Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes

Ryan S. Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

Learning analytics and educational data mining have come a long way in a short time. In this article, a lightly-edited transcript of a keynote talk at the Learning Analytics and Knowledge Conference in 2019, I present a vision for some directions I believe the field should go: towards greater interpretability, generalizability, transferability, applicability, and with clearer evidence for effectiveness. I pose these potential directions as a set of six contests, with concrete criteria for what would represent successful progress in each of these areas: the Baker Learning Analytics Prizes (BLAP). Solving these challenges will bring the field closer to achieving its full potential of using data to benefit learners and transform education for the better.

Keywords: Baker Learning Analytics Prizes, BLAP, learning analytics, educational data mining, model generalizability

The text below represents a lightly edited version of a keynote talk given by Ryan S. Baker at the 9th International Conference on Learning Analytics and Knowledge, in March 2019.

1. INTRODUCTION

Thank you all for having me here today, this is a really great honor. Learning analytics and educational data mining have been very successful in just a short time. For learning analytics, it's only been nine short years since the first conference.

In this time, student-at-risk prediction systems are now used at scale in higher ed and K12, and they're making a difference. And adaptive learning systems are now being used at-scale in higher ed and K12, and they're making a difference as well.

And there's been a steady stream of discoveries and models in a range of once difficult areas to study: collaborative learning, classroom participation and online connections, motivation and engagement, meta-cognition and self-regulated learning, and many other areas.

And I could give a talk about all of those things, full of praise and shout-outs for all the great research that has happened, and all the great people in this room. The talk would be full of warm fuzzies, and we'd all forget it by tomorrow afternoon.

So instead of talking about the last nine years, I'd like to double that and go the other direction and talk about the next 18 years, twice as long as the history of the LAK conference so far.

But before I do that, I'd like to say a word about David Hilbert. Who here has heard of David Hilbert? David Hilbert was a mathematician, a visionary, a wearer of spiffy hats (Figure 1).

Figure 1. David Hilbert and his spiffy hat. (Image taken from Wikipedia, where it is listed as being in the public domain).

In 1900, Hilbert gave a talk at the International Congress of Mathematicians (Hilbert, 1900), and at this talk, he outlined some of the problems that he thought would be particularly important for mathematicians over the following years. And this talk was, in my opinion, one most eloquent scientific speeches of all time. I encourage you to read it; it's available in its full text at (Hilbert, 1900).

Hilbert did not just give a list of general problems. He framed the problems concretely, he discussed what it would take to solve these problems, and he listed what would be necessary to demonstrate that the problems were solved. He set the goal that each problem would be clear and easy to understand, significant in impact, and difficult but not completely inaccessible. And he achieved each of these goals, except perhaps the final one. His problems were hard problems, perhaps too hard in many cases. Only 10 of his 23 problems have been solved so far.

In the years since, there's been a proliferation of these kinds of problems, with many lists of problems or grand challenges, including many in our field (e.g., Woolf et al., 2013; see a review of earlier challenges in Kay, 2012). And yet few of these have been anywhere near as influential as Hilbert's problems. Most of these challenges are very different than what Hilbert listed. They just list big, difficult problems, which is very different from what Hilbert did. Of course, there are obvious exceptions, like the Turing test/Loebner Prize (Floridi, Taddeo, & Turilli, 2009) and so on. But the fact remains, lots and lots of grand challenges, but very few that had anywhere near the impact of what Hilbert talked about.

Today I'd like to suggest a list of problems to you for learning analytics. And I know I'm no Hilbert. I do like spiffy hats (see Figure 2), but still, no Hilbert.

Figure 2. Ryan Baker and his spiffy hat.

Also, learning analytics isn't mathematics. It's a lot easier to frame concrete problems and what progress on those problems looks like in a domain like math, I think, than in a domain like learning analytics. But I hope you'll still, nonetheless, give me a few moments of your time to discuss what I see as some of the bigger upcoming challenges in our field. These challenges are not necessarily new to this talk. What is hopefully new is the conscious attempt to emulate Hilbert by trying to frame specific problems, with conditions for how we know that we'll have made concrete progress on solving these problems.

In this talk, I will frame these problems as contests – a set of contests to see who can solve them first. I'll pose a general problem, discuss a milestone on the way to solving that problem (a contest), and discuss the concrete conditions that would establish the contest has been won. Winning one of these contests won't solve the problem in full, but it will represent real and measurable progress.

2. PROBLEMS

2.1. PROBLEM 1. BLAP1. TRANSFERABILITY: THE (LEARNING SYSTEM) WALL

The first of these problems, BLAP1 (more on this acronym later in the talk), is what I call the learning system wall. Modern learning systems don't just help students learn; they learn about the student. A modern learning system learns a great deal about the student – their knowledge at minimum, and increasingly their motivation, engagement, and self-regulated learning strategies. But then the next learning system starts from scratch.

