# Objectives

The learning objectives for the project module of MSc Data Science can be described as follows (pending TC approval, changes are highlighted in boldface):  
  
*"The module should demonstrate:  
- A knowledge of programming a piece of software that goes beyond a few lines of code, applying****what was learnt in the eight taught modules****to a concrete problem.  
- The ability to develop the design of a software****solution****to a concrete problem****that can be identified as data analytics/data science****.  
- perform abstract thinking  
- exhibit abstraction skills,  
- use a coherent development process  
-****exhibit ability to validate and analyze the results****"*

# Birkbeck project:

Folder: D:\Projects\DropBox\Dropbox\Birkbeck\Project\PMResearchDatabase

# Python code

Folder: D:\Projects\RDM\DRE\ODBC\_MSAccess\_test

# Source data base:

File: December17Master.accdb (FSID2005$)

Folder: I:\DRE\Projects\Research\0004-Post mortem-AccessDB\DataExtraction

Contents: Post mortems from 1996 –

Notes:

* Post mortems with year 21 or 31 are from another location for 2012 & 13 – Ignore.

# Staging data base:

File: PMResearchReportDB.accdb

Folder: I:\DRE\Projects\Research\0004-Post mortem-AccessDB\DataExtraction

Notes:

* Linked tables from source database
* HAS tables
* Working tables
  + Lab results

# Create\_HAS\_Tables.py

* create\_has\_tables
  + DROPs and CREATEs HAS Tables in Staging database
* runTests
* CreateEvents
* CreateAttributeFromAttributes
* CreateAttributeNoOfAttributes
* CreateLabEvents
* create\_reporting\_attributes
* CreateHASCSVFiles

# Upgrade to HAS v1.08

* ha\_concepts
  + New: parent\_concept\_id
  + New: value\_type\_concept\_id
  + Update: orders code, label & term => code, term & label
  + New: note
  + Update: concept\_type => category
  + Delete: concept\_value type; replaced by value\_type\_concept\_id
* ha\_staff
  + Ordering of columns
  + Add: full\_name
  + Update: staff\_type\_id => staff\_type\_concept\_id
* ha\_patients
  + Update: birth\_date => birth\_datetime
  + Update: death\_date => death\_datetime
  + Add: deceased\_flag
  + Update: geographic\_zone => zone
  + Add: project\_code
* ha\_patient\_attributes
  + Add: parent\_patient\_attribute\_id
  + Update: patient\_attribute\_type\_id => patient\_attribute\_type\_concept\_id
  + Add: sequence\_number
  + Update: value\_date => value\_datetime
  + Add: value\_boolean
  + Update: value\_id => value\_concept\_id
* ha\_events
  + Update: event\_type\_id => event\_type\_concept\_id
  + Add: note
* ha\_event\_attributes
  + Add: parent\_event\_attribute\_id
  + Update: event\_attribute\_type\_id => event\_attribute\_type\_concept\_id
  + Add: sequence\_number
  + Update: value\_date => value\_datetime
  + Add: value\_boolean
  + Update: value\_id => value\_concept\_id

