

# **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

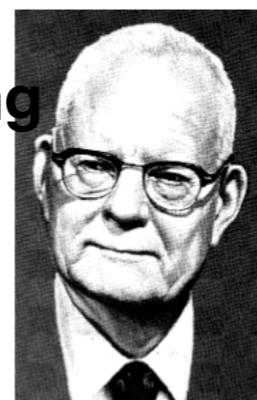
1935: "The Design of Experiments"

R.A. Fisher



1939: "Quality Control"

W.E.  
Demming

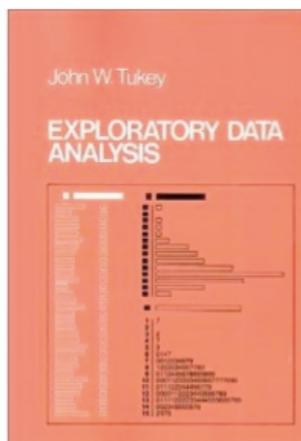


1958: "A Business Intelligence System"

Peter Luhn



1977: "Exploratory Data Analysis"



1989: "Business Intelligence"

Howard  
Dresner



1997: "Machine Learning"

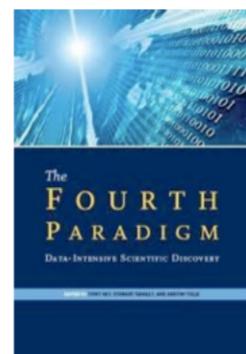


2010: "The Data Deluge"

1996: Google



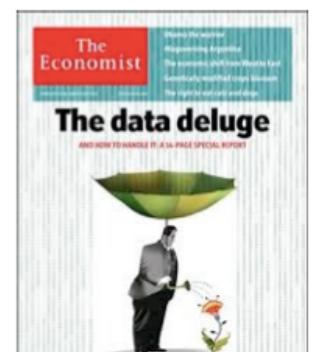
2007: "The Fourth Paradigm"



2009: "The Unreasonable Effectiveness of Data"

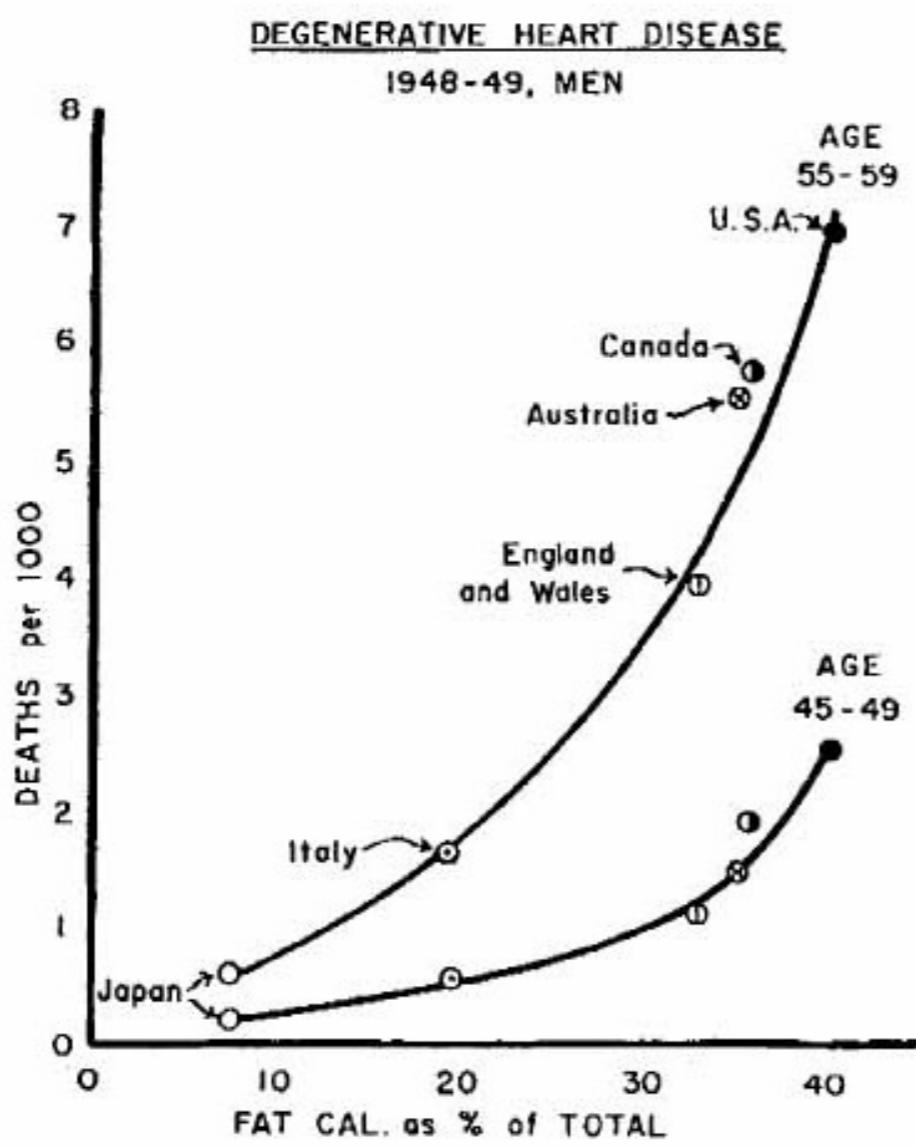


Abridged version of Jeff Hammerbacher's timeline for CS 194, 2012



# Data makes everything clearer

- See
- 13



cel K  
,28)  
1 year



# 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



the signal and the noise  
and the noise and the noise  
the noise and the noise  
noise and the noise  
why most noise are  
predictions fail to predict  
but some don't predict  
and the noise and the noise  
the noise and the noise  
nate silver noise  
noise and the noise

# 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/>

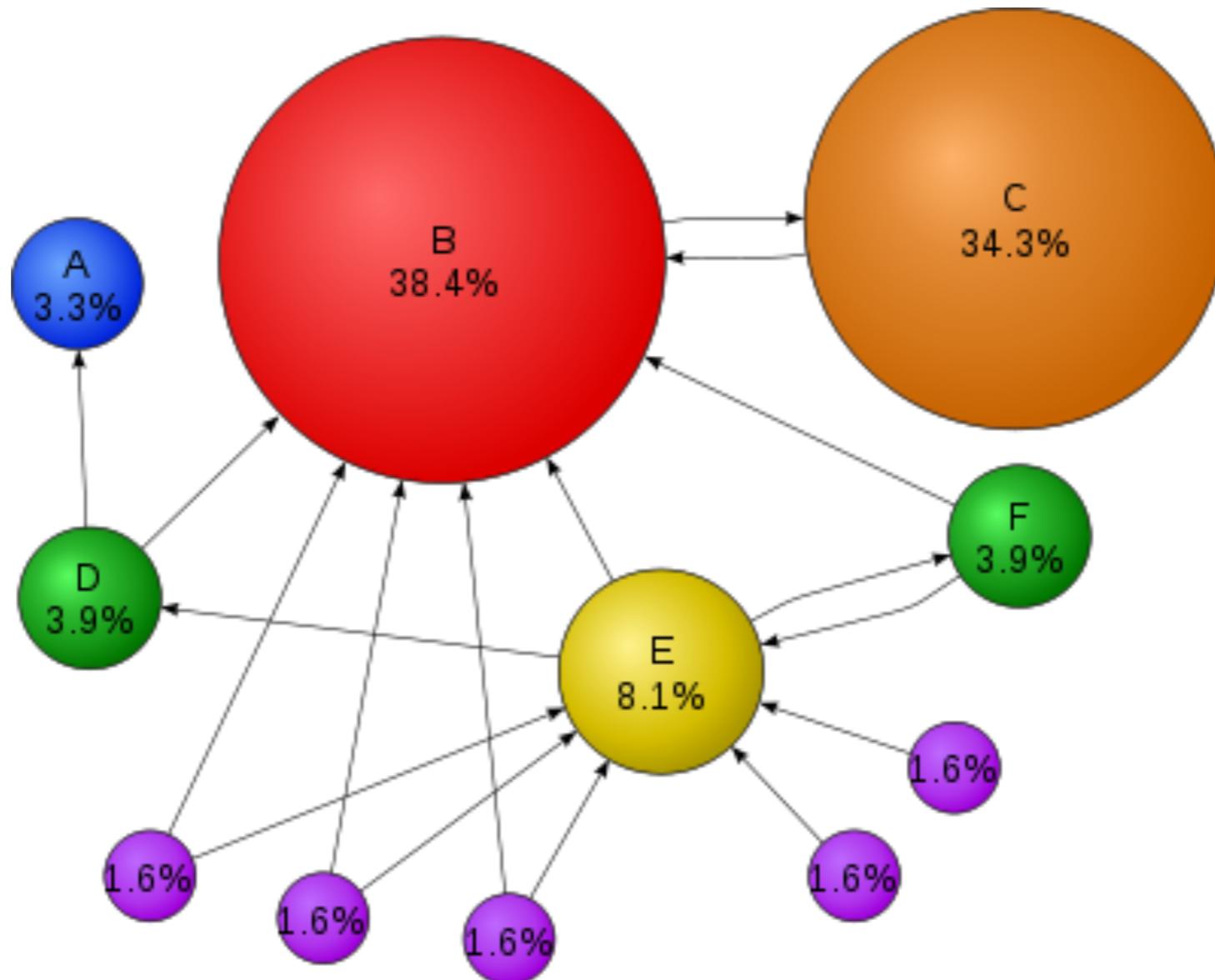
609.303 8857 backlinks 4k - none - 11/1/96

