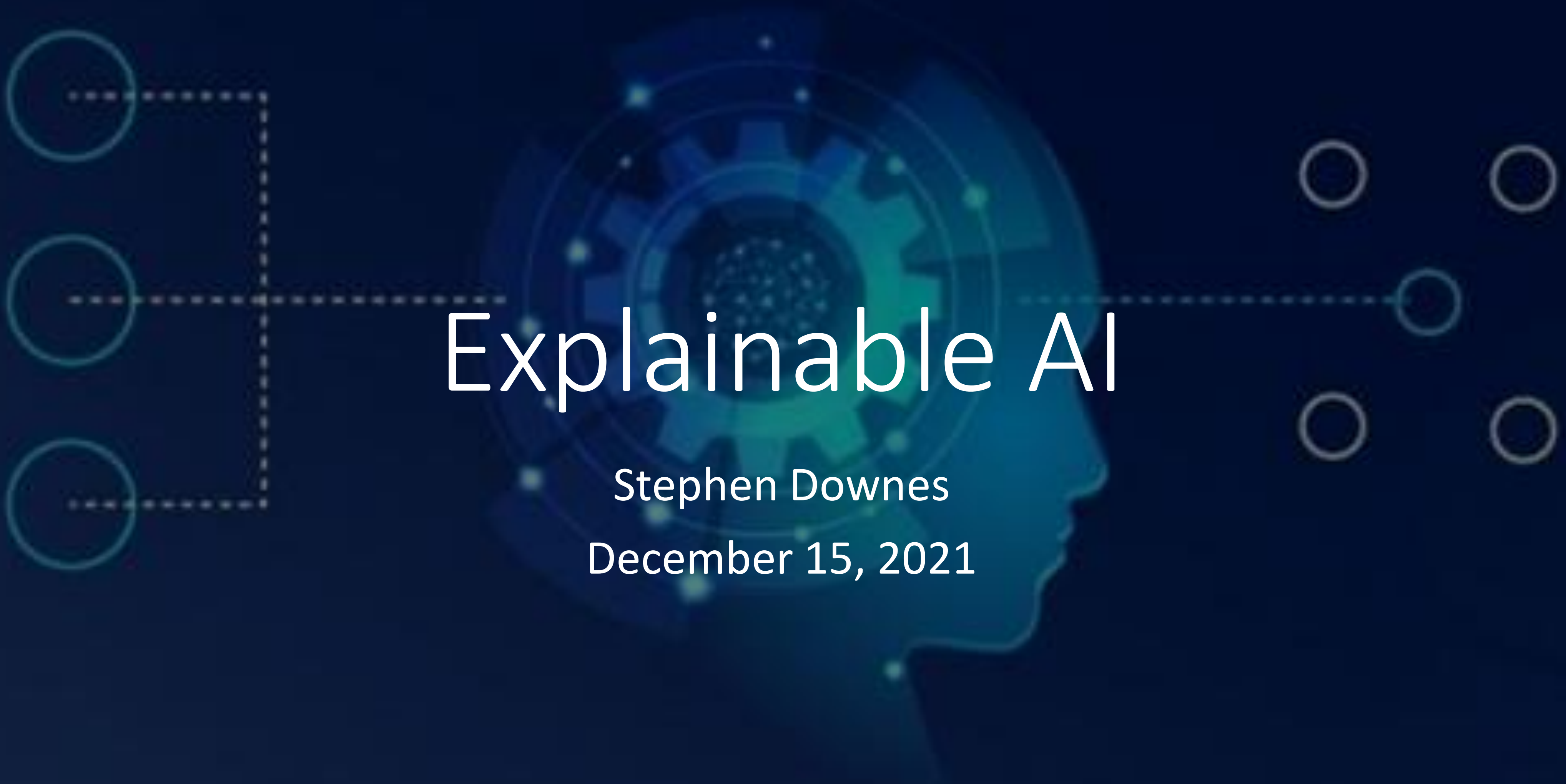


Explainable AI

Stephen Downes

December 15, 2021

Explainable AI

UK Parliament AI committee “We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual’s life, unless it can generate a full and satisfactory explanation for the decisions it will take.”

