“What Am I Doing in My Research?”

Artificial Intelligence at Los Alamos National Laboratory

Elisabeth (Lissa) Moore

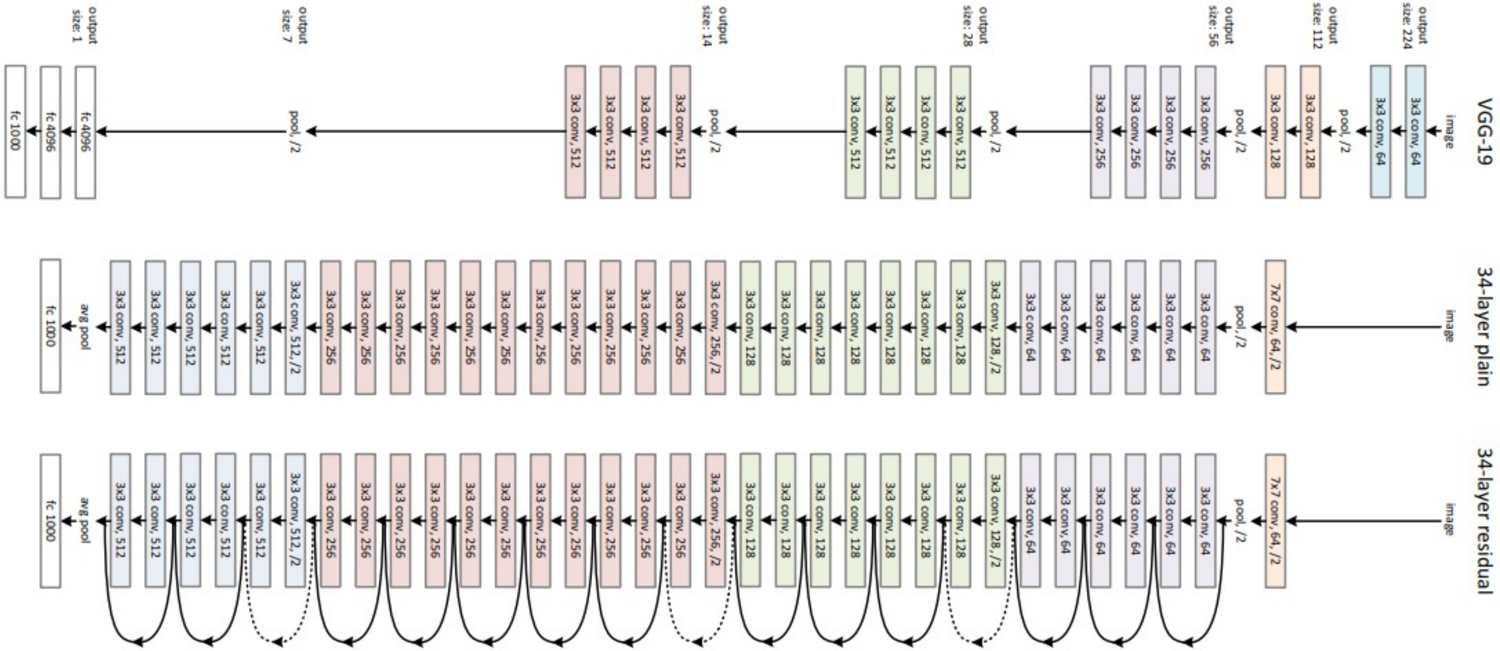
NMSU eCSR Workshop

March 20,2021

Managed by Triad National Security, LLC for the U.S. Department of Energy’s NNSA

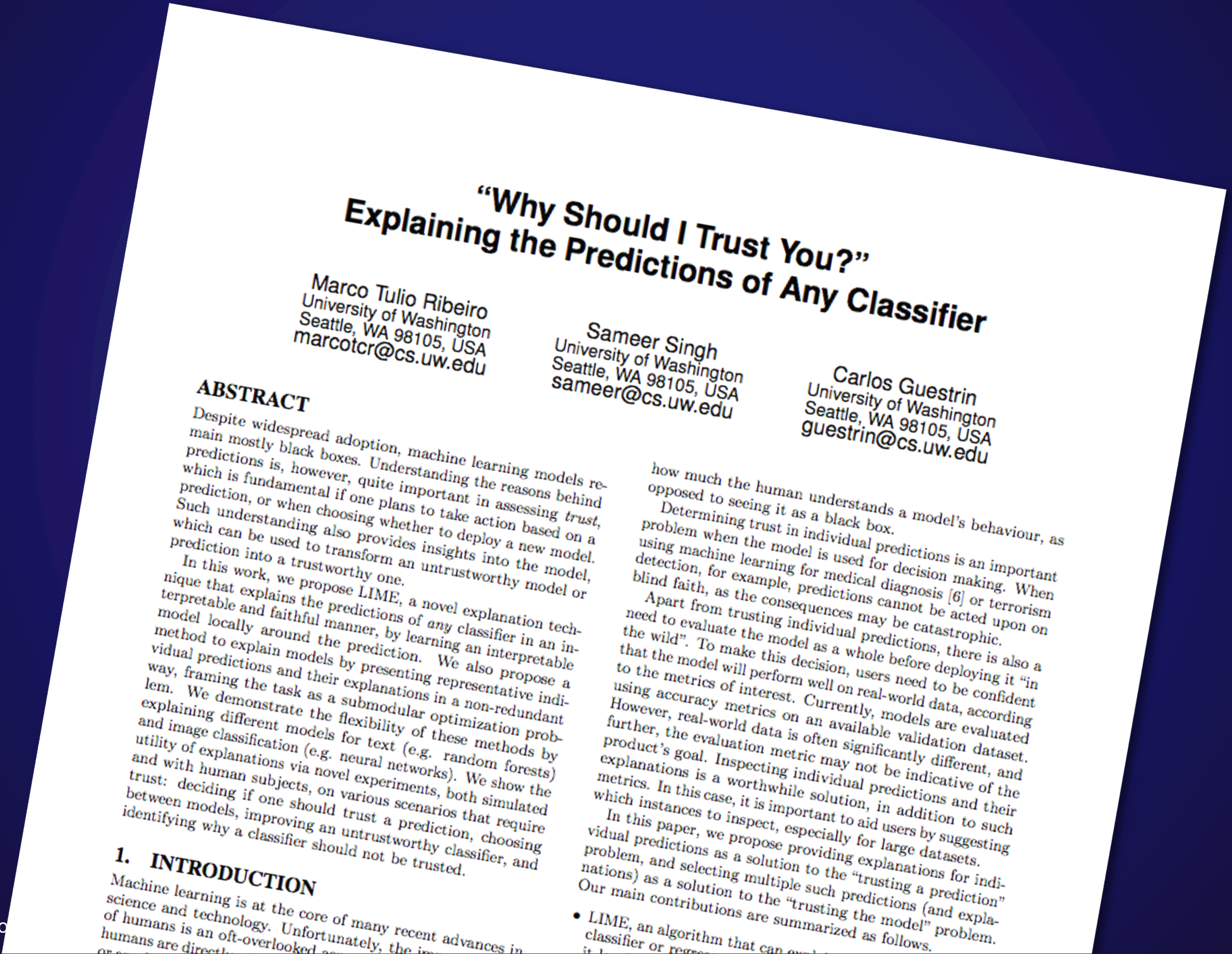
Interpretable Machine Learning

- Complex ML models have high accuracy, but we need to know *why*



- *"Why"* is crucial to safe deployment of models in the real world
