Automatic Multi-Modal Perception of Students and Classrooms

Jacob Whitehill, Worcester Polytechnic Institute
jrwhitehill@wpi.edu
Introduction
I love to teach

• Served as a volunteer math tutor for much of my adult life:
  
  • High school students in charter school (San Diego, CA)
  
  • Former gang members (Chelsea, MA)
  
  • Juvenile delinquent youth (San Diego, CA)
  
  • Adult prison inmates (Concord, MA)
I love to teach

• Taught university-level classes in computer science:
  • University of the Western Cape (South Africa)
  • Kigali Institute of Science, Technology, and Management (Rwanda)
  • University of California, San Diego (CA)
  • Worcester Polytechnic Institute (MA)
My research group

• Multi-modal machine learning for recognizing, modeling, and responding to human behavior.

• Computer vision

• Speech analysis

• Particularly interested in applications of machine learning to education.
Machine learning (ML) and deep learning (DL)

- Machine learning is a form of AI whereby intelligent behavior is learned by computer algorithms from many training examples.

- Deep learning refers to machine learning algorithms that are based on artificial neural networks.
Deep learning

• Since the past 10 years, there have been dramatic advances in machine perception due to deep learning.

AI for Education

• Multi-modal machine learning analyzes signals from different input sources, e.g.:
  • **Video** of a student at a desk
  • **Audio** of a teacher speaking in a classroom
  • **Text** written by student on an essay

• Machine learning offers many powerful opportunities to improve *educational measurement* and *feedback* to teachers and researchers.
AI for Education

• The last 10 years have seen significant growth in ML in Education for various tasks:

  • Student emotional and cognitive state recognition (Grafsgaard et al. 2013; Monkaresi et al. 2016; Woolf et al. 2010).

  • Student success prediction (Botelho et al. 2019; Gardner & Brooks 2018; Whitehill et al. 2015).

  • Automated classroom observation (Kelly et al. 2018; Ahuja et al. 2019; Owens et al. 2017).
This talk

• I will present 3 projects about ML applied to education:

  1. Automatic measurement of student engagement

  2. Exploring the relationship between thermal comfort and learning

  3. Automatic recognition of positive and negative CLASS climate

• The talk will conclude with a discussion of limitations, risks, and mitigating strategies.
Automatic measurement of student engagement

Whitehill, Serpell, Lin, Foster, and Movellan, *IEEE TAFFC 2014*
Aung & Whitehill, *IEEE FGR 2018*
Project goal

- **Student engagement** is recognized as an important outcome in its own right, not just as a correlate of learning.

- Engagement can be variously defined (Anderson et al. 2004):
  - Behavioral
  - Emotional
  - Cognitive

- How can we automatically estimate *how engaged a student appears to be* using machine learning?
Conventional methods of measuring student engagement

- Surveys and observational checklists:
  - Primacy/recency effects
  - Reluctance to give honest responses
  - Lack of temporal resolution

- Physiological sensors (EEG, GSR):
  - Obtrusive; requires physical contact
Measuring student engagement

- Our goal: develop automatic detector of student engagement using web-camera & computer vision.
  - Specifically, estimate from video how engaged the student appears to an average observer.
- Unobtrusive
- High temporal resolution.

Whitehill, Serpell, et al. 2014
Dataset

- 30 undergraduate students interacting with an iPad-based intelligent tutoring system.
Dataset

• Each student in the dataset followed the protocol:
  • **Pre-test** on a cognitive skills task ("Set")
  • 30 minutes of **training** on the task
  • **Post-test**
Our approach

- Our interdisciplinary research team annotated engagement from videos/images as:
  
  - 1 = Not engaged at all.
  
  - 2 = Nominally engaged.
  
  - 3 = Engaged.
  
  - 4 = Very engaged.
Labeling

- We labeled ~20,000 face images from the 30 participants in the dataset.

- Given images+labels, we can use machine learning to create an automatic engagement detector.
Recognition architecture

Input image

48x48 face image

Neural network

Gabor filters

Estimate of “engagement”

Whitehill, Serpell, et al. 2014
System accuracy

- Accuracy in distinguishing $E=4$ from $E=1$: 93%
System accuracy

• Accuracy on fine-grained classification tasks (e.g., “Is this student in engagement level 2?”):
  • Automated detector: 73%
  • Human-human: 72%

Whitehill, Serpell, et al. 2014
Recognition example

Whitehill, Serpell, et al. 2014
Recognition example
Recognition example
Most predictive features

- Most predictive signals of perceived low engagement:
  - Eye closure
  - Head pose:
    - Roll (in-plane rotation)
    - Pitch
Correlations with task performance

• We also investigated whether engagement is correlated with students’ test performance.

