Plenary:
Better Living Through Data

Presented By:

Amy Plent
North Carolina Bar Association
Durham, North Carolina

Ryan Graham
Peach New Media
Atlanta, Georgia

Randy Goins
TriMark Digital
Raleigh, North Carolina

Presented at:
ACLEA 51st Annual Meeting
August 1-4, 2015
Chicago, Illinois
Amy Plent
North Carolina Bar Association
Durham, North Carolina

Amy is a partner at JDS, a consulting group providing content strategy, content creation, editorial and course development advice and services to professional publishers and learning organizations. Amy helped to found JDS in 2004, sold her interest in 2007, married the new owner in 2010, and returned to employment in 2015. At present Amy is serving in a full time consulting capacity as Senior Advisor to the COO at the District of Columbia Bar. Prior roles include Director of Product Development for the American Institute of CPAs and Director of CLE for the North Carolina Bar Association. Amy is a chair of the ACLEA Executive Leadership SIG and a member of the planning committee for this 51st Annual Meeting. When not working Amy enjoys hanging out with her husband Bruce and plotting ways to buy her company back from him. Amy lives in Durham North Carolina with Bruce, her 17 year old daughter Sarah, and two lazy but entertaining dogs named Livvy and Bridget.

Ryan Graham
Peach New Media
Atlanta, Georgia

Ryan joined the original Boston Conferencing team in 2006, to develop the company’s sales organization and help expand the market presence to professional trade associations. Ryan played a key role in the 2009 merger of Boston Conferencing and Impact Media Solutions, which led to the launch of Peach New Media. Under Ryan’s sales leadership, Peach has grown to become a dominant provider of learning platform solutions in the mandatory continuing education market. Specifically, Peach has grown to become the leading learning platform provider to the mandatory continuing legal education (CLE) space, including state bars and national CLE organizations. Ryan is a frequent speaker and content contributor on topics related to mandatory continuing education and online learning. Ryan and his beautiful wife reside in Overland Park, KS, with their 2 young and energetic boys – Jackson and Peyton.
Randy Goins  
TriMark Digital  
Raleigh, North Carolina

An Appalachian State University graduate and collegiate golfer, Randy founded TriMark Digital in 2006 as a full service, integrated digital agency, headquartered in downtown Raleigh, NC. Specializing in customer acquisition and engagement strategies, TriMark Digital employs 36 full-time professionals and has been honored as an Inc 500 / 5000 for 3 consecutive years. Randy began his career in digital marketing in 2001 with CitySearch.com, a national search engine directory which pioneered local websites and search strategy. Randy currently serves as the President of TriMark Digital and when not working he enjoys spending time with his family and playing golf.
ASSOCIATION FOR CONTINUING LEGAL EDUCATION
51st Annual Meeting
Chicago, IL

Better Living Through Data

Monday, August 3, 2015
INTRODUCTIONS

PANELISTS

Ryan Graham (Peach New Media)
Randy Goins (Trimark Digital)
Amy Plent (JDS Consulting)
TODAY’S GOALS
What Does Better Living through Data Look Like?

TODAY’S GOALS
Associations, Learning & Data

TODAY’S GOALS
Considerations for CLE Providers

TODAY’S GOALS
WHAT DOES BETTER LIVING THROUGH DATA LOOK LIKE?

First off
Decisions Driven by Evidence, Not Hunches
A Consciously Engineered User Experience
Maximized Performance, Increased Opportunities
A Closer Look

ASSOCIATIONS, LEARNING & DATA
Get Personal
Shift the Focus from Compliance to Competence
Take Customer Support to New Levels of Responsiveness
An Even Closer Look

CONSIDERATIONS FOR CLE PROVIDERS
Using Data to Develop the Next Big Thing
Using Data to Improve Existing Products
Using Data to Refine CLE Marketing
Thank you for attending
TOP
DIGITAL MARKETING
SOFTWARES
Top SEO (Intermediate) Software

RAVEN  
MOZ  
SEMrush

Top SEO (Enterprise) Software

SEOprofiler.com  
webposition

seoClarity  
conductor  
BRIGHT EDGE

linkdex  
SEOlytics
Top PPC Management Software

Marin Software

Kenshoo

Oracle Marketing Cloud

DoubleClick by Google

WordStream

acquisio

Top Call Tracking Software

INVOCA

The power of talk

dialogtech

CallRail

SalesLoft

calltrackingmetrics

Marchex

MongooseMetrics

Converz

www.trimarkdigital.com  |  919.785.2275
Top Analytics/Reporting Software

- Google Analytics
- Mint
- Omniture
- KISSmetrics
- Clicktale

Top UX/CRO Software
(heatmaps, visitor recordings, etc.)

