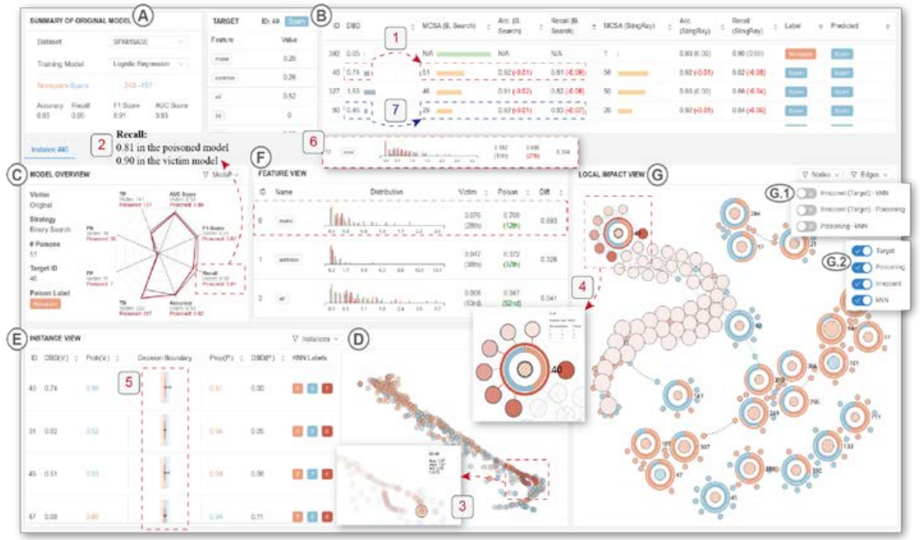
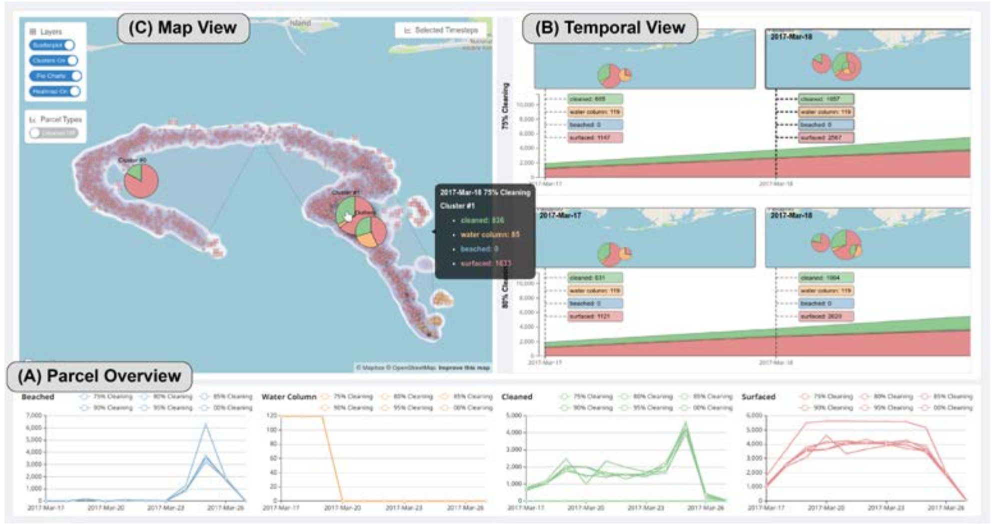
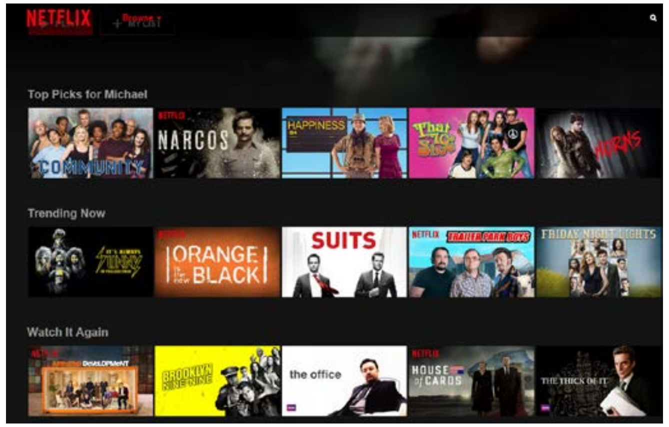


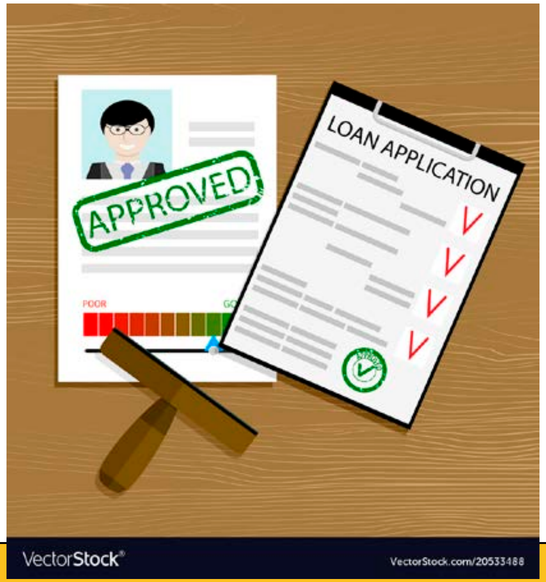


Visualization, Artificial Intelligence and Decision Making

Ross Maciejewski

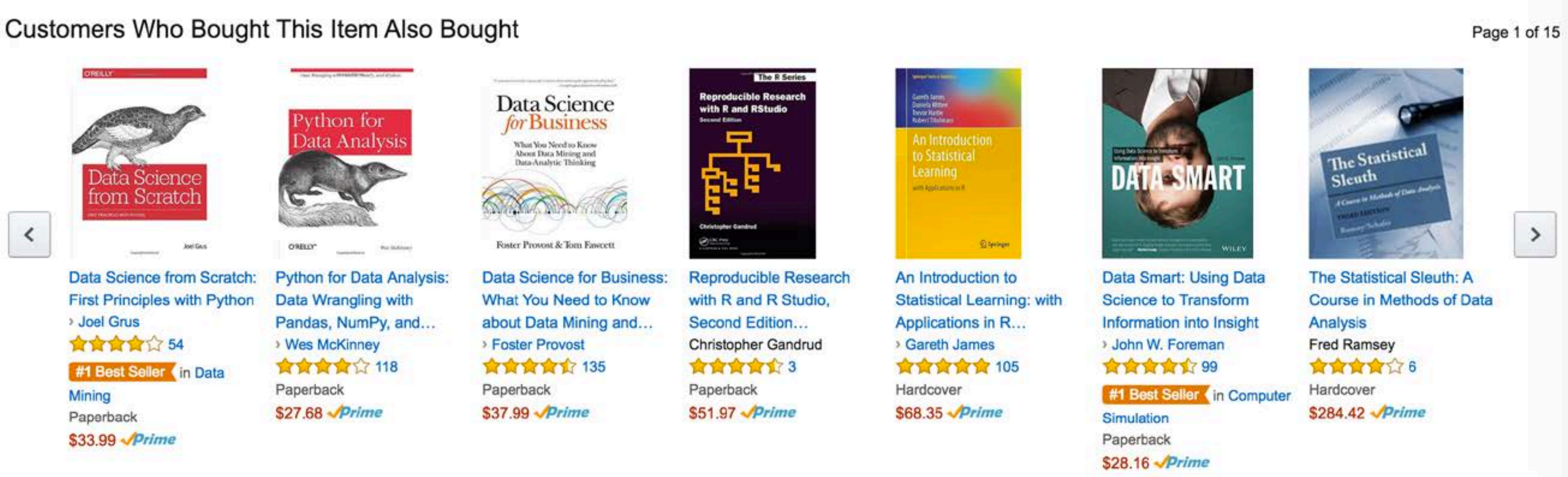
The Role of AI in Decision Making





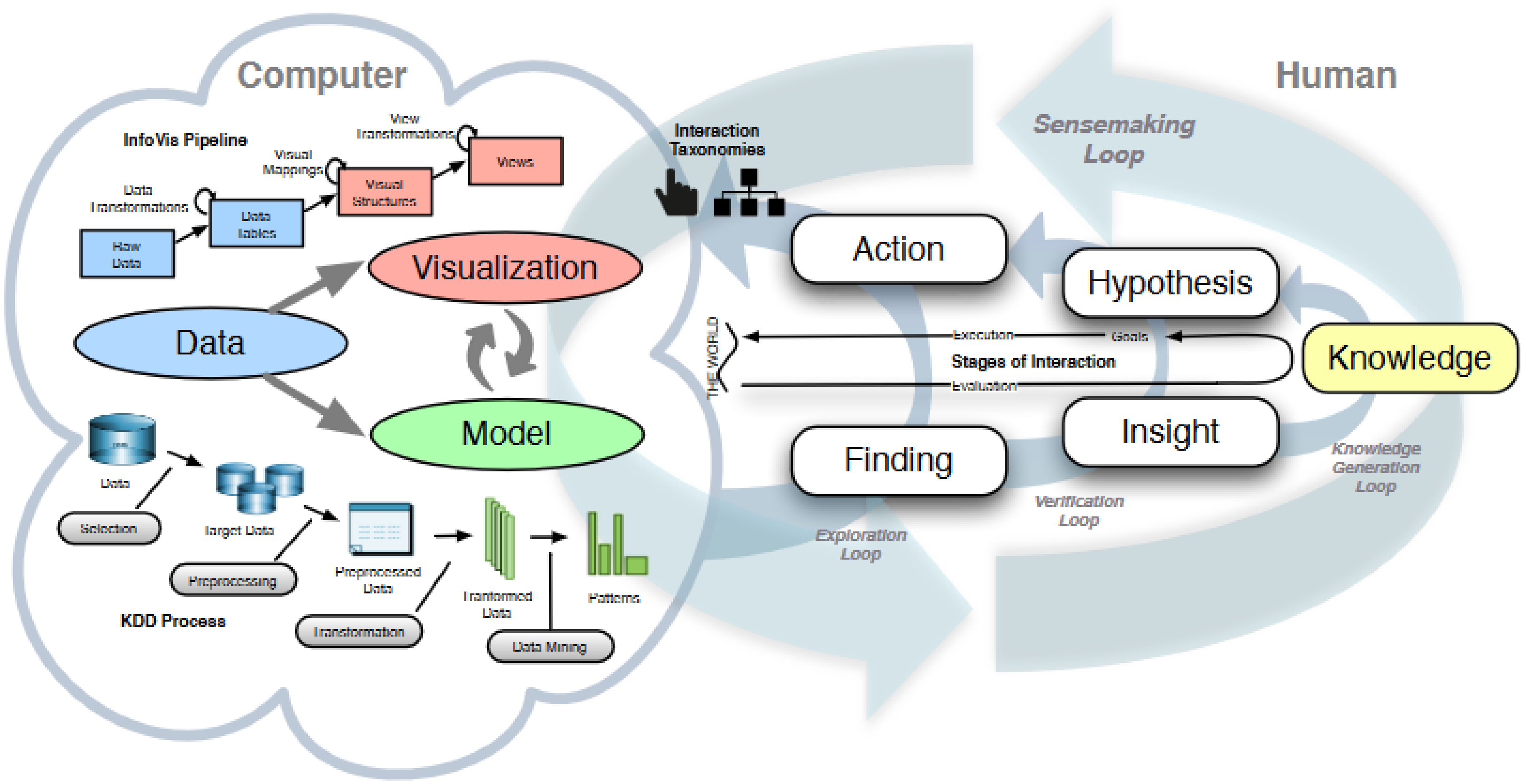
Inconsequential

Consequential



Visual Analytics

- Visual analytics is the science of analytical reasoning supported by interactive user interfaces
- Uses artificial intelligence algorithms combined with interactive visual interfaces
- This allows for the combination of domain expert knowledge with advanced analytics and data exploration to facilitate interactive decision making



D. Sacha, A. Stoffel, F. Stoffel, B.C. Kwon, G. Ellis, D.A. Keim, "Knowledge Generation Model for Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics,*

Geographic Decision Support Systems

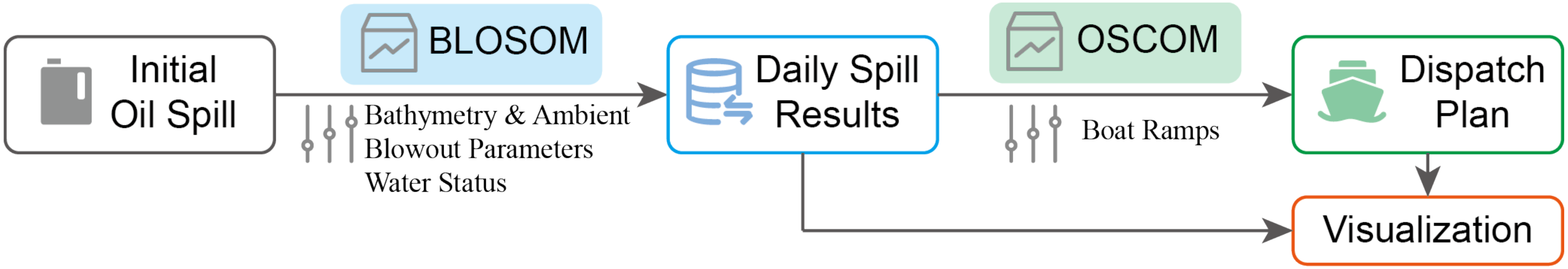
Malik, A., <u>Maciejewski, R.</u>, Collins, T., Ebert, D., T., "Visual Analytics Law Enforcement Toolkit," *IEEE International Conference on Technologies for Homeland Security*, 2010.

Malik, A., <u>Maciejewski, R.</u>, Maule, B., Ebert, D. S., "A Visual Analytics Process for Maritime Resource Allocation and Risk Assessment," *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011.

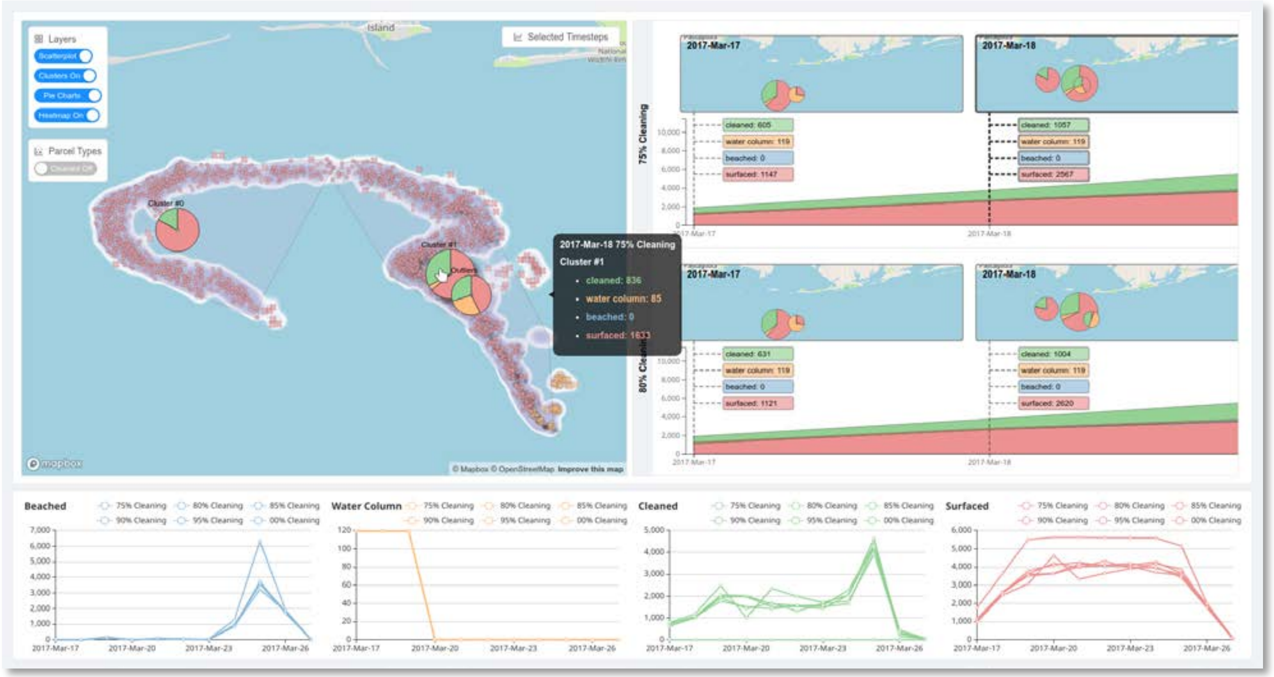
Coast Guard Meritorious Team Commendation (PROTECT), Ross Maciejewski served as a member of the United States Coast Guard Port Resilience for Operational Tactical Enforcement to Combat Terrorism (PROTECT) Team while at Purdue University's Department of Homeland Security Center of Excellence (VACCINE), May 2013.

Razip, A. M. M., Malik, A., Afzal, S., Joshi, S., <u>Maciejewski, R.</u>, Jang, Y., Elmqvist, N., Ebert, D. S., "A Mobile Visual Analytics Approach for Situational Awareness and Risk Assessment," *IEEE Pacific Visualization Symposium*, 2014.

