Atlanta
COMPUTATIONAL SOCIAL SCIENCE
Workshop 2015

ORGANIZERS

Dr. Tom Clark, Emory University
Dr. Scott Crossley, Georgia State University
Dr. Munmun De Choudhury, Georgia Institute of Technology
Dr. Jacob Eisenstein, Georgia Institute of Technology
Dr. Eric Gilbert, Georgia Institute of Technology
Dr. Adam Glynn, Emory University
Dr. Benjamin Miller, Georgia State University
Dr. Jeffrey Staton, Emory University

THANK YOU TO OUR 2015 SPONSORS

Yik Yak
Emory College of Arts & Science: Institute for Quantitative Theory and Methods
Dr. Kelly Stout, University Research Services and Administration, Georgia State University
YAHOO! Labs

SCHEDULE

11:00-11:10	Coffee & Registration
11:10-11:15	Welcome Speaker: Scott Crossley
11:15-12:25	Invited talks, Session 1 Martin Meltzer, Centers for Disease Control & Prevention (CDC) Will Lowe, Princeton University Chair: Benjamin Miller
12:25-1:15	Lunch, provided by CSS 2015
1:15-1:45	Research at Yik Yak Chair: Eric Gilbert
1:45-2:45	Doctoral consortium. Chair: Jacob Eisenstein
2:45-3:05	Coffee and Snack Break
3:05-4:15	Invited talks, Session 2 Drew Linzer, Votamatic Zeynep Tufekci, University of North Carolina Chair: Jeffrey Staton
4:15-6:00	Poster Session & Reception

INVITED SPEAKERS

Martin Meltzer, *What public health leaders expect from models during a response to an emergency*
Centers for Disease Control and Prevention (CDC)

ABSTRACT (tentative): This talk will outline the types of data that public health officials seek from mathematical models during a public health emergency. The talk will describe how modeling is integrated into CDC's emergency response structure and illustrate the type of questions that have been addressed during large-scale responses. Also discussed will be the conditions and constraints under which the response modeling teams must work under in order to provide data useful to decision makers.

BIO: Dr. Martin I. Meltzer is the Lead of the Health Economics and Modeling Unit (HEMU), and a Distinguished Consultant in the Division of Preparedness and Emerging Infections, CDC in Atlanta, GA. He received his undergraduate degree from the University of Zimbabwe in 1982, and Masters and a Doctorate in Applied Economics from Cornell University, NY, in 1987 and 1990, respectively. From 1990 to mid-1995, he was on the faculty at the College of Veterinary Medicine at the University of Florida. In 1995, he moved to CDC, where he was in the first class of Prevention Effectiveness (health economists) Fellows. He lead the modeling teams supporting CDC's response to the 2009 H1N1 influenza pandemic, including producing monthly estimates of cases, hospitalizations and deaths, as well as estimating impact of the vaccination program and use of influenza anti-viral drugs. Other responses in which he lead the modeling activities include estimating the residual risk associated with the 2012 contaminated steroid injectable products that caused fungal meningitis among patients, and the 2014 Ebola epidemics in West Africa. Examples of his research include estimating the impact of the 2009 influenza pandemic, the modeling of potential responses to smallpox as a bioterrorist weapon, and assessing the economics of controlling diseases such as rabies, dengue, hepatitis A, meningitis, Lyme, and malaria. Dr. Meltzer has published approximately 210 publications, including over 100 papers in peer-reviewed scientific journals and more than 34 software tools. These tools include FluAid, FluSurge and FluWorkLoss, designed to help state and local public health officials plan and prepare of catastrophic infectious disease events. They have been downloaded more than 100,000 times and have been used by local, state, national and international public health agencies, with jurisdictions exceeding a total of 1 billion persons. He is an associate editor for Emerging Infectious Diseases. He also supervises a number of post-doctoral health economists at CDC.

Will Lowe, *What shall we do with the humans?*
Princeton University

ABSTRACT: Humans play various roles in social science text analyses, from coding the data, to providing gold standards, to exercising expert judgment on the output of statistical models of text. Increasingly, humans are also involved in the process of training them too. This talk describes some work on removing humans from these roles where we can, using them more efficiently and humanely when we cannot, and embedding them deeper when required.

