AON
Empower Results®
Stephen Mildenhall
James Guszcza
Deloitte.

Big Data: What is it?

And what does it mean for the insurance industry?

CAS Centennial Meeting
New York City
November 10, 2014

In the past few years we have produced more data than in all of human history

Data is the new oil

Data Science: the Sexiest job of the 21st century

The 2012 presidential election was the world's first big data election.

Big data is one of the greatest sources of power in the 21st century

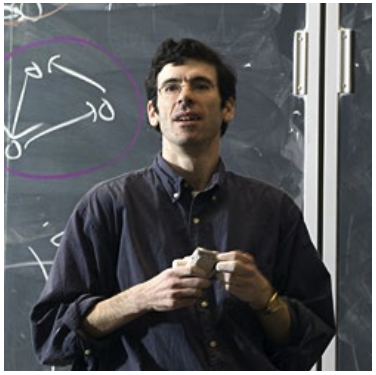
Big Data: a revolution that will transform how we live, work, and think

It will make you rich

The potential to transform everything

"The term itself is vague, but it is getting at something that is real… Big Data is a tagline for ***a process that has the potential to transform everything.***"

— Jon Kleinberg, Cornell University



nature International weekly journal of science

Computational social science: Making the links

From e-mails to social networks, the digital traces left by life in the modern world are transforming social science.

Themes

What is big data?

A few examples of big data in action

Technology and analytic tools for big data

Big data and behavioral data

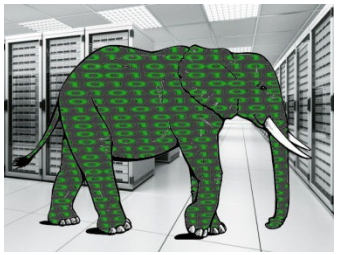
The dark side: ethics, business and liability

Data as a positive force: a new mindset for big data

So, what is it?

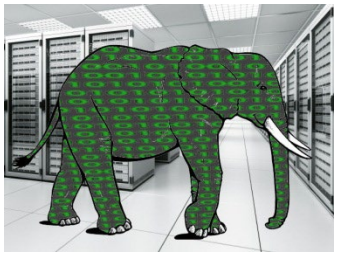
Three definitions of big data

1. Data sets with sizes beyond the capability of standard IT tools to capture, process, and analyze in reasonable time frames.



Three definitions of big data

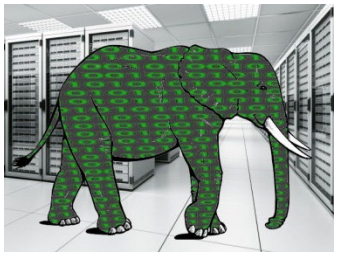
1. Data sets with sizes beyond the capability of standard IT tools to capture, process, and analyze in reasonable time frames.



2. Data with high **V**olume, **V**elocity, **V**ariety
   - Huge datasets
   - … emanating continuously from smart phones, sensors, cameras, GPS devices, computers, TVs, …
   - … involving all manner of numeric, text, photographic data

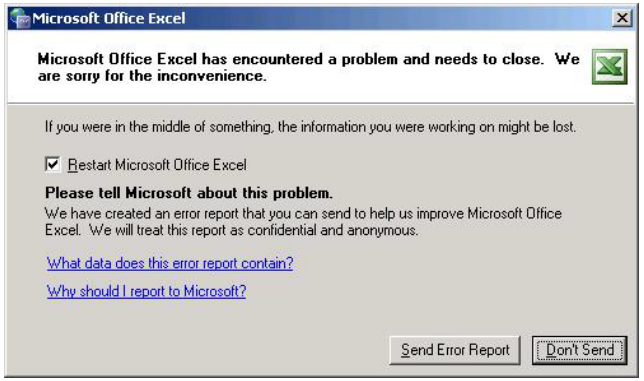
Three definitions of big data

1. Data sets with sizes beyond the capability of standard IT tools to capture, process, and analyze in reasonable time frames.



2. Data with high **V**olume, **V**elocity, **V**ariety
   - Huge datasets
   - … emanating continuously from smart phones, sensors, cameras, GPS devices, computers, TVs, …
   - … involving all manner of numeric, text, photographic data

3. "Anything that doesn't fit in Excel"



