DS 4400

Machine Learning and Data Mining I
Spring 2024

David M. Liu
Computer Science PhD Candidate
Khoury College of Computer Science
Northeastern University

Tuesday January 9, 2024

Welcome to DS 4400!

Machine Learning and Data Mining I

Today's Goals

1. Introduce myself and TAs
2. What is Machine Learning and why is it important?
3. Course goals / outline
4. Logistics / administrative information
5. Brief introduction to machine learning

DS 4400 Class – About You

- Enrollment of 87 students
- Diverse majors
  - Roughly 50/50 split between CS and DS majors
  - Joint majors: DS/Psychology, DS/Business, DS/Math, DS/Biology, etc.

DS 4400 – About Me

David M. Liu

(he/him/his)

- B.S.E. from Princeton [2018]
  - Specialized in statistics and machine learning especially the applications of these to social sciences.
- Software Engineer at Bloomberg LP [2018 – 2020]
  - Built and maintained a data pipeline to help clients make financial decisions
- PhD at NEU [2020 – Present]
  - Conducting research on the societal impact of ML (e.g. social biases of these models) and ethics of AI.
  - Advised by Professor Tina Eliassi-Rad
  - Previously interned at Meta AI and visited UC Berkeley.

* Happy to talk more about either my experiences in industry or academia!

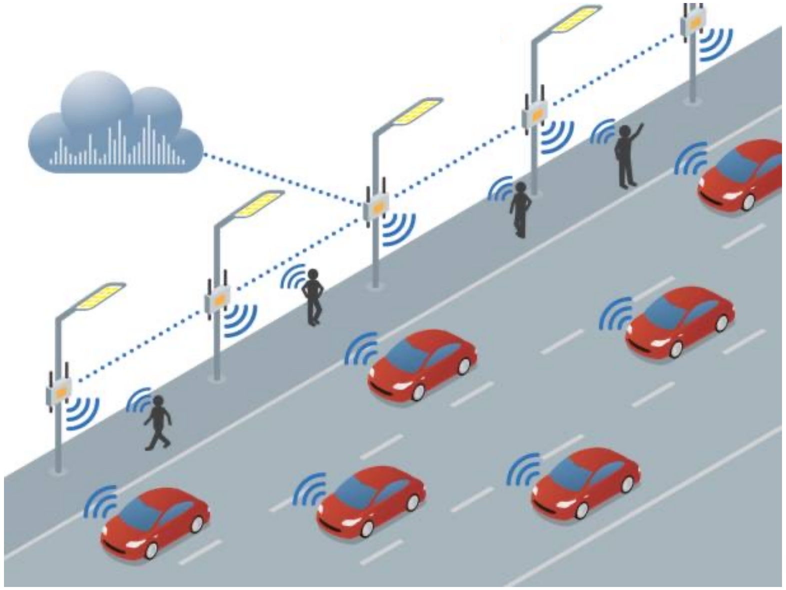
TA Introduction

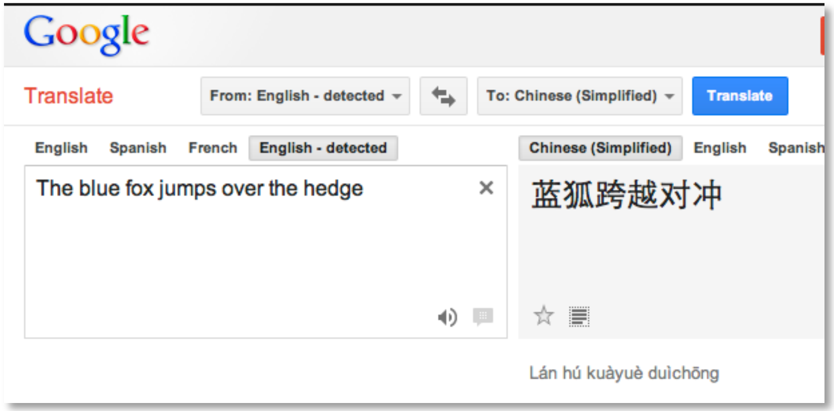
Caleb Lee
4th year CS Undergrad
(he/him/his)

Jai Amin
3rd year CS Undergrad
(he/him/his)

Dhanush Akula
1st year MS in Data Science
(he/him/his)

Machine Learning is Everywhere

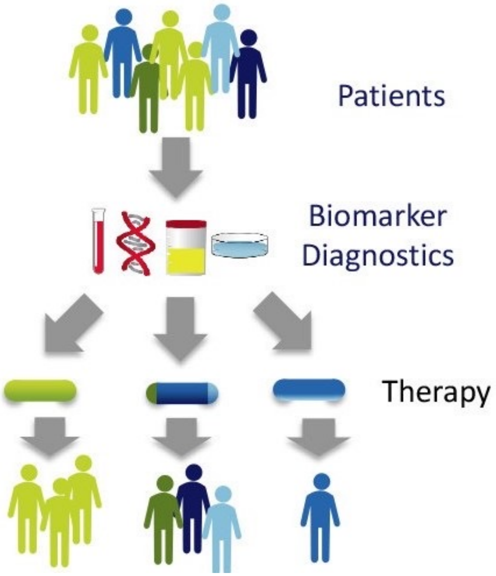




The blue fox jumps over the hedge

蓝狐跨越对冲

Lán hú kuàyuè duìchōng



Patients

Biomarker Diagnostics

Therapy

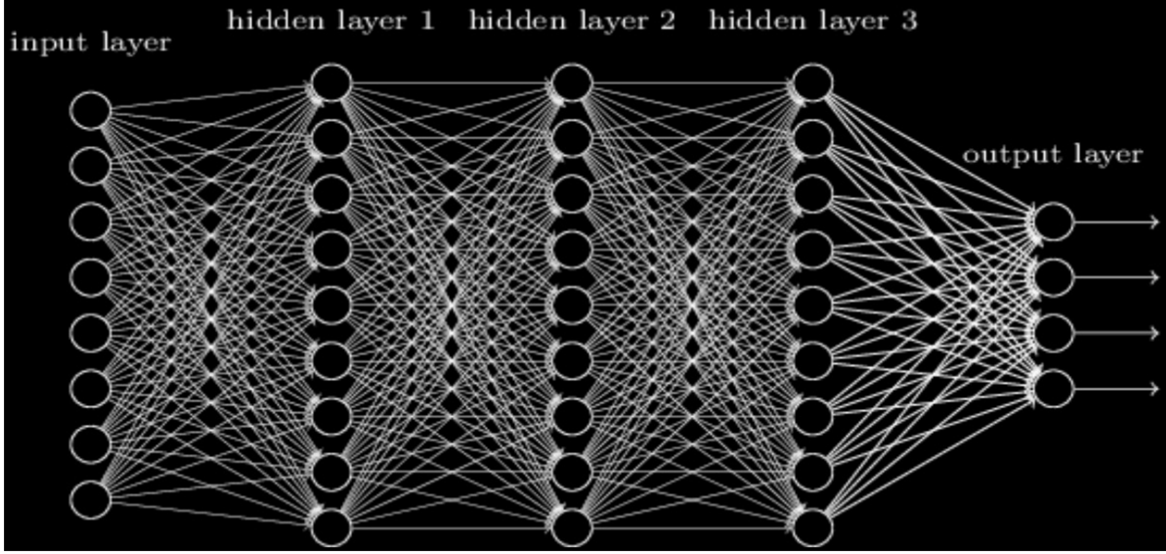
Short History

- Legendre and Gauss – linear regression / least squares, 1805
  - Astronomy applications
- Probabilistic models
  - Bayes and Laplace - Bayes Theorem, 1812
  - Markov chains, 1913
- Fisher – linear discriminant analysis for classification, 1936
  - Logistic regression, 1940
- Widrow and Hoff ADALINE neural network, 1959
- Nelder, Wedderburn, generalized linear models, 1970
- “AI winter”, limitations of perceptron and linear models, 1970
- Breiman, Friedman, Olshen, Stone, decision trees (non-linear models), 1980
- Cortes and Vapnik, SVM with kernels, 1990
- IBM Deep Blue beats Kasparov at chess, 1996
- Geoffrey Hinton, Deep learning, back propagation, 2006
- C. Szedegy: Adversarial manipulation of image classification, 2013

Deep Learning

Neural networks return and excel at image recognition, speech recognition, …

The 2018 Turing award was given to Yoshua Bengio, Geoff Hinton, and Yann LeCun.



Safety Concerns of AI

- Ethics and fairness of AI
  - Everyone is treated fairly
  - Robots will not perform harmful actions
  - Can the technology be used for nefarious purposes?
- Economic concerns
  - Might automate / displace some type of jobs in manufacturing, transportation, etc.
- Adversarial ML
  - ML can be manipulated
  - Small change in input results in different prediction

Secure and Robust ML

Image Recognition

Misreading traffic signs (Eykholt et al)

Speech recognition

Hide commands in noise (Carlini & Wagner)

Poisoning Attacks

Tay (chat bot) became inflammatory in 16 hr.

STOP
STOP
STOP
12:45

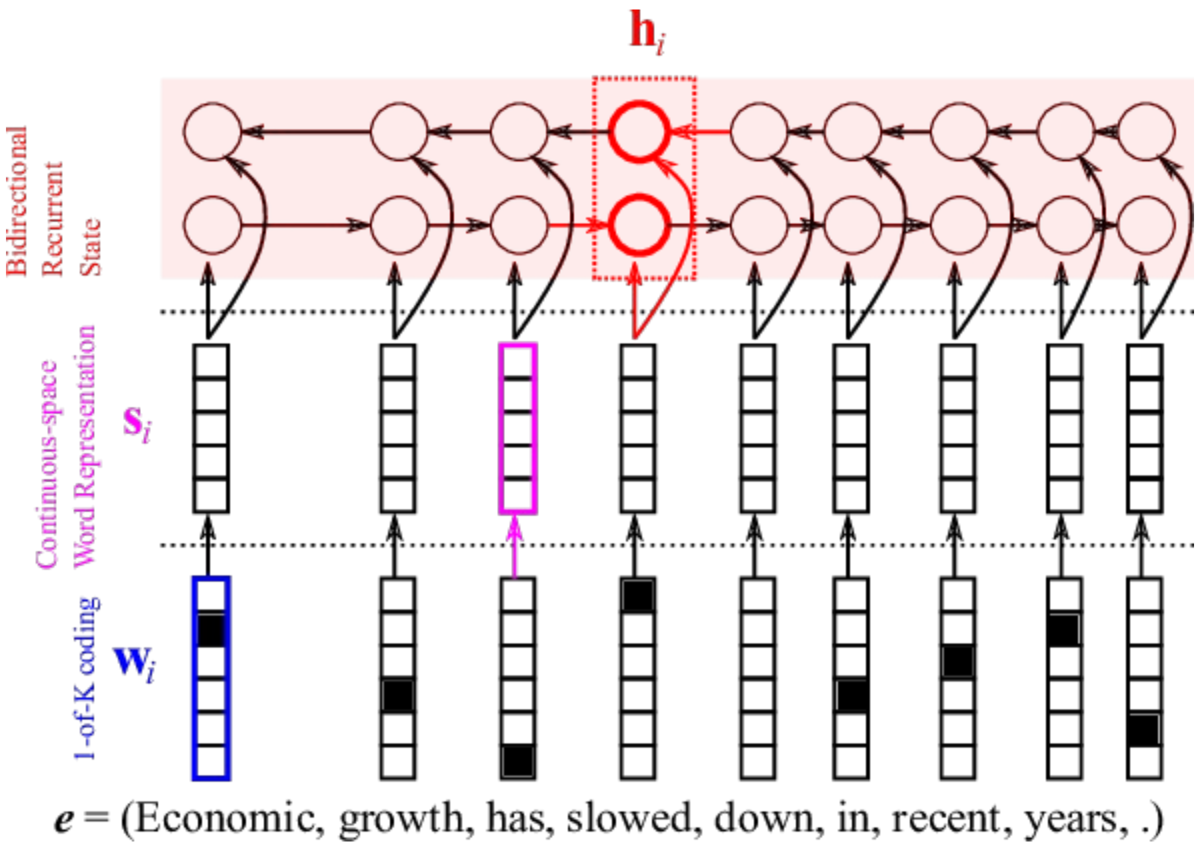
How to create safe and robust machine learning?

