Explainable AI

Dr Simone Stumpf

Reader in Responsible and Interactive AI

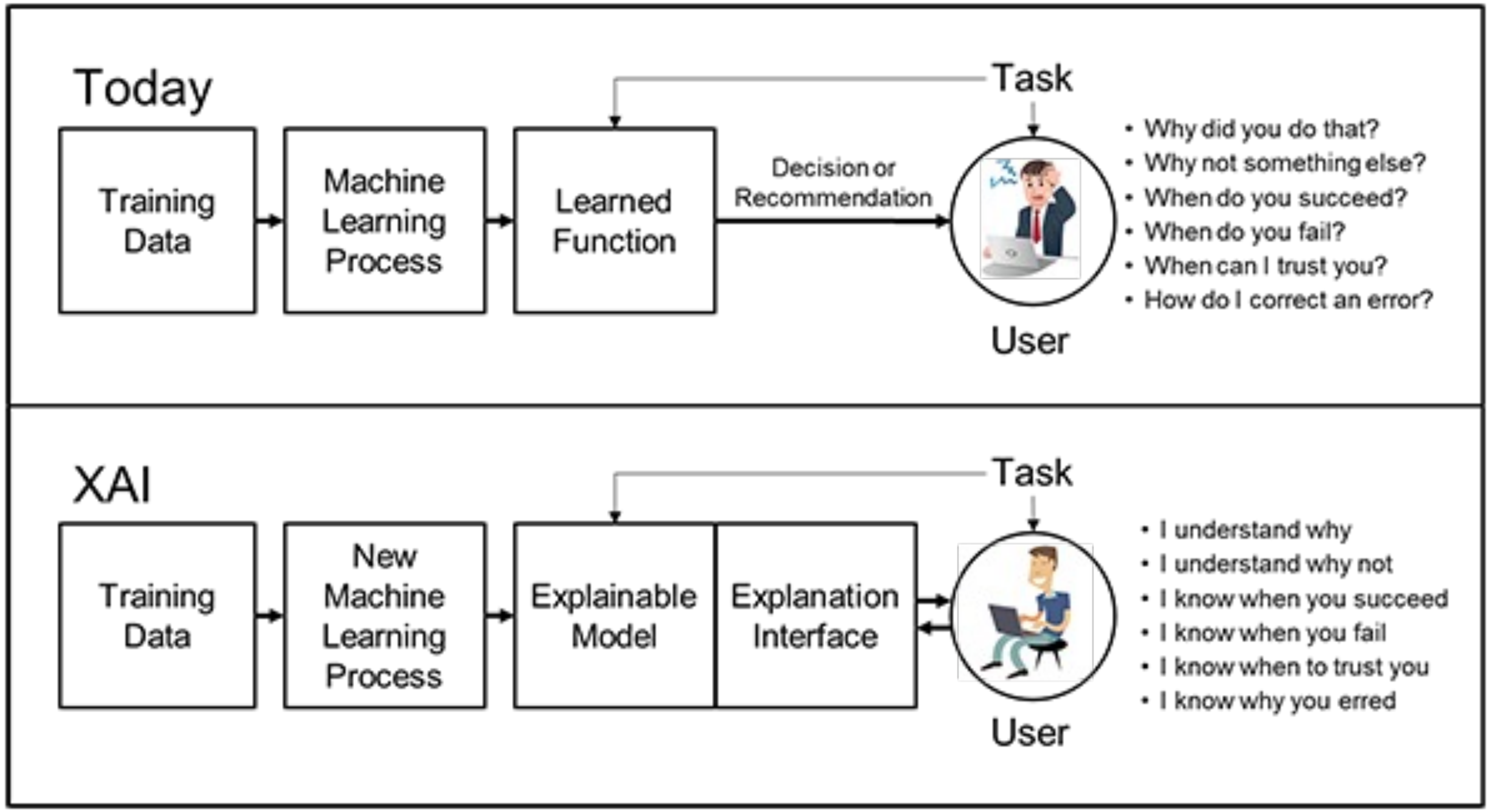
School of Computing Science

Simone.Stumpf@glasgow.ac.uk

@DrSimoneStumpf

DrSimoneStumpf@hci.social

Explainable AI (XAI) vision (2016)



Calibrated / appropriate trust

AI ≠ automation

XAI's roots

(Way back in humanities & social sciences)

- 1970s/1980s: Expert system explanations
- 1990s/2000s: Growth of machine learning
- 2016: DARPA XAI programme

Mycin: Is Meningitis a hospital acquired illness?
Doctor: why?
Mycin: [2.0] this will aid in determining the orgamisms that might be causing the infection
It has already established
[2.1] the infection that requires therapy is meningitis
[2.2] organisms were not seen on the stain of this pending csf culture
[2.3] the type of infection is bacterial
[2.4] neurosurgery is not one of the relevant items of patient 100
[2.5] the age of patient 100 is greater than 25
therefore if [2.6] the infection was acquired while the patient was in hospital then there is evidence that the organisms that might be causing the infection are E coli(0.75) staphylococcus-coag-pos(0.3) pseudomonas-aeruginosa(0.3) klebsiella-pneumoniae(0.5)
Doctor: How 2.3
Mycin: The following rules concluded about the likelihood that the type of the infection is bacterial:
[2.7] RULE148 (0.19)
[2.8] RULE500 (0.51)
[2.9] RULE501 (0.95)
[3.0] RULE502 (0.97)
[3.1] RULE526 (0.98)
[3.2] RULE504 (0.97)
[3.3] RULE524 (0.98)

Motivation

Model understanding is absolutely critical in several domains -- particularly those involving *high stakes decisions*!

Motivation: Why Model Understanding?