Stanford University: Portfolio Collection

████████ <http://www.stanford.edu/home/administration/portfolio.html>

167.919 34 backlinks

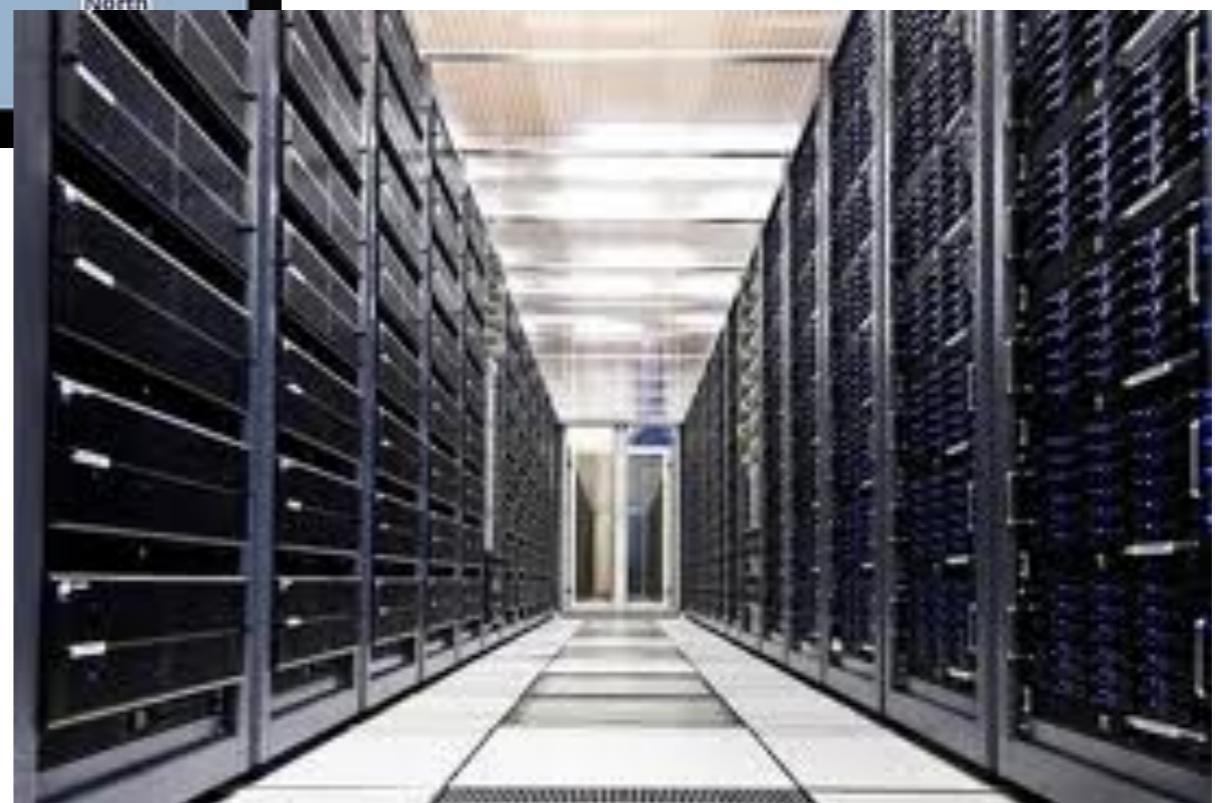
# Pagerank: The web as a behavioral dataset



# DB size = 50 billion sites



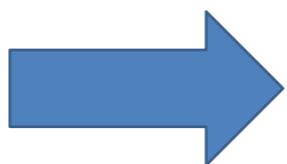
**Google server farms  
2 million machines (est)**



# 1998 – sponsored search



Overture



Google gatco towel bars Search Advanced Search

Results 1 - 10 of about 661,000 for gatco towel bars (0.33 seconds)

**Gatco Towel Bars**  
www.eFaucets.com/Gatco Low Price Guarantee, Free Shipping, 110% Price Match, No Tax, Shop Now!

**Gatco Towel Bars**  
www.AmericanHomePlus.com/Gatco 110% Low Price Guarantee. Buy Now, Free Shipping On Gatco Accessories.

**Gatco Towel Bars**  
www.PlumberSurplus.com Free S/H Available. Large Selection Gatco Towel Bars, ready to ship! Google Checkout

**Gatco Towel bars, Gatco toilet paper holders, Gatco Bathroom ...**  
Gatco Towel bars, Gatco toilet paper holders, Gatco Bathroom Accessories, Gatco robe hooks, Gatco bathroom shelves, Gatco hotel towels, Gatco double towel ...  
www.kitchensnbath.com/gatco-bath-accessories.html - Cached - Similar - ↗ ↘ ↙ ↚

**Discount Towel Bars, Towel Bars, Towel Racks**  
Discount Towel Bars offers only the highest quality bath hardware. We are factory direct distributors for Moen, Baldwin, Dynasty Hardware and Gatco. ...  
www.discounttowelbars.com/ - Cached - Similar - ↗ ↘ ↙ ↚

**Gatco at Lowe's: 24" Franciscan Chrome Double Towel Bar**  
24" Franciscan Chrome Double Towel Bar - 69656 5286.  
www.lowes.com/lws/r?action=productDetail... - Cached - Similar - ↗ ↘ ↙ ↚

**Shopping results for gatco towel bars**

**Gatco Bleu 18 in. Towel Bar - Polished Chrome**  
\$39.99 new - Sears  
**Gatco 4240 24-Inch Latitude II Towel Bar, Chrome**  
\$30.53 new - Amazon.com  
**4621 Camden Towel Bar 18 Gatco Inc**

Sponsored Links

**Gatco Bleu 18 in. Towel Bar - Polished Chrome**  
\$39.99 - Sears

**Gatco Spa Towel Rack\_3 Tier - Satin Nickel**  
\$94.99 - Sears

**Gatco Chenille 18 in. Towel Bar - Vintage ...**  
\$44.99 - Sears

**Gatco Bath Accessories**  
Lowest prices.  
All Gatco collections  
www.TheHomeDecor.net

**Towel Bars on Sale**  
Save 20%-50% Off List- New Styles  
Lowest Prices + Free Shipping!  
www.FixtureUniverse.com

Put your business here.<sup>1</sup>

Microsoft Internet Explorer

YAHOO! SEARCH jeans

Search Results

1 of 10 of about 56,000,000 for jeans

**Shop Back-to-School Boy Jeans**  
www.childrensplace.com - Shop back-to-school jeans at The Children's Place.

**Lucky Brand Jeans Official Site**  
www.luckybrandjeans.com - Shop for stylish, unique gifts for everyone this holiday season.

**7 for All Mankind**  
Designers of upscale contemporary denim jeans, as well as non-denim jackets and bottoms.  
Category: jeans > jeans  
www.7forallmankind.com - 0h - Cached - More from this site

**True Religion Brand Jeans**  
Designer jeans for men, women, and kids.  
Category: jeans > jeans  
www.truereligionbrandjeans.com - 1h - Cached - More from this site

**Diesel - Diesel's message**  
Maker of Diesel jeans, shades, fragrances, shoes, and more.  
Category: shopping > apparel > jeans  
www.diesel.com - 1k - Cached - More from this site

**Etro**  
Official site of the Etro clothing line designed by the company's founder and owner, Hidekiyo Tanaka.  
Category: shopping > apparel > jeans  
www.etro.com

**Sheplers.com: Men's & Women's Jeans**  
Casual and western-style jeans for men and women. Classic, relaxed, ...  
www.sheplers.com

**Jeans on eBay**

2002

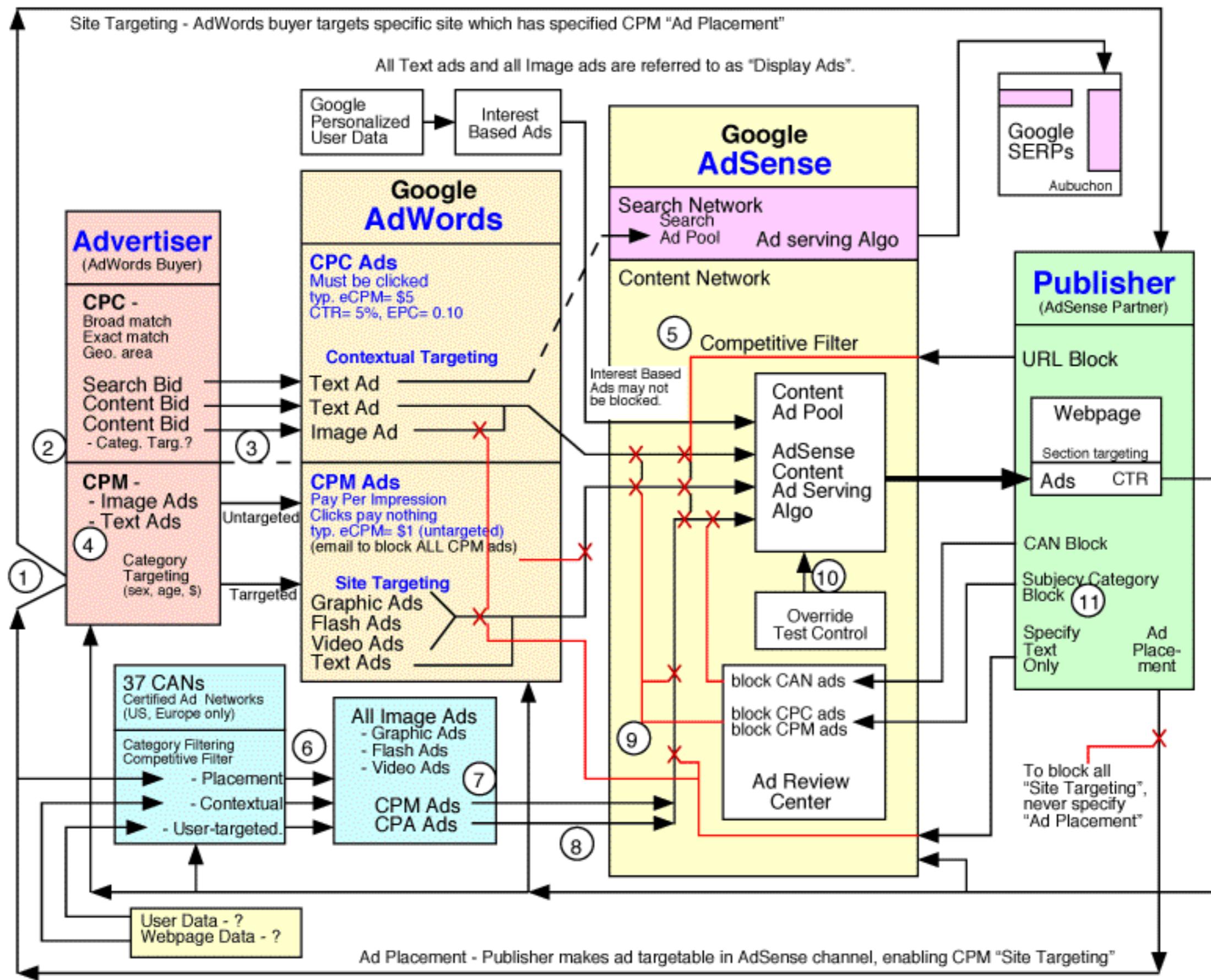
# Sponsored search

- 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.