Visual Analytics System for Oil Spill Response and Recovery



- Analytical Tasks
  - Visualization of Oil Spill
  - Visual Comparison
  - Decision Making Support
- Visual Design
  - Parcel Overview
  - Temporal View
  - Map View



Informing Coastal Community Planning and Response to Environmental Change in Regions with Offshore Oil and Gas Operations

Yuxin Ma, Prannoy Chandra Pydi Medini, Jake R. Nelson, Ran Wei, Tony Grubesic, Jorge A. Sefair, Ross Maciejewski

VADER Lab, CIDSE, Arizona State University

- **Code Available at:** https://github.com/VADERASU/BlosomAndOscom
- **Project Website:** http://vader.lab.asu.edu

Acknowledgement

Gulf Research Program – National Academies

Center for Accelerating Operational Efficiency

Operations Research & Systems Analysis

Improving process and decision time

Economic Analysis

Understanding the true cost

Homeland Security Risk Sciences

Identifying and prioritizing risk

Data Analytics

Real-time rapid response

“Center for Accelerating Operational Efficiency,” *Department of Homeland Security*, $40M, 8/17 – 8/27, PI: Ross Maciejewski; Co-PI (ASU only listed): Pitu Mirchandani, Ronald Askin, Huan Liu, Jingrui He, Hanghang Tong, Jorge Sefair, Giulia Pederelli, Gail-Joon Ahn

Artificial Intelligence in the Homeland Security Enterprise

- Technology that relies on a set of algorithms and techniques to solve problems that humans perform intuitively and near automatically
- Examples of AI in HSE
  - Verification and identification using biometrics* (face, iris, voice, fingerprints): CBP Office of Field Operations, TSA, USCIS
  - Intelligent illicit object detection

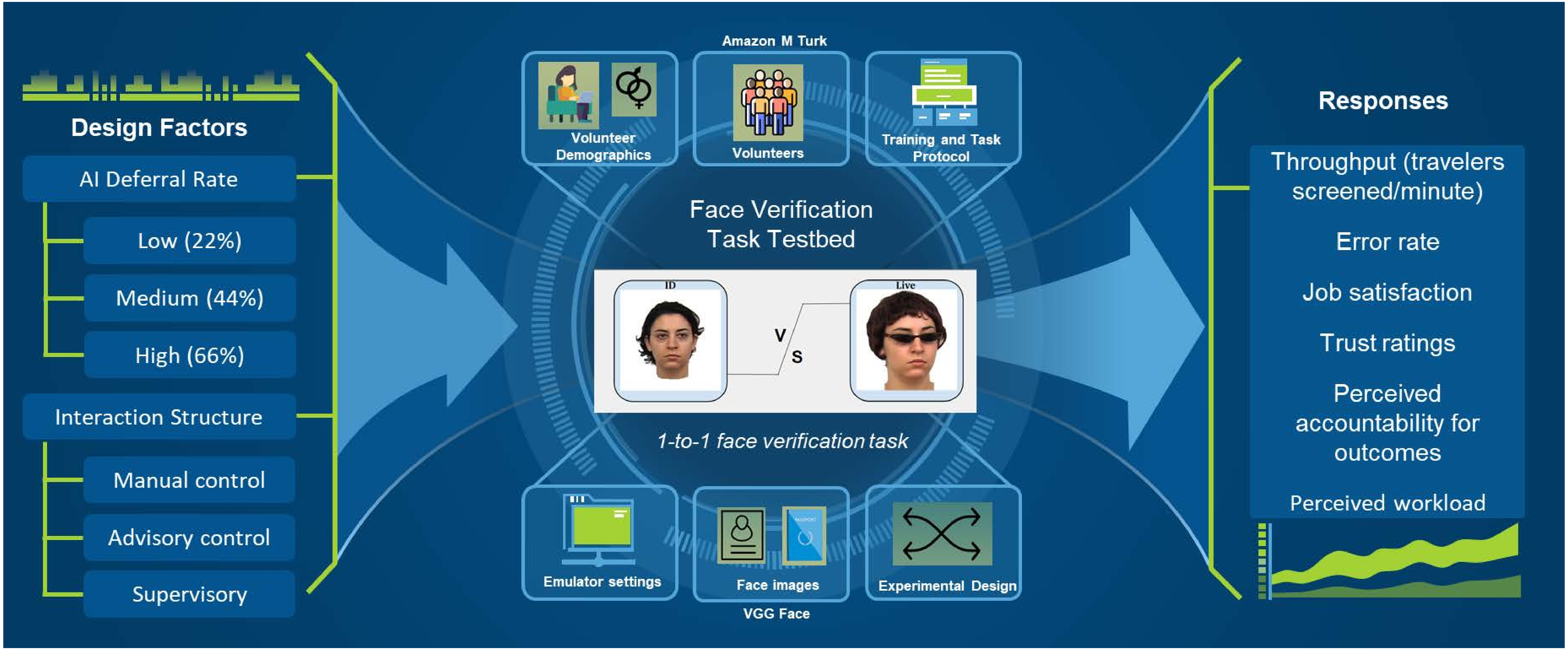
Face recognition and verification in airport security

Smiths Detection's ICMORE scanner**

**DHS Winter Study Biometrics Roadmap, 2015-2018 Final Report (2016)*

** https://www.arabianaerospace.aero/smiths-detection-highlights-weapon-recognition-in-airport-show-reveal.html

Deferring Decisions:
Effects on Human-AI Team Performance



PIs: Mickey Mancenido and Erin Chiou

Human-AI Teaming in Decision Making

- You'd think after years of using Google Maps we'd trust that it knows what it's doing. Still, we think, "Maybe taking the backroads would be faster."[1]
- People are even less trusting of algorithms if they've seen them fail, even a little. And they're harder on algorithms in this way than they are on other people.[2,3]
- An underlying goal of many visualization methods is to inject domain knowledge into the analysis and point out potential algorithmic errors to the end user for updating and correction.
- Visualization could potentially contribute to algorithmic aversion during forecasting tasks and lead to reduced performance.

1 - Walter Frick. Here's Why People Trust Human Judgment Over Algorithms. *Harvard Business Review.* February 27, 2015. https://hbr.org/2015/02/heres-why-people-trust-human-judgment-over-algorithms
2 – Berkeley J Dietvorst. 2016. People Reject (Superior) Algorithms Because They Compare Them to Counter-Normative Reference Points. 2016. https://ssrn.com/abstract=2881503
3 – Berkeley J Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1): 114-126.

How Will Humans Use Predictions?

PVA

- What is Predictive Visual Analytics

Box Office Prediction

- *VAST Grand Challenge Award: Outstanding Comprehensive Submission* for "VADER/VIS-Grand Challenge" awarded at the IEEE Conference on Visual Analytics Science and Technology (**Steptoe, M.**, Krueger, R., **Zhang, Y., Liang, X., Garcia, R., Kadambi, S.,** Ertl, T., Maciejewski, R.)

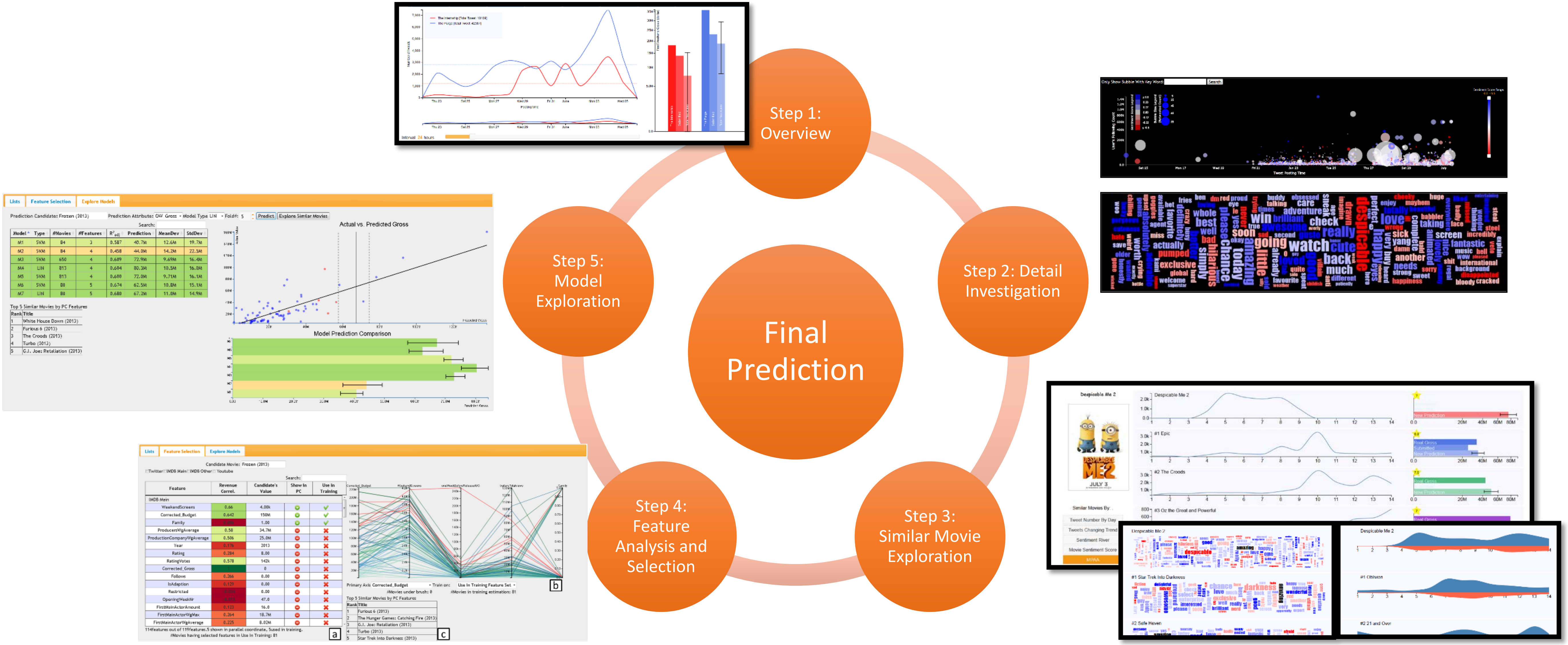
Extended Work

- Small Size User Study

Formal User Study

- Buchanan, V., **Lu, Y.**, McNeese, N., **Steptoe, M.**, Maciejewski, R., Cooke, N., "The Role of Teamwork in the Analysis of Big Data --- A Study of Visual Analytics and Box Office Prediction," *Big Data*, 5(1): 53-66, 2017.
- Human-in-the-Loop Prediction Accuracy – In Progress

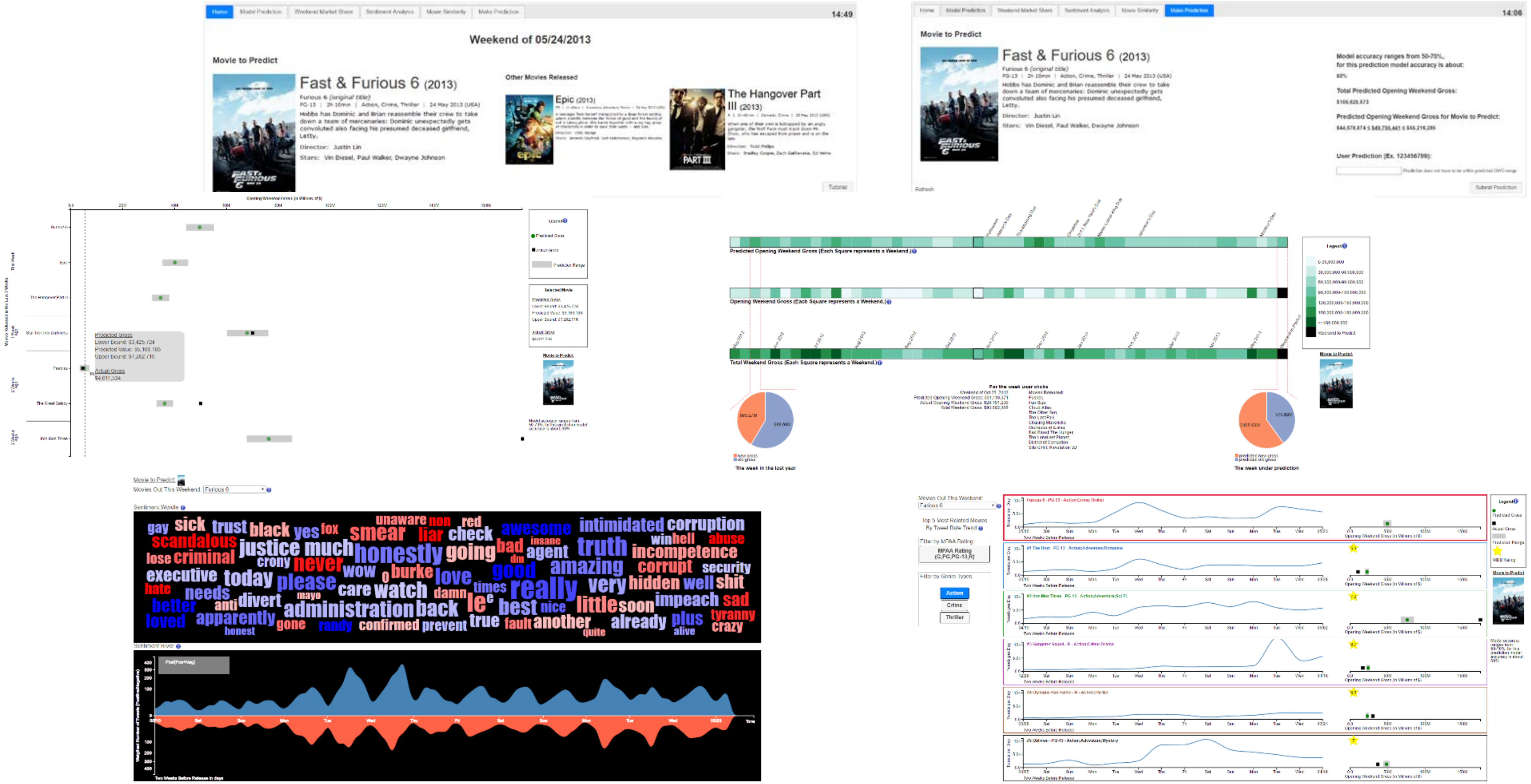
Box Office Prediction



Lu, Yafeng, Robert Krüger, Dennis Thom, Feng Wang, Steffen Koch, Thomas Ertl, and Ross Maciejewski. "Integrating predictive analytics and social media." *IEEE Conference on Visual Analytics Science and Technology,* pp. 193-202. IEEE, 2014.

Explore Prediction Accuracy

- Modify our previous system and conducted a controlled experiment
- 20 participants.
- 9 Movies
- 3 Models
- 6 interfaces

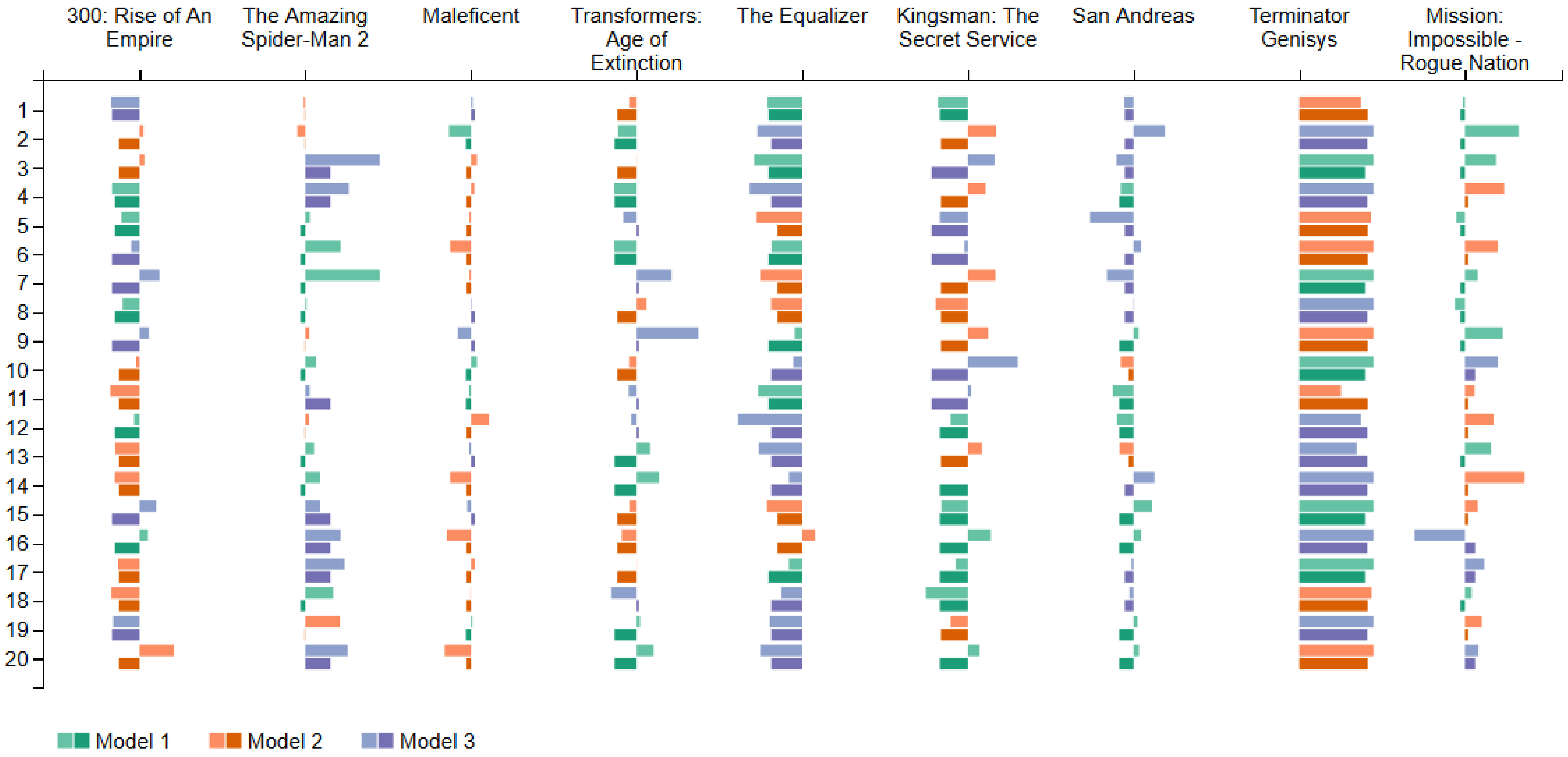


Explore Prediction Accuracy

- Procedure
  - Training
    - Use slides that covered the purpose of the study (box office prediction), the usage of the system and how to interpret the visualized information.
    - Use a quiz to test the understanding of crucial information.
  - Practice Prediction (1 movie, *Fast & Furious 6*)
  - Real Prediction (9 movies, 3 models, randomly ordered)
    - 300: Rise of An Empire, The Amazing Spider-Man 2, Maleficent, Transformers: Age of Extinction, The Equalizer, Kingsman: The Secret Service, San Andreas, Terminator Genisys, and Mission: Impossible – Rogue Nation
    - 15 min to explore and finalize their prediction
  - Questionnaires
    - What data they used to make their decisions and predictions and why.
    - Workload (NASA TLX)

Explore Prediction Accuracy

- Participant predictions were compared to model predictions



72 (40%) have a lower RAE than the model, while 108 (60%) have a higher RAE than the model.

Over 50% of the participant prediction errors are larger than the model predictions errors in all three models.

Considerations for Human-Machine Intelligence

- **Domain Knowledge Integration** - There are domains where human background knowledge is essential and where a lot of tacit knowledge which is difficult to represent in an algorithm plays a role. In such a case the human-in-the-loop approach may yield much better results.[1]

- **Visualization for Trust** - Studies report that forecasters may desire to adjust algorithmic outputs to gain a sense of ownership of the forecasts due to a lack of trust in statistical models.[2]

- **Visualization and Learning** - Typically that type of system means that the user will have some interactions that change a model, whether directly or indirectly. Getting engagement like that may really change the landscape of participation. It changes the idea of accuracy that you can test because the accuracy will evolve based on the human.

