DATA SCIENCE IMMERSIVE

GA

WHAT IS DATA SCIENCE?

GA

BUT FIRST, A STORY…

But First, a Story…

But First, a Story...

But First, a Story…

But First, a Story…

But First, a Story…

Target using her purchase history alone had predicted
his daughter's pregnancy.

UH, OK, BUT WHY...?

But First, a Story…

Research shows that one of the very few times in their lives that
people change their purchase behaviors and are receptive
to new products is with the birth of a child.
Because of this, if you can predict when
they're likely to be expecting
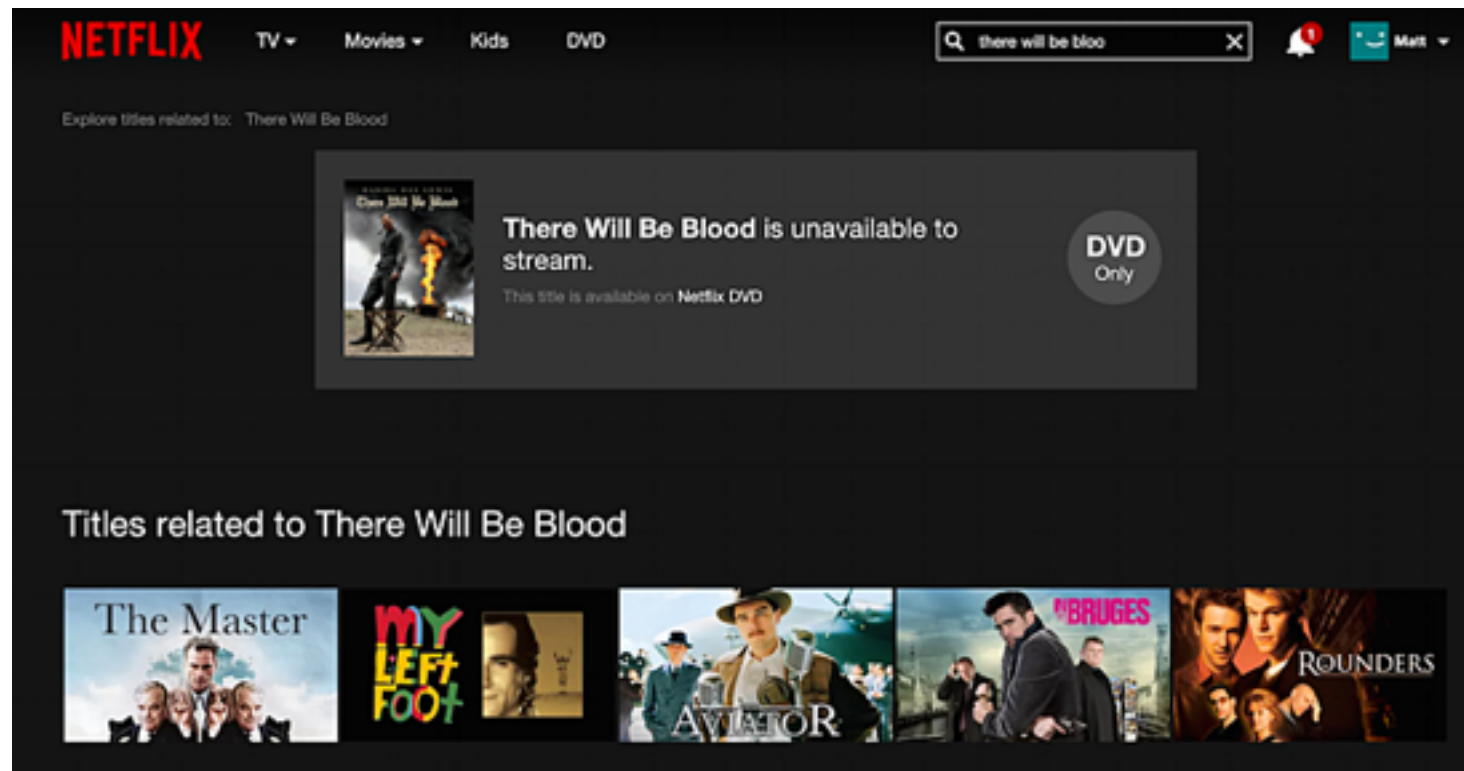you can influence
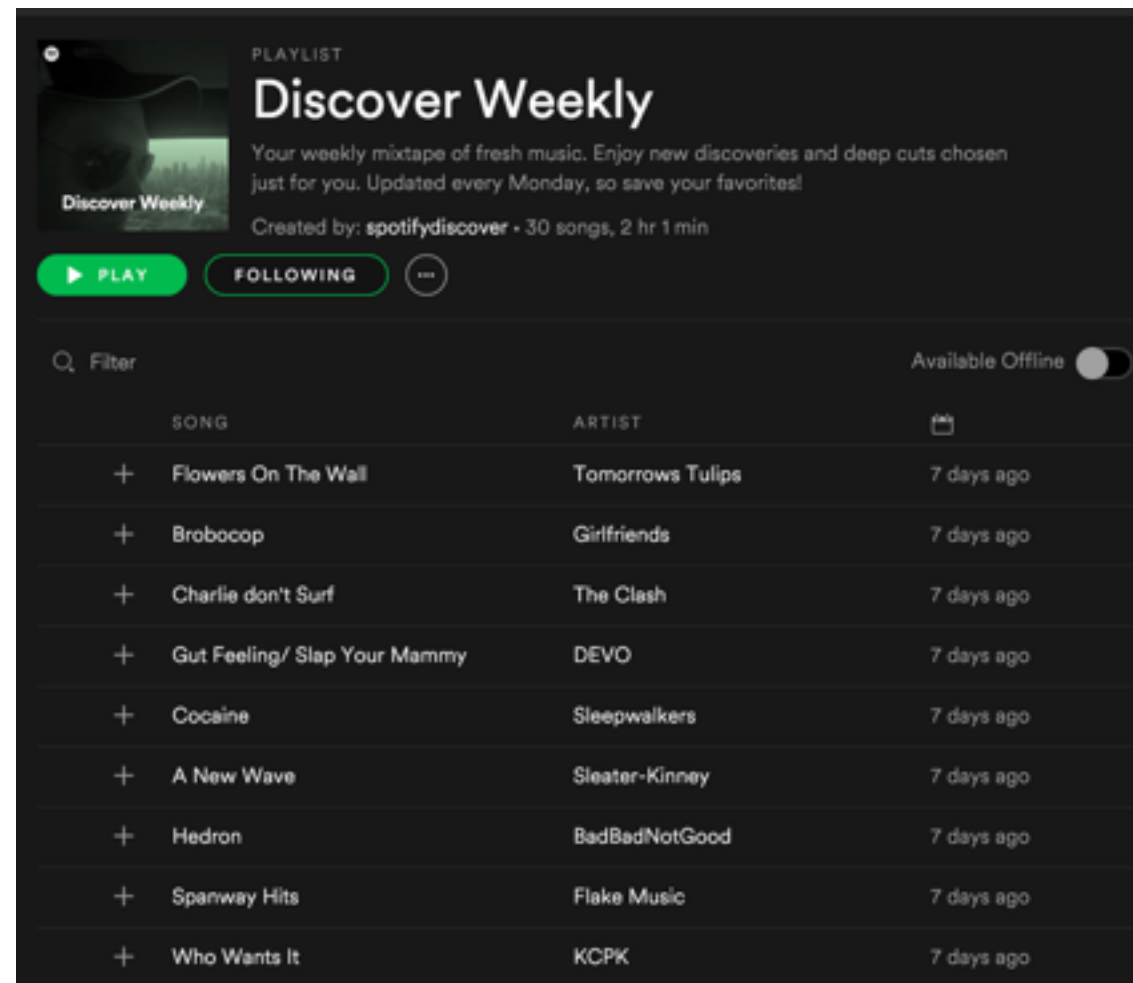their purchase
habits.