  - Especially in the national security space

Interpretable Machine Learning



“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the im of humans is an oft-overlooked a humans are directl

how much the human understands a model’s behaviour, as opposed to seeing it as a black box.

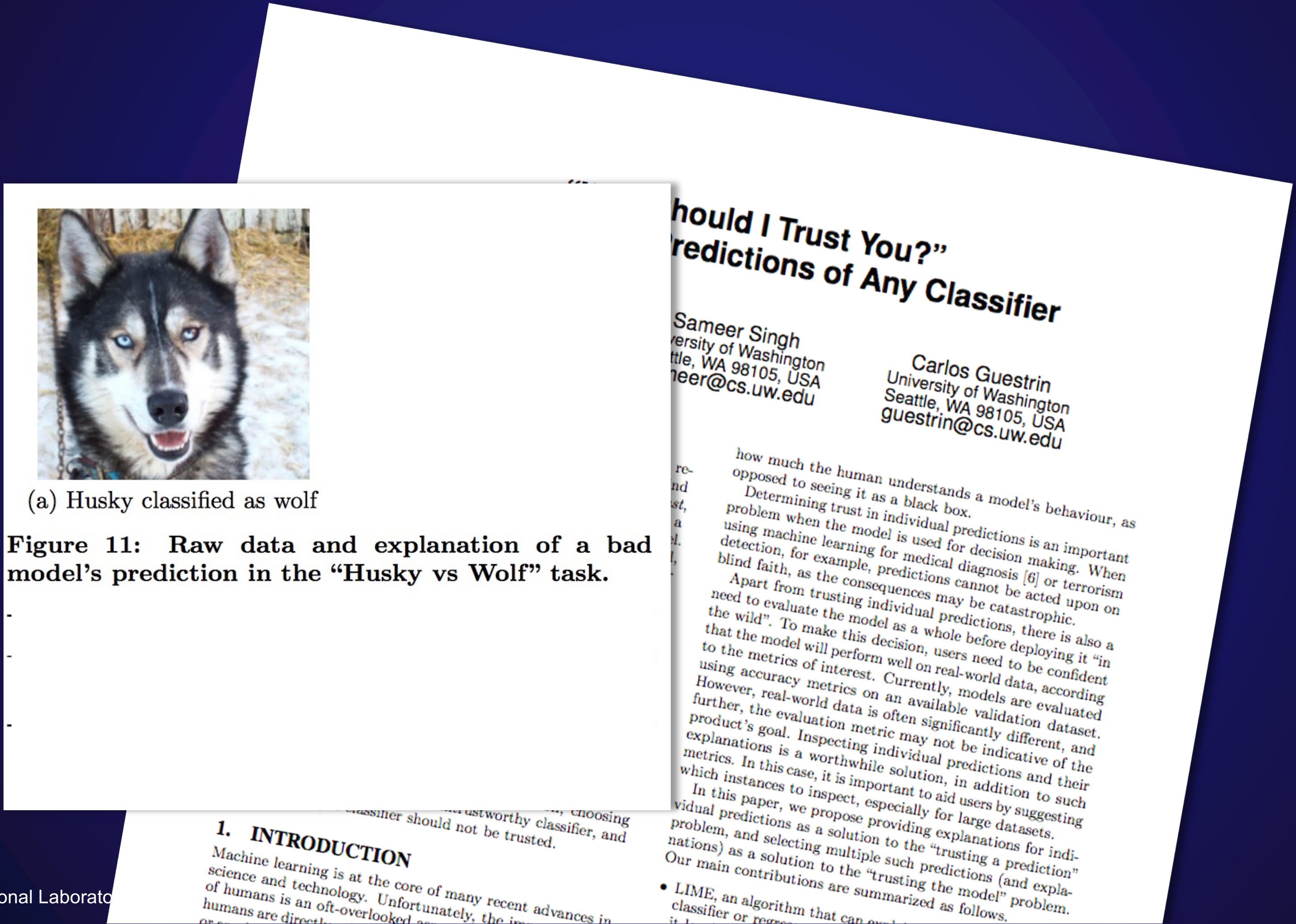
Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product’s goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can ex classifier or regre

