

# **BENGAL Lightning Talk**

10/24/2023

### Dr. Svitlana Volkova

Chief AI Scientist Office of Science and Technology

### **Dr. Robert McCormack**

Principal Research Engineer, Director Intelligent Performance Analytics Division

Aptima, Inc.

### **APTIMA®** Human-Centered Engineering®

## **Corporate Vitals**

- Multidisciplinary Small Business
  - Founded in 1995
  - Experts in cognitive and behavioral sciences, AI/ML, data, ML and software engineering
- Human-Centered Focus in Warfighter Modernization
  - Tools that bridge the gaps between people, technologies, and complex operational environments
  - Individual and collective readiness at scale
  - Human-Al integration
  - \$40M \$50M in Annual Revenues
    - Rapid recent growth
    - Move from lab research to field implementation
    - Increasing commercialization
- 26+ Years Supporting the Army
  - AFC CFTs, DEVCOM, ARI, PEOs
    - CMMC L1 compliant Aptima has completed work on all 110 controls in NIST SP 800-171 and will be pursuing CMMC Level 2 Certification though a CMMC Third Party Assessor Organization (C3PAO) in 2023
    - Approved Cost Accounting System
    - CMMI L3 recertification in progress for 1QTR23 certification



- I. AI/ML Evaluation
- II. Paradigm Shift: Evaluation of Emerging AI Behaviors
- III. Human-AI Integration and Trust
- IV. Causal Discovery and Reasoning

