ITU Courses

Introduction

In every semester following its establishment, the IT University of Copenhagen (ITU) has collected course evaluations from its students. The results of these evaluations are publicly available on the ITU website and contain scores for the different courses available during that semester.

ITU also keeps course plans publicly available on its course base website, including information such as attendees, timeslots, instructors and rooms for different courses. The question is then if any of these attributes influence the overall evaluation of the course. For example:

- Does the lecturer(s) matter?
- Does the season matter? E.g. spring or autumn.
- Do courses late in the day score differently from early courses?
- Does the number of participants matter?
- Does the language matter?
- What type of room is best: labs, auditoriums or classrooms?
- Do job prospects or workload influence the overall evaluation?

Put more succinctly: **what constitutes a good course at ITU?** That is the subject of our report.

Plan of action

The first step was to scrape the ITU website to get the evaluation data and the course base data. We wrote web scrapers in Python to fetch the data. Once the data had been downloaded it was cleaned and merged into a single dataset.

We then created charts for selected courses to see how the ratings develop over time. This was to inform us whether the evaluations vary enough with changes in attributes meriting further studies.

Once we had a single dataset to work with, we ran various algorithms on it. One of the first things we did was run the Apriori algorithm on the dataset to explore the attributes and find association rules. These rules helped inform us of what attributes to consider for other parts of the project.

Next, we tried to cluster the evaluation data to see if there were any clear relationships between the overall evaluation, time evaluation and job relevance evaluation. These clusters helped us visualise trends in the data set and get an idea of how these attributes affected course ratings.

Finally, with our knowledge of the association rules and the clusters, we created a classifier to see whether we can predict if a given course would be evaluated as being either good or bad.

Ethics

There is one ethical problem worth commenting on for this report: since we directly compare evaluations with course data such as lecturer names, certain lecturers are at risk for being exposed as being “bad” in this report.

It could be argued that since the data is publicly available and since we are using well-known algorithms, anyone could potentially find the results, making any censorship unnecessary. Furthermore, the findings could be used both as a way to hurt the credibility/reputation of a lecturer, but also as a way to commend good lecturers.

However, in the interest of keeping the peace and because we do not wish to elevate our research as being ground truth, we have omitted names of lecturers that came up during the mining process.

Pre-Processing

Scraping

ITU provides all evaluation answers for a given semester as a .csv files containing response rates and evaluations for each question.

The university has changed the set of questions they ask students at evaluations every few semesters. This means that even though some questions might be useful for mining, they do not occur in enough semesters to be used for analysis. Analysis of the data set showed that only two questions were recurring every semester from 2005 to 2014 - namely *“I think the course is relevant for my future job profile”* and *“My time consumption for this course is too high”*. These attributes - along with the average course rating, percentage of respondents, course name and study programme - were extracted from the .csv files.

Course data is not provided by ITU in an easily readable format. Information about courses from 2005 to 2014 then had to be scraped from the online course base. The course attributes varies by semester so only some of the data could be used. Some course data was also written into the course base by lecturers, which meant that this data was not structured in a uniform way and was omitted.

We decided to scrape the course name, the semester the course is in, the line of studies, ECTS points, language, minimum, expected and maximum number of participants, lecturers and time slots for lectures and exercise sessions. The data simply varied too much or was too unpredictable to calculate any sort of mean variable for the varying attributes.

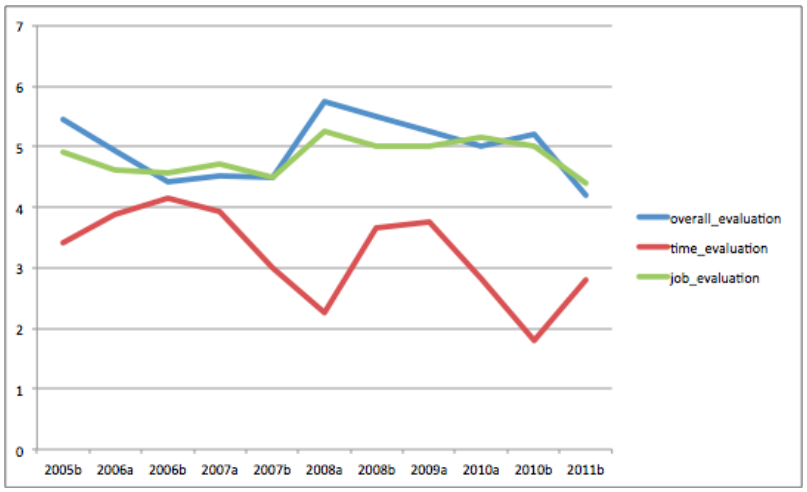
Finally, the two data sources were joined by course name, semester and line of studies into a single dataset.

