

# Big Data – Big Deal? Can Big Data Enable More Accurate Risk Estimate?

---

NATASHA BALAC, PH.D.

DATA INSIGHT DISCOVERY, INC.

MARK HERETH

BLACKSMITH GROUP/P-PIC

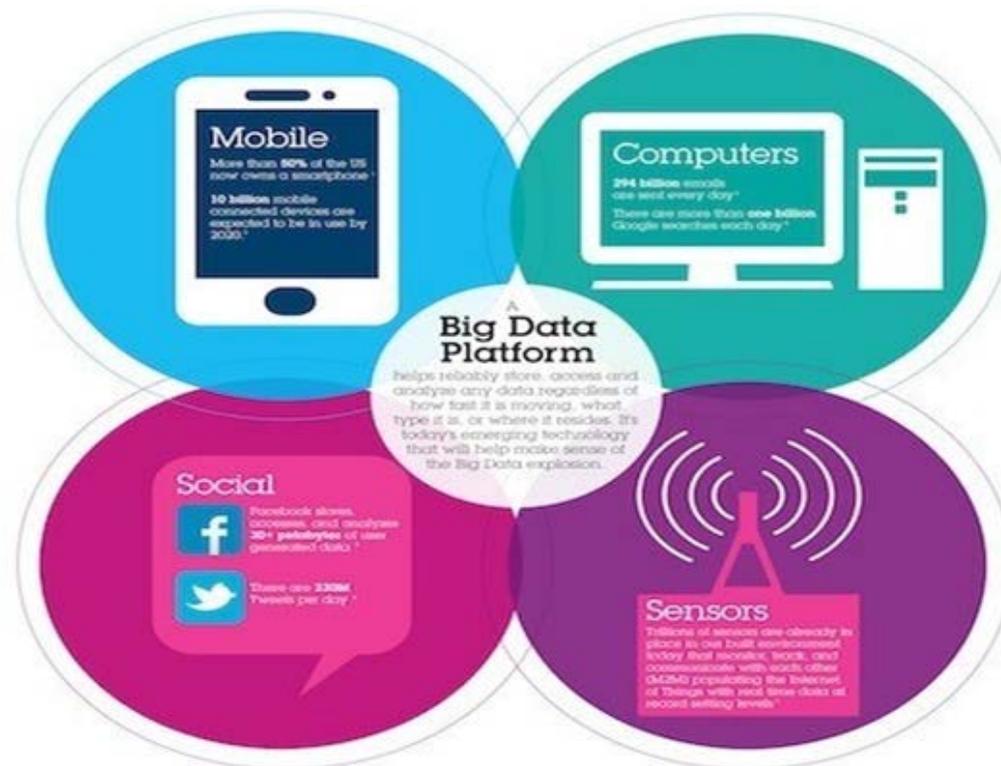
data.  
insight.  
discovery.

# Agenda

---

- Big Data
- Predictive Analytics
- How Have They Been Applied?
- How Can They Be Applied to Improve Pipeline Risk Assessment?

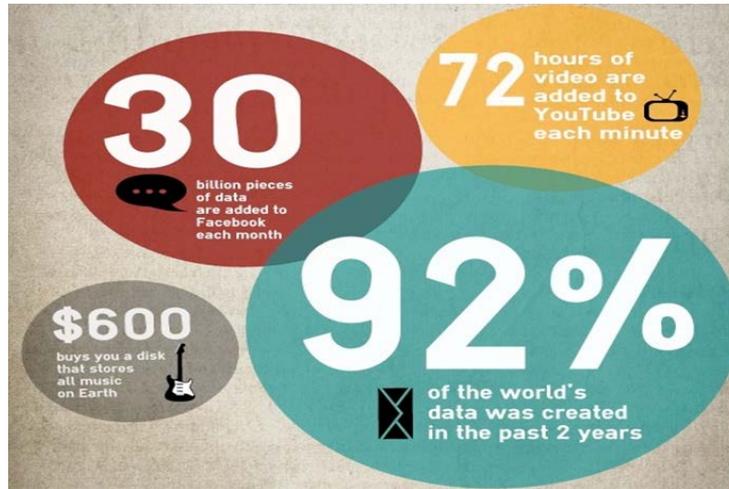
# Big Data – Making The World Go Around



**Data is moving in from a variety of sources – are you keeping up?**

data.  
insight.  
discovery.

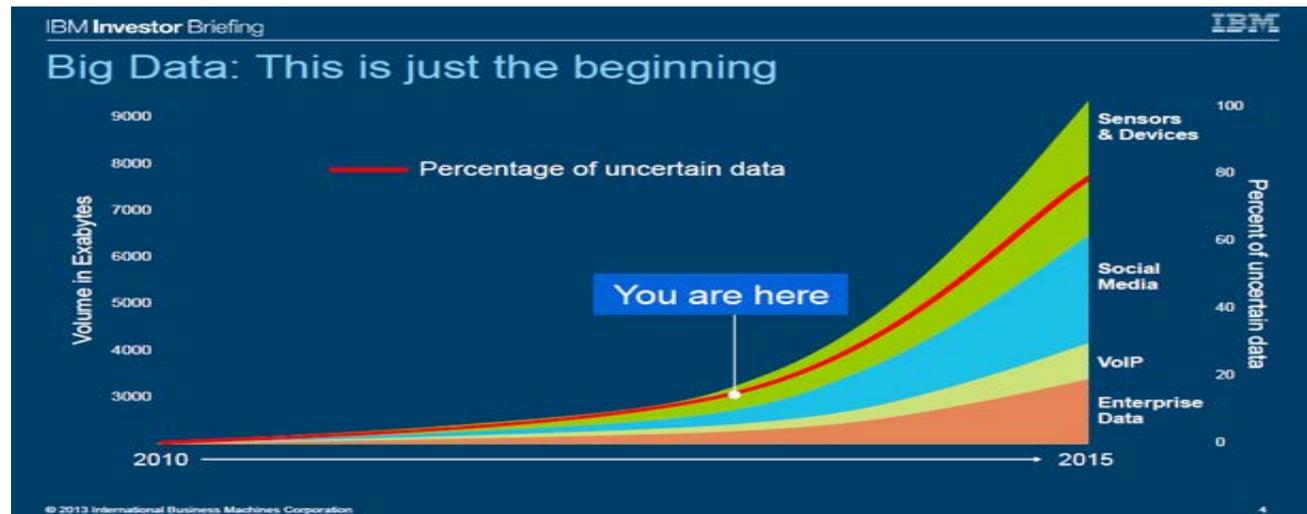
# How Big is Big Data?



