Visual Analytics Perspectives
on **Interactive** and **Explainable Machine Learning**

Mennatallah El-Assady

University Konstanz

@manunna_91

el-assady.com

Volume xx (200y), Number z, pp. 1–29

The State of the Art in Integrating Machine Learning into Visual Analytics

A. Endert[1], W. Ribarsky[2], C. Turkay[3], W. Wong[4], I. Nabney[5], I. Díaz Blanco[6], F. Rossi[7]

[1]Georgia Tech, USA
[2]University of North Carolina, Charlotte, USA
[3]City University of London, UK
[4]Middlesex University, UK
[5]Aston University, UK
[6]University of Oviedo, Spain
[7]Paris 1 Panthéon Sorbonne University, Paris

Abstract

Visual analytics systems combine machine learning or other analytic techniques with interactive data visualization to promote sensemaking and analytical reasoning. It is through such techniques that people can make sense of large, complex data. While progress has been made, the tactful combination of machine learning and data visualization is still under-explored. This state-of-the-art report presents a summary of the progress that has been made by highlighting and synthesizing select research advances. Further, it presents opportunities and challenges to enhance the synergy between machine learning and visual analytics for impactful future research directions.

Categories and Subject Descriptors (according to ACM CCS): Human-centered computing - Visualization, Visual analytics

1. Introduction

We are in a data-driven era. Increasingly more domains generate and consume data. People have the potential to understand phenomena in more depth using new data analysis techniques. Additionally, new phenomena can be uncovered in domains where data is becoming available. Thus, making sense of data is becoming increasingly important, and this is driving the need for systems that enable people to analyze and understand data.

However, this opportunity to discover also presents challenges. Reasoning about data is becoming more complicated and difficult as data scales and complexities increase. People require powerful tools to draw valid conclusions from data, while maintaining trustworthy and interpretable results.

We claim that visual analytics (VA) and machine learning (ML) have complementing strengths and weaknesses to address these challenges. Visual analytics (VA) is a multi-disciplinary domain that combines data visualization with machine learning (ML) and other automated techniques to create systems that help people make sense of data [TC05, KSF*08, Kei02, KMSZ06]. Over the years, much work has been done to establish the foundations of this area, create research advances in select topics, and form a community of researchers to continue to evolve the state of the art.

Currently, VA techniques exist that make use of select ML models or algorithms. However, there are additional techniques that can apply to the broader visual data analysis process. Doing so reveals opportunities for how to couple user tasks and activities with such models. Similarly, opportunities exist to advance ML models based on the cognitive tasks invoked by interactive VA techniques.

This state-of-the-art report briefly summarizes the advances made at the intersection of ML and VA. It describes the extent to which machine learning methods are utilized in visual analytics to date. Further, it illuminates the opportunities within both disciplines that can drive important research directions in the future. Much of the content and inspiration for this paper originated during a Dagstuhl Seminar titled, "Bridging Machine Learning with Information Visualization (15101)" [KMRV15].

Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective

Shixia Liu, Xiting Wang, Mengchen Liu, Jun Zhu

Tsinghua University, Beijing, China

Abstract

Interactive model analysis, the process of understanding, diagnosing, and refining a machine learning model with the help of interactive visualization, is very important for users to efficiently solve real-world artificial intelligence and data mining problems. Dramatic advances in big data analytics has led to a wide variety of interactive model analysis tasks. In this paper, we present a comprehensive analysis and interpretation of this rapidly developing area. Specifically, we classify the relevant work into three categories: understanding, diagnosis, and refinement. Each category is exemplified by recent influential work. Possible future research opportunities are also explored and discussed.

Keywords: interactive model analysis, interactive visualization, machine learning, understanding, diagnosis, refinement

1. Introduction

Machine learning has been successfully applied to a wide variety of fields ranging from information retrieval, data mining, and speech recognition, to computer graphics, visualization, and human-computer interaction. However, most users often treat a machine learning model as a black box because of its incomprehensible functions and unclear working mechanism [1, 2, 3]. Without a clear understanding of how and why a model works, the development of high-performance models typically relies on a time-consuming trial-and-error pro-

*Fully documented templates are available in the elsarticle package on CTAN.

Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers

Fred Hohman, *Member, IEEE*, Minsuk Kahng, *Member, IEEE*, Robert Pienta, *Member, IEEE*, and Duen Horng Chau, *Member, IEEE*

Abstract—Deep learning has recently seen rapid development and received significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the internal complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such performance are challenging and sometimes mystifying to interpret. As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement. We present a survey of the role of visual analytics in deep learning research, which highlights its short yet impactful history and thoroughly summarizes the state-of-the-art using a human-centered interrogative framework, focusing on the *Five W's and How* (Why, Who, What, How, When, and Where). We conclude by highlighting research directions and open research problems. This survey helps researchers and practitioners in both visual analytics and deep learning to quickly learn key aspects of this young and rapidly growing body of research, whose impact spans a diverse range of domains.

Index Terms—Deep learning, visual analytics, information visualization, neural networks

1 Introduction

Deep learning is a specific set of techniques from the broader field of machine learning (ML) that focus on the study and usage of *deep* artificial neural networks to learn structured representations of data. First mentioned as early as the 1940s [1], artificial neural networks have a rich history [2], and have recently seen a dominate and pervasive resurgence [3], [4], [5] in many research domains by producing state-of-the-art results [6], [7] on a number of diverse big data tasks [8], [9]. For example, the premiere machine learning, deep learning, and artificial intelligence (AI) conferences have seen enormous growth in attendance and paper submissions since early 2010s. Furthermore, open-source toolkits and programming libraries for building, training, and evaluating deep neural networks have become more robust and easy to use, democratizing deep learning. As a result, the barrier to developing deep learning models is lower than ever before and deep learning applications are becoming pervasive.

While this technological progress is impressive, it comes with unique and novel challenges. For example, the lack of interpretability and transparency of neural networks, from the learned representations to the underlying decision process, is an important problem to address. Making sense of why a particular model misclassifies test data instances or behaves poorly at times is a challenging task for model developers. Similarly, end-users interacting with an application that relies on deep learning to make critical decisions may question its reliability if no explanation is given by the model, or become baffled if the explanation is convoluted.

• *F. Hohman, M. Kahng, R. Pienta, and D. H. Chau are with the College of Computing, Georgia Tech, Atlanta, Georgia 30332, U.S.A. E-mail: {fredhohman, kahng, pientars, polo}@gatech.edu*

While explaining neural network decisions is important, there are numerous other problems that arise from deep learning, such as AI safety and security (e.g., when using models in applications such as self-driving vehicles), and compromised trust due to bias in models and datasets, just to name a few. These challenges are often compounded, due to the large datasets required to train most deep learning models. As worrisome as these problems are, they will likely become even more widespread as more AI-powered systems are deployed in the world. Therefore, a general sense of model understanding is not only beneficial, but often required to address the aforementioned issues.

Data visualization and visual analytics excel at knowledge communication and insight discovery by using encodings to transform abstract data into meaningful representations. In the seminal work by Zeiler and Fergus [10], a technique called *deconvolutional networks* enabled projection from a model's learned feature space back to the pixel space. Their technique and results give insight into what types of features deep neural networks are learning at specific layers, and also serve as a debugging tool for improving a model. This work is often credited for popularizing visualization in the machine learning and computer vision communities in recent years, putting a spotlight on it as a powerful tool that helps people understand and improve deep learning models. However, visualization research for neural networks started well before [11], [12], [13]. Over just a handful of years, many different techniques have been introduced to help interpret what neural networks are learning. Many such techniques generate static images, such as attention maps and heatmaps for image classification, indicating which parts of an image are most important to the classification. However, interaction has also been

explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady

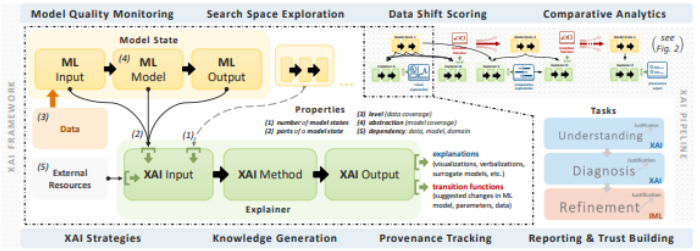


Fig. 1: Close-up view of an *explainer*, the main building-block used to construct an iterative *XAI pipeline* for the *understanding*, *diagnosis*, and *refinement* of ML models. Explainers have five properties; they take one or more model states as input, applying an XAI method, to output an explanation or a transition function. Global monitoring and steering mechanisms expand the pipeline to the full *XAI framework*, supporting the overall workflow by guiding, steering, or tracking the explainers during all steps.

Abstract— We propose a framework for interactive and explainable machine learning that enables users to (1) understand machine learning models; (2) diagnose model limitations using different explainable AI methods; as well as (3) refine and optimize the models. Our framework combines an iterative XAI pipeline with eight global monitoring and steering mechanisms, including quality monitoring, provenance tracking, model comparison, and trust building. To operationalize the framework, we present explAIner, a visual analytics system for interactive and explainable machine learning that instantiates all phases of the suggested pipeline within the commonly used TensorBoard environment. We performed a user-study with nine participants across different expertise levels to examine their perception of our workflow and to collect suggestions to fill the gap between our system and framework. The evaluation confirms that our tightly integrated system leads to an informed machine learning process while disclosing opportunities for further extensions.

Index Terms—Explainable AI, Interactive Machine Learning, Deep Learning, Visual Analytics, Interpretability, Explainability

1 Introduction

Since the first presentation of neural networks in the 1940s [47], we have seen a great increase in works on Artificial Intelligence (AI) and Machine Learning (ML). Especially within the last decade, computational resources have become cheaper and more accessible. This development has led to new state-of-the-art solutions, e.g., Deep Learning (DL), while the increasing availability of tools and libraries has led to a democratization of ML methods in a variety of domains [30]. For example, DL methods outperform traditional algorithms for image processing [56] or natural language processing [82] and can often be applied by domain experts without prior ML expertise [12].

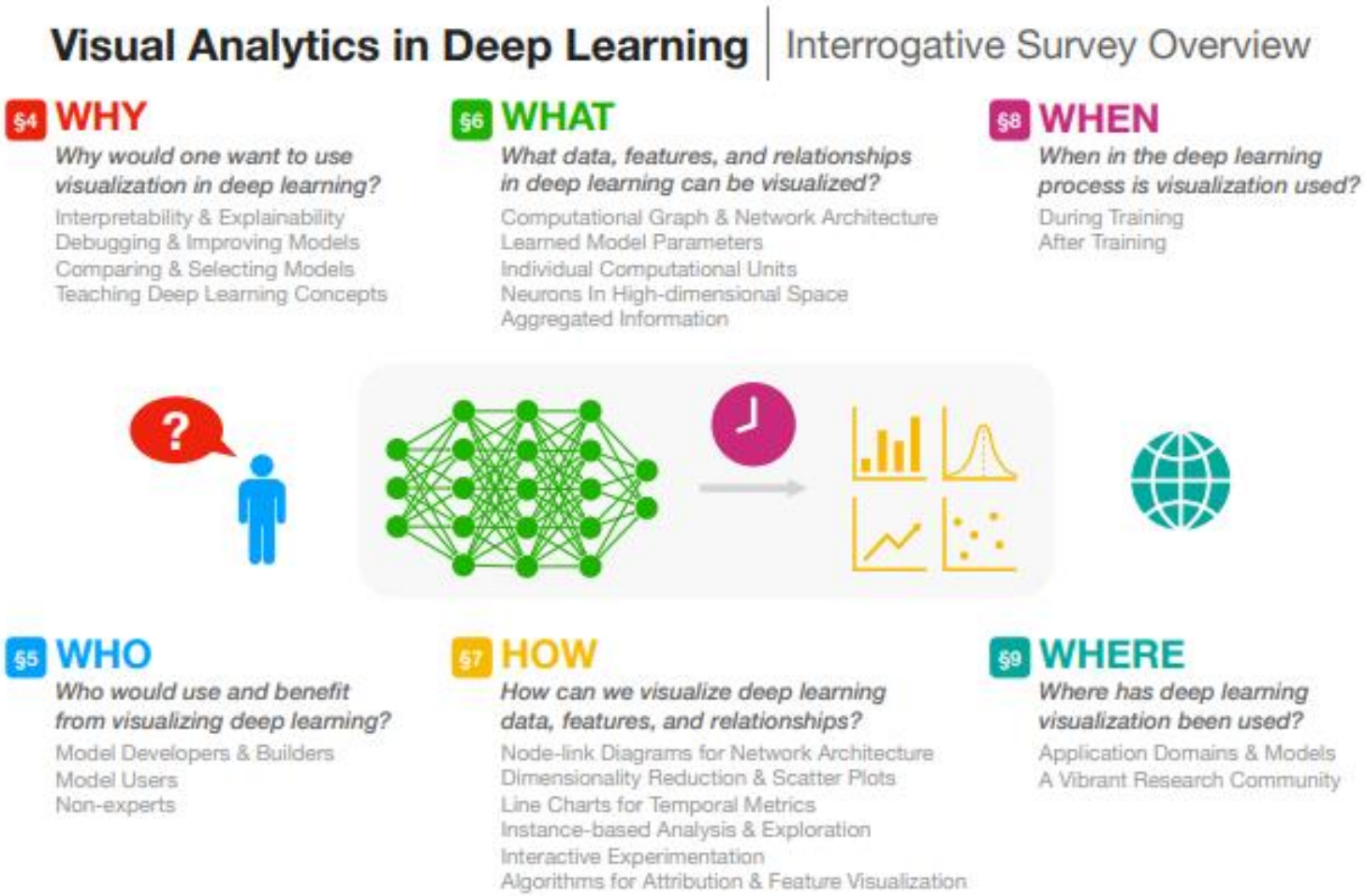
Despite the significant improvement in performance, DL models create novel challenges, due to their nature of being *black-boxes* [78]. For model developers, missing transparency in the decision-making of DL models often leads to a time-consuming trial and error process [81]. Additionally, whenever such decisions concern end-user applications, e.g., self-driving cars, trust is essential. In critical domains, this trust has to be substantiated either by reliable and unbiased decision outcomes, or convincing rationalization and justifications [66]. The growing prevalence of AI in security-critical domains leads to an ever-increasing demand for explainable and reproducible results.

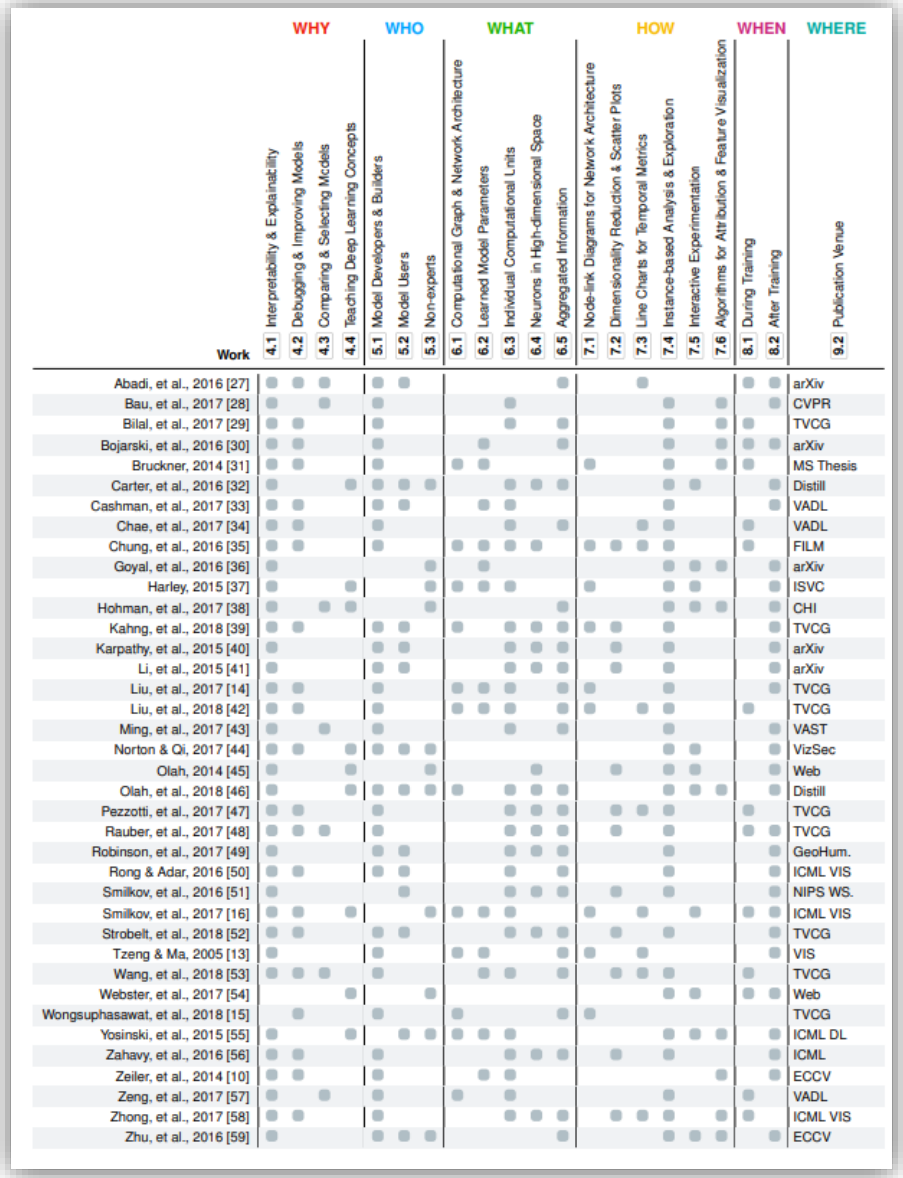
Several solutions address the problem of missing transparency in black-box models, often referred to as e**X**plainable **A**rtificial **I**ntelligence (XAI) [26]. Even though AI algorithms often cannot be directly explained [2], XAI methods aim to provide human-readable, as well as interpretable explanations of the decisions taken by such algorithms. XAI is further driven by newly introduced regulations, such as the *European General Data Protection Regulation* [77], demanding accessible justifications for automated, consumer-facing decisions, prompting businesses to seek reliable XAI solutions. A natural way to obtain human interpretable explanations is through visualizations.

More recent work focuses not only on visual design but also on interactive, mixed-initiative workflows, as provided by Visual Analytics (VA) systems [21]. Also, an exploratory workflow [60] enables a more targeted analysis and design of ML models. Visual analytics further helps in bridging the gap between user knowledge and the insights XAI methods can provide. As AI is affecting a broader range of user groups, ranging from everyday users to model developers, the differing levels of background knowledge in these user groups bring along varying requirements for the explainability.

T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady are with the University of Konstanz. E-Mail: firstname.lastname@uni-konstanz.de

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/





	WHY				WHO			WHAT					HOW							WHEN		WHERE
Work	4.1 Interpretability & Explainability	4.2 Debugging & Improving Models	4.3 Comparing & Selecting Models	4.4 Teaching Deep Learning Concepts	5.1 Model Developers & Builders	5.2 Model Users	5.3 Non-experts	6.1 Computational Graph & Network Architecture	6.2 Learned Model Parameters	6.3 Individual Computational Units	6.4 Neurons in High-dimensional Space	6.5 Aggregated Information	7.1 Node-link Diagrams for Network Architecture	7.2 Dimensionality Reduction & Scatter Plots	7.3 Line Charts for Temporal Metrics	7.4 Instance-based Analysis & Exploration	7.5 Interactive Experimentation	7.6 Algorithms for Attribution & Feature Visualization	8.1 During Training	8.2 After Training	9.2 Publication Venue
Abadi, et al., 2016 [27]	●	●	●		●	●						●				●				●	●	arXiv
Bau, et al., 2017 [28]	●		●		●					●						●		●			●	CVPR
Bilal, et al., 2017 [29]	●	●			●					●		●				●		●		●		TVCG
Bojarski, et al., 2016 [30]	●	●			●				●			●				●		●		●	●	arXiv
Bruckner, 2014 [31]	●	●			●			●	●				●			●		●		●		MS Thesis
Carter, et al., 2016 [32]	●			●	●	●	●			●	●	●				●	●				●	Distill
Cashman, et al., 2017 [33]	●	●			●	●			●	●						●					●	VADL
Chae, et al., 2017 [34]	●	●			●					●		●			●	●				●		VADL
Chung, et al., 2016 [35]	●	●			●			●	●	●	●		●	●	●	●				●		FILM
Goyal, et al., 2016 [36]	●						●		●							●	●	●			●	arXiv
Harley, 2015 [37]	●			●			●	●	●	●			●			●	●				●	ISVC
Hohman, et al., 2017 [38]	●		●	●			●					●				●	●	●			●	CHI
Kahng, et al., 2018 [39]	●	●			●	●		●		●	●	●	●	●		●					●	TVCG
Karpathy, et al., 2015 [40]	●				●	●				●	●	●		●		●					●	arXiv
Li, et al., 2015 [41]	●				●	●				●	●	●		●		●					●	arXiv
Liu, et al., 2017 [14]	●	●			●			●	●	●		●	●			●					●	TVCG
Liu, et al., 2018 [42]	●	●			●			●	●	●		●	●		●	●				●		TVCG
Ming, et al., 2017 [43]	●		●		●					●		●				●					●	VAST
Norton & Qi, 2017 [44]	●	●		●	●	●	●									●	●				●	VizSec
Olah, 2014 [45]	●			●			●				●			●		●	●				●	Web
Olah, et al., 2018 [46]	●			●	●	●	●	●		●	●	●				●	●	●			●	Distill
Pezzotti, et al., 2017 [47]	●	●			●					●	●	●		●	●	●				●		TVCG
Rauber, et al., 2017 [48]	●	●	●		●					●	●	●		●		●				●	●	TVCG
Robinson, et al., 2017 [49]	●				●	●				●	●	●				●					●	GeoHum.
Rong & Adar, 2016 [50]	●	●			●	●				●		●				●					●	ICML VIS
Smilkov, et al., 2016 [51]	●					●				●	●	●		●		●					●	NIPS WS.
Smilkov, et al., 2017 [16]	●	●		●			●	●	●	●			●		●		●			●	●	ICML VIS
Strobelt, et al., 2018 [52]	●	●			●	●				●	●	●		●		●					●	TVCG
Tzeng & Ma, 2005 [13]	●				●			●	●			●	●		●						●	VIS
Wang, et al., 2018 [53]	●	●	●		●				●	●		●		●	●	●				●		TVCG
Webster, et al., 2017 [54]				●			●									●	●			●	●	Web
Wongsuphasawat, et al., 2018 [15]		●			●			●				●	●									TVCG
Yosinski, et al., 2015 [55]	●			●		●	●	●	●	●						●	●	●			●	ICML DL
Zahavy, et al., 2016 [56]	●	●			●					●	●	●		●		●					●	ICML
Zeiler, et al., 2014 [10]	●	●			●				●	●								●			●	ECCV
Zeng, et al., 2017 [57]	●		●		●			●		●						●				●		VADL
Zhong, et al., 2017 [58]	●	●			●					●	●	●		●	●	●		●		●		ICML VIS
Zhu, et al., 2016 [59]	●				●	●	●					●				●	●	●			●	ECCV

F. Hohman, M. Kahng, R. Pienta, and D. Chau. **Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers**. *IEEE Trans. Vis. Comput. Graph.*, 2018.

Model Novice

The Newcomer.

The model novice is the 'new one' in the machine-learning class. His goal is to learn about concepts of machine learning models; he wants to understand the building blocks of the model as well as its general working. Learning resources are essential to him, either by example or by textual, visual, or external resources.

The Operator.

He is the 'user' among the users, i.e., he uses existing machine learning models to solve specific tasks. For example, this could be a domain expert - let's say a biologist - who needs to classify protein structures. To decide on a model, he wants to compare architectures, understand their underlying working, and verify his decision by executing XAI methods on some data samples.