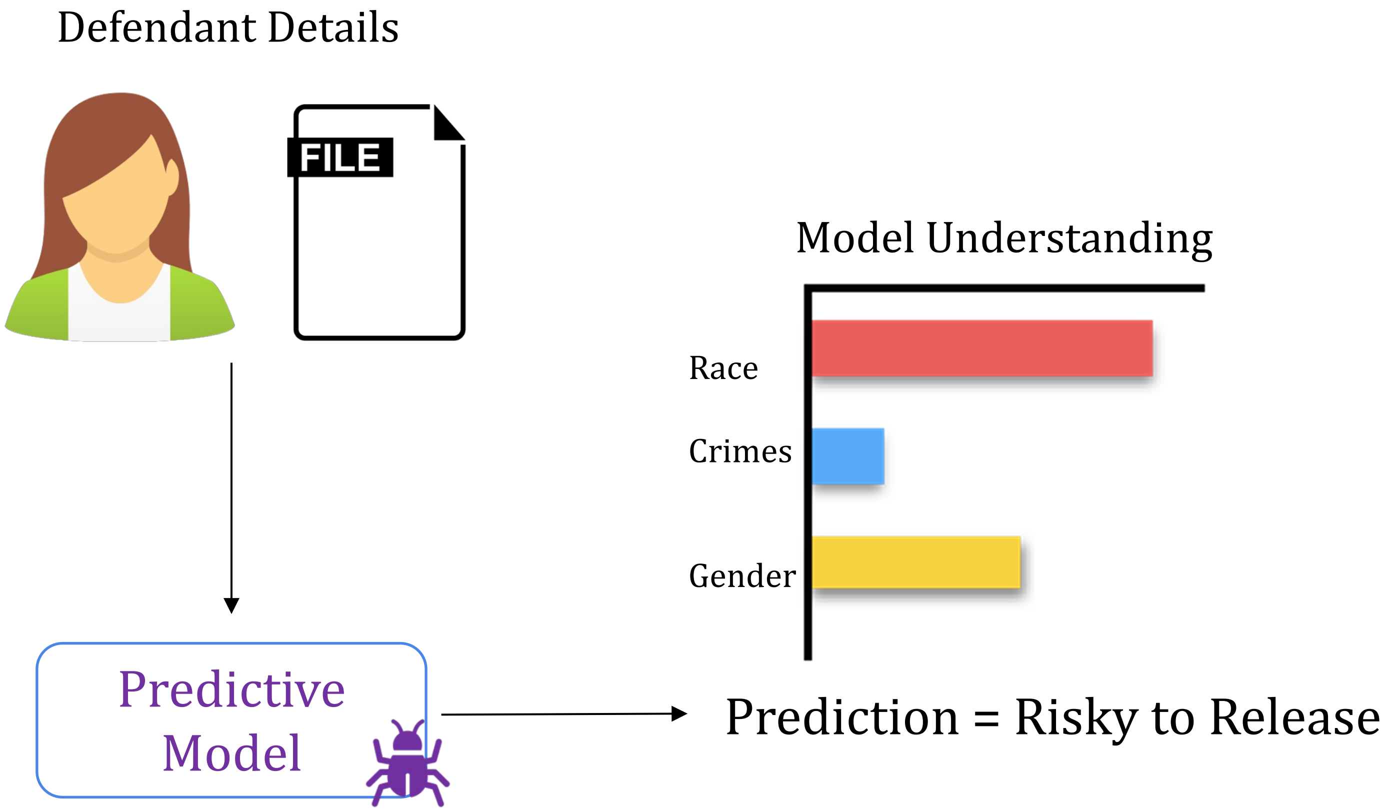
Input

Predictive Model

Prediction = Siberian Husky

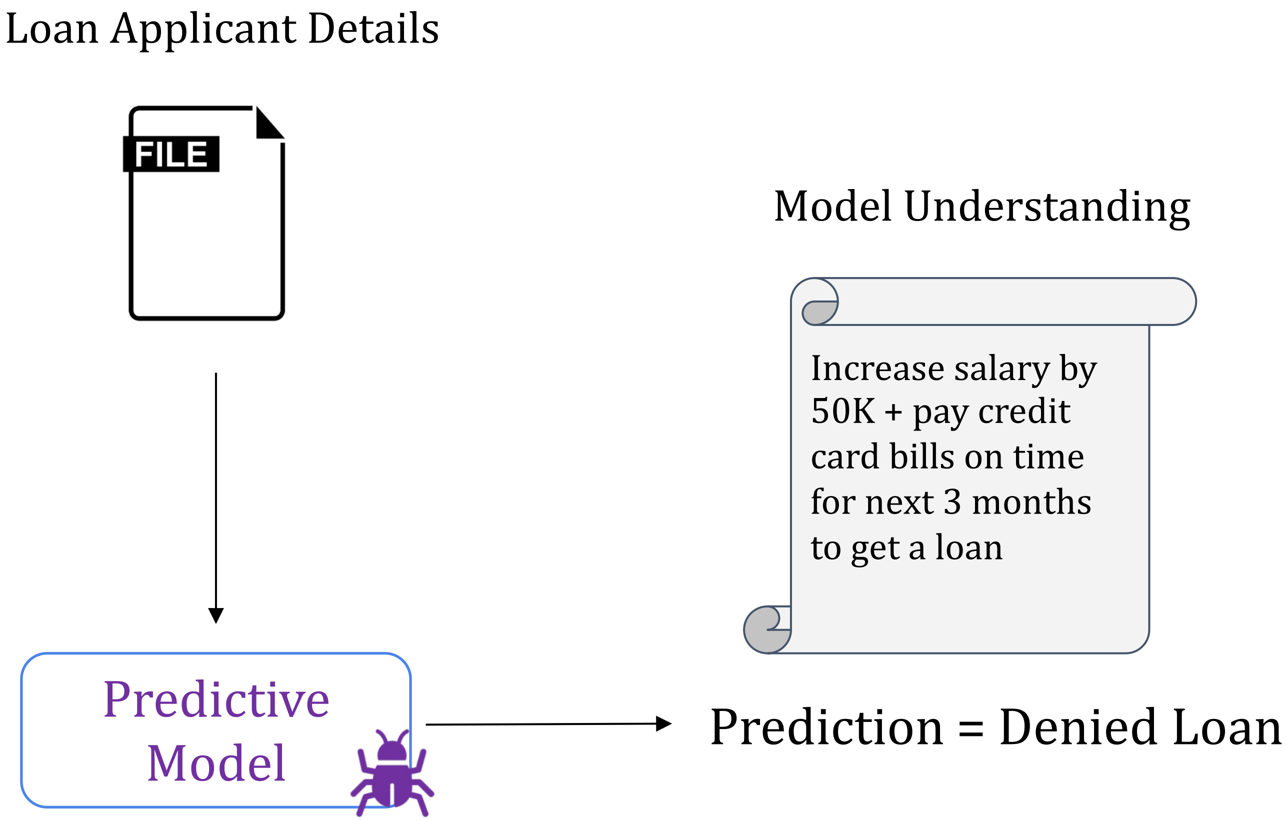
Model understanding facilitates debugging

Motivation: Why Model Understanding?



Model understanding facilitates bias detection

Motivation: Why Model Understanding?



Model understanding helps provide recourse to individuals who are affected by model predictions

Explainability versus Interpretability

- *Explainability* = ability of an AI system to explain itself
- *Interpretability (or intelligibility)* = ability of a user to build an appropriate mental model that guides interaction with the AI system
  - Understanding of how the system works
  - Being able to use the system successfully
  - Being able to ’trouble-shoot’ system and fix ‘mistakes’
- Are some algorithms (e.g. decision trees) naturally interpretable?

Mental Models

- A mental model is kind of internal representation in someone's thought process for how something works in the real world

- Based on meaning, understanding and experience

- Users build mental models to guide how they interact, behave or fix things when they go wrong

[Norman 1983, Johnson-Laird 1983]

Different stakeholders = different explanations?

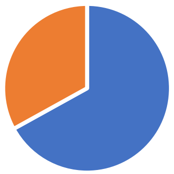
- End users / lay users (e.g. loan applicants)
- Decision makers / domain experts (e.g. doctors, judges)
- Regulatory agencies (e.g. FDA, European commission)
- Researchers, developers and engineers

Explanation content versus explanation presentation/style

- Keep apart what information is transmitted in an explanation versus its form and presentation
- E.g. based on a model it is often possible to extract the probability that the AI is assigning to a prediction of belonging to a certain class i.e. its decision confidence

0.67341

67% Accept / 33% Reject



I think it’s a little bit more likely that this application should be accepted.

Intelligibility types

[Lim and Dey CHI 2009]

- What did the system do?
- Why did the system do W?
- Why did the system not do X?
- What would the system do if Y happens?
- How can I get the system to do Z, given the current context?

