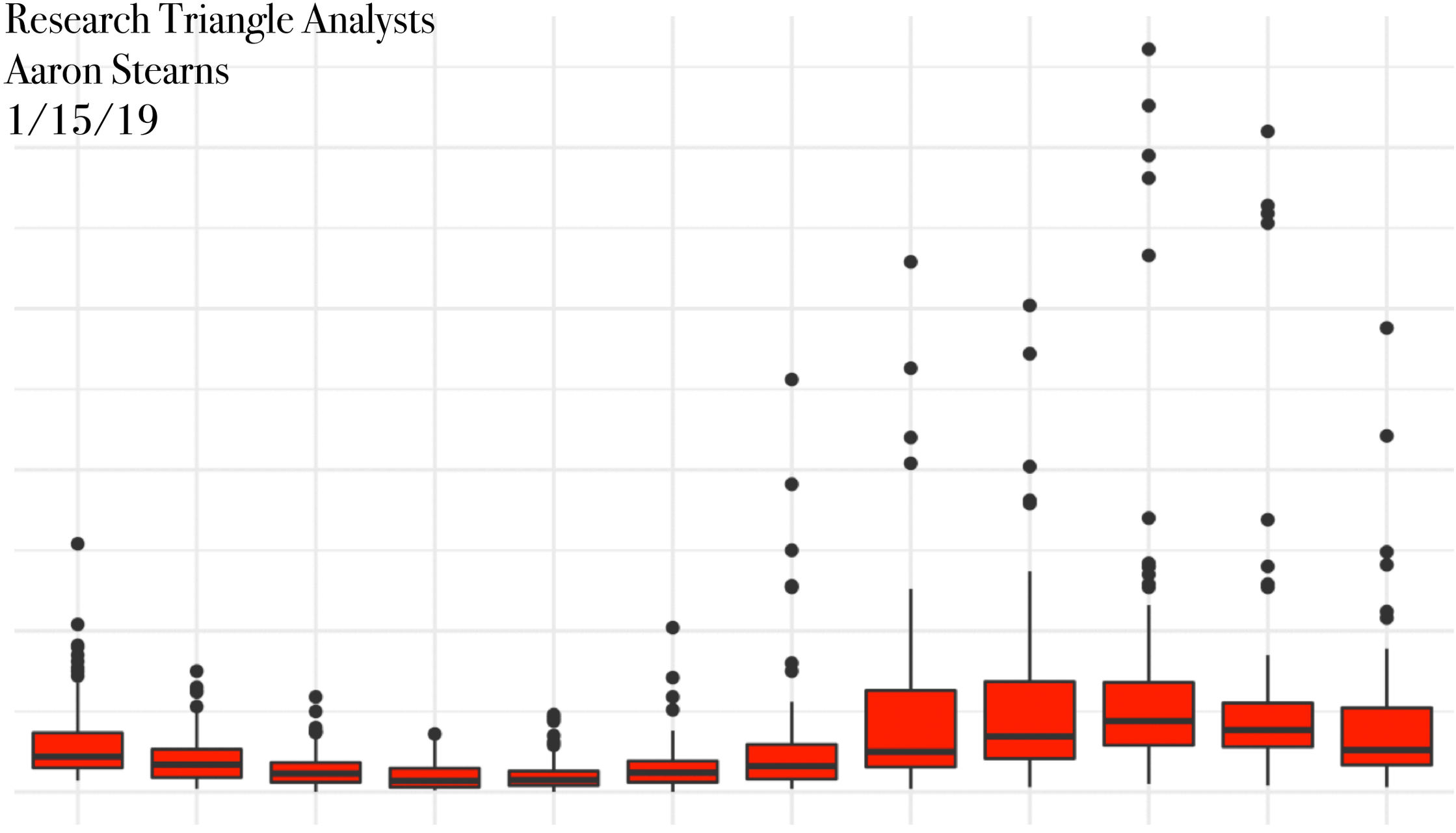
