Jelena Tešić: Research Narrative

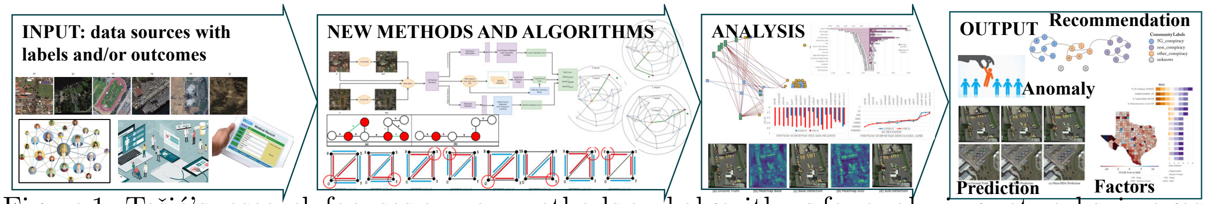
INPUT: data sources with labels and/or outcomes
NEW METHODS AND ALGORITHMS
ANALYSIS
OUTPUT
Recommendation
Anomaly
Prediction
Factors

Figure 1: Tešić's research focuses on new methods and algorithms for analyzing networks, images, videos, EHR, surveys, records, ratings, and bioinformatics data in the wild.

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** algorithms and methods for addressing such challenges of extensive collections in the wild, as illustrated in Figure 1. Tešić leads the (Data Lab@TXST] and advises 5 out of 60 Ph.D. students in the department. She has ensured over $575,000 for research support as a PI and $1,450,000 as a co-PI while at TXST and published 22 peer-reviewed research papers. The references are cited as numbered in Tešić's CV (The *student author* names are in *italic*), and papers are accessible online via (pdf) links.

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 [14] (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 [13] (pdf)), and to discover fundamental cycle basis in large signed networks [2] (pdf). The effort to compare and contrast state-of-art community discovery on real signed graphs in [9] (pdf) led to the first scalable signed benchmark comparison to date and frustration cloud-based approach for cluster boosting for high modular signed graphs [10] (pdf). The research now focuses on solving NP-hard tasks at scale such as computing frustration [29] (pdf) and finding the largest balanced subgraph [30] (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 [11]pdf). The research work compared and contrasted the analysis results of millions of Twitter data using multiple graph construction processing approaches [11] (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 [27] (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 [16] (pdf) and [27] (pdf).

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 [20] (pdf) and [22] (pdf). The project has proposed new algorithms for overhead videos to detect anomalous activities [12] (pdf) and [17]pdf). The project evolved in the Ph.D. thesis work and several innovative and efficient contrastive learning algorithms (1) to improve object classification in previously unseen highly variable overhead datasets in [4] (pdf) and [23] (pdf), (2) to propose real-time effective object detection using compression and scale prediction in [7] (pdf), and (3) to produce domain-invariant features across aerial datasets using local and global components for domain adaptation and object classification for the task using progressive domain adaptation [23] (pdf) and [26] (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 [5] (pdf) and [6] (pdf) on millions of deep descriptors. The research builds upon the approach for crowd-sensing application [31] (pdf) and improves the indexing footprint with the new stratified graph approach [25] (pdf).

Semantic Segmentation Task in The Wild Tešić'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 [3] (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 adherence and corrosion from images obtained in the NASA SpaceX-21 experiment. The project assists scientists in deciding what countermeasures will work in space to fight metal corrosion.

Predictive Modeling of Noisy Tabular Data Tešić's designed the multi-feature importance analysis algorithm. The master student applied it to large-scale analysis of public data 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 [24] (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 in Texas public schools during the COVID-19 pandemic. 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 [28] (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 [1] (pdf).

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 groups in the department and the college and proposed data-driven solutions for their specific challenges [19] (pdf), [21] (pdf), [18] (pdf) [15] (pdf), [8] (pdf), and [32] (pdf). All but two of the 22 peer-reviewed papers published, ten papers under review, and three peer-reviewed extended abstracts since joining TXST are co-authored with 25 different students. Of 25, four are Ph.D. students she advises, and six were/are Ph.D. She advised CS students, two MSEC Ph.D. graduates, two CS M.Sc. graduates, two CS M.Sc. students, five undergraduates who worked in the Data Lab, and four REU students. Tešić ensured over $575,000 for research support as a PI and over $1,450,000 as a co-PI while at TXST. She has paused the tenure clock during the COVID pandemic.