GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: WELCOME

Mark Holt
Biomedical Data Scientist

1. COURSE OVERVIEW
2. WHAT IS DATA SCIENCE?
3. LAB
    - 3.1. SETUP DEVELOPER ENVIRONMENT
    - 3.2. DATA EXPLORATION

COURSE OVERVIEW

CONTACT INFO

Mark Holt mrgh@me.com

Susan Sun sun.w.susan@gmail.com

OFFICE HOURS

by appointment

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

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

TOPICS

- LINEAR MODELS
  - GRADIENT DESCENT & REGRESSION
  - LOGISTIC REGRESSION
  - REGULARIZATION
- VISUALIZATION
- UNSUPERVISED LEARNING - CLUSTERING
- DIMENSIONALITY REDUCTION
- BAYESIAN INFERENCE & AB TESTING
- NATURAL LANGUAGE PROCESSING
- RECOMMENDATION SYSTEMS
- SUPPORT VECTOR MACHINES
- DATABASES
- GUEST SPEAKERS

SKILLS

- PROGRAMMING
- MODELLING
- PRESENTATION & COMMUNICATION

ASSIGNMENTS

- DATAEXPLORER CHALLENGES
- LAB PREP
- TERM PROJECT

TOOLS

- PYTHON DATA SCIENCE STACK
  - IPYTHON (INTERACTIVE PYTHON)
  - SCIKIT-LEARN (MACHINE LEARNING)
  - NUMPY (NUMERICAL PYTHON - LINEAR ALGEBRA)
  - SCIPY (SCIENTIFIC PYTHON)
  - PANDAS (DATA ANALYSIS LIBRARY)
  - MATPLOTLIB (VISUALIZATION * PLOTTING)

Course Etiquette & Info

- There is absolutely no such thing as a stupid question - seriously
- No-one knows everything about data science
- 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

Format of the Classes (very rough guide):

- **Homework review - 30 mins**
- **Slides for current class - 20 min**
- **Break**
- **iPython teaching notebooks - 50 min**
- **Break**
- **Lab work based on previous classes teaching notebooks - 50 min**

- **Format will change from time-to-time with project work, guest speakers, etc etc etc**

Introductions

Partner with someone next to you.
Introduce yourselves.
When we reconvene, you will introduce your classmate to the group.

Please share your partner's:

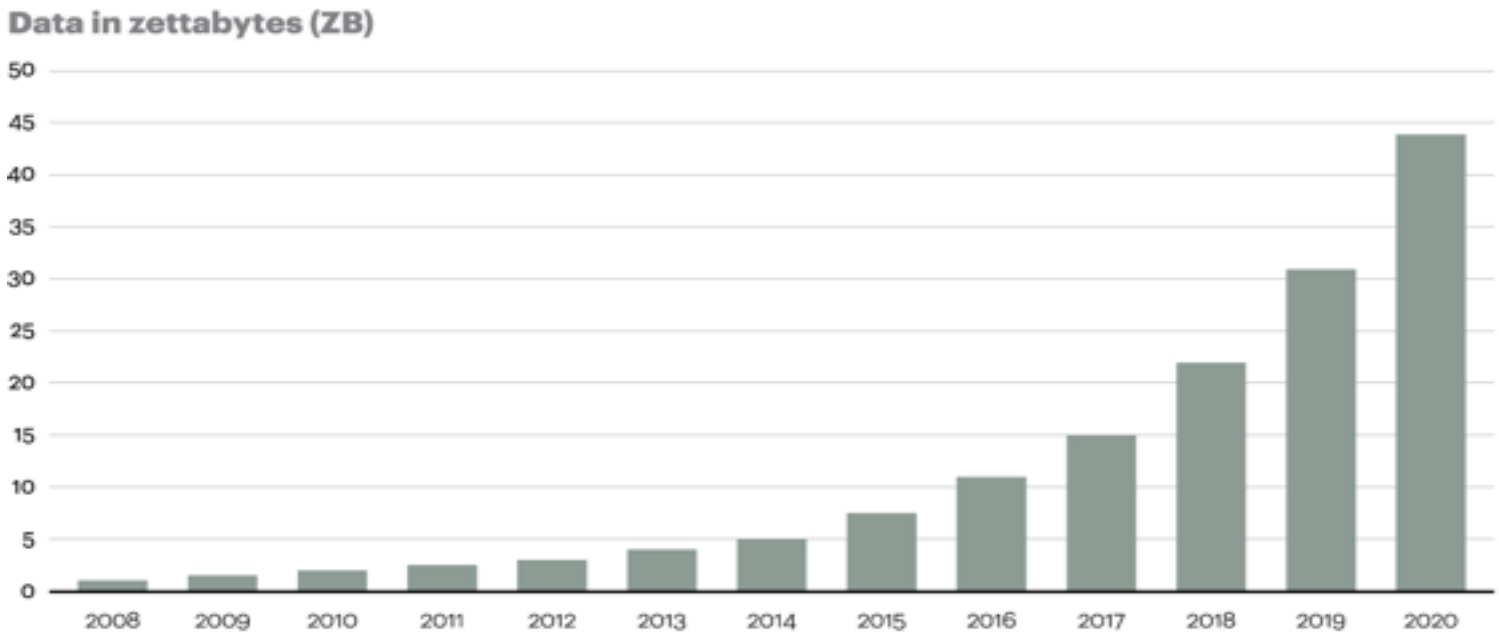
- name
- what they do
- together complete the skills questionnaire
- what s/he is most excited about learning/doing
- what s/he is most apprehensive about learning/doing
- something unusual or quirky about them to help ME remember their names

