PS132B
Machine Learning for Social Scientists

Introduction

Kirk Bansak

January 17, 2023

Why Machine Learning for Social Science or Public Policy?

Examples of Learning Problems

- Predict who will win an election, based on public opinion polls, economic data, and demographics.

Examples of Learning Problems

- Predict who will win an election, based on public opinion polls, economic data, and demographics.
- Estimate a person's expected wage based on his or her age, education, gender, and/or other personal characteristics.

Examples of Learning Problems

- Predict who will win an election, based on public opinion polls, economic data, and demographics.
- Estimate a person's expected wage based on his or her age, education, gender, and/or other personal characteristics.
- Classify online posts as either "fake news" or "real news" based on the words in the title and/or in the body of the article.

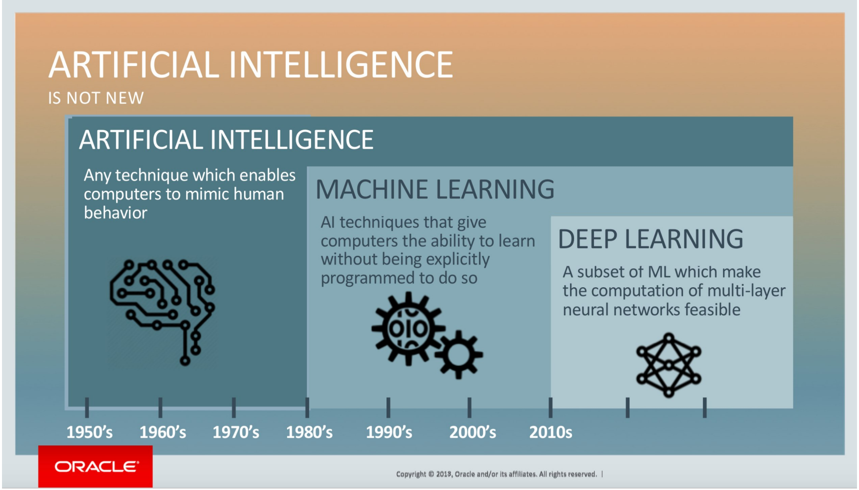
Examples of Learning Problems

- Predict who will win an election, based on public opinion polls, economic data, and demographics.
- Estimate a person's expected wage based on his or her age, education, gender, and/or other personal characteristics.
- Classify online posts as either "fake news" or "real news" based on the words in the title and/or in the body of the article.
- Identify substantive topics or themes in a collection of documents.

Machine learning refers to a vast set of tools that can learn, encode insights, and/or make predictions from data.

Supervised learning: Predict or estimate an output, usually quantitative (e.g. wage) or categorical (e.g. Republican/Democrat), based on a set of inputs.

Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures hidden in the data.


ARTIFICIAL INTELLIGENCE
IS NOT NEW
ARTIFICIAL INTELLIGENCE
Any technique which enables computers to mimic human behavior
MACHINE LEARNING
AI techniques that give computers the ability to learn without being explicitly programmed to do so
DEEP LEARNING
A subset of ML which make the computation of multi-layer neural networks feasible
1950's
1960's
1970's
1980's
1990's
2000's
2010s
ORACLE
Copyright © 2018, Oracle and/or its affiliates. All rights reserved.

From AI to Machine Learning

- "Static" algorithms of early AI: computers used to make decisions based on simple, clear, preset rules with fixed conditions

From AI to Machine Learning

- "Static" algorithms of early AI: computers used to make decisions based on simple, clear, preset rules with fixed conditions
  - e.g. Early email spam filters look for pre-specified phrases

From AI to Machine Learning

- "Static" algorithms of early AI: computers used to make decisions based on simple, clear, preset rules with fixed conditions
  - e.g. Early email spam filters look for pre-specified phrases
- Highly inflexible $\rightarrow$ Limited effectiveness

From AI to Machine Learning

- "Static" algorithms of early AI: computers used to make decisions based on simple, clear, preset rules with fixed conditions
  - e.g. Early email spam filters look for pre-specified phrases
- Highly inflexible $\rightarrow$ Limited effectiveness
- In contrast, machine learning techniques:
  - Take actual data on the phenomenon of interest
  - Have computer run algorithms to learn from the data
  - Automatically uncover a wider landscape of subtle patterns
  - Systematize those into insights (models)
  - Apply those insights (models) to new instances (new data) to make assessments or decisions