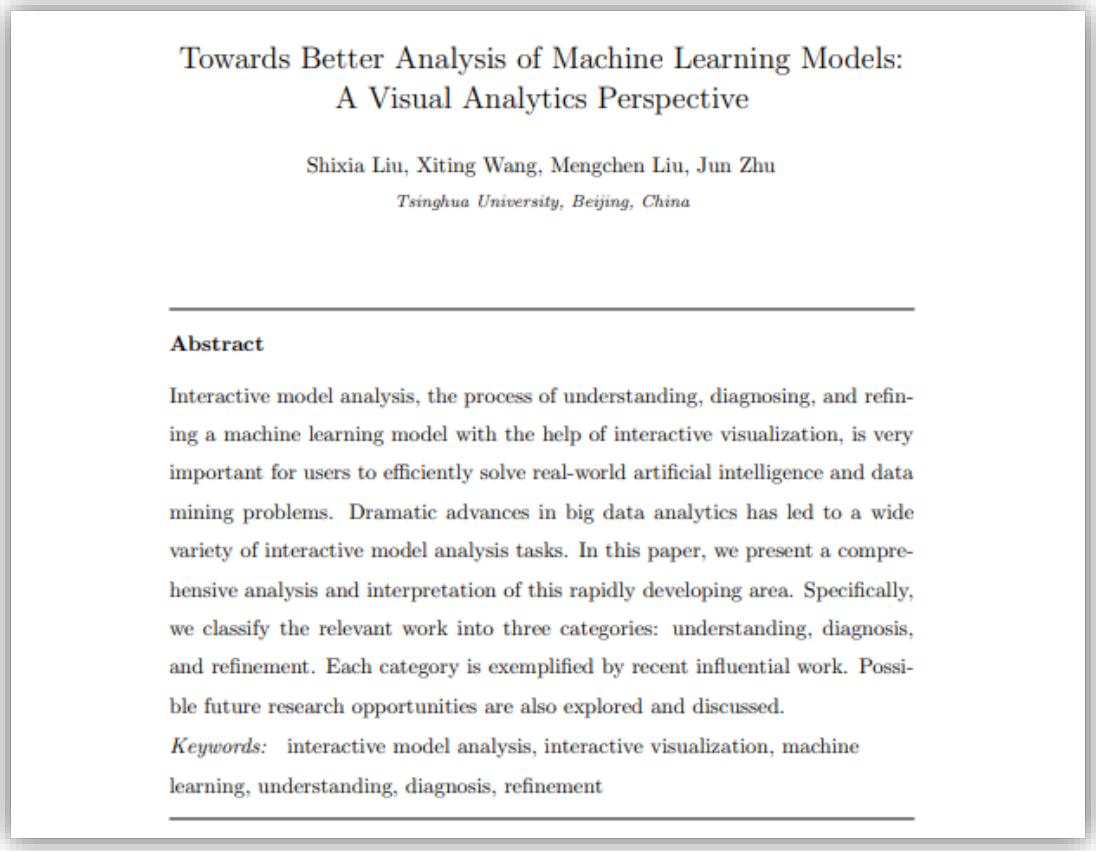
Model User

Model Developer

The Expert.

The model developer is an expert on machine learning. He develops models from scratch, refines existing models, and optimizes parameters to improve the model's performance. He is interested in the architecture of the model, including in-depth information, such as layer-sizes, initializers, and activation functions. To debug the model, explanations on all abstraction levels are relevant. His insights might lead to a model update, covering the full development and refinement process.

The **Process** of Interactive and Explainable Machine Learning



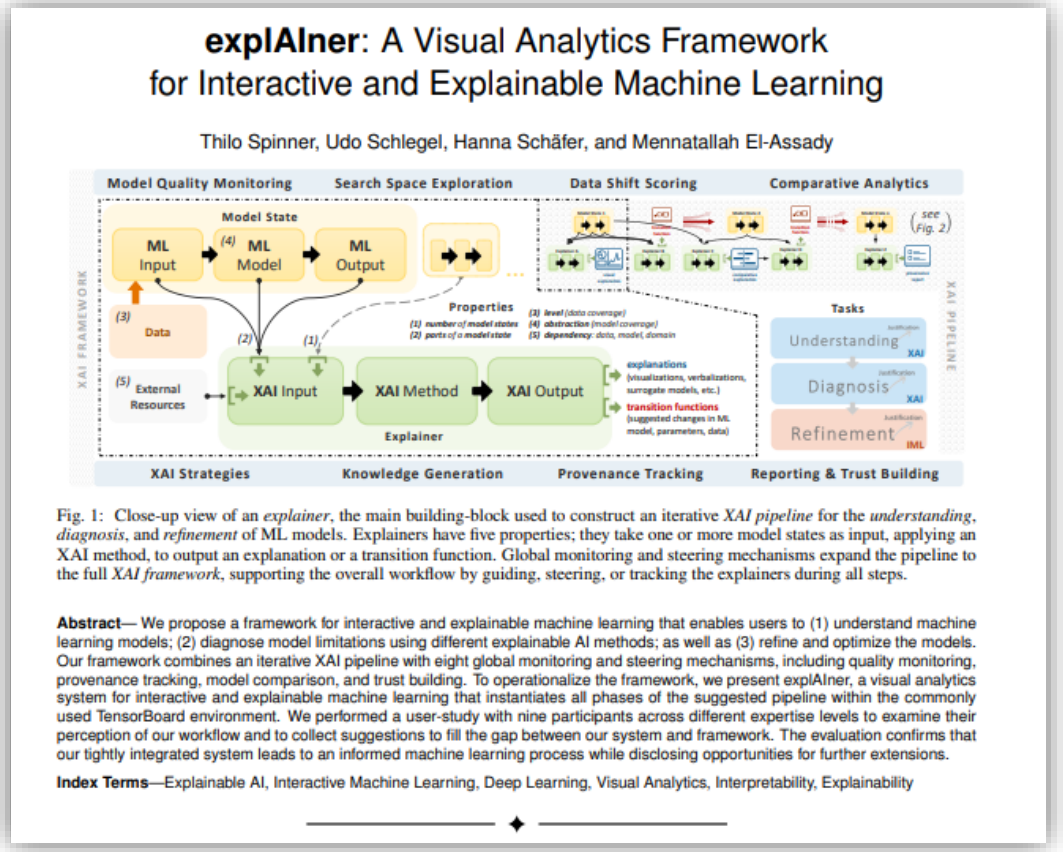
Towards Better Analysis of Machine Learning Models:
A Visual Analytics Perspective

Shixia Liu, Xiting Wang, Mengchen Liu, Jun Zhu
Tsinghua University, Beijing, China

Abstract

Interactive model analysis, the process of understanding, diagnosing, and refining a machine learning model with the help of interactive visualization, is very important for users to efficiently solve real-world artificial intelligence and data mining problems. Dramatic advances in big data analytics has led to a wide variety of interactive model analysis tasks. In this paper, we present a comprehensive analysis and interpretation of this rapidly developing area. Specifically, we classify the relevant work into three categories: understanding, diagnosis, and refinement. Each category is exemplified by recent influential work. Possible future research opportunities are also explored and discussed.

Keywords: interactive model analysis, interactive visualization, machine learning, understanding, diagnosis, refinement



explAIner: A Visual Analytics Framework
for Interactive and Explainable Machine Learning

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady

Fig. 1: Close-up view of an *explainer*, the main building-block used to construct an iterative *XAI pipeline* for the *understanding*, *diagnosis*, and *refinement* of ML models. Explainers have five properties; they take one or more model states as input, applying an XAI method, to output an explanation or a transition function. Global monitoring and steering mechanisms expand the pipeline to the full *XAI framework*, supporting the overall workflow by guiding, steering, or tracking the explainers during all steps.

Abstract— We propose a framework for interactive and explainable machine learning that enables users to (1) understand machine learning models; (2) diagnose model limitations using different explainable AI methods; as well as (3) refine and optimize the models. Our framework combines an iterative XAI pipeline with eight global monitoring and steering mechanisms, including quality monitoring, provenance tracking, model comparison, and trust building. To operationalize the framework, we present explAIner, a visual analytics system for interactive and explainable machine learning that instantiates all phases of the suggested pipeline within the commonly used TensorBoard environment. We performed a user-study with nine participants across different expertise levels to examine their perception of our workflow and to collect suggestions to fill the gap between our system and framework. The evaluation confirms that our tightly integrated system leads to an informed machine learning process while disclosing opportunities for further extensions.

Index Terms—Explainable AI, Interactive Machine Learning, Deep Learning, Visual Analytics, Interpretability, Explainability

Justification → **Understanding**

Justification → **Diagnosis**

Justification → **Refinement**

Understanding

Understanding Model Decisions

Do Convolutional Neural Networks Learn Class Hierarchy?

Bilal Alsallakh, Amin Jourabloo, Mao Ye, Xiaoming Liu, Liu Ren

Fig. 1. The user interface of
(a) The class hierarchy with a
bands indicate the selected c

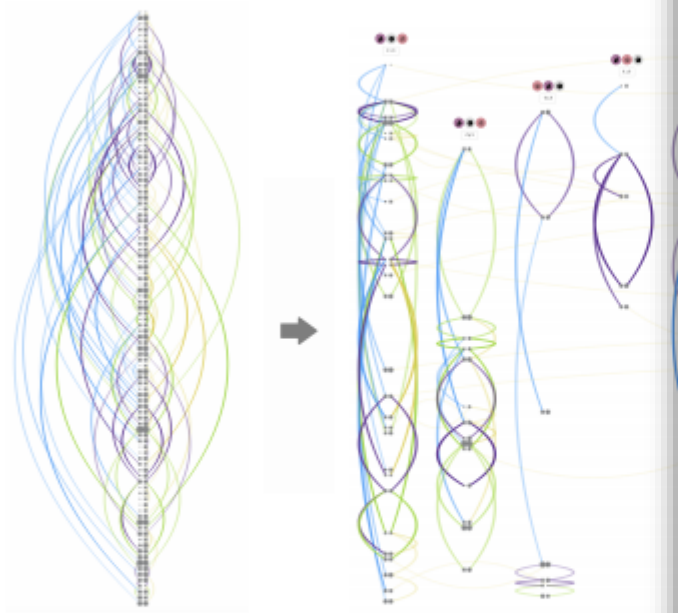
Abstract—
Convolutional Neural Networ
of classes, the accuracy usua
hierarchical structure over the
in relation with CNN-internal
furthermore dictates the learn
can separate high-level group
more epochs to develop spec
key to significant improveme
overfitting. We further demon

Index Terms—Convolutional

ThreadReconstructor: Modeling Reply-Chains to Untangle Conversational Text through Visual Analytics

Mennatallah El-Assady[1,2], Rita Sevastjanova[1]

[1]University of Kons
[2]University of Ontario Institut



(a) *Reply-Relation View* (b) *Thematic-Forest View* showing each

Figure 1: Thematic-Forest (1b) of untangled reply-chains from a full-c
compared to a random-forest model trained on 13 features (right arcs). M

Abstract

*We present ThreadReconstructor, a visual analytics approach for de
discussions, e.g., in political debates and forums. Our work is mo
in massive online conversations and verbatim text transcripts. W
models to generate a basic structure that is enriched by user-defin*

RuleMatrix: Visualizing and Understanding Classifiers with Rules

Yao Ming, Huamin Qu, *Member, IEEE*, and Enrico Bertini, *Member, IEEE*

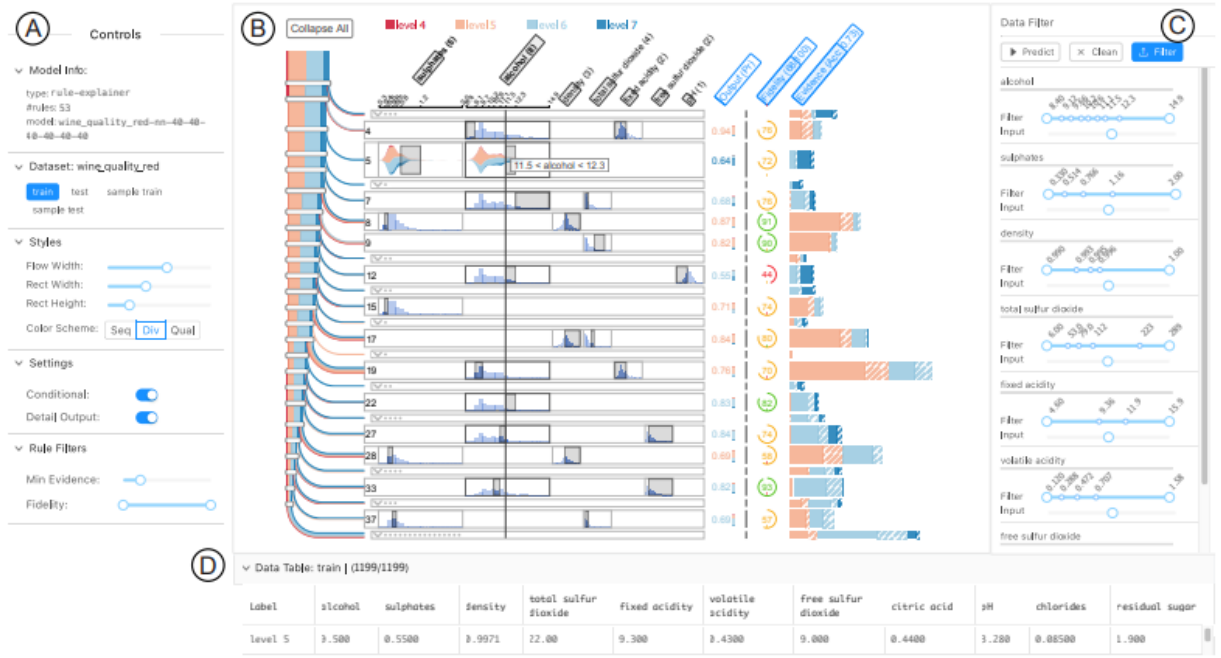


Fig. 1. Understanding the behavior of a trained neural network using the explanatory visual interface of our proposed technique. The user uses the control panel (A) to specify the detail information to visualize (*e.g.*, level of detail, rule filters). The rule-based explanatory representation is visualized as a matrix (B), where each row represents a rule, and each column is a feature used in the rules. The user can also filter the data or use a customized input in the data filter (C) and navigate the filtered dataset in the data table (D).

Abstract—With the growing adoption of machine learning techniques, there is a surge of research interest towards making machine learning systems more transparent and interpretable. Various visualizations have been developed to help model developers understand, diagnose, and refine machine learning models. However, a large number of potential but neglected users are the domain experts with little knowledge of machine learning but are expected to work with machine learning systems. In this paper, we present an interactive

Understanding

Model Architecture and Activity

Visualizing the Hidden Activity of Artificial Neural Networks

Paulo E. Rauber, Samuel G. Fadel, Alexandre X. Falcão, and Alexandru C. Telea

Abstract—In machine lear
ization from examples. Ar
typically used as black-box
observations. In this paper
representations of observa
traditional image classifica
designers. For instance, o
representations, and the pa

Index Terms—Artificial ne

1 Introduction

In machine learning, advances in
building and training (deep) arti
allowed these models to achieve
plications related to pattern reco
training ANNs is generally time
expertise [5].

In data visualization, dimen
cessfully used to compute *pro*
dimensional data in lower-dimen
preserve the data *structure*. This
bution, relationships between poi
When depicted by scatterplots, pr
the original data. Compared to o
tions, such as scatterplot matric
nates, and their variants, project
with respect to the number of dim
able with respect to the number
ter eventually becomes problema
recent dimensionality reduction
dreds of thousands of observatio

In this paper, we demonstrate
duction techniques to provide ins
Although we focus on multilayer

Understanding Hidden Memories of Recurrent Neural Networks

Yao Ming * Shaozu Cao * Ruixiang Zhang * Zhen Li * Yuanzhe Chen *
Yangqiu Song, *Member, IEEE* *

Hong Kong University of Scien

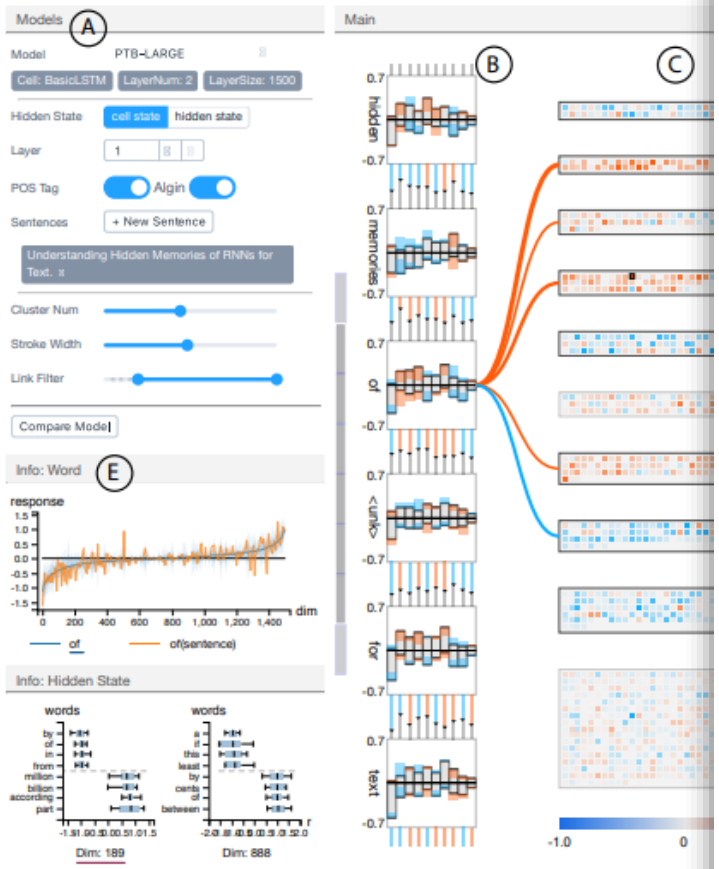


Figure 1: The interface of RNNVis. The control panel (A) shows paramet

GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation

Minsuk Kahng, Nikhil Thorat, Duen Horng (Polo) Chau, Fernanda B. Viégas, and Martin Wattenberg

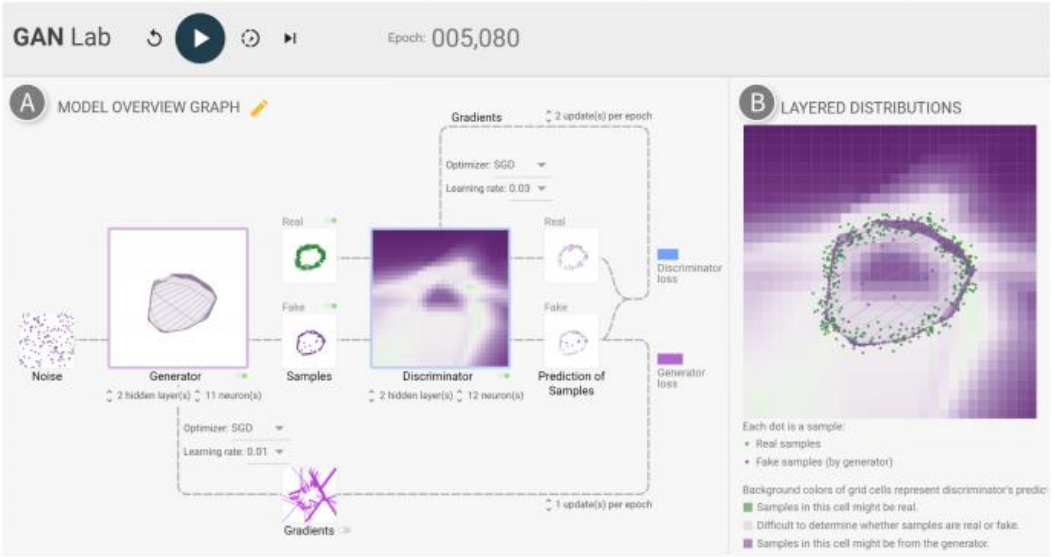


Fig. 1. With GAN Lab, users can interactively train Generative Adversarial Networks (GANs), and visually examine the model training process. In this example, a user has successfully used GAN Lab to train a GAN that generates 2D data points whose challenging distribution resembles a ring. **A.** The *model overview graph* summarizes a GAN model's structure as a graph, with nodes representing the generator and discriminator submodels, and the data that flow through the graph (e.g., fake samples produced by the generator). **B.** The *layered distributions* view helps users interpret the interplay between submodels through user-selected layers, such as the discriminator's classification heatmap, real samples, and fake samples produced by the generator.

Abstract—Recent success in deep learning has generated immense interest among practitioners and students, inspiring many to learn about this new technology. While visual and interactive approaches have been successfully developed to help people more easily learn deep learning, most existing tools focus on simpler models. In this work, we present GAN Lab, the first interactive visualization tool designed for non-experts to learn and experiment with *Generative Adversarial Networks (GANs)*, a popular class of complex deep learning models. With GAN Lab, users can interactively train generative models and visualize the dynamic training process's intermediate results. GAN Lab tightly integrates an *model overview graph* that summarizes GAN's structure, and a *layered distributions* view that helps users interpret the interplay between submodels. GAN Lab introduces new interactive experimentation features for learning complex deep learning models, such as *step-by-step* training at multiple levels of abstraction for understanding intricate training dynamics. Implemented using *TensorFlow.js*, GAN Lab is accessible to anyone via modern web browsers, without the need for

Understanding Model Behavior

LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush
– Harvard School of Engineering and Applied Sciences –

Fig. 1. The LSTMVis user interface. The user interactively *selects* a range of text specifying a hypothesis Select View (a). This range is then used to *match* similar hidden state patterns displayed in the Match View (b by specifying a start-stop range in the text (c) and an activation threshold (t) which leads to a selection of hi The start-stop range can be further constrained using the pattern plot (d). The meta-tracks below depict ex position like POS (e1) or the top K predictions (e2). The tool can then *match* this selection with similar hidd data set of varying lengths (f), providing insight into the representations learned by the model. The match vi user-defined meta-data encoded as heatmaps (g1,g2). The color of one heatmap (g2) can be mapped (h) to t allows the user to see patterns that lead to further refinement of the selection hypothesis. Navigation aids pro

Abstract— Recurrent neural networks, and in particular long short-term memory (LSTM) networks, are a rem sequence modeling that learn a dense black-box hidden representation of their sequential input. Researc understanding these models have studied the changes in hidden state representations over time and noti patterns but also significant noise. In this work, we present LSTMVis, a visual analysis tool for recurrent focus on understanding these hidden state dynamics. The tool allows users to select a hypothesis input rang changes, to match these states changes to similar patterns in a large data set, and to align these results wit from their domain. We show several use cases of the tool for analyzing specific hidden state properties on da phrase structure, and chord progressions, and demonstrate how the tool can be used to isolate patterns for fu We characterize the domain, the different stakeholders, and their goals and tasks. Long-term usage data aft revealed great interest in the machine learning community.

SEQ2SEQ-VIS : A Visual Debugging Tool for Sequence-to-Sequence Models

Hendrik Strobelt*, Sebastian Gehrmann*, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush

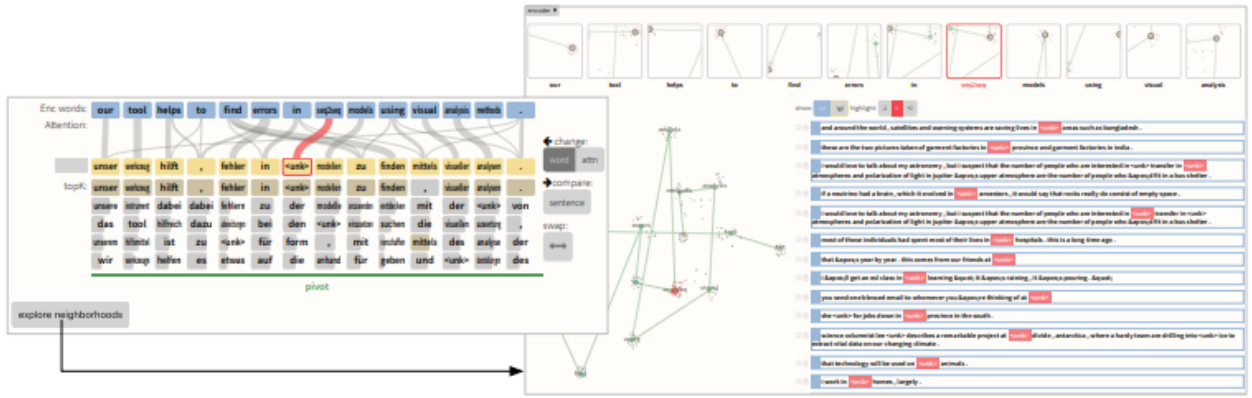


Fig. 1. Example of `Seq2Seq-Vis`. In the translation view (left), the source sequence *"our tool helps to find errors in seq2seq models using visual analysis methods."* is translated into a German sentence. The word *"seq2seq"* has correct attention between encoder and decoder (red highlight) but is not part of the language dictionary. When investigating the encoder neighborhoods (right), the user sees that *"seq2seq"* is close to other unknown words *"⟨unk⟩"*. The buttons enable user interactions for deeper analysis.

