The "Big Picture" on Big Data

Princeton Section 307
Dinner Meeting
December 11, 2013
Richard Herczeg

ASQ Princeton Section
The Global Voice of Quality™

Objective of Talk

1. Deliver a Primer on Big Data.
2. How does this emerging topic apply to Quality ?
3. Be Interactive.
4. Continue discussions on Section's LinkedIn.
5. Where can I learn more.

My Background on Big Data

1. Significant Interest.
2. Belief about the Power of Data
3. SQL Structured Data
4. Unstructured Cross Functional Projects
5. Consumer Sales – 1,000 of SKU's
6. Dashboards
7. Survey data

What is Big Data

Define Data

- Data is raw or unorganized form (such as alphabets, numbers, or symbols) that refer to, or represent, conditions, ideas, or objects. Data is limitless and present everywhere in the universe.
- **In computing**, data is information that has been translated into a form that is more convenient to move or process. Relative to today's computers and transmission media, data is information converted into binary digital form.

What is Big Data

Define Big

- There was 5 exabytes of digital data in recorded time until 2003. In 2011, the same amount of data was created in 2 days. By 2013 that time period is expected to shrink to just 10 minutes.*
- For a CSP with 100 million customers, daily location data could amount to 50 terabytes or 5 petabytes, which can no longer be discarded due to regulators requiring the retention of CDRs.

* Fortune Magazine

What is Big Data

- **Volume, Velocity, Variety, Veracity**
- **Volume**
  - Traditional relational databases (Oracle, My SQL, SQLServer) tend not to scale well past about a terabyte of data.
  - Costs for traditional solutions to go beyond this point are very high.

What is Big Data

Byte Multiples

- Kilobyte (10^3), Megabyte (10^6), Gigabyte (10^9)
- Terabyte (10^{12})
  - Sloan Digital Sky Survey 140 TB to date
- Petabyte (10^{15})
  - Walmart stores 2.5 PB of data
  - Facebook stores 7 PB of photos per month
- Exabyte (10^{18})
  - All words ever spoken by humans would consume 5 EB
  - NSA's Utah facility estimated to hold 12 EB
  - Sum of world's data storage is 300 EB
- Zettabyte (10^{21})
  - Annual global IP traffic is about 1 ZB
- Yottabyte (10^{24})

