EUROPEAN
DATA SCIENCE
ACADEMY

Project acronym:	**EDSA**
Project full name:	**European Data Science Academy**
Grant agreement no:	**643937**

D3.5 Report on the Evaluation of Course Content and Delivery 2

Deliverable Editor:	**Aba-Sah Dadzie (OU)**
Other contributors:	**Inna Novalija (JSI)**
	Joos Buijs (TU/e)
Deliverable Reviewers:	**Angi Voss (Fraunhofer), Dave Tarrant (ODI)**

Deliverable due date:	**31/01/2018**
Submission date:	**19/01/2018**
Distribution level:	**P**
Version:	**1.0**

This document is part of a research project funded
by the Horizon 2020 Framework Programme of the European Union

Change Log

Version	Date	Amended by	Changes
0.1	17/11/2017	Aba-Sah Dadzie	ToC & analysis of EDSA LRM
0.1	03/01/2018	Aba-Sah Dadzie	review version
0.2	09-11/01/2018	ALL	addressing review comments and finalising deliverable
0.3	11/01/2018	Aba-Sah Dadzie	restructuring & consolidation updating review forms
0.4	17/01/2018	Aba-Sah Dadzie	submission version
1.0	19/01/2018	Alexander Mikroyannidis	Final QA

Table of Contents

List of Tables

List of Figures

1. Executive Summary

This deliverable concludes work in T3.4 and learning analytics in the EDSA project. It continues on from the initial analysis reported in D3.3 on online courses. This deliverable examines two datasets - the JSI VideoLectures data and the TU/e FutureLearn MOOC on Process Mining. As in D3.3 the analysis looks at the delivery of material and interaction of students with this material, as part of work to identify how this impacts on student outcomes.

Three approaches were used: process mining, statistical and visual analysis, to allow different perspectives on the data, and increase ability to gain insight and identify and explore interest patterns in the data. This deliverable takes a look also at user demographics, to contribute to work to identify mappings between student backgrounds, interaction with online learning material and outcomes. At the end of the project, mainly due to limited access to student demographics, this remains work in progress.

To support both the analysis process and the presentation of results web-based and desktop visualisation tools were developed for the VideoLectures and FutureLearn datasets. Additionally, a RapidMiner workflow was built to provide a more generalised framework for learning analytics, based on project requirements. This workflow has been tested with FutureLearn data but is easily extended to other datasets and platforms.

The EDSA project in its second phase no longer directly delivers courses. Links between other project partner courses and third party online material and the EDSA project are examined, to identify synergy with other related work. The information obtained here will ultimately feed into categorising and branding online learning courses that are EDSA-recommended.

The aim of Learning Analytics in WP3 was to feed into evidence-based best practice to guide (re)design of course content and delivery and programme curricula in Data Science, to aid where necessary the tailoring of learning material and course delivery to fit particular contexts. We envisage our findings will contribute to guiding instructors in course presentation and delivery, in order to better support students in selecting courses that meet their requirements for self-improvement and as part of the processes of skill training and job-seeking.

2. Introduction

The objective of WP3 in EDSA is to:

1. Deploy the course material developed in (or in relation to) WP2 for different target groups and in different environments, comprising webinars, video lectures and face-to-face training.
2. Gather feedback about the effectiveness of learning from these courses.
3. Analyse feedback and other data generated during course delivery, to feed into improving content and/or form of deployment, as well as into the design of new courses.

Work Package 3 (WP3) produced two series of deliverables. The first series addressed objectives 1 and 3 and the second series objectives 2 and 3.

This deliverable follows on from D3.3, the first in the second set of deliverables, within Task 3.4 (T3.4). D3.3 reported on analysis of the first set of online courses delivered by the EDSA project and independently by project partners on topics falling within EDSA's remit. The aim, as with this deliverable, was to obtain overviews of the data, and carry out detailed analysis of content, in order to obtain a picture of student behaviour across all and in different courses and course types. This information was to serve two purposes - feedback to students to improve performance, and to instructors and course designers, to improve the content and presentation of a varied list of courses attended by an even more varied set of students.

D3.5, as the final deliverable for this workpackage, presents first an analysis of the course data for the second phase of the project. One key deviation from the original proposal is that EDSA no longer delivers project-specific courses, but rather provides a portal to EDSA-approved/accredited/recommended courses, delivered by project partners and other third party institutions. D3.5 examines data from the JSI VideoLectures portal, as work supported by EDSA, and the TU/e FutureLearn MOOC on Process Mining. This deliverable focuses on approaches to extracting content and information on interaction, to point to further study to meet the overall project goals. With limited access to student demographics we focus in this deliverable on analysis that allows us to identify potential areas in which to further exploration in line with both the project aims and the field of learning analytics.

Three analytical approaches were employed: statistical analysis, visual analysis and process mining. We triangulate results obtained from independent analysis of each dataset, to obtain broader reaching and more reliable conclusions on course presentation, student interaction and behaviour and the impact seen on student outcomes. Based on the analysis carried out, we revisit the proposal outlined in the D3.3 for a structured *Learning Analytics Framework* for EDSA, and present work toward achieving this goal — a workflow that may be extended to build on the outcomes obtained at the conclusion of the project.

2.1 Outline

Section 3 provides descriptions of the two event datasets analysed in this deliverable, licenses for (restricted) reuse and data publication plans as applicable, for use within and beyond the EDSA project. Section 4 details the analysis carried out for the two datasets, using one or more of three approaches: statistical analysis, visual exploratory analysis and process mining. The deliverable continues with the outcomes of work to build a framework for LA in EDSA (section 5). We conclude in section 6 with an outlook for continued research in learning analytics beyond the EDSA project.

3. Online Learning Event Datasets

Two datasets, the JSI VideoLectures data and a FutureLearn MOOC run by TU/e, are analysed. Statistical analysis, visual analytics and process mining are used to analyse data content, to extract information on course content and student interaction. A key aim is to provide data to support the generation and delivery of feedback to instructors, to feed into improving course content and delivery, and to foster engagement and aid knowledge acquisition through improved student interaction with online courses.

We summarise the datasets and detail availability for reuse in this section, and in section 4 detail the analysis carried out.

3.1 JSI VideoLectures Data

VideoLectures.NET is an open access educational video lectures repository that contains lectures recorded on video from different scientific events such as conferences, summer schools and workshops.

The Data Science category, introduced as part of the EDSA project, contains over 11,500 lectures and tutorials. EDSA deliverable D3.4 describes the current status of VideoLectures in the Data Science area and provides a set of sample videos.

In order to obtain insight into how viewers interact with Data Science VideoLectures, we have performed an analysis of the VideoLectures Apache log files from 2015 to 2017.

Figure 1 presents an example of a VideoLectures Apache log file, which contains information about viewer IP, access date and time, Request, Browser, Response, Bytes sent and Referrer.

```
IP1 - - [22/Aug/2016:17:07:53 +0200] "GET /iswc08_heat HTTP/1.0" 301 - "-" "Mozilla/5.0
(compatible; Googlebot/2.1; +http://www.google.com/bot.html)"

IP1 - - [22/Aug/2016:17:07:53 +0200] "GET /mit9000s11_gabrieli_lec20/ HTTP/1.0" 200 - "-"
"Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0;  Trident/5.0)"

IP2 - - [22/Aug/2016:17:07:53 +0200] "GET /site/translectures/ipsa09_berndtson_spsdd---1
HTTP/1.0" 404 - "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.htm

IP1 - - [22/Aug/2016:17:07:54 +0200] "GET /iswc08_heat/ HTTP/1.0" 404 - "-" "Mozilla/5.0
(compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
```

Figure 1: Example of VideoLectures Apache log (viewers' IPs are anonymized)

VideoLectures logs from 2015-17 present the most interest to the EDSA project; analysing these logs allows us to answer, e.g., 'what are the topics of interest for users in data science areas' and 'how many people came to the portal from the EDSA website'.

3.1.1 Data Publication Plan

The data publication plan for VideoLectures includes the publication of some limited detail and summary statistics. The number of views is presented with each lecture at the portal, while EDSA deliverable D3.4 and the analysis that follows present more insight into the information obtained from the data and logs.

Summaries of the VideoLectures data are available for download[1], with a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 license, in line also with requirements for reuse in EDSA.

The VideoLectures log data cannot be shared due to privacy issues.

3.2 TU/e FutureLearn Data

TU/e developed three MOOCs in total, two of which run on the UK-based FutureLearn platform[2], the EDSA-preferred MOOC platform. As discussed in D3.4, the two FutureLearn MOOCs attracted in total 4340 students since July 2016, with the last MOOC only launched in August 2017.

FutureLearn provides MOOC educators with several datasets in CSV format, as shown in Figure 2.

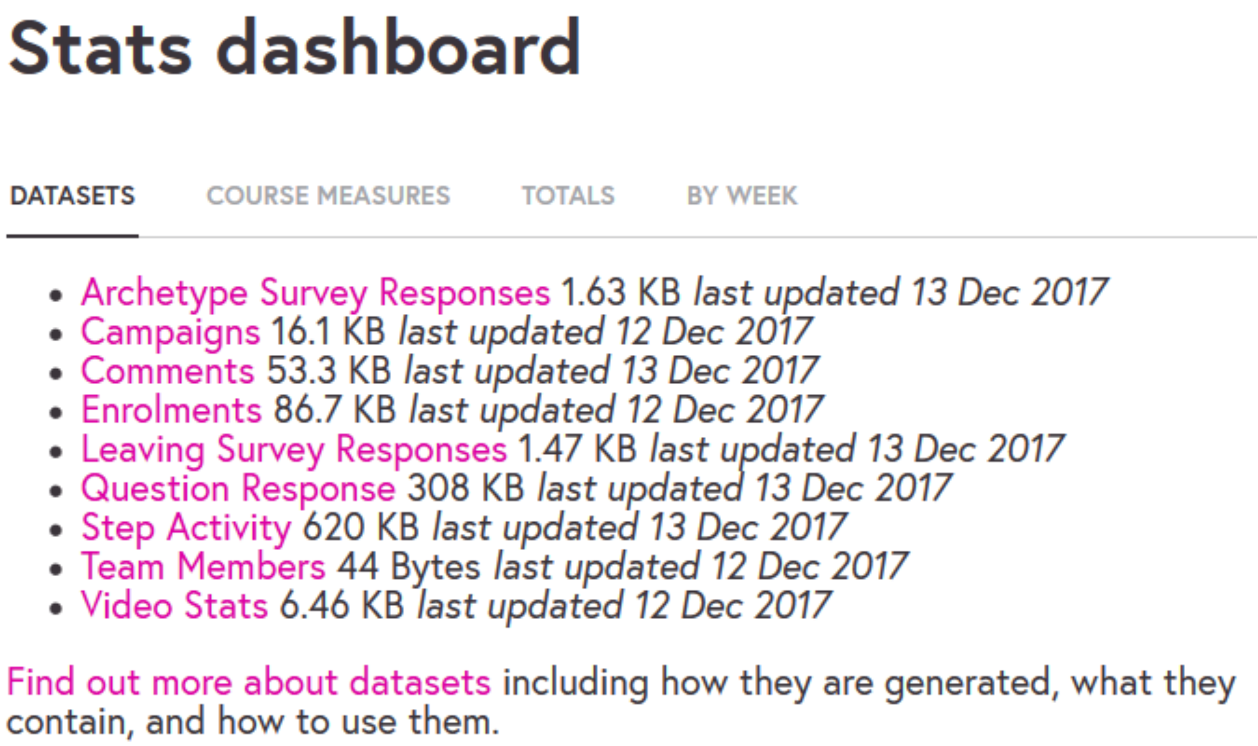


Figure 2: CSV data exports available for MOOC educators on the FutureLearn platform

The main files are the 'step activity' and 'enrolment' files, records of student activity and student details respectively. The other datasets contain information on survey responses, PR campaigns run, and comments made during the course. We focus here on the step activity and enrolment data as information that may be fed into learning analytics.

3.2.1 Data Publication Plan

The data of the TU/e MOOCs cannot be made public as defined by FutureLearn as follows:

[1] https://github.com/innanoval/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data

[2] FutureLearn is a private company wholly owned by The Open University, UK, with multiple worldwide partners. More detail is available at: https://www.futurelearn.com

"You can release any of the aggregated data related to your organisation's course. ... You cannot release any other data relating to an individual learner if there is any possibility that the individual could be identified from the data provided. This is in accordance with data protection legislation, and is especially pertinent for pre- and post-course survey data – where a learner may have provided responses from which they could be identified, without the expectation that these would ever be made public.

You cannot share data about any other partner organisation's courses without the explicit permission of the other partner, even if you are simply referring to it in comparison to your own courses."[3]

The data is however available for analysis within the EDSA project on request, in accordance with these privacy requirements.

4. Data Analysis

4.1 Learning Analytics Task employing VideoLectures

4.1.1 Basic Analysis

The basic analysis carried out on the VideoLectures log data intends to answer questions about visitors to VideoLectures in the area of data science.

We looked at the popularity of particular VideoLectures and where visitors came from. In particular, the log analysis allowed us to establish the number of accesses to VideoLectures in the data science category from 2015-17 — **271,333**.

In addition, we were able to determine the number of visitors from the EDSA website as **455**, with few visits in 2015, then growing in 2016 and 2017.

4.1.2 Visual Analysis

The VideoLectures Explorer[4] is a tool for visual analytics of information presented on Videolectures.NET. The main functionalities of the tool was described in EDSA deliverable D3.3.

The VideoLectures Explorer tool provides a landscape view for a particular query, with detail about the number of lectures found, the total number of views and categories with frequency of occurrence.

We have since, as part of Task 3.4, improved the user interface (UI) functionality, overall, and for search formulation and results presentation. We have also updated the information content.

Figure 3 presents the current look of the VideoLectures Explorer tool.

[3] From https://partners.futurelearn.com/data/what-data-can-i-talk-about-publicly (only available for FutureLearn partners).

