data blueprint

UNLOCKING BUSINESS VALUE

Demystifying Big Data

It's about separating the signal from the noise

Presented by Peter Aiken, Ph.D.

VCUbusiness

datablueprint.com

10124C West Broad Street
Glen Allen, Virginia 23060
+1.804.521.4056

The Distinguished Speakers Program is made possible by

Association for Computing Machinery

Advancing Computing as a Science & Profession

For additional information, please visit http://dsp.acm.org/

About ACM

ACM, the Association for Computing Machinery is the world's largest educational and scientific computing society, uniting educators, researchers and professionals to inspire dialogue, share resources and address the field's challenges.

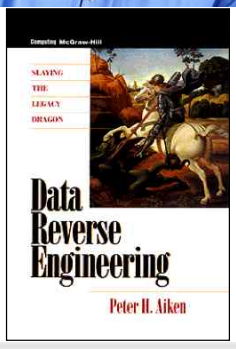
ACM strengthens the computing profession's collective voice through strong leadership, promotion of the highest standards, and recognition of technical excellence.

ACM supports the professional growth of its members by providing opportunities for life-long learning, career development, and professional networking.

With over 100,000 members from over 100 countries, ACM works to advance computing as a science and a profession.
www.acm.org

Peter Aiken, Ph.D.

- 30+ years of experience in data management
- Multiple international awards & recognition
- Founder, Data Blueprint (datablueprint.com)
- Associate Professor of IS, VCU (vcu.edu)
- (Past) President, DAMA Int. (dama.org)
- 9 books and dozens of articles
- Experienced w/ 500+ data management practices in 20 countries
- Multi-year immersions with organizations as diverse as the US DoD, Nokia, Deutsche Bank, Wells Fargo, Walmart, and the Commonwealth of Virginia



Someone Notable Thinks DM is Important

excerpt from President Obama's 2012 State of the Union Address

Adapted from John Ladley

Various Maturity Frameworks

Data is the new oil!
Data is the new (s)oil!
Data is the new bacon!

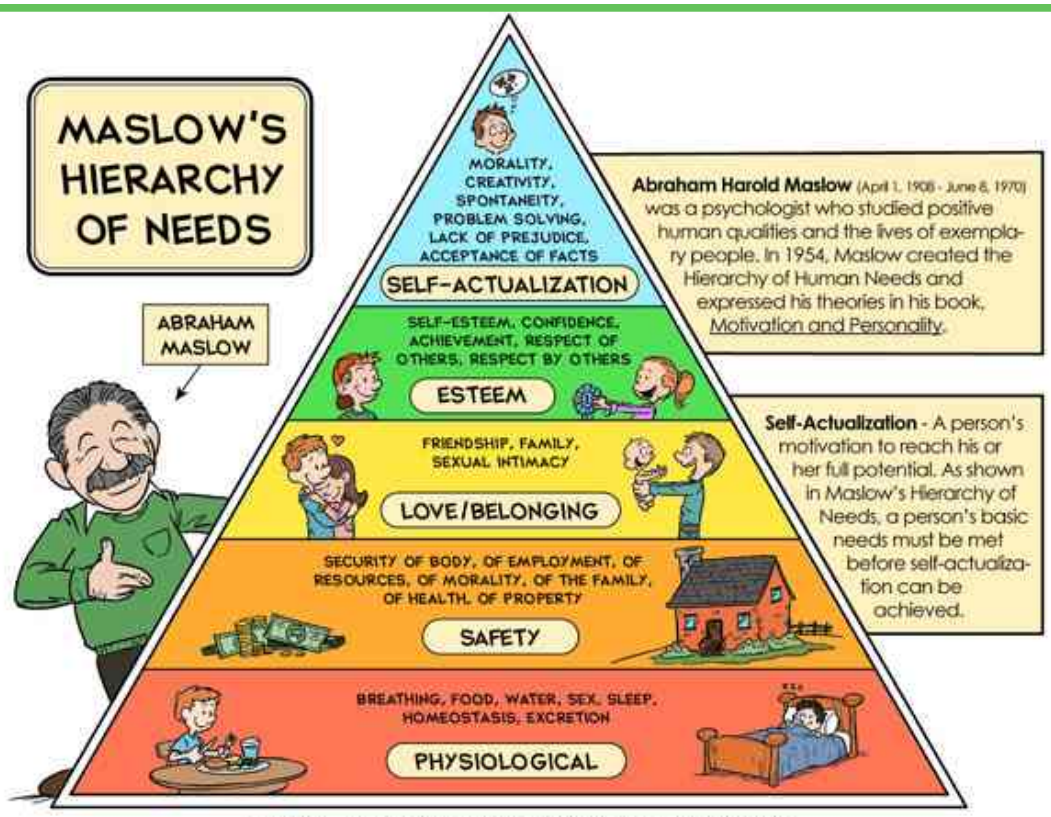
Usage basis	Make it Happen	Make it Happen Faster	What Happened ?	Why did it happen?	What will happen?	Make it happen by itself	What do I want to happen?	How do we make it happen better?	What should we do next?
Content basis	Events	Trans-actions	Reporting	Analyzing	Predictive	Operation-alize	Closed loop	Collabor - ative	Foresight
Capability basis	Initial	Repeatable		Defined	Managed		Optimized		
Organization Basis	Operate:		Consolidate		Integrate		Optimize		Innovate

Data is a lubricant!

Data becomes a fuel!

Big Data is like teenage sex

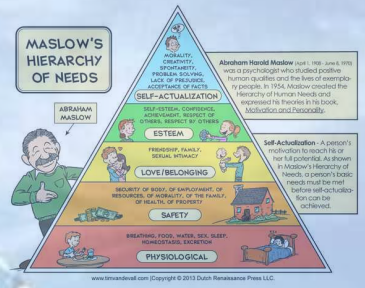
- Everyone talks about it
- Nobody really knows how to do it
- Everyone thinks everyone else is doing it
- So everyone claims they are doing it
  - Dan Ariely (via facebook)



Data Management Practices Hierarchy

You can accomplish Advanced Data Practices without becoming proficient in the Basic Data Management Practices however this will:

- Take longer
- Cost more
- Deliver less
- Present greater risk



Advanced Data Practices

- MDM
- Mining
- Big Data
- Analytics
- Warehousing
- SOA

Basic Data Management Practices

Data Stewardship

Data Development

Organizational Data Integration

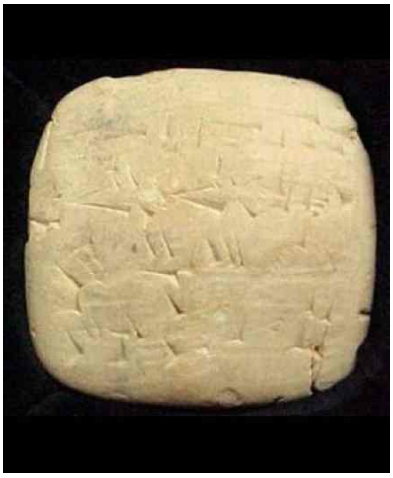
Data Program Management

Data Support Operations

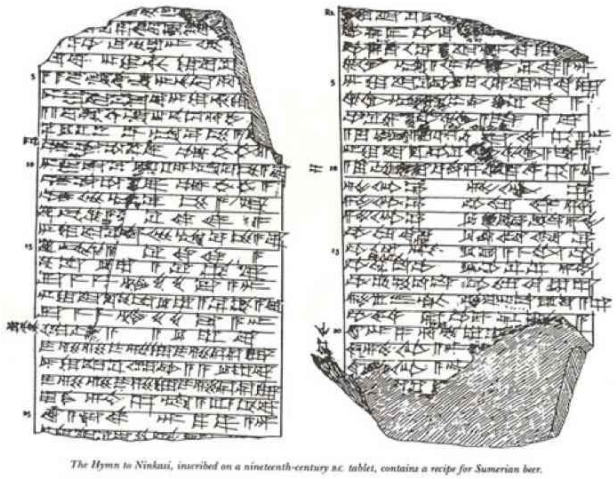
Demystifying Big Data

- Data Analysis
  - Origins
- Challenges
  - Faced by virtually everyone
- Compliment
  - Existing data management practices
- Pre-requisites
  - Necessary to exploit big data techniques
- Prototyping
  - Iterative means of practicing big data techniques
- Take Aways and Q&A

Old Beer Accounting



This records a purchase of "best" beer from a brewer, c. 2050 BC from the Sumerian city of Umma in Ancient Iraq
http://en.wikipedia.org/wiki/File:Alulu_Beer_Receipt.jpg



The Hymn to Ninkasi, inscribed on a nineteenth-century B.C. tablet, contains a recipe for Sumerian beer.

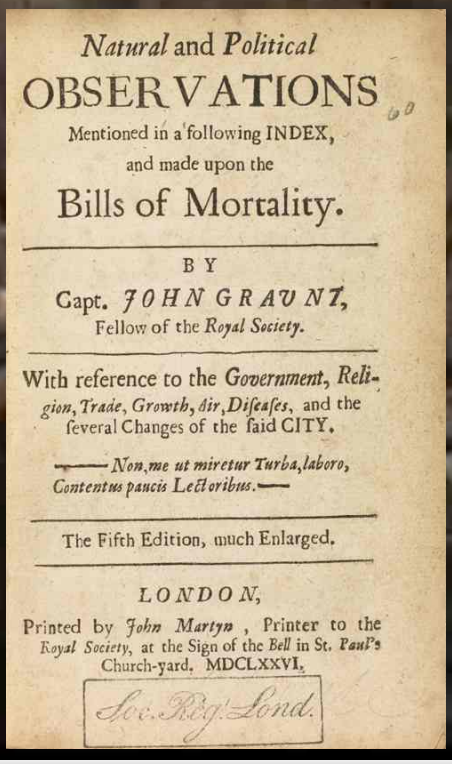
The first references to beer dates to as early as 6,000 BC. The very first recipe for beer is found on a 4,000-year-old Sumerian tablet containing the Hymn to Ninkasi, a prayer to the goddess of brewing.
http://www.neatorama.com/2009/02/18/neatolicious-fun-facts-beer/#!kN0hf

Tally Sticks

- From around 6,000 B.C.E.
- Notches in a divided stick guaranteed the authenticity of accounting data!
- Cutting notches, representing a certain amount of money, across the width of a stick
- Split stick lengthwise
  - Debtor takes one half (tally)
  - Debtee taking the other (foil)
- Notches in matching halves guaranteed the authenticity of both side's data
- The word for "contract" in written Chinese is the symbol "large tally stick"

Accounting for Beer Sales

Bills of Mortality by Captain John Graunt



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,
and made upon the

Bills of Mortality.

BY

Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government, Religion, Trade, Growth, Air, Diseases*, and the several Changes of the said CITY.

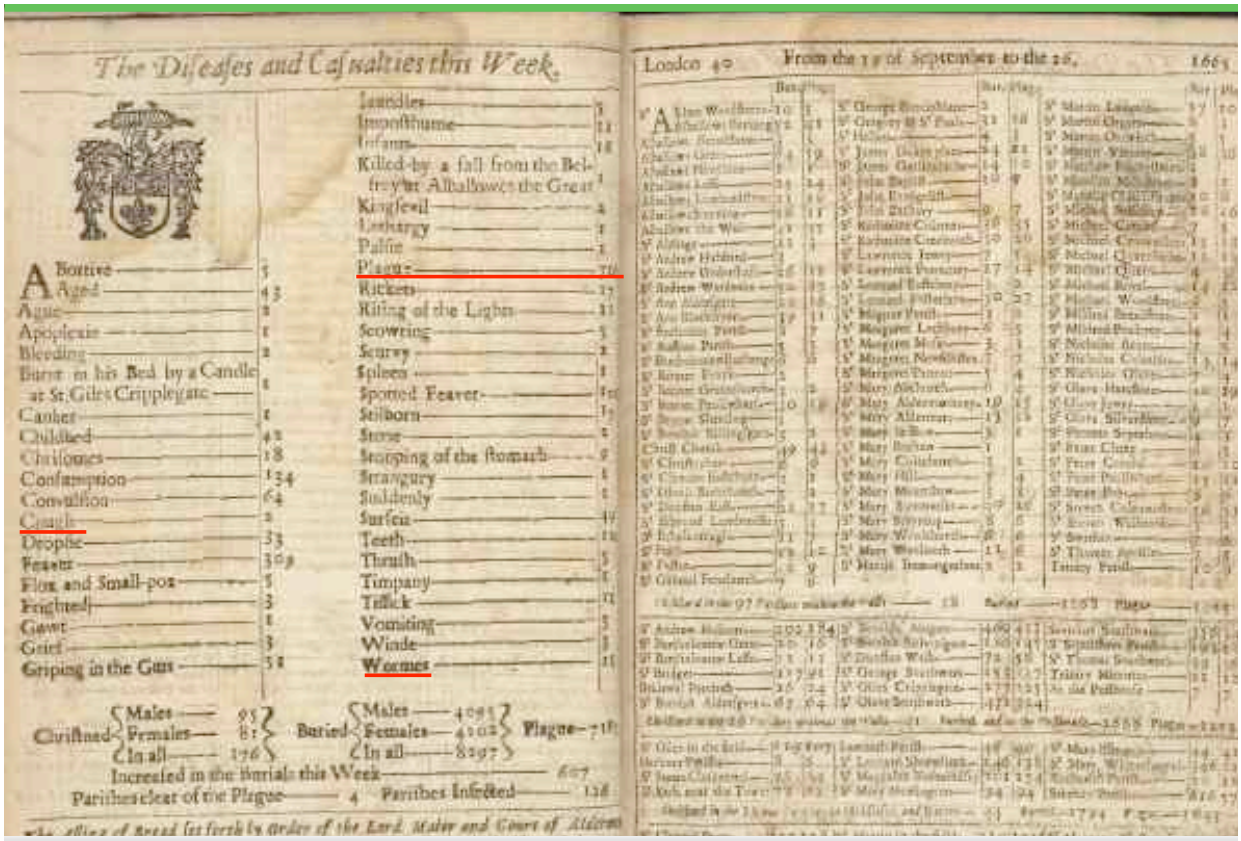
——Non me ut miretur Turba, laboro,
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

LONDON,

Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the *Bell* in St. *Paul's* Church-yard, MDCLXXVI.

Soc. Reg. Lond.



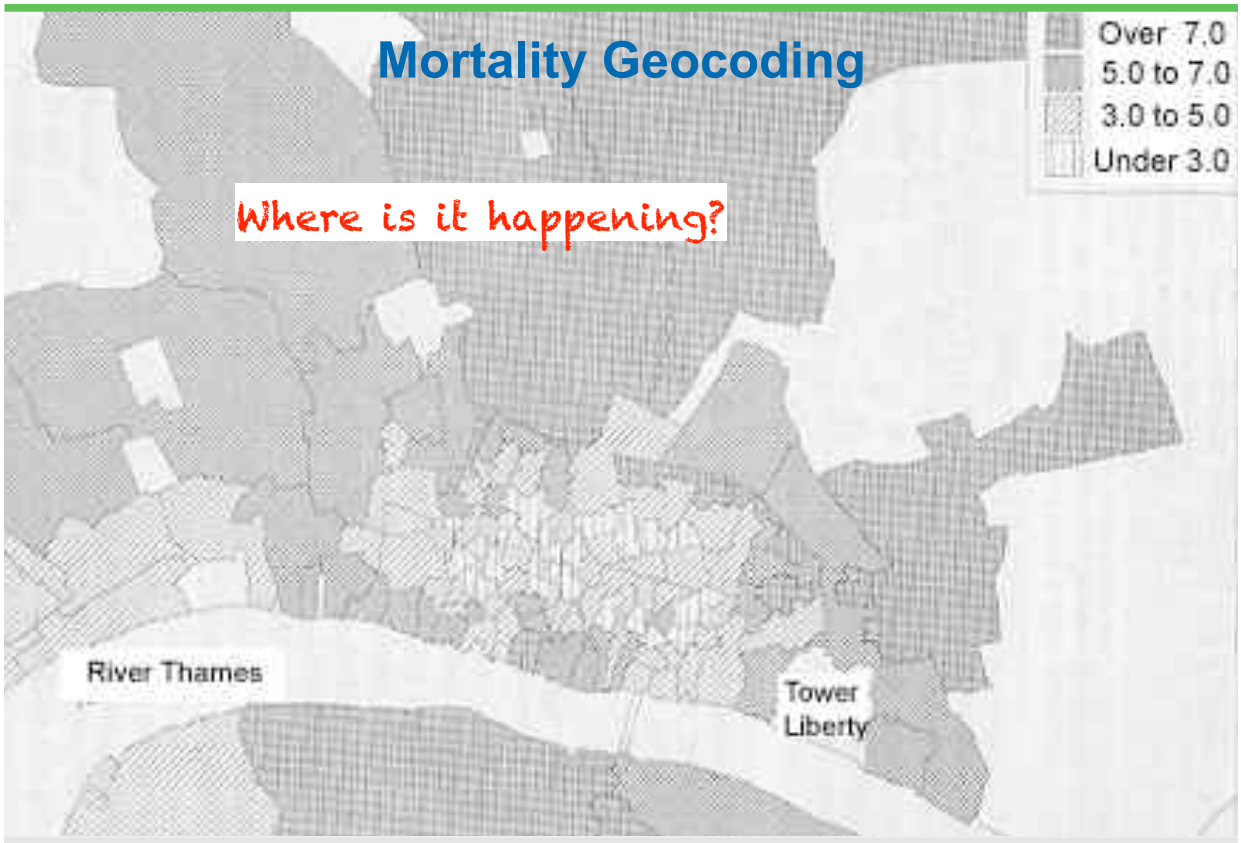
The Diseases and Casualties this Week.

