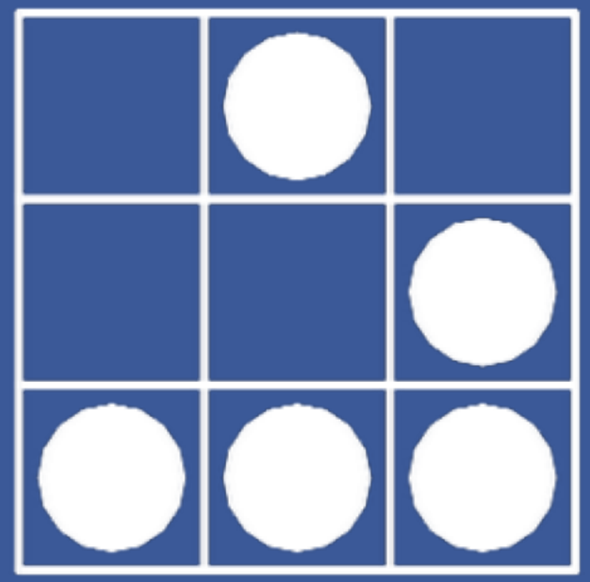
facebook

Putting the Magic in Data Science



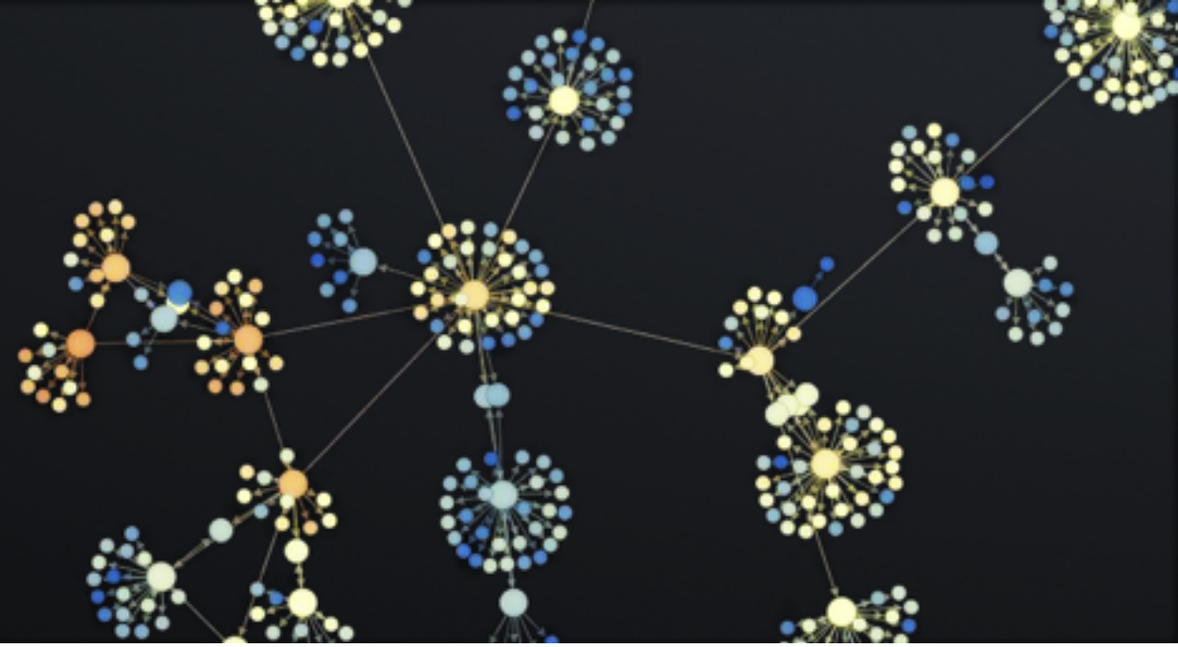
11/04/2014

Sean J. Taylor

QCon SF

https://research.facebook.com/datascience

Core Data Science @ Facebook



People

Lada Adamic

Mike Bailey

Eytan Bakshy

Moira Burke

Jonathan Chang

Ta Virot Chiraphadhanakul

David Choi

Sean Chu

Michael R. Corey

Mike Develin

Research Topics

Computational social science

Computer-mediated communication

Auction theory and mechanism design

Experimentation and statistical inference

Social influence and network externalities

User modeling

Machine learning

http://en.wikipedia.org/wiki/Pasteur's_quadrant

Applied and Basic research

Quest for fundamental understanding?	**Yes**	Pure basic research (**Bohr**)	Use-inspired basic research (**Pasteur**)
	No	–	Pure applied research (**Edison**)
		No	**Yes**
		Considerations of use?	

"The mission of CDS is to provide research and innovation that fundamentally increase the magnitude of Facebook's success."

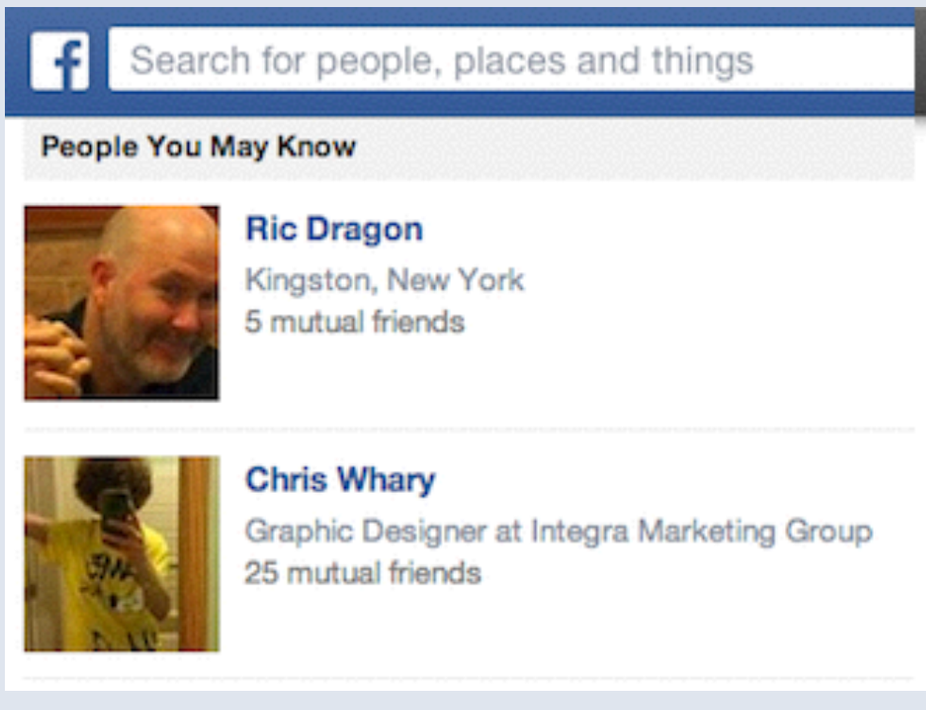
It’s not a trick, it’s an illusion.


TEXACO


Turkcell
22:46
Takvim
Kamera
Fotoğraflar
Videolar
Mesajlar
Saat
Hisse
Game Center
Hava Durumu
Borsa
Animasticiler
Facebook
Notlar
App Store
iTunes Store
Gazetelik
Ayarlar
Müzik
Pusula
Safari
Mail
Telefon


Who's in These Photos?
The photos you uploaded were grouped automatically so you can quickly label and notify friends in these pictures. (Friends can always untag themselves.)
Erik Pink
Hildy Glenn
Who is this?
Who is this?
Who is this?
Iggy Holmstrum


Customers Who Bought This Item Also Bought
Beistle Hairy Headband, Orange
2
$6.16 Prime
Dickies Unisex 40 Inch Lab Coat
399
$17.99 - $34.99
Goshman Yellow Foam Clown Noses (1 5/8")


Search for people, places and things
People You May Know
Ric Dragon
Kingston, New York
5 mutual friends
Chris Whary
Graphic Designer at Integra Marketing Group
25 mutual friends


Search...
People You May Know
Chimi Culler 2nd
Sales at C. H Robinson Worldwide
Baltimore, Maryland Area
Connect
8 shared connections
Sheena Lister 3rd
Sport Mgmt Professional


What can I help you with?
“What's the best cell phone ever”
This might answer your question:
Input interpretation
best mobile phones
by customer review average
Result
Nokia - Lumia 900 4G
Mobile Phone - Cyan (AT&T)
(5)

Any sufficiently advanced technology is indistinguishable from magic.

