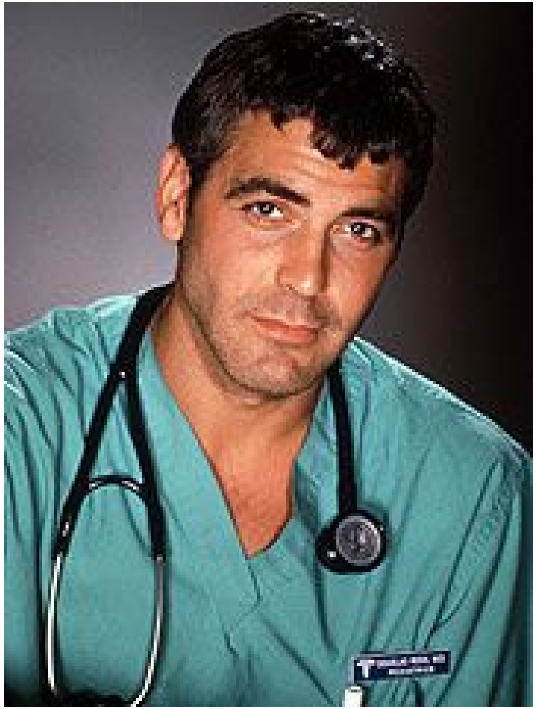
Python For Data Science

Robby Grodin



Robby Grodin
Data Engineer
Wayfair
robby@toypig.co

Goals

- Discuss What Data Science is
- Understand how to analyze data using Pandas
- Learn how to visualize data using matplotlib
- Gain a basic understanding of at least 1 machine learning algorithm

Agenda

- Python Warm-up
- Storing and accessing data in `pandas`
- Data Science Discussion
- Lunch!
- Manipulating dataframes in `pandas`
- Introduction to Data Science with `scikitlearn`
- Visualization with `matplotlib`
- Wrap-up discussion

Tools

- Python
- Pandas
- scikitlearn
- matplotlib
- Anaconda

Python Warm-up

```
names = [‘John Lennon’, ‘Paul McCartney’, ‘George
Harrison’, ‘Pete Best’]
```

1. Print out the names that contain the letter ‘a’
2. Make all of the names lowercase
3. Sort the list of names alphabetically (hint: `sorted()`)
4. Sort the list of names by length
5. Remove all instances of the letter e

Pandas

Fun Fact: The word ‘Pandas’ also refers to adorable bears.

Pandas

- Importable Python module
- Provides high performance Data Structures
- Optimized for data analysis
- Open source

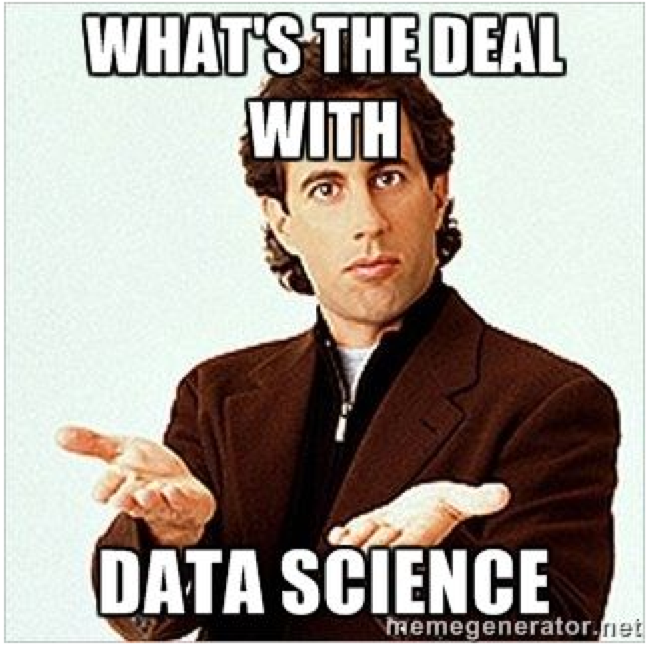
Sales Funnel

1. Open the data set in IPython
2. With a partner, discuss the data.
3. What does it represent?
4. What questions can we ask about it?
5. Is any of the data missing or poorly reported?

Please open an ipython notebook.

