GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: WELCOME

Mark Holt

Biomedical Data Scientist

I. Course Overview

II. What is Data Science?

III. Lab

-Setup Developer Environment

-Data Exploration

I. Course Overview

CONTACT INFO

		OFFICE HOURS
Mark Holt	gmmrgh@gmail.com	by appointment
Susan Sun	@gmail.com	5:30-6:30 PM, M/W

Class M/W 6:30-9:30 PM - 6/3 - 8/19

- GA West (10 E. 21st St, 4th Floor), Room

Topics

- Regression & Classification models
- Visualization & Clustering
- NLP, Bayesian inference, AB Testing
- Data Engineering

Assignments

- Dataexplorer Challenges
- Term Project

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)

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

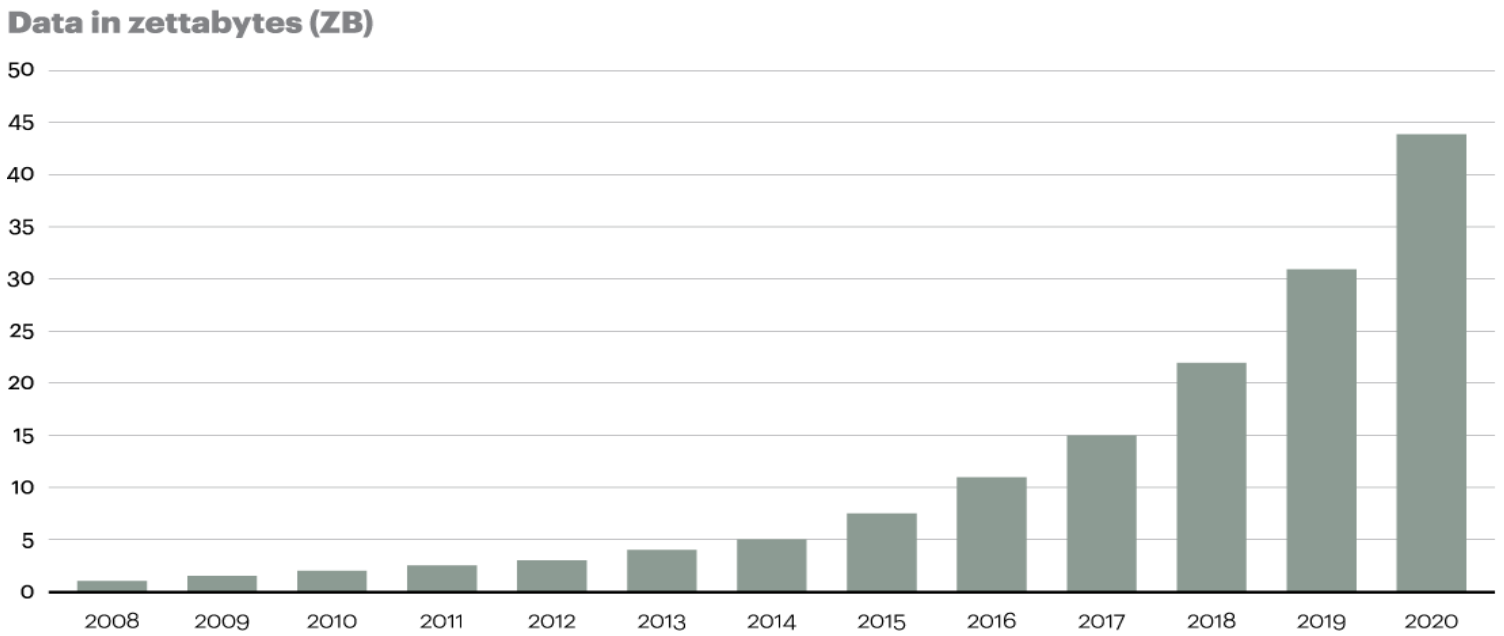
II. What is Data Science?

