Big Data

Hmmm!

"I get depressed thinking of all the data out there, much of it yearning to be processed."

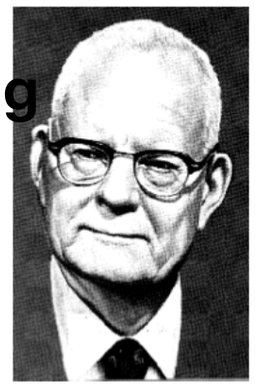
Data Analysis Has Been Around for a While

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

W.E. Demming

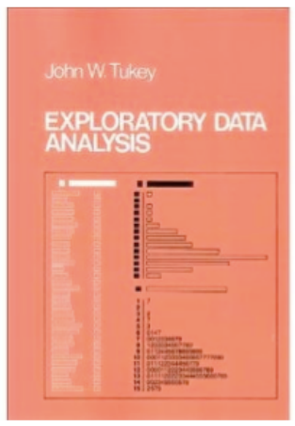


1958: "A Business Intelligence System"

Peter Luhn

1997: "Machine Learning"

1977: "Exploratory Data Analysis"

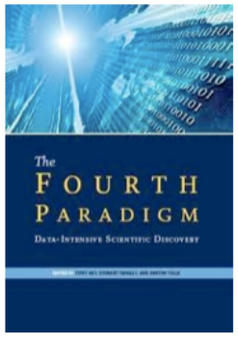


1989: "Business Intelligence"

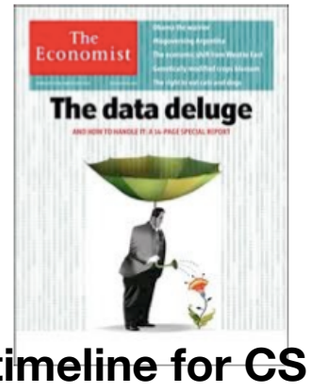
Howard Dresner

2010:"The Data Deluge"

2007:"The Fourth Paradigm"



2009:"The Unreasonable Effectiveness of Data"

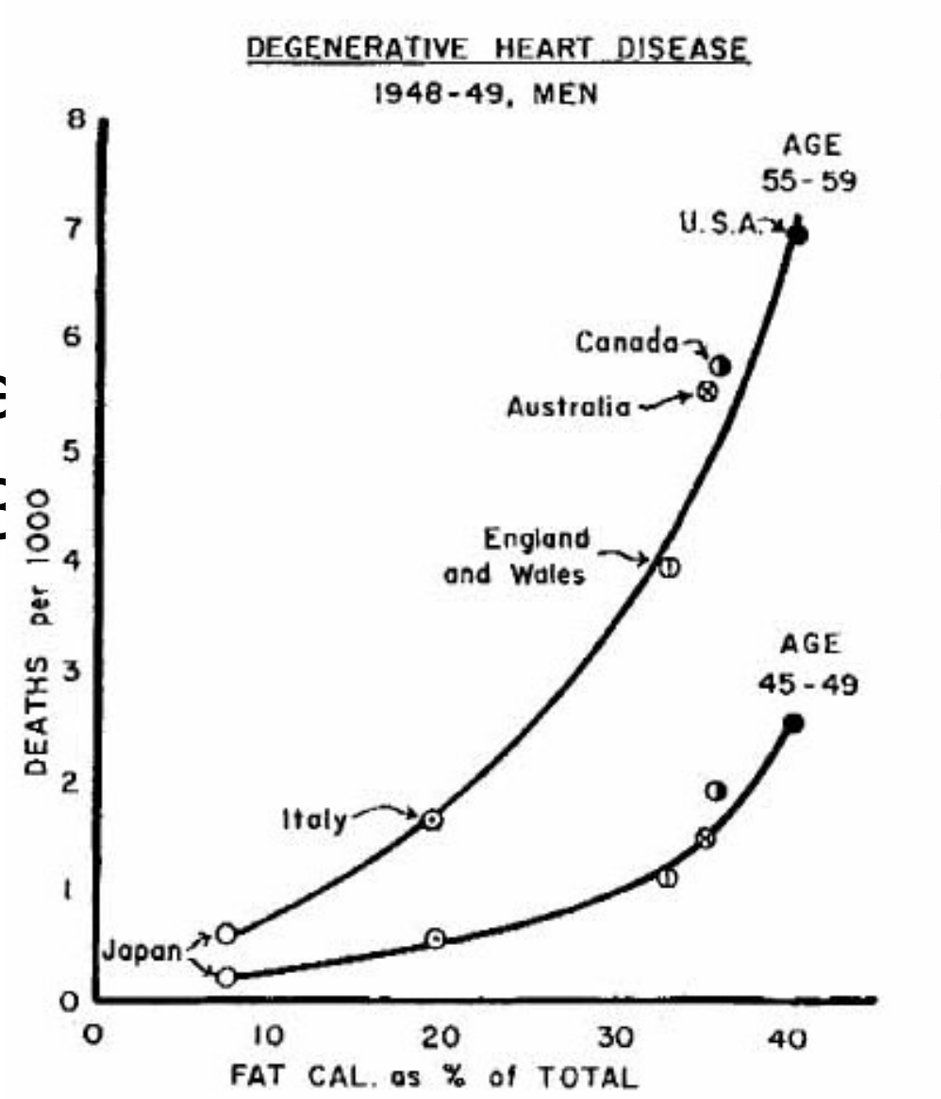


1996: Google

Google™

Abridged version of Jeff Hammerbacher's timeline for CS 194, 2012

Data makes everything clearer

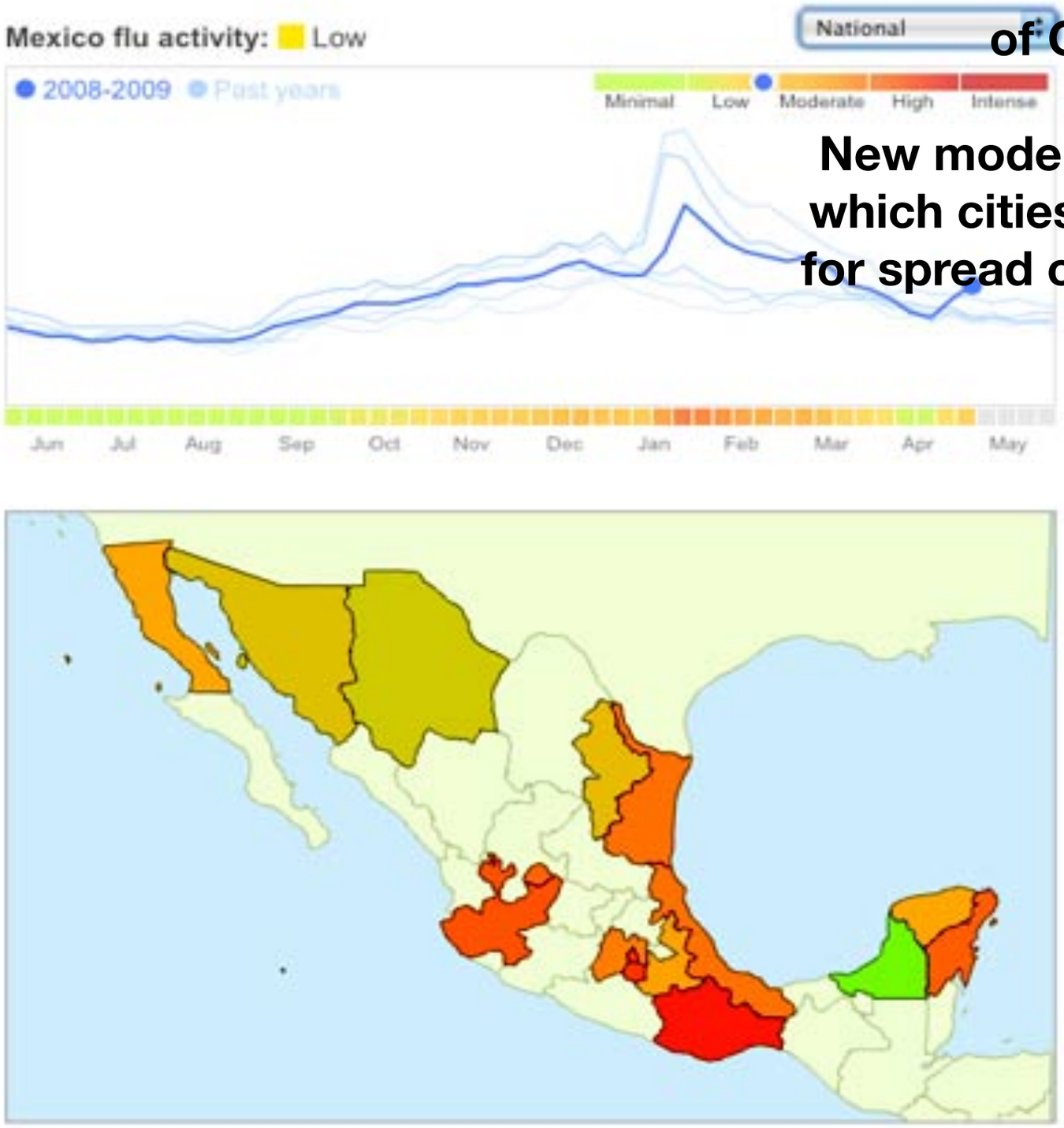


- Se cel k ,28)
- 13) yea

Data Science: Why all the Excitement?

e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data



New models are estimating
which cities are most at risk
for spread of the Ebola virus.

Why the all the Excitement?

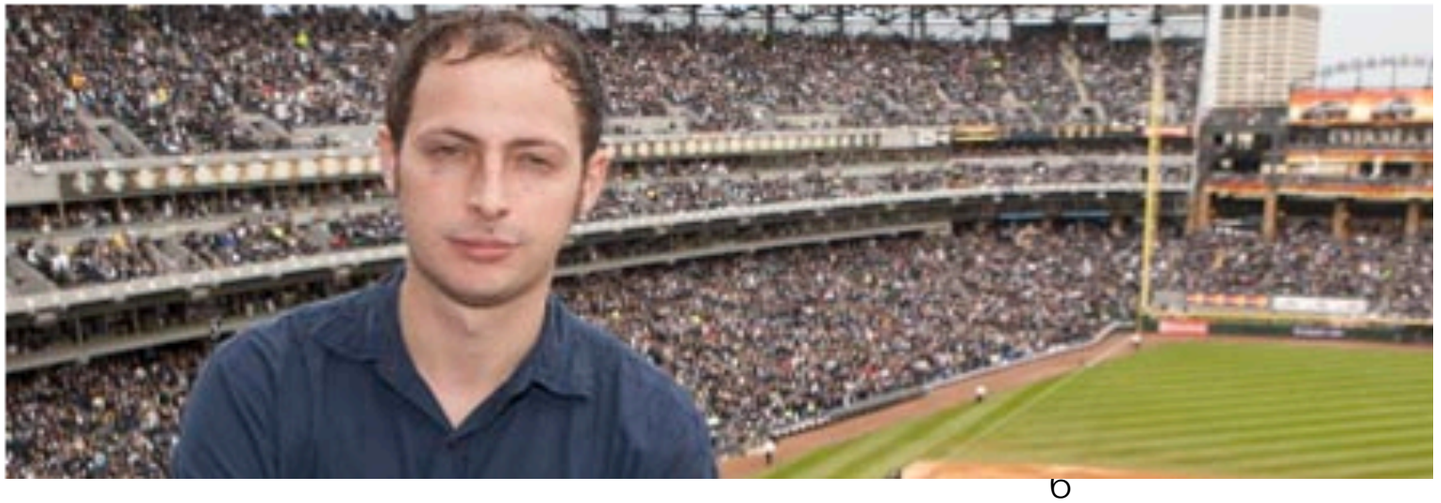
elections2012

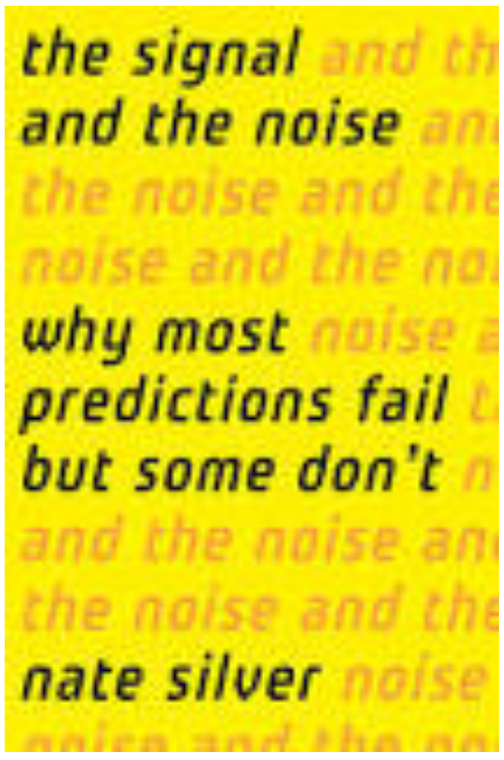
Live results | President | Senate | House | Governor | Choose your

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST





Data and Election 2012 (cont.)