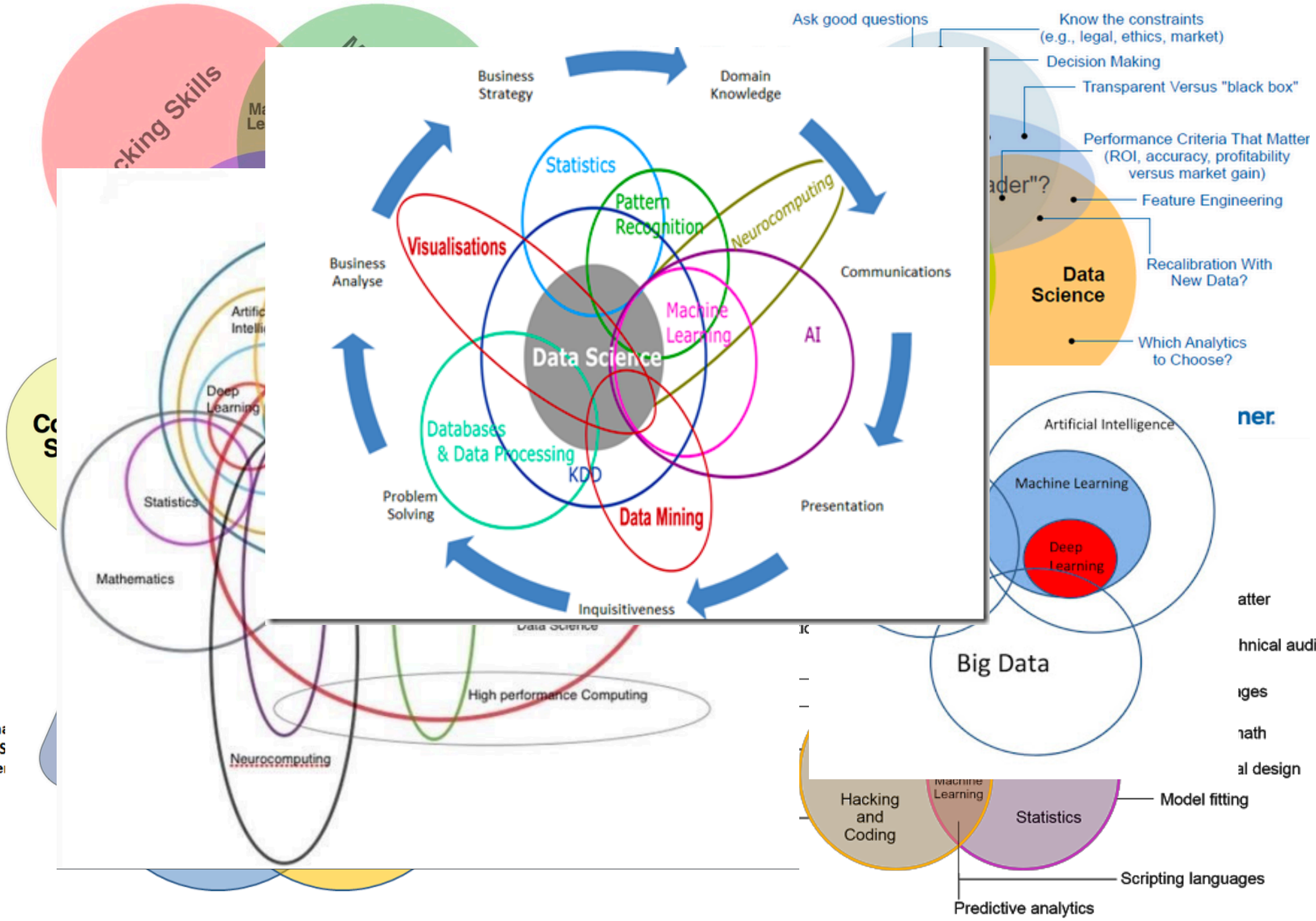
Getting Started With Data Science

