[POLS 8500] Applied Machine Learning

Professor Jason Anastasopoulos
ljanastas@uga.edu

University of Georgia

January 5, 2017

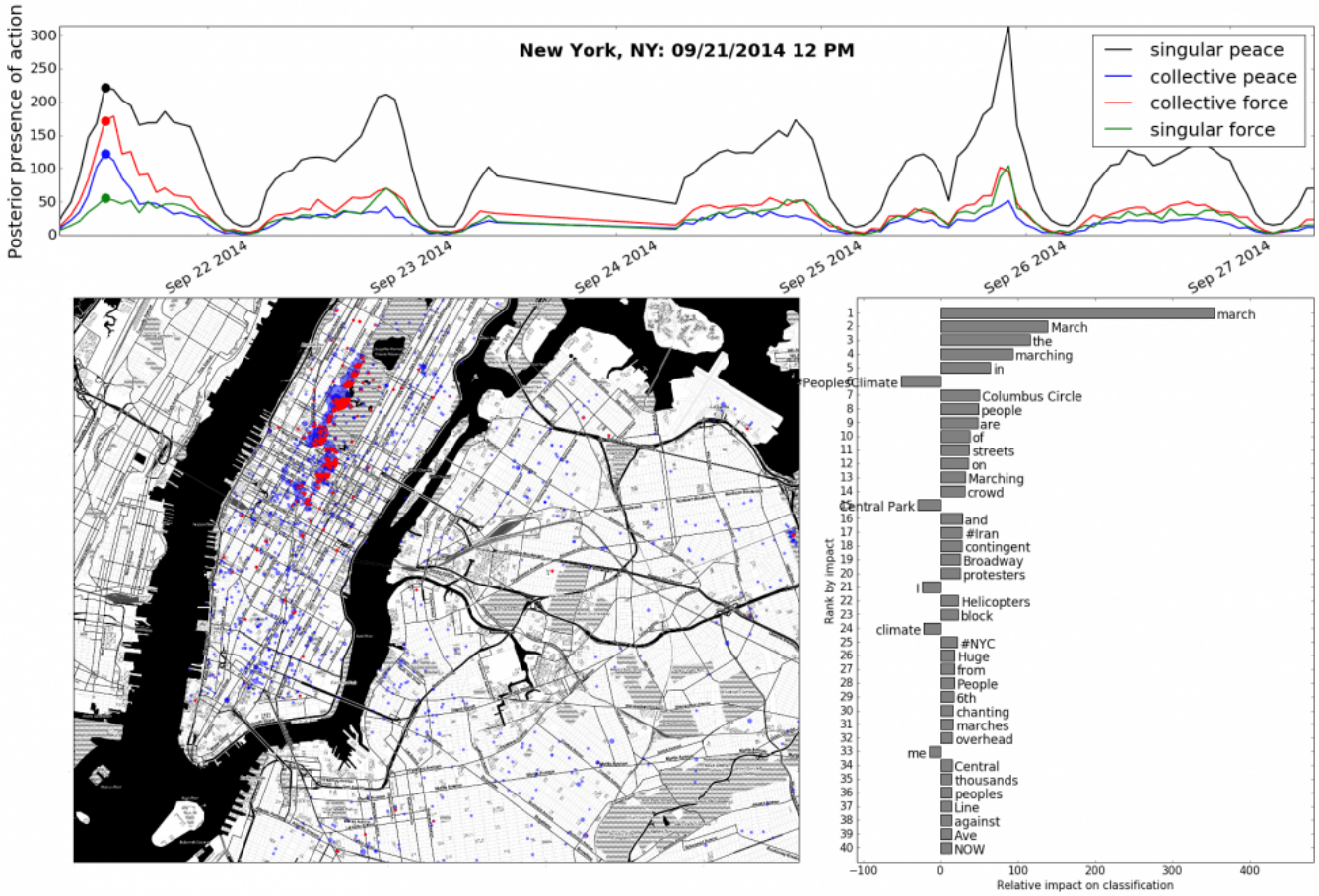
About Me

My full name has 27 letters.

About Me

I have a crazy 9 month old named Seth.

About Me

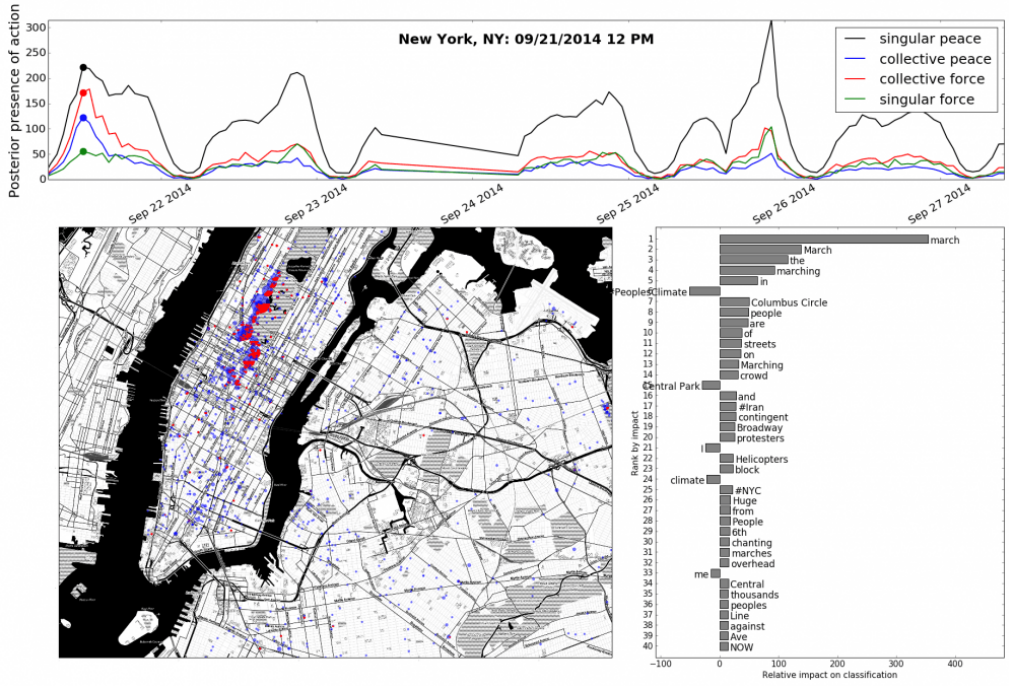


I do research on public policy, political science and machine learning.

Recent Machine Learning Projects

1. Identifying violent and non-violent protest activity using streaming Twitter data (n"aive Bayes).
2. Measuring violence in religious texts (support vector machines).

