Wharton
UNIVERSITY of PENNSYLVANIA

Workshop 1: What is Data Science?

Wharton Analytics Fellows x WUDAC

Spring 2018

Roadmap

I. Sign-in & Icebreaker

II. Purpose of Data Science

III. Breaking Down “Big Data”

IV. Machine Learning

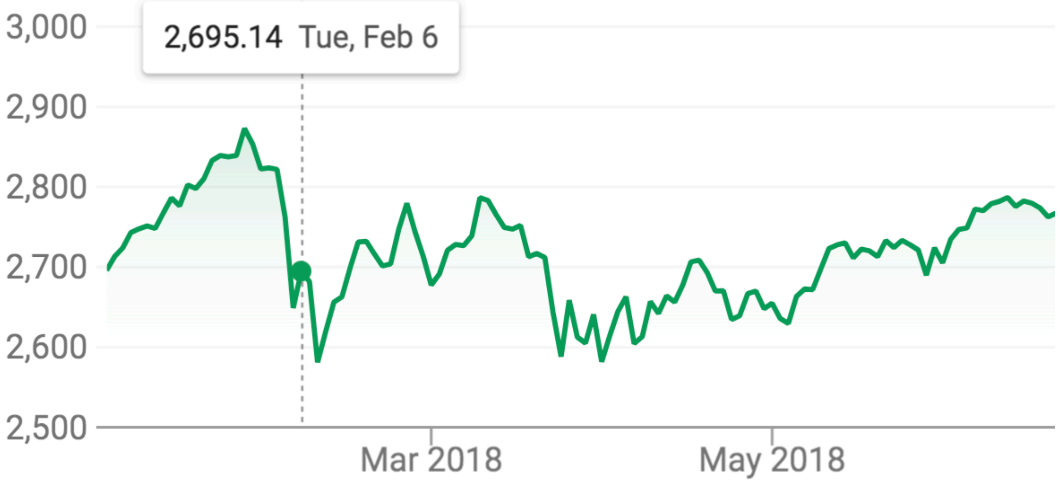
V. Bootcamp Outline

Sign-in here: goo.gl/umnhnG

Icebreaker

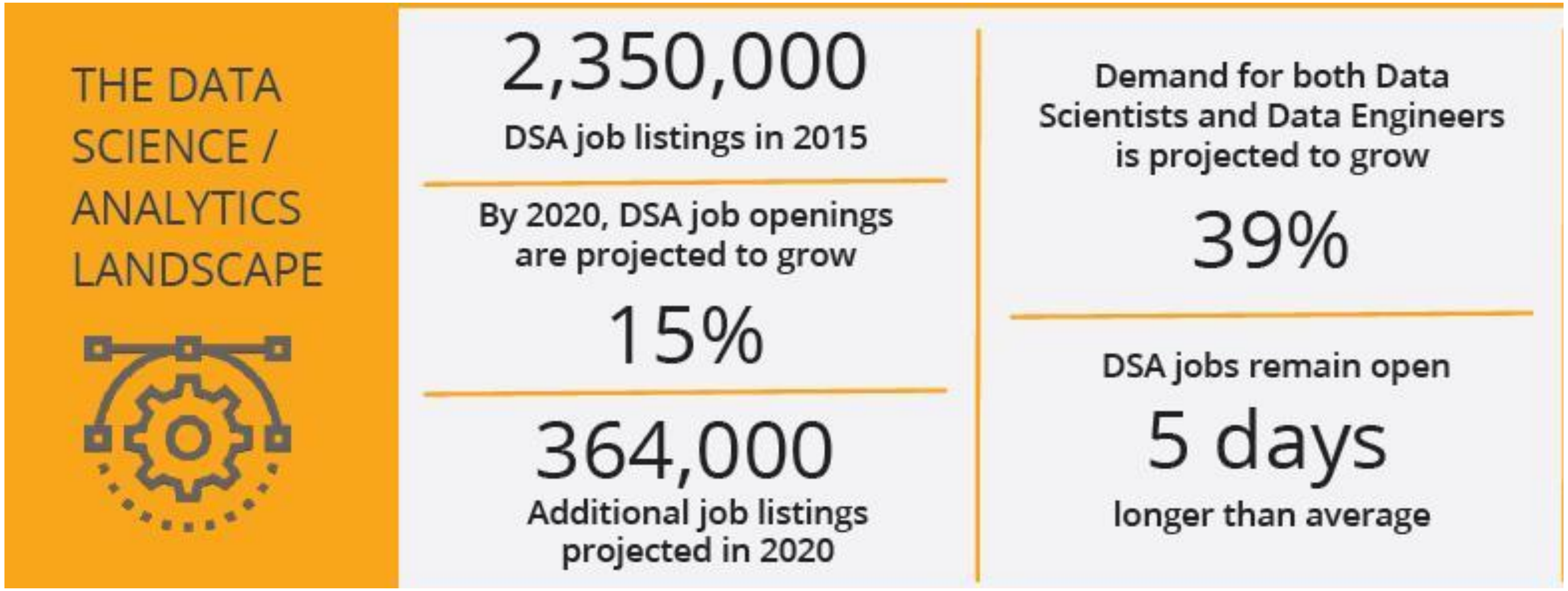
Purpose of Data Science

What will data science allow you to do?



- What do scurvy and predicting stock prices have in common?
- Answer questions, make decisions, understand hidden relationships

Nearly every field is pivoting to focus on data



THE DATA SCIENCE / ANALYTICS LANDSCAPE

2,350,000

DSA job listings in 2015

By 2020, DSA job openings are projected to grow

15%

364,000

Additional job listings projected in 2020

Demand for both Data Scientists and Data Engineers is projected to grow

39%

DSA jobs remain open

5 days

longer than average

- Data literacy is increasingly expected for not only analytics roles, but also management roles with a strategic focus
- People have always done this, but we increasingly have the tools to do this on a larger/more integrated scale

Source: https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#249f50087e3b

Careers in data science fall under a variety of titles

Data Analytics Consulting

