Introduction to Statistical Machine Learning CSC/DSCC 265/465

Lecture 1: Introduction to Course

Cantay Caliskan

Today

- What is machine learning?
- Course logistics

What is machine learning?

Areas of Specialization

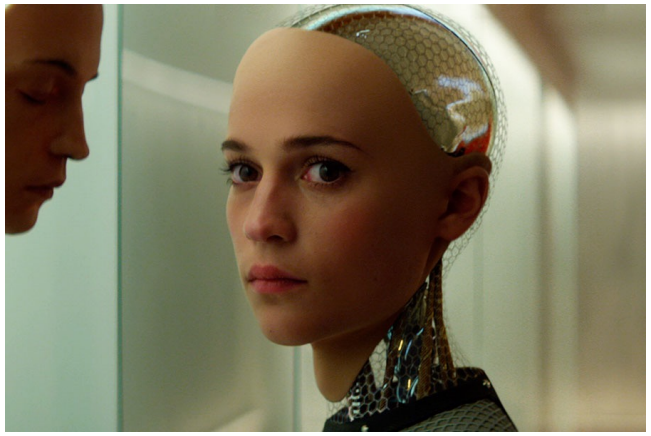
1) **Data**
   a) ‘Big’ data vs. ‘small’ data
   b) Distribution and shape
   c) Dealing with missing data
   d) Finding new sources of data
2) **Prediction**
   a) Continuous
      i) Linear regression
      ii) Time series forecasting
   b) Non-continuous
      i) Logistic regression
3) **Finding the Unknown**
   a) Clustering /Unsupervised learning
   b) Dimension reduction
4) **Learning from the Known**
   a) Classification/Supervised Learning
   b) Artificial Intelligence
5) **Algorithm Development**
   a) Improvements
   b) New algorithms: deep learning
6) **Visualization**
   a) Improvement on traditional software
   b) Visualizing the imaginable
   c) Visualizing the non-imaginable
7) **Tool and Software Development**
   a) Writing packages/libraries
   b) Developing new software

How is data used?

- **Tech**
  - Image and video recognition
  - Analyzing customer behavior
  - Speech recognition
- **Medical**
  - Protein structure prediction
  - Genomics
  - Symptom detection
  - Health studies/clustering
- **Finance**
  - Price prediction
  - Making investment decisions
  - Operations
- **Education**
  - Improving school/student performance
  - Operations
- **Consulting**
  - Improving customer experience
    - Consumer services
    - Transportation
    - Aviation
    - Biomedicals
- **Law**
  - Criminology
  - Crime prevention
  - Litigation analysis

Why do we need machine learning?

- Help automate boring, hard tasks



- How do we ***automate***?
  - Not through programming
  - ***Instead***: Let the computer determine the parameters of a ‘program’

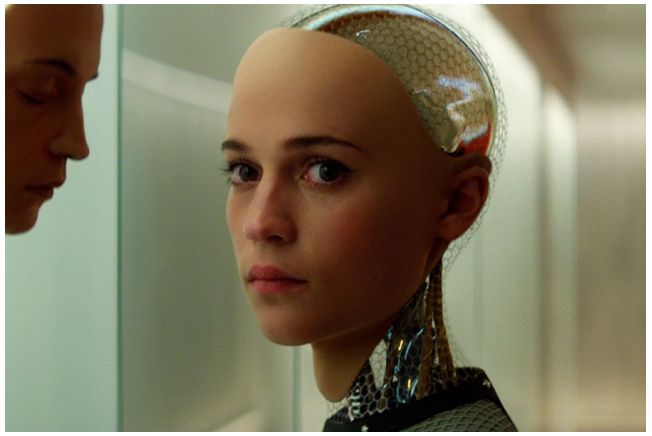
- Often takes advantage of ‘big data’!

What is machine learning?

