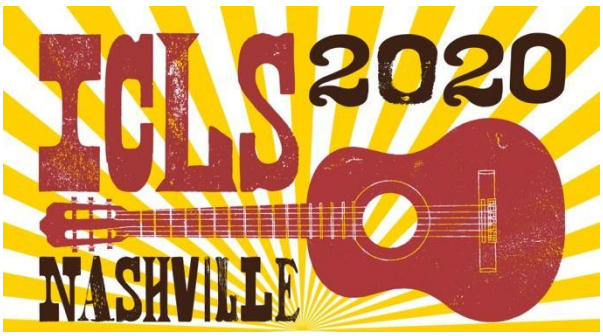

Workshop Application

1. Basic information on proposed pre-conference event

Title: Analyzing Learning with Speech Analytics and Computer Vision Methods: Technologies, Principles, and Ethics

Duration

Full day (June 19, 2010)	
Half day (June 20, 2020)	X

2. Organizer Information

Please fill the table with first name, last name, email, institution and Country of institution for the first six organizers of your pre- conference event.

First Name	Last Name	email	Institution	Country
Elizabeth	Dyer	edyer@mtsu.edu	Middle Tennessee State University	USA
Cynthia	D'Angelo	cdangelo@illinois.edu	University of Illinois, Urbana Champaign	USA
Nigel	Bosch	pnb@illinois.edu	University of Illinois, Urbana Champaign	USA
Stina	Krist	ckrist@illinois.edu	University of Illinois, Urbana Champaign	USA
Joshua	Rosenberg	jmrosenberg@utk.edu	University of Tennessee, Knoxville	USA

A maximum of 3 organizers will be able to attend the workshop—please indicate those 3 organizers who will be exempt from the pre-conference event registration fee.

First Name	Last Name	email	Institution	Country
Elizabeth	Dyer	edyer@mtsu.edu	Middle Tennessee State University	USA
Cynthia	D'Angelo	cdangelo@illinois.edu	University of Illinois, Urbana Champaign	USA
Joshua	Rosenberg	jmrosenberg@utk.edu	University of Tennessee, Knoxville	USA

Pre- conference event description

1. Please describe in a maximum of 2000 words (including references) the event you are proposing. The description should include (1) the theme and goals of the event, (2) the theoretical background, (3) the relevance to the field and conference theme, and (4) the expected outcomes and contributions of the event.

Theme, Goals, and Expected Outcomes or Contributions

We propose a half-day workshop about video and audio data collection methods that allow researchers to effectively use emerging analytical methods that leverage speech analytics and computer vision techniques, in combination with human-focused analysis (e.g. qualitative analysis). The main goals for participants attending the workshop are to:

1. Become familiar with innovative computational methods (e.g., computer vision and speech analytics) that can be used directly with audio and video data, and consider how computational methods can be used with human-focused analysis to develop new theory in the learning sciences.
2. Understand which features of audio and video data have a large influence on whether computational methods can be applied successfully.
3. Develop principles and strategies for collecting audio and video data in learning environments that increases the successful application of computational methods, including equipment positioning, recording formats and codecs, and equipment features or specifications.
4. Consider ethical implications of using innovative computational methods, both in terms of ethics of conducting research with these methods and potential uses of these methods for education practice and policy.
5. Contribute to a collective methodological research agenda and goals for future development of existing computer vision and speech analytics methods for learning sciences research.

The primary outcome of the workshop is for participants to be able to make informed decisions about collecting audio and video data of learning that will make it possible to use computational methods in analysis. The session will also produce a methodological research agenda for improving the computational methods in their applications to research questions and data used in the learning sciences.

Relevance to the Field and Conference Theme

The learning sciences has a long history of using video and audio to examine processes of human interaction that unfold over time (Goldman, Zahn, & Derry, 2014). Video enables researchers to capture longitudinal data on processes that unfold over time and at multiple time scales, leading to analyses that consider connections between micro- and macro-level phenomena. Video and audio data, in conjunction with multi-modal records of activity, continue to be a central data source in learning sciences, particularly for examining social processes of learning and development.