Traditional insurance data vs. new “big” data

Traditional Data

Guidewire
Deliver insurance your way

SAP

PeopleSoft

LexisNexis®

CoreLogic®

Internal databases

New “Big” Data

flickr™

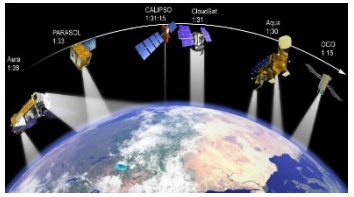
Bloomberg

Google

NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA

ebay

NETFLIX

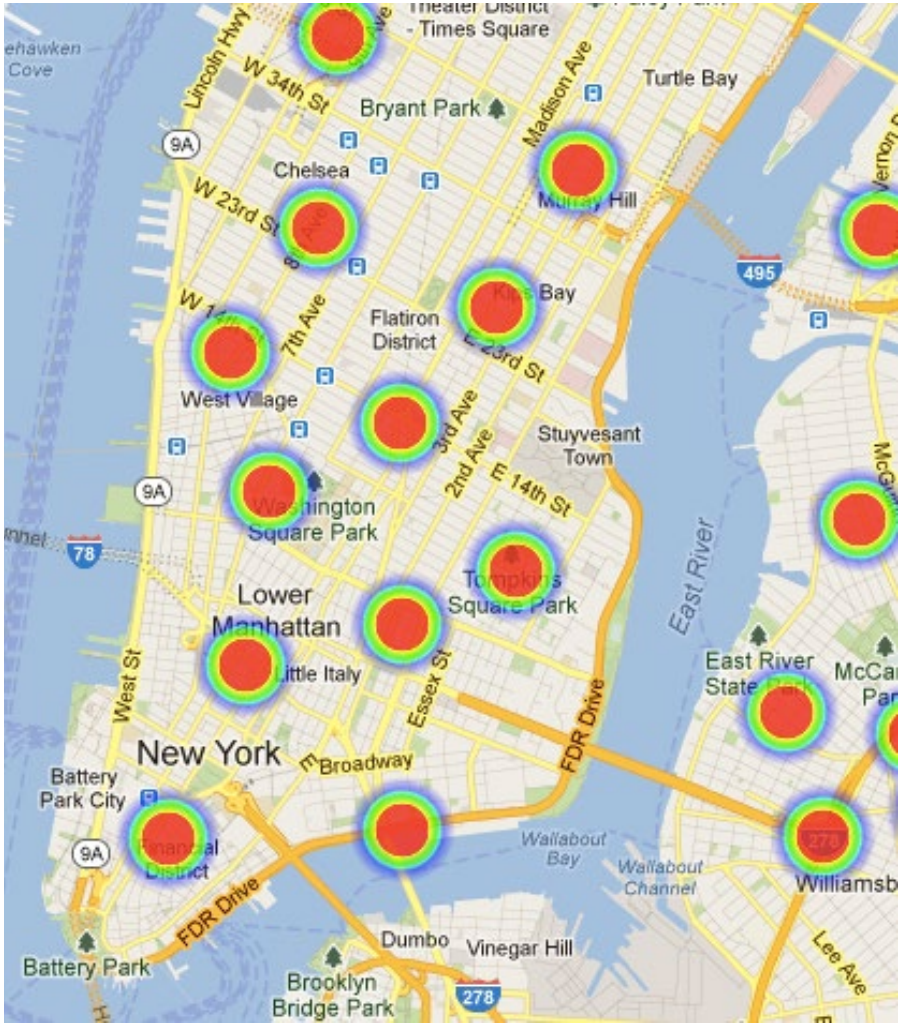


Expensive to collect, high value
Static, mixed within type

“Free” user-content, low value
Dynamic, fixed within type

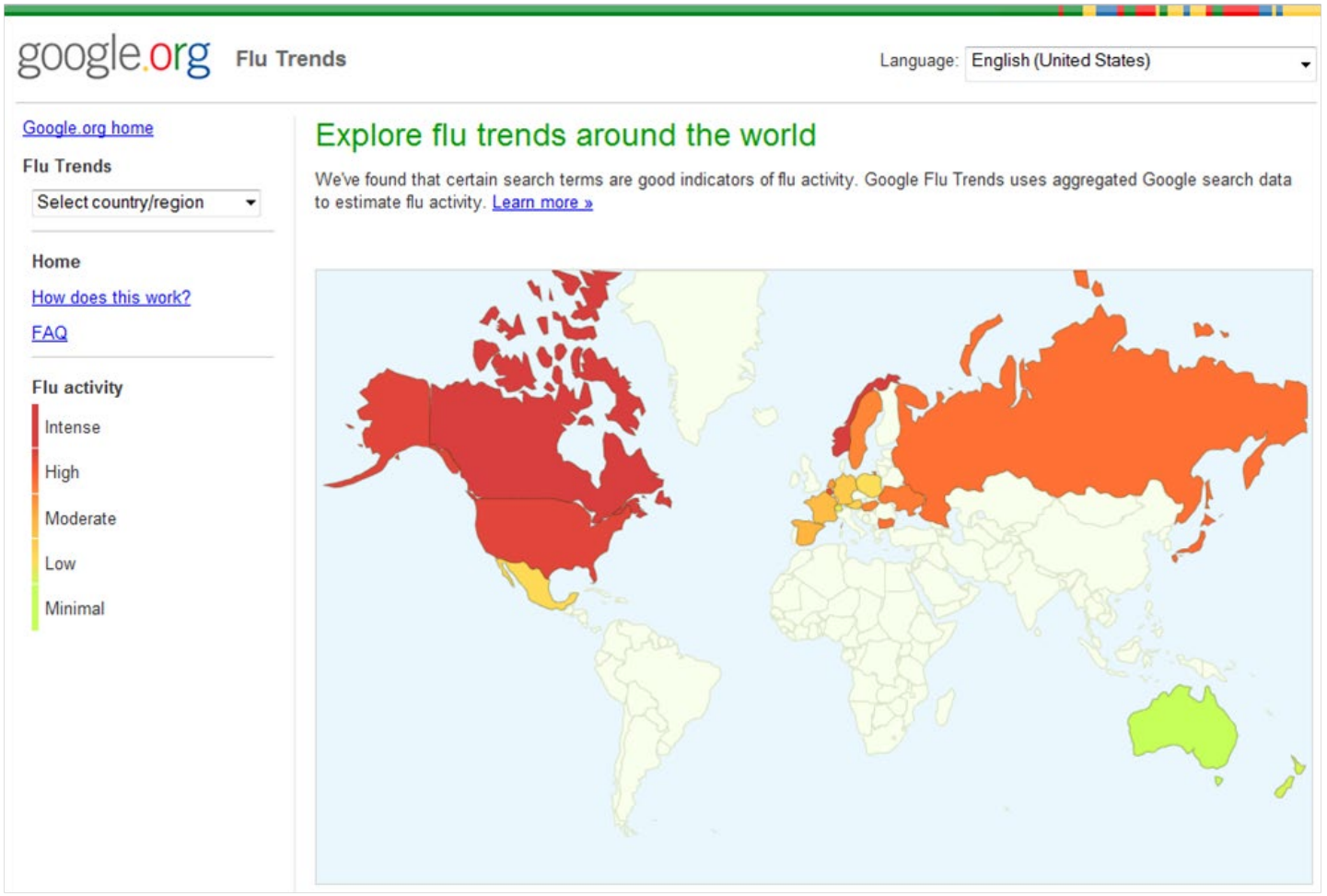
Some celebrated examples of big data in action

The big apple does actuarial science



Data-driven building inspections in New York City: Prospectively identify safety risks/violations

Taking the temperature of the population

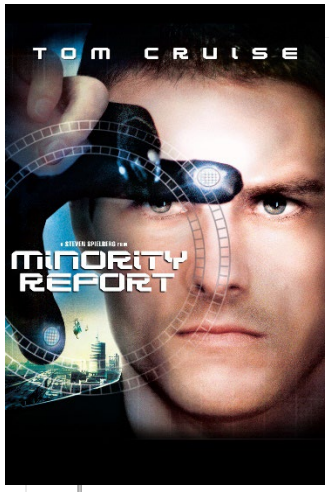


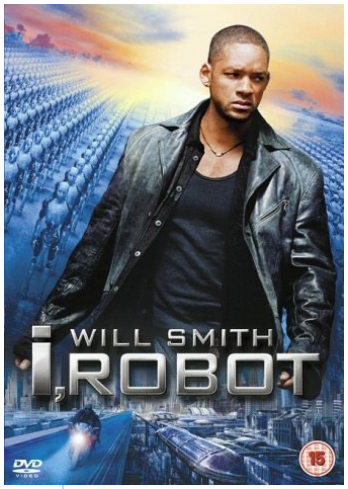
Similar ideas for economics, weather, risk hot-spots, or the "mood" of a customer base or populace

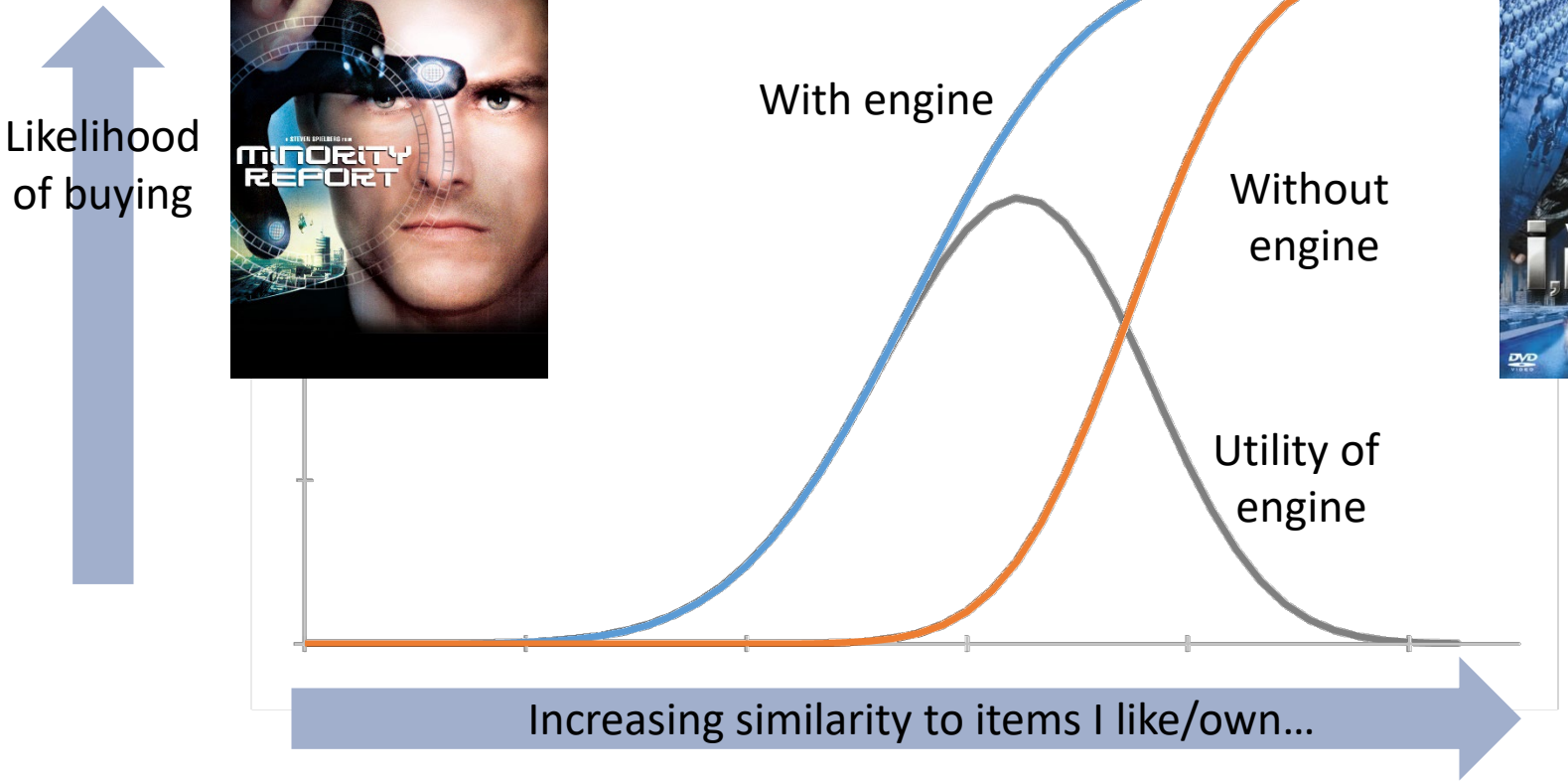
Better viewing through “datafication”

House of cards was actually built on a solid foundation of data

You may also like…







Recommendation engines must look beyond the obvious to delight and surprise customers

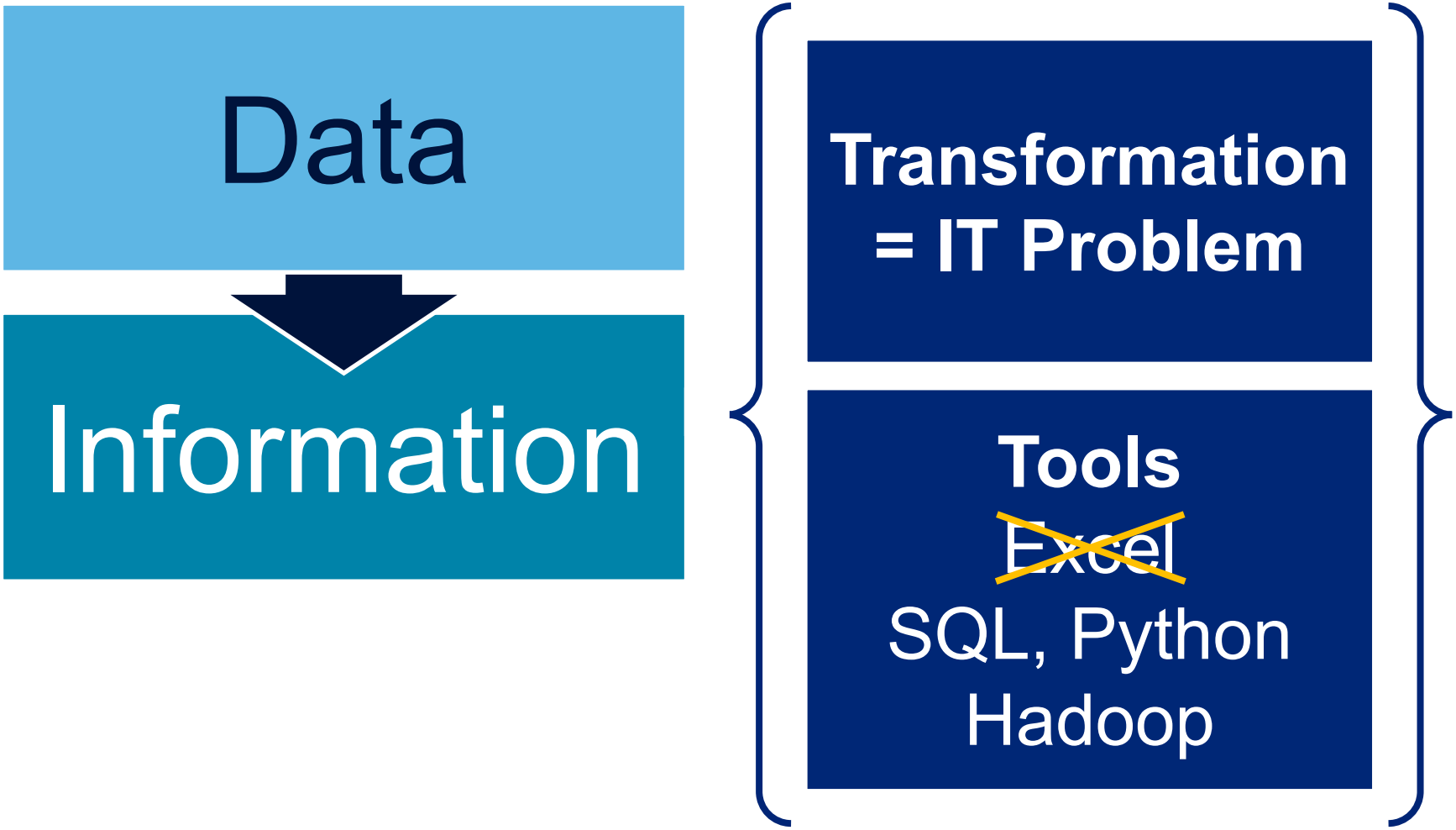
Technology and analytics transform data into insight

