

# data science @ NYT

Chris Wiggins

Aug 8/9, 2016

# Outline

1. overview of DS@NYT

# Outline

1. overview of DS@NYT
2. prediction + supervised learning

# Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL

# Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL
4. description + inference

# Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL
4. description + inference
5. (if interest) designing data products

## 0. Thank the organizers!



Figure 1: prepping slides until last minute

## Lecture 1: overview of ds@NYT

data science @ The New York Times

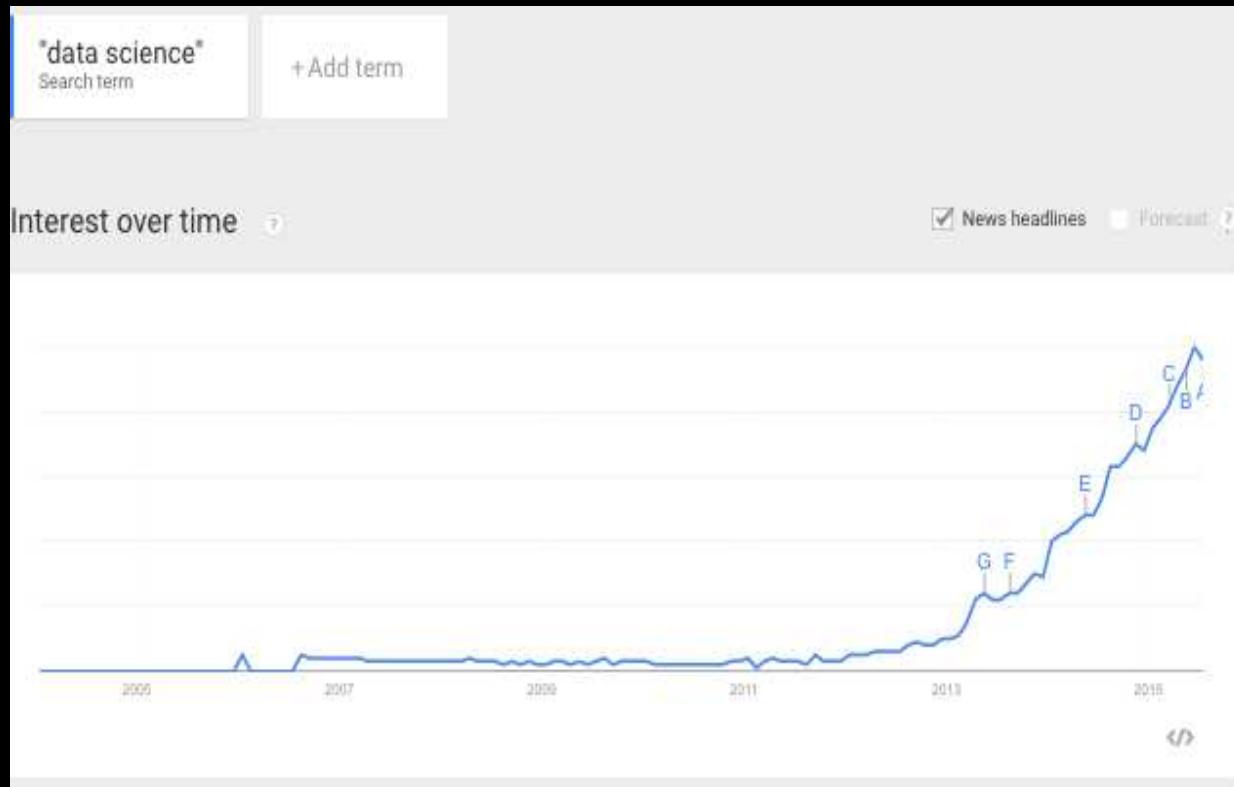


chris.wiggins@columbia.edu  
chris.wiggins@nytimes.com  
@chrishwiggins

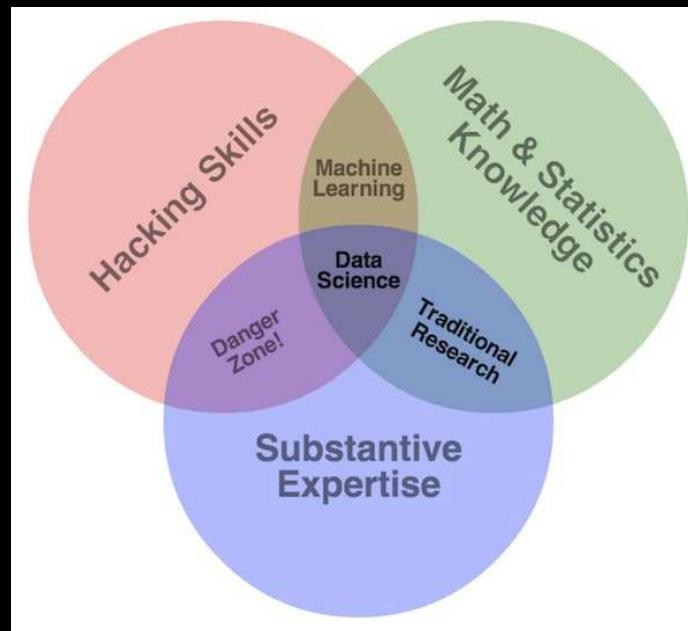
references: <http://bit.ly/stanf16>

data science

# data science: searches



## data science: mindset & toolset



drew conway, 2010

modern history:  
2009





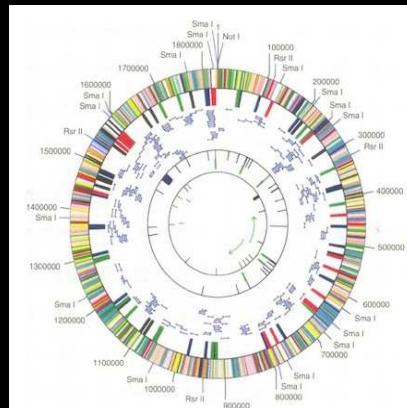
## Information Platforms and the Rise of the Data Scientist

At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn't quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization in a clear and concise fashion. To capture the skill set required to perform this multitude of tasks, we created the role of "Data Scientist."

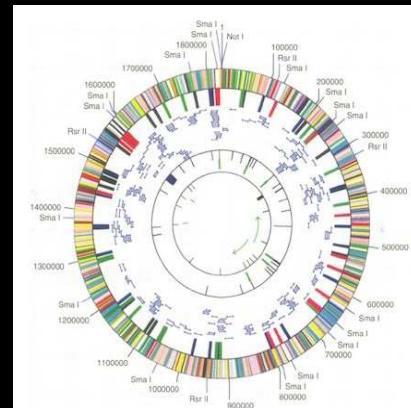
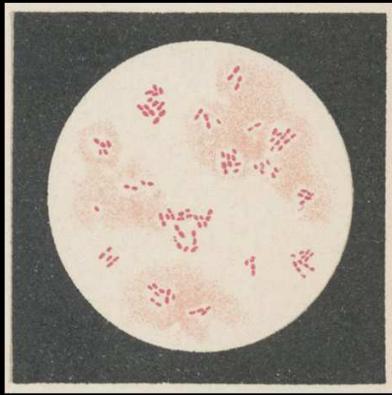
2009



## biology: 1892 vs. 1995

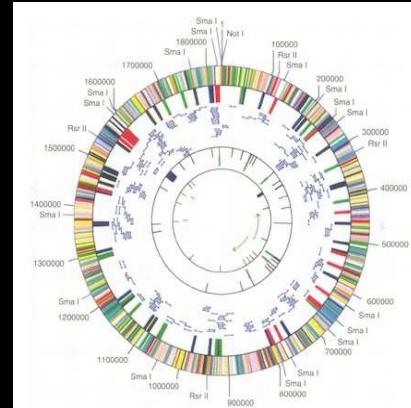


biology: 1892 vs. 1995



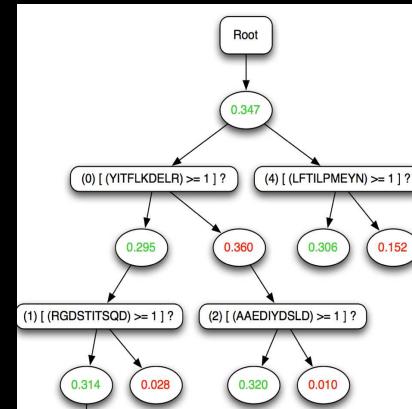
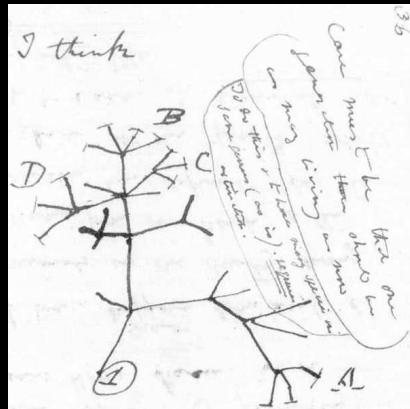
biology changed for good.

biology: 1892 vs. 1995



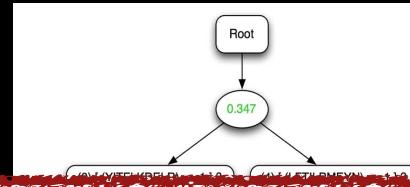
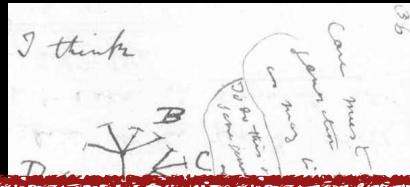
new toolset, new mindset

## genetics: 1837 vs. 2012



ML toolset; data science mindset

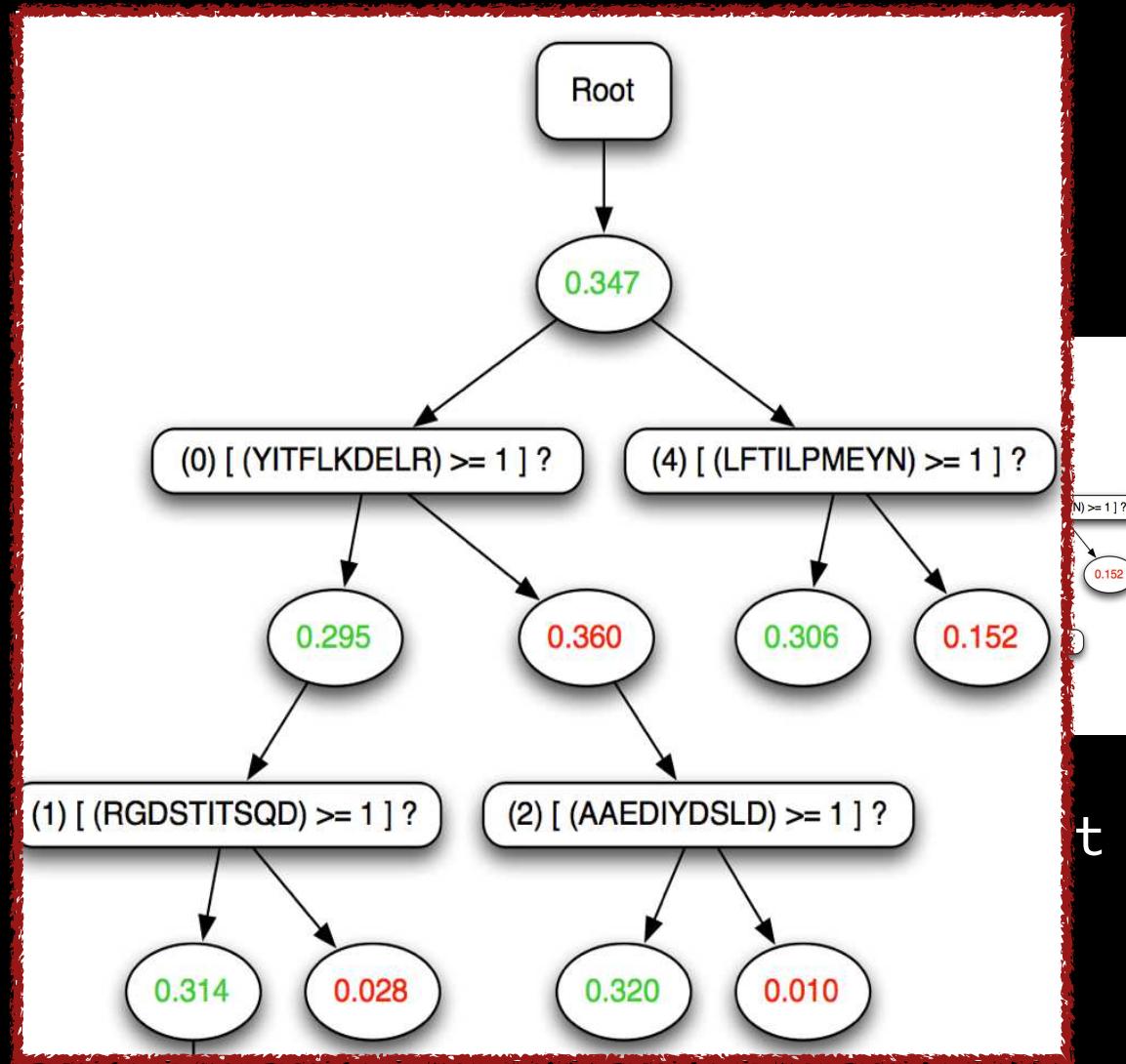
genetics: 1837 vs. 2012



*Statistical Science*  
2001, Vol. 16, No. 3, 199–231

# Statistical Modeling: The Two Cultures

Leo Breiman



*data science: mindset & toolset*



1851

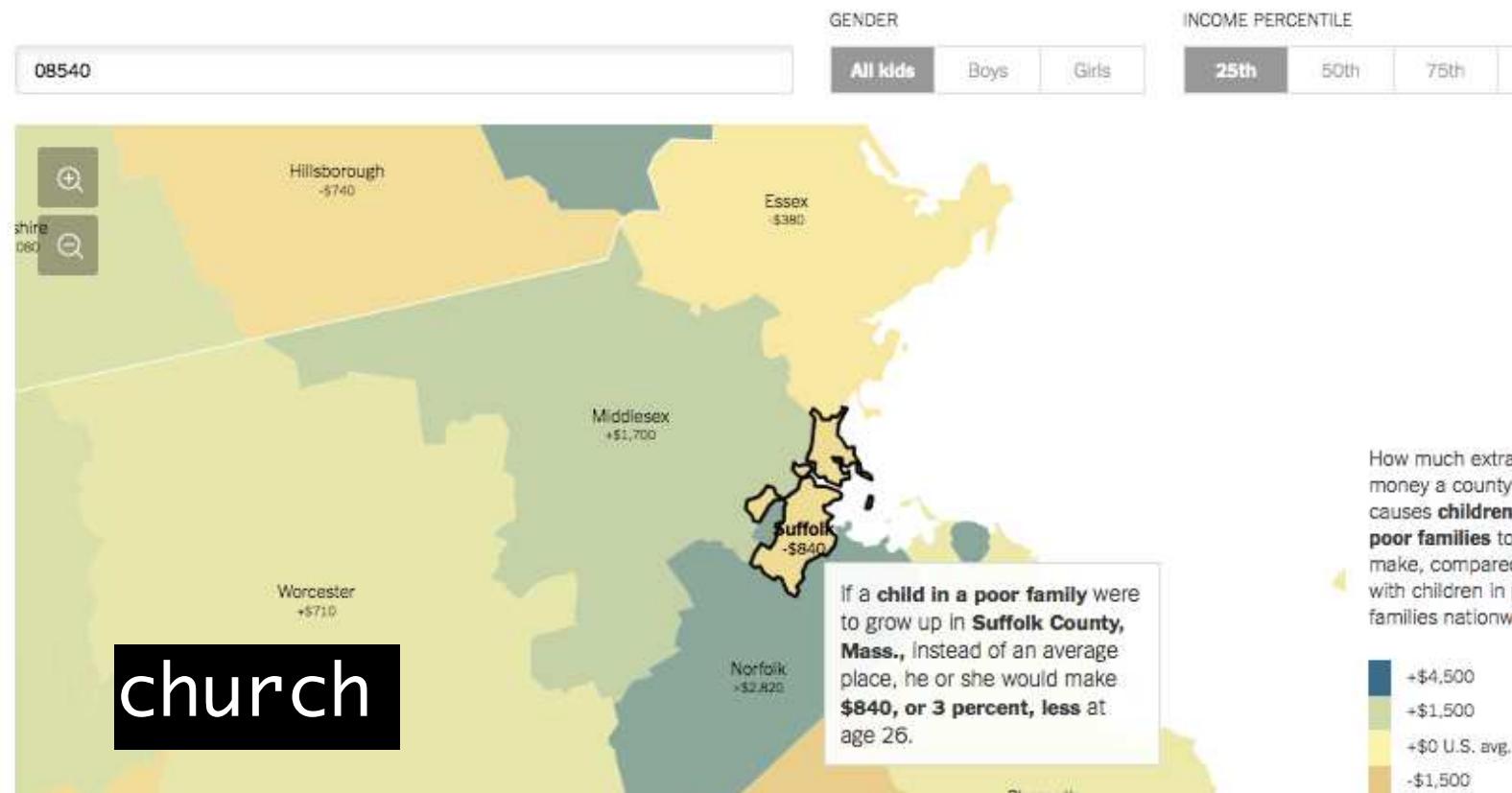
news: 20th century

church

state

## The Best and Worst Places to Grow Up: How Your Area Compares

Children who grow up in some places go on to earn much more than they would if they grew up elsewhere. [MAY 4, 2015](#) | [RELATED ARTICLE](#)



TheUpshot/leo-senate-model

<https://github.com/TheUpshot/leo-senate-model>

**GitHub** This repository Search or type a command

PUBLIC TheUpshot / leo-senate-model

church

Code and data for The Upshot's Senate model. <http://www.nytimes.com/newsgraphics/2014/senate-model/>

12 commits 1 branch 0 releases 3 contributors

branch: master / leo-senate-model / +

changing default parameters

joshkatz authored 2 hours ago latest commit 30e1af96c9

File	Description	Time
data-publisher	Include directories required for the script to generate output	11 hours ago
fundamentals	Rename file (.r -> .R) for case-sensitive filesystems (e.g. Linux extN).	11 hours ago
model	Remove dependence on the authors' directory structure	11 hours ago
output	Include directories required for the script to generate output	11 hours ago
.gitignore	Leo lives	15 hours ago
LICENSE	Like grownups	4 hours ago
README.md	added sample data output to README.md	8 hours ago
master-public.R	changing default parameters	2 hours ago

[www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html](http://www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html)

U.S.

# 2014 The Year in Interactive Storytelling, Graphics and Multimedia

Multimedia Stories | Data Visualization

[www.nytimes.com/newsgraphics/2013/12/30/year-in-interactive-storytelling/](http://www.nytimes.com/newsgraphics/2013/12/30/year-in-interactive-storytelling/)

The New York Times

# 2013: The Year in Interactive Storytelling

Multimedia Stories | Data Visualization | Explanatory Graphics | Breaking News | Visual and Interactive Features

## Multimedia Stories

From a ship in the South China Sea to the cost of health care in the United States, the range of subjects here is broad, but the common thread is the form of storytelling — an integration of text, video, photography and graphics.



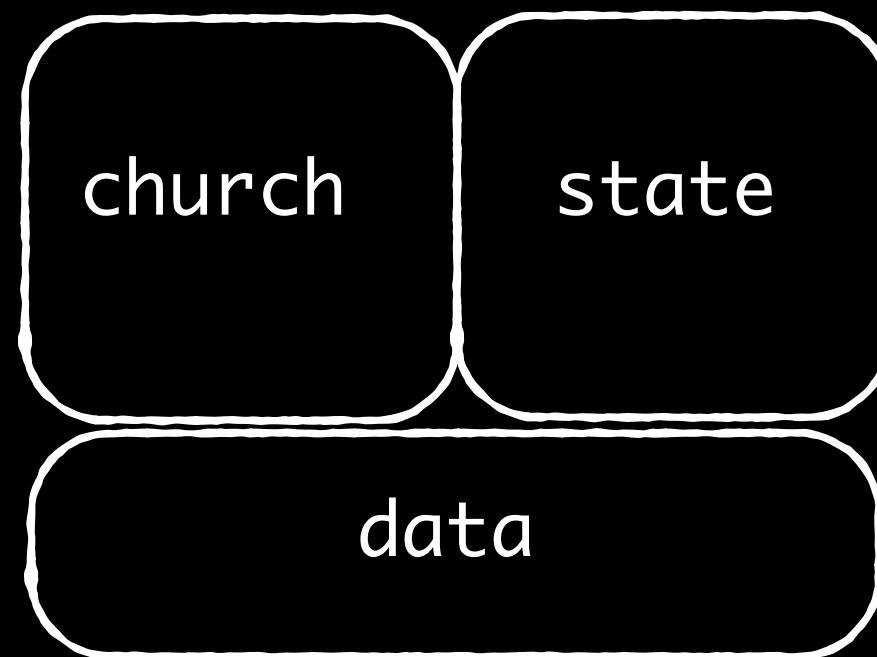
church

news: 20th century

church

state

news: 21st century



# newspapering: 1851 vs. 1996



1851

## The New York Times Introduces a Web Site

By PETER H. LEWIS  
Published: January 22, 1996

The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper's contents.

The New York Times on the Web, as the electronic publication is known, contains most of the news and feature articles from the current day's printed newspaper, classified advertising, reporting that does not appear in the newspaper, and interactive features including the newspaper's crossword puzzle.

1996

**1,615,934** site-wide views over the last hour

**1,257,958** average Sunday New York Times print circulation

**554** stories written over the last 24 hours

**206** countries with visitors in the past 25 minutes

**243,192** words written in the last 24 hours

**65** New York Times newspaper print sites globally

**733** page views from India in the last 10 minutes

2015



The lobby of The New York Times Building/Nic Lehoux

**01/09/2014**

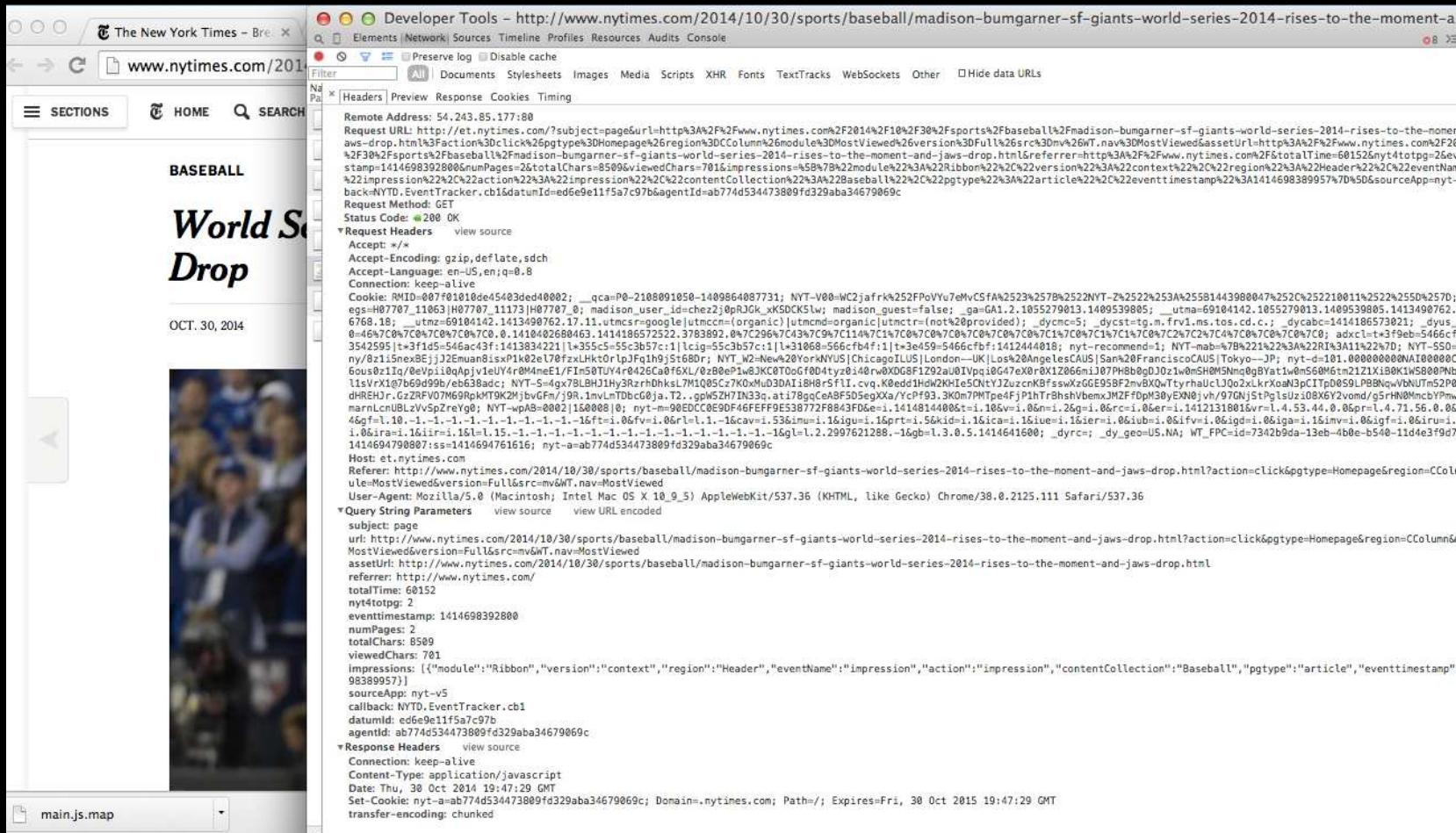
The New York Times Company to Webcast Fourth-Quarter and Full-Year 2013 Earnings Conference Call »