[4] http://explore.videolectures.net

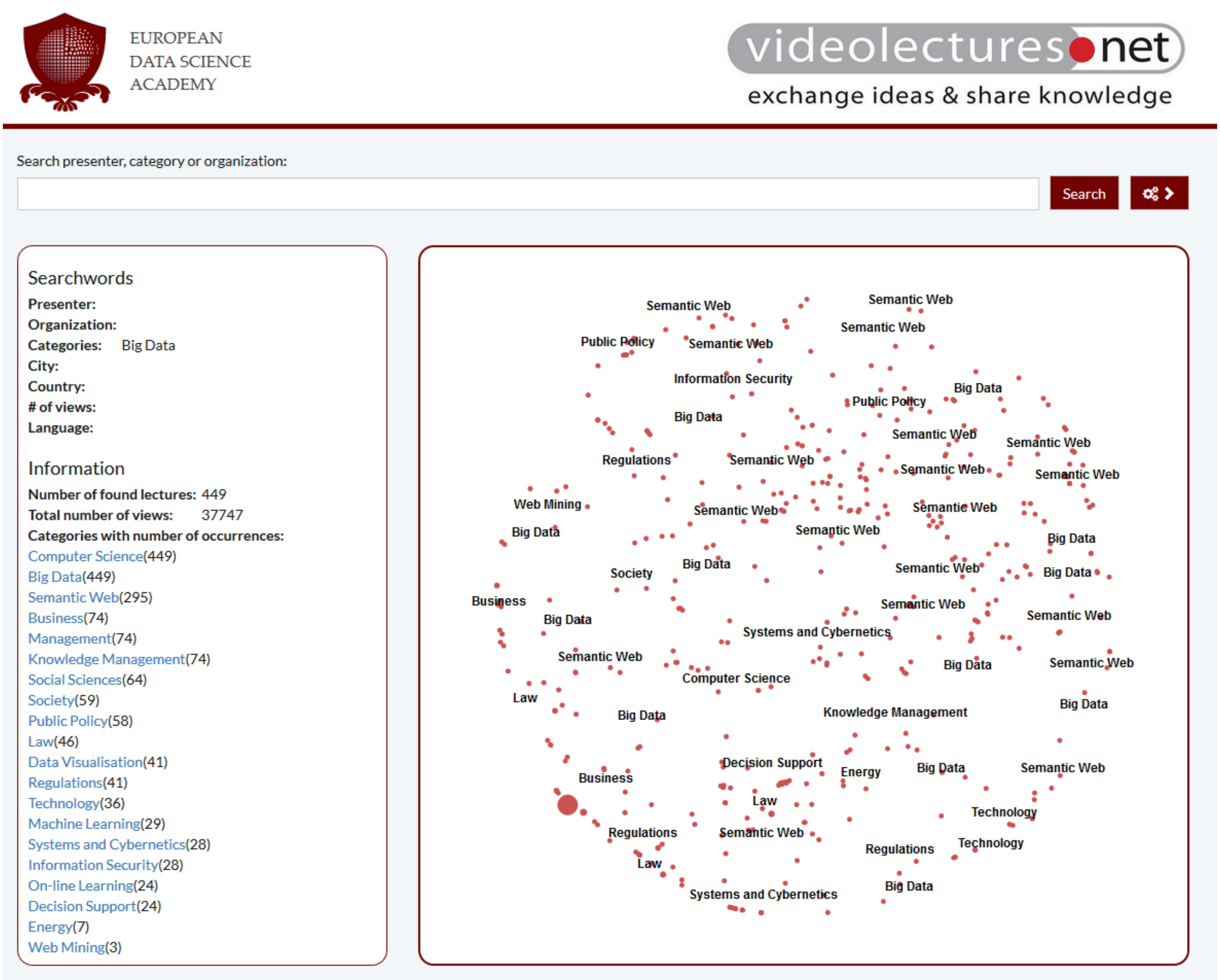


Figure 3: VideoLectures Explorer tool

Figure 4 shows the current version of the search interface, including advanced search options for, e.g., Category, Presenter, City, Organization, Country, Language, Number of views.

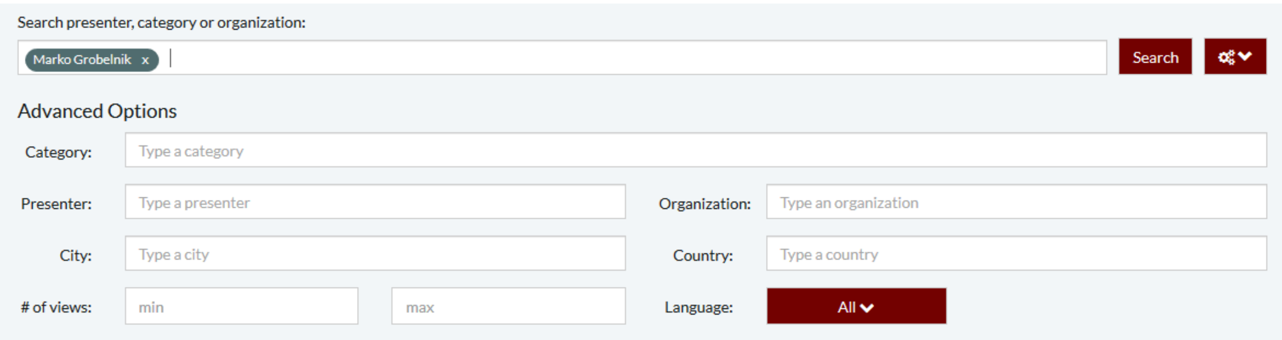


Figure 4: VideoLectures Explorer tool - Search UI

Figure 5 shows an example of search results in detail.

Figure 5: VideoLectures Explorer tool — Search results

4.1.3 Links to EDISON Project

To improve dissemination we have identified the most suitable EDISON Data Science group profile for each video lecture (see Table 1).

The Horizon 2020 EDISON project[5] (1 September 2015 – 31 August 2017) aimed to establish the 'data scientist' as a profession. This was achieved by aligning industry needs with available career paths, and supporting educational institutions in reviewing their curricula with respect to expected profiles, required expertise and professional certification. The expected results include a significant increase in the number and quality of data scientists graduating from universities and being trained by other professional education and training institutions in Europe.

Table 1: VideoLectures according to EDISON Data Science group profiles (April, 2017)

Profile code	Number of associated VideoLectures	Profile description
DSP01	968	Data Science (group) Manager, Proposes, plans and manages functional and technical evolutions of the data science operations within the relevant domain (technical, research, business). Data analytics

[5] http://edison-project.eu

		department manager, Managers
DSP02	1755	Data Science Infrastructure Manager, Proposes plans and manages functional and technical evolutions of the big data infrastructure within the relevant domain (technical, research, business). Big Data Infrastructure Manager, Managers
DSP03	1146	Research Infrastructure Manager, Proposes plans and manages functional and technical evolutions of the research infrastructure within the relevant scientific domain. Research Infrastructure data storage facilities manager, Managers
DSP04	1784	Data Scientist, Data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data - sets, and create visualisations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data. Data Analyst, Professionals
DSP05	1205	Data Science Researcher, Data Science Researcher applies scientific discovery research/process, including hypothesis and hypothesis testing, to obtain actionable knowledge related to scientific problem, business process, or reveal hidden relations between multiple processes. Data Analyst ,Professionals
DSP06	1049	Data Science Architect, Designs and maintains the architecture of Data Science applications and facilities. Creates relevant data models and processes workflows. System Architect, Applications architect, Professionals
DSP07	847	Data Science (Application) Programmer/Engineer, Designs/develops/codes large data (science) analytics applications to support scientific or enterprise/business processes. Scientific Programmer, Professionals
DSP08	1582	Data Analyst, Analyses large variety of data to extract information about system, service or organisation performance and present them in usable/actionable form, Professionals
DSP09	909	Business Analyst, Analyses large variety of data Information System for improving business performance. Business Development Manager (Data science role),Professionals

DSP10	554	Data Stewards, Plans, implements and manages (research) data input, storage, search, presentation; creates data model for domain specific data; support and advice domain scientists/ researchers, Professional (data handling/management)
DSP11	236	Digital data curator, Finds, selects, organises, shares (exhibits) digital data collections, maintains their integrity, up-to-date status and freshness, discoverability, Digital curator, digital archivist, digital librarian, Professional (data handling/management)
DSP12	6835	Digital Librarians, Selection, acquisition, organization, accessibility and preservation of digital information/library. Manages digital materials, takes a lead role in the creation, maintenance and stewardship of digital collections, including the digitization of special collections. Develops strategies for effective management and preservation of library digital assets. Digital data curator, Professional (data handling/management)
DSP13	345	Data Archivists, Maintain historically significant collections of datasets, documents and records, other electronic data, and seek out new items for archiving. Digital Archivists, Professional (data handling/management)
DSP14	363	Large scale (cloud) database designer, Designs/develops/codes large scale data bases and their use in domain/subject specific applications according to the customer needs. Large scale (cloud) database developer, Professional (database)
DSP15	104	Large scale (cloud) database administrator, Designs and implements, or monitors and maintains large scale cloud databases, Professional (database)
DSP16	166	Scientific database administrator, Designs and implements, or monitors and maintains large scale scientific databases, Large scale (cloud) database administrator, Professional (database)
DSP17	599	Big Data facilities Operator, Manages daily operation of facilities, resources, and responds to customer requests. Includes all operations related to data management and data lifecycle, Technicians and associate professionals
DSP18	381	Large scale (cloud) data storage operator, Manages daily operation of cloud storage, Including related to

		data lifecycle, and responds to requests from storage users, Technicians and associate professionals
DSP19	162	Scientific database operator, Manages daily operation of scientific databases, Including related to data lifecycle, and responds to requests from database users, Large scale (cloud) data storage operators, Technicians and associate professionals
DSP20	525	Data entry/access worker, Enter data into data management systems directly reading them from source, documents or obtained from people/users, Data entry desk/terminal worker, Clerical and support workers (general and keyboard workers)
DSP21	2069	Data entry field workers, The same work done on field when collecting data from disconnected sensors or doing direct counting or reading, Clerical and support workers (general and keyboard workers)
DSP22	722	User support data services, Provides support to users to entry their data into governmental service and user facing applications, Clerical and support workers (general and keyboard workers)

4.1.4 Summary

In summary, the results obtained from analysing the VideoLectures data show:

- Basic analysis identifies visitors interested in data science. In particular, the VideoLectures logs show that portal visitors are very interested in data science videos. Links from the Videolectures.NET portal to the EDSA project has also led to increased access, redirected from the EDSA website.
- The VideoLectures Explorer tool has been updated to provide improved overviews, employing visual analytics to improve the user experience with the VideoLectures.
- The links to EDISON profiles allowed us to identify VideoLectures that may be suitable for defined EDISON project profiles, providing additional categorisation of the data and improving access overall.

4.2 Learning Analytics Task employing TU/e FutureLearn Data

Using the data available to us as educators of FutureLearn courses, we can analyse students, student *enrolment*, and student behavior during a course run. In this section we analyse the data of the first run of the TU/e FutureLearn MOOC 'Introduction to process mining with ProM', which ran from July 11, 2016 to August 7, 2016, but remained open beyond the end of the course. In discussing the results obtained events logged by the system are highlighted using italics.

In this section we analyse the course using the statistics available to us through the FutureLearn platform (see section 4.2.1). We then look at the results of visual analysis and process mining, as provided by the FutureLearn analytics workflow developed within the EDSA project (discussed in more detail in section 5). All results in sections 4.2.2 and 4.2.3 have been obtained automatically, after running the learning analytics workflow, without manual intervention.

In section 4.2.4 further visual exploratory analysis is presented, employing an approach that includes a human in the loop, to support a progressive approach that poses new questions based on results obtained during each step.

We triangulate our findings in sections 4.2.5.

4.2.1 Basic Statistical Analysis

In total 4,875 students registered for the course. Of these 525 students (10.80%) left the course explicitly at some point (counted using the *unenrolment* event), some after completion. 28.7% of the *enrolled* students, or 1,399 students, were learners (i.e. accessed one or more steps), of which 1,000 (71.5%) active. Of these 261 (18.70%) learners were socially active, i.e., participated in commenting on the course. In the end we had 114 *fully participating* learners.

According to our expectations these numbers are very good. Note that we do not use this course in on-campus teaching at TU/e, so almost all students are doing this for their own interest. We do not push students to pass the course, but aim rather to enable students to obtain knowledge they desire or need, which could be obtain by viewing only 3 specific videos out of all the material provided for the course.

4.2.2 Visual Exploratory Analysis employing 'Automated Workflow'

The statistics discussed in section 4.2.1 can be obtained through the FutureLearn teacher interface, and do not require downloading the detailed files. However, they provide a limited view on the course, its students and their behavior. We therefore apply our learning analytics workflow on the datasets available for our courses as educator.

General note: not all students provided detailed information such as gender, age and resident country. We therefore only report on the known statistics.

Gender

Out of those who provided demographic data there were 320 male and 206 female students, with 3 with a non-binary or other gender. For the other 4,346 students no information on gender was available.

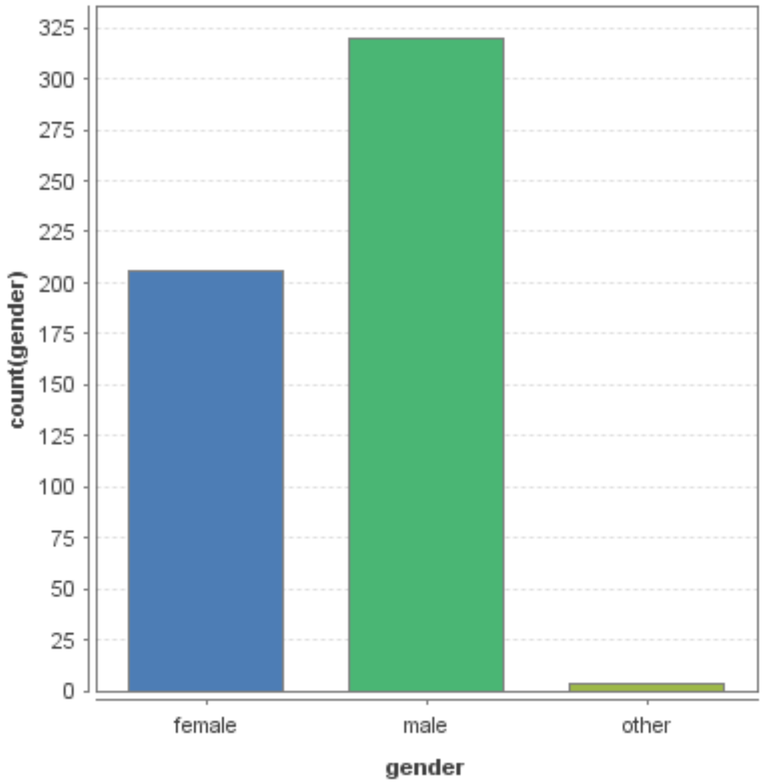


Figure 6: Gender distribution for students providing demographic data

Age

Figure 7 shows the student composition for the course in terms by age; Table 2 provides the detail on age.. Note that as for gender this only covers the sub-set that provided demographic information.

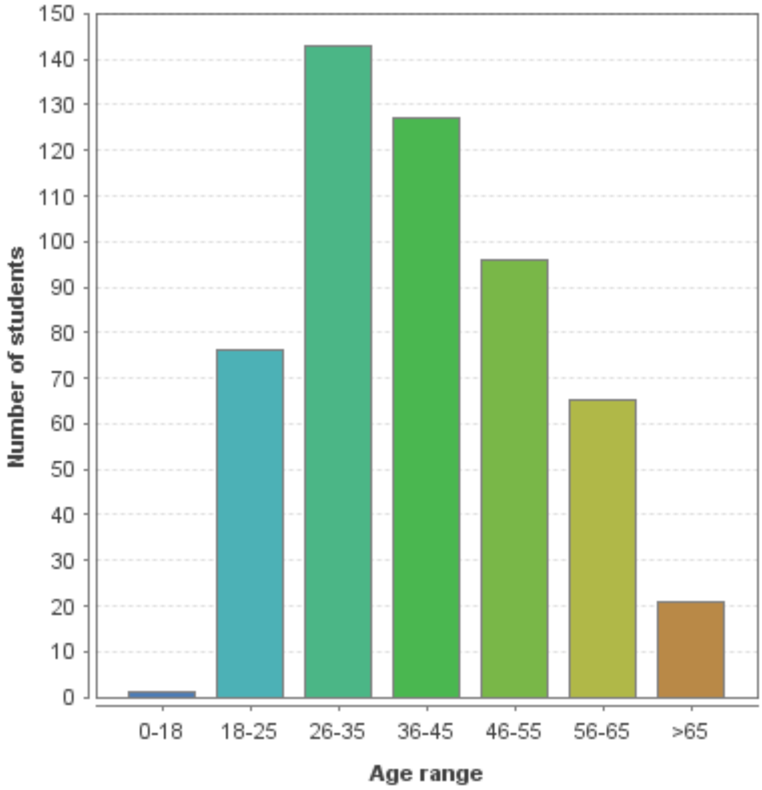


Figure 7: Age distribution for students providing demographic data

Table 2: Number of students per age range

Age range	Number of students
<18	1
18-25	76
26-35	143
36-45	127
46-55	96
56-65	65
>65	21

Country

This data table shows student composition of the course in terms of the country the student lives in for the top 10 countries in reverse order.

Table 3: Number of students per country, top 10

Country ID	Country[6]	Number of students
GB	Great Britain	142
US	United States of America	29
DE	Germany	20
IN	India	20
NL	The Netherlands	17
AU	Australia	16
BR	Brazil	15
ES	Spain	13
IT	Italy	10
PL	Poland	10

Given the base and public relations (PR) efforts of FutureLearn it is not surprising that the main country of origin is the UK. India in the fourth place is unexpected, along with Brazil in the top 10. Most other students are located in Europe, the US and Australia.

