A Practical Approach to Data Science and Data Analytics

06 March 2019

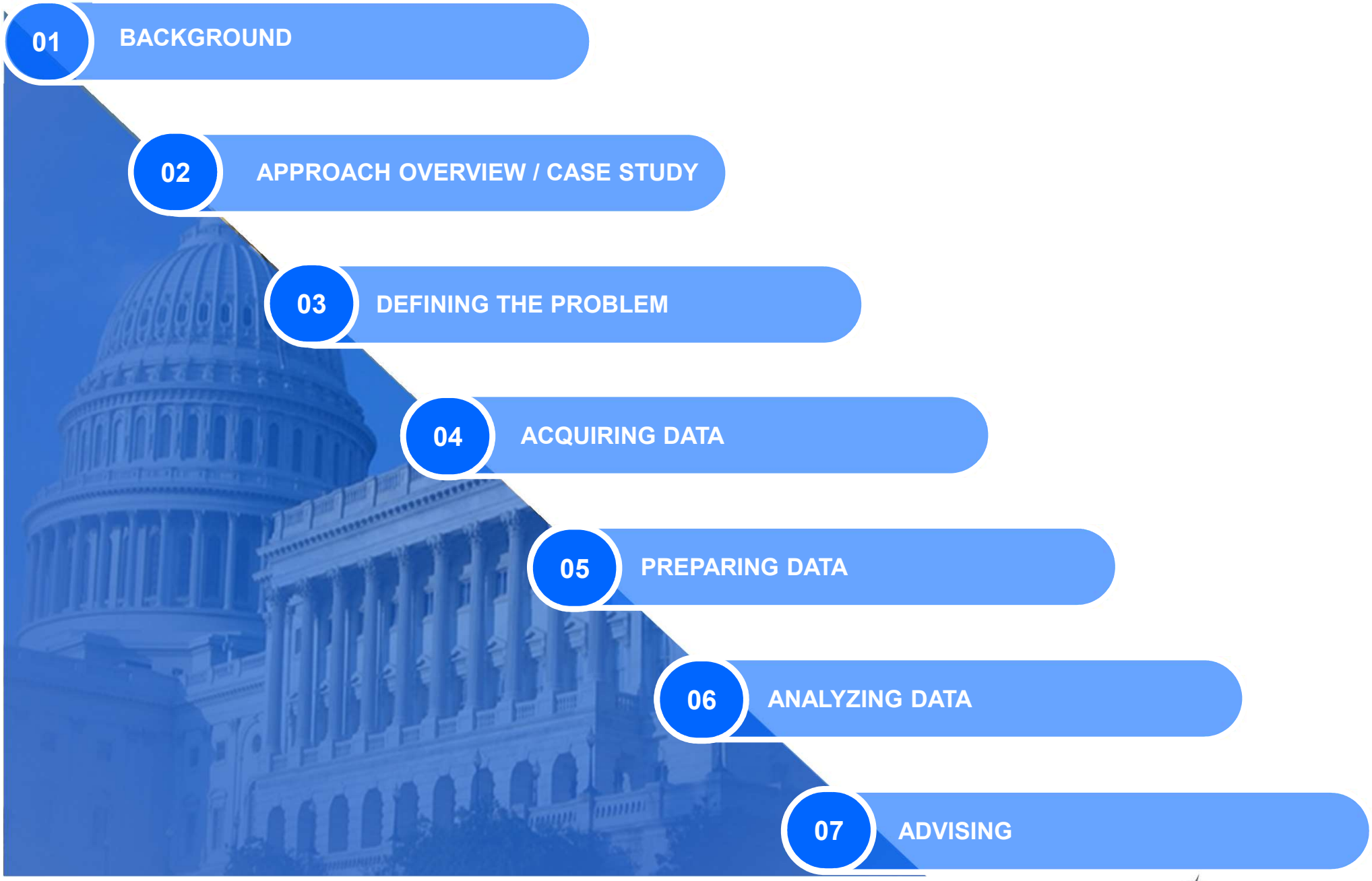
Tom Sullivan
Lead Associate | Herren Associates