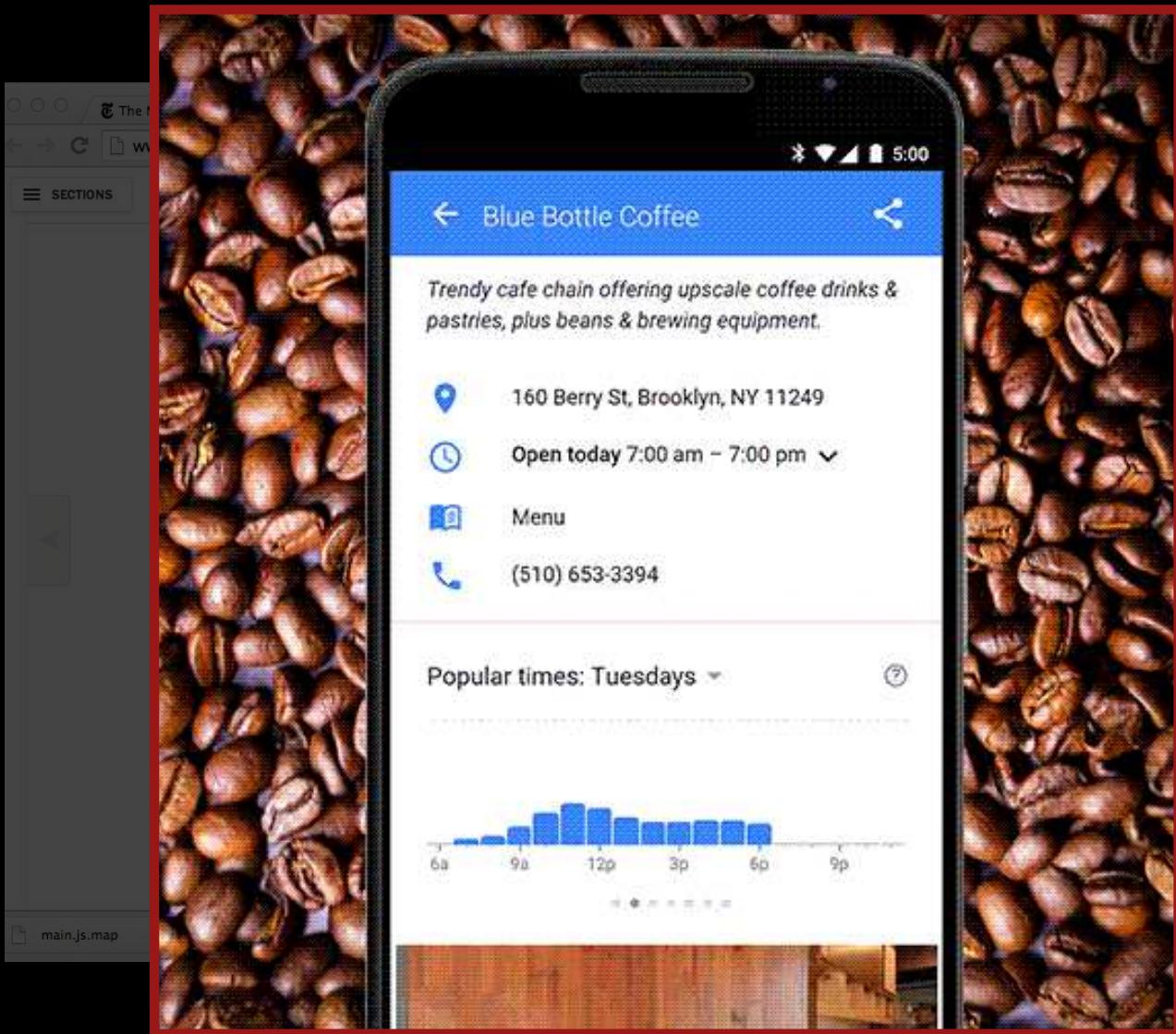
**01/02/2014**

The New York Times to Introduce Redesign of NYTimes.com to All Users Jan. 8 »

**12/12/2013**

The New York Times Company Declares Regular Quarterly Dividend »



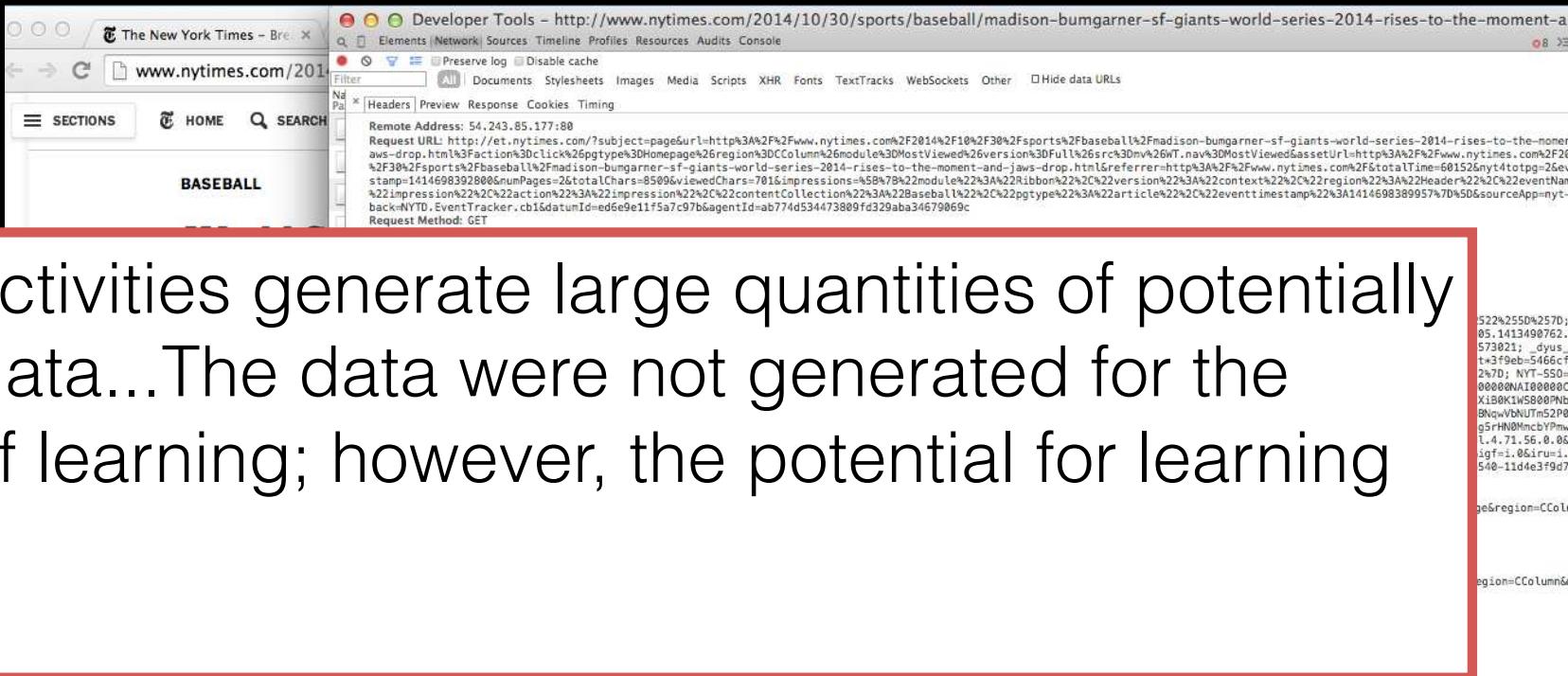


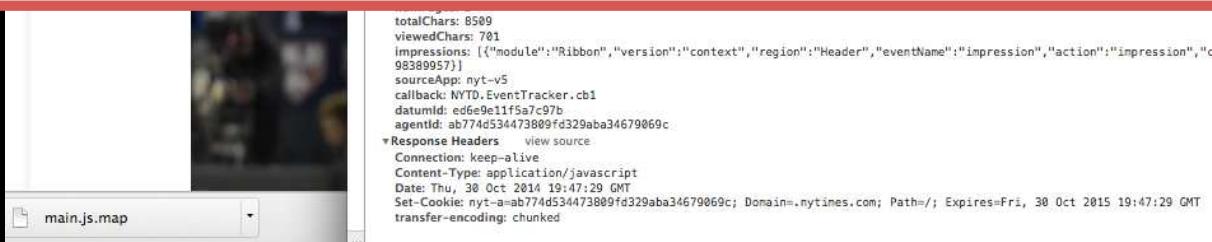
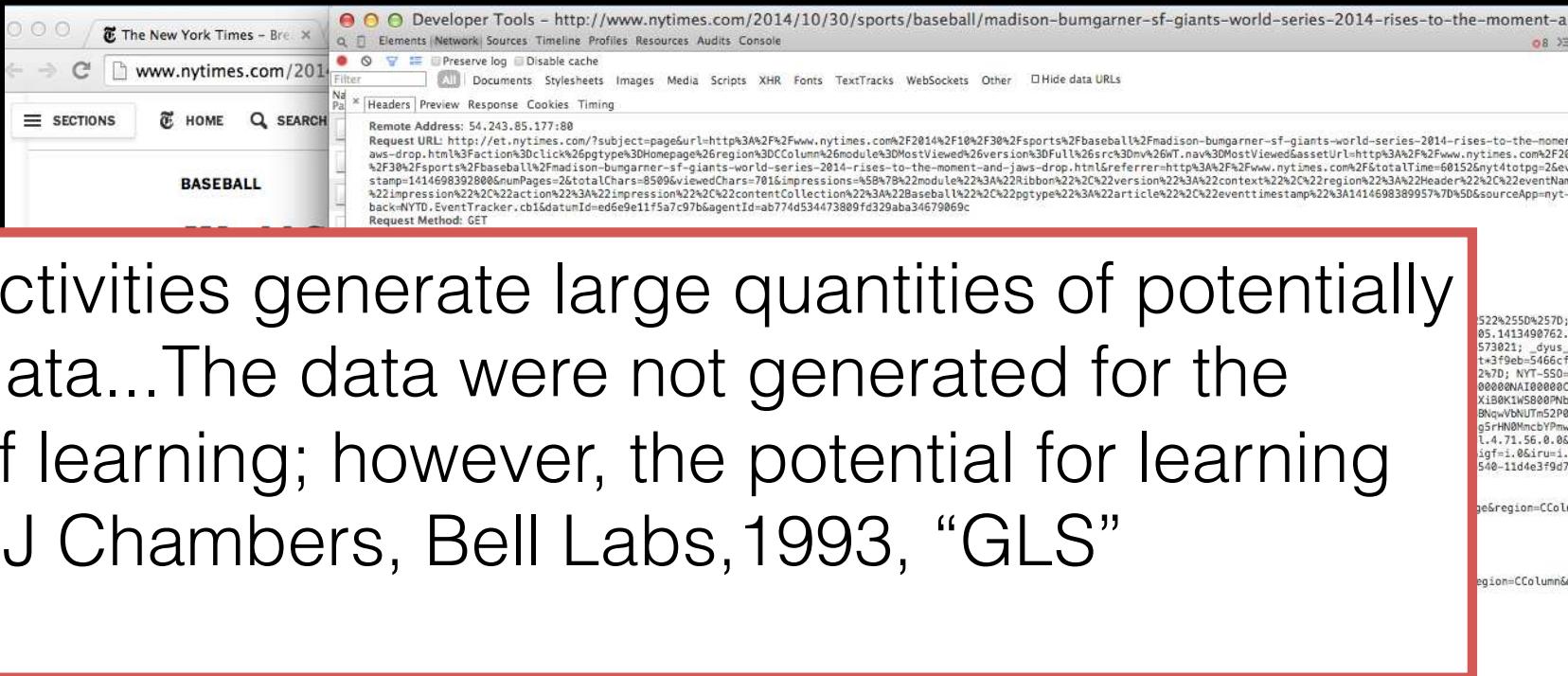
s-world-series-2014-rises-to-the-moment-a  
ungarner-sf-giants-world-series-2014-rises-to-the-mome  
30MostViewed6assetUrl=http%2F%2Fwww.nytimes.com%2F%2Fwww.nytimes.com%2F%2FtotalTfme=601525myt4totpg%2E  
2context%22%20%22region%22%3A%22Header%22%22eventNa  
2eventtimestamp%22%3A1414698389957%7D%5D&sourceApp=nyt

2522%253AV25581443980047%252C%252210011%2522%255D%257D  
9805; \_\_utma=69184142,1055279013,1409539805,1413498762,  
\_tq\_mTrv1.ms\_.tos\_cd.c; \_dyabc=1434186573921; \_dys  
147C0%7C24%7C24%7C4%7C0%7C0%7C0%7C0; advcl=t+3%9eb-5466c  
nend=1; NYT-mab=%7B%221%22%3A%22R13A11%22%7D; NYT-SSD  
20FranciscoCAUS(Tokyo--JP; nyt\_d=101,000000000NIA1000000  
5h0gD10z1wNmSH0M5mno0gRYat1w0m56M0Gtrz2121XzB9K1W58097N  
2mvlXQWtyrhalc1J0j02xLkrXoan3ptCTpD0591PRB8NvVhNUTsP2P  
F7DpM38yEN8jvh/97GNjStpqlslz108X6Y2vond/g5rHNMmcY9m  
ci\_0&er\_i\_1412131801&r=1,4,53,44,0,86r+1,4,71,56,0,86  
iHub\_1,0&fw\_i\_0&id\_i\_0&ig\_i\_1&imvi\_0&igfc\_i\_0&iru  
geo-US.NA; WT\_FPC=id-7342b9da-13eb-4b0e-b540-11d4e3f9d7

aus-drop.html?action=click&pgtype=Homepage&region=CCol  
.36  
drop.html?action=click&pgtype=Homepage&region=CCol&  
jaws-drop.html

ection": "Baseball", "pgtype": "article", "eventtimestamp":





data science: the web

data science: the web  
is your “online presence”

data science: the web

is a microscope

data science: the web  
is an experimental tool

data science: the web

is an optimization tool

# newspapering: 1851 vs. 1996 vs. 2008



1851

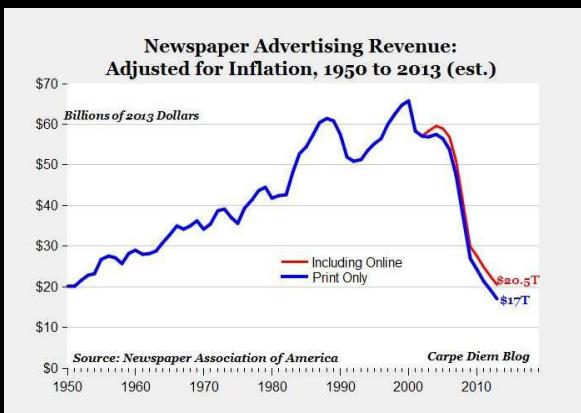
## The New York Times Introduces a Web Site

By PETER H. LEWIS  
Published: January 22, 1996

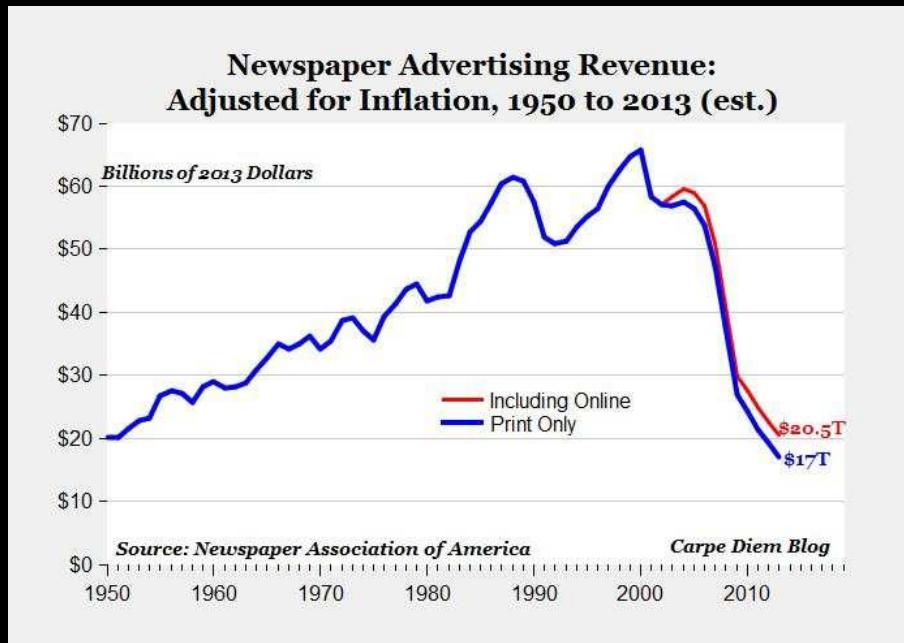
The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper's contents.

The New York Times on the Web, as the electronic publication is known, contains most of the news and feature articles from the current day's printed newspaper, classified advertising, reporting that does not appear in the newspaper, and interactive features including the newspaper's crossword puzzle.

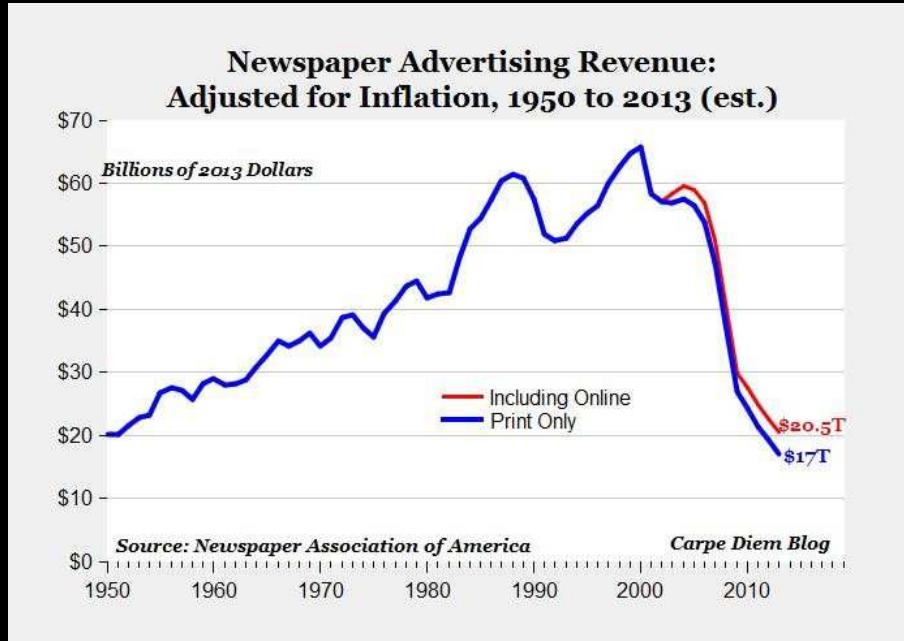
1996



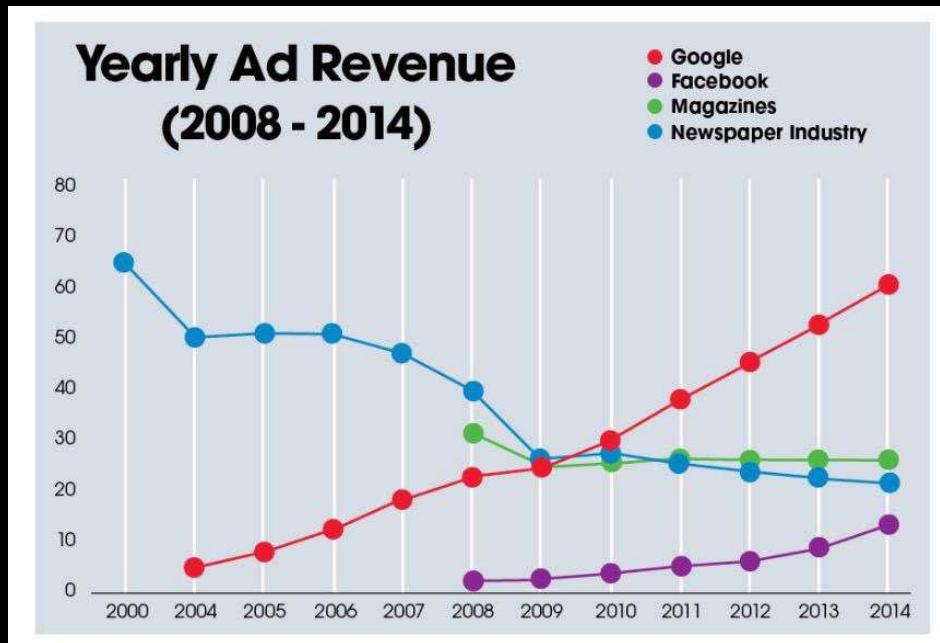
2008



“a startup is a temporary organization in search of a repeatable and scalable business model” –Steve Blank



every publisher is now a startup



every publisher is now a startup

The screenshot shows a web browser window with the following details:

- Title Bar:** Media Websites Battle Fal... (partially visible)
- Address Bar:** www.nytimes.com/2016/04/18/business/media-websites-battle-falteringad-
- Header:** BUSINESS DAY | Media Websites Battle Faltering Ad Revenue and Traffic
- Text Content:** Advertisers adjusted spending accordingly. In the first quarter of 2016, 85 cents of every new dollar spent in online advertising will go to Google or Facebook, said Brian Nowak, a Morgan Stanley analyst.

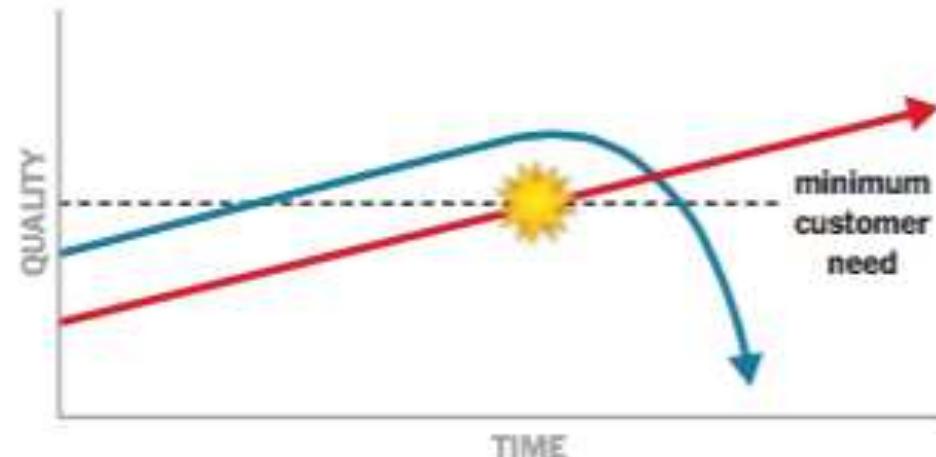
every publisher is now a startup

The New York Times

Innovation

**3.** Over time, **disruptors** improve their product, usually by adapting a new technology. The **flash-point** comes when their products become “good enough” for most customers.

They are now poised to grow by taking market share from **incumbents**.

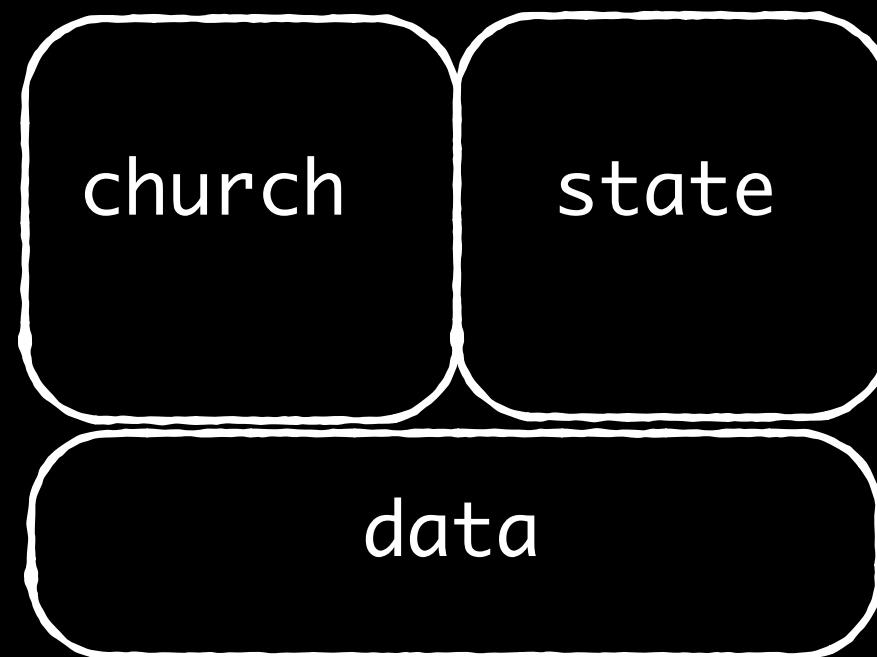


#### HALLMARKS OF DISRUPTIVE INNOVATORS

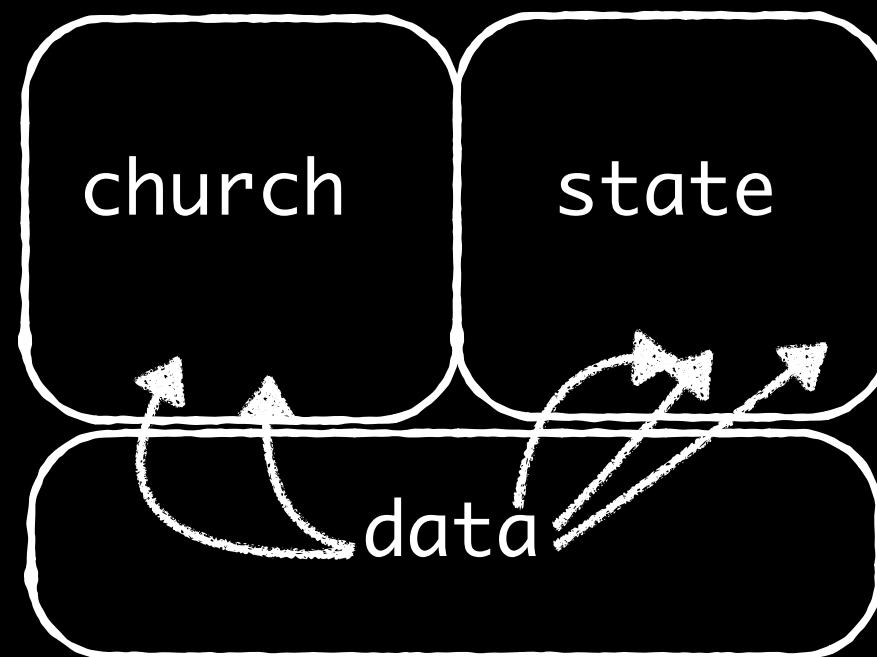
- Introduced by an “outsider”
- Less expensive than existing products
- Targeting underserved or new markets
- Initially inferior to existing products
- Advanced by an enabling technology



news: 21st century



# news: 21st century



learnings

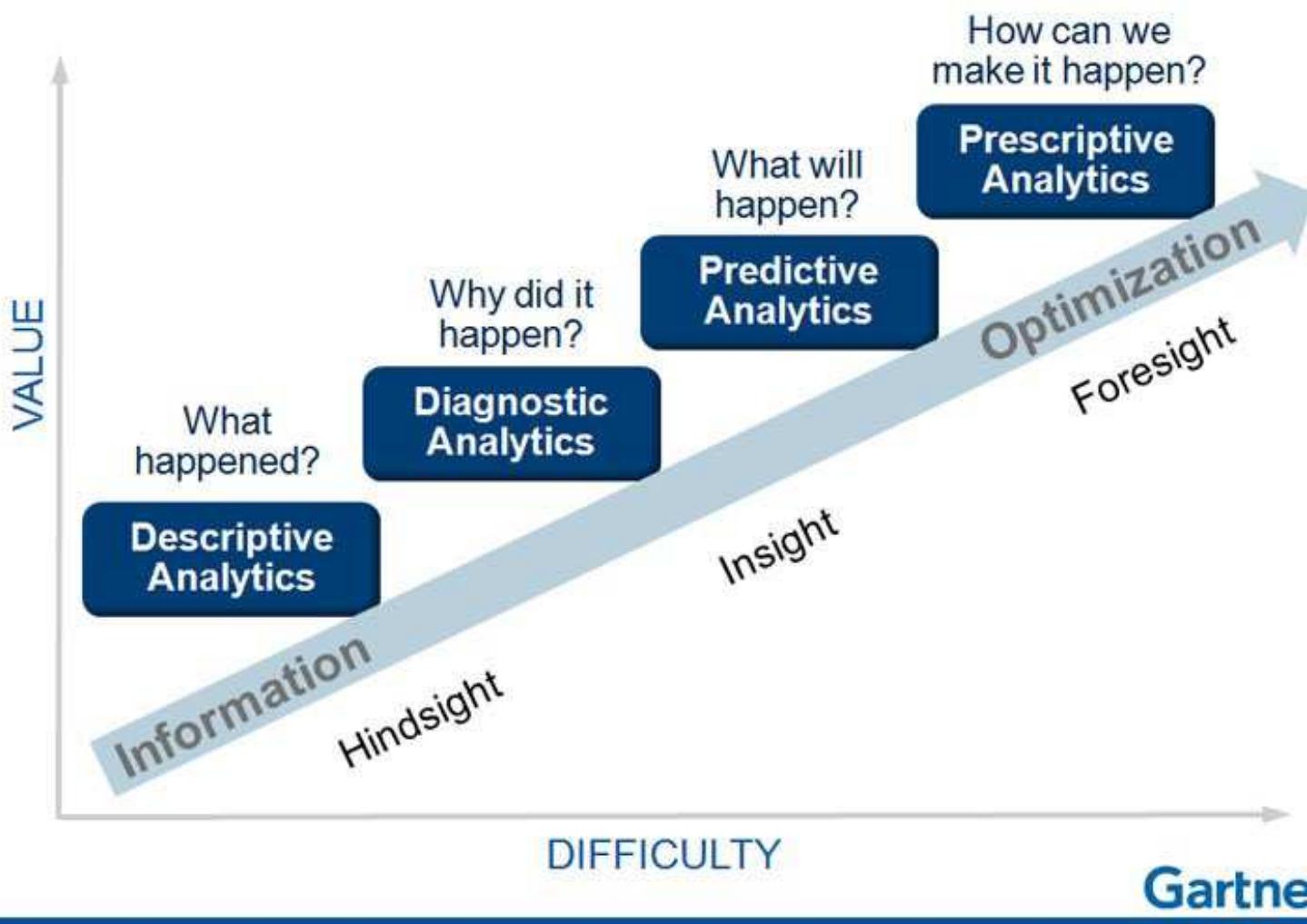
# learnings

- predictive modeling
- descriptive modeling
- prescriptive modeling

(actually ML, shhh...)