The New York Times April 14, 2013

**40**  
TRILLION GIGABYTES  
Size of digital universe by 2020, up from 130 billion in 2005.

Source: IDC/EMC



data.  
insight.  
discovery.

# Big Data Definition

## Wikipedia

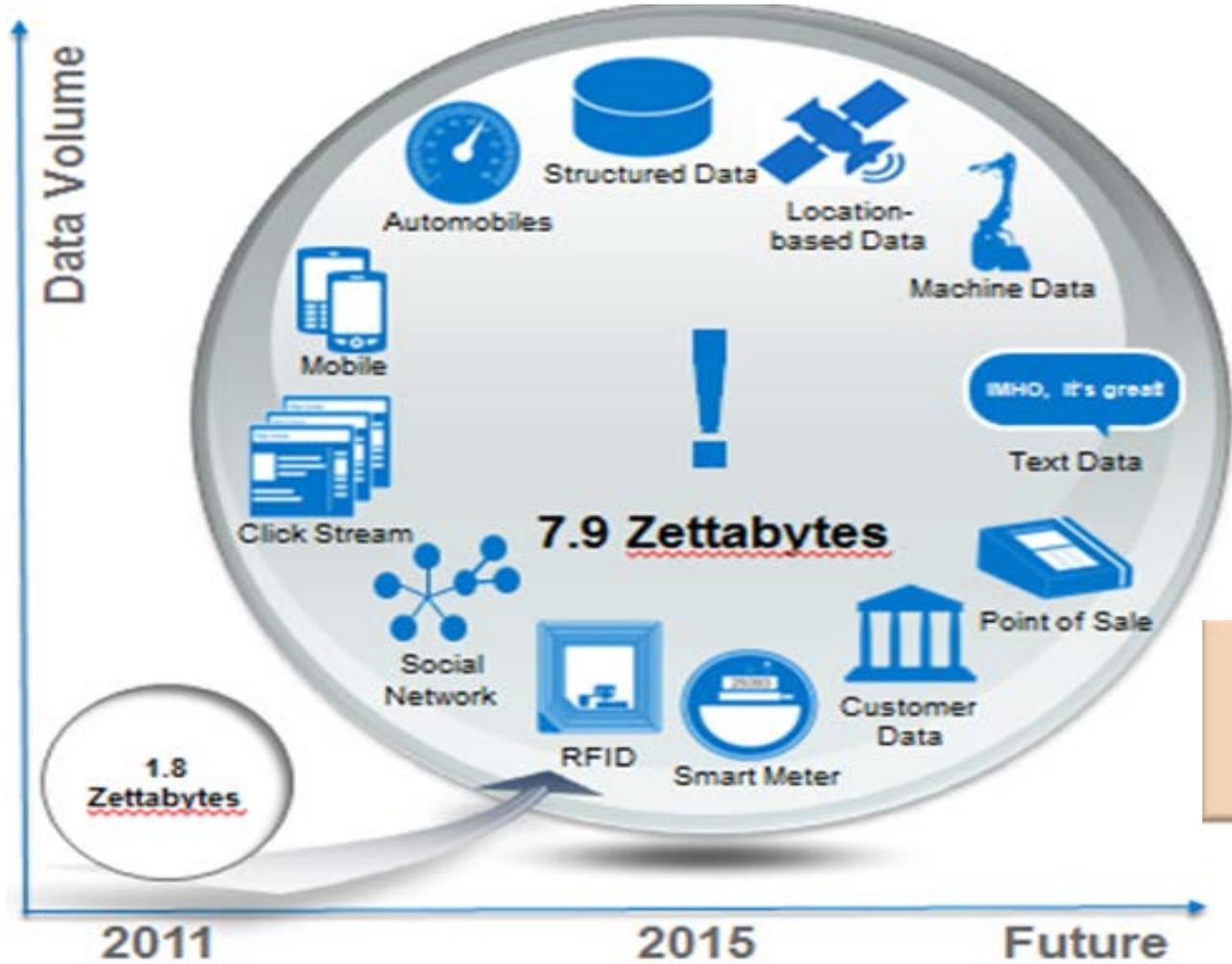
---

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using traditional database management tools or data processing applications

### **Definition of *BIG DATA* (Merriam-Webster):**

an accumulation of data that is too large and complex for processing by traditional database management tools

# Big Data Growth



- 1 Terabyte = 1024 Gigabytes
- 1 Petabyte = 1024 Terabytes
- 1 Exabyte = 1024 Petabytes
- 1 Zettabyte = 1024 Exabytes

data.  
insight.  
discovery.



1 NEW DEFINITION IS ADDED ON URBAN

1,600+ READS ON Scribd

13,000+ HOURS MUSIC STREAMING ON PANDORA

12,000+ NEW ADS POSTED ON craigslist

370,000+ MINUTES VOICE CALLS ON skype

98,000+ TWEETS

20,000+ NEW POSTS ON tumblr.

13,000+ iPhone APPLICATIONS DOWNLOADED

320+ NEW twitter ACCOUNTS

100+ NEW Linked in ACCOUNTS

1 associatedcontent NEW ARTICLE IS PUBLISHED

THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!

QUESTIONS ASKED ON THE INTERNET...