Abortive	5
Aged	43
Ague	2
Apoplexie	1
Bleeding	2
Burnt in his Bed by a Candle at St. Giles Cripplegate	1
Canker	1
Childbed	42
Chrisomes	18
Consumption	134
Convulsion	64
Cripple	2
Dropsie	33
Feaver	309
Flox and Small-pox	5
Frighted	3
Gawt	1
Grief	3
Griping in the Guts	58
Laundice	3
Imposthume	11
Infants	18
Killed by a fall from the Belfrey at Alhallowes the Great	1
Kingsevil	2
Lethargy	1
Palsie	1
Plague	70
Rickets	11
Riting of the Lights	11
Scowring	3
Scurvy	2
Spleen	1
Spotted Feaver	10
Stillborn	7
Stone	2
Stopping of the Stomach	9
Strangury	1
Suddenly	1
Surfeit	49
Teeth	11
Thrush	3
Timpany	1
Tisick	11
Vomiting	3
Winde	3
Wormes	11

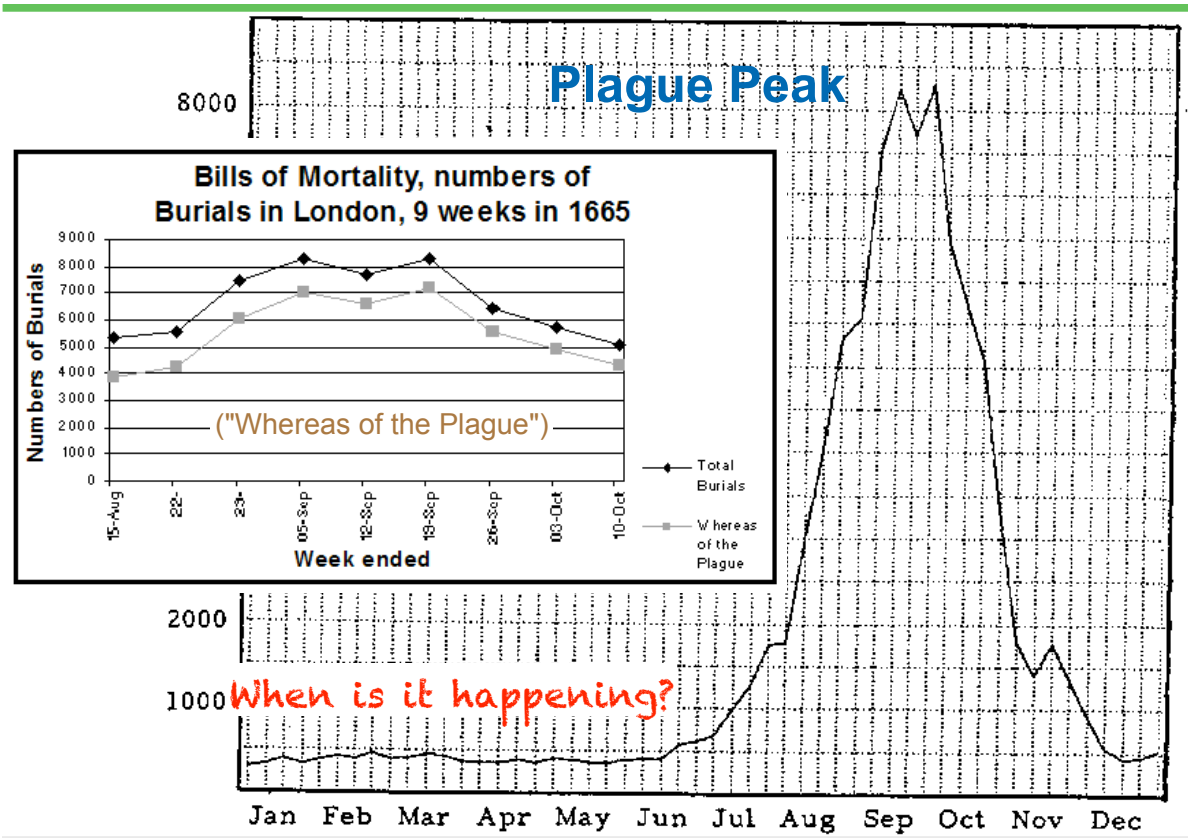
Christned: Males 91, Females 87, In all 176
Buried: Males 4081, Females 4102, In all 8297, Plague 7165
Increased in the Burials this Week 607
Parishes clear of the Plague 4 Parishes Infected 128

London 49 From the 19 of September to the 26. 1665

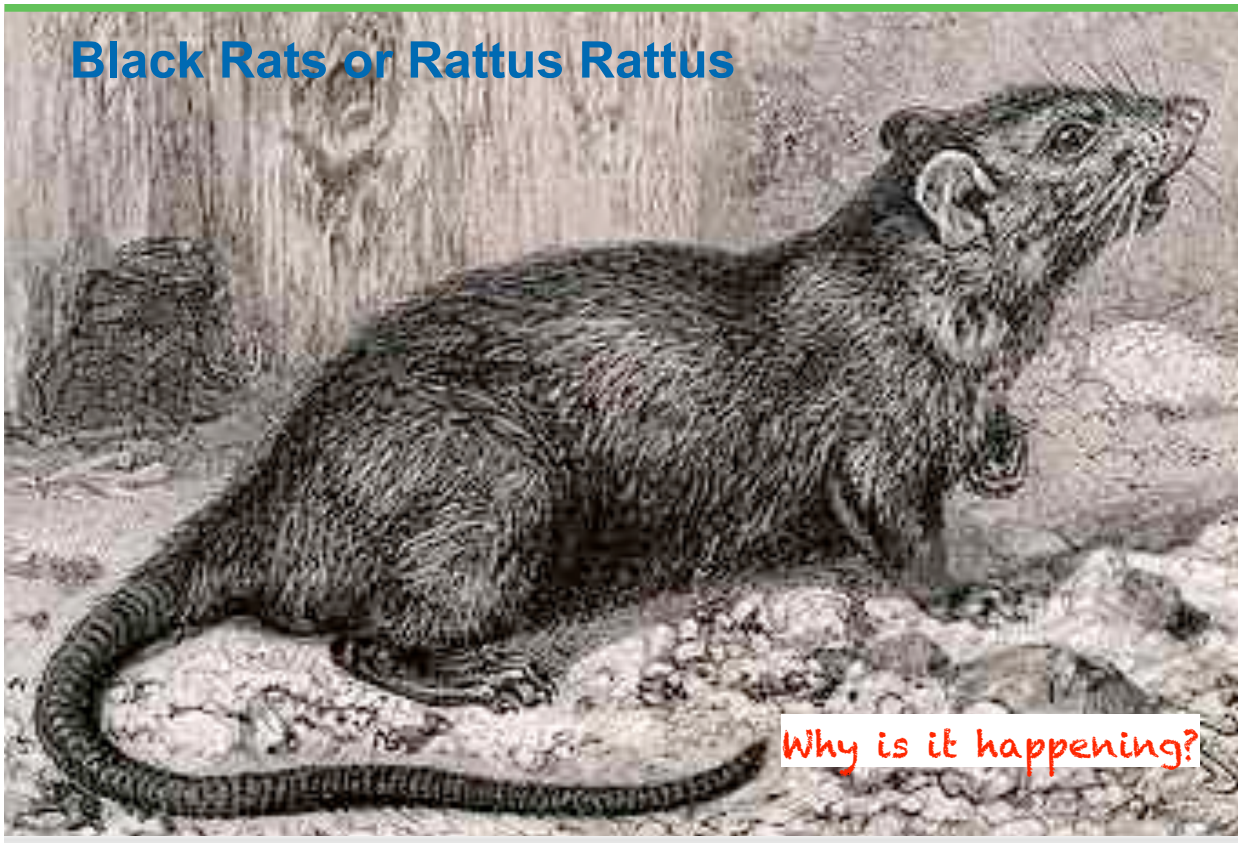
Mortality Geocoding

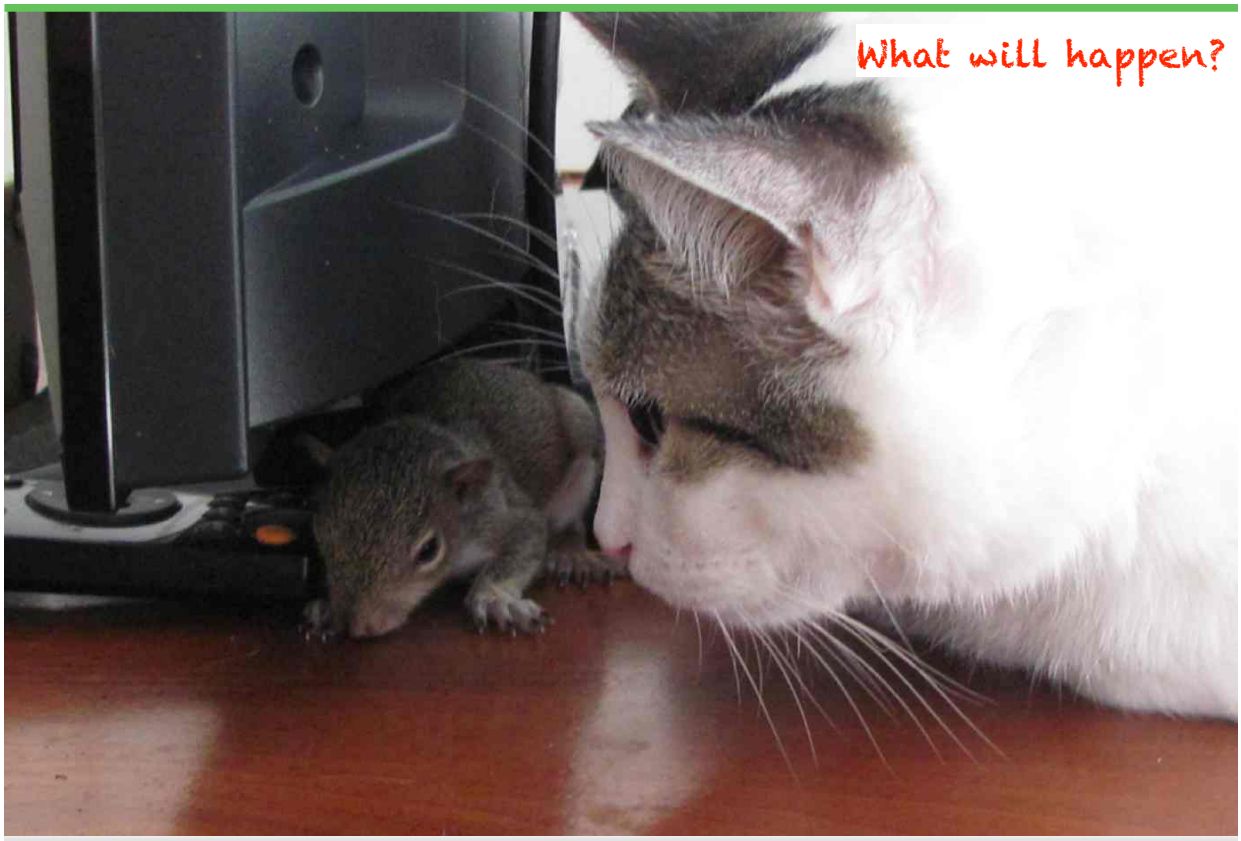


Plague Peak

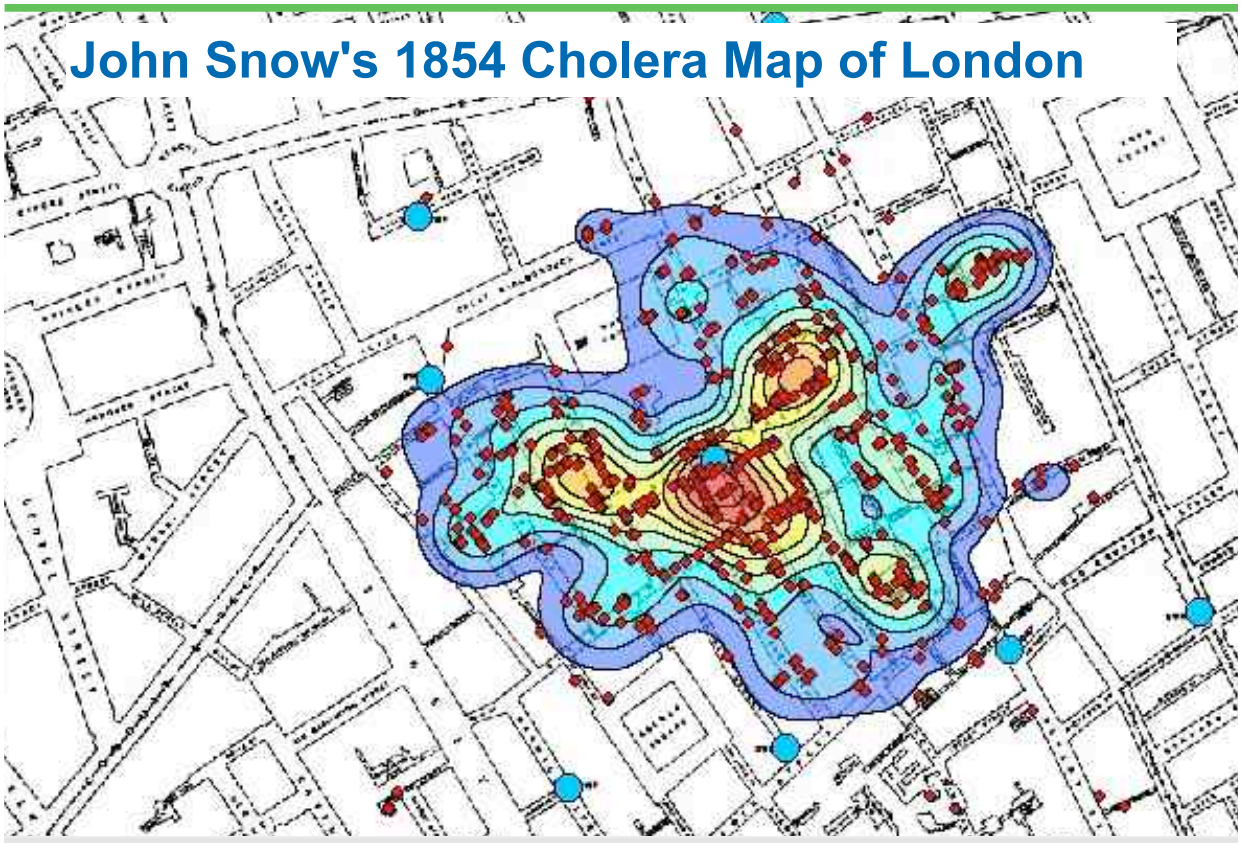


Black Rats or Rattus Rattus

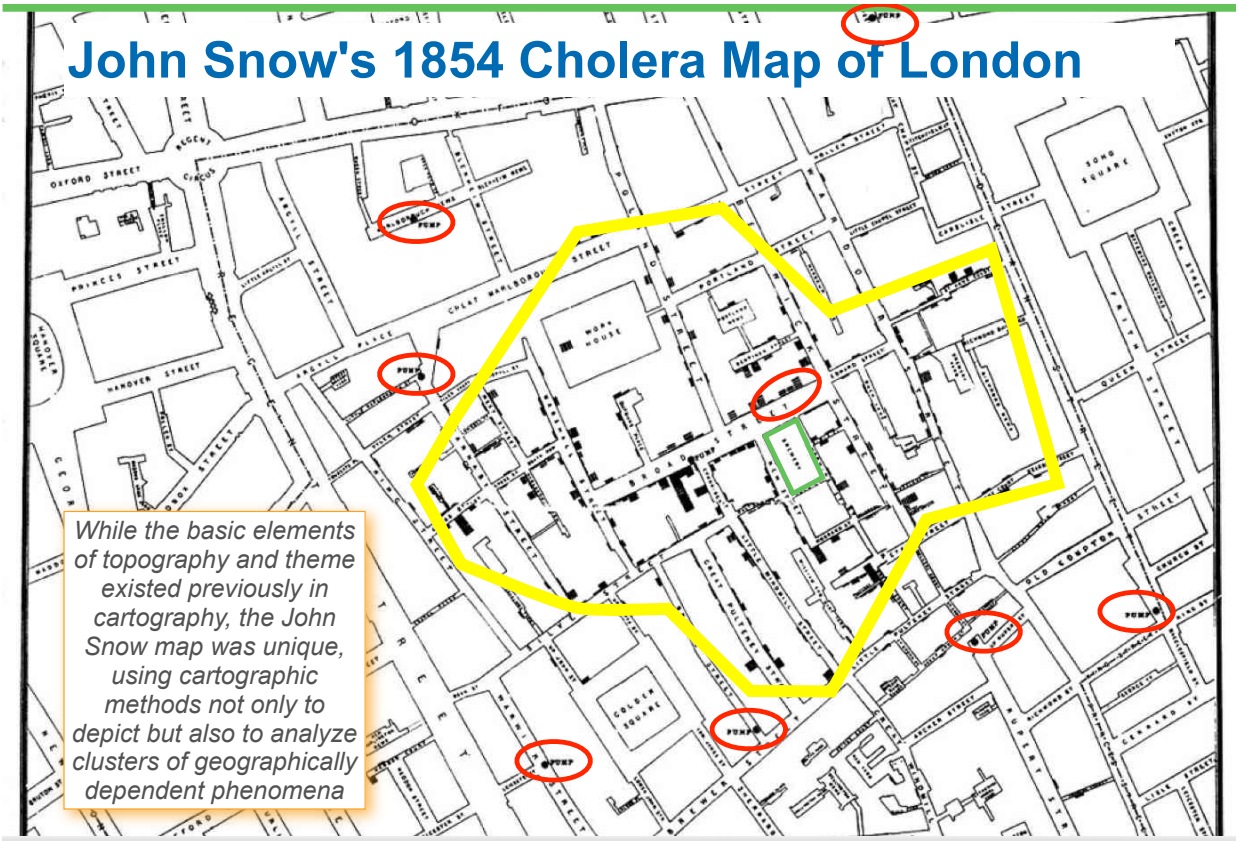




John Snow's 1854 Cholera Map of London



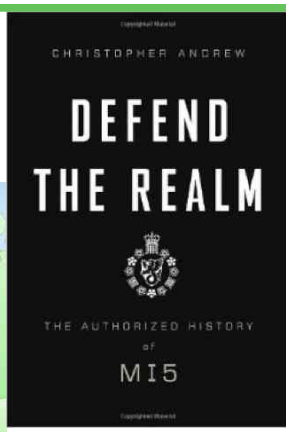
John Snow's 1854 Cholera Map of London



While the basic elements of topography and theme existed previously in cartography, the John Snow map was unique, using cartographic methods not only to depict but also to analyze clusters of geographically dependent phenomena

Formalizing Data Management

- *Defend the Realm: The authorized history of MI5* by Christopher Andrew
- World War I
- 1914
- At war with much of Europe
- 14,000,000 Germans living in the United Kingdom
- How to efficiently and effectively manage information on that many individuals?
- The Security Service is responsible for "protecting the UK against threats to national security from espionage, terrorism and sabotage, from the activities of agents of foreign powers, and from actions intended to overthrow or undermine parliamentary democracy by political, industrial or violent means."



