DATA SCIENCE
CLASS 1: INTRODUCTION AND TOOLS

WELCOME!

Instructors: Aaron Schumacher, Kevin Markham, Sinan Ozdemir

gadsdc2.hackpad.org

Course Times: 6:30pm-9:30pm, Mondays and Wednesdays (1776)

Homework / Projects

0. META-INTRO
I. WHAT IS DATA SCIENCE?
II. THE DATA MINING WORKFLOW

LAB:
III. WORKING AT THE UNIX COMMAND LINE

0. META-INTRO

LEARNING IS FOR EVERYONE

LEARNING IS A CONSEQUENCE OF THINKING

WE ARE ALL STUDENTS

WE ARE ALL TEACHERS

COMMUNICATE EARLY AND OFTEN

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-solving oriented subject.

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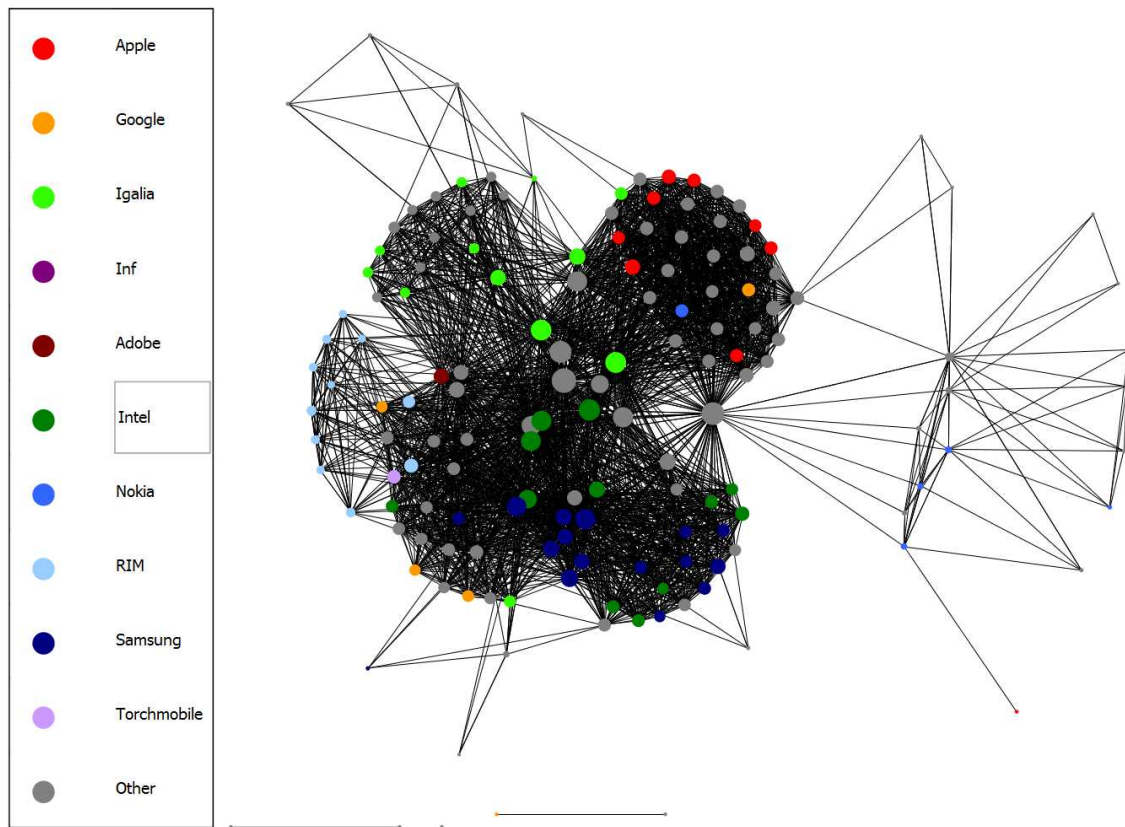
- A set of tools and techniques used to extract useful information from data.

- An interdisciplinary, problem-solving oriented subject.

- The application of scientific techniques to practical problems.

- A rapidly growing field.

- Recommending products on amazon.com

- Identifying fraudulent credit card transactions

- Recommending new musical artists

- Prioritize emergency calls in Seattle

- Many more!