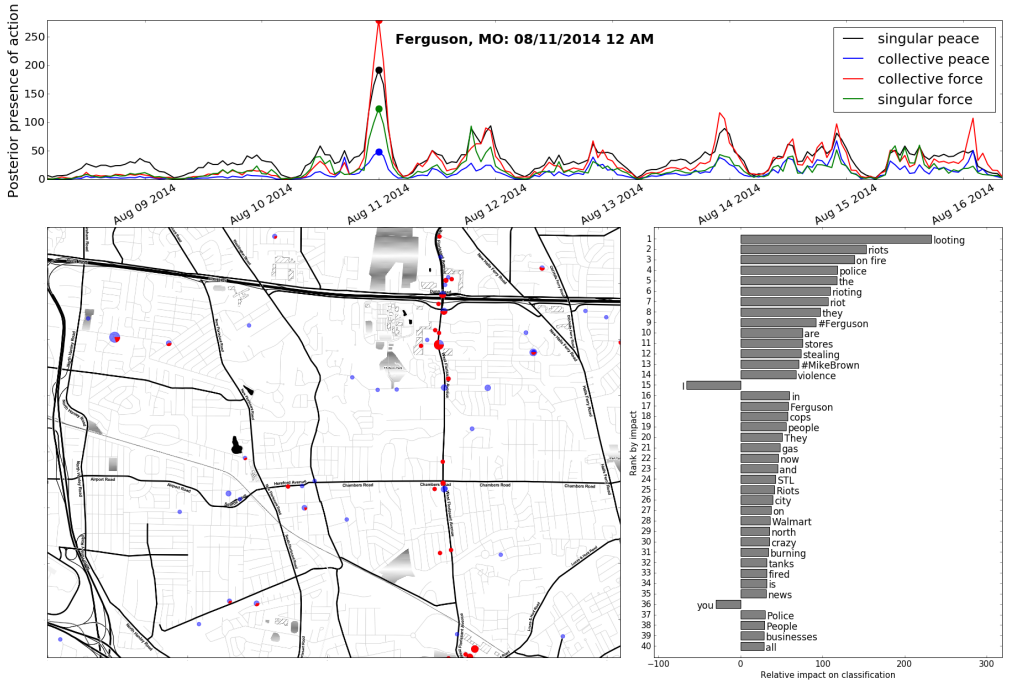
Identifying violent and non-violent protest activity using streaming Twitter data



Violent and non-violent protest activity in New York City.

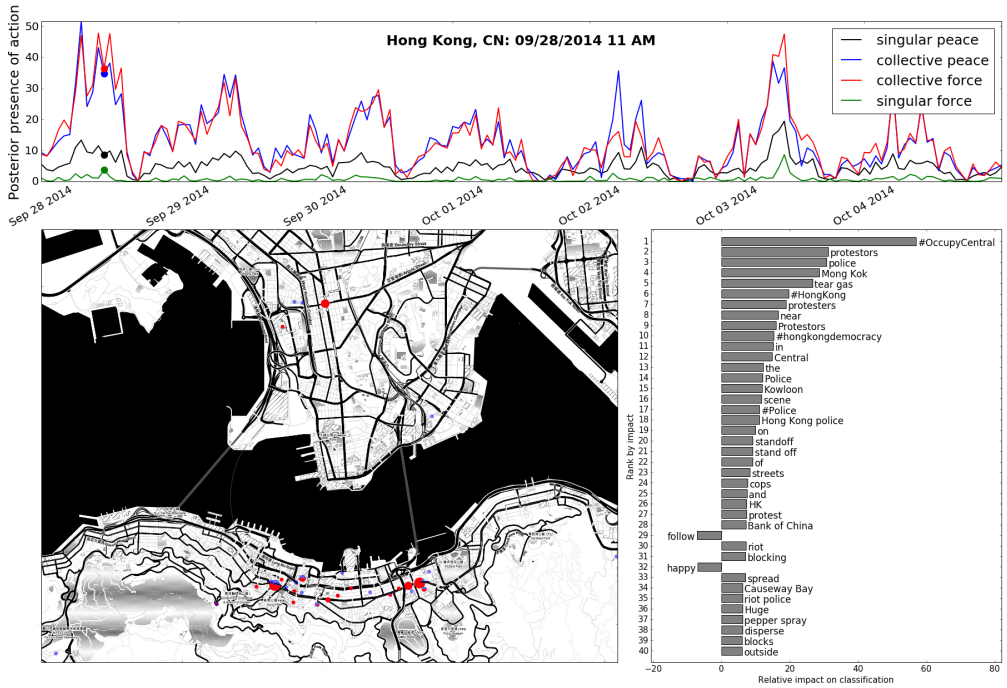
Modified näive Bayes algorithm trained on a large geo-coded Twitter database.

Identifying violent and non-violent protest activity using streaming Twitter data



Violent and non-violent protest activity in the wake of the Ferguson verdict.

Identifying violent and non-violent protest activity using streaming Twitter data



Violent and non-violent protest activity during Hong Kong Occupy protests.

