

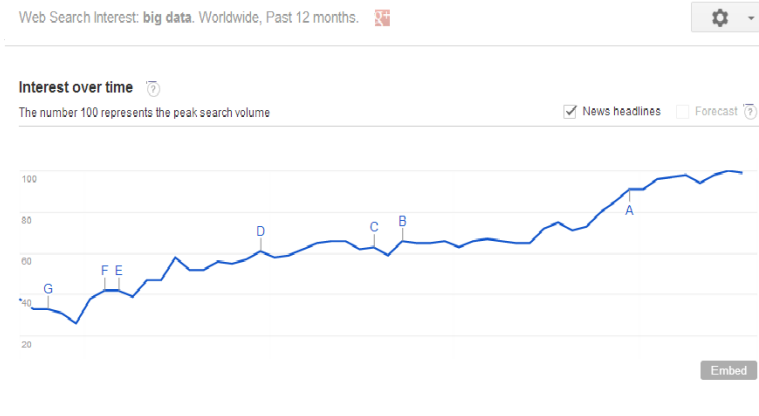
Big Data for Social Science Research: Hypes, Myths, and Realities

Jonathan Zhu

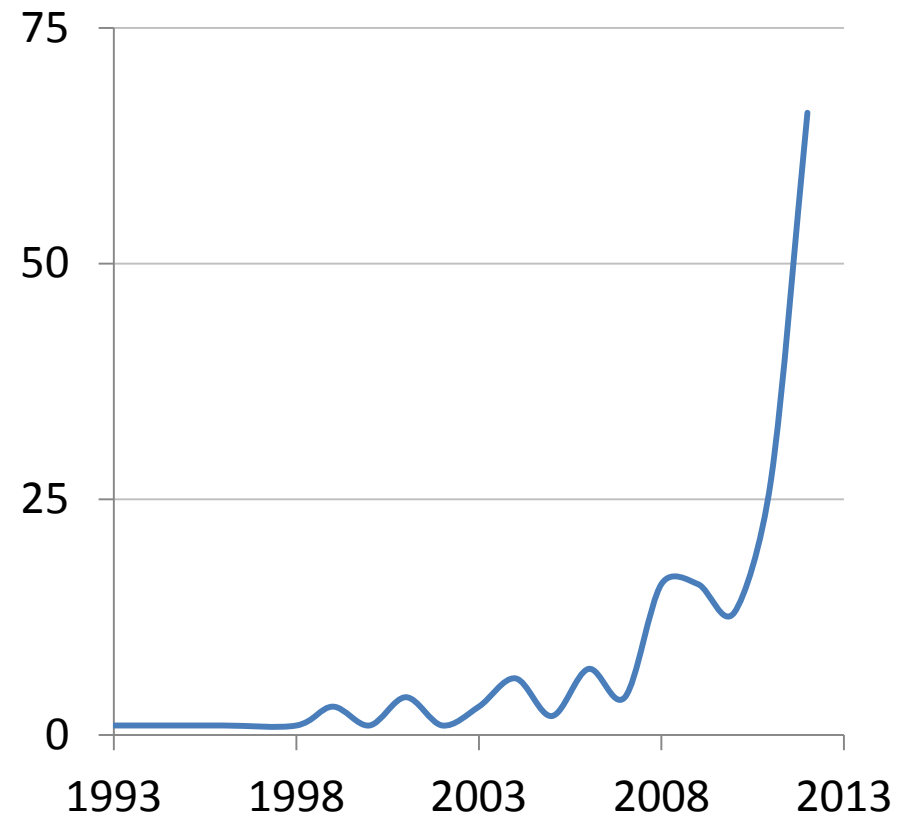
Department Research Seminar
January 21, 2013



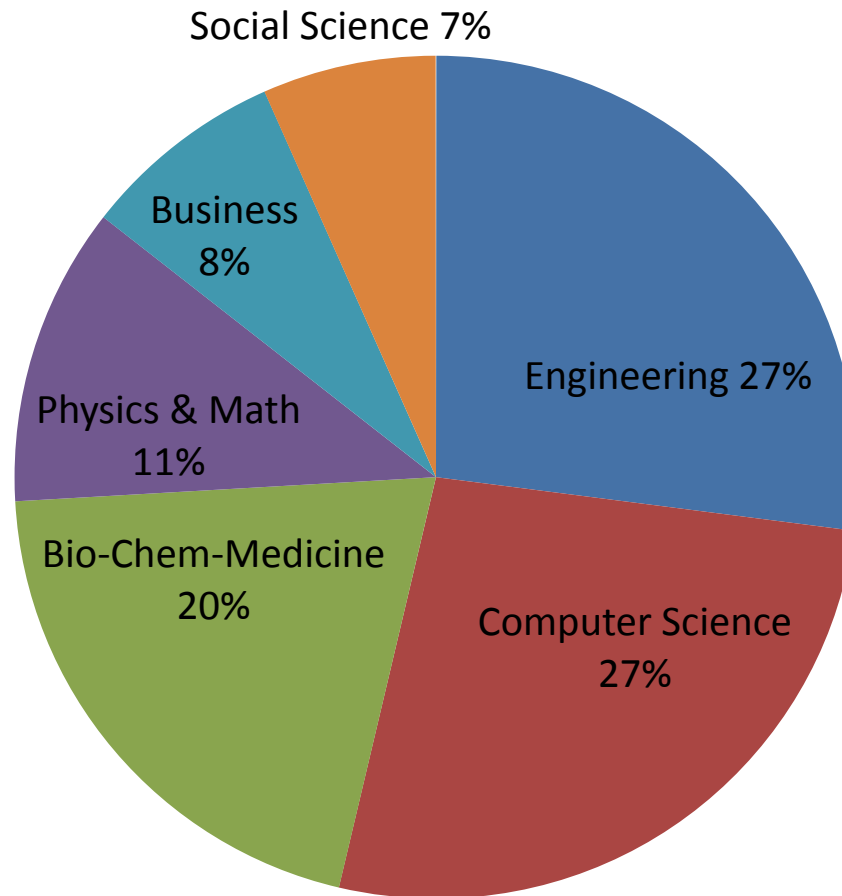
The Emergence of Big Data



N of SCI/SSCI Articles



Academic Stakeholders in Big Data



What's "Big Data"?

