

VISUAL ANALYTICS FOR DECISION MAKING IN  
PERFORMANCE EVALUATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Jieqiong Zhao

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2020

Purdue University

West Lafayette, Indiana

**THE PURDUE UNIVERSITY GRADUATE SCHOOL  
STATEMENT OF DISSERTATION APPROVAL**

Dr. David S. Ebert, Chair

School of Electrical and Computer Engineering

Dr. Niklas Elmqvist

College of Information Studies, University of Maryland

Dr. Alexander J. Quinn

School of Electrical and Computer Engineering

Dr. Edward J. Delp

School of Electrical and Computer Engineering

Dr. Melba M. Crawford

Lyles School of Civil Engineering

**Approved by:**

Dr. Dimitrios Peroulis

Head of the School Graduate Program

To my family and friends

## ACKNOWLEDGMENTS

Becoming a scientist is my childhood dream, and working on a Ph.D. degree is part of the journey. I am extremely grateful to my advisor Professor David S. Ebert for his continuous support and guidance throughout my Ph.D. study. He is always inspiring and encouraging when I come across obstacles in my research. He is an extraordinary mentor who sharpened my mind in conducting innovative research, improving clarity when presenting sophisticated ideas, and finding viable solutions to challenging research problems. His supervision helped me to develop all of these key characteristics for becoming a qualified researcher, allowing me to ultimately complete my dissertation.

I would like to express my sincere gratitude to my committee members for their dedication to my dissertation. Professor Niklas Elmqvist provided insightful and inspiring comments based on his expertise on my research topic. Professor Alex Quinn provided specific guidelines for the thesis statement and contributions. Professor Edward Delp provided valuable suggestions to improve the quality of my work. Professor Melba Crawford guided me on collaborative research on smart agriculture. Their valuable advice and comments have improved the contributions my work makes to visual analytics research and clarity of my writing.

I would like to thank all co-authors on the published papers used in the dissertation. I learned various skills and techniques from them. They are extraordinary researchers and collaborators, including: Morteza Karimzadeh, Abish Malik, Chittayong Surakitbanharn, Jiawei Zhang, Guizhen Wang, Luke S. Snyder, Hanye Xu, Ali Masjedi, Taojun Wang, Xiwen Zhang, Zhenyu Cheryl Qian, Melba M. Crawford, and David S. Ebert.

I would like to thank my VACCINE colleagues, who helped me during my Ph.D. study. They are extraordinary labmates who provided suggestions and comments

on my research, including post-docs Morteza Karimzaden, Abish Malik, Chittayong Surakitbanharn, Audrey Reinert, and Jingjing Guo; and colleagues and friends Jiawei Zhang, Guizhen Wang, Calvin Yau, Hanye Xu, Yang Yang, Chen Ma, Junghoon Chae, Sungahn Ko, and Shehzad Afzal. I also would like to thank the Purdue Terra team members who introduced me to and taught me about the amazing research on smart agriculture. It is a wonderful team that includes extraordinary researchers who conduct cutting-edge research to improve plant breeding with high-throughput field phenotyping. I would like to thank the chief, commanders, and the data analyst of local law enforcement agencies for their valuable domain feedback and suggestions on my work.

In the end, I would like to thank my lovely family members and friends. My parents and my aunt have provided me unconditional financial and mental support to pursue my dream. They taught me to become a brave, persistent, and responsible person even in unexpected situations. My younger sister and brother are the ones who understand my choice and have stood by my side. My friends Jinrui Miao, Jing Li, Di Wang, and Sriram Karthik Badam are always helpful whenever I turn to them.

This work is funded by the U.S. Department of Homeland Security VACCINE Center and the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy. The results and opinions stated in this work do not reflect the position of these entities.

## TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	ix
LIST OF FIGURES . . . . .	x
ABBREVIATIONS . . . . .	xiii
ABSTRACT . . . . .	xiv
1 INTRODUCTION . . . . .	1
1.1 Needs and Challenges in Performance Evaluation . . . . .	2
1.2 Visual Analytics for Performance Evaluation . . . . .	5
1.2.1 Organizational Employee Performance Evaluation . . . . .	5
1.2.2 Interactive Feature Selection and Model Evaluation . . . . .	8
1.3 Thesis Statement and Contributions . . . . .	10
1.4 Roadmap . . . . .	12
2 RELATED WORK . . . . .	14
2.1 Performance Evaluation in Organizations . . . . .	14
2.2 Interactive Sorting and Visualization . . . . .	15
2.3 Visual Analytics for Multi-Attribute Decision Making . . . . .	16
2.4 Feature Selection Methods . . . . .	17
2.5 Visual Analytics for Feature Selection . . . . .	18
2.6 Predictive Visual Analytics . . . . .	19
3 A VISUAL ANALYTICS APPROACH FOR EVALUATING EMPLOYEE PERFORMANCE IN PUBLIC SAFETY AGENCIES . . . . .	21
3.1 Domain Characterization . . . . .	23
3.1.1 Requirements Analysis . . . . .	23
3.1.2 Analytical Tasks . . . . .	24
3.2 Deriving Performance Metrics . . . . .	26

	Page
3.3 MetricsVis System . . . . .	30
3.3.1 Priority Adjustment View . . . . .	31
3.3.2 Performance Matrix View . . . . .	32
3.3.3 Group Performance View . . . . .	35
3.3.4 Projection View . . . . .	43
3.4 Evaluation . . . . .	44
3.4.1 Use Case 1 . . . . .	45
3.4.2 Use Case 2 . . . . .	46
3.5 Domain Expert Feedback . . . . .	48
3.6 Discussion . . . . .	49
3.7 Conclusion and Future Work . . . . .	53
4 AUTOMATIC PERFORMANCE WEIGHTS LEARNING DRIVEN BY USER- GUIDED RANKING . . . . .	54
4.1 MetricsVis II System . . . . .	55
4.1.1 Workflow . . . . .	56
4.1.2 User Interface . . . . .	56
4.1.3 Interactions . . . . .	58
4.1.4 Weights Learning . . . . .	59
4.2 Weights Comparisons by Subjective Ratings . . . . .	61
4.3 Qualitative User Evaluation . . . . .	62
4.4 Discussion . . . . .	64
4.5 Conclusion and Future Work . . . . .	65
5 INTERACTIVE FEATURE SELECTION AND REGRESSION MODEL EVALUATION FOR HYPERSPECTRAL IMAGES . . . . .	67
5.1 Background . . . . .	68
5.2 Design Goals . . . . .	69
5.3 FeatureExplorer . . . . .	71
5.3.1 Workflow . . . . .	71
5.3.2 User Interface . . . . .	73

	Page
5.3.3 Regression Models . . . . .	75
5.4 Case Study . . . . .	76
5.5 Conclusion and Future Work . . . . .	77
6 CONCLUSIONS AND FUTURE WORK . . . . .	79
REFERENCES . . . . .	83
VITA . . . . .	93

## LIST OF TABLES

Table	Page
3.1 Eight parameters in evaluation of each offense. . . . .	27
3.2 Sample survey result for weights of 27 offense categories based on a range from zero to a hundred. . . . .	29
3.3 Comparison of dandelion glyph versus other glyphs in small multiple settings.	40
5.1 Comparison of average $R^2$ for 100 trials among multiple regression models on 10 datasets. . . . .	74

## LIST OF FIGURES

Figure	Page
1.1 The visual analytics process proposed by Keim et al. [7]. The process highlights that the interactive data exploration environment integrates well designed visualizations and data models. . . . .	6
1.2 The pipeline of interactive feature selection and regression model evaluation in the FeatureExplorer system. . . . .	9
1.3 Overview of core chapters composing this dissertation. Chapter 3 introduces a visual analytics approach, MetricsVis, to visualize multiple attributes of employee performance at and between multiple levels. In Chapter 4, the MetricsVis II system is introduced, which extends the original MetricsVis system to relate subjective preferences to quantitative measurements of employee workload using a pair-wise ranking algorithm. Chapter 5 presents a visual analytics approach supporting interactive feature selection and model evaluation, which enables users to identify influential features contributing significantly to a prediction model. . . . .	13
3.1 Illustration of MetricsVis system diagram with three modules: data processing, views, and visual analytical task categories. . . . .	26
3.2 The rating distribution of two sampled severe criminal offenses, burglary and homicide, from police officers and citizens. In a histogram, the x-axis shows the rating scale from zero to one hundred, and the y-axis shows the count of each score. The black lines denote the averages. . . . .	28
3.3 A sample row in priority adjustment view: designed for law enforcement agencies. . . . .	31
3.4 The performance matrix shows the employees (columns) and job types (rows). The matrix is sorted based on the total score of employees and job types. Darker colors encode higher values. . . . .	33
3.5 The mapping of comparison tasks, visual encoding, and sorting interactions for the performance matrix view. . . . .	34
3.6 The mapping of comparison tasks, glyphs, and visual encoding for the group performance view. . . . .	36

Figure	Page
3.7 The transformation steps from a table to dandelion glyphs. (1) Get the union of the top five categories in both groups. (2) Order the categories by total in descending order. (3) Apply the logarithmic transformation to the total count. (4) Dandelion glyphs for two groups. . . . .	37
3.8 The five examples of small multiple glyph to represent the multi-dimensional data attributes of two groups. . . . .	39
3.9 The two radial layout visual representations in the group performance view: dandelion glyphs and stacked radar glyphs. The glyphs show the list of criminal incidents responded to by <i>A Day shift</i> and <i>B Night shift</i> . (a) Highlight of <i>OWI</i> incidents in dandelion glyphs. (b) Stacked radar glyphs show the contribution of each member. (c) Selection of Officer <i>1449</i> in <i>B Night shift</i> . (d) Highlight of Officer <i>1449</i> in performance matrix.	41
3.10 MetricsVis overview: The priority adjustment view (2) encodes the crowd-sourced crime severity ratings from police officers and citizens (perceived importance of factors); the red dots indicate the currently assigned weights used in the evaluation metrics. The projection view (6) shows the dimensionality reduction results. The group performance view (5) contains three visual representations that show an overview of group performance and the contribution of each member. The performance matrix view (3) displays the individual employee performance with employees in columns and job types in rows (here, employees are sorted based on their group first and then their total performance scores). The control panel shows the filters (1) and grouping method (4) applied in use case 1. . . . .	44
3.11 Day ( <i>AD, BD</i> ) and night ( <i>AN, BN</i> ) shifts have significant differences in drug abuse and <i>OWI</i> incidents for self-initiated incidents. . . . .	47
3.12 The relationship between analytical tasks (rows) and MetricsVis views and low-level interaction categories [88] (columns). Cell shading quantifies how a particular view or an interaction contributes to the analysis process of a task. . . . .	50
4.1 Additional visual representations in the MetricsVis II interface: (1) subjective ratings from multiple supervisors; (2) comparisons of total scores calculated based on updated weights and previous weights; (3) indication of selected individuals that are used in the weights learning. . . . .	57

Figure	Page
4.2 Colored cells displaying the derived weights based on rankings provided by multiple supervisors. The column header includes the rater information, and the row header shows the offense categories. The first row under the column header displays how many employees are evaluated by one rater. The first column besides the row header contains the average weights obtained by surveying police officers. The orders of both raters and offense categories are determined by a hierarchical clustering algorithm. . . . .	62
5.1 The components diagram of FeatureExplorer. . . . .	71
5.2 FeatureExplorer overview: (A) the control panel with a list of unselected features, a list of selected features, a regression button, an automatic feature selection button; (B) feature correlation panel with a correlation matrix and a scatterplot; (C) evaluation panel with a scatterplot of ground truth and predicted values, a horizontal bar chart showing the importance score of each feature, a histogram showing the frequency of used pertinent wavelengths, a table displaying the results with and without feature selection.	72
5.3 Case study using FeatureExplorer for two hyperspectral datasets. . . . .	76

## ABBREVIATIONS

CAD	Computer-Aided Dispatch
KDE	Kernel Density Estimation
LiDAR	Light Detection and Ranging
ML	Machine Learning
OWI	Operating While Intoxicated
RFE	Recursive Feature Elimination
RMS	Record Management System
RMSE	Root Mean Square Error
SVM	Support Vector Machine
SVR	Support Vector Regression
UAV	Unmanned Aerial Vehicle
VA	Visual Analytics
VNIR	Visible Near-Infrared

## ABSTRACT

Zhao, Jieqiong Ph.D., Purdue University, May 2020. Visual Analytics for Decision Making in Performance Evaluation. Major Professor: David S. Ebert.

Performance analysis often considers numerous factors contributing to performance, and the relative importance of these factors is evolving based on dynamic conditions and requirements. Investigating large numbers of factors and understanding individual factors' predictability within the ultimate performance are challenging tasks. A visual analytics approach that integrates interactive analysis, novel visual representations, and predictive machine learning models can provide new capabilities to examine performance effectively and thoroughly. Currently, only limited research has been done on the possible applications of visual analytics for performance evaluation. In this dissertation, two specific types of performance analysis are presented: (1) organizational employee performance evaluation and (2) performance improvement of machine learning models with interactive feature selection. Both application scenarios leverage the human-in-the-loop approach to assist the identification of influential factors. For organizational employee performance evaluation, a novel visual analytics system, MetricsVis, is developed to support exploratory organizational performance analysis. MetricsVis incorporates hybrid evaluation metrics that integrate quantitative measurements of observed employee achievements and subjective feedback on the relative importance of these achievements to demonstrate employee performance at and between multiple levels regarding the organizational hierarchy. MetricsVis II extends the original system by including actual supervisor ratings and user-guided rankings to capture preferences from users through derived weights. Comparing user preferences with objective employee workload data enables users to relate user evaluation to historical observations and even discover potential bias. For interactive

feature selection and model evaluation, a visual analytics system, FeatureExplorer, allows users to refine and diagnose a model iteratively by selecting features based on their domain knowledge, interchangeable features, feature importance, and the resulting model performance. FeatureExplorer enables users to identify stable, trustable, and credible predictive features that contribute significantly to a prediction model.

## 1. INTRODUCTION

Imagine that a manager has to decide which of their 10,000 employees should receive a bonus. The manager needs to evaluate the performance of all employees based on their contributions to the development of an organization. However, distinct positions and job requirements make direct comparisons among employees challenging. In smart agriculture, hyperspectral images are collected to predict the phenotypic traits of plants. Which hyperspectral features matter the most in the prediction? Performance evaluation, as described in the examples above, is used by a wide range of decision makers in a variety of scenarios, such as workforce optimization or improving the accuracy of machine learning models. During a performance evaluation, decision makers rely on their domain knowledge, recorded observations, and historical data to quantify performance. However, quantifying overall performance is difficult when a decision maker must consider both multiple performance related factors and the relative contribution of each factor. Often, there are too many factors for a decision maker to meaningfully consider and evaluate.

The problem is thus: how can a decision maker make sense of large-scale, multi-dimensional datasets during performance evaluation? Visual analytic tools and techniques offer a possible solution, as these tools are designed to aid in the analysis of large-scale multi-dimensional data [1]. So far, only limited research has been done on the possible applications of visual analytics for performance evaluation. This dissertation proposes visual analytics systems to assist domain experts when evaluating current performance and predicting future performance by identifying influential performance-related factors.

In this chapter, we present the needs and challenges of performance evaluation in Section 1.1; this is followed by an explanation of the role of visual analytics can play in

performance evaluation in Section 1.2. The thesis statement and unique contributions are described in Section 1.3. Finally, Section 1.4 lists the roadmap of this document.

## 1.1 Needs and Challenges in Performance Evaluation

Performance analysis is a multi-faceted decision-making problem. For instance, to improve the performance of a machine learning model, it is necessary to identify which features play important roles in the prediction, and the ways that changing a given feature value would impact the prediction. In organizational employee performance evaluation, evaluators need to consider various parameters and metrics that contribute to the ultimate evaluation outcome. Each factor is weighted differently during the evaluation process, and a subset of factors may have a joint effect on the performance outcome. In addition, the relative importance (weighting) of factors may vary under different conditions or scenarios. Therefore, the evaluation process becomes complicated when evaluators need to consider numerous supporting factors and determine how to prioritize them.

Traditional performance evaluation approaches rely on a fixed number of metrics to inform an evaluation. These systems are frequently unable to support diverse ranges of evaluation criteria or perform real-time evaluation. Moreover, insufficient performance measurements may result in an incomplete understanding of the performance in question, and may even introduce bias as users tend to use their own background knowledge and subjective opinions to infer the values of missing measurements. Ranking or sorting the performance of an individual data item is a broadly adopted solution that a decision maker can use to understand the contributions of elements in the system. For instance, in the context of operational management or strategic planning, an administrator needs to answer the following questions: Who are the best performers and what is the supporting evidence? Which specific activities need noticeable improvement? How can the overall performance be improved?

Such questions cannot be answered using the traditional evaluation practice of simply ranking the overall performance of individuals. It is necessary for evaluators to gain a thorough understanding of performance by interactively ranking and sorting factors contributing to the overall performance. The problem then becomes identifying methods to derive comprehensive and quantifiable performance *evaluation metrics*, composed of both performance-related factors and associated weights. Visual analytics based performance evaluation approaches offer techniques to overcome the existing drawbacks of incorporating both quantitative measurements from data records and subjective feedback from users, enabling the interactive investigation of performance. In this dissertation, we will focus on two specific types of performance analysis: (a) organizational employee performance evaluation and (b) performance improvement of machine learning models with interactive feature selection. To clarify the terms used in describing the multi-dimensional or high-dimensional data in both applications, the terms *factors*, *attributes*, and *features* are used to denote the *dimensions*.

**Organizational Employee Performance Evaluation** Performance evaluation is usually applied as a tool to understand the performance of individuals within a system. The result of a performance evaluation allows an evaluator to recognize the strengths and weaknesses of an organization and take appropriate action if improvements are needed. An effective performance evaluation system with clearly defined goals and prompt feedback is an essential tool for organizations to improve their productivity [2], especially with limited resources and personnel. Characterizing employee, unit, and organizational performance requires a decision maker to consider multiple facets, including economic return, social impact, sustainability, and team and individual productivity. This performance data may be stored as subjective reports, financial statements, or employee evaluations.

It is challenging to develop an appropriate method of integrating complex data, including qualitative, quantitative, and subjective data, into an accurate representation

of organizational performance. This task is further complicated when teams within the organization have geographically and temporally distinct workloads as well as various positions and specialties (e.g., patrol officers vs. detectives); this complication commonly occurs in public safety organizations where there are temporal differences in the number of reports a unit may respond to.

Applying standardized evaluation factors and quantitative metrics can help overcome subjective biases caused by personal traits [3,4]. Such evaluation metrics should account for the importance of each task in accomplishing organizational objectives. In addition, the evaluation metrics must accommodate the perspectives of team leaders across various departments. Thus, it can be beneficial to interactively analyze, visually explore, accurately weight, compare, and evaluate employee performance in the context of organizational hierarchy as a way to support comprehensive and holistic evaluation.

**Interactive Feature Selection and Model Evaluation** In scientific research, improving the performance of machine learning models often requires analyzing the way in which target values respond to changes in input features. Because features usually do not contribute equally to a prediction model, it is critical to identify features that are substantial in the prediction. Most machine learning models assume that input features should be independent and identically distributed; however, this is often not the case in practice. The collinearity of features can increase their redundancy in feature space, which can then reduce the performance of machine learning models as well as increasing computation time. Many automatic feature selection algorithms exist to handle that problem; however, users may question why some features are selected and others are not due to having insufficient information about the decisions. Explainable machine learning and artificial intelligence algorithms need to increase the understandability and interpretability of models in order to help users understand why certain decisions are made [5, 6]. Furthermore, it is beneficial for users to interactively examine the feature space and understand the predictability of

features. Adopting a visual analytics approach that supports comparative analysis of models and the feature space can assist researchers in gaining comprehensive insights in modeling and selecting the best-fitting models.

## **1.2 Visual Analytics for Performance Evaluation**

As mentioned in the previous section, performance evaluation is multi-faceted and dynamic in various circumstances, demanding a comprehensive analytical environment to enable detailed interactive investigation of performance. The inherent capabilities of visual analytics incorporate both exploratory data analysis and domain knowledge from user feedback, allowing it to support complicated performance evaluation requests. Keim et al. [7] define the visual analytics process (Fig. 1.1) as an interactive knowledge discovery process with the assistance of visual and automated data exploration. Specifically, human-in-the-loop or human-centered visual analytics approaches are developed to enhance users' data exploratory experience with seamless integration of visualizations and data models into their work routine [8]. Often, automated models are applied to learn from user interactions for sophisticated analysis regarding large data, such as social media data [9], time-series financial data [10], and textual data [11]. We have created interactive visualization systems to support the exploration of performance-related data records to achieve sophisticated and thorough performance evaluation goals while accounting for multiple factors and relationships among factors and data samples.

### **1.2.1 Organizational Employee Performance Evaluation**

Organizations need to evaluate their employees' performance, both at any organizational level and among groups of employees performing similar jobs at potentially different locations and time periods. A team leader may be knowledgeable about the workload and job-handling ability of each team member in their unit, based on personal interactions and job activity reports. However, a team leader may still struggle

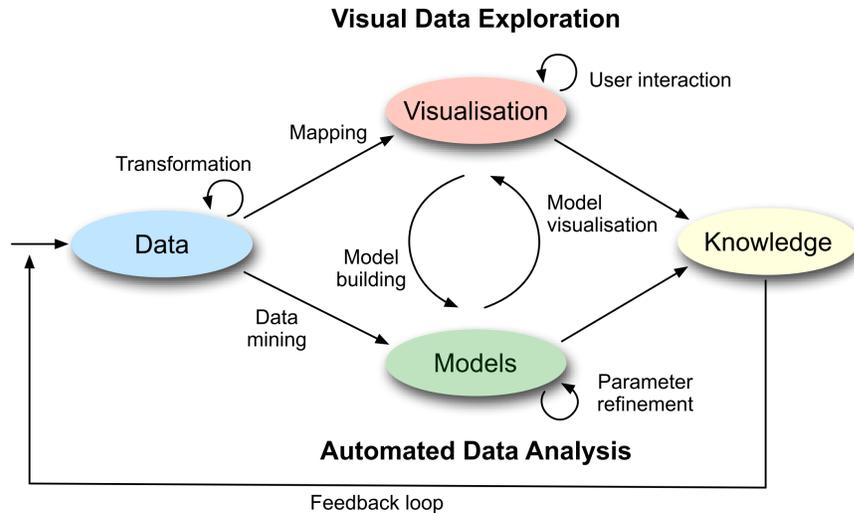


Fig. 1.1. The visual analytics process proposed by Keim et al. [7]. The process highlights that the interactive data exploration environment integrates well designed visualizations and data models.

to gain meaningful insights when comparing different aspects of ‘good’ performance both within their team and against other units; likewise, organizational leaders may struggle to compare several units or make organization-wide comparisons. In addition, the completion of different tasks can contribute differently toward ultimate organizational objectives. For example, employees may either engage in self-initiated activities, such as seeking new sales, or accomplish assigned jobs, i.e. fulfilling orders. Employee performance can be partially rated on self-initiated activities, which indicates proactivity. In law enforcement agencies, the chief and commanders have to consider the relative contributions made by an individual officer based on their effectiveness when handling emergency calls (dispatched) in addition to their ability to prevent crimes (self-initiated).

