GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: WELCOME

Robert Doherty

Lead Data Scientist/Engineer, Outbrain

AGENDA

I. Course Overview

iI. What is Data Science?

iiI. Lab

- -Setup Developer Environment
- -Data Exploration

I. COURSE OVERVIEW

Course Overview

Contact Info		Office Hours
Robert Doherty	robdoherty2@gmail.com	by appointment
David McCreary	davidfmccreary@gmail.com	5:30-6:30 PM, M/W
Jarret Petrillo	jarretpetrillo@gmail.com	4:00-6:00 PM, Sun

Class M/W 6:30-9:30 PM - 9/24 - 12/8

- 9/24 - 10/1 - GA East (902 Broadway, 4th Floor), Classroom
- 10/6 - 11/19 - GA West (10 E. 21st St, 4th Floor), Room 4A
- 11/24 - 12/8 - GA East (902 Broadway, 4th Floor), Classroom

Course Overview

Topics

- Regression models and Continuous variables
- Classification, Clustering, and Categorical variables
- Data visualization, NLP, Bayesian inference
- Data Engineering

Assignments

- Kaggle Competitions
- Dataexplor Challenges
- Term Project

Course Overview

Tools

Python Data Science Stack

- Scikit-learn (machine learning)
- Numpy, SciPy (linear algebra, numerical computation)
- Matplotlib (visualization)
- Pandas (modeling, easy-to-use data structures)
- iPython (interactive matlab-style interface)

INTRODUCTIONS

3 minutes:
Partner with someone next to you. Introduce yourselves. Then we will reconvene, and you will introduce your partner to the group.

Please share your partner’s:

- name
- occupation
- experience with Python and scikit-learn
- what s/he is most excited about learning/doing
- what s/he is most apprehensive about learning/doing
- One Weird Trick to Help Me Remember Your Name

II. What is Data Science?

Fun Fact:

- Every Day We Create 2.5 Quintillion Bytes of Data

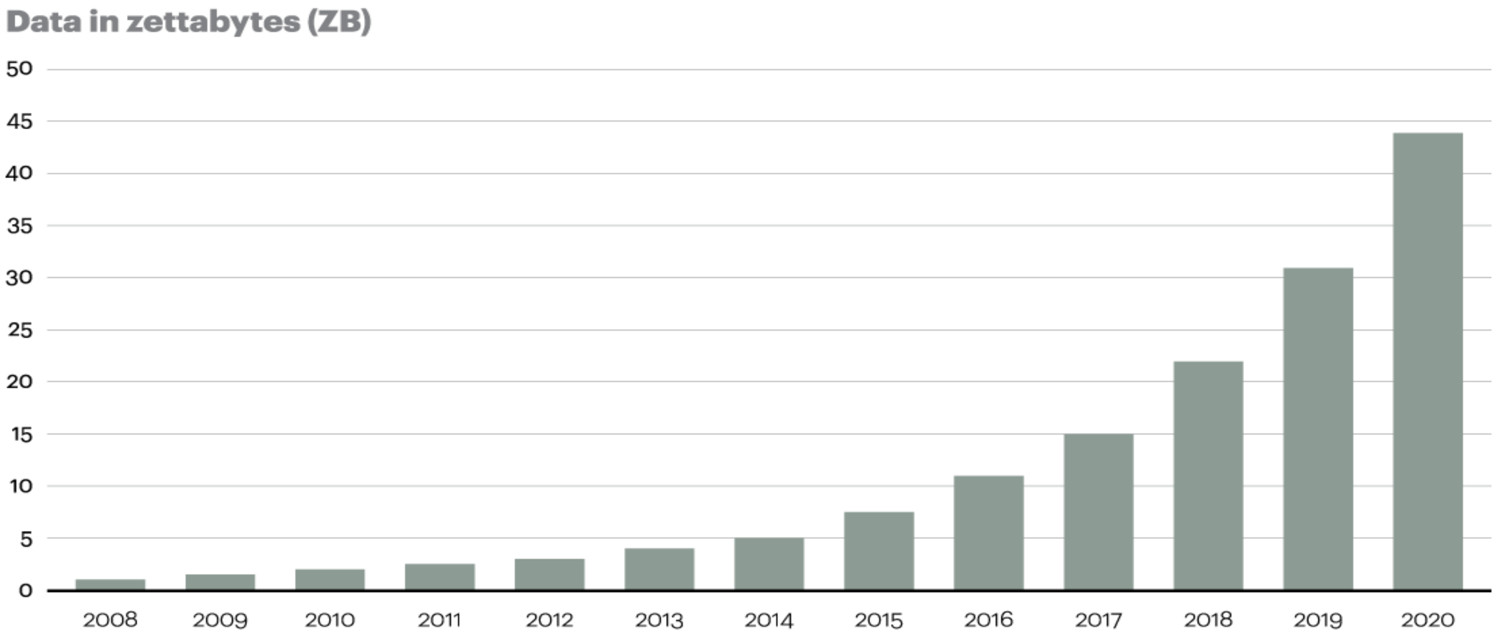
FUN FACT:

- Every Day We Create 2.5 Quintillion Bytes of Data
- 90% of current data was collected in the past two years

FUN FACT:

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



Source: Oracle, 2012

What is Data Science

- A set of tools and techniques used to extract useful information from data.

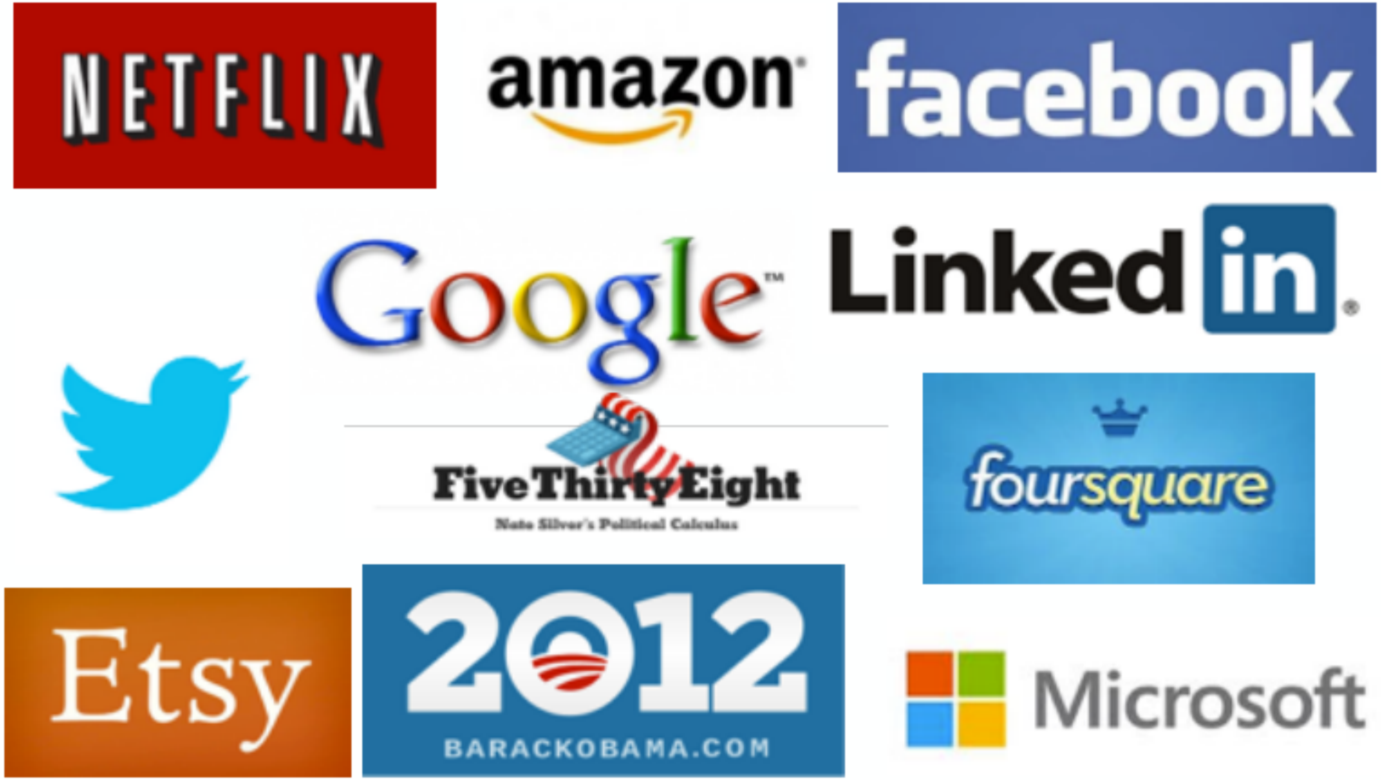
What is Data Science

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject

WHAT IS DATA SCIENCE

- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-oriented subject
- The application of the *Scientific Method* to solving business problems

WHO USES DATA SCIENCE?