SECRET INTELLIGENCE SERVICE MI6

Bringing the Computer into Intelligence Work

Some Far-out Thoughts on Computers

Orrin Clotworthy

Originally published in *Studies in Intelligence* Vol. 6, No. 4 (1962)

A Jules Verne look at intelligence processes in a coming genera

- Predicted use of not just computing in the intelligence community
- Also forecast predictive analytics
- Accompanying privacy challenges

Question: What does the size of the next coffee crop, bullfight attendance figures, local newspaper coverage of UN matters, the birth rate, the mean daily temperatures or refrigerator sales across the country have to do with who will next be elected president of Guatemala?

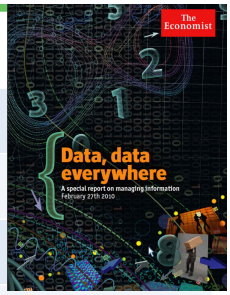
Answer: Perhaps nothing. But the question is not a frivolous one. There must be a cause behind each vote cast in an election. It may be a rational, emotional, superstitious, or accidental cause. The choice may derive from months of conscious effort to weigh the pros and cons of the aspirants to office. It may be an automatic, tradition-bound action that requires not even a cursory exercise of the thought process. Or the voter himself may not recognize why he decides as he does. But something will motivate him, and it may be closely correlative with one or more of the quantitative factors suggested in the opening question.

To learn just what the factors are, how to measure them, how to weight them, and how to keep them flowing into a computing center for continual analysis will some day become a matter of great concern to all of us in the intelligence community. I say "will" rather than "may" because it seems to me that this type of election analysis will be only the first faltering step by an infant quantified behavioral science that is going to be forced on us for its upbringing like a doorstep baby—and soon.

Demystifying Big Data

- Data Analysis
  - Origins
- Challenges
  - Faced by virtually everyone
- Compliment
  - Existing data management practices
- Pre-requisites
  - Necessary to exploit big data techniques
- Prototyping
  - Iterative means of practicing big data techniques
- Take Aways and Q&A

Data Inflation



Unit	Size	What it means
Bit (b)	1 or 0	Short for “binary digit”, after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From “thousand” in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From “large” in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From “giant” in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From “monster” in Greek. All the catalogued books in America’s Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America’s postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of The Economist
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Source: *The Economist* Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established

AT&T Mobile Data Volumes Up 8,000% Over Four Years

PB/month

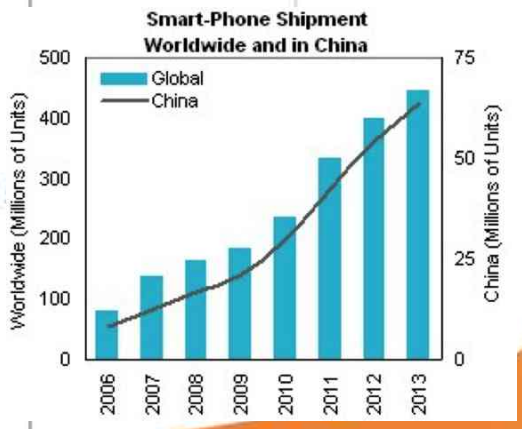
10

0

What do they mean big?

"*Every 2 days we create as much information as we did up to 2003*"

– Eric Schmidt



IP traffic will quadruple by 2015

– Asigra 2012

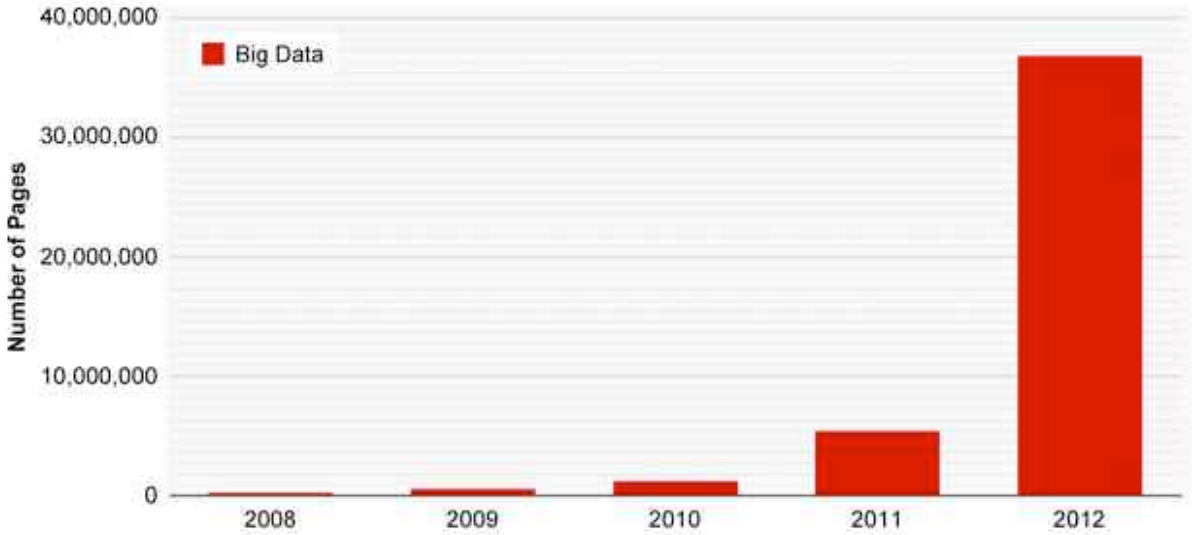
The number of things that can produce data is rapidly growing (smart phones for example)

2007 2008 2009 2010

Google Search Results for the Term "Big Data"

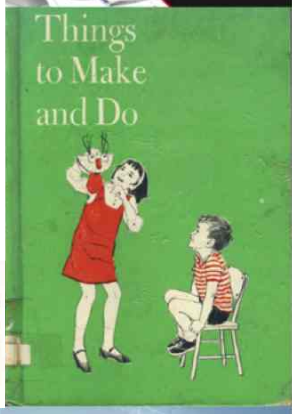
Data from Google Trends Source: Gartner (October 2012)

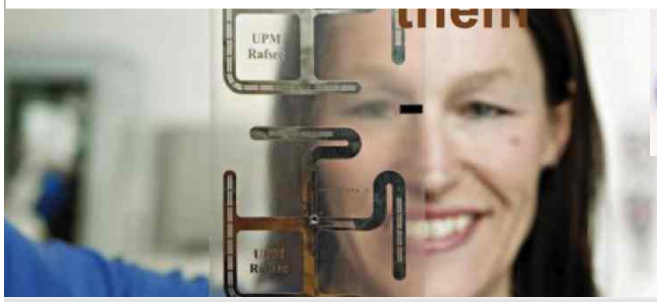
Number of Internet Pages Mentioning Big Data

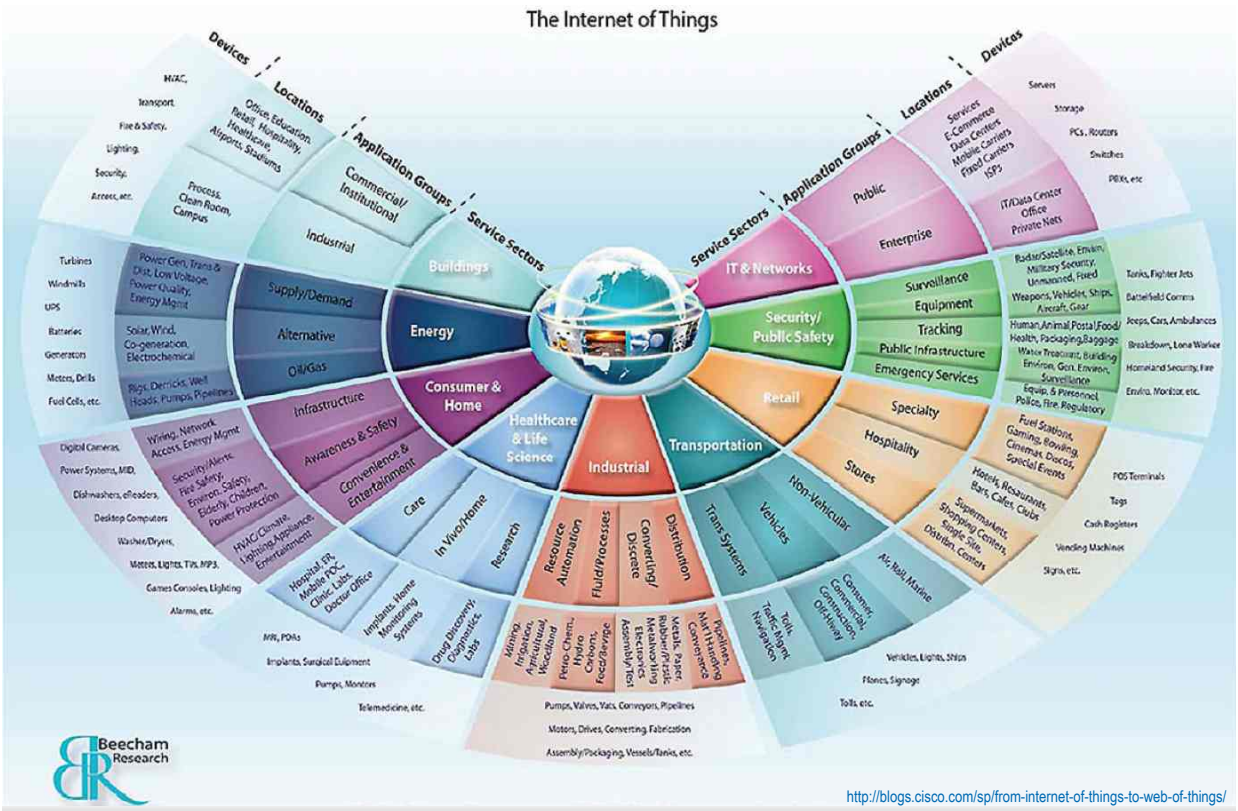


Data from Google Trends Source: Gartner (October 2012)

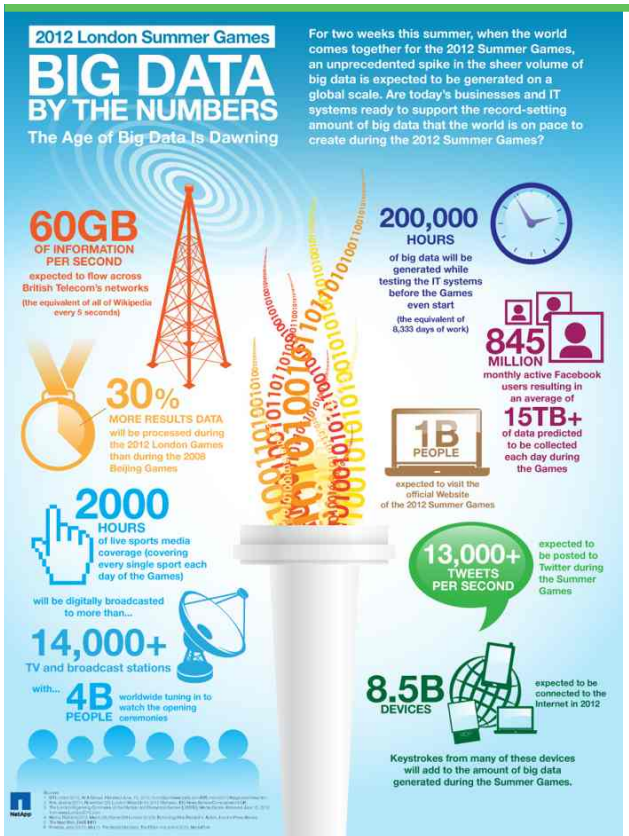
Increasingly individuals make use of the things data producing capabilities to perform services for them including context, mobile, data, sensors and location-based technology







http://blogs.cisco.com/sp/from-internet-of-things-to-web-of-things/



2012 London Summer Games

- 60 GB of data/second
- 200,000 hours of big data will be generated testing systems
- 2,000 hours media coverage/daily
- 845 million Facebook users averaging 15 TB/day
- 13,000 tweets/second
- 4 billion watching
- 8.5 billion devices connected

Data Footprints

- SQL Server
  - 47,000,000,000,000 bytes
  - Largest table 34 billion records 3.5 TBs
- Informix
  - 1,800,000,000 queries/day
  - 65,000,000 tables / 517,000 databases
- Teradata
  - 117 billion records
  - 23 TBs for one table
- DB2
  - 29,838,518,078 daily queries

Sloan Management Review/Harvard Business Review

[DATA AND ANALYTICS]

How 'Big Data' Is Different

These days, lots of people in business are talking about "big data." But how do the potential insights from big data differ from what managers generate from traditional analytics?

BY THOMAS H. DAVENPORT, PAUL BARTH AND RANDY BEAN

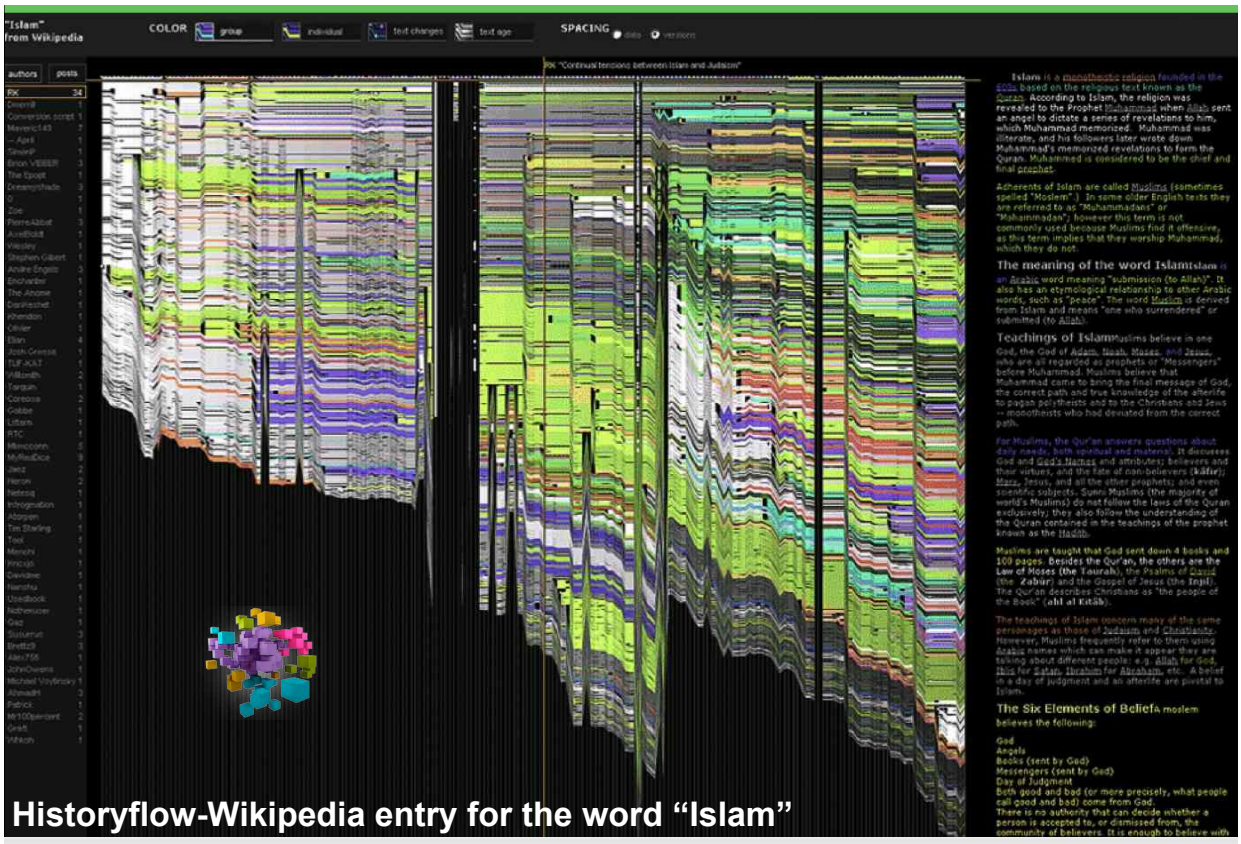
Big data encompasses everything from clickstream data from the Web to genomic and proteomic data from biological research and medicine.

These days, many people in the information technology world and in corporate boardrooms are talking about "big data." Many believe that, for companies that get it right, big data will be able to unleash new organizational capabilities and value. But what does the term "big data" actually entail, and how will the insights it yields differ from what managers might generate from traditional analytics?