Interpretable Machine Learning



(a) Husky classified as wolf

Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.

“…hould I Trust You?” …redictions of Any Classifier

Sameer Singh
…ersity of Washington
…ttle, WA 98105, USA
…neer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

re-
nd
st,
a
el.
l,

…, choosing
…ustworthy classifier, and
…assifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the im…
of humans is an oft-overlooked as…
humans are directl…
or c…

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

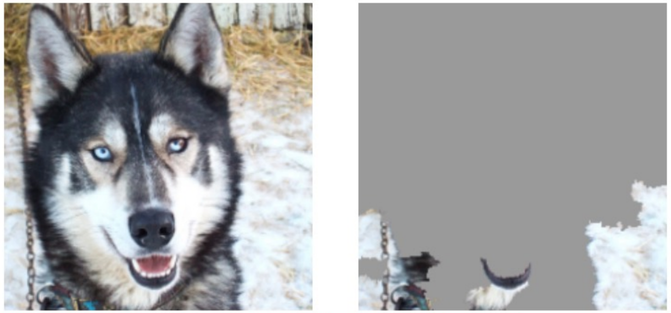
Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

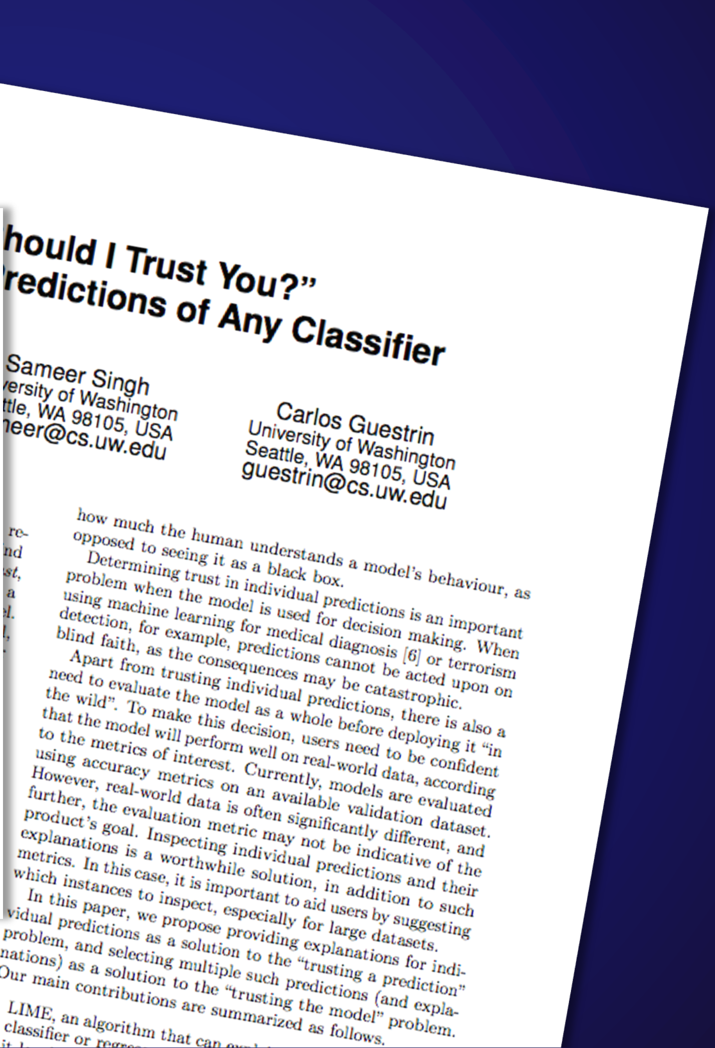
- LIME, an algorithm that can ex…
classifier or regre…

