INTRO TO DATA SCIENCE
LESSON 1: DATA EXPLORATION

WELCOME!

CONTACT

Ed Podojil   epodojil@gmail.com

Dave Goodsmith   goodsmith@scaleanalytics.com

Class M/W 6:30-9:30

OFFICE HOURS

9-9:30 and Google Hangouts (schedule accordingly)

COURSE NOTES, WIKI, AND LAB/PROJECT SUBMISSIONS

teachingdatascience.github.io/data-science-course/

I. WHAT IS DATA SCIENCE?
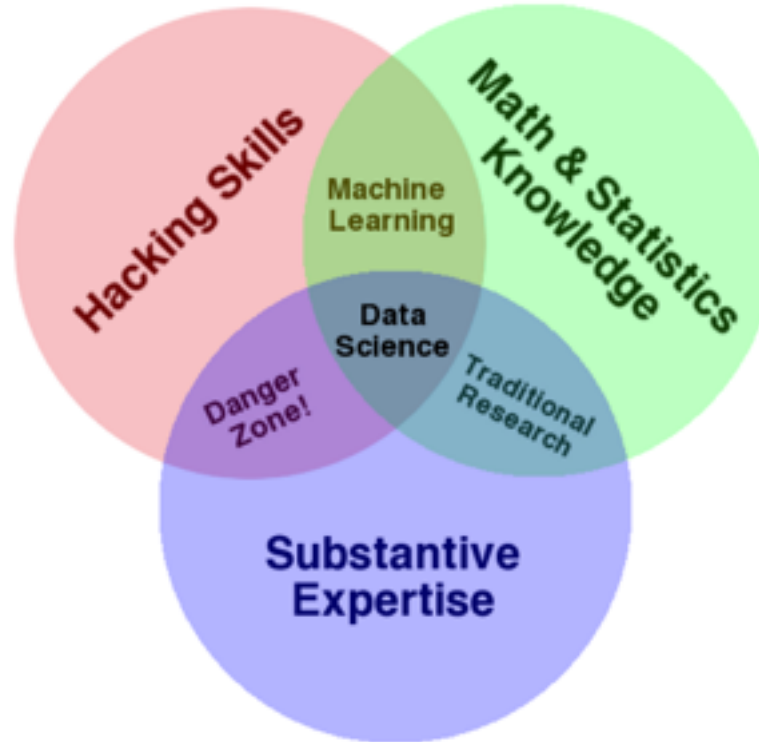# II. THE DATA MINING WORKFLOW

LAB:
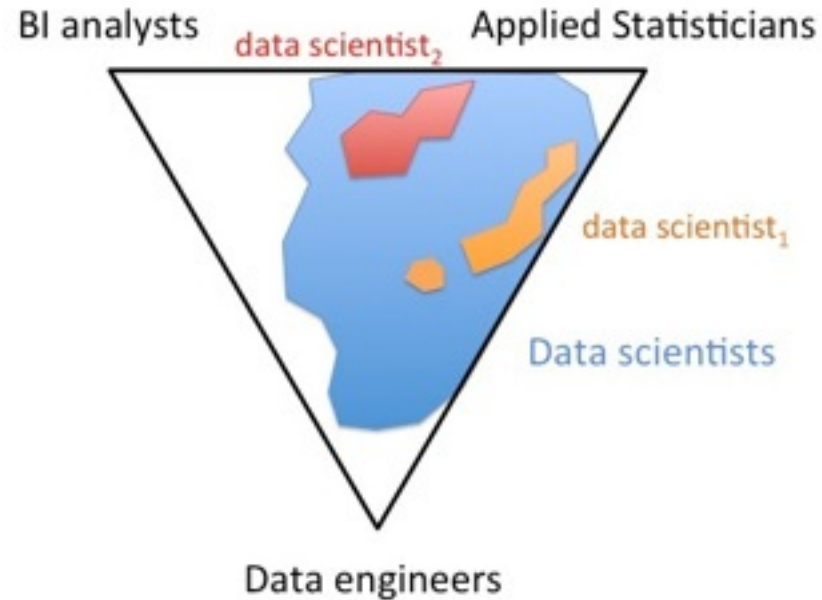# III. COMPUTER SETUP
IV. DATA PRACTICE

I. WHAT IS DATA SCIENCE?

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.

source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

source: http://www.p-value.info/2012/12/what-is-data-scientist.html

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

An interdisciplinary, problem-solving oriented subject.

The application of scientific techniques to practical problems.

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

An interdisciplinary, problem-solving oriented subject.

