



# Data: How are you going to secure it if you can't manage it?

... in the Post Big Data Era

Peter Aiken PhD, Senior Member - ACM

[datablueprint.com](http://datablueprint.com)

## Peter Aiken, Ph.D.

- 30+ years in data management
- Repeated international recognition
- Founder, Data Blueprint ([datablueprint.com](http://datablueprint.com))
- Associate Professor of IS ([vcu.edu](http://vcu.edu))
- DAMA International ([dama.org](http://dama.org))
- 9 books and dozens of articles
- Experienced w/ 500+ data management practices
- Multi-year immersions:
  - US DoD (DISA/Army/Marines/DLA)
  - Nokia
  - Deutsche Bank
  - Wells Fargo
  - Walmart
  - ...



UNLOCKING BUSINESS VALUE

# When I Grow up...



Dear Ms. Davis:

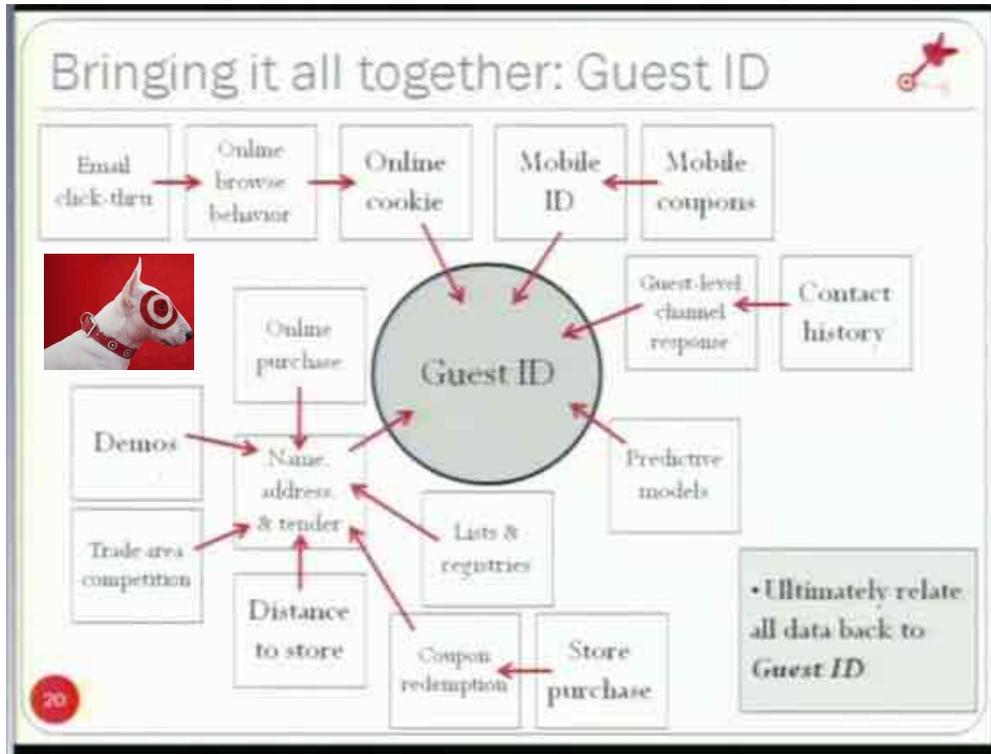
I want to be very clear on my child's illustration. It is NOT of me on a dance pole on a stage in a strip joint. I work at Target and had commented to my daughter how much money we made in the recent snowstorm. This drawing is of me selling a shovel.

I want to be Mrs. Harrington like mommy!



• Surrender To A Buyer Power

# Target Isn't Just Predicting Pregnancies



<http://mportal.performedia.com/node/1373> and <http://www.predictiveanalyticsworld.com/batimes/target-really-predict-teens-pregnancy-inside-story/> [http://mportal.performedia.com/m/paw10/gallery\\_01#1373](http://mportal.performedia.com/m/paw10/gallery_01#1373)



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 5

## Managing Data with Guidance?



**ASHLEY MADISON**<sup>®</sup>  
Life is Short. Have an Affair.<sup>®</sup>



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 6



data  
blueprint

UNLOCKING BUSINESS VALUE

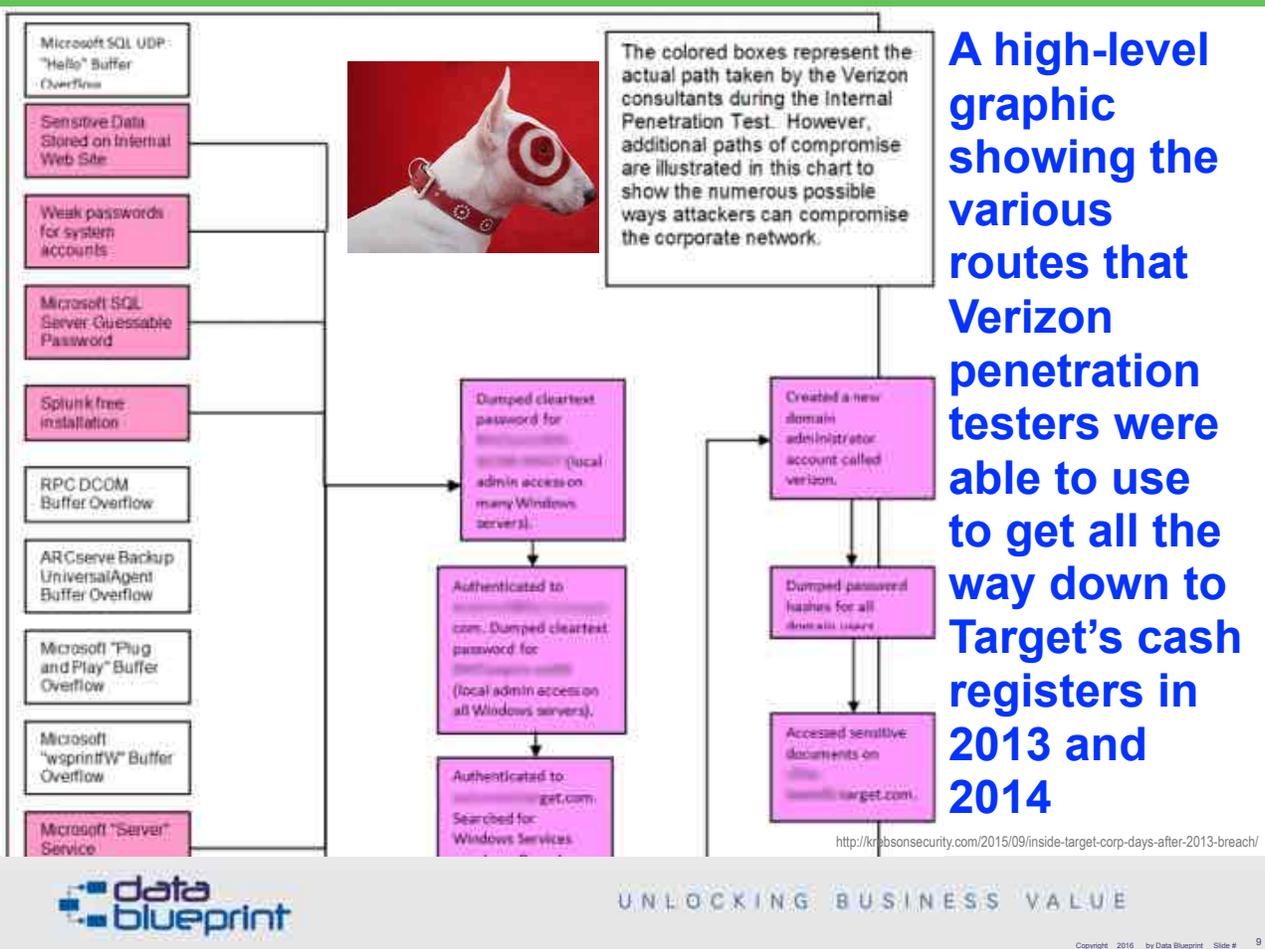
Copyright 2016 by Data Blueprint Slide # 7



data  
blueprint

UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 8



**THE WALL STREET JOURNAL.**  
U.S. EDITION - Wednesday, May 28, 2014 As of 1:58 PM EDT *These decisions have consequences!*

