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.

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-AI 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

APTIMA®
Human-Centered Engineering®

Outline

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

https://openreview.net/pdf?id=iUUvNqUzJX2

https://aclanthology.org/2022.bigscience-1.12.pdf

https://openreview.net/pdf?id=7UudBVsIrr

APTIMA®
Human-Centered Engineering®

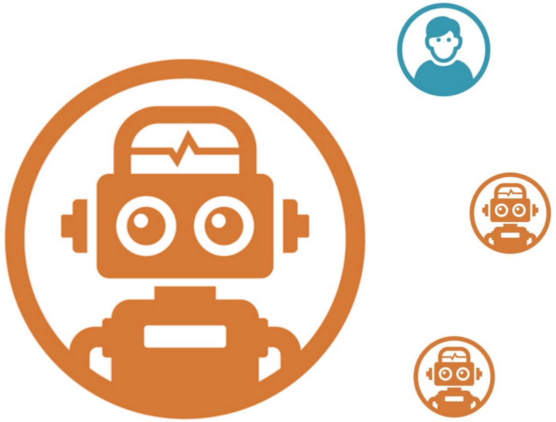
Aptima's S&T

System of Teams

Human-AI

Human

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



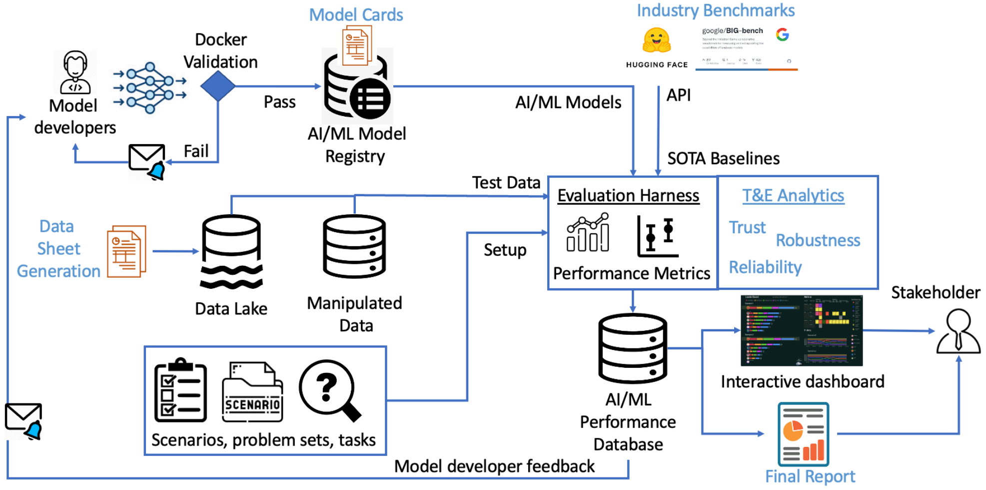
Automation (AI)

Expertise and Capabilities in AI/ML Evaluation

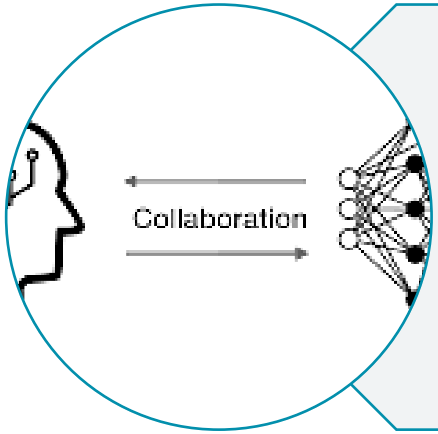
APTIMA®
Human-Centered Engineering®

- AI/ML performance evaluation against the state-of-the-art/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)



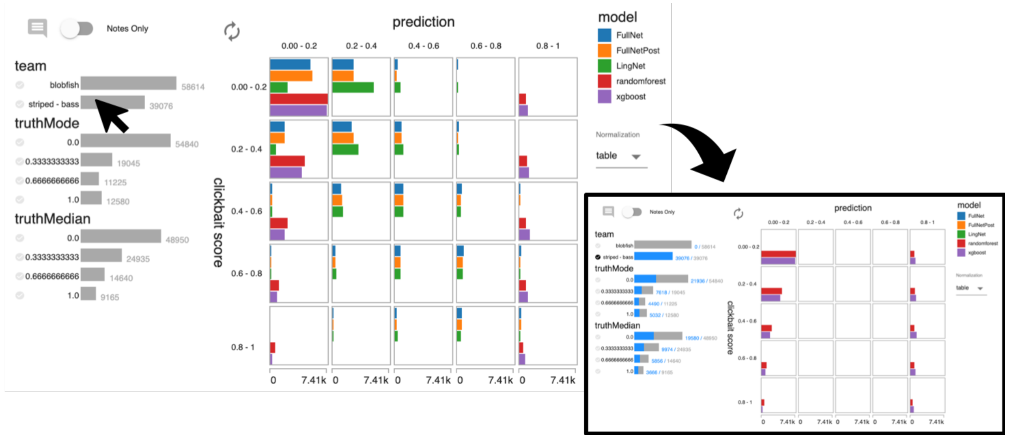
Develop a testbed to support the design and evaluation of Artificial Social Intelligence (ASI)

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

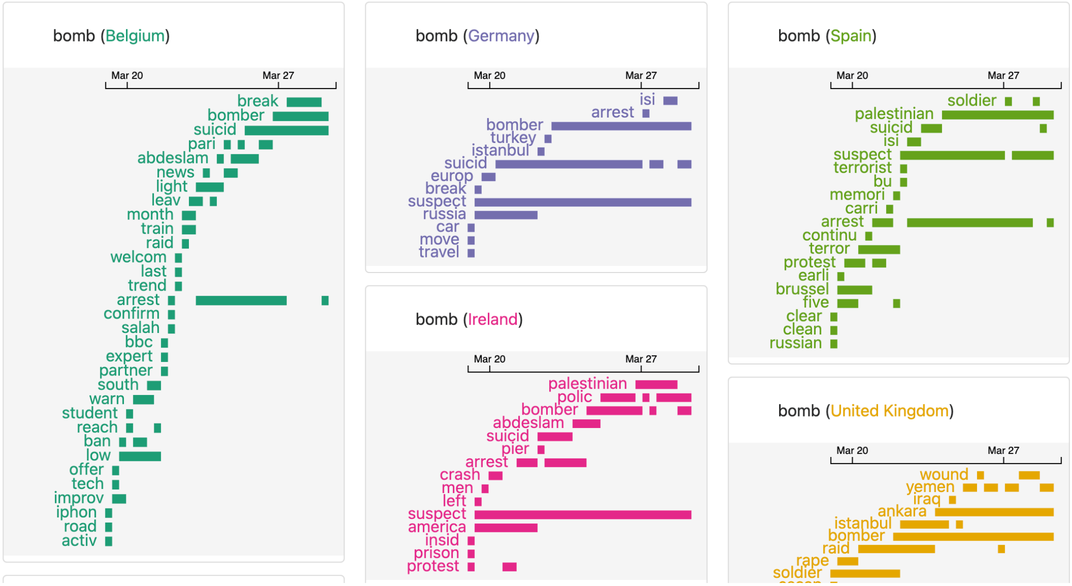
APTIMA®
Human-Centered Engineering®

Beyond Classic AI/ML Evaluation Metrics

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



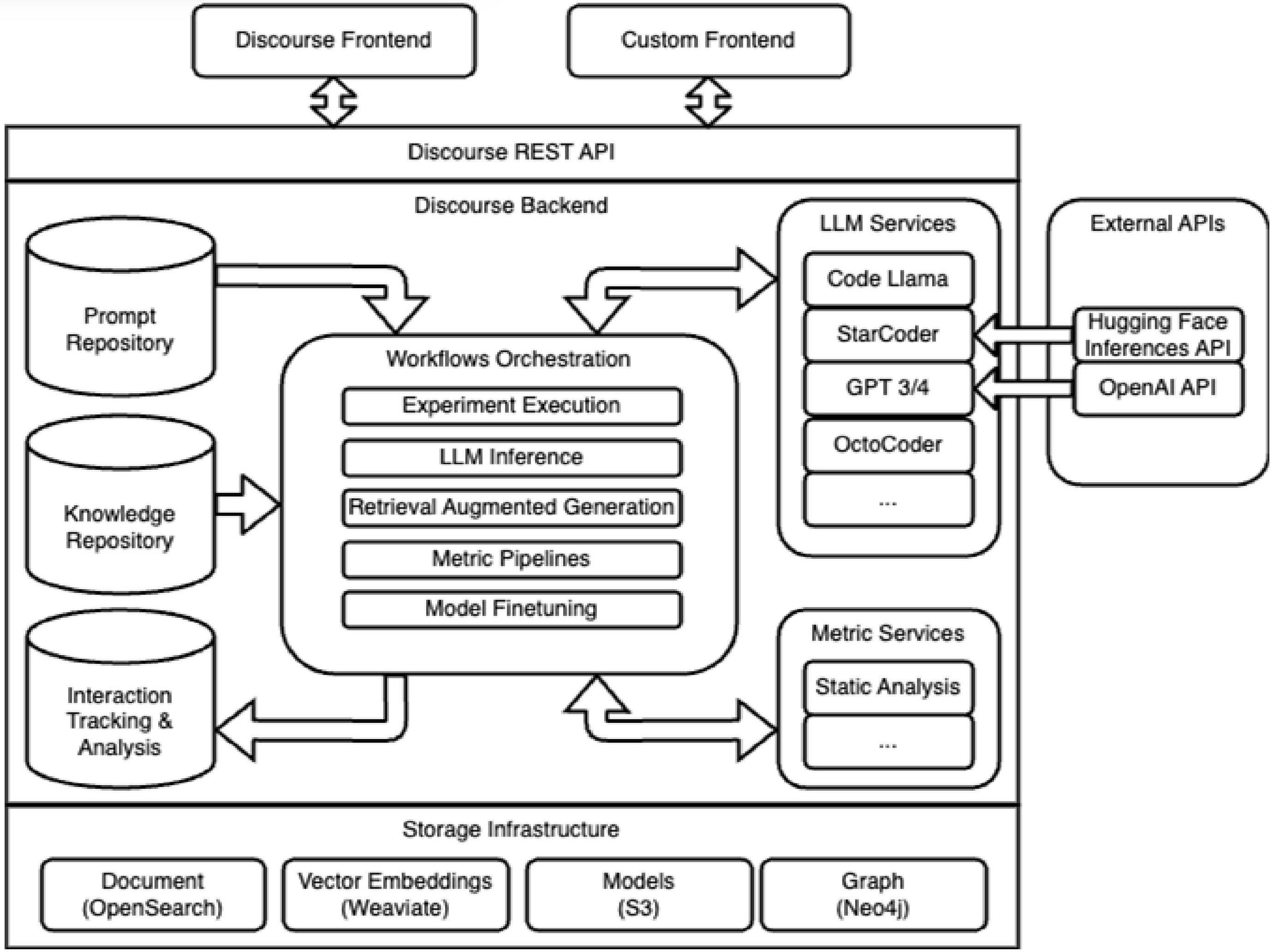
CrossCheck[1] Understanding representative data and reason for model misclassifications



ESTEEM[2] Visualizing spatiotemporal embeddings

[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

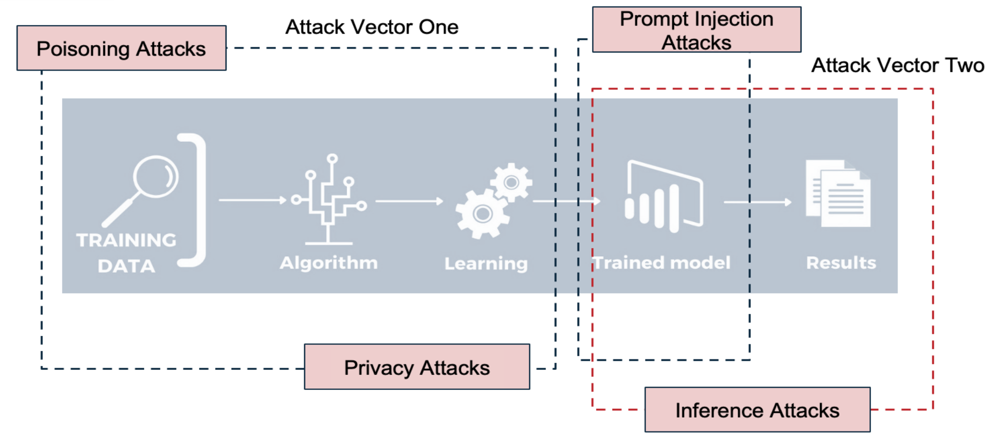
Paradigm Shift: Evaluating Emerging AI Behaviors

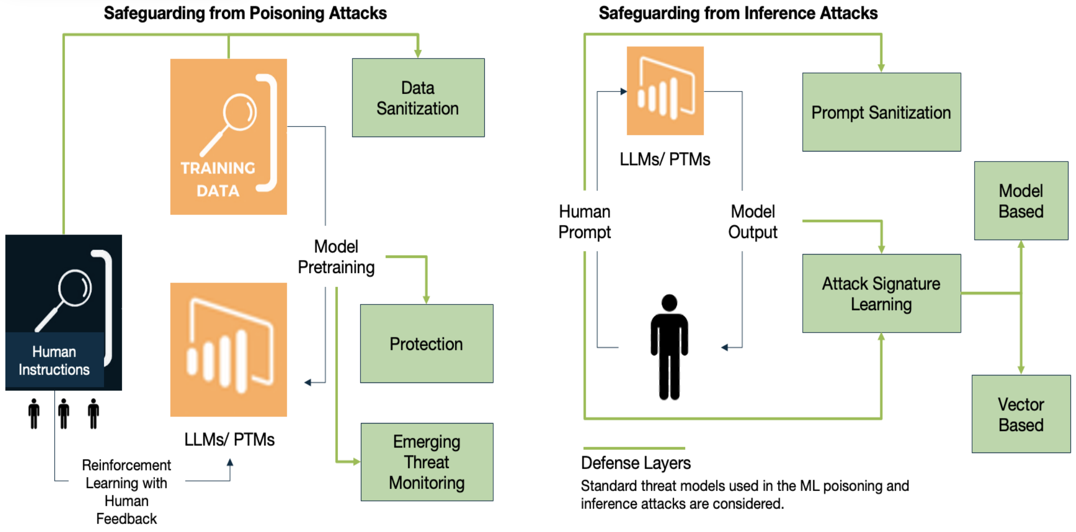


- Beyond red teaming LLMs!
- Experiment with a diverse set of prompt injection and inference attacks on LLMs, e.g., cognitive biases, data poisoning, jailbreaking privacy, and backdoor attacks
- Metrics:
    - Core trustworthy AI metrics (robustness, transparency, fairness)
    - Core cognitive framing 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)

APTIMA®
Human-Centered Engineering®

LLM Attack and Defense Strategies





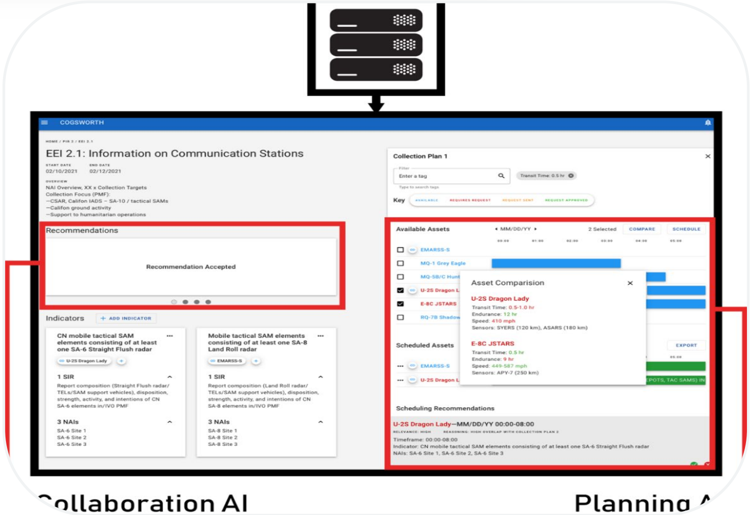
- **Contrast Effect:** Test if users change preferences between two LLM-generated ideas, one with exaggerated contrasts versus neutral.
- **Decoy Effect:** Test if introducing a decoy option shifts users' preferences between two choices described by biased versus neutral LLMs
- **Default Effect:** Test 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®
Human-Centered Engineering®

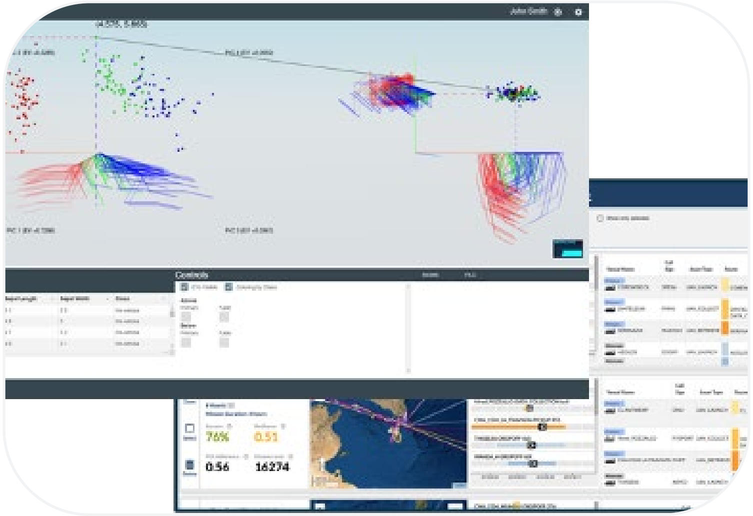
Human-AI Integration and Trust

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



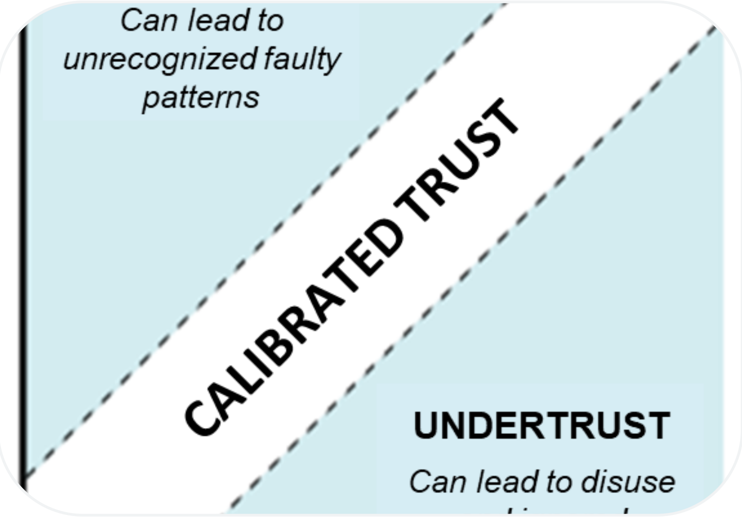
COGSWORTH

- 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-AI integration
- 6 critical design principles (Bruni, Freiman, and Riddle, 2023)

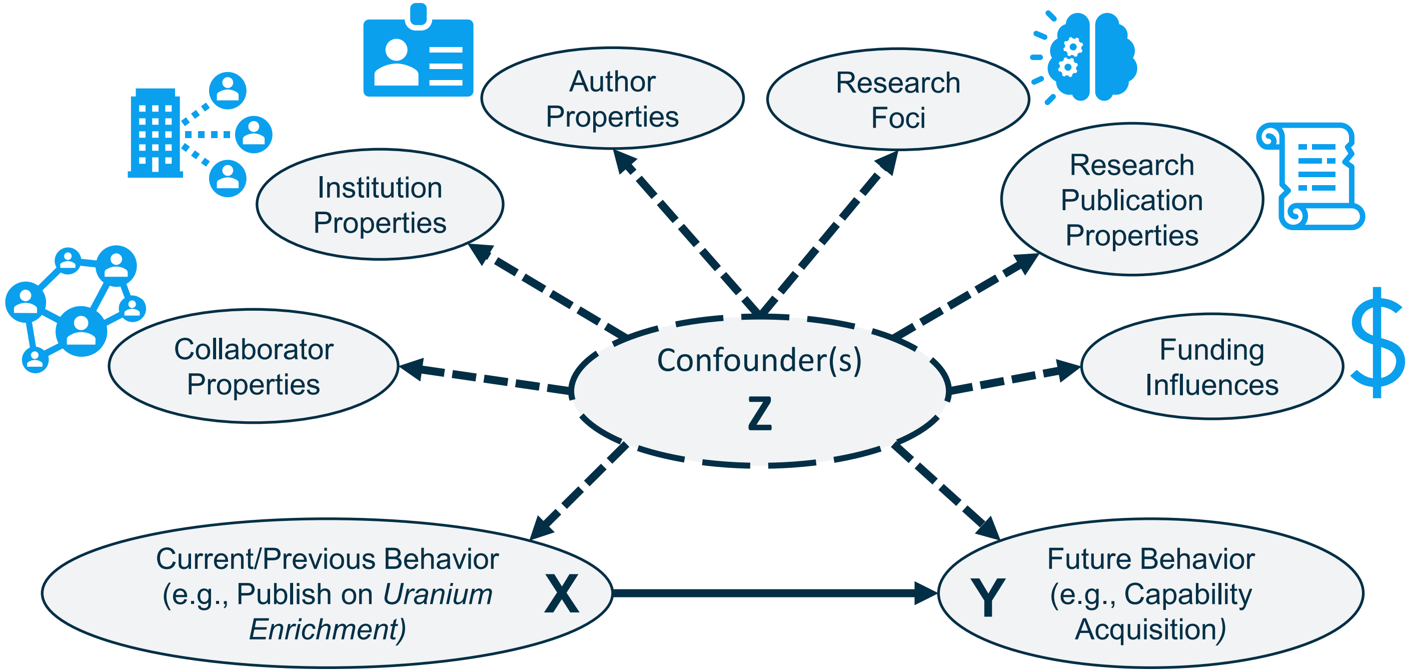


TRUST'M

- 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

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.

APTIMA®
Human-Centered Engineering®

Peer-reviewed Publications (2)

Human-AI 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.