5

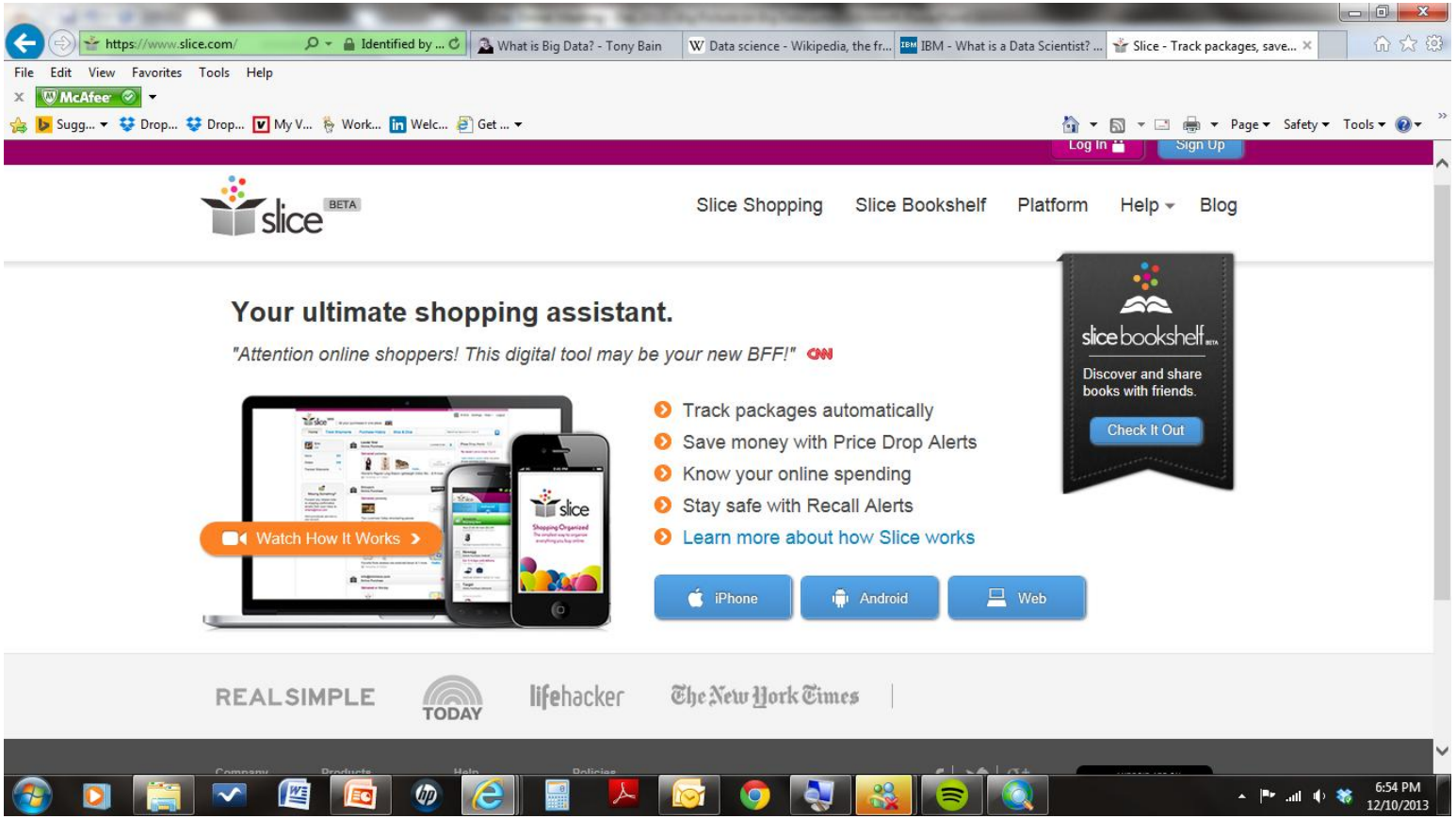
What is Big Data

- **Velocity**
  - Mobile Global Data is growing at 78% compounded growth rate and expected to exceed 10.8 exabytes per month in 2016.
  - Sometimes Big Data is static, but more often than not it is constantly streaming in, often at a high rate.
  - Efficiently updating a large repository of data with small incremental changes is technically challenging.
  - Requirement for reduced data latency – need for real-time/near time operational data.

What is Big Data

- **Variety**
  - From: Extract Data, Load in Warehouse, and Transform in Data Warehouse based on narrow variety and structured content.
  - To: Correlations of Call Center Conversations with emails, trouble tickets and social media blogs.
  - The source data can now include unstructured text, sound, and video in addition to structured data.
  - Traditional databases work best with highly structured and typed data that can fit into relational databases.
  - Increasingly, the world's data types are unstructured. Some say 80%

What is Big Data



What is Big Data

- **Veracity**
  - Big Data comes from outside our control, as a result suffers from bias, accuracy problems – calling into question both the quality of the source (its credibility) and implications to the data's target audience.
    - Example : "Likes" on Social Media placed by 3rd Parties or disgruntled employee, or quality of information from a 3rd party being used a primary input into other organizations and protocols around sharing of internal data.
  - Governance: MDM, Quality, Privacy , Data Life Cycle management.

Drivers of the Big Data Tsunami

1. Sophisticated Consumers
2. Automation
3. Monetization

What is Big Data

- "Buzz Word" – that describes the phenomenon of massive amounts of structured and unstructured data that is difficult to process using traditional database and software solutions and that can deliver insights previously thought too expensive to uncover/process.
- It's a journey.