Visualisation of Evaluations

Note that a high score for **time_evaluation** indicates that the student does not feel they have *enough* time!

- **overall_evaluation** is the score given overall for the course (higher = better).
- **time_evaluation** refers to whether or not the students feel that they do not have enough time for the course (higher = less time).
- **job_evaluation** refers to how relevant students feel that the course is for their career (higher = more relevant).

Semesters have been split into spring semesters - designated by an “a” and fall semesters designated by a “b”.



Design of User Interfaces and Data

The course “Design of User Interfaces and Data” is very stable overall. It has been taught by same lecturer for all semesters except the first two in our dataset and the overall evaluation is relatively stable.

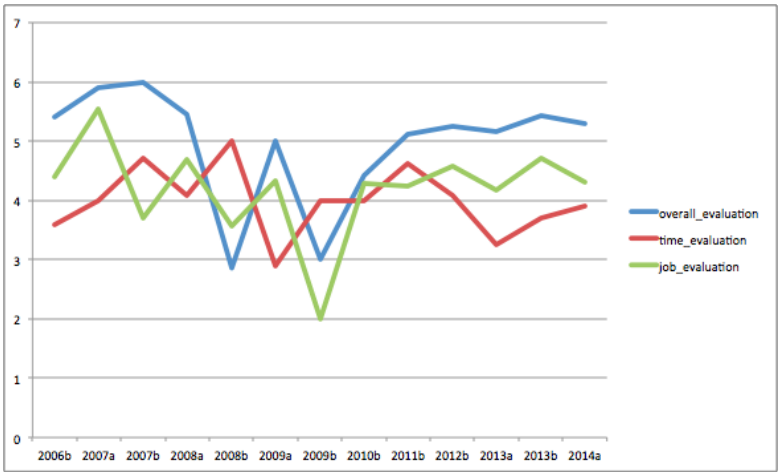
It is interesting to note that job_evaluation generally follows overall_evaluation and that the high points of overall_evaluation are when there is a low time_evaluation.

The course is taught in Danish in all semesters except for 2007a. The high point is in 2008 during spring and the low point is in 2011 during fall. Interestingly, there are 40 participants during every semester, except for the low point where there is only 12.

Introductory 3D

Now for a course which varies drastically in evaluation during semesters.

The first semesters of “Introductory 3D” are the best rated, but in the fall of 2008 a new lecturer is introduced who teaches in English and



the overall_evaluation plummets. In the spring of 2009 the lecturer is replaced by one of the original lecturers who switches back to teaching in Danish and the ratings go up again. They go down in the next semester when the “new” lecturer comes back, who still teaches in English. The lecturer then improves slightly for the following semester only to be replaced by a new lecturer in the remaining semesters where the overall_evaluation steadily improves. English is kept as a course language.

Interestingly, the time_evaluation is *not* a clear mirror image of the overall_evaluation as in the previous example.

Analysis of Selected Course Data

What these examples show is that course evaluations are likely most affected by who teaches them, yet other variables can also determine whether the overall_evaluation shifts by a point on the scale.

These other variables might of course be exogenous to some extent. Even in these two examples there are certain patterns that exist in one graph that contradict a pattern in the other graph, e.g. the time_evaluation mirroring the overall_evaluation.

It is important to note that we are merely making educated observations based on statistical data.

Association Rule Mining

We decided to do association rule mining using Apriori as we have a good understanding of how the algorithm works along with its limitations by having implemented it ourselves.

Preparation of data

The attributes were modified slightly before running the Apriori algorithm. Most importantly, course evaluations were converted from numeric values into ordinal labels as either “evaluation:good” and “evaluation:bad” split around the median. This allowed us to differentiate good courses from bad courses.

Other attributes were used in the following way:

- The number of participants was split around the median into the ordinal labels “participants:low” and “participants:high”
- Semester, language, programme was represented as nominal labels
- Rooms were represented as label combinations of room type and module type (e.g. “aud:Forelæsning”)
- Lecturers were represented as “lecturer:<name>”
- Timeslots were labeled after starting time into “time:early” (< 10), “time:mid” and “time:late” (>= 16)

We chose not to include other evaluation scores for this part, as we wanted to focus on course attributes in comparison with the overall evaluation. The data was represented as sets of labels and fed into our own implementation of the Apriori algorithm.