Measuring violence in religious texts

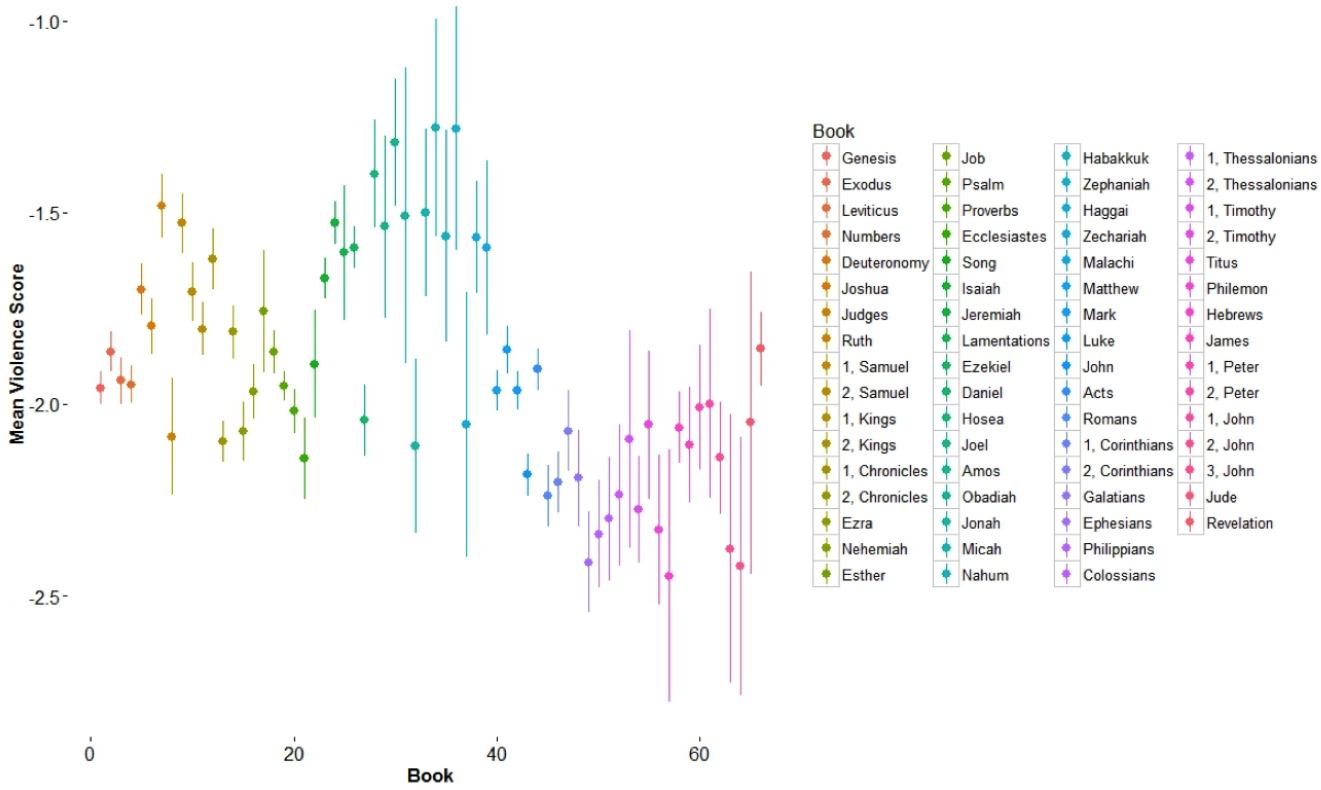
	Bible
Collective	"He arose and struck the Philistines until his hand was weary and his hand froze to the sword..." (2 Samuel 23:10)
Collective Promote	"I will set a fire in Egypt Sin shall be in great anguish..." (Ezekiel 30:16)
Interpersonal	"Even in the third year of Asa king of Judah did Baasha kill him and reigned in his place" (1 Kings 15:28)
Interpersonal Promote	"The daughter of any priest if she profanes herself by playing the prostitute she profanes her father she shall be burned with fire." (Leviticus 21:9)

Table 1: Samples of classified verses from training data.

Identification of different types of violence in verses with support vector machines.

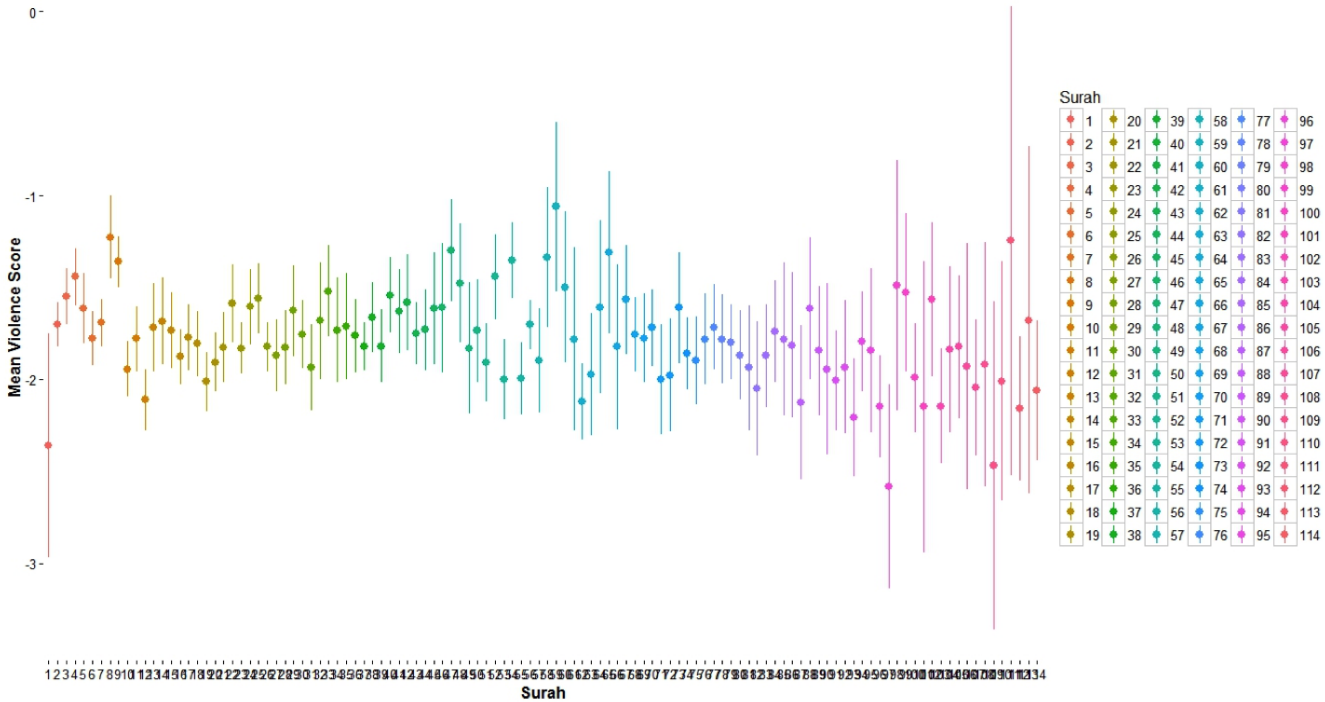
Trained on classified Quran and Bible verses.

Measuring violence in religious texts



SVM confidence scores to estimate patterns of violence in the Bible...

Measuring violence in religious texts



...and in the Quran.

