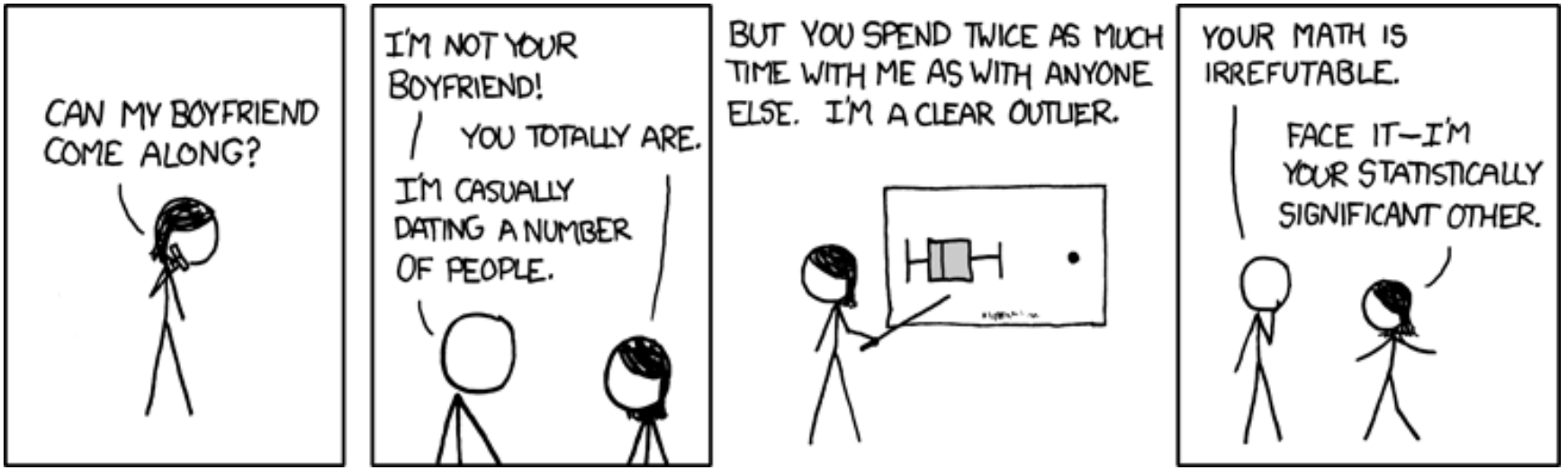
Introduction to Data Science with R

Suchana Seth

December 2014

What is Data Science ?