Results and analysis

Association rules were mined with a minimum support of 1% and a minimum confidence of 60% with a Lift of more than 1. Only frequent closed patterns were considered for mining association rules and we only looked at patterns with 2 or 1 items on the “left” side. This was done in order to limit the amount of frequent patterns to consider for the analysis, as patterns that are very specific are decidedly less relevant for deciding which courses to pick.

In the end, the following interesting rules were found (full list of rules in Appendix a):

General trends

- **Lectures** and **exercises** taking place in **auditoriums** generally imply a **bad** evaluation, even with a **low** number of participants!
- **Auditorium lectures** are particularly **bad** in the **morning**, while the **exercises** are particularly **bad later** in the day
- **Classroom lectures** and **exercises** are generally preferred to **auditoriums**, especially when there are **few** participants.
- **Lab classes** and **exercises** that take place late in the day are popular.
- **Lab classes** in **Danish** are also popular.
- Courses that take place **late** in the day in the **Spring** semester imply a **good** evaluation.
- Courses that are worth **15 ECTS** and are either scheduled **late** in the day, have **classrooms lectures** or **few** participants are generally good.
- Courses worth **7.5 ECTS** are **popular** when the lectures are in **labs** or if the instruction is in **Danish**
- Courses in **Danish** are generally evaluated well in the **fall**.

More specific results

- **KDDK** courses that have **lectures** in **labs** are **popular**.
- **KDDK** courses are also **popular** when in **classrooms**, **late** in the day or in the **Spring**.
- **KSWU** courses are **good** at **7.5 ECTS.**
- **MIND** courses are **good** when participants are **low.**
- **MINM** courses are **good** late in the **day**.
- **KSDT** courses are **bad** when scheduled **late** in the day either at **7.5 ECTS**, in the **fall** or when participants are **high**.
- **BDMD** courses are **bad** when not scheduled **early** or **late**.
- Lecturer **A** is **popular** when the instruction is in **Danish**.
- Lecturer **B** and **C** are **popular** when the instruction is in **English**.
- Lecturer **D** is popular with **7.5 ECTS** classes.
- Lecturer **X** is not **popular** when the classes take place **early** or **late**.
- Lecturer **Y** is not **popular** when instructing in **English**.

From this analysis of the rules it would seem that better courses are much easier to find in **KDDK** compared to **KSDT**. A few lecturers are worth avoiding and a few worth considering.

Generally, we can say that having courses in the **Spring** semester, in **Danish**, with **few** participants, a **late** start time or a **classroom** setting are often indicators of a good course at ITU.

Clustering

The evaluation data was examined using the K-Means clustering algorithm to see if any correlation could be found between the different evaluations. Evaluations are given on a scale from one through six, and are therefore normalised.

Several clustering algorithms were tested and the resulting clusters visualised to see if significantly different clusters could be found. The dataset was very dense with the majority of data points being in a large group rated above average.

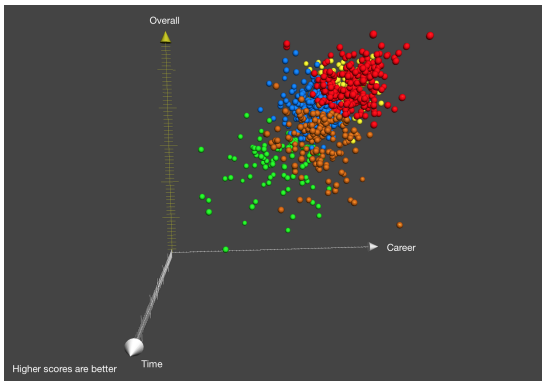
Using K-Means

The choice for using K-means came mostly from the fact that this was a well-known algorithm to us. Therefore the results were mapped using K-Means clustering with k=5 which seemed to produce the most interesting result.

Hypotheses

We wanted to test several hypotheses when looking at the found clusters:

- Time-intensive courses have overall bad ratings.
- Career-relevant courses have overall high ratings.
- Time-intensive courses with high career relevance generally have high ratings.
- Courses with a light workload result in better ratings.
- The percentage of students evaluating a course influence its overall evaluation.



Results

As mentioned earlier, most courses are in a cluster with above-average overall evaluations. The result of the clustering can be seen in the diagram in appendix C.

The green cluster (93 courses) in the diagram shows the lower-rated courses, where students feel that the course is somewhat to not at all related to their career. The time evaluations for these courses range from the students feeling that they have too little time to feeling that they are not busy at all in the course.

