University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chris Ketelsen

Lecture 1: Introduction

What is Machine Learning?

What is Machine Learning?

Seriously. What do YOU think it is?

* PATTERN DETECTION
* OPTIMIZATION
* MAKING PREDICTIONS FROM DATA
* LEARNING AUTOMATICALLY

What is Machine Learning?

Arthur Samuel (1959): Machine Learning: Field of study that gives computers the ability to learn without explicitly being programmed.

What is Machine Learning?

Arthur Samuel (1959): Machine Learning: Field of study that gives computers the ability to learn without explicitly being programmed.

What is Machine Learning?

Arthur Samuel (1959): Machine Learning: Field of study that gives computers the ability to learn without explicitly being programmed.

Tom Mitchell (1998): Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

What is Machine Learning?

Arthur Samuel (1959): Machine Learning: Field of study that gives computers the ability to learn without explicitly being programmed.

Tom Mitchell (1998): Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

What is E?

What is Machine Learning?

Arthur Samuel (1959): Machine Learning: Field of study that gives computers the ability to learn without explicitly being programmed.

Tom Mitchell (1998): Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

What is E?

E is the Data.

What is Machine Learning?

Data are Everywhere

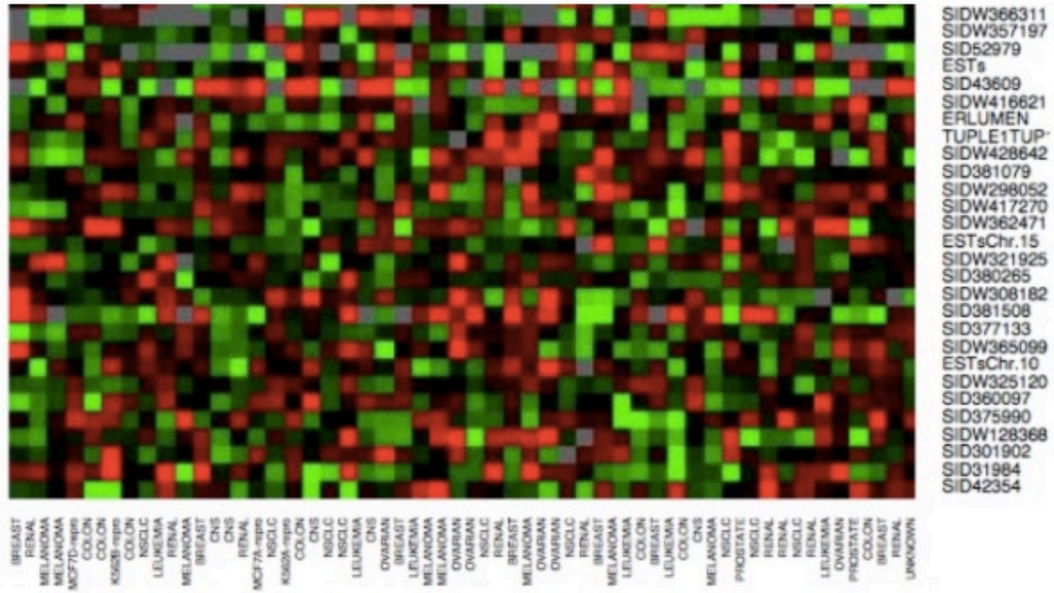
User Ratings

Silence of the Lambs	★★★★★
The Breakfast Club	★★★★☆
X-Men	★★★★☆
Jurassic Park III	★☆☆☆☆
Men of Honor	★★★★☆
The Thin Red Line	★☆☆☆☆
Best in Show	★★★★☆
Gone Baby Gone	★★★☆☆
Eastern Promises	★★★★☆
Independence Day	★★★★☆
Star Wars: Episode V: The Empire Strikes Back	★★★★★
Clear and Present Danger	★★★★☆
Star Trek: Nemesis	★★★★☆
Resident Evil	★★★☆☆

