Big Data and Education

Alex Poulovassilis
Birkbeck Knowledge Lab
Birkbeck, University of London, 5/12/2016

WHAT IS BIG DATA?

Emergent web, mobile, and pervasive digital technologies are generating data at unprecedented scales and speeds in virtually all areas of human activity. Across industry, commerce and the public sector this 'big data' is being digitally collected and computationally analysed in order to gain better understanding of providers' services and products, consumers' needs and preferences, and, more fundamentally, to expand human knowledge across the sciences, social sciences and humanities.

Originally 'big data' was taken to mean data sets that are beyond the management and analysis capabilities of traditional software tools. The generation of such data sets led to the development of new data storage and data processing paradigms, such as NoSQL data stores (Cattell 2011), massively data-parallel distributed processing frameworks (Dean & Ghemawat 2008, EMC 2015) and cloud computing platforms (Armbrust et al 2010).

Big data is distinguished from other data by exhibiting the so-called 'V' attributes, including:

- *volume* – the size of the datasets;
- *velocity* – the rapid rate at which the data may generated;
- *variety* – different types of data being generated from multiple sources, needing to be cross-referenced and combined in order to be fully exploited;
- *veracity* – the incompleteness of the data being collected, and the imprecision of inferences being made from it;
- *volatility* – data being collected or inferred may become less relevant over time.

More recently, there is a recognition that these 'V' attributes are not the whole story and that what is most important is the ability to extract *value* from such data while also complying with given time, human and technical resource constraints.

Turning specifically to big data in the Education sector, the field of **Learning Analytics (LA)** is concerned with *gathering, analysing and visualising* data about learners and learning processes so as to increase stakeholders' understanding of these, and hence *to improve learning and the environments in which it occurs* (Siemens 2012, Drachsler and Greller 2012, Ferguson 2012). This data may be collected from many different sources:

- virtual learning environments (VLEs) that track and support students' activities, interactions, reflections and progress through learning tasks;
- students' assessment activities – both formative and summative;

- students' personal records and records of prior achievement;
- learner profiling and learner modelling software;
- software supporting social networking, peer support, collaboration;
- audio and video recordings; gesture and physiological sensor recordings (e.g. heart rate, galvanic skin response, blood pressure, EEG readings);
- mobile learning apps, gathering large-scale user-centered and context-aware data.

This exceptionally broad range of data sources is allowing increasingly *individualised*, *detailed* and *longitudinal* data to be collected and analysed, bringing with it the potential to derive new insights and to provide more effective support to learners and tutors.

The field of **Educational Data Mining** (EDM) was established a few years earlier than the LA field and it, too, is concerned with gathering and analysing data so as to understand, support and improve students' learning. The LA and EDM fields have somewhat different emphases (see Siemens and Baker 2012):

- Tools to aid users in their roles (LA) as opposed to tools for automated knowledge discovery (EDM).
- Understanding learning processes as a whole (LA) as opposed to understanding specific aspects and the relationships between them (EDM).
- Tools that empower students, learners, teachers and other stakeholders to make decisions (LA) as opposed to automated personalisation and adaptation of a learning environment (EDM).

None the less, there is also much commonality between LA and EDM and they can indeed be regarded as parts of a larger interdisciplinary continuum of research and practice involving disciplines such as computer science, education and psychology, as well as teachers, learners, learning designers, policy makers and other stakeholders in learning processes from across the public and private sectors.

There is also much commonality in the computing techniques developed and applied in the LA and EDM fields. Some of the key computing techniques are listed below, and more information can be found in the references cited:

- data modelling: designing the way that data is represented within computing systems so as to facilitate the types of processing that needs be applied to the data (Chen et al 2012);
- data cleansing, transformation, integration: correcting errors and inconsistencies in the data; transforming data so that it can be more easily integrated – including converting it to expected standard formats; and creating integrated data resources that combine data from different data sources (Chen et al 2014);
- distributed data processing: designing data processing algorithms that will scale to the volumes and velocities at which data is being generated (EMC 2015);

- semantic modelling and reasoning: representing knowledge domains using specialist ontologies[1]; using ontology-based reasoning to provide personalised information to users (Siemens 2012);
- data mining: including classification, clustering, Bayesian reasoning, rule and pattern extraction (EMC 2015);
- data analytics and visualization: supporting analysis and visualization of data in ways that are useful to users;
- human-computer interaction (HCI): designing interfaces to computer systems that are appropriate for their intended users;
- learner modelling: this may use experts' knowledge, computational inferencing methods, or combinations of both approaches (Li et al 2010, Koedinger et al 2013);
- recommendation algorithms, for example to recommend a course or a learning resource to a user (Manouselis et al 2011);
- predictive modelling, for example to detect students who may be at risk of dropping out of a course so as to guide them towards resources or people who might help them (Clow 2013);
- social network analysis and discourse analysis: can be used to discover connections between learners, tutors and learning resources (Dawson 2008, De Lido et al 2011).

Figure 1 shows a general three-tier architecture for managing and exploiting data relating to learning processes and learners, showing the tier in which each of the above computing techniques might typically be utilised. We will see example applications of some of these techniques in the next section.

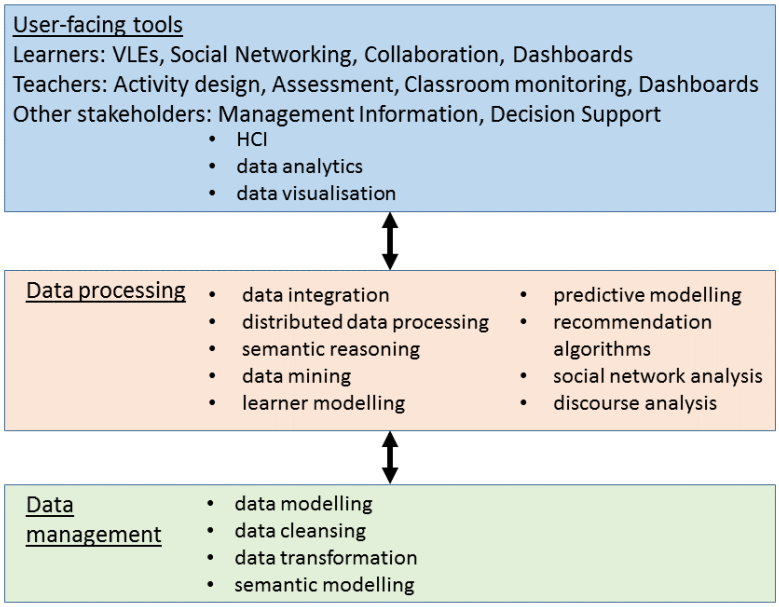

Figure 1. Data architecture and Computational Techniques

[1] See https://www.w3.org/standards/semanticweb/

WHAT IS THE POTENTIAL OF BIG DATA IN EDUCATION?

Collection and analysis of learning-related data has been used in Technology Enhanced Learning research and practice for many years. For example, *Intelligent Tutoring Systems* (ITS) are able to guide students through instructional learning activities by maintaining a computational model of the student's knowledge and skills, and using this to provide adaptive feedback to students relating to their progress on the task set and possible next steps (Koedinger et al 2013).

The recently completed LIBE Virtual Learning Environment is an ITS that also aims to support more open-ended inquiry-based learning activities. The LIBE project – "Supporting Lifelong Learning with Inquiry Based Education" (see libeproject.it/?lang=en) aims to offer young adults (16-24 years old) who may be at risk from exclusion from training and employment personalised e-learning courses that target four transversal competences: literacy, numeracy, IT literacy, and problem solving in technology-rich environments.[2] The LIBE VLE is configured with an extensible set of Learning Objectives (LOs) relating to each competence, and each activity within a course is associated with a subset of these LOs. Before a student starts on a course, a short pre-test gathers information about the student's attainment levels in the set of LOs that are targeted by that course. As the student engages in each activity, these attainment levels are updated according to the student's performance. The student's current attainment levels are used to offer personalised content to the student – for example, different levels of explanation of specialist terminology, and tests at differing levels of difficulty. Knowledge of students' attainment levels is also used to offer different levels of hints to students who may be struggling with completing an activity. Figure 2 illustrates the six currently available courses. Figure 3 shows a multiple-choice quiz from one of the courses. Figure 4 illustrates a more open-ended activity – specifically, a simulation – from another course.