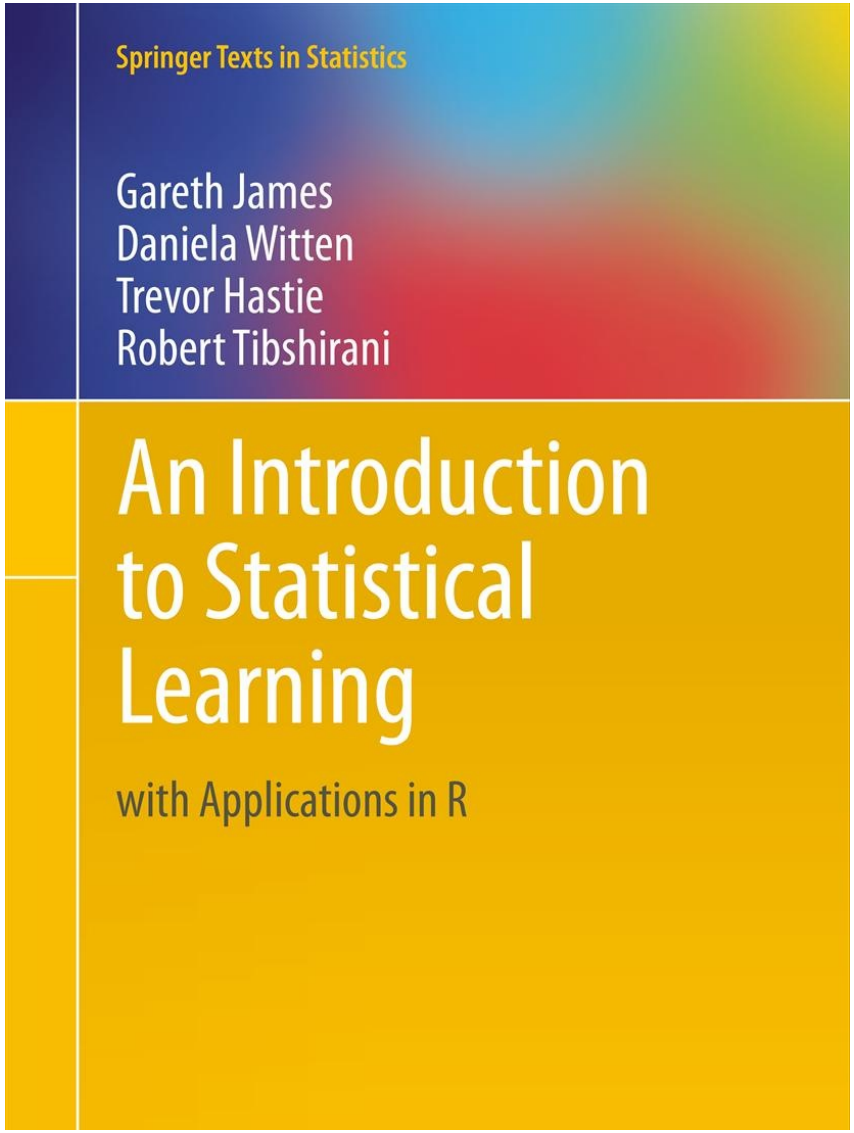
About POLS 4150

- Course goals.
- Textbooks.
- R.
- Syllabus.
- Intro to Machine Learning

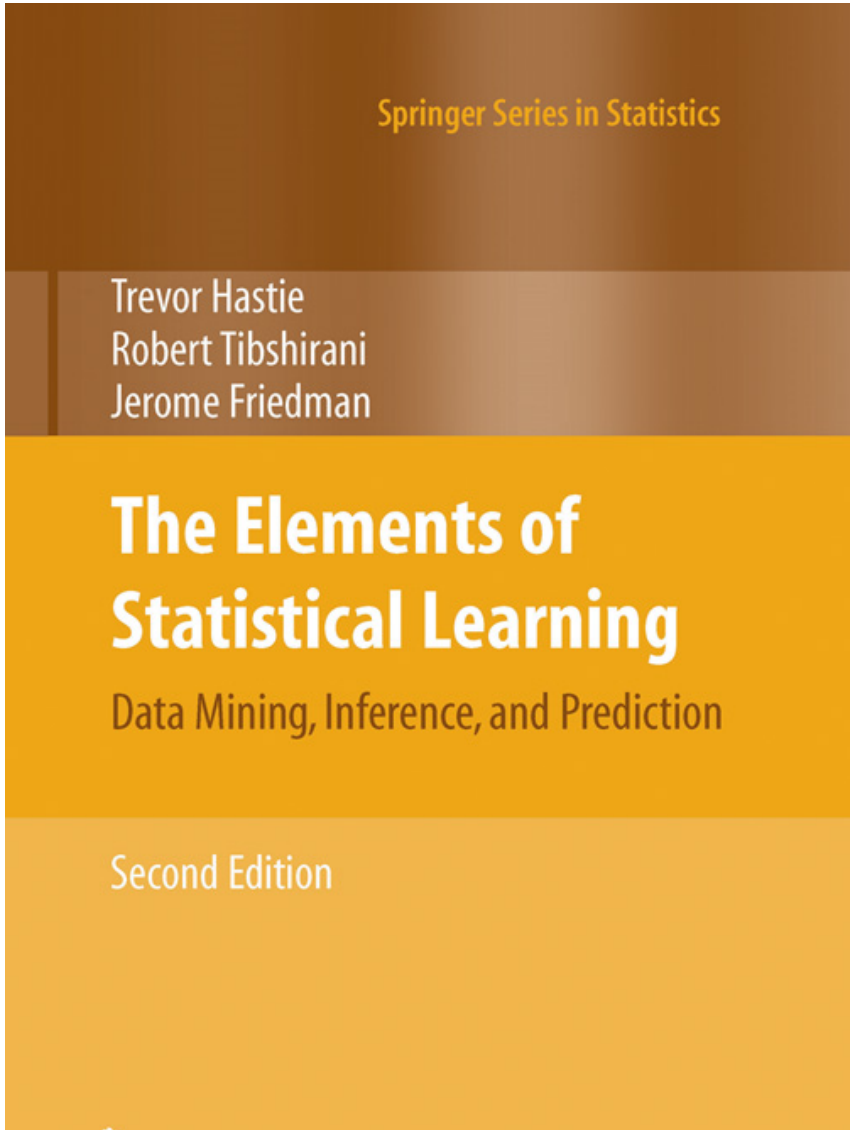
By the end of this course you will...

- Understand the statistical theory behind some of the most popular machine learning algorithms.
- Be able to apply machine learning algorithms to the analysis of text data.
- Be able to train and assess the performance of a variety of machine learning algorithms.
- Have the beginnings of a research paper which applies a machine learning algorithm to substantive problems.