Applications of ML

- Healthcare
- Vision
- NLP
- Speech recognition
- Self-driving cars
- Stock market analysis
- Recommendations
- Sentiment analysis
- Human behavior
- Quality of life

- Business
- Sports
- Bots / chatbots
- Science / engineering
- Bioinformatics
- Precision medicine
- Unsupervised learning
- Reinforcement learning

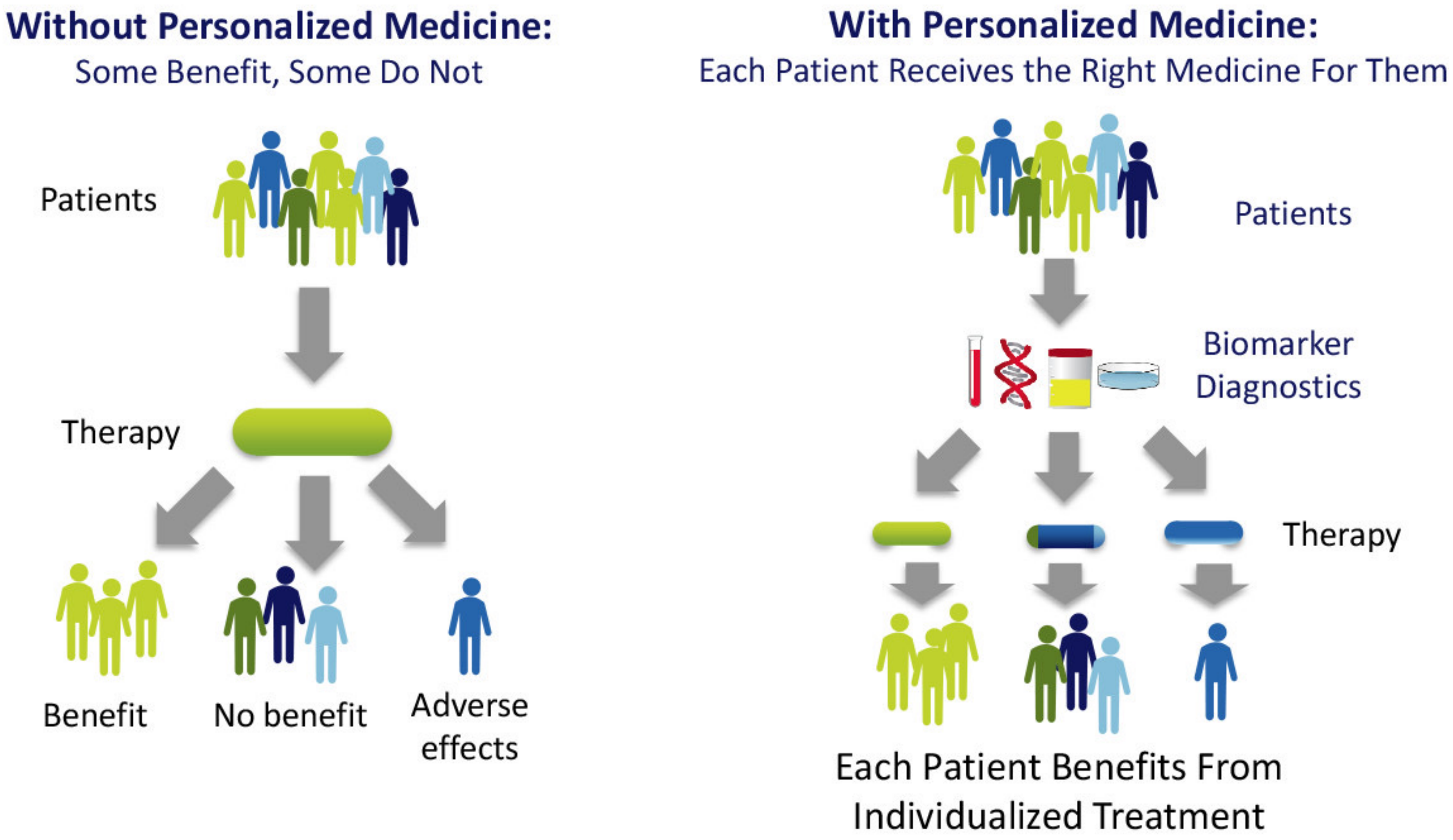
Natural Language Processing (NLP)



$\mathbf{h}_i$
Bidirectional Recurrent State
Continuous-space Word Representation
$\mathbf{s}_i$
1-of-K coding
$\mathbf{w}_i$
e = (Economic, growth, has, slowed, down, in, recent, years, .)

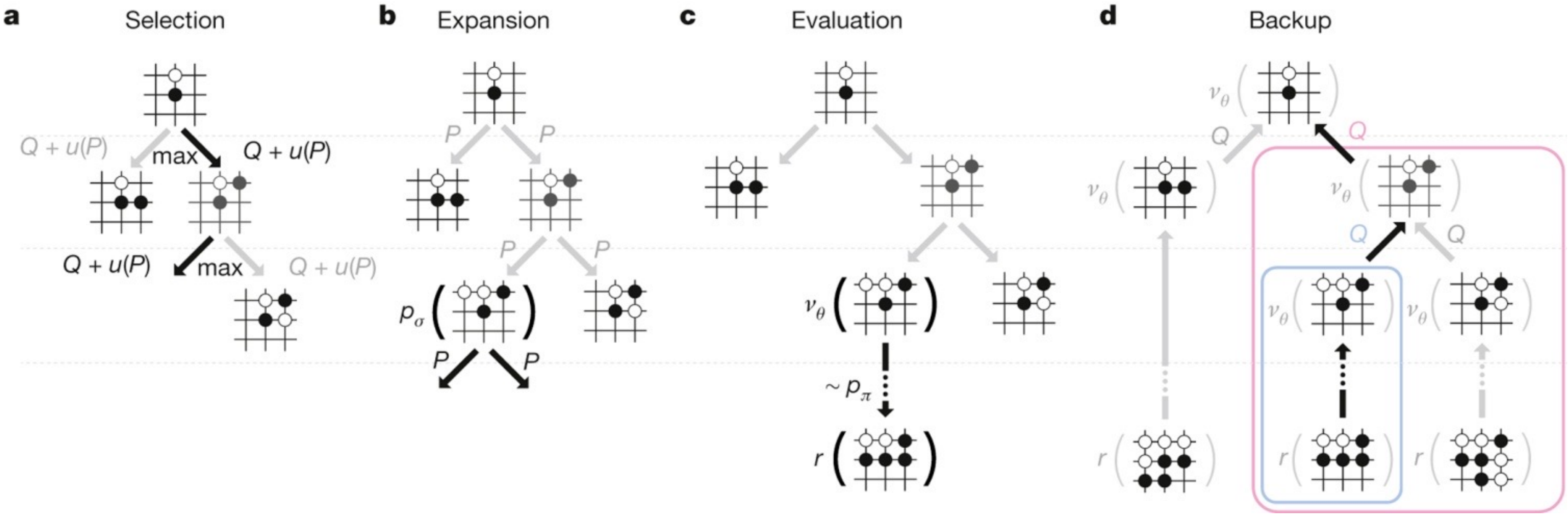
- Understand language semantics
- Real-time translation, speech recognition, question answering
- Large generative language models: BERT, GPT-2, GPT-3

Personalized medicine

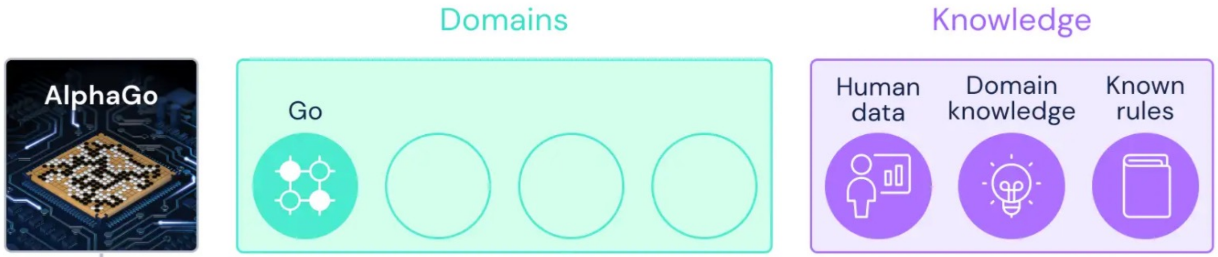

Without Personalized Medicine:
Some Benefit, Some Do Not
Patients
Therapy
Benefit
No benefit
Adverse effects
With Personalized Medicine:
Each Patient Receives the Right Medicine For Them
Patients
Biomarker Diagnostics
Therapy
Each Patient Benefits From Individualized Treatment

- Treatment adjusted to individual patients
- Predictive models using a variety of features related to patient history and genetics

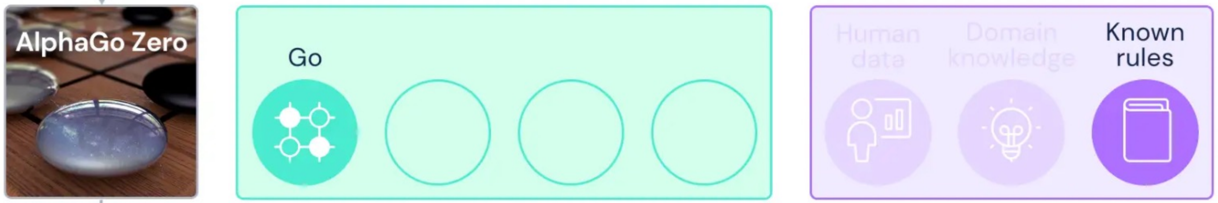
Playing games



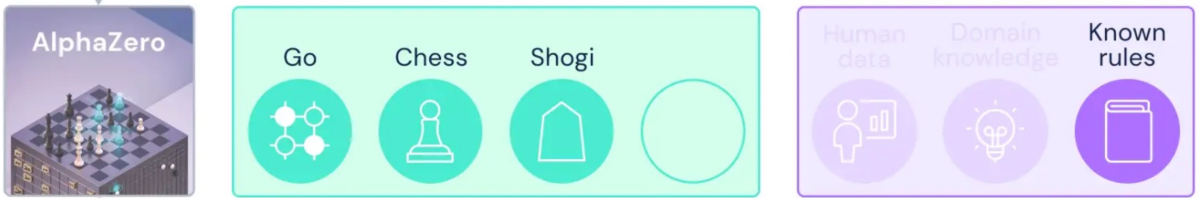
- AlphaGo: DeepMind beats world champion in 2016
- Interestingly, it discovered new, unknown strategies
- Go is the most challenging game for AI
- Algorithms based on deep reinforcement learning



