DS Visualisation and Analysis

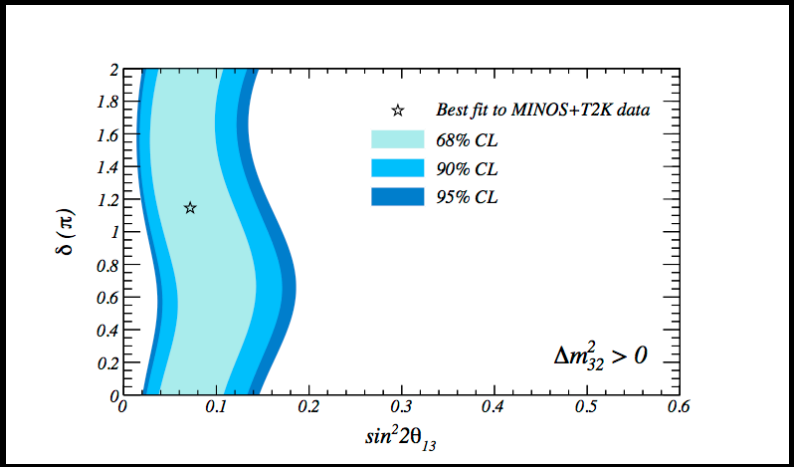
Abbey Waldron

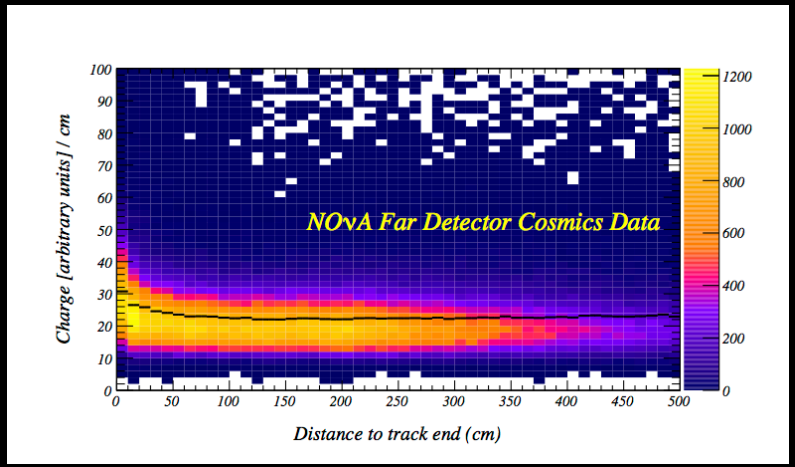
Hello

Experimental Particle Physics

- Reading data out from huge particle detectors
- Real time monitoring of data taking
- Simulating large amounts of data (Monte Carlo)
- Analyzing and visualising results