We need more than data

In general, people are not interested in data. What people want are answers.

-- David Hand, Data not Dogma

Work and some magic transforms data into information



Old & new tools and methods compared

Old World	New World
Expensive, customized hardware	**Cheap**, commodity hardware
Model data	Dump data
Pre-optimize (index) to expected queries	Post-optimize requested ad hoc queries
Early binding types	Late binding types
Compute everything, dimensional DW	Compute what you want, on demand
Move data to the code	Move the code to the data
Inflexible, changes slow	Flexible, built to change

Magic = fundamentally different operating model, 4x speed-up

How new methods can radically improve on old

Are two database records equal?

Old School

- Age1 = Age2
- Sex1 = Sex2
- Wt1 = Wt2
- SSN1 = SSN2
- Etc.

- Implement for each record type
- Slow to execute

New School

- Compare binary numbers

- Works for ALL record types
- Quick to execute: primitive CPU function

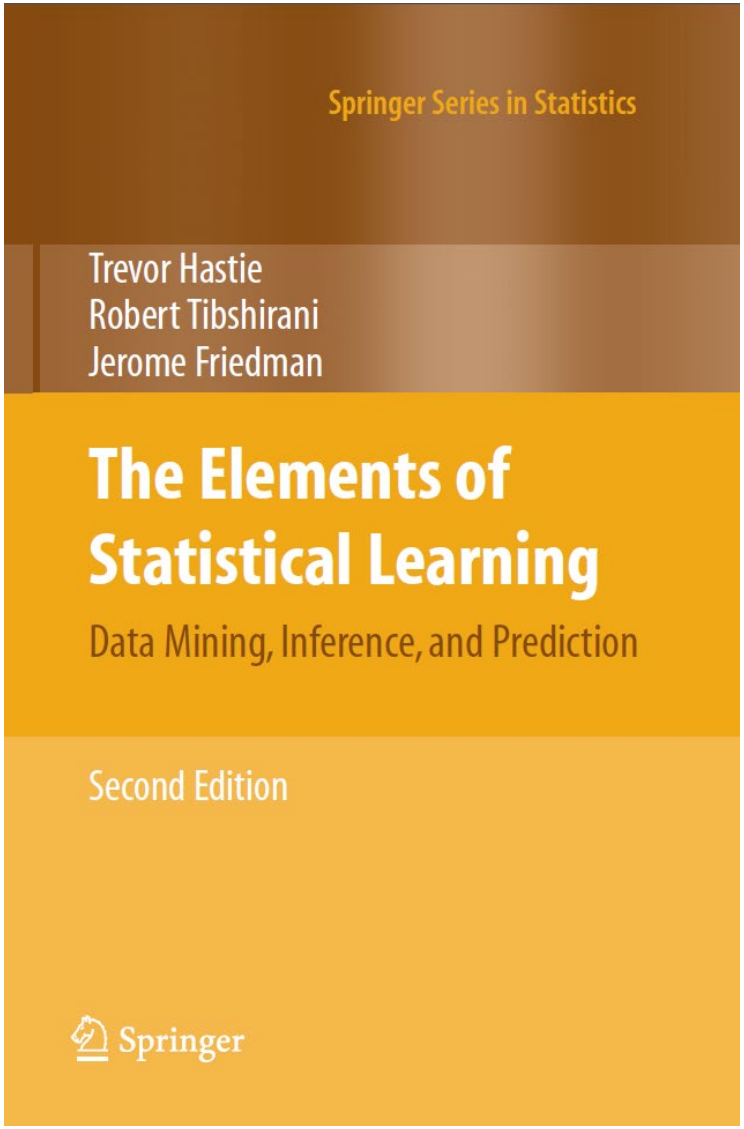
Analysis transforms information into insight

Methodology
Exploratory analysis
Data adjustments
Variable selection
Model validation
Hold-out samples
…

Algorithms
Machine Learning
SVM · NLP
Signal processing
Topological models
GAM · Splines
…