Visited course steps

Figures 8 and 9 show for each step (video, article, discussion, quiz, etc.) in the course how often it has been visited by students, and what fraction of these students marked the step as completed.

[6] This column is not present in the automatically generated table

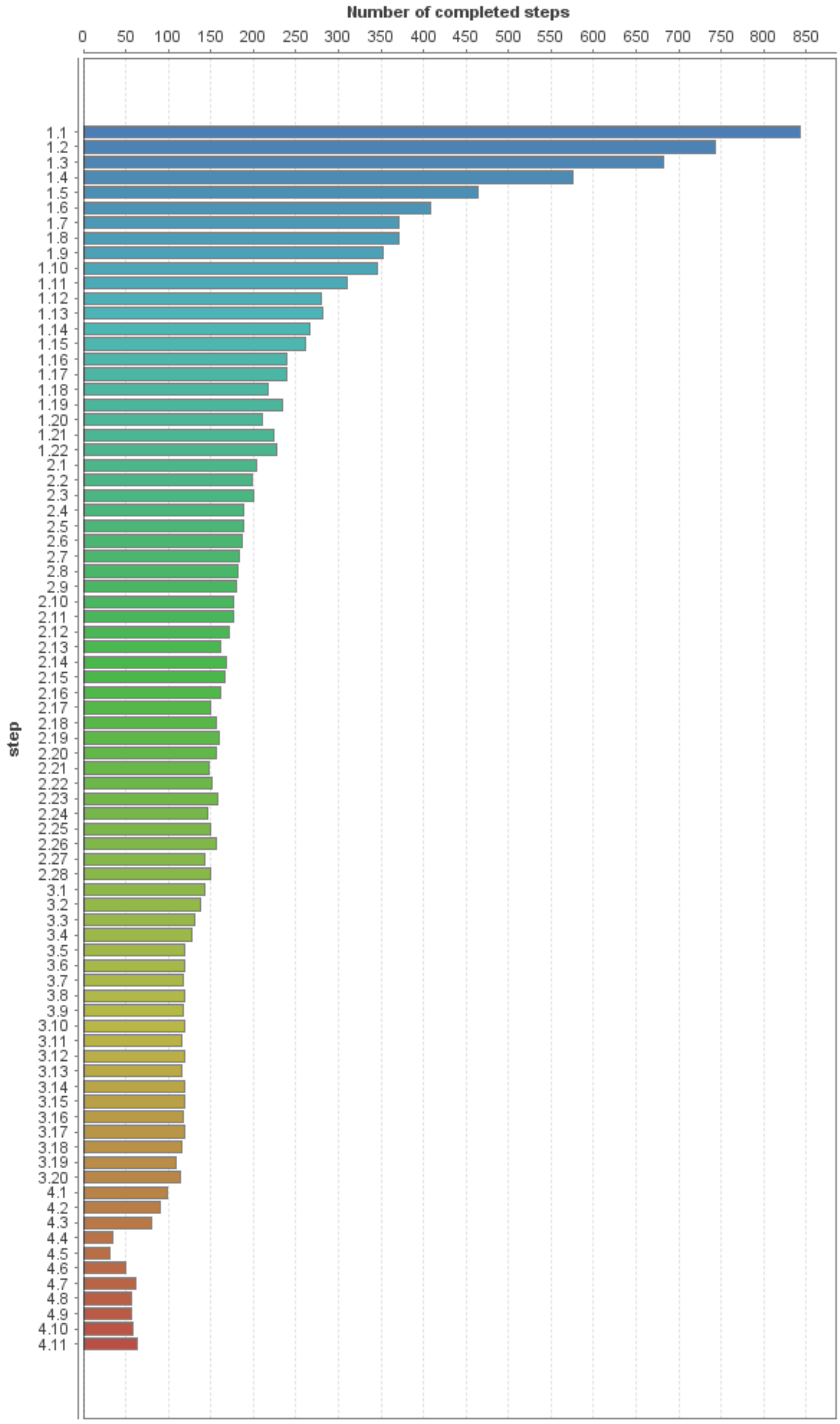


Figure 8: Number of students that marked a step as completed.

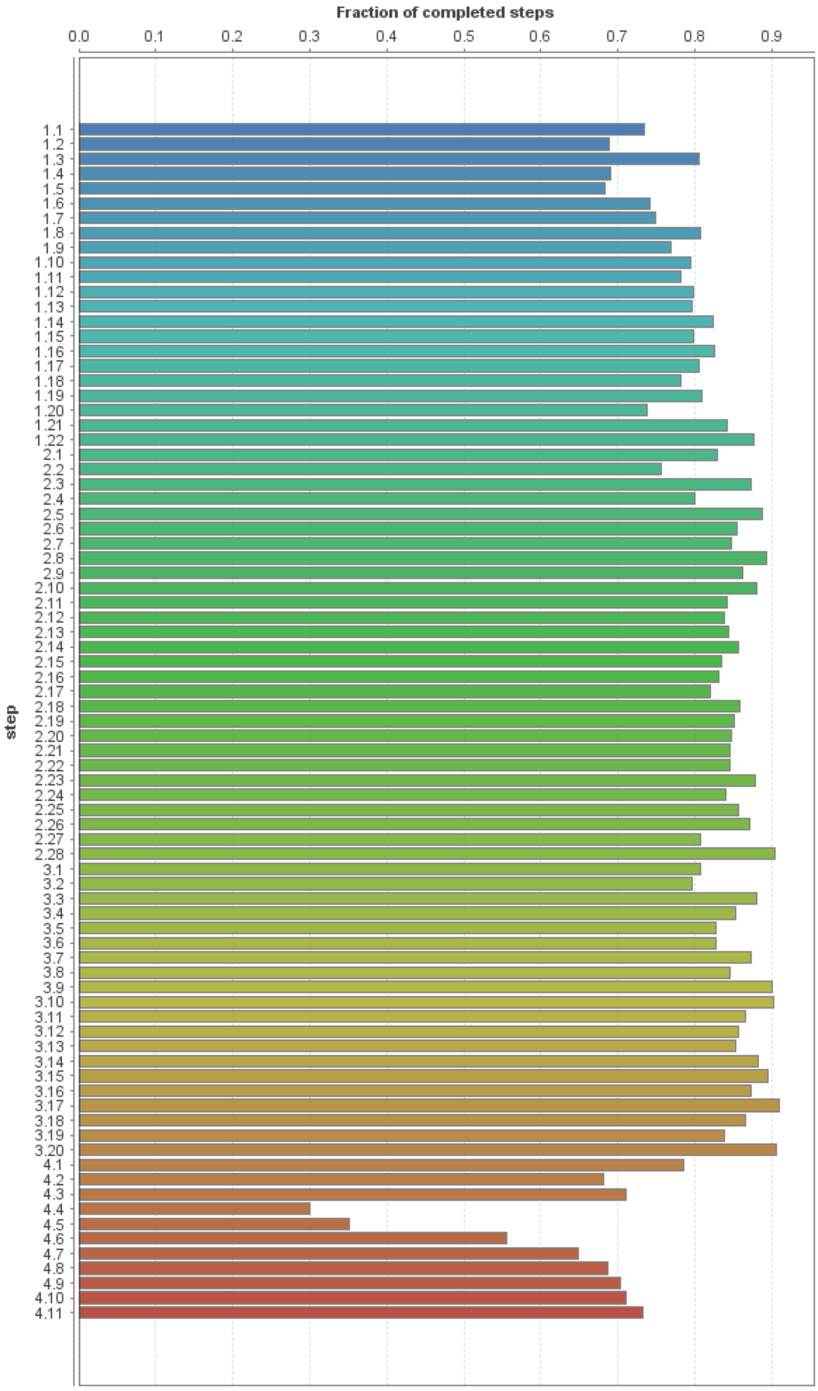


Figure 9: Fraction of learners that visit a step and mark it as complete.

From Figures 8 and 9 it becomes clear that the number of students decreases over the course run and that the first few steps are visited most. From the completion ratios it also becomes clear that not all students have the desire to complete the peer assignment which is the major part of the fourth week. This strengthens our belief that not all students that actively follow the course really want/need/require a certificate, but mainly follow it to acquire useful knowledge.

Enrolment over time

The following line chart shows *enrolment* and *unenrolment* over time.

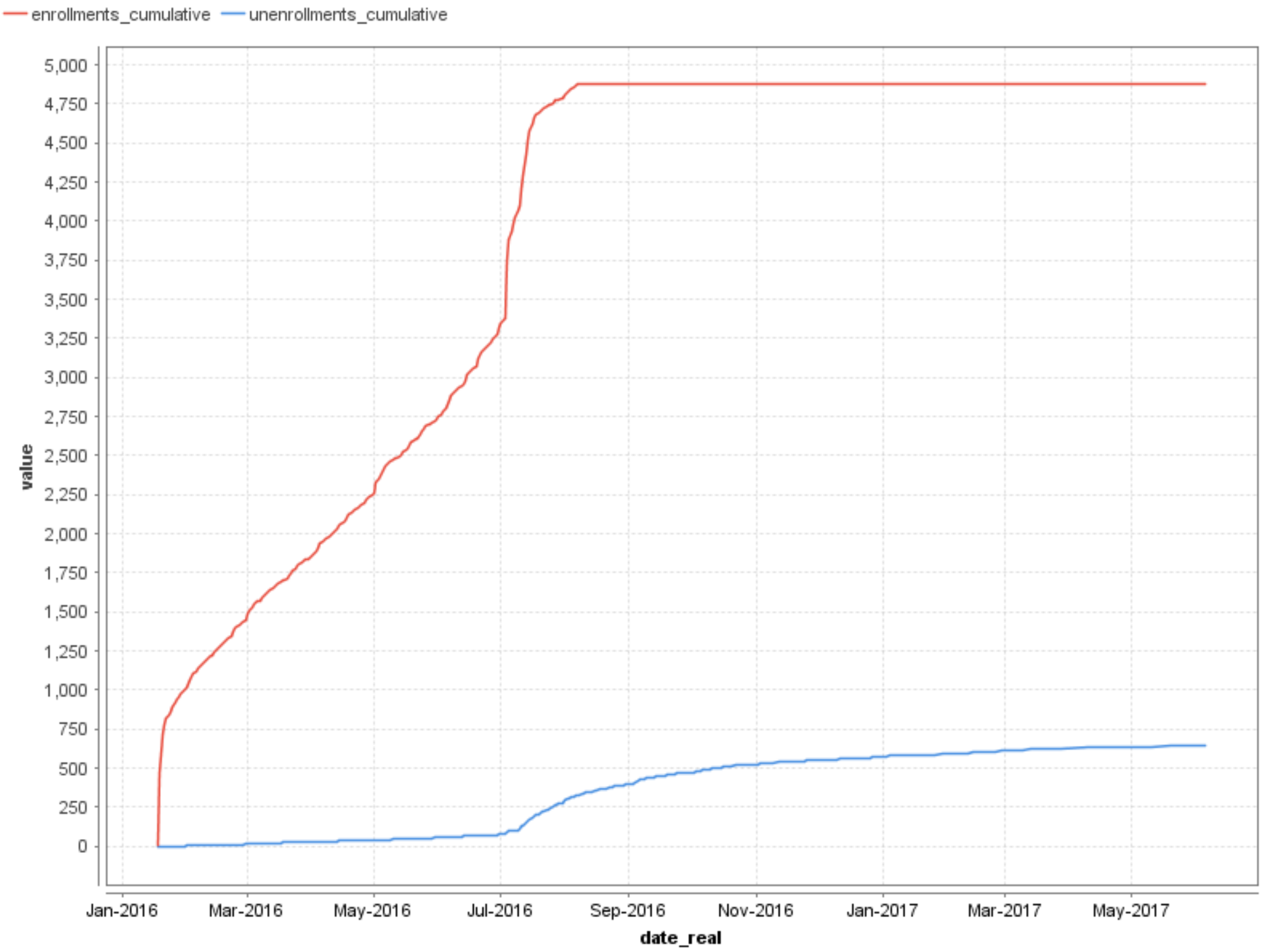


Figure 10: *(un)enrollment* of students on/off the course over time (course starts in July 2016)

Given that the course started on July 11, 2016, it becomes clear that the vast majority of students *enrol* on or around the start of the course. Another large chunk registered as soon as the course was first announced, at the beginning of 2016.

Visited steps over time, per content week

The line charts in Figure 11 show when steps are watched over time. The steps are aggregated based on the week of the course.

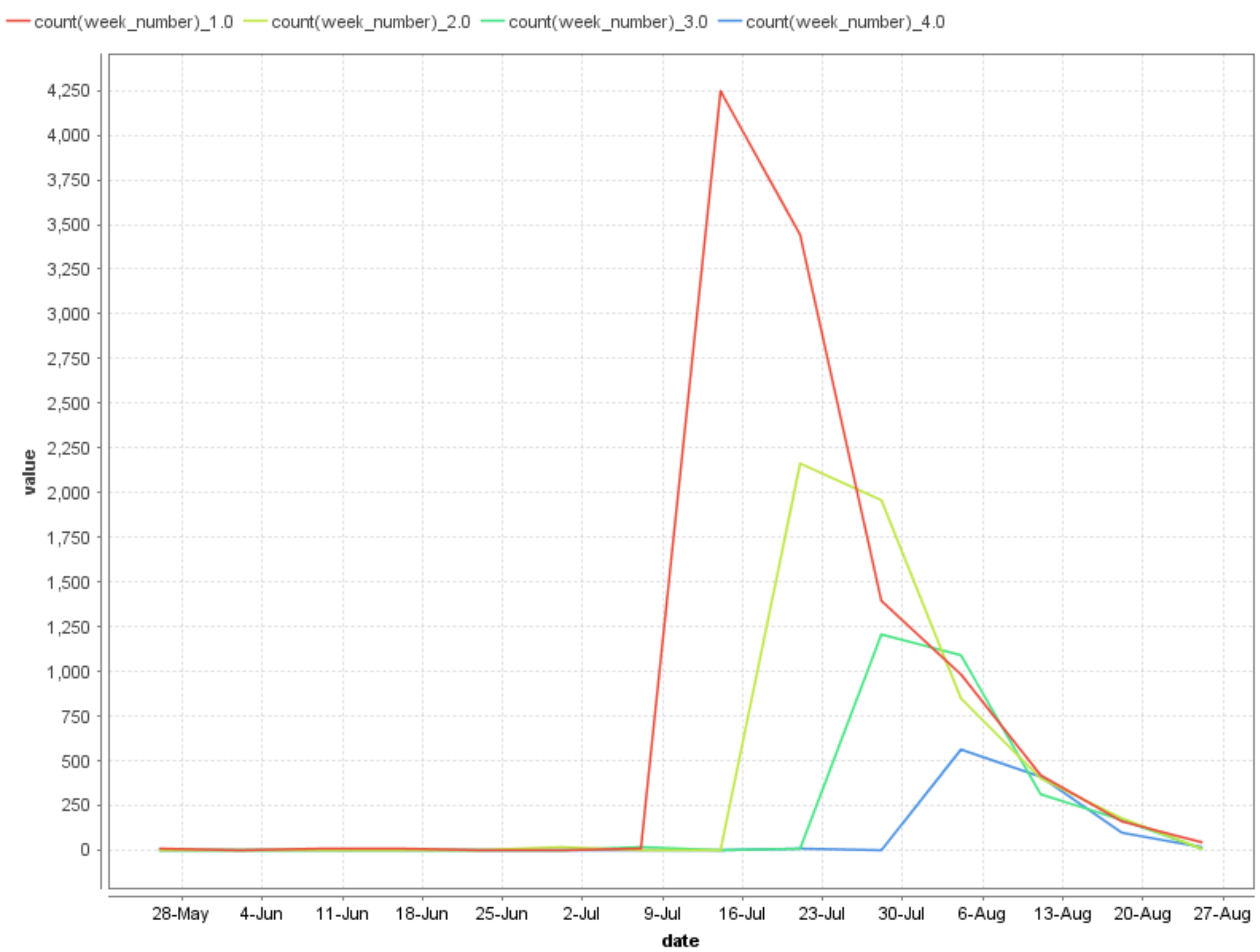


Figure 11: Number of step visits aggregated by week, over time

From the chart above it becomes clear again that students visit earlier weeks more often than later weeks. Further, we observe that the content of the first week is still visited frequently in the third and even the fourth (last) week of the course.