- **Sub-field** of artificial intelligence

- Difference is the methods used
  - Artificial intelligence: Mimicking human cognitive functions either *fully* or *partially*
    - There can be *direct* or *indirect* instruction

  - Machine learning: Using of mathematical models to help a computer learn without *direct* instruction



Related to:

Pattern recognition

Data Mining

Big Data

Deep Learning

Types of Learning

1) Unsupervised Learning (=clustering)

- When the data does not include class labels, this is called ***unsupervised learning***

2) Supervised Learning (=classification / regression)

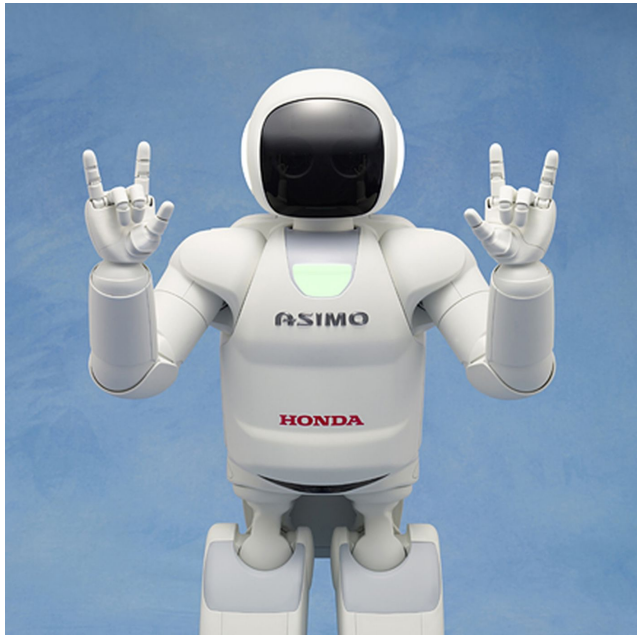
- Algorithms are given a ***training set***:
  - ***Training set***: Bunch of a data points and classification for those points
- New data is compared to the training set to make classifications

3) Reinforcement Learning (=classification / regression)

- Algorithms try to maximize the notion of reward
- A balance is sought between exploration and exploitation

Supervised Learning

- Provide a ***training set*** to your algorithm
  - Your set needs to have ***features*** and ***labels***
- Your training performance is tested with a ***test set***
- Outputs can be:
  - Categorical (***classification***)
  - Continuous (***regression***)



Examples: Supervised Learning

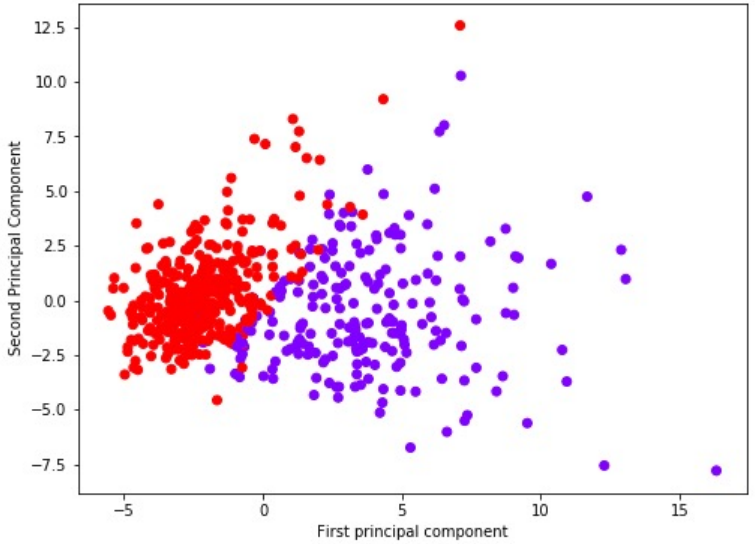
- **Recognize digits**
  - MNIST dataset!
  - What is input, what is output?

- **Predict the future prices of Tesla stock**
  - Yahoo Finance!
  - What is input, what is output?