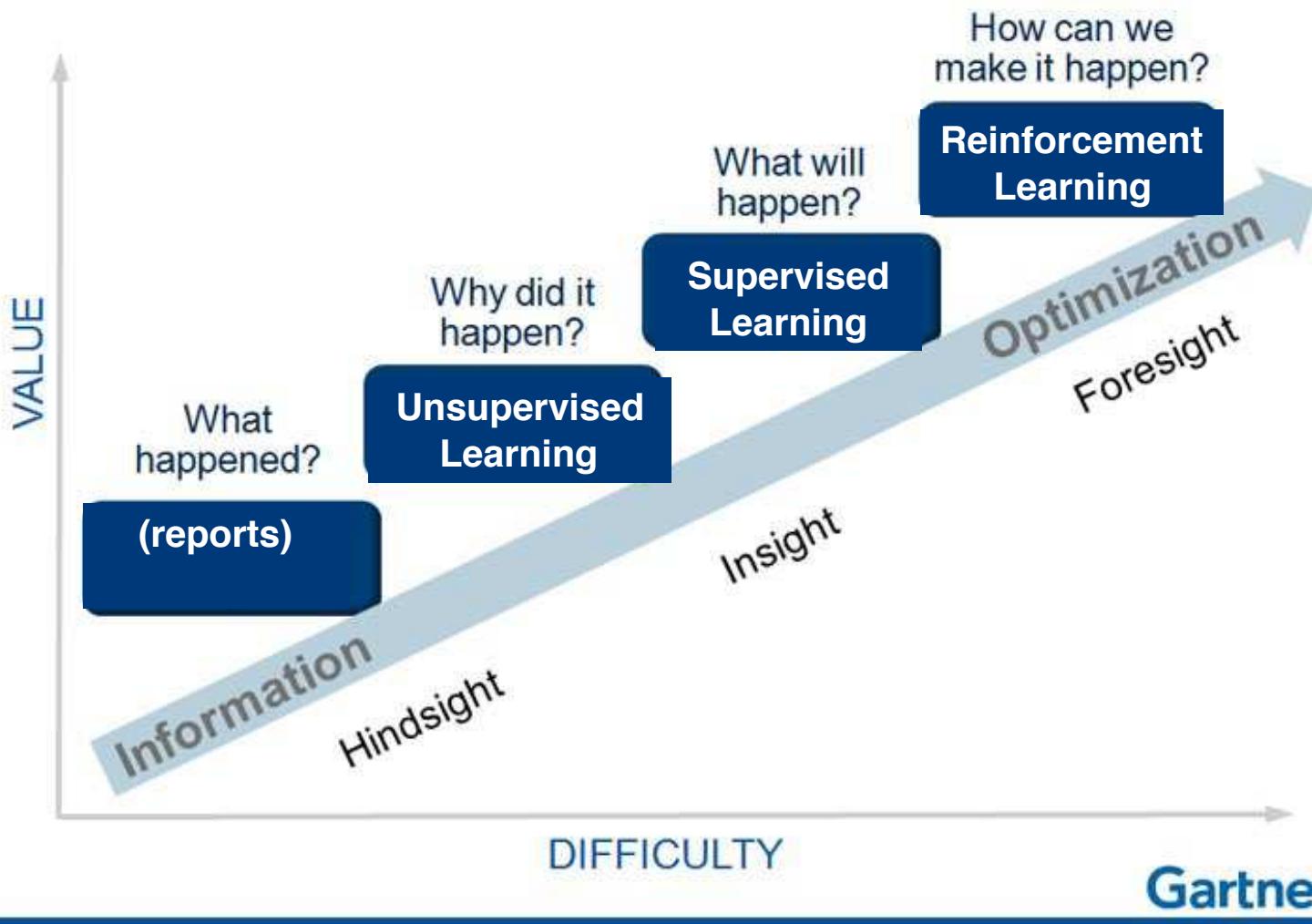
- (supervised learning)
- (unsupervised learning)
- (reinforcement learning)

# Analytic Value Escalator



h/t michael littman

# Analytic Value Escalator

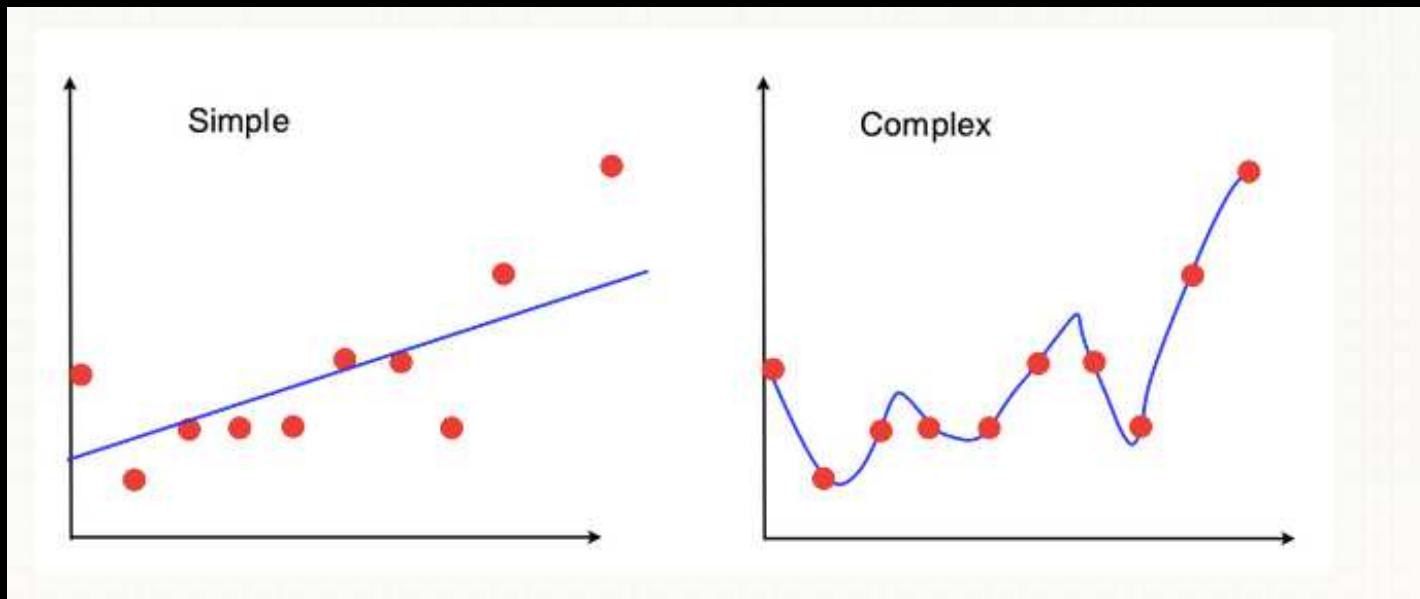


h/t michael littman

# learnings

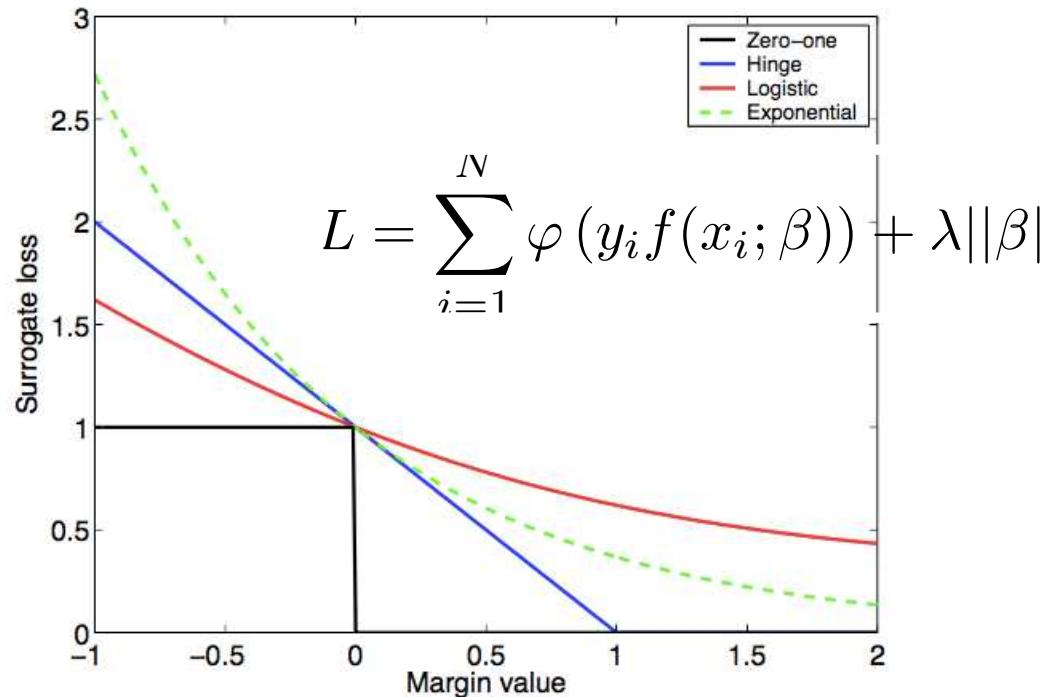
- predictive modeling
- descriptive modeling
- prescriptive modeling

cf. [modelingsocialdata.org](http://modelingsocialdata.org)



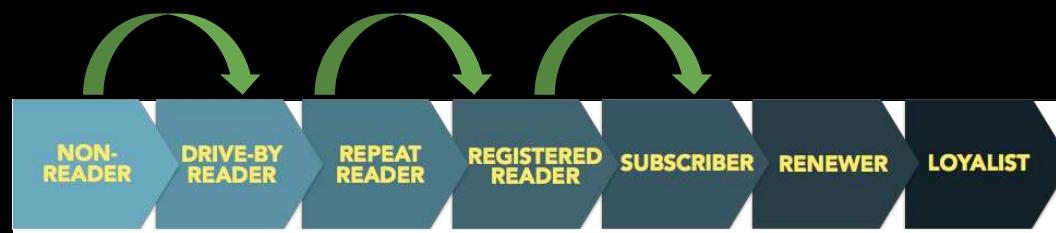
[stats.stackexchange.com](https://stats.stackexchange.com)

## Margin-Based Surrogate Loss Functions



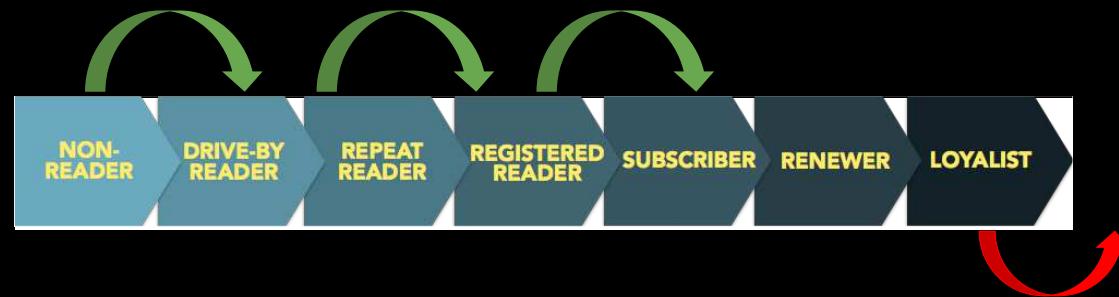
from “are you a bayesian or a frequentist”  
–michael jordan

*predictive modeling, e.g.,*



cf. [modelingsocialdata.org](http://modelingsocialdata.org)

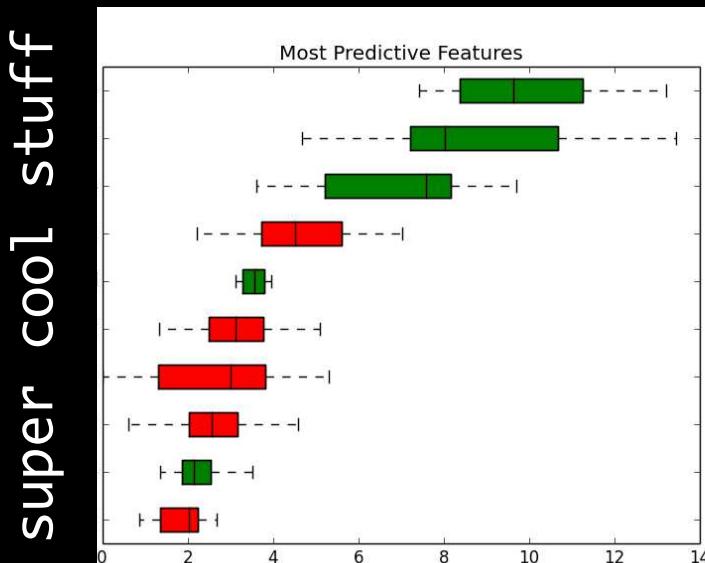
*predictive modeling, e.g.,*



“the funnel”

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

# interpretable predictive modeling

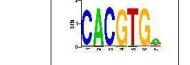
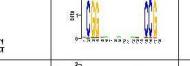
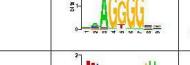
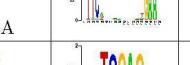
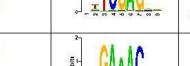
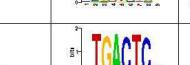
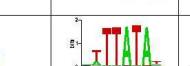
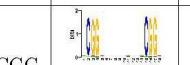
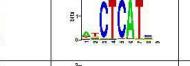
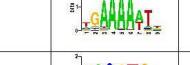
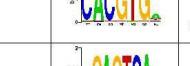
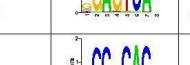
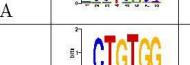
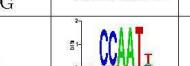
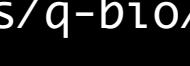


cf. [modelingsocialdata.org](http://modelingsocialdata.org)

interpreting

super cool stuff

cf. me

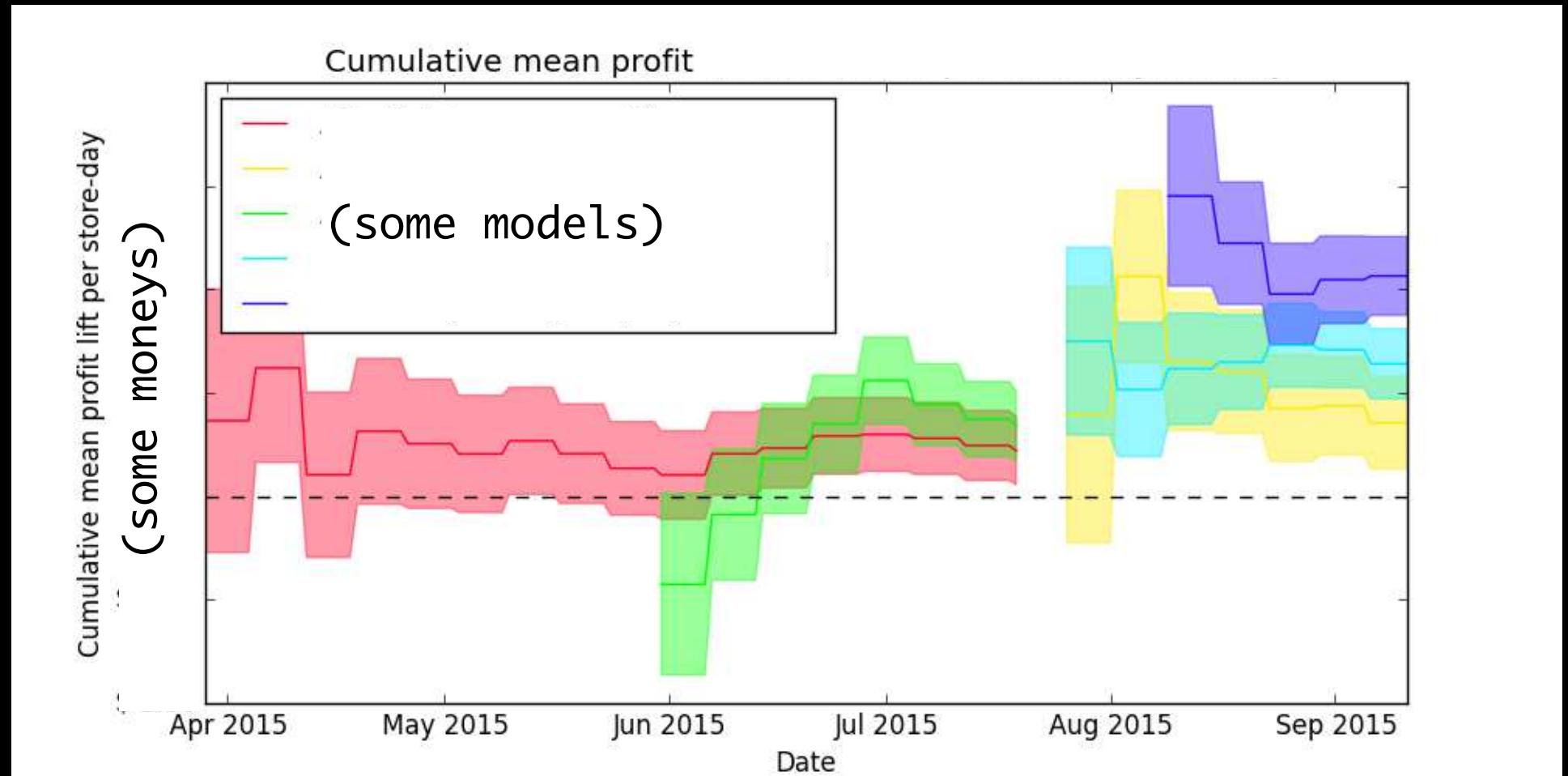
TFNAME	DB-MOTIF	MOTIF	DBNAME	d(p,q)
CBF1	CACGTG		YPD	0.032635
CGG everted repeat	CGGN*CCG		YPD	0.032821
MSN2	AGGGG		TRANSFAC	0.085626
HSF1	TTCNNNGAA		SCPD	0.102410
XBP1	TCGAG		TRANSFAC	0.140561
STE12	TGAAAC		TRANSFAC	0.256750
GCN4	TGACT		SCPD	0.292221
TBP	TTATA		TRANSFAC	0.376601
HAP1	CGGNNTWNCGG		YPD	0.423004
RAP1	RMACCCA		SCPD	0.523059
mPAC	CTCATTC		AlignACE	0.552493
mRRPE	GAAAAATTT		AlignACE	0.630740
PHO4	CACGTG		TRANSFAC	0.672961
YAP1	GAGTCA		TRANSFAC	0.777816
MIG1	CCCCCACAAA		YPD	0.799412
MET31,32	AAACTGTGG		YPD	0.84893
HAP2,3,4	CCAAT		TRANSFAC	1.070837

optimization & learning, e.g.,



“How The New York Times Works” popular mechanics, 2015

optimization & prediction, e.g.,



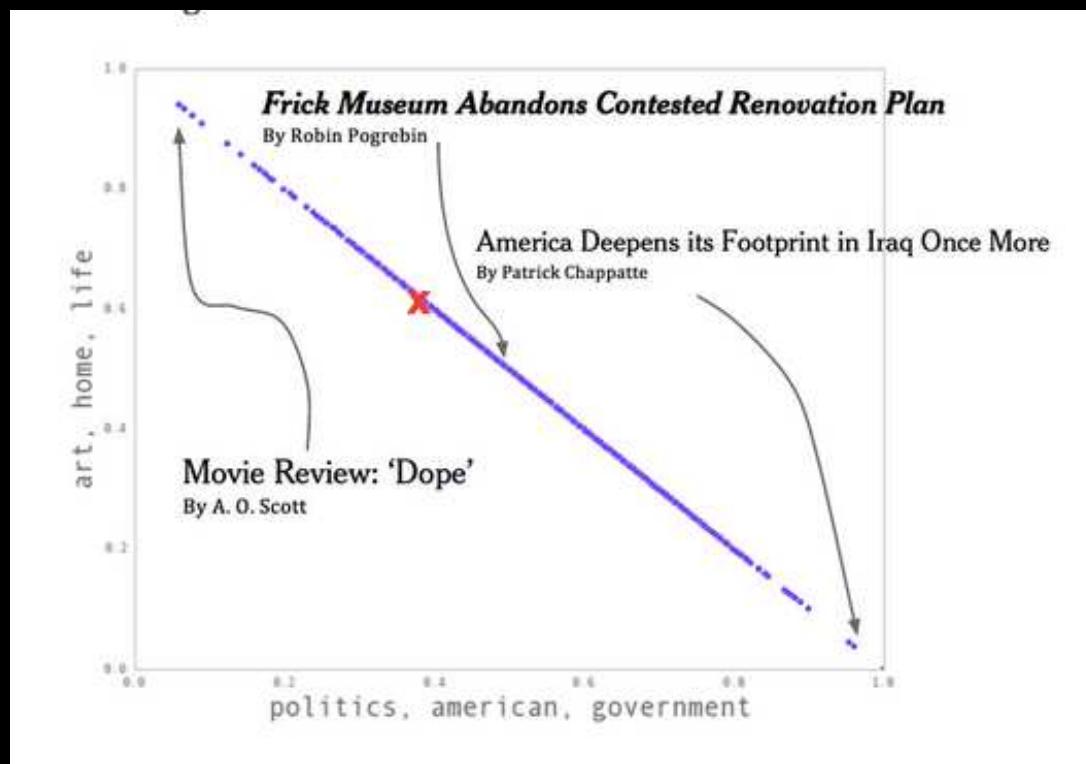
“newsvendor problem,” literally (+prediction+experiment)

# recommendation as inference

MOST EMAILED   MOST VIEWED   RECOMMENDED FOR YOU

1.	THE OUTLAW OCEAN A Renegade Trawler, Hunted for 10,000 Miles by Vigilantes	
2.	Campus Suicide and the Pressure of Perfection	
3.	As Tech Booms, Workers Turn to Coding for Career Change	
4.	Prison Worker Who Aided Escape Tells of Sex, Saw Blades and Deception	
5.	Under Oath, Donald Trump Shows His Raw Side	
6.	American Hunter Killed Cecil, Beloved Lion That Was Lured Out of Its Sanctuary	
7.	A Creature on the Loose Puts Milwaukee Residents on Edge	
8.	N.F.L. Upholds Tom Brady's Ban; Cellphone's Fate Helped Make the Call	
9.	Escalator Death in China Sets Off Furor Online	
10.	DAVID BROOKS The Structure of Gratitude	

## recommendation as inference



[bit.ly/AlexCTM](http://bit.ly/AlexCTM)

descriptive modeling, e.g,

“segments”

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

descriptive modeling, e.g.,

# “segments”

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

descriptive modeling, e.g.,

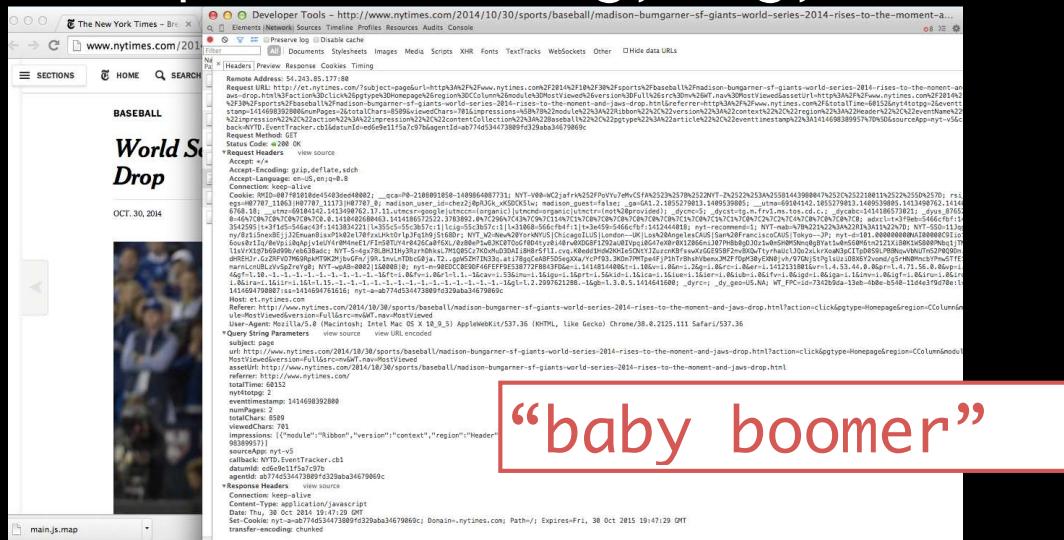
The screenshot shows a browser window with the following details:

- Title Bar:** The New York Times - Bre... (partially visible), Developer Tools - http://www.nytimes.com/2014/10/30/sports/baseball/madison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html
- Address Bar:** www.nytimes.com/2014/10/30/sports/baseball/madison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html
- Header Bar:** Preserve log, Disable cache, Elements, Network, Sources, Timeline, Profiles, Resources, Audits, Console
- Section Bar:** SECTIONS, HOME, SEARCH
- Main Content:** A large red text area displays the search query: "argmax\_z p(z|x)=14".
- Left Sidebar:** BASEBALL, World Series Drop, OCT. 30, 2014.
- Bottom Status Bar:** main.js.map

# “segments”

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

descriptive modeling, e.g.,



# “segments”

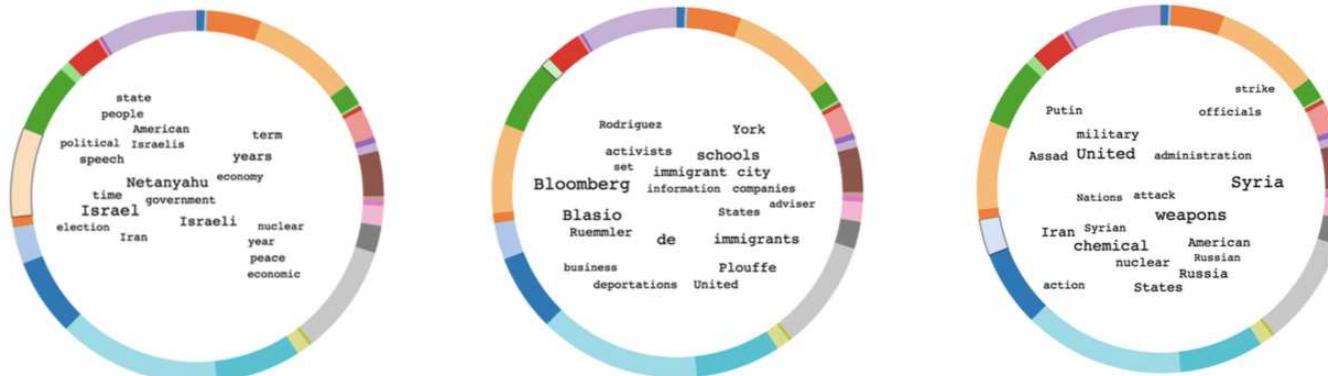
cf. [modelingsocialdata.org](http://modelingsocialdata.org)

- descriptive data product

### A Quick Refinery Demo

The New York Times Developers Article Search API v2 →  Extracting NYT articles from keyword "obama" in 2013. → 

What themes / topics defined the Obama administration during 2013?