Tools
R
SAS
Matlab
Python
Prolog
…

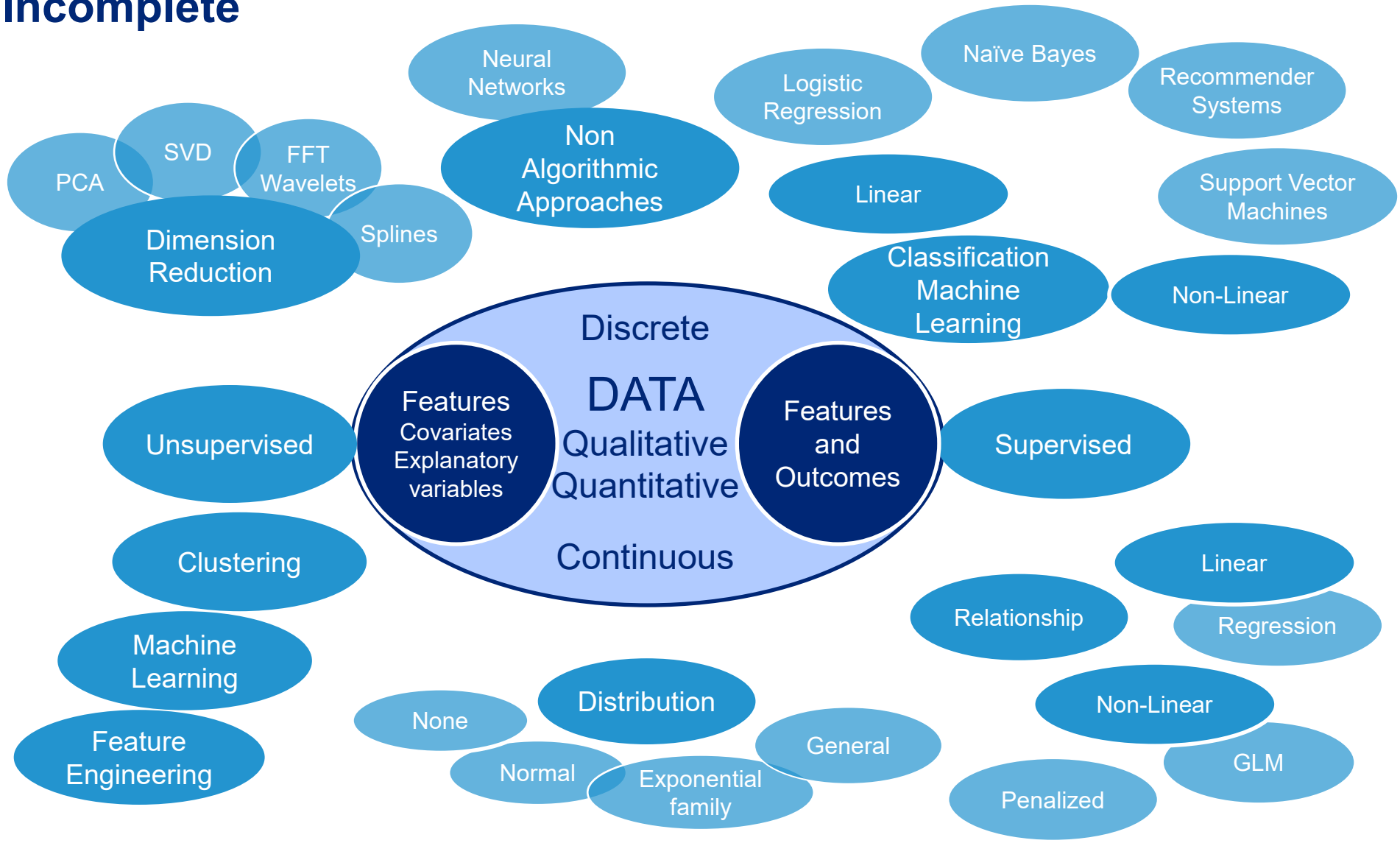
Techniques and algorithms



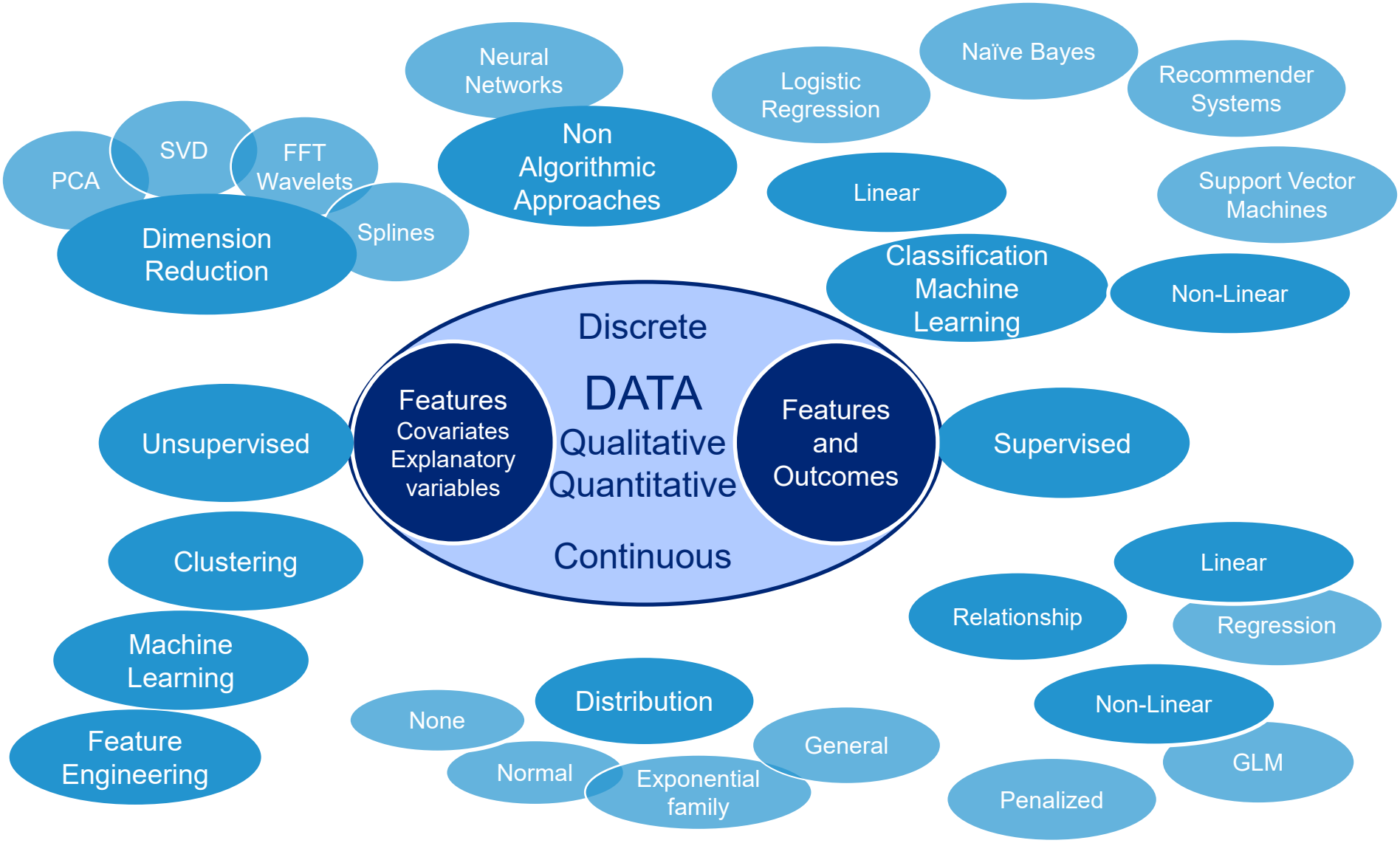
http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

Schematic of analytic techniques and algorithms

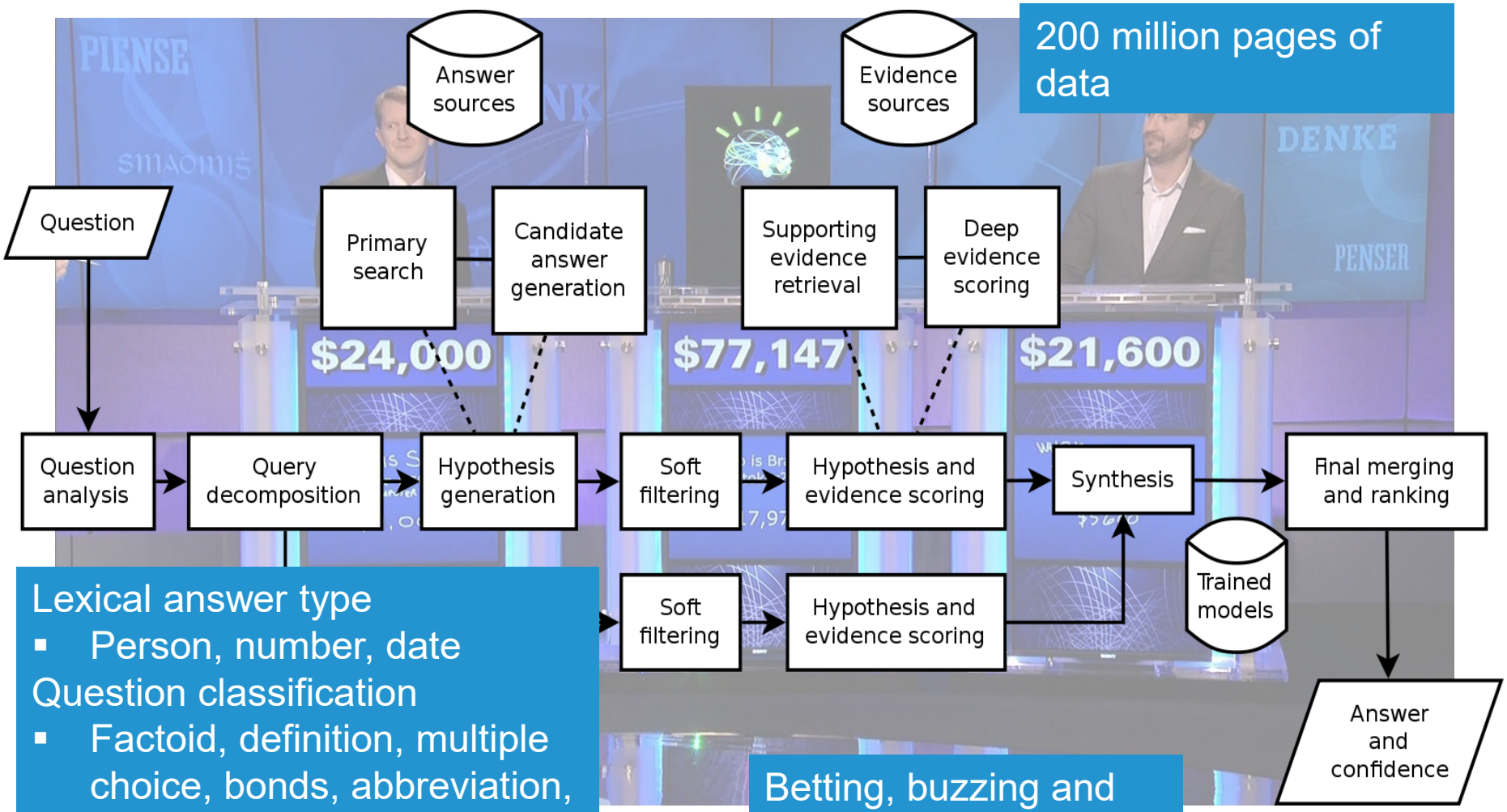
Incomplete



Schematic of analytic techniques and algorithms



Watson Wins Jeopardy! Training model a “significant effort”



Google Translate & the "unreasonable effectiveness of data"

A major big data success story!

Google Translate

"Learning from text at web scale"

- Non-parametric – no probability assumptions
- Uses unstructured text corpuses "in the wild"
- N-word matching
- "follow the data"

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

"Invariably, simple models and a lot of data trump more elaborate models based on less data…

Currently, statistical translation models consist mostly of large memorized phrase tables that give candidate mappings between specific source- and target-language phrases."

From machine translation to meme

"There is now a better way. 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."

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

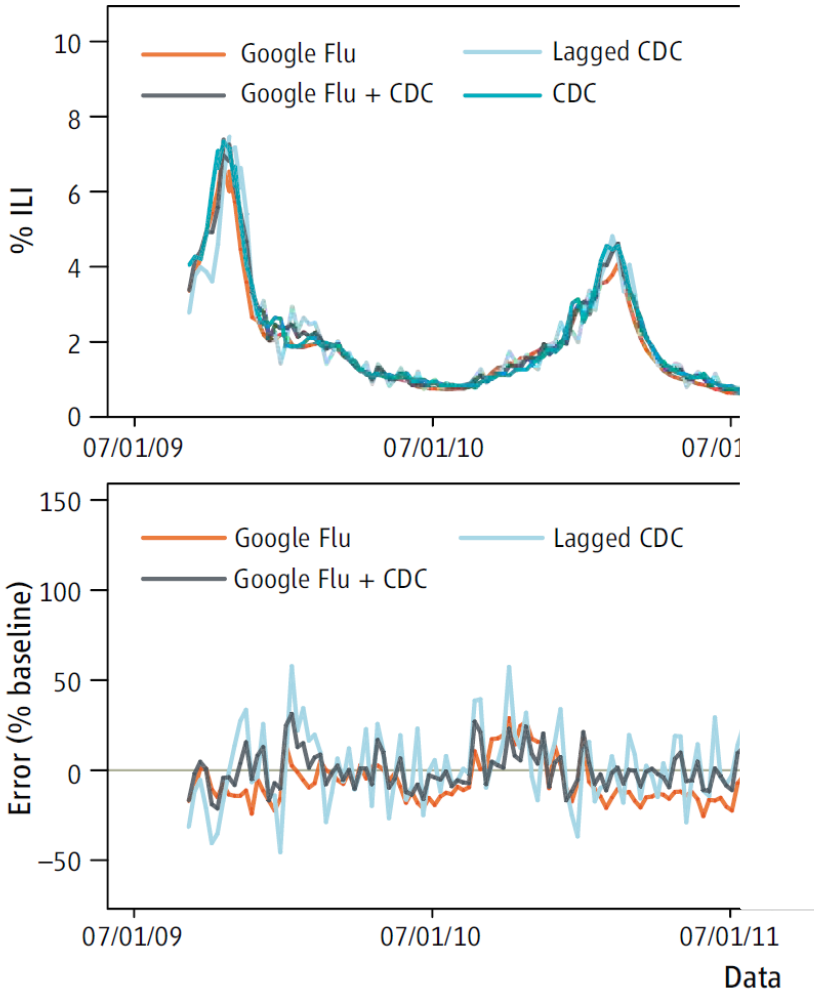
Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse,

