DS Visualisation and Analysis

Abbey Waldron

This Week - Making the right Plot!

1. Show last week's work to the class
2. Introduction to making the right plot
3. Group Challenge - guess the question
4. Individual work - 2014 ebola data

Random Student Generator

Problem Set From Last Week: Irises

1. Use help(iris) to understand what variables there are - make sure you know what they all mean.
2. Make a histogram (hist) with 20 bins of petal width for the Iris Setosa.
3. Make a scatterplot (tip: try ggplot2 ggplot) of sepal length vs petal length. Show each of the three species of Iris on the same plot with a coloured legend to separate them.
4. Make a scatterplot of sepal length vs sepal width for all Irises whose petal width is greater than 1.5.
5. Make one more plot that shows something interesting about the inter-species differences of the Irises.

What does it mean to make the right plot?

What does it mean to make the right plot?

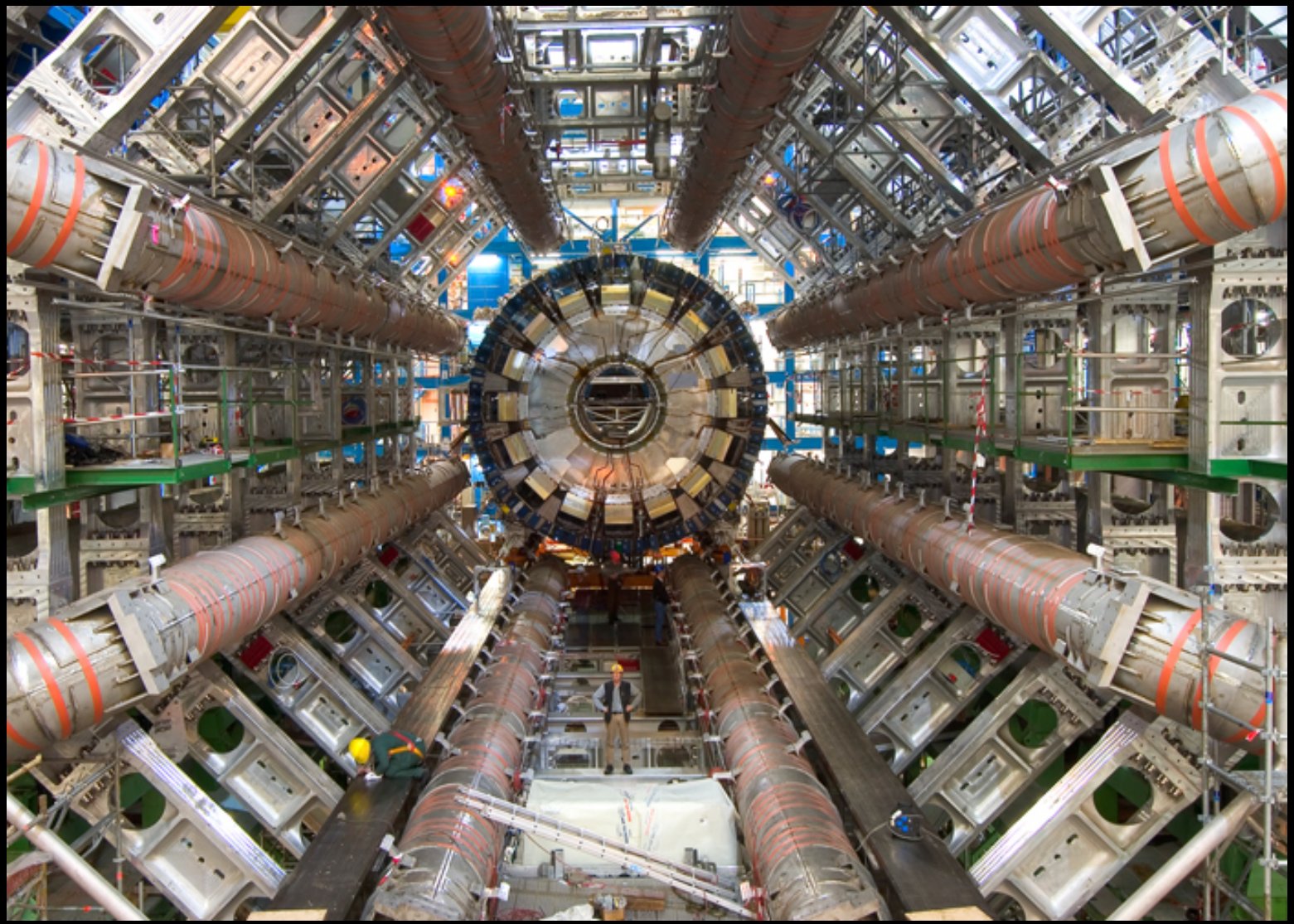
- Ask a good question
- Answer the question in one plot