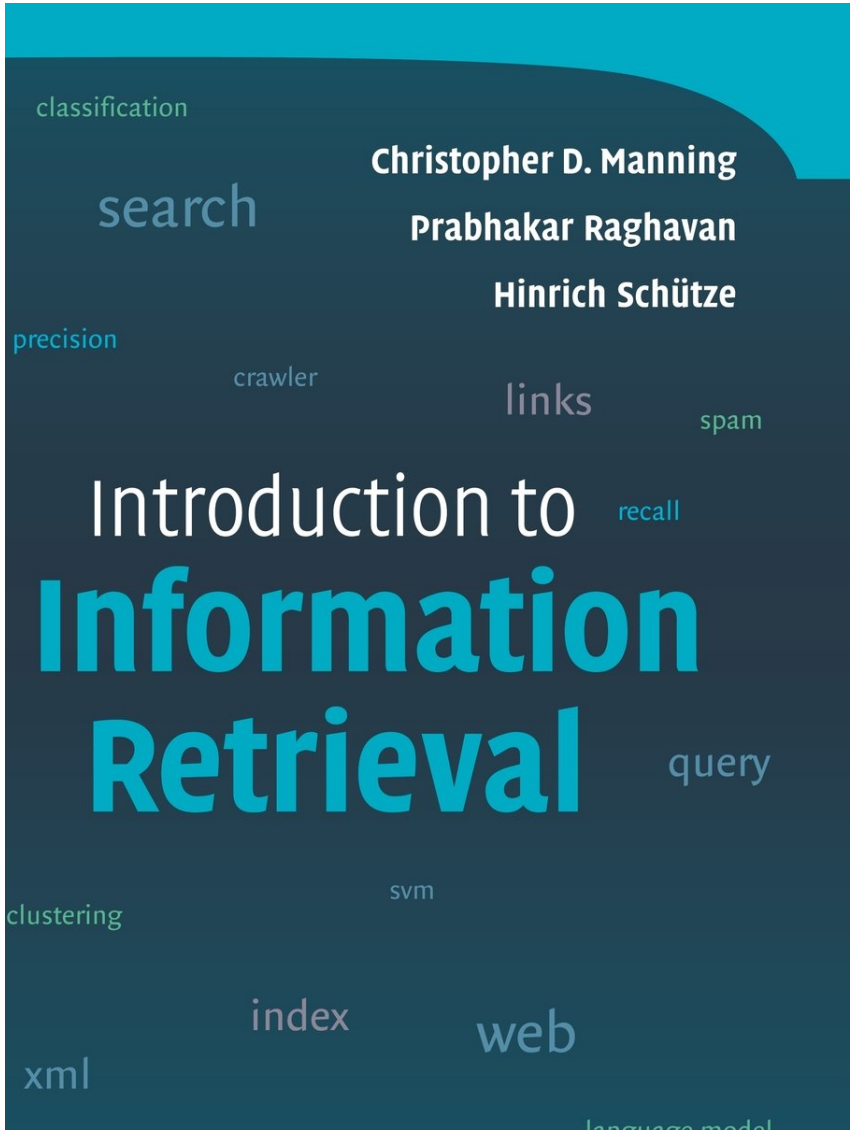
Textbooks


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R

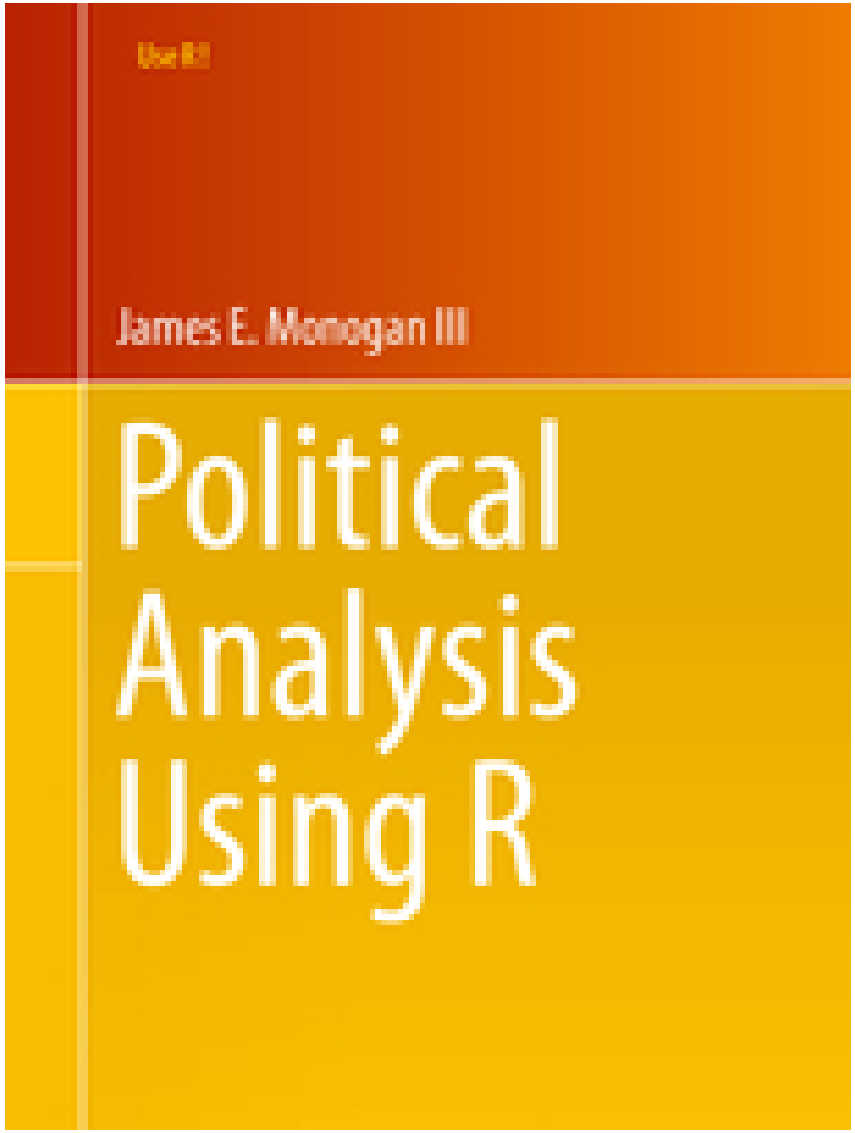
Textbooks


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition

Textbooks


classification
Christopher D. Manning
search
Prabhakar Raghavan
Hinrich Schütze
precision
crawler
links
spam
Introduction to
recall
Information
Retrieval
query
svm
clustering
index
web
xml

Textbooks



R

- `R` is a programming language and a statistical computing environment.
- There are two major components to `R`.
  1. **Code** – Instructions that you give to `R`.
  2. **Interpreter** – The software that reads the instructions and executes them.

R

- For this class you will need to download and install two things:
  1. The `R` interpreter https://cran.r-project.org/
  2. **R-Studio**– An integrated development environment for `R` which allows you simultaneously write **code** and send commands to the **interpreter**. https://www.rstudio.com/products/rstudio/#Desktop

R

- We will discuss `R` in more detail next week.

Syllabus and course website

- The syllabus and other course materials can be found on the course webpage here: http://scholar.harvard.edu/janastas/pols-8500-applied-machine-learning
- Requirements include
  1. Five problem sets.
  2. Course project.

Problem Sets

- Problem sets are 50% of your total grade.
- Problem sets will involve a combination of programming and mathematical proofs.
- I am tailoring the course to enable you to do machine learning in `R` but if you prefer using `Python`, you are welcome to do so.