The application of scientific techniques to practical problems.
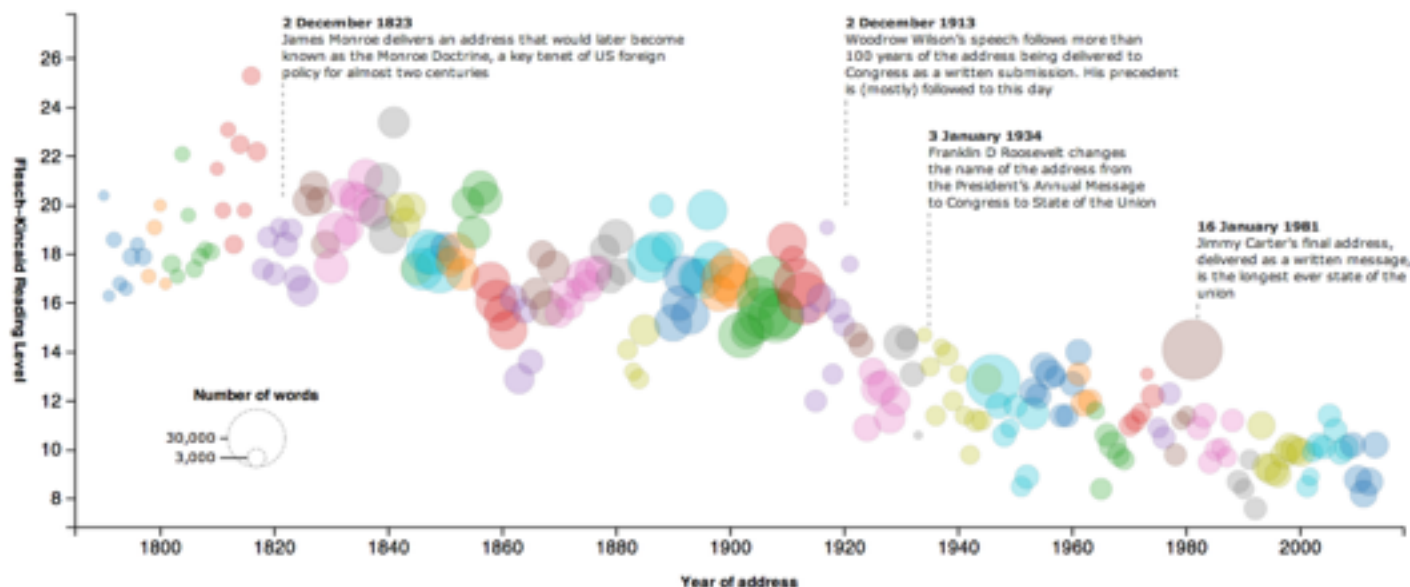
A rapidly growing field.

The state of our union is ... dumber:
How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every state of the union

🐦 "'Data Scientist' is a Data Analyst who lives in California"

🐦 "A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

🐦 "A data scientist is a business analyst who lives in New York."

🐦 "A data scientist is a statistician who lives in San Francisco."

🐦 "Data Science is statistics on a Mac."

source: http://www.youtube.com/watch?v=h9vQIPfe2uU

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

- Statistical and machine learning knowledge

- Computer Science and Engineering experience

- Academic curiosity

- Product sense

- Storytelling and communication skills

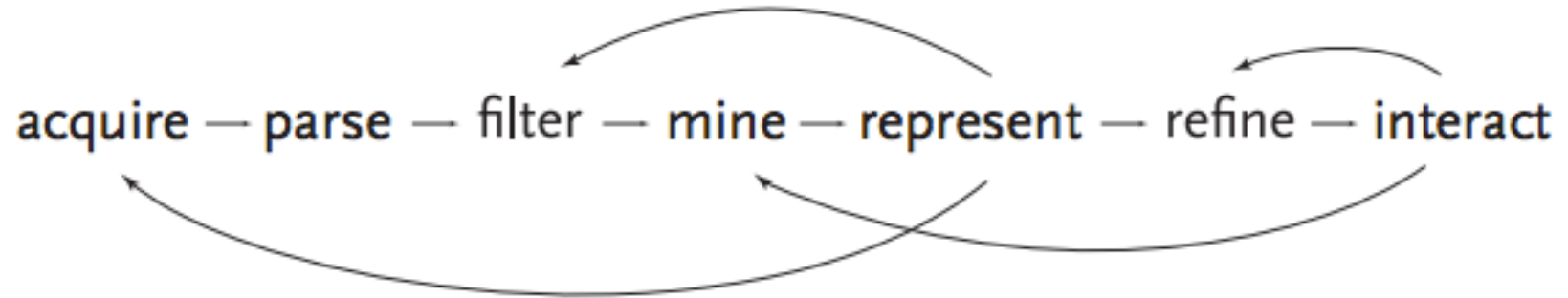# II. THE DATA SCIENCE WORKFLOW

Dataists blog

1. Obtain

2. Scrub

3. Explore

4. Model

5. Interpret

Jeff Hammerbacher: Chief Scientist, 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