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and

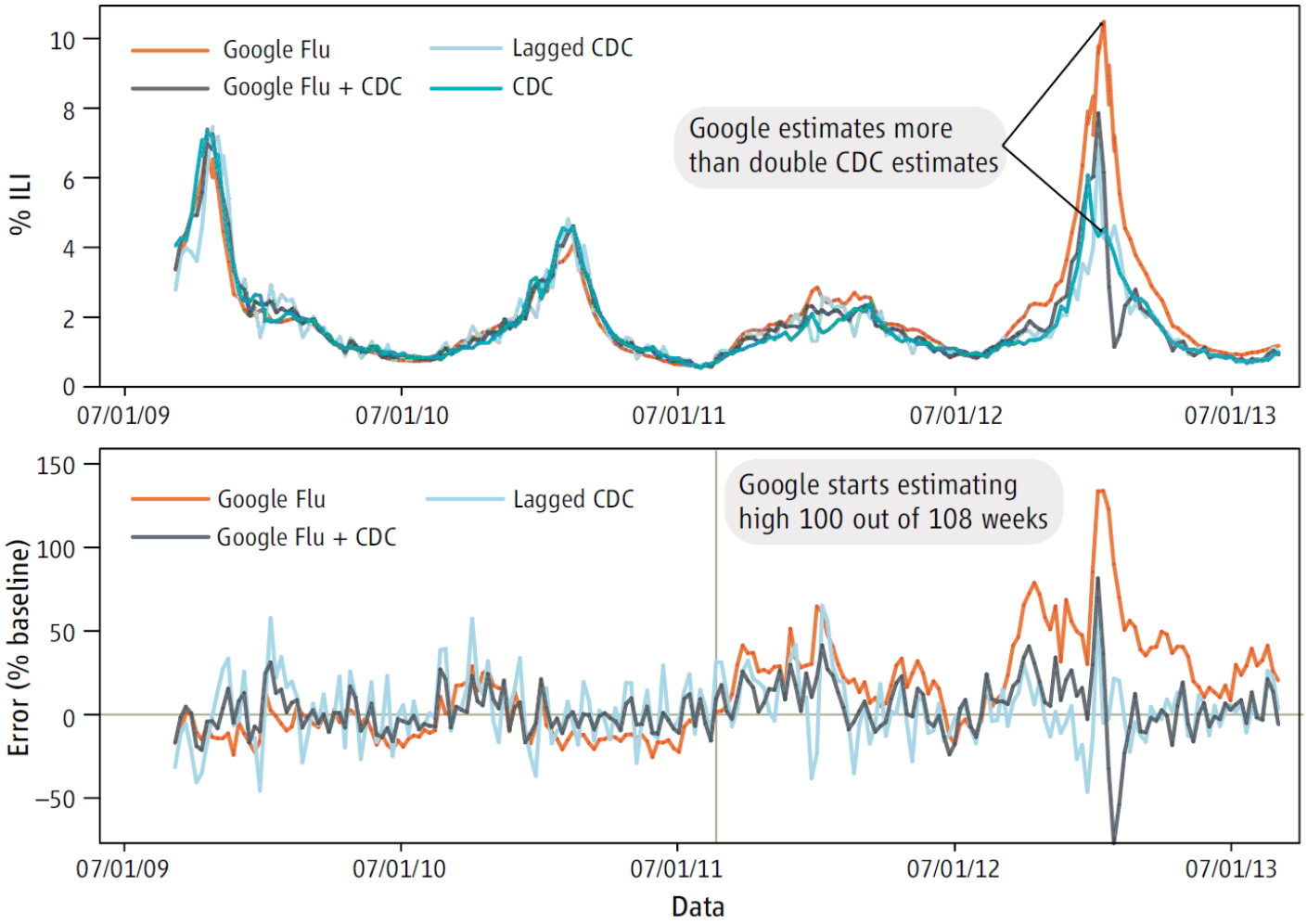
From poster child…



From poster child…to parable

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]



Google Flu Trends and “big data hubris”

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Big Data Hubris

“Big data hubris” is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

Google Flu Trends and “big data hubris”

The Parable of Google Flu: Traps in Big Data Analysis

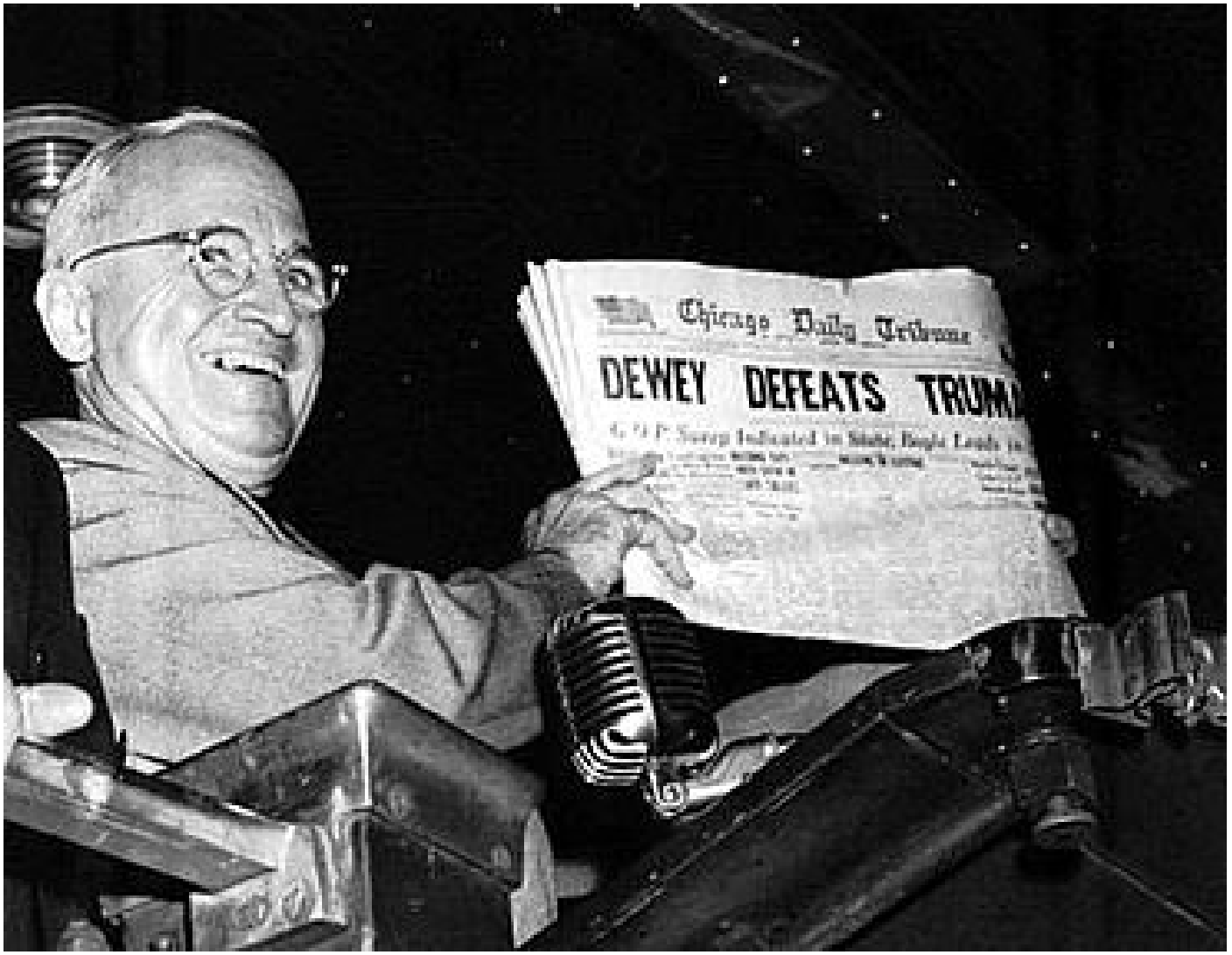
David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Big Data Hubris

