Doing Data Science

Unit 1

Faizan Javed
DataScience @ SMU

Doing Data Science, section 403

Instructor: Faizan Javed (also the coordinating faculty)
fjaved@smu.edu, faizan.javed@gmail.com

Grader: Tom (Keyue) Wang, keyuew@smu.edu

Grading:

Live Session Assignments (50%) (8 homeworks)

HW1 is due next Monday, 1 hour before live session

Case Studies (30%),

Videos and Questions during asynchronous material (BLTs) (15%),

Live Session Attendance (5%).

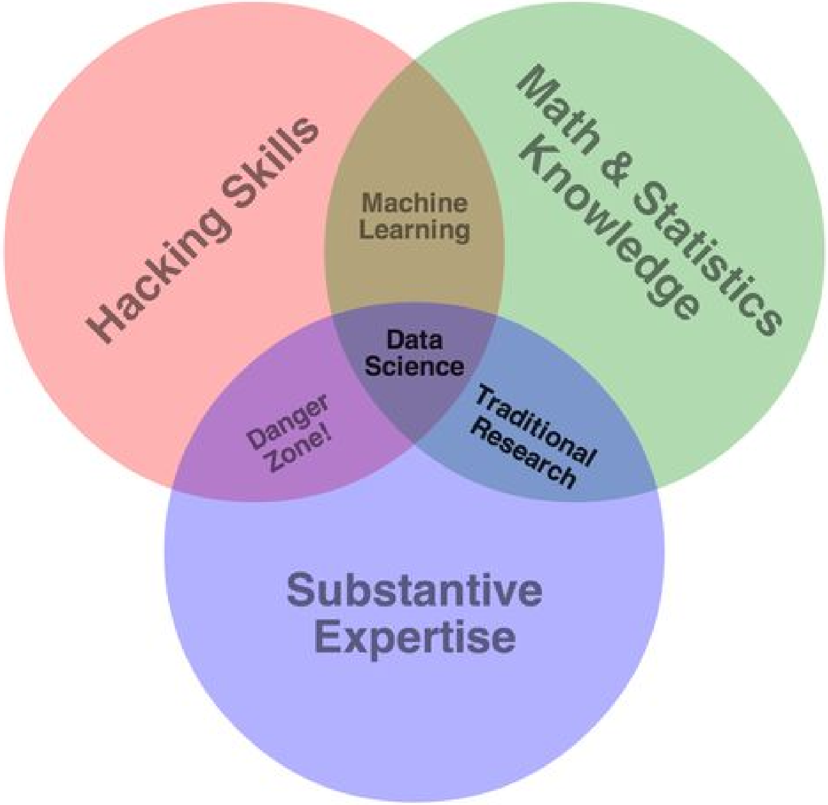
What is Data Science?