- **How can we measure the knowledge integration? What is the baseline when truly supporting human-machine tasks?**

1 - Research has shown that domain expertise diminished people's reliance on algorithmic forecasts which led to a worse performance. (Hal R Arkes, Robyn M Dawes, Caryn Christensen. 1986. Factors Influencing the Use of a Decision Rule in a Probabilistic Task. *Organizational Behavior and Human Decision Processes*. 37(1):93-110)

2 - Berkeley J. Dietvorst, Joseph P Simmons, and Cade Massey. 2016. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science.*

The Role of Visualization in AI

- “Computation and analyses are often seen as black boxes that take tables as input and output, along with set of parameters, and run to completion or error without interruption”[1]

- “… calls for more research [...] on designing analysis modules that can repair computations when data changes, provide continuous feedback during the computation, and be **steered by user interaction** when possible”[1]

1 - J.-D. Fekete. Visual Analytics Infrastructures: From Data Management to Exploration. Computer, 46(7):22–29, 2013

MÜHLBACHER T., PIRINGER H., GRATZL S., SEDLMAIR M., STREIT M.: Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations. IEEE Transactions on Visualization and Computer Graphics 20, 12 (2014), 1643–1652

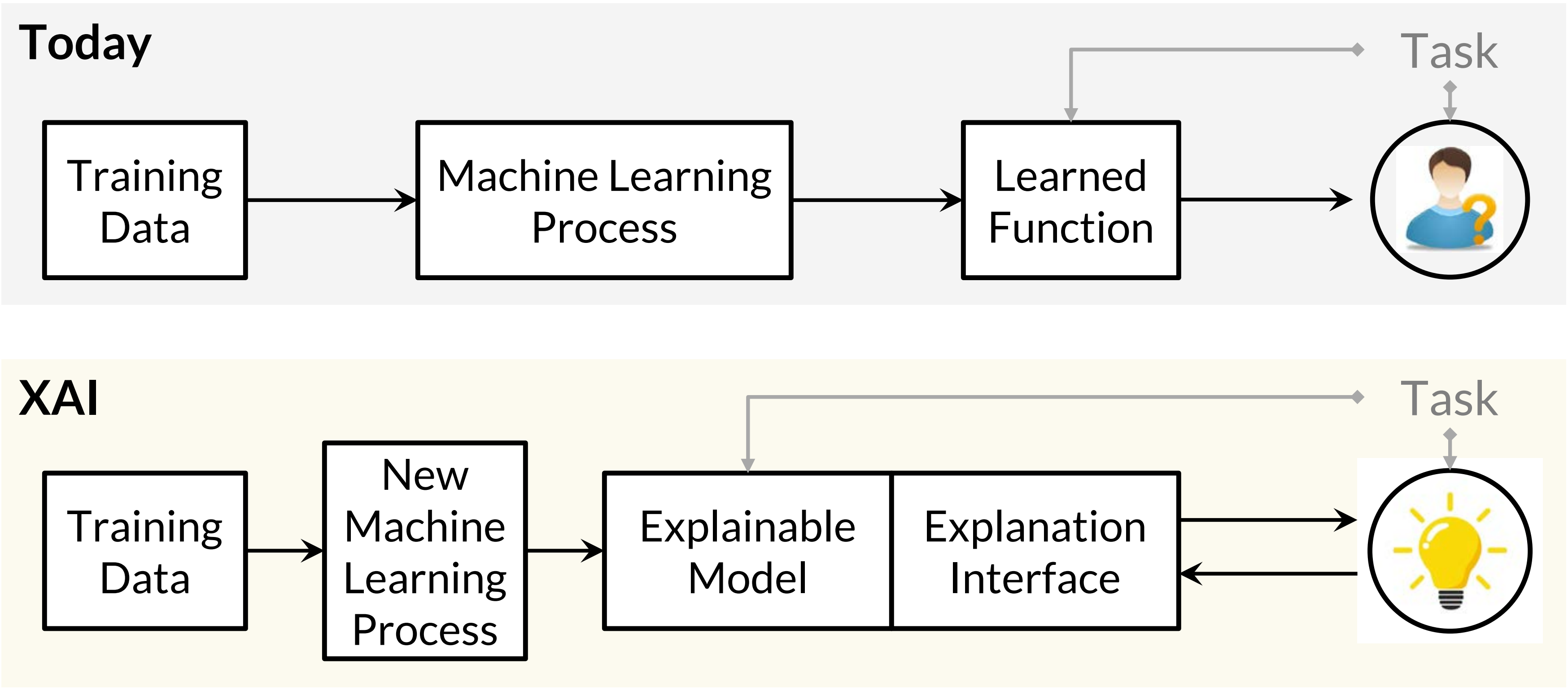
TZENG F.-Y., MA K.-L.: Opening the Black Box-Data Driven Visualization of Neural Networks. In IEEE Visualization. (2005), IEEE, pp. 383–390.

MAY T., BANNACH A., DAVEY J., RUPPERT T., KOHLHAMMER J.: Guiding Feature Subset Selection With an Interactive Visualization. In IEEE Symposium on Visual Analytics Science and Technology (2011), IEEE, pp. 111–120

LU Y., KRÜGER R., THOM D., WANG F., KOCH S., ERTL T., **MACIEJEWSKI R.**: Integrating Predictive Analytics and Social Media. In IEEE Conference on Visual Analytics Science and Technology (2014), IEEE, pp. 193–202.

Explainable AI (XAI)

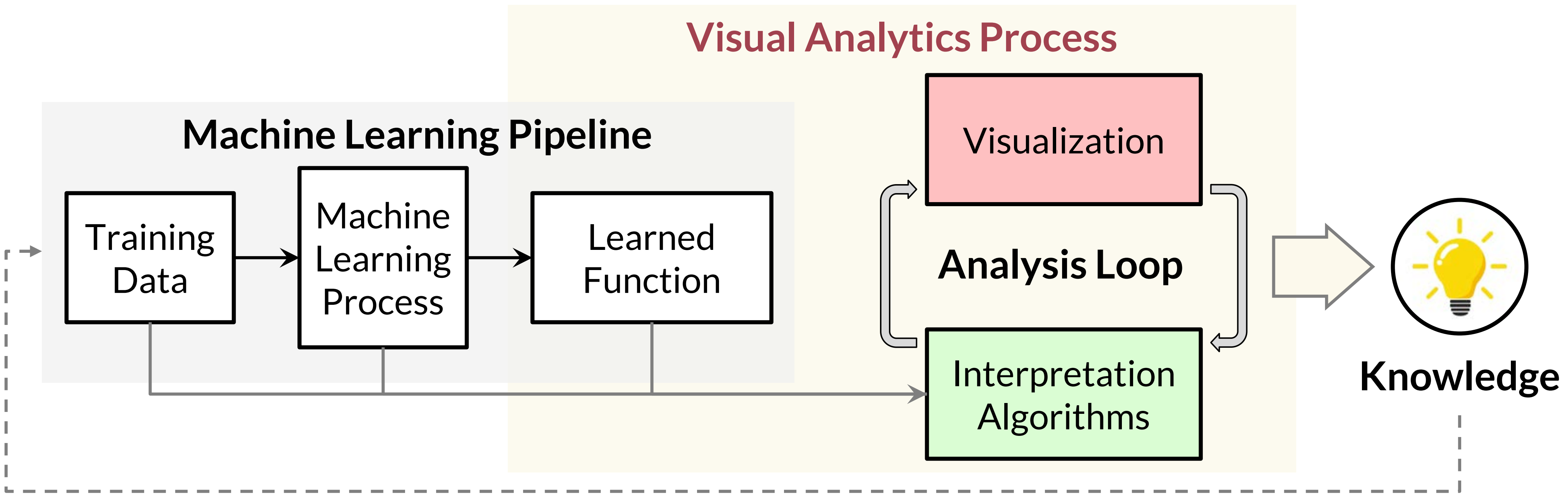
- A suite of machine learning techniques:
  - **Explainable models** with high-level performance
  - **Understand**, appropriate **trust**, and **manage AI partners**



- Questions
  - **Why** did the model **do that**?
  - **Why not** something else?
  - **When** to **succeed** and **fail**?
  - **When** to **trust**?
  - **How** to **correct errors**?

Visual Analytics in Explainable AI

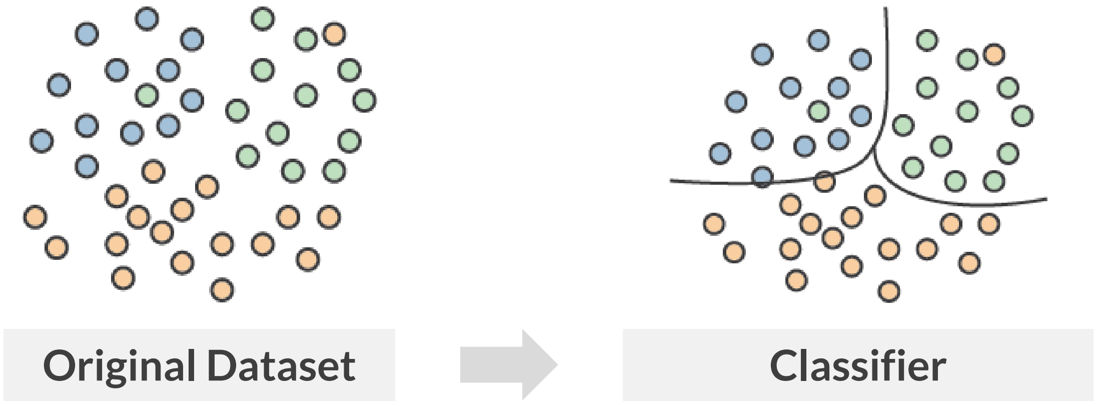
- Visual Analytics
  - **Combine** the automated analysis with interactive visualizations
  - **Enhance** the understanding, reasoning, and decision making



Classification

- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.

Task	Attribute set, x	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies



Visualizing Class Separations

- High-dimensional Labeled Dataset
  - Machine learning: Classification & Clustering

- Dimension Reduction
  - Widely-used for visualizing high-dimensional labeled datasets

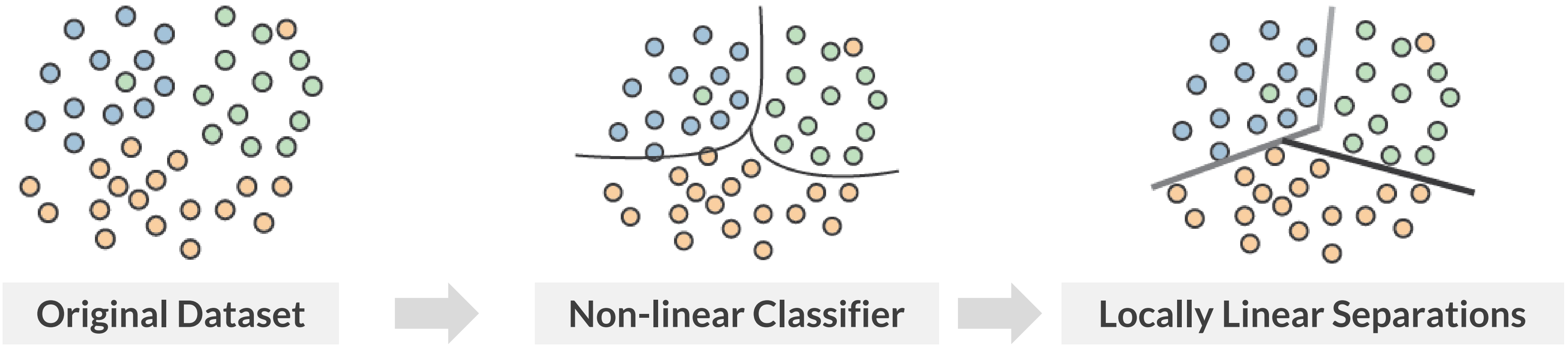
- Challenges in DR Methods

	Linear		Non-linear
+	Margins can be easily illustrated	+	Can handle non-linear separations
-	Unable to handle non-linear structures	-	Cause heavy distortions and patterns

	Supervised		Unsupervised
+	Optimized with-in class distributions	+	Less requirement for datasets (no need for labels)
-	Distortions on between-class distances	-	Difficult to depict class separation patterns

Visualizing Class Separations

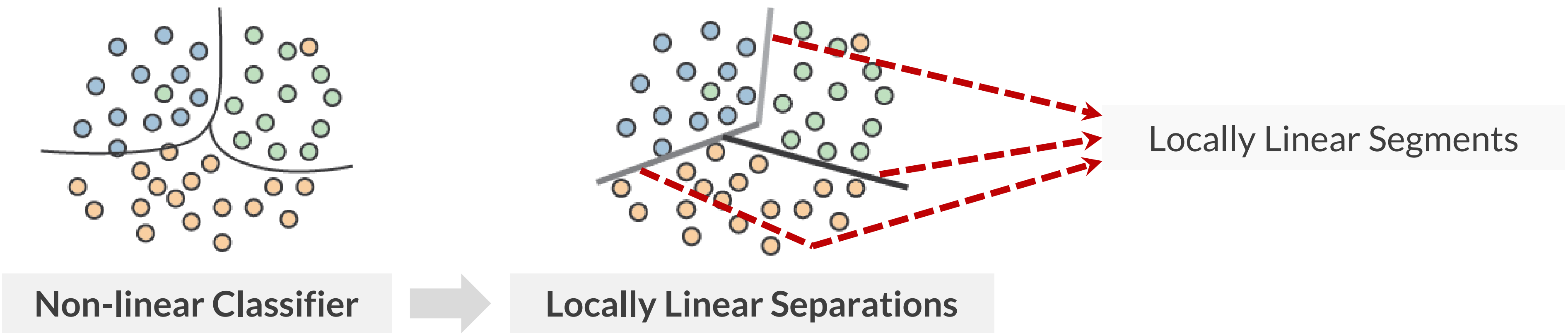
- Contributions
  - A novel approach for detecting locally linear separations in high-dimensional labeled datasets with complex class boundary structures
  - A visual analysis framework that facilitates the exploration and diagnosis of complex class boundaries
- A Way in Between: Locally Linear Separations



- **Advantages**
  - Easy-to-understand Linear Projections
  - Ability to handle complex non-linear class separations

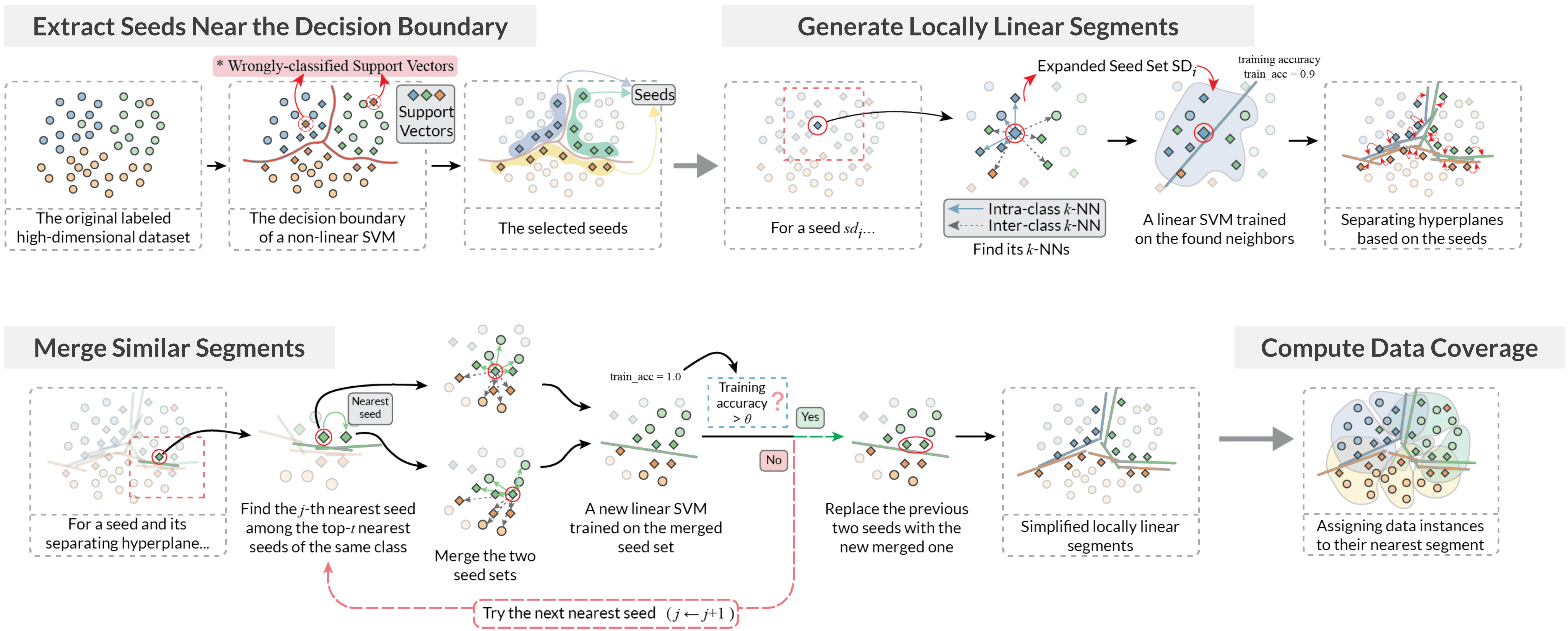
Locally Linear Segment

- Motivation
  - Decision boundaries of classifiers as a tool to describe class separations
  - Local linearity analysis in machine learning (e.g. LIME[1])
- Definition
  - A set of linear approximations extracted from the original decision boundary



[1] Ribeiro et al. “Why should I trust you?” Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

Extraction of Locally Linear Segments



Visual Analytics Framework

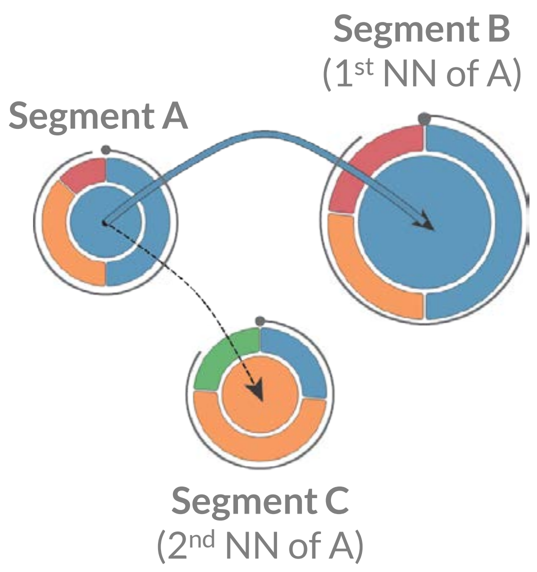
- Task 1: Macroscopic Analysis Overview of the locally linear segments
  - Show the number of segments and highlight the major ones
  - Reveal the coverage of data instances under each segment
  - Depict the locations of the segments and relationships among different segments
- Task 2: Microscopic Analysis Detailed analysis of specific segments
  - Examine the data distribution and separation near a segment
  - Show the primary features used for determining class separation
  - Exploring the neighboring segments
  - Trace a path between segments along decision boundaries