- Volume (海量)
- Velocity (快速)
- Variety (多源)
- Value (低价值)
- Storage capacity
- Computing capacity
- Seamless integration
- Optimal algorithms

Perspectives from CCF Big Data Task Force

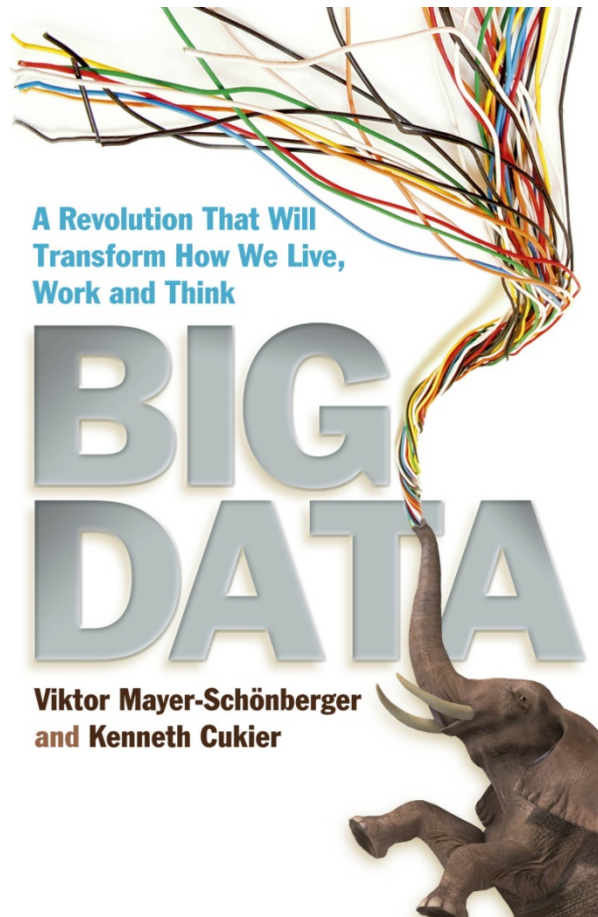
大数据的八个核心问题:

1. 数据科学与大数据的学科边界
2. 数据计算的基本模式和范式
3. 大数据的特性和数据态
4. 大数据的作用力与变化反应
5. 大数据的安全和隐私
6. 大数据对IT技术的挑战
7. 大数据的应用和产业链
8. 数据的生态环境问题

Eight Core Questions:

1. Data science as a discipline
2. Models and paradigms of data computation
3. Characteristics of big data
4. Actions and reactions of big data
5. Security and privacy of big data
6. Challenges to IT
7. Applications and business chains
8. Ecology of big data

Mayer-Schroenberger & Cukier (2013)



Big data will revolutionize

- how we live
- how we work
- how we think
 - population or sample
 - efficiency or precision
 - correlation or causation

Is Big Data a Big Lie?









Irfan Khan (CTO
of Sybase)
2012/03/21
[The Big Lie About
Big Data](#)

Massive data sets, even those including variable data types like unstructured data, can be ready for analysis in a columnar-based data warehouse. Not only are they ready, they are able to perform faster and readily scale to include as many users and as much data as is necessary to get the job done.

[Comment by Don MacLennan](#)

Columnar databases have their place in the analytics ecosystem, no doubt. ... But it skirts all of the reasons why complementary (threatening?) technologies like the Hadoop ecosystem exist. ... the author avoids the distinction between Big Data analytics and operational reporting. ... columnar databases are great for operational reporting on a price/performance basis. They don't compete on price performance in analytics domains unless they complement the Hadoop-type platforms.

Hyper, Myth or Reality?

Enthusiasts	What Do You Think?			Skeptics
	+	0	-	
Big data is better				Just more noise
Big data is a new era				A new bottle of old wine
Big data is everywhere				Inaccessible to outsiders
Tech is ready for big data				They are not ready yet
Sampling is unnecessary				Make big data small
Theory is dead				Theory wins eventually

Exemplar Studies of Big Data in Social Science

Survey

- Nate Silver's prediction of the U.S. election 2012

Experiment

- Bond et al.'s test of social message among 61 million users on Facebook

Content
Analysis

- Michel et al.'s text mining of 5 million Google books

References Cited

- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, C. J. E, & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489, 295-298.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331, 176-182.
- Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets in large data sets. *Science*, 334, 2018-2024.

Winner of Predictions in U.S. Election 2012

Nate Silver of fivethirtyeight.com emerged as the biggest winner of the prediction game. He correctly predicted the winner of 49 of the 50 states in 2008 and of every one of the 50 states in 2012. His data came from other election polls weighted by his model.



Pollster Accuracy and Bias, 2012 Presidential Election

Likely Voters Polls in Last 21 Days of Campaign

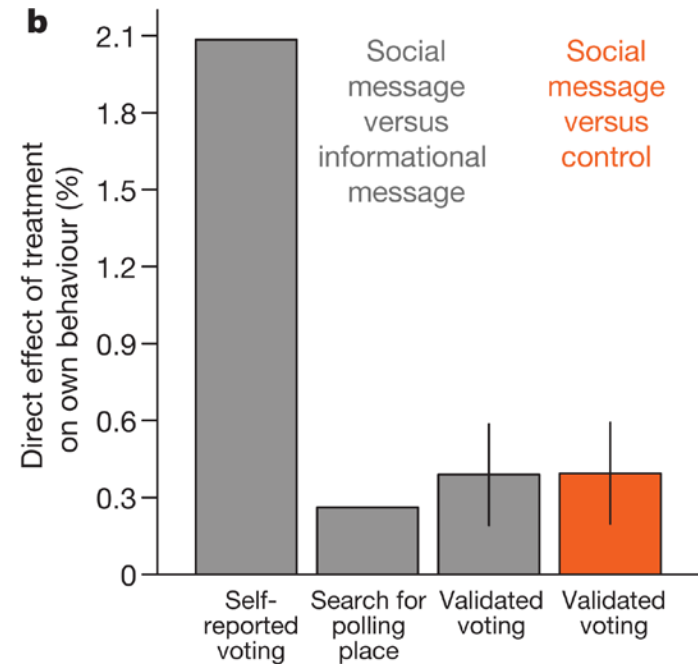
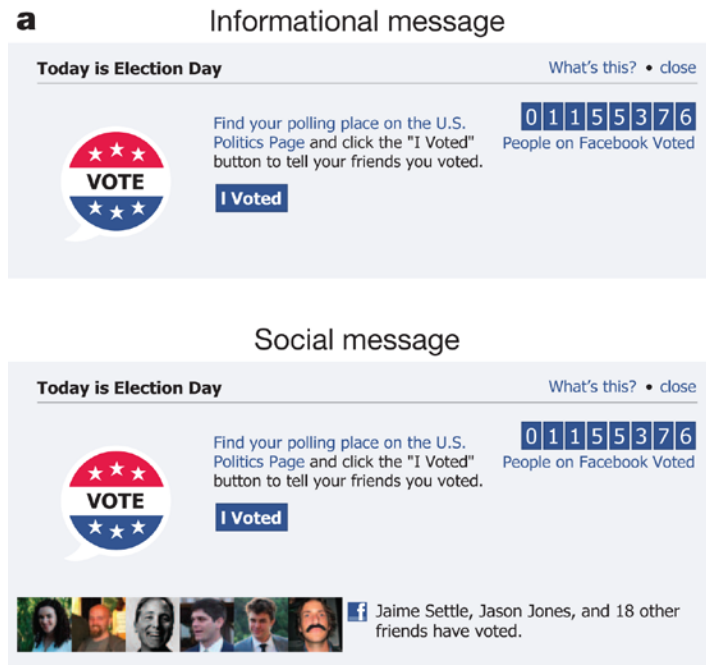
Minimum 5 Polls