From AI to Machine Learning

- "Static" algorithms of early AI: computers used to make decisions based on simple, clear, preset rules with fixed conditions
  - e.g. Early email spam filters look for pre-specified phrases
- Highly inflexible $\rightarrow$ Limited effectiveness
- In contrast, machine learning techniques:
  - Take actual data on the phenomenon of interest
  - Have computer run algorithms to learn from the data
  - Automatically uncover a wider landscape of subtle patterns
  - Systematize those into insights (models)
  - Apply those insights (models) to new instances (new data) to make assessments or decisions
  - In essence, will evaluate all of the features (subject line, sender, text, metadata, etc.) of a new email and systematically compare them to the patterns of past (spam vs. non-spam) emails to make determination.

How We Got Here...

- 1800s - 1980s: linear models

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

  More mathematical/statistical/algorithmic innovations

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

  More mathematical/statistical/algorithmic innovations

  $+$ More computational power

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

  More mathematical/statistical/algorithmic innovations
  + More computational power
  + More data

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

  More mathematical/statistical/algorithmic innovations
  + More computational power
  + More data
  + More awareness

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

  More mathematical/statistical/algorithmic innovations
  $+$ More computational power
  $+$ More data
  $+$ More awareness
  $=$ Greater power, broader applications, bigger audience

How We Got Here...

- 1800s - 1980s: linear models
- Since 1980s:

  More mathematical/statistical/algorithmic innovations
  + More computational power
  + More data
  + More awareness
  = Greater power, broader applications, bigger audience

Math & Algorithms + Computation + Data

(Navigation + Engine + Fuel)

Why Machine Learning for Social Science or Public Policy?

Why Machine Learning for Social Science or Public Policy?

Social science training, perspectives, and intuitions

Why Machine Learning for Social Science or Public Policy?

Social science training, perspectives, and intuitions

$+$

Subject matter expertise

Why Machine Learning for Social Science or Public Policy?

Social science training, perspectives, and intuitions

$+$

Subject matter expertise

$+$

Technical skills to analyze data and transform data into a tool

Why Machine Learning for Social Science or Public Policy?

Social science training, perspectives, and intuitions

$+$

Subject matter expertise

$+$

Technical skills to analyze data and transform data into a tool

$=$

Unique value across a diversity of domains

Machine Learning Applications

Industry: Increasing Revenue

- Measuring consumer opinion and behavior
- Delivering engaging content to users

Public Sector: Optimizing Public Policy Decision-making

- Predict health and safety risks
- Assist pre-trial release and parole decisions

Campaigns: Winning the Vote

- Classify voters based on likely voting, using consumer information
- Identify ideological patterns based on social media behavior

Social Science: Understanding our Social and Political World

- Polarization in political institutions: Clinton, Jackman, and Rivers (2004)
- Extent/strategy of Chinese censorship: King, Pan, and Roberts (2014, 2017)
- Public support for economic austerity: Bansak, Bechtel, and Margalit (2021)

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations).

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.).

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

Not Just for “Big Data”

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include?

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include? **1048576 possible subsets.**

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include? **1048576 possible subsets.**
- Which variables to interact?

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include? **1048576 possible subsets.**
- Which variables to interact? **190 possible 2-way interactions, 1140 possible 3-way interactions, 4845 possible 4-way interactions, ...**

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include? **1048576 possible subsets.**
- Which variables to interact? **190 possible 2-way interactions, 1140 possible 3-way interactions, 4845 possible 4-way interactions, ...**
- How can we efficiently test and assess all the different possibilities?

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include? **1048576 possible subsets.**
- Which variables to interact? **190 possible 2-way interactions, 1140 possible 3-way interactions, 4845 possible 4-way interactions, ...**
- How can we efficiently test and assess all the different possibilities?
- How do you know your model will perform well for people not in your dataset?

Not Just for "Big Data"

Imagine you have a small or medium sized dataset of individuals (e.g. 100 - 10,000 observations). You are using these data to try to predict or understand something about these people (e.g. their likely vote choice, their probability of committing a crime, their future earnings, etc.). You may have 20 explanatory variables available to make your predictions. Build the best regression model...