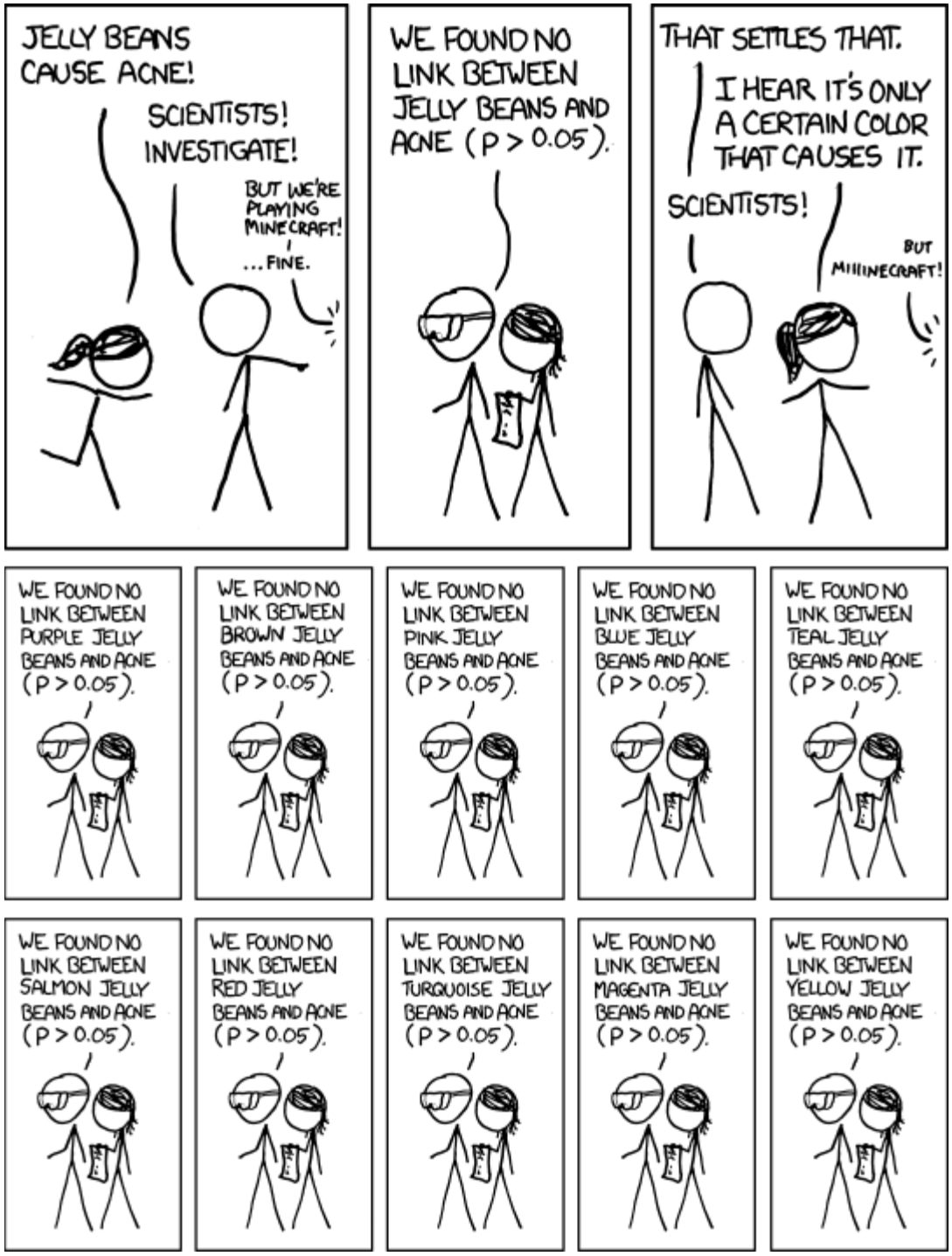
“Big data hubris” is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

Big data offer enormous possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables. We contend that these are the most exciting frontiers in studying human behavior. However, traditional “small data” often offer information that is not contained (or containable) in big data, and the very factors that have enabled big data are enabling more traditional data collection.

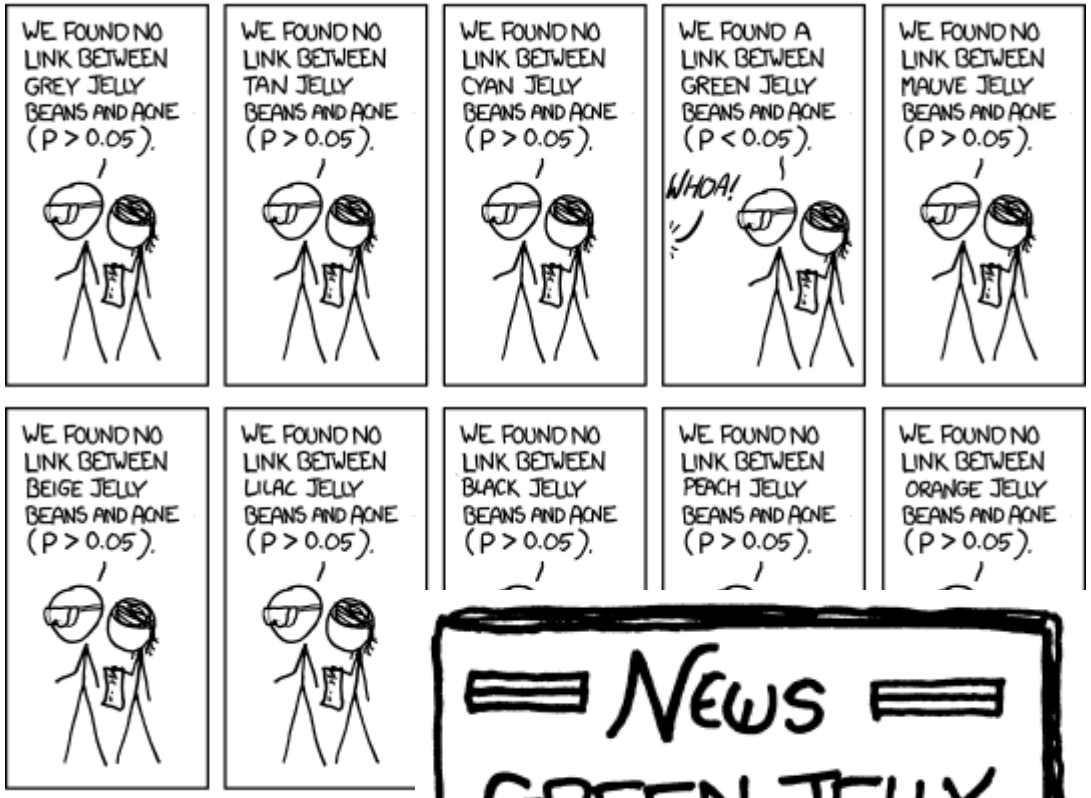
History doesn’t repeat itself but it does rhyme

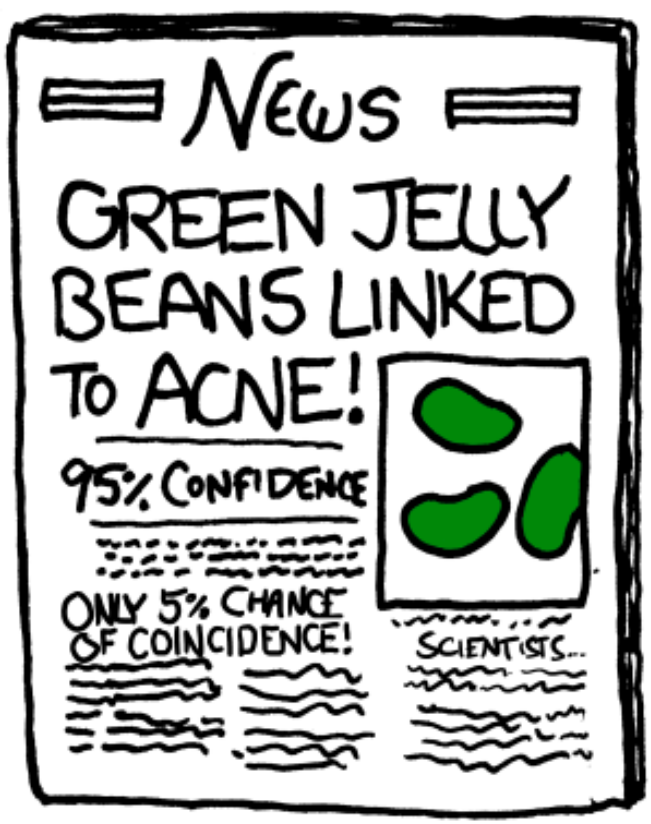


We've all bean there



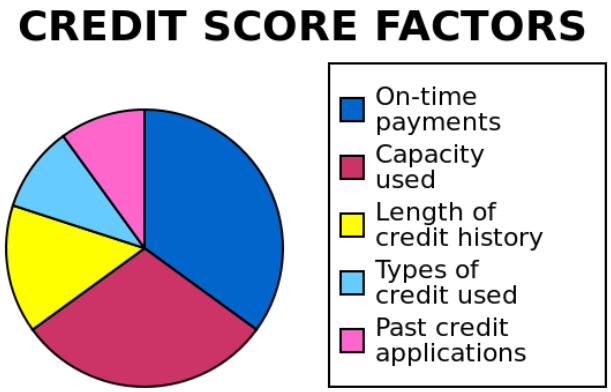
We've all bean there





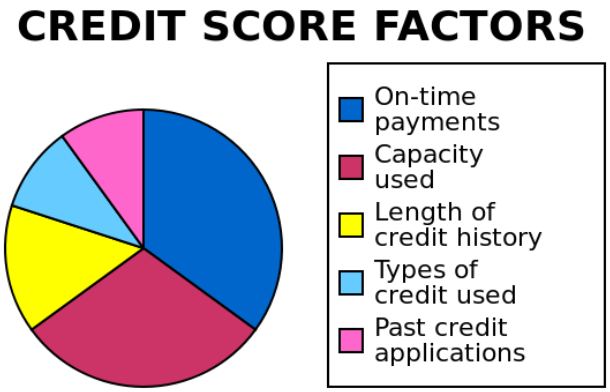
Big data and behavioral data in insurance today

