

DATA SCIENCE

CLASS 1: INTRODUCTION AND TOOLS

INTRO TO DATA SCIENCE

WELCOME!

Instructors: Aaron Schumacher, Kevin Markham, Sinan Ozdemir

gadsdc2.hackpad.com

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

▸ **META-INTRO**

**COMMUNICATE
EARLY AND
OFTEN**

INTRO TO DATA SCIENCE

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.

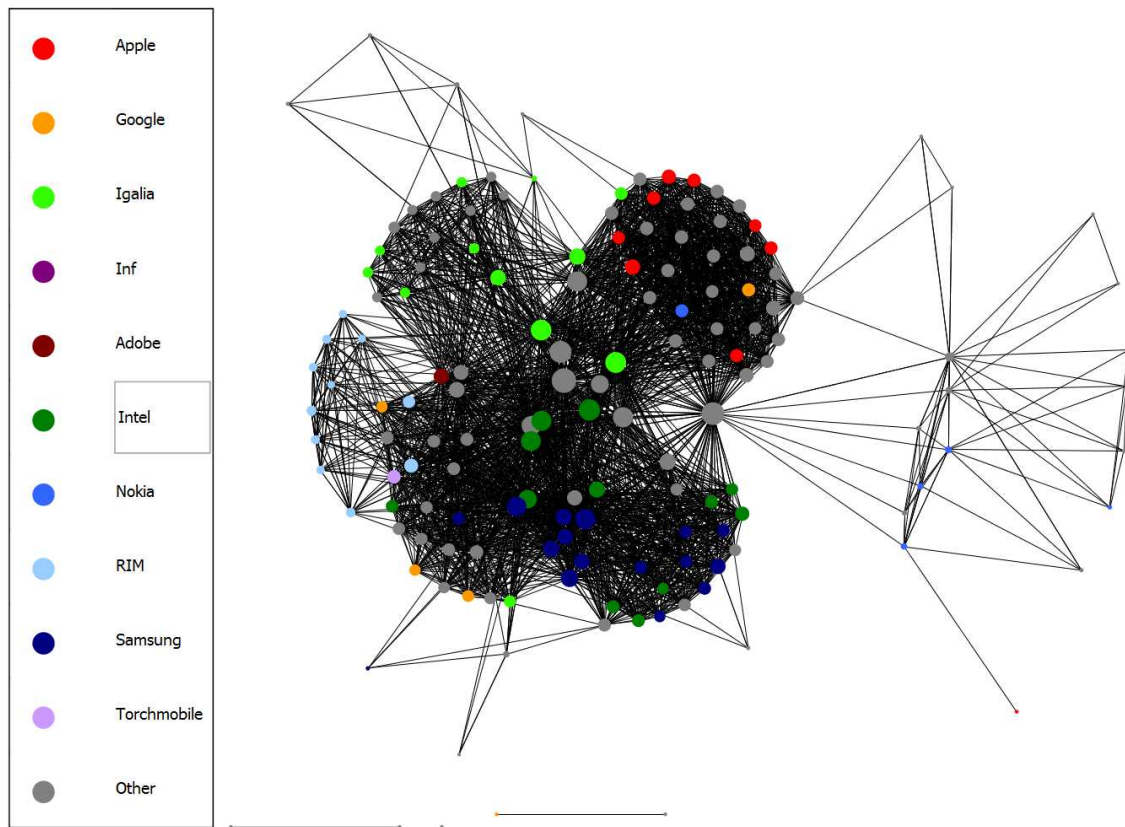
- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.



- 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*

WHO USES DATA SCIENCE?

