ICEAA Workshop 2016

Integrating Cost Estimating and Data Science Methods in R

Josh Wilson and Laura Barker

Atlanta, GA
June 2016

Outline

- Purpose / Overview
- Definition of Terms
- Cost Estimating Challenges
- Example – Installation Cost Analysis in Excel and R
- Benefits (Pros)
- Limitations (Cons)
- Conclusions
- Way Forward

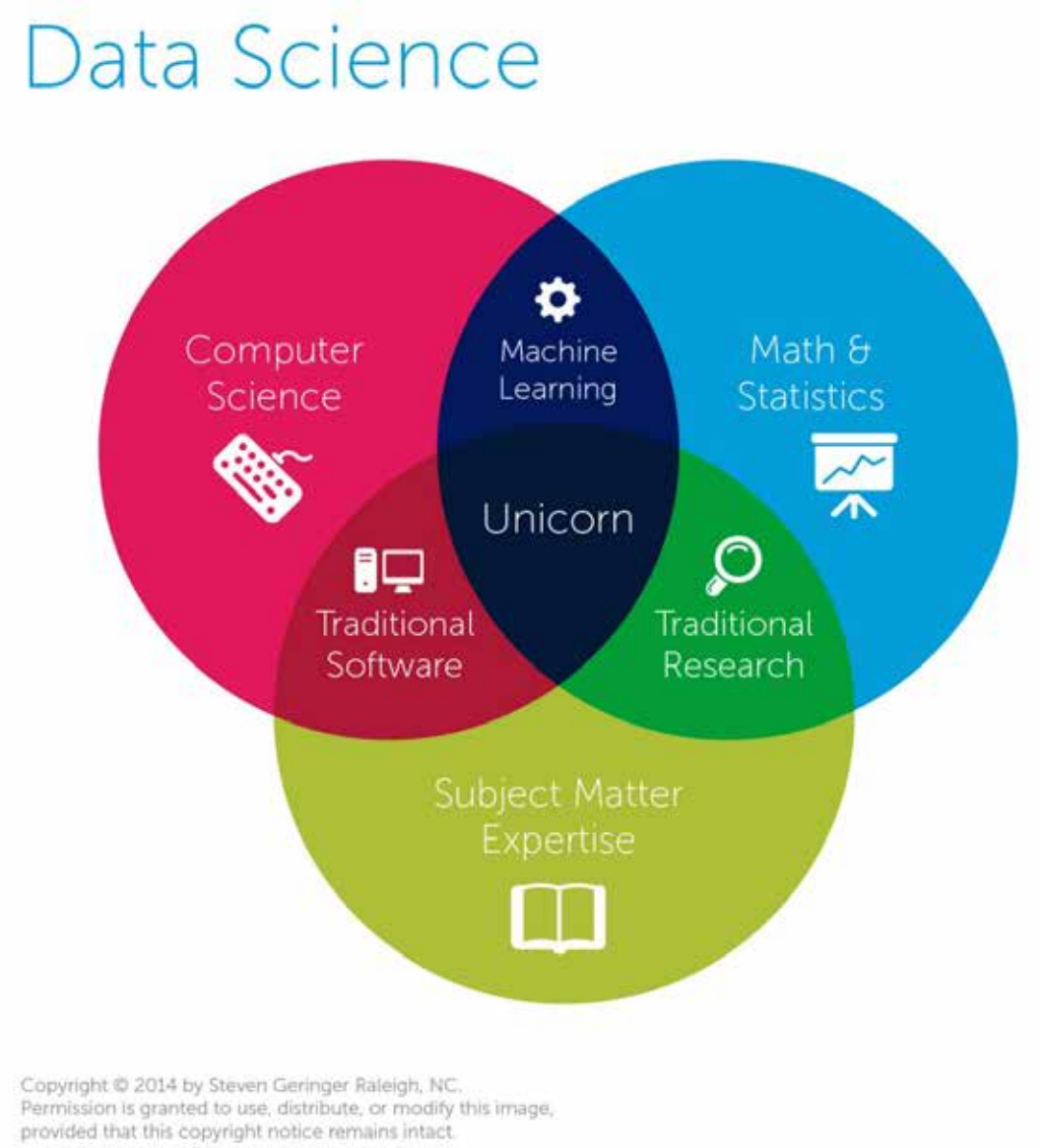
Purpose / Overview

- Data science tools can help address some basic, persistent cost estimating challenges
  - Documentation
  - Traceability
  - Reproducibility
  - Reusability
- Particularly applicable to data collection and analysis stages of the cost estimating process
- Usage of data science tools also opens the door to more advanced data science techniques

What is Data Science?

