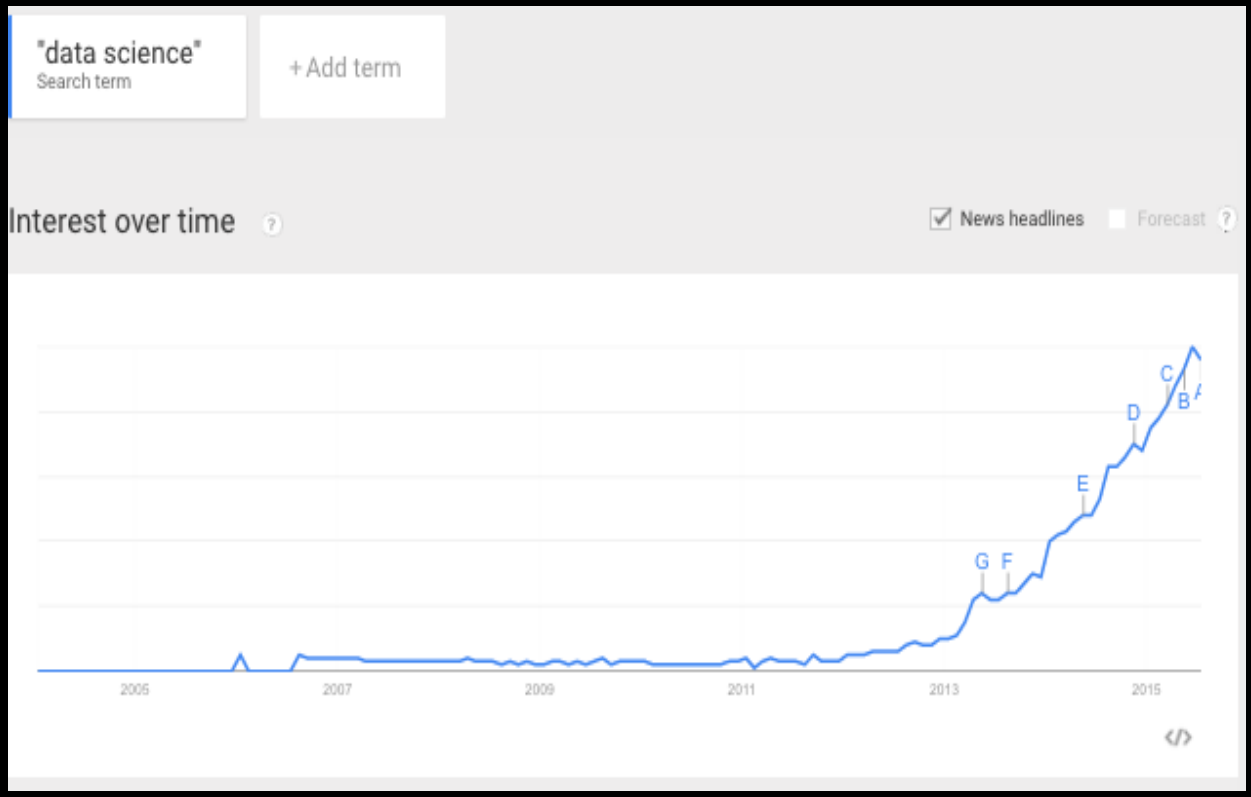
data science @ The New York Times

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

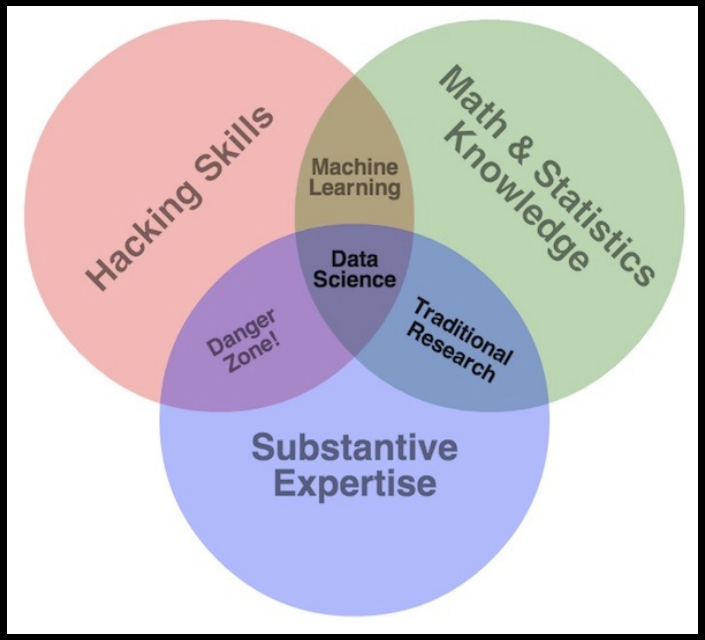
references: http://bit.ly/stanf16

data science

data science: searches



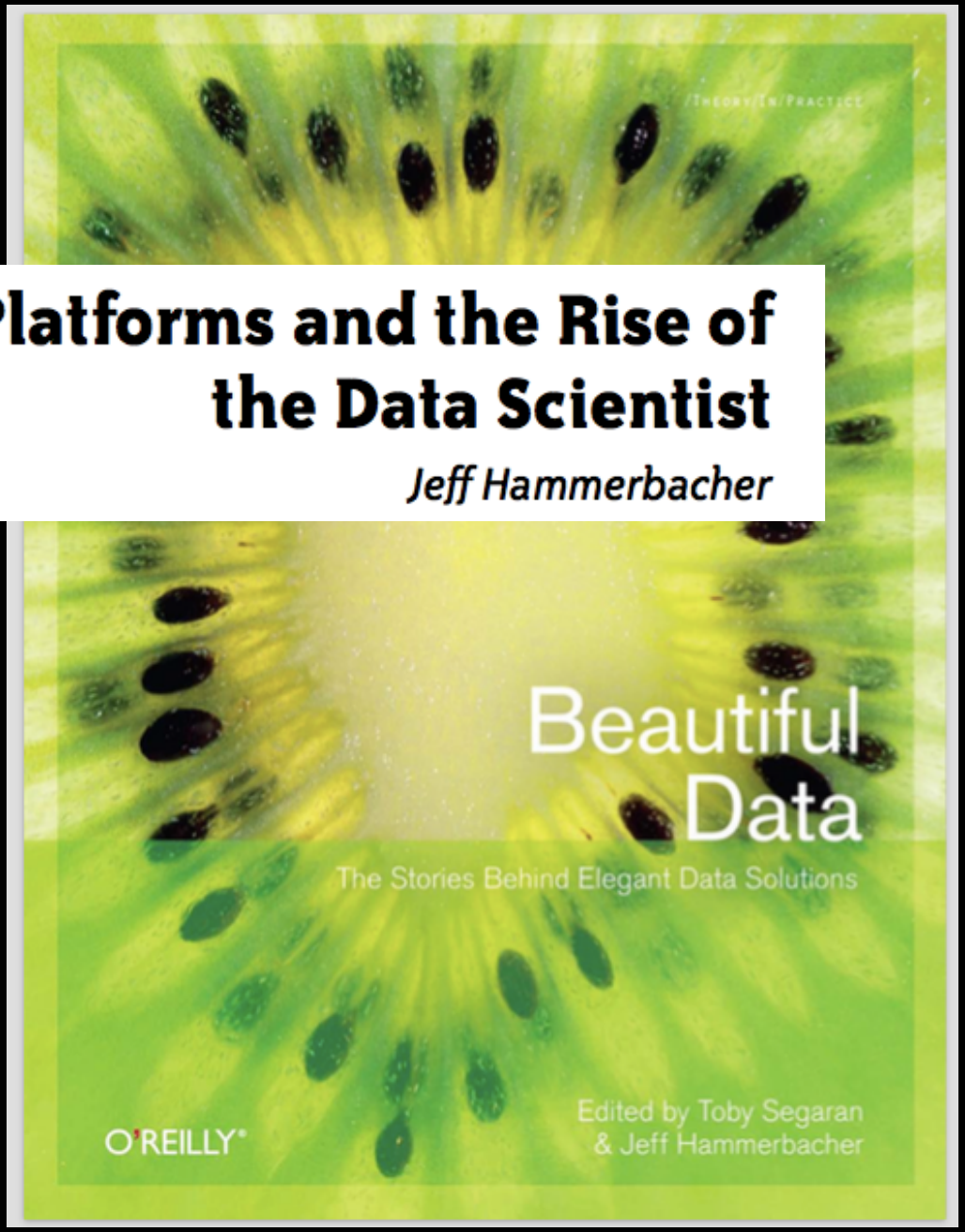
data science: mindset & toolset



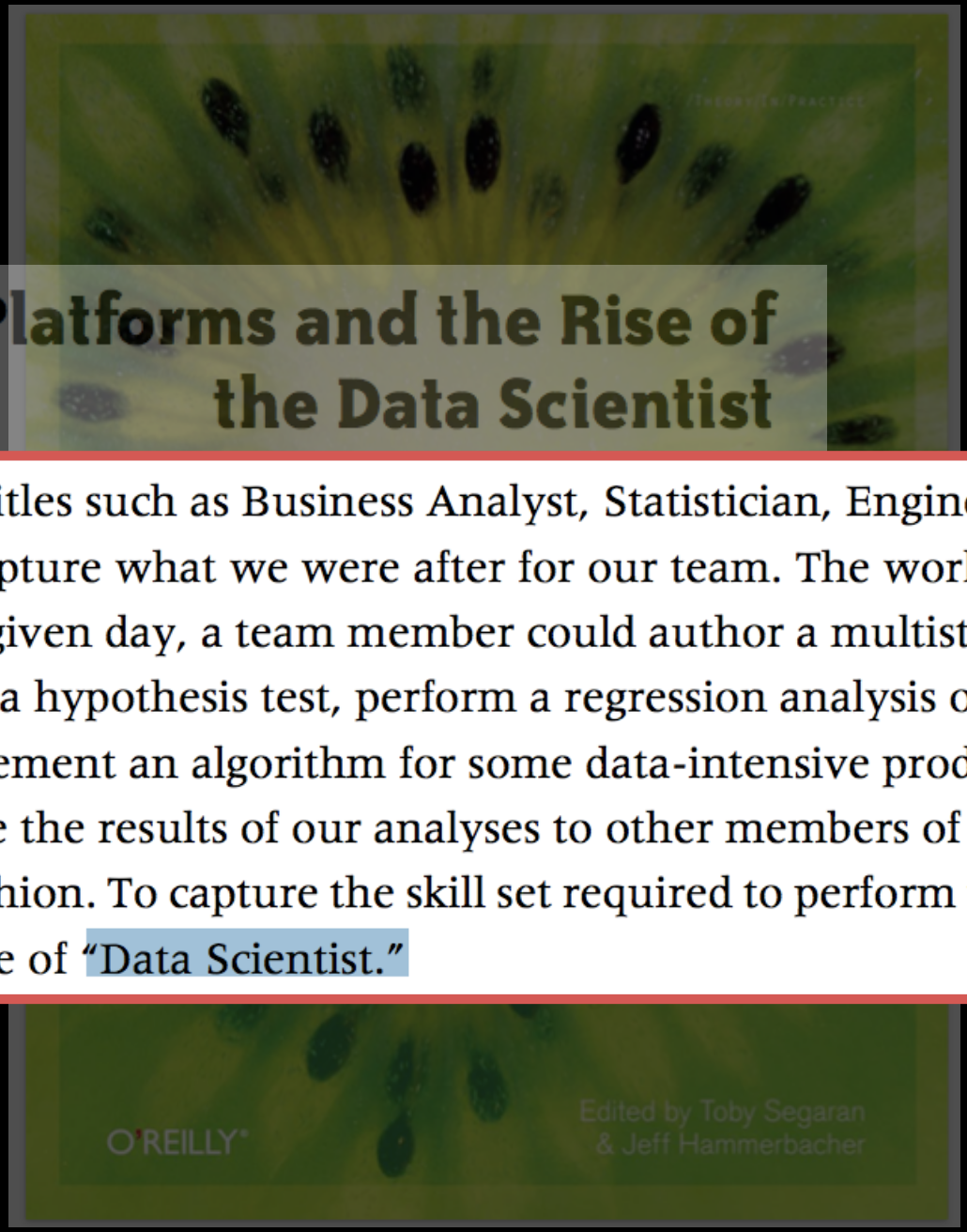
drew conway, 2010

Information Platforms and the Rise of the Data Scientist

Jeff Hammerbacher

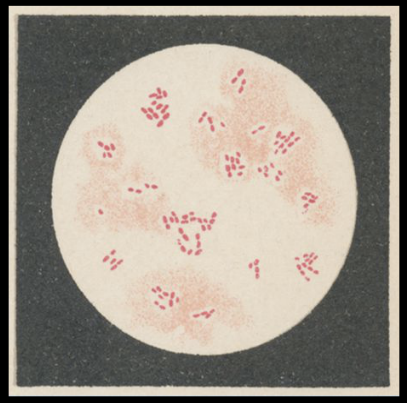


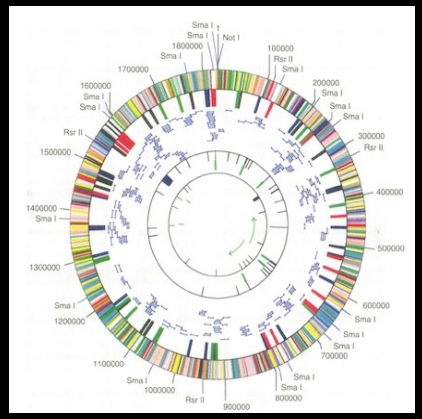
modern history:
2009



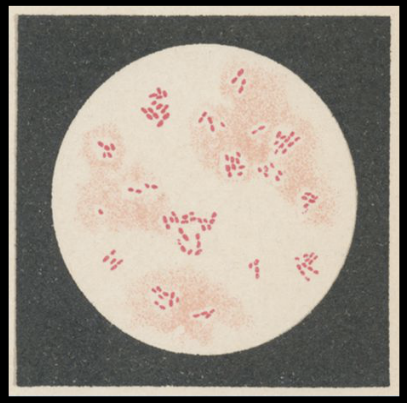
At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn't quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design 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."

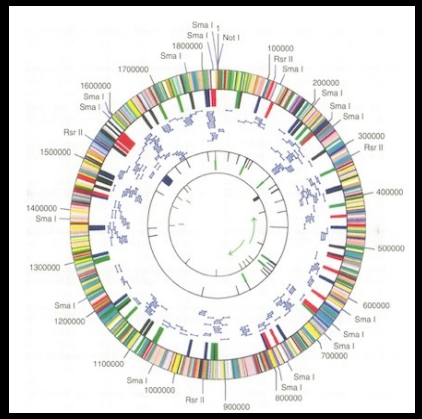
biology: 1892 vs. 1995





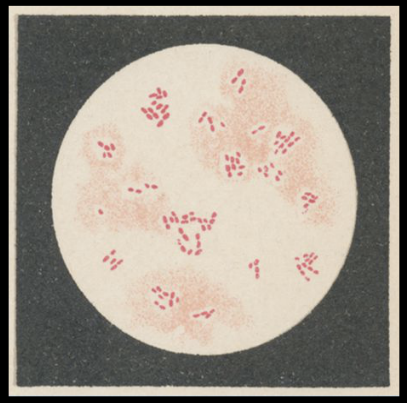
biology: 1892 vs. 1995

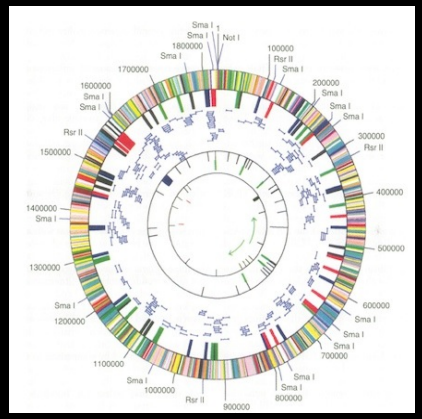




biology changed for good.