- *Collaboration in the open-source arena: The WebKit case*

Legend:
- Apple (red)
- Google (orange)
- Igalia (green)
- Inf (purple)
- Adobe (dark red)
- Intel (dark green)
- Nokia (blue)
- RIM (light blue)
- Samsung (navy)
- Torchmobile (light purple)
- Other (gray)

- Application Presentations!

- https://gadsdc1.hackpad.com/

- Statistical and machine learning knowledge

- Engineering experience

- Academic curiosity

- Product sense

- Storytelling

- Cleverness

Michael E. Driscoll
@medriscoll

Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch
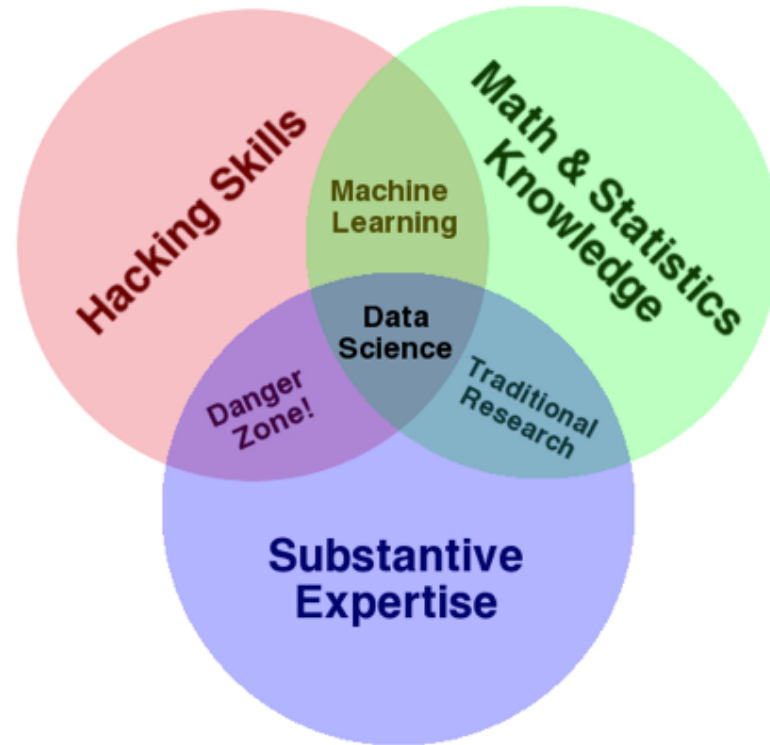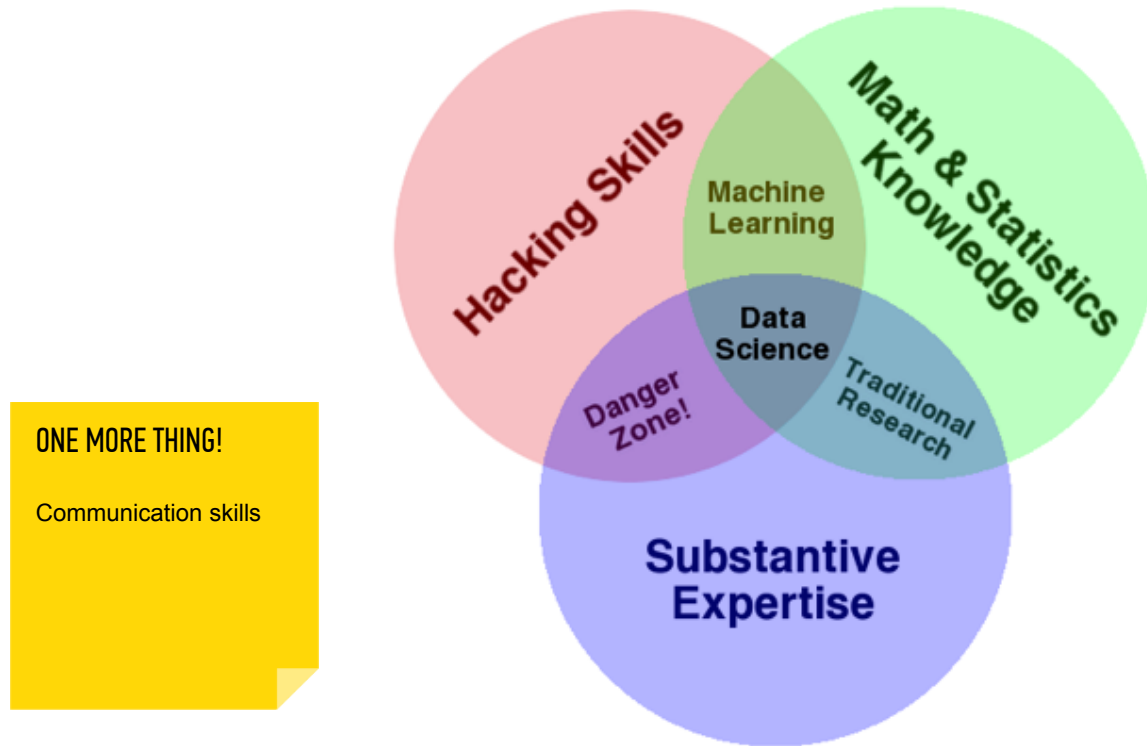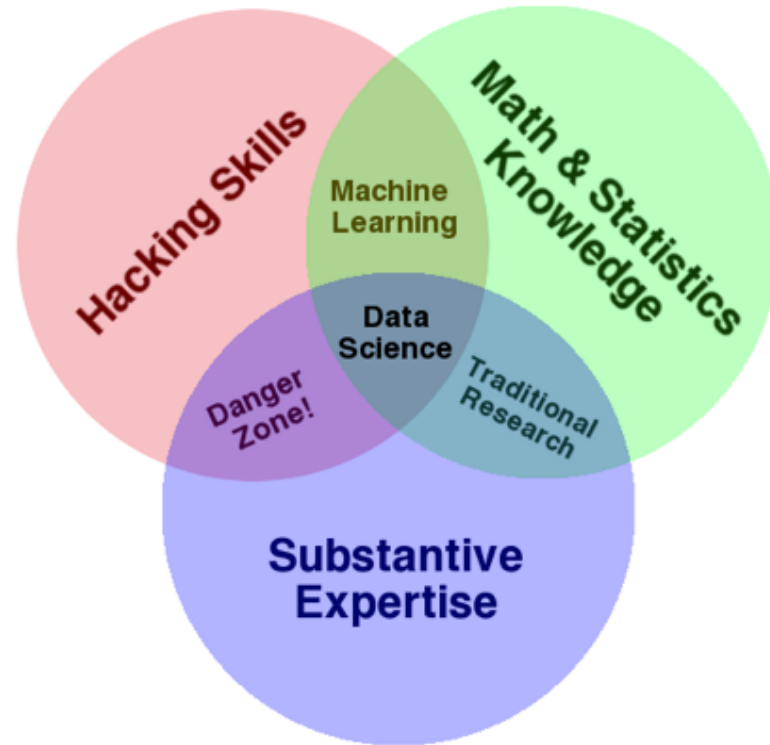
Reply   Retweet   Favorite   More   Pocket