A student might use DreamBox one year, Cognitive Tutor a couple of years later, ALEKS a couple of years after that. These are all great systems. They all learn a lot about their students, and what they learn is forgotten the second they move on to the next system.

Or imagine a student – all too commonplace these days – who is at a school that is using two learning platforms at the same time. A student might use Dreambox for some lessons and Khan Academy for others. The student is using two systems at the same time to cover the same content, and each system has to discover the exact same things about that student.

It's like there's a wall between our learning systems and no information can get in or out. If you seek better learning for students, tear down this wall!

Worse yet: This is not just a between-systems problem; it's even a problem between lessons in a single system. In most systems, a student's struggle or rapid success in one lesson usually doesn't influence the inferences the system makes in the next lesson.

There's been some early progress on this. I want to always give credit where credit's due, if I know about it. Sosnovsky et al. (2007) show that it is possible to design mappings between courses where information about student progress in one course is informative about initial knowledge for another course, within the same platform – but they were not able to actually integrate the two courses' student models in practice, due to technical limitations. Eagle et al. (2016) have shown that there could be better student models if we transferred information between lessons, within a student, and within a single platform. But it was just a secondary data analysis, and it only involved three lessons in one system. So these were in many ways still very basic demonstrations of what might be possible.

So I have a contest for you, a challenge, which I hope some of you would choose to participate in. Take a student model developed using interaction data from one learning system. Now take this model's inferences for a student, let's call her Maria, who has used that system. Take a second learning system, developed by a different team. Use system one's model inference to change system two's model inference for the same student. In other words: we have some info from the first system. We transfer it to the second system, and the second system knows something. But that's not the full challenge. We need to go beyond just changing what the system knows, because if we just change what the system knows and we don't do anything with that knowledge, who cares? So, to complete this challenge, the second system needs to change its behavior for Maria as well.

So, to summarize: the first learning system will infer something about the student, and the second learning system will adjust its inferences based upon that information, and that second system will change its behavior based on the changed inference.

The change could be just about anything. It could be different content the student starts with: maybe we just start the student in a different place. We know that Maria knows fraction division, we don't need to cover it again. It could be a different learning rate – Liu and Koedinger (2015) showed that student data could be fit better if we assume different students have different learning rates. We know that Maria is a fast learner, so we can assume that she'll learn a little faster in the second system. It could be a different interpretation of incorrect answers or other behavior. Maybe when Maria was using the first system, her behavior showed that she has a misconception around adding the denominators, and the second system can therefore interpret her errors with higher confidence.

In this first contest, BLAP1, the original model for the second system would have to be a good model for that construct. We want to improve on a model that's already decent, not a model that is at chance. It's fairly trivial to do better than a model that's at chance. By contrast, if we start with a model that's already decent on the second system, with goodness metrics on held-out data that are good enough to be published on their own, in JEDM or EDM or LAK or JLA after, let's say 2015. So for example for behavioral disengagement, like gaming the system or off-task behavior, an AUC ROC of 0.75, or 0.65 for affect, or 0.65 for latent knowledge estimation (predicting immediate correctness). These numbers are all near the state of the art for 2015. And, of course, publication in one of those venues in 2015 or after would also be good enough.

The new model for the second system would have to be able to take an entirely new set of students, and given the information from the first system, achieve better prediction than that

original model. The original model for the second system should be decent, and the new model has to do better than that.

Finally, the system behavior change would have to be able to actually run in the system. For this challenge, it's necessary to actually connect the systems and change system behavior based on the revised student model, not just conduct an analysis for the sake of publishing. After all, it would be reasonably easy to get offline data from two learning systems for the same students, link the data sets together one time, and publish it. But actually getting the two systems to talk to each other? That's where the real challenge comes from – but also where the opportunity comes from.

Problem 1: BLAP1. Transferability: The (learning system) Wall

Transfer student model from learning system A to learning system B

Improve an already-good student model in learning system B

Change behavior of learning system B in runnable fashion

2.2. Problem 2. BLAP2. Effectiveness: Differentiating Interventions and Changing Lives

The second of these problems, BLAP2, I call differentiating interventions and changing lives.

Today, we have many platforms that infer which students are at-risk on the basis of learning analytics from learning management system data, or other university or K12 data. These systems are being used at scale by instructors, teachers, guidance counselors, academic advisors, and other school/university staff to make decisions about how to better support students, for example by selecting students for targeted interventions. This has been a substantial shift over the last decade, as these systems have emerged and really started to make a difference. Increasingly, there is also evidence that these systems lead to better outcomes for students (e.g., Arnold & Pistilli, 2012; Milliron, Malcolm, & Kil, 2014).