Lots of work to make ML transparent

[Molnar 2022]

- Simplest: I give you the source code of the model

- Next simplest: I give you a representation of the model

  - Exposing the model (global explanation)
  - Exposing (combination of) features that contribute to a decision (local explanation)

Explanatory debugging principles

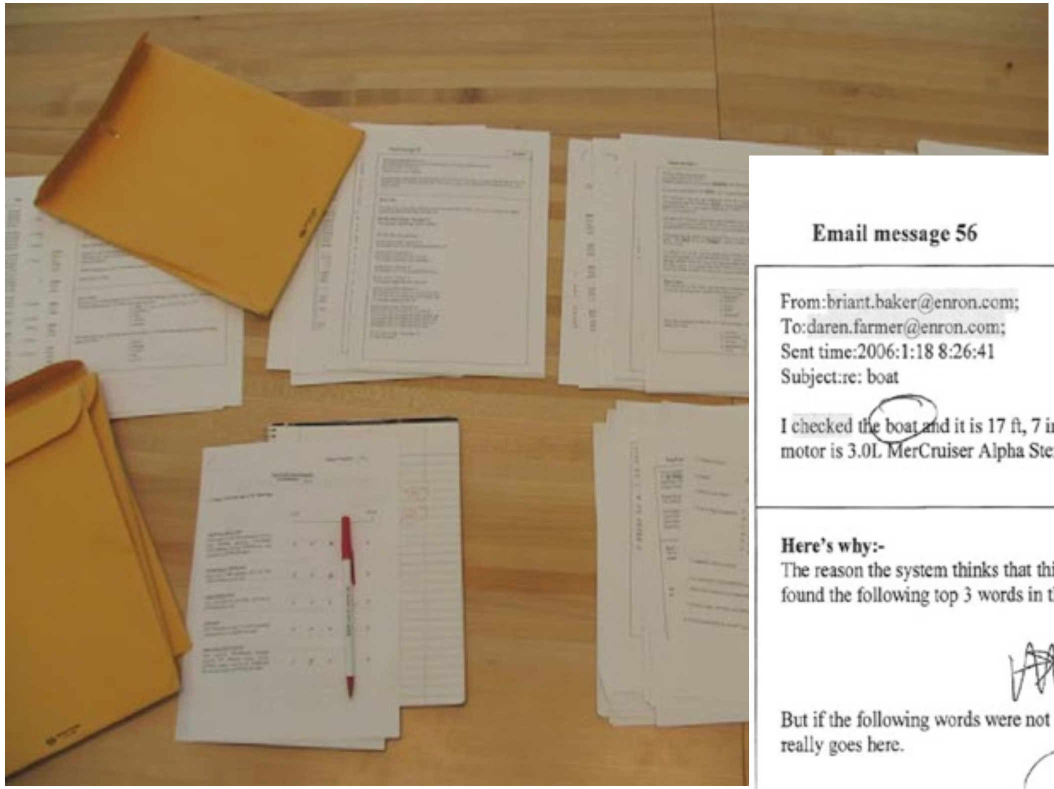
[Kulesza et al. IUI 2015]

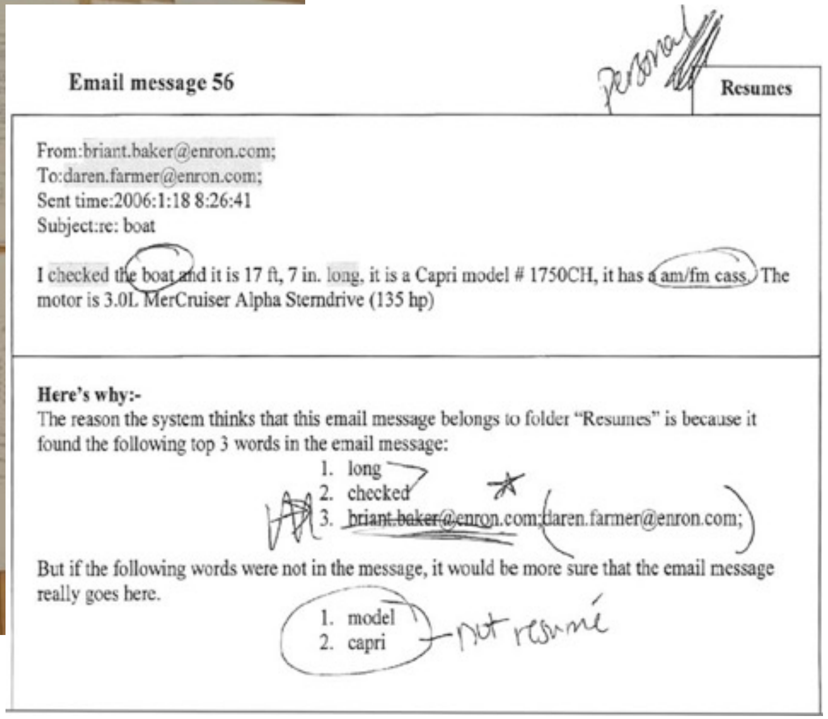
- Explanations should be
  - Iterative
  - Sound = Faithful
  - Complete
  - Don’t overwhelm

Explanation styles and feedback

- What explanation styles do end-users prefer?

[Stumpf et al. IJHCS 2009]





Email message 56 Personal Resumes

From:briant.baker@enron.com;
To:daren.farmer@enron.com;
Sent time:2006:1:18 8:26:41
Subject:re: boat

I checked the boat and it is 17 ft, 7 in. long, it is a Capri model # 1750CH, it has a am/fm cass. The motor is 3.0L MerCruiser Alpha Sterndrive (135 hp)

Here's why:-
The reason the system thinks that this email message belongs to folder "Resumes" is because it found the following top 3 words in the email message:

1. long
2. checked
3. ~~briant.baker@enron.com~~;daren.farmer@enron.com;

But if the following words were not in the message, it would be more sure that the email message really goes here.

1. model
2. capri

not resumé

Explanation styles

Keyword

Personal

From: buylow@houston.rr.com
To: j..farmer@enron.com
Subject: life in general

Good god -- where do you find time for all of that? You should w...

By the way, what is your new address? I may want to come by ... your work sounds better than anything on TV.

You will make a good trader. Good relationships and flexible pri... a few zillion other intangibles you will run into. It beats the hell o... other things.

I'll let you be for now, but do keep those stories coming we love...

The reason the system thinks that this email message belongs to folder "Personal" is because it found the following top 5 words in the email message:

1. ill
2. love
3. better
4. things
5. god

But if the following words were not in the message, it would be more sure the email message really goes here.

1. keep
2. find
3. trader
4. book
5. general

Rule

Resume

From: toni.graham@enron.com
To: daren.farmer@enron.com
Subject: re: job posting

Daren, is this position budgeted and who does it report to?
Thanks,
Toni Graham

The reason the system thinks that this email message belongs to folder "Resume" is because the highest priority rule that fits this email message was:

- **Put the email in folder "Resume" if:**
  It's from toni.graham@enron.com.

The other rules in the system are:

...

- Put the email in folder "Personal" if:
  The message does not contain the word "Enron" and
  The message does not contain the word "process" and
  The message does not contain the word "term" and
  The message does not contain the word "link".
- Put the email in folder "Enron News" if:
  No other rule applies.