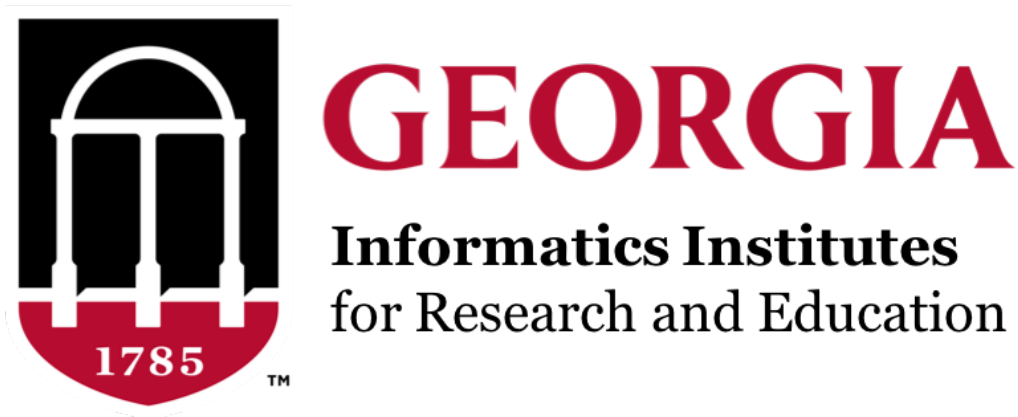
Problem Set Grading

- For the programming portions of the homework, I will provide you with starter code that you will complete.
- You must make sure that your code works before submitting it.
- If your code does not run, you will receive no credit for that particular problem.

Course project

- The course project is an opportunity to write a publishable-quality paper applying one or more of the machine learning techniques discussed in class.
- You can work alone or collaborate in groups of 2 or 3.
- It is worth 45% of your total grade and is composed of three parts:
  1. Proposal.
  2. Poster presentation.
  3. Paper.

Course project



- The poster presentation will be a poster session hosted by the Georgia Informatics Institute at the end of the semester in which members of UGA community will be invited to.

Course project examples

1. Train a classifier to identify speeches by Democrats and Republicans. Apply the trained classifier to assess the partisan leanings of other documents.
2. A study of the evolution of executive orders using topic models.
3. Measuring patterns of influence among states using state constitutional amendments and text analysis. methods.

About you

1. Name
2. Department
3. Program (BA,MA or PhD?)
4. A research topic that you are interested in.

Introduction to machine learning

- What is machine learning?
- Applications in the real world.
- Applications in social science.
- Supervised and unsupervised machine learning.

What is machine learning?

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

$$f : \arg\min_{f} \mathbb{E}[(Y - f(X))^2]$$

- Given function f that maps an input space $\mathcal{X}$ to an output space $\mathcal{Y}$
- Choose a f that minimizes the mean difference between the observed categories and the predicted .

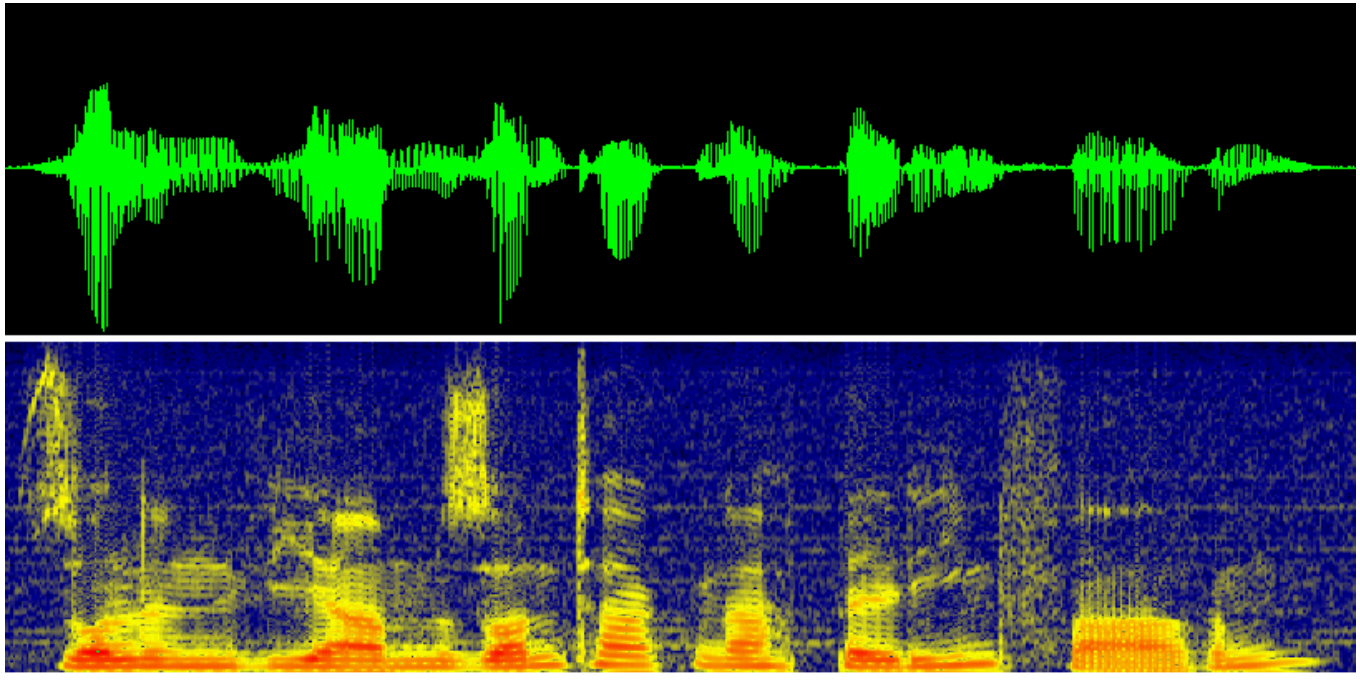
What is machine learning?

- In words, machine learning is the study of algorithms which make optimal predictions for a given task.
- **algorithm** – is a set of operations to be performed. In many ways similar to a cookbook.