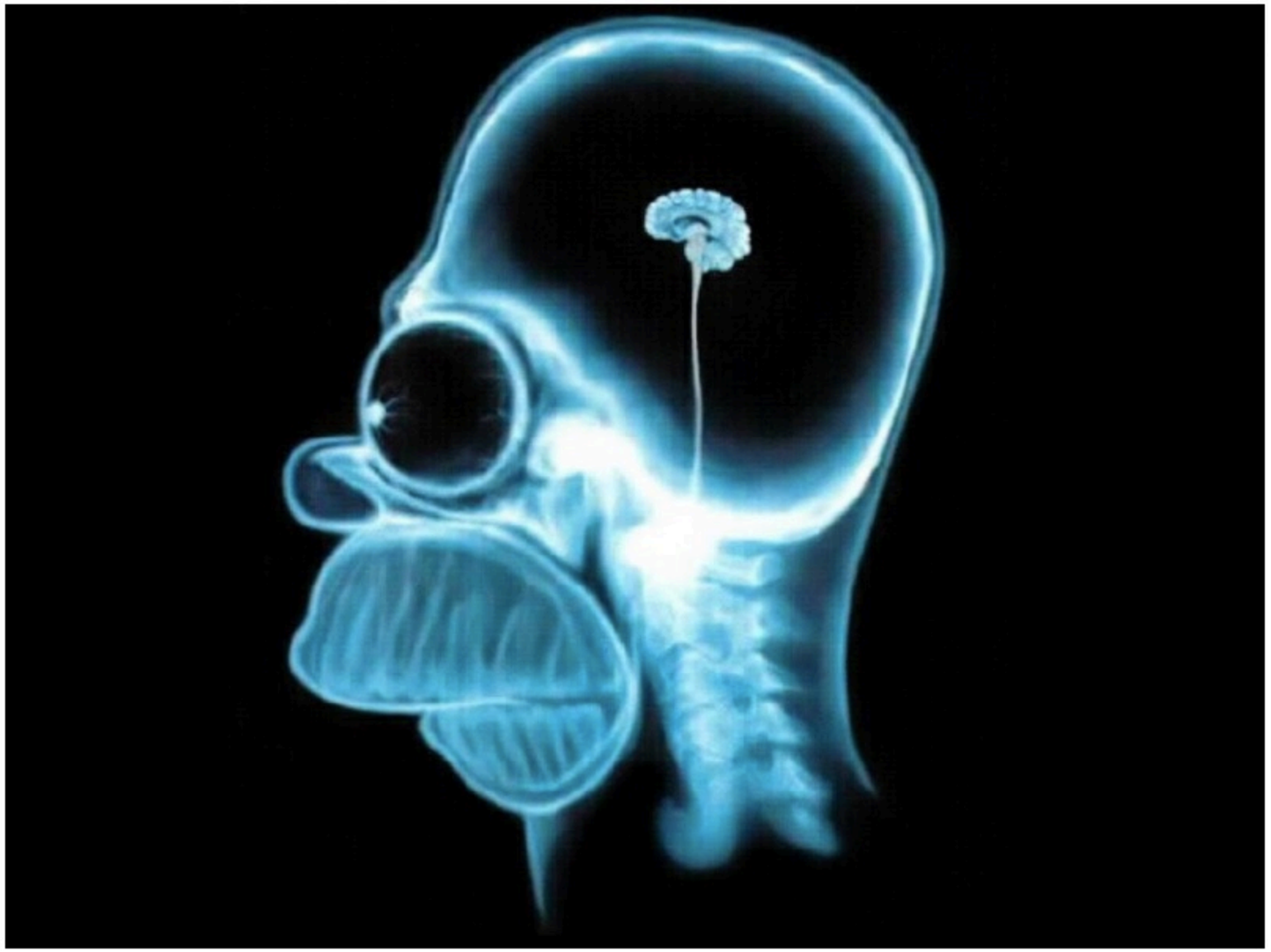
Document Collections

Social Media

The Human Genome



The Human Brain



Video

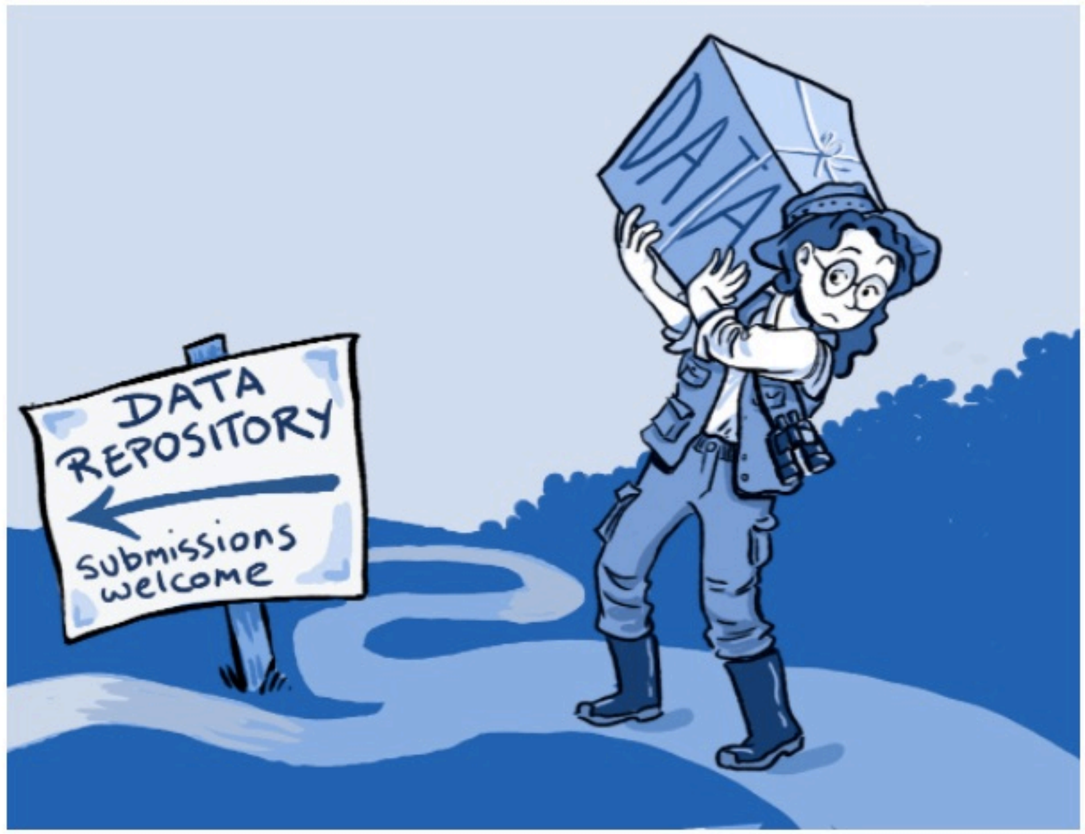
Maybe you want to use YouTube data to determine how a viral campaign affects awareness about a cause

Video

Or maybe you just want to know if a particular frame contains a cat

The Data is Out There

Today, you can find pretty much any data you want



The hard part is deciding what to do with it

Mathematical Foundation

$$\mathbf{X} \rightarrow \mathbf{Z} \rightarrow \mathbf{Y}$$

Data — *Hidden Structure* — *Answer*

- $\mathbf{X} = (X_1, X_2, \ldots, X_D)$. X_i are called **features**
- $\mathbf{Y}$ is called the **response**

- Sometimes interested in $\mathbf{Z}$, sometimes not

How Much is My House Worth?

Given: information about recently sold houses in my city

Predict: how much my house will sell for

How Much is My House Worth?

Given: information about recently sold houses in my city

Sq. Ft.	Sale Price (in thousands of $)
1100	200
2450	658
1348	305
745	255
3556	800
.	.
.	.
.	.
455	220

Predict: how much my house will sell for

How Much is My House Worth?

What is $\mathbf{X}_i$:

AREA

COLOR

LOCATION

#BRs

BATHROOMS

What is Y:

SALE PRICE

Is this Spam?

From: mrichards@therange.com
SUBJECT: $1,000 for sending an e-mail
Date: Thu, 07 Nov 2002 20:59:28 PM MST
Reply-To: rem_homosan@hotmail.com
To: christian.ketelsen@colorado.edu

We will give you $1,000 for sending an e-mail to your friends. AB Mailing, Inc. is proud to anounce the start of a new contest. Each day until January, 31 1999, one lucky Internet or AOL user who forwards our advertisement to their friends will be randomly picked to receive $1,000! You could be the winner!

Thank you for your time.

What about this?

From: [Snipped]
SUBJECT: easy camping trip?
Date: July 11, 2016 10:43:12 AM MST
To: ketelsen@colorado.edu, murray.cox@colorado.edu

Either of you down for an easy camp trip? (ie. car camping by a lake, drinking beer, and chilling?)

http://www.protrails.com/trail/398/summit-county-eagle-county-clear-creek-county-crystal-lakes

-chris

And this?

SUBJECT: Mailbox Owner
Date: July 4, 2016 10:22:25 AM MST
To: ketelsen@colorado.edu
Reply-To: euphonynet.be < upgradeteam@outlook.com >

Dear Mailbox Owner,

You have exceeded the storage limit on you mailbox. You will not be able to send receive new email until you upgrade your email quota. Advice click on the link and fill the form to upgrade your account

Admin Support
University of Colorado Boulder

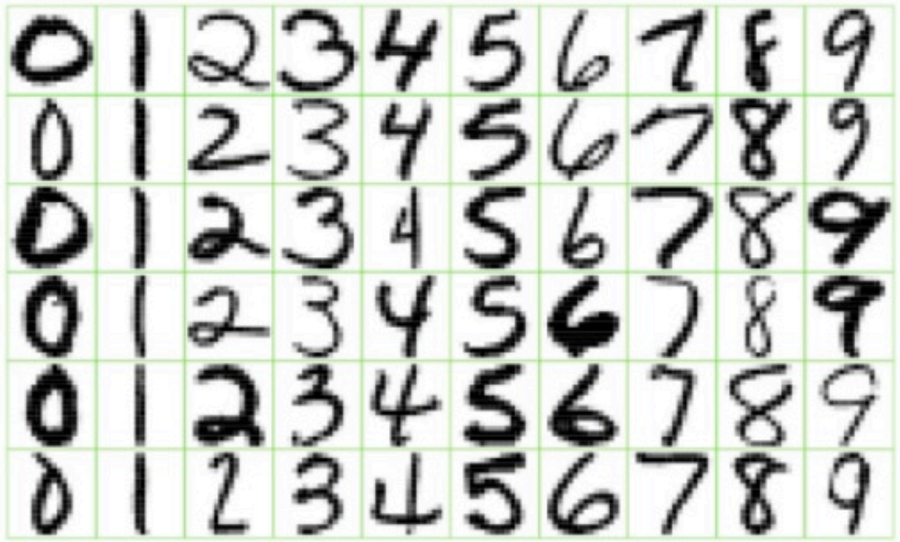
Is this Spam or Ham?

What is $\mathbf{X}_i$: WORDS

What is Y: SPAM OR HAM

Handwritten Digit Recognition

Given a bunch labeled images of numbers



Can we correctly identify the number in a new image?

Handwritten Digit Recognition

What is $\mathbf{X}_i$: pixel values in image

What is Y: $\{0, 1, 2, \cdots, 9\}$

Supervised Learning

Find patterns in **fully observed** data and then try to predict from partially observed data.

Supervised Learning - Find $f : \mathbf{X} \to \mathbf{Y}$

- **Given**: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ training examples
- **Find**: A good approximation to $f : \mathbf{X} \to Y$

$\hat{f} : X \to \hat{Y}$

Examples:

- House Market Value Prediction
  - Map (square footage, #BRs) to maket value
- Spam Detection
  - Map words in email to {Spam, Ham}
- Digit Recognition
  - Map image pixels to {0,1,2,3,4,5,6,7,8,9}

Regression vs. Classification

What type of thing am I trying to predict?

If $y \in \mathbb{R}$ we usually call this **regression**:

$$y(\mathbf{x}) = f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \epsilon \quad \Rightarrow \quad \hat{y} = \hat{f}(\mathbf{x}) = \hat{\mathbf{w}}^T\mathbf{x}$$

Example: Housing Prices

If $y \in \{1, 2, \ldots, C\}$ we usually call this **classification**:

$$\hat{y} = \hat{f}(\mathbf{x}) = \underset{c=1,\ldots,C}{\operatorname{argmax}} \; p(y = c \mid \mathbf{x}, D)$$

Example: Spam vs. Ham, Digit Recognizer

$P(y = \text{HAM} \mid \text{EMAIL}, D)$
$P(y = \text{SPAM} \mid \text{EMAIL}, D)$

Unsupervised Learning

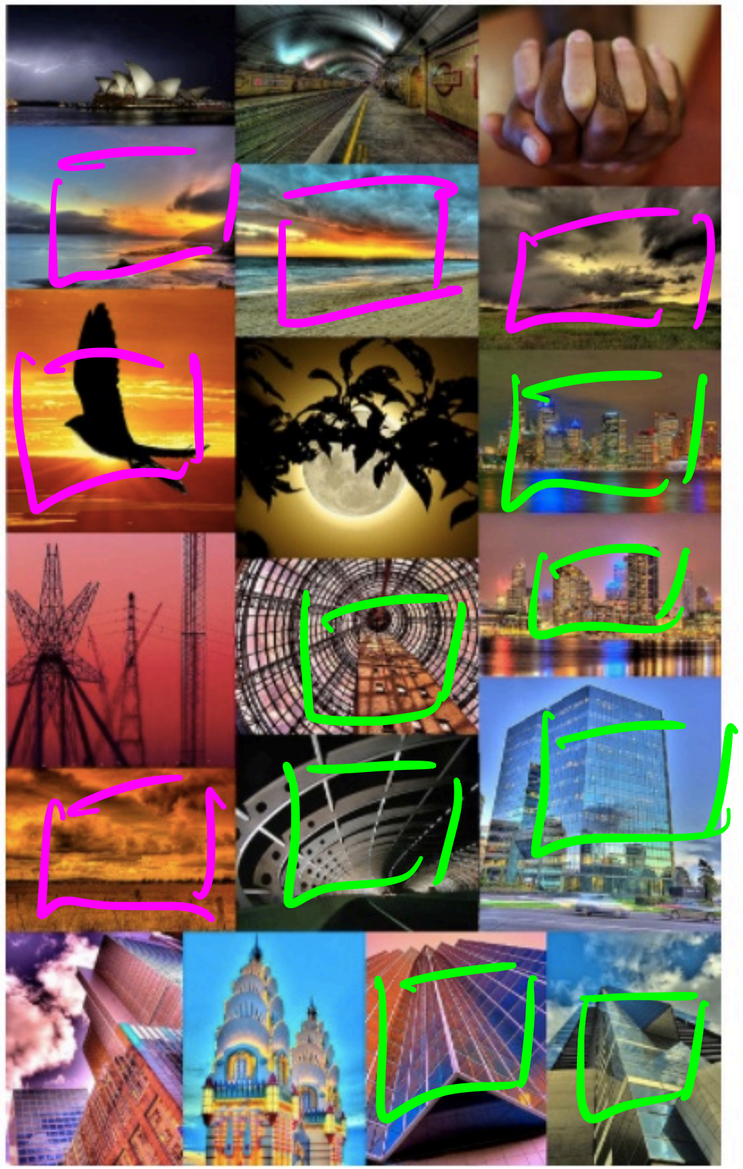
Find **hidden structure** in data, structure that we can never formally fully observe.

