GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LESSON 1: DATA EXPLORATION

WELCOME!

CONTACT

Ed Podojil	epodojil@gmail.com
Dave Goodsmith	goodsmith@scaleanalytics.com
Joe Carli	joe.carli@gmail.com

Class M/W 6:30-9:30

Mondays: GA West (10 E. 21st Street, 4th Floor)

Wednesdays: GA Annex (22 E. 17th Street, 3rd Floor)

OFFICE HOURS

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

COURSE NOTES, WIKI, AND LAB/PROJECT SUBMISSIONS

https://github.com/datadave/GADS9-NYC-Spring2014

http://bit.ly/1fBhFlX

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

LAB:
III. COMPUTER SETUP
IV. DATA PRACTICE

I. WHAT IS DATA SCIENCE?

3 Minutes:

Use paper/your computer to write down as many different definitions/ associations you know about the term ‘data science’.

3 Minutes:

Use paper/your computer to write down as many different definitions/ associations you know about the term ‘data science’.

5 Minutes:

In a small group of four around you, introduce yourselves!
Joe and I will hand out a memo pad for each group.
Review all definitions and associations from your group and write each one down on a different piece of memo paper.

5 Minutes:

In a small group of four around you, introduce yourselves!
Joe and I will hand out a memo pad for each group.
Review all definitions and associations from your group and write each one down on a different piece of memo paper.

5 Minutes:

As a group, decide how to best "cluster" your definitions and associations. You can have as many different clusters as you want!
Make sure you have a label for each cluster.

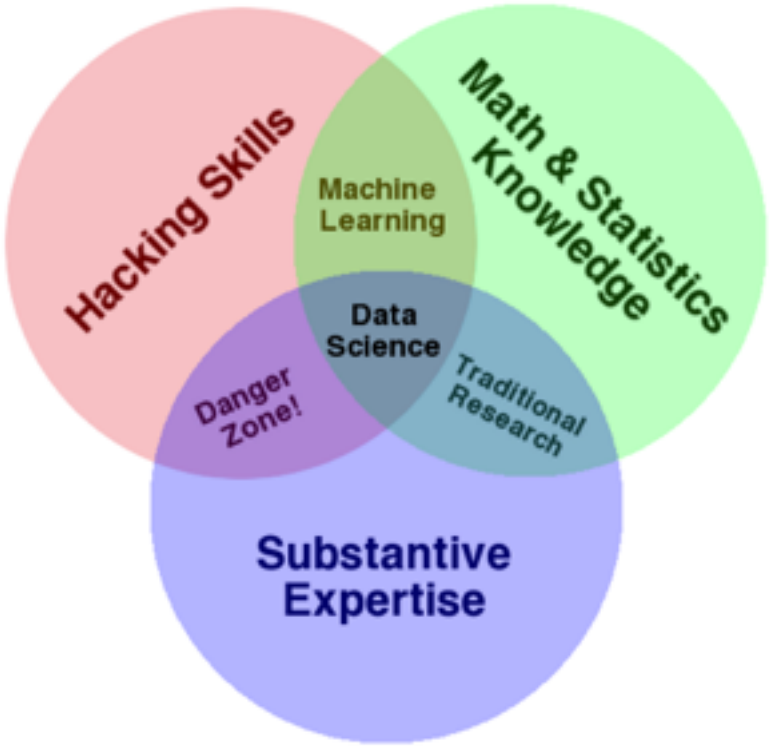
Cluster review!

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.

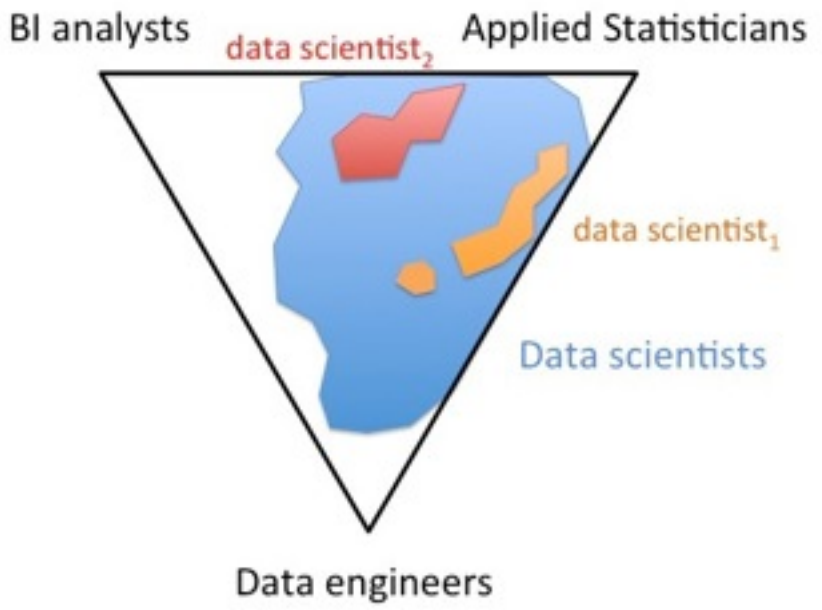
An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



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

THE QUALITIES OF A DATA SCIENTIST



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.