+

No, seriously…

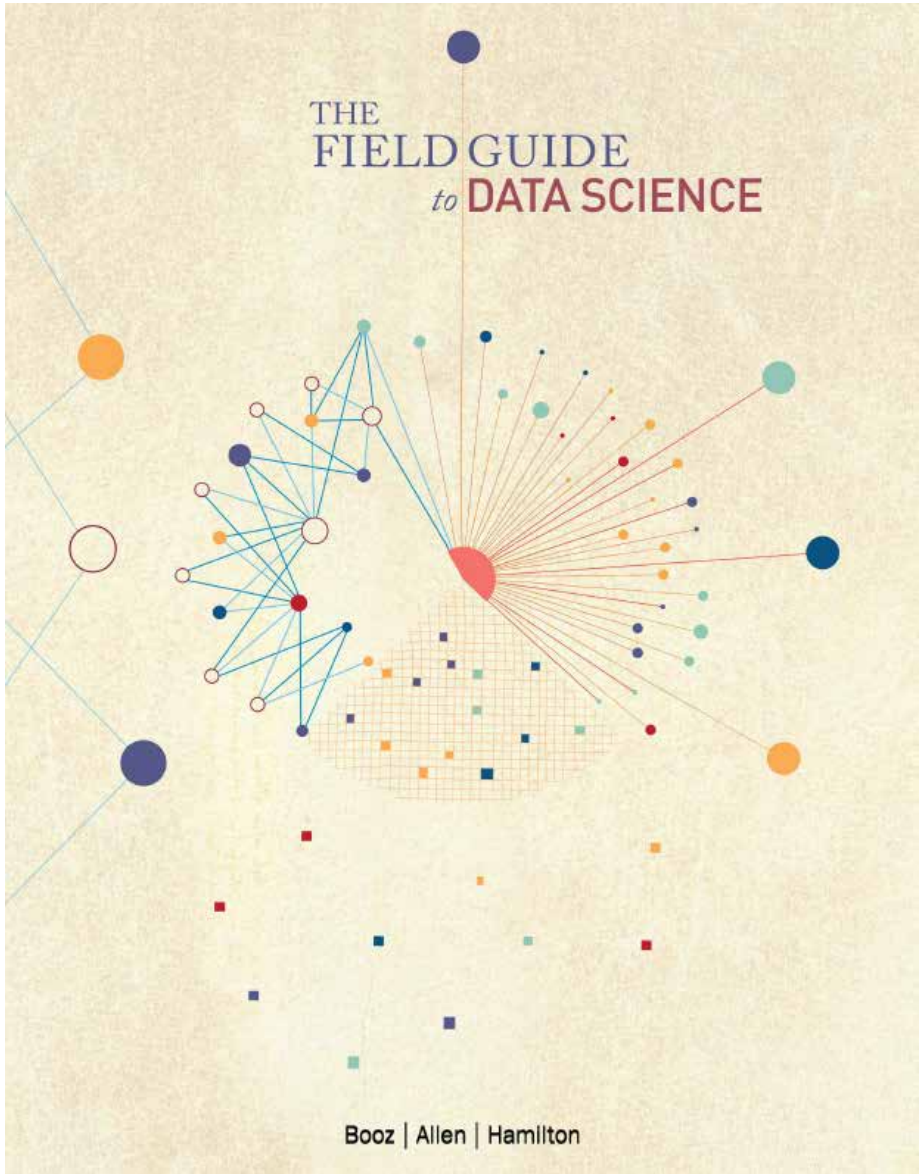


What *is* Data Science?

- "Describing Data Science is like trying to describe a sunset – it should be easy, but somehow capturing the words is impossible… Data Science is the art of turning data into actions." – The Field Guide to Data Science