- Shorter term projects with a variety of client companies
- Focus on applying existing algorithms to clients' data to answer targeted questions

BCG GAMMA

APT
A Mastercard Company

Data Scientist

- Similar to data analytics consulting, but within one company
- Apply existing algorithms to company data, sometimes develop and refine algorithms to company-specific needs
- Design own questions and "experiments"

facebook

Quora

UBER

Quantitative Researcher

- Often used as job title in financial field
- Working with alternative datasets to give investment recommendations
- Building systematic portfolios/trading strategies based on price data

DE Shaw & Co

Point72

Data Engineer

- Focus on data cleaning, manipulation, and storage
- Possibly more knowledge of database design and computer science

slack

Dropbox

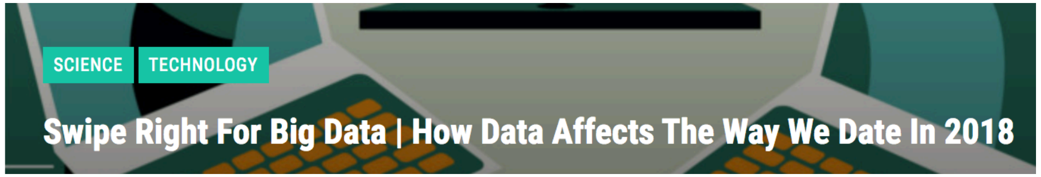
Breaking Down “Big Data”

Big Data is (unfortunately) still a huge buzzword, but…

Big data is watching you – and it wants your vote

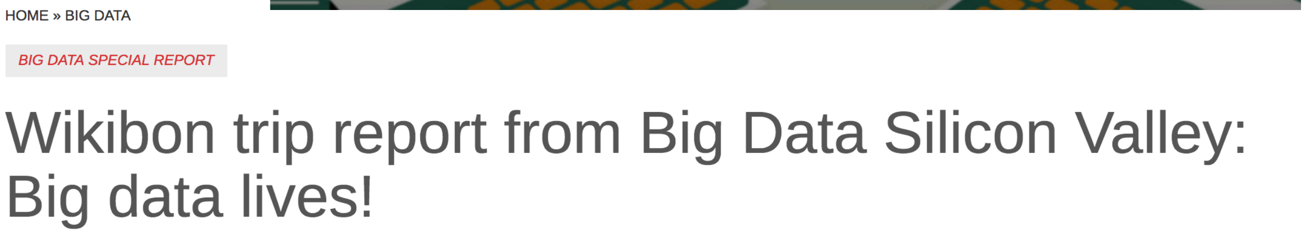
The Cambridge Analytica row shows politics moving in a disturbing direction

Jamie Bartlett



SCIENCE TECHNOLOGY

Swipe Right For Big Data | How Data Affects The Way We Date In 2018



HOME » BIG DATA

BIG DATA SPECIAL REPORT

Wikibon trip report from Big Data Silicon Valley: Big data lives!