However, it's not clear how durable a difference this type of work is making. It is worth asking the question: are we really changing lives, or are we patching short-term problems?

To answer this more conclusively, I propose a second contest, BLAP2. The contest is: take a group of undergraduates enrolled in an accredited university, whatever that means in the local context, and randomly assign them to a condition: intervention/ experimental, or no intervention/control. If that's not feasible to do (and often it isn't), take students already assigned to condition and set up a quasi-experimental comparison done to a publishably high standard. The differential treatment due to assignment to conditions could last up to a year, or it could be as brief as a single course session.

Next, assign your learning analytics-based intervention to a subset of students in the experimental condition where a machine-learned model, or a knowledge-engineered model, or some other well-defined criterion, determines which students actually receive the intervention. According to this criterion, some number of students: between, say, 3% and 50% of students in the experimental condition, will actually get the intervention. The researcher must publish or publicly declare what that model or criterion is. It's important that not all students get the intervention, because if we just give it to everybody – if it's a universal intervention – the model

wasn't needed in the first place. The value of having a model is that we can give the intervention to only the students who need it.

The next step is to identify in advance – and document, through pre-registration or some other public mechanism – four groups of students, in a two-by-two cross, shown in Table 1: experimental vs. control, and model recommends intervention vs. model does not recommend intervention. Only one group gets the intervention: the students who are assigned to the experimental condition **AND** the model thinks they should get an intervention. If the model thinks a student should get an intervention, but they're in the control condition, no intervention. If the model doesn't think a student should get an intervention, they don't get an intervention, even if they're in the experimental condition. They only get the intervention in one of the four cases.

Table 1. The conditions of the study in the BLAP2 contest

	Experimental Condition (E)	Control Condition (C)
Model thinks should receive intervention	**E*: Receives intervention**	C*: Does not receive intervention
Model does not think should receive intervention	E&: Does not receive intervention	C&: Does not receive intervention

At least three years later, collect some success outcome that matters. For this contest, that success outcome could be just about anything. It could be a standardized test score. For example, you could give a group of students an intervention in the first year of college and then see if they do better on the GRE. Or you could measure whether students attend graduate school, whether they achieve employment in their chosen field, their personal income ten years later, their personal happiness. There's a lot of possible measures that it could be. Whatever that long-term success outcome is, the goal is to make an enduring change through an intervention that is targeted to specific students. Demonstrated by showing, first, that the students in the experimental condition who received the intervention perform statistically significantly better than the students in the control condition who would have gotten the intervention based on the published model, with some reasonably high effect size. And demonstrated also by showing, second, that the students in the experimental condition who didn't get the intervention don't perform statistically better than the students in the control condition who would not have received the intervention. By testing both the possibilities, we can demonstrate that our analytics-driven intervention is what is really making the difference.

Problem 2: BLAP2. Effectiveness: Differentiating Interventions and Changing Lives

Publicize criterion for intervention

Assign students to control or experimental group

Use analytics, only within experimental group, to assign intervention

Collect longer-term outcome measure

Demonstrate that experimental/analytics-intervention group performs better than experimental/analytics-no-intervention group

But that experimental/analytics-no-intervention group does not perform better than control/analytics-no-intervention group

2.3. Problem 3. BLAP3. Interpretability: Instructors Speak Spanish, Algorithms Speak Swahili

The third problem, BLAP3, I call instructors speak Spanish, algorithms speak Swahili. It's in homage to two great world languages which don't have all that many speakers in common.

Take the case where an analytics development team has put a ton of effort into building a model that captures an important phenomenon. They've crafted the perfect recurrent neural network and validated that it has excellent predictive performance. And then this brilliant model makes a prediction that a user, an instructor for instance, finds non-intuitive. The instructor doesn't understand what that prediction means. And the development team can't explain it, because their brilliant model is utterly uninterpretable and inscrutable.

And the instructor, reasonably enough, doesn't trust the model. And then they don't use it.

That's a big problem.

If the user doesn't understand the model, and it tells them something that they don't believe, there's a risk they will just say it's wrong, even when it's right.

So the challenge is, make the decision making processes of a recurrent neural network, or a comparably advanced and complex algorithm, set to model a learning analytics phenomenon, understandable for an instructor or other similar stakeholder, for example an academic advisor, who doesn't personally have a data science background already. There's already been work to make learner models more scrutable (Bull & Kay, 2007), as well as work to make complex machine learning models more understandable (Zhang & Zhu, 2018), but I think this challenge will require building on this work to take it to a new level.

More specifically, for this challenge, the task is to take a model that predicts a learner success outcome, such as high school dropout, or college course failure, using an "advanced algorithm" with at least 100 parameters. I'm putting quotation marks around "advanced algorithm," because as most of the attendees of this talk probably know, a recurrent neural network is not going to have 100 parameters, it's going to have closer to 10 million parameters.

