Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Basics of Big Data Analytics

BRETT AMIDAN
JEFFERY DAGLE

Pacific Northwest National Laboratory

NASPI Presentation (October 23, 2014)

What is Big Data?

- Any collection of data sets **so large** and **complex** that it becomes **difficult** to process using **traditional** data processing **applications**.
- The challenges include:
  - Analyzing,
  - Capturing,
  - Curating (sorting and cleaning),
  - Searching,
  - Sharing,
  - Storing,
  - Transferring,
  - Visualizing, etc.

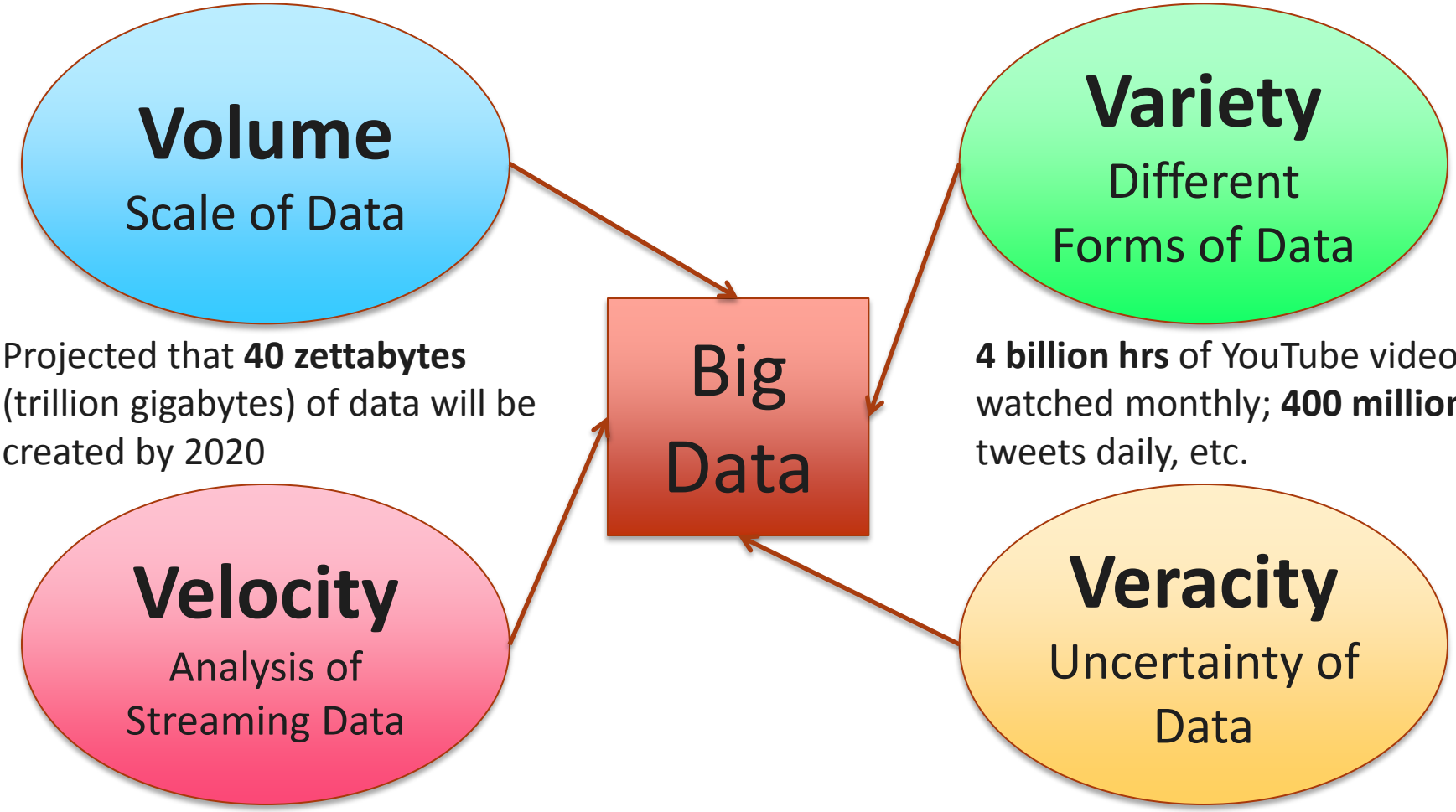
Now That's Big Data

- In January 2008 –

"Google currently processes over 20 petabytes of data per day through an average of 100,000 MapReduce jobs spread across its massive computing clusters. The average MapReduce job ran across approximately 400 machines in September 2007, crunching approximately 11,000 machine years in a single month."