- hotjar Insights
- UserTesting
- crazyegg
- luckyorange
- Optimizely
Better Living Through Data

SPECIFIC CONSIDERATIONS FOR CLE PROVIDERS

AMY PLENT
“Indigenous Data” - by product of your customers actual interactions with your products, for example:

- Navigational path to and through site
- Keyword searches
- Purchase behavior
- Abandoned carts
- Interaction with a course
- Help requests

“Farmed Data” - Metrics generated from surveying your audience. May be skewed by selective participation or anticipation of desired response.
How Data Driven Is Your CLE Business?

Where are you on the spectrum?

1. Limited/No Data Reliance: Hunches, History and Honchos
2. Minimal Data Reliance: The 3H’s set direction. Financial trends and evaluation scores are used to make minor adjustments
3. Moderate Data Reliance: Periodic customer needs/satisfaction research is added to the mix and can sometimes drive adjustments to direction
4. Heavy Data Reliance: Reporting and analytics on customer behavior/activity within your online presences is regularly reviewed and a surprising finding might drive adjustments to direction
5. Data Driven: Direction is set by the data findings and adjusted by the 3Hs

Note: #5 may not be right for your business. But making a conscious choice about where to calibrate your data reliance probably is
Subjective vs Objective Input

To what extent do your organization's decisions emerge from only one level of the pyramid?
Recipe for More Data Driven CLE: Turn Pyramid Upside Down. Shake. Repeat
Using Data to Develop the Next “Big Thing”

- Non Data Driven Methods for Developing a New Product/Service
  - Hunches
  - Brainstorming
  - Trial/Error/Refinement
  - Customer/Volunteer/Leadership Suggestions
- Can they work? Yes
- Downside
  - Can take longer to refine your way to success
  - Risk of investing in a non-starter is higher
Using Data to Develop the Next “Big Thing”

Data Driven Methods for Developing a New Product/Service

Mine the Data You are Already Generating

What are customers searching for and not finding on your website or in your LMS?

Can you track keyword searches? Run analytics for concentrations of searched terms that don’t map to an existing product?

Does your course evaluation ask about other courses people would like to see? If yes, is this information looked at only in the context of reviewing that course or is it aggregated across courses? Can you run analytics on it?

Do you run an abandoned cart report? What might it be telling you about needs the current offering may not be meeting?
Using Data to Develop the Next “Big Thing”

Data Driven Methods for Developing a New Product/Service

- Applied Google Analytics
  - Run analysis of the most common CLE related searches in Google (can be segmented to searchers in a given geographic region).
    - How does this compare to searches in your catalog/site/LMS?
    - How does it compare to your existing product offering? Are there commonly searched topics/terms for which you have no responsive product?
Using Data to Refine Existing Products/Services (or concepts)

- Ethnographic Research
  - A fancy term for watching your customers do what they do, to see if it informs a way to make it better/faster/easier
  - Shadowing or filming customers using a product has been a trusted way of evolving products for years
  - Online environments offer a new twist on this approach
    - Can your system track user navigation through a course or a platform?
    - If not, can your developer build this capability in? Can you put it on your wishlist when evaluating platforms?
    - How many clicks does it take to get from start to finish in a process? Where can this be minimized?
    - Where are users getting frustrated and abandoning the process? What can you do to mitigate this?
Using Data to Refine Existing Products/Services (or concepts)

Data as an Objective Validator of Subjective Feedback

Beware of the Vocal Minority!

Which of these is most likely to prompt an adjustment to your product?

1. The customer who complains really loud
2. The customer who complains to leadership
3. A concentration of customers with the same complaint?