An early example of business analytics

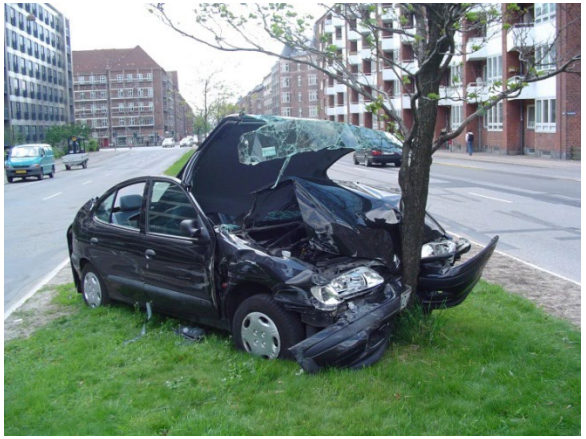


(This we know)

A more striking correlation

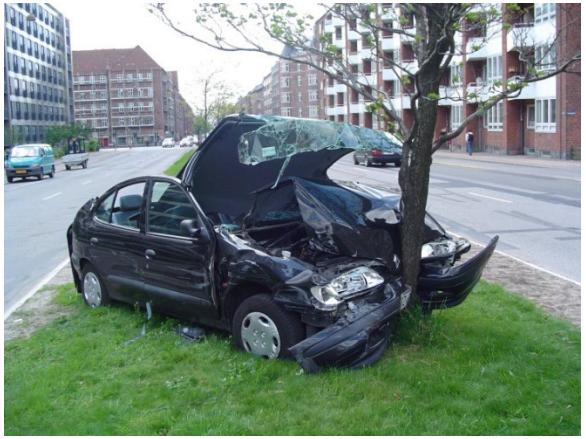


(!)



More food for thought

(!!)

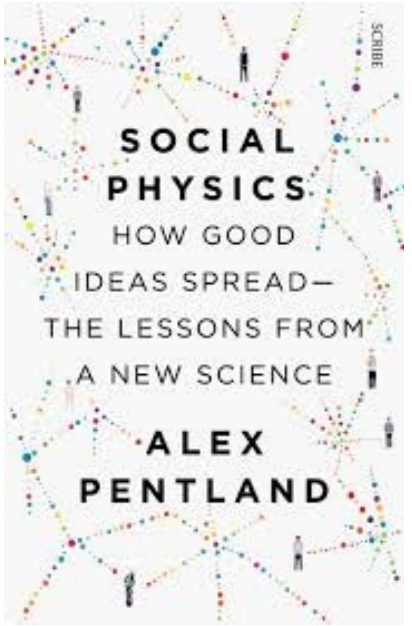


The real reason why big data is a big deal

“I believe that the power of Big Data is that it is information about people's behavior instead of information about their beliefs… This sort of Big Data comes from things like location data off of your cell phone or credit card, it's the little **data breadcrumbs** that you leave behind you as you move around in the world.

…those breadcrumbs tell… the story of your life... Big data is increasingly about real behavior, and by analyzing this sort of data, scientists can tell an enormous amount about you. They can tell whether you are the sort of person who will pay back loans. They can tell you if you're likely to get diabetes”

—Sandy Pentland, MIT Media Lab
“Reinventing Society in the Wake of Big Data”
edge.org conversation

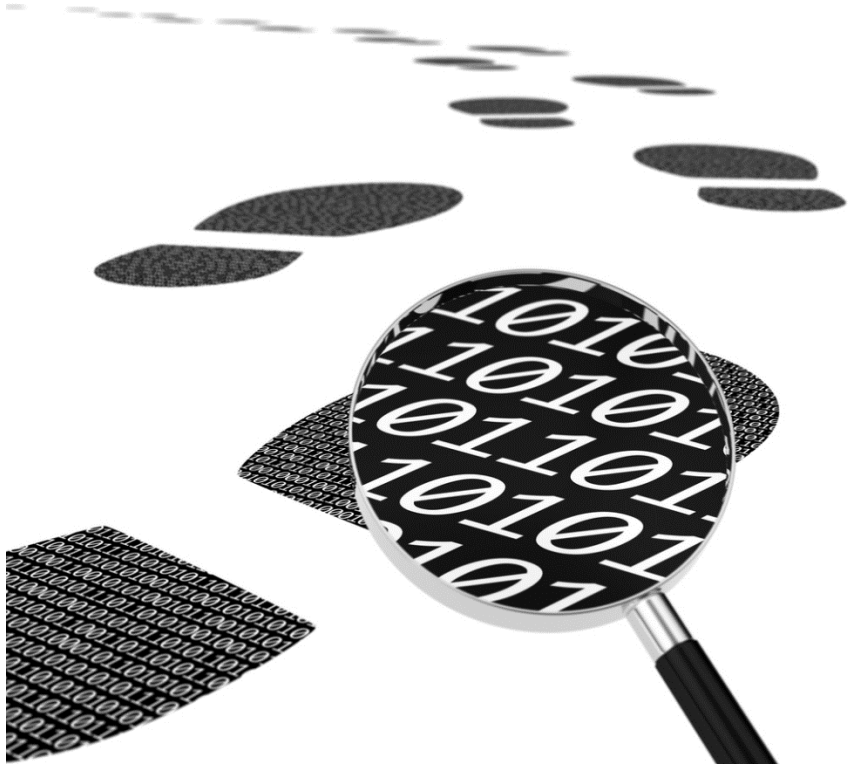


Looking ahead to …? New Mindset for Data Science

Digital breadcrumbs, today's cleaner digital exhaust

Our daily activities are increasingly ***digitally mediated****...*
We leave behind traces of

- How we drive
- What we buy
- What we eat
- What we watch, read
- What and how we opine
- Where we travel
- Who we know
- Who we call
- How we socialize
- How we surf the web
- Where we are going next
- What is really on our minds

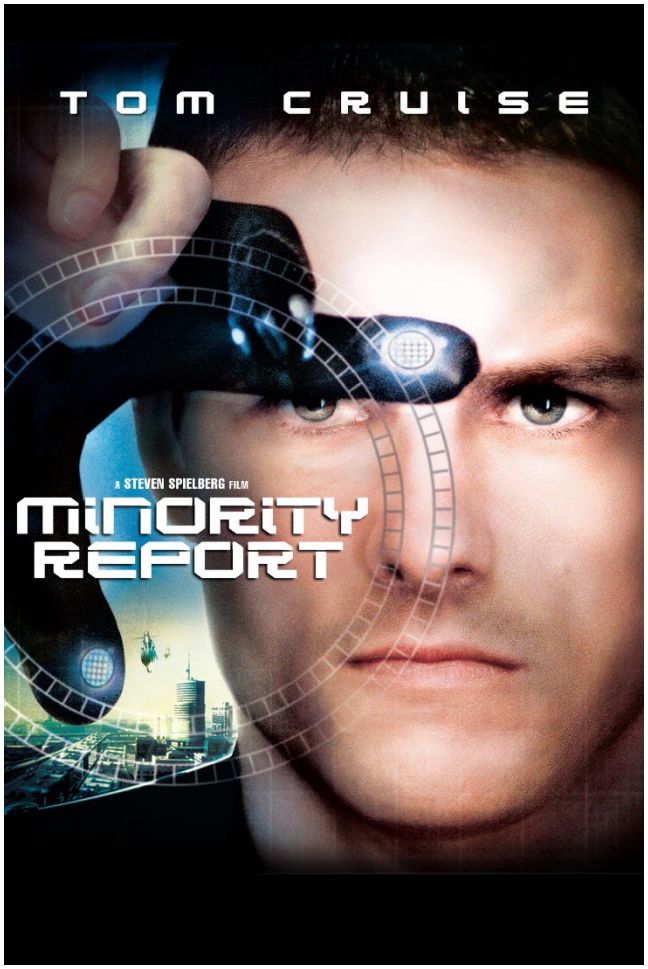


Evolving data privacy challenges

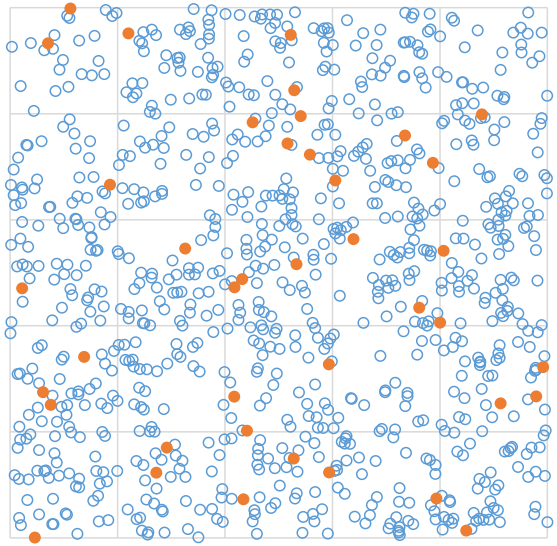
"We envision a very different privacy framework for the big-data age, one focused ***less on individual consent at the time of collection*** *and more on holding data users* ***accountable for what they do****. In such a world, firms will formally assess a particular reuse of data based on the impact is has on individuals...*

... sloppy assessments or poor implementation of safeguards will expose data users to legal liability, and regulatory actions such as mandates, fines, and perhaps even criminal prosecution."