Herren Associates, Inc.
Maritime Plaza II

1220 12th St SE, Suite 310
Washington, DC 20003

HERREN

A Practical Approach to Data Science and Analytics: Agenda



01 BACKGROUND

02 APPROACH OVERVIEW / CASE STUDY

03 DEFINING THE PROBLEM

04 ACQUIRING DATA

05 PREPARING DATA

06 ANALYZING DATA

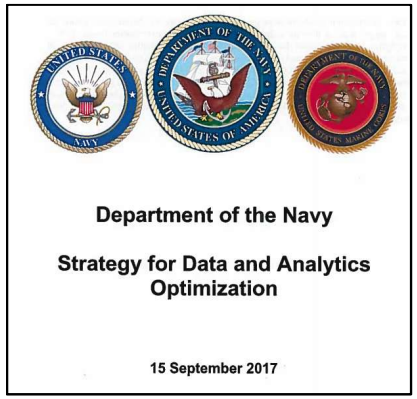
07 ADVISING

- Data science and data analytics fields are growing; whether in the commercial space, DoD fueling military readiness, or Academia with the growing presence of Universities offering master's degrees in Data Science

NETFLIX

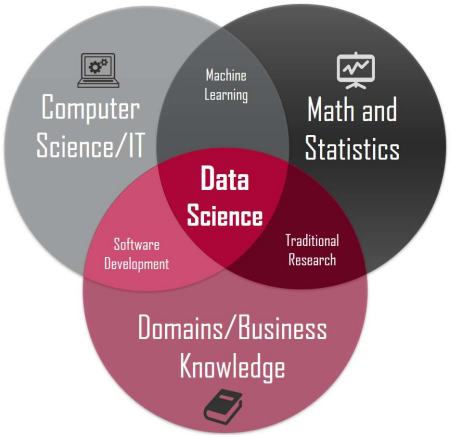
amazon Prime

UNIVERSITY OF VIRGINIA
DATA SCIENCE INSTITUTE



OPA Office of People Analytics (OPA)

- Data Science vs. Data Analytics, What is the Difference?
  - Data science and data analytics are unique fields, with the major difference being the scope and exploration
- Scope:
  - Data Science: Multidisciplinary field focused on finding insights by incorporating computer science, predictive analytics, statistics, and machine learning to parse through massive raw or unstructured data sets
  - Data Analytics: Encompasses branches of broader statistics and analysis to combine diverse sources of data and locate connections while simplifying the results



Source: Towards data science

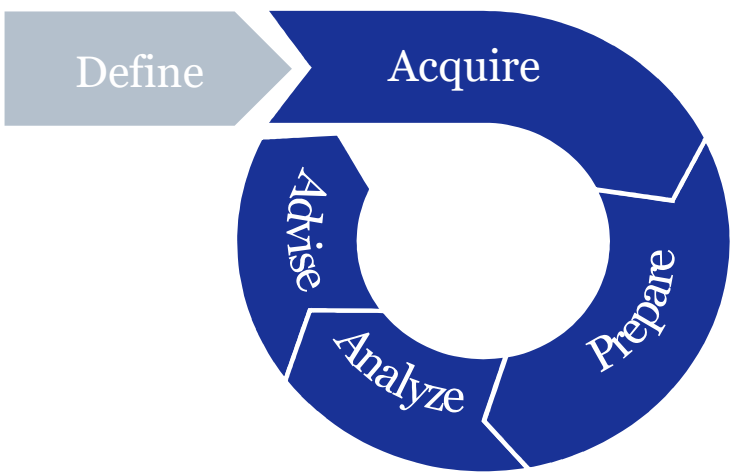
- Exploration
  - Data Science: Fixates on **unearthing answers to the things we don't know we don't know**
  - Data Analytics: Creates methods to capture, process, and organize data to **uncover actionable insights for current problems**, and establishing the best way to present this data

Data Scientists main goal is to ask questions and locate potential avenues of study, with less concern for specific answers and more emphasis placed on finding the right question to ask