Similarity

Resume

Message #2
From: 40enron@enron.com
To: All ENW employees
Subject:enron net works t&e policy
From: Greg Piper and Mark Pickering

Please print and become familiar with the updated ENW T&E P... business-first travel, with supervisor approval, for international fli... Mexico). Supervisors will be responsible for making the decision...

If you have any questions about the policy or an expense not co... Costello.

Wow! The message is really similar to the message #3 in "**Resume**" because #2 and #3 have important words in common.

Message #3
From: toni.graham@enron.com
To: lisa.csikos@enron.com, rita.wynne@enron.com, daren.farmer@enron.com
CC: renda.herod@enron.com
Subject: confirming requisitions

Confirming the open requisitions for your group. If your records indicate otherwise, please let me know.

Lisa Csikos 104355, 104001
Rita Wynne 104354
Daren Farmer 104210
Mike Eiben 104323
Pat Clynes 104285

The posting dates have all been updated to reflect a current posting date.

Results

- Explanation styles:
  - Rule-based best understood
  - Keyword-based also good but negative weights problematic (absence of features)
  - Serious understandability problems with Similarity-based
  - No clear overall preference, very individual

Local explanations

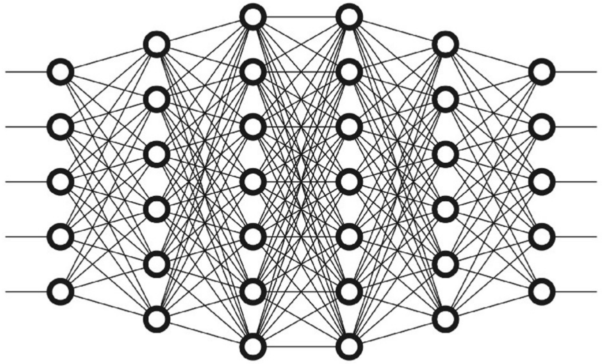
LIME: Local Interpretable Model-Agnostic Explanations

- Explains important feature that led to a decision
- Uses a post-hoc explanation on a simplified model
- Another popular method which outputs feature importances: SHAP

[Ribeiro et al. KDD 2016]

Saliency Maps

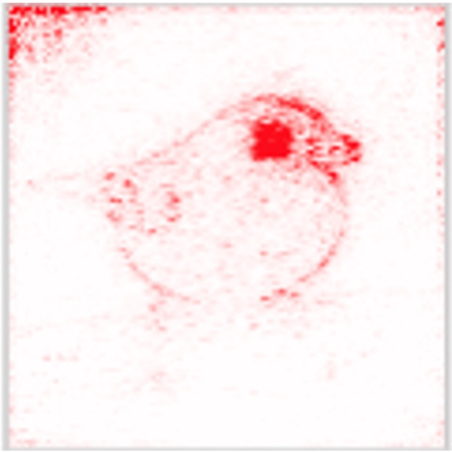
Input



Prediction

Junco Bird

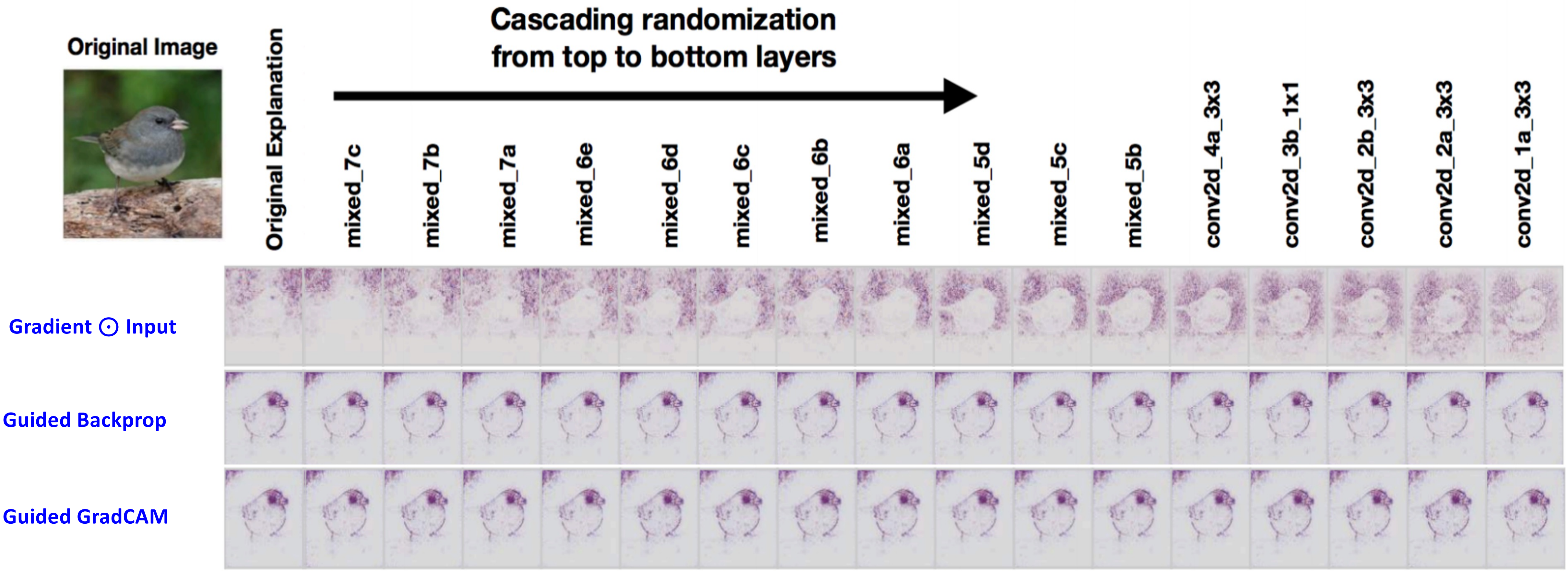
What parts of the input are most relevant for the model's prediction: 'Junco Bird'?



