Explainability for Natural Language Processing

Tutorial@AACL-IJCNLP 2020

[2] Shipi Dhanorkar
[1] Lucian Popa
[1] Yunyao Li
[1] Kun Qian
[1] Christine T. Wolf
[1] Anbang Xu

[1] IBM Research – Almaden
[2] Pennnsylvania State University

Scalable Knowledge Intelligence

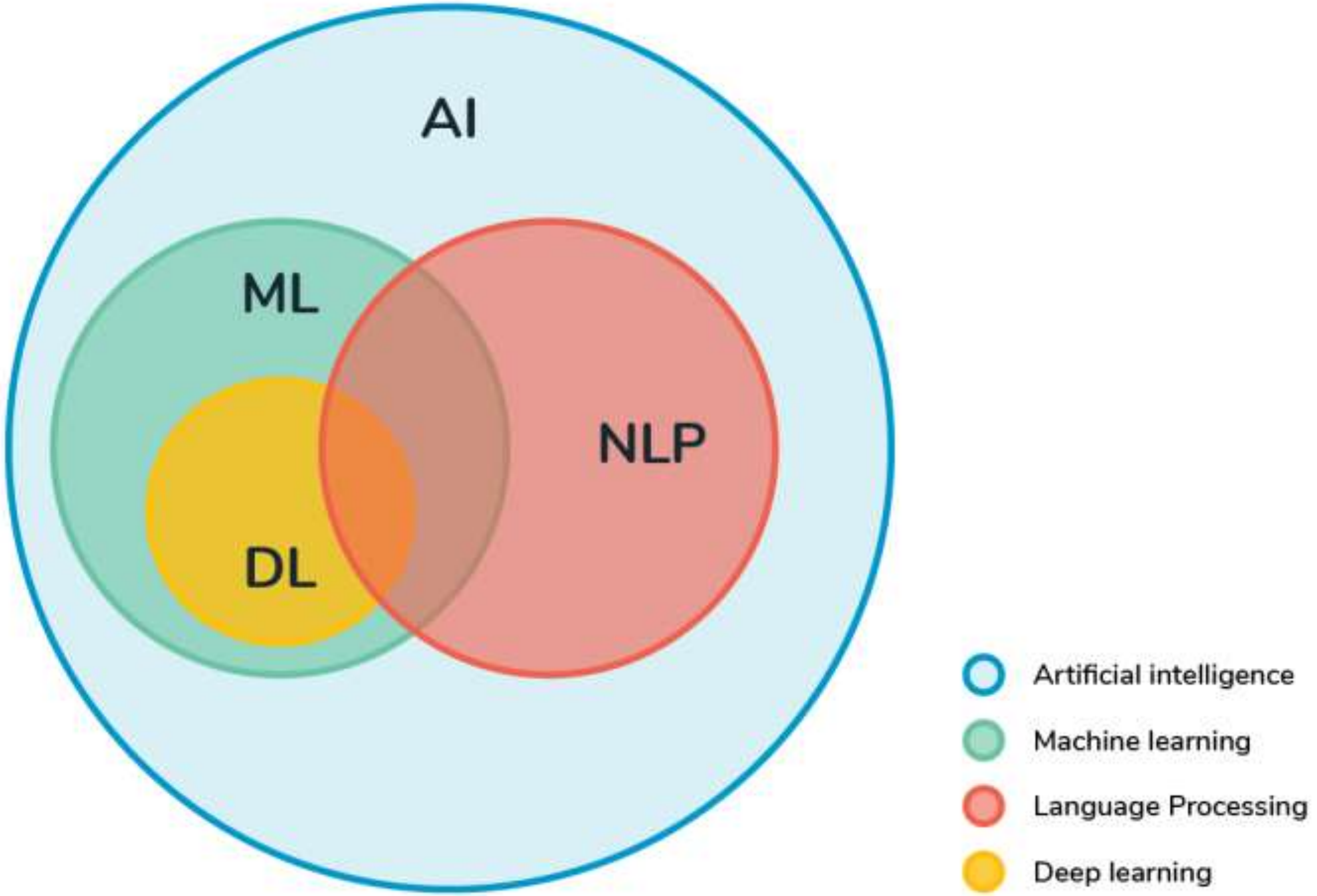
Outline of this Tutorial

- **PART I -** Introduction
- **PART II -** Current State of XAI Research for NLP
- **PART III –** Explainability and Case Study
- **PART IV -** Open Challenges & Concluding Remarks

PART I - Introduction

Yunyao Li

AI vs. NLP



source: https://becominghuman.ai/alternative-nlp-method-9f94165802ed

AI is Becoming Increasing Widely Adopted

Forbes

How Amex Uses AI To Automate 8 Billion Risk Decisions (And Achieve 50% Less Fraud)

TIMES of SAN DIEGO

UCSD Team Uses Artificial Intelligence to Predict Loneliness in Older Adults

Artificial Intelligence Is Ready For Prime Time, But Needs Full Executive Support

F Forbes | 6 days ago

Bay CTO points out foundational **AI** capabilities are quickly becoming table stakes. However, **AI** is emerging and aspiring to have broader applications when it is leveraged to augment human tasks.

CISION PR Newswire | News | Products | Contact

News in Focus | Business & Money | Science & Tech | Lifestyle & Health | Policy & Public Interest | People & Culture

TCS Deploys AI-Powered Software to Enable Safe Return to Work for its Employees

Tata Consultancy Services' Software Monitors Employee Risk Profiles on a Daily Basis, Helping Decide Who Can be in Office if Required, and Ensuring Business Continuity, Regulatory Compliance

VB | The Machine | GamesBeat | Jobs | Special Issue

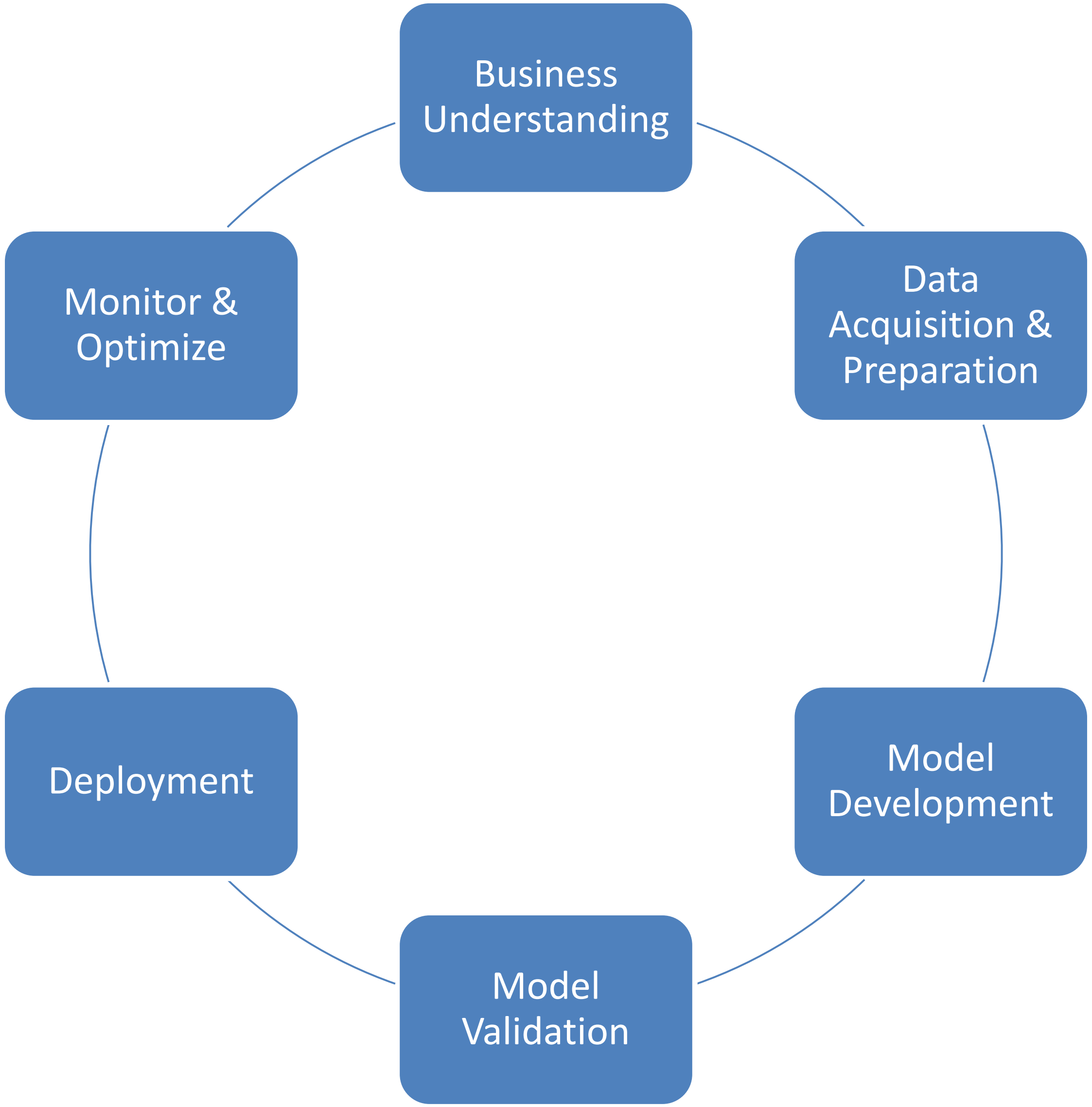
The Machine
Making sense of AI

VB Live

Mastercard's VP of AI talks bots, NLP, and why fintechs need AI for customer service (VB Live)

… …

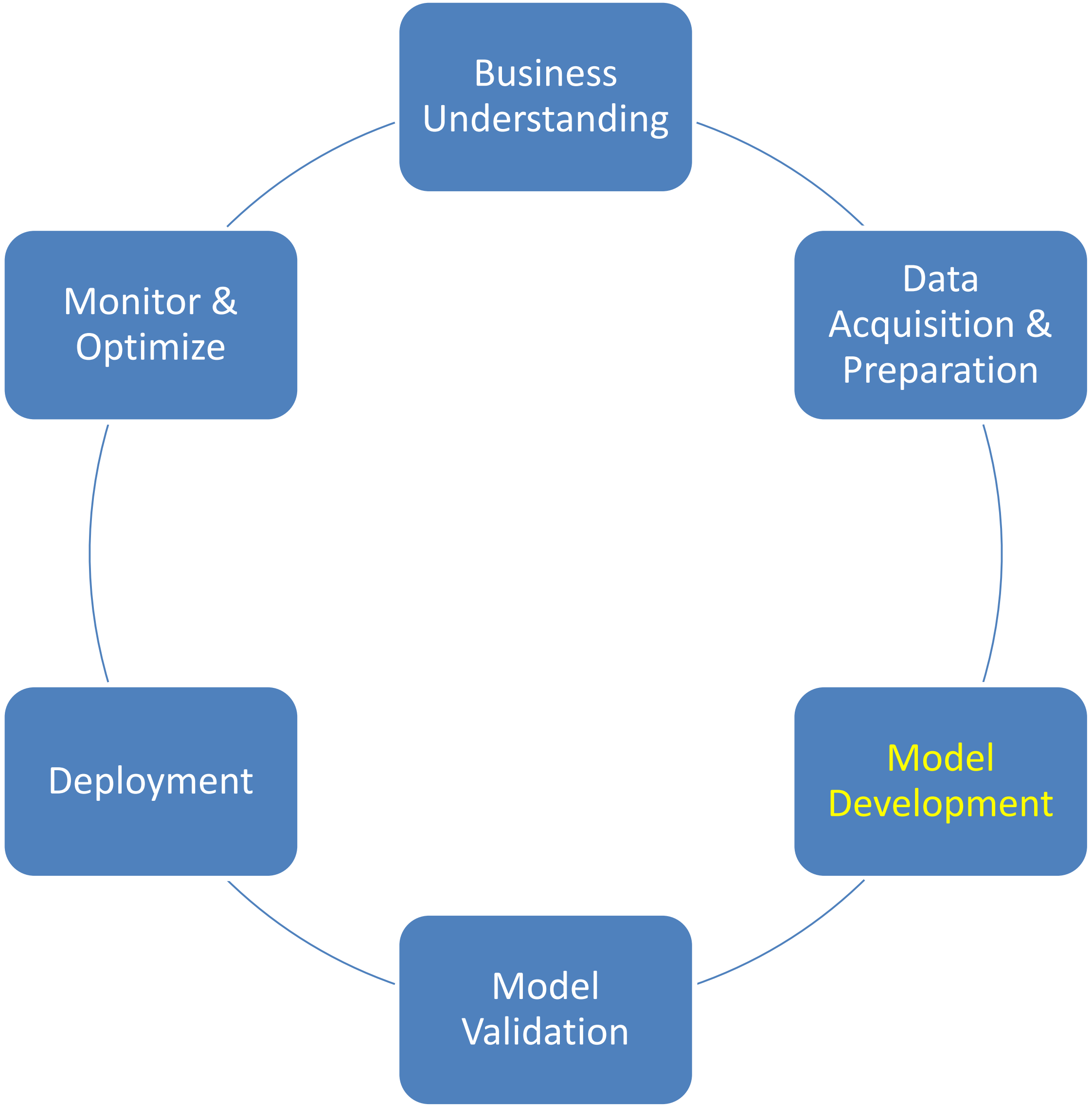
AI Life Cycle

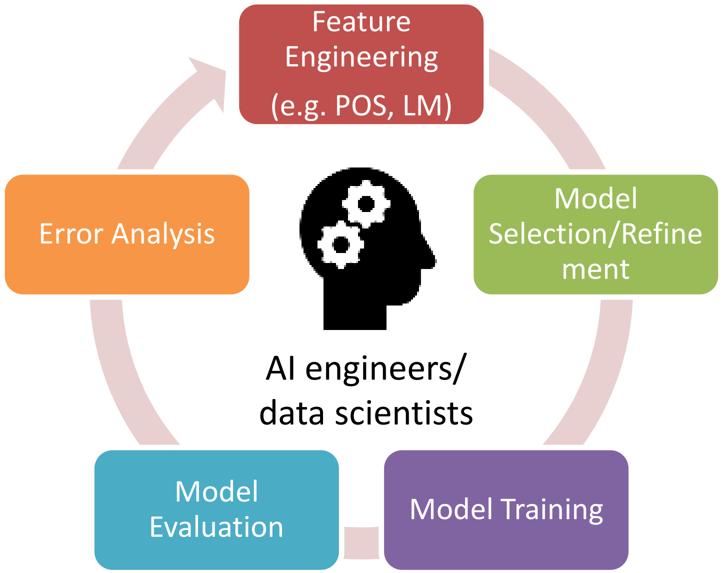


Why Explainablity?

AN ENGINEERING PERSPECTIVE

Why Explainability: Model Development





Error Analysis

e.g. *"$100 is due at signing"*
is not classified as <Obligation>

Root Cause Identification

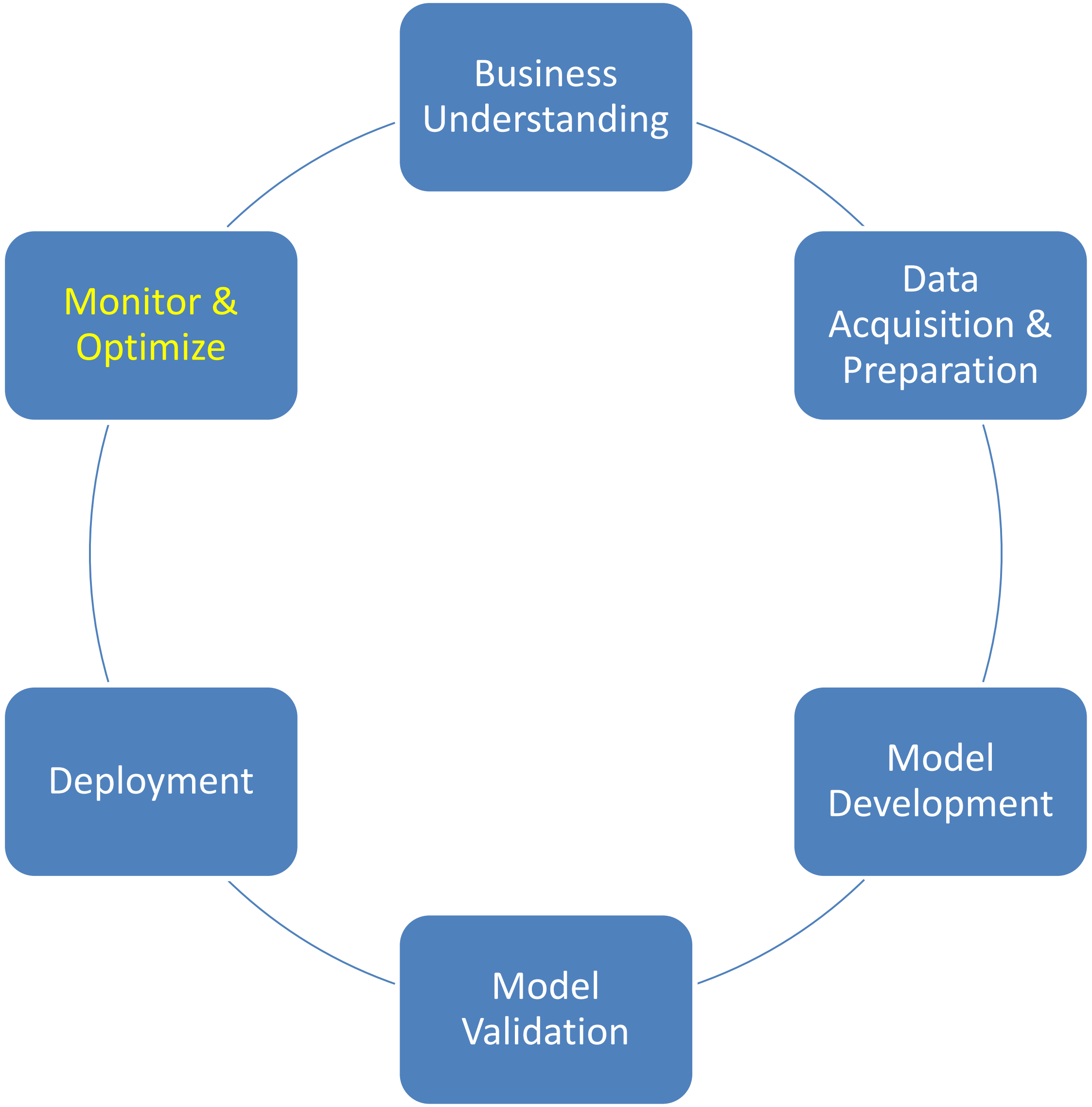
e.g. <Currency> is not considrerd in the model

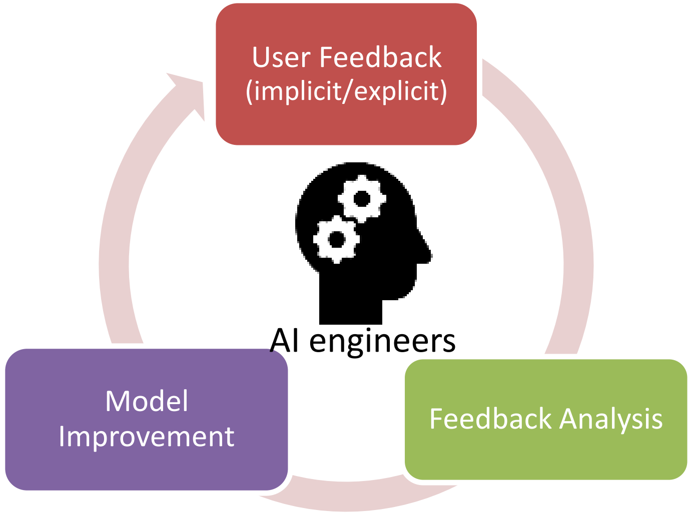
Model Improvement

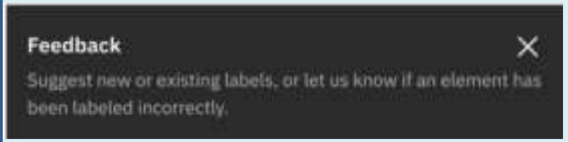
e.g. Include <Currency> as a feature / or introduce entity-aware attention

Source: ModelLens: An Interactive System to Support the Model Improvement Practices of Data Science Teams. CSCW'2019

Why Explainability: Monitor & Optimize







User Feedback

e.g. Clauses related to trademark should not classified as <IP>

Feedback Analysis

e.g. Trademark-related clauses are considered part of <IP> category by The out-of-box model

Model Improvement

e.g. Create a customized model to distinguish trademark-related clauses from other IP-related clauses.

Source: Role of AI in Enterprise Application
https://wp.sigmod.org/?p=2577

Why Explainablity?

A PRACTICAL PERSPECTIVE

Artificial intelligence

We're not ready for AI, says the winner of a new $1m AI prize

Regina Barzilay, the first winner of the Squirrel AI Award, on why the pandemic should be a wake-up call.

by **Will Douglas Heaven** September 23, 2020

The second reason is that we weren't ready. Even in normal circumstances, when people are not under stress, it is difficult to adopt AI tools into a process and make sure it's all properly regulated. In the current crisis, we simply don't have that capacity.

You know, I understand why doctors are conservative: people's lives are on the line. But I do hope that this will be a wake-up call to how unprepared we are to react fast to new threats. As much as I think that AI is the technology of the future, unless we figure out how to trust it, we will not see it moving forward.

https://www.technologyreview.com/2020/09/23/1008757/interview-winner-million-dollar-ai-prize-cancer-healthcare-regulation

Needs for Trust in AI → Needs for Explainability

What does it take?

- Being able to say with certainty how AI reaches decisions
- Building mindful of how it is brought into the workplace.

CIO | UNITED STATES | DIGITAL MAGAZINE | EVENTS | CIO100 2020 | IDG TECH(TALK) COMMUNITY | RESOURCE LIBRARY | NEWSLETTERS

Home › Emerging Technology › Artificial Intelligence

FEATURE

Risky AI business: Navigating regulatory and legal dangers to come

Artificial intelligence poses a wide range of hidden and unknown dangers for enterprises deploying the technology. Here's how to guard against the legal and compliance risks of AI.

By Bob Violino

Contributing Writer, CIO

ZDNet

Big bad data: We don't trust AI to make good decisions

The lack of trust in AI systems comes after a number of bad algorithm-driven decisions.

Why Explainable AI Must Be Grounded In Board Director's Risk Management Practices

Forbes | 30 days ago

All leaders need training on Explainable **AI**. Is your organization ready? Few are. High stakes exist in the accelerated advances in **AI**, underpinning every **industry**, every **business** and life ...

VB | The Machine | GamesBeat | Jobs | Special Issue | Become a Member | Sign In

The Machine
Making sense of AI

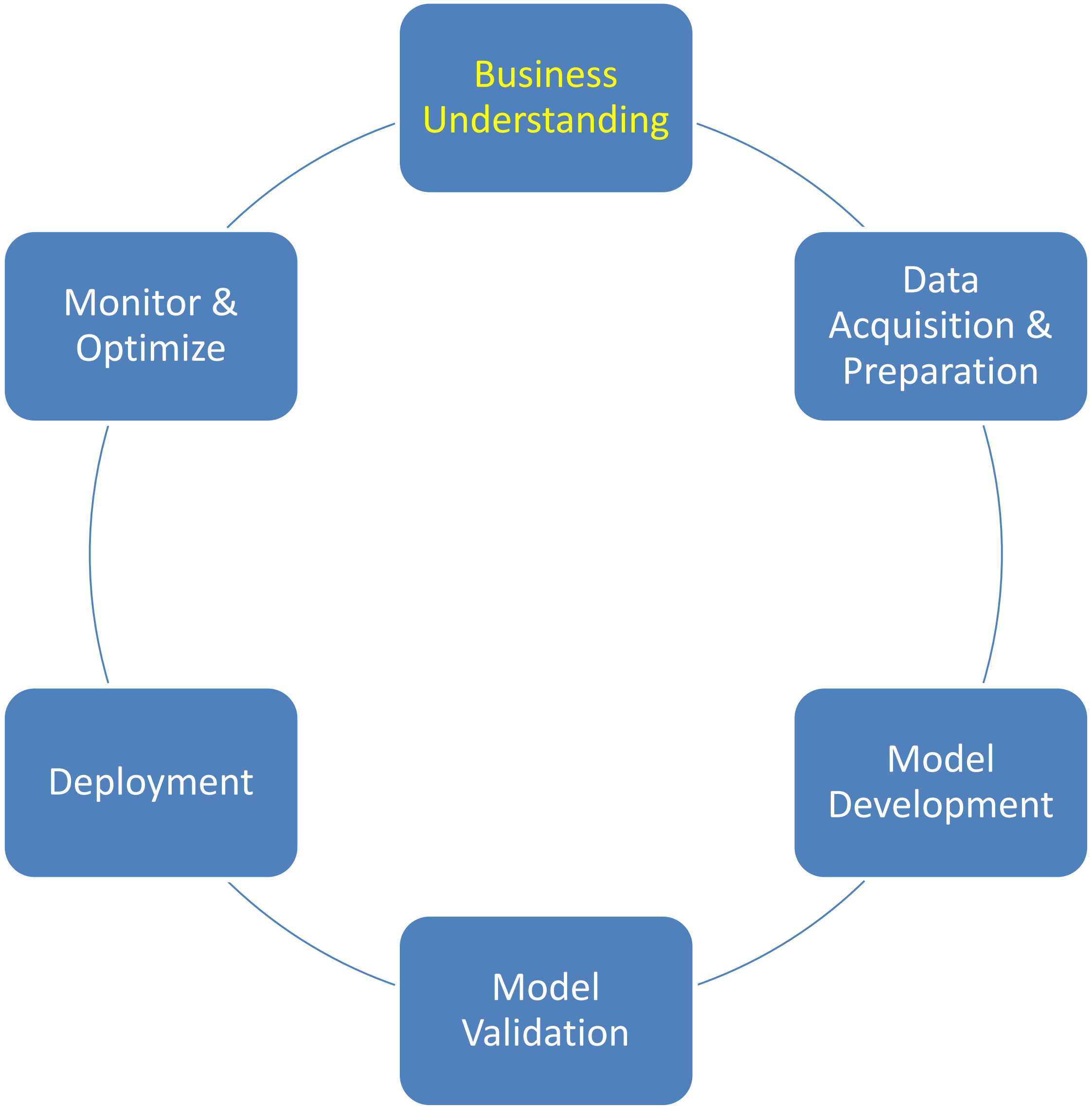
Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI

WIRED

Europe Limits Government by Algorithm. The US, Not So Much

A Dutch court halted a program to identify people more likely to commit benefits fraud. Critics said it discriminated against immigrants and low-income residents.

Why Explainability: Business Understanding



When Explainablity Not Needed

- No significant consequences for unacceptable results
  → E.g., ads, search, movie recommendations
- Sufficiently well-studied and validated in real applications
  → we trust the system's decision, even if it is not perfect

  E.g. postal code sorting, airborne collision avoidance systems

Source: *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints.* Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. EMNLP'2017

Regulations

General Data Protection Regulation (GDPR): Article 22 empowers individuals with the right to demand an explanation of how an automated system made a decision that affects them.

Algorithmic Accountability Act 2019: Requires companies to provide an assessment of the risks posed by the automated decision system to the privacy or security and the risks that contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers

California Consumer Privacy Act: Requires companies to rethink their approach to capturing, storing, and sharing personal data to align with the new requirements by January 1, 2020.

Washington Bill 1655: Establishes guidelines for the use of automated decision systems to protect consumers, improve transparency, and create more market predictability.

Massachusetts Bill H.2701: Establishes a commission on automated decision-making, transparency, fairness, and individual rights.

Illinois House Bill 3415: States predictive data analytics determining creditworthiness or hiring decisions may not include information that correlates with the applicant race or zip code.

Credit: Gade, Geyik, Kenthapadi, Mithal, Taly. *Explainable AI in Industry*, WWW’2020

FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

Using Artificial Intelligence and Algorithms

By: Andrew Smith, Director, FTC Bureau of Consumer Protection I Apr 8, 2020 9:58AM

…

EXPLAIN YOUR DECISION TO THE CONSUMER.

If you deny consumers something of value based on algorithmic decision-making, explain why. Some might say that it's too difficult to explain the multitude of factors that might affect algorithmic decision-making. But, in the credit-granting world, companies are required to disclose to the consumer the principal reasons why they were denied credit, and it's not good enough simply to say "your score was too low" or "you don't meet our criteria." You need to be specific (e.g., "you've been delinquent on your credit obligations" or "you have an insufficient number of credit references"). This means that you must know ***what*** data is used in your model and ***how*** that data is used to arrive at a decision. And you must be able to explain that to the consumer. If you are using AI to make decisions about consumers in any context, consider how you would explain your decision to your customer if asked.

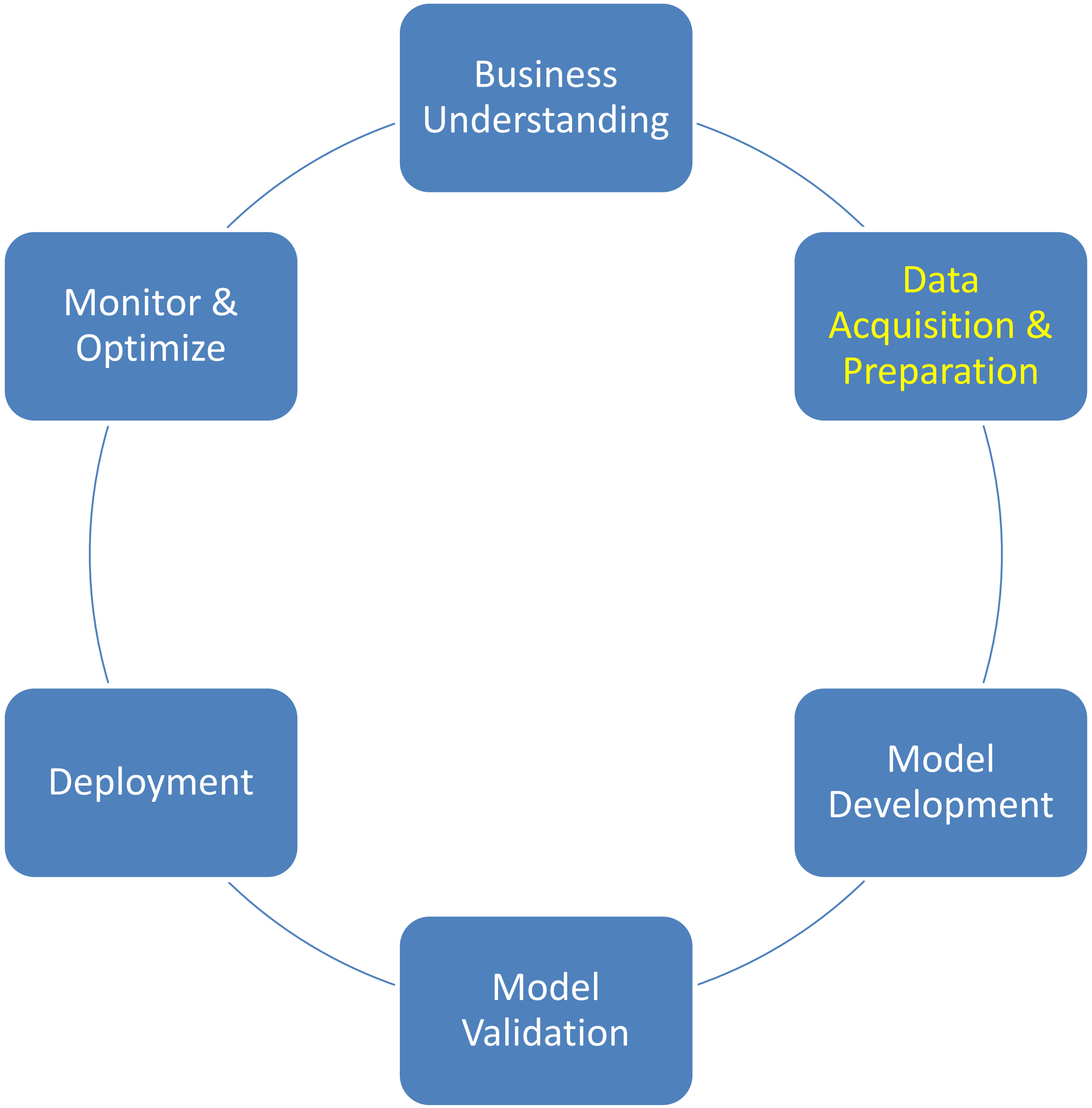
If you use algorithms to assign risk scores to consumers, also disclose the key factors that affected the score, rank ordered for importance. Similar to other algorithmic decision-making, scores are based on myriad factors, some of which may be difficult to explain to consumers. For example, if a credit score is used to deny someone credit, or offer them less favorable terms, the law requires that consumers be given notice, a description of the score (its source, the range of scores under that credit model), and at least four key factors that adversely affected the credit score, listed in the order of their importance based on their effect on the credit score.

If you might change the terms of a deal based on automated tools, make sure to tell consumers. More than a decade ago, the FTC alleged that subprime credit marketer CompuCredit violated the FTC Act by deceptively failing to disclose that it used a behavioral scoring model to reduce consumers' credit limits. For example, if cardholders used their credit cards for cash advances or to make payments at certain venues, such as bars, nightclubs, and massage parlors, they might have their credit limit reduced. The company never told consumers that these purchases could reduce their credit limit – neither at the time they signed up nor at the time they reduced the credit limit. That decade-old matter is just as important today. If you're going to use an algorithm to change the terms of the deal, tell consumers.

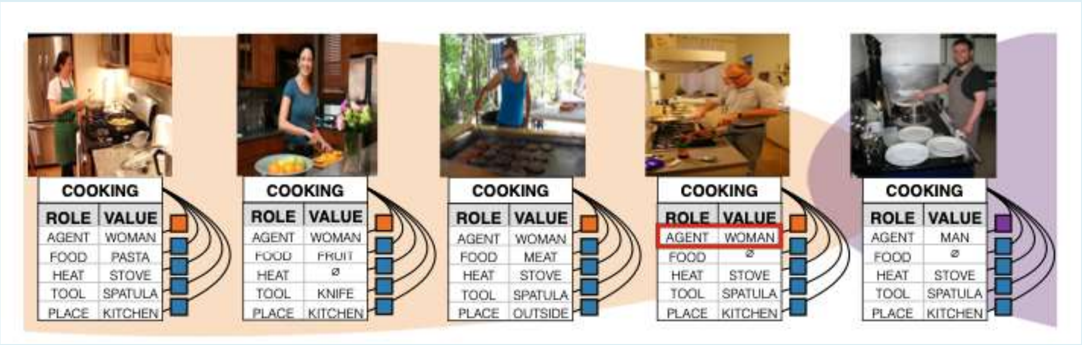
…

Source: https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms

Why Explainability: Data



- Datasets may contain significant bias
- Models can further amplify existing bias

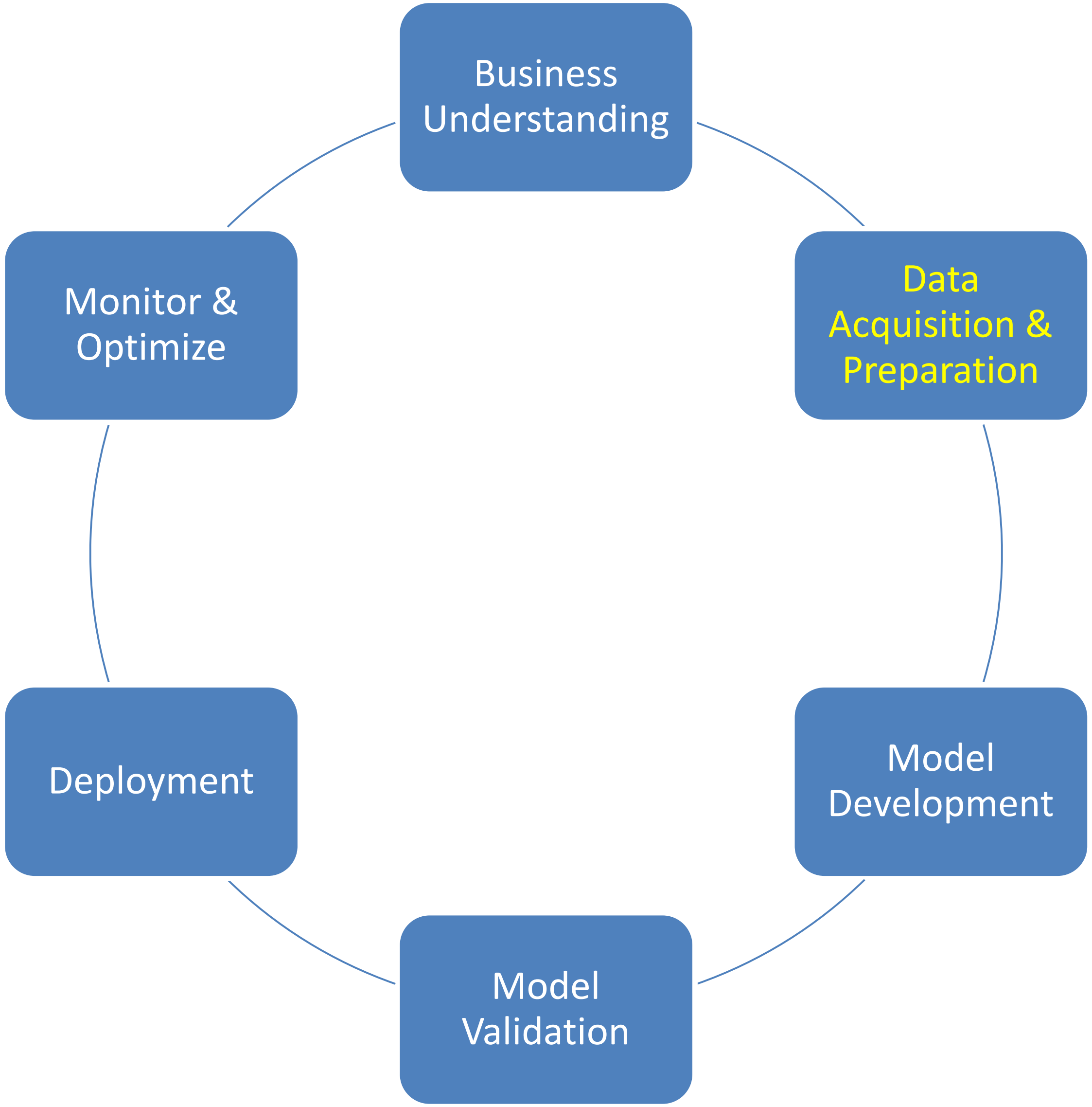


agent role in *cooking* images

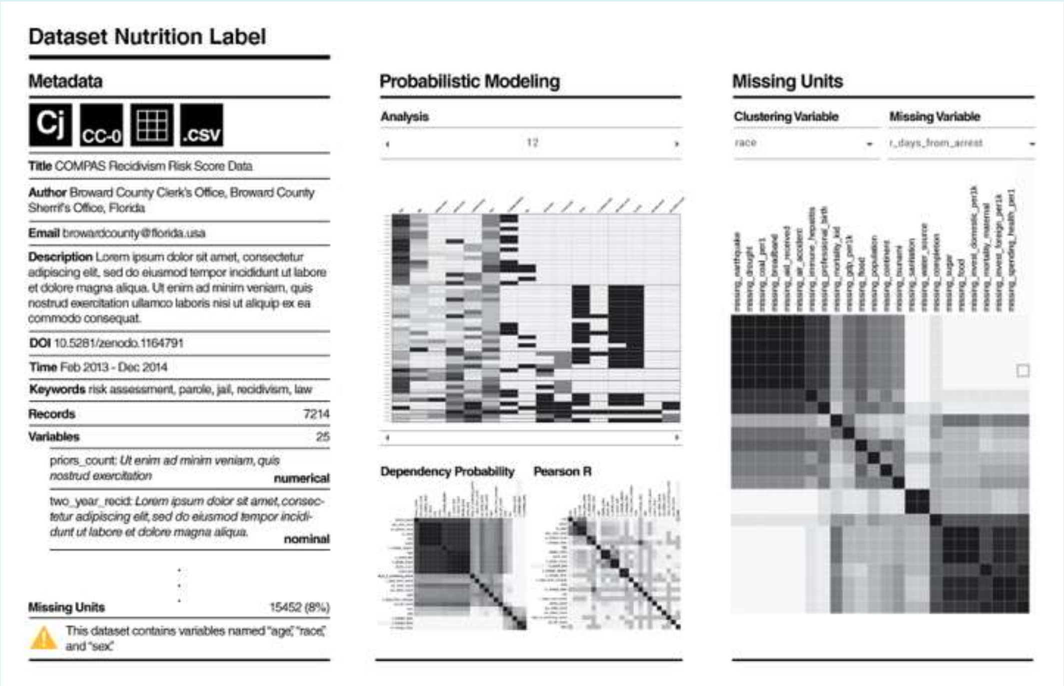
	man	woman
Training Data	33%	67%
Model Prediction	16% (↓17%)	84% (↑17%)

Source: *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints.* Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. EMNLP'2017

Why Explainability: Data



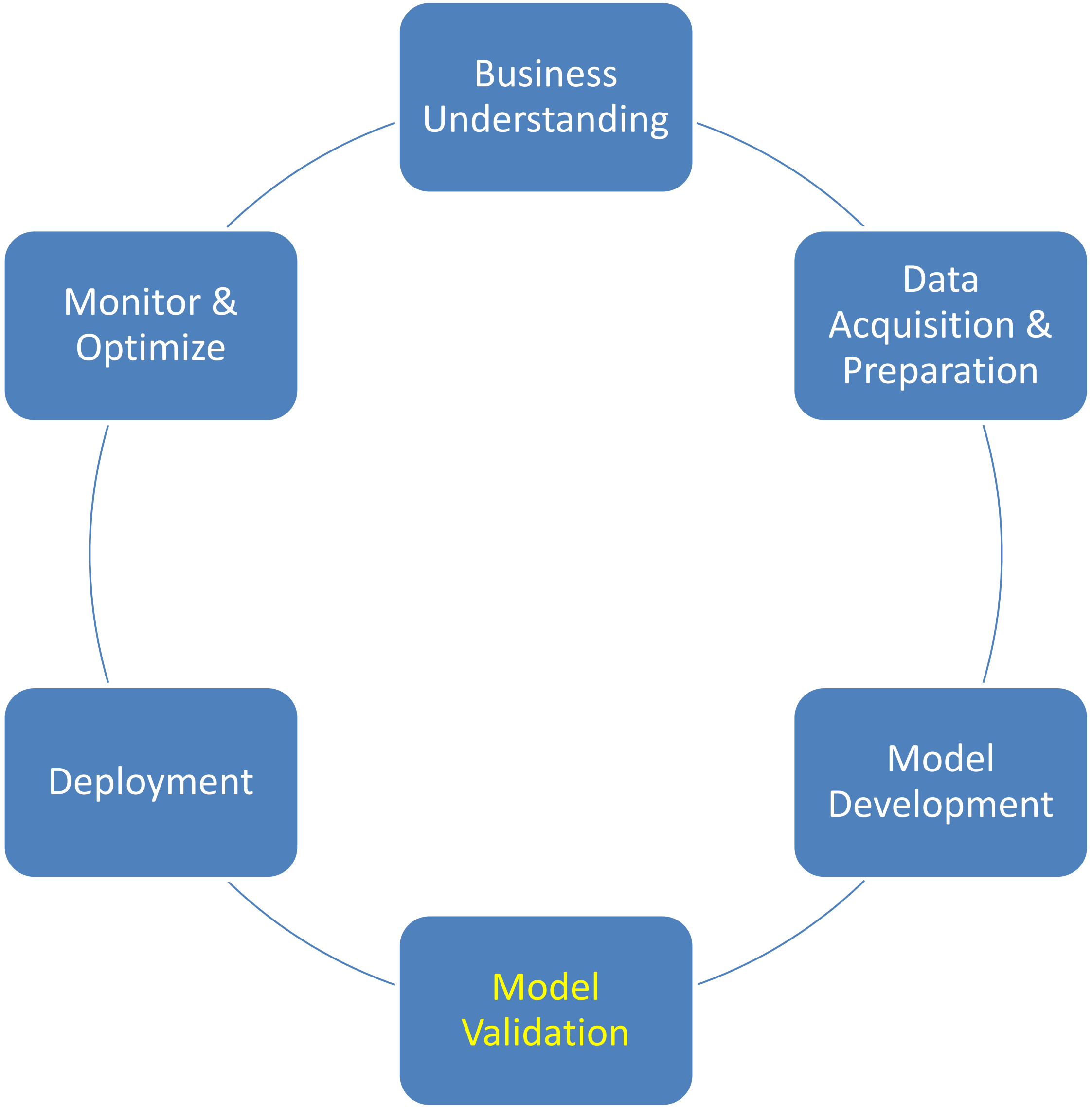
- Explicitly capture measures to drive the creation of better, more inclusive algorithms.



