CSCI 4622
Machine Learning


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

What is Machine Learning?

What is Machine Learning?

Seriously. What do **YOU** think it is?

What is Machine Learning?

Seriously. What do **YOU** think it is?

What is Machine Learning?

Arthur Samuel in 1959 defined Machine Learning as **the field of study that gives computers the ability to learn without explicitly being programmed**.

What is Machine Learning?

Arthur Samuel in 1959 defined Machine Learning as **the field of study that gives computers the ability to learn without explicitly being programmed.**

What is Machine Learning?

Arthur Samuel in 1959 defined Machine Learning as **the field of study that gives computers the ability to learn without explicitly being programmed**.

Tom Mitchell (Chair of CMU ML Dept.) in 1998 said: **in a 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 E.**

What is Machine Learning?

Arthur Samuel in 1959 defined Machine Learning as **the field of study that gives computers the ability to learn without explicitly being programmed**.

Tom Mitchell (Chair of CMU ML Dept.) in 1998 said: **in a 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 E.**

What is E?

Data are Everywhere

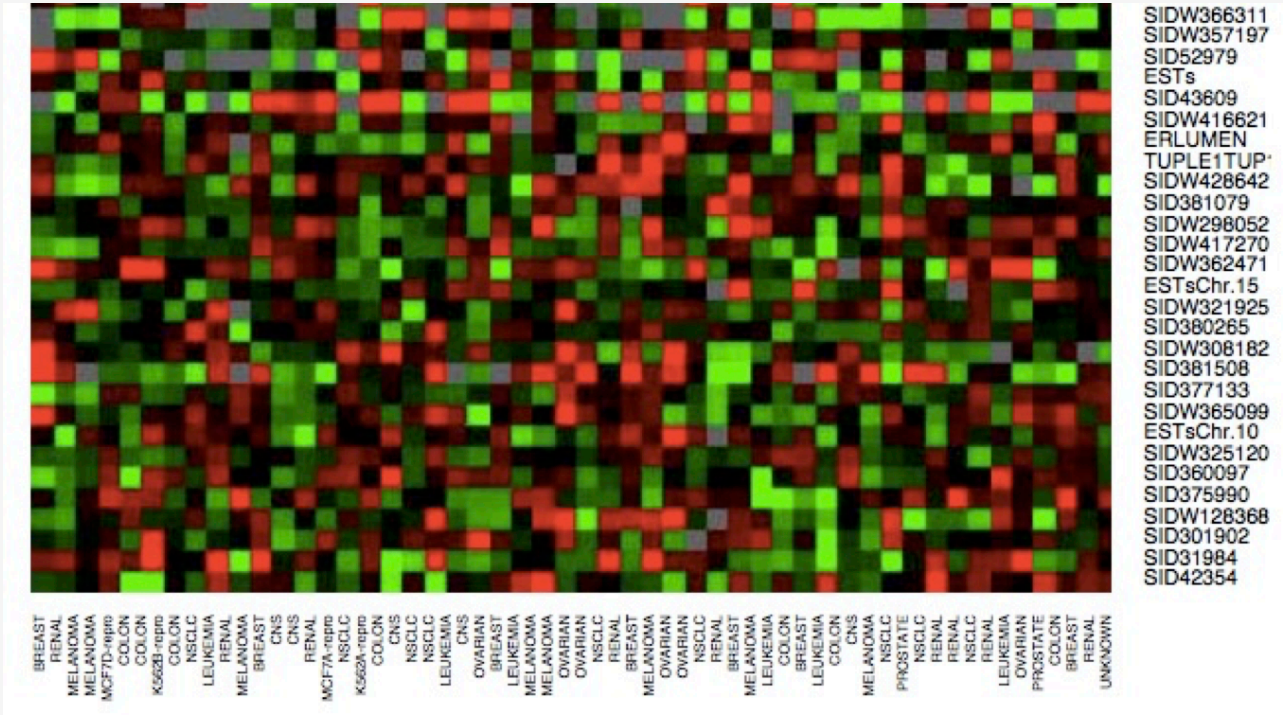
