Towards an Explainable Threat Detection Tool

Alison Smith-Renner
Decisive Analytics Corporation
Arlington, VA, USA
alison.smith@dac.us

Rob Rua
Decisive Analytics Corporation
Arlington, VA, USA
rob.rua@dac.us

Mike Colony
Decisive Analytics Corporation
Arlington, VA, USA
mike.colony@dac.us

ABSTRACT

In general, threats can be loosely divided into two categories – known threats and unknown threats. Traditional threat detection systems are limited to the identification of known threats that have been previously encountered and labeled by a security expert. These supervised learning systems are able to learn to detect and identify known threats but are unable to react to unknown threats. To this end, we have developed an unsupervised learning anomaly detection system to identify anomalous behavior without training data. Our system's interactive interface supports human-machine teaming to classify these identified anomalies as threats or benign events; however, system transparency is required to enhance operator trust and improve their feedback into the system. Transparency in this case is particularly challenging as our anomaly detection framework is based on algorithms which are inherently hard to explain (neural networks). In this paper, we introduce a real-world task and system that requires transparency, and we propose explanation methods for increasing the transparency of our threat detection tool alongside a user study for evaluating these explanations.

CCS CONCEPTS

• **Human-centered computing~Human computer interaction (HCI)** • Human-centered computing~HCI design and evaluation methods • Human-centered computing~Interactive systems and tools

KEYWORDS

Anomaly detection; explanations; transparency; human-machine teaming

ACM Reference format:

Alison Smith-Renner, Rob Rua, and Mike Colony. 2019. Towards an Explainable Threat Detection Tool. In *Joint Proceedings of ACM IUI 2019 Workshops. ACM, Los Angeles, USA, March 20, 2019, 6 pages.*

1 Introduction

Forward deployed military installations face unique challenges when automating threat detection in security monitoring systems. In particular, development of a general framework for identifying emerging security threats poses two technical obstacles: (1) the security framework must be robust to an environment where "normal" activities are initially unknown and (2) the framework must support collaboration with operators to determine what types of anomalous behavior constitute an actual threat. To this end, we have developed an unsupervised anomaly detection system, DAART, (for Detection of Anomalous Activity in Real Time), to identify anomalous behavior without training data. DAART's interactive interface supports human-machine teaming to classify these identified anomalies as threats or benign events.

Traditional threat detection systems are limited to the identification of known threats that have been previously encountered and labeled by a security expert, such as a man wielding a gun or an intruder in a restricted location. The unsupervised nature of the DAART system additionally supports identification and action on unknown threats, which is necessary to adapt to ever changing environments. A by-product of this, however, is an initial trend towards recall over precision, meaning many benign activities may be alerted to the user. DAART's Active Learning component learns from operator feedback in the form of accepting or rejecting alerts (alert-level feedback) to better distinguish benign anomalous behavior from threats. A human-in-the-loop system, such as this, requires system transparency to improve operator trust, accelerate operator workflow, and better enable operators to provide the valuable feedback required to improve the system's threat classifications.

A threat detection system may err in two distinct ways: (1) false positives in which benign behavior is predicted to be a threat and (2) false negatives in which a threat is considered benign (and therefore not alerted to the user). Operator trust is negatively affected if a system produces many false positives without explanation or if the operator cannot confirm whether the system produces false negatives. System transparency in the form of alert-level and system-level explanations therefore enhances trust, because users can better understand when and why a system makes mistakes as well as to ensure the system doesn't miss any potential threat behavior, respectively [16].

Not all anomalies are threats and not all threats are equally important. System transparency accelerates operator workflow by providing the evidence needed to quickly and accurately prioritize and determine the validity of threats. Finally, the DAART system improves with user feedback, so the goal is to get the best feedback as possible from users while minimizing the time and effort to

IUI Workshops'19, March 20, 2019, Los Angeles, USA

provide it. System transparency enhances the feedback process because users' feedback is improved when they have an understanding of how the system works and why an alert is considered anomalous. Furthermore, users' time and effort are minimized when providing feedback through the same visualizations presented to them for explanation purposes, as these are already familiar [13].