The model has to be a good model for the construct. As with BLAP1, the model needs to have goodness metrics on held-out data that are good enough to appear in an academic conference or journal in this field.

Once you've got that good model, find five data scientists and five instructors (or other model uses) who weren't part of your original development team. Design an explanation of how the algorithm works. It could be a visualization, a video, a text, an interactive system. Any of these are fine, so long as there is not a human being available to answer questions. Give the data scientists and users five case studies or examples of specific students. Ask each of them to tell you what decision the algorithm would make for each student and explain why.

Once this study's data has been collected, have two independent researchers with an appropriate methodological background code the data for the reasons why the participants thought the algorithm would make each decision, and verify that they obtain acceptable inter-rater reliability (say, Cohen's Kappa over 0.6). Do the instructors agree with the data scientists, say, at least 80% of the time? Both in terms of what the model's decision will be and what its reasons.

By doing this, by taking a data scientist and an instructor, creating a way to communicate the model to them, and making sure they have a comparable understanding of how it works, we'll have made progress towards making these inscrutable models more scrutable, more understandable, and ultimately, more trustworthy. And by doing that, we can create a dialogue. Maybe when users of these models don't trust them, they'll talk to the developers, and together both groups can figure out what the breakdown is. At the very least, even if they don't always agree on what the model *should* be, if the users understand how the model makes its decisions, they're more likely to have confidence in it and continue to use it, than if it's just a black box that makes incomprehensible predictions.

Problem 3: BLAP3. Interpretability: Instructors Speak Spanish, Algorithms Speak Swahili

Take a complex model of a learning analytics phenomenon

Develop a no-human-in-the-loop method of explaining the model

Present the explanation to five (new) data scientists and users

Ask the participants to explain what decision the model will make, and why, for five case studies

Code the explanations of the model's decisions

Verify if the data scientists and users interpret the model the same way for the case studies

2.4. Problem 4. BLAP4. Applicability: Knowledge Tracing Beyond the Screen

Fourth challenge, we're halfway there: Knowledge tracing beyond the screen.

As a community, we've been reasonably successful at producing models that can infer learner knowledge, or at least predict immediate correctness, in computer-based learning environments (e.g., Corbett & Anderson, 1995; Pavlik, Cen, Koedinger, 2009; Khajah et al., 2016; Zhang et al., 2017). And when I say reasonably successful, I mean that there are a large number of

products commercially available, that call themselves "adaptive learning", and have these models running in the background.

But mostly, these environments involve one student sitting at one computer, providing textual input: numbers, multiple choice responses, and so on. There are exceptions to the textual input – Inq-ITS, for instance, infers complex reasoning demonstrated in simulations (Sao Pedro et al., 2013), but this type of system is very much the exception. And it's still one student sitting at one computer.

The problem is that most learning doesn't take place with one student sitting at one computer. A great deal of learning occurs in collaborative project work, and discussion forums, and classrooms where teachers and students talk to each other. And certainly, there has been a considerable amount of work by EDM/LAK researchers in these contexts (e.g., Martinez-Maldonado et al., 2012; Bernger, Walker, & Ogan, 2017; Pijeira-Diaz et al., 2019). The challenge is, though, can we detect student *knowledge* within these contexts to the same quality that we can in easier to study contexts?

Therefore, I propose the following contest. Take audio, visual, and/or physical data on learning from a setting where there are at least four students in the same physical space, engaged in the same learning activity at the same time. The learning activity should involve interaction between these students that isn't wholly mediated by computers – i.e., the four students shouldn't just be sitting at four computers, interacting directly with only the computers.

With this data, build a model that can infer at least four distinct skills or knowledge components for each student. You might ask, why isn't one knowledge component enough? The reason is because I think it would be possible to infer a single knowledge component just from whether the students are engaged – just from whether they're participating. Distinguishing four different skills at the same time requires actually recognizing the cognition or the learning, and what skill it pertains to.

In terms of validation, this model has to be able to predict immediate future performance on these four or more skills, the same standard that we often use for validating knowledge models in existing one-on-one systems (e.g., Pavlik, Cen, Koedinger, 2009; Khajah et al., 2016; Zhang et al., 2017). And for a sample of new students, the model has to achieve AUC ROC in the ballpark typically seen for BKT for existing systems, say 0.65. If we can establish the same level of quality for models of student knowledge in collaborative and real-world settings as we can for one-on-one learning with a computer, we'll have taken what is the most widely used type of assessment within adaptive learning and made it accessible for a much wider range of learning activities.

Problem 4. BLAP4. Applicability: Knowledge Tracing Beyond the Screen

Take data from at least four students completing learning activity together

Model at least four distinct skills for each student

Predict immediate future performance for these skills

2.5. Problem 5. BLAP5. Generalizability: The General-Purpose Boredom Detector

Now on to the fifth challenge, generalizability: the general-purpose boredom detector.