Source: https://datanutrition.org

More: *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*. Emily M. Bender, Batya Friedman TACL'2018

Why Explainability: Model Validation



A core element of model risk management (MRM)

- Verify models are performing as intended
- Ensure models are used as intended

Defective Model

The New York Times

Knight Capital Says Trading Glitch Cost It $440 Million

BY NATHANIEL POPPER AUGUST 2, 2012 9:07 AM 356

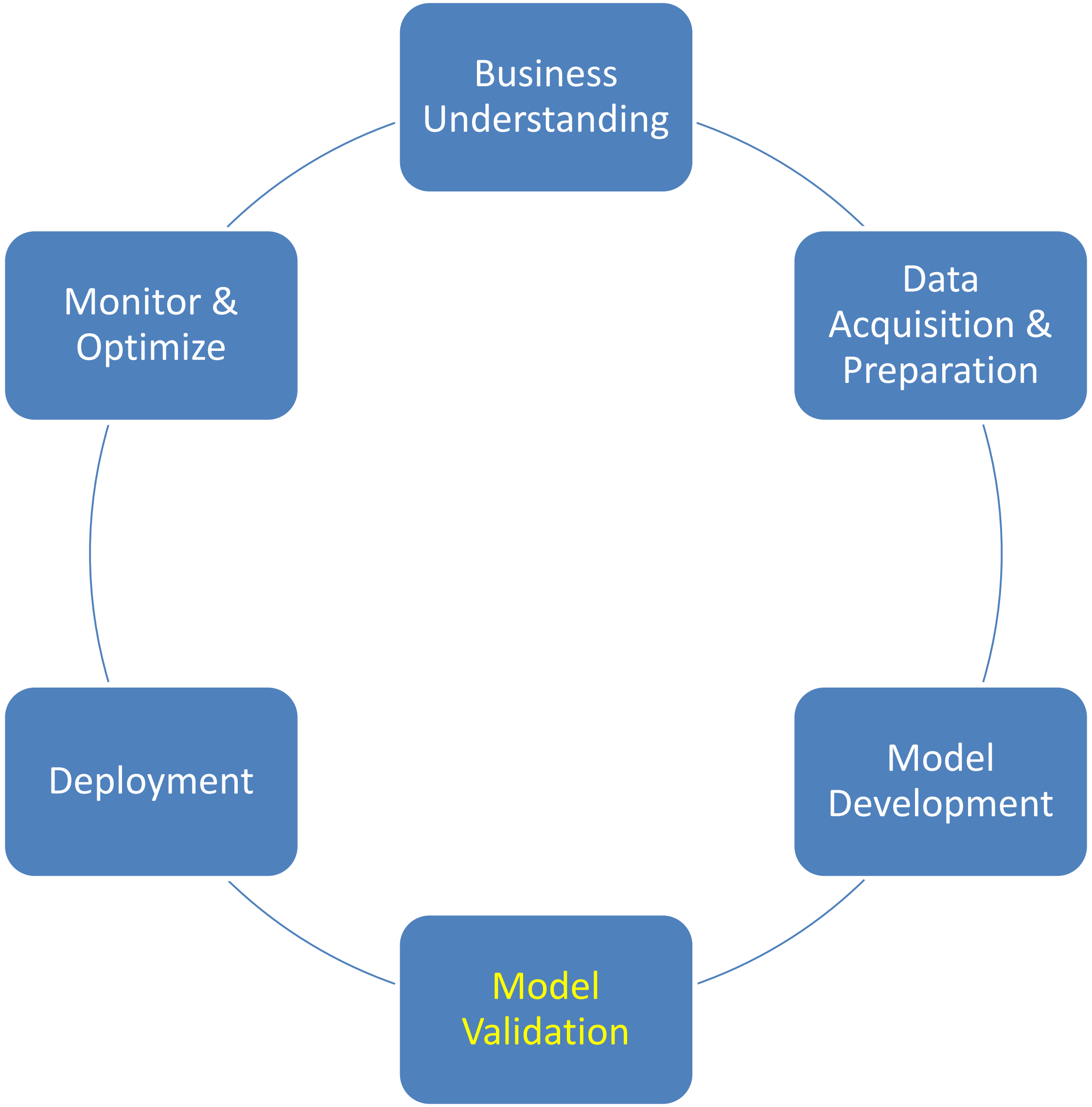
$10 million a minute.

That's about how much the trading problem that set off turmoil on the stock market on Wednesday morning is already costing the trading firm.

The Knight Capital Group announced on Thursday that it lost $440 million when it sold all the stocks it accidentally bought Wednesday morning because a computer glitch.

Source: https://dealbook.nytimes.com/2012/08/02/knight-capital-says-trading-mishap-cost-it-440-million/
https://www.mckinsey.com/business-functions/risk/our-insights/the-evolution-of-model-risk-management

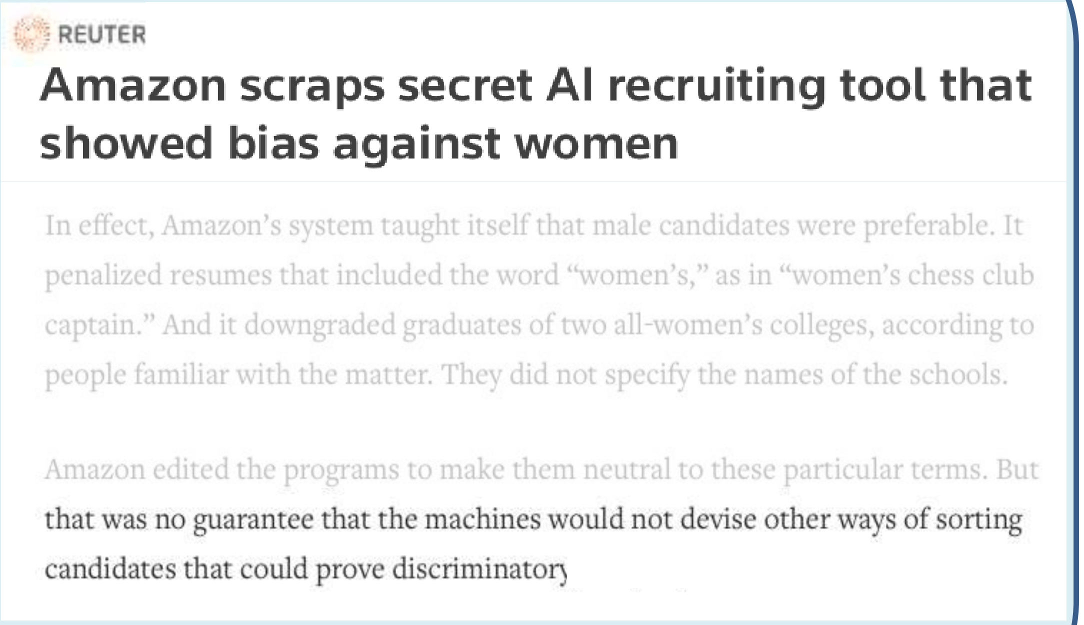
Why Explainability: Model Validation



A core element of model risk management (MRM)

- Verify models are performing as intended
- Ensure models are used as intended

Defective Model



REUTER

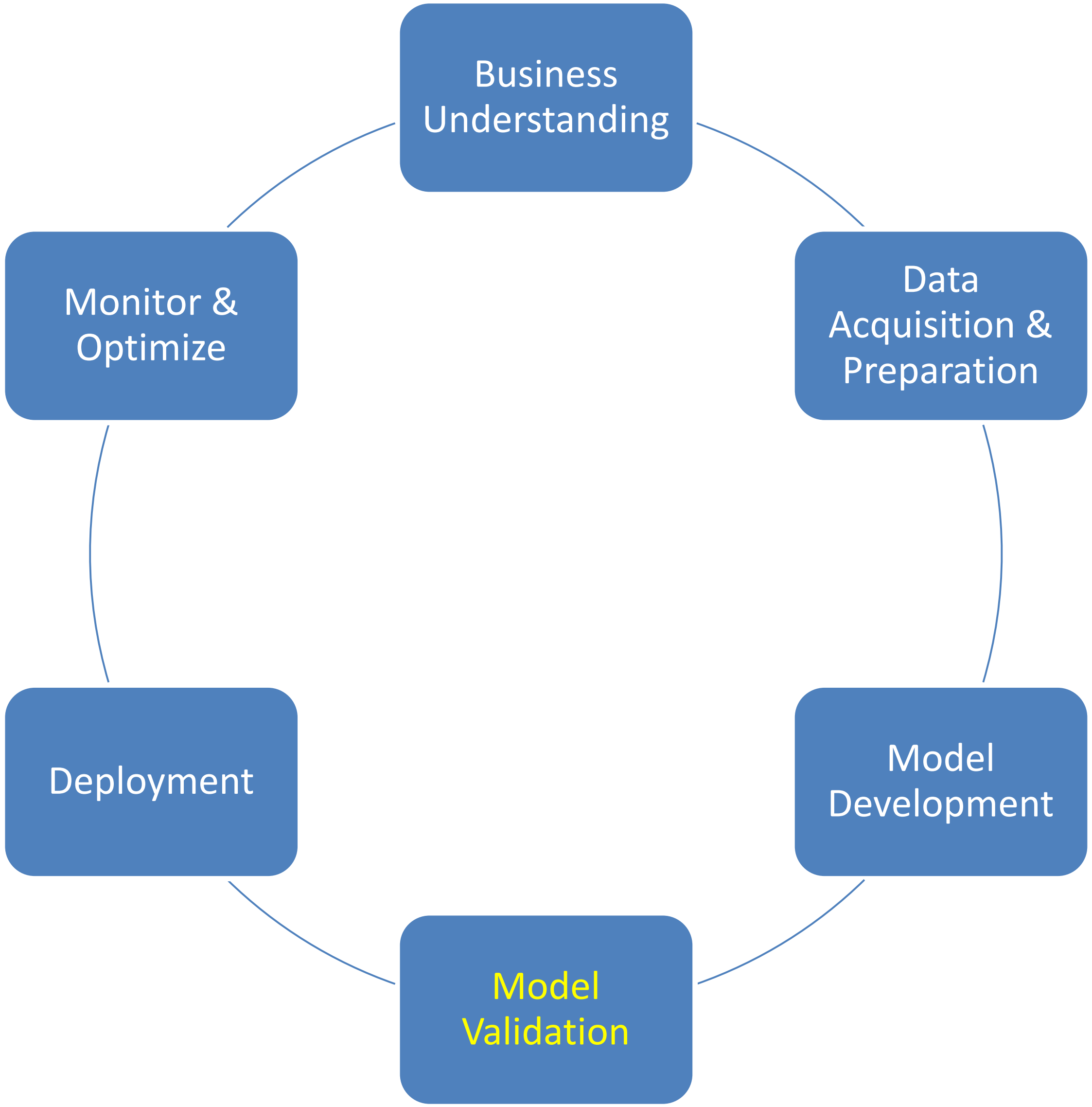
Amazon scraps secret AI recruiting tool that showed bias against women

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory

Source: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

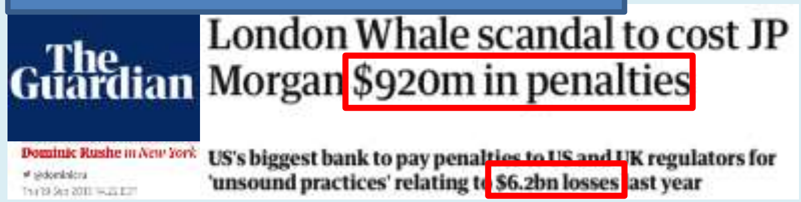
Why Explainability: Model Validation



A core element of model risk management (MRM)

- Verify models are performing as intended
- Ensure models are used as intended

Incorrect Use of Model

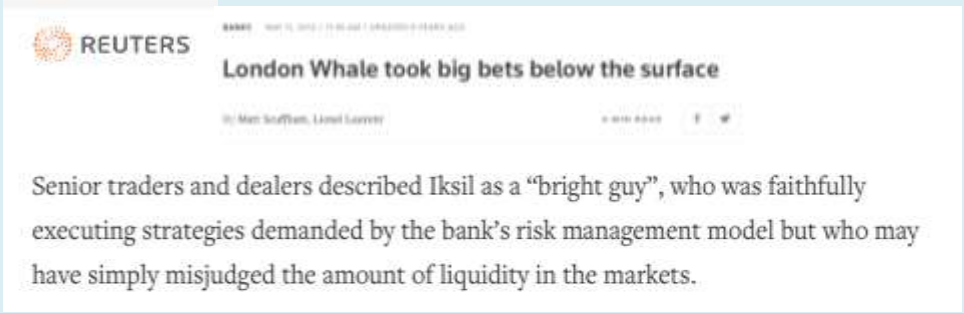


The Guardian

London Whale scandal to cost JP Morgan $920m in penalties

Dominic Rushe in New York

US's biggest bank to pay penalties to US and UK regulators for 'unsound practices' relating to $6.2bn losses last year



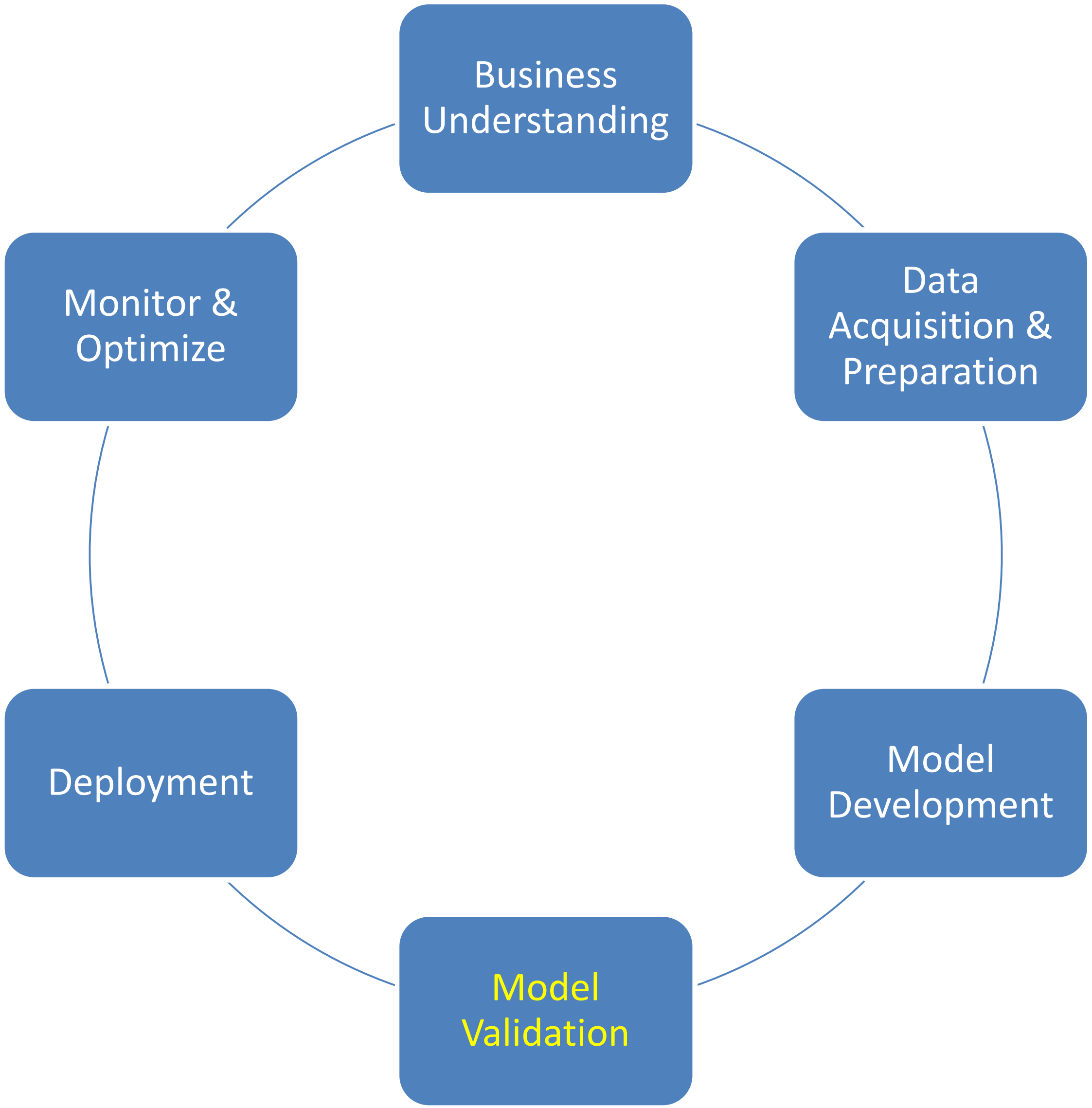
REUTERS

London Whale took big bets below the surface

Senior traders and dealers described Iksil as a "bright guy", who was faithfully executing strategies demanded by the bank's risk management model but who may have simply misjudged the amount of liquidity in the markets.

Source: https://www.theguardian.com/business/2013/sep/19/jp-morgan-920m-fine-london-whale
https://financetrainingcourse.com/education/2014/04/london-whale-casestudy-timeline/
https://www.reuters.com/article/us-jpmorgan-iksil/london-whale-took-big-bets-below-the-surface-idUSBRE84A12620120511

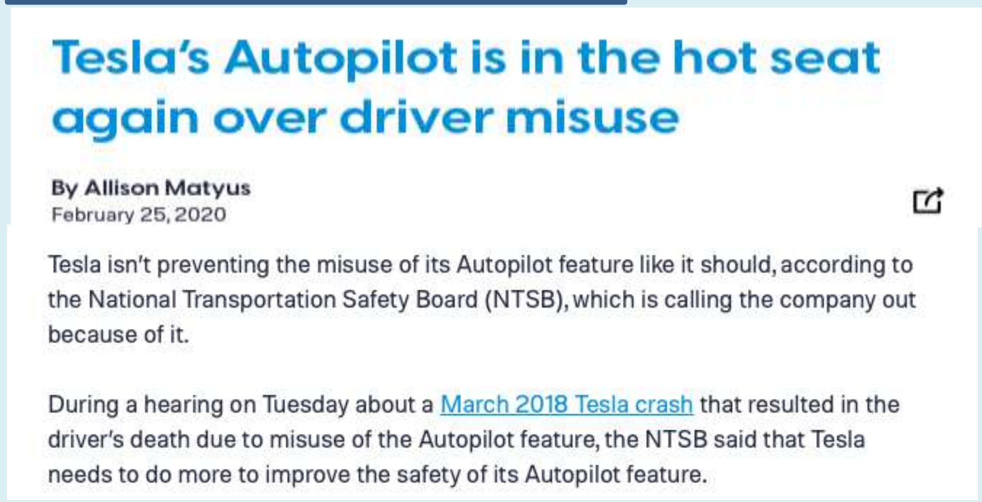
Why Explainability: Model Validation



A core element of model risk management (MRM)

- Verify models are performing as intended
- Ensure models are used as intended

Incorrect Use of Model



Tesla's Autopilot is in the hot seat again over driver misuse

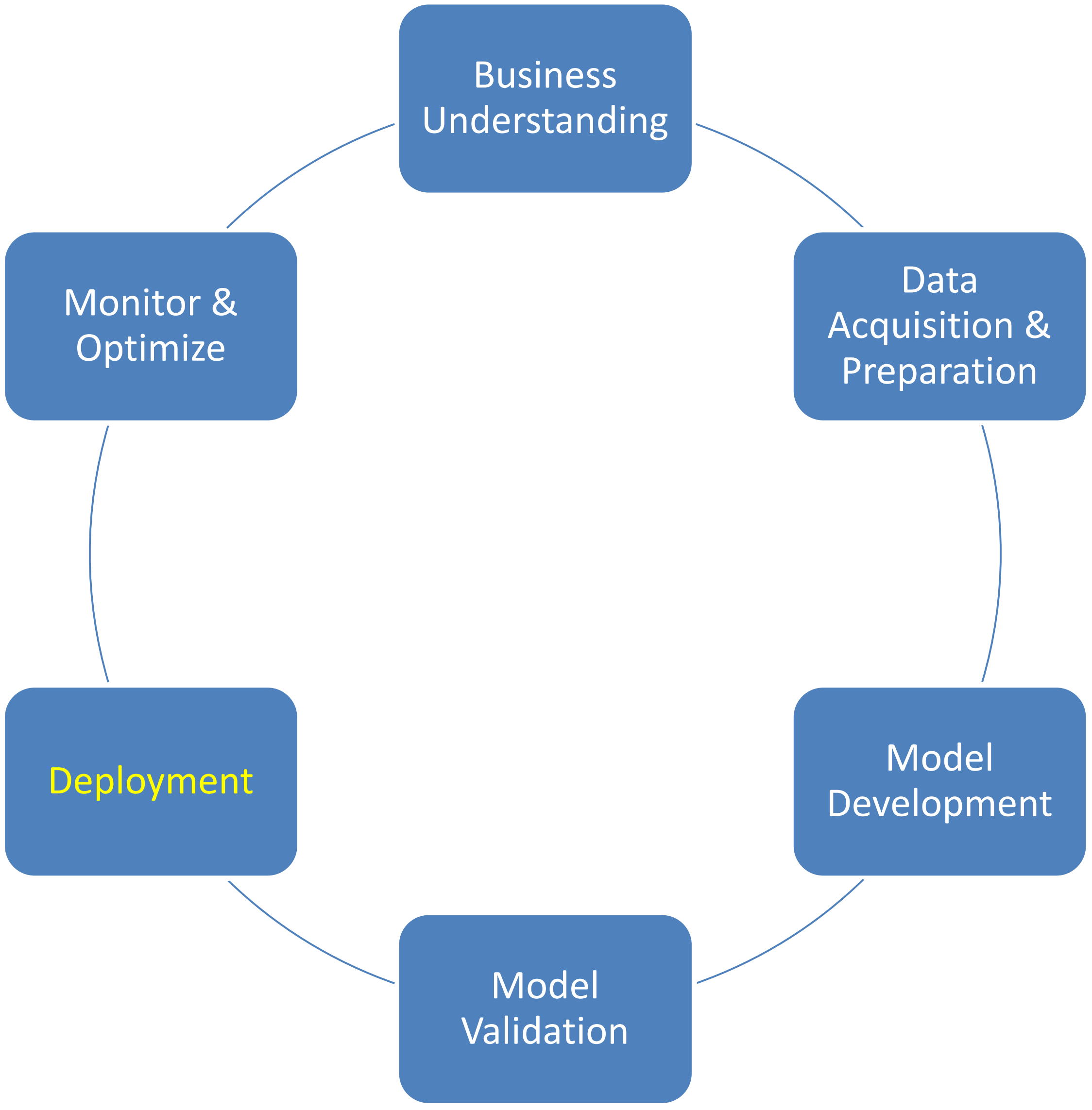
By Allison Matyus
February 25, 2020

Tesla isn't preventing the misuse of its Autopilot feature like it should, according to the National Transportation Safety Board (NTSB), which is calling the company out because of it.

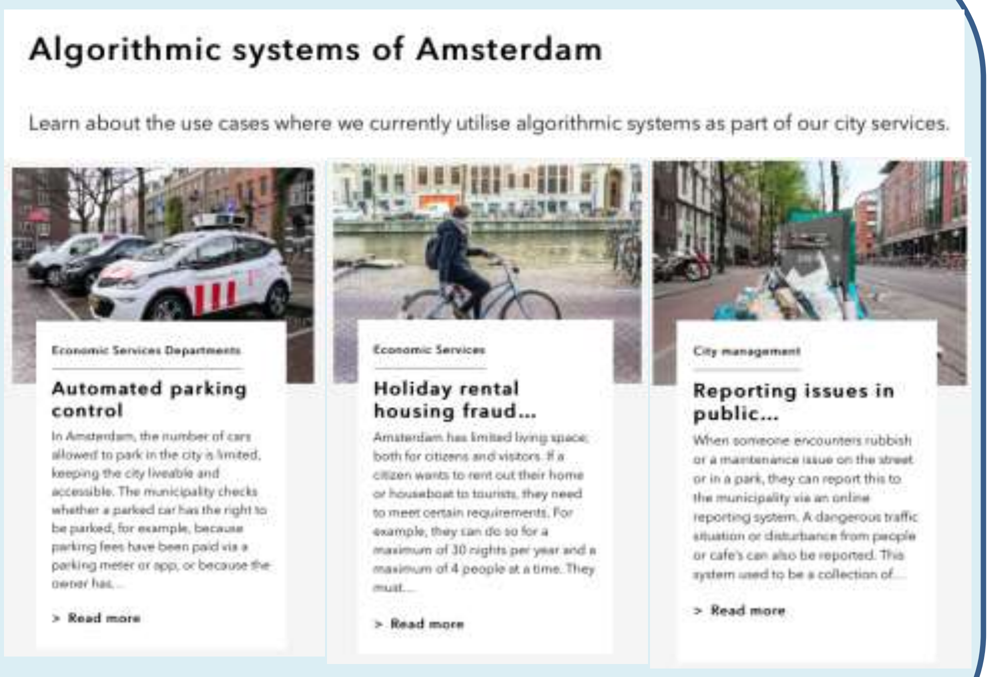
During a hearing on Tuesday about a March 2018 Tesla crash that resulted in the driver's death due to misuse of the Autopilot feature, the NTSB said that Tesla needs to do more to improve the safety of its Autopilot feature.

Source: https://www.digitaltrends.com/cars/tesla-autopilot-in-hot-seat-again-driver-misuse/

Why Explainability: Model Deployment

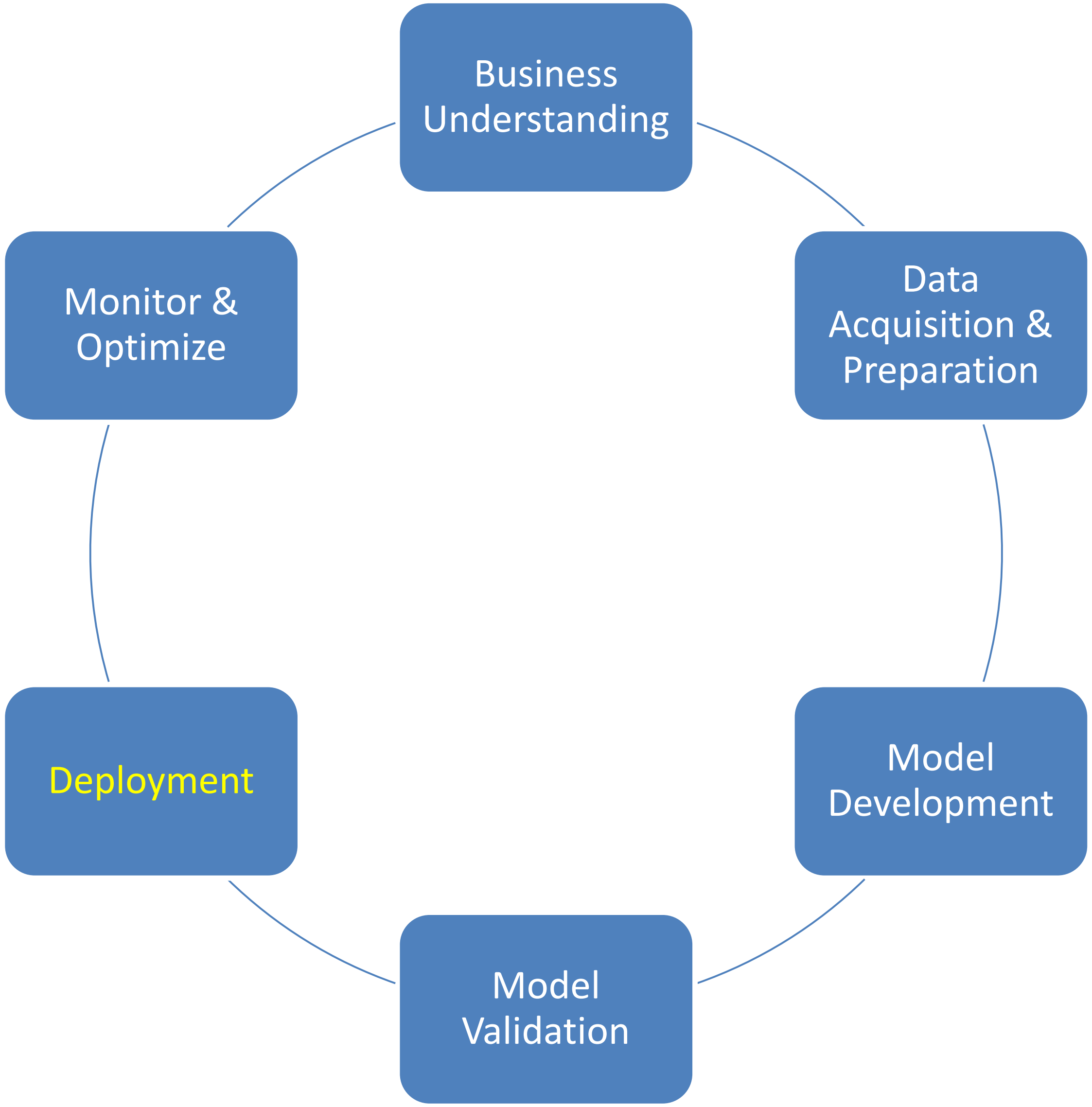


- Compliant to regulatory and legal requirements
- Foster trust with key stakeholders



Source: https://algoritmeregister.amsterdam.nl/en/ai-register/

Why Explainability: Model Deployment



- Compliant to regulatory and legal requirements
- Foster trust with key stakeholders

Algorithmic systems of Amsterdam

Learn about the use cases where we currently utilise algorithmic systems as part of our city services.

More detailed information on the system

Here you can get acquainted with the information used by the system, the operating logic, and its governance in the areas that interest you.

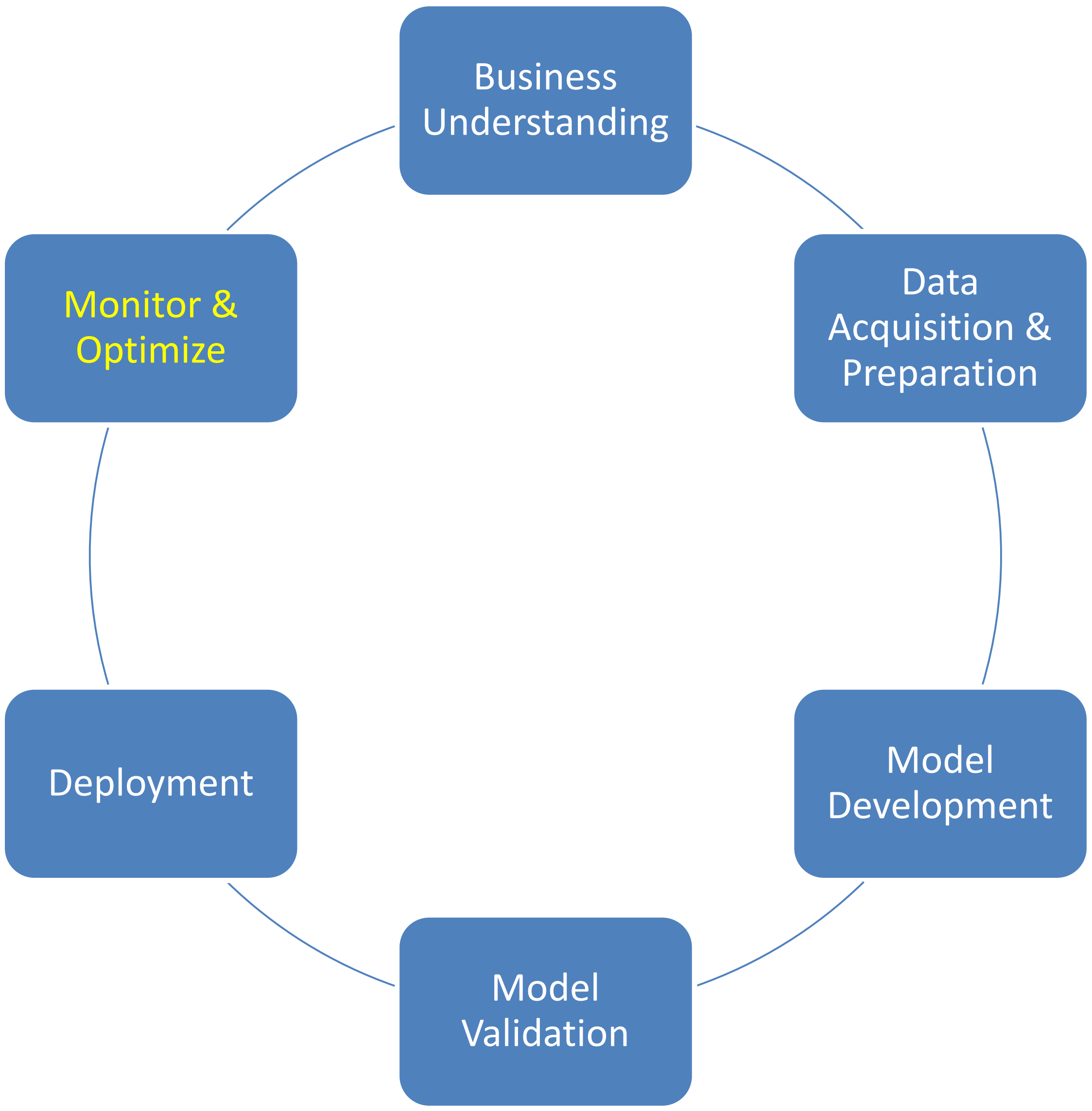
- Datasets — Show More
- Data processing — Show More
- Non-discrimination — Show More
- Human oversight — Show More
- Risks — Show More

Was this information useful?

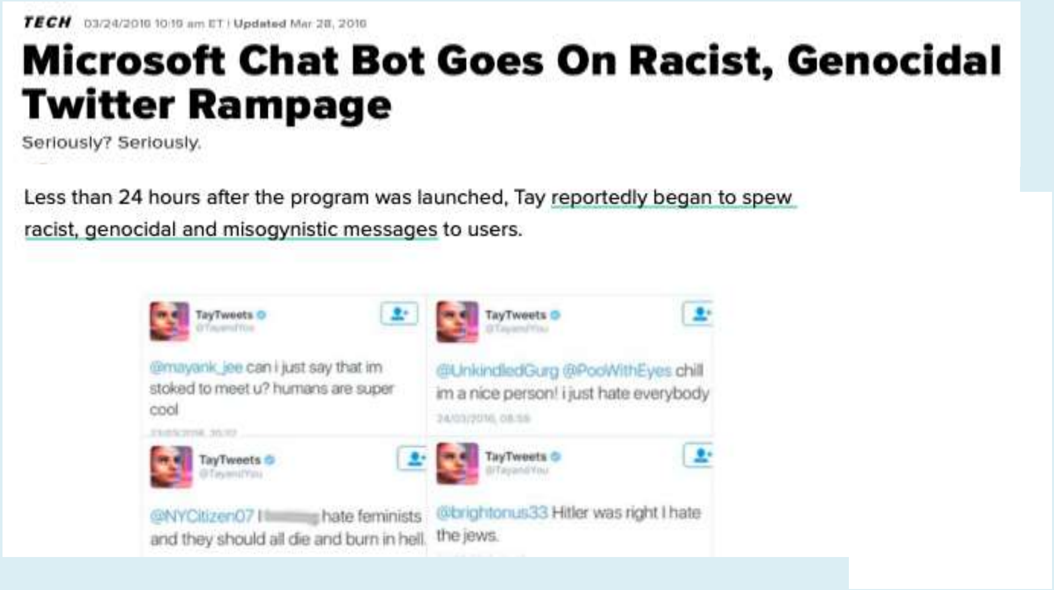
Yes, it was | Partially | Not really

Source: https://algoritmeregister.amsterdam.nl/en/ai-register/

Why Explainability: Monitor & Optimize



- Carry out regular checks over real data to make sure that the systems are working and used as intended.
- Establish KPIs and a quality assurance program to measure the continued effectiveness



TECH 03/24/2016 10:10 am ET | **Updated** Mar 28, 2016

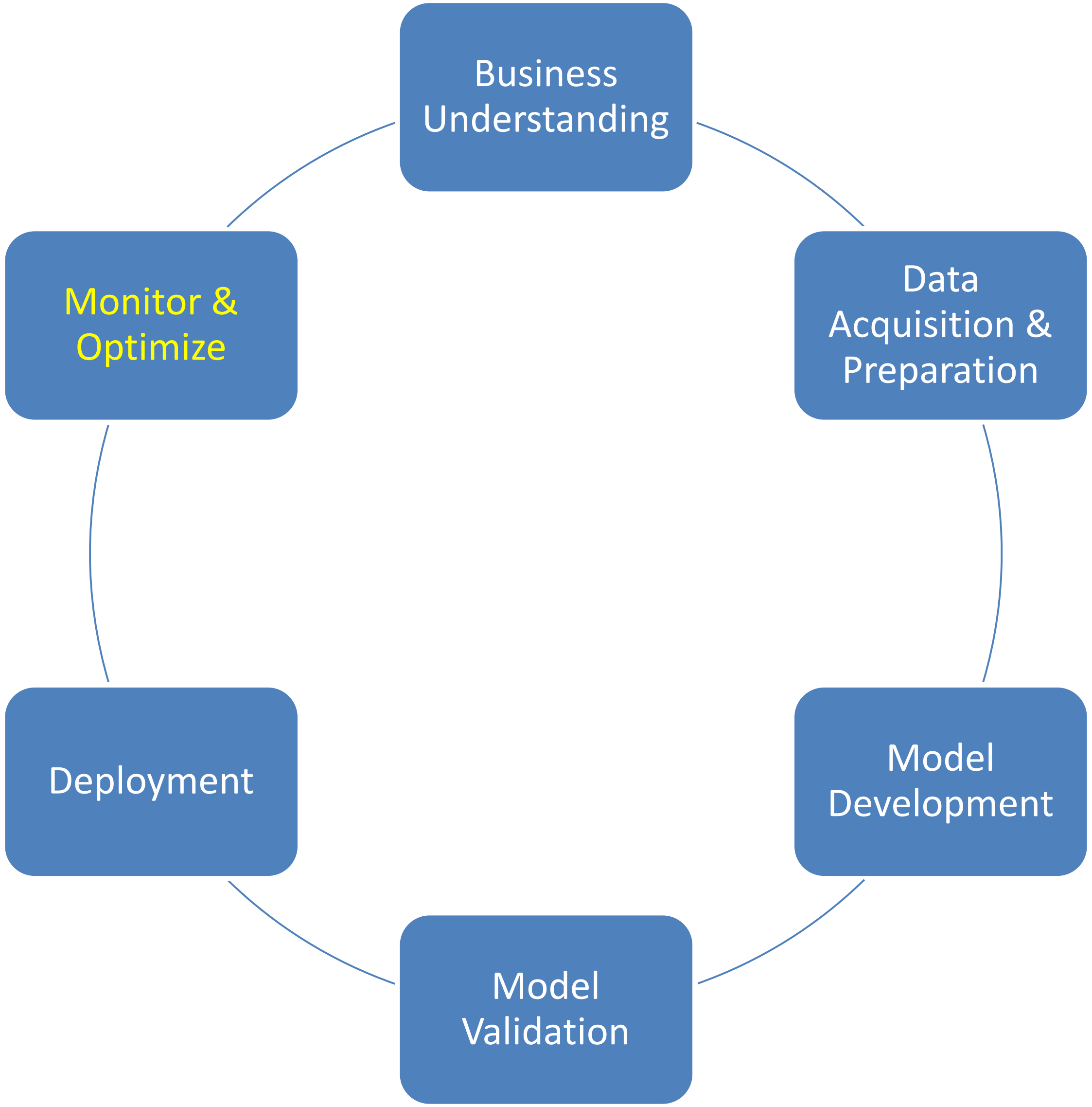
Microsoft Chat Bot Goes On Racist, Genocidal Twitter Rampage

Seriously? Seriously.