Abstract—Neural sequence-to-sequence models have proven to be accurate and robust for many sequence prediction tasks, and have become the standard approach for automatic translation of text. The models work with a five-stage blackbox pipeline that begins with encoding a source sequence to a vector space and then decoding out to a new target sequence. This process is now standard, but like many deep learning methods remains quite difficult to understand or debug. In this work, we present a visual analysis tool that allows interaction and "what if"-style exploration of trained sequence-to-sequence models through each stage of the translation process. The aim is to identify which patterns have been learned, to detect model errors, and to probe the model with counterfactual scenario. We demonstrate the utility of our tool through several real-world sequence-to-sequence use cases on large-scale models.

Diagnosis

Diagnosis

Model Outputs

A Workflow for Visual Diagnostics of Binary Classifiers using Instance-Level Explanations

Josua Krause*
NYU Tandon
School of Engineering

Aritra Dasgupta†
Pacific No
National La

Jordan Swartz‡

Yindalon Aphinyanaphongs§

Enrico Bertini¶

Abstract

Human-in-the-loop data analysis appl
transparency in machine learning mod
and trust their decisions. To this end,
ics workflow to help data scientists a
diagnose, and understand the decisions
The approach leverages "instance-level
local feature relevance that explain sin
to build a set of visual representations
investigation. The workflow is based
sentations and steps: one based on ag
data distributes across correct / incorr
explanations to understand which feat
decisions; and one based on raw data,
tial root causes for the observed patter
from a long-term collaboration with a
and healthcare professionals who use
of machine learning models they deve
this collaboration demonstrates that th
experts derive useful knowledge about t
it describes, thus experts can generate
model can be improved.

Keywords: Machine Learning, Interp

1 Introduction

In this paper we propose an interactive
interface to help data scientists and d

Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models

Jiawei Zhang, Yang Wang, Piero Mo

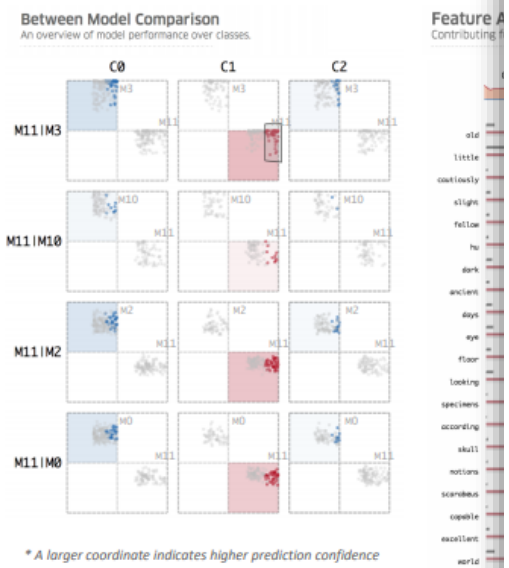
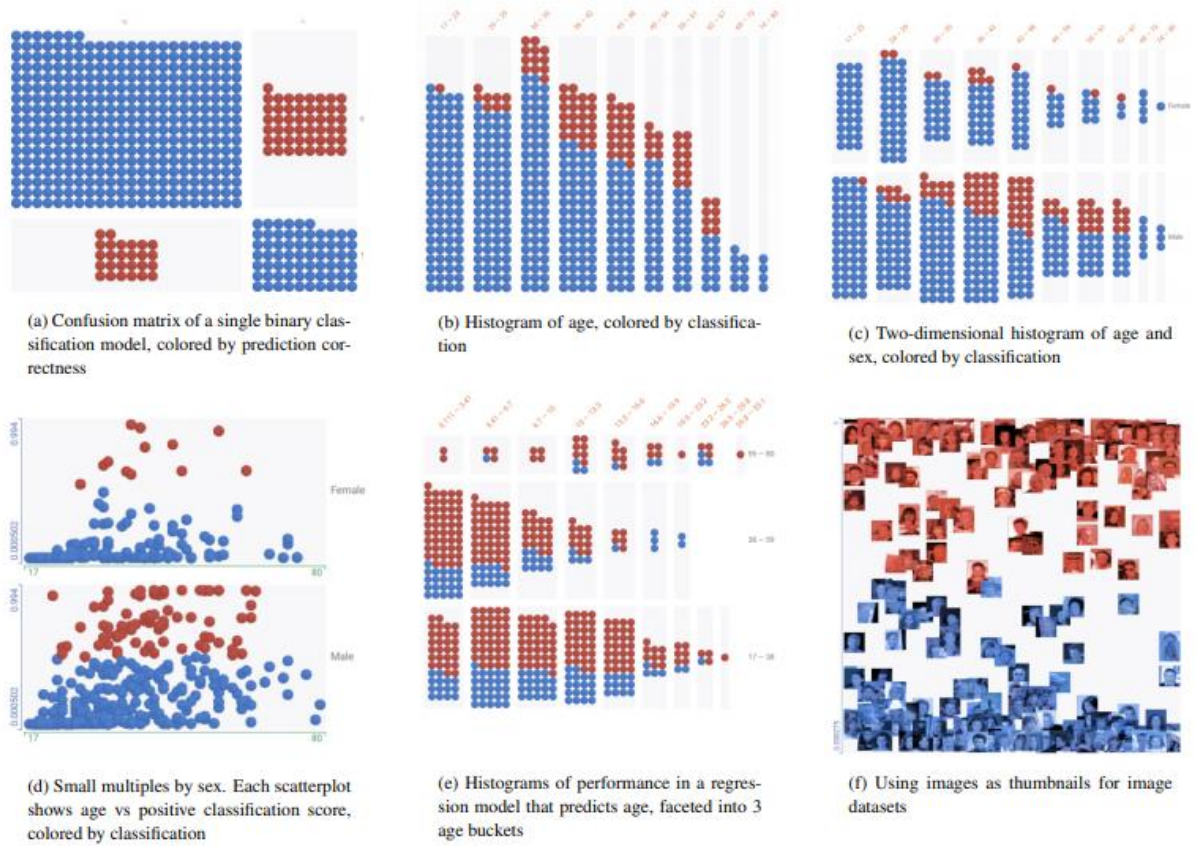


Fig. 1. Manifold consists of two interactive dialogs: a model
pairs using a small multiple design, and a local feature interp
subsets (c) and provides a similarity measure (b) of feature
most discriminative features among different subsets, i.e., so

Abstract— Interpretation and diagnosis of machine learning
in new approaches. We present Manifold, a framework tha
and comparison of machine learning models in a more tra
on visualizing the internal logic of a specific model type (i.e
scenario where different model types are integrated. To this
access the internal logic of the model and solely observes th
and probability distribution). We describe the workflow of M
commonly involved in the model development and diagnosis
(verification). The visual components supporting these task
outcome and a customizable tabular view that reveals featur

The What-If Tool: Interactive Probing of Machine Learning Models

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson



(a) Confusion matrix of a single binary classification model, colored by prediction correctness

(b) Histogram of age, colored by classification

(c) Two-dimensional histogram of age and sex, colored by classification

(d) Small multiples by sex. Each scatterplot shows age vs positive classification score, colored by classification

(e) Histograms of performance in a regression model that predicts age, faceted into 3 age buckets

(f) Using images as thumbnails for image datasets

Fig. 1: Analyzing two different data sets in the What-If Tool: US Census Income dataset from UCI (a) to (e), and CelebA dataset

Diagnosis

Robustness

Analyzing the Noise Robustness of Deep Neural Networks

Mengchen Liu*, Shixia Liu*, Hang Su†, Kelei Cao*, Jun Zhu†

*School of Software, TNList Lab, State Key Lab for Intell. Tech. Sys., Tsing
†Dept. of Comp. Sci.Tech., TNList Lab, State Key Lab for Intell. Tech. Sys., CBICR Ce

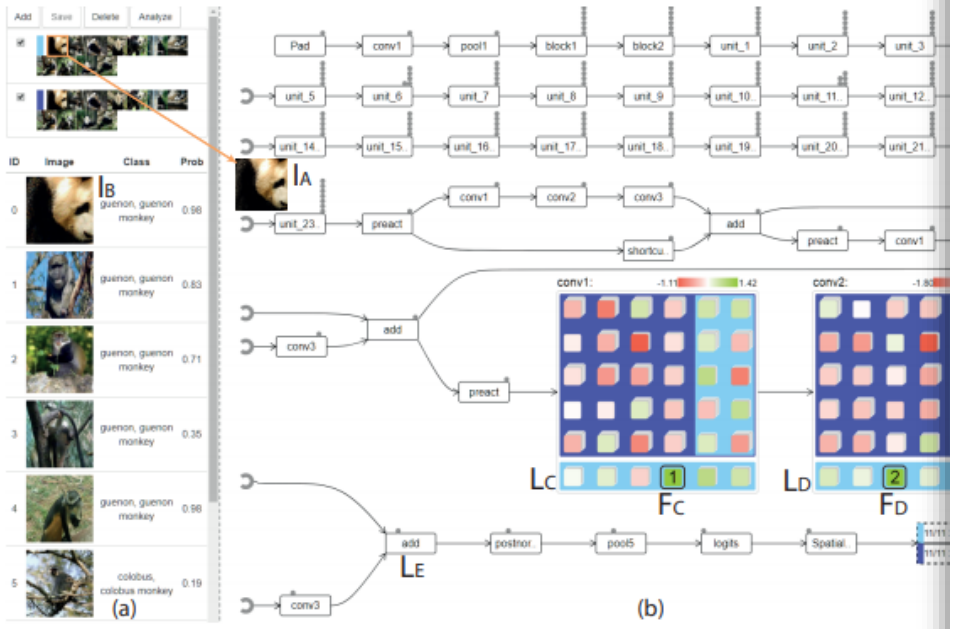


Figure 1: Explaining the misclassification of adversarial panda images. The root cause is that the
the adversarial examples (F_C), which leads to the failure of detecting a panda's face (F_D). As a
are misclassified: (a) input images; (b) datapath visualization at the layer and feature map levels

Explaining Vulnerabilities to Adversarial Machine Learning through Visual Analytics

Yuxin Ma, Tiankai Xie, Jundong Li, Ross Maciejewski, *Senior Member, IEEE*

Fig. 1. Reliability attack on spam filters. (1) Poisoning instance #40 has the largest impact on the recall value, which is (2) also depicted in the model overview. (3) There is heavy overlap among instances in the two classes as well the poisoning instances. (4) Instance #40 has been successfully attacked causing a number of innocent instances to have their labels flipped. (5) The flipped instances are very close to the decision boundary. (6) On the feature of words "will" and "email", the variances of poisoning instances are large. (7) A sub-optimal target (instance #80) has less impact on the recall value, but the cost of insertions is 40% lower than that of instance #40.

Abstract— Machine learning models are currently being deployed in a variety of real-world applications where model predictions are used to make decisions about healthcare, bank loans, and numerous other critical tasks. As the deployment of artificial intelligence technologies becomes ubiquitous, it is unsurprising that adversaries have begun developing methods to manipulate machine learning models to their advantage. While the visual analytics community has developed methods for opening the black box of machine learning models, little work has focused on helping the user understand their model vulnerabilities in the context of adversarial attacks. In this

Diagnosis

Activation

SUMMIT: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations

Fred Hohman, Haekyu Pa

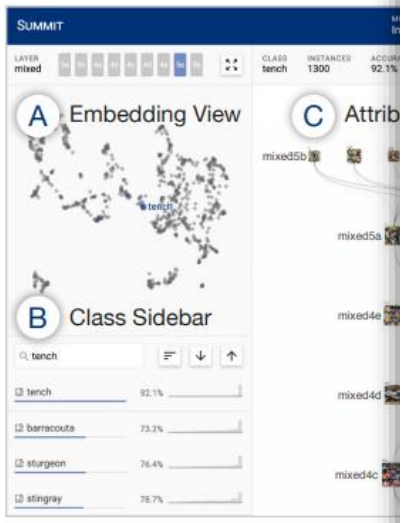


Fig. 1. With Summit, users can scalably summariz
network detects and *how* they are related. In this e
However, SUMMIT reveals surprising associations i
"tench" prediction is dependent on an intermediate "h
like *"scales," "person,"* and *"fish"*. **(A) Embedding V**
(B) Class Sidebar enables users to search, sort,
highly activated neurons as vertices ("scales," "fish"

Abstract—Deep learning is increasingly used in de
predictions remains a fundamental challenge. Exis
explaining predictions for single images or neurons.
millions of images, such explanations can easily mi
systematically summarizes and visualizes what feat
predictions. SUMMIT introduces two new scalable s
and (2) *neuron-influence aggregation* identifies rel
the novel *attribution graph* that reveals and summa
outcomes. SUMMIT scales to large data, such as
visualization and dataset examples to help users di
We present neural network exploration scenarios
large-scale image classifier's learned representation
runs in modern web browsers and is open-sourced.

Index Terms—Deep learning interpretability, visual

DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks

Junpeng Wang, Liang Gou, Han-Wei S

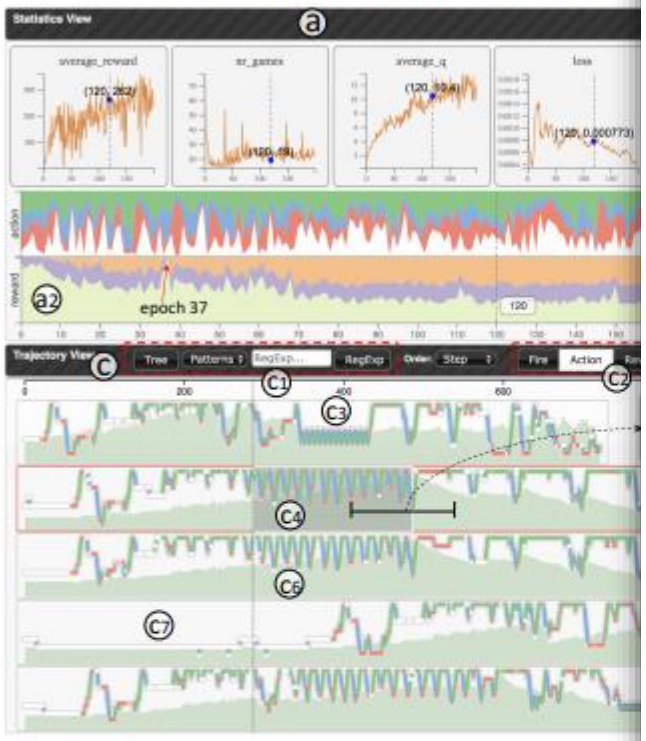


Fig. 1. *DQNViz*: (a) the *Statistics* view presents the overall training
view shows epoch-level statistics with pie charts and stacked bar c
patterns of the DQN agent in different episodes; (d) the *Segment* vie

Abstract— Deep Q-Network (DQN), as one type of deep reinforc
acquires optimal actions while interacting with an environment. The
players across many Atari 2600 games. Despite the superhuman pe
the sophisticated behaviors of the DQN agent remain to be challen
large number of experiences dynamically generated by the agent.

ACTIVIS: Visual Exploration of Industry-Scale Deep Neural Network Models

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng (Polo) Chau

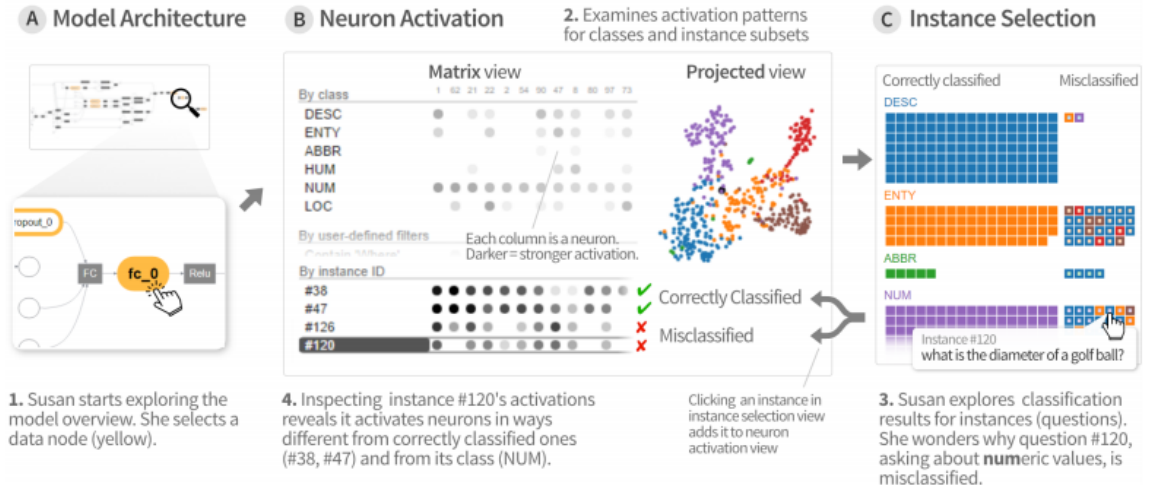


Fig. 1. ACTIVIS integrates several coordinated views to support exploration of complex deep neural network models, at both instance- and subset-level. **1.** Our user Susan starts exploring the model architecture, through its *computation graph* overview (at A). Selecting a *data node* (in yellow) displays its *neuron activations* (at B). **2.** The *neuron activation matrix view* shows the activations for instances and instance subsets; the *projected view* displays the 2-D projection of instance activations. **3.** From the *instance selection* panel (at C), she explores individual instances and their classification results. **4.** Adding instances to the matrix view enables comparison of activation patterns across instances, subsets, and classes, revealing causes for misclassification.

Abstract— While deep learning models have achieved state-of-the-art accuracies for many prediction tasks, understanding these models remains a challenge. Despite the recent interest in developing visual tools to help users interpret deep learning models, the complexity and wide variety of models deployed in industry, and the large-scale datasets that they used, pose unique design challenges that are inadequately addressed by existing work. Through participatory design sessions with over 15 researchers and engineers at Facebook, we have developed, deployed, and iteratively improved ACTIVIS, an interactive visualization system for interpreting large-scale deep learning models and results. By tightly integrating multiple coordinated views, such as a *computation graph* overview

Diagnosis Processes

Analyzing the Training Processes of Deep Generative Models

Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, Shixia Liu

Fig. 1. DGMTracker, a visual analytics tool that helps experts understand and diagnos
models (DGMs): (a) the loss changes; (b) the data flow visualization to illustrate how data
neurons influence the output of the neuron of interest; (c) visualization of the training dyn

Abstract— Among the many types of deep models, deep generative models (DGMs) p
unsupervised and semi-supervised learning. However, training DGMs requires more sk
training is more complex than other types of deep models such as convolutional neural ne
approach for better understanding and diagnosing the training process of a DGM. To
process, we first extract a large amount of time series data that represents training dyna
blue-noise polyline sampling scheme is then introduced to select time series samples,
visual clutter. To further investigate the root cause of a failed training process, we propose
how other neurons contribute to the output of the neuron causing the training failure. Tw
learning experts to demonstrate how our approach helps understand and diagnose the
how our approach can be directly used to analyze other types of deep models, such as

Index Terms—deep learning, deep generative models, blue noise sampling, credit assi

Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution

Mennatallah El-Assady[1,2], Fabian Sperrle[1], Oliver Deussen[1], Daniel Keim[1], and Christopher Collins[2]

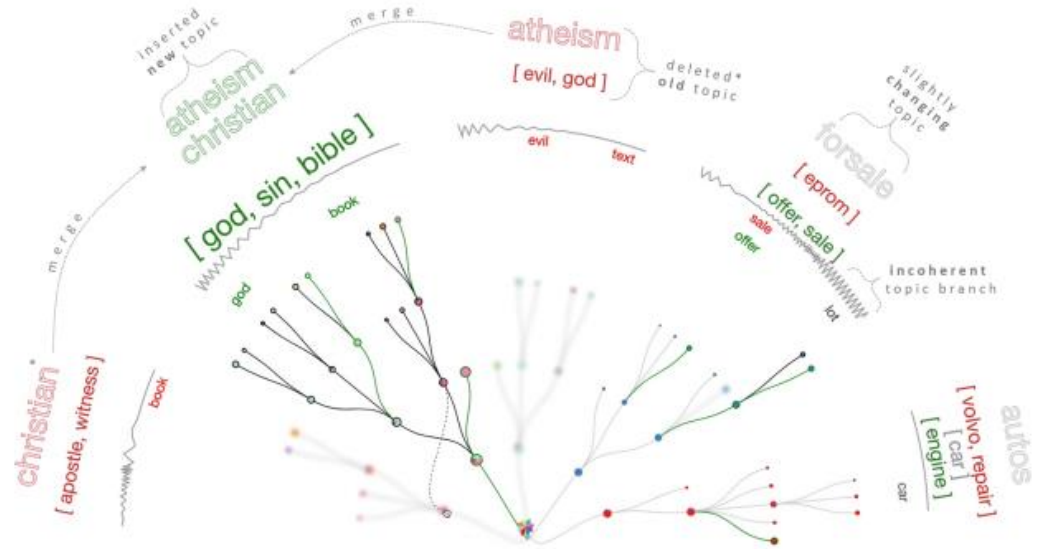


Fig. 1: The *Tree-Speculation View* is used to compare two topic models and shows the differences. Deleted branches are blurred, while **moved**, **newly added** and **removed** nodes and keywords are highlighted. To efficiently guide users towards perceivable model quality improvements, our system automatically proposes optimizations like the merge of two topics depicted here. By visualizing model uncertainties and low quality topics, we foster trust in the model and empower users to directly address these shortcomings.