Pollster	# Polls	Avg. Error	Bias	Mode	Cell?
IBD / TIPP	11	0.9	R +0.1	Live Phone	Yes
Google Consumer Surveys	12	1.6	R +1.0	Internet	N/A
Mellman	9	1.6	R +0.0	Live Phone	Yes
RAND Corporation	17	1.8	D +1.5	Internet	N/A
CNN / Opinion Research	10	1.9	R +0.6	Live Phone	Yes
Ipsos / Reuters (online)	42	1.9	R +1.4	Internet	N/A
Angus Reid	11	1.9	R +0.8	Internet	N/A
CVOTER International / UPI	13	2.0	R +2.0	Live Phone	??
Grove Insight	18	2.0	R +0.1	Live Phone	Yes
SurveyUSA	17	2.2	R +0.5	Robodial	Yes
Quinnipiac	5	2.3	D +0.3	Live Phone	Yes
Marist	11	2.5	R +1.0	Live Phone	Yes
YouGov	30	2.6	R +1.1	Internet	N/A
We Ask America	9	2.6	D +0.1	Robodial	No
Public Policy Polling	71	2.7	R +1.6	Robodial	No
Gravis Marketing	16	2.7	R +2.7	Robodial	No
JZ Analytics*	17	2.8	R +0.1	Internet	N/A
Washington Post / ABC News	16	2.8	R +2.7	Live Phone	Yes
Pharos Research Group*	14	4.0	D +2.5	Live Phone	No
Rasmussen Reports	60	4.2	R +3.7	Robo + Internet	No
American Research Group	9	4.5	R +4.5	Live Phone	Yes
Mason-Dixon	8	5.4	R +2.2	Live Phone	Yes
Gallup	11	7.2	R +7.2	Live Phone	Yes

* Not used in FiveThirtyEight forecast.

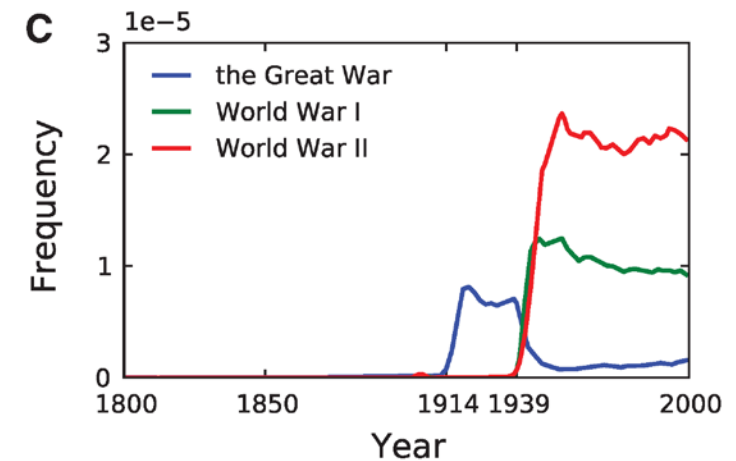
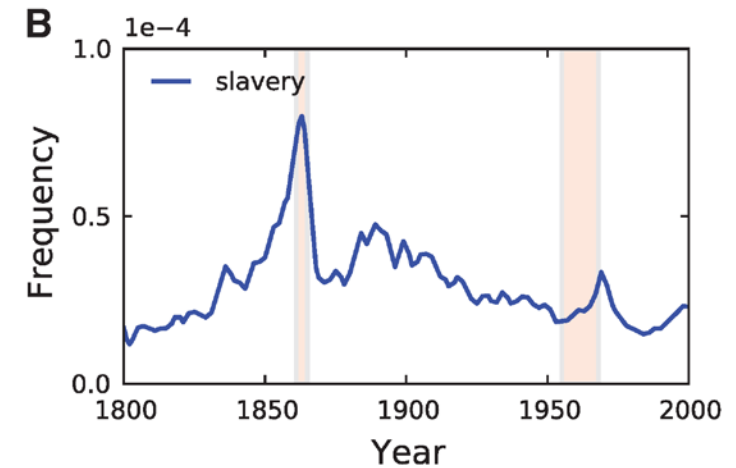
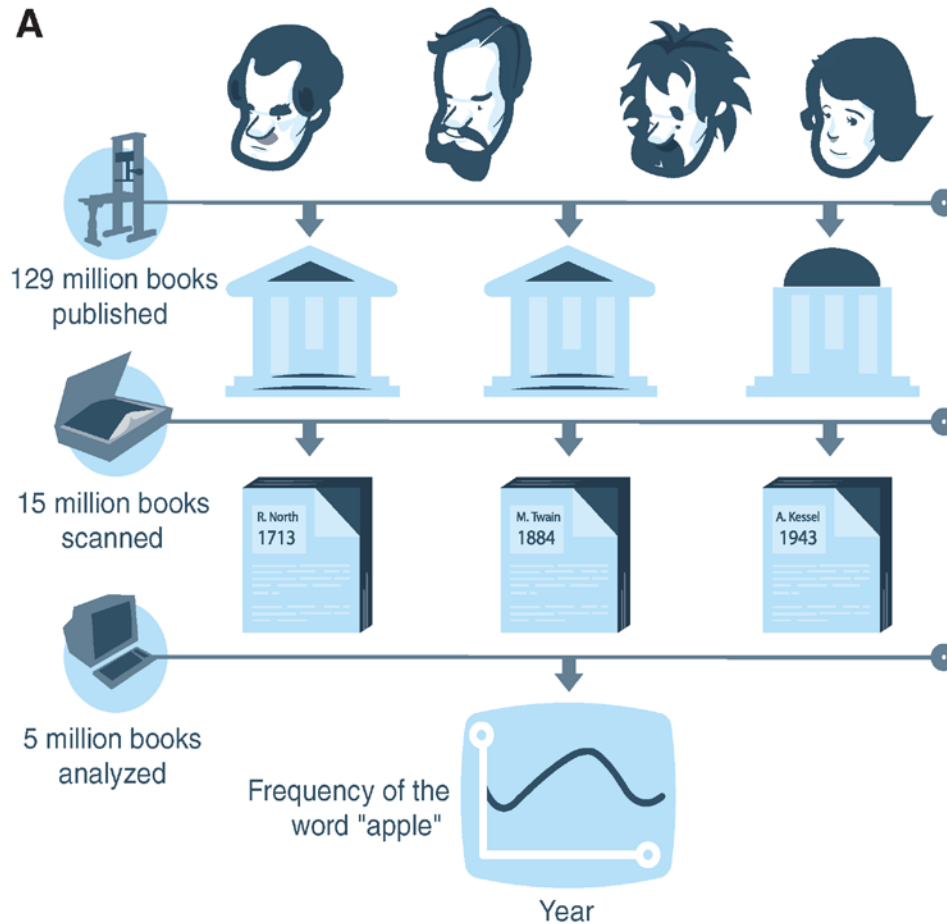
Online surveys are generally more accurate than telephone polls!

A 61-mil-person Experiment on Facebook



An effect of 0.3% amounts to 180,000 voters.
Is it statistically significant but practically trivial?