#DATASCIENCE
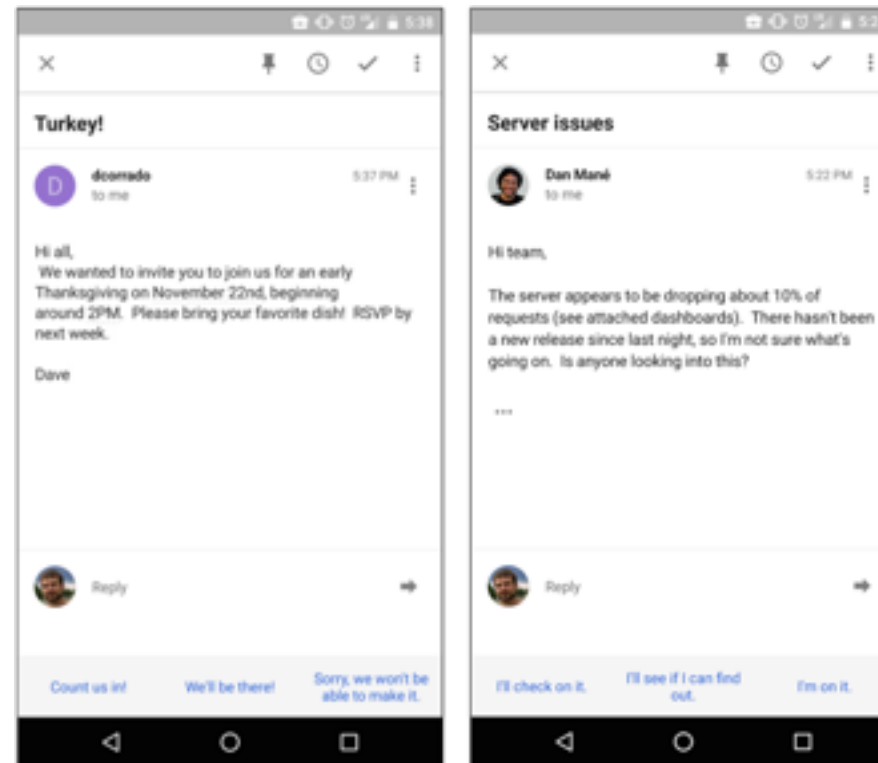
WHAT ELSE?

GA

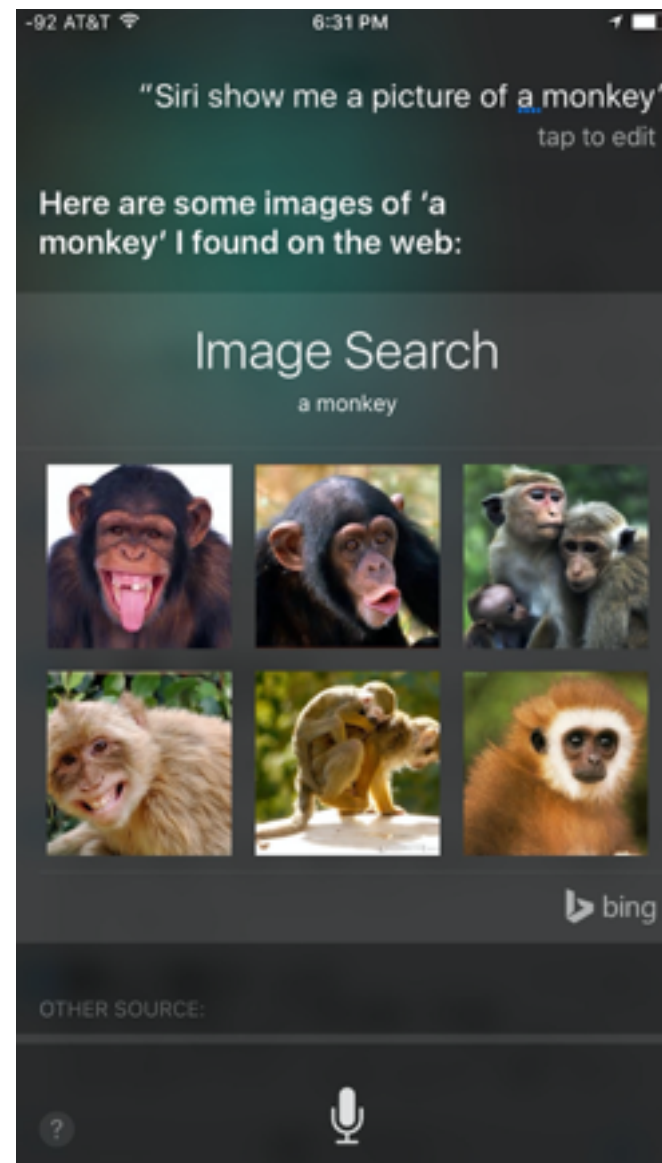## What else?