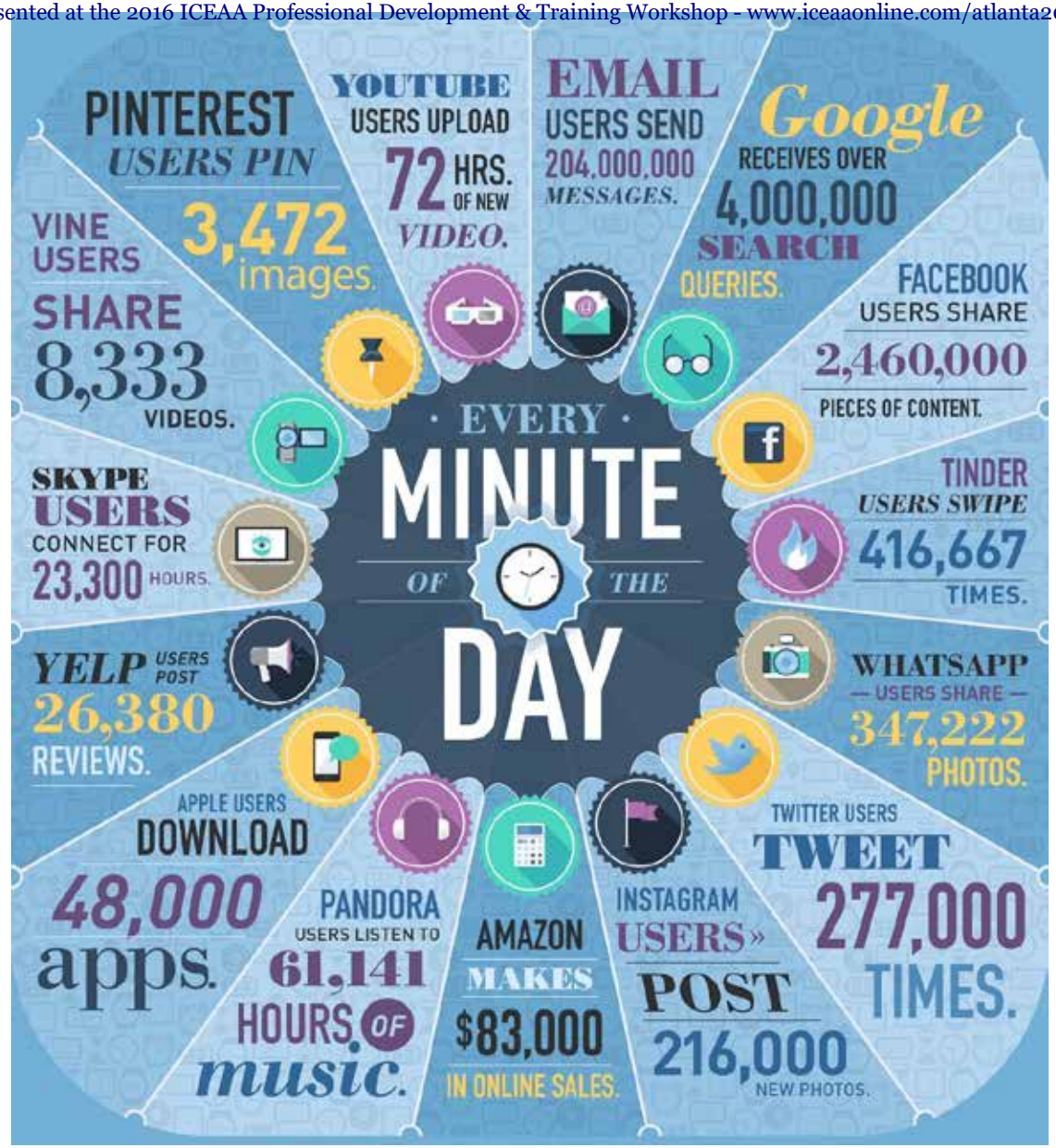
- "The key word in 'Data Science' is not Data, it is Science." – Jeff Leek

- "I think data-scientist is a sexed up term for a statistician… Statistics is a branch of science." – Nate Silver



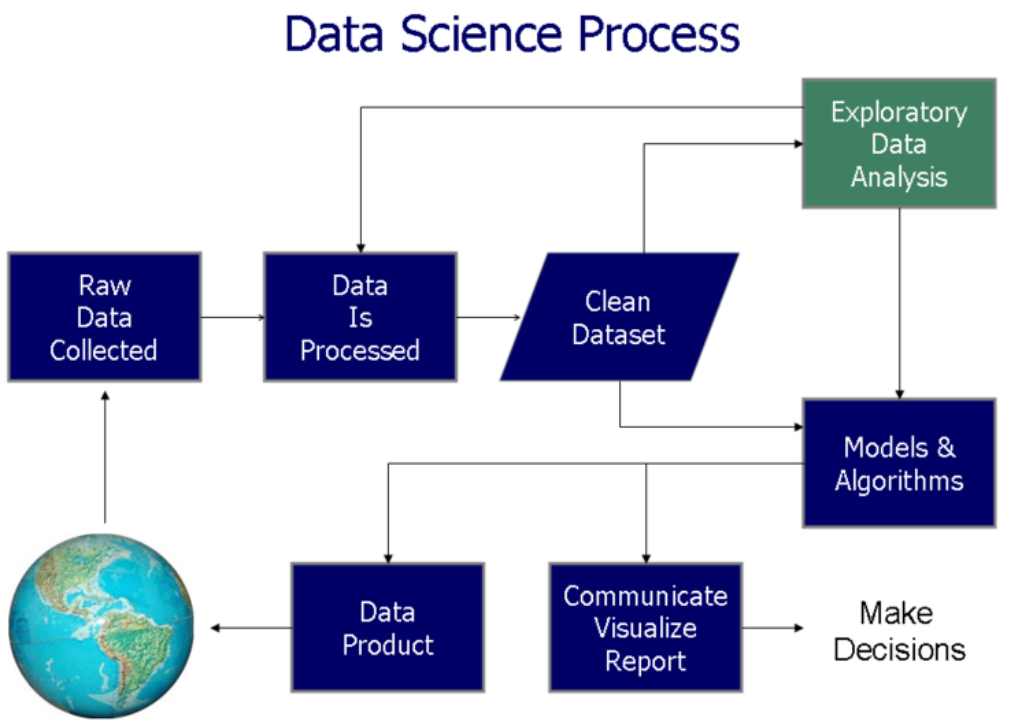
Why do you keep hearing about Data Science?

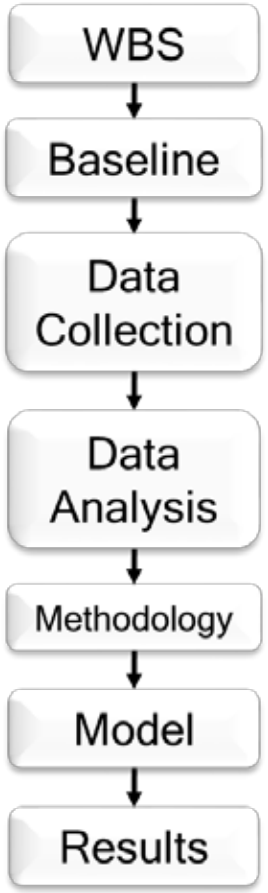
- POP QUIZ: 90% of the world's data has been created in the past ___ years.
- Data infrastructure advances
  - Everything is connected (Internet of Things)
- Commoditization of computing power
  - Data collection is getting easier and cheaper



Why are *WE* talking about Data Science?