Chances are your problem won't be as big or require as many machines; however the analysis strategies and methodologies are similar.

The Four V's of Big Data



Projected that **40 zettabytes** (trillion gigabytes) of data will be created by 2020

4 billion hrs of YouTube videos watched monthly; **400 million** tweets daily, etc.

18.9 billion network connections Projected by 2016

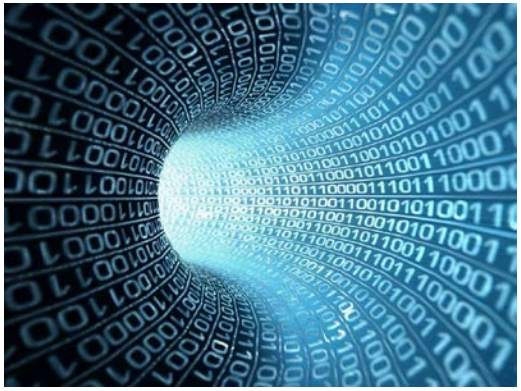
Poor data quality costs the US economy around **$3.1 trillion a year**

Why now?

- Increases in storage capacities

- Increases in processing power and cluster computing ability

- Availability of data



These are key enablers for the growth of “Big Data”

15 Most Powerful Big Data Companies*

1. IBM
2. HP
3. Teradata
4. Oracle
5. SAP
6. EMC
7. Amazon
8. Microsoft
9. Google
10. VMware
11. Cloudera
12. Hortonworks
13. Splunk
14. 10Gen
15. MapR

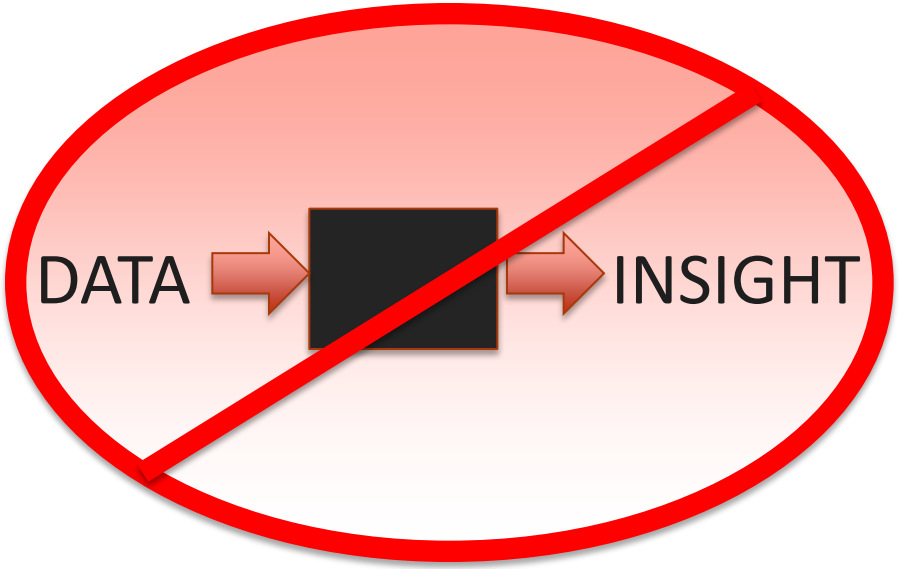
* According to NetworkWorld (in 2013)

Who Does Big Data Analytics?

- *Fortune 500* article June 13, 2014 by Katherine Noyes "These Big Data Companies Are the Ones to Watch"
- 10 industry experts were asked who the notable "Big Data" companies were. 13 companies were listed by 2 or more experts.
- Dean Abbott (co-founder of Smarter Remarketer) –

"Most of these companies will go away because the most important part of the big data movement will be how to use data operationally—to make decisions for the business rather than who can merely crunch more data faster."

What Big Data Analytics Isn't



Structured and unstructured data do NOT enter a black box application, resulting in useful insight

It just isn't that easy!

What Big Data Analytics Is
Step 1: The Plan

Big Data Analytics without decision objectives is just data storage

Often resources are spent putting data into a database, without much thought about what it'll be used for

- What would you like to do with your data?

