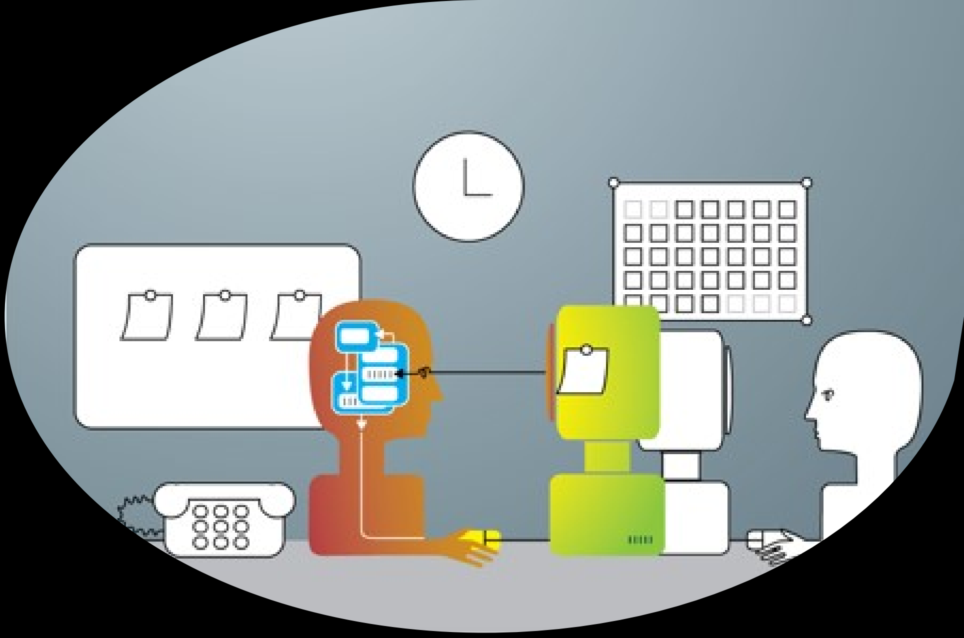
Machine-learned Computational models to Assess Ill-defined Constructs



Sidney K. D'Mello
University of Colorado Boulder

sidney.dmello@colorado.edu
www.colorado.edu/ics/sidney-dmello
August 23, 2019

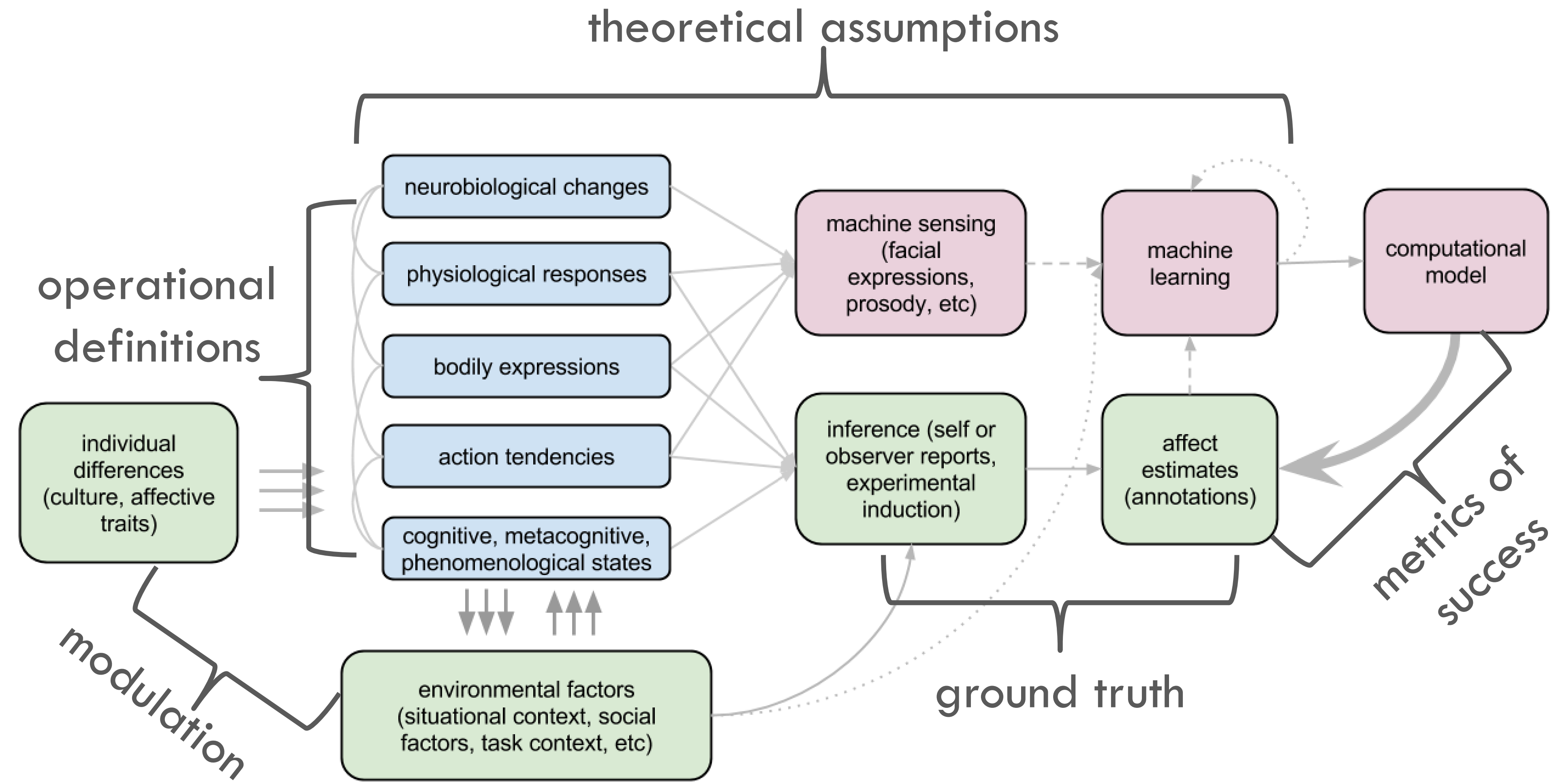
- traditional methods (classical test theory, item response theory, evidence centered design) have been invaluable for assessing a range of constructs (e.g., knowledge, skills)
- but what about "ill-defined" constructs that cannot be precisely defined, are ephemeral states, especially in situ?

machine-learned, computational models are essential

- when constructs are "ill-defined" like emotion, collaboration
- when there are no adequate theoretical mechanistic accounts
- when underlying models are "multilevel circular causal"

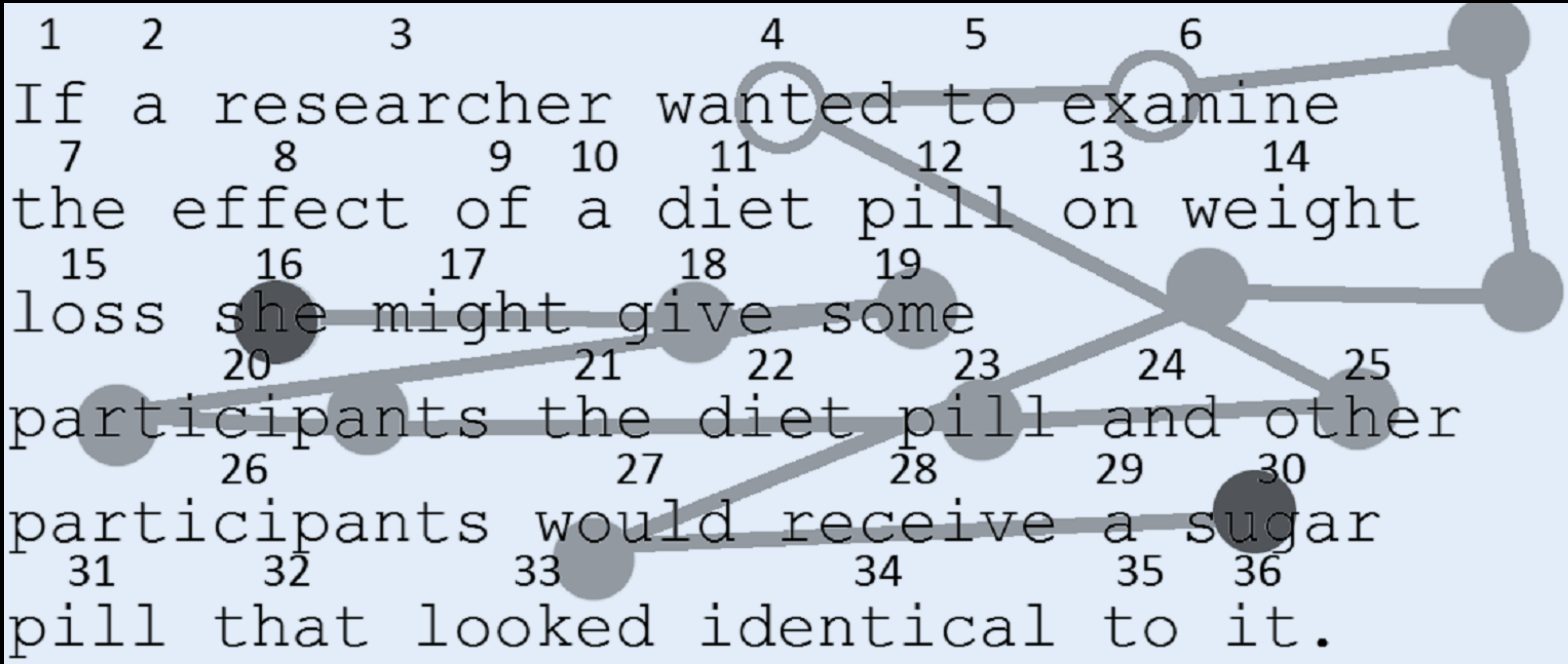
- models can promote change via intervention or reflection
- the art lies in how they are constructed and evaluated
- and in setting realistic expectations and contexts of use

claims



conceptual model [affect example]

D'Mello, Kappas, & Gratch (2018)

