P – An Open Source Personalization Platform for Higher Education

Vince Kellen, Ph.D.
Senior Vice Provost, Academic Planning, Analytics and Technologies
vkellen@uky.edu

Adam Recktenwald
Enterprise Architect
University of Kentucky

Cody Bumgardner
Enterprise Architect

1. Abstract

Adaptive and personalized learning technologies are now emerging. Early adopters of vendor solutions are currently experimenting with these tools and beginning to measure the impact on student learning. With the advent of big data systems and ubiquitous mobile devices that nearly every college student possesses, colleges and universities can begin to apply extremely fast computational approaches to enhancing student learning, advising, peer interactions and student support in real time. Since these personalization and big data technologies are relatively new, the lack of adoption within higher education has resulted in insufficient empirical findings to adequately guide further adoption across our varied curricula. Moreover, adaptive and personalized learning implementations are emerging in small silos that don’t easily share student data and personalization models across instruction, advising and student support. We propose that higher education, in partnership with industry vendors, develop largely through integration of pre-existing open source and vendor technologies, an open source personalization platform, which we call “P”. The platform should be designed to accommodate real-time personalization of content and interactivity via a variety of user interfaces, including learning management systems, MOOCs, active learning and intelligent tutoring systems, web sites and portals and mobile applications. This platform should also accommodate the need to develop adaptive and personalized learning approaches to support a variety of pedagogies in use within higher education, reduce the effort to extract and build metadata needed to develop adaptive course content and allow for several open and proprietary approaches for personalization and adaptive algorithms. More significantly, the open source platform should allow for rapid discovery and sharing of different personalization approaches to promote rapid adoption of personalized learning across multiple institutions.

2. Opportunity

Citizens and legislatures are concerned with higher education cost-effectiveness and are questioning the higher education’s adoption of technology within areas that directly impact students including teaching, learning and support. While adaptive and personalized learning technologies are now emerging, adoption is limited with early adopters of vendor solutions are currently experimenting with these tools. With the advent of big data systems and ubiquitous mobile devices that nearly every college student possesses, colleges and universities can begin to apply extremely fast computational approaches to enhancing student learning, advising, peer interactions and student support in real time in cost-effective ways.

Many colleges and universities, including University of Kentucky, are continually challenged with student graduation rates that are persistently lower than desired. Many learners fail to continue in their academic pursuits and drop out. The reasons for this, which have been researched heavily for the past few decades, vary and include: family unfamiliarity with how to prepare their children for success in college; the learner’s economic capacity to pay for college; insufficient learner preparation in K-12 schools; insufficient learner motivation and engagement in their college studied and life; work and family conflicts that pull learners away from college; individual differences in cognitive abilities; and the inability for some learners to adapt to the current form of college studies. The cost of lower student success rates is felt in two ways: learners and their families may spend and borrow significant amounts and fail to have the qualifications needed in the job market; institutions may need to spend money on additional services or aid to retain these students and lose potential tuition revenue when they drop out.
For University of Kentucky, each percentage point increase in graduation represents about $2,000,000 in additional annual revenue.

Personalization, when applied to student learning, advising and support, can impact retention. While institutions have implemented various systems to track or enable electronic communication, most of those systems require manual interventions or significant configuration effort. For example, early alert systems often require faculty and advisors to identify students who need help. Many failing students however, attempt to save face and subsequently avoid easy detection. Developing eLearning courses requires a substantial investment in faculty time to set up the course metadata. Many current adaptive learning technologies require many hours, if not hundreds of hours of configuration per each hour of online instruction. The strength of higher education -- its complex array of many focused, lower enrollment classes -- makes it difficult to augment these classes with adaptive and personalized eLearning techniques. If personalized or adaptive learning tools are to be adopted, especially across the “long tail” of courses that defines higher education today, instructors and instructional designers will need to have more automated or crowd-sourced means to develop the metadata and the progression rules for personalized instruction. Cost-cutting and insufficient resources may often prevent institutions from setting appropriate staffing levels to accomplish this. The personalization platform proposed here can automate metadata extraction, implement simpler but effective personalization and adaptive learning approaches and promote reuse of specific personalization approaches across higher education.

In order to do more with less, institutions will need to find new ways to automate components of learning and service delivery to students. By providing student with tailored information given their varied abilities and unique cognitive and non-cognitive traits, institutions can improve outcomes and reduce costs.

