Embracing Imperfection in Learning Analytics

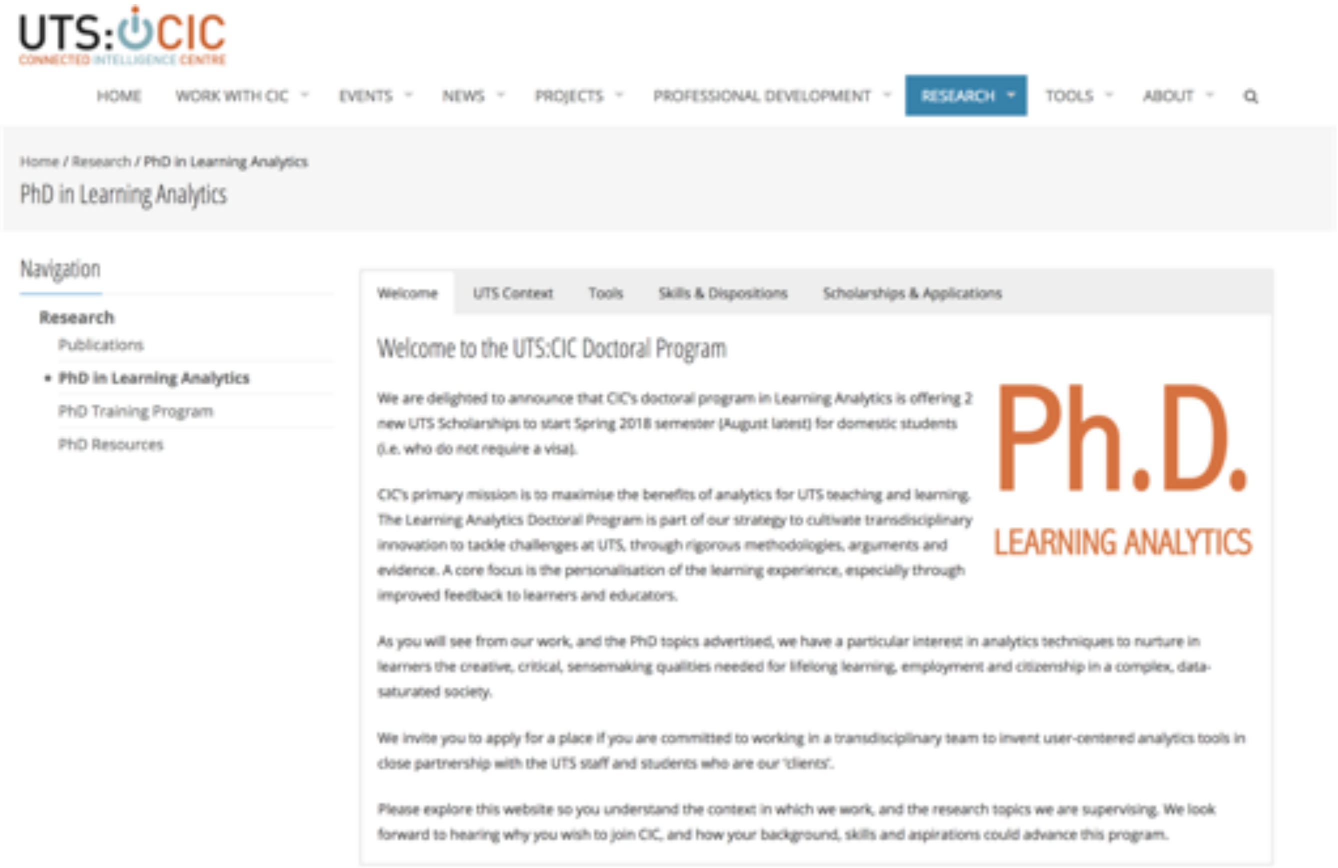
Kirsty Kitto, Simon Buckingham Shum & Andrew Gibson
UTS Connected Intelligence Centre
@KirstyKitto • kirsty.kitto@uts.edu.au
@sbuckshum • @AndrewResearch

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

UTS:CIC CONNECTED INTELLIGENCE CENTRE

https://utscic.edu.au

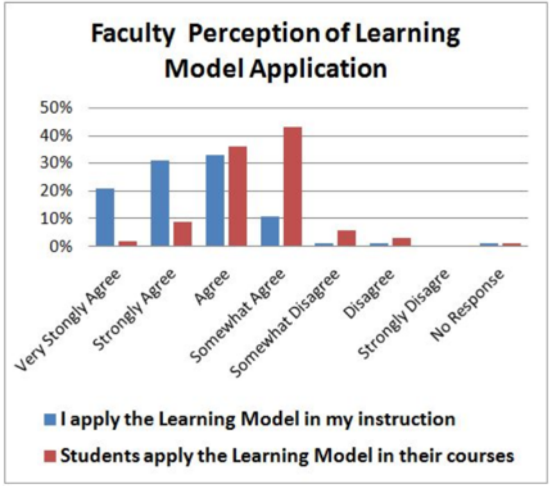
we are recruiting students!



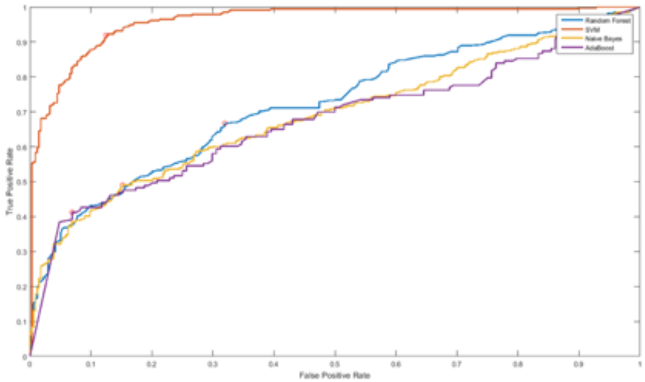
https://utscic.edu.au/research/phd/

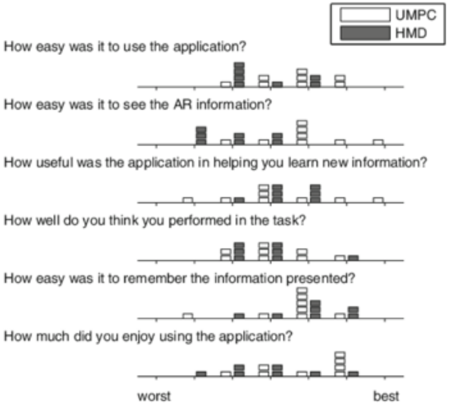
Photo by Blake Newman from Pexels https://www.pexels.com/photo/matador-786645/

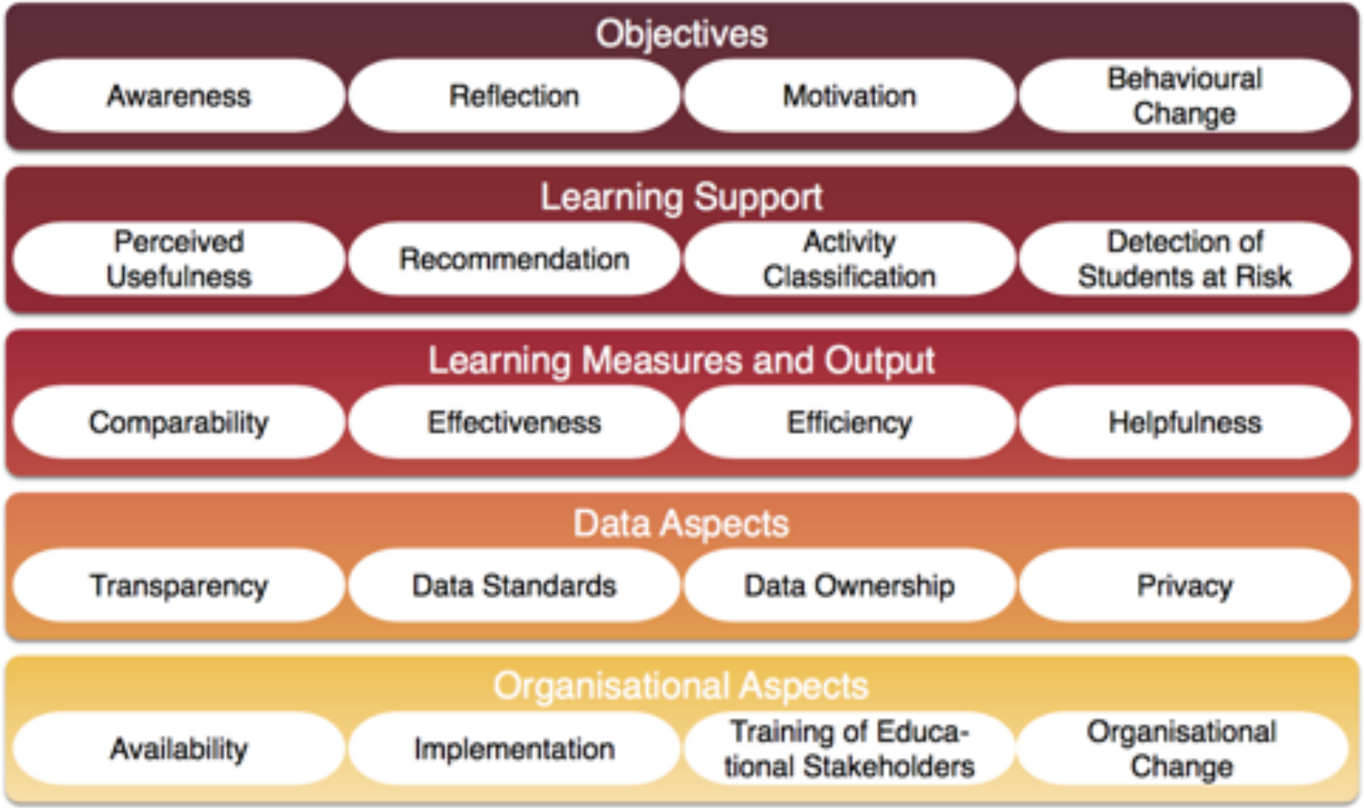
how do you evaluate LA?



how do you evaluate LA?