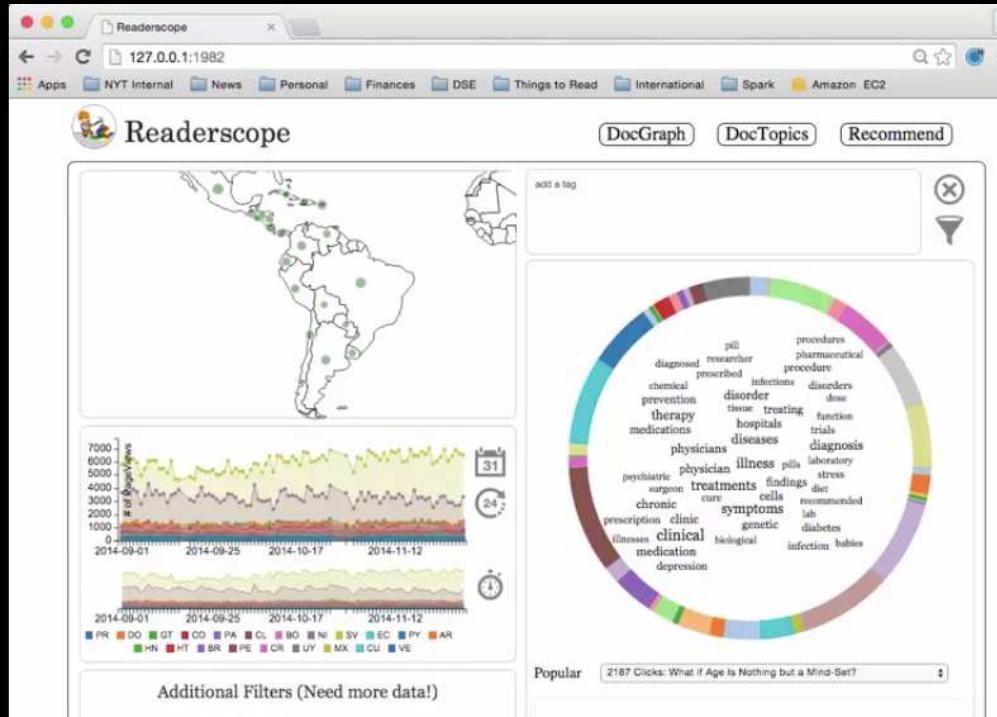


The three sunburst charts represent the most prominent themes in NYT articles about the Obama administration in 2013. The inner ring shows the main categories, and the outer ring shows specific terms.

- Chart 1 (Left):** Focuses on international relations and politics. Key terms include: state, people, American, political, speech, Netanyahu, economy, time, government, election, Israel, Iran, Israeli, nuclear, year, peace, economic, term, years.
- Chart 2 (Middle):** Focuses on domestic politics and media. Key terms include: Rodriguez, Bloomberg, Blasio, Ruemmler, de, business, deportations, Plouffe, United, activists, schools, set, immigrant, city, information, companies, adviser, immigrants.
- Chart 3 (Right):** Focuses on foreign policy and military actions. Key terms include: Putin, Assad, United, Nations, Syria, strike, officials, military, administration, attack, weapons, Iranian, Syrian, chemical, nuclear, American, Russian, Russia, States, action.

cf. [daeilkim.com](http://daeilkim.com)

descriptive modeling, e.g.,



cf. [daeilkim.com](http://daeilkim.com) ; import bnpy

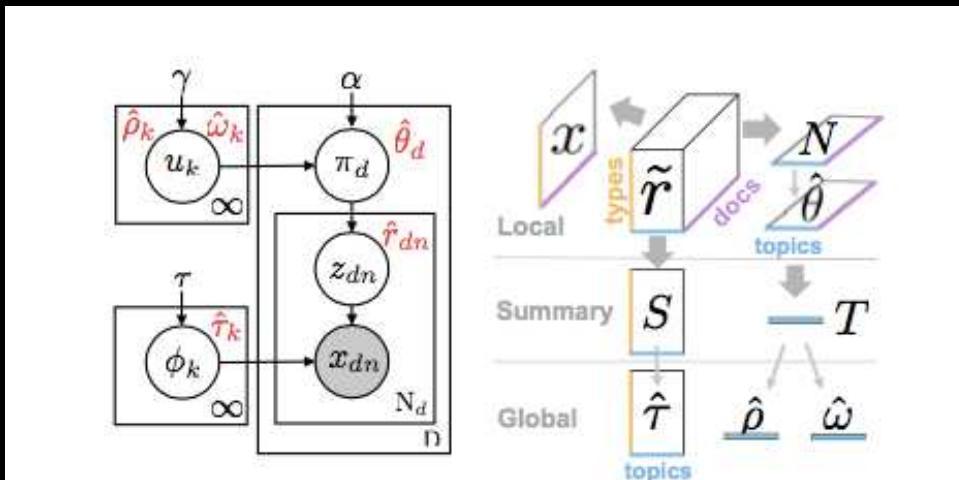


Figure 1: *Left:* Directed graphical model for the HDP admixture (Sec. 2). Free parameters for mean-field variational inference (Sec. 3) shown in red. *Right:* Flow chart for our inference algorithm, specialized for bag-of-words data, where we can use sparse type-based assignments  $\tilde{r}$  instead of per-token variables  $\hat{r}$ . We define  $\tilde{r}_{dwk}$  to be the total mass of all tokens in document  $d$  of type  $w$  assigned to  $k$ :  $\tilde{r}_{dwk} = \sum_{n=1}^{N_d} \hat{r}_{dnk} \delta_{x_{dn}, w}$ . Updates flow from  $\tilde{r}$  to global topic-type parameters  $\hat{\tau}$  and (separately) to global topic weight parameters  $\hat{\rho}, \hat{\omega}$ . Each variable's shape gives its dimensionality. Thick arrows indicate summary statistics; thin arrows show free parameter updates.

modeling your audience  
[bit.ly/Hughes-Kim-Sudderth-AISTATS15](http://bit.ly/Hughes-Kim-Sudderth-AISTATS15)

**Objective function.** Mean field methods optimize an evidence lower bound  $\log p(x|\gamma, \alpha, \tau) \geq \mathcal{L}(\cdot)$ , where

$$\mathcal{L}(\cdot) \triangleq \mathcal{L}_{data}(\cdot) + H_z(\cdot) + \mathcal{L}_{HDP}(\cdot) + \mathcal{L}_u(\cdot). \quad (4)$$

The final term  $\mathcal{L}_u(\cdot)$ , which depends only on  $q(u)$ , is discussed in the next section. The first three terms account for data generation, the assignment entropy, and the document-topic allocations. These are defined below, with expectations taken with respect to Eq. (3):

$$\mathcal{L}_{data}(\cdot) \triangleq \mathbb{E}_q[\log p(x|z, \phi) + \log \frac{p(\phi|\bar{\tau})}{q(\phi|\hat{\tau})}], \quad (5)$$

$$H_z(\cdot) \triangleq -\sum_{k=1}^K \sum_{d=1}^D \sum_{n=1}^{N_d} \hat{r}_{dnk} \log \hat{r}_{dnk},$$

$$\mathcal{L}_{HDP}(\cdot) \triangleq \mathbb{E}_q \left[ \log \frac{p(z|\pi)p(\pi|\alpha, u)}{q(\pi|\hat{\theta})} \right].$$

The forms of  $\mathcal{L}_{data}$  and  $H_z$  are unchanged from the simpler case of mean-field for DP mixtures. Closed-form expressions are in the Supplement.

modeling your audience  
(optimization, ultimately)

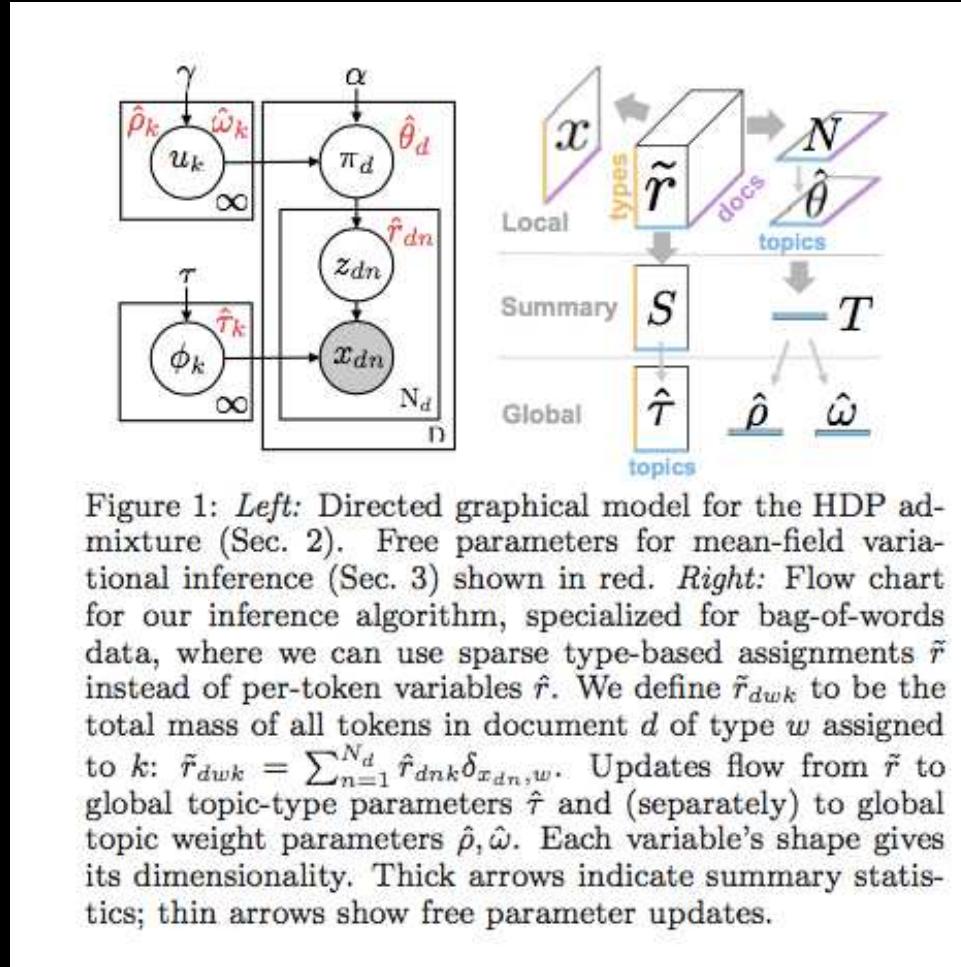


Figure 1: *Left:* Directed graphical model for the HDP admixture (Sec. 2). Free parameters for mean-field variational inference (Sec. 3) shown in red. *Right:* Flow chart for our inference algorithm, specialized for bag-of-words data, where we can use sparse type-based assignments  $\tilde{r}$  instead of per-token variables  $\hat{r}$ . We define  $\tilde{r}_{dwk}$  to be the total mass of all tokens in document  $d$  of type  $w$  assigned to  $k$ :  $\tilde{r}_{dwk} = \sum_{n=1}^{N_d} \hat{r}_{dnk} \delta_{x_{dn}, w}$ . Updates flow from  $\tilde{r}$  to global topic-type parameters  $\hat{\tau}$  and (separately) to global topic weight parameters  $\hat{\rho}, \hat{\omega}$ . Each variable's shape gives its dimensionality. Thick arrows indicate summary statistics; thin arrows show free parameter updates.

modeling your audience  
also allows insight+targeting as inference

prescriptive modeling

## prescriptive modeling

- 
- |               |   |
|---------------|---|
| descriptive:  | specify $x$ ; learn $z(x)$ or $p(z x)$ where $z$ is “simpler” than $x$      |
| predictive:   | specify $x$ and $y$ ; learn to predict $y$ from $x$                         |
| prescriptive: | specify $x, y$ , and $a$ ; learn to prescribe $a$ given $x$ to maximize $y$ |
-

prescriptive modeling

$$V = E_+(y) = \sum_{yax} y P_+(y, a, x)$$

“off policy value estimation”  
(cf. “causal effect estimation”)

$$\hat{V} = \frac{1}{N} \sum_{i=1}^{i=N} y_i \frac{1(a_i = h(x_i))}{\hat{B}(a_i|x_i)}$$

cf. Langford `08-`16;  
Horvitz & Thompson `52;  
Holland `86

“off policy value estimation”  
(cf. “causal effect estimation”)

$$\hat{V} = \frac{1}{N} \sum_{i=1}^{i=N} y_i \frac{1(a_i = h(x_i))}{\hat{B}(a_i|x_i)}$$

Vapnik's razor

“ When solving a (learning) problem of interest,  
do not solve a more complex problem as an  
intermediate step.”

# prescriptive modeling

## Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer<sup>a,1</sup>, Jamie E. Guillory<sup>b,2</sup>, and Jeffrey T. Hancock<sup>b,c</sup>

Author Affiliations 

<sup>a</sup>Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and

Departments of <sup>b</sup>Communication and

<sup>c</sup>Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

A correction has been published

A correction has been published

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

# prescriptive modeling

## Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer<sup>a,1</sup>, Jamie E. Guillory<sup>b,2</sup>, and Jeffrey T. Hancock<sup>b,c</sup>

Author Affiliations 

<sup>a</sup>Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and

Departments of <sup>b</sup>Communication and

<sup>c</sup>Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

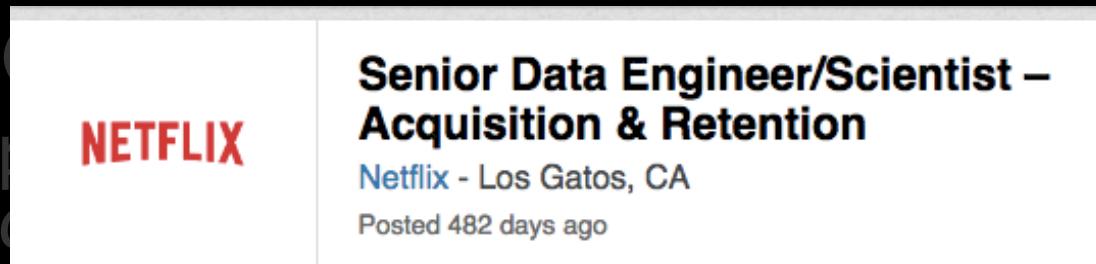
A correction has been published

A correction has been published

aka “A/B testing”;  
RCT

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

prescriptive modeling: from A/B to....



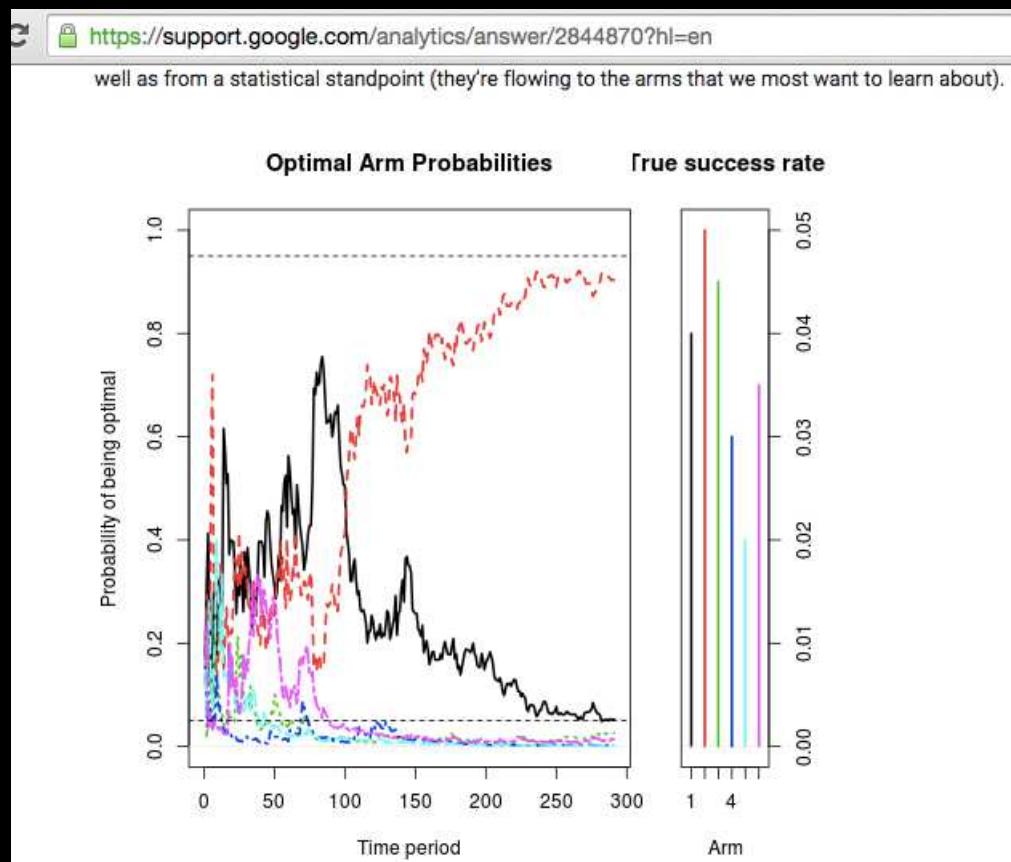
testing”; Test

Some of the most recognizable personalization in our service is the collection of “genre” rows. . . Members connect with these rows so well that we measure an **increase in member retention** by **placing the most tailored rows higher** on the page instead of lower.

business as usual Reporting

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

## real-time A/B -> “bandits”



GOOG blog:

cf. [modelingsocialdata.org](http://modelingsocialdata.org)

*prescriptive modeling, e.g.,*

prescriptive modeling, e.g.,



**Colin Russel** 10:21 AM

!blossom facebook? all



**blossombot** BOT 10:21 AM ★

Blossom has the following suggestions for your next Facebook posts:

**nytimes:** <http://www.nytimes.com/2015/08/12/opinion/frank-bruni-can-we-interest-you-in-teaching.html>

**nytopinion:** <http://www.nytimes.com/2015/08/12/opinion/when-innocence-is-no-defense.html>

**nytpolitics:** <http://www.nytimes.com/2015/08/16/magazine/president-obamas-letter-to-the-editor.html>

**upshot:** <http://www.nytimes.com/interactive/2015/07/03/upshot/a-quick-puzzle-to-test-your-problem-solving.html>

**tmagazin:** Blossom currently has no suggestions

**nytimestravel:** <http://www.nytimes.com/2015/08/13/us/politics/us-jets-meet-limit-as-iraqi-ground-fight-against-isis-plods-on.html>

**nytimesscience:** <http://www.nytimes.com/interactive/2015/07/03/upshot/a-quick-puzzle-to-test-your-problem-solving.html>

**nytfood:** Blossom currently has no suggestions

**WellNYT:** <http://www.nytimes.com/2015/08/12/sports/kaci-lickteig-climbs-toward-the-top-in-ultrarunning.html>

**NewYorkTodayNYT:** <http://www.nytimes.com/2015/08/08/nyregion/she-answered-his-ad-for-a-roommate-moved-in-and-never-left.html>



prescriptive modeling, e.g.,

Colin Russel 10:21 AM  
!blossom facebook? all

blossombot BOT 10:21 AM ★  
Blossom has the following suggestions for your next Facebook posts:  
nytimes: <http://www.nytimes.com/2015/08/12/opinion/frank-bruni-can-we-interest-you-in-teaching.html>

blossombot BOT 3:40 PM ★  
Blossom Alert!  
This piece is predicted to go viral if posted on Facebook next:  
<http://www.nytimes.com/2015/08/12/sports/football/ikemefuna-enemkpali-the-backup-who-broke-geno-smiths-jaw.html>  
Here is its Stela Link

nytimesscience: <http://www.nytimes.com/interactive/2015/07/03/upshot/a-quick-puzzle-to-test-your-problem-solving.html>

nytfood: Blossom currently has no suggestions

WellNYT: <http://www.nytimes.com/2015/08/12/sports/kaci-lickteig-climbs-toward-the-top-in-ultrarunning.html>

NewYorkTodayNYT: <http://www.nytimes.com/2015/08/08/nyregion/she-answered-his-ad-for-a-roommate-moved-in-and-never-left.html>

+ ☺

prescriptive modeling, e.g.,

Colin Russel 10:21 AM  
!blossom facebook? all

blossombot BOT 10:21 AM ★  
Blossom has the following suggestions for your next Facebook posts:  
nytimes: <http://www.nytimes.com/2015/08/12/opinion/frank-bruni-can-we-interest-you-in-teaching.html>

blossombot BOT 3:40 PM ★  
Blossom Alert!  
This piece is predicted to go viral if posted on Facebook next:  
<http://www.nytimes.com/2015/08/12/sports/football/ikemefuna-enemkpali-the-backup-who-broke-geno-smiths-jaw.html>  
Here is its Stela Link

nytimesscience: <http://www.nytimes.com/interactive/2015/07/03/upshot/a-quick-puzzle-to-test-your-problem-solving.html>

nytfood: Blossom currently has no suggestions

WellNYT: <http://www.nytimes.com/2015/08/12/sports/kaci-lickteig-climbs-toward-the-top-in-ultrarunning.html>

NewYorkTodayNYT: <http://www.nytimes.com/2015/08/08/nyregion/she-answered-his-ad-for-a-roommate-moved-in-and-never-left.html>

leverage methods which are predictive yet performant

## NB: data-informed, not data-driven

Colin Russel 10:21 AM  
!blossom facebook? all

blossombot BOT 10:21 AM ★  
Blossom has the following suggestions for your next Facebook posts:  
[nytimes: http://www.nytimes.com/2015/08/12/opinion/frank-bruni-can-we-interest-you-in-teaching.html](http://www.nytimes.com/2015/08/12/opinion/frank-bruni-can-we-interest-you-in-teaching.html)

blossombot BOT 3:40 PM ★  
**Blossom Alert!**  
This piece is predicted to go viral if posted on Facebook next:  
<http://www.nytimes.com/2015/08/12/sports/football/ikemefuna-enemkpali-the-backup-who-broke-geno-smiths-jaw.html>  
Here is its Stela Link

nytimesscience: <http://www.nytimes.com/interactive/2015/07/03/upshot/a-quick-puzzle-to-test-your-problem-solving.html>

nytfood: Blossom currently has no suggestions

WellNYT: <http://www.nytimes.com/2015/08/12/sports/kaci-lickteig-climbs-toward-the-top-in-ultrarunning.html>

NewYorkTodayNYT: <http://www.nytimes.com/2015/08/08/nyregion/she-answered-his-ad-for-a-roommate-moved-in-and-never-left.html>

+ ☺

predicting views/cascades: doable?

## **Optimizing Web Traffic via the Media Scheduling Problem**

Lars Backstrom<sup>\*</sup>  
Dept. of Computer Science  
Cornell University  
Ithaca, NY 14853.  
[lars@cs.cornell.edu](mailto:lars@cs.cornell.edu)

Jon Kleinberg<sup>†</sup>  
Dept. of Computer Science  
Cornell University  
Ithaca, NY 14853.  
[kleinber@cs.cornell.edu](mailto:kleinber@cs.cornell.edu)

Ravi Kumar  
Yahoo! Research  
701 First Ave.  
Sunnyvale, CA 94089.  
[ravikumar@yahoo-inc.com](mailto:ravikumar@yahoo-inc.com)

KDD 09: how many people are online?