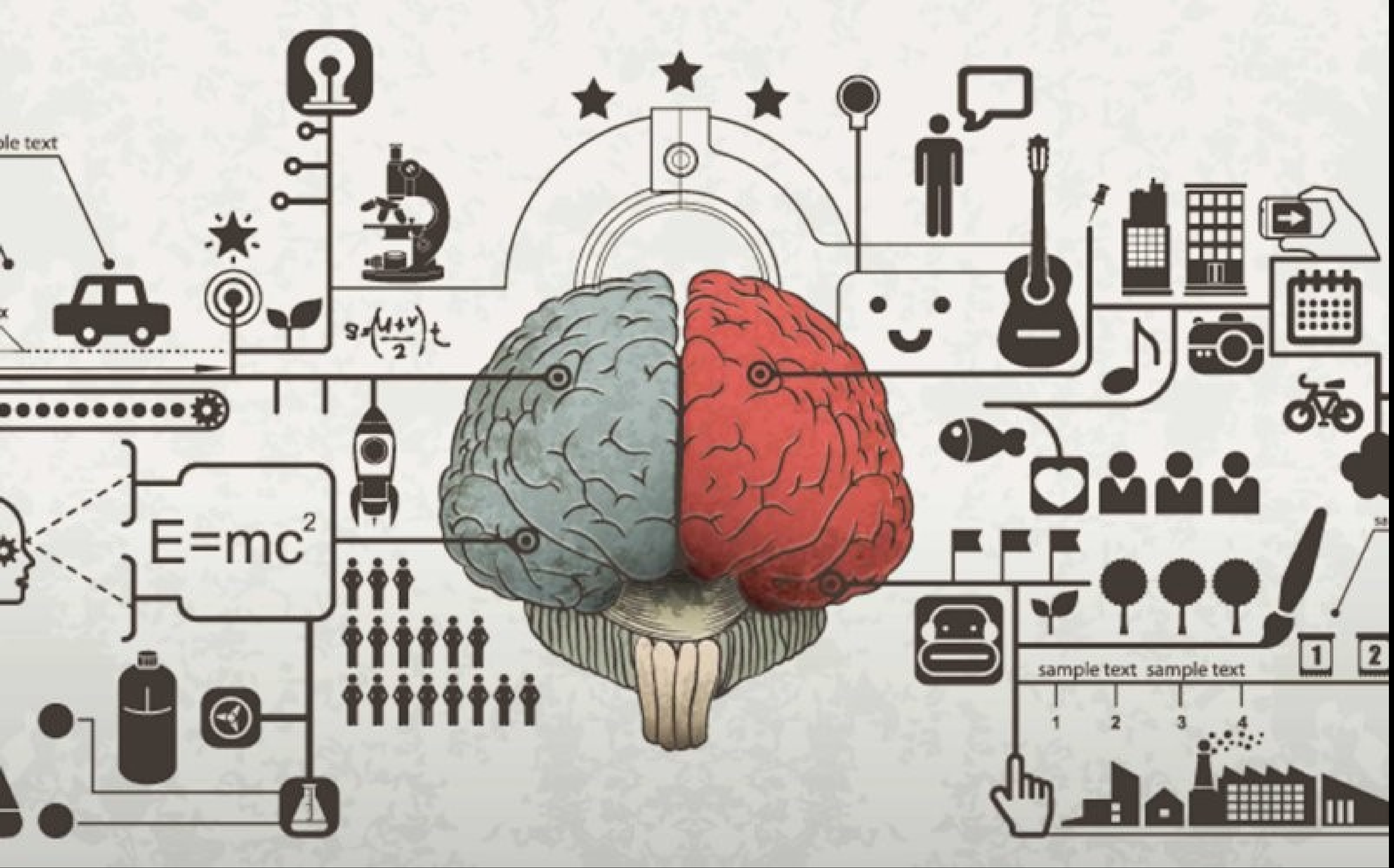




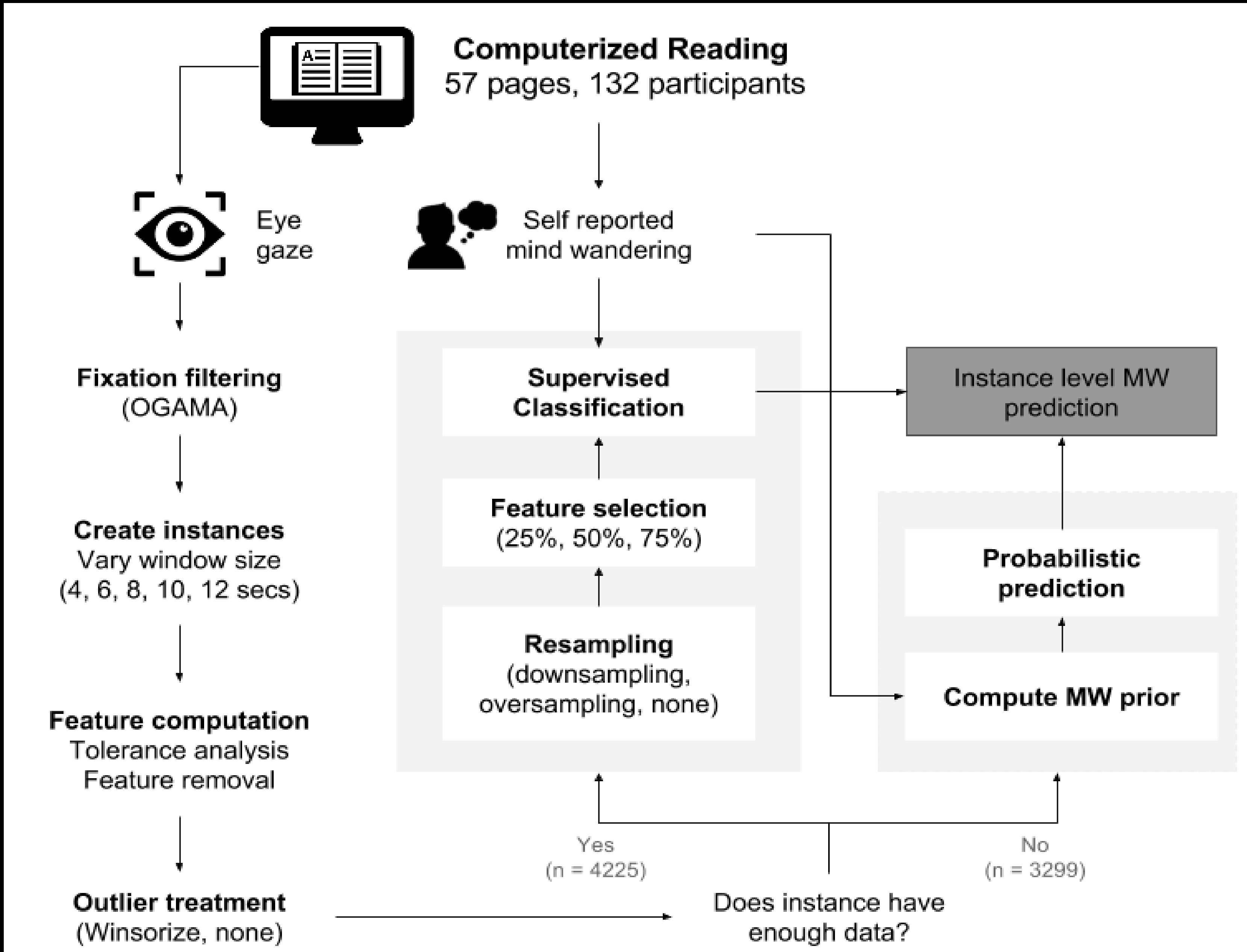




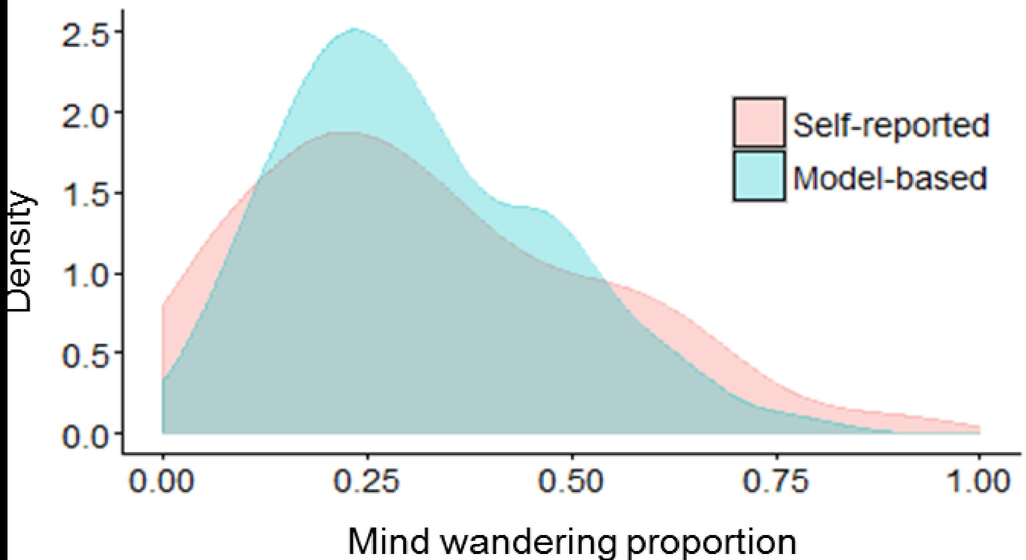

exploring the
eye-mind link
during reading



ubiquity of mind wandering



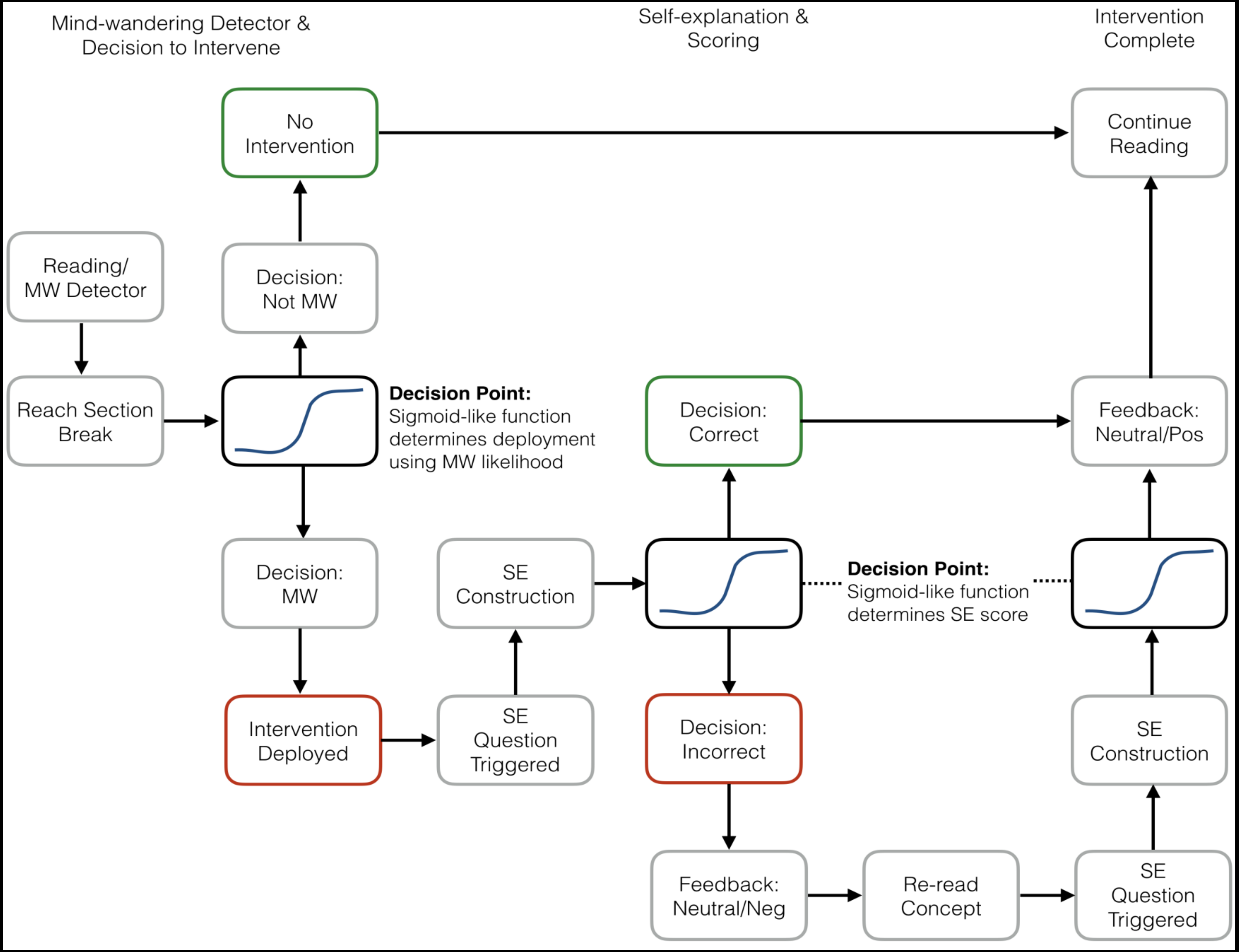
method (Faber, Bixler, & D'Mello, 2018)



- model estimates correlated with self-reported mind wandering (*r* = .400)
- correlated with comprehension (*r* = -.374) stronger than self-reports (*r* = -.208)

- models robust to missing data and internally consistent (*r* = .751)