# Project Diary

* Monday, 24th June
  + Review work to date
  + SQL structural changes to match HAS 1.08
  + Created GitHub for MSc project
  + Coded SQL changes and tested
  + Started on run\_tests
* Tuesday, 25th June
  + Problems with Access and self-referential concepts
  + Create:
    - create\_concept\_id
    - create\_core\_concepts
  + Completed run\_tests.
    - addPatientAttribute
    - addEventAttribute
  + Checked all queries in AccessDB
* Wednesday, 26th June
  + CreateEvents & CreateEventAttributes
  + Completed for first 50 cases
  + Started on creation of RDVs
    - Nearly there with all events and all columns
* Thursday, 27th June
  + Completed creation of RDV for all data
  + Added column selection
  + Added row filtering
  + Added creation of XML TDF file.
  + Created new test data set with 1,000 events
* Friday, 28th June
  + CreateCOD2\_SUMMAttributeFromCOD2Attribute.
  + CreateAttributeNoOfAttributes
  + create\_reporting\_attributes
  + Created RDV with new attributes
  + Initial decision Tree in R using an RDV produced from HAS.
* Saturday, 29th June (Traveling back from Avignon)
  + Create\_rdv\_ext\_measurements
    - Separate file for each age\_category
* Monday, 1st July
  + Added sex to RDV’s plus other minor changes T/F for Boolean, etc
  + Found bug in CreateEvents
  + Generate full data set
* Thursday, 4th July
  + Project plan
  + Organ weights
    - Remove outliers/errors
      * Check organ weights against body weights
* Friday, 5th July
  + Update\_event\_attribute\_value
    - NB Event\_id might change but Caseid always the same.
* Saturday, 6th July
  + create\_attribute\_inc\_in\_study
  + Changed INC\_IN\_STUDY from Boolean to Category
  + Manual inspection of measurements
    - Used rdv\_measurements to build up criteria to exclude from study
  + Developed exclude\_event\_attributes
  + Use of NULL as a parameter
  + Produced rdv\_study\_ext.csv
  + Produced basic Decision tree
  + Started on Random Forest, stalled as couldn’t install packages.
* Sunday, 7th July
  + Uninstall and re-install R and R studio
  + Check install packges:
    - dplyr
    - rpart
    - RandomForest
    - claret
    - e1071
    - gbm
    - xgboost
* Monday, 8th July
  + Worked on for loops for decision tree control variables
  + Determined max accuracy
  + Plot variations in accuracy as control variables change
  + Re-plot decision tree with max accuracy.
* Tuesday, 9th July (train to Newcastle)
  + Random Forest initial analysis
  + Started on XGBoost but missing packages
  + Created rdv\_study\_int1
  + Ran scripts for decision tree and random forest
* Wednesday, 10th July
  + Got RMD document creation working
  + Created rmd document dtree\_study\_ext.
* Saturday, 13th July
  + Created spreadsheet to record study results
    - Include accuracy of COD1 & COD2
  + Got gradient boost working with
    - Caret
    - XGBoost
* Sunday, 14th July
  + Recorded importance for all models
  + Run models for study\_int1
    - Record prediction accuracy & feature importance
* Monday, 15th July
  + Review with Neil & Ben
  + Started looking at one-hot encoding & z-score normalisation in python to act on RDV
* Tuesday, 16th July
  + Completed one-hot encoding & z-score normalisation modification to CSV file.
  + Column types
    - Ignore
    - Classification – one-hot encoding, include missing vales
      * Now many and what values does the column have
    - Numerics - Z-score Normalization
      * Calculate mean and standard deviation
      * Missing values, what happens if normalised?
* Wednesday
  + Started rerunning model with adjusted rdv.
* Saturday, 20th July
  + Created attributes for Macro & Histo Case and System
  + Created rdv’s for adding Macro & Histo
  + Added facility to created rdv’s for single sections rather than cumulative
  + Change NULL => nan
    - Include in TDF, check works on platform
* Sunday, 21st July
  + Bug in attribute creation – duplicate attributes same type different values was stopping rdv creation. 2 Hours!
  + Created RDVs
    - Organ weights without external measurements
    - Macroscopic examination
    - Histopathology examination
  + Spent a lot of time with models especially decision trees and XGBoost.
  + Found bug and a lot of attributes assigned to incorrect event. Corrected bug, re-created RDVs and ran model to check.
* Monday, 22nd July
  + Set up dtree\_study\_all
    - One script to process all stages
* Tuesday, 23rd July
  + Set up gboost\_study\_all
  + Worked on dtree\_all
    - Stored results in matrix
    - Set up for loop to process all stages
    - Save matrix as CSV
    - Stalled on saving feature importance
* Wednesday, 24th July
  + Implemented storing feature importance
  + Random seed based on milliseconds of now.
  + Implemented For loop and result saving for XGBoost
* FRIDAY, 26TH July
  + TDF file for adj.csv
  + Linear regression for heart\_weight vs age\_in\_days.
* Saturday, 27th July
  + Completed age\_in\_days adjustment using simple linear regression.
  + Write out parameters used for adjustment to CSV files
  + Ran decision trees and XGBoost on newly revised data.
* Sunday, 28th July
  + Heat maps for feature importance using ggplot tiles
  + Combined heat maps combining feature importance across models
  + Including saving plots to file
* Monday, 29th July
  + Added sex and age breaks in age\_in\_days normalisation
  + Removed z-score standardisation
* Tuesday, 30th July
  + Review with Nigel at Birkbeck
* Thursday, 1st August
  + Setting up project report in word. Battle with numbered headings.
* Saturday, 3rd August
  + Created structure to store all files from all models for a run in a single folder
  + Get Random Forest in structure
* Sunday, 4th August
  + Combined heatmap include RF; played with sensitivity and added displaying values
  + Modified all models to run as functions
  + Plotting Linear regression in modify.csv (python)
  + Better understood RPART library. Modified Decision Tree model.
* Monday, 5th August
  + Use one folder for all runs, add file suffix
  + Plot accuracy by run
* Tuesday, 6th August
  + Added documents to git.
* Wednesday, 7th August
  + Reviewed MscProject18-19.pdf and made notes in document outline.
* Friday, 9th August
  + Plan for the week-end
* Saturday, 10th August
  + Removed stage 5: int3\_s
  + Modify RF and XGBoost to use 4 stages and write to a single folder
  + How does prediction accuracy for each random seed change by stage:
    - X=Run, Y=Accuracy, Group = stage, Facet = model
  + How does prediction accuracy for each PM stage change by model:
    - X=Stage, Y=Accuracy, Group = Model, Facet = Run
  + Compare Feature importance change by random seed.
    - Compare for each model one stage for a number of random keys
* Sunday, 11th August
  + Modified final results comparison heat maps
  + Simplified for loops in R
  + Tuning XGBoost model with grids
  + Minor tweaks to subplots in modify\_csv.py
* Monday, 12th August
  + Saved combined results csv files
  + Minor tweaks to Linear regression plots
* Tuesday, 13th August
  + Worked on Project Report outline
* Wednesday, 14th August
  + Worked on project report outline
  + Review with project supervisor
* Sunday, 18th August
  + Return to tuning XGBoost model.
    - Developed better process
  + Started to improve feature importance plots
* Monday, 19th August
  + Completed Feature importance plot changes
  + Changed creation of train.index so all models use same data split.
  + Ran a couple of sets and results look much better.
* Tuesday, 20th August
  + Start on Fridays presentation
  + Creating plots of original data
* Wednesday, 21st August
  + Completed first draft of presentation
* Friday, 23rd August
  + Presents images to DS team & Neil
* Saturday, 24th August
  + Changed colour scheme to colour blind aware
  + Added mosaic plot for age\_category
  + Added summary\_inc\_in\_study for clarity
  + Realised error in excluding age\_category 999.
  + Re-did inc\_in\_study visualisations
  + Reproduced all RDVs inc adjusted
  + Rerun models
* Sunday, 25th August
  + Redid combined visualisations
  + Changed to box plot for combined accuracy by model by stage
  + Added table plots and included break down by C001 and C002
  + Inserted new graphics into presentation.
* Monday, 26th August – Bank holiday
  + Nothing!
* Tuesday, 27th August
  + Minor tweaks to visualisation plots.
* Friday, 30th August
  + Minor tweaks to project report outline
* Sunday, 1st September
  + Project report – Methods
* Monday, 2nd September
  + Project report
    - Data Engineering
    - Summary of Literary review