- Data science is about gaining insight from data
- Data science emphasizes:
  - Documentation
  - Traceability
  - Reproducibility
  - Reusability
  - Statistical rigor
- Data science process aligns with parts of the cost estimating process





Cost Estimating Challenges

4 Documentation

- Set of explicit instructions to define data sources and analysis steps
- Data collection and analysis process is inherently difficult to document

4 Traceability

- Ability to follow and understand the process used to perform an analysis
- Limited by poor documentation

4 Reproducibility

- Ability to perform an analysis again and get the same result
- Limited by lack of documentation and traceability

4 Reusability

- Ability to apply a completed analysis to a new situation without having to redo everything

How can Data Science enhance Cost Estimating?

Documentation – Set of instructions to define data sources and analysis steps

4 Traditional cost estimating processes and methods:

- 4 Documentation is an extra step to the analysis process
- 4 Various levels of quality
  - Source data and analysis may be included within a cost model (as an Exhibit or Appendix)
  - Source data and analysis may be in a workbook external to main cost model
  - External analysis model could have been lost
- 4 Any required Excel add-ins or macros are not typically identified

4 Cost estimating supplemented by Data Science:

- 4 Analysis scripts (code) are self-documenting
- 4 Imported packages are clearly identified
- 4 Additional documentation features are available
  - R Markdown
  - R Pubs
  - Slidify
  - Python Sphinx

How can Data Science enhance Cost Estimating?

Traceability – Ability to follow and understand the analysis process

4 Traditional cost estimating processes and methods:

- 4 Heavily dependent on having a clean model structure and good model documentation
- 4 Never ending string of Excel precedents / dependents
- 4 Lots of details are easy to miss
  - Custom filters
  - Ad hoc data cleaning
- 4 Excel Stats package output is hardcoded

4 Cost estimating supplemented by Data Science:

- 4 Scripted analysis steps are clearly defined in a sequential manner
- 4 Many popular data science tools are open-source
  - Python
  - R

How can Data Science enhance Cost Estimating?

Reproducibility – Ability to perform an analysis again and get the same result

4 Traditional cost estimating processes and methods:	4 Cost estimating supplemented by Data Science:
4 Ability to reproduce analysis dependent on quality of model documentation and traceability 4 Requires manual reproduction of analysis steps 4 Access to Excel add-ins may cause issues	4 Simply requires re-running the script 4 Required packages imported and loaded automatically

How can Data Science enhance Cost Estimating?

Reusability – Ability to easily apply a completed analysis to a new situation

4 Traditional cost estimating processes and methods:

- 4 Reuse of analysis in Excel requires reproduction of effort
  - Re-application of filters
  - Re-navigation of stats package GUIs
  - Potential manual changing of formula references
- 4 Visuals usually need to be created based on specific data

4 Cost estimating supplemented by Data Science:

- 4 Once created, a scripted analysis is readily applied to new data and/or inputs
- 4 Less "rework" required to reuse analysis script
  - Change source data
  - Change inputs
- 4 Automatically generates visuals

EXAMPLE

Question: “Is the cost to install a system different on various ship classes?”

