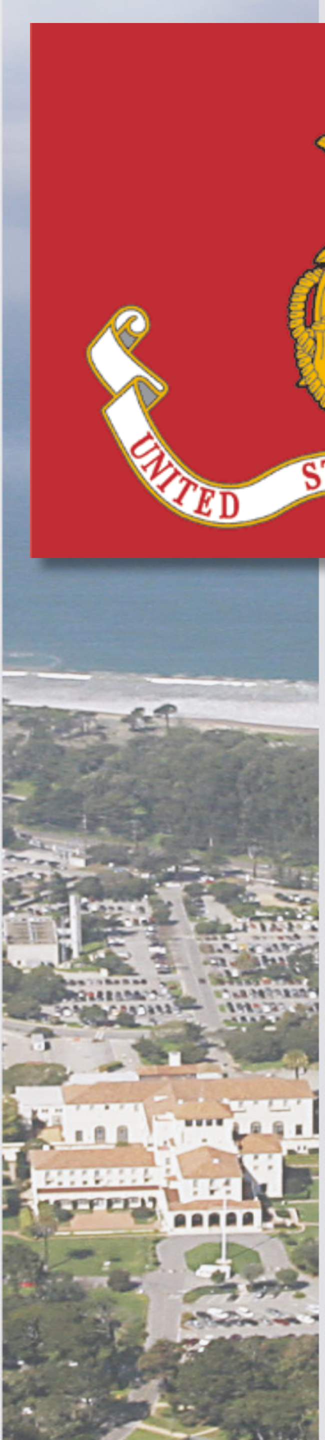


NAVAL POSTGRADUATE SCHOOL

Data Science Overview for Marine Leaders

Dr Marcus Stefanou

Naval Postgraduate School

msstefan@nps.edu 831-656-3316

9 May 17

Breckenridge Auditorium

Marine Corps Base Quantico, VA

Monterey, California

WWW.NPS.EDU

BLUF

- “Data science” is more than a buzzword – it is a way of doing business that can help you accomplish your mission more effectively and efficiently

- As a leader, you need to think about how you can grow a data science capability within your organization

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science

- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context

- Big data
- Cloud Computing

1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties

- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science
- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context
- Big data
- Cloud Computing

1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties
- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

Introduction

- **Your Background**
  - Commands represented?
  - What brought you here today?
  - What are your particular concerns?
- **My Background**
  - Assistant Professor of Computer Science at NPS
  - NPS Data Science Certificate Program Manager
  - Service History
    - 1990-2003 USMC (2nd Lt – Maj) – Aircraft maintenance, logistics information system acquisition, and Maritime Prepositioning Force
    - 2003 – 2016 USAF (Maj – Col) – Infrared sensor research and development, space-based imagery intelligence and strategic missile warning systems engineering
  - Education
    - BS Electrical Engineering, U. S. Naval Academy, 1990
    - MA Management, Webster University, 1994
    - MS Electrical Engineering, Naval Postgraduate School, 1997
    - MA International Relations, Tufts University, 2001
    - PhD Imaging Science, Rochester Institute of Technology, 2008

NPS Distinguishing Characteristics (in Data Science)

NAVAL POSTGRADUATE SCHOOL

- **Multidisciplinary graduate education**
  - Professional certifications to MS and PhD degrees
  - Resident and distance learning programs
  - Multi-service, interagency, coalition environment
- **Graduate research in all aspects of the data science**
  - Faculty expertise and interdisciplinary research projects in all areas of the data science process
- **Geographic advantages and synergies**
  - Academic institutions: UC Berkeley, Stanford
  - Government organizations: Defense Manpower Data Center, Defense Innovation Unit Experimental Industry: Silicon Valley information technology, data, and web innovation thought leaders
- **Robust research network infrastructure**
  - *.edu domain allows experimentation with emerging technologies and tools
  - Classified information processing network infrastructure

DOD needs an educated workforce that can deliver solutions informed by data science discipline

NPS Data Science Education

- **Operations Research Master's Degree Data Analytics Track**
  - Established 2015, expanded to be available to all Operations Research students in 2017
  - 7 graduates/year; focus is on consulting skills
  - Meet operational needs for formally educated (Master's degree) ops analysts
- **Data Science Certificate**
  - Provide education in the use of data science methods to gain insights from large, complex data sets
  - 4-course, 1-year distance learning sequence drawn from operations research and computer science curricula
  - First cohort of 19 National Reconnaissance Office employees graduate Sep 2017
- **USMC Analyst Community-of-Interest Short Course**
  - 11-14 Jul 17, Quantico, VA
  - Sponsored by Operations Analysis Directorate, CD&I (Mr Al Sawyers)
  - Purpose: Convey understanding of fundamental concepts, knowledge of key terms, and the ability to apply data science tools to real-world problems
  - Format: 4 days of hands-on data science for 25 students in the Marine Professional Analyst Community of Interest

Today's Purpose

- Motivate the need for data science in the Marine Corps
- Define data science
- Provide context for concepts such as “big data” and “cloud computing”
- Convey the essential elements of an organizational data science capability

Rules of Engagement for Today

- Interactive and engaging
  - I will lecture, share my insights – but I don’t have all the answers
  - I welcome your input and perspectives!
- Diverse audience from many functional areas
  - We need to make sure common understanding is established
  - Ask if something isn’t clear!
- Stay on schedule
  - Lots of material to cover
  - Afternoon session
    - “Parking lot” for questions taking too long to discuss this morning
    - Opportunity for me to listen closely to specific use cases for data science

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science
- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context
- Big data
- Cloud Computing

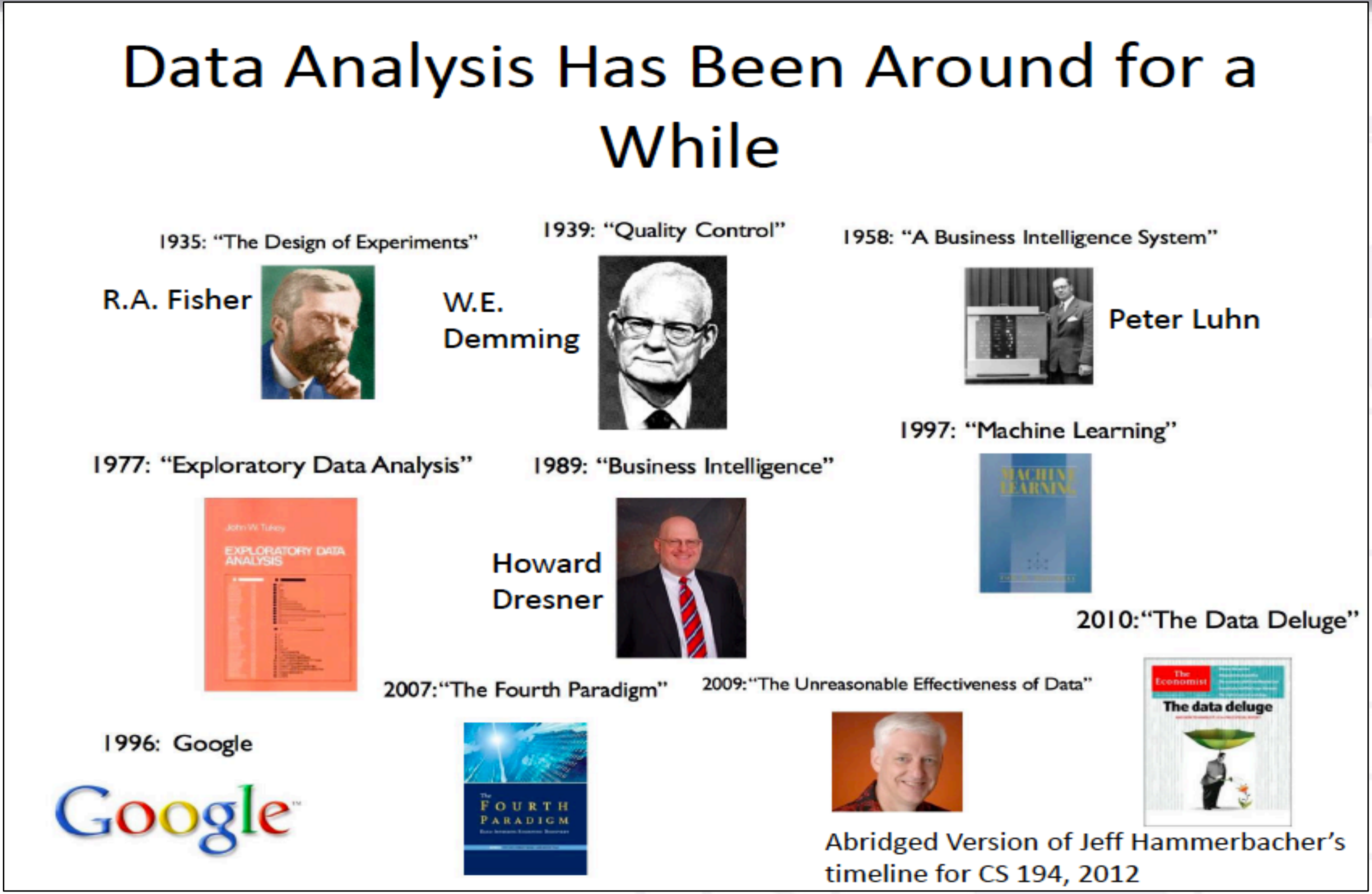
1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties
- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

There's nothing new here...



Canny, CS194

…or is there?

- We’ve been gaining insight from data for a long time, but something is fundamentally different now
- Our data now exists largely in digital form:
  - Internet-based businesses
  - Social and traditional media
  - Businesses are translating data to machine readable form
- There is a lot of data:
  - Explosion of digital sensor data about us, our environment, and our behaviors
- New technology has improved our ability to glean useful information from data:
  - Increasing compute power – allow algorithms to work with larger data sets
  - Cheap storage – continues to get cheaper
  - Robust communications infrastructure – web scale growth
  - Distributed computing – broad access to “elastic” computing capability

NAVAL POSTGRADUATE SCHOOL

What is different about data science?

Essential Element	Traditional Analysis	Data Science
Data Infrastructure	• Manage data locally • Use relational databases • Get a bigger server	• Cloud storage • Use unstructured data • Distributed computing
Statistical Modeling	• Optimize performance of best model	• Flexibility to try lots of approaches
Visualization	• Graphs, charts	• Interactive/immersive experience
Software Engineering	• Waterfall -> start with requirements and end with delivery • Contract for someone to build the tool	• Agile development sprints • Lots of customer-developer interaction • Shorter delivery timelines • “Do it yourself” government capability

Data Science is not just doing traditional analysis better, using more data, and faster – it sits the data scientist next to the decision maker to solve problems rapidly!

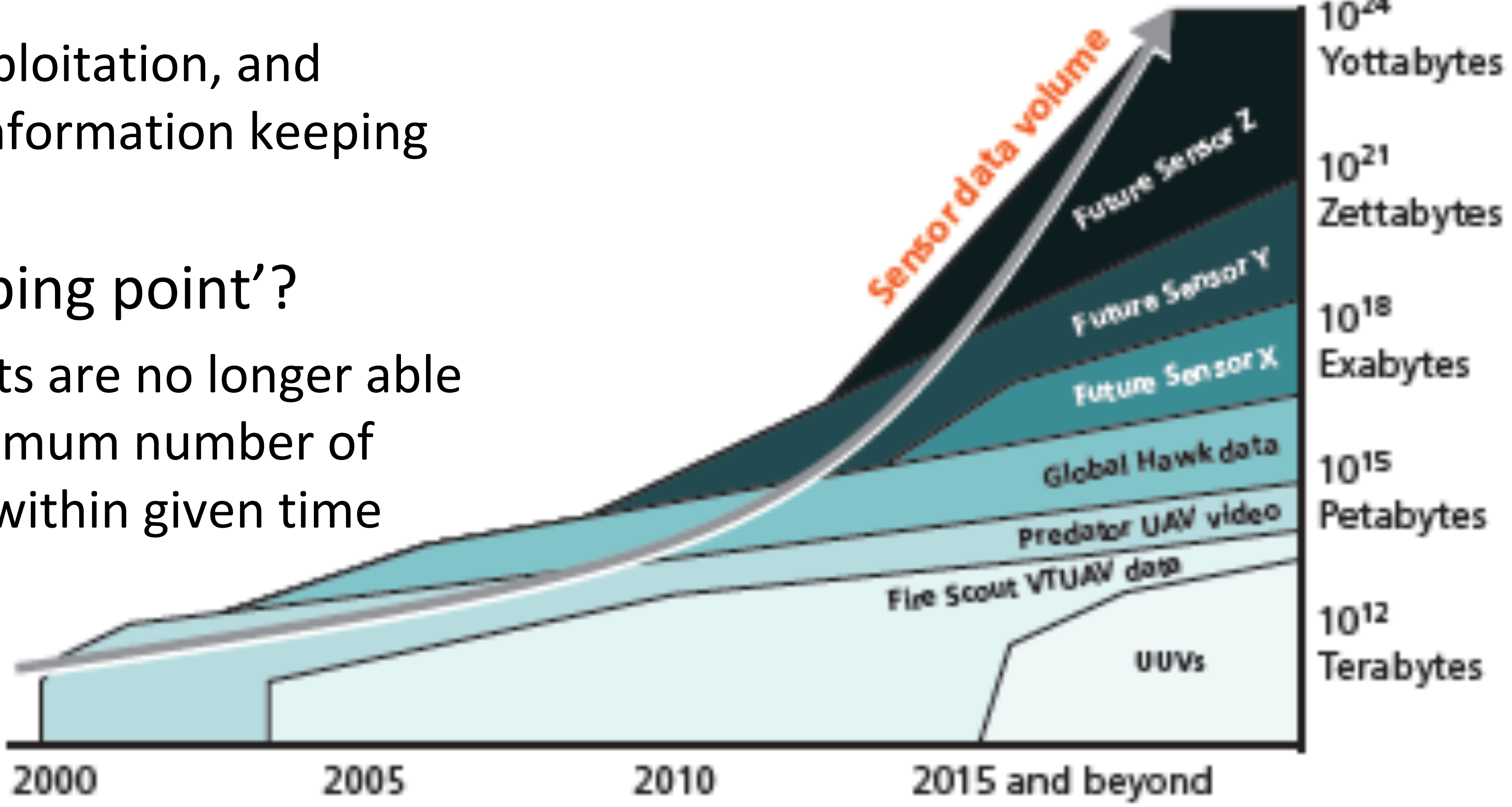
Data Science Hype Started with the Web-based Economy…

- How is DOD different than Google, Netflix, Amazon, etc?
  - Missions
  - End products
  - Business models
- Consider further:
  - Types of decisions needed
  - Speed of decision making
  - How we process data
  - How we think about data at all levels of leadership
  - How our adversaries will use data against us
  - Security implications of aggregation of large amounts of unclassified data

…but the DOD Needs Data Science Competence

- More sensors is great
  - Are processing, exploitation, and dissemination of information keeping up?
- Is there an ISR ‘tipping point’?
  - Intelligence analysts are no longer able to complete a minimum number of exploitation tasks within given time constraints

1 kilobyte KB =	10^3 bytes (B)
1 megabyte MB =	10^6 B
1 gigabyte GB =	10^9 B
1 terabyte TB =	10^{12} B
1 petabyte PB =	10^{15} B
1 exabyte EB =	10^{18} B
1 zettabyte ZB =	10^{21} B
1 yottabyte YB =	10^{24} B



SOURCE: Programs, Management, Analytics & Technologies, 2011.
NOTES: UAV = unmanned aerial vehicle; UUV = unmanned undersea vehicle; VTUAV = vertical takeoff and landing tactical unmanned aerial vehicle.

RAND MG315-1.2

Our Bosses are Asking for Help in Making Better Decisions

- **Intelligence**
  - Automated workflows
  - Predictive analytics
- **Operations**
  - Readiness effectiveness
  - Adaptive tactical decision aids
- **Network Defense/Cybersecurity**
  - Anomaly detection
  - Network traffic pattern recognition
- **Personnel Readiness**
  - Insights to recruit, train, and retain the best people
  - Training effectiveness
- **Healthcare**
  - Forecast patient health indicators for appropriate treatments
- **Planning, programing, and budgeting**
  - Budget creation and force optimization
  - Link concepts and illuminate trade spaces
  - Leverage simulations to analyze program effectiveness
  - Discover future requirements
- **Logistics**
  - Supply chain effectiveness
  - Predictive and diagnostic maintenance
  - Sparing
- **Systems Acquisition**
  - Analysis of alternatives and requirements analysis
  - Design trade studies
  - Program life cycle cost estimation
  - Test and evaluation for operational effectiveness

"Across the DoD it is clear, whether it's robotics, cyberdefense, biotech, or hypersonic engines, that data science, a cornerstone of operations research, is a critical influencer."

- Lt Gen R. S. Walsh, USMC
CG MCCDC and Deputy Commandant, CD&I
2016 MORSS address

What is "Data Science?"

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it.

- Hal Varian, Chief Economist @ Google

Making data tell its story.

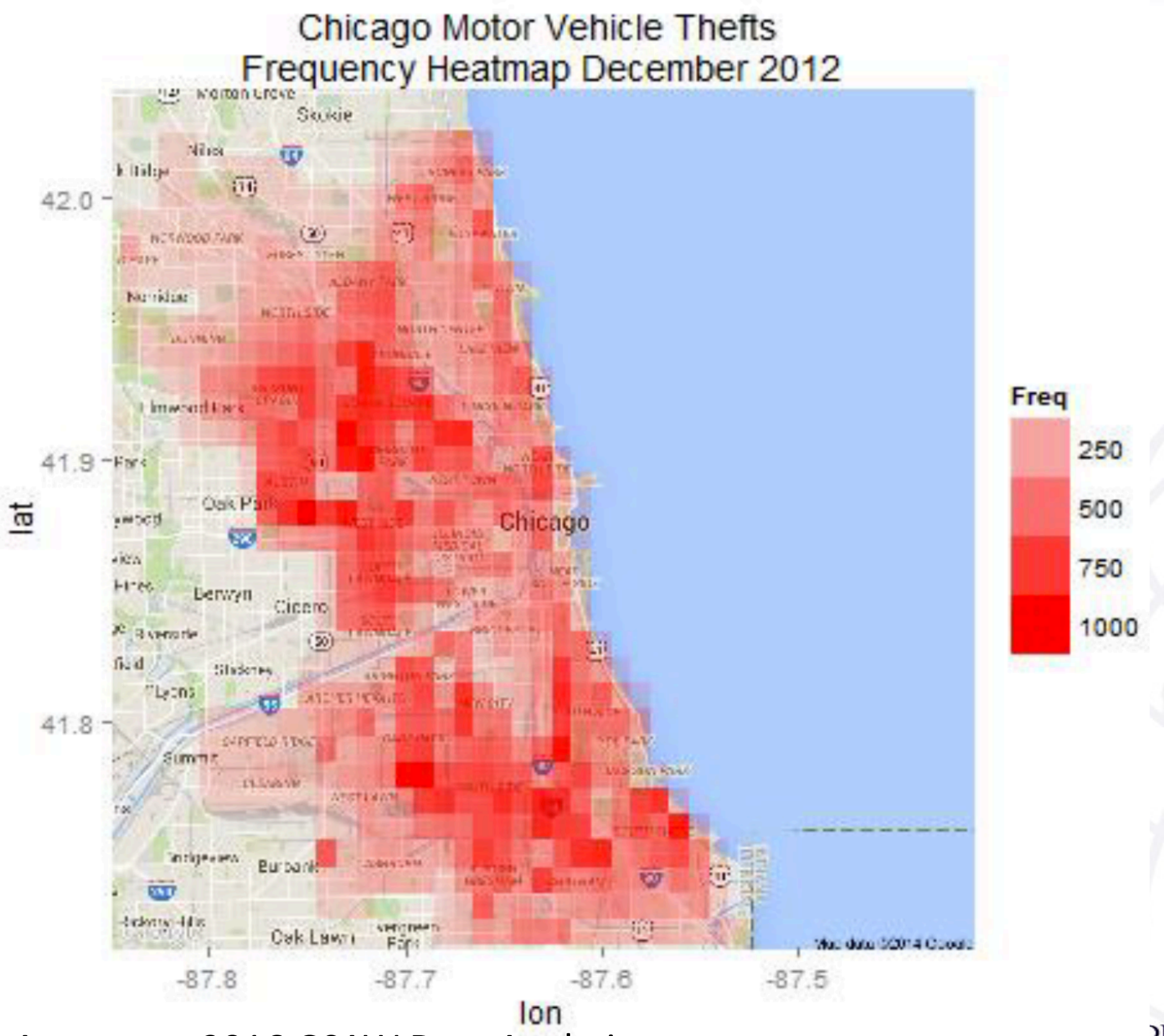
- Mike Loukides

The ability to extract knowledge and insights from large and complex data sets.

- DJ Patil, First Chief Data Scientist for US Government

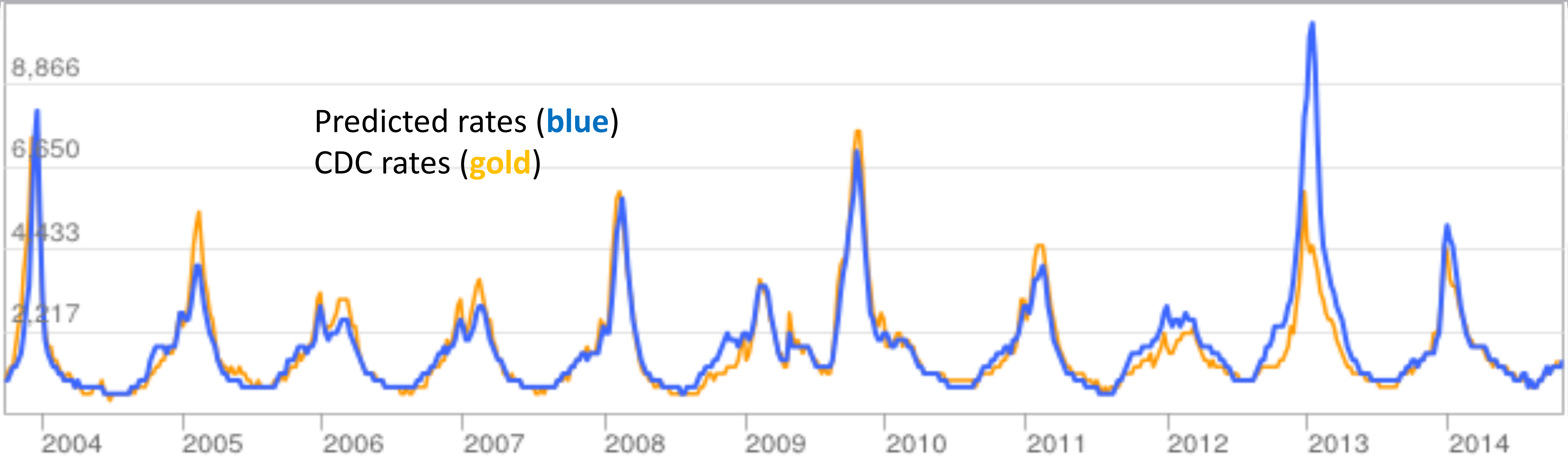
Data science is about ***organizing information*** in such a way that we can **use *models* to understand**, which often requires ***visualization* of model outputs** so that that we can **glean insights at scale**.