Interpretable Machine Learning



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.



"hould I Trust You?" redictions of Any Classifier

Sameer Singh
ersity of Washington
ttle, WA 98105, USA
eer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

re-
nd
st,
a
l.
l,

, choosing ustworthy classifier, and classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the im of humans is an oft-overlooked a humans are directl

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

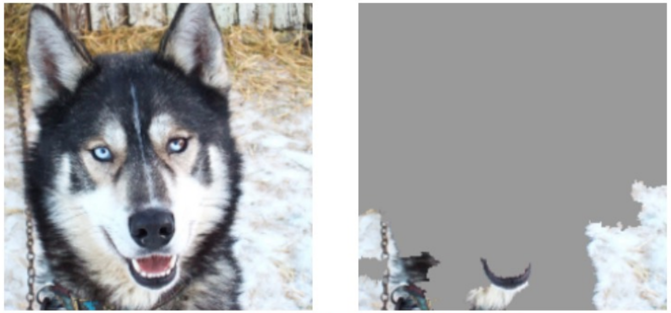
Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the "trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the "trusting the model" problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can ex classifier or regre

Interpretable Machine Learning

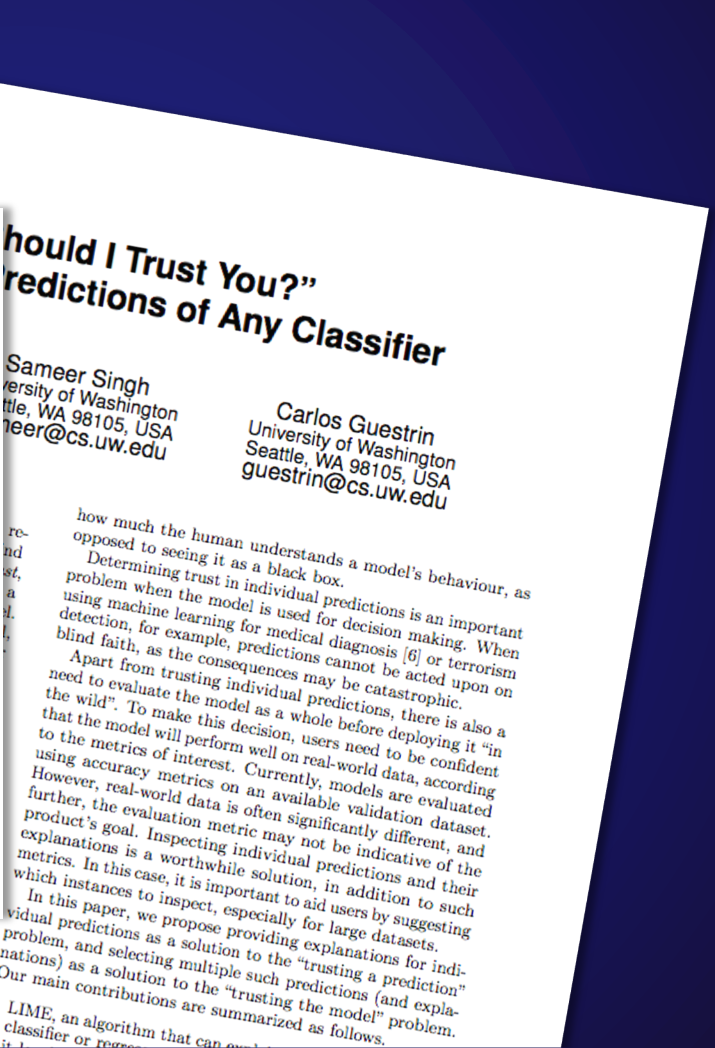


(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: “Husky vs Wolf” experiment results.



“...hould I Trust You?” ...redictions of Any Classifier

Sameer Singh
...ersity of Washington
...ttle, WA 98105, USA
...eer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

re- nd st, a el. l,

..., choosing ...ustworthy classifier, and ...assifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the im... of humans is an oft-overlooked a... humans are directl... or ...

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can ex... classifier or regre... it l...

