[1]:

```python
from matplotlib.pyplot import *
%matplotlib inline

from IPython.display import set_matplotlib_formats
set_matplotlib_formats('png', 'pdf')
```

DS1 Lecture 05

Jim Bagrow

Last week:

1. Social ratings with uncertainty

Today's plan:

1. Finish social ratings "remaining limitation" [board]
2. Data science "pipeline"
   - Overview, the big picture
3. Typology of data

Data Science

We touched on this way back in LEC01, but let's return now that we have started to dig into small bits of material.

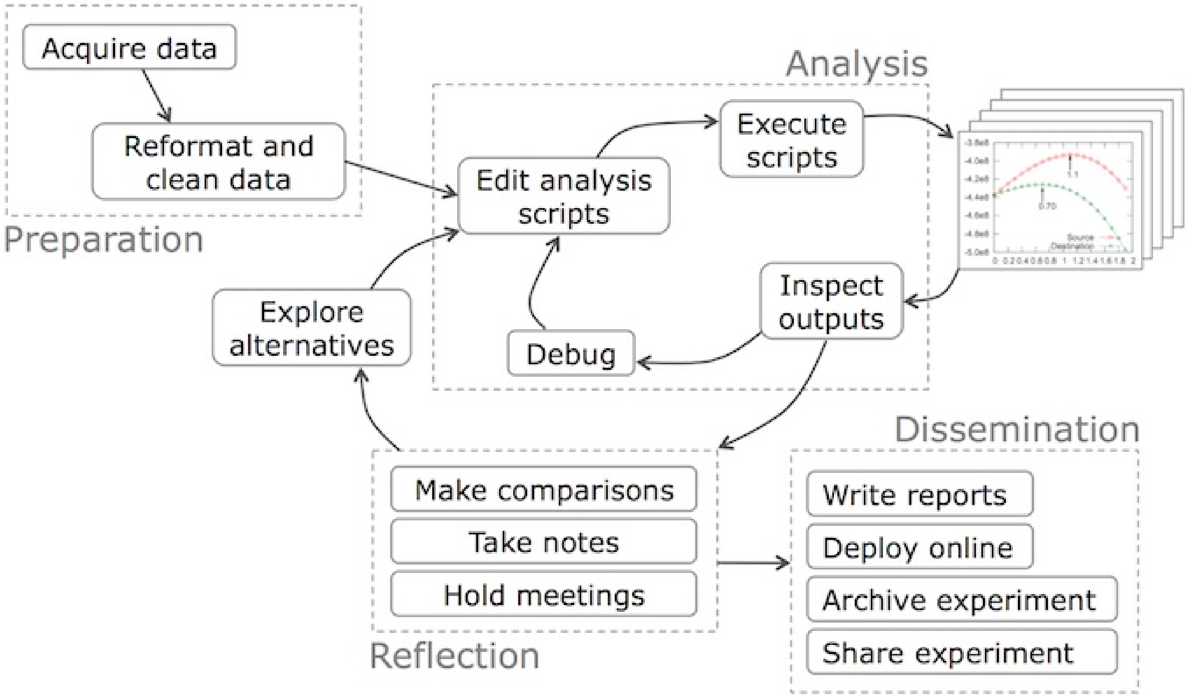
What is involved in data science? What does a data scientist do?

Here's one of the better illustrations I've seen (courtesy ACM):

[7]:

```python
from IPython.display import Image
Image(filename='rp-overview.jpg')
```

[7]:



A typology of data

1. Stevens, *Science*, 1946
2. `http://en.wikipedia.org/wiki/Level_of_measurement`

Data = measurements = types of variables

Stevens claimed that all measurements in science fall into four categories that he named:

1. Nominal,
2. Ordinal,
3. Interval,
4. Ratio.

This is somewhat different than data types (integer, float, string) in a computer program. The notion is: what do the data *mean*?

The type of data we have dictates what calculations and statistics we can apply. Let's dig in to learn what these are.

- Example statistic: **central tendency** (we'll discuss these for each level of measurement)

Nominal data

Nominal data are qualitative. It can be used for classification but can't be ordered. Examples:

- Gender
- Race
- Genre
- Species
- Language

- Names of friends

Most mathematical operations are **not meaningful** for ordinal data. It doesn't make much sense to compute english + spanish, or science fiction * comedy, etc.

- Equality makes sense: English $\neq$ Spanish.

Central Tendency

- Without addition we can't generally say the *average* (precisely, arithmetic mean) is meaningful. Although we can ask what is the most common item in the data (mode.

Ordinal Data

Similar to nominal data but with a **scale** or **ordering**. Examples:

- Healthy, Sick
- Completely Agree, Mostly Agree, Mostly Disagree, Completely Disagree

An ordering lets us measure not just equality, but that, e.g., Healthy $>$ Sick.

Central Tendency

With an ordering we can meaningful compute the "middle" of the data. This is the Median.

- The **median** is found by ordering all of the data from smallest to largest and picking the *middle* value. This is also known as the **50th percentile**.

Interval Data

Quantitative, numerical data, often continuous. Interval data capture a meaningful notion of *difference*. These data are typically numbers on a scale with a relative/arbitrary zero point. Examples:

- Dates: 200 BC to 2015 AD.
- Temperatures: 10C, 20C
- Latitude

It's meaningful to say 2015 AD - 200 BC = 2215 years. 20C is 10 degrees warmer than 10C, but it's not **twice** as warm.

- Addition and subtraction are meaningful, as are equality and ordering (>, <). Multiplication and division are not.

Central Tendency

With addition being meaningful for these data, we can now compute the typical (arithmetic) mean. The median and mode are also meaningful here.

Ratio Data

Ratio data is also quantitative and numeric, but it's defined relative to some **base quantity** or unit. That's what the "ratio" means. Examples:

- Time measured in seconds or years
- Length measured in meters, inches, etc.
- Area in square meters
- Electric charge
- Mass

Ratio data is often what we think of when we mean “numerical data”. Notice how it’s subtly different from interval data:

- **dates** are interval data, **durations** are ratio data

Addition/Subtraction, Multiplication/Division, Equality/Inequality and Ordering are all meaningful for ratio data.

Central Tendency

The mean, median and mode, the most common measures of tendency, are all meaningful here.

- In addition to the typical arithmetic mean, the geometric and harmonic means are sometimes meaningful here.

Some aspects of this grouping are controversial. Is there a difference between categorical and nominal data?

Take away: Any of these data can be and are represented on the computer with numbers, but it’s important to understand what quantities they capture so you can compute meaningful information from those numbers. Data have units!

Takeaways

The big DS picture

- We are tackling pieces of the Data Science **pipeline**, here in class and in the readings and homeworks. The big picture emerges.
- Data have meaning. Often that meaning is associated with the **units** of the data (Stevens’ levels of measurement). Use that meaning to choose appropriate (meaningful) statistics.