- …that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, helped save the president's candidacy. In Chicago, the campaign recruited a team of behavioral scientists to build an extraordinarily sophisticated database

…that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. The power of this operation stunned Mr. Romney's aides on election night, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

New York TImes, Wed Nov 7, 2012

A history of the (Business) Internet: 1997

BackRub Search: university

university Search

BackRub Query Results

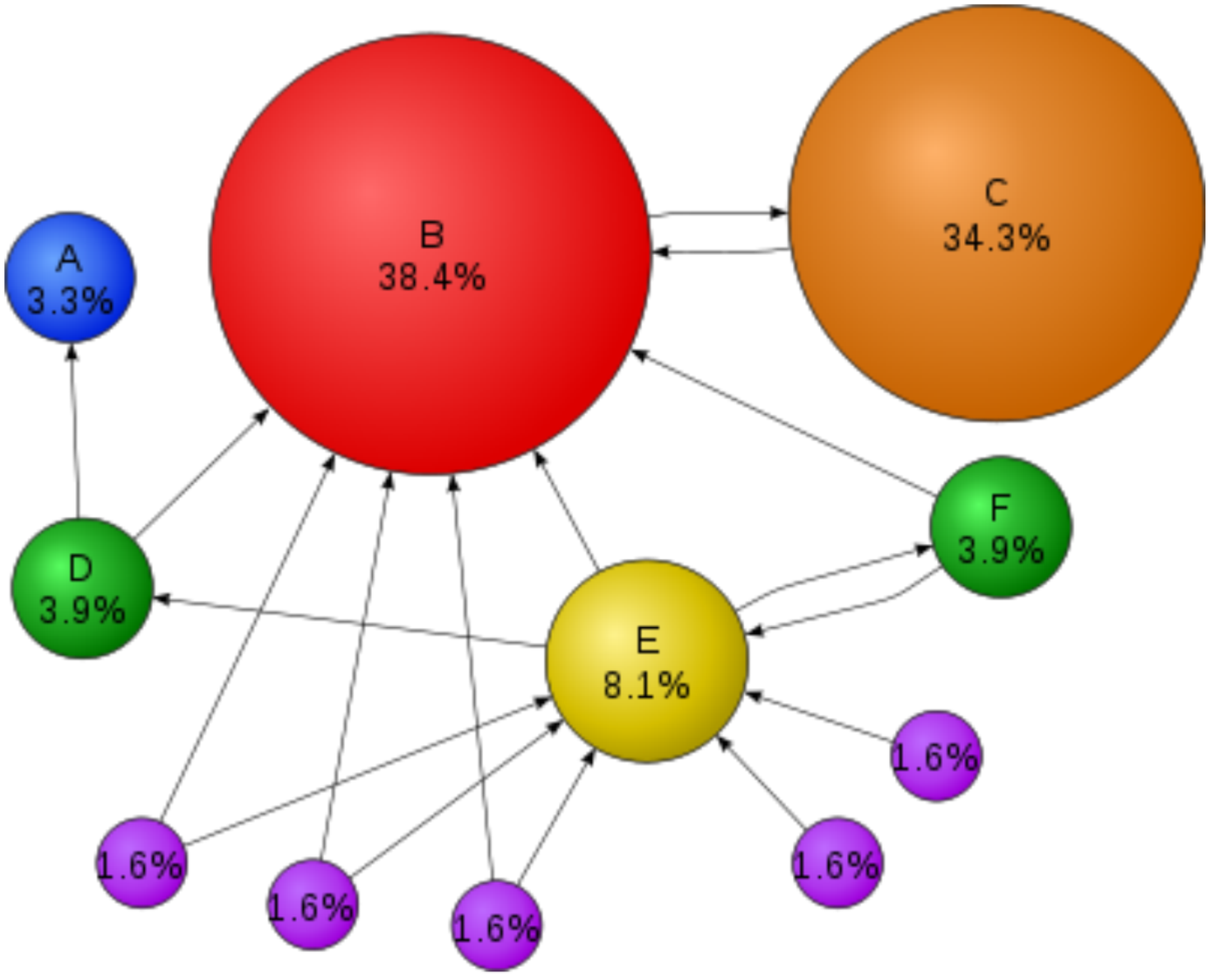
BackRub's Highest Ranked Sites

University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
694.687 8460 backlinks *12k - 10/25/96 - 11/1/96*

Stanford University Homepage
http://www.stanford.edu/
609.303 8857 backlinks *4k - none - 11/1/96*

Stanford University: Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
167.919 34 backlinks

Pagerank: The web as a behavioral dataset

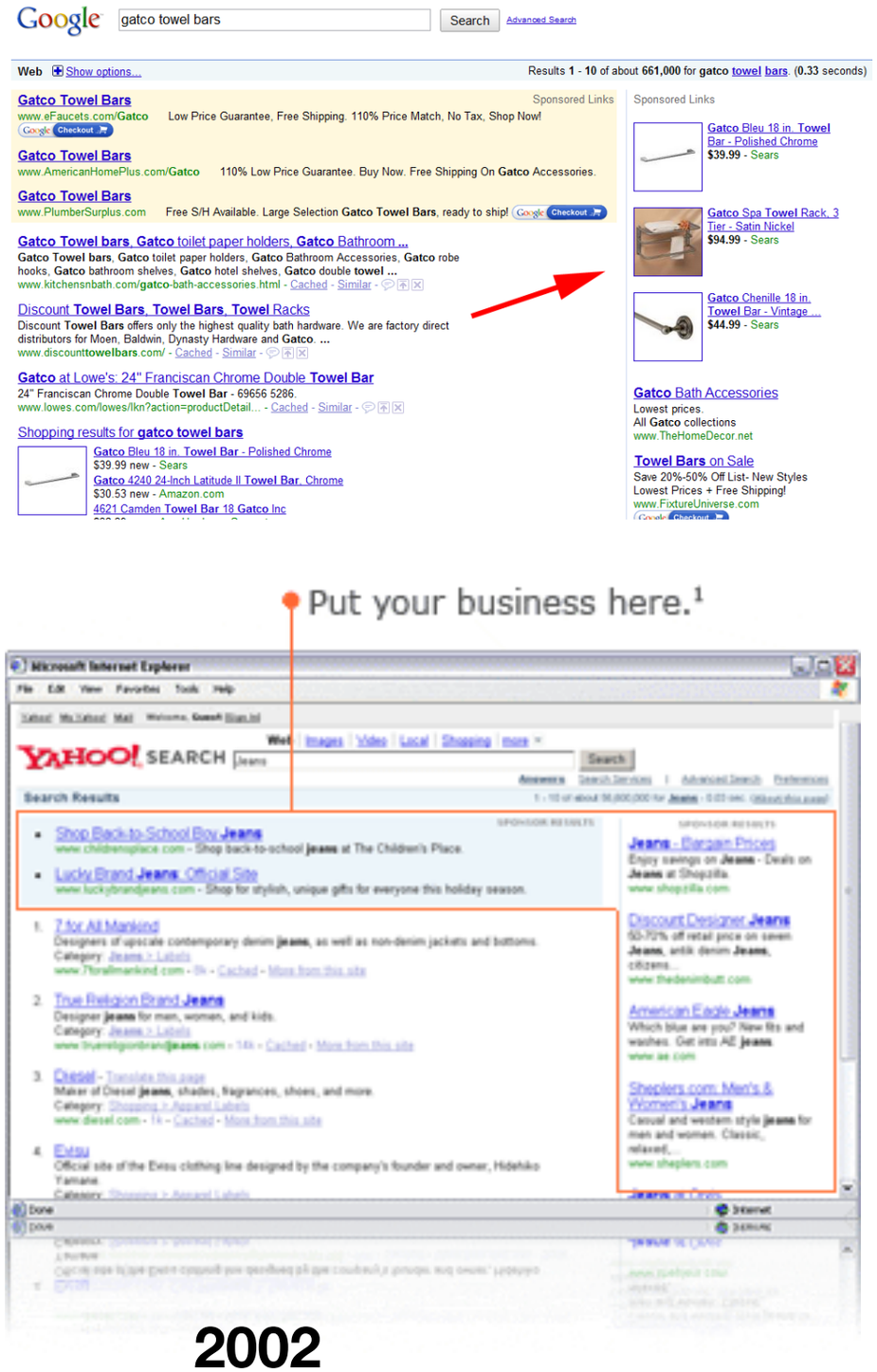


DB size = 50 billion sites

Google server farms
2 million machines (est)

1998 – sponsored search

Overture



2002

Sponsored search

- Google revenue around $50 bn/year from marketing, 97% of the companies revenue.

- Sponsored search uses an auction – a pure competition for marketers trying to win access to consumers.

- In other words, a competition for **models** of consumers – their likelihood of responding to the ad – and of determining the right bid for the item.

- There are around 30 billion search requests a month. Perhaps a **trillion events** of history between search providers.

TOP 20
Keyword Categories

Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.

#1 Insurance

24%

$54.91 Top CPC

auto **insurance** price quotes | Search

ca automobile **insurance**

building contents **insurance**

buy car **insurance** online

life **insurance** comparison quotes

Top Sample Queries

#2 Loans

$44.28 Top CPC

consolodate graduate student **loans** | Search

fixed home equity **loan** rates

cheapest homeowner **loans**

fixed rate secured **loans**

Top Sample Queries

12.8%

#3 Mortgage

9%

$47.12 Top CPC

#4 Attorney

3.6%

$47.07 Top CPC

#5

3.2%

3%

2.5%

Site Targeting - AdWords buyer targets specific site which has specified CPM "Ad Placement"

All Text ads and all Image ads are referred to as "Display Ads".

Google Personalized User Data

Interest Based Ads

Google AdSense

Search Network
Search Ad Pool
Ad serving Algo

Content Network

Google SERPs
Aubuchon

Publisher
(AdSense Partner)

URL Block

Webpage
Section targeting
Ads CTR

CAN Block

Subjecy Category Block (11)

Specify Text Only

Ad Placement

Google AdWords

CPC Ads
Must be clicked
typ. eCPM= $5
CTR= 5%, EPC= 0.10

Contextual Targeting
Text Ad
Text Ad
Image Ad

CPM Ads
Pay Per Impression
Clicks pay nothing
typ. eCPM= $1 (untargeted)
(email to block ALL CPM ads)

Site Targeting
Graphic Ads
Flash Ads
Video Ads
Text Ads

Advertiser
(AdWords Buyer)

CPC -
Broad match
Exact match
Geo. area

Search Bid
Content Bid
Content Bid
- Categ. Targ.?

CPM -
- Image Ads
- Text Ads

Category Targeting
(sex, age, $)

Untargeted
Tarrgeted

(1) (2) (3) (4)

(5) Competitive Filter

Interest Based Ads may not be blocked.

Content Ad Pool

AdSense Content Ad Serving Algo

(10) Override Test Control

block CAN ads
block CPC ads
block CPM ads

Ad Review Center

(9)

To block all "Site Targeting", never specify "Ad Placement"

37 CANs
Certified Ad Networks
(US, Europe only)

Category Filtering
Competitive Filter
- Placement
- Contextual
- User-targeted.

(6)

All Image Ads
- Graphic Ads
- Flash Ads
- Video Ads

(7)

CPM Ads
CPA Ads

(8)

User Data - ?
Webpage Data - ?

Ad Placement - Publisher makes ad targetable in AdSense channel, enabling CPM "Site Targeting"

Source: http://www.vaughns-1-pagers.com/internet/google-certified-networks.htm

filename = adsense-ad-options.gif.v17

Data Makes Everything Clearer?

Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]

1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

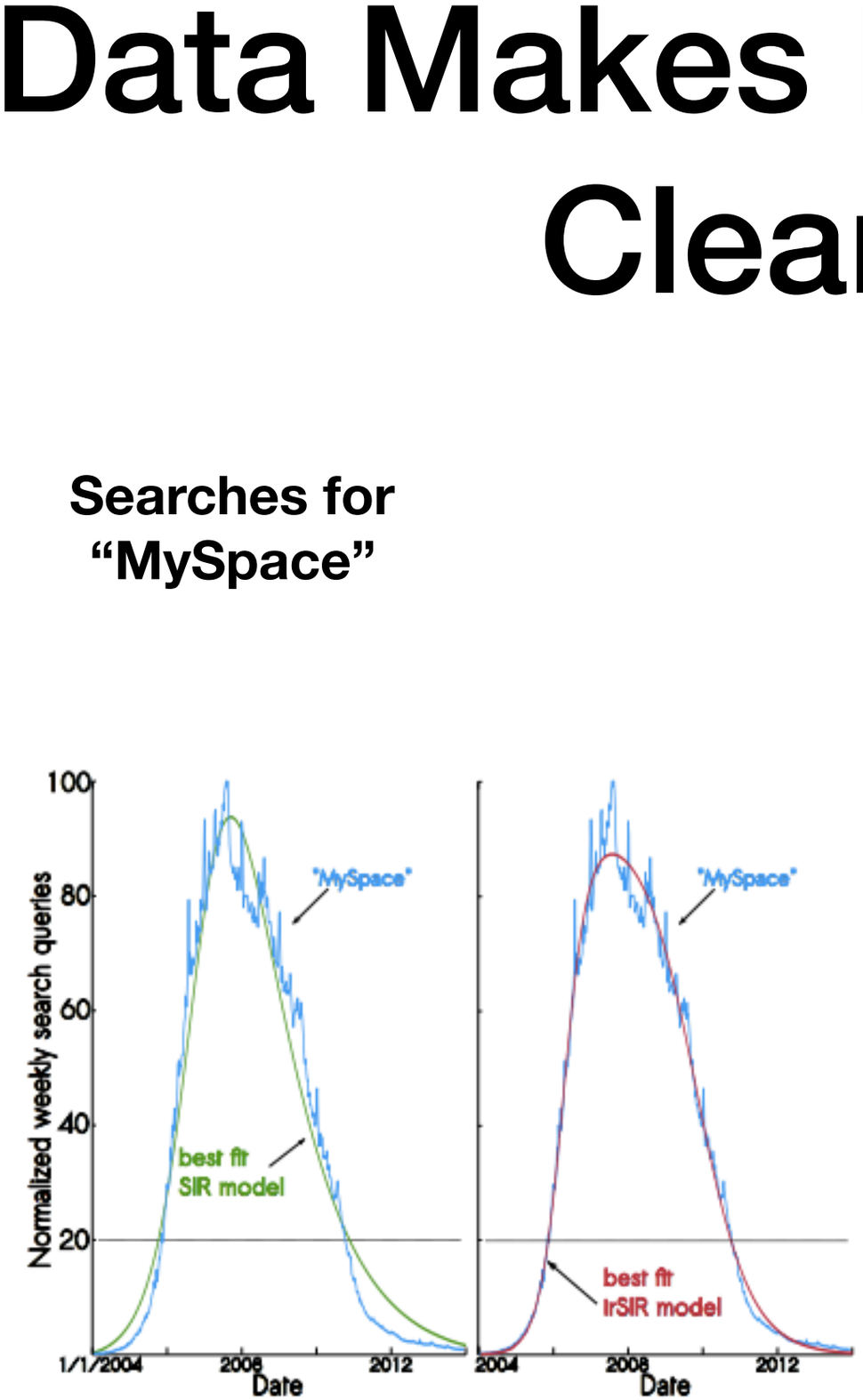
*** E-mail: Corresponding spechler@princeton.edu**