- Every Day We Create 2.5 Quintillion Bytes of Data

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

Figure 1

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



Source: Oracle, 2012

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

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

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


Google


facebook


Baidu 百度


Microsoft

WHO USES DATA SCIENCE?

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

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

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

Michael E. Driscoll
@medriscoll

Following

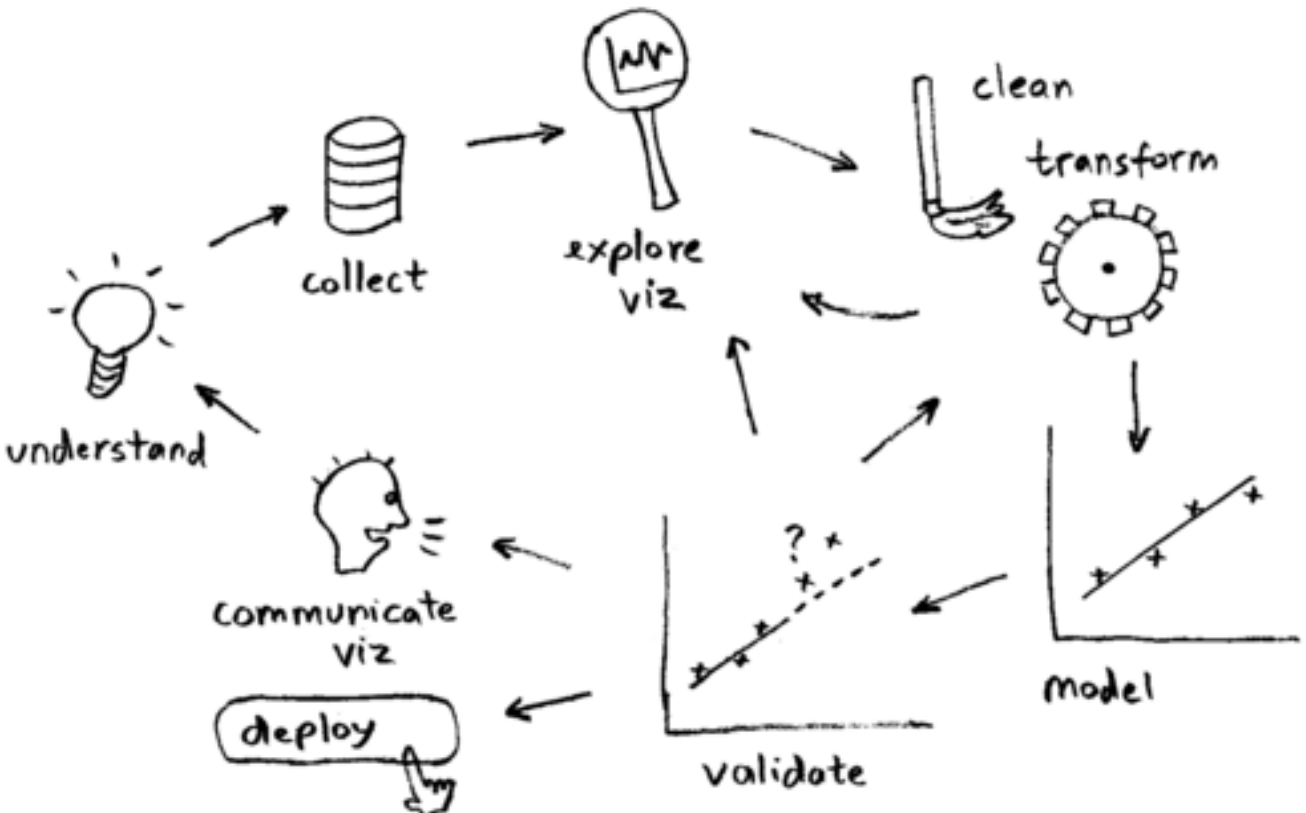
Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

Reply Retweet Favorite More Pocket

- Software Engineering
- Machine Learning
- Domain expertise
- Communication
- Visualization
- Stats
- Math

Jeff Hammerbacher, Cloudera, Mount Sinai:

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



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

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