Content Analysis of 5 Million Google Books

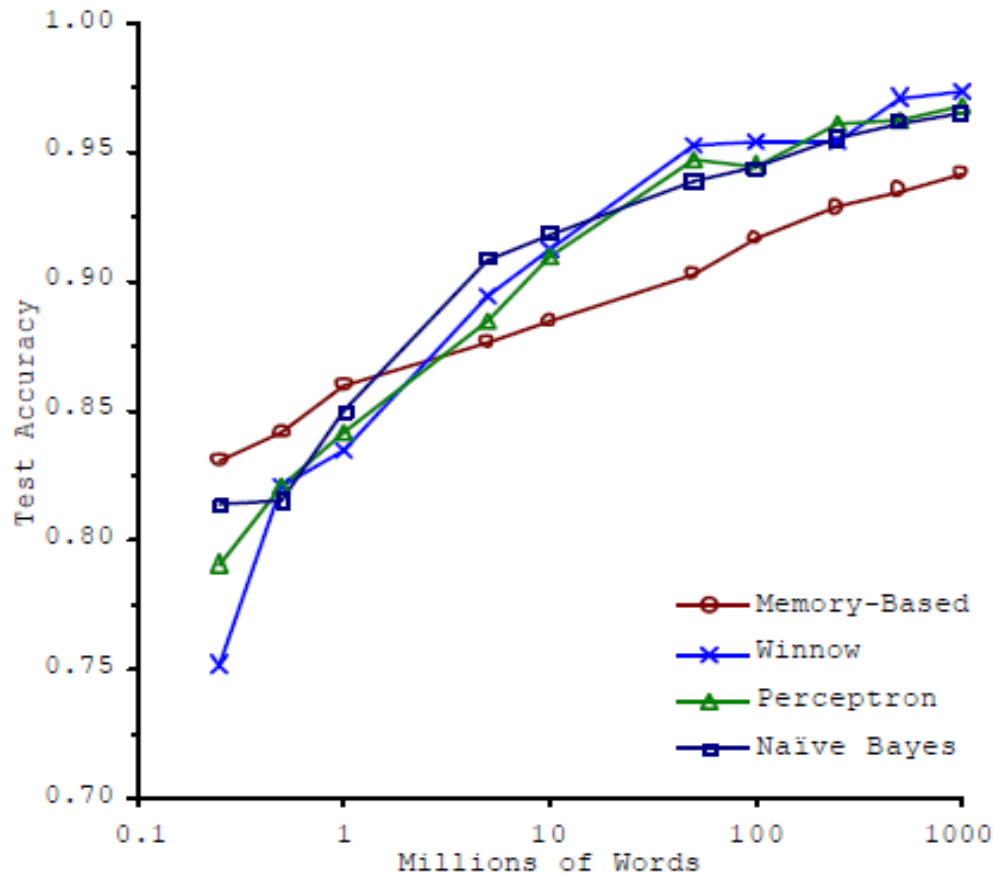


Multidisciplinary Team

1. Evolutionary Dynamics
 2. Biology
 3. Computer Science and Artificial Intelligence
 4. Medical School
 5. Organismic and Evolutionary Biology
 6. Mathematics
 7. Engineering and Applied Sciences
 8. Cultural Observatory
 9. Political Science
 10. Psychology
 11. Google
 12. Houghton Mifflin
Harcourt Publishing
 13. Encyclopedia Britannica
- We should be there, but
are absent unfortunately!**

Is more data really better?

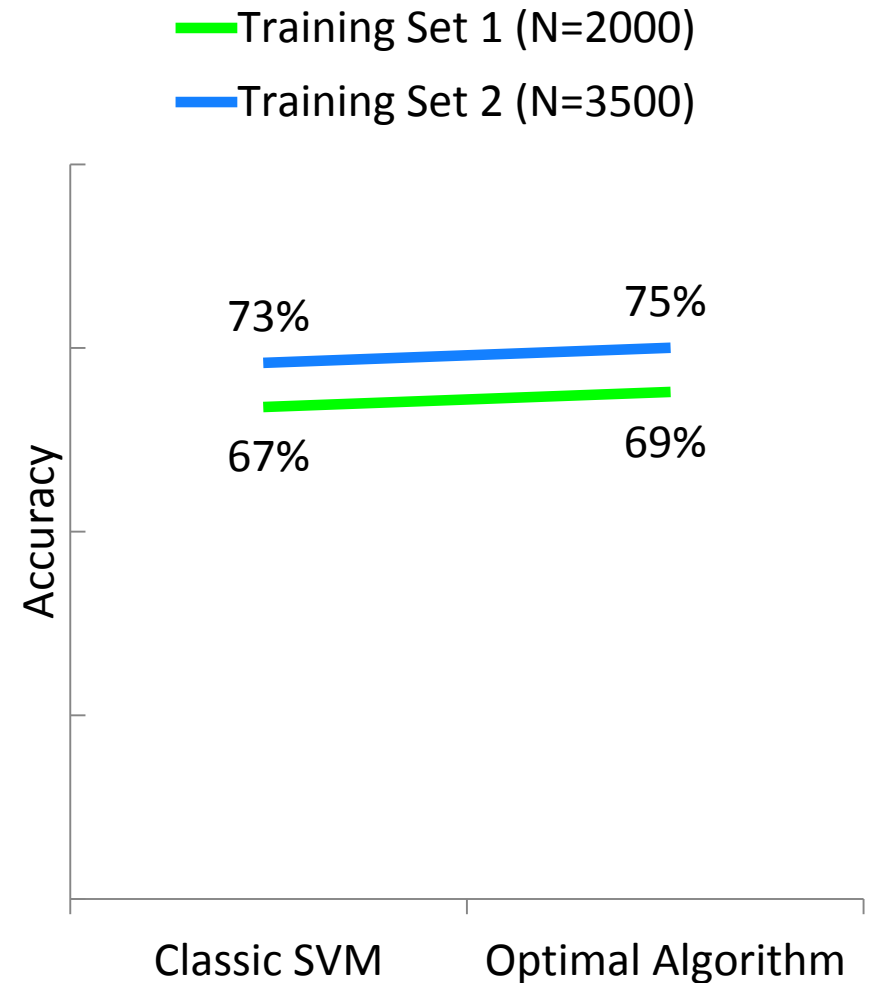
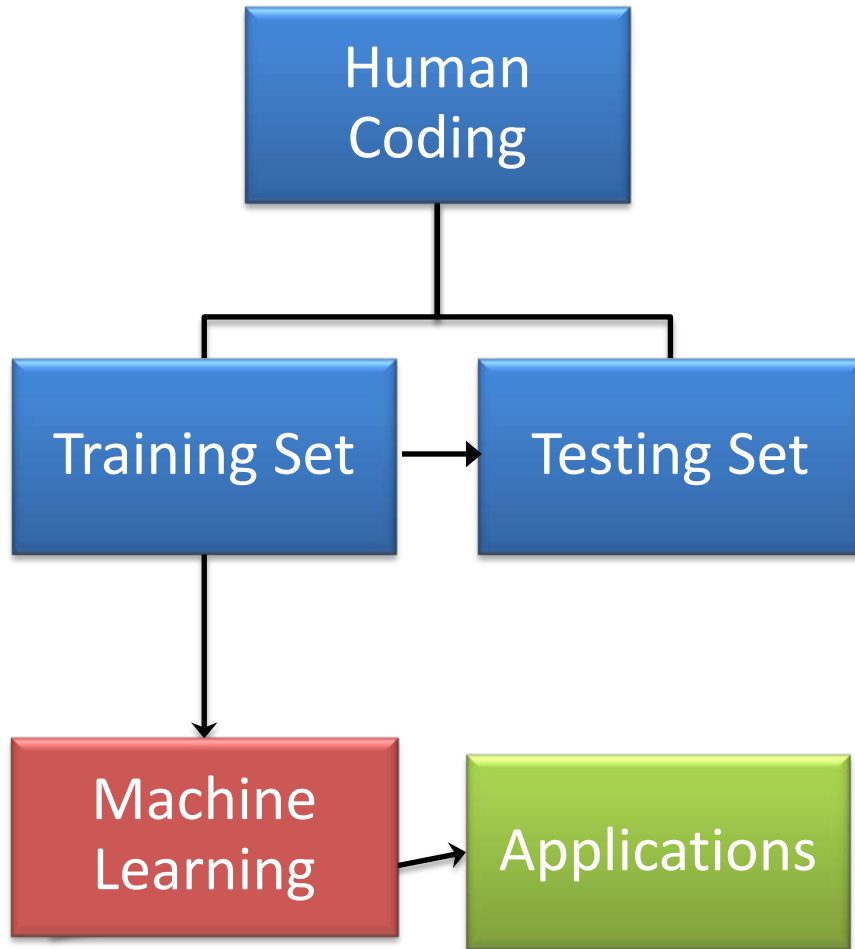
Quantity determines quality of machine learning