Macroscopic Analysis

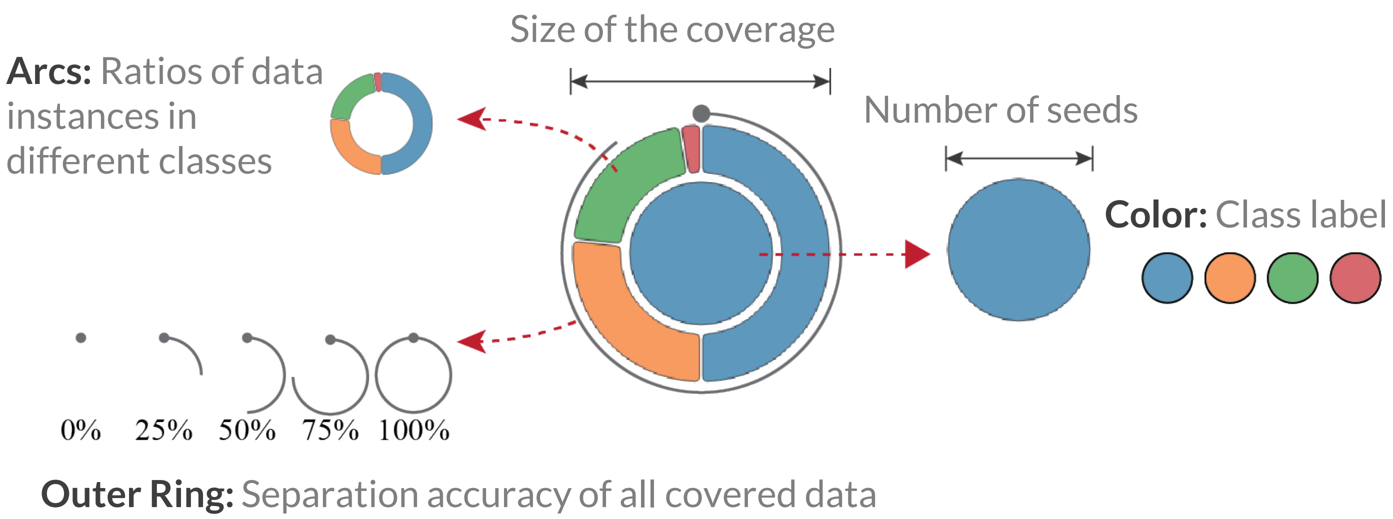
Segment Relation View

- Segment Graph
  - Visualize the segment relationships as a graph structure

Graph Structure



Segment Glyphs



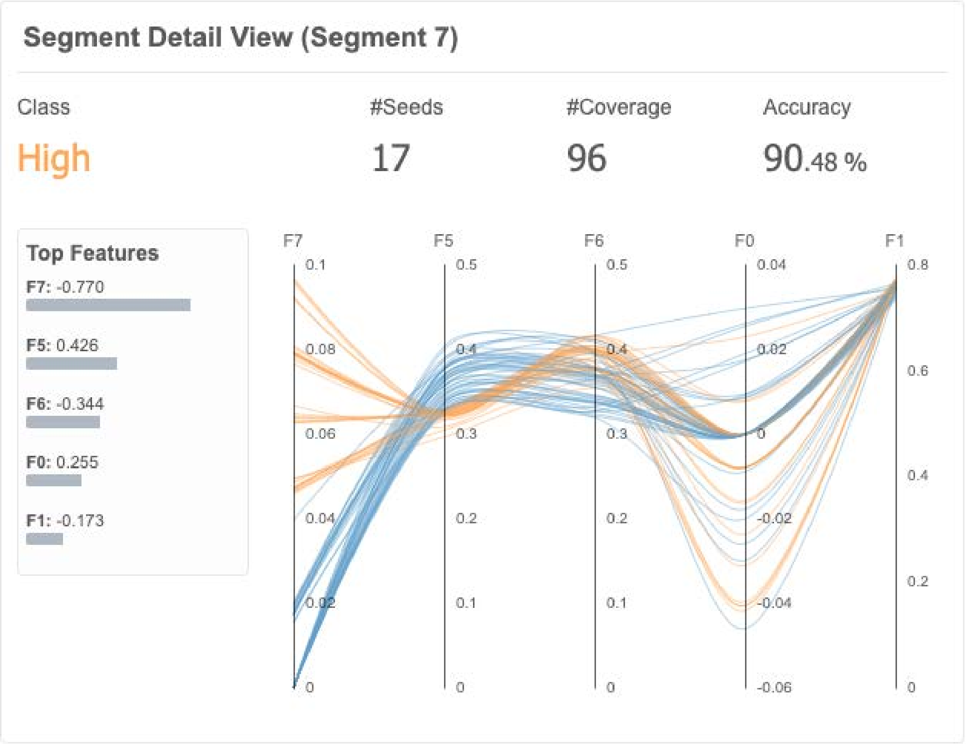
Edges

- **Visual Encoding of Edges**
  - **Curvature:** cosine of angles between the separating hyperplanes
  - **Thickness & Length:** Distances between segments

Microscopic Analysis

Segment Detail View

- Details of Covered Data Instances
  - Number of the covered instances
  - Data distribution (with PCP)
  - Dominant features for separating the local region

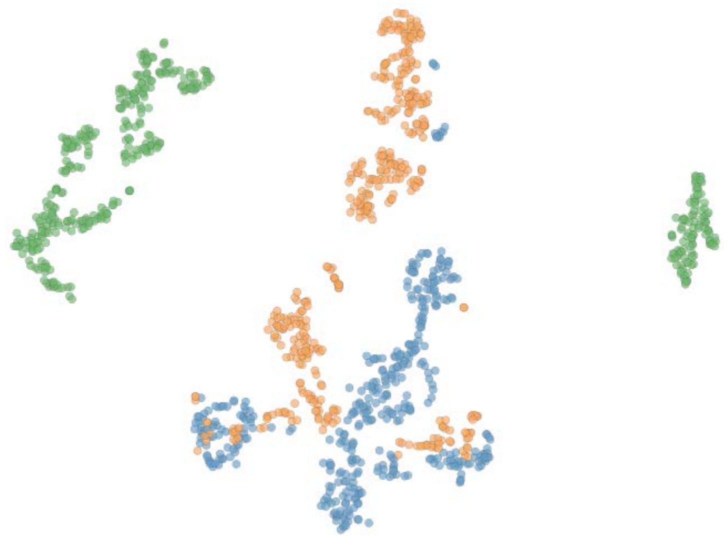


Macroscopic Analysis

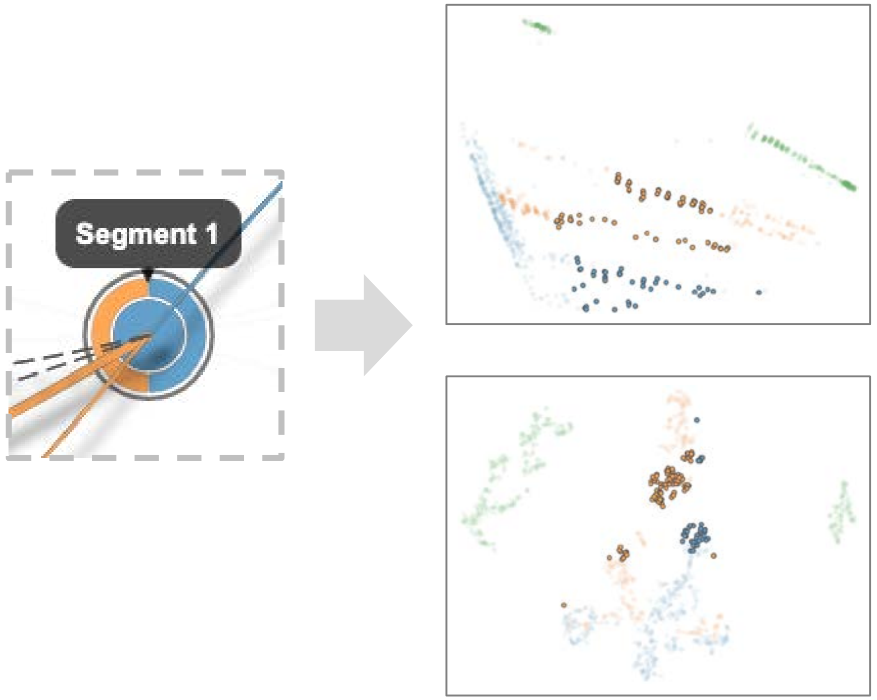
Projection View

Linear Projection (PCA)
Non-distorted view of separations

t-SNE Projection
Initial impression of the distribution



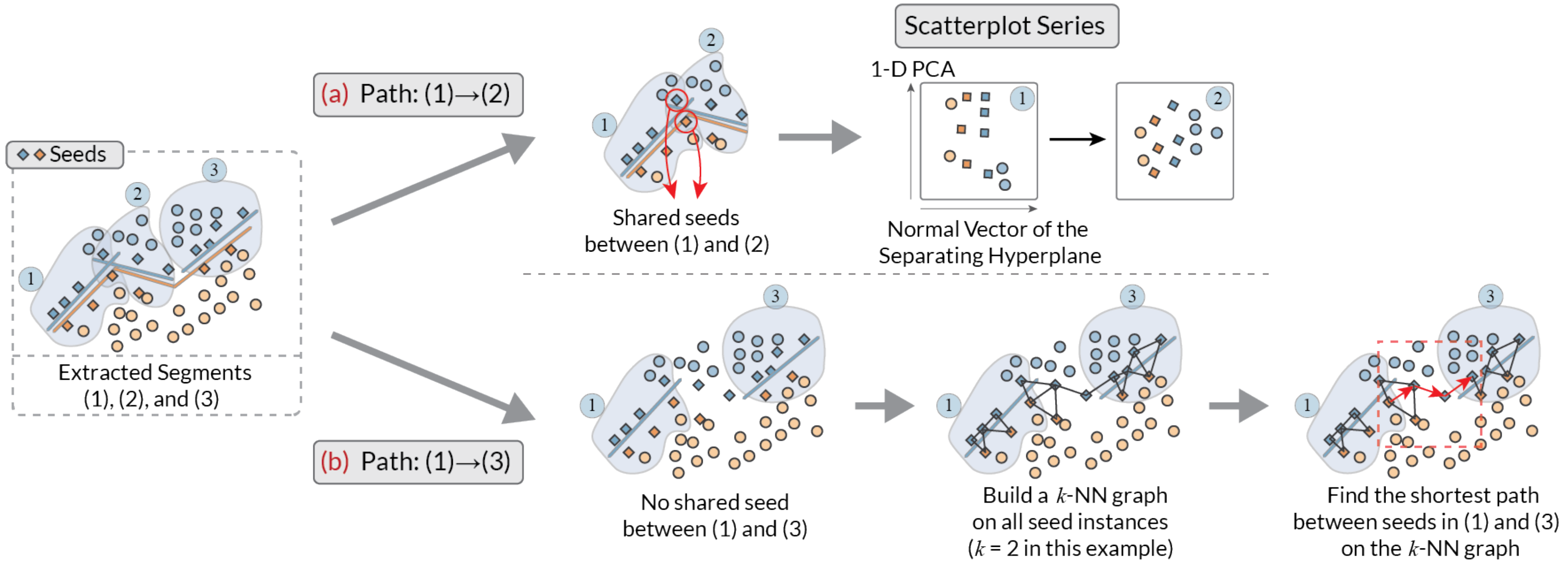
Linked Interaction
Highlight a specific segment



Microscopic Analysis

Path Exploration View

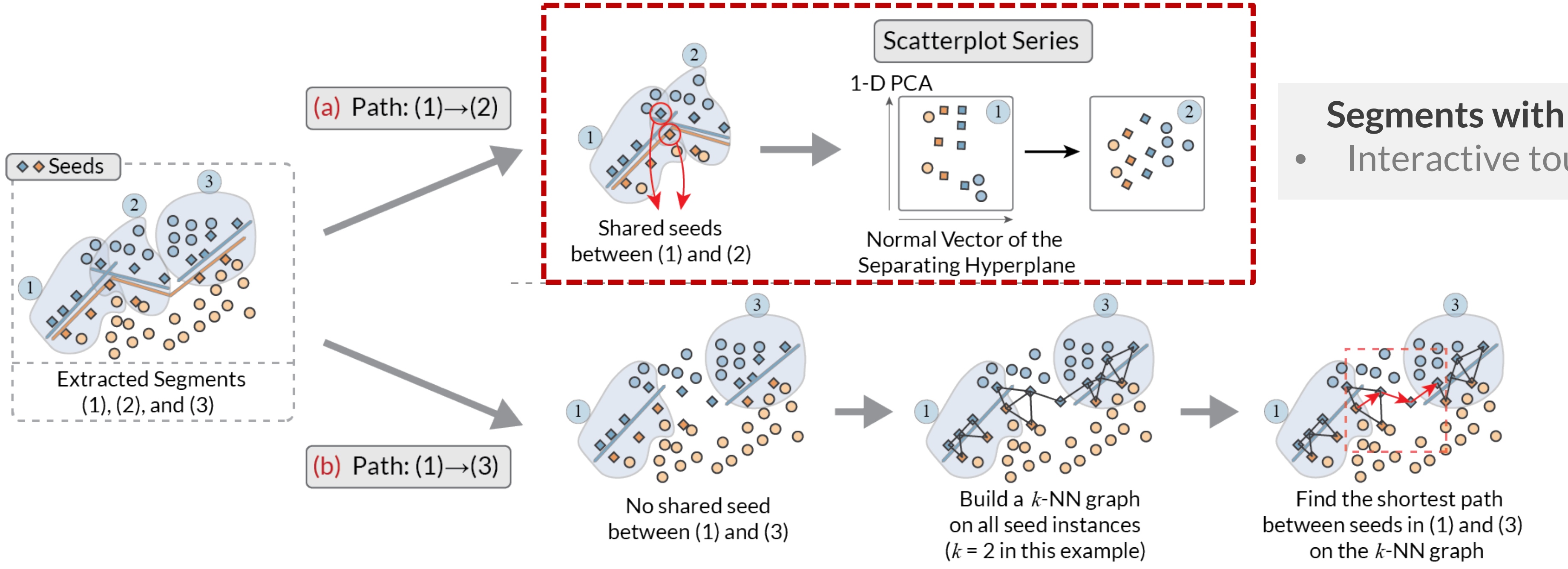
- Goal
  - Present how two segments are connected with each other
  - Flexible traverse between segments



Microscopic Analysis

Path Exploration View

- Goal
  - Present how two segments are connected with each other
  - Flexible traverse between segments



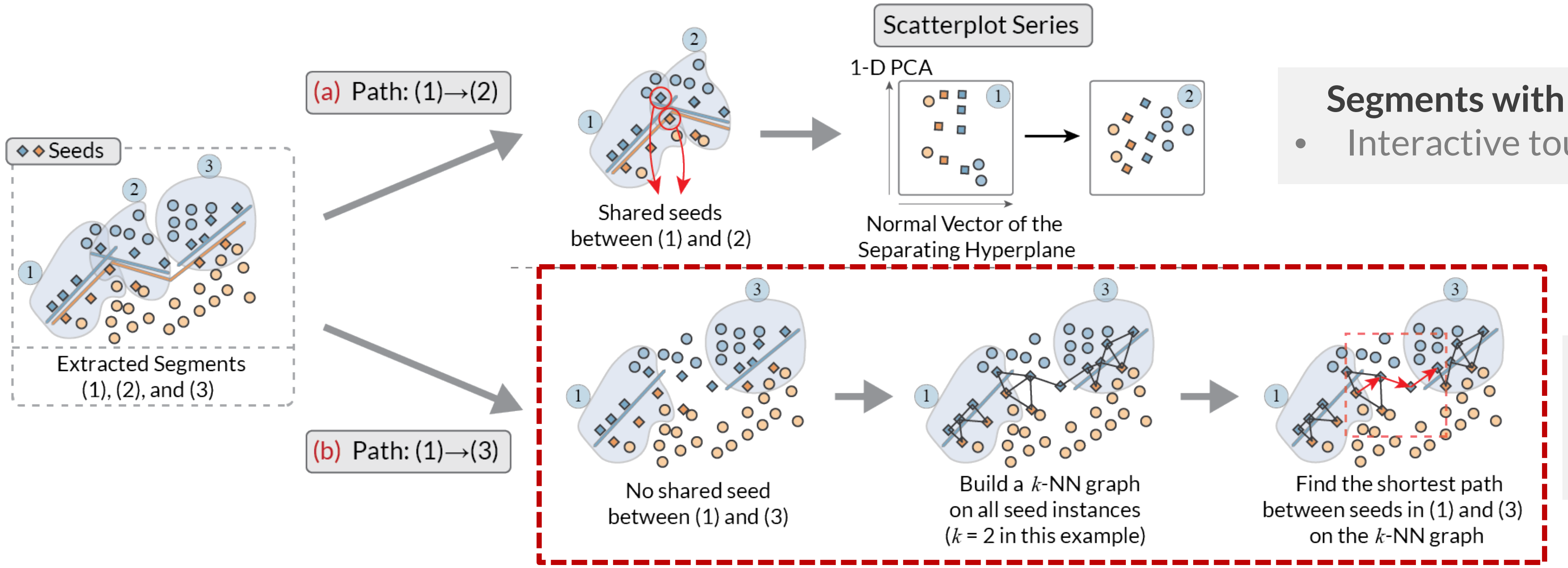
Segments with Shared Instances

- Interactive touring interaction

Microscopic Analysis

Path Exploration View

- Goal
  - Present how two segments are connected with each other
  - Flexible traverse between segments



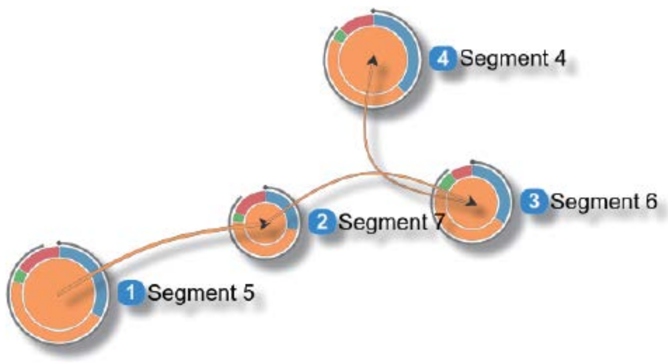
Segments with Shared Instances
- Interactive touring interaction

Segments w/o Shared Instances
- (Interpolate bridging segments)
- Interactive touring interaction

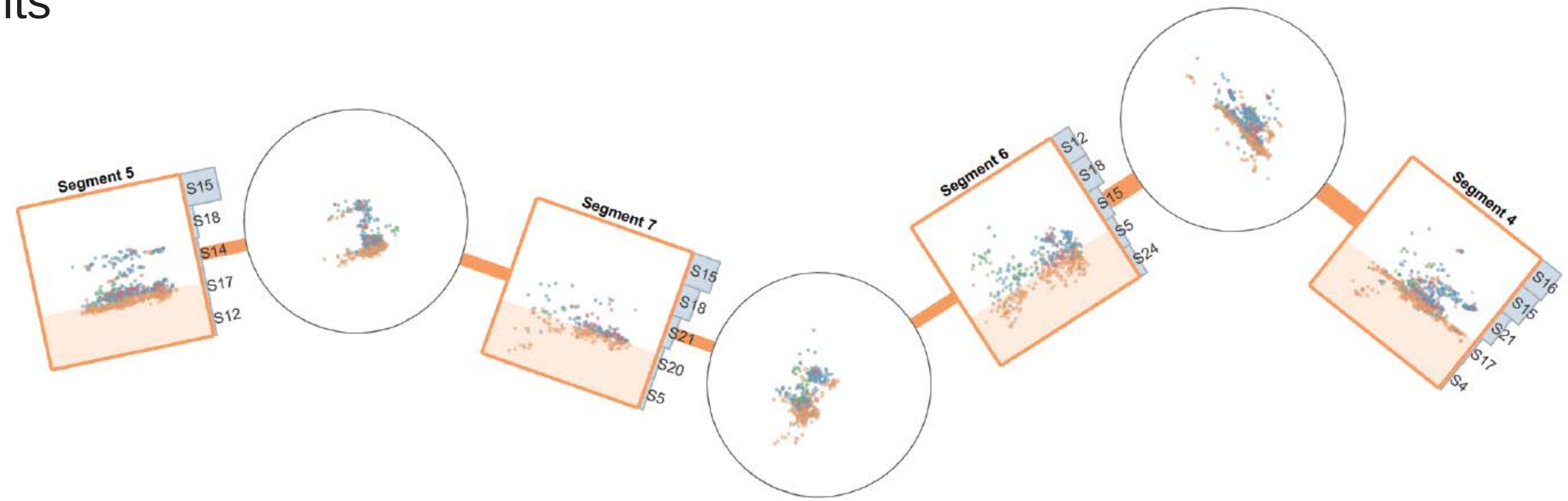
Microscopic Analysis

Path Exploration View

- Goal
  - Present how two segments are connected with each other
  - Flexible traverse between segments