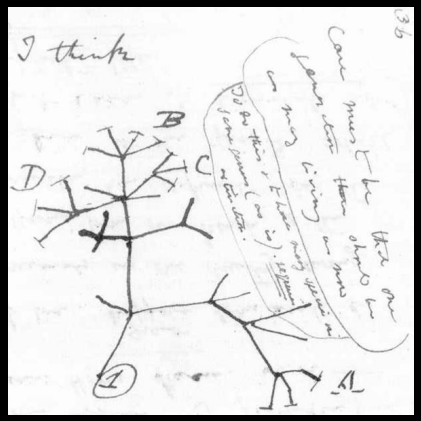
biology: 1892 vs. 1995

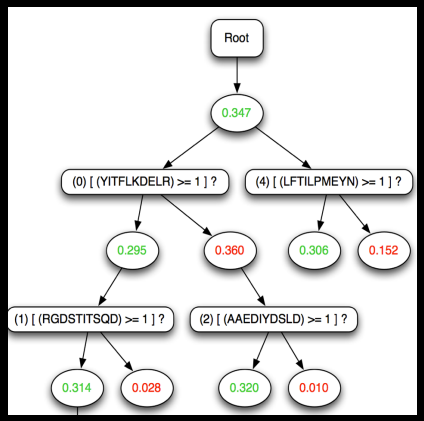




new toolset, new mindset

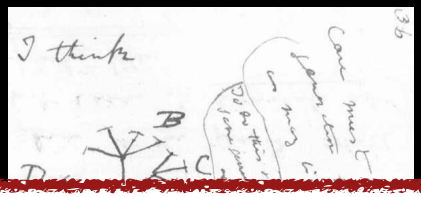
genetics: 1837 vs. 2012

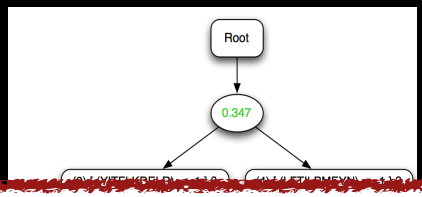




ML toolset; data science mindset

genetics: 1837 vs. 2012

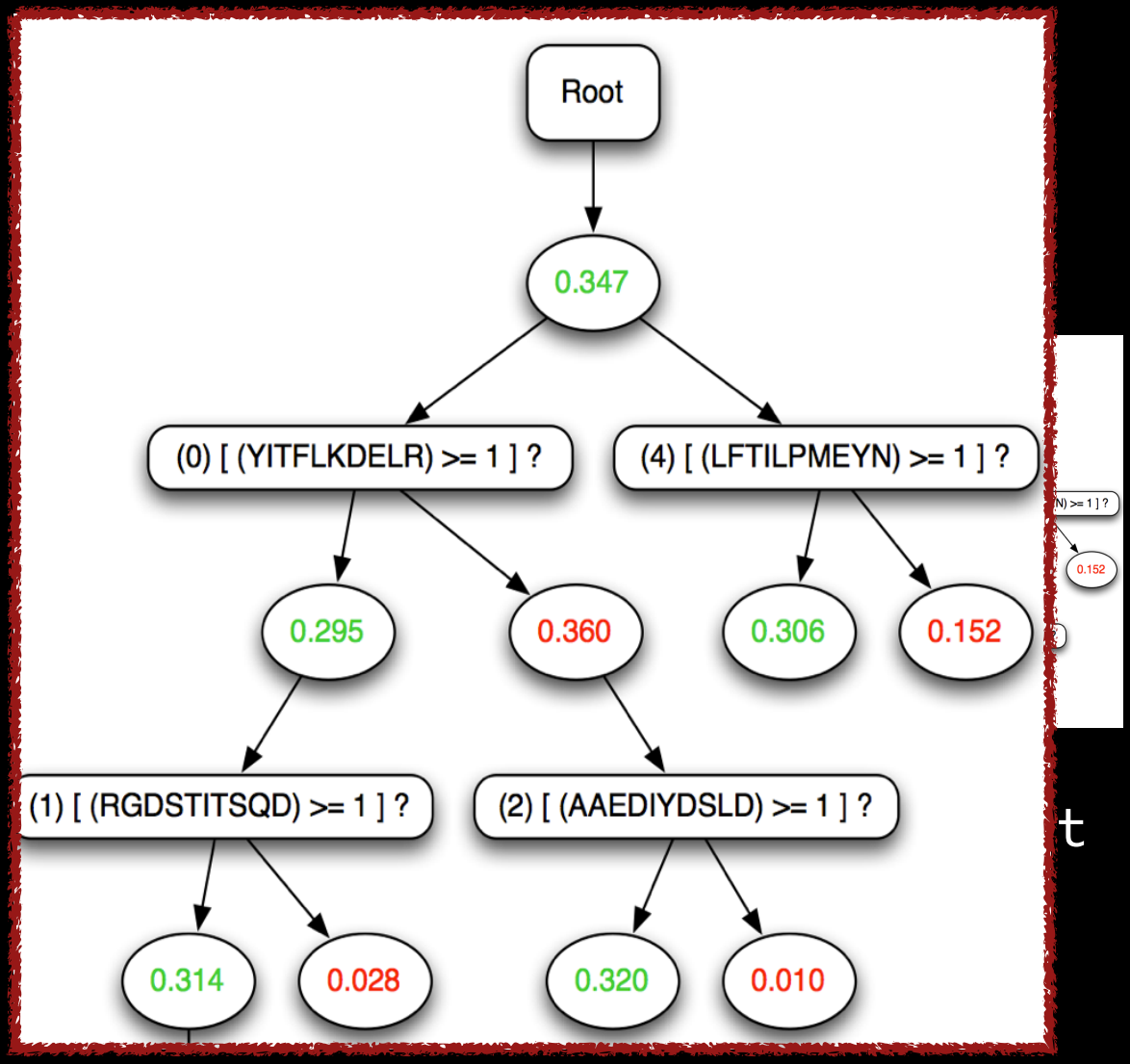




Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman



t

arxiv.org/abs/1105.5821 ; github.com/rajanil/mkboost

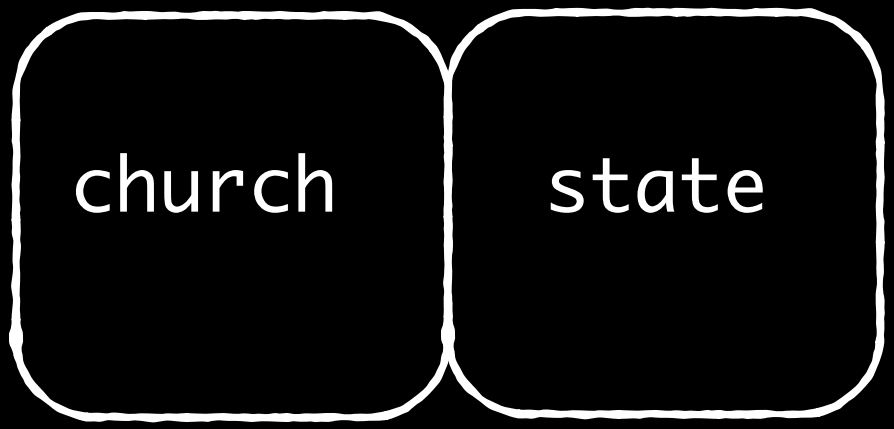
data science: mindset & toolset

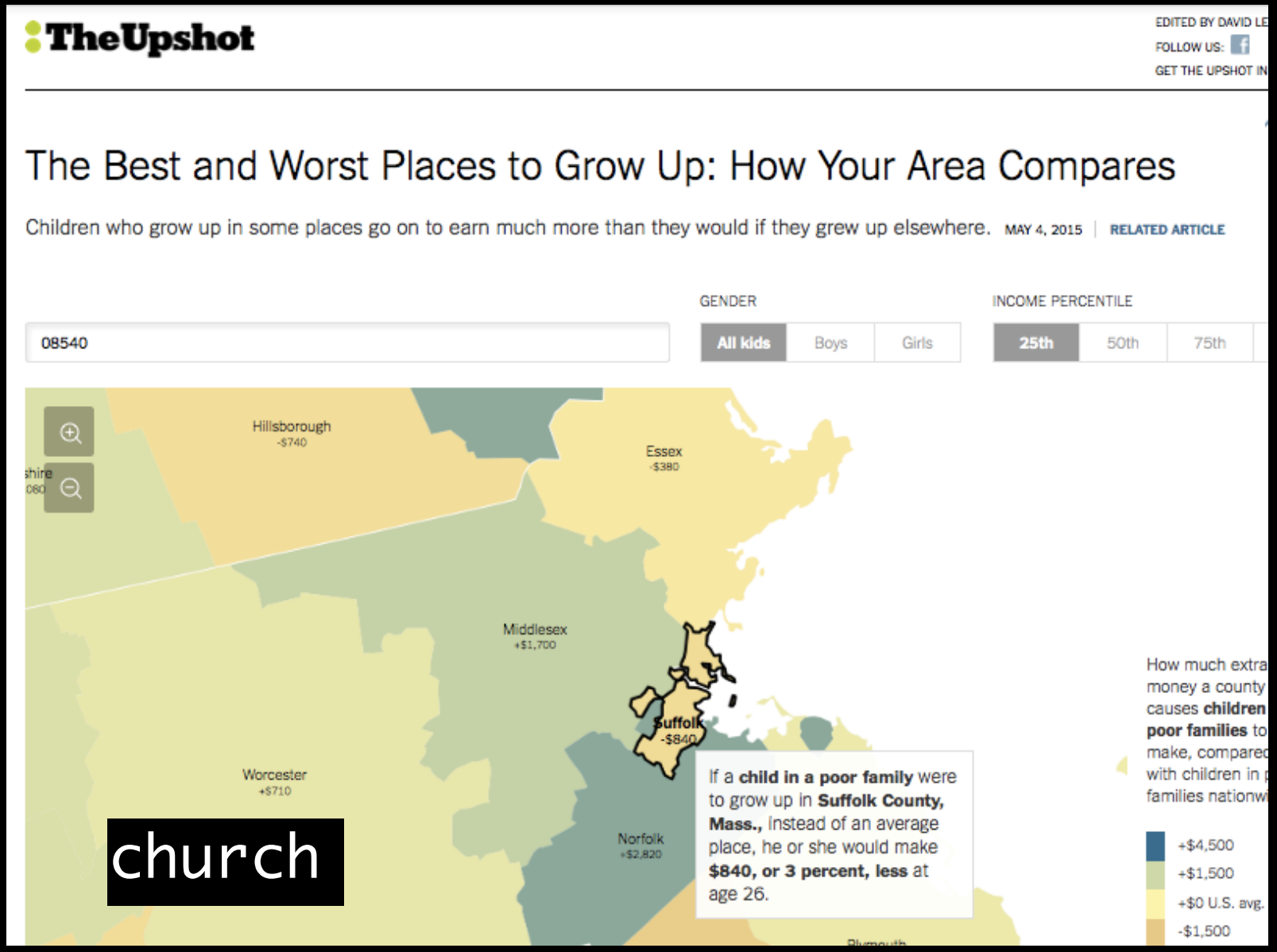
New-York Daily Times.

VOL. I....NO. 1. NEW-YORK, THURSDAY, SEPTEMBER 18, 1851. PRICE ONE CENT.

1851

news: 20th century




TheUpshot
EDITED BY DAVID LE
FOLLOW US:
GET THE UPSHOT IN
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
GENDER
INCOME PERCENTILE
08540
All kids
Boys
Girls
25th
50th
75th
Hillsborough
-$740
Essex
-$380
Middlesex
+$1,700
Suffolk
-$840
Worcester
+$710
Norfolk
+$2,820
If a child in a poor family were to grow up in Suffolk County, Mass., instead of an average place, he or she would make $840, or 3 percent, less at age 26.
How much extra
money a county
causes children
poor families to
make, compared
with children in
families nationwi
+$4,500
+$1,500
+$0 U.S. avg.
-$1,500
church


TheUpshot/leo-senate-mc
https://github.com/TheUpshot/leo-senate-model
GitHub
This repository
Search or type a command
og
PUBLIC
TheUpshot / leo-senate-model
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
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

church

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/

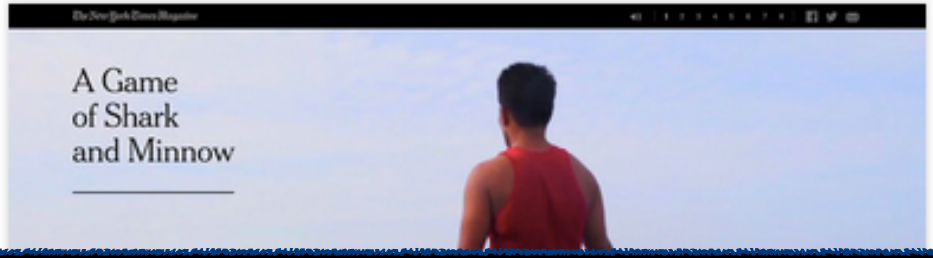
The New York Times

2013: The Year in Interactive Storytelling

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

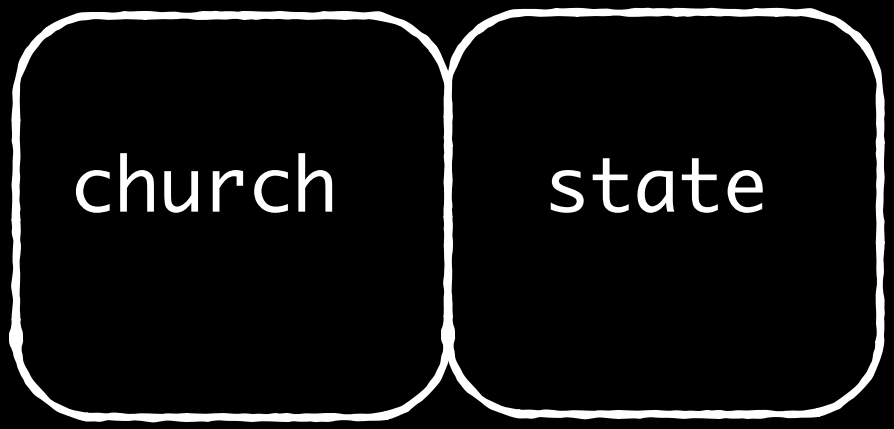
Multimedia Stories

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.

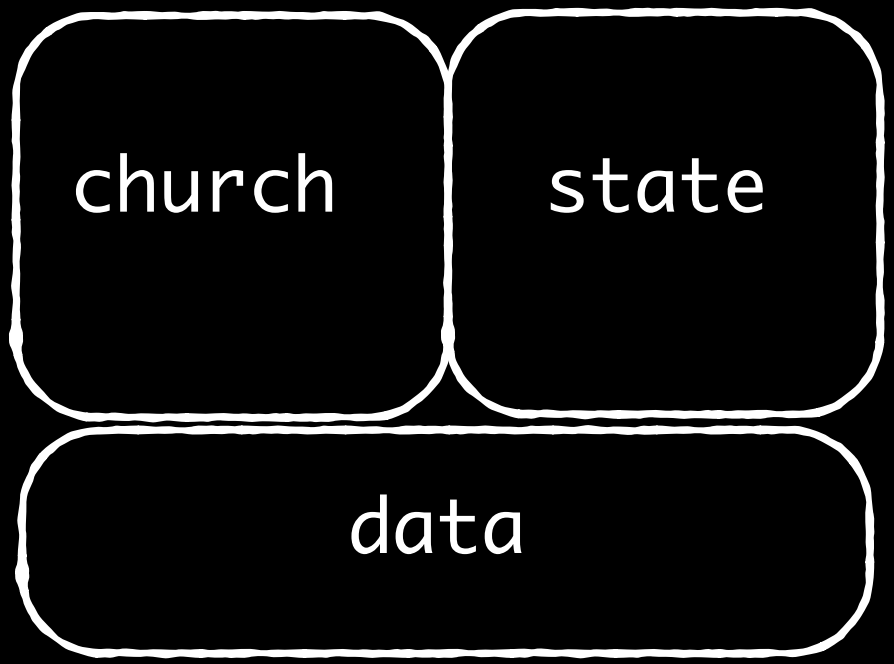


church

news: 20th century



news: 21st century



newspapering: 1851 vs. 1996

New-York Daily Times.

VOL. I., NO. 1. NEW-YORK, THURSDAY, SEPTEMBER 18, 1851. PRICE ONE CENT.

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

The New York Times

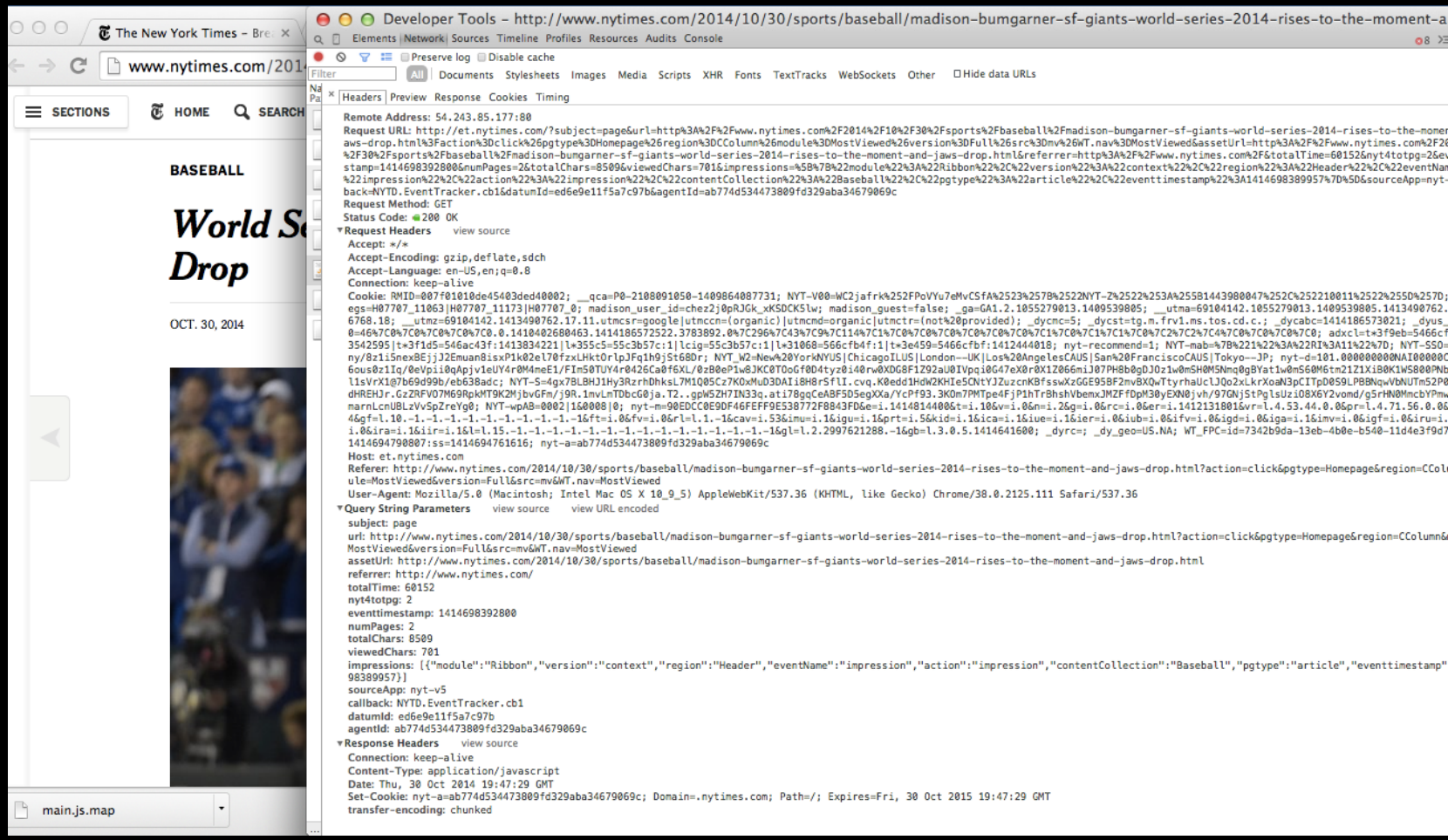
WHO WE ARE | WHAT WE DO | PRESS | CAREERS | INVESTORS | CONNECT

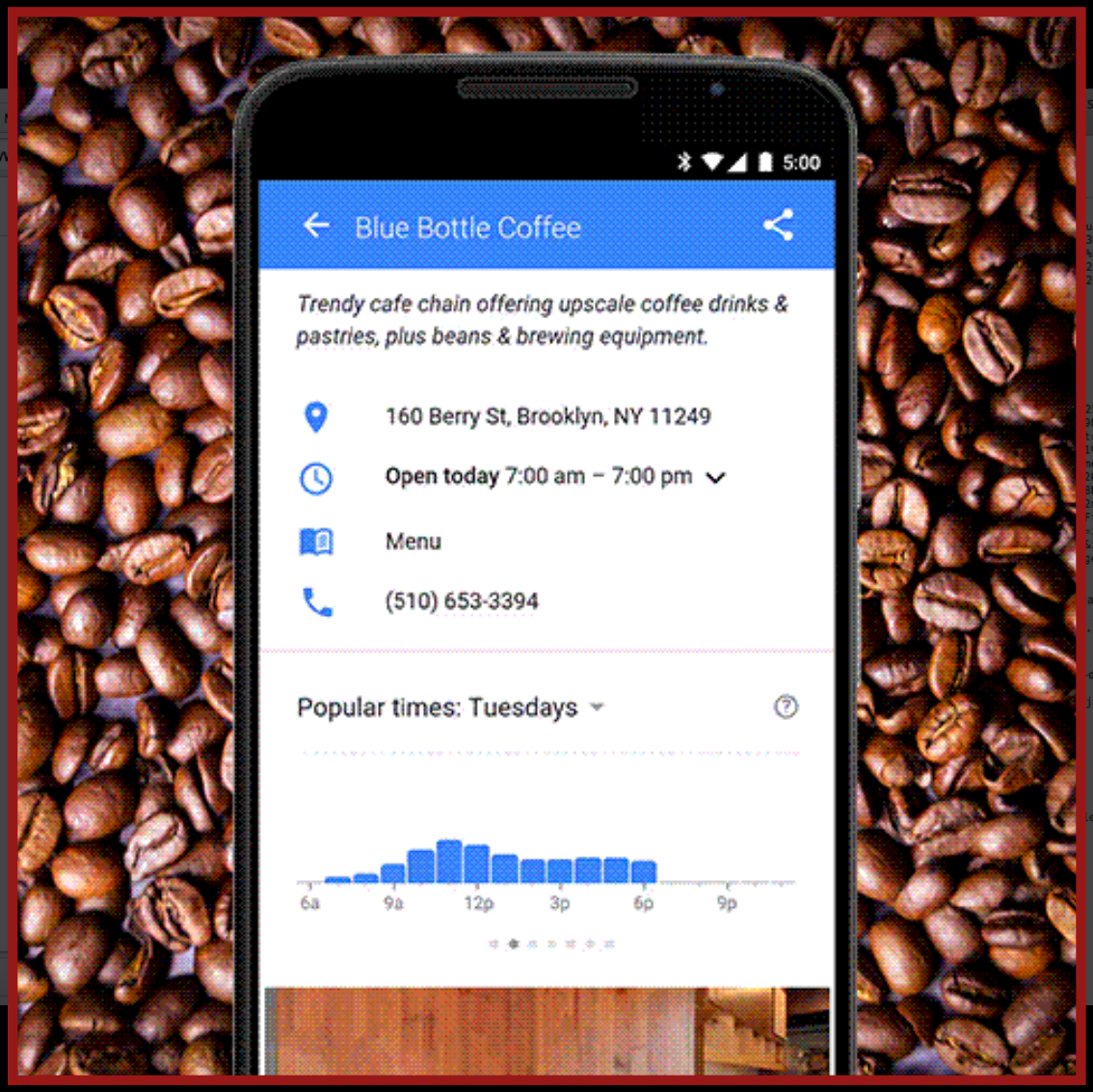
The lobby of The New York Times Building/Nic Lehoux