Data is simply $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{m}$

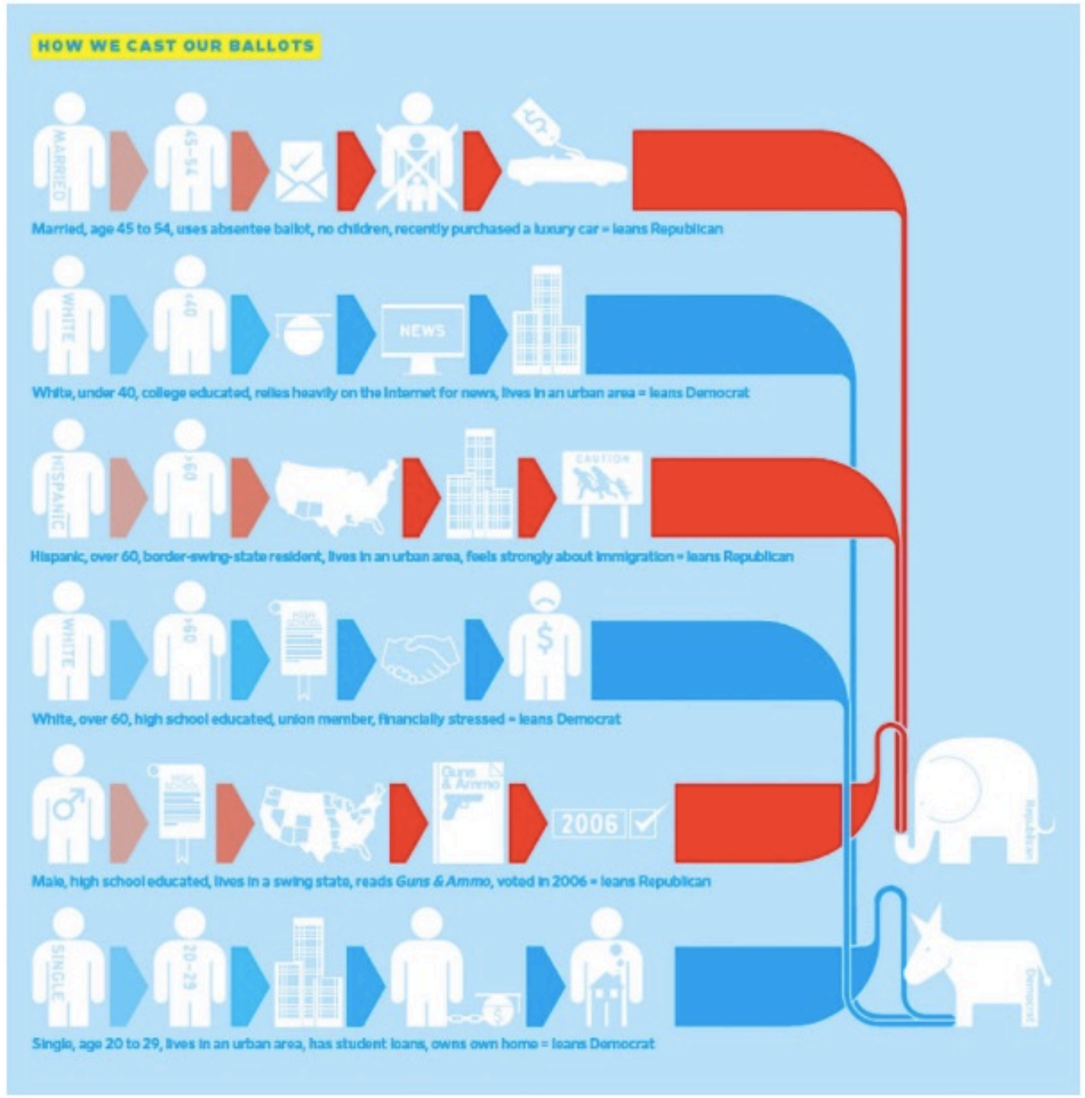
Try to get at $\mathbf{Z}$

Big Ideas: Clustering, Dimensionality Reduction

Group Images Together



Who Will Vote and for Whom?



A Useful Grouping

	discrete	*continuous*
supervised	*classification*	*regression*
unsupervised	*clustering*	*dimensionality reduction*

A Useful Grouping

	discrete	*continuous*
supervised	**classification**	*regression*
unsupervised	*clustering*	*dimensionality reduction*

Methods: KNN, Naive Bayes, Logistic Regression, SVMs

A Useful Grouping

	discrete	*continuous*
supervised	*classification*	**regression**
unsupervised	*clustering*	*dimensionality reduction*

Methods: Linear Regression, Lasso & Ridge Regression

A Useful Grouping

	discrete	*continuous*
supervised	*classification*	*regression*
unsupervised	**clustering**	*dimensionality reduction*

Methods: K-Means, Gaussian Mixture Models

A Useful Grouping

	discrete	*continuous*
supervised	*classification*	*regression*
unsupervised	*clustering*	**dimensionality reduction**

Methods: Principal Component Analysis, Kernel PCA

What's Missing?

Reinforcement Learning

- Teach a helipcopter to fly
- Teach a robot to ping-pong
- **Teach Mario to run on his own ...**

The Plan

- **Goal**: Fluency in thinking about ML problems
- We will learn about a suite of tools in modern data analysis
    - When to use them
    - What assumptions they make
    - Their capabilities and their limitations
    - A few theoretical guarantees
- After completing this course, you should be able to learn a new tool, apply it to data, and understand the meaning the result

The Plan

We'll cover (not necessarily in this order):

- Probabilistic Foundations
- Supervised Learning (lots of this)
- Unsupervised Learning (less of this)
- Methods that operate on discrete data (lots of this)
- Methods that operate on continuous data (less of this)
- Representing Data / Feature Engineering
- How to evaluate models
- Understanding the assumptions that our models make

OK, Time to Work!

Your first **Supervised Classification** algorithm...

K-Nearest Neighbor Classification

Example: Suppose your company releases a new product and you want to analyze the public response by looking at Tweets that mention the product.

Goal: You want to classify Tweets as **Positive** or **Negative**

Training Data: Suppose you made an intern go through N Tweets and label them as **Positive** or **Negative**. Giving you $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$

Prediction: Given a new Tweet, $\mathbf{x}$, how do you determine its sentiment?

Idea: Find the K Tweets in $\mathcal{D}$ that are *most similar* to $\mathbf{x}$ and assign $\mathbf{x}$ the label held by the majority.

K-Nearest Neighbor Classification

- Have $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ training examples
- $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{1, \ldots, C\}$
- Want to predict the class $\hat{y}$ for test sample $\mathbf{x}$

KNN:

1. Find $N_K(\mathbf{x}, \mathcal{D})$: the K training examples in $\mathcal{D}$ "nearest" to $\mathbf{x}$
2. Assign $\hat{y}$ the *majority* label of N_K

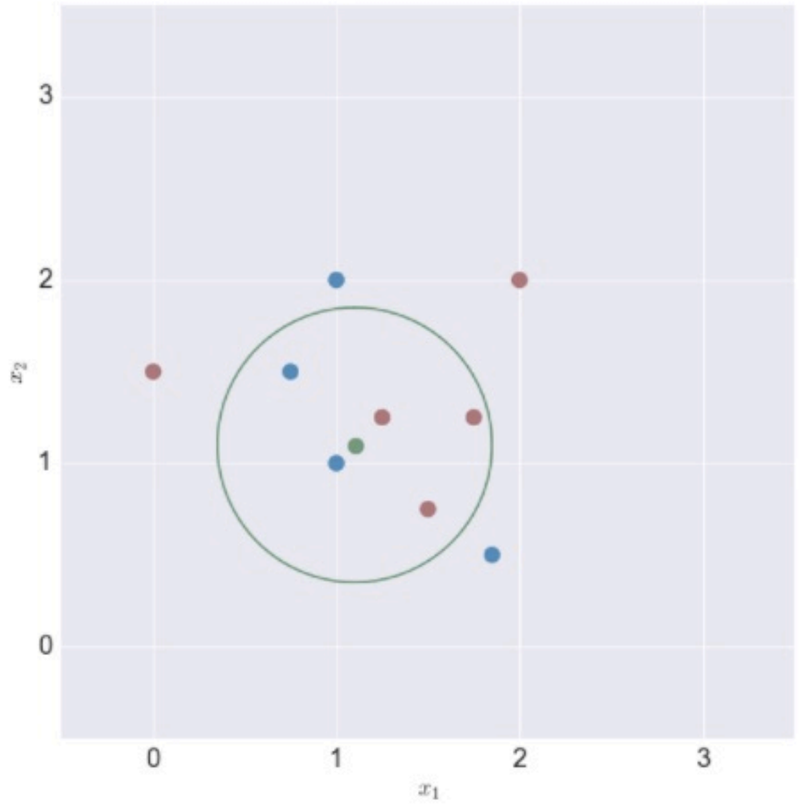
Questions:

- What does "nearest" mean?
- What do we do in the case of a tie?