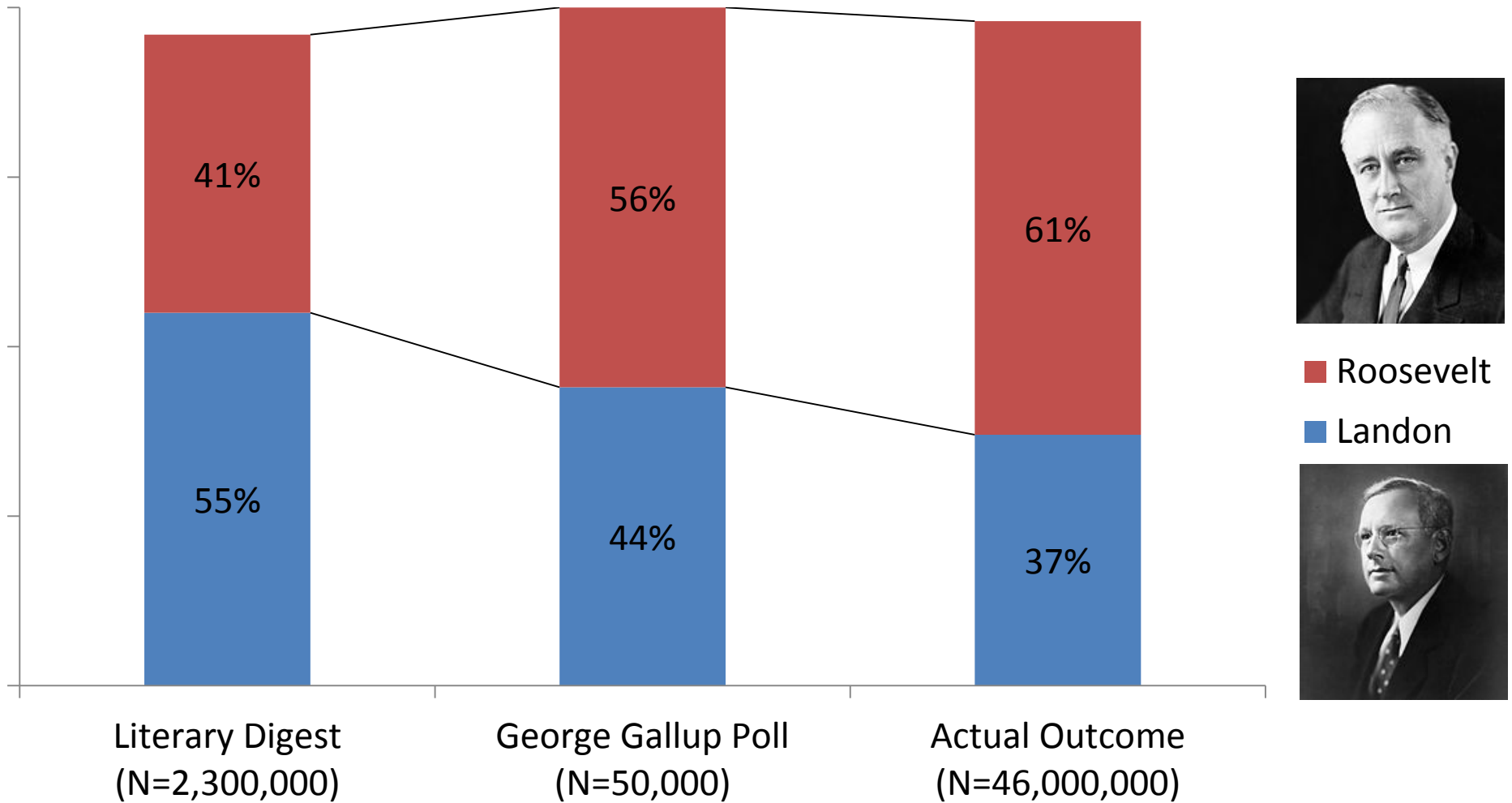
Banko and Brill (2011):

- Task: disambiguate confusion words (e.g., “than, then”)
- Learning set: varies in size up to 1 bil words
- Test set: Wall Street Journal articles in 1 mil words
- Findings: accuracy is affected by the size of learning set, not by the choice of learning algorithms

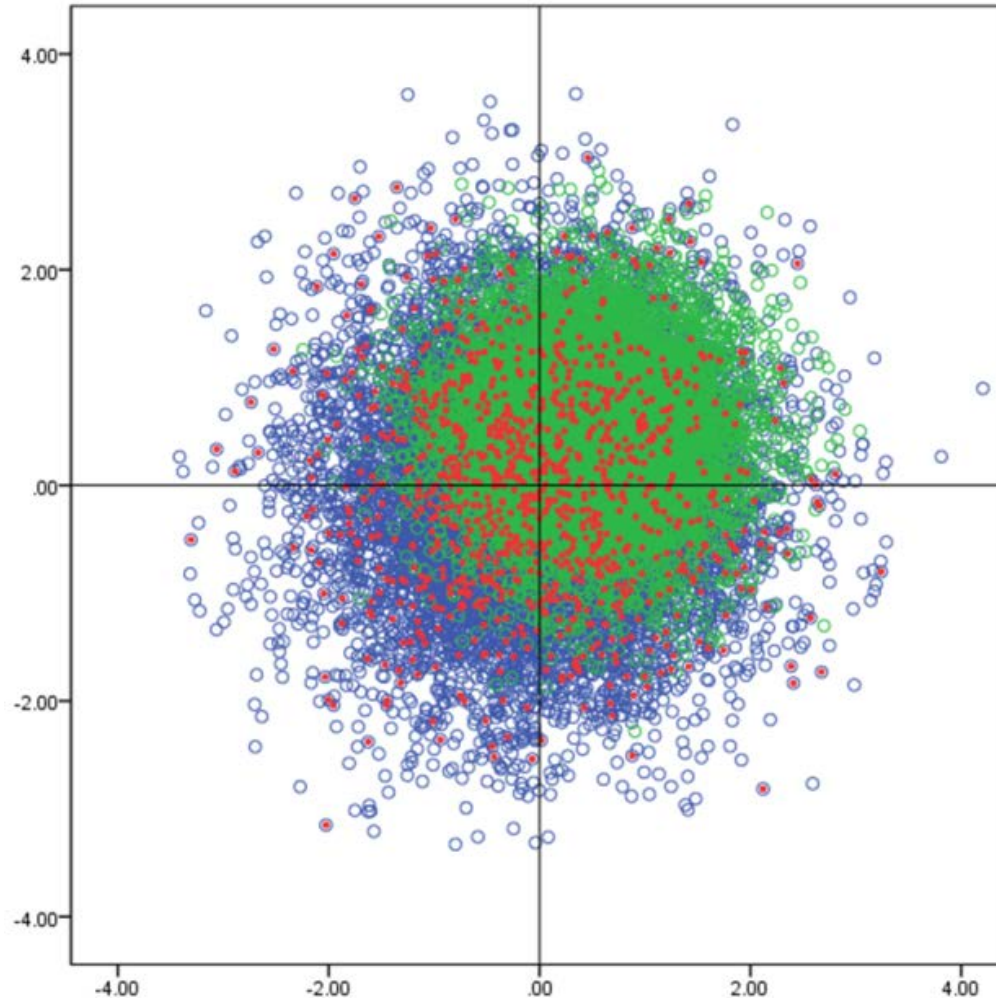
Sample Size vs. Optimal Algorithm on Quality