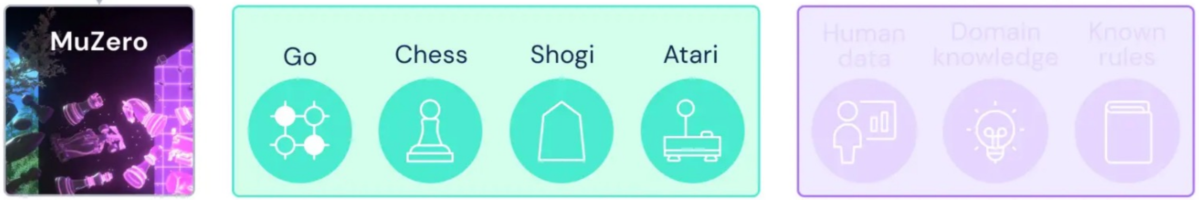
AlphaGo becomes the first program to master Go using neural networks and tree search
(Jan 2016, Nature)



AlphaGo Zero learns to play completely on its own, without human knowledge
(Oct 2017, Nature)



AlphaZero masters three perfect information games using a single algorithm for all games
(Dec 2018, Science)



MuZero learns the rules of the game, allowing it to also master environments with unknown dynamics.
(Dec 2020, Nature)

DS-4400

- What is *machine learning*?
  - The development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data
  - Design predictive algorithms that learn from data
  - Subset of Artificial Intelligence (AI): The study of "intelligent agents" that perceive their environment and take actions that maximize their chance of achieving their goals
- Machine learning is currently very successful in:
  - Machine translation
  - Voice assistants
  - Recommendation systems
  - Image recognition
- Why the hype?
  - Availability: data created/reproduced in 2010 reached 1,200 exabytes
  - Reduced cost of storage
  - Computational power (cloud, multi-core CPUs, GPUs)

DS-4400 Course objectives

- Become familiar with main machine learning tasks
  - Supervised learning vs unsupervised learning
  - Classification vs Regression
  - Focus on supervised learning
- Study most well-known algorithms
  - Regression (linear regression, spline regression)
  - Classification (SVM, decision trees, Naïve Bayes, ensembles, etc.)
  - Deep learning (different neural network architectures)
- Learn the theory and foundation behind ML algorithms and learn to apply them to real datasets
- Learn about security challenges of ML and ethical issues
  - Introduction to adversarial ML

https://dliu18.github.io/teaching/ds4400spr24/

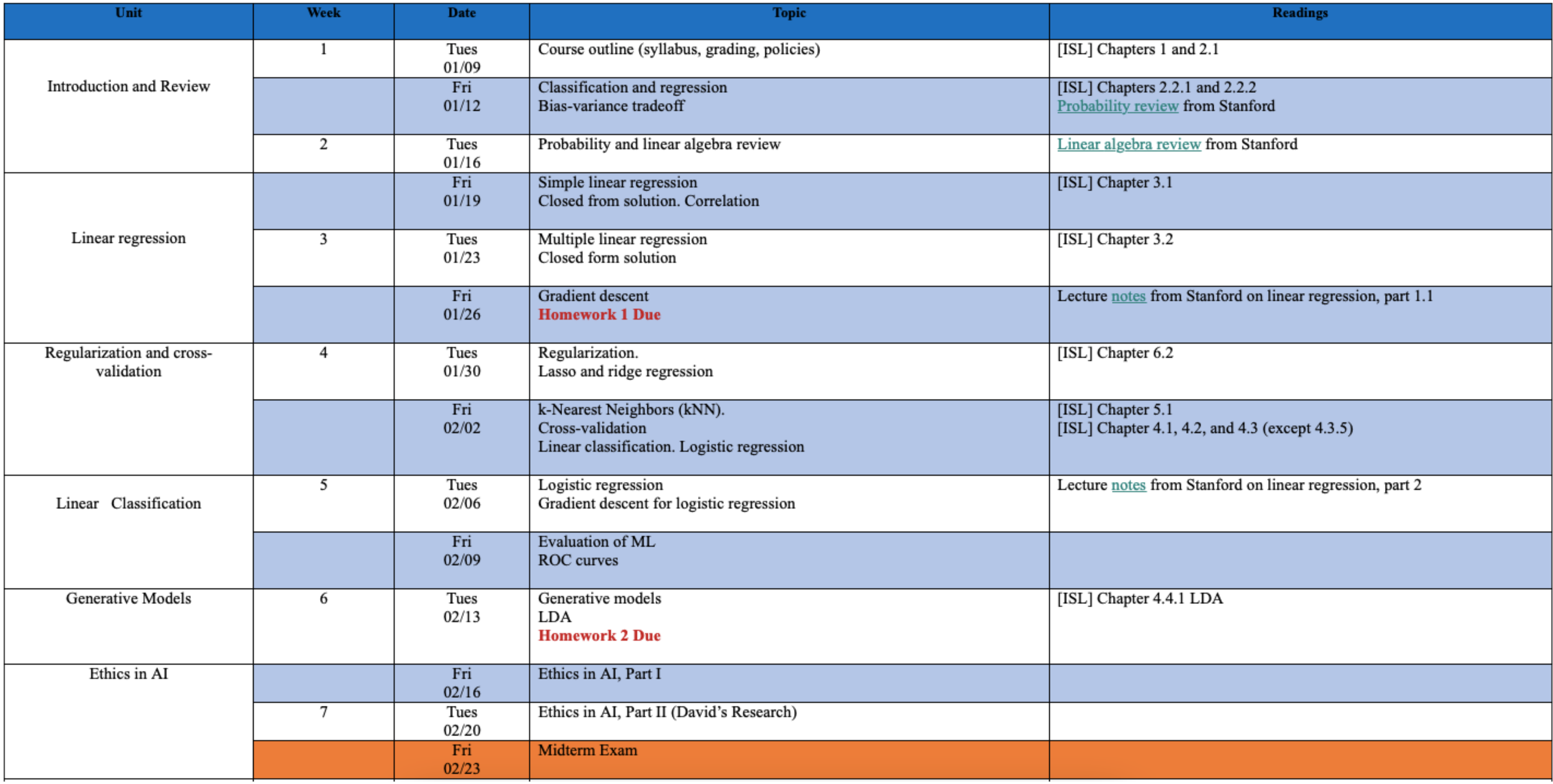
Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Linear regression and regularization – 2 weeks
- Classification - 5 weeks
  - Linear classifiers: logistic regression, LDA,
  - Non-linear: kNN, decision trees, SVM, Naïve Bayes
  - Ensembles: random forest, boosting
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  - Back-propagation, gradient descent
  - NN architectures (feed-forward, convolutional, recurrent)
- Ethics of AI – 2 lectures

Class Outline

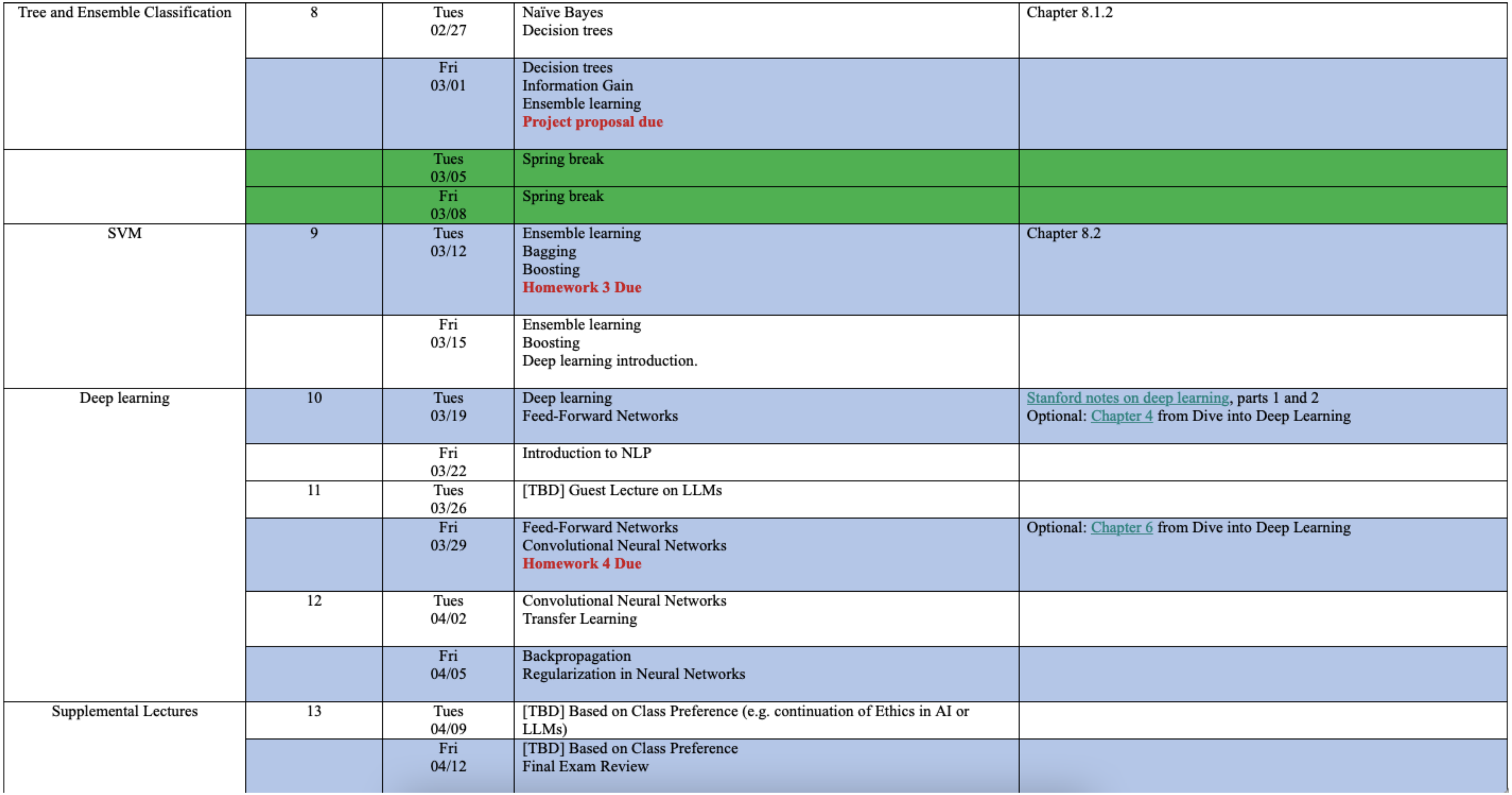
- Roughly 5 customized lectures
  - Goals will be to cover recent developments in machine learning (e.g. Large Language Models)
  - Discuss Ethics in ML/AI as well as my own research
  - Feature guest lectures

Class Outline



Unit	Week	Date	Topic	Readings
Introduction and Review	1	Tues 01/09	Course outline (syllabus, grading, policies)	[ISL] Chapters 1 and 2.1
		Fri 01/12	Classification and regression Bias-variance tradeoff	[ISL] Chapters 2.2.1 and 2.2.2 Probability review from Stanford
	2	Tues 01/16	Probability and linear algebra review	Linear algebra review from Stanford
Linear regression		Fri 01/19	Simple linear regression Closed from solution. Correlation	[ISL] Chapter 3.1
	3	Tues 01/23	Multiple linear regression Closed form solution	[ISL] Chapter 3.2
		Fri 01/26	Gradient descent **Homework 1 Due**	Lecture notes from Stanford on linear regression, part 1.1
Regularization and cross-validation	4	Tues 01/30	Regularization. Lasso and ridge regression	[ISL] Chapter 6.2
		Fri 02/02	k-Nearest Neighbors (kNN). Cross-validation Linear classification. Logistic regression	[ISL] Chapter 5.1 [ISL] Chapter 4.1, 4.2, and 4.3 (except 4.3.5)
Linear Classification	5	Tues 02/06	Logistic regression Gradient descent for logistic regression	Lecture notes from Stanford on linear regression, part 2
		Fri 02/09	Evaluation of ML ROC curves	
Generative Models	6	Tues 02/13	Generative models LDA **Homework 2 Due**	[ISL] Chapter 4.4.1 LDA
Ethics in AI		Fri 02/16	Ethics in AI, Part I	
	7	Tues 02/20	Ethics in AI, Part II (David's Research)	
		Fri 02/23	Midterm Exam	