As capturing and storing video data becomes increasingly accessible and cost-effective, video is emerging as a dominant source of "big data" in the social sciences. Large video databases allow researchers to see social phenomena first hand and provide both breadth in timespan (footage that spans weeks or months of activity) and detail (a rich moment-to-moment interactional and spatial record; Goldman et al., 2014).

Despite the promise and opportunity to use video data in new ways, analytic methods and tools for video have lagged behind innovations in data collection methods, data analytics, and visualization. Video research often relies on processes developed for text-based data (e.g., creating and analyzing transcripts), essentially hiding the temporal and visuospatial dimensions of the data. These traditional analysis methods limit the ability to apply humans' sophisticated visual processing, such as tracking movement over time or seeing relationships in spatially-aligned data points.

Recent advances in computational methods (e.g., computer vision, speech analytics) provide exciting new opportunities to improve the analysis of learning in video and audio data, particularly in large datasets. Some examples of these methods include automated detection of body positioning, emotion, gaze, collaboration, tone, speakers, and prosody. However, because these methods rely on computational power, which differs from human interpretive power, they have different requirements of the data quality and quantity. In the case of speech analytics, computers are able to do many things, but if the wrong type of audio data is collected (e.g., using a lapel mic to try to capture whole class audio), the computational methods are limited in how well they can interpret the data (Richey, D'Angelo, Alozie, Bratt, & Shriberg, 2016). Additionally, speech analytics methods benefit from high resolution audio that may be almost indistinguishable for a human. With computer vision methods, it can be difficult for computers to identify a person over time if they leave the frame at some point in the video (Wu et al., 2019). Each of these examples represents concerns relating to the quality of the audio and video data that are unique to their use with computational methods. As a result, there are new considerations and principles for collecting video and audio data that can be successfully used with new computational methods.

We, along with other scholars, argue that computational methods are most powerful when integrated with human-conducted analysis and decision-making (Baker, 2016; Berland,

Baker, & Blikstein, 2014; Nelson, 2017). These arguments come from a concern over losing the richness and complexity inherent in learning for the sake of convenience and scalability. Additionally, they recognize that humans and computers often have different analytical strengths. For example, Baker (2016) argues for relying on the computational system simply for reporting relevant, low-inference information and patterns, which humans can use for higher-inference analysis to guide future action. We believe that these computational methods provide an opportunity for greater methodological interdisciplinary when they are used in a methodological framework that combines computer- and human-focused analysis, such as computational grounded theory (Nelson, 2017).

Theoretical Background

Learning environments are complex social systems in which learning—shifts in knowledge, its collective use, and the related patterns of interaction that demonstrate knowledge development in use—is an emergent outcome. Developing theory about learning requires understanding how interactive (i.e., social and spatial) aspects of classrooms are integral parts of student learning. For example, aspects such as the nature of collaboration, use of gesture and embodiment, the nuances of discursive tone and prosody, and student positional identities are important for understanding learning (Esmonde, 2009; Roth, 2001). This work has demonstrated the need for research methodologies to capture and represent the complexity and nuance in social and spatial aspects of learning. As such, researchers have consistently argued that video and audio data are especially well-suited to capture the visuospatial and acoustic features of interactive processes.