PR campaign effectiveness

This section presents information about the way learners *enrolled* on this run of the course. Figure 12 shows the top 10 most used domains.

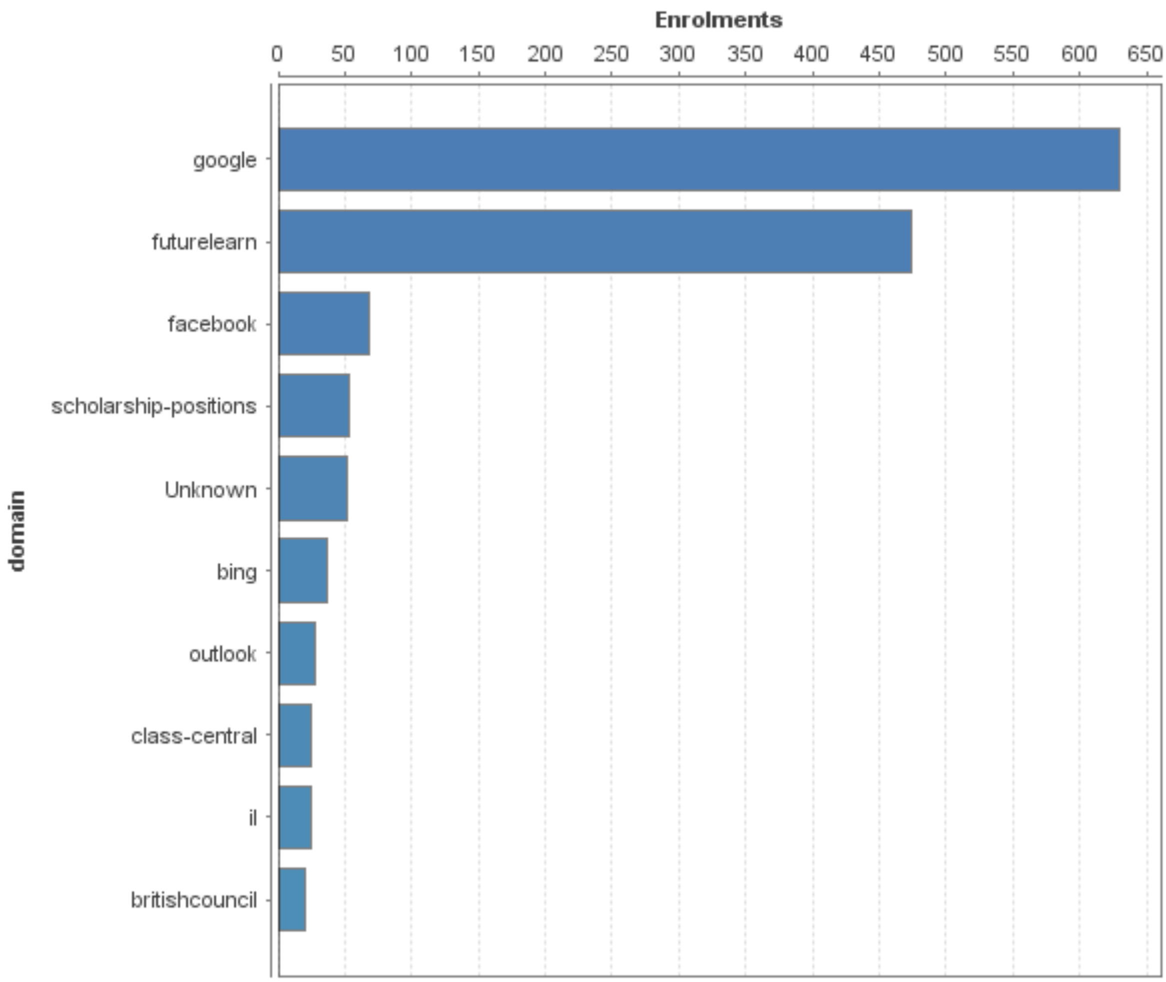


Figure 12: Sources for student *enrolment*, showing where students came to the course from

Figure 12 indicates how students 'arrived' at the course, where Google is a predominant source of enrolment, closely followed by FutureLearn directly. Google is likely to attract people that actively search for the content discussed in the course. The people joining via FutureLearn are most likely *enrolling* on the course because they see it advertised while *enrolled* on another course(s).

Note that for this course we did not buy in any ads or launch a large PR campaign. We used mainly social media and posted to some mailing lists. We however did not explicitly track effectiveness of any of these channels.

Grades per quiz question

Figures 13 and 14 show how well students answered the questions. Figure 13 shows the 10 most incorrectly answered questions and Figure 14 shows the top 10 correctly answered questions. The question number is formatted as follows: week.stepnumber.question. Hence, 2.3.1 corresponds to week 2, step 3 and question 1.

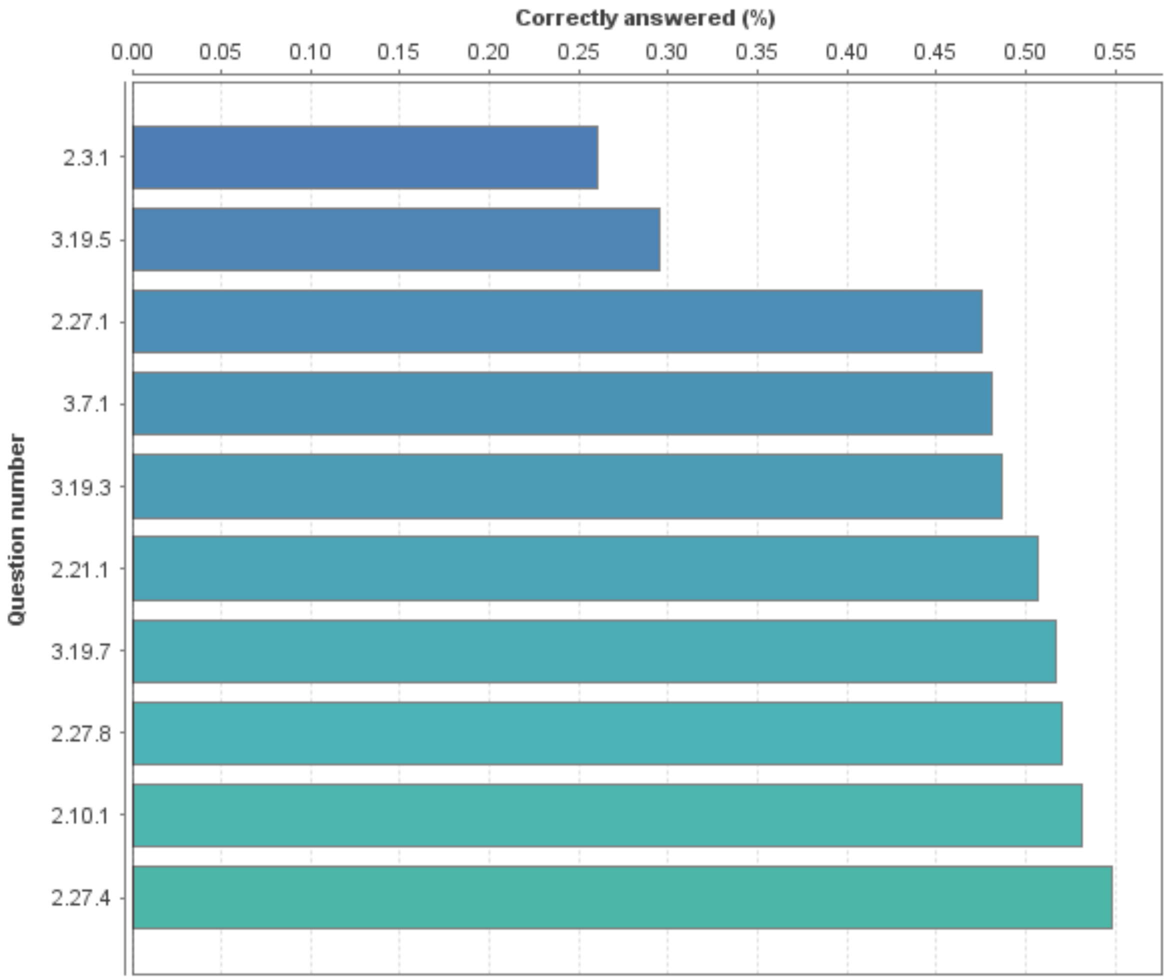


Figure 13: 10 quiz questions with the lowest percentage of correct answers provided

Such charts are of great help to educators who wish to improve the quality of their courses. They indicate which specific questions are, relatively, ‘bad’ or ‘good’.

For instance question 1 in the quiz of step 2.3 was the first question in the course where students don't need to tick one, but possibly multiple answers. We decided not to change this question as we wanted students to be aware of this feature on the platform, especially as it is often used in tests.

The second worst made question is question 5 in step 3.19, the test of week 3. In this question students are asked to use the result obtained in the previous question, with detailed instructions provided. Students are required to evaluate the results in a bit more detail, a task which proved difficult for most students. On the other hand, this topic is the most difficult one addressed in the course, so it is not surprising that this is the hardest question for most. We decided to keep the question as is and to focus on the discussion and comments made on the topic as a means for obtaining feedback for future delivery of the course.

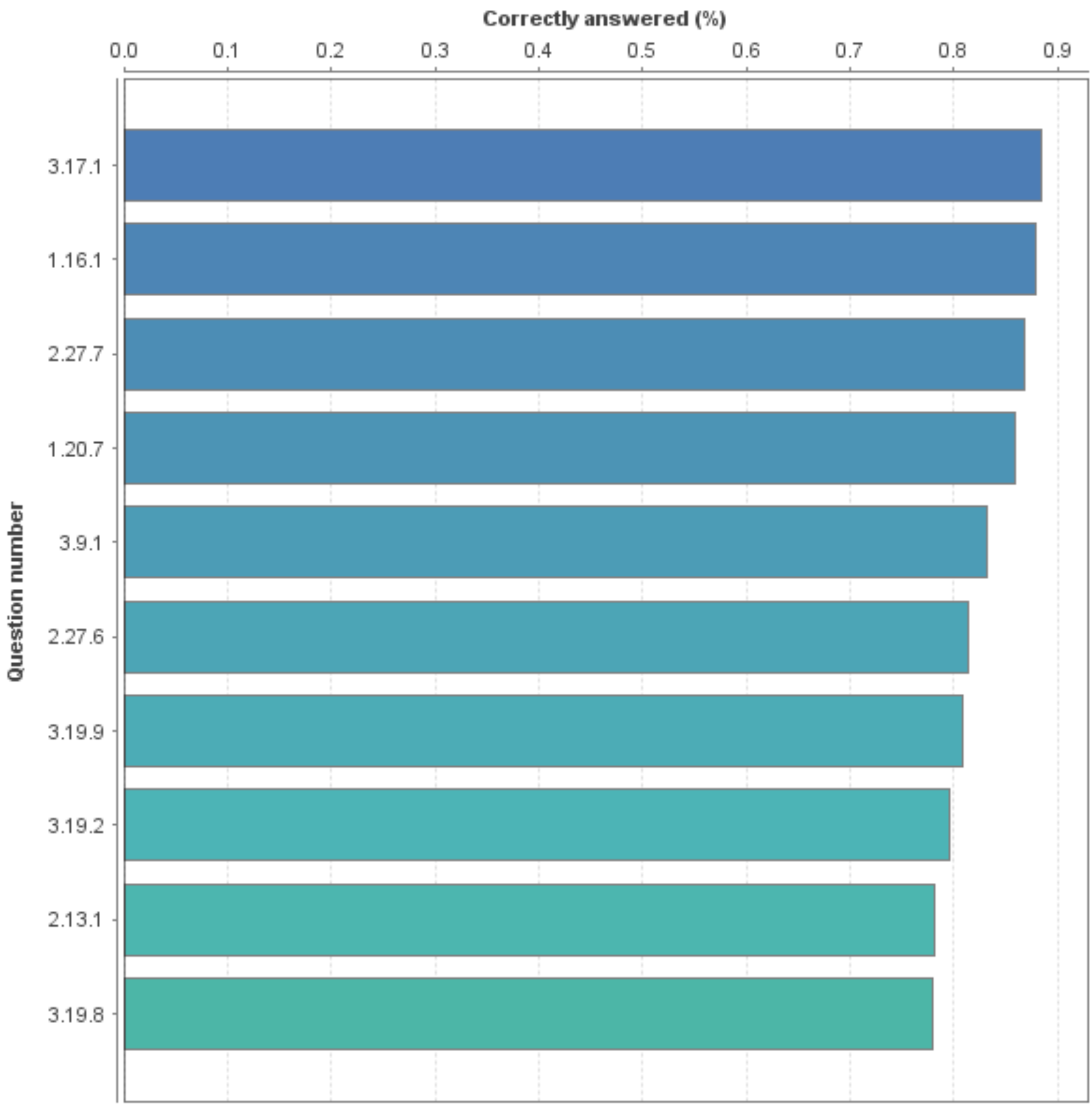


Figure 14: 10 quiz questions with the best ratio of correct answers

4.2.3 Process Mining

The automated learning analytics workflow also produces selected (interactive) process mining results.

Dotted chart analysis

The first of these is the Dotted Chart, which visualizes each learner action using a dot. The x-axis is the calendar time, while each learner's data is presented on a row. The dotted chart in Figure 15 shows that most students start from the course star (11th July 2016), while some learners record their first activity as late as August. Note that the reader is not expected to discern detail in these charts, but rather the overall patterns that result.