Selected Path on the Segment Graph

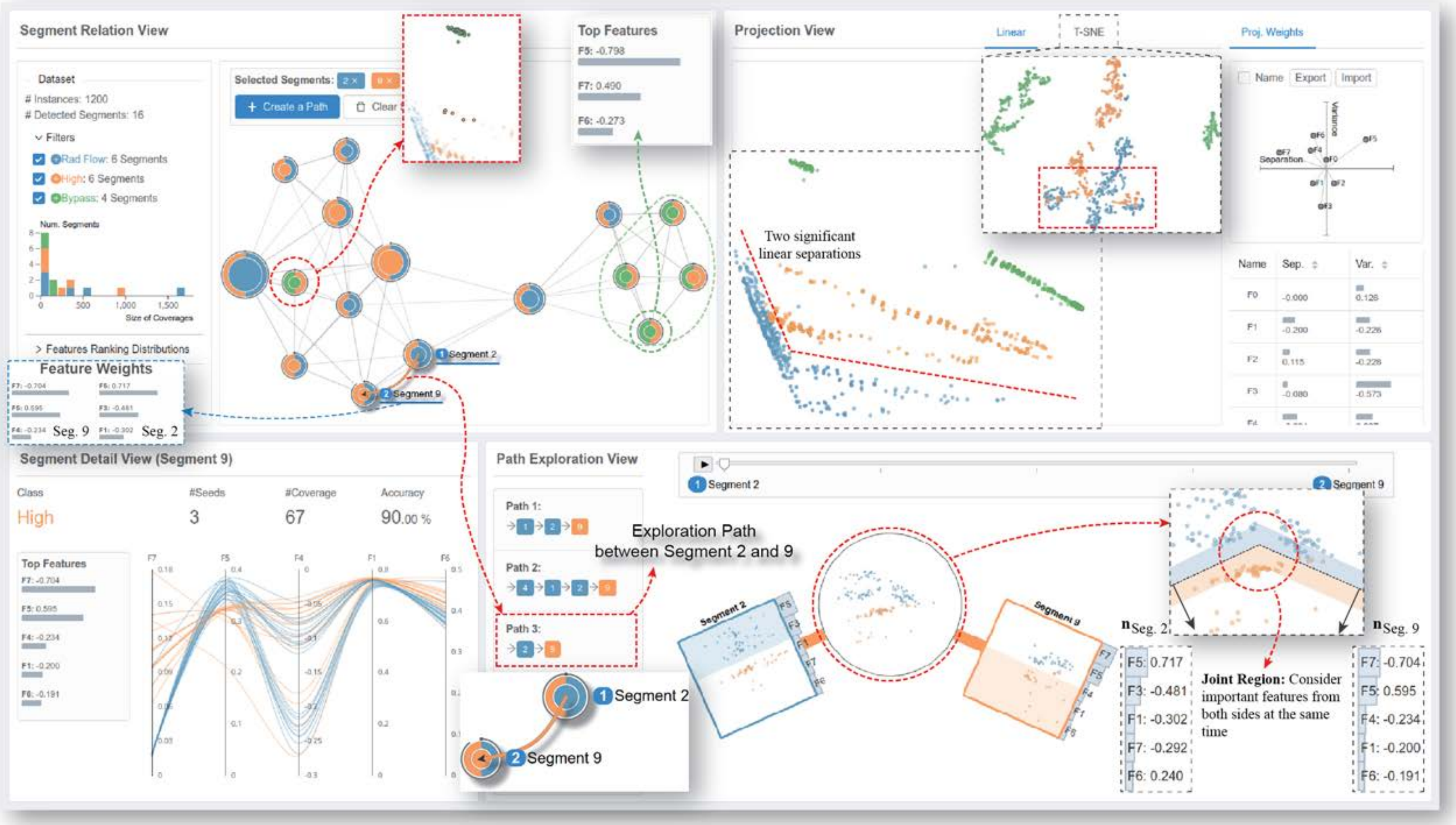


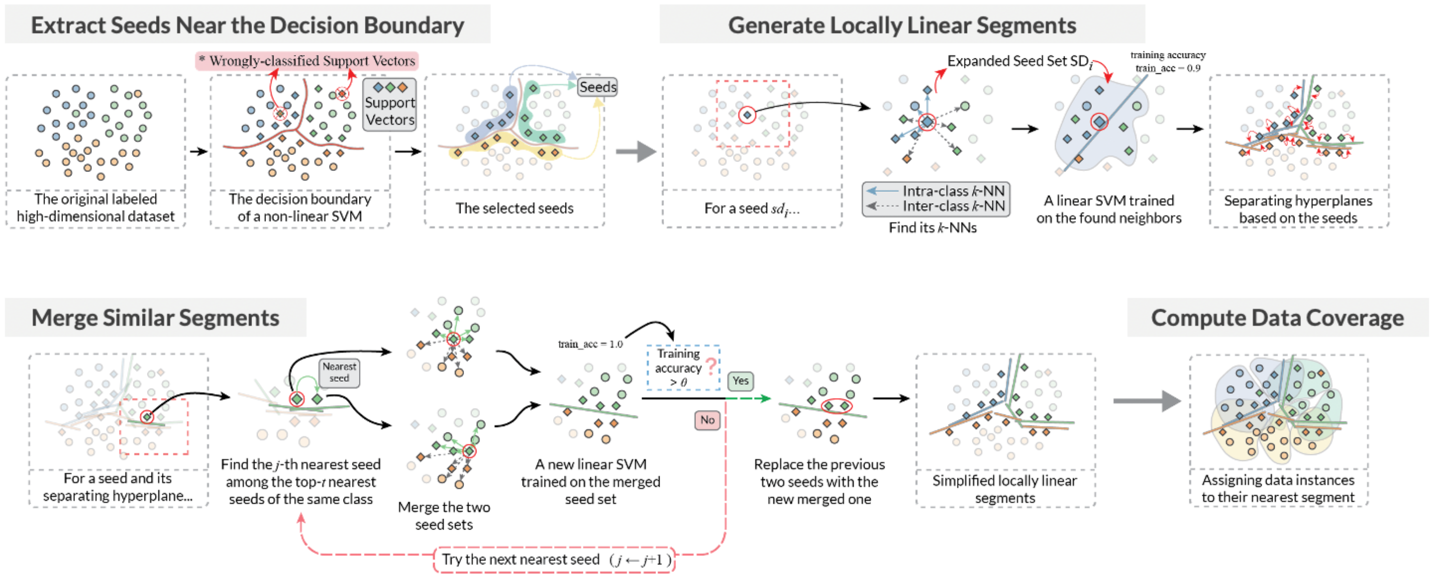
Representative Scatterplots for All Segments

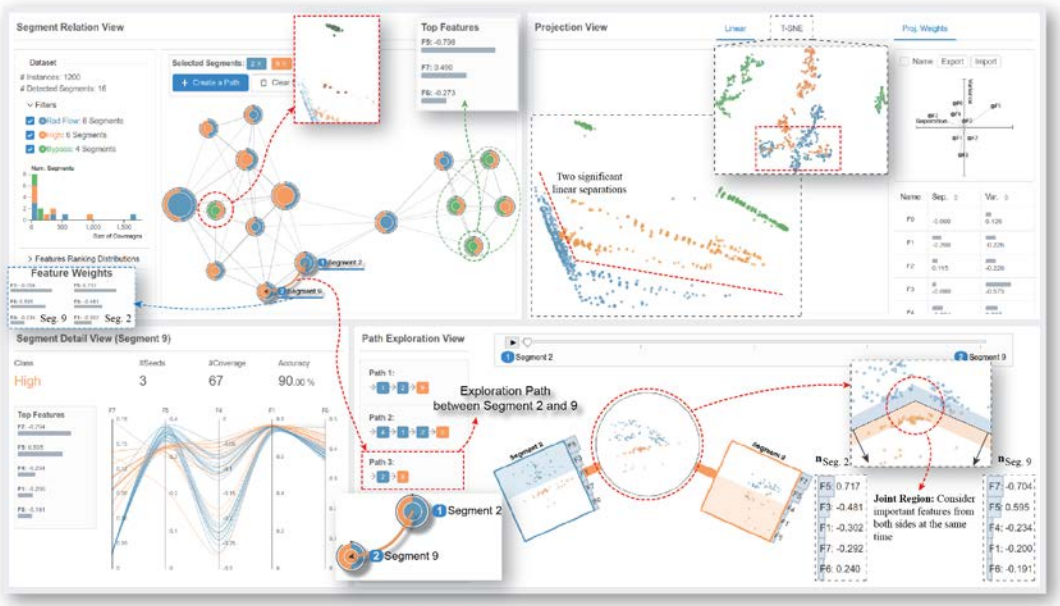
- **Layout**
  - Links of scatterplots in a zig-zag way
- **Interaction**
  - Touring interaction between representative scatterplots

Case Study

- Shuttle StatLog Dataset
  - 9 Numerical sensor readings for deciding radiator positions
  - 3 Classes: Rad Flow, High, Bypass
  - Subsampled into 400 instances for each class







Visual Analysis of Class Separations with Locally Linear Segments

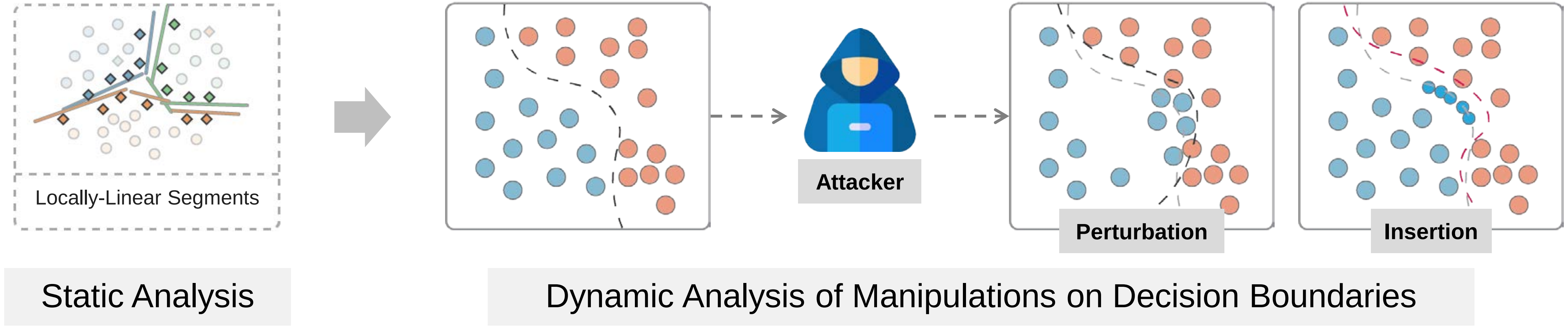
Yuxin Ma, Ross Maciejewski @ VADER Lab, CIDSE, Arizona State University

Acknowledgement U.S. Department of Homeland Security (Grant Award 2017-ST-061-QA0001 and 17STQAC00001-03-03)
National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939725

Demo Available at: https://github.com/wintericie/visual-analysis-class-boundary

Manipulating Decision Boundaries

- **Static Analysis -- Dynamic Analysis of (Malicious) Changes**
  - Comparing Decision Boundaries between Different Classifiers



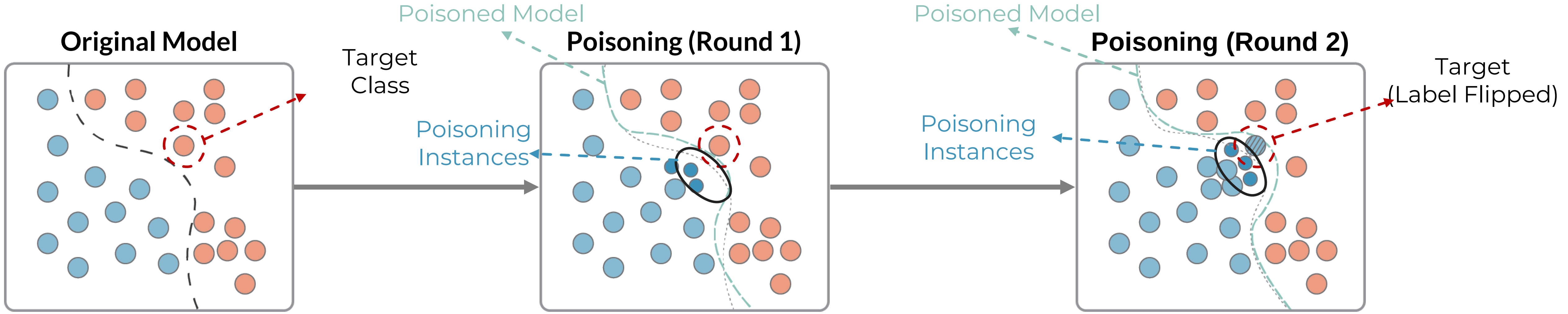
- **Manipulating Decision Boundaries of Classifiers in a Malicious Way**

  E.g. Poisoning Attack
  - Different decision boundaries (classifiers) when the training dataset is manipulated
  - Can be utilized by attackers to control the predictions from the classifiers on purpose

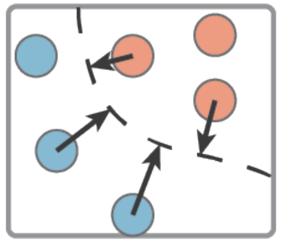
Vulnerability Analysis

Vulnerability Analysis	Vulnerability Measures for Training Instances

- **Core idea:** Prevent the target instance from being misclassified
  - **Attack Algorithms :** Binary-Search Attack & StingRay Attack



- **Vulnerability Measures**
  - **D**ecision **B**oundary **D**istances (**DBD**)
  - **M**inimum **C**ost for a **S**uccessful **A**ttack (**MCSA**)
  - Performance metrics of the poisoned model



Distance to the Decision Boundary

- **Attack 10 Times**
- **min(#insertion)**

Minimum #poison Among Multiple Attacks

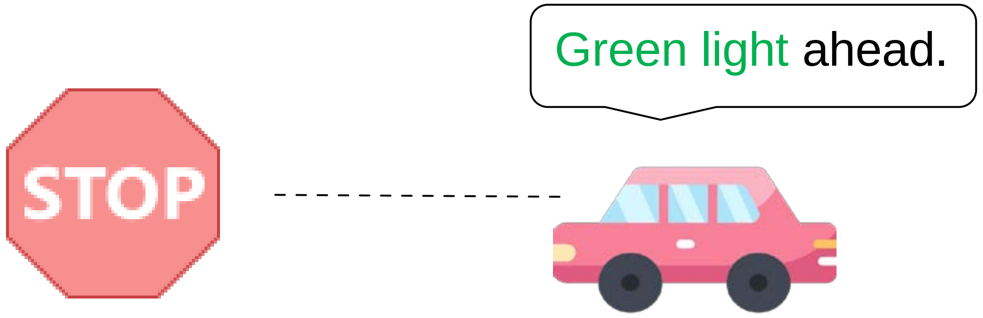
- **Training Acc.**
- **Training Recall**
- …

Model Performance Metrics (Accuracy, Recall, etc.)

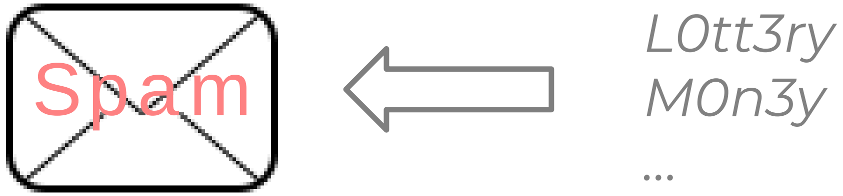
[1] Burkard et al. Analysis of causative attacks against svms learning from data streams. In Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics, pp. 31–36. ACM, 2017

[2] Suciu et al. When does machine learning fail? Generalized transferability for evasion and poisoning attacks. In Proceedings of the USENIX Security Symposium, pp.1299–1316, 2018.

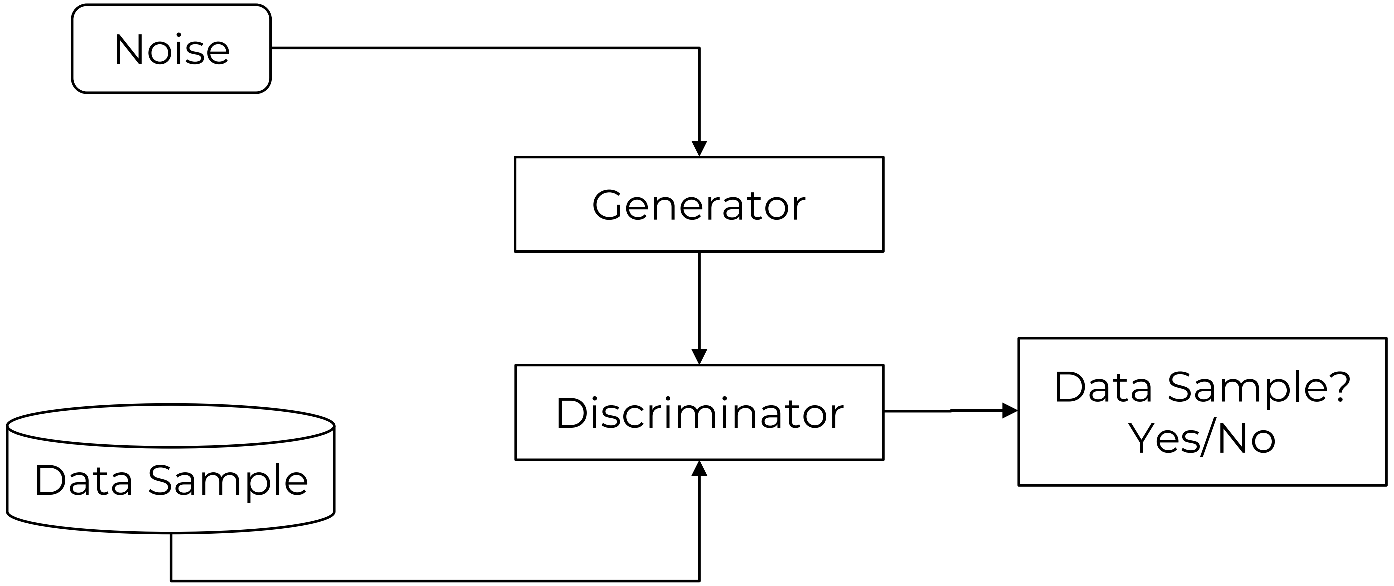
Vulnerabilities in Machine Learning



Self-driving Car



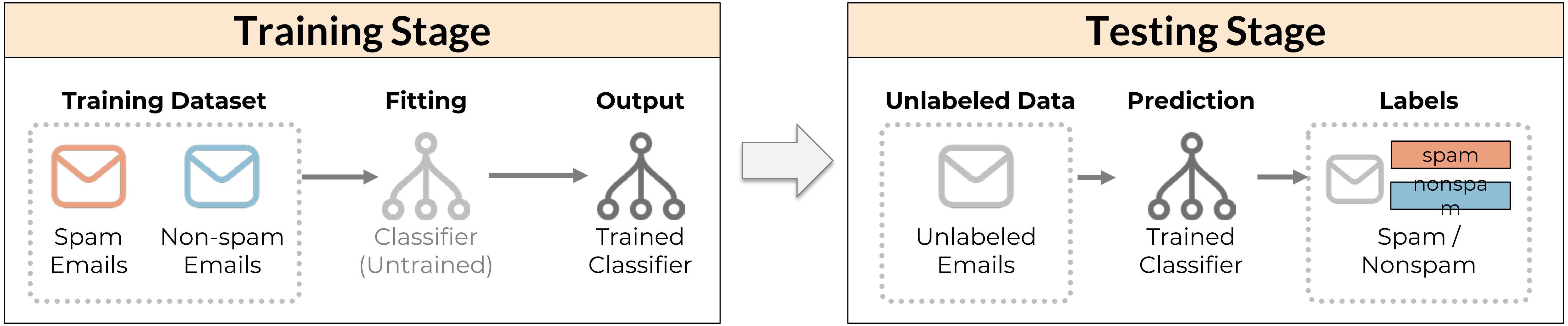
Spam FIlter



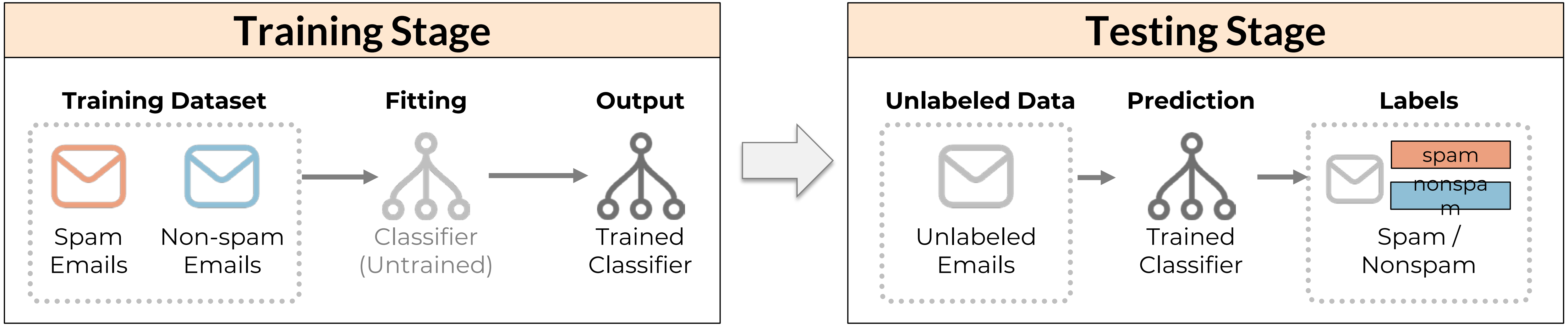
Generative Adversarial Nets (GAN)

[1] Martinez et al., Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. IEEE Transactions on Intelligent Transportation Systems,19(3):666–676, 2018.
[2] Mei et al., Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners. AAAI Conference on Artificial Intelligence, 2871–2877.
[3] Goodfellow et al., Generative Adversarial Nets. Advances in Neural Information Processing Systems. 2014.