Analytics can help you identify group 3 and provide a basis for validating (or avoiding knee jerk reactions to) feedback from categories 1 and 2

Can you envision an analytic process that would add qualitative insight to individual complaints?
Using Data to Refine Product Marketing

- **Source code tracking**
  - Embed source codes into every marketing piece
  - Track performance both in driving traffic to your store and in driving actual purchases

- **Don’t stop the analysis at what drove the customer to the site**
  - Did they buy the product that was promoted to them or something else? If the latter, what does that tell you?
  - Can you see a trend across customers? Can this reveal a hidden gem that should be made more prominent?
  - Did they abandon pursuit? When? What new information did the customer have when they abandoned purchase that was not in the promotion that drove them to your site?

- **Can you match user profiles to purchasing behavior?**
  - Are your loyal customers who you thought they were?
  - Are marketing pieces behaving differently in different demographic groups?
    - Can you split or segment your marketing in response to this?
Investigate what can be produced from existing platforms and vendors or from your organization. Add to your shopping list for new platforms and vendors:

- Access to a web analytics report like webtrends or coremetrics for all areas of the website within your responsibility
- Abandoned cart report, ideally tied to browsing history
- Search terms report for keyword search. This can be a great source of insight, particularly keyword searches in your online store
- Capability to enter and track Google analytics codes at the page level throughout your online store
- Any other reporting the system can generate that gives you a chance to analyze actual user behavior rather than asking the user about it. Brainstorm with your IT team and your vendors about what is possible.
MOOCs (Massive Open Online Courses) are ideal for data gathering and analytics due to the huge sample of online learners available to study.

Research underway at Carnegie Mellon’s LearnLab is seeking to leveraging data from MOOC students to reveal the instructional conditions and techniques that yield the most robust online learning.

Changes to the ways in which MOOC courses are structured and conducted have been requested by the researchers to make it possible to get rich data to analyze.

The findings from this research could ultimately help all providers of online learning better understand the most effective ways of teaching online.

The paper MOOCs and Technology to Advance Learning and Learning Research, which discusses this effort, has been included in your conference materials upload, with the permission of the authors.
Ubiquity Symposium

MOOCs and Technology to Advance Learning and Learning Research

Data-Driven Learner Modeling To Understand And Improve Online Learning

by Kenneth R. Koedinger, Elizabeth A. McLaughlin, and John C. Stamper

Editor’s Introduction

Advanced educational technologies are developing rapidly and online MOOC courses are becoming more prevalent, creating an enthusiasm for the seemingly limitless data-driven possibilities to affect advances in learning and enhance the learning experience. For these possibilities to unfold, the expertise and collaboration of many specialists will be necessary to improve data collection, to foster the development of better predictive models, and to assure models are interpretable and actionable. The big data collected from MOOCs needs to be bigger, not in its height (number of students) but in its width—more meta-data and information on learners’ cognitive and self-regulatory states needs to be collected in addition to correctness and completion rates. This more detailed articulation will help open up the black box approach to machine learning models where prediction is the primary goal. Instead, a data-driven learner model approach uses fine grain data that is conceived and developed from cognitive principles to build explanatory models with practical implications to improve student learning.
Ubiquity Symposium

MOOCs and Technology to Advance Learning and Learning Research

Data-Driven Learner Modeling To Understand And Improve Online Learning

by Kenneth R. Koedinger, Elizabeth A. McLaughlin, and John C. Stamper

In the midst of the recent high energy around massive open online courses (MOOCs) and other forms of online learning (e.g., Khan Academy), it is worthwhile to reflect on what these efforts may draw from and add to existing research in the learning sciences [1, 2]. Given that tens of thousands of students may complete a MOOC course, there is legitimate excitement about what we might learn from the great volumes of student interaction data that these courses are producing. However, for that excitement to become reality, computer scientists joining in this area will need to develop new expertise or forge collaborations with cognitive psychologists and educational data mining specialists.

We recommend data-driven learner modeling to understand and improve student learning. By data-driven learner modeling we mean the use of student interaction data to build explanatory models of elements of learning (e.g., cognition, metacognition, motivation) that can be used to drive instructional decision making toward better student learning. We frame this approach in contrast to the traditional pedagogical model employed in higher education and mimicked in MOOCs whereby “an expert faculty member's performance is the center of the course” [3].

