Introduction

What we're doing here

Why are we here?

❖ Understand what data scientists do

❖ Understand how to work with data scientists

❖ Get some cool tools and skills

❖ Build visualizations - for decisions, M&E, funding

❖ Stop hand-waving and start making stuff

This Weekend

* Friday: introduction, tools, Python, data

* Saturday: more data, science, communicating

* Sunday: D3, big data, continuing your journey

Focussing on Concepts

❖ Lots of data science applications and tools, very few core concepts:

   ❖ Data collection

   ❖ Data cleaning

   ❖ Visualisation

   ❖ etc

Tools change: want you to focus on the concepts

And basics of commonly-used tools

- ❖ Python, R, D3
    - ❖ Very flexible languages
    - ❖ Lots of helpful libraries
    - ❖ Huge communities

PS Ignore the holy wars - just use what works for you

Who's helping?

- ❖ Prof:
  - ❖ Sara-Jayne Terp (bodacea on github)
- ❖ Teaching Assistants:
  - ❖ Nate Brennand
  - ❖ Henrique Gubert
  - ❖ Lin He

Some of you have to leave for an hour or two

- ❖ To go to church, lectures, etc (nb "hangover" doesn't count)

- ❖ That's okay… these things happen

- ❖ All slides are online, with notes

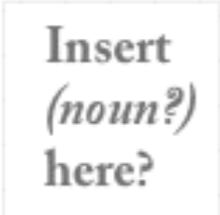- ❖ And we have "activity sessions", designed to help you get further

What is Data Science?

* "A data scientist… excels at analyzing data, particularly large amounts of data, to help a business gain a competitive edge."

* "The analysis of data using the scientific method"

* "A data scientist is an individual, organization or application that performs statistical analysis, data mining and retrieval processes on a large amount of data to identify trends, figures and other relevant information."

# The Scientific Method

❖ Ask a question

❖ Do background research

❖ Construct a hypothesis

❖ Test your hypothesis by doing an experiment

❖ Analyse your data and draw a conclusion

❖ Communicate your results

O'Neil and Schutt "Doing Data Science"

Understanding through Data

Competition Name	Reward	Teams	Deadline
limited **15.071x - The Analytics Edge Competition (Spring 2015)** Test your analytics skills by predicting which New York Times blog articles will be the most popular.	Private	528	18 days
Forest Cover Type Prediction Use cartographic variables to classify forest categories	Knowledge	1572	25 days
Billion Word Imputation Find and impute missing words in the billion word corpus	Knowledge	78	15 days
Bike Sharing Demand Forecast use of a city bikeshare system	Knowledge	2687	43 days
Random Acts of Pizza Predicting altruism through free pizza	Knowledge	384	46 days

What's a Data Scientist?
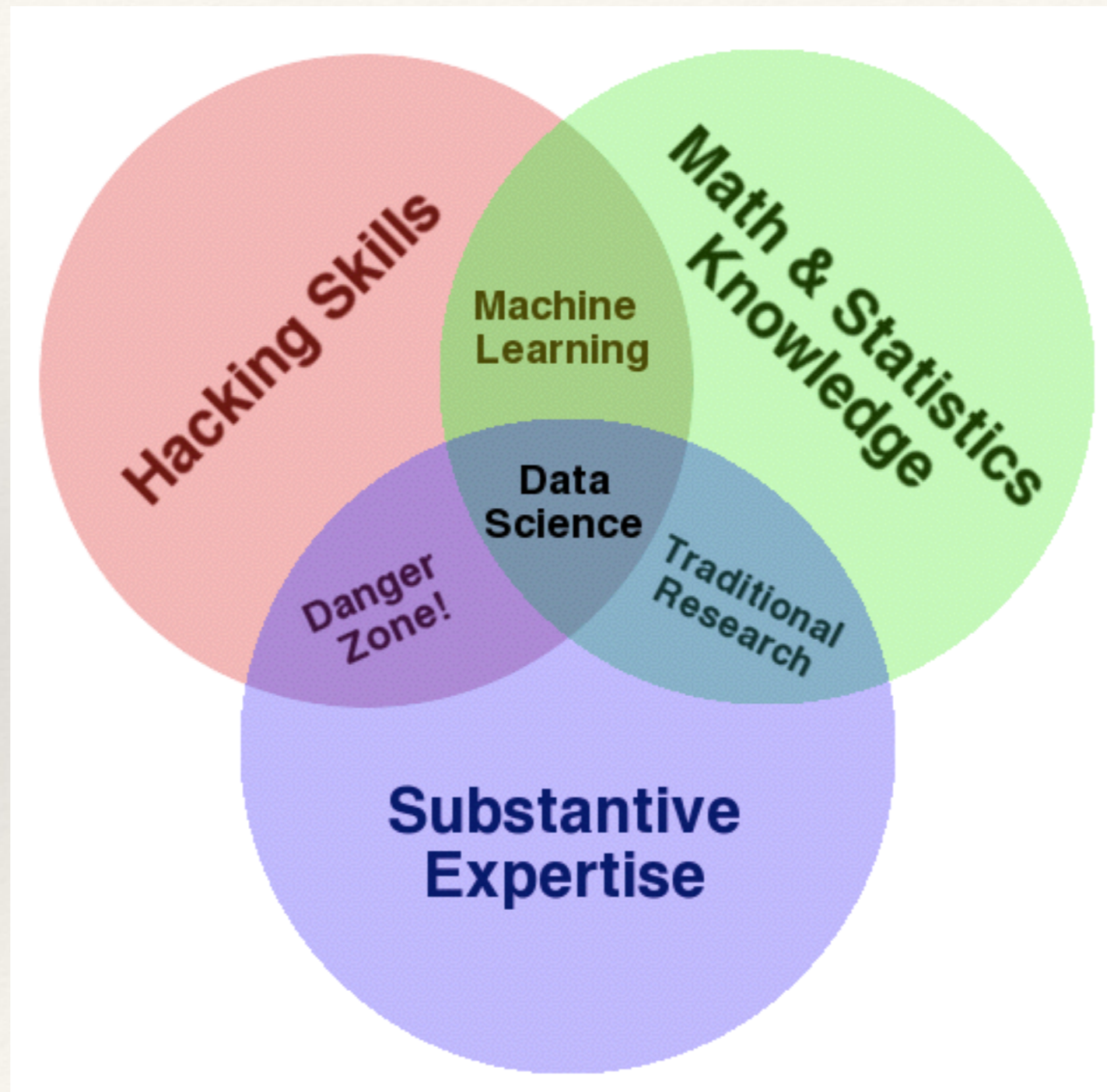
How do you become a data scientist?

Practice

❖ Here are some online places to do that…

  ❖ Kaggle - online datascience competitions

  ❖ Driven Data - social good datascience competitions

  ❖ Innocentive - some datascience challenges

  ❖ CrowdAnalytix - business datascience competitions

  ❖ TunedIt - scientific/industrial datascience challenges

Should you become a data scientist?

- Not necessarily. There are lots of data science students desperate for good problems to work on.

- You might want to become someone who can work **with** data scientists

- Which means learning how to specify data problems well

Some questions for you

❖ What do you want to get out of this weekend?

❖ What are your favorite visualisations?

❖ What's your favorite dataset?

❖ What questions do you want to answer with data?

And some answers

- Your course credits are:

  - 20% Class participation

  - 30% Coding exercises in class

  - 50% Final project: specify and work on a data science problem of your choosing, either individually or in groups.

- Final project hand-in is Monday 27th April.