• “Engagement” was positively correlated with both pre-test and post-test:

  • corr(Engagement, Pre-test): 0.57
  • corr(Engagement, Post-test): 0.47
Correlations with task performance

• We also investigated whether engagement is correlated with students’ test performance.

• “Engagement” was positively correlated with both pre-test and post-test:

  • corr(Engagement, Pre-test): 0.57
  • corr(Engagement, Post-test): 0.47
  • corr(Pre-test, Post-test): 0.44
Correlations with task performance

• We also investigated whether engagement is correlated with students’ test performance.

• “Engagement” was positively correlated with both pre-test and post-test:

  • corr(Engagement, Pre-test): 0.57
  • corr(Engagement, Post-test): 0.47
  • corr(Pre-test, Post-test): 0.44

• In other words: to predict a student’s post-test score, the average engagement level is just as informative as the pre-test score.
Engagement recognition: summary

• Our study was a proof-of-concept that an AI system can be trained using machine learning and computer vision.

• The system’s outputs are stat. sig. correlated with human codes.

• The “machine” itself may not generalize…but the algorithm used to train the machine likely does.
Limitations

• Study was small ($n=20$ students).

• This system was trained and evaluated in a specific context: African-American college students interacting with an educational game.

• For a different population or educational setting, the system would likely need to be retrained.
Limitations

• Need to be careful drawing behavioral conclusions:
  • The face and body cues we found are correlated with human codes but may not be causal.
  • Correspondence between facial expression and emotion? (Barrett et al. 2019)
  • Facial expressions consistent across cultures? (Srinivasan and Martinez 2018)
Exploring the relationship between thermal comfort and learning

Jiang, Iandoli, Van Dessel, Liu, and Whitehill, *EDM* 2019
Introduction

Impact of indoor environmental quality (IEQ) on learning

• The indoor environment quality of where people learn, study, and work can significantly impact their cognitive performance (Horr et al. 2016; Arif et al. 2016).

• Important factors include:
  • Air quality and ventilation
  • Lighting
  • Temperature and thermal comfort

• These are likely complementary to all the important pedagogical and affective factors that impact learning.
Introduction

Thermal comfort (TC)

- Thermal regulation in schools is a problem in USA, Canada, and many other countries:

  Out of the Toronto District School Board’s (TDSB) 583 schools, 128 have full air conditioning, some have partial air conditioning, and others have no air conditioning at all. (*Global News*; 30 May 2018.)

New York Times, 4 Jan 2018

https://twitter.com/i/status/1004700614071324673
Introduction

Thermal comfort (TC)

• The condition of mind that expresses **satisfaction** with the thermal environment.
• Personal and subjective.

Thermal comfort and learning

1. Does thermal comfort significantly impact students’ learning?

• A recent national study (Goodman, et al. 2018, NBER) of 10M learners who took PSAT exam concluded that:
  • 1°F temp. increase during school days => 1% less learning in classrooms without AC.
  • No effect due to hot days in summer or on weekends.
  • Effects were strongest for students from underprivileged communities.
Introduction

Thermal comfort and learning

2. Can we optimize TC for each individual student?
Introduction

Thermal comfort and learning

3. Are students the best judges of what is an optimal temperature for their own learning?

- If yes: just let them control it.
- If no: implement automatic environmental control?
## Related Work

### Indoor environmental quality (IEQ) & Learning

<table>
<thead>
<tr>
<th></th>
<th>Light</th>
<th>Air</th>
<th>Thermal comfort</th>
<th>Noise</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dorizas, et al.[10]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sarbu &amp; Cristian.[27]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Barrett, et al.[5]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Marchand, et al.[24]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Jiang, et al.[16]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Related Work

Measuring TC

• Questionnaires
  • Self-reported scores.

• Environmental sensors:
  • PMV-PPD model (Fanger 1970) based on indoor environment parameters.

• Body sensors

• Cameras
  • Web cameras of the face (Jazizadeh & Jung 2018; Jung & Jazizadeh 2018).
Research questions

1. How does the relationship between TC and learning change over time?
   • Maybe TC doesn’t matter at first but becomes more important as time goes on?