What is Big Data

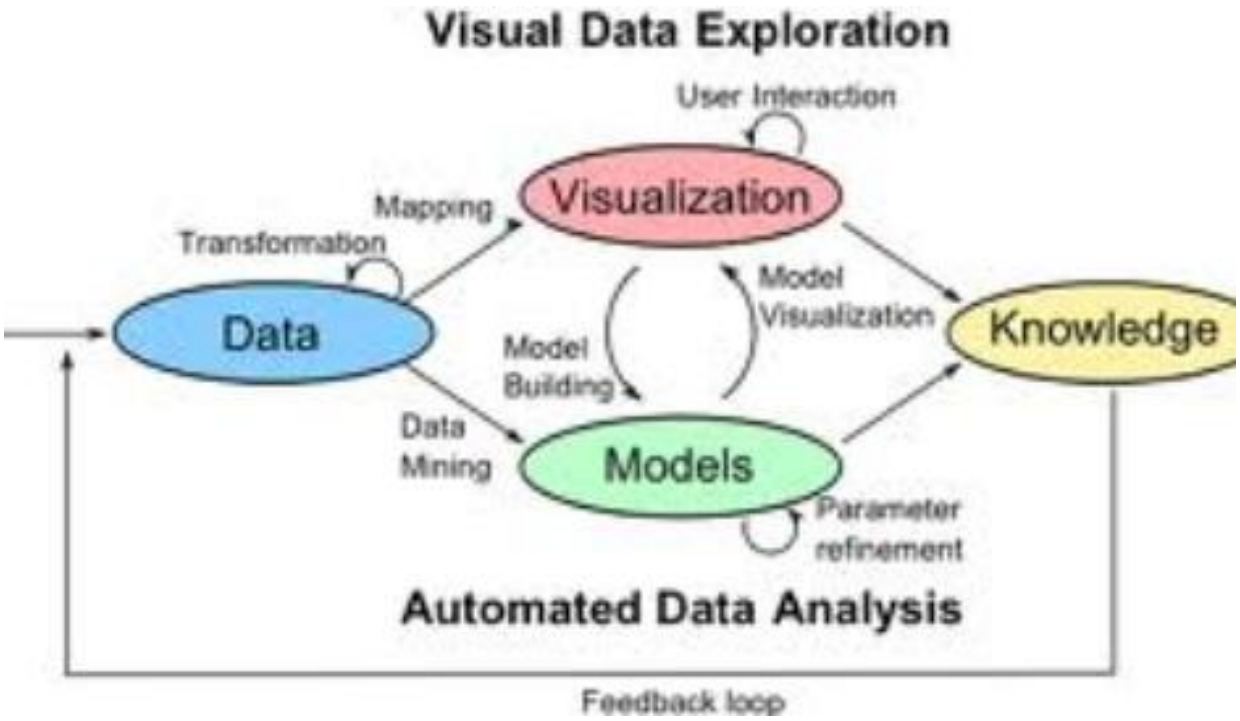
A Journey or Goal to Big Data Analytics

Princeton Section
ASQ The Global Voice of Quality™

Hadoop

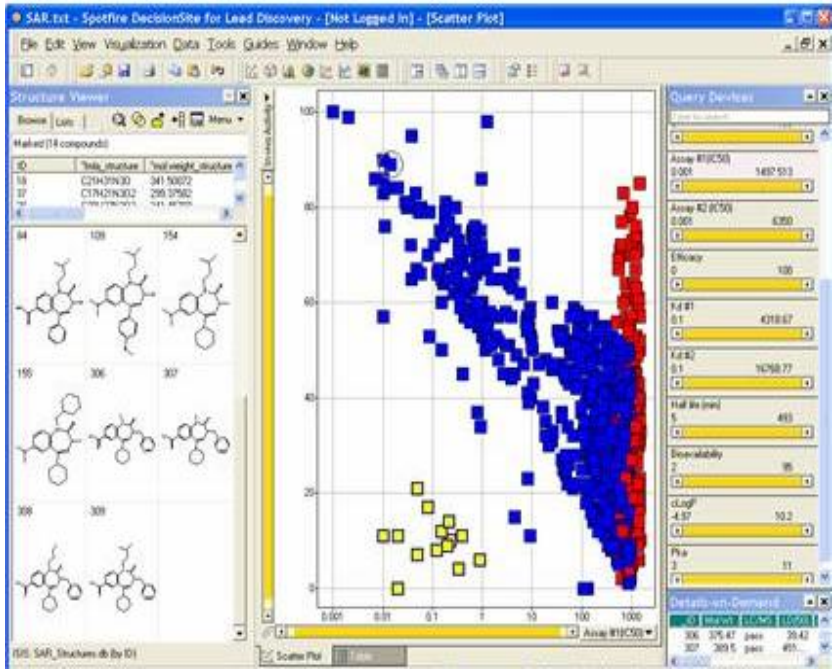
- Apache Hadoop is an open-source software framework that supports data-intensive distributed applications licensed under the Apache v2 license. It supports running applications on large clusters of commodity hardware. Hadoop derives from Google's MapReduce and Google File System papers.
- Core Platform for structuring Big data and solves the problem of making it more useful for analytics.