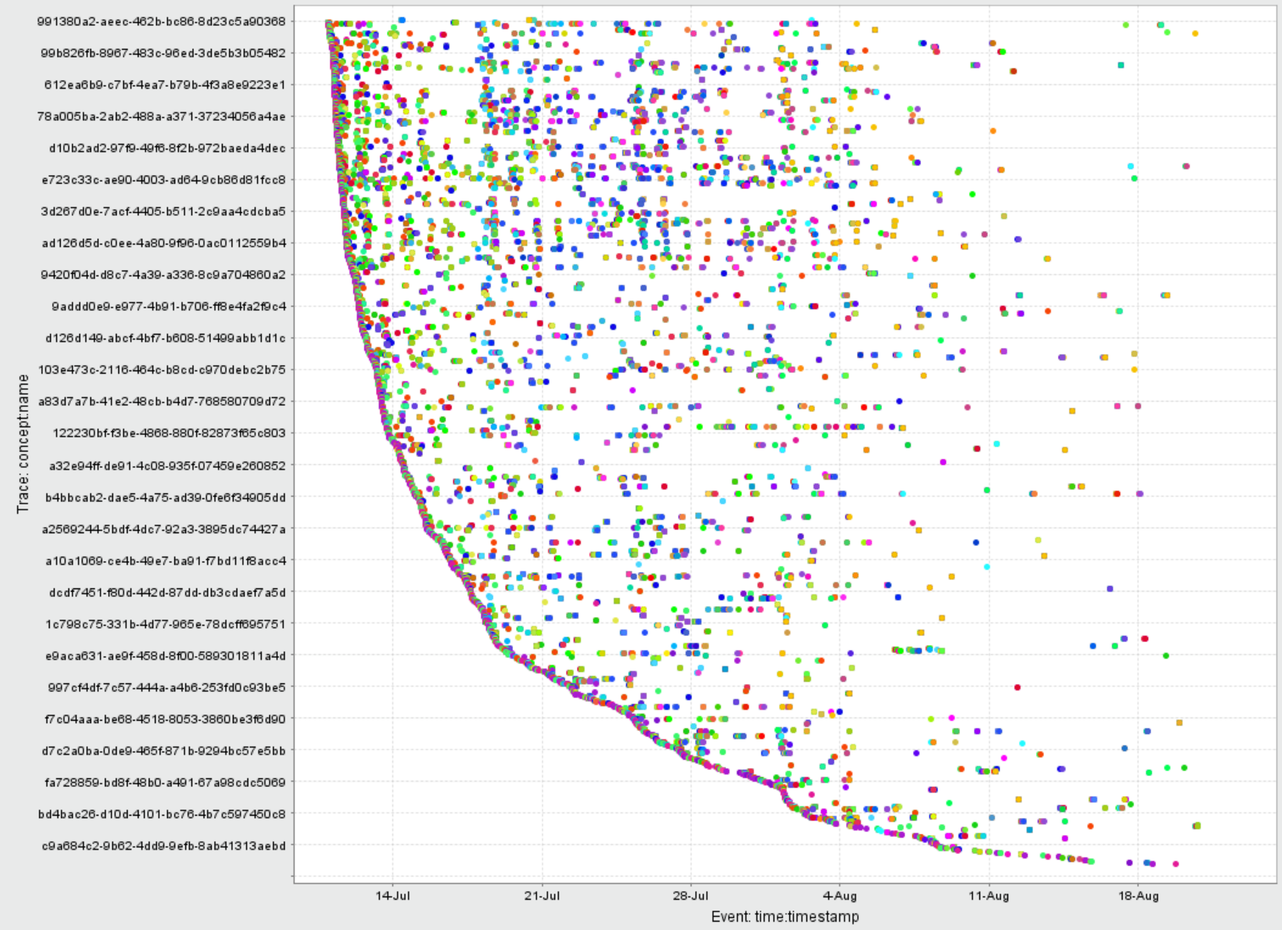


Figure 15: Dotted chart showing each action as a dot for each student over time (x-axis, from July 7 (a few days before) to August 25 (after course completion))

Figure 16 is sorted by time for the first observed action per student, then sorting student order top-bottom by shortest observation period. It is clear that most students only visit the course once and then browse some steps. Some learners spend more time on the course, up to 70 days after the first observed action. On average students visit the course over a 2 to 3 week period.

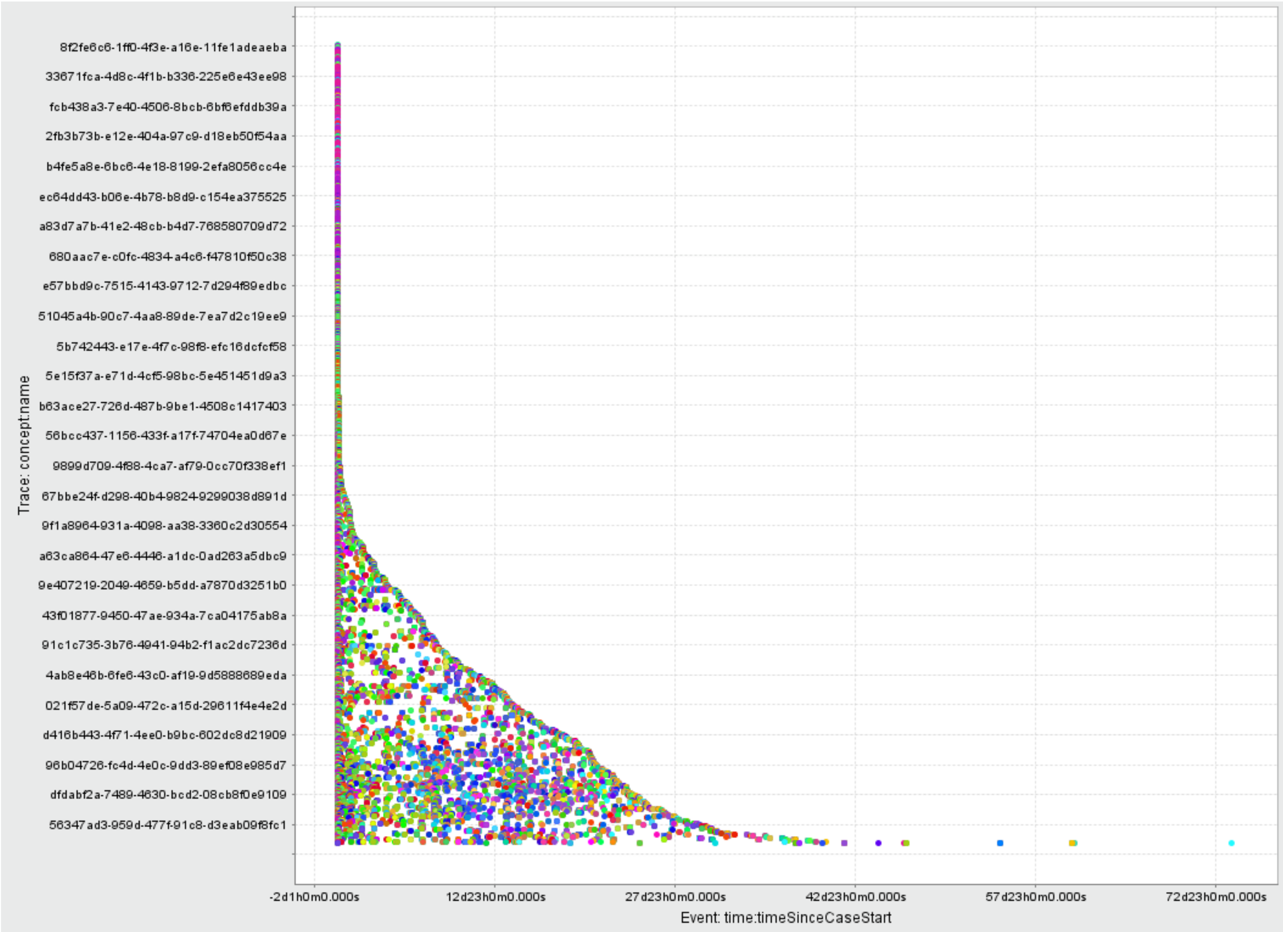


Figure 16: Dotted chart showing the actions relative to the first observed action, sorted by duration of observation (x-axis, to 73 days) per learner (rows)

In Figure 17 we change the x-axis to the observation count/position (e.g. first observation, 2nd, up to 80th). Given that most ‘columns’ have a similar color distribution it appears that most learners follow the course in the intended sequence, with only a few skips and loop-backs. We carry out further analysis to determine whether this is really the case.

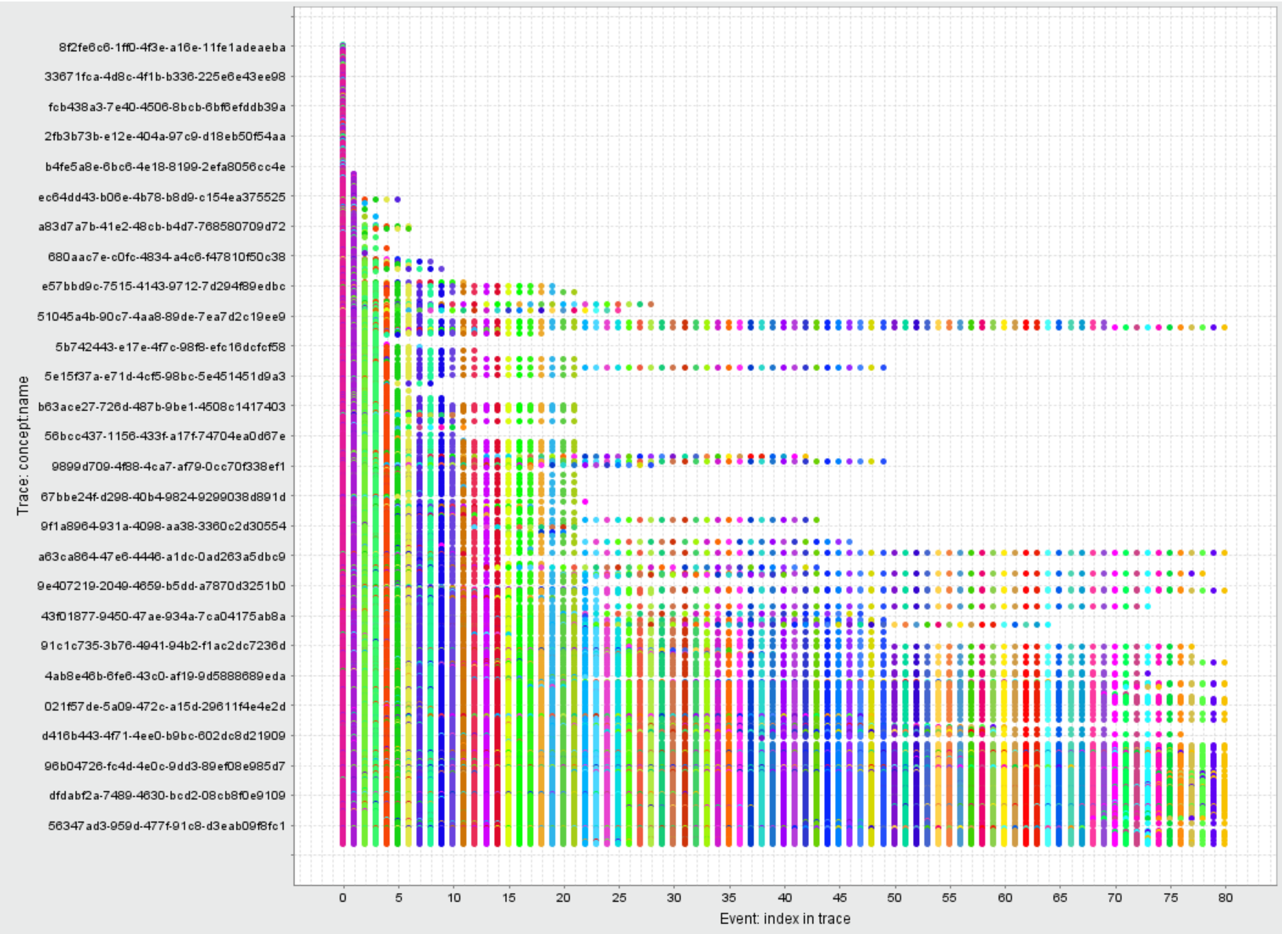


Figure 17: Dotted chart showing the observed action sequence (x-axis, up to 80 observations) per learner (y-axis), indicating some regularity in the order

Process discovery and process compliance

Teacher set up courses with an intended order for following steps. Learners are expected to start in week 1, then move on to week 2, and so on. Whether this occurs can be investigated using this data and process mining analysis: educators can examine which steps are skipped and what learning paths students follow.

Process discovery algorithms can take the data and discover/learn a process model. An example of such a process model is shown in Figure 18, which mainly consists of all steps in the intended order, but with a skip to bypass the steps.

Conformance analysis indicates that roughly 90% of all observed activity fits within this sequential behavior (including skipping lectures). Inspecting the beginning in more detail, we observe that 1,148 students visit step 1.1 first, while 185 students do not visit this step. 1,079 students visit step 1.2 in the expected order, while 254 do not. For step 1.3 the algorithm includes a ‘skip’ option: 847 students watch this step in the expected order, and for step 1.4, 834 students. A decline in students visiting steps is visible towards the end of the course.

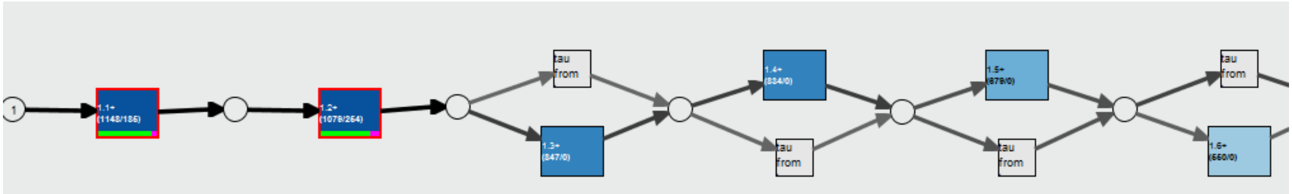


Figure 18: Discovered process model describing a sequential learning process where steps can be skipped. This figure zooms into the first part (bottom).

Applying a different algorithm provides a different result, showing far less adherence to the intended learning path, as shown in Figure 19. It shows that the first three steps are usually watched and in the intended order. We then observe a less structured process where steps are watched in almost any order, and may also be skipped. This process model is able to explain more of the observed behavior, but at the same time is less precise in doing so (as it allows for unseen behavior). It does however indicate that **most students do not follow the intended learning path**. This is not 'wrong' or 'bad', but definitely something to consider as an educator while building and running an online course.

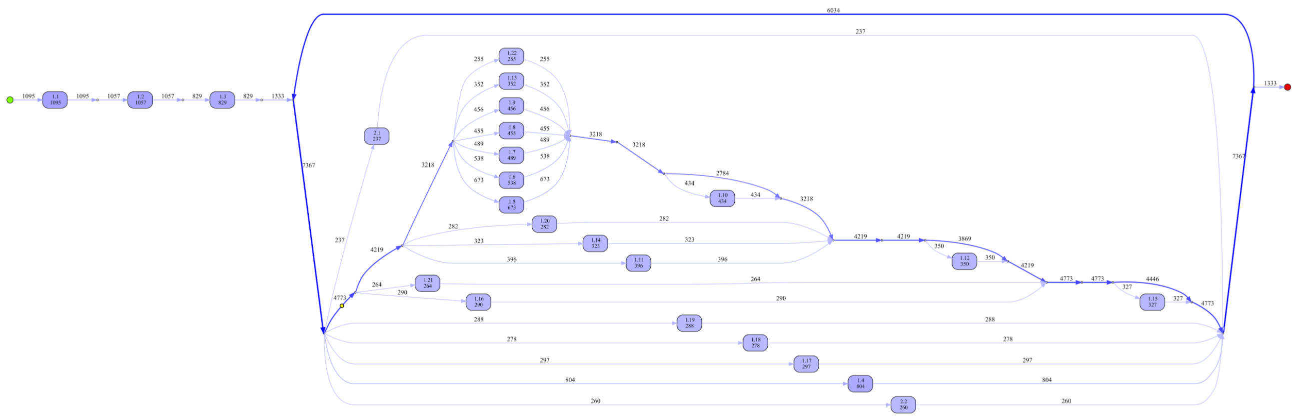


Figure 19: More advanced process model of learner behavior showing less structure

Both discovery algorithms are `correct' in a sense, they provide different results because each focuses on different characteristics in the process model. For instance, the results for the process model in Figure 18 is sequential and allows 10% of the behavior to be not explained by the model. The process model in Figure 19 tries to capture more behavior, but as a result becomes more complex and less precise (due to the recall-precision conflict in data mining). The main observation from Figure 19 is that trying to explain this 10% 'irregular' behavior is not easily done. There is no simple explanation for the deviation in learning behavior.

4.2.4 Visual Exploratory Analysis with 'Human in the Loop'

This approach used visualisation to support user-directed exploration of the data, to identify potential areas to explore in further detail. Two areas were addressed:

1. the visual analytics module applied to the EDSA online modules [EDSA Deliverable D3.3] was ported to the FutureLearn data. This provided, first, a test of reusability for other learning analytics datasets using the events model. The overview generated could further be contrasted with the analysis carried out using other approaches, to both confirm and augment results obtained in both cases.