Class Outline



Tree and Ensemble Classification	8	Tues 02/27	Naïve Bayes Decision trees	Chapter 8.1.2
		Fri 03/01	Decision trees Information Gain Ensemble learning **Project proposal due**	
		Tues 03/05	Spring break	
		Fri 03/08	Spring break	
SVM	9	Tues 03/12	Ensemble learning Bagging Boosting **Homework 3 Due**	Chapter 8.2
		Fri 03/15	Ensemble learning Boosting Deep learning introduction.	
Deep learning	10	Tues 03/19	Deep learning Feed-Forward Networks	Stanford notes on deep learning, parts 1 and 2 Optional: Chapter 4 from Dive into Deep Learning
		Fri 03/22	Introduction to NLP	
	11	Tues 03/26	[TBD] Guest Lecture on LLMs	
		Fri 03/29	Feed-Forward Networks Convolutional Neural Networks **Homework 4 Due**	Optional: Chapter 6 from Dive into Deep Learning
	12	Tues 04/02	Convolutional Neural Networks Transfer Learning	
		Fri 04/05	Backpropagation Regularization in Neural Networks	
Supplemental Lectures	13	Tues 04/09	[TBD] Based on Class Preference (e.g. continuation of Ethics in AI or LLMs)	
		Fri 04/12	[TBD] Based on Class Preference Final Exam Review	

Course Information

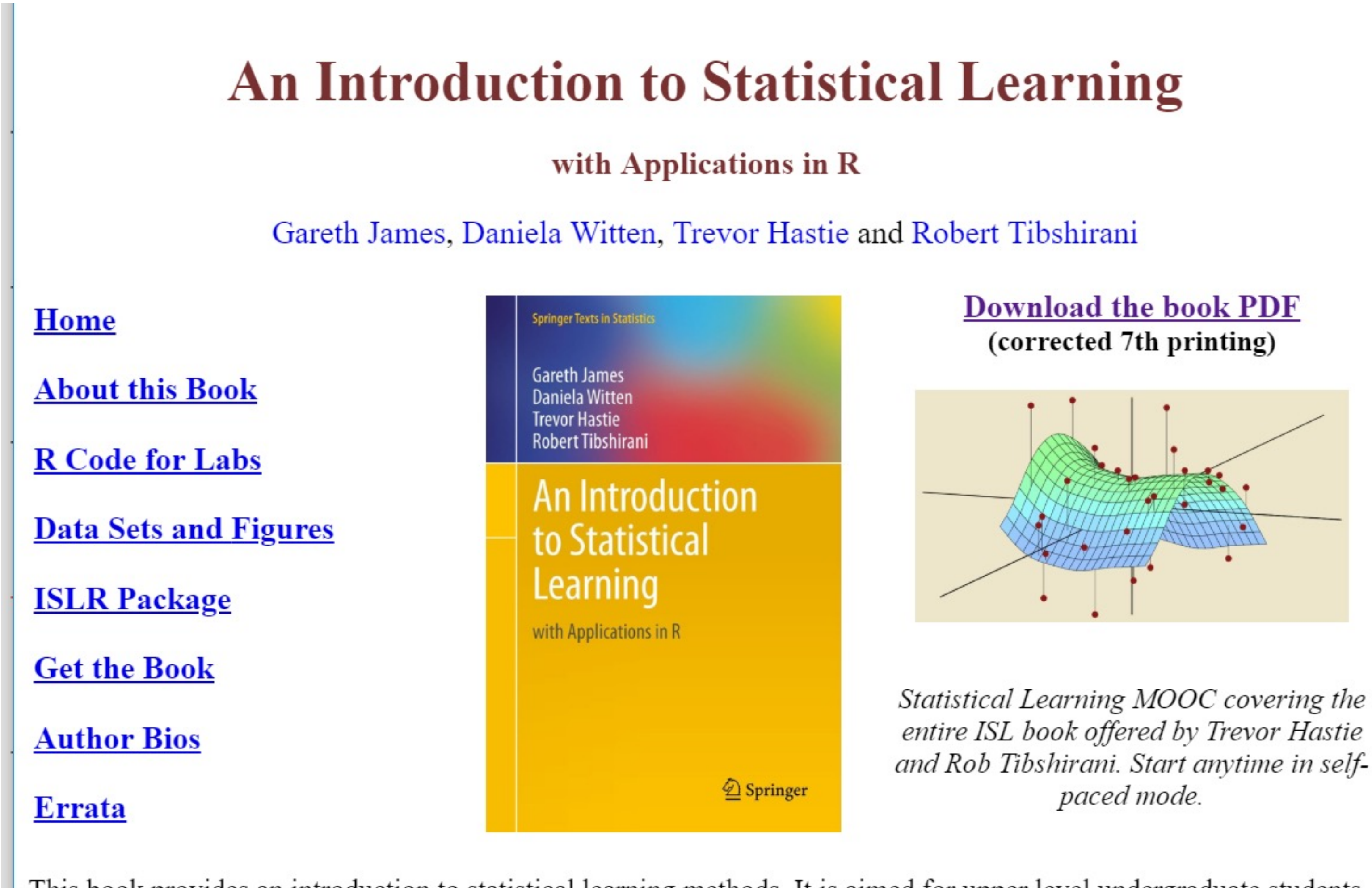
- **Website:** https://dliu18.github.io/teaching/ds4400spr24/
  - Course calendar and slides posted after lecture.

- **Canvas**
  - Assignments and grades posted here.

- **Gradescope**
  - Assignment submissions
  - Accessed via Canvas

- **Piazza**
  - Course and material discussion

Textbook



An Introduction to Statistical Learning

with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

- Home
- About this Book
- R Code for Labs
- Data Sets and Figures
- ISLR Package
- Get the Book
- Author Bios
- Errata

Download the book PDF
(corrected 7th printing)

Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students

Will utilize the Python version this semester.
Freely available online

Other resources

- Trevor Hastie, Rob Tibshirani, and Jerry Friedman, Elements of Statistical Learning, Second Edition, Springer, 2009.

- Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

- A. Zhang, Z. Lipton, and A. Smola. Dive into Deep Learning

- Lecture notes by Andrew Ng from Stanford

Schedule

- Schedule
  - Tuesday, Friday 9:50-11:30am ET, Churchill Hall 103
  - Office hours
    - Timing and location (in-person vs online) is TBD based on class input. Will ensure office hours are available most days of the week.
    - **For the first week, David will host office hours virtually on Thursday January 11 from 5-6pm. Zoom link posted on Piazza.**
- Online resources
  - Slides and lecture notes will be posted after each lecture
  - Use Piazza for questions

Policies

- Your responsibilities
  - Please be on time, attend classes, and take notes
  - Participate in interactive discussion in class
  - Submit assignments/ programming projects on time
- Late days for assignments
  - 5 total late days, after that loose 20% for every late day
  - Assignments are due at 11:59pm on the specified date
  - We will use Gradescope for submitting assignments
  - No need to email for late days

Grading

- Assignments – 25%
  - 4 assignments and programming exercises based on studied material in class
- Final project – 30%
  - Select your own project based on public dataset
  - Submit short project proposal and milestone
  - Record presentation (3 min) and written report
  - Team of 3 students
- Midterm Exam –20%
  - Tentative date: Friday, February 23
- Final Exam – 20%
  - Scheduled during finals week
- Class participation – 5%

Assignments

- Several theoretical questions and many programming exercises
- Language
  - Python
  - Jupyter notebooks recommended
  - Will share some numpy and panda tutorials
- Submission
  - Submit PDF report
  - Includes all the results, as well as link to code

Final project

- Goal: work on a larger data science project
  - Build your portfolio and increase your experience
- Requirements
  - Large dataset: at least 20,000 records (public source)
  - Not recommended to collect your own data
  - Pick application of interest
  - We will also provide a list of projects and datasets
  - Experiment with at least 4 ML models
  - Perform in-depth analysis (which features contribute mostly to prediction, which model performs best, explain results)
  - Teams of 2 students, will have a TA assigned
- Computational resources: NEU Discovery cluster, Google cloud, AWS, Google collab
- Timeline
  - Proposal: mid class; milestone 3 weeks after (Instructors will provide early feedback)
  - Final presentation (recorded video) and report (6-8 pages)

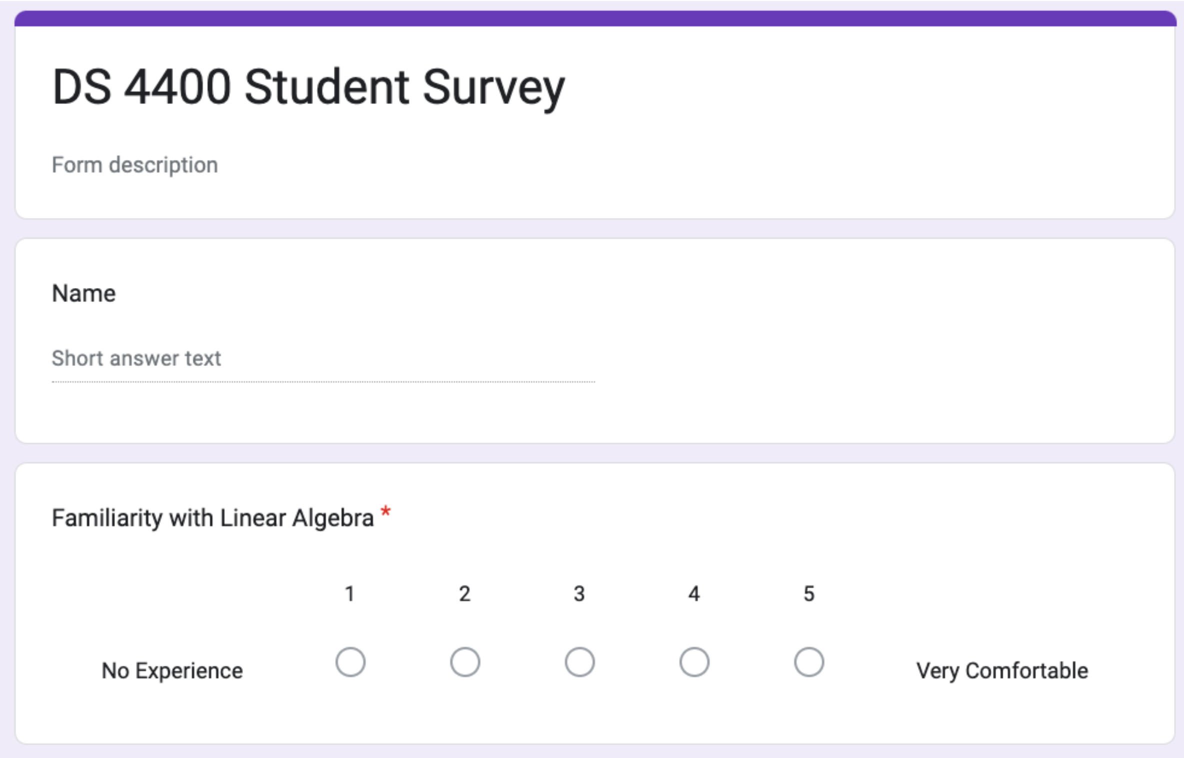
Academic Integrity

- Homework is done individually!
- Final project is done in the team!
- Rules
  - Can discuss with colleagues or instructors
  - Can post and answer questions on Piazza
  - Code cannot be shared with colleagues
  - Cannot use code from the Internet
    - Use python packages, but not directly code for ML analysis written by someone else
- NO CHEATHING WILL BE TOLERATED!
- Any cheating will automatically result in grade F and report to the university administration
- http://www.northeastern.edu/osccr/academic-integrity-policy/