3. Definition of personalization and benefits

Personalization is defined as the manipulation of content as it is delivered to individuals to improve the individual’s understanding of and receptivity to that content. Just as Google’s ecosystem personalizes advertising content specific to what it knows about the user, personalization systems in education can match specific content to individual learners based on learner cognitive (e.g., GPA, ACT/SAT test scores, cognitive ability tests, etc.) and non-cognitive (e.g., motivation, persistence, confidence, need for sensation, etc.) attributes, learner activity (e.g., click stream, submission of assignments and work, online discussions, peer interactions), pedagogical rules (e.g., progression of skill and knowledge in meeting learning objectives) and learner mastery over core concepts and skills.

Personalization can be reactive or proactive. Reactive personalization senses from the learner and alters the interface to deliver content after the learner has initiated activity. Proactive personalization prepares content based on relevant learner information already available. For example, if a student is not mastering a core concept, a personalization system can detect gaps in knowledge and deliver additional content to the learner, reactively. A personalization system can also proactively deliver content based on pre-existing data such as entrance test scores and learner demographics. Personalization can be initiated by the learner or by the system. A learning or support environment can let learners customize and configure their environment to suit their needs. In addition, a personalization system can automatically match content to the learner.

Using adaptive and personalization techniques can potentially improve student performance in classes, improve learner use of resources such as advising, tutoring and support, can improve institutional cost effectiveness by using automation instead of human labor and can enhance the institution’s understanding of how learning and support activities are best delivered. By providing an open source framework, institutions can maintain a ‘line of sight’ to the student, just as faculty and advisors do today in face-to-face interactions. Ultimately, a personalization engine is likely to contribute to improved student learning outcomes, student success and retention and institutional cost-effectiveness.
4. Student, instruction and support models

The research into adaptive learning tools, intelligent tutoring systems, intelligent pedagogical agents, artificial intelligence in education and personalization approaches to education is both broad and deep spanning several decades. The research reflects the complexity of the research problem in which matching content with students results in a combinatorial explosion of student, domain, pedagogy and user interface attributes and levels. To accommodate the diversity and complexity in the personalized learning research program today, a comprehensive personalization platform would need to handle the state data and process execution options these attributes and levels would need. As we are seeing today in current tools, this will result in a variety of computational approaches that would match content to learners.

Based on reviews of research related to individual differences in student learning and personalized learning technology (O’Connor & Paunonen, 2007; Graesser, Conley & Olney, in press; Sampson, Karagiannidis & Kinshuk, 2002; Jonassen & Grabowski, 2011; Mayer, 2009) and on current practices in higher education, a complete student data model would need to capture many different learner attributes:

- Cognitive and noncognitive traits that include working memory capacity, spatial, verbal and numeric abilities, field dependence/independence, cognitive reflection/impulsivity, focal attention, need for sensation, need for cognition, cognitive complexity, prospective memory
- Cognitive styles that include visual/haptic, visualizer/verbalizer, serialist/holist, analytic/relational, visual leveling/sharpening
- Learner emotional state including confusion, frustration, boredom, anxiety, engagement/flow
- Learning styles and their attendant attributes from various frameworks such as cognitive style mapping and Kolb’s, Dunn & Dunn, Grasha-Riechmann and Gregorc learning styles
- Personality traits such as openness to experience, conscientiousness, extraversion, neuroticism, agreeableness and the various subcomponents of these traits. Other personality traits such as ambiguity and frustration tolerance, locus of control, self-efficacy, achievement motivation and risk-taking
- Prior student knowledge within domains of study along with prior grades and grade averages, standardized test scores and subcomponents such as AP, CLEPP, ACT and SAT scores
- Current skill levels within the particular class or knowledge component, formative and summative assessments of the learner performance in the current class
- Student demographics such as age, gender, location, miles from home, high school attended, economic capability, first-generation college student, visual or other impairments, etc.
- Student behavioral data such as interactions with advisors, faculty, engagement in academic and co-curricular activities, use of informational resources such as libraries and study spaces, clickstream in learning management and other systems and other behavioral data indicating engagement or detachment from the institution

Personalization technology, including adaptive learning, intelligent tutoring systems and intelligent pedagogical software agents, can not only select among many content choices to deliver to a student, it can restructure the sequence of possible interactions for individual students. The approach to this sequencing or instructional model can vary depending on the academic domain. Not all academic domains are the same. Some domains are rich in procedures and structured problem solving and others involve more verbal and conceptual skills and knowledge acquisition. Systems today range from those that serve well-formed and procedural topics such as programming languages, geometry and algebra and others handle domains with a strong verbal foundations (Graesser, Conley & Olney, in press). Some systems today analyze verbal language using computational linguistic approaches and in the future these systems may leverage super-computing environments like IBM’s much-discussed Watson technology to handle more complex verbal domains. Different academic domains and pedagogies are likely to need different computational methods or algorithms for the varying instructional models.

