

INTRO TO DATA SCIENCE

LECTURE 1: WELCOME

Mark Holt
Biomedical Data Scientist

1.INTROS & INFO

2.SETUP DEVELOPER ENVIRONMENT

3.COURSE OVERVIEW

4.WHAT IS DATA SCIENCE?

INTRO TO DATA SCIENCE

COURSE OVERVIEW

CONTACT INFO

Mark Holt mrgh@me.com

Susan Sun sun.w.susan@gmail.com

Chris Sheehan rwc.sheehan@gmail.com

OFFICE HOURS

by appointment

to be advised

to be advised

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

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

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)**
 - **SCIPLY (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 - 0-20 min
- Break
- iPython teaching notebooks - 50 - 70 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

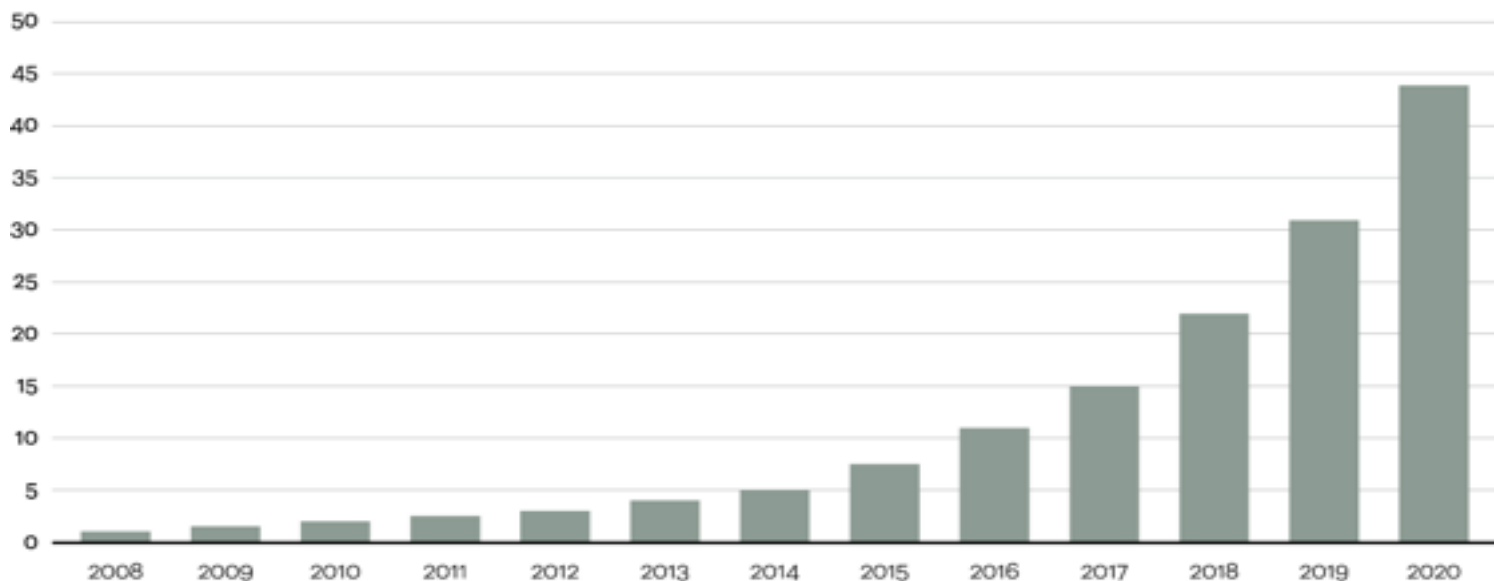
WHAT IS DATA SCIENCE?

- 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

Data in zettabytes (ZB)