To support the interactive exploration and evaluation of employee performance at multiple scales (individuals, teams, and the entire organization), we developed a visual analytics system — MetricsVis (Chapter 3) — to support the comprehensive investigation of performance measures. Discussions with domain experts in a law

enforcement agency regarding its quarterly subjective review process revealed that each supervisor only provides ratings and subjective feedback for one team. Thus, they lack reliable evidence to compare their own team's performance with other teams. A better understanding of both individual and team performance, as well as workload demand, can be the first step to improving overall organizational performance and work efficiency.

The MetricsVis system summarizes generic evaluation tasks in performance-related analyses and provides customized visualization components to support dynamic evaluations and comparisons of individual, team, and organizational performance. The system considers the in-depth requirements of performance evaluators, including numerical contributing factors in different relations, the evaluation rating knowledge of domain experts, and the organizational hierarchy. It also combines these evaluation concerns with hybrid evaluation metrics (details presented in Section 3.2) to achieve a data supported performance evaluation approach that alleviates subjective bias and incomplete understanding of subjects toward performance assessment. To demonstrate the usability of the MetricsVis system, two case studies from medium-sized law enforcement agencies are described to highlight its broader applicability to other domains.

Besides visualizing multi-dimensional performance data at multiple levels to expedite performance evaluation, dynamic adjustments of evaluation metrics are provided to satisfy the disparity between organizations. Users can interactively tune the evaluation metrics based on their preferences in order to predict future performance. However, if too many factors are involved in the manual fine-tuning process, users may find the process tedious and repetitive. To speed up this process, two methods are applied to obtain the suggested importance of performance-related factors: (a) an online survey to collect service recipients' and employees' opinions about the importance of each performance-related factor (described in Chapter 3 MetricsVis), and (b) using machine learning models to obtain the importance of each factor (discussed in Chapter 4 MetricsVis II).

### 1.2.2 Interactive Feature Selection and Model Evaluation

Machine learning models are increasingly used to analyze an overwhelming amount of multi-dimensional data and provide predictive analysis. However, many of these models are used as black boxes (primarily because of the way current computational libraries present the models/results). Therefore, domain users who do not have training in machine learning may not understand how the results are generated, and as a consequence may not trust the models. These problems are further complicated by insufficient data samples and the curse of dimensionality. Feature selection is often adopted to improve these models by identifying relevant features that make the most significant contribution to the prediction results while removing noisy, irrelevant, and less important features. We propose a visual analytics system, FeatureExplorer (Chapter 5), to support interactive feature selection by inspecting different rankings of features generated by two feature selection algorithms. Through examining the importance of features, users can discern the predictability and interchangeability of particular features. In addition, users can interactively add or remove features to investigate the impact of a subset of features on a model to verify their hypotheses.

Specifically, domain experts can select a subset of features based on their domain knowledge. When users manipulate the input features of a learning model, quantifiable measurements are necessary to indicate the effectiveness of a model, such as accuracy, root-mean-square error, relative errors, and  $R^2$  [12]. We used root-mean-square error and  $R^2$  as performance measurements for a learning model, since numerical ground truth (e.g., the biomass of plants) was used in our case studies. Thus, users can add or remove features based on their subjective judgements and then verify their hypotheses based on whether a feature is critical to the prediction or not. This working pipeline of manipulating the feature space and validating with experiments (shown in Fig. 1.2) has been incorporated into FeatureExplorer to identify the most influential subset of features. In the pipeline, two categories of feature selection methods are deployed: filter methods and wrapper methods. For filter methods,

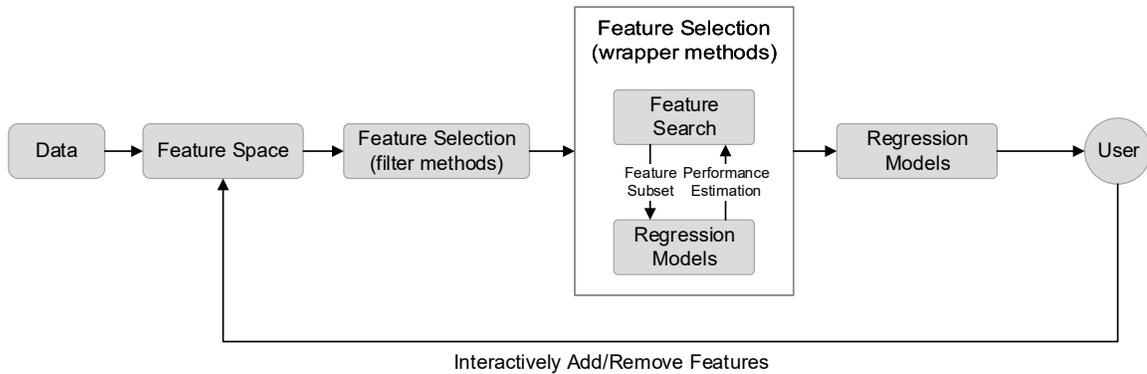


Fig. 1.2. The pipeline of interactive feature selection and regression model evaluation in the FeatureExplorer system.

Pearson’s correlation coefficient is applied to indicate the collinearity between features. For wrapper methods, the recursive feature elimination is utilized to indicate the predictability of features for a given regression model. Currently, the two specific feature selection methods are selected based on domain practice. In addition, it is an initial step toward verifying the utility of the pipeline with domain experts.

To demonstrate the use of the FeatureExplorer system, a case study conducted in collaboration with remote sensing experts is presented to show the prediction of plant biomass using features extracted from hyperspectral images. Feature mining techniques, including feature generation, feature selection, and feature extraction, are typically applied on hyperspectral images to identify the optimal feature space [13]. Remote sensing experts work intensively in feature generation and feature extraction on collected raw data in order to construct meaningful and substantial features for prediction. We aim to provide an interactive feature selection system that fulfills their requirements by extensively examining the predictability of those features. In FeatureExplorer, we focus on visualizing the importance of features provided by different feature selection methods to assist users in refining the feature space. Decreasing the number of features not only improves the performance of models but also reduces the computation complexity. Additional feature selection methods and regression

models can be included in the FeatureExplorer pipeline to satisfy the extended needs in comparing different feature selection methods and learning models. The pipeline (Fig. 1.2) can be augmented by plugging in various filter feature selectors (e.g. mutual information), wrapper feature selectors (e.g. forward selection), and regression models.

### 1.3 Thesis Statement and Contributions

This dissertation includes a set of visual analytics (VA) systems that we developed to assist the exploration of multi-dimensional performance data to enable domain experts to analyze performance-related factors and influence final performance evaluations by interactively customizing input factors. Two specific use cases in performance analysis are explored: organizational performance and feature selection for machine learning models. During an employee performance analysis (Chapter 3 MetricsVis), we developed a VA system that organization managers and supervisors can use to conduct a performance evaluation by inputting weights of performance-related factors to better reflect their opinions of the relative significance of particular measurements (e.g. the impact of a task on individual productivity). Furthermore, the hierarchical structure of data items is considered and incorporated into the visual design to illustrate the performance at and between multiple levels: individual data items, aggregated groups, and the entire dataset. The structure can be assigned by inherent relationships (e.g. members in a team) or by clustering algorithms that group similar items together by shared patterns. In interactive feature selection (Chapter 5 FeatureExplorer), machine learning algorithms are deployed to automatically provide a systematic ranking of features. Though two specific cases are explored, they share the common theme of analyzing numerous factors contributing to the ultimate performance and evolving relative importance of these factors in dynamic conditions. The thesis statement is as follows:

*Interactive exploration analysis, summarization, and visualization of multi-level and multi-dimensional data can assist in reducing the efforts required to identify influential factors and analyze multi-level comparisons necessary to improve user comprehension, performance evaluation, and decision making.*

The core contributions of this dissertation are a collection of innovative visual analytics approaches that integrate interactive analysis, novel visualizations, and predictive machine learning models to support informed decisions during exploratory performance evaluation tasks. The combination of novel visualizations displaying multi-dimensional data at multiple levels creates a unique interactive environment to increase effectiveness and comprehension of performance analysis. In addition, predictive machine learning models are deployed to enhance user experience when identifying influential factors through interactive user feedback. Specific contributions of this work include the following:

- A visual analytics approach, MetricsVis, that summarizes and visualizes employee performance at and between multiple levels of an organization (i.e., individual, group, and individual contributions to a group) to enable dynamic performance evaluation [14]. MetricsVis visualizes multiple data attributes at multiple levels to enhance a user's interpretation of the overall performance of an organization.
- In MetricsVis, hybrid evaluation metrics are applied to obtain the overall performance of individuals. The set of hybrid evaluation metrics integrates both *quantitative* measurements of achievements based on observed performance-related factors (data-driven) and *qualitative* subjective ratings for the relative importance of each factor based on online survey results collected from employees and service recipients. The performance-related factors are derived from existing job performance analysis techniques and employee activity records from public safety agencies [14, 15].

- An alternative approach to deriving the weights for performance-related factors in the hybrid evaluation metrics is applied to capture the preferences from users (Chapter 4 MetricsVis II). The approach seeks to relate subjective rankings of employees to their quantitative measures of workload. Users can manually adjust the rankings of employees, and then a ranking algorithm is utilized to learn the weights of performance-related factors based on the user supplied feedback.
- A visual analytics approach, FeatureExplorer, to support interactive feature selection and model evaluation with visualizations for ranked features provided by multiple algorithms [16]. FeatureExplorer visualizes the relationship among features and the contribution of features in machine learning models to facilitate the identification of optimal combinations of features.

## 1.4 Roadmap

This dissertation centers on applying visual analytics approaches as new capabilities to help decision makers make comprehensive and effective decisions during performance analysis. These visual analytics approaches integrate visual representations to assist multi-level comparisons and predictive machine learning models in order to expedite the identification of influential factors. Fig. 1.3 shows the structure of the main components in this dissertation. The MetricsVis system, presented in Chapter 3, supports dynamic performance evaluation at multiple organizational levels. In Chapter 4, a ranking algorithm is added into the primary MetricsVis system to automatically derive weights associated with performance-related factors, allowing users to rank some employees based on their personal preference and predict the performance of the rest. The VA system, MetricsVis II, integrates the automatic weights learning driven by user-guided ranking and is named after the original MetricsVis system. A visual analytics approach (FeatureExplorer) supporting interactive feature selection and regression model evaluation is described in Chapter 5. The core concept

of MetricsVis II and FeatureExplorer is leveraging the human-in-the-loop to identify influential factors with the aid of predictive machine learning models. In contrast to the pipeline in FeatureExplorer (shown in Fig. 1.2) which improves learning models by manipulating the feature space, MetricsVis II allows users to manipulate the ground truth (i.e., labels) of training data.

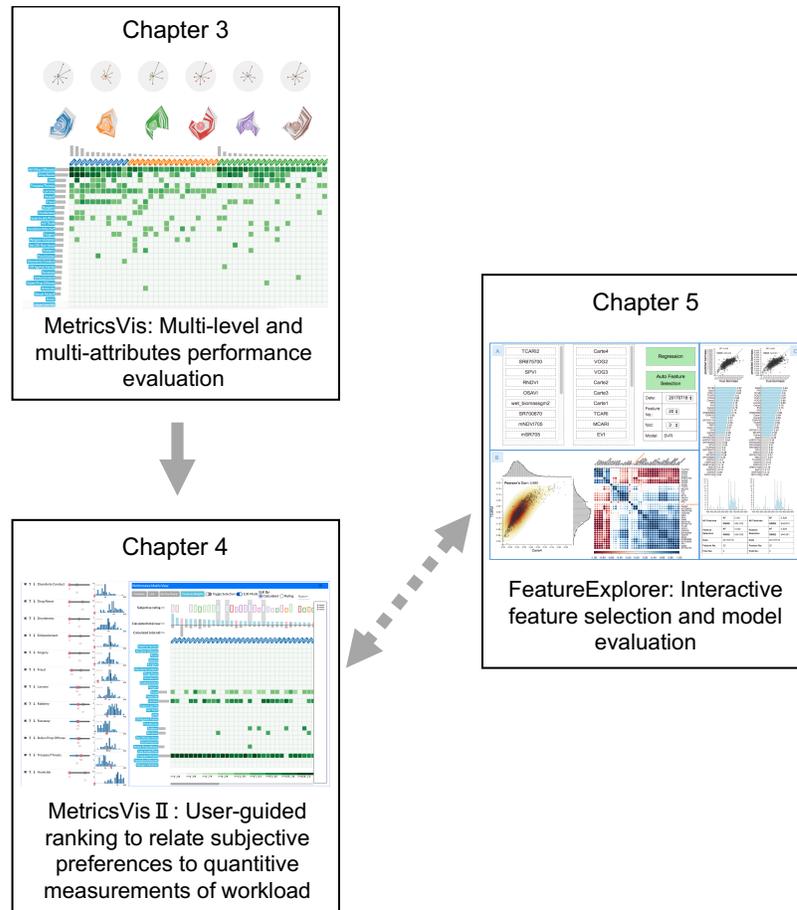


Fig. 1.3. Overview of core chapters composing this dissertation. Chapter 3 introduces a visual analytics approach, MetricsVis, to visualize multiple attributes of employee performance at and between multiple levels. In Chapter 4, the MetricsVis II system is introduced, which extends the original MetricsVis system to relate subjective preferences to quantitative measurements of employee workload using a pair-wise ranking algorithm. Chapter 5 presents a visual analytics approach supporting interactive feature selection and model evaluation, which enables users to identify influential features contributing significantly to a prediction model.

## 2. RELATED WORK

In this chapter, we review work related to applying visual analytics approaches to solve problems in two research fields: (1) evaluating organizational performance in public safety agencies and (2) interactive feature selection and evaluation of regression models. Both of these fields share the common themes of representing multi-dimensional data attributes/features and fulfilling complex exploratory tasks solicited by domain experts. For employee performance evaluation, we focus on performance evaluation in organizations (Section 2.1), interactive sorting techniques to aid in effective ranking and comparison of multi-dimensional data (Section 2.2), and analytical systems for making sense of multi-attribute data (Section 2.3). For interactive feature selection, we focus on algorithms to offer automatic feature selection (Section 2.4), and VA approaches to support interactive feature selection (Section 2.5). In addition, we describe related work in predictive visual analytics in Section 2.6.

### 2.1 Performance Evaluation in Organizations

Performance evaluation in organizations needs to compare performance observations with expectations, reveal barriers preventing the desired performance, and generate action plans for either maintenance or improvement in order to achieve organizational objectives [17,18]. Performance appraisal systems assist decision makers in realigning employee performance to meet the evolving organizational objectives [19]. An ongoing problem in organizational performance is designing metrics to measure employee effectiveness and productivity [20,21].

Several researchers have derived taxonomies to evaluate employee performance-related factors that characterize the performance of individual employees [22–28]. The results are lists of generalizable evaluation factors (e.g. task performance, or-

ganizational citizenship behavior, and counterproductive work behavior) that could be adopted in diverse evaluation scenarios. MetricsVis leverages dynamic evaluation factors which users can customize based on organizational objectives, and supports interactive variable weighting to reflect the relative importance of each task/job type (i.e., factor). In addition to evaluating individual performance, the hierarchical structure in an organization has a fundamental impact on the organization’s behavior and management [29, 30]. Our MetricsVis system supports the exploration of individual performance as well as team- and organization-level performance with respect to the organizational hierarchy.

## 2.2 Interactive Sorting and Visualization

For convenient ranking and visual comparison among data items and their attributes, many research studies have utilized bar graphs and interactive sorting techniques. Tabular visualization is widely used due to its simplicity, clarity, and familiarity. Reorderable matrices [31] are designed to efficiently explore associations between hundreds of data items and data attributes. Usually, a permutation (sorting or clustering) method is provided to highlight the pattern of multidimensional data [32]. Many tabular visualization techniques are mainly designed for numerical data and bar chart representation. For example, Table Lens [33] utilized the focus + context technique on large relational tables; ValueCharts [34] was designed to support hierarchical structure additive linear models; and LineUp [35] was designed as a multi-attribute ranking system that considers a combination of attributes and timeframes. Conversely, the Parallel Sets [36] system focused on the interactive exploration of categorical attributes. Specifically, reorderable matrices efficiently explore associations between hundreds of data items and data attributes [31]. Furthermore, permutation (sorting, clustering) methods help highlight similarity patterns in these matrices [32].

Interactive ranking and sorting is an active research area. Some techniques sort all data attributes simultaneously and use linkages across all attributes to highlight

the same data entry [37, 38]. Timespan [39] supports hierarchical reordering, which sorts data samples based on the priority of a data attribute. However, MetricsVis provides conventional sorting on a reorderable matrix that allows flexible rearranging of attributes (i.e., factors such as job categories) and data items (employees), because we found that these components are more familiar to end users.

MetricsVis computes rankings based on attribute weights provided by end users, as opposed to several recent systems that leverage user-assigned data ordering to reverse engineer the weights [40, 41]. These systems require users to interactively update the overall ranking of data samples and inspect the validity of weights. MetricsVis, however, requires domain experts to have a good understanding of the weights, which is coherent with the goal of aligning with the priorities of an organization for dynamic evaluation purpose.

### 2.3 Visual Analytics for Multi-Attribute Decision Making

Researchers have presented several VA systems to facilitate the exploration and understanding of multi-dimensional data. Zhao et al. [42] developed SkyLens, a VA solution that enables comparison of multi-dimensional data through multiple coordinated views, while filtering out inferior data candidates. LineUp [35], perhaps the work most similar to ours, performs ranking visualization of multi-attribute data, and allows users to flexibly adjust weighting parameters to identify potential relationships. However, SkyLens and LineUp do not provide interactive visualizations to support multi-level performance comparisons, such as individual to group or group to organization, which is necessary for organizational evaluation. MetricsVis was designed with this key consideration in mind. In addition, LineUp utilizes bar charts to facilitate ranking comparison; MetricsVis employs radial layouts, which have outperformed tabular layouts when comparing data attributes [43] and provide compact visualization.

The software suite Tableau [44] can provide useful individual interactive data visualizations to explore relationships, trends, and rankings among multi-attribute data, such as pie, bubble, bar charts, treemaps, and tabular visuals. However, the user may not be able to generate a visualization that communicates the data most effectively to compare multi-level performance. MetricsVis provides compact and interactively linked visualizations specifically tailored for efficient, multi-level comparison of organizational performance metrics. For instance, MetricsVis allows users to view individuals with potentially similar performance through integrated clustering. Tableau does not support this. Furthermore, while Tableau can provide a hierarchical overview of an individual’s contribution to the group (and group to organization) with treemap visualization, MetricsVis supports simultaneous comparison of multiple attributes to the overall group with stacked radar charts, which can also be used to compare groups.

## 2.4 Feature Selection Methods

Feature selection methods are widely applied to remove irrelevant features and boost the performance of machine learning models. Feature selection methods can be generally divided into four categories: filter methods, wrapper methods, embedded methods, and hybrid methods [45]. The filter and wrapper categories are relevant to our work; therefore, we will focus on them here.

Typical filter methods are Pearson’s correlation coefficient [46], mutual information [47], and features are ordered by their relationship with the dependent variable (i.e., prediction target) using statistical measurements. The advantages of filter methods are straightforward to compute and can avoid overfitting. However, filter methods may not generate the optimal subset of features for a few reasons: (1) overlooking the relationship with other features (mainly focus on the relationship with dependent variable), (2) neglecting features that are less informative by themselves than combined with other features, (3) not considering the underlying learning models.

We used Pearson’s correlation coefficient to narrow down selected features to the ones with high linear correlation with the dependent variable. However, correlated but redundant features may be selected, and the coefficient is unable to characterize nonlinear relationships.

Wrapper methods are appropriate complements for filter methods, since wrapper methods use regression or classification models to find an optimal feature subset by iteratively adding or removing features. Many search algorithms such as sequential feature selection [48, 49], greedy search [47], and genetic algorithms [50] are designed to expedite the search for an optimal subset. A compelling advantage of wrapper methods is the flexibility to apply different learning models; therefore, users can select a preferred learning model. The combination of learning models (e.g. SVR) and wrapper methods (e.g. RFE) has traditionally been used for automatic feature selection [51, 52], and we implemented the combination in the FeatureExplorer system. However, in wrapper methods, the learning models are used as black boxes and overfitting may occur. Therefore, visual examination of the feature space is necessary in order to leverage domain knowledge in choosing meaningful features.

## 2.5 Visual Analytics for Feature Selection

Several visualizations have been proposed for feature selection, including correlation matrices [53], feature clustering [54], feature ranking [55–57], scatterplot matrices [58], and dimensionality reduction [59]. A few visual analytics systems have leveraged a combination of automatic and visual feature selection techniques. RegressionExplorer [60] is one such system for inspecting logistic regression models. Other systems have been proposed to support exploring linear relationships among features [61–63]. BEAMES [64] is another multi-model system that enables users to interactively compare different types of models with various hyper-parameters (e.g., logistic regression vs. Bayesian regression models), while allowing users to interactively weigh data instances and features. INFUSE [65] enables the ensemble of multiple feature

selection methods by visualizing feature importance as determined by various feature selection methods in a radial glyph. Our focus, however, is to support domain experts in efficiently reducing a high-dimensional feature space into key feature subsets, and tracing back the features to the underlying data (wavelengths) for incorporating domain knowledge.

Partition-based visual analytics systems [66, 67] primarily focus on the interactive exploration of local structures and relationships between independent and target variables, appropriate for lower feature space dimensions. They are aimed at closer inspection of limited numbers of selected features for optimal distribution partitioning and model building. However, our focus is on high dimensions (of both data instances and feature space). Our system’s integrated hierarchical clustering and matrix visualizations facilitate the quick identification of (a) influential feature subsets (either already selected or missing) for model building, (b) the interchangeable features within those subsets, and (c) detailed feature distribution and importance.

## 2.6 Predictive Visual Analytics

Predictive analytics is defined as the process of identifying patterns in input data and predicting the output using quantitative models [68]. Machine learning algorithms are often adopted as quantitative models in predictive analytics, since they can accurately capture the relationship between input and output data. In predictive visual analytics, interactive machine learning techniques and users’ desires to be involved in the modeling process require VA systems to explain and interpretation the modeling process at 3 stages: (1) before building a model, (2) building a new model, and (3) after building a model. There are two distinct approaches widely adopted in predictive visual analytics: (1) interpreting the input-output relationship and using machine learning algorithms as black-boxes, and (2) interpreting the internal logic of machine learning algorithms.

The first approach usually involves the first stage (before building a model) and the third stage (after building a model) in a modeling process. Krause et al. [69] show that investigating the input-output relationship can improve users' understanding of prediction outcomes. Manifold [70] is another example of such a visual analytics system that assists users in interactively interpreting the input-output relationships for several models at the same time. Prospector [71] explains how features impact prediction models by interactively exploring the partial dependence. In FeatureExplorer (Chapter 5), the internal logic of the machine learning model is not demonstrated; we increase users' interpretation of the input-output relationship by applying two feature selection algorithms at two stages of the modeling process: (1) demonstrating the correlation between features before a model is built, (2) showing feature importance after a model is built.

