ORIE 4741: Learning with Big Messy Data

Limitations and Dangers of Predictive Analytics

Professor Udell

Operations Research and Information Engineering
Cornell

November 30, 2021

Announcements 11/30/21

- section this week: causal inference
- for ORIE 5741: project presentations due (as video) Friday 12/3/21 11:59pm
- project final report due Sunday 12/5/21 11:59pm
- homework 6 due 9:15am Tuesday 12/7/21
- project peer review due Sunday 12/12/21 11:59pm

Outline

Election modeling

Electoral modeling in practice

Electoral modeling: good, bad, and ugly

Weapons of Math Destruction

Analytics for presidential campaigns

goal: allocate limited resources to optimize electoral vote

- each state gets votes proportional to population (about 1 vote / 600K residents) + k
- most states allocate all votes to winner of statewide popular vote

where $k =$

A. 0
B. 1
C. 2
D. 5
E. 10

Why an electoral college?

the electoral college

- is a compromise from the constitutional convention of 1787 to get small states to join the United States
- increases power of rural states
- concentrates campaign attention in very few "battleground" states
- decides the winner of the election
  - $\neq$ popular vote winner in 1876, 1888, 2000, 2016

Why an electoral college?

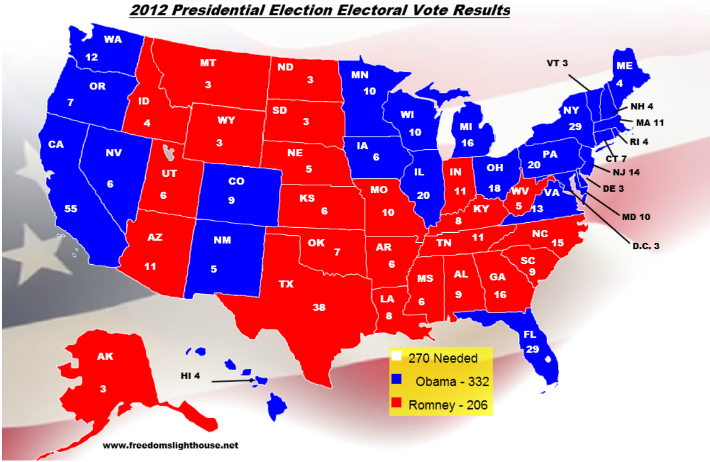
the electoral college

- is a compromise from the constitutional convention of 1787 to get small states to join the United States
- increases power of rural states
- concentrates campaign attention in very few "battleground" states
- decides the winner of the election
  - $\neq$ popular vote winner in 1876, 1888, 2000, 2016

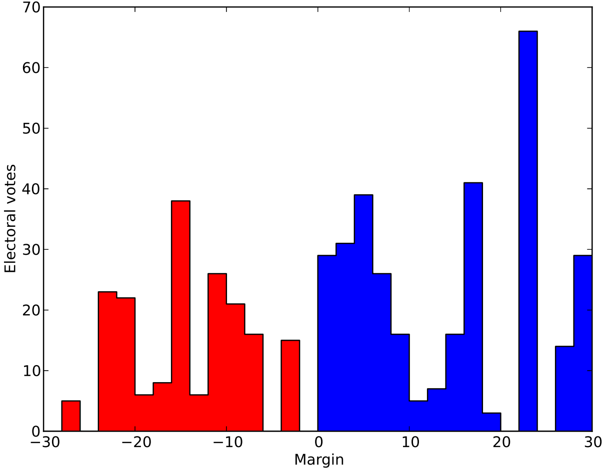
can we get rid of the electoral college?