01/09/2014
The New York Times Company to Webcast Fourth-Quarter and Full-Year 2013 Earnings Conference Call »

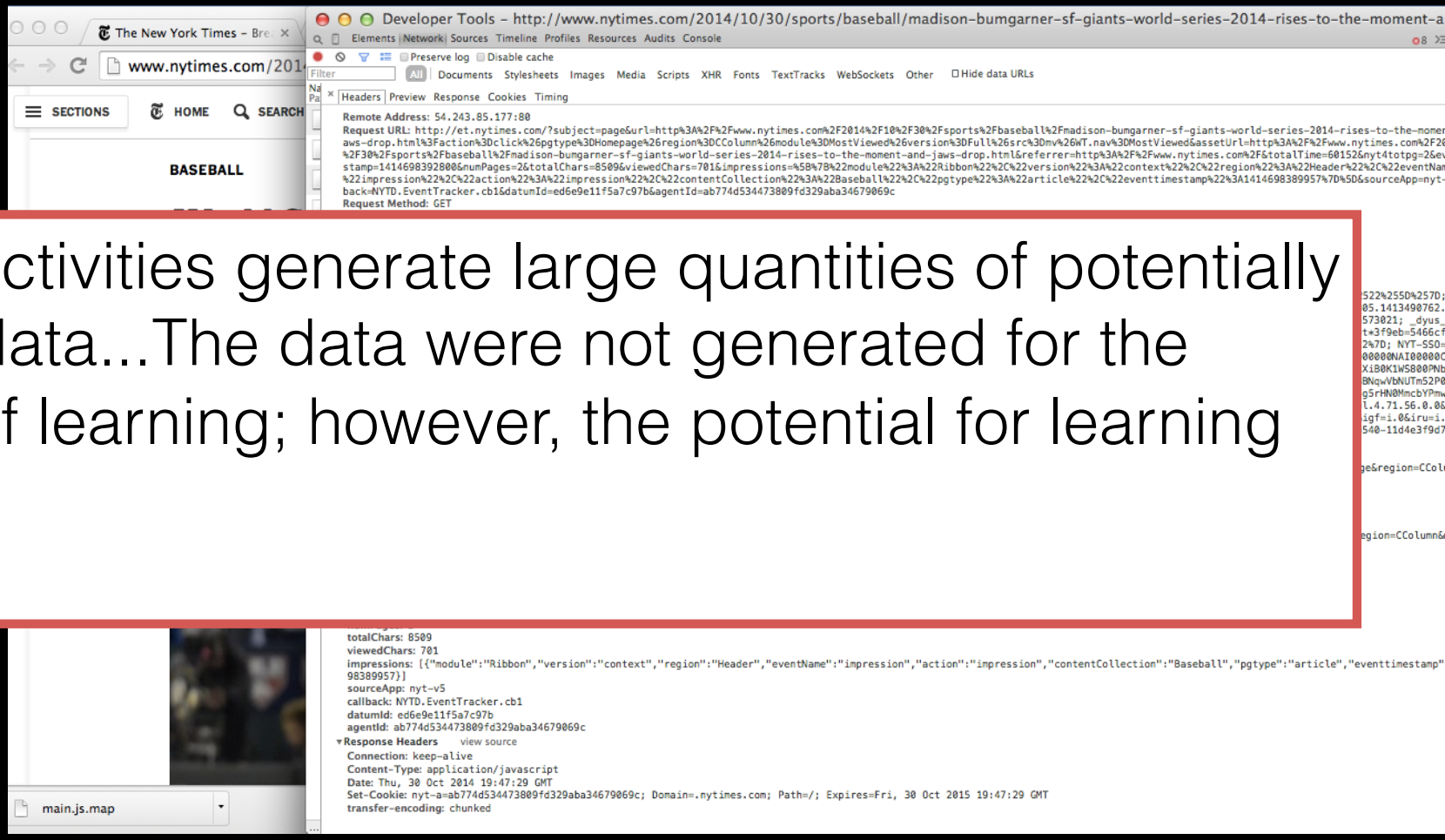
01/02/2014
The New York Times to Introduce Redesign of NYTimes.com to All Users Jan. 8 »

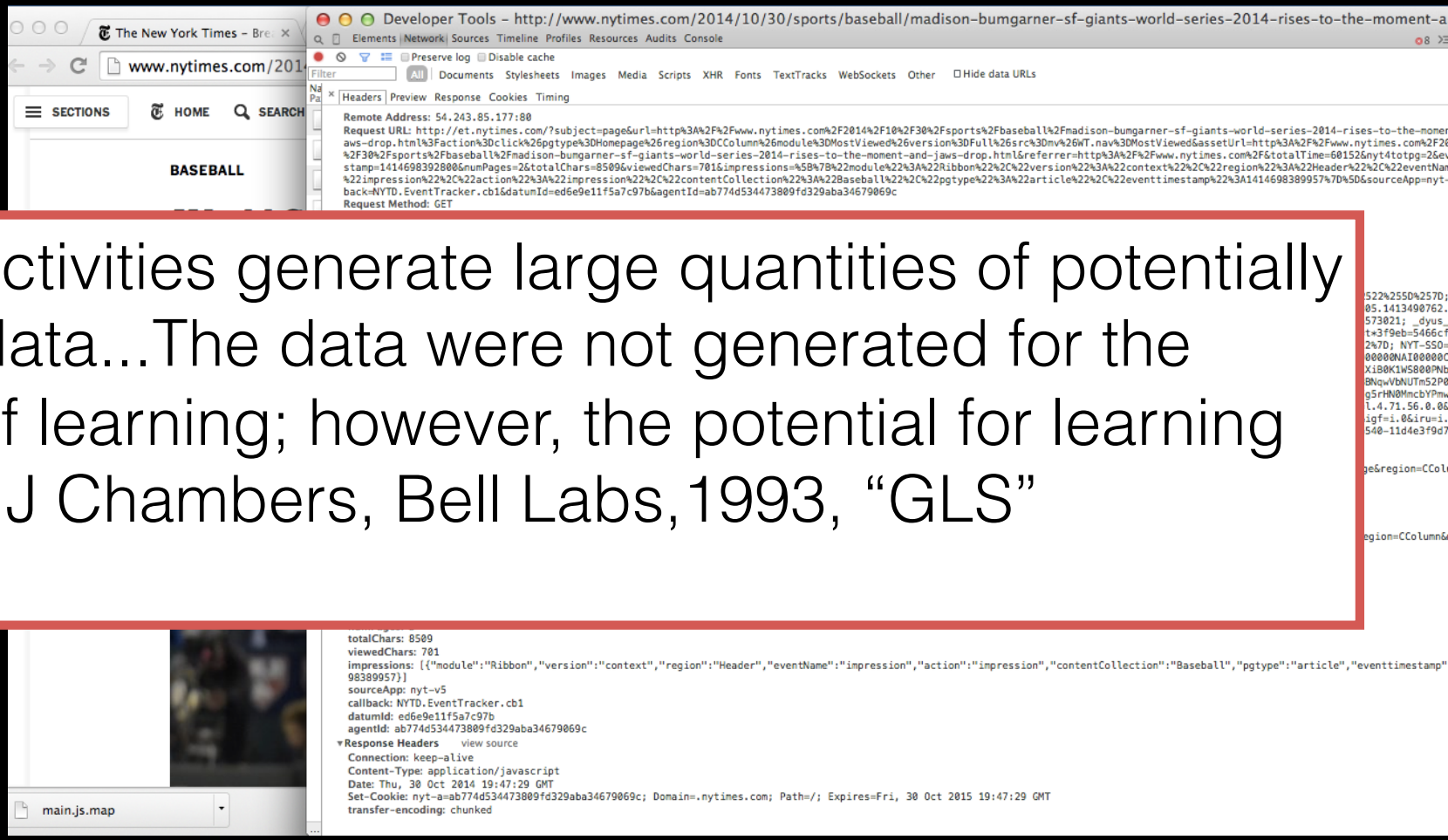
12/12/2013
The New York Times Company Declares Regular Quarterly Dividend »


The New York Times - Bre
www.nytimes.com/201
SECTIONS
HOME
SEARCH
BASELINE
World Se
Drop
OCT. 30, 2014
main.js.map
Developer Tools - http://www.nytimes.com/2014/10/30/sports/baseball/madison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-a
Elements Network Sources Timeline Profiles Resources Audits Console
Preserve log
Disable cache
Filter
All
Documents
Stylesheets
Images
Media
Scripts
XHR
Fonts
TextTracks
WebSockets
Other
Hide data URLs
Headers
Preview
Response
Cookies
Timing
Remote Address: 54.243.85.177:80
Request URL: http://et.nytimes.com/?subject=page&url=http%3A%2F%2Fwww.nytimes.com%2F2014%2F10%2F30%2Fsports%2Fbaseball%2Fmadison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment
aws-drop.html%3Faction%3Dclick%26pgtype%3DHomepage%26region%3DCColumn%26module%3DMostViewed%26version%3DFull%26src%3Dmv%26WT.nav%3DMostViewed&assetUrl=http%3A%2F%2Fwww.nytimes.com%2F2
%2F30%2Fsports%2Fbaseball%2Fmadison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html&referrer=http%3A%2F%2Fwww.nytimes.com%2F&totalTime=60152&nyt4totpg=2&ev
stamp=1414698392800&numPages=2&totalChars=8509&viewedChars=701&impressions=%5B%7B%22module%22%3A%22Ribbon%22%2C%22version%22%3A%22context%22%2C%22region%22%3A%22Header%22%2C%22eventNam
%22impression%22%2C%22action%22%3A%22impression%22%2C%22contentCollection%22%3A%22Baseball%22%2C%22pgtype%22%3A%22article%22%2C%22eventtimestamp%22%3A1414698389957%7D%5D&sourceApp=nyt-
back=NYTD.EventTracker.cb1&datumId=ed6e9e11f5a7c97b&agentId=ab774d534473809fd329aba34679069c
Request Method: GET
Status Code: 200 OK
Request Headers
view source
Accept: */*
Accept-Encoding: gzip,deflate,sdch
Accept-Language: en-US,en;q=0.8
Connection: keep-alive
Cookie: RMID=007f01010de45403ded40002; __qca=P0-2108091050-1409864087731; NYT-V00=WC2jafrk%252FPoVYu7eMvCSfA%2523%257B%2522NYT-Z%2522%253A%25581443980047%252C%252210011%2522%255D%257D;
egs=H07707_11063|H07707_11173|H07707_11173|H07707_0; madison_user_id=chez2j0pRJGk_xKSDCK5lw; madison_guest=false; _ga=GA1.2.1055279013.1409539805; __utma=69104142.1055279013.1409539805.1413490762.
6768.18; __utmz=69104142.1413490762.17.11.utmcsr=google|utmccn=(organic)|utmcmd=organic|utmctr=(not%20provided); _dycmc=5; _dycst=tg.m.frv1.ms.tos.cd.c.; _dycabc=1414186573021; _dyus_
0=46%7C0%7C0%7C0%7C0%7C0.0.1410402680463.1414186572522.3783892.0%7C296%7C43%7C9%7C114%7C1%7C0%7C0%7C0%7C0%7C0%7C1%7C0%7C1%7C1%7C1%7C0%7C2%7C2%7C4%7C0%7C0%7C0%7C0; adxcl=t*3f9eb=5466cf
3542595|t*3f1d5=546ac43f:1413834221|l*355c5=55c3b57c:1|lcig=55c3b57c:1|l*31068=566cfb4f:1|t*3e459=5466cfbf:1412444018; nyt-recommend=1; NYT-mab=%7B%221%22%3A%22RI%3A11%22%7D; NYT-SSO=
ny/8z1i5nexBEjjJ2Emuan8isxP1k02el70fzxLHktOrlpJFq1h9jSt68Dr; NYT_W2=New%20YorkNYUS|ChicagoILUS|London--UK|Los%20AngelesCAUS|San%20FranciscoCAUS|Tokyo--JP; nyt-d=101.000000000NAI00000C
6ous0z1Iq/0eVpii0qApjv1eUY4r0M4meE1/FIm50TUY4r0426Ca0f6XL/0zB0eP1w8JKC0TOoGf0D4tyz0i40rw0XDG8F1Z92aU0IVpqi0G47eX0r0X1Z066miJ07PH8b0gDJOz1w0mSH0M5Nmq0gBYat1w0mS60M6tm21Z1XiB0K1WS800PNb
l1sVrX1@7b69d99b/eb638adc; NYT-S=4gx7BLBHJ1Hy3RzrhDhksL7M1Q05Cz7KOxMuD3DAIi8H8rSfII.cvq.K0edd1HdW2KHIe5CNtYJZuzcnKBfsswXzGGE95BF2mvBXQwTtyrhaUclJQo2xLkrXoaN3pCITpD059LPBBNqwVbNUTm52P0
dHREHJr.GzZRFVO7M69RpkMT9K2MjbvGFm/j9R.1mvLmTDbcG0ja.T2..gpW5ZH7IN33q.ati78gqCeABF5D5egXXa/YcPf93.3KOm7PMTpe4FjP1hTrBhshVbemxJMZFfDpM30yEXN0jvh/97GNjStPglsUziO8X6Y2vomd/g5rHN0MmcbYPmw
marnLcnUBLzVvSpZreYg0; NYT-wpAB=0002|1&0008|0; nyt-m=90EDCC0E9DF46FEFF9E538772F8843FD&e=i.1414814400&t=i.10&v=i.0&n=i.2&g=i.0&rc=i.0&er=i.1412131801&vr=l.4.53.44.0.0&pr=l.4.71.56.0.0&
4&gf=l.10.-1.-1.-1.-1.-1.-1.-1.-1.-1.-1&ft=i.0&fv=i.0&rl=l.1.-1&cav=i.53&imu=i.1&igu=i.1&prt=i.5&kid=i.1&ica=i.1&iue=i.1&ier=i.0&iub=i.0&ifv=i.0&igd=i.0&iga=i.1&imv=i.0&igf=i.0&iru=i.
i.0&ira=i.1&iir=i.1&l=l.15.-1.-1.-1.-1.-1.-1.-1.-1.-1.-1.-1.-1.-1.-1&gl=l.2.2997621288.-1&gb=l.3.0.5.1414641600; _dyrc=; _dy_geo=US.NA; WT_FPC=id=7342b9da-13eb-4b0e-b540-11d4e3f9d7
1414694790807:ss=1414694761616; nyt-a=ab774d534473809fd329aba34679069c
Host: et.nytimes.com
Referer: http://www.nytimes.com/2014/10/30/sports/baseball/madison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html?action=click&pgtype=Homepage&region=CCol
ule=MostViewed&version=Full&src=mv&WT.nav=MostViewed
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/38.0.2125.111 Safari/537.36
Query String Parameters
view source
view URL encoded
subject: page
url: http://www.nytimes.com/2014/10/30/sports/baseball/madison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html?action=click&pgtype=Homepage&region=CColumn&
MostViewed&version=Full&src=mv&WT.nav=MostViewed
assetUrl: http://www.nytimes.com/2014/10/30/sports/baseball/madison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html
referrer: http://www.nytimes.com/
totalTime: 60152
nyt4totpg: 2
eventtimestamp: 1414698392800
numPages: 2
totalChars: 8509
viewedChars: 701
impressions: [{"module":"Ribbon","version":"context","region":"Header","eventName":"impression","action":"impression","contentCollection":"Baseball","pgtype":"article","eventtimestamp"
98389957}]
sourceApp: nyt-v5
callback: NYTD.EventTracker.cb1
datumId: ed6e9e11f5a7c97b
agentId: ab774d534473809fd329aba34679069c
Response Headers
view source
Connection: keep-alive
Content-Type: application/javascript
Date: Thu, 30 Oct 2014 19:47:29 GMT
Set-Cookie: nyt-a=ab774d534473809fd329aba34679069c; Domain=.nytimes.com; Path=/; Expires=Fri, 30 Oct 2015 19:47:29 GMT
transfer-encoding: chunked


5:00
Blue Bottle Coffee
Trendy cafe chain offering upscale coffee drinks & pastries, plus beans & brewing equipment.
160 Berry St, Brooklyn, NY 11249
Open today 7:00 am – 7:00 pm
Menu
(510) 653-3394
Popular times: Tuesdays
6a
9a
12p
3p
6p
9p

"...social activities generate large quantities of potentially valuable data...The data were not generated for the purpose of learning; however, the potential for learning is great"





"...social activities generate large quantities of potentially valuable data...The data were not generated for the purpose of learning; however, the potential for learning is great" - J Chambers, Bell Labs,1993, "GLS"

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


New-York Daily Times.

