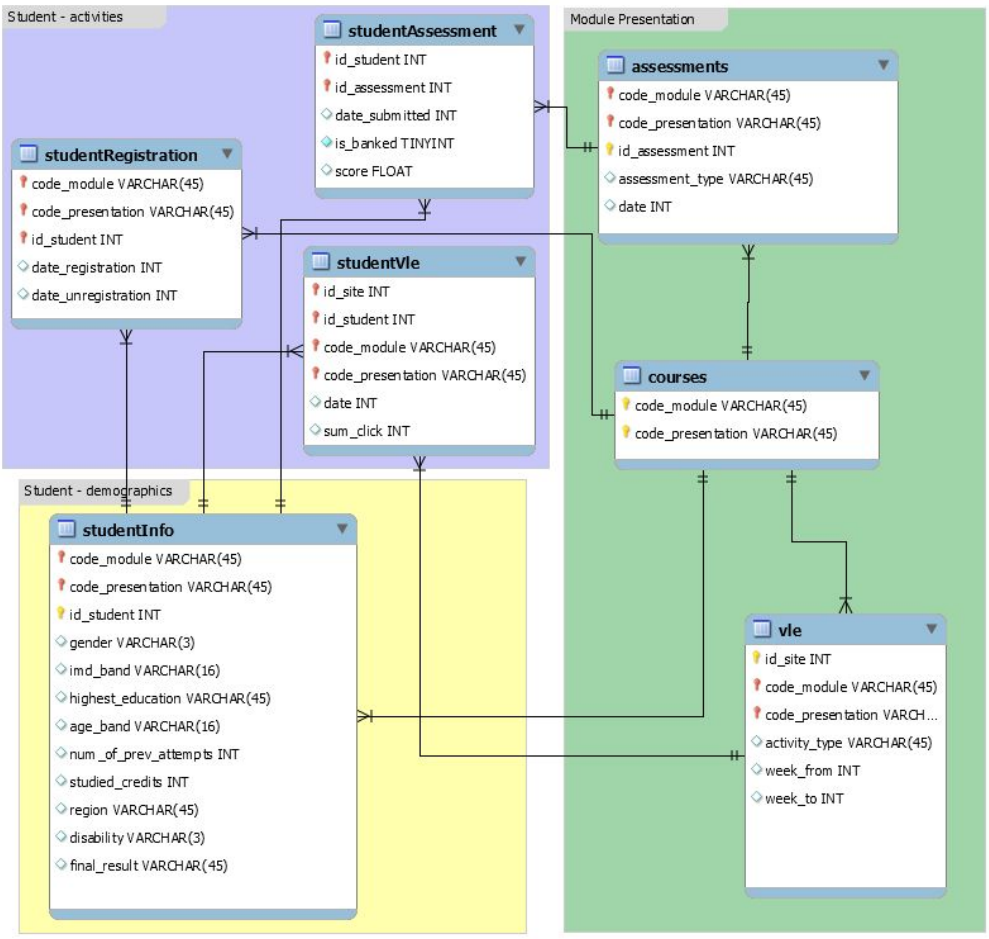
Reference:
https://github.com/feature-engineering-studio/Lecture-Slides/blob/master/HUDK5053-Lecture%206.pdf
Data collection submission: https://github.com/feature-engineering-studio/data-upload

Chad:

Project description/Logic model:
How does student engagement and resource access throughout the course affect overall student performance? My assumption is that student who interact within the course at a higher frequency tend to achieve higher scores on exams and overall performance throughout the term.

Description of data:
The modeled consists of multiple separate data sets that will need to be aggregated together. These data sets include student demographics, student interaction within the virtual learning environment (VLE), assessment information and performance, registration information, and course information.



Dataset that will be used for this project will be sourced from Open University, located here: https://analyse.kmi.open.ac.uk/open_dataset

Item level description of the data can be found here: https://github.com/cjc2238/Feature-Engineering-Project/blob/master/Data%20Upload%20Assignment/CopyOfDataset.md

Obstacle:
There is a data set that consists of student assessment performance on all three items (CMA, TMA, and final EXAM) but it doesn't seem to be that clean. Some EXAM items are incorrectly labaled as TMA/CMA and added as the last entry for that student.

The data source has this to say: "*If the information about the final exam date is missing, it is at the end of the last presentation week."*

So in order to use this assessment data I need to extract the last entry scores and convert them to the exam variable. This quick R code does exactly that:

Note ## This script feeds from this data transformation script first

```
library(plyr)
assessment_exam_df1 <- ddply(std_assessments_df, .(id_student), function(x) x[nrow(x), ])
```

The issue I am having is that some students skipped the exam, some failed, and some dropped out so the data frame this extracts may extract a CMA or TMA score. I need some third variable to determine if the score I am extracting is the final exam or not. I think including the weight or time may be the best option but a second opinion wouldn't hurt.

References:

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z. and Wolff, A. OU Analyse: Analysing At-Risk Students at The Open University. Learning Analytics Review, no. LAK15-1, March 2015, ISSN: 2057-7494.

Yaoli:
Project description/Logic model:
Educational goal/Assumption: if they pay attention(fixate on the right thing, e.g. might be semantically related) at certain point of time, they would understand the content better.
I'm developing a model that discovers students' eye fixation pattern on various video features regarding the speaker(speech-related gestures, facial expression...) and camera shooting that would predict their understanding of the video contents.

Description of data:

The data set will contain 28 subjects' 1) eye fixation data(duration and counts) over the time of watching 2 ted talk videos(each around 10 minutes long); 2)and their scoring on comprehension questions after watching the videos.

Obstacle:
I have coded video features but haven't coded how each subject's fixation is loaded onto these features over the video length. I might need to manually use a visualization tool to see if subjects' fixation x- y-coordinates maps onto visuals in the video.
And since the comprehension questions can be locked down to certain time frames/time points in the video, I'm thinking about doing the question-level model(score will be binary, right or wrong, and thus logistic) or the student-level model(score will be continuous from 0-42).

Ask me any questions if I did not clarify enough above.

Chad Response:

Instead of manually using a visualization to see where a subjects eyes are looking during the instruction, maybe you can try some EDM techniques to extract general trends/features instead.

Here are some articles that come to mind on how this can be done:

Q-Matrix Mining
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.9876&rep=rep1&type=pdf
Model Discovery
http://educationaldatamining.org/EDM2012/uploads/procs/Full_Papers/edm2012_full_10.pdf
LFA
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.7043&rep=rep1&type=pdf
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.9734&rep=rep1&type=pdf

Additionally, I think building a binary two-class logistic model would be the best way to go for now. Depending on how the information is displayed in the video, you could first build one model to detect what time frame and eye location best correlates with success on one item and then from there you can build the model to fit the other assessment items.