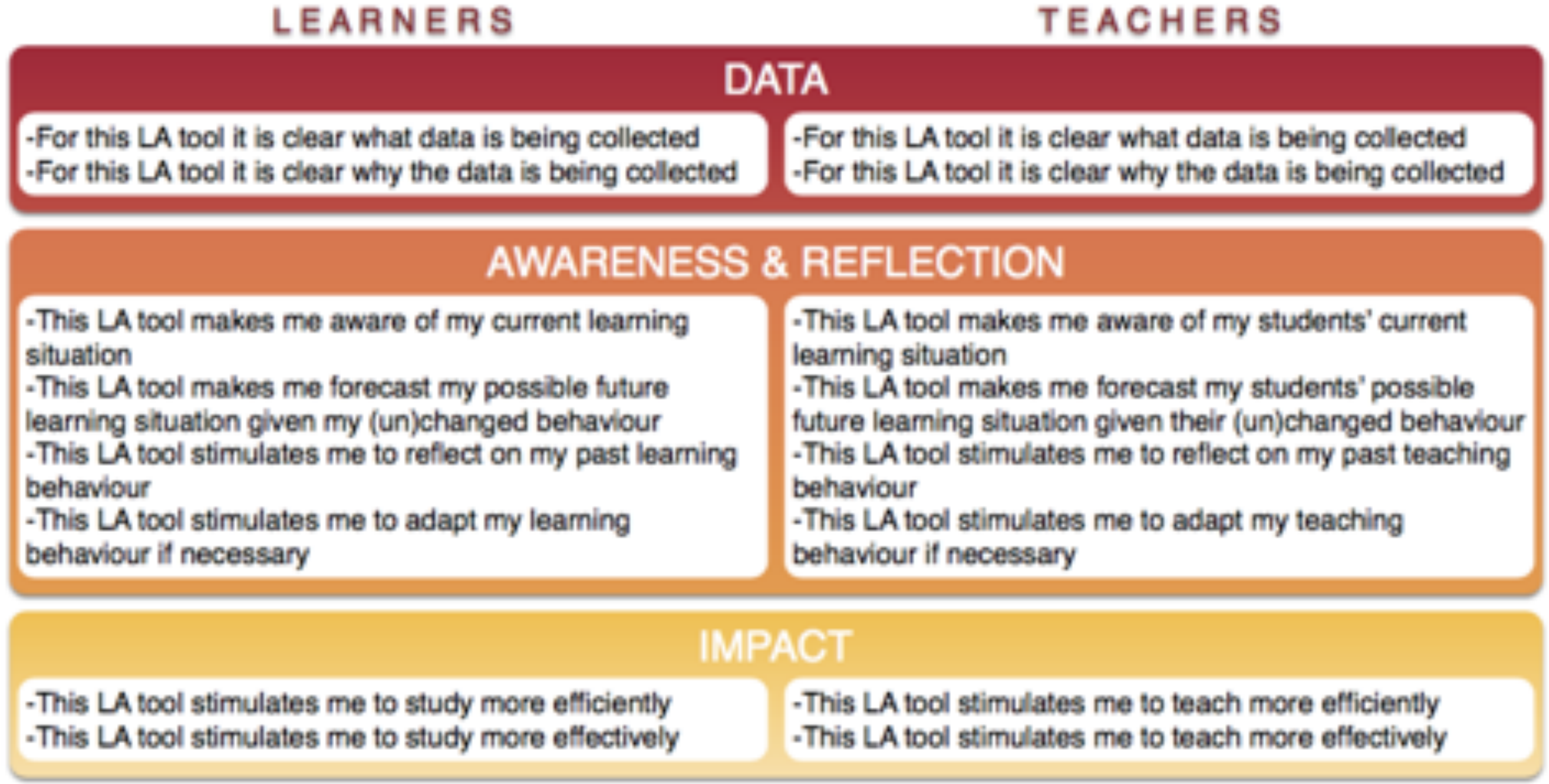






Scheffel, M., Drachsler, H., Stoyanov, S., Specht, M.: Quality Indicators for Learning Analytics. Educational Technology & Society 17(4), 117–132 (2014)

Scheffel, M. (2017). *The Evaluation Framework for Learning Analytics.*



Scheffel, M., Drachsler, H., Stoyanov, S., Specht, M.: Quality Indicators for Learning Analytics. Educational Technology & Society 17(4), 117–132 (2014)

Scheffel, M. (2017). *The Evaluation Framework for Learning Analytics.*

does LA help learning?

While EDM aims to improve learning outcomes, its

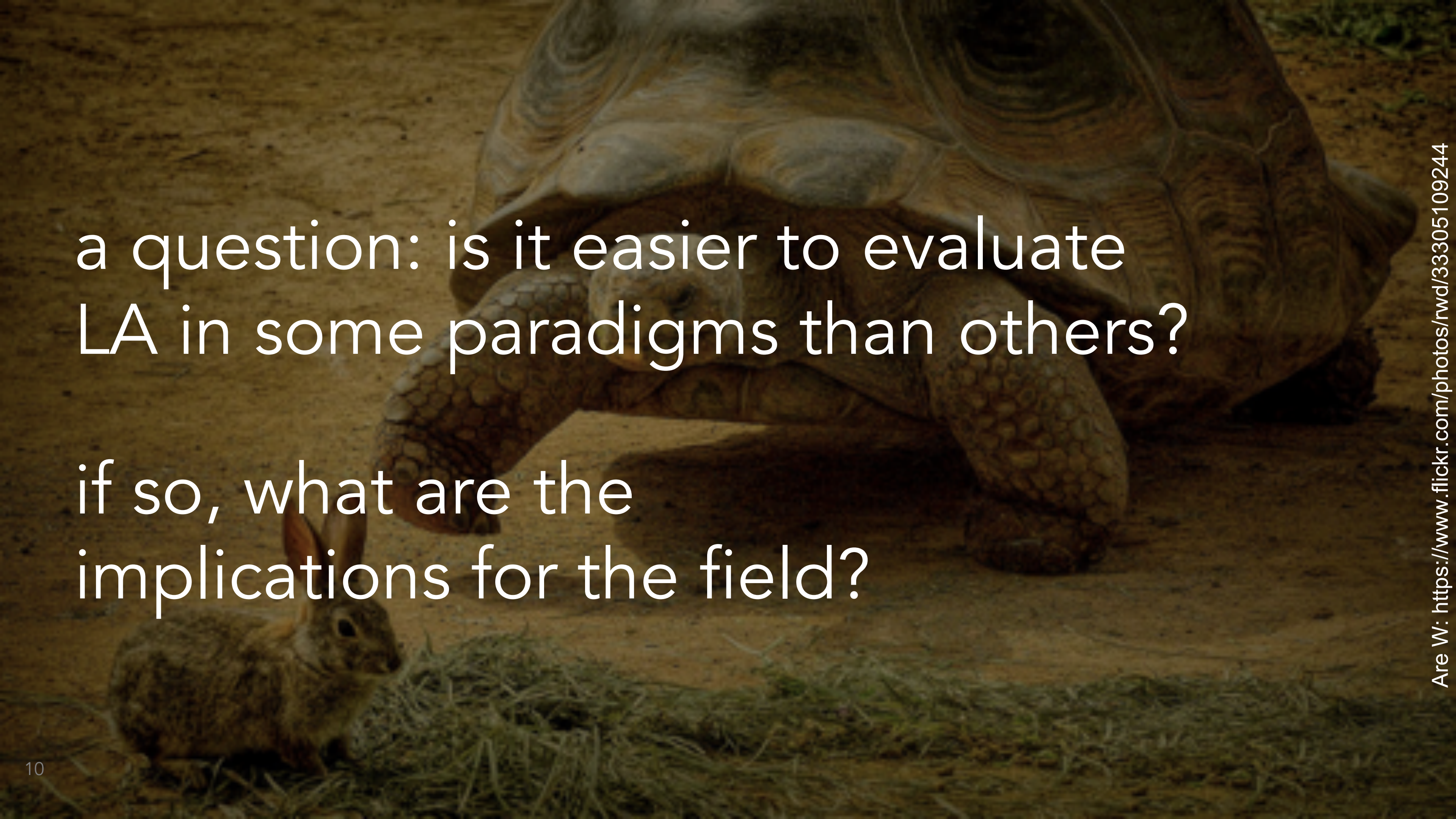
“emphasis on the ‘educational’ aspect of educational data mining has been scarce. . . One reason for this is the inclination of researchers to evaluate EDM research primarily for model fits and predictive accuracy rather than for plausibility, interpretability, and generalizable insights.”

does LA help learning?