VOL. I., NO. 1. NEW-YORK, THURSDAY, SEPTEMBER 18, 1851. PRICE ONE CENT.

1851

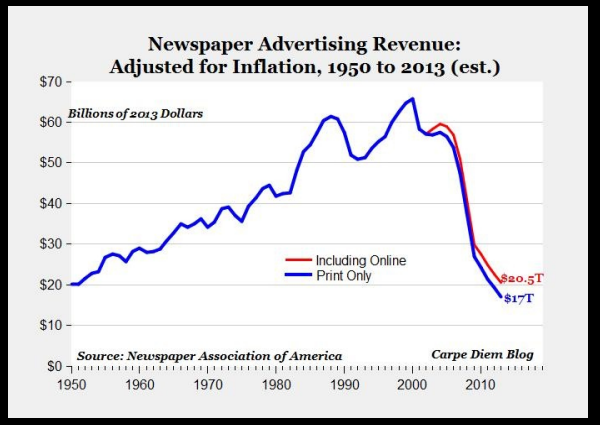
The New York Times Introduces a Web Site

By PETER H. LEWIS
Published: January 22, 1996

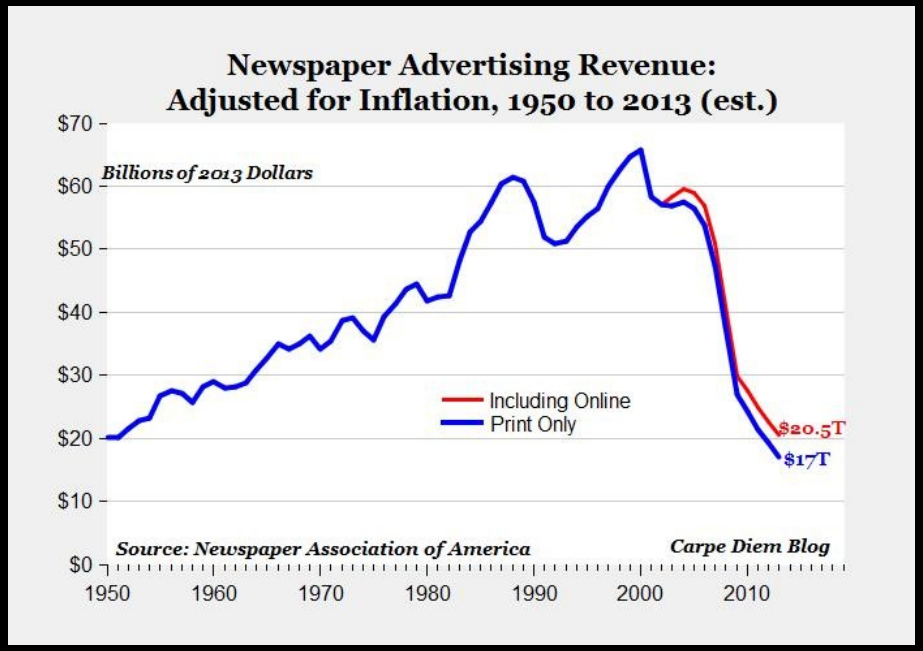
The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper's contents.

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.

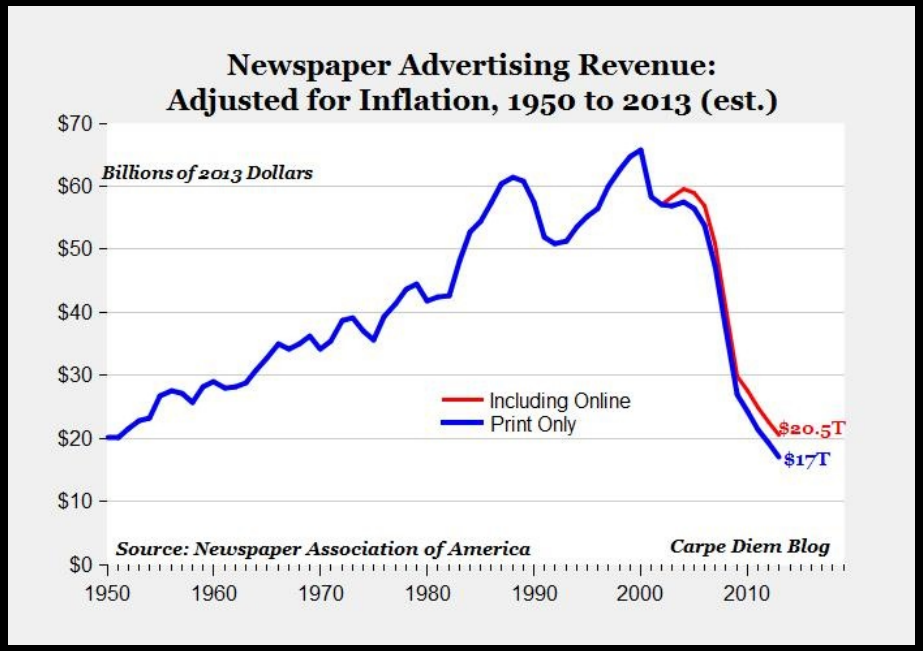
1996



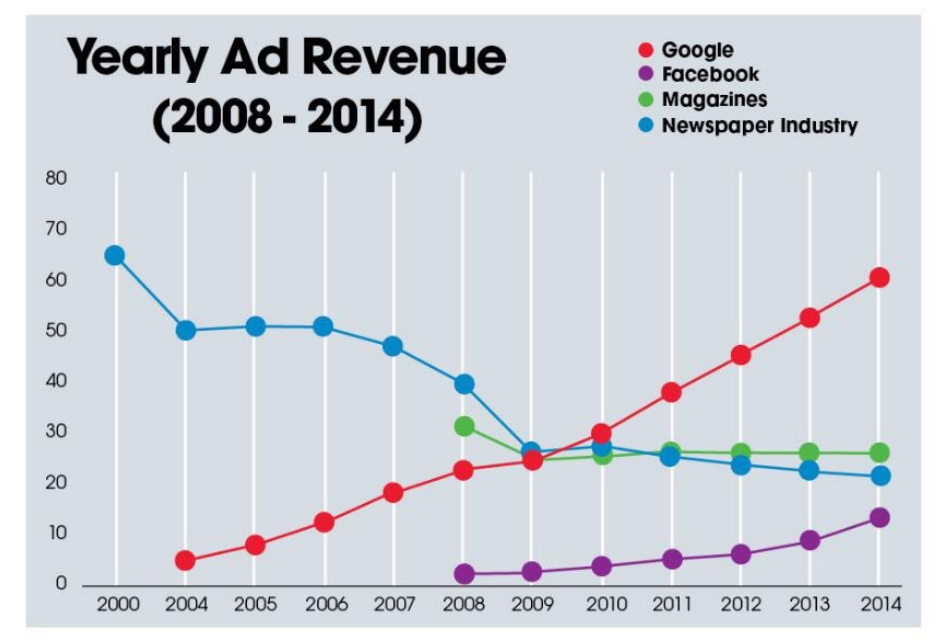
2008



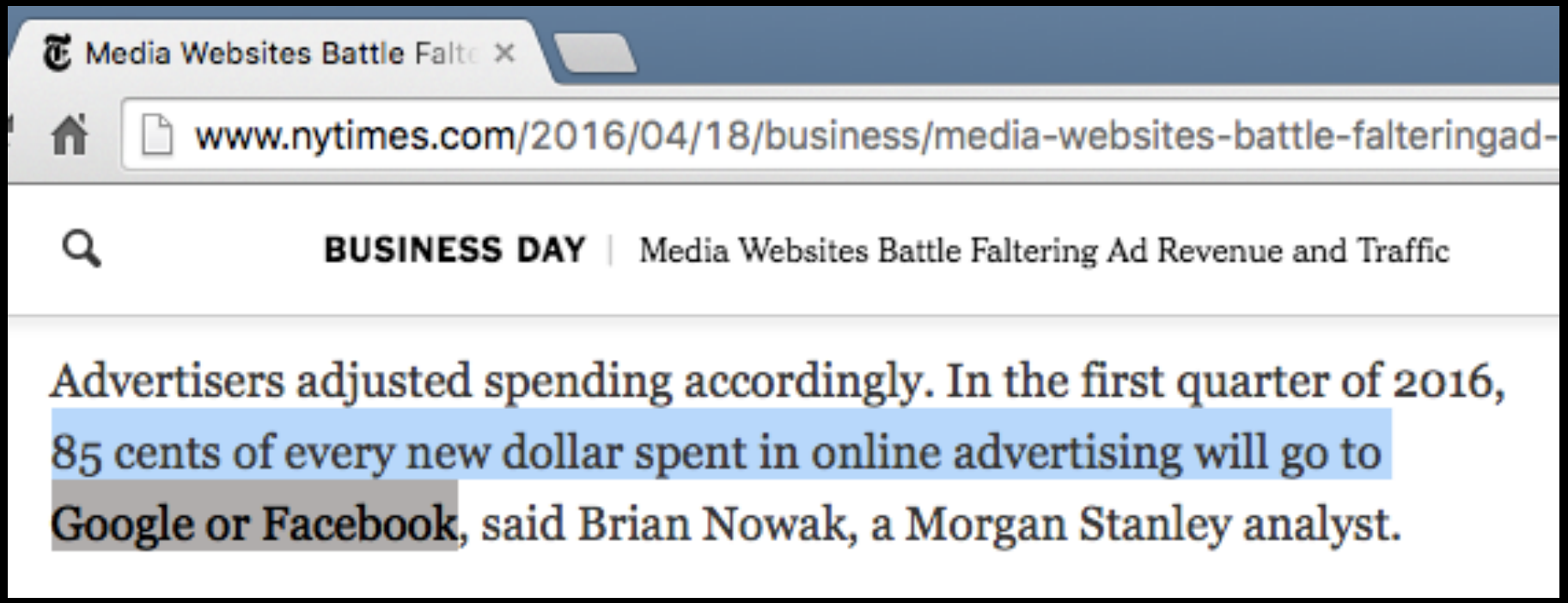
"a startup is a temporary organization in search of a repeatable and scalable business model" –Steve Blank



every publisher is now a startup



every publisher is now a startup

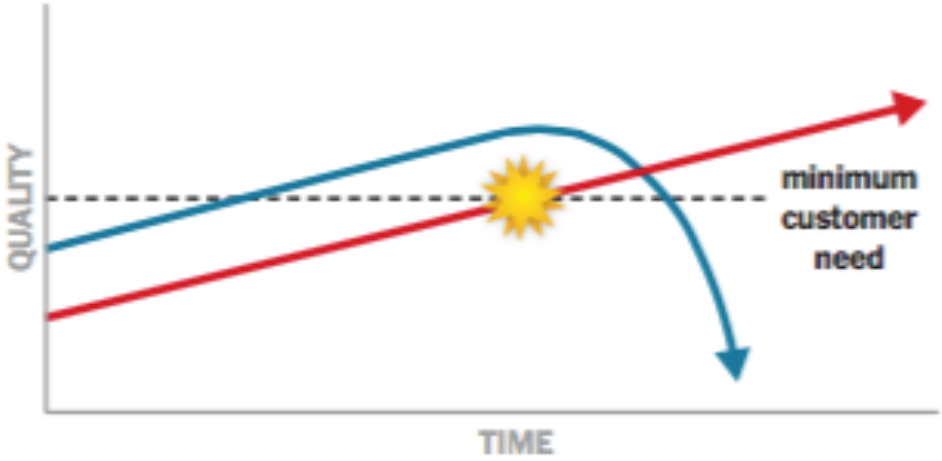


every publisher is now a startup


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.

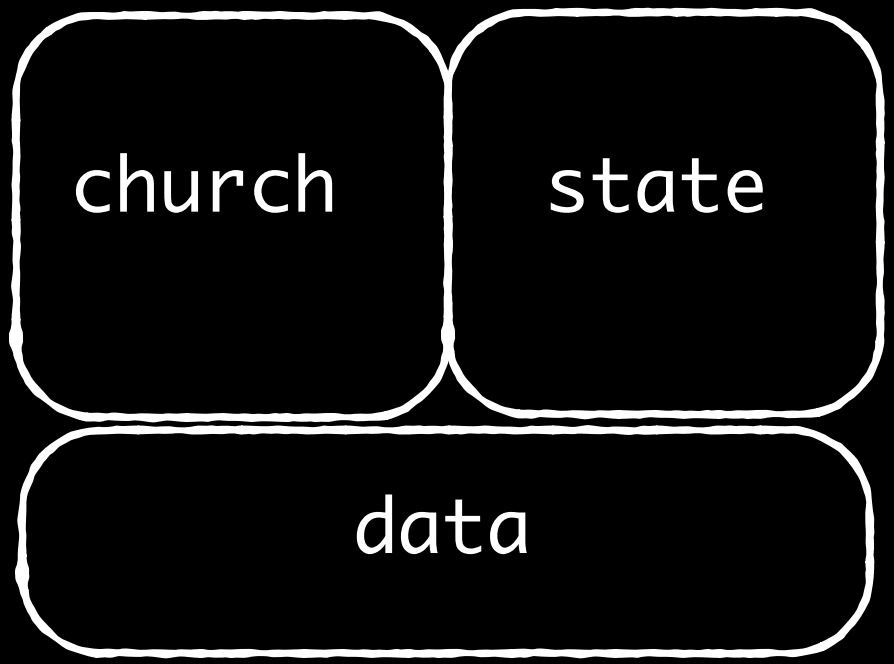


HALLMARKS OF DISRUPTIVE INNOVATORS

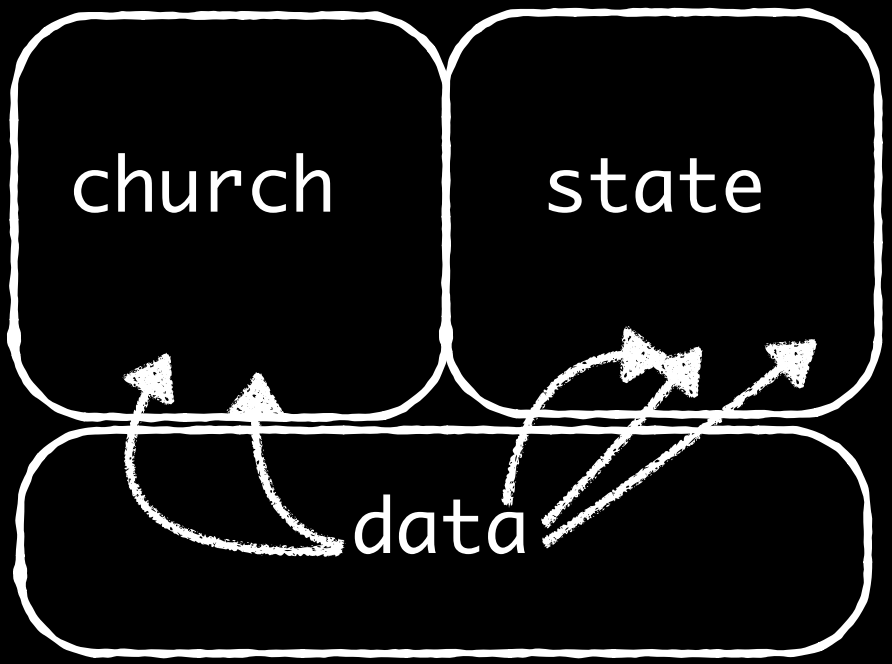
- Introduced by an "outsider"
- Less expensive than existing products
- Targeting underserved or new markets
- Initially inferior to existing products
- Advanced by an enabling technology