Research Triangle Analysts
Aaron Stearns
1/15/19

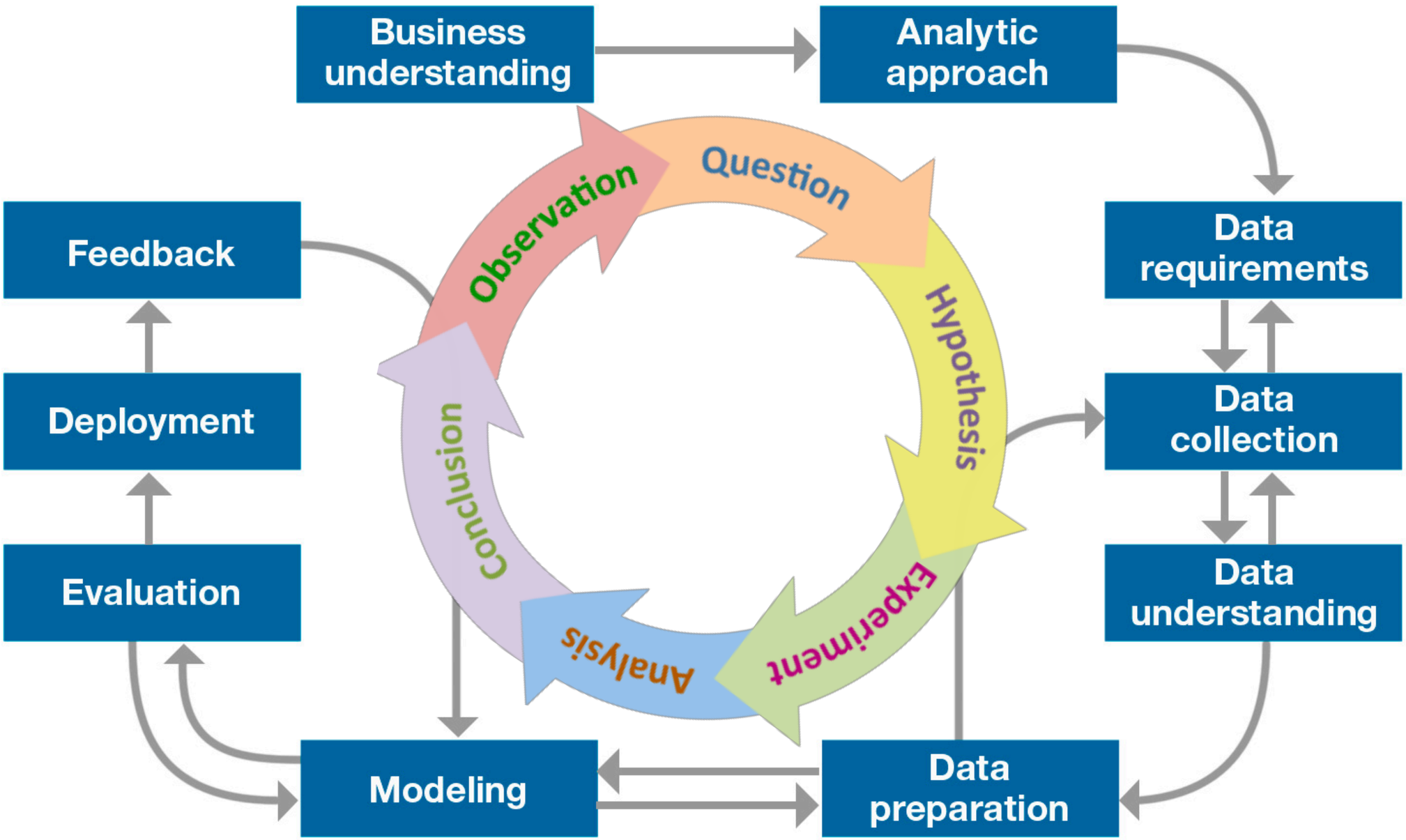


Data Science: Defined?



In clockwise order: 1) niemanlab.org 2) https://stats.stackexchange.com/users/1352/stephan-kolassa 3) https://blogs.gartner.com/christi-eubanks/three-lessons-crossfit-taught-data-science/ 4) Joel Grus 5) Dan F-M 6) https://www.gartner.com/marketing/ 7) Gregory Piatetsky-Shapiro 8) Tweeted by Mara Averick 9) https://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html

Data Science Methodology



