

# Your model is predictive but is it useful? – Theoretical and Empirical Considerations of a New Paradigm for Adaptive Tutoring Evaluation

José P. González-Brenes, Pearson



PEARSON

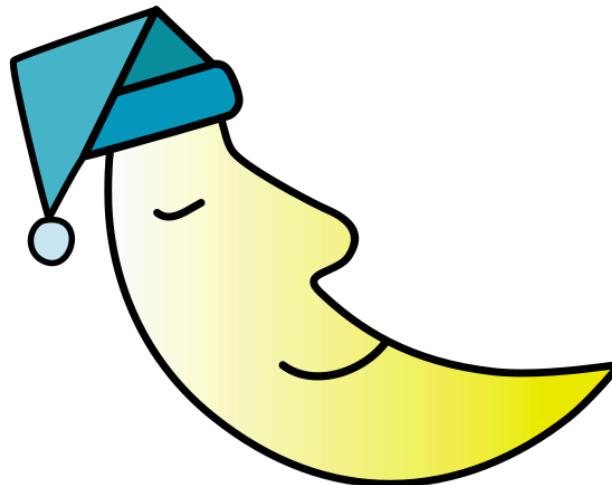
Yun Huang, University of Pittsburgh

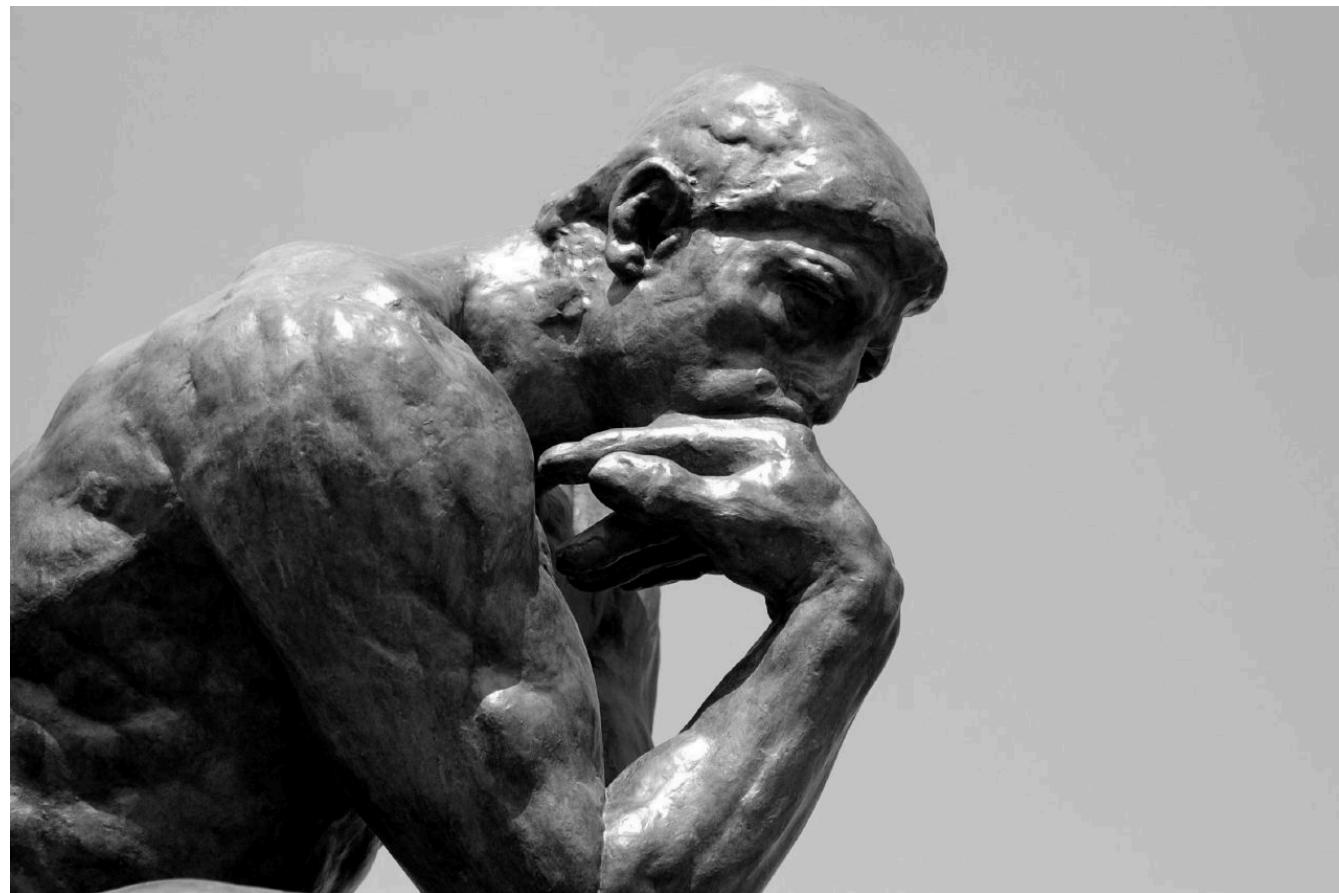


Main point of the paper:

We are not evaluating student models correctly

New paradigm, Leopard, may help





PEARSON



# Why are we here?

PEARSON





PEARSON



# Educational Data Mining

≡

# Data Mining ?

PEARSON





Should we just publish at KDD\*?