K-Nearest Neighbor Classification

Questions:

- What does "nearest" mean?
- What do we do in the case of a tie?

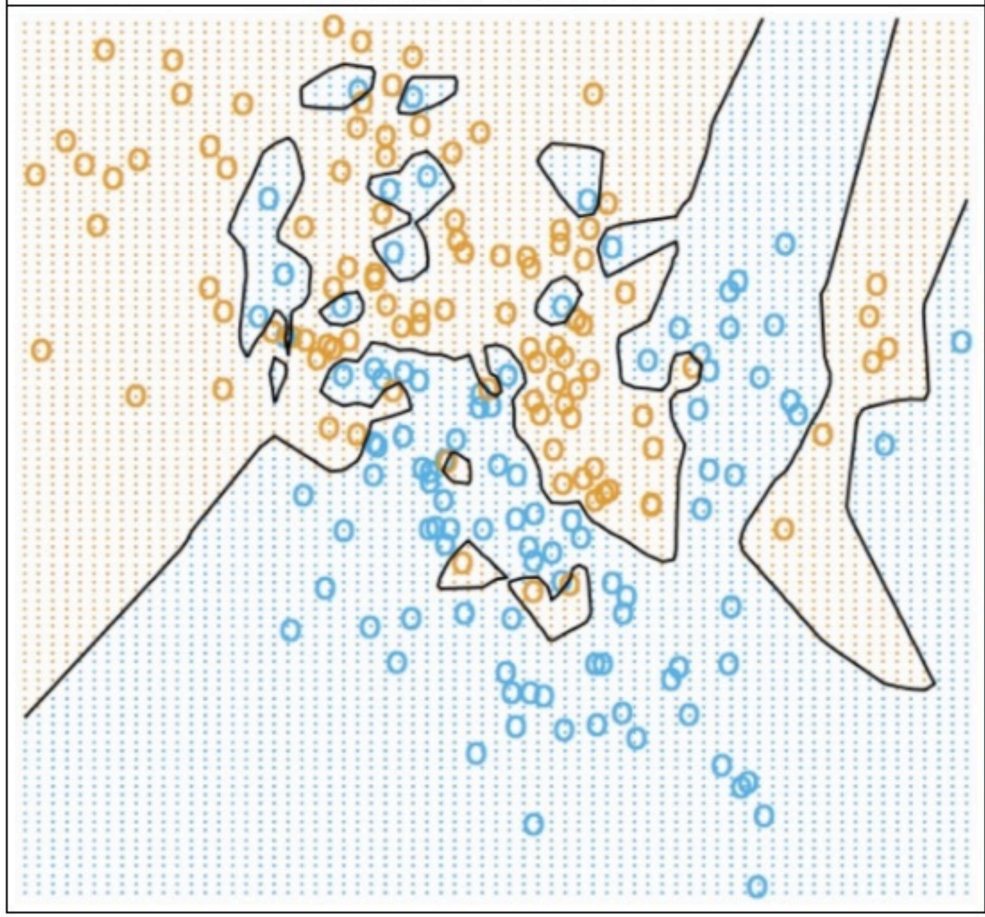
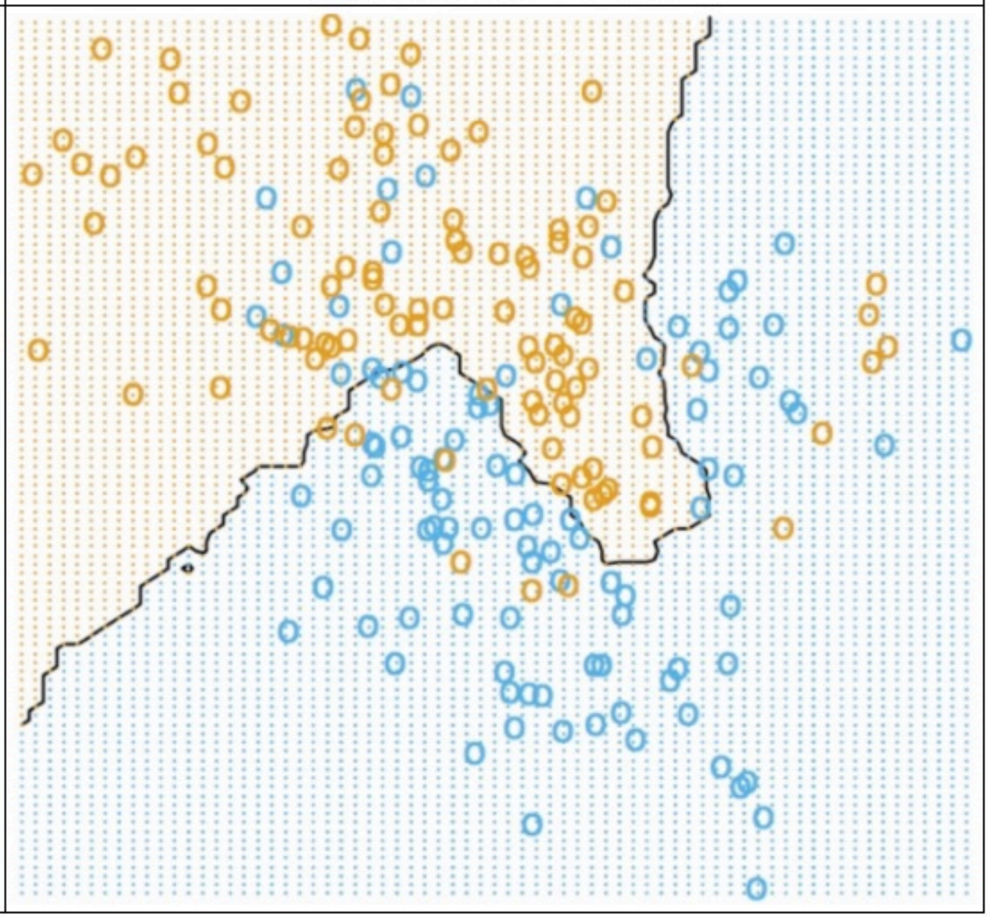


Euclidean Distance: $\|\mathbf{x}_i - \mathbf{x}\|^2$

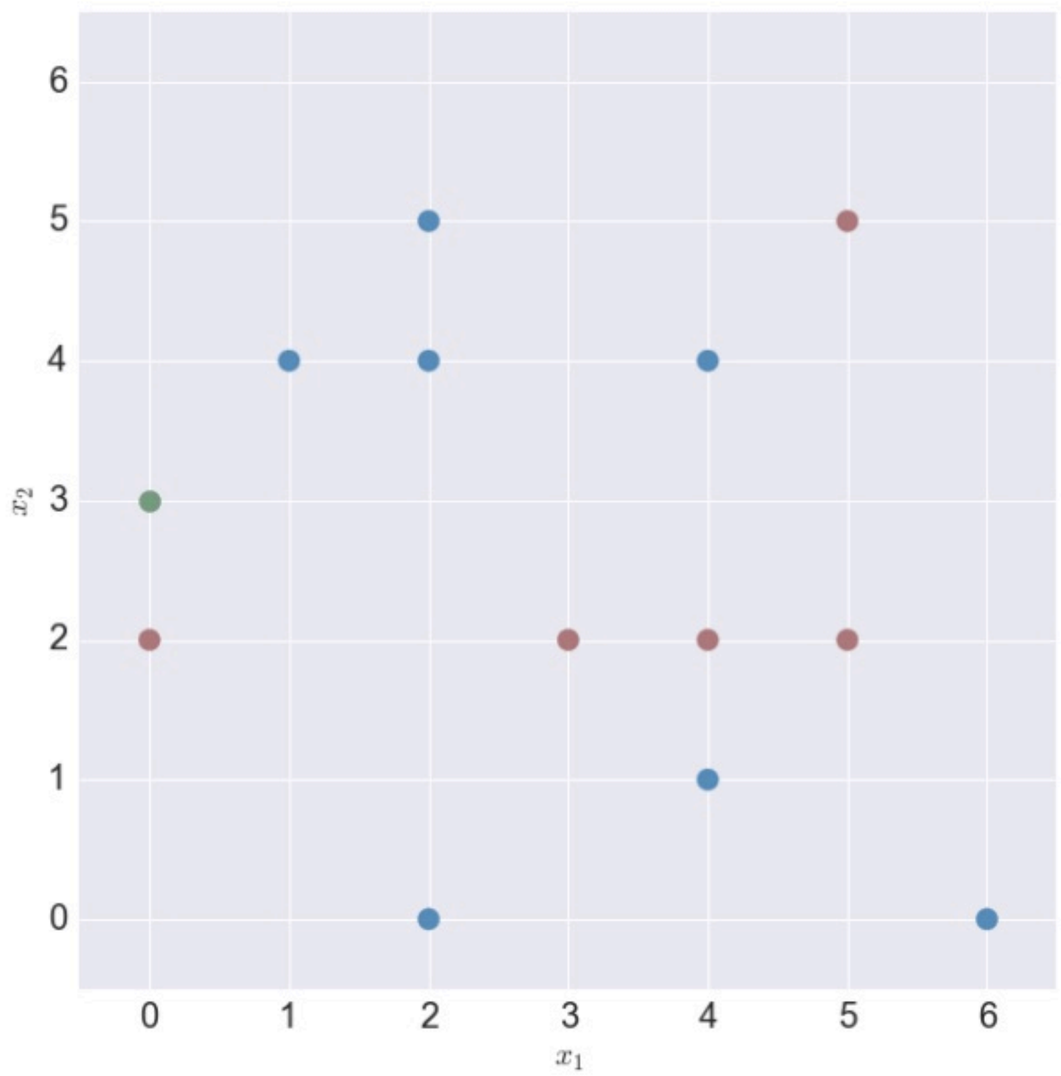
- 1-NN predicts Blue
- 2-NN prediction unclear
- 5-NN predicts Red