# Data Makes Everything Clearer?

## Epidemiological modeling of online social network dynamics

John Cannarella<sup>1</sup>, Joshua A. Spechler<sup>1,\*</sup>

<sup>1</sup> 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

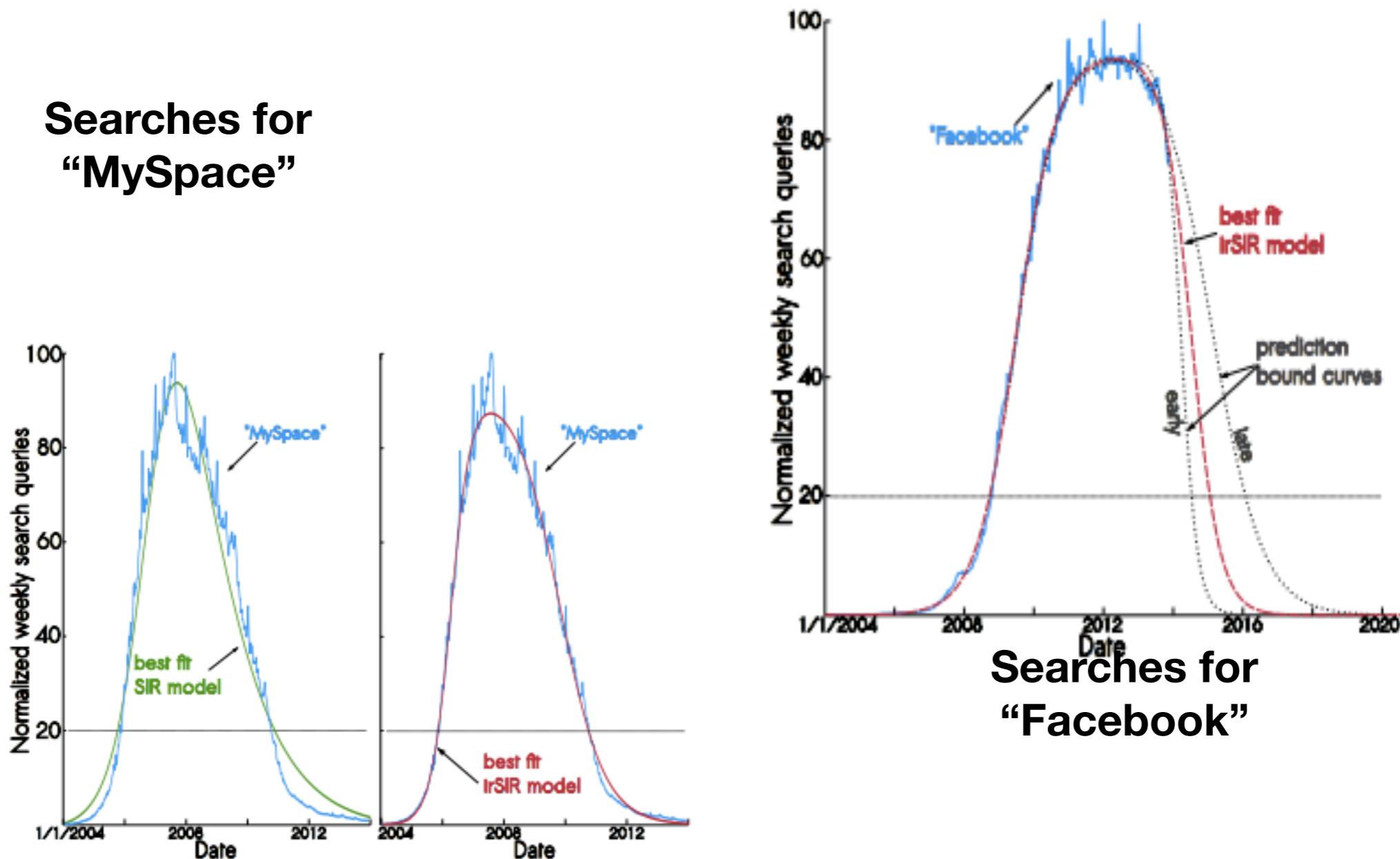
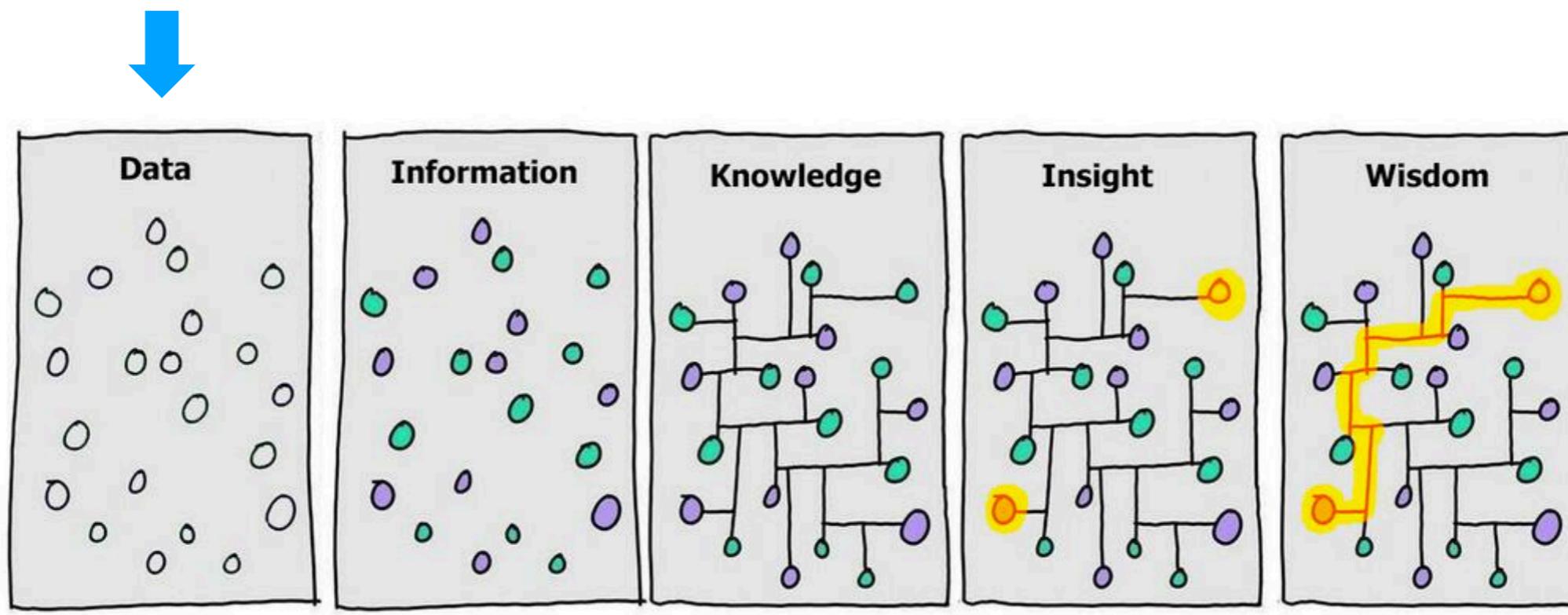


Figure 3: Data for search query “Myspace” with best fit (a) SIR and (b) irSIR models overlaid. The search query data are normalized such that the maximum data point corresponds to a value of 100.

# 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...

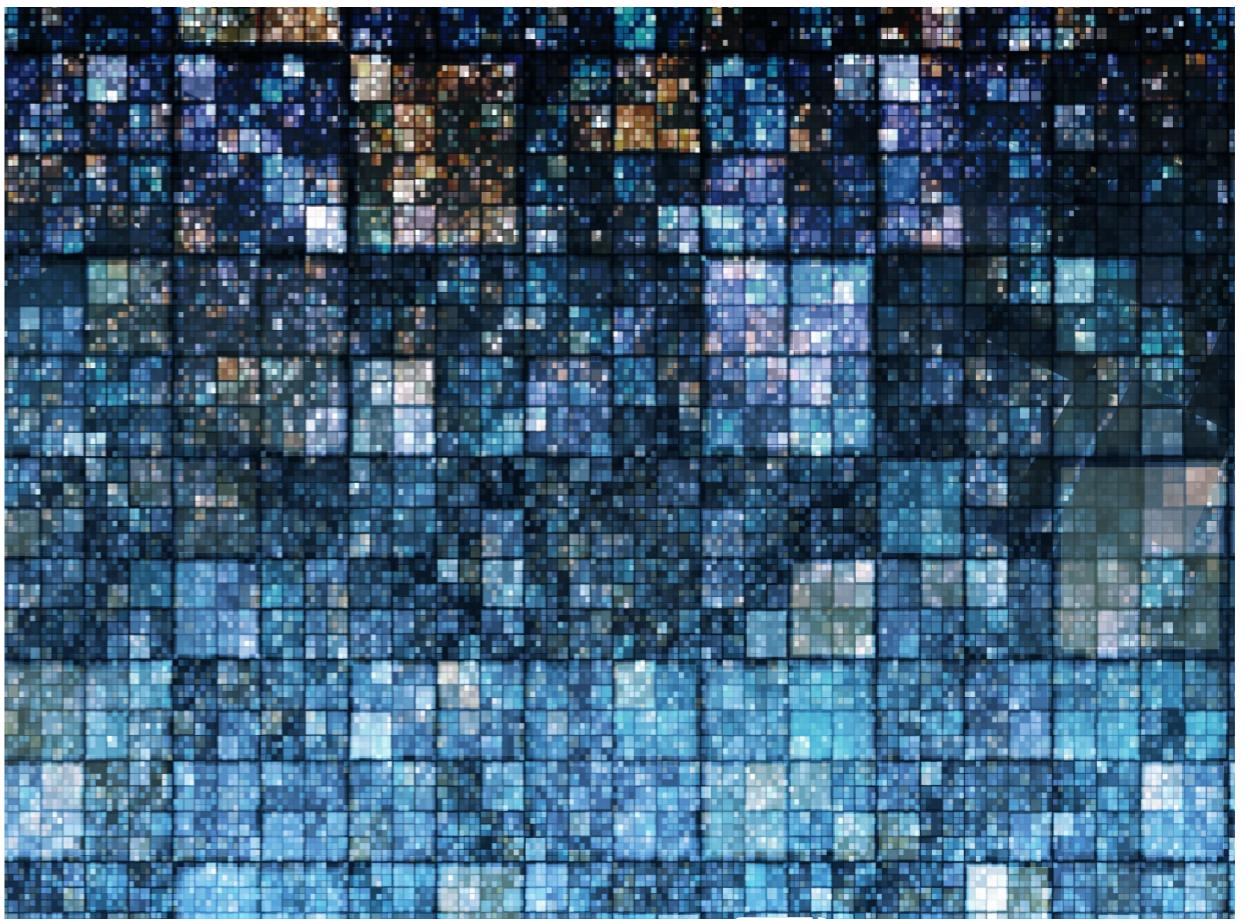


# Types of Data

- 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.