# predicting views/cascades: doable?

## Can cascades be predicted?

Justin Cheng  
Stanford University  
[jcccf@cs.stanford.edu](mailto:jcccf@cs.stanford.edu)

Lada A. Adamic  
Facebook  
[ladamic@fb.com](mailto:ladamic@fb.com)

P. Alex Dow  
Facebook  
[adow@fb.com](mailto:adow@fb.com)

Jon Kleinberg  
Cornell University  
[kleinber@cs.cornell.edu](mailto:kleinber@cs.cornell.edu)

Jure Leskovec  
Stanford University  
[jure@cs.stanford.edu](mailto:jure@cs.stanford.edu)

WWW 14: FB shares

# predicting views/cascades: features?

Content Features	
$score_{food/nature/\dots}$	The probability of the photo having a specific feature (food, overlaid text, landmark, nature, etc.)
$is\_en$	Whether the photo was posted by an English-speaking user or page
$has\_caption$	Whether the photo was posted with a caption
$lwc_{pos/neg/use}$	Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English
Root (Original Poster) Features	
$views_{0,k}$	Number of users who saw the original photo until the $k$ th reshare was posted
$orig\_is\_page$	Whether the original poster is a page
$outdeg(v_0)$	Friend, subscriber or fan count of the original poster
$age_{v_0}$	Age of the original poster, if a user
$gender_{v_0}$	Gender of the original poster, if a user
$fb\_age_{v_0}$	Time since the original poster registered on Facebook, if a user
$activity_{v_0}$	Average number of days the original poster was active in the past month, if a user
Resharer Features	
$views_{1,k-1,k}$	Number of users who saw the first $k - 1$ reshares until the $k$ th reshare was posted
$pages_k$	Number of pages responsible for the first $k$ reshares, including the root, or $\sum_{i=0}^{k-1} \mathbb{1}\{v_i \text{ is a page}\}$
$friends_{avg/90p}$	Average or 90th percentile friend count of the first $k$ reshancers, or $\frac{1}{k} \sum_{i=1}^k outdeg(v_i) \mathbb{1}\{v_i \text{ is a user}\}$
$fans_{avg/90p}$	Average or 90th percentile fan count of the first $k$ reshancers, or $\frac{1}{k} \sum_{i=1}^k outdeg(v_i) \mathbb{1}\{v_i \text{ is a page}\}$
$subscribers_{avg/90p}$	Average or 90th percentile subscriber count of the first $k$ reshancers, or $\frac{1}{k} \sum_{i=1}^k outdeg(v_i) \mathbb{1}\{v_i \text{ is a user}\}$
$fb\_ages_{avg/90p}$	Average or 90th percentile time since the first $k$ reshancers registered on Facebook, or $\frac{1}{k} \sum_{i=1}^k fb\_age_i$
$activities_{avg/90p}$	Average number of days the first $k$ reshancers were active in July, or $\frac{1}{k} \sum_{i=1}^k activity_i$
$ages_{avg/90p}$	Average age of the first $k$ reshancers, or $\frac{1}{k} \sum_{i=1}^k age_i$
$female_k$	Number of female users among the first $k$ reshancers, or $\sum_{i=1}^k \mathbb{1}\{gender_i \text{ is female}\}$
Structural Features	
$outdeg(v_i)$	Connection count (sum of friend, subscriber and fan counts) of the $i$ th resharer (or out-degree of $v_i$ on $G = (V, E)$ )
$outdeg(v'_i)$	Out-degree of the $i$ th reshare on the induced subgraph $G' = (V', E')$ of the first $k$ reshancers and the root
$outdeg(\hat{v}_i)$	Out-degree of the $i$ th reshare on the reshare graph $\hat{G} = (\hat{V}, \hat{E})$ of the first $k$ reshares
$orig\_connections_k$	Number of first $k$ reshancers who are friends with, or fans of the root, or $ \{(v_i \mid (v_0, v_i) \in E, 1 \leq i \leq k)\} $
$border\_nodes_k$	Total number of users or pages reachable from the first $k$ reshancers and the root, or $ \{(v_i \mid (v_0, v_i) \in E, 0 \leq i, j \leq k)\} $
$border\_edges_k$	Total number of first-degree connections of the first $k$ reshancers and the root, or $ \{(v_i, v_j) \mid (v_0, v_i) \in E, 0 \leq i, j \leq k\} $
$subgraph'_k$	Number of edges on the induced subgraph of the first $k$ reshancers and the root, or $ \{(v_i, v_j) \mid (v_0, v_i) \in E', 0 \leq i, j \leq k\} $
$depth'_k$	Change in tree depth of the first $k$ reshares, or $\min_\beta \sum_{i=1}^k (depth_i - \beta i)^2$
$depths_{avg/90p}$	Average or 90th percentile tree depth of the first $k$ reshares, or $\frac{1}{k} \sum_{i=1}^k depth_i$
$did\_leave$	Whether any of the first $k$ reshares are not first-degree connections of the root
Temporal Features	
$time_i$	Time elapsed between the original post and the $i$ th reshare
$time'_{1..k/2}$	Average time between reshares, for the first $k/2$ reshares, or $\frac{1}{k/2-1} \sum_{i=1}^{k/2-1} (time_{i+1} - time_i)$
$time''_{k/2..k}$	Average time between reshares, for the last $k/2$ reshares, or $\frac{1}{k/2-1} \sum_{i=k/2}^{k-1} (time_{i+1} - time_i)$
$time''_{1..k}$	Change in the time between reshares of the first $k$ reshares, or $\min_\beta \sum_{i=1}^{k-1} (time_{i+1} - time_i) - \beta i^2$
$views'_{0,k}$	Number of users who saw the original photo, until the $k$ th reshare was posted, per unit time, or $\frac{views_{0,k}}{time_A}$
$views'_{1..k-1,k}$	Number of users who saw the first $k - 1$ reshares, until the $k$ th reshare was posted, per unit time, or $\frac{views_{1..k-1,k}}{time_B}$

## predicting views/cascades: features?

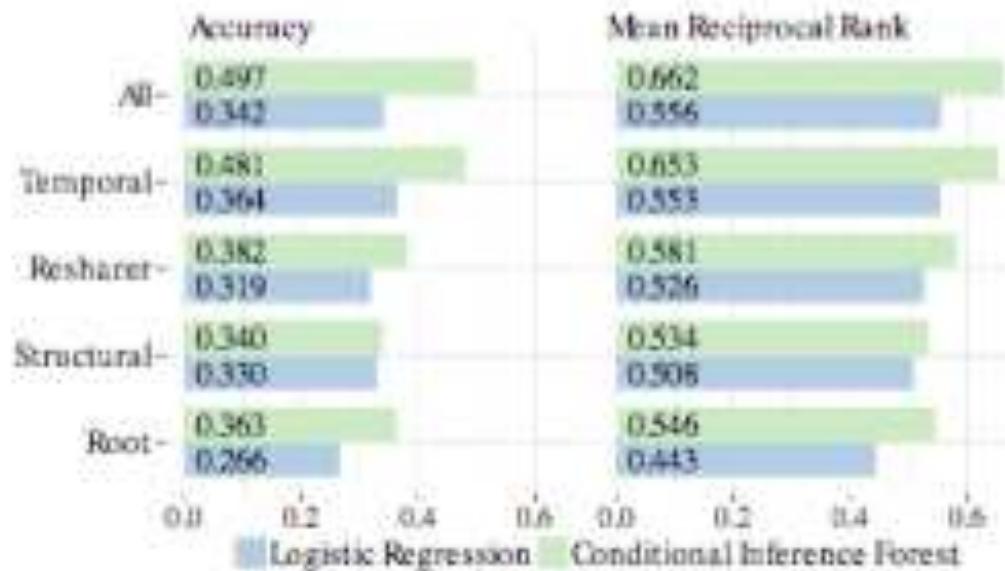


Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

# **predicting views/cascades: doable?**

## **Exploring limits to prediction in complex social systems**

Travis Martin

University of Michigan  
Dept. of Computer Science  
Ann Arbor, MI  
[travisbm@umich.edu](mailto:travisbm@umich.edu)

Jake M. Hofman

Microsoft Research  
641 6th Ave, Floor 7  
New York, NY  
[jmh@microsoft.com](mailto:jmh@microsoft.com)

Amit Sharma

Microsoft Research  
[amshar@microsoft.com](mailto:amshar@microsoft.com)

Ashton Anderson

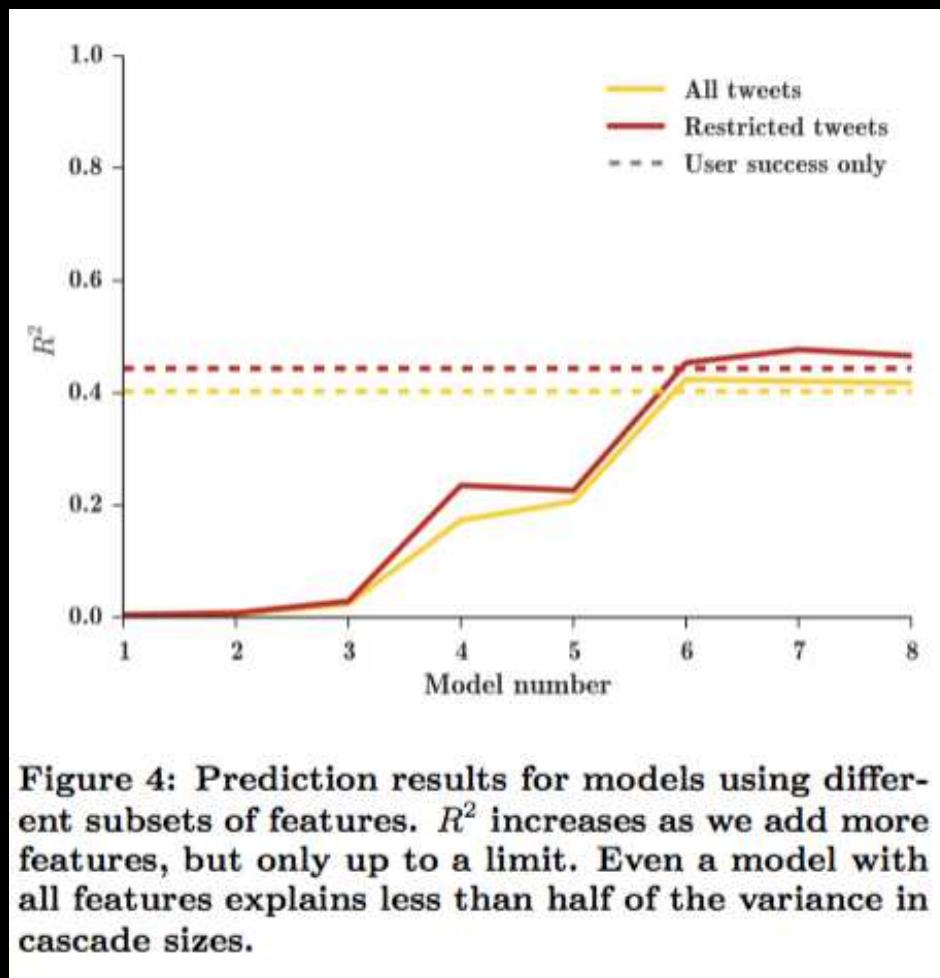
Microsoft Research  
[ashton@microsoft.com](mailto:ashton@microsoft.com)

Duncan J. Watts

Microsoft Research  
[duncan@microsoft.com](mailto:duncan@microsoft.com)

WWW 16: TWIT RT's

## predicting views/cascades: doable?



descriptive:

predictive:

prescriptive:

Explore

Learning

Test

Optimizing

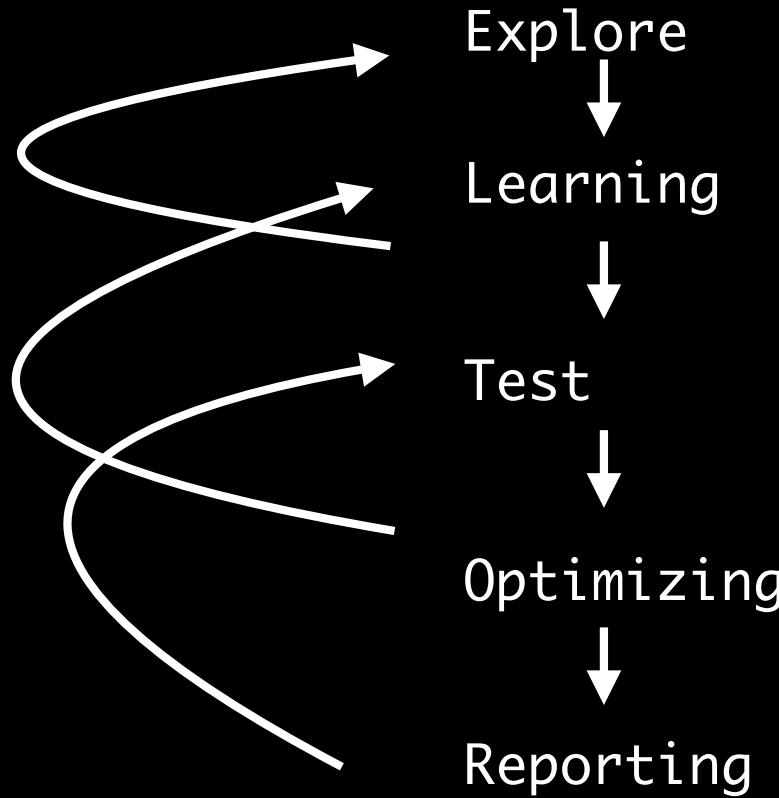
Reporting



descriptive:

predictive:

prescriptive:



things:  
what does DS team deliver?

- build data product
- build APIs
- impact roadmaps

data science @ The New York Times



chris.wiggins@columbia.edu  
chris.wiggins@nytimes.com  
@chrishwiggins

references: <http://bit.ly/stanf16>



## Lecture 2: predictive modeling @ NYT

## desc/pred/pres

---

descriptive:	specify $x$ ; learn $z(x)$ or $p(z x)$ where $z$ is “simpler” than $x$
predictive:	specify $x$ and $y$ ; learn to predict $y$ from $x$
prescriptive:	specify $x, y$ , and $a$ ; learn to prescribe $a$ given $x$ to maximize $y$

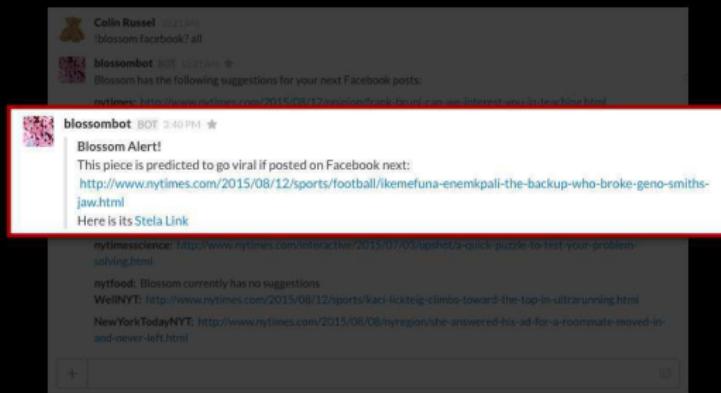
---

Figure 2: desc/pred/pres

- ▶ caveat: difference between observation and experiment. why?

# blossom example

prescriptive modeling, e.g.,

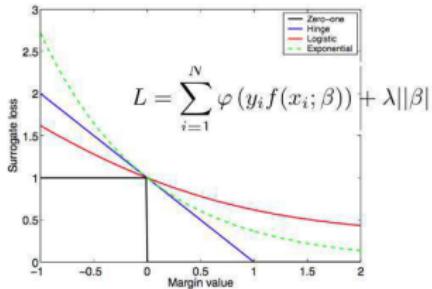


leverage methods which are predictive yet performant

Figure 3: Reminder: Blossom

## blossom + boosting ('exponential')

Margin-Based Surrogate Loss Functions



from "are you a bayesian or a frequentist"  
-michael jordan

Figure 4: Reminder: Surrogate Loss Functions

## tangent: logistic function as surrogate loss function

- ▶ define  $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$

## tangent: logistic function as surrogate loss function

- ▶ define  $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- ▶  $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$

## tangent: logistic function as surrogate loss function

- ▶ define  $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- ▶  $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- ▶  $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 (1 + e^{-y_i f(x_i)}) \equiv \sum_i \ell(y_i f(x_i))$

## tangent: logistic function as surrogate loss function

- ▶ define  $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- ▶  $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- ▶  $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + e^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$
- ▶  $\ell'' > 0, \ell(\mu) > 1[\mu < 0] \quad \forall \mu \in R.$

## tangent: logistic function as surrogate loss function

- ▶ define  $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- ▶  $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- ▶  $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + e^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$
- ▶  $\ell'' > 0, \ell(\mu) > 1[\mu < 0] \quad \forall \mu \in R.$
- ▶  $\therefore$  maximizing log-likelihood is minimizing a surrogate convex loss function for classification (though not strongly convex, cf. Yoram's talk)

## tangent: logistic function as surrogate loss function

- ▶ define  $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- ▶  $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- ▶  $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 (1 + e^{-y_i f(x_i)}) \equiv \sum_i \ell(y_i f(x_i))$
- ▶  $\ell'' > 0, \ell(\mu) > 1[\mu < 0] \forall \mu \in R.$
- ▶ ∴ maximizing log-likelihood is minimizing a surrogate convex loss function for classification (though not strongly convex, cf. Yoram's talk)
- ▶ but  $\sum_i \log_2 (1 + e^{-y_i w^T h(x_i)})$  not as easy as  $\sum_i e^{-y_i w^T h(x_i)}$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L[F] = \sum_i \exp(-y_i F(x_i))$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L[F] = \sum_i \exp(-y_i F(x_i))$
- ▶  $= \sum_i \exp(-y_i \sum_{t'}^t w_{t'} h_{t'}(x_i)) \equiv L_t(\mathbf{w}_t)$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L[F] = \sum_i \exp(-y_i F(x_i))$
- ▶  $= \sum_i \exp(-y_i \sum_{t'}^t w_{t'} h_{t'}(x_i)) \equiv L_t(\mathbf{w}_t)$
- ▶ Draw  $h_t \in \mathcal{H}$  large space of rules s.t.  $h(x) \in \{-1, +1\}$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L[F] = \sum_i \exp(-y_i F(x_i))$
- ▶  $= \sum_i \exp(-y_i \sum_{t'}^t w_{t'} h_{t'}(x_i)) \equiv L_t(\mathbf{w}_t)$
- ▶ Draw  $h_t \in \mathcal{H}$  large space of rules s.t.  $h(x) \in \{-1, +1\}$
- ▶ label  $y \in \{-1, +1\}$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i))$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces,  $L_1, L_\infty^{-1}, \dots$

---

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i))$
- ▶  $= \sum_{y=h'} d_i^t e^{-w} + \sum_{y \neq h'} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces,  $L_1, L_\infty^1, \dots$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i))$
- ▶  $= \sum_{y=h'} d_i^t e^{-w} + \sum_{y \neq h'} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-$
- ▶  $\therefore w_{t+1} = \operatorname{argmin}_w L_{t+1}(w) = (1/2) \log D_+/D_-$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces,  $L_1, L_\infty^1, \dots$

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i))$
- ▶  $= \sum_{y=h'} d_i^t e^{-w} + \sum_{y \neq h'} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-$
- ▶  $\therefore w_{t+1} = \operatorname{argmin}_w L_{t+1}(w) = (1/2) \log D_+/D_-$
- ▶  $L_{t+1}(\mathbf{w}_{t+1}) = 2\sqrt{D_+ D_-} = 2\sqrt{\nu_+(1-\nu_+)}/D$ , where  
 $0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces,  $L_1, L_\infty^1, \dots$

---

## boosting 1

$L$  exponential surrogate loss function, summed over examples:

- ▶  $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i))$
- ▶  $= \sum_{y=h'} d_i^t e^{-w} + \sum_{y \neq h'} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-$
- ▶  $\therefore w_{t+1} = \operatorname{argmin}_w L_{t+1}(w) = (1/2) \log D_+/D_-$
- ▶  $L_{t+1}(\mathbf{w}_{t+1}) = 2\sqrt{D_+ D_-} = 2\sqrt{\nu_+(1-\nu_+)}/D$ , where  
 $0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1$
- ▶ update example weights  $d_i^{t+1} = d_i^t e^{\mp w}$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces,  $L_1, L_\infty^1, \dots$

---

<sup>1</sup>Duchi + Singer “Boosting with structural sparsity” ICML ’09

## predicting people

- ▶ “customer journey” prediction

# predicting people

- ▶ “customer journey” prediction
  - ▶ fun covariates

# predicting people

- ▶ “customer journey” prediction
  - ▶ fun covariates
  - ▶ observational complication v structural models

## predicting people (reminder)

TFNAME	DB-MOTIF	MOTIF	DBNAME	d(p,q)
CBF1	CACGTG		YPD	0.032635
CGG everted repeat	CGGN*CCG		YPD	0.032821
MSN2	AGGGG		TRANSFAC	0.085626
HSF1	TTCNNNGAA		SCPD	0.102410
XBP1	TCGAG		TRANSFAC	0.140561

Figure 5: both in science and in real world, feature analysis guides future experiments

## single copy (reminder)

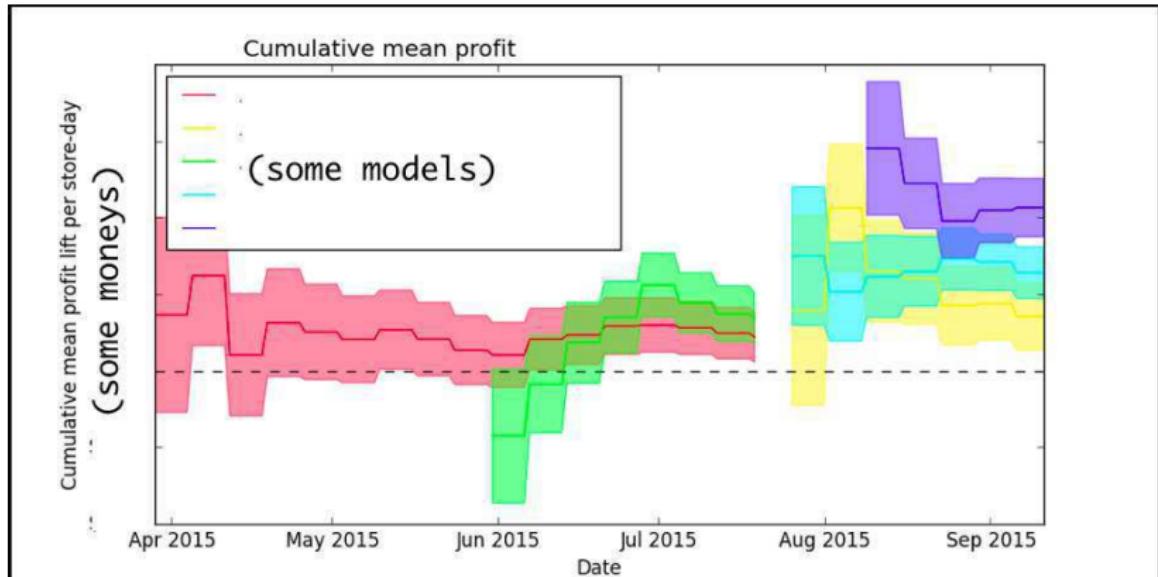


Figure 6: from Lecture 1

# example in CAR (computer assisted reporting)

[www.nytimes.com/2014/09/12/business/air-bag-flaw-long-known-led-to-recalls.html?\\_r=1](http://www.nytimes.com/2014/09/12/business/air-bag-flaw-long-known-led-to-recalls.html?_r=1)

S HOME SEARCH The New York Times BUSINESS DAY

## Air Bag Flaw, Long Known to Honda and Takata, Led to Recalls

By HIROKO TABUCHI SEPT. 11, 2014

f t



The air bag in Jennifer Griffin's Honda Civic was not among the recalled vehicles in 2008. Jim Keely

Figure 7: Tabuchi article

## example in CAR (computer assisted reporting)

- ▶ cf. Friedman's "Statistical models and Shoe Leather"<sup>2</sup>

---

<sup>2</sup>Freedman, David A. "Statistical models and shoe leather." *Sociological methodology* 21.2 (1991): 291-313.

## example in CAR (computer assisted reporting)

- ▶ cf. Friedman's "Statistical models and Shoe Leather"<sup>2</sup>
- ▶ Takata airbag fatalities

---

<sup>2</sup>Freedman, David A. "Statistical models and shoe leather." *Sociological methodology* 21.2 (1991): 291-313.

## example in CAR (computer assisted reporting)

- ▶ cf. Friedman's "Statistical models and Shoe Leather"<sup>2</sup>
- ▶ Takata airbag fatalities
- ▶ 2219 labeled<sup>3</sup> examples from 33,204 comments

---

<sup>2</sup>Freedman, David A. "Statistical models and shoe leather." *Sociological methodology* 21.2 (1991): 291-313.

<sup>3</sup>By Hiroko Tabuchi, a Pulitzer winner

## example in CAR (computer assisted reporting)

- ▶ cf. Friedman's "Statistical models and Shoe Leather"<sup>2</sup>
- ▶ Takata airbag fatalities
- ▶ 2219 labeled<sup>3</sup> examples from 33,204 comments
- ▶ cf. Box's "Science and Statistics"<sup>4</sup>

---

<sup>2</sup>Freedman, David A. "Statistical models and shoe leather." *Sociological methodology* 21.2 (1991): 291-313.

<sup>3</sup>By Hiroko Tabuchi, a Pulitzer winner

<sup>4</sup>Science and Statistics, George E. P. Box *Journal of the American Statistical Association*, Vol. 71, No. 356. (Dec., 1976), pp. 791-799.

# computer assisted reporting

## ► Impact

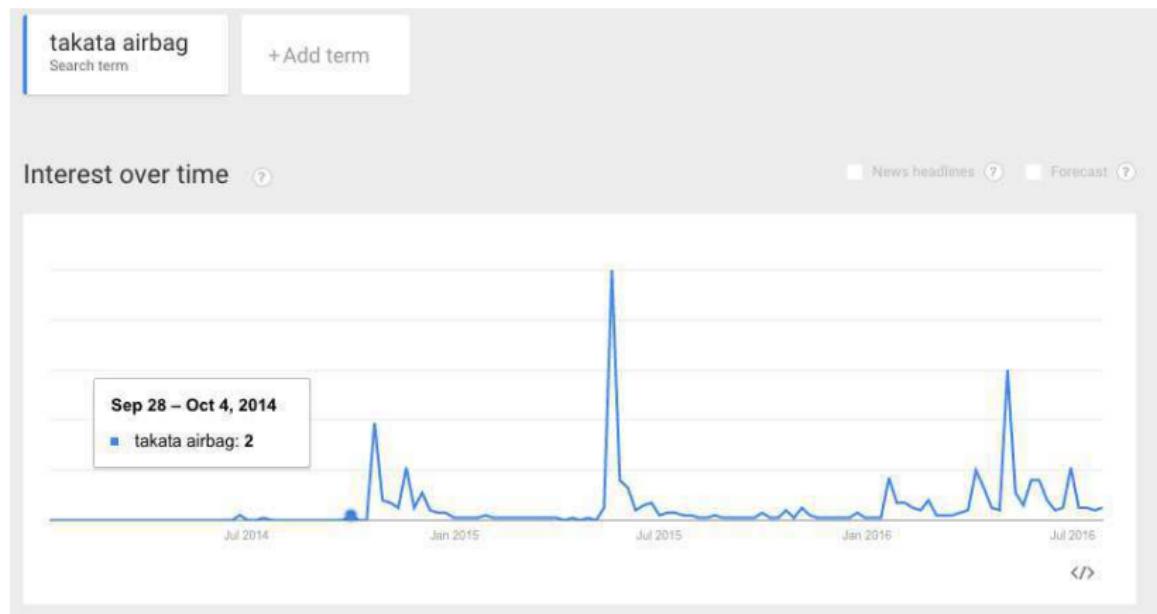


Figure 8: impact

## Lecture 3: prescriptive modeling @ NYT

## the natural abstraction

- ▶ operators<sup>5</sup> make decisions

---

<sup>5</sup>In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines

## the natural abstraction

- ▶ operators<sup>5</sup> make decisions
- ▶ faster horses v. cars

---

<sup>5</sup>In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines

## the natural abstraction

- ▶ operators<sup>5</sup> make decisions
- ▶ faster horses v. cars
- ▶ general insights v. optimal policies

---

<sup>5</sup>In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines

## maximizing outcome

- ▶ the problem: maximizing an outcome over policies...

## maximizing outcome

- ▶ the problem: maximizing an outcome over policies...
- ▶ ... while inferring causality from observation

## maximizing outcome

- ▶ the problem: maximizing an outcome over policies . . .
- ▶ . . . while inferring causality from observation
- ▶ different from predicting outcome in absence of action/policy

## examples

- ▶ observation is not experiment
-

## examples

- ▶ observation is not experiment
  - ▶ e.g., (Med.) smoking hurts vs unhealthy people smoke

---

## examples

- ▶ observation is not experiment
  - ▶ e.g., (Med.) smoking hurts vs unhealthy people smoke
  - ▶ e.g., (Med.) affluent get prescribed different meds/treatment

## examples

- ▶ observation is not experiment
  - ▶ e.g., (Med.) smoking hurts vs unhealthy people smoke
  - ▶ e.g., (Med.) affluent get prescribed different meds/treatment
  - ▶ e.g., (life) veterans earn less vs the rich serve less<sup>6</sup>

---

<sup>6</sup>Angrist, Joshua D. (1990). "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records". American Economic Review 80 (3): 313–336.

## examples

- ▶ observation is not experiment
  - ▶ e.g., (Med.) smoking hurts vs unhealthy people smoke
  - ▶ e.g., (Med.) affluent get prescribed different meds/treatment
  - ▶ e.g., (life) veterans earn less vs the rich serve less<sup>6</sup>
  - ▶ e.g., (life) admitted to school vs learn at school?

---

<sup>6</sup>Angrist, Joshua D. (1990). "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records". American Economic Review 80 (3): 313–336.