There's been a lot of success in affect detection, on detecting academic emotions through physical sensors and video (Woolf et al., 2009; AlZoubi et al., 2009) as well as solely from interactions (D'Mello et al., 2008; Pardos et al., 2014), and a combination of the two (Bosch et al., 2015; DeFalco et al., 2018). Across studies, this work has detected a range of affective states, including boredom, engaged concentration, frustration, confusion, and delight. I've put a few moments of thought into this problem myself.

A big problem with the current generation of affect detection for learning is that these current models are not generalizable across systems. They have to be rebuilt almost from scratch for new learning platforms. There are some common tools for field observation, the HART app that my lab developed (Ocumpaugh et al., 2015), and some tools for data synchronization. There's some experience in feature engineering that generalizes. Several labs have gotten faster at building new affect detectors, over the years. It's not as expensive as it used to be, but it's still a lot of work. Fiona Hollands and Ipek Bakir (2015) estimated that building these kinds of affect detector models for a new system costs between $40,000 and $75,000. That's not prohibitively expensive, but it's certainly not free either.

Even when you build these models for a system, our learning systems are always changing. It's not yet clear what changes to a learning system causes affect detectors to break down. For example, my colleagues and I have seen that interaction-based models of whether a student is confused break down when hints are removed from the learning system (Richey et al., under review), but models of gaming the system (not quite affect) seem to keep working fine even after an automated agent is added (Baker et al., 2006). Why does the model transfer in one case but not the other case? We don't entirely know yet.

In this contest, we take on the challenge of model generalizability across systems. To win this contest, you first build a model of student boredom using interaction data from one or more learning systems. Then take that model and apply it to data from an entirely new system, built by a different development team, where the interaction is not broadly identical. Not broadly identical means the student solves problems, and/or is scaffolded in different ways. Generalizing across systems developed by different development teams, who have developed systems that support the learner in different ways, is the type of approach that makes a credible case for broader generalization. To give an example, take ASSISTments and MATHia. Both have content for middle school mathematics, but the design of the activities is fairly different, so this would be fair game.

To really demonstrate generalizability, for this contest, it is necessary to apply the model with no tweaking, no refitting, no modifications at all. The model you build in the first system is the model you use in the second system. To make this feasible at all, the features used in the models would have to be defined in the same way, in some way that's general across systems. It would be OK, for instance, to say that if a student is one standard deviation slower than the mean speed in terms of that system's speed, we can use that same feature of being one standard deviation slower than the mean speed in the other system. The numbers might not be exactly the same, but the feature is generic and calculated exactly the same way for both systems. By contrast, it wouldn't be OK to refit all the cutoffs for the new system. If you did that, the detector would not longer be something that could just be applied -- you'd have to collect data, you'd have to re-train, and so on.

Once the model can be applied, the next step would be to collect ground truth of some sort, to validate it. Any reasonable, standard type of ground truth would be acceptable for this contest.

The ground truth labels could be obtained through self-report, or field observations, or video coding. They would have to meet standard expectations for reliability, of course, such as the current conventional cut-off in our field of 0.6 for Cohen's Kappa for field observations or video coding of affect. And then, during validation, you would need to demonstrate that the new model achieves the same quality of prediction in the new system as would be expected for an entirely new model today in 2019 – an AUC ROC of 0.65.

It's worth mentioning some of the early progress towards solving this challenge, already. There are two examples I'd like to highlight. The first is Paquette et al.'s (2015) work. Paquette built a detector of gaming the system on interaction data from the Cognitive Tutor system, now called MATHia. He then applied his model on data from the ASSISTments system, with no modification. It was still able to accurately predict if a student was gaming the system.

His method of developing this detector was a really interesting and unusual effort where instead of using a standard machine learning approach, he actually did very thorough knowledge engineering where he interviewed a person who'd done a lot of this coding and built a model of her process, took it back to her, talked about it with her, took it back, modified it some more, and then actually ran some data through the model, with her watching. She commented on it, and they went back and forth until they both felt that the model fully captured her reasoning. This effort took place in Cognitive Tutor, but then Paquette applied the resulting model to ASSISTments, and it worked.

This result represents evidence that system-general models are possible for learning analytics. But on the other hand, because gaming the system is something that we can verbalize, to a greater degree than how we recognize student boredom, it represents an easier challenge than this contest.

Another piece of work that represents progress towards meeting this challenge is Hutt et al.'s (2019) paper at ACM SIGCHI this year, where they built a detector of affect using machine learning for an algebra course, and then they validated that it worked on a geometry course built in the same learning system/platform. It was a different course, with different kinds of content, but still in the same learning system with the same types of interactions.