Intelligent agents are software programs that are intelligent, autonomous and act on the environment based on sensing mechanisms (Soliman & Guetl, 2011). Agents can be used in a variety of ways, principally but not
necessarily through the use of a visual representation. Agents can be used for personalizing content delivery, focused on providing emotional support or teaching meta-cognition, facilitating group learning and improving learner self-regulation (Soliman & Geutl, 2011). Intelligent agents can work independent of other intelligent agents but can also work in a distributed fashion with other intelligent agents. In short, the intelligent agent framework allows for a more distributed and loosely-couple approach for personalizing education. Given the development of intelligent agent frameworks over the past decade, a personalization platform needs to provide an environment to house and run these agents. Moreover, intelligent agents are likely to need data related to one or more student models.

Personalization technology can also make a difference in student support. Identification of students at academic or other risk are common in many institutions and prediction of potential drop-out or other risk relies on available data drawn from student demographic or behavioral models. Normal student advising and support interactions can be personalized using simpler computational approaches based on reduced student models and delivered via more widely available web and mobile software. These systems need not be separate, as they are today. Both learning and support interactions can benefit from shared student and instruction model data. For example, with the rise of social network analysis that analyzes various aspects of students relationships with others, researchers have identified student attributes related to their role and position within a social network (isolation, weak social presence, social capital) that predict student success (Eckles & Stradly, 2012; Thomas, 2000; Johnson, Hornick & Salas, 2008; Gasevic, Zouaq & Janzen, 2013). A personalization platform could gather richer student model data that can then be used by instructional and support information technology tools.

Of critical concern in these systems are the trade-offs between the setup, implementation and maintenance costs versus the benefits from the improved learning outcomes. Based on the learner and instructional model employed, some systems have a very large number of skills and production rules to configure. Others may be more parsimonious in their initial implementation (e.g., constraint-based systems, case-based reasoning systems, systems based on Bayesian network computation, semantic analysis) but may suffer from either lack of generalizability to other domains or from an inability to accommodate a broad enough range of student behavior and give sufficient student feedback (Graesser, Conley & Olney, In press). At the least, implementing personalization technology to support student learning and support will invariably give rise to difficult tensions between the a) implementation effort determined by the granularity and complexity of the student and instructional model, b) the generalizability of the approach to multiple academic domains and c) the ability to accommodate a range of student behavior appropriately. One-size will not fit all and for the foreseeable future, an approach that can foster and develop an ecosystem of tools may be needed.

5. **Challenges for adoption of MOOCs, eLearning and personalization**

Higher education has a number of challenges with regard to use of technology to affect learning. Converting existing course content to support eLearning takes time and effort. Instructional designers and instructors need to establish metadata for all course content so that the course delivery platform can select the appropriate content. For example, sections of a video segment need to be tagged as pertaining to one or more core course concepts so that related assessments can be delivered to the learner. Machine learning approaches need to be further developed to automate more of the metadata construction problem.

Different pedagogies require different rules of content and skill progression. Some disciplines require preliminary material to be covered before later material. Other disciplines approach learning in a less linear fashion. Instructors with high levels of teaching skill often have implicit approaches that are quite different than less experienced or skilled instructors. These ‘rules of personalization’ once made explicit in a personalization framework, need to be categorized and made available to other faculty and institutions so that best practices regarding personalized learning can diffuse. The open source platform will need to allow for discovery and sharing of personalization practices across a variety of institutions and pedagogies.
In order to reach learners, existing and new vendor or institution tools such as mobile applications, learning management systems and portals will need application programming interfaces (APIs) to a) let the personalization platform interrogate current systems state (e.g., where is the student in the video segment? Is the student in the middle of an assessment?); b) harvest clickstream data; c) push personalized content directly to a region within the user interface (a window of interactivity); and d) provide information to the user interface so it can determine which content to deliver next.

Finally, a personalization platform will most likely need to find ways to automate the searching of external resources such as Google Scholar, library databases, general websites or any other repository that deliver superior search results. Domain specificity can improve search results significantly. The open source platform will need to support extendable approaches to metadata extraction and external searches.