Outside: https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science Inside: https://en.wikipedia.org/wiki/Scientific_method

Getting Started

Start programming (poorly)

```
x = True
y = 4
z = 12.56
z1 = ["yes", "no"]
```

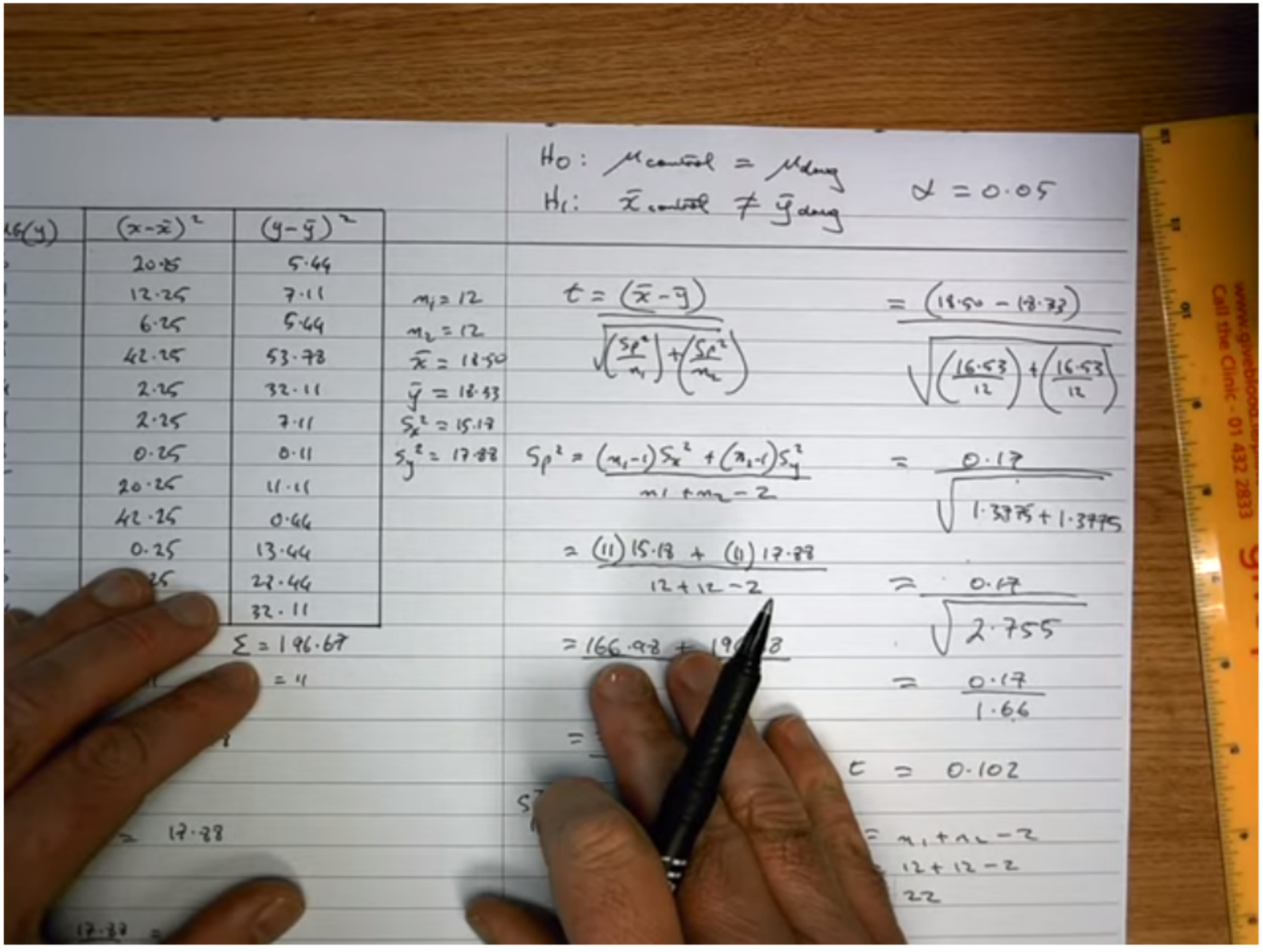
```
def function1(foo, bar):
    if not foo != bar:
        print("foobar")
    elif bar != foo:
        print("barfoo")
```

```
text = ""

for book in corpus:
    for chapter in book:
        for sentence in chapter:
            for word in sentence:
                for char in word:
                    text += char
```