While EDM aims to improve learning outcomes, its

“emphasis on the ‘educational’ aspect of educational data mining has been scarce. . . One reason for this is the inclination of researchers to evaluate EDM research primarily for model fits and predictive accuracy rather than for plausibility, interpretability, and generalizable insights.”

Ran Liu and Kenneth R Koedinger. 2017. Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. JEDM-Journal of Educational Data Mining 9, 1 (2017), 25–41.

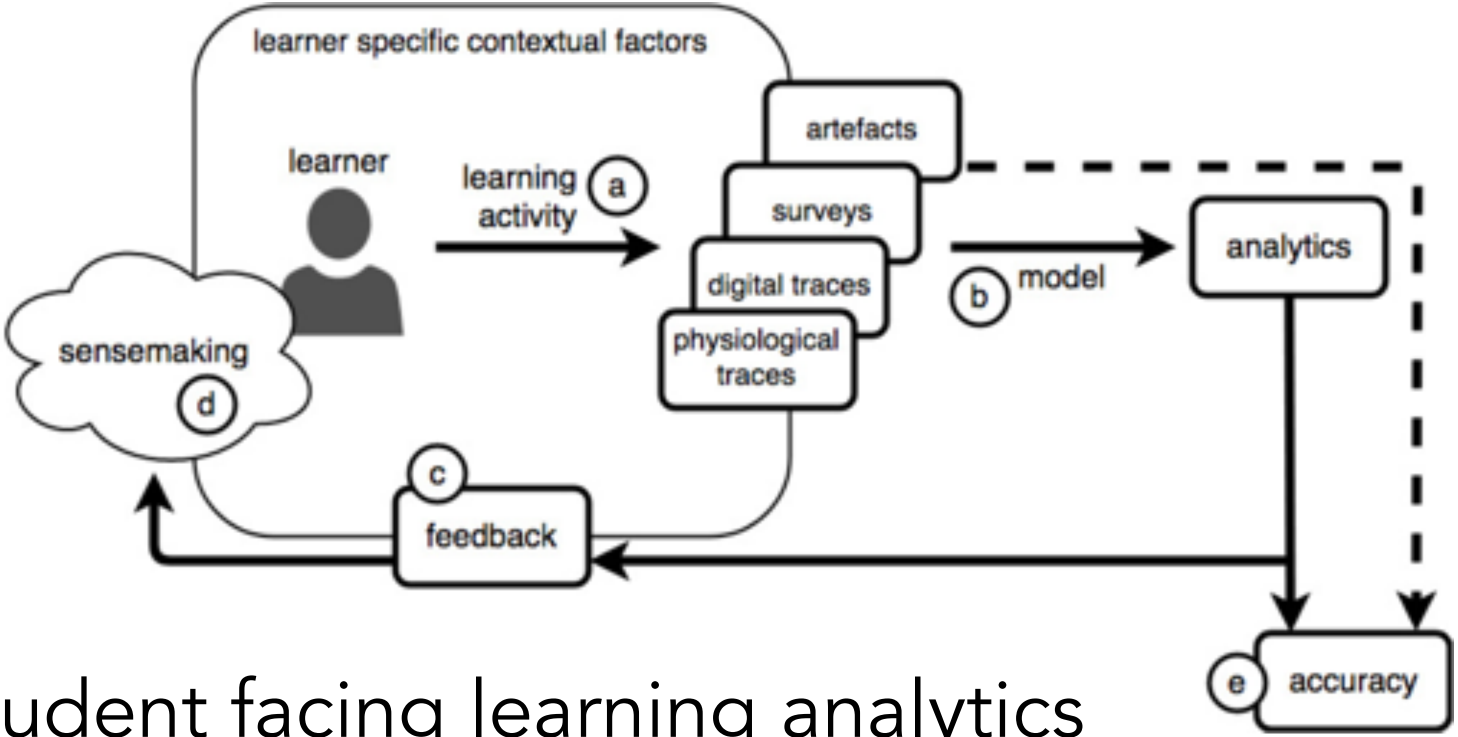


a question: is it easier to evaluate LA in some paradigms than others?

if so, what are the implications for the field?

Are W: https://www.flickr.com/photos/rwd/33305109244

but what are we evaluating here?



student facing learning analytics

but what *type* of student facing LA are we talking about?

Are students acquiring:

content and skills?

or

learning to learn?

should b for ple reason a wrong

We shold not make nanobots fore multipul reesuns. As you probibly know in the rong hands thay can be dangerus. So to fined out the rest you are going to have to reed the rest of this exsithing artikul.

cle ou ea ea ang bon

For one a nanobot could have a bug and start eeting enything cardin basked or just not work at all. Another thing is that thay may allso eat the rong substins, wich wold onle be bade in some cases. Wat is rile bad if one has a bug it cold make mor with the same problem. Now I know that you are wondering wat I am tolking abot, I meen how could it make mor of its problem inles it colud rerite uther nanobots programs. Well some sientintists are tring to figyer out how to mak it posibul for them to copy themselfs. So one might be able to bekum 100.

kes ss

Also thay are planing to make them abule to cile bakterya, and there thay might eat away at the intestens insted. But don't be werryd thay mite make it so that thay will go throw the body with the rest of th food. Also thay might program them to tern of after a serten amout of time.

Thay are also planing to make smal traking divises so kids wont get lost. I just hope thay are haker safe and thay aren't over used. I don't want the goverment to know to much. I also don't want some sikeco thraking me.

So as you can see there are lots of problems. There is bugs, hakers, goverment overyuos, and faling into the rong hands. There is good noos I think we are stile alitaule fare frome geting a lot of nanobots just yet.

news

Should the distributed intelligence of the whole system's performance (humans + technology) be the output measure?

Or, should we also be concerned with the effects on human performance when stripped of the technology?

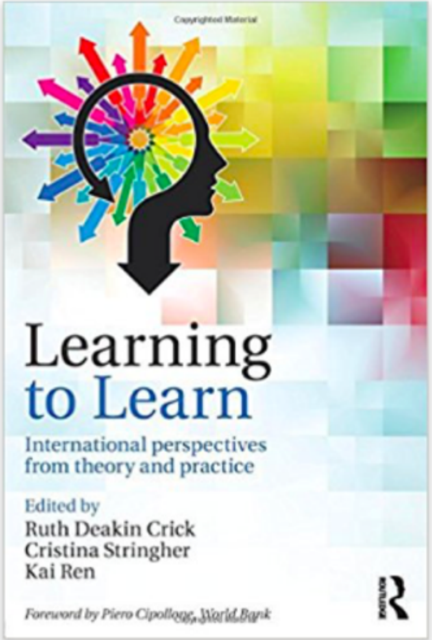
Gavriel Salomon, David N Perkins, and Tamar Globerson. 1991. Partners in cognition: Extending human intelligence with intelligent technologies. Educational researcher 20, 3 (1991), 2–9.

learning to learn

"equipping students with knowledge, skills, and dispositions that prepare them for lifelong learning, in a complex and uncertain world"

"Creativity, critical thinking, agency, curiosity, and an ability to tolerate uncertainty…"

arguably the purpose of analytics-powered pedagogy in such contexts is to provoke productive reflection on one's strengths and weaknesses — these are higher order competencies, into which a machine can have limited insight



Buckingham Shum, S. and Deakin Crick, R. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics, 3,* (2), 6–21.

authentic learning: vital but challenging for LA

wicked problems: how do we provide LA when there is no correct answer?

