

# Data Science Bootcamp

Joseph Kambourakis - Lead Technical Instructor - Databricks



## Ground Rules

- Interrupt me
- These are all my opinions and not the opinions of General Assembly or Databricks.

---

# Introductions



# Joseph Kambourakis

Lead Technical Instructor - Databricks  
Instructor / SME - General Assembly

[josephk@databricks.com](mailto:josephk@databricks.com)



WPI

Main Entrance

Worcester Polytechnic Institute





**BENTLEY**  
UNIVERSITY

# Taught Around the World

Biz+ Front Row Center  
SUNDAY 特集

震度3

棚倉町  
浅川町

玉川村  
古殿町

平田村  
小野町

データサイエンティスト  
不足で養成の動き

Denied 3/3

Approved 3/3

問題はどちらのデータが大事かだ

# Simplify Big Data and AI with Databricks

Increases Data Science  
Productivity by 5x

Eliminates Disparate Tools  
with Optimized Spark

Removes Devops &  
Infrastructure Complexity



## Unified Analytics Platform

### DATABRICKS COLLABORATIVE NOTEBOOKS

Explore Data → Train Models → Serve Models



Open Extensible API's



+ a b l e a u



### DATABRICKS RUNTIME

Reliability



Performance

Accelerates & Simplifies  
Data Prep for Analytics

### DATABRICKS SERVERLESS



Azure

Databricks Enterprise Security







You

Name

Fun Fact

Area of Study

How familiar are you with Data Science?

How familiar are you with Python?

What do you hope to get out of today?



## Agenda

- ▶ What is Data Science?
  - Skillset, Problems, and Opportunities
  - Case Studies
- ▶ Data Science Workflow
- ▶ Tools of The Trade
  - Languages and Technologies
  - Python Demo
- ▶ Exploratory Data Analysis
  - Data Visualization Techniques and Demo
- ▶ Experimental Design
- ▶ Statistical Modeling
- ▶ Closing Discussion



Sign in

[www.ga.co/signin](http://www.ga.co/signin)

---

# What is Data Science?





Skills

# What skills should a data scientist have?



## Data Science Is A Misnomer

- ▶ Scientific process involves formulating a hypothesis, gathering direct observations, and confirming/falsifying the hypothesis
- ▶ Data Science is a set of tools and techniques used to produce inference
- ▶ The marriage between statistics and software engineering

## What Is A Data Scientist?



**Josh Wills**

@josh\_wills



Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.



Reply



Retweet



Favorite

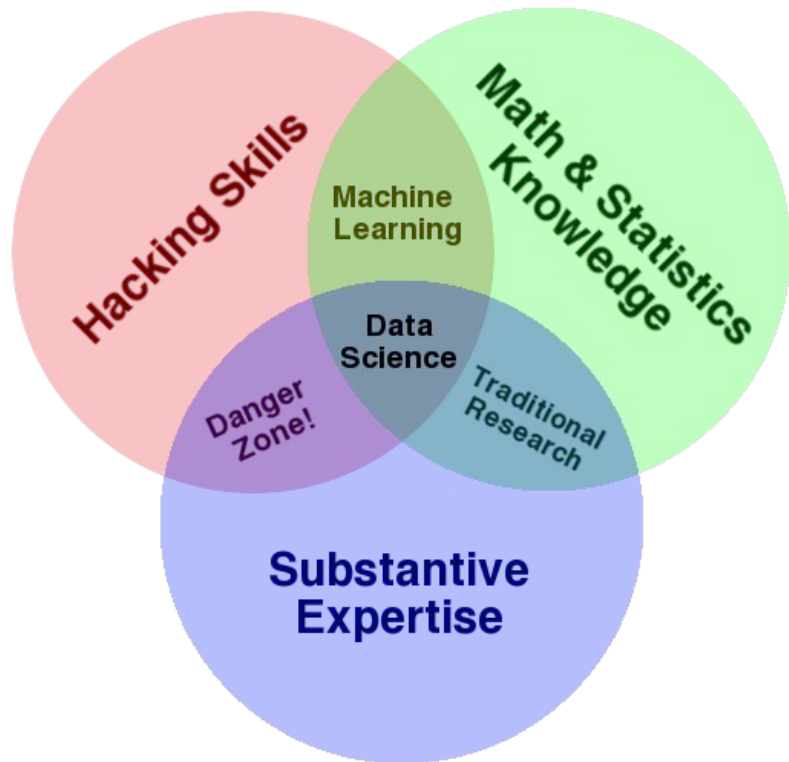


More

9:55 AM - 3 May 12



Hacking Skills + Math/Stats + Expertise







**Vicki Boykis**

@vboykis

Following



What's the difference between a data scientist and a senior data scientist? The first one tells you they're doing deep learning and they believe it. The second one tells you they're doing deep learning and you believe it.

7:08 AM - 20 Apr 2018



My definition!

Someone who uses open source tools across a multi-node environment to join disparate datasets in order to build a statistical model

## The Analyst

- Trains Models
- Answers “Why?”
- Understands Technical Stack
- Cleans Data
- Holds Domain Expertise

## The Engineer

- Builds Products
- Answers “How”?
- Understands Statistical Analysis
- Cleans Data
- Holds Domain Expertise



## Who Can Be a Data Scientist?

- ▶ PhD, or experience?
- ▶ Tech skills required
- ▶ Domain knowledge preferred
- ▶ Google put it best:
  - Get real-world experience.
  - Spend time coding.
  - Be passionate.
  - Note that you have multiple options.