Applications in the real world

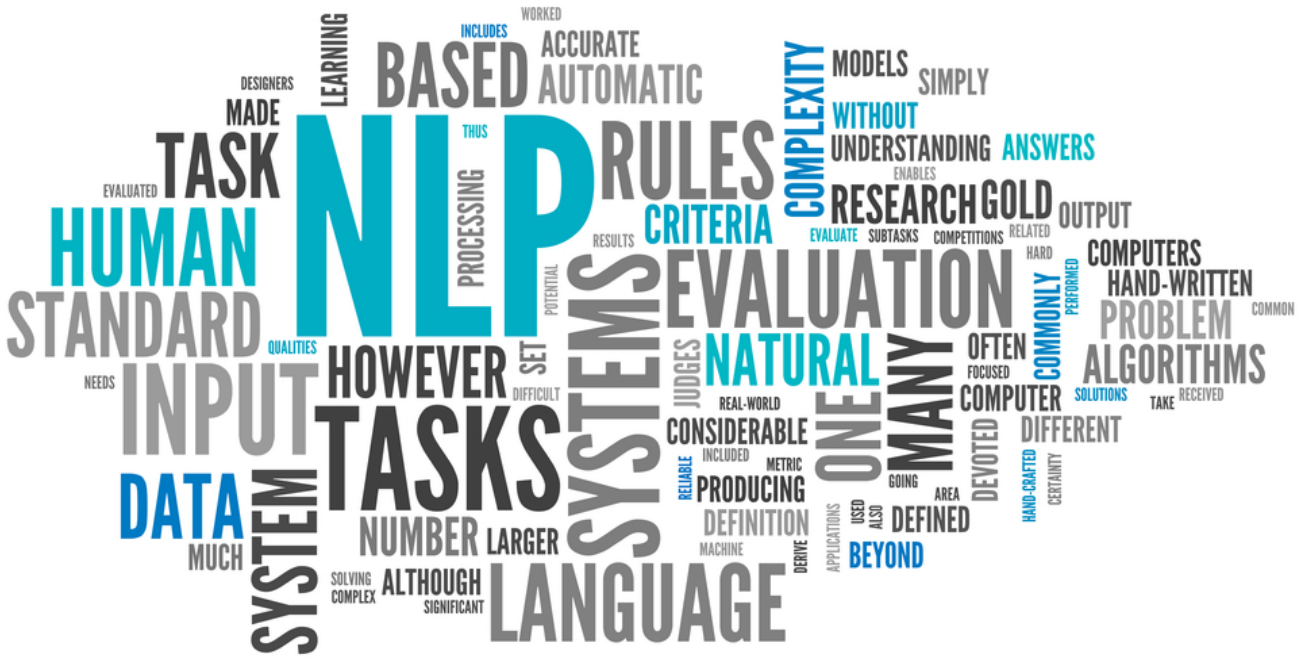
All of your recommender systems.

Applications in the real world



Speech recognition.

Applications in the real world



Natural language processing.

Applications in the real world

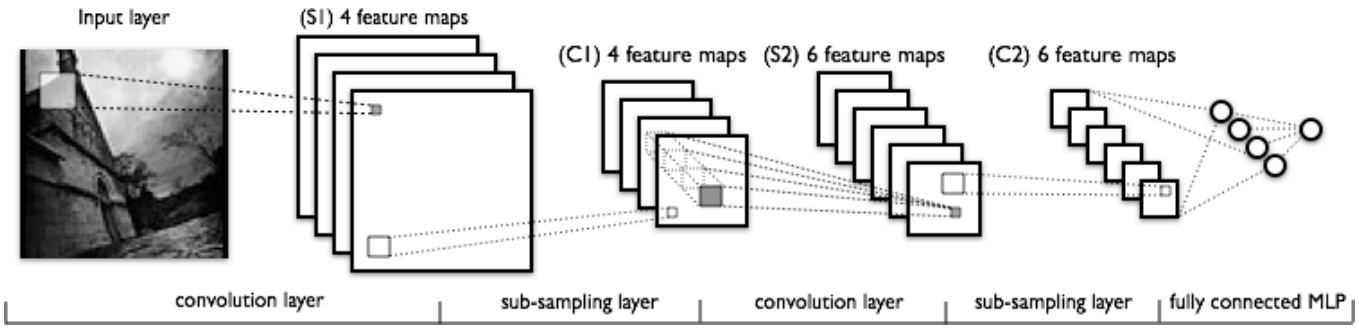
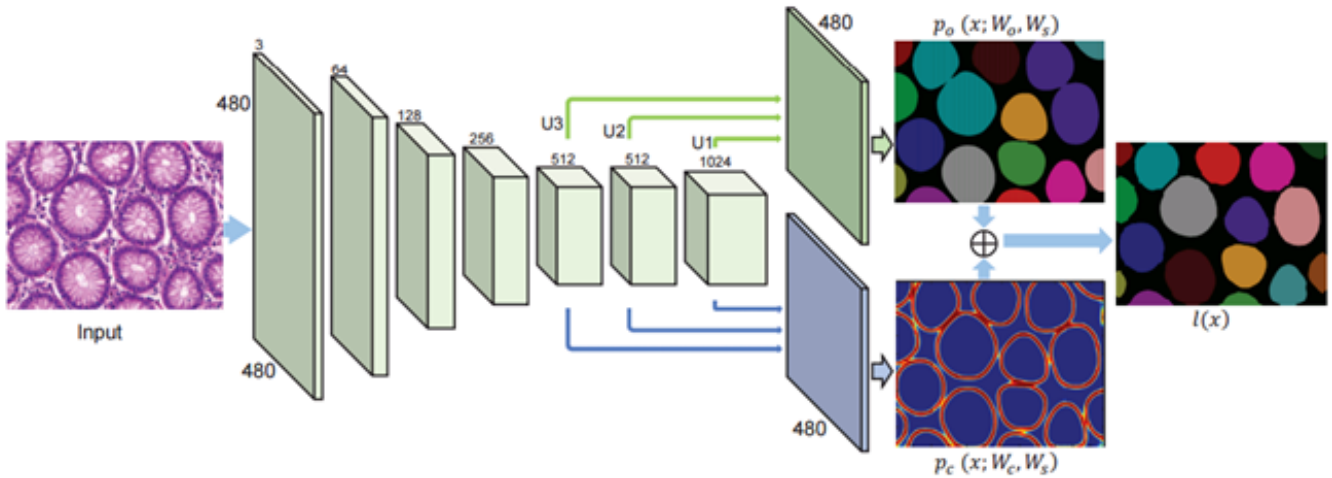