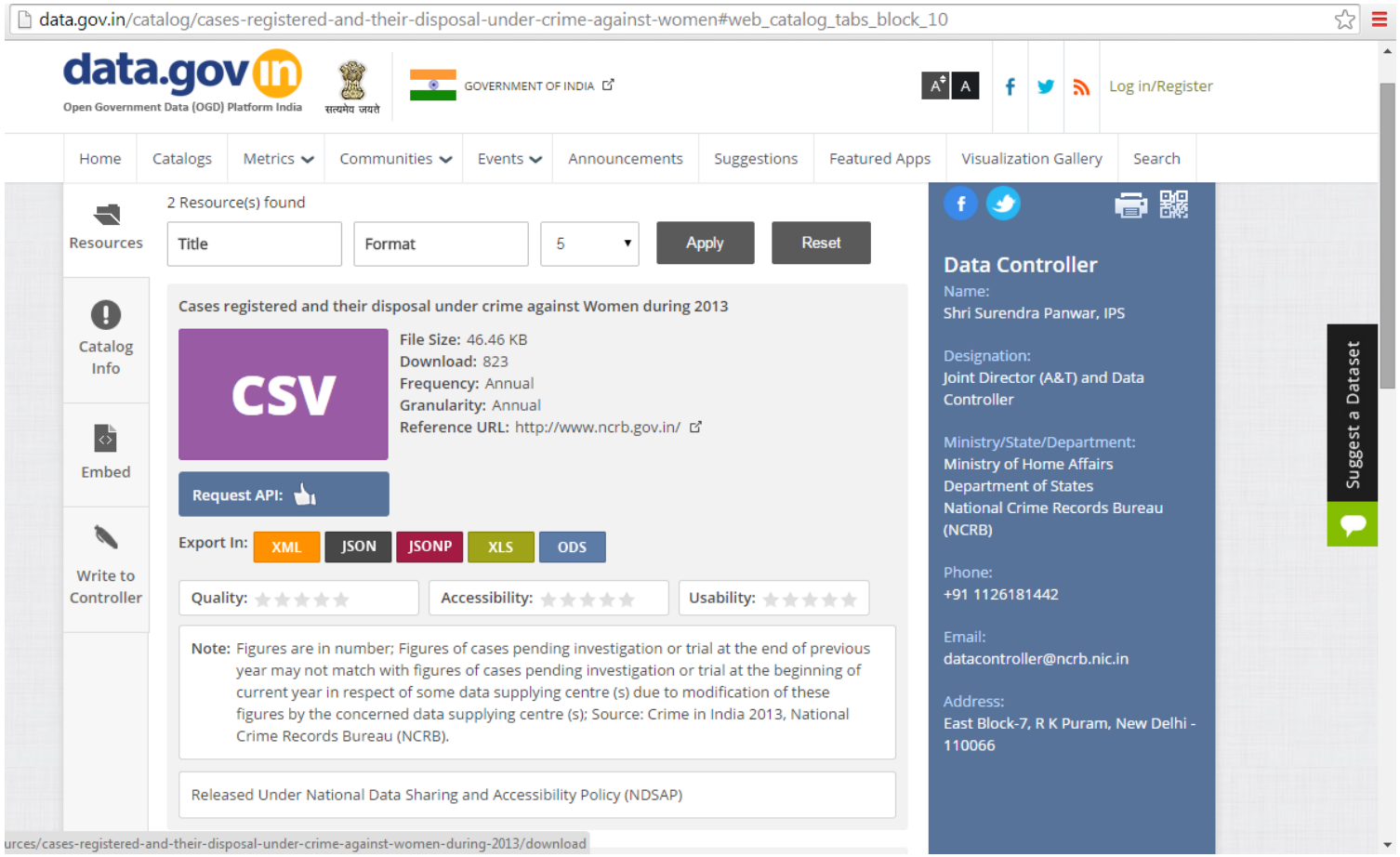
R Basics – everything is vectorized

The most basic variable in R is a `vector`. An R vector is a sequence of values of the same type. All basic operations in R act on vectors (think of the element-wise arithmetic, for example). The basic types in R are as follows.

`numeric`	Numeric data (approximations of the real numbers, $\mathbb{R}$)
`integer`	Integer data (whole numbers, $\mathbb{Z}$)
`factor`	Categorical data (simple classifications, like *gender*)
`ordered`	Ordinal data (ordered classifications, like *educational level*)
`character`	Character data (strings)
`raw`	Binary data

All basic operations in R work element-wise on vectors where the shortest argument is recycled if necessary. This goes for arithmetic operations (addition, subtraction, ...), comparison operators (==, <=, ...), logical operators (&, |, !, ...) and basic math functions like `sin`, `cos`, `exp` and so on. If you want to brush up your basic knowledge of vector and recycling properties, you can execute the following code and think about why it works the way it does.