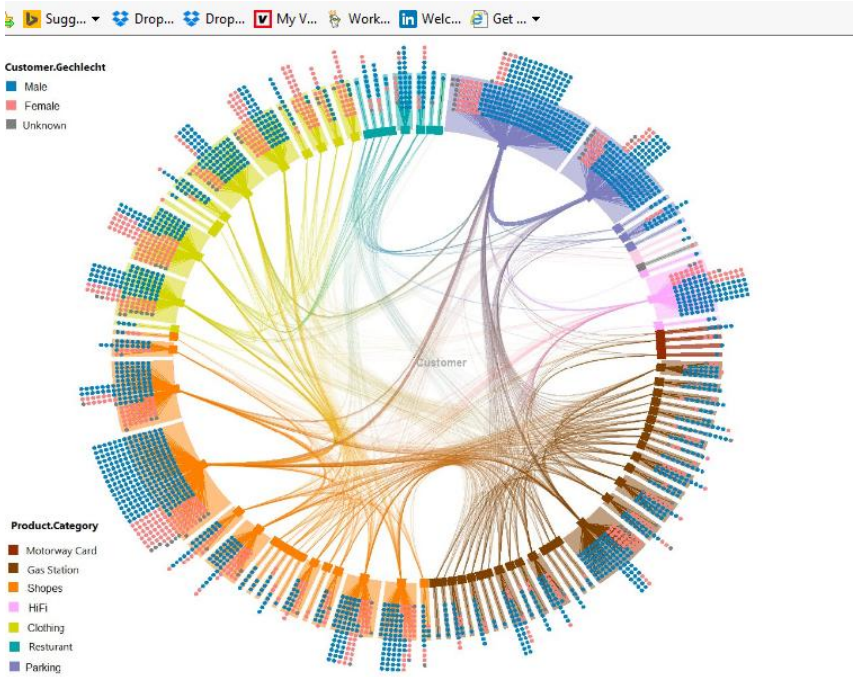
Data Visualization



Data Visualization

Is Worth a 1,00 Spreadsheets





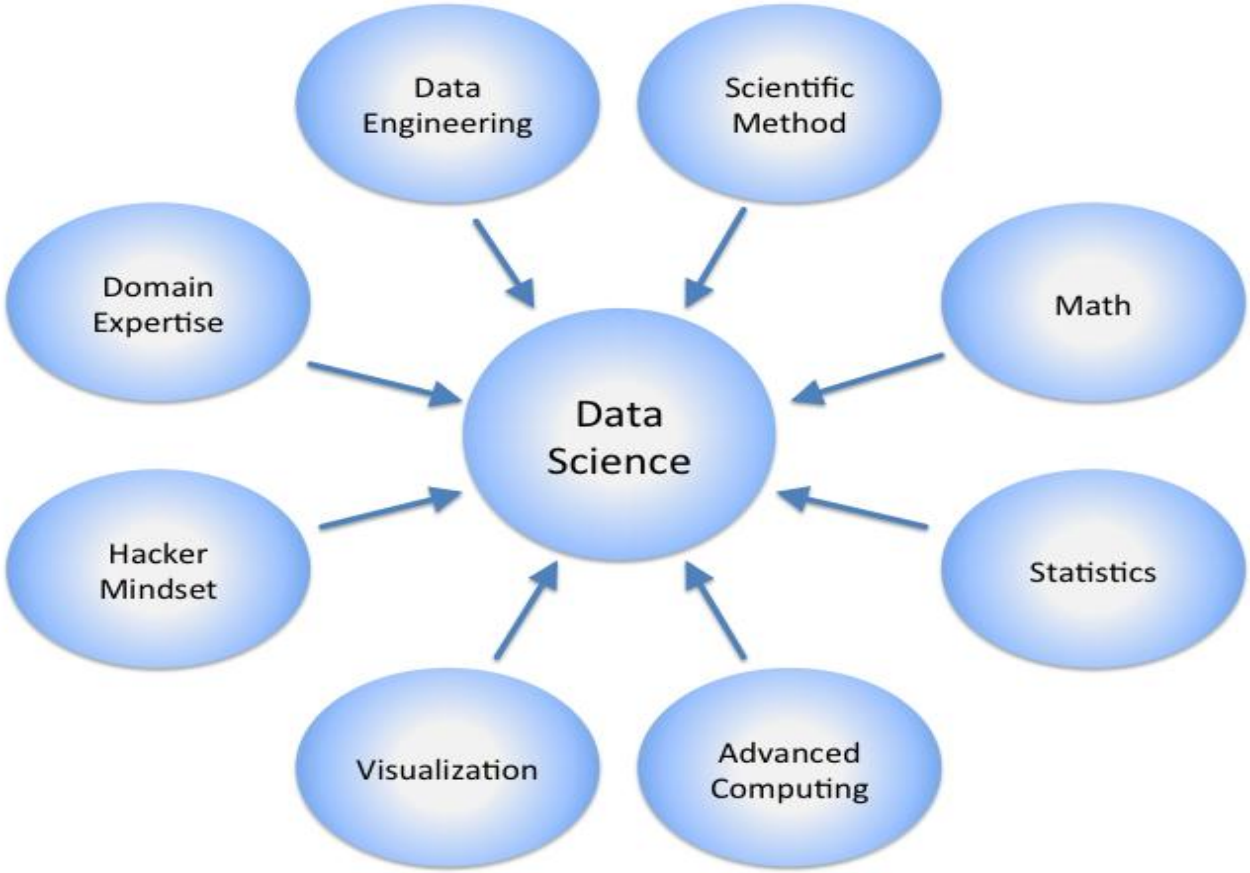
Text Analytics

Wordle – Word Count

What is the Big Quality Problem ?

- Quality Planning
- Data Collection Plans
- Separate Signal From Noise.
- How to turn Big Data into actionable information and Knowledge Management.
- Structured Data and Process.
- Data Bias, Context, Homogeneous, Quality, MSA.
- Hypothesis/ Cause & Effect
- ISO

Data Science



References

- Big Data Analytics - Arvind
- The Signal and the Noise – Nate Silver
- Super Crunchers – Ian Ayres
- Competing on Analytics – Davenport
- Analytics at Work – Davenport
- The Deciding Factor - Rosenberger