— Arthur C. Clarke

Fred Benenson
@fredbenenson

Following

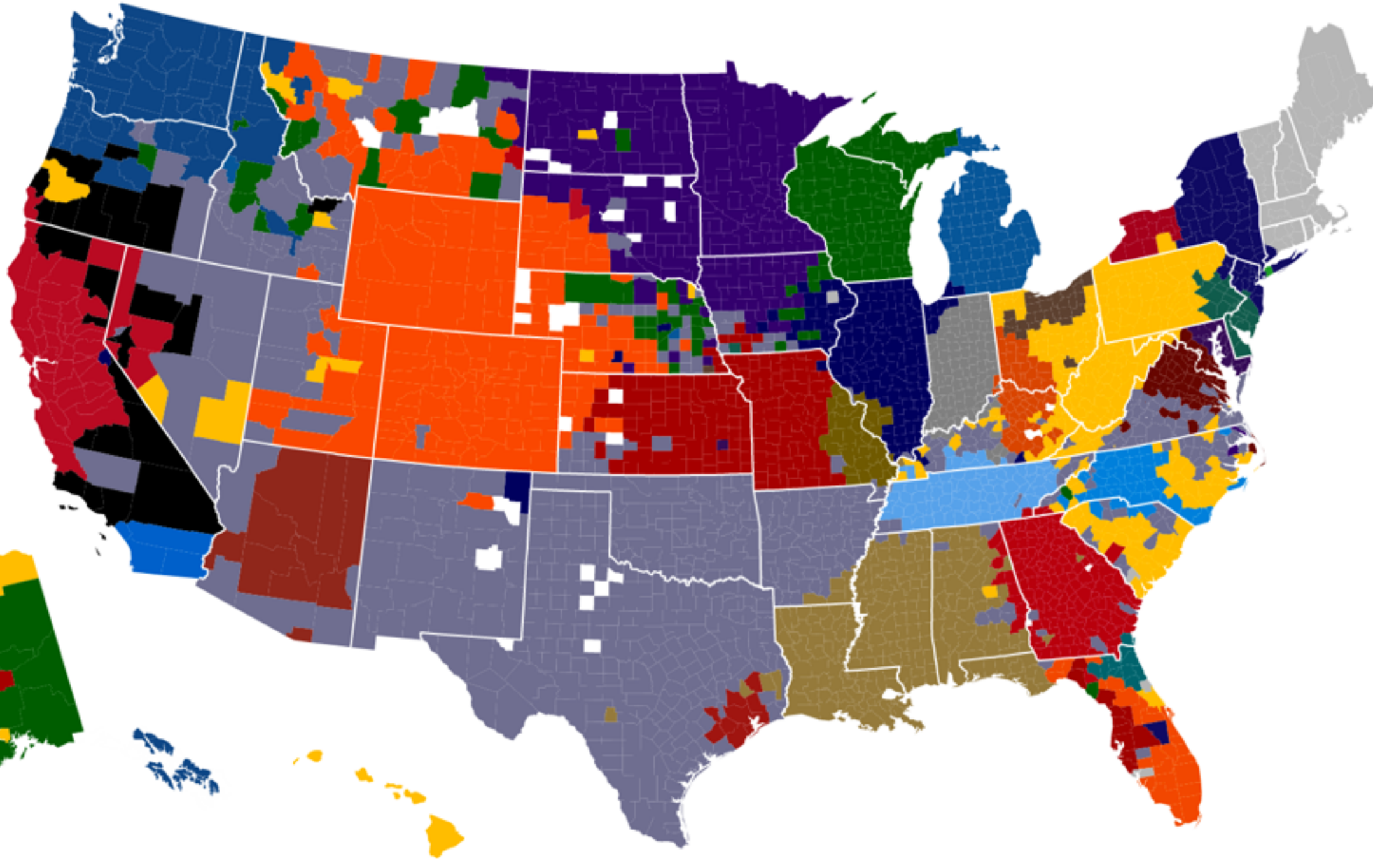
IMHO the majority of data work boils down to 3 things:

1. Counting stuff
2. Figuring out the denominator
3. The reproducibility of 1 & 2

RETWEETS 32 FAVORITES 28

12:33 PM - 21 Aug 2013

ARI
ATL
BAL
BUF
CAR
CHI
CIN
CLE
DAL
DEN
DET
GB
HOU
IND
JAC
KC
MIA
MIN
NE
NO
NYG
NYJ
OAK
PHI
PIT
SD
SEA
SF
STL
TB
TEN
WAS

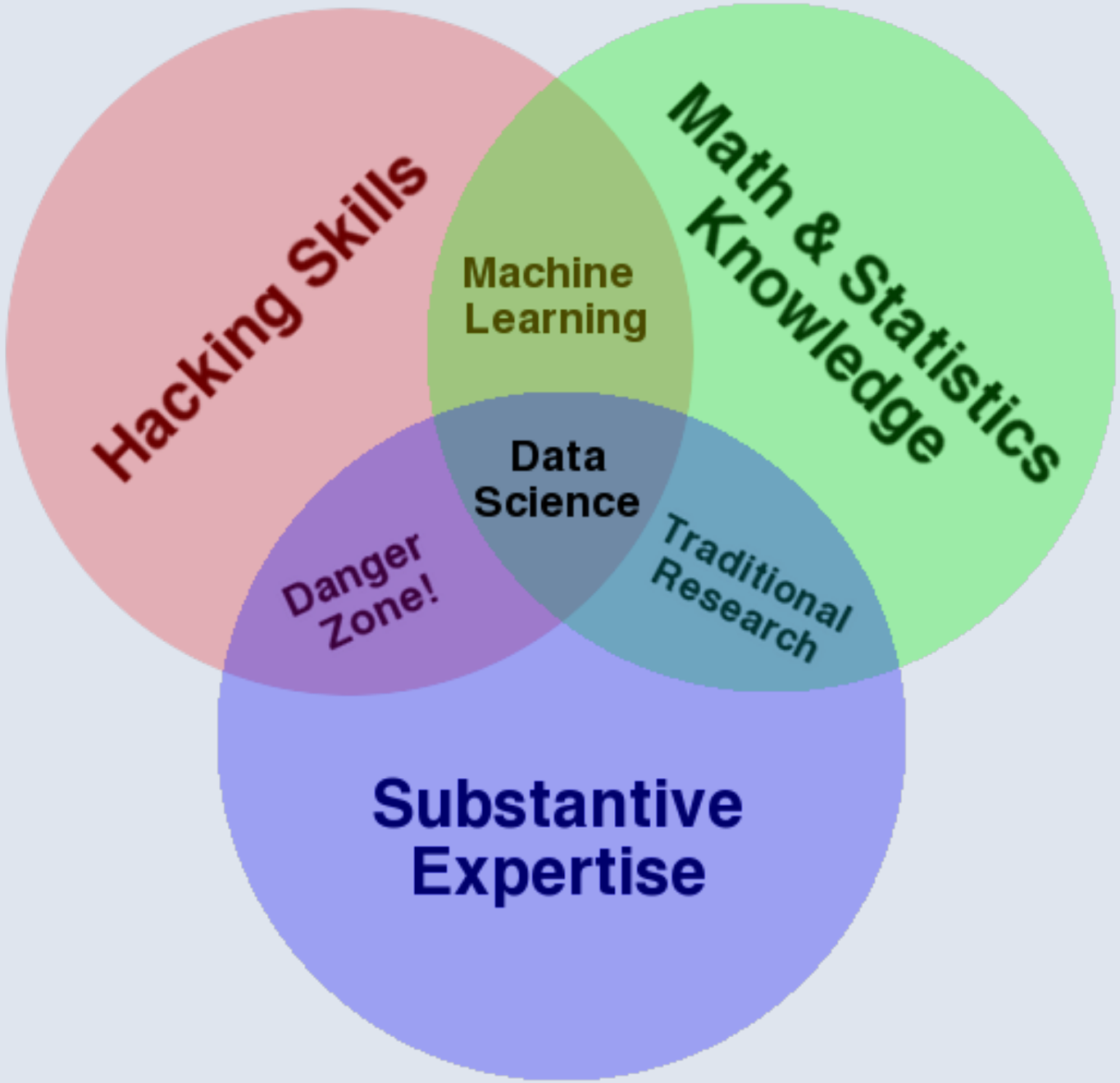



WE ARE
TERRIBLE
MAGICIANS
—MIKE VERNAL

1. create technology:
people who are not experts can use it easily with little difficulty and trust the output

2. make it "sufficiently advanced"

The Data Science Venn Diagram
Drew Conway



Basic Research	Maybe someday, someone can use this.
Applied Research	I might be able to use this.
Working Prototype	I can use this (sometimes).
Quality Code	Software engineers can use this.
Tool or Service	People can use this.

People can use it → People want to use it

Data Science Impact = Value * (Num People) * (Frequency of Use)

Very difficult to demand that people use new tech — must make a compelling value proposition for people and educate them.

What can data do?