Unsupervised Learning

- ~~**Recognize digits**~~
  - ~~MNIST dataset!~~
  - ~~What is input, what is output?~~
- Sorry, we don't have as much information ☹

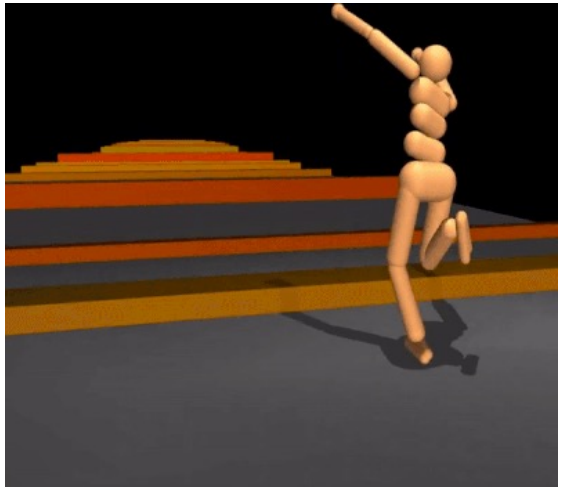
Can we still learn something?



Reinforcement Learning

- **Learn to recognize digits (?)**
  - Based on input, your algorithm is 'encouraged' to take action
    - If the action leads to prediction:
      - **Reward!**

- <u>Or a better example:</u>
  - Google's DeepMind teaches itself to **walk**



Question: Types of learning

1) Suppose you are building a machine to detect people that are about commit a crime by analyzing camera surveillance and phone conversations. Would you model it as a ***classification*** or a ***regression*** problem?

2) You are designing a machine that can learn to play a video game by playing lots of games with you and your friends. Would you use ***supervised***, ***unsupervised***, or ***reinforcement*** learning?

Course logistics

Contact Information

Instructor

- Cantay Caliskan (pronounced ‘Jantai’)
- Background in politics, computer science, statistics
- E-mail: cantay.caliskan@rochester.edu
- Office: Wegmans Hall, 1205
- Office Hours: (virtual *for now*) -> please sign up for an *appointment*
  Mondays 4:30 – 6:30 PM
  Wednesdays 6:00 – 7:00 PM

Contact Information

Sai Meka (*Head TA*)

Xinyu Cai

Ryan Hilton

Yiheng Mao

8 TAs!

Duy Pham

Amanda Pignataro

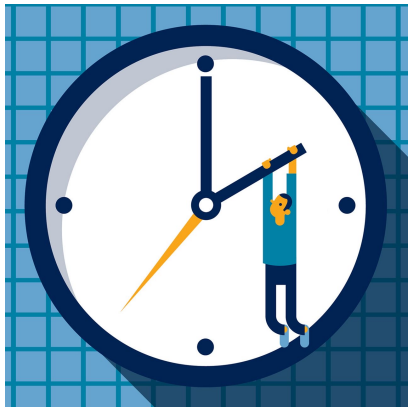
Vaarya Srivastava

Qihang Tang

Contact Information

Teaching Assistants

- Sai Meka (smeka@ur.rochester.edu) (Head TA)
- Xinyu Cai (xcai12@u.rochester.edu)
- Ryan Hilton (rhilton2@ur.rochester.edu)
- Yiheng Mao (ymao16@u.rochester.edu)
- Duy Pham (dpham4@u.rochester.edu)
- Amanda Pignataro (apignata@u.rochester.edu)
- Vaarya Srivastava (vsrivast@u.rochester.edu)
- Qihang Tang (qtang5@u.rochester.edu)



- Office: TBD
- Office Hours: TBD

Communication

You ‘should’ communicate with the **TAs**:

- Class content
- Clarification of theory
- Grading-related issues for individual assignments

You should communicate with the **Head TA**:

- Class content
- Clarification of theory
- (Bigger) grading disputes

You should communicate with the Instructor:

- Class content
- Clarification of theory

Please try to schedule an appointment if you have questions!