# Volume

The amount  
of data

# Velocity

The frequency  
of data

# Variety

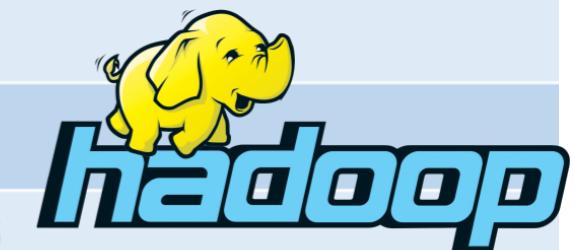
The types  
of data

# Veracity

The quality  
of data

The 4 V's  
of  
Big Data

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



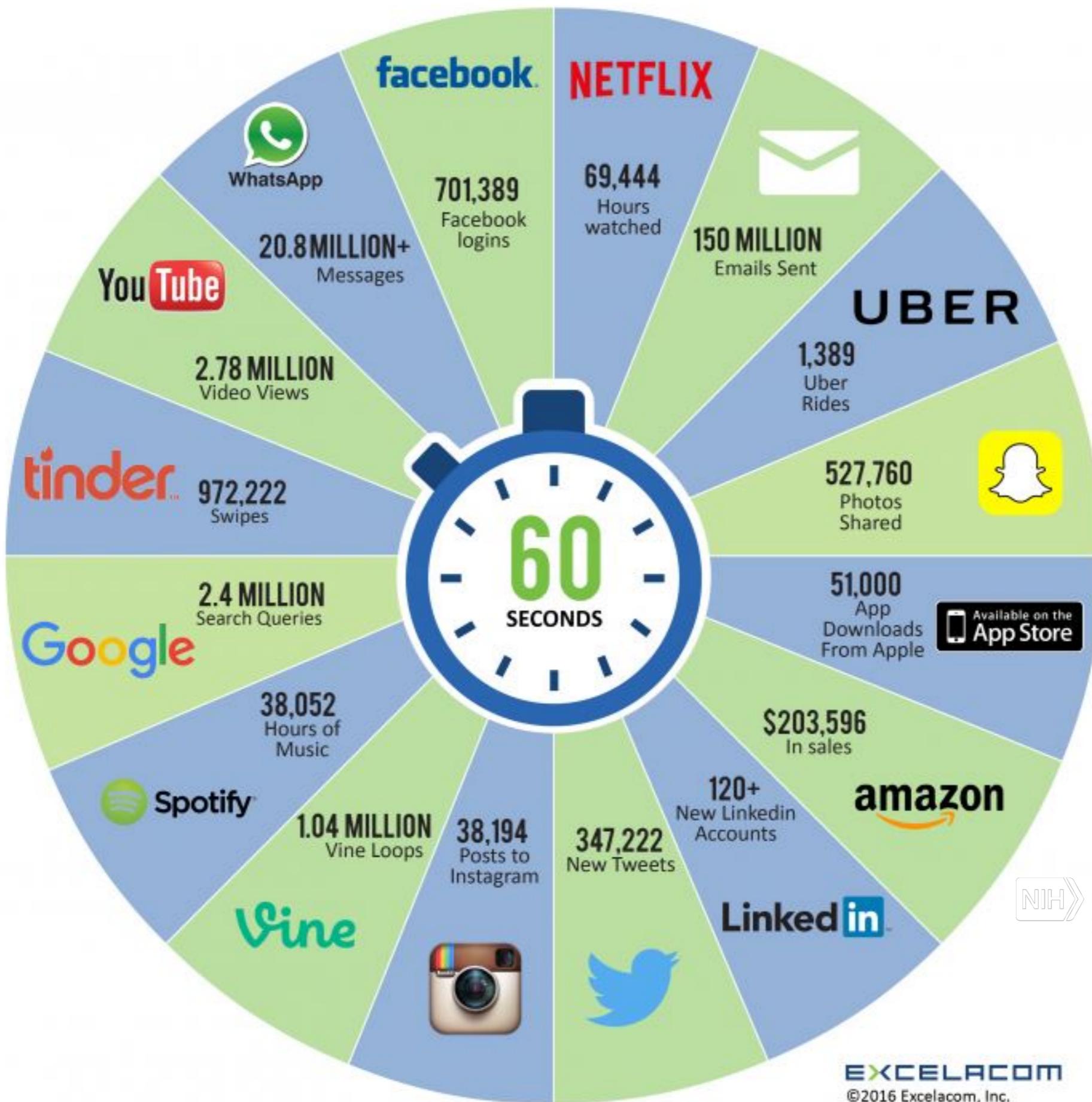
# Volume: scale of data

- 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 (40 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?



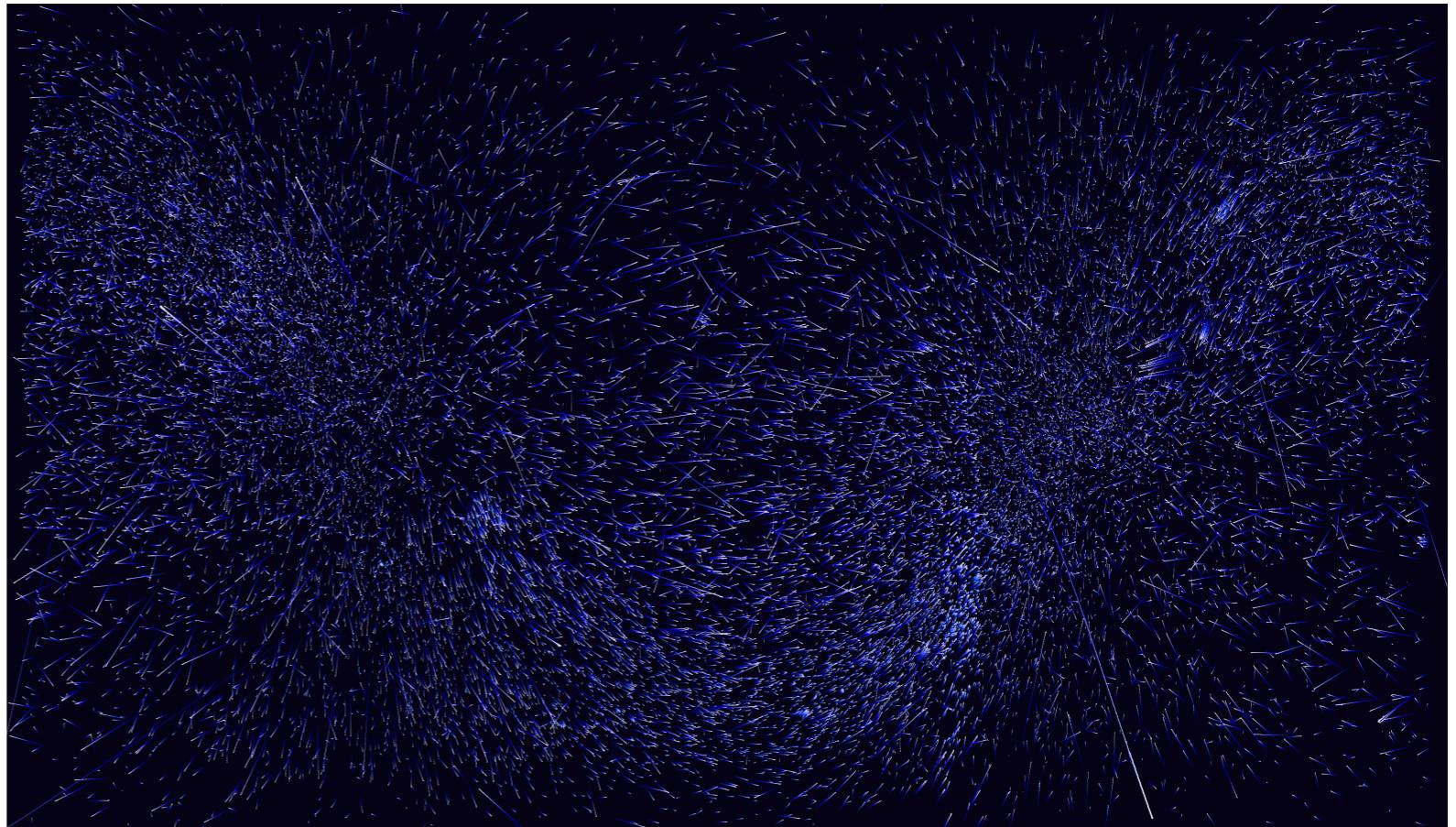
U.S. National Library of Medicine  
National Network of Libraries of Medicine  
Pacific Northwest Region

# Velocity: analysis of streaming data



# Veracity: trustworthiness of data

- Origin
- Authenticity
- Trustworthiness
- Completeness
- Integrity



## Volume

The amount  
of data

## Value

## Variety

The types  
of data

The 4 V's  
of  
Big 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!