K-Nearest Neighbor Classification

KNN on a simulated data set

KNN: K = 1	KNN: K = 15
	

KNN Example

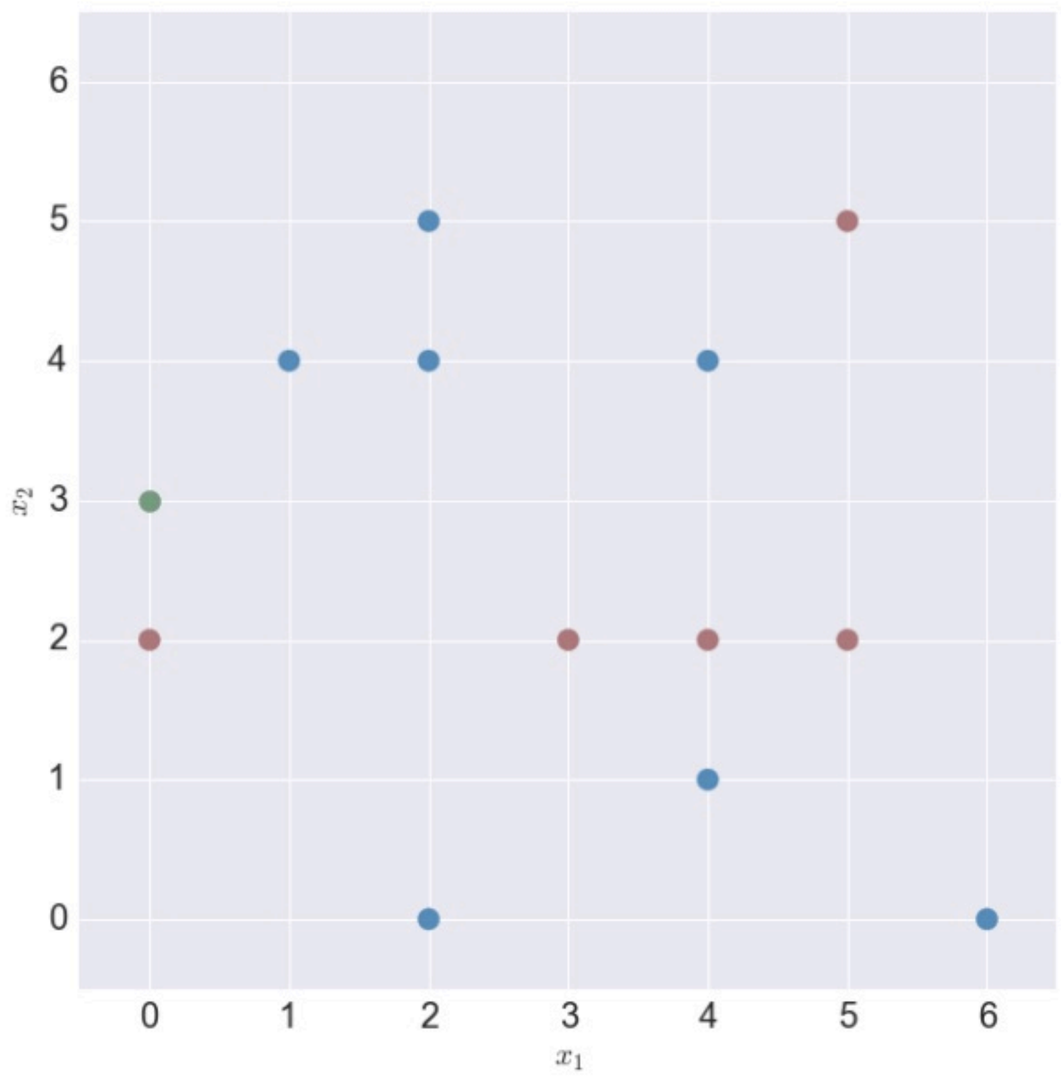


Predict $\hat{\mathbf{y}}$ with $K = 1$

Closest Points: (0,2)

Prediction: $\hat{y}$ = RED

KNN Example



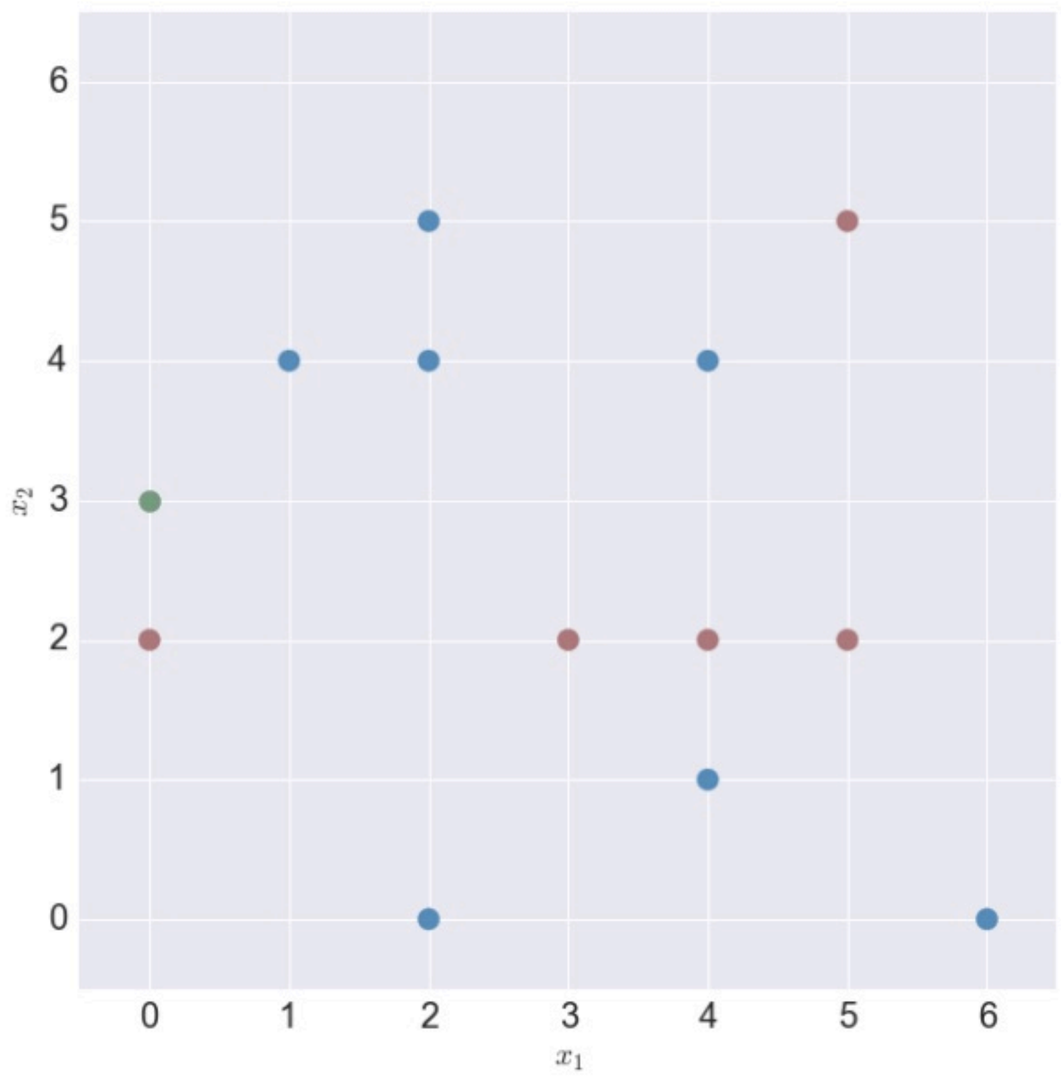
Predict $\hat{\mathbf{y}}$ with $K = 2$

Closest Points:

(0,2) (1,4)