## A Data Scientist's Job?

- ▶ Manage Business Expectations
- ▶ Develop appropriate skills and personality traits
- ▶ Internalize statistical intuition and data literacy
- ▶ Create teams with complementary skills and experience
- ▶ Encourage peer to peer support, learning and development
- ▶ Match personal growth to the business needs



## A Data Scientist's Job in Tech?

- ▶ Product Analytics
- ▶ Data Engineering
- ▶ Experimentation
- ▶ Predictive Modeling

---

# Case Studies



## “Just Google It”

- ▶ Google processes over 20 petabytes of data every day
  - That's 400 Million 4-drawer filing cabinets full of text
- ▶ PageRank algorithm determines relevance to query and quality of content
- ▶ Works by creating a graph of all pages and hyperlinks on each page
  - A page is better if it is popular
  - Personalization allows for optimal recommendations
- ▶ Data must be served FAST

## “Target Knew I Was Pregnant”

- ▶ DS team analysed buying patterns of women on baby registries
- ▶ Trends emerged:
  - Higher volume of lotion purchased near their 2nd trimester
  - Switch to scent-free products, cotton balls, wash clothes near due date
  - Colored items reveal gender (blue for boy, pink for girl)
- ▶ Marketing team used this data to target coupons

██████  
“Netflix Just Totally Gets Me”

- ▶ Public contest held to find best possible recommendations algorithm
- ▶ Customers are segmented and clustered
- ▶ Features:
  - When user watches and for how long
  - Where the user is watching
  - What device they are watching on



Almost anything you can think of!

- ▶ Everything creates data
- ▶ Endless applications in every industry!

---

# Workflow





GA Data Science Workflow

# How would you design a workflow?



## GA Data Science Workflow

A generalized approach to Data Science projects:

1. Identify Problem
2. Acquire Data
3. Parse Data
4. Mine Data
5. Refine Data
6. Model Data
7. Present Results
8. Implement

---



# Tools



## Data Science Tools

- ▶ Analysis
  - Coding Languages - R, Python, Scala, Java
- ▶ Storage
  - SQL Databases - PostgreSQL, MySQL, MsSQL, Oracle
  - Cloud AWS/Azure
- ▶ Visualization
  - Python and R have built in graphing support

## Data Science Languages

	PRO	CON
	<ul style="list-style-type: none"><li>▶ Up and running quickly with RStudio</li><li>▶ Great for exploratory data analysis, visualization</li><li>▶ Built by statisticians, for statisticians</li></ul>	<ul style="list-style-type: none"><li>▶ Very slow, poor memory optimization</li><li>▶ Difficult to run in production</li></ul>
	<ul style="list-style-type: none"><li>▶ Integrates with web applications</li><li>▶ Hundreds of freely available packages for data analysis</li><li>▶ Capable of handling large amounts of data in memory</li></ul>	<ul style="list-style-type: none"><li>▶ Missing some of the capabilities of R</li><li>▶ No environment</li></ul>



## Tools Demo

1. Create a Community Edition Account
2. Load in dbc file
  - a. <https://s3.amazonaws.com/hbsdatascience/hbs.dbc>
3. Open 01 - DataFrames
4. Feel free to follow along in code, or just watch me

## Data Storage

### ► How do we store

- On hard disks.
- On the cloud!

### ► A **database** is an o Management System

- **SQL** (Structured Q  
database



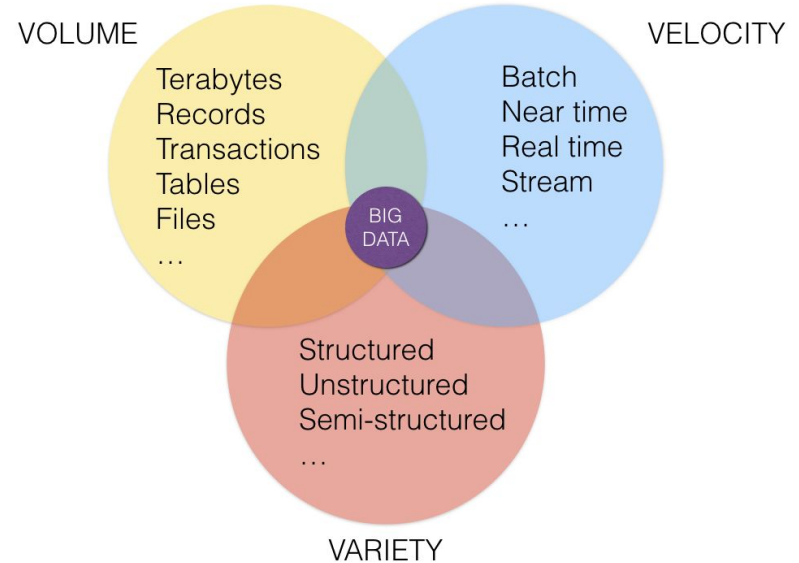
in text files, it would  
tion or update.

on a Database  
the data.

we can interact with a

## What is Big Data?

- ▶ Any amount of data that requires multiple computers to hold and process
- ▶ Big Data is everywhere, can you think of examples?
- ▶ What are the challenges in working with Big Data?
- ▶ Defined by the “3 V’s”
- ▶ Made possible by Cloud Computing





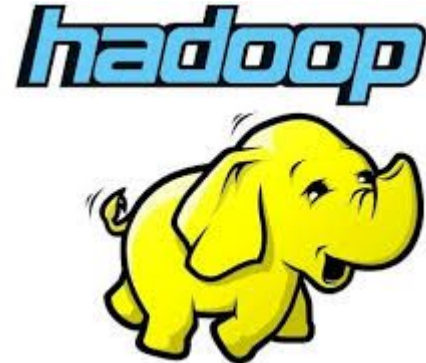


## Approaches to Big Data Computing

- ▶ Distributed processing
  - Split tasks into subtasks, compute, then recombine results
  - Divide and Conquer
  - Embarrassingly Parallel

## Big Data Computing Technologies

- ▶ Apache Spark
  - Unified analytics engine for large scale processing
- ▶ Hadoop
  - HDFS
    - Massive data storage system
  - Hive
    - Data summarization, query and analysis tool
    - Translates SQL queries into MapReduce jobs
    - Integrates with Hadoop
  - Other tools such as Pig, Zookeeper, Impala



---

# Exploratory Data Analysis



What is EDA?