  Visualizations? Analyses?
- What questions would you like to answer?
- What insights are needed to influence your business objectives?

These questions and more affect the **data** and the **analyses** that are required to provide insight

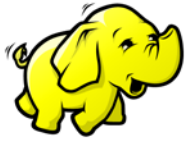
Step 2 – The Data

Understanding your objectives will help in determining answers to questions like –

- **What** data should I collect?

  Data can be structured, like sensor data streams, or unstructured, like text reports or video.

- **How** will that data be stored?



  Do you really need to store your data in a Hadoop or other database cluster? Or is there a better way to manage it in context with your business?

- Do **complete raw data streams** need to be stored, or can summaries be stored?

  Do you need every microsecond of information stored, or can summaries be created over each minute, hour, or day?

- Are there **real-time** data requirements?

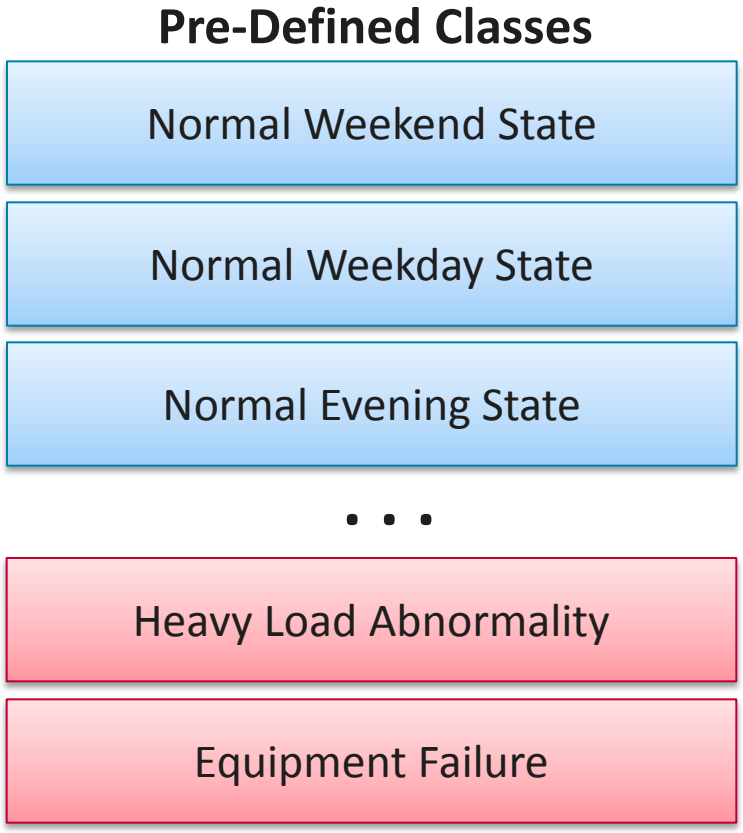
Step 3 – The Analyses

- **Data Quality!**
  - Initial analyses will identify issues
  - Narrow down data quality to the subset of data most meaningful to your objectives
  - *"Data quality can mean the difference between success and failure"*
- Visualizations & Analytics
  - Parallel processing possibilities (Chunkwise analysis)
  - Sampling the data (is it appropriate for what you are answering?)
  - Employing divide and recombine (map/reduce) techniques (influences data storage)

Analysis Possibilities

- **Classification** (Predictive)
- **Clustering** (Descriptive)
- **Regression** (Predictive)
- **Collaborative Filtering** (Predictive)
- **Longitudinal Analyses** (Descriptive/Predictive)