Answers.com 100+ 40+ YAHOO! ANSWERS

6,600+ NEW PICTURES ARE UPLOADED ON flickr

50+ WORDPRESS DOWNLOADS

25+ HOURS TOTAL DURATION

600+ NEW VIDEOS

70+ DOMAINS REGISTERED

60+ NEW BLOGS

168 MILLION EMAILS ARE SENT

694,445 SEARCH QUERIES

1,700+ Firefox DOWNLOADS

695,000+ facebook STATUS UPDATES

125+ PLUGIN DOWNLOADS

1,500+ BLOG POSTS

79,364 WALL POSTS

510,040 COMMENTS

GO-Globe.com

Google

Google Search

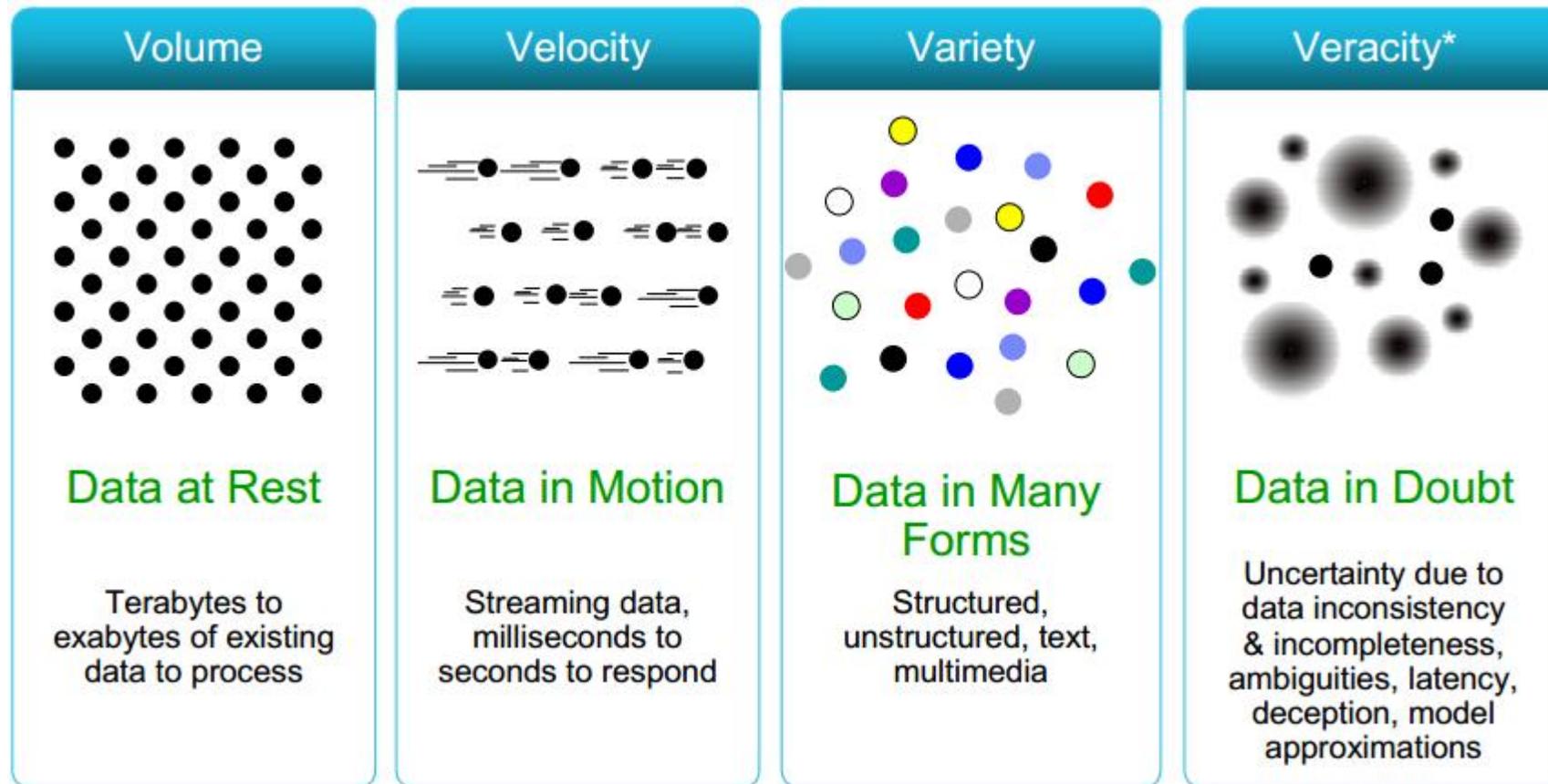
data.  
insight.  
discovery.

# Tip of the Iceberg



data.  
insight.  
discovery.