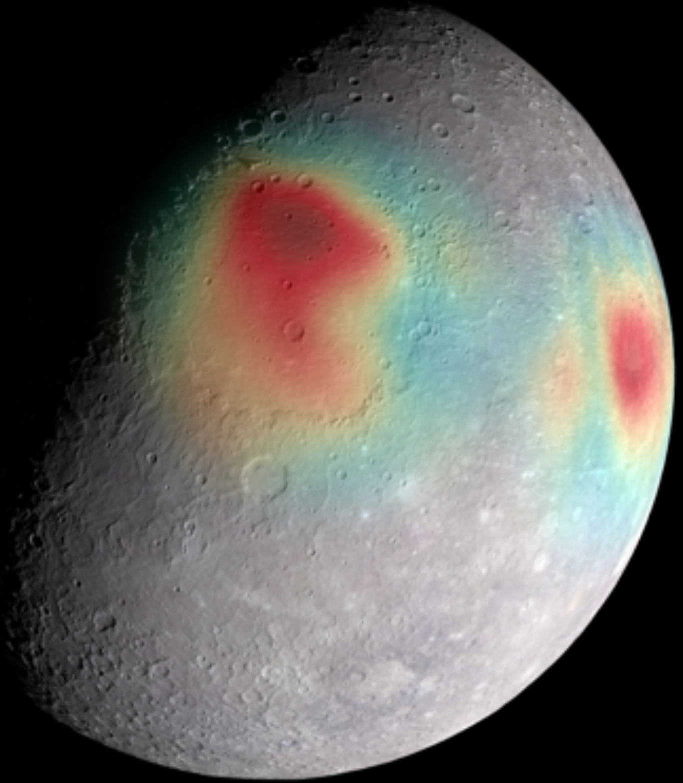
transformed perspective: the sense that a learner makes of their experience, or a shift in worldview, which by definition is not accessible to the machine, but to which a machine might have partial access

socially and psychologically complex performance: scenarios where the outcome is emergent in nature, a function of many drivers that result in unpredictable and/or unique outcomes, often because social interaction is central to the process

analytics in such contexts will in principle have a high degree of imperfection!

Act 1:

perfection is not possible



a cautionary tale from information retrieval

Turpin, Scholer (2006). User performance versus precision measures for simple search tasks. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 11-18). ACM.

User Performance versus Precision Measures for Simple Search Tasks

Andrew Turpin* Falk Scholer

School of Computer Science & IT
RMIT University, GPO Box 2476V
Melbourne, Australia, 3001.
{aht,fscholer}@cs.rmit.edu.au

ABSTRACT

Several recent studies have demonstrated that the type of improvements in information retrieval system effectiveness reported in forums such as SIGIR and TREC do not translate into a benefit for users. Two of the studies used an instance recall task, and a third used a question answering task, so perhaps it is unsurprising that the precision based measures of IR system effectiveness on one-shot query evaluation do not correlate with user performance on these tasks. In this study, we evaluate two different information retrieval tasks on TREC Web-track data: a precision-based user task, measured by the length of time that users need to find a single document that is relevant to a TREC topic; and, a simple recall-based task, represented by the total number of relevant documents that users can identify within five minutes. Users employ search engines with controlled mean average precision (MAP) of between 55% and 95%. Our results show that there is no significant relationship between system effectiveness measured by MAP and the precision-based task. A significant, but weak relationship is present for the precision at one document returned metric. A weak relationship is present between MAP and the simple recall-based task.

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Performance, Design, Experimentation, Human Factors

*Supported in part by a Discovery Project Grant from the Australian Research Council.

Keywords

Search engines, information retrieval evaluation, user study

1. INTRODUCTION

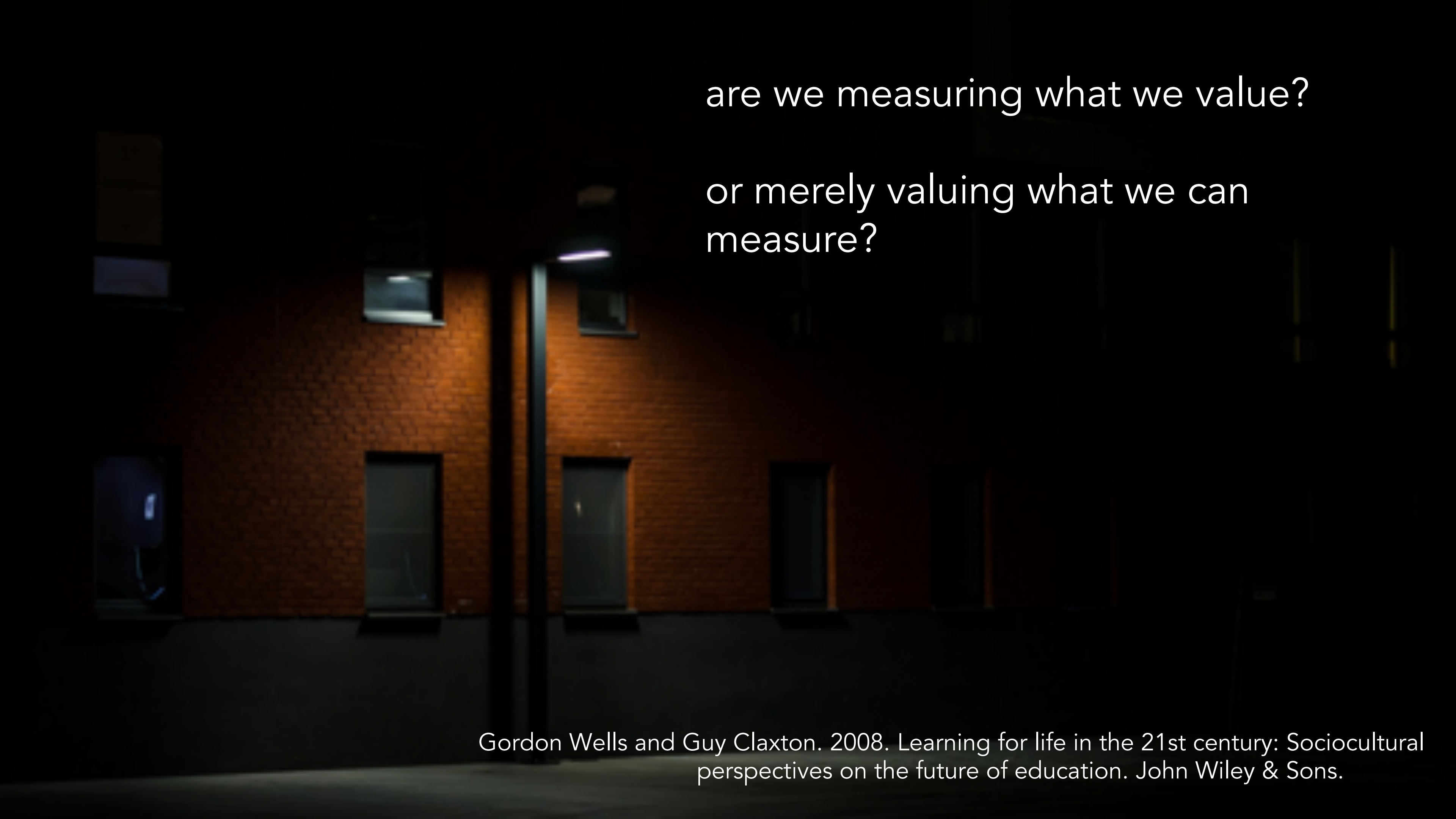
The field of information retrieval has a well-established tradition of experimental evaluation, dating back to Cleverdon's "Cranfield" experiments [7], and continuing through the ongoing series of Text REtrieval Conferences (TREC)[1]. The general approach for evaluating adhoc retrieval, where a static collection is searched for documents that are relevant to previously unknown topic, requires: a collection of documents that is to be searched; a set of queries that represent user information needs and are run against the collection; and a set of *relevance judgements* that indicate, for each query, which documents satisfy the current information need and which do not. Evaluations are typically run as a *batch process*, where the retrieval system fetches a pre-specified number of answer documents for each query, with no user interaction. Performance is quantified using a variety of metrics derived from the number of relevant answers that have been found. Commonly reported measures include mean average precision (MAP), precision at 10 documents retrieved (P@10), and bpref (these metrics are defined in Section 2). Indeed, much IR research focuses on demonstrating improvements in these metrics.

However, recent studies have demonstrated that improvements in these metrics do not translate into a direct benefit for users. A study by Hersh et al. [13] shows that instance recall – where users try to identify different aspects of a question within a limited timeframe – does not improve with small increases in mean average precision of the underlying search system on the scale that is commonly reported in IR results. Allan et al. [1] confirm this result (using bpref), but also show that for larger, specific increases in bpref, users do benefit on the instance recall task. Turpin and Hersh [17] demonstrate a lack of improvement when users are engaged in a question answering task for a small number of questions.