Understand Crime Patterns



- Use historical information to discern patterns
  - Location
  - Time of day
- Use patterns to predict future crime locations
- Provide a vector for rapidly focusing policing efforts

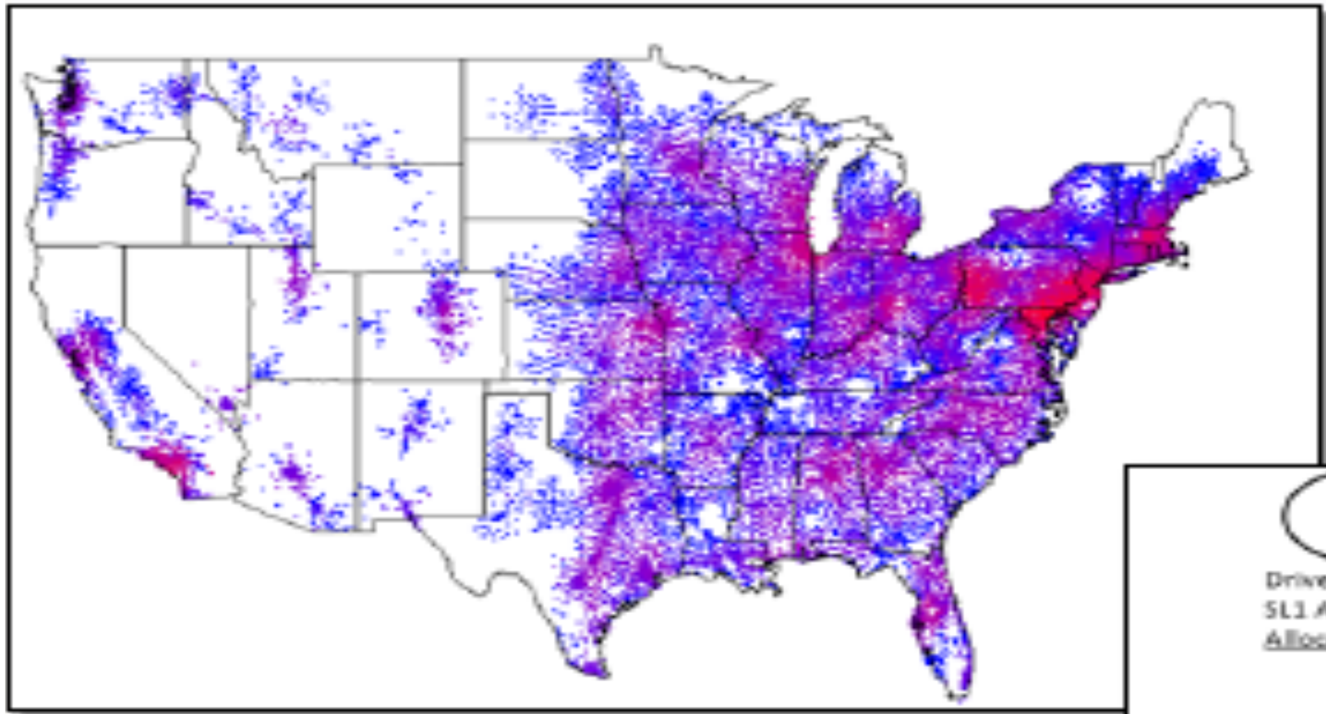
Predict Flu Trends

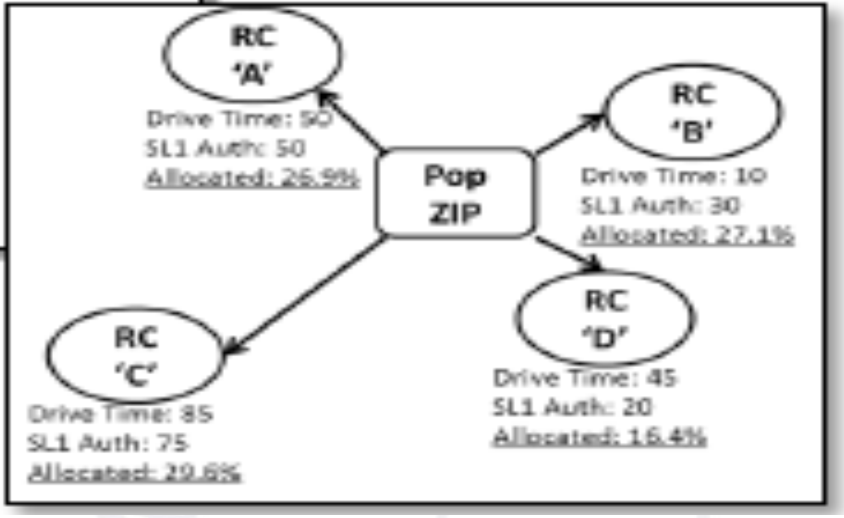


- Google Flu Trends (GFT) launched in 2008 as an effort to track flu incidents more quickly
- Idea: someone is exhibiting influenza-like symptoms if they are searching for terms relating to the flu and its symptoms
- Infers influenza incidents 2 weeks before actual Centers for Disease Control (CDC) reports

Whitaker, 2015 and CY3650 http://www.google.org/flutrends/us/

Army Reserve Stationing



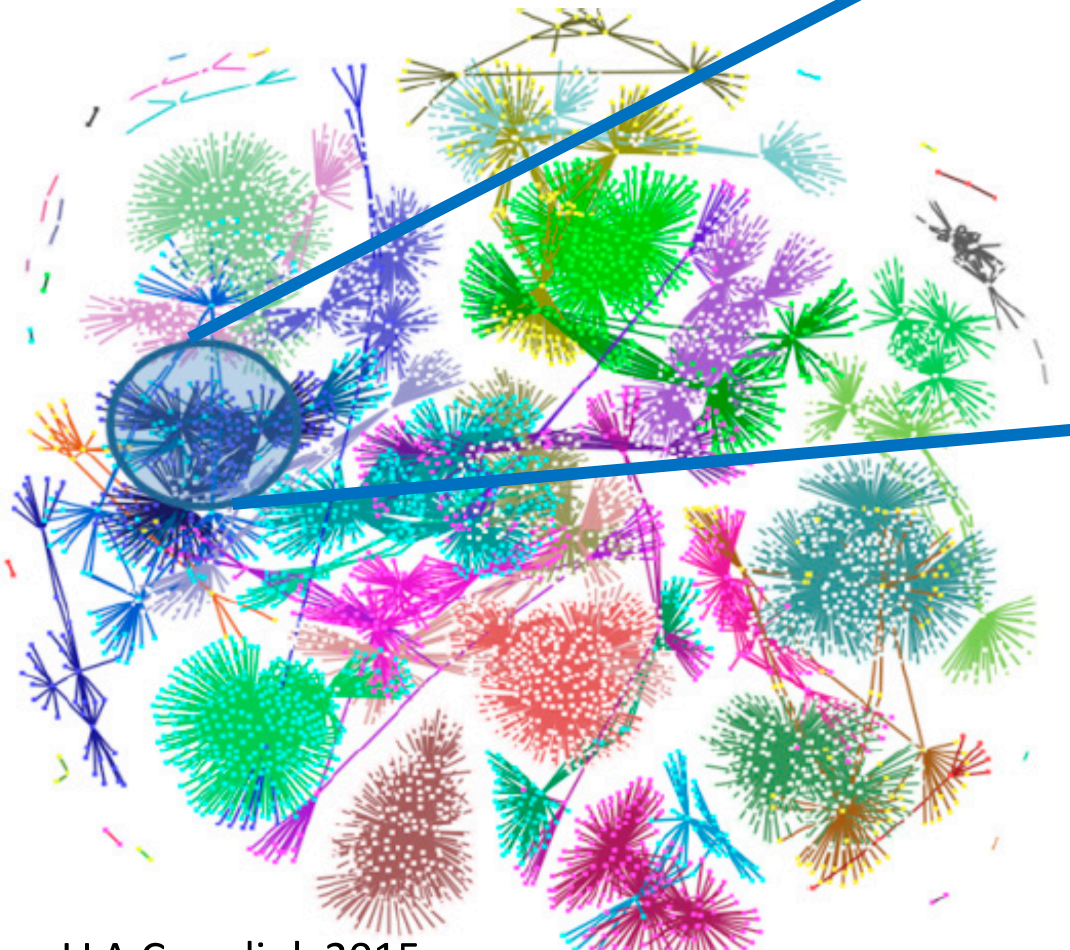


- Develop a predictive model using a potential location's recruiting market demographics
- Gives decision makers the ability to see the trade space between manning potential and other stationing criteria
- Correct placement of USAR units in sufficient recruiting markets

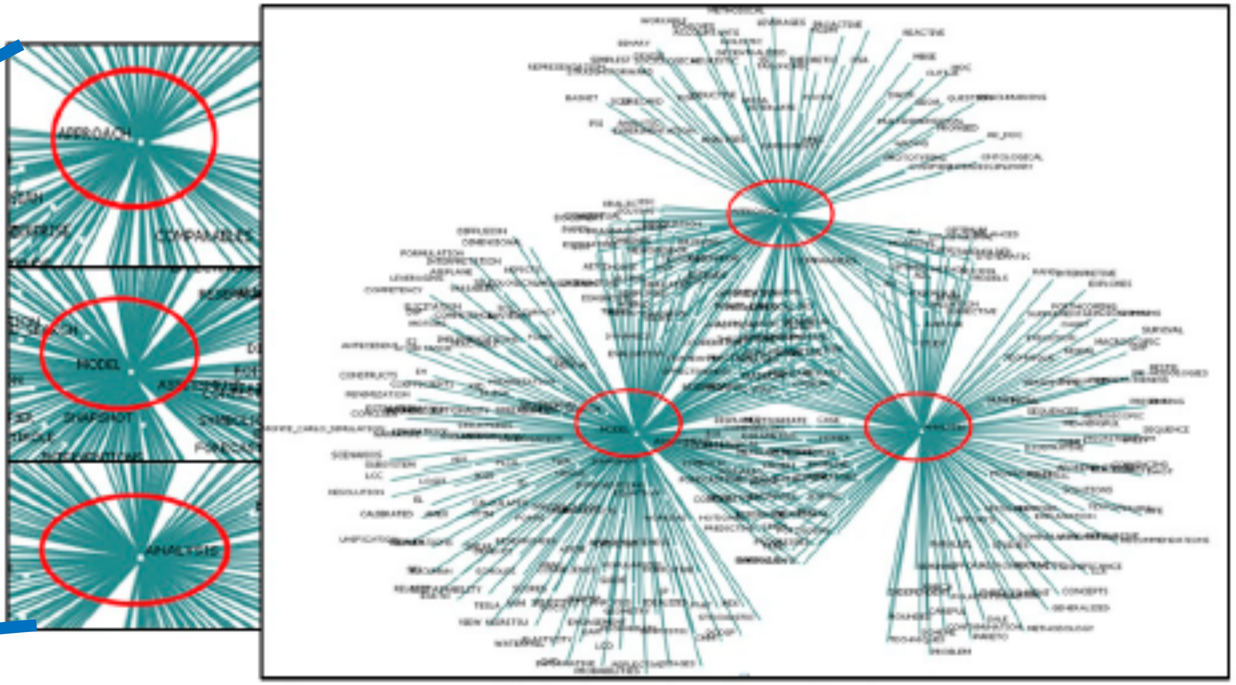
Acquisition Program Lexical Link Analysis

A complex system expressed as a word pair network for 450 acquisition programs and RDT&E budgets over 10 years. Compare:

- Urgent Need Statements with technologies
- Congressional budget documents with needs



Specific words and links in one theme of the network



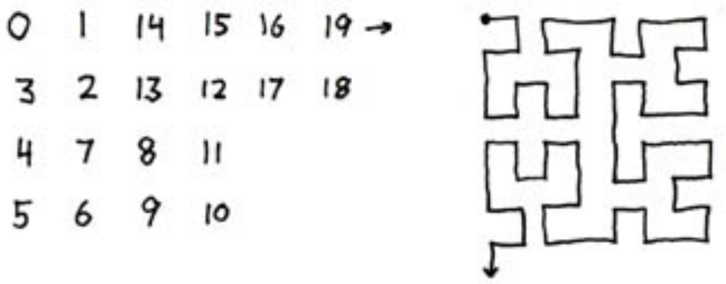
- Number of links from acquisition programs to warfighter requirements
  - Fewer links - received less budget reduction, cuts focused on large and expensive programs
  - More links - received more budget reduction, indicating a pattern of good practice of allocating resources to avoid overlapping efforts.

Zhao LLA Crosslink 2015

Internet Mapping

ip address space

Mapping the internet's "index" (ip addresses) from one dimension to two dimensions improves our understanding of internet "space".



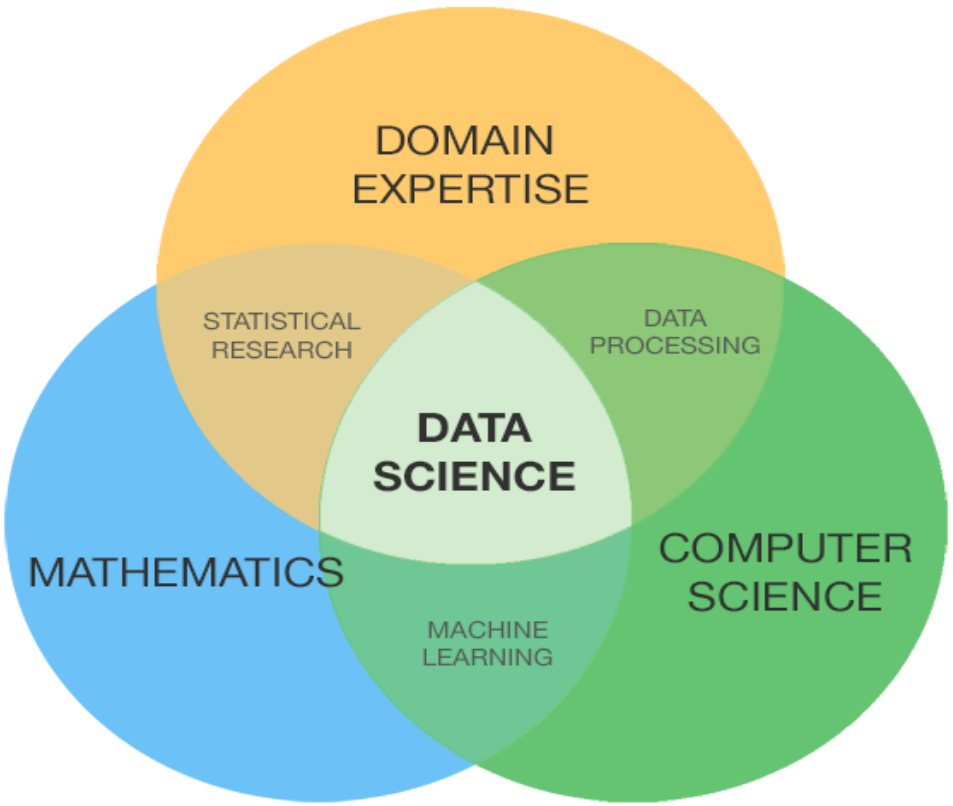
Hilbert Curve

Source: https://xkcd.com/195/

Data Science Requires Multiple Skillsets

Math and Statistics Competencies

- Statistical modeling
- Machine learning
- Bayesian inference
- Optimization
- Simulation
- Network science
- Model development



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Domain Expertise Competencies

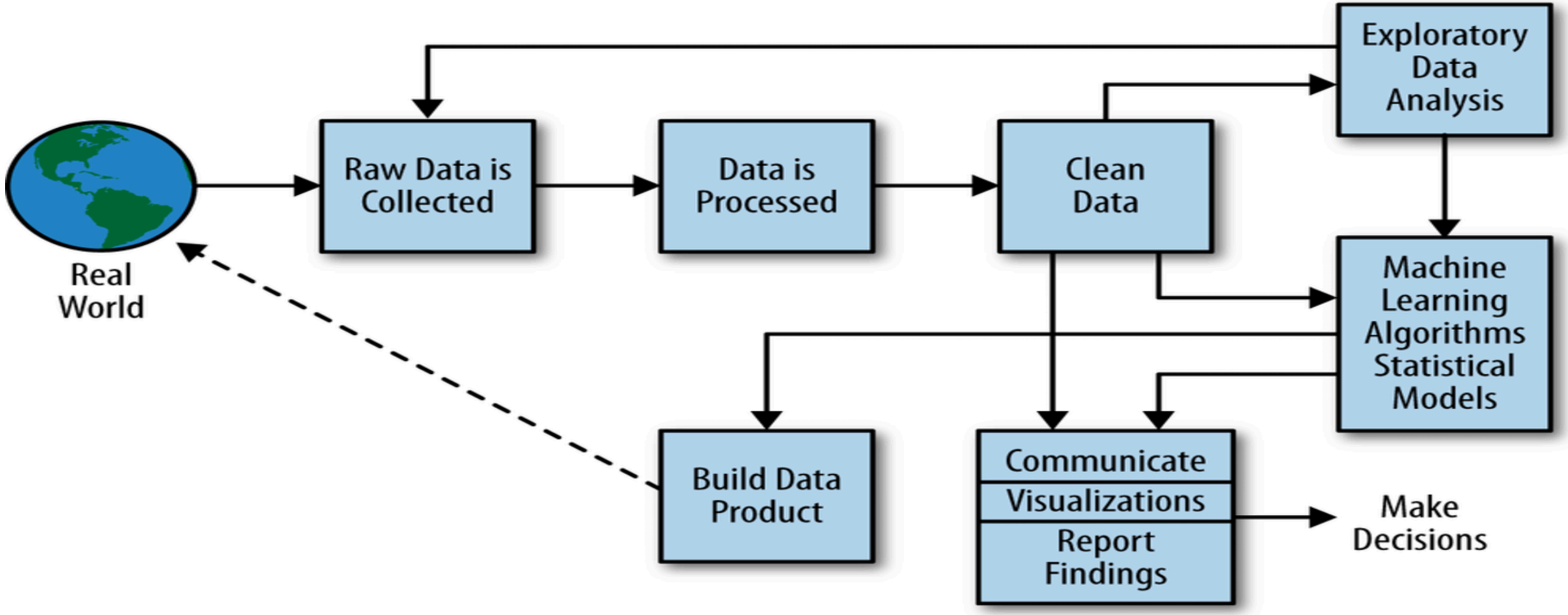
- Specific functional area
- Curious about data
- Influence with leaders
- Problem solver
- Creates narratives with data
- Visual design and communication
- Creative, innovative, and collaborative

Computer Science Competencies

- Scripting language (Python)
- Statistical computing package (R)
- Databases (SQL and NoSQL)
- Distributed storage (Hadoop Distributed File System)
- Distributed processing (MapReduce)
- Cloud computing (Amazon Web Services)
- Tool Development
- Data pipelines (Pig/Hive)

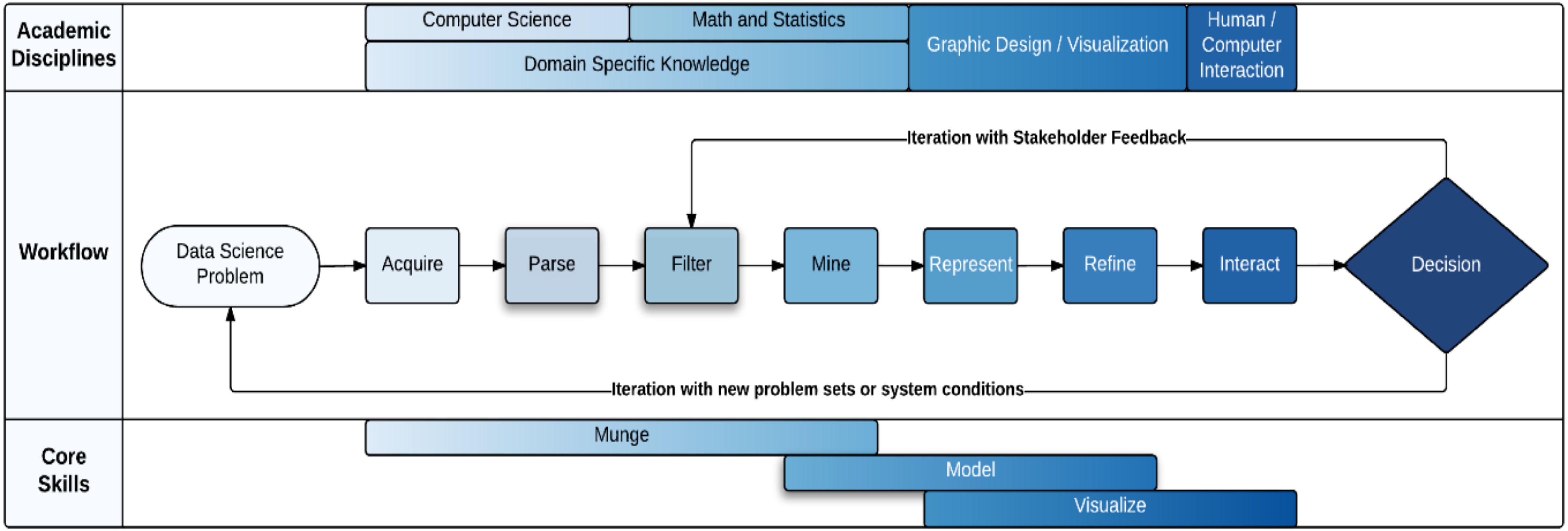
Data Science is a team sport

Data Science is a (Repeatable) Process



O'Neil and Schutt, Doing Data Science

Data Science is a (multi-disciplinary) Process



Huddleston, What is Data Science from B. Fry *Computational Information Design,* PhD Dissertation, 2004.

Data Science Process Detail

“Munge”
“Prepare”
“Exploratory Data Analysis”

“Model”
“Learn”
“Data Mine”

“Visualize”

- **Data Science Problem** – Formulate the question clearly
- **Acquire** – Obtain data
- **Parse** – Provide some structure around data meaning
- **Filter** – Remove all but the data of interest
- **Mine** – Apply methods to put data in mathematical context
- **Represent** – Determine a simple representation for the data
- **Refine** –Make it more visually engaging
- **Interact** – Add methods for manipulating the data
- **Decide** – Present findings to decision maker

Huddleston, What is Data Science from B. Fry *Computational Information Design,* PhD Dissertation, 2004.

Define the Question as a Data Science Problem

- Formulate the question in clear terms:
  - How much does sensor A contribute to force effectiveness?
  - Is there systematic bias in the data collected in my experiments?
  - Is there a connection between training schedules and training effectiveness?
  - Using maintenance logs, can I identify maintenance processes that are less effective?
  - How do I reduce my operating costs?
  - What is a “normal” demand pattern?
  - How is a particular activity related to geographic location?

Understand the Type of Question

Know (Descriptive: what happened?)

- Interactive drill down, queries
- Basic analytics and visualizations (descriptive statistics, time series, histogram, bar chart)
- Forensics, assessments, historical trends, alerts

Explain (Diagnostic: why did it happen?)

- Correlation between variables and outcomes
- Statistical analysis, data mining, classification, clustering
- Find similar items, find hubs in a graph, find frequent item sets

Predict (Predictive: what will happen? Prescriptive: what should we do?)