Current research methodologies require Herculean efforts to conduct analyses that simultaneously attend to complexity *and* nuance at a large scale, especially with video and audio data. There are strong qualitative traditions that actively attend to—and even prioritize—visuospatial and/or acoustic features (e.g., Jordan & Henderson, 1995), but these methods are incredibly arduous and time-consuming, making it all-but-impossible to carry out more than a few rich case studies. For example, qualitative studies that look across multiple contexts (e.g., comparing across 100 classrooms) and long time scales (e.g., tracking changes across multiple school years) are incredibly rare. In practice, video data are often reduced to text: transcripts of words spoken, which sometimes include meta-discursive markers or descriptions of gesture. This is a problem, as text is a poor representational form for capturing and communicating visual, spatial, and acoustic dynamics. However, the small repertoire of alternative representational practices for analysis reflected in the literature (e.g., multimodal transcription; Bezemer & Mavers, 2011) are incredibly time-consuming. These challenges to analyzing visuospatial and acoustic aspects of video are partly due to human limitations: people cannot simultaneously attend to all the multimodal dimensions of video and audio data systematically or recognize patterns in these dimensions, even with small data corpuses or a focused microanalysis. As a consequence of these challenges, we need new methodologies for analyzing the social and visuospatial dimensions of learning in video and audio data, especially with the potential to do so at scale.

Computational methods have shown promise for modeling and investigating complex phenomena with large corpora of data, including educational phenomena (Berland et al., 2014). For example, analytic techniques such as vector-space models, topic models, and deep learning/neural networks have all been applied meaningfully to educational research.

Importantly, advances in applying these models to educational data sources show their potential for increasing coding efficiency (e.g., Liu et al., 2016), making analysis of large datasets more feasible; and they can be used to detect change over time (e.g., Sherin, 2013), making longitudinal analyses more feasible.

Recent advances in computer vision, coupled with existing speech analytics methods, make it feasible to identify theoretically and practically important features from video in ways that preserve the complexity and nuance that draws educational researchers to audiovisual data—particularly with respect to visuospatial and acoustic features of learning. As an example, computer vision techniques have advanced to the extent that it is possible to use 2D cameras to identify body positioning for multiple people in real time (*OpenPose*; Cao et al., 2017). OpenPose estimates the position of up to 135 key skeletal points (e.g., location of each ankle, finger, top of head, etc.) for individuals in still images and videos. It is robust to partial occlusion, which is key for applications to many learning environments—for example, students might sit behind desks or be partially hidden from view by other students in front of them. OpenPose also produces visual representation of the skeletal points overlaid on video, as shown below.

Figure 1. Automatic detection of body positioning and visual overlay produced using *OpenPose*

Similarly, previous research in speech analytics has developed successful feature extraction techniques for a wide variety of acoustic features, including detecting speech activity and a variety of spectral, temporal, and prosodic features (e.g., Boersma, 2002; Ghosh, Tsiartas, & Narayanan, 2011). These acoustic features have also been shown to predict the quality of small-group collaborations in mathematics problem-solving using exploratory computational methods (D'Angelo et al., 2019).

References

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, *19*(1–2), 205–220.

Bezemer, J., & Mavers, D. (2011). Multimodal transcription as academic practice: A social semiotic perspective. *International Journal of Social Research Methodology*, *14*(3), 191–206.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.

D'Angelo, C. M., Smith, J., Alozie, N., Tsiartas, A., Richey, C., & Bratt, H. (2019). Mapping Individual to Group Level Collaboration Indicators Using Speech Data. *Proceedings of the 13th International Conference on Computer Supported Collaborative Learning (CSCL) 2019*, *2*, 628–631. Lyon, France: International Society of the Learning Sciences.

Esmonde, I. (2009). Explanations in mathematics classrooms: A discourse analysis. *Canadian Journal of Science, Mathematics and Technology Education*, *9*(2), 86–99.

Ghosh, P. K., Tsiartas, A., & Narayanan, S. (2011). Robust Voice Activity Detection Using Long-Term Signal Variability. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(3), 600–613.

Goldman, R., Zahn, C., & Derry, S. J. (2014). Frontiers of Digital Video Research in the Learning Sciences: Mapping the Terrain. In K. R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 213–232). New York, NY: Cambridge University Press.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The Journal of the Learning Sciences*, *4*(1), 39–103.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Nelson, L. K. (2017). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*.

Richey, C., D'Angelo, C., Alozie, N., Bratt, H., & Shriberg, E. (2016, September 8). *The SRI Speech-Based Collaborative Learning Corpus*. 1550–1554.