# To Do

* Store results of Decision tree to csv, gives details of splits.
* Plotcp(tune\_fit) for Decision Tree
* Change column names for Macro & Hist remove \_SyFiID etc.
* Visualisations of models
  + Random forest
  + XGBoost
* Data and scripts onto the platform
* Move R code to RMDs
* What is “accuracy” in RandomForest?
* What is Nodesize in random forest?
* Create measurement RDVs
  + Look at correlations - visualisations
* Continue with ReadMe.md
  + Describe more routines
* Add Lab Events
  + NB create\_reporting\_attributes needs sorting
* Value\_type\_concept\_id
  + Global variables
* Rationalise naming conventions
  + Functions either camel case or snake case
* Review with Ben, questions:
  + Why COD2\_SUMM = 003 i.e. other
  + Understand rpart and rpart.plot
    - What do they mean?
    - Plot decision boundaries
      * COD2\_summ vs Age\_in\_days
  + Visualisation of Random Forest results
  + Gradient Boosting
    - Learning rates
      * Eta too low settle into local minimum
      * Log steps: 0.001, .01,.1, etc
        + See model diagnostics
      * Max depth should match tree
        + Try 8 rather than 5
  + Try number of trees with different orders of magnitude.
  + Consistency in results for RandomForests with different random seeds?
  + Feature normalisation
    - x-mean/standard deviation =
    - <https://en.wikipedia.org/wiki/Feature_scaling>
    - <https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e>
  + Handling of categorical variables
    - one-hot encoding
      * <https://en.wikipedia.org/wiki/One-hot>
      * <https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding>