Action Item: Google Form

https://forms.gle/e9n7bkSZNKBF9iqp9

* Posted on Piazza



DS 4400 Student Survey

Form description

Name

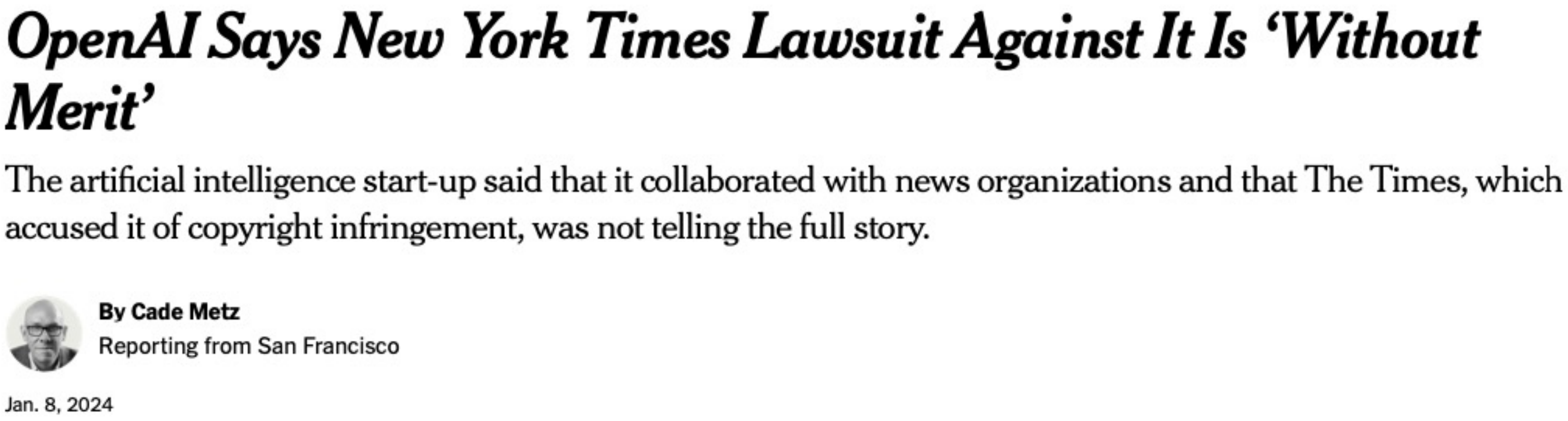
Short answer text

Familiarity with Linear Algebra *

	1	2	3	4	5	
No Experience	○	○	○	○	○	Very Comfortable

Please Share Feedback Over the Course of the Semester!

Brief Introduction to Machine Learning



OpenAI Says New York Times Lawsuit Against It Is 'Without Merit'

The artificial intelligence start-up said that it collaborated with news organizations and that The Times, which accused it of copyright infringement, was not telling the full story.

By Cade Metz
Reporting from San Francisco

Jan. 8, 2024

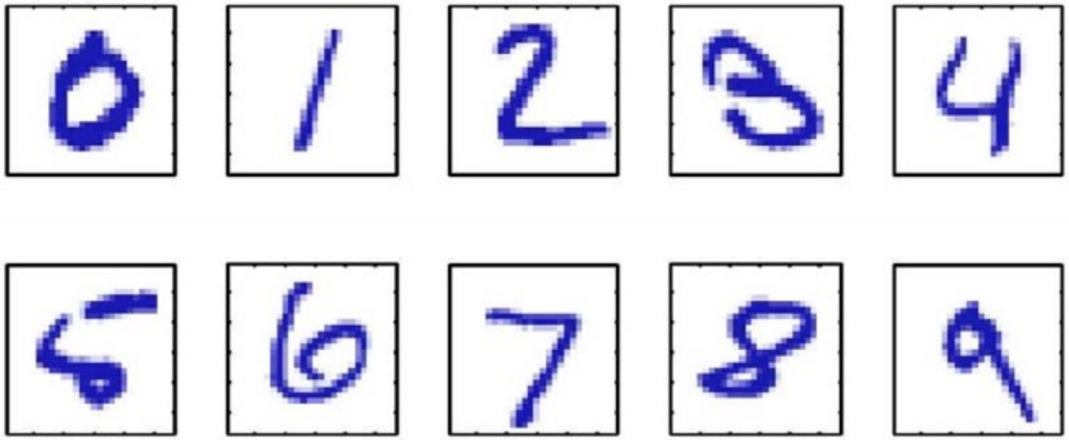
Supervised vs Unsupervised Learning

- Supervised learning
  - Classification
  - Regression
  - Examples
- Unsupervised learning
  - Clustering

Slides adapted from

- A. Zisserman, University of Oxford, UK
- S. Ullman, T. Poggio, D. Harari, D. Zysman, D Seibert, MIT
- D. Sontag, MIT
- Figures from “An Introduction to Statistical Learning”, James et al.

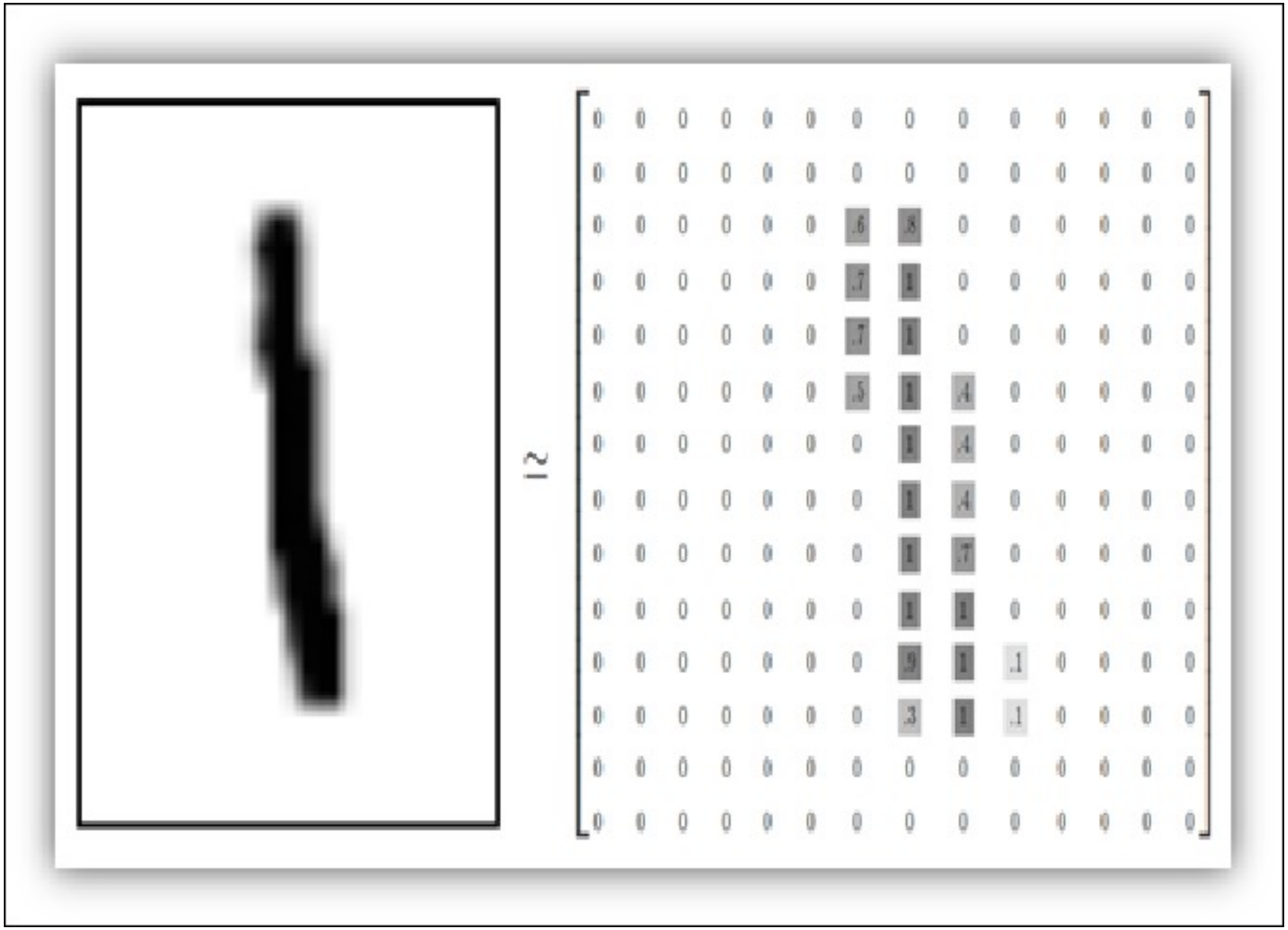
Example 1
Handwritten digit recognition



Images are 28 x 28 pixels

MNIST dataset: Predict the digit
Multi-class classifier

Data Representation

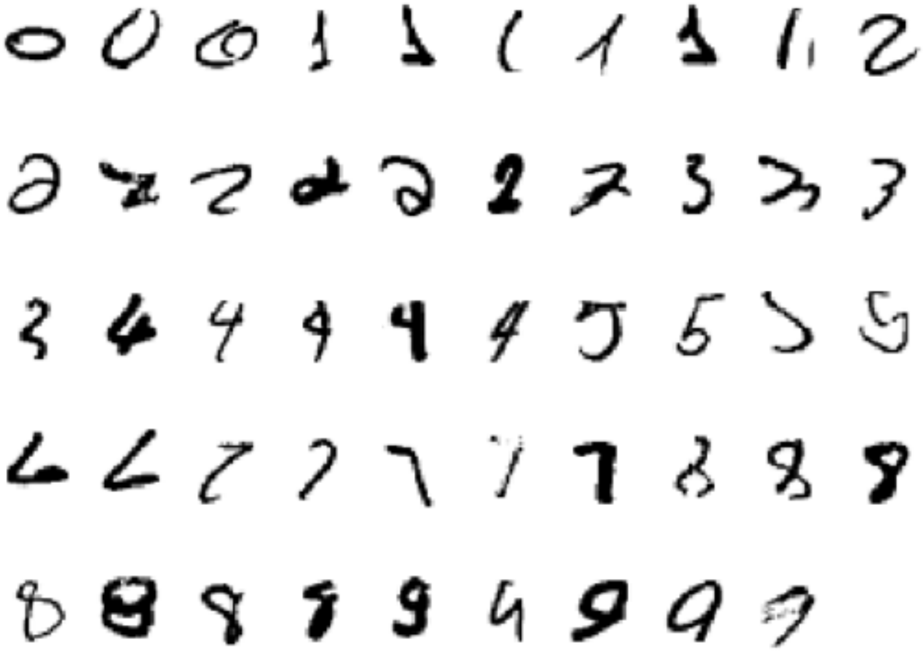


$$\approx \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .6 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .5 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .7 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .9 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Model the problem