Netflix Recommendations

What else?

Spotify Discover Weekly

What else?

Google SmartReply

Siri: Speech Recognition

What else?

WHY BE A DATA SCIENTIST?

What Bloomberg Says…

Best Jobs For 2016
Overall ranking

Job	Ranking
Data Scientist	4.7
Tax Manager	4.7
Solutions Architect	4.6
Engagement Manager	4.6
Mobile Developer	4.6
HR Manager	4.6
Physician Assistant	4.6
Product Manager	4.5
Software Engineer	4.5
Audit Manager	4.5
Analytics Manager	4.5
Software Development Manager	4.4
Product Marketing Manager	4.4
Marketing Manager	4.4
QA Manager	4.4
Finance Manager	4.3
Business Development Manager	4.3
UX Designer	4.3
Strategy Manager	4.3
Technical Account Manager	4.2
Consultant	4.2
Construction Superintendent	4.2
Nurse Practitioner	4.2
Electrical Engineer	4.2
Software Architect	4.2

Source: Glassdoor

Bloomberg

What Bloomberg Says…

Best Jobs For 2016

■ Salaries
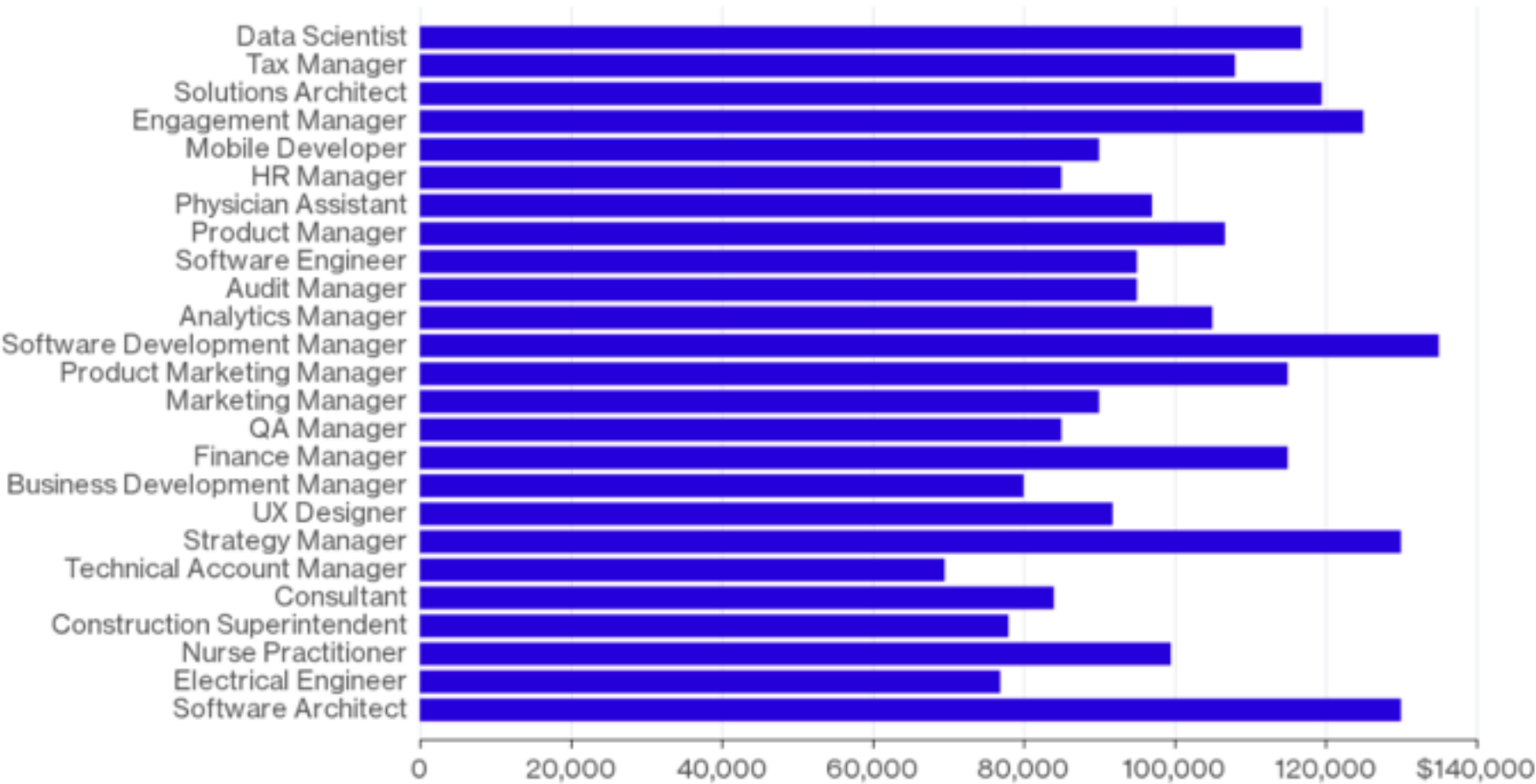
Source: Glassdoor

Bloomberg

What HBR Says…

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY   SAVE   SHARE   COMMENT   TEXT SIZE   PRINT | $8.95 BUY COPIES

WHAT SKILLS ARE REQUIRED?

GA

The Most Famous Tweet in Data Science History…
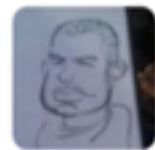
Josh Wills
@josh_wills

Following

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

The Most Famous Venn Diagram in Data Science...
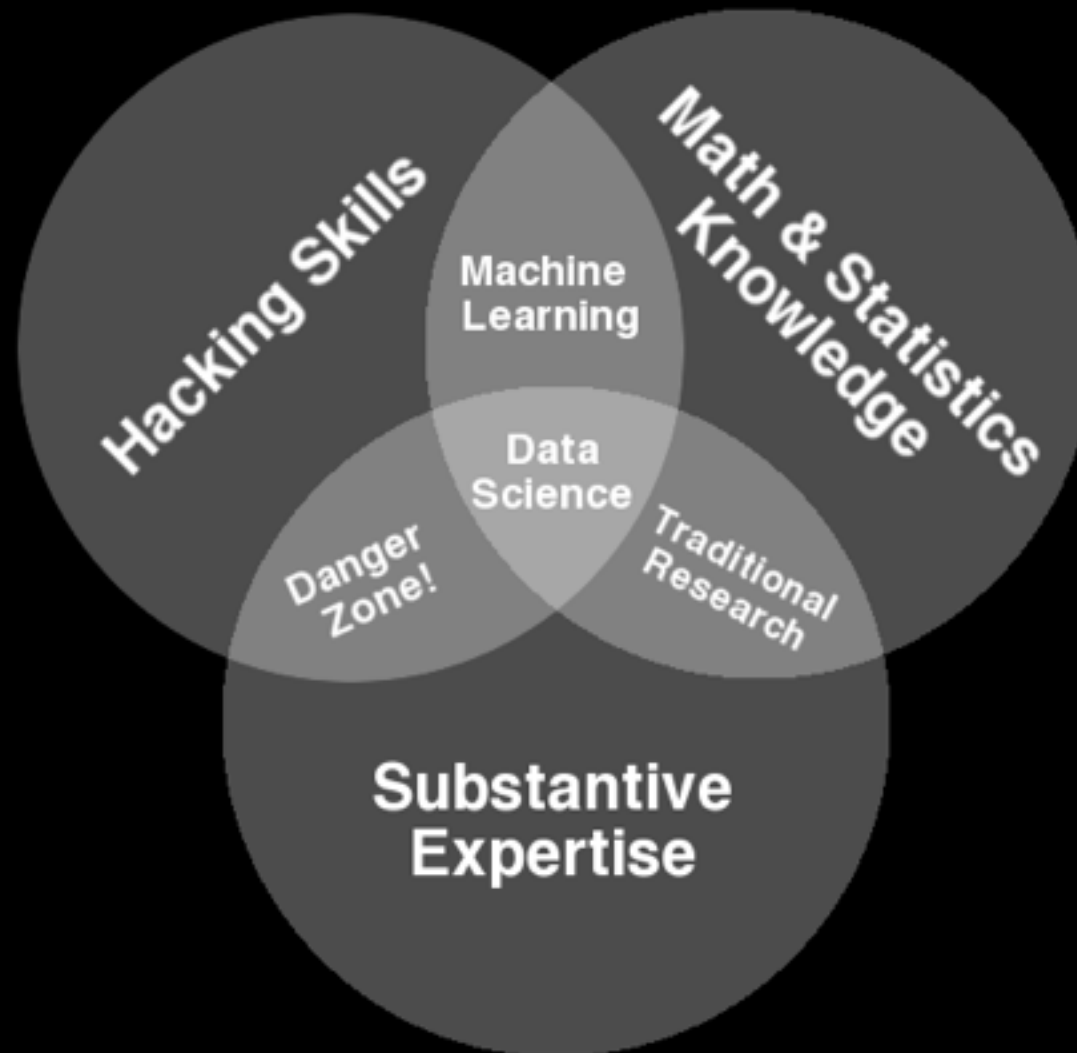
The Data Scientist's Skill Set