A Big-Data Experiment in Social Science



Population, Sub-population, Sample

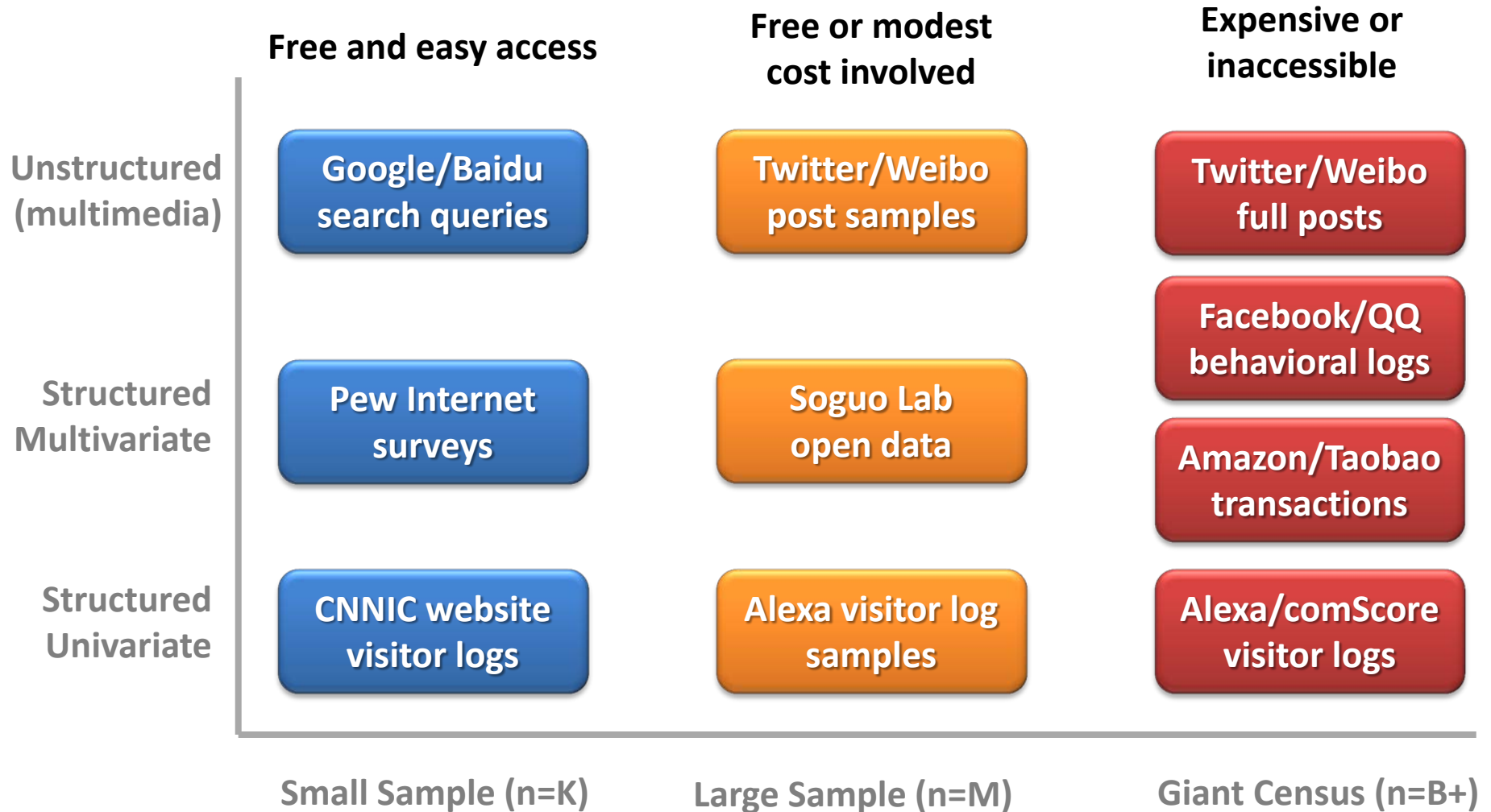


A large but biased sub-population (e.g., 80% of the population) is less informative/more harmful than a small but representative sample (e.g., only 5% of the population).

- Population (100%)
- Sub-pop (80%)
- Sample (5%)

Is big data really there?

Where Can I Find Big Data?



Cases vs. Variables

Traditional small sample: small N but many vars

ID	X1	X2	...	X _j	Y1	Y2	...	Y _k
1
...
n

Ideal big data: big N and many vars

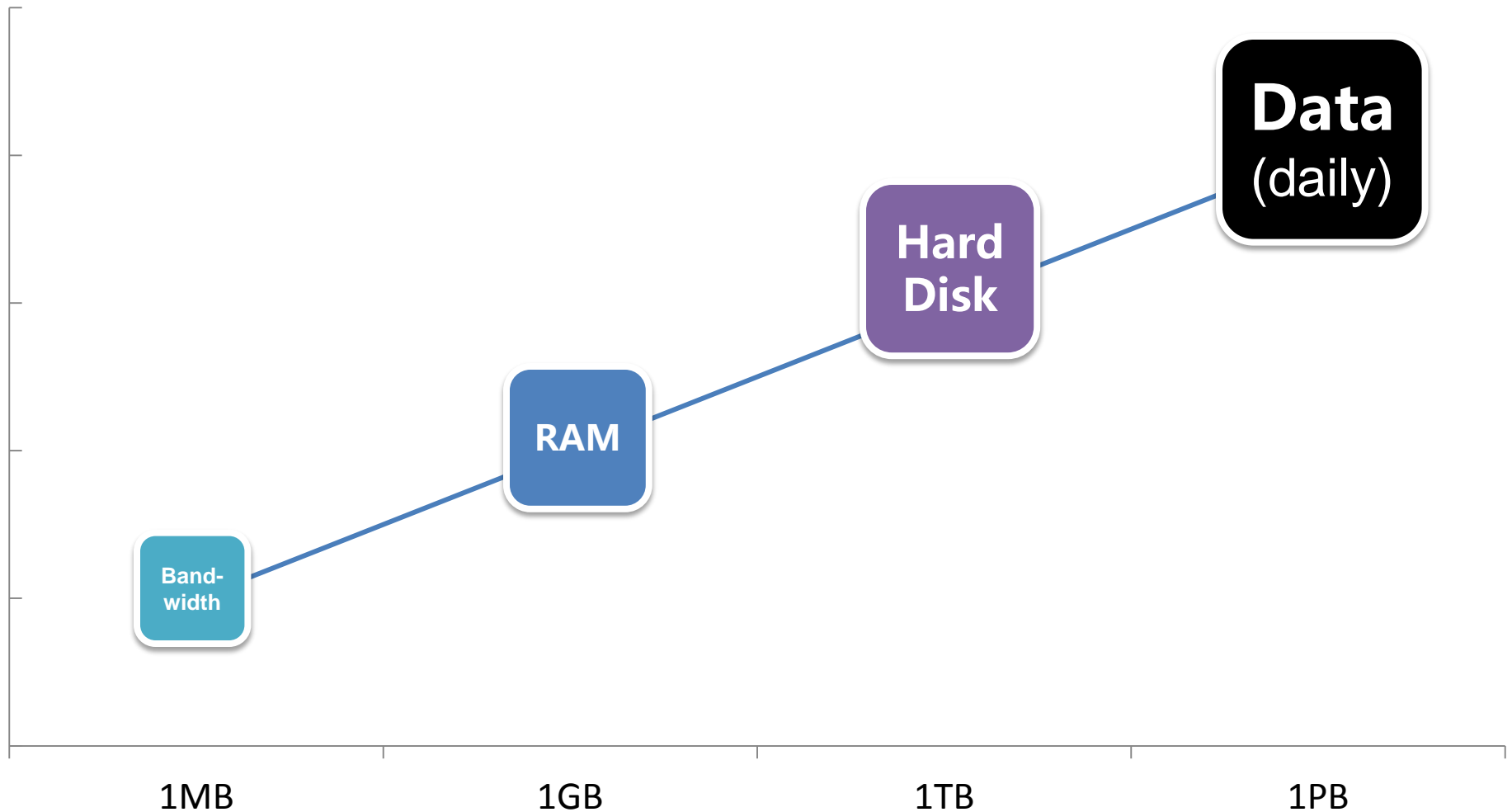
ID	X1	X2	...	X _j	Y1	Y2	...	Y _k
1
...
n
...
∞