6. **Approach**

A personalization platform would need to eventually accommodate instructional, co-instructional, advisory and at-risk support needs across mobile and fixed ubiquitous devices. The platform will need to interoperate with current and future adaptive learning and intelligent tutoring systems and intelligent software agents as well as the now-available high-speed, big data analytic tools likely to increase in adoption in higher education. The platform will need to foster and enable rapid experimentation with courses, students and academic programs of all kinds and sharing of personalization approaches and rules within and across institutions. To meet these broad and evolving needs, the approach we are taking to further develop P includes the following elements:

**Develop a metadata extraction (“distillation”) engine.** Using existing services (Green Button/Microsoft InCus/MAVIS) and tools (R and other analytic solutions), convert course lecture capture audio and any textual screen elements such as presentation slides to text. Extract key concepts from this converted text and any other course materials to derive a list of core concepts. This can speed up conversion of courses to an eLearning format and let instructional designers more easily specify personalization rules that match content to concepts. The Distillation Engine can be used to compare a corpus of course material and categorize the content. Many institutions have collected thousands of lecture captures, course documents and presentation slides but have little to no metadata describing that content.

**Implement metadata augmentation interfaces.** Develop or integrate interfaces that let instructional designers and instructors review recommended concepts from the Distillation Engine and then modify those concepts as needed for the learning objectives. The final collection of concepts and terms (the taxonomy) can then be used for multiple purposes, including using within the Personalization Engine to connect content to learners, to express learning maps and progressions used within the Personalization Engine and to send to the Distillation Engine to improve automated extraction of concepts from text. Learners themselves could potentially make additions to the taxonomy via crowd-sourced approaches. The final taxonomy can potentially improve the Distillation process. For example, the MAVIS/InCus service from Microsoft and Green Button enhances its ability to convert speech to text when it is provided a dictionary of terms. Other machine learning algorithms can be enhanced in a similar fashion.

**Document needed APIs from solution providers.** The P platform will benefit from real-time exchange of data between the personalization engine and user interfaces within learning management systems, active learning tools, MOOCs, mobile applications and web applications. These APIs should specify how P can interrogate state information including temporal location within an audio-visual player, search terms used to locate significant points in a lecture capture or any other course content and other clickstream data. The APIs should specify ‘windows of interactivity’ that the learner tool can provide. These windows of interactivity let the P platform communicate with the learner in the tool and deliver personalized messages. The APIs should also specify how the P platform can provide information to the learner tool so that the tool can coordinate and sequence course content within a learning map or progression based on data and personalization rules within P.
Include one or more personalization or adaptive learning tools. Select from the various open source tools in the categories of a) basic rule-based engines and b) machine learning algorithms that can be applied to personalization in student support and instruction. Select and support a multi-agent software platform to enable sharing and development of intelligent pedagogical agents or agents designed to enhance student success. Allow for the inclusion of one or more intelligent tutoring systems and allow data sharing and personalized messaging overlays where personalized messages may be delivered to learners in concert with or through the adaptive learning tool. Provide the administrative interfaces that enable specification of rules, configuring of machine learning approaches and writing of custom software agents for specific personalization routines.

Promote an open student model. By allowing learners to participate in the data and model concerning personalization, an open student data model can promote metacognitive activities such as planning, self-monitoring and reflection, let the learner take greater control and responsibility for their involvement and can facilitate interactions between students and instructors (Bull & Kay, 2010). Learners can participate with the institution in the design and implementation of personalization approaches and can gain access to data about themselves.

Implement personalization package discovery tool. Let faculty and institutions categorize and describe specific personalization packages so that the personalization technique can be shared within and between institutions. Publish the package metadata to external repositories and provide for federated search capabilities so within the P platform personalization packages can be discovered and applied and so that potential P installations might receive and respond to search requests from external systems or instructors and instructional designers at other institutions.

Design to promote ecosystem development. An open-source approach lets third parties and vendors use the open platform, customize it for their specific environment and add additional services and software that become valuable to higher education. For example, providers of high-speed analytic engines (e.g., SAP’s HANA) could convert portions of the P platform that would then be optimized for superior speed and performance in their analytic platform.

7. Assessment of Learning

Many popular learning management systems are also providing additional analytic capabilities that measure student involvement with the class, with specific content within the class and with performance on tests, assignments and quizzes. While these analytic capabilities are welcome, these tools typically measure interaction with a specific artifact within the course content. These tools do not measure student engagement with and mastery over specific core concepts within the class. To do this, these tools would need to allow instructors to establish the metadata and ‘tag’ specific phrases within artifacts with that metadata, which is a time-consuming process. One of the advantages of current adaptive learning tools is that these tools also provide analytics that reveal how well learners are progressing. Most of these tools lets instructors and instructional designers setup the metadata for the class (again, a time consuming process) and once set up, the analytics provide measurement of specific key concepts within the class.