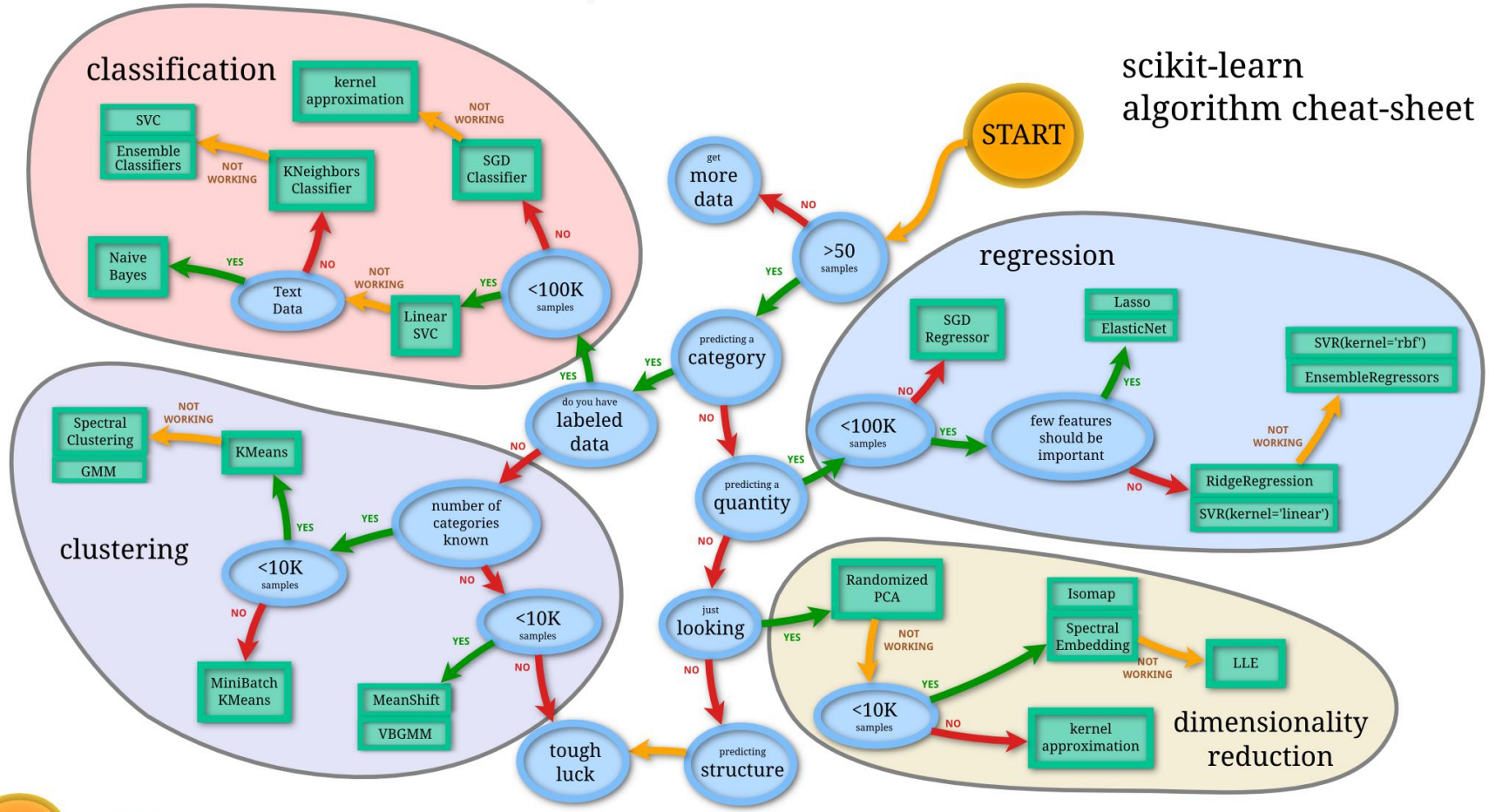
About our data set – Crimes Against Women in India in 2013



ggplot2

- **ggplot** - The main function where you specify the dataset and variables to plot
- **geoms** - geometric objects
  - geom_point(), geom_bar(), geom_density(), geom_line(), geom_area()
- **aes** - aesthetics
  - shape, transparency (alpha), color, fill, linetype.
- **scales** Define how your data will be plotted
  - *continuous*, *discrete*, *log*

How to choose an algorithm for your problem



Accuracy Measures

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Miss Rate or False Negative Rate (FNR)

$$FNR = FN/P = FN/(FN + TP)$$

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score

is the harmonic mean of precision and sensitivity

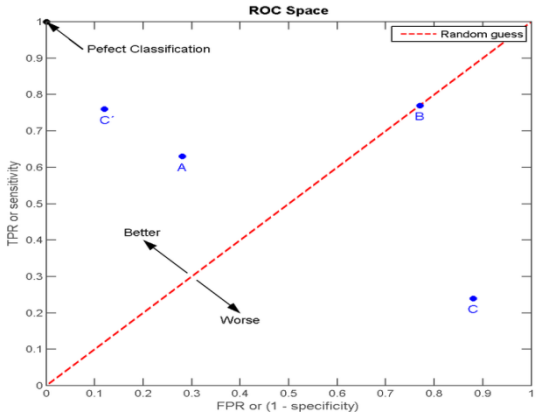
$$F1 = 2TP/(2TP + FP + FN)$$

Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

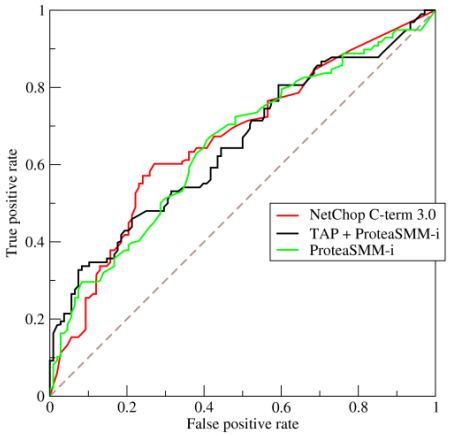
Informedness = Sensitivity + Specificity - 1