Real big data:
big N, few vars

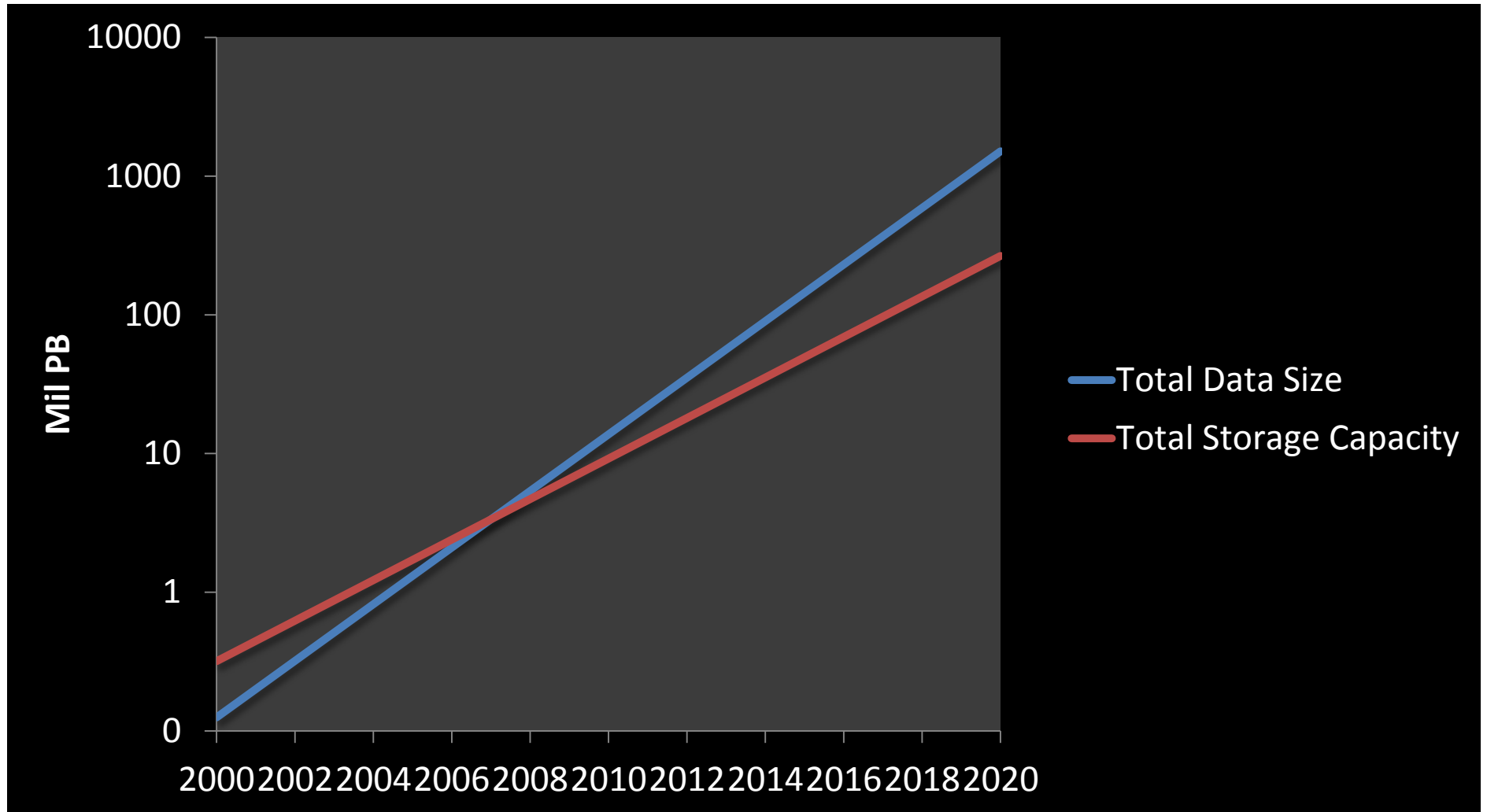
ID	X1	X2
1
...
n
...
∞

Are we **physically** ready?

Race between Data Size and the Infrastructure



Growth Gap between Data and Storage



Data Storage

A medium-size data center usually costs US\$1.5M/year



Are we **intellectually** ready?

New Statistical Methods for Big Data

David Reshef, Yarik Reshef, et al. (2011).

Science, 334, 2018-2024

Detecting novel associations in large data sets using maximal information coefficient (MIC)



A

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

MIC to big data is the same as Pearson's correlation to small data more than 100 years ago, which means that we're at a very early stage (i.e., bivariate analysis of linear relationship) of big data research.

Are we **philosophically** ready?

The End of Theory?



Wisconsin-Madison

George Box (1987):
All models are wrong, but some are useful.



Google Research

Peter Norvig (2008):
All models are wrong, and increasingly you can succeed without them.



Wired Magazine

Chris Anderson (2008):

This is a world where massive amounts of data and applied mathematics replace every other tool ... Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.