Interpretable Machine Learning

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim Martin Wattenberg Justin Gilmer Carrie Cai James Wexler
Fernanda Viegas Rory Sayres

arXiv:1711.11279v5 [stat.ML] 7 Jun 2018

Abstract

The interpretation of deep learning models is a challenge due to their size, complexity, and often opaque internal state. In addition, many systems, such as image classifiers, operate on low-level features rather than high-level concepts. To address these challenges, we introduce Concept Activation Vectors (CAVs), which provide an interpretation of a neural net's internal state in terms of human-friendly concepts. The key idea is to view the high-dimensional internal state of a neural net as an aid, not an obstacle. We show how to use CAVs as part of a technique, Testing with CAVs (TCAV), that uses directional derivatives to quantify the degree to which a user-defined concept is important to a classification result–for example, how sensitive a prediction of zebra is to the presence of stripes. Using the domain of image classification as a testing ground, we describe how CAVs may be used to explore hypotheses and generate insights for a standard image classification network as well as a medical application.

1. Introduction

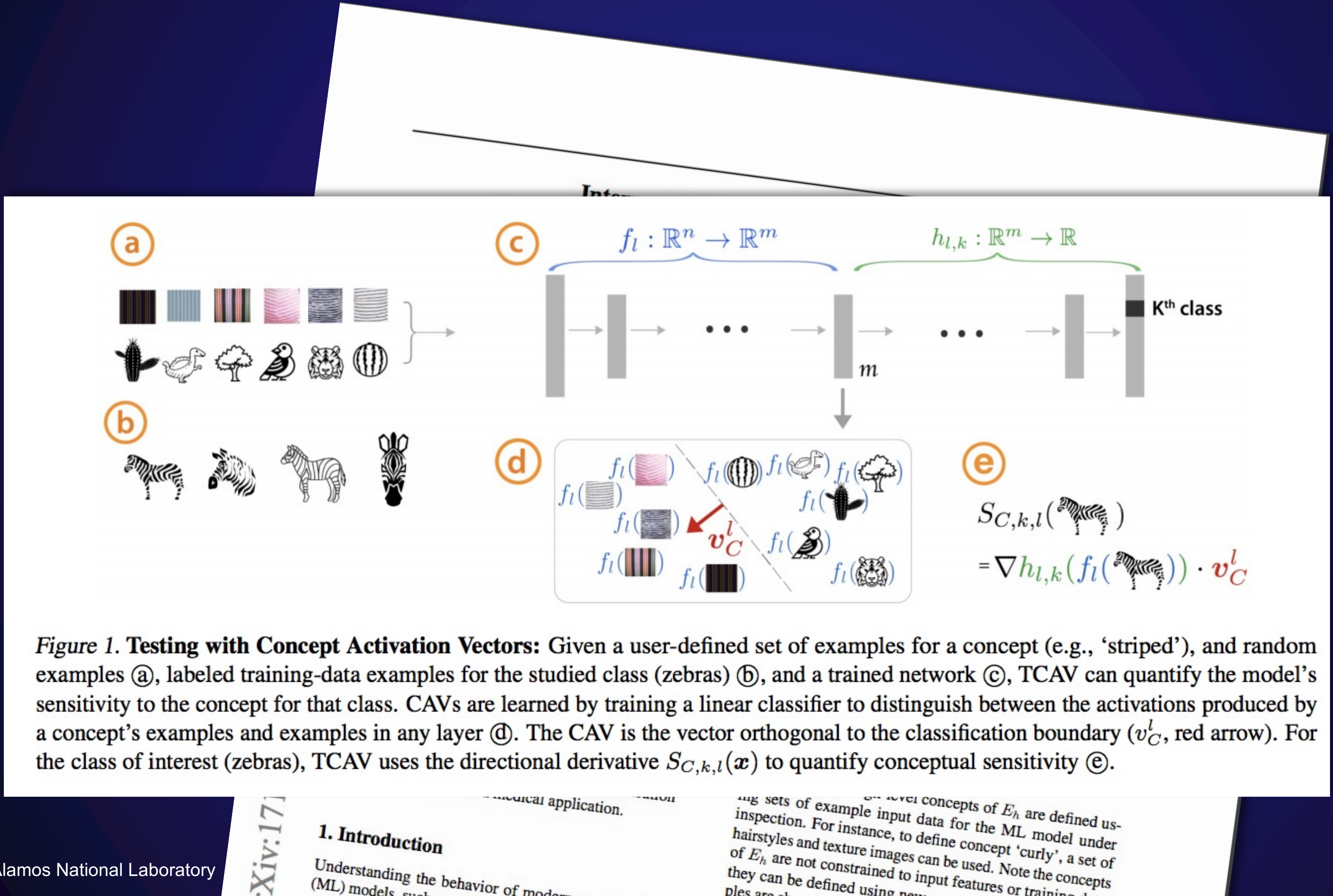
Understanding the behavior of moder
(ML) models, suc

A key difficulty, however, is that most ML models operate on features, such as pixel values, that do not correspond to high-level concepts that humans easily understand. Furthermore, a model's internal values (e.g., neural activations) can seem incomprehensible. We can express this difficulty mathematically, viewing the state of an ML model as a vector space E_m spanned by basis vectors e_m which correspond to data such as input features and neural activations. Humans work in a different vector space E_h spanned by implicit vectors e_h corresponding to an unknown set of human-interpretable concepts.

