GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: WELCOME

Mark Holt

Biomedical Data Scientist

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

Mark Holt	mrgh@me.com	by appointment
Susan Sun	sun.w.susan@gmail.com	

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

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

Course Overview

Topics

- Gradient Descent & Linear Regression
- Logistic Regression
- Basic Visualization
- Regularization
- Unsupervised Learning - Clustering
- Dimensionality Reduction
- Bayesian inference & AB Testing
- Natural Language Processing
- Recommendation Systems
- Support Vector Machines
- Databases
- Plus Guest Speakers

COURSE OVERVIEW

SKILLS

- PROGRAMMING
- PRESENTATION & COMMUNICATION

COURSE OVERVIEW

ASSIGNMENTS

- DATAEXPLORER CHALLENGES
- TERM PROJECT
- HOW MUCH HOMEWORK????

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 INTERFACE)

Course Etiquette & Info

- There is absolutely no such thing as a stupid question
- No-one knows everything about data science so please respect your fellow students
- feel free to eat/drink during class, but try to do it without disturbing the class
- Please remember that you all come from different backgrounds, what is simple for some may be challenging for others. If you are ahead of the class, or finding it easy going, say during a lab, don't disturb the rest of the class with loud side conversations
- There will be 2 breaks during the class of about 10 mins each

INTRODUCTIONS

Partner with someone next to you. Introduce yourselves. When we reconvene, 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
- something unusual or quirky about them to help ME remember your names

II. What is Data Science?

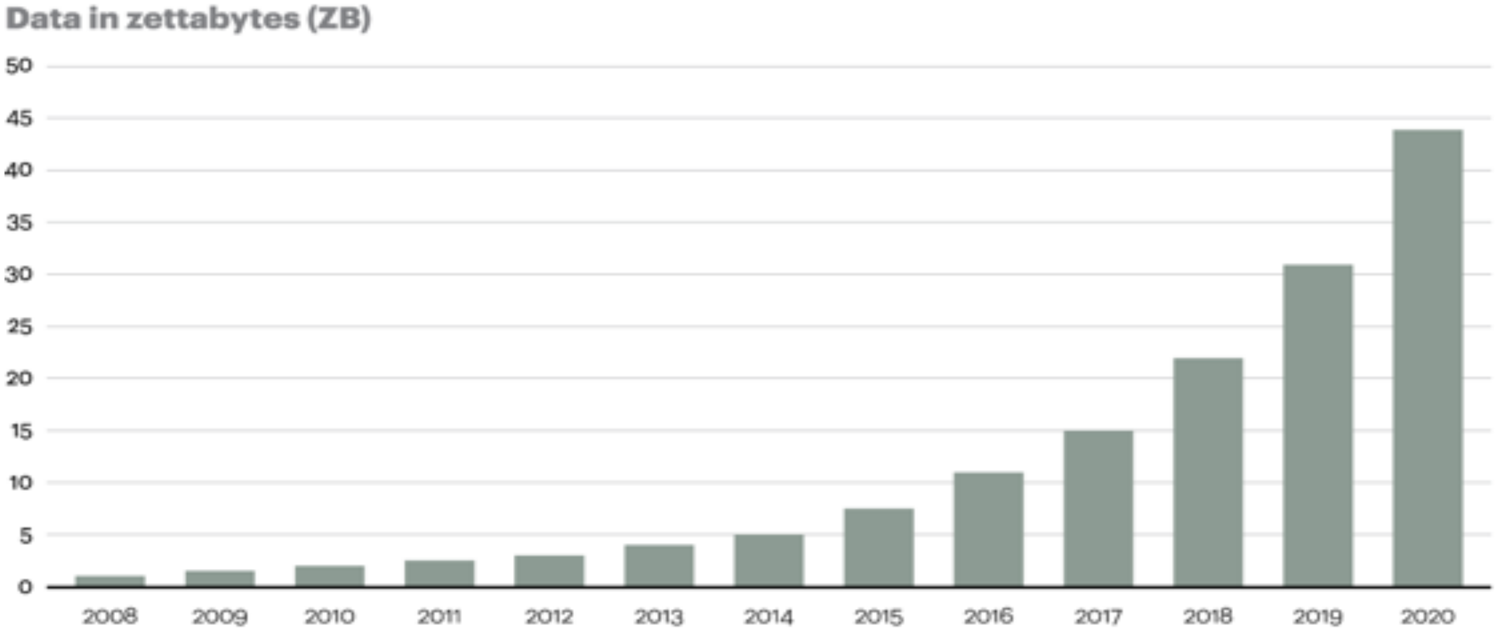
FUN FACT:

- 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 (aka knowledge) from data
- An interdisciplinary, problem-oriented subject
- The application of the *Scientific Method* to solving business problems
- An art

WHO USES DATA SCIENCE?