2. How accurately can TC be measured using different sensors?
   • Infra-red (IR) face image
   • Body sensors
   • Facial expression
Experiment

Overview

- Conducted in Aug. 2018
- IRB No.: 18-0372
- Participants: 25 university students (9 females, 16 males)
- Paid: $20 gift card
- Duration: 1.5 hours
Overview

• Participants watched 3 tutorial videos.
• After each tutorial, they took a quiz and submitted a thermal comfort survey.
• Sensor measurements (cameras and body temperature sensors) were recorded during the whole experiment.
Experiment

Procedure

• 1.5 hours in total

Adaptation 21 mins → Video 1 21 mins → Video 2 21 mins → Video 3 21 mins
Experiment

Procedure

• Adaptation Session:
  • Give informed consent.
  • Neutralize the potential impact of the outside weather conditions or on activity (e.g., running to the experiment).
  • Listen to the instructions.
  • Place skin sensors on body.
**Experiment**

**Procedure**

- **Video tutorial Session:**
  - **Watching Video**: Around 10 mins
  - **Taking Quiz**: Max 5 mins
  - **Finishing Survey**: Max 5 mins
  - **Break**: 21 mins – video – quiz – survey
Experimental Conditions

• Randomly assigned to each participant
  1. Neutral to warm (25°C to 30°C)
  2. Warm to neutral (30°C to 25°C)
Experiment

Conditions

Room Temperature

- **Condition 1:** From Neutral to Warm (25°C to 30°C)
- **Condition 2:** From Warm to Neutral (30°C to 25°C)

<table>
<thead>
<tr>
<th>Time(min)</th>
<th>Adaptation</th>
<th>Tutorial 1</th>
<th>Tutorial 2</th>
<th>Tutorial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21</td>
<td>42</td>
<td>63</td>
<td>84</td>
</tr>
<tr>
<td>30°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25°C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment

Sensors/Actuators

- Room temperature sensors
- Heaters
- AC
- Skin temperature sensors
Experiment

Sensors/Actuators

- Room temperature sensors
- Heaters
- AC
- Skin temperature sensors
Experiment

Sensors/Actuators

- Room temperature sensors
- Heaters
- AC
- Skin temperature sensors

Experiment

Materials

• 3 tutorial videos (random order) on ethics and philosophy:

Materials

- Survey
  - How sleepy/alert do you feel?
  - How easy/difficult is it to concentrate?
  - …

Experiment

Data Collection

- Videos from web-camera (30 fps)
- Infrared images (1 every 21 minutes)
- Room temperature, CO₂, and relative humidity (1 measurement every minute)
- Body temperature (1 measurement every minute for each sensor)
- Each participant’s start/end times of each tutorial video, quiz, and survey
- Quiz score and survey answers
Experiment

Data Collection

• Examples
• This was an exploratory study with many sensors (IR, web-camera, skin sensors) and dependent variables (quiz performance, engagement, TC), but modest number of participants (n=25).
• Our focus was on identifying possible hypotheses for future investigation.
• As part of a 3-year NSF grant (EHR:IUSE), we will explore these hypotheses in greater depth.
Analysis

1. The impact of room temperature on thermal comfort

- Higher temperature $\Rightarrow$ lower TC (within the 25-30°C range of the experiment).
- Pearson $r = -0.32$, $p < 0.0001$
2. The impact of thermal comfort on learning
Analysis

2. The impact of thermal comfort on learning

- Inverted “U” relationship between TC and quiz scores:
- $TC^2$ vs. QuizScore: Spearman $r = -0.235, p = 0.0042$
- This corroborates a prior result by Jiang et al. 2018.
3. TC, learning, and time

- TC² vs quiz score as a function of time (estimated with linear mixed effects model):
- Effect of TC increased over time.

<table>
<thead>
<tr>
<th>Session No.</th>
<th>Pearson corr. (r)</th>
<th>p-value</th>
<th>Effect Size (Cohen’s f²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.04</td>
<td>0.67</td>
<td>0.007</td>
</tr>
<tr>
<td>2</td>
<td>-0.08</td>
<td>0.31</td>
<td>0.044</td>
</tr>
<tr>
<td>3</td>
<td>-0.2</td>
<td>0.013</td>
<td>0.308</td>
</tr>
</tbody>
</table>
4. Relationship between thermal comfort and sleepiness.