Ben Fry: Principal, Fathom

COMPUTER SCIENCE	MATHEMATICS, STATISTICS, AND DATA MINING	GRAPHIC DESIGN	INFOVIS AND HCI
acquire parse	filter mine	represent refine	interact

source: http://benfry.com/phd/dissertation-110323c.pdf

acquire — parse — filter — mine — represent — refine — interact

Zip Decode
http://benfry.com/zipdecode/

source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge
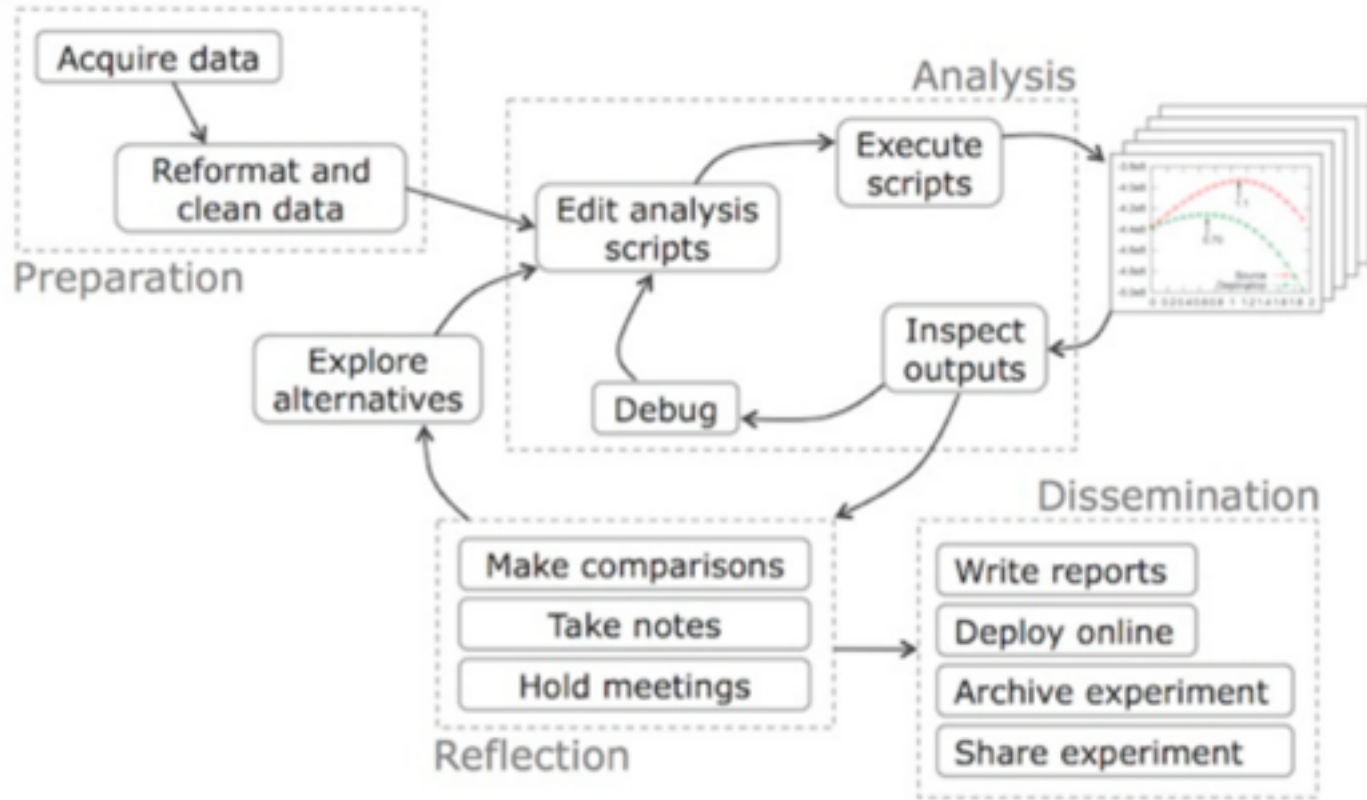
BUILDING AN ANALYTICS TEAM

1. Define the top priorities of the organization
2. Determine the data you'd like to collect

What will your greatest challenges be?
What products could you build?
What studies could you run?
How would these influence the organization?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if user would purchase again

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
   1. How many times did a user share through Facebook within a week? A month?
   2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if user would purchase again
5. Share results (and probably also go back to the drawing board)

PROBLEM: HOW TO DEFINE "MORE ITEMS TO CONSIDER" IN AMAZON?

10 Minutes: In a small group, define the flow an Amazon Data Scientist would work through to curate the "More items to consider" list for a particular user.

III. COMPUTER SETUP

LAB. DATA WORKFLOW

DISCUSSION