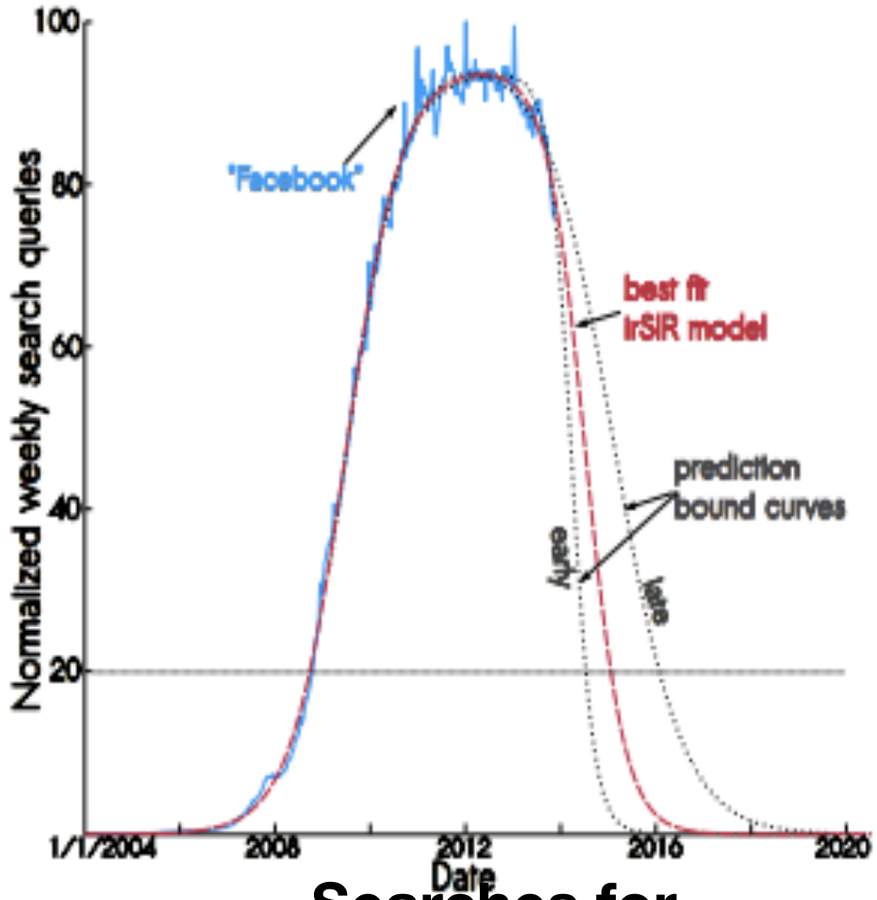
Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

Data Makes Everything Clearer

Searches for "MySpace"



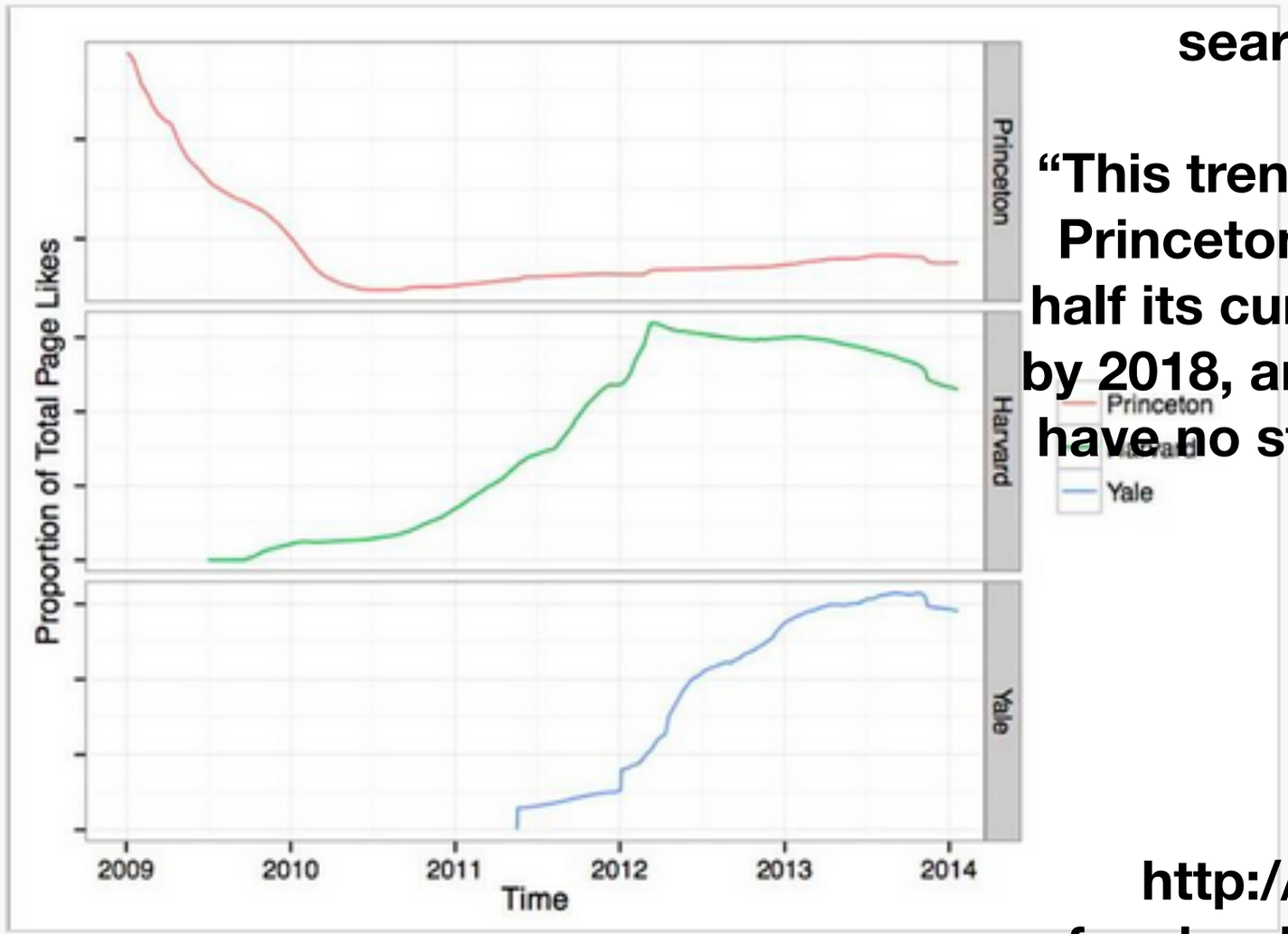


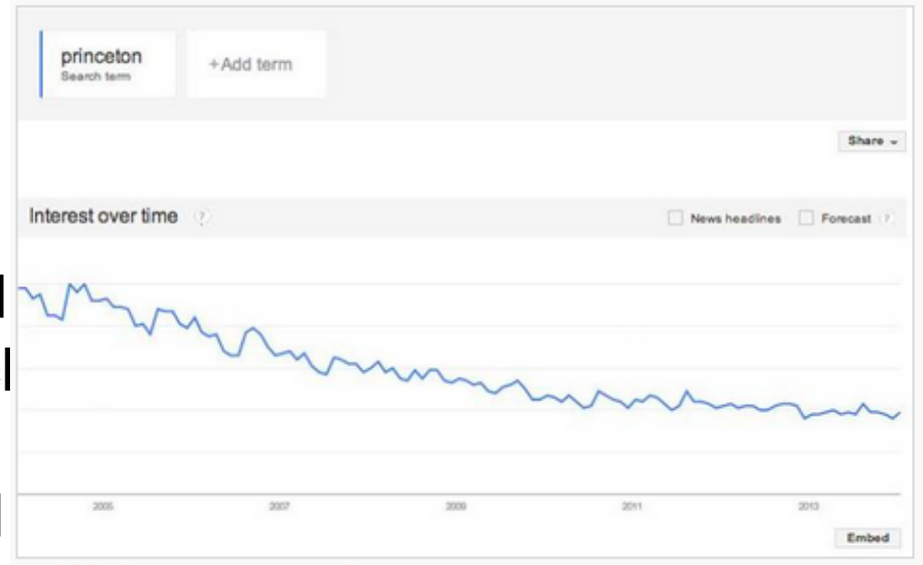
Searches for "Facebook"

Figure 3: **Data for search query "Myspace" with best fit (a) SIR and (b) irSIR models overlaid.** The search query data are normalized such that the maximum data point corresponds to a value of 100.

Data Makes Everything Clearer

In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:





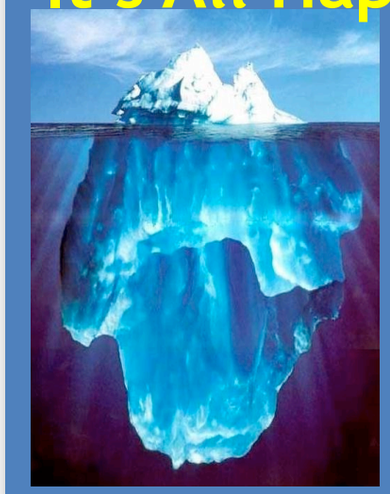
and based on search

"This trend Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…

http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/

“Big Data” Sources

It’s All Happening On-line



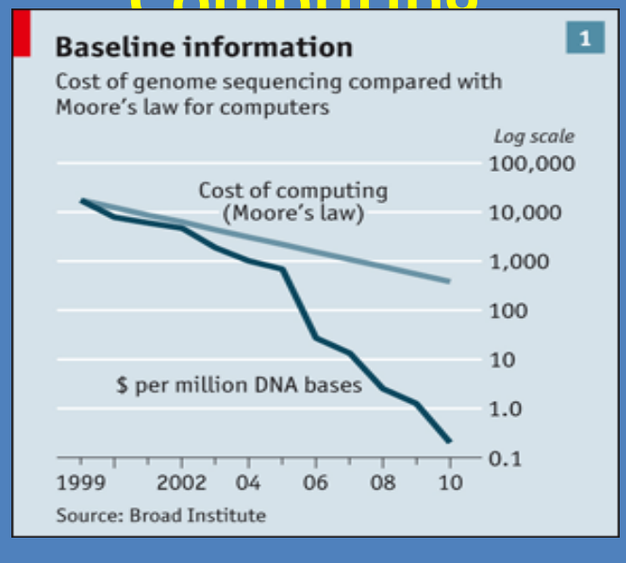
User Generated (Web & Mobile)



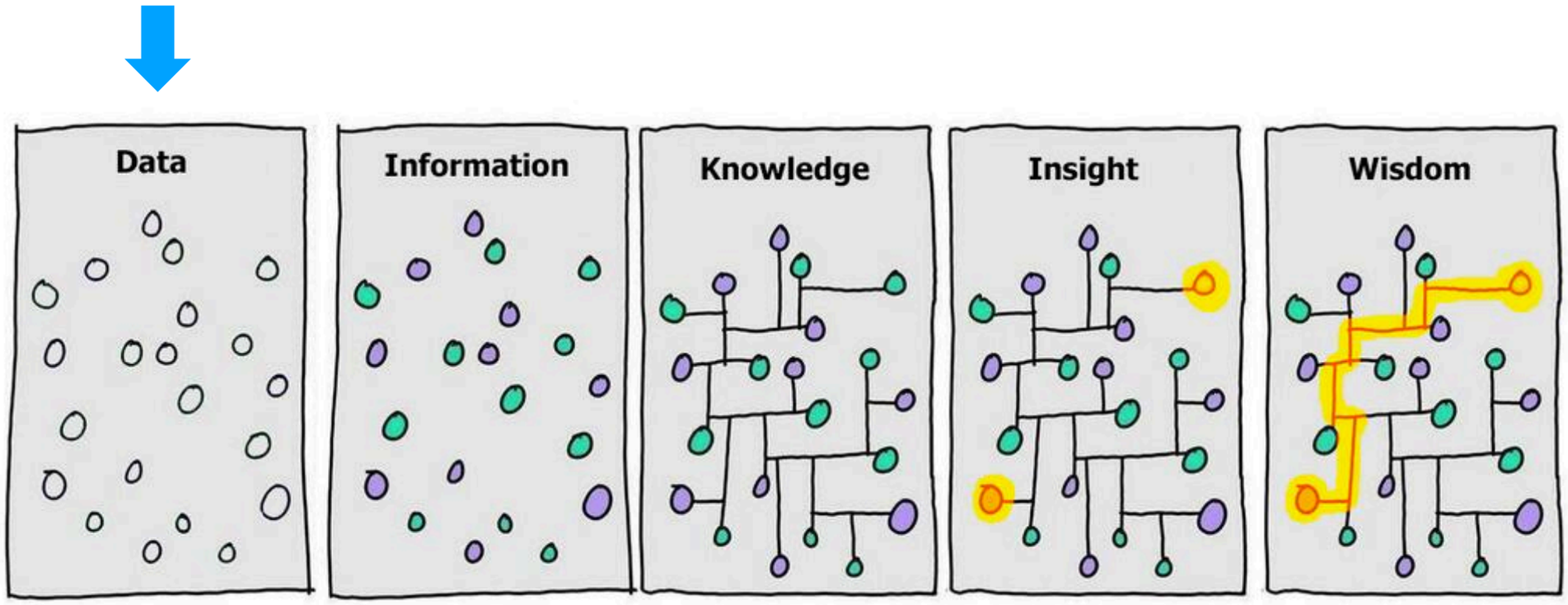

…..