There is no question that organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analyzed

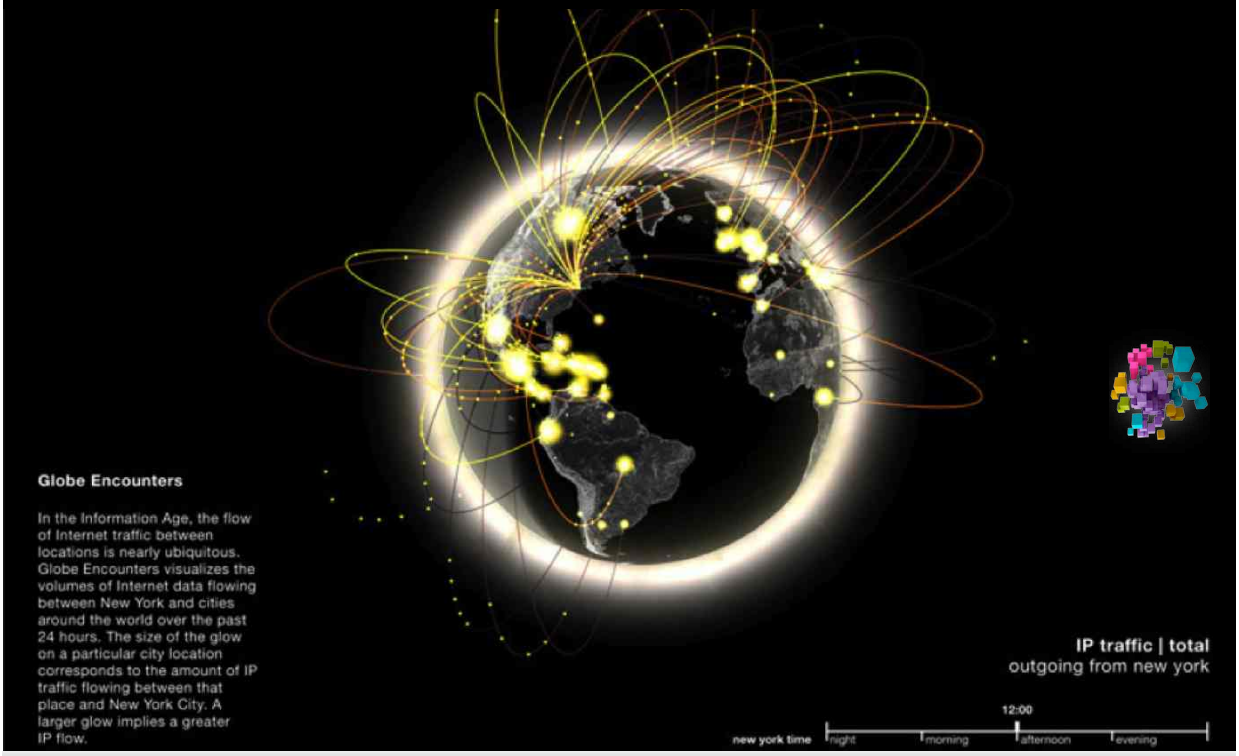
Big Data (has something to do with Vs - doesn't it?)

- **Volume**
  - Amount of data
- **Velocity**
  - Speed of data in and out
- **Variety**
  - Range of data types and sources
    - 2001 *Doug Laney*
- **Variability**
  - Many options or variable interpretations confound analysis
    - 2011 *ISRC*
- **Vitality**
  - A dynamically changing Big Data environment in which analysis and predictive models must continually be updated as changes occur to seize opportunities as they arrive
    - *2011 CIA*
- **Virtual**
  - Scoping the discussion to only include online assets
    - 2012 *Courtney Lambert*
- **Value/Veracity**
    - *Stuart Madnick (John Norris Maguire Professor of Information Technology, MIT Sloan School of Management & Professor of Engineering Systems, MIT School of Engineering)*



Historyflow-Wikipedia entry for the word "Islam"

Spatial Information Flow-New York Talk Exchange



Your goal should be to buy "wins" ...

We are card counters at blackjack table and we are going to turn the odds on the casino!

24 hour observation of all of the large aircraft flights in the world, condensed down to just over a minute

Nanex 1/2 Second Trading Data (May 2, 2013 Johnson and Johnson)

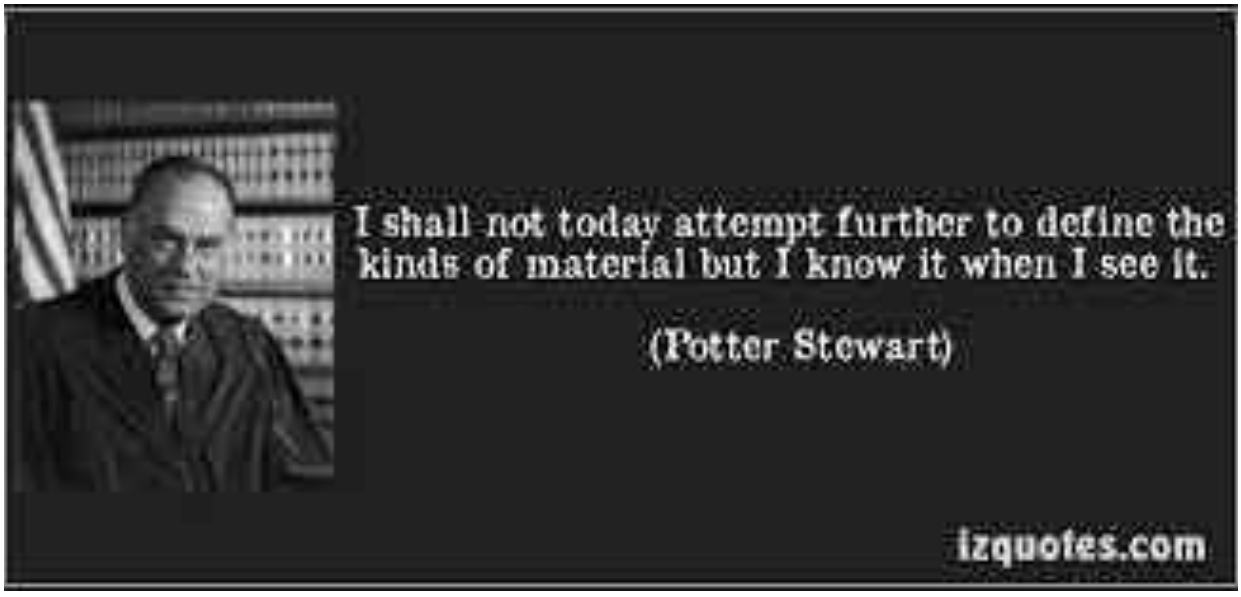
IBM's Data Baby

Surrender to a Buyer Power

Defining Big Data

- Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.
  - Gartner 2012
- Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
  - IBM 2012
- Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, curate, manage, and process the data within a tolerable elapsed time.
  - Wikipedia
- Shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.
  - NY Times 2012
- Big data is about putting the "I" back into IT.
  - Peter Aiken 2007
- *We have no objective definition of big data!*
  - Any measurements, claims of success, quantifications, etc. must be viewed skeptically and with suspicion!

I shall not today attempt further to define the kinds of material but I know it when I see it ... (Justice Potter Stewart)



Big Data

~~Big Data~~

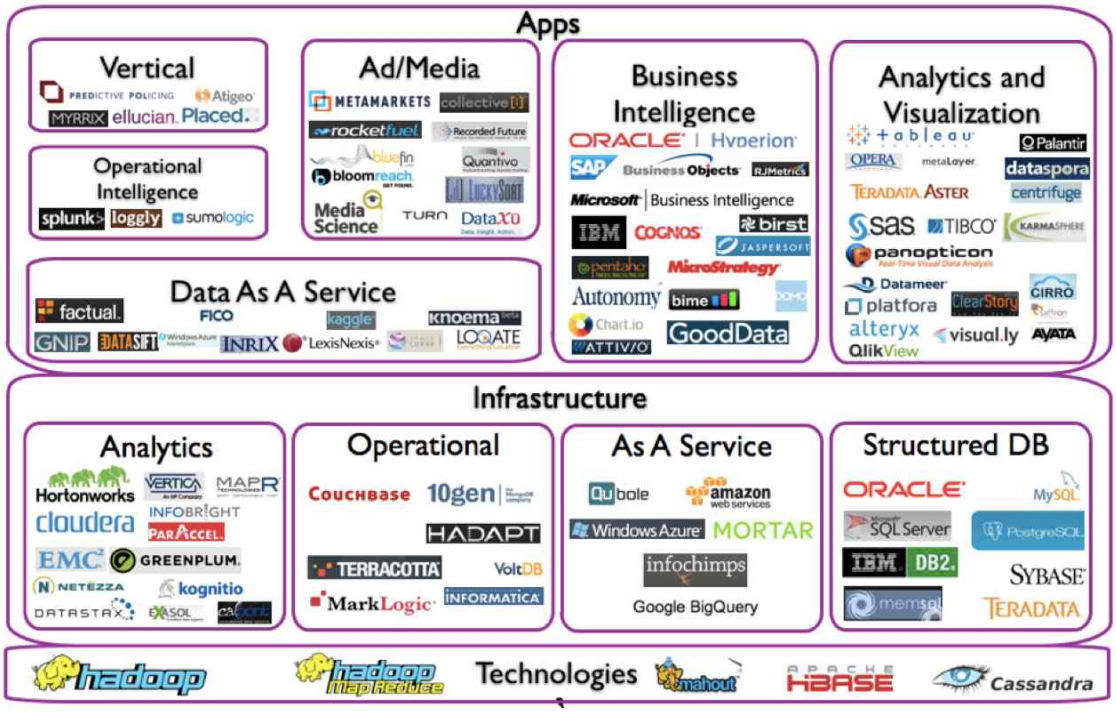
Big Data [Techniques / Technologies]

Big Data Techniques

- New techniques available to impact the productivity (order of magnitude) of any analytical insight cycle that compliment, enhance, or replace conventional (existing) analysis methods
- Big data techniques are currently characterized by:
  - Continuous, instantaneously available data sources
  - Non-von Neumann Processing (defined later in the presentation)
  - Capabilities approaching or past human comprehension
  - Architecturally enhanceable identity/security capabilities
  - Other tradeoff-focused data processing

- So a good question becomes "where in our existing architecture can we most effectively apply Big Data Techniques?"

The Big Data Landscape

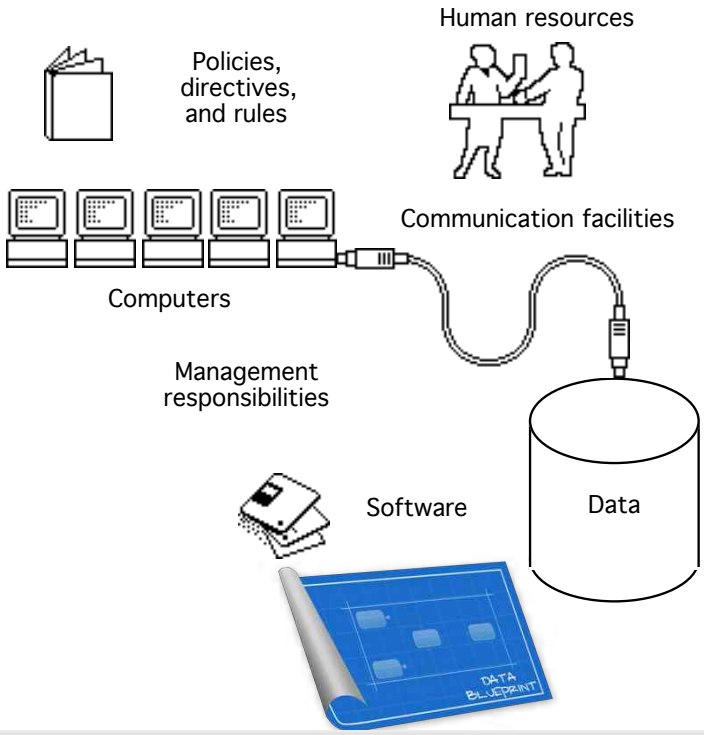


Big Data Technologies by themselves, are a One Legged Stool

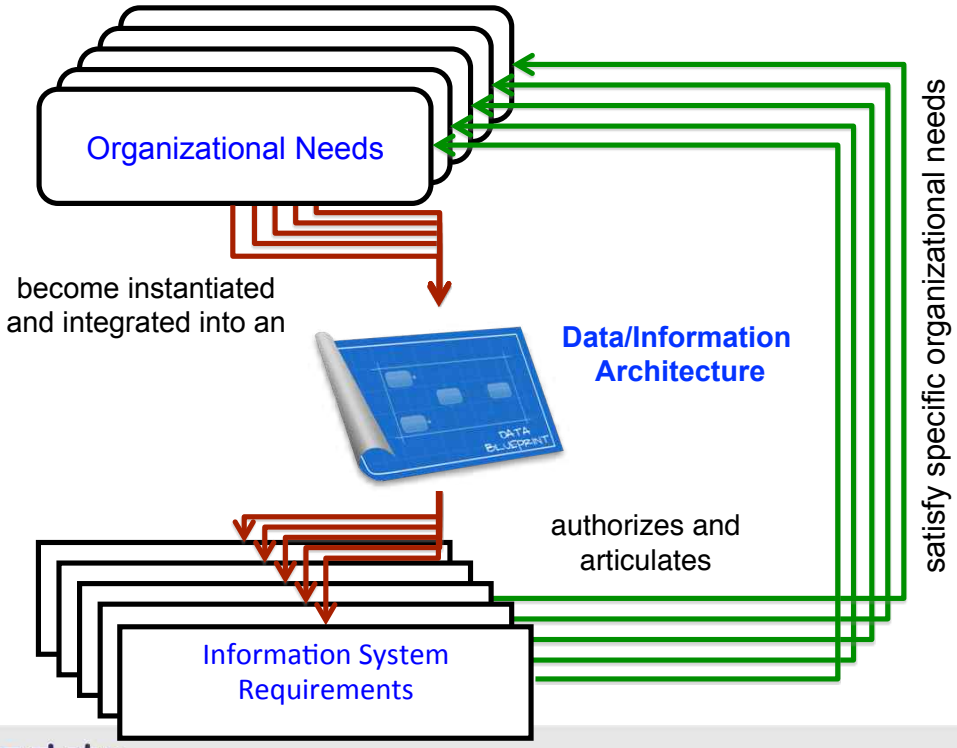
Governance is the major means of preventing over reliance on one legged stools!

What Questions Can Architectures Address?

- How and why do the components interact?
- Where do they go?
- When are they needed?
- Why and how will the changes be implemented?
- What should be managed organization-wide and what should be managed locally?
- What standards should be adopted?
- What vendors should be chosen?
- What rules should govern the decisions?
- What policies should guide the process?



Data Architectures produce and are made up of information models that are developed in response to organizational needs



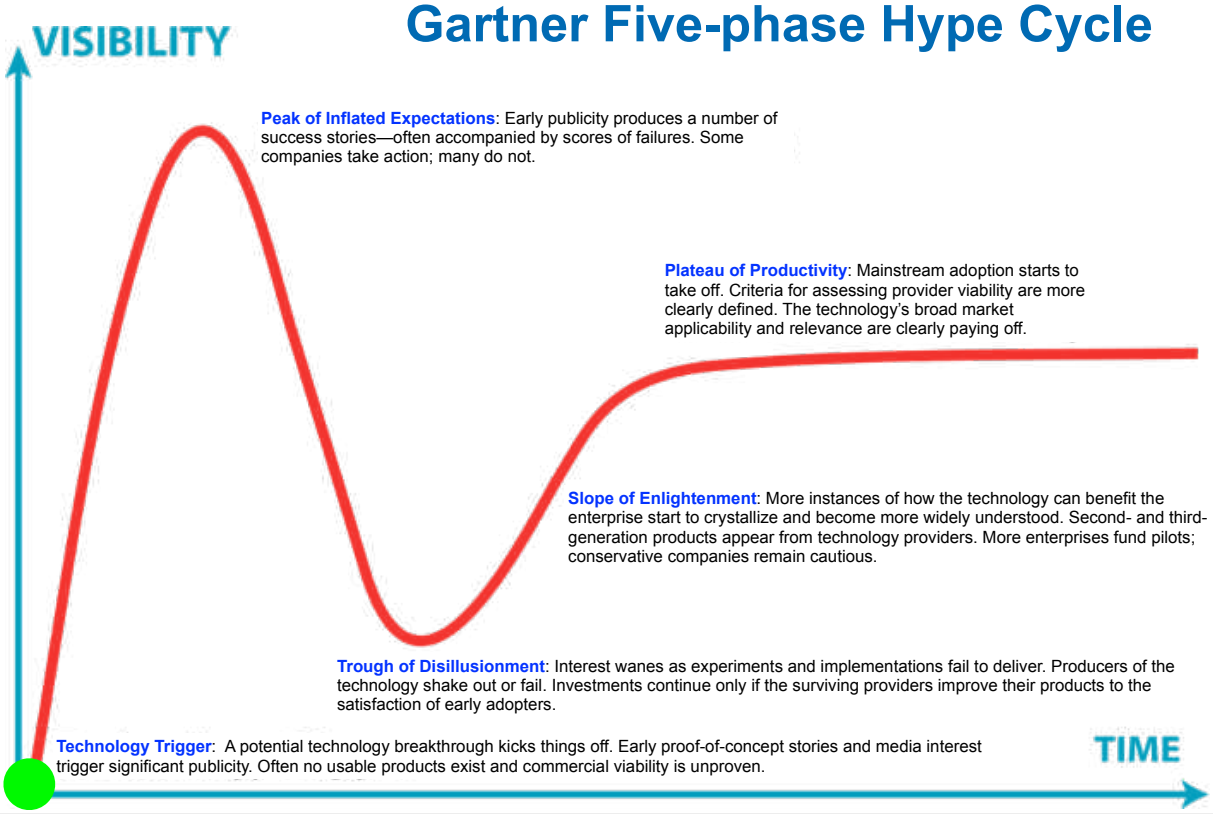
Enterprise Data Spend

- Worldwide spending on business information is now $1.1 trillion/year
- Enterprises spend an average of $38 million on information/year
- Small and Medium sized Businesses on average spend $332,000
  - http://www.cio.com.au/article/429681/five_steps_how_better_manage_your_data/

Demystifying Big Data

- Data Analysis
  - Origins
- Challenges
  - Faced by virtually everyone
- Compliment
  - Existing data management practices
- Pre-requisites
  - Necessary to exploit big data techniques
- Prototyping
  - Iterative means of practicing big data techniques
- Take Aways and Q&A

Gartner Five-phase Hype Cycle



Peak of Inflated Expectations: Early publicity produces a number of success stories—often accompanied by scores of failures. Some companies take action; many do not.