Home World U.S. Business Tech Markets Market Data Your Money Opinion Life & Culture N.Y. Real Estate

May 28, 2014, 1:58 p.m. ET

## Target Shareholders Should Oust Directors, ISS Says

Article

Email Printer Friendly Share: facebook

Text

By Paul Ziobro

A prominent proxy adviser took the unusual step of recommending that Target Corp. shareholders oust seven of the company's 10 directors, citing what it called the board's failure to manage risk and protect the retailer from a massive data breach.

Institutional Shareholder Services, which advises big shareholders like mutual funds how to vote on corporate issues, focused on directors who serve on Target's audit and corporate-responsibility committees. Those committees are tasked with overseeing and managing risk, and the data breach showed the company was inadequately prepared for the threats posed by hackers.

"It appears that failure of the committees to ensure appropriate management of these risks set the stage for the data breach, which has resulted in significant losses to the company and its shareholders," ISS wrote.

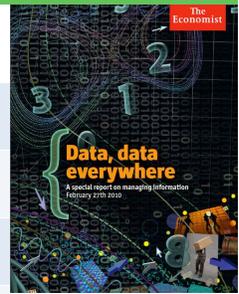
Beth Jacobs abruptly resigned in March



Target's CEO is out in wake of big security breach



# Data Inflation



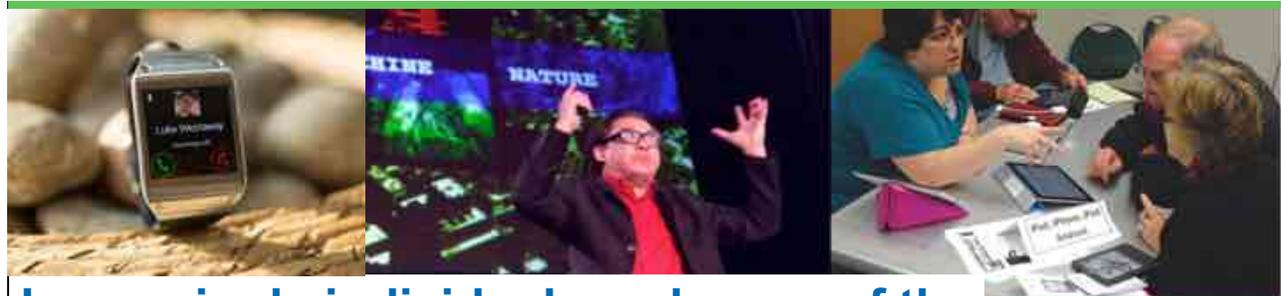
Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or $2^{10}$ , bytes	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



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 11



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



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 12



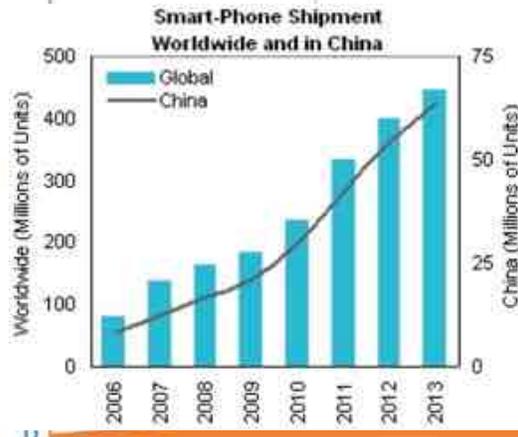
# AT&T Mobile Data Volumes Up 8,000% Over Four Years

PB/month

10

*"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)

## Take control of your data growth.



# The likely state of your data management efforts



Very Silo'ed or conflicting data sources

Multiple Data Sources

Redundancy

Multiple changes to source system

Inconsistent data definitions of common terms

ISD are data owners

Inconsistent Data Quality

Difficult to report and mine against

Lots of Data...Minimum Information



# Separating the Wheat from the Chaff



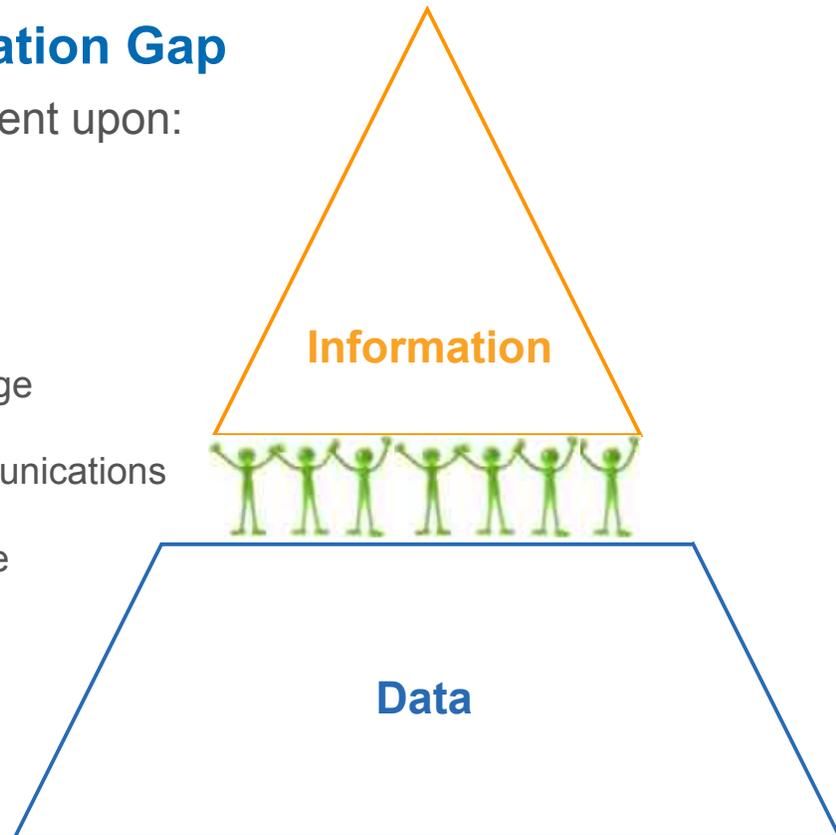
## Separating the Wheat from the Chaff

- Harvard Business School wisdom
  - 20% of your customers cause 80% of your problems
  - Eliminating those 20% (the problem customers) should increase profitability without any additional cost
  - The key question is how to identify the correct 20%
  - Very few try this strategy
- Poor data management practices are costing organizations much money/time/effort
- Data that is better organized increases in value
- Pareto analysis dictates that 80% of organizational data is ROT
  - Redundant
  - Obsolete
  - Trivial
- The question is the same - which data to eliminate?