BIO: Will Lowe is Senior Research Specialist in the Department of Politics at Princeton. He is a methodologist specializing in text analysis whose work has appeared in Political Analysis, International Organization, the Journal of Peace Research, Legislative Studies Quarterly, China Quarterly, and European Union Politics. He holds a PhD in Cognitive Science from Edinburgh University, but didn't let that stop him.

Drew Linzer, *The limitations of fundamentals-based presidential election forecasting*
Votamatic

ABSTRACT: U.S. presidential election forecasts are of widespread interest to political commentators, campaign strategists, research scientists, and the public.We argue that most fundamentalsbased political science forecasts

overstate what historical political and economic factors can tell us about the probable outcome of a forthcoming presidential election. Existing approaches generally overlook the uncertainty in coefficient estimates, decisions about model specifications, and the translation from popular vote shares to Electoral College outcomes. We introduce a Bayesian forecasting model for state-level presidential elections that accounts for each of these sources of error, and allows for the inclusion of structural predictors at both the national and state levels. Applying the model to presidential election data from 1952 to 2012, we demonstrate that, for covariates with typical levels of predictive power, the 95% prediction intervals for presidential vote shares should span approximately ±10% at the state level and ±7% at the national level.

BIO: Drew Linzer is the Chief Data Scientist at Daily Kos, located in Oakland, CA. A statistician and survey scientist, Drew was previously an Assistant Professor of Political Science at Emory University and professional pollster in California and Washington, DC. In 2012, he ran the election forecasting site votamatic.com. His research has appeared in the American Political Science Review, Journal of the American Statistical Association, International Journal of Forecasting, Political Analysis, Political Science Research and Methods, Journal of Politics, World Politics, Social Science & Medicine, and the Journal of Statistical Software. Drew holds a PhD in Political Science from the University of California, Los Angeles.

Zeynep Tufekci, *Computational social science: Considerations for a "dual-use" technology*
University of North Carolina

ABSTRACT: TBA

BIO: TBA

DOCTORAL CONSORTIUM

Kristopher Kyle, Georgia State University
Automatic analysis of syntactic development: A usage-based perspective

Tanushree Mitra, Georgia Institute of Technology
Understanding social media credibility: A deep dive into CREDBANK

Jane Lawrence Sumner, Emory University
Starving your enemies and your friends: The political consequences of the corporate provision of public goods

Dong Nguyen, University of Twente
On studying language variation in social media: Lessons learned from a public demo

POSTER SESSION

Using computational linguistic indices to predict humor in academic writing
Cynthia Berger, Stephen Skalicky, and Scott Crossley, Georgia State University
Danielle McNamara, Arizona State University

Computational indices derived from natural language processing tools were used to analyze a corpus of freshman college essays (N = 313) in order to better understand and predict humor in academic writing. First, the essays were scored for overall writing quality and creativity by human raters. Sub-scores of humor and wordplay from the analytic creativity rubric were then statistically aggregated to provide an overall *Humor* score for each essay in the corpus. Next, links between the *Humor* scores and computationally derived linguistic features were examined to investigate the potential for linguistic features to predict the Humor scores. A regression analysis identified four linguistic indices that accounted for approximately 17.5% of the variance in *Humor* scores. These indices were related to text descriptiveness (i.e., more adjective and adverb use), lower cohesion (i.e., less paragraph-to-paragraph similarity), and lexical sophistication (lower average word frequency). The findings suggest that humor is partially predicted by linguistic features in the text. Furthermore, a small but significant correlation was reported between the humor and essay quality scores, suggesting a positive link between humor and writing quality. Overall, these findings demonstrate the ability to use computational linguistic indices to predict a portion of the variance in raters' perceptions of humor and wordplay in academic writing, while contributing to a better understanding of how humor functions in academic writing.