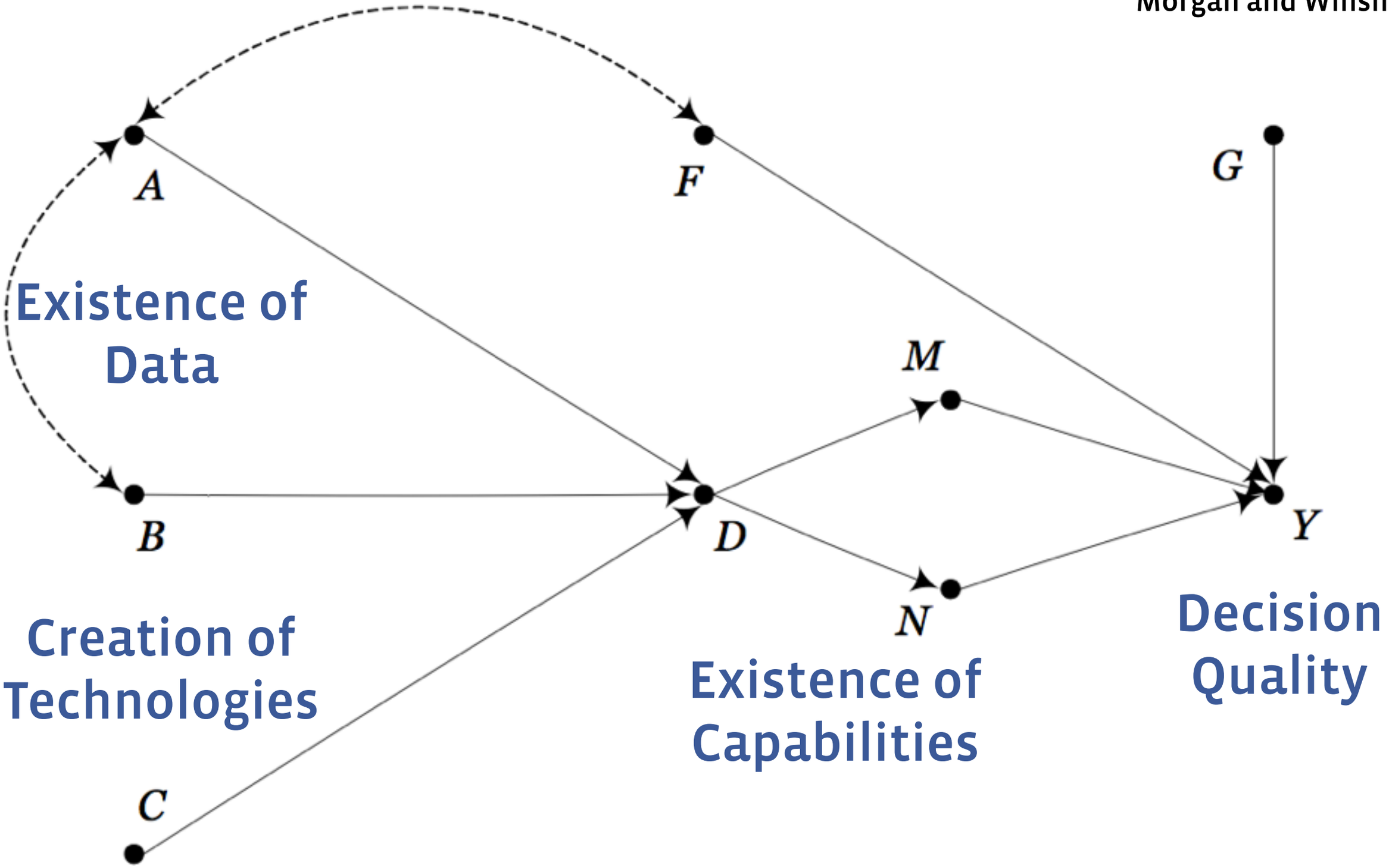
Data can't do anything.

People do things with data.
(usually they make decisions)

The Last Mile Problem

It works for you. Can you get people to use it?

Without considering this last step, all subsequent steps are useless.


A
F
G
M
Existence of Data
B
D
Y
N
Creation of Technologies
Existence of Capabilities
Decision Quality
C

Magical + Effective Data Science Tools

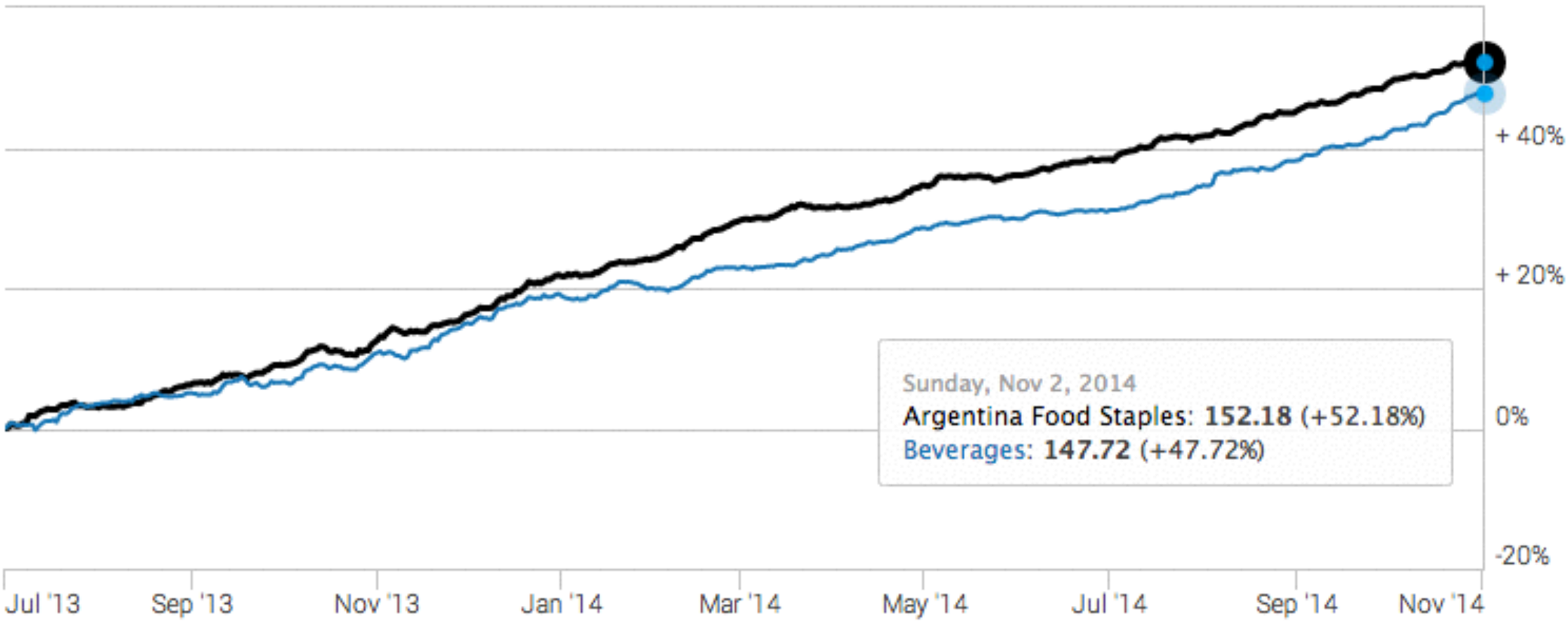
- Planout: language for expressing / deploying experimental designs
- Deltoid: analyzing the results of experiments
- ClustR: generic document clustering
- Prophet: completely automatic forecasting procedure
- Crystal Ball: large scale, interpretable regression models
- Hive / Presto / Scuba: SQL engines for different problems

Outline

1. Sources of Magic

2. Solving the Last Mile

Tricks: Sources of Data Science Magic


+ 40%
+ 20%
0%
-20%
Sunday, Nov 2, 2014
Argentina Food Staples: 152.18 (+52.18%)
Beverages: 147.72 (+47.72%)
Jul '13
Sep '13
Nov '13
Jan '14
Mar '14
May '14
Jul '14
Sep '14
Nov '14

Argentina

http://premise.com

← Beverages (15.4%)

- 1882 Fernet (1,239)
- 5 Hispanos Coffee (ground) (893)
- Americano Gancia (1,352)
- Arlistan Instant Coffee (2,018)
- Baggio Apple Juice (491)
- Brahma Beer (1,840)
- Branca Fernet (1,811)
- Cepita Apple Juice (1,608)
- Cepita del Valle Orange Juice (705)
- Coca Light Diet Coke (1,572)
- Coca-Cola Classic (Normal) (1,134)
- Del Valle Cider (336)
- Eco de Los Andes Bottled Water (Sparkling) (632)
- Glaciar Bottled Water (1,698)

1882 Fernet

relative spec weight 0.98%

Observations

51	165	489	1,239
7 day	30 day	90 day	lifetime

Date	Size	Norm. price	Lat/long	Place
10/24/2014 11:13 AM	450 ml	0.0676 ARS / ml	-31.413, -64.230	marian
10/23/2014 1:59 PM	750 g	0.0692 ARS / g	-31.464, -64.214	super
10/23/2014 12:54 PM	750 ml	0.0687 ARS / ml	-34.563, -58.459	COTO
10/23/2014 12:30 PM	750 ml	0.0714 ARS / ml	-34.559, -58.459	Carrefo
10/23/2014 12:09 PM	750 ml	0.0753 ARS / ml	-32.885, -68.849	careefo
10/23/2014 12:07 PM	450 ml	0.0768 ARS / ml	-31.456, -64.169	Carrefo
10/23/2014 10:53 AM	750 ml	0.0753 ARS / ml	-31.409, -64.170	hiper L
10/23/2014 10:37 AM	750 ml	0.0687 ARS / ml	-34.554, -58.453	Carrefo

Argentina

http://premise.com

← **Beverages** (15.4%)

- 1882 Fernet (1,239)
- 5 Hispanos Coffee (ground) (893)
- Americano Gancia (1,352)
- Arlistan Instant Coffee (2,018)
- Baggio Apple Juice (491)
- Brahma Beer (1,840)
- Branca Fernet (1,811)
- Cepita Apple Juice (1,608)
- Cepita del Valle Orange Juice (705)
- Coca Light Diet Coke (1,572)
- Coca-Cola Classic (Normal) (1,134)
- Del Valle Cider (336)
- Eco de Los Andes Bottled Water (Sparkling) (632)