Overall, though, I think it's fair to say that there has been considerable progress on this contest. I expect that this contest will be won relatively sooner than some of the other contests I've talked about today.

Problem 5. BLAP5. Generalizability: The General-Purpose Boredom Detector

Build an automated detector of affect

Demonstrate that the detector works for an entirely new learning system
with different interactions
with AUC ROC >= 0.65

2.6. Problem 6. BLAP6. Generalizability: The New York City and Marfa Problem

Let's discuss the sixth challenge, the last challenge, the New York City and Marfa problem. Learning analytics models are mostly built on the samples that we have ready at hand, whether it's the current population of students at a university developing a model, the current user base

of the adaptive learning system we're building the model for, or just students who are relatively easy to survey or observe.

But what happens when the population changes? Say you've built a model for your university in Utah, but your university starts taking in a lot of transfer students from Nevada. I worked with one university system where they had something very much like this happen. A for-profit college in their city shut down, and they got a lot of transfer students enrolling all at one time. And now they were uncertain if their graduation prediction models would still work.

There are many cases like this. Maybe you've built your learning system on data from students in the continental U, and now it's being adopted in Alaska. Will your models work? Maybe you surveyed or observed one group of students, and now it's not clear if the model will work for different groups of students. The ASSISTments system built detectors of student affect for students in Massachusetts and Maine (Ocumpaugh et al., 2014). Now we want to use those detectors in North Carolina. Do we trust the detectors?

This is not just a technical challenge, it's a challenge for inclusion, because a lot of the populations that we want to focus on including – a lot of historically underserved and underrepresented populations -- are the ones it's harder to collect data for.

I call this the New York City and Marfa problem. Why? Well, who here has tried to do research in New York City? A few hands. Those of you who raised your hands know it's hard to collect data and do research in New York City because of very restrictive rules. You have to go get fingerprinted. You have to have everybody in your lab go get fingerprinted. They have to be fingerprinted at one office in Brooklyn. You can't be fingerprinted somewhere else. If you're a researcher in Philadelphia, that's costly and time-consuming. But it's more than just that. In other communities, you get a teacher, they say they like the study, they write you a letter. Maybe they go to their assistant principal, and the assistant principal writes you a letter. Now you're all set. In New York City you're not even supposed to talk to the school until you've gone through the New York City DOE IRB process. Which adds additional challenges, because it's hard to design a study that will work well in a school when you're not allowed to talk to them. It increases inconvenience for the school. And on top of that, the NYC DOE IRB has very restrictive policies, which are much more restrictive than the federal legislation for IRBs. The people running the NYC DOE IRB are reasonable and good people, but for various reasons they've been stuck with a clunky and cumbersome set of rules they have to follow. All told, running a study in New York City takes about nine months of preparation time in addition to what you do to run a study in another city. It's possible – my team has run studies in New York City -- but it's a lot more work.

Compare that to Marfa, Texas. Marfa, Texas does not have incredibly restrictive policies on educational research. But, it's 194 miles from El Paso International Airport, which is not exactly a huge airport itself. If I wanted to run a study in Marfa, I first have to fly from Philadelphia to Dallas. Then I have to fly from Dallas to El Paso. Then I have to rent a car, and oops – not all of my grad students have drivers licenses -- and drive 194 miles through the desert, along the Texas-Mexico border. So it's not exactly the most convenient place to do research, and because of that, towns like Marfa tend to be included in research a lot less often than the suburbs of major cities.

Because of these kinds of factors, most education research in the United States is done in upper-middle-class suburbs and relatively smaller cities. From what I hear from my colleagues, I think similar patterns occur in other countries, just with differences depending on what the local conditions are.

We want our analytics to be just as valid for the students in New York City and Marfa, Texas, as for students who are easier to research.

One solution that we talk about in (Ocumpaugh et al., 2014) is to collect data from all the populations you want the model to work on and try to validate your models on all these populations. That's both a sensible solution, and completely impractical. Because, as I just brought up, some populations are difficult to get data for. It's more feasible to collect all the data in MOOCs and blended learning systems, for example. So you might say, OK, this isn't a problem for MOOCs. But even in MOOCs, we don't entirely know what the relevant populations are. Which are the groups of learners that we need to validate our models on?

So my challenge, this sixth and last challenge, is to develop a model that *just works* for new populations. The contest is to build a model to predict one of the following outcomes or measures: high school dropout, college course failure, affect, disengaged behavior, or learning strategy. I pick these because these are things that have all been relatively studied. We know that all of these can be modeled using contemporary learning analytics methods. As with the previous challenges, the model again has to be a good model for the construct, publishable at EDM or LAK after 2015.