- ▶ Getting To Know Your Data
- ▶ Requires research questions and well defined variables
- ▶ Goal is to explore data until a story emerges
- ▶ Tools/Techniques:
  - Data Viz
  - Residual Analysis
  - Data Transformation
  - Statistical Analysis



## Data Viz

A picture is worth a thousand words

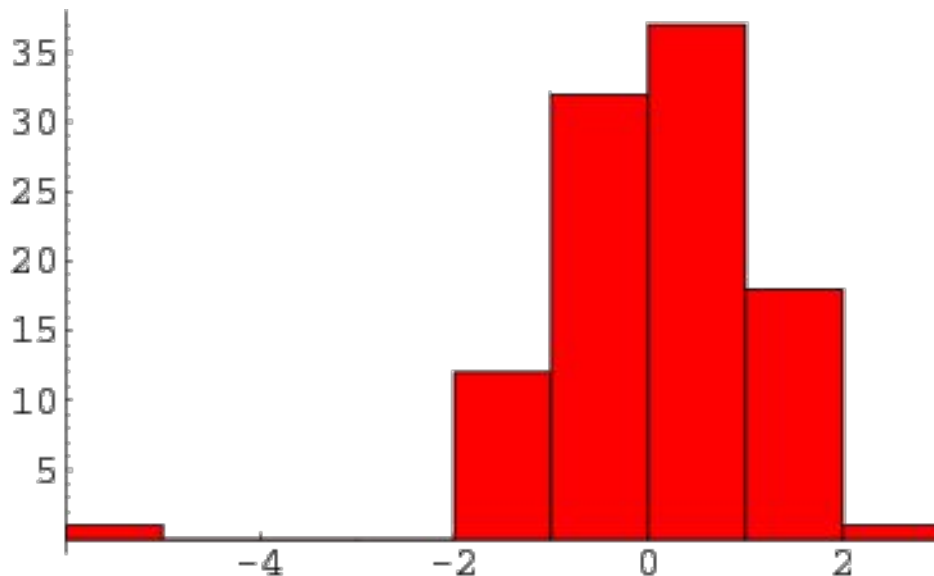
Goals:

- ▶ Spotting outliers
- ▶ Discriminating clusters
- ▶ Checking distributional and other assumptions
- ▶ Examining relationships
- ▶ Observing a time-based process



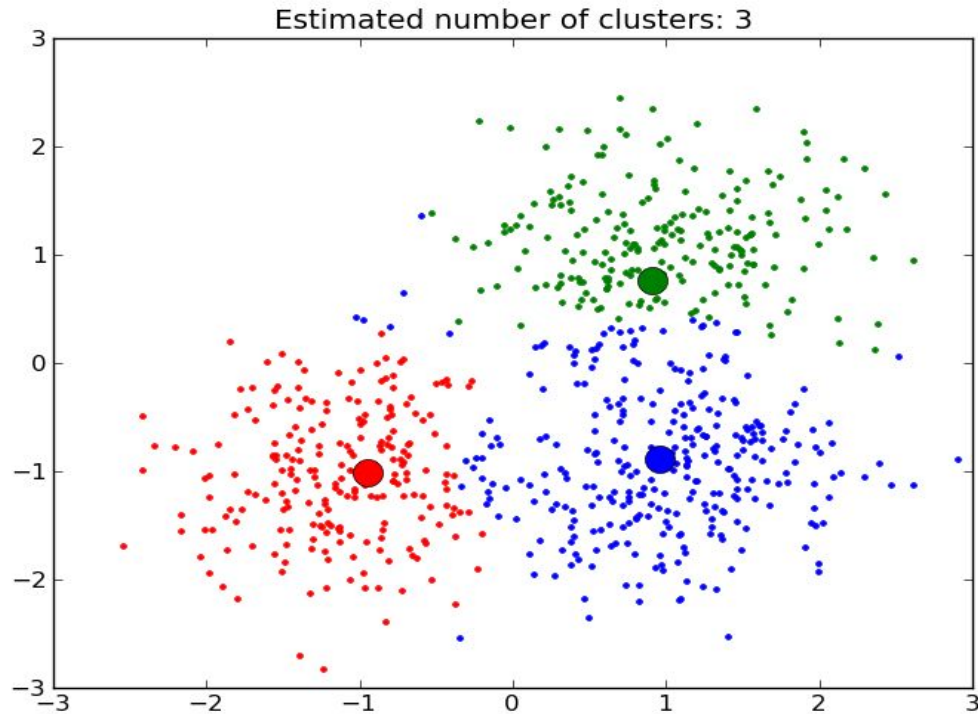
## Spotting Outliers with Histograms

- ▶ Understand the distribution of your data points
- ▶ Every outlier has a story, is it meaningful?
- ▶ Normal vs. Skew vs. Non-Normal