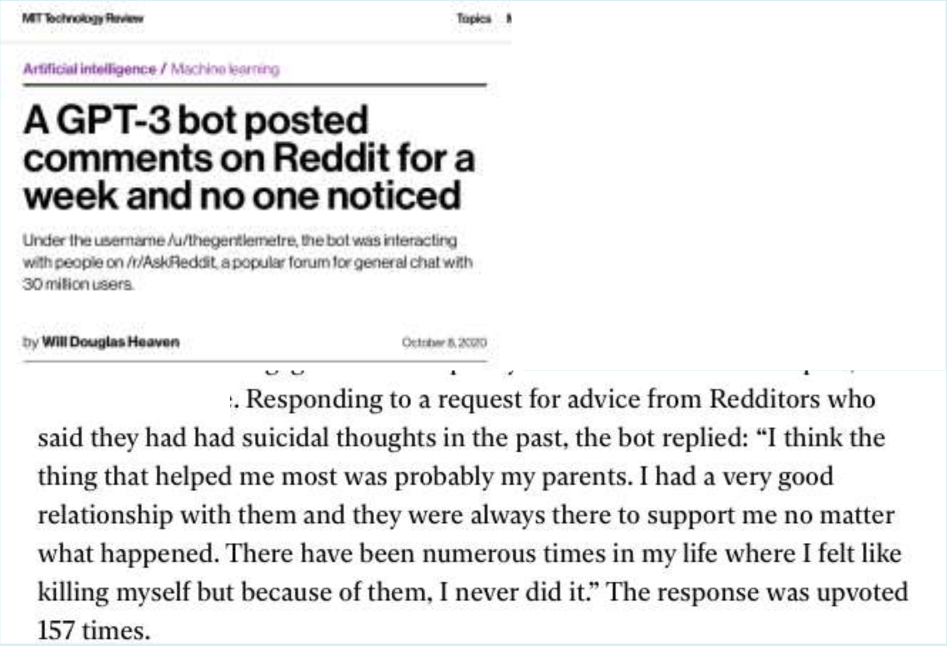
Less than 24 hours after the program was launched, Tay reportedly began to spew racist, genocidal and misogynistic messages to users.

Source: https://www.huffpost.com/entry/microsoft-tay-racist-tweets_n_56f3e678e4b04c4c37615502

Why Explainability: Monitor & Optimize



- Carry out regular checks over real data to make sure that the systems are working and used as intended.
- Establish KPIs and a quality assurance program to measure the continued effectiveness



MIT Technology Review — Topics

Artificial intelligence / Machine learning

A GPT-3 bot posted comments on Reddit for a week and no one noticed

Under the username /u/thegentlemetre, the bot was interacting with people on /r/AskReddit, a popular forum for general chat with 30 million users.

by **Will Douglas Heaven** — October 8, 2020

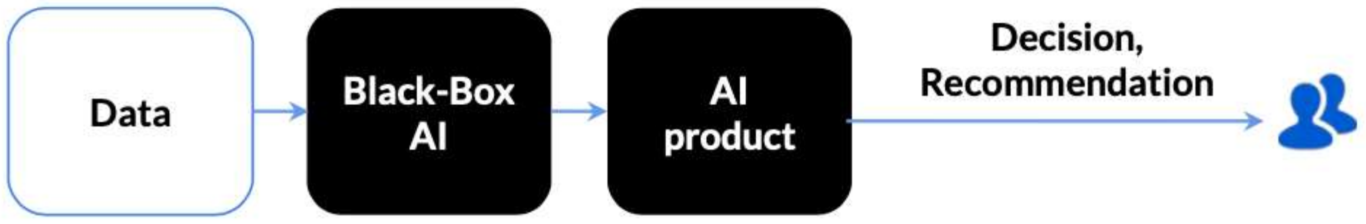
. Responding to a request for advice from Redditors who said they had had suicidal thoughts in the past, the bot replied: "I think the thing that helped me most was probably my parents. I had a very good relationship with them and they were always there to support me no matter what happened. There have been numerous times in my life where I felt like killing myself but because of them, I never did it." The response was upvoted 157 times.

Source: https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/

EXPLAIN EXPLAINATIONS

Explainable AI

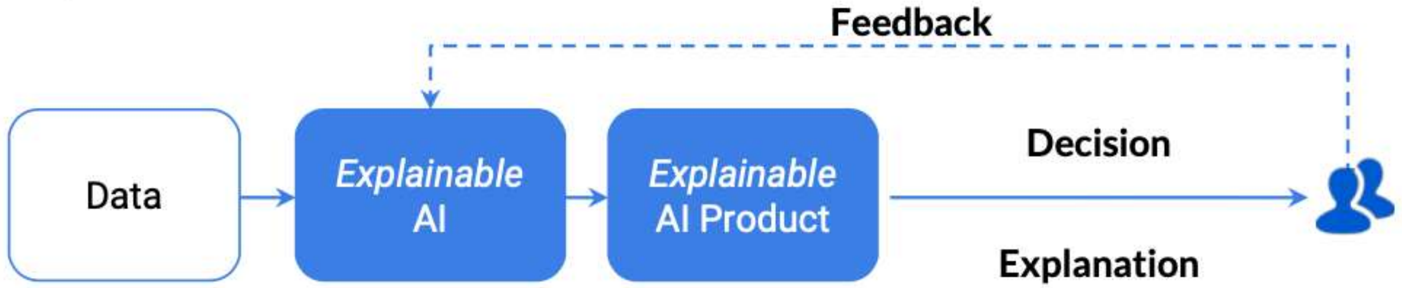
Black Box AI



Confusion with Today's AI Black Box

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

Explainable AI



Clear & Transparent Predictions

- I understand why
- I understand why not
- I know why you succeed or fail
- I understand, so I trust you

Credit: Gade, Geyik, Kenthapadi, Mithal, Taly. *Explainable AI in Industry*, WWW'2020

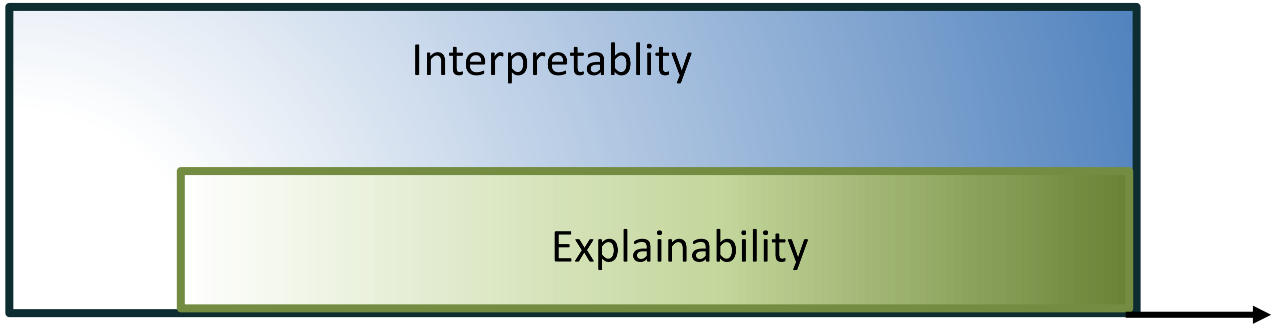
Explainablity vs. Interpretability

Interpretability
The degree to which a human can consistently **predict** the model's result

→ May not know why

Explainability
The degree to which a human can understand the **cause** of the model's result.

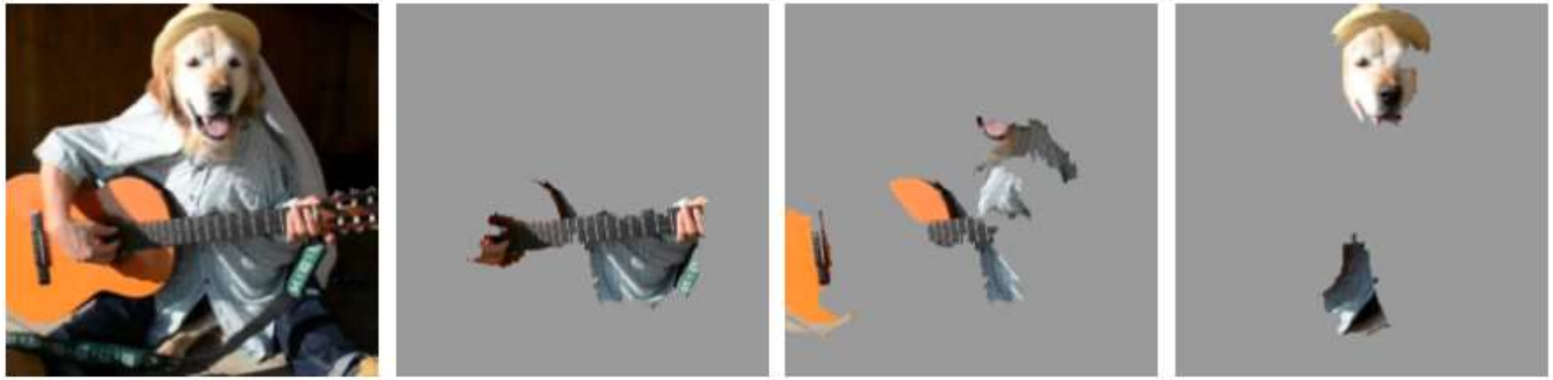
→ Know why



explainable → interpretable
interpretable !→ explainable

Source: *Examples are not enough, learn to criticize! Criticism for interpretability*. NIPS'2016
Explaining Explanations: An Overview of Interpretability of Machine Learning. DASS. 2018

Example: Explanability for Computer Vision



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Source: *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. KDD'2016

Example: Explanability for NLP

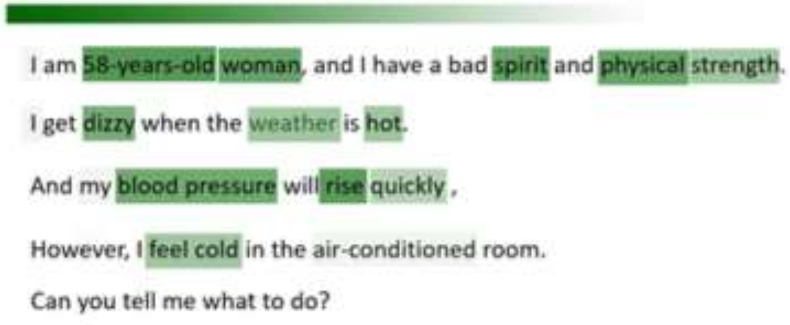


Figure 11. Attention visualization for a document labeled Gynecology.

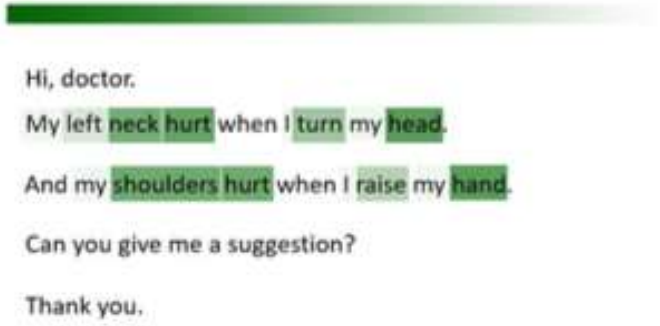
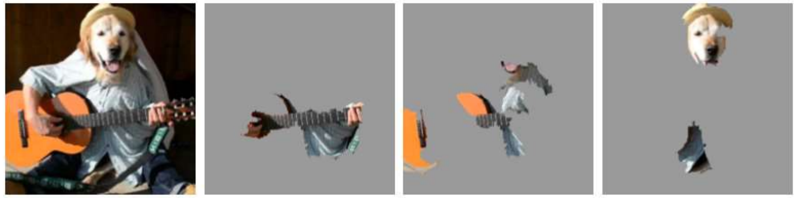


Figure 12. Attention visualization for a document labeled Orthopedics.

Source: *Outpatient Text Classification Using Attention-Based Bidirectional LSTM for Robot-Assisted Servicing in Hospital.* Information 2020

Which One Take You Longer to Understand?



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

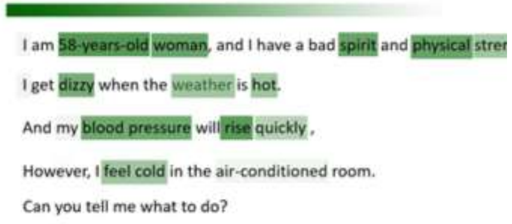


Figure 11. Attention visualization for a document labeled Gynecol

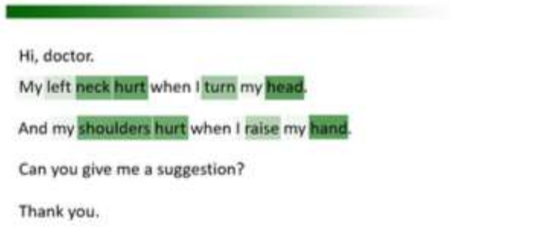
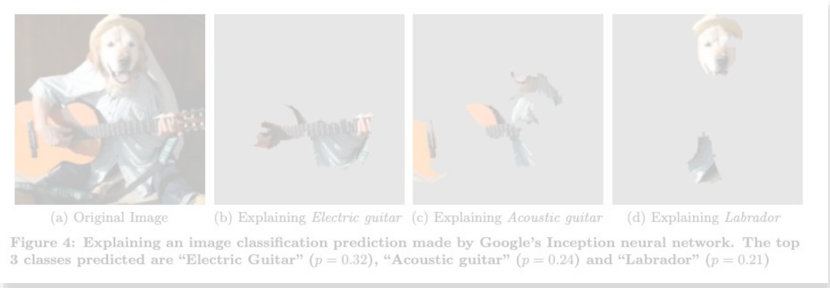


Figure 12. Attention visualization for a document labeled Orthopedics.

Which One Take You Longer to Understand?



VS

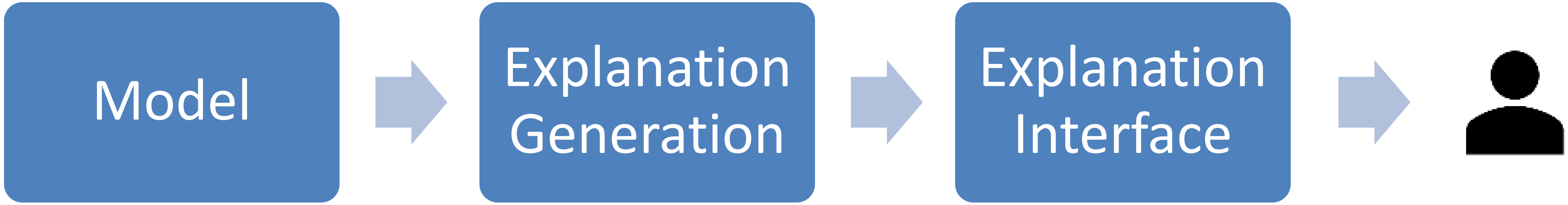
Graph comprehension

- encode the visual array and identify the important visual features
- identify the quantitative facts or relations that those features represent
- relate those quantitative relations to the graphic variables depicted

Source: Why a diagram is (sometimes) worth ten thousand words. Jill H. Larkin Herbert A. Simon. Cognitive Science. 1987

Graphs as Aids to Knowledge Construction: Signaling Techniques for Guiding the Process of Graph Comprehension November 1999. Journal of Educational Psychology

Challenges in Explainability for NLP



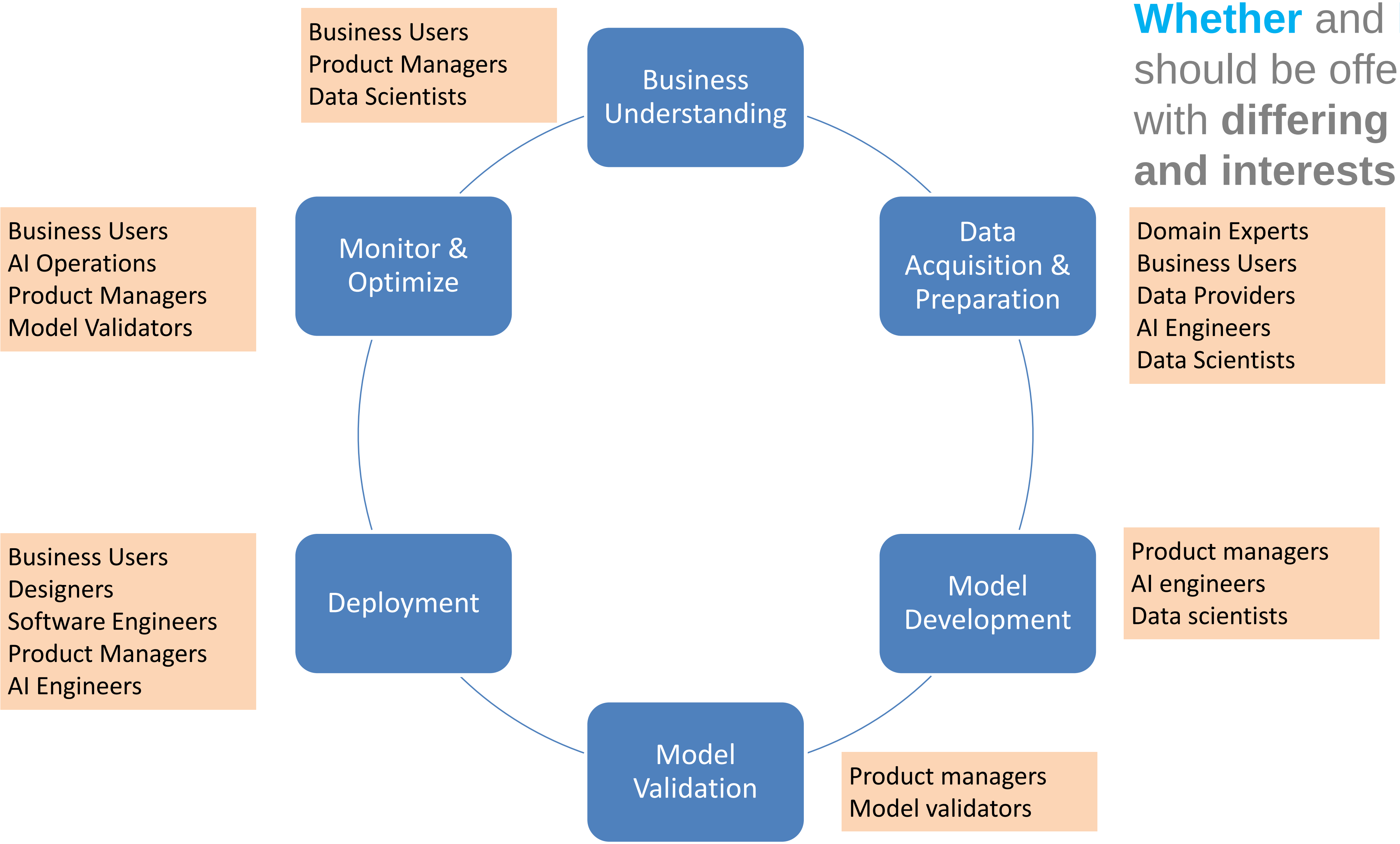
Text modality → Additional challenges in

- Generate explanation
- Display explanation
- Comprehend explanation

Source: SoS TextVis: An Extended Survey of Surveys on Text Visualization. February 2019. Computers 8(1):17

Stakeholders

Whether and **how** explanations should be offered to stakeholders with **differing levels of expertise and interests**.



PART II – Current State of Explainable AI for NLP

Outline – Part II (90-100mins)

- **Literature review methodology**

- **Categorization of different types of explanation**
  - Local vs. Global
  - Self-explaining vs. Post-hoc processing

- **Generating and presenting explanations**
  - Explainability techniques (40 mins)
    - Common operations to enable explainability
  - Visualization techniques (15 mins)
  - More insights (15 mins)

- **Other insights (XNLP website)**
  - Relationships among explainability and visualization techniques
  - Evolution of the XAI research in the past a few years

- **Evaluation of Explanations**

Outline of Part II

- **Literature review methodology**

- **Categorization of different types of explanation**
  - Local vs. Global
  - Self-explaining vs. Post-hoc processing

- **Generating and presenting explanations**
  - Explainability techniques
    - Common operations to enable explainability
  - Visualization techniques

- **Other insights (XNLP website)**
  - Relationships among explainability and visualization techniques
  - Evolution of the XAI research in the past a few years

- **Evaluation of Explanations**

Literature Review Methodology

- **The purpose is NOT to provide an exhaustive list of papers**

- **Major NLP conferences:** ACL, EMNLP, NAACL, COLING
  - And a few representative general AI conferences (AAAI, IJCAL, etc)
- **Years**: 2013 – 2019

- **What makes a candidate paper:**
  - contain XAI keywords (lemmaized) in their title (e.g., explainable, interpretable, transparent, etc)

- **Reviewing process**
  - A couple of NLP researchers first go through all the candidate papers to make sure they are truly about XAI for NLP

  - Every paper is then thoroughly reviewed by at least 2 NLP researchers
    - Category of the explanation
    - Explainability & visualization techniques
    - Evaluation methodology

  - Consult additional reviewers in the case of disagreement

Statistics of Literature Review

- **107 candidate papers – 52 of them are included**

Top 3 NLP Topics

1	2	3
Question Answering (9)	Computational Social Science & Social Media (6)	Syntax: Tagging, Chunking & Parsing (6)

Top 3 Conferences

1	2	3
EMNLP (21)	ACL (12)	NAACL (9)

Outline of Part II

- **Literature review methodology**

- **Categorization of different types of explanation**
  - Local vs. Global
  - Self-explaining vs. Post-hoc processing

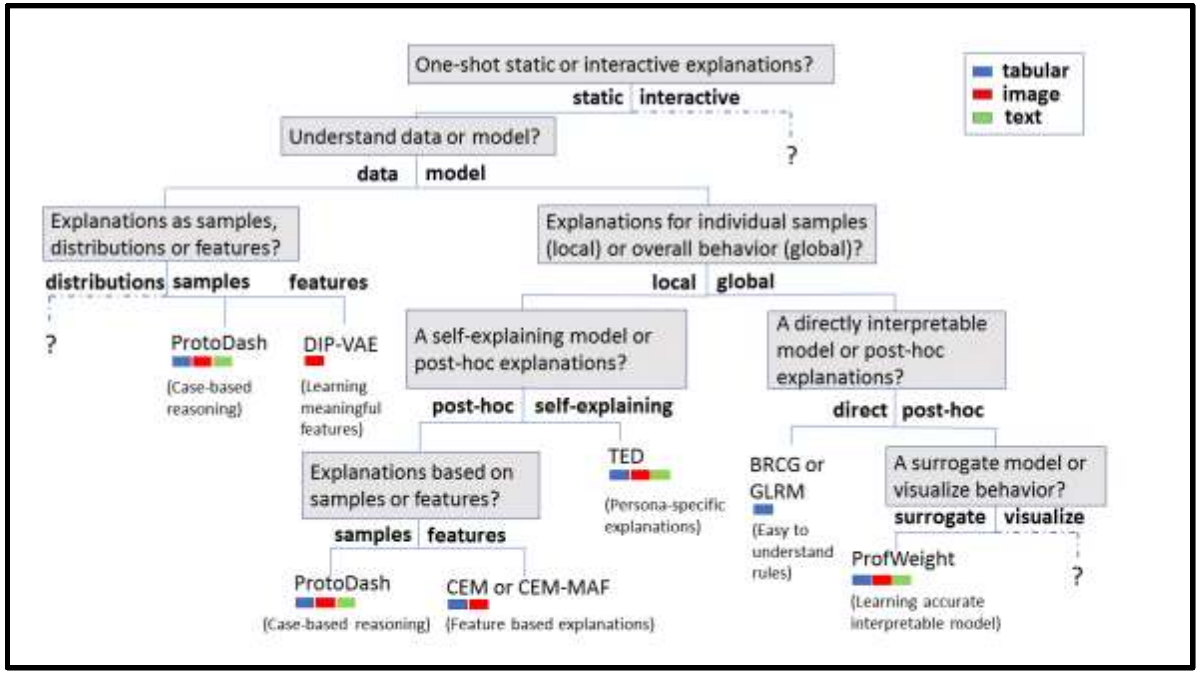
- **Generating and presenting explanations**
  - Explainability techniques
    - Common operations to enable explainability
  - Visualization techniques

- **Other insights (XNLP website)**
  - Relationships among explainability and visualization techniques
  - Evolution of the XAI research in the past a few years

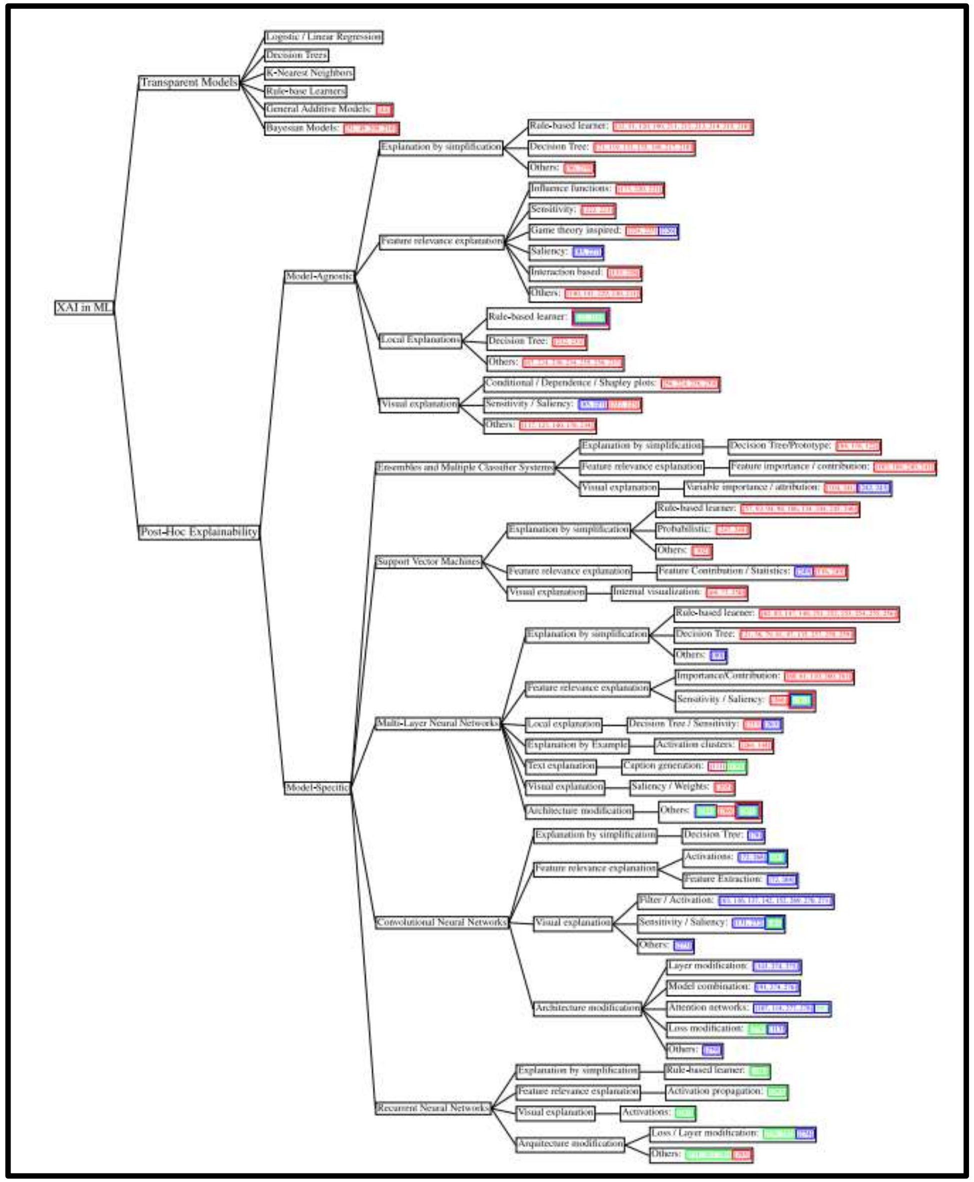
- **Evaluation of Explanations**

Task 1 - how to differentiate different explanations

- Different taxonomies exist



[Arya et al., 2019]



[Arrietaa et al, 2020]

Task 1 - how to differentiate different explanations

- Two fundamental aspects that apply to any XAI problems

Is the explanation for an individual instance or for an AI model ?

Is the explanation obtained directly from the prediction or requiring post-processing ?

Local explanation: individual instance

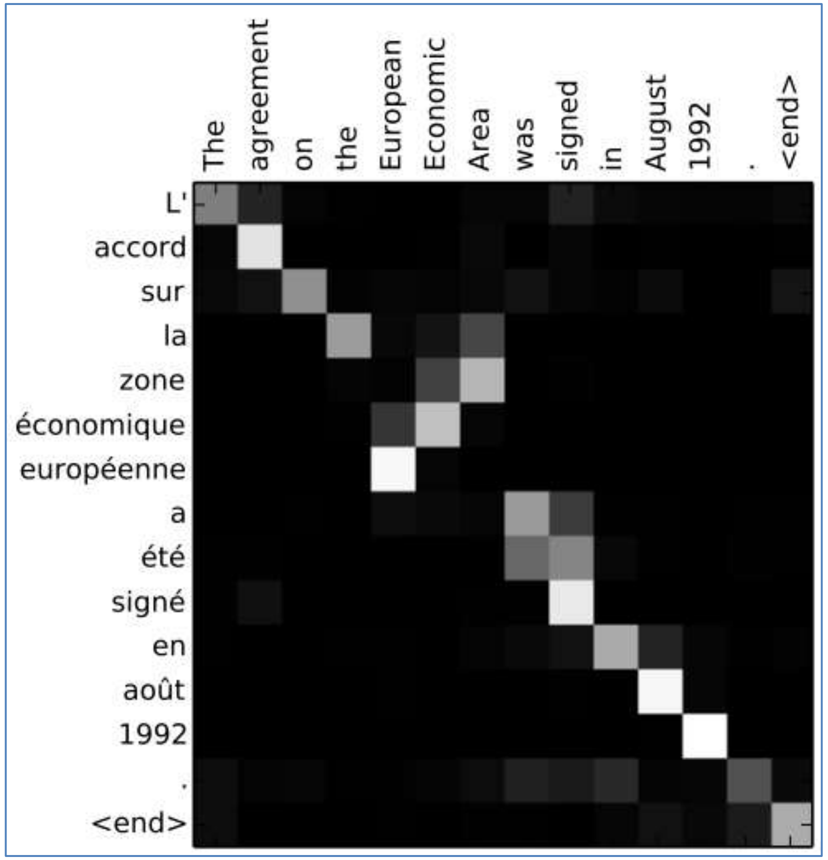
Global explanation: internal mechanism of a model

Self-explaining: directly interpretable

Post-hoc: a second step is needed to get explanation

Local Explanation [(Guidotti et al. 2018)]

- We understand only the reasons for a specific decision made by an AI model
- Only the single prediction/decision is explained



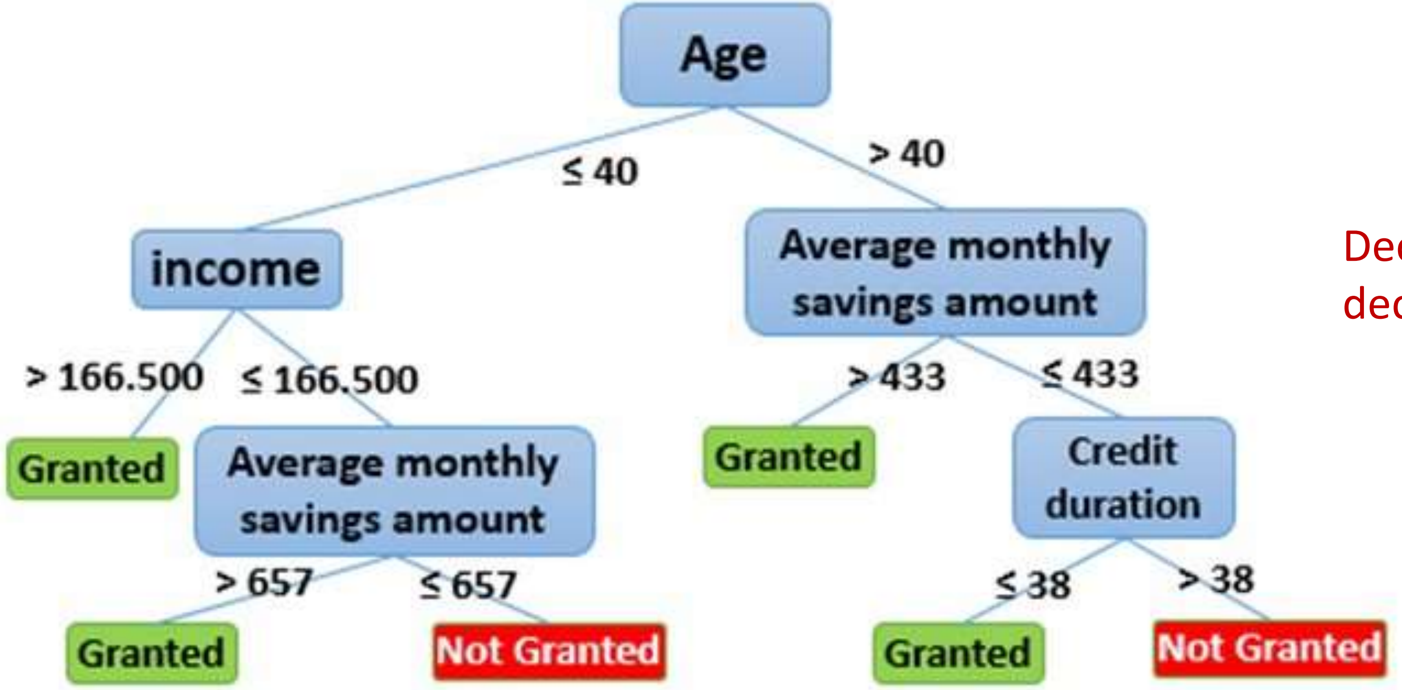
Machine translation

Input-output alignment matrix changes based on instances

[NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. Bahdanau et al., ICRL 2015]

Global explanation

- We understand the whole logic of a model
- We are able to follow the entire reasoning leading to all different possible outcomes

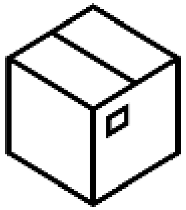


Decision tree built for credit card application decision making

Img source: https://medium.com/@fenjiro/data-mining-for-banking-loan-approval-use-case-e7c2bc3ece3

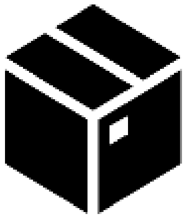
Self-explaining vs. Post-hoc

- **Self-explaining**: we can directly get the explanation with the prediction

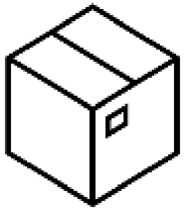


Only one model

- **Post-hoc:** explanation does not come directly with the prediction



Model 1



Model 2

Task 1 - how to differentiate different explanations

- Two fundamental aspects that apply to any XAI problems

Two orthogonal aspects

Is the explanation for an individual instance or for an AI model ?

Is the explanation obtained directly from the prediction or requiring post-processing ?

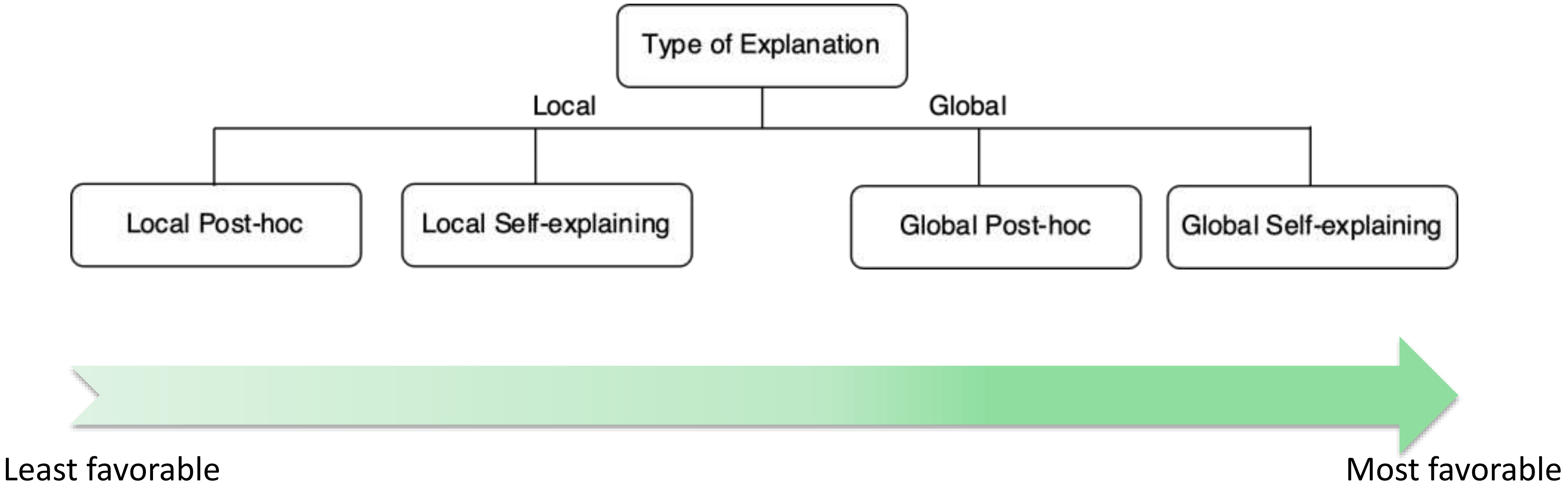
Local explanation: individual instance

Global explanation: internal mechanism of a model

Self-explaining: directly interpretable

Post-hoc: a second step is needed to get explanation

Categorization of Explanations



Outline of Part II

- **Literature review methodology**

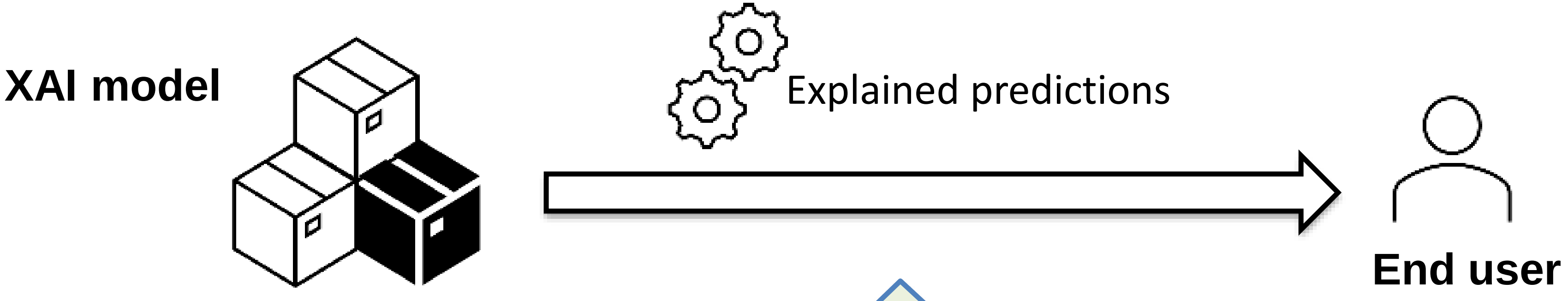
- **Categorization of different types of explanation**
  - Local vs. Global
  - Self-explaining vs. Post-hoc processing

- **Generating and presenting explanations**
  - Explainability techniques
    - Common operations to enable explainability
  - Visualization techniques

- **Other insights (XNLP website)**
  - Relationships among explainability and visualization techniques
  - Evolution of the XAI research in the past a few years

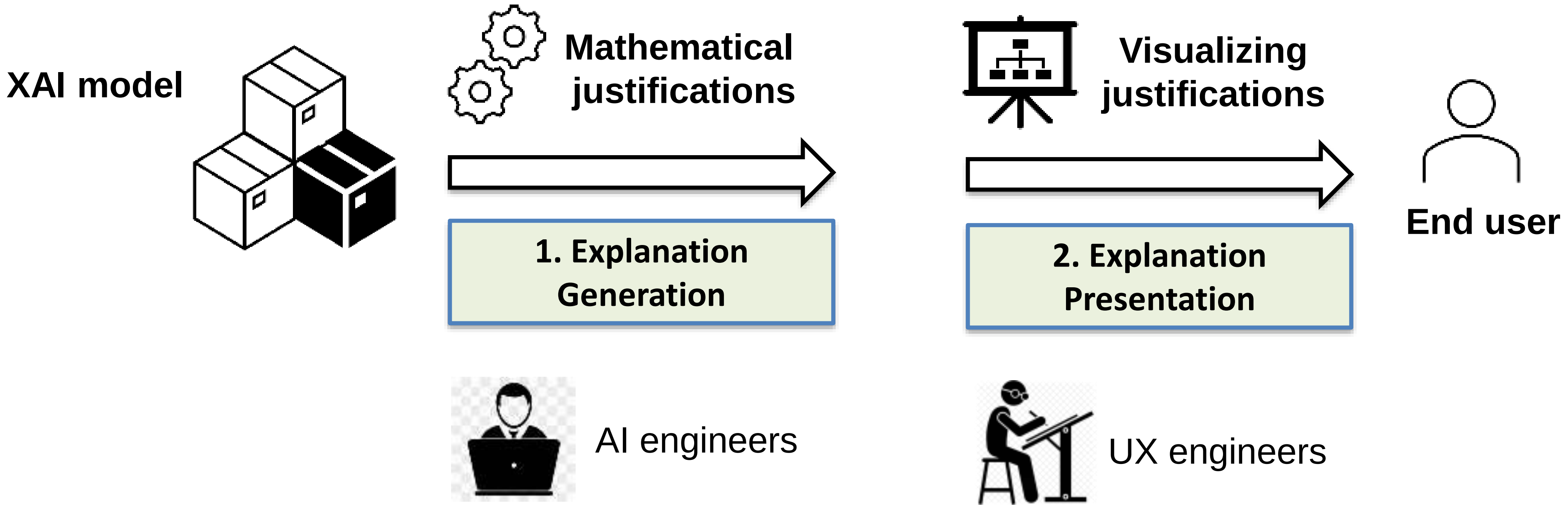
- **Evaluation of Explanations**

From model prediction to user understanding

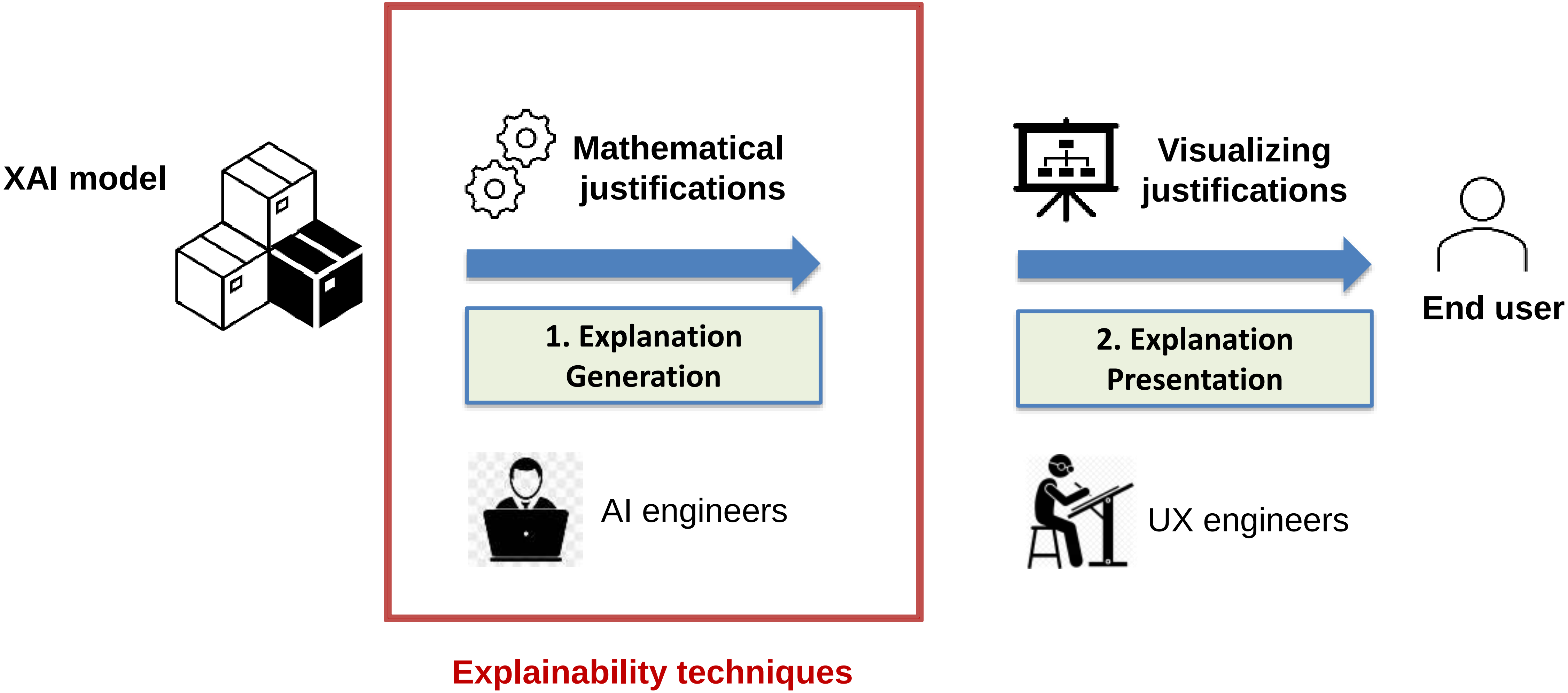


Are explanations directly sent from XAI models to users?