The main cluster can be split into four larger clusters. The lower, orange cluster (153 courses) is comprised of courses where the students feel that a course is mostly relevant to their career, but they generally have too little time for the course.

The blue cluster (232 courses) comprises courses where students feel that they have average-to-above-average time to keep up with the course, and feel that the course is relevant for their career.

The red (248 courses) and yellow clusters (165 courses) both have good overall evaluations and high career relevance. The clusters differ in the way that students feel that the courses in the yellow cluster are less time consuming compared to the red cluster.

Analysis of Results

Looking at the clusters from the career evaluation axis *(see appendix C.2)* it becomes apparent that there is a correlation between the overall course evaluation and the career relevance of a course is relevant for their career. The higher the career-relevance the higher the overall evaluation; this is in line with our cursory analysis of selected courses. It seems that the overall evaluation of courses with high career relevance is not affected by having a light workload *(see appendix C.3).*

Most courses rated above average have a low workload, but at the also have a good career rating which would affect the overall evaluation. There seems to be a slight clustering of courses with a higher than average workload scoring lower than courses with a lower one, which indicates that the workload does influence the overall evaluation slightly.

The yellow cluster would seem to be ideal courses. Students rate them as being career-relevant, overall good and the workload is not too big.

Classification

For classification we have chosen our own implementation of KNN. KNN is not the most efficient classification algorithm, but it works for the relatively small dataset we have and having implemented it ourselves, we feel confident in its results.

In order to use a classification method to help answer our research question, we run the KNN algorithm 10 times for each k in {1, 3, 5} for each possible variation of attributes and continuously shuffle the dataset and partition it into 90% training data and 10% test data. We analyse the same attributes of the dataset as when we calculated association rules.

This gives us a large set of predictions from which to calculate what combination of **k** and set of **attributes** will statistically provide the best prediction. The top five results are:

> 65.73% --> {'lecturers', 'participants', 'language', 'rooms', 'programme'}, 1
> 64.04% --> {'lecturers', 'ects', 'overall', 'language', 'programme'}, 5
> 63.71% --> {'lecturers', 'ects', 'participants', 'overall', 'language'}, 5
> 63.60% --> {'ects', 'times', 'rooms', 'programme', 'language', 'lecturers'}, 5
> 63.48% --> {'lecturers', 'ects', 'participants', 'overall'}, 3

The results show that a combination of the attributes **lecturers**, **participants**, **language**, **rooms**, and **programme** with k=1 predicts with an accuracy of more than 65% whether a course is good or bad.

The addition of **programme** - an attribute not given much thought previously - appears to be very important for getting a good result. Interestingly, in the two runner-ups, **ects** also

made a difference. The attribute **overall** can be ignored in the sets of attributes where it appears, as it was not used for calculating distance.

The five worst results were:

47.07, --> {'ects', 'participants', 'semester', 'overall'}, 1
46.62, --> {'participants', 'semester'}, 5
46.40, --> {'participants'}, 5
46.06, --> {'semester', 'language', 'programme'}, 5
45.50, --> {'ects', 'semester', 'programme'}, 5

Curiously, missing from these results are **lecturers** and **rooms**, indicating how unimportant these attributes are for accurate predictions.

We only ran 10 tests per combination since running 100 tests or 1000 tests - potentially providing more accurate results - would take a full day on the hardware we had available.

Conclusion

Both our association rules and our classifier results indicate that **rooms**, **language**, and **participants** are key to predicting whether a course is good or bad. We know the preferred choices here are **classrooms**, **Danish**, and **few participants**. While the association rules did not generate a lot of rules with lecturers in them - presumably because the support was too low - our own analysis of selected courses and the classifier results indicate that lecturers are very important for predicting evaluation scores.

Whether students feel that the curriculum of a course is relevant for their **career** seems to affect overall course ratings drastically, as shown in our clusters. The **workload** of a course seems to have a minor say in the overall evaluations of courses. The top rated courses were also those most relevant for the future career of the students.

Time did not seem to be as important in our top classifier results as it was shown to be in the association rules and while the **semester** seemed to be a good indicator in our association rules, it did not appear in the top results in our classifier tests either. Instead, the “bureaucratic” variable **programme** seems to make a significant difference and we did in fact consistently mine rules with some programmes indicating a good evaluation and some indicating a bad.

Perhaps this implies differing mentalities between students of different programmes and different levels of study - e.g. “Master”, “Kandidat”, and “Bachelor”? We would have liked to include the different student bodies for each course in this dataset. This would allow us to get an accurate representation of what share of the students came from what programme, but unfortunately this data was not available for every semester in the already limited dataset. We chose *not* to include it in order to have a more sizable dataset.