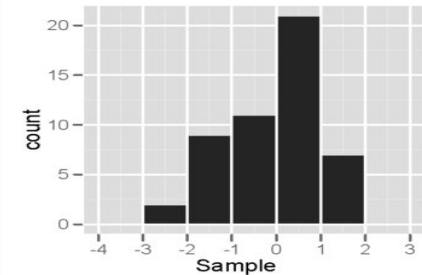
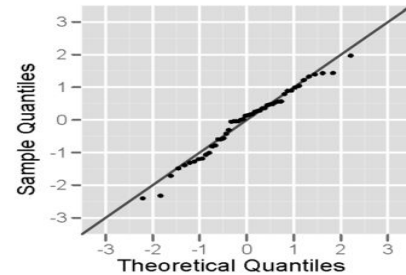
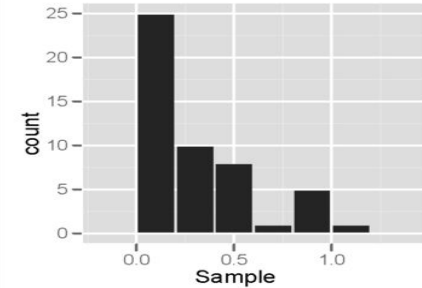
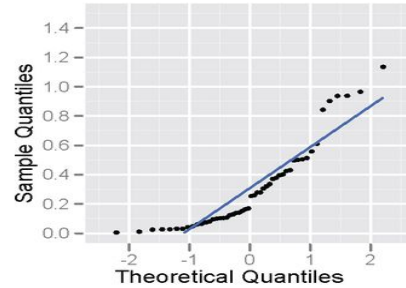
## Discriminating Clusters

- ▶ Spot latent classification
- ▶ Simple means to examine interaction effects
- ▶ Effective at describing transformations



## Evaluating Normality Assumptions

- ▶ Normal Quantile Plots
- ▶ Plots Expected Normal Score (Z-Score)
- ▶ Normally distributed data results in straight diagonal line

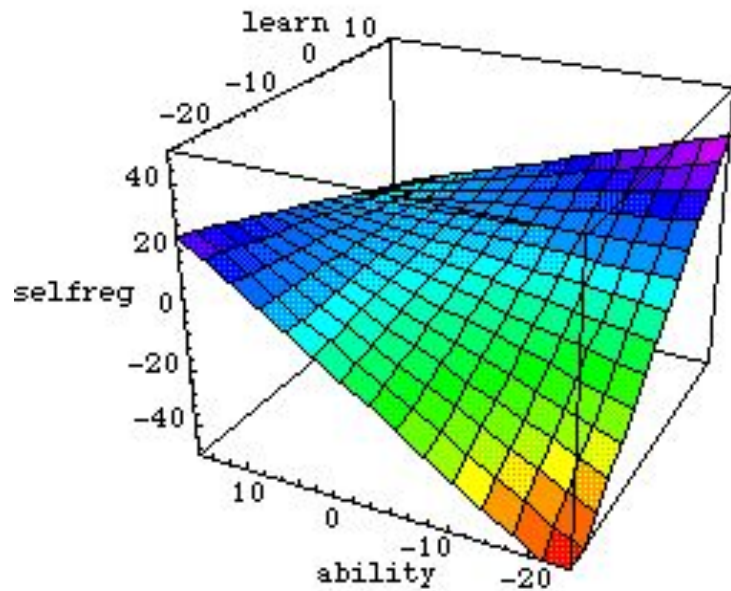






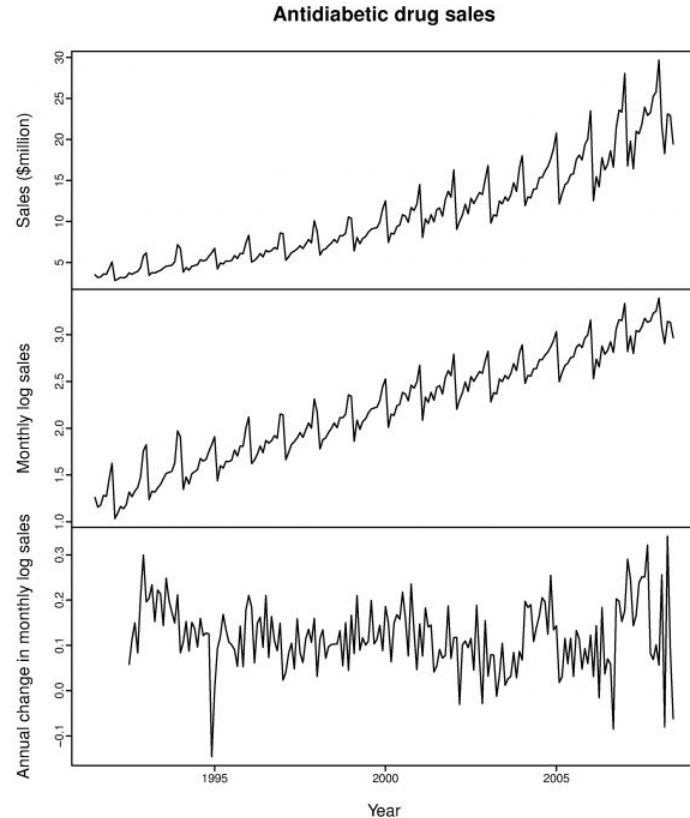
## Measuring Variable Relationships

{Family, -2.31}



## Observing Time Based Processes

- ▶ Time series analysis is a common business need
- ▶ Displays trend and seasonality
- ▶ Differencing is a common tactic to produce a stationary time series






## Visualisation Demo

1. Open 02 - Anscombe's Quartet
2. Feel free to follow along in code, or just watch me

---

# Experimental Design

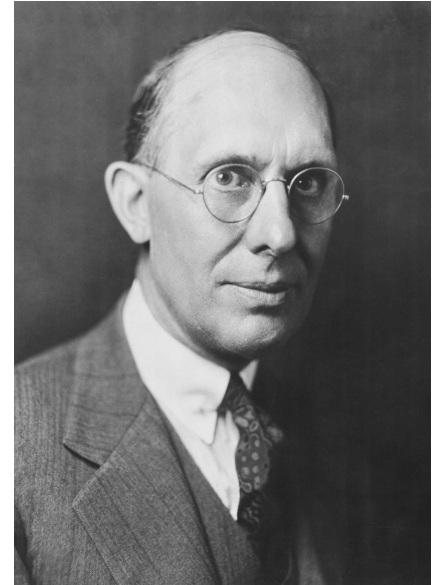


"Framing the  
question well is...  
the hardest part  
about being a  
good data  
scientist."