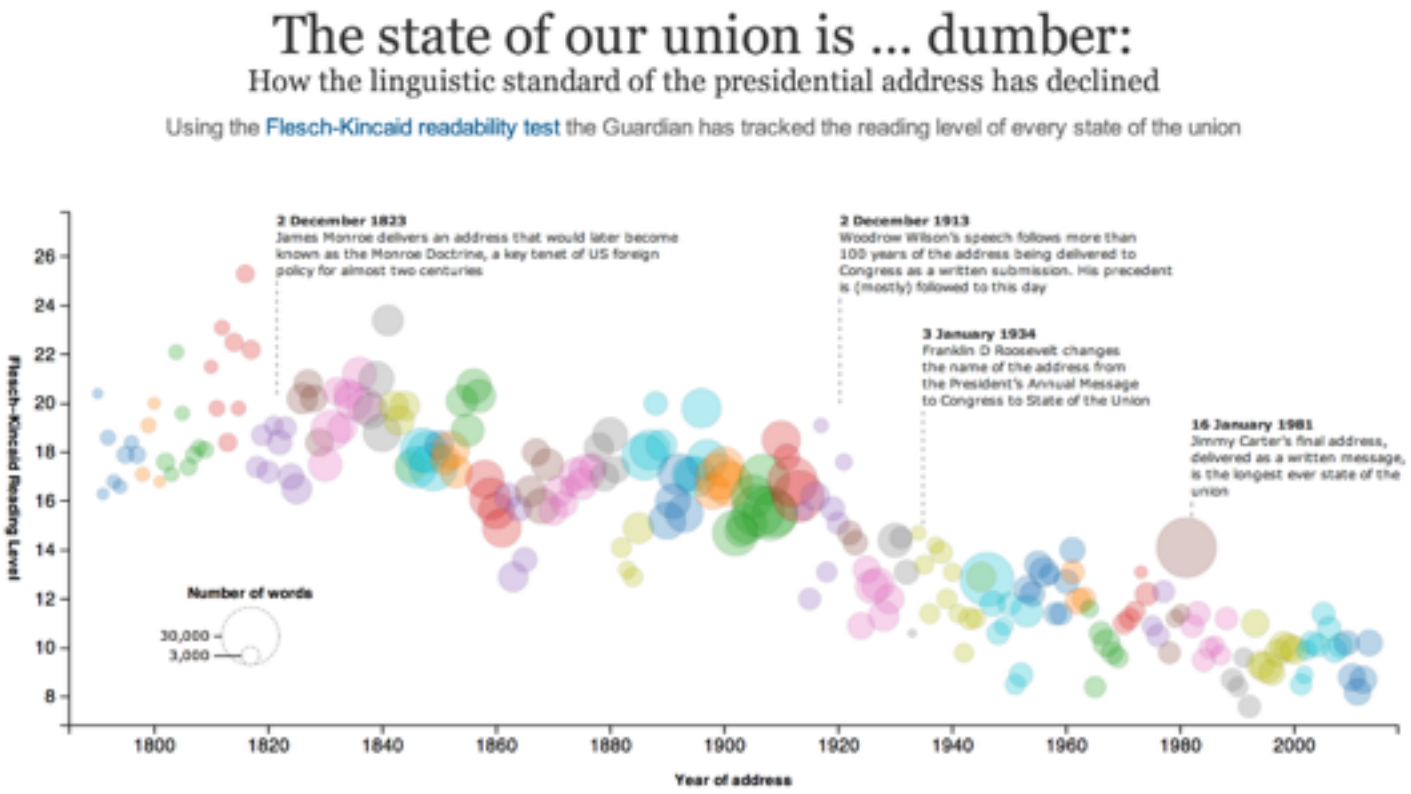
A set of tools and techniques used to extract useful information from data.

An interdisciplinary, problem-solving oriented subject.

The application of scientific techniques to practical problems.

A rapidly growing field.


NETFLIX
amazon
facebook
Google
LinkedIn
FiveThirtyEight
Nate Silver's Political Calculus
foursquare
Etsy
2012
BARACKOBAMA.COM
Microsoft


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
2 December 1823
James Monroe delivers an address that would later become known as the Monroe Doctrine, a key tenet of US foreign policy for almost two centuries
2 December 1913
Woodrow Wilson's speech follows more than 100 years of the address being delivered to Congress as a written submission. His precedent is (mostly) followed to this day
3 January 1934
Franklin D Roosevelt changes the name of the address from the President's Annual Message to Congress to State of the Union
16 January 1981
Jimmy Carter's final address, delivered as a written message, is the longest ever state of the union
Number of words
30,000
3,000
Flesch-Kincaid Reading Level
26
24
22
20
18
16
14
12
10
8
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Year of address

Music + Data:
http://bit.ly/echonest

WHAT IS A DATA SCIENTIST?