Two Aspects: Generation & Presentation



What’s next


XAI model
Mathematical justifications
1. Explanation Generation
AI engineers
Visualizing justifications
2. Explanation Presentation
UX engineers
End user
Explainability techniques

What is Explainability

- The techniques used to generate raw explanations
  - Ideally, created by AI engineers

- Representative techniques:
  - Feature importance
  - Surrogate model
  - Example-driven
  - Provenance-based
  - Declarative induction

- Common operations to enable explainability
  - First-derivative saliency
  - Layer-wise Relevance Propagation
  - Attention
  - LSTM gating signal
  - Explainability-aware architecture design

Explainability - Feature Importance

Definition The main idea of **Feature Importance** is to derive explanation by investigating the importance scores of different features used to output the final prediction.

Can be built on different types of features

Manual features from feature engineering

[Voskarideset al., 2015]

Lexical features including words/tokens and N-gram

[Godin et al., 2018]
[Mullenbachet al., 2018]

Latent features learned by neural nets

[Xie et al., 2017]
[Bahdanau et al., 2015]
[Li et al., 2015]
[Ghaeini et al., 2018]

Usually used in parallel

Feature Importance – Machine Translation [Bahdanau et al., 2015]

A well-known example of feature importance is the Attention Mechanism first used in machine translation

Traditional **encoder-decoder** architecture for machine translation

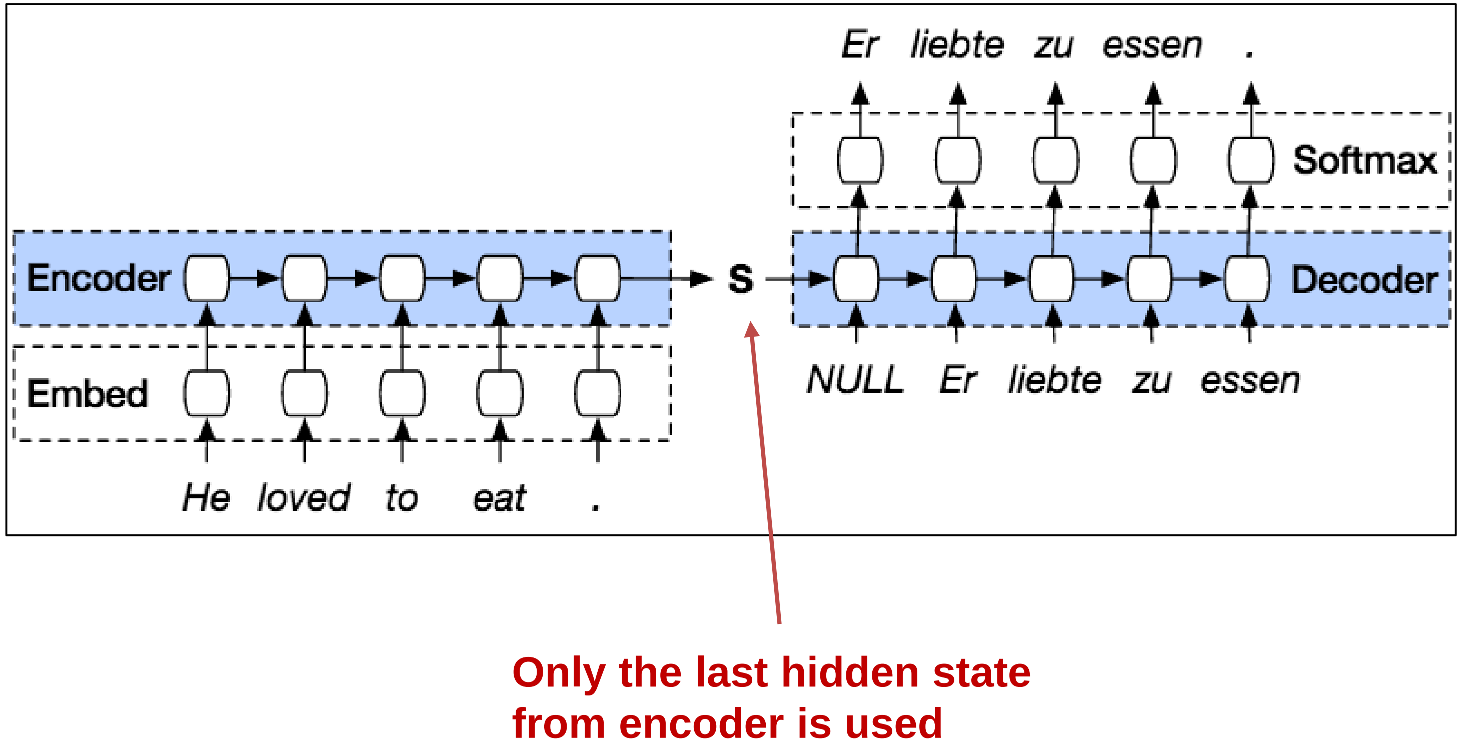


Image courtesy: https://smerity.com/articles/2016/google_nmt_arch.html

[NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. Bahdanau et al., ICLR 2015]

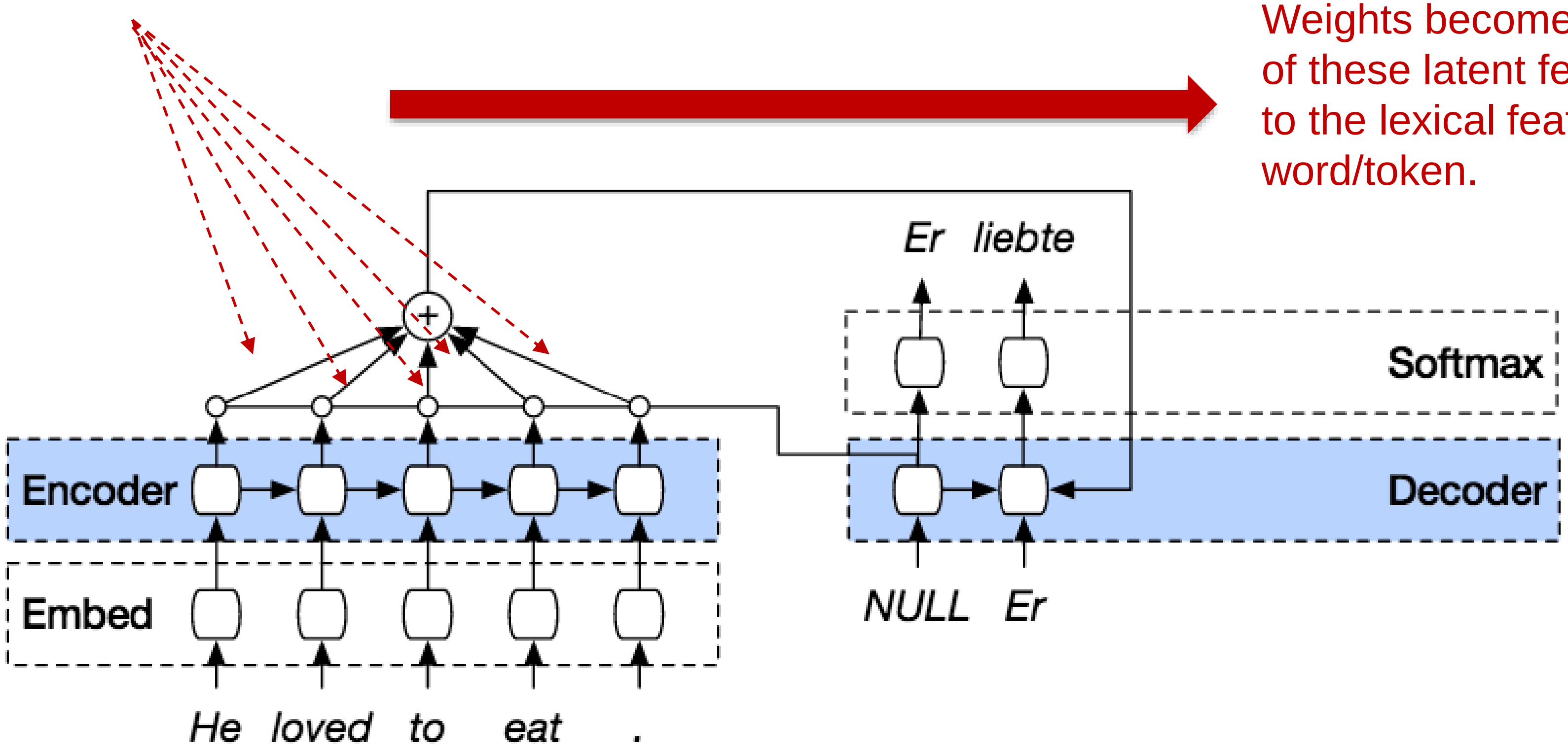
Feature Importance – Machine Translation as An Example

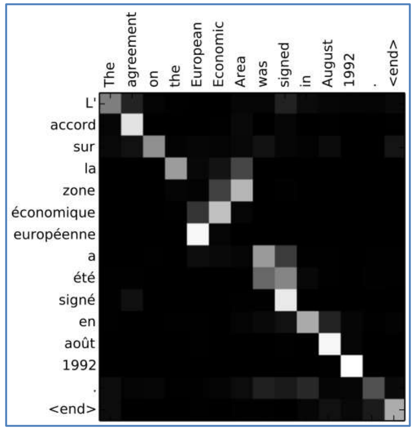
Local & Self-explaining

Attention mechanism uses the latent features learned by the encoder

- a weighted combination of the latent features

Weights become the **importance scores** of these latent features (also corresponds to the lexical features) for the decoded word/token.





Input-output alignment

Image courtesy: https://smerity.com/articles/2016/google_nmt_arch.html

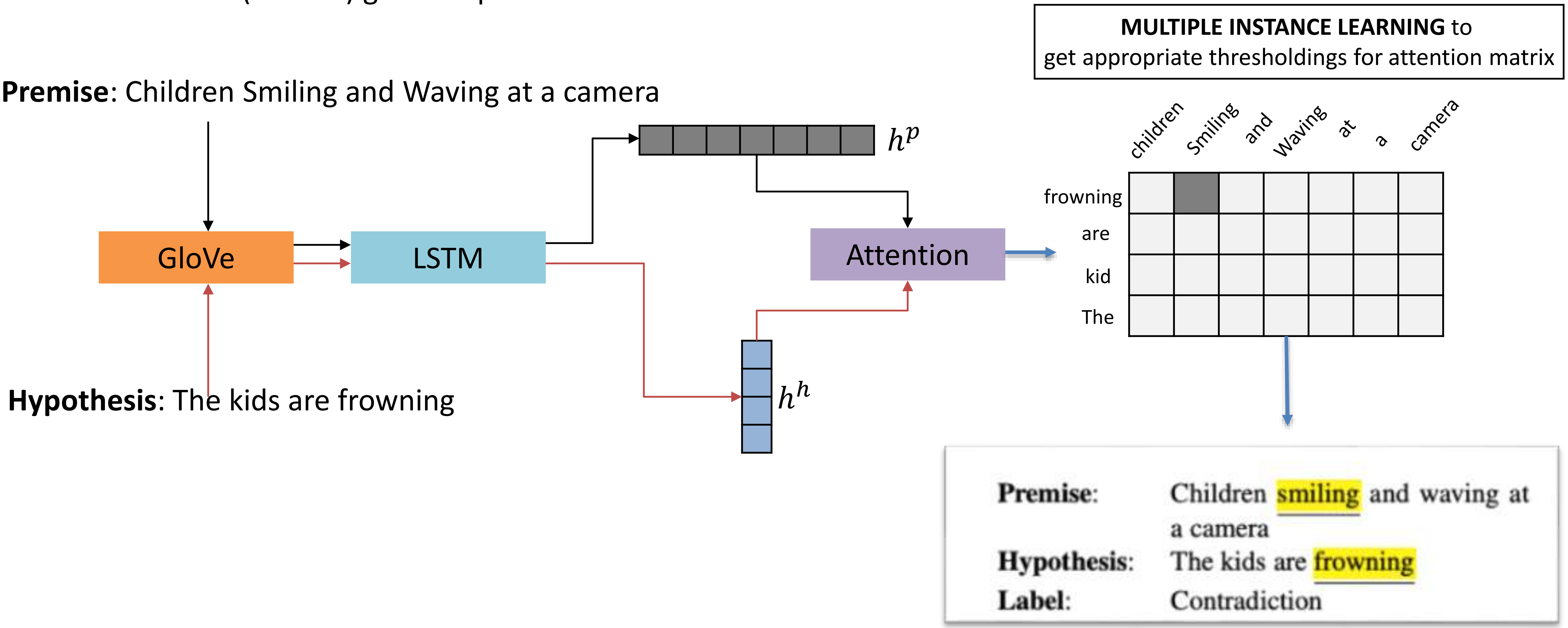
Similar approach in [Mullenbachet al., 2018]

[NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. Bahdanau et al., ICRL 2015]

Feature Importance – token-level explanations [Thorne et al]

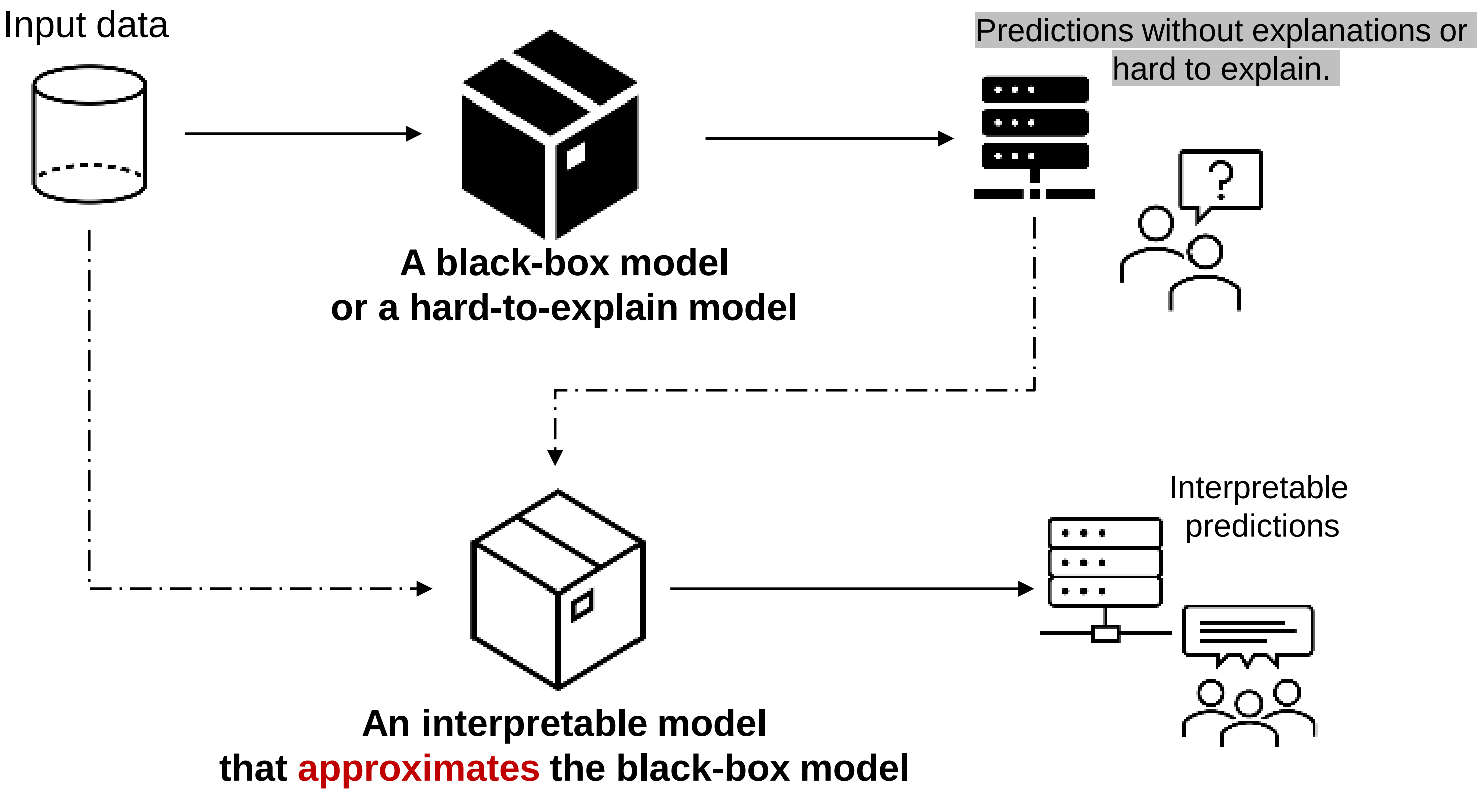
Another example that uses attention to enable feature importance

NLP Task: Natural Language Inference (determining whether a "hypothesis" is true (entailment), false (contradiction), or underdetermined (neutral) given a "premise"



Explainability – Surrogate Model

Definition	Model predictions are explained by learning a second, usually more explainable model, as a proxy.



[Ribeiro et al. KDD]

[Liu et al. 2018]

[Alvarez-melis and Jaakkola]

[Sydorova et al, 2019]

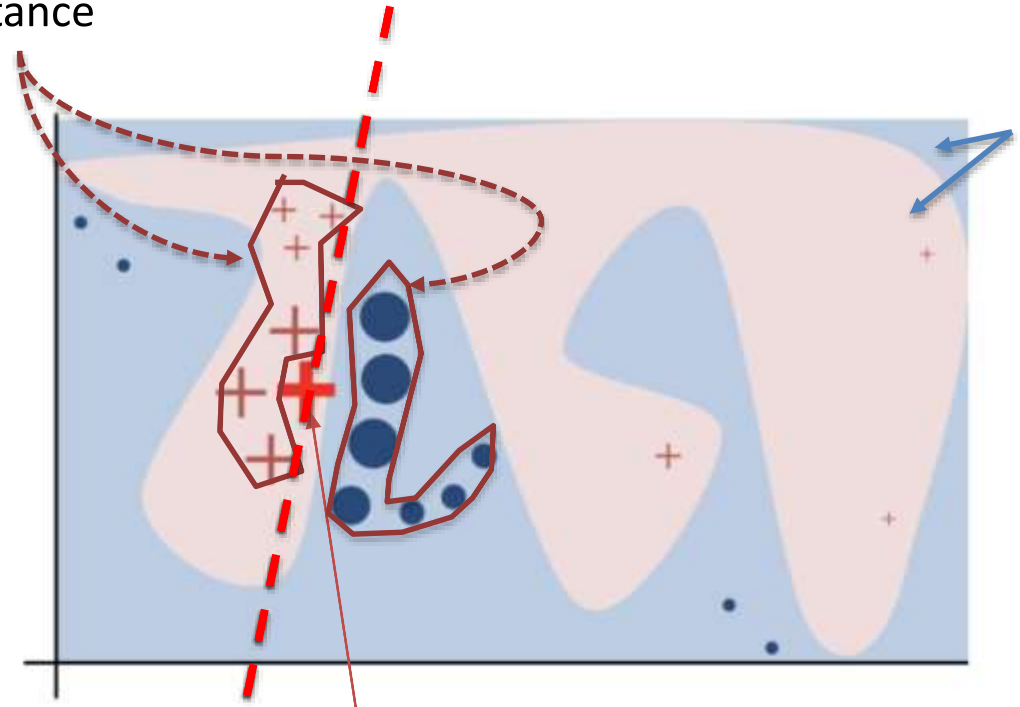
Surrogate Model – LIME [Ribeiro et al. KDD]

LIME: Local Interpretable Model-agnostic Explanations

Local & Post-hoc

Sampled instances, weighted by the proximity to the target instance

An **Interpretable classifier** with **local fidelity**



Blue/Pink area are the decision boundary of a complex model f, which is a black-box model

Cannot be well approximated by a linear function

Instance to be explained

Surrogate Model – LIME [Ribeiro et al. KDD]

Local & Post-hoc

High flexibility – free to choose any interpretable surrogate models

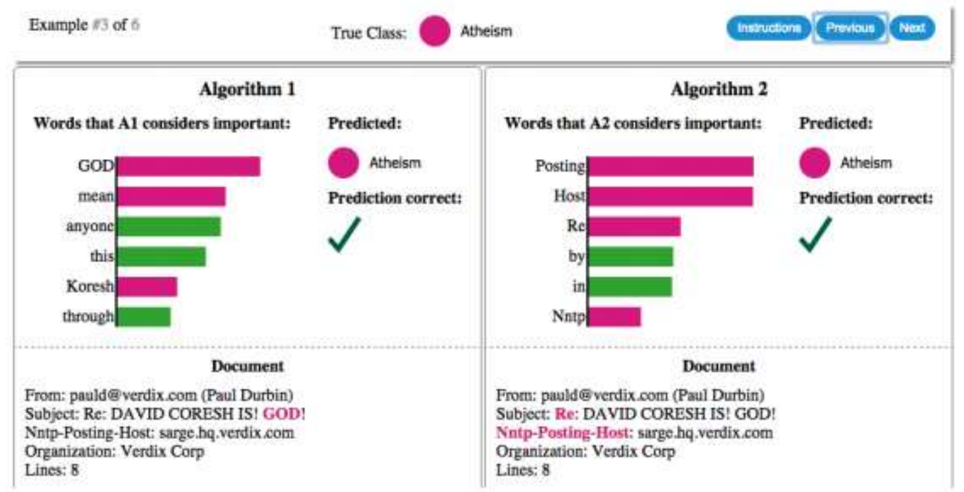


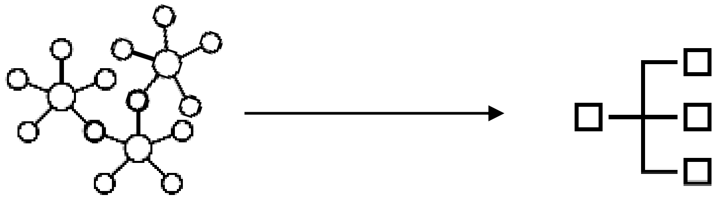
Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

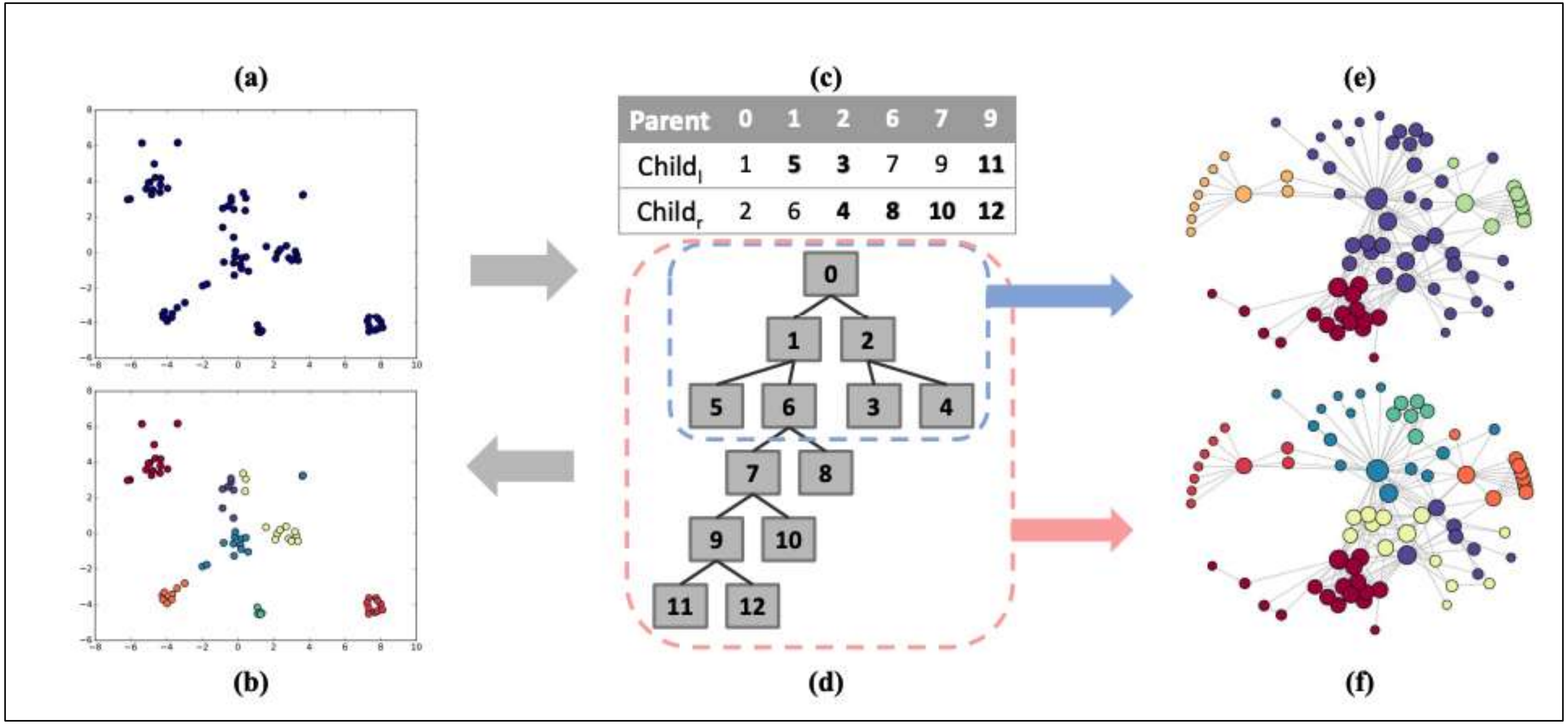
Surrogate Model - Explaining Network Embedding [Liu et al. 2018]

Global & Post-hoc



A learned
network embedding

Induce a taxonomy
from the network embedding



Surrogate Model - Explaining Network Embedding [Liu et al. 2018]

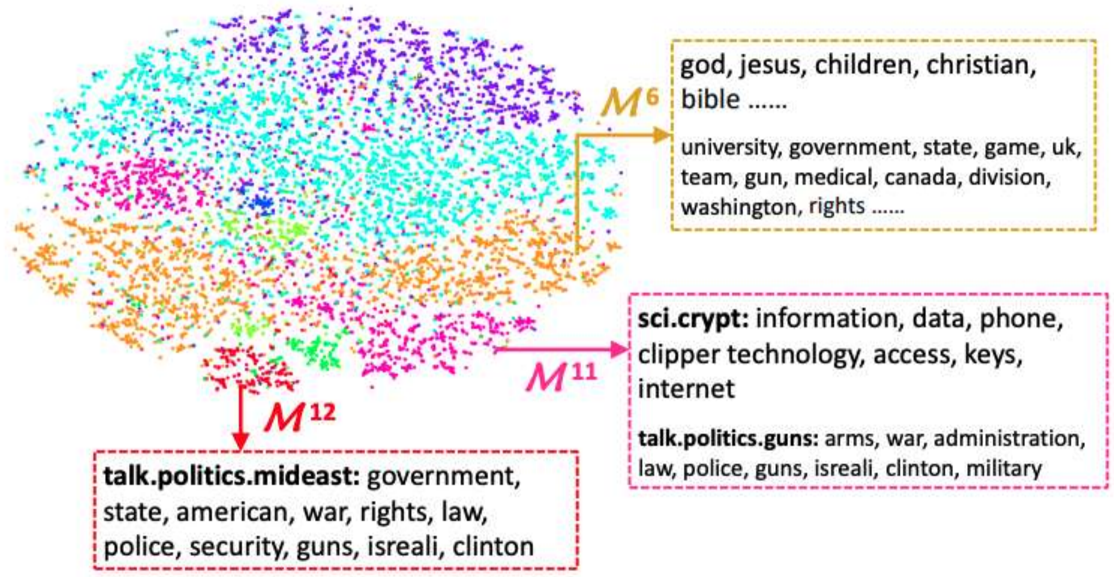


Figure 11: The inducted taxonomy with 7 leaf clusters.

Bold text are attributes of the cluster, followed by keywords from documents belong to the cluster.

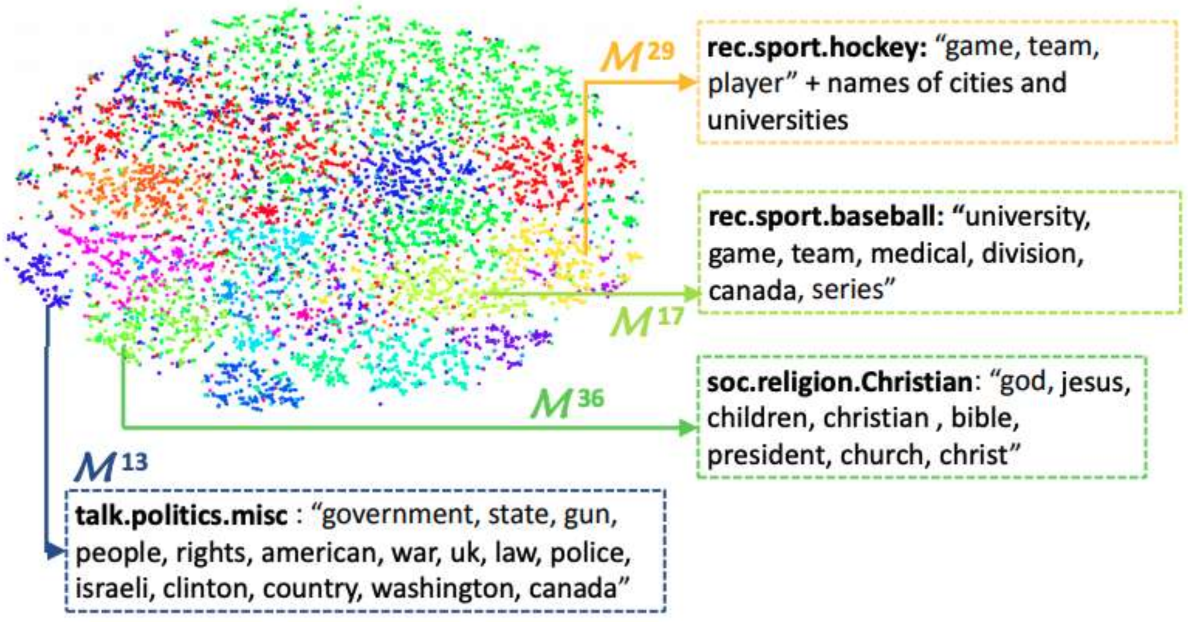


Figure 12: The inducted taxonomy with 20 leaf clusters.

Induced from 20NewsGroups Network

Explainability – Example-driven

Definition Such approaches explain the prediction of an input instance by identifying and presenting other instances, usually from available labeled data, that are semantically similar to the input instance.

- Similar to the idea of nearest neighbors [Dudani, 1976]
- Have been applied to solve problems including
  - Text classification [Croce et al., 2019]
  - Question Answering [Abujabal et al., 2017]

Explainability – Text Classification [Croce et al., 2019]

"What is the capital of Germany?" refers to a **Location**. **WHY?**

Because **"What is the capital of California?"** which also refers to a **Location** in the training data

Explainability – Text Classification [Croce et al., 2019]

Layer-wise Relevance Propagation in Kernel-based Deep Architecture

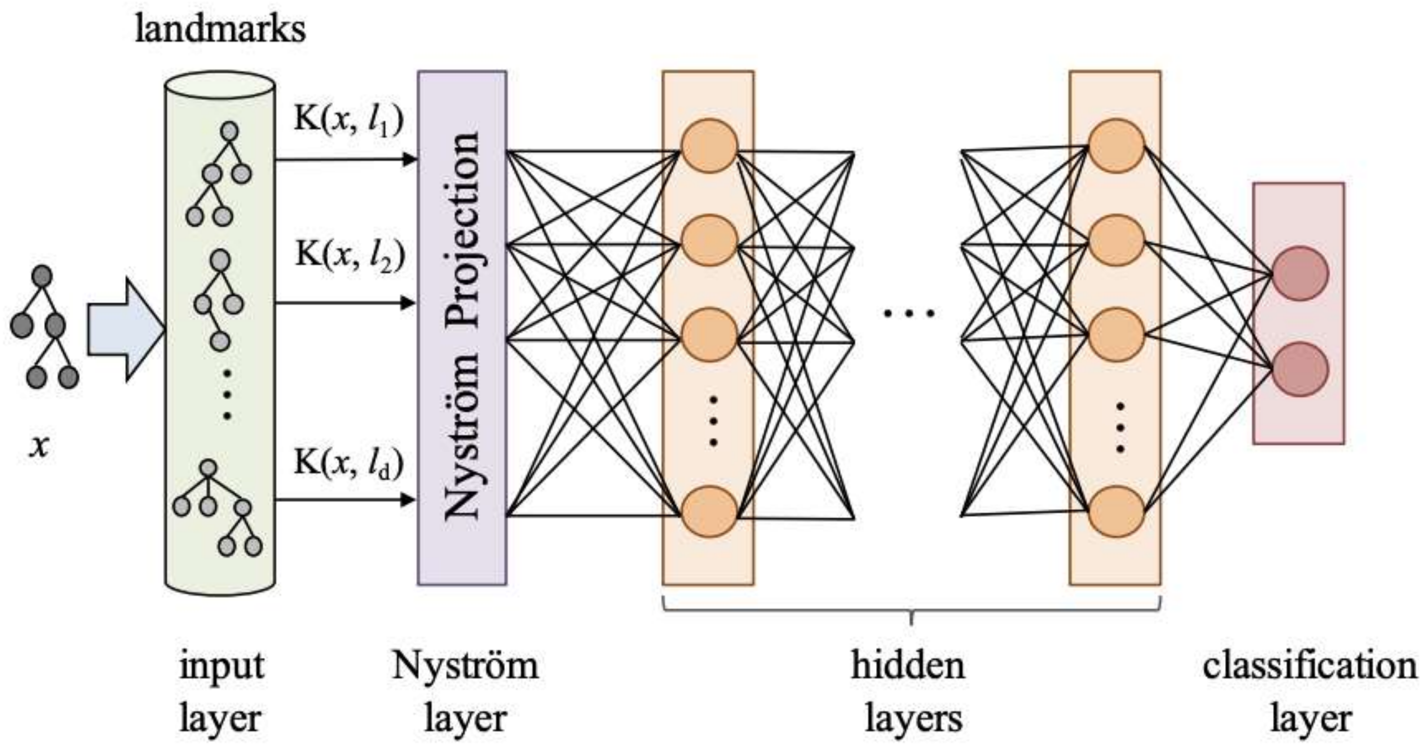


Figure 1: Kernel-based Deep Architecture.

Explainability – Text Classification [Croce et al., 2019]

Class	Questions (q_i)	$k(q_1, q_2)$	Activated Landmarks (l_i)	$k(l_1, l_2)$
LOC NUM	*"What is the capital of Ethiopia?"* *"What is the population of Nigeria?"*	0.98		
ENTY HUM	*"What was FDR 's dog 's name?"* *"What was J.F.K.'s wife 's name?"*	0.97	*"What is the name of David Letterman's dog?"* *"What was Darth Vader 's son named?"*	0.49
ENTY ENTY	*"What is the Ohio state bird?"* *"What is the pH scale?"*	0.90	*"What is the name of David Letterman 's dog?"* *"What is viscosity?"*	0.61
ENTY HUM	*"What was the first satellite to go into space?"* *"Who was the first American to walk in space?*	0.83	*"What was the first TV set to include a remote control?"* *"What 's the name of the actress who starred in the movie, Silence of the Lambs?"*	0.61
NUM NUM	*"What was the last year that the Chicago Cubs won the World Series?"* *"What is the average speed of the horses at the Kentucky Derby?"*	0.73	*"The film Jaws was made in what year?"* *"What is average salary of restaurant manager in United States?"*	0.31

Table 1: Examples of semantically similar questions in the same or different classes, with the corresponding landmarks activated during the classification.

Explainability – Provenance-based

Definition Explanations are provided by illustrating (some of) the prediction derivation process

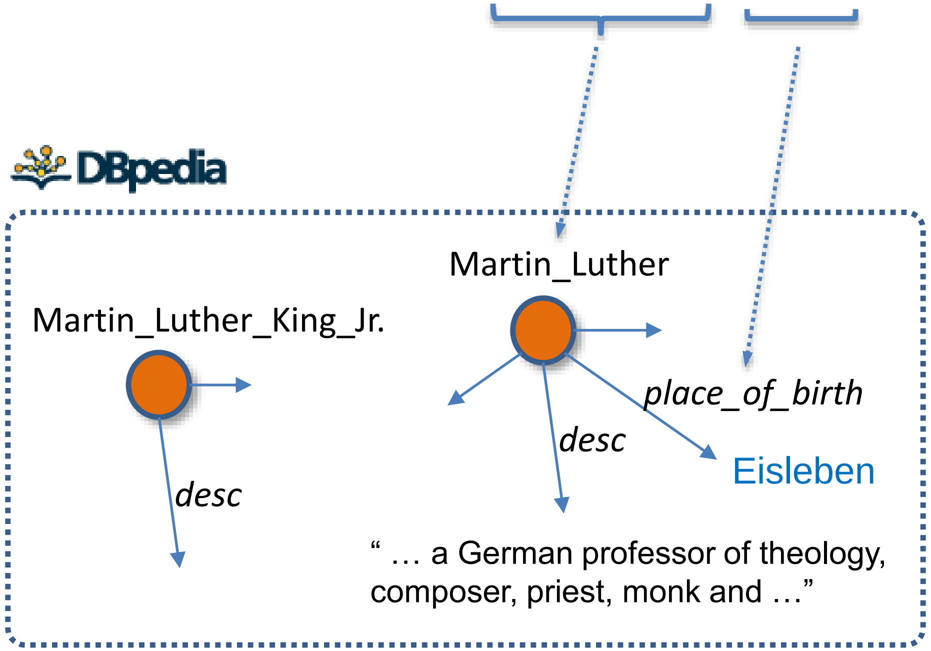
- An intuitive and effective explainability technique when the final prediction is the result of a series of reasoning steps.