Practice and THINK

Calibration (normalisation)

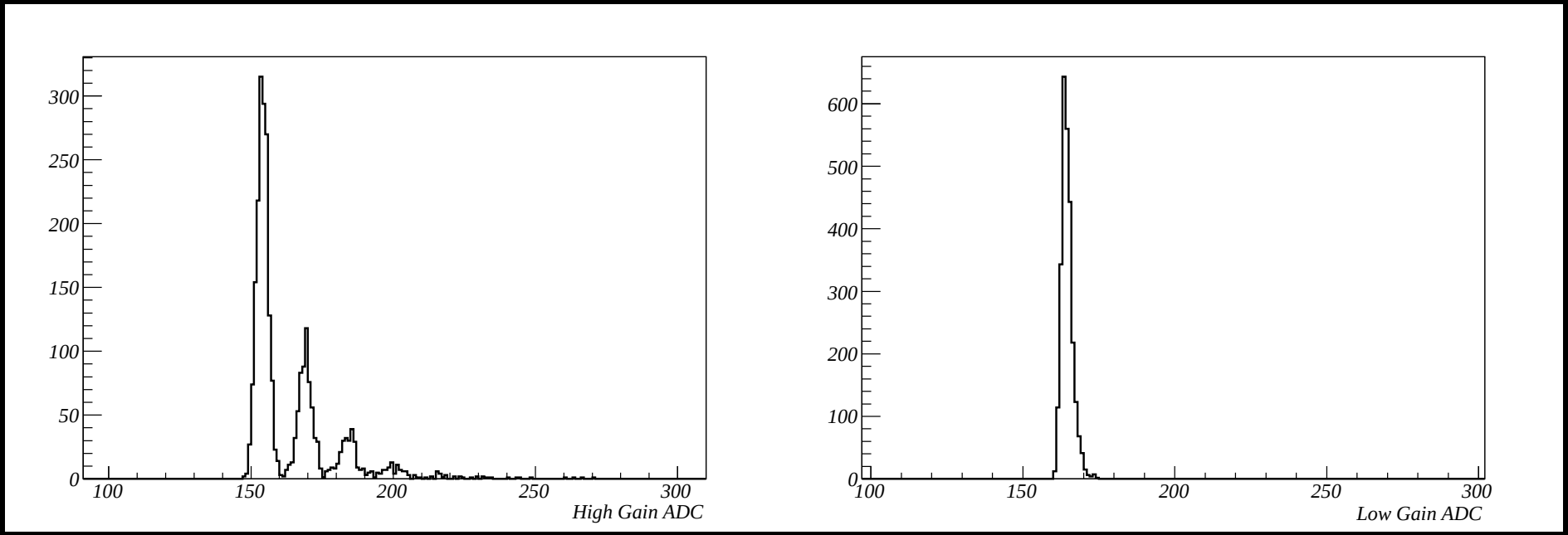
Same input $\rightarrow$ same output

A little story