From this standpoint, an "interpretation" of an ML model can be seen as function $g : E_m \to E_h$. When g is linear, we call it a **linear interpretability**. In general, an interpretability function g need not be perfect (Doshi-Velez, 2017); it may fail to explain some aspects of its input domain E_m and it will unavoidably not cover all possible human concepts in E_h.

In this work, the high-level concepts of E_h are defined using sets of example input data for the ML model under inspection. For instance, to define concept 'curly', a set of hairstyles and texture images can be used. Note the concepts of E_h are not constrained to input features or training d
they can be defined using new
ples are al

Interpretable Machine Learning



Inte
a
b
c
d
e
$f_l : \mathbb{R}^n \to \mathbb{R}^m$
$h_{l,k} : \mathbb{R}^m \to \mathbb{R}$
K^th class
m
$f_l(\;)$
v_C^l
$S_{C,k,l}(\;)$
$= \nabla h_{l,k}(f_l(\;)) \cdot v_C^l$

Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(\boldsymbol{x})$ to quantify conceptual sensitivity ⓔ.

Xiv:17

edical application.

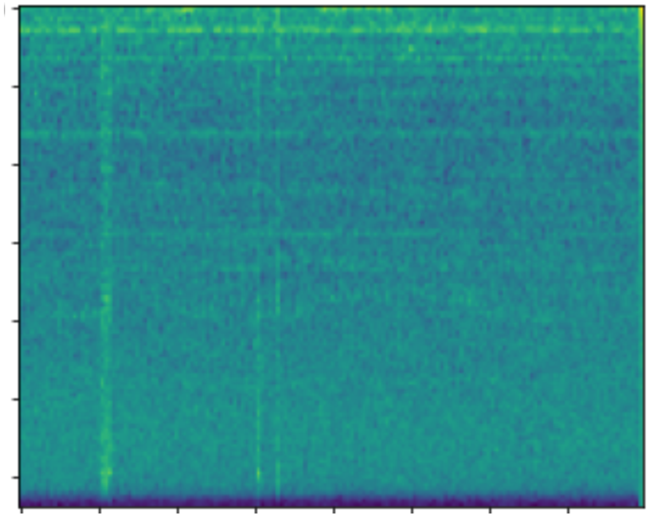
1. Introduction

Understanding the behavior of mode
(ML) models, such

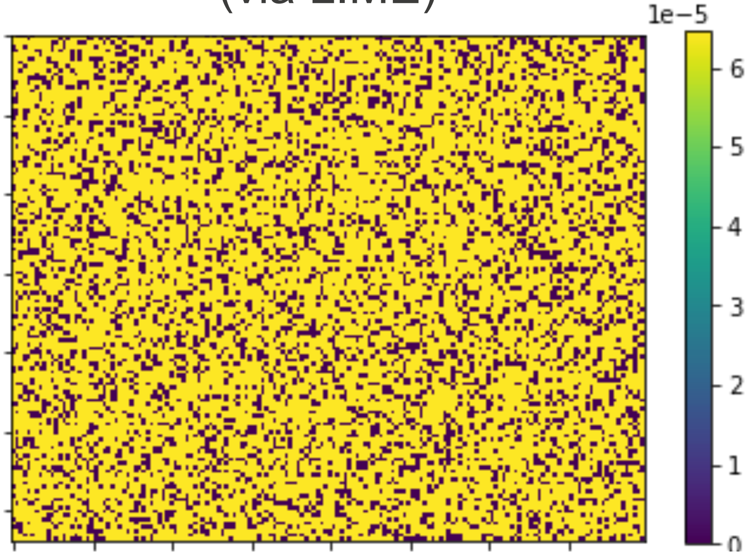
gh-level concepts of E_h are defined us-
ing sets of example input data for the ML model under
inspection. For instance, to define concept 'curly', a set of
hairstyles and texture images can be used. Note the concepts
of E_h are not constrained to input features or training d
they can be defined using new
ples are

We need to move beyond natural image explanations

Input Spectrogram



Feature Importance (via LIME)



Note very small range of color bar scale.