- Higher TC associated with more sleepiness (Pearson $r = 0.32$, $p = 0.0084$).
Automatically measuring thermal comfort

**Infrared cameras**

- Average face temperature from IR camera is negatively correlated with TC:
  - \( r = -0.34 \), \( p = 0.0029 \)
Infrared cameras

- Average face temperature from IR camera is negatively correlated with TC:
  - $r = -0.34$, $p = 0.0029$
  - This result is very similar to accuracy ($r=-0.29$) from Pavlin et al. 2017.
**Skin sensors**

<table>
<thead>
<tr>
<th>Position</th>
<th>Pearson corr ($r$)</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-0.273</td>
<td>0.018</td>
</tr>
<tr>
<td>K</td>
<td>-0.174</td>
<td>0.136</td>
</tr>
<tr>
<td>O</td>
<td>-0.186</td>
<td>0.11</td>
</tr>
<tr>
<td>Q</td>
<td>-0.28</td>
<td>0.015</td>
</tr>
</tbody>
</table>
Automatically measuring thermal comfort

Web camera

- Example
Automatically measuring thermal comfort

Web camera

- OpenFace 2.1.0 (open source toolkit for facial behavior analysis).
- Facial landmark detection and tracking
- Facial Action Unit detection

https://www.cs.cmu.edu/~face/facs.htm
Automatically measuring thermal comfort

**Web camera**

- **Facial actions units (AUs)**

<table>
<thead>
<tr>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU5</th>
<th>AU6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner brow raiser</td>
<td>Outer brow raiser</td>
<td>Brow Lowerer</td>
<td>Upper lid raiser</td>
<td>Cheek raiser</td>
</tr>
<tr>
<td>AU7</td>
<td>AU9</td>
<td>AU12</td>
<td>AU15</td>
<td>AU17</td>
</tr>
<tr>
<td>Lid tighten</td>
<td>Nose wrinkle</td>
<td>Lip corner puller</td>
<td>Lip corner depressor</td>
<td>Chin raiser</td>
</tr>
<tr>
<td>AU23</td>
<td>AU24</td>
<td>AU25</td>
<td>AU27</td>
<td></td>
</tr>
<tr>
<td>Lip tighten</td>
<td>Lip presser</td>
<td>Lips part</td>
<td>Mouth stretch</td>
<td></td>
</tr>
</tbody>
</table>

[https://www.cs.cmu.edu/~face/facs.htm](https://www.cs.cmu.edu/~face/facs.htm)
Automatically measuring thermal comfort

Web camera

• Individual face movements (17 AUs)
• Head Pose features: eye-lid distance, face size, and pitch, yaw, roll
• Significant predictors of TC:
  • AU06 (cheek raiser): Pearson $r = 0.244, p = 0.038$
  • Eye-lid distance: Spearman $r = -0.27, p = 0.02$
  • High TC => sleepiness?
We conducted an experiment to explore the relationship between thermal comfort, time on task, and learning.

Changing a room by just a few degrees can significantly influence students’ thermal comfort.

Students’ learning is best when TC is moderate (Goldilocks effect) — it is possible that students themselves are not the best judge of optimal TC for learning.

The impact of TC on learning grew over time.

IR face images, body sensors, and facial expressions can each be used to predict thermal comfort with low/moderate correlation ($r$ around 0.24 to 0.34).
Limitations/Future Work

- Modest sample size ($n=25$ participants).
- Since this was largely exploratory work, we tested lots of hypotheses (many AUs, Pearson and Spearman correlations, TC & $TC^2$, different prediction models).
  - Hence, $p$-values may be underestimated.
- The duration (~1.5 hours) is much shorter than a typical school day.
- We are planning a new study to explore how wrist-based cooling+heating devices can provide a stimulus to improve TC.
Automatic recognition of CLASS positive and negative climate

Ramakrishnan, Ottmar, LoCasale-Crouch, and Whitehill, *IEEE FG 2019*
Zylich and Whitehill, *IEEE ICASSP 2020*
Classroom interactions
Classroom interactions

• The quality of classroom interpersonal dynamics both predicts and impacts students’ downstream academic and behavioral outcomes (Burchinal et al. 2010; Ponitz et al. 2009).

• Key characteristics include:
  • Emotional support
  • Rich language exchanges between teachers & students
  • Collaborative learning opportunities.
  • …
Classroom observation

• In US schools, teachers’ classrooms are typically observed 1-2x per year.

• Observation can be conducted either live or on videorecorded classroom sessions (typically 45-60min).

