Computation+Journalism 2022

# Schedule of Talks

*Presenters' names that are bolded and black will be attending in-person, while those that are bolded and blue will be remote.*

**Invited Session — Friday, June 10, 10:00 AM - 12:00 PM EDT**

## [1] Conflict Reporting

**Behind The Civilian Casualty Files: Investigating The 'Most Precise' Air War In History - Azmat Khan** (Columbia Journalism School)

Year after year, successive U.S. administrations sold a nation weary of "forever wars" on replacing boots on the ground with "the most precise air campaign in history." The promise was that America's "extraordinary technology" would allow the military to kill the right people while taking the greatest possible care not to harm the wrong ones. In a five-year investigation based on more than 1,300 military records obtained via a lawsuit, visits to the sites of more than 100 civilian casualty incidents in Iraq, Syria and Afghanistan, and scores of interviews, Azmat Khan exposed its legacy: deeply flawed intelligence, missed targets, disproportionate destruction and scant accountability. In this talk, Khan will walk through how she did it, and how data journalists can resist approaches to journalism that flatten the experiences of those caught in conflict.

**Mapping Incidents of Civilian Harm in Ukraine - Charlotte Godart** (Bellingcat)

Bellingcat is an independent international collective of researchers, investigators and citizen journalists using open source and social media investigation to probe a variety of subjects. Since the weeks leading up and the invasion of Ukraine, we have been collecting open source content to locate and eventually verify. Currently, we have a journalistic output, our map of civilian harm in Ukraine, and a justice and accountability team that is further analysing the details of these events for potential legal proceedings. This talk will present the opportunities and challenges posed by open source collection and verification within the context of Ukraine.

**How data journalism is responding to the war: What can satellite images say? How do we detect disinformation? What other data can we use? - Roman Kulchynskyj** (Editor-in-Chief, Texty.org.ua), **Peter Bodnar, Illya Samoilovych** (Data journalists, Texty.org.ua)

How to produce quality data journalism in times of war? Roman Kulchynsky and his colleagues from TEXTY.org.ua explain what kinds of data have been generated by the recent Russo-Ukrainian war and how their team collects and analyzes them.

**Paper Sessions — Friday, June 10, 12:10 PM - 01:30 PM EDT**

## [2] Refereed Papers: Applications of Machine Learning — Lecture Hall

**Generating a Pairwise Dataset for Click-through Rate Prediction of News Articles Considering Positions and Contents - Shotaro Ishihara** (Nikkei, Inc.) and Yasufumi Nakama

In online news websites, the headlines and thumbnail images of articles are displayed in a list, and they are important navigation links to individual article pages. If we can predict the click-through rate (CTR) of readers to the article pages, we can assist the editors in creating article headlines and setting thumbnail images. However, the CTR that can be observed in the access log is heavily affected by the display position, and it is difficult to predict the CTR by machine learning using data on single articles. This paper proposes a method to construct a pairwise dataset based on the information such as similarity of the display positions and contents, and create a model to predict the CTR in the framework of pairwise learning-to-rank. In the experiment, we verified the usefulness of the proposed method by using the actual access log data and discuss the potential of the practical use of the CTR prediction as editing support.

**Detecting Stance of Tweets Toward Truthfulness of Factual Claims** - **Zhengyuan Zhu**, Zeyu Zhang, Foram Patel and **Chengkai Li** (University of Texas at Arlington)

Journalists aim to understand misinformation on social media, especially in discerning the public's opinions toward the veracity of misinformation. For that, an algorithmic tool for truthfulness stance detection can be particularly useful. This paper introduce a deep learning model we developed for detecting the stance of tweets toward the truthfulness of factual claims. The models were constructed using a dataset curated and annotated in-house. While both the models and datasets warrant further development and refinement, preliminary experiments demonstrated promising results. The model is available through both an Application Programming Interface (API) and a demonstration website.

**What's the Fairest of Them All? Aesthetic Assessment of Visuals** - Marc Willhaus, **Daniel Vera Nieto**, Clara Fernandez and Severin Klingler (Media Technology Center ETH Zurich)

Attractive images and videos are the visual backbone of journalism and social media. From trailers to teaser images to image galleries, appealing visuals have only grown in importance over the past years. Especially online, eye-catching visual content can significantly impact user engagement. However, selecting the best shots from a long video or selecting the perfect image from a vast image collection is a challenging and time-consuming task. This paper presents a system to automatically assess image and video content from the perspective of aesthetics. While this is a highly subjective task, we find that it is possible to combine expert knowledge with data-driven information to perform such an assessment. In order to do so, we identify relevant aesthetic features together with experts from the media industry and implement machine learning algorithms to infer them from the visual content. We combine the features under a single aesthetics retrieval system that allows users to sort uploaded visuals according to an aesthetic score and interact with additional photographic, cinematic, and person-specific features. The system is built into a containerized application to guarantee reproducibility. A demo video of our tool is available.

### [3] Contributed Session: AI and Investigations — Brown Institute

As investigative computational journalists, sometimes we look into the black boxes of algorithms written by other people; and sometimes we write our own AI in order to investigate social problems. Our panel will discuss the opportunities and challenges AI can bring to journalism and investigative reporting, as well as creative strategies for cracking open black boxes and discovering problems. - **Bahareh Heravi** (University of Surrey), **Meredith Broussard** (New York University), **Julia Angwin** (The Markup), **Hilke Schellmann** (New York University).

## [4] Contributed Papers: News Analysis — 607b

### **Studying Local News at Scale** - Marianne Aubin Le Quéré, Ting-Wei Chiang and Mor Naaman (Cornell Tech)

As local news outlets in the United States continue to disappear, communities are losing access to key local civic information. While we know that the loss of newspapers results in uneven access to information, our understanding of who is losing access to which types of local information is limited. In this talk, we present a public dataset of social media handles and metadata of over 10,000 local news outlets we curated and collected using semi-automated methods. We preview our method for analysis at large-scale local news data using this dataset. We show how we used this dataset to make inferences about news outlet features, including the economic state or the “localness” of an outlet. Finally, we present a Covid-19 case study, demonstrating the potential of extracting pertinent insights from analysis of such local data at scale. In our case study, we find key differences in how different types of local news outlets covered Covid-19, and how different audiences engaged with the content. For example, our results show that the more “local” an outlet is, the more they tended to cover Covid-19 statistics compared to other Covid topics. We conclude this talk by inviting collaboration and soliciting questions from other researchers and journalists, and proposing potential future work.

### **Mining Questions for Answers: How COVID-19 Questions Differ Globally** - **Jenna Sherman**, Smriti Singh, Scott Hale and Darius Kazemi (Meedan Digital Health Lab, Harvard T.H. Chan School of Public Health)

Health institutions such as the CDC and the WHO have been working to provide the public with current and accessible information on COVID-19 throughout the pandemic. A range of factors likely means these questions differ across the world; yet, research on this topic is lacking.

This study’s three aims are: 1) to examine the differences in key themes of COVID-19 questions by country, 2) to examine the differences in key themes of COVID-19 questions through a public health lens, and 3) to assess how anonymous search queries on Bing differ from pseudonymous questions on Twitter.

We are addressing these aims by analyzing a set of queries sourced from a database of COVID-19-related questions searched on Bing from January 2020 to July 2021 and comparing them by region as well as based on a set of pre-selected public health metrics such as COVID-19 case fatality rates by country and the availability of universal health coverage. We have also collected pseudonymous COVID-19 questions on Twitter and are comparing these to the Bing search queries to understand the role of anonymity.

This talk will discuss our findings on the similarities and differences in questions about COVID-19 disaggregated by geography and analyzed based on a set of public health metrics and anonymity level. The findings of this research are critical for journalists working to disseminate information in emergencies and prolonged crises to ensure health messaging is produced and dispersed equitably with emphasis on populations who are being most greatly impacted.

**Vulnerable Visualizations: How Data Visualizations Are Used to Promote Misinformation Online** - **Maxim Lisnic**, Marina Kogan and Alexander Lex (University of Utah)

Data visualizations can help explain and summarize otherwise incomprehensible amounts of data. On social media sites, users actively share and comment on data visualization posts as a form of collective sense-making. But the insights gained from these visualizations—even ones created by trusted media outlets using accurate data—can be misleading and perpetuate misconceptions and misinformation.

We analyzed a sample of over 33,000 data visualizations related to the COVID-19 pandemic shared on Twitter in 2020 and 2021. Our findings show that one of the most common ways of making sense of an ongoing crisis through visual representations data is by annotating screenshots of existing charts from reputable sources. Such annotations can potentially create or alter the chart's meaning to support misconceptions and conspiracy theories without any fake data or visual tricks. This phenomenon highlights the importance of considering audience bias in designing visualizations for data-driven journalism.

We introduce the notion of vulnerable visualizations: visualizations created from accurate data without the intention to misinform but susceptible to supporting misinformation through biased framing. We present a series of case studies of existing misinformation posts and propose design recommendations to protect charts and prevent their misinterpretation.

**Navigating Multi-perspective News Stories By Showing What Is Common, And What Is Debated** - **Philippe Laban** (Salesforce Research), Lidiya Murakhovs'Ka (Salesforce Research), Xiang Anthony Chen (UCLA) and Chien-Sheng Wu (Salesforce Research)

Modern news aggregators group all articles about a given event together, enabling news readers to easily access the “full coverage” for any news topic. The full coverage of a news story can however be overwhelming for a news reader, sometimes faced with selecting which article to read from up to a hundred sources. In such cases, a reader might select the source they are most familiar with, reducing their exposure to the desired full coverage.

We propose to organize the presentation of multi-source news stories in two parts: first present the basic facts that are common across the majority of sources, and second present specific questions on which sources diverge, reducing the effort on the news reader further to access the full coverage of a news topic. We present the automatic pipeline we use to assemble the two parts of content, based on a Question Generation, Question Answering and Summarization components, and a preliminary interface illustrating the idea.

**Invited Session — Friday, June 10, 02:30 PM - 04:30 PM EDT**

**[5] COVID Reporting — Lecture Hall**

### **How The Economist estimated the pandemic's true death toll** - **Sondre Ulvund Solstad** and Martín González (The Economist)

In May 2021, The Economist became the first and only organisation in the world to publish an estimate of the pandemic's true death toll, globally and by country (a year later the WHO would publish their own estimates, in line with ours). At the time these were the only estimates of excess deaths - the best metric of pandemic mortality, with no dependency on testing - available at the global level, and in all countries and territories. These estimates, based on machine learning models trained on over 100 variables, now update automatically every morning, and are used by people around the world. How did this project move from idea to execution, and what were the computational, statistical, and journalistic challenges involved?

We discuss how we navigated the scale of the project, the challenges in visualising uncertainty, and our interactions with international organisations relying on our work. We also try to provide a framework for how work of this scale can be done transparently, and in ways allowing others - be it academia, journalists, or government officials - to leverage it effectively.

### **The Covid Financial Crisis** - **Nick Thieme** (The Atlanta Journal-Constitution)

The COVID-19 pandemic presented a series of tremendous financial obstacles to both individuals and governments around the world. While everyone was affected to some extent, discrepancies in financial harms appeared at the household as well as municipal levels. At The Atlanta Journal-Constitution, we undertook a grant-funded, multi-year investigation into what those financial harms looked like and where those disparities were located in the state of Georgia.

To do this, we built a novel computational pipeline that, based on sampling and power calculations, automatically sampled bankruptcy cases from the PACER system of the Administrative Office of the United States Courts and used a variety of computer vision and OCR techniques to turn legal documents into tidy data. Using a combination of time series methods, general additive models, and other statistical techniques, the AJC was able to report on bankruptcy trends during the pandemic, including racial and class disparities in student loans and medical debt, with a unique level of statistical sophistication.

Additionally, with a first-of-its-kind data analysis of municipal bond issuances in Georgia from the Municipal Securities Rulemaking Board, the AJC published an article detailing the incredible cost of increased interest rates on municipal financing—in the hundreds of millions of dollars—that Georgia taxpayers are now saddled with. This article and this series, represent a step towards statistical methodology and modeling as modern data journalism.

### **Managing the challenges of data reporting and visualization in year three of the Covid-19 pandemic at The New York Times - Lisa Waananen Jones** (Washington State University/The New York Times), **Aliza Aufrichtig** (The New York Times) and **Tiff Fehr** (The New York Times)

In March 2020, as Covid-19 cases in the United States rapidly grew from dozens to thousands, the single Google spreadsheet New York Times reporters had been manually updating since January became too large to load properly. Since then, the project has required a balance of interdependent manual and automated processes to collect and publish county-level cases and deaths on a daily basis. No consistent data format ever emerged nationwide. In the third year, the project still requires full-time staffing to respond to daily shifts in data format and availability as well as evolving public information needs.

This talk places this process in the context of other long-term projects related to producing aggregated Covid-19 data sets, which fall along a spectrum of mostly manual to mostly automated processes. Because of the decentralized health department system in the United States, aggregating disparate sources on a daily basis requires ongoing methodology decisions about data definitions, averaging, and anomaly handling. The Times' approach is grounded in traditional journalism principles and techniques, which has over time revealed potential conflicts between the goals of journalism and data science.

The decentralized government system in the United States poses unique challenges and opportunities for computational journalism, and Covid-19 data demonstrates the risk of using federal data in news and research without supplementary reporting. Lessons from this project could inform large-scale dynamic data collection and publishing on other topics such as housing, education, legislation, and criminal justice.

## Paper Sessions — Friday, June 10, 04:40 PM - 06:20 PM EDT

## [6] Refereed Papers: Online Communities and Local News

### **Comparing open-ended community dialogue with local news - Hope Schroeder**, Doug Beeferman and Deb Roy (MIT Center for Constructive Communication)

In the lead-up to Boston's local elections in November 2021, the Real Talk For Change project hosted small group conversations in which local residents shared their experiences of living in Boston in response to the prompt "What is your question about the future of Boston and your place in it? What experience led you to that question?". Over 370 people from 21 of the 23 neighborhoods of Boston participated, often sharing deeply personal stories. The conversations were recorded, analyzed, and used as a basis for sharing themes and stories with the Boston community via a public portal (https://portal.realtalkforchange.org/), and as an input into public dialogues with the mayoral candidates. In this study, we apply topic modeling to a subset of the RTFC corpus of transcribed conversations to surface the inferred agenda emerging from conversations, here expressed as a distribution over topics. We compare this to the distribution of topics covered in a time-matched sample of news stories published in "The Boston Globe." We apply a semi-supervised keyword extraction method to enable quantitative analysis across the conversation and news corpora. Significant differences in the topic distributions of the two corpora reflect a mismatch between how much attention the city's largest news source gives to historically underheard residents and their expressed needs and concerns. The methodology points towards a systematic way for local news organizations to consider community experiences as an input for which topics they cover and how to cover them.

### **Local, Social, and Online: Comparing the Perceptions and Impact of Local Online Groups and Local Media Pages on Facebook - Marianne Aubin Le Quéré** (Cornell Tech), Mor Naaman (Cornell Tech) and Jenna Fields (Cornell University)

With the steady closure of local newspapers, many communities have been left without reliable news and information. Technology platforms are attempting to fill the void by providing community forums or neighborhood apps where users read and share local information. Today, Facebook groups (which include buy-and-sell, local interest, or community discussion groups) are one popular form of digital local information sharing. This study investigates how local online groups are perceived compared with more traditional local news outlets, and compares the pro-community benefits provided by each. Based on prior theoretical contributions, we developed a framework for measuring the benefits of local information presence on individual-level pro-community attitudes. In our experiment (N=170), we asked frequent Facebook users living in four U.S. cities (Boston, Columbus, Nashville, Seattle) to start following local news pages or local online groups on Facebook, and compare their perceptions of quality and changes in pro-community attitudes. We find that while posts from local news pages are perceived to be of significantly higher quality than posts from local online groups, neither led to significant changes in pro-community attitudes during our study period. We discuss implications for the future study of local news in a changing media ecology.

### Storytelling Structures in Data Journalism: Introducing the Water Tower structure - **Bahareh Heravi** (University of Surrey)

Reviewing the existing and long-established storytelling structures, this paper examines the use of the storytelling structures employed in data storytelling, specifically in the context of data journalism. For this, a large set of data stories from a variety of news outlets was collected, tagged and analysed. Accordingly, and reflecting on the results, the paper proposes a new storytelling structure for data storytelling, which addresses the unique requirements of this emerging area of study and practice, called the Water Tower structure. This proposed structure is an addition to the existing storytelling structures, and is specifically designed for and targeted at storytelling with data, with a particular focus on data journalism. While this paper is primarily focused on data storytelling in journalism, the contributions are believed to be of use and value to other domains such as Business.

## [7] Contributed Papers: Collaborations

### **The "Arab Fact-Checkers Network" led by ARIJ: How we are strengthening fact-checking to support the whole journalism ecosystem in the MENA region? - Saja Mortada** (Arab Reporters for Investigative Journalism)

When Arab Reporters For Investigative Journalism (ARIJ) established in late 2020 the "Arab Fact-Checkers Network", we believed in the idea because we noticed based on a lot of studies and research the need to build a strong community of fact-checkers in the Arab world, and the huge need to support and protect in the aim of fighting the information disorder in the MENA region.

Our main goal, is to enhance pre and post publication fact-checking in a region that lacks of skills, methodologies and the use of technology that might enhance our work in Arabic language. This goal, won't only support the fact-checking ecosystem, because ARIJ works on the intersection of fact-checking, investigative journalism and whistleblowing, and without professional fact-checking, the journalism ecosystem in the Arab world is in real danger, especially because we are suffering from suppression of freedom of expression and press, and lack of access to information & data.

ARIJ, through AFCN, is pushing towards building professional, independent & transparent journalism in the MENA region.

What challenges & obstacles we are facing? how we are overcoming them? and how we are working in the intersection of investigative journalism, fact-checking and whistleblowing, with a full digital transformation and start of technology uses in Arabic language, to fight misinformation and disinformation and enhance independent and quality journalism in a region that suffers from social, political and economic crisis and that feds up with corruption and wrongdoings?

**Exploring Together: Collaboration as a bedrock of improved use of emerging technologies in journalism** - Mattia Peretti (London School of Economics and Political Science) and **Jeremy Gilbert** (Medill, Northwestern University)

Implementing emerging technologies such as artificial intelligence in newsrooms can appear daunting, dangerous and difficult. These technologies might seem complex to non-technical news workers and inaccessible to newsrooms with limited resources. So it might make sense to work with other people to see if AI could solve your problems instead of doing it all on your own.

JournalismAI at the London School of Economics and Political Science and the Knight Lab at Medill | Northwestern University have been collaborating for over two years to help newsrooms across the world do just that: teaming up across professional and geographical boundaries to explore the opportunities offered by emerging technologies to newsrooms of all types and sizes.

Join the talk to hear what we learned through jointly managing the Collab Challenges in 2021 (https://www.lse.ac.uk/media-and-communications/polis/JournalismAI/2021-Collab-Challenges) including how to manage cooperation across time zones, the challenges of linguistic differences, newsroom resources and working styles.

You will learn how the Collab Challenges employed a design thinking approach to problem-solving and how we employed natural language processing, computer vision and other AI-related technologies to investigative journalism, modular storytelling, and audience insights. You will also learn how your team(s) can collaborate with others to make the most of emerging technologies.

**Centering the humans: Leveraging OSINT, digital authentication and rooted storytelling to elevate community voices in algorithmic accountability journalism** - **Garance Burke** (The Associated Press)

Whether in health care, policing, education, banking, immigration or the courts, predictive and surveillance tools play an outsized role in deciding who in our societies wins and loses. Despite their power, these systems remain largely invisible to the public.

The Associated Press is leading an unprecedented collaboration amongst its journalists worldwide to explore how technologies fueled by artificial intelligence are impacting our communities. In this talk, the investigative journalist heading AP's global Tracked series, Garance Burke, will share findings on how to leverage different methodologies — including deep community listening, open-source investigative and authentication techniques, as well as freedom of information requests — to center the voices of people most affected by predictive and surveillance tools.

Burke will showcase the team’s investigations of two specific algorithms: one, powering a gunshot detection technology system used by police departments throughout the U.S., and another, a tool used to predict which families should be investigated for child neglect allegations.

Both stories surfaced fresh concerns about the tools’ potential to harden racial bias into policing and the child welfare system. And both stories, which shone a light on the people who build these systems as well as the families and communities they impact, led to real-world change.

Algorithmic accountability reporting relies on exhaustive computational research into models’ statistical underpinnings, and scaling these stories for a mass audience requires journalistic dexterity. This session will focus on how to marry data science with cutting-edge journalistic techniques to unravel when such tools power breakthroughs or deepen existing biases.

### **Data as distributed content for local newsroom collaborations: Lessons from 5 multi-newsroom projects - Derek Kravitz**, **Betsy Ladyzhets** and Dillon Bergin (MuckRock), and **Mohar Chatterjee**, Smarth Gupta and Eric Fan (Brown Institute for Media Innovation)

Starting in March 2020, a group of data journalists and computer science graduate students — funded through a Brown Institute Magic Grant — started work on a new way to strategically gather public records and data and distribute content to both local and national newsroom partners.

The result has been a crash course in "distributed content" for resource-strapped newsrooms, more than 150 investigative projects with 45 different partners and the creation of a new data journalism team at MuckRock, the 10-year-old public records-focused journalism site.

The initial offshoots of this work were the Documenting COVID-19 project, a public records initiative on the pandemic, and MISSING THEM, a memorial project on New Yorkers who had died during the first two years of COVID and the disproportionate impact of the virus on certain communities. Both projects won top national awards from the Online News Association, the News Leaders Association, the Society of Professional Journalists and others.

But since then, graduate students, journalism fellows and Columbia alumni have worked to expand the public records and data journalism-focused mission of our work into new areas, with more local partners — air pollution data in California and Chicago with NPR, childcare data in Michigan with Chalkbeat and the Detroit Free-Press, "Uncounted," an award-winning ongoing project on death certificate errors and COVID-19 undercounting with the USA TODAY network. Smaller projects have garnered attention as well — on opioid abuse with KQED in the Bay Area and Biden administration financial disclosures with Forbes.

We'll explore these projects with the journalists who worked on them, what it took to conceive, collect, execute, distribute and publish this work and what the future of "distributed content" looks like.

### **Collaboration on tech for journalism and other cross-domain benefits - Cheryl Phillips** (Stanford University) and **Lisa Pickoff-White** (KQED)

In recent years, we've seen a rise in journalistic collaborations, from The Panama Papers to efforts to bridge divides between academic and journalistic work. Now, we are embarking on a new type of collaboration that could transform how journalists tap into community organizations, help give them voice and still tell critical stories through that journalistic lens.

Big Local News is a core partner in the Community Law Enforcement Accountability Network, a first-of-its-kind effort to collect, parse, analyze and share law enforcement data, beginning with records that document misconduct. Big Local has worked with the California Reporting Project – a consortium of 40 newsrooms across California, partnering on a story about police officers in Bakersfield, Calif., who broke dozens of bones when they wielded their batons; and a story on how police harmed those suffering from mental illness. Now, we are working on a story that uses machine learning to find patterns in unstructured records. This latest effort is critical because journalists represent just one component of this new type of collaboration. The Berkeley Institute for Data Science, the National Association for Criminal Defense Lawyers, the ACLU, and community advocates are all a part of the Accountability Network. Together, we are building a computational infrastructure that will enable any partner to effectively analyze police interactions. Complex and intractable problems mean we must collaborate across domains to find solutions for all parties. That means policy-making stories by journalists, access to records and the patterns within them for community organizations, attorneys and policy groups.

## [8] Contributed Papers: Stories I

### OCR Optimized for Images Created by Typesetting - Norihiko Sawa and Masaki Aota (Nikkei)

We developed a kind of Optical Character Recognition ( OCR ) system specially optimized to recognize images created by typesetting.

Accuracy by our method measured 0.97222 vs. 0.21322 by Google's Cloud Vision on images of Nikkei's replicas of print editions. The accuracy of detection of character rectangles is defined as the number of detected rectangles divided by the number of rectangles which do not cross over the bodies of characters.

As a result of its ability to recognize text characters and positioning perfectly, Nikkei's product enables users of the print edition replica app to freely highlight and copy text to clipboard.

Existing OCR engines do not extract text well on newspaper pages because of the narrow margins between lines and characters and its vertical writing.

More specifically, while conventional OCR can be applied to a flexible character arrangement, our method assumes that characters will be arranged in blocks, and detects rectangular blocks in the following order: (1) paragraphs, (2) lines, and (3) characters.

Morphological Transformations is used to find paragraphs and Loss function is to find positions of characters. Dynamic Programming and Levenshtein distance help to obtain mappings between horizontal text and detected block positions, to determine paragraph order, and to do error correction.

Replicas of print editions are just images, therefore it is difficult to add features which enhance value. It is necessary to know the accurate positions of each character within the page image to be able to develop features which improve the product and user experience.

### **Complementing language models to improve their applicability to journalism - Jeffrey Nickerson** (Stevens Institute of Technology)

Language models, of which GPT-3 and BERT are currently the most well known, are permeating both industry and academia. They are potentially applicable to journalism, because they can generate ideas, questions, summaries, and other snippets of readable text. They are useful, but can be more useful if their capabilities are augmented by other capabilities, either computational or human. This talk considers what is lacking, how they can currently be augmented by existing technology, and what implications these combinations may have on the future of journalism. In particular, these shortcomings of language models are considered: the inability to separate fact from fiction, the lack of temporal currency, the inability to remember a continuing series of interactions, and the lack of transparency. The inability to separate fact from fiction might be addressed with fact verification technologies. Many of these are based on knowledge graphs, and might provide the grounding the language models lack. The lack of currency, due to the lags between the training and deployment of language models, can be augmented by accessing knowledge bases such as Wikidata, as well as news-oriented knowledge bases. Issues of coherence and memory are more challenging. Recent work has shown that creating outlines and then feeding specific prompts to the model and reassembling the component answers mitigates some of the limitations of language models. Addressing transparency might involve a combination of the mitigating technologies just discussed, with knowledge graph probes functioning to ground and possibly explain the generated ideas. This talk reviews these issues and proposes a research agenda.

### **Developing low-code tools and pipelines for local data storytelling** - Sam Gross, Rob Powell, **Matt Albasi**, Emilia Ruzicka, Nick Devlin and Elena Cox (Stacker)

This session would explore tactics for creating and distributed localized data journalism at scale, building on the tech stack we developed at Stacker and several projects by our data reporters spanning Census, real estate, weather, sports, and other story series.

Session would cover building reproducible data projects and story frameworks, developing locational taxonomies, employing no/low-code tools for computational analysis across the newsroom, and crafting distribution strategies to maximize access to coverage.

This approach to local data journalism was honored as one of three for finalists for this year's Mega Innovation Award from America's Newspapers.

https://blog.stacker.com/how-our-story-tech-stack-delivers-journalism-to-1-000-news-organizations/
https://blog.stacker.com/introducing-stacker-local/
https://www.editorandpublisher.com/stories/from-2022megaconf-mega-innovation-award-finalist-stacker,224483

### **NewsEdits: A Roadmap for Computational Journalism - Alexander Spangher,** Xiang Ren and Jonathan May (University of Southern California), and Nanyun Peng (UCLA)

We present the NewsEdits corpus, a large corpus of news article revision histories (i.e. every time a news article gets revised online). NewsEdits is large-scale and multilingual; it contains 1.2 million articles with 4.6 million versions from over 22 English- and French-language newspaper sources based in three countries, spanning 15 years of coverage (2006-2021). It contains articles being updated with style-based updates and fact-based updates: to underscore the factual nature of many edits, we conduct analyses showing that edited parts of articles are more likely to contain updating events, main content and quotes than unchanged parts. Our vision is for this corpus to be used in computational journalism research to develop fundamental tools that can help journalists cover news in breaking news scenarios. For example, this corpus can lead to tools that recommend the right source at the right time, restructure the narrative of a breaking piece, or predict which events are likely to update. To defend this vision, we show that edits are predictable and follow patterns. We develop three predictive tasks aimed at predicting edits performed during version updates and show that these tasks are possible for expert humans but are challenging for large NLP models.

## Paper Sessions — Saturday, June 10, 10:00 - 11:20 AM EDT

### [9] Refereed Papers: Data and/as the news — Lecture Hall

**Cataloging Algorithmic Decision Making in the U.S. Government - Grace Lee**, **Jasmine Sinchai**, Daniel Trielli and Nicholas Diakopoulos (Northwestern University)

Government use of algorithmic decision-making (ADM) systems is widespread and diverse, and holding these increasingly high-impact, often opaque government algorithms accountable presents a number of challenges. Some European governments have launched registries of ADM systems used in public services, and some transparency initiatives exist for algorithms in specific areas of the United States government; however, the U.S. lacks an overarching registry that catalogs algorithms in use for public-service delivery throughout the government. This paper conducts an inductive thematic analysis of over 700 government ADM systems cataloged by the Algorithm Tips database in an effort to describe the various ways government algorithms might be understood and inform downstream uses of such an algorithmic catalog. We describe the challenge of government algorithm accountability, the Algorithm Tips database and method for conducting a thematic analysis, and the themes of topics and issues, levels of sophistication, interfaces, and utilities of U.S. government algorithms that emerge. Through these themes, we contribute several different descriptions of government algorithm use across the U.S. and at federal, state, and local levels which can inform stakeholders such as journalists, members of civil society, or government policymakers.

**News as Data for Activists: a case study in feminicide counterdata production - Rahul Bhargava** (Northeastern University), Harini Suresh (Data + Feminism Lab & CSAIL, MIT), Amelia Lee Doğan (Data + Feminism Lab, MIT), Wonyoung So (Data + Feminism Lab & DUSP, MIT), Helena Suárez Val (Feminicidio Uruguay & CIM Warwick), Silvana Fumega (ILDA) and Catherine D'Ignazio (Data + Feminism Lab & DUSP, MIT)

News articles are an important source of data for recording and aggregating a range of social phenomena. In this paper, we ask if and how technology can support civil society activists who challenge asymmetrical power relations by producing counterdata—datasets missing from mainstream counting institutions. We consider a case study centered on activists who monitor feminicide, or the lethal outcome of gender-related violence, often using news as a main source to identify and compile databases of incidents. We describe a system that we collaboratively built with activists, aimed at relieving some of the emotional and time-intensive labor this work entails. The system discovers relevant news stories on multiple systems, classifies them based on machine learning models, clusters them into groups of stories about the same incident, and delivers regular email alerts to users. Currently, 26 groups across different geographical regions are using the system, and groups who broadly monitor feminicide report that they are regularly discovering new cases. We also reflect on the short-comings of the pilot system for groups with more specific, intersectional monitoring focuses, and the implications of biased narratives or under-reporting on the system's design. This case study contributes a grounded example of computational journalism built in collaboration with, and in service of, activists working on critical human rights issues.

**Characterizing Social Movement Narratives in Online Communities: The 2021 Cuban Protests on Reddit** - **Brian Keith** (Virginia Tech), **Tanushree Mitra** (University of Washington) and Chris North

Social movements are dominated by storytelling, as narratives play a key role in how communities involved in these movements shape their identities. Thus, recognizing the accepted narratives of different communities is central to understanding social movements. In this context, journalists face the challenge of making sense of these emerging narratives in social media when they seek to report social protests. Thus, they would benefit from support tools that allow them to identify and explore such narratives. In this work, we propose a narrative extraction algorithm from social media that incorporates the concept of community acceptance. Using our method, we study the 2021 Cuban protests and characterize five relevant communities. The extracted narratives differ in both structure and content across communities. Our work has implications in the study of social movements, intelligence analysis, computational journalism, and misinformation research.

## [10] Contributed Papers: Automation — Brown Institute

**Automated Text-Based Classification of Political Campaign Expenses** - **Dylan Freedman** and **Lenny Bronner** (The Washington Post)

Campaign finance data is a critical resource for reporters and researchers in understanding how campaigns spend their money and where their support comes from. Unfortunately, this data is primarily text-based and difficult to analyze. In this talk, we present an approach to automatically classify poorly labeled campaign expenses into categories, like advertising, fundraising, travel, etc. We collected 81 million Federal Election Commission (FEC) expense records from 2016 through present and trained unsupervised text embeddings on this data. Using active learning via clustering and manually labeling a subset of data, we created a model that classifies expenses with an accuracy of over 75%. We then use this model to create a spending snapshot of active campaigns in the midterms. This talk will demonstrate how this approach works and how it can be adapted to other domains to derive insights in messy, text-based data.

**Automated news generation based on structured data of Russian local election results** - **Pavel Lebedev** (National Research University - Higher School of Economics)

Automated journalism is being used to cover local elections in the USA, UK, Finland and other countries. Prior to my research, algorithms for election coverage had not been developed by Russian newsrooms and universities, despite the need for widespread publicity about election violations.

I developed a news generation algorithm based on structured data about Russian elections so that journalists could quickly deliver news to readers based on geolocation and identify voting anomalies at electoral precincts.

In my work, I apply methods to design and analyze the resulting automated texts and news stories written by journalists. The algorithm generates short news stories based on facts and patterns. The generator uses pattern based model, a dictionary of predicates and morphological transformations.

A manual check of the program showed that 64% of the written texts had no significant flaws and only 2% had one or more obvious errors. The BLEURT score was 0.68. The algorithm can be used as a tool to automatically generate news about local election results and serve as an additional source of information for journalists. It allows quick identification of election irregularities and makes them public.

### **Designing and Deploying AI-Driven Lead Discovery Systems for Science Journalists** - **Sachita Nishal** and Nicholas Diakopoulos (Northwestern University)

Science journalists in the modern day are confronted by new time pressures that can constrain how they discover, investigate, and write about newsworthy and interesting research. These time pressures stem from the need to monitor the increasing volume of scientific research across journals, conferences, and preprint servers. Creating news stories for internet audiences also necessitates speed in the journalistic process, even as science journalists take on a variety of new roles in society, including as curators, civic educators, watchdogs, and so on. In this work, we propose the design and deployment of an AI-driven pipeline for lead discovery that can provide time and information subsidies to journalists during the news-gathering process. This pipeline is designed to recommend newsworthy scientific research from the arXiv pre-print server using a machine learning model trained on crowdsourced evaluations of past publications, and then filter and rank these recommendations based on their relevance to the specific venues where an individual science journalist wants to publish stories. We aim to evaluate the performance of this pipeline by deploying it to professional science journalists over some time period, and conducting semi-structured interviews to gain insight into how they use the system and what value it creates for them. Our aim for this work is to reduce the time and attention cost of lead evaluation, while being cognizant of journalists' resource constraints, autonomy, and agency such that more and better science communication ultimately results.

### **Archeologies of Data in Contemporary Journalism: the digital afterlives of newspapers' photo morgues - Giulia Taurino**, **Si Wu** and David Smith (Northeastern University)

Before the advent of digital photography, newsrooms processed images differently, including photo librarians and editors working in local repositories disjointed from textual resources. Published and unpublished photographs from past photo assignments and issues were stored in archives known as "photo morgues", often acquired by academic institutions, public libraries, or museums outside of the media industry. These files - including newspaper clippings, printed photographs, and unpublished negatives - have remained in storage, mostly accessed by newspaper librarians, until recent digitization efforts. This talk addresses how digitized photo morgues can be deployed in journalism studies to investigate archives as narratives, beyond single-image retrieval. It also shows how old photo-journalism images can be turned into accessible data for present-day journalistic inquiry and reporting. We take as a case study a project that uses machine learning techniques to revive the techno-cultural value of newspapers' photo morgues, based on the Boston Globe Archive. We notably propose methods that can be applied to other photo archives to facilitate the access to historical data in news coverage, as well as to improve our understanding of the selection - or suppression - process that led to current practices in journalism. By taking into account their classification systems, publication history, visual patterns, and handwritten annotations, this project offers an alternative perspective on the use of data-led practices and AI techniques in the media value chain, thus reinserting newspapers' archival practices in the conversation about the future of computational journalism - from libraries back into news and media systems.

## [11] Contributed Papers: Supporting Data Journalism — World Room

### A multi-country computational analysis of press-state-online citizen relationship - Jean Dinco (UNSW)

While social media played a central role in spreading democratic ideas during the Arab Spring; and in shaping political debates in China and Viet Nam, these findings cannot be applied in countries with a different political system. This paper argues that the emancipatory effects of social media are contingent on the country's political climate in question (Bailard, 2014). This means that if a government has restricted political rights and freedom, the positive effects of social media as a tool for change are amplified. This is because social media operates as the main, if not the only, outlet for expressing a political opinion (Barnidge, Huber, de Zuniga, & Liu, 2018, p. 166; Bolsover, 2018). This research study contributes to the body of knowledge on opinion manipulation in the Twitter network of citizens in India, Bangladesh, and the United States. To determine the extent of influence of government frames over citizen frames, I applied text-as-data methods to identify issue-specific frames. These frames undercover the main overarching communication strategies for the three countries: us-versus-them for India and Bangladesh and global responsibility for the United States. I conducted a cross-lagged correlation analysis to evaluate the alignment of frames between government and online citizens. The counts of government publications and citizen tweets in each frame category were aggregated by month. Pearson R correlation was calculated between the distribution of frames. The results contradict the overarching narratives that online communication is emancipatory and raise a point of inquiry on the hegemonic role of social media in democratic states. By concentrating on the functions of social media as a radical instrument for change in authoritarian states, scholars have overlooked the potential of social media to be used for propaganda purposes in democratic societies.

### "Open Data, Open Code, Open Knowledge:" Making Public Intent Data More Accessible and Discoverable through Visualizations and Storytelling - Tony Fujs (World Bank Group)

The World Bank's Development Data Group is committed to putting data to work for development. We do data "from farm to table," meaning we are involved in the entire data cycle, including production, collection, analysis, as well as dissemination.

Given one of our key strategic priorities is "open data, open code, open knowledge," transparency and accountability are at the core of our work. We strive to base our data products on open data and to make the code open source and our results reproducible. We put a lot of effort into to presenting our data in an easily accessible and actionable format. And we are moving beyond just opening up our data to opening up our analytics as well so that the public can reap the benefits of open knowledge.

With 5,139 datasets in our public data catalog, 20,000 indicators, and more than 30 million people a year coming to us for data on social and economic development, we engage both in computational research as well as data journalism. Through state-of-the-art visualizations and impactful storytelling (and "scrollytelling"), we constantly strive to make data more accessible and discoverable to different stakeholders.

The proposed talk would explore how we bridge disciplines through illustrative examples and lessons learned from developing some of our flagship products, including the Atlas of Sustainable Development Goals 2020 and the new Poverty and Inequality Platform. It would also raise awareness about our extensive sets of data for good among leading data journalists, scientists, and students of our time.

### **The Emergence of the Occupation of Data Journalist and its Implications for the Future of Computational Journalism** - **Keren Henderson**, Stan Jastrzebski and Kevin Crowston (Syracuse University)

Occupations are defined in part by the practice of specialized skills and deployment of specialized knowledge. As a result, new technologies often mandate new occupations to manage them. This emergence is reflected in the pattern of adopting new journalism technologies. For instance, "TV satellite truck operator" became an occupation along with the expectation to remotely transmit broadcasts.

However, over time, financial and technological barriers between occupations can shrink. This change allows managers to fold emerging skills into existing jobs. For instance, portable live video backpacks obviated the need for satellite uplinks and the trucks providing them. In cases such as these, a role may disappear, as did the satellite truck operator. On the other hand, some technologies resist simplification and so retain the occupation, such as "broadcast engineer."

In our current project, we are interviewing journalists about their use of technology, with the goal of understanding how their work is changing. One finding from this project is that the developing occupation of data journalist is currently demarcated by the specialized training required. This study will look at whether this occupation of gathering, analyzing and visualizing data fits the pattern of transitory newsroom job titles or might persist.

### **Newsroom innovation labs as 'survival entities' for journalism? Mapping the process of institutionalization at The Washington Post** - **Hannes Cools**, Baldwin Van Gorp and Michael Opgenhaffen (KU Leuven)

Newsroom innovation labs have been heralded as ‘safe testing environments’ in the news ecosystem, because they have the potential to improve where news goes and moves. At The Washington Post, these labs have been installed for three years. In the meantime, they have been developing tools like election models, fueled by AI and smart data pipelines that can possibly affect the autonomy and the tasks of the broader newsroom. By applying the process of institutionalization as a theoretical and an analytical framework, this case study enhances our understanding of how the interactions take place between these rather novel innovation labs and the broader, more established newsroom at The Washington Post. To do so, we have utilized digital ethnography in the form of newsroom observations and expert interviews with members from both the innovation labs and the broader newsroom to evaluate what forms of interaction obstruct or drive the way these tools are being (dis)trusted. Results reveal how the innovation labs justify – via interaction – the implementation of the tools to the broader newsroom and vice versa. A minority of the interviewed members of the editorial board expressed their trust in the tools. Therefore, the process of institutionalization is still in a nascent state. The journalists in the broader newsroom who did express their confidence form a heterogeneous group of techsavvy staff members. In order to increase trust across the newsroom, it seems to be essential for these innovation labs to be more transparent on the workflows and the specific output of these tools.

## Invited Session — Saturday, June 10, 11:30AM - 12:50PM EDT

## [12] Reconstructing the Built Environment — Lecture Hall

### Reconstructing the Neighborhood Destroyed by the Tulsa Race Massacre - Anjali Singhvi (The New York Times)

The New York Times analyzed archival maps, city directories, census data, newspaper clippings and survivor accounts to create a 3D visualization of Greenwood, a prosperous Black neighborhood in Tulsa, Okla., that perished at the hands of a violent white mob in 1921.

A machine-learning algorithm was trained to recognize building footprints in the 1920 Sanborn Fire Insurance Map and convert them into 3D geometry. Custom-written software allowed us to manually enter building heights based on data present in the map.

To give our readers a more intimate glimpse of the neighborhood, we manually recreated one block of Greenwood Avenue in greater detail, based on a meticulous review of archival photographs. When no visual references were available, the buildings were left as simple shapes, to avoid misrepresenting the architecture. Details like cars and a street trolley were added based on further historical research.

To give a fuller picture of the people of Greenwood, we analyzed and mapped census data to show where each Black resident lived and what they did for work. Thousands of addresses from the 1920 Census were geolocated to show the types of occupations that were held by Greenwood residents.

A dataset of Black businesses and residents was created by running the digitized pages of the 1921 Tulsa City Directory through an Optical Character Recognition process and searching the results for a “(c)” at the end of an entry (used by the directory to identify businesses and residents of color). Additional information about businesses and proprietors on our featured block was gathered through manual review of archival newspapers and documents.

With this talk, I will cover the research and technical work involved in creating this NYT interactive: https://www.nytimes.com/interactive/2021/05/24/us/tulsa-race-massacre.html

### **Forensic engineering and modeling the Surfside condo collapse - Sarah Blaskey**, Ben Conarck, Aaron Leibowitz (Miami Herald) and Dawn Lehman (University of Washington)

A behind-the-scenes look at the Miami Herald’s prize-winning forensic investigation into one of the deadliest building collapses in modern history. This session will discuss how reporters collaborated with structural engineers and used the techniques of traditional investigative journalism — conducting interviews with survivors and examining records including photos, videos, call logs, body cam footage and inspection reports — to inform a finite element analysis built in LS Dyna to identify what went so terribly wrong.

## Paper Sessions — Saturday, June 10, 01:50 PM - 03:10 PM EDT

## [13] Contributed Papers: AI and the Newsroom — Lecture Hall

### **Algorithms in the news: challenges and recommendations for an artificial intelligence with the ethical values of journalism** - Patricia Ventura-Pocino (Universitat Autònoma de Barcelona)

The use of artificial intelligence (AI) in the media is already a reality and the integration process is expected to become faster and further consolidated in the coming years. Recent reports and surveys indicate that the industry will make even greater use of AI applications and that media will increasingly adopt this technology. In the main media algorithms are already featuring in processes throughout the value chain, and the advantages that the sector perceives with regard to its potential to optimize internal workflows and the dissemination of content suggest a major transformation to journalistic routines in the immediate future.

Today it is common to delegate to algorithms such tasks as identifying newsworthy topics, analyzing and organizing source data, facilitating transcription, translations and similar processes, generating written content and infographics, choosing titles, guiding the process of writing journalistic content, moderating comments, publishing on behalf of the organization on social media accounts, customizing and recommending content to users, and many others. We have already reached the point where we can ask AI things like: What is newsworthy? What form should it take? What title to choose? And what content to highlight? In other words, AI can play a key role in decisions that are at the very core of journalism's editorial function.

This communication exposes the report made for the Catalan Media Council in which the challenges are analyzed and a series of recommendations are offered to use the algorithms with editorial criteria. The work is based on a critical analysis of the digital media ecosystem and proposes solutions to endow algorithms with principles so they can be put at the service of a digital public sphere that is governed by democratic values.

Link to the report: https://bit.ly/ethicsaijour

### **Interactive Machine Teaching: Towards Accessible Machine Learning for News Media** - **Swati Mishra** and Jeff Rzeszotarski (Cornell University)

Modern news media organizations actively employ Machine Learning (ML) algorithms to assist with the news generation process. These algorithms have the ability to automatically learn underlying relationships within data variables, and help to power applications that identify newsworthy events from large data, automatically generate stories, and assist writers in reducing polarization among media consumers. However, these ML-based systems are often built by ML experts; people with significant knowledge about the ML processes but little or no expertise in journalistic data inquiry. As a result, these models lack the domain-specific knowledge that is critical for deploying them. For instance, a classification model designed to identify "newsworthy" events from Twitter feed, may inadvertently latch on to retweet frequency. For news reporting, this might be undesirable behavior as it leads to biased reporting, or in some cases, even redundant reporting (why report an event that people already know about).

In this talk, we argue that building tools that allow ML non-experts to directly inform the design of ML-systems can help bridge this gap. We first draw on the existing literature on interactive machine learning and data visualization to formulate how non-experts make sense of ML models and their design process. We then describe the design dimensions of interactive tools that can facilitate customization, refinement and adoption of ML models. By grounding these dimensions in context to news media, we pave way for human-centered design of machine teaching tools targeted for computational journalism.

### **A practical approach for Developing Editorial Algorithms with Active Learning - Francesco Marconi, Eric Bolton** and Erin Riglin (Applied XL)

The ever-rising sea of data presents both a dramatic challenge and, up to this point, a largely untapped opportunity for the news industry. Journalism needs new methods to interpret and contextualize the enormous amount of data our society is poised to produce. We need faster and more sophisticated information gathering if we don’t want to drown in big data.

News-gathering in this news landscape requires a scaled approach that is not currently supported by traditional newsrooms, due to the need for monitoring multiple data sources simultaneously and at a pace surpassing manual capabilities.

The same way journalists vet stories produced by humans, they should be able to validate algorithmic output in an effective manner. Active learning, an often overlooked subdiscipline of machine learning, lends itself perfectly to this kind of human-in-the-loop validation, through methods meant to elicit predictions a model is most likely to get wrong. This lowers the amount of effort required to identify potentially incorrect predictions.

Depending on the task, active learning can also significantly accelerate the collection of new datasets. A good example is named entity recognition, where prioritizing samples to label using active learning can allow a model to reach the same level of precision on a dataset a quarter the size of one annotated in random order. This enables a rapid workflow that can be essential for journalists less focused on deploying long-term machine learning applications, but rather looking to tailor models to specific tasks relevant to their reporting.

Thanks to this, active learning is enabling us to scale journalistic principles in the development of our algorithms by mitigating unwanted AI biases and ensuring editorially sound information.

In this session, computational journalists Francesco Marconi, Eric Bolton, and Erin Riglin will share practical principles and best practices for building autonomous news-gathering systems.

### **AI ≥ Journalism: How the Chinese Copyright Law protects tech giants' AI innovations and disrupts the journalistic institution** - **Joanne Kuai**, Raul Ferrer-Conill and Michael Karlsson (Karlstad University)

Journalism and other institutions clash over automated news generation, algorithmic distribution and content ownership all over the world. AI policies are the main mechanisms for establishing and organizing the hierarchies among these institutions. Few studies, however, have explored the normative dimension of AI in policymaking in journalism, especially beyond the West. This case study inspects the copyright law's impact on AI innovation in newsrooms in the unexamined Chinese context. Using neo-institutional theory and policy network theory, the study investigates the third amendment to Chinese Copyright Law, exemplary court cases regarding automated journalism copyright disputes (such as Tencent v. Yingxun and Film v. Baidu), and other supporting documents. The findings show how China's copyright legal framework separates authorship and ownership; defines "originality" and "creativity" in human-machine collaboration; and prioritizes tech companies while undermining journalism autonomy. We argue that the law's eager embrace of AI may give tech companies an advantage over news organizations who do not necessarily have a strategy to adopt AI. Moreover, it favours state-owned, resource-rich official media over the private sector. An implication of this shifting power dynamic is the possibility of privately-owned news media being marginalized, resulting in even stronger state control over media production and information flow.

## [14] Contributed Papers: Stories II — Brown Institute

### **Analyzing News Frames: A Survey of Computational Approaches** - Naeemul Hassan and **Mohammad Ali** (University of Maryland, College Park)

Analyzing news frames became a prominent area of academic research in both social and computer science disciplines (D'Angelo, 2018; Guo et al., 2019). Traditionally, researchers explore frames using qualitative and quantitative approaches that require manual labor and can handle comparatively a small amount of data. With the recent advancement of computational tools and easy access to large volumes of texts, scholars started exploring news frames utilizing various computational tools in the last decade (Card et al., 2015; Guo et al., 2019; Walter & Ophir, 2019).

While manual framing analysis is a challenging task due to the lack of unified conceptualization of a frame and its analysis methods, computational approaches are in the nascent stage that drew scholars' attention for further improvement. To contribute to the ongoing discussion and methodological development in framing analysis computationally, this current paper surveyed eight distinct computational framing analysis approaches and tools. These are policy frames codebook (Boydstun et al., 2014), media frames corpus (Card et al., 2015), gun violence frame corpus (Liu et al., 2019), topic modeling (DiMaggio et al., 2013), structured topic modeling (Gilardi et al., 2021), hierarchical topic modeling (Nguyen, 2015), cluster analysis (Burscher et al., 2016).

We discussed the survey findings mainly in terms of 1) conceptualization of frames, 2) development and application of computational framing analysis approaches, and 3) future research directions. We identified a discrepancy that the approaches conceptualized frames with prominent definitions but mainly were applied to explore “topics” instead of frames. The results were discussed along with future research directions.

### A study on a face-embedded data visualization thumbnail of a news article - Joohee Kim and Sungahn Ko (Ulsan National Institute of Science and Technology)

Thumbnail, as a small-sized image delivering the summary of the content, is designed to attract the instant attention of viewers or readers. With the vast volume of data, news organizations include multiple sets of data visualization in their articles and exhibit the representative data image in the thumbnail. Among various design elements in data-driven thumbnails, some designers choose to embed a face of a famous figure, who is related to the topic, to produce a clickbait thumbnail. Similar to the framing in news titles, we hypothesize the insertion of an image can cause bias by emphasizing one facet in data interpretation. By conducting a survey, we verified that a portrait affects readers' interpretation of the data visualization. Specifically, the facial expression of the figure can contribute to understanding the trend of the graph in the manner of suggesting the tone and mood. Furthermore, the use of a certain figure may provide the reasons for reading the article as people understand the context of the visualization with the image in the thumbnail. With the finding that the image of the figure significantly contributes to the viewer’s perceived bias, we suggest an approach to computing the alignment between data trend and the figure's facial expression. The design implication for measuring the degree of possible bias in thumbnail would help thumbnail designers to deliver their intention more effectively considering the perspective of readers.

### Familiarization and self-reflection strategies in interactive data visualizations - Olga Lopes (Universidade Federal de Santa Catarina)

One of the challenges concerning digital data journalism is finding ways to address topics of public interest and maximize their circulation and impact. This work seeks to observe the argumentative strategies used in data stories that employ interactive personalization resources in visualizations to situate readers in articles that deal with complex phenomena. To understand how this customization has been applied, we conducted an exploratory content analysis in a sample of 40 narratives to observe how these projects incorporate and project users' inputs. The articles observed were published between 2018 and 2020 and collected in the weekly newsletter Data Journalism Top 10, organized by the Global Investigative Journalism Network, and on the Visualising Data blog. By analysing these examples we also discuss the role of visual rhetoric and narrative approaches found in the specialized literature and the integration of user/content interaction as part of a broader set of editorial judgments used to convey meanings in data visualizations. It was possible to verify among the analyzed datavis a prevalence of the joint use of more than one type of control to record user inputs, the presence of fast feedback loops in response to these interactions and narrative patterns linked to geographic and demographic approximation, which allows readers to identify individuals who share similar characteristics with them. Examining the changes underlying the production of interactive data visualization we acquired a deeper understanding of the contextualization strategies concerning personalization techniques that can be later replicated and improved on.

**Memory, Emotions and Heuristics: How brain functions can affect data journalism perception** - **Elina Makri** (Aristotle University of Thessaloniki)

Except the well-studied journalistic functions related to data collection methods, data cleaning and analysis as well as effective visualisations, underlying mental mechanisms are crucial for building a story, weighting on evidence and engaging in sense-making out of past events.

Evidence suggests that a better understanding of the workings of the human brain and decision-making as well as of cognitive control, emotions and psychological dynamics, may provide important insights for the data storytellers.

It could be more broadly argued that humans are not particularly good at intuitively collecting and using frequency information and probably there is less importance in studying the processes of the brain as relevant for [data] journalism. However, research on memory, the recent advances on the science of heuristics as one of the major tools for modelling decision making along with logic and statistics, the increasingly pervasive algorithms and related software, makes me believe that it is worth examining data-driven research and storytelling from a cognitive science and psychology approach in order to include perspectives from these disciplines.

I will refer to some remarkable cases and research that tests formal models of heuristic inference in election prediction, health care (statistical models vs simple human heuristics), law enforcement (CrimeStat) and legal institutions.

How reliable is a regression model than a fast brain strategy that ignores part of the information especially when parts of the information is unknown which is often the case in our uncertain, real world?

I will also talk about how memory's errors and how emotions can overshadow data.

# Workshop Schedule

### Thursday, June 9 — 2:00 - 3:15 PM EDT

**[A] Numbers for audience understanding** - Laura Santhanam (PBS NewsHour), **Jena Barchas-Lichtenstein** (Knology), John Voiklis (Knology), Erica Hendry (PBS NewsHour) and Travis Daub (PBS NewsHour)

There are tons of resources to help journalists get the numbers right...but right isn't enough if the audience doesn't get the point. In this workshop, participants will have hands-on opportunities to (a) interpret unfamiliar numbers to understand the audience perspective and (b) present numbers in ways that actually help audience members make sense of them. All of our recommendations were developed through a close collaboration between social scientists and journalists in various roles within a large public media organization. The workshop will be co-facilitated by at least two journalists (one data producer and one person in an editorial role) and two researchers; at least half of them will be women. Participants will have the chance to engage in small groups to give one another feedback as they consider how to present a wide range of numbers to a general audience. They will walk away with questions to ask themselves about statistics in news stories and strategies for writing and visualizing them that supports audience understanding. We will also share explainers we have developed that can be linked from other stories to support that understanding. This session will be valuable for anyone who writes, edits, or reviews stories that contain numbers.

**[B] AI for Everyone: Learnings from the Local News Challenge** - **Swapneel Mehta** (New York University), **Christopher Brennan** (Overtone AI), **Zhouhan Chen** (New York University) and Matt Macvey (NYC Media Lab)

Journalists and media organizations have increasingly embraced AI to offer a special kind of value to their readers, listeners, and viewers, but what are the advances that can help smaller outlets with resource constraints, focused on their local communities? NYC Media Lab organized the AI Local News Challenge this spring to bring together five teams each approaching the issue in a different way, whether through the lens of localized stories for a network of newspapers, monitoring social network interactions, tracking misinformation, curating social conversation or creating qualitative data on news articles.

This session will see the presentation of results from the teams: Gannett-Localizer, Information Tracer, Overtone, and SimPPL, focusing on their work this spring. They will share their solutions as well as their learnings from months enmeshing themselves at the intersection of local news and technological advances. All used techniques centered around artificial intelligence, including natural language processing, text generation, cybersecurity, simulation intelligence, and large-scale data analytics.

Afterwards, the discussants will offer their own ideas to dig into the complexities of questions on topics such as “what does local really mean?” and “what technologies are best able to help newsrooms that are struggling?”. The participants will also welcome questions and comments from the audience to delve further into these topics, as well as discuss the practical implications of adopting these innovations at scale.

## Thursday, June 9 — 3:30 - 4:45 PM EDT

**[C] Identifying social media manipulation with OSoMe tools** - **Kai-Cheng Yang** and Christopher Torres-Lugo (Observatory on Social Media, Indiana University)

As social media become the major platforms for discussions of important topics such as national politics, public health, and environmental policy, there is a growing concern about the manipulation of these information ecosystems and their users. Malicious behaviors include astroturf, amplification of misinformation, and trolling. Such abuses can be carried out by humans and social bots --- inauthentic accounts controlled in part by software. The resulting biased reality can fool even professionals. While journalists and researchers are increasingly interested in detecting and studying these malicious activities, there are serious challenges. First, the collection and analysis of data from social media require significant storage and computing resources. Second, knowledge, experience, and advanced computational skills are necessary to find patterns and signals of suspicious behaviors in the collected data. In this workshop, we will present free tools that aim to help researchers, journalists, and the general public combat online manipulation from the Observatory on Social Media (OSoMe) at Indiana University (osome.iu.edu/tools). We will focus on Botometer that helps detect social bots on Twitter and Hoaxy that can track and visualize the diffusion of misinformation; other useful tools from the OSoMe family will also be covered. These tools utilize the massive social media data collected by

OSoMe and they are equipped with state-of-the-art algorithms and user-friendly interfaces. They also provide public APIs to allow querying in bulk. They have helped thousands of users and served as the foundation for hundreds of research projects.

### **[D] Temporal Topologies – Making journalistic authorship transparent through in-depth tagging of newsworthy events** - Francesca Morini (Fachhochschule Potsdam)

Journalistic writing is a complex and delicate form of authorship encompassing various techniques to organize, make sense, and effectively craft stories to inform readers. Through their sense-making process, journalists decide which events are newsworthy, as well as frame them meaningfully. By choosing the news angle, journalists work to establish and render unique constellations of related events cohesively. This relational contextualization of events is what stays implicitly at the core of journalistic authorship and relies on journalists' commitment to truth and transparency. As readers struggle to orient themselves through increasingly nuanced, globally entangled phenomena, in which way can journalistic authorship become their guiding thread? We engage with this question by looking at journalistic authorship as an actionable concept for data visualization. Starting from Digital Humanities, this workshop combines analog sketching and collaging with computational thinking through the use of standardized ontologies for the Semantic Web – e.g. OWL Time (Cox & Little, 2020) – and temporal reasoning techniques like Allen's interval algebra (1983). Participants will be asked to structure a story from an archive of textual material working solely on temporal and relational instances. The workshop targets journalists interested in approaching unstructured text with diagrammatic thinking. The goal of the exercise is to produce the temporal topology of one newsworthy event: its temporal layout, extensions, and relations among parts. Participants will gain knowledge on consistent tagging languages for deep-time relational entities and will explore the act of authoring, its inflections, and the implications of exposing their sense-making process to readers.

### **[E] Using Networks Analysis and Visualization to Explain COVID-19 Spread through the Physical, Social, and Information Graphs** - Hong Qu (Harvard Kennedy School)

Understanding network science and compartmental models in epidemiology are essential to understanding, explaining, intervening, and forecasting the pandemic. This workshop introduces core concepts and skills for network data analysis and visualization using Flourish and NetworkX. We begin with simple data sets that represent the flow of migration between countries using sankey, chord, correlation matrix heatmaps and, thereby, introduce the notion of nodes and edges in a graph. Next, we introduce network models such as small-world and scale-free networks to describe typical properties of complex networks, and demonstrate how to use Flourish to produce network visualizations, encoding the size of nodes as centrality and thickness of the links as edge weight. Then, we use NetworkX to analyze network statistics and run a simple agent-based model simulation of misinformation diffusion based on a simplified SIR epidemic model. By the end of the workshop, participants will fully appreciate the network science and design principles that inspired the famous visual story created by Harry Stevens for the Washington Post: 'Why outbreaks like coronavirus spread exponentially, and how to "flatten the curve"'. We end with a brainstorming discussion of ways to combine interaction of three different layers in the multiplex network--physical mobility, social relationship, and infodemic beliefs--drive outcomes and trajectories for the ongoing pandemic.

**Thursday, June 9 — 5:00 - 6:15 PM EDT**

**[F] How can AI help you? Explore writing with a machine - Lydia Chilton** (Columbia University, Computer Science)

Groundbreaking AI has enabled computers to write; it can finish your sentences in Gmail, but it can also finish paragraphs of text and write fiction on its own. Although there is a lot of hype around AI (both positive and negative), the truth is somewhere in between. We'll demonstrate how we use AI to summarize, find background, and discover story angles. Then we'll give you keys to the latest text-generating AI, and we can explore together the good, the bad, and the ugly of writing with a machine.

**[G] Share a Story: A Daily Practice for Teaching Healthy News Habits - Blake Eskin** (Journalism + Design)

Share a Story is an application for journalism educators and their students to work on mindful news consumption and production while building community in their courses. Inspired by social reading apps, 100-day design projects, and Duolingo, Share a Story has been developed over the past five years with an undergraduate cohort that is passionate about storytelling but often lacking confidence in computational thinking. Share a Story aims to undergird an approach to journalism education focused on the individual's role as one node in a dynamic and unreliable information ecosystem rather than on the heroic reporter who goes out hunting for scoops.

Each day, Share a Story participants must share one — and only one — news story and say what happened, why it matters, and where they found it. At first students do in-class exercises based on the stories they share, such as researching the author of the story or the owner of the publication where it appeared, all while being nudged to share stories from more diverse sources. Over time, the accrued stories form a collaboratively created data set that students use to analyze their news habits, understand community information needs, look for story ideas, and prototype new news products on paper or using no code/low code tools.

This workshop session will demo a working prototype of Share a Story built in Airtable, discuss progress and challenges, and explore plans for further testing and development.