WHO USES DATA SCIENCE?

What is a Data Scientist?

- *“a data analyst who lives in California”*

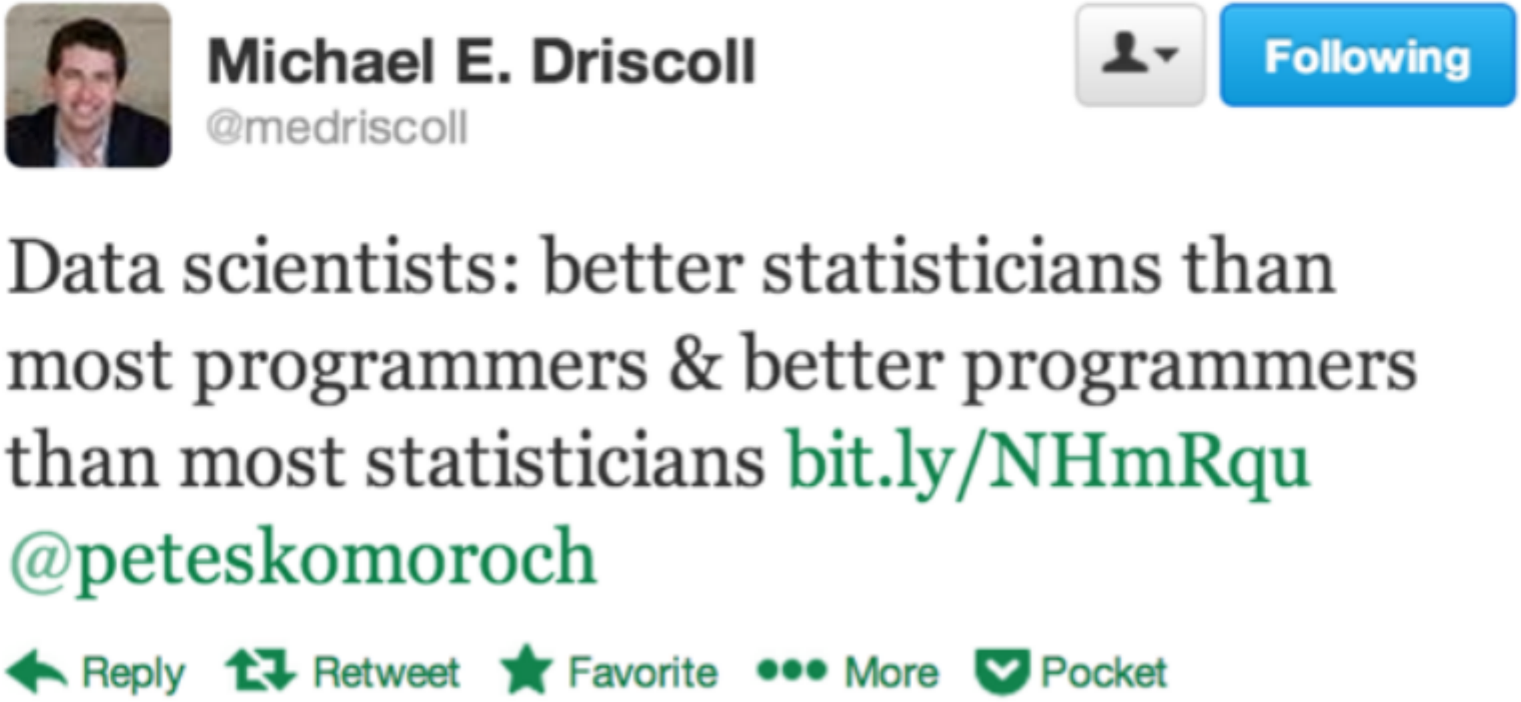
WHAT IS A DATA SCIENTIST?

- *“a data analyst who lives in California”*
- *“a business analyst who lives in New York”*

What is a Data Scientist?