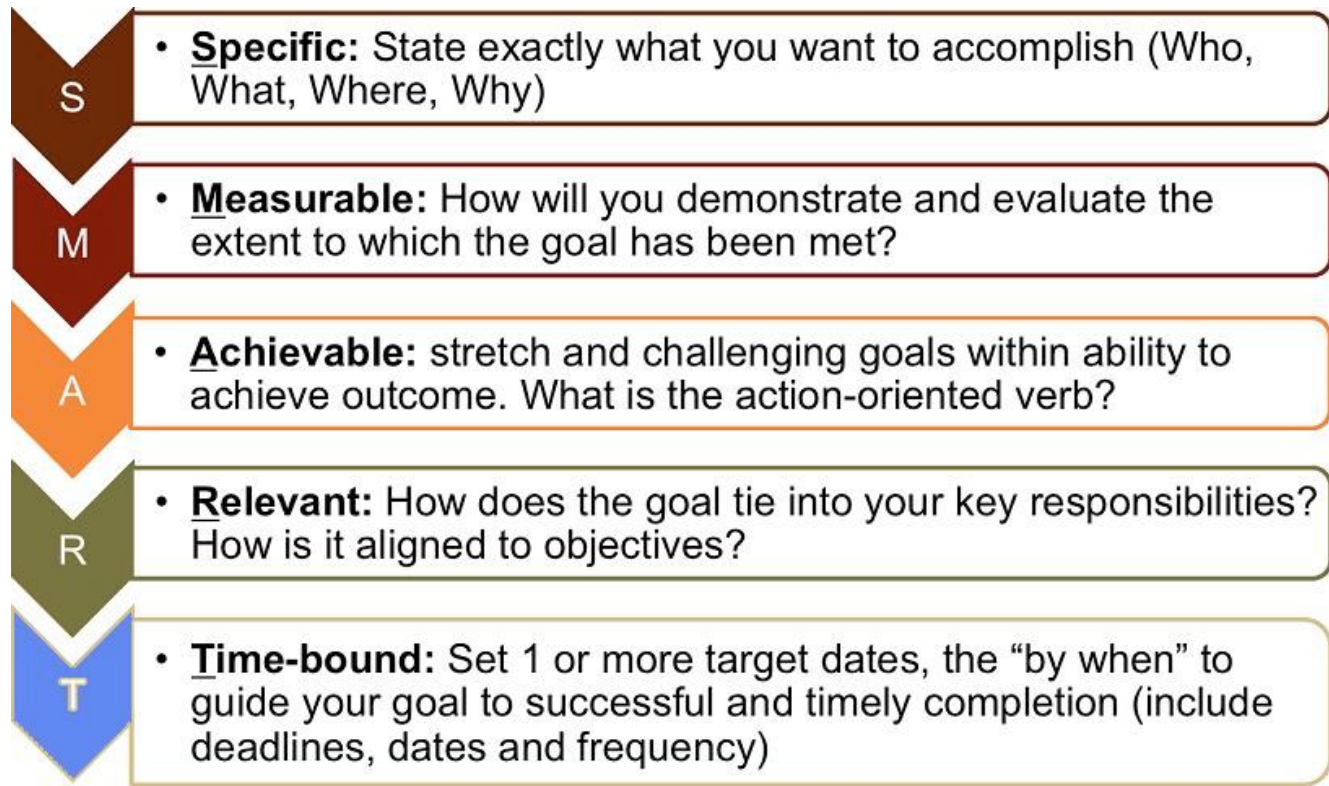
— Cathy O'Neil  
*On Being a Data Skeptic*

## Asking Good Questions

- “A problem well stated is half solved.” -Charles Kettering
- Sets yourself up for success as you begin analysis
- Establishes the basis for reproducibility
- Enables collaboration through clear goals



## What Is A Good Question?



---

# Case Study





## Asking Good Questions

**Goal:** Determine the association of foods in the home with child dietary intake.

**Data Available:** One 24-hour recall from the cross-sectional NHANES 2007-2010

**Experiment:** We will test if reported availability of certain foods available in the home increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for that food.



USDA

Children will be *more likely* to meet the USDA recommended intake level when food is always available in their home compared to *rarely or never*.



## Specific

- ▶ How data was collected:
  - 24-hour recall, self-reported
- ▶ What data was collected:
  - Fruits, dark green vegetables, low fat milk or sugar sweetened beverages, always vs. rarely available
- ▶ How data will be analyzed:
  - Using USDA recommendations as a gold-standard to measure the association
- ▶ The specific hypothesis & direction of the expected associations:
  - Children will be more likely to meet their recommended intake level



Measurable

- Determine the association of foods in the home with child dietary intake.
- We will test if the reported availability of certain foods increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for food.



Achievable

- Cross-sectional data has inherent limitations; one of the most common is that causal inference is typically not possible.
- Note that we are determining association, not causation.