Figure 2. The six courses developed by the LIBE project.

[2] The LIBE project was funded by the Lifelong Learning Programme of the European Commission, ref. no. 543058-LLP-1-2013-1-IT-KA3-KA3MP. The author thanks all partners on the project for sharing the images shown in Figures 2, 3, 4.

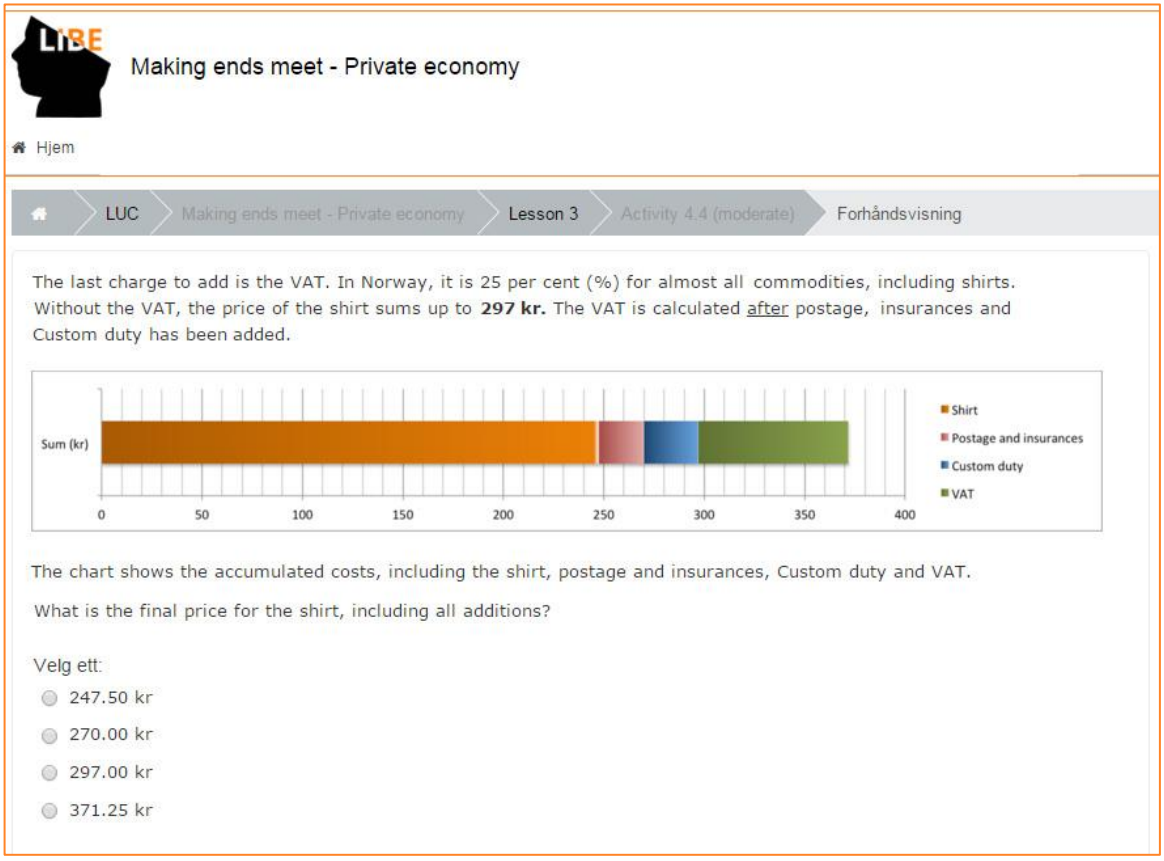

Figure 3. An example of a multiple choice quiz from the "Making ends meet" course, targeting the learning objectives "understanding of numberlines" and "calculation of percentages".

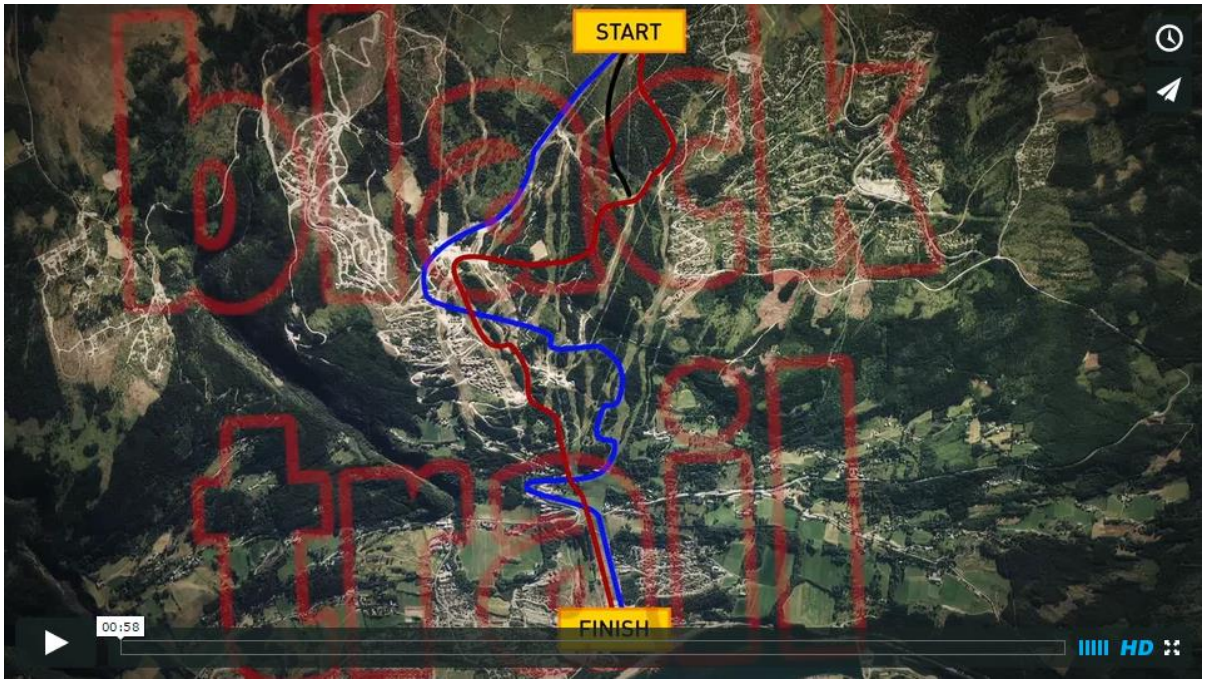

Figure 4. An example of an open-ended activity, relating to the "Mountain Biking" course. The challenge is to spend as little time as possible from the top to the bottom in a downhill bike simulation, by making smart

decisions about which trail to follow, what size bike to use, how fast to go etc. Based on their selections, and their solutions to associated maths problems, a simulation of the resulting downhill ride and the finishing time are presented using real video footage and maps. The learning objectives include "estimating distances by reading a map", "understanding the relationship between diameter and circumference", "calculating with time, distance and speed".

The six courses developed during the LIBE project were designed after conducting extensive focus groups with students and their tutors across three European countries (Italy, Norway, Portugal). Each course presents a topic that is universally appealing to young people. The courses are designed so that participants can experience increased mastery of skills as they progress through each course. The provision of personalised explanations, tests and hints aims to treat students as individuals, each with their own evolving profile of knowledge and skills, further boosting students' engagement, motivation and self-confidence. The fact that this adaptation is purely computer-generated can help to mitigate against any unconscious biases in tutors, particularly for assessment activities. For more information about LIBE, please see the LIBE e-Booklet at libeproject.it/?p=888&lang=en and more generally the extensive set of resources on the LIBE website.