Example: Filtering Spam Emails



Example: Filtering Spam Emails



Attacker

Goal	Knowledge
Make specific spams as nonspam ones	White-box setting (model, training data)

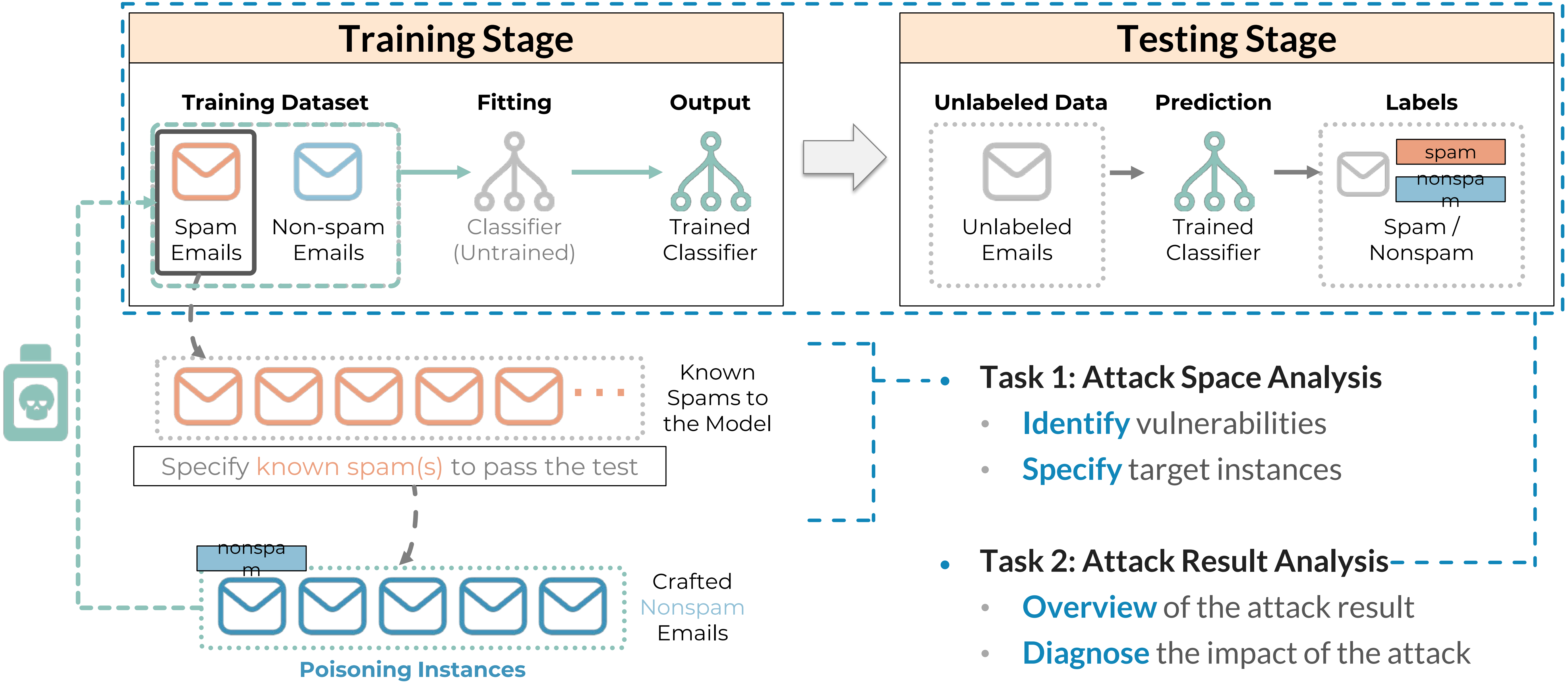
Capability	Strategy
Limited insertion to the training dataset	Find the minimum insertions for the goal

- **Explanation**
  - Identify potential vulnerabilities
  - Reveal attack processes
- **Diagnosis**
  - Inspect attack results
  - Explore different attack strategies

Visual Analytics

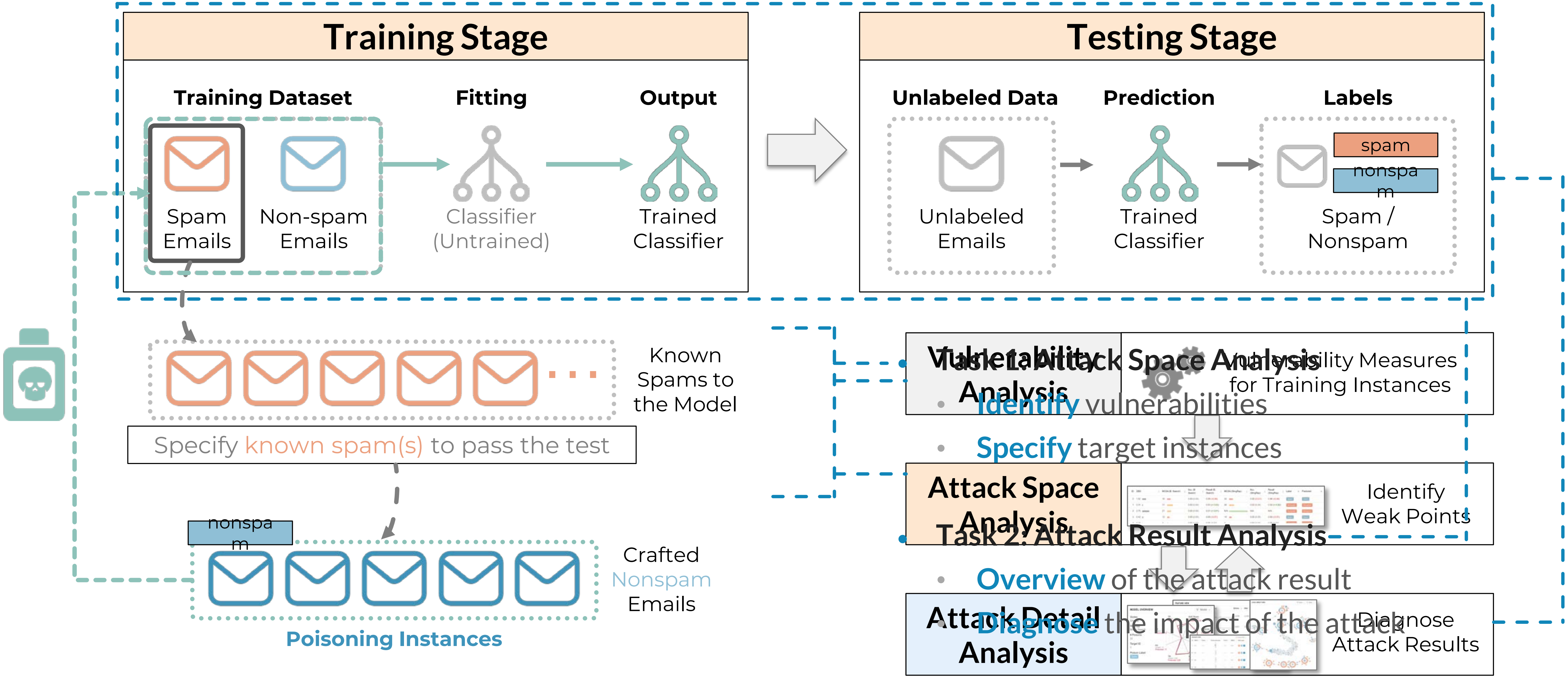
Targeted Poisoning Attack

Analytical Tasks



- **Task 1: Attack Space Analysis**
  - **Identify** vulnerabilities
  - **Specify** target instances
- **Task 2: Attack Result Analysis**
  - **Overview** of the attack result
  - **Diagnose** the impact of the attack

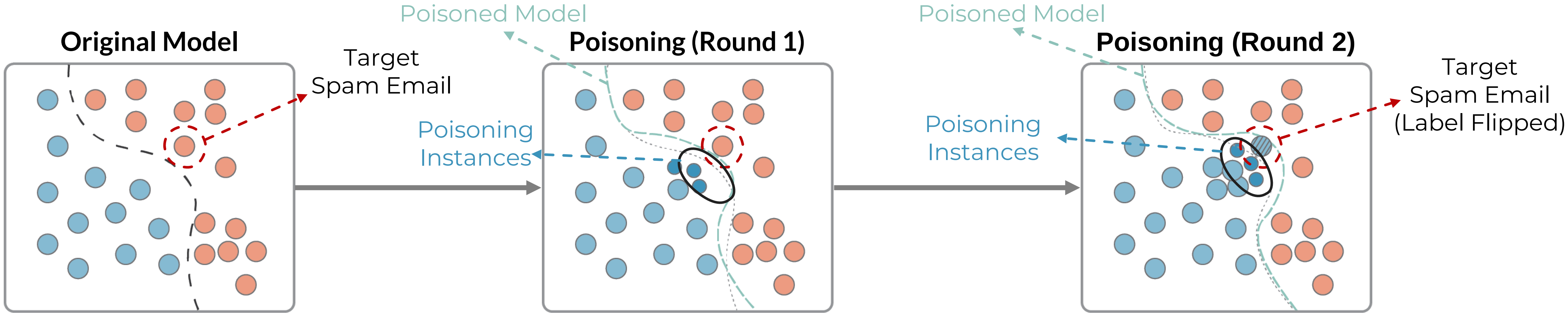
Framework Overview



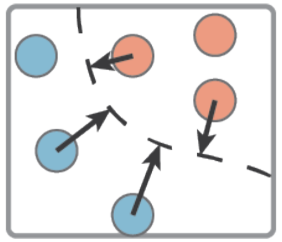
Vulnerability Analysis

Vulnerability Analysis	Vulnerability Measures for Training Instances

- **Core idea:** Prevent the target instance from being classified as Spam
  - **Attack Algorithms :** Binary-Search Attack & StingRay Attack



- **Vulnerability Measures**
  - **D**ecision **B**oundary **D**istances (**DBD**)
  - **M**inimum **C**ost for a **S**uccessful **A**ttack (**MCSA**)
  - Performance metrics of the poisoned model



Distance to the Decision Boundary

- **Attack 10 Times**
- **min(#insertion)**

Minimum #poison Among Multiple Attacks

- **Training Acc.**
- **Training Recall**
- **…**

Model Performance Metrics (Accuracy, Recall, etc.)

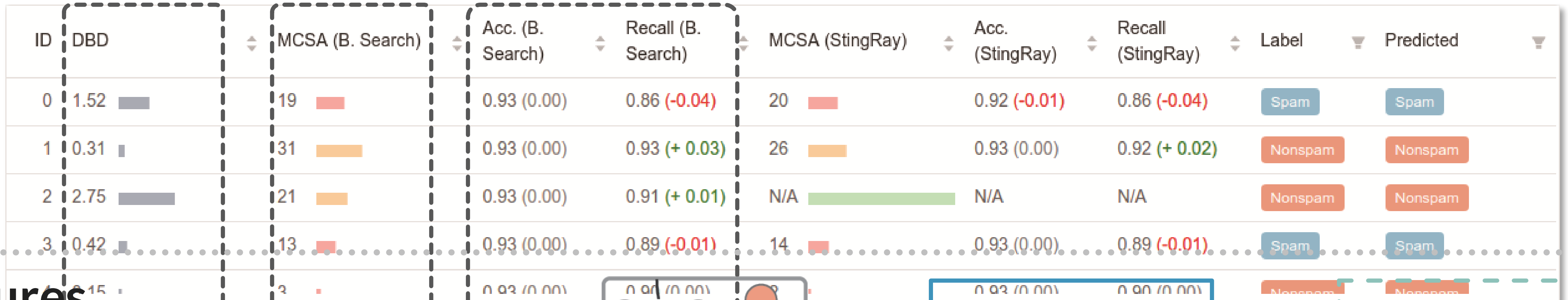
[1] Burkard et al. Analysis of causative attacks against svms learning from data streams. In Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics, pp. 31–36. ACM, 2017

[2] Suciu et al. When does machine learning fail? Generalized transferability for evasion and poisoning attacks. In Proceedings of the USENIX Security Symposium, pp.1299–1316, 2018.

Attack Space Analysis

Attack Space Analysis |  Identify Weak Points

Data Table View



ID	DBD	MCSA (B. Search)	Acc. (B. Search)	Recall (B. Search)	MCSA (StingRay)	Acc. (StingRay)	Recall (StingRay)	Label	Predicted
0	1.52	19	0.93 (0.00)	0.86 (-0.04)	20	0.92 (-0.01)	0.86 (-0.04)	Spam	Spam
1	0.31	31	0.93 (0.00)	0.93 (+ 0.03)	26	0.93 (0.00)	0.92 (+ 0.02)	Nonspam	Nonspam
2	2.75	21	0.93 (0.00)	0.91 (+ 0.01)	N/A	N/A	N/A	Nonspam	Nonspam
3	0.42	13	0.93 (0.00)	0.89 (-0.01)	14	0.93 (0.00)	0.89 (-0.01)	Spam	Spam
[illegible]	0.15	3	0.93 (0.00)	0.9[illegible] (0.00)	[illegible]	0.93 (0.00)	0.90 (0.00)	Nonspam	Nonspam

- **Vulnerability Measures**
  - Decision Boundary Distances (**DBD**)
  - Minimum Cost for a Successful Attack (**MCSA**)
  - Performance metrics of the poisoned model

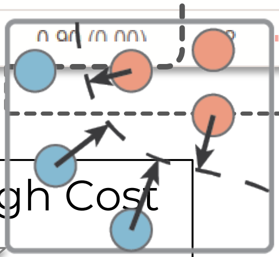
DBD: Low DBD

Easy to Flip

Distance to the Decision Boundary

MCSA: High Cost

Hard to Attack



- **Attack 10 Times**
- **min(#insertion)**

Minimum #poison Among Multiple Attacks

Performance: Metric Drop

Hard to Attack

- **Training Acc.**
- **Training Recall**

Model Performance Metrics (Accuracy, Recall, etc.)