# Project Notes

* Add comment on why I am using an over complex method to create analytic data sets.
  + Data engineering learning for main data integration project.
* Which attributes to consider
  + Am I interested in System/organ investigation?
    - Once at this level does it matter which one you do?
* Common ontology
  + Common UK child death database.
* Gestational age at birth added to external
  + How many cases will be excluded?
* Lab results
* Acknowledge Aridhia for HAS 1.08 format
* NB Thymus weight, when recorded it is important.
* <https://stackoverflow.com/questions/32108179/linear-regression-normalization-vs-standardization>
  + That makes sense because normalization and standardization do different things.
  + Normalization transforms your data into a range between 0 and 1
  + Standardization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1
  + Normalization/standardization are designed to achieve a similar goal, which is to create features that have similar ranges to each other. We want that so we can be sure we are capturing the true information in a feature, and that we dont over weigh a particular feature just because its values are much larger than other features.
  + If all of your features are within a similar range of each other then theres no real need to standardize/normalize. If, however, some features naturally take on values that are much larger/smaller than others then normalization/standardization is called for
  + If you're going to be normalizing at least one variable/feature, I would do the same thing to all of the others as well
* <https://stackoverflow.com/questions/33999512/how-to-use-the-box-cox-power-transformation-in-r>
* <https://en.wikipedia.org/wiki/Box%E2%80%93Cox_distribution>

Continuing

# Project Plan

|  |  |  |  |
| --- | --- | --- | --- |
| Week beginning | Week | Saturday | Sunday |
| 5th August |  |  |  |
| 12th August | NM Review: Thu 15th August |  |  |
| 19th August | NM Review: Thu 22nd August |  |  |
| 26th August | NM away | Cumbria  Report write up | Cumbria  Report write up |
| 2nd September | Cumbria  Report write up |  |  |
| 9th September |  |  |  |
| 16th September | Hand in report |  |  |

NB Nigel away: 18th – 29th July and 26th – 8th August

# References

Furlong, K.R., Anderson, L.N., Kang, H., Lebovic, G., Parkin, P.C., Maguire, J.L., O’Connor, D.L., Birken, C.S. and TARGet Kids! Collaboration, 2016. BMI-for-age and weight-for-length in children 0 to 2 years. *Pediatrics*, *138*(1), p.e20153809.

Coppoletta, J.M. and Wolbach, S.B., 1933. Body length and organ weights of infants and children: a study of the body length and normal weights of the more important vital organs of the body between birth and twelve years of age. *The American journal of pathology*, *9*(1), p.55.

# Links to follow:

## ggplot ideas

<https://stackoverflow.com/questions/17150183/plot-multiple-lines-in-one-graph>

<http://www.sthda.com/english/wiki/ggplot2-line-types-how-to-change-line-types-of-a-graph-in-r-software>

<https://stackoverflow.com/questions/27350243/ggplot-line-graph-with-different-line-styles-and-markers>

## Data science report

<https://career-resource-center.udacity.com/portfolio/data-science-reports>

<http://blog.kaggle.com/2016/06/29/communicating-data-science-a-guide-to-presenting-your-work/>

<https://smallbusiness.chron.com/write-data-report-61330.html>

<https://www.dataquest.io/blog/data-science-project-style-guide/>

# Create Macro & Histo summary and system attributes

* Heirachy
  + Do system ones first
  + The summary is a summary of systems not all
* What are the possible attributes
  + Which are we interested in
* Select systems
  + /EventAttribute/Observation/PostMortem
    - tblCardiovascularSystems
      * i.e. start with tbl and end with systems
    - HeartMacro\_OrFi
      * Whether an Abnormality exists
    - /EventAttribute/Observation/PostMortem/LookUp/HeartMacro\_OrFiID

| **qryEAV\_Concepts\_02** | |
| --- | --- |
| **ha\_concepts.code** | **label** |
| 001 | Normal |
| 002 | Abnormal but NOT contributed to death |
| 003 | Abnormal and POTENTIALLY contributed to death |
| 004 | Abnormal and DEFINITIVE cause of death |
| 005 | Not Examined |
| 006 | Too Autolysed |
| 999 | N/A |