Over the past decade the range of educational software has expanded considerably and now includes environments that aim to fully support exploratory learning activities. Examples of such *Exploratory Learning Environments* (ELEs) are simulators, virtual labs, microworlds and educational games. Unlike traditional ITSs, ELEs give considerable freedom to students, who may tackle the task they have been set in a variety of different ways. The tasks students are asked to undertake are open-ended in nature, may have many alternative solutions, and encourage students to follow a variety of solution approaches. ELEs aim to increase students' engagement with learning, and to foster 'deeper' learning of concepts that can be reapplied to solving new problems. Increasing a student's engagement with learning can in turn lead to improved learning outcomes as the student becomes more motivated to apply and extend their newly acquired skills and knowledge.

Research has found that considerable guidance is required to ensure learning in such open-ended contexts, but that with the provision of appropriate support ELEs can lead to more engagement and deeper learning. However, designing such support within the ELE presents two major challenges. Firstly, since the learning tasks are generally open-ended, there is not a single 'correct' answer and balance needs to be struck between allowing students the freedom to explore alternative solution approaches on the one hand and guiding them towards achieving the intended learning goals on the other. Secondly, providing support for the student is not enough; there is a need also for tools providing *support to the teacher* so as enhance the teacher's awareness of the classroom 'state' and of students' engagement and progress on the tasks set. Without such tools, a teacher can only be aware of what a small number of students are doing at any one time, and it is hard to keep track of which students are making good progress, who is off-task, and who is in difficulty and in need of additional support.

Recent research has sought to address these two challenges. Firstly, intelligent components can be designed that provide personalised, adaptive feedback to students as they are working on an exploratory learning task. This feedback is based on the detection of

a set of significant '*indicators*' as students interact with the ELE, combined with the ELE's knowledge of students' recent history of solution approaches, interactions, achievement of learning goals, affective state etc. Multidisciplinary teams involving pedagogical experts, learning designers and computer scientists can work together to identify what are significant indicators in a given learning setting. Appropriate computational techniques can then be designed and embedded into the ELE in order to detect the occurrence of these indicators, update students' learner models, and generate feedback for students.

To illustrate, Figure 5 shows a mathematical microworld called eXpresser that aims to support 11-14 year olds' learning of algebraic generalization[3]. Using eXpresser, students are asked to construct two-dimensional tiled models and associated algebraic rules. In order to build their model, students need to create 'building blocks' out of unit-square coloured tiles depending on their perception of the model's structure, and to repeat each building block in order to form a 'pattern' which forms part of their overall model. The algebraic rules they are asked to construct relate to the number of tiles of each colour required to paint each pattern and their model overall. By the end of the task students need to ensure that their models and rules are fully general, which they can accomplish only by 'unlocking' numbers to turn them into variables.

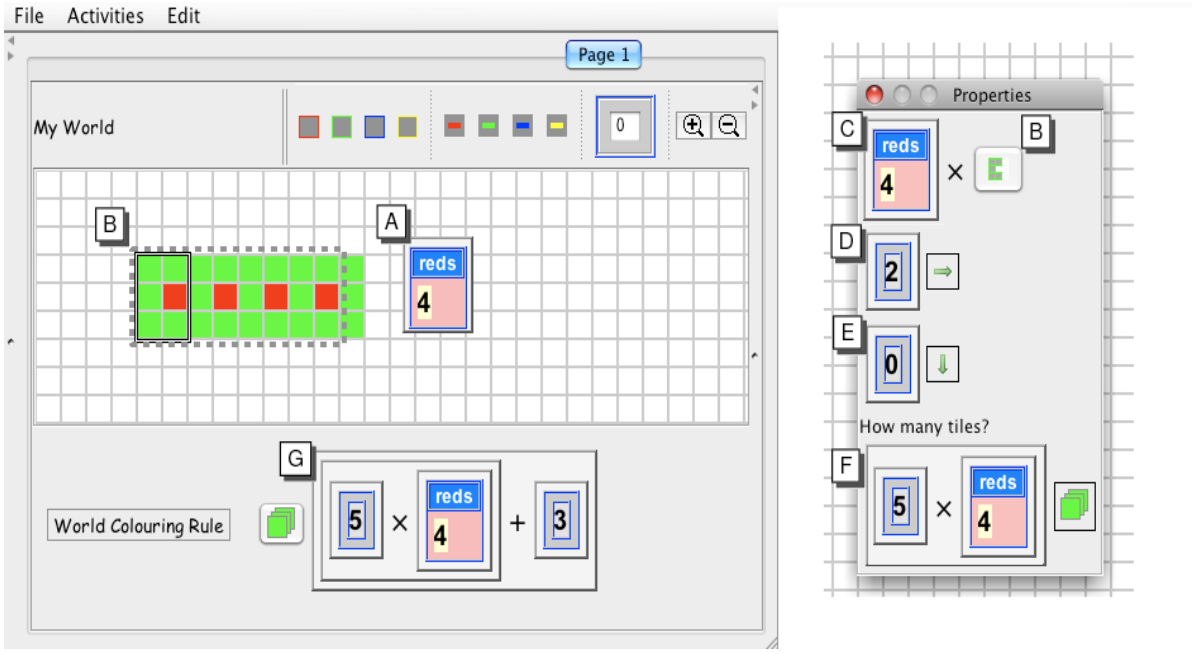

Figure 5. The eXpresser microworld. Letters highlight the main features: (A) An 'unlocked' number is given the name 'reds' and signifies the number of red (dark grey) tiles in the pattern. (B) Building block to be repeated to make a pattern. (C) Number of repetitions (in this case, the value of the variable 'reds'). (D,E) Number of grid squares to translate B to the right and down after each repetition. (F) Units of colour required to paint the pattern. (G) General expression that gives the total number of units of colour required to paint the whole pattern.

[3] eXpresser is one of a set of tools making up the MiGen system, which was developed through funding from the ESRC/EPSRC Technology Enhanced Learning programme, award no. RES-139-25-0381

Figure 6 illustrates a feedback message from the eXpresser to a student who has constructed a correct pattern and a correct colouring rule for it, nudging the student towards unlocking a number so as to now generalise their pattern and rule. A bit later on, Figure 7 shows a message of encouragement but also a stronger prompt to guide the student towards generalising their construction.