Internet of Things / M2M

Health/Scientific Computing



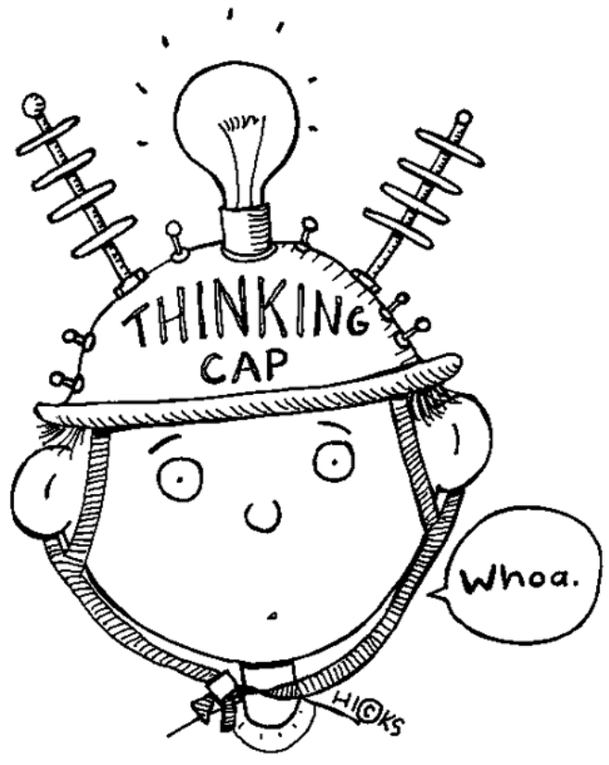
The Information Continuum



Cartoon by David Somerville, based on a two pane version by Hugh McLeod

Traditional Research

- Generate a hypothesis.
- Assemble a sample population and a control group.
- Expose both to an intervention (drug, treatment, etc.).
- Do statistical analysis to identify causal relationships.
- Rinse and repeat…



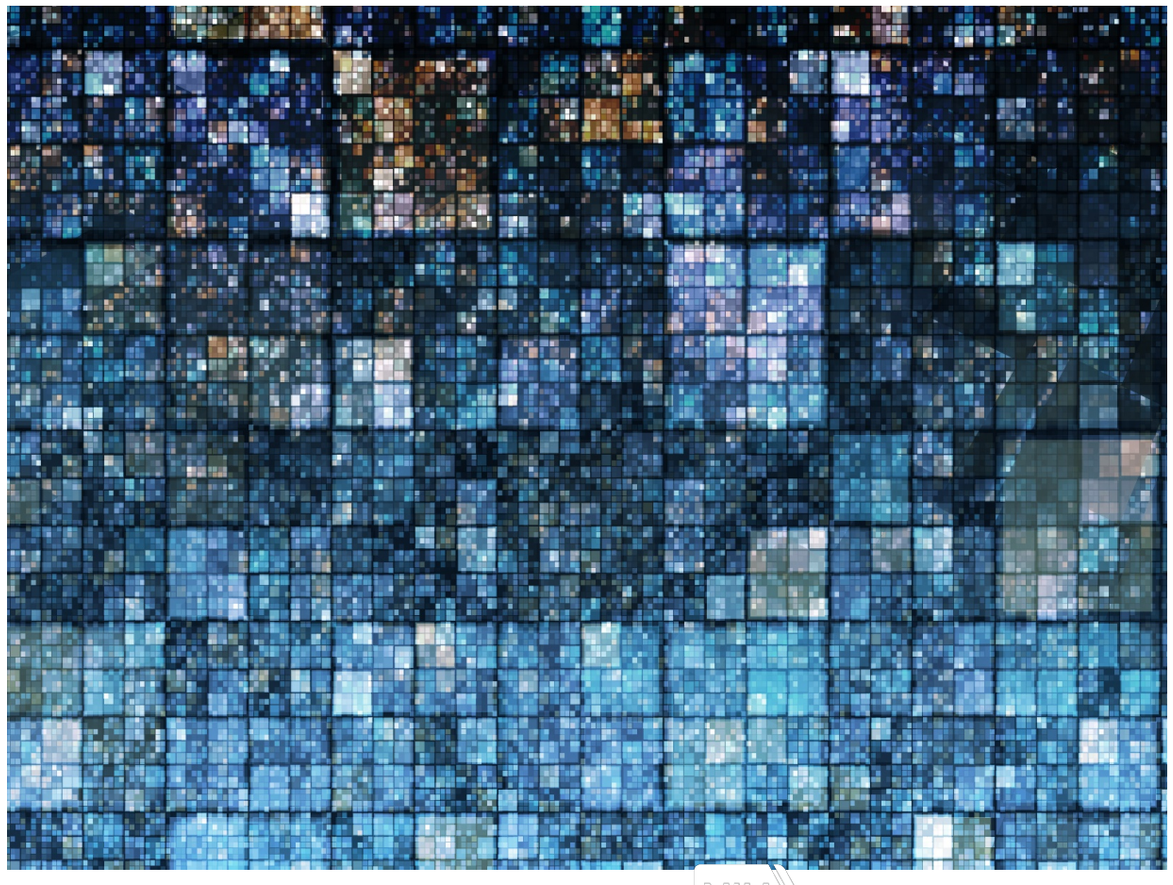
Types of Data

- Quantitative Data
  - Measurable
    - Collected through measuring things that have a fixed reality
    - Close ended

- Qualitative Data
  - Descriptive
    - Collected through observation, field work, focus groups, interviews, recording or filming conversations
    - Open ended

Big Data

- Data that is too large or too complex to be managed using traditional data processing, analysis, and storage techniques.



The 4 V's of Big Data

Volume
The amount of data

Variety
The types of data

Velocity
The frequency of data

Veracity
The quality of data

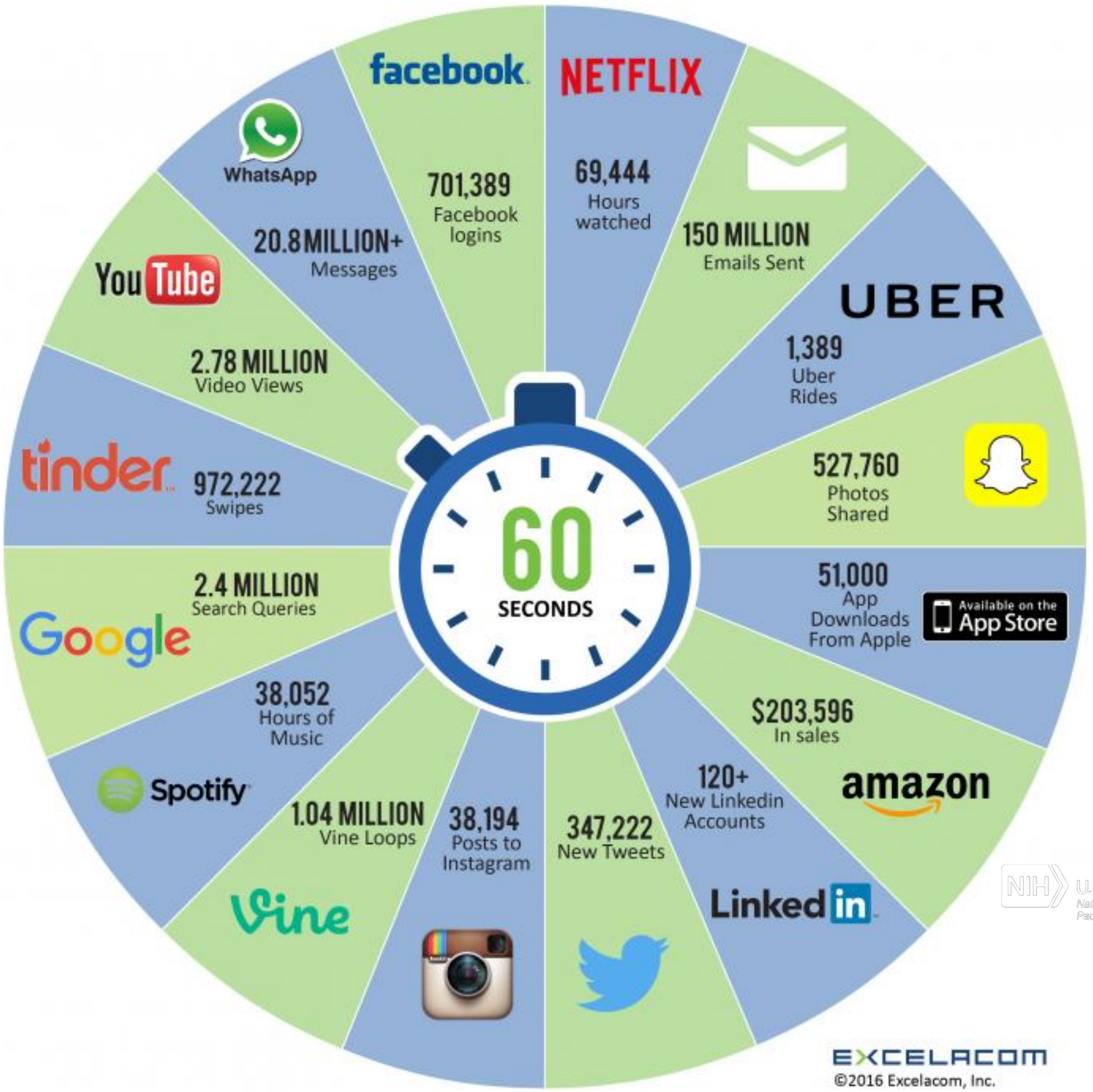
Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Volume: scale of data

- 90% of today's data has been created in just the last 2 years
- Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs
- 40 zettabytes (4o trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth
- Most companies in the US have over 100 terabytes (100,000 gigabytes) of data stored

Variety: different forms of data

2016 INTERNET MINUTE?

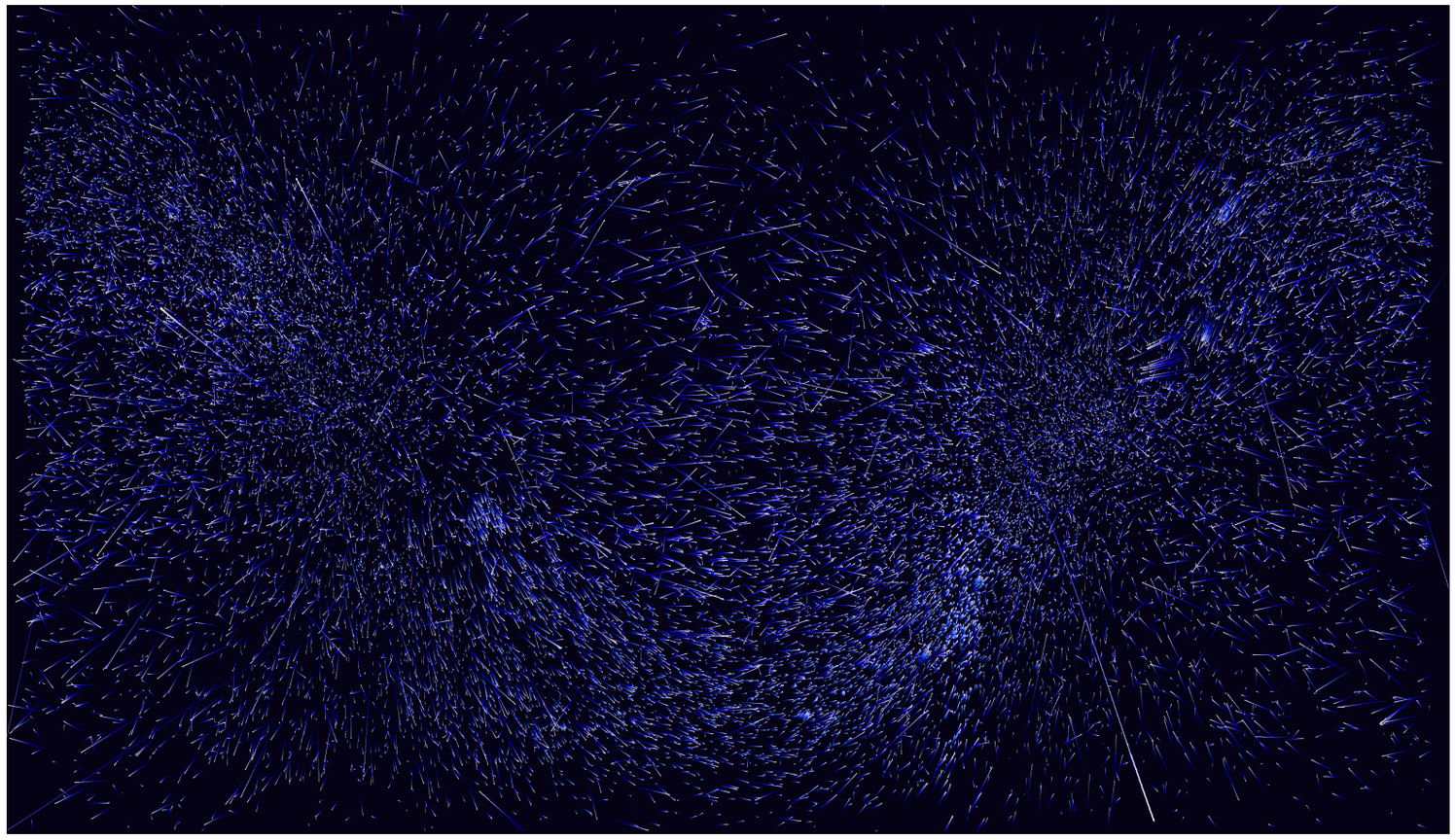


EXCELACOM

Velocity: analysis of streaming data

Veracity: trustworthiness of data

- Origin
- Authenticity
- Trustworthiness
- Completeness
- Integrity



The 4 V's of Big Data

Volume
The amount of data

Value

Variety
The types of data

Velocity
The frequency of data

Veracity
The quality of data

Or…..

- Volume:
  - How much data is really relevant to the problem solution? Cost of processing?
  - *So, can you really afford to store and process all that data?*
- Velocity:
  - Much data coming in at high speed
  - Need for streaming versus block approach to data analysis
  - *So, how to analyze data in-flight and combine with data at-rest*
- Variety:
  - A small fraction is structured formats, Relational, XML, etc.
  - A fair amount is semi-structured, as web logs, etc.
  - The rest of the data is unstructured text, photographs, etc.
  - *So, no single data model can currently handle the diversity*
- Veracity: cover term for …
  - Accuracy, Precision, Reliability, Integrity
  - *So, what is it that you don't know you don't know about the data?*
- Value:
  - How much value is created for each unit of data (whatever it is)?
  - *So, what is the contribution of subsets of the data to the problem solution?*

Big Data and Research

Big Data Mining

- Collect Big Data or obtain access to a repository.

- Perform data analysis to explore patterns (pattern recognition, predictive analytics).

- Identify potential correlations.

- Good enough!