WHAT IS DATA SCIENCE?

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


Google


facebook


Baidu 百度


Microsoft


Biogen


SPIRE


SOCURE

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

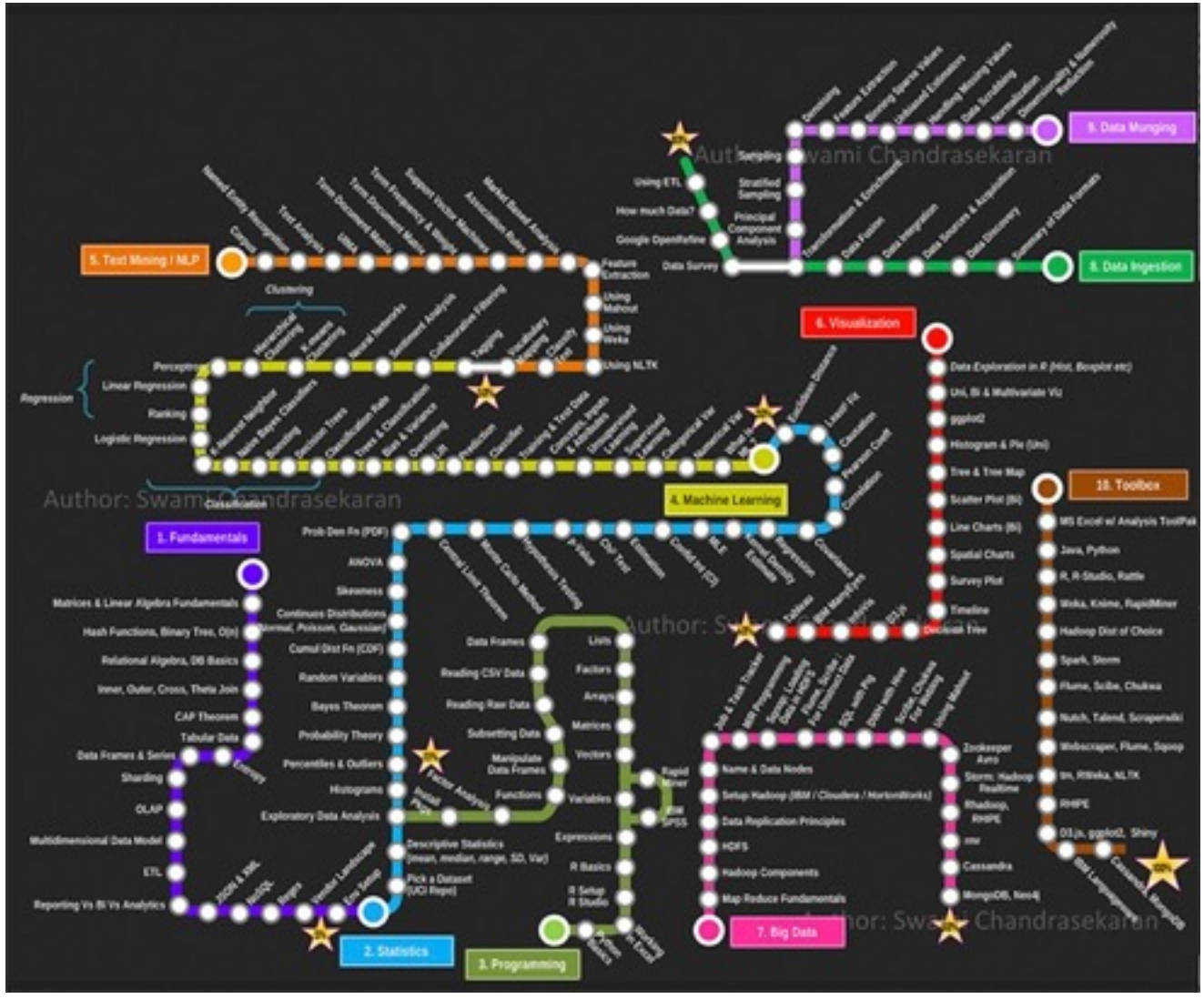
Michael E. Driscoll
@medriscoll

Following

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

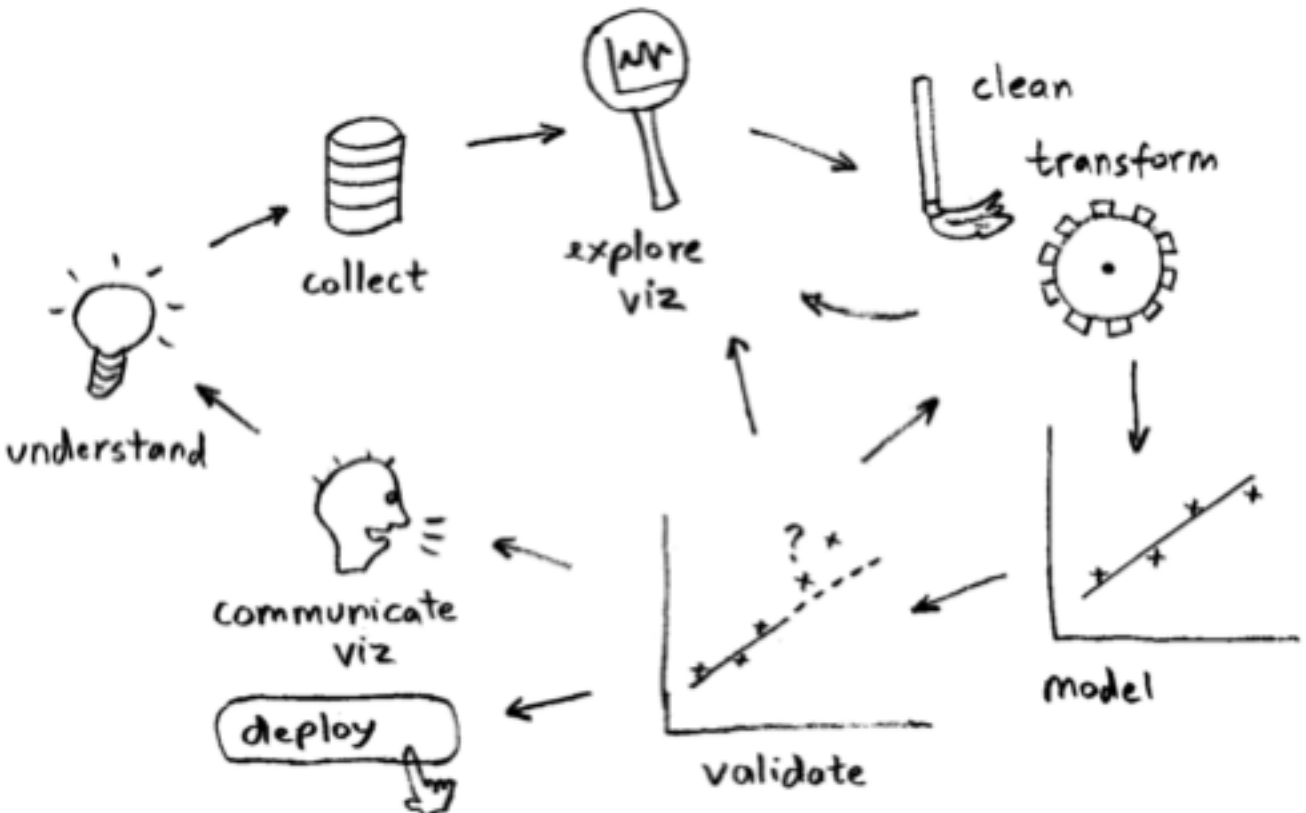
Reply Retweet Favorite More Pocket

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


Author: Swami Chandrasekaran
9. Data Munging
Sampling
Stratified Sampling
Principal Component Analysis
Using ETL
How much Data?
Google OpenRefine
Data Survey
Transformation & Enrichment
Data Frames
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
8. Data Ingestion
5. Text Mining / NLP
Clustering
Feature Extraction
Using Mahout
Using Weka
Using NLTK
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
Regression
Linear Regression
Ranking
Logistic Regression
Classification
4. Machine Learning
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scribe, Chukwa
Nutch, Talend, Scrapenutki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
d3.js, ggplot2, Shiny
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
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
Prob Den Fn (PDF)
ANOVA
Skewness
Continues Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
2. Statistics
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
SPSS
3. Programming
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper
Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
mr
Cassandra
MongoDB, Neo4j
7. Big Data

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: 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

LAB: SETUP DEVENV