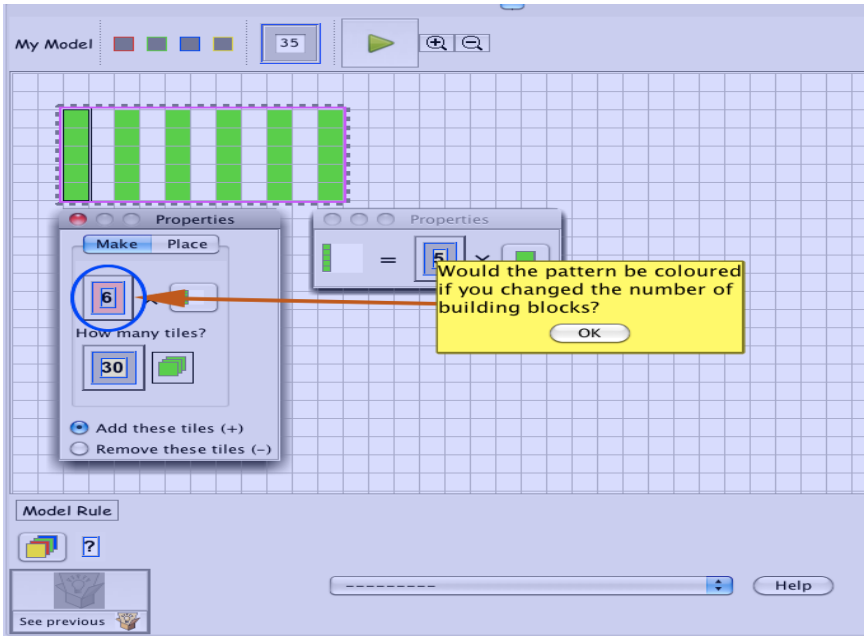

Figure 6. . A 'nudge' from the eXpresser

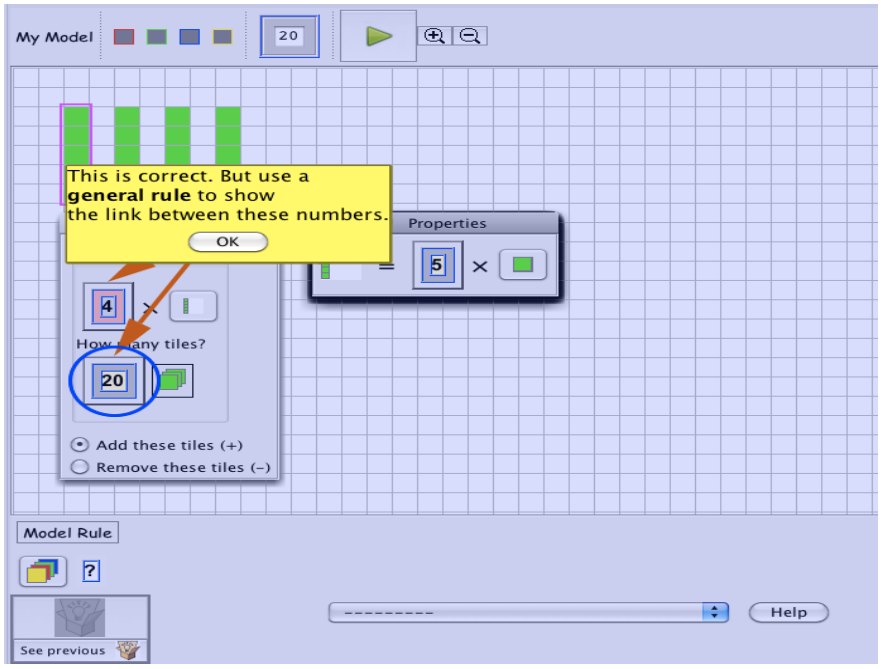

Figure 7. A message of encouragement and a stronger 'prompt' from the eXpresser

The 'indicators' that underpin the provision of such feedback to students in eXpresser are detected through a combination of case-based reasoning, rule-based reasoning and similarity metrics (see Gutierrez-Santos et al 2012 for technical details and additional references). The content, timing, and presentation of the feedback messages were determined through a series of 'Wizard-of-Oz' studies with groups of students in their classrooms, in which the system's intelligence is initially emulated by a human facilitator who supports the student remotely. Successive cycles aim to replace interventions by the human facilitator with machine-generated feedback (for details of these studies and additional references see Mavrikis et al 2013).

As another example, Figure 8 shows a microworld called FractionsLab that aims to support 8-12 year olds' learning of fractions[4]. Students are asked to construct one or more fractions and, using the affordances of the system, to compare, add or subtract fractions. In this case, the student has been asked to create a fraction, and then to create four equivalent fractions with increasingly larger denominators. So far the student has created their first fraction, but has not yet created any equivalent ones. The glowing lightbulb at the top of the screen indicates that there is help currently available from the system. If the student wishes to see it they must click on the lightbulb; this is what is termed 'low-interruption' feedback (Grawemeyer et al 2015). In this case, clicking on the lightbulb results in the feedback message shown in Figure 9, which is aiming to nudge the student towards the next step. Figure 10 shows that the student has indeed made their first equivalent function. After a period of inactivity, Figure 11 shows a message of encouragement and also an unsolicited prompt – termed 'high-interruption' feedback (Grawemeyer et al 2015) – to guide the student towards the next step.

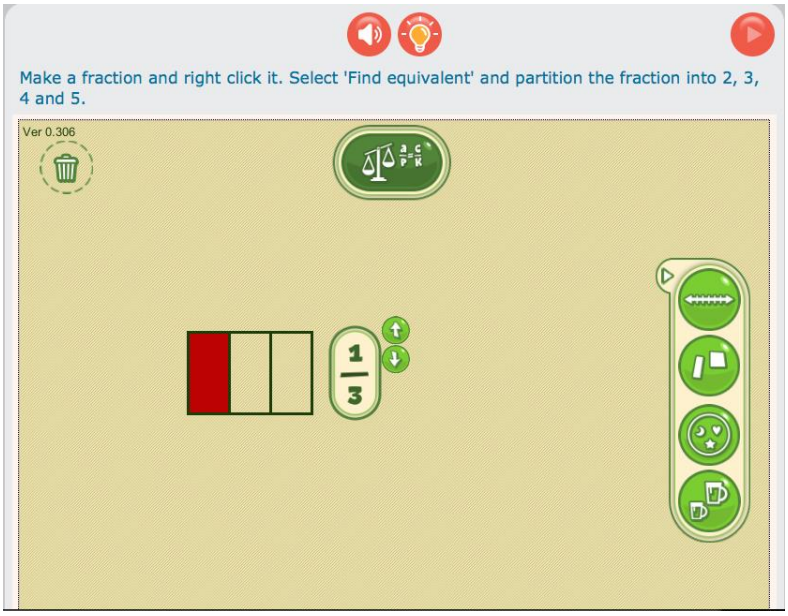

Figure 8. FractionsLab microworld, showing the availability of low-interruption feedback.

[4] FractionsLab is one of a set of tools making up the iTalk2Learn system, which was developed through funding from the EU FP7 programme, ref. no. ICT-318051. The author thanks the iTalk2Learn team for sharing the images shown in Figures 8-11.

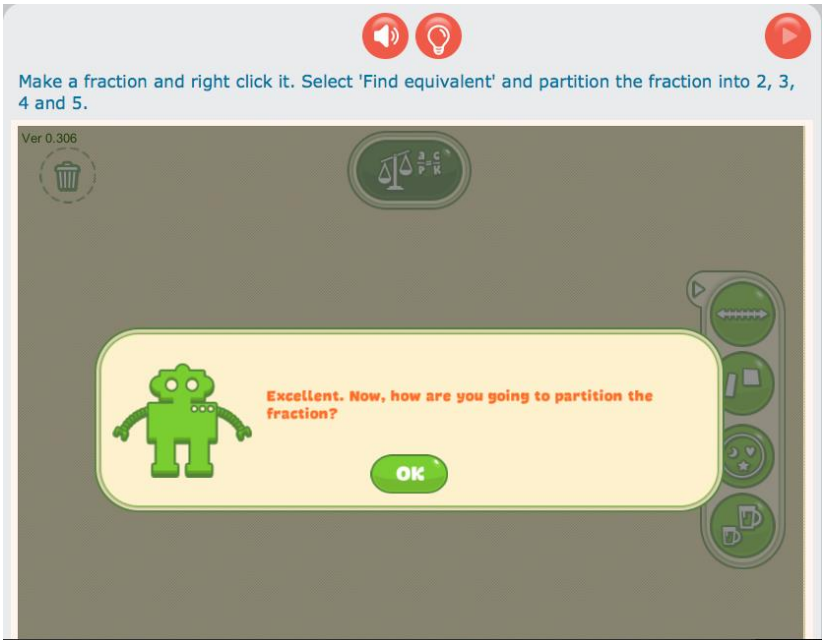

Figure 9. FractionsLab microworld, showing the elective display of low-interruption feedback.