A possible reason for the lack of correlation between underlying system effectiveness and user performance could be the nature of the search tasks that have been examined. Instance recall – as its name implies – is inherently recall-oriented [1, 13]. However, mean average precision, while including a recall component, evaluates systems predom-

[1]http://trec.nist.gov

11

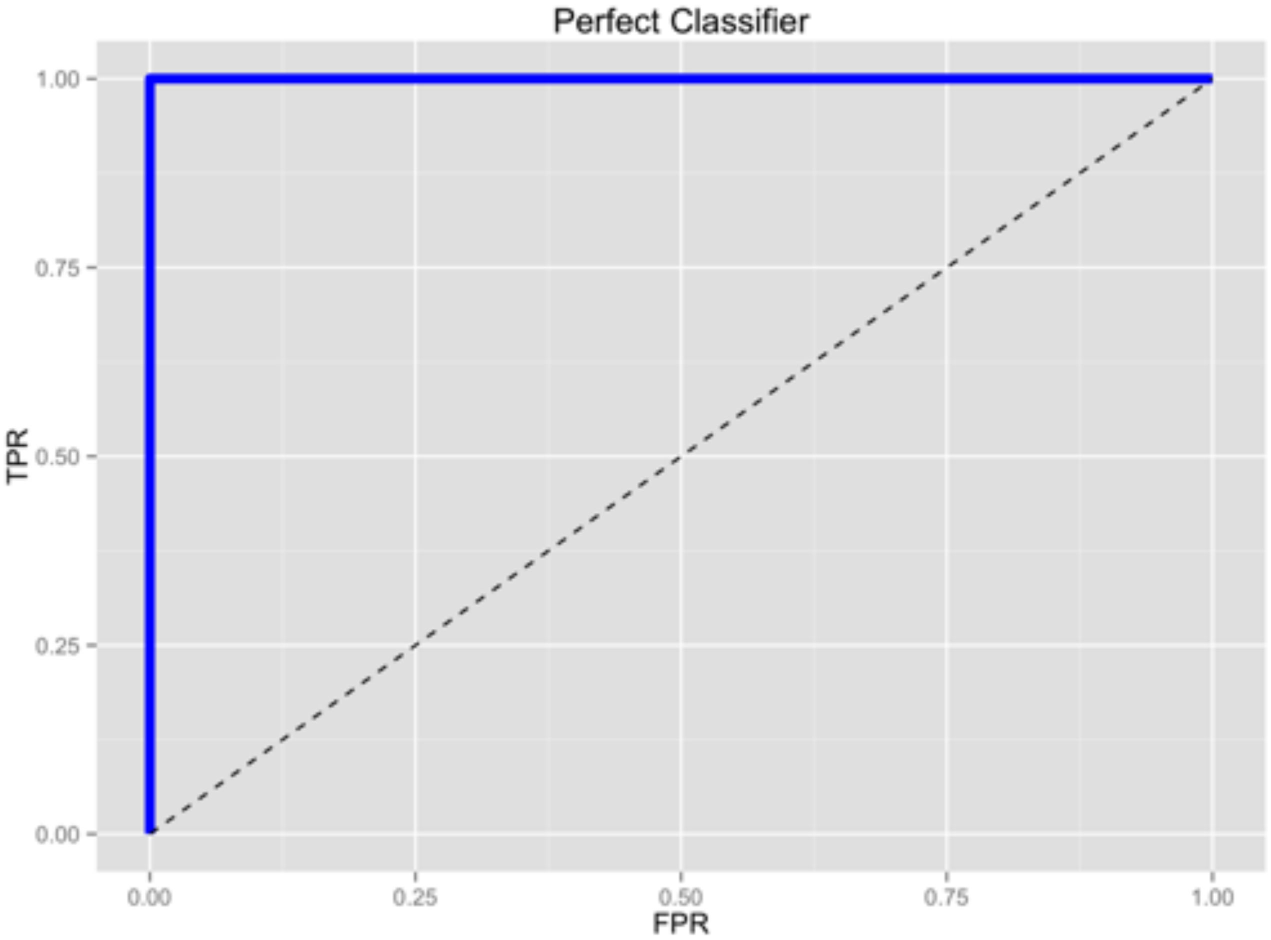


are we measuring what we value?

or merely valuing what we can measure?

Gordon Wells and Guy Claxton. 2008. Learning for life in the 21st century: Sociocultural perspectives on the future of education. John Wiley & Sons.

a thought experiment: is a perfect classifier desirable in education?



cognitive presence

"extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication."

Triggering

↓

Exploration

↓

Integration

↓

Resolution

Garrison, Anderson, Archer (2001) Critical thinking, cognitive presence, and computer conferencing in distance education. American journal of distance education, 15(1):7–23

https://plus.google.com/u/0/+StefanPSchmid/posts/4wrUbFzFwpJ we can use machine learning to classify discussion forum text using this construct


Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case

Vitomir Kovanović
School of Informatics
The University of Edinburgh
Edinburgh, UK
v.kovanovic@ed.ac.uk

Srećko Joksimović
Moray House School of Education
The University of Edinburgh
Edinburgh, UK
s.joksimovic@ed.ac.uk

Zak Waters
Queensland University of Technology
Brisbane, Australia
z.waters@qut.edu.au

Dragan Gašević
Moray House School of Education and School of Informatics
The University of Edinburgh
Edinburgh, UK
dgasevic@acm.org

Kirsty Kitto
Queensland University of Technology
Brisbane, Australia
kirsty.kitto@qut.edu.au

Marek Hatala
School of Interactive Arts and Technology
Simon Fraser University
Burnaby, Canada
mhatala@sfu.ca

George Siemens
LINK Research Lab
University of Texas at Arlington
Arlington, USA
gsiemens@uta.edu

ABSTRACT

Keywords

1. INTRODUCTION

Kovanović, Joksimović, Waters, Gašević, Kitto, Hatala, Siemens (2016). Towards automated content analysis of discussion transcripts: a cognitive presence case. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16). ACM, New York, NY, USA, 15-24.

should we use it with students yet?

how accurate does it have to be?

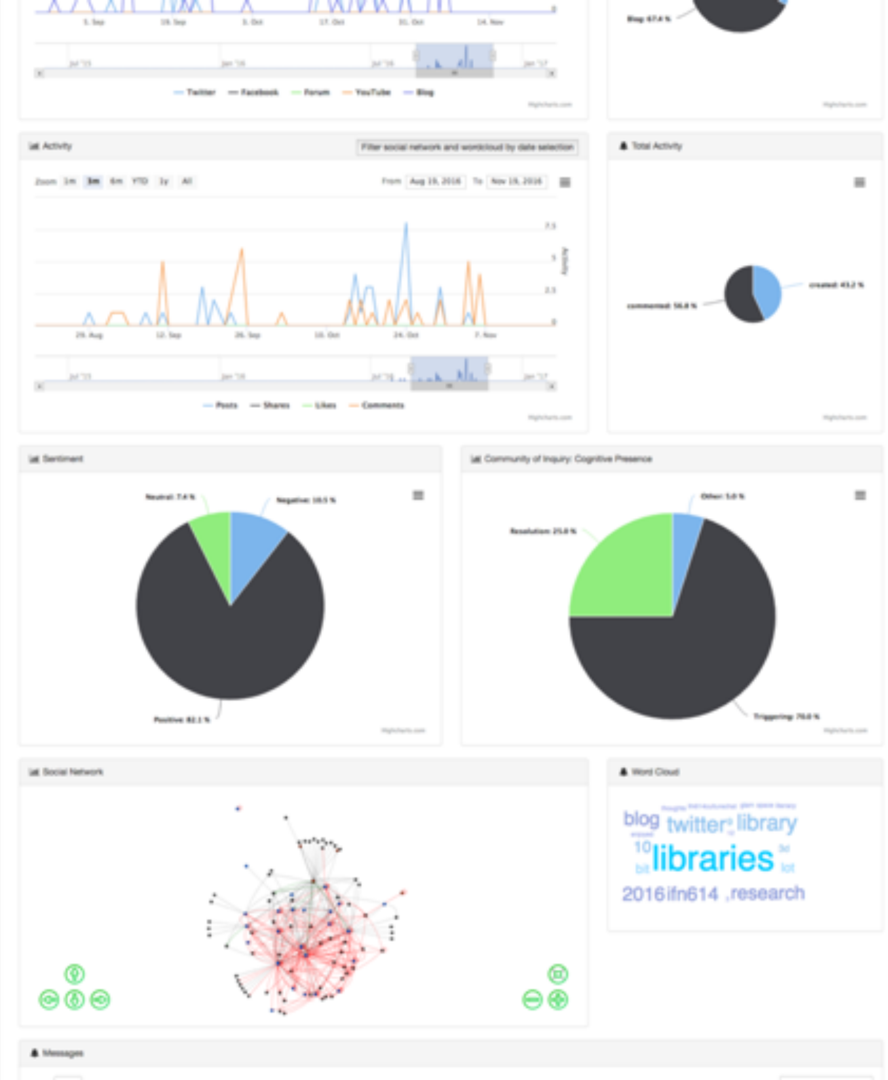
data was unbalanced (solved using boosting)

- is it overfitted for one “type” of learning scenario?
- how accurate will it be if used in another context?
- how different does a situation have to be before we retrain?

how are we going to use it?