• Classroom observations & feedback can facilitate:
  • Professional development
  • Accountability
  • Educational research
Classroom observation

• To measure classroom interactions objectively, researchers have devised observational protocols, e.g.:
  • Instructional Quality Assessment (IQA; Junker et al. 2005).
  • Protocol for Language Arts Teaching Observations (PLATO; Grossman 2009).
  • Classroom Assessment Scoring System (CLASS; Pianta, LaParo, & Hamre 2008).
Classroom Assessment Scoring System (CLASS)

• With the CLASS, coders rate (1-7 scale) the classroom quality among different dimensions within several domains.
  
• Emotional support:
  • Positive climate
  • Negative climate
  • Teacher sensitivity
  • Regard for child perspectives

• Engaged support for learning:
  • Facilitation of learning and development
  • Quality of feedback
  • Language modeling
## Classroom Assessment Scoring System (CLASS)

<table>
<thead>
<tr>
<th>CLASS Emotional Support Dimensions</th>
<th>Examples of Behavioral Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indicators</strong></td>
<td><strong>Examples of Behavioral Markers</strong></td>
</tr>
<tr>
<td><strong>POSITIVE CLIMATE</strong></td>
<td></td>
</tr>
<tr>
<td>Relationships</td>
<td>Physical proximity, matched affect</td>
</tr>
<tr>
<td>Positive Affect</td>
<td>Smiling, laughter</td>
</tr>
<tr>
<td>Positive Communication</td>
<td>Verbal affection, physical affection</td>
</tr>
<tr>
<td>Respect</td>
<td>Eye contact, warm voice</td>
</tr>
<tr>
<td><strong>NEGATIVE CLIMATE</strong></td>
<td></td>
</tr>
<tr>
<td>Negative Affect</td>
<td>Irritability, harsh voice, anger</td>
</tr>
<tr>
<td>Punitive Control</td>
<td>Yelling, threats</td>
</tr>
<tr>
<td>Sarcasm/Disrespect</td>
<td>Sarcastic voice, humiliation</td>
</tr>
<tr>
<td>Severe Negativity</td>
<td>Victimization, bullying</td>
</tr>
<tr>
<td><strong>TEACHER SENSITIVITY</strong></td>
<td></td>
</tr>
<tr>
<td>Awareness</td>
<td>Notices difficulties, anticipates problems</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>Acknowledges emotions, provides comfort</td>
</tr>
<tr>
<td>Addresses Problems</td>
<td>Helps resolve problems</td>
</tr>
<tr>
<td>Student Comfort</td>
<td>Students seek support and guidance from teacher</td>
</tr>
<tr>
<td><strong>REGARD FOR STUDENT/CHILD PERSPECTIVES</strong></td>
<td></td>
</tr>
<tr>
<td>Flexibility and Student Focus</td>
<td>Shows flexibility, incorporates students' ideas</td>
</tr>
<tr>
<td>Support for Autonomy and Leadership</td>
<td>Allows choice, gives students responsibilities</td>
</tr>
<tr>
<td>Student Expression</td>
<td>Encourages student talk</td>
</tr>
<tr>
<td>Restriction of Movement</td>
<td>Allows movement, is not rigid</td>
</tr>
</tbody>
</table>
Limitations of human coding

• While manual coding of classrooms observations is useful, it has important limitations:

• **Cost:** Careful coding of videos is very laborious and can take many person-hours (Berlin, 2012).

• **Reliability:** Scores can vary widely across coders, and they are prone to primacy/recency effects.

• **Specificity:** Scores are usually assigned to long segments (15-20 min), not specific people or interactions.
Potential for AI-based classroom observation

• Al-based classroom observation, as enabled by machine learning, can partially overcome some of these challenges.

• Machines are particularly adept at perceiving “low-level” attributes of the classroom (e.g., who is where, when, doing what, with whom).

• What kinds of new analyses and feedback mechanisms could result?
Our vision for automated classroom observation

A. Observation measure

Positive Climate (PC)

Aggregate Score: 5/7

B. Score

C. Temporal Heatmap (blue=low, red=high)

D. Key Moments

E. Key Participants
Our vision for automated classroom observation

Frame-by-Frame Interactions

Tracking Changes over Time

Social Network of Eye Gaze

Social Network of Speech

Kent, Happy

Ashley, Sad.