## reinforcement/machine learning/graphical models

- ▶ key idea: model joint  $p(y, a, x)$

## reinforcement/machine learning/graphical models

- ▶ key idea: model joint  $p(y, a, x)$
- ▶ explore/exploit: family of joints  $p_\alpha(y, a, x)$

## reinforcement/machine learning/graphical models

- ▶ key idea: model joint  $p(y, a, x)$
- ▶ explore/exploit: family of joints  $p_\alpha(y, a, x)$
- ▶ “causality”:  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$  “ $a$  causes  $y$ ”

## reinforcement/machine learning/graphical models

- ▶ key idea: model joint  $p(y, a, x)$
- ▶ explore/exploit: family of joints  $p_\alpha(y, a, x)$
- ▶ “causality”:  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$  “a causes y”
- ▶ nomenclature: ‘response’, ‘policy’/‘bias’, ‘prior’ above

in general

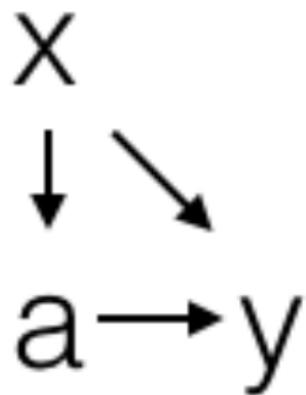


Figure 9: policy/bias, response, and prior define the distribution

also describes both the 'exploration' and 'exploitation' distributions

## randomized controlled trial

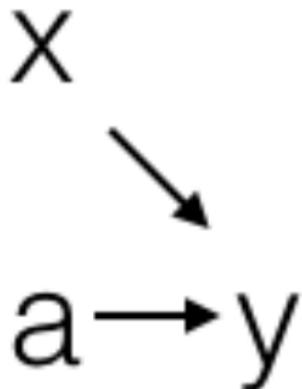


Figure 10: RCT: ‘bias’ removed, random ‘policy’ (response and prior unaffected)

also Pearl's ‘do’ distribution: a distribution with “no arrows” pointing to the action variable.

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation
  - ▶ aka “off policy estimation”

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation
  - ▶ aka “off policy estimation”
  - ▶ role of “IPW”

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation
  - ▶ aka “off policy estimation”
  - ▶ role of “IPW”
- ▶ reduction

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation
  - ▶ aka “off policy estimation”
  - ▶ role of “IPW”
- ▶ reduction
- ▶ normalization

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation
  - ▶ aka “off policy estimation”
  - ▶ role of “IPW”
- ▶ reduction
- ▶ normalization
- ▶ hyper-parameter searching

## POISE: calculation, estimation, optimization

- ▶ POISE: “policy optimization via importance sample estimation”
- ▶ Monte Carlo importance sampling estimation
  - ▶ aka “off policy estimation”
  - ▶ role of “IPW”
- ▶ reduction
- ▶ normalization
- ▶ hyper-parameter searching
- ▶ unexpected connection: personalized medicine

## POISE setup and Goal

- ▶ “ $a$  causes  $y$ ”  $\iff \exists$  family  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$

## POISE setup and Goal

- ▶ “a causes y”  $\iff \exists$  family  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- ▶ define off-policy/exploration distribution  
 $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$

## POISE setup and Goal

- ▶ “a causes y”  $\iff \exists$  family  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- ▶ define off-policy/exploration distribution  
 $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
- ▶ define exploitation distribution  
 $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$

## POISE setup and Goal

- ▶ “a causes y”  $\iff \exists$  family  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- ▶ define off-policy/exploration distribution  
 $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
- ▶ define exploitation distribution  
 $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$
- ▶ Goal: Maximize  $E_+(Y)$  over  $p_+(a|x)$  using data drawn from  $p_-(y, a, x)$ .

## POISE setup and Goal

- ▶ “a causes y”  $\iff \exists$  family  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- ▶ define off-policy/exploration distribution  
 $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
- ▶ define exploitation distribution  
 $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$
- ▶ Goal: Maximize  $E_+(Y)$  over  $p_+(a|x)$  using data drawn from  $p_-(y, a, x)$ .

## POISE setup and Goal

- ▶ “a causes y”  $\iff \exists$  family  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- ▶ define off-policy/exploration distribution  
 $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
- ▶ define exploitation distribution  
 $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$
- ▶ Goal: Maximize  $E_+(Y)$  over  $p_+(a|x)$  using data drawn from  $p_-(y, a, x)$ .

notation:  $\{x, a, y\} \in \{X, A, Y\}$  i.e.,  $E_\alpha(Y)$  is not a function of  $y$

## POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

- ▶  $E_+(Y) \equiv \sum_{yax} y p_+(y, a, x)$

## POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

- ▶  $E_+(Y) \equiv \sum_{yax} y p_+(y, a, x)$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(y, a, x) / p_-(y, a, x))$

## POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

- ▶  $E_+(Y) \equiv \sum_{yax} y p_+(y, a, x)$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(y, a, x) / p_-(y, a, x))$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(a|x) / p_-(a|x))$

## POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

- ▶  $E_+(Y) \equiv \sum_{yax} y p_+(y, a, x)$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(y, a, x) / p_-(y, a, x))$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(a|x) / p_-(a|x))$
- ▶  $E_+(Y) \approx N^{-1} \sum_i y_i (p_+(a_i|x_i) / p_-(a_i|x_i))$

## POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

- ▶  $E_+(Y) \equiv \sum_{yax} y p_+(y, a, x)$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(y, a, x) / p_-(y, a, x))$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(a|x) / p_-(a|x))$
- ▶  $E_+(Y) \approx N^{-1} \sum_i y_i (p_+(a_i|x_i) / p_-(a_i|x_i))$

## POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

- ▶  $E_+(Y) \equiv \sum_{yax} y p_+(y, a, x)$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(y, a, x) / p_-(y, a, x))$
- ▶  $E_+(Y) = \sum_{yax} y p_-(y, a, x) (p_+(a|x) / p_-(a|x))$
- ▶  $E_+(Y) \approx N^{-1} \sum_i y_i (p_+(a_i|x_i) / p_-(a_i|x_i))$

let's spend some time getting to know this last equation, the importance sampling estimate of outcome in a “causal model” (“a causes y”) among  $\{y, a, x\}$

## Observation (cf. Bottou<sup>7</sup> )

- ▶ factorizing  $P_{\pm}(x)$ :  $\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{+\text{but not-}}(x)}{P_{-\text{but not+}}(x)}$

## Observation (cf. Bottou<sup>7</sup> )

- ▶ factorizing  $P_{\pm}(x)$ :  $\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{+\text{but not-}}(x)}{P_{-\text{but not+}}(x)}$
- ▶ origin: importance sampling  $E_q(f) = E_p(fq/p)$  (as in variational methods)

## Observation (cf. Bottou<sup>7</sup> )

- ▶ factorizing  $P_{\pm}(x)$ :  $\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{+\text{but not-}}(x)}{P_{-\text{but not+}}(x)}$
- ▶ origin: importance sampling  $E_q(f) = E_p(fq/p)$  (as in variational methods)
- ▶ the “causal” model  $p_{\alpha}(y, a, x) = p(y|a, x)p_{\alpha}(a|x)p(x)$  helps here

## Observation (cf. Bottou<sup>7</sup> )

- ▶ factorizing  $P_{\pm}(x)$ :  $\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{+\text{but not-}}(x)}{P_{-\text{but not+}}(x)}$
- ▶ origin: importance sampling  $E_q(f) = E_p(fq/p)$  (as in variational methods)
- ▶ the “causal” model  $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$  helps here
- ▶ factors left over are numerator ( $p_+(a|x)$ , to optimize) and denominator ( $p_-(a|x)$ , to infer if not a RCT)

## Observation (cf. Bottou<sup>7</sup> )

- ▶ factorizing  $P_{\pm}(x)$ :  $\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{+\text{but not-}}(x)}{P_{-\text{but not+}}(x)}$
- ▶ origin: importance sampling  $E_q(f) = E_p(fq/p)$  (as in variational methods)
- ▶ the “causal” model  $p_{\alpha}(y, a, x) = p(y|a, x)p_{\alpha}(a|x)p(x)$  helps here
- ▶ factors left over are numerator ( $p_+(a|x)$ , to optimize) and denominator ( $p_-(a|x)$ , to infer if not a RCT)
- ▶ unobserved confounders will confound us (later)

---

<sup>7</sup>Counterfactual Reasoning and Learning Systems, arXiv:1209.2355

## Reduction (cf. Langford<sup>8, 9, 10</sup> ('05, '08, '09 ))

- ▶ consider numerator for deterministic policy:  
 $p_+(a|x) = 1[a = h(x)]$

---

## Reduction (cf. Langford<sup>8, 9, 10</sup> ('05, '08, '09 ))

- ▶ consider numerator for deterministic policy:  
 $p_+(a|x) = 1[a = h(x)]$
- ▶  $E_+(Y) \propto \sum_i (y_i / p_-(a|x)) 1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)]$

---

## Reduction (cf. Langford<sup>8, 9, 10</sup> ('05, '08, '09 ))

- ▶ consider numerator for deterministic policy:  
 $p_+(a|x) = 1[a = h(x)]$
  - ▶  $E_+(Y) \propto \sum_i (y_i / p_-(a|x)) 1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)]$
  - ▶ Note:  $1[c = d] = 1 - 1[c \neq d]$
-

## Reduction (cf. Langford<sup>8, 9, 10</sup> ('05, '08, '09 ))

- ▶ consider numerator for deterministic policy:  
 $p_+(a|x) = 1[a = h(x)]$
- ▶  $E_+(Y) \propto \sum_i (y_i / p_-(a|x)) 1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)]$
- ▶ Note:  $1[c = d] = 1 - 1[c \neq d]$
- ▶  $\therefore E_+(Y) \propto \text{constant} - \sum_i w_i 1[a \neq h(x)]$

---

## Reduction (cf. Langford<sup>8, 9, 10</sup> ('05, '08, '09 ))

- ▶ consider numerator for deterministic policy:  
 $p_+(a|x) = 1[a = h(x)]$
- ▶  $E_+(Y) \propto \sum_i (y_i / p_-(a|x)) 1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)]$
- ▶ Note:  $1[c = d] = 1 - 1[c \neq d]$
- ▶  $\therefore E_+(Y) \propto \text{constant} - \sum_i w_i 1[a \neq h(x)]$
- ▶  $\therefore$  reduces policy optimization to (weighted) classification

---

<sup>8</sup>Langford & Zadrozny "Relating Reinforcement Learning Performance to Classification Performance" ICML 2005

<sup>9</sup>Beygelzimer & Langford "The offset tree for learning with partial labels" (KDD 2009)

<sup>10</sup>Tutorial on "Reductions" (including at ICML 2009)

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i 1[a_i \neq h(x_i)]$

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- ▶ weight  $w_i = y_i / p_-(a_i|x_i)$ , inferred or RCT

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i \mathbf{1}[a_i \neq h(x_i)]$
- ▶ weight  $w_i = y_i / p_-(a_i|x_i)$ , inferred or RCT
- ▶ destroys measure by treating  $p_-(a|x)$  differently than  $1/p_-(a|x)$

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- ▶ weight  $w_i = y_i / p_-(a_i | x_i)$ , inferred or RCT
- ▶ destroys measure by treating  $p_-(a|x)$  differently than  $1/p_-(a|x)$
- ▶ normalize as  $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}{\sum_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}$

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- ▶ weight  $w_i = y_i / p_-(a_i | x_i)$ , inferred or RCT
- ▶ destroys measure by treating  $p_-(a|x)$  differently than  $1/p_-(a|x)$
- ▶ normalize as  $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}{\sum_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}$
- ▶ destroys lovely reduction

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- ▶ weight  $w_i = y_i / p_-(a_i | x_i)$ , inferred or RCT
- ▶ destroys measure by treating  $p_-(a|x)$  differently than  $1/p_-(a|x)$
- ▶ normalize as  $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}{\sum_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}$
- ▶ destroys lovely reduction
- ▶ simply<sup>11</sup>  $L(\lambda) = \sum_i (y_i - \lambda) 1[a_i \neq h(x_i)] / p_-(a_i | x_i)$

---

<sup>11</sup>Suggestion by Dan Hsu

## Reduction w/optimistic complication

- ▶ Prescription  $\iff$  classification  $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- ▶ weight  $w_i = y_i / p_-(a_i | x_i)$ , inferred or RCT
- ▶ destroys measure by treating  $p_-(a|x)$  differently than  $1/p_-(a|x)$
- ▶ normalize as  $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}{\sum_i 1[a_i \neq h(x_i)] / p_-(a_i | x_i)}$
- ▶ destroys lovely reduction
- ▶ simply<sup>11</sup>  $L(\lambda) = \sum_i (y_i - \lambda) 1[a_i \neq h(x_i)] / p_-(a_i | x_i)$
- ▶ hidden here is a 2nd parameter, in classification,  $\therefore$  harder search

---

<sup>11</sup>Suggestion by Dan Hsu

## POISE punchlines

- ▶ allows policy planning even with implicit logged exploration data<sup>12</sup>

---

<sup>12</sup>Strehl, Alex, et al. "Learning from logged implicit exploration data." Advances in Neural Information Processing Systems. 2010.

## POISE punchlines

- ▶ allows policy planning even with implicit logged exploration data<sup>12</sup>
- ▶ e.g., two hospital story

---

<sup>12</sup>Strehl, Alex, et al. "Learning from logged implicit exploration data." Advances in Neural Information Processing Systems. 2010.

## POISE punchlines

- ▶ allows policy planning even with implicit logged exploration data<sup>12</sup>
- ▶ e.g., two hospital story
- ▶ “personalized medicine” is also a policy

---

<sup>12</sup>Strehl, Alex, et al. “Learning from logged implicit exploration data.” Advances in Neural Information Processing Systems. 2010.

## POISE punchlines

- ▶ allows policy planning even with implicit logged exploration data<sup>12</sup>
- ▶ e.g., two hospital story
- ▶ “personalized medicine” is also a policy
- ▶ abundant data available, under-explored IMHO

---

<sup>12</sup>Strehl, Alex, et al. “Learning from logged implicit exploration data.” Advances in Neural Information Processing Systems. 2010.

## tangent: causality as told by an economist

different, related goal

- ▶ they think in terms of ATE/ITE instead of policy

## tangent: causality as told by an economist

different, related goal

- ▶ they think in terms of ATE/ITE instead of policy
  - ▶ ATE

## tangent: causality as told by an economist

different, related goal

- ▶ they think in terms of ATE/ITE instead of policy
  - ▶ ATE
    - ▶  $\tau \equiv E_0(Y|a=1) - E_0(Y|a=0) \equiv Q(a=1) - Q(a=0)$

## tangent: causality as told by an economist

different, related goal

- ▶ they think in terms of ATE/ITE instead of policy
  - ▶ ATE
    - ▶  $\tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0)$
  - ▶ CATE aka Individualized Treatment Effect (ITE)

## tangent: causality as told by an economist

different, related goal

- ▶ they think in terms of ATE/ITE instead of policy
  - ▶ ATE
    - ▶  $\tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0)$
  - ▶ CATE aka Individualized Treatment Effect (ITE)
    - ▶  $\tau(x) \equiv E_0(Y|a = 1, x) - E_0(Y|a = 0, x)$

# tangent: causality as told by an economist

different, related goal

- ▶ they think in terms of ATE/ITE instead of policy
  - ▶ ATE
    - ▶  $\tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0)$
  - ▶ CATE aka Individualized Treatment Effect (ITE)
    - ▶  $\tau(x) \equiv E_0(Y|a = 1, x) - E_0(Y|a = 0, x)$
    - ▶  $\equiv Q(a = 1, x) - Q(a = 0, x)$

## *Q*-note: “generalizing” Monte Carlo w/kernels

- ▶ MC:  $E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i)$

## *Q*-note: “generalizing” Monte Carlo w/kernels

- ▶ MC:  $E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i)$
- ▶ K:  $p \approx N^{-1} \sum_i K(x|x_i)$

## *Q*-note: “generalizing” Monte Carlo w/kernels

- ▶ MC:  $E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i)$
- ▶  $K$ :  $p \approx N^{-1} \sum_i K(x|x_i)$
- ▶  $\Rightarrow \sum_x p(x)f(x) \approx N^{-1} \sum_i \sum_x f(x)K(x|x_i)$

## *Q*-note: “generalizing” Monte Carlo w/kernels

- ▶ MC:  $E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i)$
- ▶  $K$ :  $p \approx N^{-1} \sum_i K(x|x_i)$
- ▶  $\Rightarrow \sum_x p(x)f(x) \approx N^{-1} \sum_i \sum_x f(x)K(x|x_i)$
- ▶  $K$  can be any normalized function, e.g.,  $K(x|x_i) = \delta_{x,x_i}$ , which yields MC.

## *Q*-note: “generalizing” Monte Carlo w/kernels

- ▶ MC:  $E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i)$
- ▶  $K$ :  $p \approx N^{-1} \sum_i K(x|x_i)$
- ▶  $\Rightarrow \sum_x p(x)f(x) \approx N^{-1} \sum_i \sum_x f(x)K(x|x_i)$
- ▶  $K$  can be any normalized function, e.g.,  $K(x|x_i) = \delta_{x,x_i}$ , which yields MC.
- ▶ multivariate  
$$E_p(f) \approx N^{-1} \sum_i \sum_{yax} f(y, a, x) K_1(y|y_i) K_2(a|a_i) K_3(x|x_i)$$

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶  $Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶ 
$$Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$$
- ▶ 
$$= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x)$$

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶  $Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$
- ▶  $= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x)$
- ▶ stratify  $x$  using  $z(x)$  such that  $\cup z = X$ , and  $\cap z, z' =$

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶  $Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$
- ▶  $= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x)$
- ▶ stratify  $x$  using  $z(x)$  such that  $\cup z = X$ , and  $\cap z, z' =$
- ▶  $n(x) = \sum_i 1[z(x_i) = z(x)]$  = number of points in  $x$ 's stratum

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶  $Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$
- ▶  $= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x)$
- ▶ stratify  $x$  using  $z(x)$  such that  $\cup z = X$ , and  $\cap z, z' =$
- ▶  $n(x) = \sum_i 1[z(x_i) = z(x)]$  = number of points in  $x$ 's stratum
- ▶  $\Omega(x) = \sum_{x'} 1[z(x') = z(x)]$  = area of  $x$ 's stratum

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶  $Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$
- ▶  $= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x)$
- ▶ stratify  $x$  using  $z(x)$  such that  $\cup z = X$ , and  $\cap z, z' =$
- ▶  $n(x) = \sum_i 1[z(x_i) = z(x)]$  = number of points in  $x$ 's stratum
- ▶  $\Omega(x) = \sum_{x'} 1[z(x') = z(x)]$  = area of  $x$ 's stratum
- ▶  $\therefore K_3(x|x_i) = 1[z(x) = z(x_i)]/\Omega(x)$

## *Q*-note: application w/strata+matching, setup

Helps think about economists' approach:

- ▶  $Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)}$
- ▶  $= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x)$
- ▶ stratify  $x$  using  $z(x)$  such that  $\cup z = X$ , and  $\cap z, z' =$
- ▶  $n(x) = \sum_i 1[z(x_i) = z(x)]$  = number of points in  $x$ 's stratum
- ▶  $\Omega(x) = \sum_{x'} 1[z(x') = z(x)]$  = area of  $x$ 's stratum
- ▶  $\therefore K_3(x|x_i) = 1[z(x) = z(x_i)]/\Omega(x)$
- ▶ as in  $MC$ ,  $K_1(y|y_i) = \delta_{y,y_i}$ ,  $K_2(a|a_i) = \delta_{a,a_i}$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

“matching” means: choose each  $z$  to contain 1 positive example & 1 negative example,

- ▶  $p_-(a|x) \approx 1/2, n(x) = 2$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

“matching” means: choose each  $z$  to contain 1 positive example & 1 negative example,

- ▶  $p_-(a|x) \approx 1/2, n(x) = 2$
- ▶  $\therefore \tau(a, x) = Q(a=1, x) - Q(a=0, x) = y_1(x) - y_0(x)$

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

“matching” means: choose each  $z$  to contain 1 positive example & 1 negative example,

- ▶  $p_-(a|x) \approx 1/2, n(x) = 2$
- ▶  $\therefore \tau(a, x) = Q(a=1, x) - Q(a=0, x) = y_1(x) - y_0(x)$
- ▶  $z$ -generalizations: graphs, digraphs, k-NN, “matching”

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

“matching” means: choose each  $z$  to contain 1 positive example & 1 negative example,

- ▶  $p_-(a|x) \approx 1/2, n(x) = 2$
- ▶  $\therefore \tau(a, x) = Q(a=1, x) - Q(a=0, x) = y_1(x) - y_0(x)$
- ▶  $z$ -generalizations: graphs, digraphs, k-NN, “matching”
- ▶  $K$ -generalizations: continuous  $a$ , any metric or similarity you like,...

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

“matching” means: choose each  $z$  to contain 1 positive example & 1 negative example,

- ▶  $p_-(a|x) \approx 1/2, n(x) = 2$
- ▶  $\therefore \tau(a, x) = Q(a=1, x) - Q(a=0, x) = y_1(x) - y_0(x)$
- ▶  $z$ -generalizations: graphs, digraphs, k-NN, “matching”
- ▶  $K$ -generalizations: continuous  $a$ , any metric or similarity you like,...

## *Q*-note: application w/strata+matching, payoff

- ▶  $\sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- ▶  $p(x) \approx (n(x)/N) \Omega(x)^{-1}$
- ▶  $\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

“matching” means: choose each  $z$  to contain 1 positive example & 1 negative example,

- ▶  $p_-(a|x) \approx 1/2, n(x) = 2$
- ▶  $\therefore \tau(a, x) = Q(a=1, x) - Q(a=0, x) = y_1(x) - y_0(x)$
- ▶  $z$ -generalizations: graphs, digraphs, k-NN, “matching”
- ▶  $K$ -generalizations: continuous  $a$ , any metric or similarity you like,...

IMHO underexplored

## causality, as understood in marketing

- ▶ a/b testing and RCT

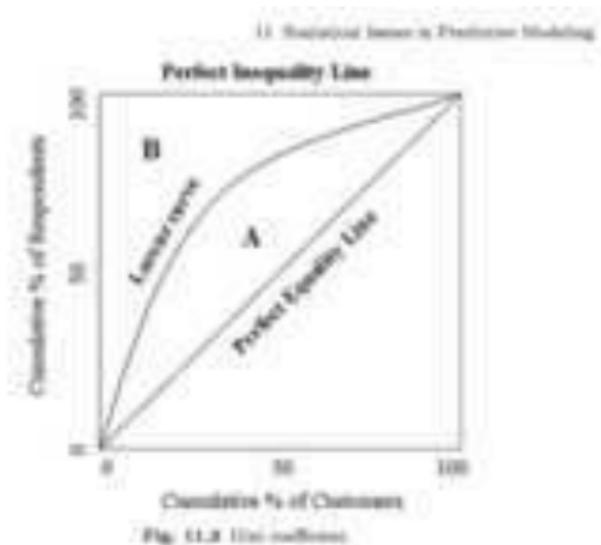


Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin.  
Database Marketing, Springer New York, 2008

## causality, as understood in marketing

- ▶ a/b testing and RCT
- ▶ yield optimization

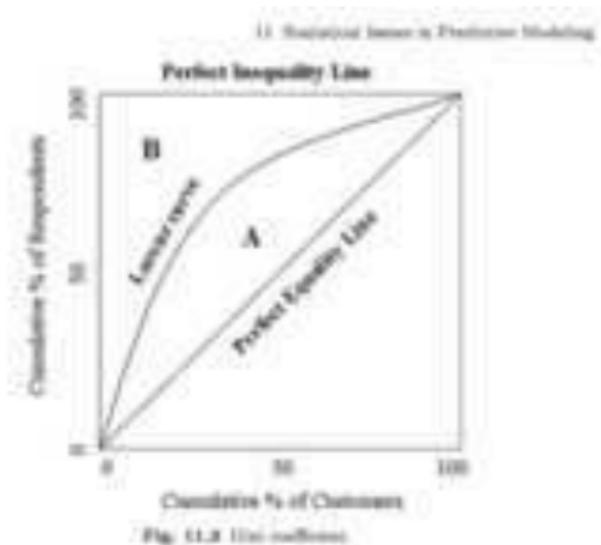


Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin.  
Database Marketing, Springer New York, 2008

## causality, as understood in marketing

- ▶ a/b testing and RCT
- ▶ yield optimization
- ▶ Lorenz curve (vs ROC plots)

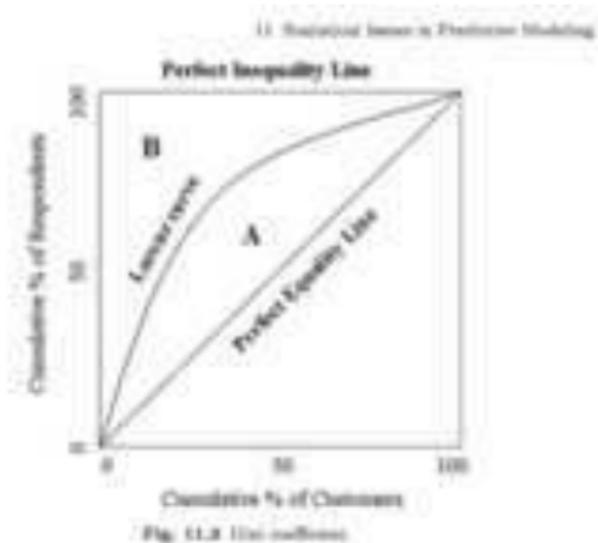


Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin.  
Database Marketing, Springer New York, 2008

## unobserved confounders vs. “causality” modeling

- ▶ truth:  $p_\alpha(y, a, x, u) = p(y|a, x, u)p_\alpha(a|x, u)p(x, u)$

## unobserved confounders vs. “causality” modeling

- ▶ truth:  $p_\alpha(y, a, x, u) = p(y|a, x, u)p_\alpha(a|x, u)p(x, u)$
- ▶ but:  $p_+(y, a, x, u) = p(y|a, x, u)p_-(a|x)p(x, u)$

## unobserved confounders vs. “causality” modeling

- ▶ truth:  $p_\alpha(y, a, x, u) = p(y|a, x, u)p_\alpha(a|x, u)p(x, u)$
- ▶ but:  $p_+(y, a, x, u) = p(y|a, x, u)p_-(a|x)p(x, u)$
- ▶  $E_+(Y) \equiv \sum_{yaxu} y p_+(yaxu) \approx N^{-1} \sum_{i \sim p_-} y_i p_+(a|x) / p_-(a|x, u)$

## unobserved confounders vs. “causality” modeling

- ▶ truth:  $p_\alpha(y, a, x, u) = p(y|a, x, u)p_\alpha(a|x, u)p(x, u)$
- ▶ but:  $p_+(y, a, x, u) = p(y|a, x, u)p_-(a|x)p(x, u)$
- ▶  $E_+(Y) \equiv \sum_{yaxu} y p_+(yaxu) \approx N^{-1} \sum_{i \sim p_-} y_i p_+(a|x) / p_-(a|x, u)$
- ▶ denominator can not be inferred, ignore at your peril

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)

---

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)
  - ▶  $x$ : gender ( $x=1$ : female,  $x=0$ : male)
-

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)
  - ▶  $x$ : gender ( $x=1$ : female,  $x=0$ : male)
  - ▶ lawsuit (1973):  $.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = .35$
-

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)
- ▶  $x$ : gender ( $x=1$ : female,  $x=0$ : male)
- ▶ lawsuit (1973):  $.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = .35$
- ▶ 'resolved' by Bickel (1975)<sup>13</sup> (See also Pearl<sup>14</sup> )

---

<sup>13</sup>P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* 187 (4175): 398–404

<sup>14</sup>Pearl, Judea (December 2013). "Understanding Simpson's paradox". UCLA Cognitive Systems Laboratory, Technical Report R-414.