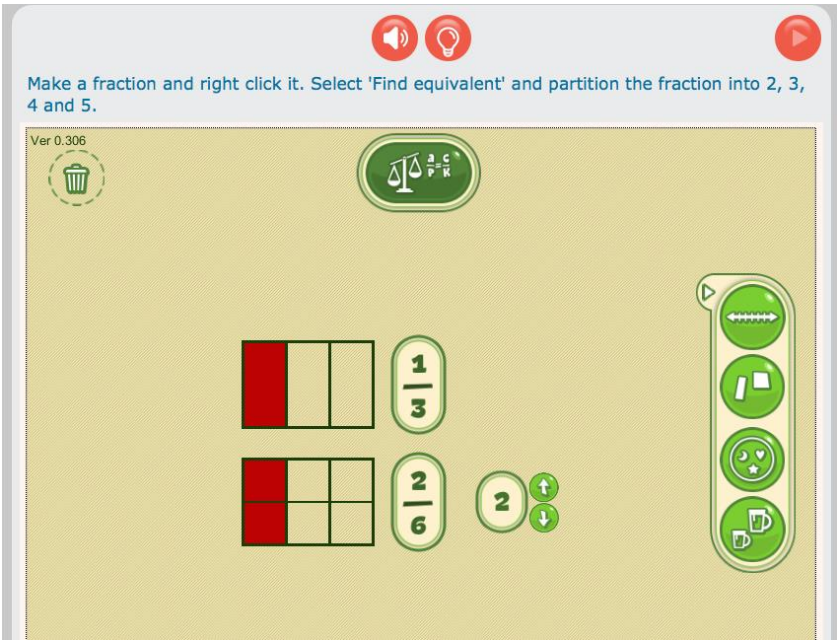

Figure 10. FractionsLab microworld, showing that the student has progressed to the next step.

Figure 11. FractionsLab microworld, showing a message of encouragement and also a stronger prompt to guide the student towards accomplishing the subsequent step.

Detection of students' *affective state* can further enhance learning, by means of nudges that move students out of negative states such as boredom or frustration that inhibit learning into positive states such as engagement or enjoyment. Affective state can be detected through analysis of speech, facial expressions, eye tracking, body language, physiological data, or combinations of these (D'Mello and Kory 2015). In the iTalk2Learn system, students' affective state is determined through detection of keywords and prosodic features in their speech as they talk aloud when interacting with the system (Grawemeyer et al 2015b, Grawemeyer et al, in press). Giving messages of encouragement improves the affective state of students who are struggling (D'Mello et al 2012). Recent research in the iTalk2Learn project has found that high-interruption feedback is more effective than low-interruption feedback when students are in a negative affective state; but that if students are in a positive affective state then low-interruption feedback is preferable (Grawemeyer et al 2015). We have recently remodeled the system-student interaction data arising from iTalk2Learn using graph-based methods so as to more easily investigate the effectiveness of the intelligent support being provided by the system. Figure 12 illustrates one possible visualization of a how a student's affective state changes during a learning task.

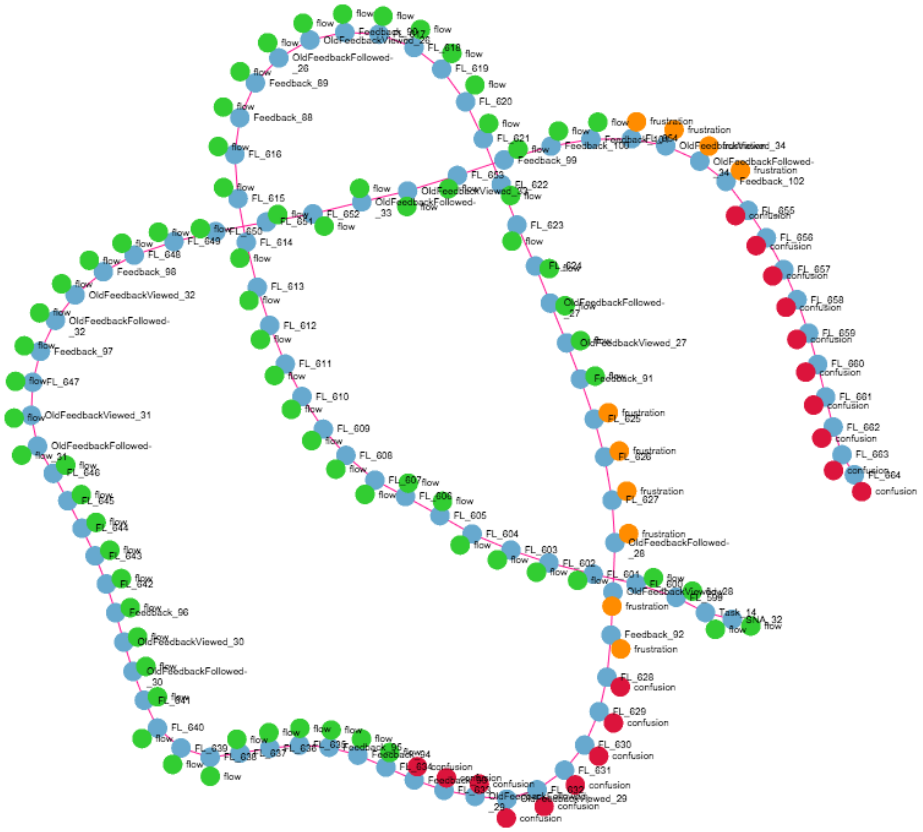

Figure 12. Graph-based modelling and visualisation of students' interactions, illustrating how a student's affective state changes between states of Engagement (green), Frustration (amber) and Confusion (red). Successive events are shown in blue and linked to each other by red edges.

A further benefit of providing intelligent feedback to students through the ELE, as illustrated in Figures 6-11, is that it frees up the teacher to provide additional, more complex or nuanced, guidance that is beyond the capabilities of the ELE's computational intelligence. It also helps the teacher to support larger classes.

We can see from the above descriptions that the data gathered and generated by intelligent ELEs such as MiGen (Noss et al 2012) and iTalk2Learn (Grawemeyer et al 2015b, Grawemeyer et al, in press) are many and varied. The data include:

- Event-based data: log data of students' actions in the ELE; students' reflections, e.g. through text (in MiGen) or speech (in iTalk2Learn); occurrence of key indicators as students interact with the ELE; generation and provision of feedback by the ELE.
- Students' constructions: the models and mathematical expressions being constructed by students, including a full history of how each was constructed.
- Task information: task descriptions, task learning goals, common solution approaches to each task.
- Learner models: information about students' level of attainment of concepts and skills, recent history of interactions with the system, progress with tasks set, achievement of learning goals, affective states.

This data exhibits all of the 'V' attributes that we discussed earlier. As well as its evident volume and velocity, under the 'variety' attribute we have unstructured data (e.g. the students' reflections), semi-structured data (e.g. the log data, task information, and students' constructions) and structured data (e.g. the learner models and indicator data). Under 'veracity' there is the inherent imprecision of the inferences being made by the system's intelligent components, e.g. in the detection of task-dependent indicators (Gutierrez-Santos et al 2012) or students' affective states (Grawemeyer et al 2015). Under 'volatility', a student's history of interactions, inferred indicators and affective states may become less relevant with time.

The rich range of data that can be collected and inferred by an ELE provides not only the possibility to generate personalised feedback for the learner, as discussed earlier, but also the opportunity to design visualisation and notification tools for the teacher. We noted earlier that there is indeed a need to provide tools for the teacher when ELEs are used in the classroom in order to enhance the teacher's awareness of the classroom state, students' progress on the task set, and students' achievement of learning goals. The provision of such awareness information can help the teacher to formulate her own interventions to support both individual students and the class as a whole. To be fully effective in the classroom, such tools need to be designed by multi-disciplinary teams involving teachers, pedagogical experts and computer scientists. In our own work in this area, we have used an iterative participatory methodology, comprising successive phases of prototyping, requirements elicitation, incremental development and evaluation (for details see Gutierrez-Santos et al 2012, Mavrikis et al 2016).

To illustrate, Figures 13 and 14 show two of the Teacher Assistance tools we designed for the MiGen system: the Classroom Dynamics (CD) tool and the Goal Achievements (GA) tool. In the CD tool, each student present in the classroom is represented by a circle containing their initials. At the outset of the lesson, the teacher can drag-and-drop these circles so that their position on the screen reflects the students' spatial positioning in the classroom. The colour of a student's circle reflects the student's current activity status, as inferred by the system. Green indicates a student working productively on the task set. Amber indicates a student who has not interacted with eXpresser for some time (by default, five minutes). Red indicates a student who has requested help from the system in a situation where the intelligent support cannot help any further: in such cases the eXpresser displays the message "The teacher will come to help you now" to the student, and the student's circle becomes coloured red to attract the attention of the teacher.