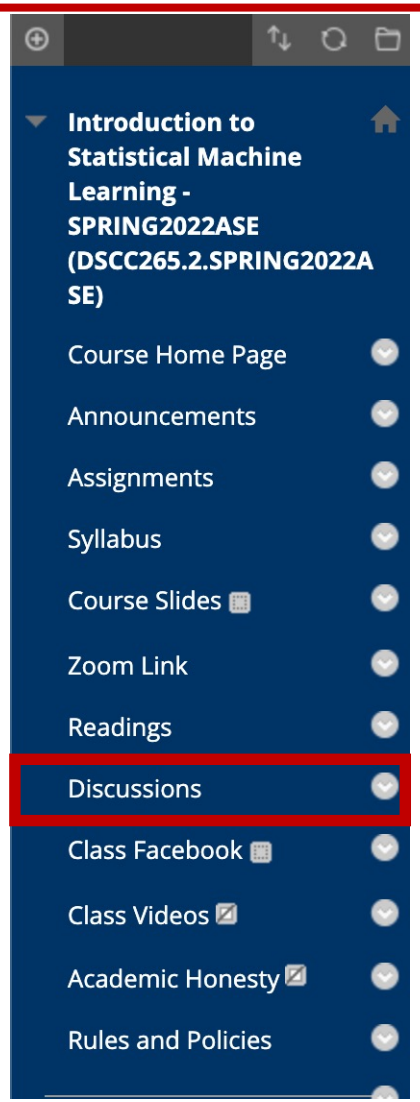
I prefer virtual appointments over email.

I may not be able to respond to emails quite timely (we have 165 students!)

Discussion

Discussion between students is highly encouraged!

- Please check the '*Discussions*' tab on BlackBoard:
  - To ask new questions
  - To answer others' questions
  - Might be better than sending an email



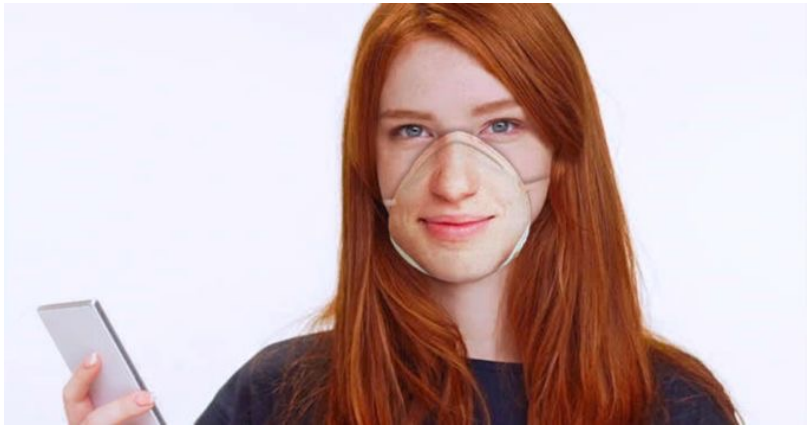
Your first homework is coming today!

And, it is ***extra credit***!

Your first assignment -> Extra credit

Please write a paragraph about yourself and talk about (posted as Introductions on *BlackBoard*):

1. **Add your picture**
2. **What is your first name and last name?**
3. **What is your email address?**
4. **Where are you from?**
5. **What is/are your field(s) of study?**
6. **What are the Data Science topics you are interested in?**
7. **Do you want to see any Data Science applications in particular in this course?**
8. **Anything you want to surprise us and make us smile!**
9. **Send the file in .pdf format!**



And add your picture without the mask!

Course Goals

- Understand the mathematical background of algorithms
- Code algorithms from scratch and comparing your code's performance to pre-packaged algorithms
- Apply your code to empirical problems
- Develop an understanding of different statistical philosophies
- Handle missing data with both simple and sophisticated techniques
- Describe structures in data using unsupervised algorithms
- Reduce dimensionality of data for easier storage, computation, and analysis
- Estimate algorithm run times and efficiently handle large computations
- Understand both the 'breadth' and the 'depth' of the material
- Develop an understanding of what a good ML problem/puzzle might be

Who should take this course?