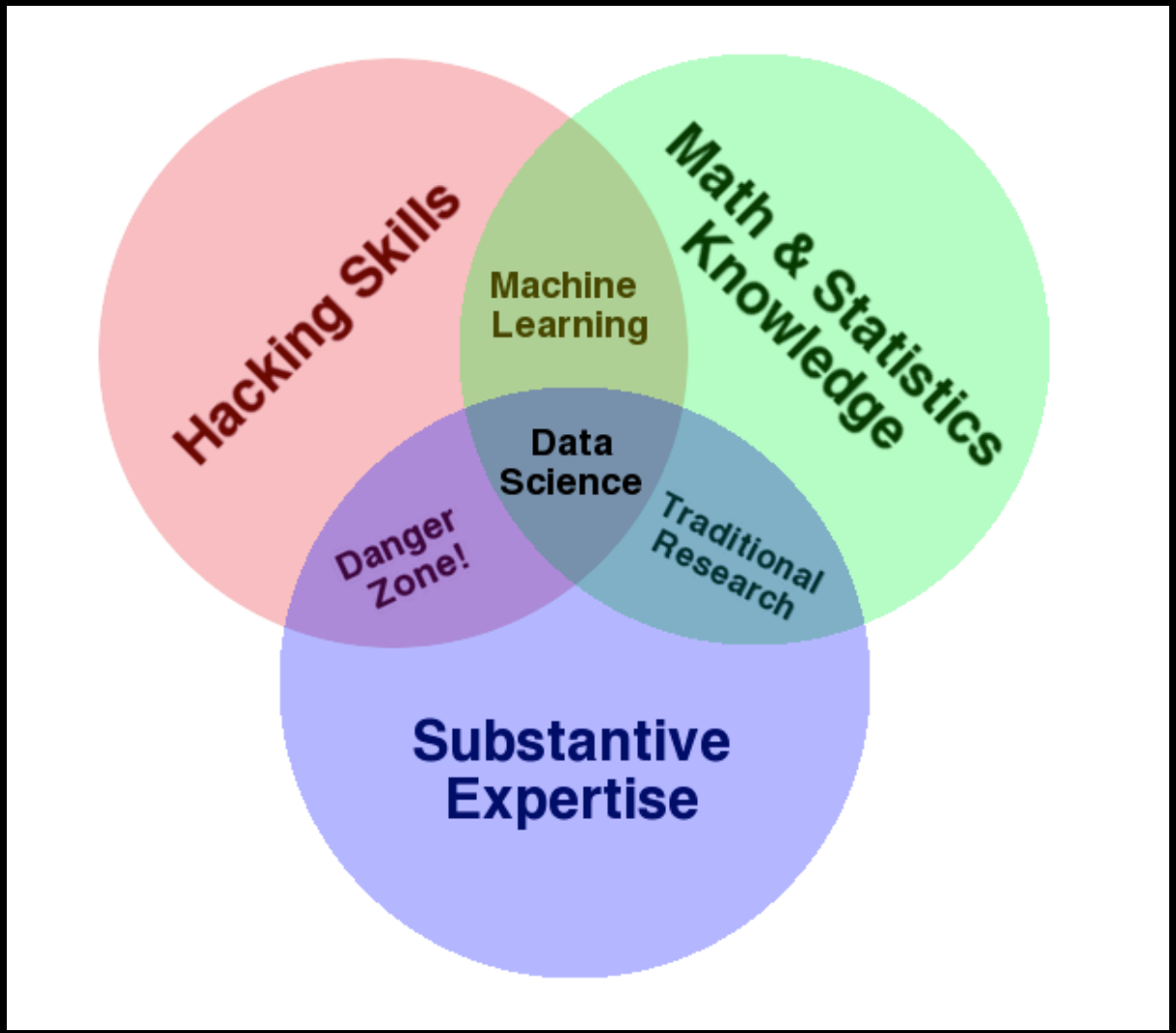
For Example


Best fit to MINOS+T2K data
68% CL
90% CL
95% CL
$\Delta m^2_{32} > 0$
$\delta\,(\pi)$
$sin^2 2\theta_{13}$


NOνA Far Detector Cosmics Data
Charge [arbitrary units] / cm
Distance to track end (cm)

So, you want to be a data scientist?

Data Science...