Once we've built that model, the next step is to collect data for a new population that's substantially different from the original population. It's not enough to just hold out a little bit of the data in the first place, to win this contest you have to actually go collect a new data set. Because as you all probably know, even when you hold out the data set, it's kind of easy to snoop your data and use the test data to see how you're doing. It's hard to really prevent yourself from doing so, and hard to verify as a reviewer that the author didn't do this. By contrast, if you collect a new data set that wasn't available to you when the original model was being developed, then you can put a check on yourself and prevent honestly motivated behaviors that nonetheless call the results into question.

Regarding this new data set that is collected. The new data set should be substantially different from the original one. The population should be more than half belonging to some group that was rare in the original data set, say 10% or less, consisting of one of the following dimensions defined however the census categories are defined in your country: race, ethnicity, native language, citizenship, poverty, degree of urbanicity – rural versus urban versus suburban. In other words, some demographic difference that the broader public or policy environment thinks makes a difference. Collect a new sample that is dominant in terms of that group, which was not well represented in your original population.

The model's prediction for your new population has to have limited degradation, less than 0.1 in AUC ROC or Pearson or Spearman correlation. The predictions also have to remain better than chance. If you go down from say 0.58 to point 0.49 in AUC ROC, you haven't actually accomplished much. If you go down from 0.85 in AUC ROC to 0.63, maybe you're still better than chance, but there's something that's changed a lot for your model's applicability, and it's worth finding out what it is.

Problem 6. BLAP6. Generalizability: The New York City and Marfa Problem

Build an automated detector for a commonly-seen outcome or measure

Collect a new population distinct from the original population

Demonstrate that the detector works for the new population
with degradation of quality under 0.1 (AUC ROC, Pearson/Spearman correlation)
and remaining better than chance

So to reprise these challenges, briefly. BLAP1, the learning system wall. Build a model in one system and have it transfer information to another system in a way that makes that system's predictions better and changes the system's behavior. BLAP2, differentiating interventions, and changing lives. Do a rigorous test that shows that learners who received your intervention have better outcomes in the longer term than learners who didn't, in a controlled study, where you also demonstrate that it's not simply a case where everyone in the experimental condition benefitted more. BLAP3, instructors speak Spanish, algorithms speak Swahili. Build an explanation of a complex model that practitioners can understand as well as data scientists. BLAP4, knowledge tracing beyond the screen. Get the same quality of knowledge tracing for learning that occurs in groups, or in classrooms, as for learning that occurs in one student, using one computer. BLAP5, the general-purpose boredom detector. A detector of boredom that just works in an entirely new system. And finally, BLAP6, the New York City and Marfa problem. Build a model that just works on a new population. These represent a set of challenges that represent a span of problems that our field needs to solve to reach its full potential -- although clearly, they don't cover everything.

3. PRIZES

I'd like to announce a prize here, today at Learning Analytics and Knowledge 2019, that will go to the teams that are first to solve each of these problems.

But first, a word on how this prize was established. We all know that there are many generous billionaires out there in this day and age, who strive to give back to the world what they've earned. Who yearn to better support education however they can, marshaling their resources, their intelligence, their efforts, to make this world a better place. Individuals for whom $45,000 is the merest pocket change, not even worth picking up if it fell on the streets.

Unfortunately, none of them will take my calls.

So, I'd like to announce the Baker Learning Analytics Prizes, BLAP. Successfully completing one of these contests will involve an award of – drum roll please – one US dollar.

Yes, if you solve one of these six challenges, I will give you $1. $1 for each challenge.

4. CONCLUSIONS

In this talk, I proposed six contests, representing six problems whose solution I believe would bring our field forward, and conditions under which we would know there's been progress on each of these problems. I hope you found these ideas compelling or at least thought-provoking.

Ultimately, the field moves forward if it takes on big goals that make a difference. One of the things we have to watch out for is becoming obsessed with tiny optimizations on small problems. It's really easy to say, "My cool new algorithm can do 0.003 better on a data set that everyone agrees is a good data set to apply algorithms to."

You can get published for that pretty easily. And incremental progress can be valuable. But it shouldn't be the primary focus of a field. In this talk, I've tried to propose some alternate goals, some challenges that will be harder to solve (although, to return to Hilbert, difficult but not completely inaccessible).

What I've presented today might not be the right big goals. If not, I hope I have provoked you to think about what the right big goals would be. If it's not generalizability, interpretability, applicability, transferability, and effectiveness, then what are the right challenges? What problems should we be taking on, as a field?

Whatever you think those are, I encourage you to go out and pursue those.

I look forward to seeing you all in 2037, when I hope we'll have achieved all these goals, or everyone will agree that these are really bad ideas in the first place.

Thank you.

5. FINAL NOTE

For further details on the BLAP competitions, their winners, progress towards BLAP, and other resources, please see http://www.upenn.edu/learninganalytics/blap.html . I will also make major announcements through my lab Twitter feed, @BakerEDMLab.