$O(n^?)$

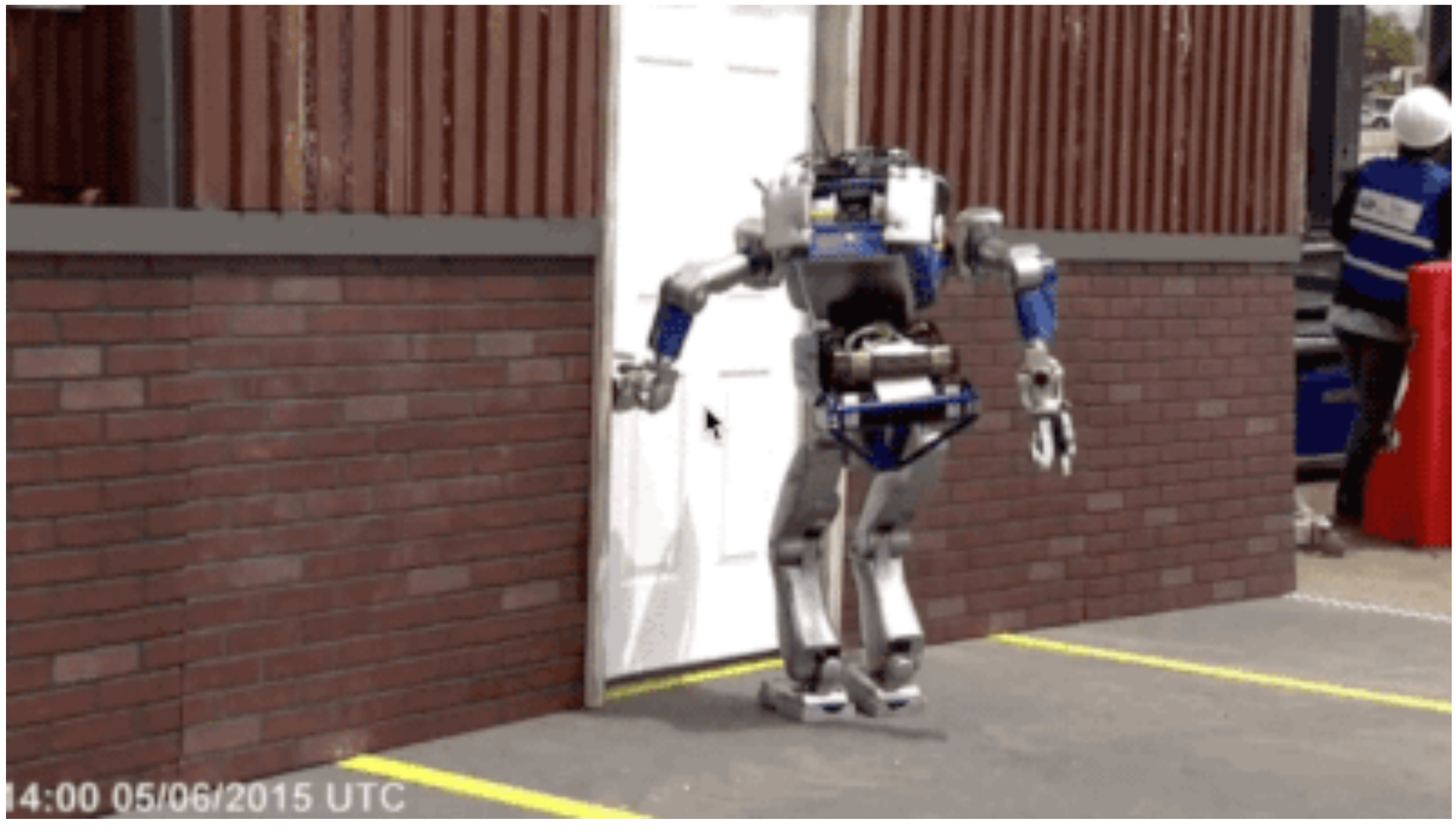
Brushing Up On Your Stats



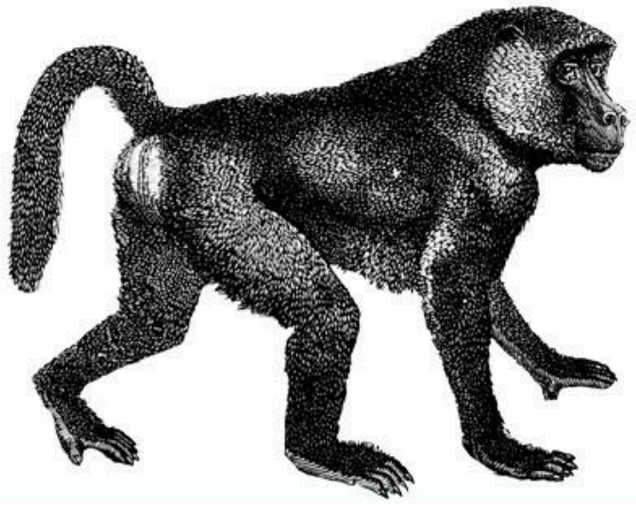
Eugene O’Loughlin's [video lecture series]

- Pick a language and start practicing
- Read other people's code
- Start learning some theory and how it is applied
    - Find some data that interests you
        - Some ideas:
            - Download your checking account transaction history and investigate your purchasing patterns.
            - Download your text messages / Facebook messages and learn about your messaging habits
            - Find an 'open data portal' and investigate the goings-on in your area
                - Durham: https://live-durhamnc.opendata.arcgis.com/
                - Cary: https://data.townofcary.org/pages/homepage/
                - Raleigh: https://data-ral.opendata.arcgis.com/
                - Chapel Hill: https://www.chapelhillopendata.org/page/home1/