In this paper we present the DAART system for identifying anomalous behavior without training data, and we propose interactive explanation methods for improved operator trust, accelerated workflow, and enhanced operator feedback through system transparency. In particular, we propose methods for determining and displaying explanation information, such as multi-modal localization (or attention) and normalcy exemplar, and an interactive explanation interface to present these and other simple explanation types (system confidence, alternate classifications, features) to users for promoting transparency and providing a means for user feedback. We additionally propose a user study to evaluate these various interactive explanation methods for the DAART system.

2 Background

2.1 Anomaly Detection

Detecting anomalies in sensor data requires a standardized feature representation of the incoming data. Traditionally, these features are defined by expert scientists who specialize in particular sensor modalities. More recently, supervised machine learning models have been able to outperform expert-defined features in their descriptiveness about the original sensor data [12]. DAART improves on this, leveraging state-of-the-art research in unsupervised convolutional feature learning [1,6] to generate comparably discriminative features without the need for human-labeled training data. While these extracted features are not as easily understood by a human as expert-specified features, they have more expressive power when used for tasks such as anomaly detection. Importantly, this approach is sensor agnostic, meaning it can be applied to any sensor data, including but not limited to, EO and IR video, audio, and acoustic sensors.

2.2 User Feedback

Interactive machine learning systems incorporate end-user feedback to re-train underlying algorithms and improve their output. Users may provide this feedback in the form of interactively labeling data [21], as part of an interactive training phase [7,8], to fix specific system mistakes [22], or to inject their domain expertise into the system [11].

The system we present here builds on interactive machine learning techniques, such as accepting and rejecting system's output [25] and interactive clustering [15] for improved threat classification. We additionally propose to enhance the system with support for richer user feedback, such as modifying feature weights [18].

2.3 System Transparency

There is growing interest in system transparency, or explainable artificial intelligence (XAI), driven in part by both DARPA's XAI initiative [10] and the European Union's data protection law for "right to explanation" [24]. We aim for transparency in our anomaly detection system as it supports operator decision making [23], improves trust [14,16], and aids users in better providing feedback to the system [13,20] as well as motivating them to do so [19].

System transparency can be provided through explanations or visualizations that provide insight into what the system is doing and why it is doing it (see [3] for a survey). In particular, prior work has identified explanation types [13,17] to improve end user understanding of complex systems. We propose to implement and evaluate these explanation types in the DAART system.

3 DAART

Figure 1 shows the DAART system overview. The DAART system ingests multi-modal sensor data (audio, video, radar, etc.), which is converted to discriminative features for use in anomaly detection. The anomaly detection component utilizes the feature data to continually learn normalcy baselines against which it performs anomaly detection in real time. The user governs the anomaly detection process through the creation of Scopes, which define specific parameterizations (or filters) on the data. Detected anomalies are provided to the user via an interactive threat classification component which leverages user feedback to learn to classify anomalies as threats. We describe these components in more detail in the following sections.

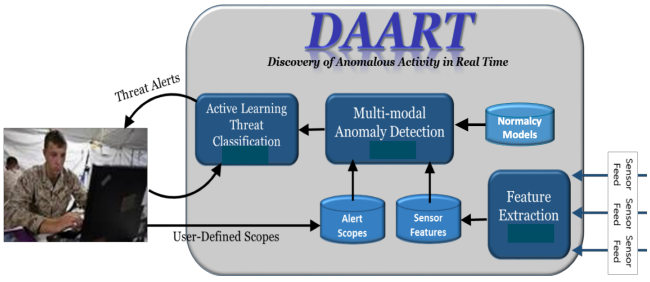

Figure 1: DAART system overview. Data enters the system from various sensor feeds, and sensor features are extracted. These features feed the multi-modal anomaly detection component along with user-defined scopes. A normalcy model is trained to represent normal behavior to which new, possibly anomalous behavior is compared. Identified anomalies are alerted to the user through the active learning threat classification component, which supports users in vetting or rejecting anomalous alerts as threats as well as specifying the threat's class.