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)
- ▶  $x$ : gender ( $x=1$ : female,  $x=0$ : male)
- ▶ lawsuit (1973):  $.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = .35$
- ▶ 'resolved' by Bickel (1975)<sup>13</sup> (See also Pearl<sup>14</sup> )
- ▶  $u$ : unobserved department they applied to

---

<sup>13</sup>P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* 187 (4175): 398–404

<sup>14</sup>Pearl, Judea (December 2013). "Understanding Simpson's paradox". UCLA Cognitive Systems Laboratory, Technical Report R-414.

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)
- ▶  $x$ : gender ( $x=1$ : female,  $x=0$ : male)
- ▶ lawsuit (1973):  $.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = .35$
- ▶ 'resolved' by Bickel (1975)<sup>13</sup> (See also Pearl<sup>14</sup> )
- ▶  $u$ : unobserved department they applied to
- ▶  $p(a|x) = \sum_{u=1}^{u=6} p(a|x, u)p(u|x)$

---

<sup>13</sup>P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* 187 (4175): 398–404

<sup>14</sup>Pearl, Judea (December 2013). "Understanding Simpson's paradox". UCLA Cognitive Systems Laboratory, Technical Report R-414.

## cautionary tale problem: Simpson's paradox

- ▶  $a$ : admissions ( $a=1$ : admitted,  $a=0$ : declined)
- ▶  $x$ : gender ( $x=1$ : female,  $x=0$ : male)
- ▶ lawsuit (1973):  $.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = .35$
- ▶ 'resolved' by Bickel (1975)<sup>13</sup> (See also Pearl<sup>14</sup> )
- ▶  $u$ : unobserved department they applied to
- ▶  $p(a|x) = \sum_{u=1}^{u=6} p(a|x, u)p(u|x)$
- ▶ e.g., gender-blind:  $p(a|1) - p(a|0) = p(a|u) \cdot (p(u|1) - p(u|0))$

---

<sup>13</sup>P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* 187 (4175): 398–404

<sup>14</sup>Pearl, Judea (December 2013). "Understanding Simpson's paradox". UCLA Cognitive Systems Laboratory, Technical Report R-414.

## confounded approach: quasi-experiments + instruments<sup>17</sup>

- ▶ Q: does engagement drive retention? (NYT, NFLX, ...)

---

## confounded approach: quasi-experiments + instruments<sup>17</sup>

- ▶ Q: does engagement drive retention? (NYT, NFLX, ...)
  - ▶ we don't directly control engagement

---

## confounded approach: quasi-experiments + instruments<sup>17</sup>

- ▶ Q: does engagement drive retention? (NYT, NFLX, ...)
  - ▶ we don't directly control engagement
  - ▶ nonetheless useful since many things can influence it

---

- ▶ Q: does engagement drive retention? (NYT, NFLX, . . .)
  - ▶ we don't directly control engagement
  - ▶ nonetheless useful since many things can influence it
- ▶ Q: does serving in Vietnam war decrease earnings<sup>15</sup>?

---

<sup>15</sup>Angrist, Joshua D. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." *The American Economic Review* (1990): 313-336.

- ▶ Q: does engagement drive retention? (NYT, NFLX, . . .)
  - ▶ we don't directly control engagement
  - ▶ nonetheless useful since many things can influence it
- ▶ Q: does serving in Vietnam war decrease earnings<sup>15</sup>?
  - ▶ US didn't directly control serving in Vietnam, either<sup>16</sup>

---

<sup>15</sup>Angrist, Joshua D. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." *The American Economic Review* (1990): 313-336.

<sup>16</sup>cf., George Bush, Donald Trump, Bill Clinton, Dick Cheney...

- ▶ Q: does engagement drive retention? (NYT, NFLX, . . .)
  - ▶ we don't directly control engagement
  - ▶ nonetheless useful since many things can influence it
- ▶ Q: does serving in Vietnam war decrease earnings<sup>15</sup>?
  - ▶ US didn't directly control serving in Vietnam, either<sup>16</sup>
- ▶ requires **strong assumptions**, including linear model

---

<sup>15</sup>Angrist, Joshua D. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." *The American Economic Review* (1990): 313-336.

<sup>16</sup>cf., George Bush, Donald Trump, Bill Clinton, Dick Cheney...

<sup>17</sup>I thank Sinan Aral, MIT Sloan, for bringing this to my attention

## IV: graphical model assumption

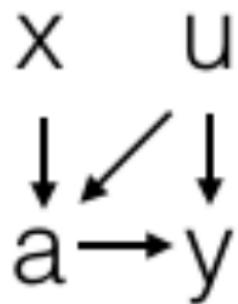


Figure 12: independence assumption

## IV: graphical model assumption (sideways)

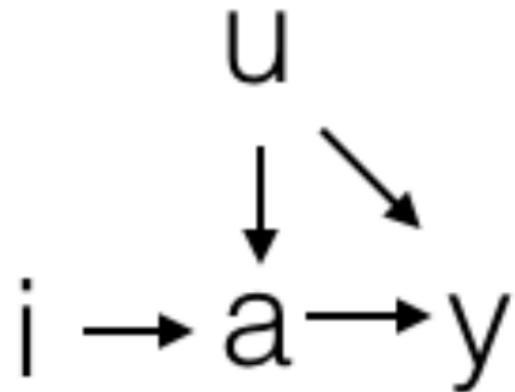


Figure 13: independence assumption

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ *a endogenous*

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ *a endogenous*
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ *a endogenous*
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$
- ▶ linear ansatz:  $y = \beta^T a + \epsilon$

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ *a endogenous*
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$
- ▶ linear ansatz:  $y = \beta^T a + \epsilon$
- ▶ if *a exogenous* (e.g., OLS), use  $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$   
(note that  $E[A_j A_k]$  gives square matrix; invert for  $\beta$ )

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ *a endogenous*
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$
- ▶ linear ansatz:  $y = \beta^T a + \epsilon$
- ▶ if *a* exogenous (e.g., OLS), use  $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$   
(note that  $E[A_j A_k]$  gives square matrix; invert for  $\beta$ )
- ▶ add *instrument*  $x$  uncorrelated with  $\epsilon$

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ a endogenous
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$
- ▶ linear ansatz:  $y = \beta^T a + \epsilon$
- ▶ if  $a$  exogenous (e.g., OLS), use  $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$   
(note that  $E[A_j A_k]$  gives square matrix; invert for  $\beta$ )
- ▶ add instrument  $x$  uncorrelated with  $\epsilon$
- ▶  $E[YX_k] = E[\beta^T AX_k] + E[\epsilon]E[X_k]$

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ a endogenous
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$
- ▶ linear ansatz:  $y = \beta^T a + \epsilon$
- ▶ if  $a$  exogenous (e.g., OLS), use  $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$   
(note that  $E[A_j A_k]$  gives square matrix; invert for  $\beta$ )
- ▶ add instrument  $x$  uncorrelated with  $\epsilon$
- ▶  $E[YX_k] = E[\beta^T AX_k] + E[\epsilon]E[X_k]$
- ▶  $E[Y] = E[\beta^T A] + E[\epsilon]$  (from ansatz)

## IV: review s/OLS/MOM/ ( $E$ is empirical average)

- ▶ a endogenous
  - ▶ e.g.,  $\exists u$  s.t.  $p(y|a, x, u), p(a|x, u)$
- ▶ linear ansatz:  $y = \beta^T a + \epsilon$
- ▶ if  $a$  exogenous (e.g., OLS), use  $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$   
(note that  $E[A_j A_k]$  gives square matrix; invert for  $\beta$ )
- ▶ add instrument  $x$  uncorrelated with  $\epsilon$
- ▶  $E[YX_k] = E[\beta^T AX_k] + E[\epsilon]E[X_k]$
- ▶  $E[Y] = E[\beta^T A] + E[\epsilon]$  (from ansatz)
- ▶  $C(Y, X_k) = \beta^T C(A, X_k)$ , not an “inversion” problem, requires  
“two stage regression”

## IV: binary, binary case (aka “Wald estimator”)

- ▶  $y = \beta a + \epsilon$

## IV: binary, binary case (aka “Wald estimator”)

- ▶  $y = \beta a + \epsilon$
- ▶  $E(Y|x) = \beta E(A|x) + E(\epsilon)$ , evaluate at  $x = \{0, 1\}$

## IV: binary, binary case (aka “Wald estimator”)

- ▶  $y = \beta a + \epsilon$
- ▶  $E(Y|x) = \beta E(A|x) + E(\epsilon)$ , evaluate at  $x = \{0, 1\}$
- ▶  $\beta = (E(Y|x=1) - E(Y|x=0))/(E(A|x=1) - E(A|x=0))$ .

## bandits: obligatory slide

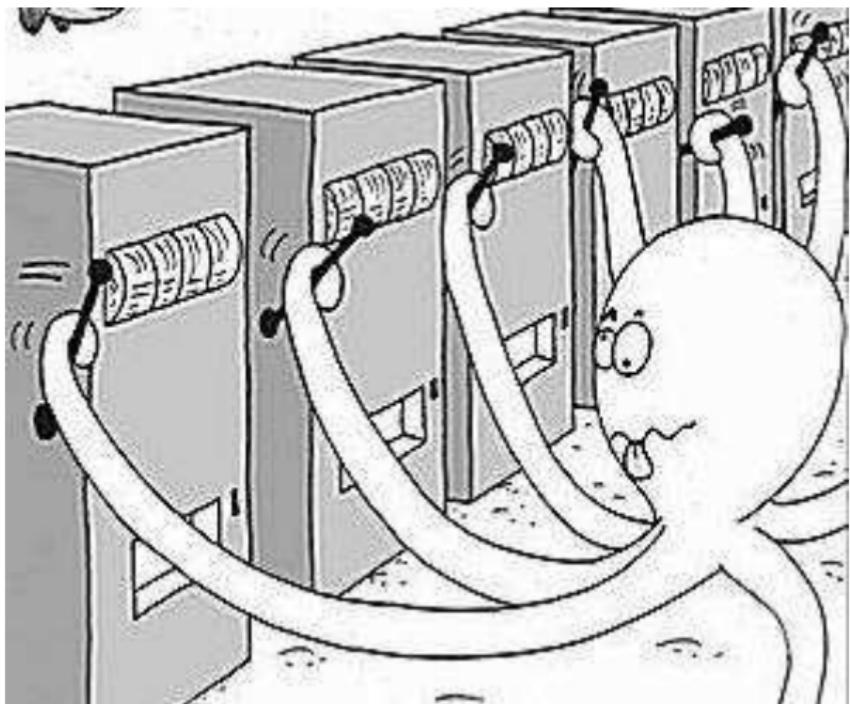


Figure 14: almost all the talks I've gone to on bandits have this image

## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .



## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .
  - ▶ replace meetings with code
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .
  - ▶ replace meetings with code
  - ▶ requires software engineering to replace decisions with, e.g., Javascript
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .
  - ▶ replace meetings with code
  - ▶ requires software engineering to replace decisions with, e.g., Javascript
  - ▶ most useful if decisions or items get “stale” quickly
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .
  - ▶ replace meetings with code
  - ▶ requires software engineering to replace decisions with, e.g., Javascript
  - ▶ most useful if decisions or items get “stale” quickly
  - ▶ less useful for one-off, major decisions to be “interpreted”
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .
  - ▶ replace meetings with code
  - ▶ requires software engineering to replace decisions with, e.g., Javascript
  - ▶ most useful if decisions or items get “stale” quickly
  - ▶ less useful for one-off, major decisions to be “interpreted”
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, . . .
- ▶ replace meetings with code
- ▶ requires software engineering to replace decisions with, e.g., Javascript
- ▶ most useful if decisions or items get “stale” quickly
- ▶ less useful for one-off, major decisions to be “interpreted”

## examples

- ▶  $\epsilon$ -greedy (no context, aka ‘vanilla’, aka ‘context-free’)
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, ...
- ▶ replace meetings with code
- ▶ requires software engineering to replace decisions with, e.g., Javascript
- ▶ most useful if decisions or items get “stale” quickly
- ▶ less useful for one-off, major decisions to be “interpreted”

## examples

- ▶  $\epsilon$ -greedy (no context, aka ‘vanilla’, aka ‘context-free’)
  - ▶ UCB1 (2002) (no context) + LinUCB (with context)
-

## bandits

- ▶ wide applicability: humane clinical trials, targeting, ...
- ▶ replace meetings with code
- ▶ requires software engineering to replace decisions with, e.g., Javascript
- ▶ most useful if decisions or items get “stale” quickly
- ▶ less useful for one-off, major decisions to be “interpreted”

### examples

- ▶  $\epsilon$ -greedy (no context, aka ‘vanilla’, aka ‘context-free’)
- ▶ UCB1 (2002) (no context) + LinUCB (with context)
- ▶ Thompson Sampling (1933)<sup>18, 19, 20</sup> (general, with or without context)

---

<sup>18</sup> Thompson, William R. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. *Biometrika*, 25(3–4):285–294, 1933.

<sup>19</sup> AKA “probability matching”, “posterior sampling”

<sup>20</sup> cf., “Bayesian Bandit Explorer” ([link](#))

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods
- ▶ In the economics approach we focus on

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods
- ▶ In the economics approach we focus on
- ▶  $\tau(\dots) \equiv Q(a=1, \dots) - Q(a=0, \dots)$  “treatment effect”

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods
- ▶ In the economics approach we focus on
  - ▶  $\tau(\dots) \equiv Q(a=1, \dots) - Q(a=0, \dots)$  “treatment effect”
  - ▶ where  $Q(a, \dots) = \sum_y y p(y| \dots)$

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods
- ▶ In the economics approach we focus on
  - ▶  $\tau(\dots) \equiv Q(a=1, \dots) - Q(a=0, \dots)$  “treatment effect”
  - ▶ where  $Q(a, \dots) = \sum_y y p(y| \dots)$

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods
- ▶ In the economics approach we focus on
  - ▶  $\tau(\dots) \equiv Q(a=1, \dots) - Q(a=0, \dots)$  “treatment effect”
  - ▶ where  $Q(a, \dots) = \sum_y y p(y| \dots)$

In Thompson sampling we will generate 1 datum at a time, by

- ▶ asserting a parameterized generative model for  $p(y|a, x, \theta)$

## TS: connecting w/“generative causal modeling” 0

- ▶ WAS  $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- ▶ These 3 terms were treated by
  - ▶ response  $p(y|a, x)$ : avoid regression/inferring using importance sampling
  - ▶ policy  $p_\alpha(a|x)$ : optimize ours, infer theirs
  - ▶ (NB: ours was deterministic:  $p(a|x) = 1[a = h(x)]$ )
  - ▶ prior  $p(x)$ : either avoid by importance sampling or estimate via kernel methods
- ▶ In the economics approach we focus on
  - ▶  $\tau(\dots) \equiv Q(a=1, \dots) - Q(a=0, \dots)$  “treatment effect”
  - ▶ where  $Q(a, \dots) = \sum_y y p(y| \dots)$

In Thompson sampling we will generate 1 datum at a time, by

- ▶ asserting a parameterized generative model for  $p(y|a, x, \theta)$
- ▶ using a deterministic but averaged policy

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$
- ▶ idea: use prior data  $D = \{y, a, x\}_1^t$  to define *non-deterministic* policy:

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$
- ▶ idea: use prior data  $D = \{y, a, x\}_1^t$  to define *non-deterministic* policy:
  - ▶  $p(a|x) = \int d\theta p(a|x, \theta) p(\theta|D)$

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$
- ▶ idea: use prior data  $D = \{y, a, x\}_1^t$  to define *non-deterministic* policy:
  - ▶  $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
  - ▶  $p(a|x) = \int d\theta 1[a = \operatorname{argmax}_{a'} Q(a', x, \theta)]p(\theta|D)$

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$
- ▶ idea: use prior data  $D = \{y, a, x\}_1^t$  to define *non-deterministic* policy:
  - ▶  $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
  - ▶  $p(a|x) = \int d\theta 1[a = \operatorname{argmax}_{a'} Q(a', x, \theta)]p(\theta|D)$
- ▶ hold up:

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$
- ▶ idea: use prior data  $D = \{y, a, x\}_1^t$  to define *non-deterministic* policy:
  - ▶  $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
  - ▶  $p(a|x) = \int d\theta 1[a = \operatorname{argmax}_{a'} Q(a', x, \theta)]p(\theta|D)$
- ▶ hold up:
  - ▶ Q1: what's  $p(\theta|D)$ ?

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 1

- ▶ model true world response function  $p(y|a, x)$  parametrically as  $p(y|a, x, \theta^*)$
- ▶ (i.e.,  $\theta^*$  is the true value of the parameter)<sup>21</sup>
- ▶ if you knew  $\theta$ :
  - ▶ could compute  $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$  directly
  - ▶ then choose  $h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)$
  - ▶ inducing policy  $p(a|x, \theta) = 1[a = h(x; \theta) = \operatorname{argmax}_a Q(a, x, \theta)]$
- ▶ idea: use prior data  $D = \{y, a, x\}_1^t$  to define *non-deterministic* policy:
  - ▶  $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
  - ▶  $p(a|x) = \int d\theta 1[a = \operatorname{argmax}_{a'} Q(a', x, \theta)]p(\theta|D)$
- ▶ hold up:
  - ▶ Q1: what's  $p(\theta|D)$ ?
  - ▶ Q2: how am I going to evaluate this integral?

---

<sup>21</sup>Note that  $\theta$  is a vector, with components for each action.

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$
  - ▶  $= p(\theta|\alpha) \prod_t p(y_t|a_t, x_t, \theta)$

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$
  - ▶  $= p(\theta|\alpha) \prod_t p(y_t | a_t, x_t, \theta)$
  - ▶ warning 1: sometimes people write “ $p(D|\theta)$ ” but we don't need  $p(a|\theta)$  or  $p(x|\theta)$  here

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$
  - ▶  $= p(\theta|\alpha) \prod_t p(y_t | a_t, x_t, \theta)$
  - ▶ warning 1: sometimes people write “ $p(D|\theta)$ ” but we don't need  $p(a|\theta)$  or  $p(x|\theta)$  here
  - ▶ warning 2: don't need historical record of  $\theta_t$ .

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶ 
$$p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$$
  - ▶ 
$$\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$$
  - ▶ 
$$= p(\theta|\alpha) \prod_t p(y_t | a_t, x_t, \theta)$$
  - ▶ warning 1: sometimes people write “ $p(D|\theta)$ ” but we don't need  $p(a|\theta)$  or  $p(x|\theta)$  here
  - ▶ warning 2: don't need historical record of  $\theta_t$ .
  - ▶ (we used Bayes rule, but only in  $\theta$  and  $y$ .)

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$
  - ▶  $= p(\theta|\alpha) \prod_t p(y_t|a_t, x_t, \theta)$
  - ▶ warning 1: sometimes people write “ $p(D|\theta)$ ” but we don't need  $p(a|\theta)$  or  $p(x|\theta)$  here
  - ▶ warning 2: don't need historical record of  $\theta_t$ .
    - ▶ (we used Bayes rule, but only in  $\theta$  and  $y$ .)
- ▶ A2: evaluate integral by  $N = 1$  Monte Carlo

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$
  - ▶  $= p(\theta|\alpha) \prod_t p(y_t | a_t, x_t, \theta)$
  - ▶ warning 1: sometimes people write “ $p(D|\theta)$ ” but we don't need  $p(a|\theta)$  or  $p(x|\theta)$  here
  - ▶ warning 2: don't need historical record of  $\theta_t$ .
    - ▶ (we used Bayes rule, but only in  $\theta$  and  $y$ .)
- ▶ A2: evaluate integral by  $N = 1$  Monte Carlo
  - ▶ take 1 sample “ $\theta_t$ ” of  $\theta$  from  $p(\theta|D)$

## TS: connecting w/“generative causal modeling” 2

- ▶ Q1: what's  $p(\theta|D)$ ?
- ▶ Q2: how am I going to evaluate this integral?
- ▶ A1:  $p(\theta|D)$  definable by choosing prior  $p(\theta|\alpha)$  and likelihood on  $y$  given by the (modeled, parameterized) response  $p(y|a, x, \theta)$ .
  - ▶ (now you're not only generative, you're Bayesian.)
  - ▶  $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - ▶  $\propto p(\{y\}_1^t | \{a\}_1^t, \{x\}_1^t, \theta) p(\theta|\alpha)$
  - ▶  $= p(\theta|\alpha) \prod_t p(y_t | a_t, x_t, \theta)$
  - ▶ warning 1: sometimes people write “ $p(D|\theta)$ ” but we don't need  $p(a|\theta)$  or  $p(x|\theta)$  here
  - ▶ warning 2: don't need historical record of  $\theta_t$ .
    - ▶ (we used Bayes rule, but only in  $\theta$  and  $y$ .)
- ▶ A2: evaluate integral by  $N = 1$  Monte Carlo
  - ▶ take 1 sample “ $\theta_t$ ” of  $\theta$  from  $p(\theta|D)$
  - ▶  $a_t = h(x_t; \theta_t) = \operatorname{argmax}_a Q(a, x, \theta_t)$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$
  - ▶  $F_a \equiv$  number of failures flipping coin  $a$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$
  - ▶  $F_a \equiv$  number of failures flipping coin  $a$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$
  - ▶  $F_a \equiv$  number of failures flipping coin  $a$

Then

- ▶  $p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$
  - ▶  $F_a \equiv$  number of failures flipping coin  $a$

Then

- ▶  $p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)$
- ▶  $= \left( \prod_a \theta_a^{\alpha-1} (1 - \theta_a)^{\beta-1} \right) \left( \prod_{t,a_t} \theta_{a_t}^{y_t} (1 - \theta_{a_t})^{1-y_t} \right)$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$
  - ▶  $F_a \equiv$  number of failures flipping coin  $a$

Then

- ▶  $p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)$
- ▶  $= \left( \prod_a \theta_a^{\alpha-1} (1 - \theta_a)^{\beta-1} \right) \left( \prod_{t,a_t} \theta_{a_t}^{y_t} (1 - \theta_{a_t})^{1-y_t} \right)$
- ▶  $= \prod_a \theta_a^{\alpha+S_a-1} (1 - \theta_a)^{\beta+F_a-1}$

That sounds hard.

No, just general. Let's do toy case:

- ▶  $y \in \{0, 1\}$ ,
- ▶ no context  $x$ ,
- ▶ Bernoulli (coin flipping), keep track of
  - ▶  $S_a \equiv$  number of successes flipping coin  $a$
  - ▶  $F_a \equiv$  number of failures flipping coin  $a$

Then

- ▶  $p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)$
- ▶  $= \left( \prod_a \theta_a^{\alpha-1} (1 - \theta_a)^{\beta-1} \right) \left( \prod_{t,a_t} \theta_{a_t}^{y_t} (1 - \theta_{a_t})^{1-y_t} \right)$
- ▶  $= \prod_a \theta_a^{\alpha+S_a-1} (1 - \theta_a)^{\beta+F_a-1}$
- ▶  $\therefore \theta_a \sim \text{Beta}(\alpha + S_a, \beta + F_a)$

# Thompson sampling: results (2011)

---

## An Empirical Evaluation of Thompson Sampling

---

**Olivier Chapelle**

Yahoo! Research

Santa Clara, CA

[chap@yahoo-inc.com](mailto:chap@yahoo-inc.com)

**Lihong Li**

Yahoo! Research

Santa Clara, CA

[lihong@yahoo-inc.com](mailto:lihong@yahoo-inc.com)

Figure 15: Chaleppe and Li 2011

## TS: words

In the realizable case, the reward is a stochastic function of the action, context and the unknown, true parameter  $\theta^*$ . Ideally, we would like to choose the action maximizing the expected reward,  $\max_a \mathbb{E}(r|a, x, \theta^*)$ .

Of course,  $\theta^*$  is unknown. If we are just interested in maximizing the immediate reward (exploitation), then one should choose the action that maximizes  $\mathbb{E}(r|a, x) = \int \mathbb{E}(r|a, x, \theta) P(\theta|D) d\theta$ .

But in an exploration / exploitation setting, the probability matching heuristic consists in randomly selecting an action  $a$  according to its probability of being optimal. That is, action  $a$  is chosen with probability

$$\int \mathbb{I} \left[ \mathbb{E}(r|a, x, \theta) = \max_{a'} \mathbb{E}(r|a', x, \theta) \right] P(\theta|D) d\theta,$$

where  $\mathbb{I}$  is the indicator function. Note that the integral does not have to be computed explicitly: it suffices to draw a random parameter  $\theta$  at each round as explained in Algorithm 1. Implementation of the algorithm is thus efficient and straightforward in most applications.

Figure 16: from Chaleppe and Li 2011

---

**Algorithm 1** Thompson sampling

---

$D = \emptyset$   
**for**  $t = 1, \dots, T$  **do**  
    Receive context  $x_t$   
    Draw  $\theta^t$  according to  $P(\theta|D)$   
    Select  $a_t = \arg \max_a \mathbb{E}_r(r|x_t, a, \theta^t)$   
    Observe reward  $r_t$   
     $D = D \cup (x_t, a_t, r_t)$   
**end for**

---

Figure 17: from Chaleppe and Li 2011

## TS: Bernoulli bandit p-code<sup>22</sup>

---

### Algorithm 2 Thompson sampling for the Bernoulli bandit

---

**Require:**  $\alpha, \beta$  prior parameters of a Beta distribution

$S_i = 0, F_i = 0, \forall i.$  {Success and failure counters}

**for**  $t = 1, \dots, T$  **do**

**for**  $i = 1, \dots, K$  **do**

        Draw  $\theta_i$  according to  $\text{Beta}(S_i + \alpha, F_i + \beta).$

**end for**

    Draw arm  $\hat{i} = \arg \max_i \theta_i$  and observe reward  $r$

**if**  $r = 1$  **then**

$S_{\hat{i}} = S_{\hat{i}} + 1$

**else**

$F_{\hat{i}} = F_{\hat{i}} + 1$

**end if**

**end for**

---

## TS: Bernoulli bandit p-code (results)

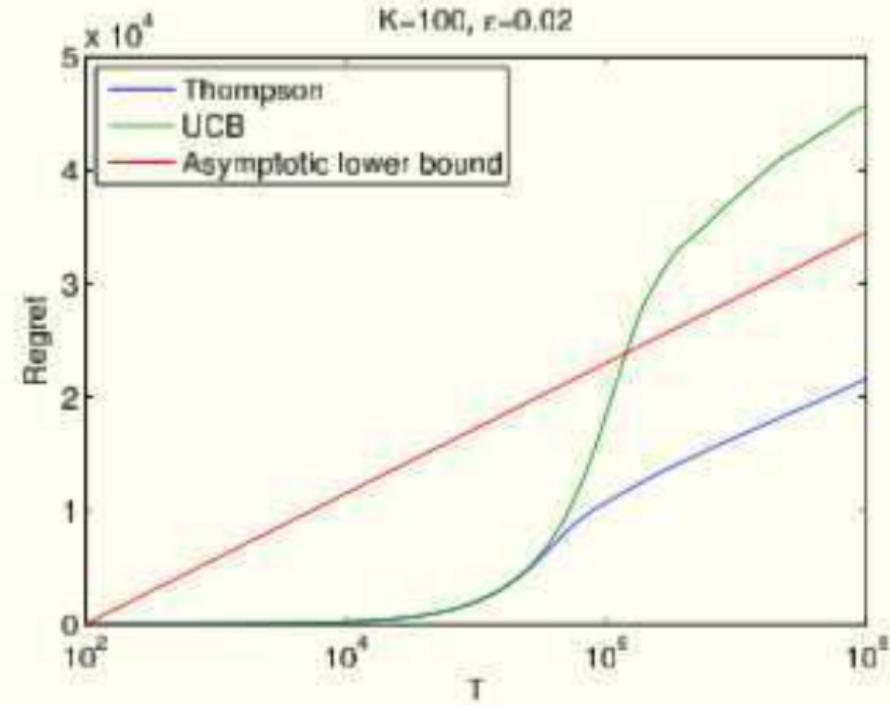


Figure 19: from Chaleppe and Li 2011

## UCB1 (2002), p-code

**Deterministic policy: ucb1.**

**Initialization:** Play each machine once.

**Loop:**

- Play machine  $j$  that maximizes  $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$ , where  $\bar{x}_j$  is the average reward obtained from machine  $j$ ,  $n_j$  is the number of times machine  $j$  has been played so far, and  $n$  is the overall number of plays done so far.