Abstract— To effectively assess the potential consequences of human interventions in model-driven analytics systems, we establish the concept of *speculative execution* as a visual analytics paradigm for creating user-steerable preview mechanisms. This paper presents an explainable, mixed-initiative topic modeling framework that integrates speculative execution into the algorithmic decision-making process. Our approach visualizes the model-space of our novel incremental hierarchical topic modeling algorithm, unveiling its inner-workings. We support the active incorporation of the user's domain knowledge in every step through explicit model manipulation interactions. In addition, users can initialize the model with expected topic seeds, the backbone priors. For a more targeted optimization, the modeling process automatically triggers a speculative execution of various optimization strategies, and requests feedback whenever

Diagnosis

Fairness

FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning

Ángel Alexander Cabrera Will Epperson Fred Hohman Minsuk Kahng
Jamie Morgenstern Duen Horng (Polo) Chau*

Georgia Institute of Technology

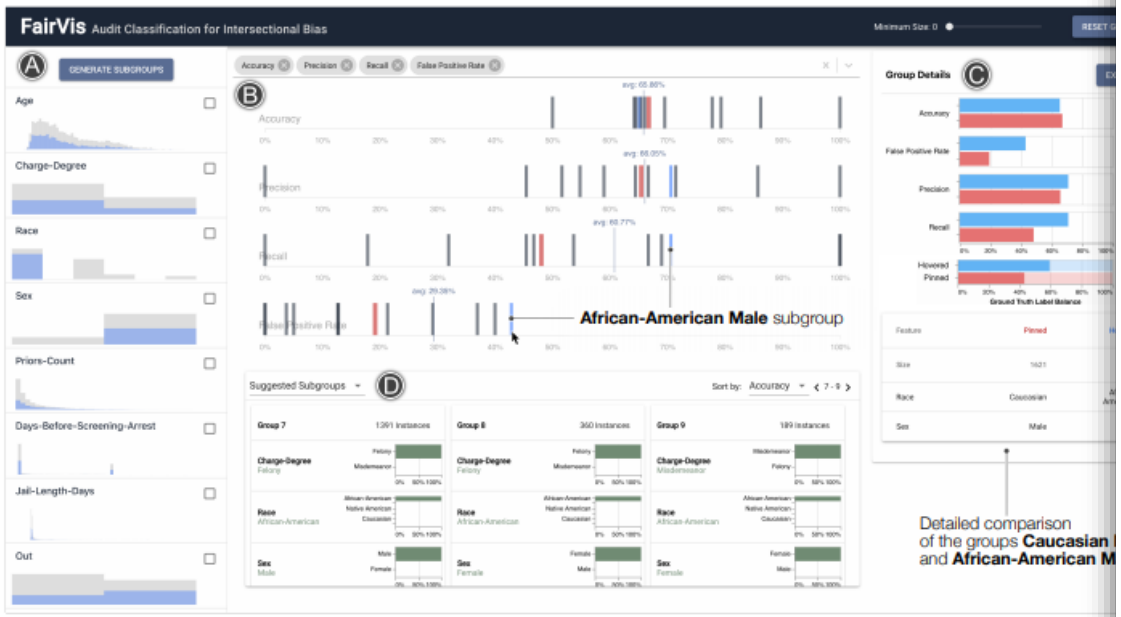


Figure 1: FairVis integrates multiple coordinated views for discovering intersectional bias. Above, our user investig intersectional subgroups of *sex* and *race*. **A.** The *Feature Distribution View* allows users to visualize each feature's distrib generate subgroups. **B.** The *Subgroup Overview* lets users select various fairness metrics to see the global average p and compare subgroups to one another, e.g., pinned **Caucasian Males** versus hovered **African-American Males**. The *Recall* and *False Positive Rate* show that for African-American Males, the model has relatively high recall but also the hig positive rate out of all subgroups of sex and race. **C.** The *Detailed Comparison View* lets users compare the details of tw and investigate their class balances. Since the difference in False Positive Rates between Caucasian Males and African- Males is far larger than their difference in base rates, a user suspects this part of the model merits further inquiry. **D.** The *S and Similar Subgroup View* shows suggested subgroups ranked by the worst performance in a given metric.

FairSight: Visual Analytics for Fairness in Decision Making

Yongsu Ahn, Yu-Ru Lin

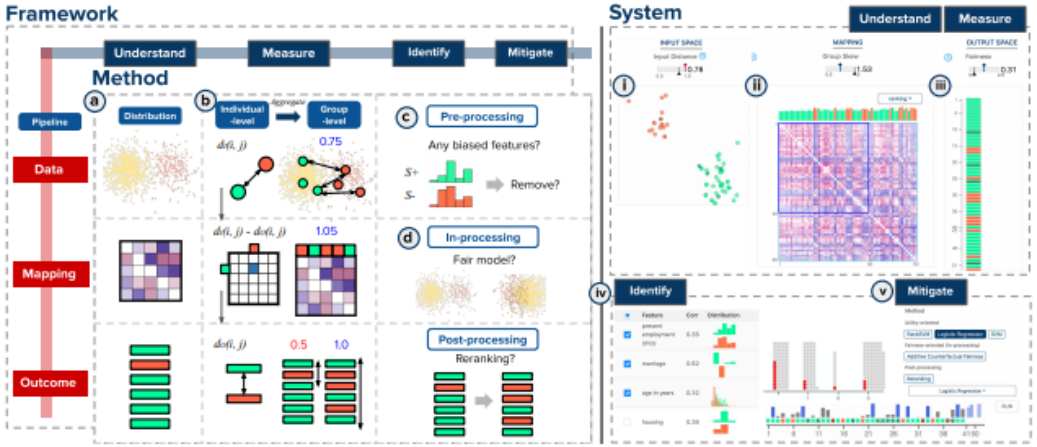


Fig. 1. We propose a design framework to protect individuals and groups from discrimination in algorithm-assisted decision making. A visual analytic system, *FairSight*, is implemented based on our proposed framework, to help data scientists and practitioners make fair decisions. The decision is made through ranking individuals who are either members of a protected group (orange bars) or a non-protected group (green bars). (a) The system provides a pipeline to help users understand the possible bias in a machine learning task as a *mapping* from the *input* space to the *output* space. (b) Different notions of fairness – *individual* fairness and *group* fairness – are measured and summarized numerically and visually. For example, the individual fairness is quantified by how pairwise distances between individuals are preserved through the mapping. The group fairness is quantified by the extent to which it leads to fair outcome distribution across groups, with (i) a 2D plot, (ii) a color-coded matrix , and (iii) a ranked-list plot capturing the pattern of potential biases. The system provides diagnostic modules to help (iv) identify and (v) mitigate biases through (c) investigating features before running a model, and (d) leveraging fairness-aware algorithms during and after the training step.

Abstract— Data-driven decision making related to individuals has become increasingly pervasive, but the issue concerning the potential discrimination has been raised by recent studies. In response, researchers have made efforts to propose and implement fairness measures and algorithms, but those efforts have not been translated to the real-world practice of data-driven decision making. As such, there is still an urgent need to create a viable tool to facilitate fair decision making. We propose *FairSight*, a visual analytic system to address this need; it is designed to achieve different notions of fairness in ranking decisions through identifying the required actions – understanding, measuring, diagnosing and mitigating biases – that together lead to fairer decision making. Through a case study and user study, we demonstrate that the proposed visual analytic and diagnostic modules in the system are effective in understanding the fairness-aware decision pipeline and obtaining more fair outcomes.

Index Terms—Fairness in Machine Learning, Visual Analytic

Refinement

Refinement

Single Model

DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks

Nicola Pezzotti, Thomas Höllt, Jan

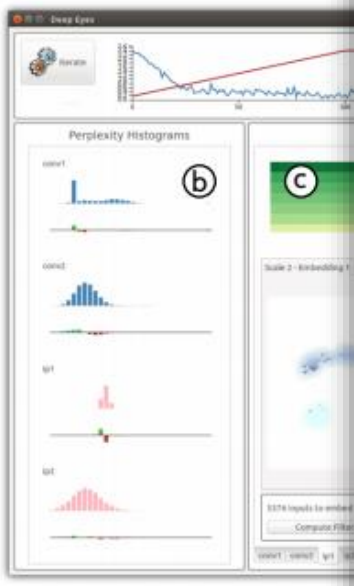


Fig. 1. **DeepEyes** is a Progressive Visua
on the training is given by the commonly
that allows the detection of stable layers.
filters are detected in the Activation Heatm
relationships among the filters in a layer a

Abstract—Deep neural networks are now
classifiers, where features are handcrafte
handcrafting the features, it is now the net
the number of layers or the number of filter
design guidelines exist, designing a neur
due to the large datasets used for training
the design of neural networks during trair
stable set of patterns and, therefore, are o
superfluous filters or layers, and informatio
through multiple use cases, showing how

Index Terms—Progressive visual analytic

RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records

Bum Chul Kw
Young Bin Kim

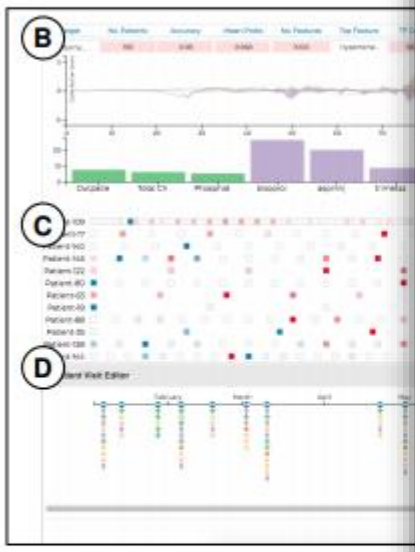


Fig. 1. A screenshot of RetainVis consisti
summary view (right) of patients. (B) *Patien*
individual patients in a row of rectangles.
Users can open (D) *Patient Editor* to condu

Abstract— We have recently seen many su
(EMRs), which contain histories of patients'
states of patients. Despite the strong perfo
particular prediction. Such *black-box* natu
established methods to interactively leverag
our design study aims to provide a visual a
medical experts, artificial intelligence scient
experts, we design, implement, and evaluat
and interactive RNN-based model called R
tasks. Our study shows the effective use of
predictions, using EMRs of patients with he

Towards Better Analysis of Deep Convolutional Neural Networks

Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, Shixia Liu

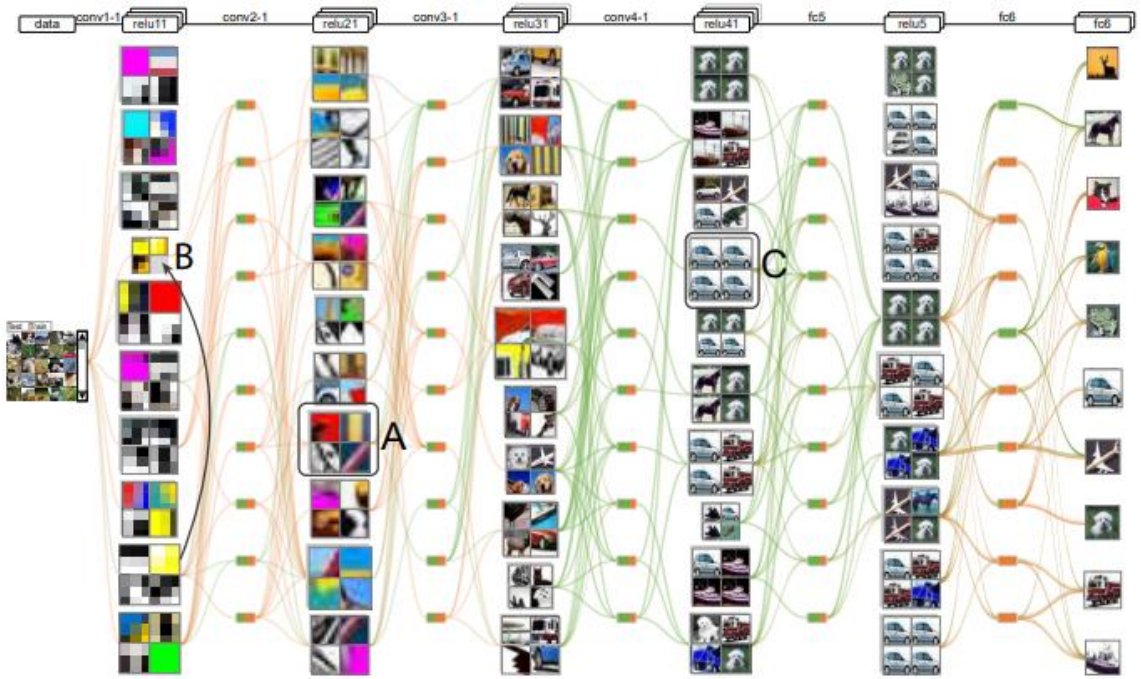


Fig. 1. CNNVis, a visual analytics toolkit that helps experts understand, diagnose, and refine deep CNNs.

Abstract— Deep convolutional neural networks (CNNs) have achieved breakthrough performance in many pattern recognition tasks such as image classification. However, the development of high-quality deep models typically relies on a substantial amount of

Refinement

Multi Model

DeepCompare: Visual and Interactive Comparison of Deep Learning Model Performan

Sugeerth Murugesan
University of California, Davis

Tuan Manh La
Purdue Universi

Sana Malik, Fan Du, and Eunyee Koh
Adobe Research

***Abstract*—Deep learning models have become the state-of**
text sentiment analysis to facial image recognition. Howev
certain models perform better than others or how one mode
another is often difficult yet critical for increasing their effe
prediction accuracy, and enabling fairness. Traditional me
efficacy, such as accuracy, precision, and recall provide a
performance; however, the qualitative intricacies of why o
than another are hidden. In this paper, we interview machin
understand their evaluation and comparison workflow. Fro
a visual analytic approach, DeepCompare, to systematicall
deep learning models, in order to provide insight into the m
interactively assess tradeoffs between two such models. T
evaluate model results, identify and compare activation pa

Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework

Mennatallah El-Assady[1,2], Rita Sevastjanova[1], Fabian Sperrle[1], Daniel Keim[1], and Christopher Collins[2]

[1]University of Konstanz, Germany
[2]University of Ontario Institute of Technology, Canada

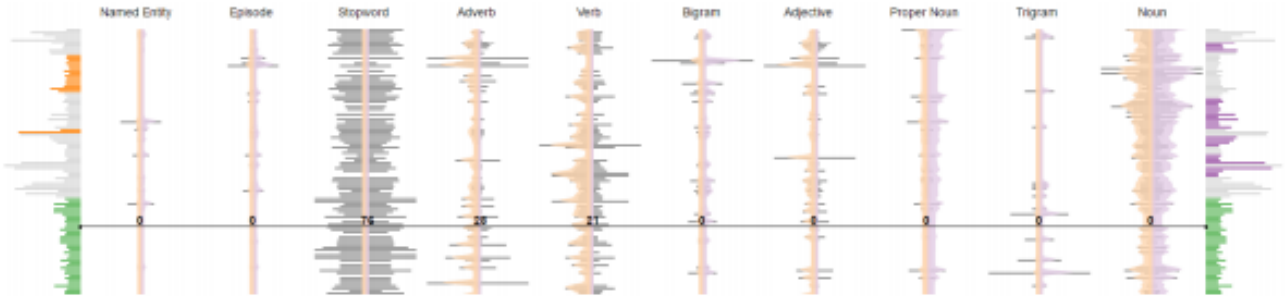


Fig. 1. *Parameter Distribution View* using comparative bar charts. This compact visualization technique enhances the comparison of two parameter distributions using mirrored bar-charts as a baseline and two asymmetrical violin-style plots as distribution estimates. The plots are scaled using the ratio between the two compared assortments (on both sides). The larger value is scaled to the full width of the baseline and the smaller value is scaled proportionally. This figure depicts the comparison of the utterance descriptor features of the second US presidential debate between Obama and Romney in 2012. All utterances are sorted according to their topic coherence.

Abstract— Topic modeling algorithms are widely used to analyze the thematic composition of text corpora but remain difficult to interpret and adjust. Addressing these limitations, we present a modular visual analytics framework, tackling the understandability and adaptability of topic models through a user-driven reinforcement learning process which does not require a deep understanding of the underlying topic modeling algorithms. Given a document corpus, our approach initializes two algorithm configurations based on a parameter space analysis that enhances document separability. We abstract the model complexity in an interactive visual workspace for exploring the automatic matching results of two models, investigating topic summaries, analyzing parameter distributions, and reviewing documents. The main contribution of our work is an iterative decision-making technique in which users provide a document-based relevance feedback that allows the framework to converge to a user-endorsed topic distribution. We also report feedback from a

Refinement

Iterative

Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making

Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, Michael Terry
Google Brain, Google Health
Mountain View, CA
{cjcai,ereif,hegde,hipp,beenkim,smilkov,wattenberg,viegas,gcorrado,mstumpe,m

ABSTRACT

Machine learning (ML) is increasingly being used in image retrieval systems for medical decision making. One application of ML is to retrieve visually similar medical images from past patients (e.g. tissue from biopsies) to reference when making a medical decision with a new patient. However, no algorithm can perfectly capture an expert's ideal notion of similarity for every case: an image that is algorithmically determined to be similar may not be medically relevant to a doctor's specific diagnostic needs. In this paper, we identified the needs of pathologists when searching for similar images retrieved using a deep learning algorithm, and developed tools that empower users to cope with the search algorithm on-the-fly, communicating what types of similarity are most important at different moments in time. In two evaluations with pathologists, we found that these refinement tools increased the diagnostic utility of images found and increased user trust in the algorithm. The tools were preferred over a traditional interface, without a loss in diagnostic accuracy. We also observed that users adopted new strategies when using refinement tools, re-purposing them to test and understand the underlying algorithm and to disambiguate ML errors from their own errors. Taken together, these findings inform future human-ML collaborative systems for expert decision-making.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**;

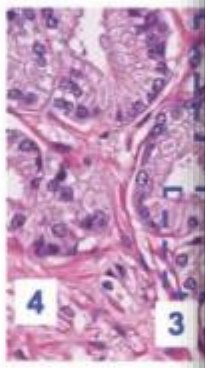


Figure 1: Medical image features, such as cellular interaction between cor (4), and many more. It c search algorithm to perf similarity, because what important differs from dependent.

ACM Reference Format:
Carrie J. Cai, Emily Reif, N
Daniel Smilkov, Martin Wat
rado, Martin C. Stumpe, M
Tools for Coping with Imper
Making. In *CHI Conference*
tems Proceedings (CHI 2019
ACM, New York, NY, USA, 1
3300234

Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections

Mennatallah El-Assady[1,2], Rebecca Kehlbeck[1], Christopher Collins[2], Daniel Keim[1], and Oliver Deussen[1]
[1] University of Konstanz, Germany. [2] University of Ontario Institute of Technology, Canada.

Fig. 1: Guided relevance feedback for the targeted refinement of incoherent areas in the *Semantic Conecpt Space*. This user guidance component tours through the space and highlights potentially uncertain areas, suggesting a recommended action for refinement.

Abstract— We present a framework that allows users to incorporate the semantics of their domain knowledge for topic model refinement while remaining model-agnostic. Our approach enables users to (1) *understand* the semantic space of the model, (2) *identify* regions of potential conflicts and problems, and (3) *readjust* the semantic relation of concepts based on their understanding, directly influencing the topic modeling. These tasks are supported by an interactive visual analytics workspace that uses word-embedding projections to define *concept regions* which can then be refined. The user-refined concepts are independent of a particular document collection and can be transferred to related corpora. All user interactions within the concept space directly affect the semantic relations of the underlying vector space model which, in turn, change the topic modeling. In addition to direct manipulation, our system guides the users' decision-making process through recommended interactions that point out potential improvements. This targeted refinement aims at minimizing

Beyond VIS

Minions, Sheep, and Fruits: Metaphorical Narratives to Explain Artificial Intelligence and Build Trust

Wolfgang Jentner[1],

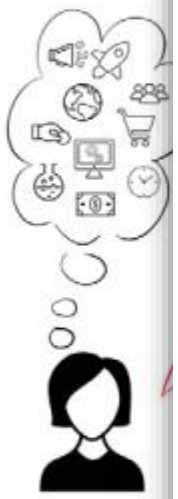


Fig. 1. A transitive
the complexity of t

Abstract— Advan
bringing along the
the underlying arti
ineffective and hesi
Therefore, we prop
and domain expert
complex models to
building and an ade
is known to provide
fields and theories,

Shall we play? – Extending the Visual Analytics Design Space through Gameful Design Concepts

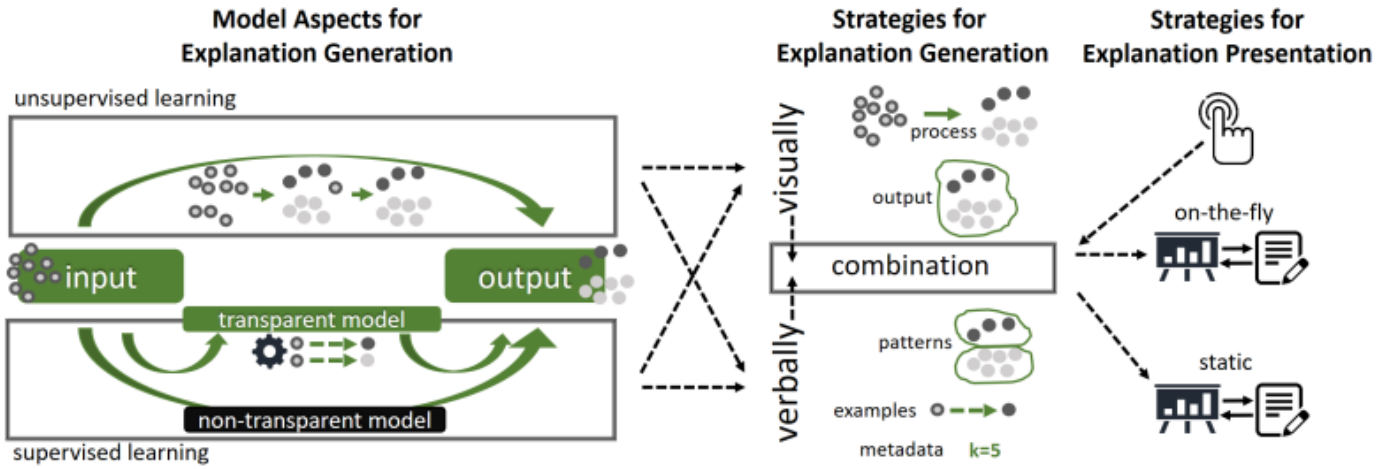
Rita Sevastjanova*
University of Konstanz

Figure 1: We describe a model for
When does the challenging task oc
the users do to design an engaging
Why do people do those challengi
have impact, and to be accepted.
What are the **game mechanics** su

Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models

Rita Sevastjanova[1], Fabian Beck[2], Basil Ell[3], Cagatay Turkay[4], Rafael Henkin[4], Miriam Butt[1], Daniel Keim[1], Mennatallah El-Assady[1,5]

[1]University of Konstanz, Germany
[2]University of Duisburg-Essen, Germany
[3]Bielefeld University, Germany
[4]City University of London, UK
[5]University of Ontario Institute of Technology, Canada



explAIner

The framework for explainable AI and interactive machine learning. Making XAI accessible.

https://explainer.ai/

explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning

Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady

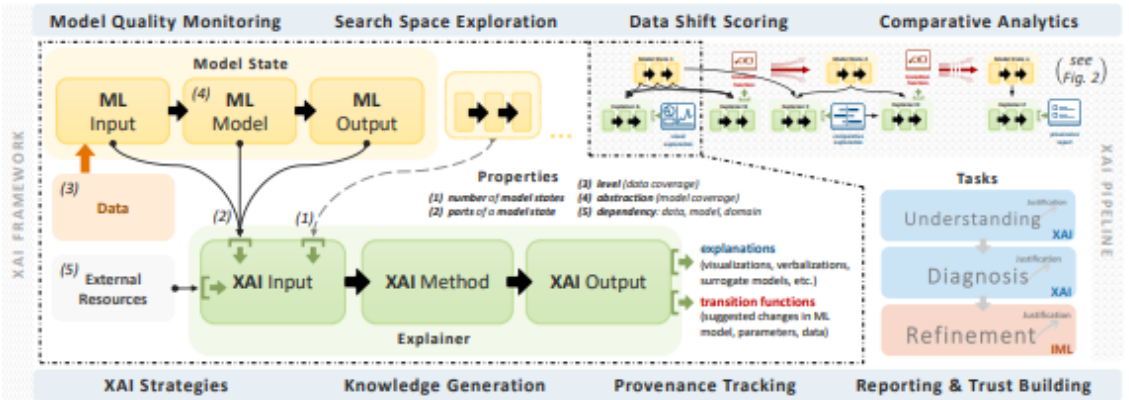
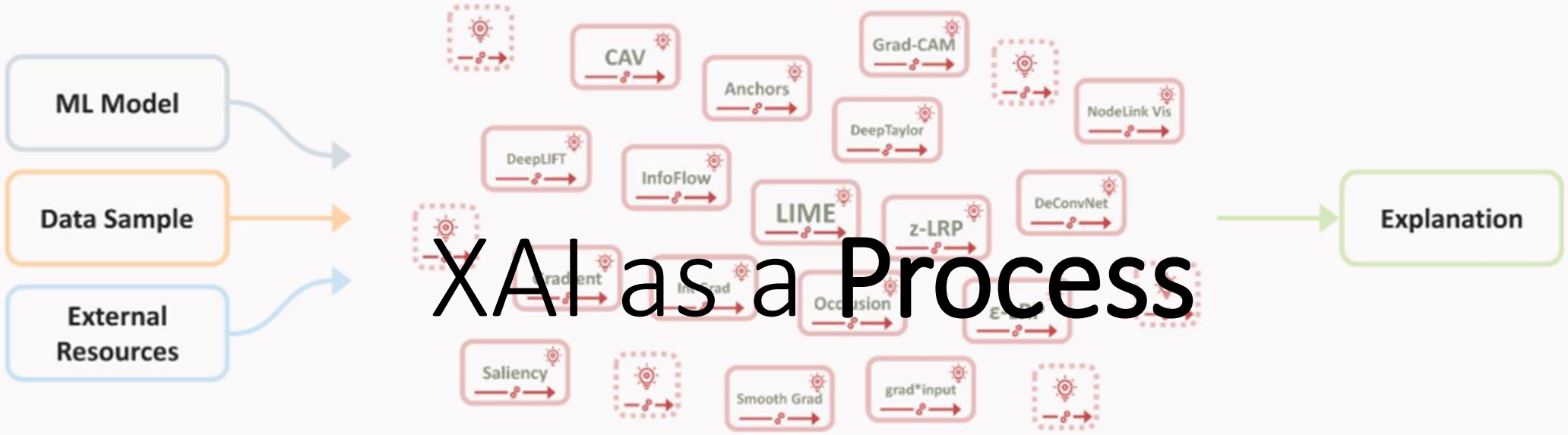


Fig. 1: Close-up view of an *explainer*, the main building-block used to construct an iterative *XAI pipeline* for the *understanding*, *diagnosis*, and *refinement* of ML models. Explainers have five properties; they take one or more model states as input, applying an XAI method, to output an explanation or a transition function. Global monitoring and steering mechanisms expand the pipeline to the full *XAI framework*, supporting the overall workflow by guiding, steering, or tracking the explainers during all steps.