==========

A measurement model to assess bias, construct validity, and effectiveness in peer evaluations of writing
Lee Branum-Martin and Melissa Patchan, Georgia State University

Peer learning is often used in classrooms to support knowledge and skill acquisition. One form of peer learning, peer assessment, involves the quantitative (i.e., peer ratings) or qualitative (i.e., peer feedback) evaluation of a learner's performance by another learner among students. While we might be concerned about the quality of the author's writing in the assignment, we might also be concerned about the quality of the evaluation given by peers (i.e., bias). Each evaluation in such a design represents two sources of variability: the author and the peer. Questions therefore become complicated regarding the validity of multiple measures to indicate writing quality, as well as relations with external measures, such as self-ratings or instructor's ratings. A multilevel structural equation framework can clarify these sources of variability. We analyzed ratings from 313 students in an undergraduate psychology course who each rated five classmates' essays in three aspects: content, style, and conformity to the assignment. Results suggest writing quality may be well represented by these three aspects. Approximately 30% of the variance in ratings was due to the author, and 10% of the variance was due to peer critics. Relations of peer ratings with self-ratings were positive, but low, suggesting that these students might not see the same aspects of quality in their own writing as in the writing of others. We explain how this modeling framework can clarify important questions of measurement validity and effective instruction and intervention, with implications for peer assessment designs more generally.

==========

Where does it hurt?: Analysis of health-related community questions and answers
Pavel Braslavski, Emory University
Alexander Beloborodov, Ural Federal University

The web has become an important source of health information for lay people. Many users post their health-related inquiries on community question answering (CQA) sites such as Yahoo! Answers. CQA platforms have collected a vast amount of data that is freely available online, indexed by search engines, and can be re-used for a variety of tasks.

Our ongoing project investigates versatile aspects of health-related CQA data: topical structure, content quality, and user behavior. In our study, we use a sizable dataset from a popular Russian CQA site Otvety@Mail.Ru. We apply topic modeling to the data to uncover underlying topics (diseases and symptoms), compare topics dynamics with real-life events such as flu epidemics and weather conditions, and discover regular topic-to-topic transitions in subsequent questions posted by the same user. About 1,500 question-answer pairs were manually evaluated by medical professionals; in addition an automatic evaluation based on reference disease-medicine pairs was performed. We conducted two surveys: on motivation of CQA answerers and on attitudes of physicians toward health information online and on CQA sites in particular.

The outcomes of the project can be used to improve policies and content quality of CQA in health domain, to advance in re-using CQA data in various tasks, to perform large-scale health and drug usage surveillance, as well as to identify common misconceptions and beliefs, which in turn can guide health education.

==========

Network analysis of pro-eating disorder Instagram communities
Stevie Chancellor, Julia Deeb, and Munmun De Choudhury, Georgia Institute of Technology

Research shows that online communities help those struggling with illnesses, like cancer and diabetes, by being an always-available source of information and support. However, some online communities promote self-injury, unstable thoughts, and disordered habits derived from a false belief that mental illnesses are alternative lifestyles. One community in particular, the pro-eating disorder (pro-ED) community, promotes disordered eating and exercise habits, self-injury, and suicidal tendencies. Instagram hosts a prominent pro-ED community that is tightly connected and very active.

This project explores this community through the lens of network analysis. Instagram's robust tag network and follower/following structure exposes the interconnections within the pro-ED community that may offer insights into how the community connects with others and changes over time. Our dataset contains over 170,000 users and 33 million posts of people who have posted to pro-ED Instagram tags.

We ask the following questions: what clusters of users and tags form? How do these clusters change over time? Do certain clusters experience higher levels of mental illness severity than others? To answer some of these questions, we will model the tag relation network with a temporal dynamic graph that uses monthly time intervals. We hope to visualize important clusters in the community as well as identify areas of the community that may be more susceptible to mental illness severity than others. This research could lead to new strategies in interventions as well as informing social computing researchers how such deviant communities form and change over time.

==========

Lexicality and lying: Lexical observations of deceptive language
Meredith D'Arienzo, Georgia State University
Nicholas Duran, Arizona State University
Scott Crossley, Georgia State University

Deception in spoken language is an area that has received little attention from previous applied linguistic research. This study examines the ways in which deception manifests itself lexically in spoken interaction. Previous studies, in particular Newman et al. (2003), have found that deception is revealed through the use of fewer first- and third-person singular pronouns, more negative emotion words, fewer exclusive words, and more motion verbs. This study examines 49 pairs of conversations between undergraduate students on a variety of controversial topics; in each conversation, one of the students was told to lie about their point of view. By comparing truthful and deceptive language using a

computational tool, we will investigate whether factors previously found to reveal deception are present and whether additional lexical factors such as the use of n-grams, lexical range, and psycholinguistic word properties are prominent. In addition, this study will examine whether there are noticeable effects on the interlocutor's language when speaking with a deceptive partner.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003) Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, 29(5), 665-675.