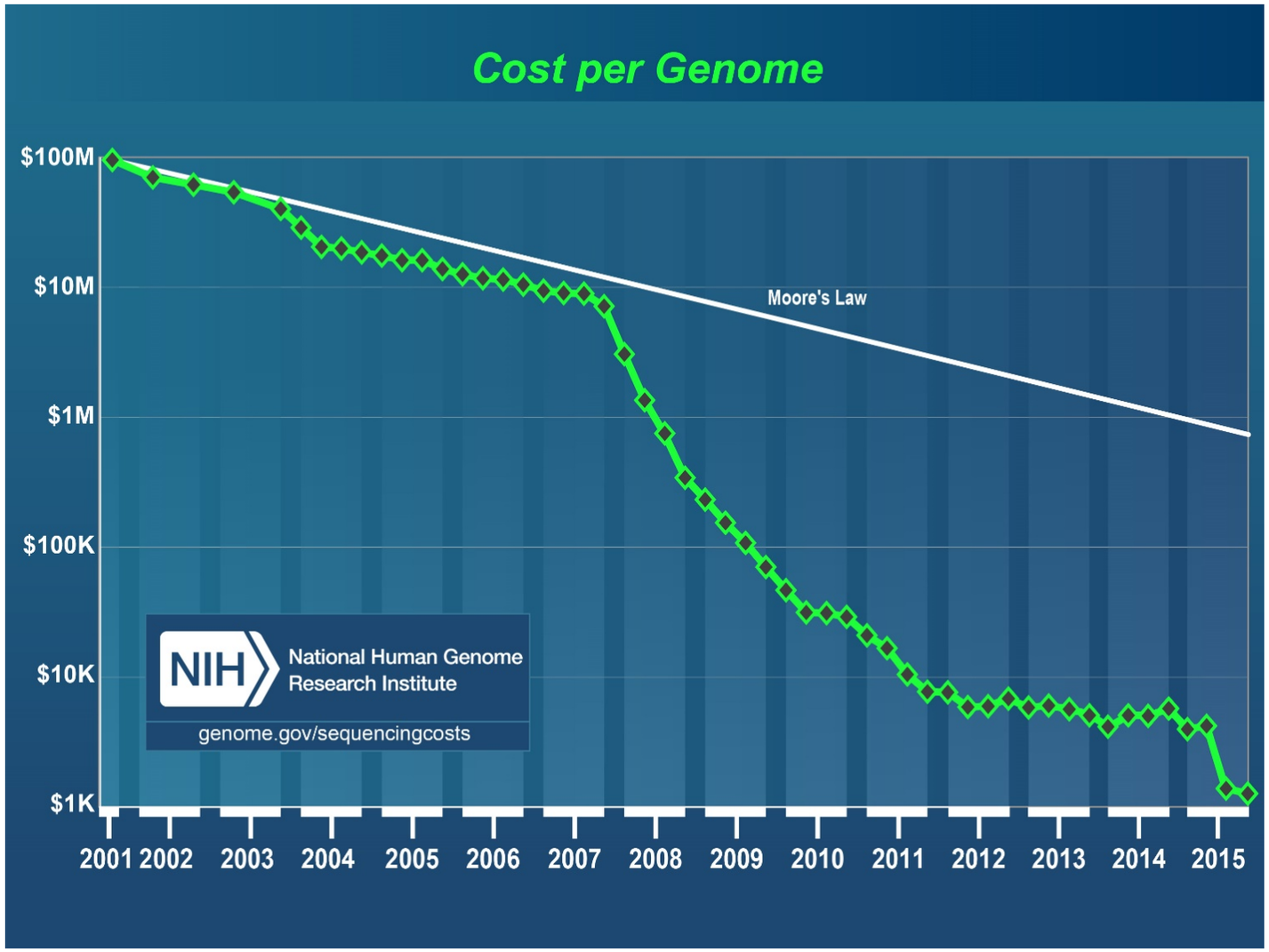
©Rina Piccolo

Big Data in Health Care

- Faster and cheaper technology and data storage
- Widespread sensing devices
- An increase in “born” digital data
- Greater availability of data via repositories
- Data sharing mandates

Faster and cheaper technology and data storage

- The cost to sequence a whole human genome sequence has fallen from +$100 million to less than $1,000 over the past 15 years.



Sensing devices

- Smartwatches
- Smart jewelry
- Fitness trackers
- Sport watches
- Smart glasses
- Smart clothing…

An increase in “born” digital data

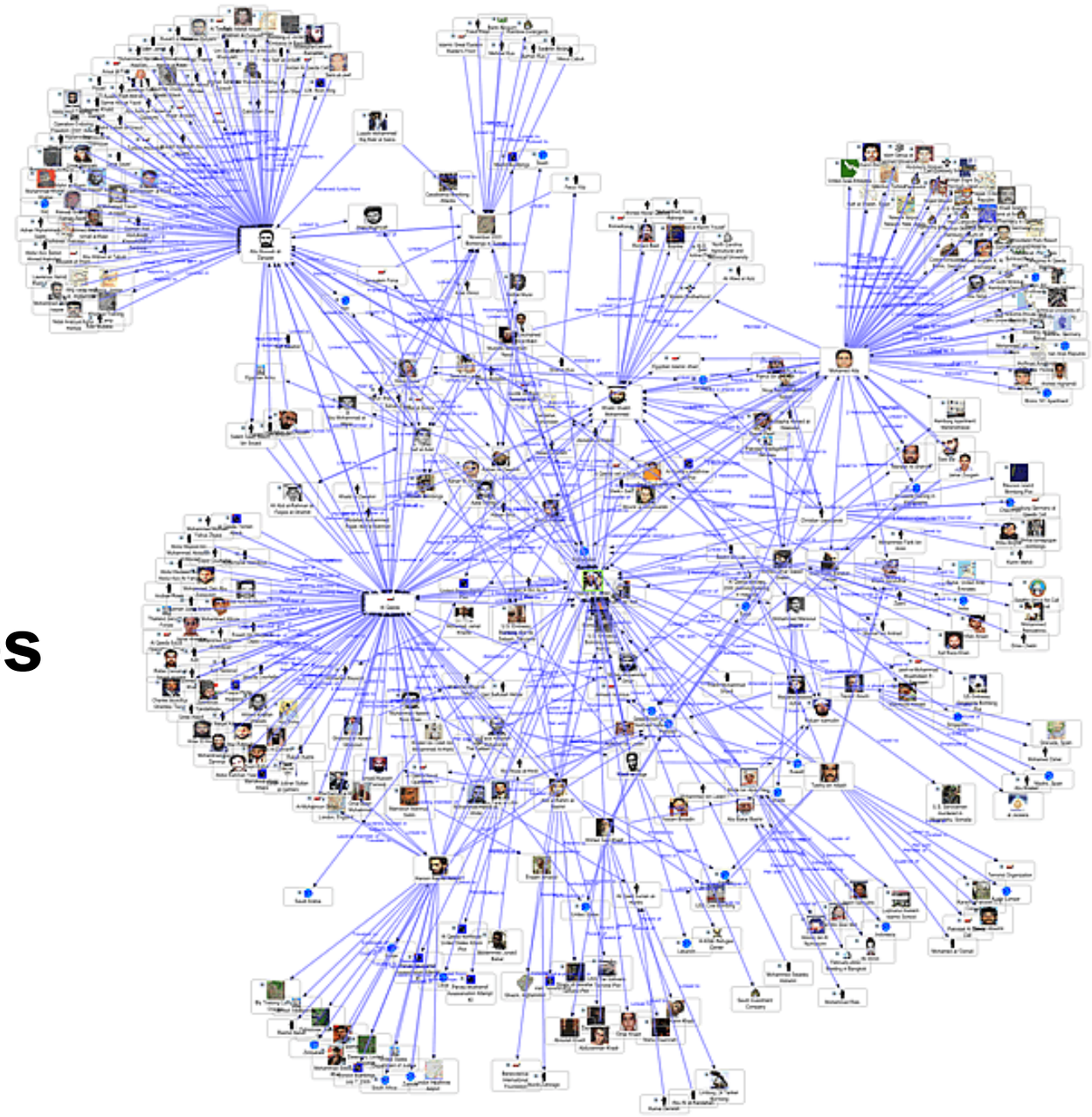
- Data that originates as digital data, rather than being converted or digitized later is proliferating. Think digital electronic medical records, implanted medical devices, diagnostic imaging technology…

Graph Data

Lots of interesting data has a graph structure:

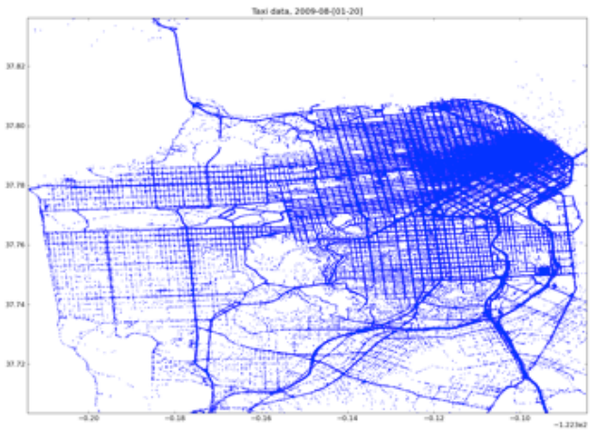
- **Social networks**
- **Communication networks**
- **Computer Networks**
- **Road networks**
- **Citations**
- **Collaborations/Relationships**
- **…**

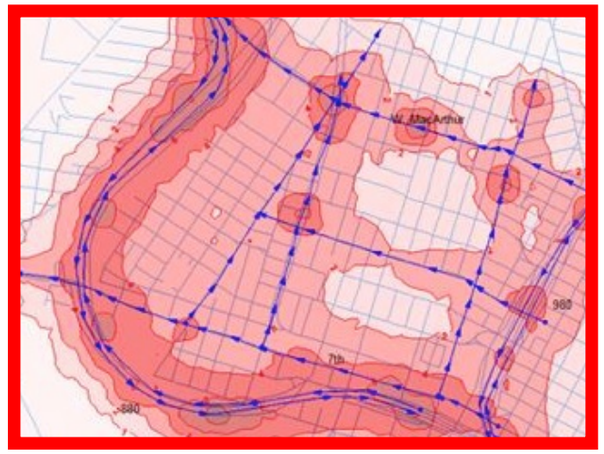
Some of these graphs can get quite large (e.g., Facebook* user graph)

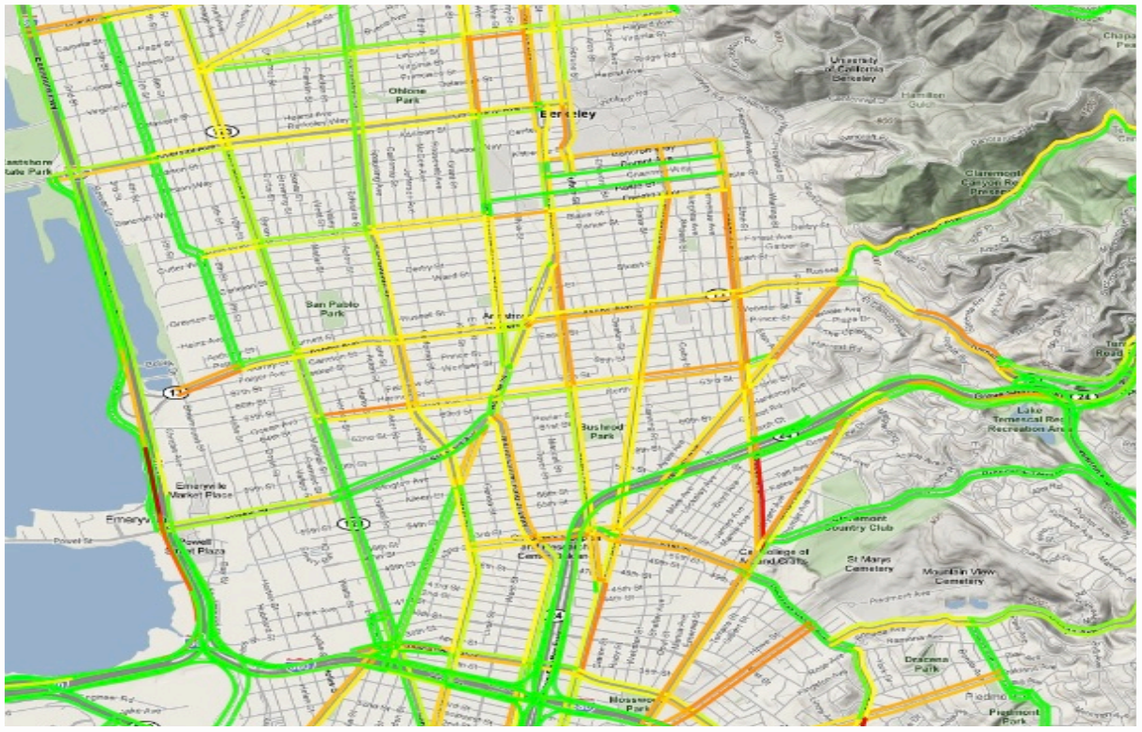


What can you do with the data?

Crowdsourcing + physical modeling + sensing + data assimilation







“Big Data” is so 2012

- “… the sexy job in the next 10 years will be statisticians,” Hal Varian, Google Chief Economist
- the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute’s June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in “Big Data Science”

Data Science – What IS IT?

“Data Science” an

O’Reilly Radar report

Types of Analytics

- ***Descriptive***: A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance.
- ***Diagnostic***: A set of techniques for determine what has happened and why
- ***Predictive***: A set of techniques that analyze current and historical data to determine what is most likely to (not) happen
- ***Prescriptive***: A set of techniques for computationally developing and analyzing alternatives that can become courses of action – either tactical or strategic – that may discover the unexpected
- ***Decisive***: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making when presented with a set of alternatives.