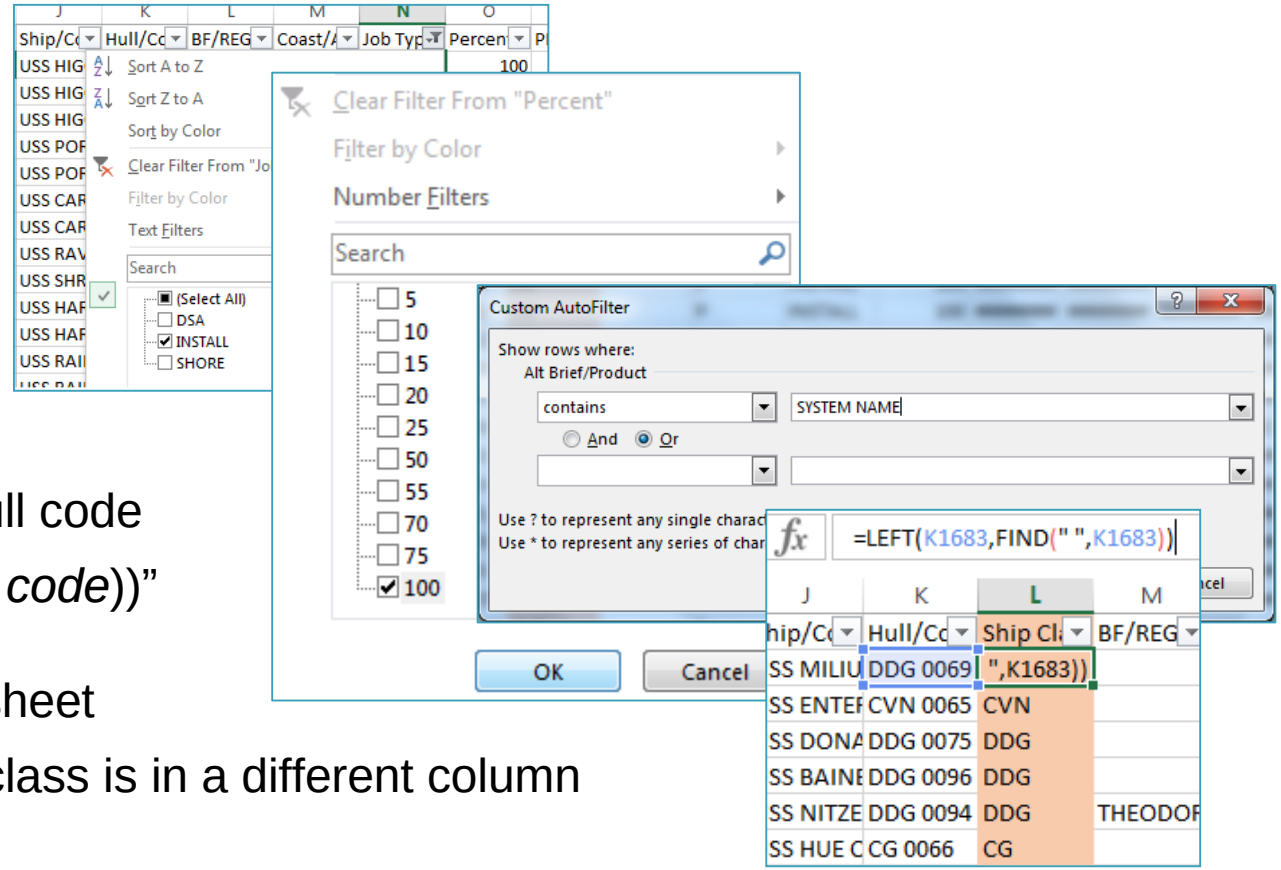
Example: Analysis of Install Cost Data in Excel

1. Obtain raw data pull

2. Apply filters
   - "Job Type" equals "INSTALL"
   - "Percent" equals "100"
   - "Product" contains text "SYSTEM NAME"

3. Add column to extract ship class info from hull code
   - Use formula "LEFT(*hull code*, FIND(" ", *hull code*))"

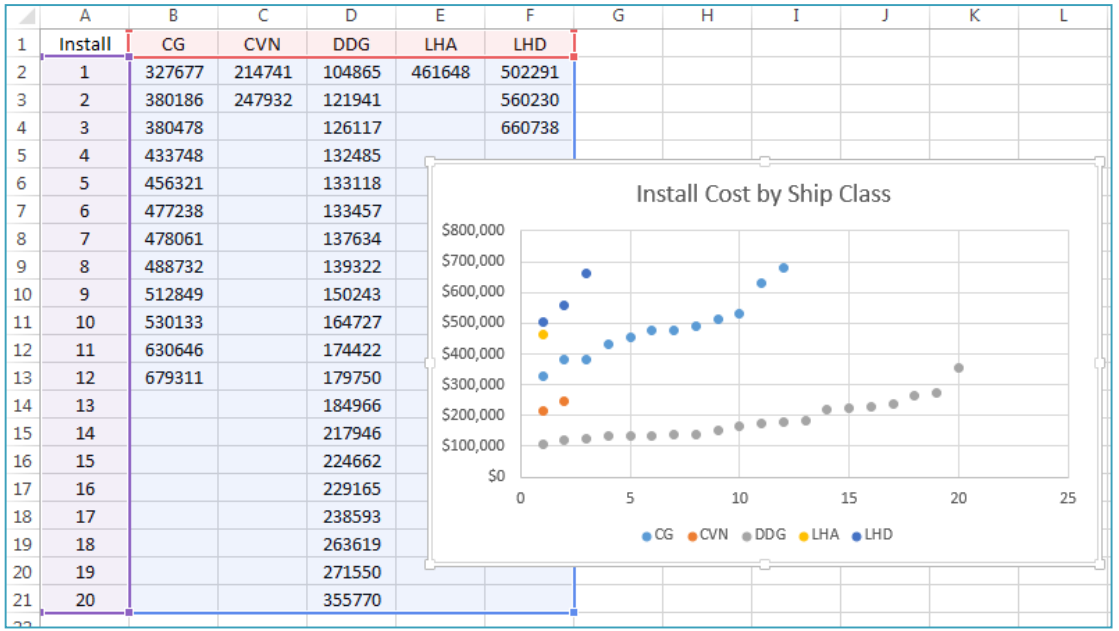
4. Manually copy and paste values to another sheet
   - Manipulate data so that data for each ship class is in a different column
   - Necessary to simplify generation of graphs



Example: Analysis of Install Cost Data in Excel (cont'd)

5. Scatter plot install cost
   - Separate ship class columns put each ship class in an individual data series
   - Note that some GUI navigation steps here were skipped / i.e. not documented