Reproducible

- With all the specifics, it would be straightforward to pull the data from NHANES and reproduce the analysis.



## Time-Bound

- Using one 24-hour recall from NHANES 2007-2010, we will determine the factors associated with food available in the homes of American children and adolescents.



## Why Data Types Matter

- Different data types have different limitations and strengths.
- Certain types of analyses aren't possible with certain data types.





## Cross-Sectional Data

- All information is determined at the same time; all data comes from the same time period.
- Issues: There is no distinction between exposure and outcome



## Cross-Sectional Data

### ▸ Strengths

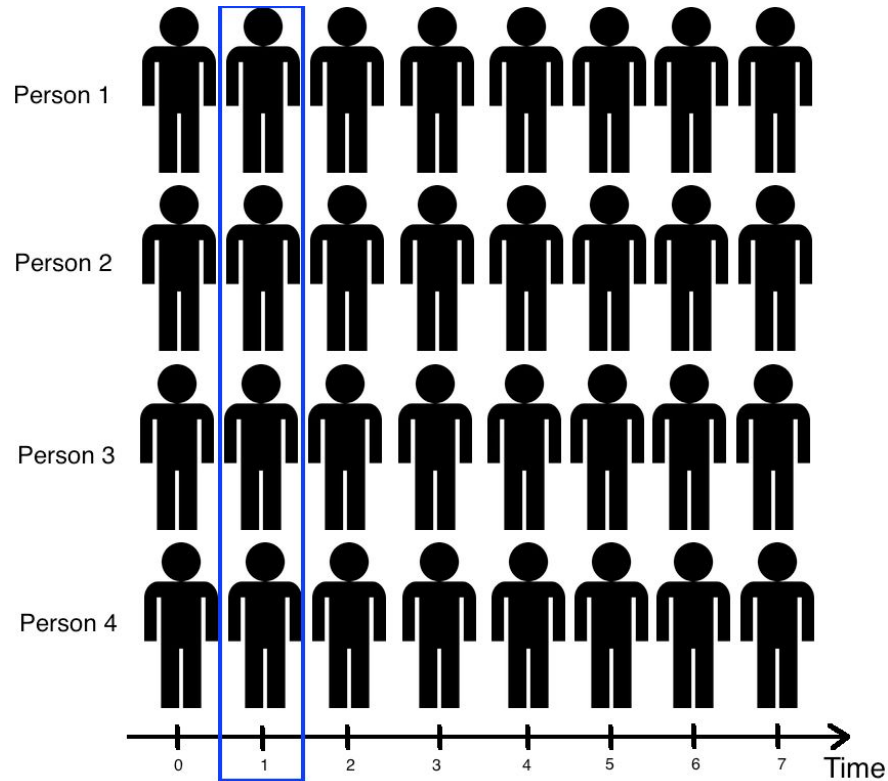
- Often population based
- Generalizability
- Reduce cost compared to other types of data collection methods

### ▸ Weaknesses

- Separation of cause and effect may be difficult (or impossible)
- Variables/cases with long duration are over-represented



## Cross-Sectional Data

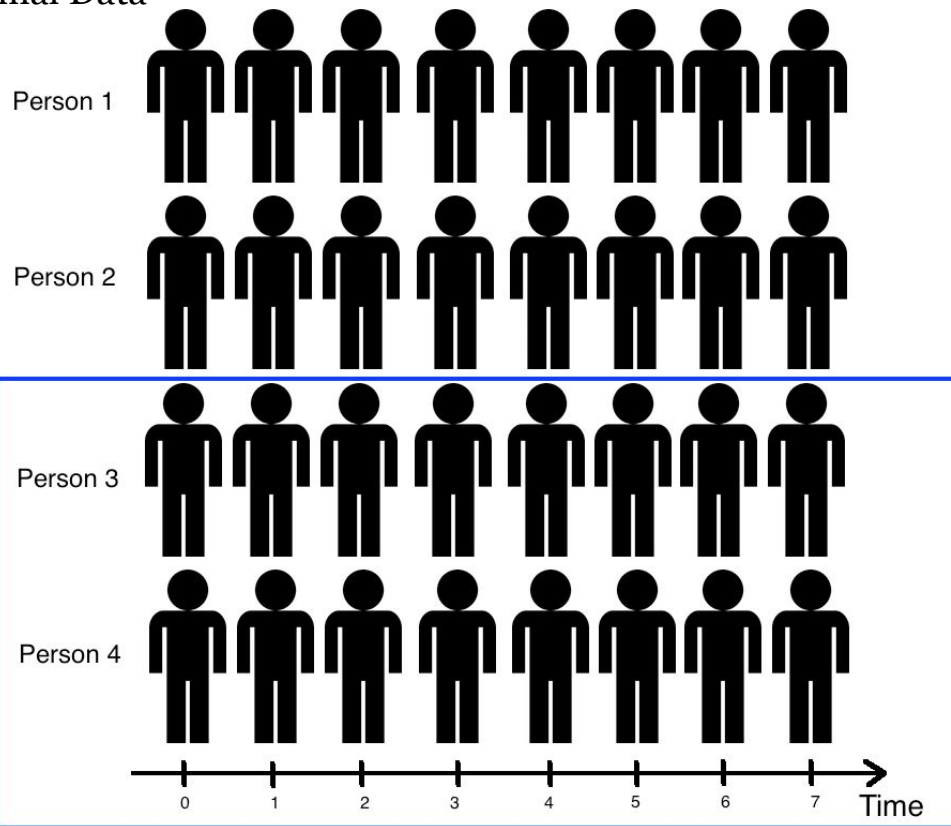




## Time-Series/Longitudinal Data

- The information is collected over a period of time
- Strengths
  - Unambiguous temporal sequence - exposure precedes outcome
  - Multiple outcomes can be measured
- Weaknesses
  - Expense
  - Takes a long time to collect data
  - Vulnerable to missing data