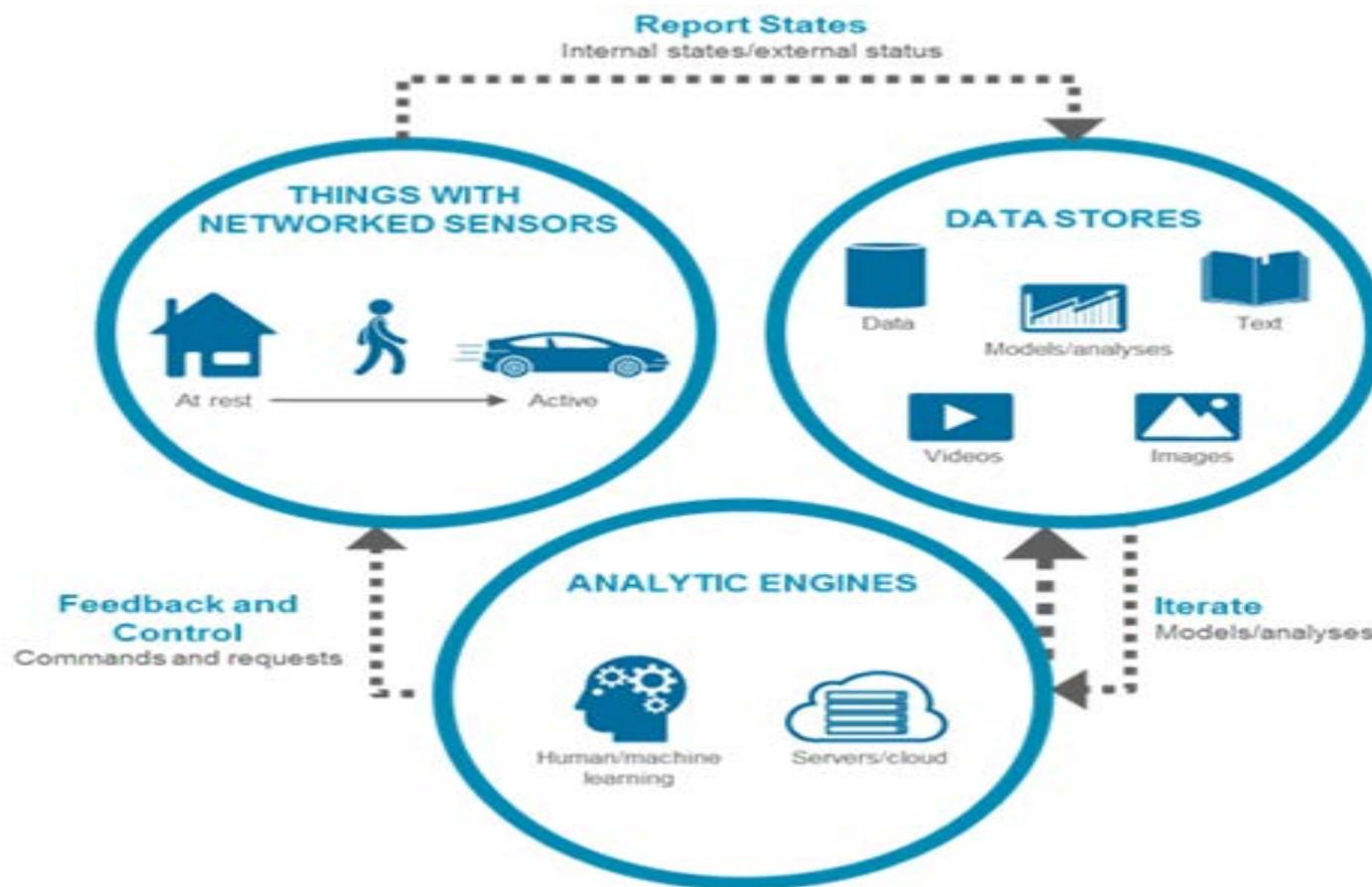
# 4 V's of Big Data



IBM, 2012

data.  
insight.  
discovery.

# THE INTERNET OF THINGS

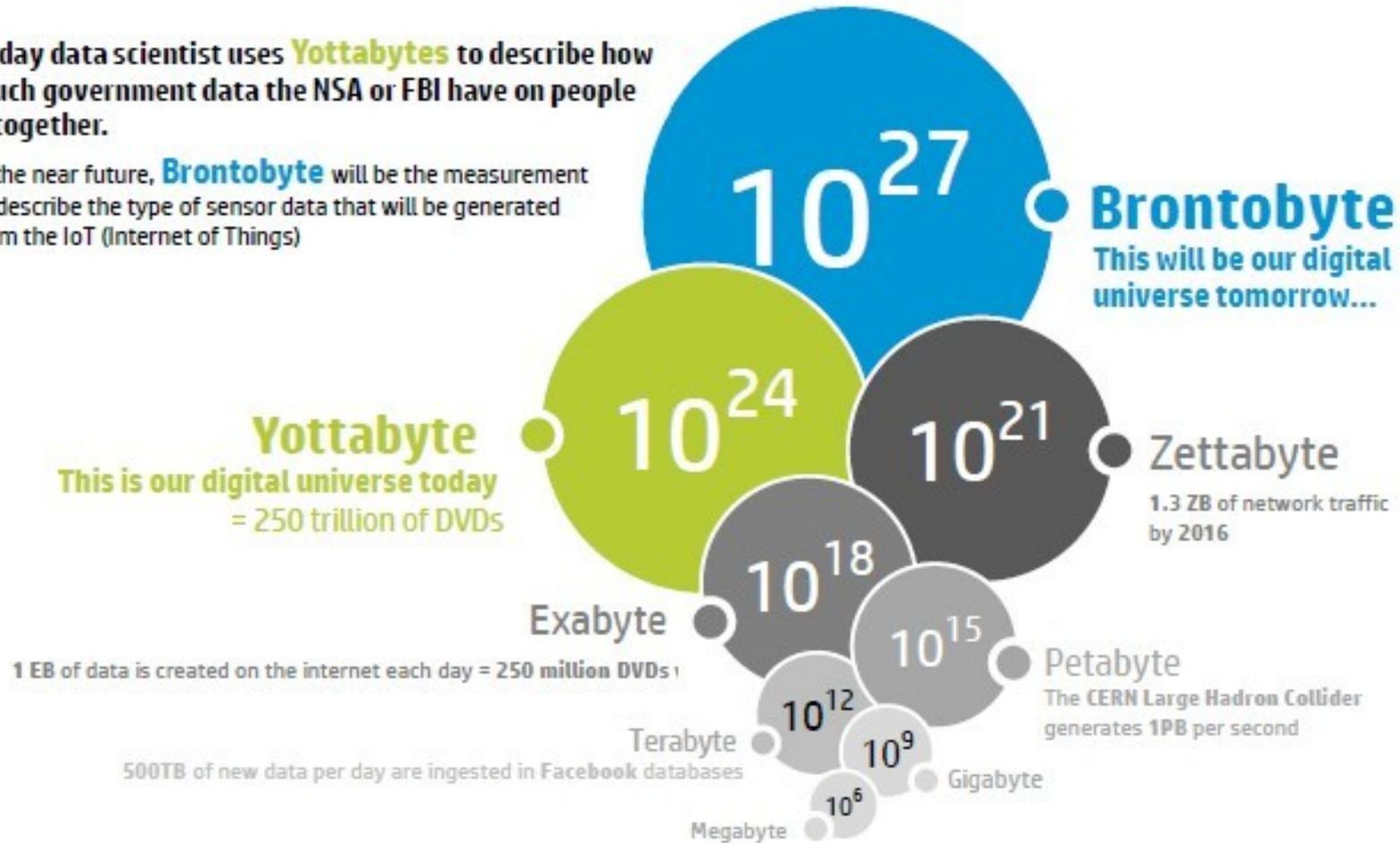


data.  
insight.  
discovery.

# Growing Internet of Things Data

Today data scientist uses **Yottabytes** to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)



# Why Predictive Analytics?

---

Build upon the strength of index-based models by defining relationships that reflect increasing risk and ultimately failure

- Increase the value of “system condition” data
- Reduce (or eliminate) confounding and bias of multiple parameters
- Identify relationships not otherwise discernable

Build upon the “certainty” of probabilistic models by defining the relationships that lead to failure

- Increase the value of quantification
- Enable threats to be viewed on a common basis