For the second approach, the interpretability of a model is increased by visualizing the internal logic of machine learning models, which is more conventional in predictive visual analytics [72]. Building a model such that its internal logic can be easily understood by end-users who are not familiar with the mechanisms of predictive models is critical. Two types of models are straightforward to understand: (1) rule-based models (e.g., decision trees adopted in BaobabView [73]), and (2) linear models [40]. In MetricsVis II (Chapter 4), a linear model is included to externalize users' preferences using weights. A higher value on specific weights indicates that an organizational performance evaluator may show a stronger preference for a given category, e.g. employees who handled more cases in a given job category.

### 3. A VISUAL ANALYTICS APPROACH FOR EVALUATING EMPLOYEE PERFORMANCE IN PUBLIC SAFETY AGENCIES

This chapter is based on papers published in IEEE TVCG 2020 and IEEE HST 2017: J. Zhao, M. Karimzadeh, L. S. Snyder, C. Surakitbanharn, Z. C. Qian, and D. S. Ebert, “MetricsVis: A visual analytics system for evaluating employee performance in public safety agencies,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1193–1203, Jan 2020. doi: 10.1109/TVCG.2019.2934603

J. Zhao, A. Malik, H. Xu, G. Wang, J. Zhang, C. Surakitbanharn, and D. S. Ebert, “MetricsVis: A visual analytics framework for performance evaluation of law enforcement officers,” in *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, April 2017, pp. 1–7. doi: 10.1109/THS.2017.7943468

Performance evaluation is critical for supervisors and managers to understand the performance of employees to improve the overall productivity of an organization. Currently, specialized visual analytics tools for harnessing multi-dimensional organizational data to facilitate effective employee performance evaluation are lacking. Current performance evaluation practices often apply subjective supervisory ratings with data tables listing the simple statistical summaries of entire departments and details regarding tasks completed by individual employees. Existing visual analytics applications support either multi-dimensional data visualizations or multi-criteria decision-making [35, 42] that treat individuals uniformly, ignoring the inherent hierarchical relationships and different teams or task types typical of public safety and many other organizations.

In this chapter, we present MetricsVis (Fig. 3.10), a visual analytics system for evaluating the performance of individual employees, teams, and the entire organization in public safety agencies. We designed the system iteratively with users from two medium-sized law enforcement agencies (representing similar-sized organizations in our study). We rooted our metrics in the existing organizational performance literature and adaptively tailored MetricsVis to meet the requirements of public safety organizations with groups of employees performing similar jobs but at different locations and times, resulting in different workloads that impact their contribution to organizational goals. Additionally, we formalized the analytical tasks, goals, and metrics; derived metrics; and surveyed organizational personnel and the public to decide on priorities of evaluation. We implemented multiple coordinated views in MetricsVis to support efficient, effective, and dynamic performance evaluation for multiple levels of an organization.

The MetricsVis system enables a holistic evaluation of organizational priorities versus actual achievements, and helps identify opportunities for improvement. Additionally, it facilitates the evaluation of strategic goals, expedites resource allocation (e.g. understanding which employees may need additional training or would be good trainers), and improves workload balance and individual employee performance. Specific contributions of our research and design are as follows:

- The mapping of the analysis of public-safety organizational performance evaluation into four visual analytical task categories.
- A novel system supporting interactive visual organizational performance analysis in public safety agencies based on hybrid evaluation metrics that integrate quantitative employee data and qualitative subjective feedback, and appropriate visual representations to support the four aforementioned visual task categories.
- A system evaluation with domain experts from two medium-sized law enforcement agencies to validate system usability.

In the remainder of this chapter, we first present the visual analytics task categories that were distilled through reviewing literature and collaboration with domain experts. What follows is a detailed description of deriving the hybrid evaluation metrics. Next, we describe the MetricsVis system in detail and demonstrate its utility through two case studies. Finally, the generalization and limitations of the MetricsVis system are discussed.

### **3.1 Domain Characterization**

We identified the general requirements for an effective and efficient performance evaluation system by reviewing the literature [17–21, 25, 29, 30, 74, 75], which informed our discussions with domain experts from law enforcement. We then mapped the refined requirements into four visual analytical task categories, as explained below.

#### **3.1.1 Requirements Analysis**

Assessing employee performance involves considering and integrating multiple performance-contributing factors to enable accurate comparison against organizational objectives. To satisfy the diverse and evolving requirements of different organizations, we decided to adopt dynamic evaluation metrics that can be refined based on user preferences. For clarification, evaluation metrics in our context are comprised of two aspects: the performance-contributing factors and their weights. Choosing performance factors is a challenging task specific to each domain. We derived these factors from a combination of (a) unstructured interviews with commanders and chiefs at law enforcement agencies and (b) taxonomies of task performance in work settings [22–25, 74, 75].

Organizational hierarchy affects the performance evaluation process. Evaluators' differing perspectives can hinder comparisons across the entire department, especially when traditional performance evaluation practices rely heavily on subjective ratings from management. Leaders may analyze their team's workload and performance

quality through personal interaction and job activity reports; however, it may still be difficult to compare different aspects of satisfactory performance between units across the organization. Though the evaluators can ensure unbiased judgment within their teams, variation across multiple evaluators is inevitable. Besides comparisons at the same level (individual versus individual, group versus group), we also need to evaluate the contribution of an individual to its group and a group to its organization. Therefore, an effective performance management system must support the performance evaluation at and between multiple levels of the organizational hierarchy.

We summarize these requirements from three independent perspectives: (1) dynamic performance metrics that can be adjusted by users to align with organizational objectives; (2) multiple levels including individual, team, and the entire organization; and (3) two relational contexts including comparisons at the same level as well as between two levels.

### 3.1.2 Analytical Tasks

The goal of MetricsVis is to enable the evaluation of individual employee, team, and organization effectiveness through the exploration of performance measures derived from digital activity records (quantitative, qualitative, and subjective). To accomplish this goal, MetricsVis was designed to address several visual analytical task categories for performance evaluation:

**T1 Evaluate individual employee performance:** The first challenge a team manager may encounter is aggregating and transforming activity reports and statistics into measures of subordinate performance. One approach to evaluating performance is to determine the frequency of different jobs accomplished by an employee, the difficulty and effort required for a certain category of job, and whether the job was self-initiated or dispatched. Since not all job types are equal in difficulty and effort, the option to weight each job type should be

incorporated. Additionally, a supervisor needs to select and compare multiple employees' performance to find patterns in low- and high-performing employees.

**T2 Evaluate group and team performance:** The second challenge is to understand the most influential factors that create successful teams for a variety of jobs. Some factors that may impact team effectiveness include location, manager, shift time of day, personnel proficiency levels, and time spent working. For instance, assigning police officers with experience in a certain geographic area to respond to calls in that area may increase patrolling effectiveness due to additional tacit knowledge providing an advantage over someone unfamiliar with that area. Additionally, understanding how these issues affect workload balance and morale serves to help optimize personnel allocation.

**T3 Investigate organizational workload:** Managers want to understand resource and personnel allocation strategies, pattern changes in services, and the effect on workload balance to increase overall organizational effectiveness. Exploring and comparing grouping factors (e.g. locations, time periods, and servicing patterns) can enable understanding of whether resource expenditure is aligned with organizational goals, discover unexpected drains on resources, and find excess personnel capacity in certain areas or during certain time periods. This information is crucial for advanced resource allocation strategies (e.g. dynamic allocation, request-based allocation), and evaluating the effectiveness of alternative strategies.

**T4 Evaluate department priorities:** If the priorities of the organization shift over time due to increased requests based on a particular service, the managers will be able to reflect these changes through adjusting the weights of the evaluation metrics. Stakeholders and managers may have different opinions about the importance of a job type or activity, and a good performance evaluation system should allow the administrator and managers to investigate the impacts of applying different evaluation criteria.

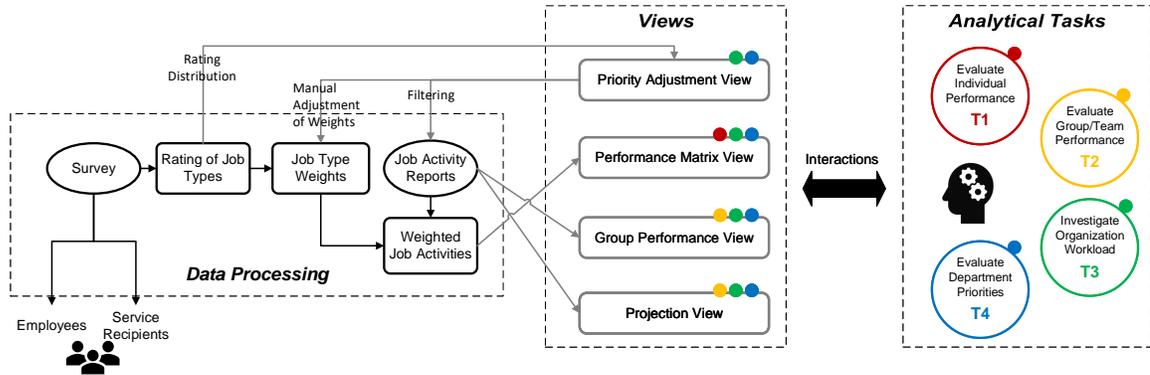


Fig. 3.1. Illustration of MetricsVis system diagram with three modules: data processing, views, and visual analytical task categories.

### 3.2 Deriving Performance Metrics

One of the key challenges of a successful performance appraisal system is quantifying the workload of employees and then deciding the contribution of specific jobs to the team or organizational objectives with appropriate scoring to reflect the priorities of a given organization. We describe our method, which (a) transforms job activity reports to workload descriptors and (b) transforms subjective feedback from employees and communities to qualitative measurements of contributions (shown in Fig. 3.1 Data Processing). To demonstrate the performance data extraction stage, we utilize activity records of law enforcement agencies as an example and explain the transformation process in detail. For that, we need to briefly describe the data source and several terms related to law enforcement agencies.

Law enforcement agencies typically use two related databases: computer-aided dispatch (CAD) and record management system (RMS). The CAD tables contain calls for service event information such as call nature, address, time, patrol units dispatched, etc. The calls usually fall into two categories: dispatched and self-initiated. The self-initiated calls are usually started by officers on their patrolling duties, whereas dispatched calls are assigned to the officers. RMS tables are concerned

with criminal **incidents** that have been written into reports by officers, including parameters such as date, location, offense committed, etc. We denote the calls ending with patrol service but not resulting in criminal incident report as *Call for Service events*, which need to be considered as a separate category. Dispatched activities and self-initiated activities should not be treated equally, since self-initiated activities are proactive behavior to prevent crimes and dispatched activities are responses to citizen requests; the option of filtering activities by behavior types (self-initiated versus dispatched) is extremely useful for evaluating the performance of patrol officers. Both tables store data in a multi-dimensional format in which every entry contains a report with its associated metadata.

Based on the taxonomy of major indicators of individual employee task performance [22–25, 74, 75], the top three common performance-related factors are **job completion, work quantity, and work quality**. Rooted in these factors and based on feedback from our users, MetricsVis utilizes offense categories from law enforcement agencies as the diversity of job completion, the number of cases responded to by officers as work quantity, and the crowdsourced survey rating for the seriousness of offense categories as a practical substitute for work quality.

Table 3.1.  
Eight parameters in evaluation of each offense.

Economic Loss to Victim
Economic Loss to Group
Economic Loss to Government
Economic Loss to Private Organization
Impact on Culture
Impact on Victim’s Mental Well-being
Impact on Victim’s Physical Well-being
Risk to Officer’s Life

We conducted this crowdsourcing online survey with two participating groups: police officers (59 participants) and community citizens (33 participants), each rating

the severity and economic impact of 27 offense categories on a Likert scale. The severity and economic impact were assessed by eight parameters listed in Table 3.1. A final rating of an offense category was assigned by the aggregated score of the eight parameters and scaled to a range from zero to one hundred. Each participant was asked to rate 216 questions in total (27 offense categories \* 8 parameters). During the survey, participants might skip some survey questions that they found difficult to rate. We replaced missing ratings with the average rating from all other participants. The average ratings of the majority of offense categories for both citizens and officers are shown in Table 3.2.

MetricsVis transformed these crowdsourced ratings to weights, which can be assigned either based on the average rating from the survey or on interactive adjustment from end-users. Ultimately, the overall performance of an officer is calculated as the summation of these weighted offense categories. In summary, the evaluation metrics are denominated as a set of hybrid evaluation metrics that contain (a) the quantitative measurement of employee achievements based on activity reports with respect to classified job types (data-driven) and (b) qualitative ratings from surveys or dynamic input from end-users (subjective input).

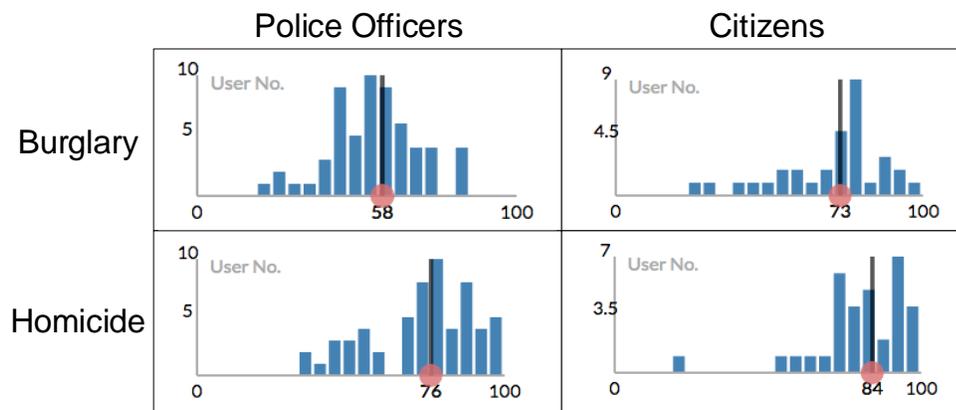


Fig. 3.2. The rating distribution of two sampled severe criminal offenses, burglary and homicide, from police officers and citizens. In a histogram, the x-axis shows the rating scale from zero to one hundred, and the y-axis shows the count of each score. The black lines denote the averages.

Table 3.2.  
Sample survey result for weights of 27 offense categories based on a range from zero to a hundred.

<b>Offense Category</b>	<b>Police</b>	<b>Citizen</b>
Homicide	76.44	83.80
Robbery	66.62	74.43
Drug Abuse	64.24	77.01
Offense Against Family	59.06	69.35
Arson	58.8	66.38
Burglary	58.01	73.17
Operating While Intoxicated (OWI)	56.82	70.22
Assault	56.37	72.59
Fraud	54.48	67.59
Weapon Violation	53.82	67.28
Embezzlement	53.26	66.58
Motor Vehicle (MV) Theft	51.71	64.64
Stolen Property Offense	49.57	61.96
Forgery	49.57	61.96
Larceny	49.43	61.79
Drunkenness	46.8	58.50
All Other Offenses	43.26	54.08
Liquor Law Violation	42.49	53.11
Vandalism/Mischief	40.63	50.79
Runaway	38.73	48.41
Trespass/Threats	35.92	44.90
Disorderly Conduct	35.61	44.51
Suspicious Incident/Person	35.32	44.15

The ratings from citizens are much higher than those of officers for all offense categories. One possible interpretation is that since officers are exposed to a wide range of crimes on a daily basis, they have a less-biased viewpoint, whereas citizens usually only experience crimes from the position of a victim or witness. Both officers

and citizens weighted homicide as the top crime. Fig. 3.2 shows the rating comparison between police officers and citizens for *Burglary* and *Homicide*. The rating from citizens has especially high values for both categories, but the rating from police officers is normally distributed.

To compare the public’s opinion with that of law enforcement towards different types of offenses, the chief from a partner law enforcement agency applied both of these weightings from officers and citizens. He found that the ranking based on total performance score did not diverge significantly from their administrative goals. However, he noted the difference between citizens’ concern towards some types of crimes and the officers’ understanding of these crimes. Although these differences did not affect the overall performance rating, they can be used in community meeting discussions to help align both groups’ priorities.

Rating weights for other organizations can be obtained directly from managers or supervisors. Organizations can also survey employees and service recipients to obtain initial estimates of job category importance. If it is easier for end users to rank the employees, initial weights can also be reverse-calculated using machine learning algorithms. However, weights obtained through such methods could be difficult to explain. As shown in Fig. 3.1, the derived evaluation performance metrics and the job activity records are populated into designated views to show the performance of employees within and across multiple levels.

### 3.3 MetricsVis System

MetricsVis is implemented as a web-based application that utilizes Redux [76] to manage asynchronous calls between the client and server for data consistency, React [77] to support efficient updates of visualizations when data are modified, and D3 [78] to render the customized graphical interface. MetricsVis contains four views: (1) a priority adjustment view displaying the domain-dependent evaluation metrics, (2) the performance matrix showing the details of individuals, (3) the group per-

formance view showing the summarized results of groups as well as an individual's contribution to its group, and (4) the projection view supporting similarity pattern analysis of team members. In this section, each view is described based on its usage purpose and visual representations. To demonstrate the visual representations in an example context, the views are rendered with datasets provided by a law enforcement agency.

### 3.3.1 Priority Adjustment View

The priority adjustment view encodes the evaluation metrics that consider the diversity of evaluation factors in an organization. Its main role is to support the dynamic selection of evaluation factors and the adjustment of associated weights to match organizational priorities (T4). As evaluation metrics in appraisal systems evolve due to rapid changes in service requests, we adopt a priority adjustment view to illustrate the contribution of each evaluation factor by associated weights. After the dynamic modification of evaluation metrics (filtering of evaluation factors or tuning of weights), users can observe the impact on individual and group performance in all other views.

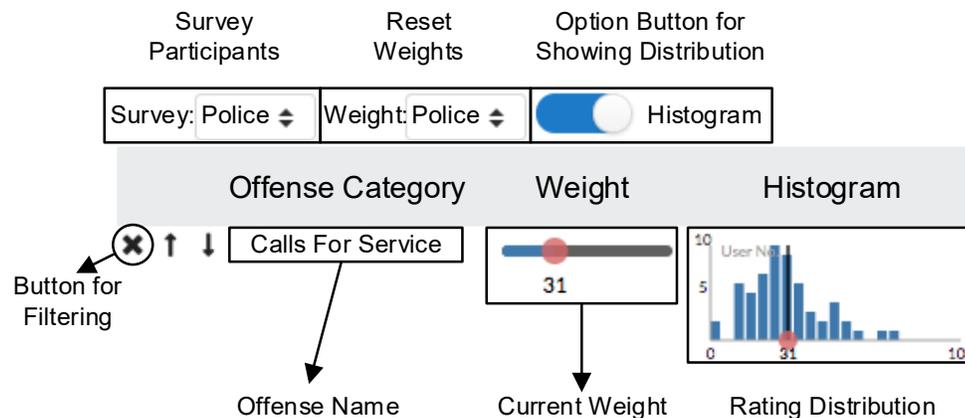


Fig. 3.3. A sample row in priority adjustment view: designed for law enforcement agencies.

In the domain-dependent priority adjustment view (Fig. 3.3), each offense category appears as a row. Each row has a slider bar to modify the current weight and a histogram to illustrate the rating distribution from either police officers or citizens. In the histogram, the x-axis shows the rating weight scale from zero to one hundred, and the y-axis shows how many participants provided each rating score. Placing the rating distribution beside the slider bar provides extra visual cues [79] as to the severity of each offense category. The rating distribution indicates the variation of opinions among survey participants. The initial recommended weights are the average ratings from either police officers or citizens, with the exact value indicated by a red dot along the x-axis. Users can dynamically adjust the weight by dragging the red dot in a slider bar. If the priorities of the organization change, users can appropriately tune the weights of offense categories until the performance scores reflect the change in goals for the department.

### 3.3.2 Performance Matrix View

To efficiently evaluate and compare the performance of employees for the entire organization (T1, T3), our performance matrix (Fig. 3.4) is designed to show the detailed job completion status of all employees in a holistic view. We adopted a color-coded reorderable matrix for this purpose because the matrix (1) occupies a compact screen space to encode all employees, (2) provides a variation of a table in order to maintain visual familiarity, and (3) provides flexible sorting interactions. In the matrix, employees and job types are represented by the columns and rows, respectively. The column heading located at the top with gray bars shows the total performance score of individuals. The row heading located at the left side shows the total score of all performance-related factors. Each cell shows the performance score based on the hybrid evaluation metrics. Each cell's color is defined by the score of a job category accomplished by an employee. When users mouse over a cell, a

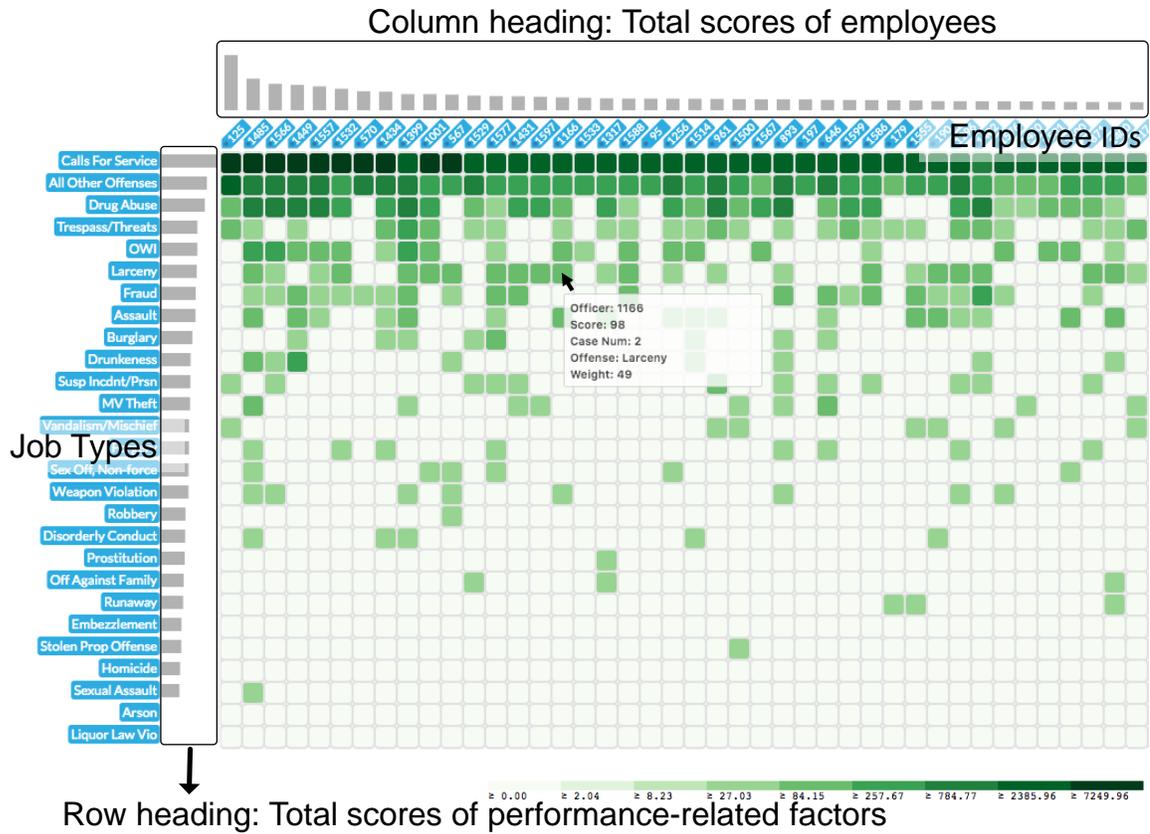


Fig. 3.4. The performance matrix shows the employees (columns) and job types (rows). The matrix is sorted based on the total score of employees and job types. Darker colors encode higher values.

tooltip shows the precise score, number of completed jobs, and weight. Clicking on an employee or job category re-sorts the table with transitional animations.