- via constitutional amendment: need 2/3 of states to agree
- via state compact: need 105 more electoral votes

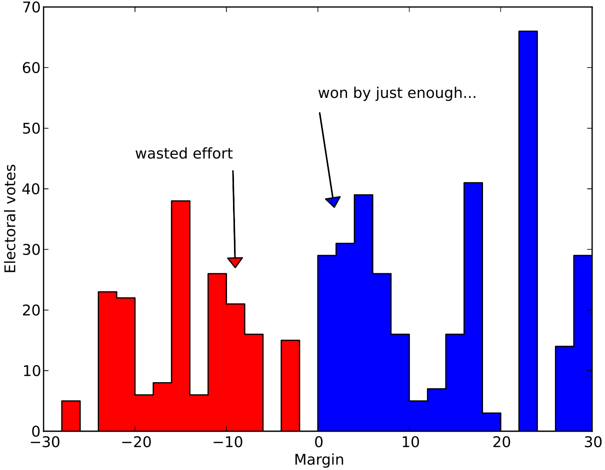
2012 electoral map



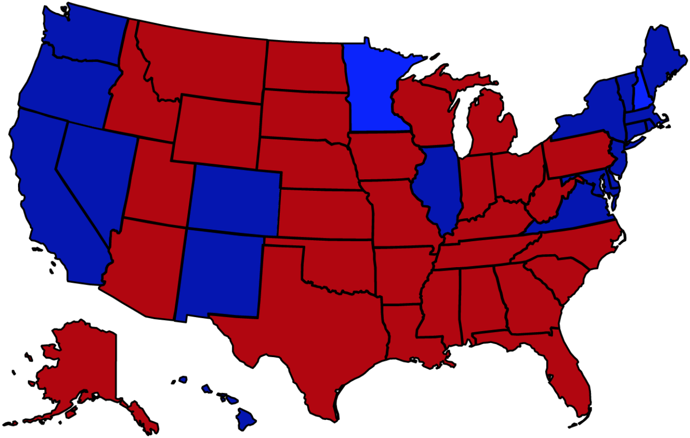
Optimization on the Obama 2012 campaign



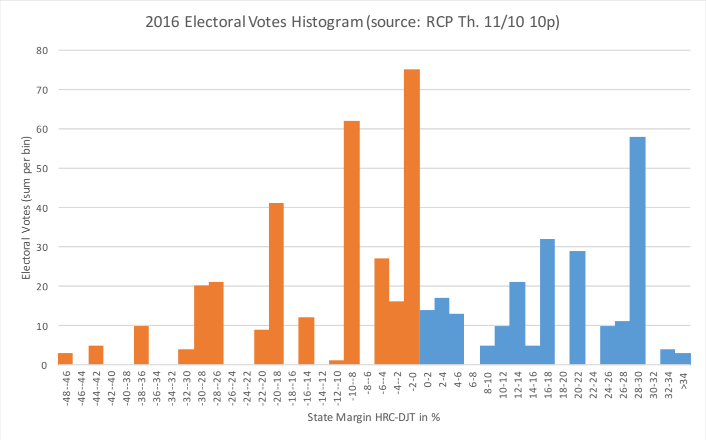
Optimization on the Obama campaign



2016 electoral map



Less (effective) optimization on the 2016 Clinton campaign



Election 2020

big questions:

- polls were **quite** wrong (a bit less than 2016):
  9% predicted margin $\rightarrow$ around 4–5% actual margin (national popular vote)
- is the US electoral system democratic? (good/bad?)
  - first-past-the-post
  - electoral college
  - senate
  - supreme court
  - relevant reading: They Don't Represent Us, by Lawrence Lessig

What choices can a campaign make?

campaigns can control

- which states the candidate visits
- how many ads for the candidate are produced
  - TV
  - yard sign
  - internet
  - facebook
  - merchandise
- voter registration targeting
- get-out-the-vote (GOTV) targeting
- candidate policy statements

to maximize probability of electoral win

Outline

Data for political campaigns

data sources:

- (public) voter files
- (private) purchased information
- polling data
  - previous years
  - primary election
  - surveys (phone, internet, in-person, ...)
  - exit polls
- economic data

Analytics for political campaigns

goal: allocate limited resources to optimize electoral vote

three key components:

- support
- persuasion
- turnout

Levels of modeling

three kinds of models:

- agent level predictive model (Obama 2012)
- demographic level predictive model (NYT turnout model)
- aggregated polls-based model (Nate Silver and 538)