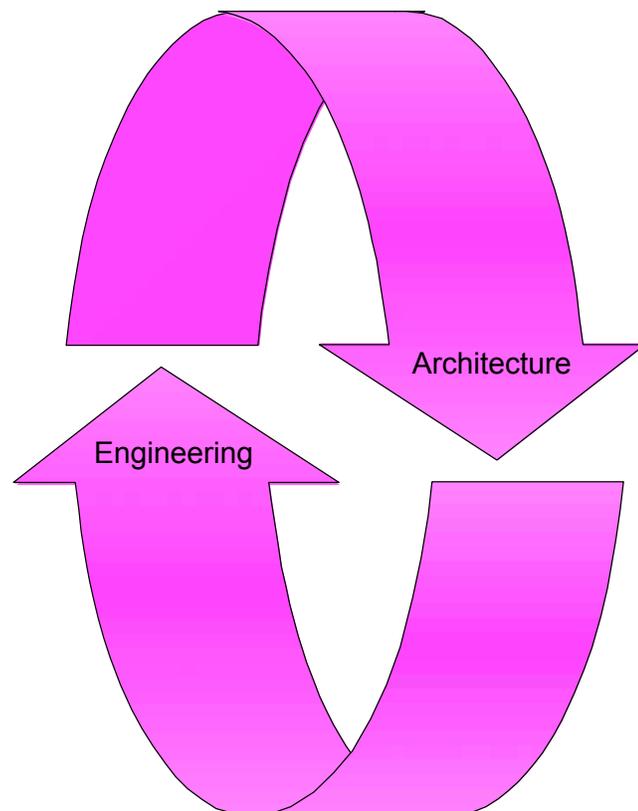
## Data / Information Gap

- Overly dependent upon:
  - Human-beings
  - Wetwear
  - Tribal knowledge
  - Informal communications
  - Non-repeatable practices



## Architecture and Engineering

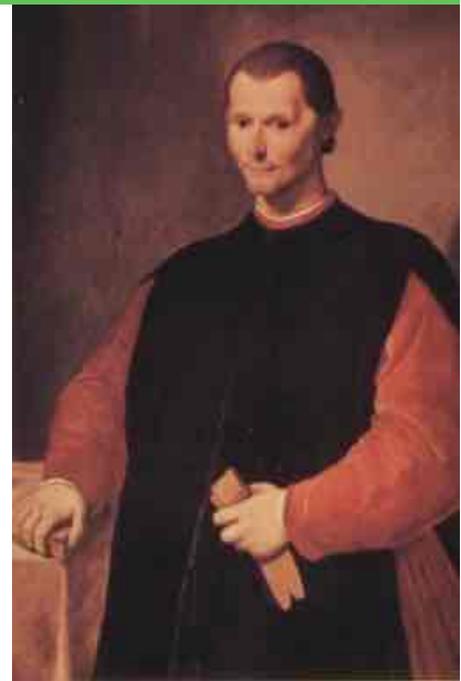
- Architecture enables complex "things" to be built
- Engineering ensures a disciplined approach to development





## Niccolo Machiavelli

(1469-1527)



*He who doesn't lay his foundations before hand, may by great abilities do so afterward ...*

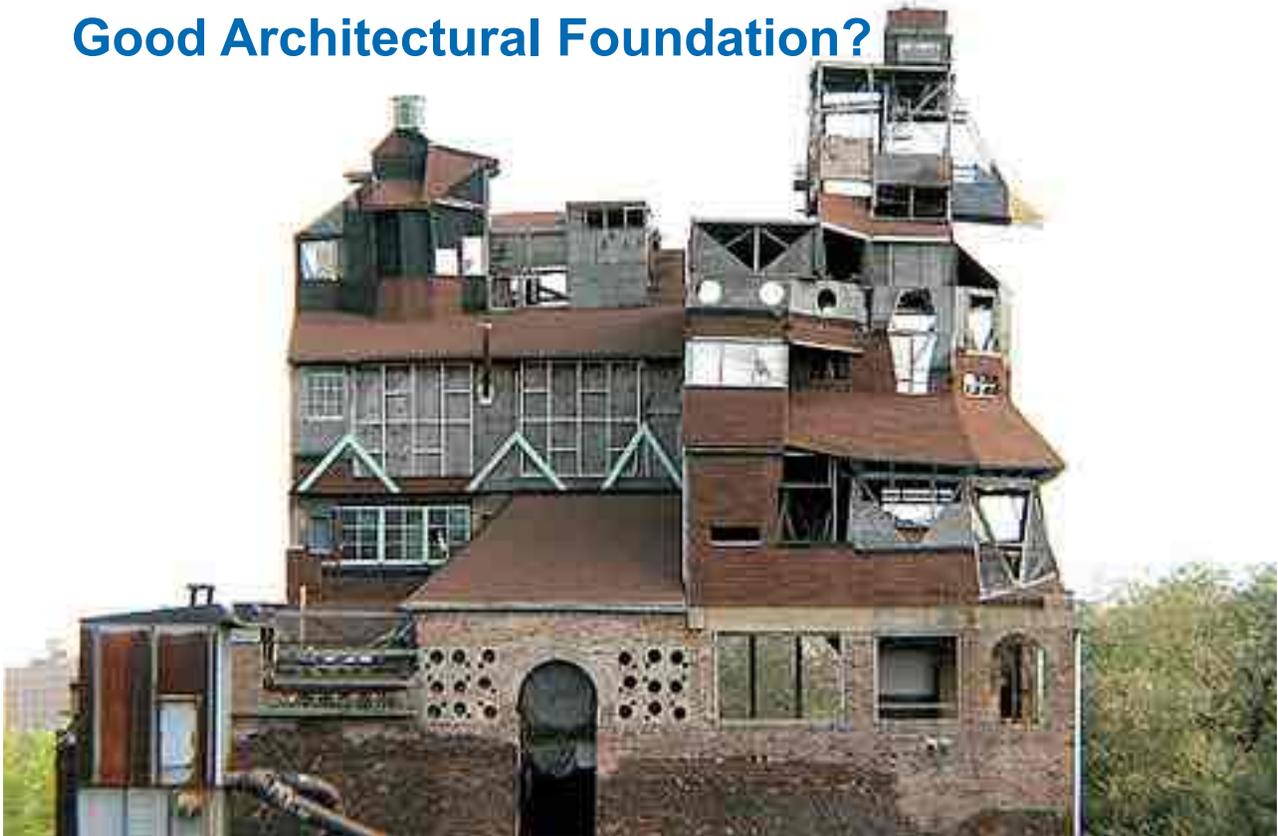
*... although with great trouble to the architect and danger to the building.*

Machiavelli, Niccolò. *The Prince*. 19 Mar. 2004 <http://pd.sparknotes.com/philosophy/prince>