To satisfy the requirement of comparing multi-dimensional data at the individual level to the entire organization, the mapping of high- and low-level comparison tasks, visual encoding, and sorting interactions are listed in Fig. 3.5. Two sorting interactions are provided in the performance matrix: (1) sort by total score of employees or job categories, and (2) sort by an individual employee or a particular job category.

We use event and incident records from a law enforcement agency as an example; the sorted results are shown in Fig. 3.4. Because some job types have low occurrences

High-level Comparison Tasks	Low-level Comparison Tasks	Visual Encoding	Sorting Interaction
Comparison of aggregated values	Compare data items based on total scores	The length of bars to encode the total scores of data items	Sort the total scores of data items
	Compare total scores of attributes	The length of bars to encode the total scores of attributes	Sort the total scores of attributes
Comparison of multi-dimensional data	Compare data items by single attribute	Color-coded cell to indicate the magnitude of single attribute for a data item	Sort all data items by values of an attribute
	Compare attributes by one data item	Color-coded cell to indicate the magnitude of single attribute for a data item	Sort all attributes by one data item
Similarity pattern analysis	Compare data items in the context of groups	Both bar chart and color-coded cells	Sort the total scores of data items with the constraint of groups (actual groups or clusters)

Fig. 3.5. The mapping of comparison tasks, visual encoding, and sorting interactions for the performance matrix view.

(e.g., arson, liquor law violation), the data in the performance matrix is relatively sparse. In order to minimize the visual impact of this uneven data distribution and increase contrast within the matrix, we applied quantile mapping after a logarithmic transformation of the original scores and used nine sequential green colors recommended by ColorBrewer [80]. Our color mapping method is built on a data binning procedure that first normalizes the data using a power transformation and then applies equal interval binning on the transformed space [81]. We employed green colors because humans can perceive more shades of green than red or blue color tones [82].

As shown in Fig. 3.4, the performance matrix is sorted by the total scores of officers in a descending order so that users can easily observe the officers with top

performance scores. Moreover, users can investigate the top performing officers who are dispatched, those who self-initiate the call response, or a combination of both (T1). With offense categories sorted by a selected officer, users can observe the officers' relative workload across different offense categories and where they focused their self-initiated work. This helps commanders understand the strengths and weaknesses of an officer. The performance matrix includes all members across the organization, which provides comparisons in an organizational context. Users can observe how an officer ranks in the organization. Selection interactions are supported to simplify officer comparison; for instance, users can select any officers that they are interested in, and then those officers will be aligned on the left side of the matrix. With selection operations, commanders can evaluate and compare the officers in their teams and explore the different types of incidents responded to by individuals, their team, and the organization.

Sorting by total score of offense categories demonstrates the overall workload needed to be addressed by an agency (T3). Comparing the total score of offense categories in two time frames, such as between consecutive months, can indicate changes in the prevalence of certain crimes. Ranking officers by a given offense category can directly reveal the most experienced officers for dealing with such incidents. If the police department wants to target a specific crime category, the commanders can determine the officers most suited for the task.

### **3.3.3 Group Performance View**

Most organizations have employees working in teams; as a result, for effective performance evaluation, it is essential to understand the performance within these groups. To support comparisons among groups (T2), our system provides two grouping methods: (1) group by team assignments and (2) group by a clustering algorithm. We implemented three visual representations in our overall group performance view to support this comparison and analysis of team performance: (1) a table list, (2)

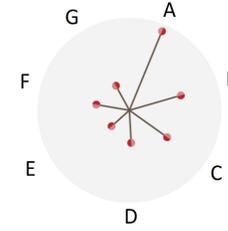
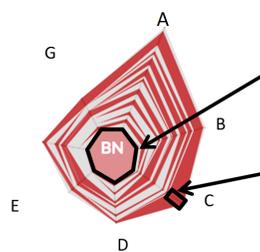
High-level Comparison Tasks	Low-level Comparison Tasks	Recommended Glyphs	Visual Encoding
Group comparison	<ul style="list-style-type: none"> <li>Explore the variation among groups</li> <li>Identify outliers and extreme attribute values</li> </ul>	<p><b>Dandelion Glyph</b></p> 	<ul style="list-style-type: none"> <li>Axes to encode different attributes (categorical data)</li> <li>Length of axes to encode attribute values (numerical data)</li> <li>Encode different groups with color</li> <li>The circle encodes the value and direction along an axis</li> </ul>
Individual to group	<ul style="list-style-type: none"> <li>Identify subordinate data items' contribution to their upper level</li> <li>Correlate attributes to identify performance patterns</li> </ul>	<p><b>Stacked Radar Chart</b></p> 	<ul style="list-style-type: none"> <li>Axes to encode different attributes</li> <li>The center to encode the group ID</li> <li>Each colored ribbon encodes one data item</li> <li>The length of a ribbon on one axis encodes the proportional ratio of one data item to the total of an attribute</li> </ul>
Group to entire organization			

Fig. 3.6. The mapping of comparison tasks, glyphs, and visual encoding for the group performance view.

dandelion glyphs, and (3) stacked radar charts. The group performance view provides an overview of the aggregated multi-dimensional performance data items for groups. For high-level comparison tasks (Fig. 3.6), the group performance view demonstrates (1) performance evaluation and comparison at the group level (within the same level), and (2) each individual data item's performance contribution to its group and performance contribution of a group to the entire organization (across two levels). For low-level comparison tasks, the customized dandelion glyphs provide an efficient simultaneous comparison for a set of data attributes, and identification of outliers and correlation among attributes. The combination of the dandelion glyph and stacked radar chart enables retention of inherent hierarchical relationships among employees and supports high- and low-level comparison tasks. In addition, the dandelion glyph

displays an overview of a group, with the details of individual employees expanded on-demand in the stacked radar chart.

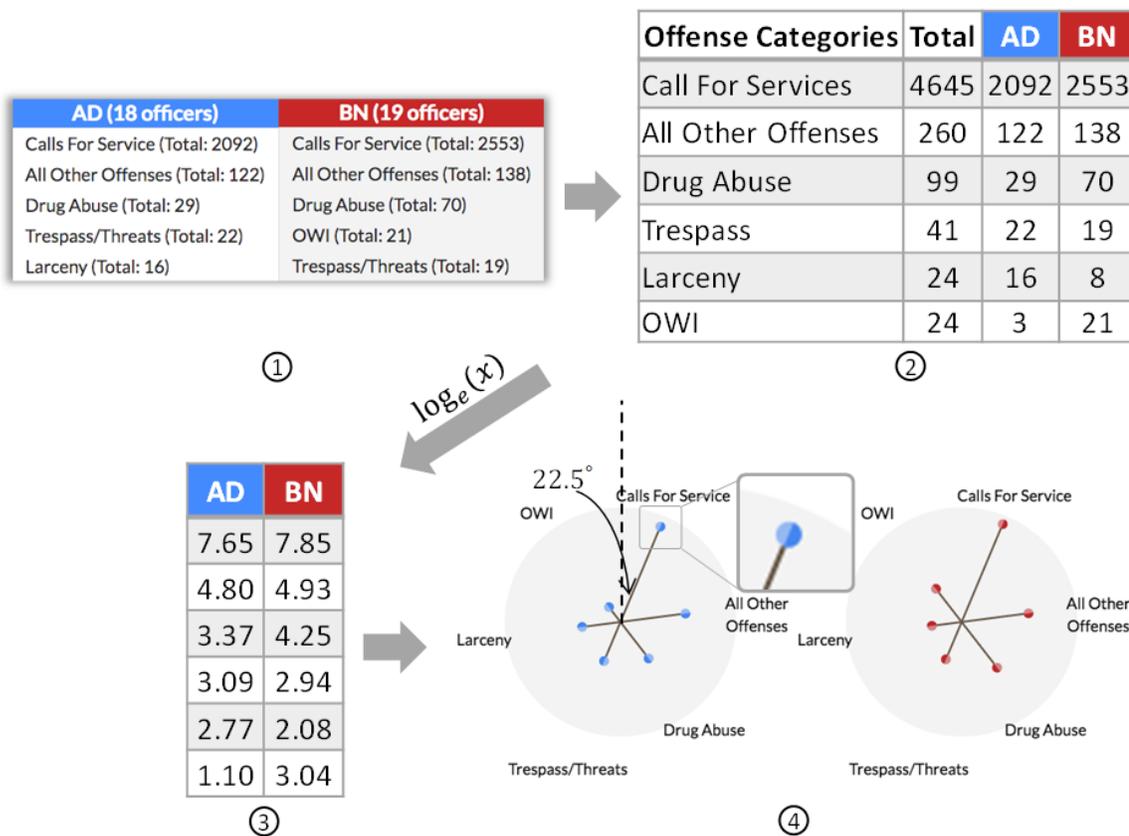


Fig. 3.7. The transformation steps from a table to dandelion glyphs. (1) Get the union of the top five categories in both groups. (2) Order the categories by total in descending order. (3) Apply the logarithmic transformation to the total count. (4) Dandelion glyphs for two groups.

### Table

The table at the top of the group performance view shows the overall performance of each group (Fig. 3.7(1) group by assigned shifts). With a summary of jobs accomplished by employees, the table lists the ranking of job categories based on

their counts, making it intuitive for users to examine the workload of each team. For instance, patrol officers that work in law enforcement agencies need to constantly monitor designated areas to ensure the safety of the community and are usually assigned by shifts and districts. The *A shift* and *B shift* split the days of week (alternating days). Each day is broken down into a day shift and a night shift. Fig. 3.7(1) shows the top five offense categories for *A Day shift* **AD** and *B Night shift* **BN**. For both teams, officers spent the most time on *Calls for Service* events that did not generate criminal case reports or incidents belonging to the *All Other Offenses* category. *Calls for Service* events are not considered a high priority, but generate a large portion of the workload. Our law enforcement partner agency found that this view provided the insight that they needed to break down the *All Other Offenses* category and examine which offenses in this category should receive further examination. With the ordering of offense categories for each group, users can easily determine which tasks utilize the most resources from each team and shift. However, it is not as easy to compare the different groups with only the table listing. Thus, we created a dandelion glyph to enable convenient comparison among such groups.

### Dandelion Glyph

Small multiple glyphs are expressive and use screen space effectively to illustrate large data [84]. Thus, we incorporated characteristics of various small multiple glyph designs (Fig. 3.8) into the design of the dandelion glyph. Inspired by previous research indicating that star plots with radial layout outperform tabular displays for comparing attribute values [43], we also adopted the radial layout into our dandelion glyph. In our dandelion glyph, the axes encode different attributes (categorical data) and the length of the axes encode the attribute values (numerical data). We compare our dandelion glyph with the graphs shown in Fig. 3.8 and Table 3.3. Dandelion glyphs have high data-to-ink ratio and are intuitive, visualizing the differences among groups effectively.

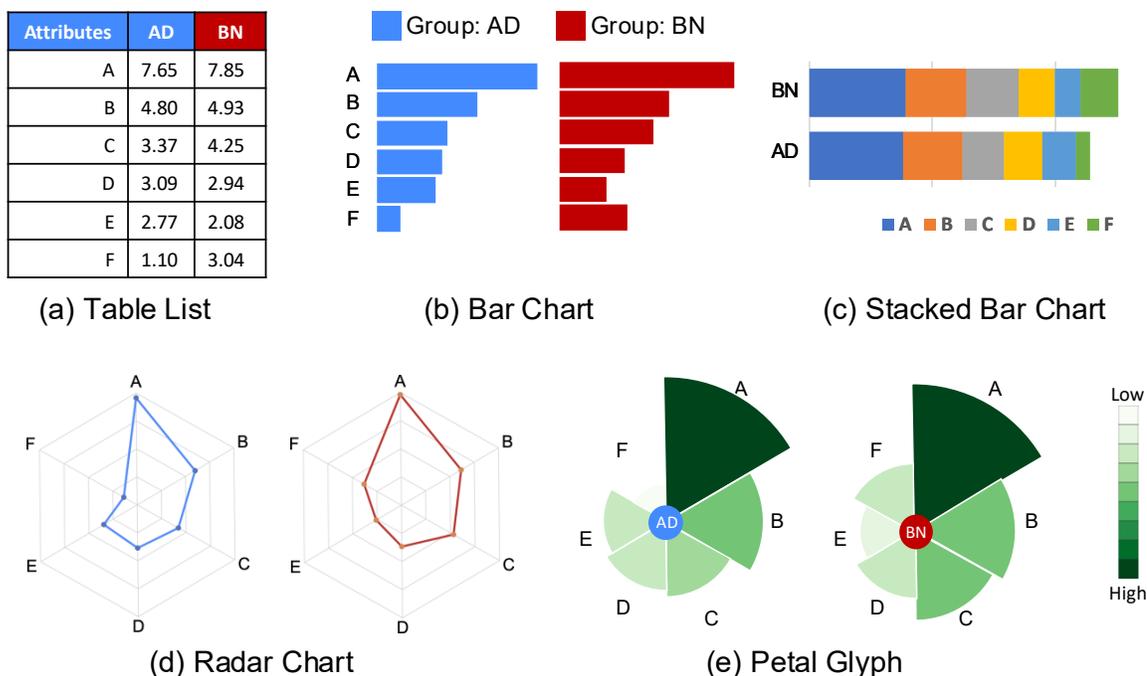


Fig. 3.8. The five examples of small multiple glyph to represent the multi-dimensional data attributes of two groups.

The transformation process from tabular display to dandelion glyphs is shown in Fig. 3.7. In step 2, we took the union of offense categories across all groups, and then ordered the offense categories based on the total count. Finally, the order of axes was determined based on the total count of each attribute. A logarithmic transformation of the total count was applied to the dandelion glyph, since the values of each category axes vary extensively in our dataset. However, datasets with minimal variance (e.g., agencies in which group performance categories contain similar values) might only require linear transformations. The transformation enables users to perceive the contribution of each job category. Notably, the dandelion glyph is a simplified version of star coordinates. Munzner [83] discussed the suitable scenarios of applying radial layout and importantly mentioned the inappropriate representation effect that

Table 3.3.  
Comparison of dandelion glyph versus other glyphs in small multiple settings.

Visualization	Advantages	Disadvantages for Increased Data Size
Dandelion Glyph	<ul style="list-style-type: none"> <li>• Radial layout<sup>1</sup></li> <li>• High data-ink ratio</li> </ul>	<ul style="list-style-type: none"> <li>• Scalability problem<sup>3</sup></li> </ul>
Table List	<ul style="list-style-type: none"> <li>• Precise values</li> <li>• Commonly understood</li> </ul>	<ul style="list-style-type: none"> <li>• Less efficient in comparison tasks</li> <li>• Large pixel size for single data item</li> </ul>
Bar Chart	<ul style="list-style-type: none"> <li>• Rectangular layout<sup>2</sup></li> </ul>	<ul style="list-style-type: none"> <li>• Complex in comparison tasks than radial layout: harder to locate identical attribute</li> </ul>
Stacked Bar Chart	<ul style="list-style-type: none"> <li>• Easy to compare the sum of all attributes</li> </ul>	<ul style="list-style-type: none"> <li>• Hard to compare the bars in the middle</li> </ul>
Radar Chart	<ul style="list-style-type: none"> <li>• Radial layout<sup>1</sup></li> </ul>	<ul style="list-style-type: none"> <li>• Scalability problem<sup>3</sup></li> <li>• The connections at the end of axes are unnecessary</li> <li>• Unequal importance among attributes</li> </ul>
Petal Glyph	<ul style="list-style-type: none"> <li>• Radial layout<sup>1</sup></li> <li>• Double encoding (length and color) for values</li> </ul>	<ul style="list-style-type: none"> <li>• Scalability problem<sup>3</sup></li> <li>• More pixels on the screen are used for each attribute</li> </ul>

<sup>1</sup>Efficient in comparison tasks for large data [43], <sup>2</sup>Simple layout to indicate the data variance,

<sup>3</sup>Only appropriate to show a dozen data attributes or less [83]

symmetric axes have on the same value. To eliminate symmetric impressions in our dandelion glyphs, we rotated the glyph by a small amount ( $\frac{1}{8}\pi$ ).

Although the dandelion glyph is most suitable for displaying up to 10 to 12 attributes, users can interactively adjust the number of top categories in each group. Users also can interactively explore the total count among groups with selection interaction. Comparing two groups' performance in Fig. 3.7(4), the significant difference of *OWI* incidents is readily apparent. To further confirm the exact numerical difference, users can select the *OWI* axis, and the corresponding axes in all dandelion glyphs are highlighted with precise values (Fig. 3.9(a)). The dandelion glyphs represent an overview to avoid initial visual clutter and can be expanded to stacked radar charts to show moderate details of individuals on-demand.

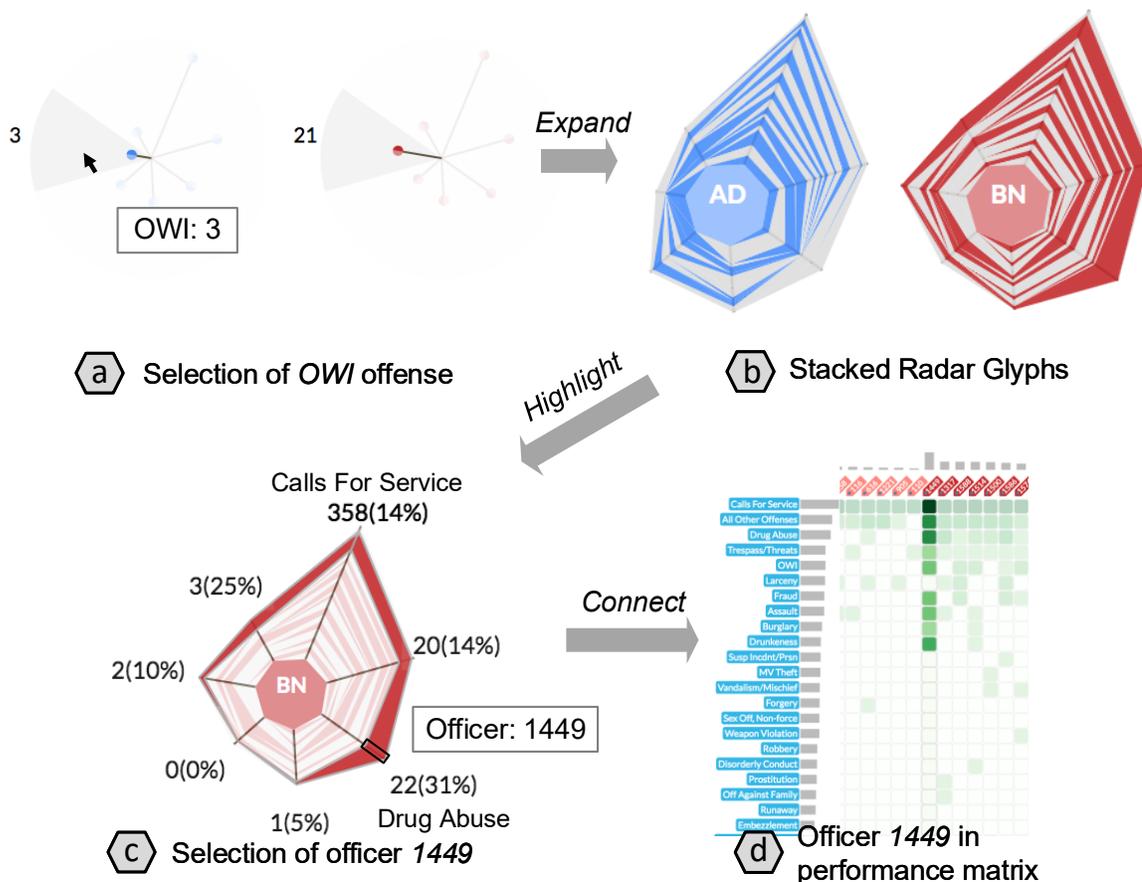


Fig. 3.9. The two radial layout visual representations in the group performance view: dandelion glyphs and stacked radar glyphs. The glyphs show the list of criminal incidents responded to by *A Day shift* and *B Night shift*. (a) Highlight of *OWI* incidents in dandelion glyphs. (b) Stacked radar glyphs show the contribution of each member. (c) Selection of Officer 1449 in *B Night shift*. (d) Highlight of Officer 1449 in performance matrix.

### Stacked Radar Chart

The stacked radar chart is customized to illustrate contributions of subordinate individuals to their upper levels/groups, and it holds the same contour as dandelion glyph to keep familiarity and consistency. It can be applied to show connections between two levels and preserve the information of aggregated upper groups and

show moderate details of subsequent levels in a compact space. An example of a stacked radar chart can be found in Fig. 3.9(b). For a group member, the values of axes are shown as colored ribbons in the radial layout. As shown in Fig. 3.9(c), the selected officer *1449* dealt with 22 *Drug Abuse* incidents, which is around 31.42% ( $\frac{22}{70}$ ) of the entire group. The proportion of pixels along one axis is calculated based on the ratio between the value of a member to the group total. Using the link to the performance matrix, we can observe that, unsurprisingly, officer *1449* had the top performance score in his or her group (Fig. 3.9(d)).

The stacked radar chart allows users to inspect the contribution ratio of each member of a group. Diehl et al. [85] found that using a radial layout to encode data attributes by sectors outperforms Cartesian coordinates (i.e., matrix) when focusing on one dimension. This observation from Diehl et al. was made based on an evenly distributed radial grid layout with a single grid highlighted. In our scenario, colored ribbons are adopted to show the variations across multiple attributes simultaneously. We chose to use this method because it allows users to not only identify which members contribute significantly to a group, but to compare performance pattern with those of other members as well. However, while the stacked radar chart effectively demonstrates individual contributions within a group context, users should use it with caution. The length of axes cannot be compared directly since a logarithmic transformation (a non-linear monotonic function) is applied in the dandelion glyph generation process, yet values within an axis are linearly mapped. As discussed in the previous section, the transformation is necessary due to the skewed nature of the original input dataset. Our approach is a tradeoff between encoding the actual value and providing appropriate visual perception. In conclusion, we believe the advantage of using stacked radar charts outweighs the side effects caused by the transformation. To compensate for the uneven spatial distribution of the radial layout, we add a null inner circle (Fig. 3.9(c)) to reduce the bias introduced in the connection between axes.

Compared with matrix and tabular visualizations, the stacked radar chart is less precise regarding showing exact values. The alternating neighboring colors are used

to separate individuals in a compact screen space; therefore, only a limited number of items can be shown. Filtering interactions (showing only a few of the members) and keyboard selection of a single data item mitigate the scalability issues of the stacked radar chart. In our informal interview with domain experts, they confirmed the advantages of using stacked radar chart as following: easy and quick identification of high-contributing individuals and extreme attribute values.

### 3.3.4 Projection View

The projection view contains a scatterplot showing the projected distance among data items. In this view (Fig. 3.10(b)), each data item is shown as a solid dot with an identifiable label, and its color encodes the group information. For instance, the officers close to each other in the projection view have handled similar types of offenses, and their performance is highly correlated. During shift planning, team commanders can build a new team of employees with similar experiences addressing specific types of crime.