-- Viktor Mayer-Schönberger and Kenn Cukier



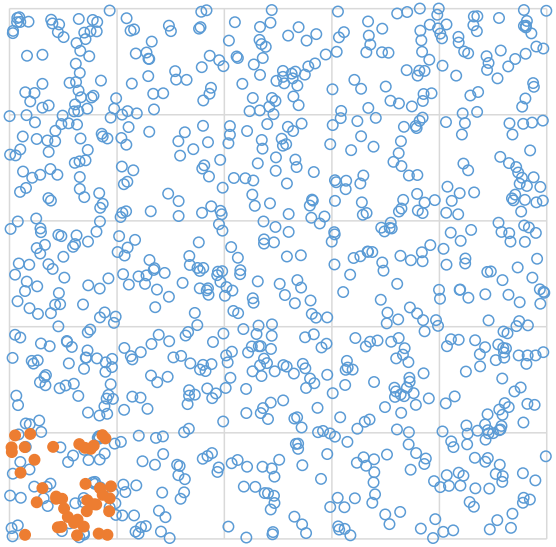
Big data and insurance: be careful what you wish for



Insurable

Old School

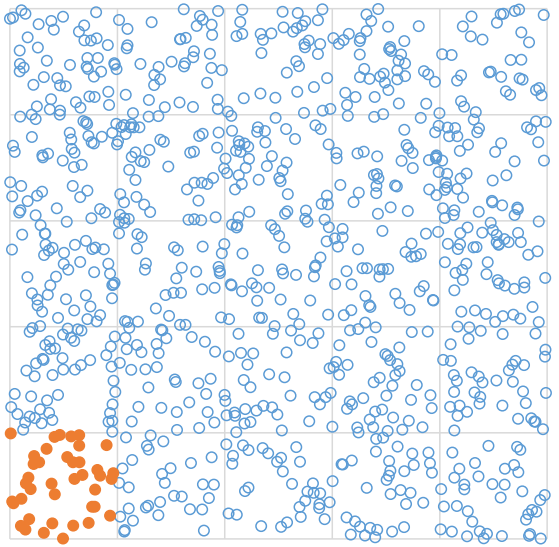
4% of 100%



Insurable, but expensive

Flood

50% of 8%



Not Insurable

Genetics

100% of 4%

The potential to transform everything

Driving behavioral change

Actuaries now use telematics data to better segment and price insurance policyholders in terms of their utilization and riskiness

But could the data be used to create new products and services…

… periodic or real-time reports that serve as behavioral nudges…

Ideas

- *Detailed feedback reports to help student drivers learn and older drivers stay behind the wheel longer and safer*

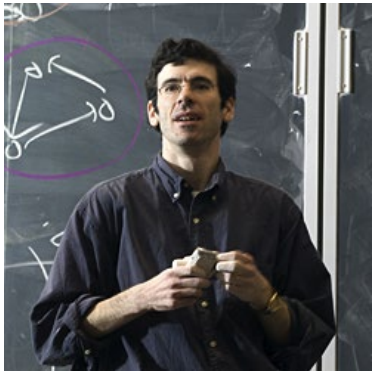
- *Feedback prompting carbon footprint improvements through peer effects*

A healthy regard for one's policyholders

The potential to transform everything

"The term itself is vague, but it is getting at something that is real… Big Data is a tagline for ***a process that has the potential to transform everything****."*

— Jon Kleinberg, Cornell University



nature International weekly journal of science

Computational social science: Making the links

From e-mails to social networks, the digital traces left by life in the modern world are transforming social science.

Contact Information

James Guszcza

US Chief Data Scientist
Deloitte Consulting, LLP

Los Angeles

jguszcza@deloitte.com

Stephen Mildenhall

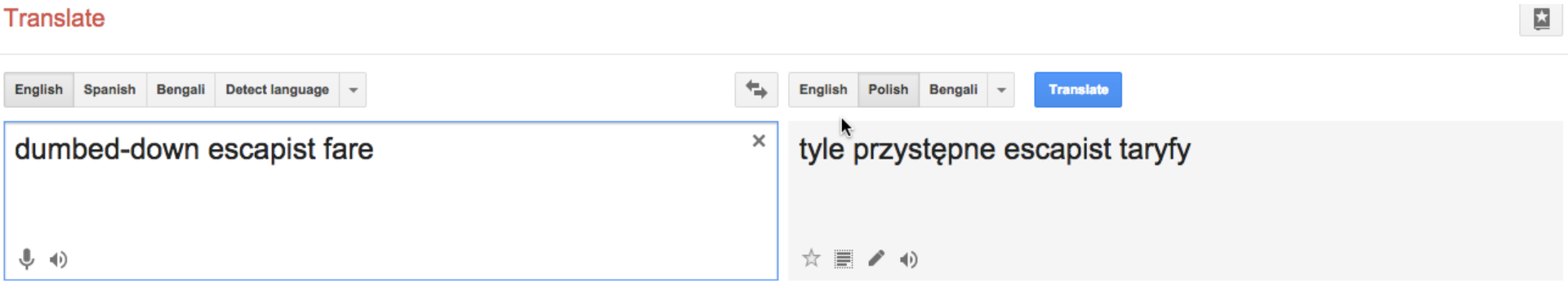
Global CEO of Analytics Aon Center for Innovation and Analytic, Singapore

stephen.mildenhall@aon.com

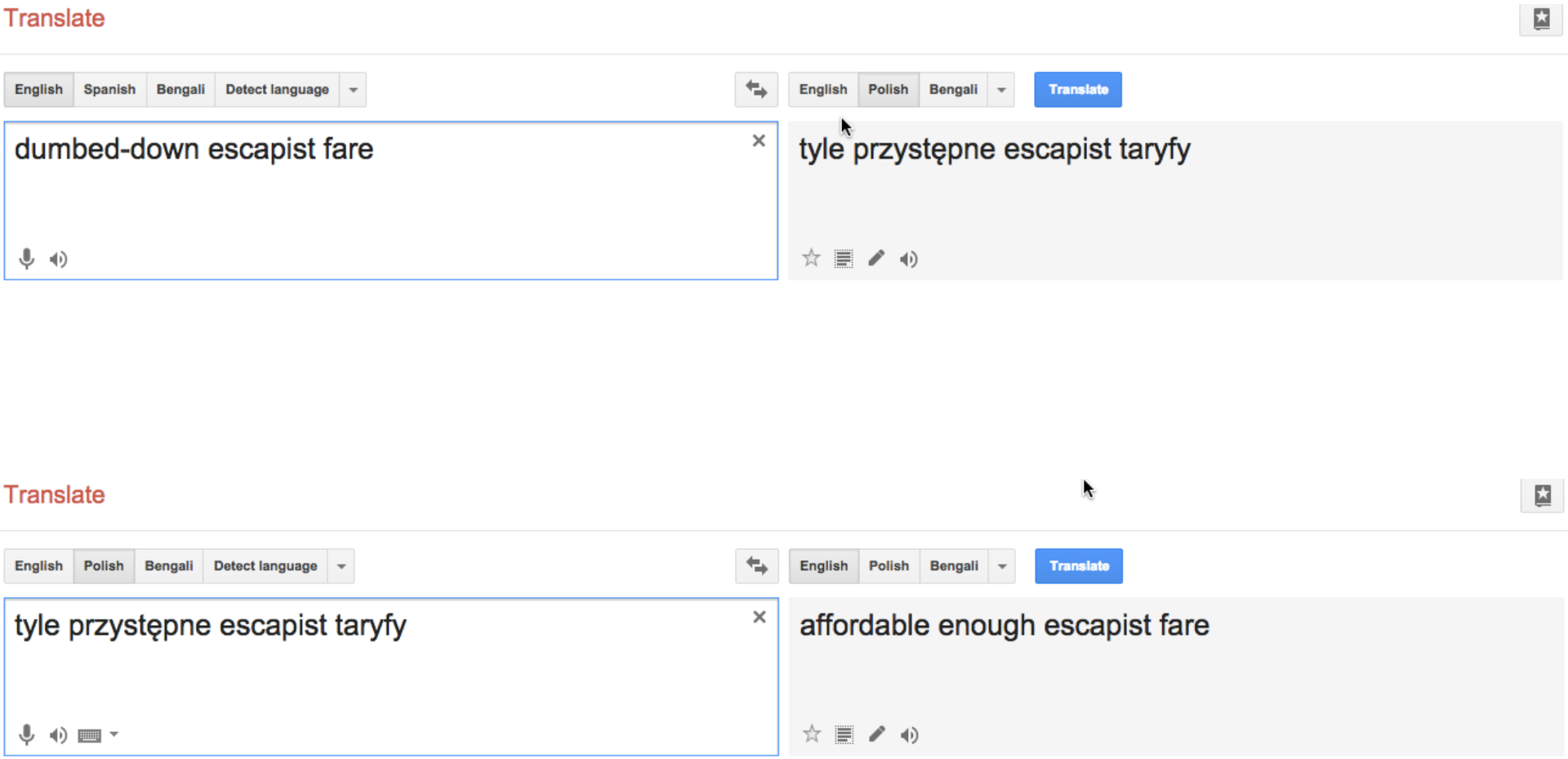
Sing cell: +65 9233 0670

US cell: +1 312 961 8781

Incredibly useful…



Incredibly useful… though not perfect



Credit as a psychobehavioral proxy

"... we delineate that basic chemical and psychobehavioral characteristics (e·g· a sensation-seeking personality type) are common to individuals exhibiting both higher insured automobile loss costs and poorer credit scores..."

© *The Journal of Risk and Insurance*, 2007, Vol. 74, No. 1, 23-63

BIOLOGICAL AND PSYCHOBEHAVIORAL CORRELATES OF CREDIT SCORES AND AUTOMOBILE INSURANCE LOSSES: TOWARD AN EXPLICATION OF WHY CREDIT SCORING WORKS

Patrick L. Brockett
Linda L. Golden

ABSTRACT

The most important new development in the past two decades in the personal lines of insurance may well be the use of an individual's credit history as a classification and rating variable to predict losses. However, in spite of its obvious success as an underwriting tool, and the clear actuarial substantiation of a strong association between credit score and insured losses over multiple methods and multiple studies, the use of credit scoring is under attack because there is not an understanding of *why* there is an association.