6. Formulate hypotheses based on visual inspection of data
   - Install costs for DDG and CVN appear indistinguishable
   - Install costs for CG, LHA, LHD also appear indistinguishable
   - Install costs for {CG, LHA, LHD} > {DDG, CVN}

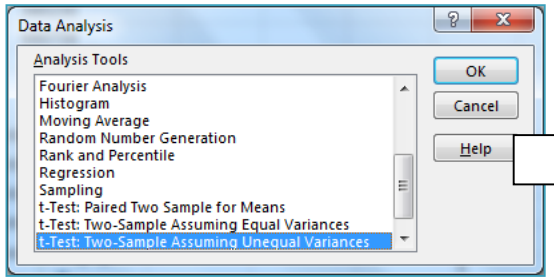


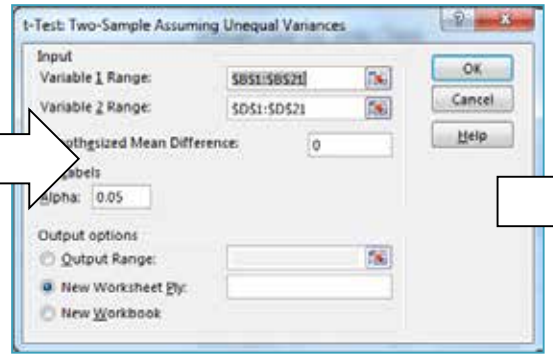
	A	B	C	D	E	F
1	Install	CG	CVN	DDG	LHA	LHD
2	1	327677	214741	104865	461648	502291
3	2	380186	247932	121941		560230
4	3	380478		126117		660738
5	4	433748		132485		
6	5	456321		133118		
7	6	477238		133457		
8	7	478061		137634		
9	8	488732		139322		
10	9	512849		150243		
11	10	530133		164727		
12	11	630646		174422		
13	12	679311		179750		
14	13			184966		
15	14			217946		
16	15			224662		
17	16			229165		
18	17			238593		
19	18			263619		
20	19			271550		
21	20			355770		

* DDG = Destroyer, CVN = Carrier, CG = Cruiser, LHA / LHD = Types of Amphibious Assault Ships

Example: Analysis of Install Cost Data in Excel (cont'd)

7. Hypothesis testing via multiple t-tests
   - Need (N*(N+1))/2 individual pairwise tests to compare N ship classes (5 ship classes = 10 comparisons)
   - Below is t-test of CG vs. DDG cost





	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*CG*	*DDG*
4	Mean	481281.6667	184217.6
5	Variance	10168099331	4124384208
6	Observations	12	20
7	Hypothesized Mean Difference	0	
8	df	16	
9	t Stat	9.152090711	
10	P(T<=t) one-tail	4.64306E-08	
11	t Critical one-tail	1.745883676	
12	P(T<=t) two-tail	9.28612E-08	
13	t Critical two-tail	2.119905299	

8. Determine conclusion – two install cost inputs are needed
   - CVN / DDG
   - CG / LHA / LHD

Example: Analysis of Install Cost Data in Excel (cont'd)

4 Questions the cost analyst may be asked during presentation of results

- What is the supporting data?
- What were the steps of the analysis?
- Does the analysis process make sense?
- Can the analysis results be reproduced?
- Can the analysis results be easily updated when new data is available?
- Can the analysis be applied to other systems?

- Documentation
- Traceability
- Reproducibility
- Reusability

Example: Analysis of Install Cost Data in Excel (cont'd)

- Documentation issues
  - Lots of small, manual steps are easy to overlook
  - GUI navigation steps require screenshots
  - Required add-ins and macros not typically identified

- Traceability issues
  - Copying and pasting specific values loses any reference to source data
  - Hardcoded t-test results eliminate visibility to test input data

- Reproducibility issues
  - Difficult due to documentation and traceability issues

- Reusability issues
  - Requires re-filtering, re-copying and pasting values, re-navigation of stats GUIs
  - Manual changes to update plots

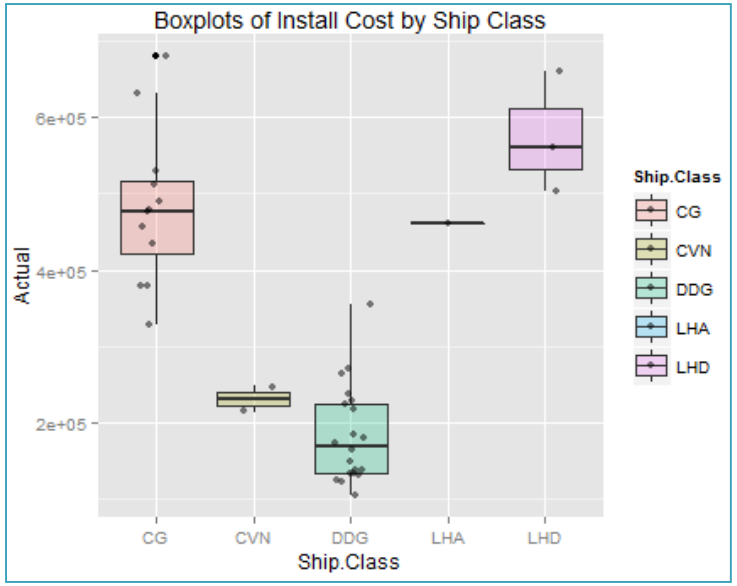
Example: Analysis of Install Cost Data in R