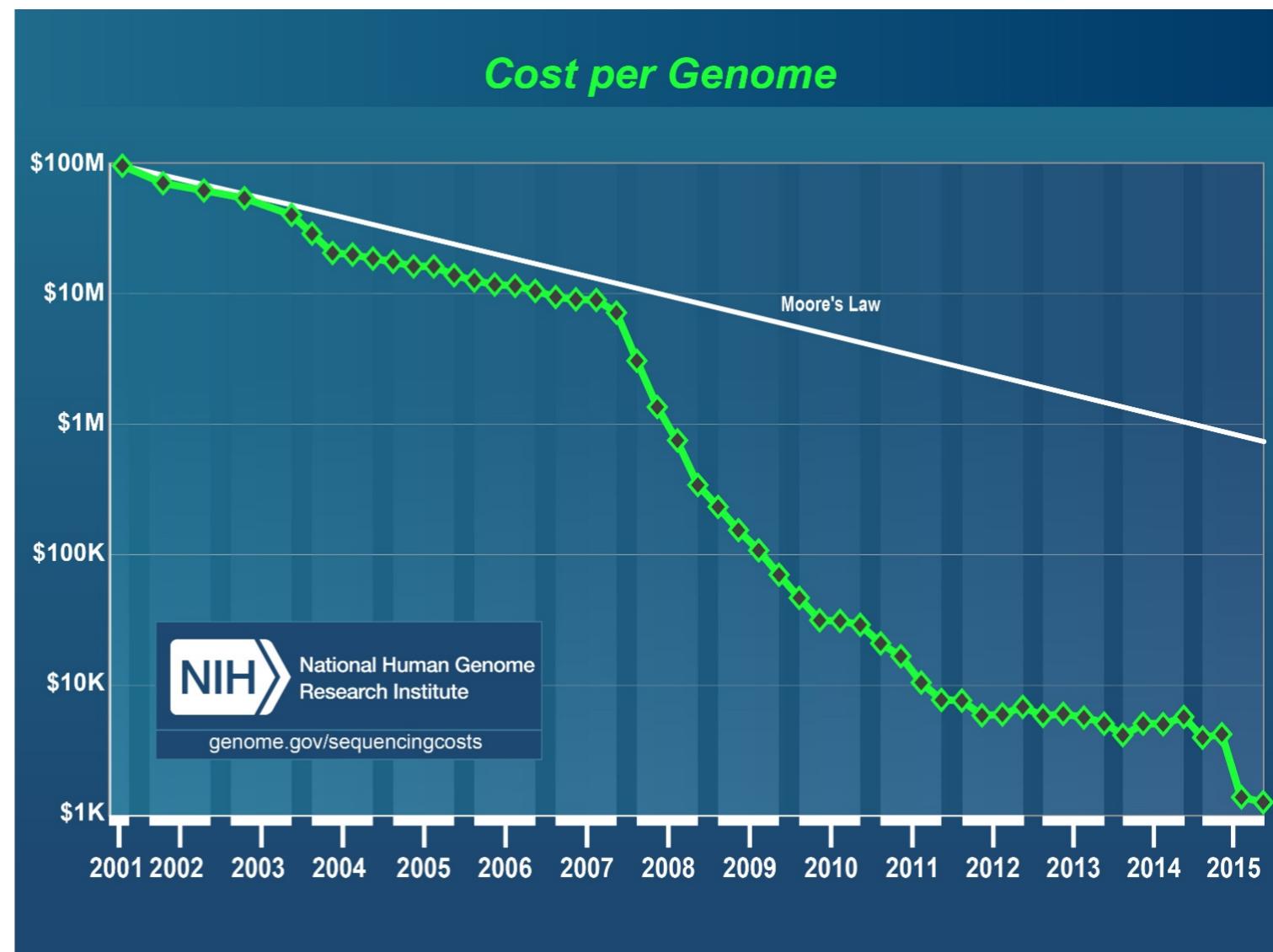


# Big Data in Health Care

- 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.



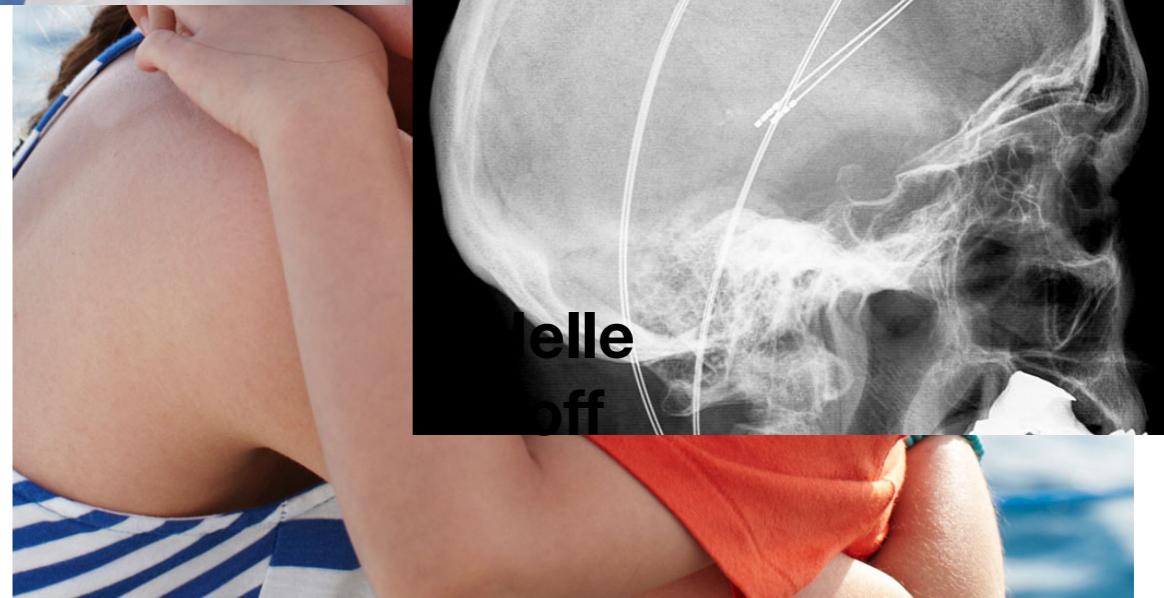
# Sensing devices

- 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...