- Forecast/extrapolation
- Decision models, neural networks, supervised learning, optimization
- Weather forecast, translation, user profile, traffic flows

Different Functional Areas Have Different Questions

- Analytical/ Systems Acquisition Communities
  - Evaluate effectiveness of education, training, logistics, personnel, and business processes
  - Given changes affecting system performance (new or modified equipment, organization, or TTPs), how will these changes impact operational effectiveness?
- Intel Community – Understand enemy intentions, connect multiple sources of data

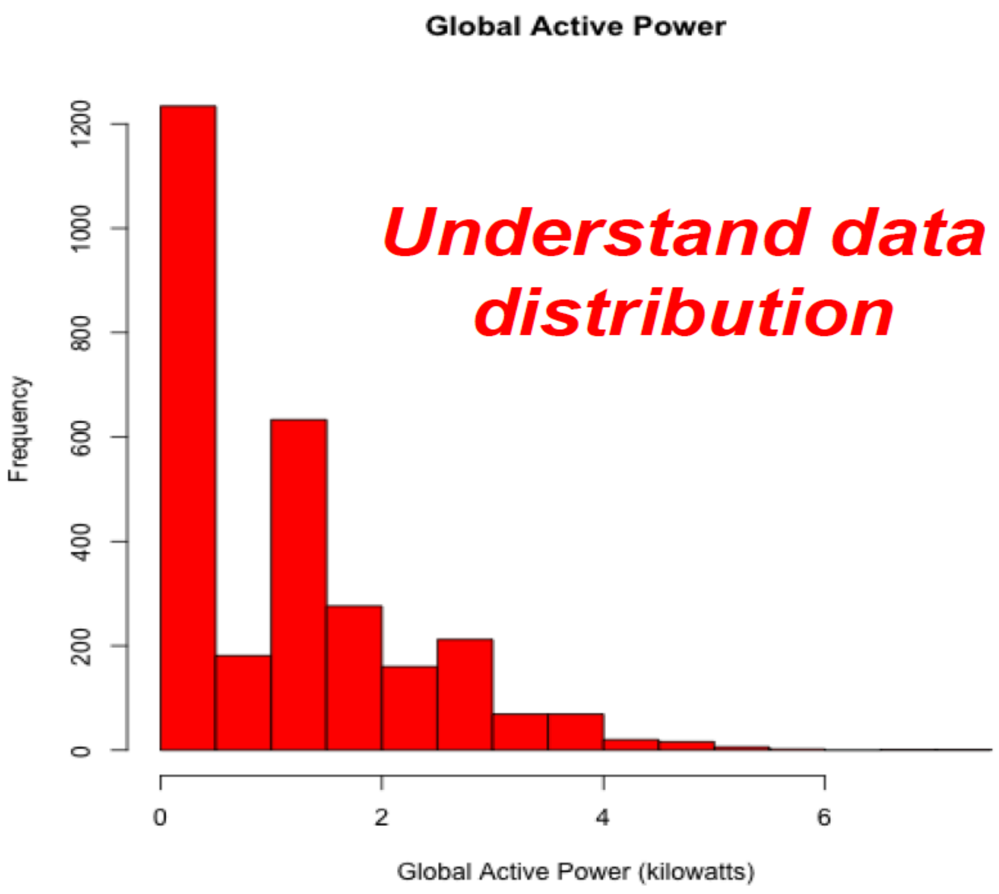
Elements	Intel	Analysis/Acquisition
Problem type	• Find the needle in the haystack	• Optimizations • Simulations
Data control	• Raw "in the wild" • Multi-source	• More control over data source (simulation-based)
Urgency	• Fleeting targets • Patterns of life	• Studies • Analyses of alternatives
Data volume	• Lots • Streaming	• Governed by amount of compute power to run more cases

Exploratory Data Analysis

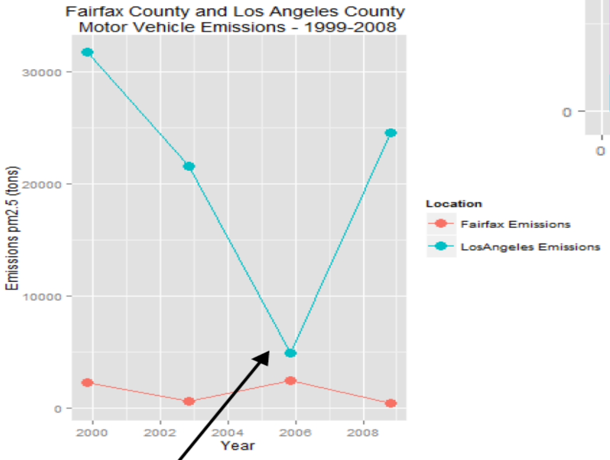
- Explore the data and define measurable goals
- Don’t worry about a hypothesis or model at this point
- Use plots, graphs, and summary statistics to gain an understanding and intuition about the data
- Sanity check of the data – distribution, range, scale, units, outliers, missing data, errors, interesting correlations between variables
- Forms the basis for cleaning and preparing the data

> Exploratory Data Analysis is an attitude, a state of flexibility, a willingness to look for things that we believe are not there as well as those things we believe are there.
> -John Tukey

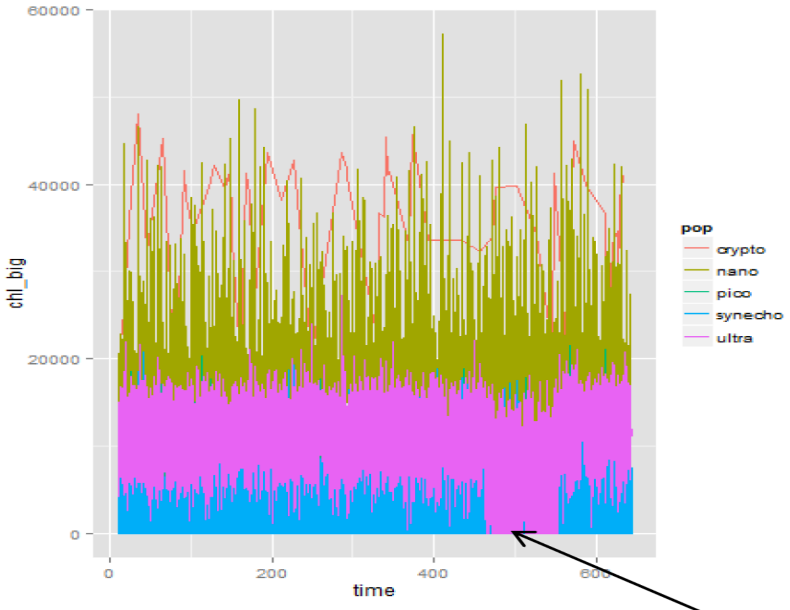
Exploratory Data Analysis



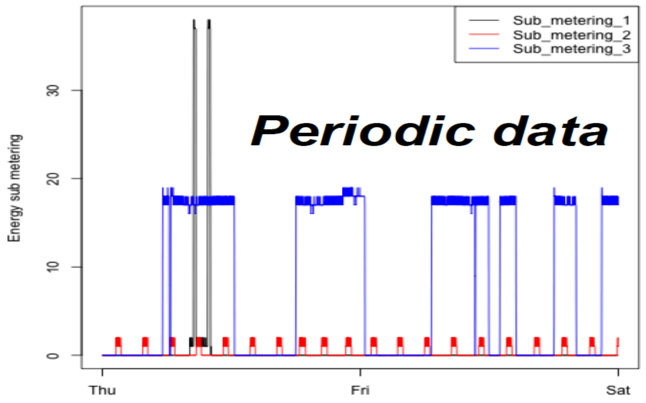
Understand data distribution



Identify outliers



Miscalibrated Data

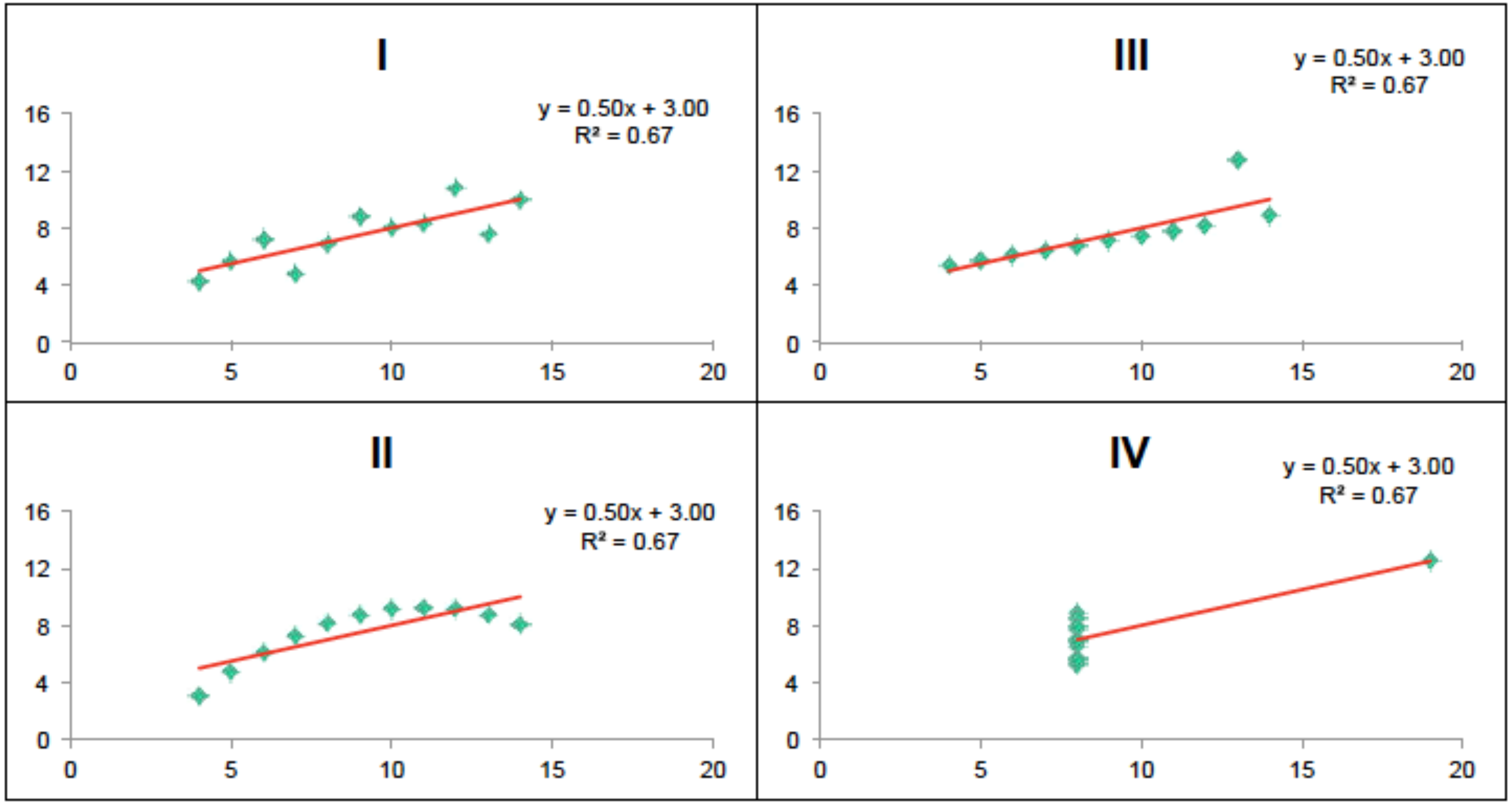


Periodic data

Know your Data – Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Using line plots reveals the differences among the data sets



This quartet is used as an example of the importance of *looking* at your data before analyzing it in Edward Tufte's book, *The Visual Display of Quantitative Information.*

Matos PHALANX 2015

Data Preparation

Missing data

State	Year	Rasmussen	SurveyUSA	DiffCount	PropR	Republican
Alabama	2004	11	18	5	1	1
Alabama	2008	21	25	5	1	1
Alaska	2004			1	1	1
Alaska	2008	16		6	1	1
Arizona	2004	5	15	8	1	1
Arizona	2008	5		9	1	1
Arizona	2012	8		4	0.833333	1
Arkansas	2004	7	5	8	1	1
Arkansas	2008	10		5	1	1
Arkansas	2012			2	1	1
California	2004	-11	-11	-8	0	0
California	2008	-27	-24	-5	0	0
California	2012		-14	-6	0	0
Colorado	2004	5	3	9	1	1
Colorado	2008	-4		-15	0	0
Colorado	2012	3	-2	-5	0.307692	0
Connecticut	2004			-3	0	0
Connecticut	2008	-17	-16	-4	0	0
Connecticut	2012	-7	-13	-8	0	0
Delaware	2004			-2	0	0
Delaware	2008	-15	-30	-4	0	0
Florida	2004	3	1	0	0.5	1
Florida	2008	1	-3	-13	0.157895	0
Florida	2012	2	0	6	0.666667	0
Georgia	2004		12	4	1	1
Georgia	2008	5	7	9	1	1
Georgia	2012		8	4	1	1
Hawaii	2004			2	0.75	0
Hawaii	2008	-41		-1	0	0
Hawaii	2012			-2	0	0
Idaho	2004			1	1	1
Idaho	2008	39		1	1	1
Idaho	2012			1	1	1
Illinois	2004	-11	-12	-5	0	0

- Identify missing or anomalous entries
- Tabulate categories, dates
- Put rare entries together into “Other”
- “bin” numerics into several big groups
- Detect and remove redundant columns
- Deal with numerous formats
- Understand what the data can/can’t tell you
- Transform data (rescale or normalize)

Build a Model to Answer the Question

- How do you know which model to use?
  - Type of problem – classification, scoring, clustering
  - Type of data – numeric, categorical
  - Number of variables – a few or many
  - Variable and outcome interaction – linear, non-linear, correlated inputs
  - Targets/outcomes – known or unknown
  - Desired level of interpretability - black box or intuitive understanding
  - Quality of answer – quick and dirty or detailed validation
  - Ground truth – plentiful or nonexistent
- How do you implement the model?
- How do you evaluate the "goodness" of the model?

All models are wrong but some are useful.
-George E. P. Box

Basics Structure of Models

- The relationship between a variable of interest, y, and several observed attributes, x's, so that in the future, given new x's, without a y, we can predict y
- To start, we might think of the y's as being generated by a model like this:

$$y = f(X_1, \ldots, X_p) + \varepsilon$$

A function of the attributes which we might know something about, but not everything.

Random "error" or noise from all the stuff we can't measure

- Our job is to estimate the structural or systematic part of the right hand side of the equation:

Predicted or fitted y

$$\hat{y} = \hat{f}(x_{new,1}, \ldots, x_{new,p})$$

Estimated, fitted, learned, trained function

Whitaker, short course 2015

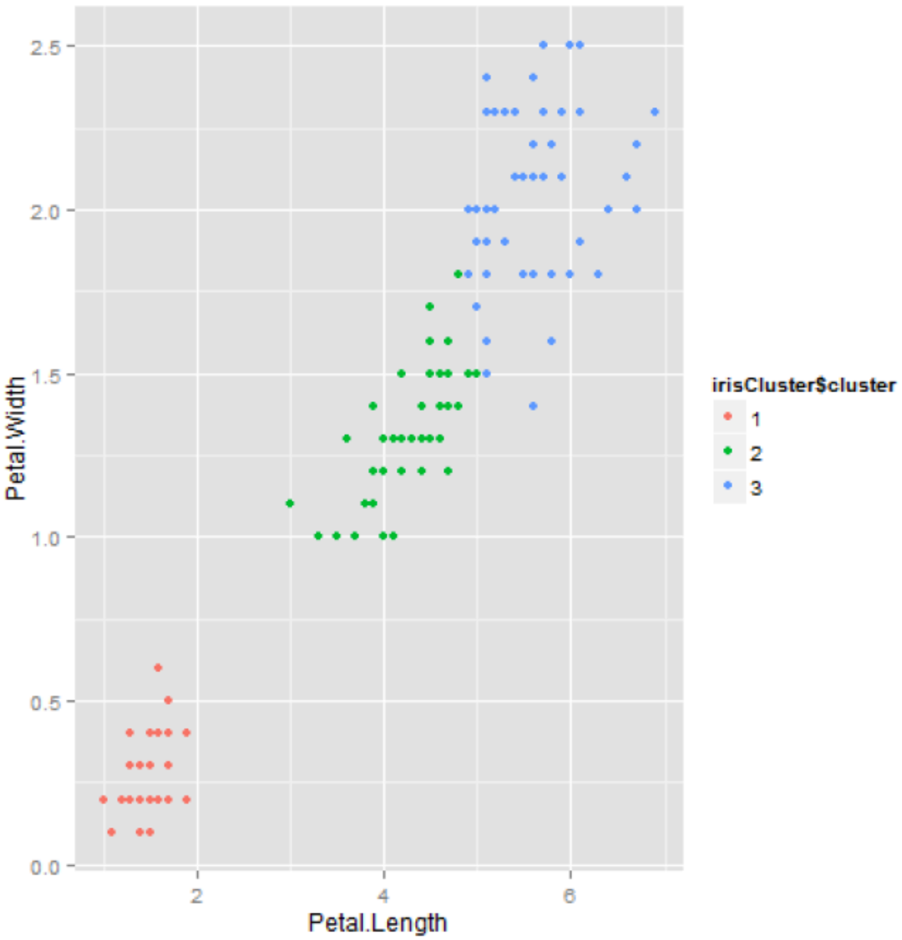
Approaches

- We are basically trying to do one of three things:
  - Understand what is happening
  - Predict what will happen
  - Find some pattern of interest
- Unsupervised Techniques
  - No truth observations are available for labeling the data
  - All we have are the observations (the x's) but we don't have corresponding y's
  - Algorithm needs to find the answer in the data
  - Techniques group similar things based on common features in the data – "Clustering"
- Supervised Techniques
  - Assumed that a 'training set' of ground truth or labeled data exists
  - This is used to train your algorithm to recognize or "label" data
  - We observe y and x's for the training data but for future data, we will want to predict y from the x's
  - Techniques: linear regression, logistic regression, classification trees, support vector machines, neural networks

Whitaker, short course 2015

Unsupervised Learning

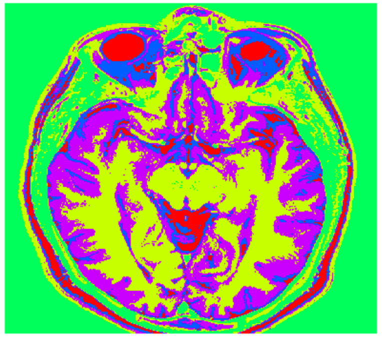
Clusters of data using two attributes



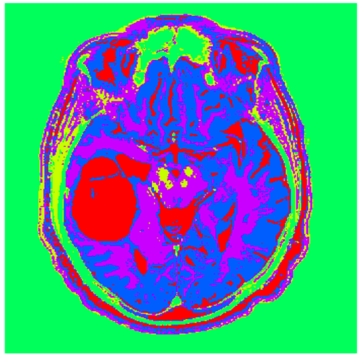
Example – *k*-means Clustering

Segmenting MRI images to identify tumors

- Specify number of clusters *k*
- Randomly assign each data point to a cluster
- Compute cluster centroids
- Reassign points to closest cluster centroid
- Recompute cluster centroids
- Repeat last two steps until no changes



Healthy Image

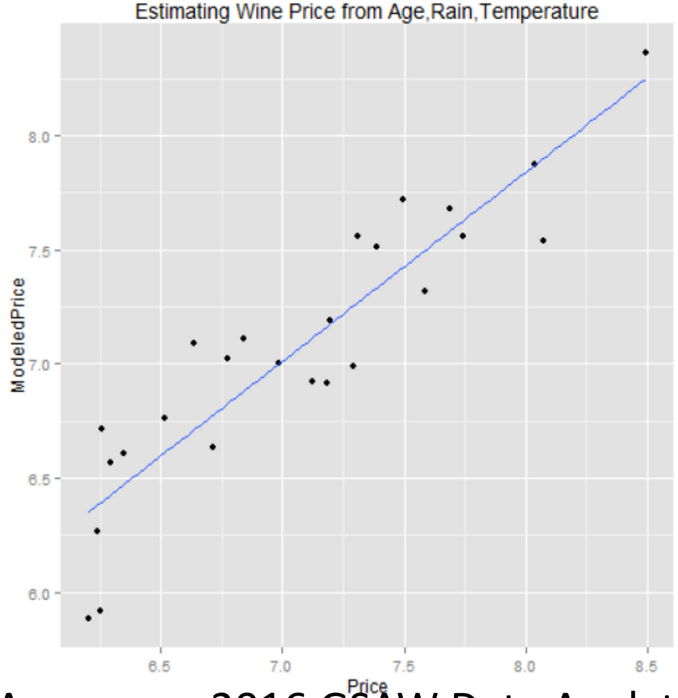


Tumor Image

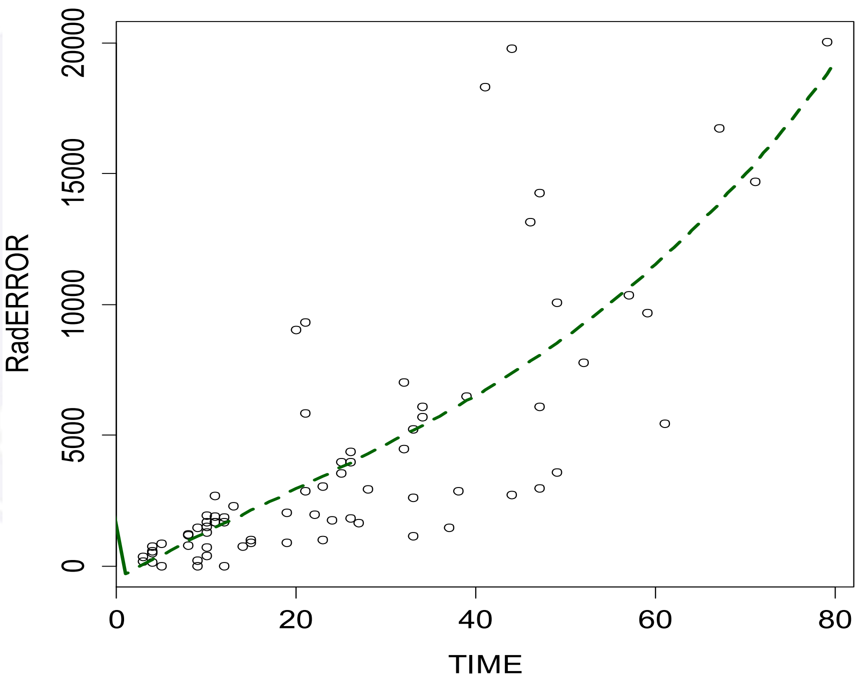
Supervised Learning – Linear Regression

- The linear model is the foundation of many models
- The goal is to find the model coefficient that will minimize the error between prediction and observed truth
- This model is more flexible that it seems - it accommodates non-linear functions of the attributes