Future research might also include a keyword analysis of the course description and title in order to discover whether certain subjects are more likely to receive good evaluations.

Appendix

Appendix A: Association Rules

{classroom:Forelæsning, ects:1500} => {overall:good}
{classroom:Forelæsning, language:Dansk} => {overall:good}
{classroom:Forelæsning, participants:low} => {overall:good}
{classroom:Forelæsning, programme:KDDK} => {overall:good}
{classroom:Forelæsning, time:late} => {overall:good}
{classroom:Øvelser, programme:KDDK} => {overall:good}
{classroom:Øvelser, time:late} => {overall:good}
{ects:1500, participants:low} => {overall:good}
{ects:1500, time:late} => {overall:good}
{ects:750, lab:Forelæsning} => {overall:good}
{ects:750, language:Dansk} => {overall:good}
{ects:750, lecturer:D} => {overall:good}
{ects:750, programme:KSWU} => {overall:good}
{lab:Forelæsning, language:Dansk} => {overall:good}
{lab:Forelæsning, programme:KDDK} => {overall:good}
{lab:Forelæsning, time:late} => {overall:good}
{lab:Øvelser, time:late} => {overall:good}
{language:Dansk, lecturer:A} => {overall:good}
{language:Dansk, participants:low} => {overall:good}
{language:Dansk, programme:KDDK} => {overall:good}
{language:Dansk, semester:Efterår} => {overall:good}
{language:Dansk, time:late} => {overall:good}
{language:Engelsk, lecturer:B} => {overall:good}
{language:Engelsk, lecturer:C} => {overall:good}
{participants:low, programme:KDDK} => {overall:good}
{participants:low, programme:MIND} => {overall:good}
{participants:low, semester:Forår} => {overall:good}
{programme:KDDK, semester:Forår} => {overall:good}
{programme:KDDK, time:late} => {overall:good}
{programme:MINM, time:late} => {overall:good}
{semester:Forår, time:late} => {overall:good}

{aud:Forelæsning, aud:Øvelser} => {overall:bad}
{aud:Forelæsning, ects:750} => {overall:bad}
{aud:Forelæsning, language:Engelsk} => {overall:bad}
{aud:Forelæsning, participants:high} => {overall:bad}
{aud:Forelæsning, participants:low} => {overall:bad}
{aud:Forelæsning, semester:Efterår} => {overall:bad}
{aud:Forelæsning, semester:Forår} => {overall:bad}
{aud:Forelæsning, time:early} => {overall:bad}
{aud:Forelæsning, time:mid} => {overall:bad}
{aud:Forelæsning} => {overall:bad}

{aud:Øvelser, language:Engelsk} => {overall:bad}
{aud:Øvelser, participants:high} => {overall:bad}
{aud:Øvelser, participants:low} => {overall:bad}
{aud:Øvelser, semester:Efterår} => {overall:bad}
{aud:Øvelser, semester:Forår} => {overall:bad}
{aud:Øvelser, time:late} => {overall:bad}
{aud:Øvelser, time:mid} => {overall:bad}
{aud:Øvelser} => {overall:bad}
{ects:750, programme:KSDT} => {overall:bad}
{language:Engelsk, lecturer:Y} => {overall:bad}
{lecturer:X, time:mid} => {overall:bad}
{participants:high, programme:KSDT} => {overall:bad}
{programme:BDMD, time:mid} => {overall:bad}
{programme:KSDT, semester:Efterår} => {overall:bad}
{programme:KSDT, time:late} => {overall:bad}