- Several ***question answering*** papers adopt this approach.
  - [Abujabal et al., 2017] – Quint system for KBQA (knowledge-base question answering)
  - [Zhou et al., 2018] – Multi-relation KBQA
  - [Amini et al., 2019] – MathQA, will be discussed in the declarative induction part (as a representative example of **program generation**)

Provenance-Based: Quint System

A. Abujabal et al. *QUINT: Interpretable Question Answering over Knowledge Bases*. EMNLP System Demonstration, 2017.

"Where was Martin Luther raised?"

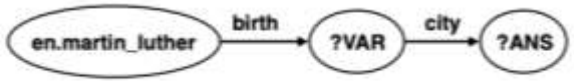


- When QUINT answers a question, it shows the **complete derivation sequence**:

  how it understood the question:

  - Entity linking: entity mentions in text ⟼ actual KB entities
  - Relation linking: relations in text ⟼ KB predicates

  the SPARQL query used to retrieve the answer:



"Eisleben"

Learned from **query templates** based on structurally similar instances ("Where was Obama educated?")

Derivation provides insights towards reformulating the question (in case of errors)

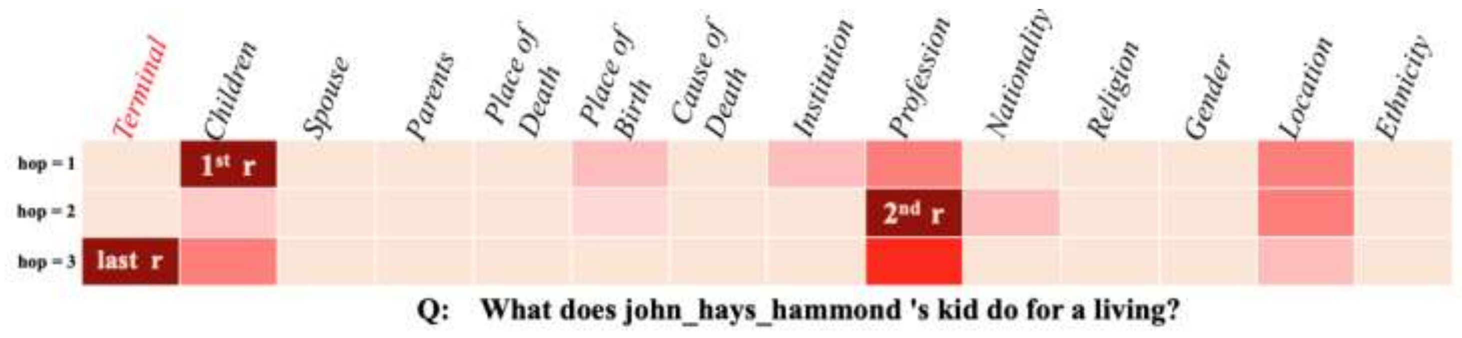
Provenance-Based: Multi-Relation Question Answering

IRN (Interpretable Reasoning Network):

- Breaks down the KBQA problem into multi-hop reasoning steps.
- Intermediate entities and relations are predicted in each step:

$$e_s \xrightarrow{r_1} e_1 \xrightarrow{r_2} \dots \xrightarrow{r_n} a.$$

M. Zhou et al. *An Interpretable Reasoning Network for Multi-Relation Question Answering*. COLING, 2018.



Facilitates **correction by human (e.g., disambiguation)**

- multiple answers in this example (3 different children)

Reasoning path: John_Hays_Hammond $\xrightarrow{\text{Child}}$ e $\xrightarrow{\text{Profession}}$ a

- dbr:John_Hays_Hammond,_Jr. → dbr:Inventor
- dbr:Natalie_Hays_Hammond → dbr:Author, dbr:Costume_designer, dbr:Miniaturist
- dbr:Richard_Pindle_Hammond

Explainability – Declarative Induction

Definition The main idea of **Declarative Induction** is to construct human-readable representations such as **trees**, **programs**, and **rules**, which can provide local or global explanations.

Explanations are produced in a *declarative* specification language, as humans often would do.

Trees

[Jiang et al., ACL 2019] - express alternative ways of reasoning in multi-hop reading comprehension

(Specialized) Programs:

[A. Amini et al., NAACL, 2019] – MathQA, math programs for explaining math word problems

Examples of local explainability (per instance)

Rules:

[A. Ebaid et al., ICDE Demo, 2019] - Bayesian rule lists to explain entity matching decisions

[Pezeshkpour et al, NAACL 2019] - rules for interpretability of link prediction

[Sen et al, EMNLP 2020] - linguistic expressions as sentence classification model

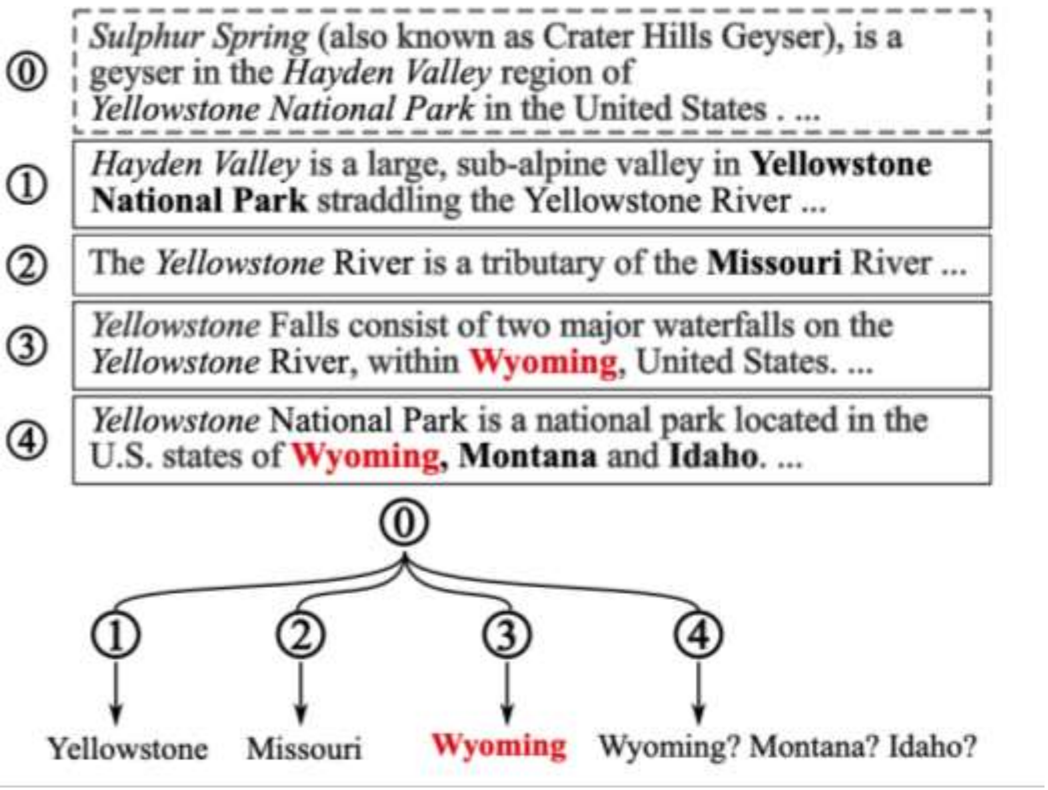
[Qian et al, CIKM 2017] - rules as entity matching model

Examples of global explainability (although rules can also be used for local explanations)

Declarative Induction: Reasoning Trees for Multi-Hop Reading Comprehension

Task: Explore and connect relevant information from multiple sentences/documents to answer a question.

Question subject: “Sulphur Spring”
Question body: located in administrative territorial entity



Reasoning trees play dual role:

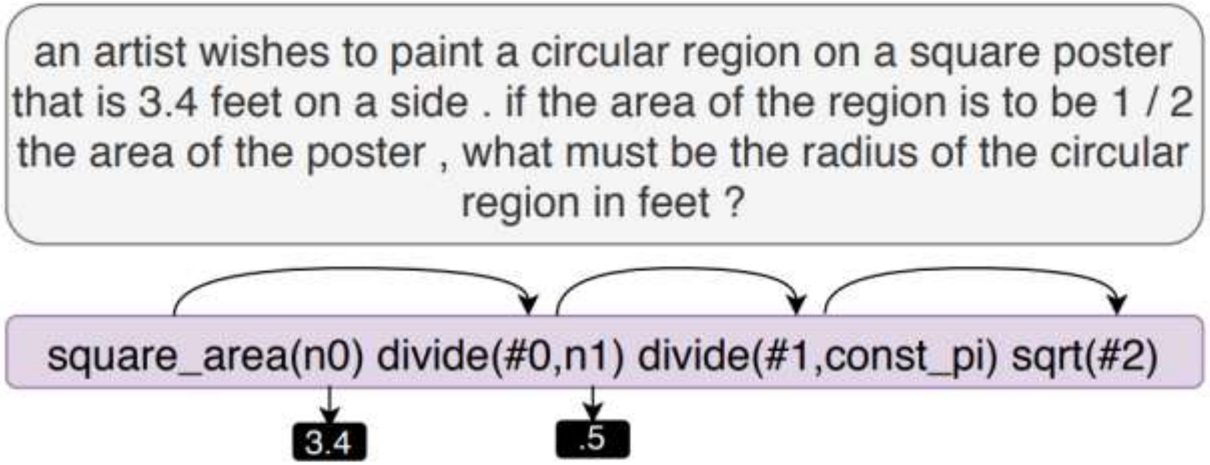
- Accumulate the information needed to **produce the answer**
  - Each root-to-leaf path represents a possible answer and its derivation
  - Final answer is aggregated across paths
- **Explain the answer** via the tree

Y. Jiang et al. *Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension*. ACL, 2019.

Declarative Induction: (Specialized) Program Generation

Task: Understanding and solving math word problems

A. Amini et al. *MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms*. NAACL-HLT, 2019.



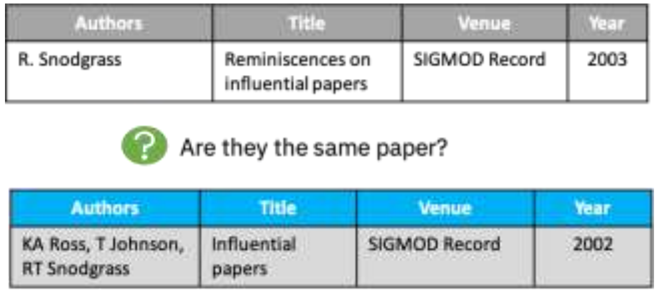
Domain-specific program based on math operations

- **Seq-2-Program** translation produces a human-interpretable representation, for each math problem
- Does not yet handle problems that need complicated or long chains of mathematical reasoning, or more powerful languages (logic, factorization, etc).
- Also fits under **provenance-based** explainability:
  - provides answer, and how it was derived.

- Representative example of a system that provides:
  - Local (“input/output”) explainability
  - But no global explanations or insights into how the model operates (still “black-box”)

Declarative Induction: Post-Hoc Explanation for Entity Resolution

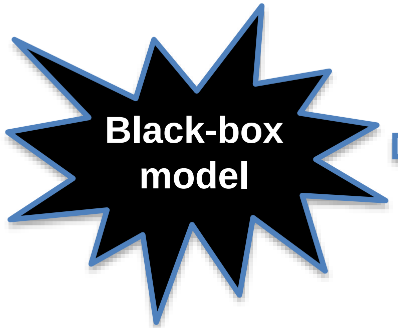
Entity resolution (aka ER, entity disambiguation, record linking or matching)



Authors	Title	Venue	Year
R. Snodgrass	Reminiscences on influential papers	SIGMOD Record	2003

Are they the same paper?

Authors	Title	Venue	Year
KA Ross, T Johnson, RT Snodgrass	Influential papers	SIGMOD Record	2002

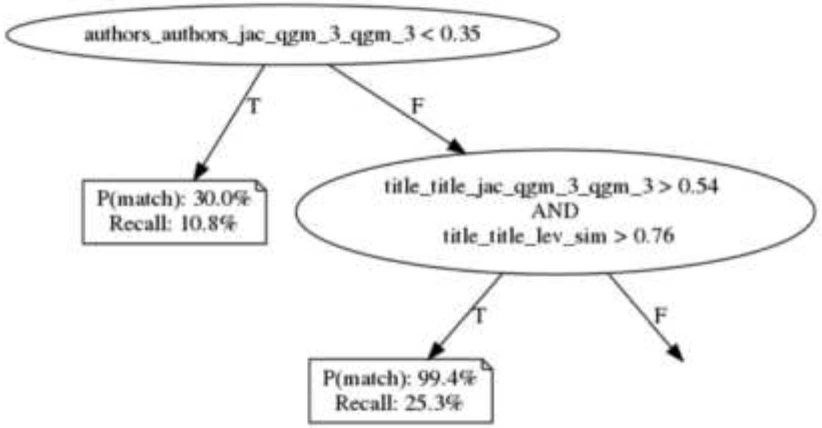


ExplainER

- Local explanations
  (e.g., how different features contribute to a single match prediction – based on LIME)
- Global explanations
  Approximate the model with an interpretable representation (Bayesian Rule List -- BRL)
- Representative examples
  Small, diverse set of <input, output> tuples to illustrate where the model is correct or wrong
- Differential analysis
  Compare with other ER model (disagreements, compare performance)

M. T. Ribeiro et al. *Why should I trust you: Explaining predictions of any classifier*. In SIGKDD, pp. 1135—1144, 2016.

B. Letham et al. *Interpretable classifiers using rules and Bayesian analysis.* Annals of Applied Stat., 9(3), pp. 1350-1371, 2015.



Can also be seen as a **tree**
(of probabilistic IF-THEN rules)

A. Ebaid et al. *ExplainER: Entity Resolution Explanations*. ICDE Demo, 2019.

Declarative Induction: Rules for Link Prediction

Task: Link prediction in a knowledge graph, with focus on robustness and explainability

Rule Body, $R_1(a,c) \wedge R_2(c,b) \Rightarrow$	**Target**, $R(a,b)$
Common to both	
isConnectedTo$(a,c) \wedge$ isConnectedTo(c,b)	isConnectedTo
isLocatedIn$(a,c) \wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
isMarriedTo$(a,c) \wedge$ hasChild(c,b)	hasChild
only in DistMult	
playsFor$(a,c) \wedge$ isLocatedIn(c,b)	wasBornIn
dealsWith$(a,c) \wedge$ participatedIn(c,b)	participatedIn
isAffiliatedTo$(a,c) \wedge$ isLocatedIn(c,b)	diedIn
isLocatedIn$(a,c) \wedge$ hasCapital(c,b)	isLocatedIn
only in ConvE	
influences$(a,c) \wedge$ influences(c,b)	influences
isLocatedIn$(a,c) \wedge$ hasNeighbor(c,b)	isLocatedIn
hasCapital$(a,c) \wedge$ isLocatedIn(c,b)	exports
hasAdvisor$(a,c) \wedge$ graduatedFrom(c,b)	graduatedFrom

Uses adversarial modifications (e.g., removing facts):

1. Identify the KG nodes/facts most likely to influence **a link**
2. Aggregate the individual explanations into an **extracted set of rules** operating at the level of entire KG:
   - Each rule represents a frequent pattern for predicting the target link

Works on top of existing, black-box models for link prediction:

DistMult: B. Yang et al. *Embedding entities and relations for learning and inference in knowledge bases.* ICLR 2015
ConvE: T. Dettmers et al. *Convolutional 2d knowledge graph embeddings*. AAAI 2018

Global explainability able to pinpoint common mistakes in the underlying model (**wrong inference** in ConvE)

P. Pezeshkpour et al. *Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications*. NAACL 2019.

- Previous approach for rule induction:
  - The result is a set of logical rules (Horn rules of length 2) that are built on existing predicates from the KG (e.g., Yago).

- Other approaches also learn logical rules, but they are based on different vocabularies:
  - Next:
    - Logical rules built on linguistic predicates for Sentence Classification
  - [SystemER – see later under Explainability-Aware Design]
    - Logical rules built on similarity features for Entity Resolution

Declarative Induction: Linguistic Rules for Sentence Classification

Task: Sentence classification in a domain (e.g., legal/contracts domain)

Notices **will be transmitted** electronically, by registered or certified mail, or courier

Label

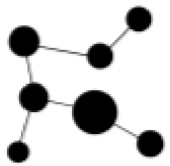
"Communication"

"Terms and Termination"

...

...

Syntactic and Semantic NLP Operators



(Explainable) Deep Learning

Rules (predicates based on syntactic/semantic parsing)

$\exists$ `verb` $\in S$:
`verb.theme` $\in$ `ThemeDict`
and `verb.voice = passive`
and `verb.tense = future`

Dictionaries

ThemeDict
communication notice

- Representative example of system that uses a ***neural model*** to learn another model (***the rules***)
  - The rules are used for classification but also act as global explanation
- Joint learning of both rules and domain-specific dictionaries
- Explainable rules generalize better and are more conducive for *interaction with the expert* (see later – **Visualization**)

Sen et al, *Learning Explainable Linguistic Expressions with Neural Inductive Logic Programming for Sentence Classification.* EMNLP 2020.

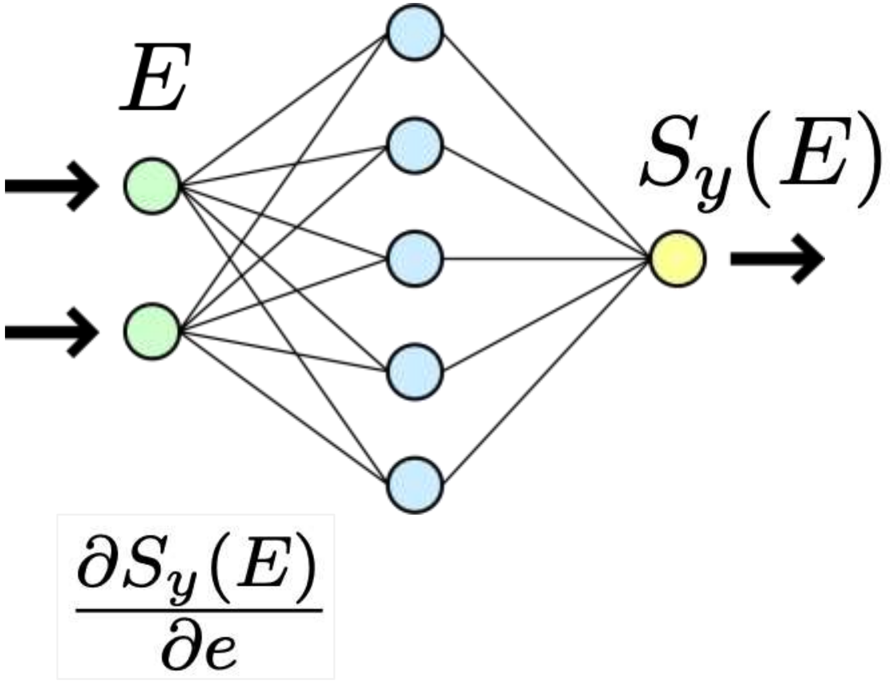
Common Operations to enable explainability

- **A few commonly used operations that enable different explainability**
  - First-derivative saliency
  - Layer-wise relevance propagation
  - Input perturbations
  - Attention
  - LSTM gating signal
  - Explainability-aware architecture design

Operations – First-derivative saliency

- Mainly used for enabling **Feature Importance**
  - [Li et al., 2016 NAACL], [Aubakirova and Bansal, 2016]

- Inspired by the NN visualization method proposed for computer vision
  - Originally proposed in (Simonyan et al. 2014)
  - How much each input unit contributes to the final decision of the classifier.



E Input embedding layer

$S_y(E)$ Output of neural network

$\frac{\partial S_y(E)}{\partial e}$ Gradients of the output wrt the input

Operations – First-derivative saliency

Naturally, it can be used to generate explanations presented by saliency-based visualization techniques

Sentiment Analysis

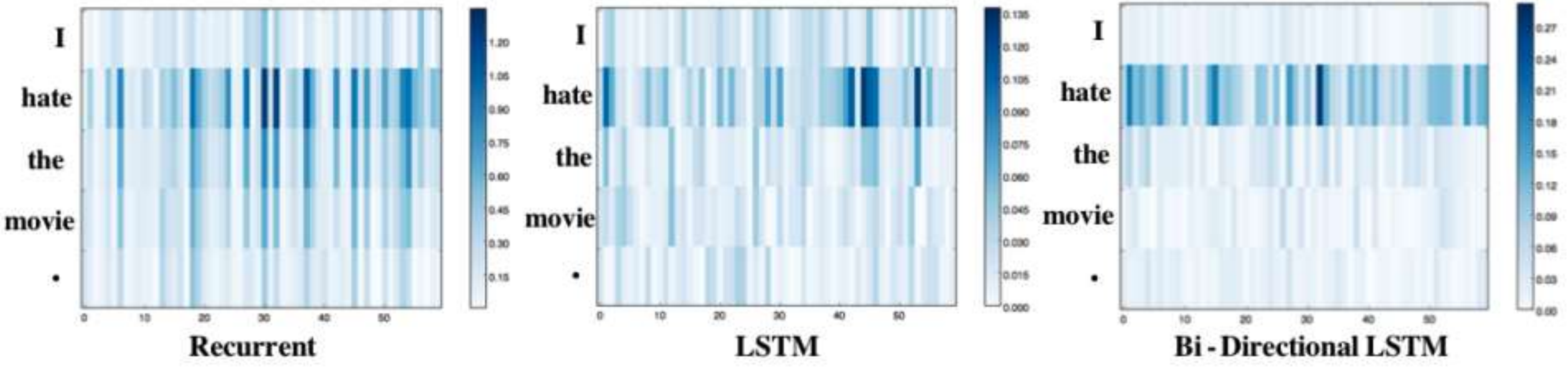


Figure 5: Saliency heatmap for for "I hate the movie ." Each row corresponds to saliency scores for the correspondent word representation with each grid representing each dimension.

[Li et al., 2016 NAACL]

Operations – Layer-wise relevance propagation

- For a good overview, refer to [Montavon et al., 2019]

- An operation that is specific to neural networks
  - Input can be images, videos, or **text**

- Main idea is very similar to first-derivative saliency
  - Decomposing the prediction of a deep neural network computed over a sample, down to relevance scores for the single input dimensions of the sample.

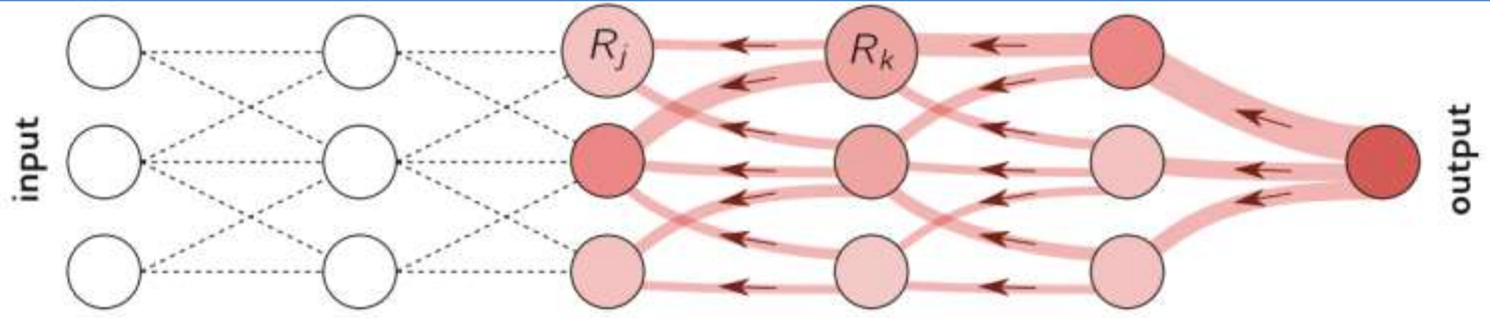


Fig. 10.2. Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer.

[Montavon et al., 2019]

Operations – Layer-wise relevance propagation

- Key difference to first-derivative saliency:
  - Assigns to each dimension (or feature), x_d, a relevance score $R_d^{(1)}$ such that [Croce et al]

$$f(x) \approx \sum_d R_d^{(1)}$$

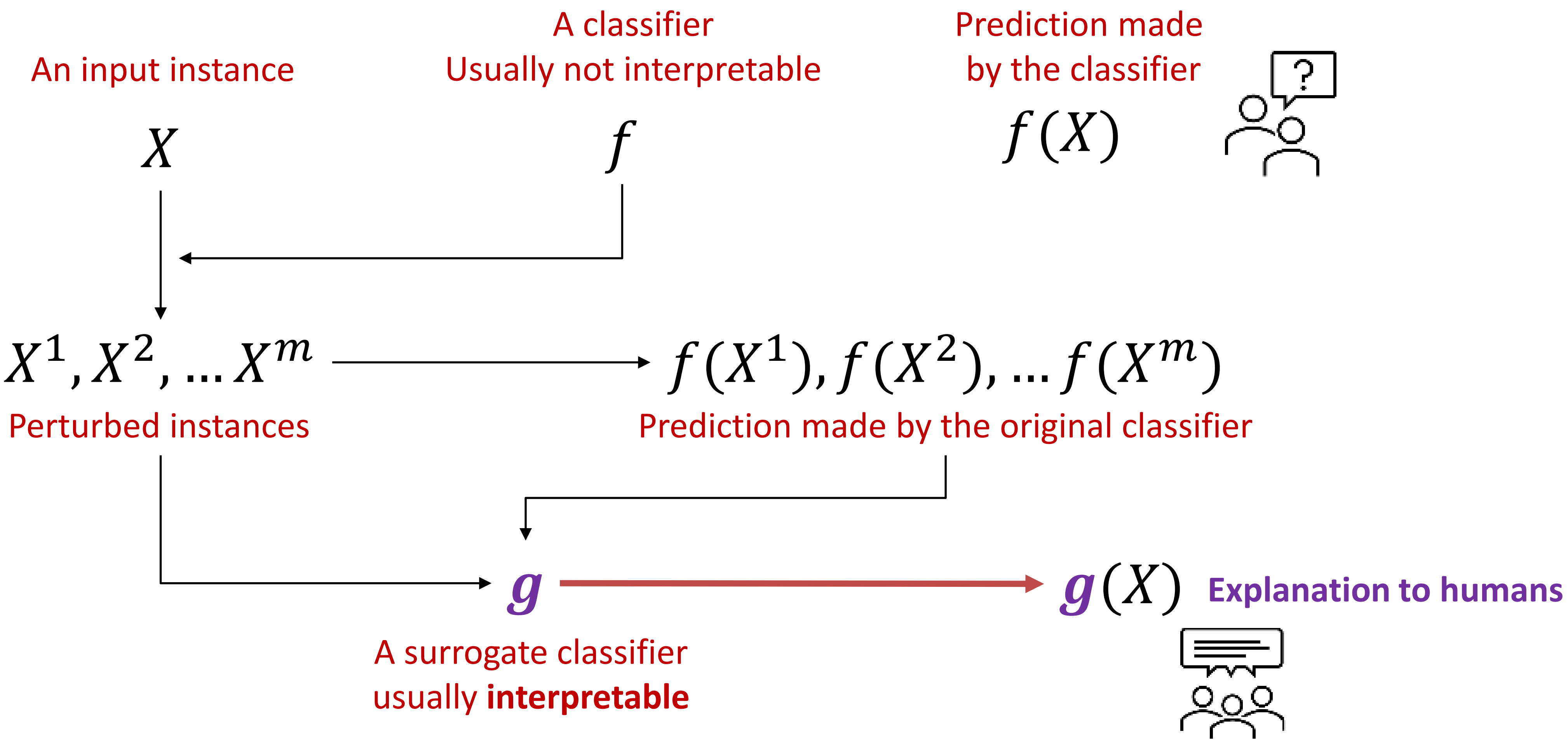
$R_d^{(1)} > 0$ – the feature **in favor** the prediction

$R_d^{(1)} < 0$ – the feature **against** the prediction

- Propagate the relevance scores using purposely designed local propagation rules
  - Basic Rule (LRP-0)
  - Epsilon Rule (LRP-ϵ) [Montavon et al., 2019]
  - Gamma Rule (LRP-γ)

Operations – Input perturbation

- Usually used to enable local explanation for a particular instance
  - Often combined with surrogate models



Operations – Input perturbation

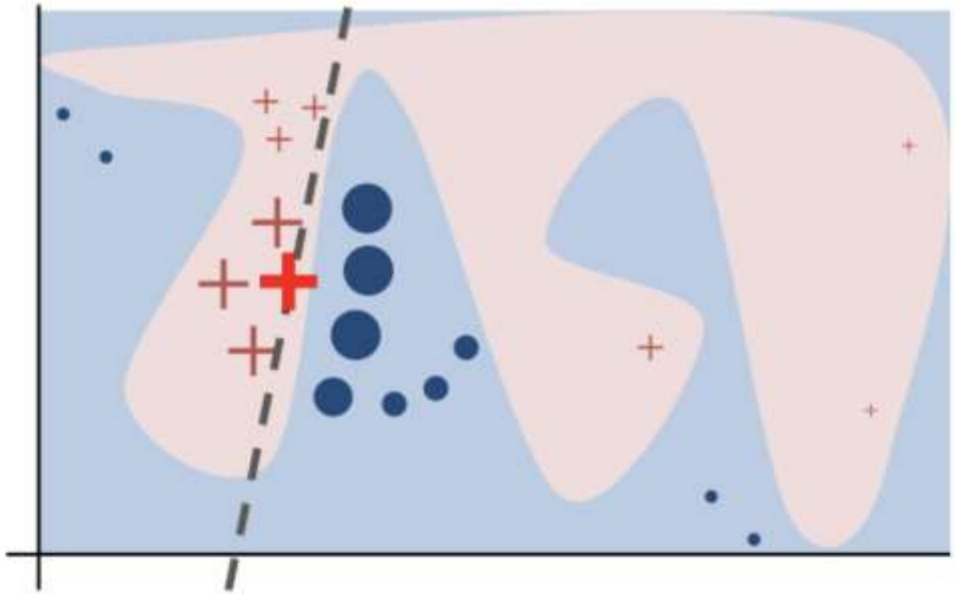
- Different ways to generate perturbations
  - Sampling method (e.g., [Ribeiro et al. KDD] - LIME)
  - Perturbating intermediate vector representation (e.g., [Alvarez-Melis and Jaakkola, 2017])

- **Sampling method**
  - Directly use examples within labeled data that are semantically similar

- **Perturbating intermediate vector representation**
  - Specific to neural networks
  - Introduce minor perturbations to the vector representation of input instance

Input perturbation – Sampling method

[Ribeiro et al. KDD]



Find examples that are both semantically “**close**” and “**far away**” from the instance in question.

- Need to define a proximity metric
- Will not generate instance

Perturbating intermediate Vector Representation

[Alvarez-Melis and Jaakkola, 2017])

- Explain predictions of any black-box structured input – structured output model
  - Around a specific input-output pair

- An explanation consists of groups of input-output tokens that are causally related.

- The dependencies (input-output) are inferred by querying the blackbox model with perturbed inputs

Perturbating intermediate Vector Representation

[Alvarez-Melis and Jaakkola, 2017])

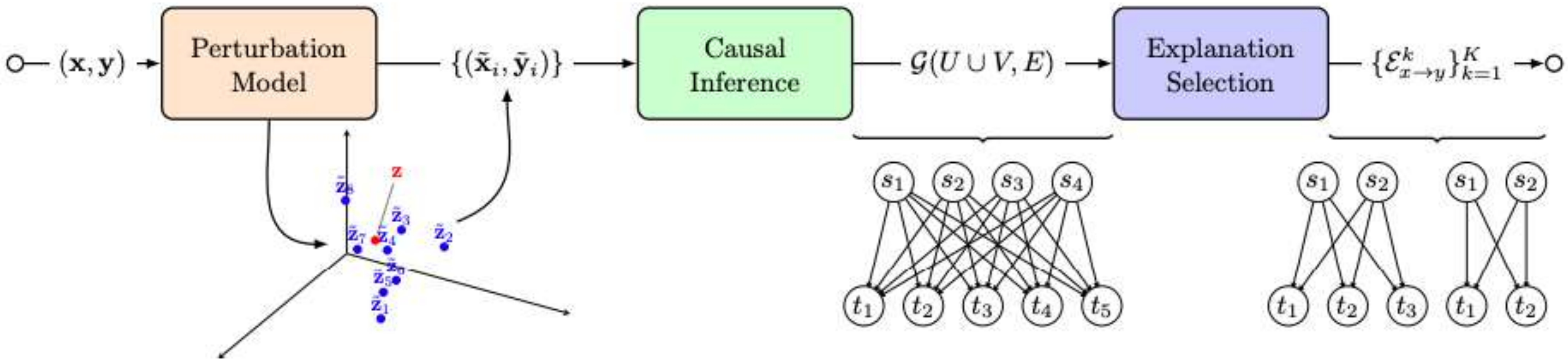
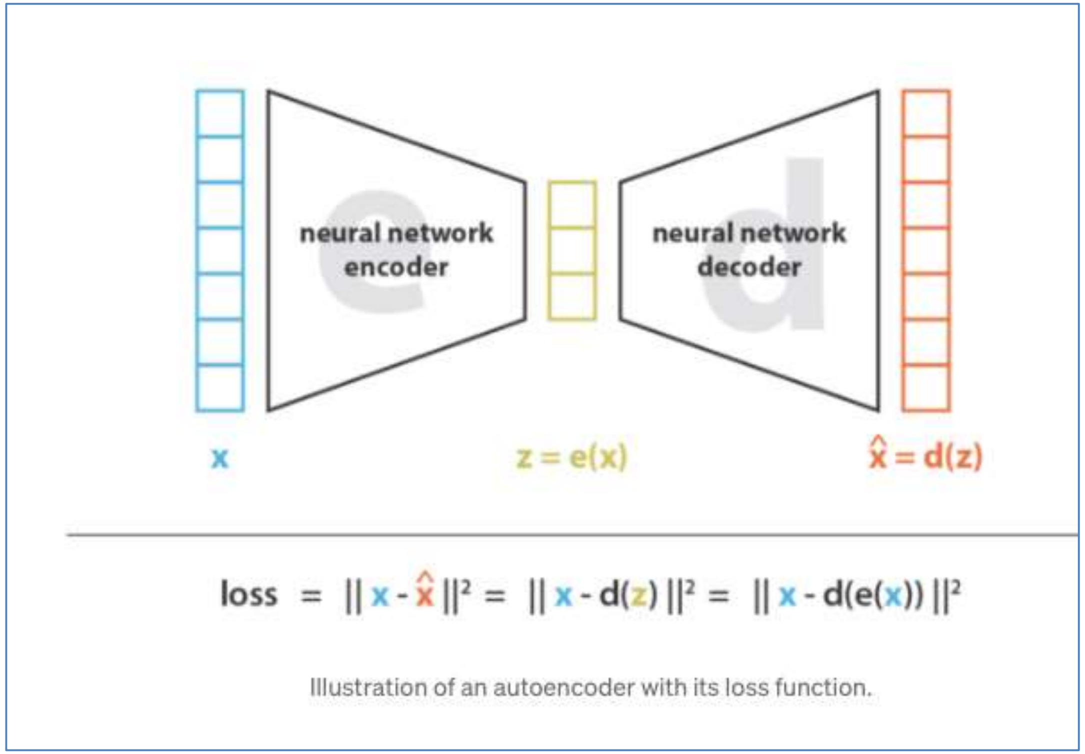


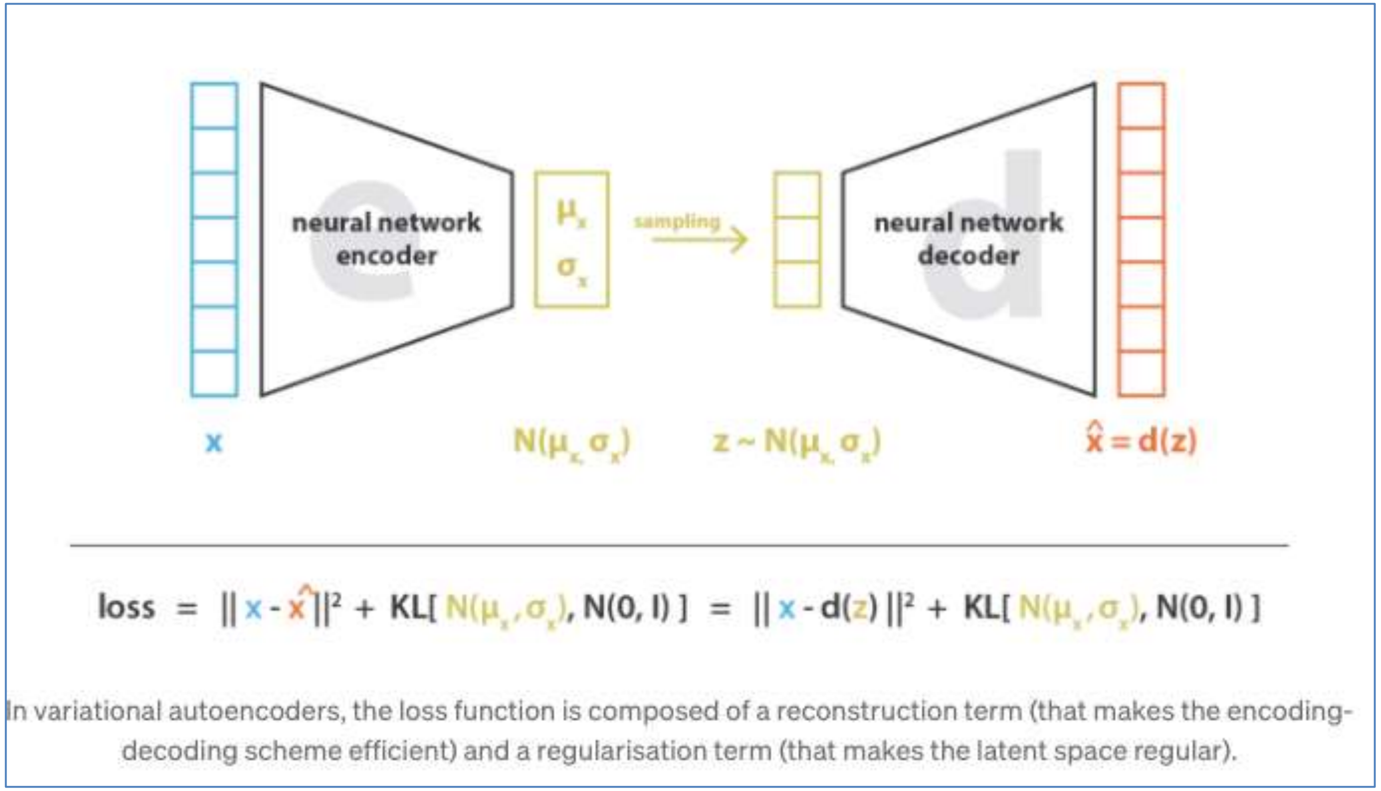
Figure 1: A schematic representation of the proposed prediction interpretability method.

Perturbating intermediate Vector Representation

[Alvarez-Melis and Jaakkola, 2017])



Typical autoencoder



Variation autoencoder
(generative model)

Image source: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Operations – Attention

[Ghaeini et al., 2018]

Naturally, can be used to enable feature importance

Attention mechanism uses the latent features learned by the encoder

- a weighted combination of the latent features

Weights become the **importance scores** of these latent features (also corresponds to the lexical features) for the decoded word/token.

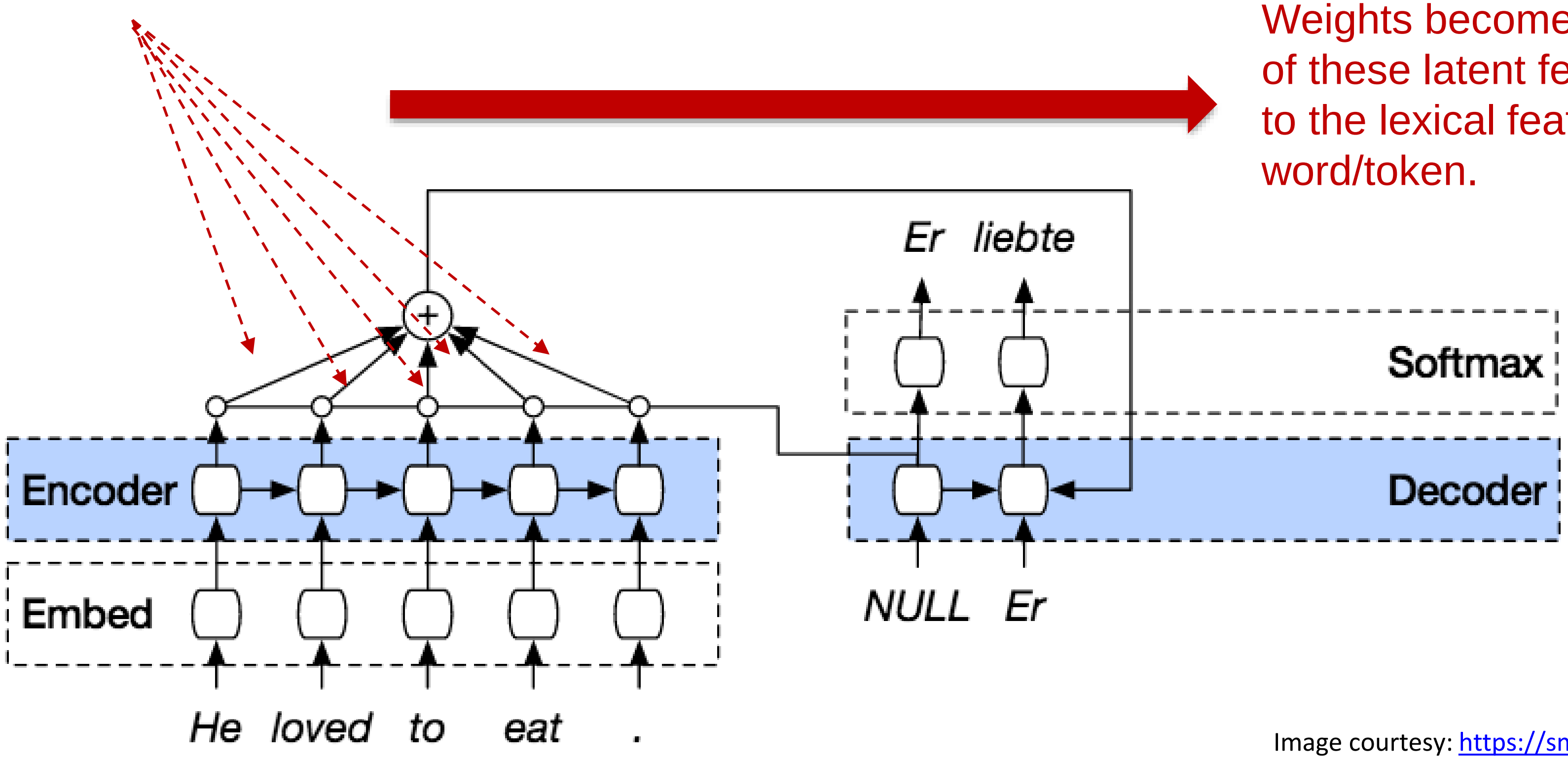


Image courtesy: https://smerity.com/articles/2016/google_nmt_arch.html

[NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. Bahdanau et al., ICRL 2015]

Operations – LSTM gating signals

- LSTM gating signals determine the flow of information [Ghaeini et al., 2018]
  - How LSTM reads the word sequences and how the information from different parts is captured and combined

- Gating signals are computed as the partial derivative of the score of the final decision wrt each gating signal.