Predict the price of wine based on three variables



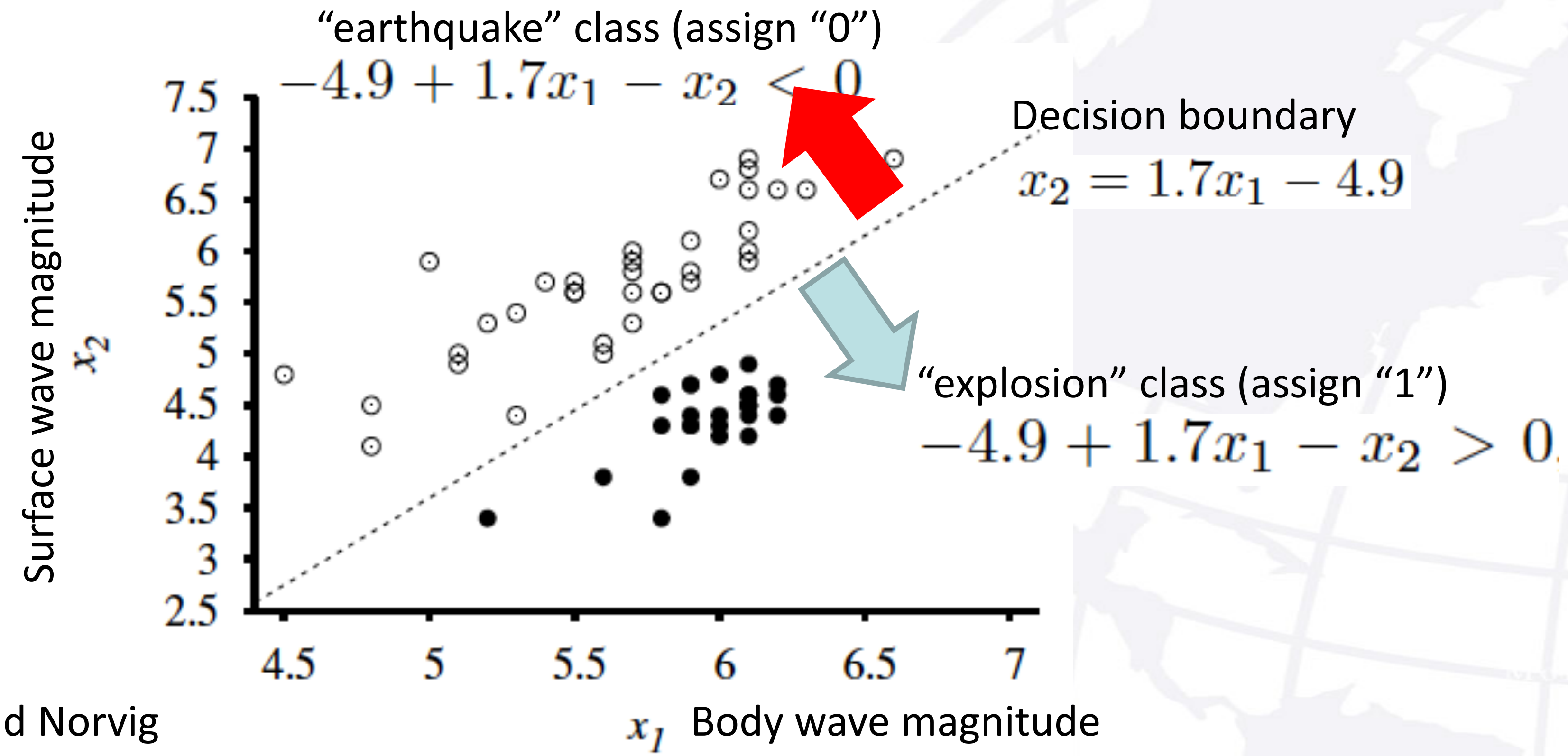
Predict error as a function of time



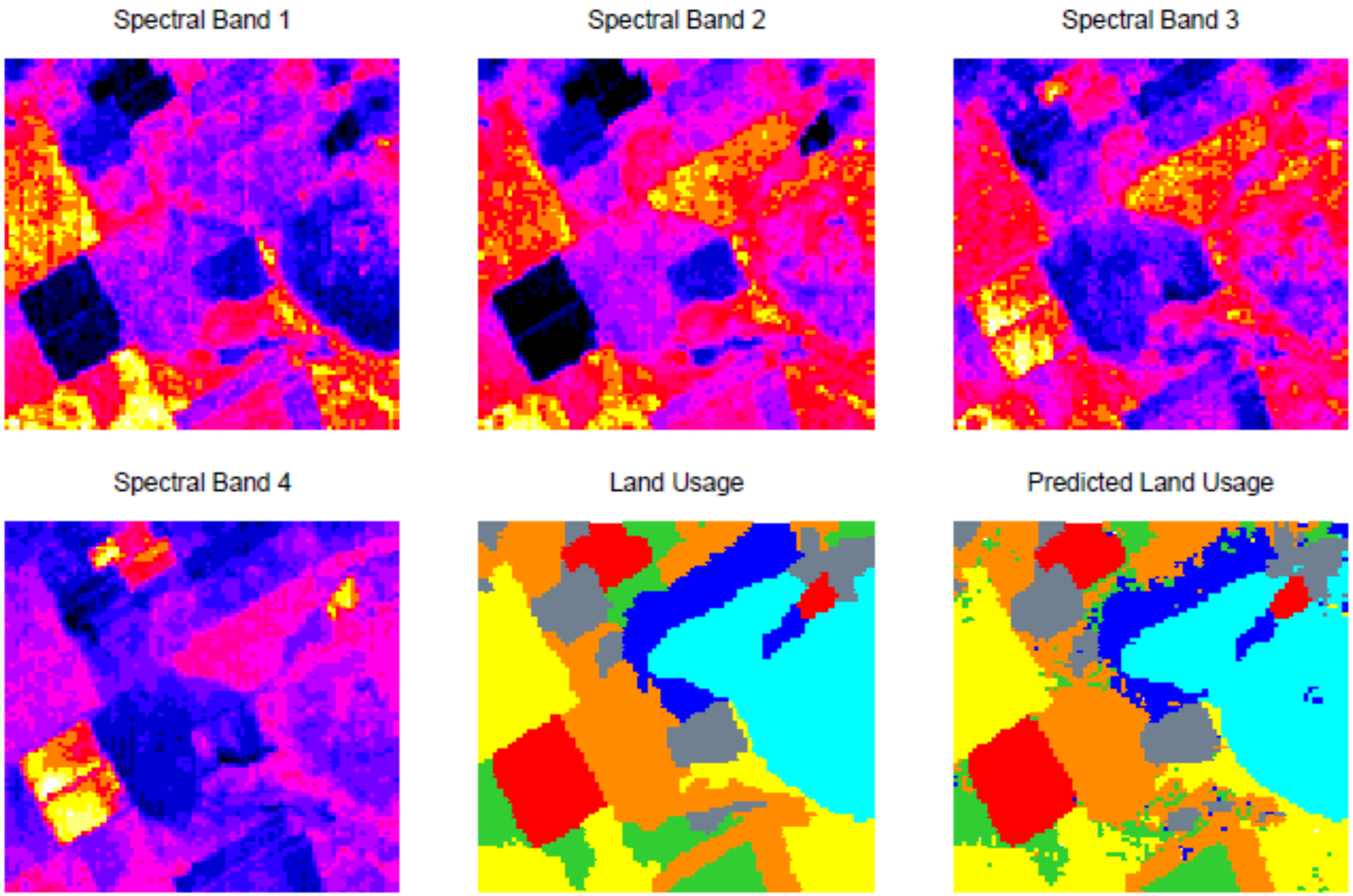
Aerospace 2016 GSAW Data Analytics and Whitaker, 2015

Linear Classification

- Task: Learn a hypothesis that will take new observations and correctly classify the data
- Decision boundary: Line (or hyperplane for > two dimensions) that separates classes



Classify Land Use in a Satellite Image



Land Use Classes:

- *Red soil*
- *Cotton*
- *Vegetation stubble*
- *Mixture*
- *Gray soil*
- *Damp gray soil*

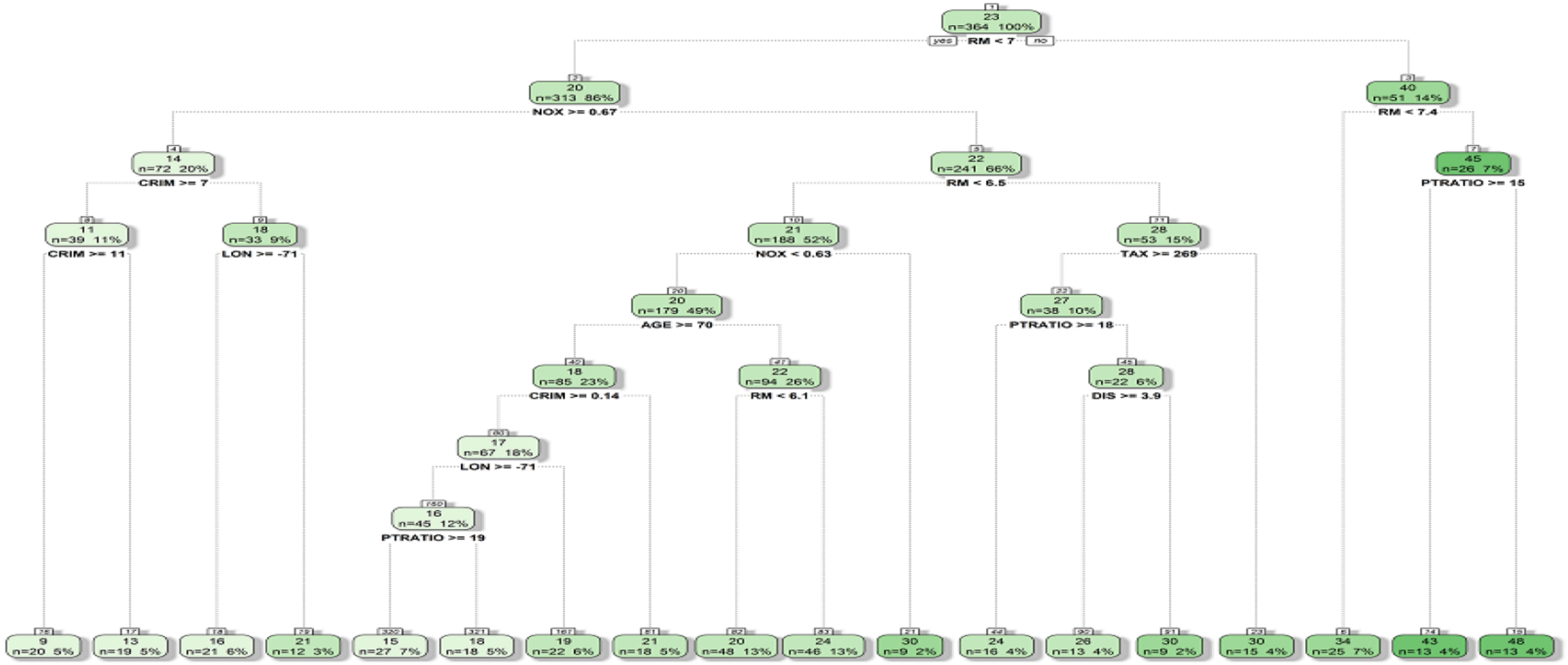
Whitaker, short course 2015

Decision Trees

Boston House Prices*

- Determine the median price of a house (in thousands) based upon:
  - *Location (latitude, longitude), Crime, % Zoning for Large Properties, % Industry in area, NOx emissions, Number of rooms, Age, Proximity to highways, Taxes, Proximity to Charles River, Pupil-teacher Ratio*

- Non linear classification
- Intuitive understanding of branches



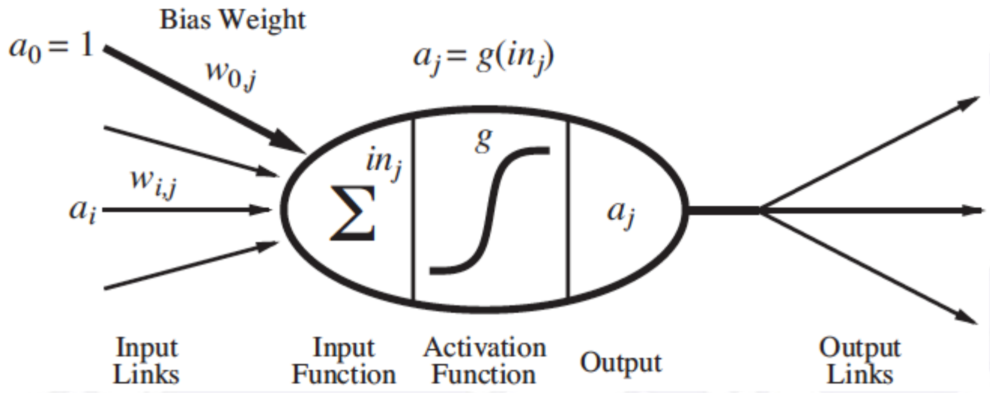
Data source: Harrison, D. and Rubinfeld, D.L. "Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol. 5, 81-102, 1978.

Models based on Neural Networks

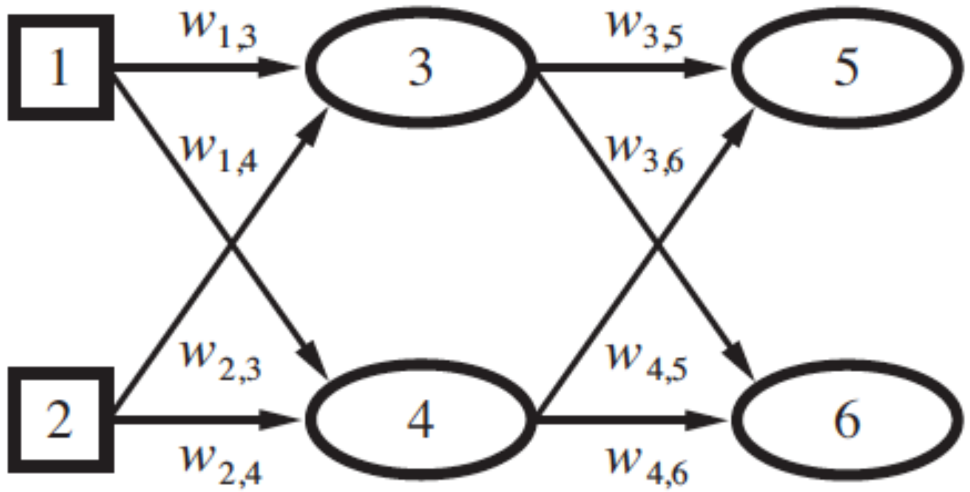
- Early technique in artificial intelligence field
- Designed to mimic human brain neurons
- Accomplish non-linear classification
- Learn the combination of weights and activation functions to minimize error between predictions and observations

- More complex networks with better performance (convolutional neural networks, deep learning)
  - Faster computers
  - Large training sets
- “black box” – no intuition into “how” classification happens

Neuron in the Network



Network of Connected Neurons



Neural Network Tasks

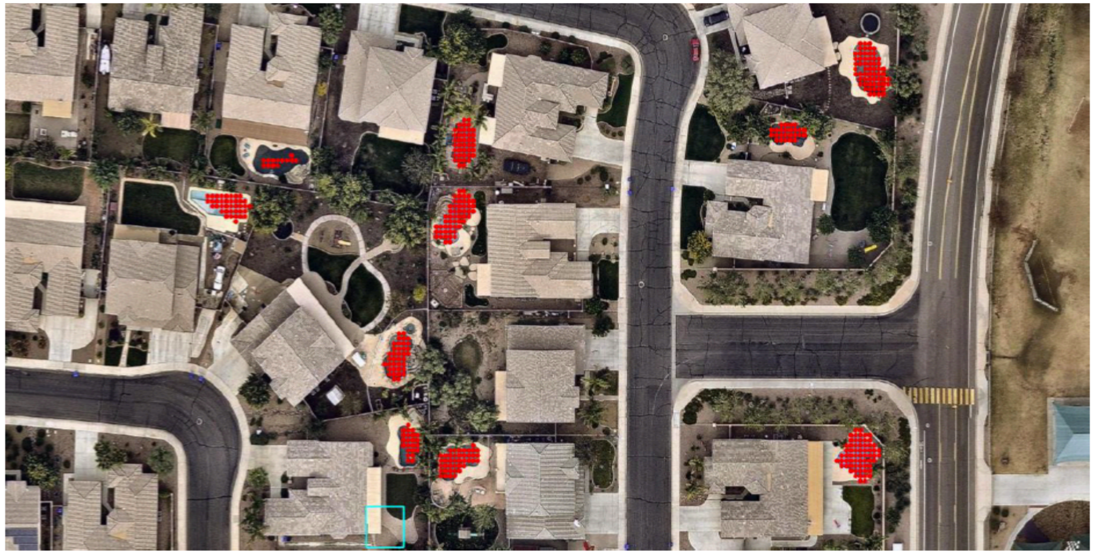
GSAW 2016 data analytics

Count the vehicles

- Image is a parking lot of 88 cars
- Classifies vehicles by multiple red-dots
- Dots are clustered and counted as unique objects

Count the swimming pools

- Distinguishes between pools and pool-shaped objects



Neural Network Image Classification Tasks

Train neural net to classify a new image based on 60,000 labelled thumbnail images

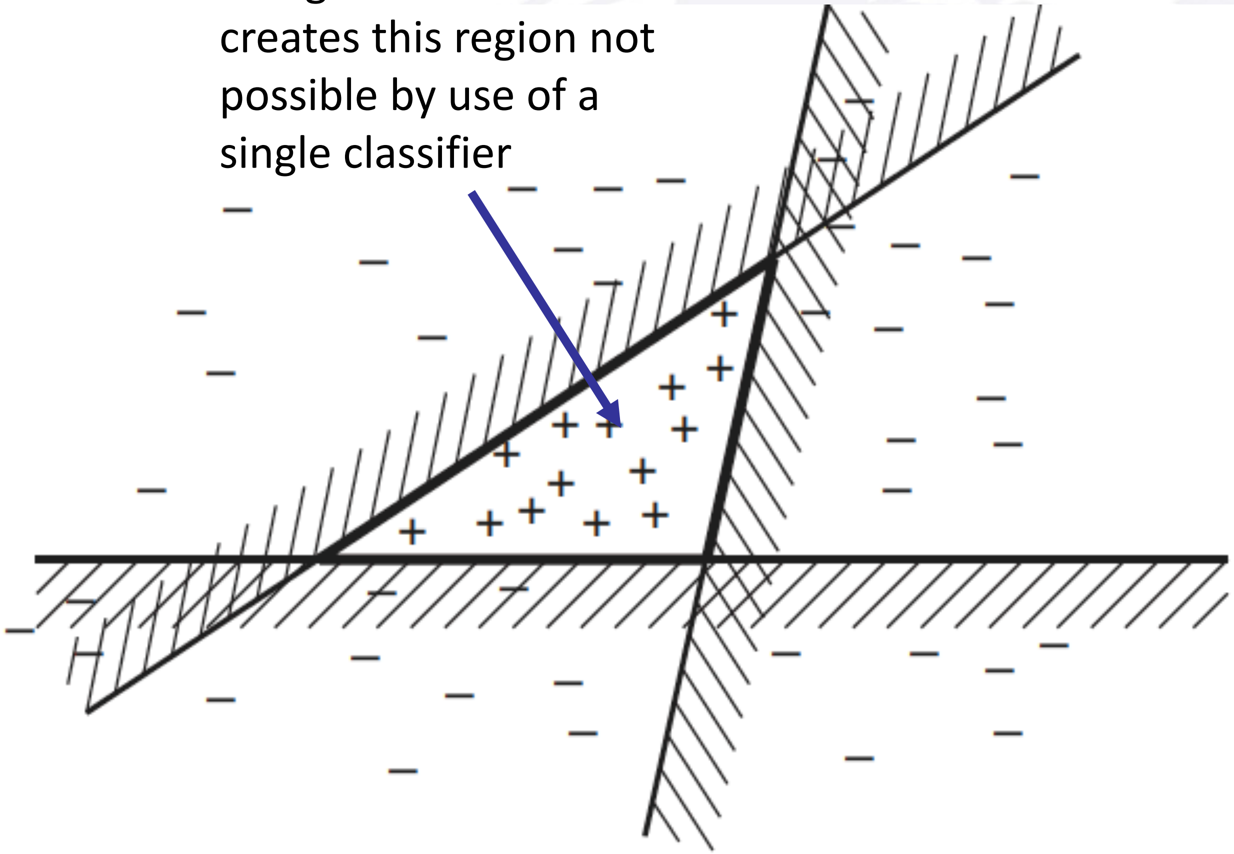
Classes include things such as:

- Frog
- Aircraft
- Bird
- Horse
- etc

Ensemble Learning

- Select a collection of models and combine their predictions
- Reduce misclassification below a single model
- Can be time consuming
- Models are not intuitive
- Ex: Random forests

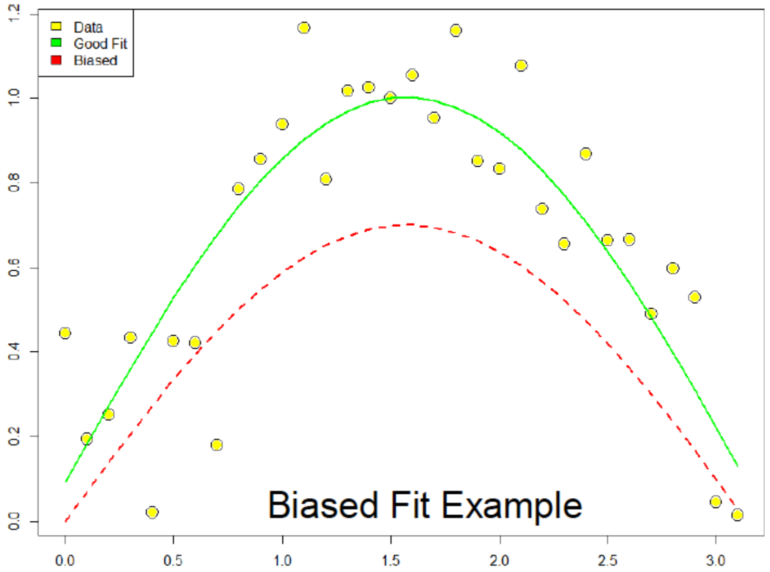
Using 3 linear classifiers creates this region not possible by use of a single classifier

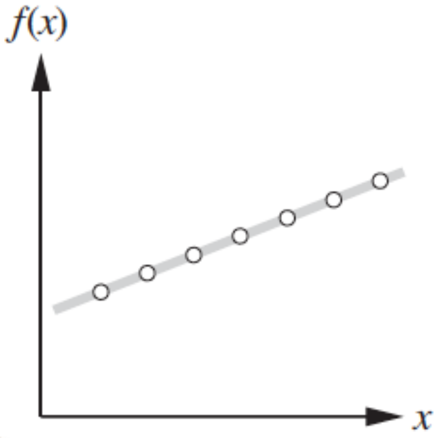


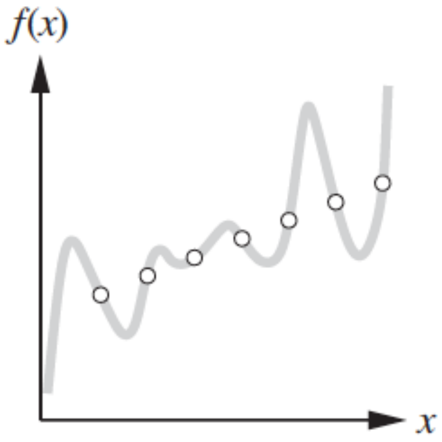
Evaluating Model Performance

- **Error** rate of model – proportion of mistakes it makes
  - Lots of ways to measure and summarize error
- **Validation** – testing the model on a set of examples not yet seen (called "validation" set)
  - Different than the "training set" that is used to generate the models
  - Different than the "test set" that is used to evaluate the final model we choose
  - Validation set tells us how well our models generalize
- Must avoid using the test data to influence the learning
  - Keep test set separate until you are totally done training the algorithm

Common Model Errors







- Bias
  - Systematic errors, always over or under predicting
- Variance
  - Large but non-systematic differences between model and predictions
- Overfitting
  - Over sensitivity to features in the training set but not in the general training set
- Non-significance
  - Apparent relationship shown in the model that is irrelevant in the general data set