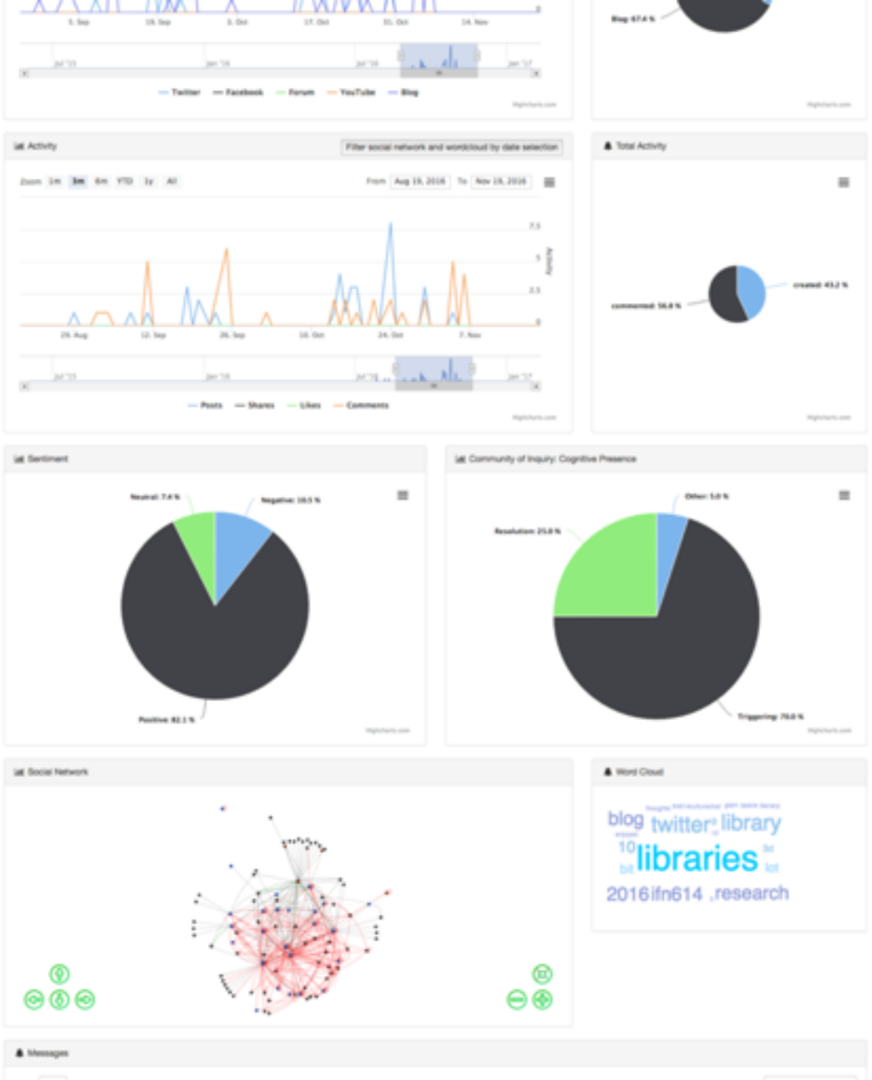
- who sees the classifications?
- what happens if the classifier is wrong?

well we have already… and it wasn't even the state of the art classifier…



well we have already… and it wasn't even the state of the art classifier…

should we worry?



A very strong reflection from most recent trial?

In Week 2 I was very aspirational about the role I wanted to play; 'I would like my profile to be professional, respectful, organised, connected and visible. I aim to be an active participant within "reflection and critical discourse that is the core dynamic of a community of inquiry". I achieved my aim of being an active participant as I made over 75 comments on my peers' posts, averaging over 5 per week. **However I feel I did not participate fully in all 4 phases of the cognitive presence in the Practical [sic] Inquiry Model; triggering event, exploration, integration and resolution – despite having sentence openers taped next to my computer!** Triggering events and some exploration were met by sharing an interesting article relevant to a post I had read and also asking some questions, but I felt a lot of my posts were agreeing with and complimenting upon the erudite musings of my peers. I was definitely wary of confronting differing ideas and promoting a critical discourse. **This participation in all cognitive phases needs improving** so the sentence openers will remain up! [score=4]

Kirsty Kitto, Mandy Lupton, Kate Davis, and Zak Waters. 2017. Designing for student-facing learning analytics. *Australian Journal of Educational Technology*, 33, 5 (2017), 152–168.

Act 2:

perfection is not desirable

the Navajo rug

In a Navajo rug there is always an imperfection woven into the corner. And interestingly enough, it's where “the Spirit moves in and out of the rug.” The pattern is perfect and then there's one part of it that clearly looks like a mistake …

Perfection is not the elimination of imperfection. That's our Western either/or, need-to-control thinking. Perfection, rather, is the ability to incorporate imperfection!

Breathing Under Water: Spirituality and the 12 Steps,
by Richard Rohr

the Navajo rug

In a Navajo rug there is always an imperfection woven into the corner. And interestingly enough, it's where "the Spirit moves in and out of the rug." The pattern is perfect and then there's one part of it that clearly looks like a mistake …

Perfection is not the elimination of imperfection. That's our Western either/or, need-to-control thinking. Perfection, rather, is the ability to incorporate imperfection!

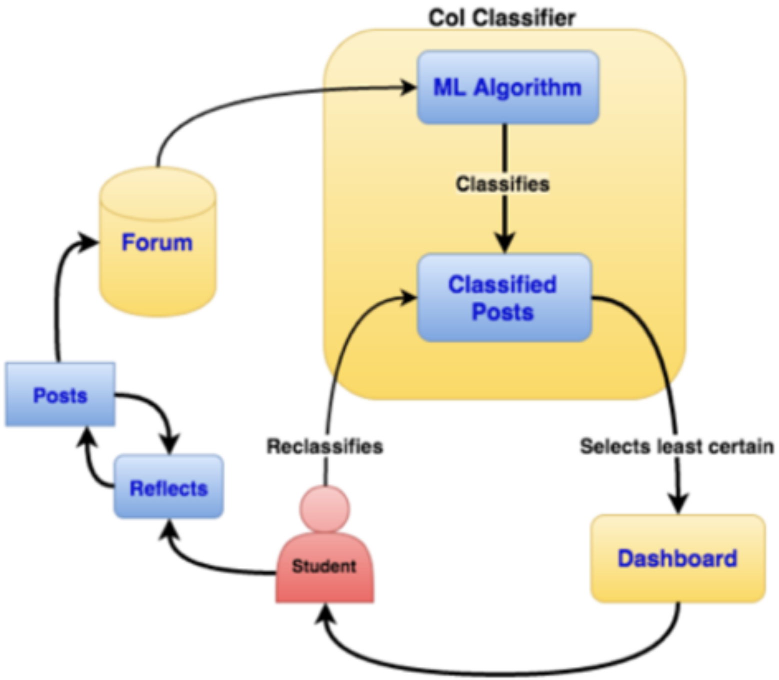
Breathing Under Water: Spirituality and the 12 Steps,
by Richard Rohr

http://exhibitions.kelsey.lsa.umich.edu/less-than-perfect/navajo.php

active learning squared (AL^2)

the student trains the classifier…

…while it is training the student…



Kirsty Kitto, Mandy Lupton, Kate Davis, and Zak Waters. 2017. Designing for student-facing learning analytics. *Australian Journal of Educational Technology*, 33, 5 (2017), 152–168.

Connected Learning Analytics Toolkit

Community of Inquiry Classification

Community of Inquiry Classifications

Want to learn about your participation within your learning community?

When you start this activity, you will see one of your posts. We have used machine learning to catgorise your cognitive presence according the Community of Inquiry model.

However, our machine learning tool is still learning and it could be wrong. We would like you to:

1. Think about how your post was classified
2. Choose what category you believe your post belongs to
3. If you like, you may highlight text from your post that you used in making your decision, or add remarks to the text-box about what helped you come to your conclusion
4. You can view your history below

What is Cognitive Presence?

Cognitive presence has four phases: Triggering, Exploration, Integration, and Resolution.

Triggering Phase initiates discussion about a particular issue/topic for inquiry.
Exploration Phase posts explore the issue at hand by exchanging knowledge between members of the community.
Integration Phase interactions build upon the ideas shared and explored in the Exploration phase and begin to construct understanding or a solution about a topic or issue.
Resolution Phase are messages in a discussion that test the solutions or understanding developed in the Integration phase.

Begin

Connected Learning Analytics Toolkit

Community of Inquiry Classification

Community of Inquiry Classifications

Want to learn about your participation within your learning community?

When you start this activity, you will see one of your posts. We have used machine learning to catgorise your cognitive presence according the Community of Inquiry model.