When Science Meets Big Data

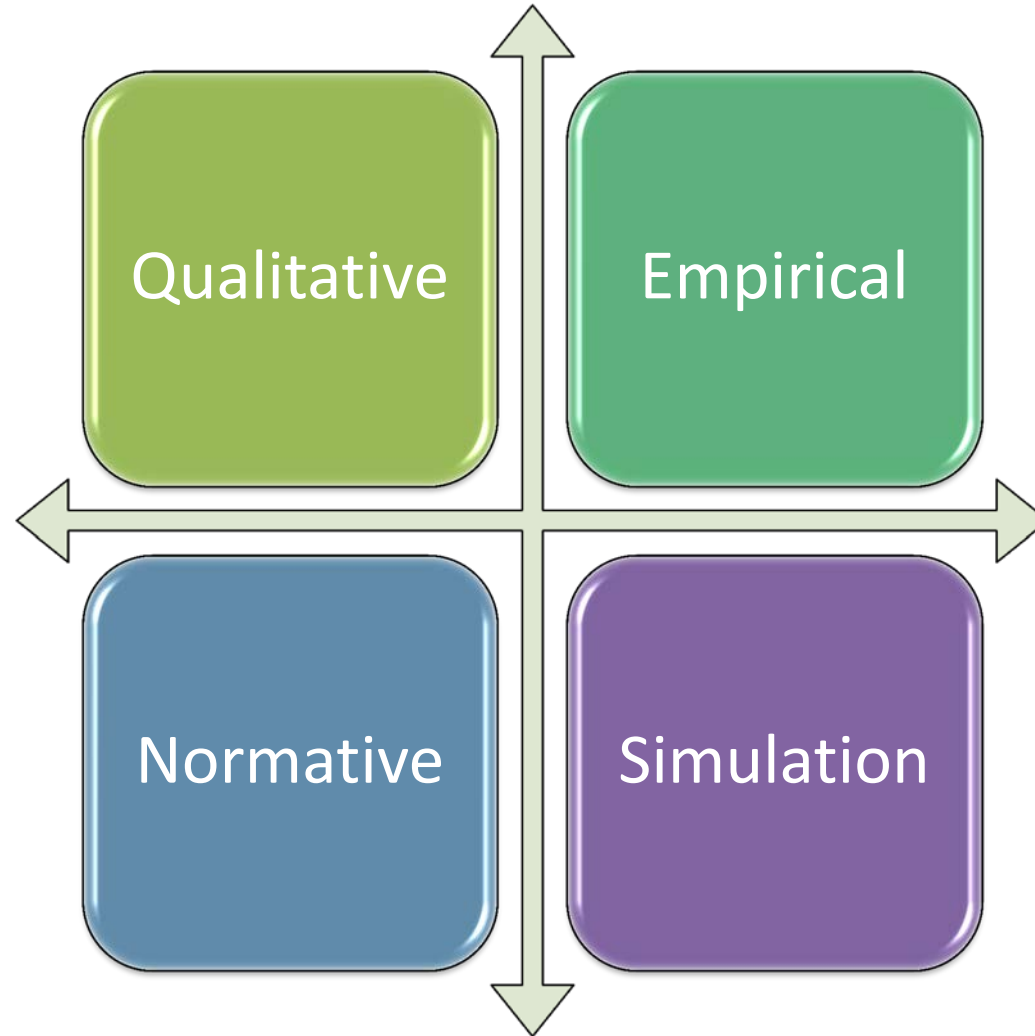
Classic Paradigm:

- The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works.
- Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

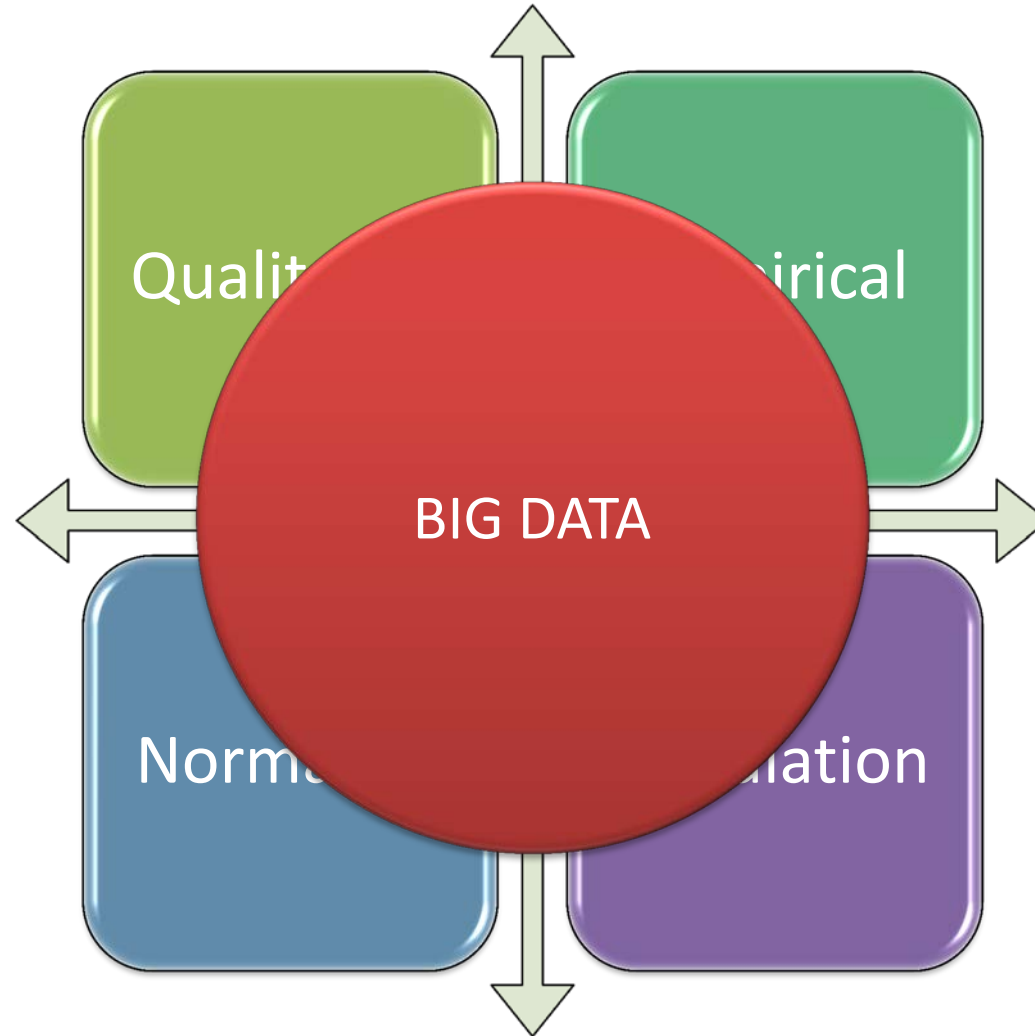
Big Data Paradigm:

- Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.
- The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

Existing Paradigms of Social Science



Challenges to Existing Paradigms



What Should We Do?

Q: Do we have to use big data?

A: No. Big data is not a necessity but an option, a costly and beneficial option.

Q: What will happen if we stay with small data?

A: We'll be marginalized and bypassed by those who play big data.

Concluding Remarks

1. Big data is indeed everywhere.
2. Everything else being equal, big data does have advantages over small data.
3. Social scientists will have access to only a small fraction of the data ocean in the foreseeable future.
4. There are a lot of barriers and constraints on constraints on data storage, processing, analysis, and applications, caused by economic, scientific, technical, and legal reasons.
5. Data scientists are in short supplies, especially those who can bridge information science and social science.
6. Social scientists should and could participate in big data, by contributing our experience with small data.

THANK YOU & CONTACT US @

j.zhu@cityu.edu.hk

weblab.com.cityu.edu.hk

weibo.com/weblabcityu