“AI in the UK: ready, willing and able?,” UK Parliament (House of Lords) Artificial Intelligence Committee. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm
Cover image: https://www.softwebsolutions.com/resources/explainable-ai-for-business.html

What Is An Explanation?

- Causality (An explanation is an assignment of causal responsibility)
  - Internal and external causes
  - Causal chains
  - Necessary and sufficient conditions
- Product (an explanation is an answer to a why–question (van Fraassen))
  - Based on presuppositions and alternative possibilities
- Abduction (inference to the best explanation (Pierce, Harman))
  - Various criteria employed to choose among hypotheses
- Justification
  - Observers can understand the *reason* for (e.g.) a decision

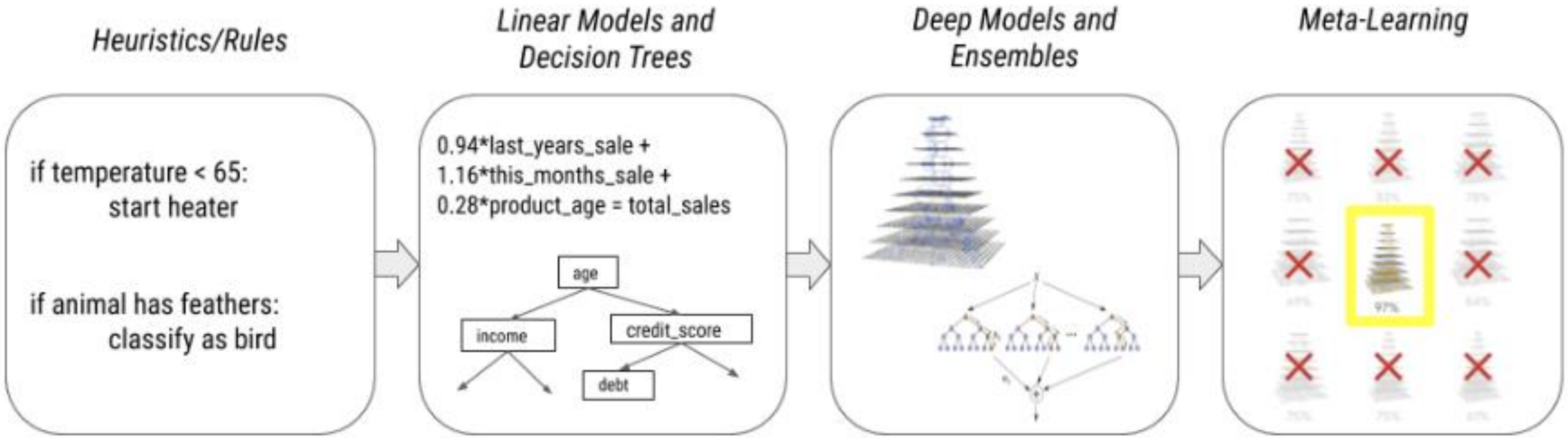
Miller, 2018 https://arxiv.org/pdf/1706.07269.pdf

The Need for Explainable AI

- Transparency: we need explanations in terms, format and language we can understand
- Causality: can the model also provide us with some explanation for underlying phenomena?
- Bias: How can we ensure that the AI system isn’t biased?
- Fairness: Can we verify that decisions were made fairly?
- Safety: Can we be confident in the reliability of our AI system?