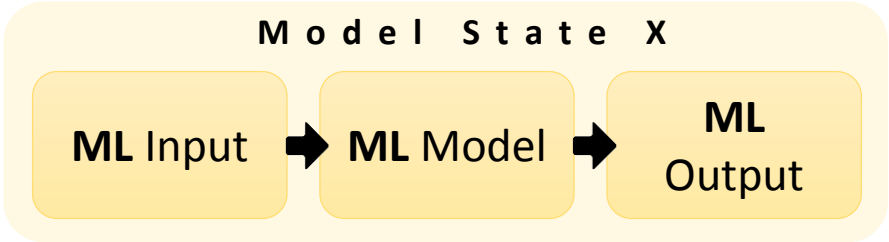
Abstract— We propose a framework for interactive and explainable machine learning that enables users to (1) understand machine learning models; (2) diagnose model limitations using different explainable AI methods; as well as (3) refine and optimize the models. Our framework combines an iterative XAI pipeline with eight global monitoring and steering mechanisms, including quality monitoring, provenance tracking, model comparison, and trust building. To operationalize the framework, we present explAIner, a visual analytics system for interactive and explainable machine learning that instantiates all phases of the suggested pipeline within the commonly used TensorBoard environment. We performed a user-study with nine participants across different expertise levels to examine their perception of our workflow and to collect suggestions to fill the gap between our system and framework. The evaluation confirms that our tightly integrated system leads to an informed machine learning process while disclosing opportunities for further extensions.

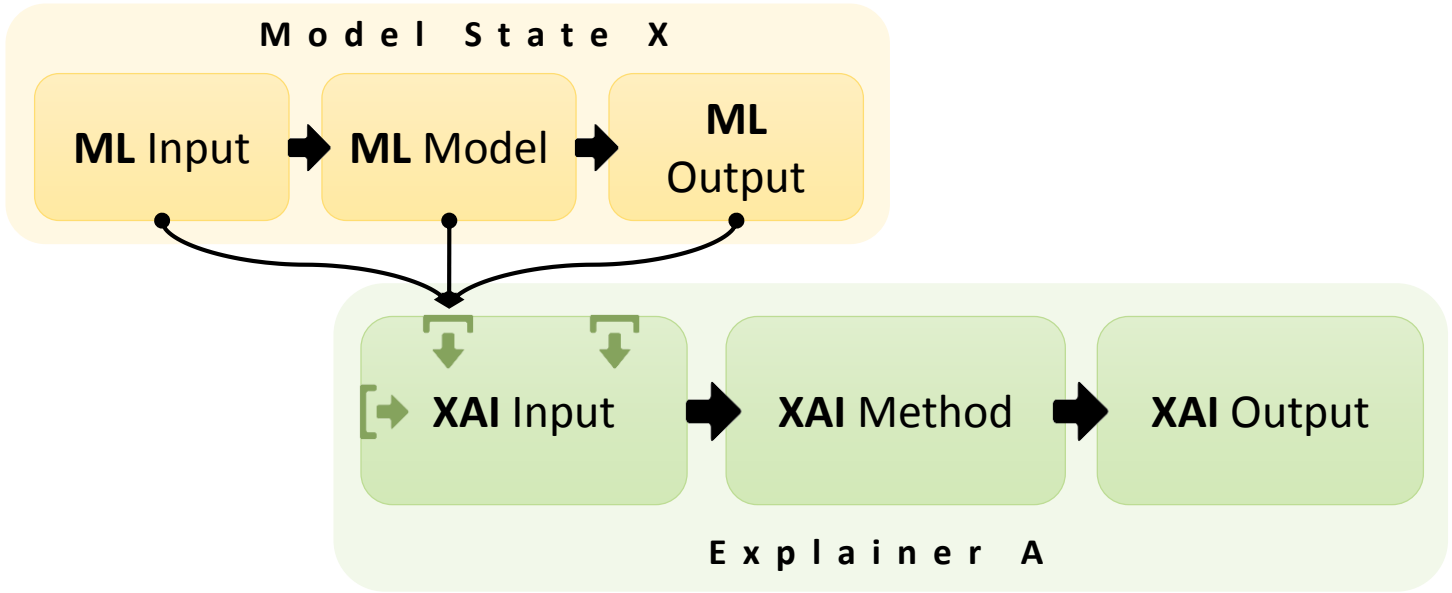
Index Terms—Explainable AI, Interactive Machine Learning, Deep Learning, Visual Analytics, Interpretability, Explainability

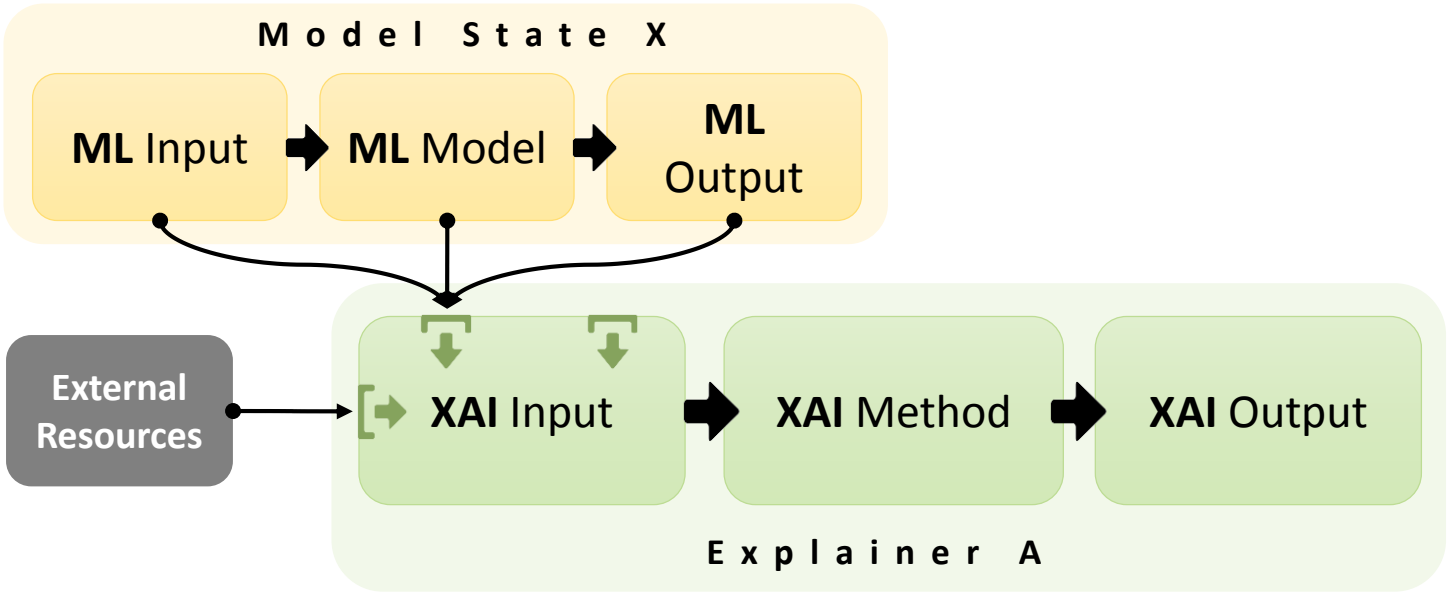
XAI as a Process

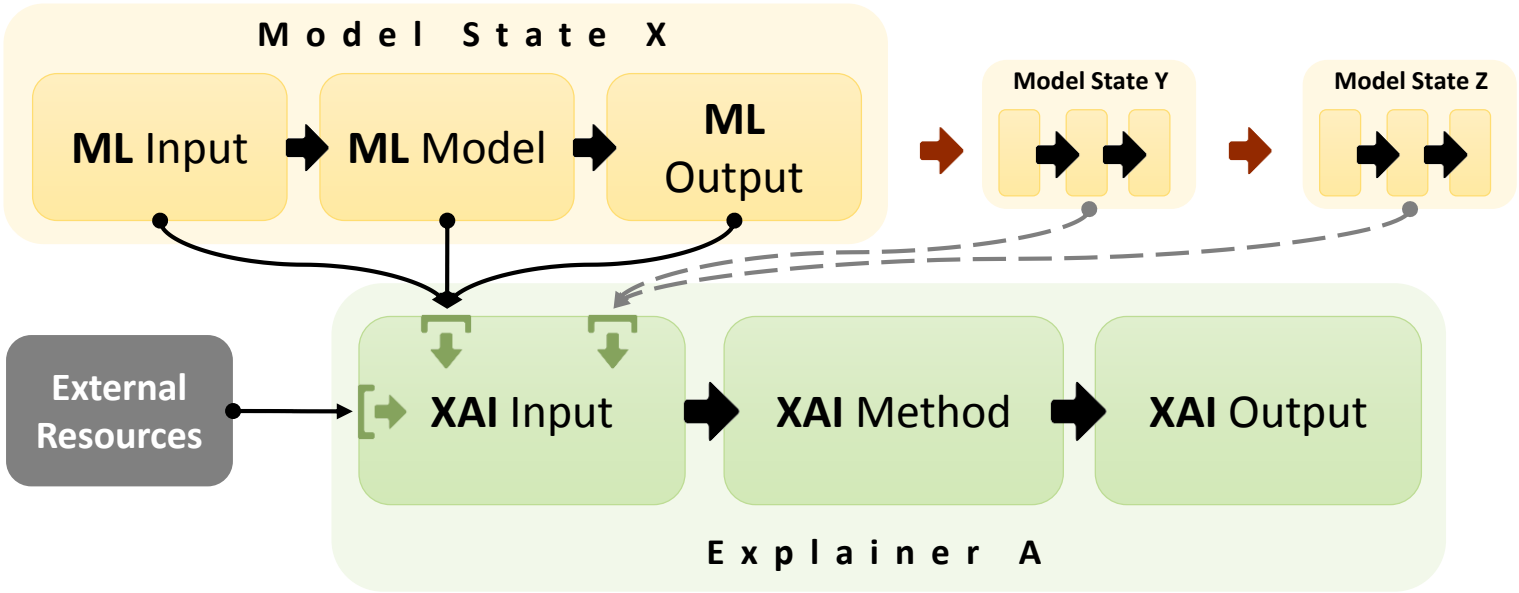


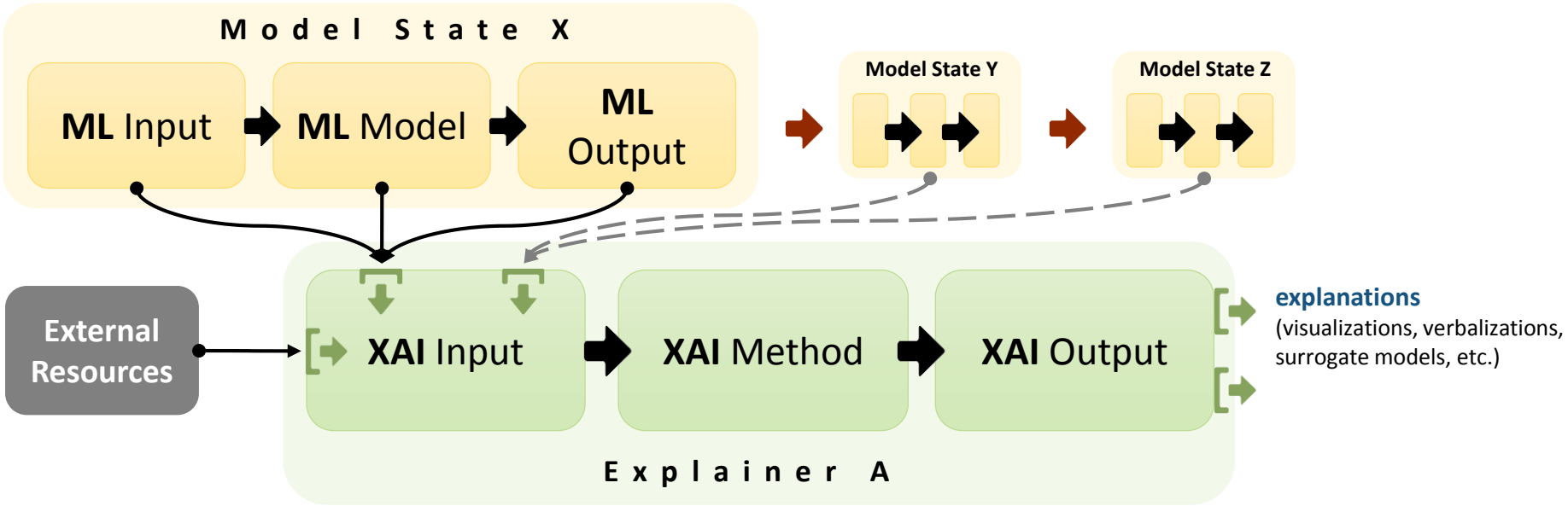
XAI as a **Process**

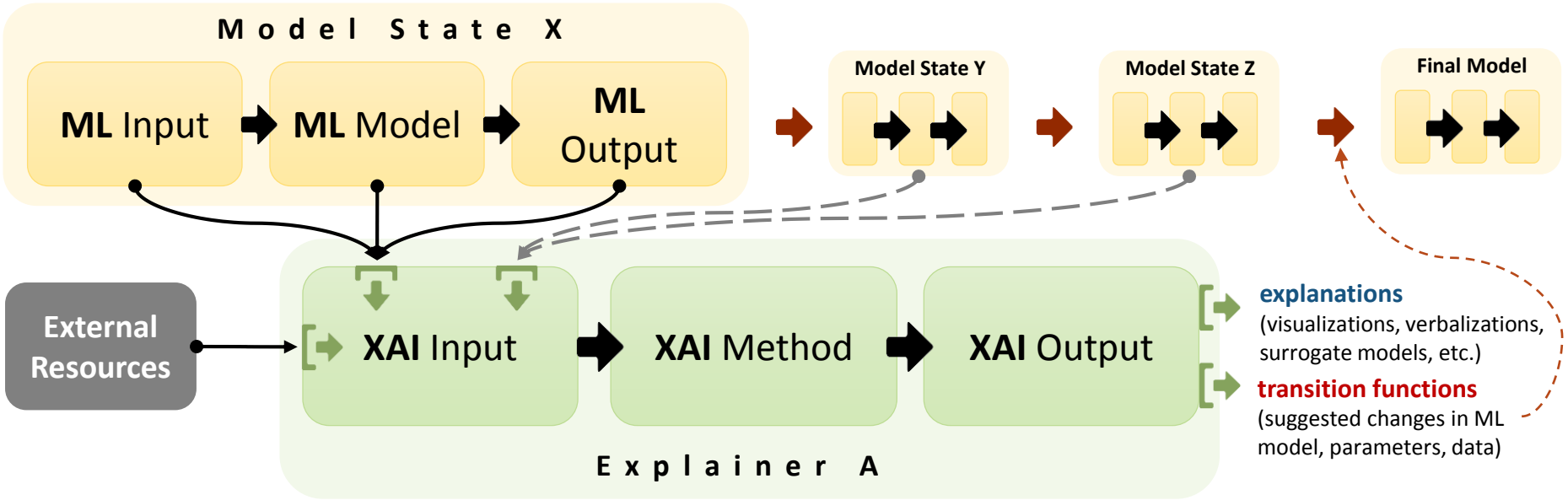

Model State X
ML Input
ML Model
ML Output

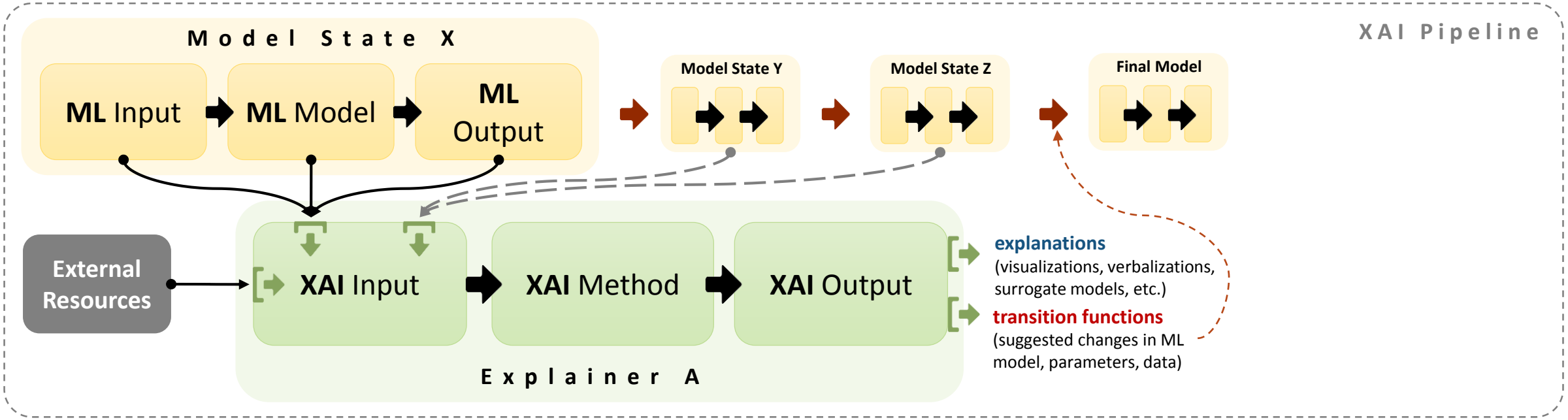

Model State X
ML Input
ML Model
ML Output
XAI Input
XAI Method
XAI Output
Explainer A

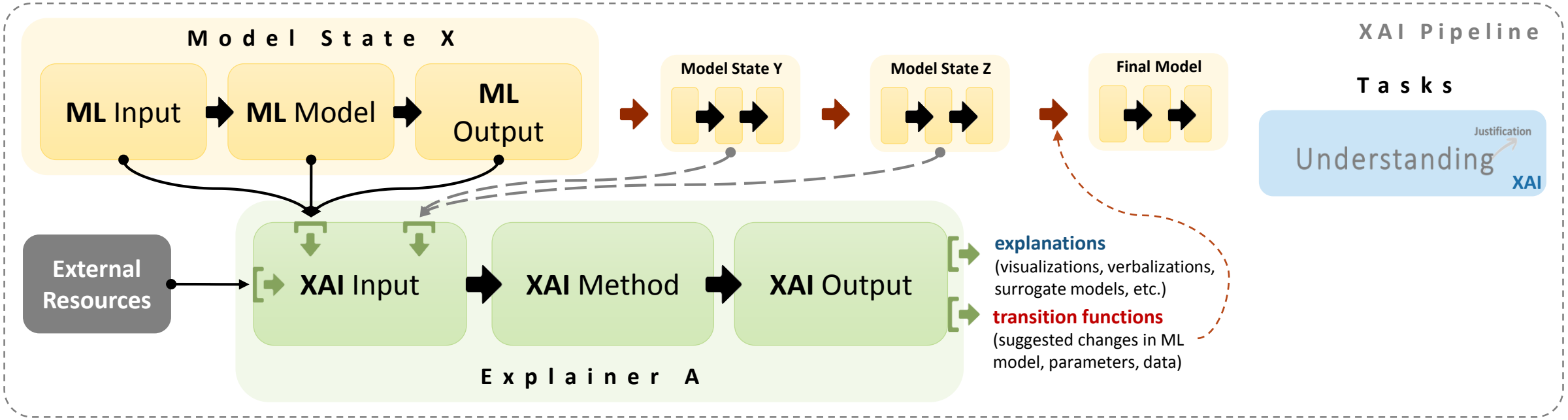

Model State X
ML Input
ML Model
ML Output
External Resources
XAI Input
XAI Method
XAI Output
Explainer A

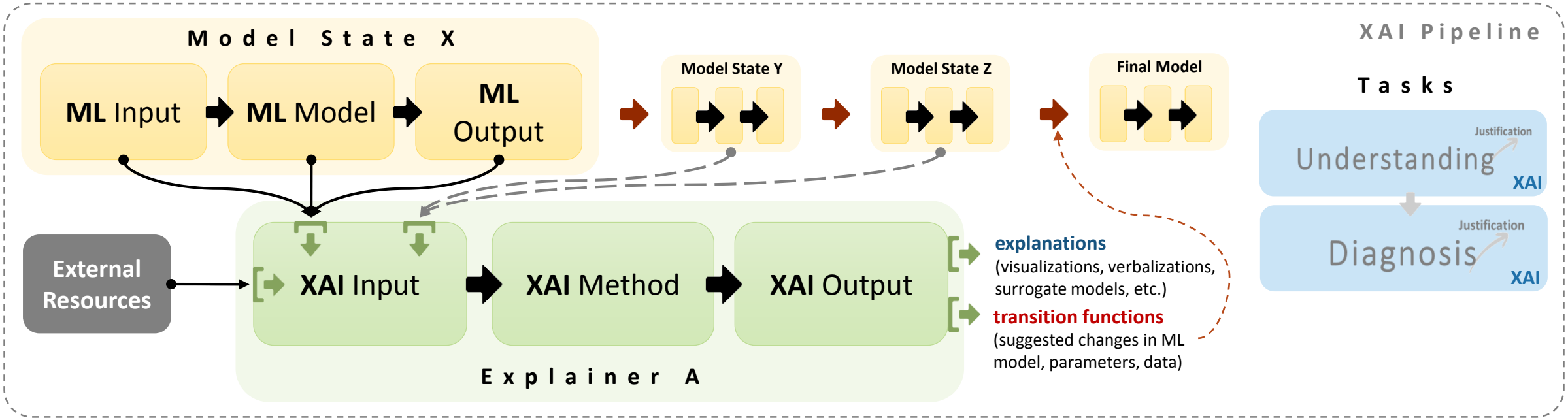

Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
External Resources
XAI Input
XAI Method
XAI Output
Explainer A