Agent level predictive model

age	gender	state	income	education	voted?	support	$\cdots$
29	F	CT	\$53,000	college	yes	Biden	$\cdots$
57	?	NY	\$19,000	high school	yes	?	$\cdots$
?	M	CA	\$102,000	masters	no	Trump	$\cdots$
41	F	NV	\$23,000	?	yes	Trump	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Demographic level predictive model

county	demographics	vote %	support %	$\cdots$
tompkins	white male	$\vdots$	$\vdots$	$\vdots$
tompkins	asian female	$\vdots$	$\vdots$	$\vdots$
tompkins	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Aggregated polls-based model

weighted average predictor:

- poll peopple from each demographic group i to measure average support $\bar{v}_i$
- predict turnout t_i for each group i using historical data
- compute weighted average

$$\text{predicted vote share} = \frac{1}{\sum_i t_i} \sum_i t_i \bar{v}_i$$

Aggregated polls-based model

weighted average predictor:

- poll peopple from each demographic group i to measure average support $\bar{v}_i$
- predict turnout t_i for each group i using historical data
- compute weighted average

$$\text{predicted vote share} = \frac{1}{\sum_i t_i} \sum_i t_i \bar{v}_i$$

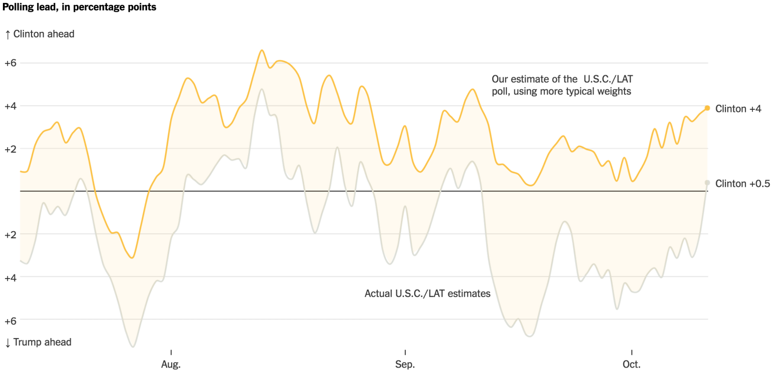
sources of error:

- statistical error
- systematic error
- non-response bias

How to choose the weights?

two ingredients produce weights:

- definition of demographic group
- turnout prediction per group



The trouble with polls

Q: do you pick up your phone when an unknown number calls?

A. always
B. usually
C. sometimes
D. never

The trouble with polls

Q: do you pick up your phone when an unknown number calls?

A. always
B. usually
C. sometimes
D. never

A: . . . varies by demographic; eg, older americans are more likely to answer

The trouble with polls

Q: do you pick up your phone when an unknown number calls?

A. always
B. usually
C. sometimes
D. never

A: . . . varies by demographic; eg, older americans are more likely to answer

Q: could we build a model to predict which of you picks up?

The trouble with polls

Q: do you pick up your phone when an unknown number calls?

A. always
B. usually
C. sometimes
D. never

A: . . . varies by demographic; eg, older americans are more likely to answer

Q: could we build a model to predict which of you picks up?

The trouble with polls

Q: do you pick up your phone when an unknown number calls?

A. always
B. usually
C. sometimes
D. never

A: . . . varies by demographic; eg, older americans are more likely to answer

Q: could we build a model to predict which of you picks up?

Q: are people who respond to polls like people who don't?

The trouble with polls

Q: do you pick up your phone when an unknown number calls?

A. always
B. usually
C. sometimes
D. never

A: . . . varies by demographic; eg, older americans are more likely to answer

Q: could we build a model to predict which of you picks up?