Google

facebook

Bai du 百度

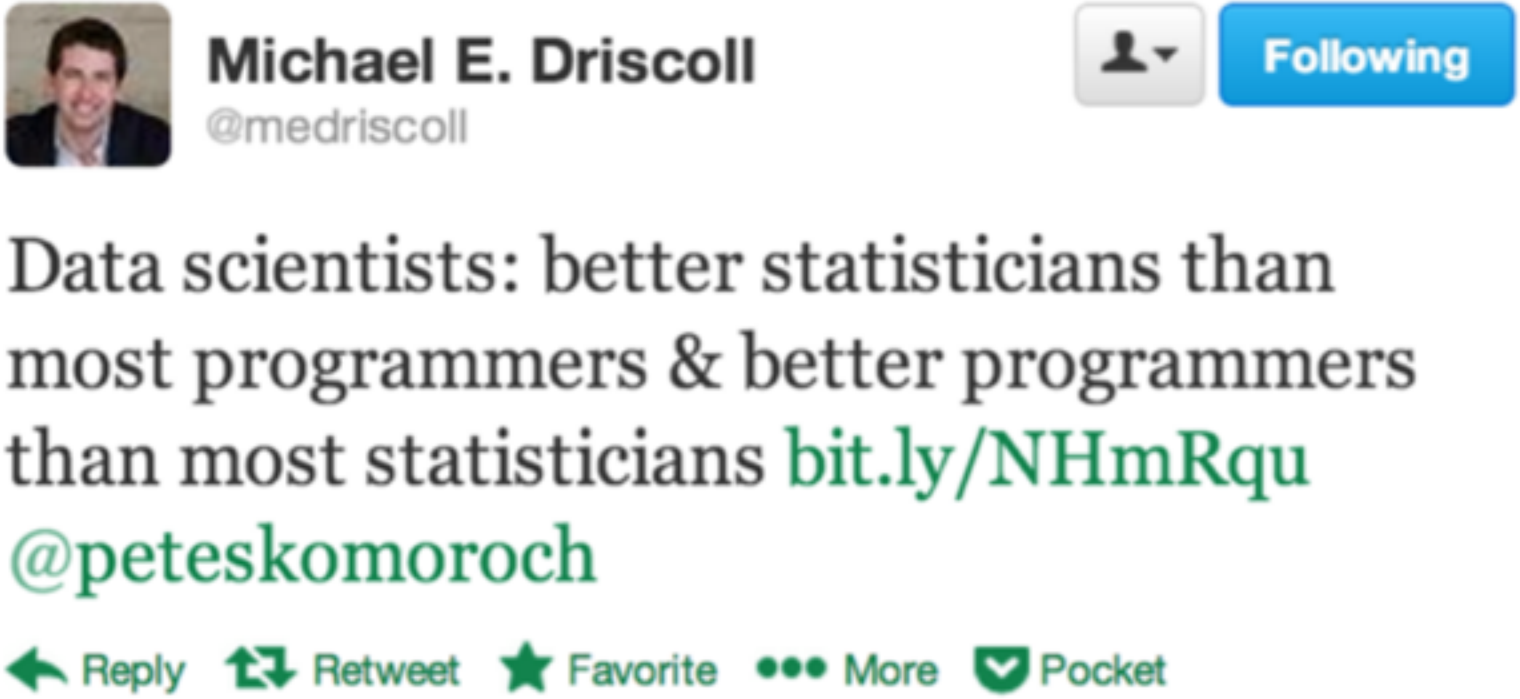
Microsoft

WHO USES DATA SCIENCE?

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?