Machine Learning



Geeks México



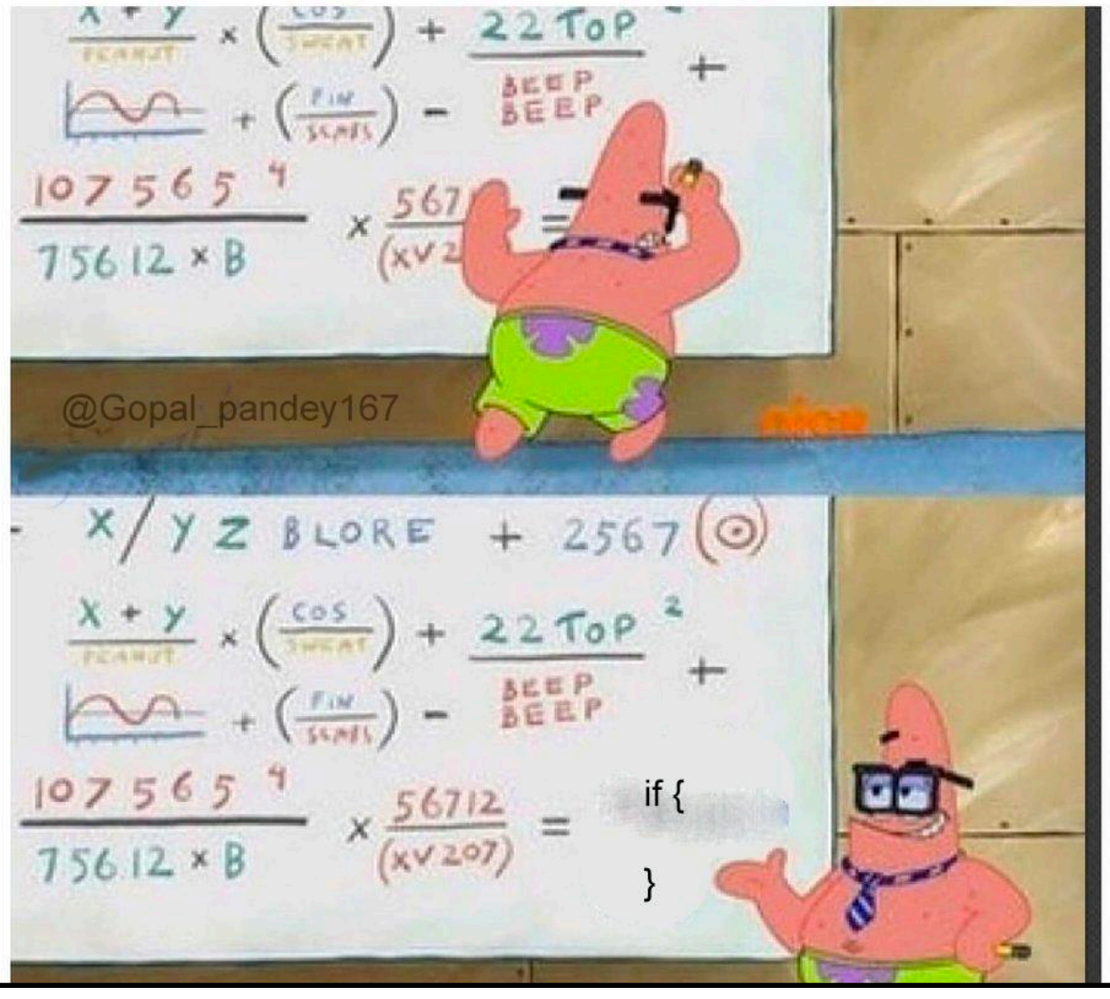
AI based on if / else statements

The Definitive Guide

O RLY?

@raidentrance

When you are making AI



I Am Devloper @iamdevloper · Feb 10

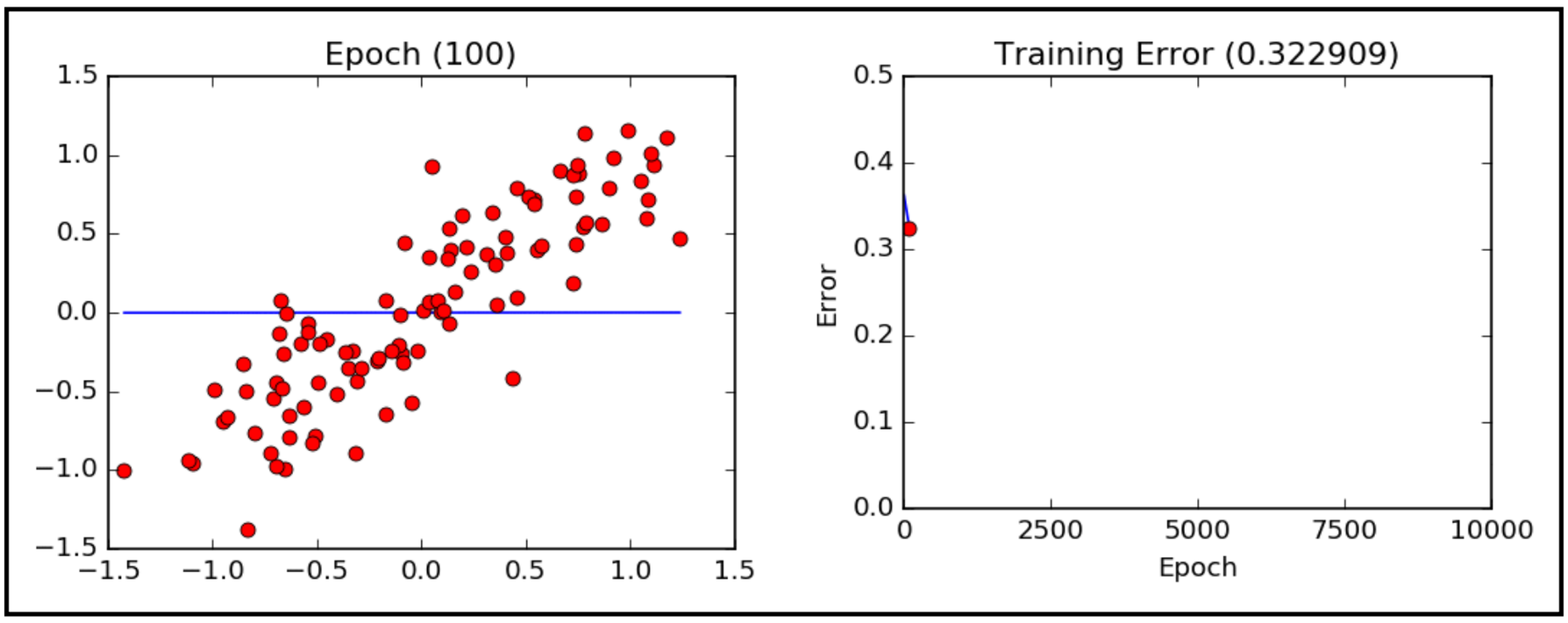
You say: "We added AI to our product"
I hear: "We added a bunch more IF statements to our codebase"