3.1 Discriminative Features and Anomaly Detection

We generate discriminative features and perform anomaly detection using an approach based on Generative Adversarial Networks (GANs) [9].

Generative Adversarial Networks (GANs) have demonstrated the ability to generate images from a random noise vector that are able to fool another model attempting to determine if the image is real or fake. A GAN consists of two competing models. A Generator (G) model learns how to transform random noise into a fake image. A Discriminator (D) model then tries to determine if the fake image is real or fake. Over time both models are trained until the fake images are indistinguishable from the real images. An Adversarial Learned Inference (ALI) [5] model is an adaptation of a GAN that can be exploited for anomaly detection. The ALI model modifies a standard GAN by adding an Encoder (E) that simultaneously learns to generate a latent input vector that will allow the Generator (G) model to fool the Discriminator (D) model. To generate discriminative features in DAART, we use an approach based on GANomaly [1], an extension of ALI.[1]

GANomaly extends ALI to perform anomaly detection as part of the feature extraction process. This approach consists of three sub-networks: A Generator (G), Encoder (E), and Discriminator (D). During the training process, the GANomaly model is trained only on normal images, in essence learning a model of normalcy. The images (video frames, or vectors from other sensor types) are fed into the Generator, which learns two things: (1) a lower dimensional mapping (z), and (2) to reconstruct the image (x′). The reconstructed image is then fed into both the Encoder and Discriminator. The Encoder learns a second lower dimensional mapping of the reconstructed image (z′), and the Discriminator learns to tell the difference between real and fake images. During the training process, the Generator learns by minimizing the loss between the original image and the fake image (x - x′). The Encoder learns by minimizing the loss between the first and second lower dimensional spaces (z - z′).

Once trained, GANomaly is used in live operations to test each input vector (e.g. video frame) to compute an anomaly score. Since the model was trained only on normal behavior, an anomaly score that determines whether an input is anomalous or not can be computed based on the L1-normalized Euclidean difference between the first lower dimensional mapping learnt by the encoder and the lower dimensional mapping learnt from the reconstructed image (z - z′).

3.2 Scopes and Normalcy Model

Scopes define specific filters on sensor and facility conditions that the user wants the anomaly detection system to be restricted to when discovering anomalies. When a Scope is specified, all data that matches the defined filters will be processed by the anomaly detection algorithm, which calculates an n-dimensional probability distribution of the data over the features learned during feature extraction. The result is a baseline normalcy model defining what sensor data is considered normal activity.

This baseline normalcy model is incrementally updated each time new sensor data is ingested into the DAART system. We use this normalcy model to compute a strangeness metric proposed in prior work [2]: incoming sensor data is compared against the baseline model to determine how strange the observation is compared to normal behavior. Once this strangeness metric has been calculated for each individual sensor modality, the metrics are merged across all modalities to determine whether an incident observed by multiple sensors is anomalous.

3.3 Interactive Threat Classification

When the DAART system identifies anomalous activity, it alerts the user. An example of the DAART system upon identifying anomalous activity is shown in Figure 2.

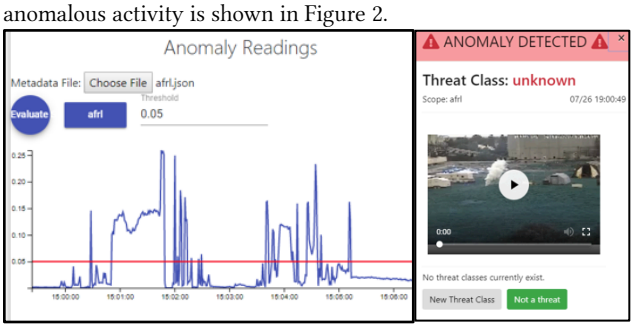

Figure 2: The DAART system's interactive threat classification involves alerting anomalous behavior to the security operator who can then reject the alert or select an appropriate threat class.