Q: are people who respond to polls like people who don't?

anecdote from 2016:

> *There is a 19-year-old black man in Illinois who has no idea of the role he is playing in this election.*
>
> *He is sure he is going to vote for Donald J. Trump.*
>
> *In some polls, he's weighted as much as 30 times more than the average respondent, and as much as 300 times more than the least-weighted respondent.*

Correct biased sample

two types of people

- type A always fill out all questions
- type B leave question 3 blank half the time

question 1	question 2	question 3	question 4	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	?	no	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	?	no	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	

estimate population mean of question 3

- excluding missing entries: 2.5
- imputing missing entries: 2

Correct biased sample

two types of people

- type A always fill out all questions
- type B leave question 3 blank half the time

question 1	question 2	question 3	question 4	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	?	no	$\cdots$
2.7	yes	4	yes	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	1	no	$\cdots$
9.2	no	?	no	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	

estimate population mean of question 3 if the type B people have two subtypes:

- one that answers "1" to question 3
- another that doesn't answer, but whose true answer is "27"

How does this apply to election models?

simple model: suppose that in each demographic group,

- there are some Trump and some Biden supporters
- the Trump supporters respond to pollsters at lower rates (or lie about their support)

there is **no way** to detect this from polling data!

How does this apply to election models?

simple model: suppose that in each demographic group,

- there are some Trump and some Biden supporters
- the Trump supporters respond to pollsters at lower rates (or lie about their support)

there is **no way** to detect this from polling data!

confidence intervals: (computed eg via bootstrap or analytical methods)

- account for **statistical** error
- do not account for **systematic** error

Conditions for unbiased estimation

to ensure that estimate is unbiased, need **outcome** to be independent of **missingness** conditional on **covariates**

e.g., estimate for predicted vote share is unbiased if (turnout model is correct and)

$$\text{support for Trump} \perp \text{non-response} \mid \text{demographics}$$

Dealing with systematic bias

problem with systematic bias:

- even if you know it **exists**, you don't know **how much**!

Dealing with systematic bias

problem with systematic bias:

- even if you know it **exists**, you don't know **how much**!

modeling systemic bias?

Dealing with systematic bias

problem with systematic bias:

- even if you know it **exists**, you don't know **how much**!

modeling systemic bias?

- use the bias from previous years to infer bias this year

Dealing with systematic bias

problem with systematic bias:

- even if you know it **exists**, you don't know **how much**!

modeling systemic bias?

- use the bias from previous years to infer bias this year

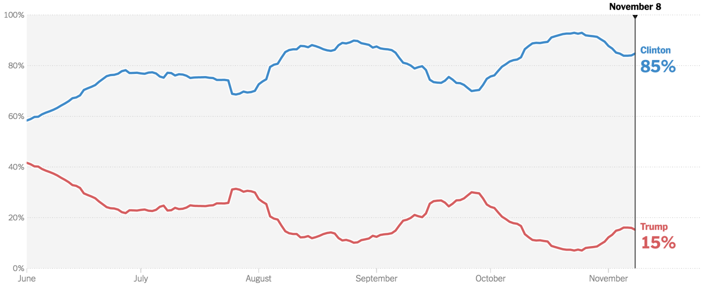
problem: level of bias may vary with other (never-before-seen) factors, *e.g.*(from 2016 and 2020)

- female candidate
- candidate with no experience
- candidate who endorses unconstitutional policies
- COVID
- shutdowns

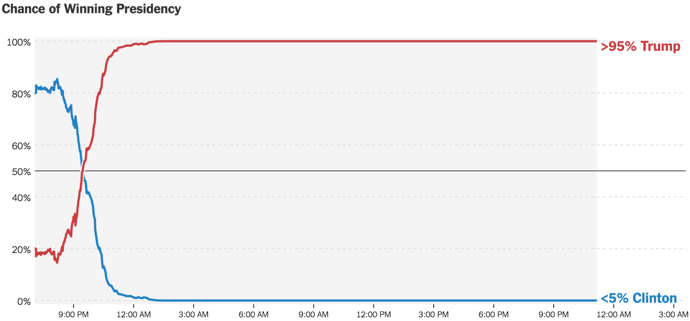
we have no data to estimate these effects!

Outline

NYT 5 month prediction



NYT night-of prediction



Assessing the quality of analytics

as of the Monday night before the election, 11/8/16, predictions:

- **good.** Nate Silver and 538: 65% Clinton to 35% Trump
- **bad.** New York Times: 84% Clinton to 16% Trump
- **ugly.** Princeton Election Consortium: 99% Clinton to 1% Trump (!)