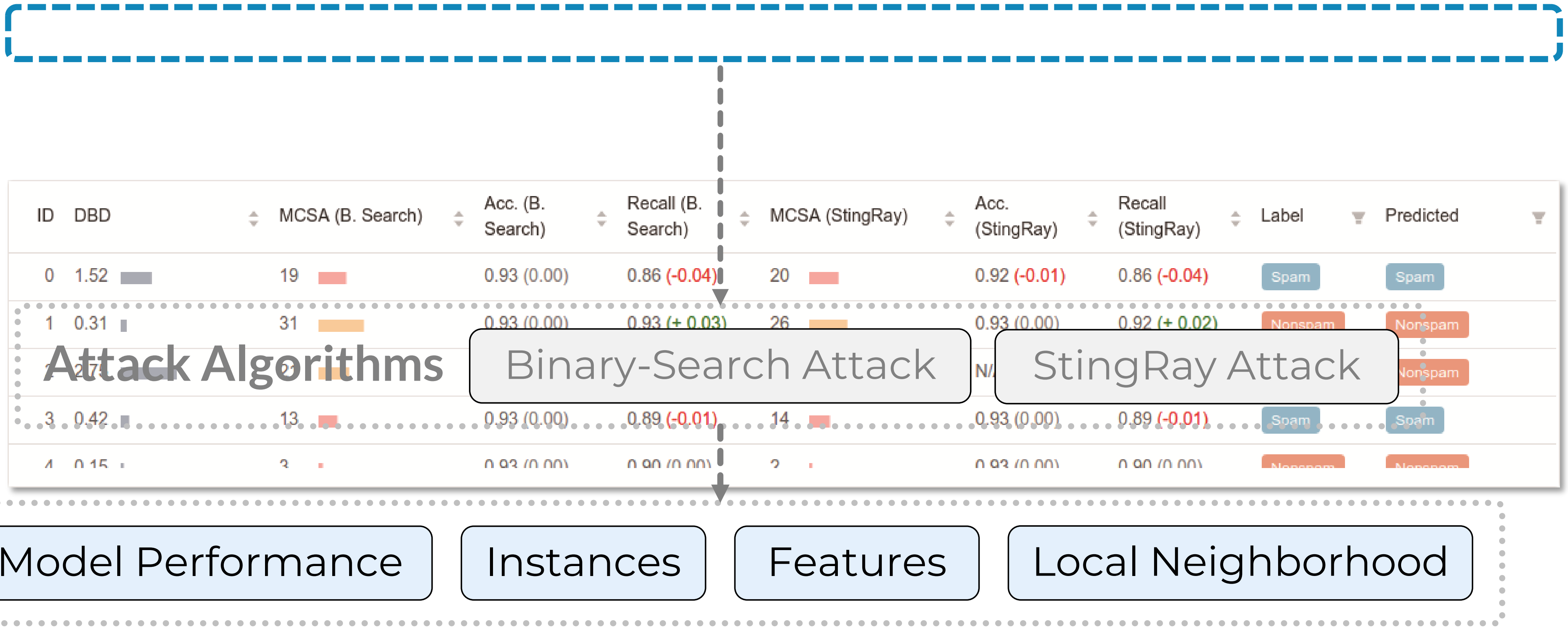
Attack Detail Analysis

Attack Detail Analysis

Diagnose Attack Results

Data Table View

Detailed Diagnosis

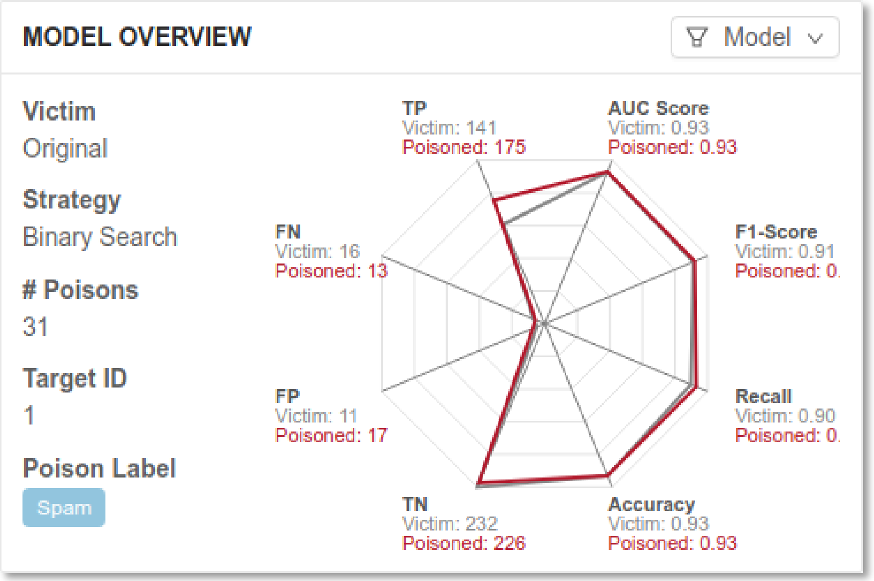


ID	DBD	MCSA (B. Search)	Acc. (B. Search)	Recall (B. Search)	MCSA (StingRay)	Acc. (StingRay)	Recall (StingRay)	Label	Predicted
0	1.52	19	0.93 (0.00)	0.86 (-0.04)	20	0.92 (-0.01)	0.86 (-0.04)	Spam	Spam
1	0.31	31	0.93 (0.00)	0.93 (+ 0.03)	26	0.93 (0.00)	0.92 (+ 0.02)	Nonspam	Nonspam
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	N/A	[illegible]	[illegible]	Nonspam
3	0.42	13	0.93 (0.00)	0.89 (-0.01)	14	0.93 (0.00)	0.89 (-0.01)	Spam	Spam
4	0.15	3	0.93 (0.00)	0.90 (0.00)	2	0.93 (0.00)	0.90 (0.00)	Nonspam	Nonspam

Attack Detail Analysis

Model Performance | Instances | Features | Local Neighborhood

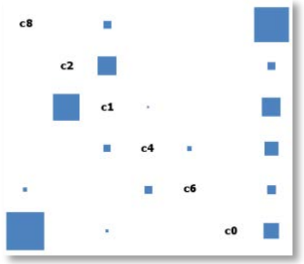
Model Overview



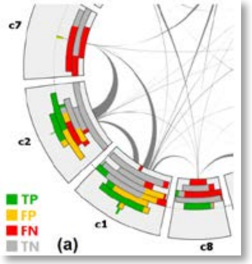
Radar Chart for Model Performances

- **Comparison**
  - Victim ⟷ Poisoned
  - Performance metrics

(Alternatives)



Confusion Matrix



Confusion Wheel

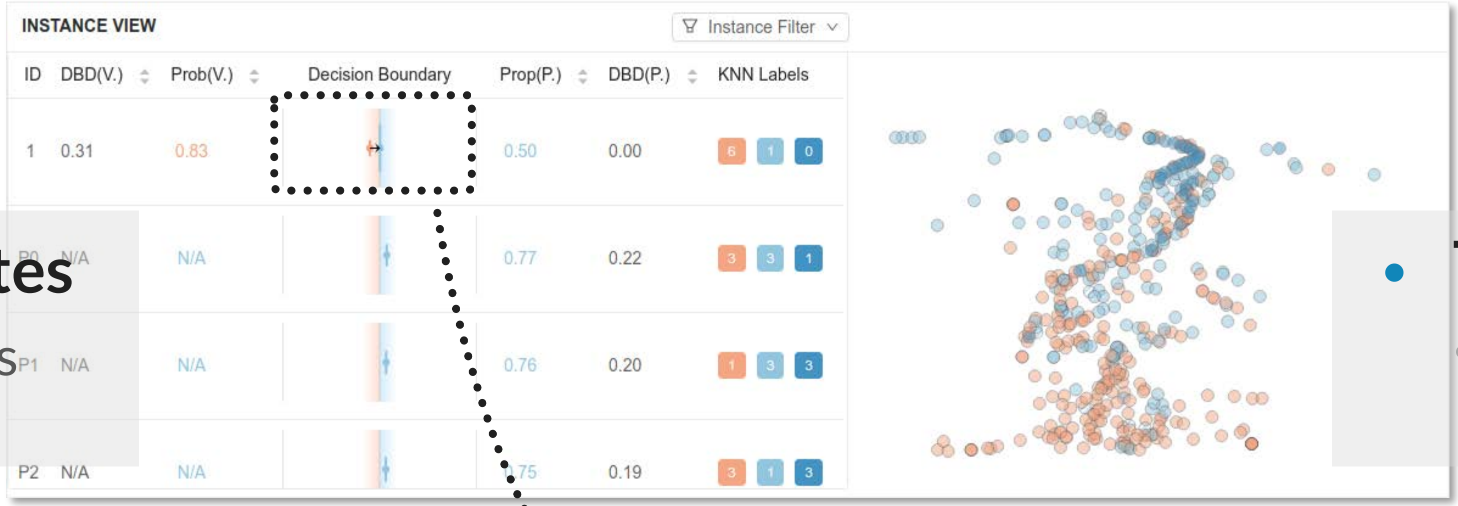
- **Design Rationale**
  - Suitable for comparing differences
  - Easy to use

[1] Alsallakh et al. (2012). Reinventing the contingency wheel: Scalable visual analytics of large categorical data. IEEE TVCG.
[2] Alsallakh et al. (2014). Visual Methods for Analyzing Probabilistic Classification Data. IEEE TVCG.

Attack Detail Analysis

Model Performance | Instances | Features | Local Neighborhood

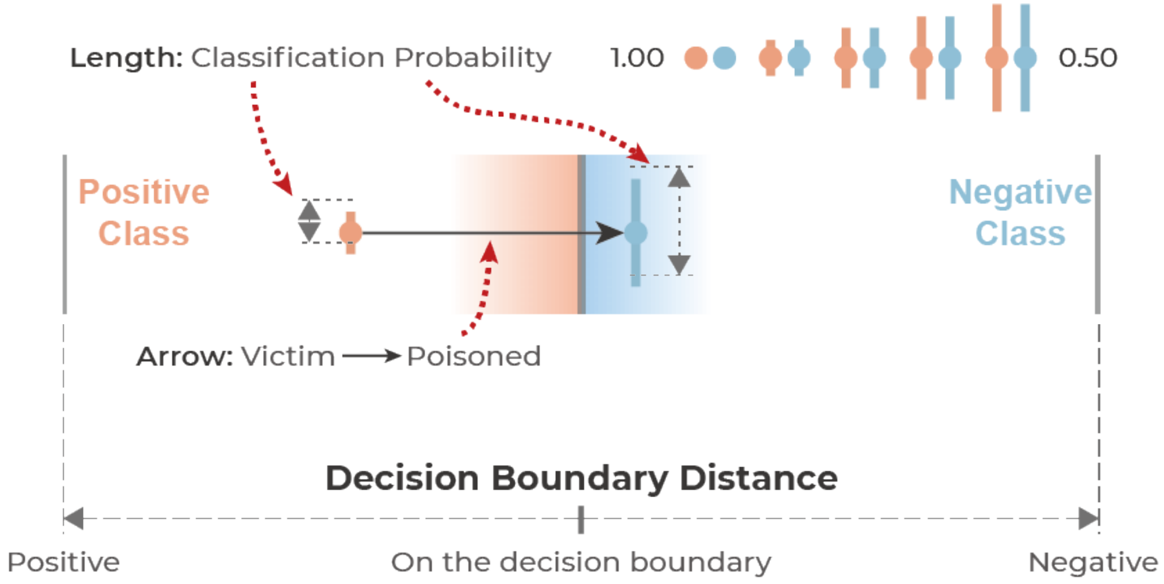
Instance View



- Instances Attributes
  - Details of instances

- T-SNE Projection
  - Overview of the data distribution

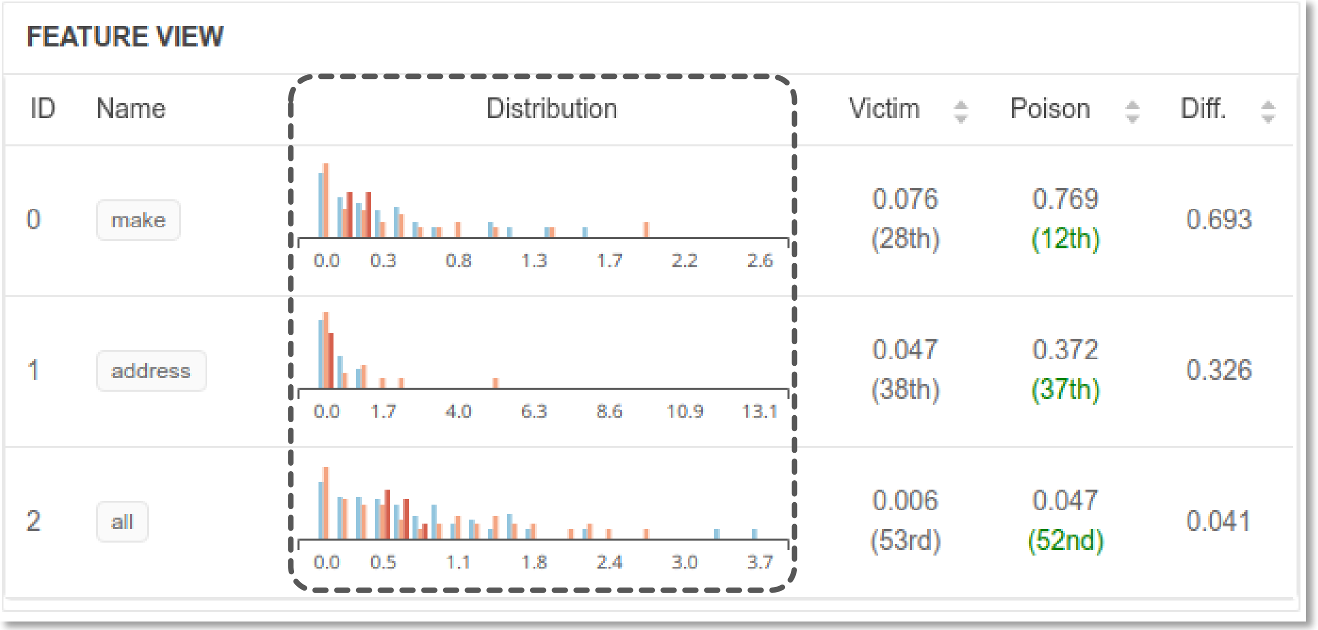
- Key Attributes for Instances
  - Decision Boundary Distances
  - Classification Probabilities
  - Labels of k-NNs



Attack Detail Analysis

Model Performance | Instances | Features | Local Neighborhood

Feature View


FEATURE VIEW

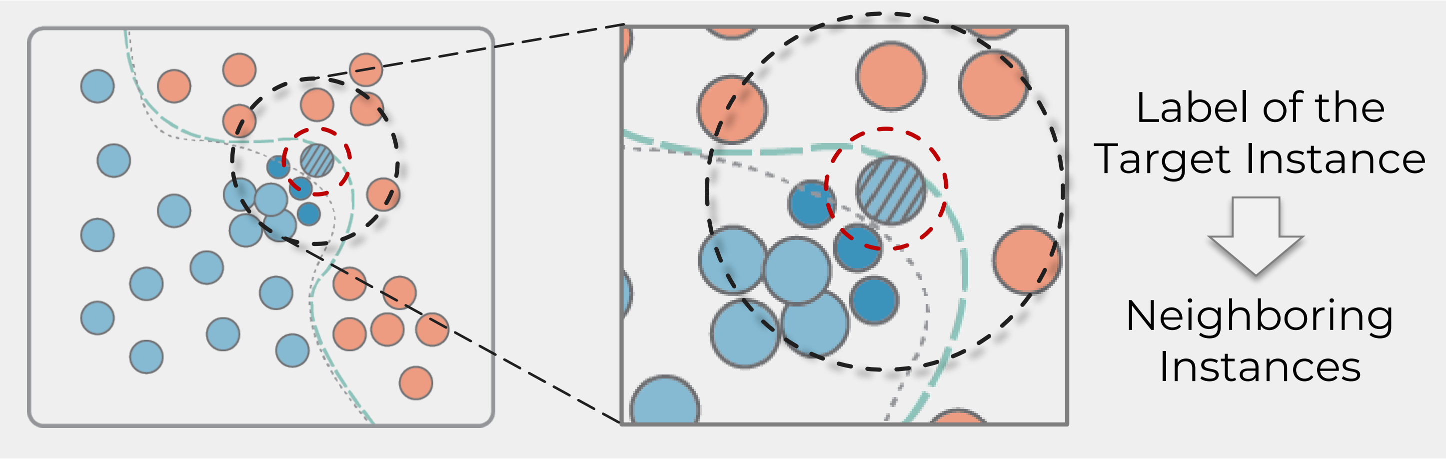
ID	Name	Distribution	Victim	Poison	Diff.
0	make	0.0 0.3 0.8 1.3 1.7 2.2 2.6	0.076 (28th)	0.769 (12th)	0.693
1	address	0.0 1.7 4.0 6.3 8.6 10.9 13.1	0.047 (38th)	0.372 (37th)	0.326
2	all	0.0 0.5 1.1 1.8 2.4 3.0 3.7	0.006 (53rd)	0.047 (52nd)	0.041

- **Data Distributions on Features**
  - Instances in the spam / nonspam classes
  - Poisoning Instances
- **Feature Importance Rankings**
  - In the victim model
  - In the poisoned model
  - Differences

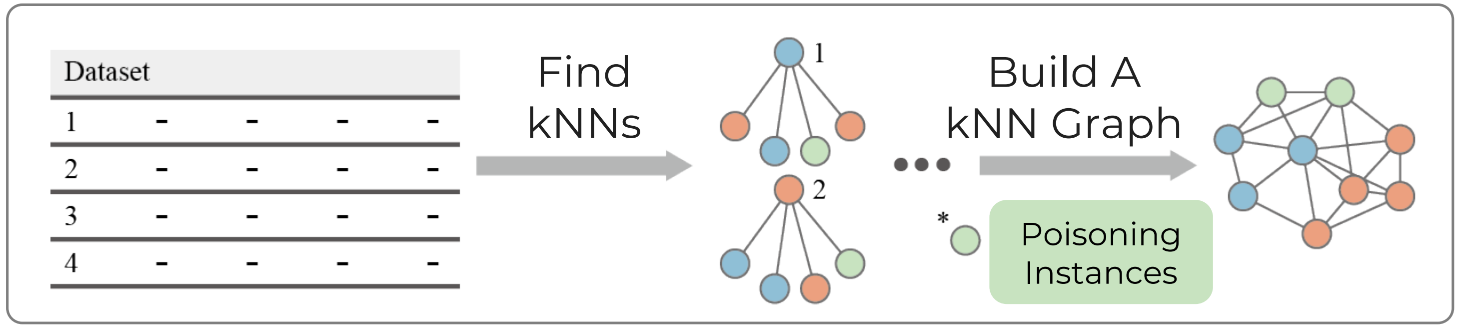
Attack Detail Analysis

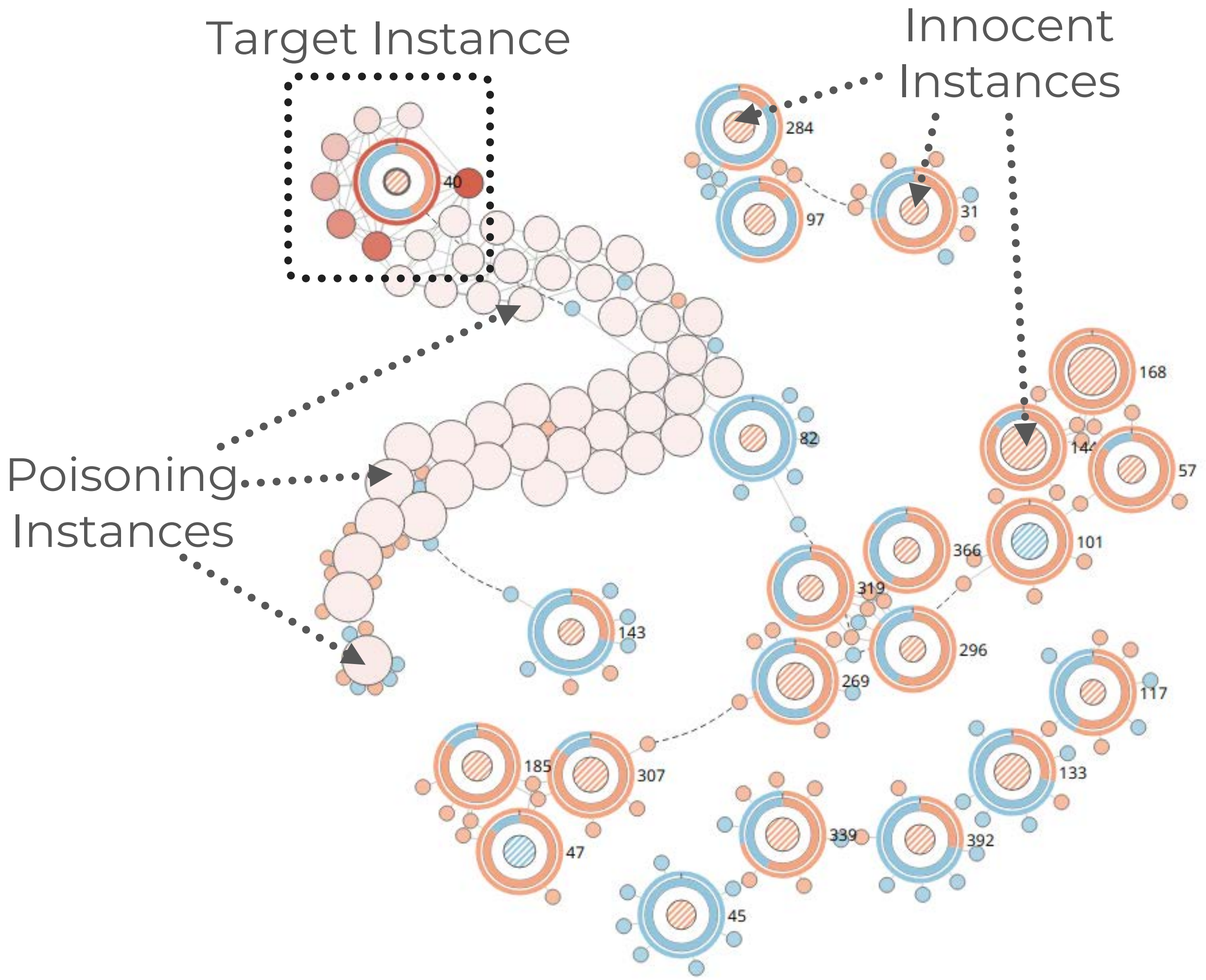
Model Performance | Instances | Features | **Local Neighborhood**