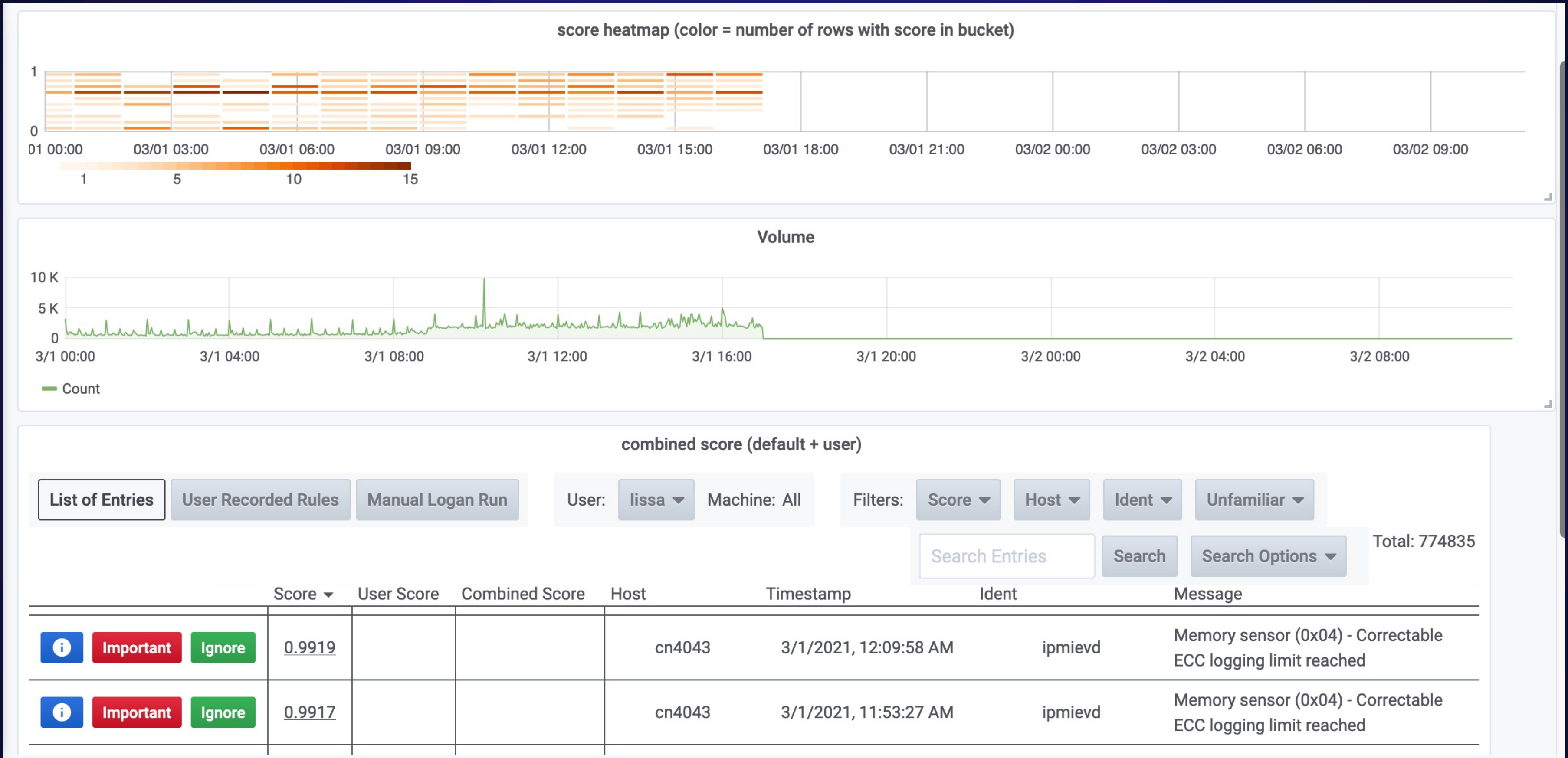
We need to move beyond natural image explanations

- Majority of available interpretability techniques focus on natural image classification

- National security data is multi-modal: images, text, time series, etc.

- National security projects involve more than classification: anomaly detection, knowledge discovery, etc.