Appendix B: Predictions

65.73033707865168, --> {'lecturers', 'participants', 'language', 'rooms', 'programme'}, 1
64.04494382022472, --> {'lecturers', 'ects', 'overall', 'language', 'programme'}, 5
63.70786516853933, --> {'lecturers', 'ects', 'participants', 'overall', 'language'}, 5
63.59550561797754, --> {'ects', 'times', 'rooms', 'programme', 'language', 'lecturers'}, 5
63.48314606741573, --> {'lecturers', 'ects', 'participants', 'overall'}, 3
63.37078651685394, --> {'times', 'lecturers', 'overall', 'rooms', 'programme'}, 1
63.37078651685393, --> {'ects', 'times', 'rooms', 'programme', 'participants', 'semester', 'lecturers', 'overall'}, 3
63.25842696629214, --> {'lecturers', 'participants', 'language', 'rooms', 'programme'}, 3
63.25842696629214, --> {'times', 'rooms', 'programme', 'semester', 'language', 'lecturers', 'overall'}, 5
63.146067415730345, --> {'lecturers', 'ects', 'language', 'rooms', 'programme'}, 3
63.146067415730345, --> {'ects', 'times', 'rooms', 'programme', 'language', 'semester'}, 3
63.03370786516854, --> {'rooms', 'programme', 'language', 'semester', 'lecturers', 'overall'}, 1
62.92134831460674, --> {'times', 'rooms', 'programme', 'participants', 'lecturers', 'overall'}, 5
62.58426966292134, --> {'ects', 'rooms', 'times', 'language', 'semester', 'lecturers'}, 3
62.58426966292134, --> {'rooms', 'programme', 'language', 'semester', 'lecturers', 'overall'}, 5
62.35955056179774, --> {'lecturers', 'participants', 'overall', 'rooms', 'programme'}, 5
62.35955056179774, --> {'ects', 'times', 'rooms', 'programme', 'language', 'lecturers', 'overall'}, 3

(... skipped about 1500 results)

49.10112359550562, --> {'ects', 'participants', 'semester', 'times'}, 5
49.10112359550562, --> {'ects', 'semester', 'times', 'overall'}, 1
48.98876404494382, --> {'overall', 'times'}, 1
48.98876404494382, --> {'overall', 'language'}, 5
48.98876404494382, --> {'ects', 'semester', 'times'}, 1
48.98876404494382, --> {'participants', 'overall', 'language'}, 5
48.988764044943816, --> {'language', 'semester', 'times', 'overall'}, 3
48.87640449438202, --> {'ects', 'overall', 'language', 'programme'}, 1
48.76404494382022, --> {'times'}, 3
48.651685393258415, --> {'ects', 'participants', 'semester'}, 1
48.42696629213483, --> {'ects', 'overall', 'rooms'}, 5
48.42696629213483, --> {'ects', 'participants', 'overall', 'times'}, 1
48.20224719101125, --> {'semester', 'overall'}, 5
48.20224719101123, --> {'semester', 'times'}, 1
48.20224719101123, --> {'participants', 'semester', 'language', 'programme'}, 1
48.08988764044943, --> {'participants', 'semester', 'rooms', 'overall'}, 1
47.64044943820225, --> {'participants', 'semester', 'overall', 'language'}, 3
47.64044943820224, --> {'participants', 'overall', 'language'}, 1

47.30337078651685, --> {'ects'}, 5
47.07865168539326, --> {'ects', 'participants', 'semester', 'overall'}, 1
46.62921348314607, --> {'participants', 'semester'}, 5
46.40449438202247, --> {'participants'}, 5
46.06741573033709, --> {'semester', 'language', 'programme'}, 5
45.50561797752809, --> {'ects', 'semester', 'programme'}, 5