Prediction: UNCLEAR

KNN Example



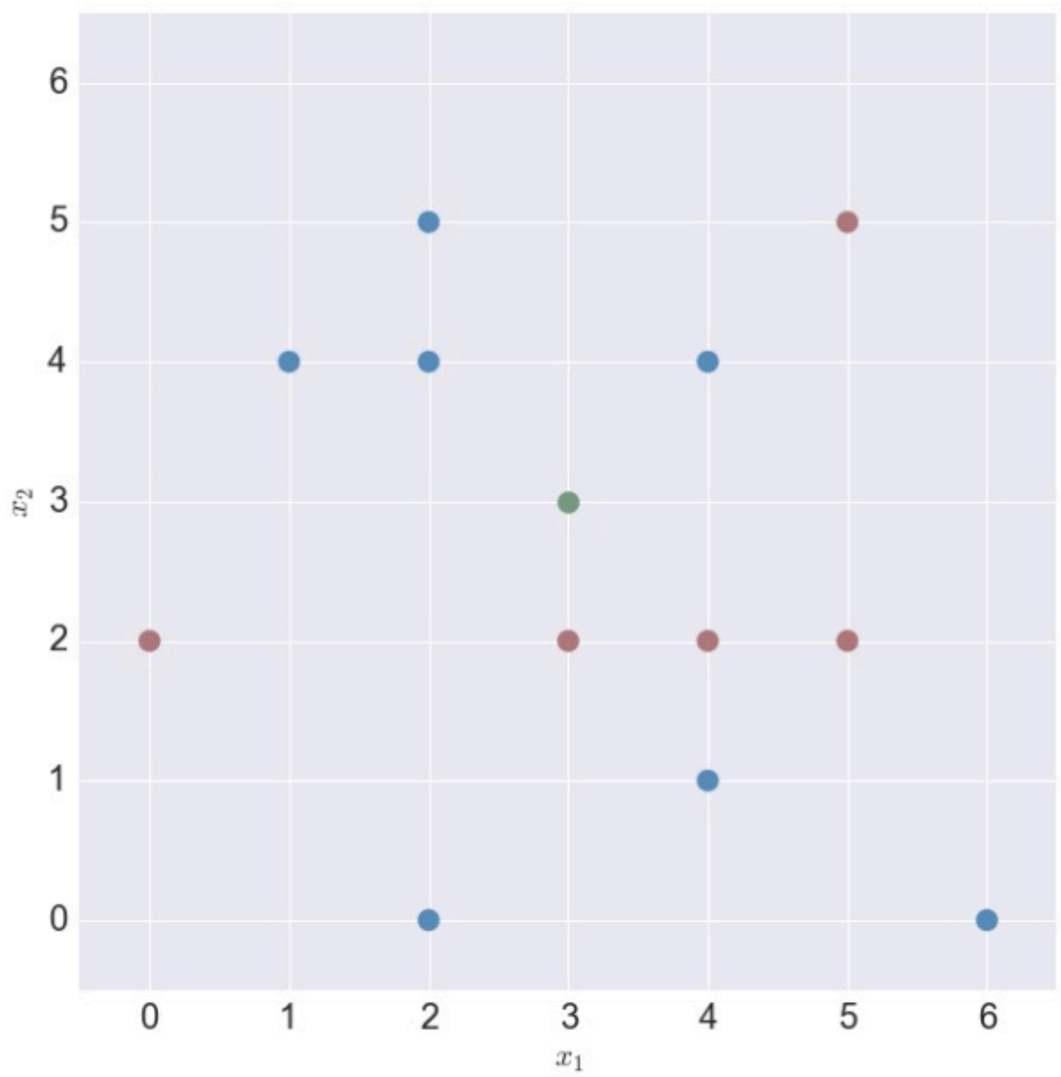
Predict $\hat{\mathbf{y}}$ with $K = 3$

Closest Points:

(0,2) (1,4) (2,4)

Prediction: $\hat{y}$ = BLUE

KNN Example



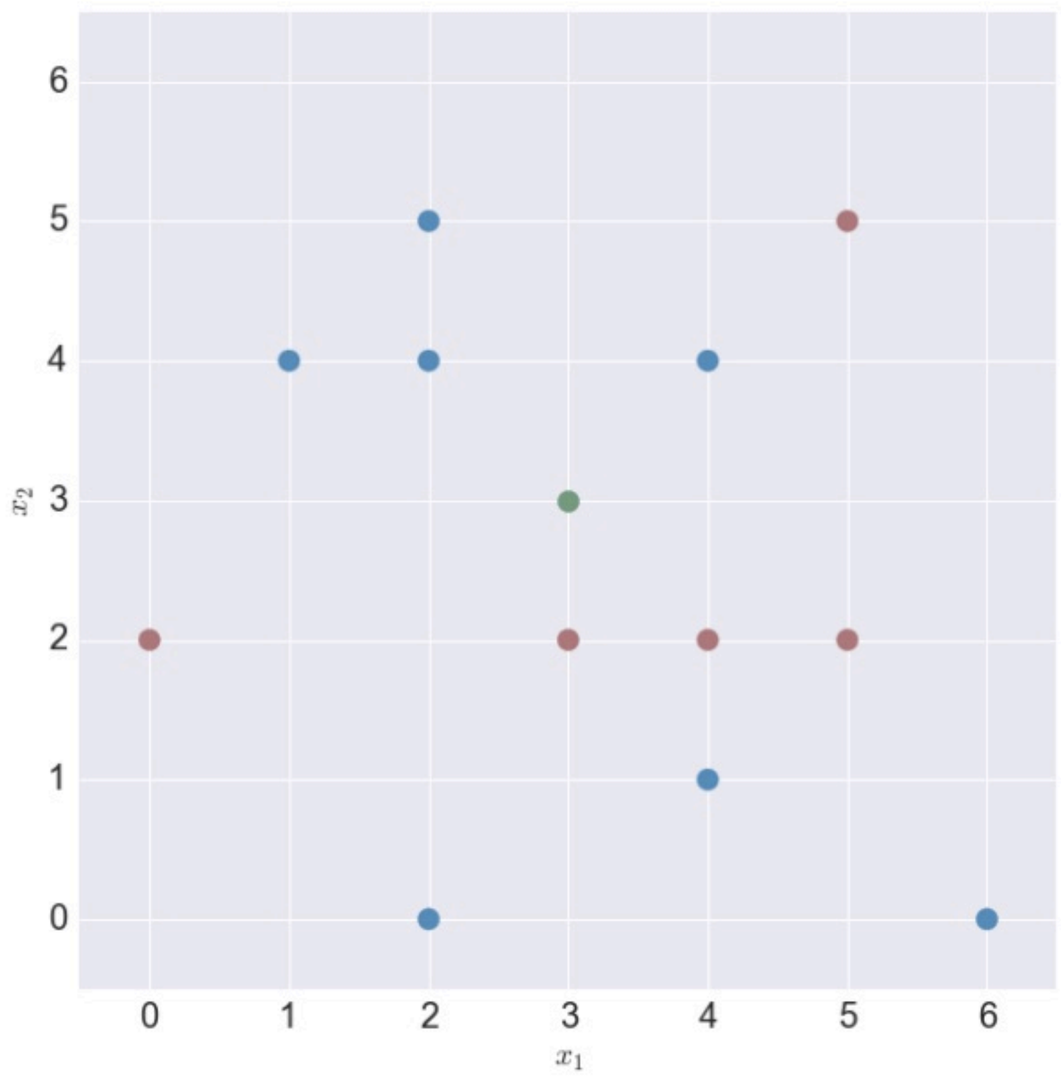
Predict $\hat{\mathbf{y}}$ with $K = 1$

Closest Points:

(3, 2)