- Which variables to include? **1048576 possible subsets.**
- Which variables to interact? **190 possible 2-way interactions, 1140 possible 3-way interactions, 4845 possible 4-way interactions, ...**
- How can we efficiently test and assess all the different possibilities?
- How do you know your model will perform well for people not in your dataset?

Automated machine learning model building and assessment methods can help with small, medium, and big data problems!

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)

Not Just for “Big Data”

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions
- Big Number:

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions
- Big Number:

  7 Billion researchers

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions
- Big Number:

  7 Billion researchers

  Impossibly Fast (enumerate one clustering every millisecond)

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions
- Big Number:

  7 Billion researchers
  Impossibly Fast (enumerate one clustering every millisecond)
  Working Around the Clock (24/7/365)

Not Just for "Big Data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions
- Big Number:

  7 Billion researchers
  Impossibly Fast (enumerate one clustering every millisecond)
  Working Around the Clock (24/7/365)
  $\approx 2.16 \times 10^{94}$ years!

Not Just for "Big Data"

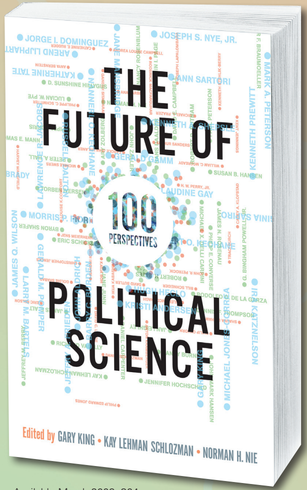
Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB; A B)
- Bell(3) = 5 (ABC; AB C; A BC; AC B; A B C)
- Bell(5) = 52
- Bell(100) = $4.75x10^{115}$ partitions
- Big Number:

  7 Billion researchers
  Impossibly Fast (enumerate one clustering every millisecond)
  Working Around the Clock (24/7/365)
  $\approx 2.16 \times 10^{94}$ years!

Automated methods can help with small, medium, and large problems!

This Can Actually Work!



Available March 2009: 304pp
Pb: 978-0-415-99701-0: **$24.95**
www.routledge.com/politics

THE FUTURE OF POLITICAL SCIENCE

100 Perspectives

Edited by **Gary King**, Harvard University, **Kay Lehman Schlozman**, Boston College and **Norman H. Nie**, Stanford University

"The list of authors in *The Future of Political Science* is a 'who's who' of political science. As I was reading it, I came to think of it as a platter of tasty hors d'oeuvres. It hooked me thoroughly."
—Peter Kingstone, University of Connecticut

"In this one-of-a-kind collection, an eclectic set of contributors offer short but forceful forecasts about the future of the discipline. The resulting assortment is captivating, consistently thought-provoking, often intriguing, and sure to spur discussion and debate."
—Wendy K. Tam Cho, University of Illinois at Urbana-Champaign

"King, Schlozman, and Nie have created a visionary and stimulating volume. The organization of the essays strikes me as nothing less than brilliant. . . It is truly a joy to read."
—Lawrence C. Dodd, Manning J. Dauer Eminent Scholar in Political Science, University of Florida

Routledge

Evaluators' Rate Machine Choices Better Than Their Own (Grimmer and King)

Generate pairs of similar documents: Humans vs Machines

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2

Evaluators' Rate Machine Choices Better Than Their Own (Grimmer and King)

Generate pairs of similar documents: Humans vs Machines

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Random Selection	1.38	1.16	1.60

Evaluators' Rate Machine Choices Better Than Their Own (Grimmer and King)

Generate pairs of similar documents: Humans vs Machines

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Random Selection	1.38	1.16	1.60
Hand-Coded Clusters	1.58	1.48	1.68

Evaluators' Rate Machine Choices Better Than Their Own (Grimmer and King)

Generate pairs of similar documents: Humans vs Machines

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Random Selection	1.38	1.16	1.60
Hand-Coded Clusters	1.58	1.48	1.68
Hand-Coded Pairings	2.06	1.88	2.24