==========

Characterizing dietary choices, nutrition, and language in food deserts via social media
Munmun De Choudhury, and Sanket Sharma, Georgia Institute of Technology
Emre Kiciman, Microsoft Research

Social media has emerged as a promising source of data for public health. This paper examines how these platforms can provide empirical quantitative evidence for understanding dietary choices and nutritional challenges in "food deserts" — Census tracts characterized by poor access to healthy and affordable food. We present a study of 3 million food related posts shared on Instagram, and observe that content from food deserts indicate consumption of food high in fat, cholesterol and sugar; a rate higher by 5-17% compared to non-food desert areas. Further, a topic model analysis reveals the ingestion language of food deserts to bear distinct attributes. Finally, we investigate to what extent Instagram ingestion language is able to infer whether a tract is a food desert. We find that a predictive model that uses ingestion topics, socioeconomic and food deprivation status attributes yields high accuracy (>80%) and improves over baseline methods by 6- 14%. We discuss the role of social media in helping address inequalities in food access and health.

==========

Rhetorical patterns in legislative speech
Jacob Eisenstein, Georgia Institute of Technology

Discourse describes linguistic phenomena beyond the sentence level, explaining how information and arguments are organized, both semantically and rhetorically. Discourse therefore can quantify not only what is said in political speech, but how it is said --- potentially identifying similarities in argumentative style across disparate topics and political issues, and revealing which forms of argumentation may be brought to bear in various political situations.

To explore the relationship between discourse structures and political context, I apply a state-of-the-art discourse parser to the speech of members of the United States House of Representatives, using a dataset of congressional floor speeches. I then link rhetorical motifs in these texts to political variables relating to the speaker, the bill under discussion, and the larger legislative context. The purpose of this analysis is to quantify how discourse structure reflects political factors, and conversely, to assess the predictive power of discourse as an issue-neutral signal of political ideology and intention. A further contribution is the quantitative comparison of the two best known models of discourse structure, in terms of their applicability to the domain of political speech.

==========

Understanding the diffusive and resistive nature of lexical innovations in social media
Rahul Goel and Sandeep Soni, Georgia Institute of Technology

Lexical innovations (new word usage) spread through a social network over time to different places, having originated in one place. An example is the usage of the emoticon -_- on Twitter, which started from coastal US but spread to other

parts over time. However, we also observe that some words remain bound to a region over long time. An example is the word ard which has remained localized in or near Maryland. We hypothesize that the variation in such diffusion is because of the combination of social network and dialect effects. The overall spread is due to assistance and resistance from the social network structure and regional dialect.

We model usage of words as events being generated by an evolutionary point process. In our study, we consider the multi dimensional hawkes process where the next usage depends on past occurrences by every user. We model the influence between every pair of users as a linear combination of the social structure features and shared geographic features between the pair. By learning the weights over these features, we can understand how the social network structure and the geography play a role together in diffusion.

Our dataset consists of 37 new words and corresponding timestamped tweets from a 3 year period between Mar 2011 to June 2014 mentioning these words. We also have a mentions social network from twitter from the same period. Additionally, we have geographic data for some users to help identify the metropolitan area to which a user belongs.

==========

Extracting cohesive indices of deceptive language using TAACO
Ali Heidari and Scott A. Crossley, Georgia State University
Nicholas Duran, Arizona State University

The language people use and the way they use it reveals a great deal about their mental state in deceptive situations (Duran et al., 2009). Therefore, it is expected that language produced in deceptive situations will be qualitatively different from that of truthful circumstances. However, research to date has not been successful in providing a reliable set of linguistic features that is distinctive of different mental states in truthful versus deceptive situations. The present research intends to apply a natural language processing tool called the Tool for the Automatic Analysis of Cohesion (TAACO) to analyze cohesive indices of deceptive and truthful language. TAACO incorporates over 150 classic and recently developed indices related to local, global and overall textual cohesion. Applied to a computer mediated corpus of deceptive and truthful language, TAACO output revealed that there were differences between deceptive and non-deceptive language. It showed that compared to truthful language, deceitful language contained less repeated content words, bigrams and trigrams and fewer function words. The findings from this study will be discussed to provide insight into differences between truthful and deceptive language.