46 3.2K 5.3K

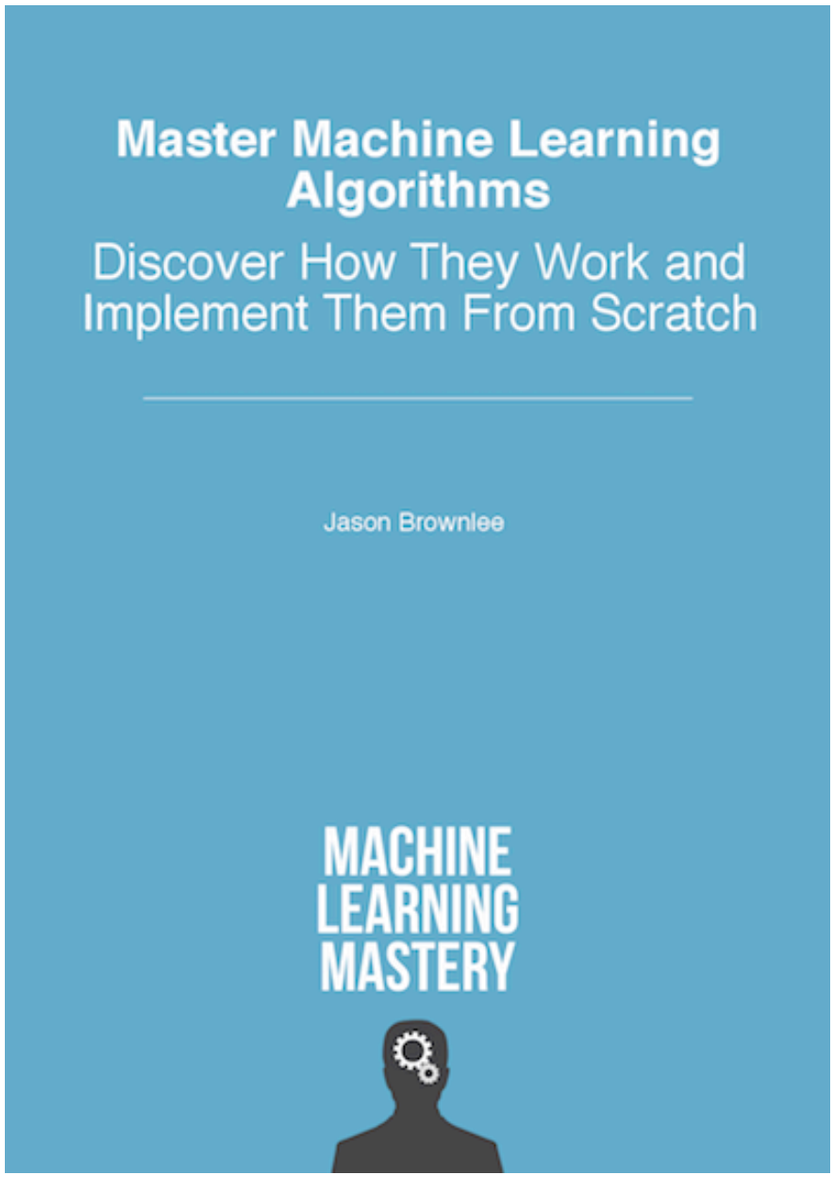
A.I. Taking Over The World

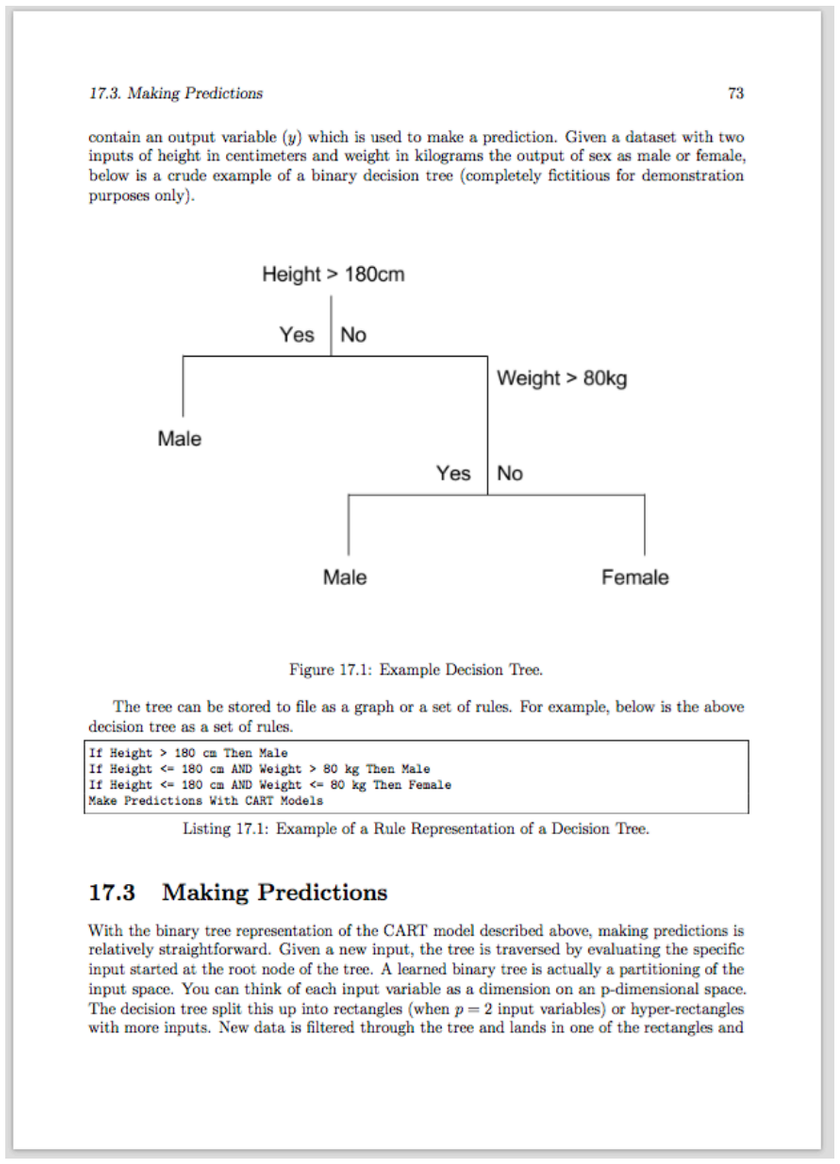
My CPU is a neural-net processor; a learning computer