- page-level predictions moderate – precision of 72.2%; recall of 67.4%
- fewer but longer fixations and fewer horizontal saccades related to mind wandering

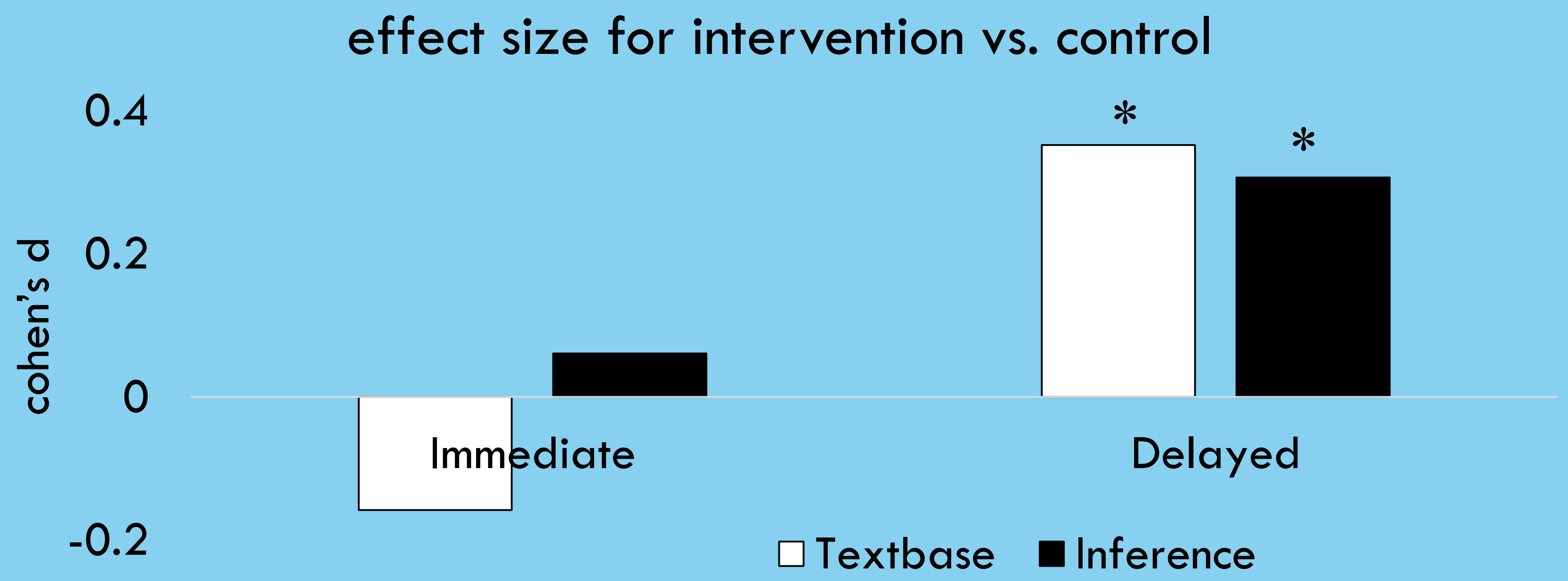
key results



real-time intervention (Mills, et al., in review)

method

- 70 participants read book on surface tension in liquids
- randomly assigned to intervention or yoked-control
- tested for text- AND inference- level comprehension after reading AND one week later (parallel forms)