Appendix C: Clustering

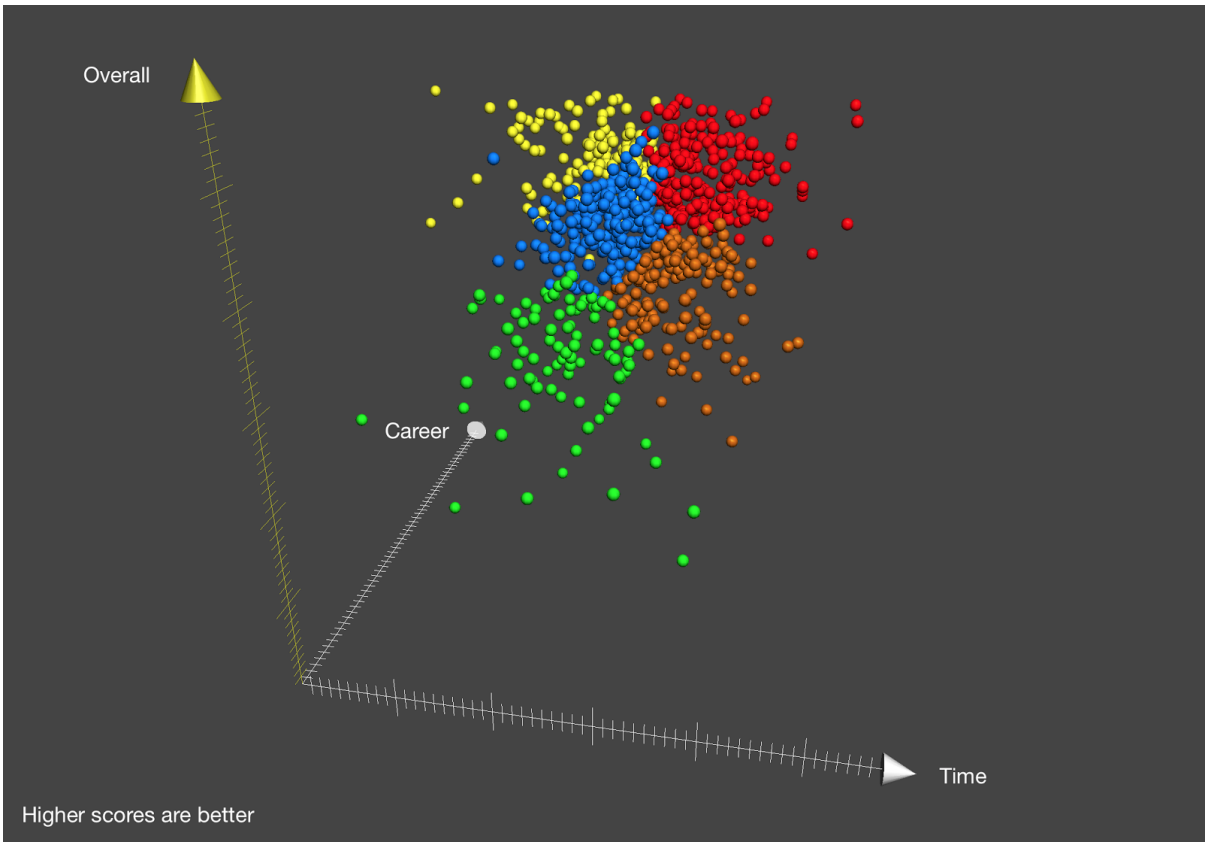


Fig. C.1
Visualised clusters of course evaluations.