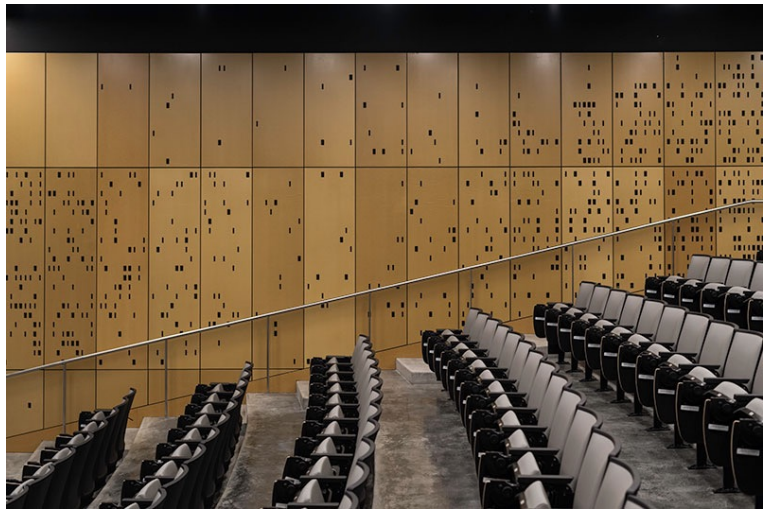
It is essential to know the mechanics of the algorithms!

This course is for you if:

- You are a Data Science major / Data Science Master's student
- You want to learn about the mathematical background of the ***data science project*** algorithms that you are already using
- You want to prepare yourself to have a theoretical understanding of Data Science
- You still want to be involved in empirical problems and visualizatio

Course Summary

- **Math Review**
  - Linear algebra and probability theory
- **Supervised Learning: Regression**
  - Regression, Gradient Descent
- **Supervised Learning: Classification**
  - Finding patterns, Regularization
- **Unsupervised Learning: Clustering**
  - K-means, Gaussian Mixtures
- **Unsupervised Learning: Dimension Reduction**
  - PCA, t-SNE
- **Probabilistic Models**
  - LDA, Generative vs. Discriminative Models



Course Summary

- **Support Vector Machines**
  - Methods, kernels, loss functions
- **Application: Feature Selection**
  - Feature selection and feature pre-processing
- **Application: Missing Value Estimation**
  - Frequentist and Bayesian Approaches
- **Application: Ethics in Machine Learning**
  - Bias, fairness, transparency, accountability
- **Network Analysis: Descriptive Concepts**
  - Understanding social network data
- **Network Analysis: Models**
  - Clustering, tie prediction
- **Final Project**

General Rules and Policies

- **Please use the laptops only for class related purposes!**
- **Please no cellphones during class**
- **No photography**

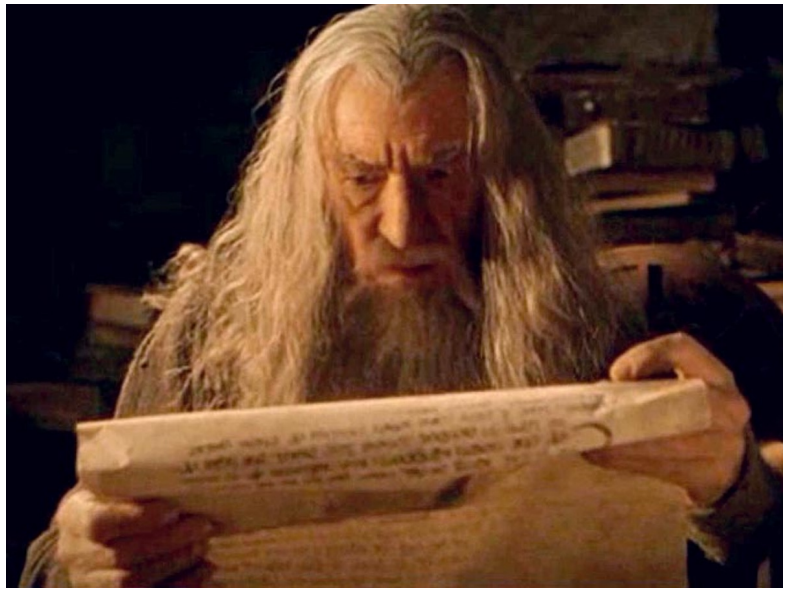
- You can ask your questions to the TAs or to me:
  - Please be patient, it may take a few hours or more for us to respond
- No intellectual theft -> please read the *Code of Academic Integrity*
  - Discussing ideas is fine!
- Curved grading might be possible for some assignments