*\*or other data mining venue*

PEARSON



Claim:

Educational Data Mining helps learners



PEARSON

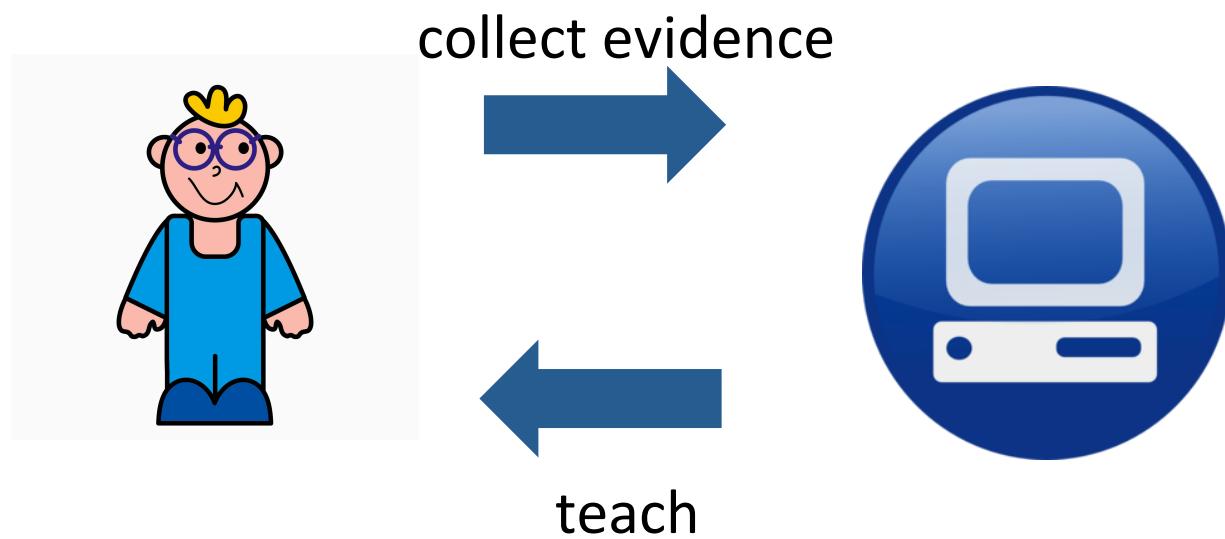


# Is our research helping learners?

PEARSON



# Adaptive Intelligent Tutoring Systems: Systems that teach and adapt content to humans



Paper writing: Researchers quantify the improvements of the systems compared

Not a purely academic pursuit:  
Superintendents to choose between alternative technology

Teachers choose between systems

Randomized Controlled Trials may measure the *time* students spent on tutoring, and their *performance* on post-tests

# Difficulties of Randomized Controlled Trials:

- IRB approval
- experimental design by an expert
- recruiting (and often payment!) of enough participants to achieve statistical power
- data analysis



# How do other A.I. disciplines do it?

PEARSON



Bleu [Papineni et al '01]:  
machine translation systems



Rouge [Lin et al '02]:  
automatic summarization systems



Paradise [Walker et al '99]:  
spoken dialogue systems



Using automatic metrics can be very positive:

- Cheaper experimentation
- Faster comparisons
- Competitions that accelerate progress