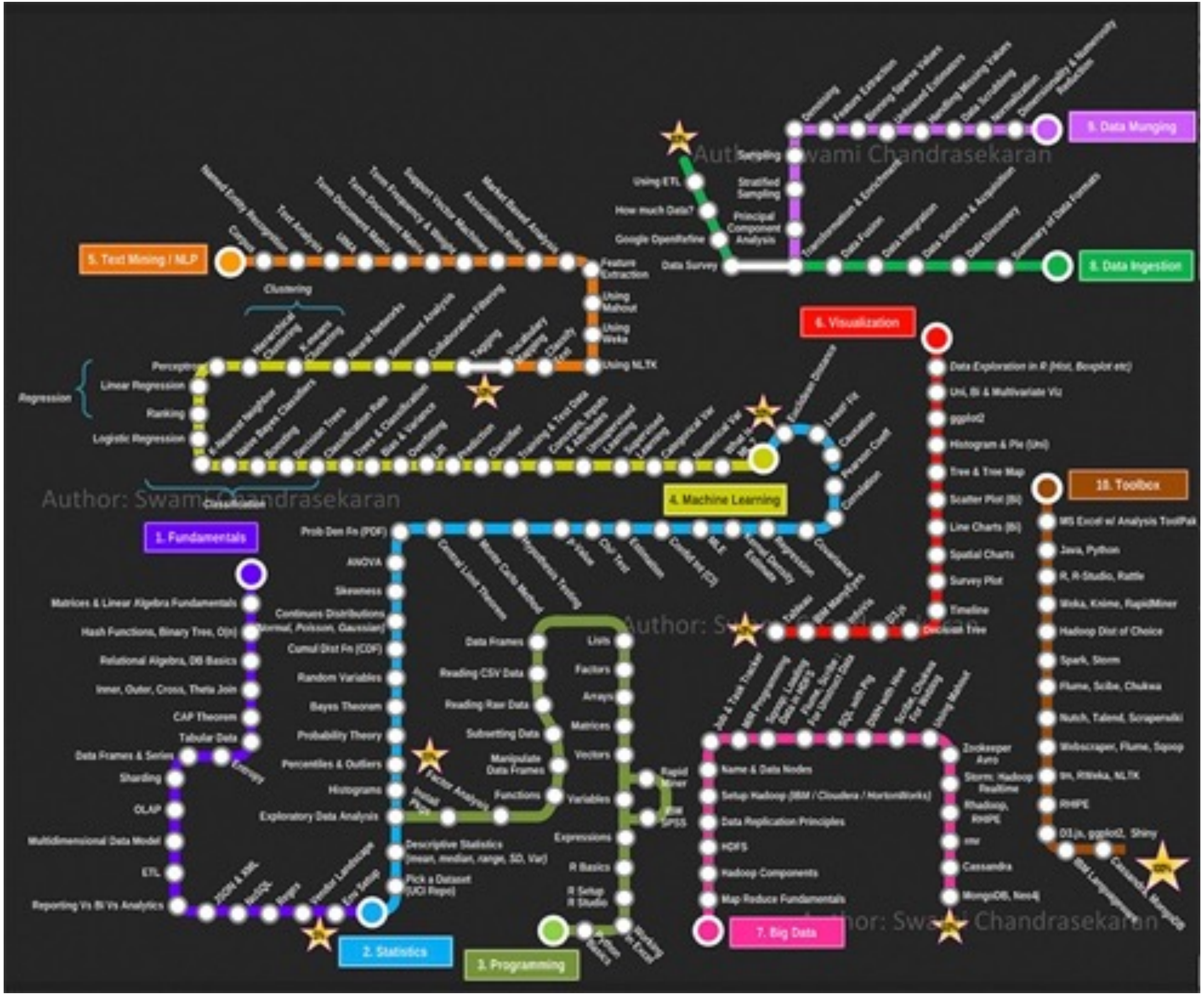


Michael E. Driscoll
@medriscoll

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

What makes a good Data Scientist?