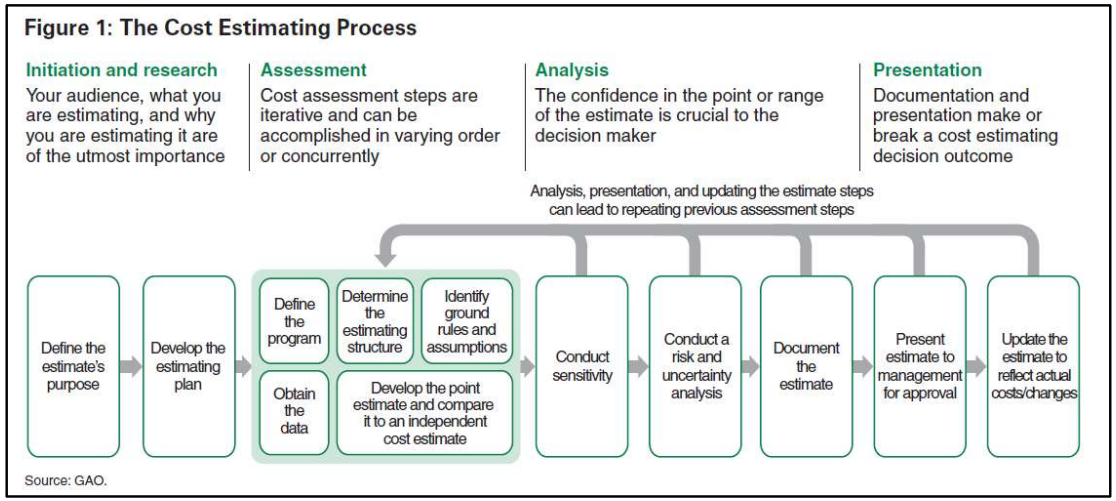
Source: SISENSE

APPROACH OVERVIEW / CASE STUDY

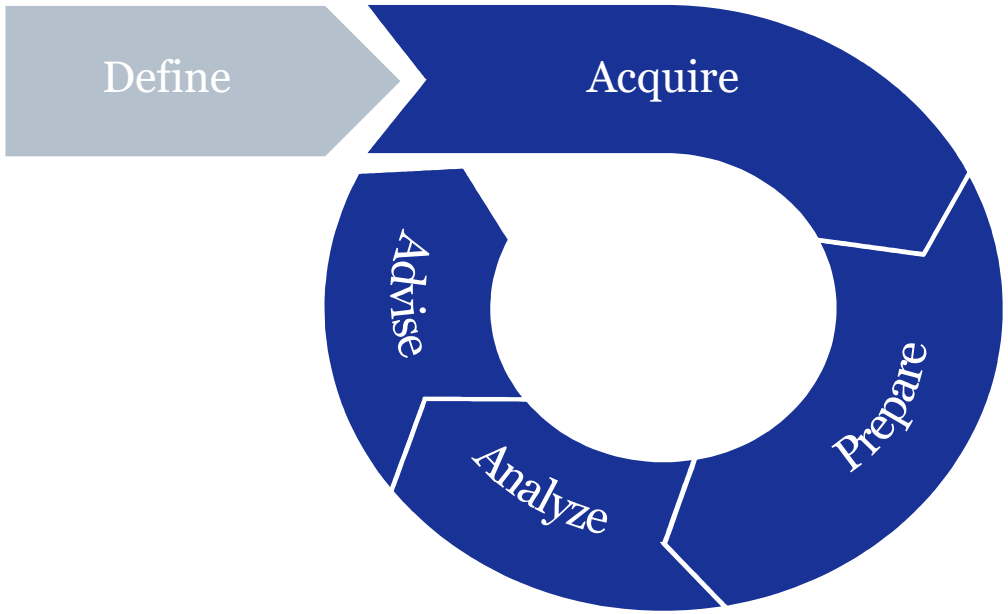
- Herren leverages a systematic and iterative approach to tackling data science and data analytics challenges including cost estimating
  - **Define:** Define the problem and data requirements plan
  - **Acquire:** Obtain necessary data
  - **Prepare:** Structure / Restructure data to fit analytic needs
  - **Analyze:** Interpreting data using statistical or analytical techniques
  - **Advise:** Present data to stakeholders in easily interpretable format



Herren Approach to Data Science and Data Analytics



- *Case Study: The Navy is reviewing their maintenance program and wants to project costs for a ship class. The Navy is interested in gaining better insight on the following cost drivers:*
  - *Port location where work is performed*
  - *Variance in costs of systems, and*
  - *Impact of contracting strategies*



DEFINING THE PROBLEM

- Define the Problem
  - Engage stakeholders to gain an understanding of their needs
    - What types of issues do stakeholders have?
  - Identify key opportunities
    - Solving which issue will make the greatest impact

- *Case Study Problem: The Navy wants to better understand their cost drivers*

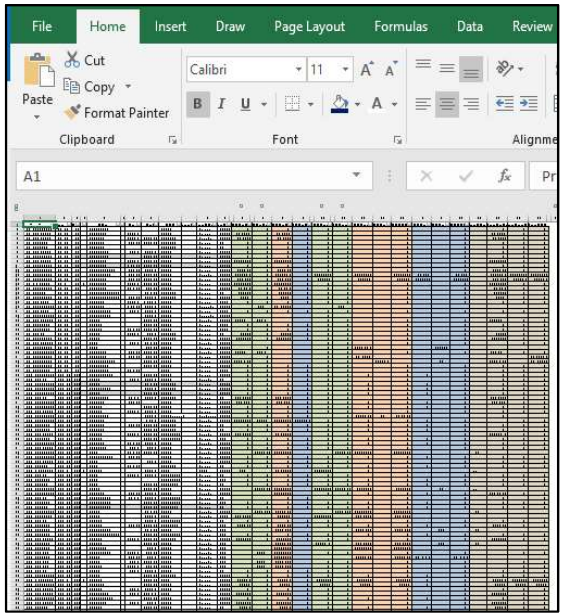
- Develop Hypothesis
  - Determine what you expect from your analysis
    - Is the issue worth pursuing? If so, use this as your defined hypothesis
- *Case Study Hypothesis: Port location of maintenance work is being done is the biggest cost driver for the availability*
  - What data would you need to gain an in-depth understanding of the issues?
    - Begin thinking about whether the data exists to solve your defined problem