- What is R?
  - Free, open source programming language
  - Used for data manipulation, data analysis, statistical modeling, data visualization, and predictive modeling
  - Users include data scientists, statisticians, analysts
  - Has gained substantial popularity in recent years
  - Easily extended via packages
- Analysis in R is a set of instructions
  - Script to read, manipulate, and analyze data

```
1  ###### INSTALL / LOAD PACKAGES ######
2  if(!require("dplyr"))
3        install.packages("dplyr")
4
5  if(!require("ggplot2"))
6        install.packages("ggplot2")
7
8  ###### READ DATA ######
9
10 ## change wd to the folder in which you have "SPIDER_san2.csv" saved
11 ## ("SPIDER_san2.csv" is the source data file)
12
13 wd <- "C:/Users/534300/Desktop/Work Files/Data Science/DSWG/Install Ana
14 setwd(wd)
15 all_data <- read.csv("SPIDER_san2.csv") %>% tbl_df
16
17 ###### FILTER DATA ######
18
19 ## filter inputs
20 filter_job_type <- "INSTALL"
21 filter_percent <- 100
22
23 ## product names have been sanitized by replacing with NES game titles
24 filter_product <- "Mega Man"
25 # filter_product <- "The Addams Family"
26 # filter_product <- "The Incredible Crash Dummies"
27 # filter_product <- "M.U.L.E."
28 # filter_product <- "Kung-Fu"
29
30 ## filter data
31 filtered_data <- all_data %>%
32       filter(Job.Type == filter_job_type) %>%
33       filter(Percent == filter_percent) %>%
34       filter(grepl(filter_product, Alt.Brief.Product))
35
36 ## add ship class from hull code
37 hull_code <- filtered_data$Hull.Code
38 hull_code2 <- strsplit(as.character(hull_code), " ")
39 hull_code3 <- sapply(hull_code2, function(x) x[1])
40 filtered_data$Ship.Class <- hull_code3
41 remove(hull_code, hull_code2, hull_code3)
42
43 ###### EXPLORATORY PLOTS ⟺
72
73 ###### HYPOTHESIS TESTING ⟺
```

Example: Analysis of Install Cost Data in R (cont'd)

- Documentation
  - Source data clearly identified
  - Imported packages clearly identified
  - All steps clearly documented
- Traceability
  - Less than 100 lines of code
  - R is open-source (as is Python)
- Reproducibility
  - Re-running script on source data reproduces the analysis
- Reusability
  - Only requires changing source data and/or filter parameters
  - Automates generation of plots and t-test results



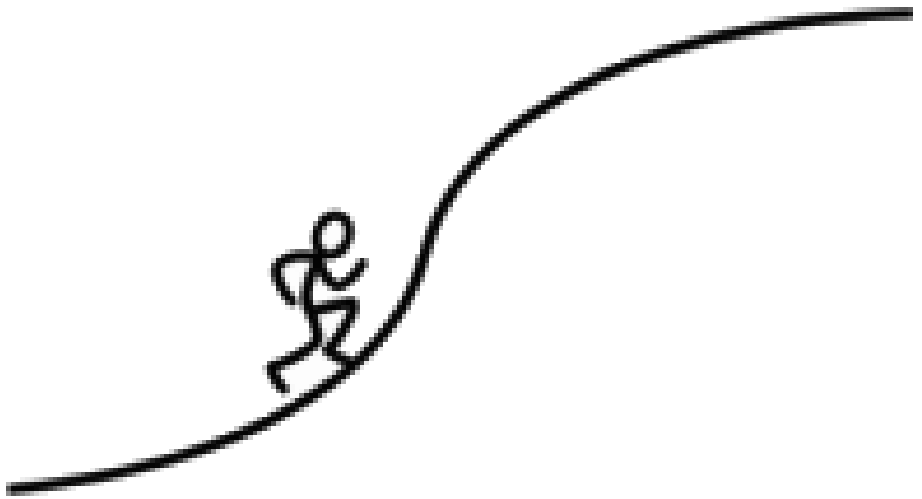
	Ship_Class.1	Ship_Class.2	p_value
1	DDG	CG	7.394639e-08
2	DDG	LHD	8.599988e-03
3	CVN	CG	4.013784e-05
4	CVN	LHD	1.109978e-02

Benefits

- Code is self-documenting
- Additional documentation features are available (R markdown, R pubs, Slidify, Sphinx)
- Some of the most popular data science tools are open source (R, Python)
- Function code in open source languages can be inspected
- Traceability is maximized, making data analysis process more transparent
- Reproducibility is automatic once the analysis script is complete
- Reusing an analysis requires less effort
- More time can be spent improving existing analysis scripts in lieu of performing same analysis on new data
- Robust statistical functionality is built-in
- Opens the door to more advanced analysis techniques

Limitations

- Learning curve
- Unfamiliarity with tools may impact ability to communicate analysis process
- Commonality of Excel versus data science tools
- Every cost estimate is unique
- Garbage in, garbage out



Conclusions & Way Forward

- Data science can be applied anywhere there is data, including cost estimating
- Data science tools and practices can be used to address some common cost estimating challenges
- Data science tools will not replace cost estimating tools or methods, but are more appropriate in some cases
- Utilizing data science tools opens the door to advanced analysis techniques
  - Clustering algorithms to identify truly analogous programs or efforts
  - Textual analysis of contractor monthly status reports to predict cost overruns
  - k-NN (k Nearest Neighbors) to address missing info in messy data sources
  - Interactive and dynamic visuals (R Shiny package)

Questions?