Evaluators' Rate Machine Choices Better Than Their Own (Grimmer and King)

Generate pairs of similar documents: Humans vs Machines

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Random Selection	1.38	1.16	1.60
Hand-Coded Clusters	1.58	1.48	1.68
Hand-Coded Pairings	2.06	1.88	2.24
Machine Pairings	2.24	2.08	2.40

Evaluators' Rate Machine Choices Better Than Their Own (Grimmer and King)

Generate pairs of similar documents: Humans vs Machines

- Scale: (1) unrelated, (2) loosely related, or (3) closely related
- Table reports: mean(scale)

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Random Selection	1.38	1.16	1.60
Hand-Coded Clusters	1.58	1.48	1.68
Hand-Coded Pairings	2.06	1.88	2.24
Machine Pairings	2.24	2.08	2.40

P.S. The hand-coders were themselves the evaluators!

This Can Actually Work!

Bansak et al. (2018) develop an algorithm for optimizing the geographic assignment of refugees to boost their employment outcomes.

This Can Actually Work!

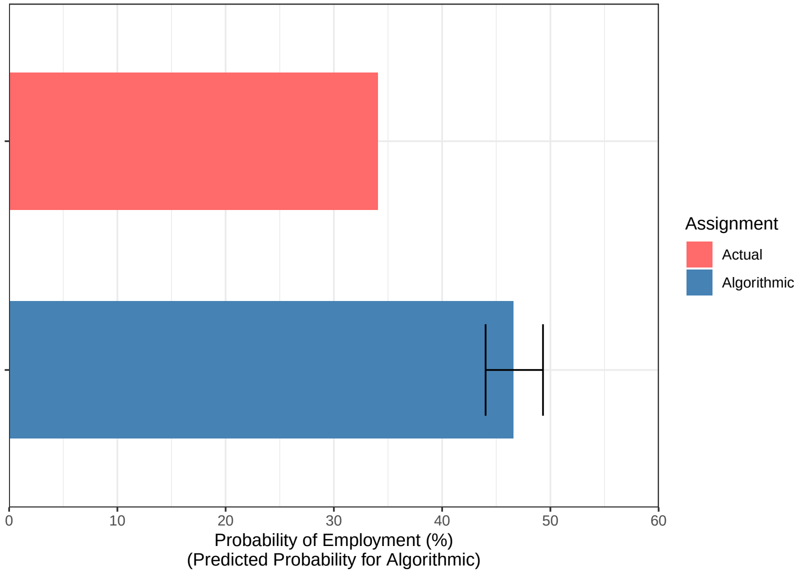
Bansak et al. (2018) develop an algorithm for optimizing the geographic assignment of refugees to boost their employment outcomes.

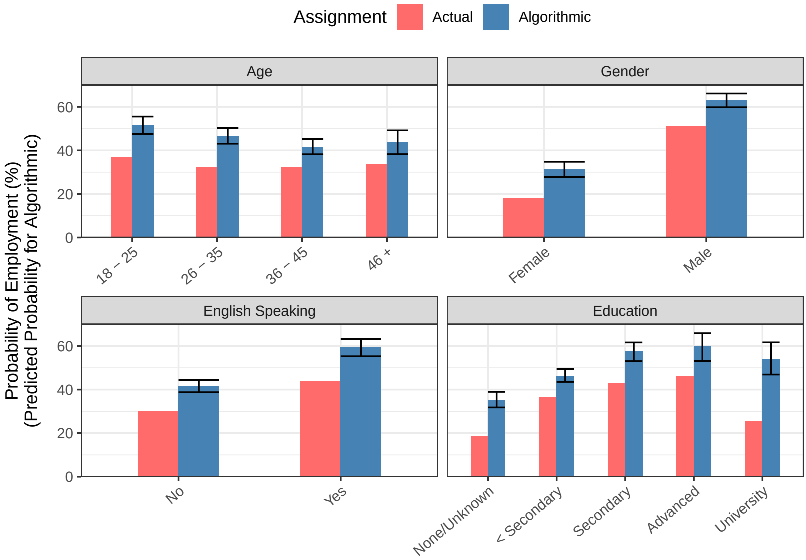
- Two-stage algorithm
  1. Modeling: machine learning models to predict refugees' employment outcomes at their different possible geographic destinations
  2. Matching: assignment of refugees to optimal locations, subject to capacity and other constraints

This Can Actually Work!