# Automatic metrics do not replace RCTs

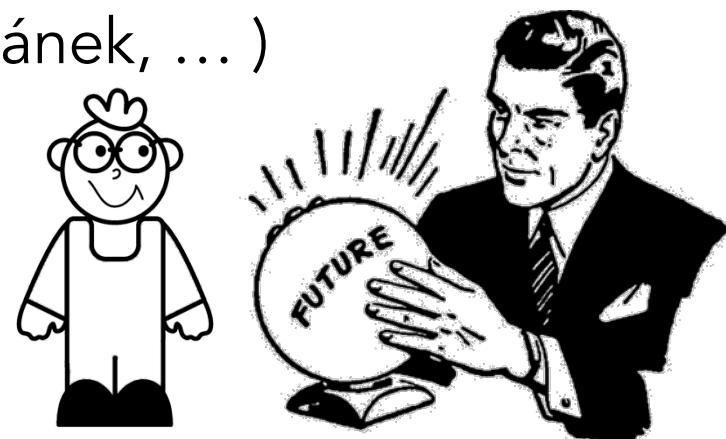
PEARSON



# What does the Educational Data Mining community do?

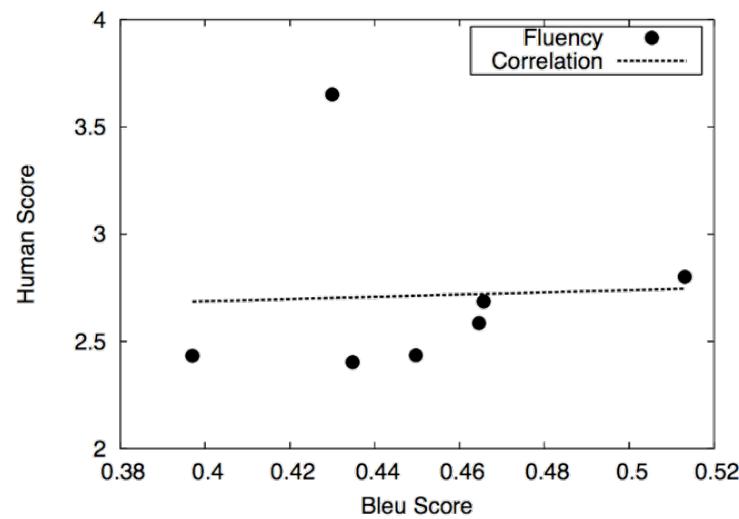
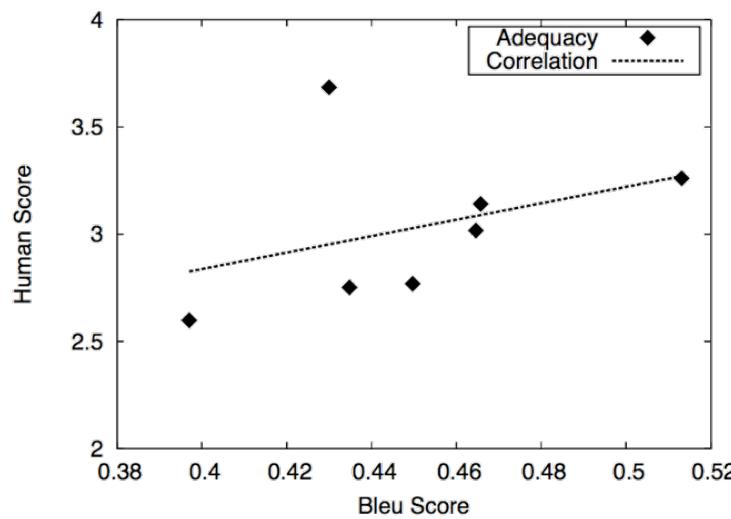
Evaluate the student model using classification accuracy metrics like RMSE, AUC of ROC, accuracy...

(Literature reviews by Pardos, Pelánek, ... )



Other fields verify that automatic metrics correlated with the target behavior

[Eg.: Callison-Burch et al '06]



PEARSON



Ironically, we have a growing body of evidence that classification evaluation metrics are a BAD way to evaluate adaptive tutors

*Read Baker and Beck papers on limitations / problems of these evaluation metrics*



Surprisingly, in spite of all of the evidence against using classification evaluation metrics, their use is still very widespread in the adaptive literature\*

Can we do better?

\* ExpOppNeed [Lee & Brunskill] is an exception

The rest of this talk:

- Leopard Paradigm
  - Teal
  - White
- Meta-evaluation
- Discussion

# Leopard



© AP/ Chicago Zoological Society, Jim Shultz

PEARSON



- Leopard: Learner Effort-Outcomes Paradigm
- Leopard quantifies the *effort* and *outcomes* of students in adaptive tutoring

- Effort: Quantifies how much practice the adaptive tutor gives to students. Eg., number of items assigned to students, amount of time...
- Outcome: Quantifies the performance of students after adaptive tutoring

- Measuring effort and outcomes is not novel by itself (e.g, RCT)
- Leopard's contribution is measuring both without a randomized control trial
- White and Teal are metrics that operationalize Leopard

# White: Whole Intelligent Tutoring system Evaluation

White performs a counterfactual simulation (“What Would the Tutor Do?”) to estimate how much practice students receive

Design desiderata:

Evaluation metric should be easy to use

Same, or similar input than conventional metrics

predicted performance				
actual performance				
t	student <i>u</i>	skill <i>q</i>	$\hat{c}_{u,q,t+1}$	$y_{u,q,t}$
0	Alice	s1	.6	
1	Alice	s1	.5	0
2	Alice	s1	.5	1
3	Alice	s1	.6	1
<hr/>				
0	Bob	s1	.4	
1	Bob	s1	.7	1
2	Bob	s1	.7	1
3	Bob	s1	.7	1
4	Bob	s1	.8	0
4	Bob	s1	.9	1
6	Bob	s1	.9	1

PEARSON



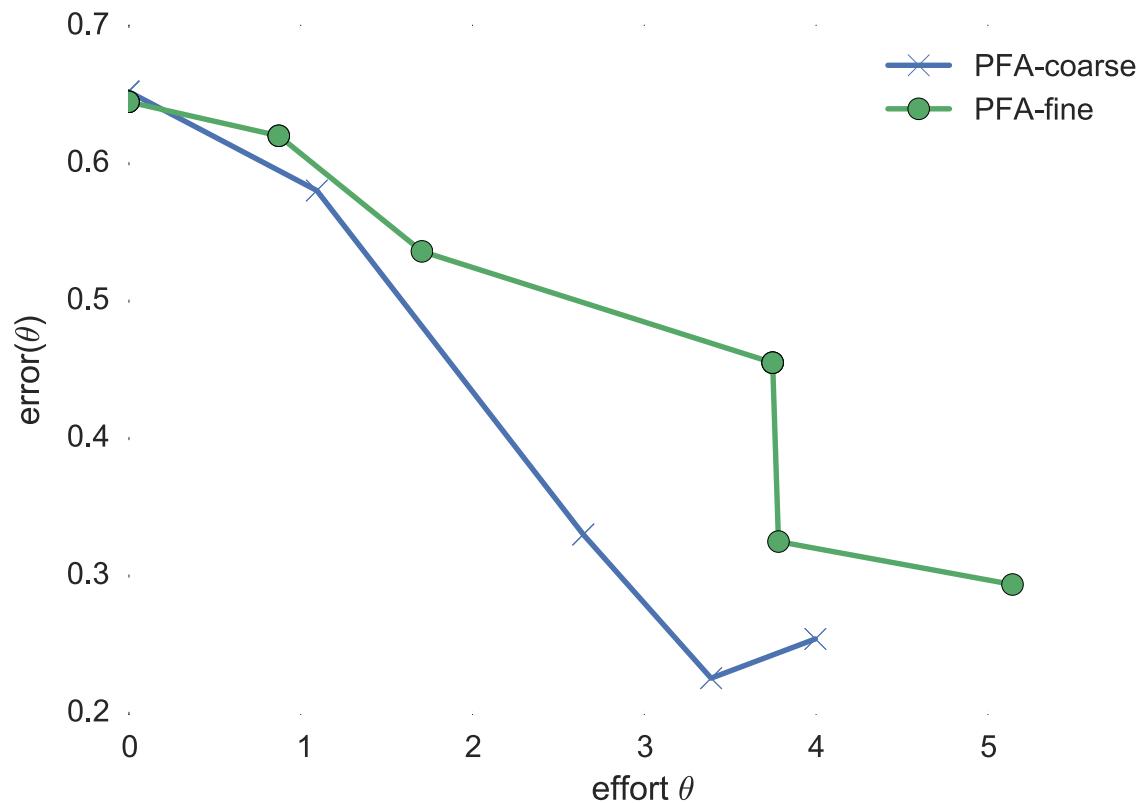
predicted performance				
actual performance				
t	student $u$	skill $q$	$\hat{c}_{u,q,t+1}$	$y_{u,q,t}$
effort=0	Alice	s1	.6	
	Alice	s1	.5	0
	Alice	s1	.5	1
	Alice	s1	.6	1
effort=1	Bob	s1	.4	
	Bob	s1	.7	1
	Bob	s1	.7	1
	Bob	s1	.7	1
	Bob	s1	.8	0
	Bob	s1	.9	1
	Bob	s1	.9	1