However, our machine learning tool is still learning and it could be wrong. We would like you to:

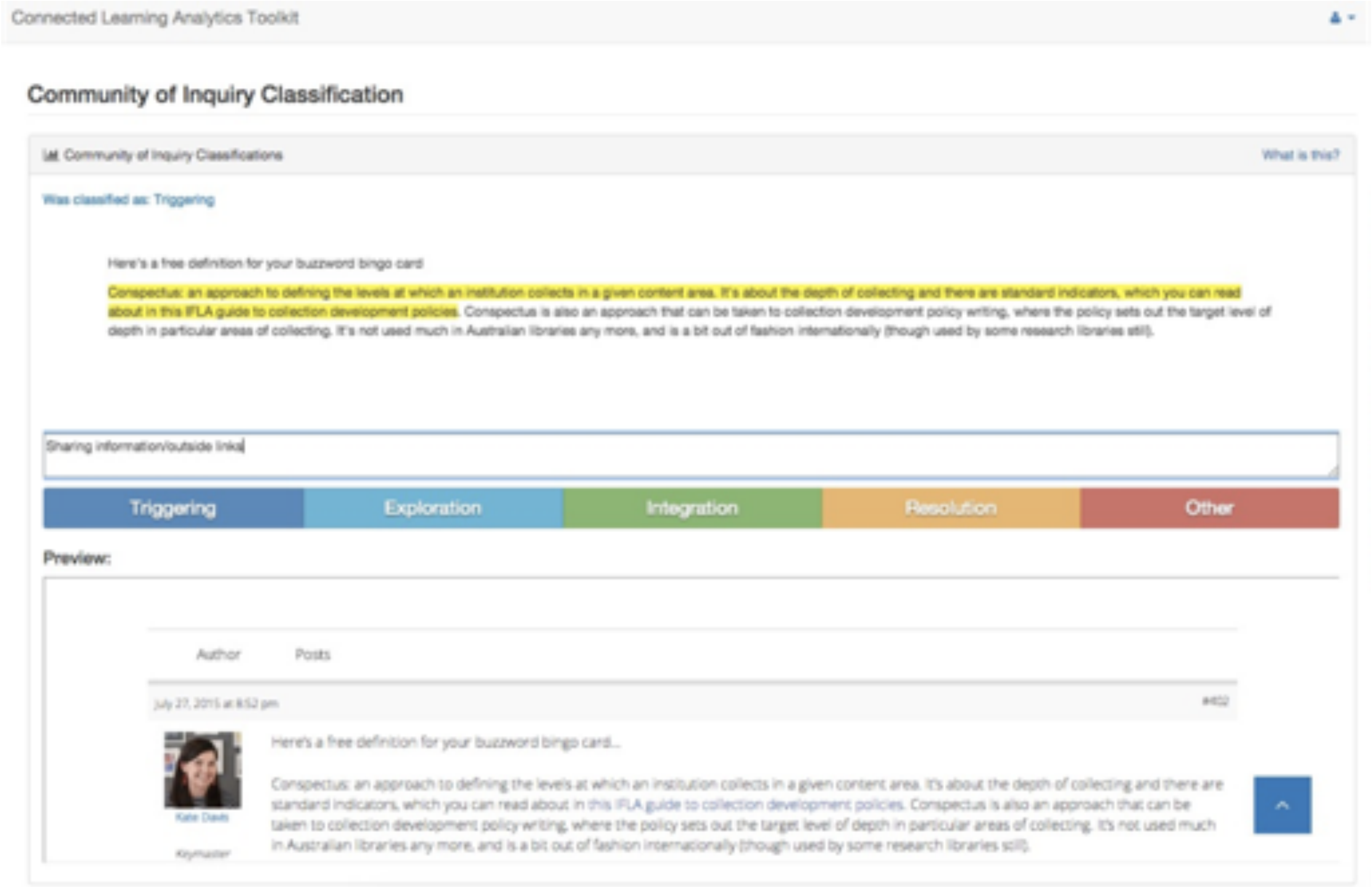
1. Think about how your post was classified
2. Choose what category you believe your post belongs to
3. If you like, you may highlight text from your post that you used in making your decision, or add remarks to the text-box about what helped you come to your conclusion
4. You can view your history below

What is Cognitive Presence?

Cognitive presence has four phases: Triggering, Exploration, Integration, and Resolution.

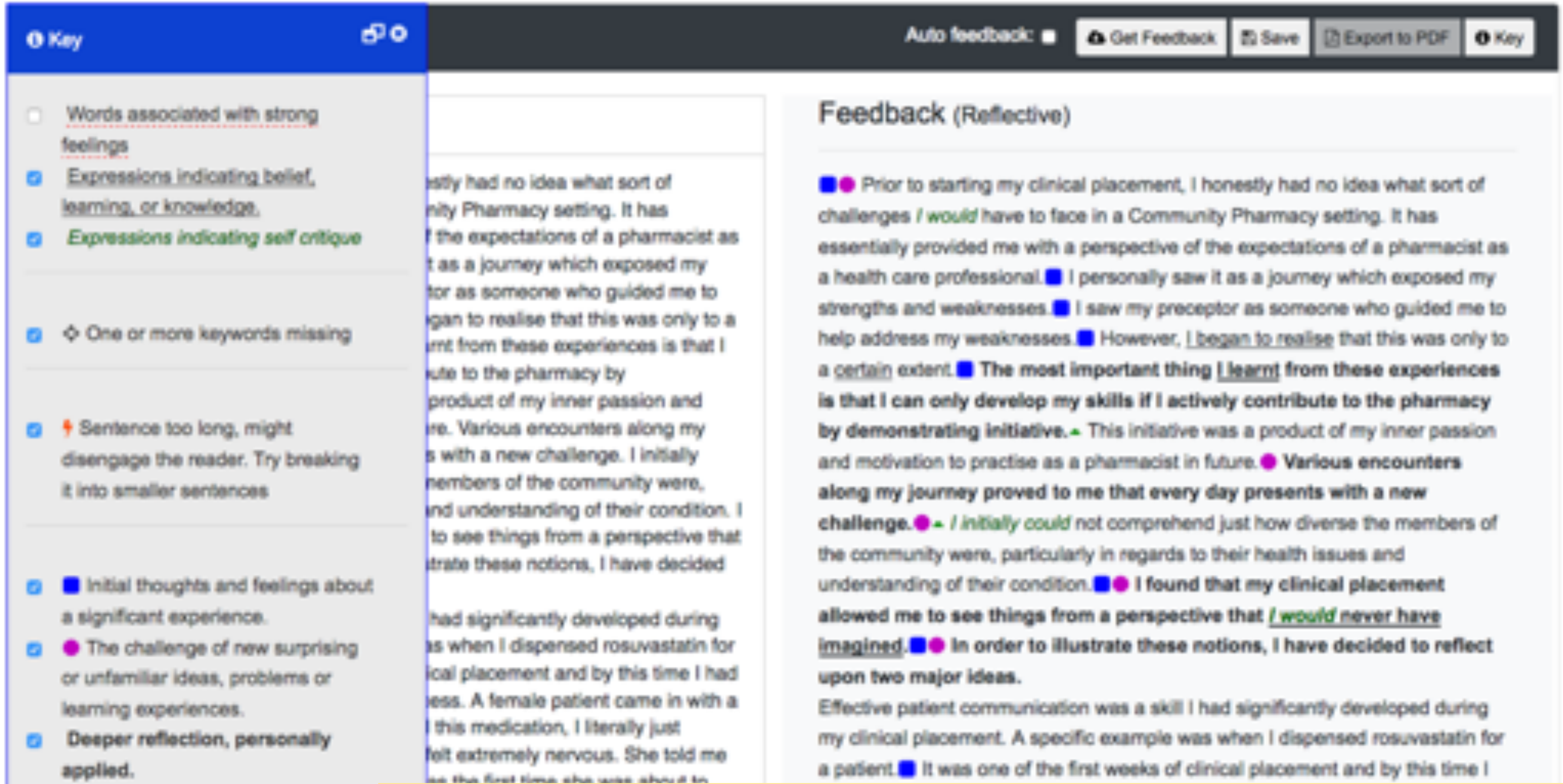
Triggering Phase initiates discussion about a particular issue/topic for inquiry.
Exploration Phase posts explore the issue at hand by exchanging knowledge between members of the community.
Integration Phase interactions build upon the ideas shared and explored in the Exploration phase and begin to construct understanding or a solution about a topic or issue.
Resolution Phase are messages in a discussion that test the solutions or understanding developed in the Integration phase.