Home » Opinion

Is Big Data a threat to free democratic choice?

MAR 21, 2018 @ 10:31 PM 762

The Little Black Book

How Cambridge Analytica Used Big Sleaze To Mine Big Data

BRIEF

Big Data the 'new gold rush,' report predicts

Big Data in Space

What *is* “Big Data”, and why do we care now?

Because of:

① Increased computational power

② Increased collection of data

We now can:

① Run machine learning algorithms that were theoretically useful, but not feasible in practice, ex. random forest

② Use large amounts of data that contain information about behaviors not previously tracked, and feed data-intensive algorithms to find patterns

What *is* “Big Data”, and why do we care now?

Volume: Quantity of data available for analysis

- Gigabytes
- Terabytes
- Petabytes

Variety: Types of data available for analysis

- Structured (financial metrics, demographic data, etc.)
- Unstructured (conversational transcripts, social media, etc.)

Velocity: How quickly data is available for analysis

- Real Time
- Near-Real Time
- Batch

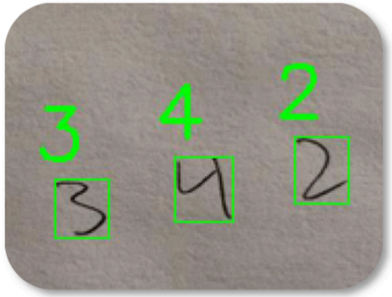
Source: https://github.com/ntlind/Principles_of_AI_and_ML/blob/master/Principles%20of%20AI%20%26%20ML%20vD.pdf

Machine Learning

Machine learning is teaching computers to find patterns in data on their own.

Supervised

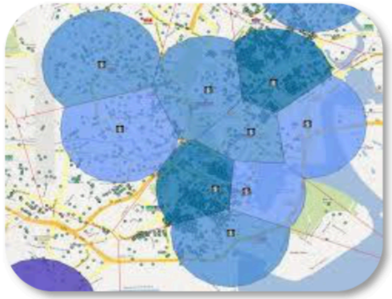
Generate predictions by training on **labeled datasets**



Handwriting Recognition

Unsupervised

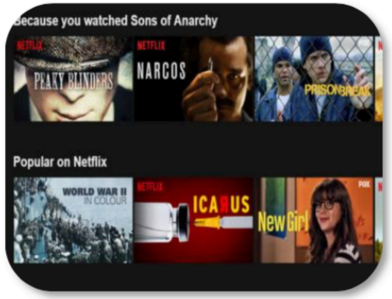
Expose and **visualize** hidden relationships and anomalies in **unlabeled datasets**



Geospatial Market Segmentation

Semi-Supervised

Generate predictions using a **small amount of labeled data** within a larger pool of unlabeled data



Interactive Recommendations

Reinforcement

Create an agent capable of taking environmental actions to **maximize utility** over time

Self-Driving Vehicles

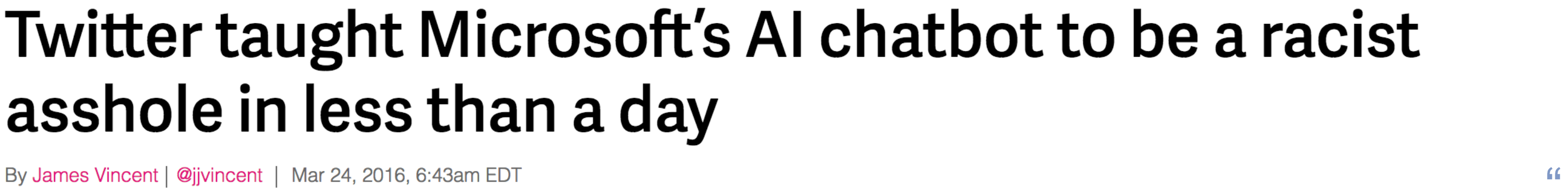
Source: https://github.com/ntlind/Principles_of_AI_and_ML/blob/master/Principles%20of%20AI%20%26%20ML%20vD.pdf

Importantly, what can’t it do (yet)?