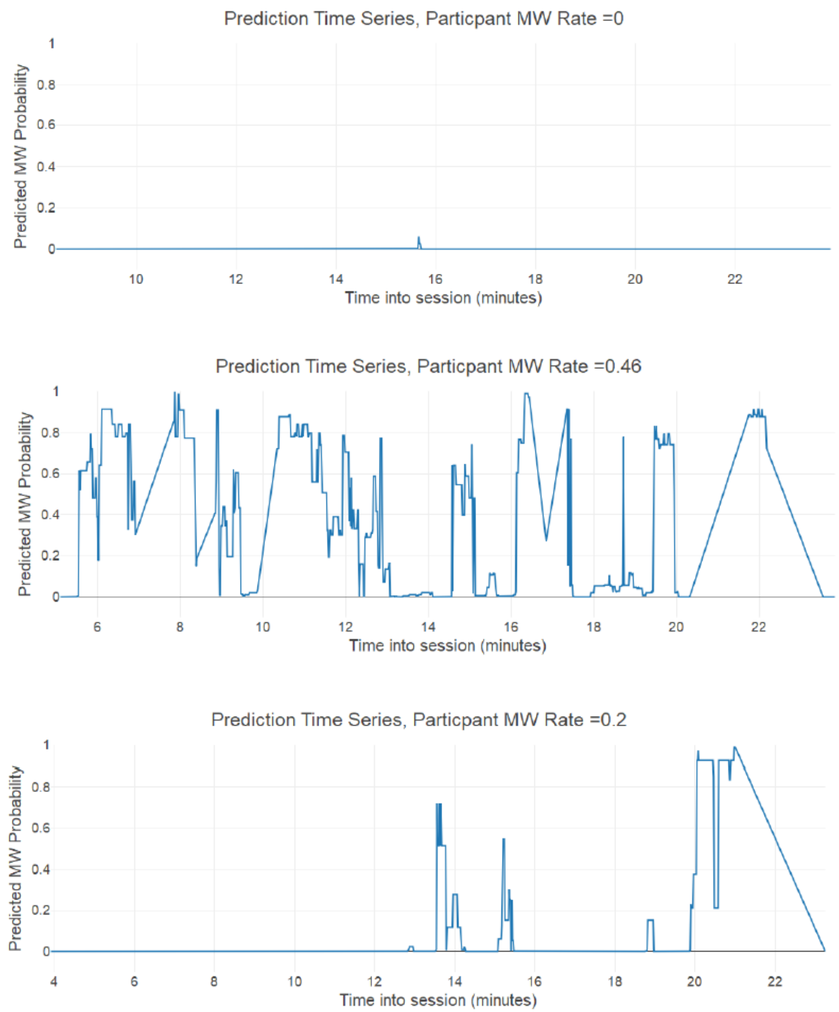


experimental validation

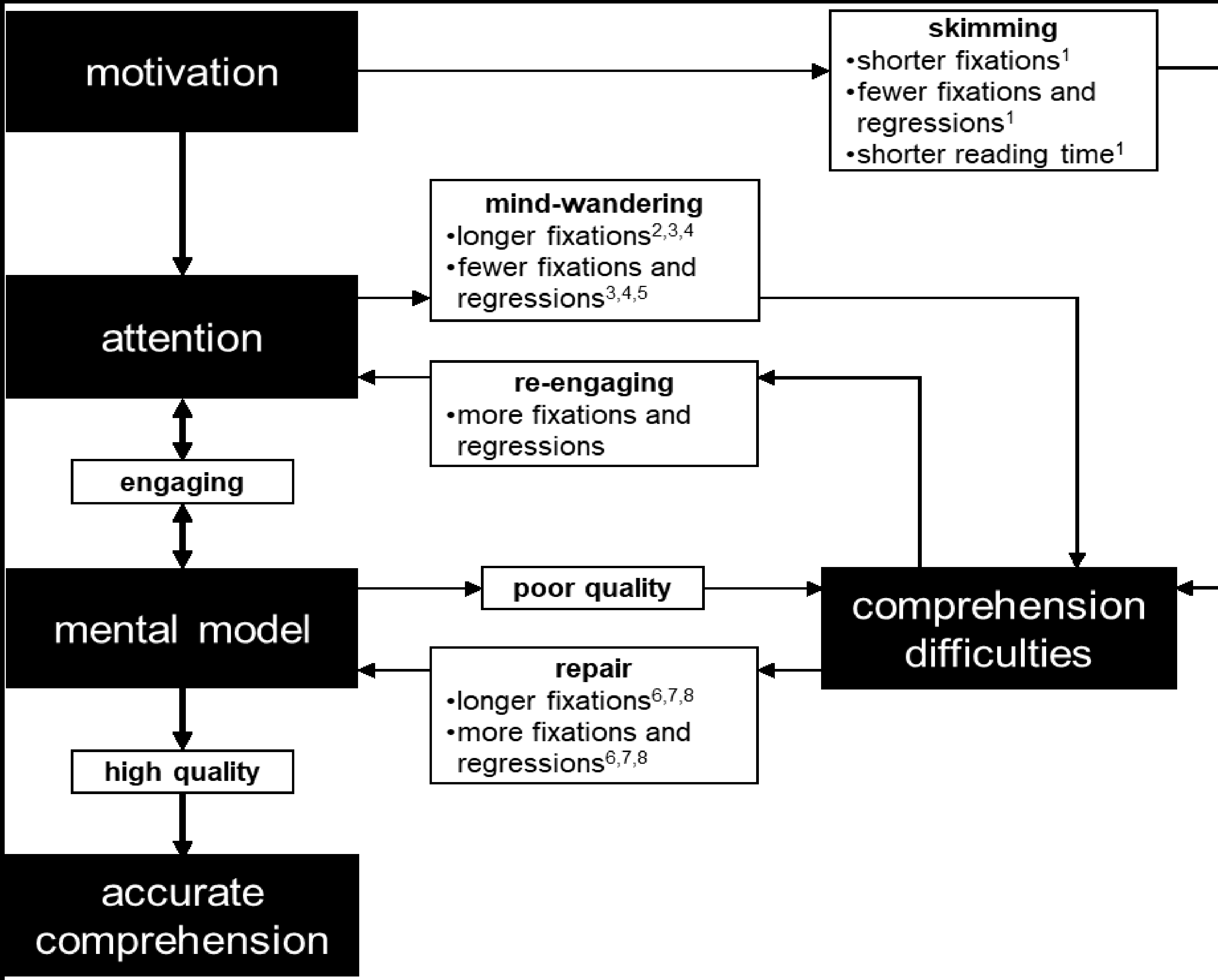
out of the lab and into the wild (Hutt et al., 2019)

- tracking validity between 75% (both eyes) and 95% (one eye)
- moderately accurate at mind wandering detection (precision .55; recall .65)
- model predictions correlated with learning (r = -20)

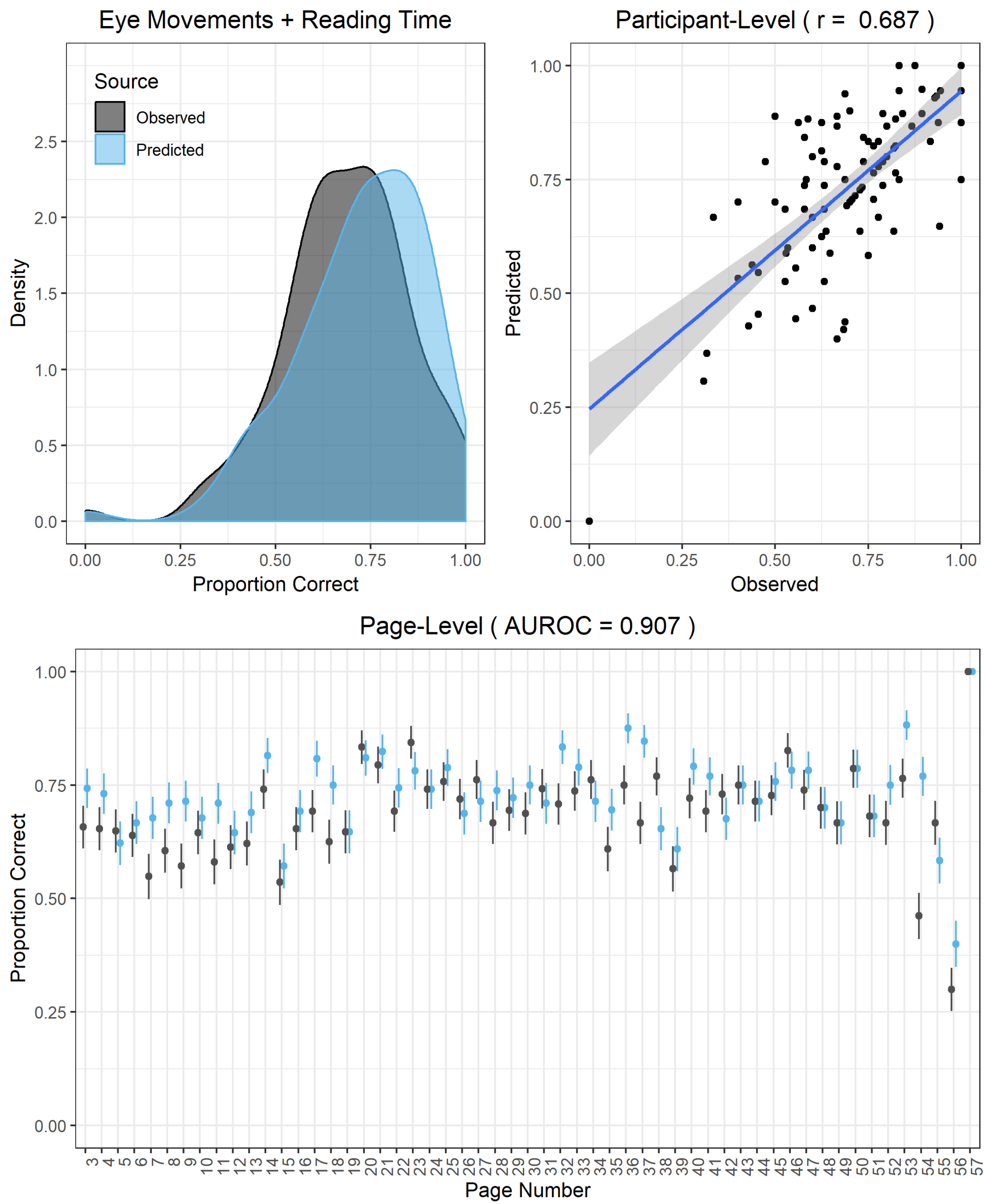
using models for interventions



key results



can eye movements predict comprehension?



very accurate for textbase-level comprehension assessed during reading (AUROC = 0.9; r = 0.68)

Gregg & D'Mello (in review)

motivation

- surprising lack of consistency in literature
- very little research on long connected texts, especially after reading
- tested weak vs. strong association hypotheses (*Rsq*. of 1% vs. 10%)

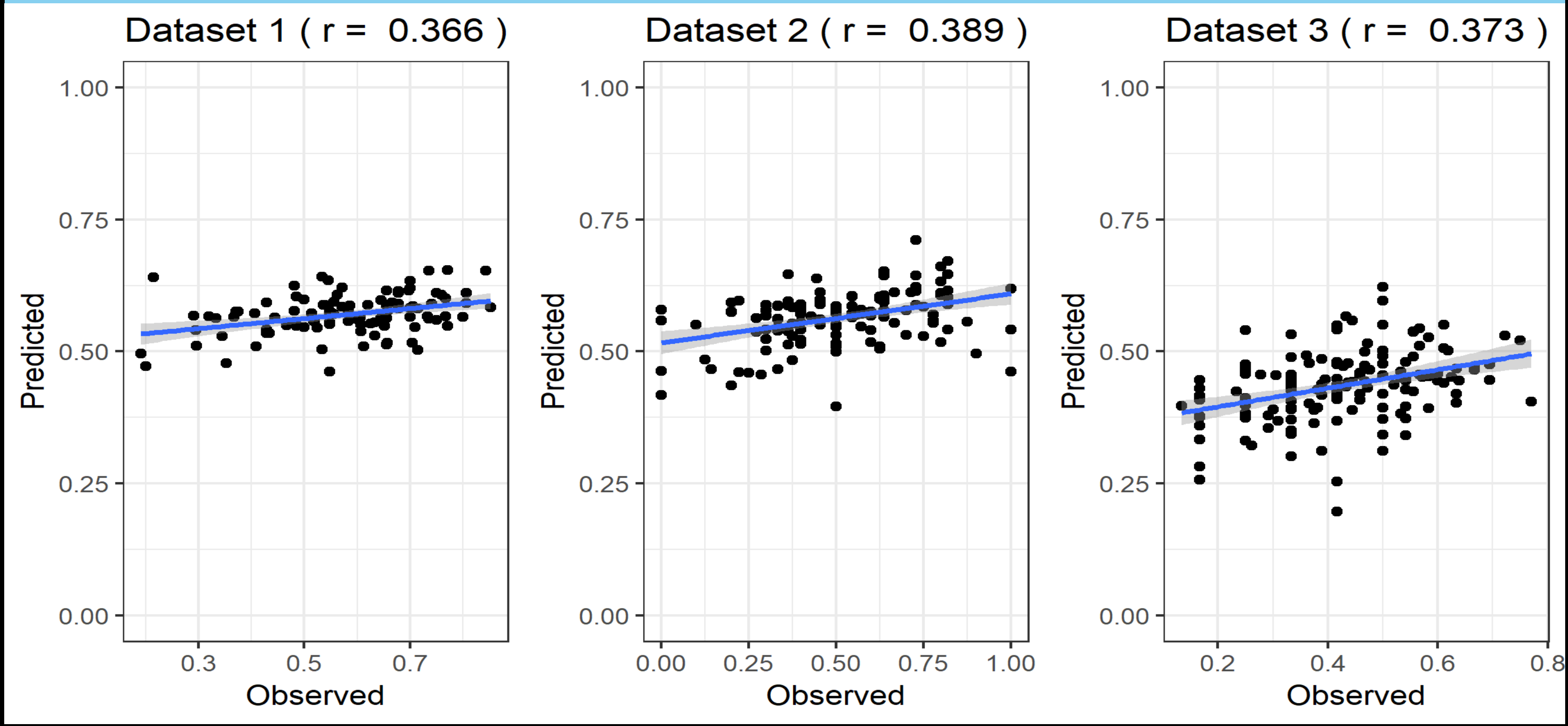
methods

- datasets 1 and 2: predict textbase-level comprehension 30-mins after reading one long connected text
- dataset 3: predict textbase- and inference-level comprehension after reading upto 8 short texts
- focused on seven eye gaze features and reading times
- simple cross-validated regression models