Reproducible vs Replicable

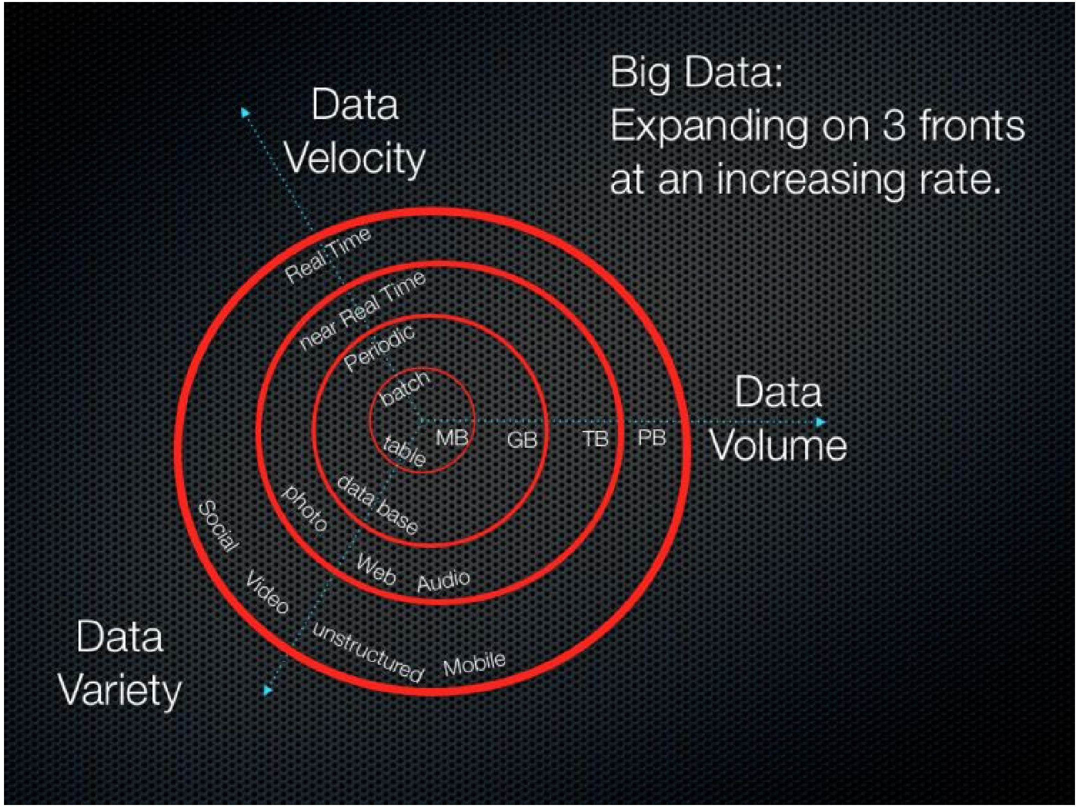
Tools for Data Science

Data Science Venn Diagram

https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html



Big Data: the 3 Vs



Data Science profile

Computer Programming

Math

Statistics

Machine Learning

Domain Expertise

Communication, Presentation Skills

Data Visualization

Reproducible research

Use specific set of computational functions/analyses (usually specified in terms of code) and original data to recreate findings

Replicable: recreate findings with new data

A study is only replicable if you perform the exact same experiment (at least) twice, collect data in the same way both times, perform the same data analysis, and arrive at the same conclusions.

http://languagelog.ldc.upenn.edu/nll/?p=21956

Why is replicability important?

Open Source

Public research

Business/Private Sector

Tools for reproducible research

R

Knitr

RMarkdown

RStudio

Section 1.5.1, install R and RStudio, and all packages on page xix

Workflow of Reproducible research

Data Gathering

Data Analysis

Results Presentation

Machine learning

Training data (e.g. implicit and explicit feedback on job ads)

Target function (e.g. probability of user clicking and/or applying for a job)

Metric (e.g. precision vs. recall, or any ranking metric that correlates to AB test metrics)

Practical tips

Document everything

Ensure compatibility of your software/libraries/packages (troubleshooting assistance)

Comment your code (variable/argument names should not be cryptic)

Source code comment header

Getting started with R

#Print R session info -- why is this useful?

```
sessionInfo()
```

# c(combine) function : create vectors

```
NumVec ← c (2, 3, 4)

CharVec ← c (“doing”, “data”, “science”)
```

# Reassign row.names

```
row.names(StringNumObject) ←- c("First", "Second", "Third")
```

# data.frame() : create an object with rows and columns

StringNumObj ← data.frame(NumVec, CharVec)

# cbind()/rbind() : combine vectors side-by-side

StringNumObj ← cbind(NumVec, CharVec)

Why use data.frame when we have cbind()/rbind()?

```
# $ : component selection

NewNumeric ←  StringNumObject$NumericVect



# head()/tail() : select first/last few rows

head(cars)



# [rows,columns] subscript operators, : sequence operator

cars[3:7, ]
```

# str()

compactly display the structure of an R object

# summary()

display summary statistics for analysis (mean, quantiles, etc)

# dim()

retrieve or set the dimensions of an object (array, matrix, dataframe)

Missing/extreme values:

NA = not available

NaN = undefined

Inf = extremely small/large (infinity)

What did you learn today?

Questions?