ACQUIRING DATA

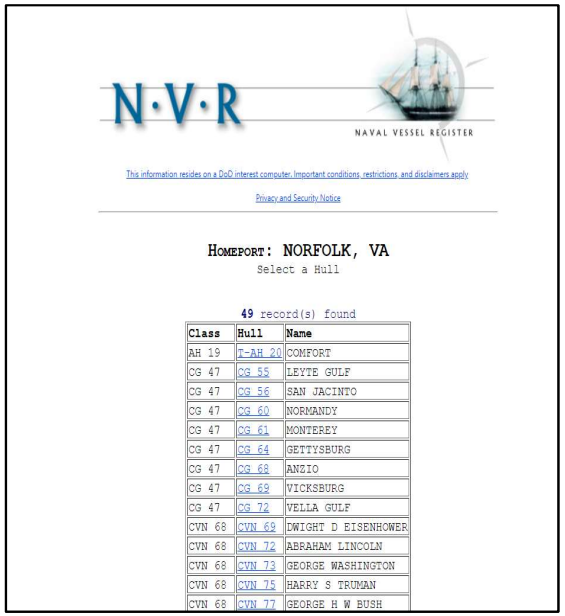
HERREN

- Explore the different sources and structures of data available and determine if they would help accept or refute the hypothesis
- Identify ground rules and assumptions
- Obtain all necessary data to see the whole picture
  - Structured Data: Data that resides in fixed fields
  - Semi-structured Data: Data that contains tags and other markers to separate data elements
  - Unstructured Data: Data that does not reside in fixed fields
- *Case Study: Leverage different data sources and structures for future analysis*

Structured:
Ex. Database Spreadsheets

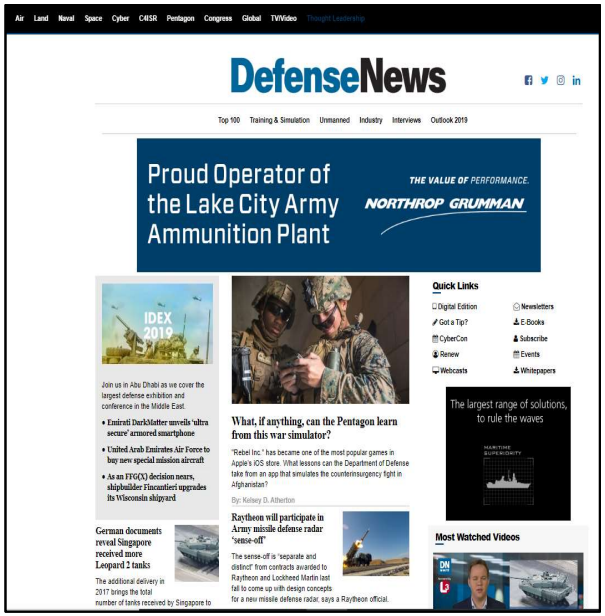


Semi – Structured:
Ex. XML or HTML Spreadsheets



Class	Hull	Name
AH 19	T-AH 20	COMFORT
CG 47	CG 55	LEYTE GULF
CG 47	CG 56	SAN JACINTO
CG 47	CG 60	NORMANDY
CG 47	CG 61	MONTEREY
CG 47	CG 64	GETTYSBURG
CG 47	CG 68	ANZIO
CG 47	CG 69	VICKSBURG
CG 47	CG 72	VELLA GULF
CVN 68	CVN 69	DWIGHT D EISENHOWER
CVN 68	CVN 72	ABRAHAM LINCOLN
CVN 68	CVN 73	GEORGE WASHINGTON
CVN 68	CVN 75	HARRY S TRUMAN
CVN 68	CVN 77	GEORGE H W BUSH

Unstructured:
Ex. Books, emails, untagged audio, images, and video



PREPARING DATA

- Data Validation: Process of ensuring data has undergone data cleansing so the quality of the data is useful and correct

7. Best Practices Checklist: Data

☐ As the foundation of an estimate, data

- ✓ Have been gathered from historical actual cost, schedule and program, and technical sources;
- ✓ Apply to the program being estimated;
- ✓ Have been analyzed for cost drivers;
- ✓ Have been collected from primary sources, if possible, and secondary sources as the next best option, especially for cross-checking results;
- ✓ Have been adequately documented as to source, content, time, units, assessment of accuracy and reliability, and circumstances affecting the data;
- ✓ Have been continually collected, protected, and stored for future use;
- ✓ Were assembled as early as possible, so analysts can participate in site visits to understand the program and question data providers.