Visualization

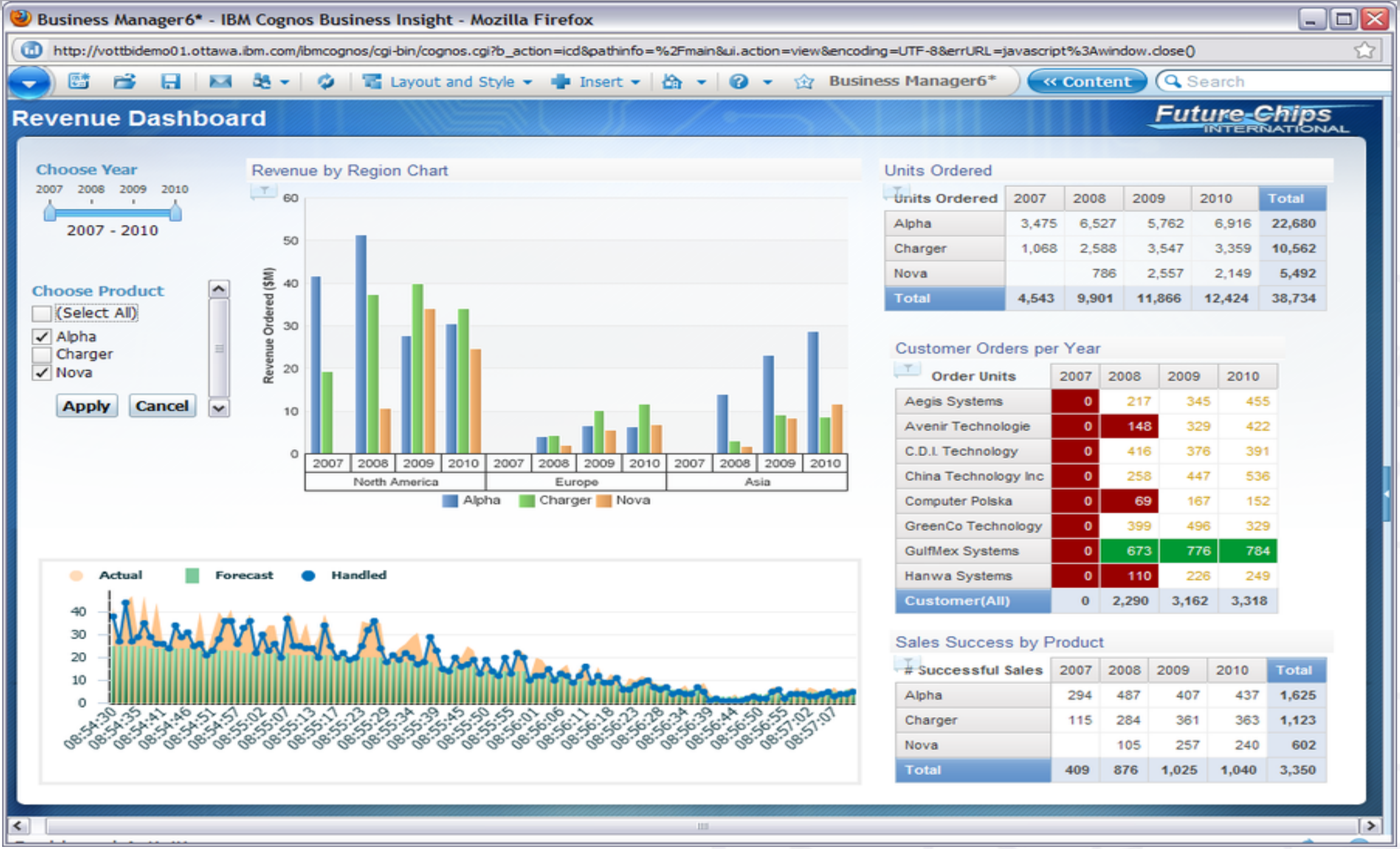
- Improve cognition with graphics to enhance the human visual system's ability to see patterns and trends
- Limit to how much information can be shown at once without overwhelming the viewer
- Readability of the data is critical:
  - Font size
  - Contrast
  - Overlaps between display components
  - Rate of display change
- Reduce redundancy by including a "legend" for repeated format, shape, or color encodings
- Ambiguity of display components can cause lost opportunities and serious errors

If done correctly, the conclusion is self-evident without the need for a lot of conversation

Visualization Thought Process

- What are the results of the analysis?
- What is the best means to display them?
  - Table, graph, both, neither
- Where will the variables be displayed?
- Where would you place other objects?
- Is there particular data to highlight?
- Is there a message I need to convey?
- How can I get the decision maker to interact with the data to better understand the behavior of the key attributes?
- Did I answer the question that was asked?

Charts and Tables

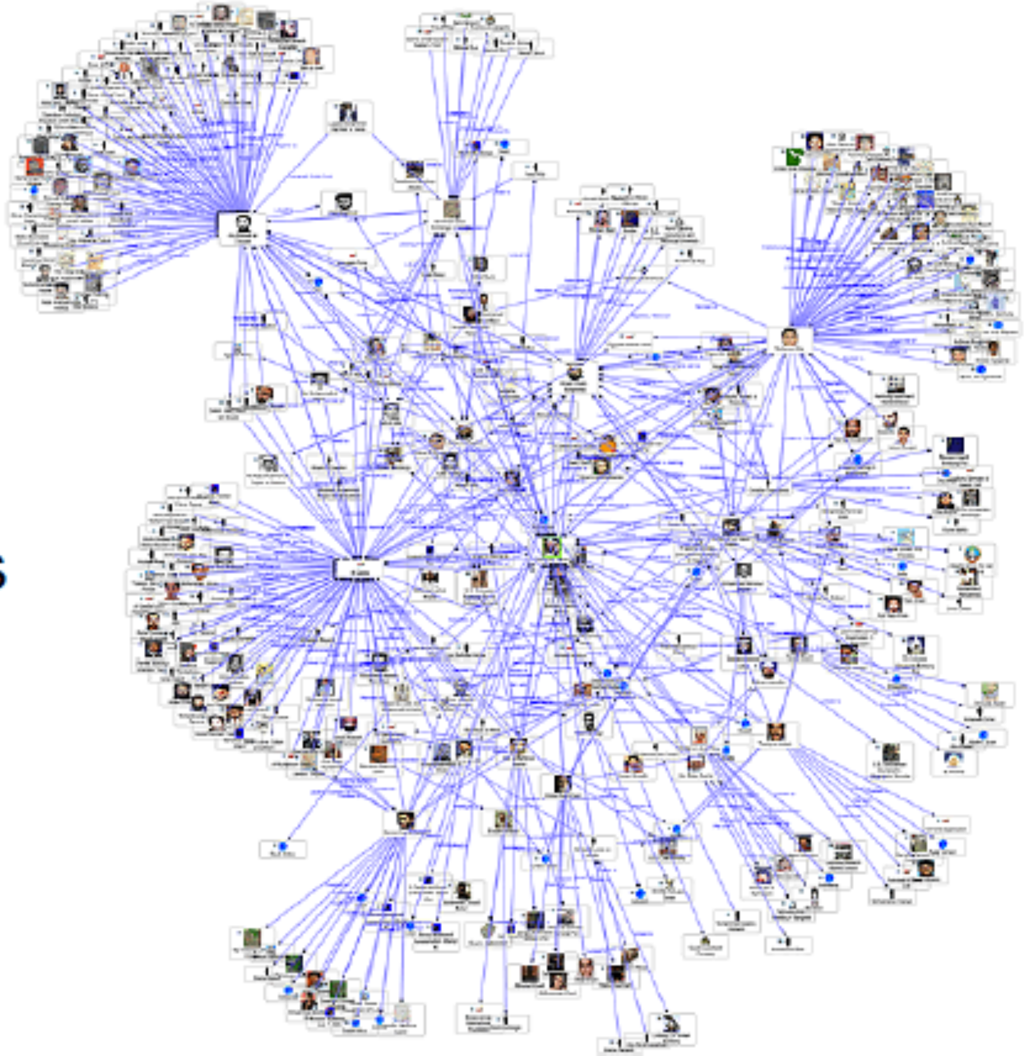


Graphs

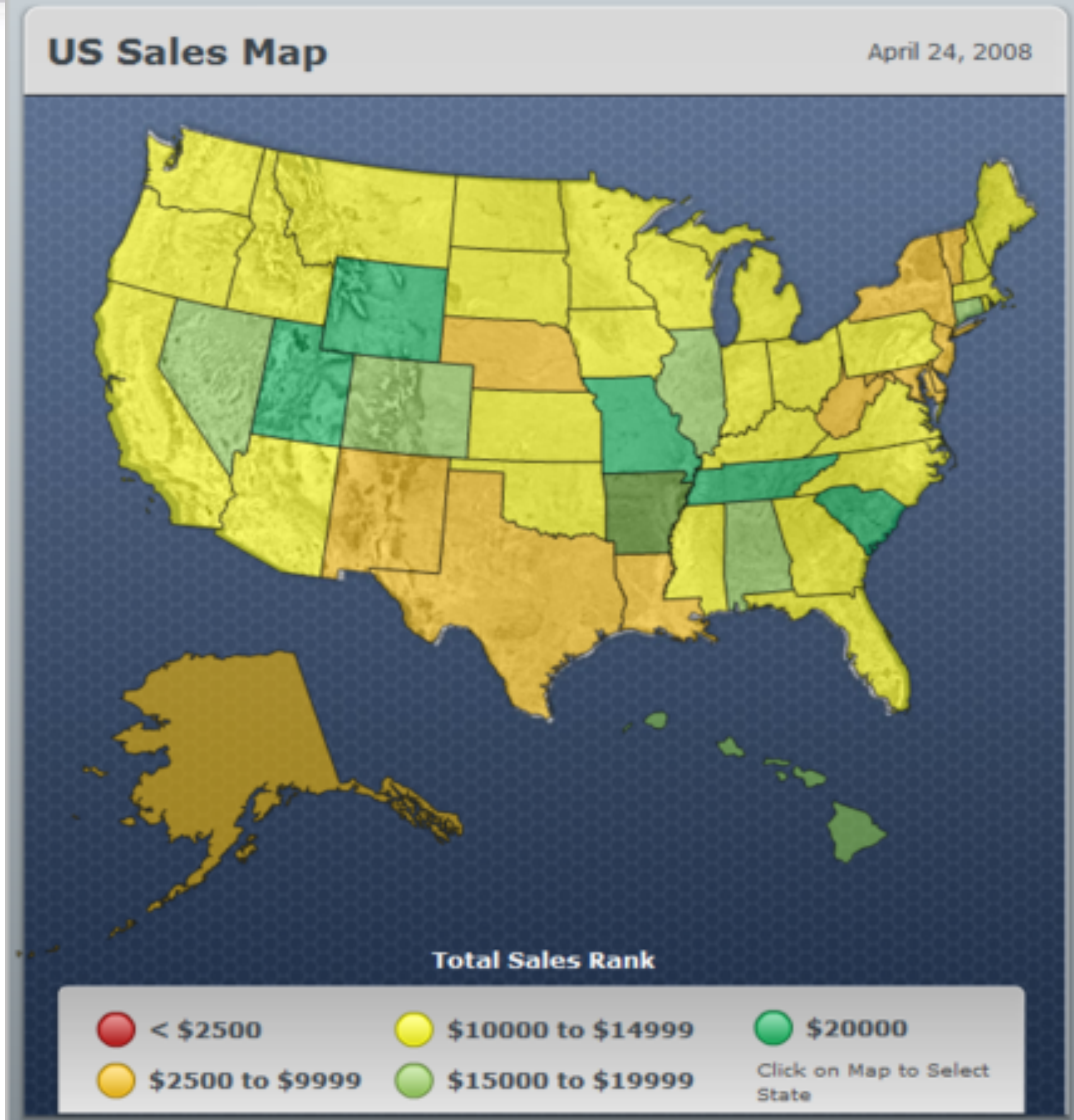
Lots of interesting data has a graph structure:

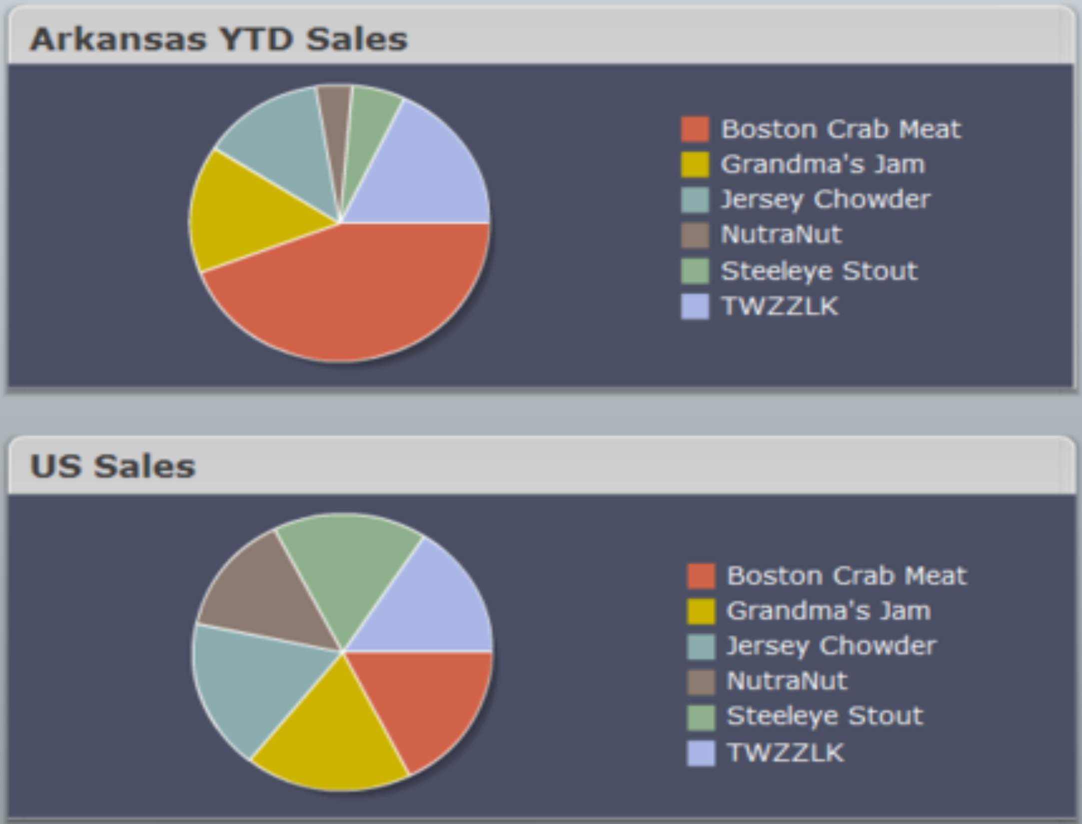
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

Some of these graphs can get quite large (e.g., Facebook* user graph)



Varied Interrelated Results

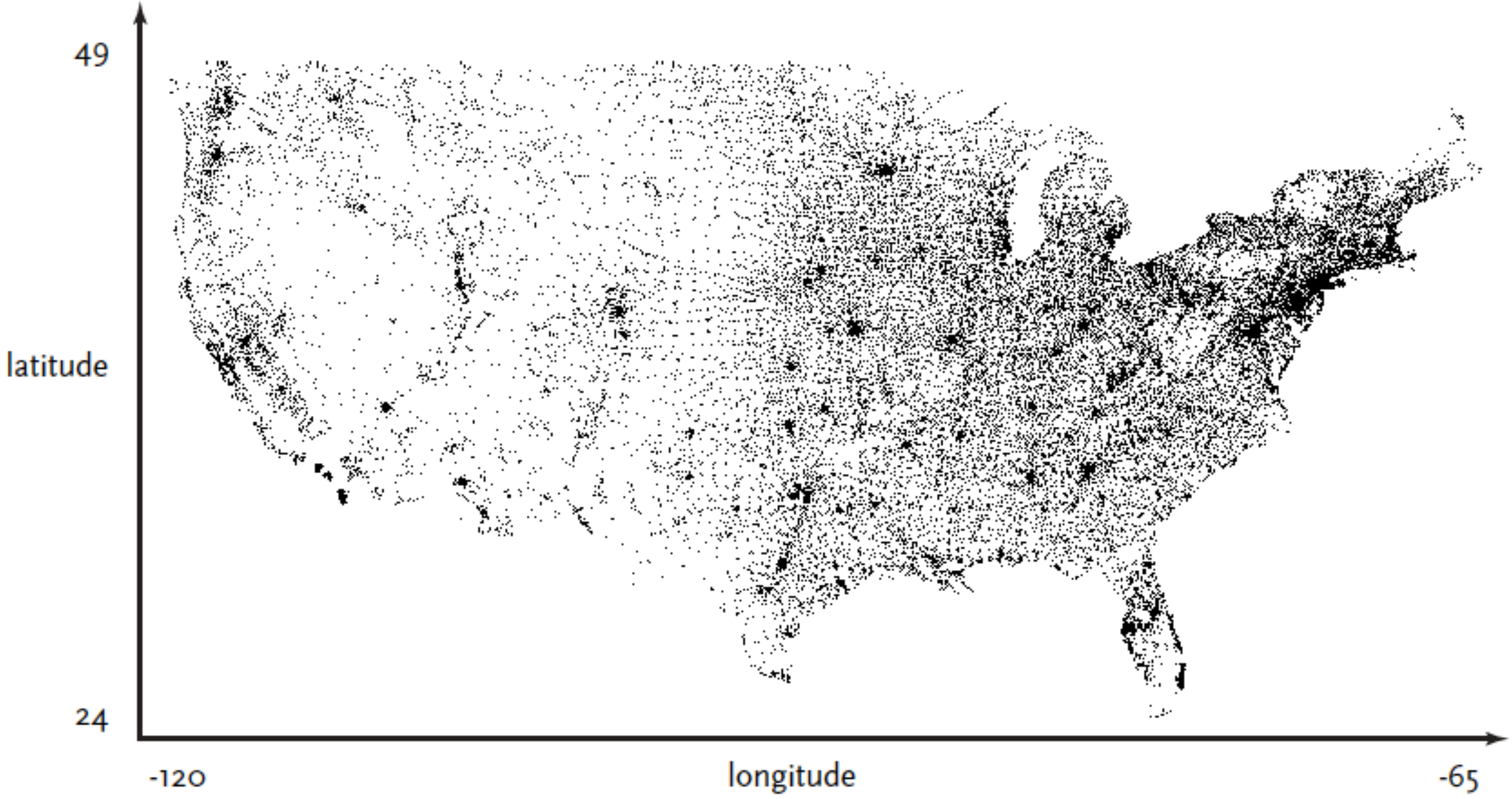




Arkansas	YTD Sales	Product Share	US Sales	Product Share
Boston Crab Meat	$7,744	44%	$119,736	18%
Grandma's Jam	$2,621	15%	$123,226	18%
Jersey Chowder	$2,324	13%	$119,391	18%
NutraNut	$689	4%	$96,086	14%
Steeleye Stout	$1,007	6%	$113,523	17%
TWZZLK	$3,136	18%	$107,464	16%
TOTAL	$17,520	100%	$679,427	100%

Interacting with Data (1 of 3)

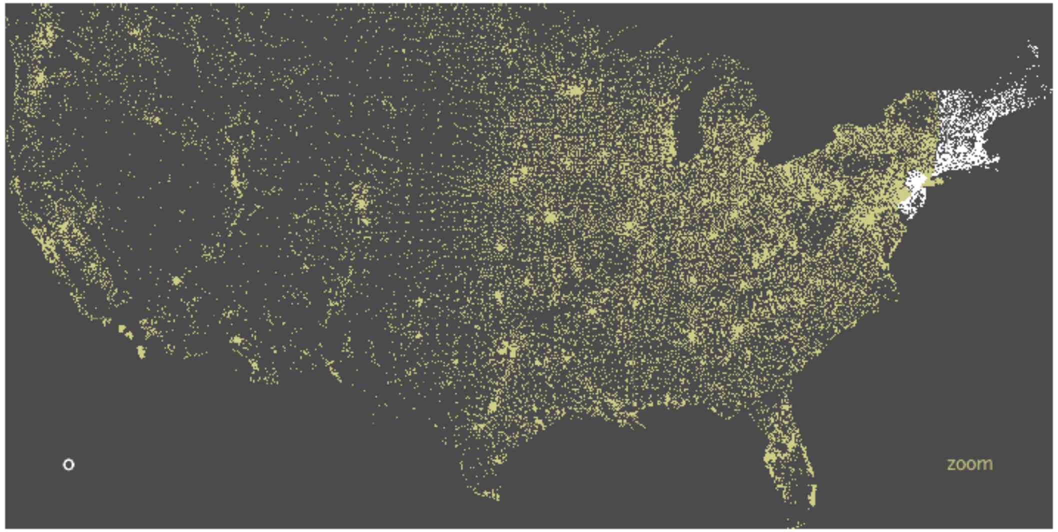
All Zip Codes in US



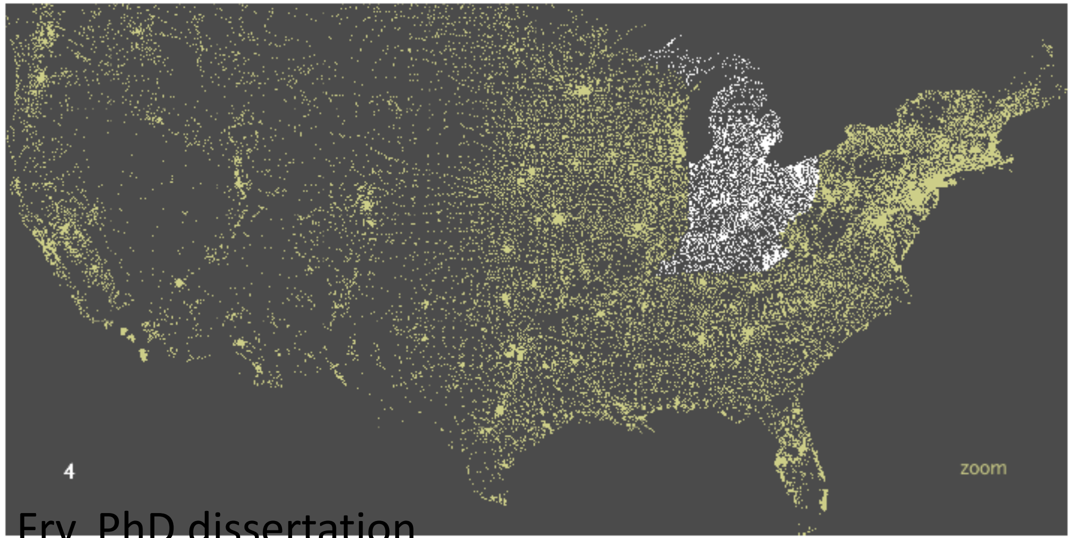
Fry, PhD dissertation

Interacting with Data (2 of 3)