Saliency Map

But beware: “explanation” might be misleading

Model parameter randomization test



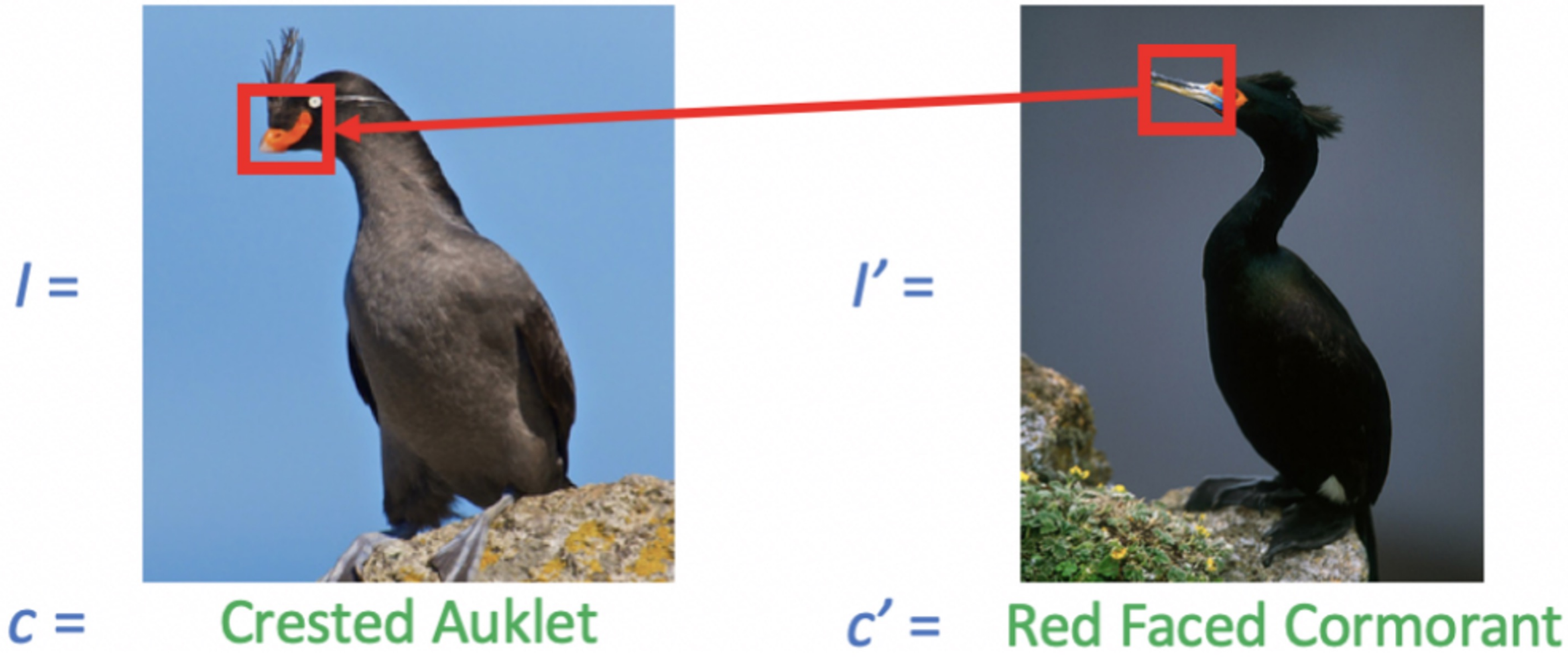
Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Prototypes/Example

- Use examples (synthetic or natural) to explain individual predictions
    - Identify instances in the training set that are responsible for the prediction of a given test instance
    - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest

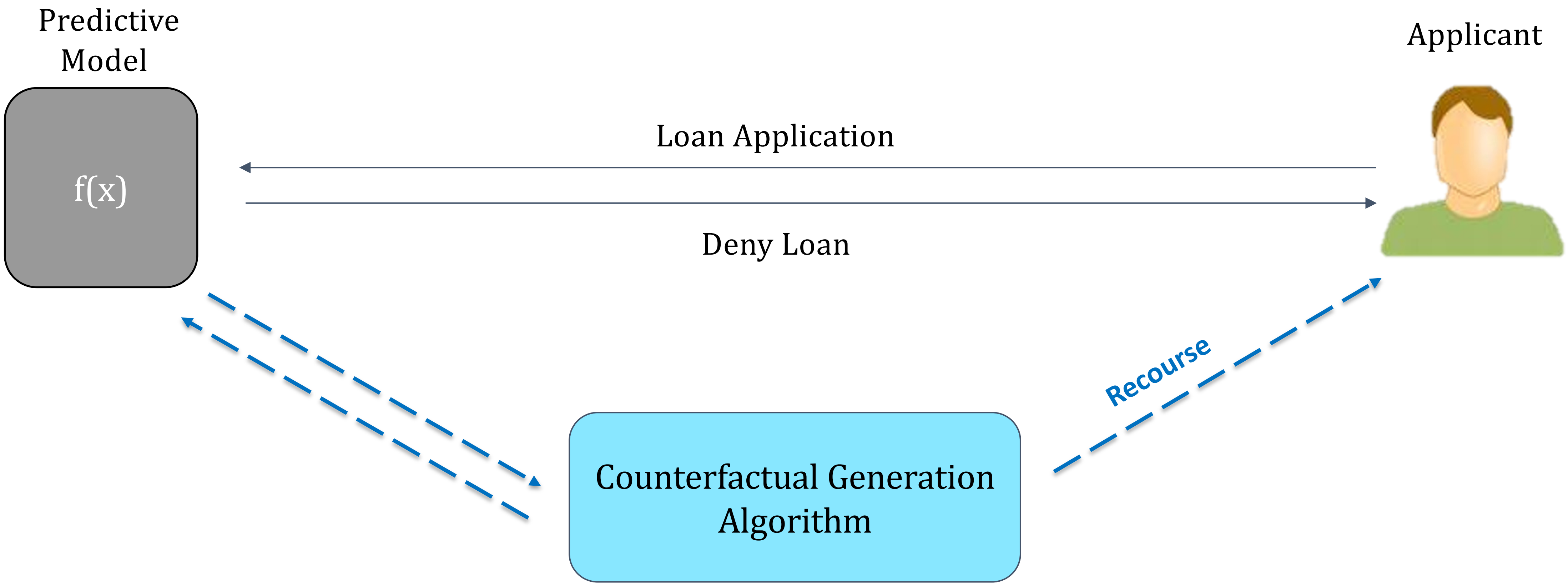
Counterfactual Explanations

What features need to be changed and by how much to flip a model's prediction?



[Mothilal et al 2020]

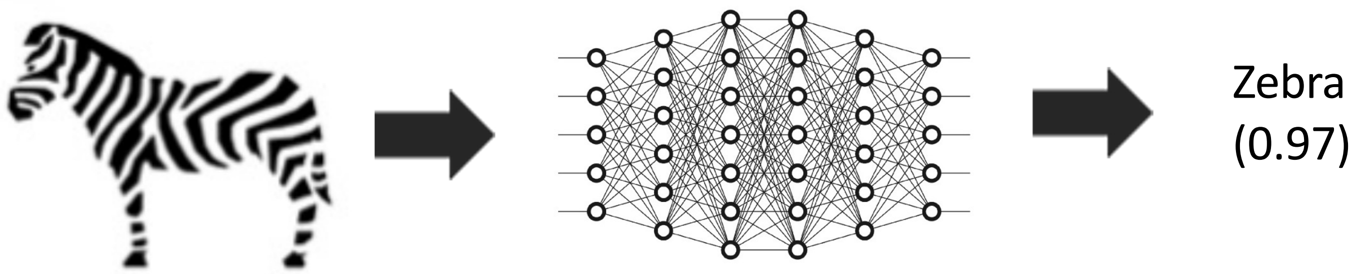
Counterfactual Explanations



Recourse: Increase your salary by 50K & pay your credit card bills on time for next 3 months