To assist with designing resource allocation strategies that balance workload and the skill set of a group, we applied a K-means clustering method [86]. The scatterplot displays the results of a manifold dimensionality reduction algorithm t-SNE [87], which can reduce the multi-dimensional data into a lower number of dimensions to reveal the relationship among data items. The clustering results are marked in the scatterplot through group colors. Users may adjust the number of clusters to get rid of outliers, since the K-means algorithm is sensitive to noise. For K-means, the input data attributes are the number of cases in offense categories. A normalization of input data (maps the original range of one attribute to the range 0 to 1) is applied to guarantee each data attribute contributes properly to the final clustering results.

### 3.4 Evaluation

To demonstrate how our partner agency utilized MetricsVis in exploring event and incident records as well as evaluating patrol officers' effectiveness, we describe two example use cases.

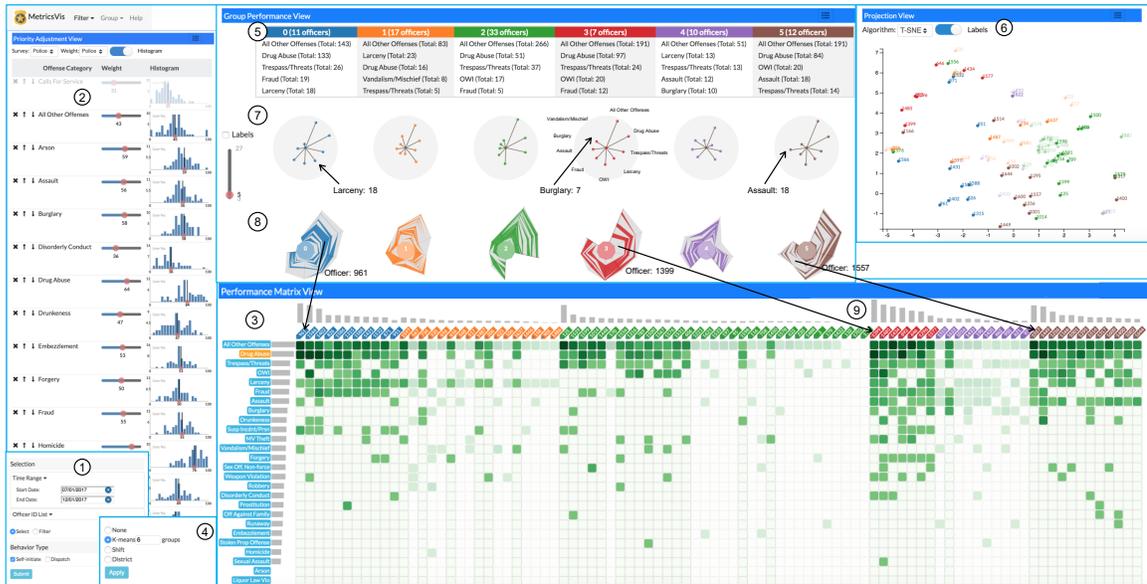


Fig. 3.10. MetricsVis overview: The priority adjustment view (2) encodes the crowdsourced crime severity ratings from police officers and citizens (perceived importance of factors); the red dots indicate the currently assigned weights used in the evaluation metrics. The projection view (6) shows the dimensionality reduction results. The group performance view (5) contains three visual representations that show an overview of group performance and the contribution of each member. The performance matrix view (3) displays the individual employee performance with employees in columns and job types in rows (here, employees are sorted based on their group first and then their total performance scores). The control panel shows the filters (1) and grouping method (4) applied in use case 1.

### 3.4.1 Use Case 1

The chief of a law enforcement agency needs to build provisional specialized anti-drug teams. Before forming the new teams, he wants to know the historical workforce performance of handling drug abuse incidents. He is interested in exploring five months (July 1st to Dec 1st, 2017) of incident records (Fig. 3.10(1)). He selects all officers and filters out the dispatched cases and call for service events (Fig. 3.10(2)), since the majority of drug abuse incidents are self-initiated and result in a criminal report.

He first examines the ranking of officers' total scores in the performance matrix view. He observes that the top 3 officers responded to 88, 74, and 67 total cases, respectively. Next, he examines the most prevalent crimes through sorting by offense categories. He finds that drug abuse is the second most frequent offense category (Fig. 3.10(3)) with an initial weighting of 64 (average rating from police officers). In examining the precise numbers, he notices that the top 3 officers handled 36, 33, and 15 drug abuse incidents, which required 40.90%, 44.59%, and 22.39% of their self-initiated workload. To explore drug abuse cases more closely, he directly sorts the officers by drug abuse offense category and discovers that 52 officers were involved in a total 383 cases (ranging from 1 to 36 by individual officer). With 3 officers handling over 20% of the cases, when creating an anti-drug team, these officers and officers with similar performance across all cases are good potential candidates.

Since offense categories are not independent and drug abuse is highly correlated with 80% of crimes, he explores the activity patterns of the 52 officers more closely using the automatic grouping generated by our clustering algorithms and the visualization results in the group performance view, projection view, and performance matrix view. After a few trials, he finds that K-means clustering with six clusters (Fig. 3.10(4)) provides a good grouping of the results to understand the performance pattern among officers. The majority of the 52 officers are scattered into four clusters, and officers in three clusters responded to the majority of the total number of drug

abuse incidents: the blue cluster 0 of 11 officers handled 133 drug abuse incidents, the red cluster 3 of 7 officers handled 97 drug abuse incidents, and the brown cluster 5 of 12 officers handled 84 drug abuse incidents (Fig. 3.10(5)). The combination of these three clusters of 43 officers responded to 81.98 % of drug abuse incidents. He also inspects the clustering results in the projection view to observe the similarity pattern among clusters, where he finds that green 2 and red 3 clusters are farther apart in the projection space than blue 0 and brown 5 clusters, which can also be observed in the group performance view (Fig. 3.10(6)). He digs into the details among these 3 clusters by first examining the dandelion glyphs (Fig. 3.10(7)). Besides large numbers of overlapping cases (e.g. trespass/threats, operating [a vehicle] while intoxicated (OWI)), he finds that officers in the blue cluster 0 also dealt with many larceny cases, officers in the red cluster 3 dealt with many burglary cases, and officers in the brown cluster 5 handled many assault cases. The stacked radar chart (Fig. 3.10(8)) shows the patterns among the three groups and distinguishable officers in each group. By further examining the officers in the performance matrix (Fig. 3.10(9)), the chief identifies the officers that are most experienced with combinations of different offenses with drug abuse. “This tool provides [commanders] with objective data to assist in resource deployment decision making rather than solely relying on subjective, ‘best guess’, practices that are the norm in law enforcement,” commented the chief.

### 3.4.2 Use Case 2

The department currently evaluates each officer by their supervisors’ scores, which contain subjective metrics that are time-consuming and possibly biased. The chief wants to know if he can utilize data-driven officer metrics in combination with MetricsVis to more effectively and efficiently evaluate performance. He applies the average weighting initially provided by police officers to each incident type. He uses the same time frame as Use Case 1, as well as all call events and crime incidents for both

dispatched and self-initiated activities. He now compares his view and his command staff's view of the top performing officers versus the results shown in our system. Interestingly, the top-ranked officer in the performance matrix view does not match their internal evaluation results. Interactively exploring factors and ideas about what they consider characteristics of the best officers, he decides to consider only criminal incidents that exclude the call events. He finds some officers that match his understanding of good performance get better rankings in the performance matrix under this system. Exploring deeper, he proceeds to filter out dispatched incidents, because he thinks self-initiated incidents are a key component of a top officer. He finally finds that a ranking using only self-initiated incidents matches his command team's understanding of top individual officer performance.

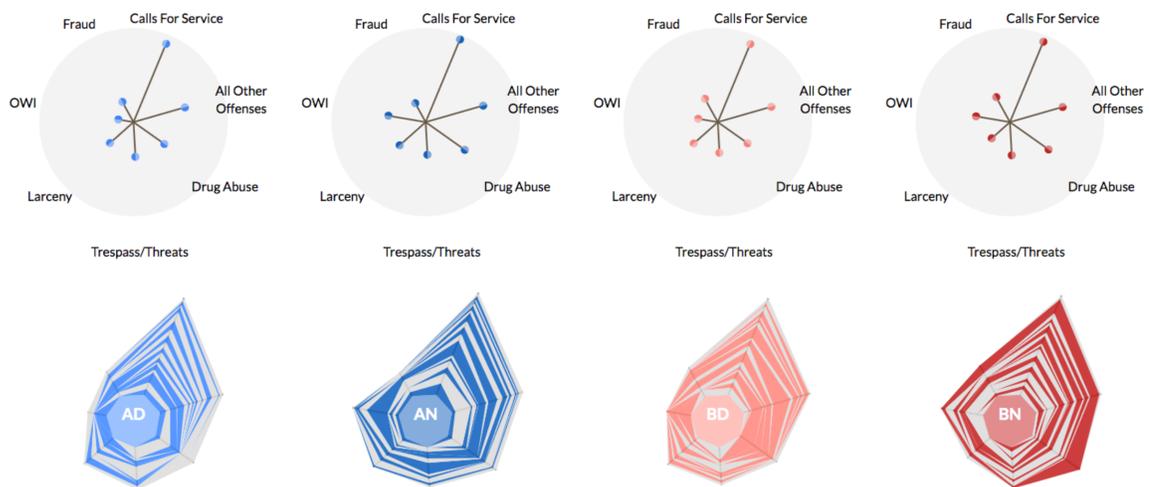


Fig. 3.11. Day ( $AD$ ,  $BD$ ) and night ( $AN$ ,  $BN$ ) shifts have significant differences in drug abuse and OWI incidents for self-initiated incidents.

With the confirmation of the effectiveness of the collected evaluation metrics, the chief is interested in investigating the difference between shifts and districts. He continues with shifts grouped using self-initiated incidents. (As mentioned in Section 3.3.3,  $A$  shift and  $B$  shift are alternating by days, and each day is broken into

a day shift and a night shift. Some patrol officers are not assigned to a specific shift.) It is not surprising that the day shifts exhibit a similar pattern and the night shifts show another trend (Fig. 3.11). Based on the dandelion glyphs, he notices that the significant difference between day shift and night shift is the number of drug abuse and OWI cases. He also wants to compare the dispatched incidents between shifts, and expects the four shifts to have very similar patterns and the workload to be evenly distributed across all shifts for dispatched incidents. He also wonders about the workload across different districts. Even for dispatched incidents, the difference is noticeable. Therefore, these differences can be used to guide effective policing on each shift and district and also must be factored into an officer's performance evaluation, since an officer should not be scored poorly because they are assigned to a low crime time period and area. A lieutenant from highway patrol recognizes this: "MetricsVis would enable [commanders] to look at the total impact of officers and teams and not just sums of cases/incidents. This enables them to assess team and organization level performance in achieving their goals."

### **3.5 Domain Expert Feedback**

We deployed the system to a local police chief, shift commanders, and a crime analyst. The local chief stated that MetricsVis is a valuable visual analytics tool that supports a broad view of the entire organization and provides the possibility to break stereotypes and overcome bias in understanding organizational performance. MetricsVis has also revealed new insights into staff workload and which quantitative metrics (e.g. self-initiated incidents) relate to supervisor's subjective evaluation of top officers. Moreover, the chief noticed the necessity of deconstructing the All Other Offenses category, which contains around 50 % of criminal incidents. Drilling down on this generic offense category can improve the comprehensiveness of evaluation metrics in aligning with organizational objectives.

The command staff have now used the tool during their last four quarterly performance reviews and have provided very positive feedback. They expressed that the tool enables them to ground their evaluations, and quickly and effectively explore understandable quantitative metrics. It also indicates role models and activity types for officers to use as guidelines for improving their performance. Another noted valuable aspect of MetricsVis is its ability to convey the most effective and experienced officers for handling certain incident types. This information is helpful in preparing shifts and training sessions.

A crime analyst who was engaged in the development process of MetricsVis provided valuable interpretation of the data (e.g. night shifts often deal with more self-initiated incidents even though there are fewer calls after midnight, since officers during the day are largely occupied with dispatched cases; day and night shifts usually have very different working patterns), as well as helped validate datasets and define questions of interest. He has identified additional factors that contribute to organizational performance for inclusion for future improvement of MetricsVis (e.g. days worked, arrests, traffic stops).

### 3.6 Discussion

**Tasks, Views, and Interaction Mapping** To accomplish each task, a number of views and several interaction categories (proposed by Yi et al. [88]) are required. Fig. 3.12 outlines the role of the views and interactions needed for each task, and the shaded cells were colored based on the frequency of using views and interaction categories to accomplish tasks during the interactive sessions with domain experts (police and commanders). To efficiently evaluate individual employee performance (T1), the performance matrix view is frequently used to explore the details of all employees in a holistic view. Users can highlight (select) a subset of employees to compare and rank by performance with sorting and reconfiguring interactions. To support comparisons among groups (T2), the group performance view shows the aggregated results of

groups (group level comparison) as well as the contribution of group members (across individual and group level comparison). Abstract/elaborate interaction categories are frequently used to show the overview among groups first and on-demand details of individuals within one group in the group performance view. Select, explore, filter, and connect are the basic interaction categories for linking employees to their groups and identifying prominent patterns (e.g. anomalies with low/high performance). Evaluating organizational workload (T3) and priorities (T4) are more comprehensive tasks that require exploration with all views. Analyzing the workload across the entire organization (T3) involves the summation of all completed jobs during a certain period. The performance matrix aggregates all jobs completed by selected employees, showing the productivity outcome of the entire organization. The options to select, filter, and reconfigure interaction categories provide the flexibility to investigate the

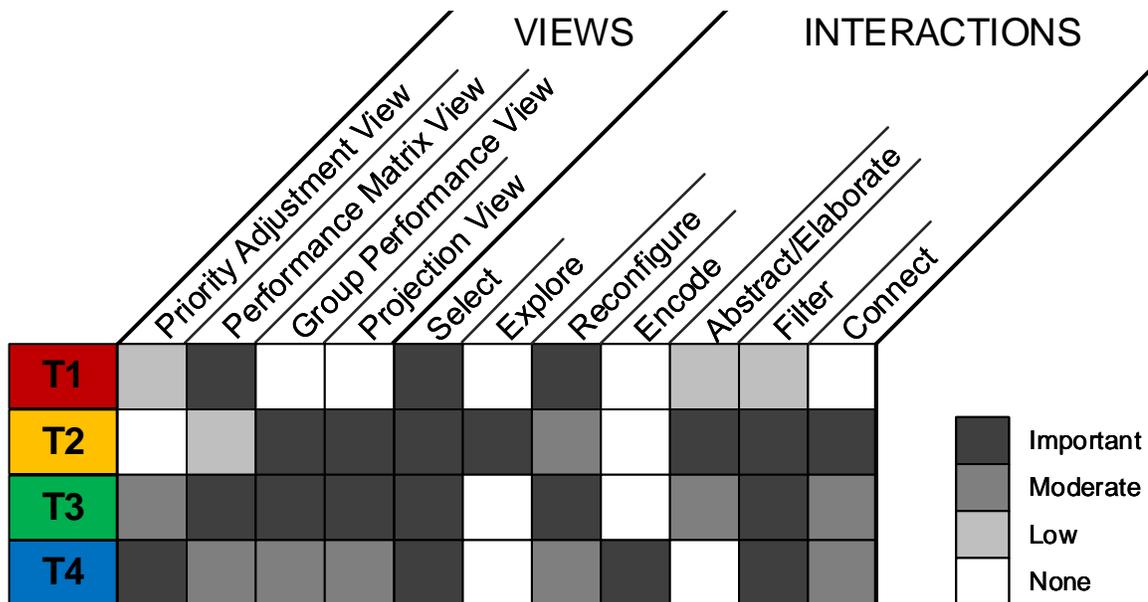


Fig. 3.12. The relationship between analytical tasks (rows) and MetricsVis views and low-level interaction categories [88] (columns). Cell shading quantifies how a particular view or an interaction contributes to the analysis process of a task.

overall performance over different time frames, locations, and alternative team assignment. To verify the alignment of evaluation metrics vs. department priorities (T4), the priority adjustment view is heavily used for the filtering of job types (filter) and tuning of weights, and then the corresponding changes are reflected in all other views (encode).

**Evaluation Metrics** We started by trying to understand the general characteristics of employees, teams, and shifts with different workloads in various organizations while consulting the literature. Our collaboration with domain experts from law enforcement agencies enabled us to better understand the importance of refining the evaluation metrics. We considered using the number of cases responded to per officer to represent the quantitative measurement of productivity. However, the effort required to resolve each case is different. After consultation with police supervisors, we adopted the idea of substituting the effort of handling a case with the severity of the crime. The severity somewhat reflects the relative importance of responding to a case. Based on the initial weights determined through surveys, domain experts can dynamically investigate the overall performance, which is derived using additive weighting. The goal of MetricsVis is not micromanagement (deciding who is the best officer), but a systematic approach to investigating the effectiveness of an organization at and across multiple levels. The optimization of evaluation metrics is an ongoing area of research, but with the assistance of MetricsVis, domain experts can investigate different sets of evaluation metrics to identify the best match with their organizational objectives.

**Visual Designs** For visualization design, we deduced that a tabular visualization summarizing all employees provided better utility than graphs of individual statistics. Besides the individual performance, we noticed the importance of providing an overview on the aggregated group results. After examination of different designs in small multiple settings, we adopted a dandelion glyph, which is a variation of star coordinates. We added the stacked radar chart to bridge the gap between dandelion

glyph (group) and performance matrix (individual). The stacked radar chart shows all members within a group as well as their performance-related factors in limited screen space. Compared with treemaps [89] and node-linked graphs that usually focus on displaying the hierarchical relationships among data items, the stacked radar chart allows simultaneous comparison for multiple attributes using continuous shape instead of separated ones. The radial layout can show only a limited number of visually differentiable categories; however, the number of common job types across different teams is limited, and filtering interactions and selection by keyboard can improve the usability.

**Generalization** We believe that the four visual analytical tasks categories identified in this paper are applicable to other team- or shift-based organizations that use automatic systems to record employee activities, such as delivery drivers, nurses, and emergency medical services. In addition, MetricsVis, although implemented for public safety agencies, was designed with individual and group performance evaluation in mind and, therefore, we expect that the system can be extended to similar type organizations.

**Limitations** There are several limitations in our current system. For instance, officers who work fewer shifts cannot be directly compared with officers working full shifts. Also, the number of hours officers work each shift is not currently logged. The time required to respond to each type of incident needs to be incorporated as a weighting factor when computing metrics of performance. Currently, our system is designed for organizations with only a few hundred employees and dozens of job categories. Scalability of the system for larger organizations may be an issue as the number of dimensions for similarity pattern analysis increases; additional hierarchical modeling and filtering may be a solution for scaling to higher dimensions.

### 3.7 Conclusion and Future Work

We presented MetricsVis, an interactive visual analytics system for organizational performance evaluation. Our system contains four visual components to support interactive visual analysis of organizational performance with a set of hybrid evaluation metrics, integrating subjective ratings and quantifiable outcomes of job activities at multiple grouping granularities. The usability of MetricsVis was demonstrated with two use cases that leverage the designed features and their use for real-world problems: new group staffing and actual group assignments to shifts and districts.

To optimize and improve the evaluation metrics, we plan to incorporate more activity records (e.g. number of arrests, traffic stops). Another possible improvement is to include the time associated with job types as another contributing factor in the final performance outcome, since the time to complete a particular problem is of interest regardless of the domain. Furthermore, the actual performance ratings from supervisors can be used as potential rankings of officers to reverse engineer the evaluation factors/weights to investigate potential biases.

## 4. AUTOMATIC PERFORMANCE WEIGHTS LEARNING DRIVEN BY USER-GUIDED RANKING

In this chapter, we present MetricsVis II, a visual analytics system supporting the learning of weights through user-guided ranking. As previously mentioned in Chapter 3, a set of hybrid evaluation metrics that combine (a) quantitative measurements of employee achievements and (b) qualitative subjective feedback on relative contributions are applied to demonstrate the performance of individuals. A simple additive weighting [90] is applied to the evaluation metrics to derive the overall scores of officers. Specifically, the quantitative measurements of employee workload are used as data attributes, and subjective feedback on the importance of each task is transformed into weights. The quantitative measurements of employee workload are extracted from historical employee activity records (Section 3.2). For instance, the number of responded cases for different offense categories is utilized to produce the data attributes in case studies with law enforcement agencies. What could be an appropriate approach to applying similar evaluation metrics in organizations that do not have an accurate estimation of the weights for different job categories? In particular, a method of collecting satisfying weights that precisely reflect the contribution of different tasks is not adequately addressed in the original MetricsVis system, since the system focuses on performance evaluation across multiple levels according to organizational hierarchy.

Compared to the former weights obtained through a survey of employees and service recipients, we recommend another approach to obtain weights through user-guided ranking. Supervisors can provide the rankings of some employees that they are familiar with and empower learning algorithms to determine the weights, and then predict the performance for the rest. In addition, the actual subjective ratings of employees are provided in the MetricsVis II system as supplementary information

for obtaining potential weights. We are leveraging the human-in-the-loop approach in order to learn the weights interactively. The weights, automatically determined by either the subjective ratings or user-guided rankings, can provide meaningful insights to users in two perspectives: (1) relate their evaluation to the workload of different offense categories, and (2) reveal potential unintended biases in their rankings.

We summarize the contributions of this chapter as follows:

- An interactive performance evaluation system supporting the dynamic learning of weights based on users' adjustments of ranking orders and visual comparisons of ranking results.
- Comparisons of learned weights based on subjective ratings provided by multiple supervisors, and analyze preferences between supervisors.
- A qualitative user evaluation conducted with domain users to collect their feedback on using MetricsVis II to evaluate employee performance as general, and their preferences between the two weighting methods (survey method vs. interactive user-guided ranking).

#### **4.1 MetricsVis II System**

The MetricsVis II system, which extends the original MetricsVis system, encourages the interactive learning of weights by training ranking algorithms dynamically. New visual components are incorporated into the existing matrix view that allows users to modify input ranking orders, investigate derived weights, and analyze predicted overall performance ranking orders. Furthermore, subjective ratings by multiple supervisors are integrated into the system to enable comparisons among different rating groups.

### 4.1.1 Workflow

Compared to the survey method that directly acquires the weights from employees and service recipients as indicators of importance for different job tasks, the MetricsVis II system takes relative rankings of employee performance as input to compute the weights. Users can provide their preferred order of employee ranking by dragging and dropping a few selected employee entries to the proper ranking positions, and can then obtain weights calculated by machine learning algorithms. Based on the derived weights, users can verify the generalizability of the weights through further inspection of the predicted performance of other employees. This interactive refinement of ranking orders can provide an alternative approach to identifying influential factors in performance prediction driven by user preference. In addition, subjective ratings from supervisors are included in the system to support the investigation of uniformity and variety among supervisors. Taking subjective ratings from a supervisor as an example, users can drag and drop employees to positions that comply with the order of subjective ratings from a given supervisor, then obtain learned weights that reflect the preference of that supervisor. Similarly, users can repeat these steps to obtain weights from different supervisors, and further investigate similarity and diversity among supervisor preferences.