Drew Conway

Course Overview

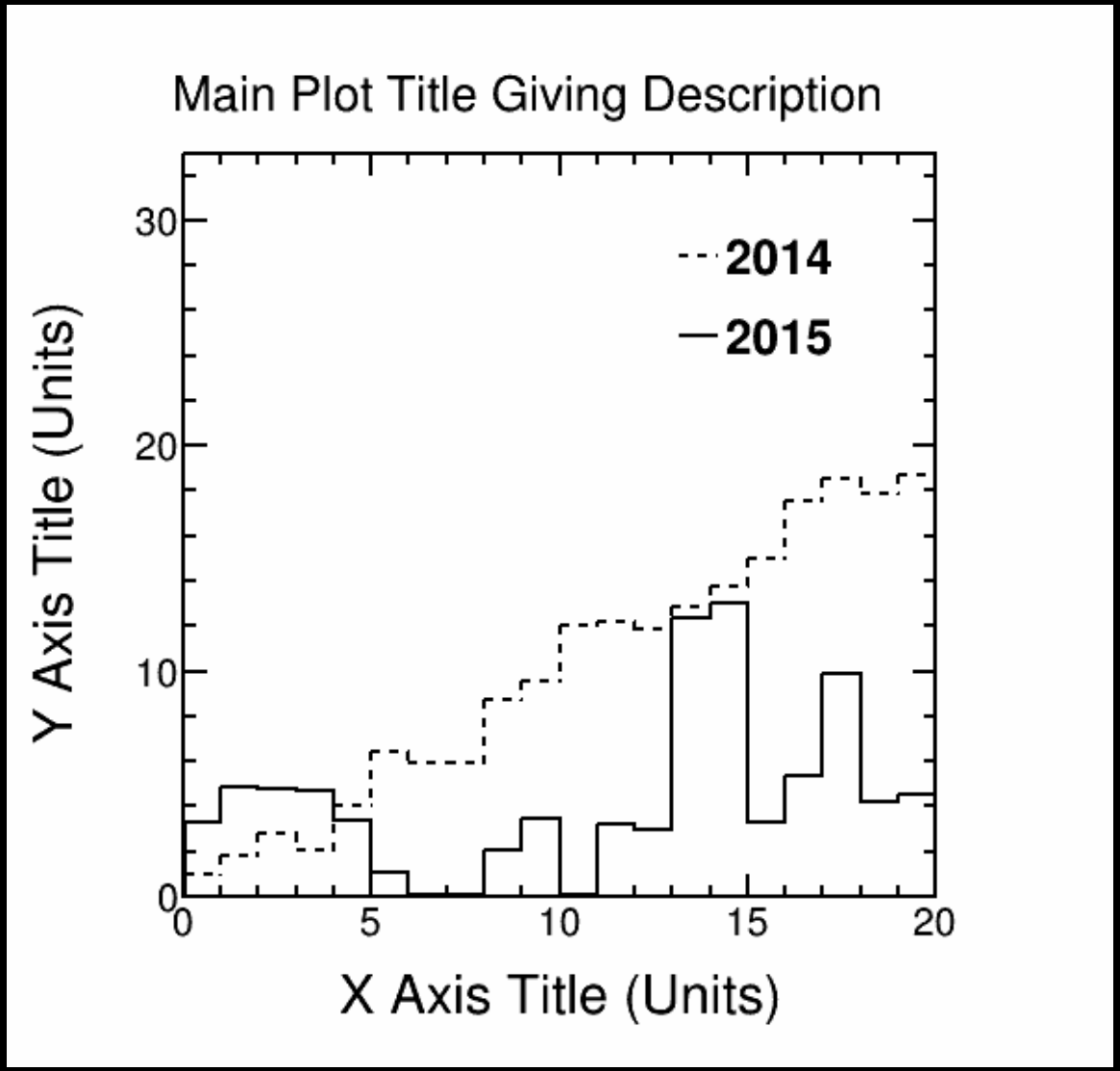
Go over document and starting points…

Week One: Introduction to Visualisation

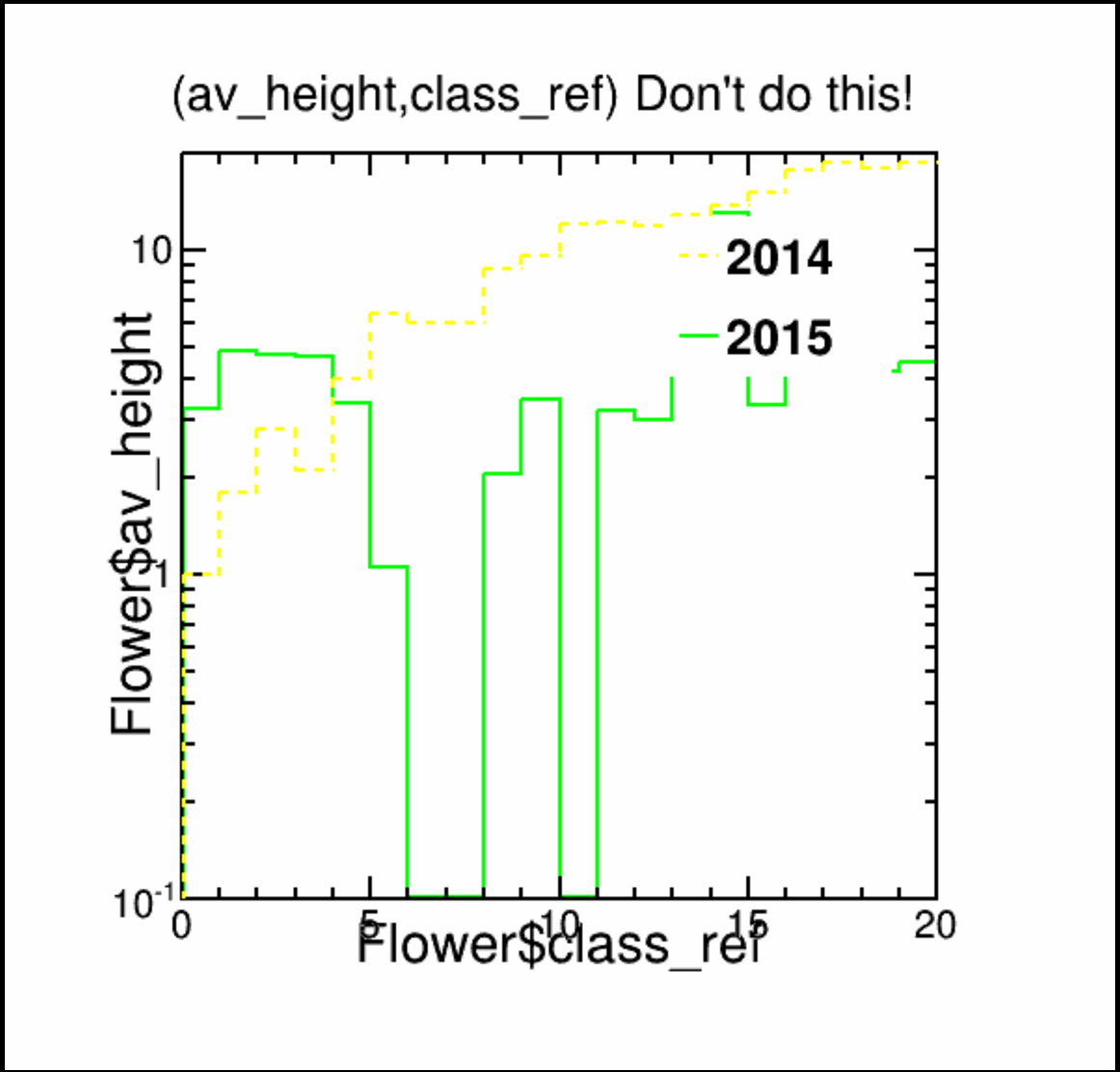
A Good Plot

- Title
- Axis titles
- Numbers and units on all axes
- Legend labelling all lines if more than one
- Clear what is plotted
- Legible
- Colours visible on projectors and in print

Plot Outline



Bad Plot...


(av_height,class_ref) Don't do this!
2014
2015
Flower$av_height
Flower$class_ref
10
1
10^-1
0
5
10
15
20

Assignments

Will involve making descriptive plots for one of the R example data sets. The details of the assignment will be given during the class. You will be given some specific plots to make to gain familiarity with the plotting functions in R.

Optional Prep

If you are not already familiar with R then I recommend you familiarize yourself with the basics before the start of the course. You can find some introductory material free online, I suggest working through this tutorial:

- https://www.nceas.ucsb.edu/files/scicomp/Dloads/RProgramming/BestFirstRTutorial.pdf

There are also lots of screen-casts available on YouTube, for example this series:

- https://www.youtube.com/watch?v=UYcImg1_KLk

Working with built in R data

Display available data sets: `data()`

Once you chose one you can get help, to find its structure: `str(<data>)` and its description: `help(<data>)`.

You can use the `help()` function in general to see documentation

Basic Plotting

Two options you will most commonly use, I think are `hist()` and `plot()`.

Also see the package ggplot2: http://docs.ggplot2.org/current/

Accessing Data Frames

To begin with look at what variables you have: `str(<dataframe>)`

To see elements: `head(<dataframe>)`, or matrix-like access e.g. show all columns first five elements: `dataframe[1:5,]`

Get the content of each row by variable name:
`dataframe$variable`

Problem 1: Irises

Use the built in data frame ``iris''. For all of the plots, make sure that you have human-readable titles and clear labelling (i. e. no variable names please)

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.

Backup Slides