Use Expert Knowledge to test and refine analyses

Consider a broad slate of analytical tools

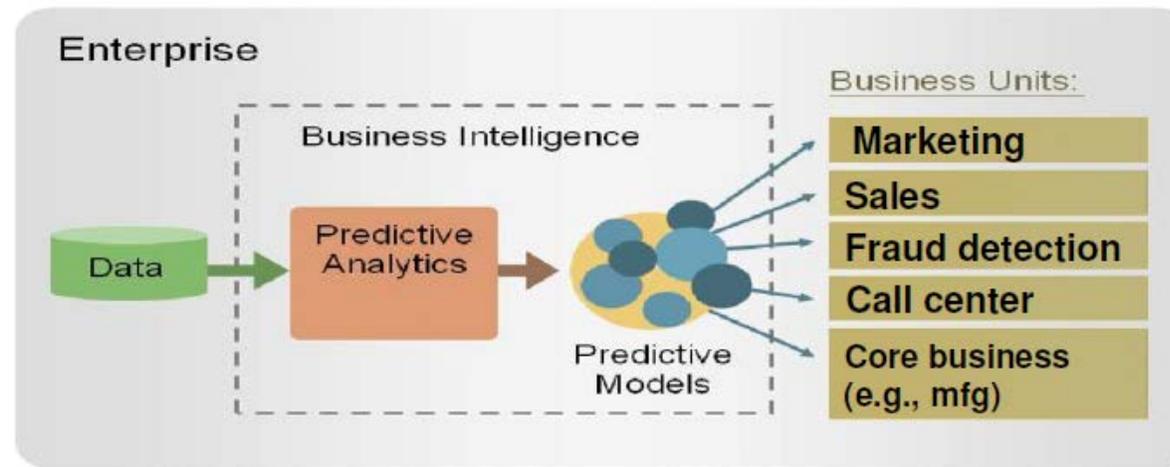
- Not just Monte Carlo Simulations
- Draw upon analogous situations from other industries
  - Box Jenkins – managing sensor data sets in Finance, Environmental and Energy
  - Probabilistic based models
  - Artificial Neural Networks

# From Data to Smart Decisions

Employee data? Consumer data? **Sensor Data? Inspection Data? Third party data?**

With so much raw data available today, every organization must harness the most relevant data to drive real-time

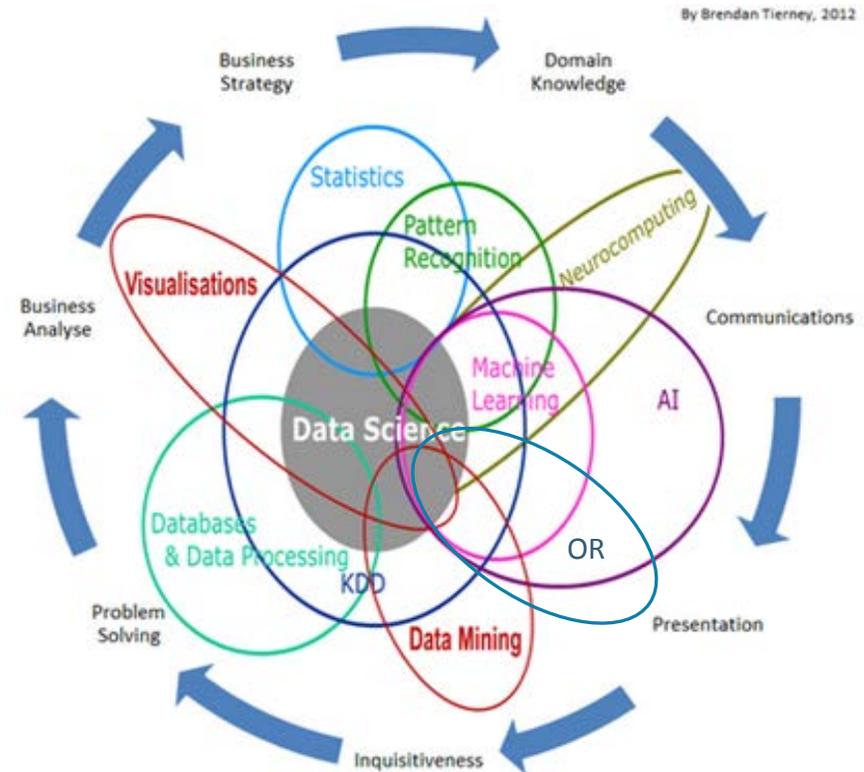
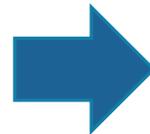
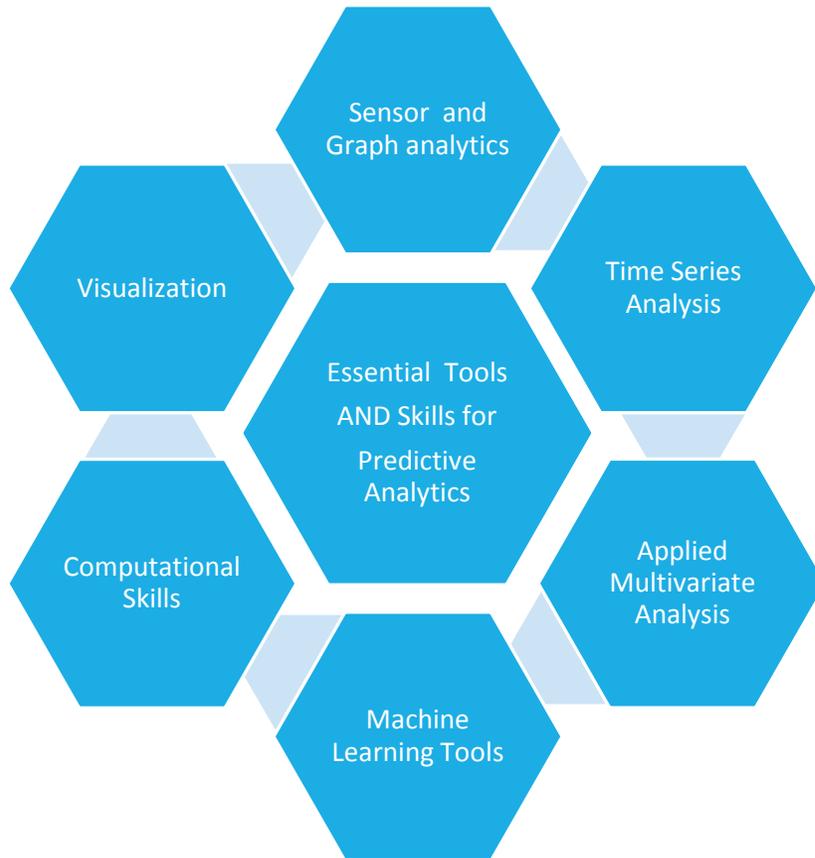
- Insightful decision-making
- Maximizing investment



Reference: PAW, 2012

data.  
insight.  
discovery.

# Predictive Analytics Skills



data.  
insight.  
discovery.