Meetings

- For the moment, we will be meeting online through ***Zoom***: https://rochester.zoom.us/j/94823489174

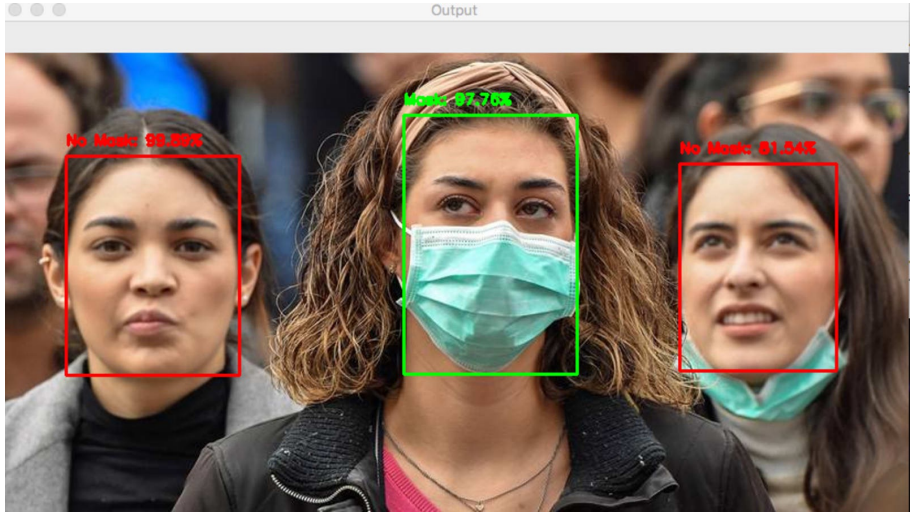
- We will have two ***lectures*** per week**:**
  - Mondays and Wednesdays (2:00 PM – 3:15 PM)
  - Normally: *Wegmans Hall 1400*

- No lab sessions are planned for the class.



Social Distancing Policies (in-person)

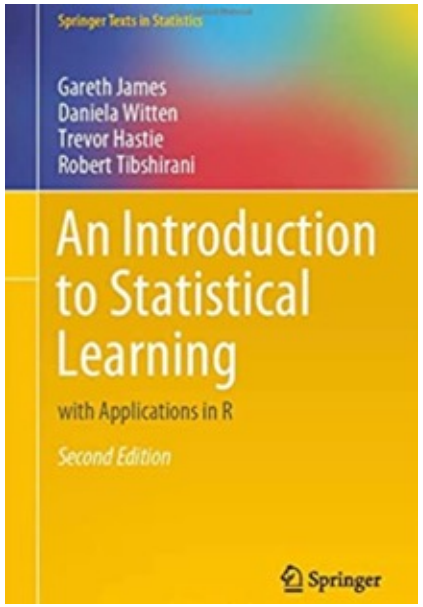
- Masks need to be on in all lectures
- Drinking water is fine ☺
- Please be careful about eating, your classmates may not be comfortable…



Textbook

- Textbook:
  - **An Introduction to Statistical Learning: with Applications in R**

  by James, Witten, Hastie, and Tibshirani (2nd edition, 2021
  - The book is free! (and uploaded on BlackBoard)



- Necessary files will be uploaded on *BlackBoard*
- Datasets, tutorials, and online links will be shared
- Recordings of the lectures will be posted for the first few weeks

Equipment

- It is much better, if you have:
- A PC
- OR a MacBook
- Than having ...
- Chromebook
- Or no laptop

Course Website

BlackBoard: http://learn.rochester.edu

- Please check the class page!
- Let me know if you have any questions!