User Ratings

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	★★★☆☆

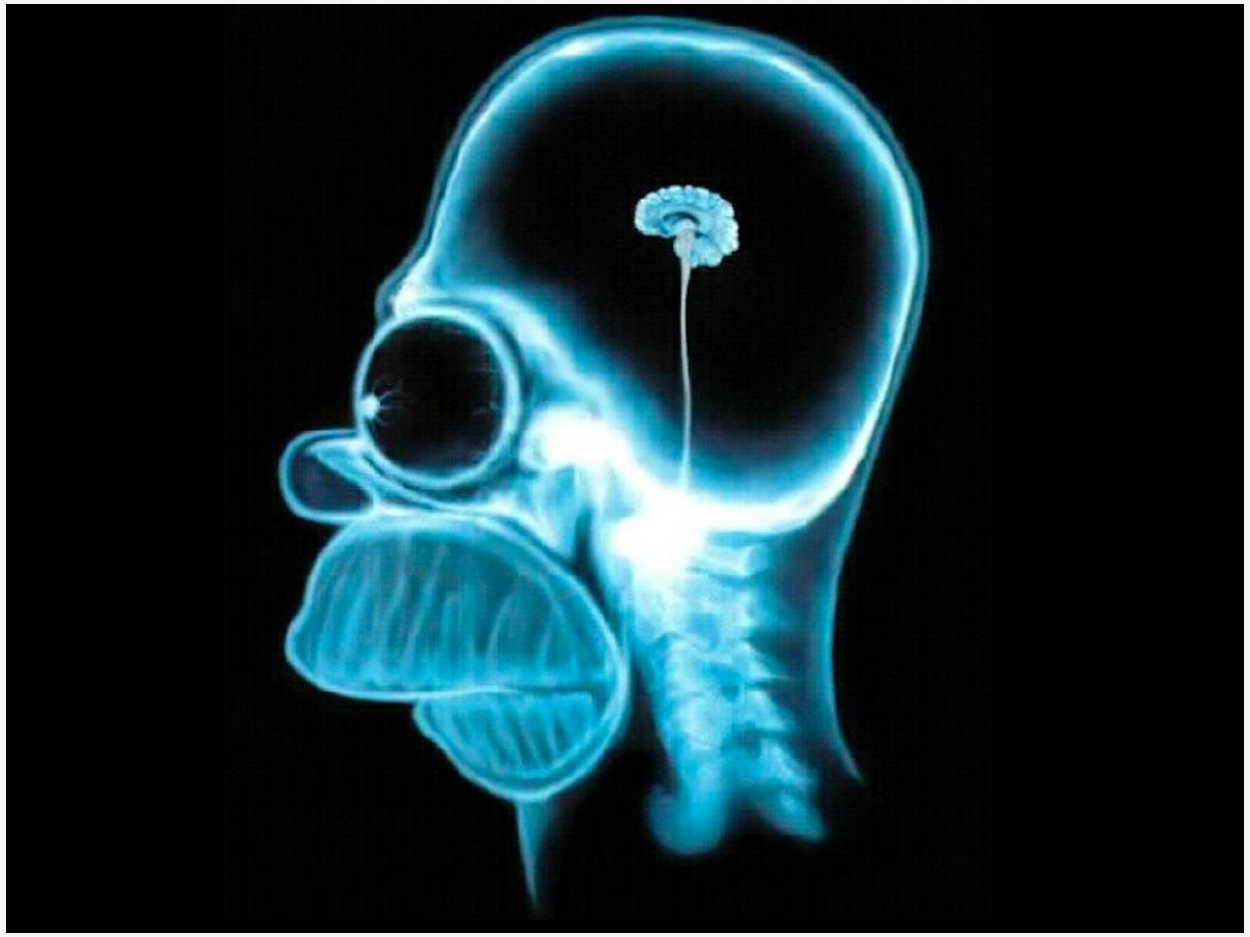
Document Collections

Social Media

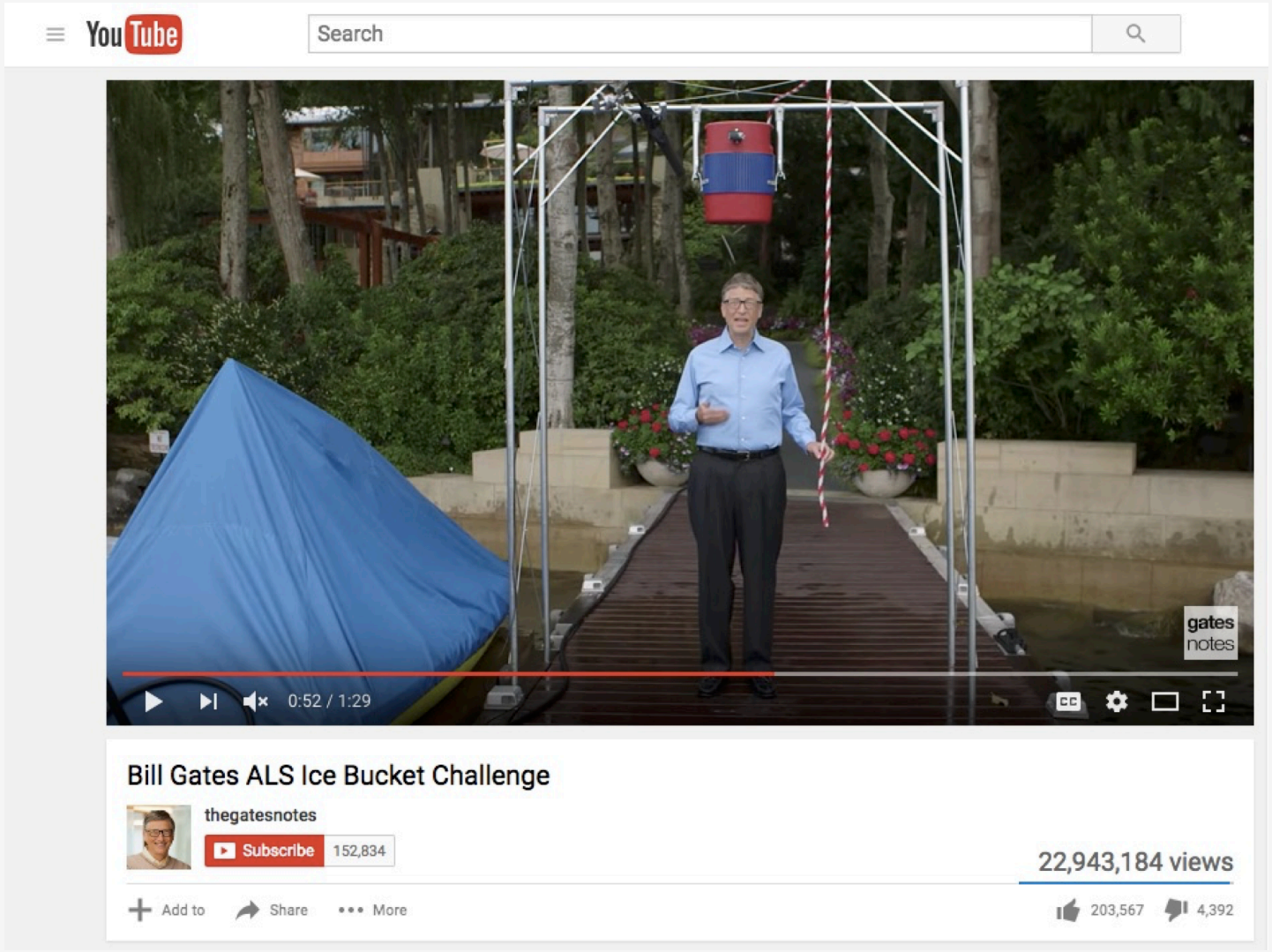
The Human Genome



The Human Brain



Commerce and Philanthropy

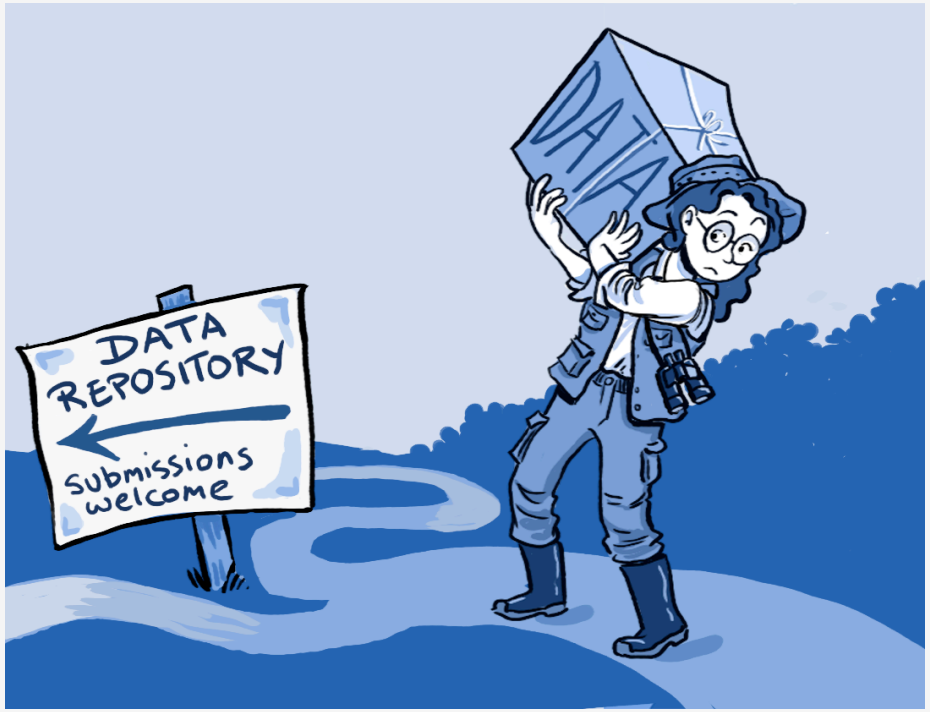


Cats


Funny Cats Compilation [Most See] Funny Cat Vid...
0:00 / 14:02
YouTube

The Data is Out There

Today, you can find pretty much any data that you want.



The hard part is deciding what to do with it.

Mathematical Foundation

$$\mathbf{X} \quad \Rightarrow \quad \mathbf{Z} \quad \Rightarrow \quad \mathbf{Y}$$