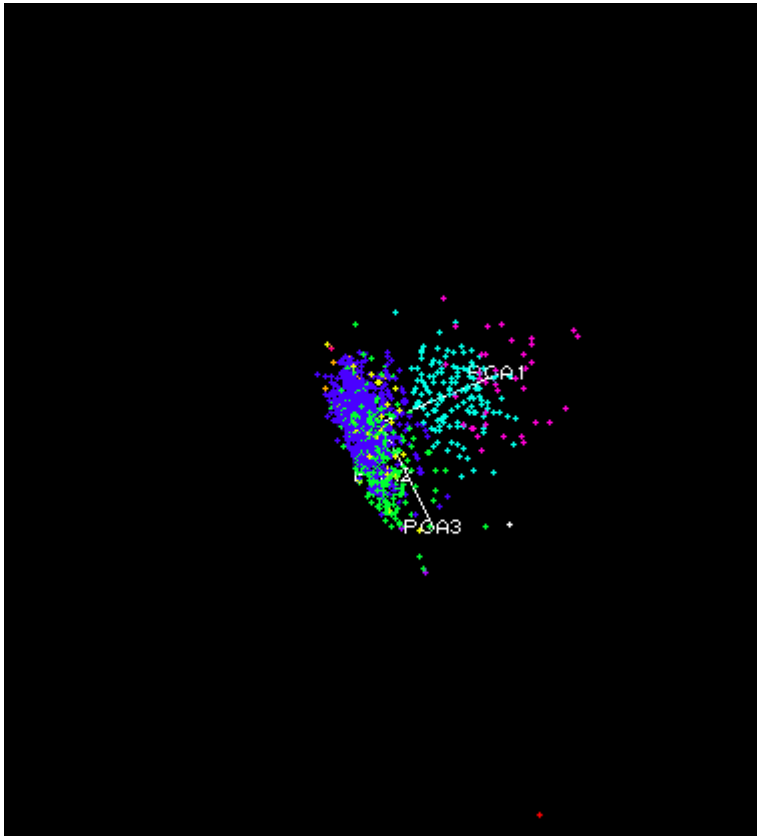
Statistical Classification



- Grouping new, previously unseen records into already defined classes (defined by historical data and outcomes)

Statistical Clustering

- Grouping records into groups using an unsupervised approach (letting the data organize itself into the groups)