This instructor-centered model typically includes, as do many MOOCs, questions to check for student understanding, problems to apply ideas in practice, and perhaps even learn-by-doing scenarios/simulations that require adapting concepts and skills in support of deeper learning. However, these activities are still “instructor-centered” in that they are primarily designed based on the intuitions of instructors, their conscious reflection on their expertise, and their beliefs about what students should know.
Our experience is that too much online course development is guided merely by instructor intuitions. These intuitions are clouded by what we have called “expert blind spot” [4]—the notion that experts are often unaware of the cognitive processes they utilize when performing in their specialty area. Much expertise is tacit knowledge used in pattern recognition, problem solving, and decision making, and experts’ self-reflections are often inaccurate about the nature of their own tacit knowledge. For example, while most math educators judge story problems to be more difficult for beginning algebra students than matched equations [5], student data indicates the opposite: Students perform better on story problems (70 percent correct) than on matched equations (42 percent correct; [4]). Therefore, in contrast to instructional design based purely on instructor intuition, course development should also be informed by the kind of data that has repeatedly revealed flaws and limitations in the models of student learning implicit in course designs.

Using data-driven models to develop and improve educational materials is fundamentally different from the instructor-centered model. In data-driven modeling, course development and improvement is based on data-driven analysis of student difficulties and of the target expertise the course is meant to produce; it is not based on instructor self-reflection as found in purely instructor-centered models. To be sure, instructors can and should contribute to interpreting data and making course redesign decisions, but should ideally do so with support of cognitive psychology expertise. Course improvement in data-driven modeling is also based on course-embedded in vivo experiments (multiple instructional designs randomly assigned to students in natural course use, also called “A/B testing”) that evaluate the effect of alternative course designs on robust learning outcomes. In courses based on cognitive science and data-driven modeling, student interaction is less focused on reading or listening to an instructor’s delivery of knowledge, but is primarily about students’ learning by example, by doing and by explaining.

Successes in Data-Driven Course Improvement
Both qualitative and quantitative techniques that combine subject matter expertise with cognitive psychology have been developed and successfully applied to educational data in numerous domains. We provide several examples to illustrate how data can be used to inform instruction.

Cognitive task analysis based on qualitative analysis of verbal data. Cognitive task analysis (CTA) focuses on the psychological processes behind task performance. More specifically, CTA uses a variety of techniques to elicit the knowledge of experts and differentiate between the critical decision making of experts and novices. It is a proven method for discovering latent
variables and unraveling some of the complexities of domain-specific learning. This method incorporates elements of cognitive psychology and domain expertise, and requires a high level of human interpretation of the data. By increasing the volume (e.g., as collected from MOOCs) and density (e.g., more frequent well-designed observations bearing on learners’ cognitive and self-regulatory states) of data, the need for human interpretation and the potential for subjectivity can be reduced but it cannot be eliminated.

The success of various types of cognitive task analyses has been demonstrated in a variety of courses where newly discovered factors led to course modifications and better student learning. Velmahos et al. used a CTA with surgeons to make improvements to a course on catheter insertion for medical interns [6]. When compared with the pre-existing course, the data-driven course redesign resulted in higher posttest scores and better surgery results (e.g., 50 percent fewer needle insertions). Lovett employed a CTA with statistics’ experts and discovered a “hidden skill” (variable type identification) as part of planning a statistical analysis [7]. We say “hidden” because expert instructors were not consciously aware of performing this planning step, nor were they aware of students’ difficulty with it. In using data-driven insights like these, interactive activities were designed for students to practice such hidden skills [8]; these activities are also a key part of the Open Learning Initiative’s online “Probability and Statistics” course. A randomized trial comparing blended use of this online course in a half semester to the preexisting full-semester course found students using the online course not only spent half the time learning, but learned more as demonstrated on a post-test measuring their conceptual understanding of statistics [8].