Local Impact View



kNN Graph Building





Case Study

- **Spambase Email Data**
  - Task: Binary Classification
  - 57 Dimensions
    - BoW vectors
  - Subsampled 400 Emails
    - 243 non-spam emails
    - 157 spam emails

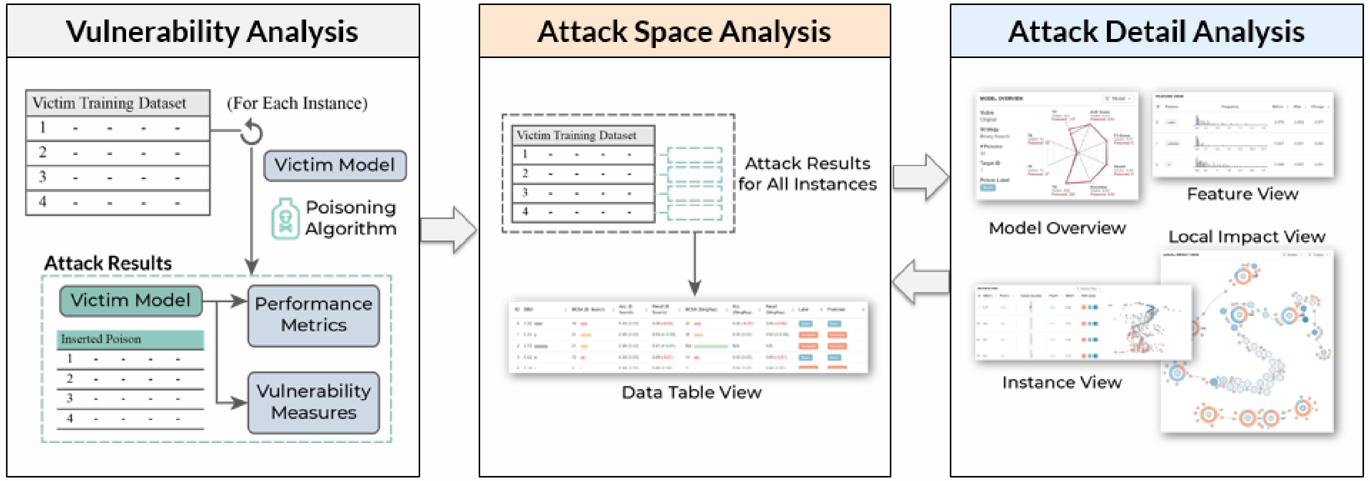
Explaining Vulnerabilities to Adversarial Machine Learning through Visual Analytics

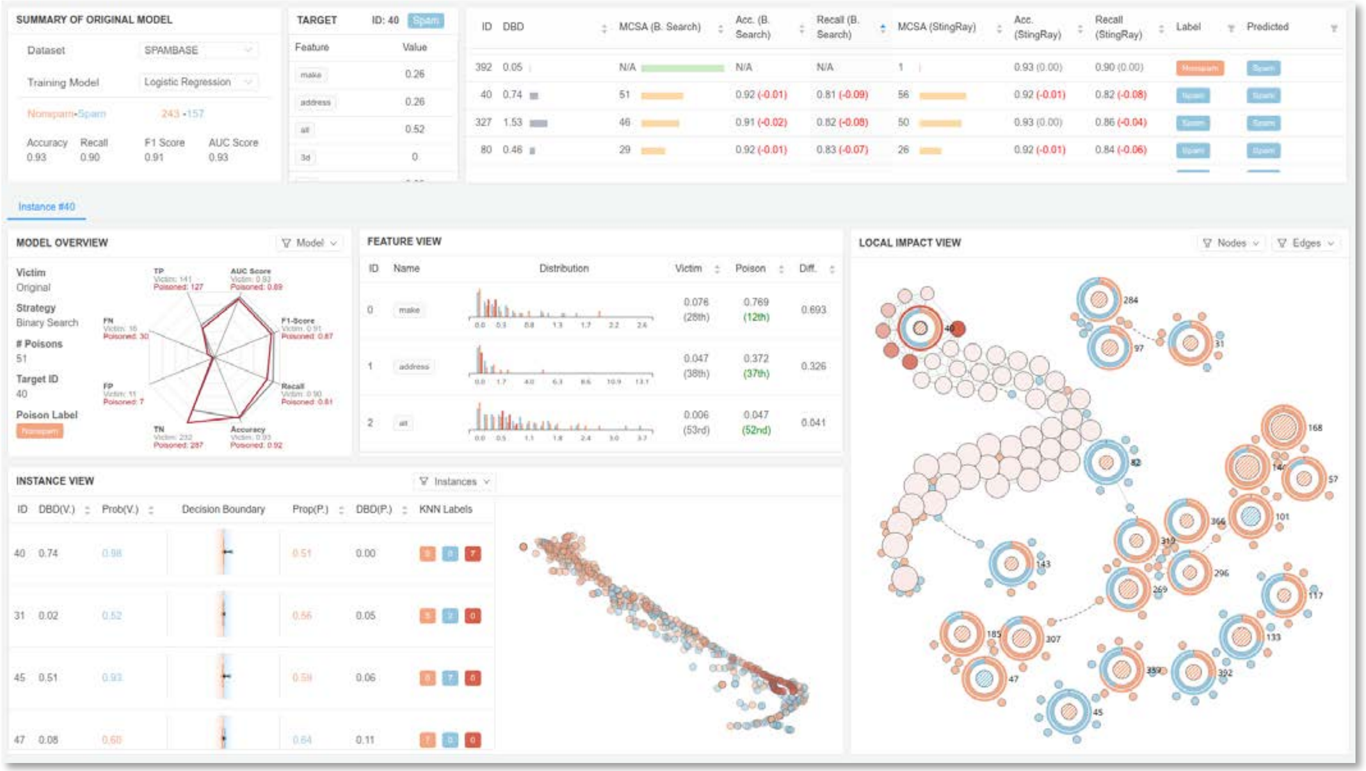
Yuxin Ma, Tiankai Xie, Jundong Li,
Ross Maciejewski
VADER Lab, CIDSE, Arizona State University

- **Code Available at:** https://github.com/VADERASU/visual-analytics-adversarial-attacks
- **Project Website:** http://vader.lab.asu.edu

Acknowledgement

U.S. Department of Homeland Security (Grant Award 2017-ST-061-QA0001)



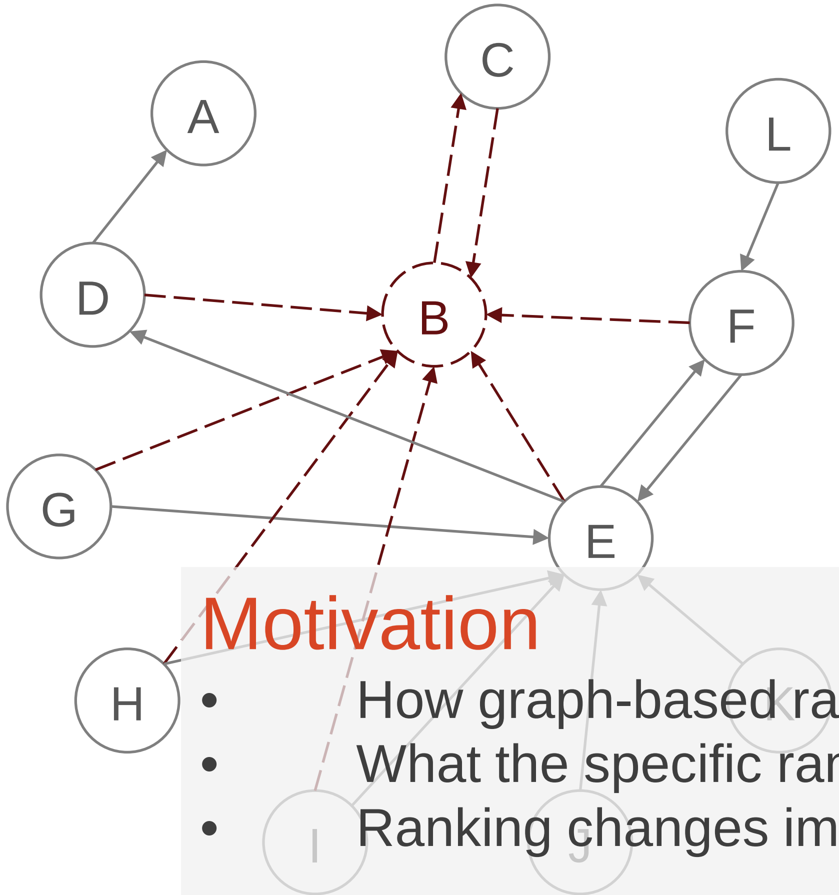


Graph-based Ranking

Search Engine Ranking **Recommendation System**

[1] Page, Lawrence, et al. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999.
[2] Gori, Marco, et al. "Itemrank: A random-walk based scoring algorithm for recommender engines." *IJCAI*. Vol. 7. 2007.

Remove a Node



PageRank

Rank	Webpage	PR value
-	B	removed
4⬇	C	0.032⬇
1⬆	E	0.308⬆
2⬆	D	0.163⬆
3⬆	F	0.163⬆
5⬆	G	0.032⬆
	…	…

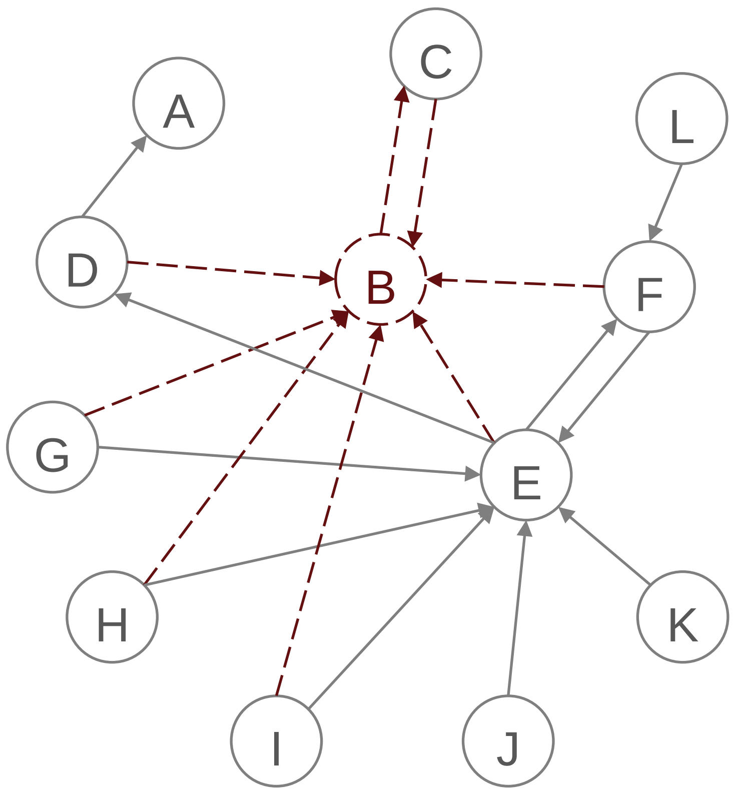
Motivation

- How graph-based ranking methods are sensitive to the perturbation of graph elements?
- What the specific ranking changes would a certain removed node cause?
- Ranking changes impact the exposure in real-world applications

[1] "Pagerank". En.Wikipedia.Org, 2020, https://en.wikipedia.org/wiki/PageRank. Accessed 17 Aug 2020.
[2] Singh, Ashudeep, and Thorsten Joachims. "Fairness of Exposure in Rankings." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018):

Sensitivity Index

- The degree of the ranking method's sensitivity to the perturbation (removal)
- Given any graph-ranking method



$$SI_B = L\,(r\,,\,r')$$

Distance metric (L1 norm)

Rank	Webpage
1	B
2	C
3	E
4	D
5	F
6	G
…	…

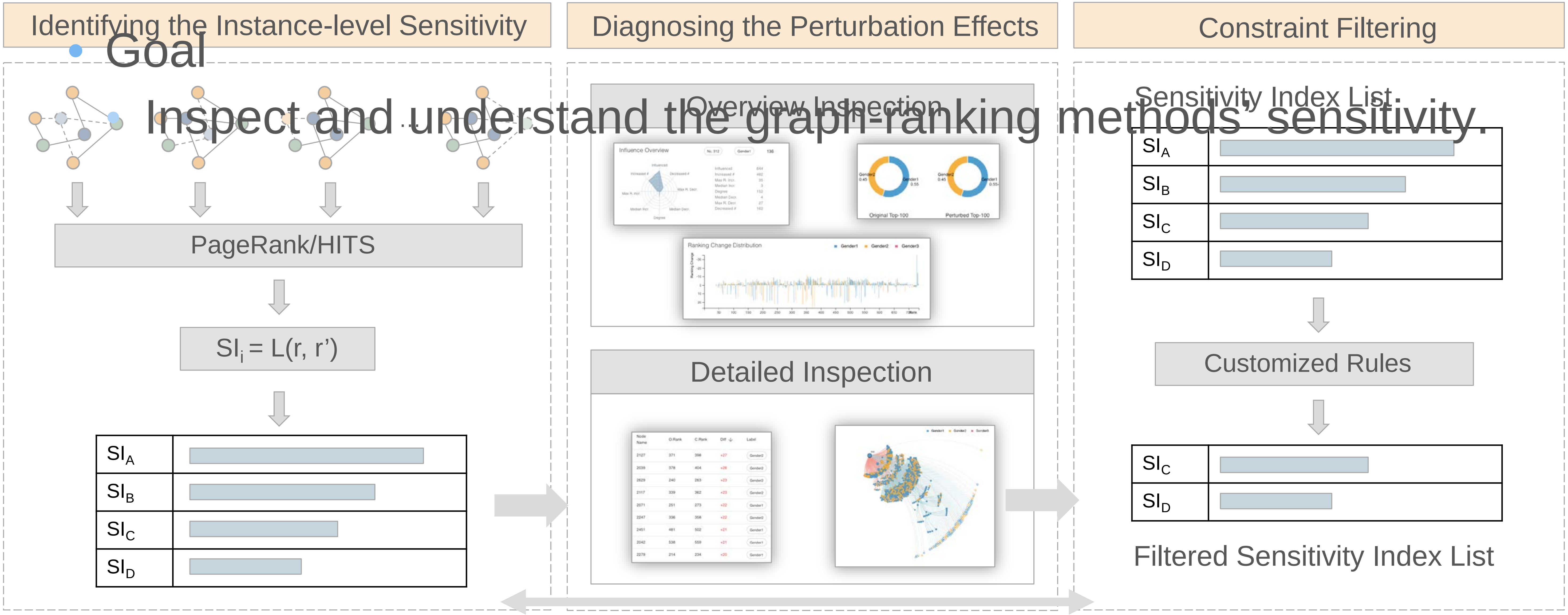
Ranking Positions (original)

Rank	Webpage
-	B
4	C
1	E
2	D
3	F
5	G
…	…

Ranking Positions (perturbated)

Visual Analytics Framework

- Goal
  - Inspect and understand the graph-ranking methods' sensitivity.



Auditing the Sensitivity of Graph-based Ranking with Visual Analytics

Tiankai Xie[1], Yuxin Ma[1], Hanghang Tong[2], My T. Thai[3], Ross Maciejewski[1]

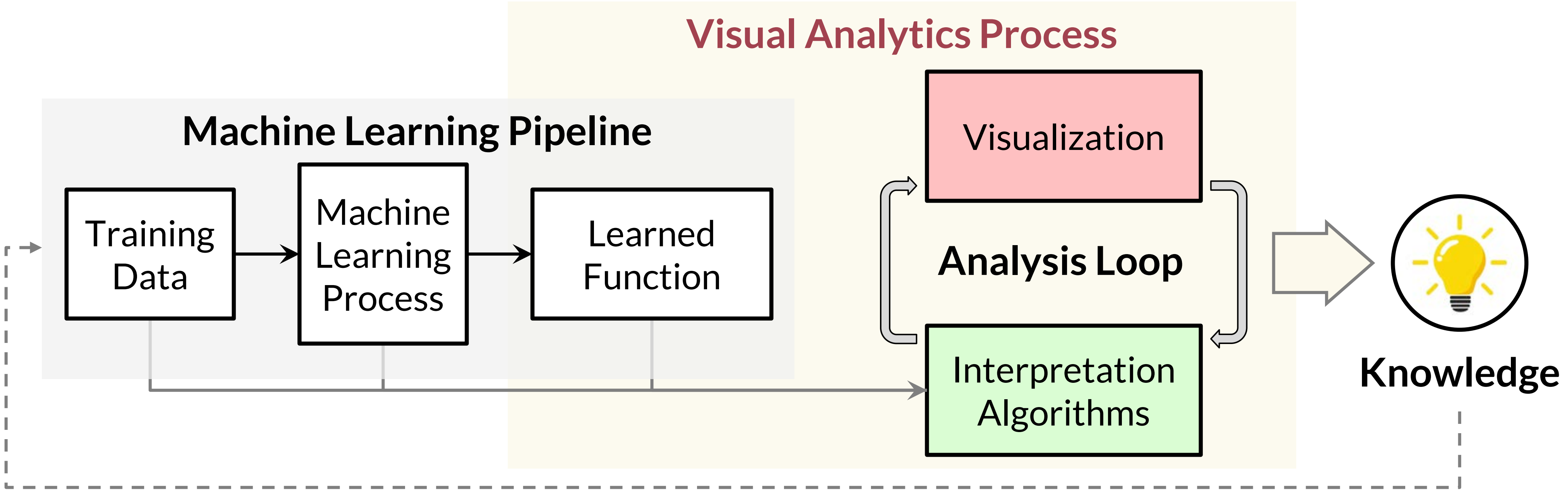
1. VADER Lab, CIDSE, Arizona State University
2. University of Illinois at Urbana-Champaign
3. University of Florida

- **Code Available at:** https://github.com/VADERASU/auditing-sensitivity-graph-ranking
- **Project Website:** http://vader.lab.asu.edu

Acknowledgement

U.S. Department of Homeland Security under Grant Award 2017-ST-061-QA0001 and 17STQAC00001-03-03, and the National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939725

Visual Analytics in Explainable AI



- **Explainable AI (XAI)** in the VADER Lab @ ASU
  - **Data Preprocessing**: Visual Inspection of Decision Boundaries (**TVCG 2020**)
  - **Interpretable Model Training**: Open-box Exploration of SVMs (**CVMJ 2017**)
  - **Security**: Visual Explanation of Adversarial Machine Learning (**VAST 2019**), Graph Auditing (**VAST 2020**)
  - **Reusability**: Visual Analysis of Transfer Learning Processes (**VAST 2020**)