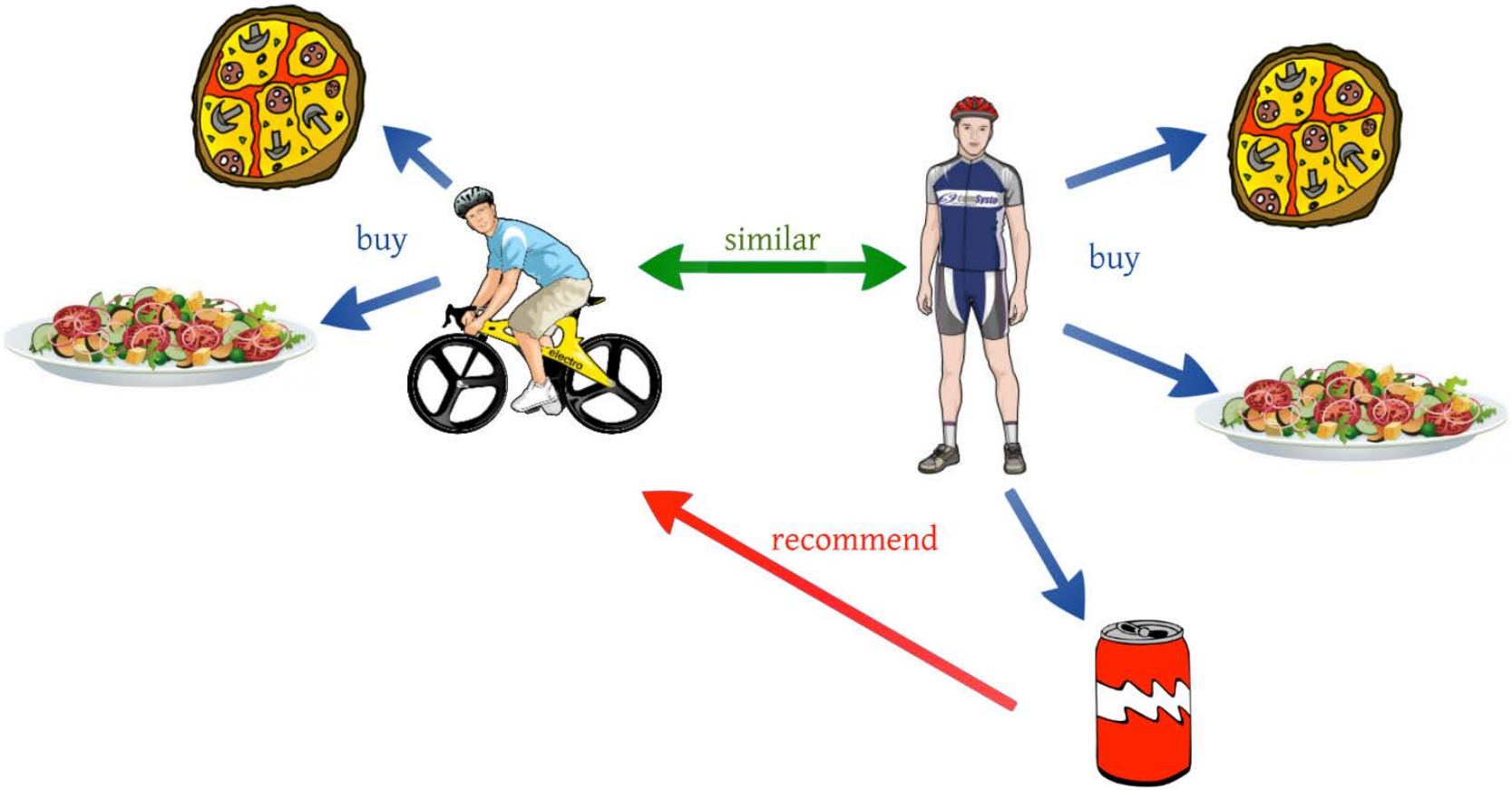
Regression

- Determining trends and/or correlations in the data

Although these examples are simple linear regression, regression can include many predictor variables and can be non-linear.