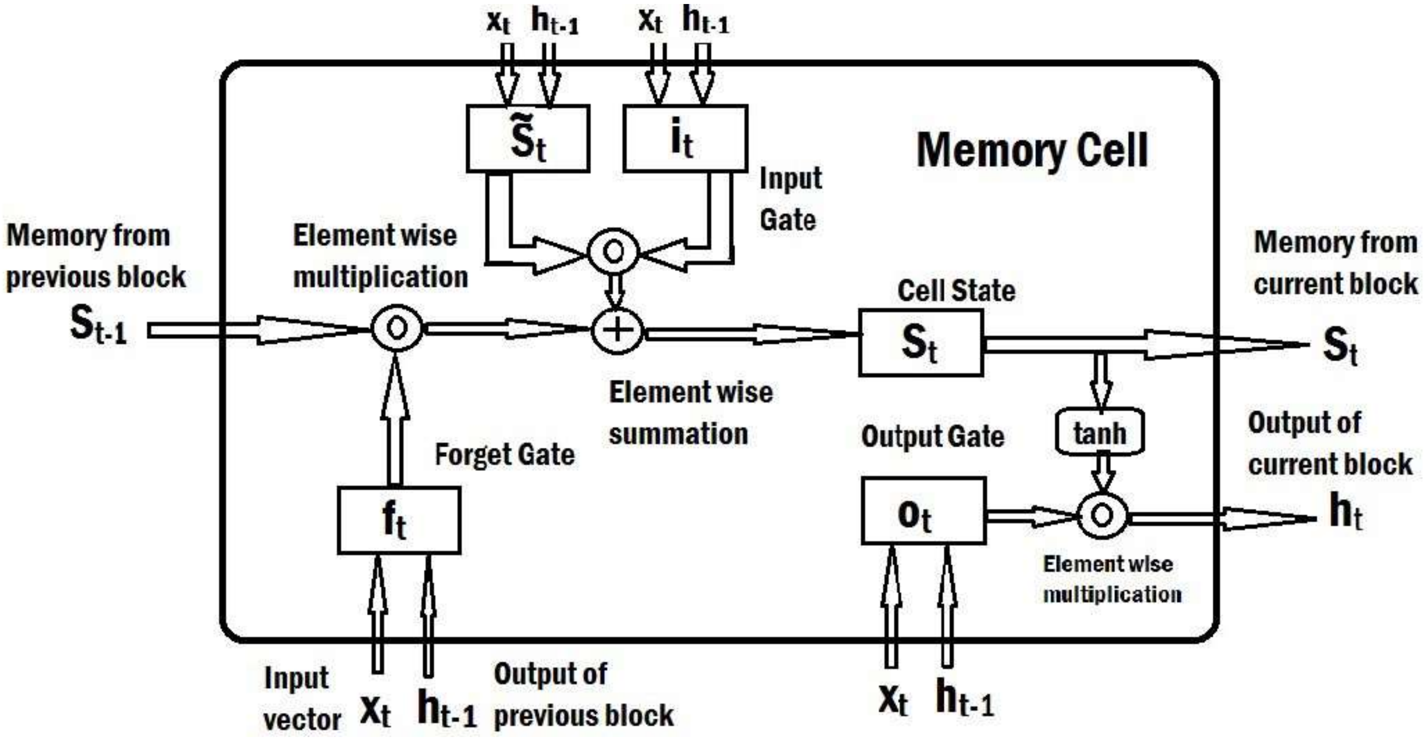


Image source: https://blog.goodaudience.com/basic-understanding-of-lstm-539f3b013f1e

Operations – LSTM gating signals

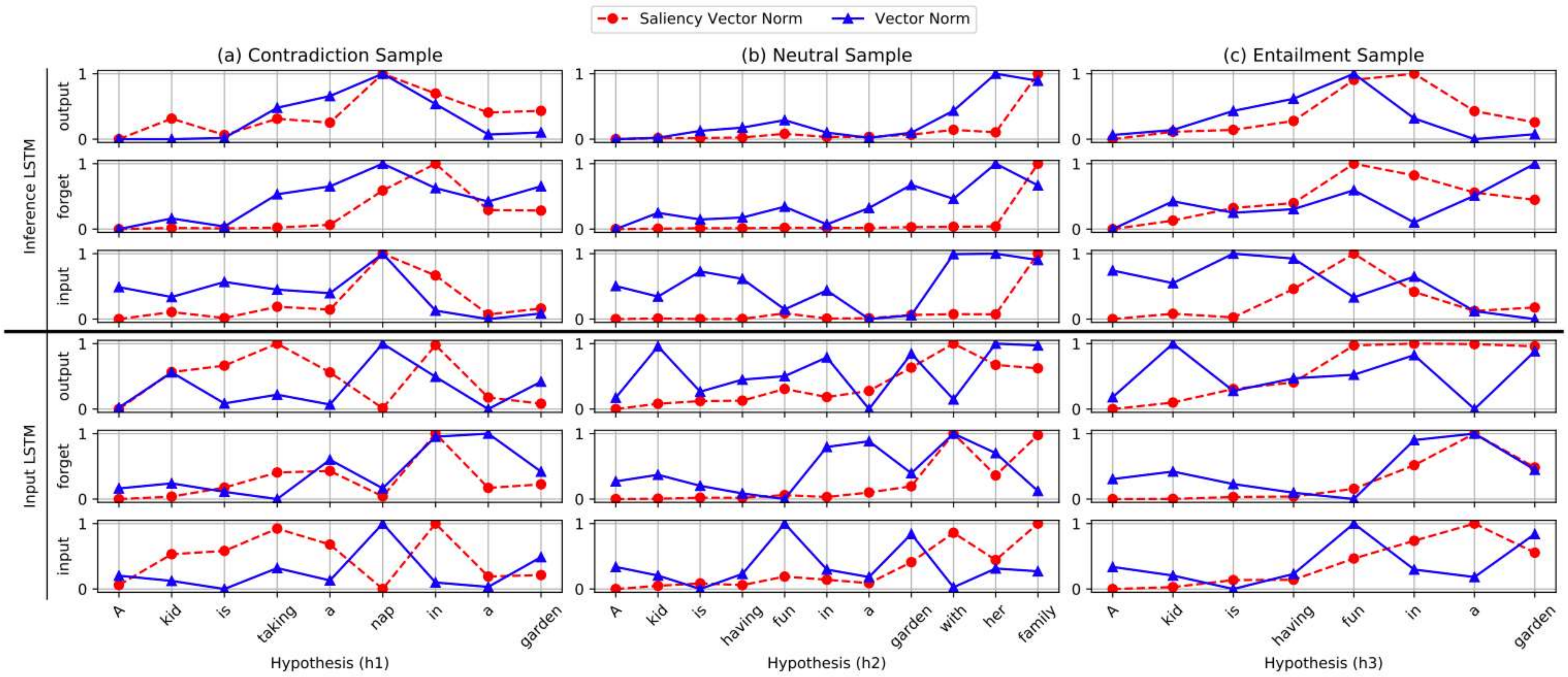


Figure 3: Normalized signal and saliency norms for the input and inference LSTMs (forward) of ESIM-50 for three examples. The bottom (top) three rows show the signals of the input (inference) LSTM. Each row shows one of the three gates (input, forget and output).

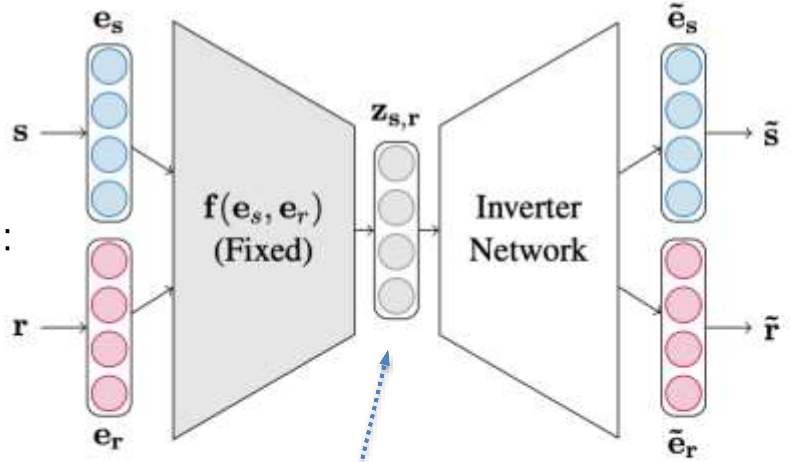
Explainability-Aware: Rule Induction Architecture

P. Pezeshkpour et al. *Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications*. NAACL 2019.

Works on top of existing, black-box models for link prediction, via a decoder architecture (i.e., from embeddings back to graph)

Rule Body, $R_1(a,c) \wedge R_2(c,b) \Rightarrow$	**Target,** $R(a,b)$
isConnectedTo$(a,c)\wedge$ isConnectedTo(c,b)	isConnectedTo
isLocatedIn$(a,c)\wedge$ isLocatedIn(c,b)	isLocatedIn
isAffiliatedTo$(a,c)\wedge$ isLocatedIn(c,b)	wasBornIn
isMarriedTo$(a,c)\wedge$ hasChild(c,b)	hasChild

Knowledge graph facts:
wasBornIn (s, o)



Decode back the vector $z_{s'r'}$ to actual KG:
isLocatedIn (s', o)

Optimization problem in R^d: find the most influential vector $z_{s'r'}$ whose removal would lead to removal of wasBornIn (s, o)

Rules of length 2 are then extracted by identifying frequent patterns

isAffiliatedTo (a, c) ∧ isLocatedIn (c, b) → wasBornIn (a,b)

Explainability-Aware Design: Rule-Based Active Learning

Qian et al, *SystemER: A human-in-the-loop System for Explainable Entity Resolution. VLDB’17 Demo* (also full paper in CIKM’17)

Iteratively learns an ER model via **active learning** from few labels.

- **Rules**: explainable representation for *intermediate* states and *final* model
- **Rules**: used as mechanism to generate and explain the examples brought back for labeling

Large datasets to match

corporation	Zip	City	address
Trader Joe’s	95118	San Jose	5353 Almaden Express Way

company	Zip	City	address
Trader Joe’s	95110	San Jose	635 Coleman Ave

Rule Learning Algorithm

Candidate rule learning

R.corporation = S.company and R.city = S.city

refinement

R.corporation = S.company and R.city = S.city and R.zip = S.zip

Example selection

corporation	Zip	City	address
KFC	90036	LA	5925 W 3rd St.

company	Zip	City	address
KFC	90004	LA	340 N Western Ave

Domain expert labeling

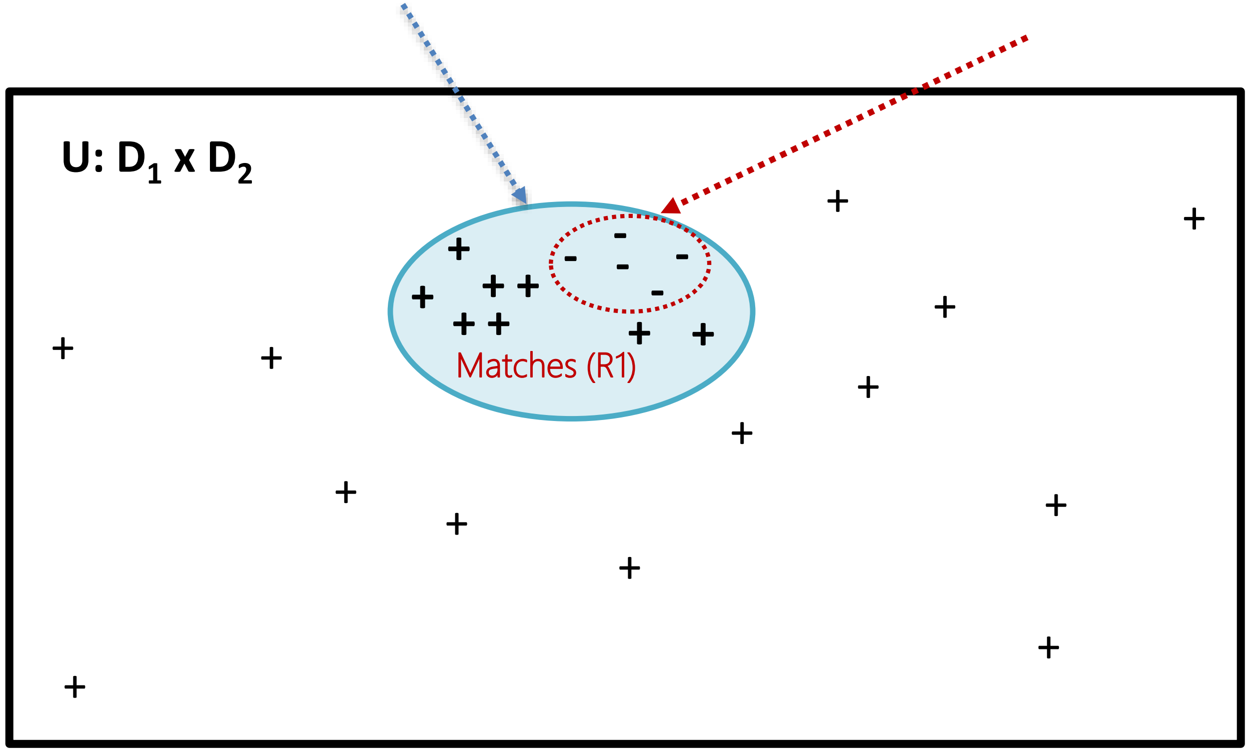
✕ ***Not a match***

SystemER: Example Selection (Likely False Positives)

Candidate rule R1

To refine R1 so that it becomes high-precision, need to find examples from the matches of R1 that are likely to be false positives

Negative examples will enhance precision



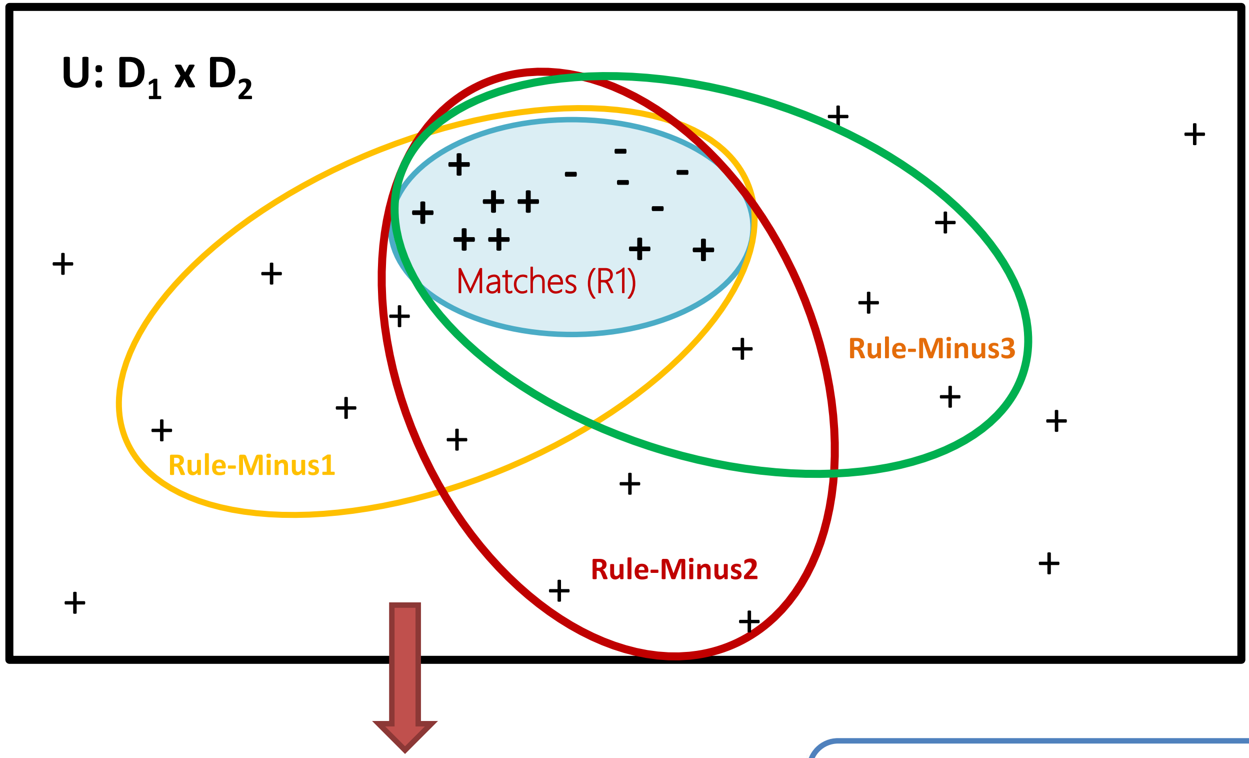
SystemER: Example Selection (Likely False Negatives)

Candidate rule R1

To refine R1 so that it becomes high-precision, need to find examples from the matches of R1 that are likely to be false positives

Negative examples will enhance precision

Rule-Minus heuristic: explore beyond the current R1 to find what it misses (likely false negatives)



New positive examples will enhance recall

R1:

	i.firstName = t.first
AND	**i.city = t.city**
AND	**i.lastName = t.last**

Rule-Minus1

	~~**i.firstName = t.first**~~
AND	**i.city = t.city**
AND	**i.lastName = t.last**

Rule-Minus2

	i.firstName = t.first
~~**AND**~~	~~**i.city = t.city**~~
AND	**i.lastName = t.last**

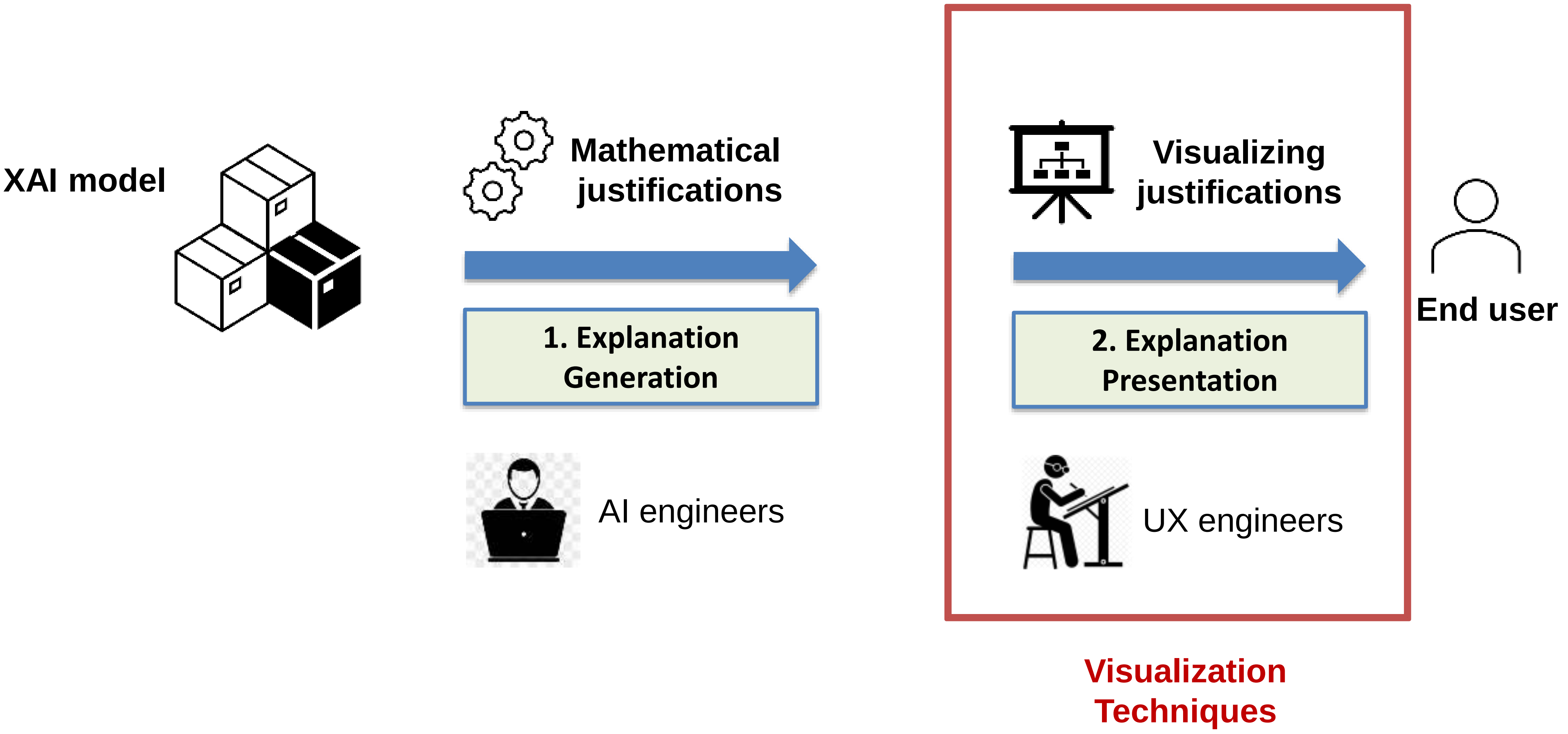
Rule-Minus3

	i.firstName = t.first
AND	**i.city = t.city**
~~**AND**~~	~~**i.lastName = t.last**~~

At each step, examples are explained/visualized based on the predicates they satisfy or not satisfy

Part II – (b) Visualization Techniques

What's next


XAI model
Mathematical justifications
1. Explanation Generation
AI engineers
Visualizing justifications
2. Explanation Presentation
UX engineers
End user
Visualization Techniques

Visualization Techniques

- Presenting raw explanations (mathematical justifications) to target user

- Ideally, done by UX/UI engineers

- Main visualization techniques:
  - Saliency
  - Raw declarative representations
  - Natural language explanation
  - Raw examples

Visualization – Saliency

Definition	has been primarily used to visualize the importance scores of different types of elements in XAI learning systems

- One of the most widely used visualization techniques
  - Not just in NLP, but also highly used in computer vision

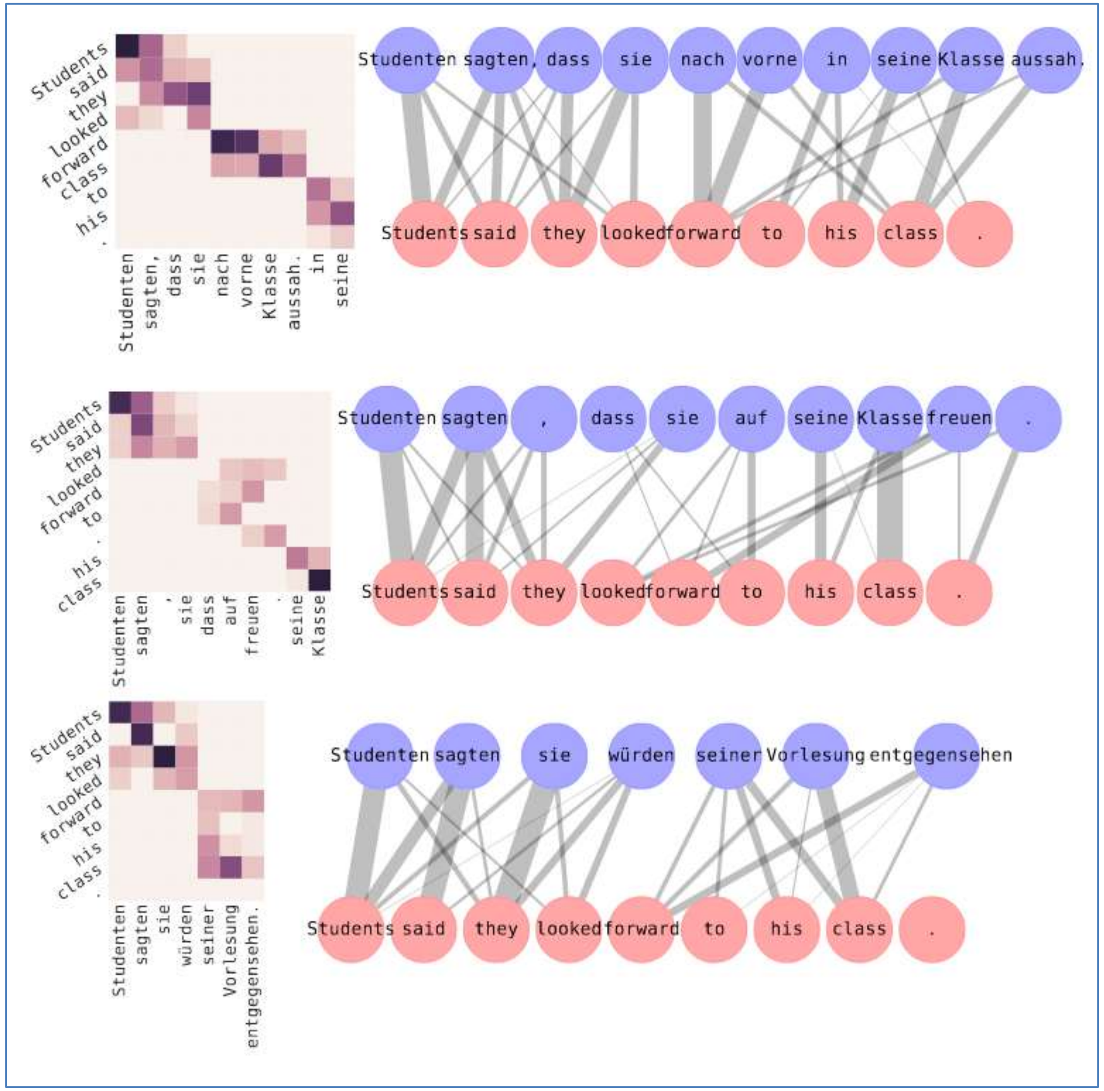
- Saliency is popular because it presents visually perceptive explanations
  - Can be easily understood by different types of users

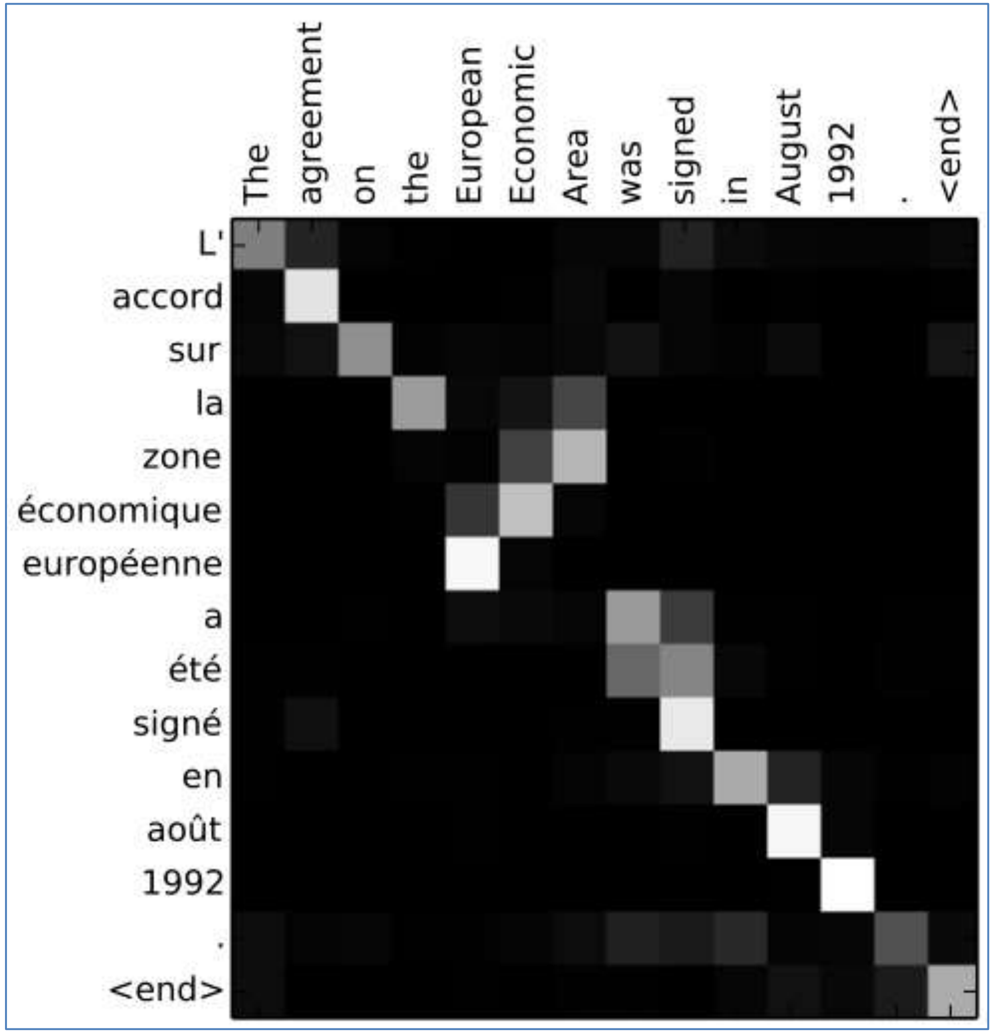
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

A computer vision example

Visualization – Saliency Example





[Bahdanau et al., 2015]

Input-output alignment

Visualization – Saliency Example

Source: " das trifft uns schwer , Rama ist ein herber Verlust " .

Machine (2.304, 0.993): " this strikes us hard , Rama is a severe loss ! "

Human (-2.625, 0.007): " it 's hit us hard . Rama is a bitter loss . "

Source: Jazz wurde , auch wenn er nicht direkt verboten war , nicht gespielt .

Machine (2.97, 0.998): jazz was not played , even if it was not directly banned .

Human (-3.032, 0.002): jazz , too , without exactly being proscribed , wasn ' t played .

Source: Jumbo - Hersteller streiten im Angesicht großer Bestellungen über Sitzbreite

Machine (3.013, 0.999): Jumbo manufacturers argue over seat width in the face of large orders .

Human (-3.618, 0.001): jet makers feud over seat width with big orders at stake

[Schwarzenberg et al., 2019]

Prediction	Input	Saliency Map
Contradiction	Premise	a young boy reaches for and touches the propeller of a vintage aircraft.
	Hypothesis	a young boy swims in his pool.
Entailment	Premise	a brown a dog and a black dog in the edge of the ocean with a wave under them boats are on the water in the background.
	Hypothesis	the pets are sleeping on the grass..
	Premise	man in a blue shirt standing in front of a structure painted with geometric designs.
Entailment	Hypothesis	a man is wearing a blue shirt.
Entailment	Hypothesis	a man is wearing a black shirt.
	Color Legend	Positive Impact Negative Impact

[Wallace et al., 2018]

Great book for travelling Europe : I currently live in Europe , and this is the book I recommend for my visitors . It covers many countries , colour pictures , and is a nice starter for before you go , and once you are there .

Figure 1: Contributions to positive classification.

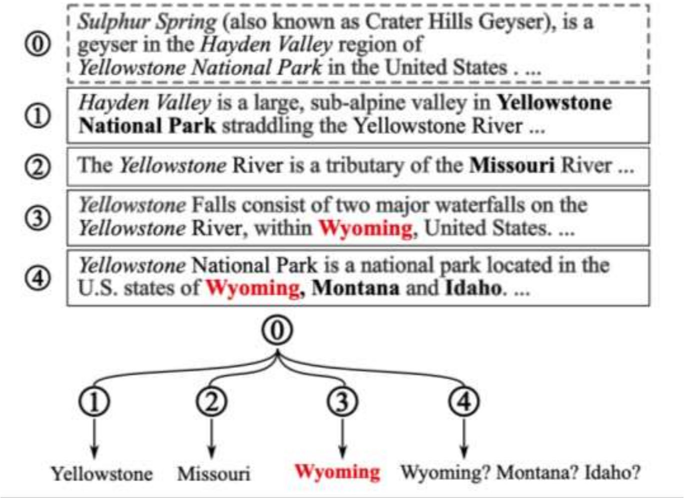
[Harbecke et al, 2018]

Word	Left	Right	Examples
0.24	**0.38**	**0.38**	<BOS> *langmore* **, a persistent campaigner for interventionist economic**
0.15	**0.59**	0.26	<BOS> **australian parliamentarian john** *langmore* has formally resigned from his lower house
0.15	**0.61**	0.24	**had received today from mr john vance** *langmore* , a letter resigning his place as
0.15	**0.69**	0.16	<BOS> **rtrs - australian mp john** *langmore* formally resigns . <EOS>
0.17	**0.40**	**0.43**	<BOS> *langmore* **, 57 , announced in november that**

[Garneau et al., 2018]

Visualization – Declarative Representations

- Trees:
  [Jiang et al., ACL 2019]
  – Different paths in the tree explain the possible reasoning paths that lead to the answer

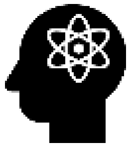


- Rules:

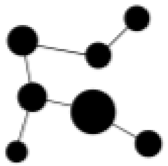
  [Pezeshkpour et al, NAACL 2019]
  – Logic rules to explain link prediction
  (e.g., isConnected is often predicted because of transitivity rule, while hasChild is often predicted because of spouse rule)

isConnected(a,c) $\wedge$ isConnected (c,b) $\Longrightarrow$ isConnected(a,b)
isMarriedTo (a,c) $\wedge$ hasChild (c,b) $\Longrightarrow$ hasChild(a,b)

  [Sen et al, EMNLP 2020]
  – Linguistic rules for sentence classification:
    - More than just visualization/explanation
    - Advanced HCI system for **human-machine co-creation** of the final model, built on top of rules
      (Next)



Human feedback



```
∃ verb ∈ S:
    verb.theme ∈ ThemeDict
    and verb.voice = passive
    and verb.tense = future
```

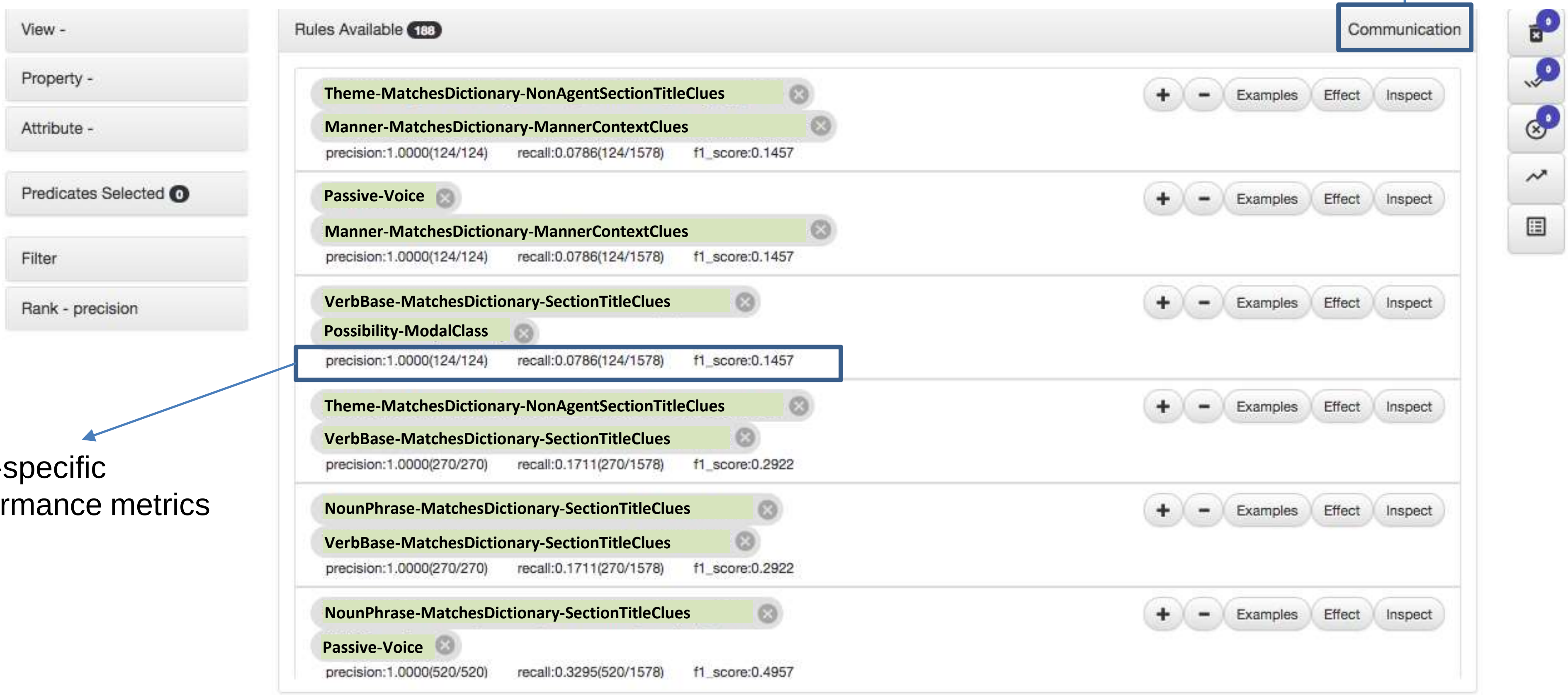
Machine-generated rules

Linguistic Rules: Visualization and Customization by the Expert

Sentence classification rules learned by [Sen et al, *Learning Explainable Linguistic Expressions with Neural Inductive Logic Programming for Sentence Classification.* EMNLP 2020]

Label being assigned

Various ways of selecting/ranking rules

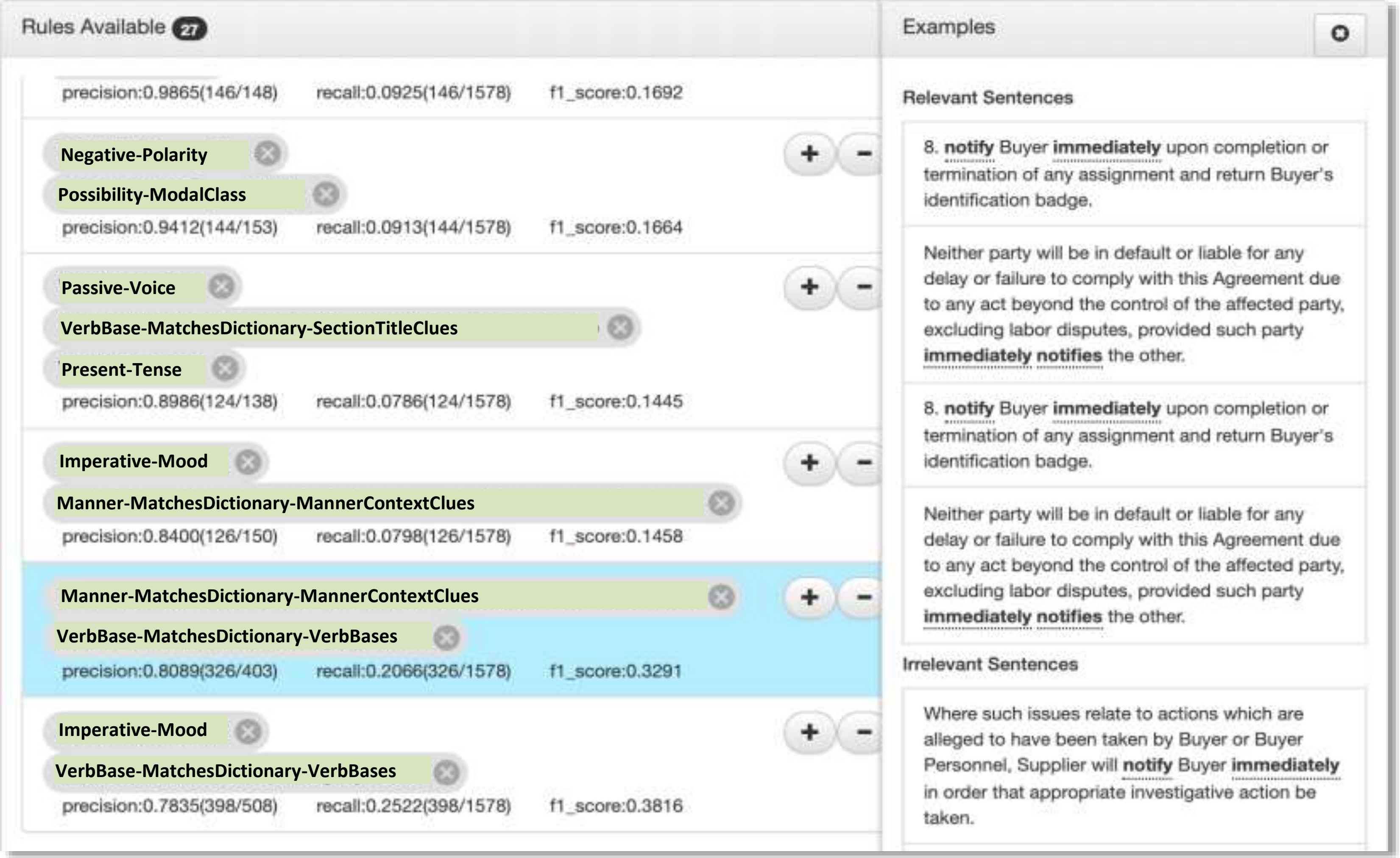


Rule-specific performance metrics

Y. Yang, E. Kandogan, Y. Li, W. S. Lasecki, P. Sen. ***HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop.*** **ACL, System Demonstration, 2019.**

Linguistic Rules: Visualization and Customization by the Expert

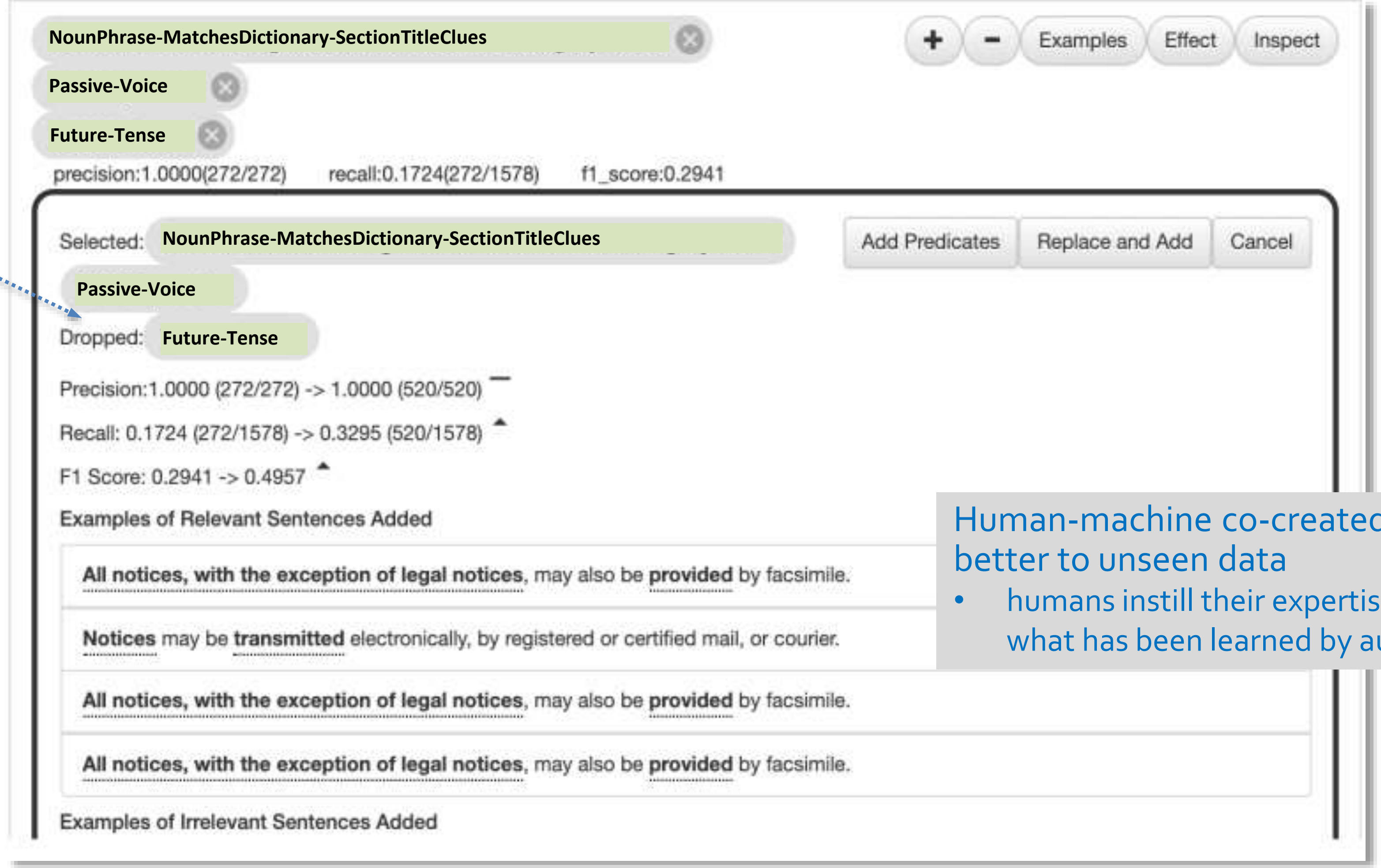
Visualization of examples per rule



Y. Yang, E. Kandogan, Y. Li, W. S. Lasecki, P. Sen. ***HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop.*** ACL, System Demonstration, 2019.