6. ACKNOWLEDGMENTS

I would like to thank several individuals for their suggestions and helpful feedback on this talk, and the ideas in it, including Alex Bowers, Christopher Brooks, Heeryung Choi, Neil Heffernan, Shamya Karumbaiah, Yoon Jeon Kim, Richard Scruggs, Stephanie Teasley, and all of the people who commented on this talk during the session, after the talk at the conference, and on twitter and through email. All the bad ideas remain wholly mine.

7. REFERENCES

AlZoubi, O., Calvo, R. A., & Stevens, R. H. (2009, December). Classification of EEG for affect recognition: an adaptive approach. In *Australasian Joint Conference on Artificial Intelligence*, Springer, Berlin, Heidelberg, 52-61.

Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, ACM, 267-270

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006) Adapting to when students game an intelligent tutoring system. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

Bergner, Y., Walker, E., & Ogan, A. (2017). Dynamic Bayesian network models for peer tutoring interactions. In *Innovative Assessment of Collaboration*, Springer, Cham, 249-268.

Bosch, N., Chen, H., Baker, R., Shute, V., D'Mello, S. (2015) Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 17th International Conference on Multimodal Interaction*, 267-274.

Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI:() open learner modeling framework. *International Journal of Artificial Intelligence in Education*, *17*, 2, 89-120.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 4, 253-278.

DeFalco, J.A., Rowe, J.P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B.W., Baker, R.S., Lester, J.C. (2018) Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence and Education, 28*, 2, 152-193.

D'Mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18,* 1-2, 45-80.

Eagle, M., Corbett, A., Stamper, J., McLaren, B., Baker, R.S. (2016) predicting individual differences for learner modeling in intelligent tutors from previous learner activities. In *Proceedings of the 24th Conference on User Modeling, Adaptation, and Personalization,* 55-63.

Floridi, L., Taddeo, M., & Turilli, M. (2009). Turing's imitation game: still an impossible challenge for all machines and some judges—an evaluation of the 2008 Loebner contest. *Minds and Machines, 19,* 1, 145-150.

Hilbert, D. (1900) Mathematical problems. Presentation to the International Conference of Mathematicians. Paris, France. Text retrieved 3/12/2019 from https://mathcs.clarku.edu/~djoyce/hilbert/problems.html

Hollands, F., & Bakir, I. (2015). Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods. *New York, NY: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.*

Kay, J. (2012). AI and education: grand challenges. *IEEE Intelligent Systems, 27,* 5, 66-69.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? In *Proceedings of the 9th International Conference on Educational Data Mining,* 94-101.

Liu, R., & Koedinger, K. R. (2015). Variations in learning rate: Student classification based on systematic residual error patterns across practice opportunities. In *Proceedings of the 8th International Conference on Education Data Mining,* 420–423.

Martinez-Maldonado, R., Kay, J., Yacef, K., & Schwendimann, B. (2012). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In *International Conference on Intelligent Tutoring Systems*, Springer, Berlin, 482-492.

Milliron, M. D., Malcolm, L., & Kil, D. (2014). Insight and action analytics: Three case studies to consider. *Research & Practice in Assessment, 9,* 70-89.

Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T., Salvi, A. van Velsen, M., Aghababyan, A., Martin, T. (2015). HART: The human affect recording tool. In *Proceedings of the ACM Special Interest Group on the Design of Communication (SIGDOC),* 24:1-24:6.

Paquette, L., Baker, R.S., de Carvalho, A., Ocumpaugh, J. (2015) Cross-system transfer of machine learned and knowledge engineered models of gaming the system. In *Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization*, 183-194.

Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics, 1,* 1, 107-128.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis: A new alternative to knowledge tracing. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, 531-538.

Pijeira-Díaz, H. J., Drachsler, H., Järvelä, S., & Kirschner, P. A. (2019). Sympathetic arousal commonalities and arousal contagion during collaborative learning: How attuned are triad members? *Computers in Human Behavior, 92,* 188-197.

Sao Pedro, M. A., Baker, R. S., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, *23*, 1, 1-39.

Sosnovsky, S., Dolog, P., Henze, N., Brusilovsky, P., & Nejdl, W. (2007) Translation of overlay models of student knowledge for relative domains based on domain ontology mapping. In R. Luckin, K. R. Koedinger and J. Greer (eds.) *Proceedings of 13th International Conference on Artificial Intelligent in Education*, IOS, 289-296.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognizing and responding to student affect. *International Journal of Learning Technology*, *4*. 3-4, 129-164.

Woolf, B. P., Lane, H. C., Chaudhri, V. K., & Kolodner, J. L. (2013). AI grand challenges for education. *AI Magazine*, *34*, 4, 66-85.

Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on the World Wide Web*, 765-774.

Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, *19*, 1, 27-39.