☐ Before being used in a cost estimate, the data were

- ✓ Fully reviewed to understand their limitations and risks;
- ✓ Segregated into nonrecurring and recurring costs;
- ✓ Validated, using historical data as a benchmark for reasonableness;
- ✓ Current and found applicable to the program being estimated;
- ✓ Analyzed with a scatter plot to determine trends and outliers;
- ✓ Analyzed with descriptive statistics;
- ✓ Normalized to account for cost and sizing units, mission or application, technology maturity, and content so they are consistent for comparisons;
- ✓ Normalized to constant base-year dollars to remove the effects of inflation, and the inflation index was documented and explained.

Source: GAO Cost Estimating and Assessment Guide

HERREN

- Data Re/structuring: If data isn't cleanly structured, it can inhibit analysis
  - Increasing efficiencies in data structuring decreases the overall amount of data analysis
  - Determine what fields are necessary and data flexibility needs
  - Typical data cleaning activities include outlier checking, data parsing, null values, and value imputation
    - All of these are made easier with well structured data
  - Precursor to relational databases & analysis tools (R, Python, Access)
    - Work best with data that is in a specific structure
  - Real-world data is almost never in the correct form for analysis. As such, it requires structing beforehand

Color Key	
Manual	
Pick List	
Calculated	

Selected Date

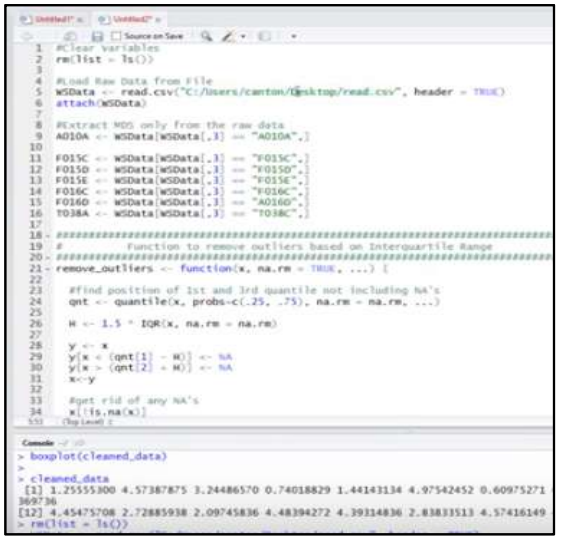
									1/1/2018	2/1/2018	3/1/2018	4/1/2018
						$ 44,791,734	$ -	$ -	$ -	$ -	$ -	$ -
Fiscal Yea	TPS Year	TPS Number	Project	Owner	Performing Activit	Oct-17	Nov-17	Dec-17	Jan-1	Feb-1	Mar-1	Apr-1
FY17	FY17	TI12H_SIM2	Facilities	David Bateman	NSWC DD	$ -	$ -	$ -				
FY17	FY17	TI16_CDS00	Facilities	David Bateman	LM	$ -	$ -	$ -				
FY17	FY17	TI16_CPS00	Facilities	David Bateman	DRS TECH	$ -	$ -	$ -				
FY18	FY18	20P0_DDS00	Facilities	David Bateman	CDSA DAM NECK	$ -	$ -	$ -				
FY18	FY18	20P0_ATD00	Facilities	David Bateman	CDSA DAM NECK	$ 150,000	$ -	$ -				
FY18	FY18	20P0_CPS00	Facilities	David Bateman	DRS TECH	$ -	$ -	$ -				
FY18	FY18	20P0_CDS00	Facilities	David Bateman	LM	$ -	$ -	$ -				

80% of data analysis is spent cleaning and preparing data – before any analysis is done!