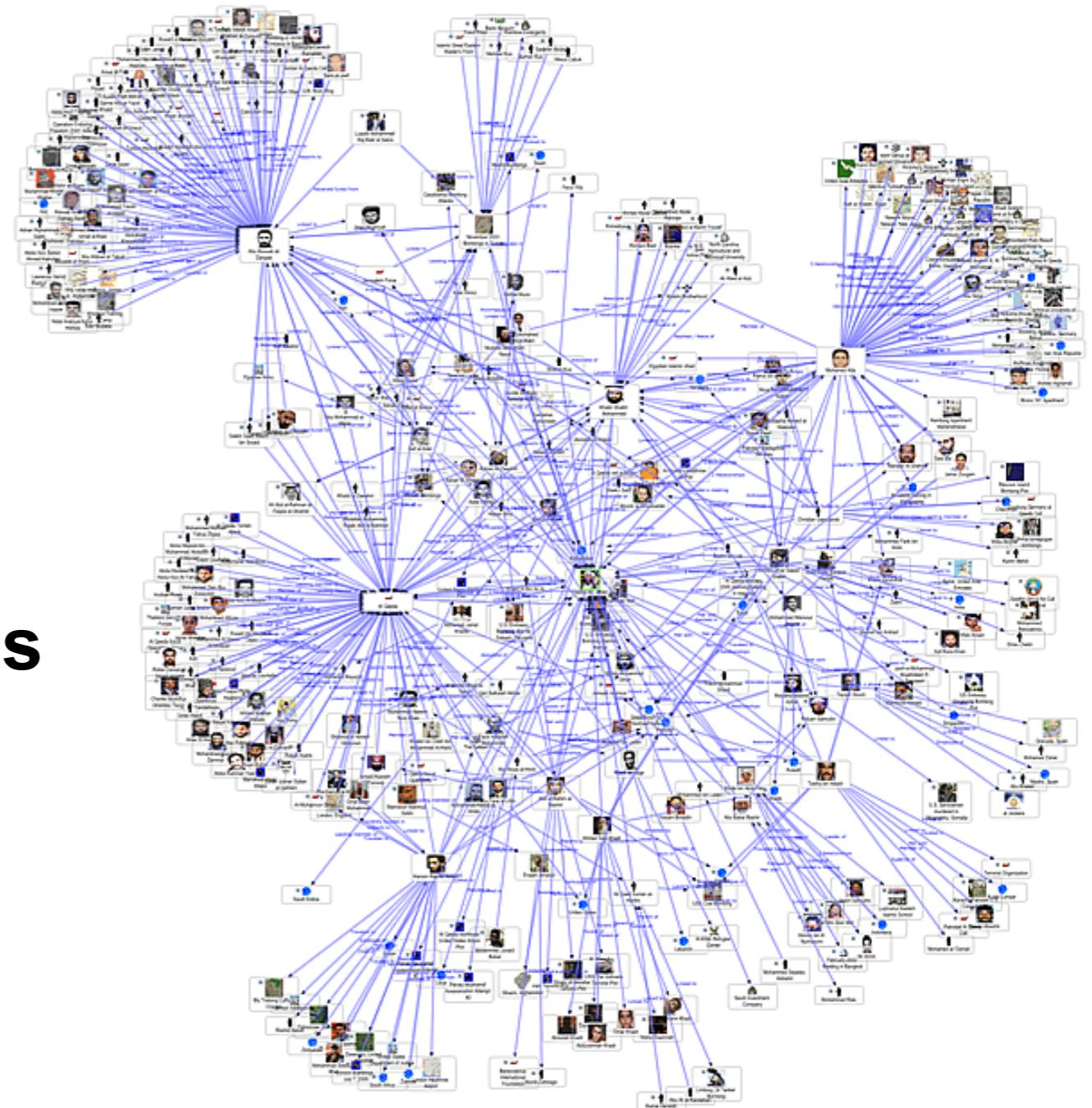


# Graph Data

**Lots of interesting data  
has a graph structure:**

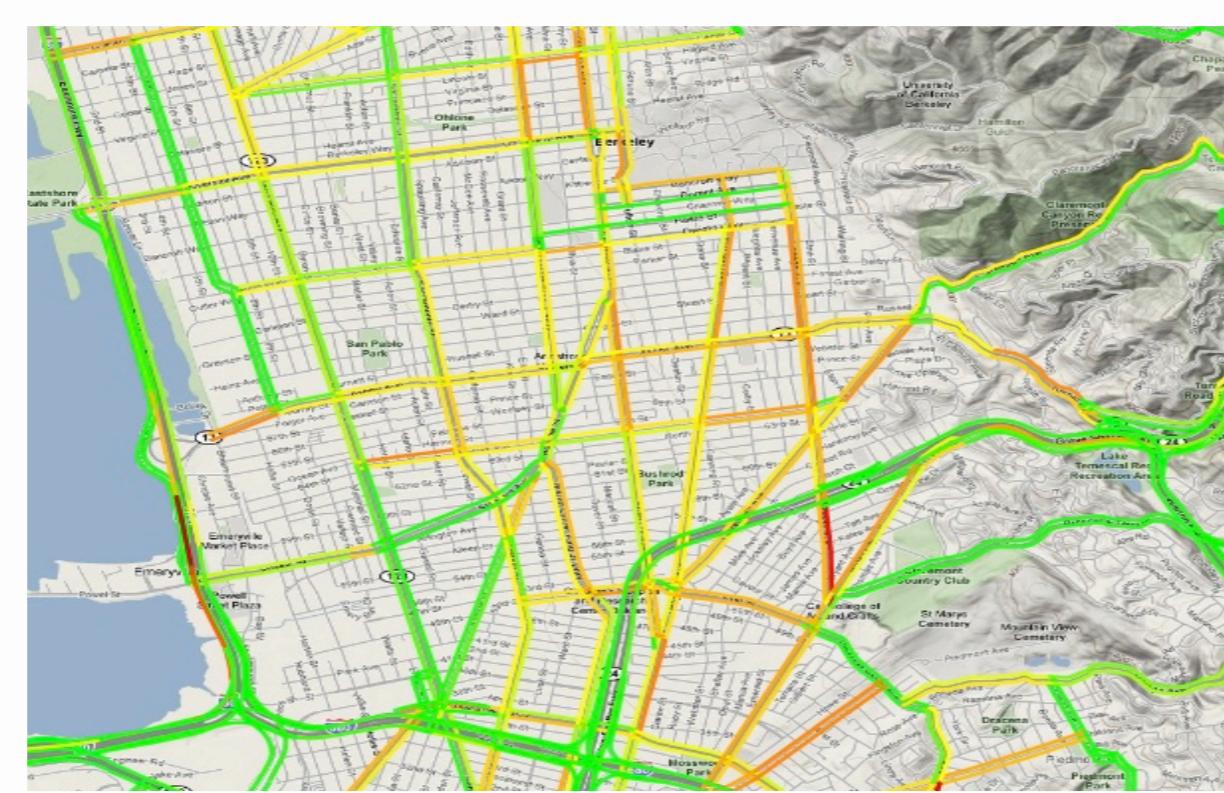
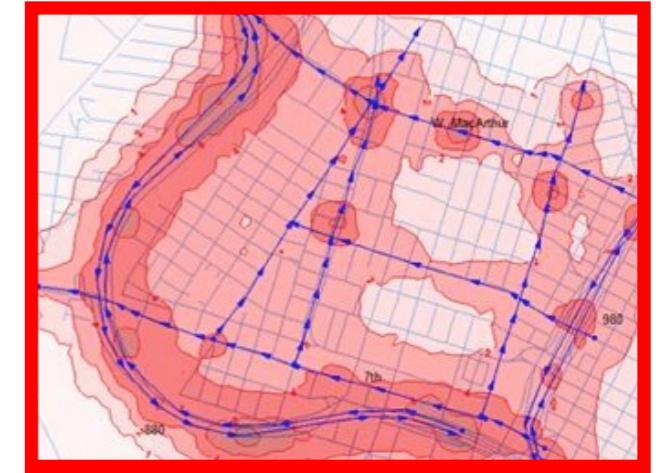
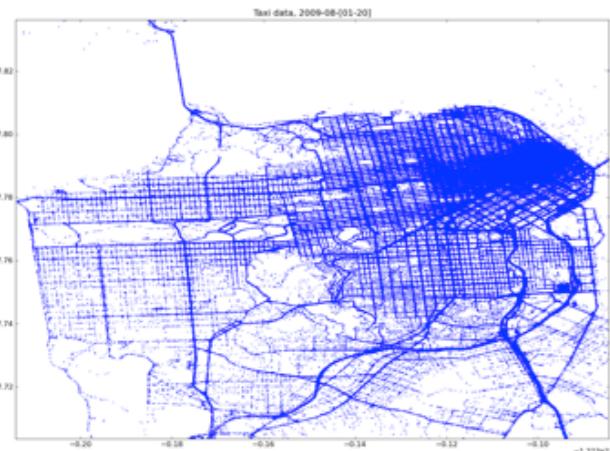
- **Social networks**
- **Communication networks**
- **Computer Networks**
- **Road networks**
- **Citations**
- **Collaborations/Relationships**
- ...

**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?**

# ***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

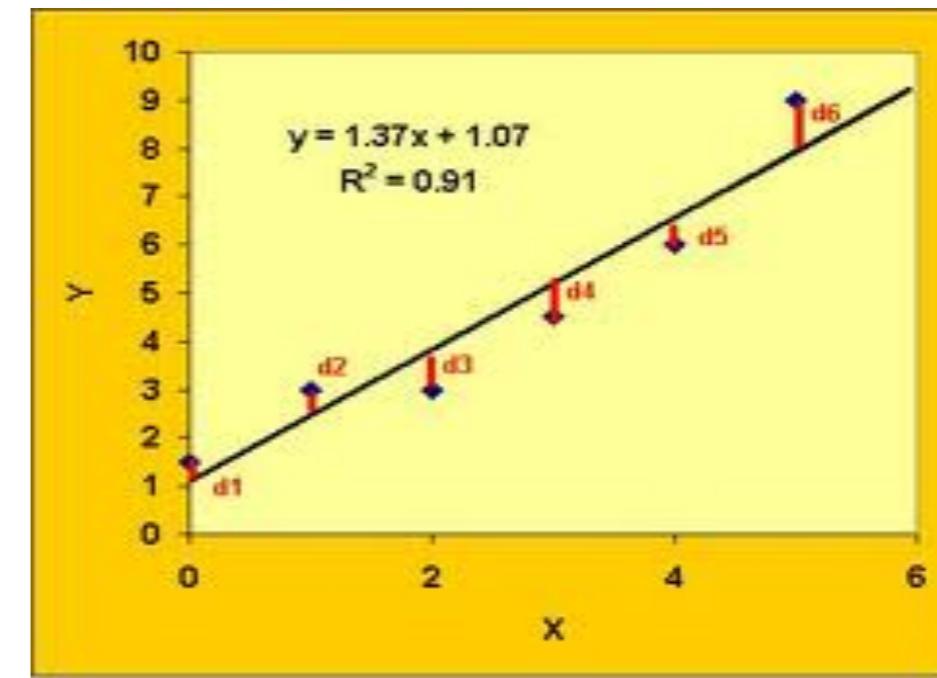
<http://v1shal.com/content/25-cartoons-give-current-big-data-hype-perspective/>



# *Diagnostic Analytics*

---

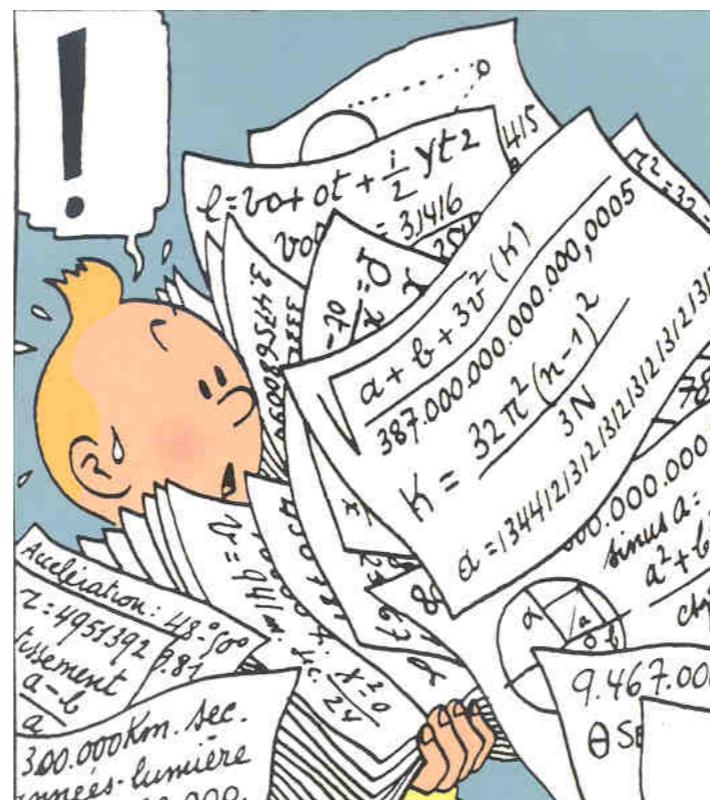
- **Process:**
  - Begin with descriptive analytics
  - Extract patterns from large data quantities via data mining
  - Correlate data types for explanation of near-term behavior – past and present
  - Estimate linear/non-linear behavior not easily identifiable through other approaches.
- Example: by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