### 4.1.2 User Interface

To satisfy the interactive learning of weights and verification of predicted overall performance, we include three additional visual representations (Fig. 4.1) in the performance matrix view (the original performance view is described in Section 3.3.2):

1. The ground truth subjective ratings (i.e., total rating scores) provided by supervisors are added at the top as extra reference for interactive adjustments. Each supervisor only provides ratings for a few employees. Border colors are used to denote different supervisors to distinguish the subjective ratings from calculated total scores. The list of supervisors is found in a dropdown list, and

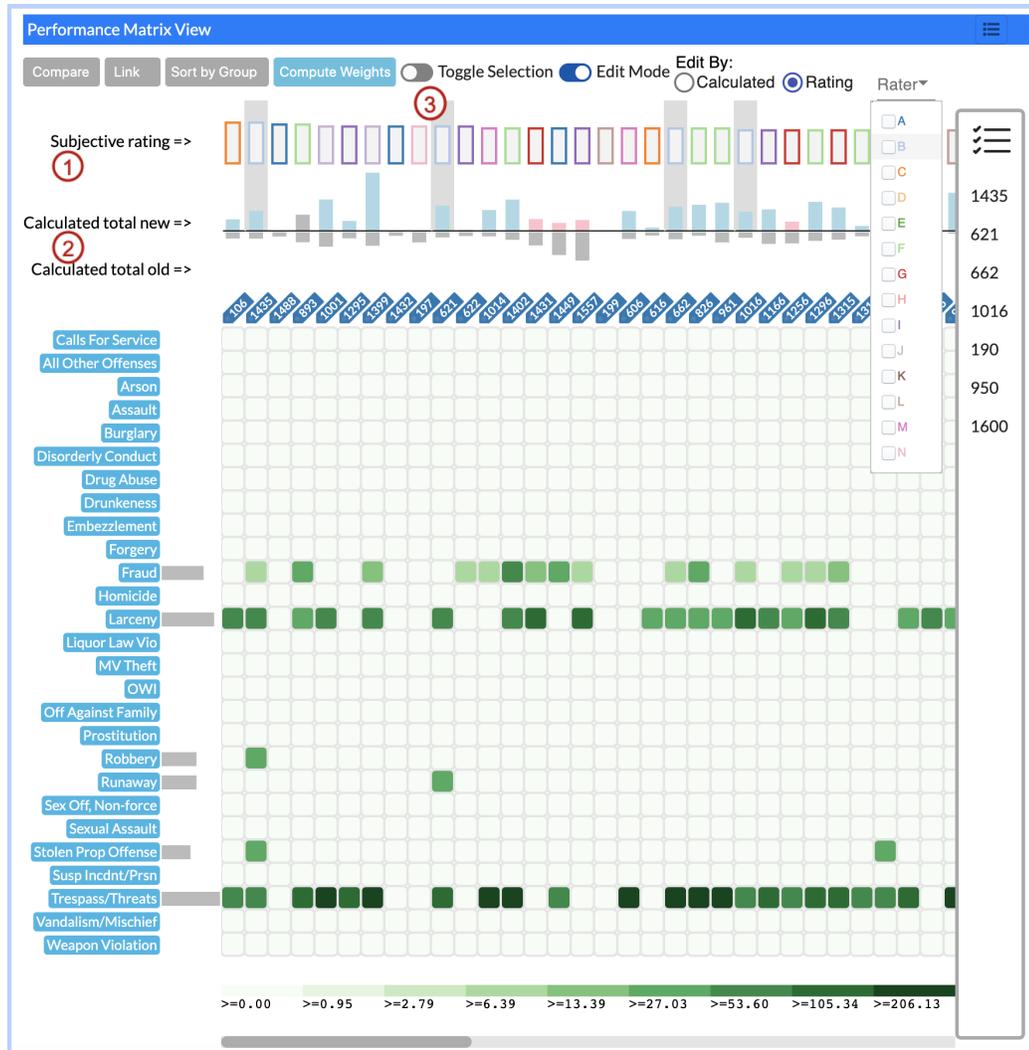


Fig. 4.1. Additional visual representations in the MetricsVis II interface: (1) subjective ratings from multiple supervisors; (2) comparisons of total scores calculated based on updated weights and previous weights; (3) indication of selected individuals that are used in the weights learning.

users can filter out employees by their supervisors. To ensure the confidentiality of supervisors, only the categorical information is kept in the system. We use rater A and rater B in the MetricsVis II system to denote specific supervisors who provide the ratings for a group of employees. Furthermore, users

can compare the diversity of employee performance based on different grouping strategies (e.g., shift, location, or rater).

2. Back-to-back dual rectangles are included to illustrate the total scores derived from updated weights and previous weights for an individual. Therefore, users can observe the changes of rankings and total scores. Since the changing of weights is not limited to a single performance-related factor, the direct comparisons between updated and previous total scores are not reasonable. Additionally, the ranking is more insightful than total scores alone; it indicates the performance of an employee in the context of the entire department. Thus, we use color encoding on the updated total scores to indicate higher (blue) or lower (red) rankings compared with the previous total scores.
3. The selected individuals that are applied as a preferred ranking to obtain the updated weights are highlighted with gray shadows. Users can keep track of selected individuals and then further inspect their positions among the rest.

### 4.1.3 Interactions

**Editing Mode** An *editing mode* is introduced in MetricsVis II to enable the interactive adjustment of ranking while retaining existing sorting interactions. Once the editing mode is turned on, the sorting interactions are suspended to ensure the relative orders among selected individuals are preserved. Users can interactively drag an individual to the left or right to indicate a higher or lower ranking. The same color encoding is applied to illustrate relocating to a higher (blue) or lower (red) ranking.

**Default Ranking** Two type of rankings can be used as the default ranking: (1) ranking by subjective ratings and (2) ranking by latest calculated total scores. Users can choose either one as the default ranking, which can be regarded as a starting point to manipulate the positions of individuals. If users choose ranking by subjective ratings as the initial stage, they can take advantage of predefined ranking. More

importantly, they can examine fewer individuals filtering by their raters. For instance, if a supervisor as a user of the system is only interested in his/her group, the ranking of that selected group can be used to derive a ranking. If none of the raters are selected in the dropdown list, all the data will be displayed. The latest calculated total is another default ranking that can be used to verify whether the input ranking is accurately reflected by the automatically derived weights.

**Compute Weights** When users are satisfied with the ranking of selected individuals, they can click the compute weights button to initiate the learning of weights. Then the updated weights are used to obtain the latest calculated total scores, and the former weights become previous weights. We adopted the pairwise learning to rank approach that was advocated by Podium [40] and Ranking SVM [91] in order to create the list of training data samples. The pairwise approach only considers relative ranking by two data samples; users are more confident providing their relative preferences for two samples rather than an absolute ranking of all data samples [92]. Therefore, we use entirely pairs of selected individuals to generate the training data samples. Pairwise approaches usually outperform standard regression and classification approaches, because pairwise approaches consider the relative relationship between a pair of data samples (i.e., the order of two data samples), which follows the natural practice of ranking [93].

#### 4.1.4 Weights Learning

**Ranking SVM** The Ranking SVM [91] algorithm was initially proposed to obtain weights constrained by pairwise preference in information retrieval applications. Later, Podium [40] uses the algorithm to obtain weights for general ranking applications (e.g., ranking of football teams based on offense and defense statistics, ranking of movies) in an interactive visualization system. Ultimately, the ranking problem is reduced into a two-class classification problem using standard SVM with a par-

ticular transformation on the feature space. Next, we briefly explain how to use performance-related factors to derive the proper input.

We use a row vector  $\mathbf{x}_i = \langle x_{i1}, x_{i2}, \dots, x_{id} \rangle$  to denote one data sample  $i$ , where  $d$  equals number of dimensions. A tuple of  $(\mathbf{x}_i, y_i)$  represents a data sample and its corresponding label, and  $y_i$  equals a ranking score that is higher for a preferred data sample. For instance, the top ranked data sample will have the highest numerical value. For a pair of data samples  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , the difference between the two vectors  $(\mathbf{x}_i - \mathbf{x}_j)$  is the input data attributes for standard SVM, and the label equals  $sign(y_i - y_j)$ . As a result, for any pair of data samples, a tuple in Equation 4.1 represents an input training sample for SVM. Generally  $sign(y_i - y_j) \in \{-1, 0, +1\}$  has three classes, we ignore all the 0s in our application, and excluding  $i = j$ , wherein a pair of data samples have equal importance are incomparable for a ranking algorithm. Therefore, Ranking SVM becomes a two-class classification problem using standard SVM.

$$\left( (\mathbf{x}_i - \mathbf{x}_j), sign(y_i - y_j) \right) \quad (4.1)$$

For any two data samples  $i \neq j$ , if  $sign(y_i - y_j) = 1$ , the input tuple (Equation 4.1) belongs to the +1 class, and vice versa. To balance the number of training samples for each class (+1, -1), the combination of any two data samples are tested before determining the input tuples. The label of a input tuple can be alternated by rearranging the position of  $x_i$  and  $x_j$ , such as  $\left( (\mathbf{x}_j - \mathbf{x}_i), sign(y_j - y_i) \right)$ .

**Non-negative Weights** The hybrid evaluation metrics adopted in the MetricsVis system integrate both quantitative measurements of observed achievements externalized using performance-related factors, and subjective feedback on the relative importance of each factor (more details can be found in Section 3.2). Specifically, each weight is in a range between 0 and 100. To maintain consistency and familiarity, the same performance-related factors are applied in the MetricsVis II system, and the range of each weight remains unchanged. The non-negative weights are required due to two reasons: (1) each data attribute (i.e., performance-related factor) is initially designed to positively contribute to the prediction target (i.e., the overall

performance of employees); (2) negative weights can mean inverse relationship between an attribute and the prediction target, which means less service is better. We can argue that negative weights are meaningful to some extent, since they show less preferred job categories. However, it conflicts with the initial design of the performance evaluation metrics. More precisely, we use  $\mathbf{w} = \langle w_1, w_2, \dots, w_d \rangle$  to denote the row vector of weights, where  $w_i \in [0, 100]$ . To comply with the non-negative weights constraint, we used the VarSVM<sup>1</sup> as the two-class linear classifier to obtain the non-negative weights. VarSVM is a variation of standard SVM, and it sets extra constraints on non-negative weights in each iterative coordinate descent step using dual form of linear SVM and Lagrangian multiplier.

## 4.2 Weights Comparisons by Subjective Ratings

The weights obtained by applying a ranking algorithm for the subjective ratings from all supervisors are shown in Fig. 4.2. Each supervisor provides subjective ratings for a group of officers, and these officers are usually working on the same shift. For instance, supervisor B works closely with 7 officers, and their subjective ratings are in a range between 14 and 17 out of 20 as maximum. The 7 officers can be ranked by their subjective ratings. Here, we use only the default ranking without drag and drop interactions, and the underlying data applied in the analysis is all self-initiated incidents for the second half year in 2017 (July 1st to Dec 31st, 2017). The time frame explicitly matches with the semi-annual supervisor ratings time window. Then we click the compute weights button to obtain automatically calculated weights by the ranking algorithm. Only 6 non-negative weights are returned, 5 of which are relatively high (*Trespass/Threats*: 56, *Larceny*: 37, *Stolen Property Offense*: 42, *Robbery*: 42, *Runaway*: 42) and 1 of which has a lower value (*Fraud*: 11). After referring back to the raw incidents, the officers who have high supervisor ratings dealt with more incidents in these 6 offense categories. The returned weights can almost reflect the

<sup>1</sup><https://github.com/statmlben/Variant-SVM>

Rater	Survey	J	B	M	A	E	C	G	F	I
Number of Employees	59	7	7	9	10	8	8	8	6	9
Burglary	58	91	0	0	0	18	0	0	38	0
Trespass/Threats	36	40	56	62	0	0	27	12	27	42
Larceny	49	0	37	0	0	0	0	48	0	0
Stolen property Offense	50	0	42	0	30	0	0	0	0	0
Robbery	67	0	42	0	0	0	0	0	0	0
Runaway	39	0	42	0	0	0	0	0	0	0
Sexual Assault	65	0	0	0	0	0	0	0	0	0
Offense Against Family	59	0	0	0	0	0	0	0	0	0
Liquor Law Violation	43	0	0	0	0	0	0	0	0	0
Arson	59	0	0	0	0	0	0	0	0	0
Homicide	76	0	0	0	0	0	0	0	0	0
Prostitution	45	0	0	0	0	0	0	27	0	0
Suspicious Incident/Person	35	0	0	0	0	0	0	25	17	0
All Other Offenses	43	0	0	0	0	10	0	8	0	32
Weapon Violation	54	0	0	0	0	14	0	0	0	24
Calls For Service	31	0	0	0	14	22	0	15	4	0
Disorderly Conduct	36	0	0	0	0	23	0	0	0	0
Embezzlement	53	0	0	0	0	26	0	0	0	0
Sex Off, Non-force	55	0	0	69	0	26	0	18	0	0
Vandalism/Mischief	41	0	0	35	30	0	0	0	15	0
Drug Abuse	64	10	0	0	20	0	68	20	41	55
Forgery	50	0	0	0	0	0	68	19	0	0
Assault	56	0	0	0	4	54	0	42	0	0
Drunkenness	47	0	0	0	22	57	0	0	51	0
Motor Vehicle Theft	52	0	0	14	41	26	0	25	38	0
Fraud	55	0	11	0	22	0	0	48	39	57
OWI	57	0	0	0	70	11	0	10	0	20

Fig. 4.2. Colored cells displaying the derived weights based on rankings provided by multiple supervisors. The column header includes the rater information, and the row header shows the offense categories. The first row under the column header displays how many employees are evaluated by one rater. The first column besides the row header contains the average weights obtained by surveying police officers. The orders of both raters and offense categories are determined by a hierarchical clustering algorithm.

preference of supervisors who often work on a day shift. Comparing among all the supervisors, supervisor J has a strong preference for officers who dealt with a high volume of *Burglary* incidents; and supervisors G, F, and I have similar preferences, which may be due to all of them working on night shifts.

### 4.3 Qualitative User Evaluation

We conducted online pairwise analytics [94] evaluations with administrative personnel in a partner law enforcement agency to collect their feedback from management

perspectives. Though the participants were quite familiar with the system, we still demonstrated the four views and then specifically spent more time on explaining the new automatic weights learning functionalities in the performance matrix view. The instruction session took about 15 minutes, and users spent another 20 minutes exploring the performance of individuals and groups. During the exploratory analysis, participants could consult the VA expert with any questions about the system, and the VA expert might ask questions about insights discovered by domain experts. After the analysis session, a short interview was conducted concerning three topics: (1) how does MetricsVis II fit the general objectives of their organizations, (2) which views are more relevance to their daily work, and (3) which weighing method is preferred (survey vs. user-guided ranking).

We collected feedback from three participants: a police chief and two commanders. Participants provided their opinions on the aforementioned topics. In describing the uses of MetricsVis II to address the objectives of their organization, they listed several objectives, including achieving systematic and objective performance evaluations, identifying proactive and efficient officers, performing timely evaluations, and identifying low performers and improving their contributions. They all agreed that the usage of case numbers is sufficient as a first step to improve the fairness of evaluations, rather than relying on pure subjective ratings. Regarding the usage frequency of different views, they all stated that the performance matrix is most frequently used in their evaluations.

One of the valuable features of MetricsVis II highlighted by all participants is the new capability to comprehensively analyze the objective data while incorporating subjective feedback with a holistic understanding of the entire department. After comparing officer workloads and their subjective ratings, it is quite common for the volume of responses to not correlate well with the subjective ratings. MetricsVis II is appreciated for its first attempt to bring the quantitative measurement of officers' workload and supervisors' subjective ratings together and relate them. For the comparisons of two weighting methods, one participant prefers the crowdsourced survey,

and the other two prefer the user-guided ranking. The advocates of the user-guided ranking method believe that the weights derived from user-guided rankings are helpful in indicating the preferences of different supervisors. In addition, the different working patterns on day and night shifts also largely impact the weights derived from different supervisors, as shown by comparing groups based on supervisors and shifts in Section 4.2. The participant who prefers the crowdsourced survey also confirmed the benefits of having user-guided rankings, which provide novel perspectives based on user or supervisor preference; however, the direct manipulation is also very important to their work routine. He expressed the need to apply standard evaluation metrics across the entire department, and stated that all supervisors should agree upon the evaluation metrics. In addition, participants mentioned the outdated subjective ratings; one clear next step based on this complaint would be to incorporate digitized subjective ratings into the real-time system. Other advantages have already been discussed in the domain expert feedback section in Chapter 3 (Section 3.5), such as reducing subjective bias based on objective data, evaluating performance at multiple scales, understanding performance across teams, assigning officers to particular incidents, identifying officers who need additional training, etc. After deploying the MetricsVis II system, domain experts can further explore the relationship between subjective ratings and objective data. They can even dig into other ‘soft criteria’ such as quality of written reports, call handling, interview skills, effective of arrests, physical abilities, safety tactics, and job knowledge, which have not been digitally recorded or measured.

#### 4.4 Discussion

The requirement of non-negative weights lowers the performance of the applied ranking algorithm, since the false interpretation of the workload is more misleading. For example, *Call for Service* events are usually recorded in non-emergency circumstances. If the weight of *Call for Service* is negative, then someone working

on a larger number of *Call for Service* events might be penalized even when handling more cases. All performance-related factors in our case are designed based on a positive relationship with the overall performance. Mathematically, collinearity among the input features are the main causes for negative weights. If two features are highly correlated, then a slight increase on one feature will cause another feature to be negative in a linear machine learning model. As we discussed in FeatureExplorer (Chapter 5), that collinearity can be mitigated by feature selection. However, applying feature selection does not significantly improve the performance. There are two main reasons for this, relate to the intrinsic characteristics of the dataset we used to represent officer performance: the sparsity of data and the skewed distributions of offense categories. Some offense categories consume more than half the workload of officers; therefore, these categories need to be broken down.

Nonetheless, VarSVM performs accurately on non-negative weights. Compared with standard SVM, VarSVM does not shift the signs of negative weights but pushes them to be zero in each coordinate descent step, and it can highlight the preferred features by user favored rankings.

The user-guided ranking method requires more training than direct manipulation on weights, but the domain users can drag and drop data samples and interpret the derived weights after few trails. Thus, users can capture the contradictory results between objective data and subjective ratings and find out possible explanations through further digging into soft factors such as the quality of written reports or evaluation comments from supervisors. In addition, it is beneficial to provides weights comparisons between different supervisors or users.

#### 4.5 Conclusion and Future Work

The MetricsVis II system extends the original system by including subjective ratings from supervisors and automatic weights learning through user-guided ranking. The manipulation of the ranking of preferred employees provides new insights

through derived weights, where higher weights for a performance-related factor means preferred employees contribute to the factor. The adoption of subjective ratings from supervisors allows domain experts to compare the different preference from different supervisors. Based on qualitative evaluation with domain users, the disparity between objective data and subjective ratings encourages the organization managers to take advantage of our system to gradually mitigate subjective bias and understand where the bias originated.

To improve the predictability and interpretability of learned weights, additional data pre-processing methods besides normalization should be incorporated to reduce the collinearity among features. The fairness of the weights should be further studied across more domain users and different law enforcement agencies. The MetricsVis II system can be deployed to other shift- or team- based organizations to assist organizational performance evaluation.

## 5. INTERACTIVE FEATURE SELECTION AND REGRESSION MODEL EVALUATION FOR HYPERSPECTRAL IMAGES

This chapter is based on the paper published in 2019 IEEE Visualization Conference: J. Zhao, M. Karimzadeh, A. Masjedi, T. Wang, X. Zhang, M. M. Crawford, and D. S. Ebert, “FeatureExplorer: Interactive feature selection and exploration of regression models for hyperspectral images,” in *2019 IEEE Visualization Conference (VIS)*, Oct 2019, pp. 161–165. doi: 10.1109/VISUAL.2019.8933619

In this chapter, we present FeatureExplorer, a visual analytics system to support interactive feature selection and model evaluation for remotely-sensed data. To design this system, we collaborated with remote sensing experts and plant scientists whose goal was to predict plants’ wet biomass using data recorded in hyperspectral imagery. These domain experts needed to identify the predictive ability and interchangeability of key features derived from hyperspectral images (and their underlying wavelengths) for biomass prediction. It was challenging to investigate such high-dimensional datasets and regression models without visual analytics tools, which motivated the design of FeatureExplorer. It enables experts to trace the regression models back to the key contributing features (hyperspectral indices), and ultimately the pertinent image wavelengths (among a large number of bands), along with options for interactive manipulation, feature selection, and model evaluation based on domain knowledge.

Our system supports integrated visual exploration and selection of features through the analysis of: (1) linear relationships among features using a correlation matrix; (2) distribution of any pair of two features using a scatterplot enhanced with Kernel Den-

sity Estimation (KDE) visualizations; (3) feature importance ranking for non-linear relationships based on a combination of a feature selection method (Recursive Feature Elimination (RFE)) and a regression model (Support Vector Regression (SVR)).

We summarize the contributions of this chapter as follows:

- An interactive system supporting dynamic feature exploration and selection based on univariate and multivariate feature analysis with integrated regression models, reducing the large number of features to a few key ones that can be used for improved modeling and future data collection and analysis.
- Experimental results comparing various machine learning methods for predicting biomass using hyperspectral indices.
- A workflow for identifying key hyperspectral indices and the original reflectance values used in index calculations.
- A case study of the use of the platform by domain experts for hyperspectral image analysis to predict plant wet biomass.

In the remainder of this chapter, we first describe the background information related with features and the prediction target in remote sensing and plant breeding domains. Then, we present the design goals identified collaboratively with remote sensing experts. This is followed by a detailed description of FeatureExplorer system and a case study to demonstrate its potential usage. Lastly, we summarize the work and discuss possible future directions.

## 5.1 Background

Biomass is an important plant characteristic that helps with crop monitoring, yield estimation, and indicating plant growing conditions, and is quantified based on the above-ground weight of a plant before and after dehydration (i.e., wet biomass and dry biomass). In the case of sorghum (the crop in our study), biomass determines

the amount of ethanol product. To identify superior plant varieties for breeding and determine the development of plants, biomass can be manually measured during a growing season; however, this traditional method is time consuming, expensive, and retrospective. Instead, hyperspectral images collected by Unmanned Aerial Vehicles (UAVs) throughout the season can potentially be used to predict the final biomass. Remote sensing experts in our team collected high-resolution hyperspectral images multiple times (from June to Sept.) over 14 acres of experimental sorghum fields with 830 varieties in the 2017 growing season. The ground truth wet biomass applied as the prediction target was measured at the end of the growing season (Oct. 15<sup>th</sup>).