HERREN

Ship Classes	FY17	FY18	FY19	FY20	FY21	FY22	FY23	Total
CVN	2	1	2	2	2	1	1	11
CG 47	3	3	4	5	3	1	3	22
DDG 51	9	10	9	10	9	10	9	66
LCS	1	2	2	2	3	3	2	15
SSBN	4	4	4	4	4	4	4	28
LHA 6	0	0	0	1	0	0	0	1
LHD 1	1	1	1	1	2	1	1	8
LPD 17	2	1	2	1	1	2	2	11
LSD 41	1	2	1	1	1	1	1	8
LSD 49	1	0	1	0	0	1	1	4
MCM 1	1	1	2	2	2	2	1	11
AS 39	1	0	1	0	0	0	0	2
ESB	0	1	0	0	0	0	0	1

Data for illustrative purposes and does not reflect fielding profile

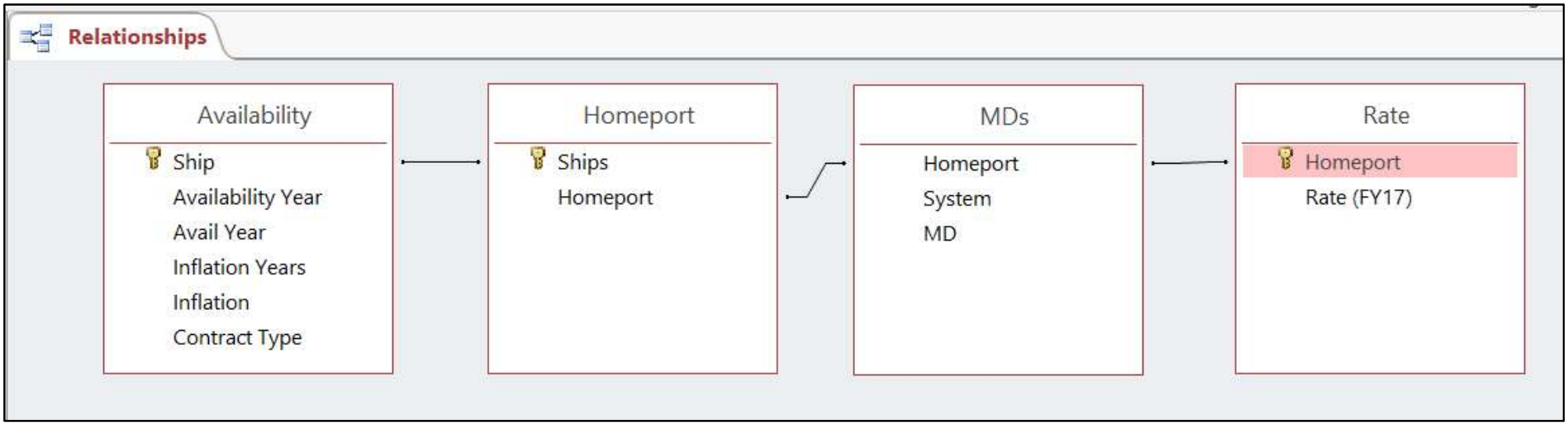


R script to transform data

- **Problem: Fiscal years are used as the header of the table.**
- The "messy" data presented above is often used for presentation purposes.
- Often used for displaying information over time (previous financial management dashboard example)
- More easily readable for humans, but is less useful for computers and software
- **Solution:** "melt" columns into rows

Ship Classes	Year	Availabilities
CVN	FY17	2
CVN	FY18	1
CVN	FY19	2
CVN	FY20	2
CVN	FY21	2
CVN	FY22	1
CVN	FY23	1
CG 47	FY17	3
CG 48	FY18	3