2. preliminary exploration of the use of the discussion forums by students. Among the topics of interest in learning analytics for online courses is the impact of student-student and student-instructor interaction on performance, engagement and dropout, and the use of social interaction as a learning tool in itself (see, e.g., [Buckingham Shum 2012, Wise 2014]). This study looked for patterns within the data based on student contributions to discussions, to determine, ultimately, how
    - and for what purpose discussions are initiated,
    - discussions develop,
    - level of activity and discussion content relate to course content — which in turn maps to point in time (course), recorded as week and step numbers,
    - the relationship between online interaction and student outcomes.

The analysis that follows is based on the run of the FutureLearn course (described in section 3.2). Starting with the overview dot plots in Figure 20, ordered from top-bottom by activity count, we observe that most activity, barring enrolment, is clustered in the top fifth of the plot. Gradually filtering out less active users, we finally concentrate on the top 500 most active users — just over 1/10th of the total enrolled; outside this sub-set most activity pertains to enrolment and unenrolment. The next most active batch of 500 records a very small number of events, predominantly *quiz* activity and contributions to *discussions*. However, proportionally this is so few as to distort the picture obtained overall when compared to the top 500 alone.

The sequence of snapshots that follow demonstrate patterns seen in the overview, and how these change based on successive filtering from the ordered overview. This action provides the equivalent of zooming in slowly to the ROIs revealed in the interactive prototype. Note the focus here is on patterns in the overview rather than the detail in individual events or learners. The figure captions highlight changes to follow in each subsequent snapshot.

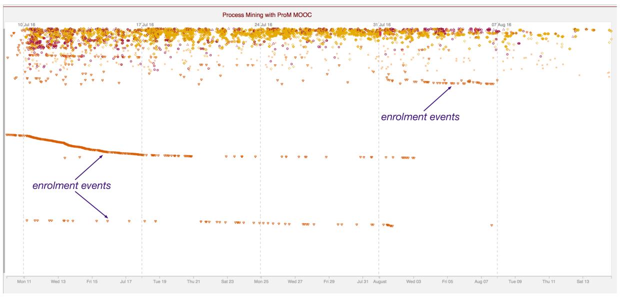


Figure 20: The dot plot for all students (4875 including 8 admin), ordered top-bottom by activity count. Three chains of enrolment events can be seen, below the main activity.

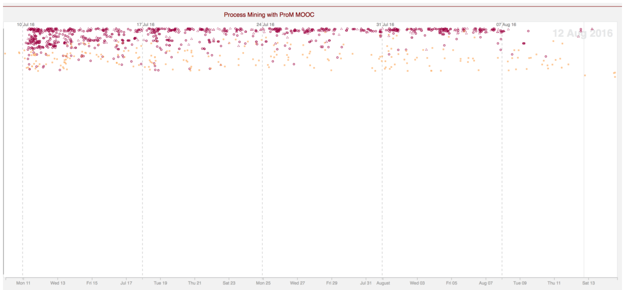
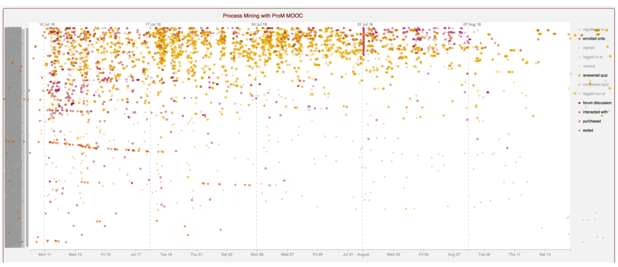


Figure 21: Filtering out the enrolment activity in Figure 20 shows that most other activity occurs in the top fifth of the plot.



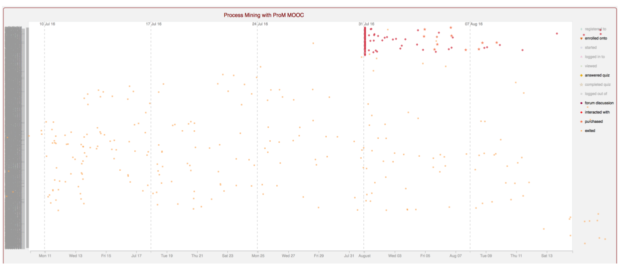


Figure 22: Filtering out all but the top 1000 most active (left), then leaving only *purchased*, *full interaction* and *unenrolment* events (right) - most activity in the bottom half regards enrolment and unenrolment.

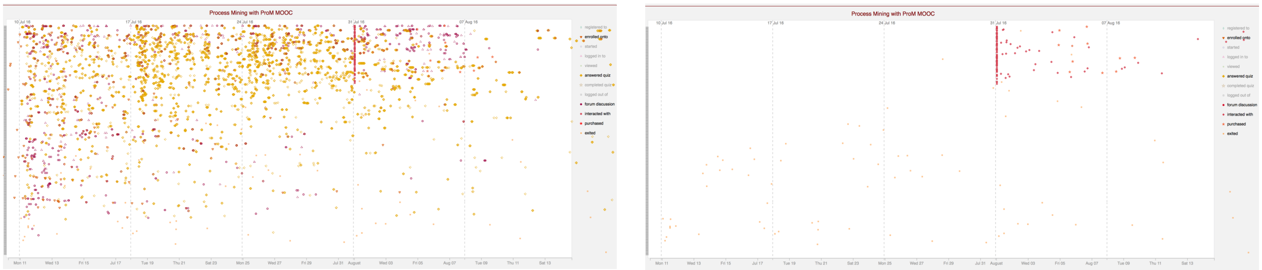


Figure 23: Filtering further to focus on the top 500 most active learners (left), then only events *purchased*, *full interaction* and *unenrolment* (right), the top third contains all of the first two events and the bottom 2/3 most of the *unenrolment*.

Filtering down to the top 500 (Figure 24), then 150 most active learners (Figure 25) we are able to hone in to the patterns for *full interaction* (encircled, top, Figure 24) — occurring mostly at the start of the final week (4), and most of the *purchase* events scattered though the second half of week 4 and into the start of week 5, after the course ended. A small number of unenrolment events occurs even for these learners, mostly toward the bottom, with falling activity. A selection of *purchase* events in the final week (4) is highlighted. Further work will investigate whether this also maps to (perceived) level of difficulty of assessment.

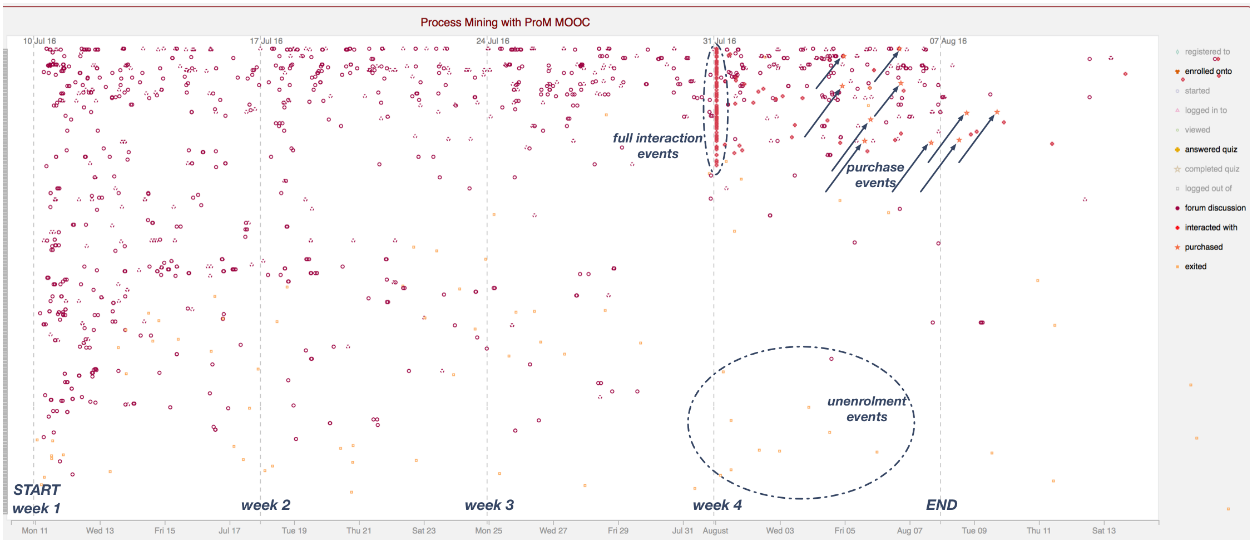


Figure 24: Focusing on the discussions for the top 500, the top third shows interaction throughout the course, with a few comments persisting into the following week. For the rest of the plot we see interaction falling away after the first week, then sharply after the halfway point. Contributions seldom occur beyond the third week. Most of the much smaller number of unenrolment events for this sub-set occurs in the areas of sparse activity.

Zooming in to this ROI and applying a filter, the *purchased* (stars) and *full interaction* (diamonds) events only for the data set shows more clearly where this cluster lies, at the start of week 4, then scattered over the rest of week and into week 5, beyond the end of the course.

Highlighted at the top is the *purchased* event for the learner highlighted in Figure 27 — the most active overall.

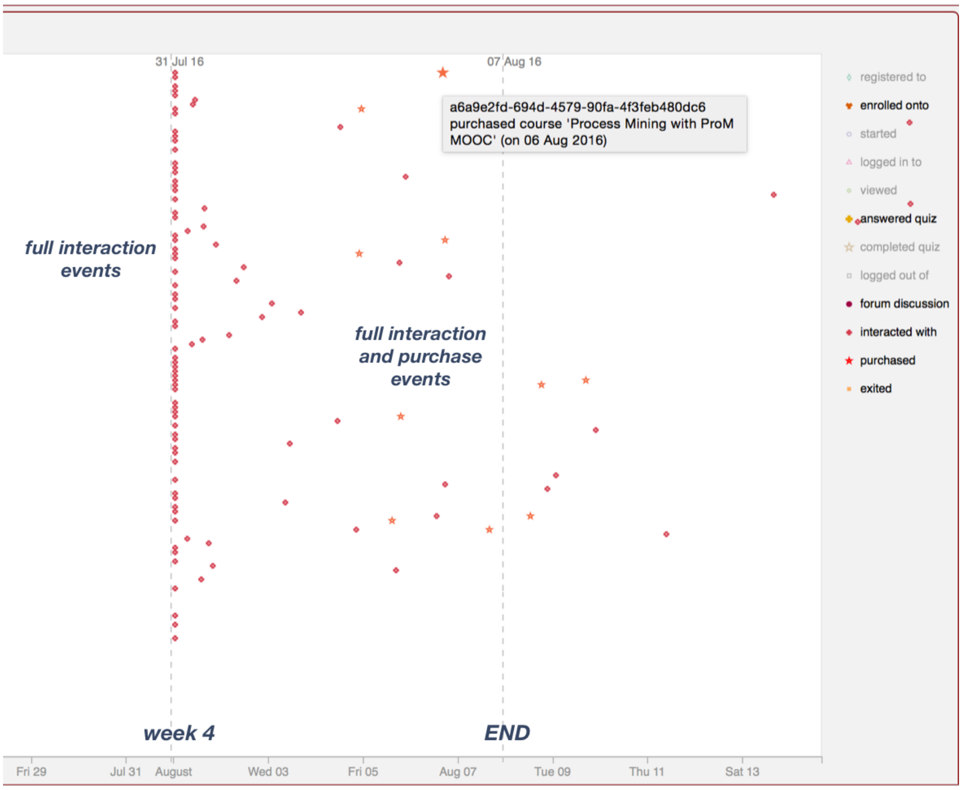


Figure 25: Zooming in to the top 150 (the ROI at the top right, Figure 24)

The next step in the study was to identify networks of (interacting) individuals and topic clusters. This was to allow us to look also at the larger picture of student behaviour, and in future work, relate this, where provided, to any mappings that may exist between student demographics and

- motivation to study a topic,
- whether interest is maintained and, if so, what supports this,
- the level of interaction with the course material and
- actual performance.

To this end, first, two main visualisations are being used to supplement the overview dot plot:

- a discussion-centric network rooted on the forum entry that initiates a new discussion thread (or trail). Visual cues are used to distinguish the focus (selected entry), the discussion root and all other contributions, each linked to its parent (as shown in the legend superimposed on Figure 26).
  This view always results in a tree of depth one as the FutureLearn platform only allows responses to the root of a discussion. Additional visual cues may be used to indicate order of contribution; further analysis of content is required to determine where entries that follow are actually targeted at a subsequent, rather than the root entry. This will be tackled as part of future work; the discussion-centric network provides a very restricted view on interaction, we focus here rather on the alternative structure.

- an author-centric network rooted on the actor for a selected forum entry. The graph is grown out to a pre-specified depth to include, first, all other contributors to the selected discussion thread, then, iteratively, all other actors each interacts with throughout the course, linked to each actor node instance. This results in clusters of actors who contribute often to the same discussions, and small sub-networks and leaf nodes — those who interact with few others, pushed outward toward the edges. An acknowledged challenge with graph visualisation, however, is that high cross-linking quickly leads to a dense network that is difficult to read. To counter this the network is coupled with a reorderable (2D) matrix that shows the relationships between actor pairs.

Interaction Network Overview

Figure 26 shows the plot for the 257 actors taking part in 779 distinct discussion threads/trails for this run of the course. Outside the central cluster of connected individuals, 132 started a trail but either never received a response or were the only other respondents. These nodes are pushed to the boundaries of the plot, along with the 5 pairs who interacted only with each other.

Colour coding is used to distinguish:

- the first post to the forum (deep cyan — also used to encode the root node in selected discussion trails) — a lone/disconnected node,
- admin, i.e., instructors or other platform admin (nearly black),
- all (parent) nodes for which the learner started at least one trail (greyish blue),
- leaf nodes — contributed to but never started a trail (pale pastel blue).

Nodes with a deep red border represent learners who provided at least one of the entries for demographic data. Links are colour-coded based on the parent node.

Node size is mapped to number of discussions participated in, using a log scale to differentiate the relatively wide scale. The most active learner (overall) posted 74 entries to 53 distinct conversations, and can be seen at the centre of the densest part of the network. This represents a node of interest, within its wider region of interest (ROI), as a set of learners and discussions to investigate further.

Link width is mapped to the number of times a learner (child node) responded to a trail started by any one parent (including itself), using a linear scale — the maximum is 12.