2/3 score

Alternatively,  
we can  
model error

4/5 score

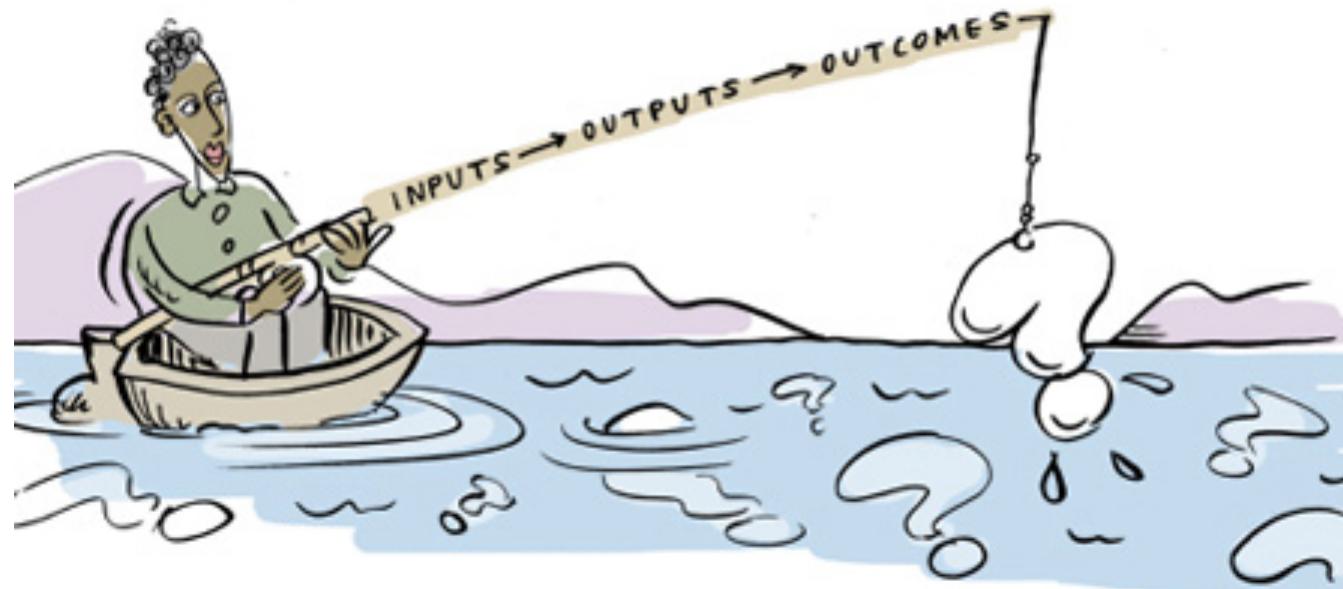
# *Future direction: Present aggregate results*



Q/ What if student does not achieve target performance?

A: “Visible” imputation

# Meta-Evaluation



Compare:

- Conventional classification metrics
- Leopard metrics (White)