* + - HeartHisto\_orFi
      * Whether an Abnormaility identified in histo exam
    - /EventAttribute/Observation/PostMortem/LookUp/HeartHisto\_OrHiID

| **qryEAV\_Concepts\_02** | |
| --- | --- |
| **ha\_concepts.code** | **label** |
| 001 | Normal |
| 002 | Abnormal but NOT contributed to death |
| 003 | Abnormal and POTENTIALLY contributed to death |
| 004 | Abnormal and DEFINITIVE cause of death |
| 005 | Not Examined |
| 006 | Too Autolysed |
| 007 | Pres norm, organ-specific histo taken NOT reported |
| 008 | Pres norm, "PM histo" taken but organ NOT reported |
| 999 | N/A |

* + - HeartMacroAbn\_HeMaID
      * Actual abnormalities
* What systems?
  + Category = ‘/EventAttribute/Observation/PostMortem’
  + Code Like ‘tbl\*systems’
* What Macro/Histo exams in system
  + Category = ‘/EventAttribute/Observation/PostMortem/tbl[system\_name]systems’
  + Code Like ‘\*Macro\_OrFiID’
  + Concepts
    - Category = /EventAttribute/Observation/PostMortem/LookUp/[organ\_name]Macro\_OrFiID
  + Code Like ‘\*Histo\_OrHiID’
  + Concepts
    - Category = /EventAttribute/Observation/PostMortem/LookUp/[organ\_name]Histo\_OrHiID
* Create Concepts
  + Group codes
    - Macro & Histo
      * ‘001’ Normal
      * ‘002’ Abnormal but not COD
      * ‘003’ & ‘004’ Abnormal COD
      * Other N/A
  + Get List of Systems
    - Category = ‘/EventAttribute/Observation/PostMortem’
    - Code Like ‘tbl\*Systems’
  + ~~Get List of Organs for System~~
    - ~~Category = ‘/EventAttribute/Observation/PostMortem/tbl[system\_name]Systems’~~
    - ~~Code = ‘\*Macro\_OrFiID’~~
    - ~~Code = ‘\*Histo\_OrHiID’~~
  + Codes Case
    - Category = /EventAttribute/Observation/PostMortem
    - Code = CaseMacro\_CsFiID
    - Category = /EventAttribute/Observation/PostMortem
    - Code = CaseHisto\_CsHiID
  + Concepts Case
    - Category = /EventAttribute/Observation/PostMortem/LookUp/CaseMacro\_CsFiID
    - Category = /EventAttribute/Observation/PostMortem/LookUp/CaseHisto\_CsHiID
  + Codes System
    - Category = /EventAttribute/Observation/PostMortem
    - Code = [system\_name]
    - Category = /EventAttribute/Observation/PostMortem
    - Code = [system\_name]Macro\_SyFiID
    - Category = /EventAttribute/Observation/PostMortem
    - Code = [system\_name]Histo\_SyHiID
  + Concepts Systems
    - Category = /EventAttribute/Observation/PostMortem/LookUp/[system\_name]Macro\_SyFiID
    - Category = /EventAttribute/Observation/PostMortem/LookUp/[system\_name]Histo\_SyHiID
* Assign concepts
  + Query for all defined Macro codes by event\_id orderd by system and organ
    - Only returns events that have Macro codes assigned.
  + Add attributes as appropriate
  + Ditto Histo

# Create RDV Modelling plan

* Issues
  + Straight values & adjust with one-hot encoding and z-score normalisation
  + 4 basic types
    - Ext, int1 (incl organ weights), int2 (incl Macro )and int3 (Incl Histo)
  + For Macro & Histo
    - Case summary vs System Summary
    - Can I include both?
  + Model tuning
    - Do I get best vales for parameters then use for all rdvs or do I have to find the best parameter values for each rdv?
    - If one-set of parameters which rdv do I use ext or int3
    - If int3 do I use case and
  + Added features
    - Season – should make no difference
    - Gestational\_age\_at\_birth
      * A lot of ‘nan ‘ values so reduces data set
      * Could be a good example for setting ‘nan’ values to -99
* Step 1
  + Do basic Decision Tree on all rdvs