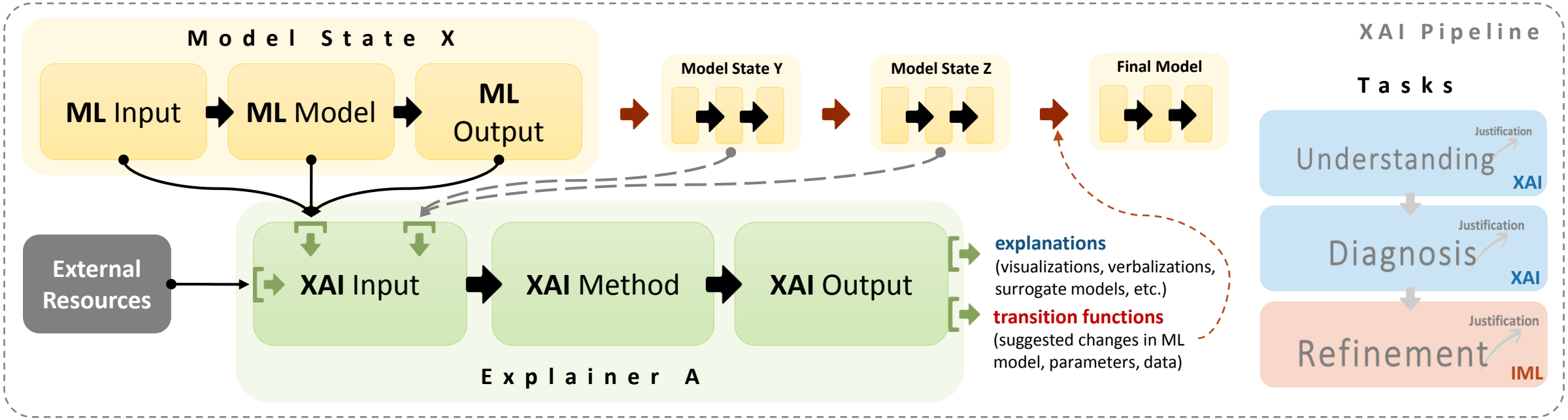

Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
Explainer A

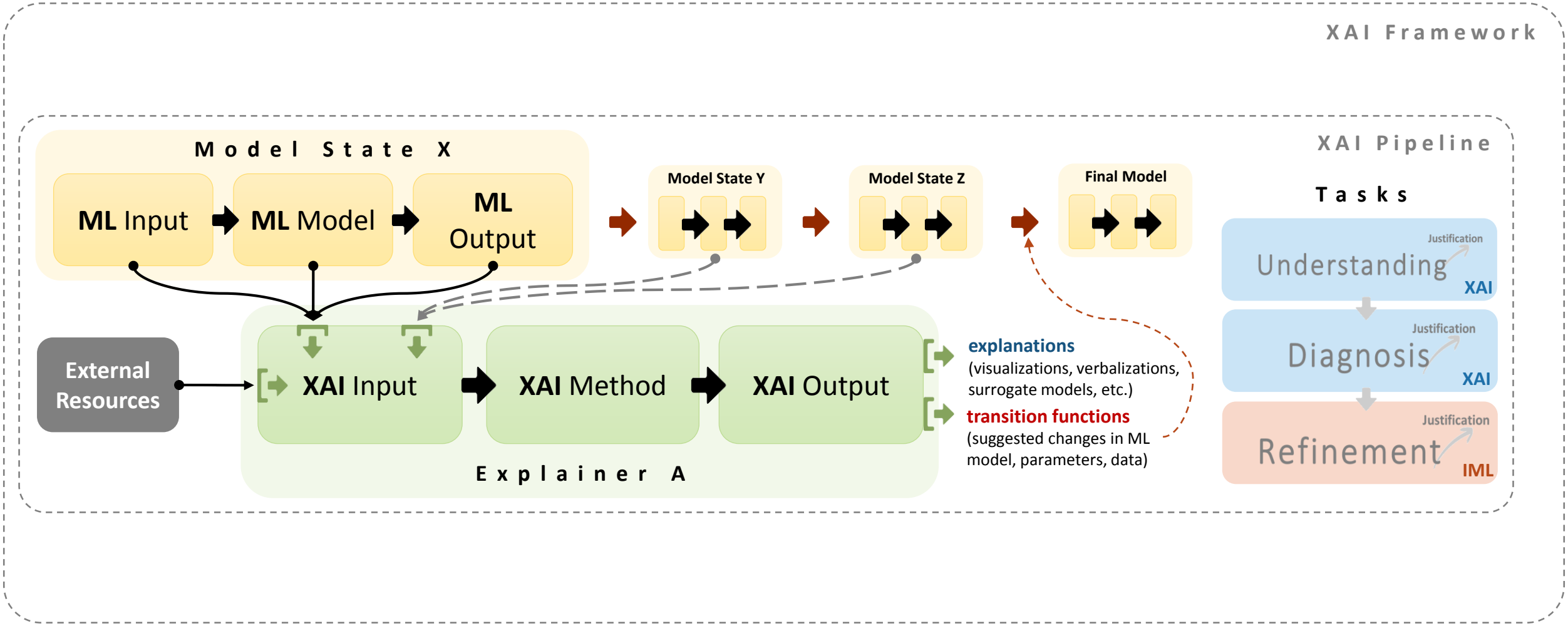

Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
External Resources
XAI Input
XAI Method
XAI Output
Explainer A
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)

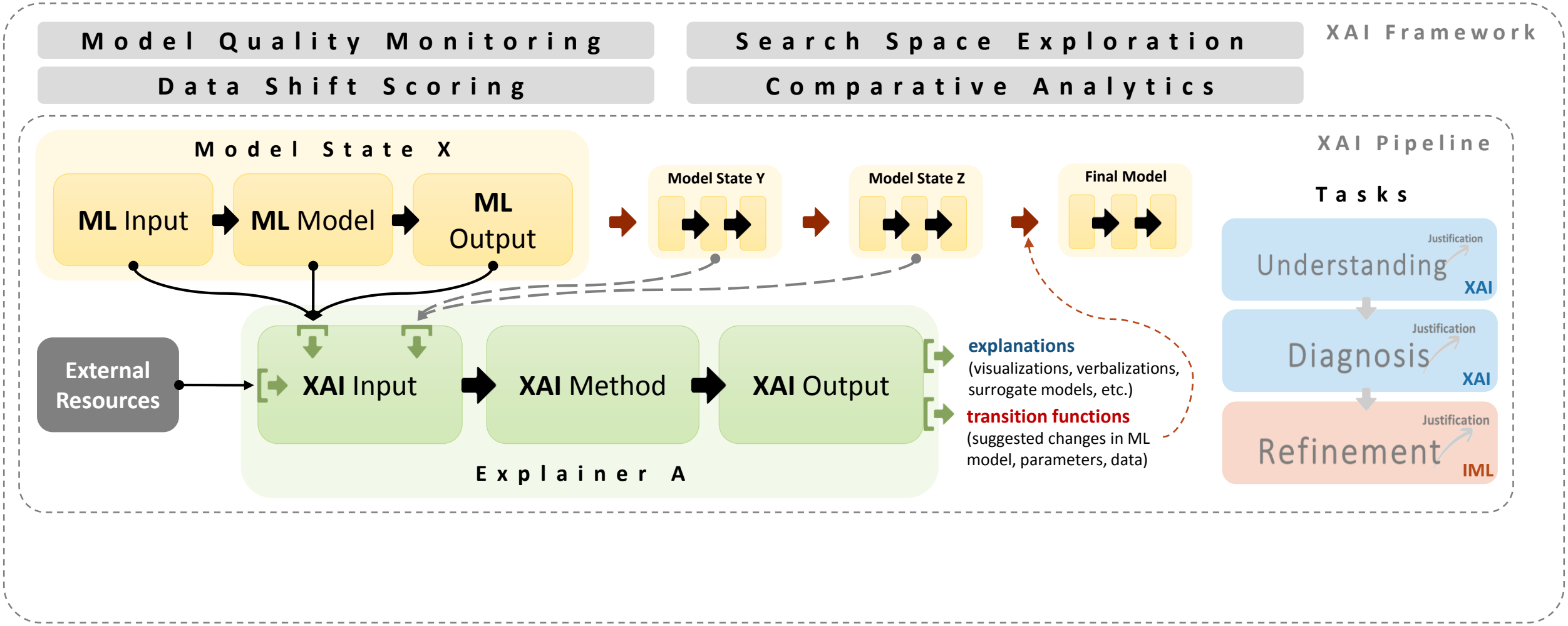

XAI Pipeline
Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A

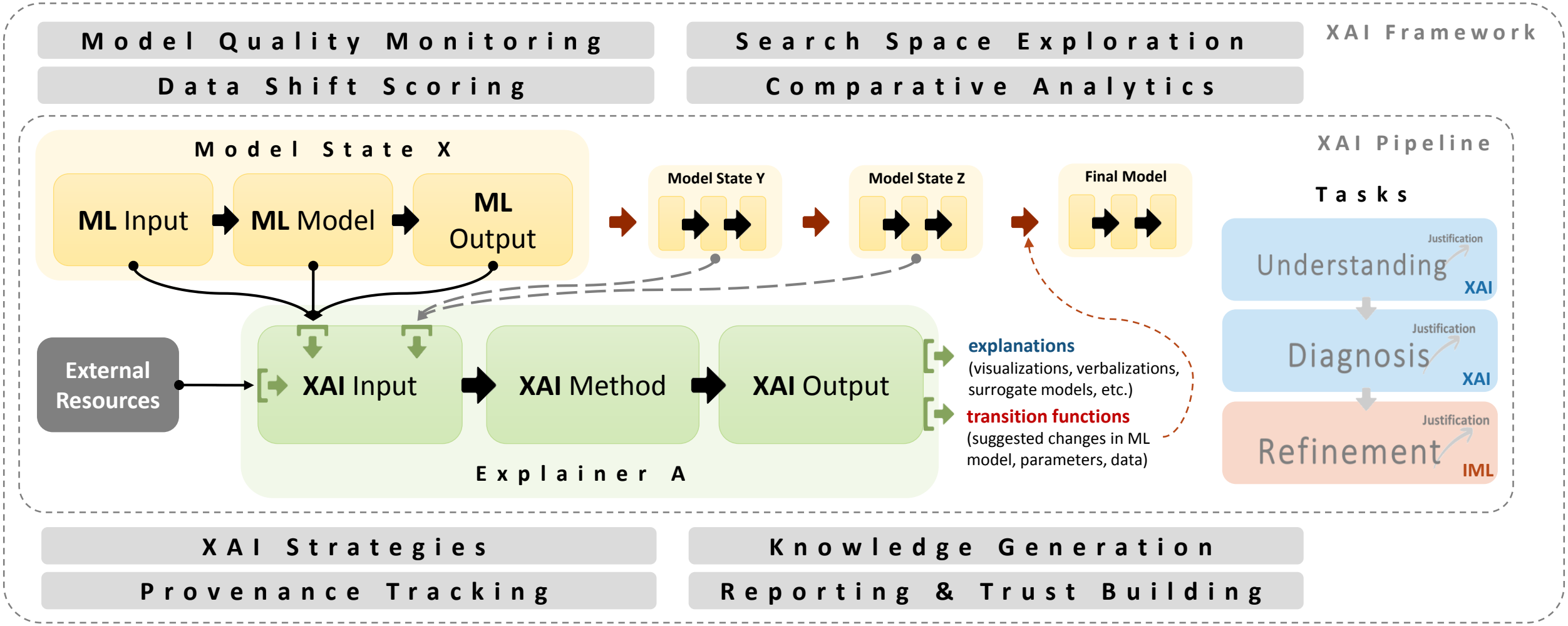

Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
XAI Pipeline
Tasks
Justification
Understanding
XAI
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A

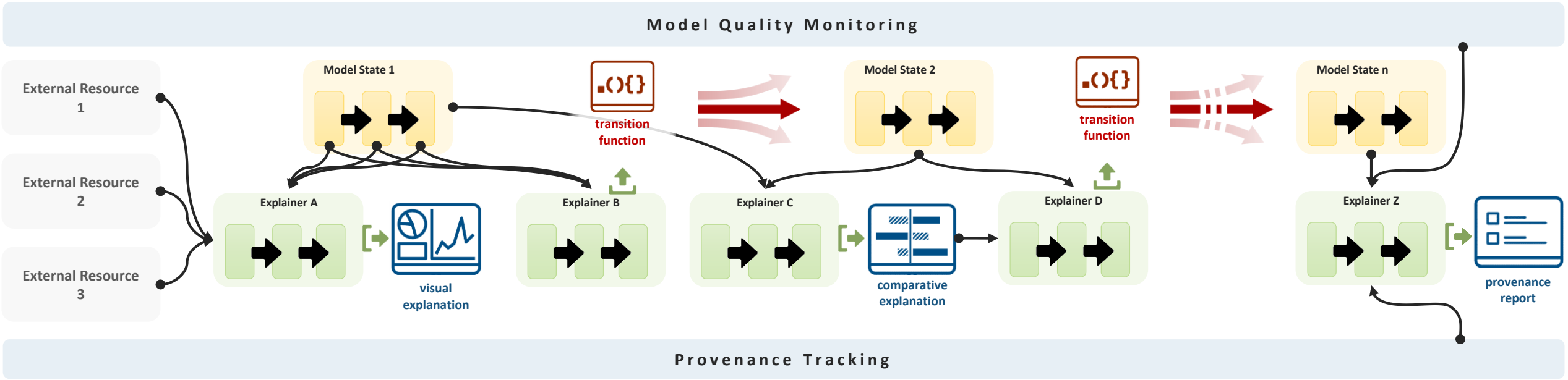

Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
XAI Pipeline
Tasks
Justification
Understanding
XAI
Justification
Diagnosis
XAI
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A


Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
XAI Pipeline
Tasks
Justification
Understanding
XAI
Justification
Diagnosis
XAI
Justification
Refinement
IML
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A


XAI Framework
XAI Pipeline
Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
Tasks
Justification
Understanding
XAI
Justification
Diagnosis
XAI
Justification
Refinement
IML
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A


Model Quality Monitoring
Search Space Exploration
XAI Framework
Data Shift Scoring
Comparative Analytics
XAI Pipeline
Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
Tasks
Justification
Understanding
XAI
Justification
Diagnosis
XAI
Justification
Refinement
IML
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A


Model Quality Monitoring
Search Space Exploration
XAI Framework
Data Shift Scoring
Comparative Analytics
XAI Pipeline
Model State X
ML Input
ML Model
ML Output
Model State Y
Model State Z
Final Model
Tasks
Justification
Understanding
XAI
Justification
Diagnosis
XAI
Justification
Refinement
IML
External Resources
XAI Input
XAI Method
XAI Output
explanations
(visualizations, verbalizations, surrogate models, etc.)
transition functions
(suggested changes in ML model, parameters, data)
Explainer A
XAI Strategies
Knowledge Generation
Provenance Tracking
Reporting & Trust Building


Model Quality Monitoring
External Resource 1
External Resource 2
External Resource 3
Model State 1
transition function
Model State 2
transition function
Model State n
Explainer A
visual explanation
Explainer B
Explainer C
comparative explanation
Explainer D
Explainer Z
provenance report
Provenance Tracking

Model Quality Monitoring

Search Space Exploration

Data Shift Scoring

Comparative Analytics

XAI FRAMEWORK

Model State

(4) ***abstraction*** *(model coverage): high (whole model), low (parts)*

Data

(3) ***level*** *(data coverage): global (all), local (subset)*

ML Input

ML Model

ML Output

(2) ***parts*** *of the* ***model state*** *considered*

(1) ***number*** *of* ***model states*** *considered*

External Resources

(5) ***dependency****: data, model, domain*

XAI Input

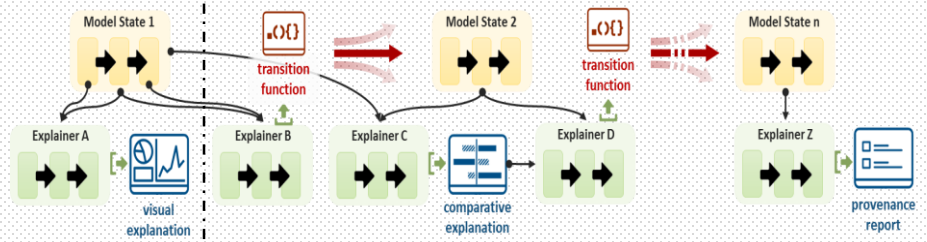
XAI Method

XAI Output

Explainer

explanations

transition functions



XAI PIPELINE

Tasks

Justification
Understanding
XAI

Justification
Diagnosis
XAI

Justification
Refinement
IML

XAI Strategies

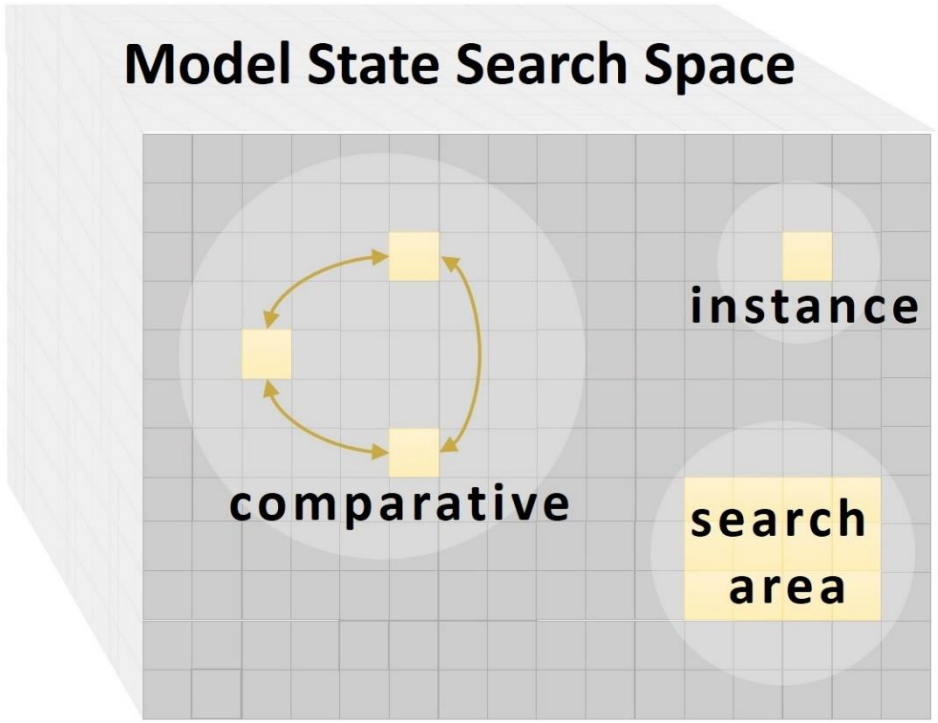
Knowledge Generation

Provenance Tracking

Reporting & Trust Building

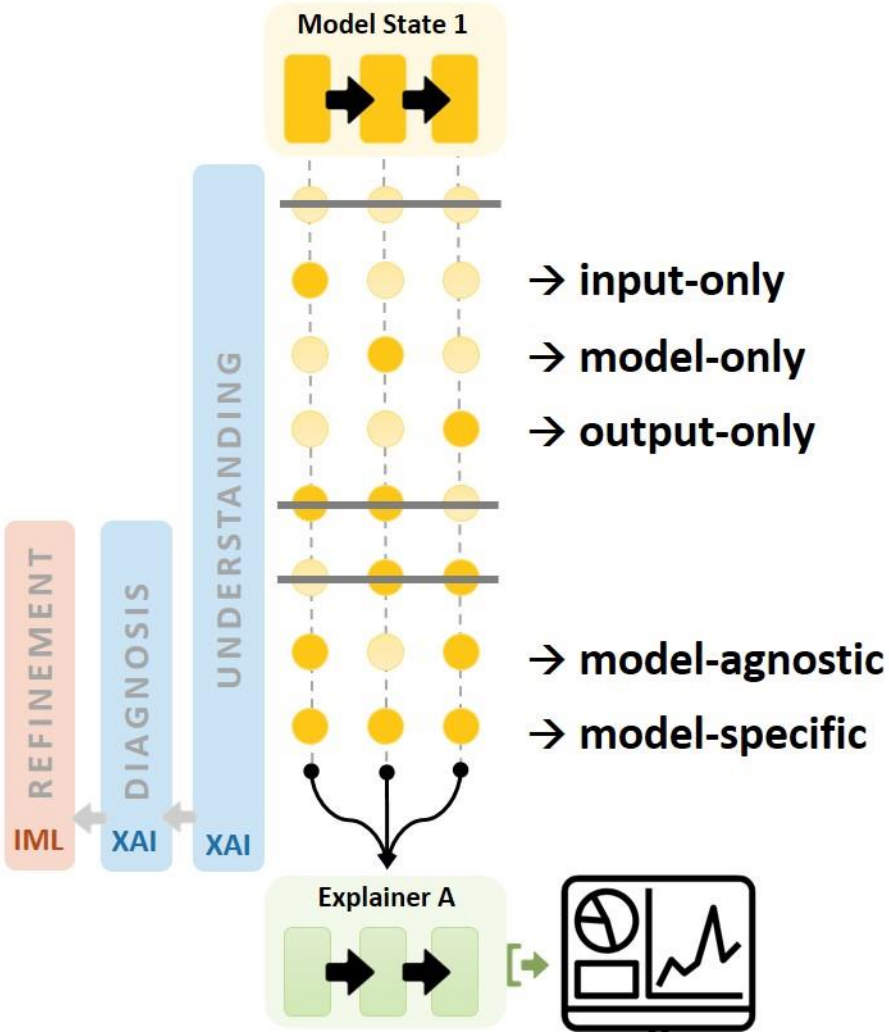
Explainer Properties

***(1) number** of **model states** considered*
***(2) parts** of the **model state** considered*
***(3) level** (data coverage): global, local*
***(4) abstraction** (model coverage): high, low*
***(5) dependency**: data, model, domain*



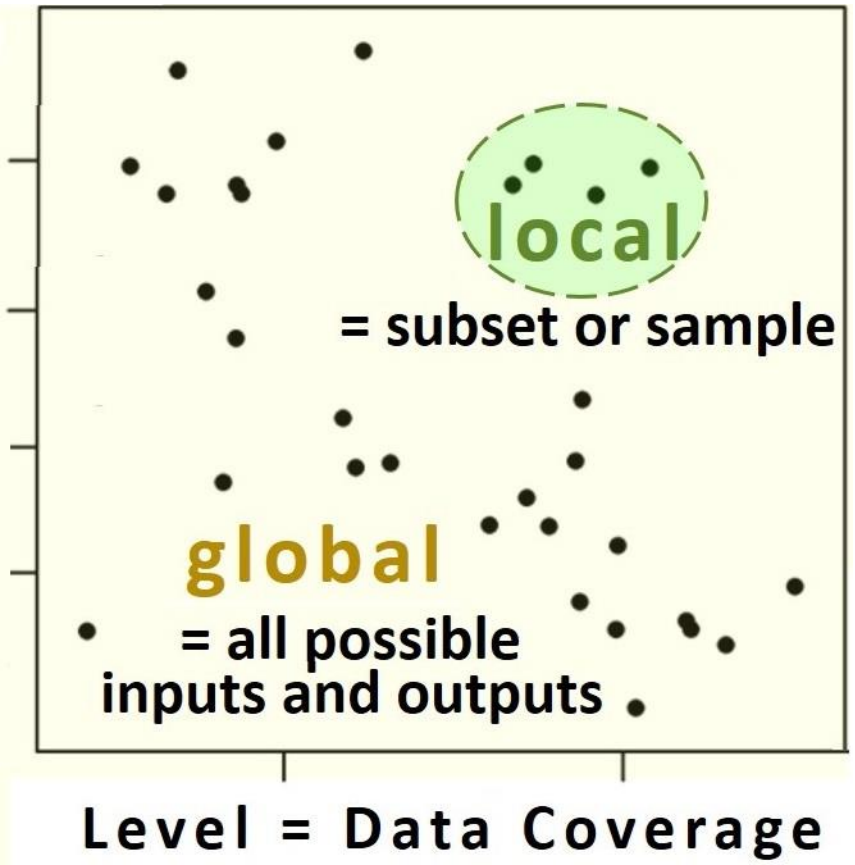
Explainer Properties

(1) ***number*** *of* ***model states*** *considered*
(2) parts *of the* ***model state*** *considered*
(3) ***level*** *(data coverage): global, local*
(4) ***abstraction*** *(model coverage): high, low*
(5) ***dependency****: data, model, domain*