Assessing the quality of analytics

as of the Monday night before the election, 11/8/16, predictions:

- **good.** Nate Silver and 538: 65% Clinton to 35% Trump
- **bad.** New York Times: 84% Clinton to 16% Trump
- **ugly.** Princeton Election Consortium: 99% Clinton to 1% Trump (!)

why did pollsters predict the 2016 election so poorly?
`http://fivethirtyeight.com/features/the-polls-missed-trump-we-asked-pollsters-why/`

Assessing the quality of analytics

as of the Monday night before the election, 11/8/16, predictions:

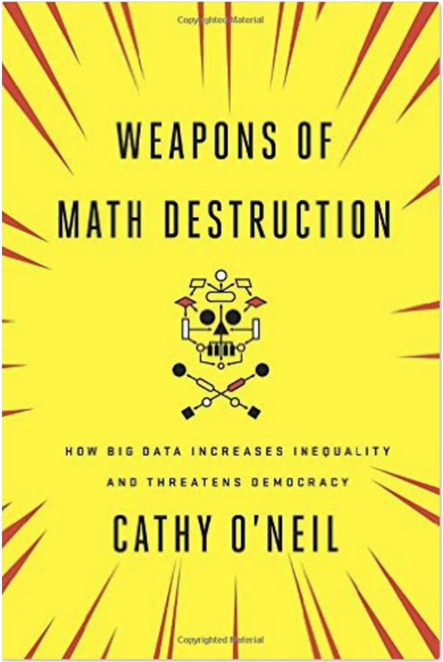
- **good.** Nate Silver and 538: 65% Clinton to 35% Trump
- **bad.** New York Times: 84% Clinton to 16% Trump
- **ugly.** Princeton Election Consortium: 99% Clinton to 1% Trump (!)

why did pollsters predict the 2016 election so poorly?
`http://fivethirtyeight.com/features/the-polls-missed-trump-we-asked-pollsters-why/`

from the data we have, can we conclude predictions did poorly?

need to **calibrate** prediction accuracy

Outline


Copyrighted Material
WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL
Copyrighted Material

Weapons of Math Destruction

what is a WMD? a predictive model

- whose outcome is **not** easily measurable
- whose predictions can have negative consequences
- that creates self-fulfilling (or defeating) feedback loops

CERN is not a WMD

CERN is not a WMD

why are predictive models from CERN ok?

- they are running an experiment
- they rigorously separate train and test sets
- they have a lot of data and can get more
- they have generative models that quantitatively agree with experimental measurements
- they have multiple experiments that measure the same phenomenon of interest using completely different tools

College rankings are a WMD

why are college rankings a WMD?

College rankings are a WMD

why are college rankings a WMD?

- can't measure "test error" for college rankings
- high rankings **cause** college "quality" to increase: attracts
  - students
  - faculty
  - donors
- college rankings shape behaviour and pervert incentives
  - tuition increases
  - spending on sports

Parole models are a WMD

- parole models predict **recidivism** (probability of committing another crime)
- automatically decide if a prisoner can be released early

Parole models are a WMD

- parole models predict **recidivism** (probability of committing another crime)
- automatically decide if a prisoner can be released early

why are parole models a WMD?

Parole models are a WMD

- parole models predict **recidivism** (probability of committing another crime)
- automatically decide if a prisoner can be released early

why are parole models a WMD?

- what data do they use? race? neighborhood?
- keeping people in prison longer may **cause** them to recidivate
    - hard to find a job
    - increase familiarity with crime
    - effects may spread within neighborhoods or communities

Are election models a WMD?

are (presidential) election models a WMD?

- is the outcome measurable?
  - yes, but only every four years!
  - or more frequently, but with unquantifiable systematic error
- do predictions affect results?
  - via turnout: yes
  - via support: ?
  - via persuasion: ?
- do predictions create negative feedback cycles?
  - yes: contacting only people likely to be on your side deepens extremism
  - using race as a feature thus deepens racial divides

Don't create WMDs

is your project a WMD?

- are outcomes hard to measure?
- could its predictions harm anyone?
- could it create a feedback loop?