Plateau of Productivity: Mainstream adoption starts to take off. Criteria for assessing provider viability are more clearly defined. The technology's broad market applicability and relevance are clearly paying off.

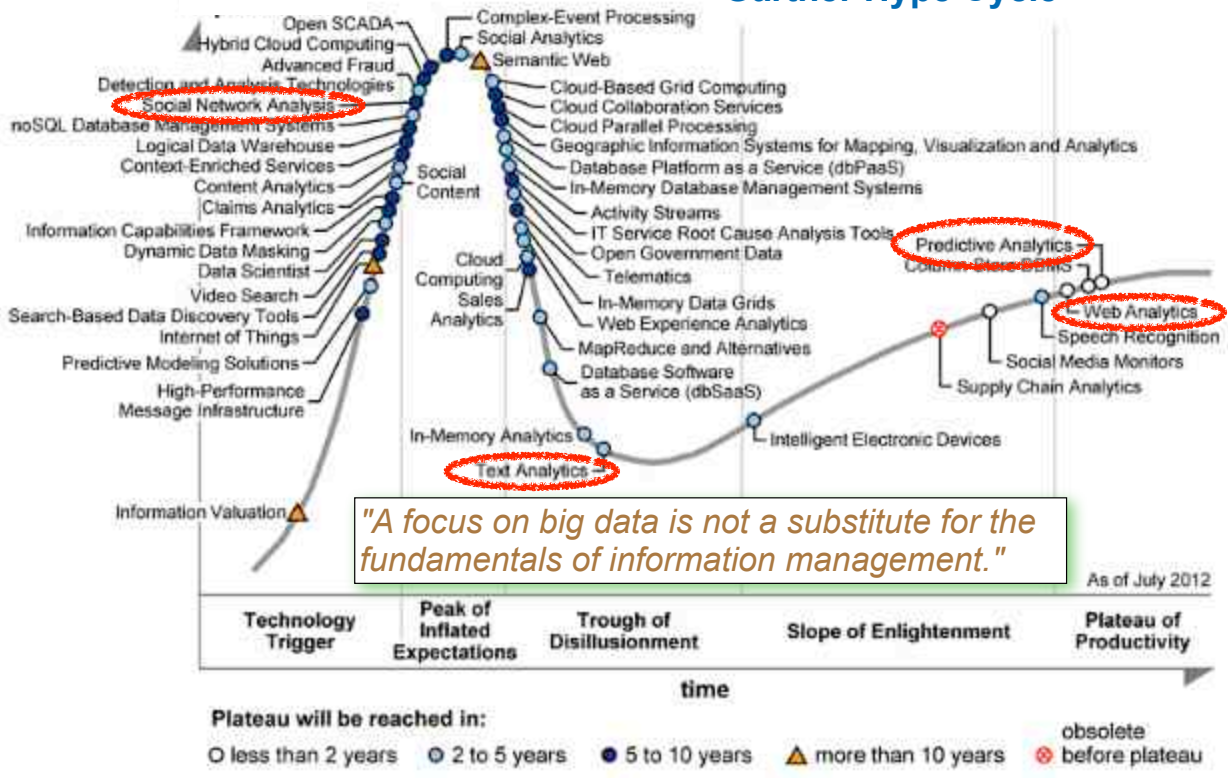
Slope of Enlightenment: More instances of how the technology can benefit the enterprise start to crystallize and become more widely understood. Second- and third-generation products appear from technology providers. More enterprises fund pilots; conservative companies remain cautious.

Trough of Disillusionment: Interest wanes as experiments and implementations fail to deliver. Producers of the technology shake out or fail. Investments continue only if the surviving providers improve their products to the satisfaction of early adopters.

Technology Trigger: A potential technology breakthrough kicks things off. Early proof-of-concept stories and media interest trigger significant publicity. Often no usable products exist and commercial viability is unproven.

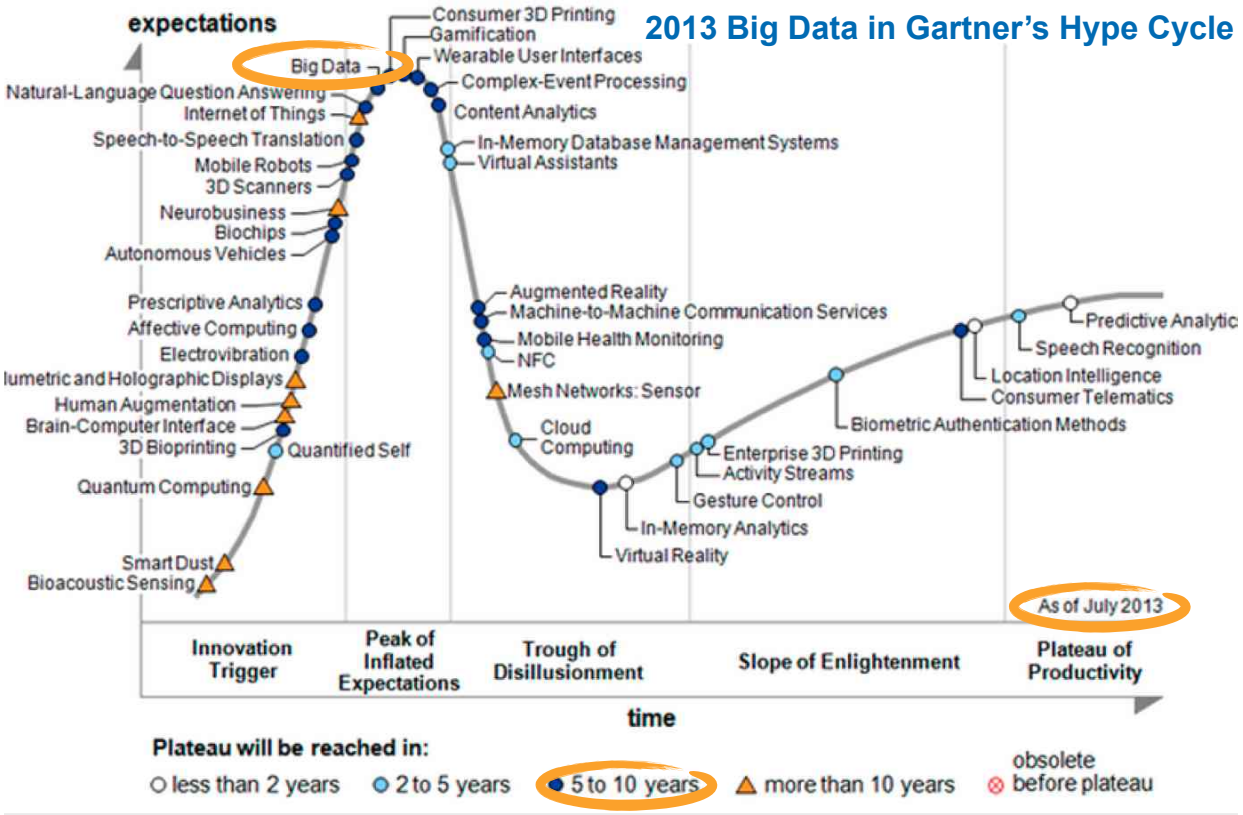
http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Gartner Hype Cycle

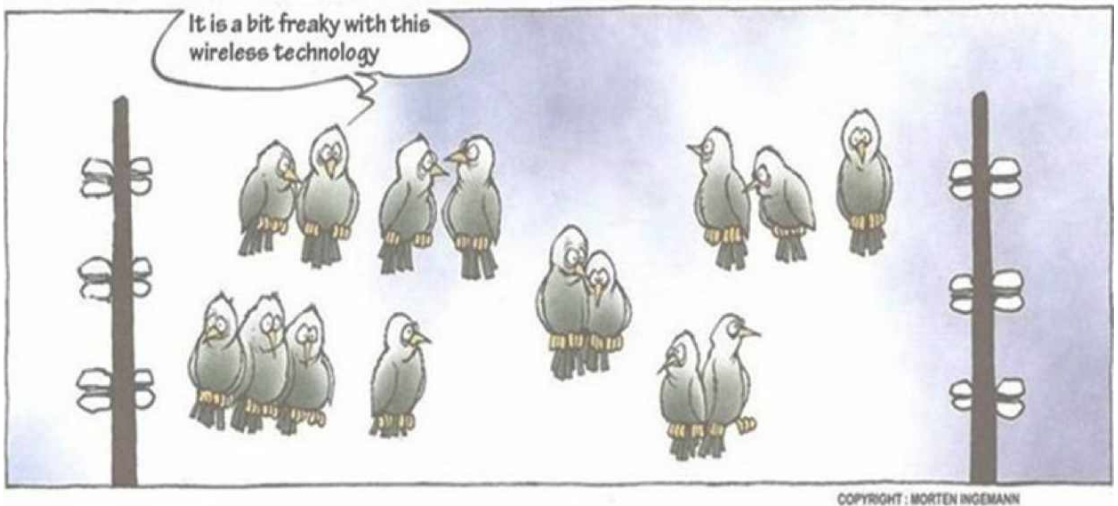


2012 Big Data in Gartner's Hype Cycle

2013 Big Data in Gartner's Hype Cycle

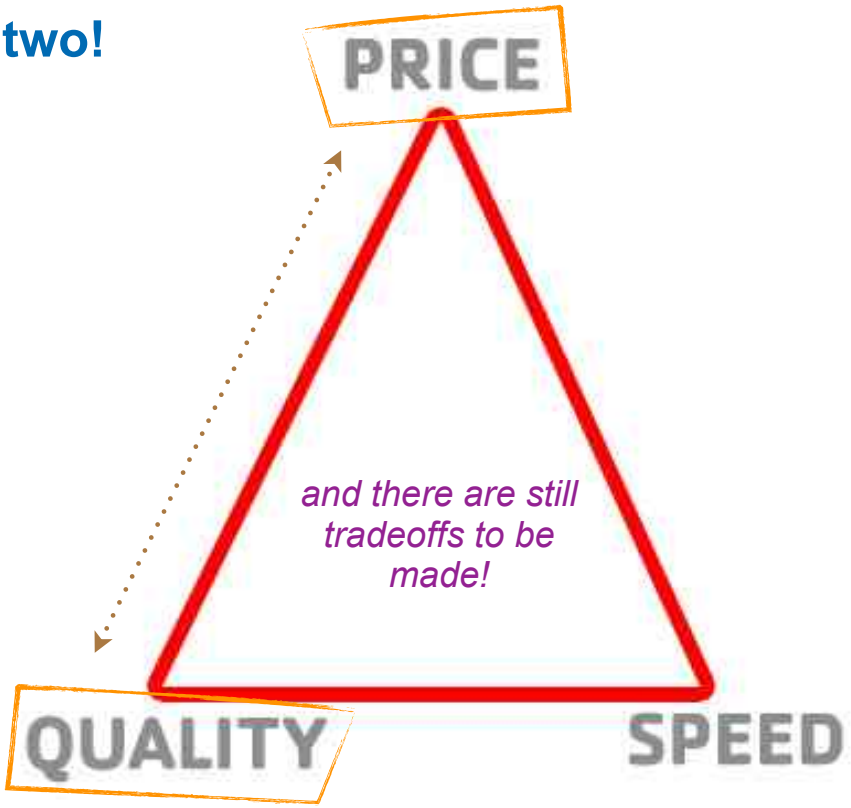


Technology Continues to Advance



- (Gordon) Moore's law
  - Over time, the number of transistors on integrated circuits doubles approximately every two years

Pick any two!



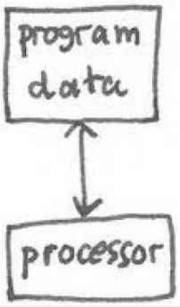
"There's now a blurring between the storage world and the memory world"

- Faster processors outstripped not only the hard disk, but main memory
  - Hard disk too slow
  - Memory too small
- Flash drives remove both bottlenecks
  - Combined Apple and Yahoo have spend more than $500 million to date
- Make it look like traditional storage or more system memory
  - Minimum 10x improvements
  - Dragonstone server is 3.2 tb flash memory (Facebook)
- Bottom line - new capabilities!

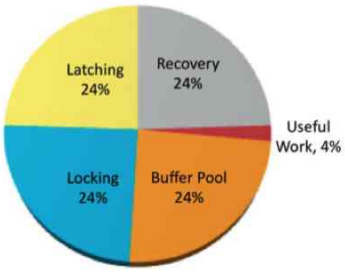
Non-von Neumann Processing/Efficiencies

- von Neumann bottleneck (computer science)
  - "An inefficiency inherent in the design of any von Neumann machine that arises from the fact that most computer time is spent in moving information between storage and the central processing unit rather than operating on it"

[http://encyclopedia2.thefreedictionary.com/von+Neumann+bottleneck]



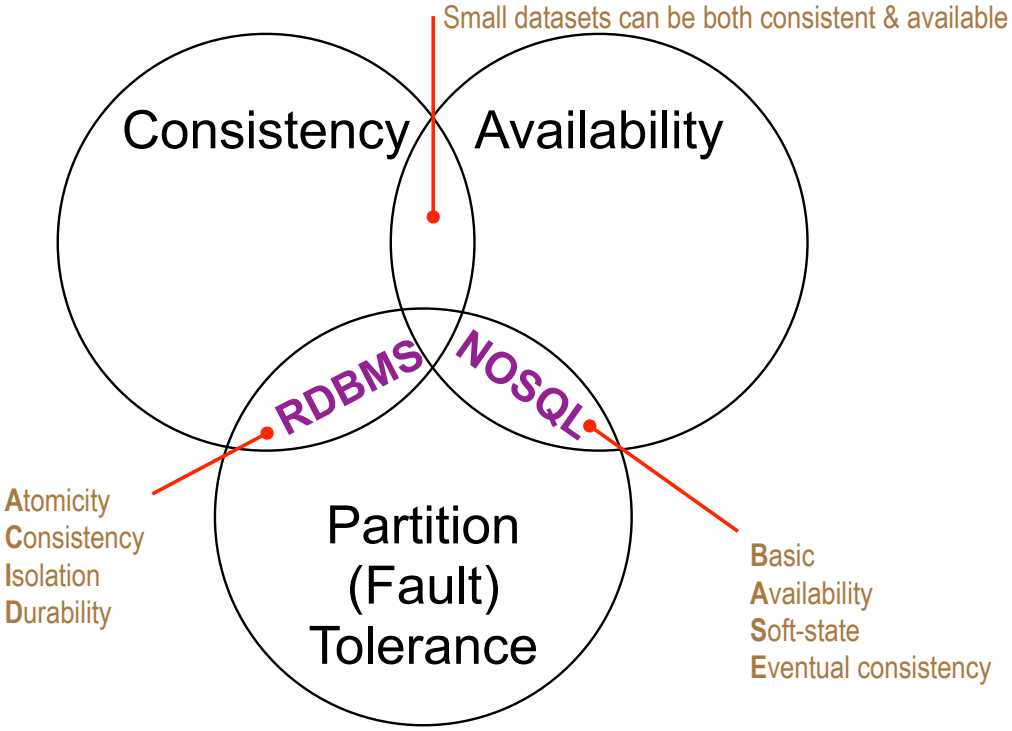
- Michael Stonebraker
  - Ingres (Berkeley/MIT)
  - Modern database processing is approximately 4% efficient



- Many "big data architectures are attempts to address this, but:
  - Zero sum game
  - Trade characteristics against each other
    - Reliability
    - Predictability
  - Google/MapReduce/ Bigtable
  - Amazon/Dynamo
  - Netflix/Chaos Monkey
  - Hadoop
  - McDipper
- Big data exploits non-von Neumann processing

Potential Tradeoffs:

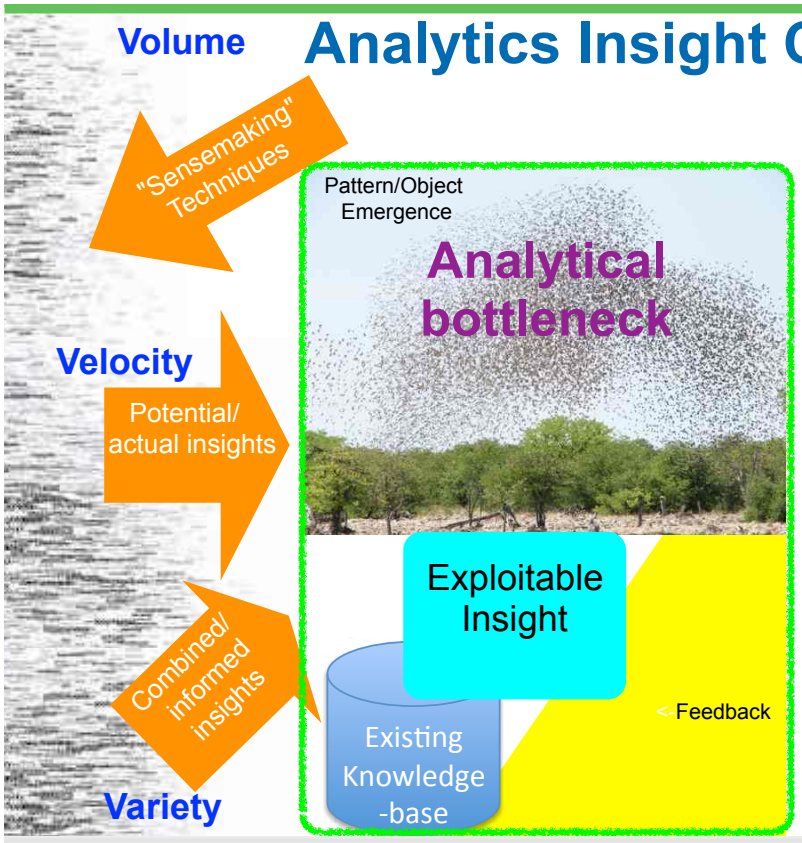
CAP theorem: consistency, availability and partition-tolerance