The approach here is to automate that metadata extraction process by using commonly available services and machine learning algorithm to convert recorded speed to text and convert video capture streams to structured data using optical character recognition methods. This can help extend personalization techniques to the “long tail” of necessary, but smaller classes.

Assessment of learning can be achieved by standardized tests at key points in the student’s progression (first year and fourth year) or by assessing student work produced in the class. The latter, often called authentic student learning assessment, typically requires a pool of faculty raters who can examine student work and assess how well students are mastering complex skills, such as critical thinking or quantitative reasoning. These assessment regimes provide high-quality data regarding student skills across courses and programs. While these assessments
are necessary, many faculty do not find sufficient value for them to engage in either preparing their courses for this type of assessment or to participate as a faculty rater in the assessment protocol.

The course concept assessment method advocated here could be a “missing link” between instructor specific assessment (grades) and comprehensive, complex skill assessment across different classes or programs. The assessment data P will make available would be data specific to sections that shared a similar set of concepts, such as introductory classes into statistics, chemistry, biology, calculus, computer science and many other subjects. Advanced level courses and professional and graduate program courses also frequently have a discrete set of concepts common across sections that student must master.

To illustrate the assessments possible, imagine an introductory course into computer science with 10 sections and 400 students. If the computer science department can agree to a set of common concepts in the specific course (usually by adopting a common text book or some common tests), using common tools such as lecture capture and learning management systems, the P personalization platform would identify where in the lecture video file key concepts are being discussed, where in the LMS students are discussing specific concepts, where in the presentation slides (and other documents) those concepts reside. Through the APIs specific in the architecture, the P personalization platform will collect student engagement data (via click streams and logging data) and through the assessments (quizzes, tests and assignments), P can identify student mastery over these key concepts.

Using real-time data analytics available today, the P personalization platform could present department chairs and faculty with up-to-the-minute data on how well their cohort of students are engaged with specific concepts and what their level of mastery may be, providing the faculty with the ability to intervene early either through direct face-to-face intervention, or through the use of personalization and adaptive learning facilities within P. This assessment data would provide a rich set of data that would be highly actionable within short time frames and provide institutions the opportunity to quickly and iteratively improve the quality of instruction in a highly measured way.

8. **Project Status and Partners**

University of Kentucky. UK is one of the first universities globally to adopt high-speed analytics supporting student success. The University has procured and implemented SAP’s high-speed analytic platform, HANA and has developed, in partnership with Dell, a higher education reference model for student data. UK has used the student data model to delivery access to high-speed analytic views across campus to support community analysis of student success data. UK has established and advanced analytics community of practice to encourage college and department analysis of student success issues. So far, UK has done the following:

- Developed and deployed a mobile platform that nearly 100% of the student population has installed and is deploying a “student health indicator” metric directly to students (K-Score) to provide a means for personalized recommendations to students
- Complete initial research and proof-of-concepts for automatic partial transcriptions of captured lectures to improve video way-finding and provide means for delivering personalized interactions to students while interacting with online course material. These techniques use services from Microsoft (MAVIS) and GreenButton (InCus) to performance speech-to-text conversion.
- Currently conducting tests of automatic content classification and taxonomy extraction from course content using open-source tools. UK is continuing to develop the P platform using existing staff but the project is reaching the point where broader collaboration with industry partners is needed to share specific knowledge and accelerate development.
- Collaborating with Echo 360 on API specification to support personalized and adaptive learning using the P platform.
Dell. Dell is supporting this project and is willing to make in-kind and other contributions. Dell staff co-developed the higher education student data reference model with UK. Dell’s engineering experience in the inBloom initiative will be helpful in ensuring the P platform can reuse and interoperate with inBloom architecture.

Echo 360. This company sells active learning and lecture capture tools to higher education and supports this project. UK is currently collaborating with Echo 360’s active learning tool expert, Dr. Perry Samson (University of Michigan faculty) on specifying APIs that can enable personalized messages delivered within the active learning tools and advanced analytics than can help match external content to learners using the active learning tool. Echo 360 is in support of this project.

SAP. Executives in SAP’s public services division are in support of this initiative. SAP and UK are currently collaborating on developing an industry-university big data analytics community. SAP is supporting this project.
9. References