As a supervised classification problem

Start with training data, e.g. 6000 examples of each digit



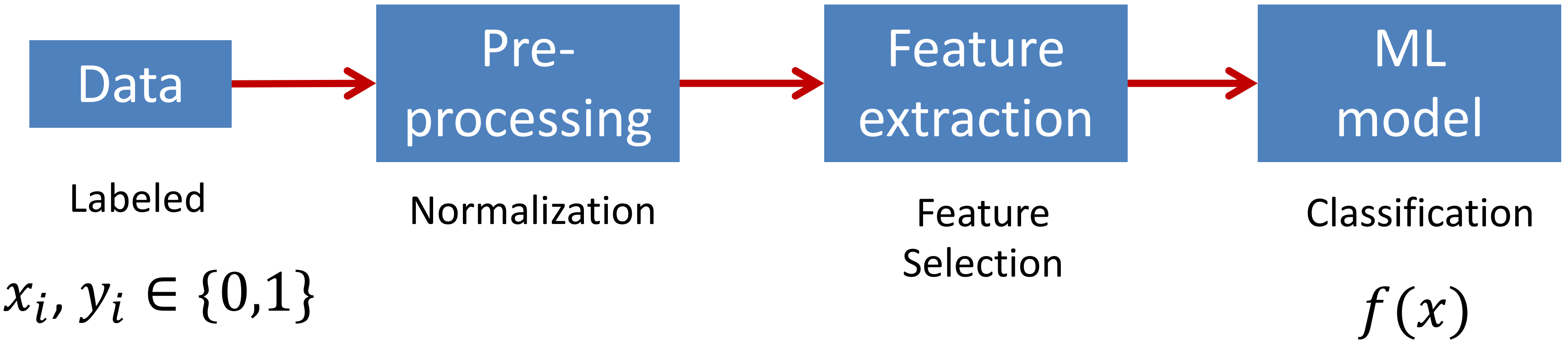
- Can achieve testing error of 0.4%
- One of first commercial and widely used ML systems (for zip codes & checks)

Other examples

- Spam classification
  - Is my email spam or not? Binary classification
  - Is the attachment safe?
- Weather prediction
  - Will it rain tomorrow or not?
- Healthcare classification
  - Is the patient sick or not?
- Image classification
  - What object does the image depict?
  - Where is the object in the image?

Supervised Learning: Classification

Training



Data → Pre-processing → Feature extraction → ML model

Labeled

$x_i, y_i \in \{0,1\}$

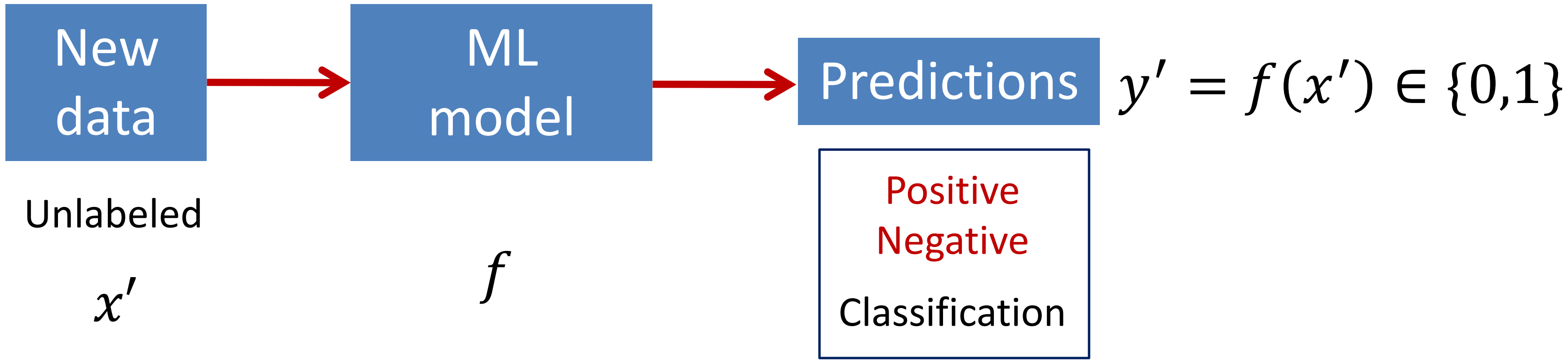
Normalization

Feature Selection

Classification

$f(x)$

Testing



New data → ML model → Predictions $y' = f(x') \in \{0,1\}$

Unlabeled

x'

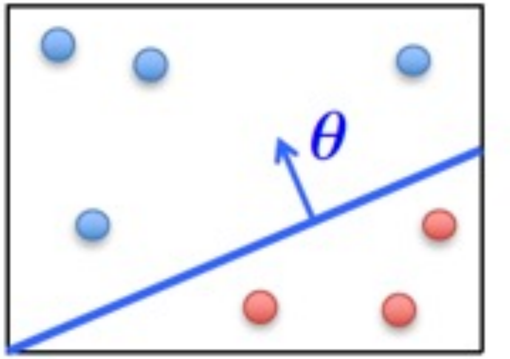
f

Positive
Negative

Classification

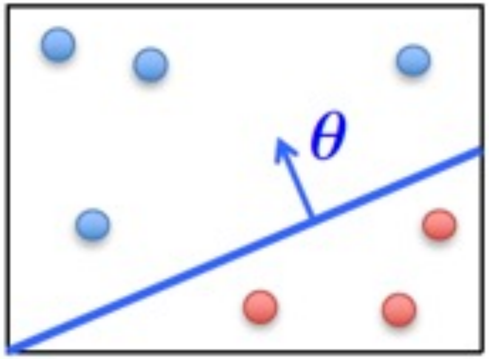
Classification

- Training data
  - $x_i = [x_{i,1}, \dots x_{i,d}]$: vector of image pixels (features)
  - Size $d = 28\text{x}28 = 784$
  - y_i: image label
- Models (hypothesis)
  - Example: Linear model (parametric model)
    - $f(x) = wx + b$
  - Classify 1 if $f(x) > \mathrm{T}$; 0 otherwise

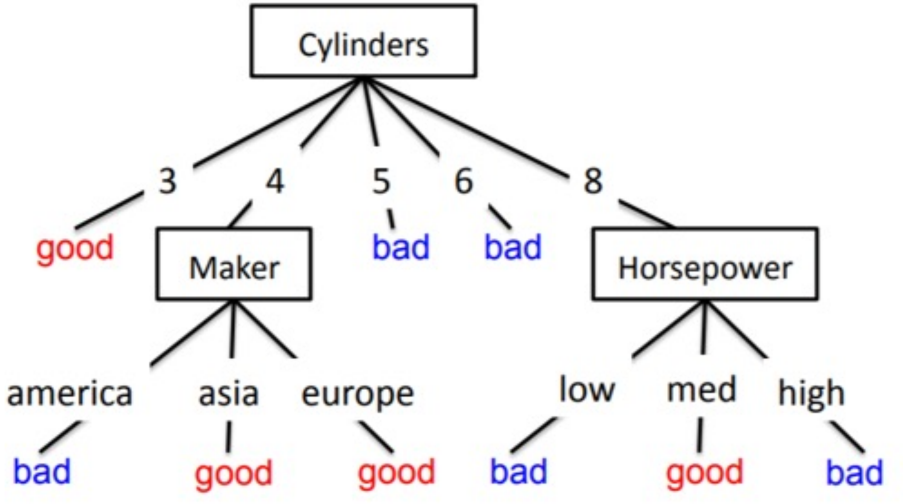


- Classification algorithm
  - Training: Learn model parameters w, b to minimize error (number of training examples for which model gives wrong label)
  - Output: “optimal” model
- Testing
  - Apply learned model to new data and generate prediction $f(x)$

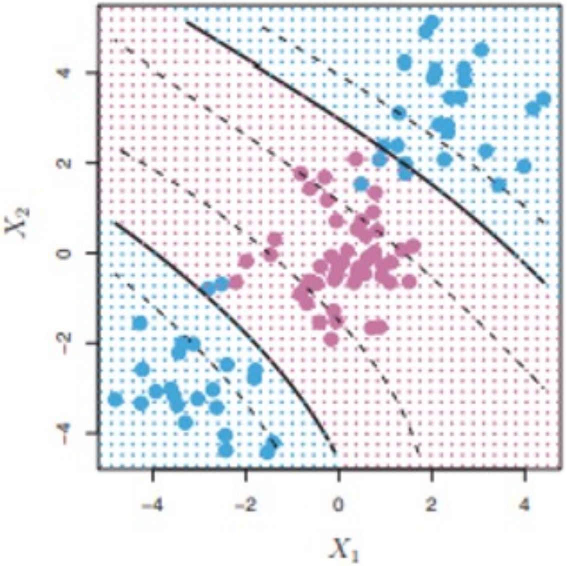
Example Classifiers



Linear classifiers: logistic regression, SVM, LDA



Decision trees



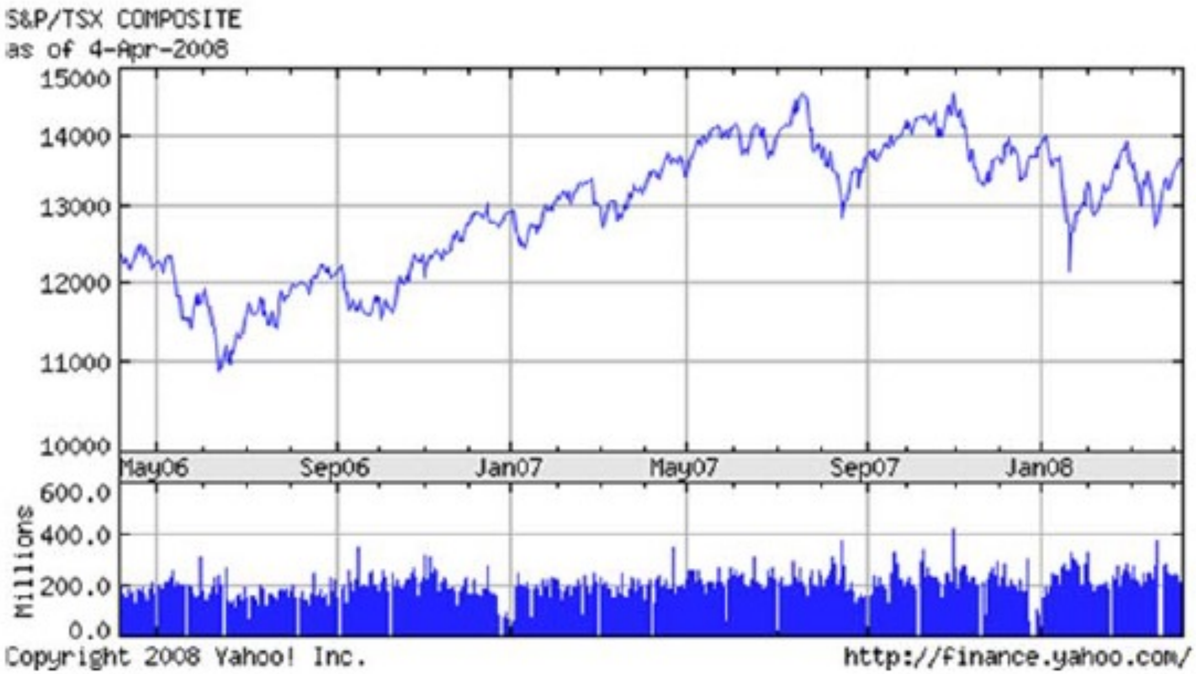
SVM polynomial kernel

Why Multiple Models?