Markedness = Precision + NPV - 1



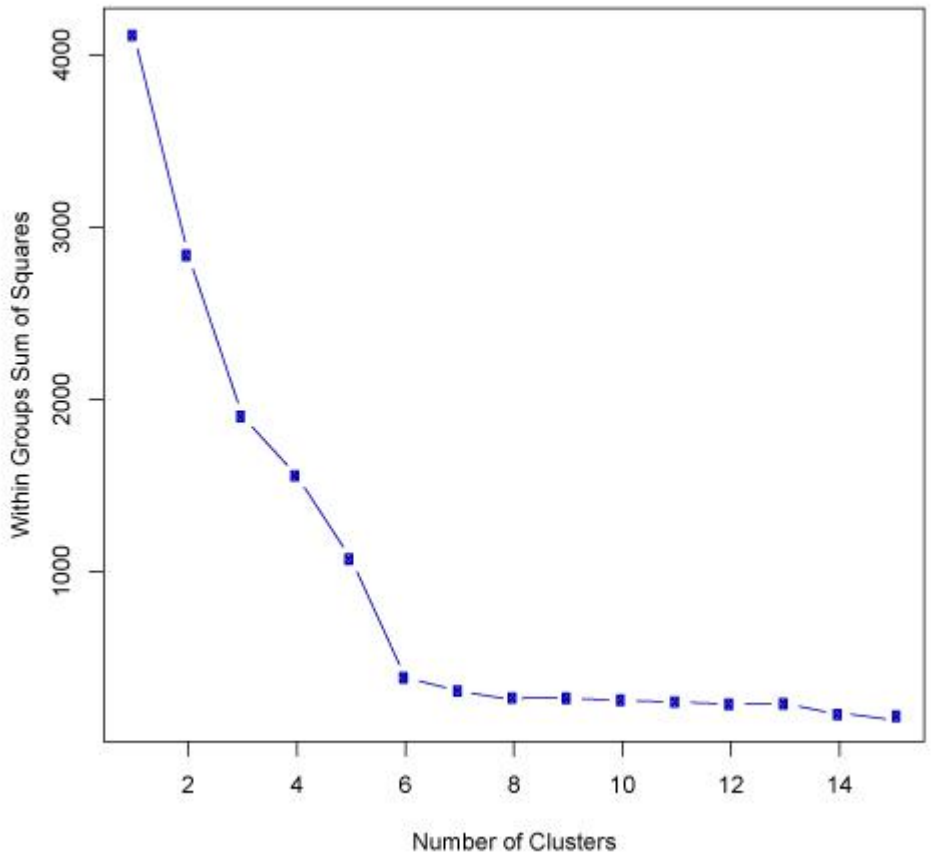
Theoretical

Actual example



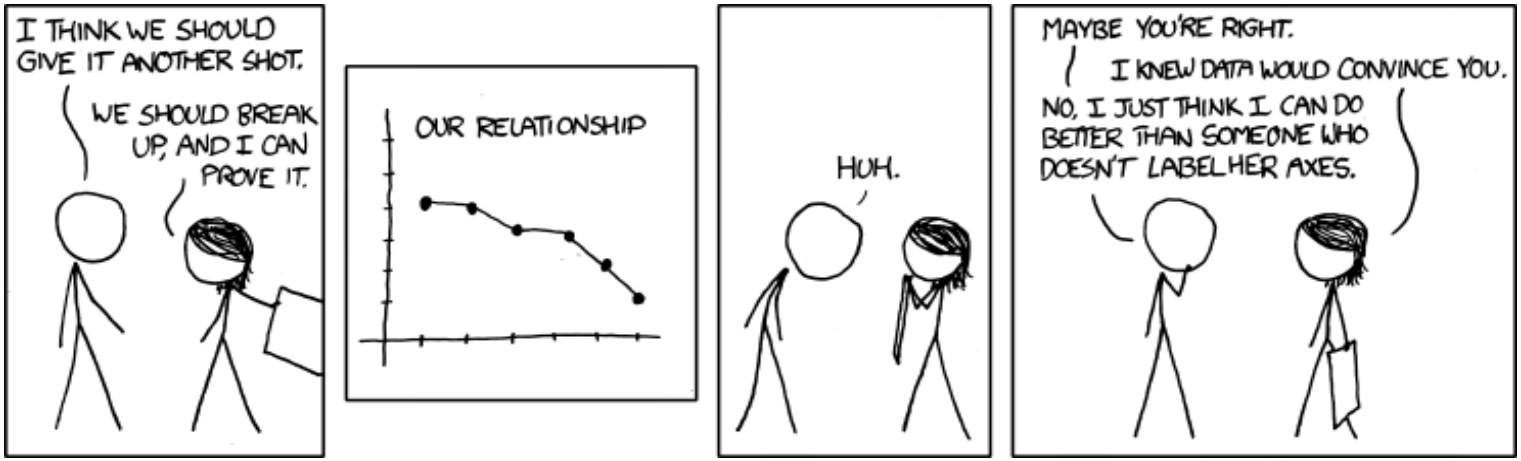
Choosing the appropriate number of clusters