Global explanations

Representation Based Explanations



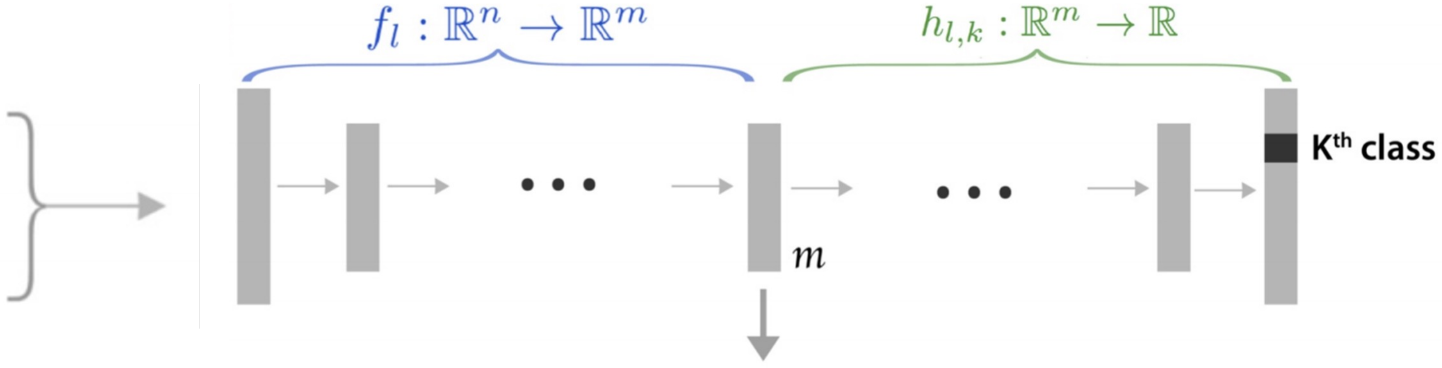
How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]

Representation Based Explanations: TCAV

Examples of the concept “stripes”

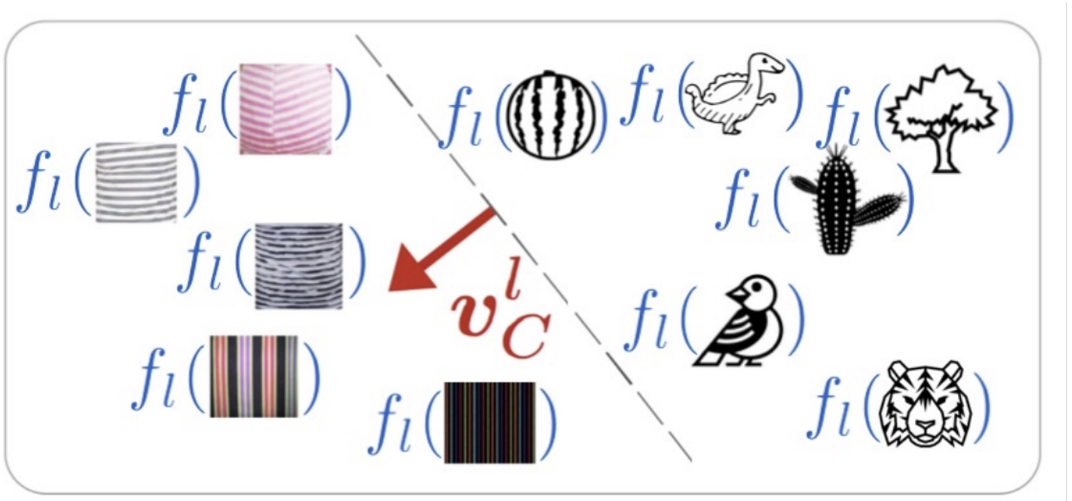
Random examples



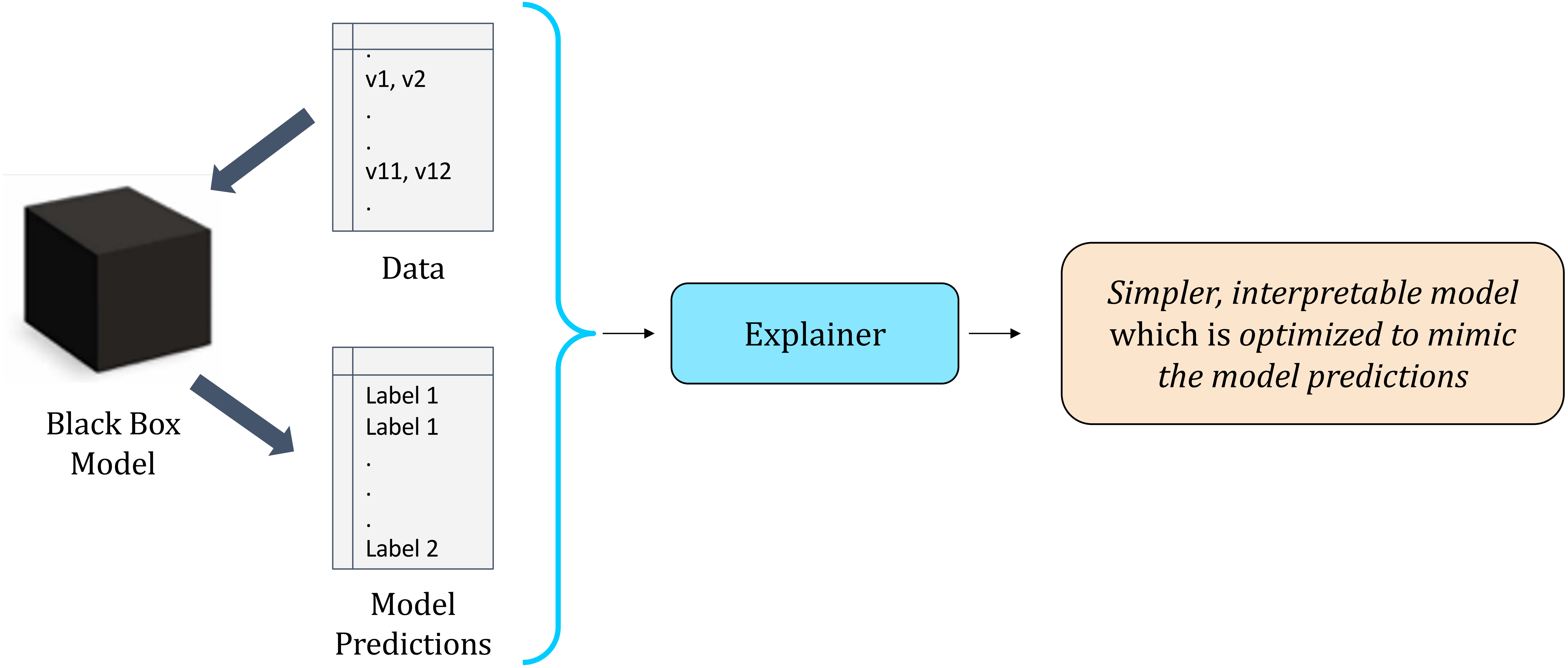
Train a linear classifier to separate activations

The vector orthogonal to the decision boundary denotes the concept “stripes”