1882 Fernet

relative spec weight 0.98%

Observations

7 day	30 day	90 day	lifetime
51	165	489	1,239

MAGIC The Gathering®
Of Data

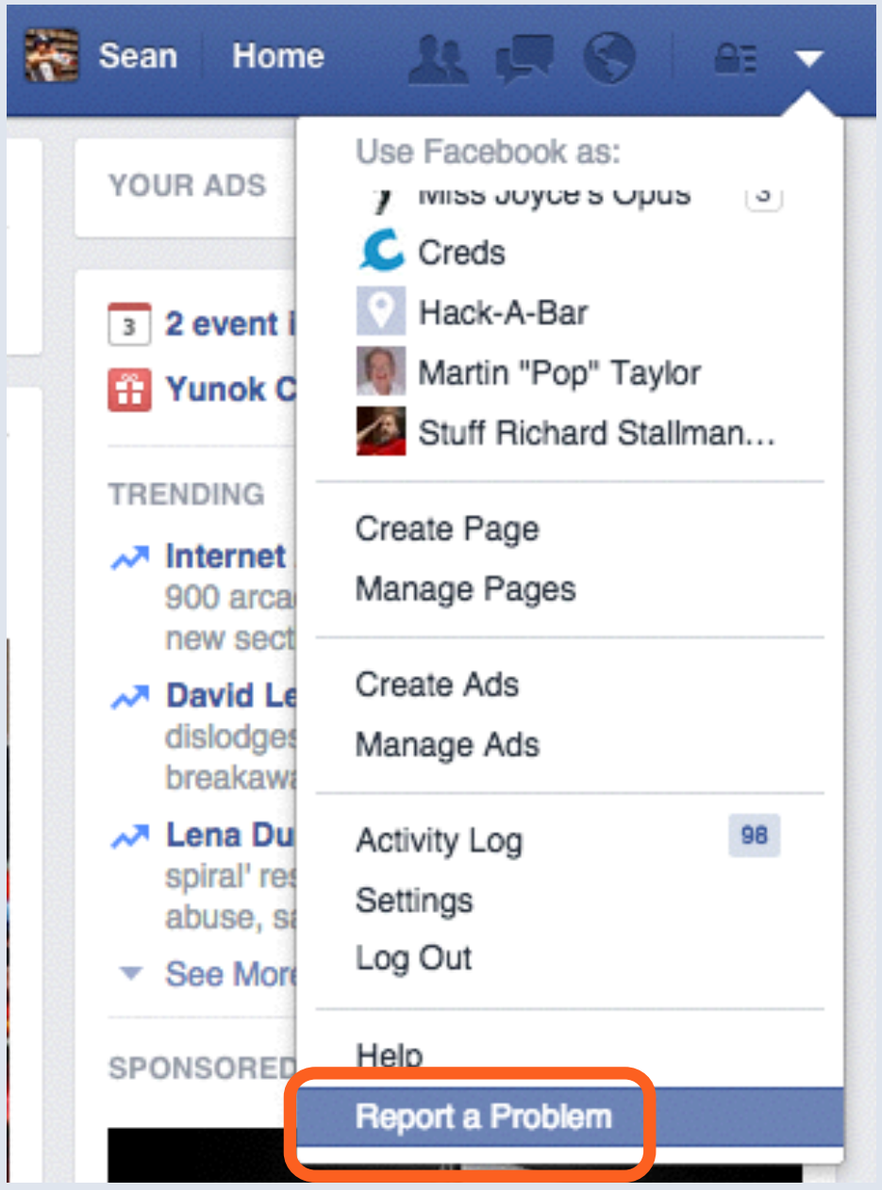
Trick 1: invest in data collection

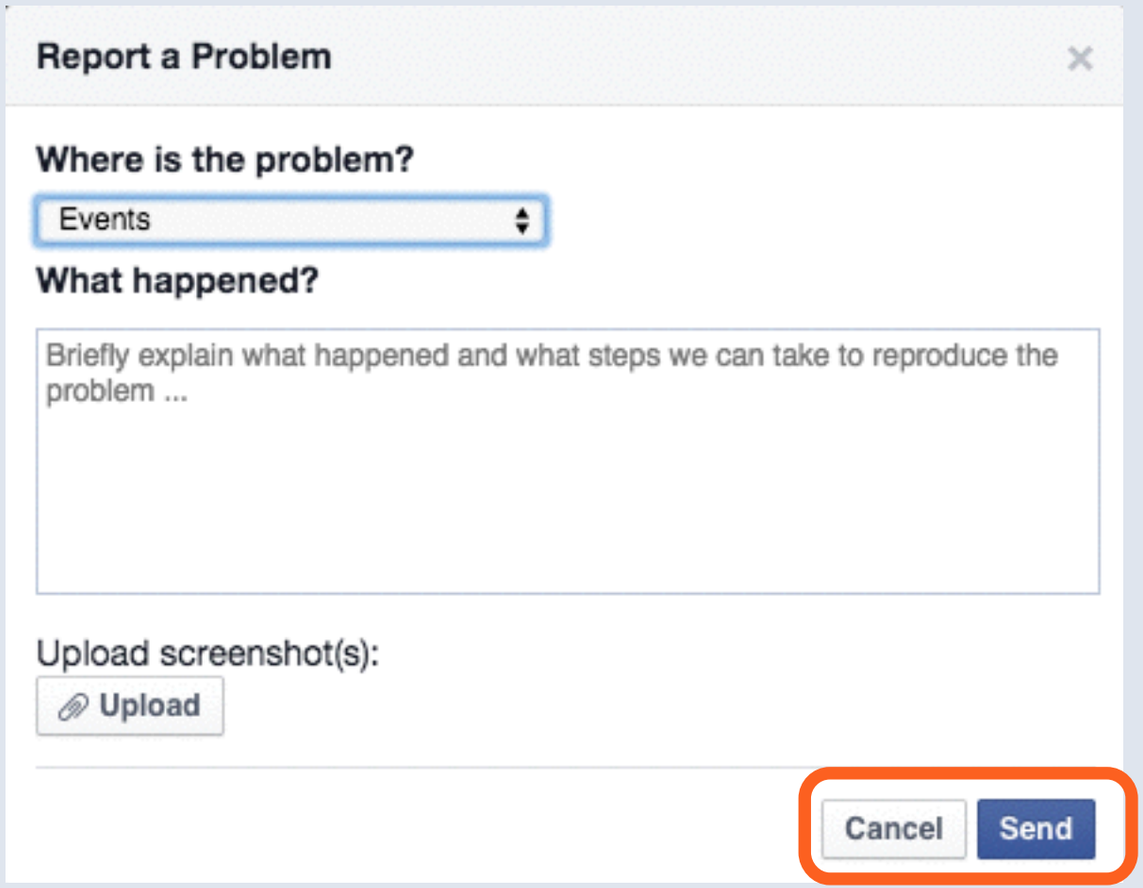
Novel sources of data are magic.

Making your own quality data is better than being a data alchemist.

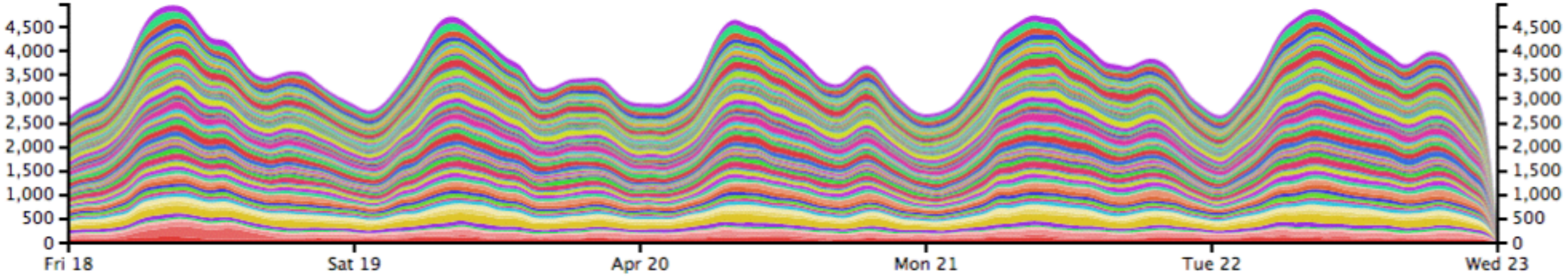
>

[PDF] **Twitter** mood **predicts the stock market**. - arXiv
arxiv.org/pdf/1010.3003& ▾ arXiv ▾
by J Bollen - 2010 - Cited by 605 - Related articles
Oct 14, 2010 - Index Terms—**stock market** prediction — **twitter** — social media (blogs, **Twitter** feeds, etc) to **predict** changes in vario