# Predictive Analytics Case Study: Netflix

---



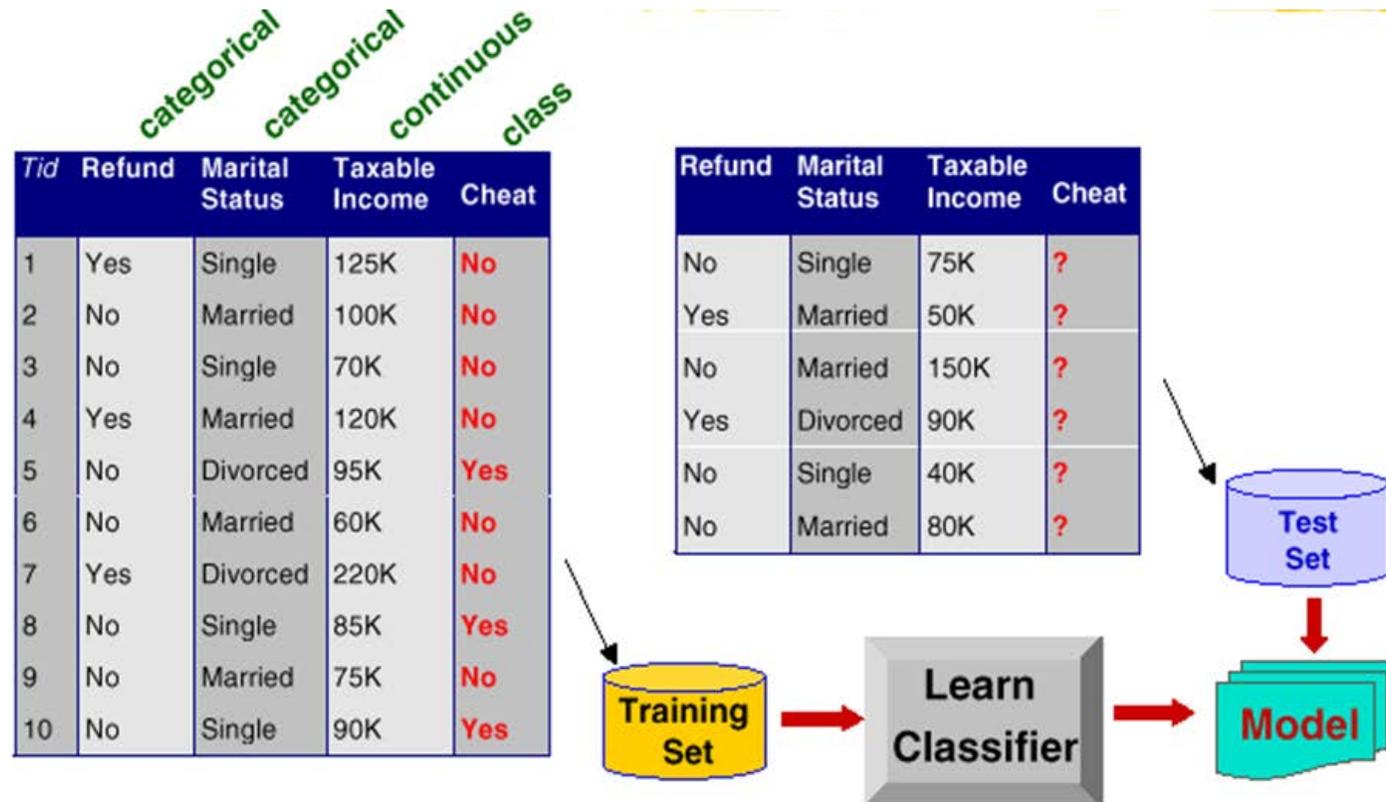
Credit: Paul Sakuma, associated Press

- NETFLIX, a US movie delivery company
- Competition: improve Netflix's ability to predict what movies users would like by 10%
- From \$5 million revenue in 1999 reached \$3.2 billion revenue in 2011 as a result of analytics
- Analyzing customer behavior and buying patterns created a recommendation engine optimizing both customer preferences and inventory
- Predictive Analytics: Collaborative Filtering

# Classification Example

- Data defined in terms of attributes, one of which is the class
- Find a model for class attribute as a function of the values of other(predictor) attributes, such that previously unseen records can be assigned a class as accurately as possible
- Training Data: used to build the model
- Test data: used to validate the model (determine accuracy of the model)
- Given data is usually divided into training and test sets.

# Classification Example



# Risk Analysis

---

If you have several risks at once, are their impacts additive?

Can we integrate qualitative and quantitative methods to see the issue from a joint perspective?

How important is the qualitative information about a process and how it fits in a quantitative analysis?

How can you use historical data to predict future behavior?

# Risk Modelling in Other Industries?

---

- Marketing – Uplift modelling, churn
- Finance - Credit Card
- Insurance – Fraudulent claims
- Transportation - Accidents
- Utilities – Blackouts/Events
- Hospitality - Airbnb

# Transportation



data.  
insight.  
discovery.

# Electric Power Transmission PMU Monitoring

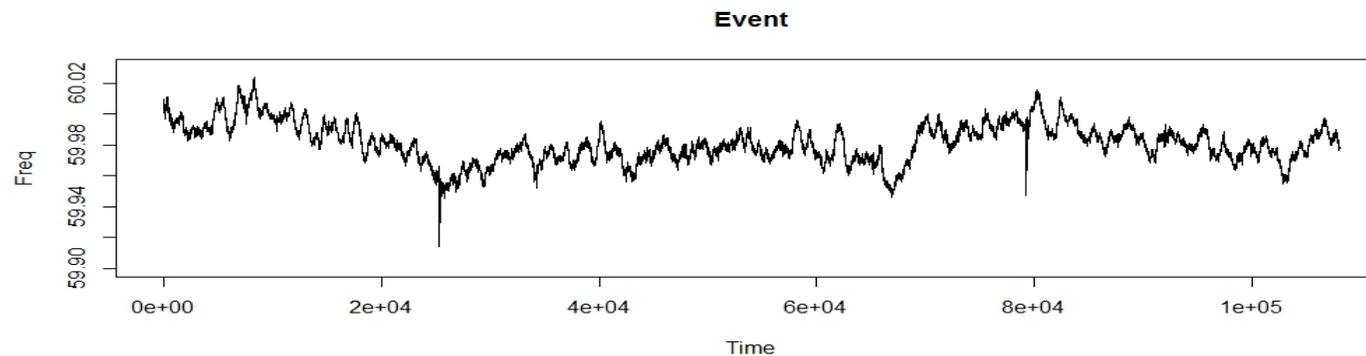
---

Power outages tripled from the mid to late 2000s – publics' tolerance for outages is zero