***You cannot architect after implementation!***



## Good Architectural Foundation?



## Poor Architectural Foundation



# What they think they are purchasing!



# What is this?



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



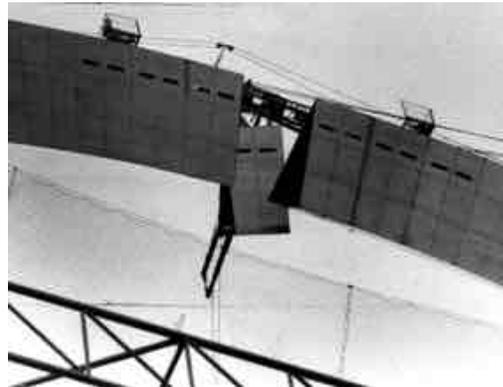


## New River Bridge

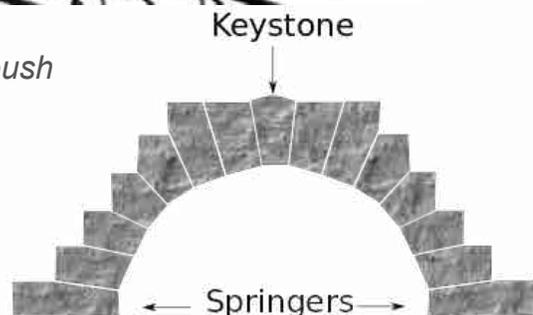


[http://www.stltoday.com/news/local/metro/a-look-back-arch-workers-had-to-stretch-push-legs/article\\_0bc89bca-0433-11e2-be00-001a4bcf6878.html](http://www.stltoday.com/news/local/metro/a-look-back-arch-workers-had-to-stretch-push-legs/article_0bc89bca-0433-11e2-be00-001a4bcf6878.html)

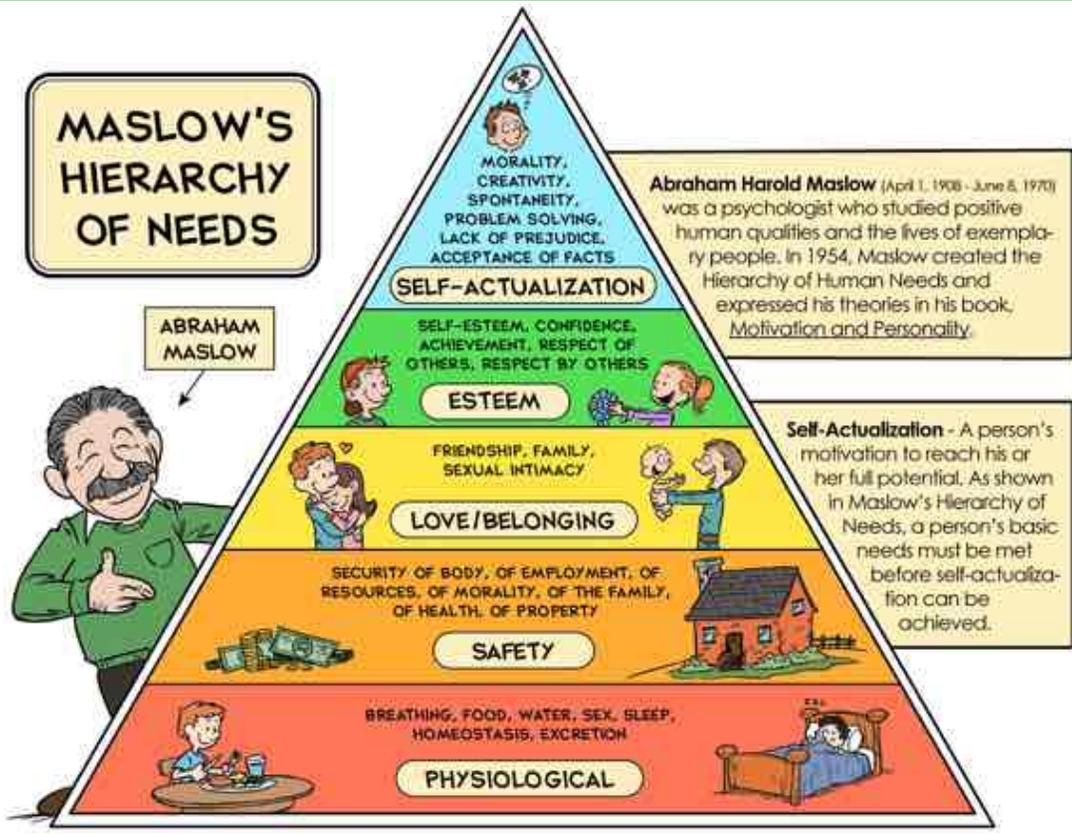
## Oct. 28, 1965 Keystoning the St. Louis Gateway Arch



- *A Look Back* • Arch workers had to stretch push legs apart to fit final piece in
- 1 week before, earthquake
- Every fire truck in the state
- Arch could swell 1' in the 40 degree temps
- Jacked legs 8.5' apart - 6" to spare
- 8' keystone provided structural integrity