**Cognitive task analysis based on quantitative analysis of educational technology data.**

Traditional CTA techniques use qualitative data (e.g., interviews with instructors and students) to assist in making pedagogical decisions for course improvement. We have developed quantitative approaches to conducting CTA that are more efficient and scalable. The data generated from observing student performance is used to discover hidden skills and support course improvements. Early work of this kind involved comparison of student performance on systematically designed task variations designed to pinpoint what tasks (problems/questions/activities) cause students the most difficulty. These so-called “difficulty factor assessments” have led to many discoveries, perhaps the most striking of which is, in contrast to math educators predictions, algebra students are actually better at solving story problems than matched equations [5]. Results such as these have been critical to the design and continual improvement of the “Cognitive Tutor Algebra” course, now in use by some 600,000 middle and high school students a year. The most recent of many full-year randomized
field trials involved 147 schools and showed students using the “Cognitive Tutor Algebra” course learned twice as much as students in traditional algebra courses [9].

Heffernan and Koedinger employed a difficulty factors assessment that suggested the skill of composing multiple-operator expressions (e.g., as exercised in tasks like “substitute 40x for y in x=800-y”) is, surprisingly, a hidden component of translating story problems to equations [10]. Koedinger and McLaughlin ran an in vivo study (an A/B test) replacing some story problem practice with such substitution tasks [11]. They demonstrated significantly better learning on complex story problem translation for students who had more opportunities to practice substitution than those who did not. Stamper and Koedinger used a learning curve analysis of geometry tutor data to discover a hidden planning skill on problems that cannot be solved by simply applying a single formula [12]. Koedinger, Stamper, McLaughlin, and Nixon redesigned the tutor based on this discovery and compared it with the prior tutor in an in vivo experiment [13]. Students using the redesigned tutor reached tutor-determined mastery in 25 percent less time and did better on a paper post-test, especially on difficult problems requiring the hidden planning skill that was discovered.

The previous examples are just a few illustrations of the power of using data to improve instruction. A key question then is how learning environments and data collection systems can be best designed to “yield data that transform into explanatory models of a student’s learning, and also support course improvement” [3]? Having an understanding, an explanation, of how and why a model better predicts puts one in a much better position to use that understanding to make specific course redesigns.

**Opportunities for Improving MOOCs**

Before addressing the question of how to use data for improving MOOCs and courses in general, we first address the crucial question of designing learning activities to enhance data collection. Good course instrumentation for data gathering requires presenting complex tasks that represent learning objectives and identify students’ intermediate thinking processes as they perform these difficult tasks. Using strategies that emphasize student activity and scaffold student reasoning processes will help improve the quality of data, improve the inferences that can be made from data, and thus lead to better instructional design decisions (e.g., instructional modifications, re-sequencing of tasks).

Data-gathering. To build explanatory and actionable models, we need data that is fine grained in time and in thinking units. Observations that are finer grained in time provide more gradable
student actions per minute. Observations that are finer grained in thinking units help unpack how students are thinking, reasoning, or arriving at decisions. Many activities in MOOCs and online courses (e.g., multiple choice questions about simple facts) are too simple to provide much insight into student understanding and ability to apply what they have learned. Other activities are more complex, but only solicit final answers without recording intermediate reasoning steps. In Figure 1, for example, we see two entered answers (Parts A and B) for a complex physics problem without any of the steps, such as drawing the free body diagram or entering intermediate equations, leading up to these answers. The volume of such coarse grain data that are coming out of MOOCs will be of limited value, even if vast.

In Figure 2, we see a physics problem similar to the problem in Figure 1. However, in this case, students enter intermediate steps such as drawing vectors and writing a sequence of equations before offering a final solution. Such finer grained data provide more meaningful assessment beyond proficiency or completion, providing potential insights into aspects of reasoning or problem solving that are particularly challenging for students. Note in both examples students can make multiple (incorrect) attempts before arriving at a final (correct) solution.