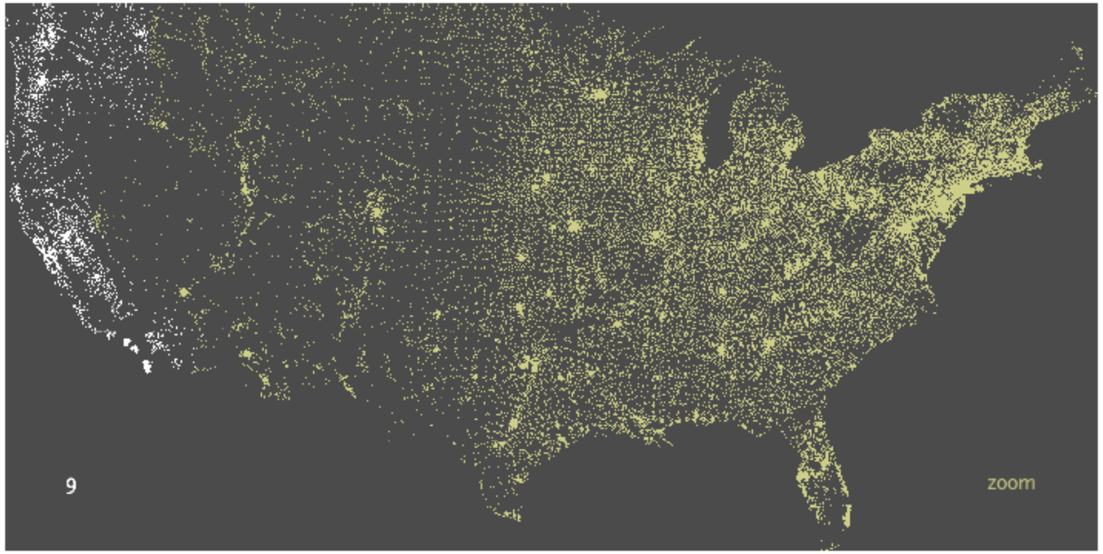
First Zip Code Digit = 0



First Zip Code Digit = 4



First Zip Code Digit = 9

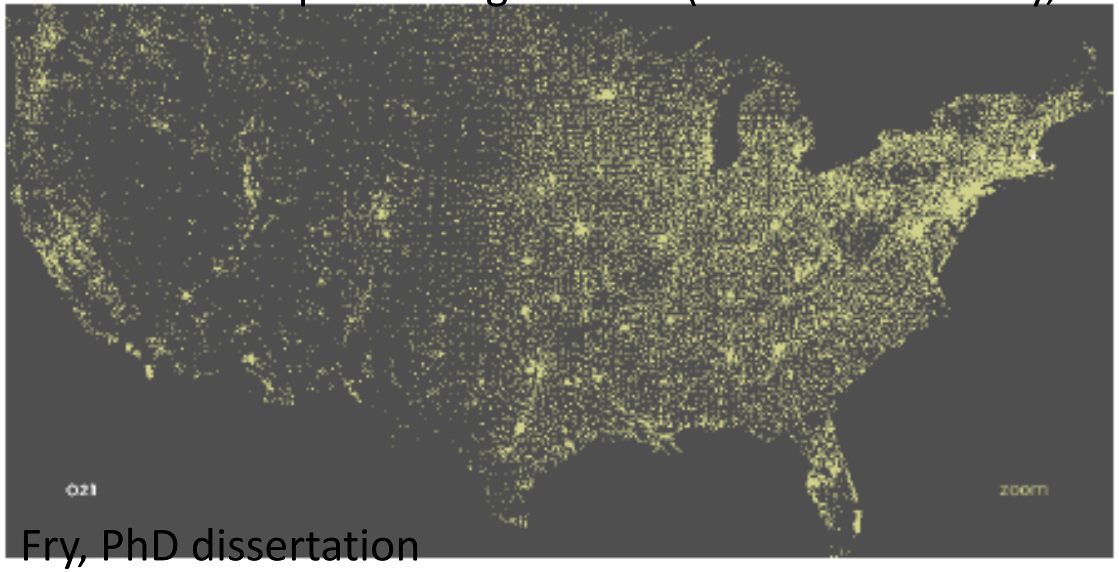


Fry, PhD dissertation

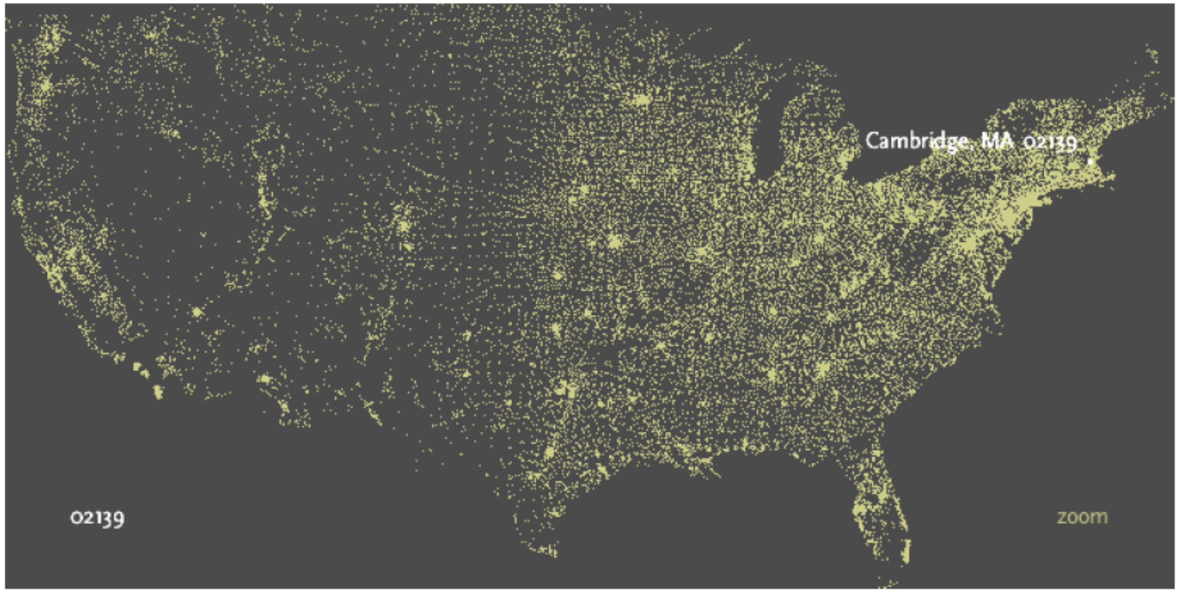
Interacting with Data (3 of 3)

First Two Zip Code Digits = 02 (Eastern Massachusetts)

First Three Zip Code Digits = 021 (Middlesex County, MA)



Zip Code Digits = 02139 (Cambridge, MA)



Fry, PhD dissertation

Data Science Takeaways

- Turns data into decisions
- Provides actionable information without exposing decision-makers to underlying data or analysis
- Helps you know your data
- Models need not be static – they can be updated and improved
- Key enabler of competitive advantage
  - Data-driven decisions instead of gut instinct, loudest voice, or best argument
- Data science capability is built over time (to be covered in last section)
- Data science is a repeatable process
- Data science is a multi-disciplinary team sport

BAH 2015 Field Guide to Data Science

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science

- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context

- Big data
- Cloud Computing

1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties

- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science

- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context

- Big data
- Cloud Computing

1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

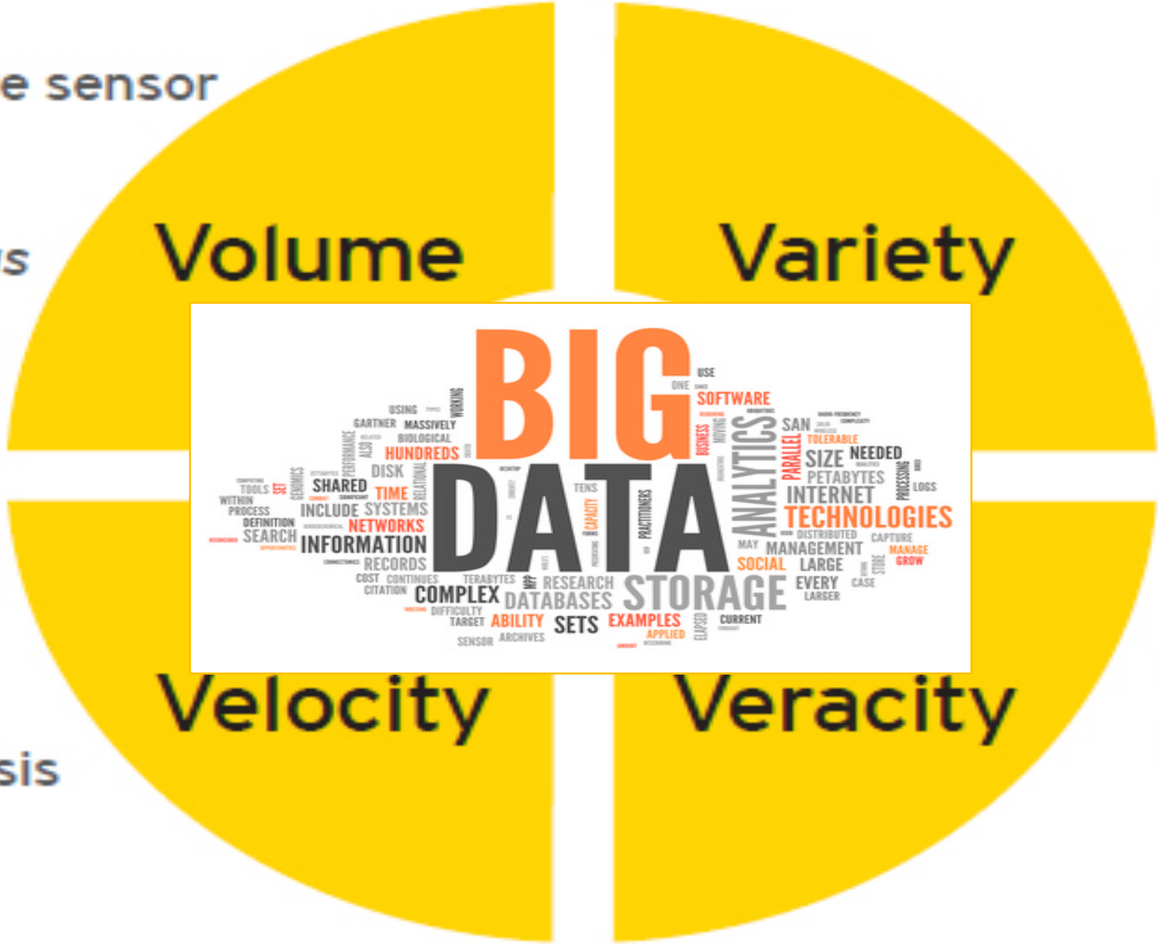
1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties

- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

What is "Big Data?"

- Click stream
- Active/passive sensor
- Log
- Event
- Printed *corpus*
- Speech
- Social media
- Traditional



- Unstructured
- Semi-structured
- Structured

- Speed of generation
- Rate of analysis

- Untrusted
- Uncleansed

Large, complex data that push technology limits and are difficult to store, process and analyze via traditional methods

Significance of Big Data

- Provides significant competitive advantages and supports disruptive innovation
- Enormous value to be extracted from the ever-increasing pile of transaction logs being aggregated by mission-critical systems
- Making faster, better, more confident business decisions
- Necessitates making connections between various types of structured and unstructured data

Challenges of Dealing with Big Data

- Doesn't fit in 1,048,576 rows and 16,384 columns)?
- Doesn't fit in memory (constraining factor for ?
- Doesn't fit on a single machine (starts at ~1TB)?
- Requires starting around 5-10TB)?

– **Storage**: At 1TB each, it takes 1000 computers to store 1 PB

– **Movement**: Assuming a 10Gb network, it takes 2 hours to copy 1TB, or 83 days to copy 1PB

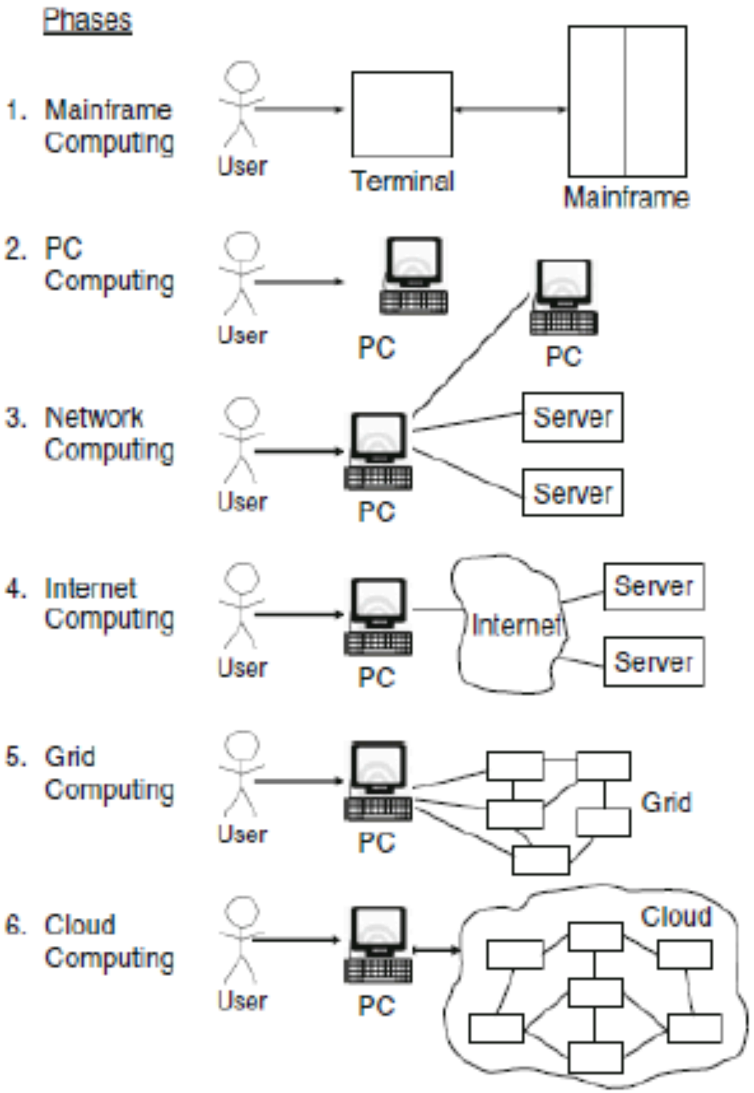
– **Searching**: Assuming each record is 1KB and one machine can process 1000 records per sec, it needs 277 CPU days to process 1TB and 785 CPU years to process 1PB

– **Processing**:
  - How do we convert existing algorithms to work on large data
  - How do we create new algorithms?

Performance of the traditional applications is becoming inadequate to process and analyze Big Data in a time- or cost-efficient manner

INSCOM brief, 2016

What is "Cloud Computing?"



- On-demand computing infrastructure consisting of virtual computers
- Computing, storage, and software as a service
- Treat as a utility, like power or phone
- Common characteristics:
  - Pay-per-use
  - Elastic capacity (scalable hardware resources)
  - Self-service interface
  - Resources that are abstracted/virtualized

CS3200, Handbook of Cloud Computing

“-as-a-Service” Models

- **Software as a Service (SaaS)**
  Applications are accessible from various client devices through a thin client interface

- **Platform as a Service (PaaS)**
  User deploys applications *using programming languages and tools supported by the provider*

- **Infrastructure as a Service (IaaS)**
  User provisions processing, storage, and networks to deploy software

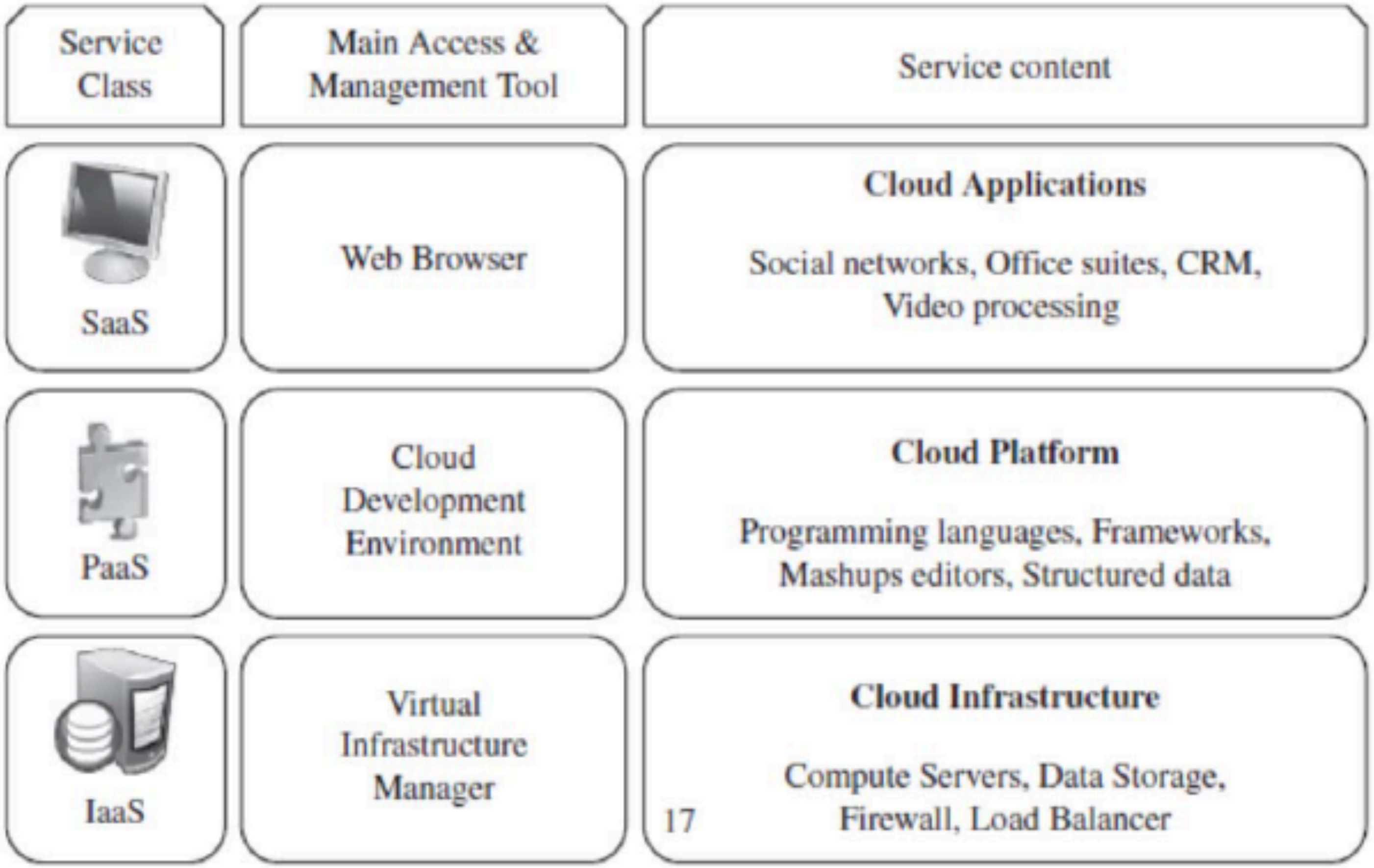


FIGURE 1.3. The cloud computing stack.

Amazon Web Services (AWS)

- Elastic Cloud Compute (EC2):
  - Virtual Machines (Computers) that Reside in the Cloud (just like a real computer, you choose RAM and Storage size)
  - Choose Linux of Microsoft image

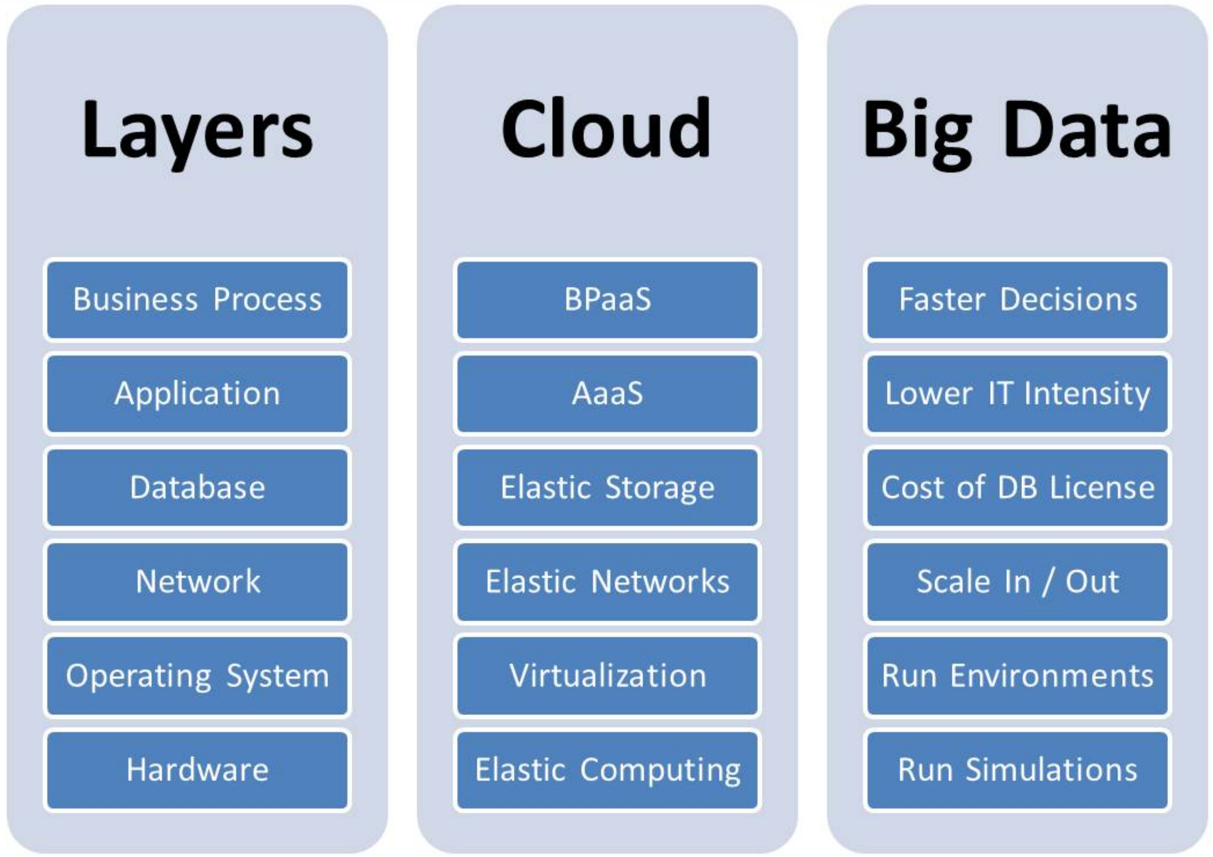
- Simple Storage Solution (S3)
  - “Buckets” that can store files
  - Think of this as an infinitely expandable Dropbox in which you only pay for the storage used

Scalability

- Simplicity for growing or shrinking some operation in response to demand
- Use same tools for data, large and small
- Methodology that lets us build once, rather than re-engineer continuously

Big Data and Cloud Computing

- Data science applications can capitalize on the benefits provided by cloud environment
  - No significant improvements in “scaling up” (upgrade a single node)
  - Lots of improvements with “Scaling out” (add more nodes)



Credit: Watalon.com : Big Data and Cloud

As data sizes grow, it is impossible to economically scale a single machine to meet the demand => must distribute across cluster of machines

Why Big Data and Cloud Computing are Significant to Data Science

- Old way of thinking about data: static
  - Store data in a data warehouse and chip away little bits of it to study
  - Use a single-node database to collect, store, and run queries
- New perspectives on data: dynamic
  - Distributed computational tools are more accessible
  - View data as flowing from source to destinations
- Open source data technologies provide the opportunity to combine different tools to deal with lots of data efficiently
  - Data pipelines to transform data
  - Clusters of inexpensive machines

Reference: M. Manoochehri, Data Just Right: Introduction to Large-Scale Data and Analytics, 2014.