	Passive		Active
Deductive	Descriptive		Diagnostic
Inductive	Predictive		Prescriptive

Descriptive Analytics

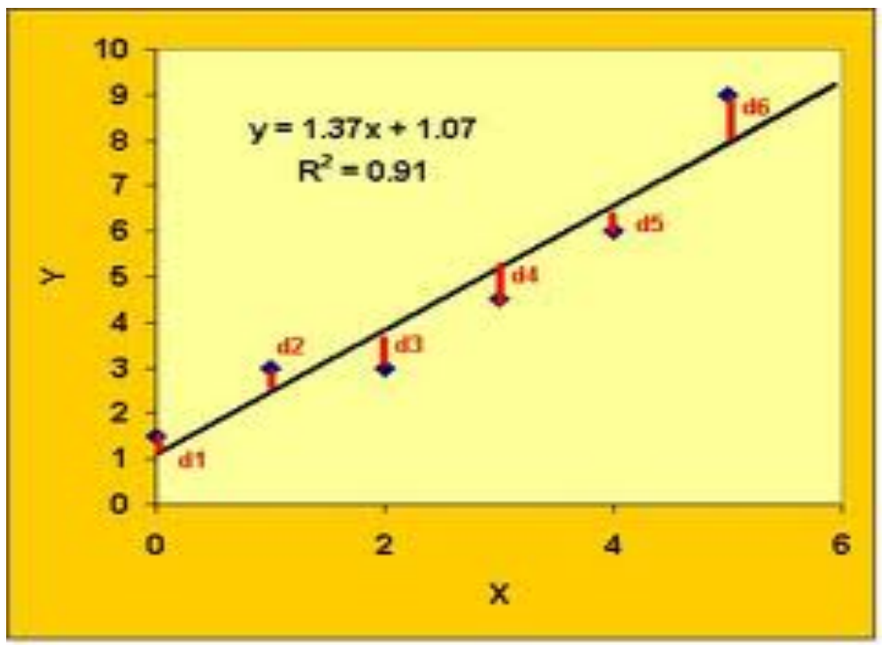
- **Process**:
  - Identify the attributes, then assess/evaluate the attributes
  - Estimate the magnitude to correlate the relative contribution of each attribute to the final solution
  - Accumulate more instances of data from the data sources
  - If possible, perform the steps of evaluation, classification and categorization quickly
  - Yield a measure of adaptability within the OODA loop
- At some threshold, crossover into diagnostic and predictive analytics

http://v1shal.com/content/25-cartoons-give-current-big-data-hype-perspective/

"Data don't make any sense,
we will have to resort to statistics."

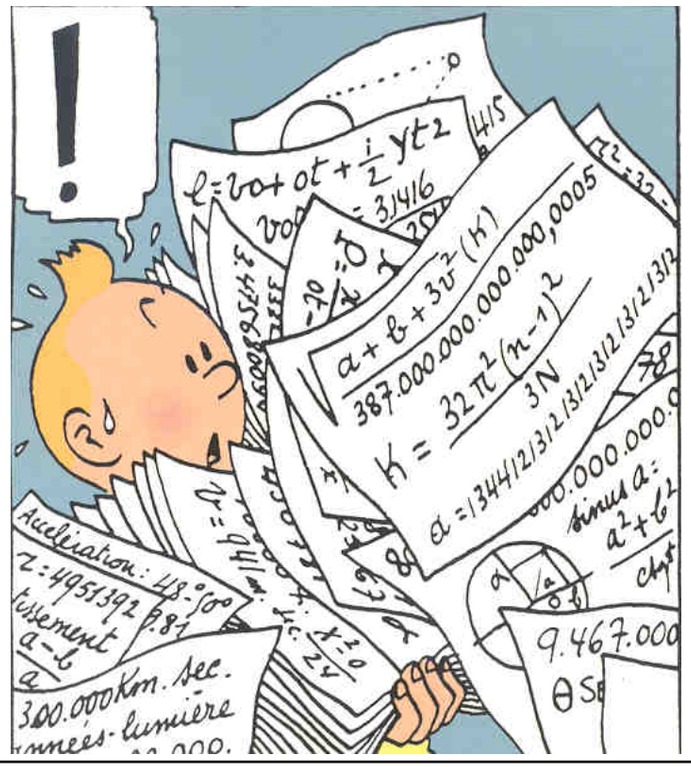
Diagnostic Analytics

- Process:
  - Begin with descriptive analytics
  - Extract patterns from large data quantities via data mining
  - Correlate data types for explanation of near-term behavior – past and present
  - Estimate linear/non-linear behavior not easily identifiable through other approaches.
- Example: by classifying past insurance claims, estimate the number of future claims to flag for investigation with a high probability of being fraudulent.



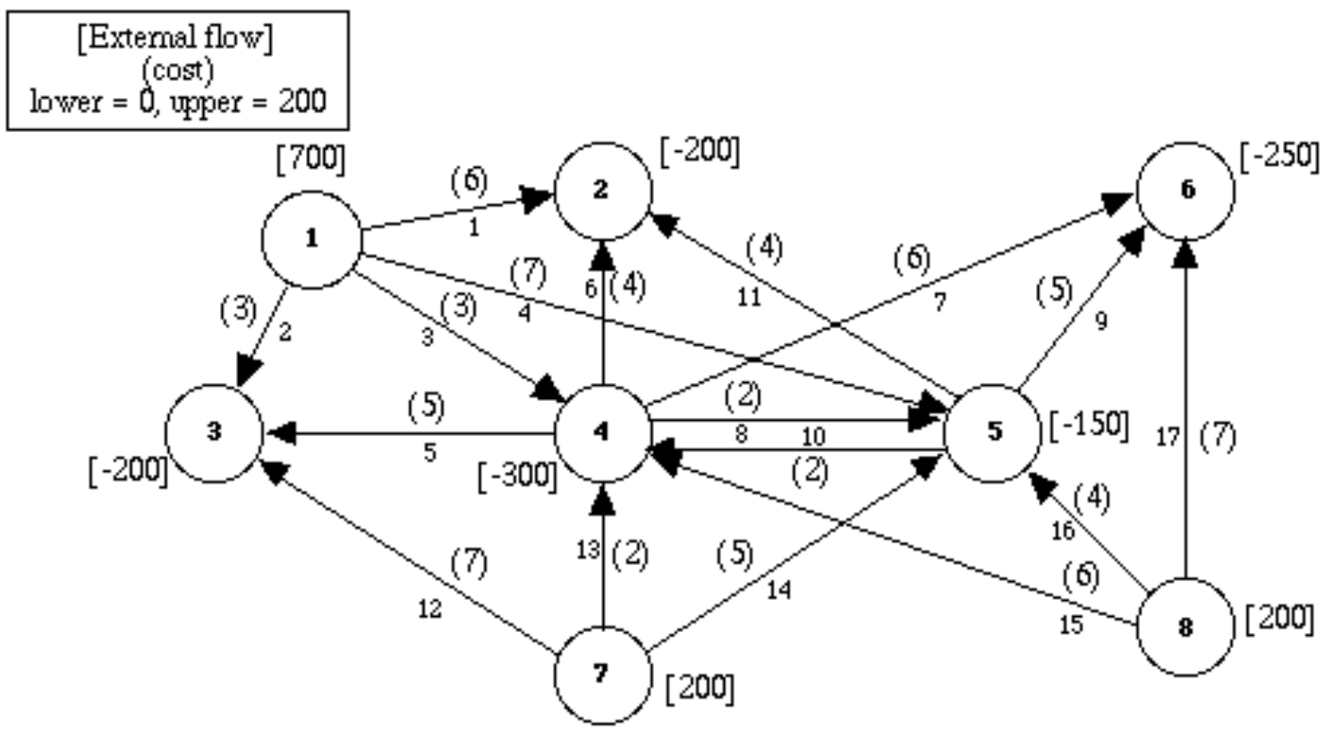
Predictive Analytics

- Process:
  - Begin with descriptive AND diagnostic analytics
  - Choose the right data based on domain knowledge and relationships among variables
  - Choose the right techniques to yield insight into possible outcomes
  - Determine the likelihood of possible outcomes given initial boundary conditions
  - Remember! Data driven analytics is non-linear; do NOT treat like an engineering project



Prescriptive Analytics

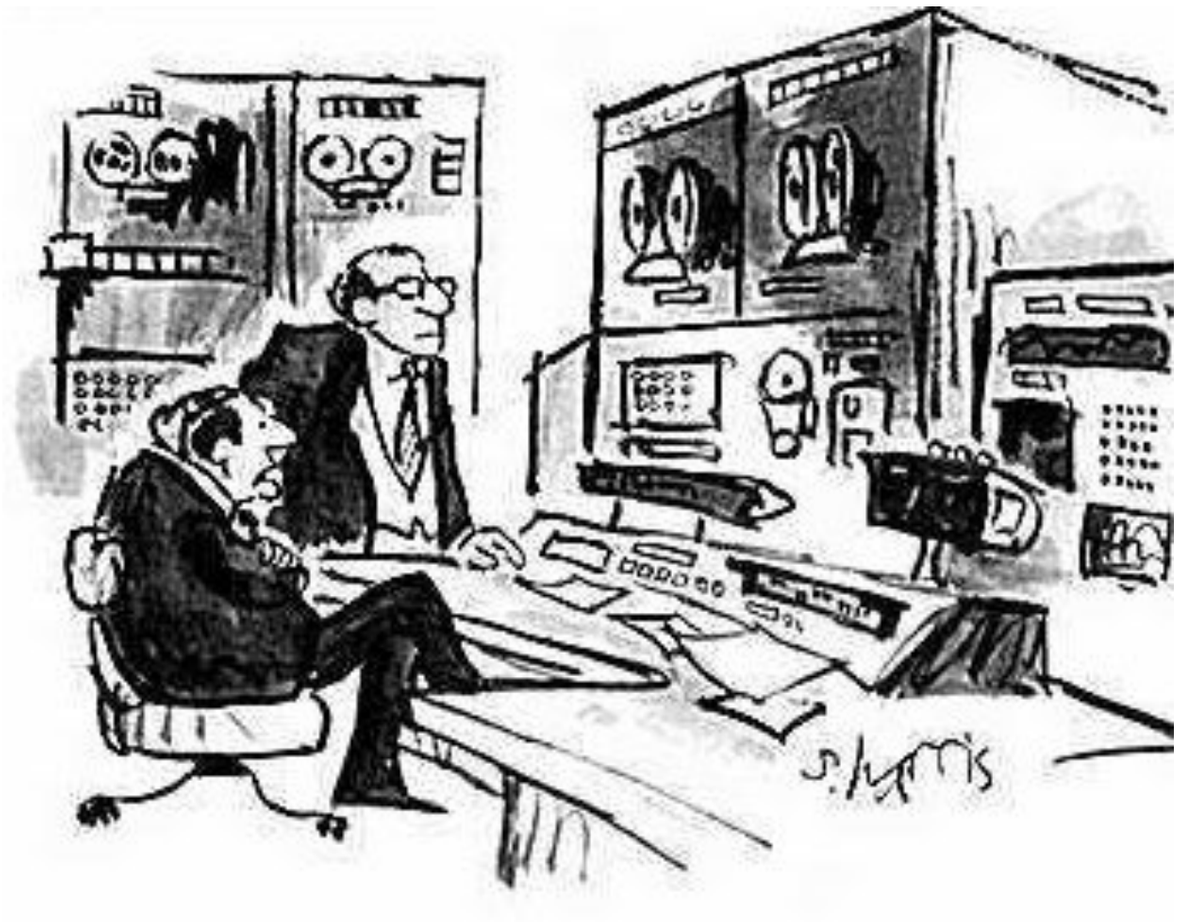
- **Process**:
  - Begin w/ predictive analytics
  - Determine what should occur and how to make it so
  - Determine the mitigating factors that lead to desirable/undesirable outcomes
  - "What-if" analysis w/ local or global optimization
  - Ex: Find the best set of prices and advertising frequency to maximize revenue
  - Ex: And, the right set of business moves to make to achieve that goal



"Make it so"

Decisive Analytics

- Process:
  - Given a set of decision alternatives, choose the one course of action to do from possibly many
  - But, it may not be the optimal one.
  - Visualize alternatives – whole or partial subset
  - Perform exploratory analysis – what-if and why
    - How do I get to there from here?
    - How did I get here from there?



"What it comes down to is this thing is capable of telling us a lot more than we really want to know."

The Role of Analytics

- "Tools and techniques that gear the analyst's mind to apply higher levels of critical thinking can substantially improve analysis… structuring information, challenging assumptions, and exploring alternative interpretations."

  Richards Heuer, Jr., "The Psychology of Intelligence Analysis"

- Beware Frege's Caution:
  - Converse Problems:
    - If you magnify on details, you are losing the overview
    - If you focus on the overview, you don't see the details
  - Problem with Data Mining:
    - Applying statistics to understand the trends causes a loss of grounding in the data

The Analytics Continuum

- Analytics problems span a continuum:
  - Short-term analysis leads to quick fixes and quick results, which may be unsustainable
  - What are the disruptive innovations in the middle-term that provide near-term domain leadership?
  - Long-term leads to strategic changes and innovations that provide sustainable domain dominance.

Finding a Needle
In a Haystack
Top-Down Analysis
Deductive

A little bit of this; a little bit of that
Middle-Out Analysis
Abductive?