Begin

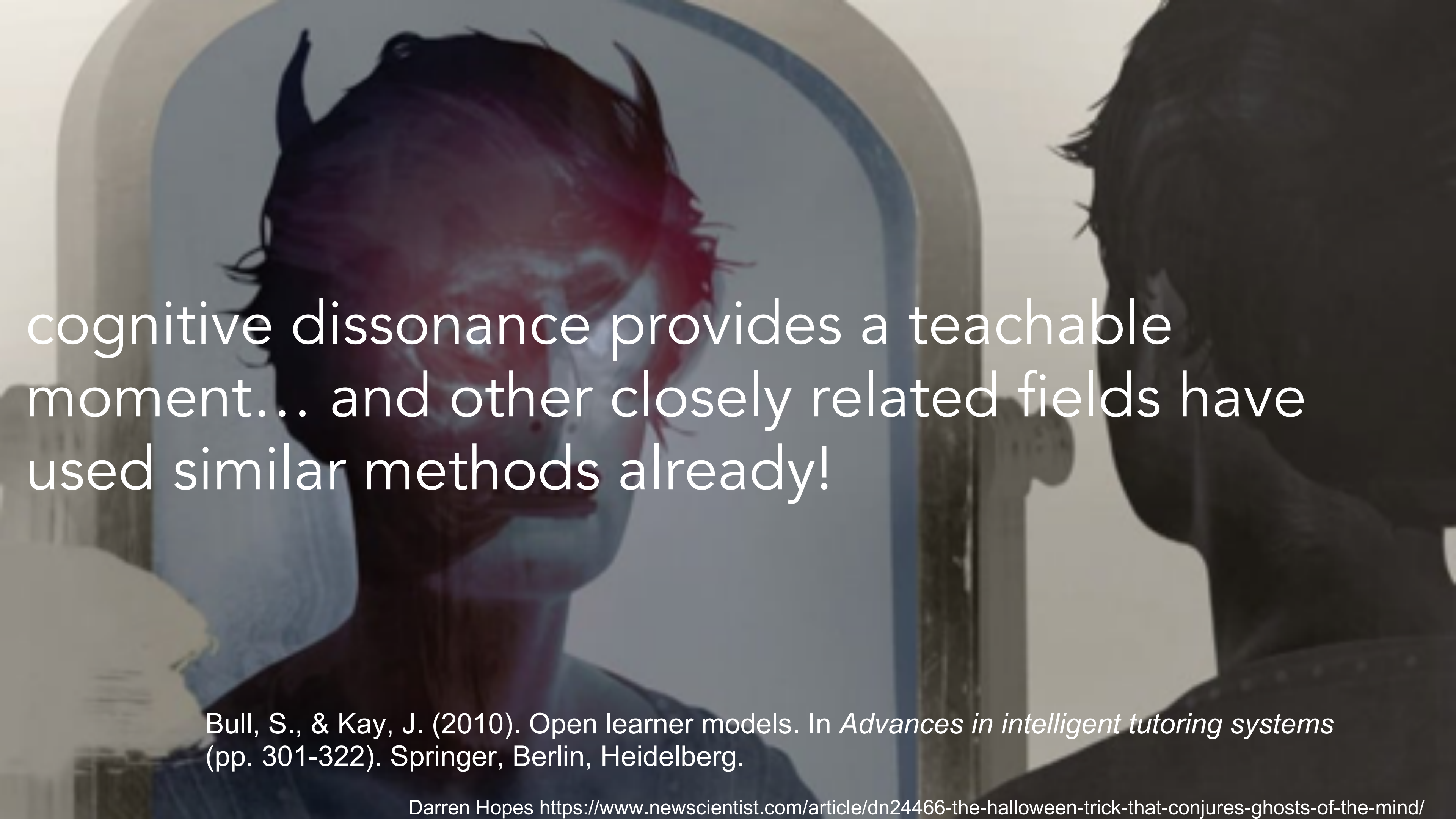

Connected Learning Analytics Toolkit
Community of Inquiry Classification
Community of Inquiry Classifications
What is this?
Was classified as: Triggering
Here's a free definition for your buzzword bingo card
Conspectus: an approach to defining the levels at which an institution collects in a given content area. It's about the depth of collecting and there are standard indicators, which you can read about in this IFLA guide to collection development policies. Conspectus is also an approach that can be taken to collection development policy writing, where the policy sets out the target level of depth in particular areas of collecting. It's not used much in Australian libraries any more, and is a bit out of fashion internationally (though used by some research libraries still).
Sharing information/outside links
Triggering
Exploration
Integration
Resolution
Other
Preview:
Author
Posts
July 27, 2015 at 8:52 pm
Here's a free definition for your buzzword bingo card...
Conspectus: an approach to defining the levels at which an institution collects in a given content area. It's about the depth of collecting and there are standard indicators, which you can read about in this IFLA guide to collection development policies. Conspectus is also an approach that can be taken to collection development policy writing, where the policy sets out the target level of depth in particular areas of collecting. It's not used much in Australian libraries any more, and is a bit out of fashion internationally (though used by some research libraries still).
Keymaster

see the paper for second example of imperfection:

automated formative feedback on reflective writing



LAK17 Best Paper / Academic Writing Analytics: https://utscic.edu.au/tools/awa



cognitive dissonance provides a teachable moment… and other closely related fields have used similar methods already!

Bull, S., & Kay, J. (2010). Open learner models. In *Advances in intelligent tutoring systems* (pp. 301-322). Springer, Berlin, Heidelberg.

Darren Hopes https://www.newscientist.com/article/dn24466-the-halloween-trick-that-conjures-ghosts-of-the-mind/

embracing imperfection

so imperfection in our LA tools opens up new opportunities

- teachable moments
- intelligence augmentation
- mindful engagement with automated feedback
- learning to challenge computational decisions
- accelerates presence of more advanced LA in education

but to get to this point we need to ensure that mature LA tools are evaluated holistically!

as machine intelligence reduces, we can increase human agency (and learning) through good LD

"nonautomatic, effortful and thus metacognitively guided processes"

Gavriel Salomon, David N Perkins, and Tamar Globerson. 1991. Partners in cognition: Extending human intelligence with intelligent technologies. Educational researcher 20, 3 (1991), 2–9.

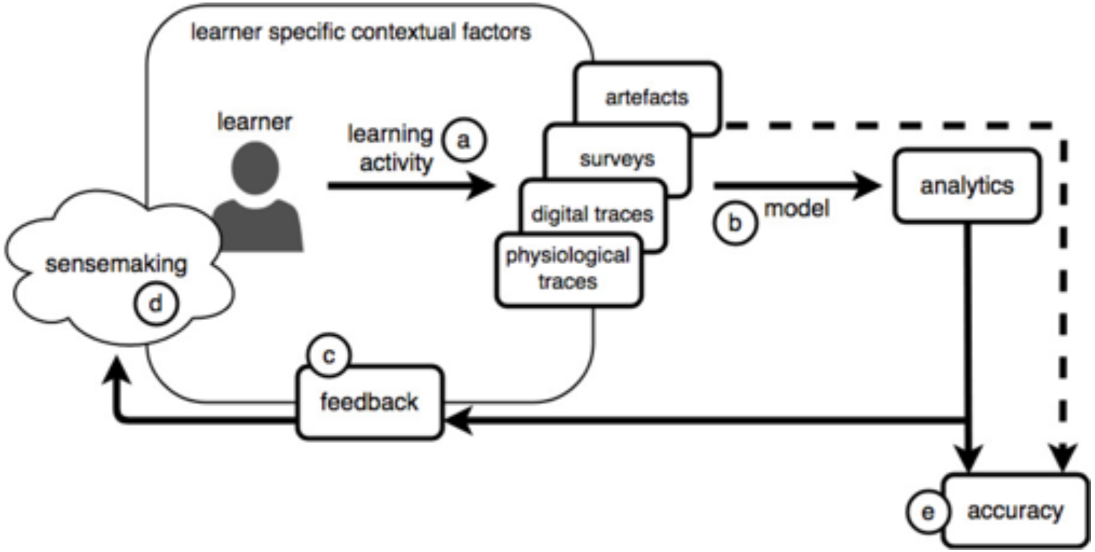
but then how can we evaluate success?

towards comprehensive evaluation for LA

mature student facing LA (that aims to help students learn how to learn) needs to be evaluated across a range of criteria

in the paper we explore:

1. Learning design
2. Model
3. Feedback
4. Sensemaking/gain
5. Accuracy



applying this to AL^2

Learning design: this learning design aims to teach (i) data literacy (i.e. that ML can be wrong) and (ii) a basic educational construct

Model: dual process model of cognition

Feedback: automatic classifications are appended to student comments and presented in a new display

Sensemaking/gain: The interface allows the student to (i) change the classification of their post, (ii), highlight components of the post that they feel are indicative of the classification they have chosen, (iii) leave a comment about why they chose that classification.

Accuracy: to date - very low in pilot trials (30.2%!)

conclusions

- perfect accuracy in LA is unlikely to be possible in a wide range of authentic learning scenarios…
- … nor is it always desirable — embracing imperfection opens up new possibilities for teachable moments!
- imperfection is sometimes a feature not a bug

Thanks!

Debate!

@kirstykitto
@sbuckshum
@andrewresearch

Open Access reprint:
http://bit.ly/NavajoRugLAK18