![Figure 1. Coarse-grain data collection is illustrated in an online physics homework system called “MasteringPhysics,” where students use the keyboard to enter a final answer to a problem (e.g., in Part B an incorrect expression for the magnitude of a force $F_{ww}$ is provided by a student). In a single activity (a problem to solve), just two gradable student steps are observed and stored in the data log for later analysis. Source: Pearson MasteringPhysics.](http://ubiquity.acm.org)
Figure 2. Fine-grain data collection is illustrated in a physics intelligent tutor, Andes, where students use mouse clicks and the keyboard to draw diagrams (e.g., the coordinates and vectors are drawn over the given problem image in the middle left); define quantities (e.g., the variable T0 is defined as “the instant depicted” in the upper right); and enter equations (e.g., a first equation, “F_g_y+F_1_y=0”, which is incorrect, is shown in the middle right). In a single activity (a problem to solve), about 20 gradable student steps are observed live by the tutoring system, and stored in the data log for later analysis. Source: The Andes Project.

Activities can be more finely instrumented by providing workspaces (e.g., a free body diagramming tool or an equation solving worksheet) as illustrated in Figure 2 or by adding scaffolding that prompts for intermediate solution products (e.g., asking for converted fractions before their final sum). Intelligent tutoring systems [14] and some online courses [8] do support this finer grain action collection and, further, add data about timing, correctness, and amount of instructional help needed. Such data are much more informative than simple single-answer correctness. It is precisely this kind of fine-grained, multi-featured observations that are found in datasets in DataShop [15].
Much like the intricacies of a CTA that uncover the cognitive processes behind an observable task, making student thinking visible in an online activity is about more than having them "show their work" as they would on paper. First, there is the challenge of designing an interface or AI technology such that the work is computer interpretable. Second, it is often desirable to have students indicate their thinking in ways that they might not normally do on paper. As an example of both, consider asking students to explain the steps they take in solving geometry problems (e.g., by entering "the sum of the angles of a triangle is 180" to justify that 70 is the value of a given angle). Prompting students to perform such “self-explanation” has been experimentally demonstrated to enhance learning in math and science domains. Computer interpretation of student explanations has been achieved through both structured interfaces (e.g., menus of alternate explanations) and natural language processing technology. Structured interfaces are effective in enhancing student learning [16], but it remains an open question whether the extra effort of implementing natural language processing leads to further learning gains [17].

**Beyond black box models: From predictive to explanatory and actionable models.** In addition to avoiding the pitfall of developing interactive activities that do not provide enough useful data to reveal student thinking, MOOC developers and data miners must avoid potential pitfalls in the analysis and use of data. One such pitfall is the application of sophisticated statistical and machine learning techniques to educational data without understanding or contributing to relevant cognitive and pedagogical principles. This “black box model” approach focuses on improving prediction without regards to understanding what is happening cognitively (i.e., inside the box). Such understanding provides a means for instructional improvement as we illustrate below. It requires using analysis methods that focus on developing explanatory models to produce interpretable insights. Such insights advance understanding of learning and produce recommendations for improved educational practices.

One step toward explanatory models is annotating data sets with theoretically motivated labels or semantic features. DataShop helps researchers label student actions, such as the steps in Figure 2, with factors that might cause students difficulties in doing or learning. For example, one such hypothesized difficulty is whether steps in a geometry tutor require students to apply an area formula “backwards” (i.e., when the area is given) rather than the usual forward application (i.e., finding the area). To our surprise, a model incorporating this distinction across all formulas did not predict student data better than one without this distinction. However, we then used a model discovery algorithm [18] to find a particular situation, the circle area formula, where making a backward-forward distinction did improve model prediction. We return to the question of why in a moment, but it is worth highlighting that this learning factors
analysis (LFA) algorithm has been used on many DataShop data sets to discover better cognitive models of domain skills across a variety of domains (math, science, and language) and technologies (tutors, online courses, and educational games) [18]. LFA is an instance of a quantitative CTA discussed above.

Having well-labeled variables is important for developing explanatory models, but it is not sufficient. A second step toward explanatory and actionable models is applying psychological theory to interpret data-driven discoveries in terms of underlying cognitive processes. We started to interpret the better prediction of a model that splits, rather than merges, the forward versus backward applications of circle area by first verifying that the circle-area-to-radius (backward) steps were harder than circle-radius-to-area (forward) steps. An explanation, then, should indicate a cognitive process needed for the harder task that is not needed for the easier task: In this case, to undo the area formula ($A = \pi r^2$) to find $r$ requires knowing when to employ the square root operation. Finally, armed with such an explanation, a course developer can take action. The suggested action is to develop instruction and problems that better teach and practice the process of determining when to use the square root operation.