- “‘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

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

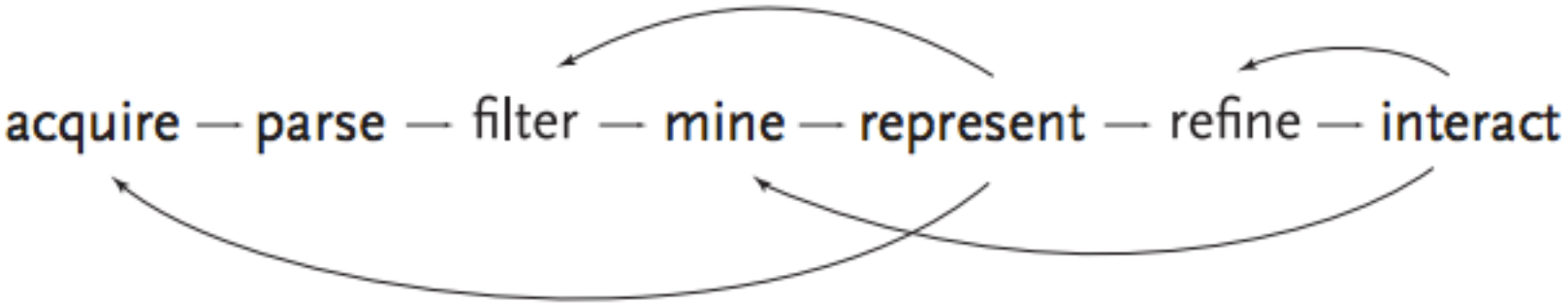
Jeff Hammerbacher: Chief Scientist, Cloudera

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

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ben Fry: Principal, Fathom

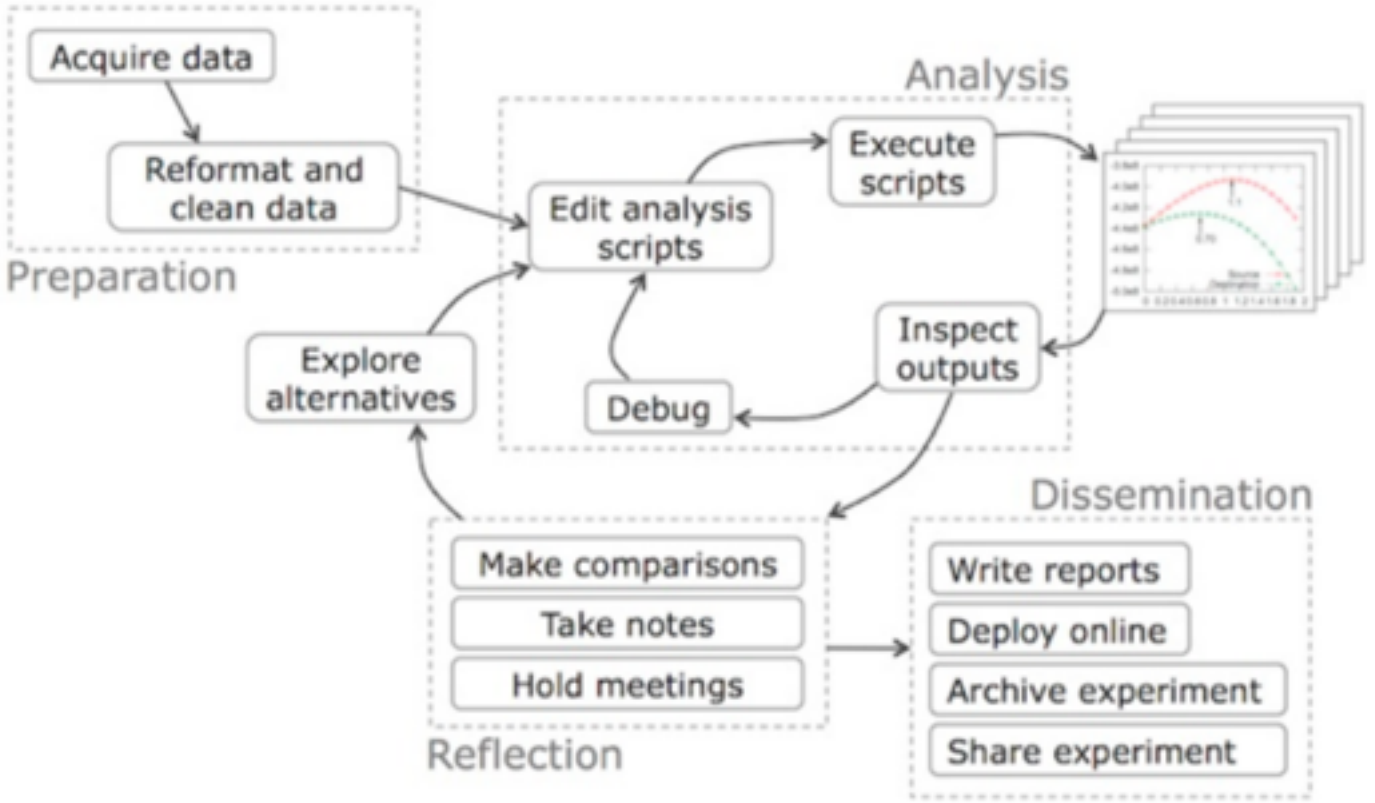
source: http://benfry.com/phd/dissertation-110323c.pdf



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

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 the same 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