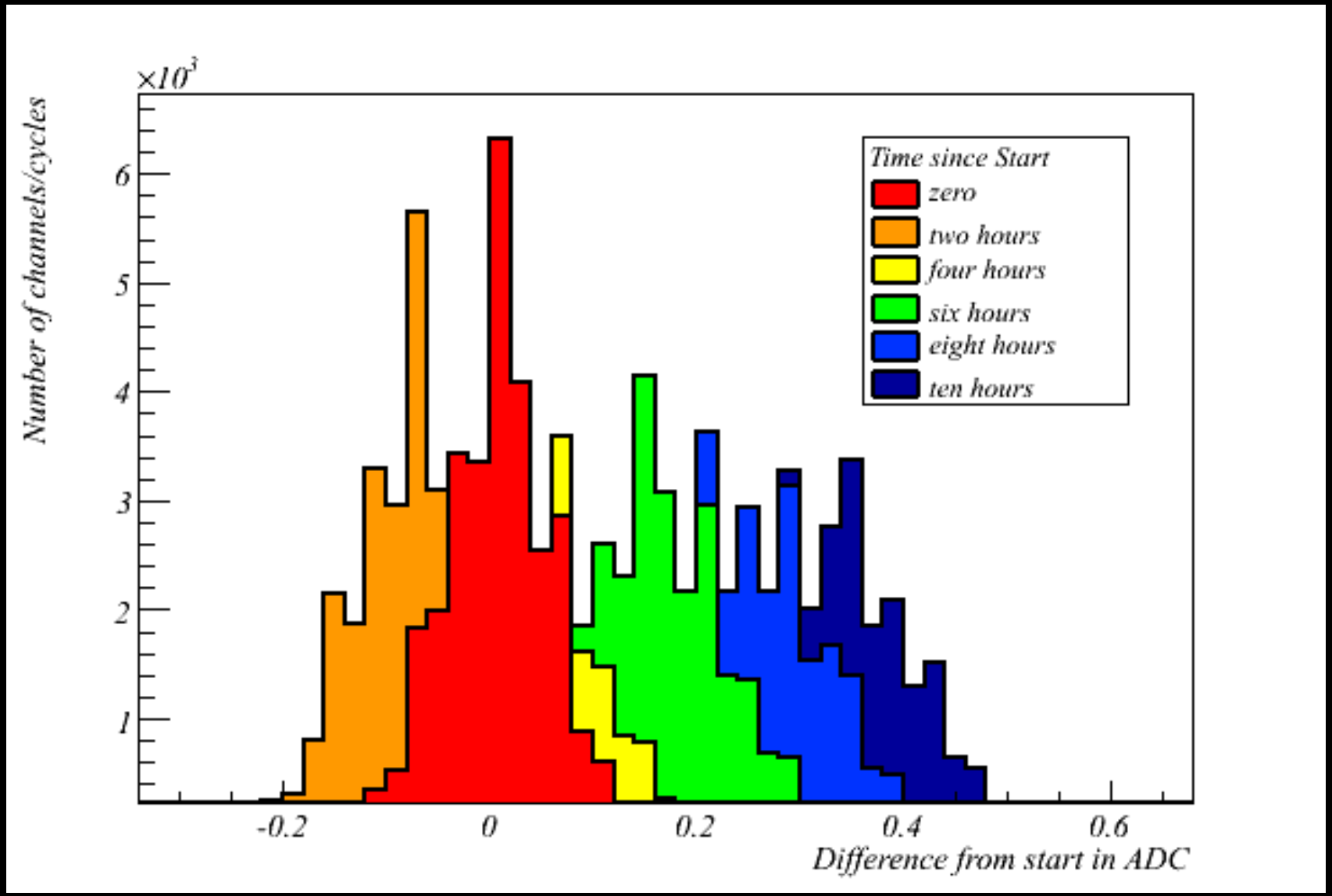


demaco.nl

Noise


300
250
200
150
100
50
0
100
150
200
250
300
High Gain ADC
600
500
400
300
200
100
0
100
150
200
250
300
Low Gain ADC