Data — Hidden Structure — Answer

- $\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$. Each X_k is called a **feature**
- $\mathbf{Y}$ 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 when it goes on the market

How Much is My House Worth?

What are the **features** $\mathbf{X} = (X_1, X_2, \ldots, X_p)$?

What is the **response** $\mathbf{Y}$?

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 ouradvertisement to their friends will be randomly picked to receive $1,000! You could be the winner!

Thank 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

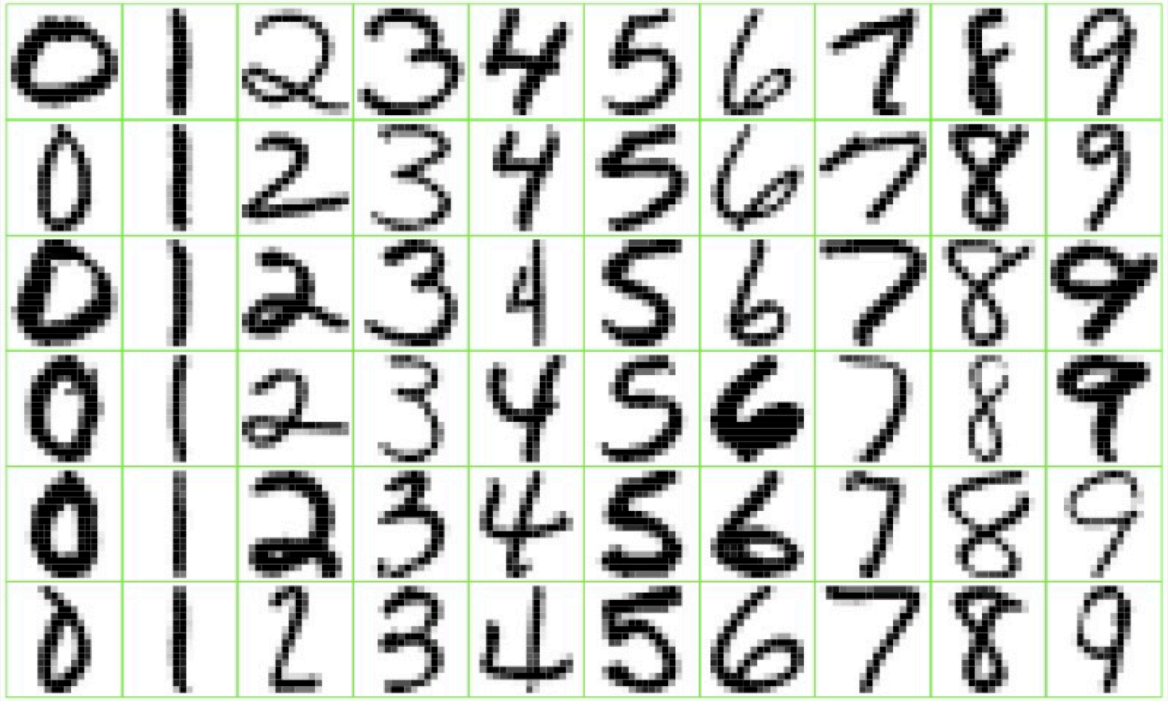
Is This Spam?