## Datasets:

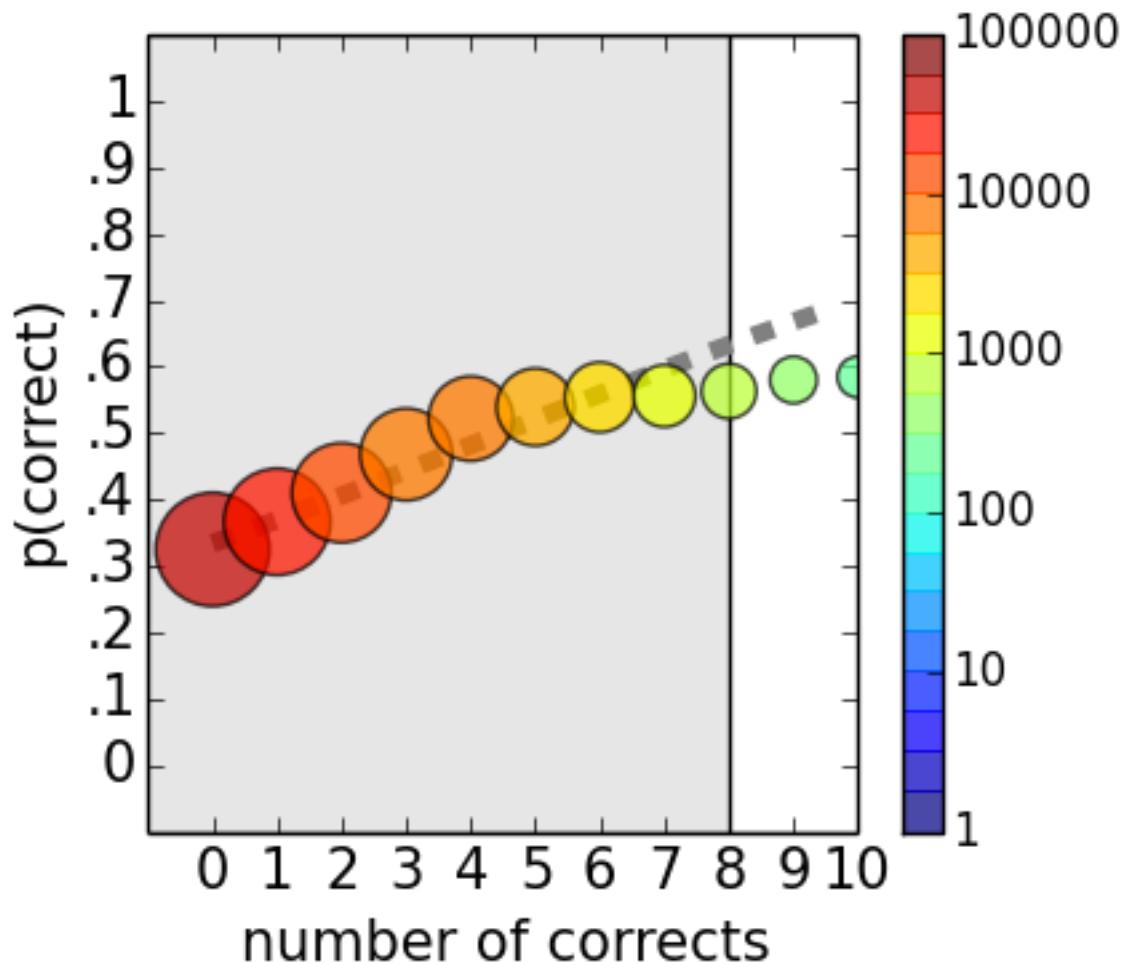
- Data from a middle-school Math commercial tutor
  - 1.2 million observations
  - 25,000 students
  - Item to skill mapping:
    - Coarse: 27 skills
    - Fine: 90 skills
    - (Other item-to-skill model not reported)
- Synthetic data

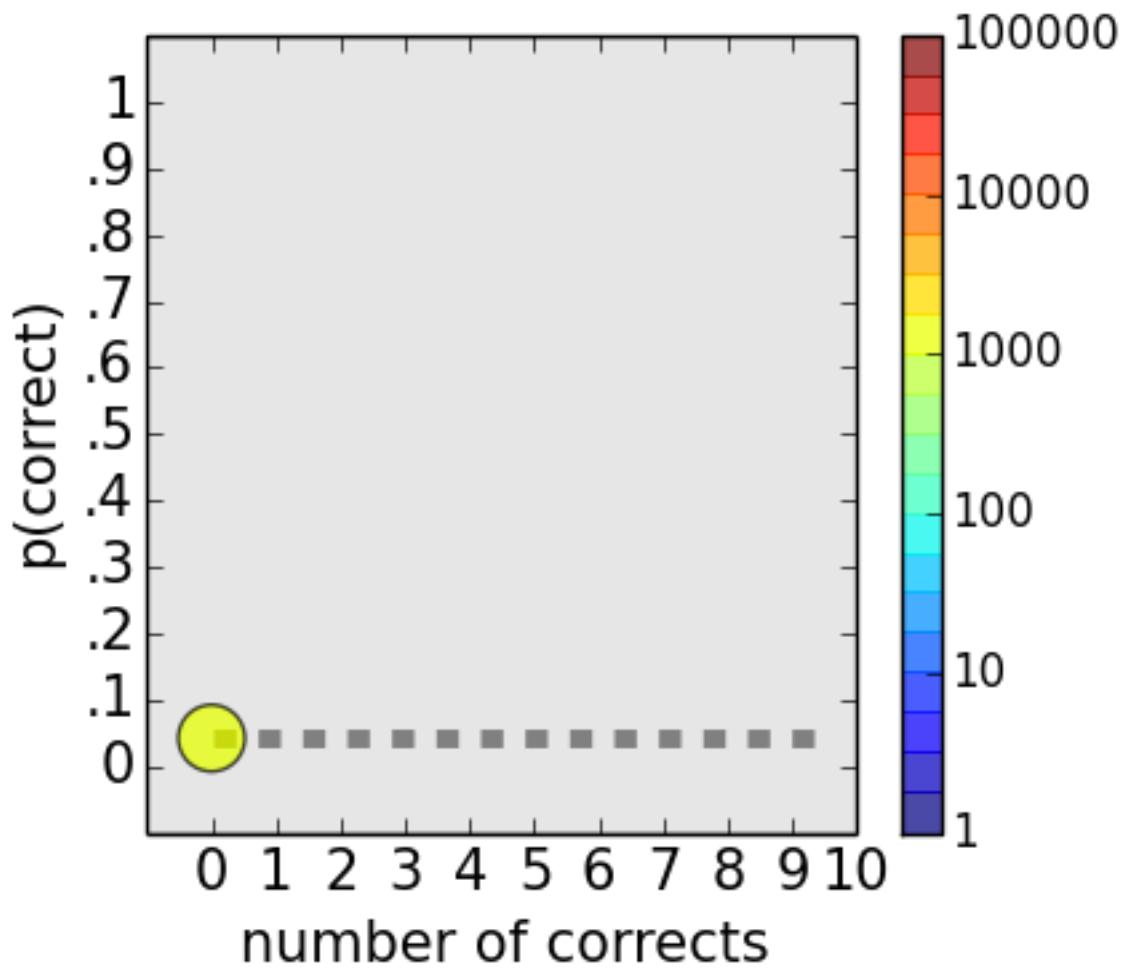
Assessing an evaluation metric with real student data is difficult because we often do not know the ground truth