Spinning Straw into Gold
Bottom-Up Analysis
Inductive

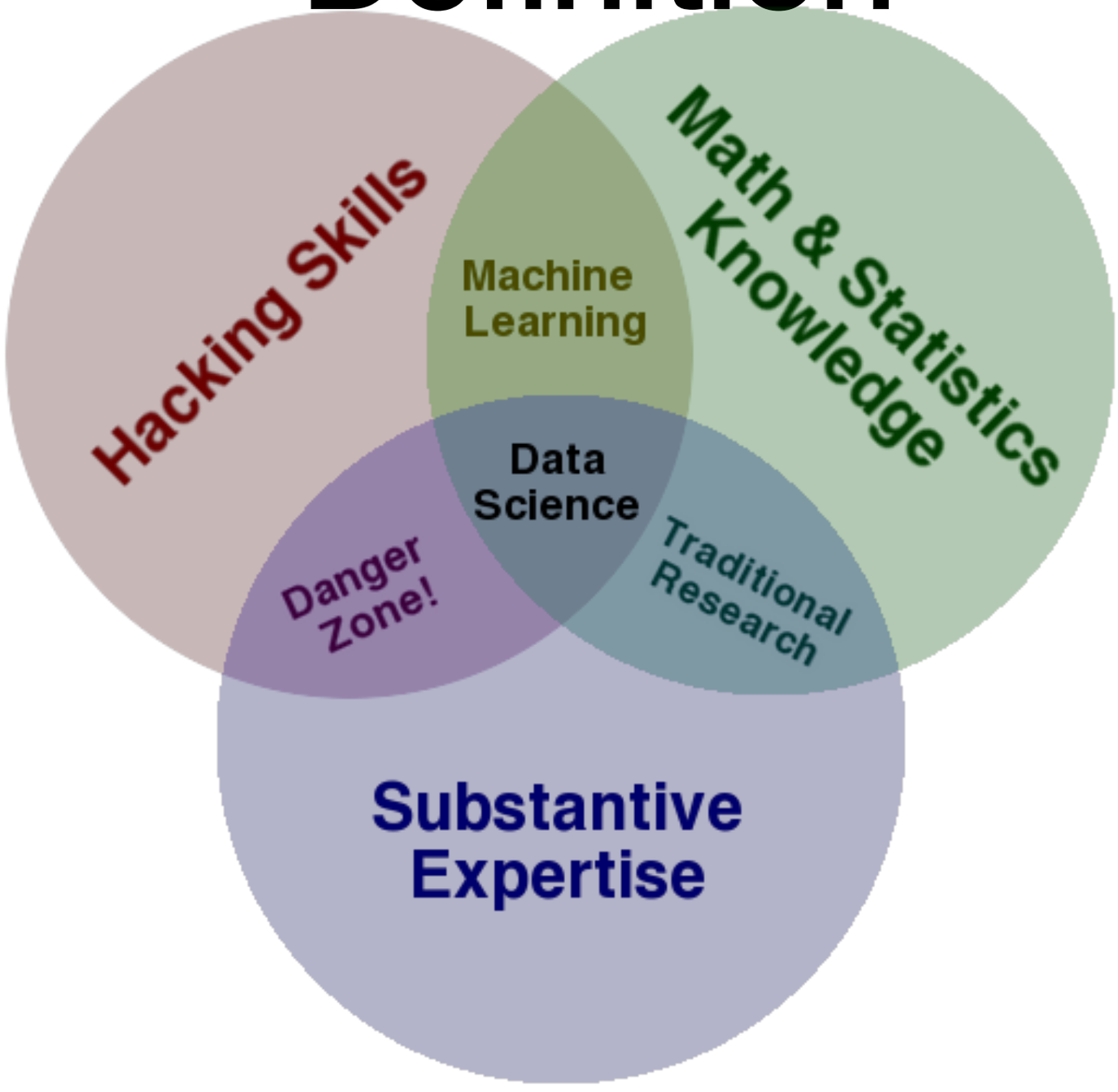
Some recent ML Competitior

kaggle

Active Competitions

	Flight Quest 2: Flight Optimization Final Phase of Flight Quest 2	33 days Coming soon $220,000
	Packing Santa's Sleigh He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams $10,000
Genentech	**Flu Forecasting** Predict when, where and how strong the flu will be	41 days 37 teams
GALAXY ZOO	**Galaxy Zoo - The Galaxy Challenge** Classify the morphologies of distant galaxies in our Universe	2 months 160 teams $16,000
DENIED APPROVED	**Loan Default Prediction - Imperial College Lon...** Constructing an optimal portfolio of loans	52 days 82 teams $10,000
	Dogs vs. Cats Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag

Data Science – One Definition



Contrast: Databases

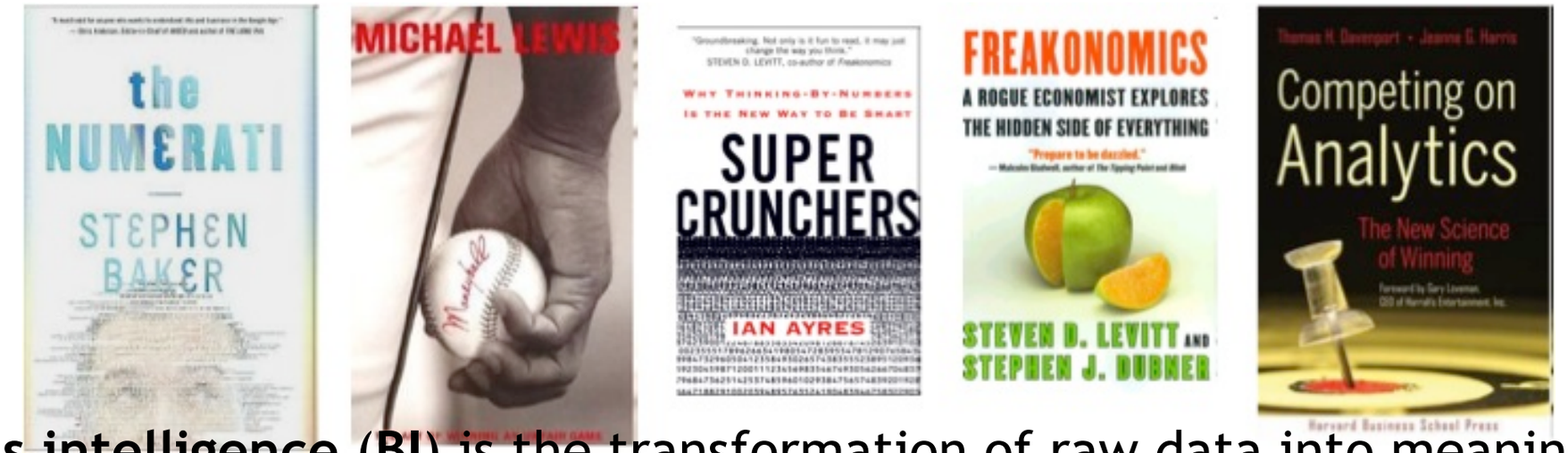
	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery,	Speed, Availability,
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3),
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB,

ency, Isolation and Durability

CAP = Consistency, Availability, Partition Tolerance

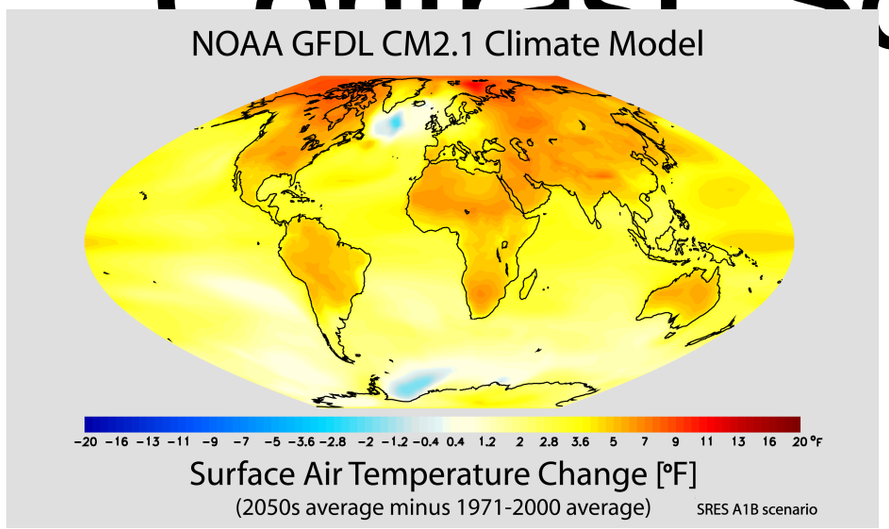
Contrast: Databases

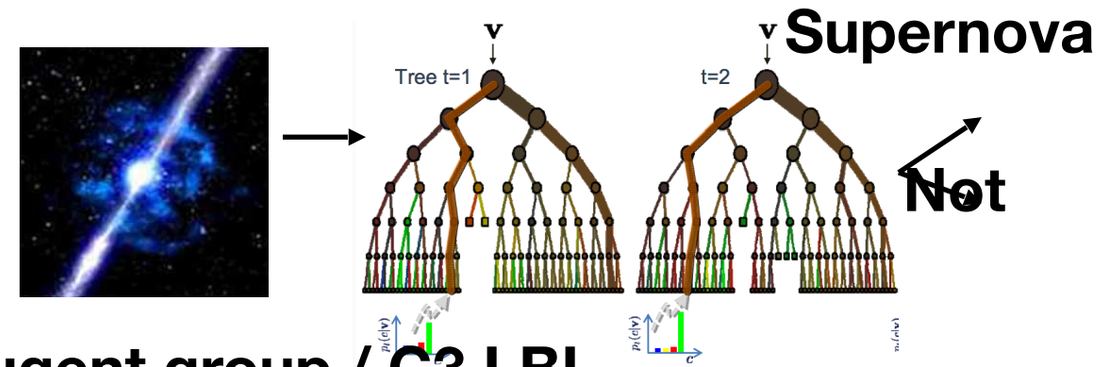
Databases	Data Science
Querying the past	Querying the future



Business intelligence (**BI**) is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

Contrast: Scientific Computing vs. General-purpose classifier





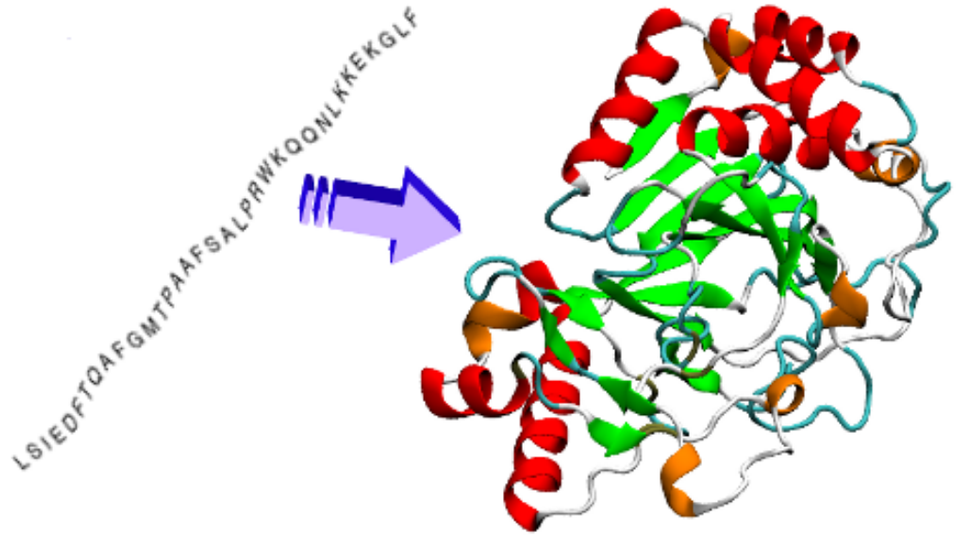
Scientific Modeling
Physics-based models
Problem-Structured
Mostly deterministic, precise
Run on Supercomputer or High-end Computing Cluster

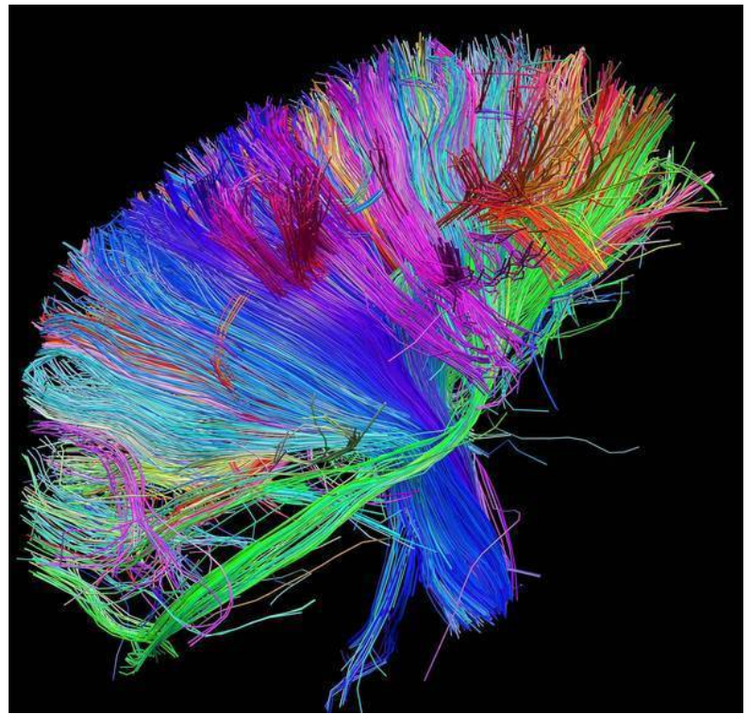
Data-Driven Approach
General inference engine replaces model
Structure not related to problem
Statistical models handle true randomness, and **unmodeled complexity**.
Run on cheaper computer Clusters (EC2)

Contrast: Computational Science

Worldwide Distributed Computing
Folding Contest

Brain Mapping at Multiple Institutes
White House, Berkeley





Quark
Rich, Complex Energy Models
Faithful, Physical Simulation

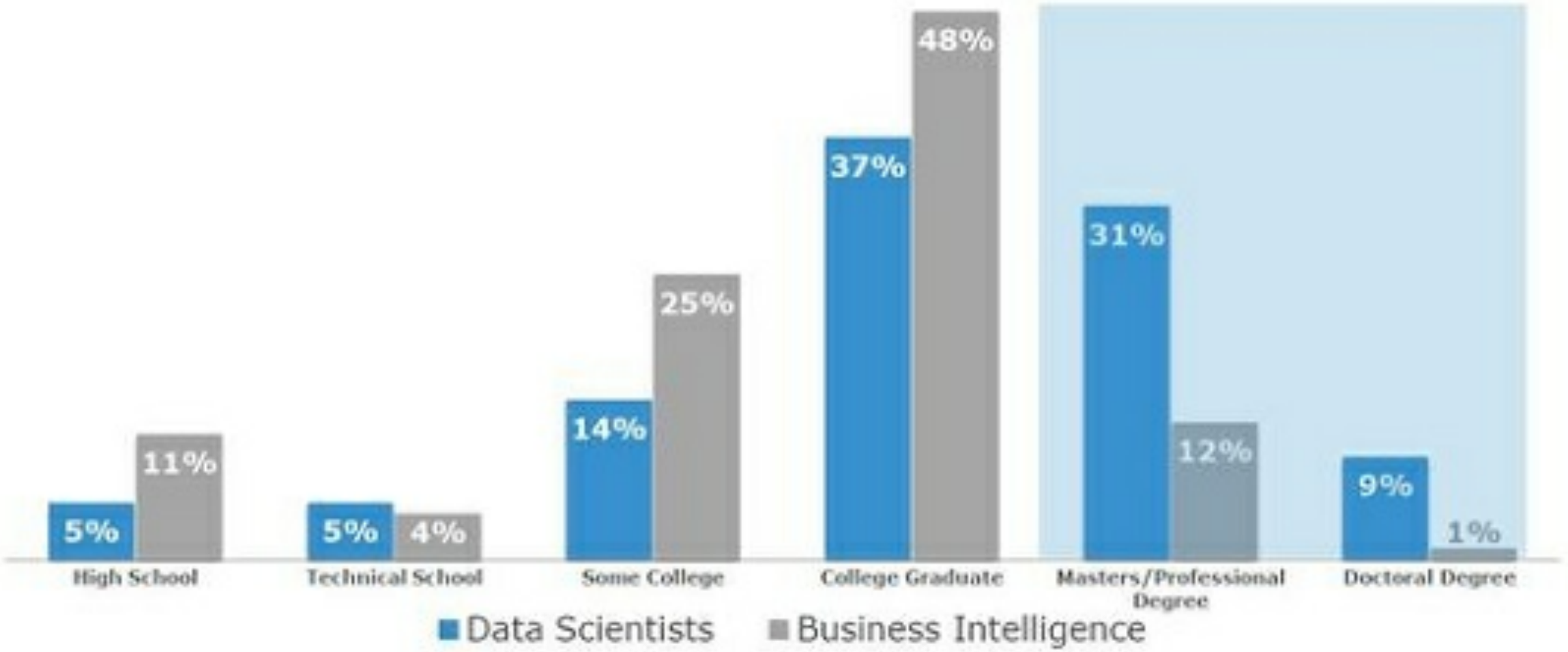
Raptor-X
Data-intensive, general ML models
Feature-based inference
Conditional Neural Fields

Techniques (Massive ML)
Principal Component Analysis
Independent Component Analysis
Sparse Coding
Spatial (Image) Filtering

Contrast: Machine Learning

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

Data science requires greater education



40% of data science professionals have an advanced degree – and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD.