Most of the time, the teacher will have the CD tool selected for display on her handheld computer. When students show as amber, she can approach them and encourage them to resume working on the task set. If students who are not showing as red call out for help she can encourage them to first seek help from the system, knowing that if the intelligent support cannot help the student's circle will automatically appear as red in the CD tool. If a student does appear as red, the teacher can click on the student's circle on her way over to the student so as to see their current model and rule, which helps her to prepare her feedback for the student.

From time to time, the teacher will consult also the GA tool. This comprises a tabular display of students and task goals. Each row of the table shows the progress of one student (identified by their initials) in completing the task goals. A white cell indicates a goal that has not yet been achieved by the student. A green cell indicates that the goal is currently being achieved by the student's construction. An amber cell indicates that the goal was achieved at some point, but is not currently being achieved by the student's construction Knowing which students have accomplished all the task goals allows the teacher to set them additional activities, for example comparing their construction approach with that of a peer (see below). Other students may be advancing more slowly towards completing the task goals; knowing this allows the teacher to set them homework so that they can catch up with their peers. If the GA tool shows that many students are not achieving a particular task goal, the teacher can interrupt the lesson to help all students at the same time.

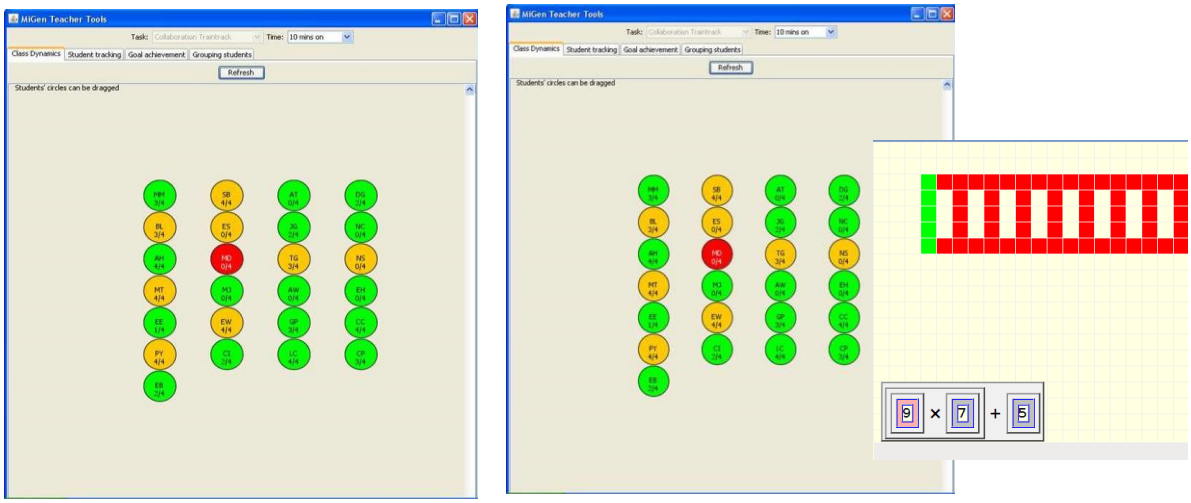

Figure 13. MiGen's Classroom Dynamics tool. On the left, a classroom with the students sitting at benches in rows. On the right, the teacher has clicked on the 'red' student to see their construction and rule on the way over to help them.

Figure 14. MiGen's Goal Achievements tool. We see that some students have achieved all or most task goals, some students have not made any progress yet, and some students are moving back and forth.

Another of MiGen's teacher tools – the Grouping Tool (GT) (Gutierrez-Santos et al, in press) – supports the teacher in managing group discussion activities after students have finished their individual construction activities, by automating the pairing of students based on their constructions. To generate fruitful discussions, students with *dissimilar* constructions are paired together (since discussing your construction with a student who has solved the task in a similar way will not result in much additional reflection or insight). Identifying appropriate pairs would be very time-consuming for the teacher to do manually during a lesson: it would require the teacher to investigate every student's construction, identify pairs of constructions that are sufficiently dissimilar, and then put the students into pairs, taking also into account interpersonal factors. The GT is designed to aid the teacher in this task by automatically generating an initial set of pairings, aiming to minimise the overall similarity across all pairings. The degree of similarity of two constructions is calculated by the system by comparing aspects such as the building blocks used, the positioning of building blocks within patterns, and the number of variables used (see Noss et al 2012, Gutierrez-Santos et al In Press). The proposed pairings are presented visually to the teacher, who can then confirm or change each pairing – see Figure 15 (we note that in the case of an odd number of students, one of the 'pairings' generated will be a triplet!). In the

GT, students are represented by their initials within a circle. The degree of similarity between pairs of constructions is represented by a small green rectangle for low similarity; medium-sized yellow rectangle for moderate similarity; or large red rectangle for high similarity. The teacher can select students' circles and drag them into different groups in order to change the pairings suggested by the system so as to take into account factors that are beyond the system's knowledge, such as students' interpersonal relationships.

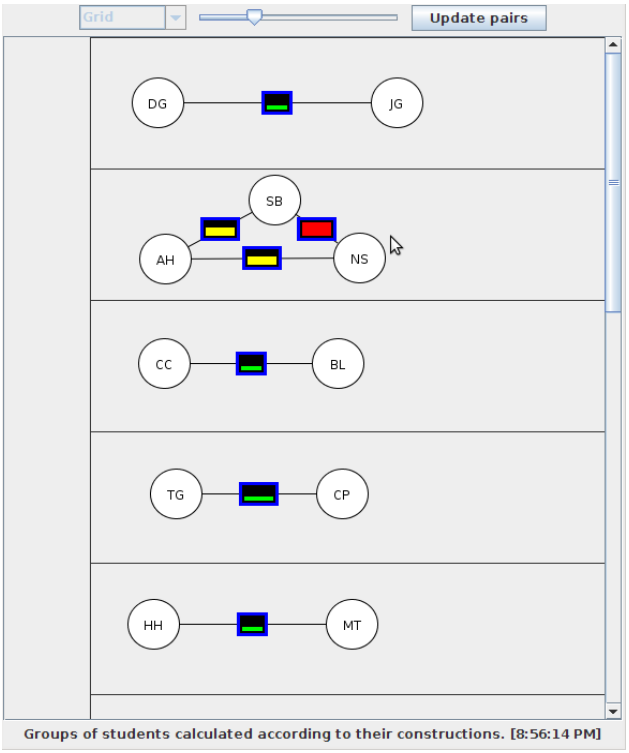

Figure 15. MiGen's Grouping Tool.

Another of MiGen's teacher tools – the Student Tracking (ST) tool (see Gutierrez-Santos et al 2012) – provides the teacher with information about all occurrences of significant indicators as students are working on a task using eXpresser. The ST tool can be consulted by the teacher during the course of a lesson in order to examine in more detail a particular student's solution approach and interactions with the system, for example so as to formulate more detailed feedback for that student than would be possible by consulting only the CD and GA tools. The information displayed by the ST tool can also be viewed after the lesson, for after-class analysis of what students have achieved, how they have approached the tasks set in the lesson, and how they have interacted with the eXpresser microworld. This information can inform the teacher's support of individual students and the class as a whole in the next lesson.

The immediacy of the information available through MiGen's teacher tools can help teachers formulate their interventions during the current lesson, set additional homework, plan the next lesson, as well as adjust the design of future tasks to be set for a given class of students. The availability of such tools allows teachers to use ELEs in the classroom in new ways because they provide a greater sense of awareness than is possible with general-

purpose student monitoring tools. Moreover, such tools can support teachers in providing evidence of students' learning, even in a context that is less subject to formal assessment, and to engage in their own enquiry into more conceptual student learning.