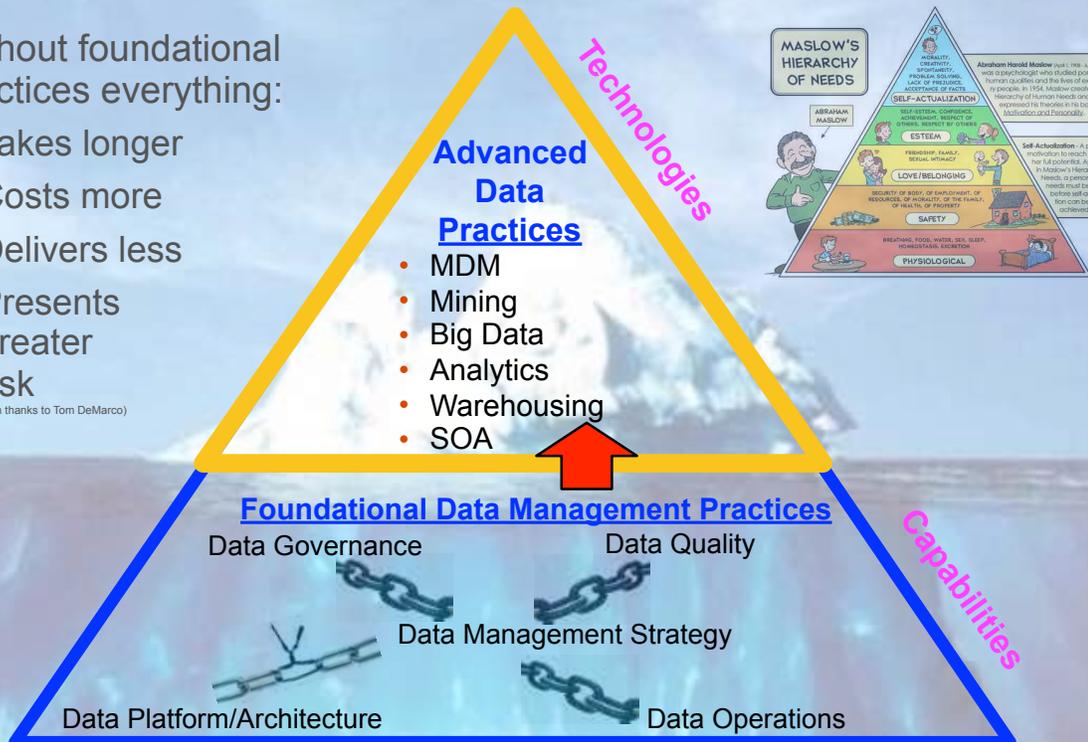


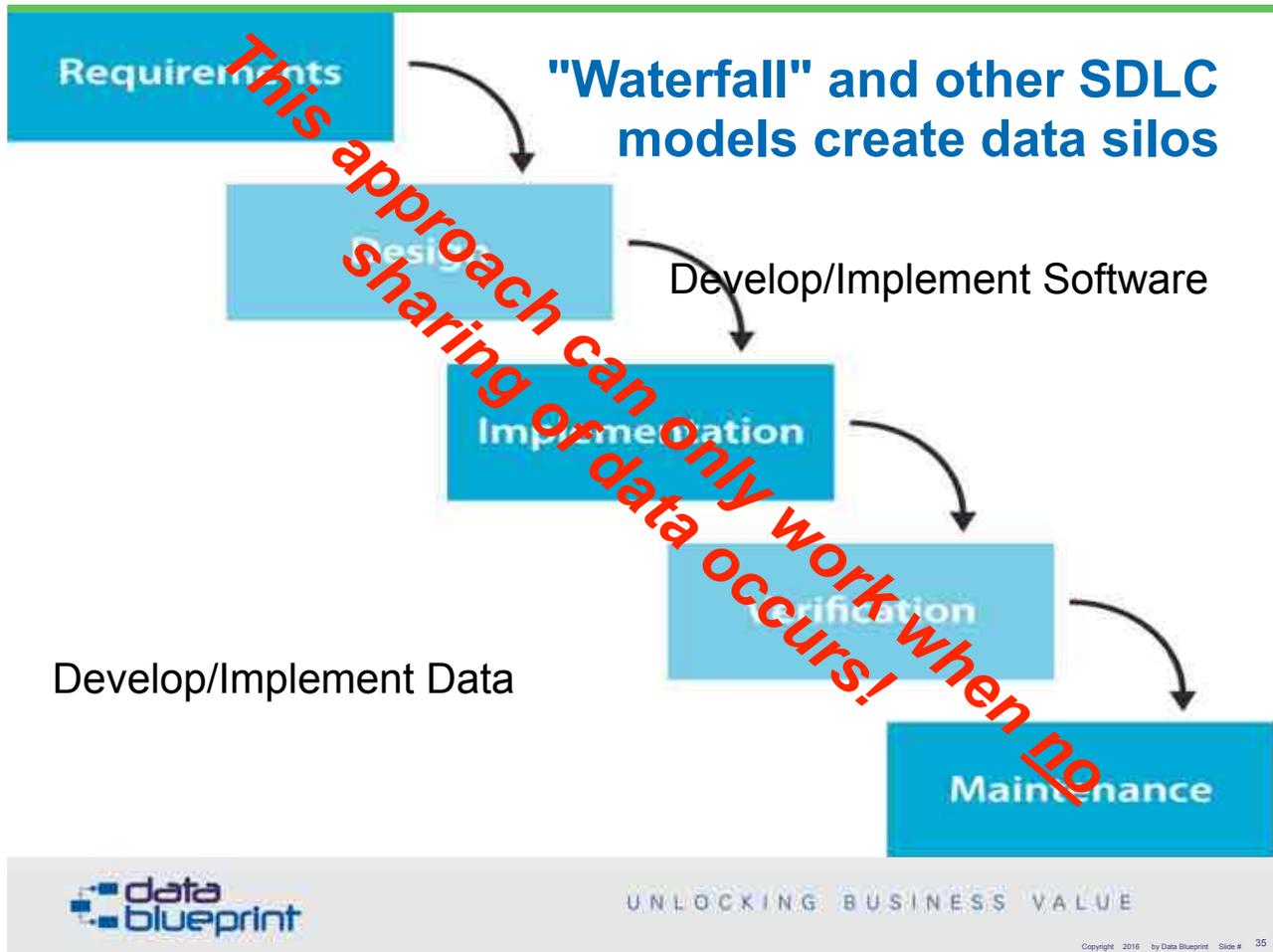
## Data Management Practices Hierarchy

Without foundational practices everything:

- Takes longer
- Costs more
- Delivers less
- Presents greater risk

(with thanks to Tom DeMarco)





## Evolving Data is Different than Creating New Systems

Common Organizational Data  
(and corresponding data needs requirements)



Future State



Evolve

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

Systems Development Activities



Create

New Organizational Capabilities



# Information Security Maturity Model

Reactive

Proactive

## Blocking & Tackling

- Lack of Executive support
- Underfunded
- Understaffed
- Lack of metrics for reporting
- Set up for failure

## Compliance Driven

- Control-based security approach
- Align to mandatory regulations
  - EU/PII Data protection
  - FFIEC
  - HIPAA
  - ISO 2700x
  - PCI
  - NCUA

## Risk-Based Approach

- Multi-layered security and risk-based approach
- Using behavior analytics and evaluating new technologies frequently
- Linking events across multiple disciplines

<http://krebsonsecurity.com/2015/04/whats-your-security-maturity-level/>



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 37

## Who wrote this ... ?

- *In considering any new subject, there is frequently a tendency first to overrate what we find to be already interesting or remarkable, and secondly - by a sort of natural reaction - to undervalue the true state of the case.*
- Augusta Ada King, Countess of Lovelace - aka Ada Lovelace, publisher of the first computing program