Potential either/or Tradeoffs

SQL	Big Data
Privacy	Big Data
Security	Big Data
?	Massive High-speed Flexible

Analytics Insight Cycle

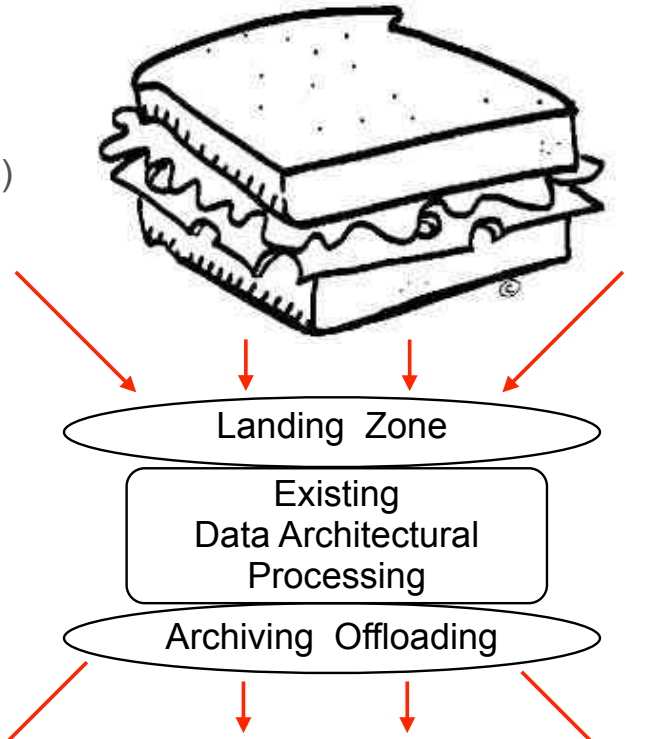


- Things are happening
  - Sensemaking techniques address "what" is happening?
- Patterns/objects, hypotheses emerge
  - What can be observed?
- Operationalizing
  - The dots can be repeatedly connected
  - "Big Data" contributions are shown in orange
- Margaret Boden's computational creativity
  - Exploratory
  - Combinational
  - Transformational

Big Data: Two prominent use cases

- Sandwich offers a good analogy of the big data and existing technologies
- Landing Zone (less expensive)
  - Especially useful in cases were data is highly disposable
- Existing technologies are the
  - Contents sandwiched and complemented landing zone and archival capabilities
- Archiving/Offloading (less need for structure)
  - "Cold" transactional and analytic data

Adapted from Nancy Kopp:
http://ibmdatamag.com/2013/08/relishing-the-big-data-burger/



Demystifying Big Data

- Data Analysis
  - Origins
- Challenges
  - Faced by virtually everyone
- Compliment
  - Existing data management practices
- Pre-requisites
  - Necessary to exploit big data techniques
- Prototyping
  - Iterative means of practicing big data techniques
- Take Aways and Q&A

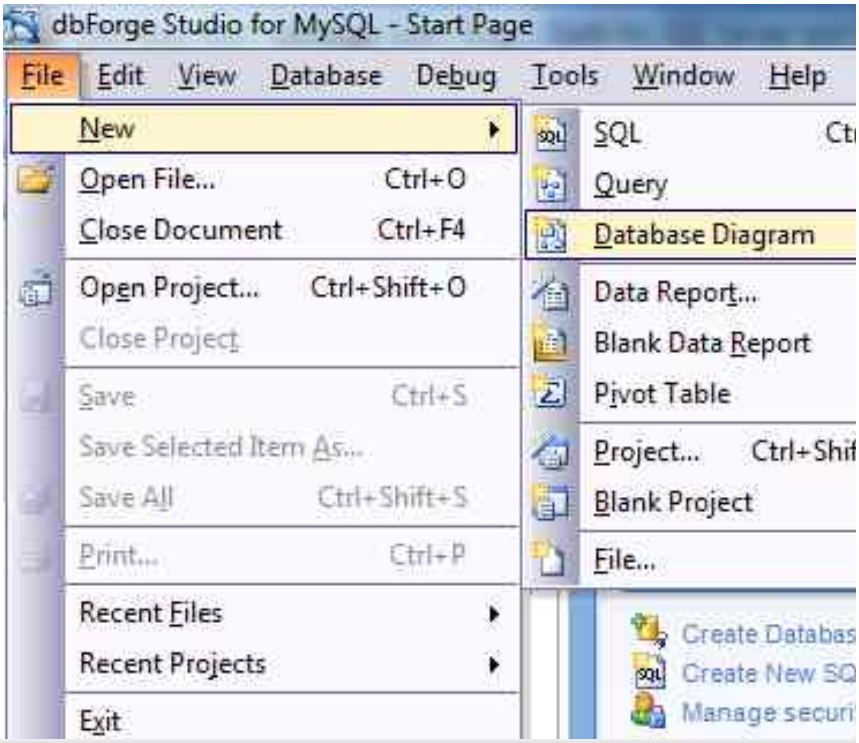
What do we teach business people about data?

What percentage of the deal with it daily?

100%

What do we teach IT professionals about data?

- 1 course
  - How to build a new database
  - 80% if IT expenses are used to improve existing IT assets
- What impressions do IT professionals get from this education?
  - Data is a technical skill that is used to develop new databases
- This is not the best way to educate IT and business professionals - every organization's
  - Sole, non-depletable, non-degrading, durable, strategic asset



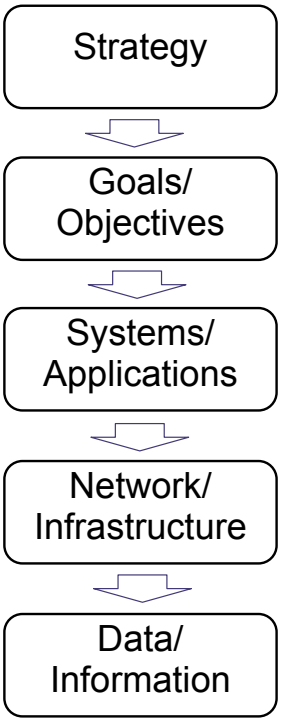
You cannot architect after implementation!

What is this?

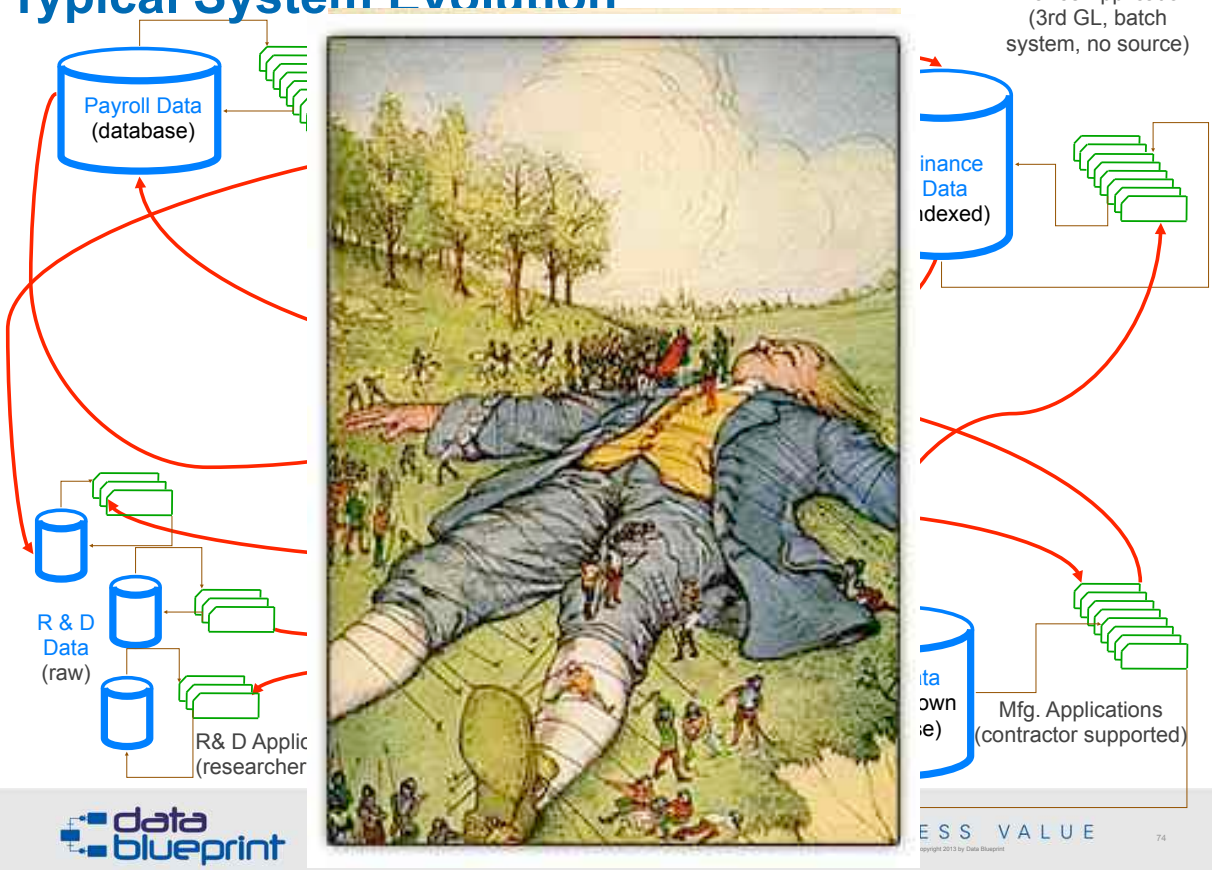
- It is tall
- It has a clutch
- It was built in 1942
- It is still in regular use!

Application-Centric Development

- In support of strategy, organizations develop specific goals/objectives
- The goals/objectives drive the development of specific systems/applications
- Development of systems/applications leads to network/infrastructure requirements
- Data/information are typically considered after the systems/applications and network/ infrastructure have been articulated
- Problems with this approach:
  - Ensures data is formed to the applications and not around the organizational-wide information requirements
  - Process are narrowly formed around applications
  - Very little data reuse is possible



Typical System Evolution



healthcare.gov

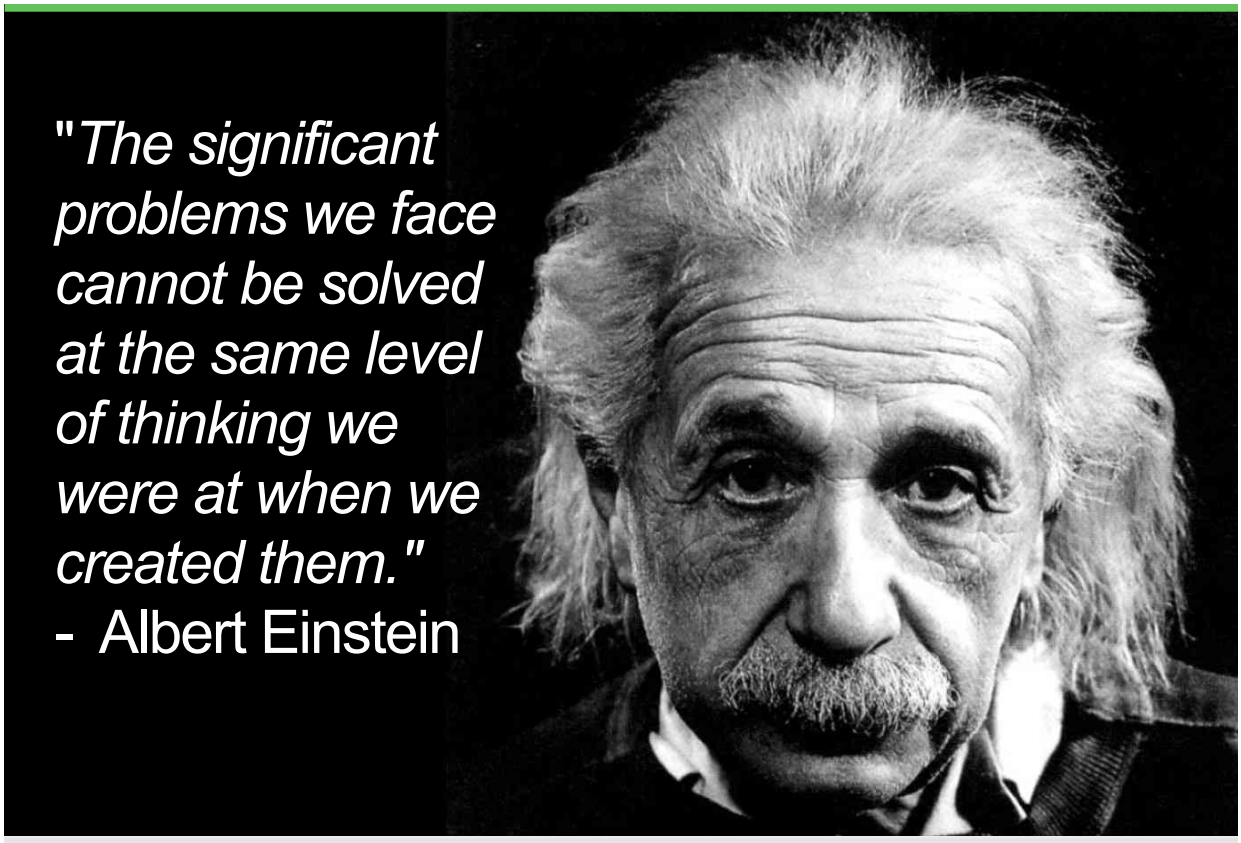
- 55 Contractors!
- "Anyone who has written a line of code or built a system from the ground-up cannot be surprised or even mildly concerned that Healthcare.gov did not work out of the gate,"

  Standish Group International Chairman Jim Johnson said in a recent podcast.

- "The real news would have been if it actually did work. The very fact that most of it did work at all is a success in itself."

- Software programmed to access data using traditional data management technologies
- Data components incorporated "big data technologies"

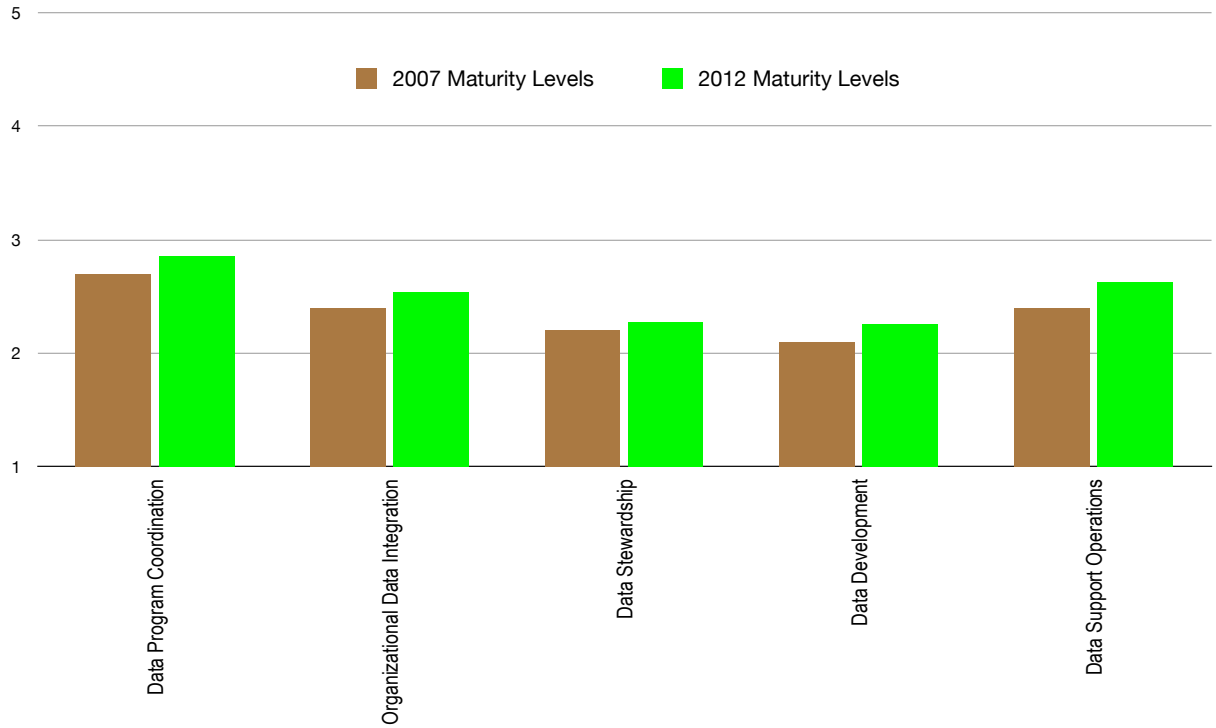
http://www.slate.com/articles/technology/bitwise/2013/10/problems_with_healthcare_gov_cronyism_bad_management_and_too_many_cooks.html



What does it mean to treat data as an organizational asset?