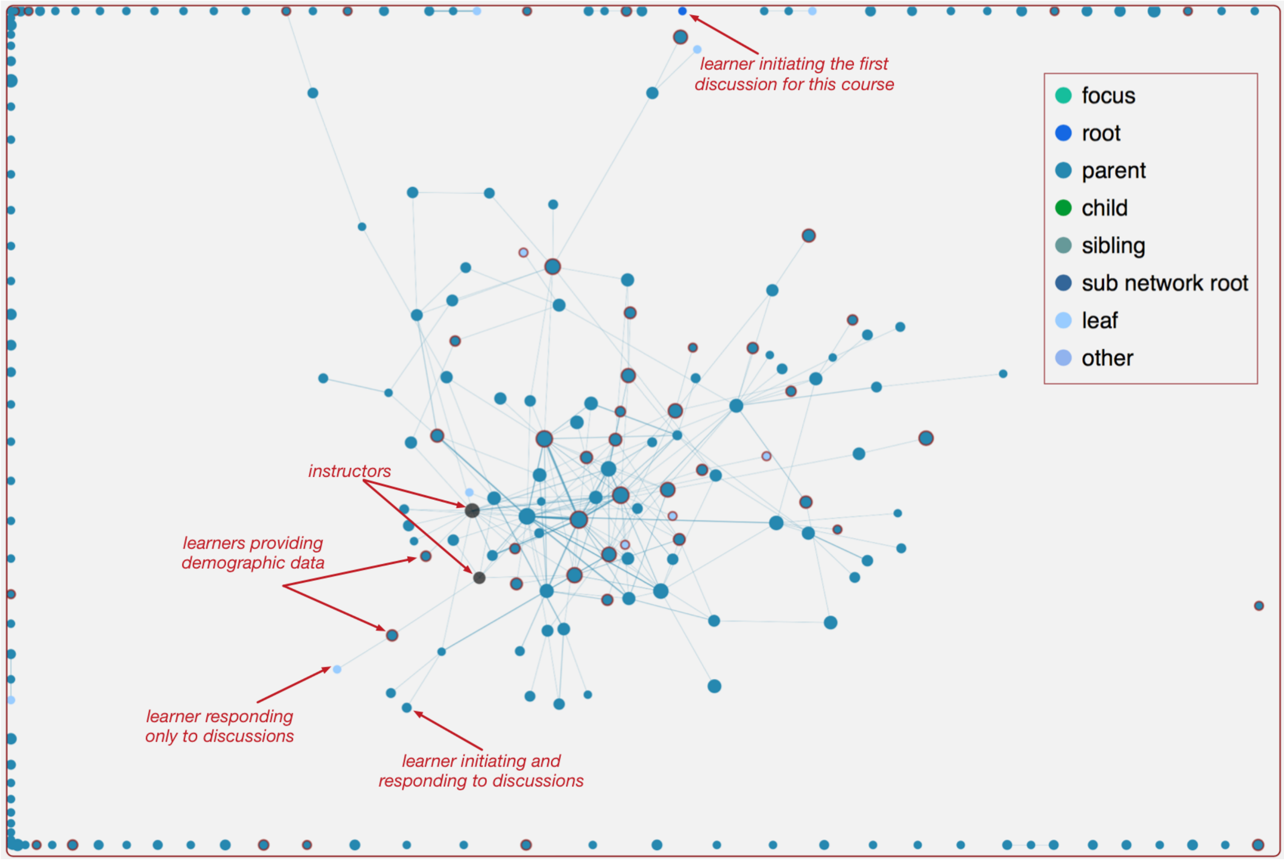


Figure 26: The interaction graph for the 255 learners and 2 instructors who took part in the online discussions. 132 "lone" participants and 5 pairs are pushed to the boundaries, outside the core that contains learners interacting with more than one other learner or represent leaves in a sub-network. The legend is superimposed on the top, right; where there is no selected/focus node parent|child|sibling do not apply - only the (start node - as root), sub-network root, leaf and other node encoding is applicable.

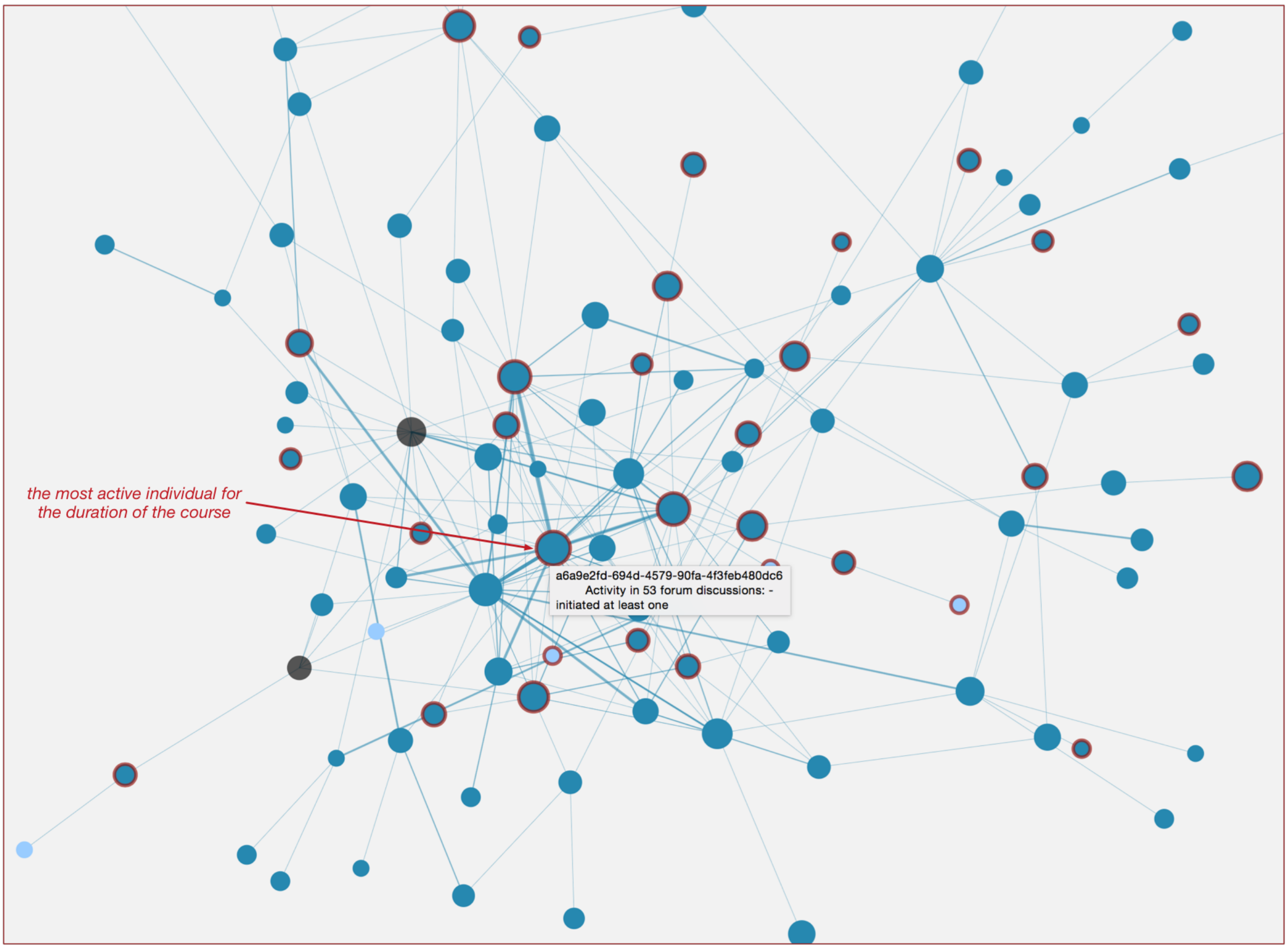


Figure 27: Zooming in to the central cluster of nodes (in Figure 26) shows the densest part of the graph, with interaction between learners also at its highest overall.

From the dot plot forum entries may be selected; this allows the interaction graph to be drawn from the discussion trail and author of interest. Selecting from an entry in the first week for the most active user, the interaction graph extends to 5 levels from this learner, and includes 103 additional learners (11 fewer than the central graph in Figure 26).[7]

The coupled matrix maps opacity to relative interaction count between node pairs. Along the diagonal opacity is mapped to total number of contributions to discussions for the node for the course. The matrix may also be reordered on a set of facets: Figures 29, 30 and 31 show reordering on *unenrolment date*, *purchase date* and total discussion trails participated in, respectively.

[7] Up to 110 nodes are obtained for between 4 and 5 levels for a range of test discussions selected, the maximum depth for each graph. These, compared to the 115 in the central connected cluster, therefore provide representative data samples to analyse in more detail.

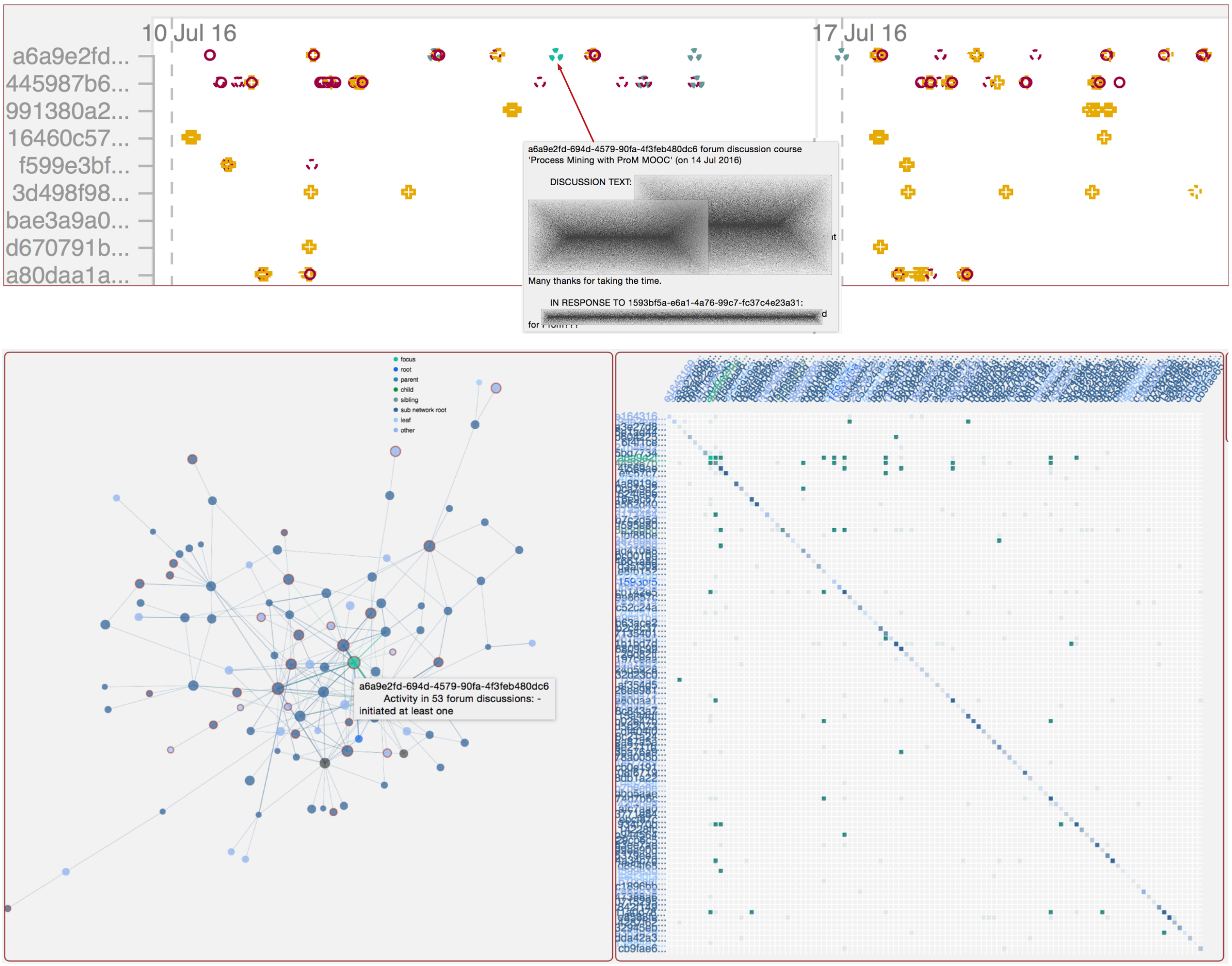


Figure 28: The interaction graph in the bottom figure, containing 104 nodes over 6 levels, results from the forum entry selected (light green) in the dot plot above. Note the discussion text is deliberately obscured for privacy.
Network odes with a dark red ring represent learners who provided demographic data. The coupled matrix is ordered by enrolment date, and maps opacity to relative interaction between learner pairs. Colour coding for both visualisations is as per the legend at the top of the graph.

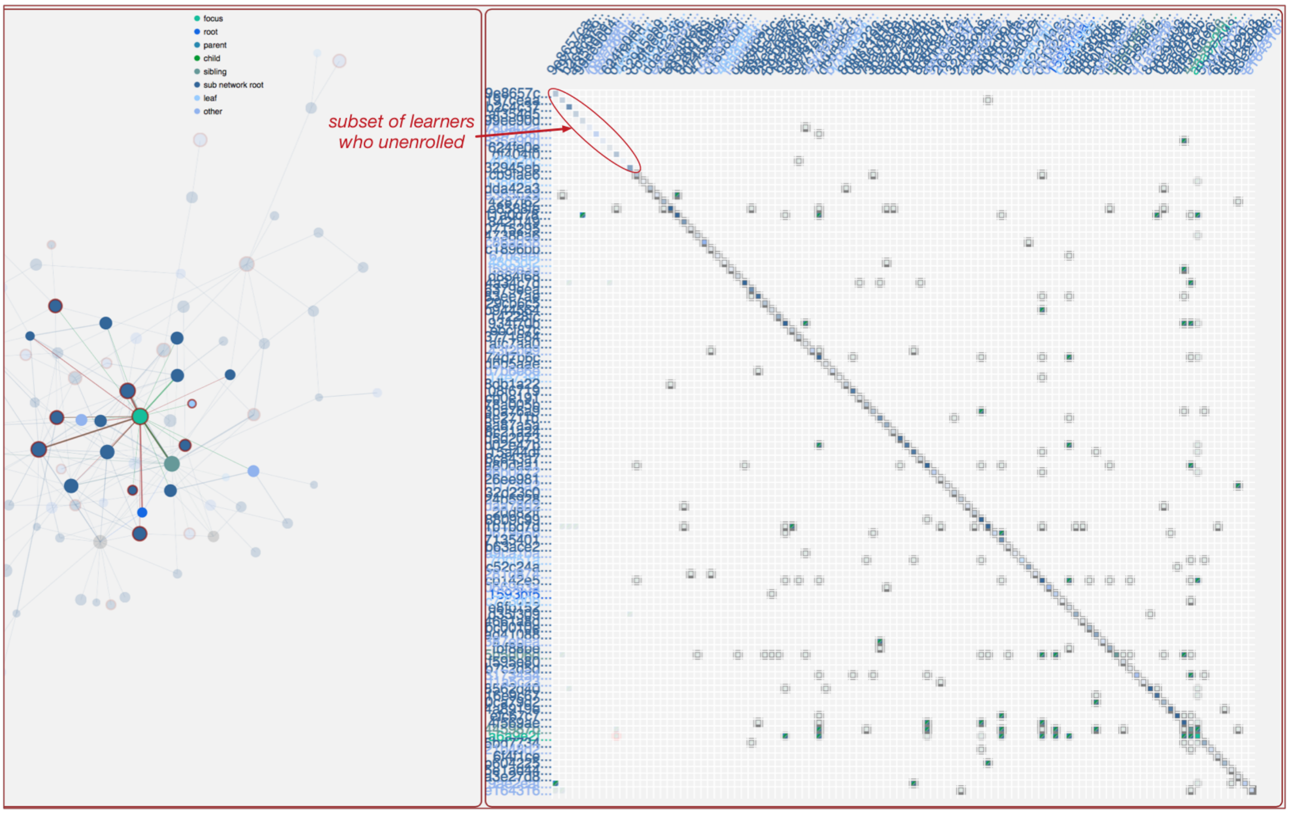
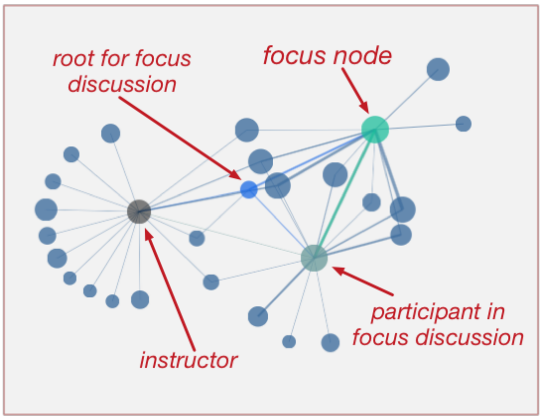


Figure 29: The sub-network of interest is highlighted, with links from the selected node to parents in brownish red and the reverse in green. The matrix is reordered by unenrolment date, and for all nodes not recording the event by latest enrolment date — the latter, most of these most active learners, have a grey border.