Insight: Use data that we know a priori its behavior in an adaptive tutor

For adaptive tutoring to be able to optimize when to stop instruction, the student performance should increase with repeated practice (the learning curve should be increasing)

Decreasing /flat learning curve = bad data





## Procedure:

1. Select skills with decreasing/flat learning curve (aka bad data)
2. Train a student model on those skills
3. Compare classification metrics with Leopard

	F1	AUC	Score	Effort
Bad student model	.79	.85		
Majority class	0	.50		



PEARSON



	F1	AUC	Score	Effort
Bad student model	.79	.85	.18	10.1
Majority class	0	.50	.18	11.2



## What does this mean?

- High accuracy models may not be useful for adaptive tutoring
- We need to change how we report results in adaptive tutoring

# Solutions

- Report classification accuracy averaged over skills (for models with 1 skill per item)
  - ✖ Not useful for comparing or discovering different skill models
- Report as “difficulty” baseline
  - ✖ Experiments suggest that models with baseline performance can be useful
- Use Leopard

Let's use all data, and pick an item-to-skill mapping:

	AUC	Score	Effort
Coarse (27 skills)	.69		
Fine (90 skills)	.74		



Let's use all data, and pick an item-to-skill mapping:

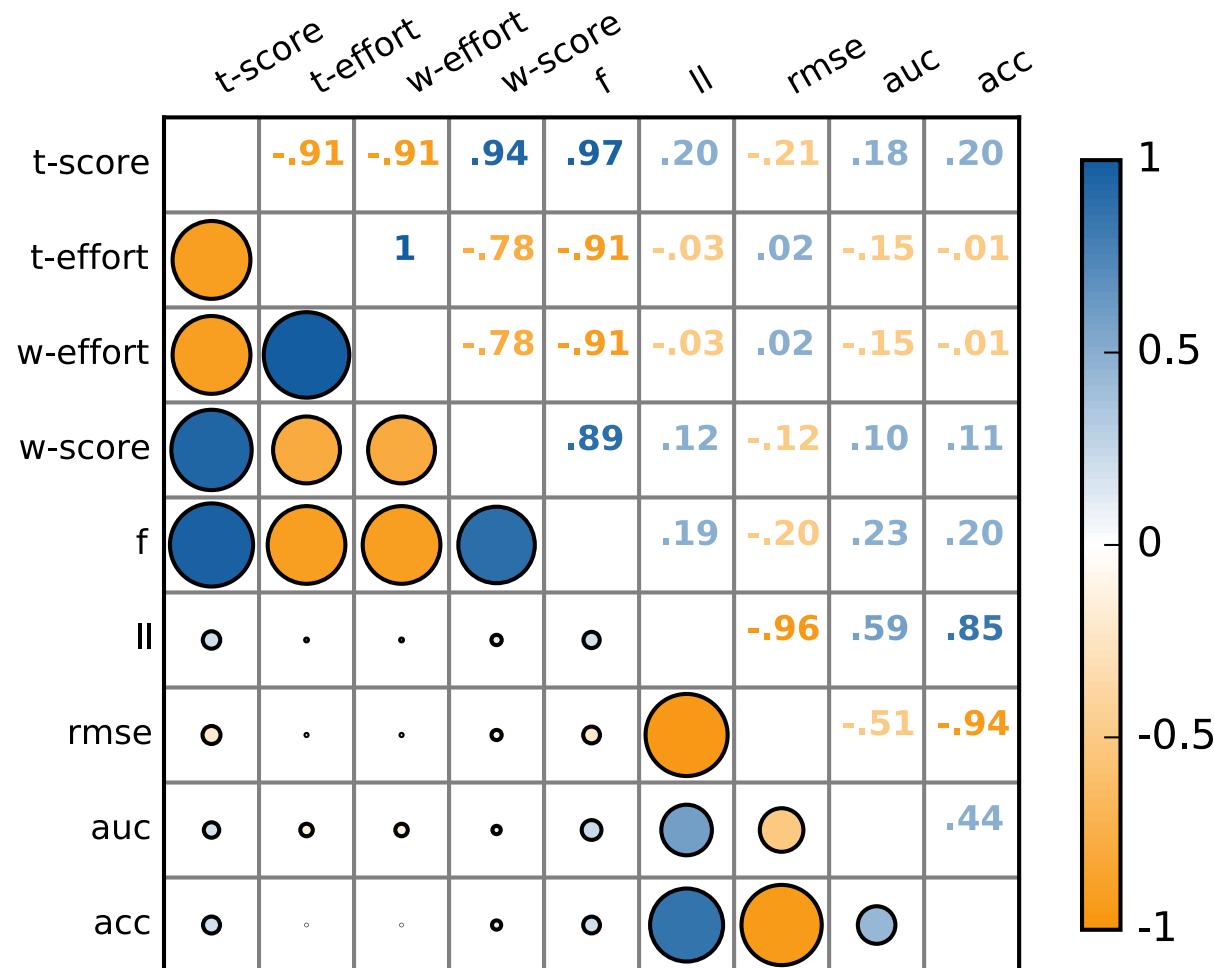
	AUC	Score	Effort
Coarse (27 skills)	.69	.41	55.7
Fine (90 skills)	.74	.36	88.1