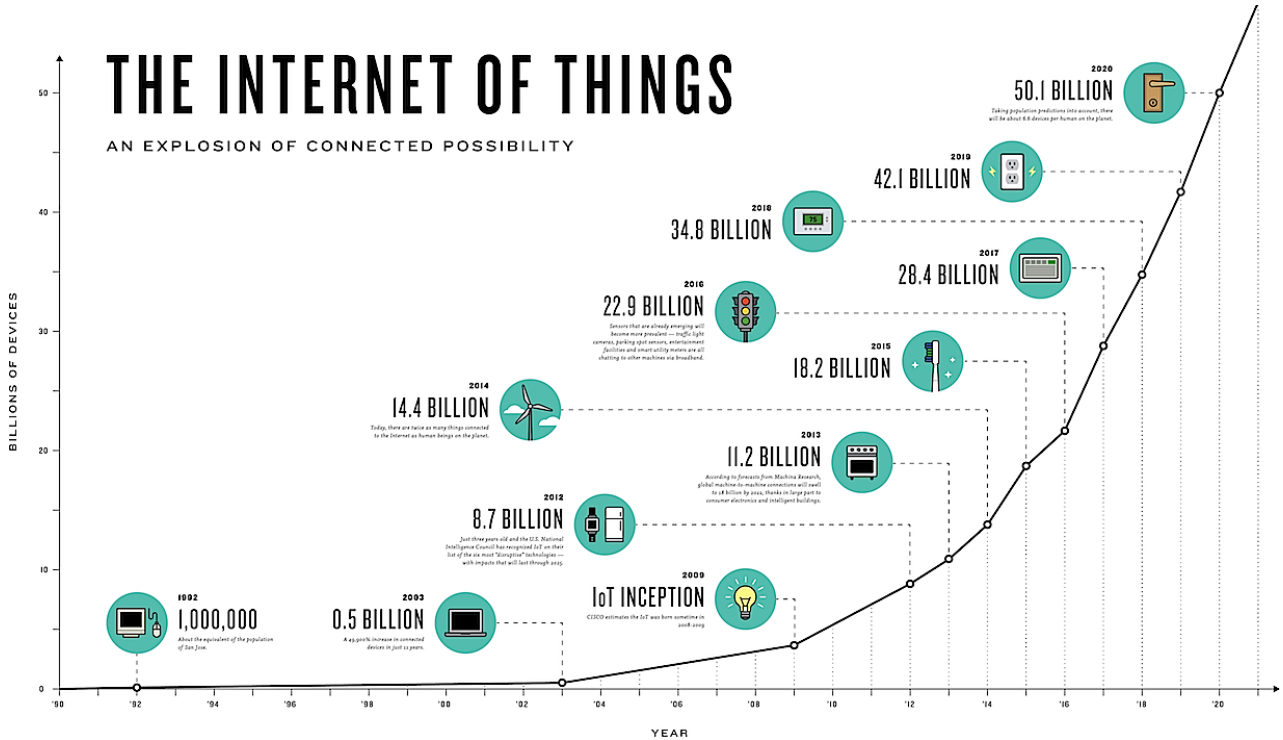
Image feature detection and object recognition.

Applications in the real world



Medical diagnoses.

Applications in the real world

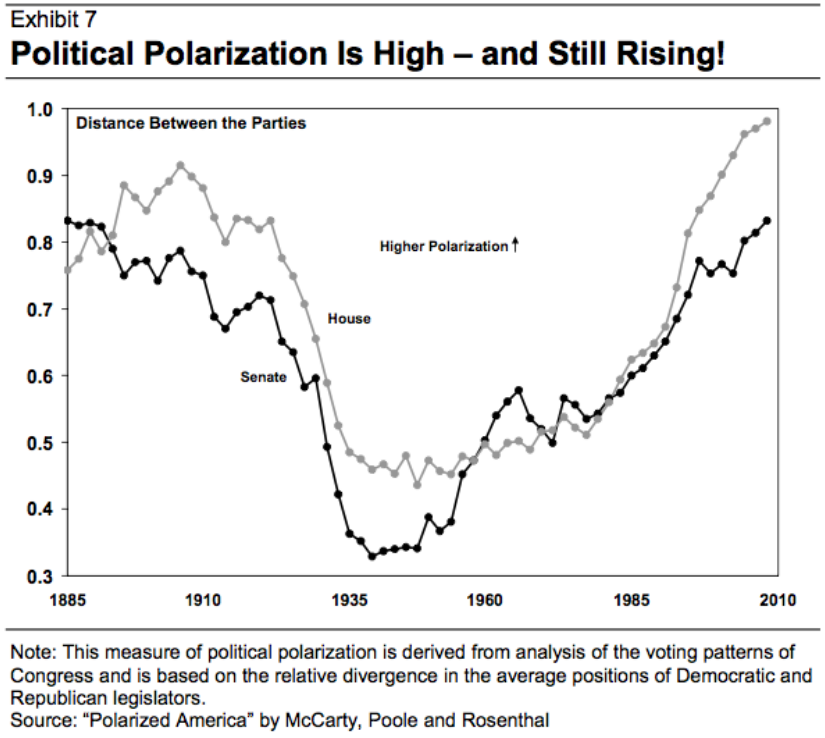


Internet of things.

Applications in social science

- Applications in social science $\subset$ Applications in the real world.
- Mostly used for natural language processing.
- In political science, for analysis of political documents.

Applications in social science

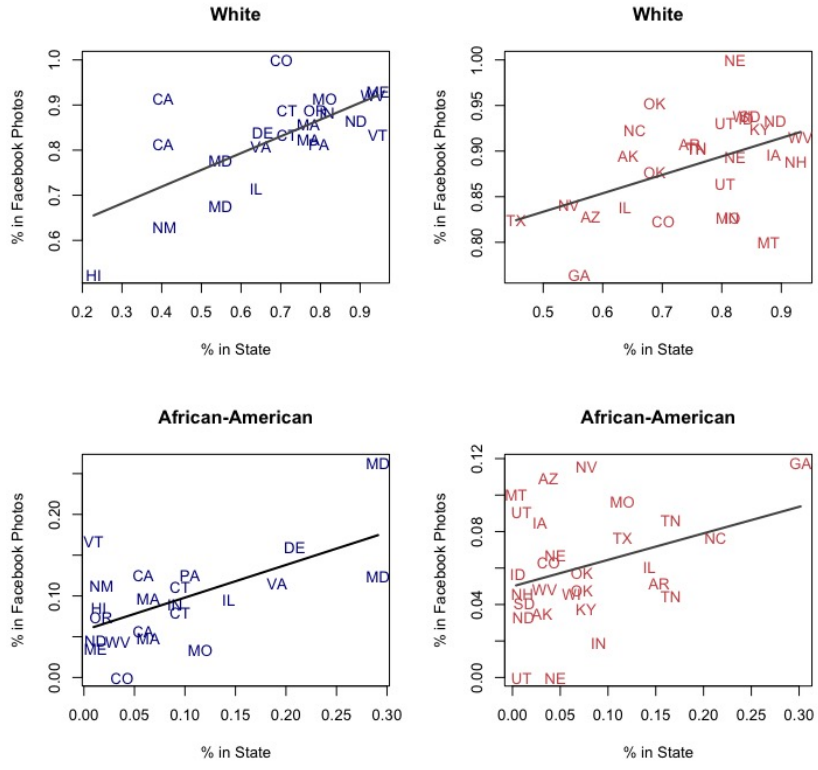


- Measuring of polarization in Congress using the NOMINATE dimensionality reduction technique.

Applications in social science

Language of partisanship.

Applications in social science



The use of images by politicians.

Supervised and unsupervised learning

Machine learning algorithms can either rely on data or theoretical constructs to perform classification.

Supervised learning – involves using data to "train" (estimate parameters for) a machine learning algorithm.

- **Unsupervised learning** – involves algorithms that can automatically classify data without requiring training data.

Supervised learning

- Examples include linear regression, naive Bayes, support vector machines, neural networks.
- Vast majority of social science and data applications use supervised learning algorithms.

Unsupervised learning

- Examples include K-Means clustering, Dirichlet processes, topic models, multidimensional scaling, principal components analysis, Google PageRank.
- Image segmentation, web search etc. etc.