In general, more sophisticated algorithms need to be developed that unleash the potential instructional and learning benefits available from the big data obtained from MOOCs. But, to simply offer improved predictions (the standard goal in machine learning) without meaningful scientific discovery and practical implications is not sufficient. Instead, explanatory models of students are needed that uncover critical insights (e.g., that beginning algebra students are better at story problems than matched equations) or important nuances of student learning (e.g., that equations with “-x” terms are harder than ones with terms with numeric coefficients such as “3x”) [19] so as to support improved instruction and learning.

A model that makes more accurate predictions may not be insightful or actionable. Conversely, a model that only produces small prediction improvements may nevertheless produce actionable insights. In fact, the models produced by LFA typically yield only small (but reliable) reductions in prediction error. Nevertheless, such models have been usefully interpreted to suggest modifications to improve educational materials. Randomized controlled experiments have demonstrated that such modifications can yield reliable and substantial improvements in student learning efficiency and post-instruction effectiveness [13].

**Conclusion**

There are great opportunities for improving MOOCs through data-driven learner modeling. However, the computer science community needs to better recognize and engage the existing
state of knowledge in learning science and educational data mining. If not, the rich volume of principles of learning and instruction already produced by learning science research is at risk of being very slowly rediscovered by the new players to online course development. Educational data mining research has established a great potential for insights on student cognition, metacognition, motivation, and affect [20]. These insights have been possible only because the data used to derive them has come from student learning interactions that are both complex and fine-grained—of the kind produced in intelligent tutoring systems or other online activities involving multi-step interfaces (e.g., simulations, games, mini-tutors) [21]. Further, these insights have been used to make design changes in online systems and, in some cases, experiments have demonstrated significant improvements in student learning, metacognition, or motivation by comparing the redesigned system to the original one [13].

Using data-driven learner models to improve courses contrasts with the instructor-centered model in three key ways. First, course development and improvement is based not solely on instructor self-reflection, but on a data-driven analysis of student difficulties and of the target expertise the course is meant to produce. Second, course improvement is based on course-embedded in vivo experiments that evaluate the effect of alternative course designs on robust learning outcomes. Third, course interaction is not centrally about instructor’s delivery knowledge, but about student learning by example, by doing and by explaining.

For data-driven learner modeling to yield greater understanding and improvement of student learning, we recommend more emphasis on (a) exploratory data analysis in addition to machine learning, (b) simpler models with fewer parameters as well as highly complex models, (c) use of explicit research questions to drive analyses, and (d) inclusion of cognitive psychology expertise to guide online activity designs that make thinking visible and to aid interpretation of model results.

More finely instrumented activities are not only valuable for making thinking visible and improving data for cognitive diagnosis, such fine-grain data are crucial to systems that are now making reliable inferences about students’ motivations and affective states [20]. While the examples above have emphasized monitoring and analyzing cognitive functions, such as reasoning and problem solving, other educational data mining research has investigated roles of metacognition, motivation, and social dialogue in learning.

There are some good signs of recent progress in useful mining of MOOC data [22, 23], but more such work is needed. MOOCs and other forms of online learning provide a tremendous opportunity to enhance education and diversify learning if productive collaborations are formed and pitfalls of insufficient data-gathering and black box prediction are avoided. The
more than 450 datasets already available in DataShop offer opportunities to develop data-driven models of learners that include conceptual understanding, cognitive skills, metacognitive and learning skills, general dispositions and motivations toward learning, and specific states of affect (e.g., confusion or flow) during learning [15]. These data-driven learner models provide great potential to advance both learning science and educational practice.

References


**About the Author**

Ken Koedinger is Professor of Human-Computer Interaction and Psychology at Carnegie Mellon. He directs LearnLab, which leverages cognitive and computational approaches to support researchers in investigating the instructional conditions that cause robust student learning.

Elizabeth A. McLaughlin is a Research Associate at the Human-Computer Interaction Institute at Carnegie Mellon University.

John Stamper is a member of the research faculty at the Human-Computer Interaction Institute at Carnegie Mellon University and is the Technical Director of LearnLab’s DataShop.

**DOI:** 10.1145/2591682