What are the **features** $\mathbf{X} = (X_1, X_2, \ldots, X_p)$?

What is the **response** $\mathbf{Y}$?

Handwritten Digit Recognition

Given: an image of a handwritten digit



Predict: which digit it is

Handwritten Digit Recognition

What are the **features** $\mathbf{X} = (X_1, X_2, \ldots, X_p)$?

What is the **response** $\mathbf{Y}$?

Supervised Learning

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

Supervised Learning

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

- **Assume** there is some true functional relationship between data and response: $f : \mathbf{X} \rightarrow \mathbf{Y}$
- **Given** a set of training examples: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- **Learn** a good approximation to $f : \mathbf{X} \rightarrow \mathbf{Y}$

Supervised Learning

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

- **Assume** there is some true functional relationship between data and response: $f : \mathbf{X} \rightarrow \mathbf{Y}$
- **Given** a set of training examples: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- **Learn** a good approximation to $f : \mathbf{X} \rightarrow \mathbf{Y}$
- **Housing Market Value Prediction**
  - Map (square footage, #BRs, Age) to market 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**:

$$y(\mathbf{x}) = f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon \quad \Rightarrow \quad \hat{y} = \hat{f}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_i$$

Example: Housing Prices

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

$$\hat{y} = \hat{f}(\mathbf{x}) = \underset{c}{\operatorname{argmax}} \;\; p(y = c \mid \mathbf{x})$$

Examples: Spam vs Ham, Digit Recognition

Unsupervised Learning

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

- Data is simply $\{\mathbf{x}_i\}_{i=1}^n$
- Try to get at $\mathbf{Z}$
- **Big Ideas**: Clustering, Similarity, Dimensionality Reduction