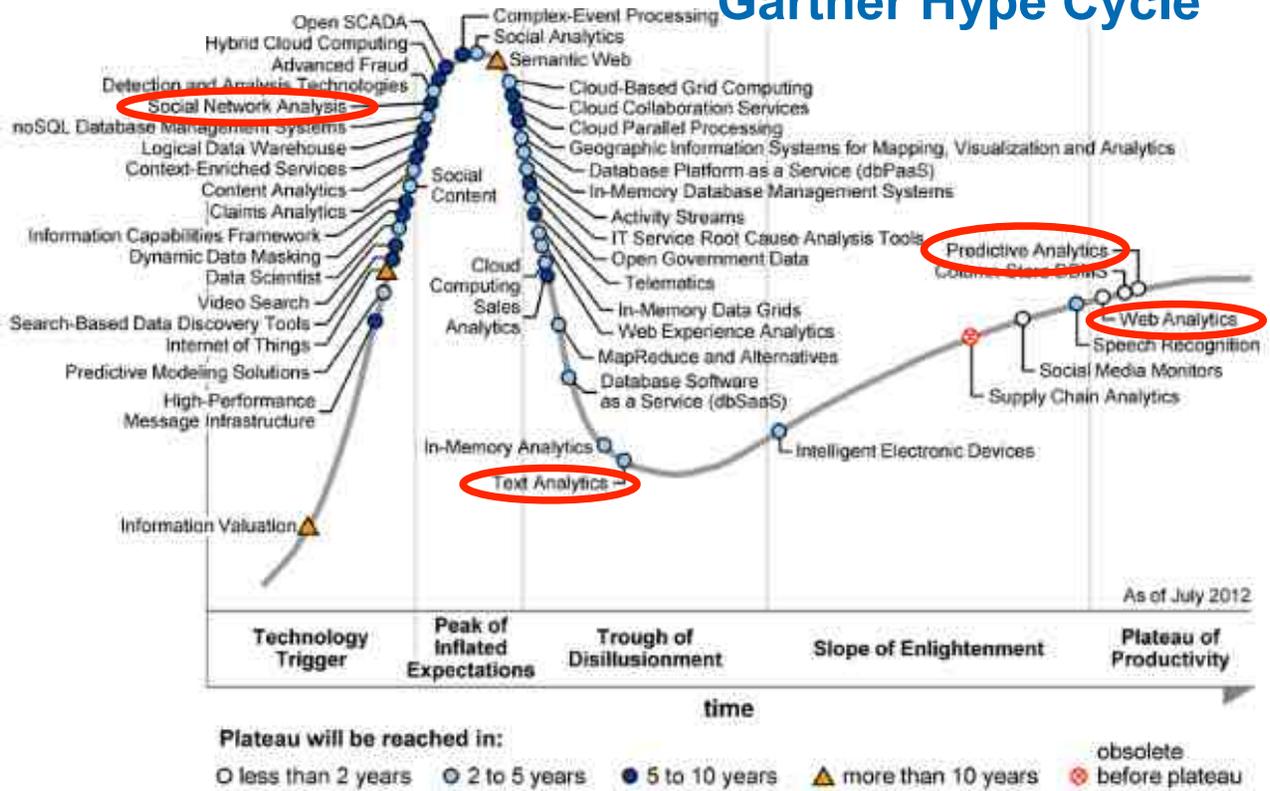


UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 38



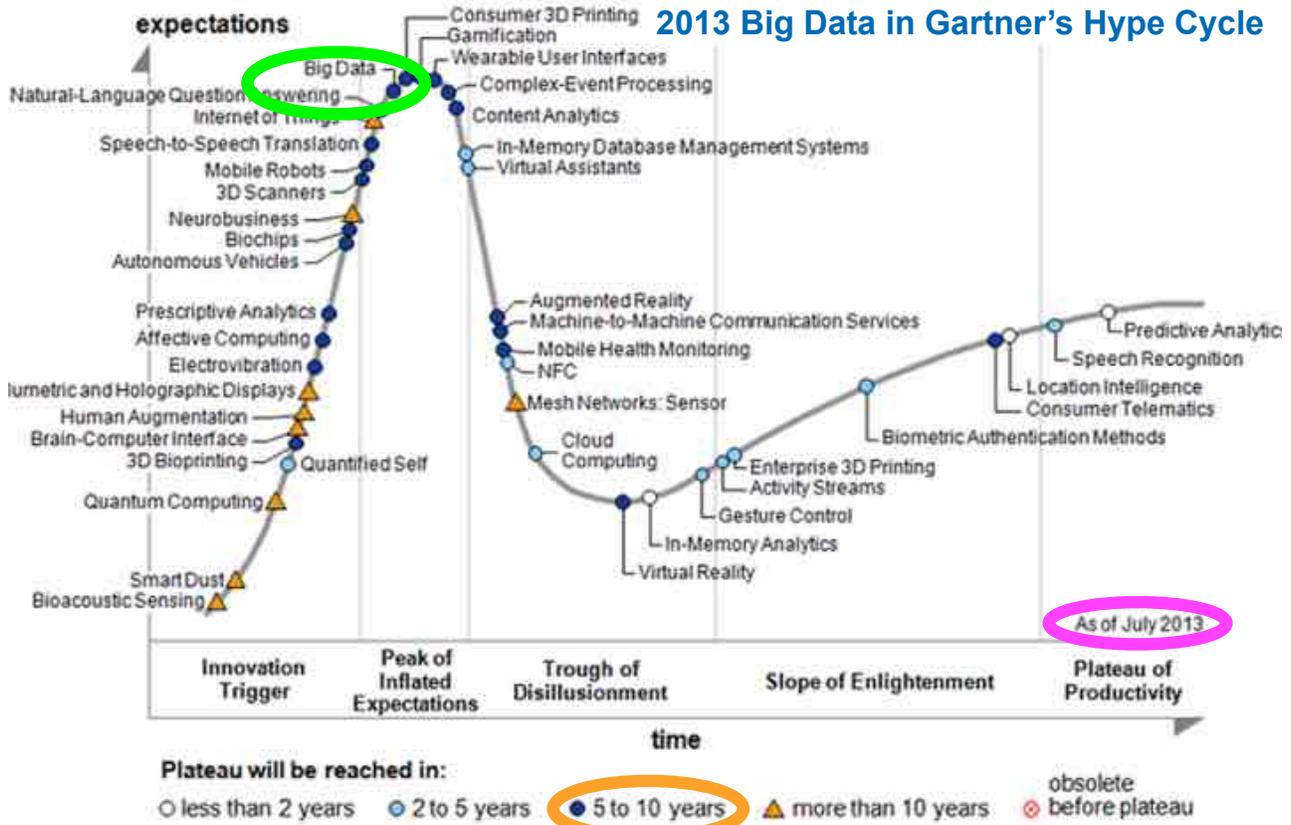
# Gartner Hype Cycle



# 2012 Big Data in Gartner's Hype Cycle



## 2013 Big Data in Gartner's Hype Cycle



## 2014 Big Data in Gartner's Hype Cycle



## 2015 Big Data in Gartner's Hype Cycle



*The Demise of Big Data, Its Lessons and the State of Things to Come*

## 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
- An all-encompassing term for any collection of datasets so large and complex that it becomes difficult to process using traditional data management tools or traditional data processing applications
  - Wikipedia 2014
- Shorthand for advanced trends in technology that open the door to a new approach to understanding the world and making decisions.
  - NY Times 2011
- The broad range of new and massive data types that have appeared over the last decade
  - Tom Havenport 2014
- Definition of a very large file, typically to the extent that its manipulation and management present significant logistical challenges.”
  - Oxford English Dictionary 2014
- 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!*

## "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!

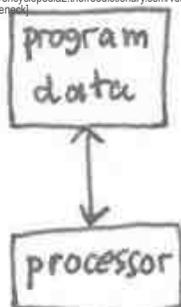


## Non-von Neumann Processing/Efficiencies

- von Neumann bottleneck (computer science)

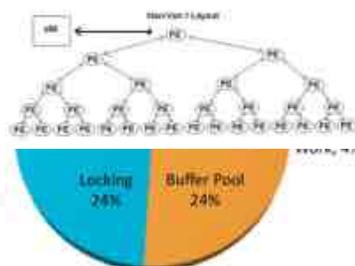
- "An inefficiency inherent in the design of any von Neumann machine that arises from the fact that most computer time is spent in moving information between storage and the central processing unit rather than operating on it"

(<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



- Decomposition
- Reassembly
  - not optional!