- There is no free lunch in statistics / ML!

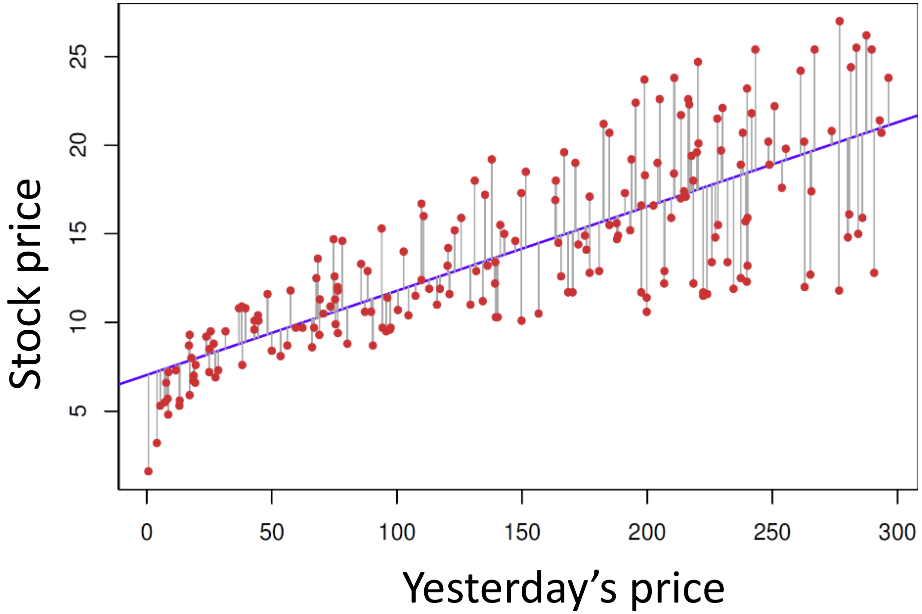
- There is no single model that dominates all
- Performance depends on many things, such as:
  - Data distribution
  - Data dimensionality
  - Quality of data and labeling

Example 2
Stock market prediction



- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

Regression



Linear regression
1 dimension

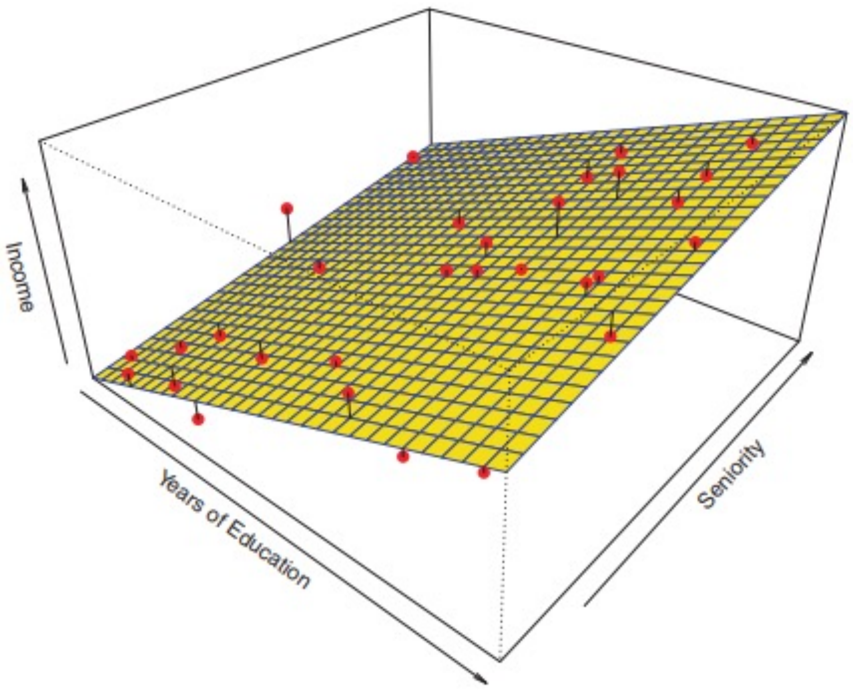
- Suppose we are given a training set of N observations

$$(x_1, \ldots, x_N) \text{ and } (y_1, \ldots, y_N)$$

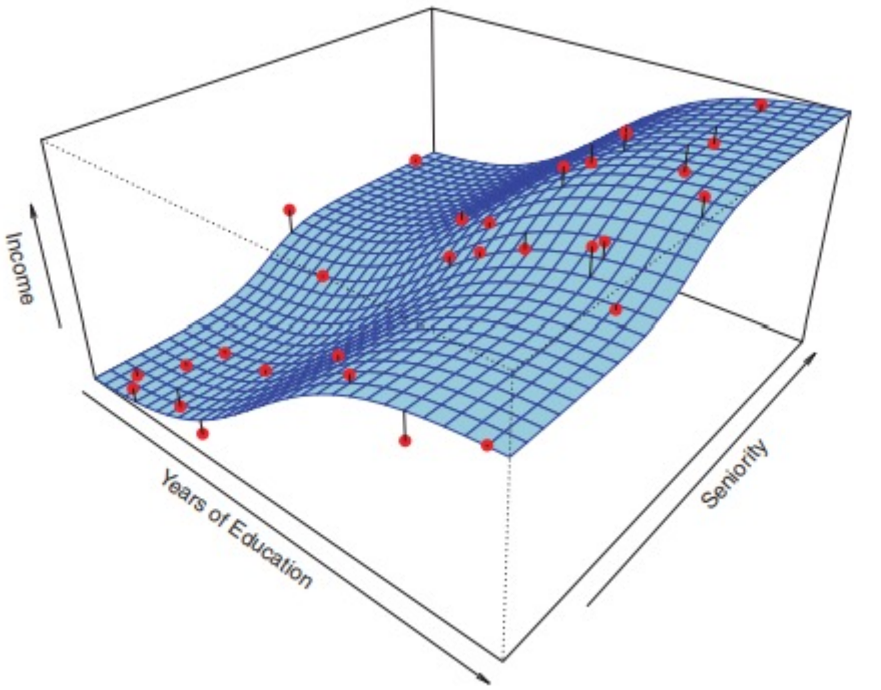
- Regression problem is to estimate y(x) from this data

$x_i = (x_{i1}, \ldots, x_{id})$ - d predictors (features)
y_i - response variable, numerical

Income Prediction



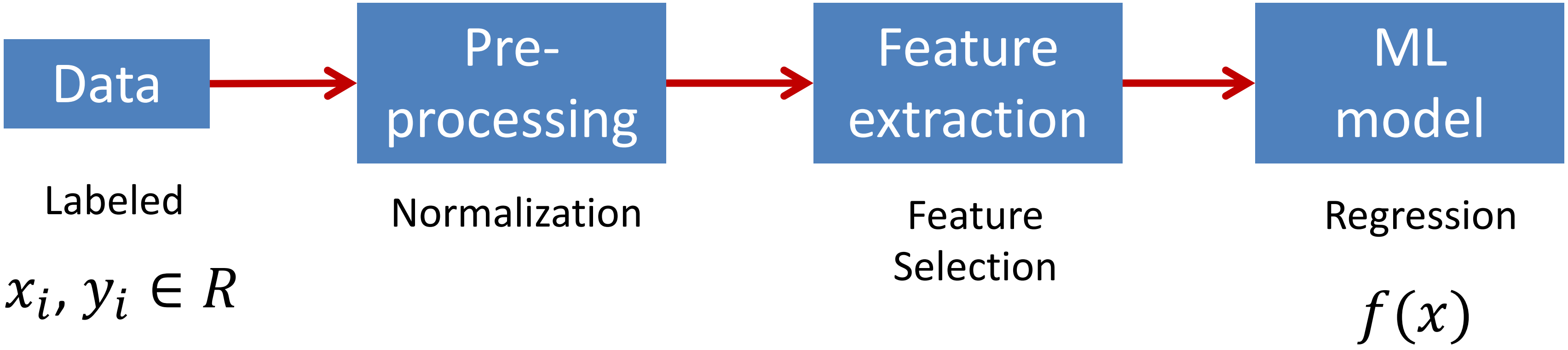
Linear Regression



Non-Linear Regression
Polynomial/Spline Regression

Supervised Learning: Regression

Training



Data → Pre-processing → Feature extraction → ML model

Labeled

$x_i, y_i \in R$

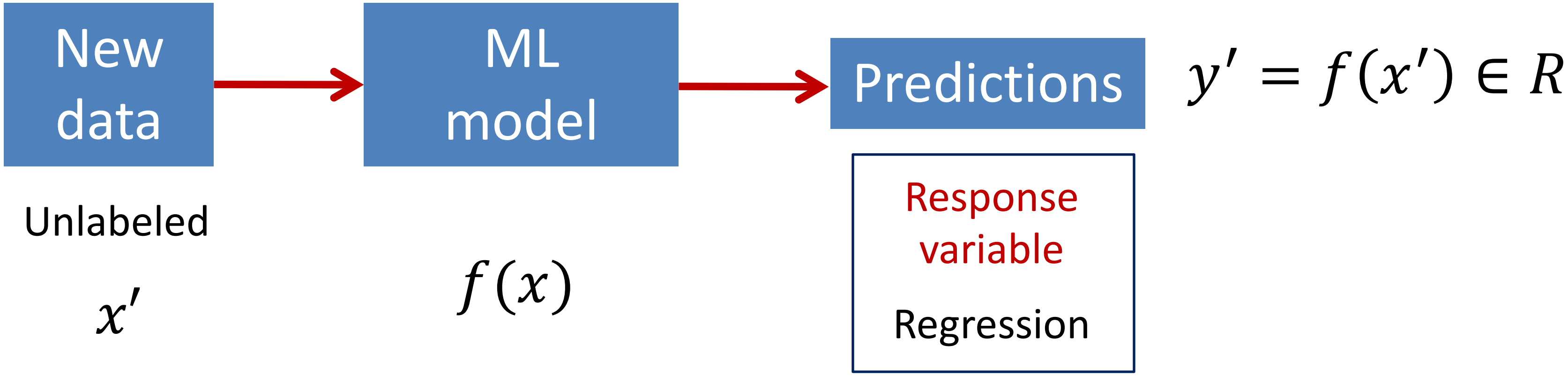
Normalization

Feature Selection

Regression

$f(x)$

Testing



New data → ML model → Predictions

$y' = f(x') \in R$

Unlabeled

x'