So far, we have seen examples of educational software in which data volume and velocity arise from the fact that the majority of the data are being generated by the system as users interact with it. There are other categories of system (most notably, social networking and collaboration software) in which high data volume and velocity arise from the sheer numbers of users and where the majority of the data are user-generated. Our own research in the L4All and MyPlan projects[5] provides an example of this latter category of system. The prototype L4All system developed by these projects aimed to support adult learners in exploring learning opportunities and in planning and reflecting on their learning. The system allows users to create and maintain a chronological record of their learning, work and personal episodes—their timelines. Users' timelines are encoded as RDF triples, compliant with an RDFS ontology[6]. There are some 20 types of episode, each belonging to one of four categories: Educational, Occupational, Personal, Other. Educational and Occupational episodes can be annotated by the user with a primary and possibly a secondary classification when they create the episode (these classifications are drawn from standard United Kingdom occupational and educational taxonomies[7]). Figure 16 illustrates a fragment of the overall L4All ontology.

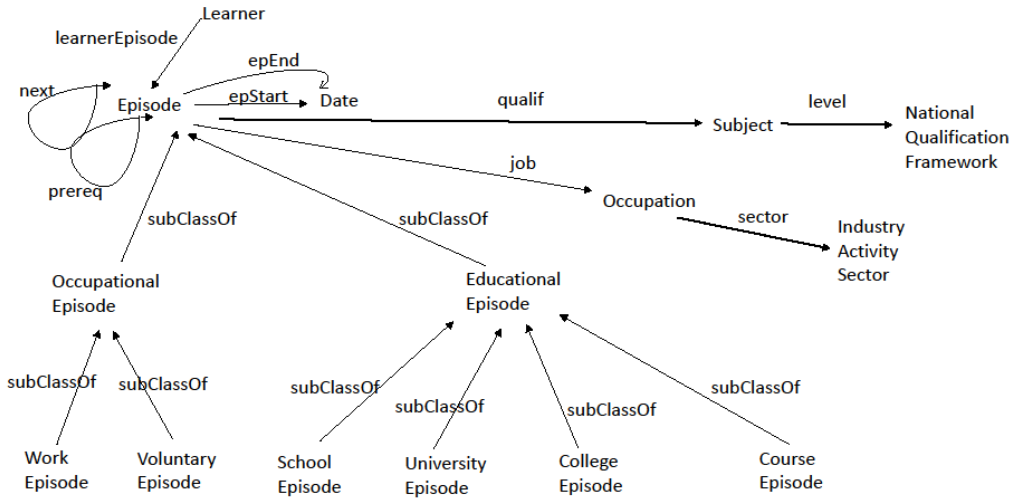

Figure 16. Fragment of the L4All ontology. Each instance of the Episode class is: linked to other episode instances by edges labelled `next' or `prereq' (indicating whether the earlier episode simply preceded, or was necessary in order to be able to proceed to, the later episode; linked either to an Occupation or to an

[5] L4All – Lifelong Learning in London for All; MyPlan – Personal Planning for Learning throughout life. Funded by JISC Distributed e-learning Pilot Call, 2005 – 2008.

[6] See https://www.w3.org/standards/semanticweb/ for information about RDF and RDFS.

[7] See Labour Force Survey User Guide, Vol 5, http://www.ons.gov.uk/ons/guide-method/method-quality/speci_c/labour-market/labour-market-statistics/index.html

educational qualification (Subject) by means of an edge labelled 'job' or 'qualif'. Each occupation is linked to an instance of the Industry Activity Sector class by an edge labelled 'sector'. Each qualification is linked to an instance of the National Qualification Framework (NQF) class by an edge labelled 'level'. The Occupation, Subject, Industry Activity Sector and NQF hierarchies are drawn from standard United Kingdom occupational and educational taxonomies (see Labour Force Survey User Guide, Vol 5, http://www.ons.gov.uk/ons/guide-method/method-quality/speci_c/labour-market/labour-market-statistics/index.html).

Users can choose to make their timelines 'public' and thus accessible by other users. This sharing of timelines exposes future learning and work possibilities that may otherwise not have been considered, positioning successful learners as role models to inspire confidence and a sense of opportunity. The system's interface provides screens for the user to enter their personal details, to create and maintain their timeline (see Figure 17), and to search over the timelines of other users based on a variety of search criteria.

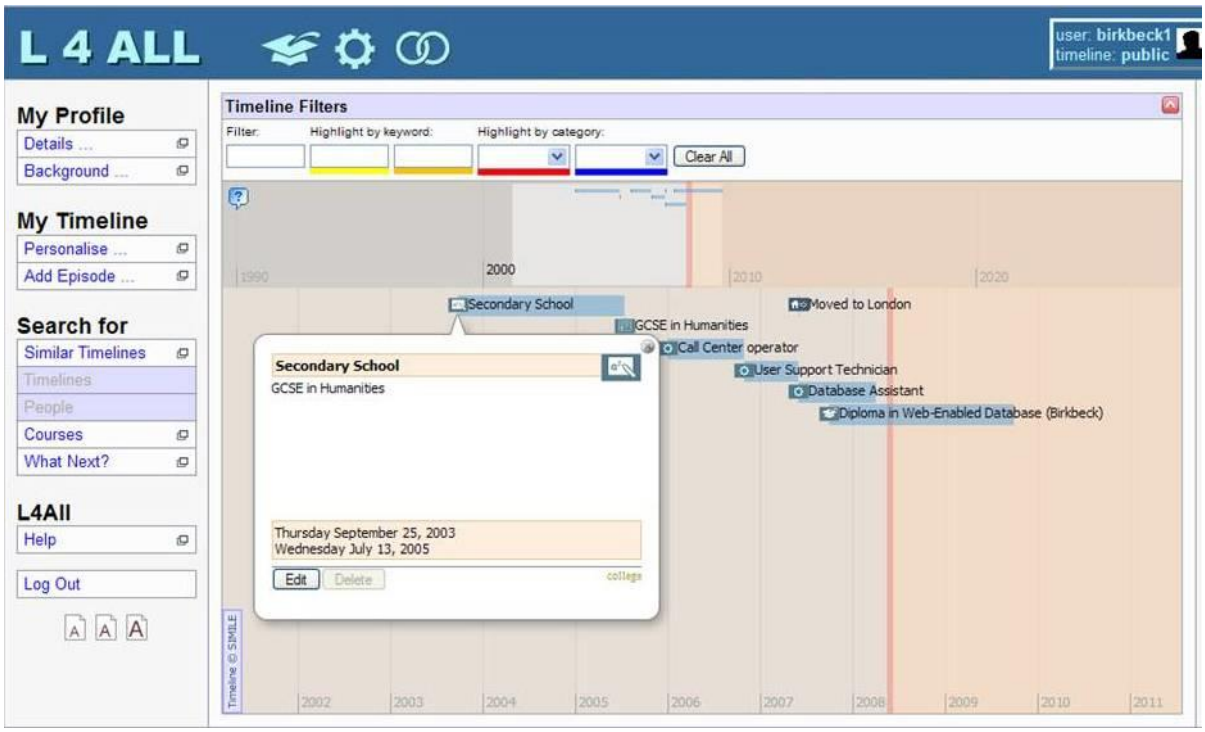

Figure 17. The main screen of the L4All system. At its centre is a visual representation of the user's timeline, and the system functionalities are organised around this. Each episode of learning or work is displayed in chronological order, depicted by an icon specific to its type and a horizontal block representing its duration. Details of an episode can be viewed by clicking on the block representing it, which pops-up more detailed information about the episode (dates, description), as well as access to edit and deletion functions.

Van Labeke et al. 2009, 2011 describe two of the search facilities provided by the system, one to search for "people like me" and another to find recommendations of "what to do next". The latter is illustrated in Figure 18 where we see one of the recommended timelines being displayed beneath the user's own, for easy visual comparison. Specific episodes within that timeline that are being recommended to the user as a source of possible inspiration for their own future learning and career development are shown in orange.

Figure 18. The "What Next" user interface. Episodes in the recommended (lower) timeline that match episodes in the user's own (upper) timeline are shown in blue; episodes that start after all blue episodes are shown in orange – these are deemed by the system to be relevant for this user as they occur after the matching episodes, and thus represent possible choices that the user may be inspired to explore further for their future learning and career development; episodes that occur earlier than all blue episodes or have no matches within the user's own timeline, are shown in grey.