One common method of choosing the appropriate cluster solution is to compare the sum of squared error (SSE) for a number of cluster solutions. SSE is defined as the sum of the squared distance between each member of a cluster and its cluster centroid. Thus, SSE can be seen as a global measure of error. In general, as the number of clusters increases, the SSE should decrease because clusters are, by definition, smaller. A plot of the SSE against a series of sequential cluster levels can provide a useful graphical way to choose an appropriate cluster level. Such a plot can be interpreted much like a scree plot used in factor analysis. That is, an appropriate cluster solution could be defined as the solution at which the reduction in SSE slows dramatically. This produces an "elbow" in the plot of SSE against cluster solutions. In the example shown below, there is an "elbow" at the 6 cluster solution suggesting that solutions >6 do not have a substantial impact on the total SSE.



http://www.mattpeeples.net/kmeans.html

How (Not) To Do Data Science ?

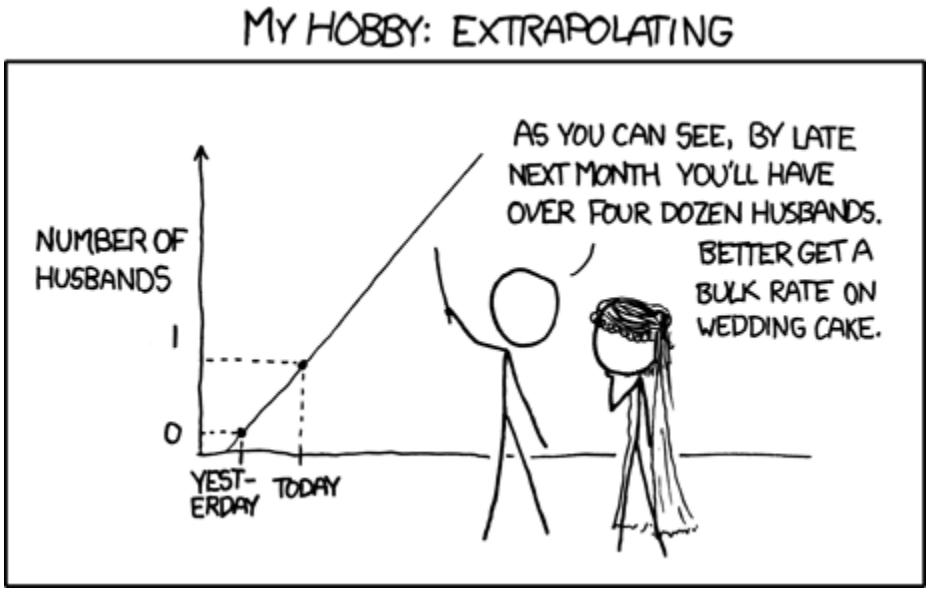


Label your axes

Don’t compare apples to oranges

Don’t extrapolate

Be sceptical – especially about your pet theories



Resources

- shailesh kumar google - fifth elephant talk on data science - https://hasgeek.tv/fifthelephant/all-videos
- Del Harvey's Twiiter & Data at Scale talk on ted - http://www.ted.com/talks/del_harvey_the_strangeness_of_scale_at_twitter?language=en
- By Sharon Machlis - http://www.computerworld.com/article/2497143/business-intelligence-beginner-s-guide-to-r-introduction.html
- coursera
- stack overflow
- quora
- r-bloggers
- conferences / events -
- fifth elephant
- grace hopper
- data kind
- + books – Bishop | Hal Daume | Tiibshirani
- somewhat math heavy but great
- hastie & tibshirani's books + vids -
- http://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/
- data viz
- http://www.cookbook-r.com/Graphs/
- fun things to do for devs -
- r shiny
- build r packages
- Art of R Programming – great book