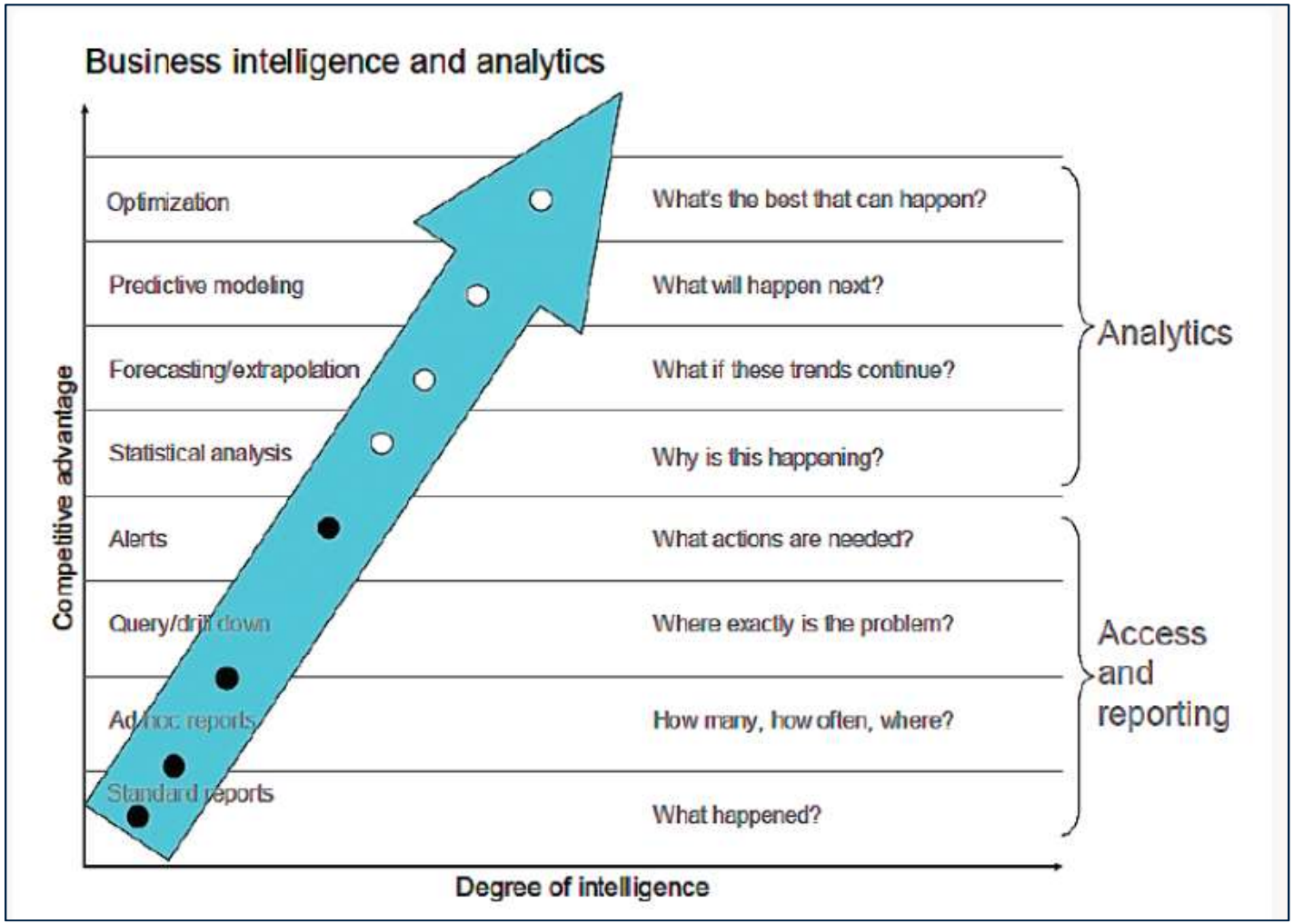
- **Relational Database –** Utilize numerous tables with relationships connecting all information and allowing simple extraction and reporting of information across tables
  - Reduced data redundancy
  - Scales well – stores large amounts of data efficiently
  - Powerful data manipulation enabled using SQL
- *Case Study: Leverage structured data sets in tables to form relationship allowing for queries and more efficient analysis*



Relationships illustrated in Microsoft Access

ANALYZING DATA

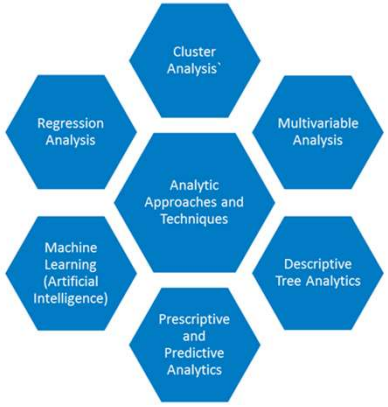
- Data Analytics: Davenport and Harris characterize data analytics as beginning with statistical analysis ("Why is this happening") and enhancing a competitive advantage up through predictive modeling and optimization

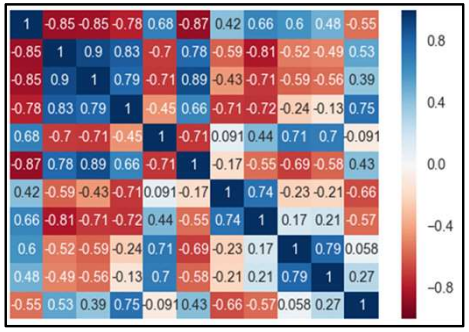


Source: Davenport and Harris (2007): *Competing on Analytics: The New Science of Winning.* Boston: Harvard Business Review Press