circa vox BUSINESS INSIDER H POLITICO BuzzFeed

news: 21st century



news: 21st century



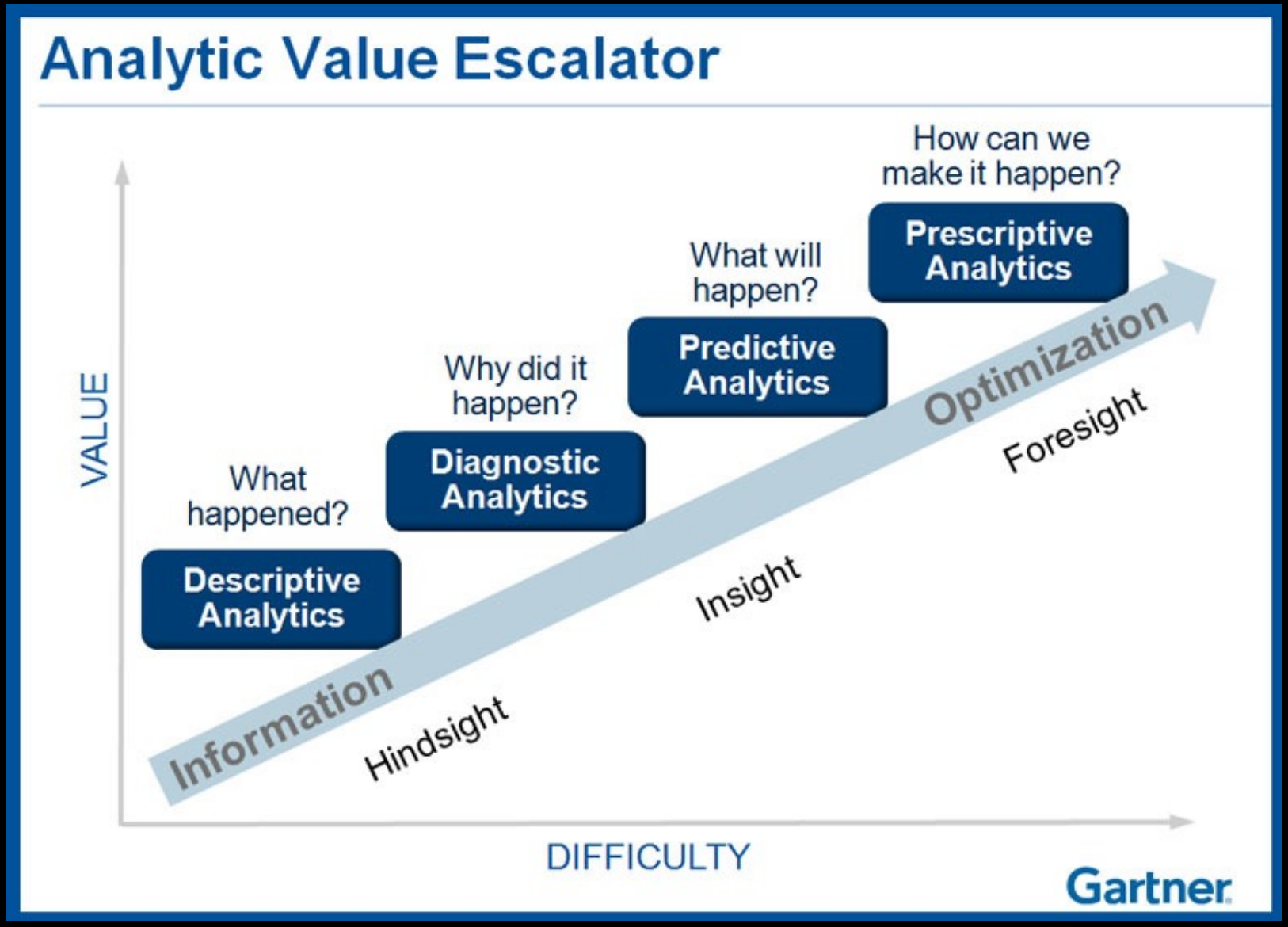
learnings

learnings

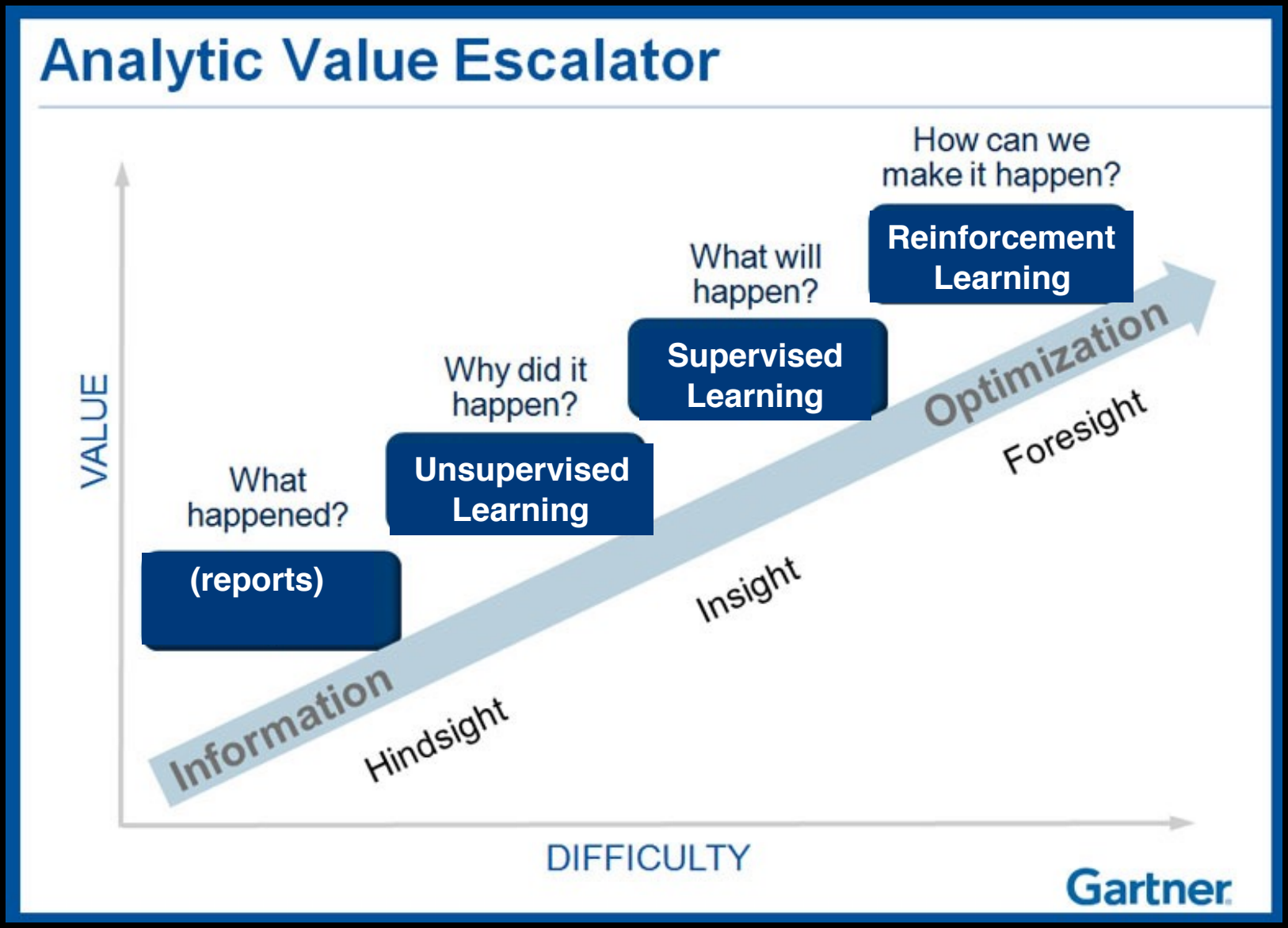
- predictive modeling
- descriptive modeling
- prescriptive modeling

(actually ML, shhhh…)

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



h/t michael littman

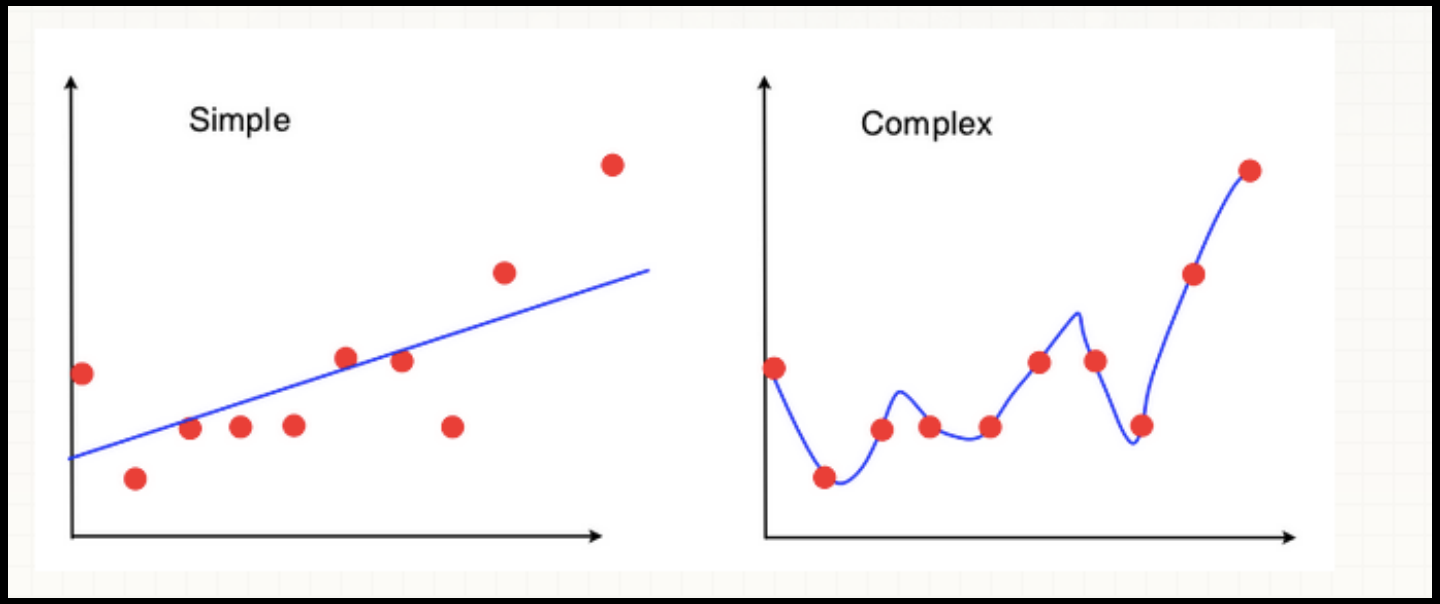


h/t michael littman

learnings

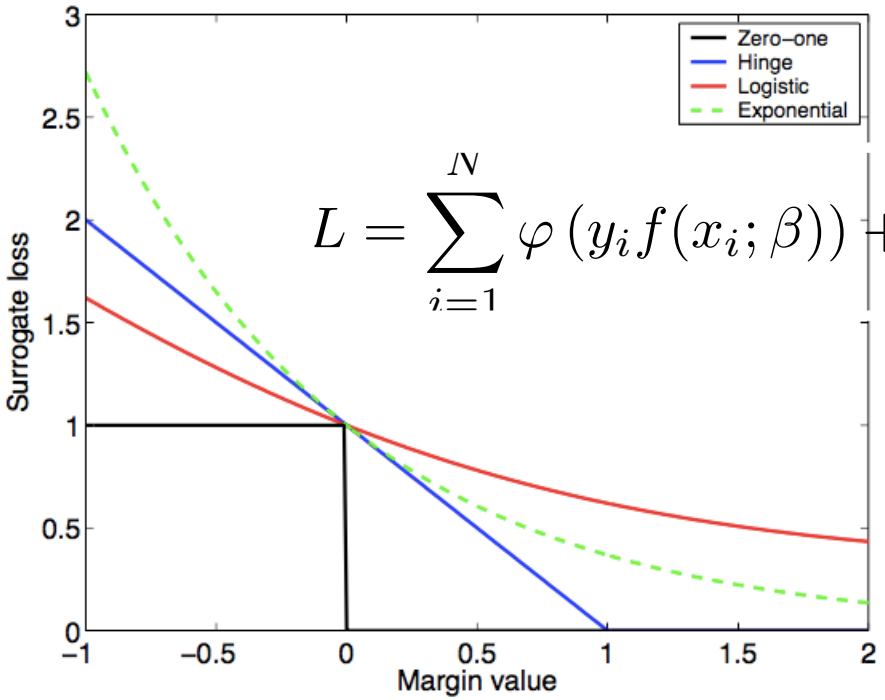
- predictive modeling
- descriptive modeling
- prescriptive modeling

cf. modelingsocialdata.org



stats.stackexchange.com

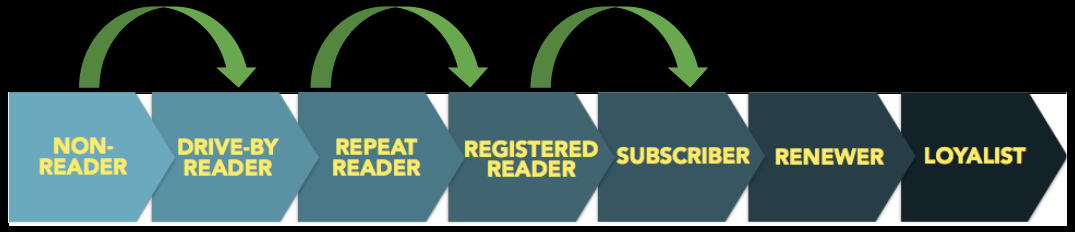
Margin-Based Surrogate Loss Functions



$$L = \sum_{i=1}^{N} \varphi\left(y_i f(x_i;\beta)\right) + \lambda||\beta|$$

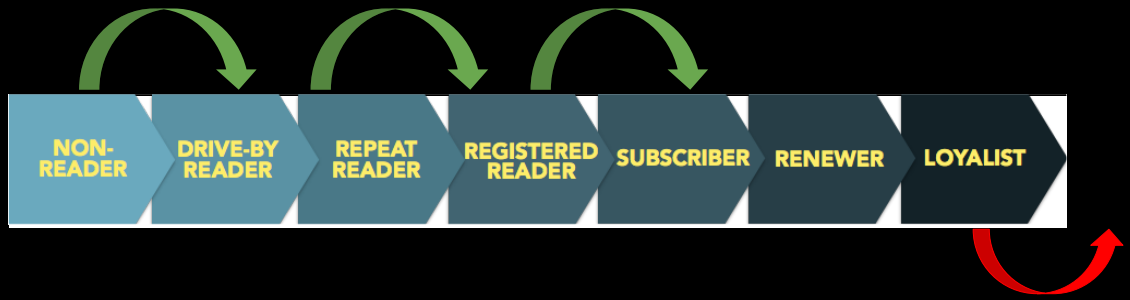
from “are you a bayesian or a frequentist”
–michael jordan

predictive modeling, e.g.,



cf. modelingsocialdata.org

predictive modeling, e.g.,

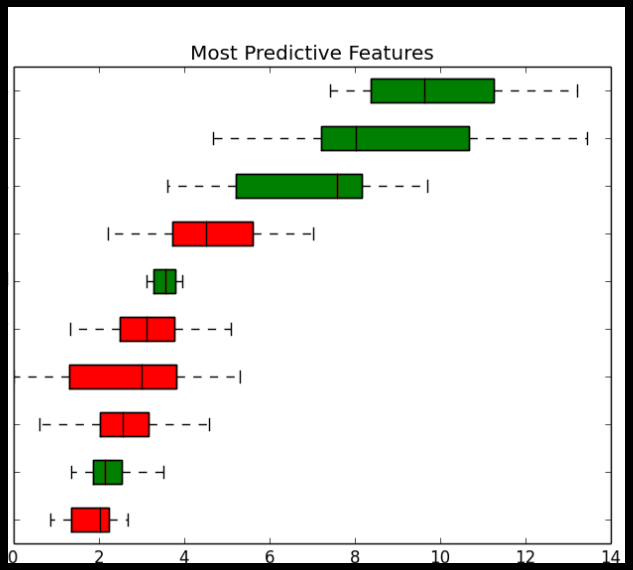


“the funnel”

cf. modelingsocialdata.org

interpretable predictive modeling

super cool stuff

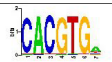
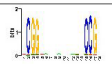
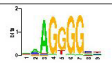
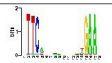
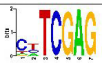
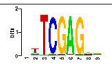
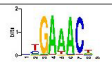
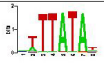
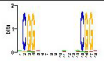
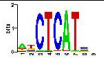
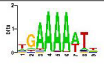
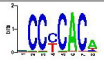
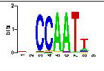


cf. modelingsocialdata.org

interpr ng

super cool stuff

cf. m

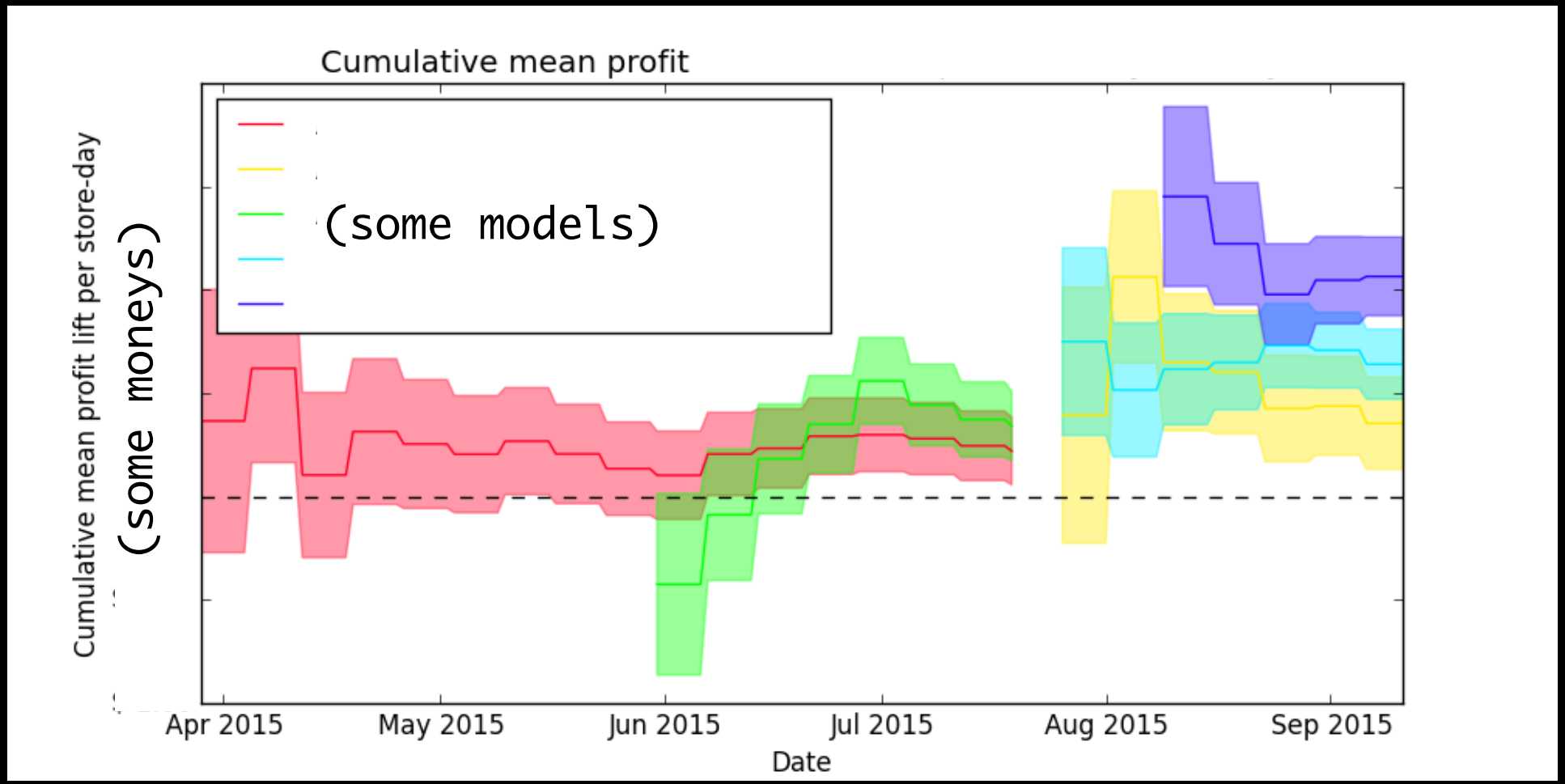
TFNAME	DB-MOTIF	MOTIF	DBNAME	d(p,q)
CBF1	CACGTG		YPD	0.032635
CGG everted repeat	CGGN*CCG		YPD	0.032821
MSN2			TRANSFAC	0.085626
HSF1	TTCNNNGAA		SCPD	0.102410
XBP1			TRANSFAC	0.140561
STE12			TRANSFAC	0.256750
GCN4			SCPD	0.292221
TBP			TRANSFAC	0.376601
HAP1	CGGNNNTWNCGG		YPD	0.423004
RAP1	RMACCCA	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