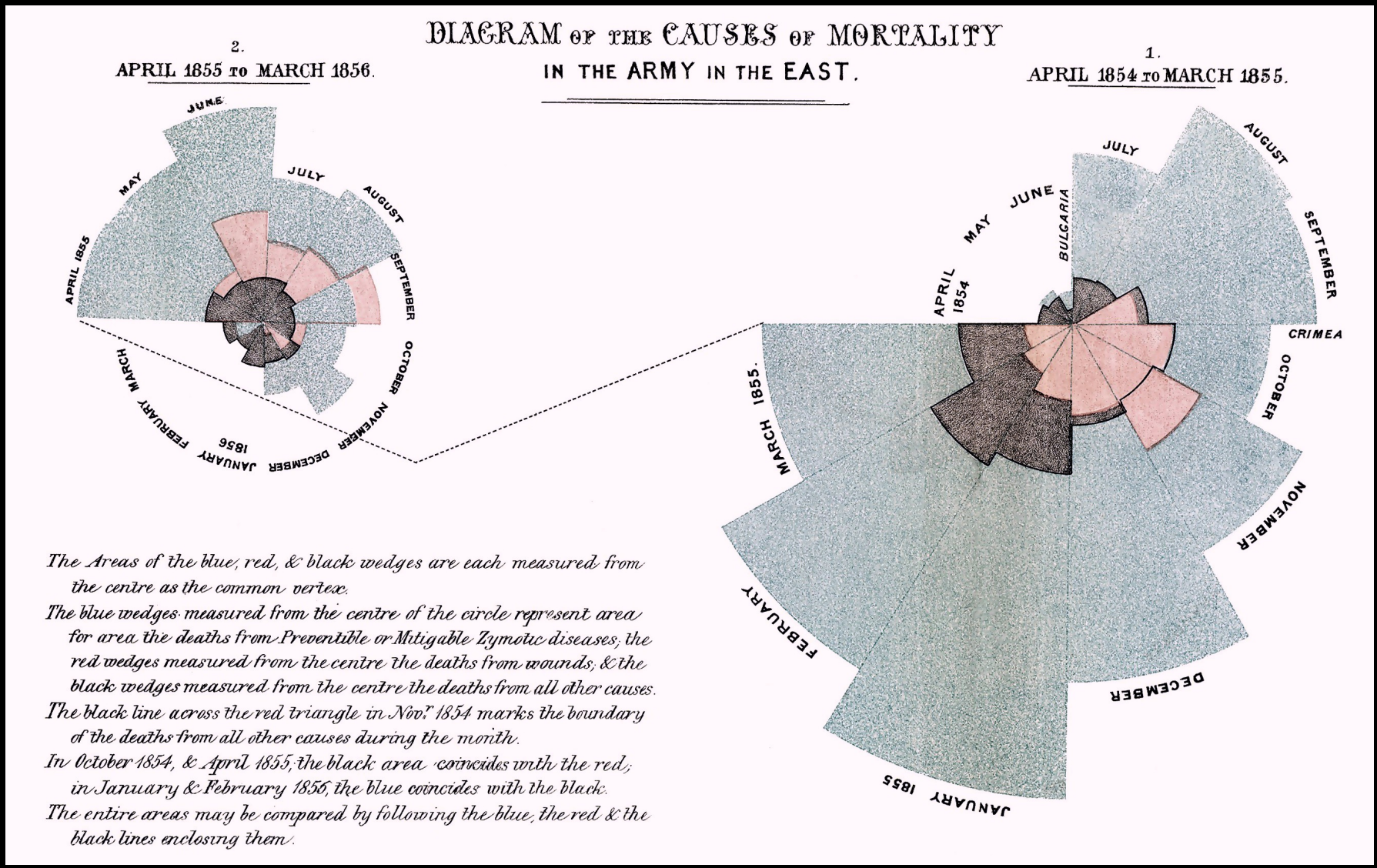
What Worked


×10^3
Number of channels/cycles
6
5
4
3
2
1
-0.2
0
0.2
0.4
0.6
Difference from start in ADC
Time since Start
zero
two hours
four hours
six hours
eight hours
ten hours