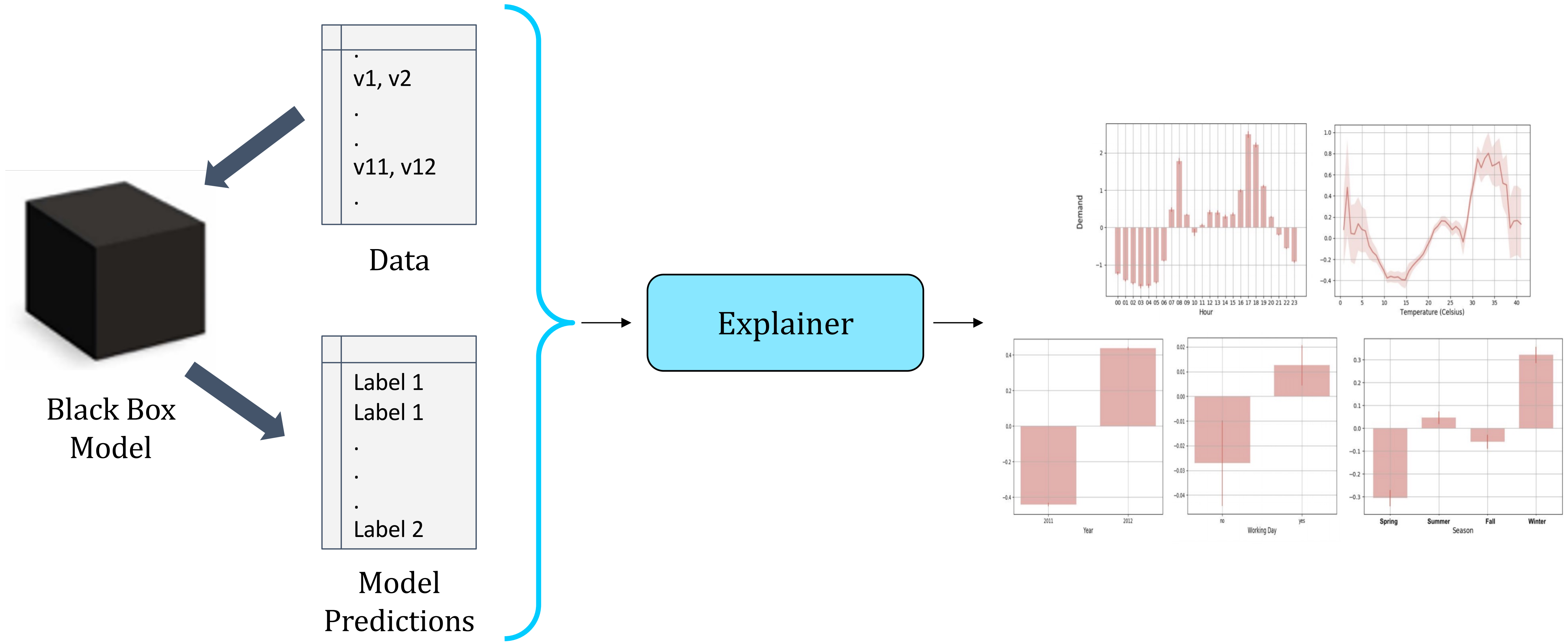
Compute gradient w.r.t. this vector to determine how important is the notion of stripes for a prediction



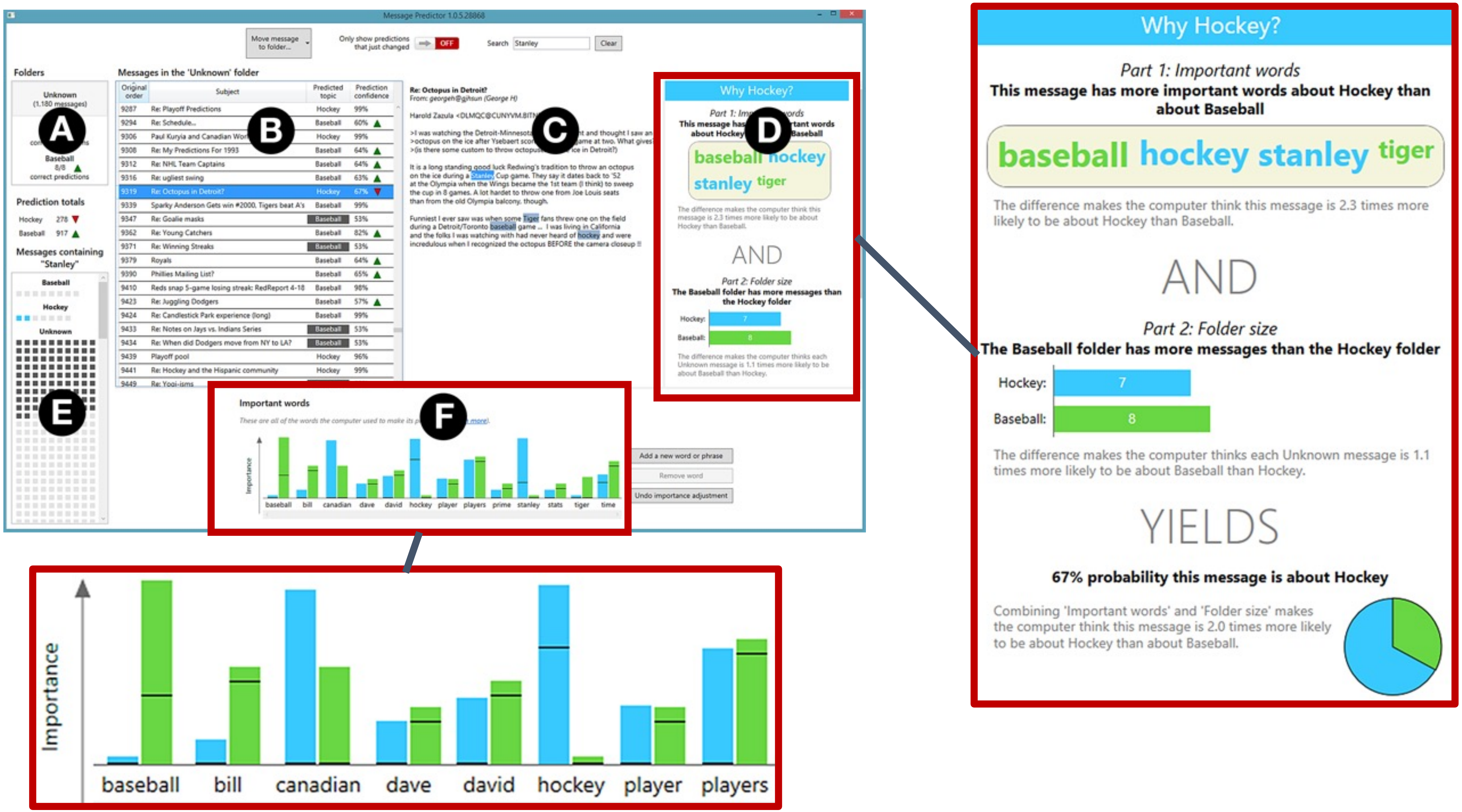
Model Distillation



Model Distillation Using Generalized Additive Models



Kulesza et al. IUI 2015



Study setup

- 77 participants split into two groups: 40 using EluciDebug, 37 using a version without explanations and advanced feedback
- 20 Newsgroup data set (Hockey and Baseball): initial system training on 5 messages for each subject, 1850 unlabeled messages to sort
- 30 minutes to “make the system as accurate as possible”
- Measures: accuracy, amount of feedback given, mental model scores, perceived workload
- Multinomial Naïve Bayes, retrained after every feedback

Results

- More accurate system accuracy with less effort
  - 85% for our system versus 77% without explanations at end of study
  - Made adjustments to 47 messages while without explanations had to label 182 messages

- With better understanding
  - 15.8 mental model score versus 10.4
  - The more you understand, the better you can make the system