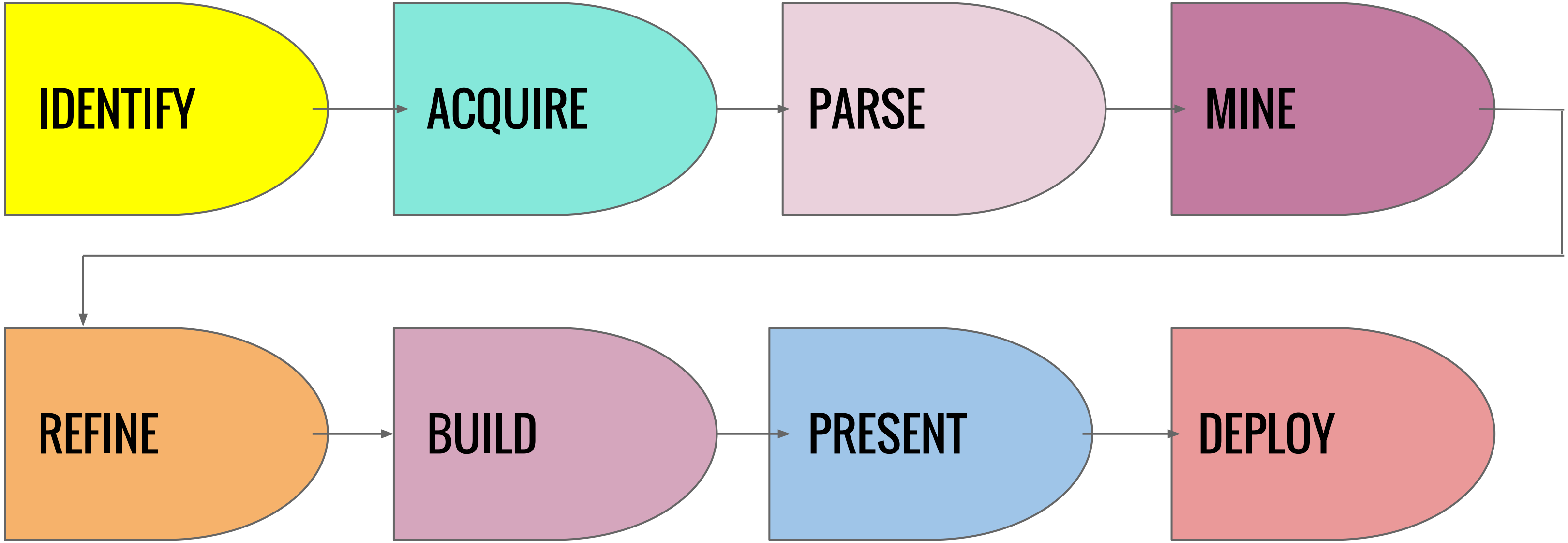
Data Science

The means by which we apply statistical inference to a corpus of data in order to extract insights about the data.



The Analyst	The Engineer
‣ Trains Models	‣ Builds Products
‣ Answers “Why?”	‣ Answers “How”?
‣ Understands Technical Stack	‣ Understands Statistical Analysis
‣ Cleans Data	‣ Cleans Data
‣ Holds Domain Expertise	‣ Holds Domain Expertise

Data Science Workflow



Data Science At Wayfair

- Marketing Analysis
- Business Intelligence
- Personalization
- ???

wayfair

How 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 is really good at recommending movies

- Customers are segmented and clustered
- Features:
  - When user watches and for how long
  - Where the user is watching
  - What device they are watching on
- Neural Networks implement Collaborative Filtering

Cross Validation Analysis

Terms

Feature: A piece of measurable data, i.e. age, height, gender

Target: The value your model is trained to predict

Dependent Variable: Variables whose values depend on the value of Independent Variables

Model: “A specification of a mathematical (or probabilistic) relationship that exists between different variables.” *Grus 2015*

Terms

Cross Validation: The process of splitting data into training and test sets

Training Set: A set of observed data given to an algorithm to provide the basis for a prediction model

Test Set: A set of data whose independent variables are used by the model to produce predictions, which are then compared to the true values to score the model.

Boston Housing

1. Open the data set in IPython
2. Open the data description in your browser
3. With a partner, discuss the data.
4. What does it represent?
5. What questions can we ask about it?
6. Is any of the data missing or poorly reported?

Linear Regression

- Measures the relationship between a scalar dependent variable and one or more independent variables

- Estimated coefficients produce a ‘best fit line’, aka *regression line*
    - $Y = a + b(X)$

- Is scored by judging the sum of the square of the errors in predictions

Statistical Classification

Terms

Classification: The determination of which category(s) an item falls under

Regression: “...the more general problem of fitting any kind of model to any kind of data. This use of the term 'regression' is a historical accident; it is only indirectly related to the original meaning of the word.” *Downey, 2014*

Linear Regression: The process of finding a linear relationship in data that doesn’t naturally line up.

Popular Classification Algorithms

- Random Forest
- Logistic Regression
- Support Vector Machines
- Neural Networks
- k Nearest Neighbors

K Neighbors Classification

- Classification algorithm that clusters based on a system of distance based weighting
- Requires tuning while searching for optimal value of K
- Can be visualized as a Voronoi diagram

Breast Cancer

1. Open the data set in IPython
2. Open the data description in your browser
3. With a partner, discuss the data.
4. What does it represent?
5. What questions can we ask about it?
6. Is any of the data missing or poorly reported?

Next Steps:

➔ Read up on stats, data and engineering

➔ Use Anaconda to play with data in Jupyter

➔ GeneralAssemb.ly/Boston