Linguistic Rules: Visualization and Customization by the Expert

Playground mode allows to add/drop predicates



Human-machine co-created models generalize better to unseen data

- humans instill their expertise by extrapolating from what has been learned by automated algorithms

Y. Yang, E. Kandogan, Y. Li, W. S. Lasecki, P. Sen. ***HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop.*** **ACL, System Demonstration, 2019.**

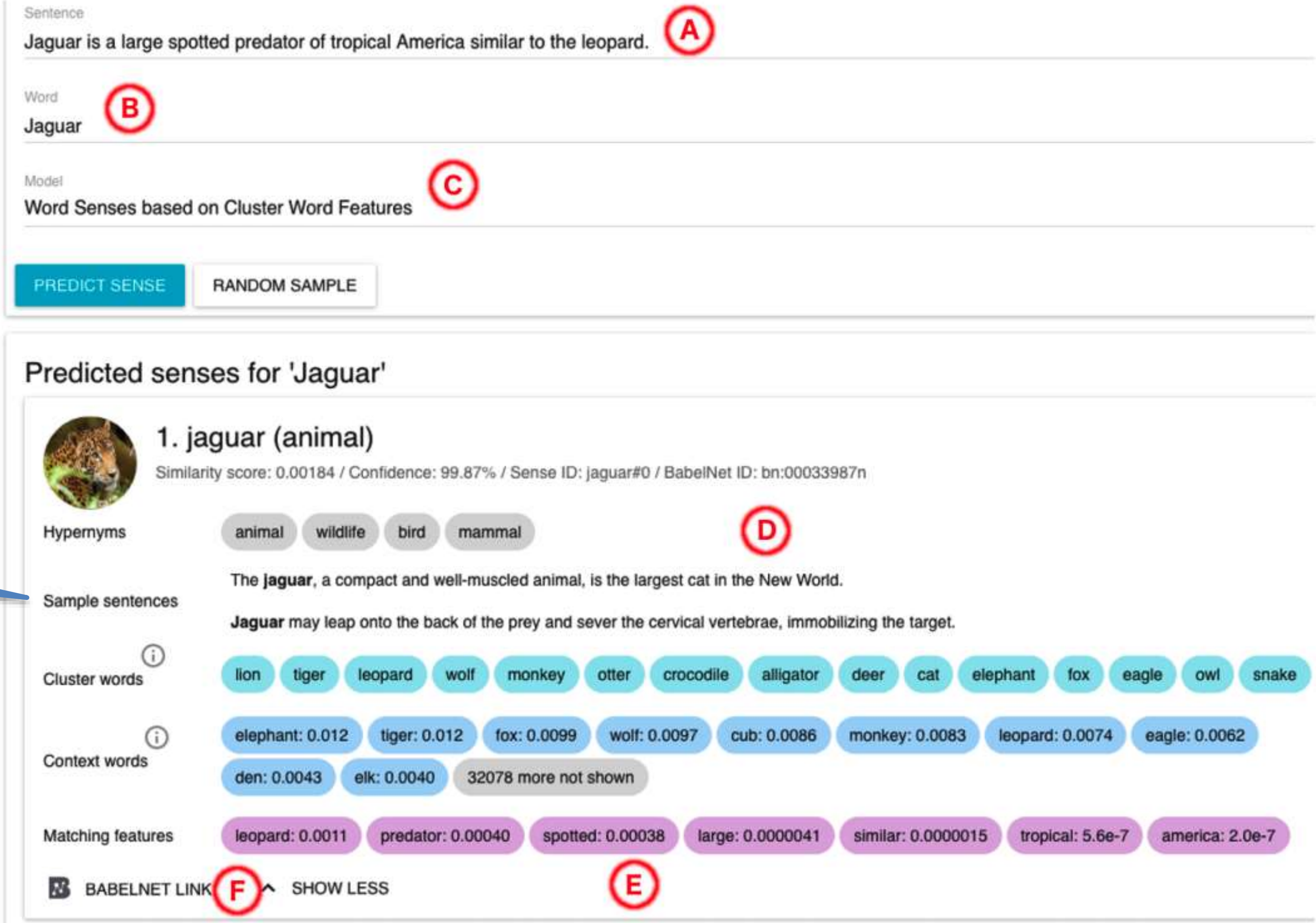
Visualization – Raw Examples

Definition	Explaining by presenting one (or some) instance(s) (usually from the training data) that is semantically similar to the instance needs to be explained

- Use existing knowledge to explain a new instance

Jaguar is a large spotted predator of tropical America similar to the leopard.

Raw examples



Word sense disambiguation

[Panchenko et al., 2016]

Visualization – Raw Examples

Sentences taken from the labeled data

I think **"this plate"** *is* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *in "Can you* PUT ***the soap*** *in the washing machine?"*.

Explain with one instance (basic model)

I think **"this plate"** *is* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *in "Can you* PUT ***"the soap"*** *in the washing machine?" and it is also similar to* **"my coat"** *in "*HANG ***my coat*** *in the closet in the bedroom"*.

Explain with > 1 instance (multiplicative model)

I think **"this plate"** *is the* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *which is in "Can you* PUT ***the soap*** *in the washing machine" and it is not the* GOAL *of* PLACING *since different from* **"on the counter"** *in "*PUT *the plate* ***on the counter***".

Explain with both positive and negative instance (contrastive model)

Classification
[Croce et al, 2018]

Visualization – Natural language explanation

Definition	The explanation is verbalized in human-comprehensible language

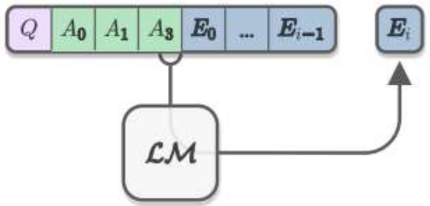
- Can be generated by using sophisticated deep learning approaches
  – [Rajani et al., 2019]

- Can be generated by simple template-based approaches
  – [Abujabal et al,, 2017 - QUINT]
  – [Croce et al, 2018]
  – Many declarative representations can be naturally extended to natural language explanation

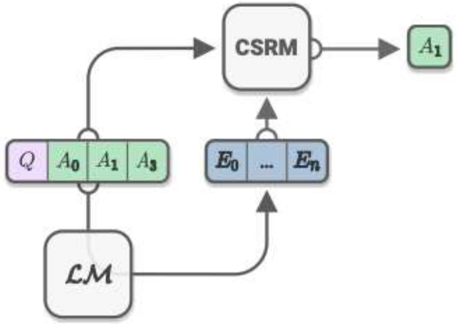
Visualization – Natural language explanation

[Rajani et al., 2019]

A language model trained with human generated explanations for commonsense reasoning



(a) One time-step of training a CAGE language model to generate explanations from CoS-E. It is conditioned on the question tokens Q concatenated with the answer choice tokens A_1, A_2, A_3 and previously generated tokens $E_1, \ldots, E_{i-1}$. It is trained to generate token E_i.



(b) A trained CAGE language model is used to generate explanations for a downstream commonsense reasoning model (CSRM), which itself predicts one of the answer choices.

Question:	They were getting ready for a really long hike, he put the food in his what?
Choices:	recycling center, house, **backpack**
CoS-E:	Backpacks are used on hikes
Reason	a backpack is a place to store food and supplies.
Rationale:	a backpack is used to carry food and supplies

Open-ended human explanations

Explanation generated by **CAGE**

Visualization – Natural language explanation

Sentences taken from the labeled data

I think **"this plate"** *is* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *in "Can you* PUT ***the soap*** *in the washing machine?"*.

Explain with one instance (basic model)

I think **"this plate"** *is* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *in "Can you* PUT ***"the soap"*** *in the washing machine?" and it is also similar to* **"my coat"** *in "*HANG ***my coat*** *in the closet in the bedroom"*.

Explain with > 1 instance (multiplicative model)

I think **"this plate"** *is the* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *which is in "Can you* PUT ***the soap*** *in the washing machine" and it is not the* GOAL *of* PLACING *since different from* **"on the counter"** *in "*PUT *the plate* ***on the counter"***.

Explain with both positive and negative instance (contrastive model)

Classification
[Croce et al, 2018]

Visualization – Natural language explanation

$$M(e, \mathcal{L}_k) = \begin{cases} \text{'}s \text{ is } C \text{ since it is similar to } \ell\text{'} \\ \forall \ell \in \mathcal{L}_k^+ \quad \text{if } \tau > 0 \\ \\ \text{'}s \text{ is not } C \text{ since it is different} \\ \text{from } \ell \text{ which is } C\text{'} \\ \forall \ell \in \mathcal{L}_k^- \quad \text{if } \tau < 0 \\ \\ \text{'}s \text{ is } C \text{ but I don't know why'} \\ \text{if } \mathcal{L} \equiv \emptyset \end{cases}$$

Template-based explanatory model

I think **"this plate"** *is* THEME *of* PLACING *in "Robot* PUT ***this plate*** *in the center of the table" since similar to* **"the soap"** *in "Can you* PUT ***the soap*** *in the washing machine?"*.

Classification
[Croce et al, 2018]

Outline of Part II

- **Literature review methodology**

- **Categorization of different types of explanation**
  - Local vs. Global
  - Self-explaining vs. Post-hoc processing

- **Generating and presenting explanations**
  - Explainability techniques
    - Common operations to enable explainability
  - Visualization techniques

- **Other insights (XNLP website)**
  - Relationships among explainability and visualization techniques
  - Evolution of the XAI research in the past a few years

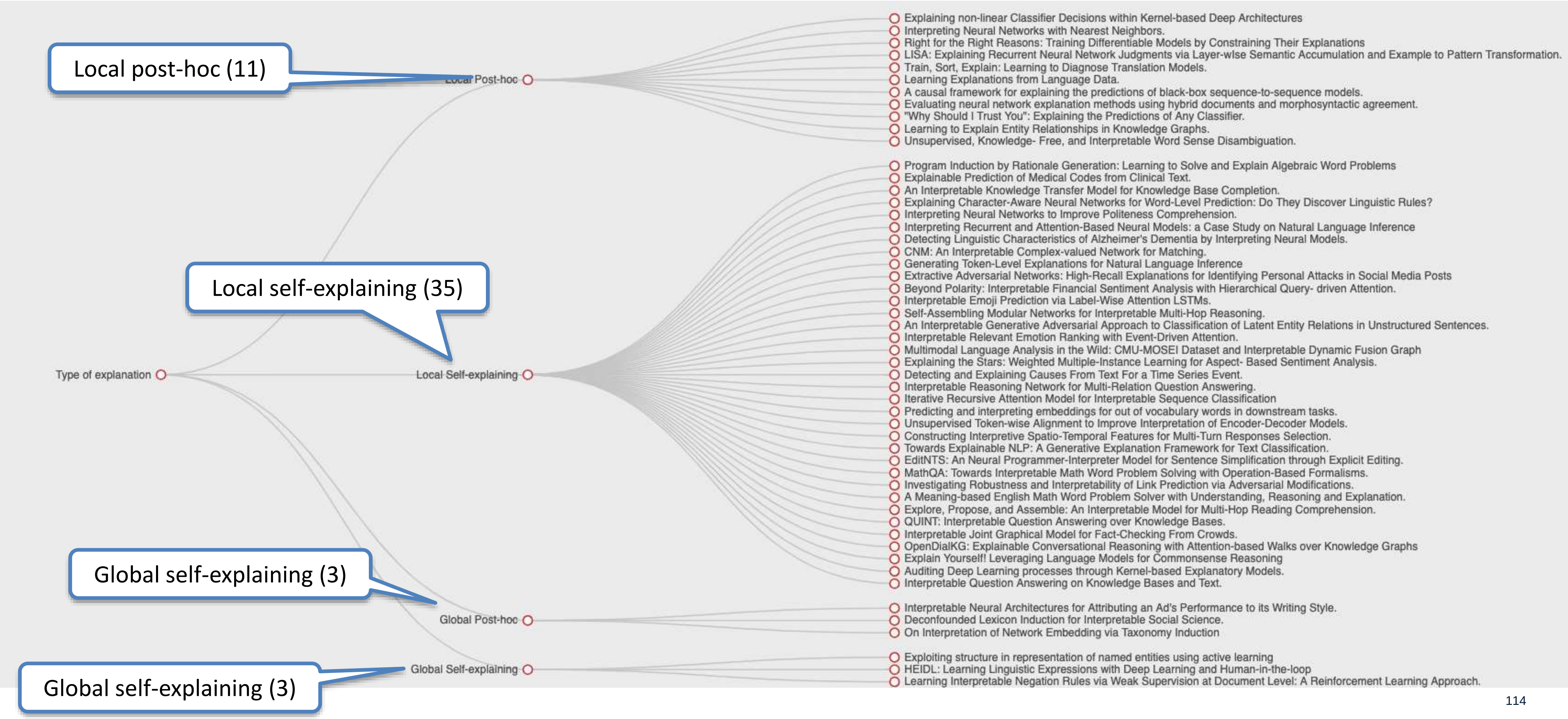
- **Evaluation of Explanations**

XNLP

- We built an interactive website for exploring the domain
- It's self-contained
- It provides different ways for exploring the domain
  - Cluster View (group papers based on explainability and visualization)
  - Tree view (categorize papers in a tree like structure)
  - Citation graphs (show the evolution of the field and also show influential works)
  - List view (list the set of papers in a table)
  - Search view (support keyword search and facet search)
- Link to the website: https://xainlp2020.github.io/xainlp/home

XNLP - Tree View

- https://xainlp2020.github.io/xainlp/clickabletree



XNLP – List View

- https://xainlp2020.github.io/xainlp/table

Showing **6** papers for **Global**

(Columns are sortable, and rows are clickable)

Year ↑	Title	Main Explainability	Main Visualization ↑	Citations ↑	Venue ↑	NLP Topic ↑
2019	Heidi: Learning Linguistic Expressions With Deep Learning And Human-in-the-loop	Induction	Raw Declarative	0	ACL	Sentence Classification
2018	Learning Interpretable Negation Rules Via Weak Supervision At Document Level: A Reinforcement Learning Approach.	Induction	Raw Declarative	4	NAACL	Sentiment Analysis, Stylistic Analysis, And Argument Mining
2018	Exploiting Structure In Representation Of Named Entities Using Active Learning	Induction	Raw Declarative	3	COLING	Syntax: Tagging, Chunking And Parsing
2018	Interpretable Neural Architectures For Attributing An Ad's Performance To Its Writing Style.	Feature Importance	Saliency	4	EMNLP	Computational Social Science And Social Media
2018	Deconfounded Lexicon Induction For Interpretable Social Science.	Feature Importance	Saliency	6	NAACL	Computational Social Science And Social Media
2018	On Interpretation Of Network Embedding Via Taxonomy Induction.	Surrogate Model	Other	8	KDD	Computational Social Science And Social Media

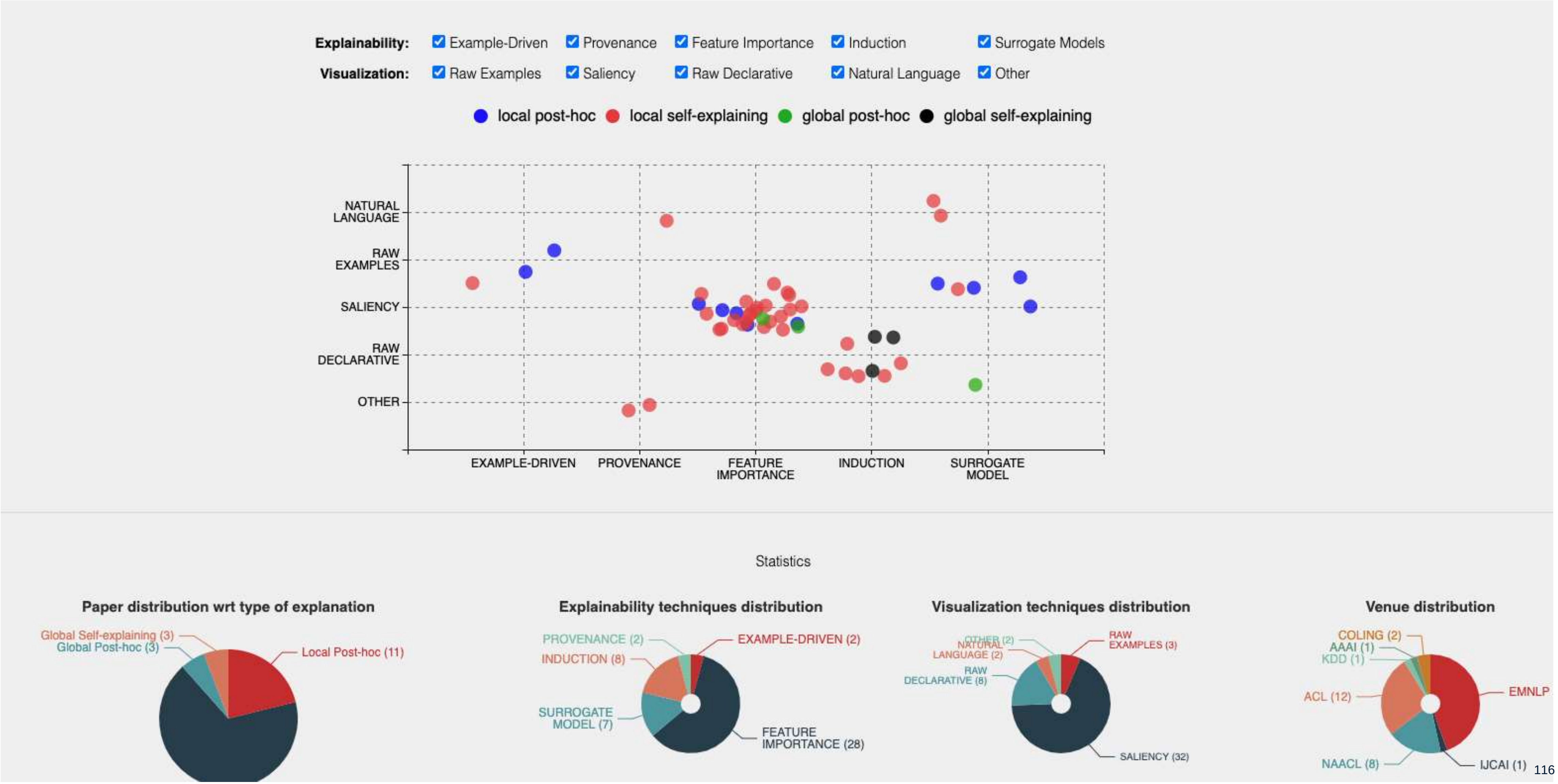
Showing **46** papers for **Local**

(Columns are sortable, and rows are clickable)

Year ↑	Title	Main Explainability	Main Visualization ↑	Citations ↑	Venue ↑	NLP Topic ↑
2017	Unsupervised, Knowledge- Free, And Interpretable Word Sense Disambiguation.	Example-driven	Raw Examples	9	EMNLP	Semantics: Lexical
2018	Explaining Non-linear Classifier Decisions Within Kernel-based Deep Architectures	Example-driven	Raw Examples	5	EMNLP	Semantics: Sentence Level
2019	Auditing Deep Learning Processes Through Kernel-based Explanatory Models.	Example-driven	Raw Examples	0	EMNLP	Syntax: Tagging, Chunking And Parsing
2014	Explaining The Stars: Weighted Multiple-instance Learning For Aspect- Based Sentiment Analysis.	Feature Importance	Saliency	32	EMNLP	Sentiment Analysis, Stylistic Analysis, And Argument Mining
2018	Interpreting Neural Networks To Improve Politeness Comprehension.	Feature Importance	Saliency	22	EMNLP	Computational Social Science And Social Media
2017	Right For The Right Reasons: Training Differentiable Models By Constraining Their Explanations.	Feature Importance	Saliency	117	IJCAI	Semantics: Textual Inference And Other Areas Of Semantics
2017	Detecting And Explaining Causes From Text For A Time Series Event.	Feature Importance	Saliency	11	EMNLP	Nlp Applications
2017	An Interpretable Knowledge Transfer Model For Knowledge Base Completion.	Feature Importance	Saliency	33	ACL	Nlp Applications
2018	Learning Explanations From Language Data	Feature Importance	Saliency	1	EMNLP	Sentiment Analysis, Stylistic Analysis, And Argument Mining
2018	Lisa: Explaining Recurrent Neural Network Judgments Via Layer-wise Semantic Accumulation And Example To Pattern Transformation.	Feature Importance	Saliency	5	EMNLP	Information Extraction
2018	Interpreting Neural Networks With Nearest Neighbors.	Feature Importance	Saliency	10	EMNLP	Semantics: Textual Inference And Other Areas Of Semantics
2018	Unsupervised Token-wise Alignment To Improve Interpretation Of Encoder-decoder Models.	Feature Importance	Saliency	2	EMNLP	Summarization
2018	Predicting And Interpreting Embeddings For Out Of Vocabulary Words In Downstream Tasks.	Feature Importance	Saliency	2	EMNLP	Syntax: Tagging, Chunking And Parsing
2018	Iterative Recursive Attention Model For Interpretable Sequence Classification.	Feature Importance	Saliency	3	EMNLP	Sentiment Analysis, Stylistic Analysis, And Argument Mining
2018	Interpretable Reasoning Network For Multi-relation Question Answering.	Feature Importance	Saliency	11	COLING	Question Answering

XNLP – Cluster View

- https://xainlp2020.github.io/xainlp/viz



XNLP – Citation Graph

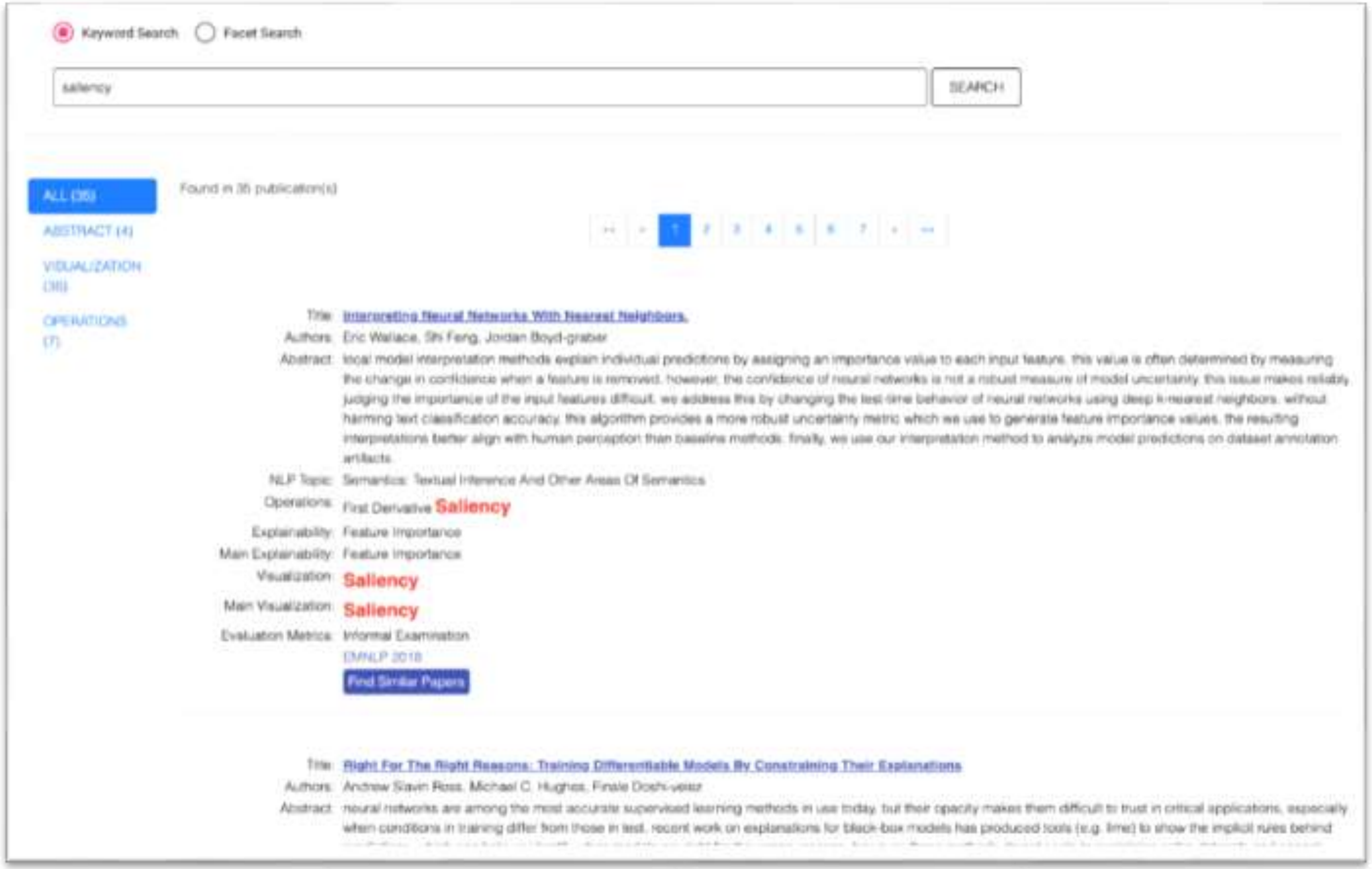
- https://xainlp2020.github.io/xainlp/network

XNLP – Citation Graph

- https://xainlp2020.github.io/xainlp/network

XNLP – Search View

- https://xainlp2020.github.io/xainlp/search



More advanced ways to explore the field can be found in our website

https://xainlp2020.github.io/xainlp

Outline of Part II

- **Literature review methodology**

- **Categorization of different types of explanation**
  - Local vs. Global
  - Self-explaining vs. Post-hoc processing

- **Generating and presenting explanations**
  - Explainability techniques
    - Common operations to enable explainability
  - Visualization techniques

- **Other insights (XNLP website)**
  - Relationships among explainability and visualization techniques
  - Evolution of the XAI research in the past a few years

- **Evaluation of Explanations**

How to evaluate the generated explanations

- An important aspect of XAI for NLP.
  - No just evaluating the accuracy of the predictions
- Unsurprisingly, **little agreement** on how explanations should be evaluated
  - 32 out of 52 lack formal evaluations
- Only a couple of evaluation metrics observed
  - Human ground truth data
  - Human evaluation

None or Informal Examination only	Comparison to Ground Truth	Human Evaluation
32	12	9

Table 3: Common evaluation techniques and number of papers adopting them, out of the 50 papers surveyed (note that some papers adopt more than one technique)

Evaluation – Comparison to ground truth

- Idea: compare generated explanations to ground truth explanations

- Metrics used
  - Precision/Recall/F1 (Carton et al.,2018)
  - BLEU (Ling et al., 2017; Rajani et al., 2019b)

- Benefit
  - A quantitative way to measure explainability

- Pitfalls
  - Quality of the ground truth data
  - Alternative valid explanations may exit

- Some attempts to avoid these issues
  - Having multiple annotators (with inter-annotator agreement)
  - Evaluating at different granularities (e.g., token-wise vs. phrase-wise)

Evaluation – Human Evaluation

- Idea: ask humans to evaluate the effectiveness of the generated explanations.

- Benefit
  - Avoiding the assumption that there is only one good explanation that could serve as ground truth

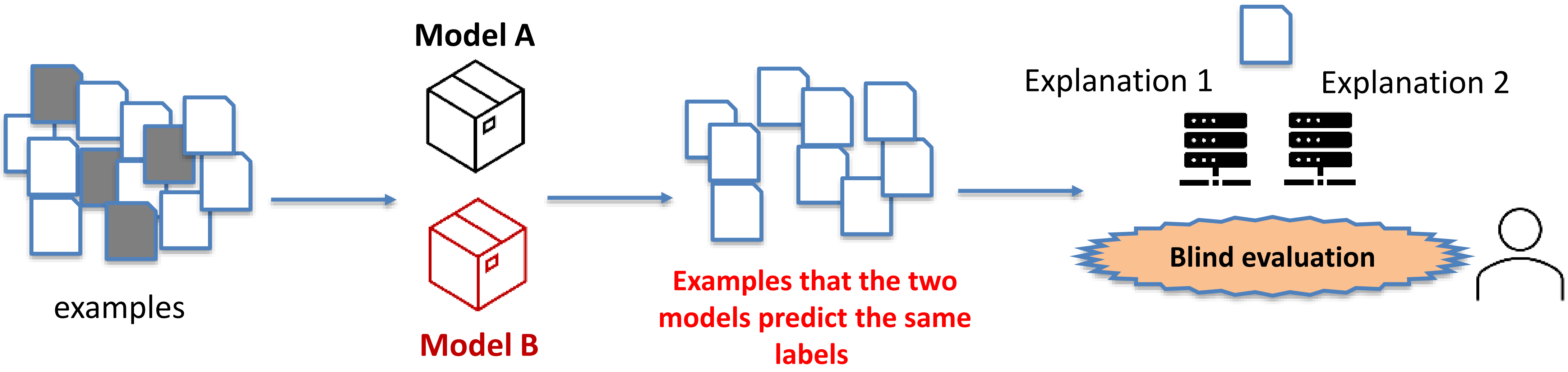
- Important aspects
  - It is important to have multiple annotators (with inter-annotator agreement, avoid subjectivity)

- Observations from our literature review
  - Single-human evaluation (Mullenbach et al., 2018)
  - Multiple-human evaluation (Sydorova et al., 2019)
  - Rating the explanations of a single approach (Dong et al., 2019)
  - Comparing explanations of multiple techniques (Sydorova et al., 2019)

- Very few well-established human evaluation strategies.

Human Evaluation – Trust Evaluation (Sydorova et al., 2019)

- Borrowed the trust evaluation idea used in CV (Selvaraju et al. 2016)
- A comparison approach
  - given two models, find out which one produces more intuitive explanations



Human Evaluation – Trust Evaluation (Sydorova et al., 2019)

Figure 2: Interface for the human annotation study.

Quality of Explanation – Predictive Process Coverage

- The predictive process starts from input to final prediction
- Not all XAI approaches cover the whole process
- Ideally, the explanation covers the whole process or most of the process
- However, many of the approaches cover only **a small part** of the process
  - The end user has to fill up the gaps

Quality of Explanation – Predictive Process Coverage

- Take MathQA (Amini et al. 2019) as an example

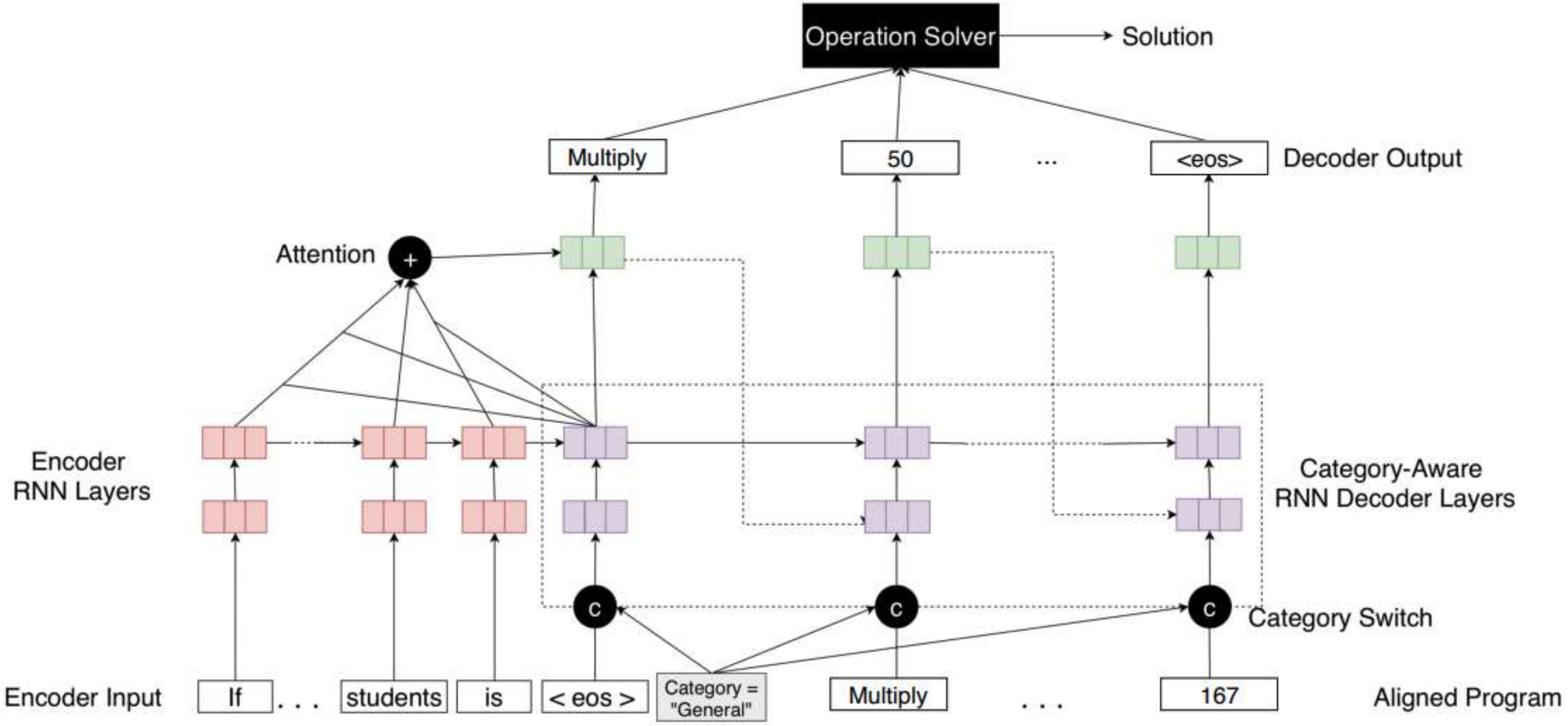
Problem : How long does a train 110m long running at the speed of 72 km/hr takes to cross a bridge 132m length?
Operations : `add(110, 132), multiply(72, const_0.2778), divide(#0, #1), floor(#2)`

Induced mathematical operations

Table 5: Problems solved correctly by Seq2prog+cat model.

Quality of Explanation – Predictive Process Coverage

- Take MathQA [(Amini et al. 2019)](#) as an example



Fidelity

- Often an issue arises in surrogate model
- Surrogate model is flexible since its model-agnostic
  - Local fidelity vs. global fidelity
- Logic fidelity
  - Surrogate model may use completely different reasoning mechanism
- ***Is it really wise to use a linear model to explain a deep learning model?***

PART III – Explainability & Case Study

Explainability

Human-Centered Problem

Human-centered perspective on XAI

Human-AI interaction guidelines (Amershi et al., 2019)

Explanations in social sciences (Miller, 2019)

Leveraging human reasoning processes to craft explanations (Wang et al., 2019)

Explainability scenarios (Wolf, 2019)

Broader ecosystem of actors (Liao et al., 2020)

Explanations for human-AI collaboration during onboarding (Cai et al, 2019)

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019, May). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.
Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-15).
Wolf, C. T. (2019, March). Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 252-257).
Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, *3*(CSCW), 1-24.

Human-Centered approach to Explainabilty

User Study + Case Study

Explainability in Practice

RQ1: What is the nature of explainability concerns in industrial AI projects?

RQ2: Over the course of a model's pipeline, when do explanation needs arise?

Interview Participants

AI Engineers -7

Data Scientists - 13

Designers – 2

Information Vis Specialists - 3

Product Managers - 3

Technical Strategists - 5

*some individuals were overlapping in two roles

Domains

Drug safety, key opinion leader mining, information extraction from business communications e.g., emails, contract document understanding, HR, chatbots, public health monitoring ….

Our approach

Who touches the AI model?

What are their informational needs?

When do explanations get warranted?