Figure 20: UCB1

from Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer.  
“Finite-time analysis of the multiarmed bandit problem.” Machine learning 47.2-3 (2002): 235-256.

## TS: with context

---

**Algorithm 3** Regularized logistic regression with batch updates

---

**Require:** Regularization parameter  $\lambda > 0$ .

$m_i = 0, q_i = \lambda$ . {Each weight  $w_i$  has an independent prior  $\mathcal{N}(m_i, q_i^{-1})$ }

**for**  $t = 1, \dots, T$  **do**

    Get a new batch of training data  $(\mathbf{x}_j, y_j), j = 1, \dots, n$ .

    Find  $\mathbf{w}$  as the minimizer of:  $\frac{1}{2} \sum_{i=1}^d q_i (w_i - m_i)^2 + \sum_{j=1}^n \log(1 + \exp(-y_j \mathbf{w}^\top \mathbf{x}_j))$ .

$m_i = w_i$

$q_i = q_i + \sum_{j=1}^n x_{ij}^2 p_j (1 - p_j), p_j = (1 + \exp(-\mathbf{w}^\top \mathbf{x}_j))^{-1}$  {Laplace approximation}

**end for**

---

Figure 21: from Chaleppe and Li 2011

## LinUCB: UCB with context

---

**Algorithm 1** LinUCB with disjoint linear models.

---

```
0: Inputs:  $\alpha \in \mathbb{R}_+$ 
1: for  $t = 1, 2, 3, \dots, T$  do
2:   Observe features of all arms  $a \in \mathcal{A}_t$ :  $\mathbf{x}_{t,a} \in \mathbb{R}^d$ 
3:   for all  $a \in \mathcal{A}_t$  do
4:     if  $a$  is new then
5:        $\mathbf{A}_a \leftarrow \mathbf{I}_d$  ( $d$ -dimensional identity matrix)
6:        $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$  ( $d$ -dimensional zero vector)
7:     end if
8:      $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$ 
9:      $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$ 
10:   end for
11:   Choose arm  $a_t = \arg \max_{a \in \mathcal{A}_t} p_{t,a}$  with ties broken arbitrarily, and observe a real-valued payoff  $r_t$ 
12:    $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$ 
13:    $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$ 
14: end for
```

---

Figure 22: LinUCB

## TS: with context (results)

Table 2: CTR regrets on the display advertising data.

Method	TS	LinUCB			$\epsilon$ -greedy			Exploit	Randor
Parameter	0.25	0.5	1	0.5	1	2	0.005	0.01	0.02
Regret (%)	4.45	3.72	3.81	4.99	4.22	4.14	5.05	4.98	5.22

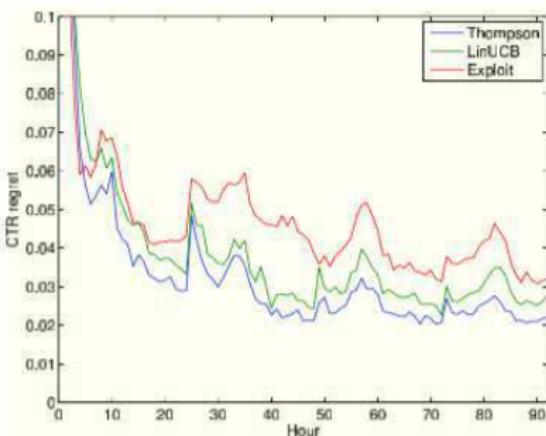


Figure 4: CTR regret over the 4 days test period for 3 algorithms: Thompson sampling with  $\alpha = 0.5$ , LinUCB with  $\alpha = 2$ , Exploit-only. The regret in the first hour is large, around 0.3, because the algorithms predict randomly (no initial model provided).

## Bandits: Regret via Lai and Robbins (1985)

**THEOREM 2.** Assume that  $I(\theta, \lambda)$  satisfies (1.6) and (1.7) and that  $\Theta$  satisfies (1.9). Fix  $j \in \{1, \dots, k\}$ , and define  $\Theta_j$  and  $\Theta_j^*$  by (2.1). Let  $\varphi$  be any rule such that for every  $\theta \in \Theta_j^*$ , as  $n \rightarrow \infty$

$$\sum_{i \neq j} E_\theta T_n(i) = o(n^a) \quad \text{for every } a > 0, \quad (2.2)$$

where  $T_n(i)$ , defined in (1.2), is the number of times that the rule  $\varphi$  samples from  $\Pi_i$  up to stage  $n$ . Then for every  $\theta \in \Theta_j$  and every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P_\theta \left\{ T_n(j) \geq (1 - \epsilon)(\log n)/I(\theta_j, \theta^*) \right\} = 1, \quad (2.3)$$

where  $\theta^*$  is defined in (1.4), and hence

$$\liminf_{n \rightarrow \infty} E_\theta T_n(j)/\log n \geq 1/I(\theta_j, \theta^*).$$

Figure 24: Lai Robbins

# Thompson sampling (1933) and optimality (2013)

**Theorem 2.** For any instance  $\Theta = \{\mu_1, \dots, \mu_N\}$  of Bernoulli MAB,

$$R(T, \Theta) \leq (1 + \epsilon) \sum_{i \neq I^*} \frac{\ln(T) \Delta_i}{KL(\mu_i, \mu^*)} + O(N/\epsilon^2)$$

Recall that we have  $\lim_{T \rightarrow \infty} \frac{R(T, \Theta)}{\ln(T)} \geq \sum_{i \neq I^*} \frac{\Delta_i}{KL(\mu_i, \mu^*)}$ . Above theorem says that Thompson Sampling matches this lower bound. We also have the following problem independent regret bound for this algorithm.

**Theorem 3.** For all  $\Theta$ ,

$$R(T) = \max_{\Theta} R(T, \Theta) \leq O(\sqrt{NT \log T} + N)$$

For proofs of above theorems, refer to [2].

Figure 25: TS result

from S. Agrawal, N. Goyal, "Further optimal regret bounds for Thompson Sampling", AISTATS 2013.; see also Agrawal, Shipra, and Navin Goyal. "Analysis of Thompson Sampling for the Multi-armed Bandit Problem." COLT. 2012 and Emilie Kaufmann, Nathaniel Korda, and R'emi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In Algorithmic Learning Theory, pages 199–213. Springer, 2012.

## other 'Causalities': structure learning

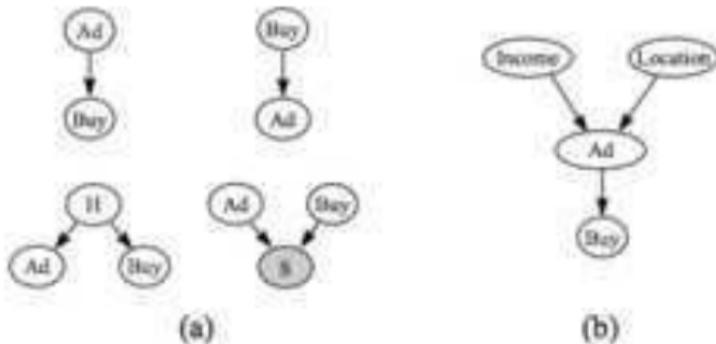


Figure 9: (a) Causal graphs showing for explanations for an observed dependence between *Ad* and *Buy*. The node *H* corresponds to a hidden common cause of *Ad* and *Buy*. The shaded node *S* indicates that the case has been included in the database. (b) A Bayesian network for which *A* causes *B* is the only causal explanation, given the causal Markov condition.

Figure 26: from heckerman 1995

D. Heckerman. A Tutorial on Learning with Bayesian Networks.  
Technical Report MSR-TR-95-06, Microsoft Research, March, 1995.

## other ‘Causalities’: potential outcomes

- ▶ model distribution of  $p(y_i(1), y_i(0), a_i, x_i)$

---

## other ‘Causalities’: potential outcomes

- ▶ model distribution of  $p(y_i(1), y_i(0), a_i, x_i)$
- ▶ “action” replaced by “observed outcome”

---

## other ‘Causalities’: potential outcomes

- ▶ model distribution of  $p(y_i(1), y_i(0), a_i, x_i)$
- ▶ “action” replaced by “observed outcome”
- ▶ aka Neyman-Rubin causal model: Neyman ('23); Rubin ('74)

---

## other ‘Causalities’: potential outcomes

- ▶ model distribution of  $p(y_i(1), y_i(0), a_i, x_i)$
- ▶ “action” replaced by “observed outcome”
- ▶ aka Neyman-Rubin causal model: Neyman ('23); Rubin ('74)
- ▶ see Morgan + Winship<sup>23</sup> for connections between frameworks

---

<sup>23</sup>Morgan, Stephen L., and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

## Lecture 4: descriptive modeling @ NYT

## review: (latent) inference and clustering

- ▶ what does kmeans mean?

## review: (latent) inference and clustering

- ▶ what does kmeans mean?
  - ▶ given  $x_i \in R^D$

## review: (latent) inference and clustering

- ▶ what does kmeans mean?

- ▶ given  $x_i \in R^D$
- ▶ given  $d : R^D \rightarrow R^1$

## review: (latent) inference and clustering

- ▶ what does kmeans mean?

- ▶ given  $x_i \in R^D$
- ▶ given  $d : R^D \rightarrow R^1$
- ▶ assign  $z_i$

## review: (latent) inference and clustering

- ▶ what does kmeans mean?
  - ▶ given  $x_i \in R^D$
  - ▶ given  $d : R^D \rightarrow R^1$
  - ▶ assign  $z_i$
- ▶ generative modeling gives meaning

## review: (latent) inference and clustering

- ▶ what does kmeans mean?
  - ▶ given  $x_i \in R^D$
  - ▶ given  $d : R^D \rightarrow R^1$
  - ▶ assign  $z_i$
- ▶ generative modeling gives meaning
  - ▶ given  $p(x|z, \theta)$

## review: (latent) inference and clustering

- ▶ what does kmeans mean?
  - ▶ given  $x_i \in R^D$
  - ▶ given  $d : R^D \rightarrow R^1$
  - ▶ assign  $z_i$
- ▶ generative modeling gives meaning
  - ▶ given  $p(x|z, \theta)$
  - ▶ maximize  $p(x|\theta)$

## review: (latent) inference and clustering

- ▶ what does kmeans mean?
  - ▶ given  $x_i \in R^D$
  - ▶ given  $d : R^D \rightarrow R^1$
  - ▶ assign  $z_i$
- ▶ generative modeling gives meaning
  - ▶ given  $p(x|z, \theta)$
  - ▶ maximize  $p(x|\theta)$
  - ▶ output assignment  $p(z|x, \theta)$

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$   
(cf. importance sampling)

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$
  - ▶ analogy to free energy in physics

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$
  - ▶ analogy to free energy in physics
- ▶ alternate optimization on  $\theta$  and on  $q$

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$
  - ▶ analogy to free energy in physics
- ▶ alternate optimization on  $\theta$  and on  $q$ 
  - ▶ NB:  $q$  step gives  $q(z) = p(z|x, \theta)$

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$
  - ▶ analogy to free energy in physics
- ▶ alternate optimization on  $\theta$  and on  $q$ 
  - ▶ NB:  $q$  step gives  $q(z) = p(z|x, \theta)$
  - ▶ NB:  $\log P$  convenient for independent examples w/ exponential families

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$
  - ▶ analogy to free energy in physics
- ▶ alternate optimization on  $\theta$  and on  $q$ 
  - ▶ NB:  $q$  step gives  $q(z) = p(z|x, \theta)$
  - ▶ NB:  $\log P$  convenient for independent examples w/ exponential families
  - ▶ e.g., GMMs:  $\mu_k \leftarrow E[x|z]$  and  $\sigma_k^2 \leftarrow E[(x - \mu)^2|z]$  are sufficient statistics

## actual math

- ▶ define  $P \equiv p(x, z|\theta)$
- ▶ log-likelihood  $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q$   
(cf. importance sampling)
- ▶ Jensen's:  
$$L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$
  - ▶ analogy to free energy in physics
- ▶ alternate optimization on  $\theta$  and on  $q$ 
  - ▶ NB:  $q$  step gives  $q(z) = p(z|x, \theta)$
  - ▶ NB:  $\log P$  convenient for independent examples w/ exponential families
  - ▶ e.g., GMMs:  $\mu_k \leftarrow E[x|z]$  and  $\sigma_k^2 \leftarrow E[(x - \mu)^2|z]$  are sufficient statistics
  - ▶ e.g., LDAs: word counts are sufficient statistics

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

$$\blacktriangleright -U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$$

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- ▶  $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- ▶  $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- ▶ define  $r_{ik} = \sum_z q_i(z) 1[z_i = k]$

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- ▶  $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- ▶ define  $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
- ▶  $-U_x = \sum_i r_{ik} \log p(x_i|k)$ .

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- ▶  $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- ▶ define  $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
- ▶  $-U_x = \sum_i r_{ik} \log p(x_i|k).$
- ▶ Gaussian<sup>24</sup>  
 $\Rightarrow -U_x = \sum_i r_{ik} \left( -\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

---

<sup>24</sup>math is simpler if you work with  $\lambda_k \equiv \sigma^{-2}$

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- ▶  $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- ▶ define  $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
- ▶  $-U_x = \sum_i r_{ik} \log p(x_i|k).$
- ▶ Gaussian<sup>24</sup>  
 $\Rightarrow -U_x = \sum_i r_{ik} \left( -\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

---

<sup>24</sup>math is simpler if you work with  $\lambda_k \equiv \sigma^{-2}$

## tangent: more math on GMMs, part 1

Energy  $U$  (to be minimized):

- ▶  $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- ▶  $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- ▶  $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- ▶ define  $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
- ▶  $-U_x = \sum_i r_{ik} \log p(x_i|k).$
- ▶ Gaussian<sup>24</sup>  
 $\Rightarrow -U_x = \sum_i r_{ik} \left( -\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

simple to minimize for parameters  $\vartheta = \{\mu_k, \lambda_k\}$

---

<sup>24</sup>math is simpler if you work with  $\lambda_k \equiv \sigma^{-2}$

## tangent: more math on GMMs, part 2

- $-U_x = \sum_i r_{ik} \left( -\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

## tangent: more math on GMMs, part 2

- ▶  $-U_x = \sum_i r_{ik} \left( -\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$
- ▶  $\mu_k \leftarrow E[x|k]$  solves  $\sum_i r_{ik} = \sum_i r_{ik} x_i$

## tangent: more math on GMMs, part 2

- ▶  $-U_x = \sum_i r_{ik} \left( -\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$
- ▶  $\mu_k \leftarrow E[x|k]$  solves  $\sum_i r_{ik} = \sum_i r_{ik} x_i$
- ▶  $\lambda_k \leftarrow E[(x - \mu)^2|k]$  solves  $\sum_i r_{ik} \frac{1}{2}(x_i - \mu_k)^2 = \lambda_k^{-1} \sum_i r_{ik}$

tangent: Gaussians  $\in$  exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$

---

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$

---

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$
  - ▶  $T_2 = x^2$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$
  - ▶  $T_2 = x^2$
  - ▶  $\eta_1 = \mu/\sigma^2 = \mu\lambda$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$
  - ▶  $T_2 = x^2$
  - ▶  $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  - ▶  $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$
  - ▶  $T_2 = x^2$
  - ▶  $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  - ▶  $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$
  - ▶  $A = \lambda\mu^2/2 - \frac{1}{2}\ln\lambda$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$
  - ▶  $T_2 = x^2$
  - ▶  $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  - ▶  $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$
  - ▶  $A = \lambda\mu^2/2 - \frac{1}{2}\ln\lambda$
  - ▶  $\exp(B(x)) = (2\pi)^{-1/2}$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

## tangent: Gaussians $\in$ exponential family<sup>26</sup>

- ▶ as before,  $-U = \sum_i r_{ik} \log p(x_i|k)$
- ▶ define  $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- ▶ e.g., Gaussian case<sup>25</sup>,
  - ▶  $T_1 = x,$
  - ▶  $T_2 = x^2$
  - ▶  $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  - ▶  $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$
  - ▶  $A = \lambda\mu^2/2 - \frac{1}{2}\ln\lambda$
  - ▶  $\exp(B(x)) = (2\pi)^{-1/2}$
- ▶ note that in a mixture model, there are separate  $\eta$  (and thus  $A(\eta)$ ) for each value of  $z$

---

<sup>25</sup>Choosing  $\eta(\theta) = \eta$  called ‘canonical form’

<sup>26</sup>NB: Gaussians  $\in$  exponential family, GMM  $\notin$  exponential family! (Thanks to Eszter Vértes for pointing out this error in earlier title.)

tangent: variational joy  $\in$  exponential family

- ▶ as before,  $-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$

---

## tangent: variational joy $\in$ exponential family

- ▶ as before,  $-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$
- ▶  $\eta_{k,\alpha}$  solves  $\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}$  (canonical)

---

## tangent: variational joy $\in$ exponential family

- ▶ as before,  $-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$
- ▶  $\eta_{k,\alpha}$  solves  $\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}$  (canonical)
- ▶  $\therefore \partial_{\eta_{k,\alpha}} A(\eta_k) \leftarrow E[T_{k,\alpha}|k]$  (canonical)

---

## tangent: variational joy $\in$ exponential family

- ▶ as before,  $-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$
- ▶  $\eta_{k,\alpha}$  solves  $\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}$  (canonical)
- ▶  $\therefore \partial_{\eta_{k,\alpha}} A(\eta_k) \leftarrow E[T_{k,\alpha}|k]$  (canonical)
- ▶ nice connection w/physics, esp. mean field theory<sup>27</sup>

---

<sup>27</sup>read MacKay, David JC. *Information theory, inference and learning algorithms*, Cambridge university press, 2003 to learn more. Actually you should read it regardless.

## clustering and inference: GMM/k-means case study

- ▶ generative model gives meaning and optimization

## clustering and inference: GMM/k-means case study

- ▶ generative model gives meaning and optimization
- ▶ large freedom to choose different optimization approaches

## clustering and inference: GMM/k-means case study

- ▶ generative model gives meaning and optimization
- ▶ large freedom to choose different optimization approaches
  - ▶ e.g., hard clustering limit

## clustering and inference: GMM/k-means case study

- ▶ generative model gives meaning and optimization
- ▶ large freedom to choose different optimization approaches
  - ▶ e.g., hard clustering limit
  - ▶ e.g., streaming solutions

## clustering and inference: GMM/k-means case study

- ▶ generative model gives meaning and optimization
- ▶ large freedom to choose different optimization approaches
  - ▶ e.g., hard clustering limit
  - ▶ e.g., streaming solutions
  - ▶ e.g., stochastic gradient methods

## general framework: E+M/variational

- ▶ e.g., GMM+hard clustering gives kmeans

## general framework: E+M/variational

- ▶ e.g., GMM+hard clustering gives kmeans
- ▶ e.g., some favorite applications:

## general framework: E+M/variational

- ▶ e.g., GMM+hard clustering gives kmeans
- ▶ e.g., some favorite applications:
  - ▶ hmm

## general framework: E+M/variational

- ▶ e.g., GMM+hard clustering gives kmeans
- ▶ e.g., some favorite applications:
  - ▶ hmm
  - ▶ vbmod: arXiv:0709.3512

## general framework: E+M/variational

- ▶ e.g., GMM+hard clustering gives kmeans
- ▶ e.g., some favorite applications:
  - ▶ hmm
  - ▶ vbmod: arXiv:0709.3512
  - ▶ ebfret: ebfret.github.io

## general framework: E+M/variational

- ▶ e.g., GMM+hard clustering gives kmeans
- ▶ e.g., some favorite applications:
  - ▶ hmm
  - ▶ vbmod: arXiv:0709.3512
  - ▶ ebfret: ebfret.github.io
  - ▶ EDHMM: edhmm.github.io

## example application: LDA+topics

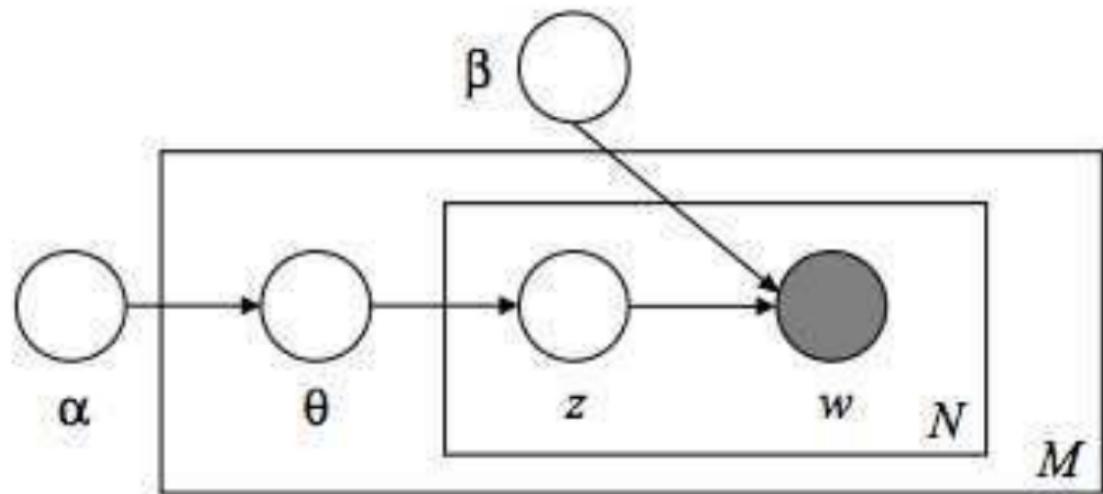


Figure 27: From Blei 2003

## rec engine via CTM<sup>28</sup>

item latent vector  $v \sim \mathcal{N}(\theta, \lambda_v^{-1} I_K)$

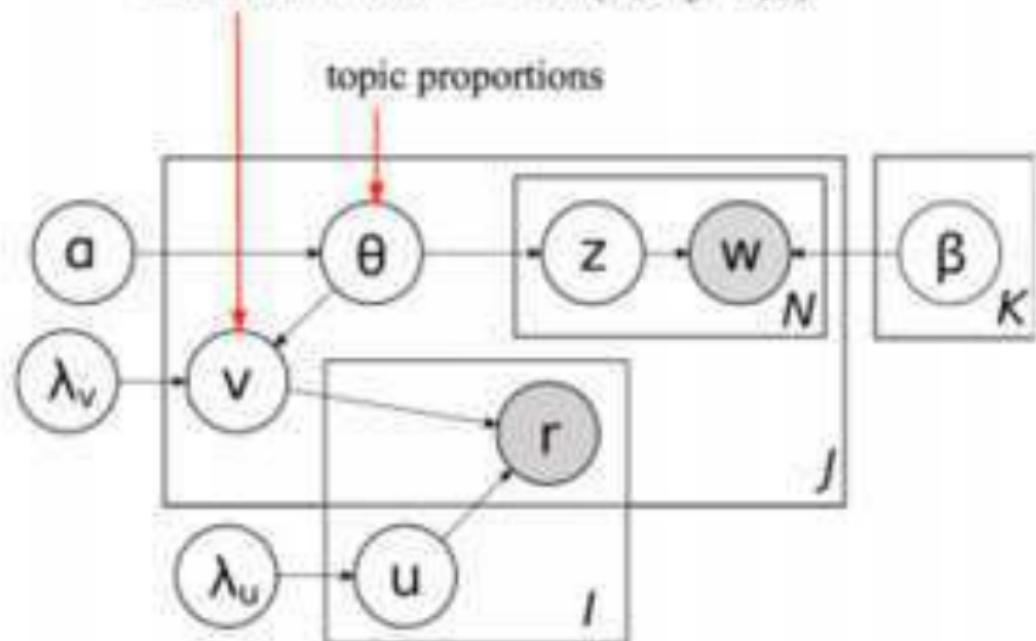


Figure 28: From Blei 2011

recall: recommendation via factoring

$$\min_{U,V} \sum_{i,j} (r_{ij} - u_i^T v_j)^2 + \lambda_u \|u_i\|^2 + \lambda_v \|v_j\|^2,$$

Figure 29: From Blei 2011

## CTM: combined loss function

Maximization of the posterior is equivalent to maximizing the complete log likelihood of  $U$ ,  $V$ ,  $\theta_{1:J}$ , and  $R$  given  $\lambda_u$ ,  $\lambda_v$  and  $\beta$ ,

$$\begin{aligned}\mathcal{L} = & -\frac{\lambda_u}{2} \sum_i u_i^T u_i - \frac{\lambda_v}{2} \sum_j (v_j - \theta_j)^T (v_j - \theta_j) \quad (7) \\ & + \sum_j \sum_n \log (\sum_k \theta_{jk} \beta_{k,w_{jn}}) - \sum_{i,j} \frac{c_{ij}}{2} (r_{ij} - u_i^T v_j)^2.\end{aligned}$$

Figure 30: From Blei 2011

## CTM: updates for factors

$$u_i \leftarrow (VC_iV^T + \lambda_u I_K)^{-1} VC_i R_i \quad (8)$$

$$v_j \leftarrow (UC_jU^T + \lambda_v I_K)^{-1} (UC_j R_j + \lambda_v \theta_j). \quad (9)$$

Figure 31: From Blei 2011

## CTM: (via Jensen's, again) bound on loss

$$\begin{aligned}\mathcal{L}(\theta_j) &\geq -\frac{\lambda_v}{2}(v_j - \theta_j)^T(v_j - \theta_j) \\ &+ \sum_n \sum_k \phi_{jnk} (\log \theta_{jk} \beta_{k,w_{jn}} - \log \phi_{jnk}) \\ &= \mathcal{L}(\theta_j, \phi_j).\end{aligned}\tag{10}$$

Figure 32: From Blei 2011

## Lecture 5 data product

# data science and design thinking

- ▶ knowing customer

## data science and design thinking

- ▶ knowing customer
- ▶ right tool for right job

## data science and design thinking

- ▶ knowing customer
- ▶ right tool for right job
- ▶ practical matters:

# data science and design thinking

- ▶ knowing customer
- ▶ right tool for right job
- ▶ practical matters:
  - ▶ munging

# data science and design thinking

- ▶ knowing customer
- ▶ right tool for right job
- ▶ practical matters:
  - ▶ munging
  - ▶ data ops

# data science and design thinking

- ▶ knowing customer
- ▶ right tool for right job
- ▶ practical matters:
  - ▶ munging
  - ▶ data ops
  - ▶ ML in prod

# Thanks!

Thanks MLSS students for your great questions; please contact me @chrishwiggins or chris.wiggins@{nytimes,gmail}.com with any questions, comments, or suggestions!