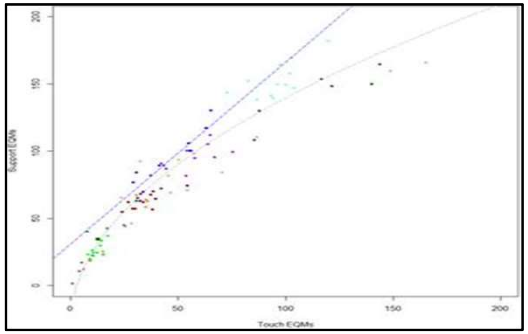
HERREN

- Numerous analytical approaches and techniques exist and are dependent on the previously defined problem and data available
- Numerous tools exist that support the efficiency and effectiveness of analytical approaches and techniques; including, but not limited to:
  - Python: Interpreted, object-oriented, high-level programming language with dynamic semantics
  - R: Free software environment for statistical computing and graphics
- ***Case Study: Leverage Python and R to analyze correlations of variables impacting cost drivers and predict maintenance costs***
  - Leverage Python modules and packages, to run data correlations efficiently across large data sets and multiple variables
  - Leverage R in support of the proposal evaluation process, including linear and non-linear regressions, and sensitivity analysis





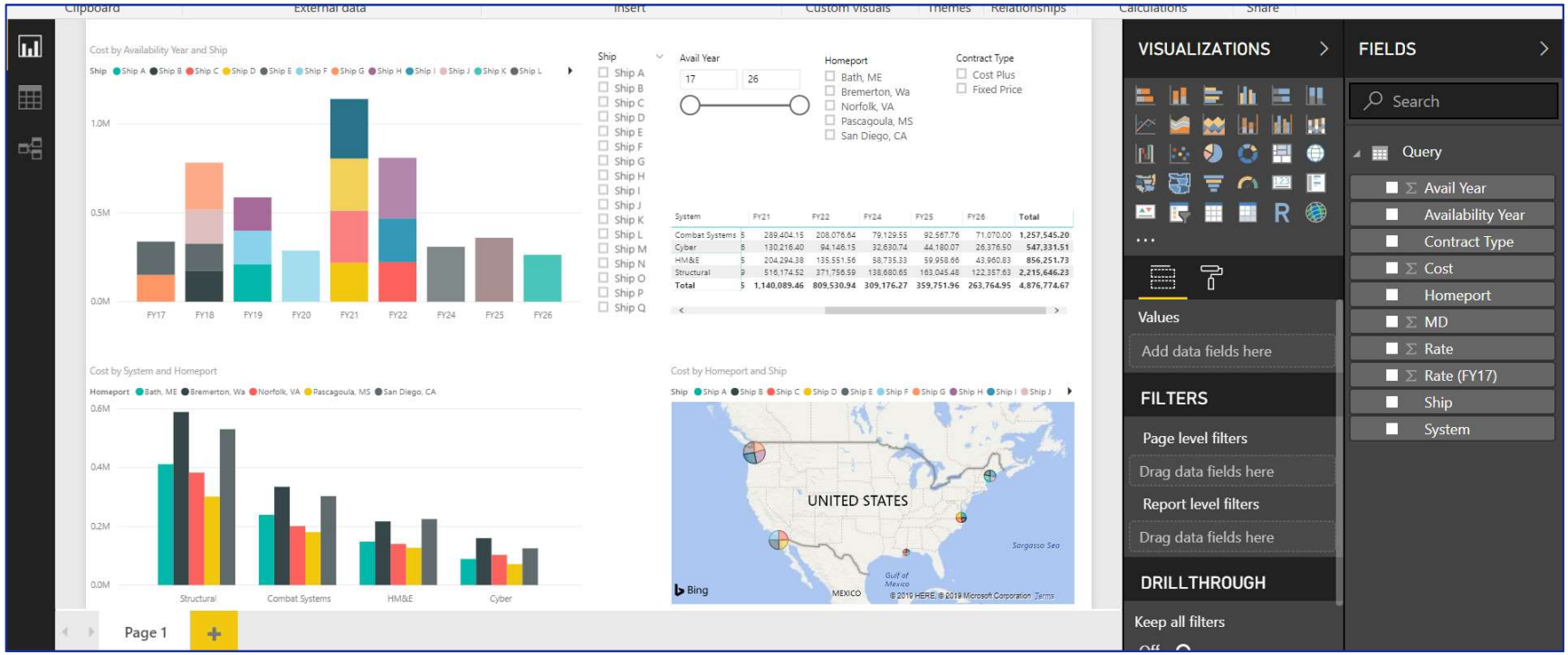
Python



R

ADVISING

- Summarize your findings and conclusions
  - Present data to stakeholders in easily interpretable format
  - Develop charts and graphs that support your thesis
  - Consolidate your findings into a report for the stakeholders
- *Case Study: Data Visualization*
  - *Link insights to financial and operational metrics to show the impact*
  - *Ensure your results align with the stakeholder's strategy*
  - *Provides real-time scenario analysis and collaboration with stakeholders*



Microsoft PowerBI

QUESTIONS