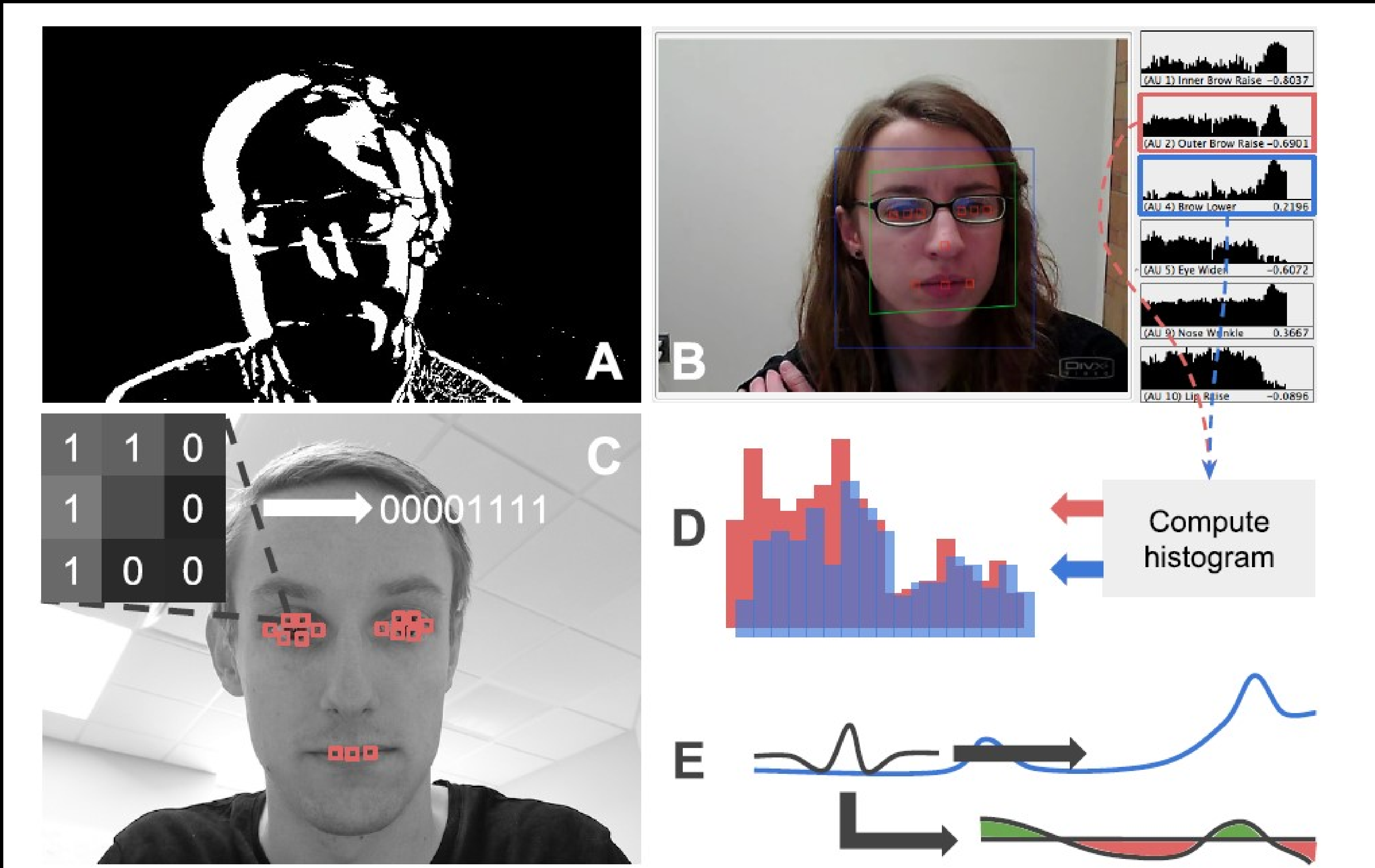
what about comprehension *after* reading?

(Gregg, Bixler, & D'Mello, in review)

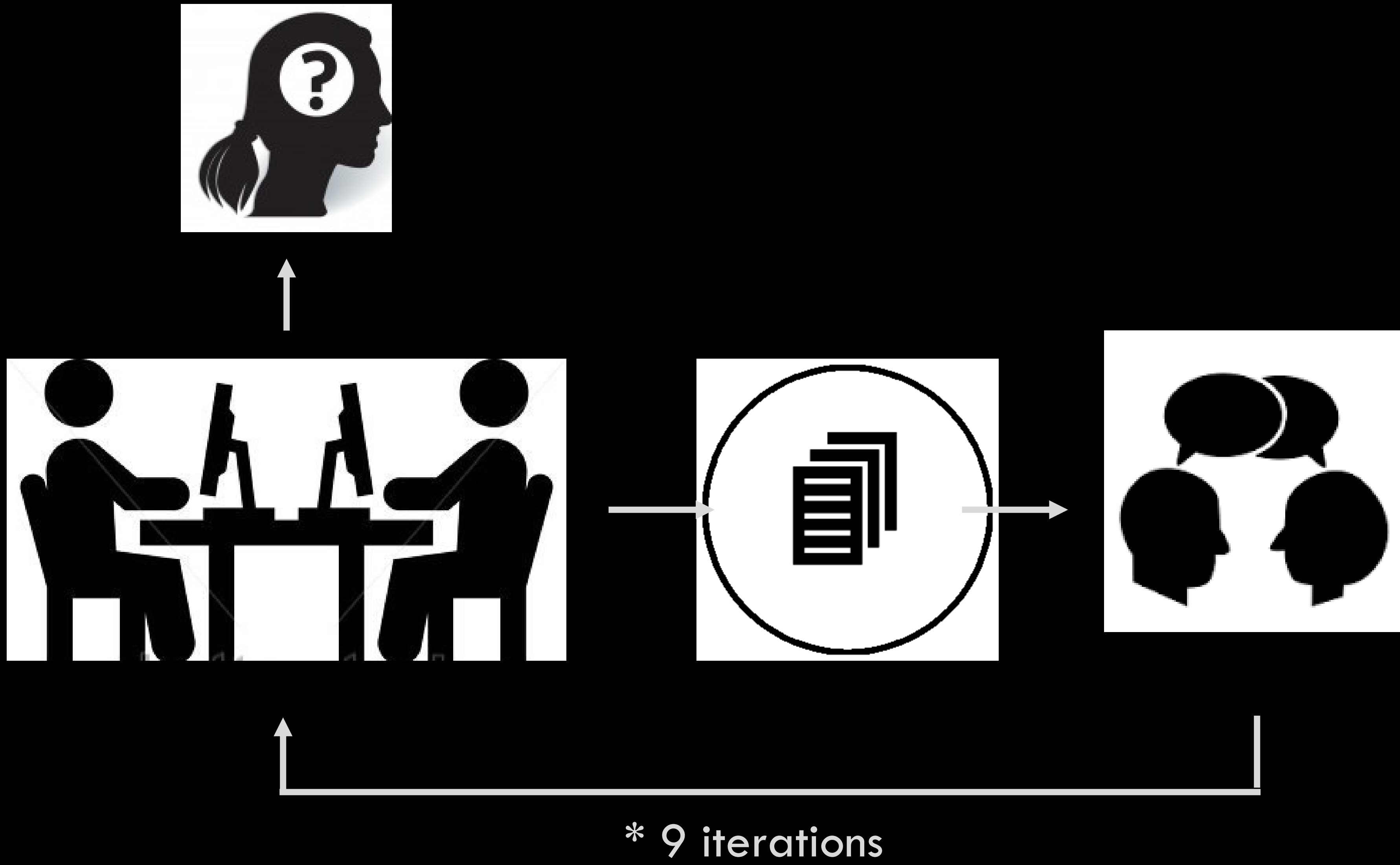
- moderate cross-validated correlations between observed and predicted comprehension
- models from one study generalized to another
- more, but shorter, fixations predicted comprehension
- results hold after accounting for mind wandering and exposure to print (author recognition test)



data support the strong association hypothesis machine-learned computational models of eye movements can assess reading processes and outcomes & can drive intervention



video-based modeling
of affect and attention



does frame-of-reference coding help

D’Mello (2016)

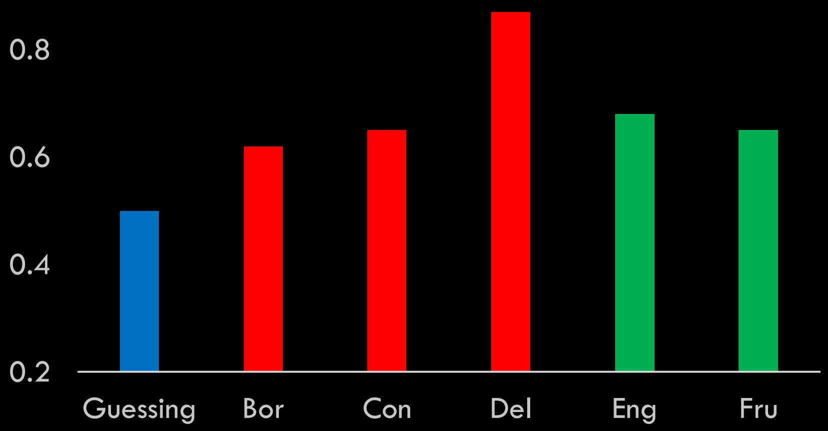

0.4
0.2
0
Kappa
between judges
agreement with self
0
2
4
6
8
10
Iteration

it depends...