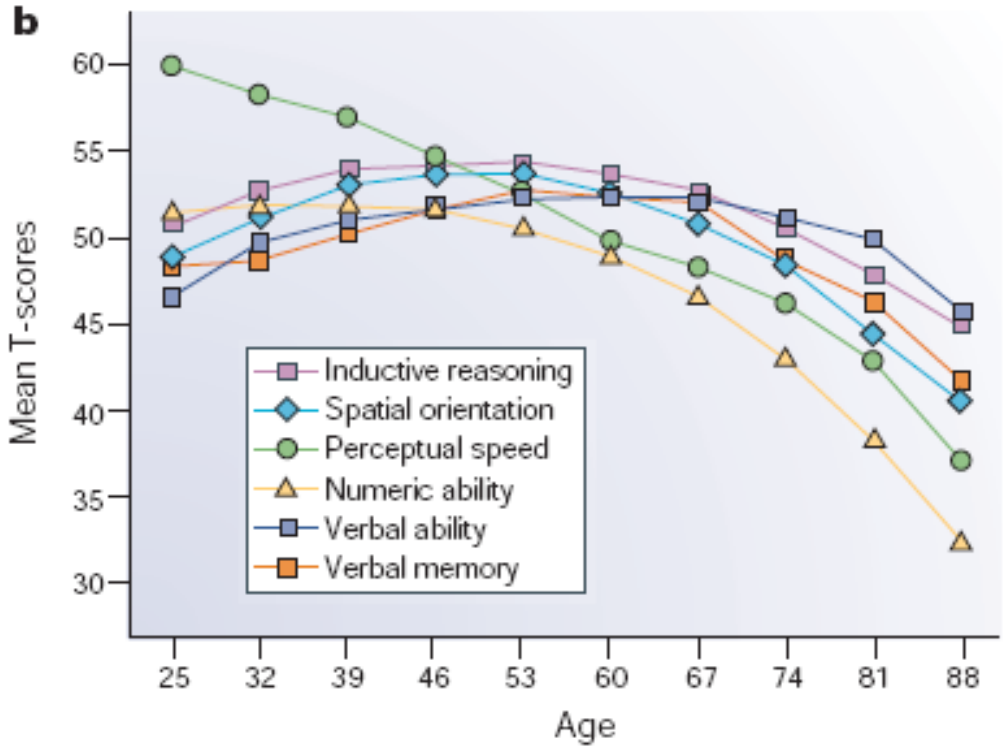
Collaborative Filtering

- Filtering information to find the information that is of interest

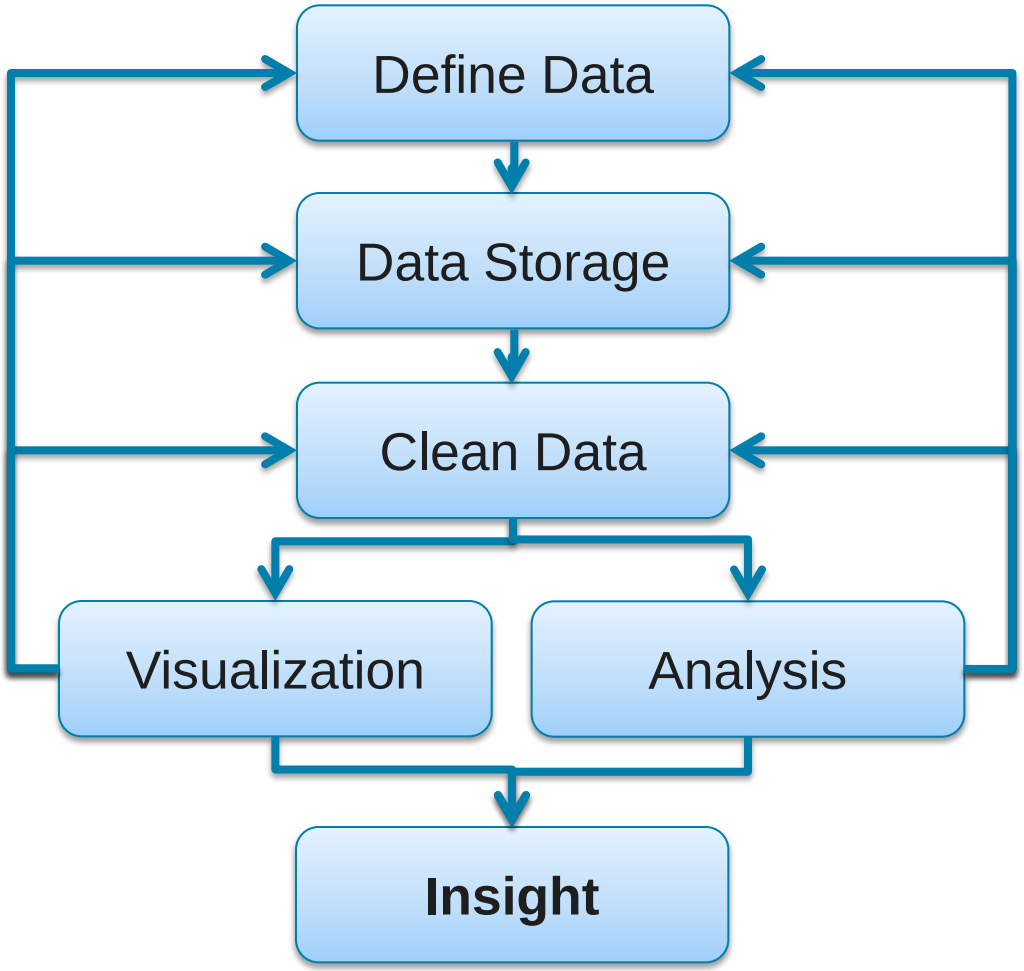


Longitudinal Analyses

- Discovering trends and patterns in time-based data



It’s an Iterative Process



7 Key Insights Concerning Big Data*

1. Data exist in every industry & business function & are now an important factor of production, alongside labor & capital.
2. Using big data creates value in 5 ways -
    1. Makes information transparent and usable;
    2. Can collect more accurate and detailed performance information on everything, exposing variability and boost performance;
    3. Leads to more precisely tailored products or services;
    4. Sophisticated analytics can improve decision-making; &
    5. Can improve the development of next generation products & services.
3. Big data will become a key basis of competition & growth for individual firms.
4. Using big data will underpin new waves of productivity growth and consumer surplus.

* McKinsey & Company (May 2011)

7 Key Insights Continued

5. Use of big data will matter across sectors, but some sectors are set for greater gains (i.e. computer and information sectors, finance and insurance, healthcare, government, etc).
6. There will be a shortage of talent necessary to take advantage of big data (by 2018, a projected shortage of 140,000 to 190,000 data scientists).
7. Several issues will need to be addressed to capture the full potential of big data (i.e. privacy, security, intellectual property, liability)

The Future of Big Data Analytics

We are currently only scratching the surface

- **Data Quality** will continue to be one of the most important issues.
- Present-day successes (i.e. retail, search engines, etc.) will lead to many **more industries** looking to apply big data analytics to provide needed insight.
- **More packaged tools and technologies** will be developed as big data analytics becomes more mainstream.