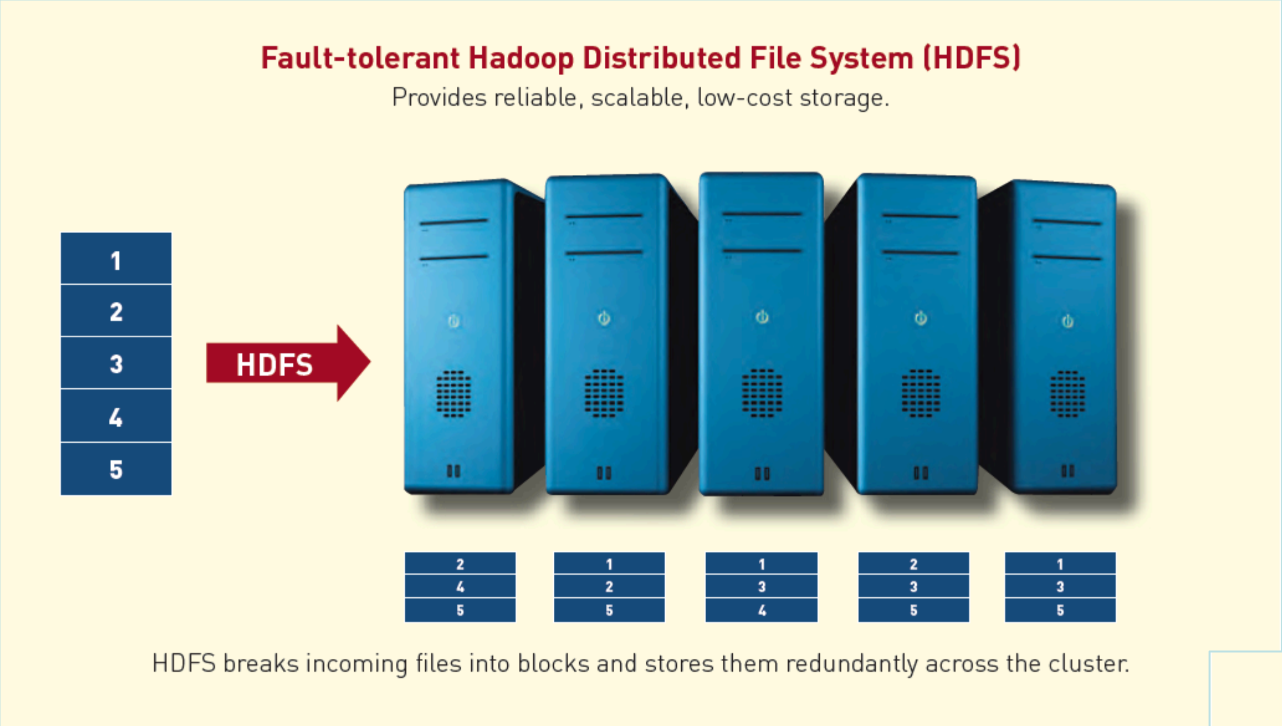
Big Data Technologies

- Popular Technologies
  - NoSQL databases:
    - Database systems optimized to process large unstructured and semi-structured data sets
    - Short for "Not only SQL"
  - MapReduce and Hadoop frameworks:
    - Store and process large unstructured and semi-structured data
    - Based on a distributed computing paradigm
- Common characteristics:
  - Commodity hardware enables scale-out
  - Parallel processing techniques
  - Non-relational data storage capabilities
  - Advanced analytics and data visualization technologies

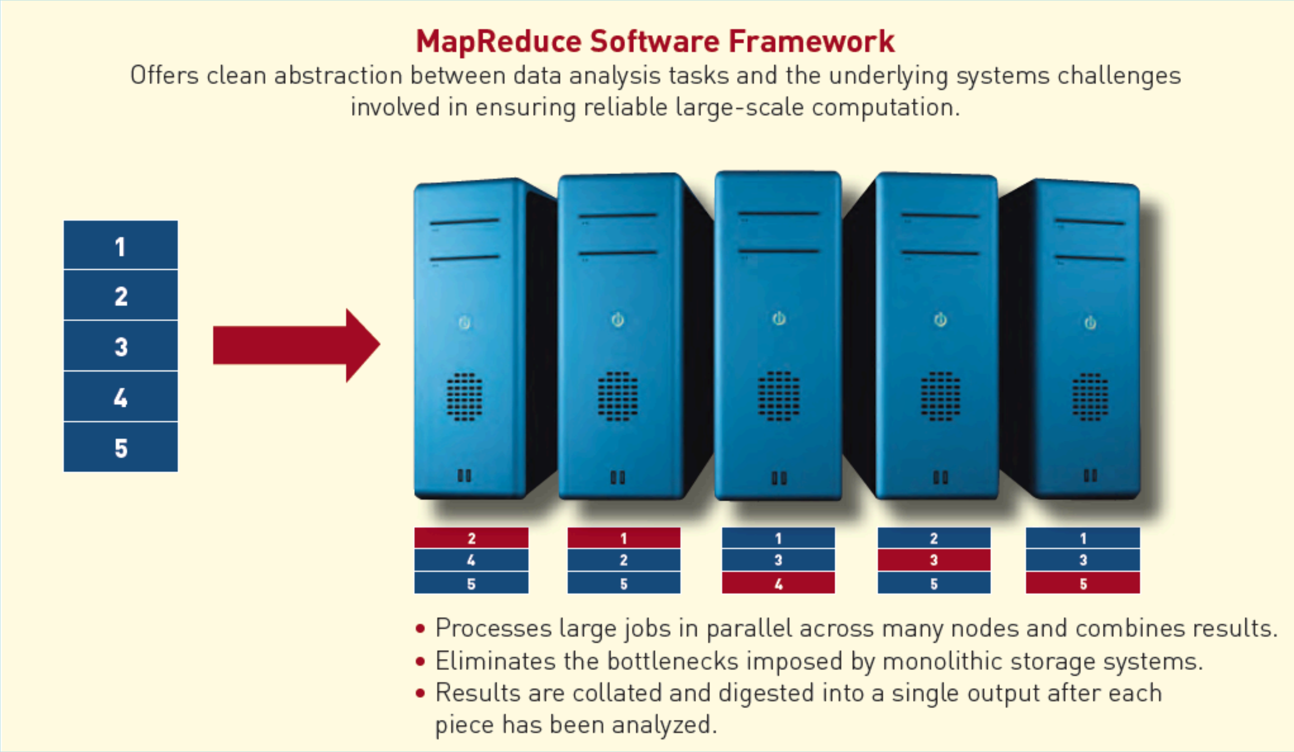
MapReduce

- Do large-scale data transformations in a reasonable amount of time
  - Developed by Google in 2004
  - Move the program to the data rather than the data to the program
  - Divide and conquer approach: Map phase + Reduce phase
- Use Cases:
  - Given a set of records, collect all records that meet a condition
  - Given a set of records, transform each record into another representation (text parsing, value extraction, convert from one format to another)
- Example:
  - The New York Times spun up 100 Amazon EC2 instances
  - 4TB of scanned articles (11 million articles from 1851-1980 in TIFF form) were sent to Amazon S3
  - S3 cluster converted these into 1.5TB of PDF documents
  - Required re-scaling, gluing articles and converting to PDF
  - Each page could be worked on separately, in a fairly natural divide-and-conquer approach
  - Took <24 hrs compared to what would have taken weeks of work

Hadoop



- Introduced in 2009
- Open source implementation of MapReduce
- Hadoop is built on two parts:
    1. Hadoop Distributed File System (HDFS) – inexpensive, reliable, distributed file storage
    2. MapReduce – parallel data processing system that exploits the distributed storage of HDFS



Ref: M. Olson, "HADOOP: Scalable, Flexible Data Storage and Analysis "IQT Quarterly, Vol1, No 3, Spring 2010, pp.14-18.

Working with Big Data

TRADITIONAL APPROACH:

Data stored in one place

DISTRIBUTED APPROACH:

Data stored in many places

Manipulating the data

SQL

MapReduce

Parallel DBMSs

Data structure and namespace

Relational database models

Nonrelational database models: BigTable and HBase

Persistent Storage of Data

File Systems

Distributed Files Systems: GFS and HDFS

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science

- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context

- Big data
- Cloud Computing

1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties

- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

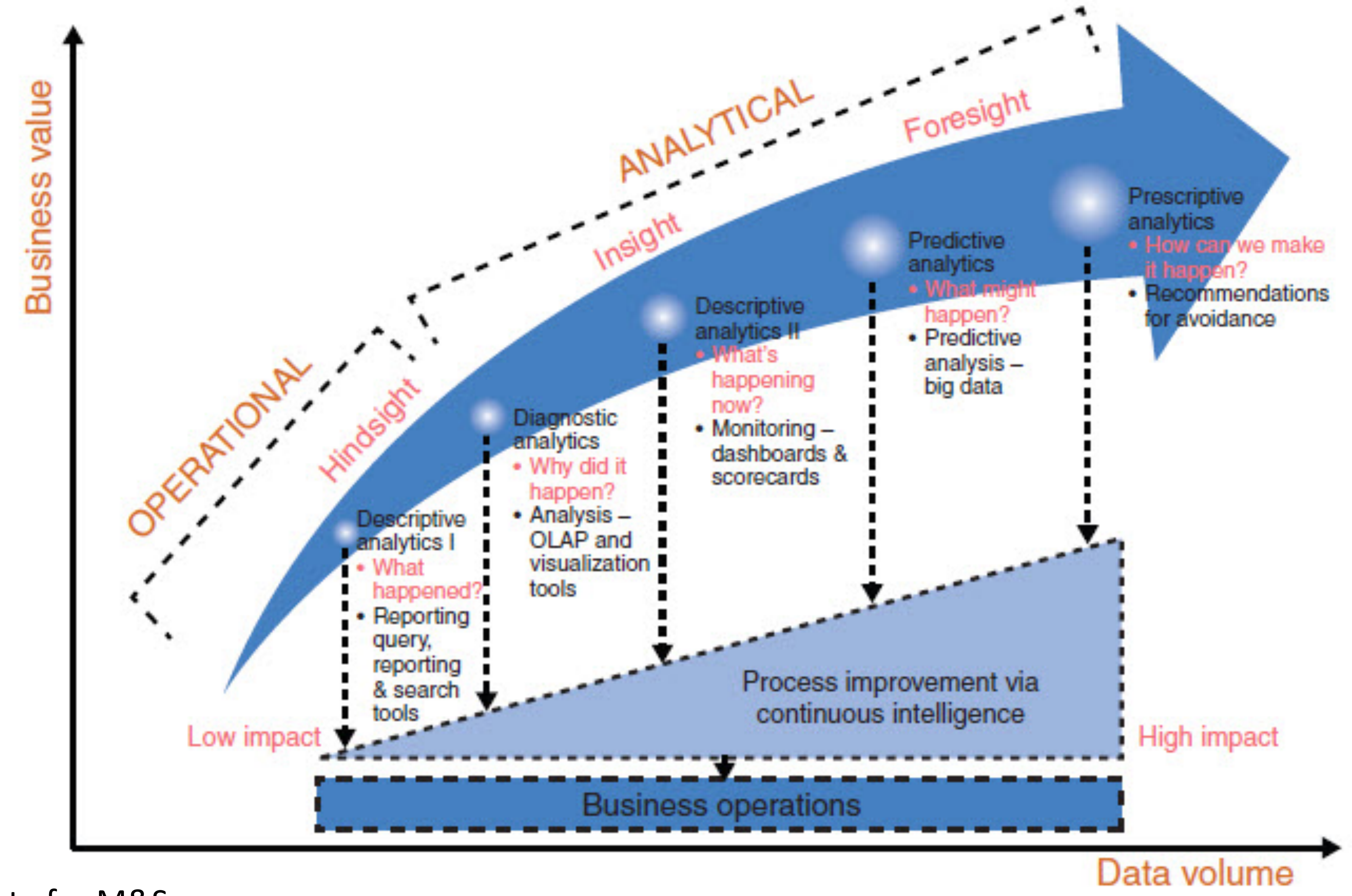
How will you Interact with Data Science?

- As **leaders**, you create the environment for data science:
  - Interdisciplinary team with the proper training and education
  - Compute/storage/network environments and infrastructure
  - Policies that facilitate access to data and agile tool development
- As **practitioners**, you use data science tools to drive data-driven decisions through analytical insights

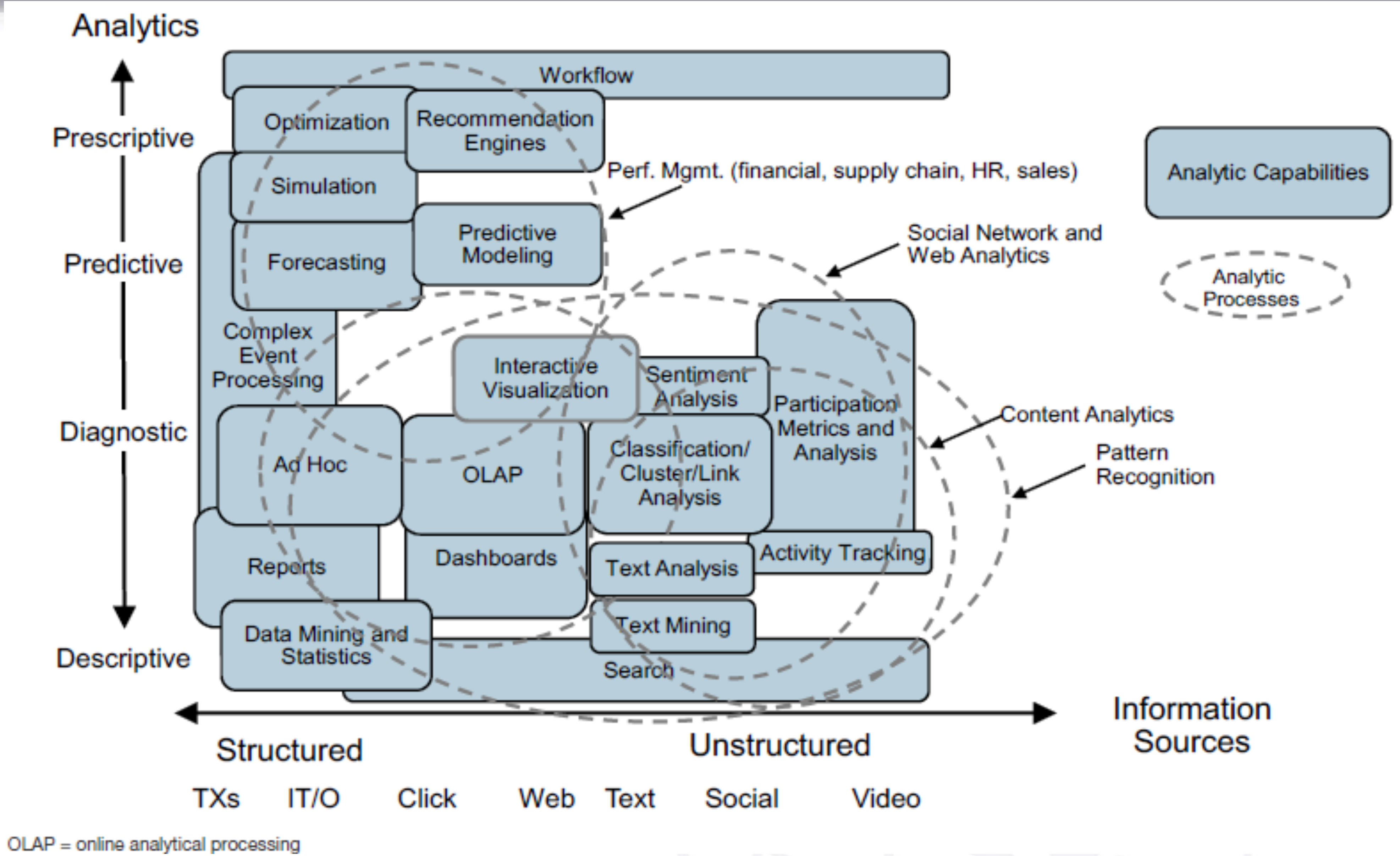
Building Data Science Capability Takes Time

- Crawl
  - Solve real problems using data science process
  - Data set may be small, but if it solves the problem, who cares?
  - Simple tools (excel) are fine
  - Important thing is to use the data to drive decisions
  - Education
- Walk
  - More sophisticated tools that allow analysts to write executable code (R and python)
  - Better data handling ability
  - Automation of workflows
  - Education
- Run
  - Creating robust, cloud-enabled capabilities to prototype proof-of-concept solutions
  - Putting repeatable, production dataflows in place
  - Rapidly constructing applications to serve user needs
  - High functioning, data science teams
  - Education

Understand the Types of Questions Data Your Organization Has (or should have)

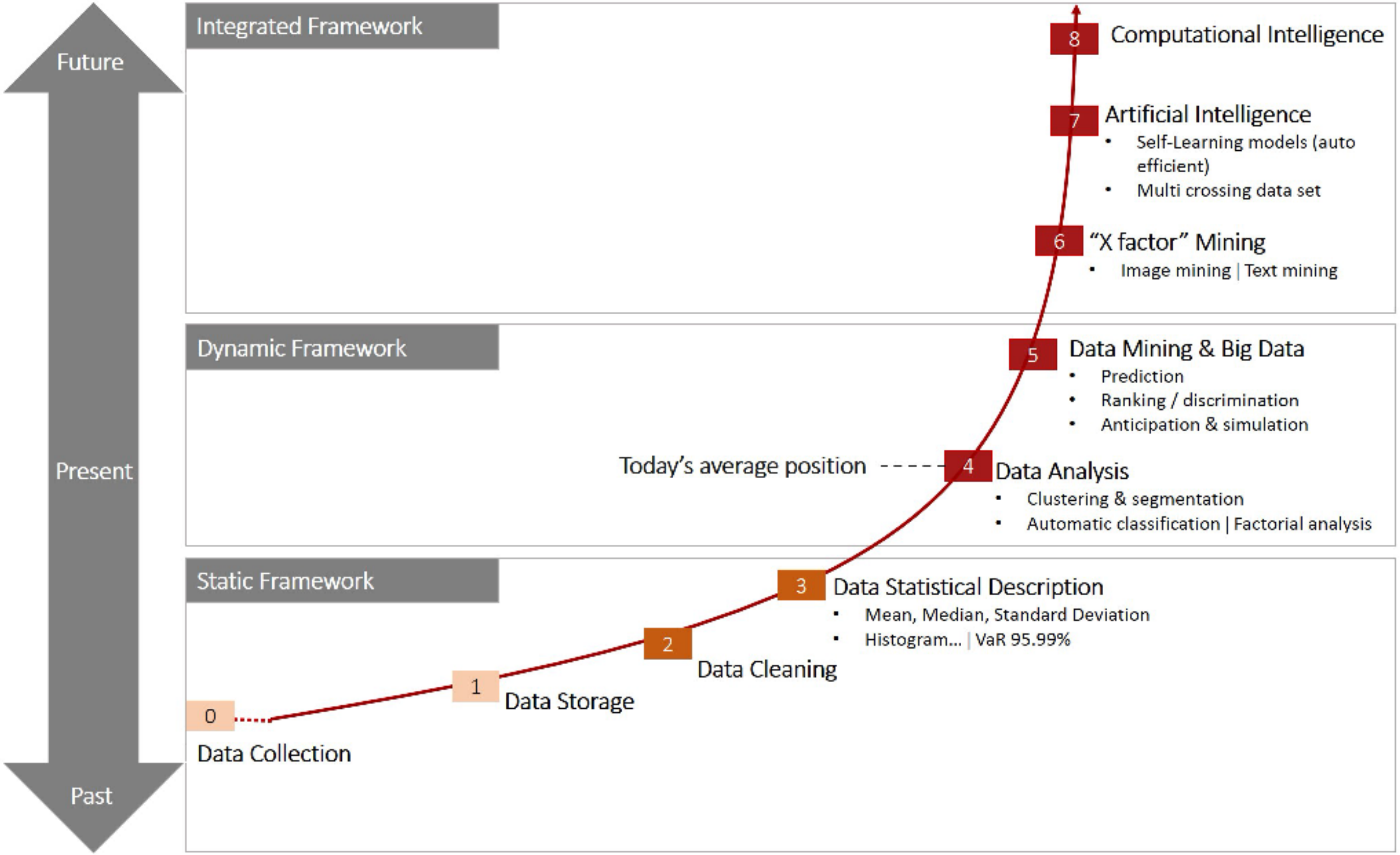


Understand the Variety of Data Science Methods



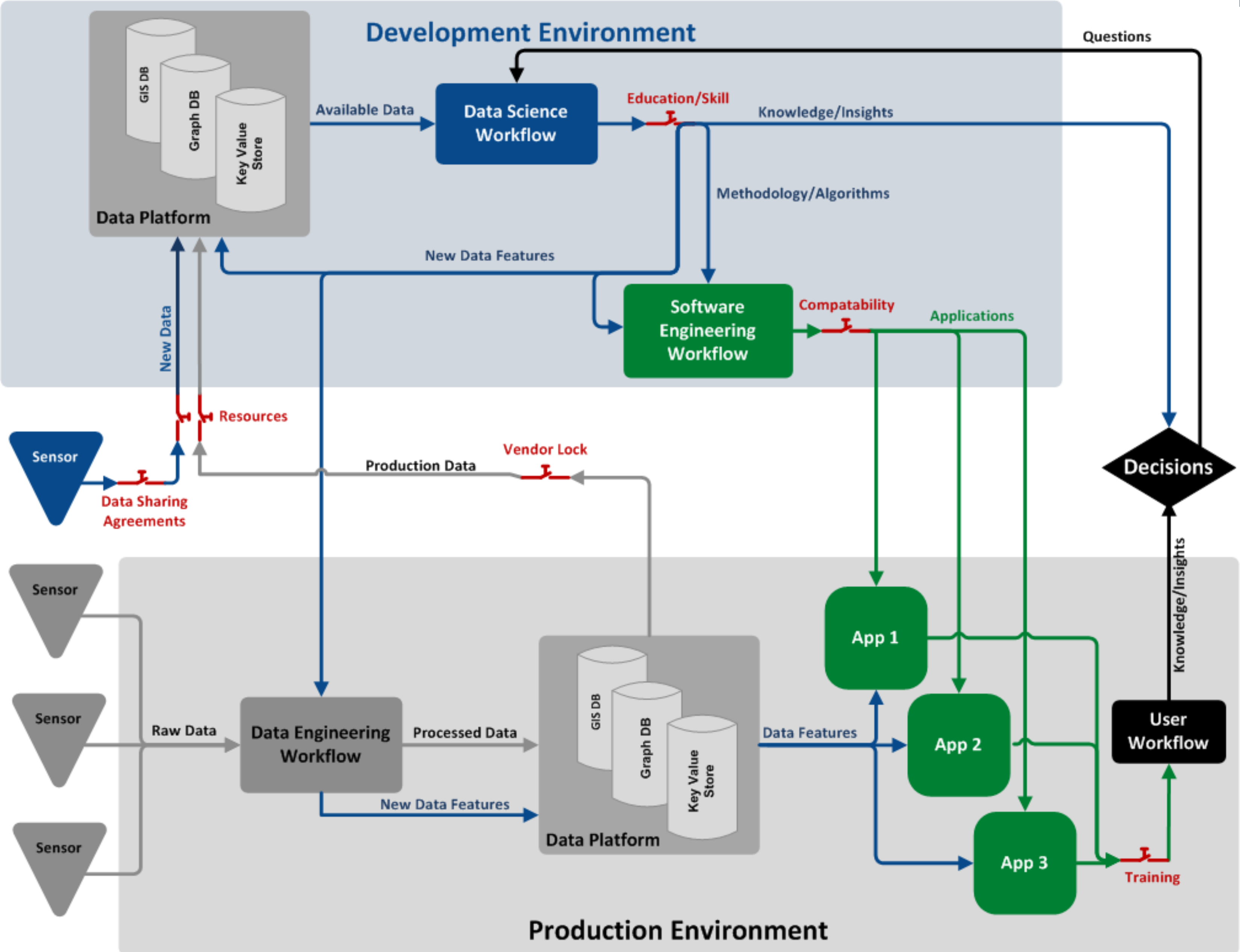
Gartner Group, 2012

Understand your Organization's Data Science Capability



Developed by Nii Attoh-Okine and Chappuis Halder
Available at http://interfacejournal.com/archives/1407