Figure 2 (right) shows the alert, which includes a video clip of the anomalous activity and a timestamp at which the activity occurs. Users interact with an alert to either specify that it is "Not a threat" or provide a threat class for it. In addition to viewing the clip, the system also explains the anomaly using a timeseries chart showing the strangeness score of the anomaly compared to prior readings as shown by Figure 2 (left). Users can additionally modify the strangeness threshold above which an anomaly is alerted. We propose additional methods for explanation and user feedback in the following section.

4 Explanation Methods

4.1 Localization (Attention)

For operators to better understand the anomalies and threats that DAART alerts them to, it would be ideal to be able to isolate which part of the sensor reading was anomalous. In the case of EO video, this could mean showing the user a bounding box which identifies where in the video stream the anomalous activity is occurring. This type of functionality is extremely valuable in helping the operator decide what threat label to assign to new unknown threats, and to help them better determine what course of action is reasonable in response to a threat.

Because of the fully unsupervised GAN-based approach DAART uses for anomaly detection, localization of the anomalous activity in sensor readings is non-trivial. Unlike many supervised approaches, in which the detection of specific objects or actions are

[1] We performed a qualitative comparison of GANomaly and ALI and found that GANomaly demonstrated super results and stability in the training phase.

triggers for anomaly or threat alerts, the current GANomaly-based unsupervised approach uses a more context-oriented approach which examines the entire sensor reading at once.

Recently, however, because of the popularity of GAN-based techniques for unsupervised machine learning, approaches have been developed for fully unsupervised object detection and localization using GANs [4]. These approaches use introspection of the hidden layers of the GAN feature extractor, mapped back to the original input space, to identify which areas in the input data are contributing to the network recognizing an object.

We propose to integrate this introspection-based approach into its unsupervised GAN-based feature learning, allowing anomalies detected using those learned features to be localized. The operator-facing DAART explanation interface will then be updated to show which parts of an anomalous sensor reading are most responsible for the reading being considered anomalous.

4.2 Normalcy Exemplar

We propose to generate normalcy exemplars that can be displayed to operators to compare against detected threats. These normalcy explanations allow the system to describe what the situation typically looks like to help explain why a new instance is deemed anomalous or threatening.

We propose two techniques for generating normalcy exemplars for this comparison. The first technique is to simply determine the existing normal exemplar (non-anomalous prior reading) that is most similar to the anomalous input using the features space. A side-by-side view displays that exemplar sensor reading against the detected anomalous reading for comparison. Furthermore, bounding boxes can be added to highlight differences in the anomalous input by utilizing the localization information. One limitation to this technique is that not all differences between the normal and anomalous scenes are important. The second, and more complex, proposed technique accounts for this limitation by generating a synthetic "normal" feature vector that is similar to the anomalous reading, but without the features that make it an anomaly. The GAN then generates a synthetic sensor reading from feature vector. In this case, the only difference between the two displayed exemplars are the elements of the input that make it anomalous.

We can similarly use the GAN to determine normalcy exemplars for other data types, such as audio and acoustic sensor data, but a challenge of this task will be determining appropriate ways to expose this information to operators. Audio can be handled similarly to imagery, for example, by providing two audio clips the operator can listen to for comparison. However, for the other data types, we will work with operators to determine what view of each modality fits best into their existing threat detection workflow as part of this task.

4.3 Interactive Explanation Interface

In addition to the explanation information discussed in prior sections, prior work has introduced simple explanation types [13,17] shown to improve end user understanding of complex algorithm processes. These types include the classification, system confidence in the classification, human-understandable features of the classifier, and alternate classifications. We propose to implement a set of explanation information within a DAART interactive explanation interface to best support system transparency and user feedback. Figure 3 is a notional representation of a sample anomaly input and the explanation information that might be displayed to the user. How many and which of these explanation types to display to the user must be chosen to maximize transparency without overwhelming or confusing the operator [14]. In particular, prior work has shown that confidence should only be displayed when it is high or else will result in negative impacts on trust [16]. We propose a user study for evaluating these explanation types in the following section.