With synthetic data we can use Teal as the ground truth

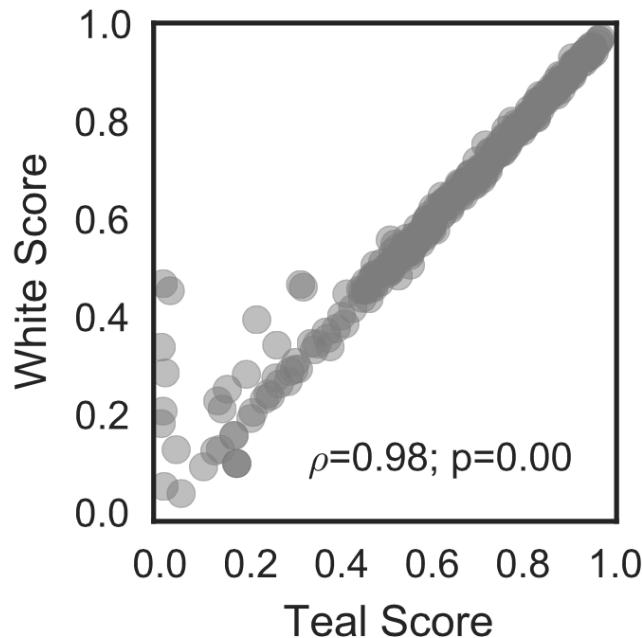
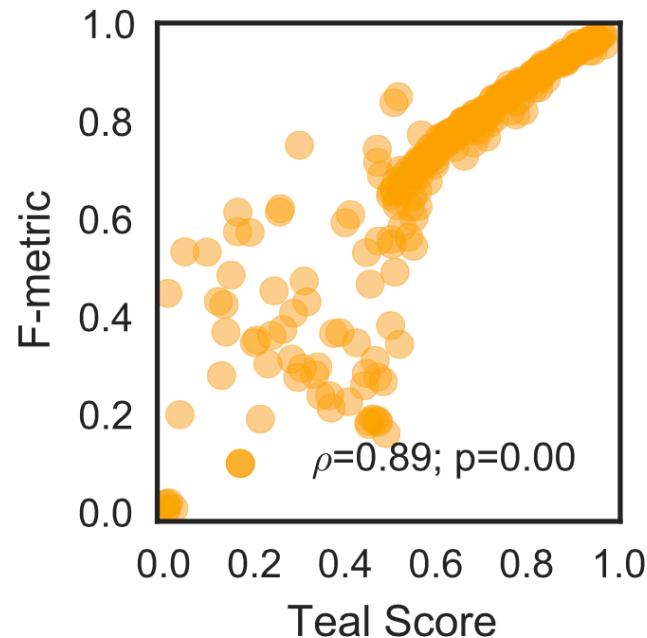
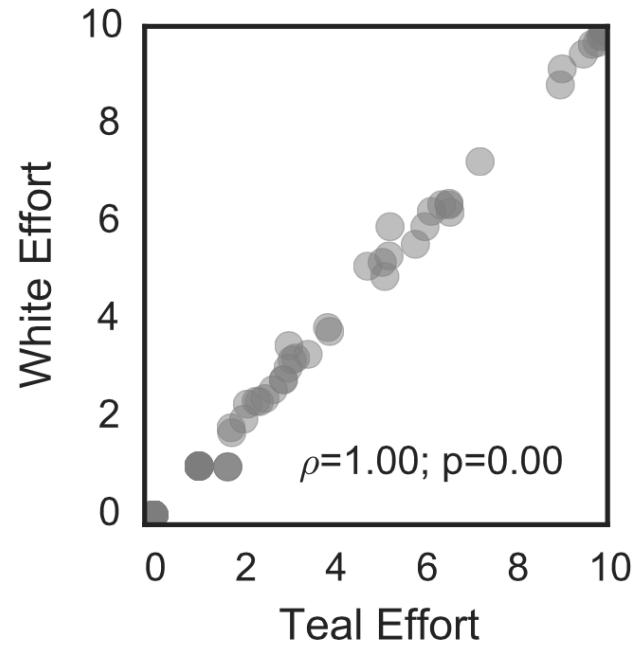
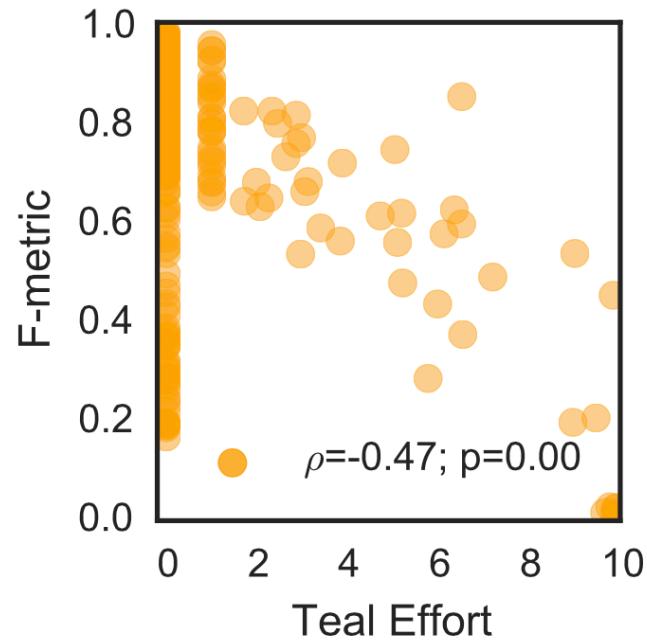
We generate 500 synthetic datasets with known Knowledge Tracing Parameters