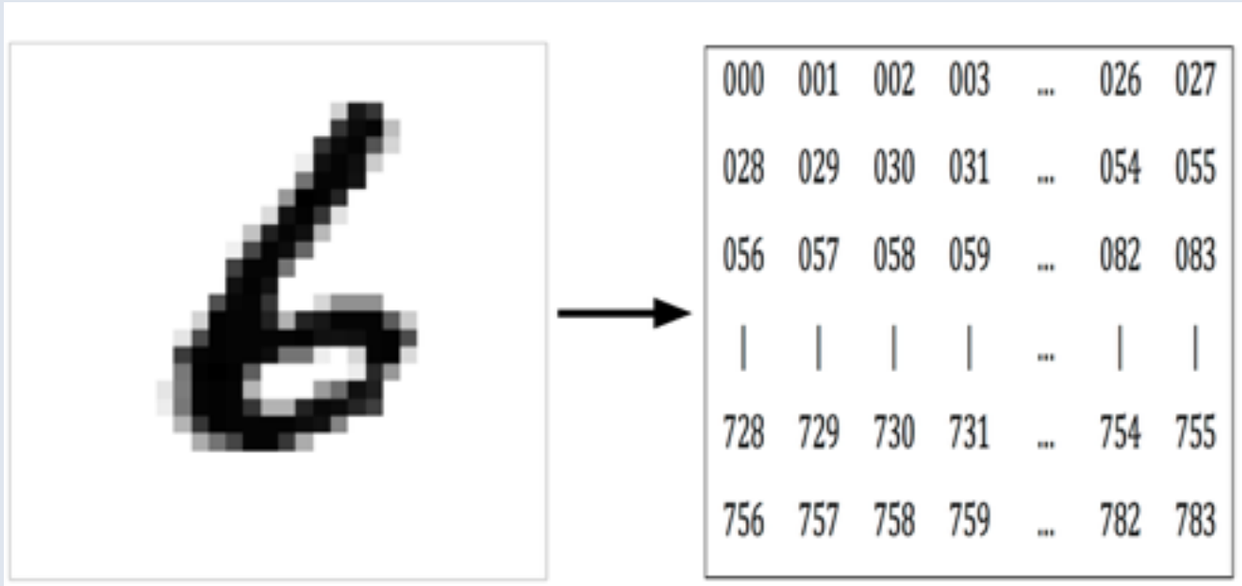
Let's say you have a billion users... and you want to listen to them all



0	1	2	3	4	5	6
[Top Docs]	[Top Docs]	[Top Docs]	[Top Docs]	[Top Docs]	[Top Docs]	[Top Docs]
[Sampled Docs]	[Sampled Docs]	[Sampled Docs]	[Sampled Docs]	[Sampled Docs]	[Sampled Docs]	[Sampled Docs]
# Docs: 10724	# Docs: 5497	# Docs: 9885	# Docs: 5395	# Docs: 3712	# Docs: 7920	# Docs: 15995
photo	**share**	**photo**	**photo**	**let**	**name**	**message**
upload	photo	**download**	friends#photo	**won**	change#name	send
photo#upload	photo#share	**save**	page#photo	let#won	**change**	message#send
file	instagram	photo#save	album	wont	friends#name	friends#message
profile#upload	post#share	download#photo	album#photo	know#let	name#page	message#sent
pictures#upload	link	mobile	cover	trying	account#name	sent
problem#upload	page#share	download#video	help#photo	let#message	name#profile	message#receive
camera	friends#share	download#pictures	photo#send	friends#let	help#name	receive
phone#upload	see#share	gallery	photo#thank	let#see	name#use	help#message
page#upload	instagram#photo	application	cover#photo	let#like	like#name	message#thank
trying#upload	like#share	nn	tag	let#trying	user	friends#send
help#upload	share#video	pictures#save	photo#tag	know	name#thank	know#message
post#upload	pictures#share	install	photo#use	let#page	know#name	message#says

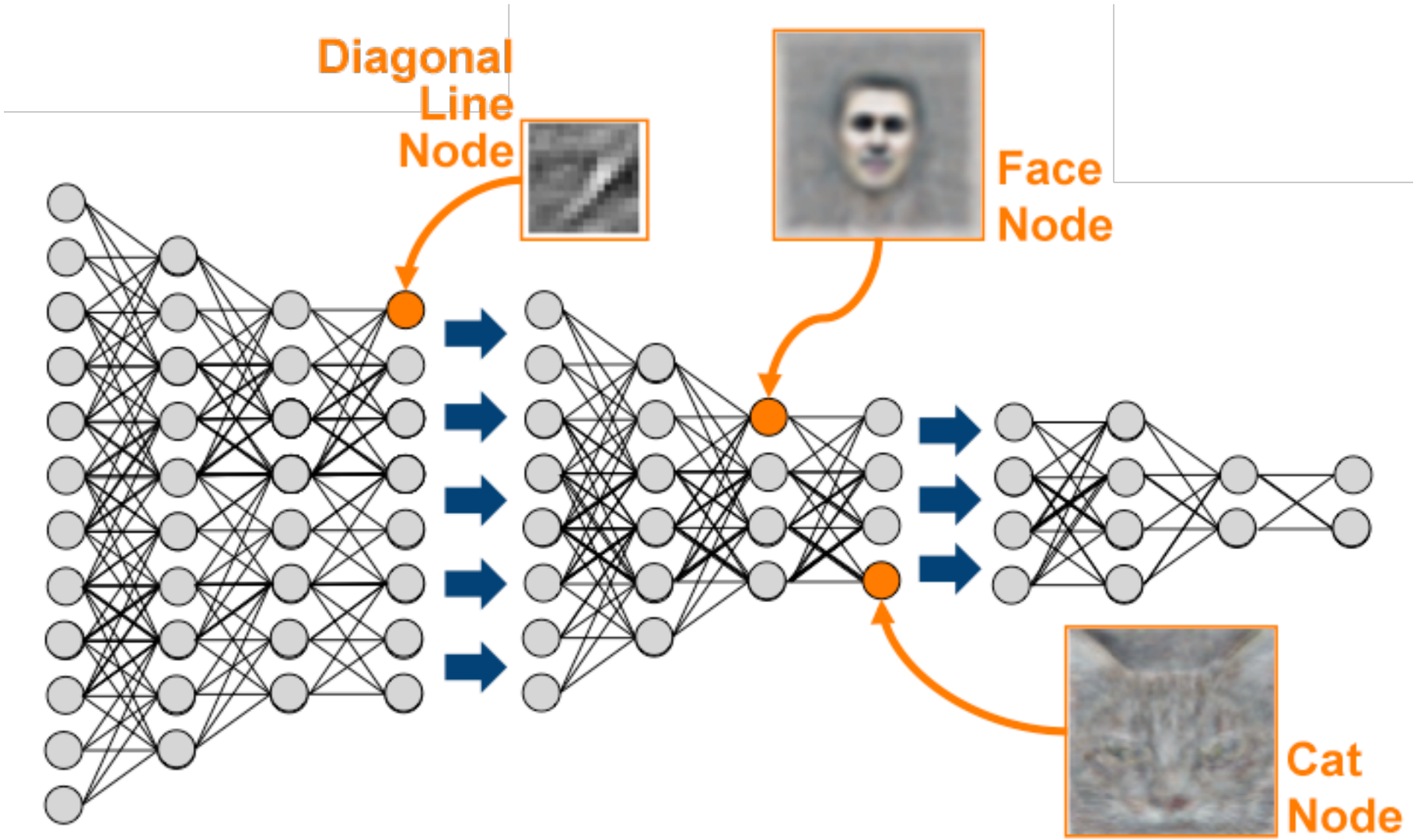
Trick 2: Dimensionality Reduction

Increasingly individual observations can be very high dimensional: text documents, images, audio.



Clustering and classification techniques can find/extract a smaller dimensional representation that retains meaning.

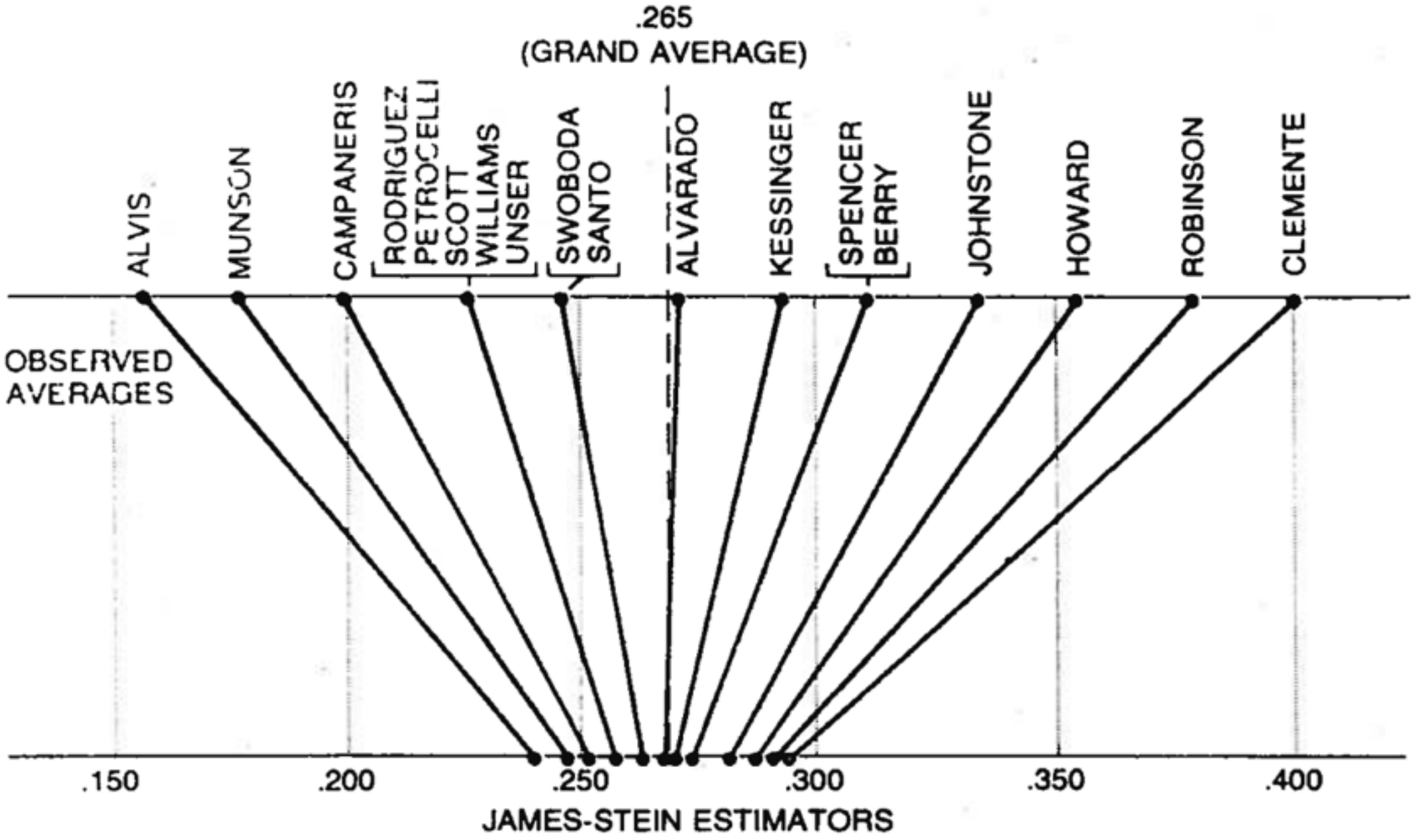
Deep Learning is just (very) fancy dimensionality reduction



Problem:

Estimate the probability of rare events or events pertaining to new objects.