Operationalizing Data Science



Graphic: Samuel H. Huddleston and Isaac Faber "Making the Leap from Analysis to Analytics" working paper, 2017

Operationalizing Data Science in an organization requires connecting:

- Data to questions
- Sensors to decisions
- Development to production environments
- Users to apps

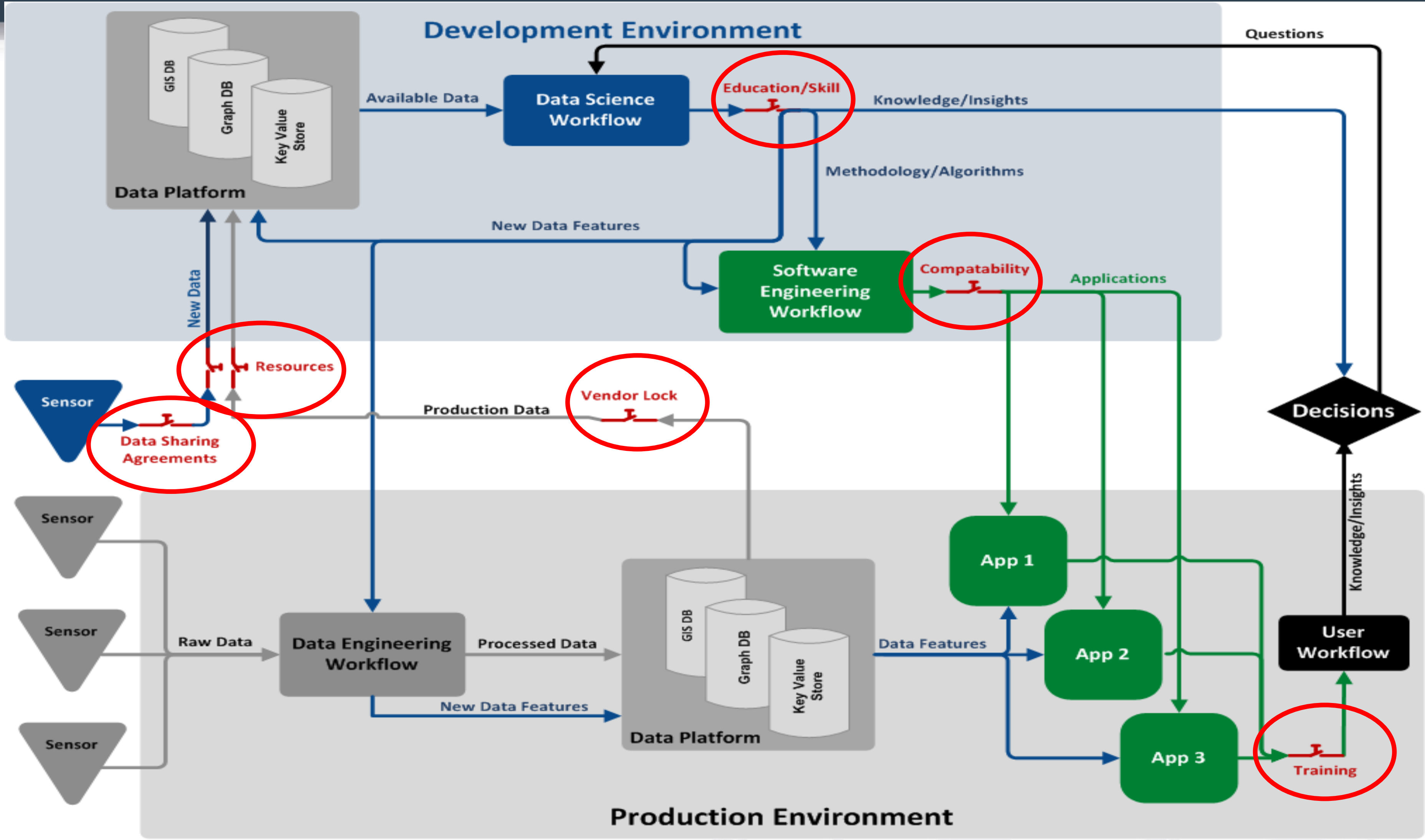
Data science is a process which involves:

- Policies
- People
- Not just buying a software suite

Data science in an organization is making the munge-model-visualize cycle repeatable and sustainable

NAVAL POSTGRADUATE SCHOOL

Clear the Obstacles to Connecting Data to Decisions



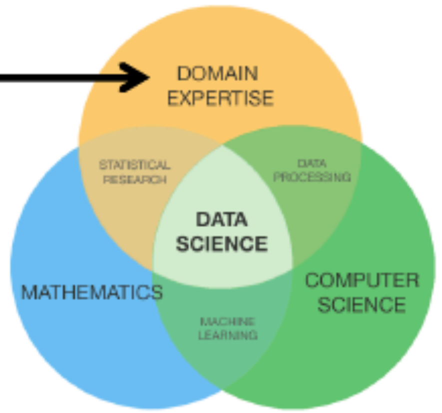
Graphic: Samuel H. Huddleston and Isaac Faber "Making the Leap from Analysis to Analytics" working paper, 2017

Leaders and managers need to clear the chokepoints in order to create a data science capability in their organizations

Assemble the Team (1 of 4)

People: Users/Consumers

- Analytics (Data Science) is a *team sport*.
- Analytic tools cannot replace the people currently performing the functions they are designed to support.
- Instead, ***analytic tools multiply the effectiveness*** of the people currently performing those functions.
- The users/consumers of the analytic tools are integral to their development; they have all the *domain expertise*.

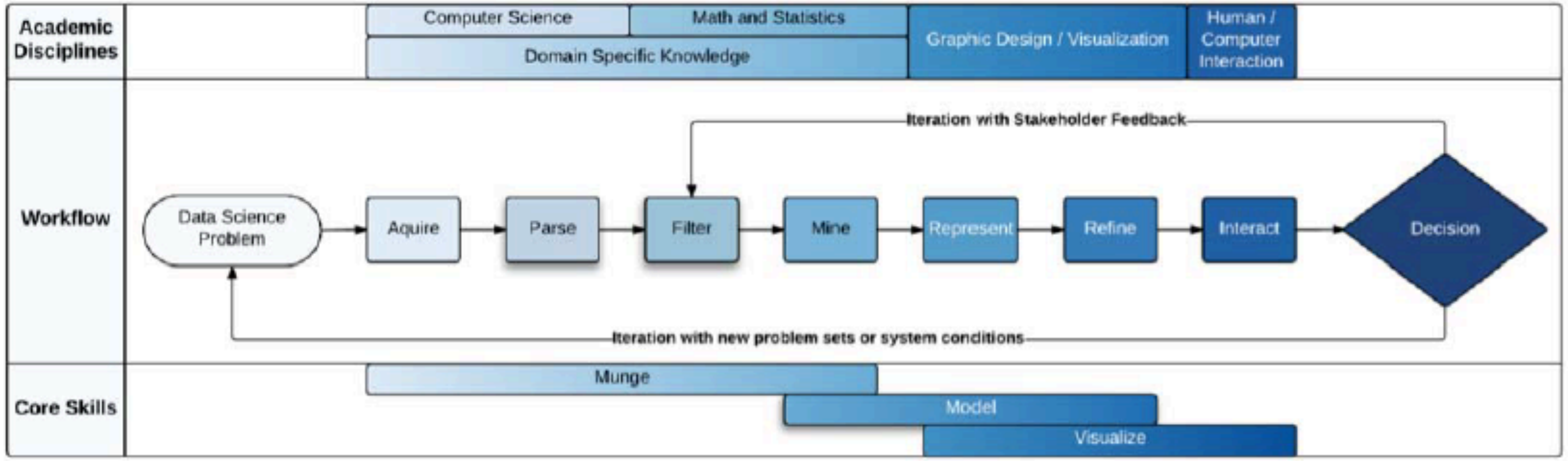


Source: Palmer, Shelly. *Data Science for the C-Suite.* New York: Digital Living Press, 2015. Print.

Integrate your highest performing analysts into your analytics teams (and make everyone better).

Assemble the Team (2 of 4)

People: Data Scientists



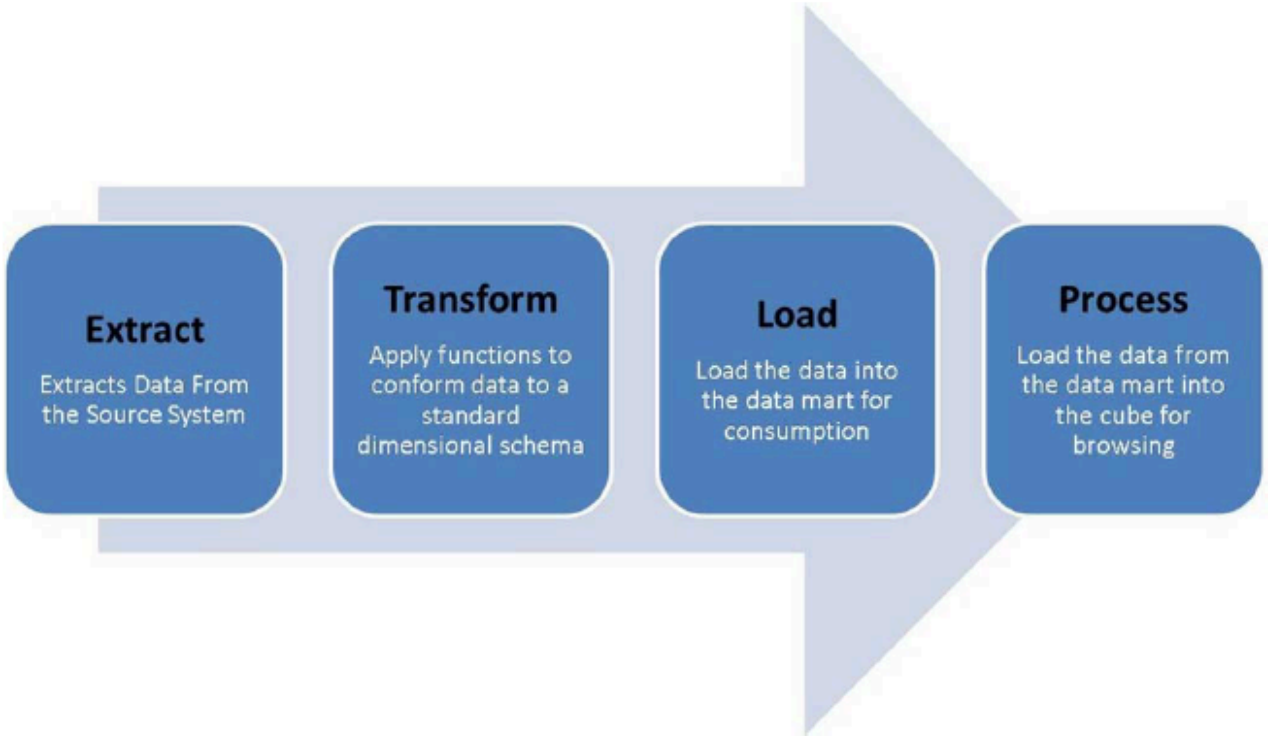
Data Science: the ability to extract knowledge and insights from large and complex data sets

Dr. DJ Patil, Chief Data Scientist , US Government

Data Scientists provide the *methodology* (answer the question!) for transforming raw data into meaningful decision tools.

Assemble the Team (3 of 4)

People: Data Engineers

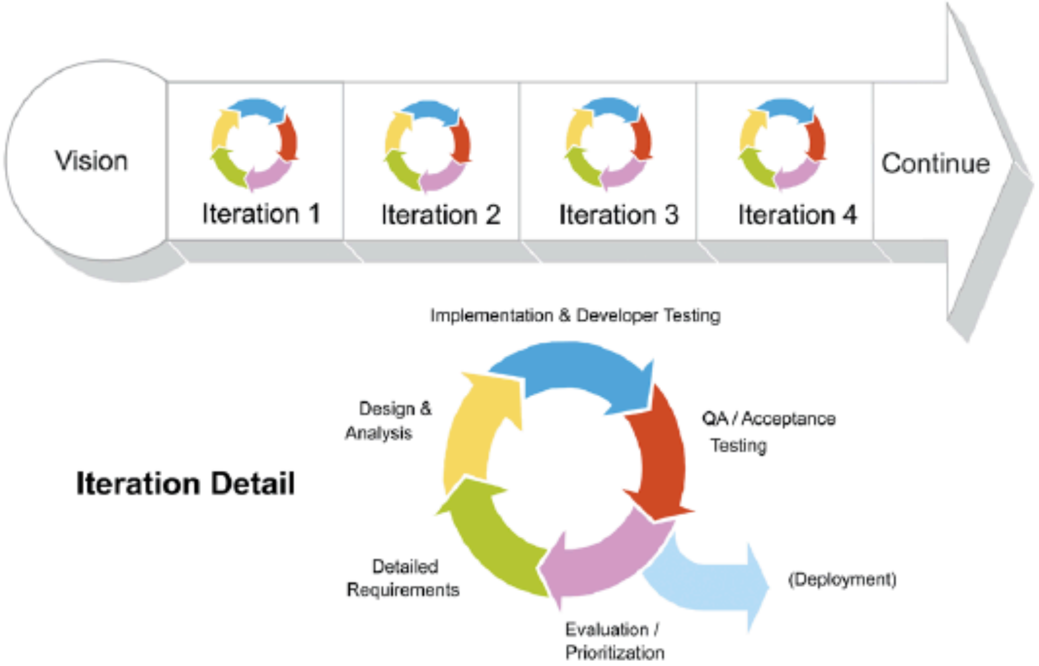


Data Engineers ensure that data is *always available* in a *consistent format* conducive to supporting both analysis and analytics (which are probably different formats).

Image Source: http://blogs.technet.com/b/servicemanager/archive/2012/02/03/olap-cubes-in-the-scsm-data-warehouse-olap-cube-processing.aspx

Assemble the Team (4 of 4)

People: Software Engineer



Software Engineering: the application of a systemic, disciplined, quantifiable approach to the development, operation, and maintenance of software *IEEE Standards 610.12 - 1990*

Software Engineers develop robust (scale, speed, reliability, usability) implementations of the data science methodology in a format *compatible with the data platform(s) and production environment.*

Image Source: http://scrumreferencecard.com/scrum-reference-card/

Decide on the Infrastructure

Platforms ("Stack")

- Production vs. Development Environments
- Cloud vs. Traditional Architecture
- Components of a Platform [Example: LAMP]
  – Operating System [Linux]
  – Server(s) [Apache web server]
  – Database(s) [MySQL database management system]
  – Compute/Scripting Language(s) [Python, Perl, or PHP]

Determine the Right Policies

Policies

Data Architecture: models, policies, rules, or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations *The Business Dictionary*

- Policies = *documented* data architecture
- Policies balance the inherent risks of conducting the analytics process against the payoff of executing it successfully.
- You can develop a list of the necessary policies for conducting the analytics process in your organization by analyzing each *function* and *gate* in the analytics process (diagram) and identifying:
  - *Who* will perform this function or close this gate
  - *Under what conditions* (when/where/how) can that function/gate be executed
  - *Who is responsible* to ensure that the function is performed correctly and efficiently
  - *Who is responsible* to provide the *resources* for performing that function

Most analytics failures are due to failures in policy vs. technical limitations. The key is *organizational understanding* of the analytics process.

What Success Looks Like

Characteristics of Rapid Analytic Development

- **Computationally lightweight:** should not require costly resources to execute routinely
- **Easy to understand and use:** end users should not require significant training to adopt
- **Easy to develop by a small team:** minimal programming languages and a team as small as one person
- **Easy to change:** small variations in versions should be simple to include and update
- **Compatible with other similar applications:** analytic solutions should, themselves, share data in a common format
- **Disposable:** resources to develop should be so low that walking away should have minimal sunk costs
- **Augment existing work flow:** applications should not make radical change their user's processes, this leads to poor adoption
- **Supplant analytic work flow:** applications should, if possible, seek to automate existing work flow(s) to free up valuable analyst time

Prioritize adaptability, situational awareness and speed over completeness.

Huddleston, Making the Analytics Leap

Listen to Lessons Learned

Analytic Implementation

(The Hard Part)

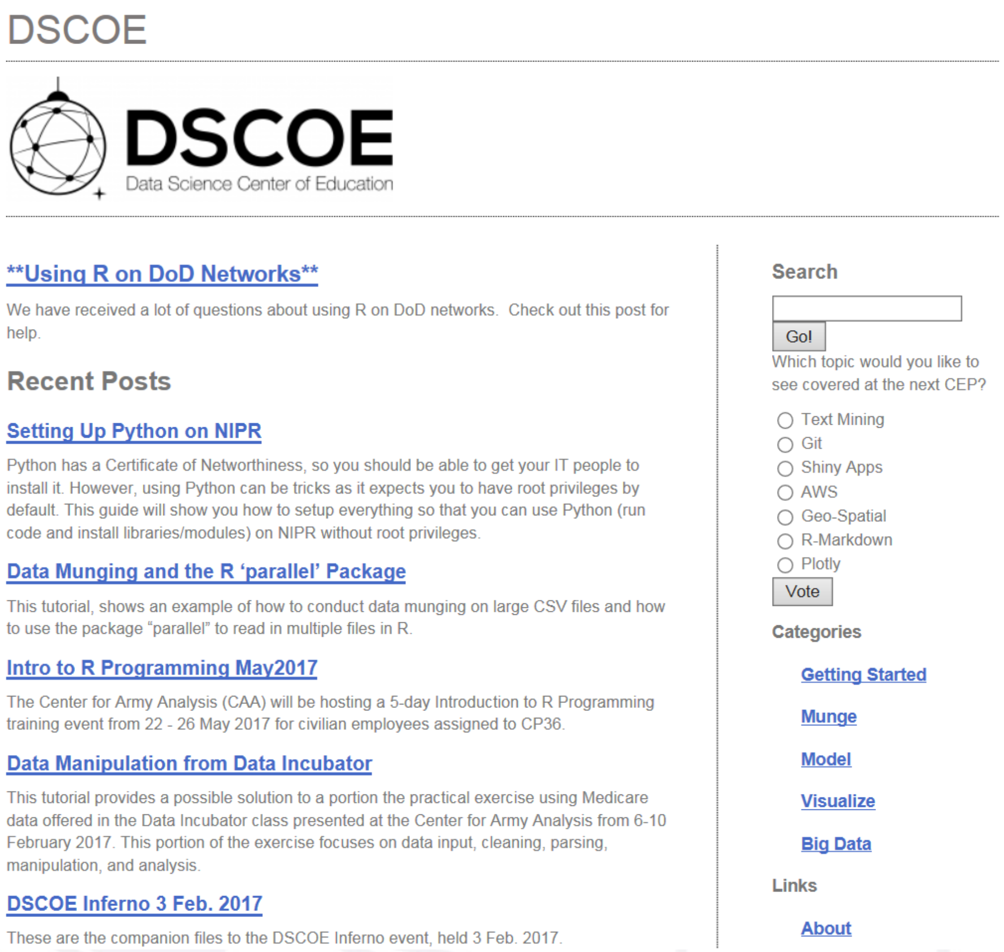
- If stakeholder mapping is not done correctly during problem definition, the tool/results may be ignored (who *wants* it vs. who *uses* it).
- The tool may highlight decisions that need to be made that the organization will not have considered (disruptive): "What do we do about this?"
- Think about how to integrate the tool into the organization's existing processes before the time comes (part of iterative problem definition).
- Your analytic tools may make it *very clear* that things are not going so well, inducing organizational defensiveness/hostility.
- Allocate time/resources for demos/training/user guide/documentation.
- Who is responsible for reviewing/updating/removing tools (analytic lifecycle management)?

Just because you build it doesn't mean they'll (want to) use it.

Huddleston, Making the Analytics Leap

Army Data Science Center of Excellence

- Army has established a web site for its data scientists
  - https://dscoe.army.mil
- Army worked through information assurance issues (e.g., certificate of networthiness for R statistical software tools)
- See Sponsor's Corner article in March 2017 PHALANX by Dr. Forrest Crain, CAA



Leadership Perspective

- Think end-to-end – data science requires a holistic approach
- Be patient in growing the capability
- Appreciate the challenges your analysts will face
- Work constructively with your IT organization to reap the benefits of the big data tools and activities of data science
- Understand your organization and its culture
- When in doubt, ask – there is no shortage of success stories from industry

NAVAL POSTGRADUATE SCHOOL

Key Takeaways in Creating a Data Science Capability

- Data Science capability can be built over time
  - Organizations mature in their data science capability in stages
  - Gains are found at every stage
  - Building an understandable, repeatable process for the organization
- Data Science is a team sport
  - Needs a broad view of the organization
  - Requires people, infrastructure, and policy alignment
- Leaders need to be advocates for:
  - Finding and removing the obstacles
  - Getting to the data
  - Connecting the disparate parts of the organization
  - Getting buy-in

BAH 2015 Field Guide to Data Science

Agenda

0830 – 0845 Introduction and Purpose

0845 – 1000 Data Science

- Motivation and Utility
- Definitions

1000 - 1015 Break

1015 – 1045 Context

- Big data
- Cloud Computing

1045 – 1145 Essential elements of a data science capability

1145 – 1200 Wrap up

1200 – 1300 Break

1300 – 1600 Small group discussions with any interested parties

- What specific problems lend themselves to trying data science solutions?
- How can we improve the Marine Corps' data science capability?
- How can NPS tailor its data science programs to better meet the needs of the Marine Corps?

Summary

- "Data science" is more than just doing better analysis – it is a mindset that sits the data scientist next to the decision maker and encourages interaction to solve problems rapidly!
  - It is a team sport
  - Provides actionable information without exposing decision-makers to underlying data or analysis
  - Data science is a repeatable process
- Open source data technologies (like cloud computing and Hadoop) provide the opportunity to combine different tools to deal with big data efficiently
- Data science capability is built over time
- As a leader and manager, you can create a data science capability within your organization

References

- Shaw, Paul, "Big Data for M&S" SPAWAR brief, Sep 2015.
- Huddleston, Sam, "What is Data Science" Center for Army Analysis Brief, 2016.
- Buttrey, Sam, "Data Curation Brief for NAVAIR" NPS OR Dept Brief, Apr 2016.
- Huddleston, Sam. "Making the Leap from Analysis to Analytics" Center for Army Analysis Brief, 2015.
- Rangachar, Ramesh, "Big Data and Hadoop" Aerospace Ground Systems Architecture Workshop, March 2013.
- Henningsen, Cavender, Muccio, McQuade, Herbranson, and Moore, "Big Data and Data Analytics," Phalanx, Mar 2014.
- Agile Data Science, Safari Books.
- The Field Guide to Data Science, 2nd ed, Booze Allen and Hamilton report, 2015.
- Kerr, S., Roettiger K., Vogel, S., "Data Analytics – From Data to Knowledge" Aerospace Ground Systems Architecture workshop Brief, Feb 2016.
- Schutt, R. and O'Neil, C., Doing Data Science O'Reilly, 2015.
- Perris, Wayne, "Data Science: An Elemental & Critical Component for Achieving Naval Information Dominance" ONR brief, 11 Feb 2014.
- Knopp, B., Beaghley, S., Frank, A., Orrie, R., and Watson, M., "Defining the Roles, Responsibilities, and Functions for Data Science Within the Defense Intelligence Agency," RAND Study, 2016.
- Porche, R., WilsonB., Johnson E., Tierney S., and Saltzman E., "Data Flood: Helping the Navy Address the Rising Tide of Sensor Information" RAND Study, 2014.
- National Academy of Sciences, " Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions" Oct 2016.