The technical basis for both the "people like me" and the "what to do next" facilities is the users' annotation of their episodes with concepts drawn from the L4All ontology. The availability of this metadata allows similarity algorithms to be used to compare the user's own timeline with all other timelines (see Van Labeke et al 2009, 2011, and also Poulovassilis et al 2012).

CONCLUDING REMARKS

We have presented here examples of some of the opportunities that 'big data' brings to childrens' and adults' learning:

- The provision of personalised and adaptive feedback to students working with an Intelligent Tutoring System or an Exploratory Leaning Environment can enhance students' engagement, motivation and self-confidence, leading to improved learning outcomes.
- Provision of individual feedback to students through an ITS or ELE also frees up time for the teacher to formulate more complex or nuanced support for students. The automated production of feedback for the most common situations can help the teacher to support larger numbers of students learning in larger classes.
- Exploratory Learning Environments, in particular, have the potential to foster greater engagement and deeper learning. To achieve effective support for students in such environments, multi-disciplinary teams of pedagogical experts, learning designers and computer scientists must work together in order to identify what are significant indicators in a given learning setting and to design computational techniques for detecting such indicators and generating feedback for students.

- Detecting students' affective state has the potential to further support learners, through nudges that can move them out of negative states that inhibit learning. Messages of encouragement are particularly important.
- The rich range of data that can be collected by an ELE provides also the opportunity to design visualisation and notification tools for the teacher. Such tools increase the teacher's awareness of the classroom state and of individual students' progress on the task set; and hence help the teacher in supporting both individual students and the class as a whole. To be fully effective in the classroom, such Teacher Assistance (TA) tools need to be designed by multi-disciplinary teams involving teachers, pedagogical experts and computer scientists.
- Moreover, with the increased emphasis on evidence-based teaching, such TA tools can help teachers provide evidence of students' learning, even in a context such as exploratory learning, and to engage further in their own professional development.
- Similarity algorithms can be used to aid teachers in grouping students for productive reflection and discussion activities. Such activities also provide opportunities for students to support each other, i.e. peer support.
- Semantic modelling of learners and learning resources can be used to implement fine-grained similarity algorithms that can underpin sophisticated search and recommendation functionalities, for example for identifying and reflecting on possible next steps in learning and on career trajectories.

Despite these opportunities, there are still many challenges to fully exploiting the potential of big data in education. These challenges include:

- *pedagogical challenges:* Understanding what information is useful to whom and in what learning contexts.
- *technical challenges:* Designing methods for collecting, managing, integrating, analysing and visualising big data in a way that is both practically feasible and pedagogically meaningful in specific learning contexts and settings.
- *socio-technical challenges:* Ensuring that teachers, learners and other stakeholders are sufficiently empowered, involved, and trained to make effective use of the information that can be obtained from big data. Framing agreements between different educational stakeholders so as to allow sharing of learning-related data for the benefit of learners.
- *ethical challenges:* These are numerous and include questions such as: What data about an individual should require their explicit consent in order to be collected, combined, used and shared? Likewise, what knowledge should be allowed to be inferred from the data, and what uses of such knowledge should be permitted? What levels of information and explanation are needed so that individuals can make fully informed decisions? What are appropriate anonymization, privacy, authorisation and preservation policies for both data and inferred knowledge in different contexts of usage? From the opposite perspective, what inequalities may be faced by students (for example from less advantaged backgrounds) whose learning-related data is *not* being collected and utilised to offer them enhanced educational opportunities?

In our own research projects, we address the pedagogical and socio-technical challenges through close collaboration between researchers, developers, students, teachers, and other stakeholders. The technical challenges we face we address by drawing on multi-disciplinary expertise from across computer science, the learning sciences and education. In the absence as yet of sufficiently broad and robust ethical frameworks, we address ethical challenges on a project-by-project basis, fully engaging with our institutions' processes for ethical review of research, and also aiming to inform and shape these going forwards into the 'big data' era.

References

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A. et al. (2010). A view of cloud computing. *Communications of the ACM*, *53*(4), 50-58.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, *39*(4), 12-27.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, *36*(4), 1165-1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, *19*(2), 171-209.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683-695.

Dawson, S. (2008). A study of the relationship between student social networks and sense of community. *Educational Technology & Society*, *11*(3), 224-238.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

De Liddo, A., Shum, S. B., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). Discourse-centric learning analytics. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 23-33). ACM.

D'Mello, S., & Graesser, A. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 23.

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 43.

Drachsler, H., & Greller, W. (2012). The pulse of learning analytics understandings and expectations from the stakeholders. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 120-129). ACM.

EMC Education Services. (2015). *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Grawemeyer, B., Holmes, W., Gutiérrez-Santos, S., Hansen, A., Loibl, K., & Mavrikis, M. (2015). Light-bulb moment?: towards adaptive presentation of feedback based on students' affective state. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 400-404). ACM.

Grawemeyer, B., Gutierrez-Santos, S., Holmes, W., Mavrikis, M., Rummel, N., Mazziotti, C. J., & Janning, R. (2015 b). Talk, tutor, explore, learn: intelligent tutoring and exploration for robust learning. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED)*.

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutierrez-Santos, S., Wiedmann, M., Rummel, N., (In Press). Improving engagement and enhancing learning with affect-aware feedback.

Gutierrez-Santos, S., Geraniou, E., Pearce-Lazard, D., & Poulovassilis, A. (2012). Design of teacher assistance tools in an exploratory learning environment for algebraic generalization. *IEEE Transactions on Learning Technologies*, *5*(4), 366-376.

Gutierrez-Santos, S., Mavrikis, M., Geraniou, E., & Poulovassilis, A. (In Press). Similarity-based Grouping to Support Teachers on Collaborative Activities in Exploratory Learning Environments. *IEEE Transactions on Emerging Topics in Computing.*

Koedinger, K. R., Brunskill, E., Baker, R. S., McLaughlin, E. A., & Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, *34*(3), 27-41.

Li, N., Cohen, W., Koedinger, K. R., & Matsuda, N. (2010). A machine learning approach for automatic student model discovery. In *Educational Data Mining 2011*.

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., & Koper, R. (2011). Recommender systems in technology enhanced learning. In *Recommender systems handbook* (pp. 387-415). Springer US.

Mavrikis, M., Gutierrez-Santos, S., Geraniou, E., Hoyles, C., Magoulas, G., Noss, R., & Poulovassilis, A. (2013). Iterative context engineering to inform the design of intelligent exploratory learning environments for the classroom. *Handbook of Design in Educational Technology*, 80-92.

Mavrikis, M., Gutierrez-Santos, S., & Poulovassilis, A. (2016). Design and evaluation of teacher assistance tools for exploratory learning environments. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 168-172). ACM.

Noss, R., Poulovassilis, A., Geraniou, E., Gutierrez-Santos, S., Hoyles, C., Kahn, K., Magoulas, G.D., & Mavrikis, M. (2012). The design of a system to support exploratory learning of algebraic generalisation. *Computers & Education*, *59*(1), 63-81.

Poulovassilis, A., Selmer, P., & Wood, P. T. (2012). Flexible querying of lifelong learner metadata. *IEEE Transactions on Learning Technologies*, *5*(2), 117-129.

Siemens, G. (2012). Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 4-8). ACM.

Siemens, G., & d Baker, R. S. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.

Van Labeke, N., Magoulas, G. D., & Poulovassilis, A. (2009). Searching for “people like me” in a lifelong learning system. In *European Conference on Technology Enhanced Learning* (pp. 106-111). Springer Berlin Heidelberg.

Van Labeke, N., Magoulas, G. D., & Poulovassilis, A. (2011). Personalised search over ligelong learners’ timelines using string similarity measures. Technical Report BBKCS-11-01, Birkbeck. http://www.dcs.bbk.ac.uk/research/techreps/2011/bbkcs-11-01.pdf, 2011.