Prediction: $\hat{y}$ = RED

KNN Example



Predict $\hat{\mathbf{y}}$ with $K = 2$

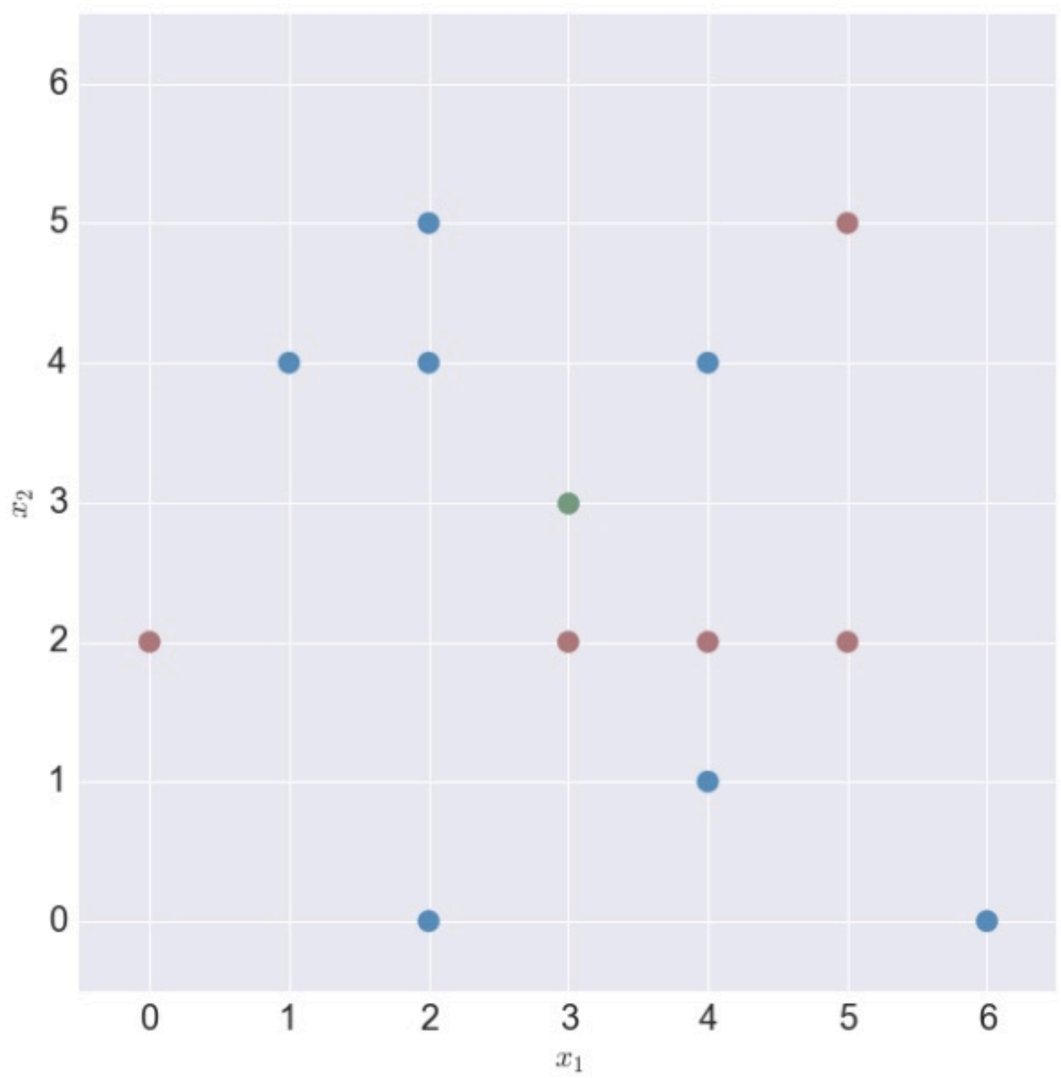
Closest Points:

(3,2) (2,4)

(4,2) (4,4)

Prediction: Unclear

KNN Example



Predict $\hat{\mathbf{y}}$ with $K = 3$

Closest Points:

(3,2) (2,4)

(4,2) (4,4)

Prediction: UNCLEAR

K-Nearest Neighbor Classification

How much does KNN cost?

A single query with m training examples, each with D features

Naive KNN:

- $\mathcal{O}(Dm)$ in time
- $\mathcal{O}(Dm)$ in space

Tree-Based KNN:

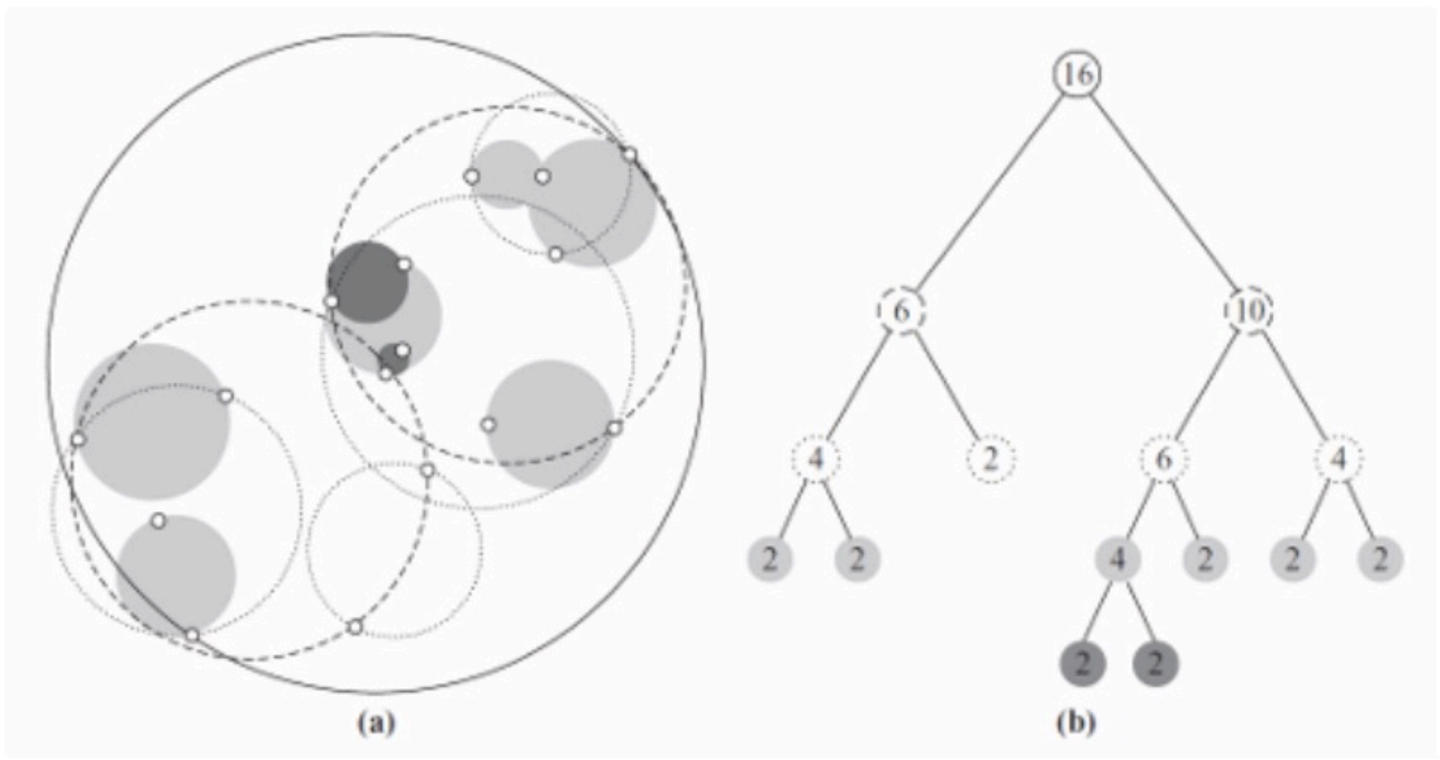
- Faster single queries, but have to build data structure
- More storage required for data structure

K-Nearest Neighbor Classification

How much does KNN cost?

A single query with m training examples, each with D features

Tree-Based KNN:



K-Nearest Neighbor Classification

How much does KNN cost?

Optimal algorithm depends largely on m, D, and distribution of training data.

Check out the scikit-learn documentation for their Nearest Neighbors Classifier for a detailed discussion:

http://scikit-learn.org/stable/modules/neighbors.html

K-Nearest Neighbor Classification

Wait. Aren't we supposed to be thinking *probabalistically*?

K-Nearest Neighbor Classification

Like most learning techniques we'll look at, KNN has both an algorithm-y and probabalistic interpretation.

Given a test example $\mathbf{x}$, let the probability that $\mathbf{x}$ belongs to class $c \in \{1, \ldots, C\}$ be modeled by

$$p(y = c \mid \mathbf{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x}, \mathcal{D})} \mathbf{I}(y_i = c)$$

where here $\mathbf{I}$ is the indicator function.

Assign to $\mathbf{x}$ the label $\hat{y} = c$ with highest probability.

Course Logistics

What You Need for this Course

- You must use Python 3
- You must have a willingness to do some math
    - multivariate calculus
    - basic linear algebra
    - probability and statistics
    - prove some things
- You must have some computer/programming skills
    - must know how to program
    - manipulate data (mostly text files)
    - algorithms relatively simple

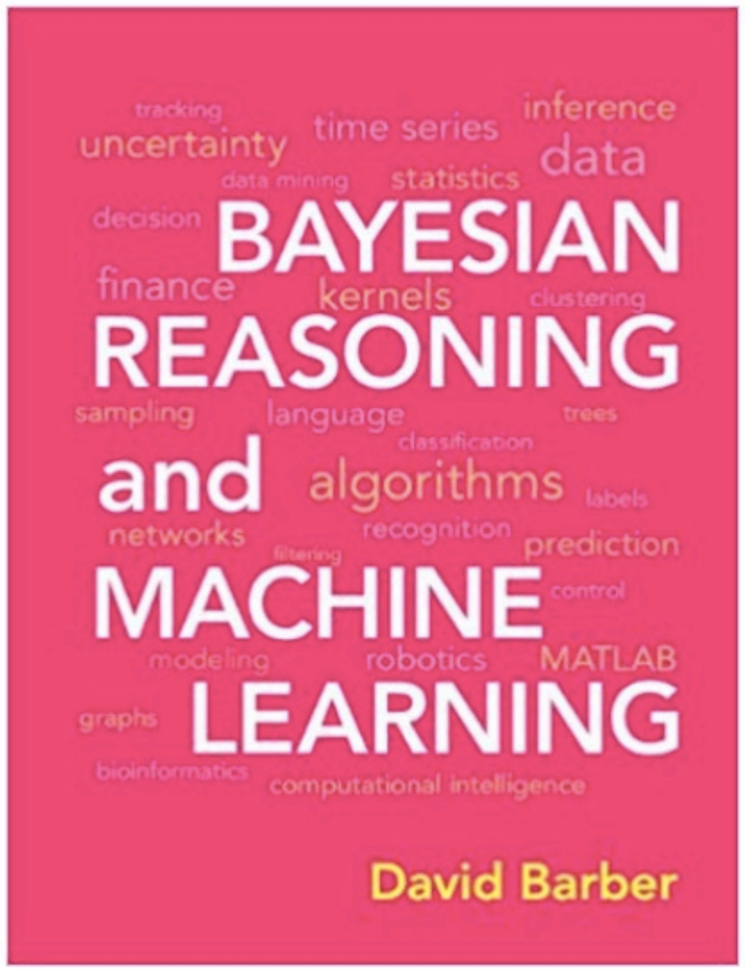
Course Logistics

- Before Class: Do assigned reading
- In Class: Lectures, Discussion, Occasional Hands-On
- Helpful to bring a laptop with Jupyter Notebook installed