# **Predictive Analytics**

- **Process:**

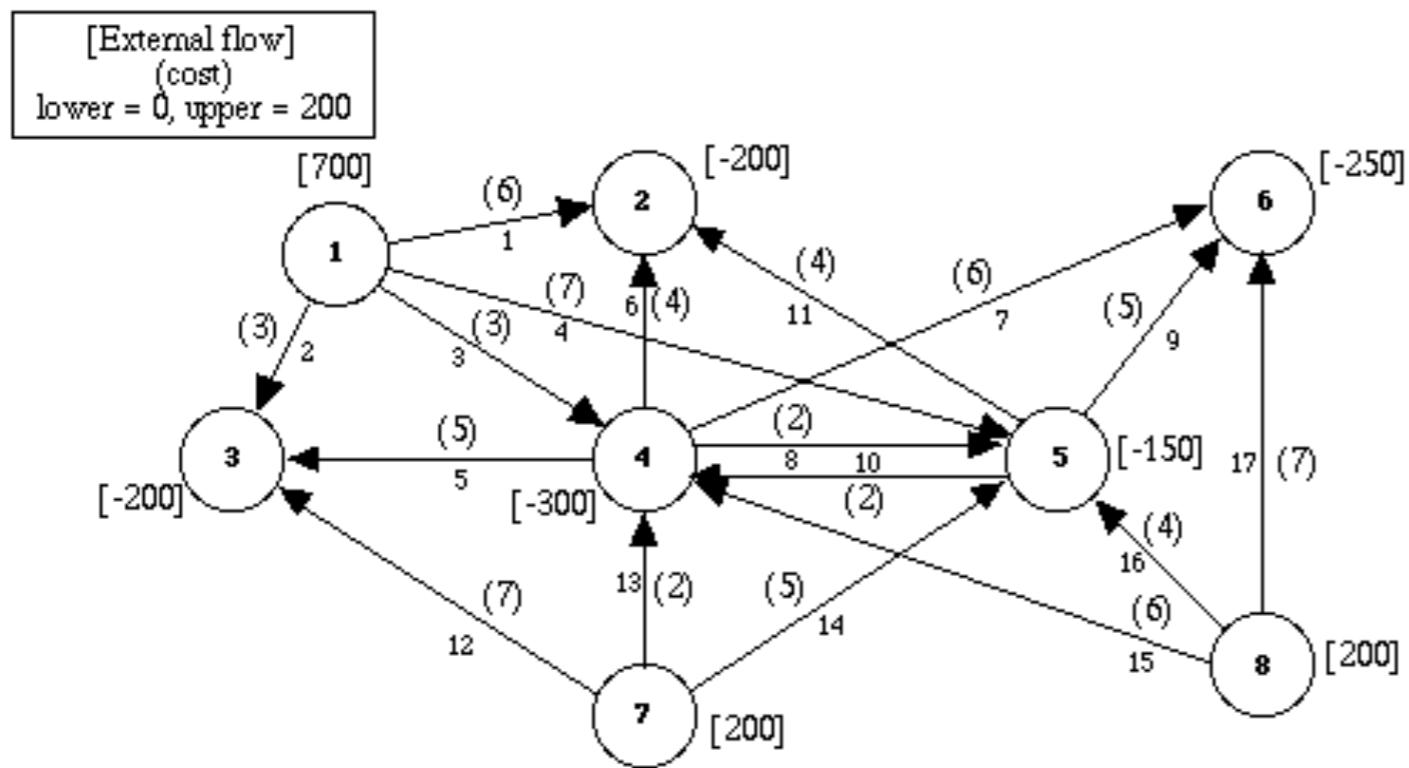
- Begin with descriptive AND diagnostic analytics
  - Choose the right data based on domain knowledge and relationships among variables
  - Choose the right techniques to yield insight into possible outcomes
  - Determine the likelihood of possible outcomes given initial boundary conditions
  - Remember! 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



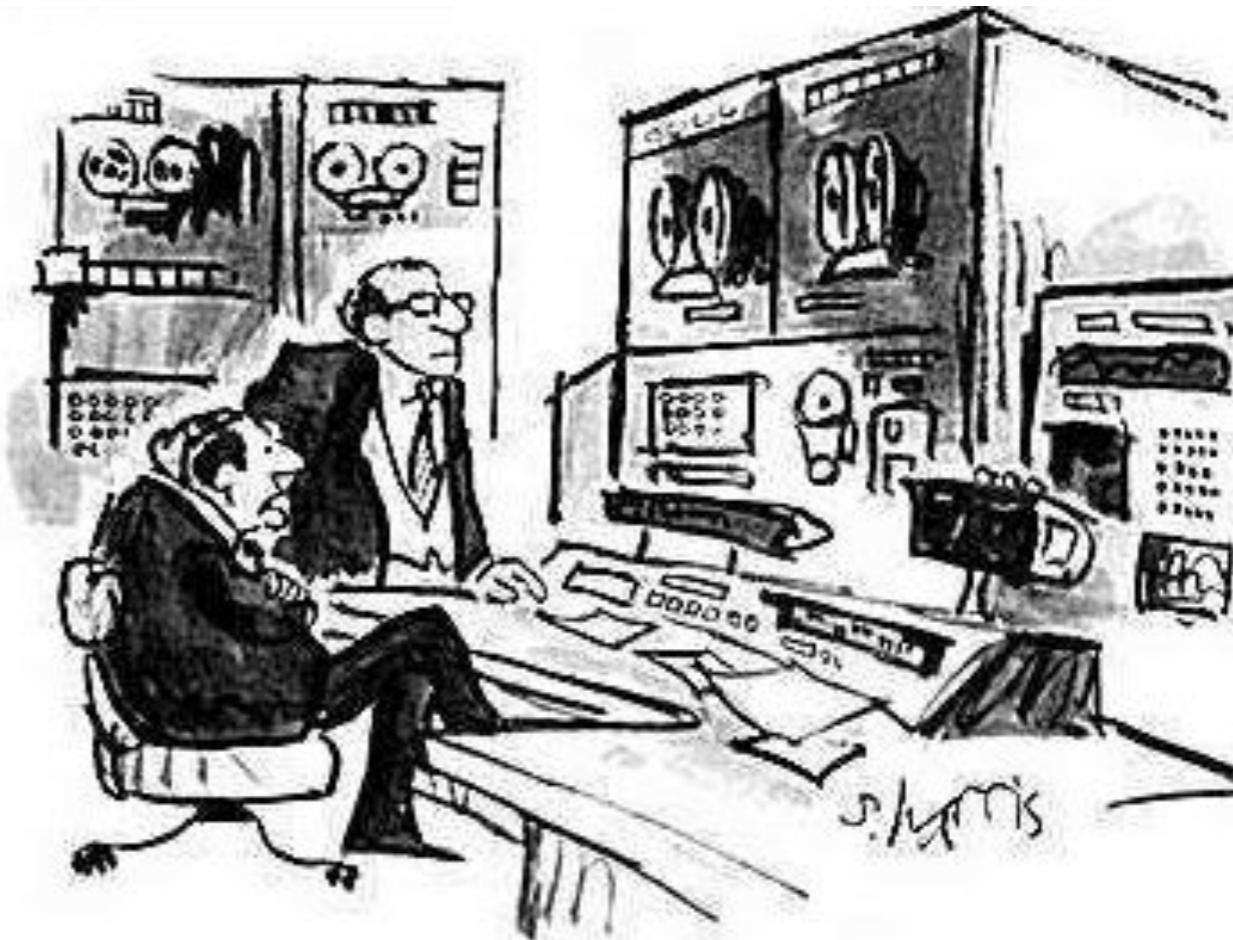
“Make it so”

# *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.”

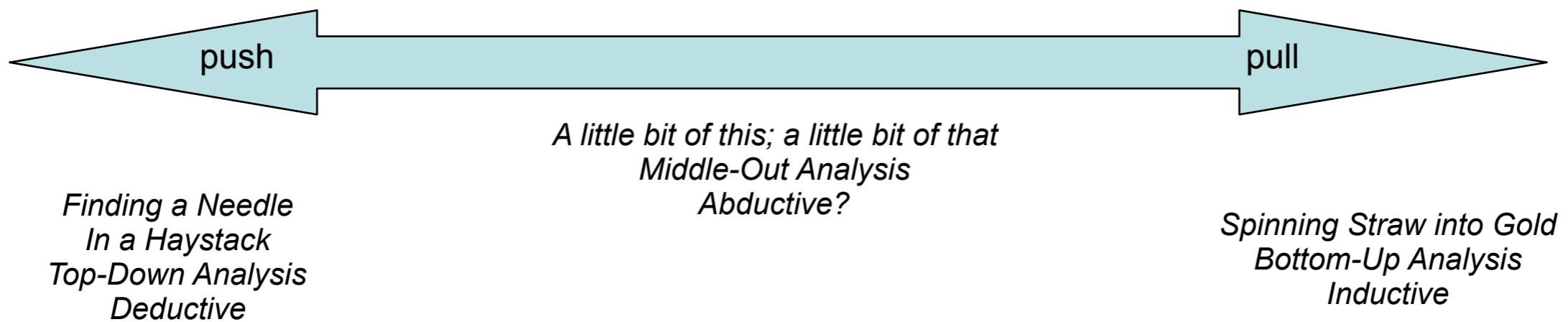
Richards Heuer, Jr., “The Psychology of Intelligence Analysis”