==========

An enriched game-theoretic model of International Politics: Mining historical data to spot hidden patterns
Hamid İncidelen, Bilkent University
Dilara Soylu, Georgia Institute of Technology

Game theory establishes a formal way in international politics to analyze cooperation and conflict among states by assumptions about their preferences and examination of historical cases. As Duncan Snidal points out, "game theory often seems to demand more information than can be feasibly obtained". Indeed, by ignoring a bulk of data for the sake of simplicity, analysts devise models that are mostly too plain to capture the historical richness, relevant actors or choices available to actors. Such shortcomings led some analysts into failure to predict the outcomes in international politics.

Our demo offers the application of computational social science tools which enables us to analyze and process the big data to encompass a larger body of information to be used in game-theoretic simulations. An enriched agent-based model helps analysts to give the meaning of historical trends and social phenomena that are placed -but ignored- in the data. Being able to include more data in the model results in more accurate outcomes regarding the players and their

choices.

It would surely take a long time for analysts to handle such a great amount of data which is not necessarily coherent in itself. Instead, machine learning may work more efficiently through algorithms to mine data and look for meaningful patterns that are present in the data to adjust it into a game-theoretic simulation.
Our demo uses tools such as machine learning and big data analysis to efficiently benefit data to obtain more accurate results in game-theoretic analyses of international politics.

==========

Distribute giveaways effectively in mobile opportunistic social networks
Chenguang Kong and Xiaojun Cao, Georgia State University

When one user meets another user in Mobile Opportunistic Social Networks (MOSNs), they can establish a temporary connection for information exchange. For instance, a book author may use these opportunistic and intermittent connections (or encounters) in MOSNs for distributing book samples to interested readers. In this work, we study how to effectively distribute limited giveaways effectively through the opportunistic connections among mobile users in MOSNs. We first derive a Social Connection Pattern (SCP) to statistically describe the interest distribution of the users connected. The SCP is applied to predict the interests of possible connected users in future. We then propose a giveaways distribution algorithm based on the Social Connection Pattern to calculate the best number of content copies to distribute when two users meet. Our dataset based simulation shows that the proposed algorithm is effective and efficient to distribute giveaways in MOSNs.

==========

Connection recommendation for efficient information acquisition in social networks
Chenguang Kong, Georgia State University
Guangchun Luo and Ling Tian, University of Electronic Science and Technology of China
Xiaojun Cao, Georgia State University

Social networks such as Twitter and Facebook have become important sources for users to acquire information. In these social networks, users can obtain information through the posts/reposts from their social connections. To acquire information efficiently, users are motivated to connect to users that offer attractive and timely information. In this work, we study how to recommend social connections for a user to maximize the efficiency of its information acquisition. We define this problem as Connection Recommendation for efficient Information Acquisition (CRIA). To measure the information acquisition efficiency, we analyze the information accuracy/timeliness, and propose a novel MapReduce-based Connection Set Selection (CSS) algorithm to efficiently solve the CRIA problem. Our study shows that the proposed CSS can recommend connections for a user to obtain information with high accuracy, low spam rate and low latency.

==========

Characterizing the dynamics of a protracted activist movement via social media
Shagun Jhaver, Munmun De Choudhury and Benjamin Sugar, Georgia Institute of Technology

Social media platforms have emerged as powerful platforms of expression for individuals from diverse groups as well as around a variety of mundane and sensitive real-world happenings. We study how data shared on social media around the Black Lives Matter protests and the associated incidents of police brutality in Ferguson, NYC and Baltimore may reflect the broader community's perception of these events. We are particularly interested in studying how the protracted nature of these events are impacting people's psychological expression on Twitter and the linguistic and

social adaptations people make in their Twitter feeds while conversing about this sensitive and highly contested topic. Our initial investigations reveal that increased participation in these topics on Twitter is associated with habituation of negative emotions, such as decreased anger and anxiety, although we observe the evolution of a collective identity over time. We situate our findings in the context of literature in psychology that examines prolonged upheavals in the society and their impact on the well-being of affected communities.

==========

Varieties of democratic stability
Richard Morgan, Jian Xu, and Adam Glynn, Emory University

From the consolidation of power to the breakdown of political order, the relationship between institutional design and political stability is a mainstay in the political science literature. However, despite the important role political institutions play in our theories of stability and transition, ambiguity remains concerning which features of these institutions have the greatest influence on the fragility and longevity of a regime. By using over 300 indicators of political institutions from the Varieties of Democracies (VDem) project, this project moves towards this goal. Informed by well-established theories of regime stability and longevity, this project uses these indicators and random forest techniques to predict democratic transitions and breakdowns as well as civil conflict onset and low-level social unrest. This allows us to answer the following questions: Which design features of a state's political institutions have the greatest association with the likelihood that a state will transition to a democracy? Do these same features also affect democratic breakdown and a state's slide towards autocracy? Further, which institutional features have the largest influence on the likelihood of internal unrest and the use of political violence?

==========

Automatic detection of narrative similarity
Dong Nguyen , University of Twente

Red Riding Hood, or the urban legend about the Vanishing Hitchhiker are stories that most of us are familiar with. However, when asking people to recall a specific story, everyone tells his or her own version. Variations of such stories appear due to their oral transmission over time: locations can change, characters can be added, or complete events introduced or left out. Automatically detecting variations of the same story is useful for humanities scholars as well as to support the digitization of folk narrative collections.

I will present two studies on narrative similarity based on data from the Dutch Folktale Database. In the first study, narratives are automatically classified according to their story type. A story type represents a collection of 'similar' stories often with recurring plot and themes. For example, an example story type is Red Riding Hood, which is classified as story type ATU 333 in the Aarne-Thompson-Uther type-index. We experiment with approaches inspired by distributed information retrieval in a learning to rank framework, and demonstrate high performance gains compared to a baseline method.

In the second study a crowdsourcing experiment was performed in which the perception of narrative similarity between folktale experts and crowd workers was compared. While experts focus mostly on the plot, characters and themes of narratives, non-experts also pay attention to dimensions such as genre and style. The results show that a more nuanced view is needed of narrative similarity than captured by story types alone.

==========

A causal inference approach to study emoji usage on Twitter
Umashanthi Pavalanathan and Jacob Eisenstein, Georgia Institute of Technology

Non-verbal cues present in face-to-face communication provide rich contextual information about the utterance such as author intention and affective state. Online writing lacks these cues and people are changing their writing to express themselves in online settings through non-standard orthographies such as emoticons, expressive lengthening, and non-standard punctuations. Recently, emojis have been introduced to social media such as Twitter and Instagram and are becoming increasingly popular. Are these colorful and expressive pictographs going to replace earlier orthographic methods of paralinguistic communication? To shed some light on this question, we test whether the adoption of emojis on Twitter causes individual users to use fewer emoticons in their tweets. Using a matching approach from causal inference we find evidence for the hypothesis that the emojis compete with emoticons, and that the introduction of emojis can lead to a decline in orthographic variation. We will further investigate whether the adoption of predefined pictographic characters makes online writing more standard, by looking at the differences in standard word usage rates by users adopting emojis on Twitter.

==========

A learning analytics approach for studying students' science epistemologies
Melanie Peffer, Maggie Renken, and Caleb Lewis, Georgia State University

Science epistemology, or beliefs about what it means to do science and how science knowledge is generated, is an integral part of authentic science inquiry. Although the development of a sophisticated science epistemology is critical for attaining science literacy, epistemology remains an elusive construct to precisely and quantitatively evaluate. Our previous work demonstrated that when students are engaged in the computer-based non-linear authentic inquiry experience provided by Science Classroom Inquiry (SCI) simulations, their inquiry practices are unique to each user and diverse. Diverse inquiry practices, such as running multiple investigations, seeking additional information, and coordinating evidence with theory, may provide insight into students' science epistemology. Here we propose the use of learning analytics to assess students' science epistemology by investigating user practices when engaged in a SCI simulation. We found that although undergraduate students score high with little variance on an established metric of science epistemology, we detected differences in actions performed by each student. Students displayed a preference for either investigative actions (e.g. generating hypotheses, performing tests) or information-seeking actions (e.g. Internet searches, use of the simulation library function). This preference for investigative or information-seeking actions was predictive of total decisions made and information-seeking actions, but not investigative actions. We also detected a putative effect of gender on the number of actions made. This study underscores the potential for the use of learning analytics in simulated authentic inquiry to provide a novel and valuable method of assessing inquiry practices and related epistemologies.