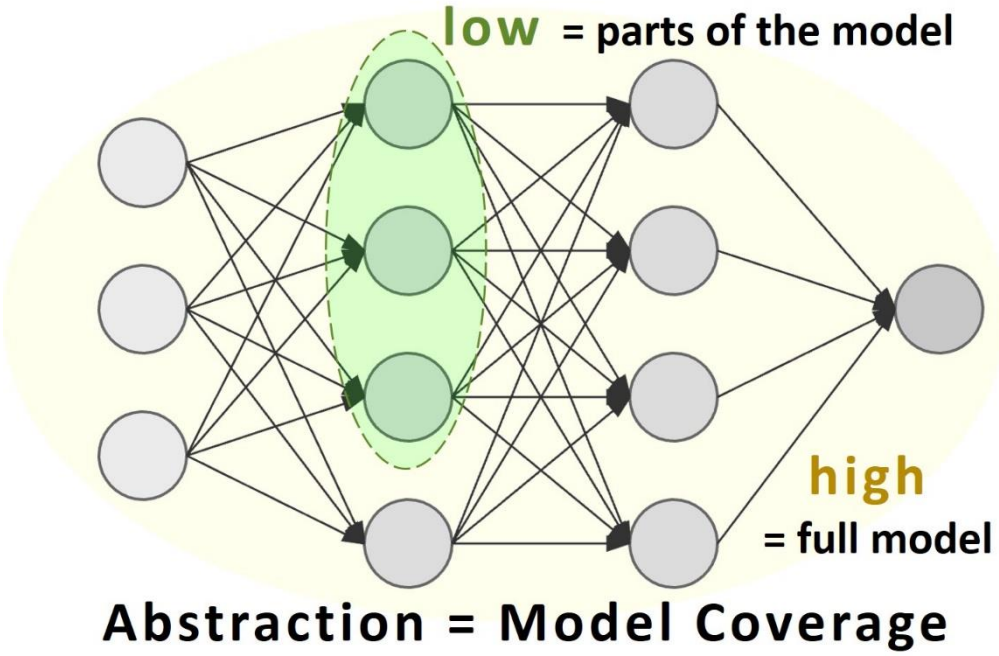
Explainer Properties

(1) ***number*** *of* ***model states*** *considered*
(2) ***parts*** *of the* ***model state*** *considered*
(3) level *(data coverage): global, local*
(4) ***abstraction*** *(model coverage): high, low*
(5) ***dependency****: data, model, domain*



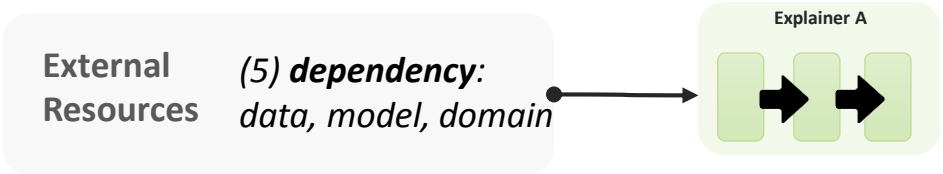
Explainer Properties

(1) ***number*** *of* ***model states*** *considered*
(2) ***parts*** *of the* ***model state*** *considered*
(3) ***level*** *(data coverage): global, local*
(4) ***abstraction*** *(model coverage): high, low*
(5) ***dependency****: data, model, domain*



Explainer Properties

(1) ***number*** *of* ***model states*** *considered*
(2) ***parts*** *of the* ***model state*** *considered*
(3) ***level*** *(data coverage): global, local*
(4) ***abstraction*** *(model coverage): high, low*
(5) dependency*: data, model, domain*



TensorBoard UNDERSTANDING DIAGNOSIS REFINEMENT REPORTING GRAPHS INACTIVE

Global Surrogate Models
- Lime
- Anchors
- Cav

Global Gradient Propagation Methods
- Grad*Input
- IntGrad
- Gradient
- SmoothGrad
- Input Times Gradient
- Integrated Gradients
- DeepLift
- Saliency
- DeConvNet

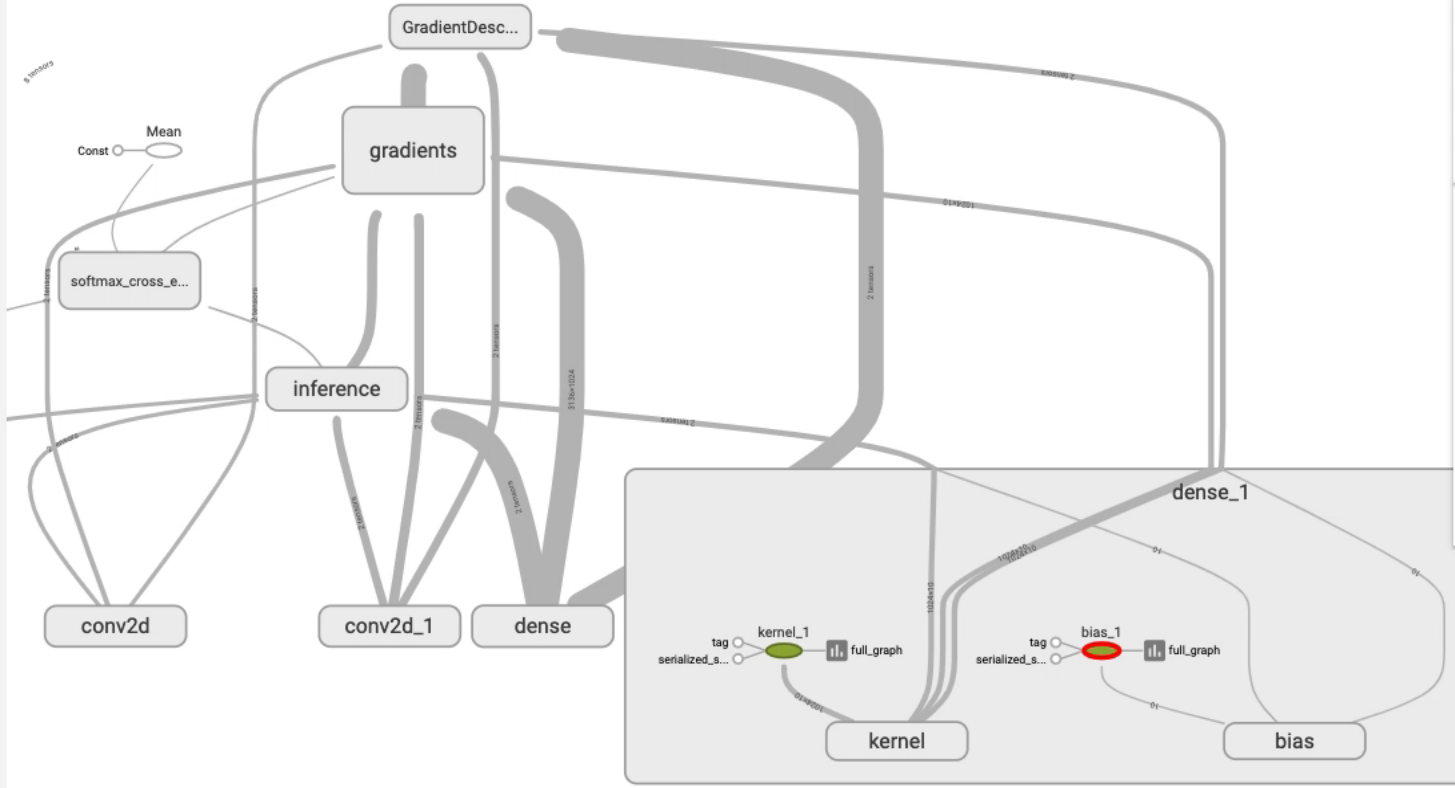
Global Layer-wise Relevance Propagation
- z-LRP
- e-LRP (fast)
- e-LRP (slow)

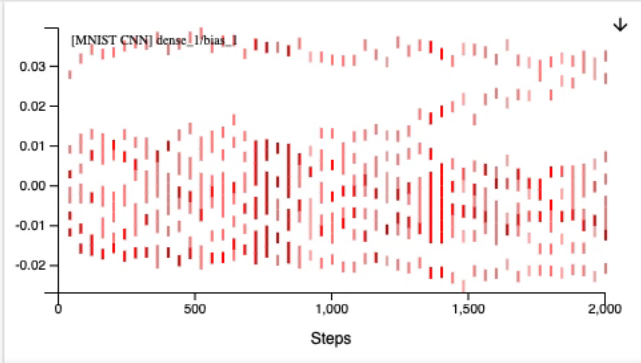
Model
- FunctionExplainer
- LossDecay

Graph Flow
- ImageInterpreter
- InformationFlow

Weights
- HistoTrend
- MinMax
- DeadWeight
- SaturatedWeight

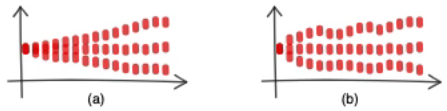
Try it out on **explainer.ai**

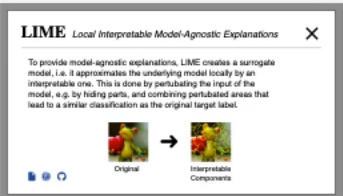


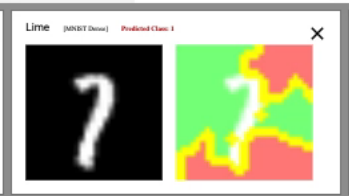


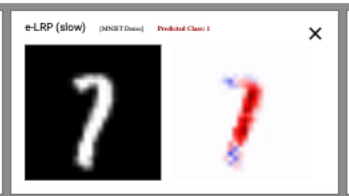
HistoTrend *Histograms, over Time*

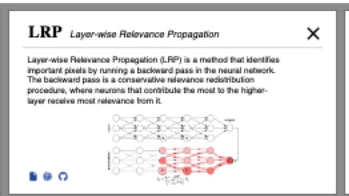
The HistoTrend-Tool plots the Histograms of a tensor, one for each logged step. It shows how the distribution of values changes over time and can reveal interesting patterns, such as binning (a) or weight oscillations (b).

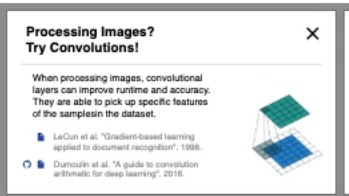










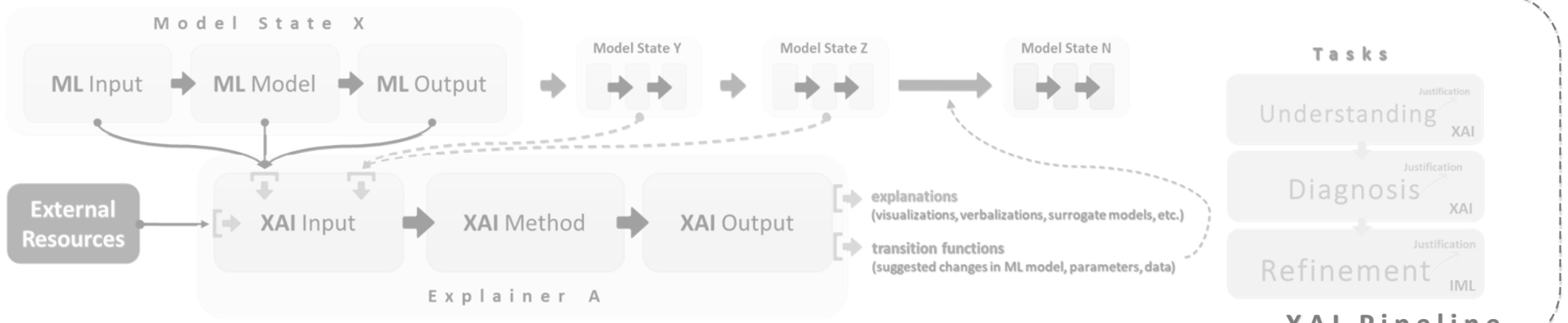


Model Quality Monitoring
(Metrics, Bias, Uncertainty, Performance)

Search Space Exploration
(Speculative Execution, Targeted Optimization)

Data Shift Scoring
(Training -- Testing Data Continuum)

Comparative Analytics
(Model Selection, Recommendation)



XAI Pipeline

XAI Strategies
(User Guidance, Explanation Mediums)

Knowledge Generation
(ML Validation, Learning from User Interactions)

Provenance Tracking
(Temporal Evolution, Interaction Tree)

Reporting & Trust Building
(Education , Storytelling, Justification)

XAI Framework

Global Monitoring and Steering Mechanisms


Pedagogy


Storytelling


Dialog & Argumentation

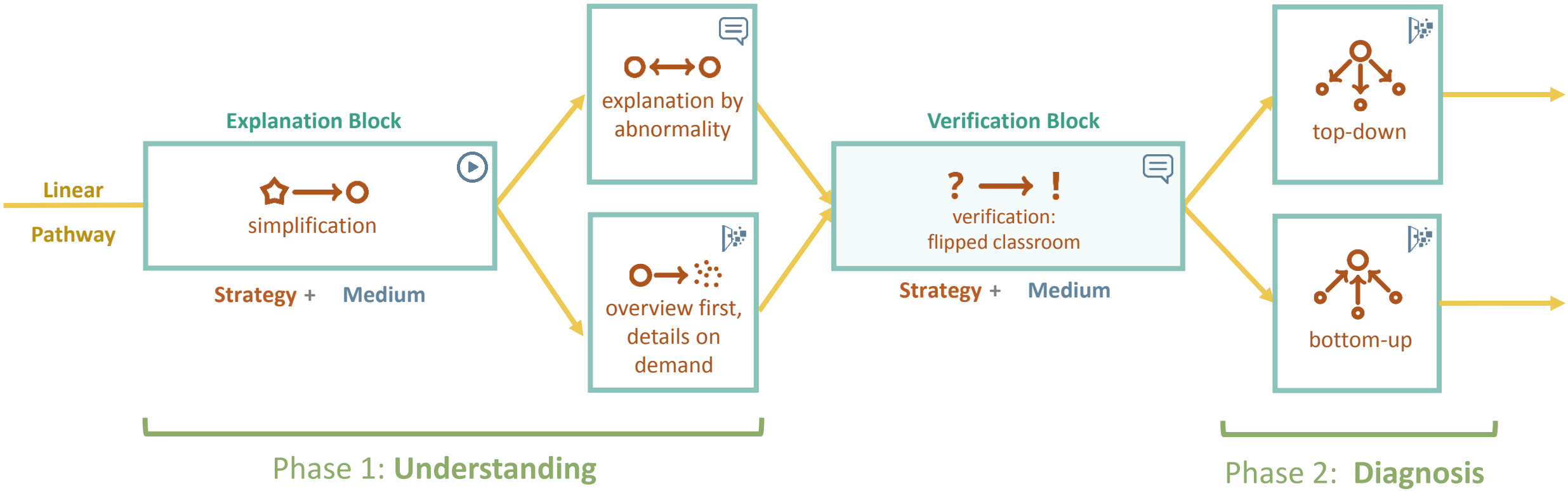
Towards XAI

Structuring the Processes of Explanations

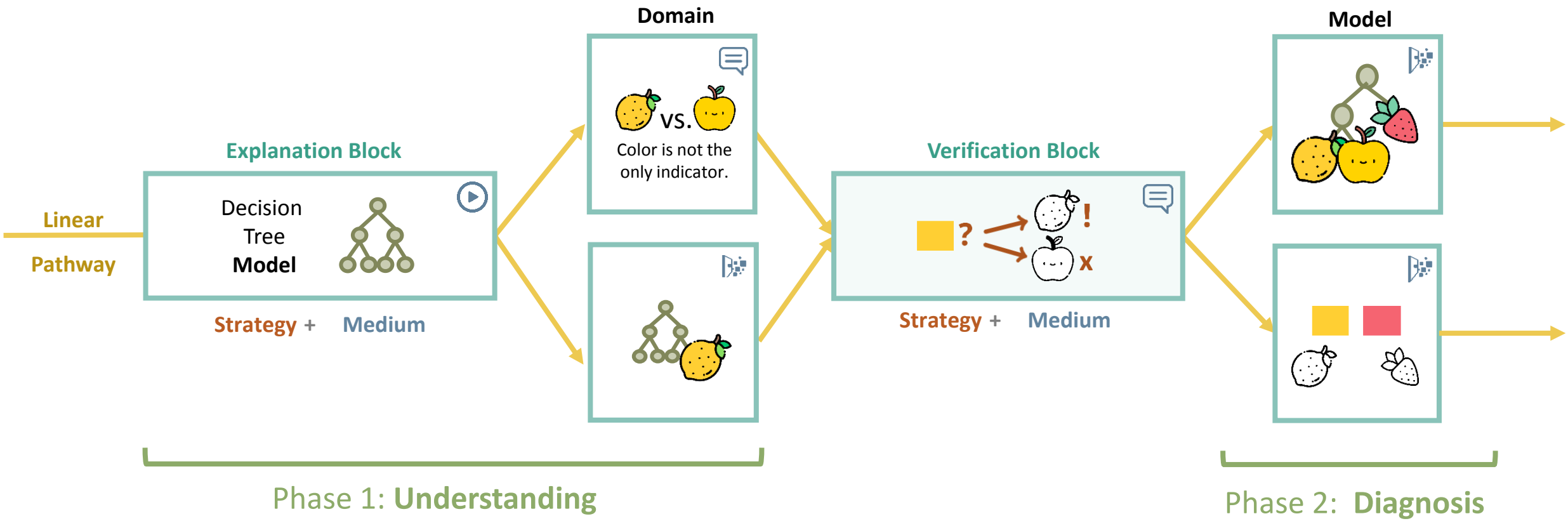

Programming


Trust Building

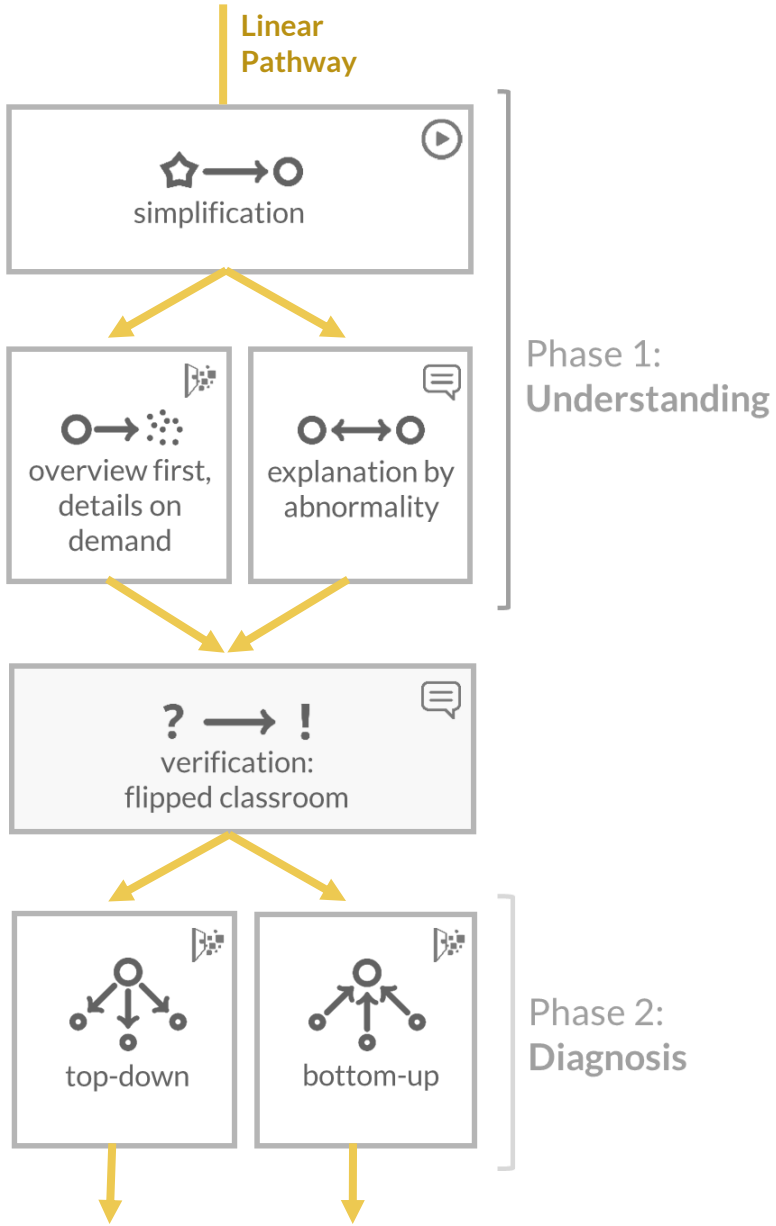

Gamification



Explanation Process Model



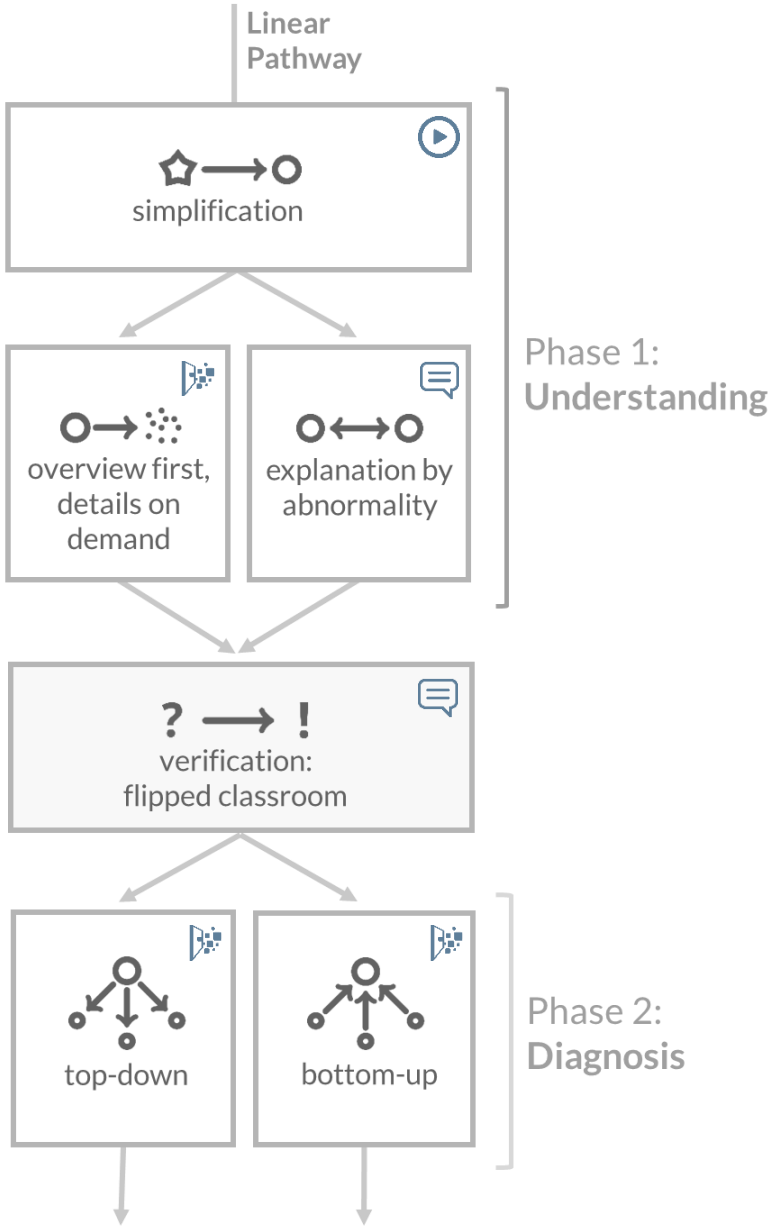
Explanation Process Model



Pathways

Linear vs. **Iterative**

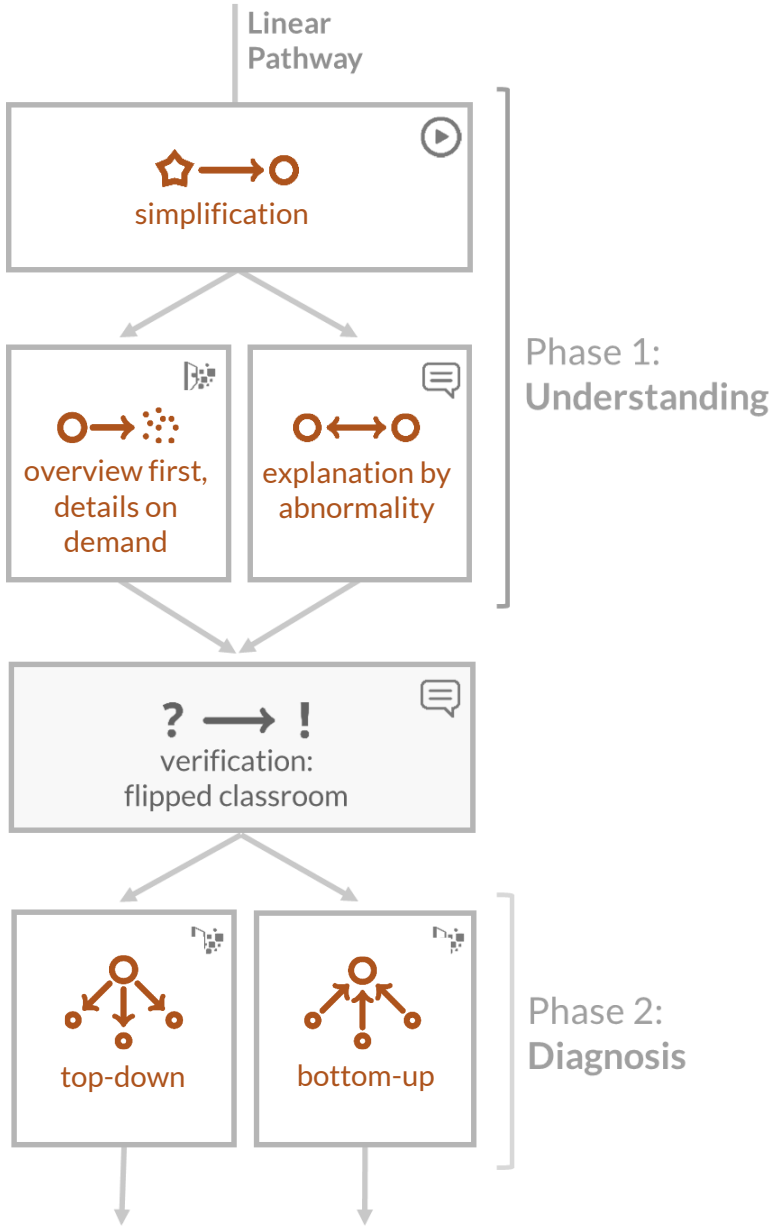
Guided vs. **Serendipitous**



Mediums

- Visualizations
- Verbalizations
- Infographics
- Illustrated Text
- Comics
- Videos
- Audios
- Images
- Video Games
- Dialog Systems