- 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

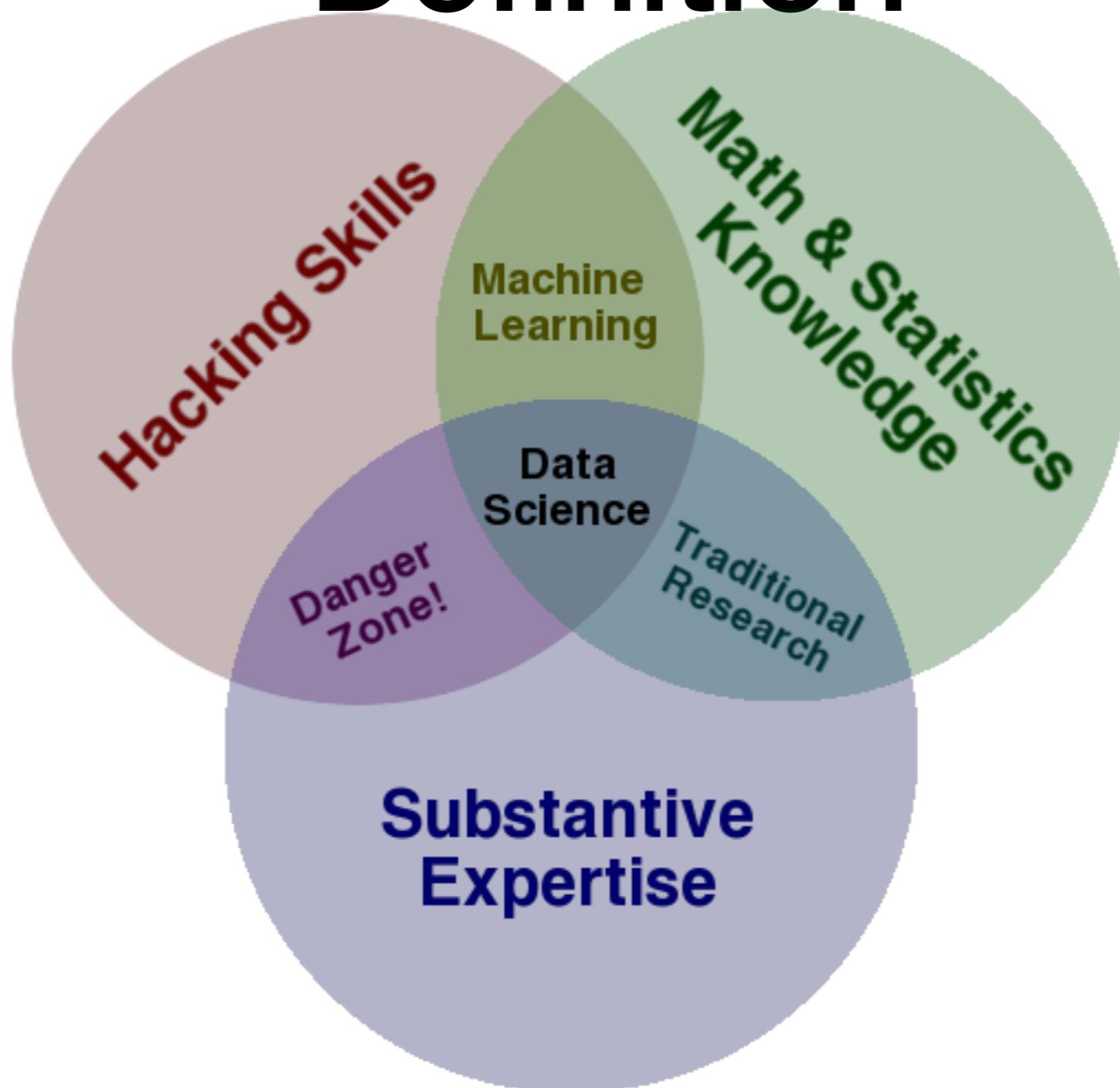
# *The Analytics Continuum*

---

- 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.



# 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), NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB,
Realizations	SQL	

# **Doing Data SciENCE**

# Ben Fry's Model

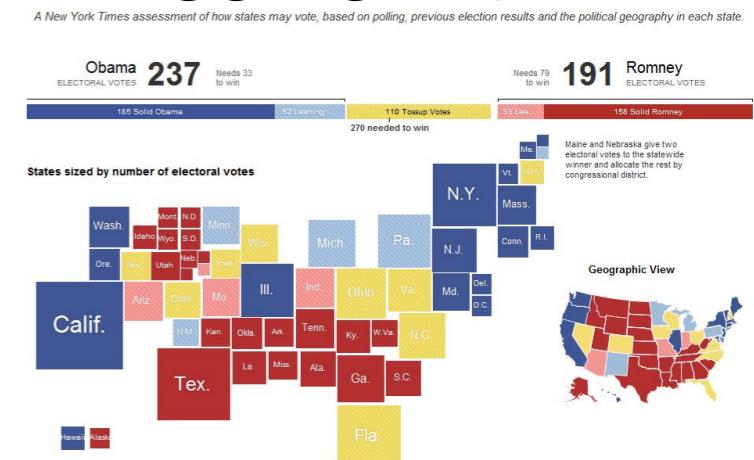
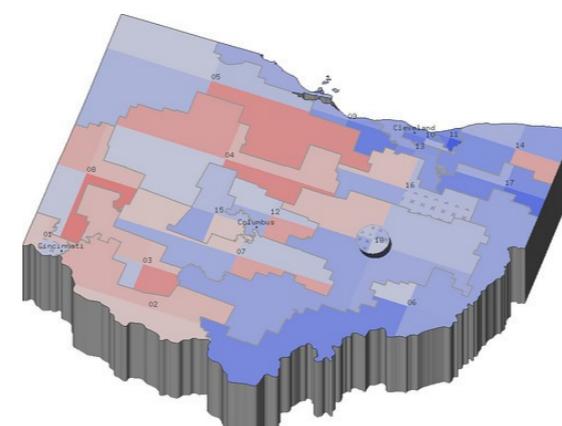
1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

# 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

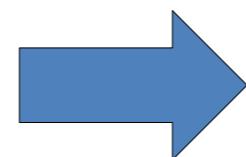


# Data Scientist's Practice



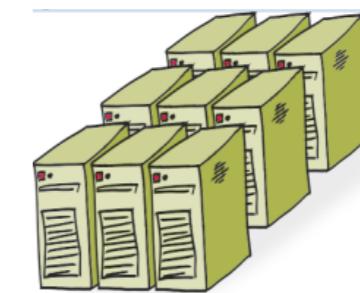
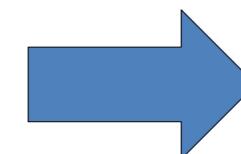
Digging Around  
in Data

Clean,  
prep

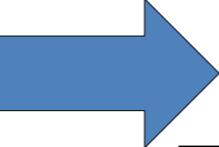
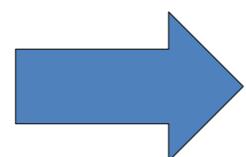


$$\begin{bmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

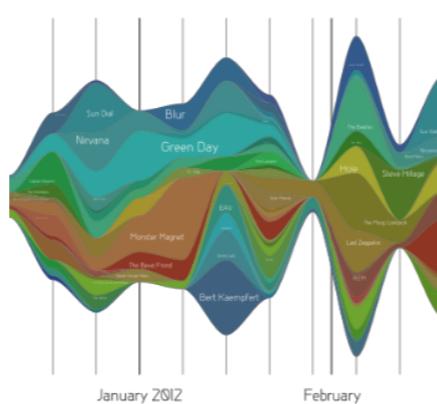
Hypothesize  
Model



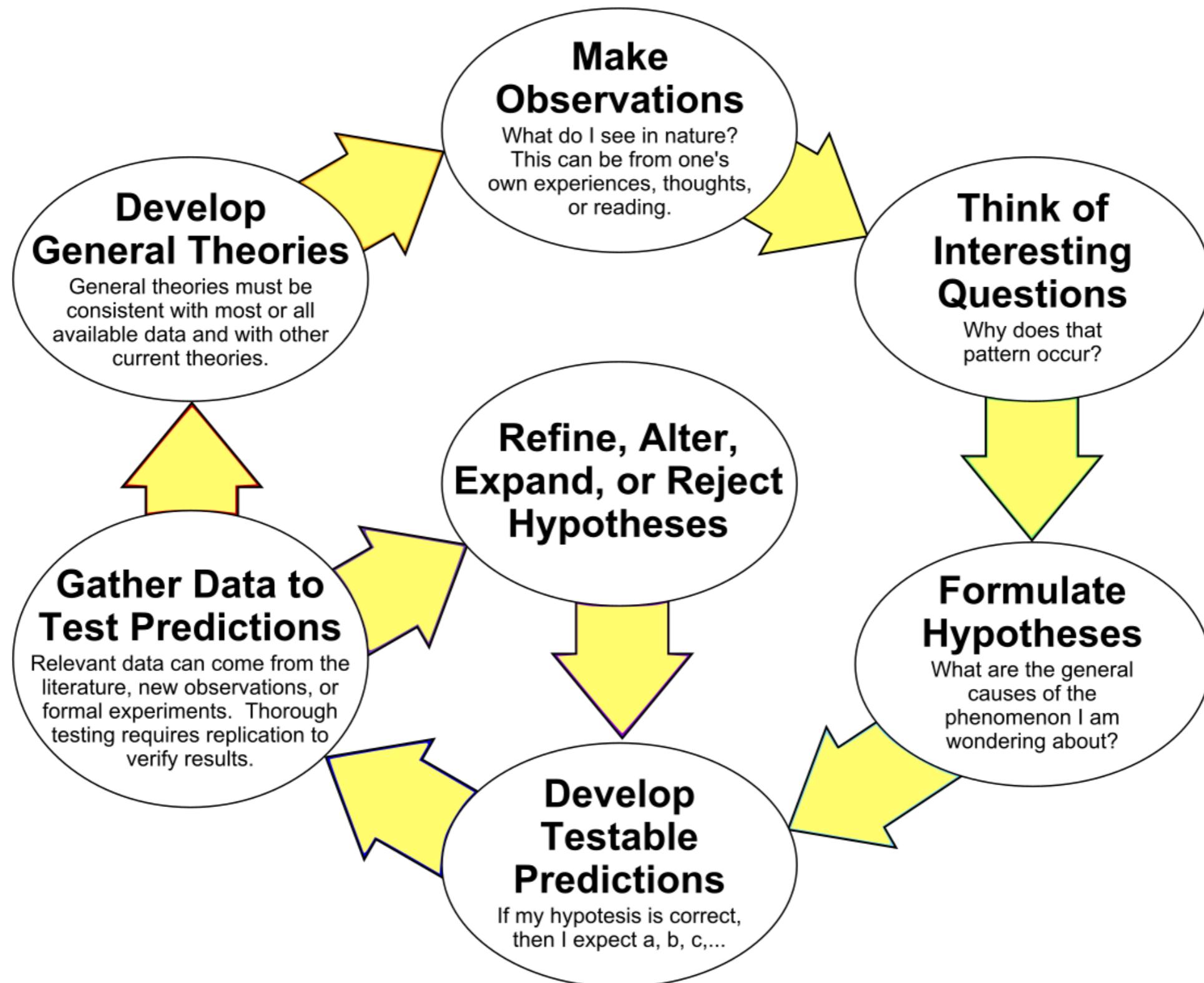
Large Scale  
Exploitation



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?)