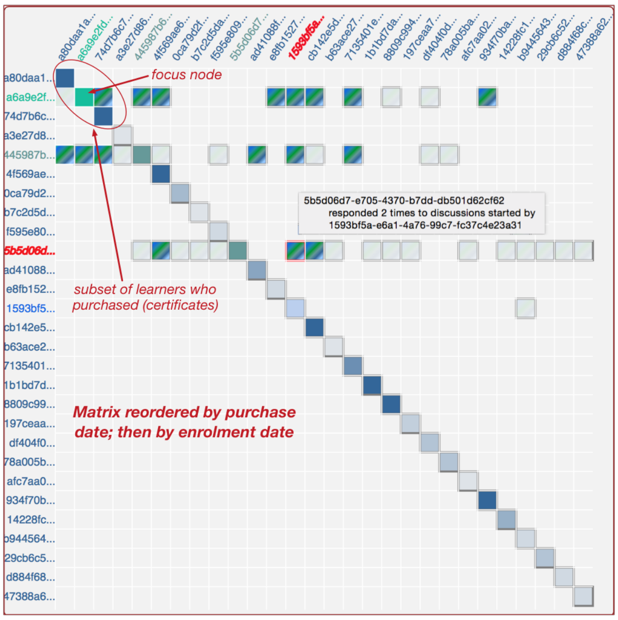
Reducing the graph in (Figure 29) to two levels from the central node (as shown on the right) allows a focus on the discussion initially selected (which includes an instructor) and only other learners interacting with this subset in the matrix. Three clusters can now be seen, connected by one other node, in addition to the discussion root.



Figures 30 and 31 show reordering on *purchase date* and total discussion trails participated in per learner. Only three here, including the focus, purchased a certificate (non-matches have a grey border).

The remainder of the plot is ordered by enrolment date — while heavier interaction falls toward the top of the plot we see some dispersion in terms of total contribution to discussions. Reordering by total discussion sees more top-heavy interaction, as would be expected. Both activity count and number of other learners interacted with (encoded as opacity) fall away toward the bottom, right.

Node-node interaction is unbalanced to show direction of interaction, for a node responding to another, with respondents on the left axis. As an example, Figure 30 shows a node pair highlighted (bold, red text and red cell border - the popup text shows the order of interaction — left to top node). While interaction appears to be predominantly directional, i.e., some learners tend to start discussions while others are more likely to respond, in some cases two-way interaction occurs (see, e.g., top, left, Figure 31). This appears to occur for the most active nodes — indicating that the more active contributors are both more likely to start a new discussion and engage in discussions started by other learners. Further analysis is required to confirm this, and any impact this has on student outcomes, as part of the content analysis to be completed.



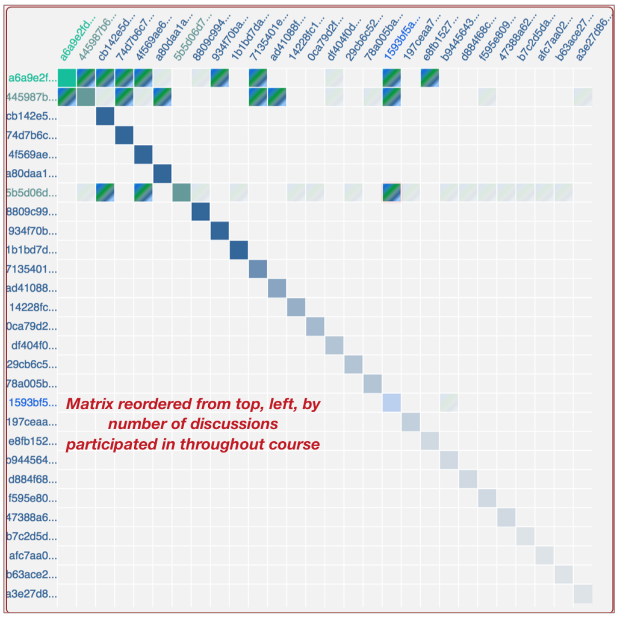


Figure 30: The graph is reduced to depth 2 for the same network and the matrix reordered by *purchase* date — with such a small percentage, however, no conclusions can be drawn about any relationship to use of the discussion forum.

Figure 31: Graph reordered by number of discussions participated in for the whole course. As expected, especially along the diagonal, node opacity fades out toward the bottom, right.

The results obtained, based on the patterns observed, are mixed. While the first set of plots clearly show a relationship between *purchase*, *full interaction* and contribution to *discussions*, once filtered down to the most active users these relationships are not straightforward. Very few of the most active users *unenrolled* from the course — an expected pattern. However, while learners who *fully participate* and/or *purchase* a certificate are most likely to be among the most active users, the reverse does not necessarily apply. Davies (2005) records similar findings — while the study showed low interaction was associated with students with failing grades, for students with passing grades, high interaction did not necessarily map to high grades.

Text analysis is on-going, to extract topics of interest and identify what mappings, if any, exist between discussion content and quizzes and other class material. Browsing ROIs we find that the disconnected nodes are often reasons given for taking the course, for instance, part of the first contribution is:[8]

> "…. *I really want to learn process mining but it is hard…*"

[8] Note text is deliberately truncated or otherwise masked for privacy, and also brevity for reporting purposes.

Other similar comments include:

"*I'm involved with a marketing firm that collaborates with a wide variety of partners. In speaking with colleagues about this course …*"

"*From the presentation on the video, I can see [...] process mining used widely and worldly. For instance; This is a tool any government could …*"

Others lone nodes, like those in connected threads, voice a problem:

"*Took some time to start up and I got error codes forcing me to reload. Now seems to be working …*"

or affirm a comment made previously, but without linking to it:

"*Thank you. Very interesting approach!*"

(Connected) threads provide a better picture of challenges learners encountered during the course and suggestions from other learners, as well as, in some cases, more specific information from instructors to resolve them. These concern setting up tools, how to approach the quizzes and specific topics presented during the course. Further analysis is required to determine why some queries do not receive a response, beyond those for which previous answers may have been provided. Further analysis will look at whether clusters are formed based on actors, topics of interest or both.

Finally, 546 learners — 11.2% of the total, provided demographic data. 55 in the 257 (21.4%) of these contributed to the discussion forum, and 31, 12.2%, of contributors to discussions, are in the central cluster of connected learners, comprising ~30% of this cluster. Analysis of this data will be used to categorise learners, to identify what additional information this provides on motivation to take the course, degree and participation, tools employed, and how this maps to students' results.

4.2.5 Summary

Our analysis has shown that many interesting insights can, and should, be gained based on the data available for a FutureLearn course. This gives crucial insight into the behavior, and therefore also intentions of learners. We were able to observe which steps were visited more often and which relatively fewer times. Such information may be used to improve steps where necessary, with an aim to improve student interaction with course material and, therefore, knowledge and new skills acquired. Similarly, by examining question statistics, we were able to observe which questions are too simple or too difficult, and therefore warrant review. As demonstrated, in investigating the learning process we see that learners do not follow the intended sequential learning path.

Finally, by examining learner-learner interaction we observe that relatively more active students, both overall and also focusing on use of the discussion forums, are likely to stay through to the end of the course. As in the learning analytics literature this points toward potential to harness discussion forums as a way to engage and retain students. Further investigation is required to conclusively determine the impact of such interaction on final student outcomes.

5. A 'Learning Analytics Framework' for EDSA

Within the EDSA project we developed an initial learning analytics workflow based on the data available to educators on the FutureLearn platform. The workflow, built using RapidMiner[9], a visual scientific workflow tool, is available at: www.win.tue.nl/~jbuijs/files/EDSA-LA demo

[9] https://rapidminer.com

report/EDSA_LAtemplate.rmp A demo report can be accessed at www.win.tue.nl/~jbuijs/files/EDSA-LA demo report.

The workflow contains routines for basic statistical processing and a selection of process mining techniques and results. The workflow was designed to be extensible, with the aim to encourage contribution of additional learning analytics techniques to further enrich the report it generates.

The high level workflow is shown in Figure 32.

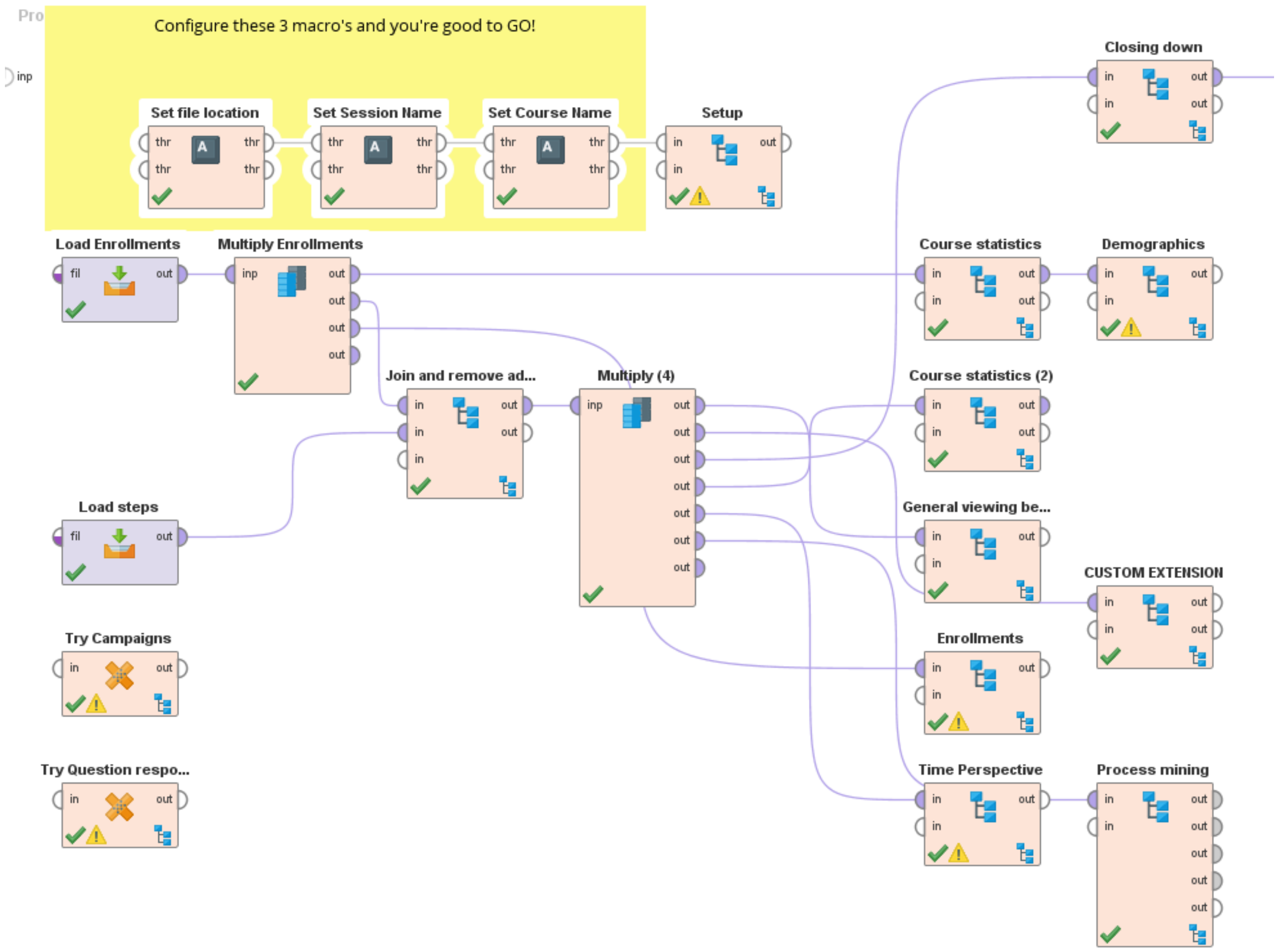


Figure 32: Overview of the RapidMiner workflow in use to analyse FutureLearn course data

The user sets the data directory, session name and course name using the three operators in the yellow box on the top, left. This loads the data into the following building blocks and starts generating results. To extend the workflow to add further insights new blocks may be built and added to the workflow.

The current workflow requires the *enrolment* and *steps* data files to be loaded, the *campaigns* and *question response* data are optional. The result is twofold: an HTML report containing key statistics of the course, as presented in section 4.2.2. The results are more detailed than that currently provided by the FutureLearn platform to educators. Several interactive results become available — discussed in section 4.2.3 — which provide process mining insights into the data. For example, a dotted chart is shown, as well as process models discovered.

We also developed an overview workflow that, given a directory of FutureLearn data files, runs this workflow on each course session and collects the derived metrics (e.g. number of students, male/female counts, start time, number of steps), in order to aid comparison between courses.

6. Conclusions

This deliverable concludes work in Task 3.4, on the “design and deployment of learning analytics”. D3.5 continues on from the work reported in D3.3, to investigate student interaction with online learning material, a topic of interest to the EDSA project and the learning analytics community.

To support the learning analytics task a set of text analysis visual analysis and presentation modules for both datasets were developed and/or extended from existing tools and APIs. Visual exploration and quantitative analysis of the Videolectures.NET portal log files was carried out, from the perspectives of the end user and the lecturer. The log analytics tools provided insight into user behaviour, and the explorer provided information on the content and semantics of lectures given, topics addressed and authors of the learning material. Process mining and visual analytics provided insight into evolving student behaviour over the course of the the FutureLearn MOOC. This provided also insight into user (student) behaviour, mapped onto the structure developed by course instructors.

The information obtained in both cases may be fed back to course presenters to improve course design and delivery. Further, by tailoring the information collected to individual students and the cohort as a group we aim to foster interaction with and between students beyond the formal presentation of course material, as a way of increasing engagement and peer learning, and therefore value obtained from taking part in online learning.

Key learning points from the analysis reported in D3.5 include areas to investigate further, beyond the project, in order to improve student interaction with course material and with other students and instructors. This is important in contributing to acknowledged challenges in the field of learning analytics — among others, retaining engagement and effectively supporting students with varied demographics and motivation for online learning, so that both students and instructors obtain higher benefit from remote, online learning courses and independent learning materials.

The work carried out through T3.4 has fed into finalising a framework that should serve to improve course delivery, by providing instructors with a simple means to obtain an overview of student interaction with a course, based on a set of predefined metrics. While this is currently set up for the FutureLearn platform the approach taken allows extension of the workflow to other online learning datasets.

6.1 Outlook

Over the course of the project we have been able to analyse a number of online and hybrid courses and other learning material, giving us insight into different approaches to presenting courses to a variety of students, and into student interaction with the learning material. At the close of the EDSA project we have built a framework that may provide a guide for instructors, by providing an overview of student behaviour and outcomes, mapped to course design, information that may be fed into improving course delivery and designing new courses.

Beyond the project, we are currently working on submission of a publication to a journal or to the LAK or TEL conferences, based on the joint work in T3.4.

7. References

[Buckingham Shum 2012] Buckingham Shum, S., & Ferguson, R. (2012). Social Learning Analytics. Educational Technology & Society, 15(3), pp.3–26.

[EDSA Deliverable D3.3] Aba-Sah Dadzie, Inna Novalija, Rémi Brochenin, Joos Buijs, & Alexander Mikroyannidis (2016). EDSA Deliverable D3.3 Report on the Evaluation of Course Content and Delivery 1.

[EDSA Deliverable D3.4] Angi Voss, EDSA Deliverable D3.4 Report on the delivery of video-lectures, webinars and face-to-face trainings.

[Davies 2005] Davies, Jo, & Graff, Martin (2005). Performance in e-learning: online participation and student grades, British Journal Of Educational Technology, 36(4), pp. 657–663,

[Wise 2014] Alyssa Friend Wise, Yuting Zhao, & Simone Nicole Hausknecht (2014). Learning Analytics for Online Discussions: Embedded and Extracted Approaches. Journal of Learning Analytics, 1(2), 48-71.