Roth, W. M. (2001). Gestures: Their role in teaching and learning. *Review of Educational Research*, *71*(3), 365–392.

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences*, *22*(4), 600–638.

Wu, D., Zheng, S.-J., Zhang, X.-P., Yuan, C.-A., Cheng, F., Zhao, Y., … Huang, D.-S. (2019). Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, *337*, 354–371.

2. Please describe in a maximum of 1000 words an overview of the activities that are planned for the workshop described above.

Activity Outline

- Introduction to the workshop & Getting to know the participants (45 minutes)
- Overview of existing computer vision and speech analytics methods (30 minutes)
- Develop research questions through radical innovation (30 minutes)
- Principles for audiovisual data collection (30 minutes)
- Applying principles for data collection and use (30 minutes)
- Ethical considerations (30 minutes)

- Constructing a Methodological Research Agenda (45 minutes)

Activities described below.

Introduction

Goal: Help us and participants get to know one another and our interests better, as well as communicate what to expect in the workshop.

In this activity we will introduce the workshop organizers, the goals for the session, and an outline of activities. Then we will ask participants discuss the following in small groups:

- What research questions and/or constructs are you most interested in studying with audio or video data?
- What research questions and/or constructs do you wish you could study, but it isn't feasible?

Each group will also be assigned to discuss one of the following:

- What methods do you use/plan/like to use for studying those things in audio or video data? How would you diagram your analysis workflow?
- What challenges do you have for using the methods you've selected to answer research questions and/or analyze particular constructs?
- What ethical considerations or issues do you think are important when doing research with audio or video data?

Groups will write the ideas shared on post-its and place on the wall/posters.

Overview of existing methods

Goal: Introduce participants to the wide variety of existing methods and show a few examples of their use in analysis workflows.

We'll share an overview of the approach to integrating human and computational methods (i.e., computational grounded theory), computational techniques (e.g., *OpenPose*, *Praat*, *OpenSMILE, OpenFace*), what constructs they might be useful for analyzing, how to access them, and what output they produce. We will provide a summary handout, and ask participants to share any additional techniques.

We will then share two examples of using these techniques that show the process of running the techniques, output, and how their use modified analysis workflows.

Finally, we will ask groups to discuss the implications of the tools in relation to the assigned question topic from the first activity (i.e., analysis workflows, analytical challenges, ethical considerations) and add these new implications to the post-it wall/posters.

Develop research questions

Goal: Engage in radical innovation to imagine research questions, which benefit from integrating human- and computer-focused methods, that they may not have considered or discarded previously.

Participants will work individually or in groups to brainstorm new research questions or revise the research questions they wrote down previously to leverage the analytical

techniques presented in the last activity. They will also be asked to write down new questions and concerns about using the techniques, which we will use to guide future workshop activities.

Principles for data collection

Goal: Share considerations and guidelines for collecting audio and video data that are specific to using computational methods.

We will present information in the table below and show examples of how the data features influence the techniques' output. Summary handouts will be provided.

	Features of data	**Guidelines and considerations for data collection**
Speech analytics	● Resolution - beyond what a human is able to easily notice ● Background noise - beyond what a human can filter out ● Capture different volumes - differences in volume of different speech are analytically useful to capture	● **Save original/raw files** ● WAV audio format at 24 bits ● Use the recording equipment that captures the type of audio you are interested in (e.g., close-talking mics for single speakers, table mics for small groups, microphone arrays for larger groups) ● Utilize separate audio sources for individuals or groups when possible ● Don't use features like auto leveling/equalizing on equipment (or in pre-processing)
Computer vision	● Resolution & bitrate - computers need more "data points" than humans to identify objects ● Occlusion & movement out of field of view - computers struggle to re-identify or infer position during frames when objects are not visible ● Contrast - differences in brightness and hue help computers, and techniques are typically less effective for people of color	● **Save original/raw files** ● Use camera angles to reduce occlusion (e.g. high positions) ● Use wide enough fields of view to reduce the need for re-identification (e.g. wide angle lenses) ● Use higher resolutions/bit rates when possible ● Place cameras to maximize the size of objects in the field of view ● Position cameras to increase contrast (e.g., avoid positions with strong backlighting)

Applying principles

Goal: Consider how to apply these principles to their interests, future research studies, or existing data.