## Time Series/Longitudinal Data



---

# Machine Learning Models

**Not hotdog!**



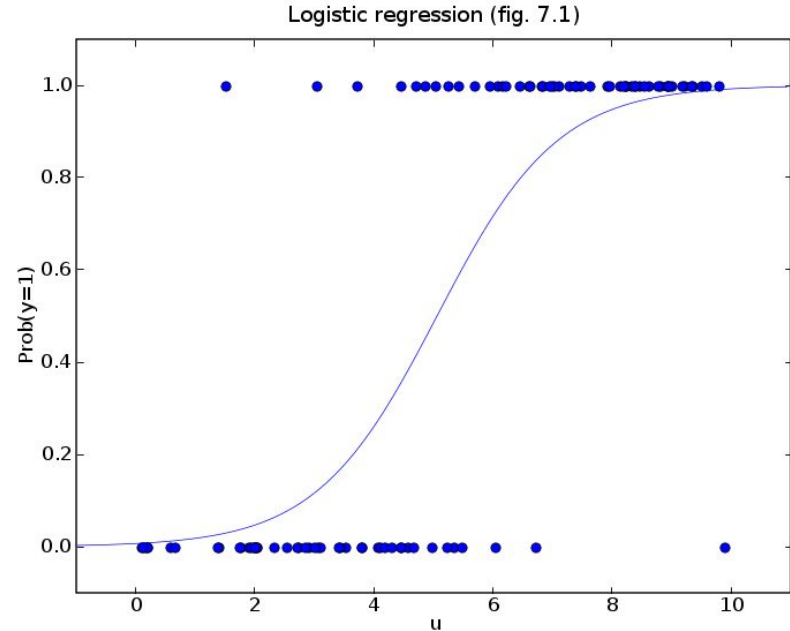
**Hotdog!**



## Build Some Models 2: Selection

### ► Binary Classification

- “Is this X an instance of Y, or Z?”
- Trained using Logistic Regression





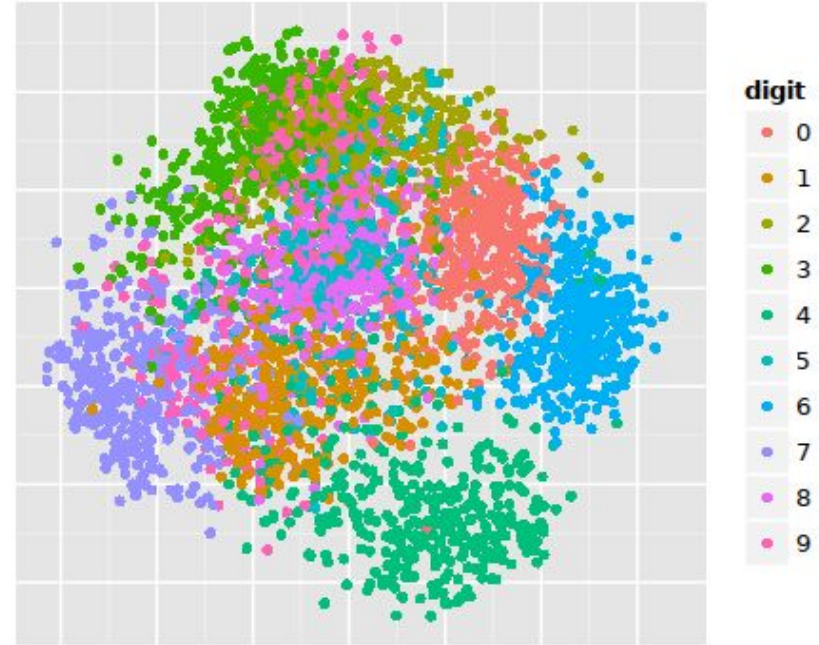


## Classification Demo

1. Open 03-Classification
2. Feel free to follow along in code, or just watch me

## Build Some Models 2: Selection

- ▶ Multiclass Classification
  - “Is this X an instance of A, B, C, D ...?”
  - Trained using Clustering Algorithms (i.e. K-Nearest Neighbors)



# Regression

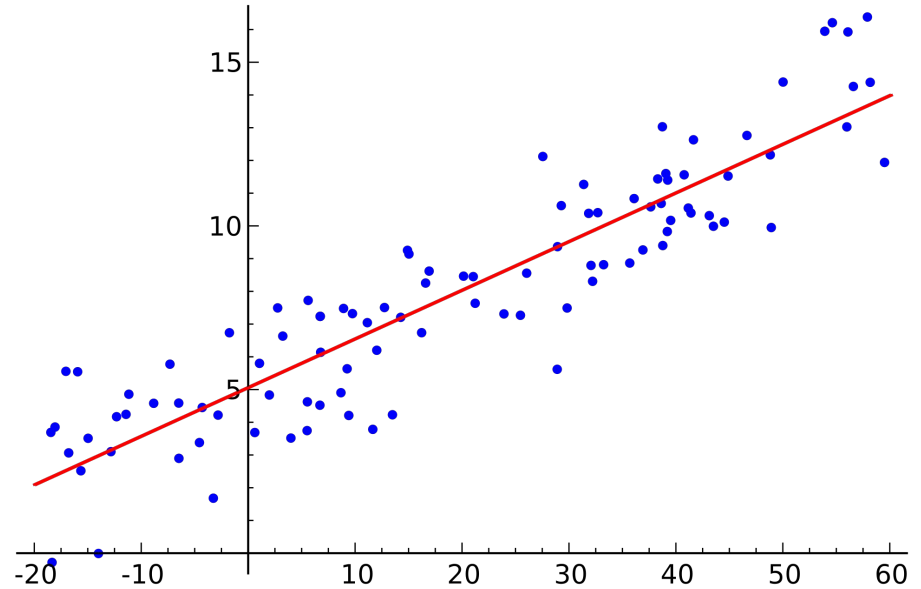
## FAMILY HEIGHTS. *from R.F.F.* (add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18.5	7.0	13.2 <small>5.5</small>	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 12.5 <small>2.0 3.0</small>	5.5, 5.5
3	15.0	about 4.0	11.0 <small>4.0</small>	8.0
4	15.0	4.0	10.5, 8.5 <small>4.5 6.5</small>	7.0, 4.5, 3.0
5	15.0	-1.5	12.0, 9.0, 8.0 <small>3.0 6.0 7.0</small>	6.5, 2.5, 2.5