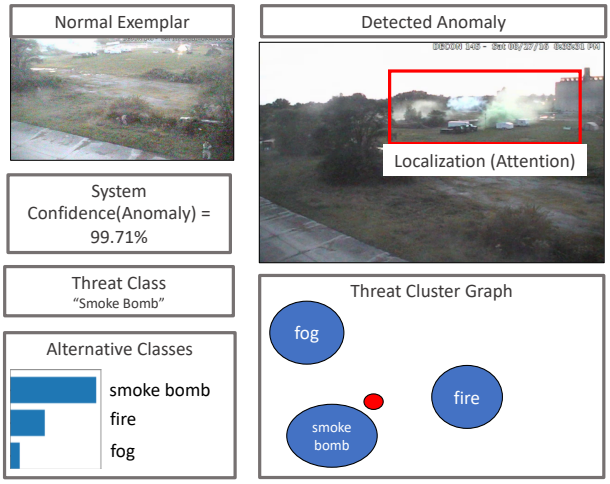

Figure 3: Notional explanation interface showing a detected anomaly and the varied explanation information that could be displayed to the user. The normal exemplar paired with the localization information provide the operator with quick understanding of what about the input is anomalous. The exposed system confidence and underlying feature information alongside the threat classifications resulting from prior operator feedback provide a more detailed understanding of how the system works. Finally, exposing the threat cluster graph provides a global understanding of previously recorded threatening and benign behavior.

These explanation presentations additionally provide an intuitive means for user feedback [13], which yields more and better feedback from the user in the loop. Operators can provide alert-level feedback to correct classifications by interacting with the assigned classification (accepting or rejecting the class) and furthermore by interacting with the alternative classes to select the correct class if it exists. Operators can provide system-level feedback by interacting with the features or localization information from the input that resulted in the classification. Additionally, as the DAART tool utilizes clusters of previously classified anomalies to perform classification, we propose to expose these clusters to the user as part of the interactive explanation interface. This view provides a global explanation of all threat data and how it is understood by the system. An operator may notice that two separate clusters can be

merged to represent a single threat type or that a single cluster should be split to represent two distinct threat types.

5 User Study

We outline a user study of our proposed explanation interface to evaluate the effects of varied explanation information on trust, feedback quality, and overall human-machine team performance.

5.1 Research Questions

The goal of the proposed study is to answer the following research questions:

Q1: Which explanation information yields the optimal human-machine team?

Similar to [20], we hope to determine a set of the explanation information to provide to the user to maximize performance while reducing system complexity.

Q2: How is trust affected by varied explanation information?

As trust is particularly important to the adoption of a system such as ours in the military domain, we intend to evaluate the effect of varied explanation information on system trust

5.2 Method

To support examination of the identified research questions, we propose a crowdsourced user study. We will identify a dataset and specific task that is representative of real system usage, but also approachable to non-security experts. This might include video on a street corner for which the human-machine team is tasked with identifying suspicious behavior or video replay from a tower defense-style game[2] for which the human-machine team is tasked with identifying aggressive behavior towards a base. In the study, we will hold all aspects of the DAART system constant, but simply vary the explanation information shown to the user during the task. After the task we will evaluate the human-machine team performance, as well as ask users to score the system in terms of trust, frustration, complexity. In this way, we can study the effects of the various explanations on user experience.

6 Conclusion

In this paper we present an unsupervised anomaly detection system, DAART, that identifies anomalies from normal behavior and classifies those anomalies as threats through interaction with operators. We additionally propose and explanation interface towards the goal of a DAART system that the user not only understands and trusts but is maximally accurate due to increased, improved user feedback. Our proposed user study aims to evaluate the explanation interface to increase effectiveness and reduce complexity.

ACKNOWLEDGMENTS

This work was supported by AFRL contract FA8650-18-P-1628 and advised by Dr. Olga Mendoza-Schrock and Mr. Todd Rovito. This publication was cleared for public release via 88ABW-2019-0665.

