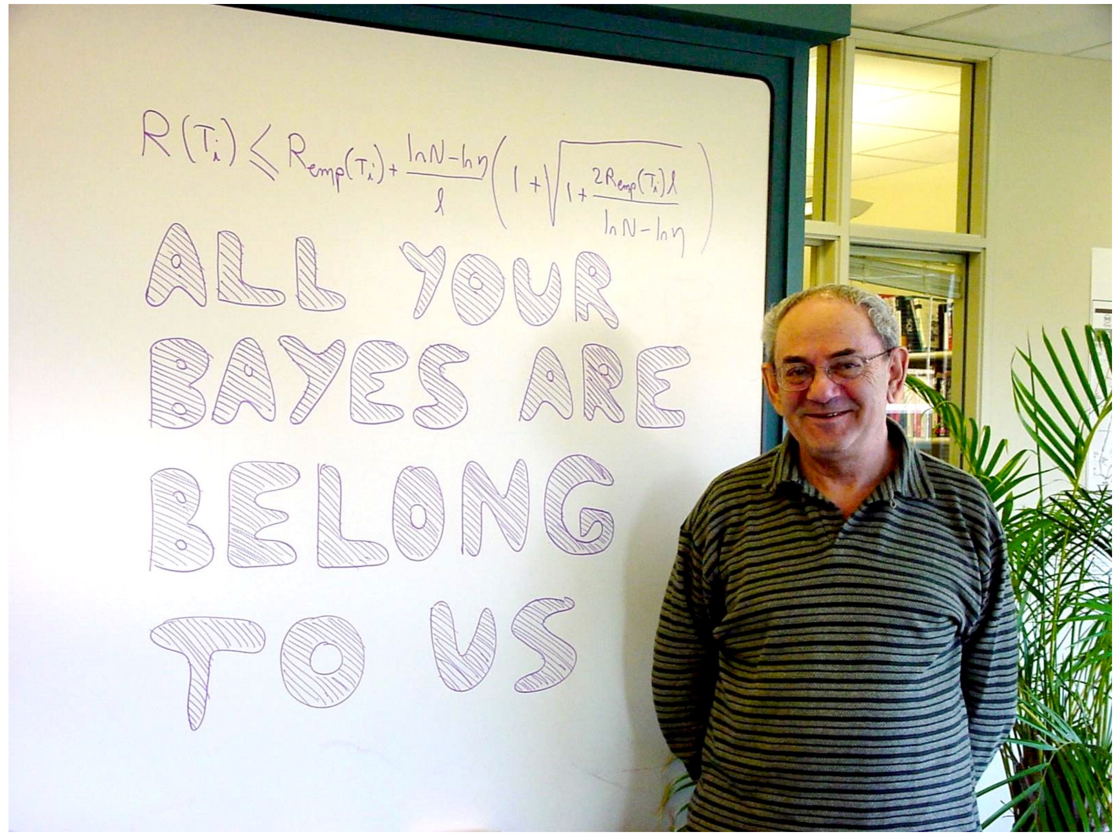
$$\mathrm{Gen}(\mathcal{H}) := \sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(h(x_i) \neq y_i) - \mathbb{E}[\mathbf{1}(h(x) \neq y)] \right|$$

(** some text uses “generalization error” as a synonym as “test error”, which has created much confusion. The above is the definition we adopt.)

Statistical Learning Theory

TL;DR: Proving that the generalization error → 0, thus showing that ML works.





Closely related to Empirical Process Theory, Computational Learning Theory.

Summary of today's lecture

- Machine learning overview

- Supervised learning: Spam filtering as an example
  - Features, feature extraction
  - Models, hypothesis class
  - Choosing an appropriate hypothesis class
  - Performance metric
  - Overfitting and generalization

Next lecture: more on supervised learning

- Supervised learning: problem setup

- Loss function, Risk, Empirical Risk
  - Risk decomposition
  - Analysis of linear regression under linear gaussian model
  - Analysis of supervised learning with bounded loss

- Model-selection: Holdout, Cross-validation