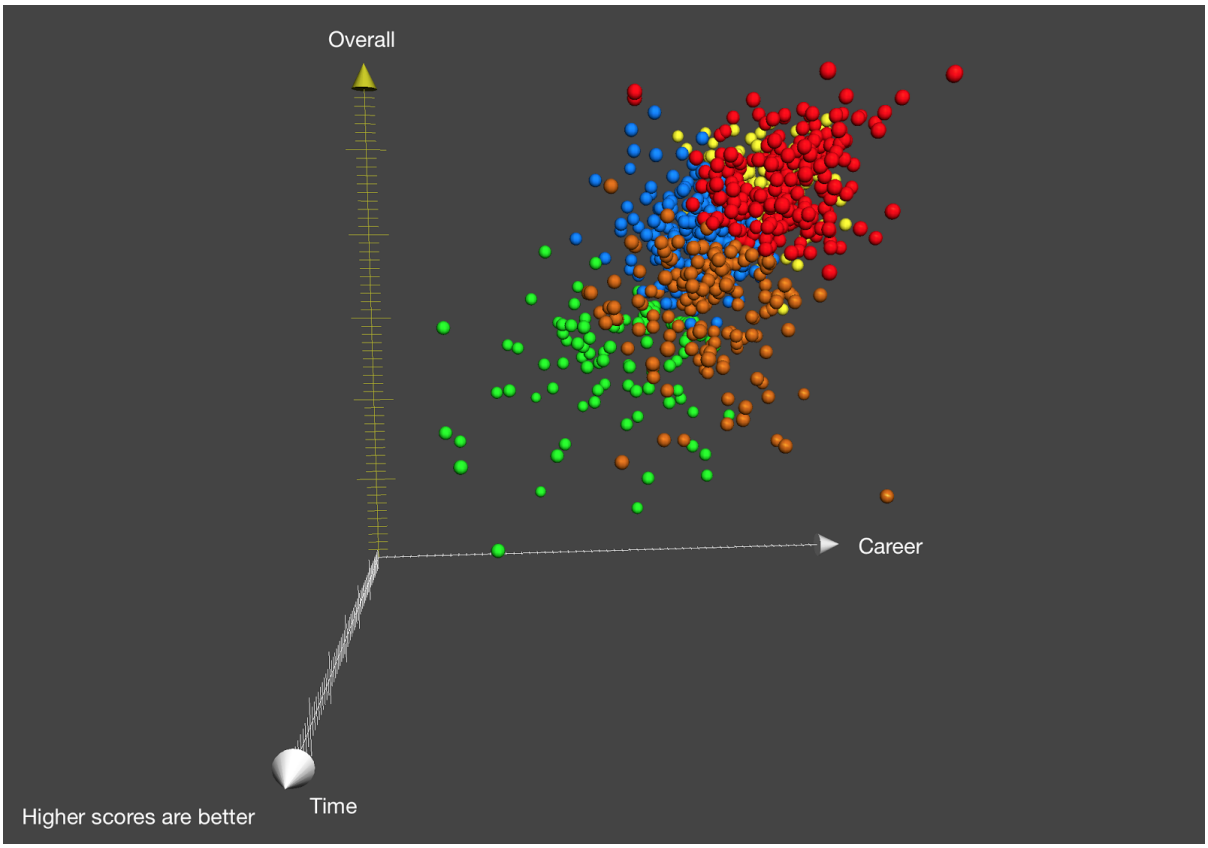


Fig. C.2
Visualised clusters of course evaluations
viewed along the “Career Evaluation” axis.

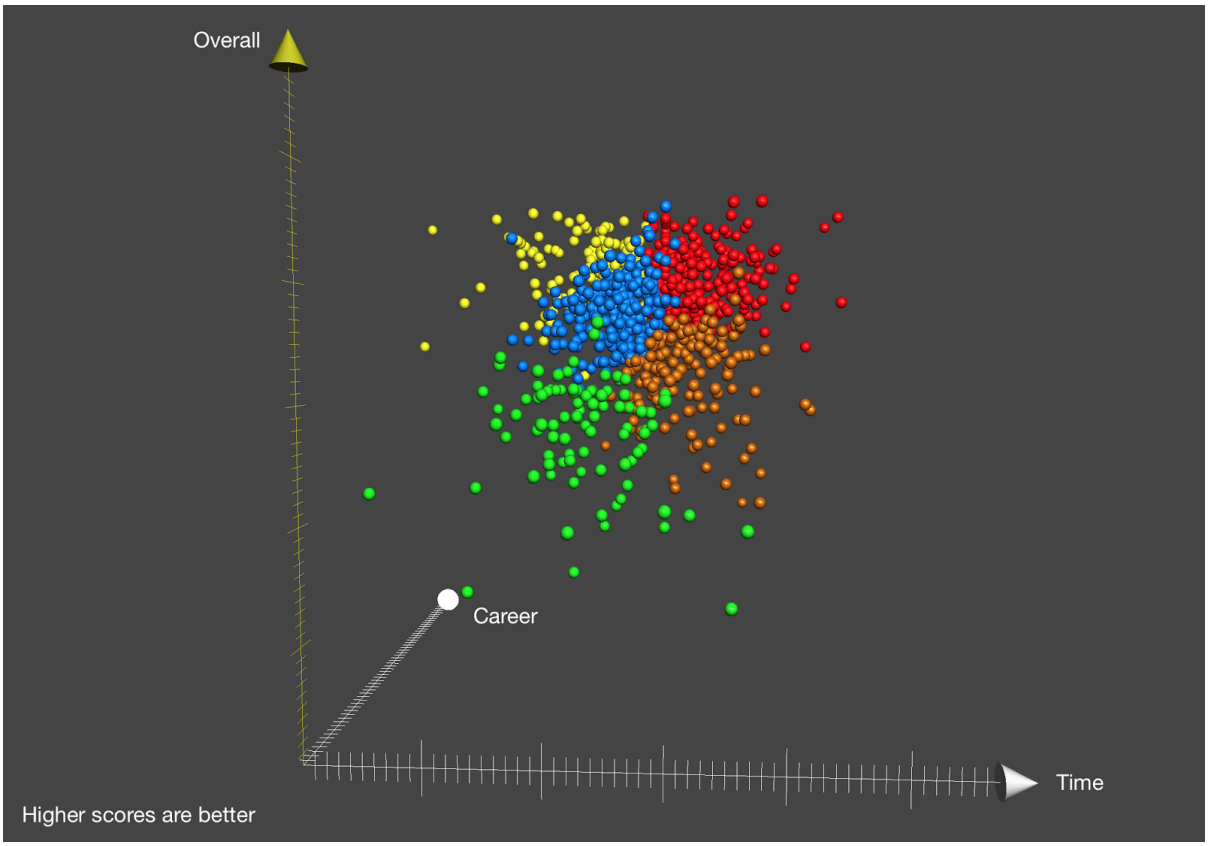


Fig. C.3
Visualised clusters of course evaluations
viewed along the “Time Evaluation” axis.