Lessons from Developing Multimodal Models with Code and Developer Interactions			
Nicholas Botzer University of Notee Dame Rootzer@od.edu	Sameera Herawalavithana Pacific Neethwest National Laboratory maneera. horawalavithana\$panlgov		
Tim Weninger University of Notre Dame tweeiinger@nd.edu	Svitlana Volkova Pacific Noefhwent National Laboratory avitlana. volkova@penl.gov		
	hstract preservation has been the time of language the the addity of these models in field and the second		
1 Introduction Annual devices and advances and a second a seco	Opportunities, Challen	ific Knowledge for Chemistry: ess and Lessons Learned	
products via Githab Copilor. Despite this carb containes to seek innovative ways to improve moves beyond the straightforward modeling, syntax tree of the code during the pre-trainin akility so fill-in missing argments of code, m into consideration models that can handle bot PLRART [21a] cad CoddTS [21].	theirest	rton, Shivam Sharma, Scott Howland, in Cosbey, Maria Glenski, Svitlama Volkovu I Laboratory, Richland, WA	
These models are quickly being adopted by e- with little regard to the accurate prediction engoing work by Ausre et al. seeks to investing models (13, Auditional work by Pareze et al. type of Computer Weakness Enzamentions (C I Can't Believe It's Not Better Workshop at Next	ADDUTING Freedotion modely previousl on large compose demonstrate significant gains across many sam- nal language processing tasks and domains e.g., law, beathoure, education, etc. However, only larized efforts have singuistic and accessing applications. In models is a science and accessing applications. In	2021) and Preview (Jorgio et al., 2021), estable block a pareaffer shift in Artificia Intelligence (AD. These foundation models, also called non- ori platform, at built using self-supervised pre- training at scale. They are then able to be staily adapted to a vide range of downstream tasks via transfer learning (Homassan et al., 2021) and Ine- tuning (Let et al., 2019).	
	A contrast the state of the s	The wide community adjection of foundation models can be explained by their key properties, two of which are energened behavior and konsong- englatation - which also make foundation models appealing for adjustion across science and security domains. Energeneous, or emergent behavior, reflect new behaviors that a model introduces or is capable. Under review as a conference opper at ICLR 2023	
	daph, and trape-ratio on seekl performance in sensitive in tools taking definitions, seek tools and the sensitive sensitive sensitive set in the sensitive sensitive sensitive sensitive performance where we have a sensitive sensitive performance reserve for data gamma (at 2011) performance reserve for data gamma (at 2011) per boots and performance with for sequelity performance performance of the sensitivative data boots taked from performa (at 2011) densitiative data boots taked from performa (at 2011) densitiative data boots taked from performa (at 2011) densitiative data data boots taked from performa (at 2011) densitiative data data data data data data data data	MOLJET: MULTIMODAL TRANSFORMER FOR CONDIT ULAR DESIGN AND MULTI-PI	
	pasks, r.g., 3:6Q, and (4) models pro-donized from scratch perform better on in-donain tasks than those tuned from general-purpose models like Open AFs GPT2. <ol> <li>Introduction</li> </ol>	Anonymous authors Paper under double-blind review	
	1 Introduction The emergence of foundation models (Bom- masani et al., 2020) such as large-scale autoen- coding models (e.g., BERT (Devins et al., 2018), RoBBRTs (Liu et al., 2019) and autoergensive larguage models (e.g., BERT (Devins et al., 2014), 2019). OPT-3 (Brows et al., 2020). Megatreen- Turing (Smith et al., 2022, and Cophor (Eds et al., 2021) as well as multimodal vision and larguage models, such as FLANC (Singht et al., 2018).	ABSTRA Multi-property constrained optimization of design models is vital for the successful age towards materials and drug discovery. Yet ported performance of such models in the real world design accurates, Furthermore, e ble to chemistis without an extensive buckgrue for the successful accurate of the successful age Ranbedding Transformer (MoiJITT), which desired molecular distributions based on buc	CT molecules using generative do novo lication of Artificial Intelligence (AI) there remains a gap between the re- iterature and their practical utility in axisting models are largely inaccessi- nod in computer science. To address and in computer science. To address a performs conditional generation of man-interpretable chemistry prompts
l	Proceedings of Biglisinov Epitode of — Weakboy on Challenger May 27, 2022 (2002) Associat	ADDITION TO A STATE OF A DISTUNCT AND A DISTUNCT AN	on the standard benchmarks available transworks. These includes structure: transworks the structure of the structure of the objective given realistic property con- revised peterining. MOLIET to outper- using zero-shot inferences and beats recover, the performance of MOLIET of a multimodal approach to molec- ture-based de-nove molecular design odels and should serve as a building sible AI for chemists.
		1 INTRODUCTION	
		Energying crises in climate, disease and human bealt ity and must be actively new with recative contributions. A discovery of innovative functional materials or novel flow interactive, the visibility or using redox. How baselers redox potential and high solubility (Zhang et al., 2013), the climate space (Foldschutz et al., 2013), the of climatical phase space (Foldschutz et al., 2013), the of climatical interaction of the state of the state state of the state of the state of the state of the state state of the state of the state of the state of the state state of the sta	so (RPBs) for insolvenous weat optimine pre- so (RPBs) for long-term and large-scale with fast electrochemical kinetics, a 18). Due to the immense size and con a search for suitable materials is far fro- miterative modifications to existing c 96). ugun to look towards generative de nove
		structures are often far too low (Koho & Beratan, 19 To address this user, researchers have increasingly by models to efficiently anvigate the wast molecular pha are evaluated to their ability to generate a diverse an taneously basing them towards a desired property di the ubiquity of uting based models and property di the ubiquity of uting based models are presentable design. For instance, transformer architectures have prediction tasks that require quantum-level accuracy ( increase the diversity of candidates sampled from m et al., 2021).	aw opace (Meyers et m., 2024). These ary of novel molecular structures while istribution (Polykowskiy et al., 2020), ns (Weininger, 1988; Krenn et al., 20, een saccessfully applied to <i>de novo</i> mis a shieved state-of-the-art results on p Ross et al., 2021) and have also been al schine-learned molecular distributions



## Aptima's S&T

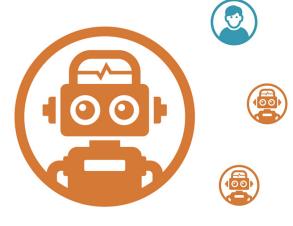
## System of Teams



### Human-Al



Aptima's focus is on technology innovation for measuring and optimizing human, AI, and integrated human-AI performance



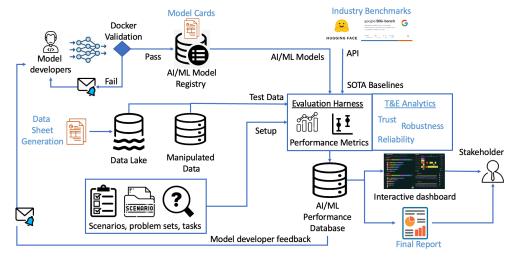
Automation (AI)



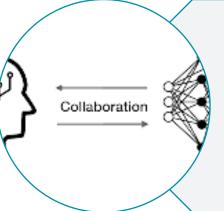
# **Expertise and Capabilities in AI/ML Evaluation**

- AI/ML performance evaluation against the state-of-theart/baselines
- Interoperability with industry-led benchmarking platforms and performance metrics
- A comprehensive set of AI/ML performance metrics ranging from scalars to vectors and distributions
- Interactive visualizations to support both quantitative (metrics based) and qualitative evaluation of model behavior
- The ability to rapidly perform repeated evaluations and metrics calculations to quantify uncertainty and report error bars
- The ability to experiment across multiple datasets and problem sets to perform an "apple-to-apple" comparison of AI/ML model performance
- Support "out-of-distribution" evaluation
- Quantitatively evaluate AI/ML robustness, reliability, explainability and trust

### AI/ML T&E (DARPA SEMAFOR)



### Human-AI T&E (DARPA ASIST)



Develop a testbed to support the design and evaluation of Artificial Social Intelligence (ASI)

Conduct human-in-the-loop team experiments to assess analytics components that predict and measure teamwork and ASI agents that intervene to improve teamwork.

# **Beyond Classic AI/ML Evaluation Metrics**

 Accountability: Relative reliability when applied in key circumstances (operational datasets, problem sets, tasks, and scenarios)

Human-Centered Engineering®

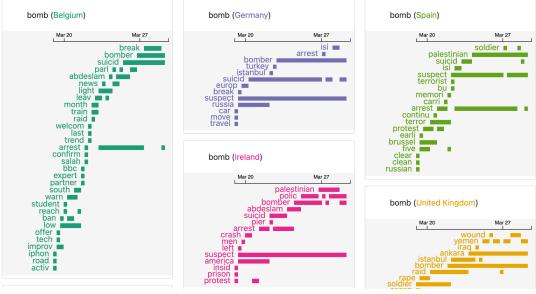
- Robustness: Quantitative metrics for testing resiliency to variations in data input
- Explainability: Transparency of model behavior and identify points of failure through data inputs and model predictions trust in AI/ML in operational setting
  - [1] CrossCheck: Rapid, Reproducible, and Interpretable Model Evaluation. Arendt, D., Shaw, Z., Shrestha, P., Ayton, E., Glenski, M., and Volkova, S. ACL Workshop on Data Science with HITL. 2021.
  - [2] ESTEEM: A Novel Framework for Qualitatively Evaluating and Visualizing Spatiotemporal Embeddings in Social Media. D. Arendt, and S. Volkova. ACL'17

# **CrossCheck**<sup>[1]</sup> Understanding representative data and reason for model misclassifications

741k0 741k0 741k0 74

FullNe

FullNetPo: LingNet



0.2 - 0.4

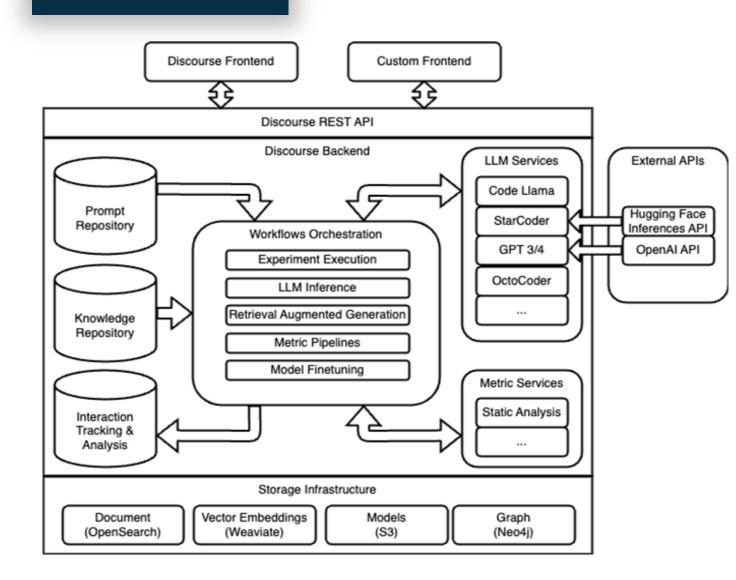
0.8 - 1

7.41k 0

0.4 - 0.6

**ESTEEM**<sup>[2]</sup> Visualizing spatiotemporal embeddings

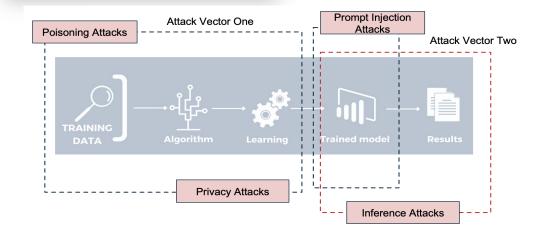
## Paradigm Shift: Evaluating Emerging AI Behaviors



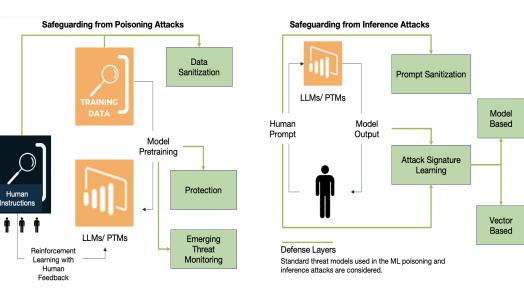
- Beyond red teaming LLMs!
- Experiment with a diverse set of prompt injection and inference attacks on LLMs, e.g., cognitive biases, data poisoning, jailbreaking privacy, and backdoor attacks
- Metrics:
  - Core trustworthy AI metrics
     (robustness, transparency, fairness)
  - Core cognitive framing effect measures (contrast, decoy, default, distinction)
  - Measuring the effect of LLM attacks
     on downstream task performance
- Learning from human-AI interaction data (observational and interventional) at scale
- Achieving robust LLM behavior by chaining multiple feedback processes (self-critique → self-refine → self-revision)

Human-Centered Engineering®

# LLM Attack and Defense Strategies



Human-Centered Engineering®



- **Contrast Effect:** Test if users change preferences between two LLM-generated ideas, one with exaggerated contrasts versus neutral.
- **Decoy Effect:** Test if introducing a decoy option shifts users' preferences between two choices described by biased versus neutral LLMs
- **Default Effect:** Test how often users accept the default choice described as recommended by a biased LLM versus a neutral one.
- **Distinction Bias:** Test if users evaluate two options differently when presented separately versus simultaneously after reading exaggerated versus consistent LLM descriptions.

Wijesekera, P., et al. (2017). The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. 2017 IEEE Symposium on Security and Privacy (SP). IEEE.

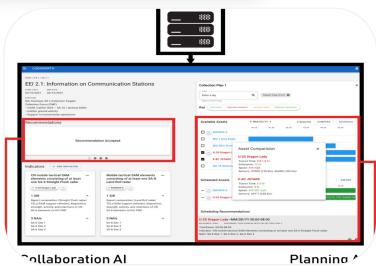
Wijesekera, P., et al. (2018). Contextualizing privacy decisions for better prediction (and protection). Proceedings of the CHI Conference on Human Factors in Computing Systems.

Oliver, S., Reimann, M., & Cook, K. S. (2021). Trust in social relations. Annual Review of Sociology, 47, 239-259. Reimann, M., Oliver S., & Cook, S. (2017). Trust is heritable, whereas distrust is not. Proceedings of the National Academy of Sciences, 114(27), 7007-7012.

### **APTIMA**<sup>®</sup> Human-Centered Engineering<sup>®</sup>

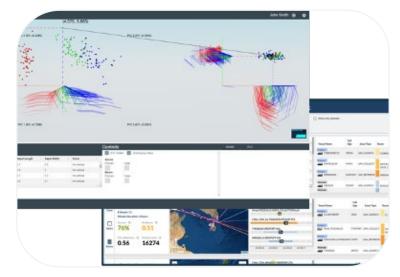
# **Human-Al Integration and Trust**

Intel analysts, aircraft maintainers, mission planners, sonar watchstanders, UAV pilots, simulation data analysts



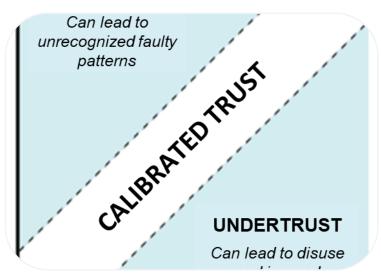
### COGSWORTH

 Proactive and adaptive cognitive assistant for collection managers that ingests, parses, and reports collection plan requirements, indicators, and detailed asset information to improve situational awareness in evolving combat scenarios.



#### Sidekick™

- Systems for Interactive Discovery and Exploitation of Knowledge and Insights with Contextual Kinetics approach to human-machine teaming and human-Al integration
- 6 critical design principles (Bruni, Freiman, and Riddle, 2023)



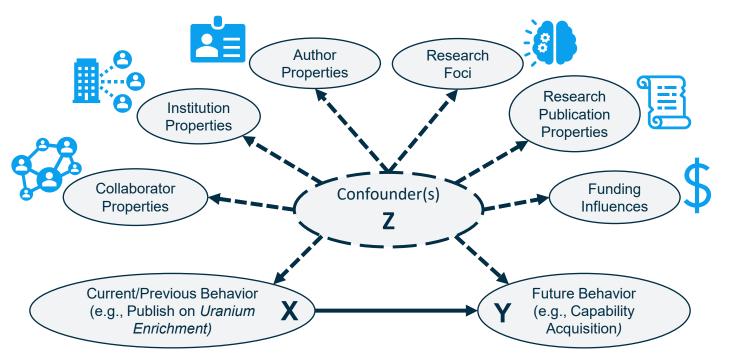
### TRUST'M

 Novel approach that allows to intervenes at the right time in the right way updating its recommendations based upon user interactions – allowing it to evolve with its human teammate's evolving needs during analyst's information search (Rebensky et al., 2022).



## **Causal Discovery – Causal Structural Learning Techniques**

Structural causal models are graph representations of how relevant features of the world interact with each other, i.e., the mechanisms by which data (observations) is generated.



Causal Effect of X on Y is calculated by: P(Y = y | do(X = x)) =

 $\sum_{z} P(Y = y | X = x, PA = z) P(PA = z),$ 

Algorithms used include:

- Constraint-Based
- Score-Based
- Causal Ensemble (as published in NeurIPS Workshop on Causal Discovery & Causality-inspired ML)

**Volkova, S.**, et al. (2023). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods. Computational and Mathematical Organization Theory, 29(1).

Saldanha, ... S. Volkova. 2020. "Evaluation of Algorithm Selection and Ensemble Methods for Causal Discovery." In Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, December 2020.

Glenski M.F., S. Volkova. Identifying Causal Influences on Publication Trends and Behavior: A Case Study of the Computational Linguistics Community. In EMNLP Workshop on Causal Inference & NLP. 2020.





Dr. Svitlana Volkova

https://www.linkedin.com/in/svitlanavolkova/

svolkova@aptima.com



Dr. Robert McCormack

https://www.linkedin.com/in/robert-mccormack-1321488/

rmccormack@aptima.com

www.aptima.com

### **APTIMA**<sup>®</sup> Human-Centered Engineering<sup>®</sup>

## **Peer-reviewed Publications (1)**

#### LLM and Foundation Model Training and Fine-Tuning

- Horawalavithana, S., ... & Volkova, S. (2022a). Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In Proceedings of BigScience--Workshop on Challenges & Perspectives in Creating Large Language Models (pp. 160-172).
- Dollar, O. W., Horawalavithana, S., Vasquez, S., Pfaendtner, W. J., & Volkova, S. (2022). MolJET: Multimodal Joint Embedding Transformer for Conditional de novo Molecular Design and Multi-Property Optimization.
- Botzer, N., Horawalavithana, S., Weninger, T., & Volkova, S. (2022). Lessons from Developing Multimodal Models with Code and Developer Interactions. In I Can't Believe It's Not Better Workshop: Understanding Deep Learning Through Empirical Falsification.

#### Robustness

- W. Wu, D. Arendt, S. Volkova. (2021) Evaluating Neural Machine Comprehension Model Robustness to Noisy Inputs and Adversarial Attacks. EACL'21.
- M. Glenski, E. Ayton, R. Cosbey, D. Arendt, and **S. Volkova**. (2020). Towards Trustworthy Deception Detection: Benchmarking Model Robustness across Domains, Modalities, and Languages. International Workshop on Rumors and Deception in Social Media at COLING.
- M. Glenski, E. Ayton, R. Cosbey, D. Arendt, and S. Volkova. (2021). Evaluating Deception Detection Model Robustness To Linguistic Variation." In International Workshop on Natural Language Processing for Social Media (SocialNLP).

#### Transparency

- Volkova, S., Ayton, E., Arendt, D., Huang, Z., & Hutchinson, B. (2019). Explaining Multimodal Deceptive News Prediction Models. Proceedings of the International AAAI Conference on Web and Social Media.
- Arendt, D., & **Volkova, S.** (2017). ESTEEM: A novel framework for qualitatively evaluating and visualizing spatiotemporal embeddings in social media. Proceedings of ACL 2017, System Demonstrations.

#### Reliability

- Arendt, D., Huang, Z., Shrestha, P., Ayton, E., Glenski, M., & Volkova, S. (2021). CrossCheck: Rapid, Reproducible, and Interpretable Model Evaluation.
   Workshop on Data Science with Human-in-the-loop: Language Advances (DaSH-LA) colocated with NAACL 2021.
- Arendt, D., E. Grace, and **S. Volkova**. (2018). Interactive machine learning at scale with CHISSL. In Thirty-Second AAAI Conference on Artificial Intelligence.
- S., Emily, L. M. Blaha, A. V. Sathanur, N. Hodas, **S. Volkova**, and M. Greaves. (2019). Evaluation and Validation Approaches for Simulation of Social Behavior: Challenges and Opportunities." In Social-Behavioral Modeling for Complex Systems.

#### © 2023 Aptima, Inc.

### **APTIMA**<sup>®</sup> Human-Centered Engineering<sup>®</sup>

## **Peer-reviewed Publications (2)**

#### Human-Al Integration and Trust

- Carmody, K., Ficke, C., Nguyen, D., Addis, A., Rebensky, S., & Carroll, M. (2022). A Qualitative Analysis of Trust Dynamics in Human-Agent Teams (HATs).
   In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 66, No. 1, pp. 152-156). Sage CA: Los Angeles, CA: SAGE Publications.
- Rebensky, S., Carmody, K., Ficke, C., Carroll, M., & Bennett, W. (2022). Teammates instead of tools: The impacts of level of autonomy on mission performance and human-agent teaming dynamics in multi-agent distributed teams. Frontiers in Robotics and AI, 102.
- Rebensky, S., et al. (2021) Whoops! Something went wrong: Errors, trust, and trust repair strategies in human agent teaming. Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings. Cham: Springer International Publishing.
- Ezer, N., Bruni, S., Cai, Y., Hepenstal, S. J., Miller, C. A., & Schmorrow, D. D. (2019). Trust engineering for human-AI teams. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 63, No. 1, pp. 322-326). Sage CA: Los Angeles, CA: SAGE Publications.
- Salinas, A., Shah, P. V., Huang, Y., **McCormack, R.**, & Morstatter, F. (2023). The Unequal Opportunities of Large Language Models: Revealing Demographic Bias through Job Recommendations. arXiv preprint arXiv:2308.02053.

#### **Causal Discovery and Reasoning**

- Volkova, S., et al. (2023). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods. Computational and Mathematical Organization Theory, 29(1).
- Saldanha, ... **S. Volkova. (**2020). Evaluation of Algorithm Selection and Ensemble Methods for Causal Discovery. In Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems.
- Glenski, M., & Volkova, S. (2021). Identifying Causal Influences on Publication Trends and Behavior: A Case Study of the Computational Linguistics Community. In Proceedings of the First Workshop on Causal Inference and NLP (pp. 83-94).
- Guo, G., Glenski, M.F., Shaw, Z.H., Saldanha, E.G., Endert, A., Volkova, S., & Arendt, D.L. (2021). VAINE: Visualization and AI for natural experiments. IEEE VIS 2021.
- Cottam, J.A., Glenski, M.F., Shaw, Z.H., Rabello, R.S., Golding, A.J., Volkova, S., & Arendt, D.L. (2021). Graph comparison for causal discovery. Visualization in Data Science 2021.