1 Machine learning requires **lots of data** – usually, more data/cleaner data will give you better results than a more complex algorithm

> “We don’t have better algorithms than anyone else; we just have more data.” – Peter Norvig, Google

2 Machine learning can only observe what patterns seem to exist in the provided data: it’s positive, not normative.



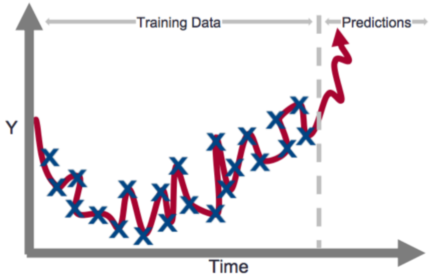
Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day

By James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT

“

3 There’s always a danger of overfitting to the provided data.

Overfitting (High Variance)

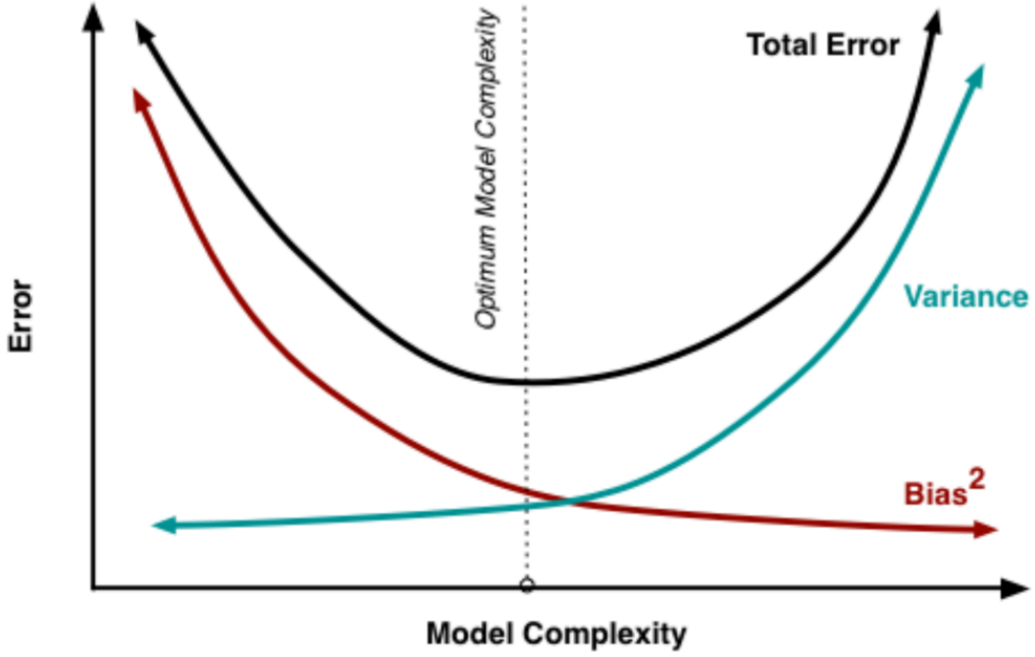


Different algorithms have different pros and cons – understanding how/when to use an appropriate algorithm is essential.

For a given problem, pick the right algorithms…

Supervised		Semi-Supervised
Regression	Classification	Clustering
Linear Regression	Logistic Regression	K-Nearest Neighbors
Multivariate Linear Reg.	Multinomial Logistic Reg.	HCA
Random Forests		PCA
Gradient Boosted Machines		LLE
Support Vector Machines		t-SNE
Multi-Layer Neural Networks		LDA
Recurrent Neural Networks		DBSCAN
Convolutional Neural Networks		Autoencoders
…		….

… to optimize the bias-variance trade-off



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Data science requires familiarity with a variety of tools.