Business Users clients of AI model

AI Engineers train/fine-tune the model

AI Operations gather important metrics on AI models

Data Providers gatekeepers of public, private, third party data

Designers develop the user experience of the AI model

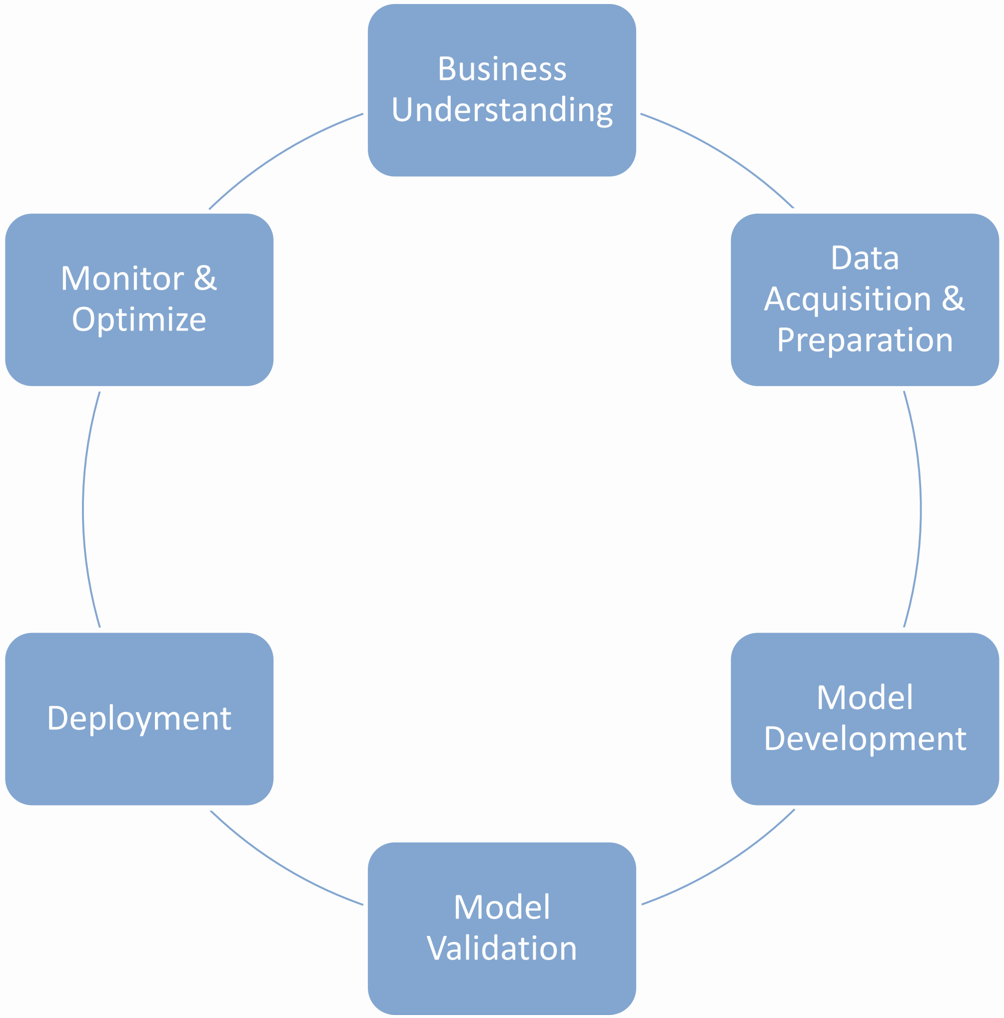
Data Scientists develop new models, design new algorithms

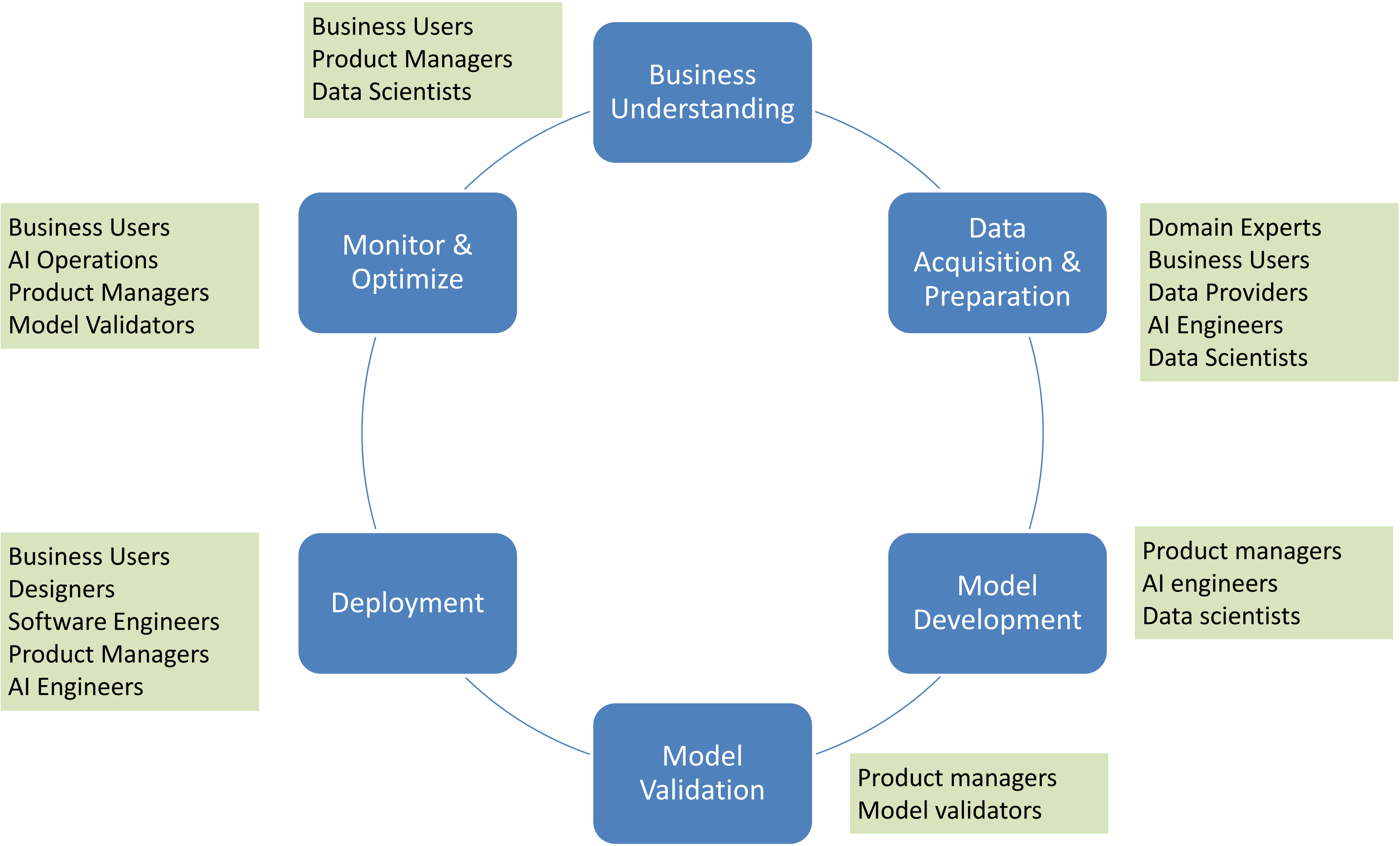
Domain Experts subject matter experts in business domain helping in labeling data

Model Validators debug, see if model continues to operate running as intended, detect model drift

Product Managers interface between business users and data scientists, AI engineers

Software Engineers support AI model deployment




Business Understanding
Business Users
Product Managers
Data Scientists
Data Acquisition & Preparation
Domain Experts
Business Users
Data Providers
AI Engineers
Data Scientists
Model Development
Product managers
AI engineers
Data scientists
Model Validation
Product managers
Model validators
Deployment
Business Users
Designers
Software Engineers
Product Managers
AI Engineers
Monitor & Optimize
Business Users
AI Operations
Product Managers
Model Validators

Explanations during model development

Understanding AI models inner workings

“*Explanations can be useful to NLP researchers explore and come up with hypothesis about why their models are working well…*
If you know the layer and the head then you have all the information you need to remove its influence... by looking, you could say, oh this head at this layer is only causing adverse effects, kill it and retrain or you could tweak it perhaps in such a way to minimize bias (I-19, Data scientist)

“*Low-level details like hyper-parameters would be discussed for debugging or brainstorming. (I-12, Data Scientist)*

Explanations during model validation

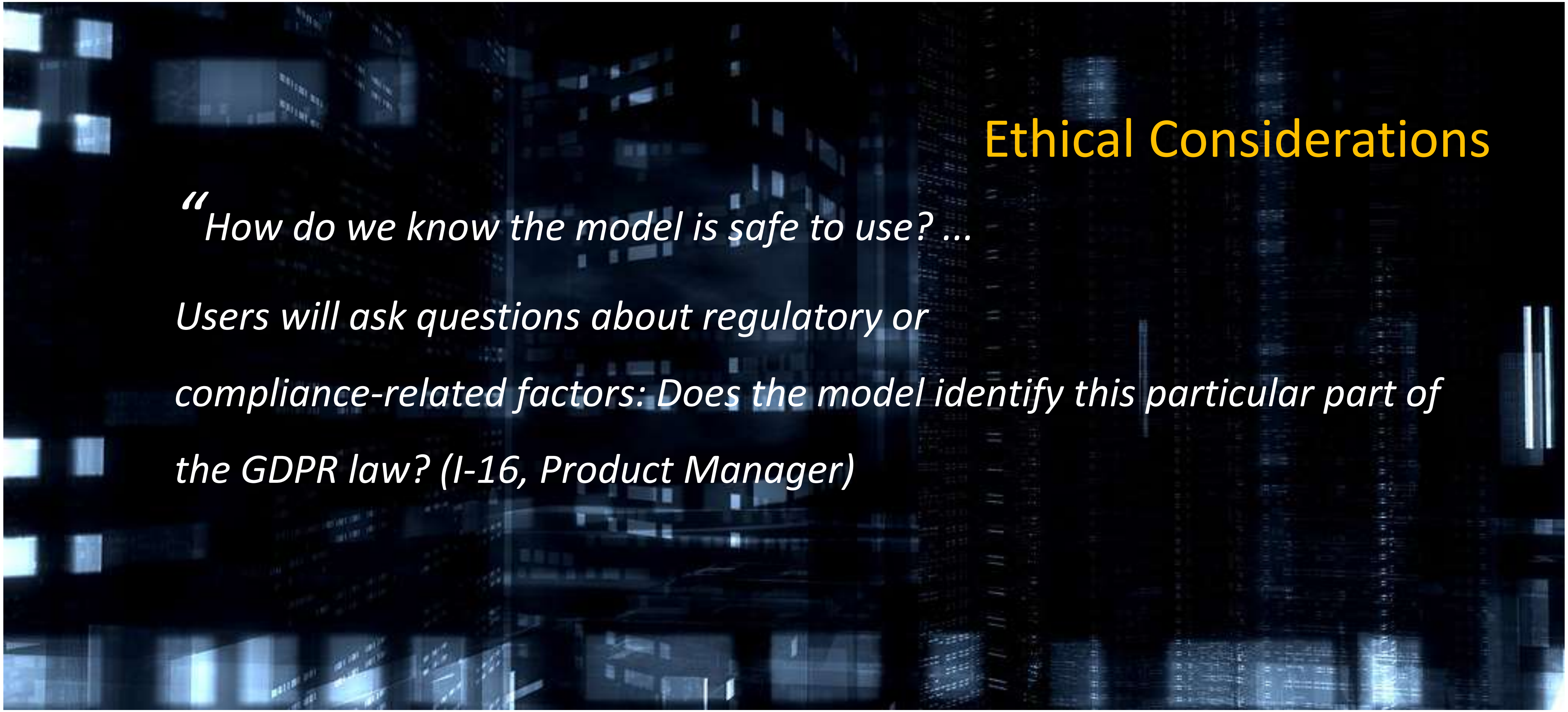
Details about data over which model is built

“Domain experts want to know more about the public medical dataset that the model is trained on to gauge if it can adapt well to their proprietary data (I-4, Data Scientist)

Model design at a high level

“Initially we presented everything in the typical AI way (i.e., showing the diagram of the model). We even put equations but realized this will not work... After a few weeks, we started to only show examples and describing how the model works at a high level with these examples (I-4)

Explanations during model validation



Ethical Considerations

"How do we know the model is safe to use? ...

Users will ask questions about regulatory or compliance-related factors: Does the model identify this particular part of the GDPR law? (I-16, Product Manager)

Explanations during model in-production

Expectation mismatch

“Data quality issues might arise resulting from expectation mismatch, but the list of recommendations must at least be as good as the manual process ... If they [the clients] come back with data quality issues ... we need to [provide] an explanation of what happened (I-5, Technical Strategist)

“Model mistakes versus decision disagreements” (I-28, Technical Strategist)

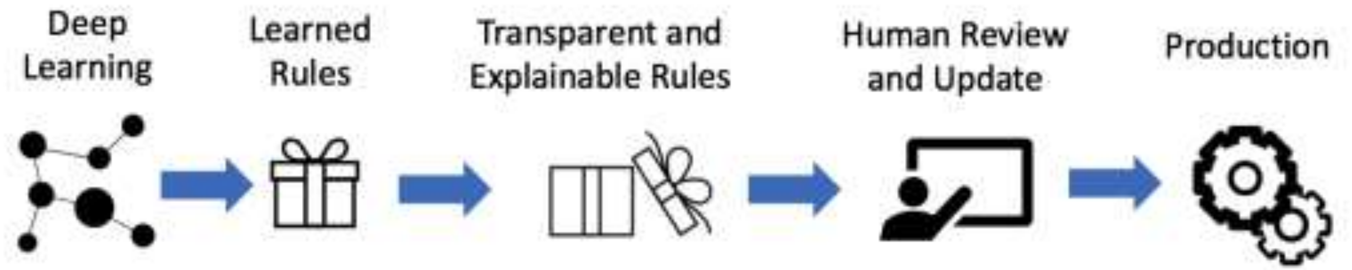
Explanation in service of business actionability

“Is the feature it is pointing to something I can change in my business? If not, how does knowing that help me? (I-22, AI Engineer)

AI Lifecycle Touchpoints	Initial Model building	Model validation during proof-of-concept	Model in-production
Audience (Whom does the AI model interface with)	Model developers	Data Scientists Product Mangers Domain experts Business owners Business IT Operations	Model developers Data Scientists | Technical Strategists Product managers | Design teams Business owners/users Business IT Operations
Explainability Motivations (Information needs)	- Peeking inside models to understand their inner workings - Improving model design (e.g., how should the model be retrained, re-tuned) - Selecting the right model	- Characteristics of data (proprietary, public, training data) - Understanding model design - Ensuring ethical model development	- Expectation mismatch - Augmenting business workflow and business actionability

Explainability in HEIDL

HEIDL (Human-in-the loop linguistic Expressions wIth Deep Learning)



HEIDL video

Explainability takes on two dimensions here:

models are fully explainable

users involved In model co-creation with immediate feedback

Design Implications of AI Explainability

Balancing external stakeholders needs with their AI knowledge

"We have to balance the information you are sharing about the AI underpinnings, it can overwhelm the user (I-21, UX Designer)

Group collaboration persona describing distinct types of members

"Loss of control making the *support and maintenance of explainable models hard (I-8, Data Scientist)*

Simplicity versus complexity tradeoff

"The design space for (explaining) models to end users is in a way more constrained ... you have to assume that you have to put in a very, very very shallow learning curve (I-27 HCI Researcher)

Privacy

"Explanatory features can reveal identities (e.g., easily inferring employee, department, etc.) (I-24, HCI researcher)

Design Implications of AI Explainability

Balancing external stakeholders needs with their AI knowledge

“We have to balance the information you are sharing about the AI underpinnings, it can overwhelm the user (I-21, UX Designer)

Group collaboration persona describing distinct types of members

“Loss of control making the *support and maintenance of explainable models hard (I-8, Data Scientist)*

Simplicity versus complexity tradeoff

“The design space for (explaining) models to end users is in a way more constrained ... you have to assume that you have to put in a very, very very shallow learning curve (I-27 HCI Researcher)

Privacy

“Explanatory features can reveal identities (e.g., easily inferring employee, department, etc.) (I-24, HCI researcher)

Design Implications of AI Explainability

Balancing external stakeholders needs with their AI knowledge

"We have to balance the information you are sharing about the AI underpinnings, it can overwhelm the user (I-21, UX Designer)

Group collaboration persona describing distinct types of members

"Loss of control making the *support and maintenance of explainable models hard (I-8, Data Scientist)*

Simplicity versus complexity tradeoff

"The design space for (explaining) models to end users is in a way more constrained ... you have to assume that you have to put in a very, very very shallow learning curve (I-27 HCI Researcher)

Privacy

"Explanatory features can reveal identities (e.g., easily inferring employee, department, etc.) (I-24, HCI researcher)

Design Implications of AI Explainability

Balancing external stakeholders needs with their AI knowledge

“We have to balance the information you are sharing about the AI underpinnings, it can overwhelm the user (I-21, UX Designer)

Group collaboration persona describing distinct types of members

“Loss of control making the *support and maintenance of explainable models hard (I-8, Data Scientist)*

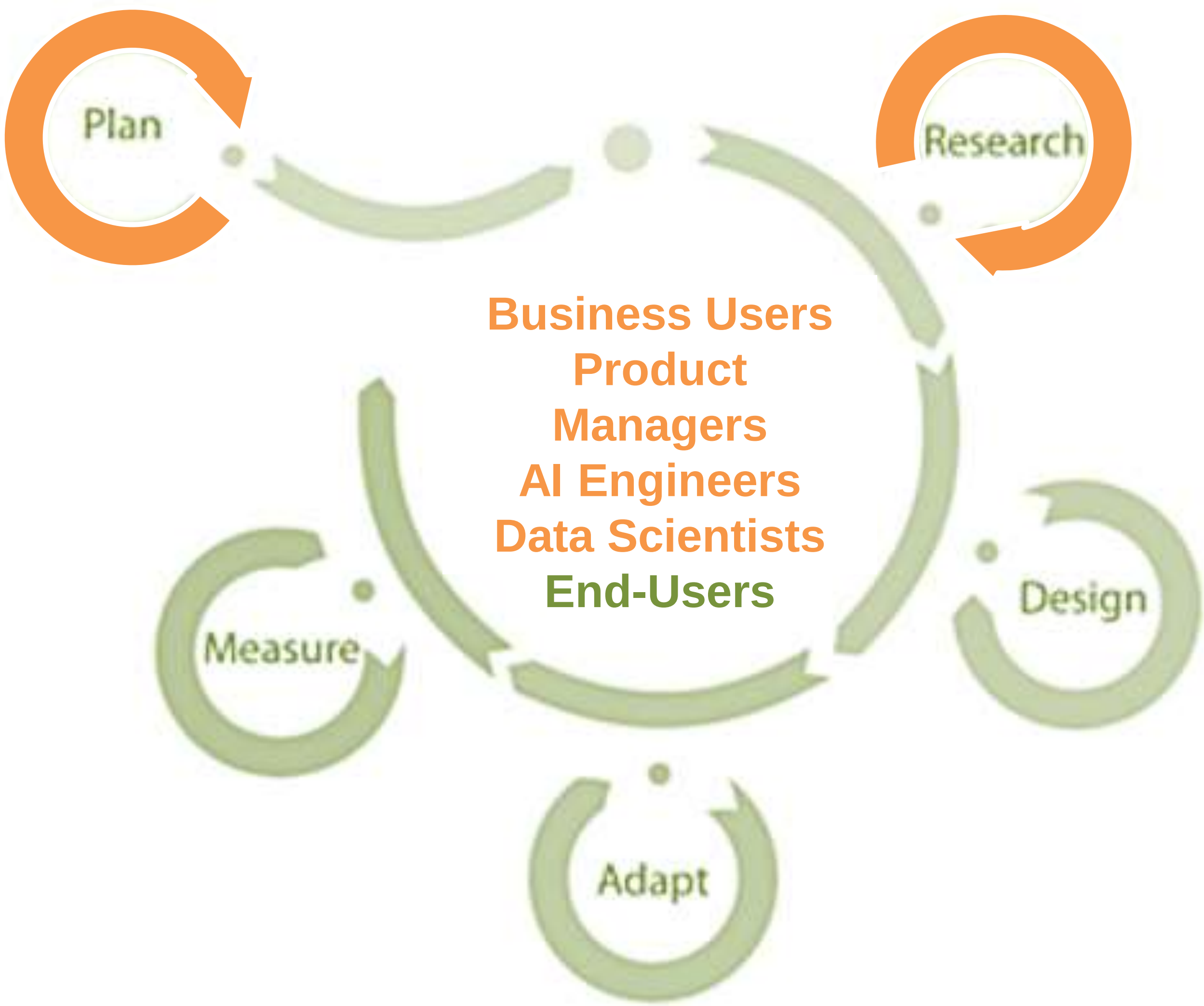
Simplicity versus complexity tradeoff

“The design space for (explaining) models to end users is in a way more constrained ... you have to assume that you have to put in a very, very very shallow learning curve (I-27 HCI Researcher)

Privacy

“Explanatory features can reveal identities (e.g., easily inferring employee, department, etc.) (I-24, HCI researcher)

Future Work - User-Centered Design of AI Explainability



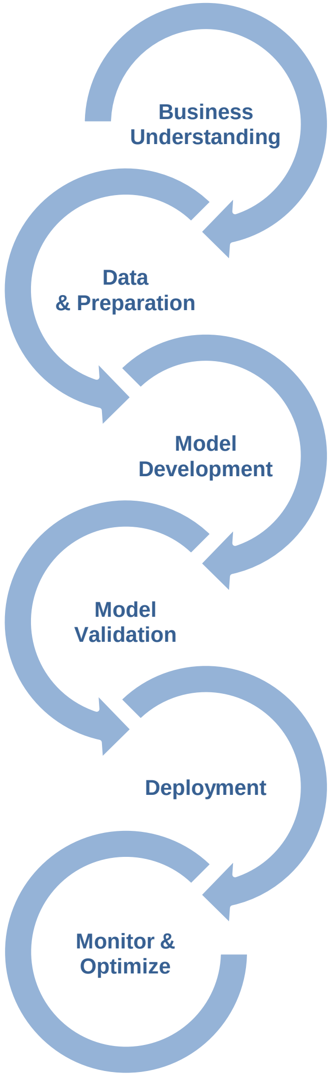


Image source: https://www.slideshare.net/agaszostek/history-and-future-of-human-computer-interaction-hci-and-interaction-design

Future Work - User-Centered Design of AI Explainability

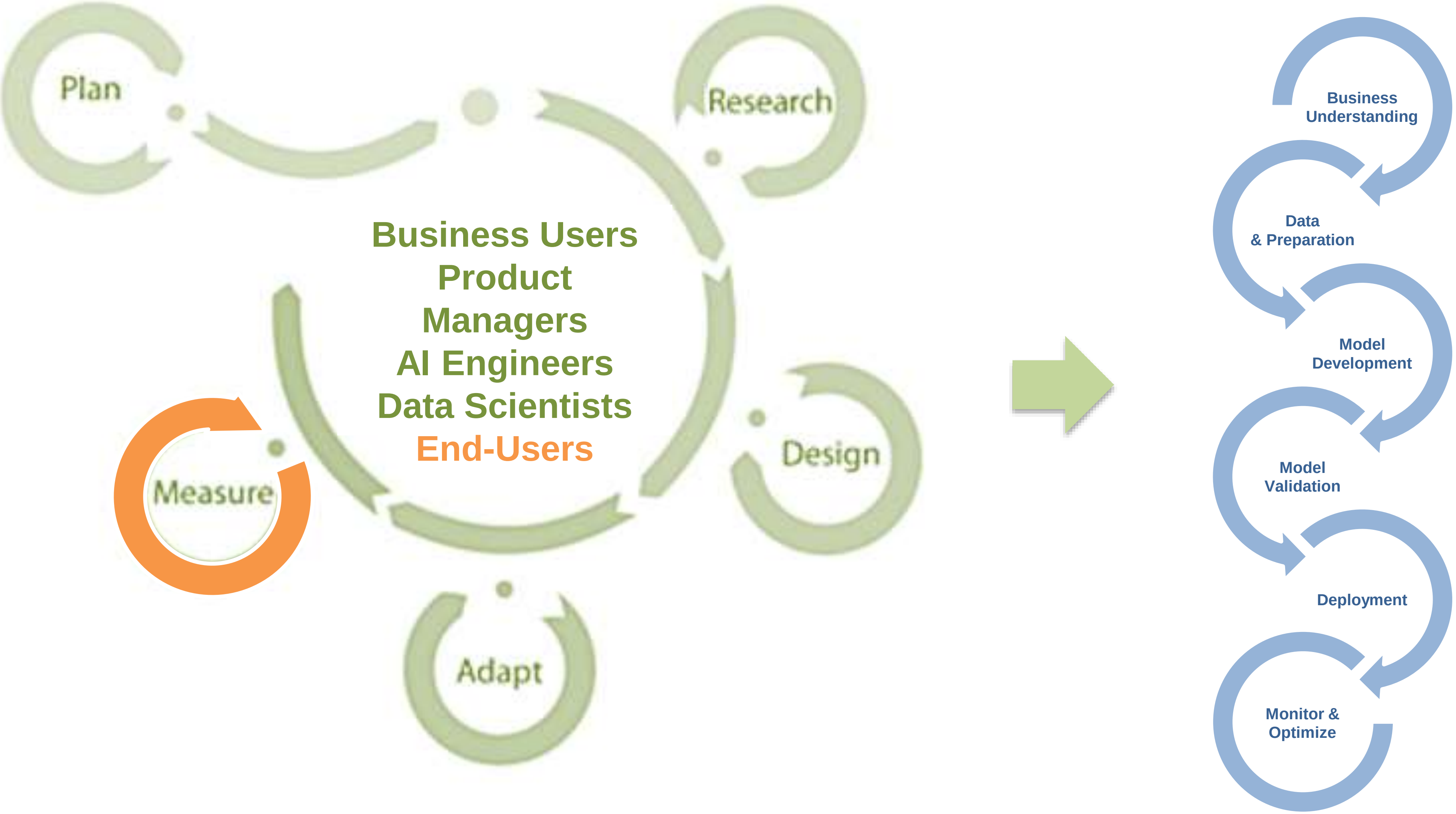
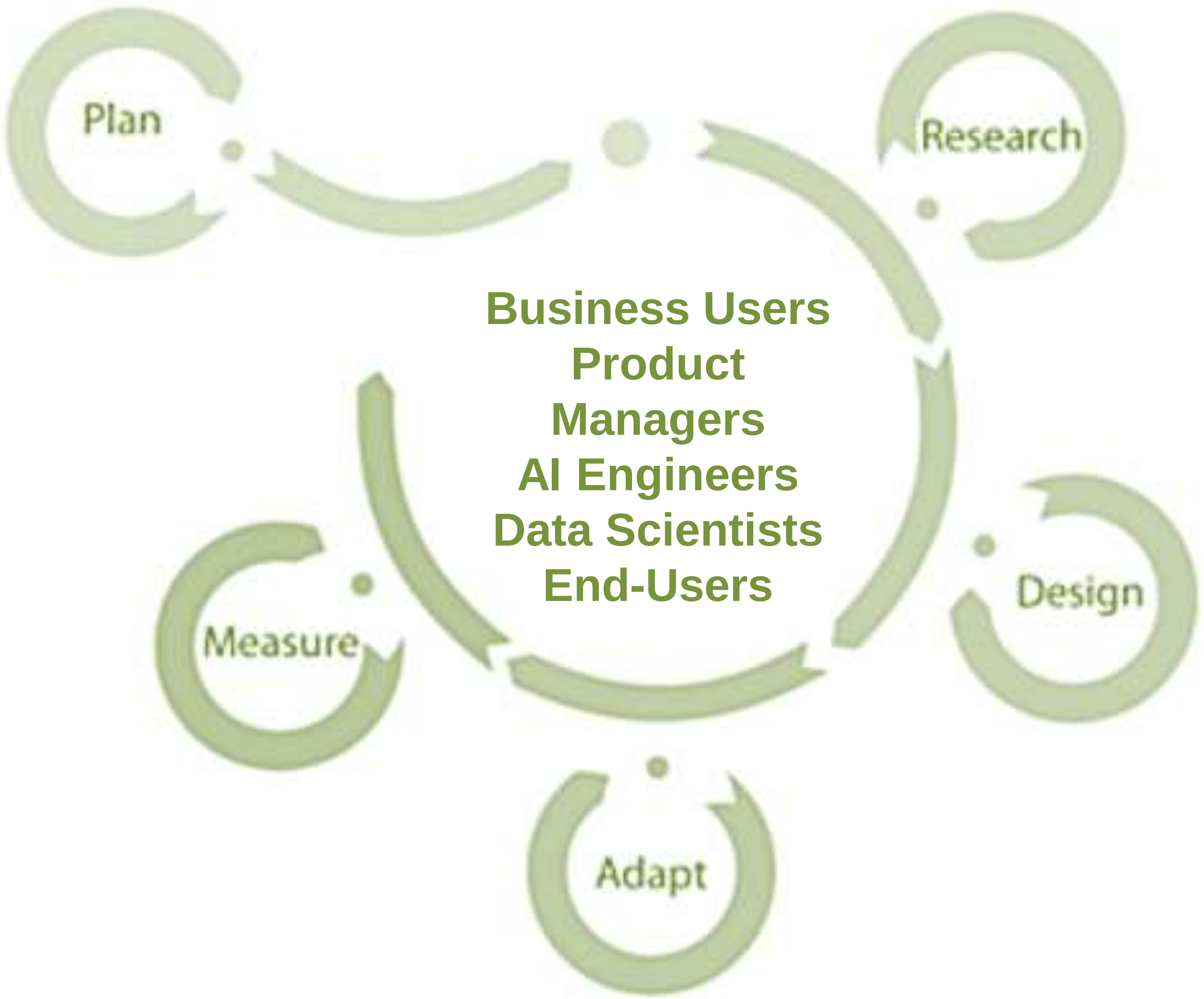


Image source: https://www.slideshare.net/agaszostek/history-and-future-of-human-computer-interaction-hci-and-interaction-design

Future Work - User-Centered Design of AI Explainability



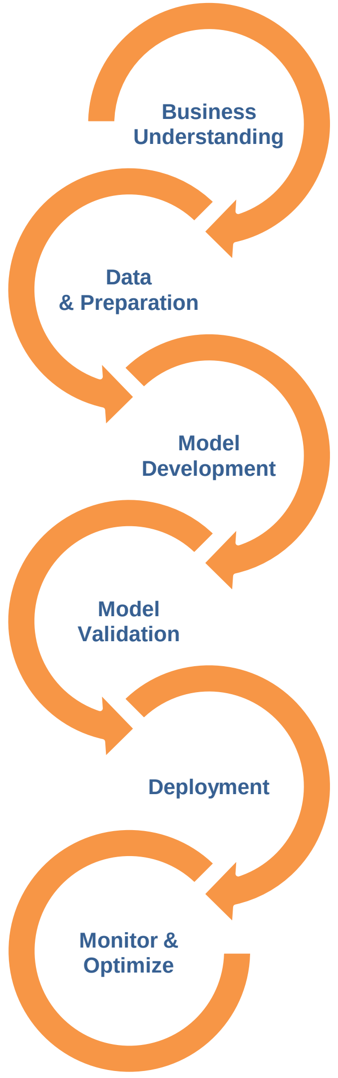


Image source: https://www.slideshare.net/agaszostek/history-and-future-of-human-computer-interaction-hci-and-interaction-design

PART IV – Open Challenges & Concluding Remarks

Explainability for NLP is an Emerging Field

- AI has become an important part of our daily lives
- Explainability is becoming increasingly important
- Explainability for NLP is still at its early stage

Challenge 1: Standardized Terminology

Comprehensibility ...

Understandability Gray-box

Interpretability transparency Directly interpretable

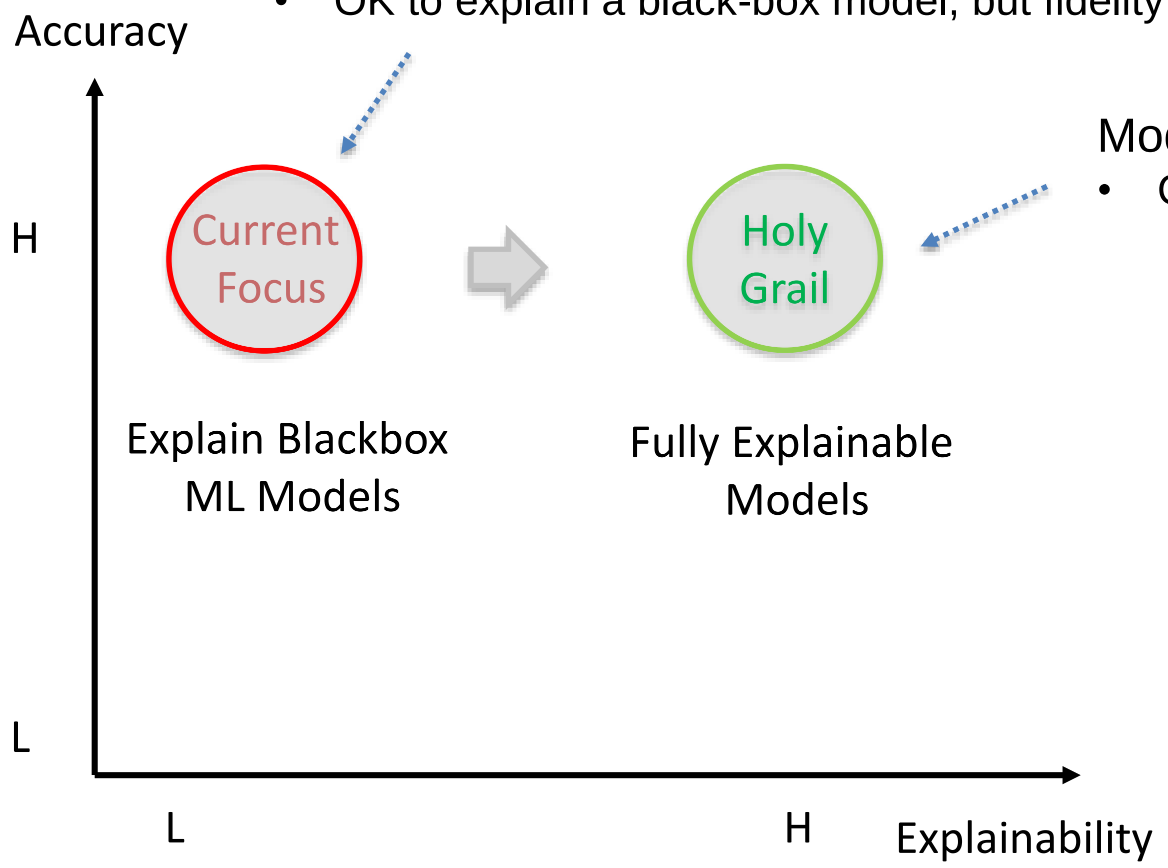
explainability Self-explaining White-box

... Black-box-ness ...

Challenge 2: Explainability AND Accuracy

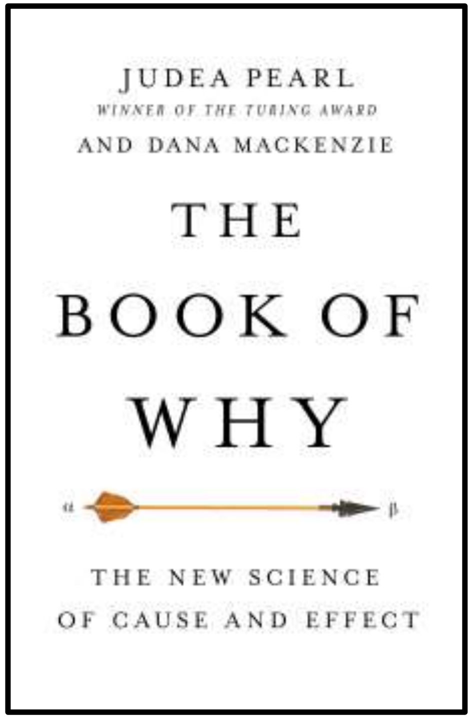
Surrogate models:

- OK to explain a black-box model, but fidelity suffers



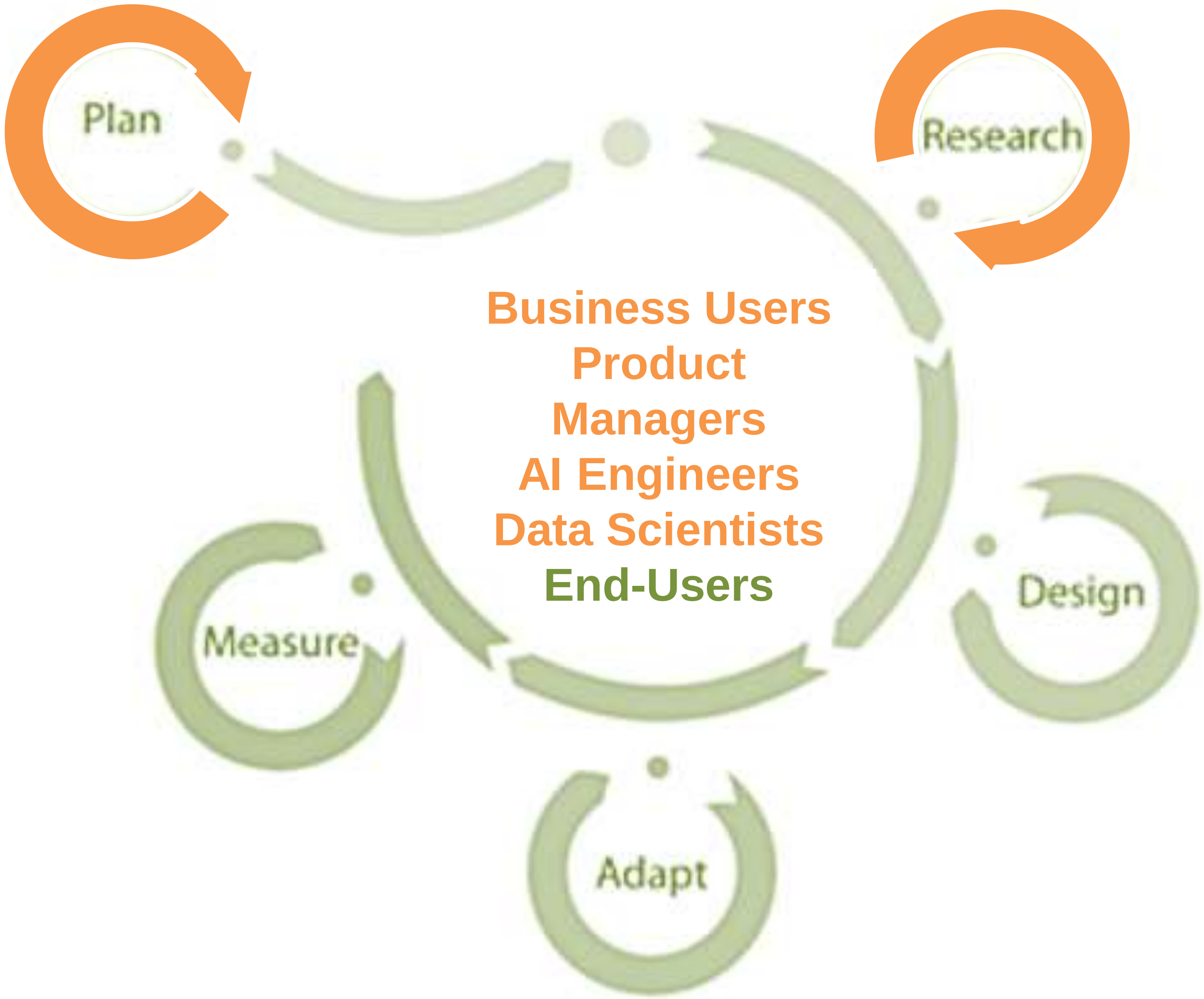
Model design geared towards explainability

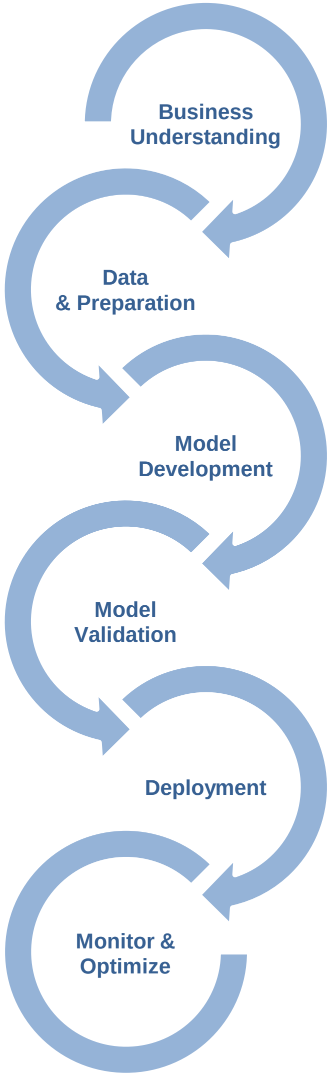
- Can be both explainable and high-accuracy!



Source: *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature. Machine Intelligence.

Challenge 3: Stakeholder-Oriented Explainability





Challenge 3: Stakeholder-Oriented Explainability

Different stakeholders have different expectations and skills

Statement of Work

Company A
Company B
…
Time is of essence of this Agreement. ….

```
label = Risky
Contains($sentence, $time) = False
```

AI engineer

Statement of Work

Company A
Company B
…
Time is of essence of this Agreement. ….
…

Risky. Need to define time frame to make this clause meaningful.

Potential risk: If no time specified, courts will apply a "reasonable" time to perform. What is a "reasonable time" is a question of fact - so it will up to a jury.

legal professional

Challenge 4: Evaluation Methodology & Metrics

- Need more well-established evaluation methodologies

- Need more fine-grain evaluation metrics
  - Coverage of the prediction
  - Fidelity

- Take into account the stakeholders as well!

Path Forward

- Explainability research for NLP is an interdisciplinary topic
  - Computational Linguistics
  - Machine learning
  - Human computer interaction

- NLP community and HCI community can work closer in the future
  - Explainability research is essentially user-oriented
  - but most of the papers we reviewed did not have any form of user study
  - Get the low-hanging fruit from the HCI community

This tutorial is on a hot area in NLP and takes an innovative approach ***by weaving in not only core technical research but also user-oriented research****. I believe this tutorial has the potential to draw a reasonably sized crowd.*
– anonymous tutorial proposal reviewer

"Furthermore, ***I really appreciate the range of perspectives (social sciences, HCI, and NLP)*** *represented among the presenters of this tutorial."*
– anonymous tutorial proposal reviewer

NLP people appreciate the idea of bringing NLP and HCI people together

Thank You

- Attend our other presentation at AACL 2020:
  "A Survey of the State of Explainable AI for Natural Language Processing"
  Session 9A 23:30-23:40 UTC+8 December 5, 2020

- We are an NLP research team at IBM Research – Almaden
  - We work on NLP, Database, Machine Learning, Explainable AI, HCI, etc.
  - Interested in collaborations? Just visit our website!
    http://ibm.biz/scalableknowledgeintelligence

- Thank you very much for attending our tutorial!