arxiv.org/abs/q-bio/0701021

optimization & learning, e.g.,

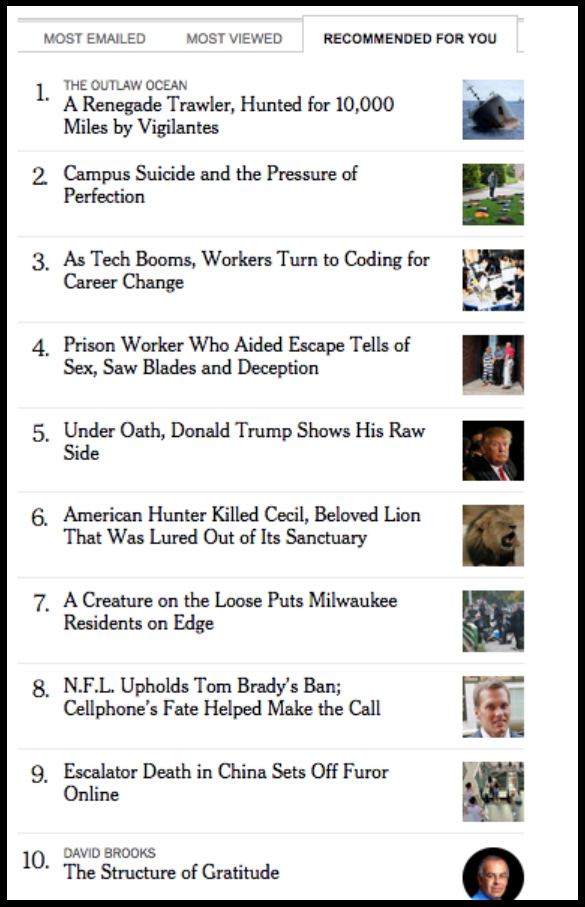
“How The New York Times Works “popular mechanics, 2015

optimization & prediction, e.g.,

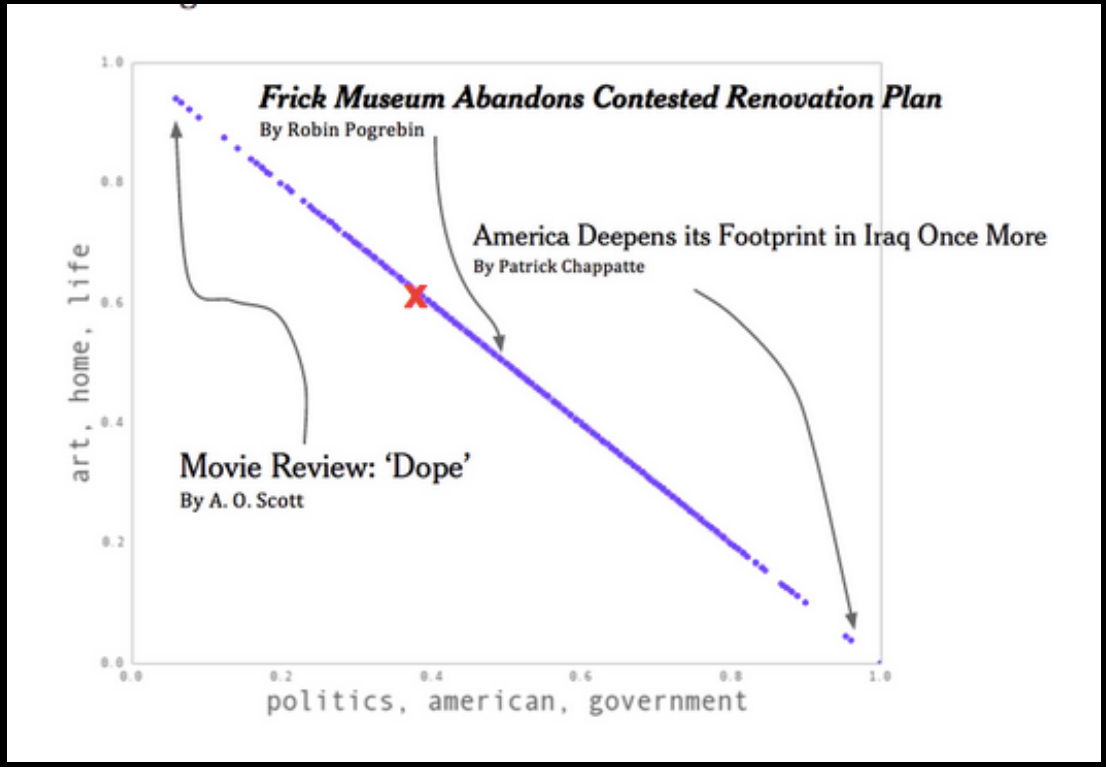


“newsvendor problem,” literally (+prediction+experiment)

recommendation as inference



recommendation as inference



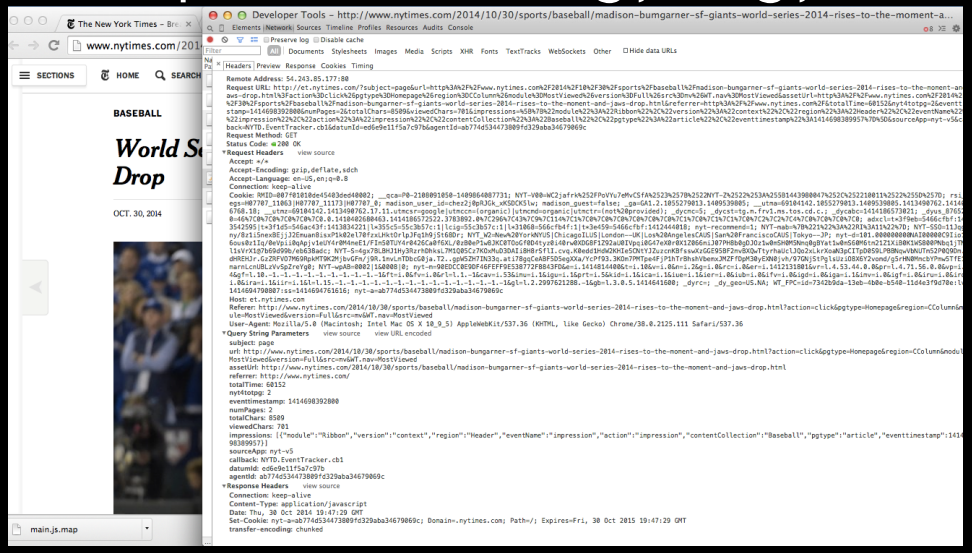
bit.ly/AlexCTM

descriptive modeling, e.g,

“segments”

cf. modelingsocialdata.org

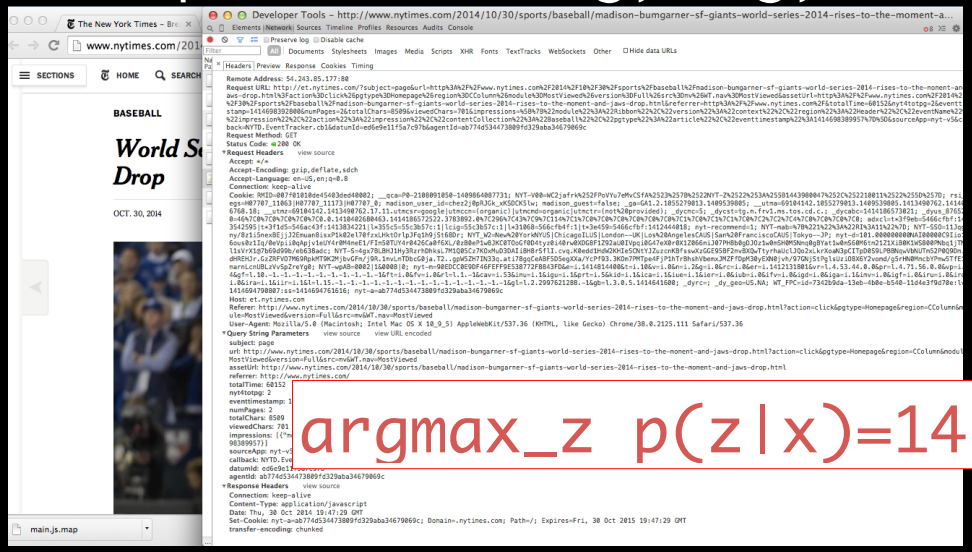
descriptive modeling, e.g,



“segments”

cf. modelingsocialdata.org

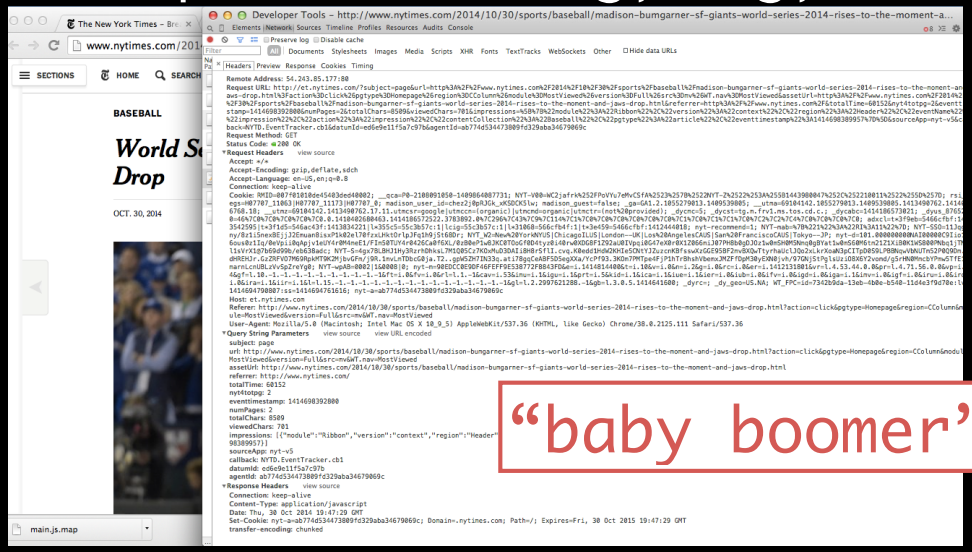
descriptive modeling, e.g,



“segments”

cf. modelingsocialdata.org

descriptive modeling, e.g,



“segments”

cf. modelingsocialdata.org

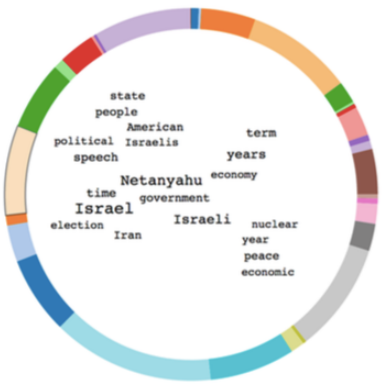
- descriptive data product

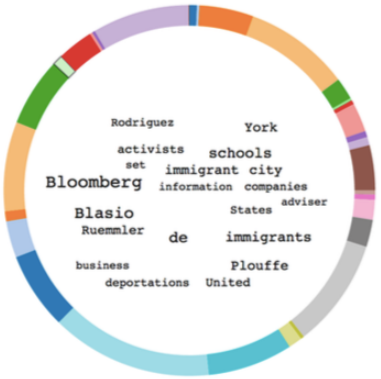
A Quick Refinery Demo

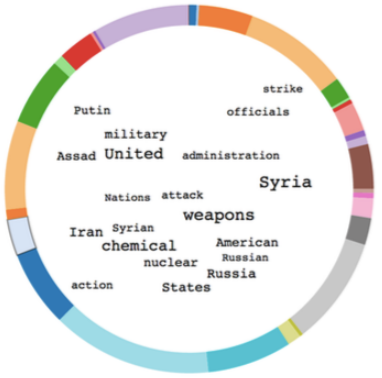
The New York Times Developers
Article Search API v2

Extracting NYT articles from keyword “obama” in 2013.

What themes / topics defined the Obama administration during 2013?

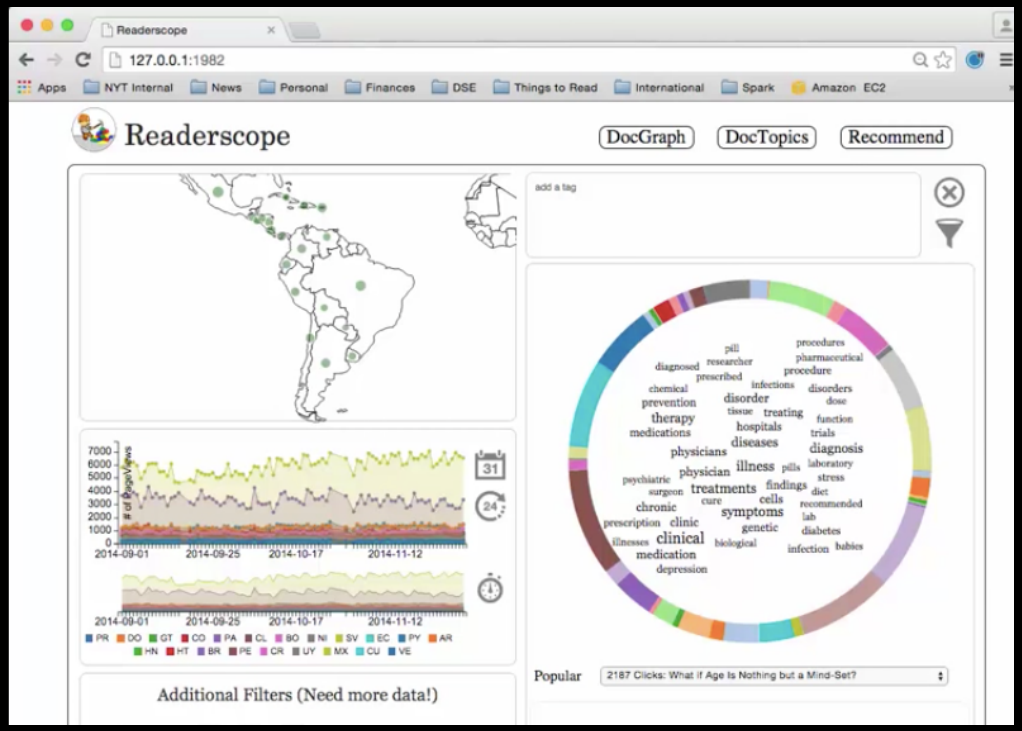




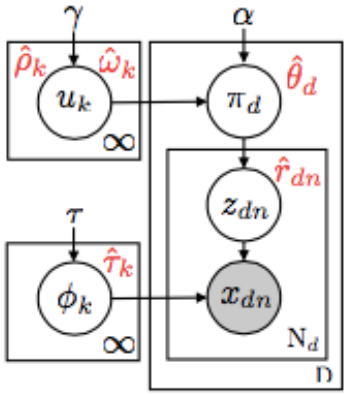


cf. daeilkim.com

descriptive modeling, e.g,



cf. daeilkim.com ; import bnpy



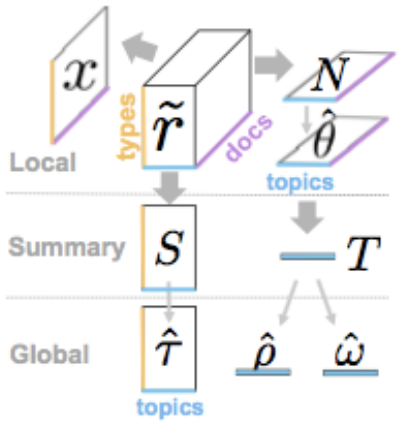


Figure 1: *Left:* Directed graphical model for the HDP admixture (Sec. 2). Free parameters for mean-field variational inference (Sec. 3) shown in red. *Right:* Flow chart for our inference algorithm, specialized for bag-of-words data, where we can use sparse type-based assignments $\tilde{r}$ instead 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

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):

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

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

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

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)

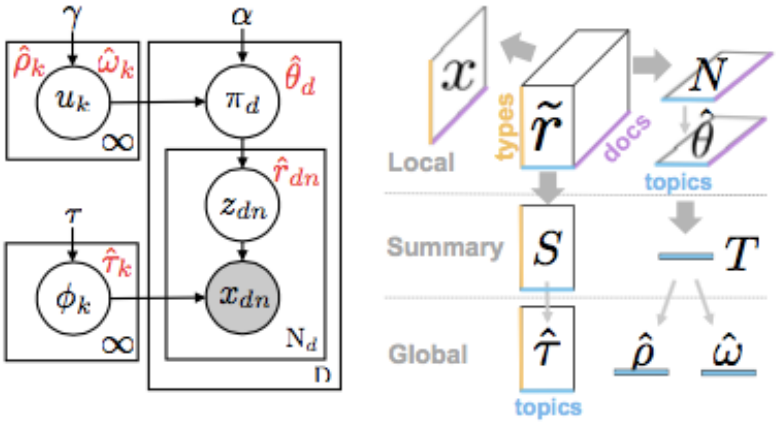


Figure 1: *Left:* Directed graphical model for the HDP admixture (Sec. 2). Free parameters for mean-field variational inference (Sec. 3) shown in red. *Right:* Flow chart for our inference algorithm, specialized for bag-of-words data, where we can use sparse type-based assignments $\tilde{r}$ instead 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

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

"off policy value estimation"
(cf. "causal effect estimation")

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

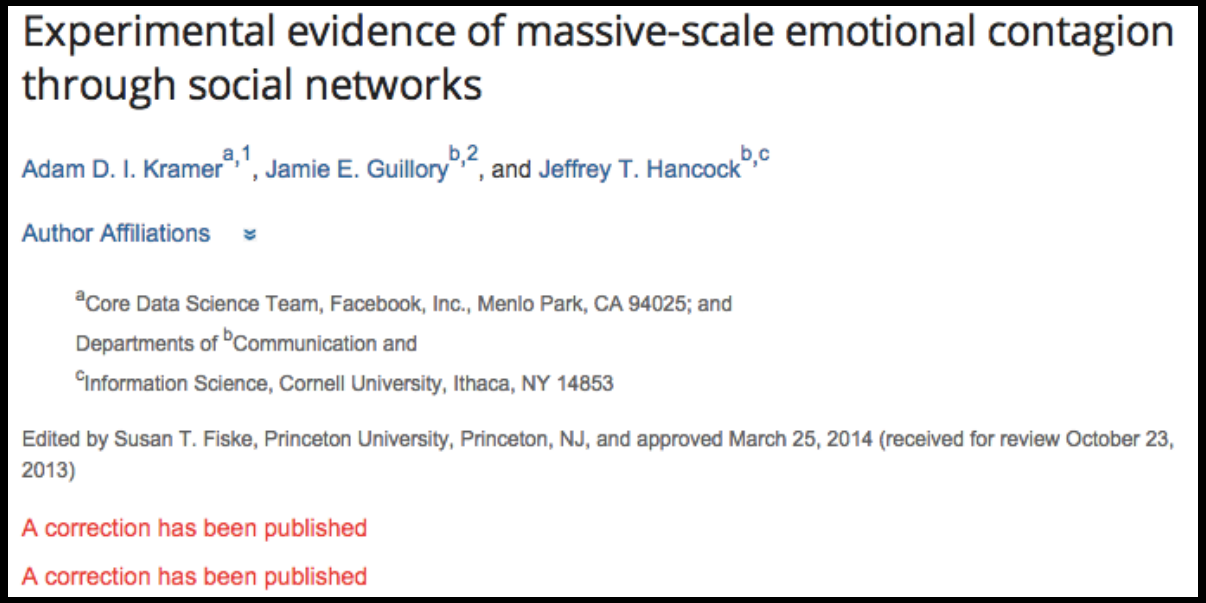
cf. Langford `08-`16;
Horvitz & Thompson `52;
Holland `86