Course Logistics

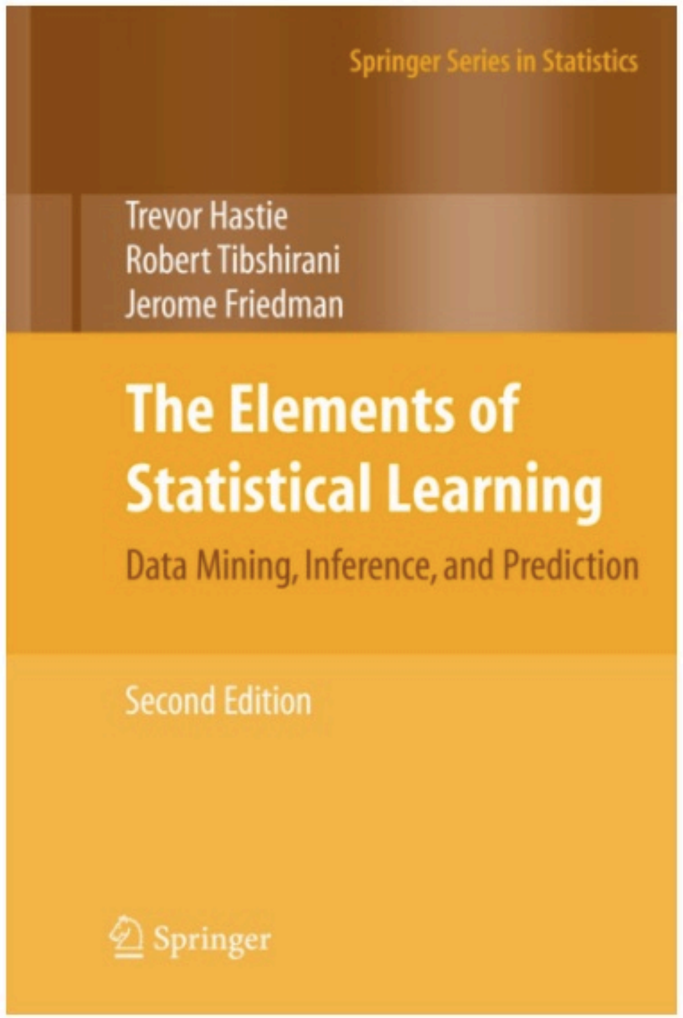
- Keep track of course webpage (Piazza)
- Homework Assignments, 3 Free Late Days
- Reading Quizzes on Moodle
- Midterm Exam
- Semester Group Project
- Read the Syllabus for More Details!
  - Grade Breakdown
  - Course Policies
- Let me know about special needs in a timely manner

Course Reading



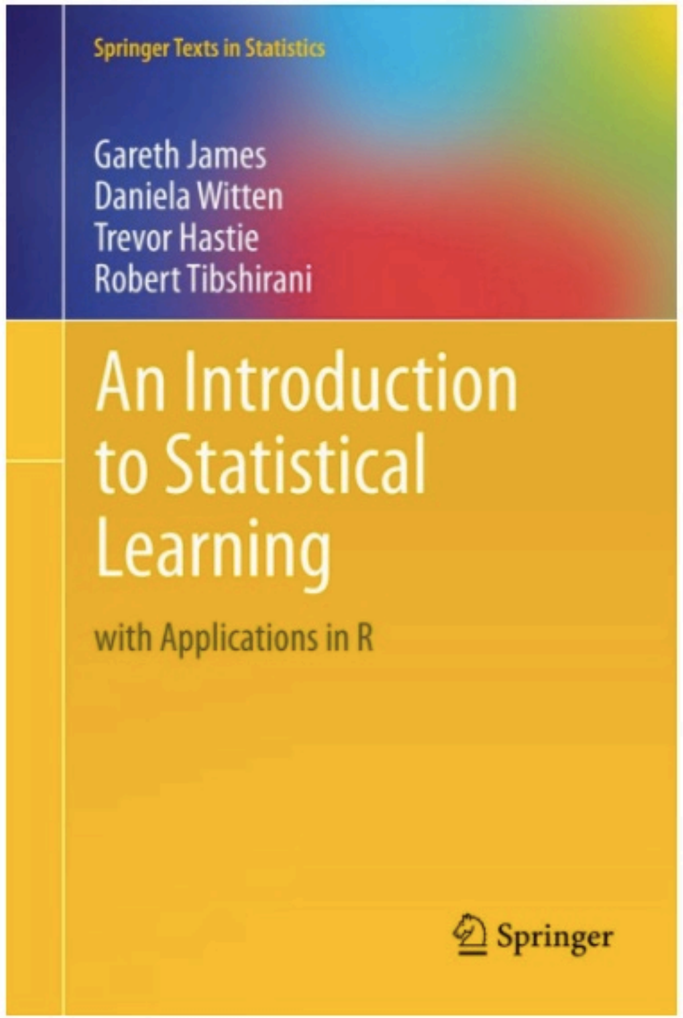
- ML from a Bayesian perspective
- A little mathy. Give it your best shot, but default to notation/rigor presented in videos
- **Free Online!**

Course Reading



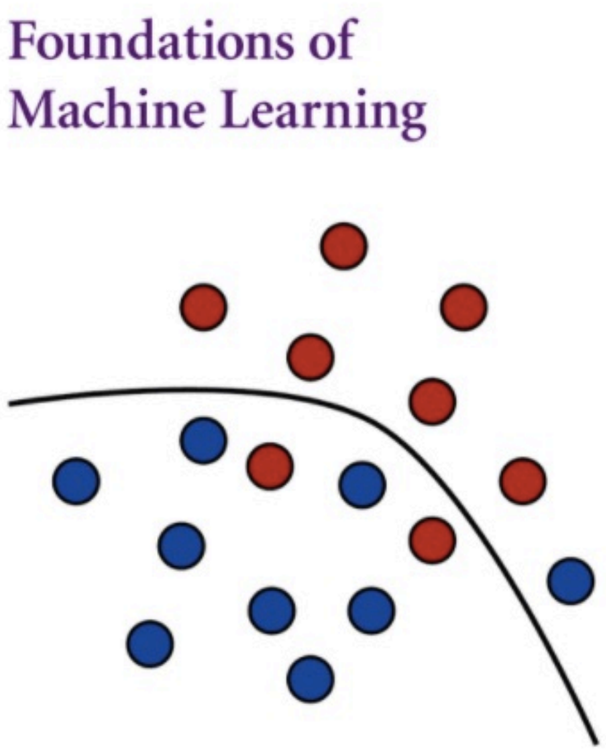
- Awesome book
- **Free Online!**

Course Reading



- Undergraduate version of ESL
- **Free Online!**
- All examples in R :\

Course Reading



- We'll use for theory unit
- Super-duper mathy
- Check it out if you like theory

Communicating with Piazza

We will use Piazza to manage all communication

http://piazza.com/colorado/spring2018/csci5622

- Announcements and Course Material posted here
- Ask technical and logistics questions in Q & A forum
- Discuss assigned work but don't post answers/vital code
- **Send private messages to faculty instead of email**

How to Ask for Help

- Explain what you're trying to do
- Give a minimal example of the problem
    - Should be easy to replicate
    - Shouldn't require info that only you have
- Explain what you **think** should happen
- Explain what you get instead (copy/pastes or screenshots are awesome)
- Explain what else you've tried

About Me

- Fifth year Instructor at CU (first three in Applied Math)
    - Specialize in Numerics, Data Science, ML coursers
- Before CU, at Lawrence Livermore Nat'l Lab
- Before that, PhD in Applied Math at CU
- Research: UQ, Multilevel Monte Carlo, Multigrid Methods
- Please call me **Chris** or Dr. Ketelsen
- **Office Hours**: MW 4:30-5:30pm and Th 12:30-2pm in ECOT 731

First Homework Assignment

- Implement KNN to classify handwritten digits
- Designed to be not too difficult
- Figure out our Python / Jupyter environment
- Figure out how to submit work to Moodle
- Due 26 January

Next Time...

- *Probability*
- First Probabilistic Learning Algorithm
    - Naive Bayes
- Application: Spam Filters

Acknowledgements

Many of the slides in this presentation were adopted from Jordan Boyd-Graber, Lauren Hannah, and Dave Blei

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani