Jelena Tešić: Research Narrative

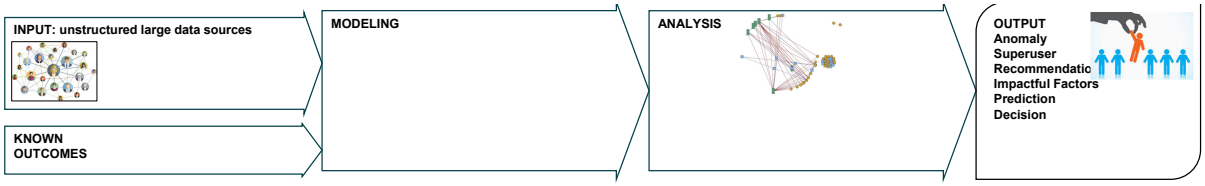

Figure 1: DataLab research focuses on modeling and analysis of large unstructured noisy data.

The arching quest of Tešić's research in the Computer Science department is to **identify** analytics tasks where unstructured data or task at hand do not adhere to underlying assumptions of state-of-art algorithms and to **design** efficient, effective, intuitive, and responsible algorithms for addressing the challenges of such extensive collections in the wild, as illustrated in Figure 1. Tešić leads the [Data Lab@TXST] and currently advises 5 out of 60 Ph.D. students in the department. and the research is grouped by the tasks it solves below.

Scaling Signed Graph Tasks through Fundamental Cycle Basis The research proposes the novel balance theory approach to signed social network graph analysis: a frustration cloud view of the signed graph where the vertices and edges are quantified through statistics and validated the approach for multiple real networks [30] [pdf]. Next, the research expanded to the development of an algorithm to efficiently compute the fundamental cycle basis in large, unstructured graphs to scale the frustration cloud computation [8][pdf], and to discover fundamental cycle basis in large signed networks [23] [pdf]. The effort to compare and contrast state-of-art community discovery on actual signed graphs in [14] [pdf] led to the first scalable signed benchmark comparison to date, and frustration cloud-based approach for cluster boosting for high modular signed graphs [15] [pdf]. Two master's and five undergraduate students contributed to this research project to date. The research now focuses on solving NP-hard tasks at scale such as computing frustration [25] [pdf] and finding the largest balanced subgraph [24] [pdf] for graphs with millions of nodes and edges derived from sensor, agent, and gene networks.

Modeling Social Network Relations The multifaceted interconnectivity of users and content on Twitter through user connections, replies, quotes, hashtags, and shared content makes it an exciting medium for research on the effectiveness of the representation and methods used. The project has introduced a scalable end-to-end Twitter network data management pipeline that gathers, stores, and models rich relationships from Twitter networks [12] [pdf]. The research work compared and contrasted the analysis results of millions of Twitter data using multiple graph construction processing approaches [13] [pdf]. The community-based modeling, where the tweet is classified on the content and also on the retweets, replies, quotes, hashtags, and the author, yields precision, recall, and accuracy comparable to lexical classifiers [22] [pdf]. The project proposes new multi-modal approaches that consistently deliver the most robust outcomes and exhibit the highest performance measures for network graphs constructed based on Twitter interactions related to the COVID-19 pandemic [1] [pdf] and [22] [pdf]. One undergraduate, one master's, and one Ph.D. student participated in the research work for this research project.

Vision Tasks for Highly Variable Overhead Videos The project has proposed multiple domain adaptation approaches to alleviate the degradation of object identification in previously unseen overhead datasets with significant domain gaps and dominant small objects [3] [pdf] and [26] [pdf]. The project has proposed new algorithms for overhead videos to detect anomalous activities [7, 31]. Five undergraduates, two masters, and three Ph.D. students participated in this project sponsored by NAVAIR. The project evolved in the Ph.D. thesis work, and the team

proposed innovative and efficient contrastive learning algorithms to improve object classification in previously unseen highly variable overhead datasets [4] [pdf] and [6] [pdf]. The team is now introducing progressive domain adaptation to produce domain-invariant features across aerial datasets using local and global components for domain adaptation and object classification for the task [5] [pdf] and [18] [pdf]. The second part of the project has introduced a new indexing and search algorithm for deep descriptor databases that have up to four times lower memory usage and higher effectiveness than state-of-art [19] [pdf] and [20] [pdf] on millions of deep descriptors. The study is building upon the approach for crowd-sensing application [18] [pdf] and plans to improve the indexing footprint while keeping the search effectiveness by proposing a new stratified graph approach [21] [pdf].

Semantic Segmentation Task in The Wild Teši'c's contribution to semantic segmentation focuses on pavement distress detection for transportation and road maintenance and NASA spaceflight sample images. The project has quantified the main influencing factors that affect the performance of deep learning models in pavement distress detection pipelines and proposed a semantic segmentation algorithm that significantly improves the accuracy of localizing pavement cracks [9] [pdf]. Current research explores the domain adaptation approaches from Section 3 to improve the algorithmic performance for specialized fields and how pixel-wise segmentation can improve the explainability of the model. In parallel, the project focuses on a semantic segmentation pipeline to automatically classify Bacterial Adhesion and Corrosion from images obtained in the NASA SpaceX-21 experiment. Automatically identifying corrosion in tens of thousands of images of samples flown into space will help researchers streamline the samples' data collection and help them reach conclusions faster on what countermeasures will work in space. Two undergraduates, two masters, and one Ph.D. student are contributing to the project.

Predictive Modeling of Noisy Tabular Data Tešić's team has designed the multi-feature importance analysis algorithm and applied it to large-scale analysis of public data from the National Center for Education Statistics (NCES) to provide data-driven insights into teacher attrition challenges. The study discovered that the race and sex of the principal, the type of school, and the school's location impact teacher retention rates the most and that modeling historical data resulted in a predicted attrition rate of over 10%, aligning closely with the current prevalent attrition rates in the USA [11] [pdf]. The project has developed an interpretable data-driven scoring fusion to discover the most critical factors from an extensive collection of heterogeneous public data sources on learning loss during the COVID-19 pandemic in Texas public schools. The robust approach found that the number of students on school campuses was the most impactful predictor of how the students would perform on the standardized test in mathematics and reading in the Spring of 2021 in Texas [16] [pdf]. The work introduced new cascade enhancement to ensure effectiveness and the prediction coverage of our modeling pipeline to predict long COVID in N3C data [17] [pdf]. Two undergraduate, two master, and one Ph.D. student contributed to the project.

Summary The research focuses on providing new algorithms based on mathematics, computer science, and statistics theory while considering the specifications (efficiency, scalability, usability, interpretability) of the task at hand, data characteristics, and domain applicability. To this end, Tešić has collaborated with other Labs at our department and proposed data-driven solutions for their specific challenges [10, 27, 2, 28, 29]. Out of 31 papers published or under review since joining TXST, 29 papers are co-authored with TXST students (see pdf links and supplemental documentation). Tešić is currently working on new algorithms and methods for solving challenging analytics tasks stemming from the nature and size of unstructured data in climate analysis and precision medicine.

* *author* - names of the the Data Lab students are in *italic*

References Cited

[1] *Andrew Magill*, *Lia Nogueira de Moura*, *Maria Tomasso*, *Mirna Elizondo*, and **Jelena Tešić**. Enriching content analysis of tweets using community discovery graph analysis. In *Proceedings of the MediaEval 2020 Workshop*, volume 2882, 2020.

[2] *Blake W Ford*, Apan Qasem, **Jelena Tešić**, and Ziliang Zong. Migrating software from x86 to arm architecture: An instruction prediction approach. In *2021 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 1–6, 2021.

[3] *David Heyse*, *Nicholas Warren*, and **Jelena Tešić**. Identifying maritime vessels at multiple levels of descriptions using deep features. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 423 – 431. SPIE, 2019.

[4] *Debojyoti Biswas*, *MMM Rahman*, Ziliang Zong, and **Jelena Tešić**. Improving the energy efficiency of real-time dnn object detection via compression, transfer learning, and scale prediction. In *The IEEE 16th International Conference on Networking, Architecture, and Storage (NAS 2022)*, September 2022.

[5] *Debojyoti Biswas* and **Jelena Tešić**. Domain adaptation with contrastive learning for object detection in satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing (**under review**)*, 2023.

[6] *Debojyoti* Biswas and Jelena Tešić. Progressive domain adaptation with contrastive learning for object detection in the satellite imagery. In *arXiv:2209.02564*, 2023.

[7] *George E. Strauch*, *Jiajian (Jax) Lin*, and **Jelena Tešić**. Overhead projection approach for multi-camera vessel activity recognition. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5626–5632, 2021.

[8] *Ghadeer Alabandi*, **Jelena Tešić**, Lucas Rusnak, and Martin Burtscher. Discovering and balancing fundamental cycles in large signed graphs. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA, 2021. Association for Computing Machinery.

[9] *Haitao Gong*, **Jelena Tešić**, *Jueqiang Tao*, Xiaohua Luo, and Feng Wang. Automated pavement crack detection with deep learning methods: What are the main factors and how to improve the performance? *Transportation Research Record*, page 03611981231161358, 2023.

[10] *Hanie Samimi*, **Jelena Tešić**, and Anne Hee Hiong Ngu. Patient-centric data integration for improved diagnosis and risk prediction. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 185–195, Cham, 2019. Springer International Publishing.

[11] *June Yu*, Li Feng, and **Jelena Tešić**. Mitigating u.s. public school teacher attrition crisis: A data science approach. *Information Processing & Management (**under review**)*, 2023.

[12] *Lia Nogueira de Moura*. Social network analysis at scale: Graph-based analysis of twitter trends and communities. Master's thesis, TXST, 2020. Advisor: **Jelena Tešić**.

[13] *Lia Nogueira de Moura* and **Jelena Tešić**. pytwanalysis: Twitter data management and analysis at scale. In *2021 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2021.

[14] *Maria Tomasso*, Lucas Rusnak, and **Jelena Tešić**. Advances in scaling community discovery methods for signed graph networks. *Journal of Complex Networks*, 10(3), 06 2022.

[15] *Maria Tomasso*, Lucas Rusnak, and **Jelena Tešić**. Cluster boosting and data discovery in social networks. In *Proceedings of the 37th ACM/SIGAPP Symposium On Applied Computing (SAC)*, 2022.

[16] *Mirna Elizondo*, *June Yu*, *Daniel Payan*, Li Feng, and **Jelena Tešić**. Data driven analysis of intervention effectiveness for covid learning loss in texas public schools. *(**Under Submission**)*, 2023.

[17] *Mirna Elizondo*, Rasim Musal, *June Yu*, and **Jelena Tešić** on behalf of N3C. Long covid challenge: Predictive modeling of noisy clinical tabular data. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, 2023.

[18] *MMM Rahman*, *Debojyoti Biswas*, and **Jelena Tešić**. Evirec: Efficient visual indexing and retrieval for edge crowd-sensing. In ***Submitted to a Conference***, 2023.

[19] *MMM Rahman* and **Jelena Tešić**. Evaluating hybrid approximate nearest neighbor indexing and search (hannis) for high-dimensional image feature search. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6802–6804, 2022.

[20] *MMM Rahman* and **Jelena Tešić**. Hybrid approximate nearest neighbor indexing and search (hannis) for large descriptor databases. In *2022 IEEE International Conference on Big Data*, pages 3895–3902, 2022.

[21] *MMM Rahman* and **Jelena Tešić**. Stratified graph indexing for efficient search in deep descriptor databases. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (**under review**)*, 2023.

[22] *Muhieddine Shebaro*, *Lia Nogueira de Moura*, and **Jelena Tešić**. Improving association discovery throughmultiview analysis of social networks. *Social Network Analysis and Mining (**under review**)*, 2023.

[23] *Muhieddine Shebaro* and **Jelena Tešić**. Identifying stable states of large signed graphs. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*, 2023.

[24] *Muhieddine Shebaro* and **Jelena Tešić**. Abcd: Algorithm for balanced component discovery in signed networks. In ***Submitted to a Conference***, 2023.

[25] *Muhieddine Shebaro* and **Jelena Tešić**. Scaling frustration index and corresponding balanced state discovery for real signed graphs. In ***Submitted to a Conference***, 2023.

[26] *Nicholas Warren*, *Ben Garrard*, Elliot Staudt, and **Jelena Tešić**. Transfer learning of deep neural networks for visual collaborative maritime asset identification. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, pages 246–255, Oct 2018.

[27] *Noah Dunstatter*, *Alireza Tahsini*, Mina Guirguis, and **Jelena Tešić**. Solving cyber alert allocation markov games with deep reinforcement learning. In Tansu Alpcan, Yevgeniy Vorobeychik, John S. Baras, and György Dán, editors, *Decision and Game Theory for Security*, pages 164–183, Cham, 2019. Springer International Publishing.

[28] *Taylor Mauldin*, Anne H. Ngu, Vangelis Metsis, Marc E. Canby, and **Jelena Tešić**. Experimentation and analysis of ensemble deep learning in iot applications. *2019 VLDB DMAH*, 5(1):133–149, 2019.

[29] Andreas Lommatzsch, Benjamin Kille, Özlem Özgöbek, Yuxiao Zhou, **Jelena Tešić**, Cláudio Bartolomeu, David Semedo, Lidia Pivovarova, Mingliang Liang, and Martha Larson. Newsimages: Addressing the depiction gap with an online news dataset for text-image rematching. In *Proceedings of the 13th ACM Multimedia Systems Conference*, MMSys '22, pages 227–233, New York, NY, USA, 2022. Association for Computing Machinery.

[30] Lucas Rusnak and **Jelena Tešić**. Characterizing attitudinal network graphs through frustration cloud. *Data Mining and Knowledge Discovery*, 6, November 2021.

[31] **Jelena Tešić**, Dan Tamir, S. Neumann, Naphtali Rishe, and Abe Kandel. Computing with words in maritime piracy and attack detection systems. In Dylan D. Schmorrow and Cali M. Fidopiastis, editors, *Augmented Cognition. Human Cognition and Behavior*, pages 434–444, Cham, 2020. Springer International Publishing.