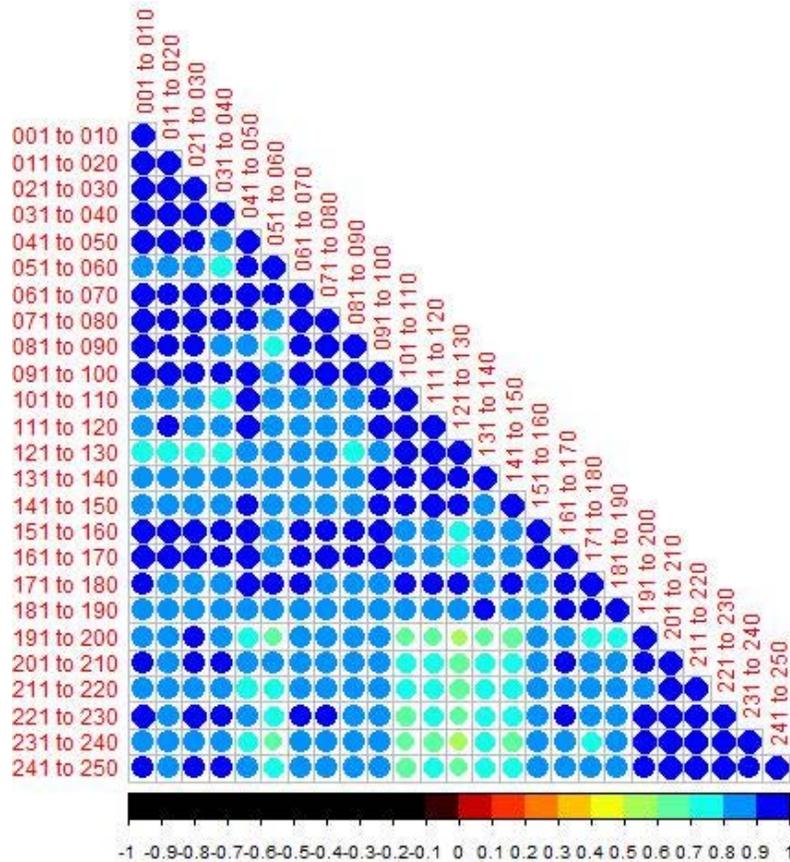
Challenges of scale and interconnectedness of transmission grid

Phasor measurement units (PMUs) provide a measure of system stability

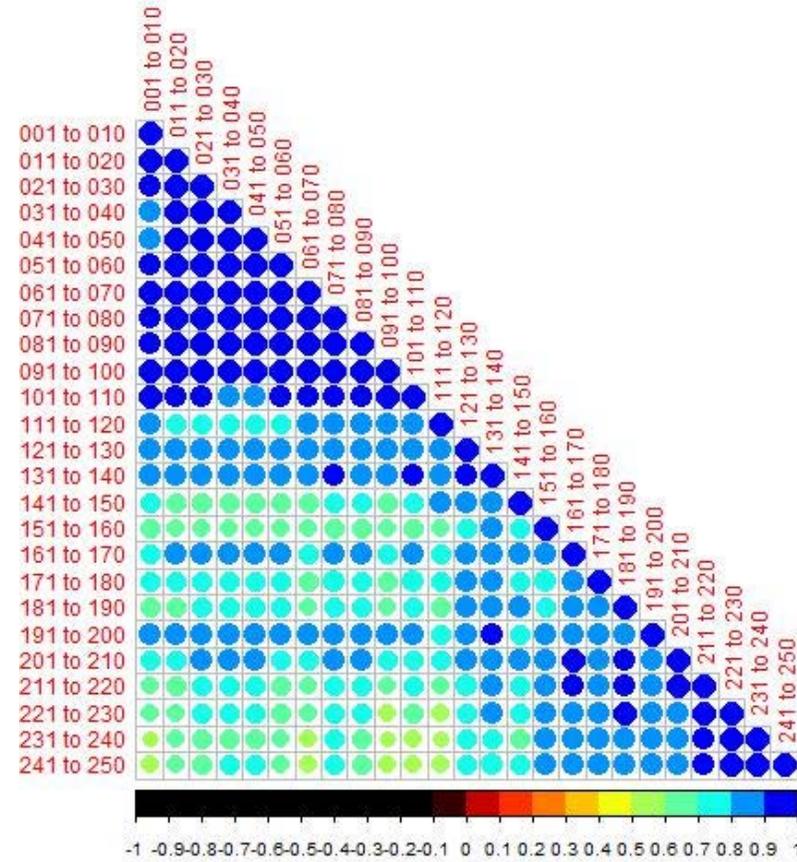
Predictive analytics provide tools to identify trends towards instability