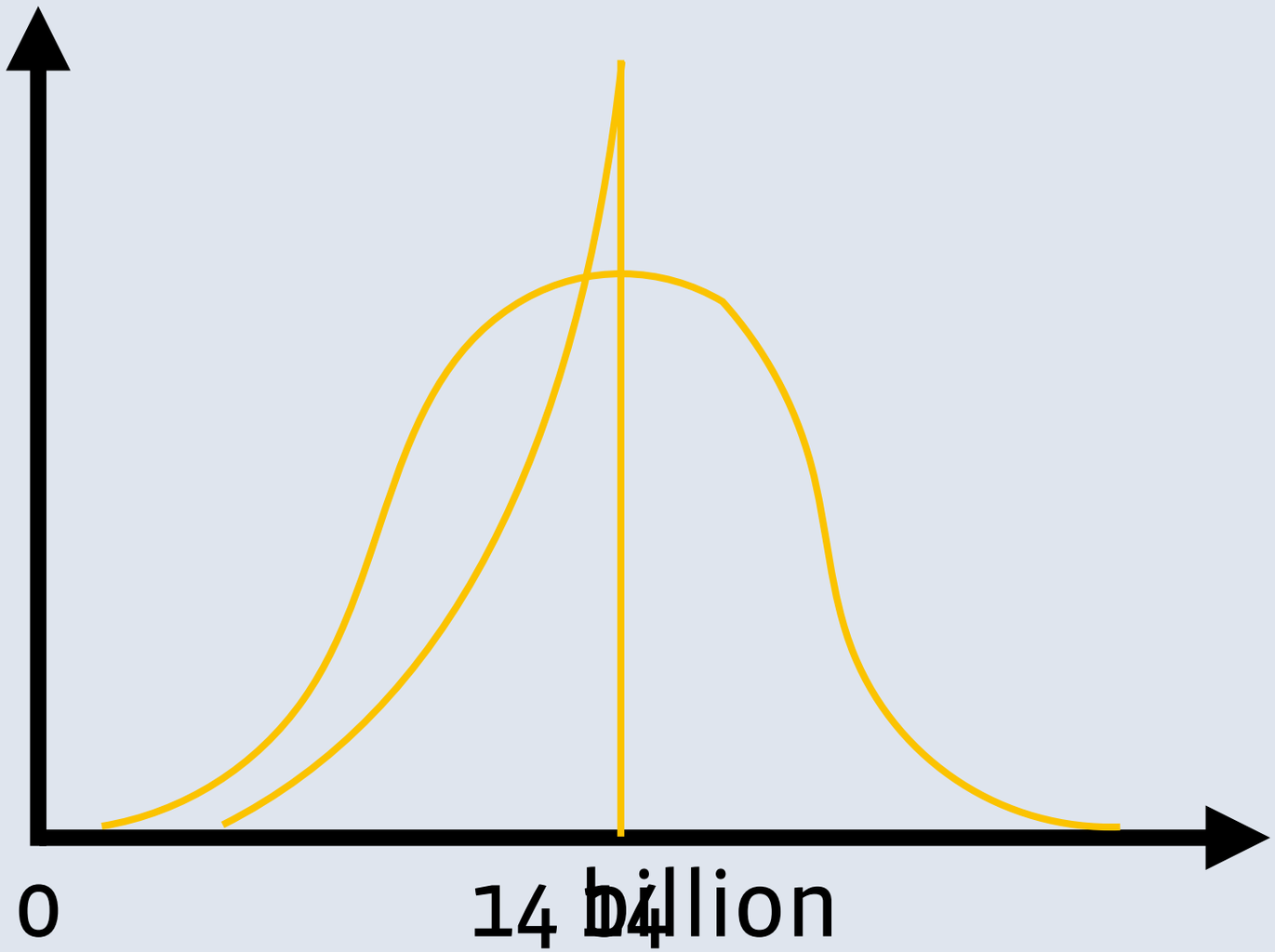
E.g. click, like, comment, share

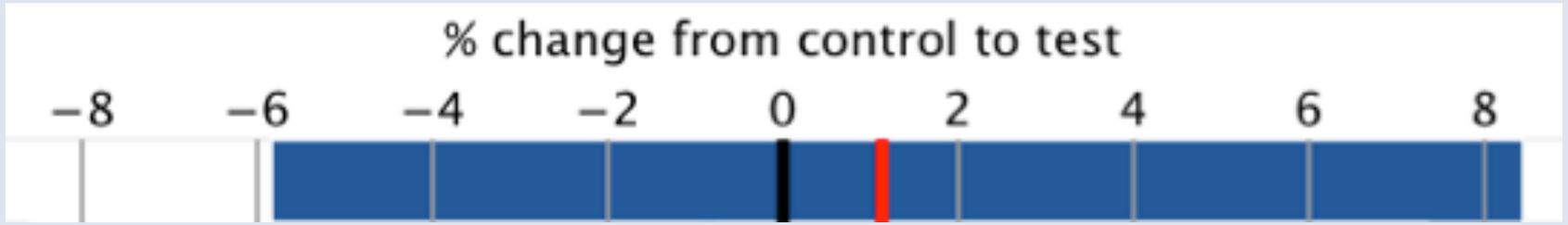


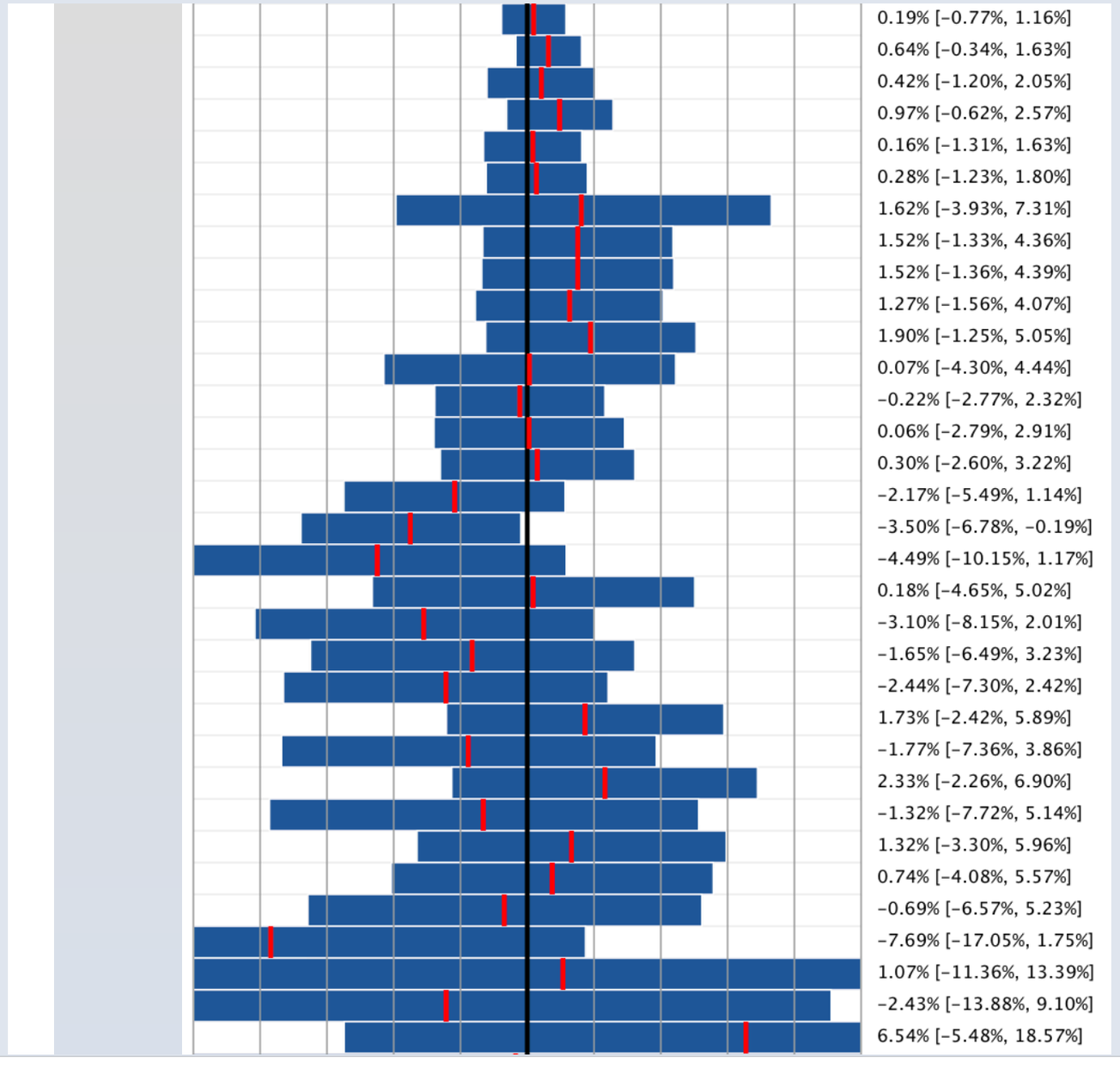
JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Trick 3: Be a (Practical) Bayesian

- If you have rare or new things you'd like to learn about, it's often hard to say much.
- But it's sometimes easy to think of cases which are similar to the one you are trying to predict.
- James-Stein estimators demonstrate that weighted averages including related observations will help improve predictions.