17



- Application Presentations!
- <https://gadsdc2.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



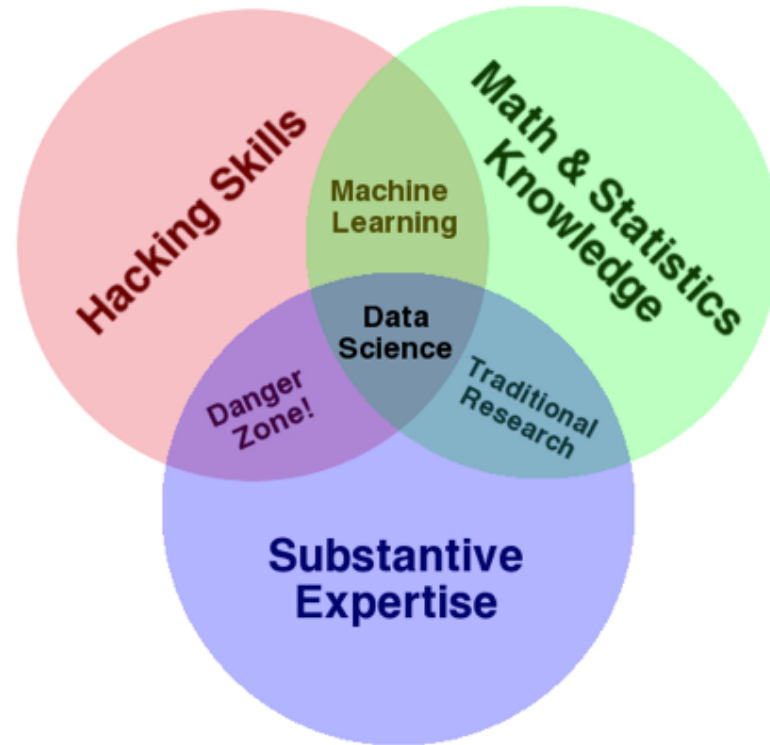
Favorite

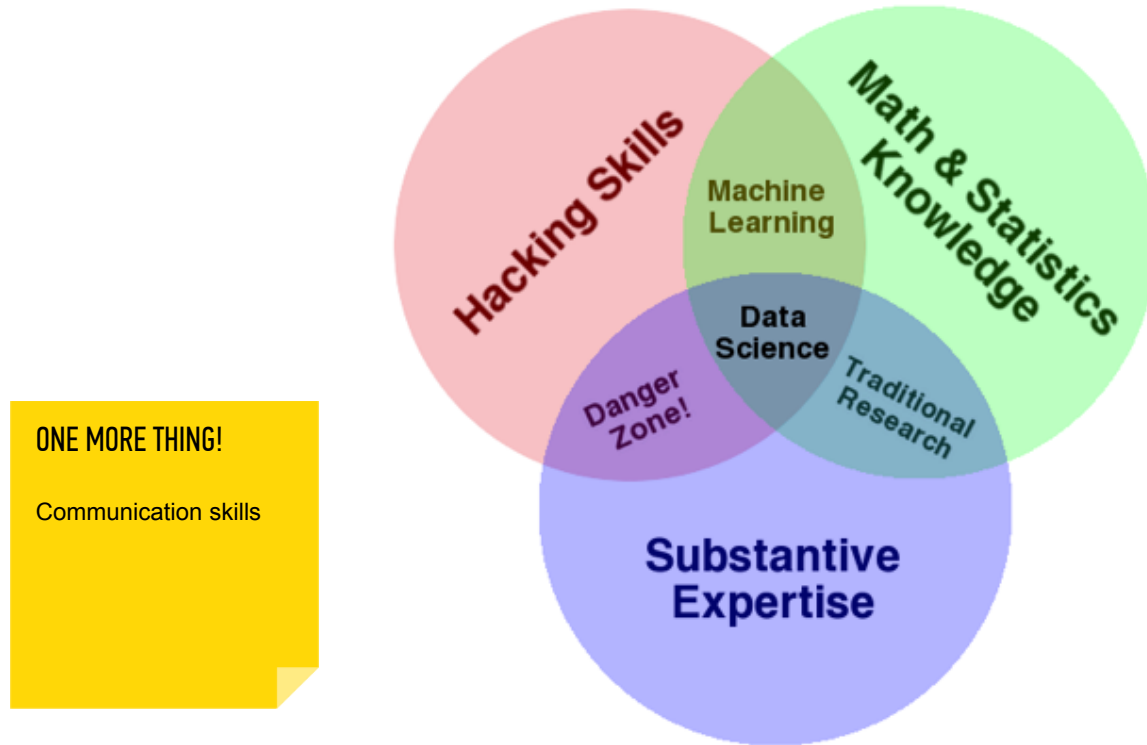