# FFT Frequency Correlation Matrix



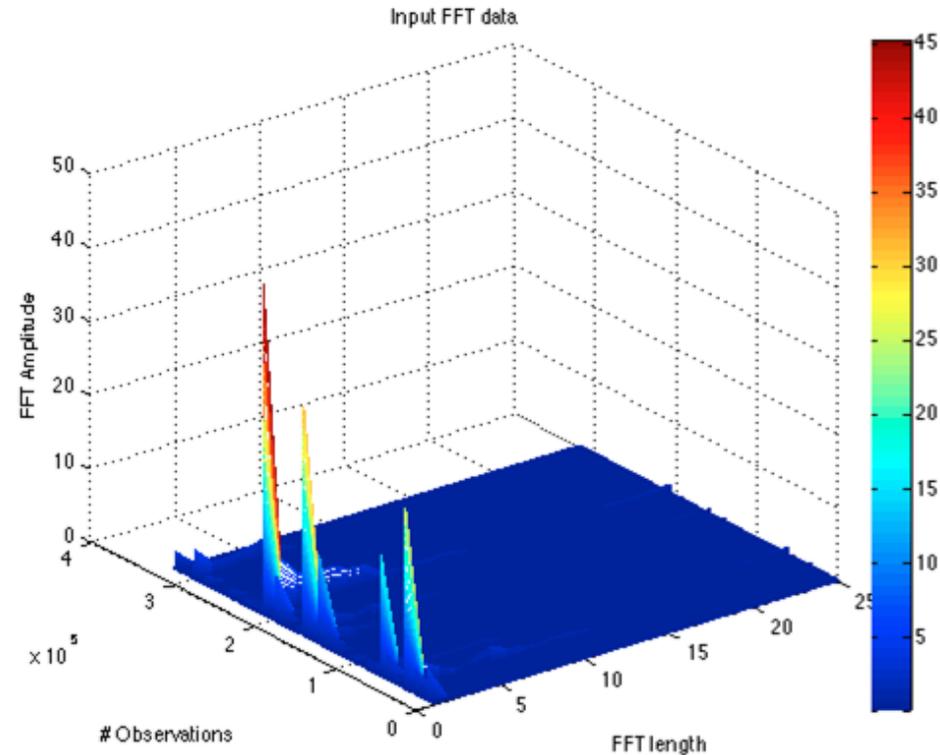
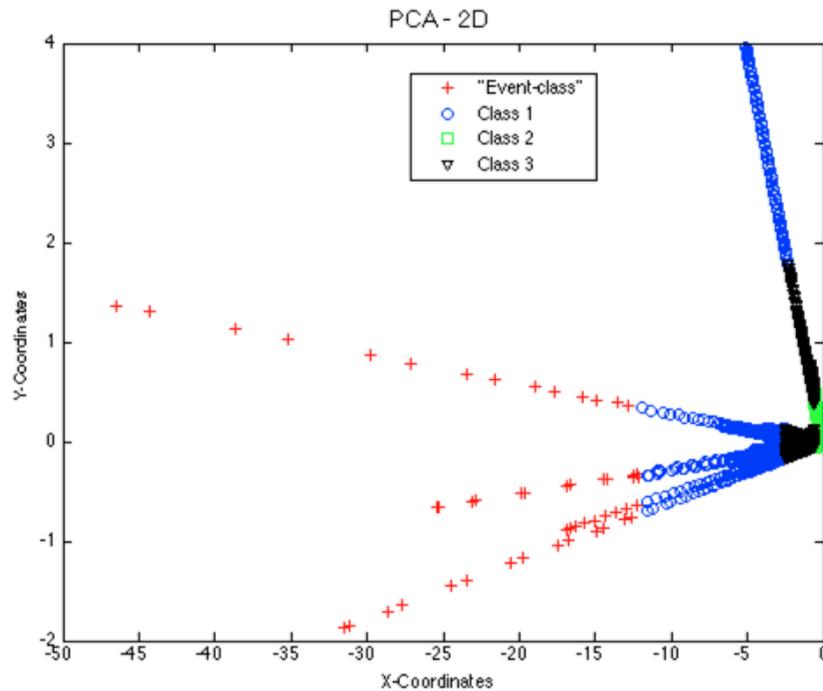
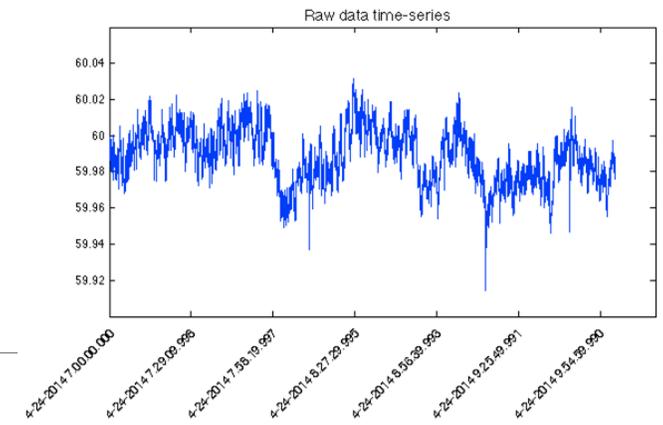
Normal Operations



Event – Power Failure

# 2D/3D Plots of FFT Clustering

- Detection of outliers at distant coordinates

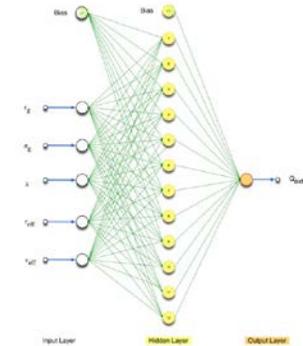
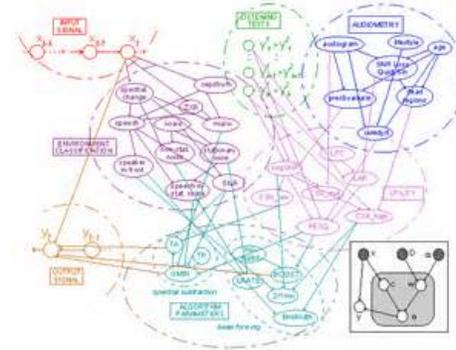


Cluster FFT slides on the direction of the "#Observations" - axis

# Potential Applications - Risk Assessment of Stress Corrosion Cracking

Combine diverse data sources

- Sensor-networks
- Databases
- Expert information
- Inspection data
- Diverse assessment methods



Utilize Machine Learning methods

Feature reduction and selection - parameters affecting pipeline failure, quantify uncertain data

Quantify expert knowledge and combine with statistical data

Predicts events by including probabilistic distributions of data

Ensemble Approach -multiple risk modelling methods combined in one unified method

# Possible Applications – Interacting Threats

---

Historical methods have relied on experts to define potential interactions and apply index-based modeling approaches

GRI considered interactions using combinations of threats and probabilities of failures

Predictive analytics can be used to identify relationships among parameters that increase risk and lead to failure

- Example: Weather and ground movement – what are the conditions that contribute to increased ground movement
  - Pipeline operators core business is not monitoring weather
  - Aggregate precipitation, soils, and location data to identify potential of increased risk of ground movement
- Example: Weather and loss of cover on pipeline water crossings
  - Recognize that it is impossible to sample during periods of high flow – sensor data set
  - Use precipitation, river levels and velocity, location data among others to identify periods of increased risk of loss of cover



**KEEP  
CALM  
AND  
ANALYZE  
BIG DATA**



**KEEP  
CALM  
AND  
USE DATA  
WISELY**

data.  
insight.  
discovery.



# Thank you!

# Questions?

Natasha Balac

[natasha@datainsightdiscovery.com](mailto:natasha@datainsightdiscovery.com)