Gaze Interaction

Ashley, Sad

Ashley, Happy

Trent

Alice

Teacher

61 Gazes
12 Gazes
32 Gazes
20 Gazes
52 Gazes
61 Gazes

2 Utterances
3 Utterances
13 Utterances
6 Utterances
29 Utterances
29 Utterances
Multi-modal machine learning approach

• For the past 3 years, our WPI+UVA team has investigated how to develop multi-modal machine learning architectures to estimate CLASS scores automatically.

• We are focusing on what are arguably the “easiest” dimensions to automate: positive climate (PC) and negative climate (NC).

• **PC**: evidence of mutual respect and support for learning.

• **NC**: evidence of overt negativity in the classroom.
Multi-modal machine learning approach

- Our system uses two kinds of information:
  1. **Visual:**
     - Facial expression of every student and teachers.
     - Analysis of the entire visual scene.

![Multi-modal machine learning approach](image-url)
Multi-modal machine learning approach

- Our system uses two kinds of information:

2. **Auditory**:
   - Analysis of Fourier features of classroom speech.
   - Identifying key phrases ("thank you", "good job", etc.).
Context

• We are focusing on estimating CLASS scores for pre-schools (toddlers):
  • Extremely cluttered visual scene
  • Noisy, with multiple overlapping voices
  • Most speech is from teachers
  • Students and teachers may leave & re-enter the scene
Datasets

• For training and testing our system, we use:

  • UVA Toddler dataset: ~200 CLASS-coded videos of pre-school classrooms.

  • Measures of Effective Teaching (MET) dataset: ~10,000 CLASS-coded videos of middle-school classrooms.
Results: UVA Toddler

- Compared to human codes, our automated system achieves an accuracy (Pearson r, on videos not used for training):
  - Positive climate: 0.59
  - Negative climate: 0.66
Results: UVA Toddler

- Compared to human codes, our automated system achieves an accuracy (Pearson $r$, on videos not used for training):
  - Positive climate: 0.59
  - Negative climate: 0.66
- These correlations are similar to inter-coder reliability; however, the machine sometimes makes mistakes that a human would never make.
Example outputs

- Estimated smile scores of teachers, students:

![Example outputs](image-url)
Example outputs

- Estimated eye gazes:

Aung, Ramakrishnan, and Whitehill (2018)
Results: MET

• Due to privacy restrictions enforced by the University of Michigan, we could only train a relatively simple audio-based system for CLASS score estimation.

• How much signal does audio alone give us about CLASS scores?

• Since MET is huge, we can examine how accuracy improves as we train on bigger datasets.
Results: MET

Number of Decision Trees vs Positive and Negative Climate

Pearson Correlation Score

- **Pos_Climate**
- **Neg_Climate**
Automated CLASS score estimation: summary

- We have developed a proof-of-concept automated CLASS score estimation system.

- The system’s outputs for PC and NC are stat. sig. correlated with human codes.

- The system can observe individual people and estimate low-level cues (e.g., utterances, eye gazes) with high temporal resolution.
Automated CLASS score estimation: next steps

• How to devise new teacher training experiences based on automated feedback?

• PC and NC are semantically “low-level” dimensions; need to explore how to estimate higher-level dimensions (e.g., language modeling).
Privacy, accuracy, and fairness
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?

2. Machines can be highly accurate on average but fail miserably on individual people/moments, or on specific subgroups.
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?
2. Machines can be highly accurate on average but fail miserably on individual people/moments, or on specific subgroups.
3. AI-based systems are often highly confident in their predictions, even when they are wrong.
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?

2. Machines can be highly accurate on average but fail miserably on individual people/moments, or on specific subgroups.

3. AI-based systems are often highly confident in their predictions, even when they are wrong.

4. Machine learning is typically used to maximize the correlation between predictions and human-coded scores, but the resulting model may not be causal.
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?

2. Machines can be highly accurate on average but fail miserably on individual people/moments, or on specific subgroups.

3. AI-based systems are often highly confident in their predictions, even when they are wrong.

4. Machine learning is typically used to maximize the correlation between predictions and human-coded scores, but the resulting model may not be causal.

5. Machine learning systems depend on the quality of the datasets they are trained on.
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?

2. Machines can be highly accurate on average but fail miserably on individual people/moments, or on specific subgroups.

3. AI-based systems are often highly confident in their predictions, even when they are wrong.

4. Machine learning is typically used to maximize the correlation between predictions and human-coded scores, but the resulting model may not be causal.