<b>Vertical Apps</b> 	<b>Ad/Media Apps</b> 	<b>Business Intelligence</b> 	<b>Analytics and Visualization</b> 
<b>Log Data Apps</b> 	<b>Data As A Service</b> 	<b>Operational Infrastructure</b> 	<b>Structured Databases</b> 
<b>Analytics Infrastructure</b> 	<b>Infrastructure As A Service</b> 	<b>Technologies</b> 	



# Big Data Technologies by themselves, are a One Legged Stool



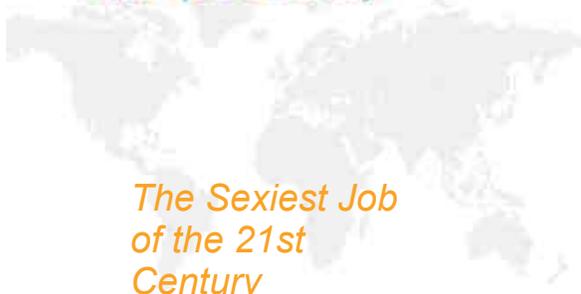
*Governance is the major means of preventing over reliance on one legged stools!*

## Data Science

McKinsey Global Institute



Big data: The next frontier for innovation, competition, and productivity



*The Sexiest Job of the 21st Century*

*Big data—capturing its value*

**\$300 billion**  
potential annual value to US health care—more than double the total annual health care spending in Spain

**€250 billion**  
potential annual value to Europe's public sector administration—more than GDP of Greece

**\$600 billion**  
potential annual consumer surplus from using personal location data globally

**60%** potential increase in retailers' operating margins possible with big data

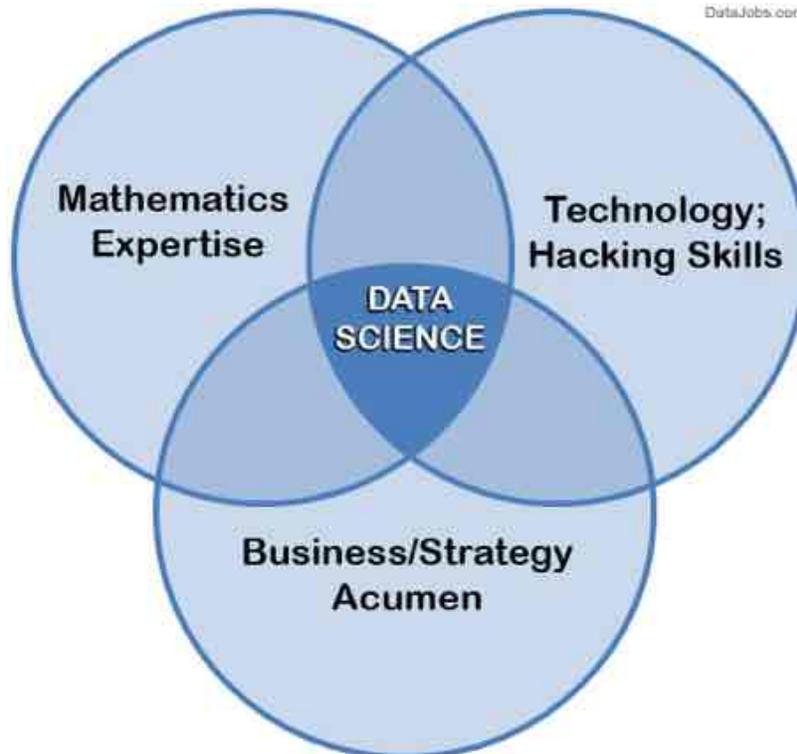
**140,000–190,000**  
more deep analytical talent projected by 2020

**1.5 million**  
more data-savvy managers needed to take full advantage of big data in the United States

# What is a Data Scientist?



# Data Scientist?



# Data Scientist?

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

### MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



### PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages e.g. R
- ☆ Databases: SQL and NoSQL
- ☆ Relational systems
- ☆ Parallel databases and parallel query processing
- ☆ Multidimensional concepts
- ☆ Hadoop and Hadoop/MapReduce
- ☆ Custom reducers
- ☆ Experience with AWS, like AWS



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint. Slide # 57

# Data Scientist?

Data Science allows firms to better predict what and when consumers are likely to buy. The ability to write algorithms that find relationships in datasets is unable to provide actionable insight.



## Urgent Need

Data Scientists - those with the technical savvy and analytical chops to derive meaning from all the information - are in high demand

## Skills by the Numbers

The skills and talents that make a fantastic Data Scientist

Complex Formulas	40%
Consumer Psychology	25%
Business Acumen	25%
Programming Languages	10%

## The Challenge

- Data Mining
- Analysis
- Communication

## Industry Niche Titles

- Financial Institutions/ Decision Scientist
- Retailers/Omni Channel Expert
- Marketing Agencies/Consumer Behaviour Analyst
- E-commerce/Analytics Expert

## Did you Know?

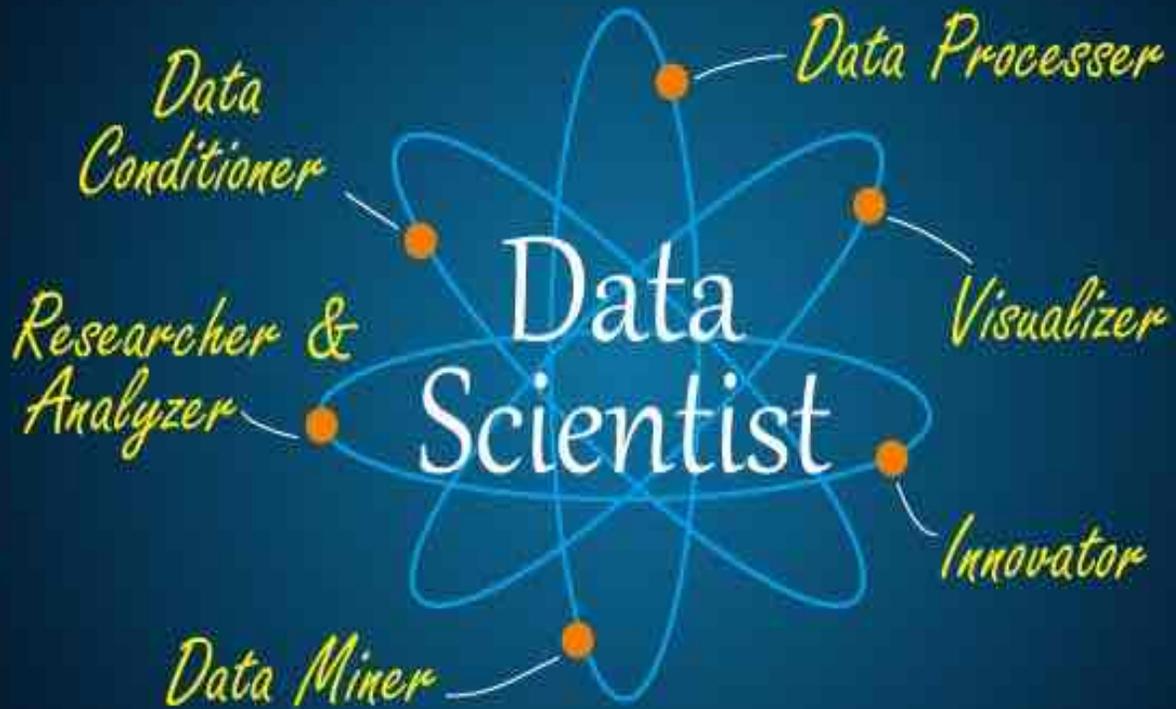
Google's Eric Schmidt claims that every two days now we create as much information as we did from the dawn of civilization up until 2003.