A hyperspectral image used in prediction is obtained by a camera that covers the visible near-infrared (VNIR) range, which ranging from 400 nm to 1000 nm in 2.2 nm increments for each pixel (272 bands). The recorded spectra can be applied to distinguish different materials (e.g., plants, soils, water) due to the uniqueness of acquired signals. However, the hyperspectral signatures emulate contiguous narrow bands, which are highly correlated with neighboring ones. To mitigate dependency among original bands and reduce dimensionality, we adopted hyperspectral indices based on domain practice. Specifically, we utilize the 36 hyperspectral vegetation indices listed in [95]. Each index is typically derived from two or three band values and based on a unique plant biophysical meaning such as leaf chlorophyll and nutrition content [96], photosynthesis status [97]. As biomass is a more comprehensive indicator of plant growth, we investigate the joint prediction effect of all 36 indices. However, some indices are closely-related and can provide redundant information in prediction. More information about the sensors, data pre-processing, and feature extraction is available in [98–100].

## 5.2 Design Goals

We collaborated with three remote sensing experts: two Ph.D. students and a senior faculty member with expertise in hyperspectral image analysis for agronomy.

Traditionally, they predict biomass using feature reduction techniques (including feature selection and feature extraction) and regression models. Often, optimally tuning these algorithms requires large numbers of data samples, which are expensive to collect. It is challenging to build a model that performs well for all kinds of hybrid varieties, plants in different locations, or at different growing stages/conditions with limited samples. Therefore, domain experts needed to identify the key hyperspectral features to achieve stable, credible, and accurate prediction results, using both automated methods and their domain knowledge to inspect the relationships between features and the feature importance, and trace the hyperspectral indices back to the biophysical space. Hyperspectral indices indicate meaningful chemical concentrations in plants, which can be applied to differentiate plant varieties. The domain experts also expressed the need for clustering features, dynamic feature selection, and model performance comparisons with and without feature selection. We derived the following design goals to fulfill these requirements:

- DG1** Interactive exploration of features, including feature density distributions and relationships among multivariate features.
- DG2** Identification of important features such as influential hyperspectral indices and the underlying wavelengths that contribute to the prediction of wet biomass.
- DG3** Direct manipulation and refinement on subsets of features through interactively adding and removing specific features.
- DG4** Evaluation of regression results with ground truth for subset of selected features versus the full set of features.

These requirements were formalized into design mock-ups using visualizations already familiar to domain users based on their requests. We then implemented the design, and made minor modifications according to feedback from domain experts, as described below.

### 5.3 FeatureExplorer

In this section, we first explain how our system addresses the design goals, and then elaborate on the frontend user interface and backend analytics components of FeatureExplorer.

#### 5.3.1 Workflow

Figure 5.1 presents the system components in FeatureExplorer, and our process. As shown in Figure 5.1, FeatureExplorer supports the analysis of both linear and non-linear relationships (DG1, DG2). To visualize feature relationships, a correlation matrix serves as an overview to render the Pearson’s correlation coefficient for all pairs of features. Users can click on any cell for a detailed inspection of any particular pair of features. For non-linear relationship analysis, Support Vector Regression and Recursive Feature Elimination (SVR + RFE) provide feature importance rankings. Users can compare and analyze the ranking results and use the synthesized information to add or remove features (DG3).  $R^2$  and Root Mean Square Error (RMSE) are calculated to show the regression models’ performance with the selected subset of features (DG4). After initial implementation, users requested the capability to adjust the number of folds in cross validation, to compare the performance of regression models with a selected subset of features versus with all features, and to map

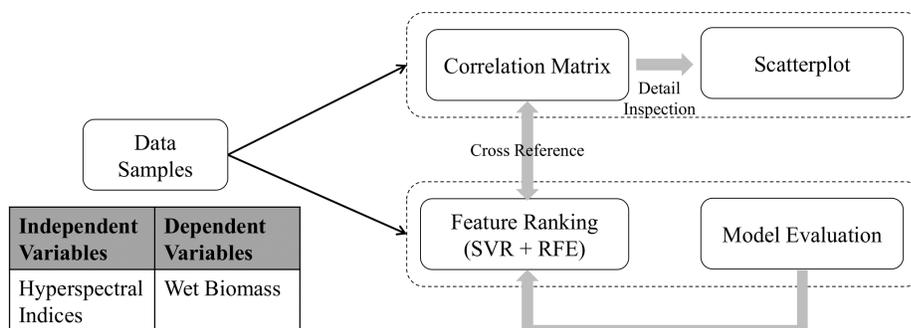


Fig. 5.1. The components diagram of FeatureExplorer.

hyperspectral indices to original wavelengths. This way, users can utilize the gained insights from the interactive exploration process to identify the underlying pertinent wavelengths, and strategize ways to collect only pertinent data in the future to save cost and time.

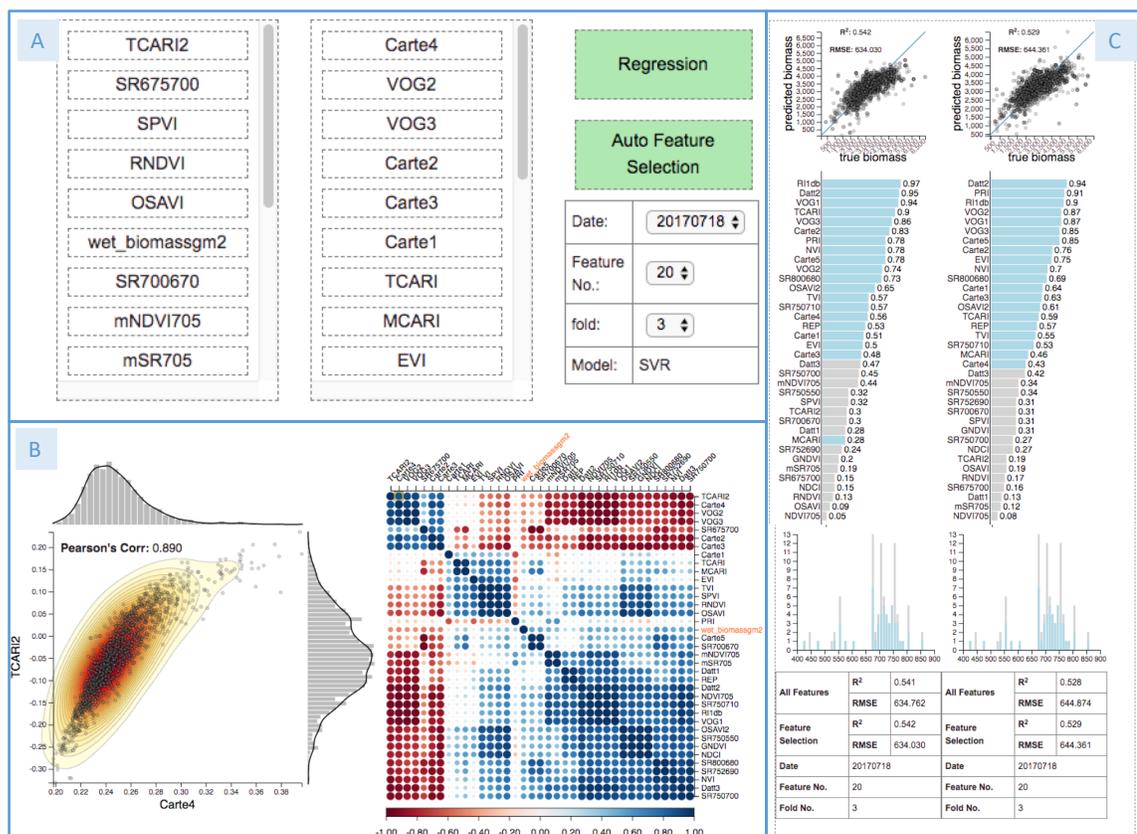


Fig. 5.2. FeatureExplorer overview: (A) the control panel with a list of unselected features, a list of selected features, a regression button, an automatic feature selection button; (B) feature correlation panel with a correlation matrix and a scatterplot; (C) evaluation panel with a scatterplot of ground truth and predicted values, a horizontal bar chart showing the importance score of each feature, a histogram showing the frequency of used pertinent wavelengths, a table displaying the results with and without feature selection.

### 5.3.2 User Interface

Figure 5.2 illustrates the user interface that contains three panels: (A) a control panel, (B) a correlation panel, and (C) an evaluation panel. As we described in the previous section, the two latter panels are separated based on the linearity of the relationship between input features and predicted variables. In this section, we describe the views individually, and will showcase the integrated use of these views in a use case in Section 5.4.

In the correlation panel, a correlation matrix shows the Pearson’s correlation coefficient between any pair of features. The coefficient value is double-encoded using two visual channels (color and radius) for better usability. Hierarchical clustering groups features based on the similarity of correlations to other features. This helps users identify representative pairs from each cluster, while minimizing the chances of including other similarly correlated pairs. While providing a good overview, a single correlation value does not provide sufficient information for interpreting the relationship between two features. To address this, users can click on any cell to see the scatterplot of the selected two features. The system uses both histograms and KDE to illustrate the marginal distribution of univariate features at the edge of the histogram. We also overlaid a 2D KDE on the scatterplot to better visualize the distribution of two features. The marginal distributions and KDE contours are beneficial in understanding general data patterns. The domain users pointed out that exploring the hyperspectral index vs. wet biomass scatterplot could help them investigate whether the index captures the variation across high and low biomass values.

At the top part of the evaluation panel, a scatterplot shows ground truth values against predicted results along with  $R^2$  and RMSE values. With this graph, domain users identified that the regression model does not perform well on extremely high or low biomass values. The horizontal bar graphs show the feature importance score for each input feature (using SVR + RFE), and the light blue rectangles indicate

selected features. The histogram beside the bar graphs shows the frequency of using pertinent reflectance (raw data) to derive the indices in the subset of selected features over the wavelength range of 400 nm to 900 nm. This enables domain experts to trace back the selected features to the wavelengths that are utilized to derive the indices. Moreover, a table shows performance comparisons for a subset of selected features versus all features based on the same data partition (training vs. testing) and regression model.

As we mentioned before, the correlation matrix and the SVR + RFE bar graphs provide different rankings, the former for linear relationships and the latter for non-linear models. Users can refer to both to adjust the subset of selected features. In the control panel, the leftmost list shows unused features, and the list in the middle shows selected ones. Users can drag and drop features between these two lists and evaluate the results on the fly. To avoid exhaustive feature searching at the beginning by users, the system enables an initial automatic feature selection method based on SVR + RFE.

Table 5.1.  
Comparison of average  $R^2$  for 100 trials among multiple regression models on 10 datasets.

Date	Ridge	Elastic Net	Partial Least Square	SVR	Random Forest	AdaBoost
06/21/2017	<b>0.20</b>	0.13	<b>0.20</b>	<b>0.20</b>	<b>0.20</b>	0.15
06/27/2017	<b>0.25</b>	0.16	<b>0.25</b>	0.24	0.23	0.18
07/04/2017	<b>0.27</b>	0.17	<b>0.27</b>	<b>0.27</b>	0.26	0.19
07/18/2017	0.51	0.23	0.51	<b>0.53</b>	0.44	0.36
07/30/2017	0.51	0.28	0.52	<b>0.55</b>	0.49	0.45
08/08/2017	0.53	0.34	0.53	<b>0.56</b>	0.50	0.45
08/14/2017	0.53	0.35	0.53	<b>0.54</b>	0.51	0.45
08/23/2017	<b>0.54</b>	0.34	<b>0.54</b>	<b>0.54</b>	<b>0.54</b>	0.50
09/10/2017	<b>0.52</b>	0.32	<b>0.52</b>	<b>0.52</b>	<b>0.52</b>	0.47
09/24/2017	0.51	0.35	0.51	<b>0.52</b>	0.51	0.45

### 5.3.3 Regression Models

After testing several regression models including Ridge, Elastic Net, Partial Least Squares, SVR, Random Forest, and AdaBoost, we found that SVR [101] outperforms other models for predicting biomass from hyperspectral indices for most dates. The results of  $R^2$  for these regression models are listed in Table 5.1. Since  $R^2$  and RMSE are highly correlated (higher  $R^2$  means lower RMSE), we only report the  $R^2$ . Based on the results, we decided to integrate SVR + RFE (for automatic feature selection) into the system.

The system runs k-fold cross validation for model evaluation. For each training of the SVR model, the system first runs a grid search with a Radial Basis Function (RBF) [102] kernel to select the best model hyperparameters that maximize  $R^2$ , and then performs initial feature selection on that model [103]. The RFE ranks the features based on their contributions in the regression model, and the system transforms these ranks to scores in the range of [0, 1], 0 meaning no contribution and 1 meaning the most important feature in the model.

We use Equation 5.1 to compute the ranking score of a feature, where k is the number of folds, d is the number of dimensions in the feature space, and r denotes the ranking determined by RFE. The RFE method outputs the feature ranking in sequential order from the most important to least; the most important feature has a ranking of 1 and the least important feature has a ranking of d. The numerator of Equation 5.1 sums the normalized ranking (mapping values in [1, d] to [0, 1]), which is then divided by k to calculate the average of these scores for multiple runs (in k-fold cross validation). We use this RankingScore in feature importance visualization (the horizontal bar graphs).

$$RankingScore = \frac{\sum_{i=1}^k \frac{(d+1-r_i)-1}{d-1}}{k} \quad (5.1)$$

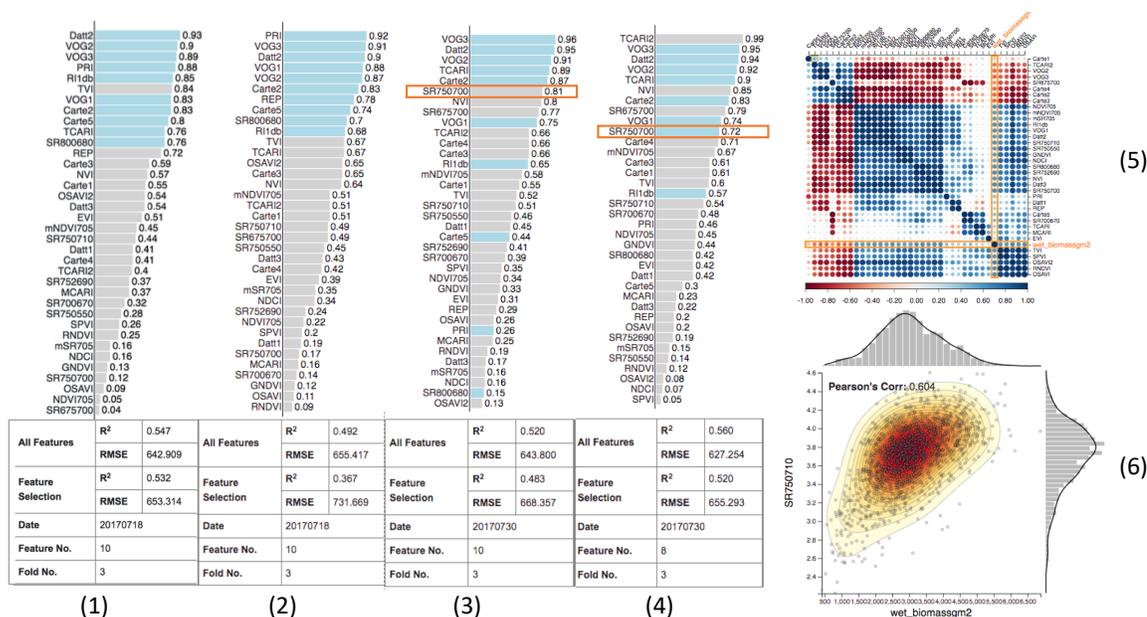


Fig. 5.3. Case study using FeatureExplorer for two hyperspectral datasets.

## 5.4 Case Study

A remote sensing expert in our team used FeatureExplorer to investigate hyperspectral indices for biomass prediction. He aimed to determine which indices were the most predictive ones, and if he could reduce a combination of 36 features down to 10 key features while understanding their biophysical meanings in collaboration with a plant scientist. He used 10 hyperspectral images collected from June 21st to Sept. 24th in 2017 to investigate whether the important subset of hyperspectral indices changed in each image set. First, he started with one dataset (July 18th) and applied automatic feature selection for 20 features (out of 36 total), and found that the performance using 20 features was slightly better than when using all 36 features. Then, he applied automatic feature selection, limiting to 3 features. The regression performance ( $R^2$ ) dropped significantly (higher RMSE). Based on ranked feature sets and the correlation matrix, he added 4 features that had high importance scores and low correlation among them. These 4 features were selected from different clusters

in the correlation matrix, since he wanted the regression model to learn useful information from diverse features. The performance of the model improved. After adding up to 10 features, the performance of the regression model was almost equivalent to its performance when using 20 features (Figure 5.3(1)). He then tested whether applying automatic selection limited to 10 features would lead to similar results; it turned out that the manually selected features outperformed the automatic selection (Figure 5.3(2)).

Next, he applied the same subset of features on another hyperspectral image (July 30th) that was captured 12 days after the first one. He found that wet biomass had stronger correlations with most hyperspectral indices (the correlation matrix shown in Figure 5.3(5)) compared with the first dataset (the correlation matrix shown in Figure 5.2(B)). The regression model performed better on the second dataset than the first one because the plants were at a different growing stage [104] and their reflectance had changed [105]. Tuning the regression model on the second dataset with the 10 features selected during analyzing the first dataset did not improve the prediction results; however, the performance of the regression model did not drop dramatically (Figure 5.3(3)). By carefully examining the correlation matrix for the second dataset, he found 3 features that did not have high correlations with biomass. After removing these 3 features and adding another feature which had a high importance score and high correlation with biomass, the model's performance improved significantly (Figure 5.3(4)). This indicates the human-in-the-loop can improve the predictive performance of the regression model.

## 5.5 Conclusion and Future Work

We presented a visual analytics system for the exploration, ranking, and selection of features in integrated regression models supporting analysis on linear and non-linear relationships. The system provides initial automated feature selection, and enables users to dynamically change, compare and evaluate models' performance based on

user-specified subsets of features. We demonstrated the successful use of the system by remote sensing experts to identify important hyperspectral indices at various plant growth stages for predicting the biomass at the end of the growing season, as well as tracing these indices back to the underlying wavelengths for each growing stage. This enables more targeted data collection and analysis in the future. FeatureExplorer can also be applied to other sensor data (e.g., multispectral, LiDAR) that possess similar properties to hyperspectral indices (e.g. high dimensions, derived correlated features), to predict variables other than biomass. Our system can also be adjusted to include different regression models, since the underlying model will not intrinsically impact the feature exploration workflow.

Future visual analytics research should investigate the dynamic generation of features based on raw input data, e.g. customized features based on different formulations of hyperspectral indices. Also, one can improve the feature selection workflow by visually highlighting potential features in clusters that are ranked high importance (or low), for faster subgroup inclusion/exclusion. Feature selection in regression models for spatially and temporally heterogeneous data is also an open area for research. Specifically, the geovisualization of feature importance for spatial regression methods has not been adequately addressed. Finally, time series analysis can be incorporated to model temporally variable feature contributions, e.g. in a sequence of hyperspectral images with temporally variable wavelength reflectances at different plant growing stages.

## 6. CONCLUSIONS AND FUTURE WORK

In this dissertation, we have presented three visual analytics systems to assist users with evaluating performance in two application scenarios: (1) organizational employee performance evaluation, and (2) improving the performance of machine learning models through interactive feature selection. All of these approaches have been developed by working closely with domain users and are designed to integrate domain users' work processes into these systems. In particular, MetricsVis has been deployed to a partner agency to assist with quarterly performance reviews. Furthermore, FeatureExplorer integrates feature selection methods, which is a substantial step forward in remote sensing experts' feature mining pipeline and provides assistance with the assessment of feature predictability and model performance. MetricsVis and its extension MetricsVis II are novel VA approaches with customized visual representations that demonstrate employee performance at multiple scales, enabling dynamic, effective, and comprehensive performance evaluations. FeatureExplorer and MetricsVis II both leverage the human-in-the-loop approach to improve either the performance of machine learning models or predictions of employee performance. We restate our main contributions in the following two perspectives:

- **Identify influential factors:** In FeatureExplorer, we applied two feature selection methods in two stages of building regression models: (1) before building a model, analyze the relationship between features using a correlation matrix; (2) after building a model, obtain the relative importance of features by applying recursive feature elimination on regression models. The hierarchical clustered correlation matrix can expedite the identification of (1) influential features in different clusters, and (2) interchangeable features within one cluster. The importance scores provided by RFE + SVR are crucial indicators of feature

predictability. With visual representations to effectively demonstrate feature importance, users can iteratively add or remove features (i.e. alter the feature space) by incorporating their domain knowledge in order to achieve higher model performance.

In MetricsVis II, users can manipulate the ranking of data samples (i.e., ground truth from training data) to derive the weights of different performance-related factors. These weights can reflect users' preferences for certain features. Both applications adopt human-in-the-loop approaches in VA systems, incorporating user feedback in identifying influential factors.

- **Multi-level comparisons:** In MetricsVis, employee performance is evaluated at individual, group, and organizational levels. Specifically, the groups can be assigned by (1) shifts, (2) locations, (3) supervisors, and (4) clustering algorithms. Customized visual representations are designed to show employee performance at multiple levels: (1) a reorderable matrix showing individual performance, (2) a dandelion glyph showing aggregated group performance, (3) a stacked radar chart showing an individual's contribution to a group. The comparisons across multiple levels can help create new group assignments and investigation of the total impact of individual, teams as well as the organization. Unlike MetricsVis, where multiple levels are reflected by the relationship between data samples, FeatureExplorer illustrates features at multiple levels. The hyperspectral indices (i.e., input features) are hierarchically clustered in a correlation matrix, and the detailed distributions of any two features can be inspected in a scatterplot. In addition, users can trace the hyperspectral indices back to their original wavelength. Based on a clustered matrix, users can rapidly narrow down the number of features into a key subset. Further details about each feature can help users understand the data and dig into potential reasons to select or remove a feature.

In summary, a collection of VA approaches is presented to assist with informed decision making in organizational performance evaluation and to improve the performance of machine learning models by feature selection. In our future work, we intend to extend these systems by including additional data sources and improving the interpretability of applied machine learning models. We discuss future directions in the context of two performance analysis scenarios:

- **Organizational employee performance evaluation:** An additive weighting method is applied to derive the overall performance of employees based on hybrid evaluation metrics. A simple extension that would include additional impact factors (e.g., the time spent on each case, the subjective feedback on the quality of service, the quality of written reports) is to deploy a linear hierarchical model [34] to accommodate more complex descriptions of work quality. Though non-negative weights are required in MetricsVis II, it is still worthwhile to explore other popular pairwise ranking algorithms (e.g., RankNet [106], IR SVM [107], Lambda Rank [108], LambdaMART [109]) and additional data pre-processing techniques that reduce collinearity between features to improve the quality of prediction models. Based on the domain user evaluation, the next step is to deploy the MetricsVis II system with our partner agency. Domain users can further compare subjective ratings with objective workload measurements in order to reveal the subjective preferences of evaluators and discover potential biases.
- **Interactive feature selection and model evaluation:** A straightforward extension is to include other types of remote sensing data (e.g., LiDAR data) and additional ground truth data that depicts other appearance characteristics of plants (e.g., height, canopy cover). Our domain users also suggested the inclusion of additional regression models (e.g., partial least squares regression, random forest regression) to provide additional flexibility for the machine learning models without extensive updates to the system, since recursive fea-

ture elimination can work well with any models. For faster subgroup inclusion/exclusion, the system should incorporate the optimal cluster number and automatically highlight potential features with high importance scores in each cluster. Another appropriate extension is the inclusion of time-series analysis across multiple datasets, since the remote sensing data are collected regularly across the entire growing season. It is necessary for domain users to distinguish features that are performing well and poorly at different stages of plant growth.