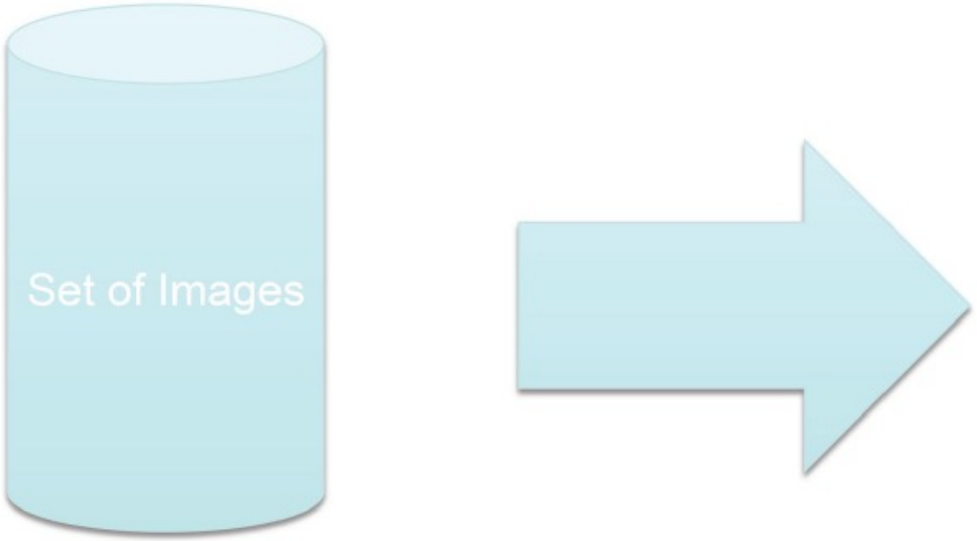
$f(x)$

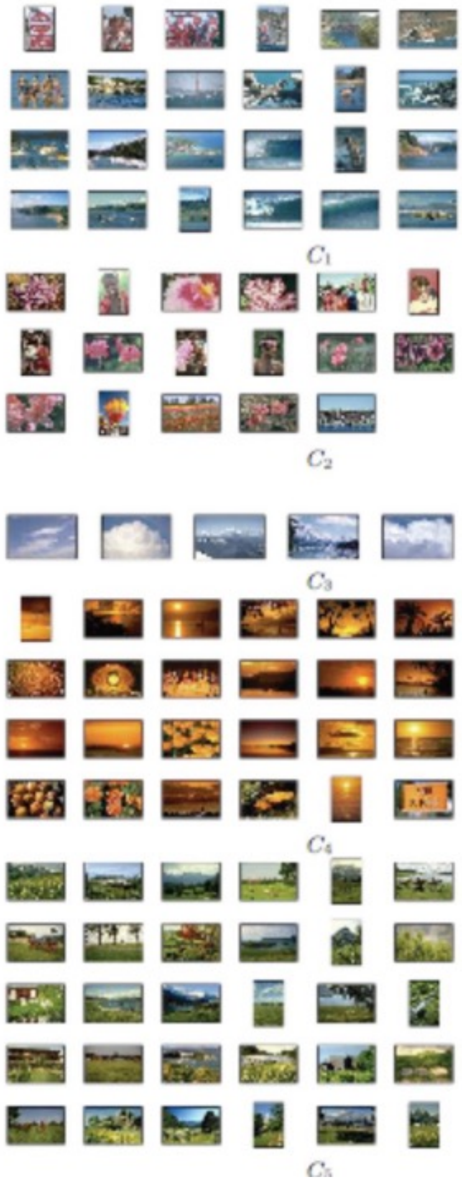
Response variable

Regression

Example 3: image search

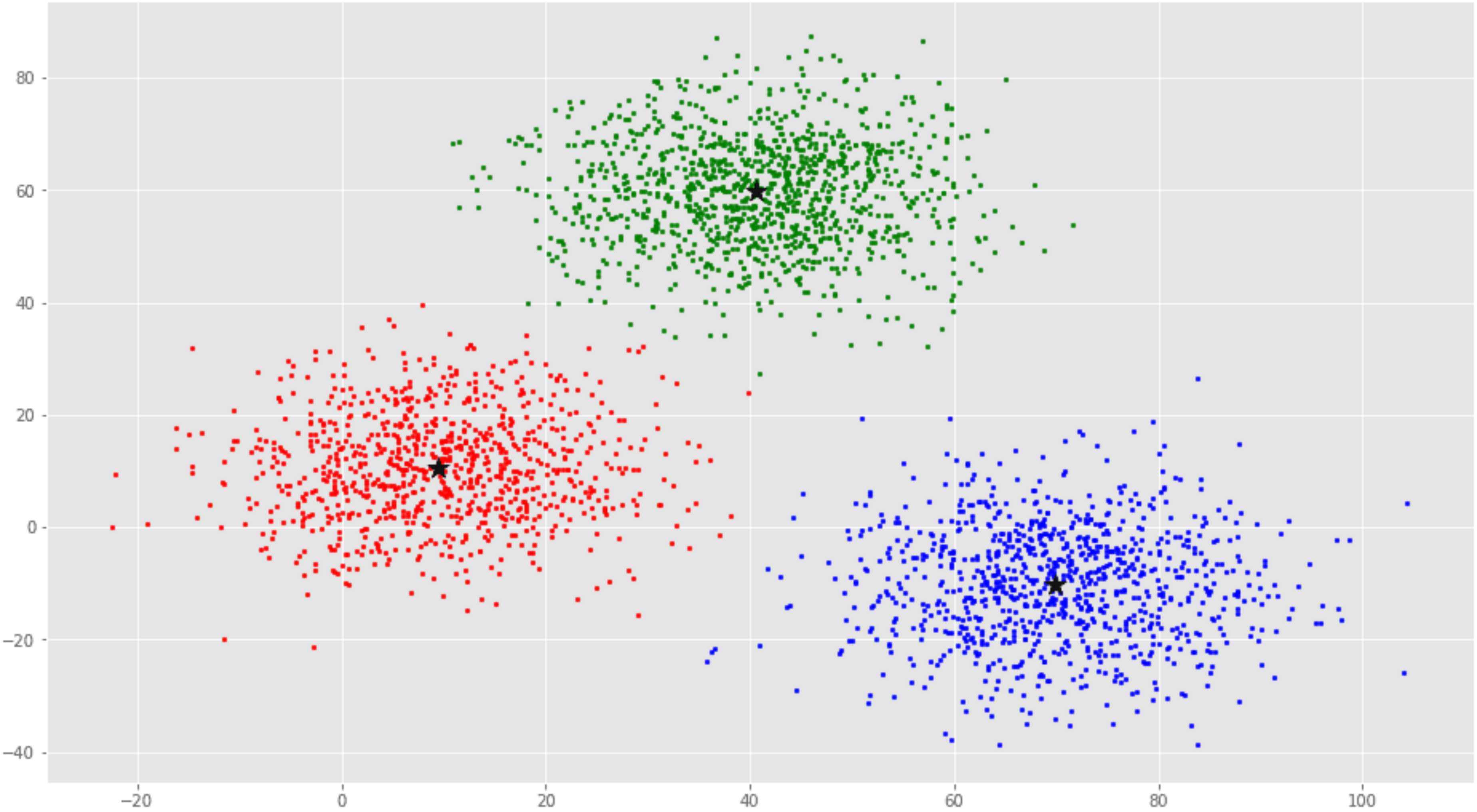
Clustering images





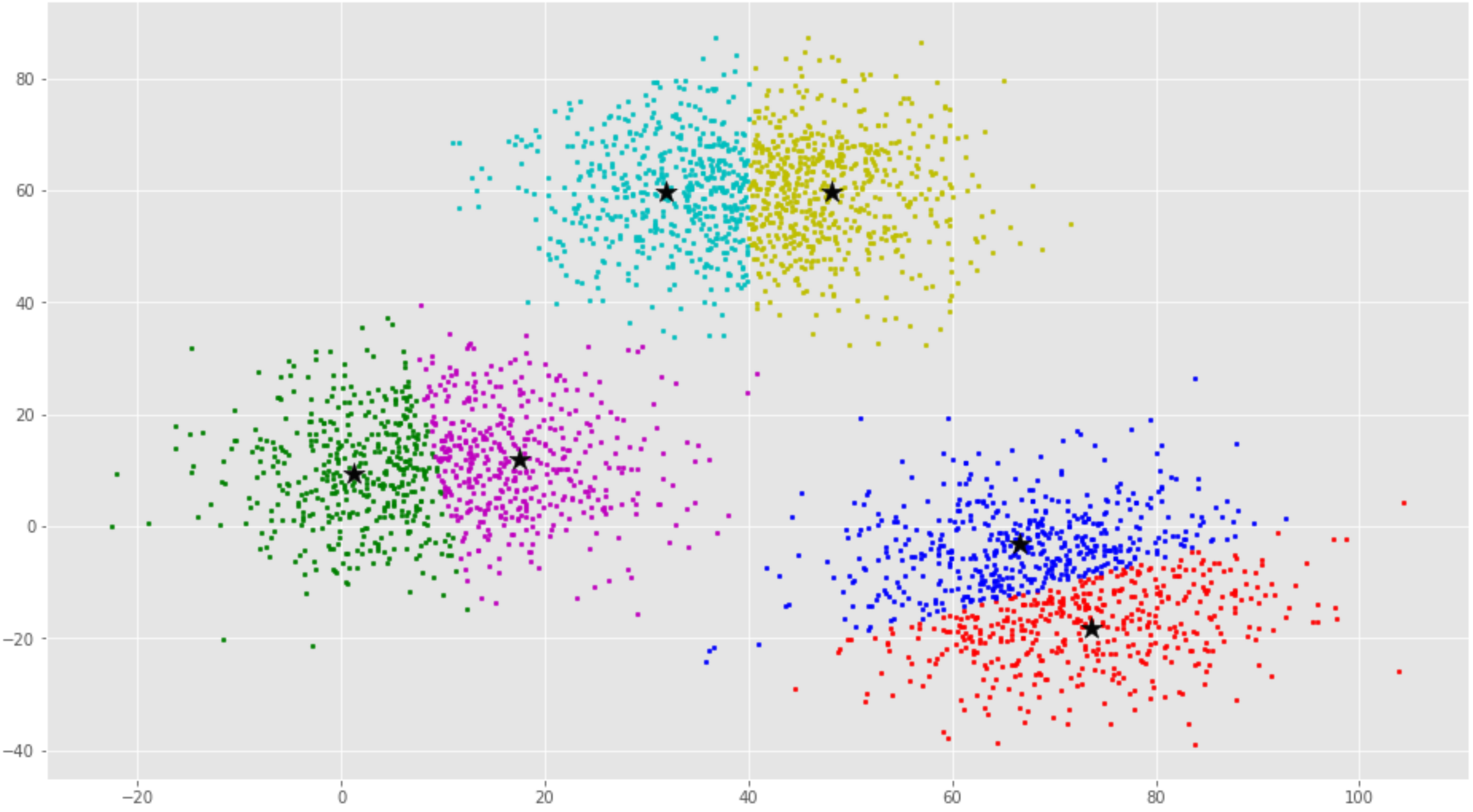
Find similar images to a target one

K-means Clustering



K=3

K-means Clustering



K=6

Unsupervised Learning

- Clustering
  - Group similar data points into clusters
  - Example: k-means, hierarchical clustering, density-based clustering
- Dimensionality reduction
  - Project the data to lower dimensional space
  - Example: PCA (Principal Component Analysis)
- Feature learning
  - Find feature representations
  - Example: Autoencoders

Supervised Learning Tasks

- Classification
  - Learn to predict class (discrete)
  - Minimize classification error 1/N $\sum_{i=1}^{N}[y_i \neq f(x_i)]$
- Regression
  - Learn to predict response variable (numerical)
  - Minimize MSE (Mean Square Error)
  - $1/N \sum_{i=1}^{N}[y_i - f(x_i)]^2$
- Both classification and regression
  - Training and testing phase
  - “Optimal” model is learned in training and applied in testing

Learning Challenges

- Goal
  - Classify well new testing data
  - Model generalizes well to new testing data
  - Minimize error (MSE or classification error) in testing
- Variance
  - Amount by which model would change if we estimated it using a different training data set
  - More complex models result in higher variance
- Bias
  - Error introduced by approximating a real-life problem by a much simpler model
  - E.g., assume linear model (linear regression), then error is high
  - More complex models result in lower bias

Bias-Variance tradeoff