[2] https://web.archive.org/web/20160329012303/http://gameranx.com/features/id/13529/article/best-tower-defense-games/

REFERENCES

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. 2018. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. 1–16. https://doi.org/arXiv:1805.06725v3

[2] Daniel Barbará, Carlotta Domeniconi, and James P. Rogers. 2006. Detecting outliers using transduction and statistical testing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*. https://doi.org/10.1145/1150402.1150413

[3] Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. *1st Workshop on Explainable Artificial Intelligence*.

[4] Junsuk Choe, Joo Hyun Park, and Hyunjung Shim. 2018. Generative Adversarial Networks for Unsupervised Object Co-localization. Retrieved from http://arxiv.org/abs/1806.00236

[5] Jeff Donahue and Trevor Darrell. 2017. Adversarial Feature Learning. *Iclr*. https://doi.org/10.1038/nphoton.2013.187

[6] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2016. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2015.2496141

[7] Jerry Alan Fails and Dan R Olsen. 2003. Interactive machine learning. *Proceedings of the 8th international conference on Intelligent user interfaces IUI 03*: 39–45. https://doi.org/10.1145/604045.604056

[8] Rebecca Fiebrink, Dan Trueman, and Perry R Cook. 2009. A metainstrument for interactive, on-the-fly machine learning. In *Proceedings of New Interfaces for Musical Expression (NIME)*, 3.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets (NIPS version). *Advances in Neural Information Processing Systems 27*. https://doi.org/10.1001/jamainternmed.2016.8245

[10] Dave Gunning. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*. Retrieved October 10, 2018 from https://www.darpa.mil/program/explainable-artificial-intelligence

[11] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning* 95, 3: 423–469. https://doi.org/10.1007/s10994-013-5413-0

[12] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)*: 1–9. https://doi.org/10.1109/5.726791

[13] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, 126–137. https://doi.org/10.1145/2678025.2701399

[14] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[15] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum* 31, 3pt3: 1155–1164. https://doi.org/10.1111/j.1467-8659.2012.03108.x

[16] Brian Y. Lim and Anind K. Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, 415. https://doi.org/10.1145/2030112.2030168

[17] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*: 2119–2129. https://doi.org/10.1145/1518701.1519023

[18] Hema Raghavan, O Madani, and Rosie Jones. 2006. Active learning with feedback on features and instances. *Journal of Machine Learning Research*. https://doi.org/10.1016/S0022-460X(70)80001-3

[19] Al M. Rashid, Kimberly Ling, Regina D. Tassone, Paul Resnick, Robert Kraut, and John Riedl. 2006. Motivating Participation by Displaying the Value of Contribution. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 955–958. https://doi.org/10.1145/1124772.1124915

[20] Stephanie L. Rosenthal and Anind K. Dey. 2010. Towards Maximizing the Accuracy of Human-Labeled Sensor Data. In *Proceedings of the International*

Conference on Intelligent User Interfaces, 259–268. https://doi.org/10.1145/1719970.1720006

[21] Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.1021/jp048641u

[22] Michael Shilman, Desney S Tan, and Patrice Simard. 2006. CueTIP: a mixed-initiative interface for correcting handwriting errors. *Proceedings of the ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/1166253.1166304

[23] Kimberly Stowers, Nicholas Kasdaglis, Michael Rupp, Jessie Chen, Daniel Barber, and Michael Barnes. 2017. Insights into human-agent teaming: Intelligent agent transparency and uncertainty. In *Advances in Intelligent Systems and Computing*, 149–160. https://doi.org/10.1007/978-3-319-41959-6_13

[24] The European Parliament and The Council of The European Union. 2016. *General Data Protection Regulation*. https://doi.org/http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf

[25] Kevin Ward and Jack Davenport. 2017. Human-machine interaction to disambiguate entities in unstructured text and structured datasets. In *SPIE Conference on Next-Generation Analyst V*.