Josh Wilson
Associate

Booz | Allen | Hamilton

Booz Allen Hamilton Inc.
1615 Murray Canyon Road
Suite 900
San Diego, CA 92108
Tel (619) 278-3855
Mobile (619) 820-6226
wilson_joshua@bah.com

Laura Barker
Senior Consultant

Booz | Allen | Hamilton

Booz Allen Hamilton Inc.
1615 Murray Canyon Road
Suite 900
San Diego, CA 92108
Tel (619) 680-4857
Mobile (770) 990-6905
barker_laura@bah.com

BACKUP

What is R?

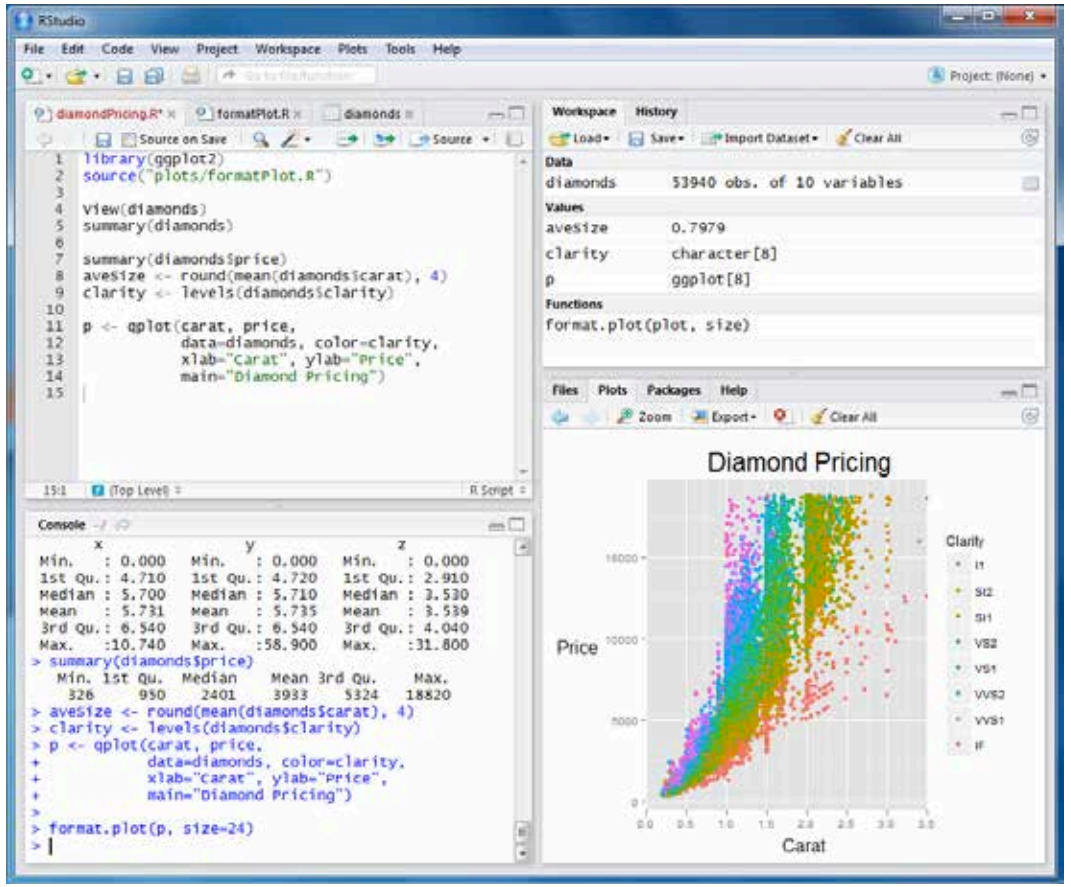
- R is a free, open source programming language and software environment for statistical computing and graphics

- The R language is widely used among statisticians and data miners for developing statistical software and data analysis, and has gained substantial popularity in recent years

What is RStudio?

- RStudio is a free Integrated Development Environment (IDE) for R that has added functionality and more user-friendly interface

- Includes:
  - console
  - syntax-highlighting editor that supports direct code execution
  - easier file navigation
  - tools for plotting, history, debugging, and workspace management



Why choose R over other Data Science tools?

- Free
- Open source
- Large user base
- (Relatively) easy to learn
- Built-in tools and packages for statistical analysis
  - Linear and nonlinear modelling
  - Classical statistical tests
  - Time-series analysis
  - Classification
- Data visualization capabilities