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

ONE MORE THING!

Communication skills

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

ONE MORE THING!

Communication skills

ANOTHER THING!

Answer a question!

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

What's big data?

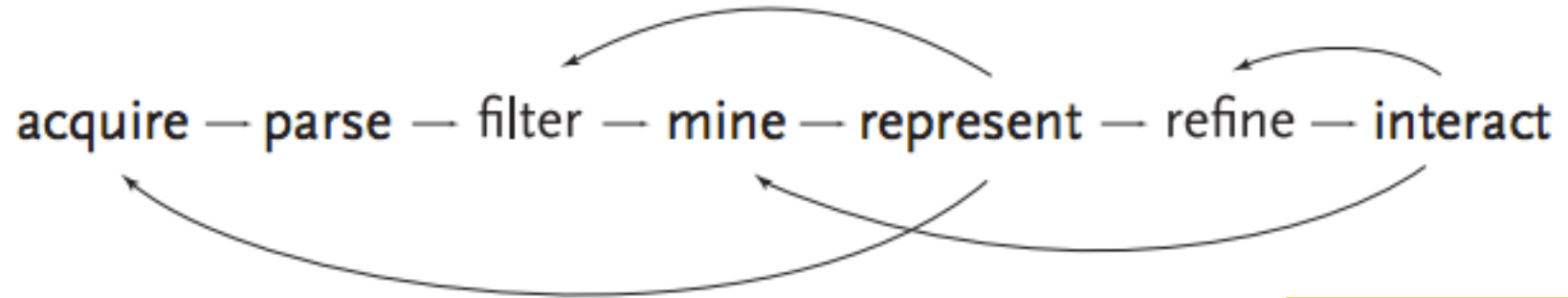The practical viewpoint:

1. $O(n^2)$ algorithm feasible: small data
2. Fits on one machine: medium data
3. Doesn't fit on one machine: big data

source: http://people.cs.umass.edu/~mcgregor/stocworkshop/langford.pdf
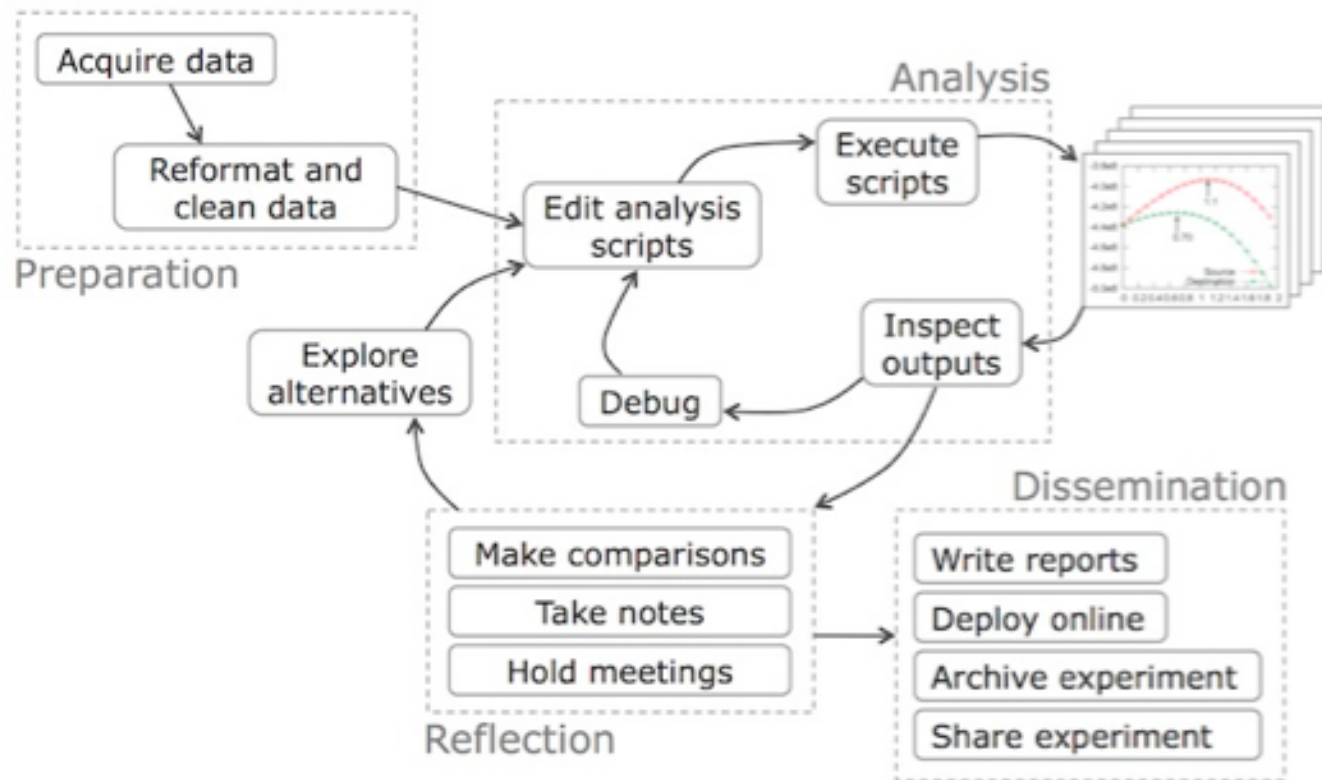
II. THE DATA SCIENCE WORKFLOW

from Jeff Hammerbacher:

- 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

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

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

ALSO:

scale

source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge

III. WORKING AT THE UNIX COMMAND LINE

KEY OBJECTIVES

- Navigate the filesystem

- Create, move, copy, and delete files & directories

- View & search files

- Edit & interact with files

- Combine steps

- Learn more

TOOLS

- ls, cd

- cat, touch, mv, cp, mkdir, rm, rmdir

- head, tail, less, cat, grep

- vim, tr, sort, uniq, wc

- pipe (|)

- man, apropos

NOTE

Being comfortable at the command line makes your life much easier!

GIT

LINE-ORIENTED PIPELINES

INTRO TO DATA SCIENCE