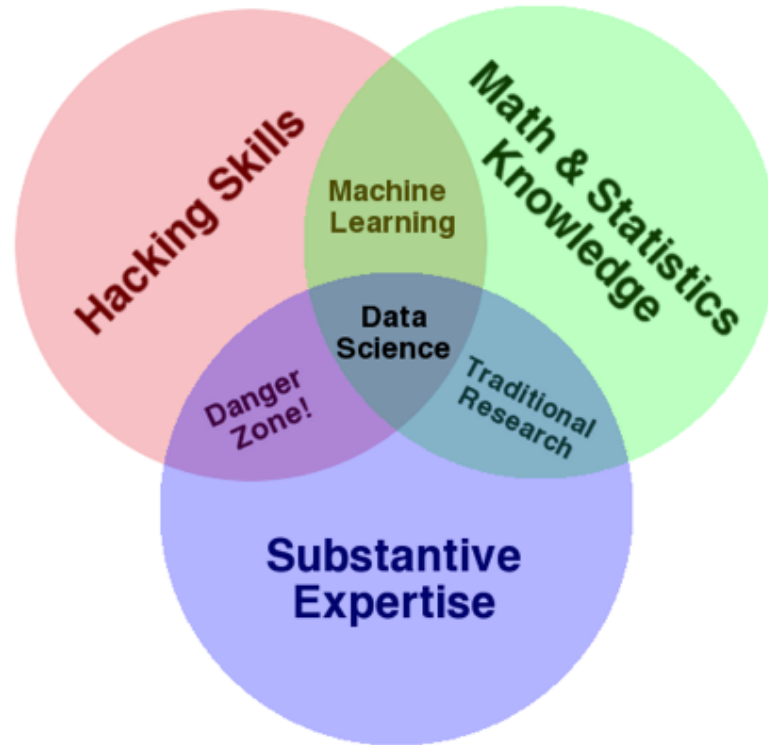
More



Pocket







ONE MORE THING!

Communication skills

ANOTHER THING!

Answer a question!

What's big data?

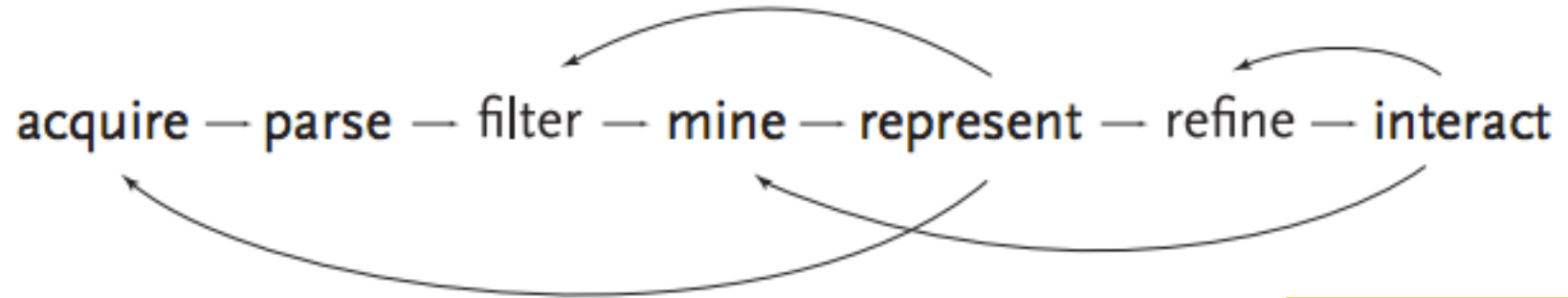
The practical viewpoint:

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

II. THE DATA SCIENCE WORKFLOW

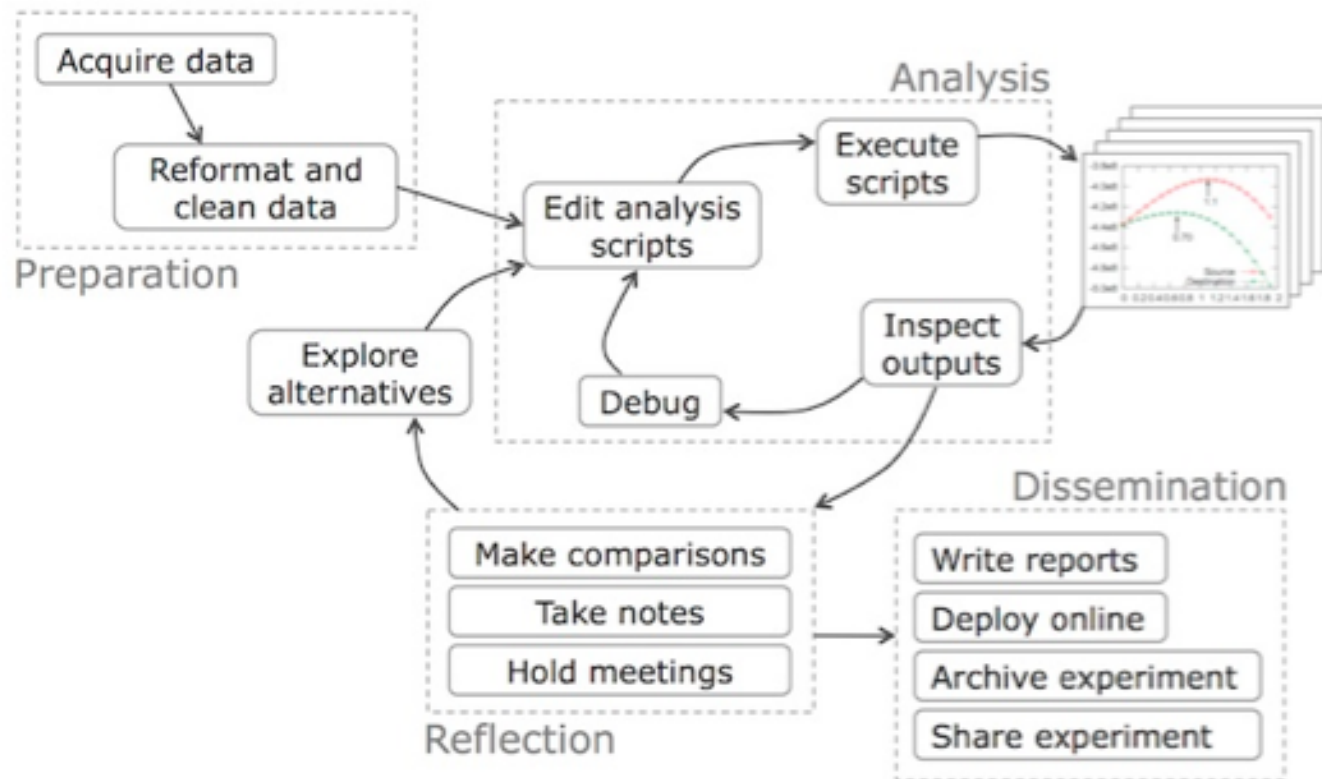
from Jeff Hammerbacher:

- 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



ALSO:

scale



III. WORKING AT THE UNIX COMMAND LINE

EXERCISE – WORKING AT THE UNIX COMMAND LINE

30

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!

▸ WORKING AT THE UNIX COMMAND LINE

GIT

LINE-ORIENTED PIPELINES

INTRO TO DATA SCIENCE