EMC²

Analyzing the Analysts

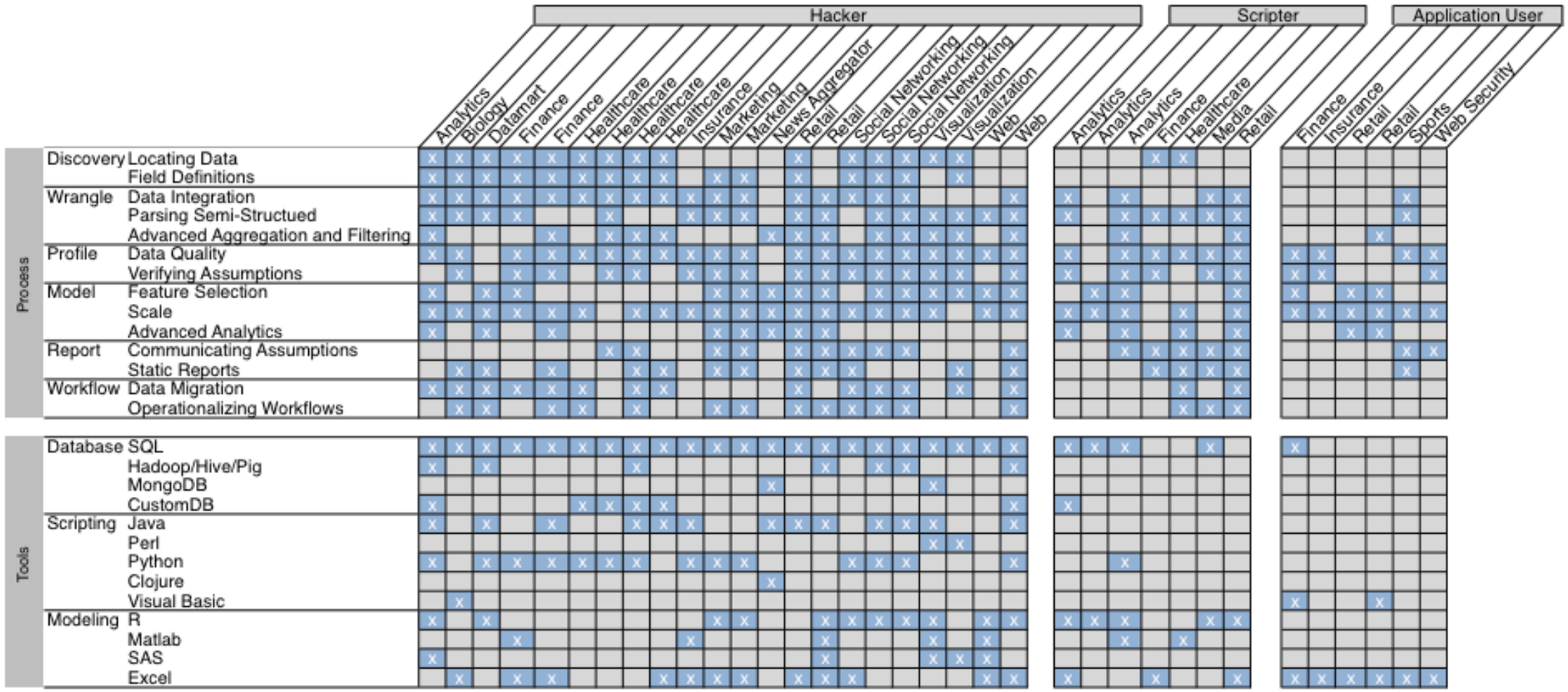


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

From Kandel, Paepcke, Hellerstein and Heer, "Enterprise Data Analysts and Visualization: An Interview Study", IEEE VAST 2012

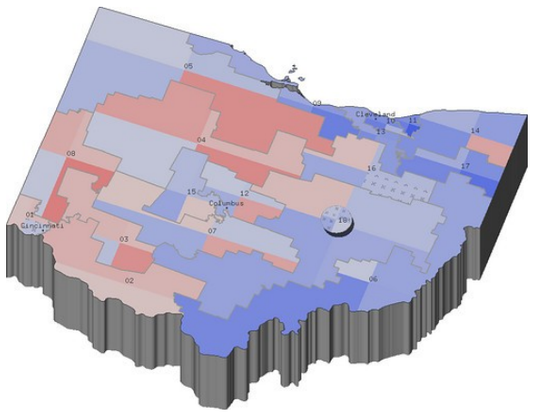
Doing Data SciENCE

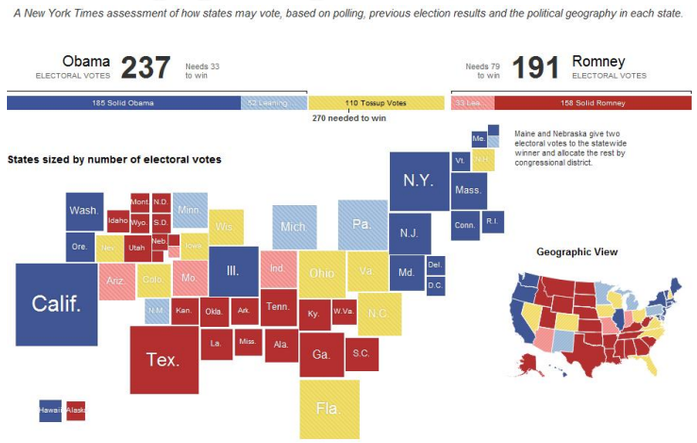
Ben Fry's Model

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Jeff Hammerbacher's Model

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results





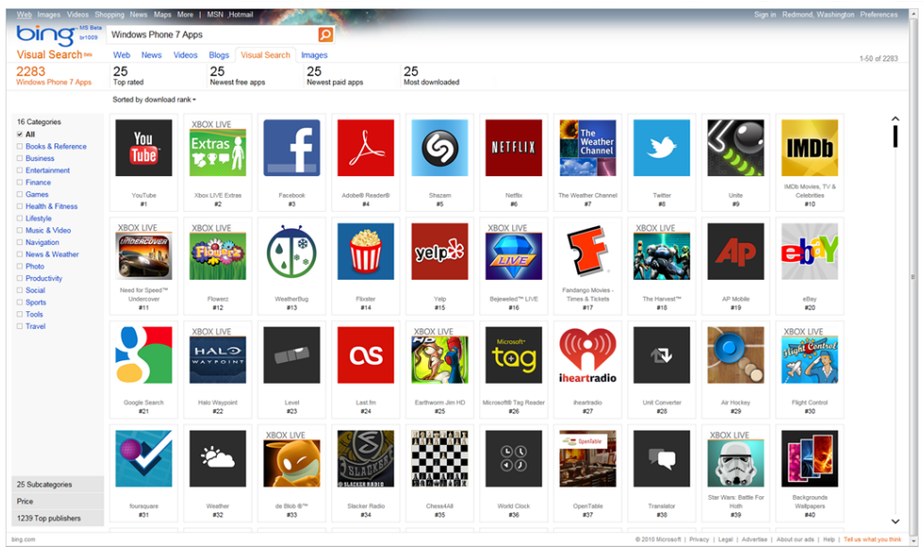
From the Trenches

Yahoo [KDD 2009, best app. paper]

Ebay [SIGIR 2011, hon. mention]

Quantcast [2012]

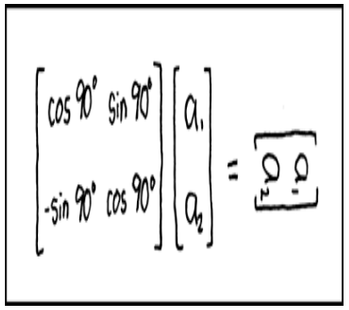
Microsoft [CIKM 2014]



Data Scientist's Practice

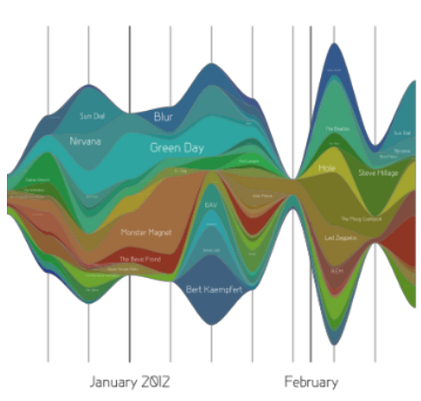
Digging Around in Data

Clean, prep



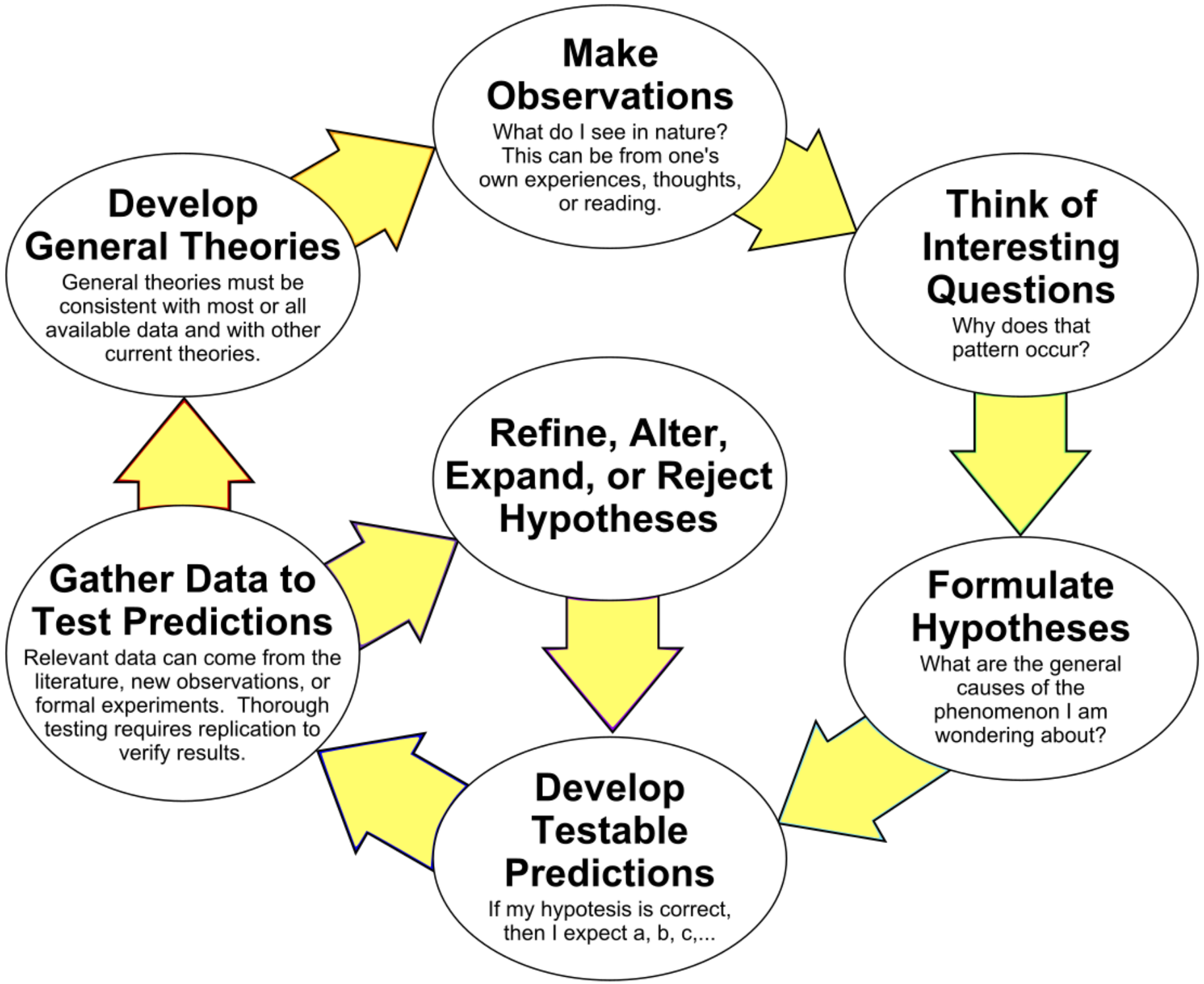
Hypothesize Model

Large Scale Exploitation



Evaluate Interpret

The Scientific Method as an Ongoing Process



What's Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

A data analysis exercise:

English Premier League Soccer: Everton vs. W. Bromwich Albion

Predict the outcome:

Everton Goals

W. Bromwich Goals	0	1	2+
0			
1			
2+			

Analysis

What kinds of data will you use?

- Almost anything is OK, except other predictions.
- History: individual or pair-wise?
- Team or players?
- Numerical or text?
- What kind of model will you build?
- What assumptions are safe to make?