My learning rate is 0.001



Get Started With Machine Learning





17.3. Making Predictions 73

contain an output variable (y) which is used to make a prediction. Given a dataset with two inputs of height in centimeters and weight in kilograms the output of sex as male or female, below is a crude example of a binary decision tree (completely fictitious for demonstration purposes only).

Figure 17.1: Example Decision Tree.

The tree can be stored to file as a graph or a set of rules. For example, below is the above decision tree as a set of rules.

```
If Height > 180 cm Then Male
If Height <= 180 cm AND Weight > 80 kg Then Male
If Height <= 180 cm AND Weight <= 80 kg Then Female
Make Predictions With CART Models
```

Listing 17.1: Example of a Rule Representation of a Decision Tree.

17.3 Making Predictions

With the binary tree representation of the CART model described above, making predictions is relatively straightforward. Given a new input, the tree is traversed by evaluating the specific input started at the root node of the tree. A learned binary tree is actually a partitioning of the input space. You can think of each input variable as a dimension on an p-dimensional space. The decision tree split this up into rectangles (when $p = 2$ input variables) or hyper-rectangles with more inputs. New data is filtered through the tree and lands in one of the rectangles and

Get the book

Live coding time!

OneR - Simple Classification Algorithm Tutorial:

github.com/AaronStearns/OneR-tutorial

ML-Powered Bowling Ball Calculator App:

github.com/AaronStearns/Bowling-Ball

Resources

LEARN R

Interactive Online Course:
https://www.udemy.com/data-science-and-machine-learning-bootcamp-with-r/

Kaggle Kernels:
https://www.kaggle.com/philippsp/exploratory-analysis-instacart

Free e-books:
R For Data Science
https://r4ds.had.co.nz/
Advanced R
http://adv-r.had.co.nz/

LEARN PYTHON

Interactive Online Course:
https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp/

Kaggle Kernels:
https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-instacart

LEARN SQL / NoSQL

Learn SQL
https://zh.sqlzoo.net/

Learn MongoDB
https://university.mongodb.com/courses/M001/about

Resources (cont'd)

Probability and statistics:

Eugene O'Loughlin "How to… Statistics by hand" YouTube video lecture series:
https://www.youtube.com/watch?v=JM94I5q84hc&list=PLfGMkZaH76AmBzace1FjpQveu80dmC-80

Introduction to Probability (free):
https://www.edx.org/course/introduction-probability-science-mitx-6-041x-2

Practical Statistics for Data Scientists
https://www.amazon.com/Practical-Statistics-Data-Scientists-Essential/dp/1491952962

Machine Learning:

Understand and implement ML algorithms from scratch, step-by-step, in a spreadsheet:
https://machinelearningmastery.com/master-machine-learning-algorithms/

An Introduction to Statistical Learning (free e-book and online course):
https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about

Machine learning with R:
https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r

Data analysis, visualization, and machine learning with R:
https://machinelearningmastery.com/machine-learning-with-r/

Resources (cont'd)

Data visualization:

Data Visualization: A Practical Introduction (Using R and ggplot2)
https://www.amazon.com/Data-Visualization-Introduction-Kieran-Healy/dp/0691181624

The Visual Display of Quantitative Information
https://www.amazon.com/Visual-Display-Quantitative-Information/dp/1930824130

Plotly for R
https://plot.ly/r/

R Graph Gallery
https://www.r-graph-gallery.com/

Python Visualizations Gallery
https://matplotlib.org/gallery.html

NLP / Text Analytics:

Text Mining With R (free)
https://www.tidytextmining.com/

SentDex - Python NLP tutorial series
https://www.youtube.com/watch?v=FLZvOKSCkxY

GitHub Link to download this presentation:

github.com/AaronStearns/gettingStartedWithDataScience

Aaron Stearns

stearnsconcepts@gmail.com

linkedin.com/in/aaronstearns