5. Machine learning systems depend on the quality of the datasets they are trained on.

6. Complex deep learning models contain millions (or more!) of parameters, and are very difficult to interpret.
Privacy, accuracy, and fairness

1. Do we want large-scale observation of every teacher and every student at every moment in time?

2. Machines can be highly accurate on average but fail miserably on individual people/moments, or on specific subgroups.

3. AI-based systems are often highly confident in their predictions, even when they are wrong.

4. Machine learning is typically used to maximize the correlation between predictions and human-coded scores, but the resulting model may not be causal.

5. Machine learning systems depend on the quality of the datasets they are trained on.

6. Complex deep learning models contain millions (or more!) of parameters, and are very difficult to interpret.

7. Machine learning can enable a new generation of scientific instruments, but we must be careful in interpreting their outputs.
Automated classifiers for educational measurement: a methodological risk

Whitehill & Ramakrishnan, *ICML 2019*
Machine learning to advance basic science

• Machine perception can advance basic science in:
  • Psychology
  • Education
  • Medicine

• ...by providing automatic classifiers as new scientific instruments, e.g.:
  • Automatic stress detectors from wrist monitors instead of questionnaires.
  • Facial action unit detectors from video instead of electromyography.
  • Student engagement detectors from video instead of observational protocols.
Correlation study

• Suppose a researcher wishes to measure the relationship between two constructs $U$ and $V$, e.g.:
  • $U = \text{stress}$
  • $V = \text{academic performance}$.

• Standard methodology:
  • Use a standard measurement tool (e.g., survey, observational protocol) to estimate the values of $U$ and $V$ from a sample of $n$ participants.
  • We can represent these measurements as vectors $u$ and $v$.
  • The cosine of the angle between them determines their correlation $r$.

\[ r = \cos \angle(u, v) \]
Correlation study

• But what if the researcher instead uses an **automatic stress detector** \( d \) whose accuracy is known to be \( q \)?

• Instead of \( u \), the researcher obtains a vector \( \hat{u} \).

• What kind of spurious deductions about the correlation between \( U \) and \( V \) could result?
Trivariate correlation

- Suppose $u$ and $v$ are ground-truth values of $U$ and $V$.

- The correlation between $u$ and $v$ is $r = \cos(105^\circ) = -.259$. 

\[
\begin{array}{c}
\text{Trivariate correlation} \\
\includegraphics[width=0.5\textwidth]{correlation_diagram}
\end{array}
\]
• Using a detector $d$, the researcher might obtain $\hat{u}$, whose correlation with $u$ is $q$.

• The correlation between $\hat{u}$ and $v$ is $\cos(135^\circ) = -0.707$ — much larger than, but same sign as, the ground-truth correlation.
Trivariate correlation

- But they might also obtain vector \( \hat{u}' \), whose correlation with \( u \) is also \( q \).

- The correlation between \( \hat{u}' \) and \( v \) is \( \cos(75^\circ) = +.259 \) — this is the \textbf{opposite sign} as the ground-truth correlation.

- We call this a \textbf{false correlation}. 

WPI
Main results

1. We derive a formula $h$ for the probability of obtaining a false correlation from an automatic detector $d$.

2. The probability of a false correlation can be substantial even when $n$ is moderately large (100+), and even if the correlation is statistically significant.

3. We present 2 case studies, based on recent educational and behavioral science research papers, to put these results into context.
Case study: Student engagement vs. cognitive task performance

• Whitehill et al. 2014 measured student engagement using (1) observational protocol and (2) automatic engagement detector $d$ ($q=0.50$).

• Using hand-coded labels, $\text{corr}(U, V)$ was estimated as $r=0.37$.

• Given $n$, $q$, $r$, what is probability of false correlation from $d$?
Talk summary
Talk summary

- Machine learning-based artificial intelligence offers many powerful ways to make educational measurement:
  - Cheaper
  - Faster
  - More reliable
  - More specific
- In turn, this can enable richer forms of teacher training and professional development.
Talk summary

- But automated classroom observation has big risks that are magnified when conducted at large-scale:
  - Privacy
  - Algorithmic bias against subgroups
  - Ill-founded conclusions
Talk summary

• But automated classroom observation has big risks that are magnified when conducted at large-scale:
  • Privacy
  • Algorithmic bias against subgroups
  • Ill-founded conclusions

• Important to understand both opportunities and risks so that stakeholders can hold machine researchers accountable to improve both educational quality and equity.
Thank you