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

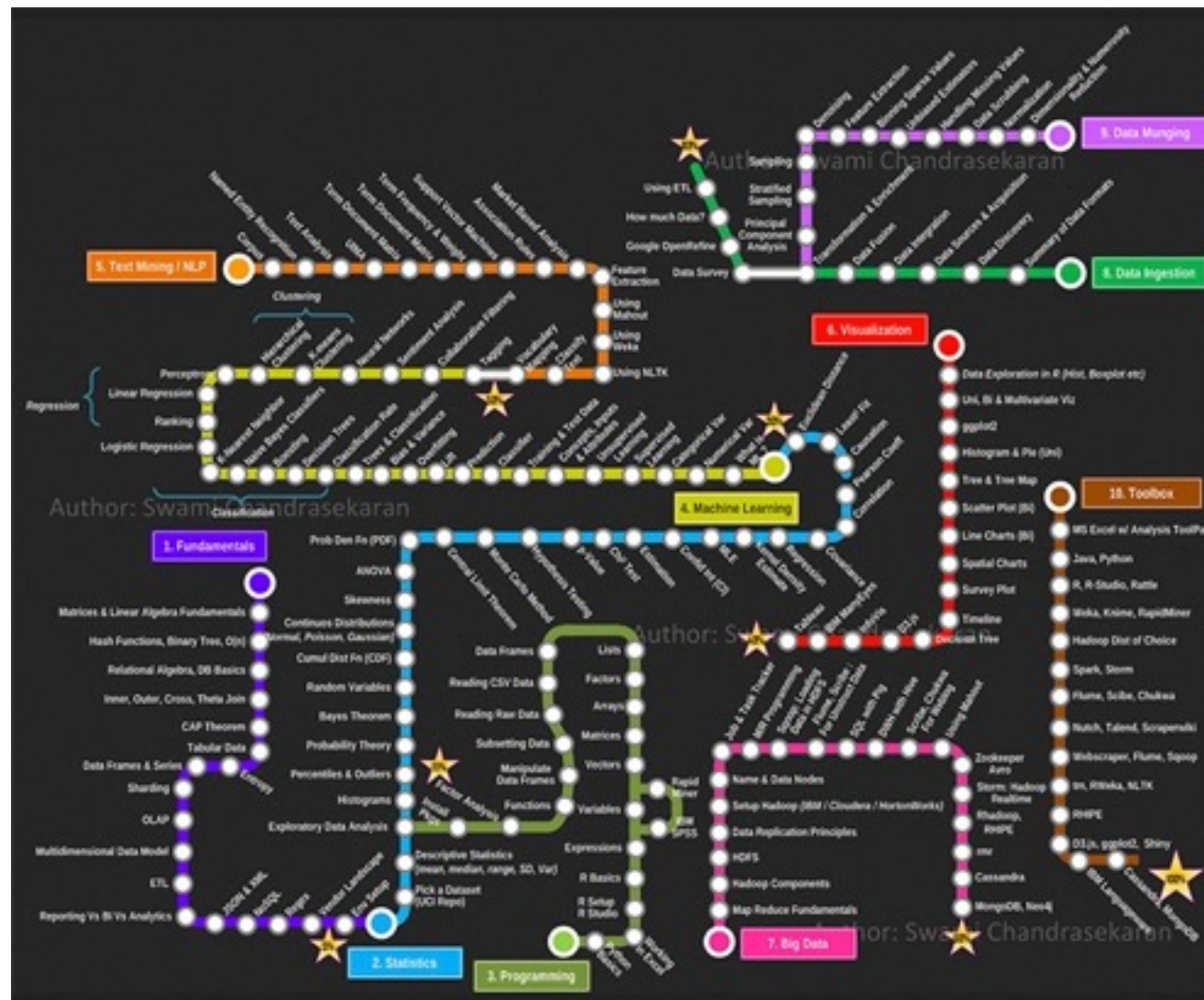


Following

Data scientists: better statisticians than
most programmers & better programmers
than most statisticians bit.ly/NHmRqu
[@peteskomoroch](#)

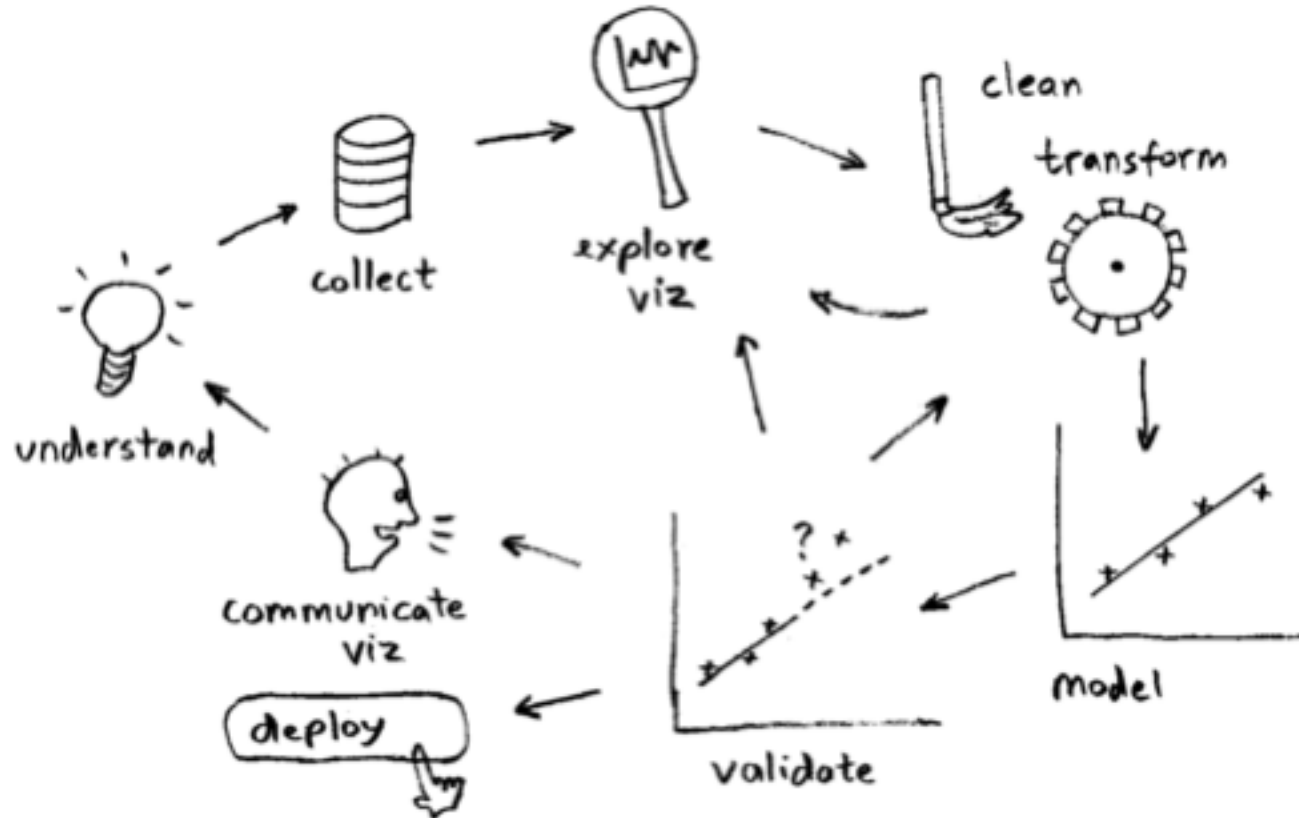
Reply Retweet Favorite More Pocket

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



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



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