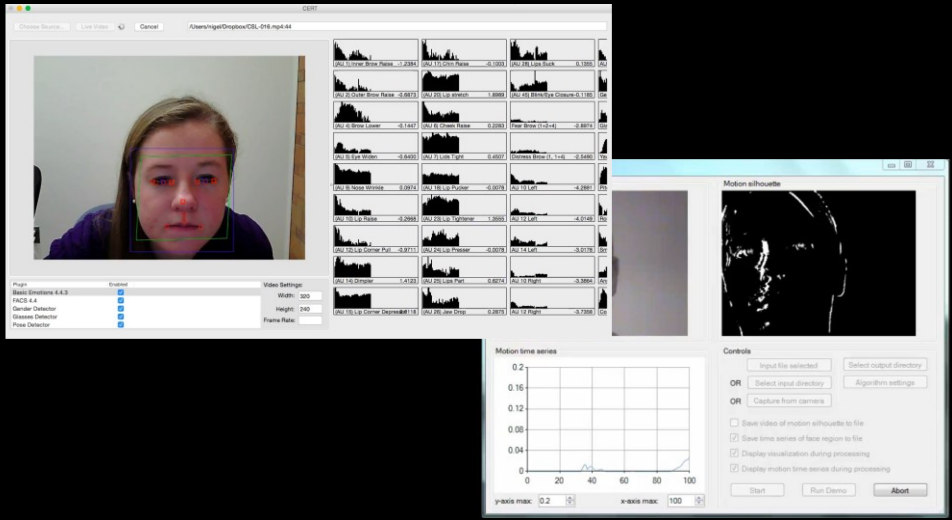
Physics Playground

online observations

results (AUC)

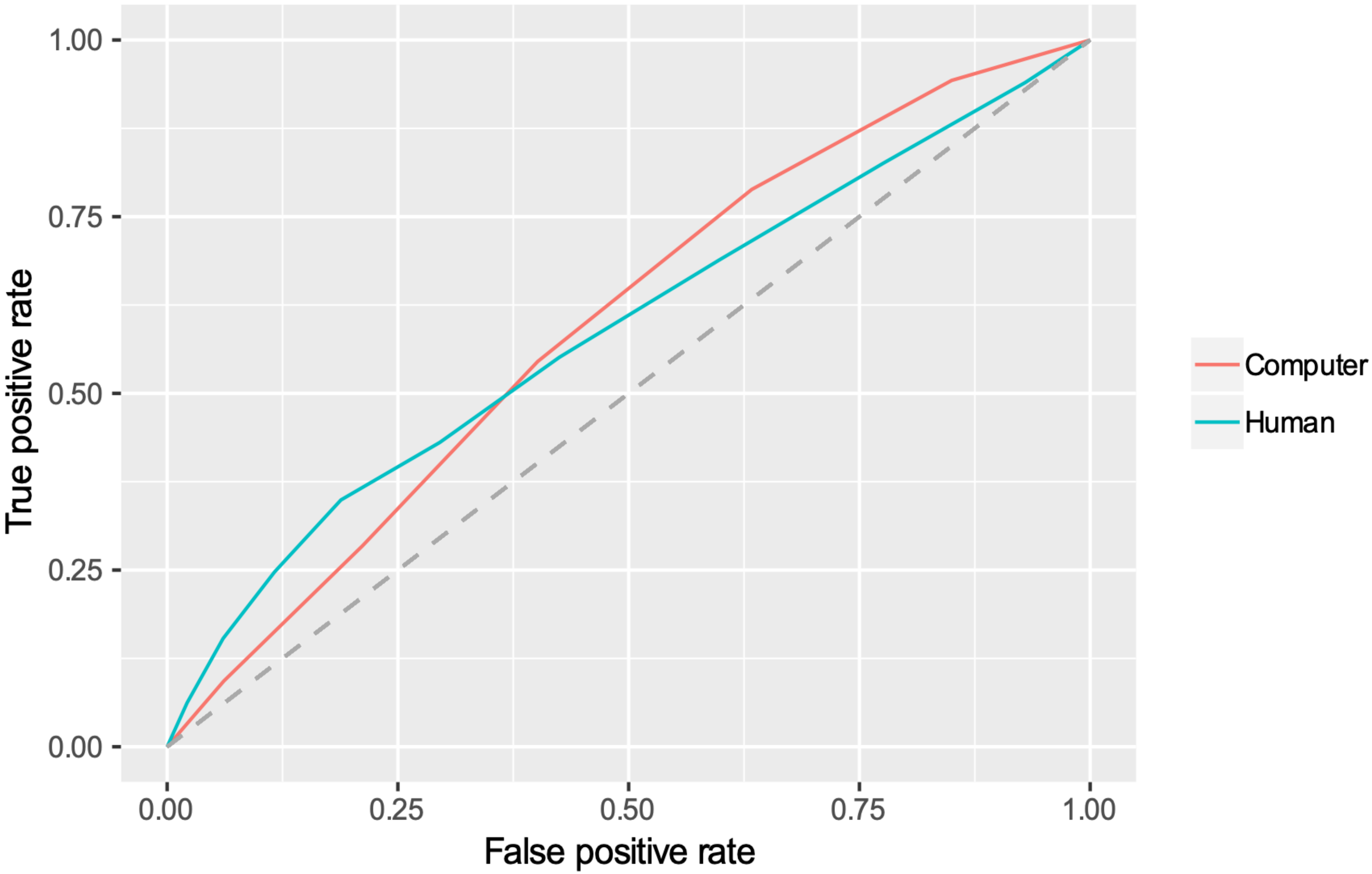


facial features + body movements



modeling affect from video

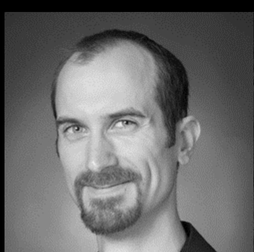
(Bosch, et al., 2016)



video-based mind wandering detection

Bosch & D’Mello (2019)

video-based models can provide human-comparable results for affect and attention




speech and language processing for discourse analysis

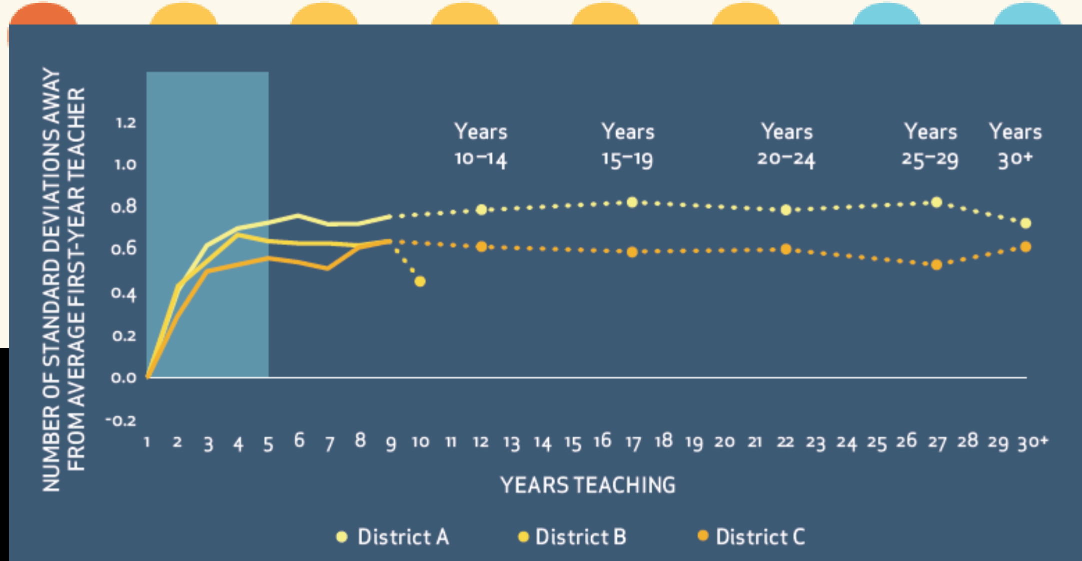
2015

TNTP reimagine teaching

The districts we studied spend an average of nearly $18,000 per teacher, per year on

Give teachers a clear, deep understanding of their own performance and progress.

Confronti



chers
stantial
t.

Which of these would you consider authentic?

Teacher: *"How does a person become a noble?"*

Student: "They're born into it"

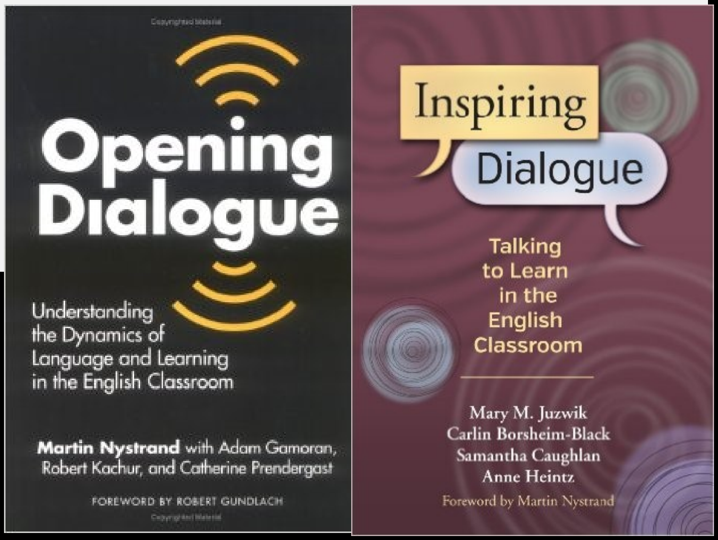
Teacher: *"They're born into it, right? It's by family. It gets passed down*

….