Cluster Images Together

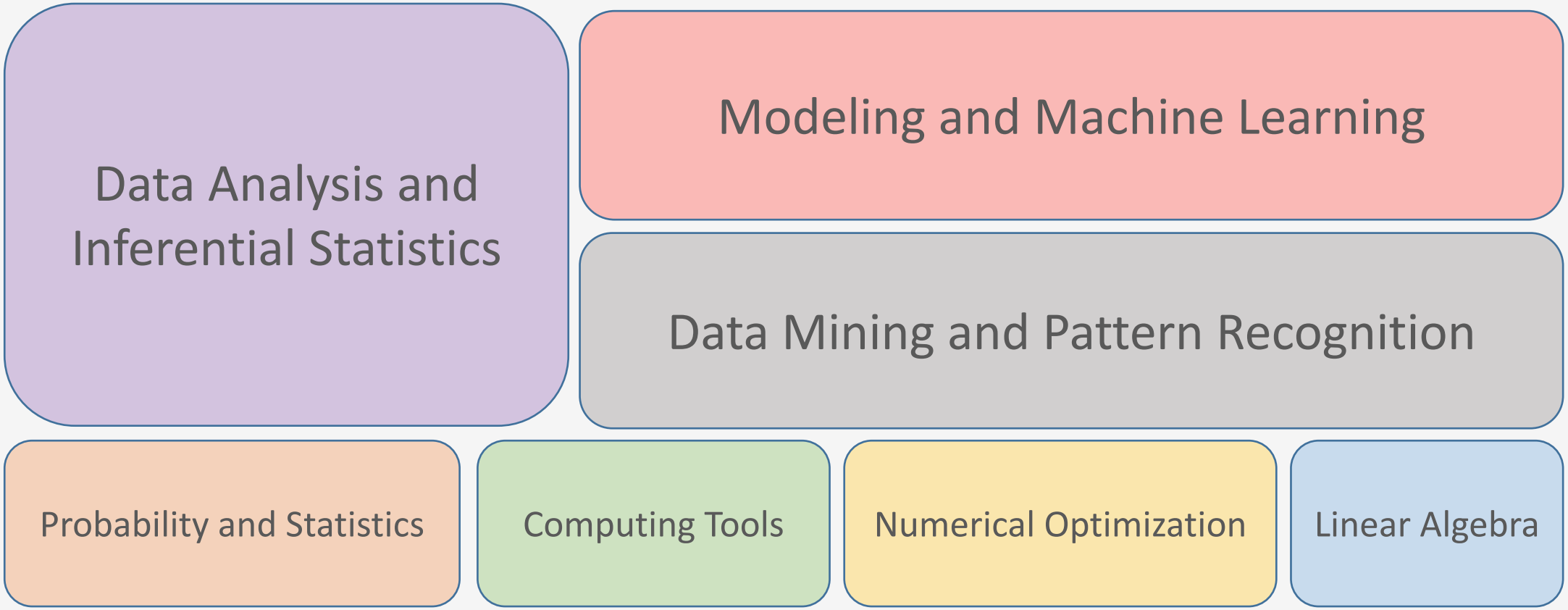
One Decomposition of Data Science

Data Analysis and Inferential Statistics

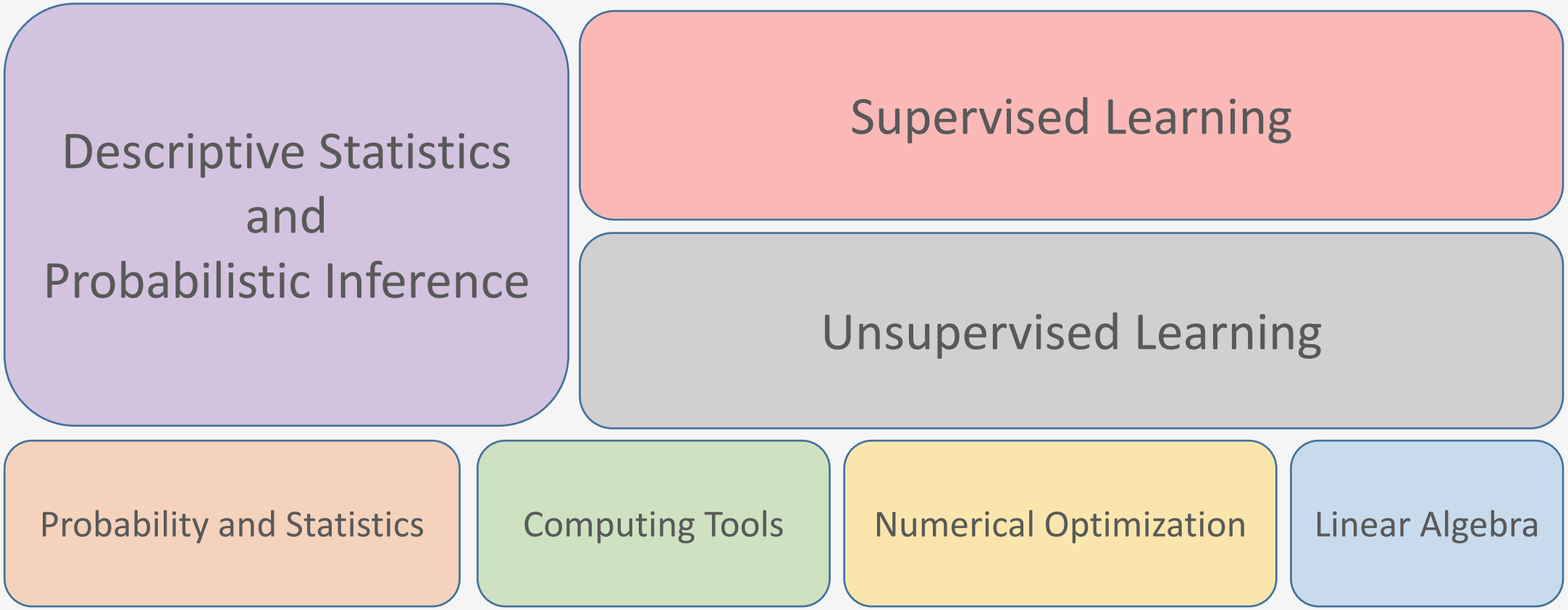
Modeling and Machine Learning

Data Mining and Pattern Recognition

One Decomposition of Data Science



One Decomposition of Data Science



One Decomposition of Data Science

CSCI 3022:
Intro to Data Science
with Prob-Stats

CSCI 4622: Machine Learning

CSCI 4022: Advanced Data Science

Probability and Statistics

Computing Tools

Numerical Optimization

Linear Algebra

The Plan

Goal: Fluency in the theoretical and computational aspects of supervised learning methods

At the end of this course you'll be able to

1. Identify the correct learning algorithm for your target problem
2. Prepare your data in appropriate way for your chosen algorithm
3. Implement your algorithm, both from scratch and with tools like Scikit-Learn
4. Critically analyze your results and make improvements
5. Understand the mathematical foundations behind your learning method