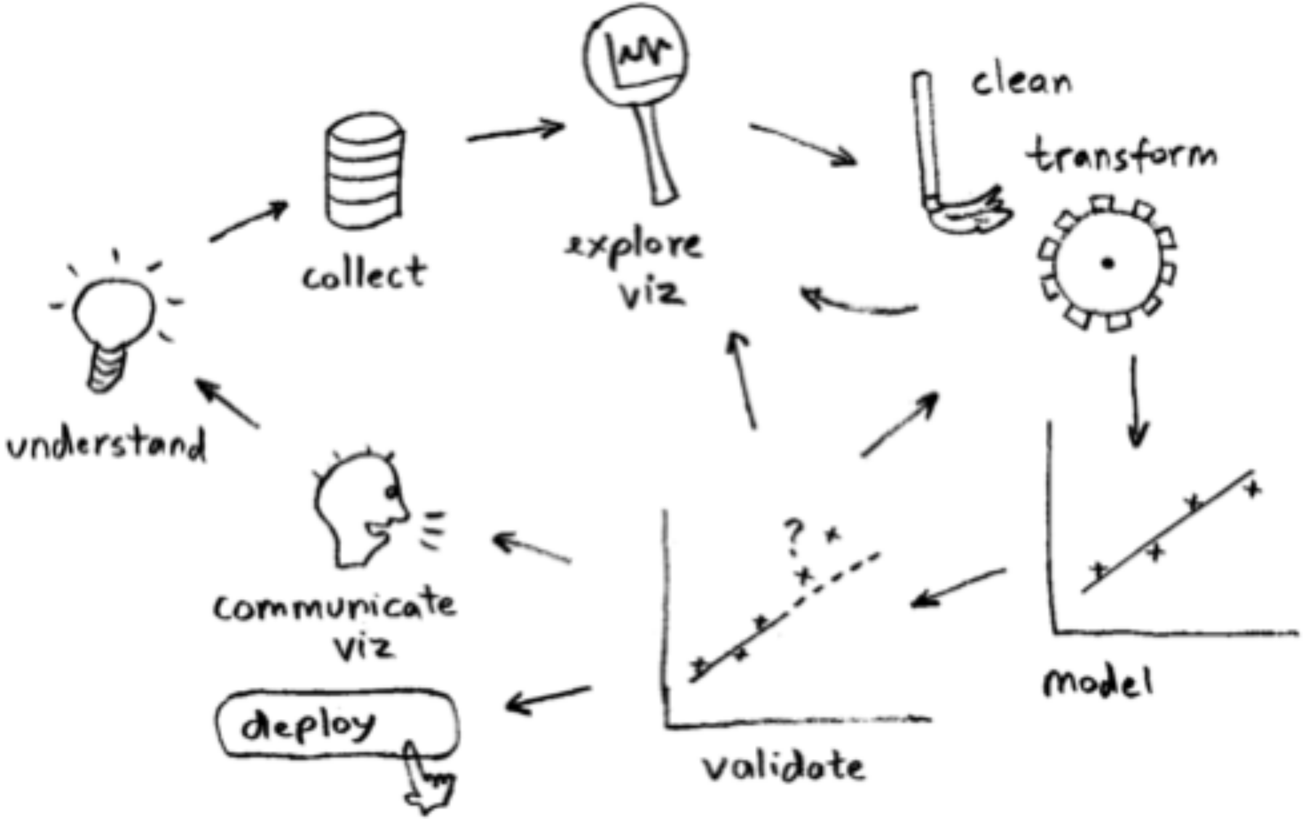
- Software Engineering
- Machine Learning
- Domain expertise
- Communication
- Visualization
- Stats
- Math
- Intuition


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Entropy
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
2. Statistics
Pick a Dataset (UCI Repo)
Descriptive Statistics (mean, median, range, SD, Var)
Exploratory Data Analysis
Histograms
Percentiles & Outliers
Probability Theory
Bayes Theorem
Random Variables
Cumul Dist Fn (CDF)
Continues Distributions (Normal, Poisson, Gaussian)
Skewness
ANOVA
Prob Den Fn (PDF)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
Pearson Coeff
Causation
Least² Fit
Euclidean Distance
3. Programming
Python Basics
Working in Excel
R Setup
R Studio
R Basics
Expressions
Variables
Vectors
Matrices
Arrays
Factors
Lists
Rapid Miner
IBM SPSS
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Factor Analysis
Scatter Plots
4. Machine Learning
What is ML?
Numerical Var
Categorical Var
Supervised Learning
Unsupervised Learning
Concepts, Inputs & Attributes
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Classification
Classification Rate
Decision Trees
Boosting
Naive Bayes Classifiers
K-Nearest Neighbor
Classification
Logistic Regression
Ranking
Linear Regression
Regression
Perceptron
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
Classify Text
Vocabulary Mapping
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
D3.js
InfoVis
IBM ManyEyes
Tableau
7. Big Data
Map Reduce Fundamentals
Hadoop Components
HDFS
Data Replication Principles
Setup Hadoop (IBM / Cloudera / HortonWorks)
Name & Data Nodes
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe: For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
8. Data Ingestion
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichment
Data Survey
Google OpenRefine
How much Data?
Using ETL
9. Data Munging
Dimensionality & Numerosity Reduction
Normalization
Data Scrubbing
Handling Missing Values
Unbiased Estimators
Binning Sparse Values
Feature Extraction
Denoising
Sampling
Stratified Sampling
Principal Component Analysis
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
Author: Swami Chandrasekaran

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

DATA SCIENCE WORKFLOW



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

DISCUSSION:

Problem: You are a top Internet news site competing hard in a tough environment. You have a good story but need a good headline to get readers to enter your site to read the story. Your advertising dollars depend on your headline attracting the most readers in the shortest possible time. How would you approach solving this problem??

In a small group, consider how you might go about solving this problem

III. Lab: Setup DevEnv