==========

Footprint design: The role of ontology in visual representations of human behavioral data
Peter Polack, Georgia Institute of Technology

In analyzing human behavioral data, analysts may describe an individual's daily routine as a set of tendencies, where at any given time, a person has a characteristic probability of returning to particular activities or locations. From this perspective, analysts can infer footprints of regular human behavior, as well as deviations from normal human behavior, and best determine how to prepare for and react to these variations with behavioral models and adaptive interventions. However, although this strategy can successfully classify instances of certain normal and irregular behaviors, it cannot accurately determine whether consistent or anomalous behaviors are, in fact, meaningful ones. In the case of spatiotemporal data, an individual's presence in a location new to them may or may not entail a meaningful, introspective change in behavior. Therefore, acknowledging that the characteristics of human behavior depend on unmeasurable external circumstances and internal motivations, we demonstrate how a reliance on semantic ground truths to understand human behavioral patterns can reveal more about the data than about the sampled behaviors

themselves. Further, we indicate how varying representational models of human activities affect interpretations of behavioral character, and how an overly objective segmentation of behavioral activities can contribute to inaccurate conclusions about human behavior. To accomplish this, we visualize spatiotemporal behavioral patterns as hierarchies of location-based activities, with the intent to discover, understand, and exemplify the constraints of various objective representations of human behavioral data. Through this demonstration, we encourage a dialogue about alternate, human-centered means of representing and using behavioral data for behavioral analysis and designing adaptive interventions.

==========

Effects of interface or user ability? Interpreting data collected with a computer-based tool
Maggie Renken, Ilya Goldin and Ellen Litkowski, Georgia State University

We developed a computer-based instrument to measure students' skill identifying sources of knowledge in explanations of science phenomena. Due to the complexity of the user interface task, creating this instrument presented challenges for interaction design. The basic task of the instrument is for users to categorize segments of a brief passage. Categorization decisions center on five prescribed sources of knowledge (SoK). Thus, these decisions must be elicited and collected at a specific grain size. Because the instrument is meant to measure a developmentally progressive skill, it needed to be equally accessible to elementary through undergraduate students. This has implications for how children can interact with the passage and task. Because the instrument is meant to measure a contextually situated skill, it needed to include multiple domains. We will demonstrate requirements for text presentation (especially to children) and for the categorization task. We will further describe two options in Qualtrics, a web-based survey application, we pursued for the design. Finally, we will report on the relevant results of two studies. In Study 1, 338 elementary (grades 2-5) and middle (grades 6-8) students completed the instrument using a highlighting tool. In Study 2, 81 elementary (grades 2-5) students completed the instrument using a grid selection tool. Differences in children's propensity to categorize a segment of text as a given SoK as a function of instrument design will be illustrated. Implications for parsing the influence of interface design and user ability when interpreting data collected with computer-based tools will be emphasized.

==========

Measuring creativity using computational models: A linguistic approach
Stephen Skalicky and Scott Crossley, Georgia State University
Danielle McNamara, Arizona State University
Kasia Muldner, Carleton University

This study investigates the linguistic features of creativity during interactive problem solving using computational text analysis tools that measure the lexical, grammatical, and psycholinguistic properties of words. Data was gathered from thirty-nine pairs of undergraduate students (n = 78). Each pair completed three problem-solving tasks using computer-mediated communication. Task instructions informed the participants to generate as many solutions to the problems as possible.

Participant interactions were measured for creativity by expert raters using an analytic rubric with seven sub-scales designed to capture overall idea generation and style. A principle component analysis yielded two factors: creativity and elaboration. Measurements of the linguistic features of the participant interactions were then obtained from text analysis tools TAALES, TAACO, and WAT; these measurements were input into stepwise regression analyses in order to develop computational models predictive of the creativity or elaboration scores.

Results demonstrated that linguistic features were able to account for 83% of the variance in scores for the creativity component and 71% of the variance in scores for the elaboration component. Analysis of the coefficients revealed that

creativity contained more different word types, modal verbs, words with high amount of associations and fewer frequent words, academic words, and private verbs (e.g., think, know). The elaboration component was marked by more repeated function words, exemplifications (e.g., for example) and fewer social words (e.g., family), connections between independent clauses, and differences in paragraph-to-paragraph semantic similarity. These results demonstrate that computational models can successfully be used to predict linguistic creativity.