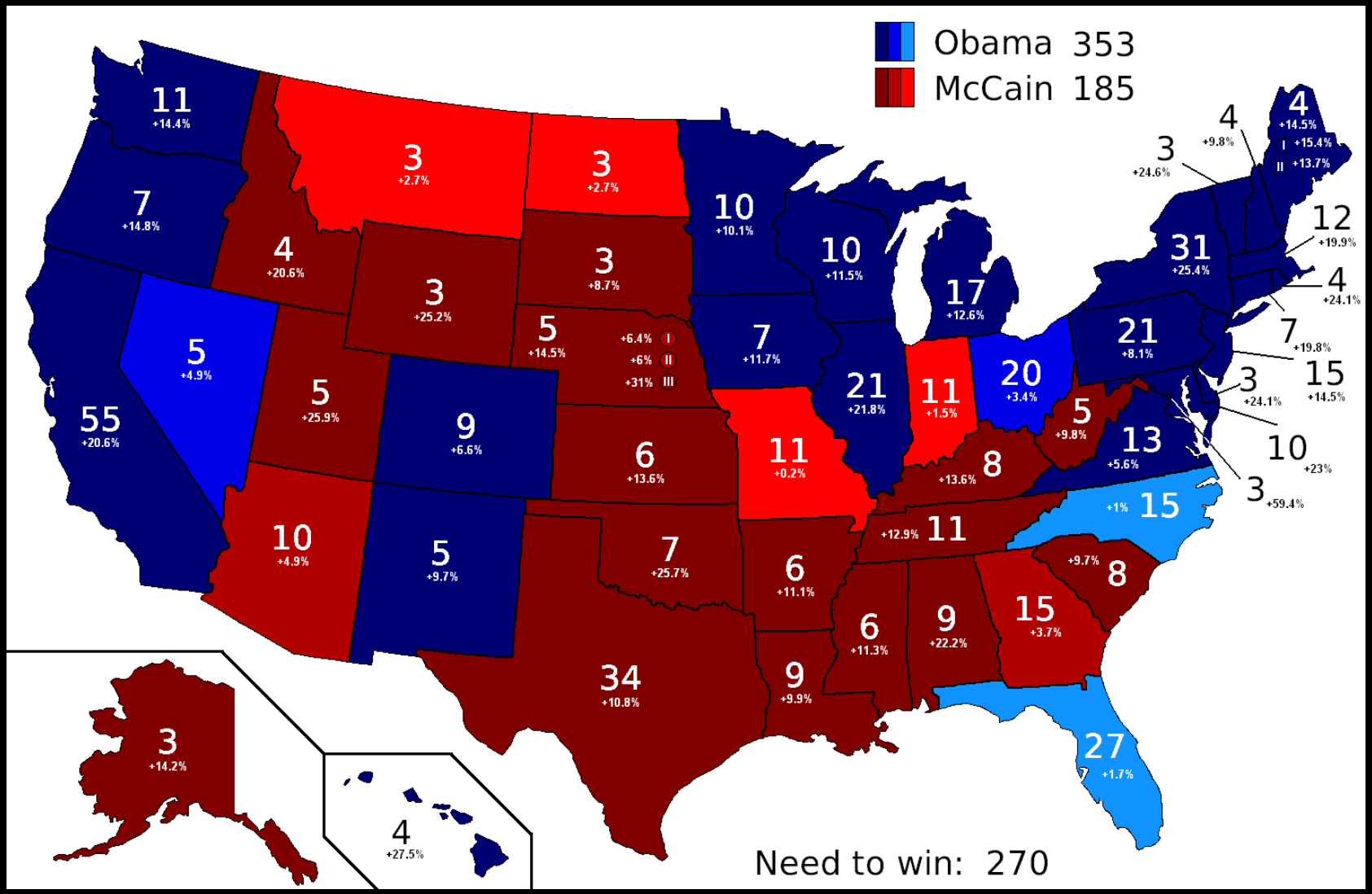
Group Challenge

Here are some plots: what is the question they answer?