Logan: Computer-Generated Text Log Anomaly Detection



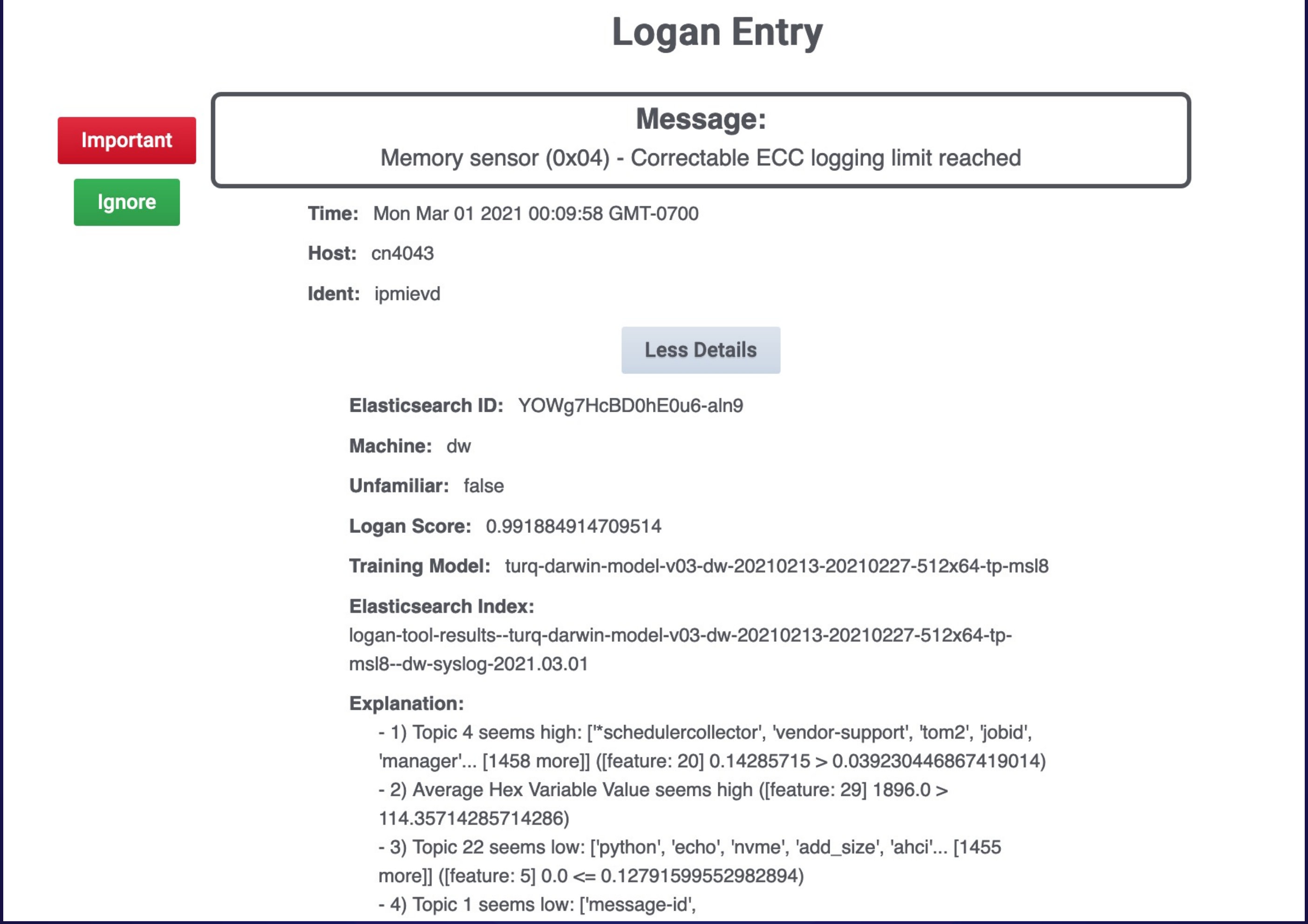
User: lissa

Machine: All

Total: 774835

	Score	User Score	Combined Score	Host	Timestamp	Ident	Message
i Important Ignore	0.9919			cn4043	3/1/2021, 12:09:58 AM	ipmievd	Memory sensor (0x04) - Correctable ECC logging limit reached
i Important Ignore	0.9917			cn4043	3/1/2021, 11:53:27 AM	ipmievd	Memory sensor (0x04) - Correctable ECC logging limit reached

Logan: Computer-Generated Text Log Anomaly Detection



Logan Entry

Message:

Memory sensor (0x04) - Correctable ECC logging limit reached

Time: Mon Mar 01 2021 00:09:58 GMT-0700

Host: cn4043

Ident: ipmievd

Elasticsearch ID: YOWg7HcBD0hE0u6-aln9

Machine: dw

Unfamiliar: false

Logan Score: 0.991884914709514

Training Model: turq-darwin-model-v03-dw-20210213-20210227-512x64-tp-msl8

Elasticsearch Index:
logan-tool-results--turq-darwin-model-v03-dw-20210213-20210227-512x64-tp-msl8--dw-syslog-2021.03.01

Explanation:

- 1) Topic 4 seems high: ['*schedulercollector', 'vendor-support', 'tom2', 'jobid', 'manager'... [1458 more]] ([feature: 20] 0.14285715 > 0.039230446867419014)
- 2) Average Hex Variable Value seems high ([feature: 29] 1896.0 > 114.35714285714286)
- 3) Topic 22 seems low: ['python', 'echo', 'nvme', 'add_size', 'ahci'... [1455 more]] ([feature: 5] 0.0 <= 0.12791599552982894)
- 4) Topic 1 seems low: ['message-id',

Why Los Alamos National Laboratory?

- Middle-ground between academia and industry
- Focus on important, hard problems neglected by academia and industry
- Opportunities to work on a variety of projects
  - HPC
  - Social Network Analysis
  - Adversarial Defense
  - Method Development
  - Quantum Computing
- Diverse, friendly environment

- Offers summer and year-round research internships to students at all levels

Thank you!

lissa@lanl.gov