==========

Spanish-English bilingualism as human capital?: Discourses of multilingualism on CareerBuilder
Nicholas Subtirelu , Georgia State University

The value of individual multilingualism is often framed in economic terms, as a form of human capital increasing individuals' value on the labor market (Heller & Duchêne, 2012). Nonetheless, assertions that multilingualism offers an economic advantage appear to overlook other factors contributing to the social valuation of languages in society, which impact the degree to which labor market rewards accrue to multilingual individuals. Indeed, Spanish-English bilinguals in the US have commonly been found to earn less than their English monolingual counterparts, suggesting a penalty for bilingualism (Alarcón et al., 2014). Such findings, however, contrast with employers' reports that they treat knowledge of Spanish as a desirable trait (Porras et al., 2014).

Such apparent contradictions between analyses of Census Bureau income data and employer surveys suggest that closer examination of how employers conceptualize multilingualism is warranted. In this study, I take a mixed methods computational approach, analyzing, both quantitatively and qualitatively, a corpus of job advertisements collected from the website CareerBuilder. I show that advertisements mentioning Spanish are somewhat common on CareerBuilder: about 3% of all advertisements, although this frequency varies in expected ways (e.g., geographically). I also show that the mention of Spanish within an advertisement is associated with lower advertised wages. Finally, I examine how advertisements mentioning Spanish construct linguistic competence differently across three levels of pay, arguing that lower wage positions treat multilingualism as a crucial aspect of the work to be performed in contrast to higher wage advertisements that treat multilingualism as an index of managerial cosmopolitanism.

==========

Inferring latent user characteristics for analyzing political discussions in social media
Yu Wang, Eugene Agichtein, and Tom Clark, Emory University
Mark Dredze, John Hopkins University
Jeffrey Staton, Emory University

Twitter and other social media platforms, which carry opinions from millions of users, have become an important resource for political science analysis. While this has afforded enormous research opportunities, the nature of the data provide challenges, as demographic characteristics -- a key factor in political science research -- are unavailable. One approach relies on supervised learning to infer traditional demographic characteristics, such as age, gender and ethnicity. Unfortunately, this method must identify, in advance, which demographic characteristics are important for each issue, and create a classifier accordingly; additionally, some issues cut across demographic boundaries, making traditional demographic distinctions less important.

We propose another approach: learning latent user characteristics and interests directly from Twitter data, by mining user self-descriptions from profiles with unsupervised learning. These latent user characteristics complement and sometimes replace traditional demographic attributes for grouping users. We present preliminary results of analysis of two major political issues surrounding US Supreme Court decisions, to demonstrate the effectiveness of our approach. Our results indicate that our proposed approach may be more effective than using manually coded, or automatically inferred, traditional demographic characteristics for political sentiment analysis tasks.

==========

Future orientation and wellbeing at the population level: A preliminary analysis based on twitter posts

Phillip Wolff, Emory University
Bridget Copley, CNRS / Paris 8
Eugene Agichtein, Robert Thorstad, Allen Nie, and Reid Kilgore, Emory University

Thinking about the future—or prospection--may be good for you. Prior research suggests that future thinking is associated with higher levels of physical and mental health, academic achievement, economic success, and social engagement. In this research, we extend these findings by examining how prospection and wellbeing may be related to each other at the population level. A temporal orientation parser was developed for categorizing linguistic expressions as referring to the past, present or future. Using this parser, we automatically detected the temporal orientation of 8.5 million tweets sampled from each of the states in United States. This allowed us to estimate the overall future orientation of each state – and in turn to compare each state's temporal orientation to the health and economic wellbeing statistics of several surveys reported in the Gallup-Healthways State of the States series.

Our preliminary results show that future-orientation of states correlated strongly with various state-wide measures of mental and behavioral health, including wellbeing, exercise frequency, community recognition and concern for the future. In a further analysis, we analyzed the future-oriented tweets to estimate how far into the future they refer. Interestingly, this future distance was associated with a very different set of mental and behavioral practices: States with Twitter users making reference to the far future were more likely to eat vegetables, avoid drunk driving and invest in their highway and educational systems. The results provide the strongest evidence available to date that future-oriented thinking may be good for a population's mental and physical health.

==========