- Do not overwhelm
  - No difference in workload measures

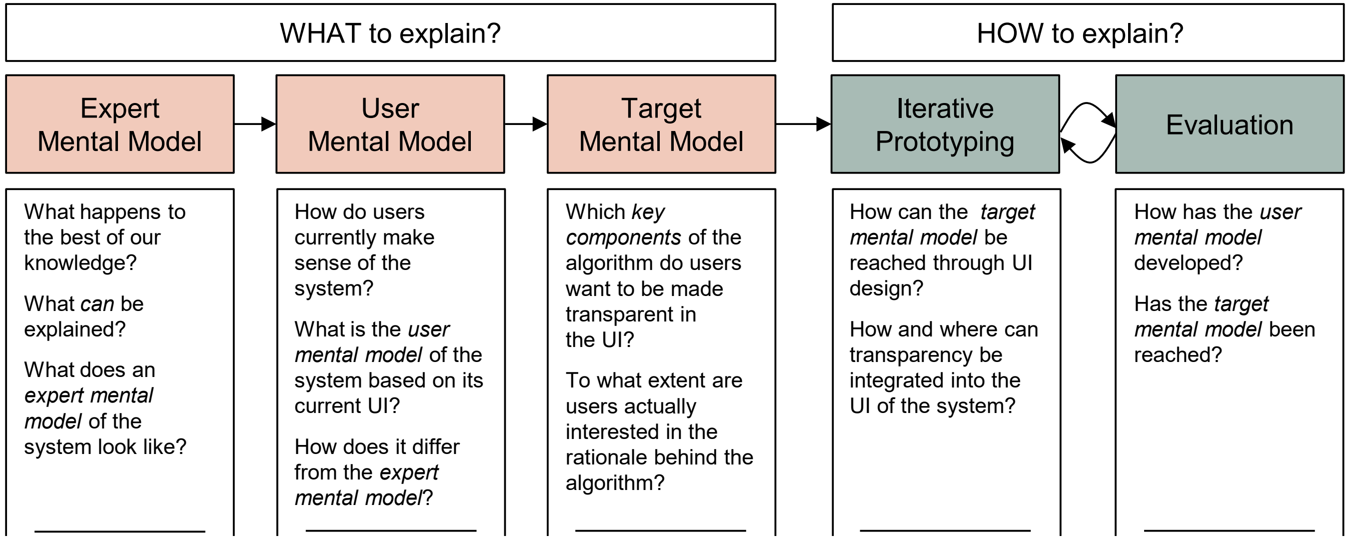
AI explanation design

- Need to know who the user is

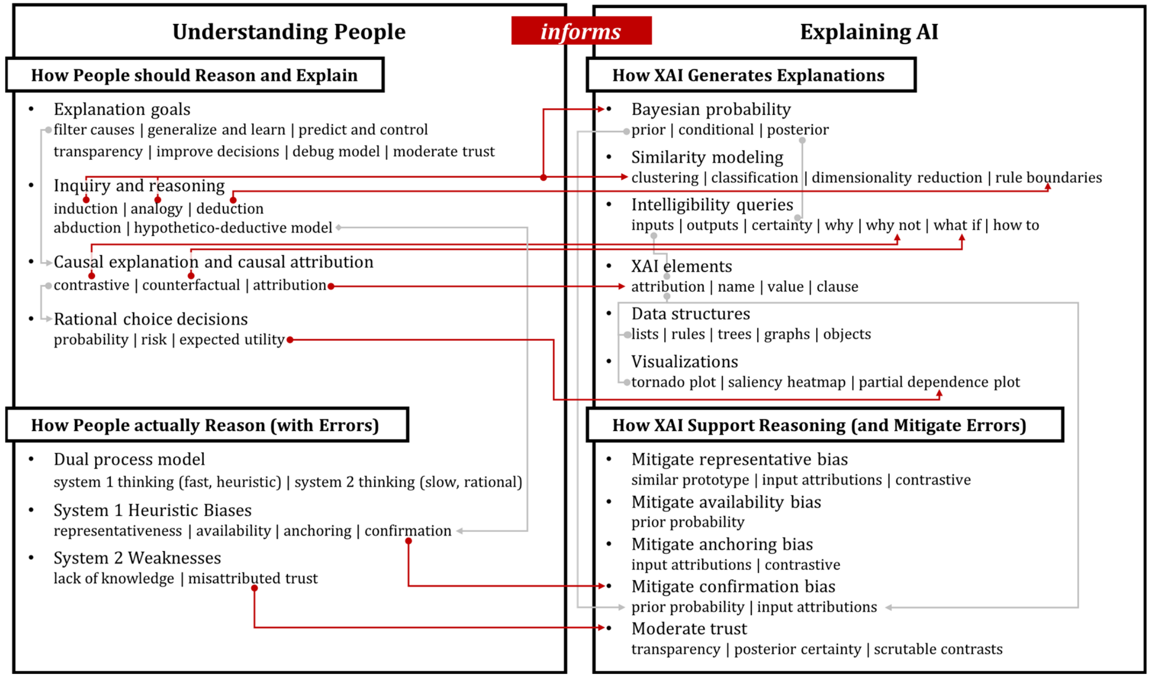
- Global or local explanations or both?

- Global explanations
  - How the model works
  - The accuracy of the model
  - Important features
- Local explanations
  - Important features for this decision
  - Decision confidence

Designing for Intelligibility



[Eiband et al. IUI 2018]



[Wang et al. CHI 2019]

- Essentially hand-crafted for each user group and each AI system

Challenges

- No explanations desired for certain tasks and contexts [Bunt et al. IUI 2012]
- Different people need different explanations [Gunning et al. Science Robotics 2019]
- “Placebic” explanations [Eiband et al. CHI 2019]
- Explanations calibrate trust and reliance [Bussone et al. ICMI 2015, Holliday et al. IUI 2016, Nourani et al. HCOMP 2019]
- Explanations might be outside of the ML [Ehsan et al. CHI 2021]

Summary

- Interpretability important for understanding how an AI system works
- Two different ways: global and local explanations
- Various approaches to provide these but
  - Local explanations = give features that make a difference to a specific prediction
  - Global explanations = show how model works overall
- Some challenges ahead in terms of providing the right explanations at the right time in the right way to whoever needs them

Resources

- Don Norman. 1983. Some observations on mental models. Lawrence Erlbaum Associates, Hillsdale,New Jersey,US.
- David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. Science Robotics 4, 37. https://doi.org/10.1126/scirobotics.aay7120
- Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. 2119–2128. https://doi.org/10.1145/1518701.1519023
- Christoph Molnar. Interpretable Machine Learning. Retrieved February 5, 2020 from https://christophm.github.io/interpretable-ml-book/
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI ’15), 126–137. https://doi.org/10.1145/2678025.2701399
- Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Hum.-Comput. Stud.* 67, 8: 639–662.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16), 1135–1144. https://doi.org/10.1145/2939672.2939778
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (Un)reliability of Saliency Methods. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen and Klaus-Robert Muller (eds.). Springer International Publishing, Cham, 267–280. https://doi.org/10.1007/978-3-030-28954-6_14
- Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT* ’20), 607–617. https://doi.org/10.1145/3351095.3372850
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In International Conference on Machine Learning, 2668–2677. Retrieved December 11, 2018 from http://proceedings.mlr.press/v80/kim18d.html
- Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’15), 1721–1730. https://doi.org/10.1145/2783258.2788613

- Thanks to Hima Lakkaraju and her tutorial on XAI!