In small groups, participants will discuss the implications of the previous activity for data collection to answer their research questions (including planned future studies), or use with existing data. Workshop organizers will answer questions that are raised.

Ethical considerations

Goal: Consider the ethical aspects of using these methods, especially with the increasing role of evaluation and surveillance in educational contexts.

Participants will brainstorm ethical considerations and questions in small groups, writing ideas on post-it notes. Then we will facilitate whole-group discussion. We anticipate that IRB, data reuse, how to handle non-consenting participants, and dehumanization are likely topics of discussion.

Methodological research agenda

Goal: Summarize key ideas from the workshop and develop new directions for the field.

We will share general take-aways and highlight the resources provided in the workshop. Depending on participants' interests, we will either hold a whole-group Q&A (if participants need more time to ask questions about the methods and data collection) or ask the group to discuss the methodological research they would like to see done for using these methods with learning sciences data. Potential research agenda questions include:

- What would make these computational methods most useful for you?
- What data collection guidelines are most prohibitive for applications to learning sciences research (and are worth making the computational methods better to address them)?

3. What is the maximum participants to be accepted to your pre-conference event?

40

3. What is the minimum participants to conduct the pre-conference event (if greater than 6).

6

5. Which of the four conference strands does your proposed workshop/event fit into? (*Teaching & Teacher Learning; Learning and Identity; Design; or Scale*) If it is a cross-cutting workshop/event, please just write "*Cross-Cutting*." You will have to pick one strand for upload, but we will note the nature of your workshop and make sure it gets reviewed appropriately.

Scale

Call for participants

Please use up to 500 words to announce your event. This will be posted with event and organizer information on the conference website to advertise the event.

This half-day workshop focuses on video and audio data collection methods that allow researchers to effectively use emerging computer-focused analytical methods (e.g., speech analytics and computer vision techniques) in combination with human-focused analysis (e.g., qualitative analysis). Video and audio recordings are an increasingly common data source for examining the complexities and nuances of learning in situ. To date, analysis of video and audio data of learning has been unable to fully leverage computational methods that take

advantage of this richness, especially with visuospatial and acoustic features (as opposed to textual extractions; e.g., transcripts).

Recent advances in computer vision, coupled with existing speech analytics methods, make it feasible to identify theoretically and practically important features from video that matter for examinations of learning. Additionally, these computational methods for video and audio data are likely to be most powerful when integrated with human-conducted analysis and decision-making, such as the computational grounded theory methodological framework. However, these new computational methods require different technical specifications for video and audio data than human-focused analysis, many of which must be decided and set before recording occurs.

In this workshop, we will share new principles for collecting audio and video data so that they can be used with innovative computational methods. Specifically, participants in this workshop will:

1. Become familiar with innovative computational methods (e.g., computer vision and speech analytics) that can be used directly with audio and video data (e.g., OpenPose: automated detection of body positioning in video), and consider how computational methods can be used with human-focused analysis to develop new theory in the learning sciences.
2. Understand which features of audio and video data have a large influence on whether computational methods can be applied successfully.
3. Develop principles and strategies for collecting audio and video data in learning environments that increases the successful application of computational methods, including equipment positioning, recording formats and codecs, and equipment features or specifications.
4. Consider ethical implications of using innovative computational methods, both in terms of ethics of conducting research with these methods and potential uses of these methods for education practice and policy.
5. Contribute to a collective methodological research agenda and goals for future development of existing computer vision and speech analytics methods for learning sciences research.

WORKSHOP PROPOSAL DEADLINE: DECEMBER 15, 11:59pm CENTRAL
Submit at: https://icls2020.exordo.com