## REFERENCES

## REFERENCES

- [1] K. A. Cook and J. J. Thomas, “Illuminating the path: The research and development agenda for visual analytics,” Pacific Northwest National Lab.(PNNL), Richland, WA (United States), Tech. Rep., May 2005.
- [2] G. P. Latham and K. N. Wexley, *Increasing Productivity through Performance Appraisal*. Reading, MA: Addison-Wesley, 1981.
- [3] J. W. Smither, “Lessons learned: Research implications for performance appraisal and management,” in *J. W. Smither (Ed.), Performance Appraisal: State of the Art in Practice*. San Francisco, CA: Jossey-Bass, 1998.
- [4] H. J. Bernardin and R. W. Beatty, *Performance Appraisal: Assessing Human Behavior at Work*. Boston, MA: Kent Publishing Company, 1984.
- [5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, Aug 2018. doi: 10.1145/3236009
- [6] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, 2016, p. 1135–1144. doi: 10.1145/2939672.2939778
- [7] D. A. Keim, J. Kohlhammer, F. Mansmann, T. May, and F. Wanner, *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Association, 2010, ch. 2.2, pp. 10–11.
- [8] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews, “The human is the loop: New directions for visual analytics,” *Journal of Intelligent Information Systems*, vol. 43, no. 3, pp. 411–435, 2014. doi: 10.1007/s10844-014-0304-9
- [9] L. S. Snyder, Y. Lin, M. Karimzadeh, D. Goldwasser, and D. S. Ebert, “Interactive learning for identifying relevant tweets to support real-time situational awareness,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 558–568, Jan 2020. doi: 10.1109/TVCG.2019.2934614
- [10] S. K. Badam, J. Zhao, S. Sen, N. Elmqvist, and D. Ebert, “TimeFork: Interactive prediction of time series,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16, 2016, p. 5409–5420. doi: 10.1145/2858036.2858150
- [11] A. Endert, P. Fiaux, and C. North, “Semantic interaction for visual text analytics,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’12, 2012, p. 473–482. doi: 10.1145/2207676.2207741

- [12] T. O. Kvalseth, “Cautionary note about  $R^2$ ,” *American Statistician*, vol. 39, no. 4, pp. 279–285, 1985. doi: 10.2307/2683704
- [13] X. Jia, B.-C. Kuo, and M. M. Crawford, “Feature mining for hyperspectral image classification,” *Proceedings of the IEEE*, vol. 101, no. 3, pp. 676–697, March 2013. doi: 10.1109/JPROC.2012.2229082
- [14] J. Zhao, M. Karimzadeh, L. S. Snyder, C. Surakitbanharn, Z. C. Qian, and D. S. Ebert, “MetricsVis: A visual analytics system for evaluating employee performance in public safety agencies,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1193–1203, Jan 2020. doi: 10.1109/TVCG.2019.2934603
- [15] J. Zhao, A. Malik, H. Xu, G. Wang, J. Zhang, C. Surakitbanharn, and D. S. Ebert, “MetricsVis: A visual analytics framework for performance evaluation of law enforcement officers,” in *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, April 2017, pp. 1–7. doi: 10.1109/THS.2017.7943468
- [16] J. Zhao, M. Karimzadeh, A. Masjedi, T. Wang, X. Zhang, M. M. Crawford, and D. S. Ebert, “FeatureExplorer: Interactive feature selection and exploration of regression models for hyperspectral images,” in *2019 IEEE Visualization Conference (VIS)*, Oct 2019, pp. 161–165. doi: 10.1109/VISUAL.2019.8933619
- [17] I. Guerra-López, *Evaluating Impact: Evaluation and Continual Improvement for Performance Improvement Practitioners*. Amherst, MA: Human Resource Development, 2007.
- [18] I. Guerra-López, *Performance Evaluation: Proven Approaches for Improving Program and Organizational Performance*. San Francisco, CA: Jossey-Bass, 2008.
- [19] L. M. Coutts and F. W. Schneider, “Police officer performance appraisal systems: How good are they?” *Policing: An International Journal of Police Strategies & Management*, vol. 27, no. 1, pp. 67–81, 2004. doi: 10.1108/13639510410519921
- [20] R. M. Kanter and D. Brinkerhoff, “Organizational performance: Recent developments in measurement,” *Annual Review of Sociology*, vol. 7, pp. 321–349, 1981. doi: 10.1146/annurev.so.07.080181.001541
- [21] B. Becker and B. Gerhart, “The impact of human resource management on organizational performance: Progress and prospects,” *The Academy of Management Journal*, vol. 39, no. 4, pp. 779–801, 1996. doi: 10.2307/256712
- [22] W. C. Borman and D. H. Brush, “More progress toward a taxonomy of managerial performance requirements,” *Human Performance*, vol. 6, no. 1, pp. 1–21, 1993. doi: 10.1207/s15327043hup0601\_1
- [23] R. D. Arvey and K. R. Murphy, “Performance evaluation in work settings,” *Annual Review of Psychology*, vol. 49, no. 1, pp. 141–168, 1998. doi: 10.1146/annurev.psych.49.1.141

- [24] L. Koopmans, C. M. Bernaards, V. H. Hildebrandt, W. B. Schaufeli, H. C. de Vet, and A. J. van der Beek, “Conceptual frameworks of individual work performance: A systematic review,” *Journal of Occupational and Environmental Medicine*, vol. 53, no. 8, pp. 856–866, Aug 2011. doi: 10.1097/JOM.0b013e318226a763
- [25] W. F. Cascio and H. Aguinis, *Applied Psychology in Human Resource Management*, 7th ed. Upper Saddle River, NJ: Prentice Hall, 2011.
- [26] J. P. Campbell, R. A. McCloy, S. H. Oppler, and C. E. Sager, “A theory of performance,” in *Personnel Selection in Organizations*, N. Schmitt and W. C. Borman, Eds. San Francisco: Jossey-Bass, 1993, pp. 35–70.
- [27] W. C. Borman and S. M. Motowidlo, “Expanding the criterion domain to include elements of contextual performance,” in *Personnel Selection in Organizations*, N. Schmitt and W. C. Borman, Eds. San Francisco: Jossey-Bass, 1993, pp. 35–70.
- [28] P. R. Sackett and C. J. Devore, “Counterproductive behaviors at work,” in *Handbook of Industrial, Work & Organizational Psychology - Volume 1: Personnel Psychology*, N. Schmitt and W. C. Borman, Eds. London: Sage Publications Ltd, 2001, pp. 145–164. doi: 10.4135/9781848608320.n9
- [29] J. M. Ivancevich and M. T. Matteson, *Organizational Behavior and Management*, 5th ed. Irwin/McGraw-Hill, Boston, MA, 1999.
- [30] J. T. Delaney and M. A. Huselid, “The impact of human resource management practices on perceptions of organizational performance,” *Academy of Management journal*, vol. 39, no. 4, pp. 949–969, 1996. doi: 10.5465/256718
- [31] P. McLachlan, T. Munzner, E. Koutsofios, and S. North, “LiveRAC: Interactive visual exploration of system management time-series data,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08, 2008, pp. 1483–1492. doi: 10.1145/1357054.1357286
- [32] M. Behrisch, B. Bach, N. Henry Riche, T. Schreck, and J.-D. Fekete, “Matrix reordering methods for table and network visualization,” *Computer Graphics Forum*, vol. 35, no. 3, pp. 693–716, 2016. doi: 10.1111/cgf.12935
- [33] R. Rao and S. K. Card, “The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '94, 1994, pp. 318–322. doi: 10.1145/191666.191776
- [34] G. Carenini and J. Loyd, “ValueCharts: Analyzing linear models expressing preferences and evaluations,” in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '04, 2004, pp. 150–157. doi: 10.1145/989863.989885
- [35] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “LineUp: Visual analysis of multi-attribute rankings,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2277–2286, Dec 2013. doi: 10.1109/TVCG.2013.173

- [36] R. Kosara, F. Bendix, and H. Hauser, “Parallel Sets: Interactive exploration and visual analysis of categorical data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 558–568, July 2006. doi: 10.1109/TVCG.2006.76
- [37] J. Xia, Y. Hou, Y. V. Chen, Z. C. Qian, D. S. Ebert, and W. Chen, “Visualizing rank time series of wikipedia top-viewed pages,” *IEEE Computer Graphics and Applications*, vol. 37, no. 2, pp. 42–53, March 2017. doi: 10.1109/MCG.2017.21
- [38] I. Hur and J. S. Yi, “SimulSort: Multivariate data exploration through an enhanced sorting technique,” in *Human-Computer Interaction. Novel Interaction Methods and Techniques*, ser. HCI 2009, J. A. Jacko, Ed., vol. 5611, 2009, pp. 684–693. doi: 10.1007/978-3-642-02577-8\_75
- [39] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale, “TimeSpan: Using visualization to explore temporal multi-dimensional data of stroke patients,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 409–418, Jan 2016. doi: 10.1109/TVCG.2015.2467325
- [40] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert, “Podium: Ranking data using mixed-initiative visual analytics,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 288–297, Jan 2018. doi: 10.1109/TVCG.2017.2745078
- [41] S. H. Zanakis, A. Solomon, N. Wishart, and S. Dublisch, “Multi-attribute decision making: A simulation comparison of select methods,” *European Journal of Operational Research*, vol. 107, no. 3, pp. 507 – 529, 1998. doi: 10.1016/S0377-2217(97)00147-1
- [42] X. Zhao, Y. Wu, W. Cui, X. Du, Y. Chen, Y. Wang, D. L. Lee, and H. Qu, “SkyLens: Visual analysis of skyline on multi-dimensional data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 246–255, Jan 2018. doi: 10.1109/TVCG.2017.2744738
- [43] M. Keck, D. Kammer, T. Gründer, T. Thom, M. Kleinsteuber, A. Maasch, and R. Groh, “Towards glyph-based visualizations for big data clustering,” in *Proceedings of the 10th International Symposium on Visual Information Communication and Interaction*, ser. VINCI ’17, 2017, pp. 129–136. doi: 10.1145/3105971.3105979
- [44] “Tableau,” <https://tableau.com/>.
- [45] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan 2014. doi: 10.1016/j.compeleceng.2013.11.024
- [46] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994. doi: 10.1109/72.298224
- [47] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994. doi: 10.1109/72.298224

- [48] P. Pudil, J. Novovičová, and J. Kittler, “Floating search methods in feature selection,” *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119 – 1125, 1994. doi: 10.1016/0167-8655(94)90127-9
- [49] J. Reunanen, “Overfitting in making comparisons between variable selection methods,” *Journal of Machine Learning Research*, vol. 3, pp. 1371–1382, March 2003.
- [50] O. Soufan, D. Kleftogiannis, P. Kalnis, and V. B. Bajic, “DWFS: a wrapper feature selection tool based on a parallel genetic algorithm,” *PLOS ONE*, vol. 10, no. 2, Feb 2015. doi: 10.1371/journal.pone.0117988
- [51] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, “Multiple SVM-RFE for gene selection in cancer classification with expression data,” *IEEE Transactions on NanoBioscience*, vol. 4, no. 3, pp. 228–234, Sep 2005. doi: 10.1109/TNB.2005.853657
- [52] J. Ding, J. Shi, and F.-X. Wu, “SVM-RFE based feature selection for tandem mass spectrum quality assessment,” *International Journal of Data Mining and Bioinformatics*, vol. 5, no. 1, pp. 73–88, Feb 2011. doi: 10.1504/IJDMB.2011.038578
- [53] M. Friendly, “Corrgrams: Exploratory displays for correlation matrices,” *The American Statistician*, vol. 56, no. 4, pp. 316–324, Jan 2002. doi: 10.1198/000313002533
- [54] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner, “Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets,” in *IEEE Symposium on Information Visualization*, ser. InfoVis ’03, Oct 2003, pp. 105–112. doi: 10.1109/INFVIS.2003.1249015
- [55] J. Seo and B. Shneiderman, “A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections,” in *IEEE Symposium on Information Visualization*, ser. InfoVis ’04, Oct 2004, pp. 65–72. doi: 10.1109/INFVIS.2004.3
- [56] H. Piringer, W. Berger, and H. Hauser, “Quantifying and comparing features in high-dimensional datasets,” in *12th International Conference Information Visualisation (InfoVis ’08)*, July 2008, pp. 240–245. doi: 10.1109/IV.2008.17
- [57] S. Johansson and J. Johansson, “Interactive dimensionality reduction through user-defined combinations of quality metrics,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 993–1000, Nov 2009. doi: 10.1109/TVCG.2009.153
- [58] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, “Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1539–1148, Nov 2008. doi: 10.1109/TVCG.2008.153
- [59] C. Turkey, P. Filzmoser, and H. Hauser, “Brushing dimensions - A dual visual analysis model for high-dimensional data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2591–2599, Dec 2011. doi: 10.1109/TVCG.2011.178

- [60] D. Dingen, M. van't Veer, P. Houthuizen, E. H. J. Mestrom, E. H. H. M. Korsten, A. R. A. Bouwman, and J. van Wijk, "RegressionExplorer: Interactive exploration of logistic regression models with subgroup analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 246–255, Jan 2019. doi: 10.1109/TVCG.2018.2865043
- [61] Z. Guo, M. O. Ward, and E. A. Rundensteiner, "Model space visualization for multivariate linear trend discovery," in *IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST '09, Oct 2009, pp. 75–82. doi: 10.1109/VAST.2009.5333431
- [62] H. Piringer, W. Berger, and J. Krasser, "HyperMoVal: Interactive visual validation of regression models for real-time simulation," *Computer Graphics Forum*, vol. 29, no. 3, pp. 983–992, Aug 2010. doi: 10.1111/j.1467-8659.2009.01684.x
- [63] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs, "Multivariate visual explanation for high dimensional datasets," in *IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST '08, Oct 2008, pp. 147–154. doi: 10.1109/VAST.2008.4677368
- [64] S. Das, D. Cashman, R. Chang, and A. Endert, "BEAMES: Interactive multi-model steering, selection, and inspection for regression tasks," *IEEE Computer Graphics and Applications*, vol. 39, no. 5, pp. 20–32, June 2019. doi: 10.1109/MCG.2019.2922592
- [65] J. Krause, A. Perer, and E. Bertini, "INFUSE: Interactive feature selection for predictive modeling of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1614–1623, Dec 2014. doi: 10.1109/TVCG.2014.2346482
- [66] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer, "Guiding feature subset selection with an interactive visualization," in *IEEE Conference on Visual Analytics Science and Technology*, ser. VAST '11, Oct 2011, pp. 111–120. doi: 10.1109/VAST.2011.6102448
- [67] T. Muhlbacher and H. Piringer, "A partition-based framework for building and validating regression models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1962–1971, Dec 2013. doi: 10.1109/TVCG.2013.125
- [68] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski, "The state-of-the-art in predictive visual analytics," *Computer Graphics Forum*, vol. 36, no. 3, pp. 539–562, 2017. doi: 10.1111/cgf.13210
- [69] J. Krause, A. Perer, and E. Bertini, "Using visual analytics to interpret predictive machine learning models," in *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, ser. WHI 2016. ArXiv e-prints, June 2016.
- [70] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 364–373, 2019. doi: 10.1109/TVCG.2018.2864499

- [71] J. Krause, A. Perer, and K. Ng, “Interacting with predictions: Visual inspection of black-box machine learning models,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16, 2016, p. 5686–5697. doi: 10.1145/2858036.2858529
- [72] J. Lu, W. Chen, Y. Ma, J. Ke, Z. Li, F. Zhang, and R. Maciejewski, “Recent progress and trends in predictive visual analytics,” *Frontiers of Computer Science*, vol. 11, no. 2, pp. 192–207, 2017. doi: 10.1007/s11704-016-6028-y
- [73] S. van den Elzen and J. J. van Wijk, “BaobabView: Interactive construction and analysis of decision trees,” in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011, pp. 151–160. doi: 10.1109/VAST.2011.6102453
- [74] J. I. Sanchez and E. L. Levine, “The rise and fall of job analysis and the future of work analysis,” *Annual Review of Psychology*, vol. 63, no. 1, pp. 397–425, Nov 2011. doi: 10.1146/annurev-psych-120710-100401
- [75] M. A. Campion, A. A. Fink, B. J. Ruggeberg, L. Carr, G. M. Phillips, and Ronald B Odman, “Doing competencies well: Best practices in competency modeling,” *Personnel Psychology*, vol. 64, no. 1, pp. 225–262, March 2011. doi: 10.1111/j.1744-6570.2010.01207.x
- [76] “Redux,” <https://redux.js.org/>.
- [77] “React,” <https://reactjs.org/>.
- [78] “D3,” <https://d3js.org/>.
- [79] W. Willett, J. Heer, and M. Agrawala, “Scented widgets: Improving navigation cues with embedded visualizations,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1129–1136, Nov 2007. doi: 10.1109/TVCG.2007.70589
- [80] “Color brewer,” <http://colorbrewer2.com>.
- [81] R. Maciejewski, A. Pattath, Sungahn Ko, R. Hafen, W. S. Cleveland, and D. S. Ebert, “Automated Box-Cox transformations for improved visual encoding,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 1, pp. 130–140, Jan 2013. doi: 10.1109/TVCG.2012.64
- [82] J. G. Nicholls, A. R. Martin, P. A. Fuchs, D. A. Brown, M. E. Diamond, and D. A. Weisblat, *From Neuron to Brain*, 5th ed. Sunderland, MA: Sinauer Associates, Inc., 2001.
- [83] T. Munzner, *Visualization Analysis and Design*. CRC Press, 2014, ch. 7.6.3, pp. 166–168.
- [84] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg, “Evaluation of alternative glyph designs for time series data in a small multiple setting,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, ser. CHI ’13, 2013, pp. 3237–3246. doi: 10.1145/2470654.2466443

- [85] S. Diehl, F. Beck, and M. Burch, “Uncovering strengths and weaknesses of radial visualizations—an empirical approach,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 935–942, Nov 2010. doi: 10.1109/TVCG.2010.209
- [86] J. A. Hartigan and M. A. Wong, “A k-means clustering algorithm,” *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979. doi: 10.2307/2346830
- [87] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [88] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko, “Toward a deeper understanding of the role of interaction in information visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, Nov 2007. doi: 10.1109/TVCG.2007.70515
- [89] B. Shneiderman and C. Plaisant, “Treemaps for space-constrained visualization of hierarchies,” <http://www.cs.umd.edu/hcil/treemap-history>, 1998.
- [90] M. Velasquez and P. T. Hester, “An analysis of multi-criteria decision making methods,” *International Journal of Operations Research*, vol. 10, no. 2, pp. 56–66, 2013.
- [91] T. Joachims, “Optimizing search engines using clickthrough data,” in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02, July 2002, p. 133–142. doi: 10.1145/775047.775067
- [92] B. Carterette, P. N. Bennett, D. M. Chickering, e. C. Dumais, Susan T.”, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, “Here or there preference judgments for relevance,” in *Advances in Information Retrieval*, 2008, pp. 16–27. doi: 10.1007/978-3-540-78646-7\_5
- [93] H. LI, “A short introduction to learning to rank,” *IEICE Transactions on Information and Systems*, vol. E94.D, no. 10, pp. 1854–1862, 2011. doi: 10.1587/transinf.E94.D.1854
- [94] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, “Pair analytics: Capturing reasoning processes in collaborative visual analytics,” in *2011 44th Hawaii International Conference on System Sciences*, Jan 2011, pp. 1–10. doi: 10.1109/HICSS.2011.339
- [95] L. Liang, L. Di, L. Zhang, M. Deng, Z. Qin, S. Zhao, and H. Lin, “Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method,” *Remote Sensing of Environment*, vol. 165, pp. 123–134, 2015. doi: 10.1016/j.rse.2015.04.032
- [96] C. Wu, Z. Niu, Q. Tang, and W. Huang, “Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation,” *Agricultural and Forest Meteorology*, vol. 148, no. 8, pp. 1230 – 1241, 2008. doi: 10.1016/j.agrformet.2008.03.005

- [97] J. A. Gamon, J. Peñuelas, and C. B. Field, “A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency,” *Remote Sensing of Environment*, vol. 41, no. 1, pp. 35 – 44, 1992. doi: 10.1016/0034-4257(92)90059-S
- [98] A. Masjedi, J. Zhao, A. M. Thompson, K.-W. Yang, J. E. Flatt, M. M. Crawford, D. S. Ebert, M. R. Tuinstra, G. Hammer, and S. Chapman, “Sorghum biomass prediction using UAV-based remote sensing data and crop model simulation,” in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS '18)*, July 2018, pp. 7719–7722. doi: 10.1109/IGARSS.2018.8519034
- [99] M. Elbahnasawy, T. Shamseldin, R. Ravi, T. Zhou, Y.-J. Lin, A. Masjedi, E. Flatt, M. Crawford, and A. Habib, “Multi-sensor integration onboard a UAV-based mobile mapping system for agricultural management,” in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS '18)*, July 2018, pp. 3412–3415. doi: 10.1109/IGARSS.2018.8517370
- [100] Z. Zhang, A. Masjedi, J. Zhao, and M. M. Crawford, “Prediction of sorghum biomass based on image based features derived from time series of UAV images,” in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS '17)*, July 2017, pp. 6154–6157. doi: 10.1109/IGARSS.2017.8128413
- [101] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT press, 2002.
- [102] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT press, 2009.
- [103] Q. Liu, C. Chen, Y. Zhang, and Z. Hu, “Feature selection for support vector machines with RBF kernel,” *Artificial Intelligence Review*, vol. 36, no. 2, pp. 99–115, Aug 2011. doi: 10.1007/s10462-011-9205-2
- [104] T. Gerik, B. Bean, and R. Vanderlip, “Sorghum growth and development,” *Texas AgriLife Extension publication*, 2003.
- [105] Z. N. Brandão, V. Sofiatti, J. R. Bezerra, G. B. Ferreira, J. C. Medeiros *et al.*, “Spectral reflectance for growth and yield assessment of irrigated cotton,” *Australian Journal of Crop Science*, vol. 9, no. 1, pp. 75–84, Jan 2015.
- [106] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma, “FRank: A ranking method with fidelity loss,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '07. Association for Computing Machinery, 2007, p. 383–390. doi: 10.1145/1277741.1277808
- [107] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, “Adapting ranking svm to document retrieval,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06, Aug 2006, p. 186–193. doi: 10.1145/1148170.1148205
- [108] C. J. C. Burges, R. Ragno, and Q. V. Le, “Learning to rank with nonsmooth cost functions,” in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06, Dec 2006, p. 193–200.

- [109] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao, “Adapting boosting for information retrieval measures,” *Information Retrieval*, vol. 13, no. 3, pp. 254–270, Sep 2010. doi: 10.1007/s10791-009-9112-1

VITA

## VITA

Jieqiong Zhao is a Ph.D. student in the School of Electrical and Computer Engineering at Purdue University in West Lafayette, IN, USA. Her research interests include visual analytics, information visualization, and human-computer interaction. She received her master's degree in computer science in 2013 from Tufts University in Medford, MA, USA. She received her bachelor's degree in software engineering from Zhejiang University of Technology in Hangzhou, China.