"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[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Author Affiliations

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and

Departments of [b]Communication and

[c]Information Science, Cornell University, Ithaca, NY 14853

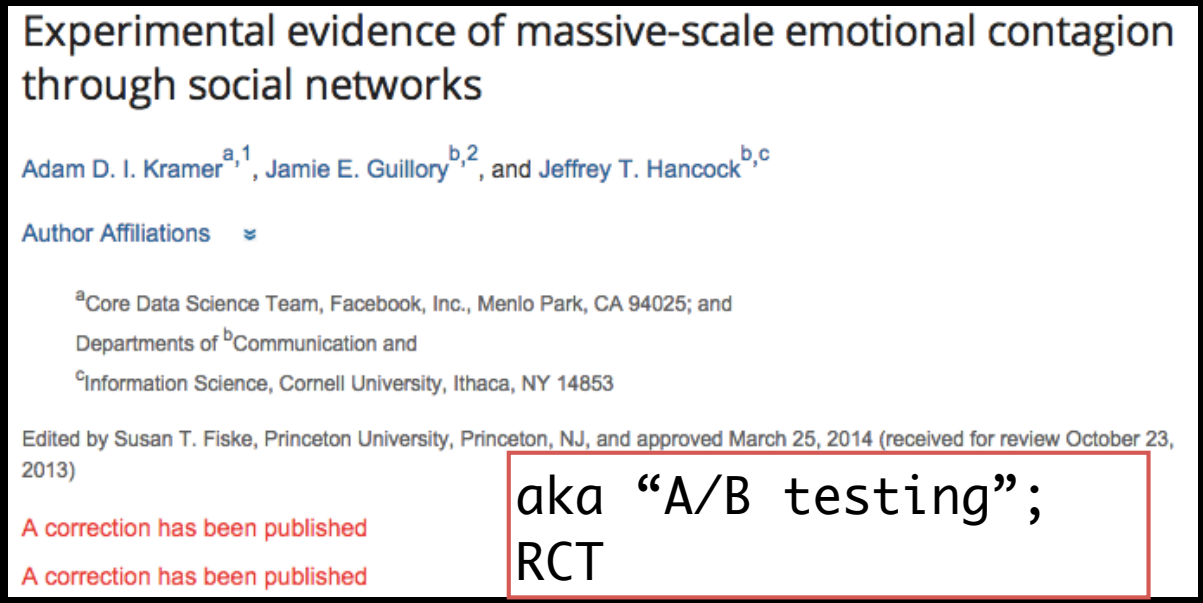
Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

A correction has been published

A correction has been published

cf. modelingsocialdata.org

prescriptive modeling



Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Author Affiliations

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and

Departments of [b]Communication and

[c]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)

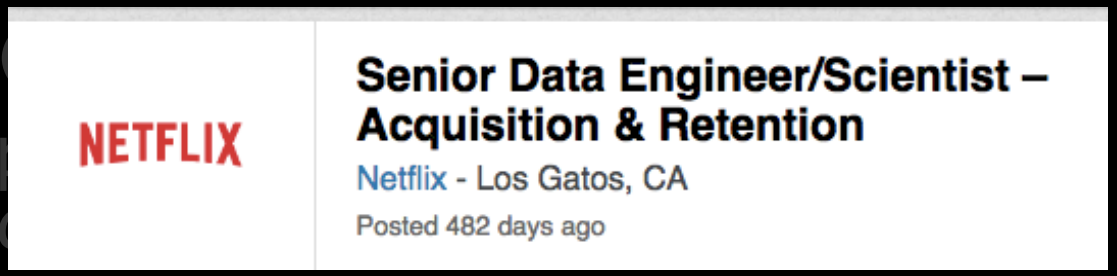
A correction has been published

A correction has been published

aka “A/B testing”; RCT

cf. modelingsocialdata.org

prescriptive modeling: from A/B to….

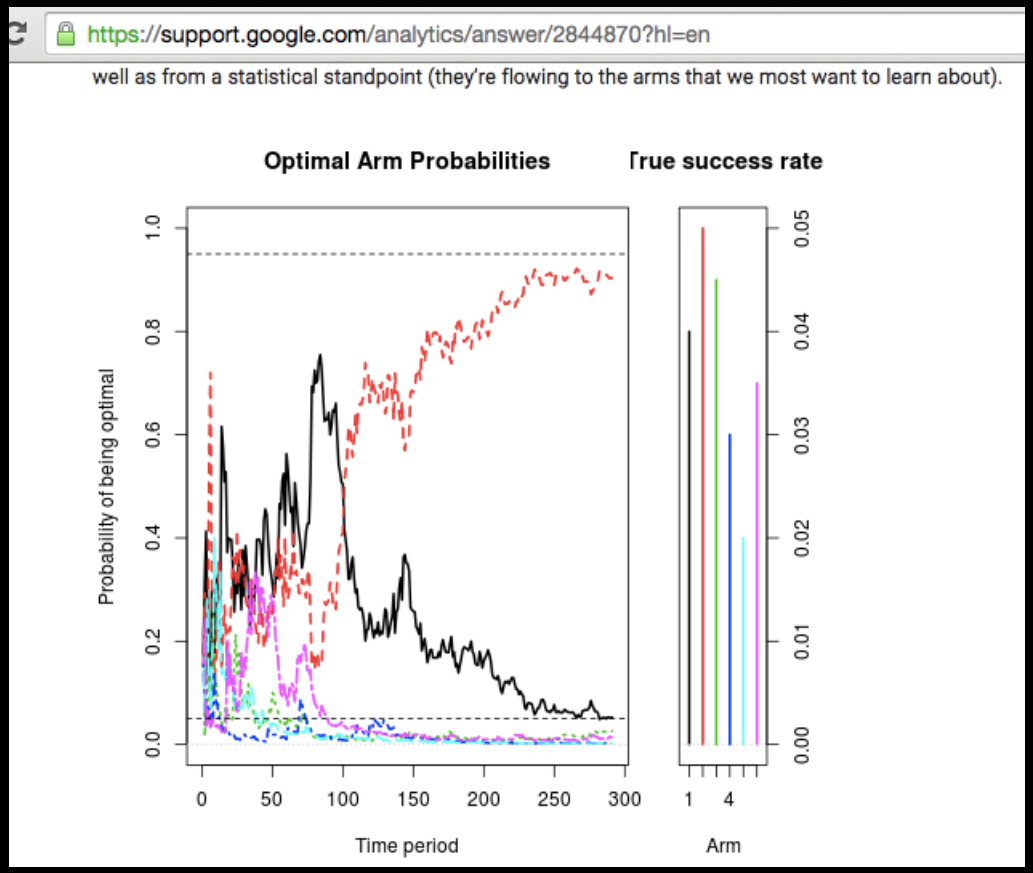


Some of the most recognizable personalization in our service is the collection of “genre” rows. …Members connect with these rows so well that we measure an **increase in member retention** by **placing the most tailored rows higher** on the page instead of lower.

cf. modelingsocialdata.org

real-time A/B -> “bandits”

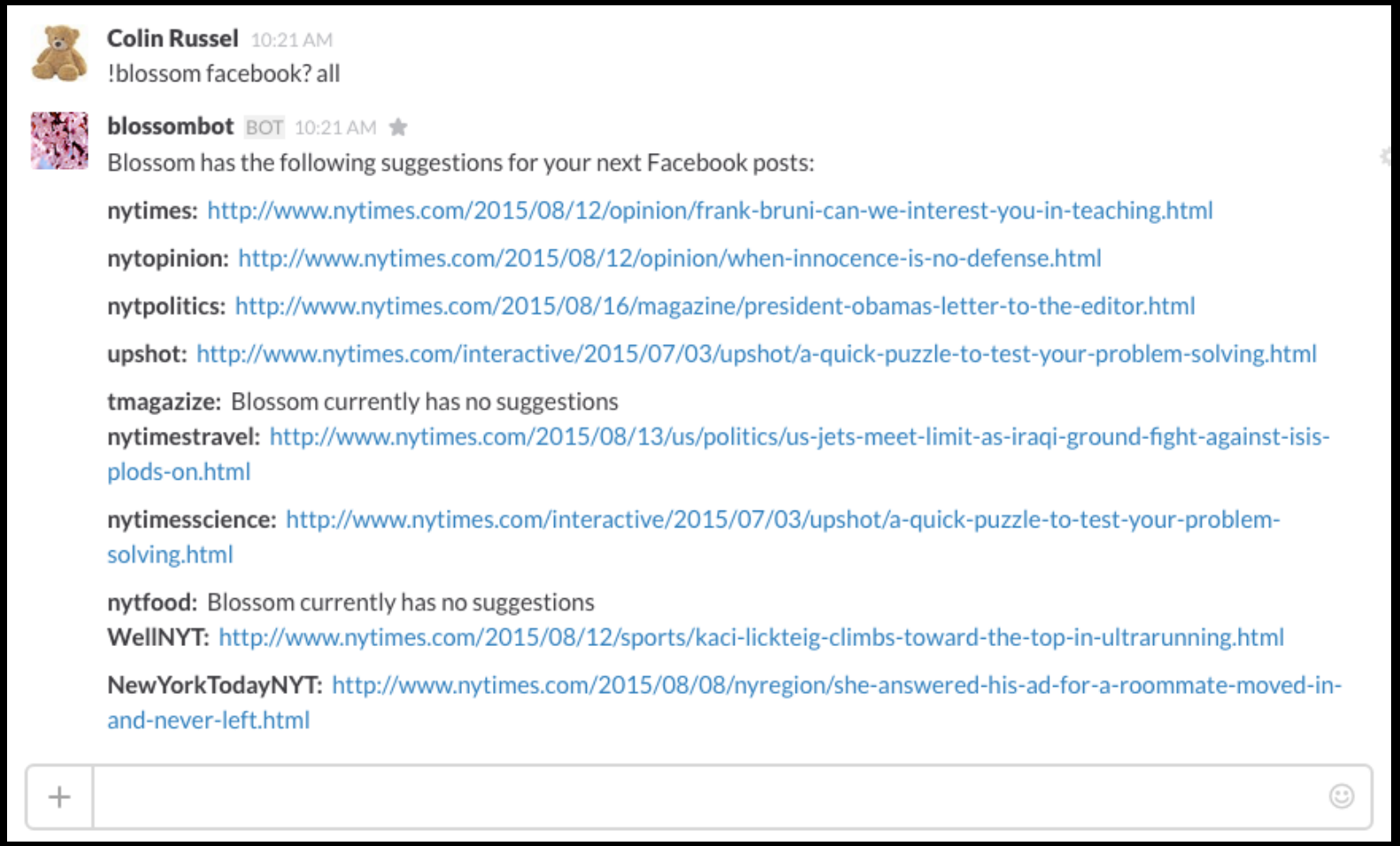
GOOG blog:



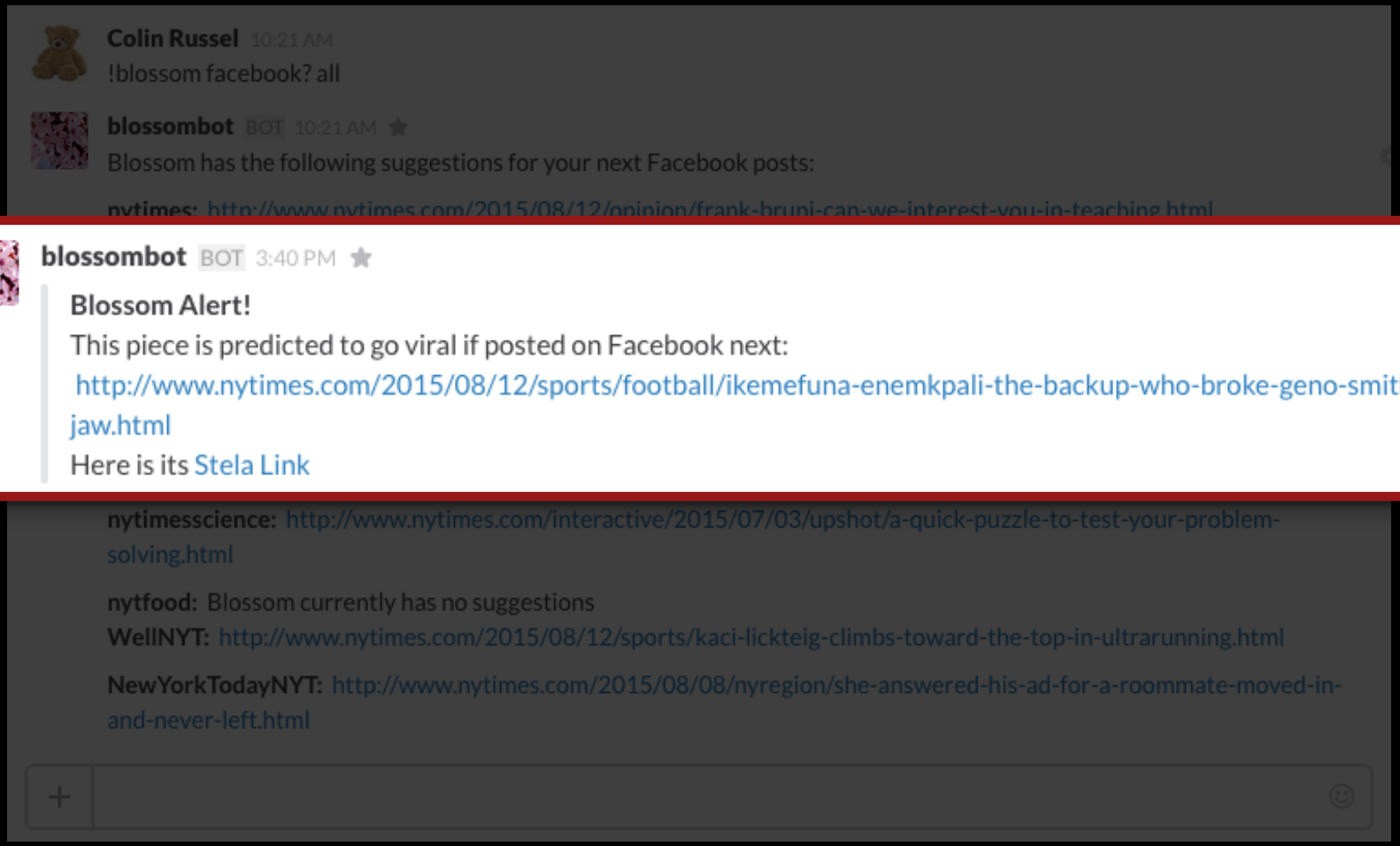
cf. modelingsocialdata.org

prescriptive modeling, e.g,

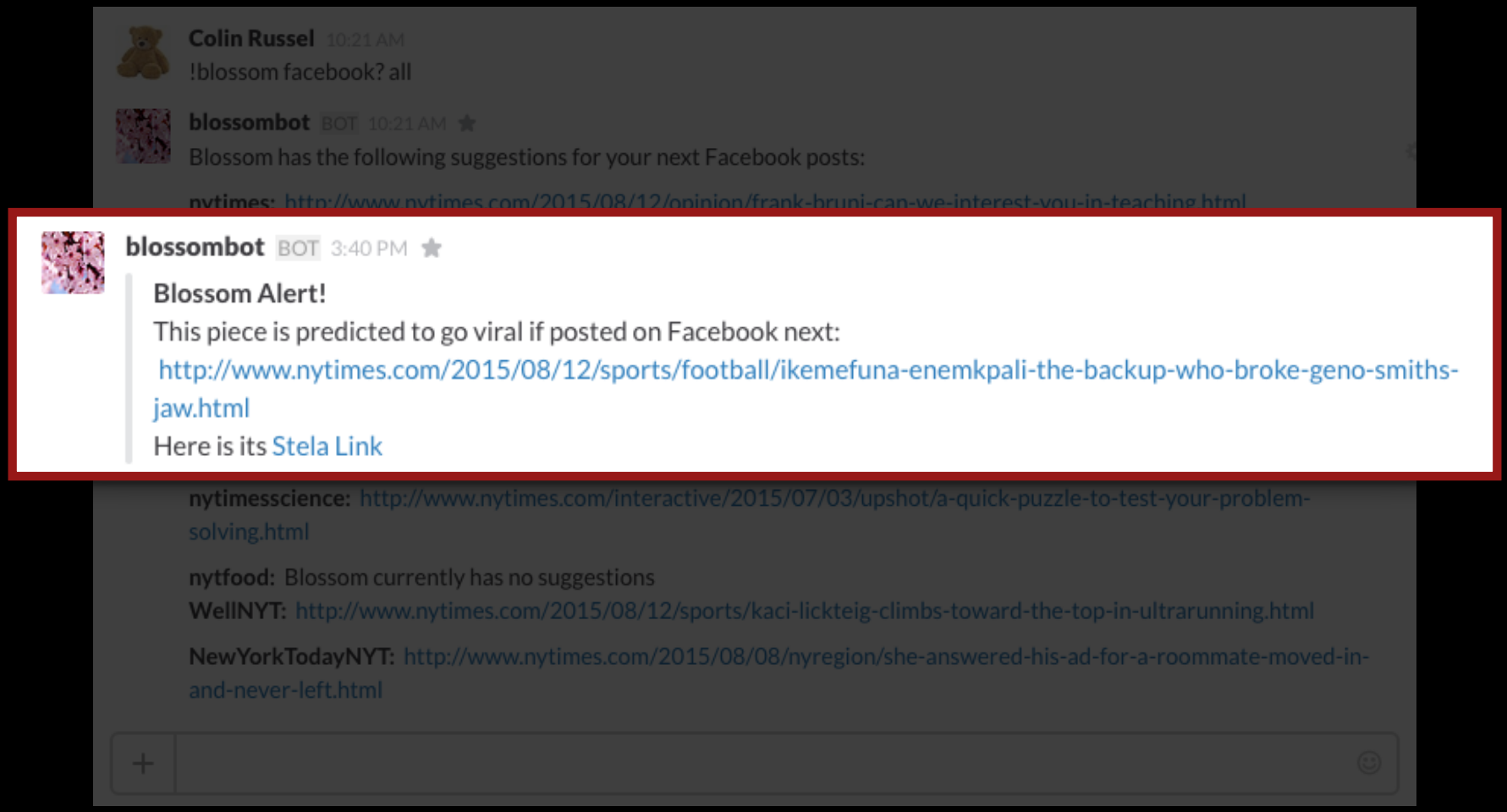
prescriptive modeling, e.g,



prescriptive modeling, e.g,

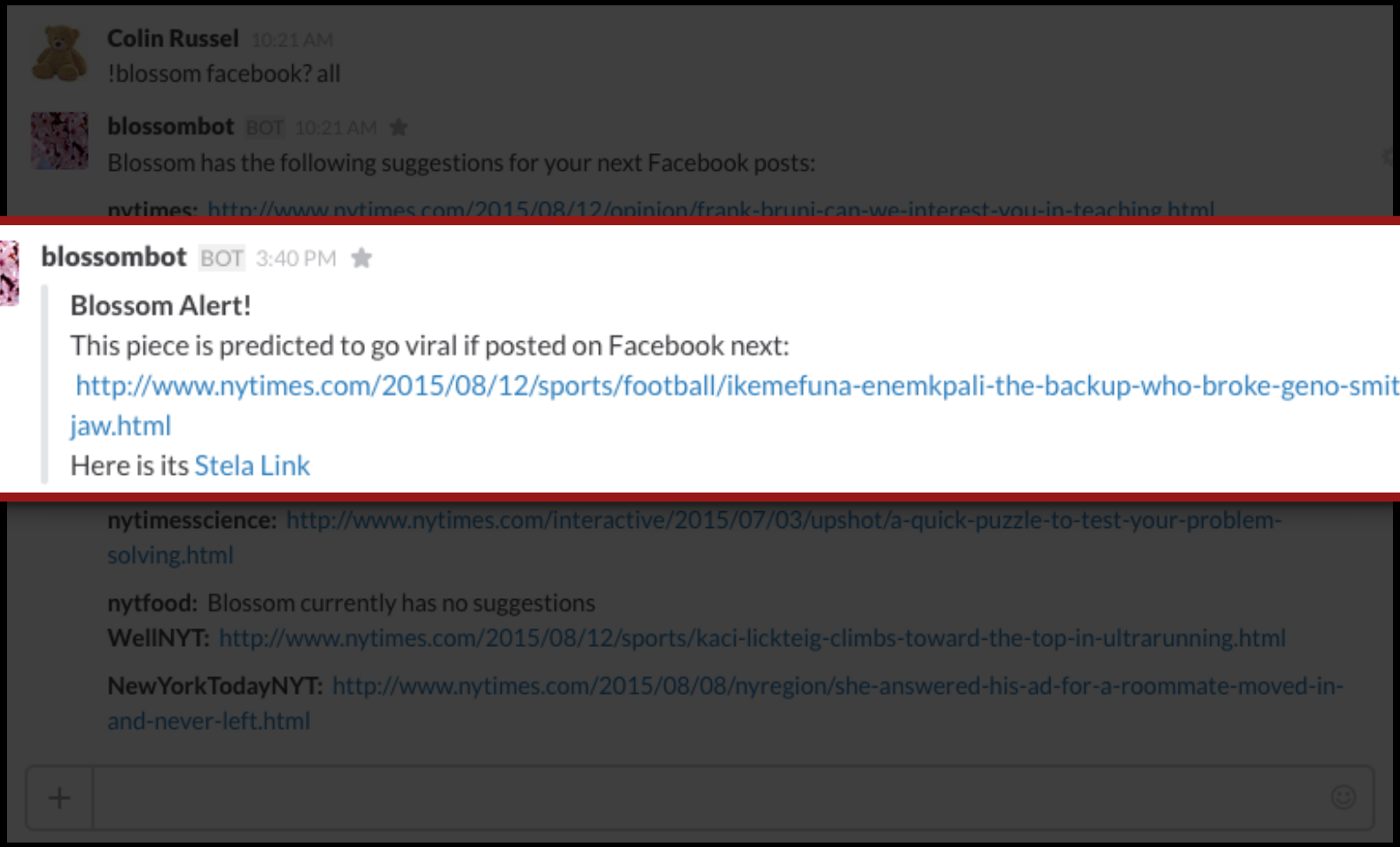


prescriptive modeling, e.g,



leverage methods which are predictive yet performant

NB: data-informed, not data-driven



predicting views/cascades: doable?

Optimizing Web Traffic via the Media Scheduling Problem

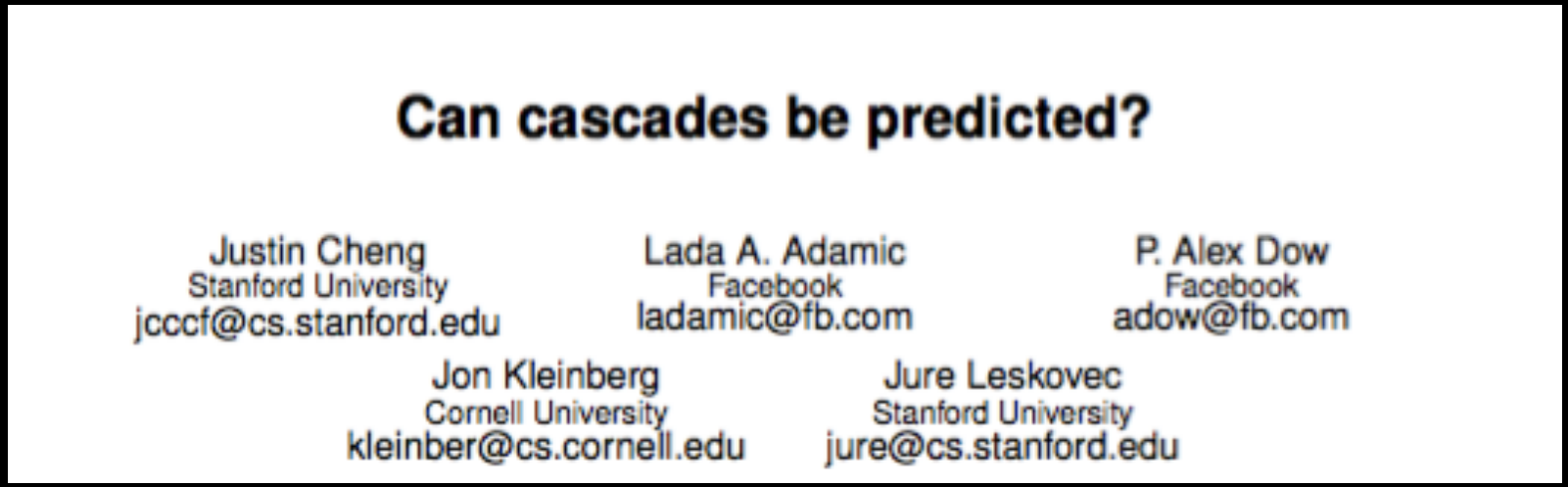
Lars Backstrom*
Dept. of Computer Science
Cornell University
Ithaca, NY 14853.
lars@cs.cornell.edu

Jon Kleinberg†
Dept. of Computer Science
Cornell University
Ithaca, NY 14853.
kleinber@cs.cornell.edu

Ravi Kumar
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089.
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

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

WWW 14: FB shares

predicting views/cascades: features?

Content Features	
$score_{food/nature/...}$	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
$liwc_{pos/neg/soc}$	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 kth 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_0	Age of the original poster, if a user
$gender_0$	Gender of the original poster, if a user
fb_age_0	Time since the original poster registered on Facebook, if a user
$activity_0$	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 kth reshare was posted
$pages_k$	Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} \mathbb{1}\{v_i \text{ is a page}\}$
$friends_k^{avg/90p}$	Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fans_k^{avg/90p}$	Average or 90th percentile fan count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg(v_i)\mathbb{1}\{v_i \text{ is a page}\}$
$subscribers_k^{avg/90p}$	Average or 90th percentile subscriber count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{subscriber}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fb\_ages_k^{avg/90p}$	Average or 90th percentile time since the first k resharers registered on Facebook, or $\frac{1}{k}\sum_{i=1}^{k} fb\_age_i$
$activities_k^{avg/90p}$	Average number of days the first k resharers were active in July, or $\frac{1}{k}\sum_{i=1}^{k} activity_i$
$ages_k^{avg/90p}$	Average age of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} age_i$
$female_k$	Number of female users among the first k resharers, 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 ith resharer (or out-degree of v_i on $G=(V,E)$)
$outdeg(v_i')$	Out-degree of the ith reshare on the induced subgraph $G'=(V',E')$ of the first k resharers and the root
$outdeg(\hat{v}_i)$	Out-degree of the ith reshare on the reshare graph $\hat{G}=(\hat{V},\hat{E})$ of the first k reshares
$orig\_connections_k$	Number of first k resharers who are friends with, or fans of the root, or $\lvert\{v_i \mid (v_0,v_i)\in E,\ 1\le i\le k\}\rvert$
$border\_nodes_k$	Total number of users or pages reachable from the first k resharers and the root, or $\lvert\{v_i \mid (v_i,v_j)\in E,\ 0\le i,j\le k\}\rvert$
$border\_edges_k$	Total number of first-degree connections of the first k resharers and the root, or $\lvert\{(v_i,v_j) \mid (v_i,v_j)\in E,\ 0\le i,j\le k\}\rvert$
$subgraph_k'$	Number of edges on the induced subgraph of the first k resharers and the root, or $\lvert\{(v_i,v_j) \mid (v_i,v_j)\in E',\ 0\le i,j\le k\}\rvert$
$depth_k'$	Change in tree depth of the first k reshares, or $\min_\beta \sum_{i=1}^{k}(depth_i-\beta i)^2$
$depths_k^{avg/90p}$	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 ith 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 kth reshare was posted, per unit time, or $\frac{views_{0,\,k}}{time_k}$
$views_{1..k-1,\,k}'$	Number of users who saw the first $k-1$ reshares, until the kth reshare was posted, per unit time, or $\frac{views_{1..k-1,\,k}}{time_k}$

predicting views/cascades: features?

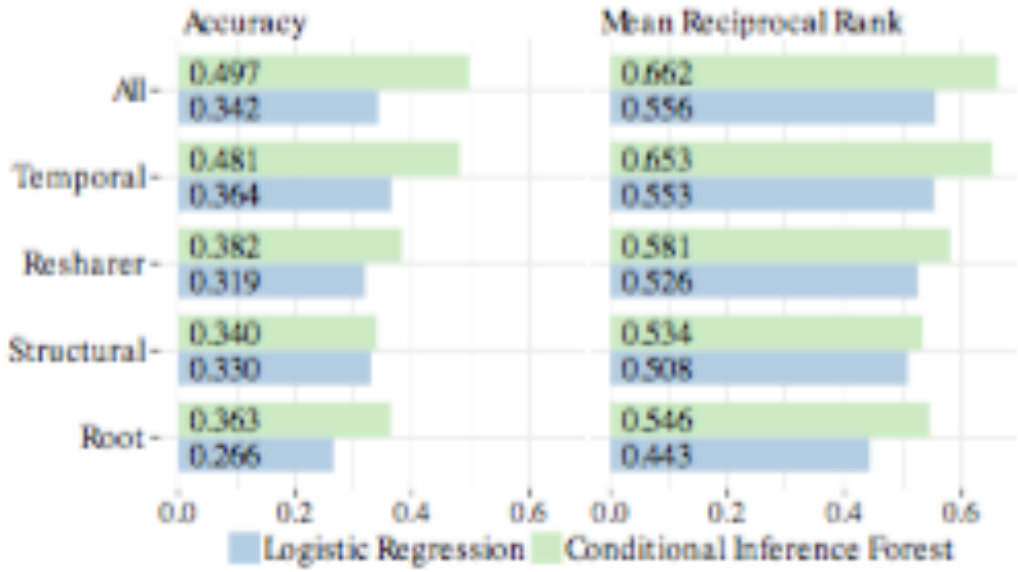


Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

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

Jake M. Hofman
Microsoft Research
641 6th Ave, Floor 7
New York, NY
jmh@microsoft.com

Amit Sharma
Microsoft Research
amshar@microsoft.com

Ashton Anderson
Microsoft Research
ashton@microsoft.com

Duncan J. Watts
Microsoft Research
duncan@microsoft.com

WWW 16: TWIT RT's

predicting views/cascades: doable?

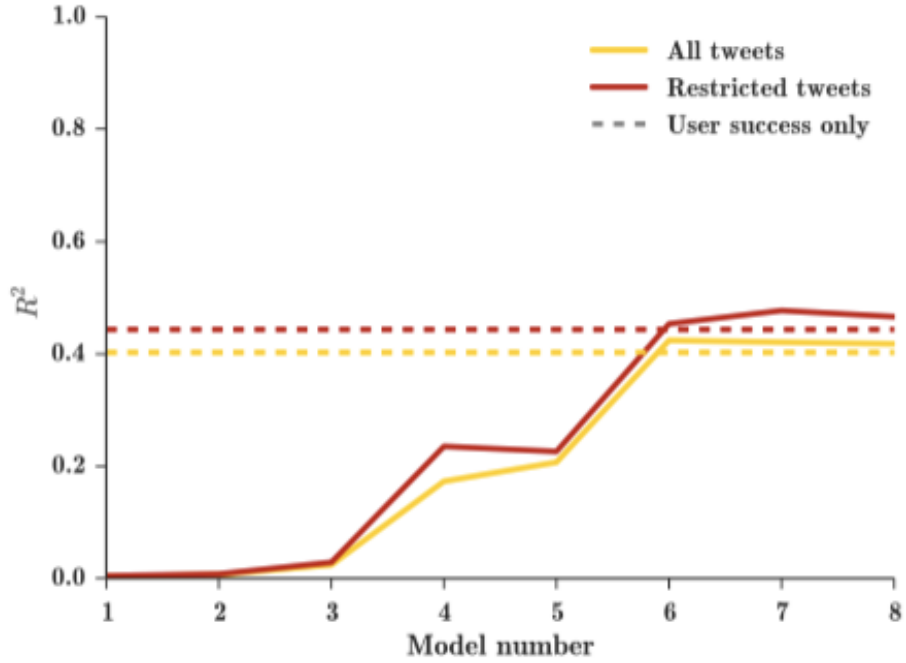
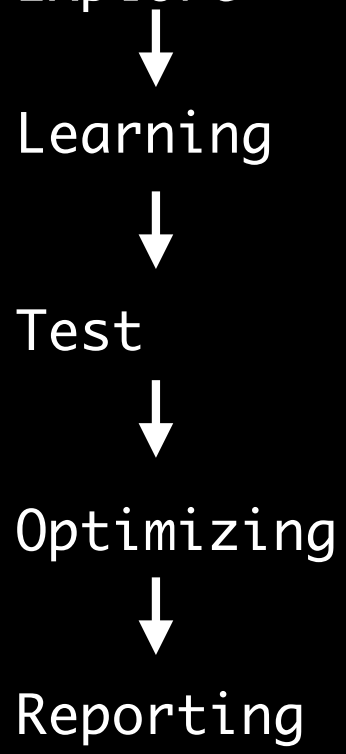


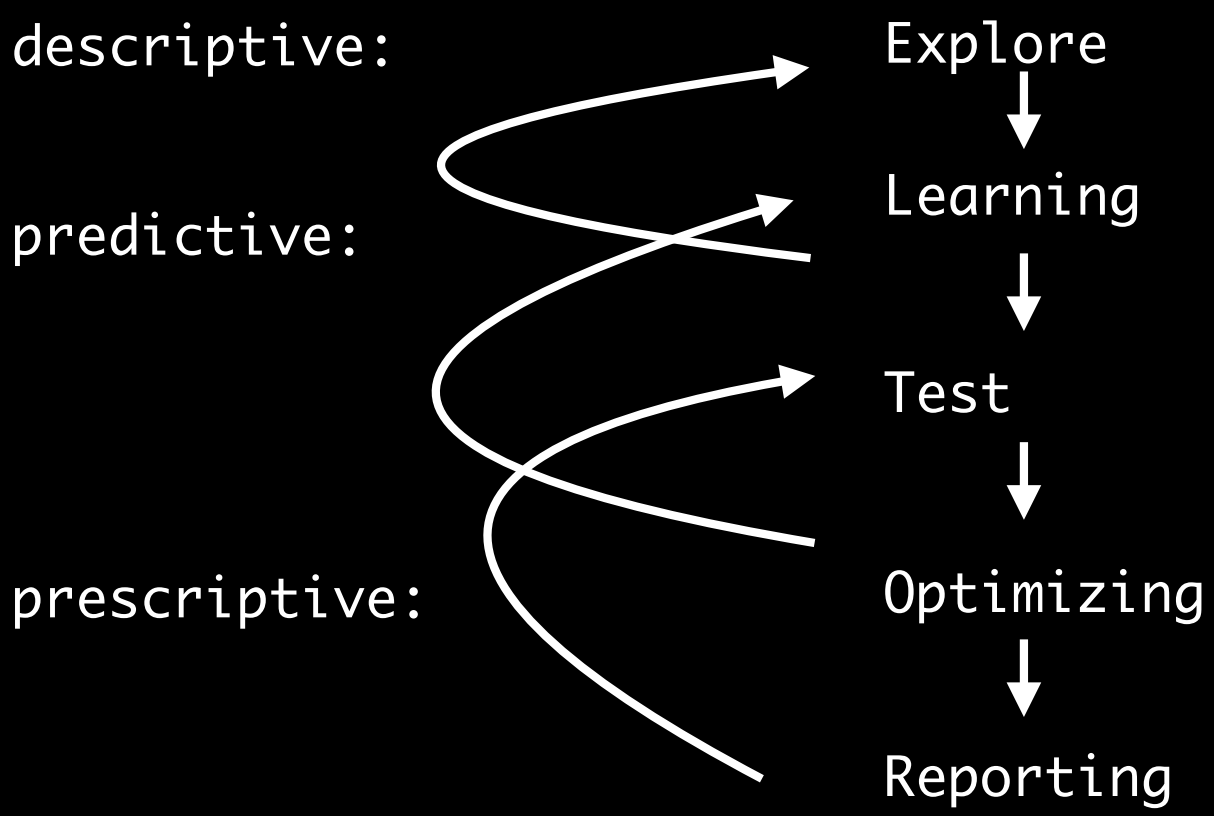
Figure 4: Prediction results for models using different subsets of features. R^2 increases as we add more features, but only up to a limit. Even a model with all features explains less than half of the variance in cascade sizes.

descriptive:

predictive:

prescriptive:




descriptive:
predictive:
prescriptive:
Explore
Learning
Test
Optimizing
Reporting

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