Explanation Strategies



Inductive (Bottom Up)

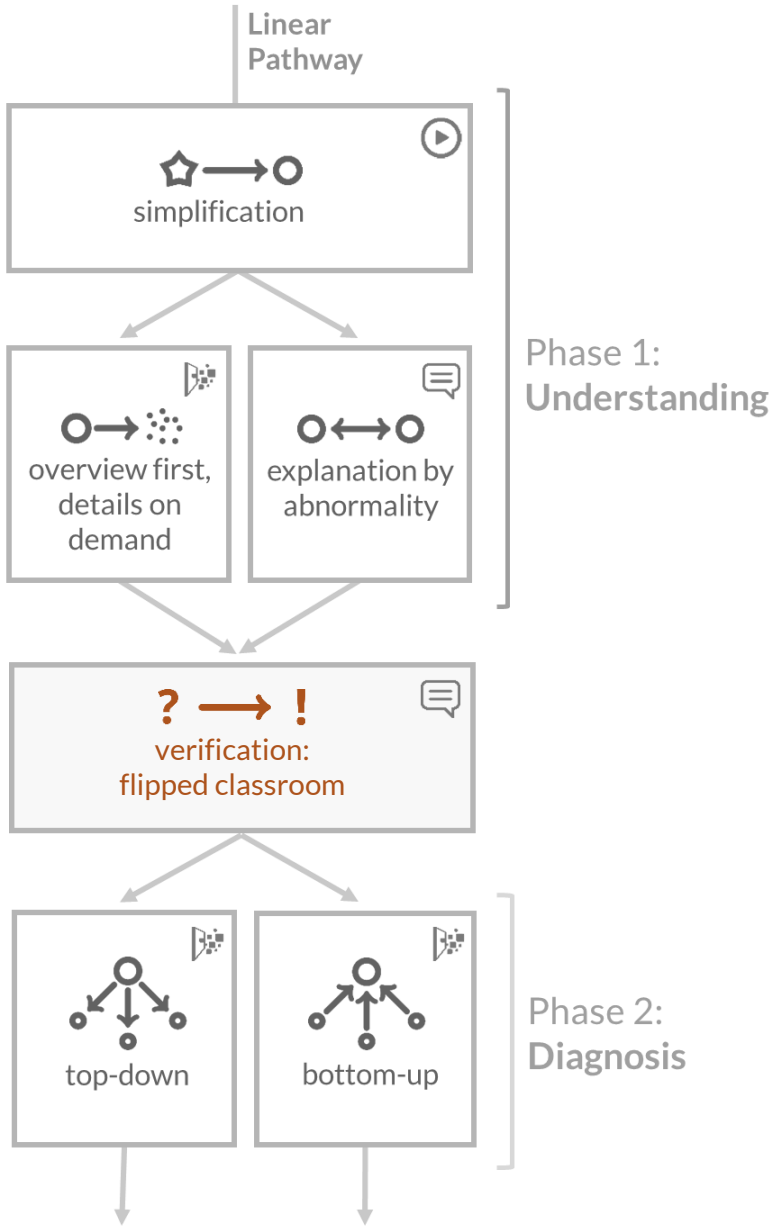
simplification
metaphorical narrative
divide and conquer
explanation by example
dynamic programming
depth first - breadth first
describe and define
teaching by categories

Deductive (Top Down)

transfer learning
teaching by association
overview first, details on demand
drill down story
define and describe

Contrastive (Comparison)

opposite and similar
example by abnormality



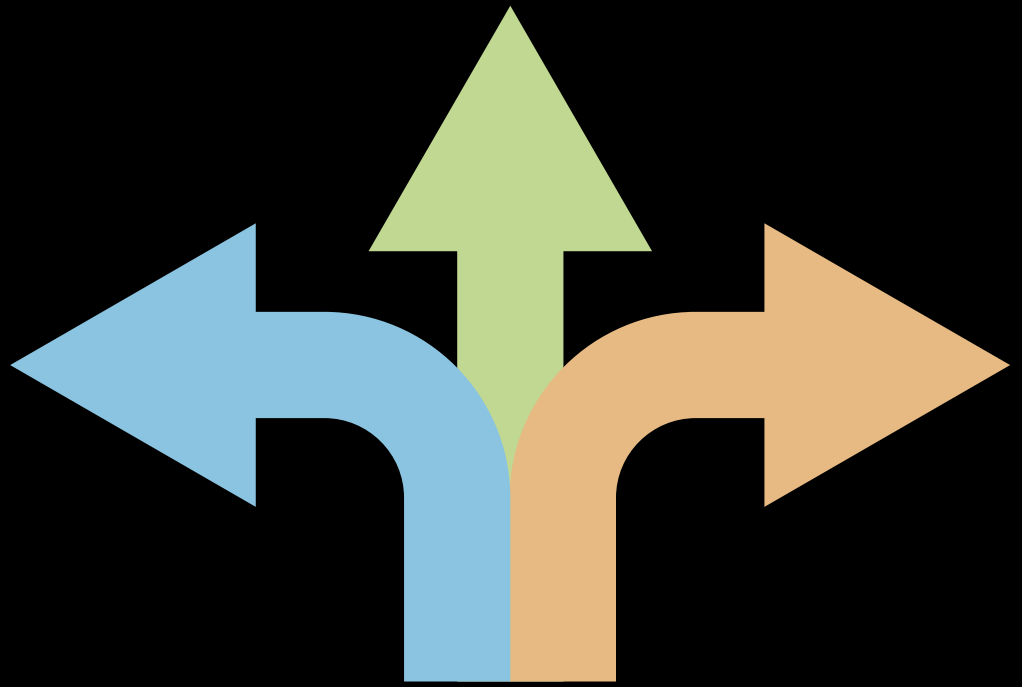
Verification Strategies

? ⟶ ! flipped classroom
reproduction
transfer

Verification

? ⟶ !

verification:
flipped classroom

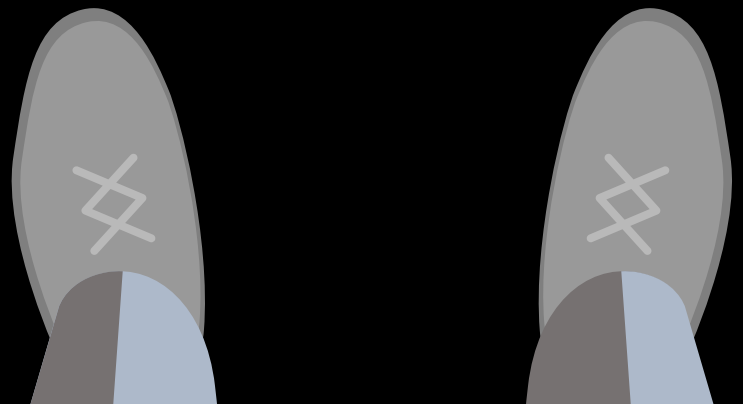


Top-Down Explanation

top-down

Bottom-Up Explanation

bottom-up



Design Guidance

Robust XAI Methodology

(through observing other domains)

Research Project Overview

el-assady.com

LingVis

Integrating Computational Linguistics and Visual Analytics.

https://lingvis.io/

VisArgue

Analyzing Successful Rhetoric and Argumentation in Debates.

http://visargue.uni.kn/

explAIner

Developing Explainable and Interactive Machine Learning.

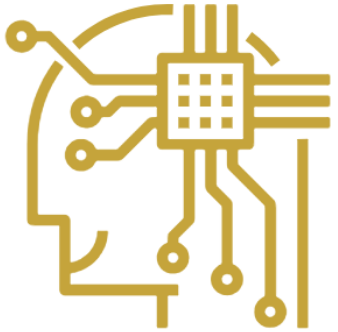
https://explainer.ai/

Visual Musicology

Exploring the Intersection of Musicology and Visual Analytics.

https://visual-musicology.com/

What's the role of humans in interpretability?



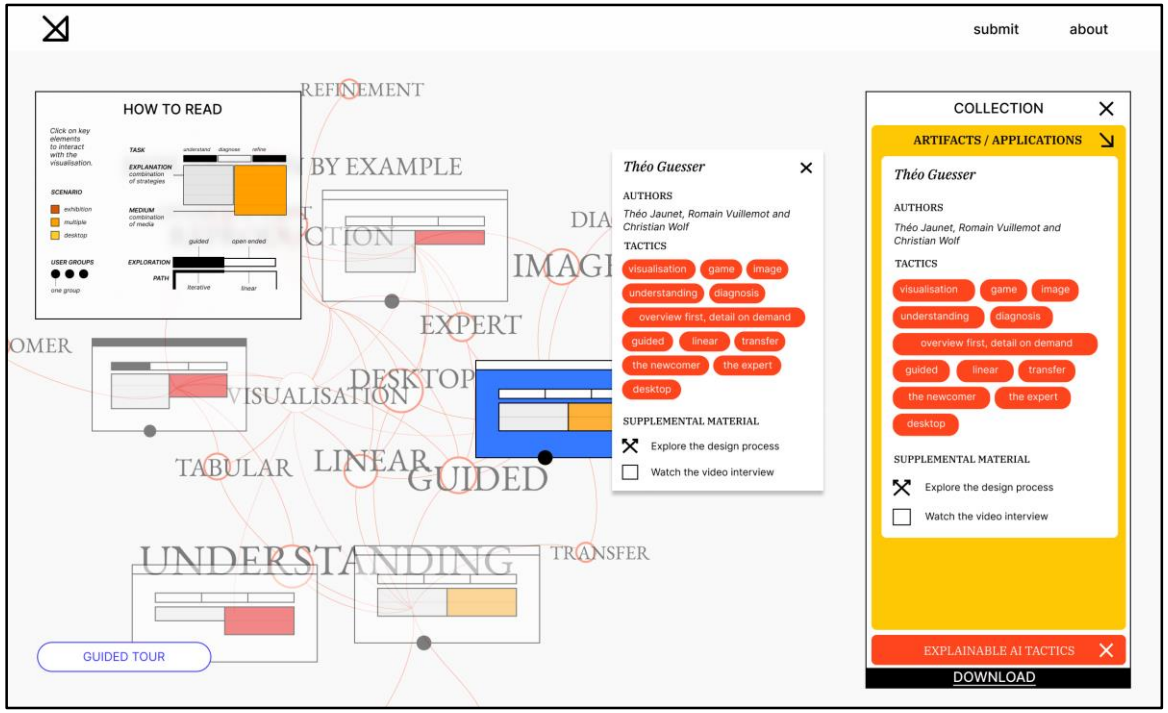
How much should we open the black-box?

How much guidance is needed to enable effective XAI?

work with: **Beatrice Gobbo**

Structuring XAI using a Museum Metaphor

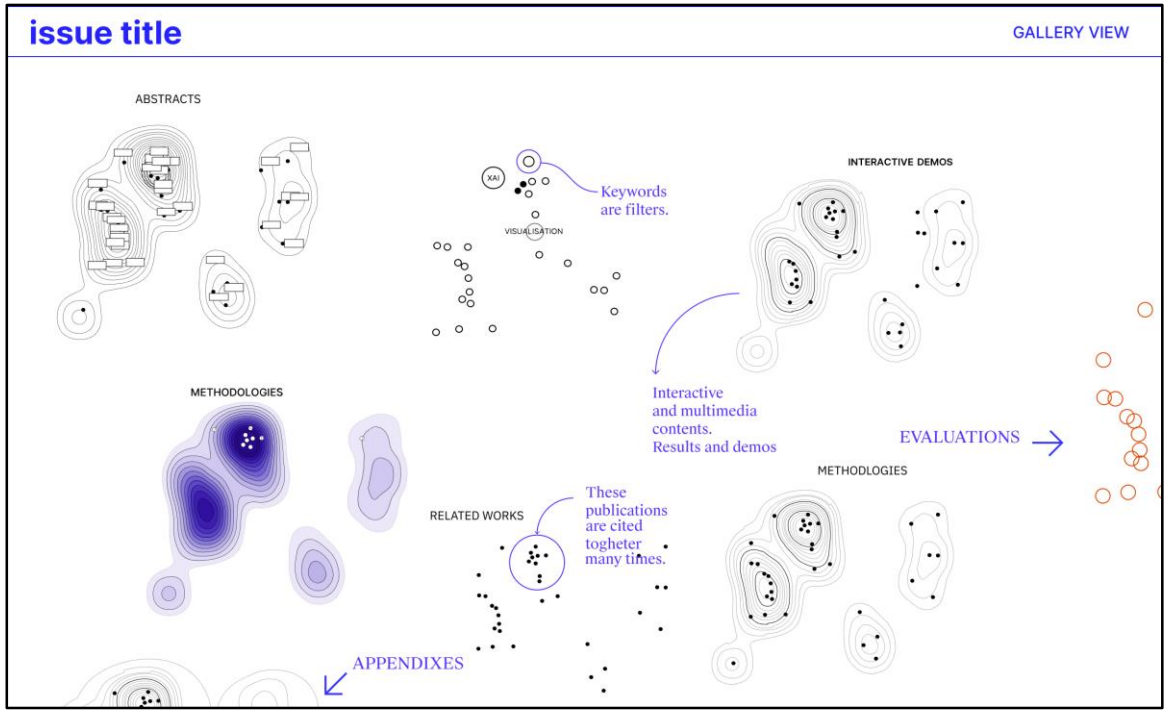
XAI PRIMER



http://bit.ly/xaiprimer

CHI21 Workshop
Operationalizing Human-Centered Perspectives in XAI
Sat. May 8 & Sun May 9, 2021
@ 1300 EDT/ 1900 CEST

PAPER MUSEUM

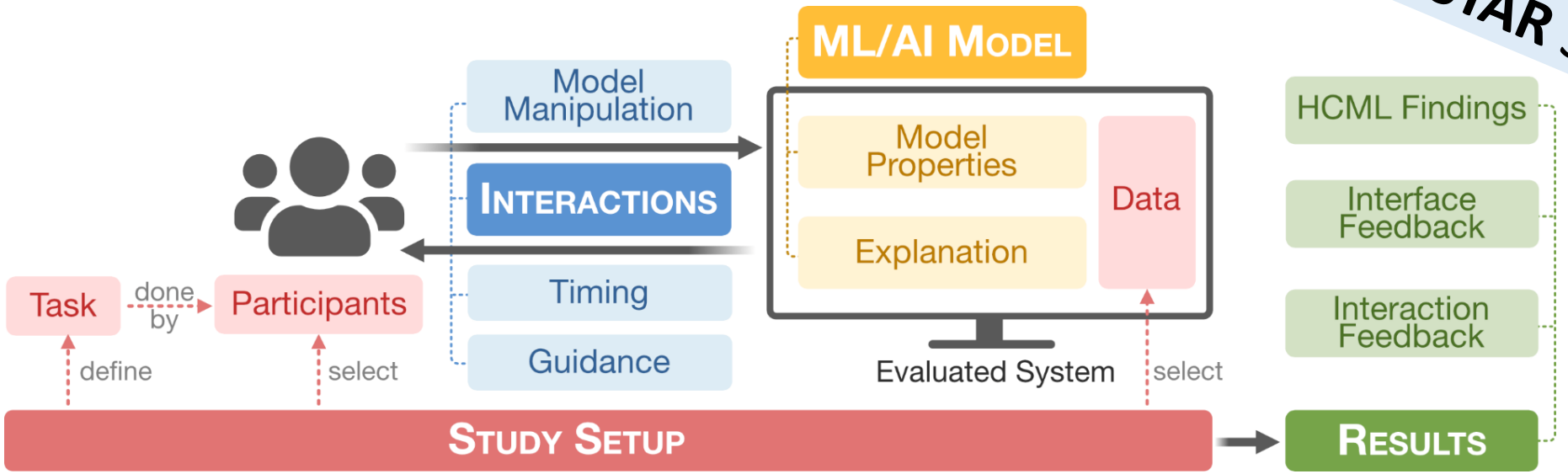


http://bit.ly/papermuseum

ICLR 2021 Workshop
Rethinking ML Papers
Friday, May 7 2021
@ 800 EDT/ 1900 CEST

Human-Centered Evaluation

Check out the **EuroVis'21 STAR** Sessions

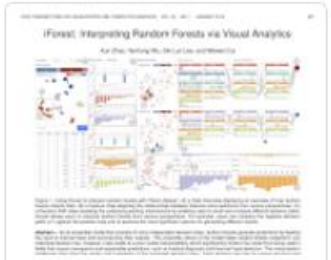


Interdisciplinarity **Complexity** **Participant Diversity** **Evaluation Focus**

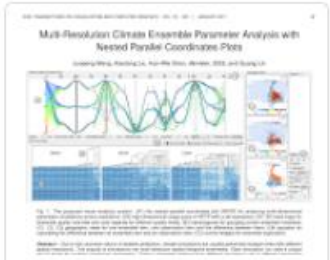
A Survey of Human-Centered Evaluations in Human-Centered Machine Learning
Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, Duen Horng Chau, Alex Endert, Daniel Keim
Computer Graphics Forum, 2021 (to appear)

Files

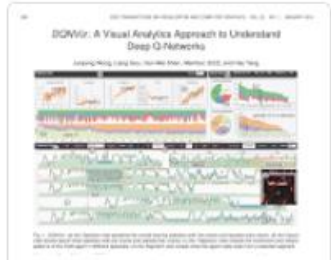
Name


Zhao et al. - 2018 - iFor...

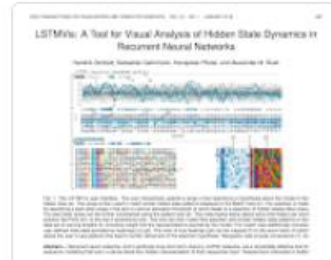

Zhang et al. - 2018 - Ma...


Wang, Shen and Lin - 20...

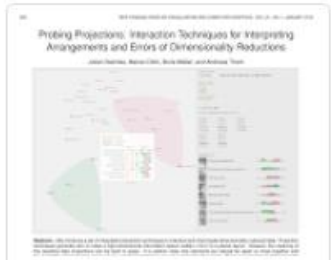

Wang et al. - 2019 - Vis...


Wang et al. - 2018 - DQ...


Strobelt et al. - 2018 - S...


Strobelt et al. - 2017 - L...

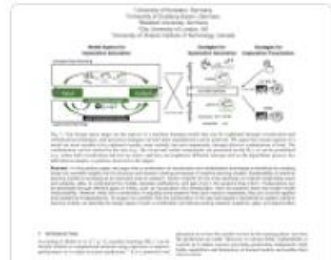

Stoffel et al. - 2015 - Fe...


Stahnke et al. - 2015 - P...

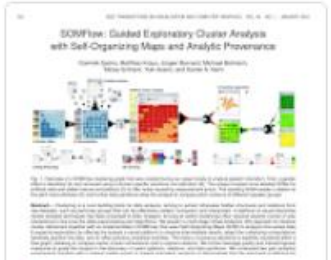

Spinner, Korner, Gortier, ...


Spinner et al. - 2019 - e...

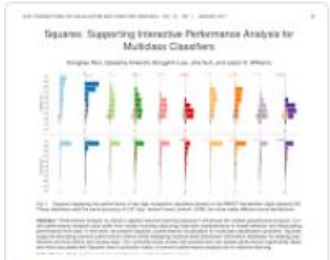

Sevastjanova et al. - 20...


Sevastjanova et al. - 20...


Schall et al. - 2018 - Vis...

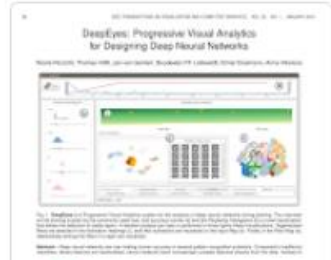

Sacha et al. - 2017 - SO...


Sacha et al. - 2016 - Vis...


Ren et al. - 2016 - Squar...


Rauber et al. - 2016 - Vi...


Preston and Ma - 2015 -...


Pezzotti et al. - 2017 - D...


Pezzotti et al. - 2016 - A...

http://bit.ly/VISmeetsML

4th Workshop on Visualization for AI Explainability

October 24th or 25th, 2021 at IEEE VIS in New Orleans, Louisiana

The role of visualization in artificial intelligence (AI) gained significant attention in recent years. With the growing complexity of AI models, the critical need for understanding their inner-workings has increased. Visualization is potentially a powerful technique to fill such a critical need.

The goal of this workshop is to initiate a call for *"explainables" / "explorables"* that explain how AI techniques work using visualization. We believe the VIS community can leverage their expertise in creating visual narratives to bring new insight into the often obfuscated complexity of AI systems.

Example interactive visualization articles that explain general concepts and communicate experimental insights when playing with AI models. (a) **A Visual Exploration of Gaussian Processes** by Görtler, Kehlbeck, and Deussen; (b) **The Beginner's Guide to Dimensionality Reduction** by Conlen and Hohman; (c) **What if we Reduce the Memory of an Artificial Doom Player?** by Jaunet, Vuillemot, and Wolf; (d) **Comparing DNNs with UMAP Tour** by Li and Scheidegger; (e) **The Myth of the Impartial Machine** by Feng and Wu; (f) **FormaFluens Data Experiment** by Strobelt, Phibbs, and Martino.

Contribute to VISxAI!

http://visxai.io/

Join the XAI Slack conversation!

http://bit.ly/xai-slack

Visual Analytics Perspectives
on **Interactive** and **Explainable Machine Learning**

Mennatallah El-Assady

University Konstanz

@manunna_91

el-assady.com