0
14 billion
Phillies Win
Facebook Launches
Eagles Win

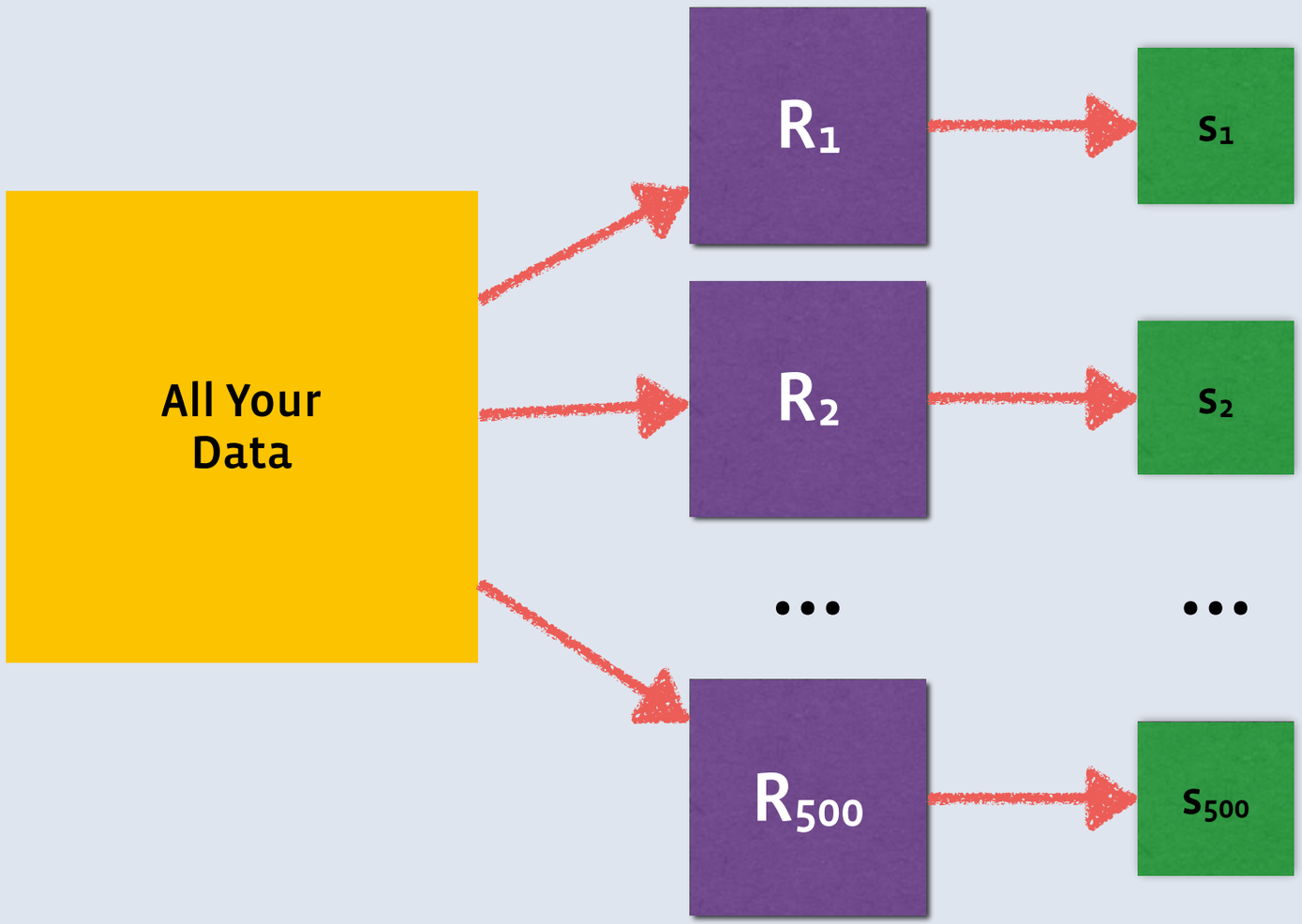

% change from control to test
−8
−6
−4
−2
0
2
4
6
8


0.19% [−0.77%, 1.16%]
0.64% [−0.34%, 1.63%]
0.42% [−1.20%, 2.05%]
0.97% [−0.62%, 2.57%]
0.16% [−1.31%, 1.63%]
0.28% [−1.23%, 1.80%]
1.62% [−3.93%, 7.31%]
1.52% [−1.33%, 4.36%]
1.52% [−1.36%, 4.39%]
1.27% [−1.56%, 4.07%]
1.90% [−1.25%, 5.05%]
0.07% [−4.30%, 4.44%]
−0.22% [−2.77%, 2.32%]
0.06% [−2.79%, 2.91%]
0.30% [−2.60%, 3.22%]
−2.17% [−5.49%, 1.14%]
−3.50% [−6.78%, −0.19%]
−4.49% [−10.15%, 1.17%]
0.18% [−4.65%, 5.02%]
−3.10% [−8.15%, 2.01%]
−1.65% [−6.49%, 3.23%]
−2.44% [−7.30%, 2.42%]
1.73% [−2.42%, 5.89%]
−1.77% [−7.36%, 3.86%]
2.33% [−2.26%, 6.90%]
−1.32% [−7.72%, 5.14%]
1.32% [−3.30%, 5.96%]
0.74% [−4.08%, 5.57%]
−0.69% [−6.57%, 5.23%]
−7.69% [−17.05%, 1.75%]
1.07% [−11.36%, 13.39%]
−2.43% [−13.88%, 9.10%]
6.54% [−5.48%, 18.57%]

Trick 4: Bootstrap all the statistics

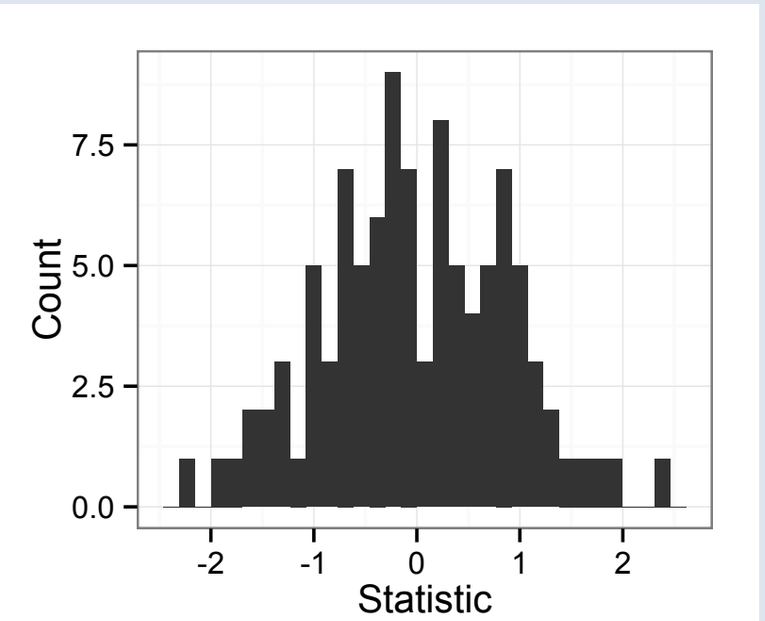
- The bootstrap allows you to get a sampling distribution over almost any statistic you can compute from your data.
- Embarrassingly parallelizable / computable online.

Bootstrapping in Practice

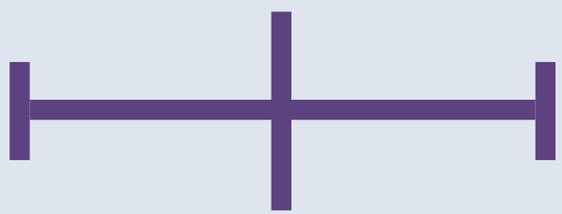


Generate random sub-samples

Compute statistics or estimate model parameters



Get a distribution over statistic of interest (usually the prediction)



- take mean
- CIs == 95% quantiles
- SEs == standard deviation

Grab bag of tricks

- Everything is linear if you use enough features.
- Matrix factorizations: NMF, PCA.
- Probabilistic data structures: LSH, min-hash.
- Exploit distributed, online algorithms as much as possible.
- “A little bit of ridge never hurts.” — Dean Eckles
- Label propagation: use data about network neighbors.
- Data reduction: create bins and analyze per-bin stats.