Florence Nightingale

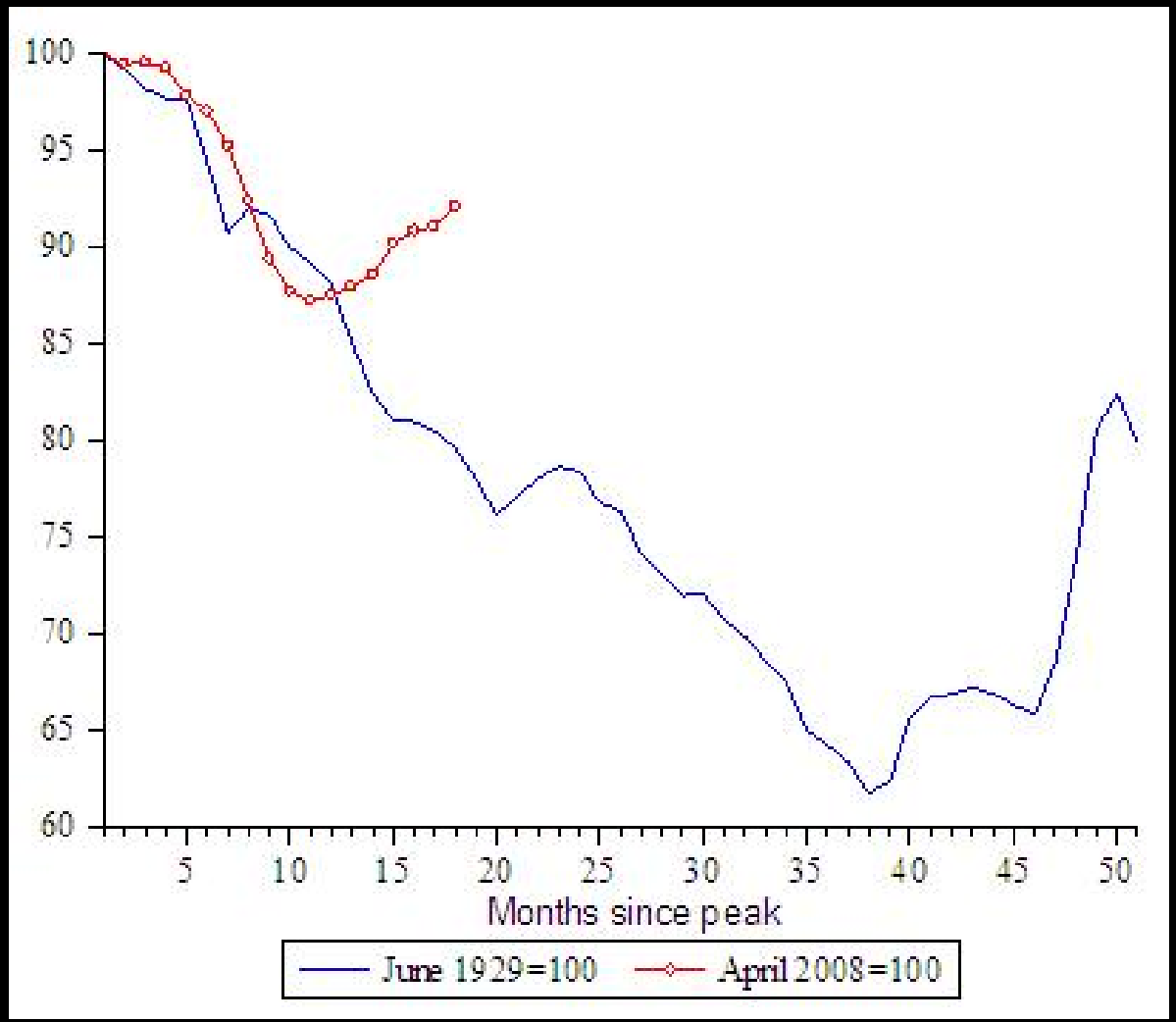


US Poll Results pre 2008 election



Wikipedia commons

Guess



voxeu.org

This week's problems

We will work with data from the West Africa ebola outbreak of 2013 to present.

Still ongoing, so far 28 000 cases are suspected to have occurred.

Notes on the data

Data is from the WHO, cite it! The "total" number is only given when data of confirmed, probable and suspected cases or deaths is not known.

Think carefully about which numbers you want to use for the different questions:

- Confirmed: laboratory tests confirm ebola
- Probable: health workers see symptoms consistent with ebola
- Suspected: other cases that may be ebola

R Code Hints

Some things you may need:

- `read.csv()`
- `lapply`, `tapply` etc.

Check out the swirl stats tutorial if you don't know the last ones.

Backup Slides