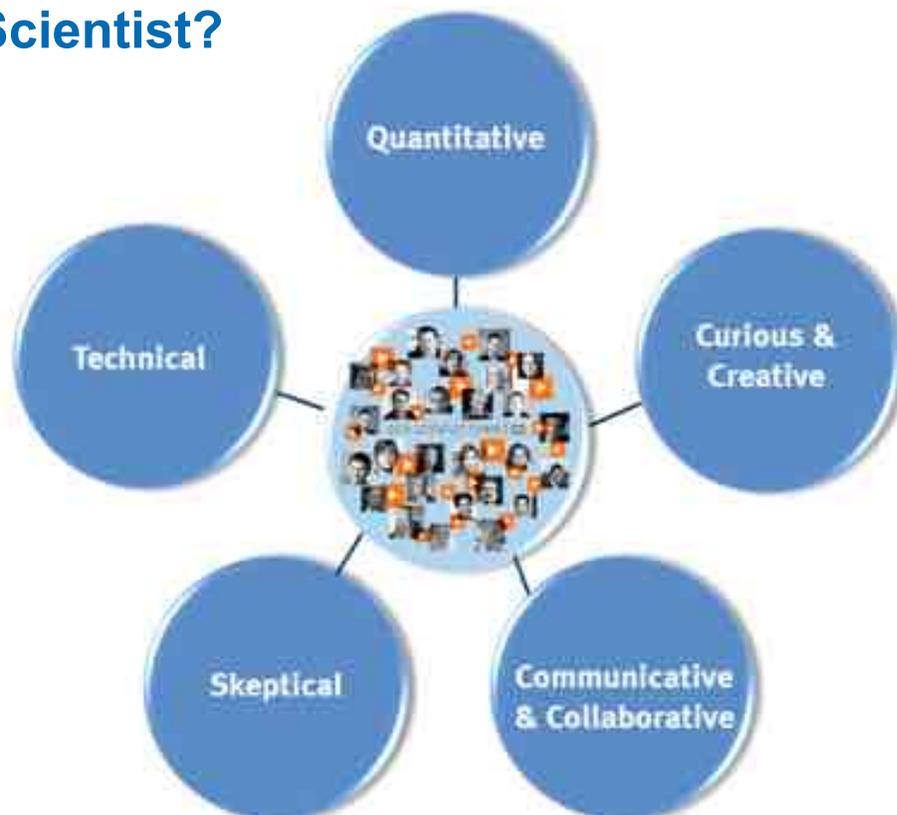
UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint. Slide # 58

## Data Scientist?



## Data Scientist?

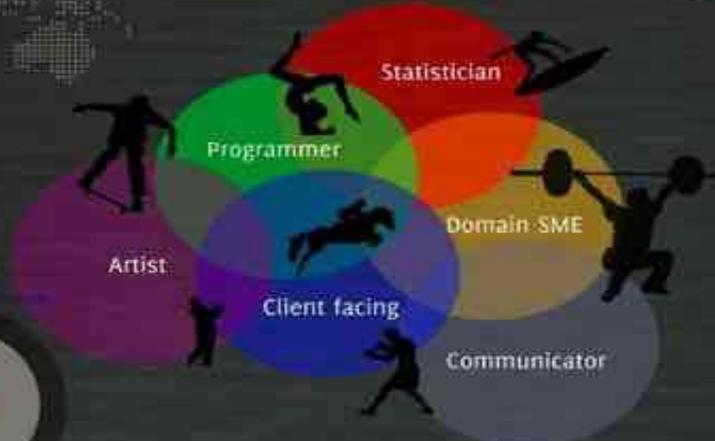


# Data Scientist?



# DATA SCIENTIST

Analysts. Industry Leaders. Trend Setters. Academicians and almost everyone has opinion on Data Scientist's must-have skills...



Question is does someone really exist with all these skills and if yes why would the person work for someone else?



# Data Scientist?

# BIG DATA SCIENTIST





**Customer**

Current Customer      Potential Customer

VIP-Customer?

Ex-Customer?

**data blueprint**

UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 65

## Data Scientist?

- Wrong level of abstraction
  - Actuarial Data Scientist
  - Forensic Data Scientist
  - Financial Data Scientist
  - Forestry Data Scientist
  - Marine Data Scientist
  - Chemical Data Scientist
  - Canine Data Scientist
  - Economic Data Scientist
  - Manufacturing Data Scientist
  - FDA Data Scientist
  - Cancer Data Scientist
  - Diabetes Data Scientist
  - Weather Data Scientist
  - Metadata Scientist



**data blueprint**

UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 66

## We believe ...

**Asset:** A resource controlled by the organization as a result of past events or transactions and from which future economic benefits are expected to flow [Wikipedia]

- Today, data is the most powerful, yet underutilized and poorly managed organizational asset
- Data is your
  - Sole
  - Non-depletable
  - Non-degrading
  - Durable
  - Strategic
- Asset
  - Data is the new oil!
  - Data is the new (s)oil!
  - Data is the new bacon!
- Our mission is to unlock business value by
  - Strengthening your data management capabilities
  - Providing tailored solutions, and
  - Building lasting partnerships

	Data Assets	Financial Assets	Real Estate Assets	Inventory Assets
Non-depletable	Available for subsequent use	Can be used up		Can be used up
Non-degrading	✓	✓	Can degrade over time	Can degrade over time
Durable	Non-taxed		✓	✓
Strategic Asset	✓	✓	✓	✓

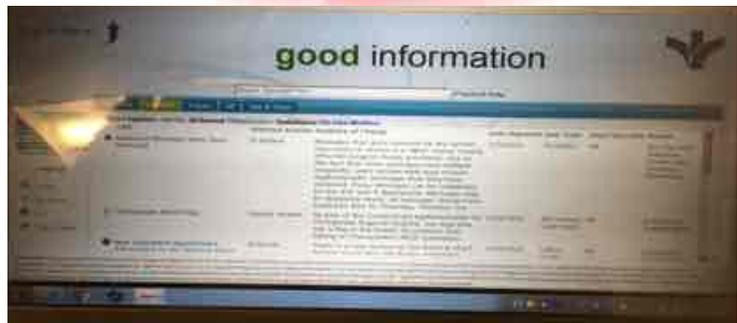
## Data Assets Win!



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 67

- A certain university
- The Egg Man
- Admission Date



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 68

# The Case for the Top Data Job (TDJ)

Rethinking the Essence of a Critically  
Lacking Business Function





Peter Aiken

# MONETIZING DATA MANAGEMENT



*Unlocking the Value in Your Organization's  
Most Important Asset.*

**PETER AIKEN WITH JUANITA BILLINGS**  
FOREWORD BY JOHN BOTTEGA



UNLOCKING BUSINESS VALUE

Copyright 2016 by Data Blueprint Slide # 69

## Questions?



It's your turn!



UNLOCKING BUSINESS VALUE



[datablueprint.com](http://datablueprint.com)

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