# Thesis Ideas

**Thesis workflow**

**Data extraction**

* Explain the dataset: instances, features, which features are important (by literature and CFS)
* Explain the process to extract and clean the data. In particular, explain which variables of the dataset were used to work out interaction sequences.
* Explain any other processes performed on the data e.g. aggregation, de-aggregation
* Explain what variables were removed
* Note: relate discussion of the dataset to literature

**Data transformation**

* Explain how the chosen data variables were transformed/encoded into interaction sequences.
* Explain the **different encodings**, and their strengths and weaknesses e.g. time complexity of the different algorithms.

**Feature extraction**

* Explain how features are extracted from the interaction sequences, and relate the selection of these features back to literature.
* Explain the positives and negatives of **different features** and metrics.
* Perform an analysis on feature correlation.

**Prediction**

* Predict student performance with different metrics (e.g. final exam mark, final course mark, module mark, outcome).
* Use different prediction methods: clustering, Naive Bayes, Decision Trees etc. – look into different methods.
* **Other prediction tasks**
  + Using early behaviour to predict later behaviour
  + Predicting at risk behaviour
  + Think about other ML tasks

**Evaluation and comparison**

* Compare
  + Different encodings
  + Different features extracted
  + Different performance metrics
  + Different prediction methods
  + Different datasets and problems

**Note:** Any analysis by module can be aggregated later

# TIMELINE

**July**

* **6 July**
  + **Binary encoding (Thursday)** 
    - Extend the current analysis (distribution) to the rest of the module
    - Work out a method to extend the analysis to any number of individual slides (more flexible, regardless of structure of each model)
    - Extend the current analysis to the rest of the challenge
    - Compare results across modules and challenges
    - Gather better insights about the data, and record them down in a data insights journal (relate to Qs at the end)
  + **Different encodings (Friday)** 
    - Read about different encodings
    - Design 2-3 different encodings
  + **Extract outcomes (Saturday)** 
    - Extract different performance outcomes for students
    - Design and implement interaction features/metrics for the encoding.
    - Look at submission stats
  + **Prediction (Monday)** 
    - Set up the prediction process, and predict outcomes based on interaction sequences for the binary encoding example only. Do it for the individual problem, module and challenge and compare. ☐
* 13 July
  + **Encoding A (Thursday)** 
    - Implement the encoding for a given module and challenge.
    - Analyse the distribution of behaviours for the given encoding.
    - Design and implement interaction features/metrics for the encoding.
    - Set up the prediction process, and predict outcomes based on interaction sequences.
    - Compare results across modules and challenges
  + **Encoding B (Saturday)** 
    - Implement the encoding for a given module and challenge.
    - Analyse the distribution of behaviours for the given encoding.
    - Design and implement interaction features/metrics for the encoding.
    - Set up the prediction process, and predict outcomes based on interaction sequences.
    - Compare results across modules and challenges
  + **Encoding C (Monday)** 
    - Implement the encoding for a given module and challenge.
    - Analyse the distribution of behaviours for the given encoding.
    - Design and implement interaction features/metrics for the encoding.
    - Set up the prediction process, and predict outcomes based on interaction sequences.
    - Compare results across modules and challenges
* **20 July**
* **27 July**

**August**

* Bryn comes back
* Start write up

**September**

* Continue write up
* Finalise EVERYTHING

**October**

* Prepare seminar presentation
* Oct 17 – Seminar Day
* Finish write up

**November (THESIS DUE)**

* **Thesis due 6 November**

**Appendix**

**Encodings**

* An ideal binary encoding might look like 1 1 1 1 1 for 5 different slides indicating all slides were completed. However, someone with the interaction sequence 0 0 0 0 0 may also get high marks – they might have found the material easy
  + **Strengths:** Easy and simple to implement
  + **Weakness:** Lose information about the order in which slides are completed. Binary representation is a less rich representation of student slide interactions.
* **One example:** We might want to measure the **linearity/completion** of someone’s progress. Ideally, the most diligent students will complete the slides in order (e.g. Slide 0, 1, 2, 3, 4). Note, this is different to the order in which slides may be accessed/viewed.

**Metrics**

* **Normalised metric**
  + We could award a point for completion
    - Point for doing some of the activities
    - Point for completing the activities
    - Do a score per slide and then normalise against available points for the module
    - 1 – complete attempt at all the material in the module 0 – not attempting any slides

**Visualisations**

* You can plot a variety of interesting visualisations (different colours, different groups and clusters), visualisations that show movement over time
* Potential graph
  + Two dimensions: Completion vs. linearity
    - Could encode colours
    - And then plot how students move
    - Could do that across modules of the course and then make some observations at that level
    - Fewer dimensions – the easier it is to cluster

**Other questions**

* Other educational phenomenon to explore if there is time
  + Does engaging with slides improve outcome of the module?
  + Is there a change in how students engage with the slides (e.g. they may whiz through the material initially and then find it more difficult towards the end of semester e.g. 0 0 0 0 2 at the start)? Is slide engagement consistent?
  + What are the different groups/clusters of students?
    - Students who excel with little effort spent engaging with material
    - Students who excel with dedicated effort spent engaging with material (there may be a linear relationship here where the more effort you put in, the higher your results)
    - Students who do not do well at all despite engaging with the material.
    - Students who do not do well and spent little effort engaging with the material
* Considerations
  + Students get points if they pass all validation tests.
  + Points available to them drop if they make too many submissions
  + **Student motivation is increasing points, rather than learning and developing skill**