Last Mile of Data Science Magic

Principle 1: Reliability

"60% of the time, it works every time"

Test-driven data science

Learn how to build reliable data science systems from software engineers.

1. Write test fixtures with simulated or case-study data sets.

2. Write automated tests that check that your system works on fixtures, and add new ones when it doesn't.

3. (Bonus) Test input data to ensure it meets all assumptions.

Principle 2: Latency + Interactivity

"how many hypotheses per second are you testing/generating?"

Answer more questions

People have good intuitions and tend to search effectively given understandable tools.

First order effect of speed: **more answers per second.**

Second order effect of speed: **more questions asked.**

Deltoid: effortless experimentation

Scuba: in-memory, distributed, sampled database.

Presto: aggressive caching, distributed SQL query engine

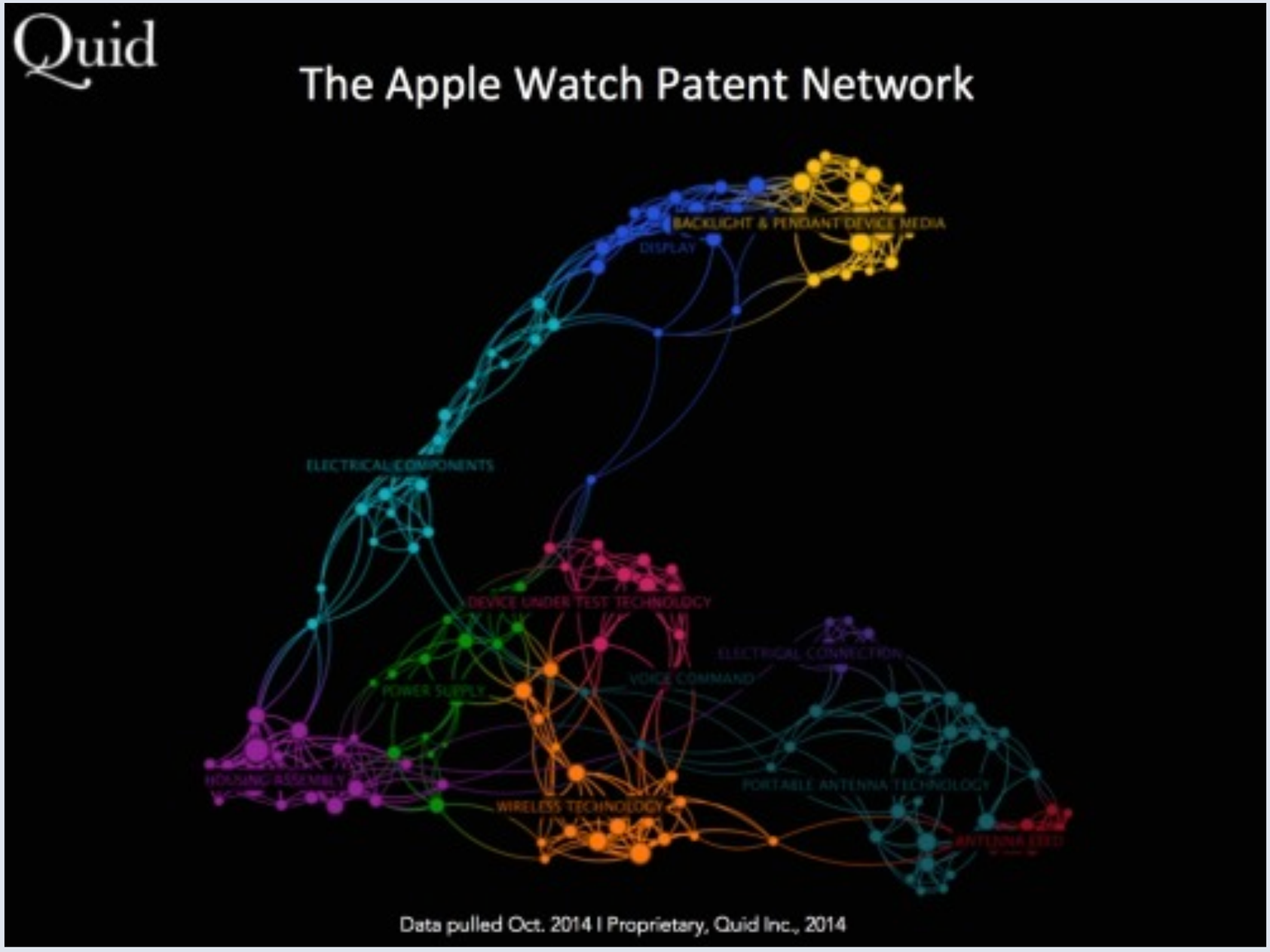
Principle 3: Simplicity + Modularity

Choose one thing to do very well

- It makes it easier to optimize your technology.
- It makes it easier for people to understand what it does.
- It makes people more likely to build around it.

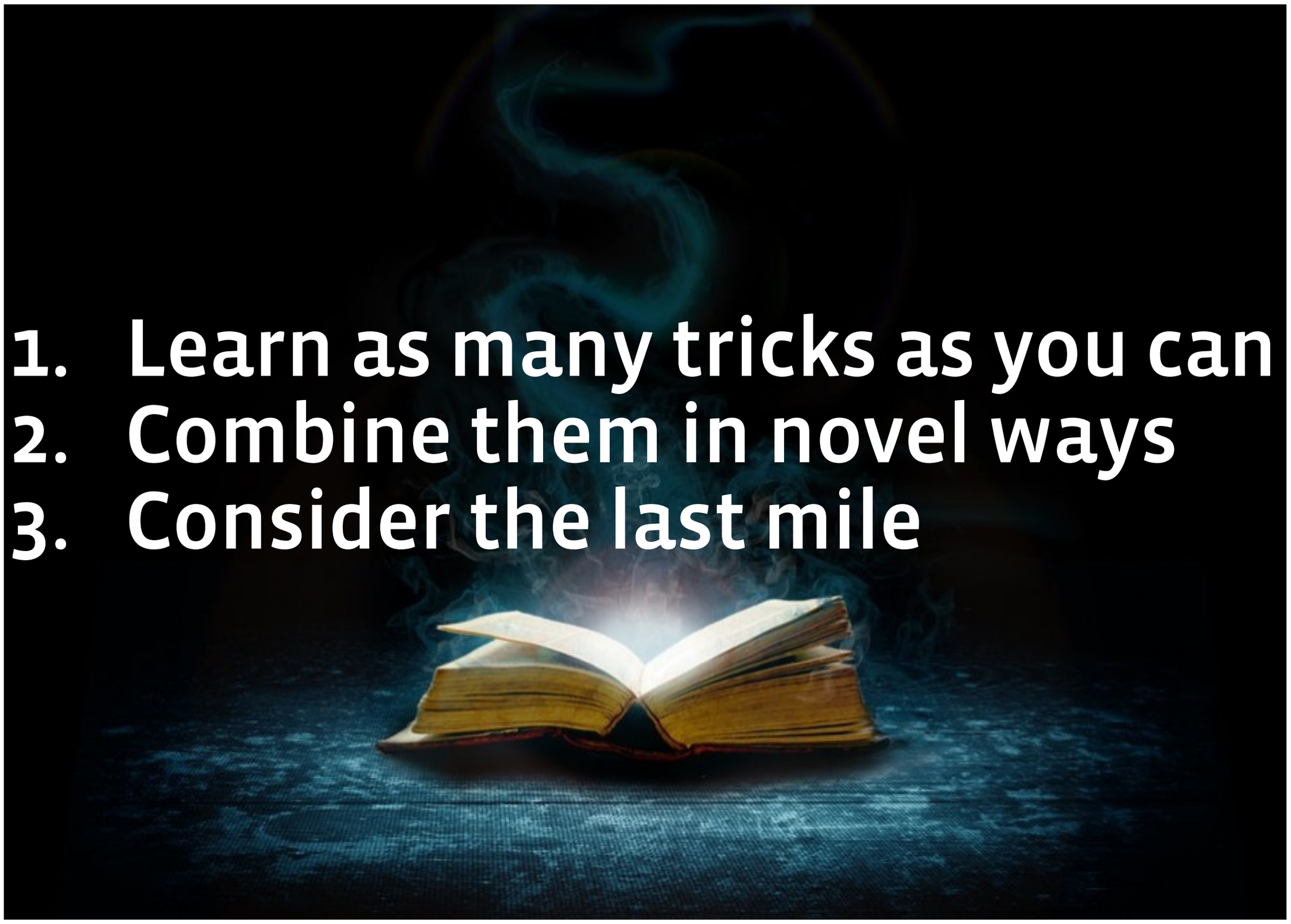
Principle 4: Unexpectedness

Show people the most interesting things



Tricks Explained

- Planout: simplicity + modularity
- Deltoid: effortless experiment analysis + bootstrap
- ClustR: dimensionality reduction + interactivity
- Prophet: everything's linear + basis expansion + new data
- Crystal Ball: everything's linear + regularization + speed
- Hive / Presto / Scuba: reliability/latency tradeoffs


1. Learn as many tricks as you can
2. Combine them in novel ways
3. Consider the last mile

sjt@fb.com
http://seanjtaylor.com

facebook