Bansak et al. (2018) develop an algorithm for optimizing the geographic assignment of refugees to boost their employment outcomes.

- Two-stage algorithm
  1. Modeling: machine learning models to predict refugees' employment outcomes at their different possible geographic destinations
  2. Matching: assignment of refugees to optimal locations, subject to capacity and other constraints
- Backtests using data in United States and Switzerland show that algorithmic assignment can lead to 40-70% gains in employment relative to status quo assignment procedure.


0
10
20
30
40
50
60
Probability of Employment (%)
(Predicted Probability for Algorithmic)
Assignment
Actual
Algorithmic


Assignment
Actual
Algorithmic
Age
Gender
English Speaking
Education
Probability of Employment (%)
(Predicted Probability for Algorithmic)
0
20
40
60
18 – 25
26 – 35
36 – 45
46 +
Female
Male
No
Yes
None/Unknown
< Secondary
Secondary
Advanced
University

Presumptions for this Course

1. Machine learning is relevant and useful in a wide range of academic and non-academic fields.
2. We will be able to broadly understand the models, intuitions, and strengths and weaknesses of the various approaches.
3. While it is important to know what general job is performed by each cog, it is not necessary to have the skills to completely construct the underlying algorithms.
4. Applying machine learning methods to "real-world problems" requires not only quantitative skills but also conceptual reasoning and subject matter understanding.

Core Objectives

Ultimate Goal: introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely.

Two Broad Categories of Machine Learning Covered:

- Supervised Learning
- Unsupervised Learning

Proximate Goals:

1) Learn a variety of machine learning techniques and how to effectively choose between them and use them with real-world data.
2) Learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems.
3) Develop programming abilities in R.
4) Introduce substantive problems in lectures, homework, and sections.

Core Objectives

This class is not a course on:

1) Classical statistical inference or causal inference
2) Full technical details behind machine learning methods, such as optimization algorithms and theoretical properties
3) The full universe of machine learning
4) How to become a professional programmer

Prerequisites

PS 3 or Data 8 (or equivalent coursework). This includes:

- A mechanical understanding of regression
- A brief introduction to statistical inference
- Experience with `R` or `Python`

⋆ We will exclusively use `R`, but previous knowledge of `Python` will enable getting up to speed on `R`

As our primary reference, we will use the book listed below:

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, **Second Edition**, 2021.

Logistics

Teaching Staff

Sections

Evaluation

- Problem Sets: Six assignments, 40% of final grade
  - Should be submitted in R markdown
  - Can think through and discuss problem in small groups (of 2-3), but (a) you must write your own answers and code, and (b) you must specify whom you worked with
- Challenges: Two group challenges, 25% of final grade
  - Challenge 1: Predicting recidivism
  - Challenge 2: Analyzing political text
- Midterm exam (in-class): 15%
- Final exam (official exam schedule): 20%

Ed Discussion

Course Plan

- Introductory Week

1. R
2. Supervised Learning
3. Unsupervised Learning

- Concluding Week

Introduction	Week 1	01/17	Introduction
	Week 1	01/19	A Machine Learning Focus on Regression *Read: ISLR pp. 15 - 39*
Unit 1: **R**	Week 2	01/24	Data and Datasets *Read: Kelleher and Tierney (2018); Kaplan (2009)*
	Week 2	01/26	(Re)Introduction to `R` *Do: Install/Update `R` and `RStudio` on your computer*
	Week 3	01/31	Introduction to `R Markdown` *Do: Install the `tidyverse` and `knitr` packages* **PSet 1 assigned**
	Week 3	02/02	Diving Deeper into `R`: Core Functionality *Read: Venables et al. (2022), Chapters 2, 3, and 6*
	Week 4	02/07	Diving Deeper into `R`: Data Visualization and Exploration *Read: Wickham and Grolemund (2017), Chapter 3* **PSet 1 due, PSet 2 assigned**
	Week 4	02/09	Diving Deeper into `R`: Functions and Iteration

For Next Class

Read: ISLR pp. 15 - 39