6. Tell **compelling stories** about data and machine learning

Course Logistics

Keep track of course webpages (Piazza, GitHub, Moodle)

- Piazza: https://piazza.com/colorado/spring2018/csci4622
  - Send me private messages on Piazza, rather than emails
  - Address message specifically to me if necessary

- GitHub: https://github.com/chrisketelsen/csci4622
  - Slides, notebooks, and homework will be posted here
  - Good Git tutorial if you're unfamiliar: http://rogerdudler.github.io/git-guide/

- Moodle: CSCI 4831-002 - Ketelsen - Machine Learning
  - Enrollment Key: csci4831-002-S18
  - Submitting homework, taking Reading Quizzes, and recording grades

Course Logistics

Course Work:

- Homework assignments every two weeks (30%)
  - Lowest homework score dropped
  - 3 total late days (1min – 23hr 59min late = 1 late day)
- Reading Quizzes on Moodle (10%)
- Midterm Exam (20%)
- Practicum (15%)
- Final Exam (25%)

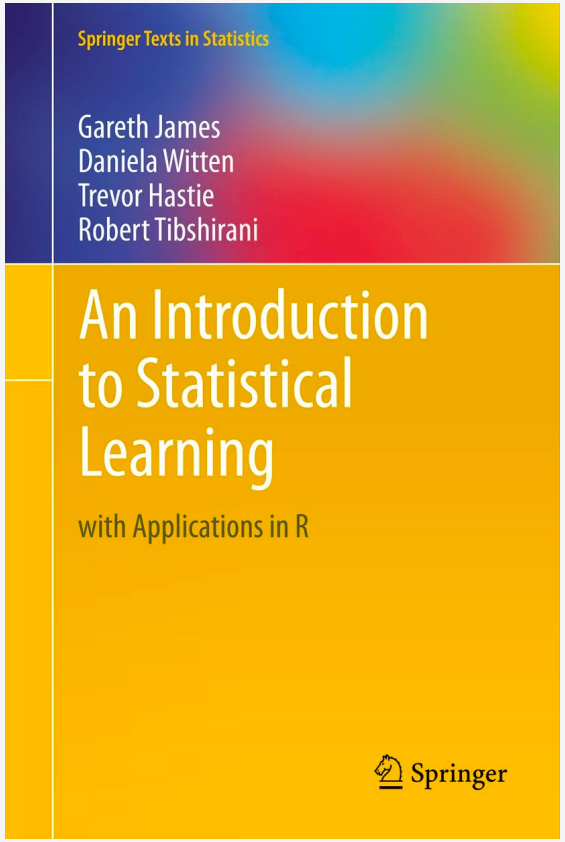
- **Exam Cutoff**: Must earn at least 55% on average on exams to get C- or better