## Build Some Models 2: Selection

### ► Regression Model

- “What will X be for a known value of Y?”
- Industry Standard is Linear Regression





## Visualisation Demo

1. Open 04 - Linear Regression
2. Feel free to follow along in code, or just watch me

---

# Review



What do you hope to get out of today?



Please Share Your Feedback

[www.ga.co/surveyhbs](http://www.ga.co/surveyhbs)



---

# Final Thoughts



## Job Types

### ▶ Data Scientist

- Type A: Analytics, Statistics, Strong Business Understanding
- Type B: Builder, Engineer, Highly Technical

### ▶ Business Intelligence Analyst

- Strong Business Understanding, SQL Skills, Visualization

### ▶ Data Analyst

- Project Dedicated, Gathering Data, Defining Specifications

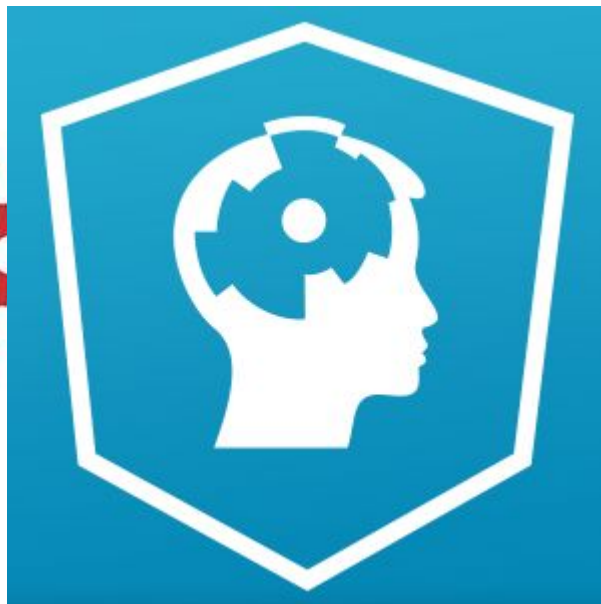
### ▶ Product Manager

- Guide The “Doers”, Evangelise Projects, Ensure Product Viability

## Further Learning



**GENERAL ASSEMBLY**



## Further Learning

### Doing Data Science

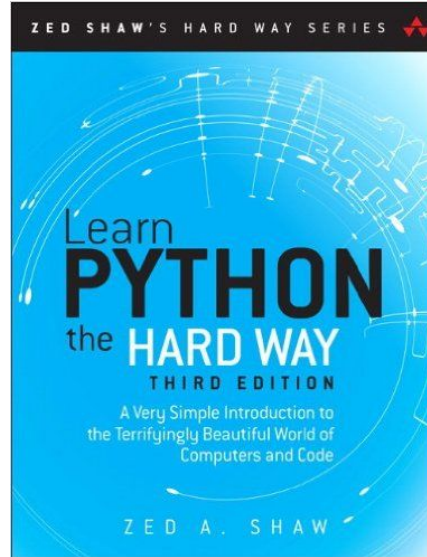
Straight Talk from the Frontline

By [Cathy O'Neil](#), [Rachel Schutt](#)



### Learn Python The Hard Way

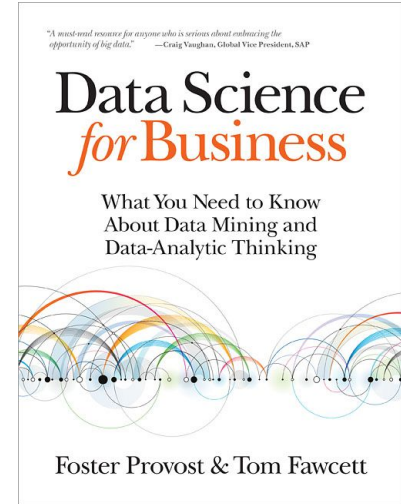
By [Zed A. Shaw](#)



### Data Science for Business

What You Need to Know about  
Data Mining and Data-Analytic  
Thinking

By [Foster Provost](#), [Tom Fawcett](#)





## Further Learning

- ▶ [reddit.com/r/datascience](https://reddit.com/r/datascience)
- ▶ [reddit.com/r/machinelearning](https://reddit.com/r/machinelearning)
- ▶ [Kdnuggets.com](https://Kdnuggets.com)
- ▶ [Kaggle.com](https://Kaggle.com)
- ▶ [datascience.stackexchange.com/](https://datascience.stackexchange.com/)
- ▶ [Quora.com](https://Quora.com)
- ▶ [Meetup.com](https://Meetup.com)