Machine Learning for Social Sciences

Justin Grimmer

Associate Professor
Department of Political Science
University of Chicago

January 3rd, 2018

Machine Learning in the Social Sciences

- Discovery
- Measurement
- Causal Inference

Course Evaluation Plan

Three (Equal) Parts to Evaluation

1) 5 homeworks.
   - Collaborate with folks in class
   - But write up your own work
   - Goal: (1) deeper understanding of the statistical methods (2) develop programming skills and (3) learn how to apply techniques from class to your own work
2) Class Participation
3) Poster Session + Paper

Poster Session + Paper

Goal: create *publishable* research output

Work in groups (2-3 people), apply methods from the class

Sequence:

- Initial project selection/question: January 17th.
- Data set collected, ready to analyze: February 7th
- Initial analyses/Write Up: February 19th
- Final Meeting with me to discuss project: February 28th
- Poster Session: March 7th
- Paper due by the end of final exam period

I want to work with you to make publishable research

Course Content

Prerequisites:

1) Must have: Linear Regression, Mathematical Statistics, background in `R`, `Python` or related language
2) (Very) Nice to have: Likelihood Theory, Causal Inference, and related courses

Technical class:

- Hard work: time spent on programming, problem sets, and research
- Time consuming: please set aside time to work on this class
- Everyone can succeed

Questions: Smartest person in the room rule

Machine learning $\rightsquigarrow$ powerful, but important to recognize limitations

Online Advertisements

- Online ads: billions of revenue
- Last click attribution: ads "get credit" if last thing you see before you buy
- Goal: optimize probability my ad is the last one clicked

Optimized, but for the task you choose

Voter Targeting Decisions

Campaigns: exert effort to mobilize voters

- Voter lists, consumer data, and proprietary surveys to target
- Hersh 2015: limitations to voter file, depends on state
- Merge: hard to combine data from different sources
- Clean: hard to know if someone has moved or just not voting
- Target: hard to run experiment during campaign to determine who to target

You work with the data you have

Machine Learning and "Bias"

Machine learning methods can mitigate bias in decision making

- Kleinberg et al "Human Decisions and Machine Predictions" $\rightsquigarrow$ Make better bail decisions using machine learing
- Bansak et al $\rightsquigarrow$ machine learning places refugees in better areas

Machine learning methods can inherent (and amplify) biases in decision making

- Caliskan et al "Semantics derived automatically from language corpora contain human-like biases" $\rightsquigarrow$ machine learning can inherent human biases

Machine learning is not a panacea for human biases

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
    - Hard to find

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
    - Hard to find
    - Time Consuming
    - Not generalizable (each new data set...new coding scheme)
    - Difficult to store/search

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)
  - Difficult to store/search
  - Idiosyncratic to coders/researcher

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
    - Hard to find
    - Time Consuming
    - Not generalizable (each new data set...new coding scheme)
    - Difficult to store/search
    - Idiosyncratic to coders/researcher
    - Statistical methods/algorithms, computationally intensive

A post-2000's view of text in social science:

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...
- Facebook posts, tweets, emails, cell phone records, ...

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...
- Facebook posts, tweets, emails, cell phone records, ...
- Newspapers, magazines, news broadcasts, ...

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...
- Facebook posts, tweets, emails, cell phone records, ...
- Newspapers, magazines, news broadcasts, ...
- Foreign news sources, treaties, sermons, fatwas, ...

Why?

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $=$ 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions
  - Press Releases

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions
  - Press Releases

What Can Text Methods Do?

Haystack metaphor:

What Can Text Methods Do?

Haystack metaphor: Improve Reading

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
    - Humans: amazing (Straussian political theory, analysis of English poetry)
    - Computers: struggle

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
    - Humans: amazing (Straussian political theory, analysis of English poetry)
    - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\rightsquigarrow$ Organizing hay stack

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts$\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing$\rightsquigarrow$ largely what we'll discuss today

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing $\rightsquigarrow$ largely what we'll discuss today

What automated text methods don't do:

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing $\rightsquigarrow$ largely what we'll discuss today

What automated text methods don't do:

- Develop a comprehensive statistical model of language
- Replace the need to read
- Develop a single tool + evaluation for all tasks

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?
- Who is the `We`?

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?
- Who is the `We`?
- What is the `mountaintop` (literal?)

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?
- Who is the `We`?
- What is the `mountaintop` (literal?)

Texts $\rightsquigarrow$ high dimensional, not self contained

Texts are Surprisingly Simple

(Lamar Alexander (R-TN) Feb 10, 2005)

Word	No. Times Used in Press Release
department	12
grant	9
program	7
firefight	7
secure	5
homeland	4
fund	3
award	2
safety	2
service	2
AFGP	2
support	2
equip	2
applaud	2
assist	2

Texts are Surprisingly Simple (?)

`US Senators Bill Frist (R-TN) and Lamar Alexander (R-TN) today applauded the U S Department of Homeland Security for awarding a $8,190 grant to the Tracy City Volunteer Fire Department under the 2004 Assistance to Firefighters Grant Program's (AFGP) Fire Prevention and Safety Program...`

Not just for "big data"

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)

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) $\approx 4.75 \times 10^{115}$ 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) $\approx 4.75 \times 10^{115}$ 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)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs

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)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  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)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  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) $\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times$

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)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$

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)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$ 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) $\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14{,}000{,}000{,}000)$ years

Machine Learning methods can help with even small problems

Course Plan

- Preliminaries: Acquiring Text and Feature Engineering
- Discovery
  - Regular Expressions and Vector Space Model of Text
  - Unsupervised Clustering
  - Topic Models
  - Embeddings
  - Fictitious Prediction Problems
- Measurement
  - Hand Coding
  - Dictionary Methods
  - LASSO and Ridge
  - Naive Bayes and ReadMe
  - Boosting, Bagging, and Ensembles
  - Structural Topic Models for Measurement
- Causal Inference
  - Text as Intervention
  - Text as Response and as Covariate

Five principles for Machine Learning and Social Science

Social Science Inferences are Necessarily Sequential

Social Science Inferences are Necessarily Sequential

Story of KPR

The g (codebook) function is central: Text as data methods are about compression

The g (codebook) function is central: Text as data methods are about compression

Text based experiments

There is no general theory of language, nor globally best method.

Text as Data Methods Do Not Replace Humans, They Augment Them

Validate,Validate,Validate