Which metrics correlate best to the truth?



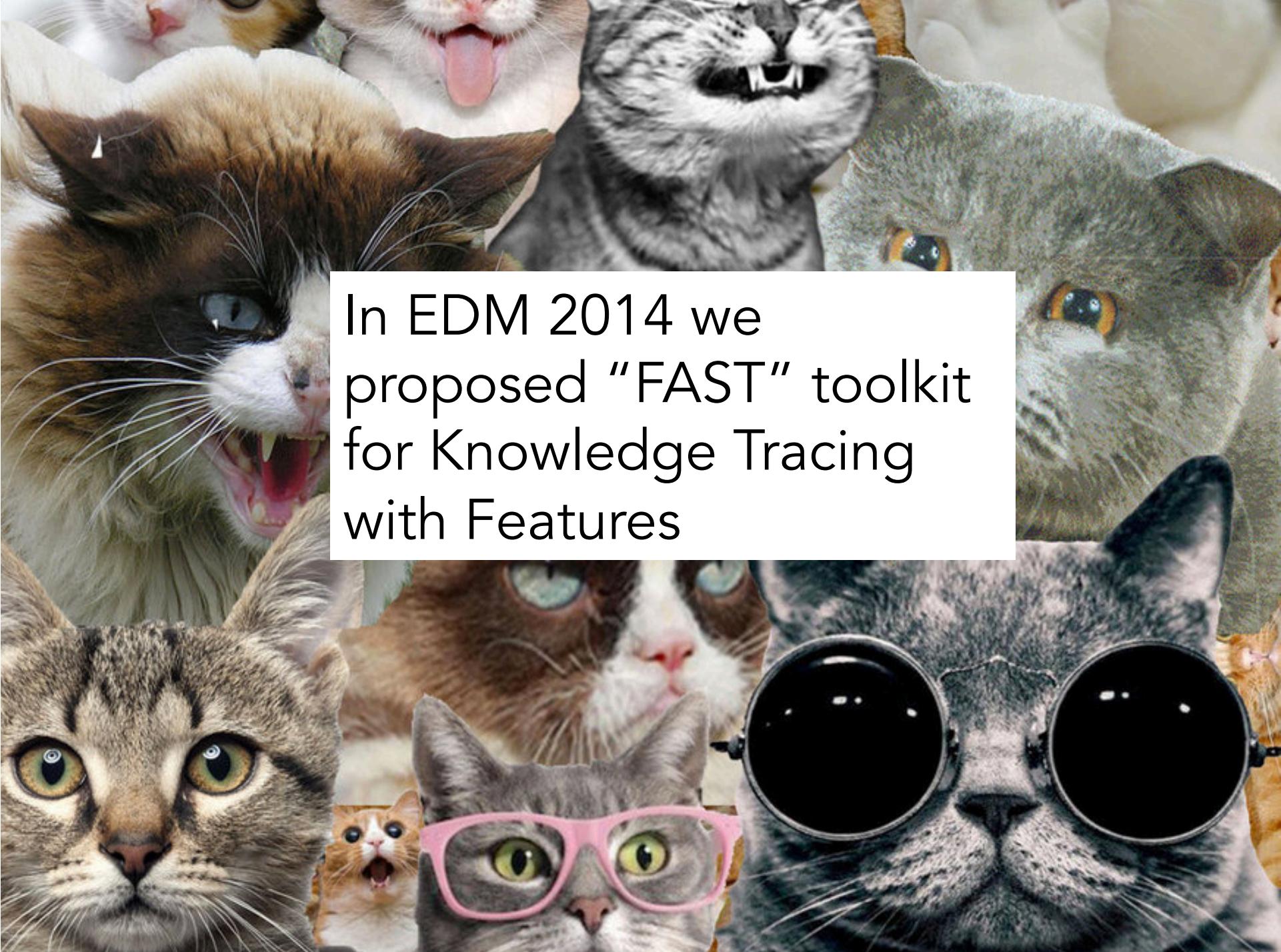
PEARSON





# Discussion





In EDM 2014 we proposed “FAST” toolkit for Knowledge Tracing with Features

“FAST model improves 25% AUC of ROC”





PEARSON



PEARSON



# Input

Teal	White
Knowledge Tracing Family parameters	Student's correct or incorrect response
Sequence length	Student models' prediction of correct/ incorrect
Target Probability of correct	Target Probability of correct