Course Logistics

Collaboration Policy:

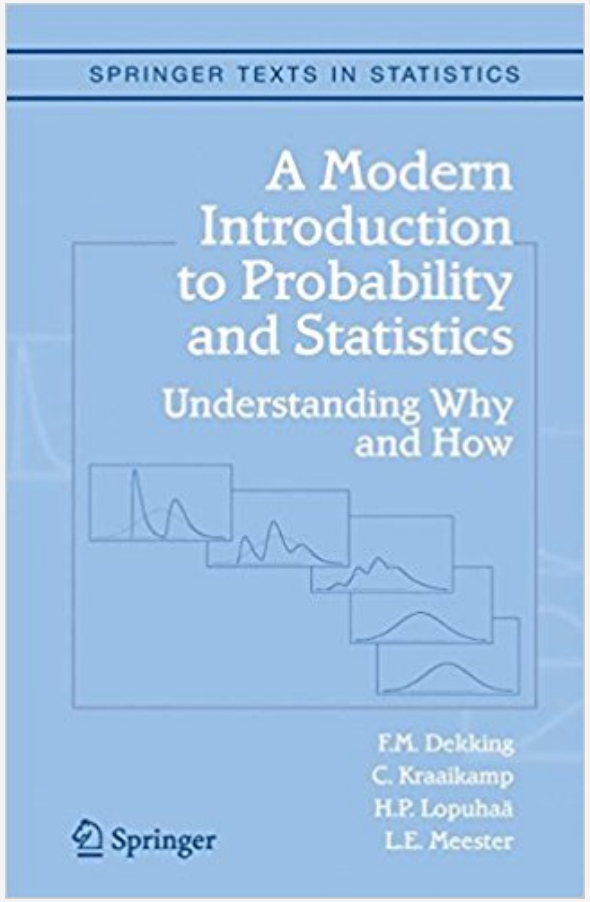
- Discuss homework with your classmates.
- But do your own work. **Write solutions and CODE on your own**.
- Give **hints**, not solutions, on Piazza.
- Make repositories containing your homework **private** (GitHub, Azure)
- Do **NOT** search for solutions online or post to things like Chegg, Reddit, Stack-Exchange
- For more details, see syllabus

Course Reading



- Good book with useful examples and exercises
- Free PDF from Book Page (link on syllabus)
- Overly **mathy** sometimes
- Only responsible for what we cover in class
- Does things in slightly different order than us
- Does everything in R :\

Course Reading



- Supplemental Text on Prob-Stats
- Only need if you want prob-stats refresher
- Free PDF CU Library! (link on syllabus)

Computing

- We will use Python 3 and in particular Numpy/Scikit-Learn
- Lot's of great data science libraries and decent plotting
- We'll exclusively work in Jupyter Notebooks
- Jupyter is ubiquitous DS collaboration and communication tool
- Easiest way to get **both** is **Anaconda Python 3.6**
- We **strongly recommend** you install local copy
- Often work on problems in groups in class
- Bring a laptop or have a buddy with a laptop

My Quirks and Expectations

- Be on time
- Stay until the end unless you're bleeding from the face
- Leave laptops in your bag unless we're working on a notebook

About Me

- 5th year as an instructor at CU (first 3 years in APPM, last 2 years in CS)
- Specialize in the **Mathy** courses (Discrete, Lin. Alg., Data Science, Machine Learning)
- Before CU, at Lawrence Livermore National Lab
- Before that, PhD in Applied Math at CU
- Before that, taught Philosophy at Washington State
- **Research**: Numerical Linear Algebra and **Stochastic Simulation**
- Please call me **Chris** or Dr. Ketelsen
- **Office Hours**: WF 12-1pm and Th 1:30-3pm in ECOT 731, or by appointment