The Data Science Workflow

DATA SCIENCE WORKFLOW

IDENTIFY THE PROBLEM
- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA
- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA
- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA
- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA
- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL
- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

PRESENT RESULTS
- ☐ Summarize with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up issues for future analysis

DEPLOY AND VALIDATE
- ☐ Write unit tests and documentation
- ☐ Deploy stable production-ready code
- ☐ Retrain and validate models over time

LET'S TALK CURRICULUM & PROJECTS

Curriculum

Unit Breakdown

Unit	Title	Lessons	Topics
Unit 1	Data Science Foundations	Weeks 1-3	Programming Fundamentals, Pandas & EDA, Linear Regression
Unit 2	Supervised Learning Algorithms	Weeks 4-6	Logistic Regression, Classification, Databases, Ensemble Models
Unit 3	Advanced Modeling Techniques	Weeks 7-9	Unsupervised Learning, Bayes, Time Series
Unit 4	Data Science Careers	Weeks 10-12	Spark & Big Data, Portfolios, Presentations

Projects

Projects

- Project 1: SAT Scores + Summary Statistics
- Project 2: Billboard Hits + Data Munging
- Project 3: Liquor Sales + Linear Regression
- Project 4: Web Scraping + Logistic Regression
- Project 5: Disaster Reflief + Classification
- Project 6: IMDB API + Random Forests
- Project 7: Airport Delays + Cluster Analysis

- Part 1: Capstone Topic + Dataset Validation
- Part 2: Problem Statement + Exploratory Data Analysis
- Part 3: Progress Report + Preliminary Findings
- Part 4: Report Writeup + Technical Analysis
- Part 5: Presentation + Recommendations

Feedback - Rubric

Requirements	Incomplete (0)	Does Not Meet Expectations (1)	Meets Expectations (2)	Exceeds Expectations (3)
Identify: Discuss Data				
Acquire: Load Data with Python				n/a
Parse: Print Data				n/a
Mine: Sort Data				n/a
Refine: Analyze Data				n/a
Present: Visualize Data				Bonus: Use Tableau
Bonus! Present: Create blog post summary				

Q&A

THANK YOU.

‣ alex.combs@ga.co
‣ winston.featherly-bean@ga.co
‣ daniel.yawitz@ga.co