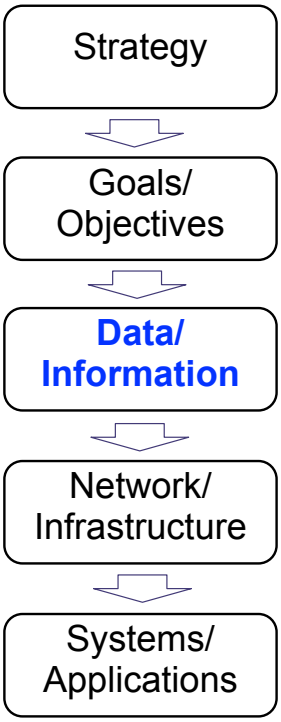
- Assets are economic resources
  - Must own or control
  - Must use to produce value
  - Value can be converted into cash
- An asset is a resource controlled by the organization as a result of past events or transactions and from which future economic benefits are expected to flow to the organization [Wikipedia]
- With assets:
  - Formalize the care and feeding of data
    - Cash management - HR planning
  - Put data to work in unique/significant ways
    - Identify data the organization will need [Redman 2008]

Comparison of DM Maturity 2007-2012



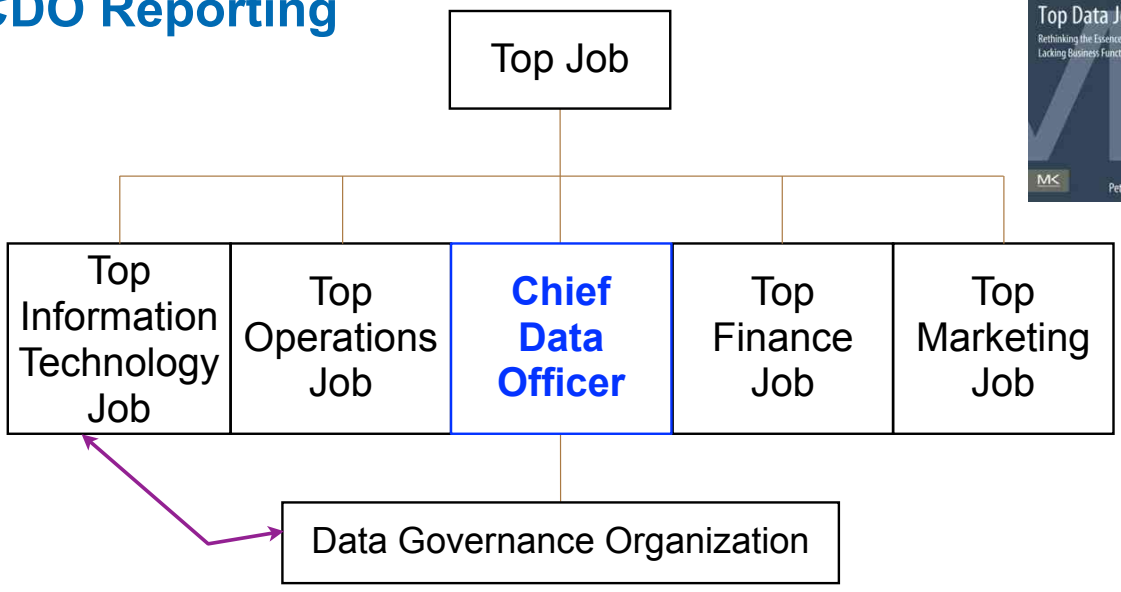
Data-Centric Development

- In support of strategy, the organization develops specific goals/objectives
- The goals/objectives drive the development of specific data/information assets with an eye to organization-wide usage
- Network/infrastructure components are developed supporting organizational data use
- Development of systems/applications is derived from the data/network architecture
- Advantages of this approach:
  - Data/information assets are developed from an organization-wide perspective
  - Systems support organizational data needs and compliment organizational process flows
  - Maximum data/information reuse



Original articulation from Doug Bagley @ Walmart

CDO Reporting



- There is enough work to justify the function
- There is not much talent
- The CDO provides significant input to the Top Information Technology Job

CDO Reporting Particulars

1. Report outside of IT and the current CIO altogether;

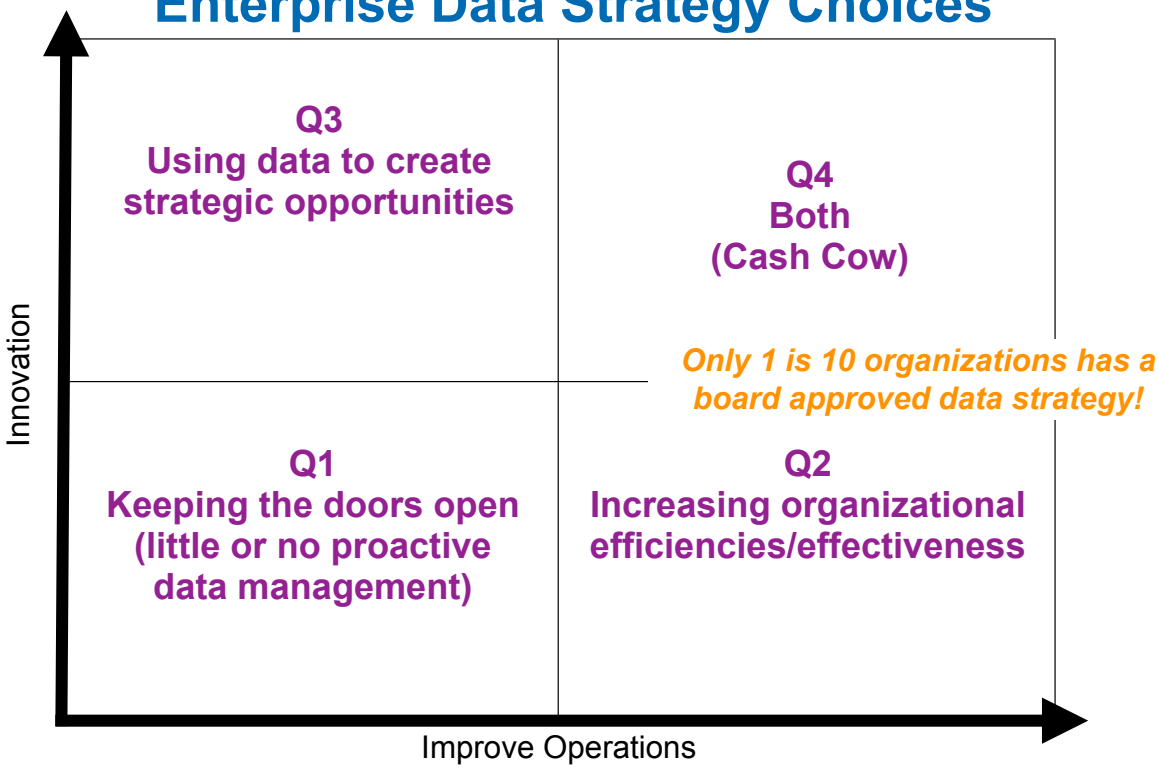
2. Report to the same organizational structure that the CFO and other "top" jobs report into; and

3. Focus on activities that are outside of (and more importantly) upstream from any system development lifecycle activities (SDLC).

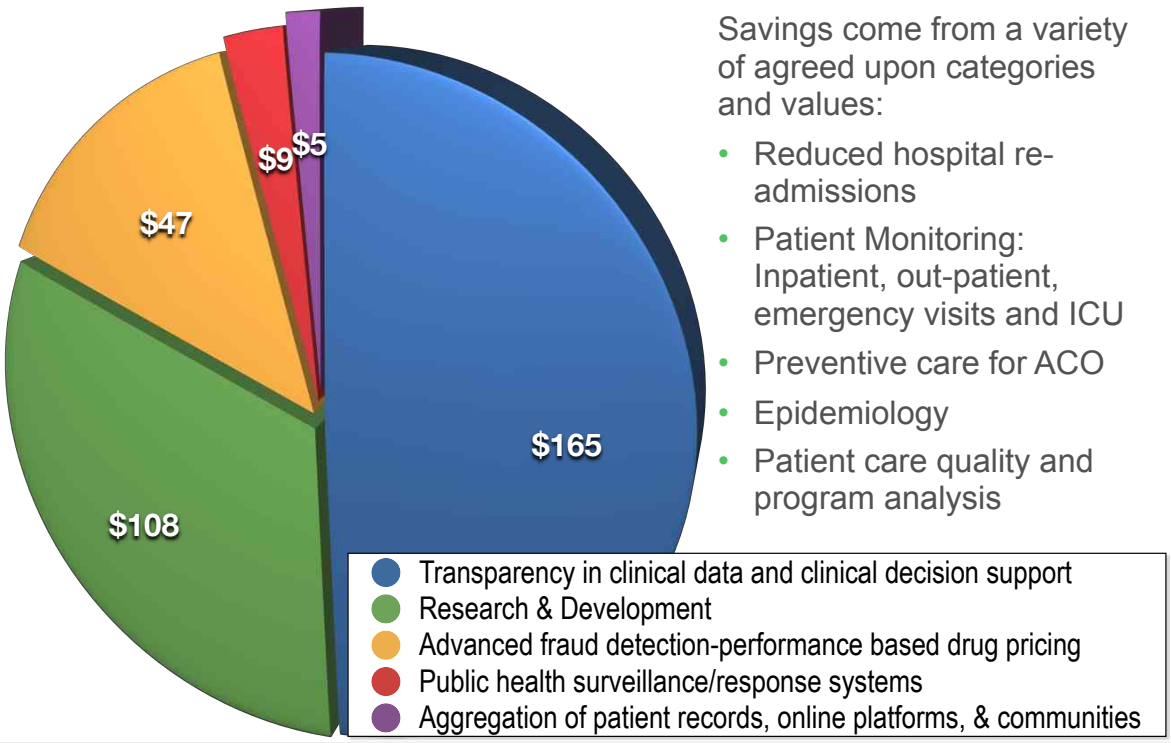
Data-Centric Perspective

- Measure success differently
- Same project
- Same process
- Different measures for success
  - Asking if our data is correct;
  - Valuing data more than we value "on time and within budget";
  - Valuing correct data more than correct processes; and
  - Auditing data rather than project documents.
    - Articulation by Linda Bevolo

Enterprise Data Strategy Choices



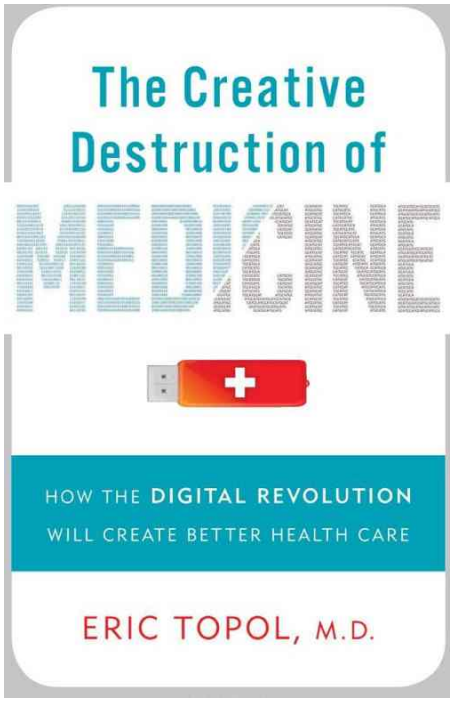
$300 billion is the potential annual value to health care



Savings come from a variety of agreed upon categories and values:

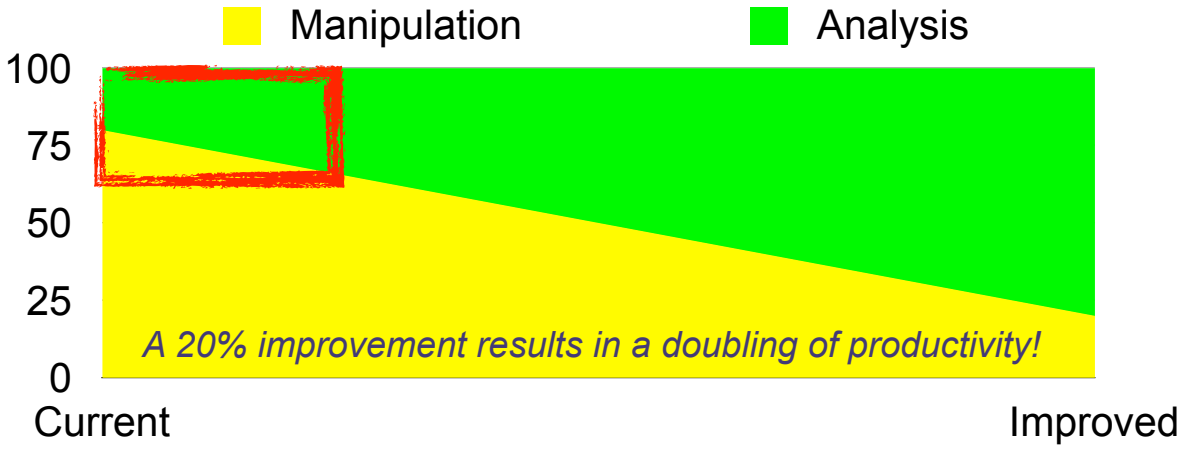
- Reduced hospital re-admissions
- Patient Monitoring: Inpatient, out-patient, emergency visits and ICU
- Preventive care for ACO
- Epidemiology
- Patient care quality and program analysis

Book Recommendation



- Permits the reorientation of medicine
  - From populations
  - To individuals
- Big Data Capture
  - Wireless sensors
  - Genome sequencing
  - Printing organs

Reversing The Measures

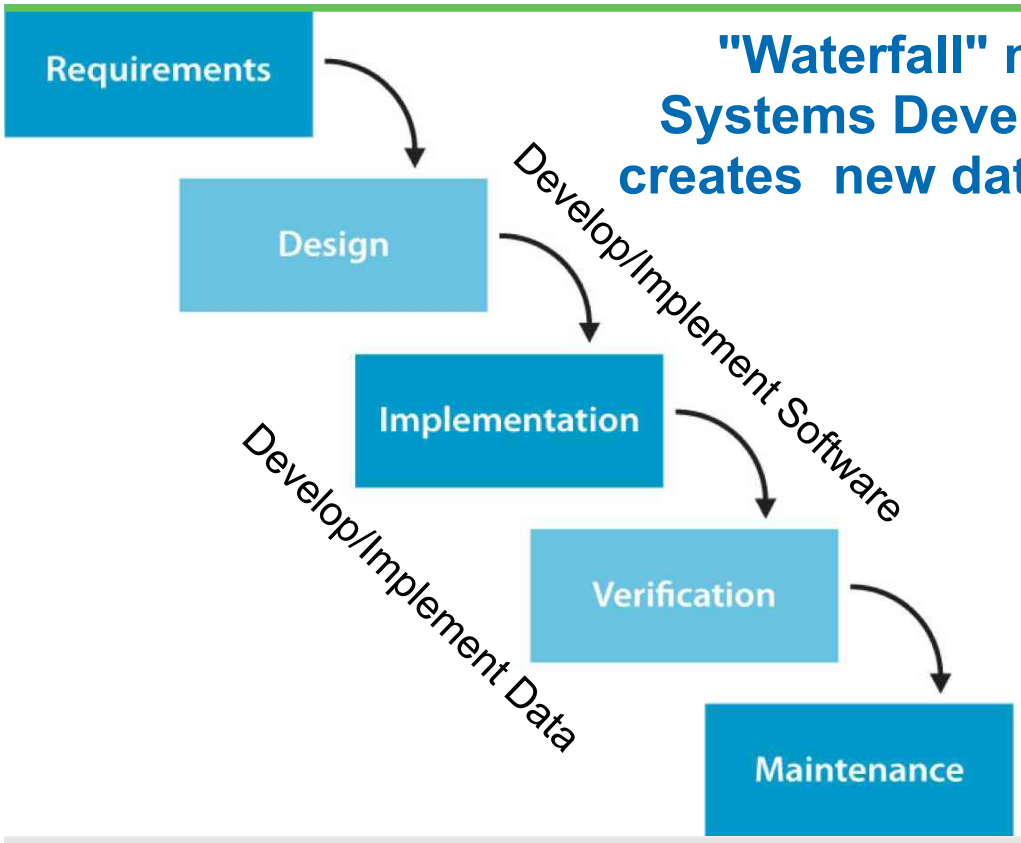


- Currently:
  - Analysts spend 80% of their time manipulating data and 20% of their time analyzing data
  - Hidden productivity bottlenecks
- After rearchitecting:
  - Analysts spend less time manipulating data and more of their time analyzing data
  - Significant improvements in knowledge worker productivity

Demystifying Big Data

- Data Analysis
  - Origins
- Challenges
  - Faced by virtually everyone
- Compliment
  - Existing data management practices
- Pre-requisites
  - Necessary to exploit big data techniques
- Prototyping
  - Iterative means of practicing big data techniques
- Take Aways and Q&A

"Waterfall" model of Systems Development creates new data siloes



Evolving Data is Different than Creating New Systems

Common Organizational Data
(and corresponding data needs requirements)

Future State

(Version +1)

Evolve

Data evolution is separate from, external to, and precedes system development life cycle activities!