1 Storing/Pulling Data

- Hadoop
- SQL

2 Building Models

- C/C++
- Python
- R
- Matlab

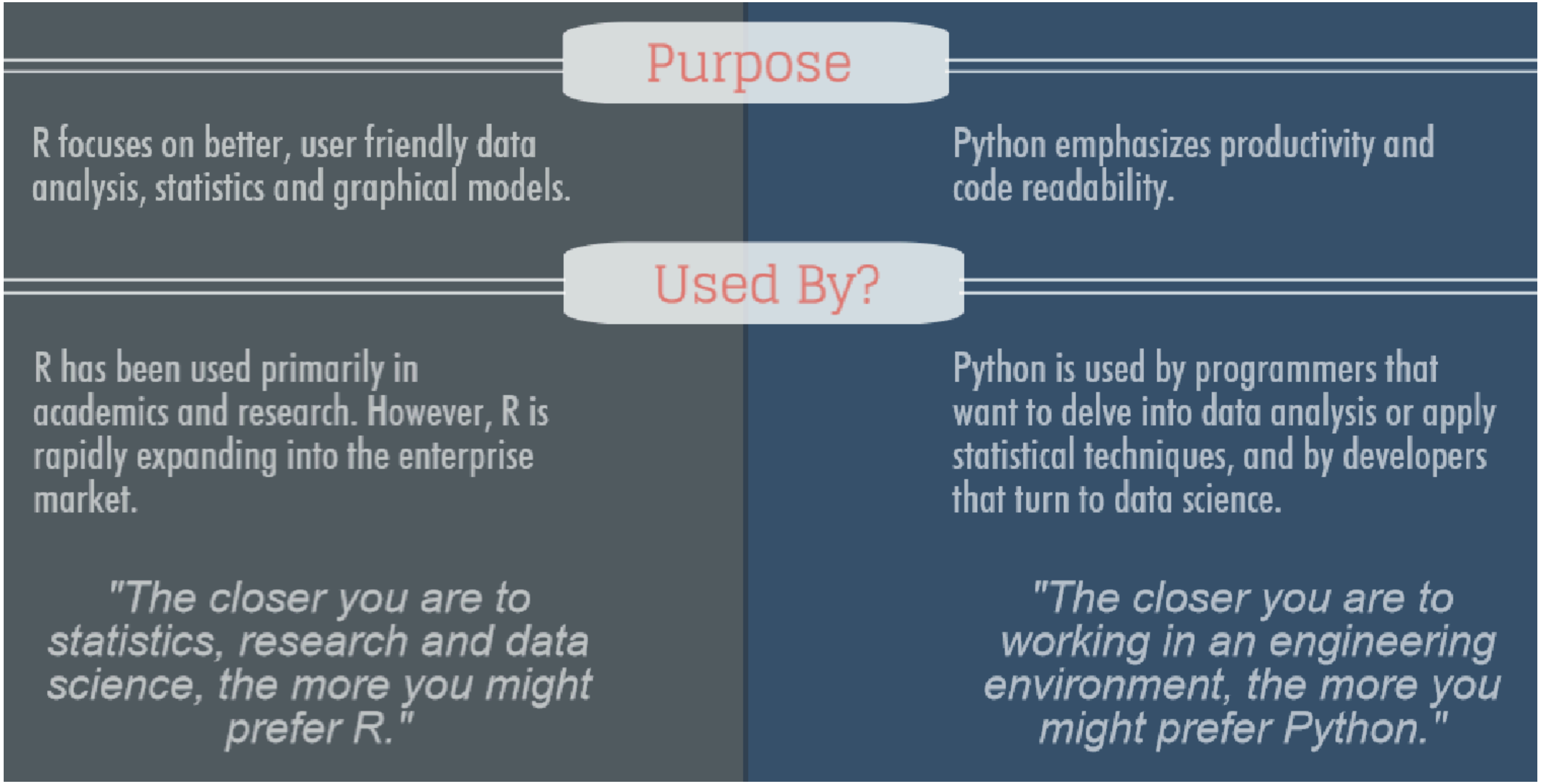
3 Running Models

- Hadoop
- Python
- R
- Matlab
- Spark

4 Data Viz and Presentation

- Python
- R
- Tableau
- Excel
- Powerpoint

Where to start - R or Python?



Purpose

R focuses on better, user friendly data analysis, statistics and graphical models.

Python emphasizes productivity and code readability.

Used By?

R has been used primarily in academics and research. However, R is rapidly expanding into the enterprise market.

Python is used by programmers that want to delve into data analysis or apply statistical techniques, and by developers that turn to data science.

"The closer you are to statistics, research and data science, the more you might prefer R."

"The closer you are to working in an engineering environment, the more you might prefer Python."

Source: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

Bootcamp Outline