Teacher: *"How did that make you guys feel, I mean what was your gut reaction to all that?"* authentic

Student: "Ashamed"

Teacher: *"Ashamed in what way?"*



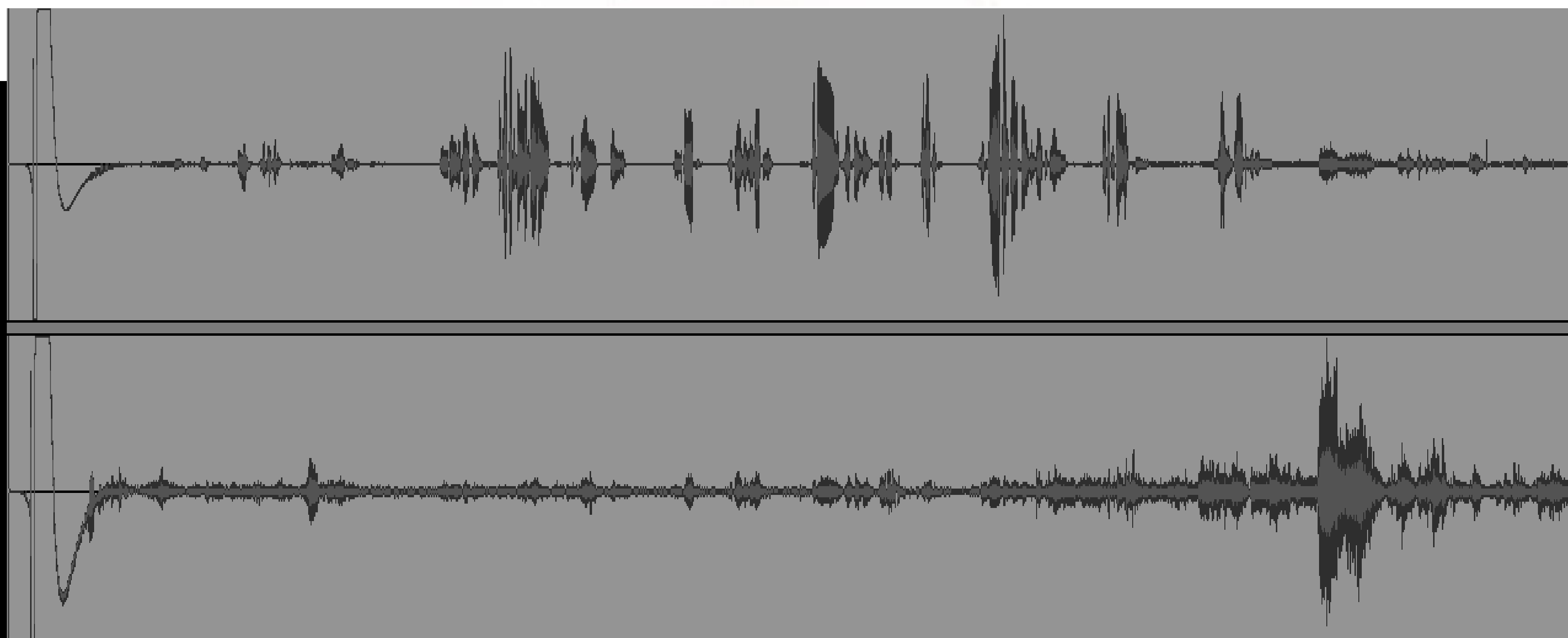
authentic questions

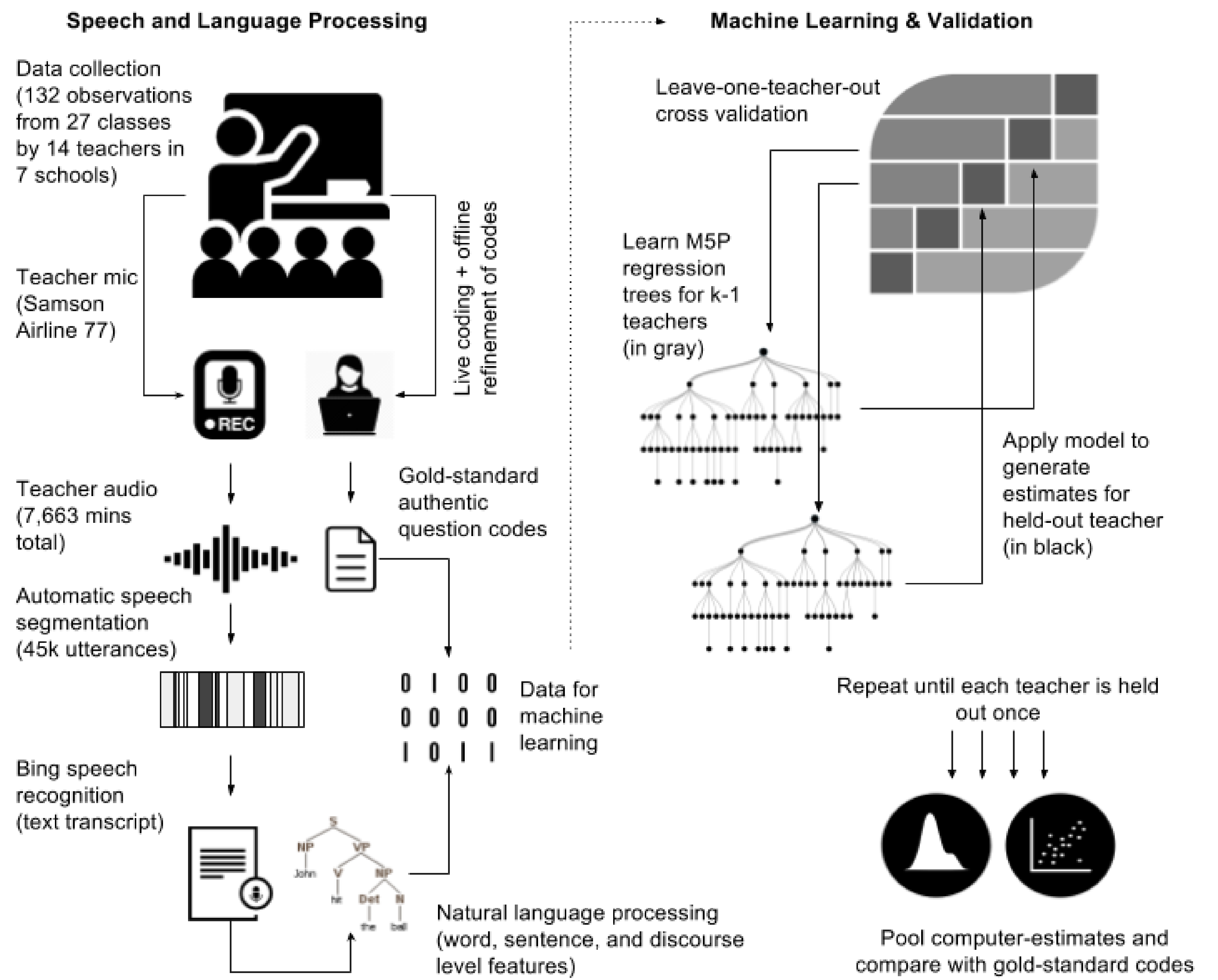
teacher mic (Samson Airline 300)

Classroom mic

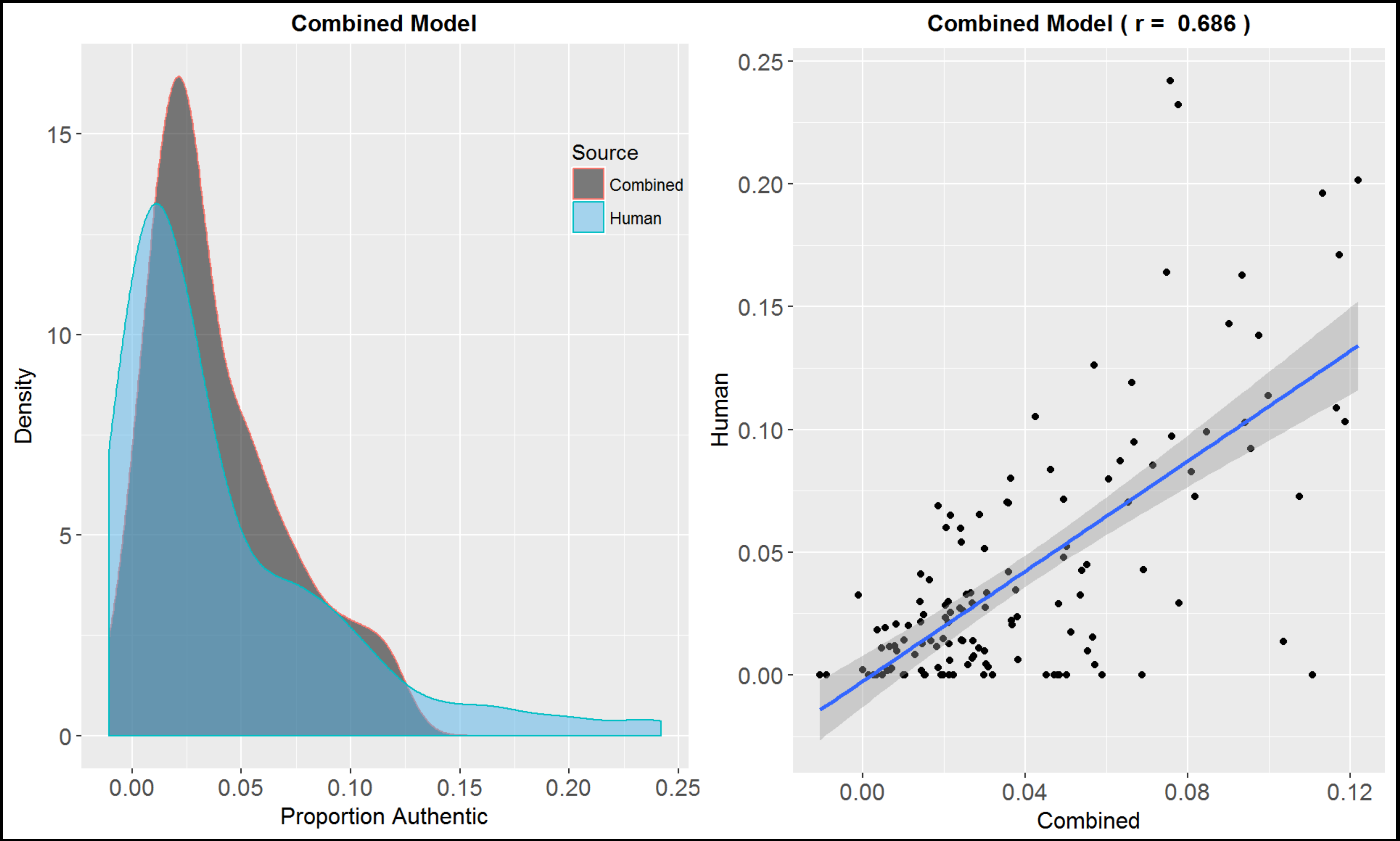
mixer (M-Audio M-Track)

1





Speech and Language Processing
Machine Learning & Validation
Data collection (132 observations from 27 classes by 14 teachers in 7 schools)
Teacher mic (Samson Airline 77)
REC
Live coding + offline refinement of codes
Teacher audio (7,663 mins total)
Gold-standard authentic question codes
Automatic speech segmentation (45k utterances)
Data for machine learning
Bing speech recognition (text transcript)
S
NP
VP
John
V
NP
hit
Det
N
the
ball
Natural language processing (word, sentence, and discourse level features)
Leave-one-teacher-out cross validation
Learn M5P regression trees for k-1 teachers (in gray)
Apply model to generate estimates for held-out teacher (in black)
Repeat until each teacher is held out once
Pool computer-estimates and compare with gold-standard codes



computer scores of authenticity correlated with human codes (r = .686)

Streamlined recording

- Teacher mic only
- Moving to lapel/smartphone

Teachers collect data

- 127 class sessions from 16 teachers in western Penn.