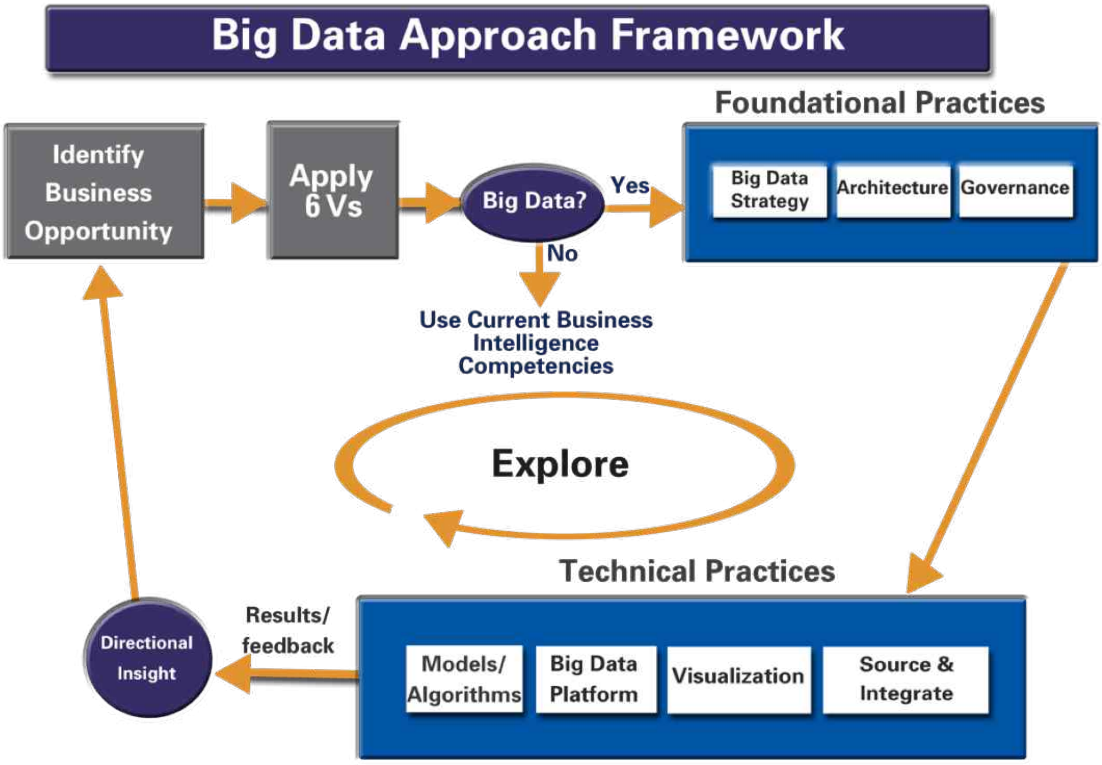
Systems
Development
Activities

Create

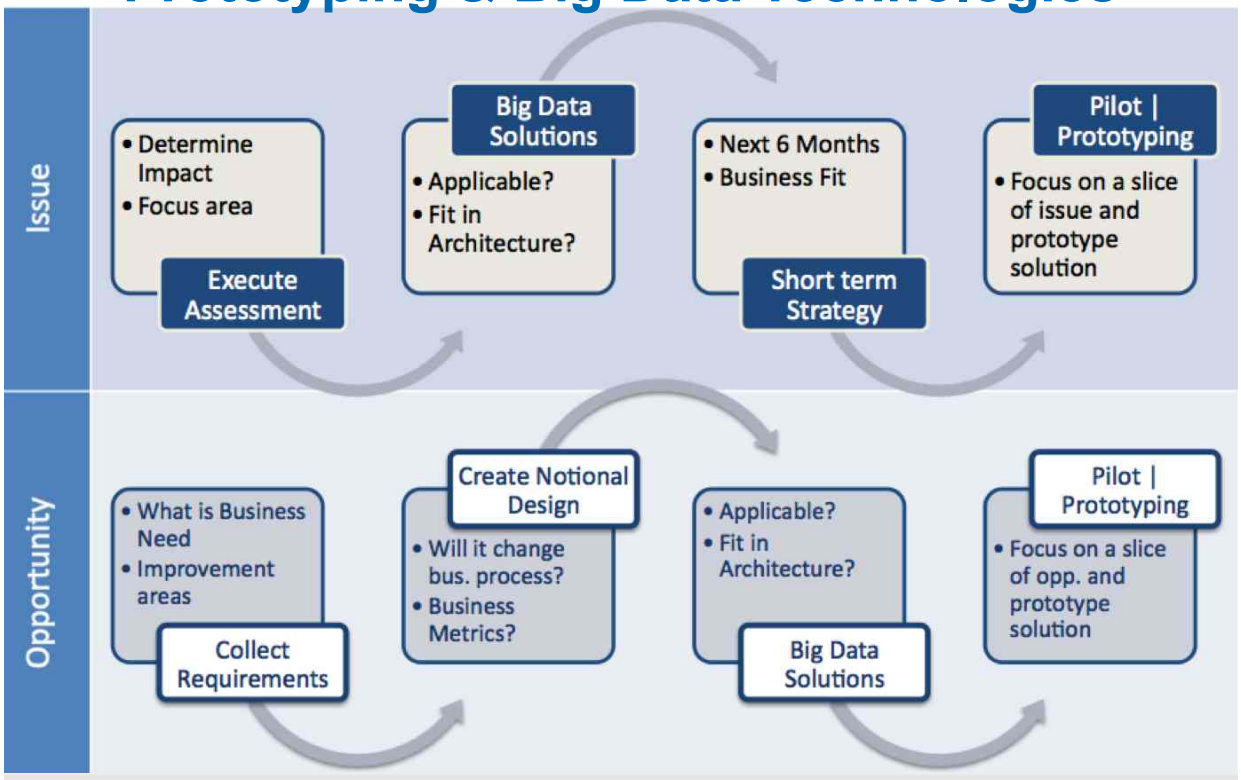
New Organizational
Capabilities

"Why IT Fumbles"

Traditional IT Project	Analytics or Big Data Project
Typical Projects	
Install an ERP system Automate a claims-handling system process Optimize supply chain performance	Develop a new, shared understanding of customers' needs and behaviors Predict future growth markets
Typical Overarching Goals	
Improve efficiency Lower costs Increase productivity	Change thinking about data Challenge the assumptions and biases Use new insights to serve customers better, build new businesses, and predict outcomes
Project Structure	
Traditional Project Management: Define desired outcomes Redesign work processes Specify technology needs Develop detailed plans to deploy IT, manage organizational change, and train users Implement plans	Discovery-driven: Develop theories Build hypotheses Identify relevant data Conduct experiments Refine hypotheses in response to findings Repeat process
Competencies Required	
IT professionals with engineering, computer science, and math backgrounds People who know the business	In addition: Data scientists Cognitive and behavioral scientists
What does success look like?	
Project come in on time, to plan, and within budget Project achieves the desired process change	Base decisions on data and evidence Use data to generate new insights in new contexts



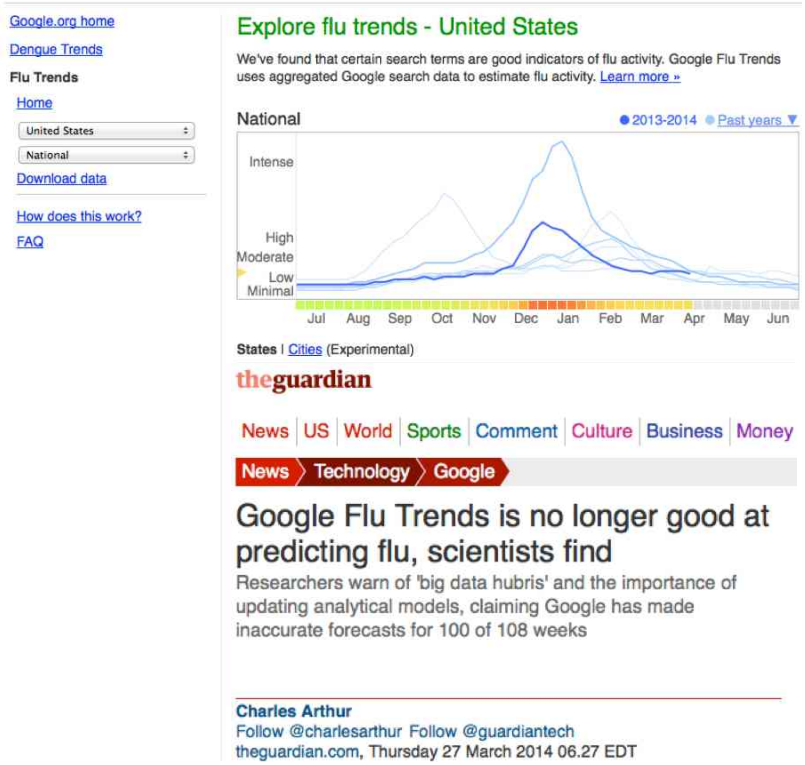
Prototyping & Big Data Technologies



Demystifying Big Data

- Data Analysis
  - Origins
- Challenges
  - Faced by virtually everyone
- Compliment
  - Existing data management practices
- Pre-requisites
  - Necessary to exploit big data techniques
- Prototyping
  - Iterative means of practicing big data techniques
- Take Aways and Q&A

Four Articles of Big Data Faith



- Cheerleaders for big data have made four exciting claims, each one reflected in the success of Google Flu Trends:
  1. that data analysis produces uncannily accurate results;
  2. that every single data point can be captured, making old statistical sampling techniques obsolete;
  3. that it is passé to fret about what causes what, because statistical correlation tells us what we need to know; and
  4. that scientific or statistical models aren't needed because, to quote "The End of Theory", a provocative essay published in Wired in 2008, "with enough data, the numbers speak for themselves"

http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html

The New York Times

The Opinion Pages

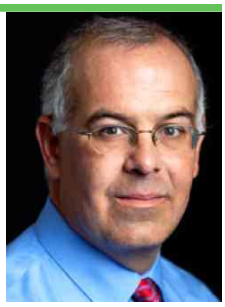
WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY

OP-ED COLUMNIST

What Data Can't Do

By DAVID BROOKS

Published: February 18, 2013 | 254 Comments

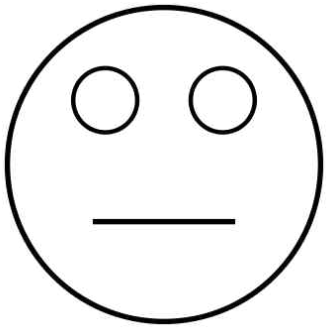


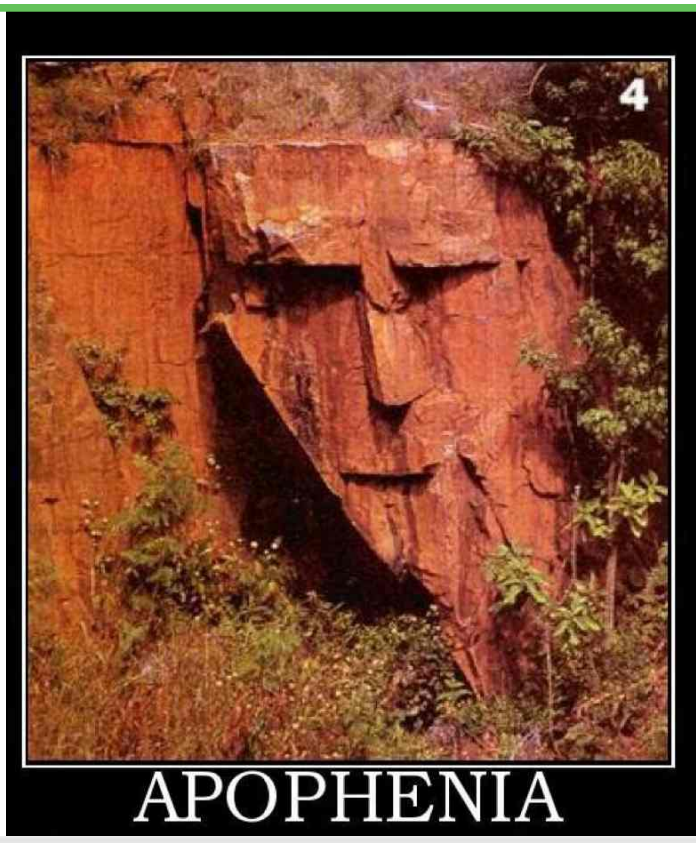
Some Big Data Limitations

- Data analysis struggles with the social
  - Your brain is excellent at social cognition - people can
    - Mirror each other's emotional states
    - Detect uncooperative behavior
    - Assign value to things through emotion
  - Data analysis measures the quantity of social interactions but not the quality
    - Map interactions with co-workers you see during work days
    - Can't capture devotion to childhood friends seen annually
  - When making (personal) decisions about social relationships, it's foolish to swap the amazing machine in your skull for the crude machine on your desk
- Data struggles with context
  - Decisions are embedded in sequences and contexts
  - Brains think in stories - weaving together multiple causes and multiple contexts
  - Data analysis is pretty bad at
    - Narratives / Emergent thinking / Explaining
- Data creates bigger haystacks
  - More data leads to more statistically significant correlations
  - Most are spurious and deceive us
  - Falsity grows exponentially greater amounts of data we collect
- Big data has trouble with big problems
  - For example: the economic stimulus debate
  - No one has been persuaded by data to switch sides
- Data favors memes over masterpieces
  - Detect when large numbers of people take an instant liking to some cultural product
  - Products are hated initially because they are unfamiliar
- Data obscures values
  - Data is never raw; it's always structured according to somebody's predispositions and values

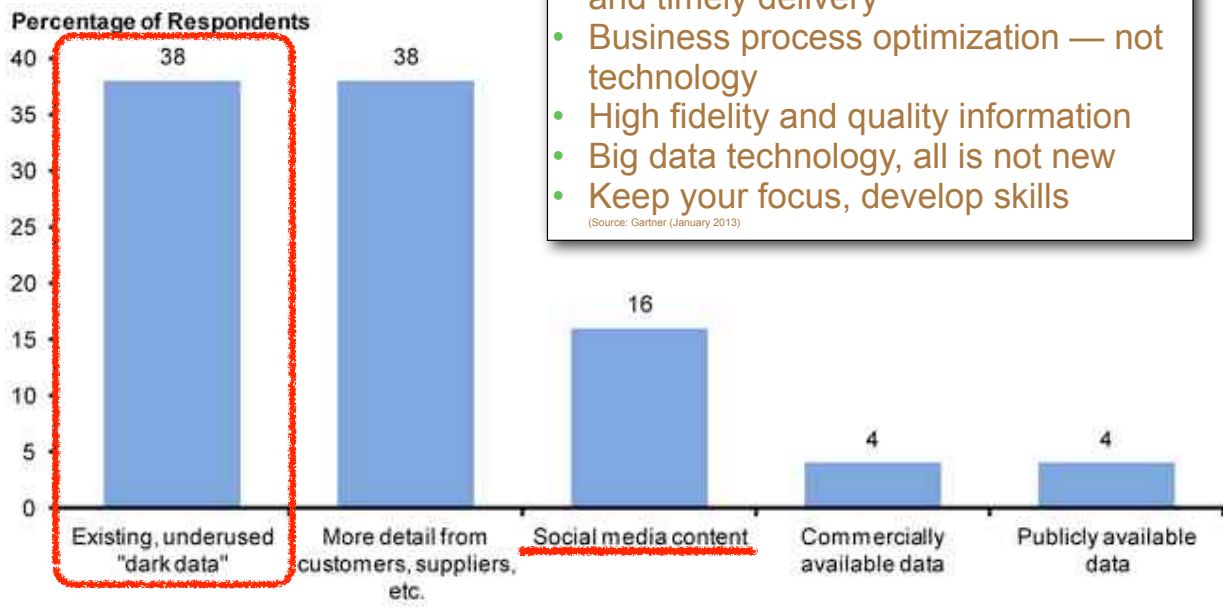
Apophenia

- Spontaneous perception of connections and meaningfulness of unrelated phenomena
  [http://skepdic.com/apophenia.html]
- *Nothing is so alien to the human mind as the idea of randomness*
  [John Cohen]





Which Source of Data Represents the Most Immediate Opportunity?



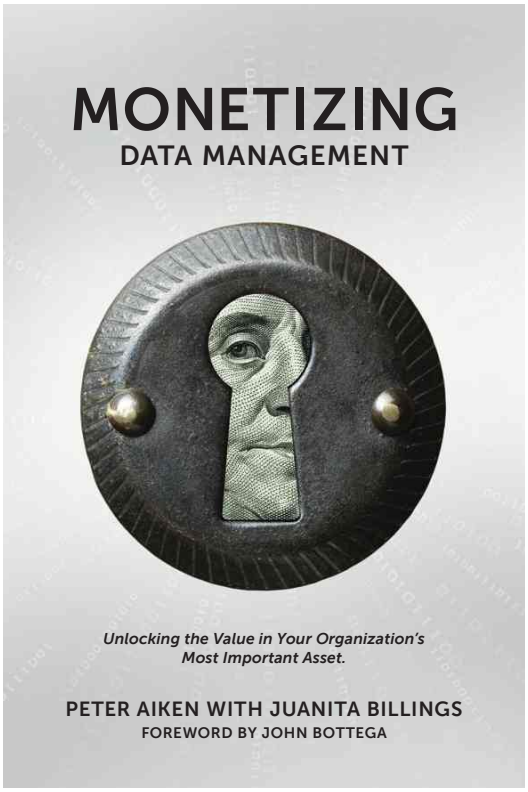
Guidance

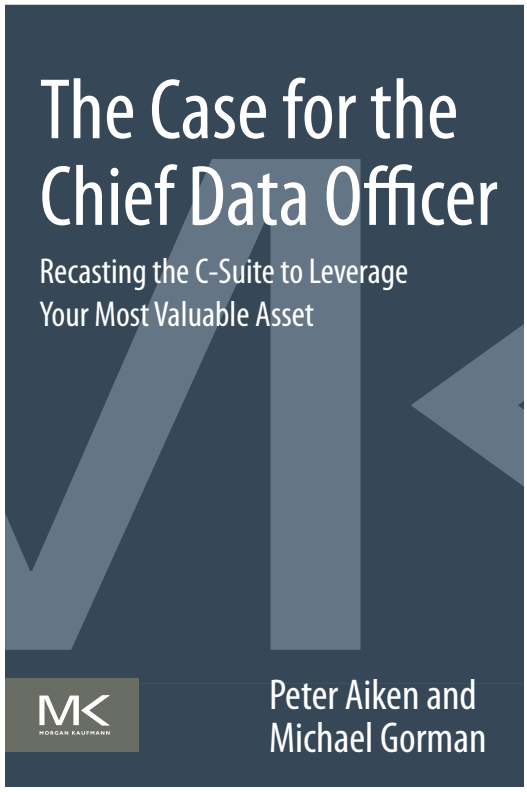
- The real change: cost-effectiveness and timely delivery
- Business process optimization — not technology
- High fidelity and quality information
- Big data technology, all is not new
- Keep your focus, develop skills

(Source: Gartner (January 2013)

Gartner Recommendations

Impacts	Top
Some of the new analytics that are made possible by big data have no precedence, so innovative thinking will be required to achieve value	Treat big data projects as innovation projects that will require change management efforts. The business will take time to trust new data sources and new analytics
Creative thinking can unearth valuable information sources already inside the enterprise that are underused	Work with the business to conduct an inventory of internal data sources outside of IT's direct control, and consider augmenting existing data that is IT 'controlled.' With an innovation mindset, explore the potential insight that can be gained from each of these sources
Big data technologies often create the ability to analyze faster, but getting value from faster analytics requires business changes	Ensure that big data projects that improve analytical speed always include a process redesign effort that aims at getting maximum benefit from that speed





data blueprint

UNLOCKING BUSINESS VALUE

datablueprint.com

10124 W. Broad Street, Suite C
Glen Allen, Virginia 23060
804.521.4056