BlackBoard will be used for:

- Assignments
- Questions
- Discussions
- Sharing reading materials

Problem Sets

- There will be ‘many’ of them (**40% of the grade**)
  - **Eight short problem sets (5% each)**
  - **One per week for the first 2/3 of the class**
- Will be on the weekly topics and cover different themes
- You can work in a group of **up to two (2) people**
- Guidance will be provided
- You will have one week to complete them

Midterm

Midterm:

- **20% of your grade**
- Will be scheduled for the week after *Spring Break*
- Will mainly test your theoretical understanding of the class material in a conventional written / short paragraph format
- If the class is still held virtually -> there will be a *take-home* midterm
- It will be completed **individually**

Challenge

Challenge:

- **10% of your grade**
- Will be introduced in the last **1/3** of the class
- A programming Project that will give you a chance to apply your knowledge and develop an ML solution to a real-world problem
- Will be organized as a **Kaggle** competition
- Will also test your theoretical understanding of the class material
- More information will be provided later.

Final Project

- A project that you will design on your own (in a team of **~5 people**)
- Can be ***descriptive*** / ***explanatory*** / ***predictive***
- Will be an empirical project that uses advanced ML algorithms to solve a puzzle / to explain a phenomenon
- You can get your inspiration from problem sets / your external projects
- **30%** of your grade
  - Final prospectus (**5%**) + Final presentation (**10%**) + Final report (**15%**)
  - **Graduates and undergraduates will be graded differently**

Let’s take a look at the schedule

Weeks	Topics	Readings	Assignments
Week 1 Jan. 12	**Introduction to Course and Review** Topics, policies, questions	Chapter 1	No assignment
Week 2 Jan. 19	**Math Review** Review of linear algebra and probability theory	Chapter 2	PS 1
Week 3 Jan. 24, 26	**Supervised Learning I: Regression** Regression, linear hypothesis, gradient descent	Chapter 3	PS 2
Week 4 Jan. 31, Feb. 2	**Supervised Learning II: Classification/Regularization** Finding patterns, bias/variance	Chapter 4	PS 3
Week 5 Feb. 7, 9	**Unsupervised Learning I: Clustering** Clustering, k-means, Gaussian mixtures	Chapter 12	PS 4
Week 6 Feb. 14, 16	**Unsupervised Learning II: Dimension Reduction** PCA, T-SNE	**No reading**	PS 5
Week 7 Feb. 21, 23	**Application I: Practical ML** Feature selection and feature pre-processing	Article on feature selection	PS 6
Week 8 Feb. 28, Mar. 2	**Probabilistic Models: LDA and Bayesian Methods** Gen. linear models, generative vs. discriminative	Revise Chapter 4	PS 7
Week 9 Mar. 7, 9	**SPRING BREAK**		No assignment
Week 10 Mar. 14, 16	**Support Vector Machines** Methods, kernels, loss functions	Chapter 9	**Midterm**
Week 11 Mar. 21, 23	**Application II: Missing Value Estimation** Frequentist and Bayesian approaches	Article on missing value imputation	PS 8
Week 12 Mar. 28, 30	**Application III: Ethics in Machine Learning** Bias, fairness, transparency, accountability	Article on bias and fairness in ML	Challenge
Week 13 Apr. 4	**Network Analysis I: Descriptive Concepts** Understanding the data – **Class is cancelled on Apr. 6!**	**No reading**	Challenge
Week 14 Apr. 11, 13	**Network Analysis II: Models** Clustering, tie prediction	**No reading**	Final Project Preparation
Week 15 Apr. 18, 20	**Review, Final Project (Presentations)**	**No reading**	**Final Project**
Week 16 Apr. 25, 27	**Final Project (Presentations)**	**No reading**	**Final Project**

Special Accommodation

- If you have needs for special accommodation, let me know as soon as possible (after the class or through e-mail).

Read the syllabus!

- Please go through the syllabus for more info
  - It may contain some additional information that we weren't able to cover here

Please do the following until next week!

Start with your 'homework' and do the reading:

- Write a short paragraph about yourself and send a **.pdf** document with your picture
- Do the reading for this week: *Chapter 1* of the textbook
- Start with the reading for the next week: *Chapter 2*

And install:

- The Anaconda distribution for *Python*:
- https://www.anaconda.com/distribution/#download-section

Install the Python 3.9 version using the graphical installer!

Questions?