Utterance-level coding

- Cloud-based automatic speech recognition generates utterances

Spreadsheet-based coding

- Direct coding of utterances into using excel macros and foot pedals for audio

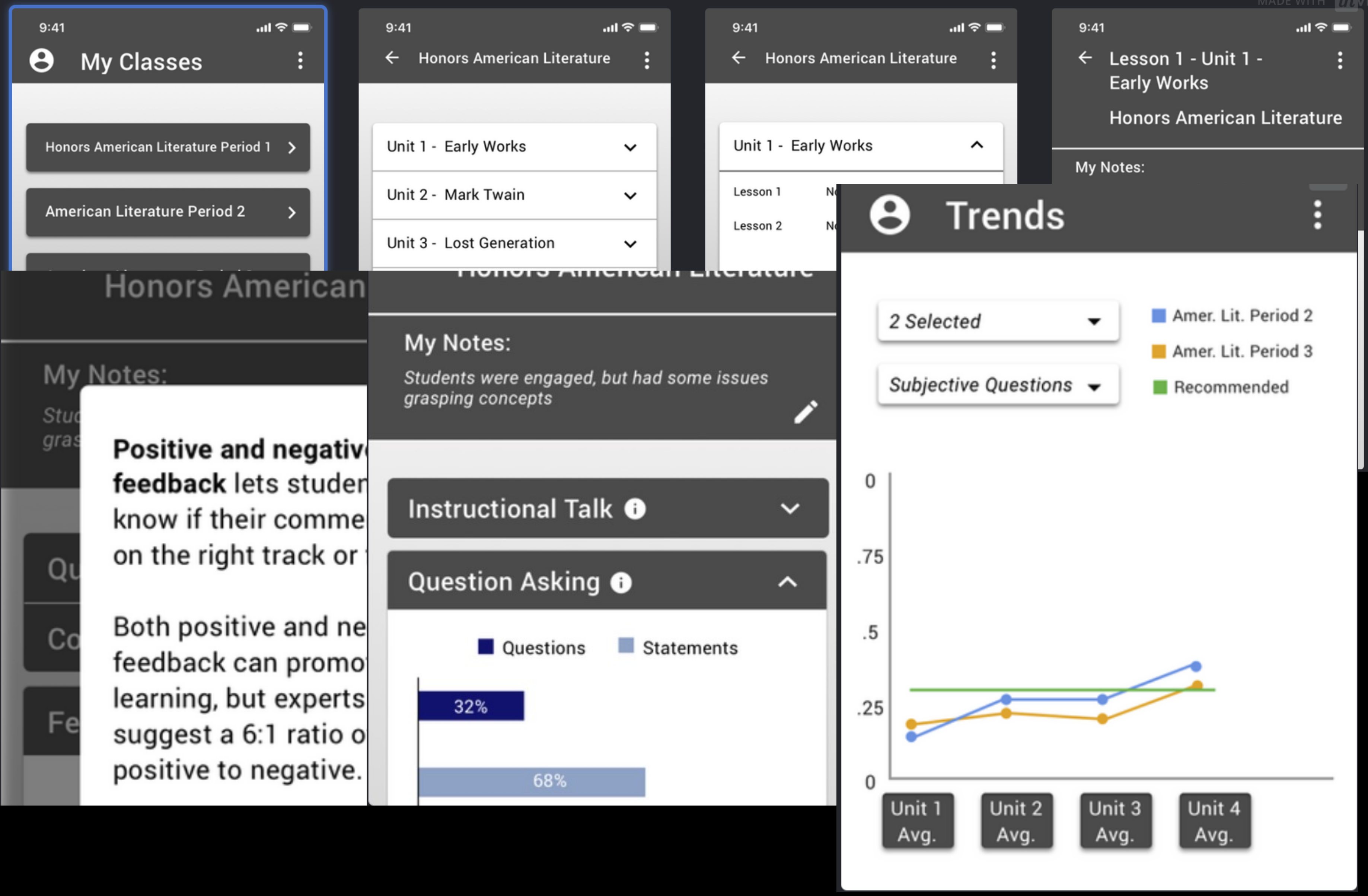
Expanded codes

- Expansive set of codes including teacher-led and transactional discourse

Multilevel modeling

- Modeling at teacher, class, and utterance levels

new approach



design of feedback app

models of spoken language can capture complex aspects of discourse in noisy environments

description

- funded by Intelligence Advanced Research Projects Activity (IARPA)
- challenge was to robustly predict psychological traits, health/well being, and job performance in the real-world from sensors alone?
- target correlation of 0.5 on a blinded sample
- do it all in 16 months

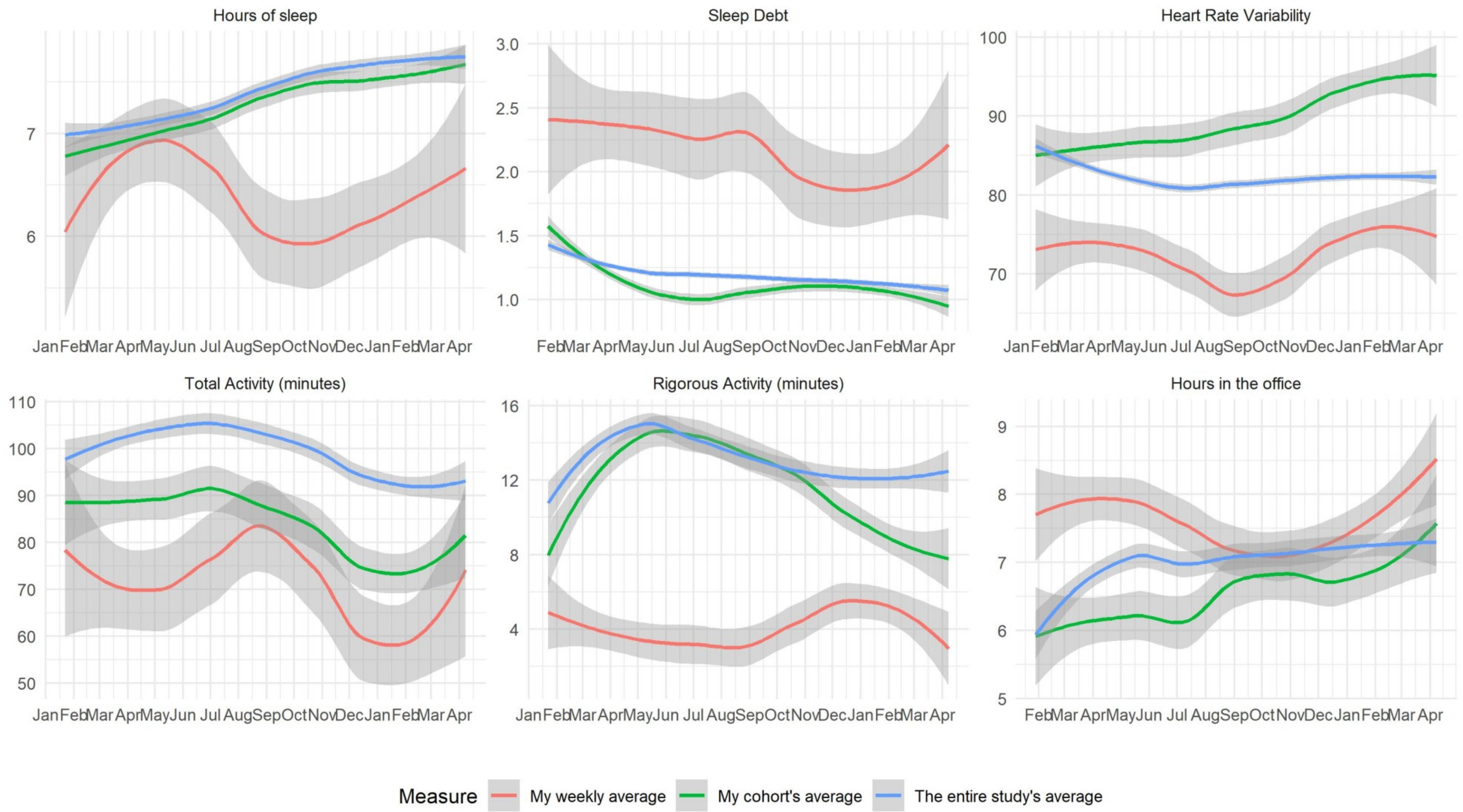
our approach

- Project Tessarae - 10 PIs from 8 universities
- collected data from 757 US information workers for 1-year
- four sensors (wearable, Beacons, phone agent, social media)

results

- modeling social, lifestyle, tech use, physiology/behavior, & context
- ensemble-based machine learning approach for robustness
- average correlation of 0.21 [0.08 to 0.41] on 14 constructs

The MOSAIC Program



These results are based on an average of 63 weeks of data

patterns of life

machine-learned, computational models can enhance assessment

- machine-learning when theory/mechanisms are sparse
- data is abundant and sufficiently complex (nonlinearities)
- models can promote change with intervention and/or reflection

tips on constructing models

- reliance on theory without being overly constrained by it
- striving for parsimony rather than chasing fads (deep learning)
- important to go beyond minimizing validation loss
- explainability, real-time applicability, fairness, & generalizability

summary

things to consider when assessing ill-defined constructs

- defining constructs – *don't really need precise definitions*
- reliability concerns – *reliability important but not a show stopper*
- quantify performance – *external sources critical*
- what is good performance? – *beyond chance probabilistic*
- how good is good enough? – *good for what purpose?*

looking into the future

- standardized testing
- game-based assessments & performance tasks
- machine-learned computational models for specific tasks

- is the future robust multimodal sensing in context?

concluding thoughts

www.colorado.edu/ics/sidney-dmello
sidney.dmello@colorado.edu