- *“a data analyst who lives in California”*
- *“a business analyst who lives in New York”*
- *“a statistician who lives in San Francisco”*

What is a Data Scientist?



Michael E. Driscoll
@medriscoll

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

What makes a good Data Scientist?

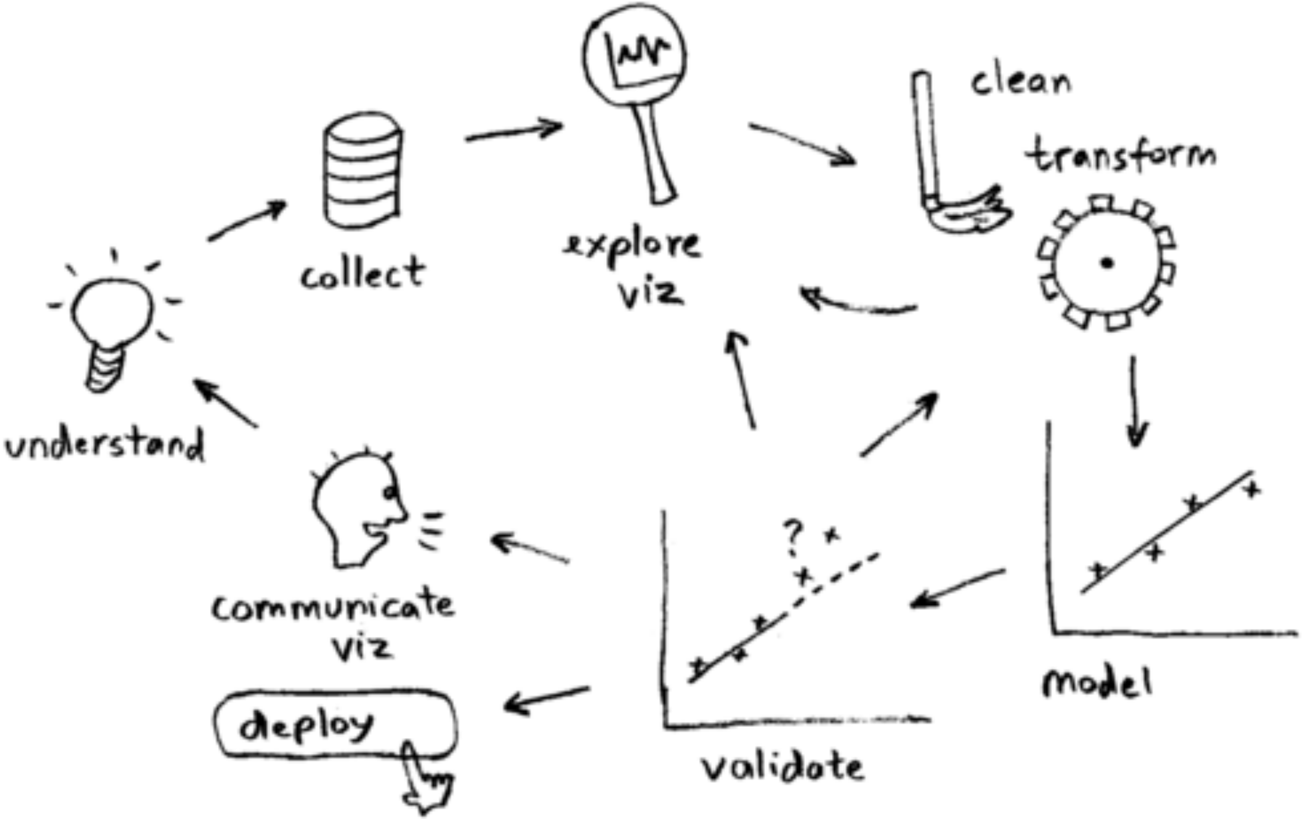
- Statistical inference and Machine Learning knowledge
- Computer Science and Engineering Experience
- Domain expertise
- Communication skills
- Data visualization skills

DATA SCIENCE WORKFLOW

Jeff Hammerbacher, Cloudera:

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

Data Science Workflow



source: DATA SCIENCE TOOLBOX SURVEY RESULTS… SURPRISE! R AND PYTHON WIN

Discussion:

Problem: How would you implement “More items to consider” on Amazon.com?

In a small group, define the process an Amazon Data Scientist would work through to curate the “More items to consider” list for a particular user.

III. Lab: Setup DevEnv