Hagras, 2018 https://www.researchgate.net/publication/328088140_Toward_Human-Understandable_Explainable_AI

The Evolution of Machine Learning



Heuristics/Rules

```
if temperature < 65:
        start heater

if animal has feathers:
        classify as bird
```

Linear Models and Decision Trees

```
0.94*last_years_sale +
1.16*this_months_sale +
0.28*product_age = total_sales
```

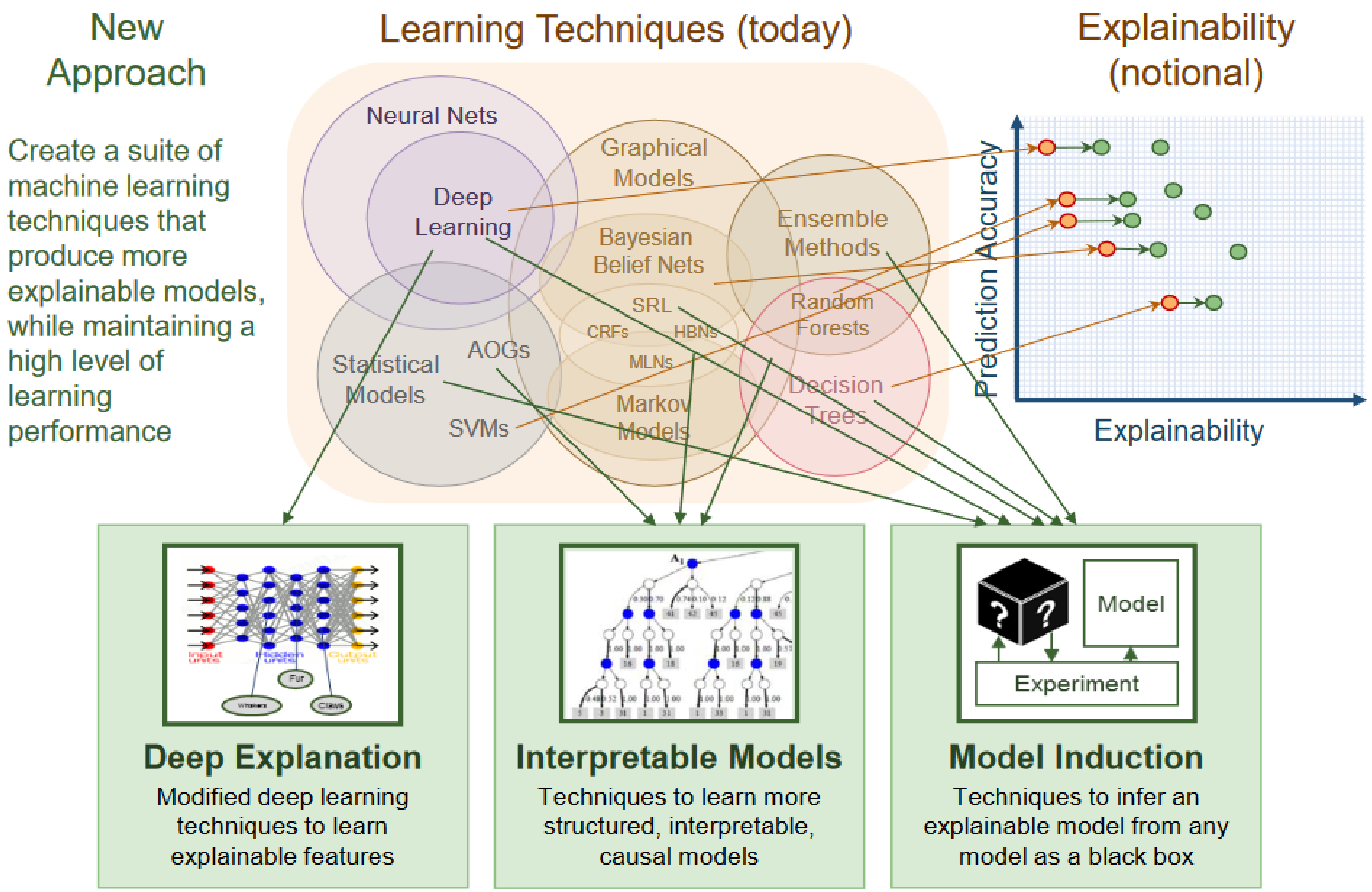
Deep Models and Ensembles

Meta-Learning

As machine learning evolves, it moves further and further away from Explainability, but the public perception of Explainability hasn't shifted

https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf

What *Is* Explainability?



New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Deep Explanation

Modified deep learning techniques to learn explainable features

Interpretable Models

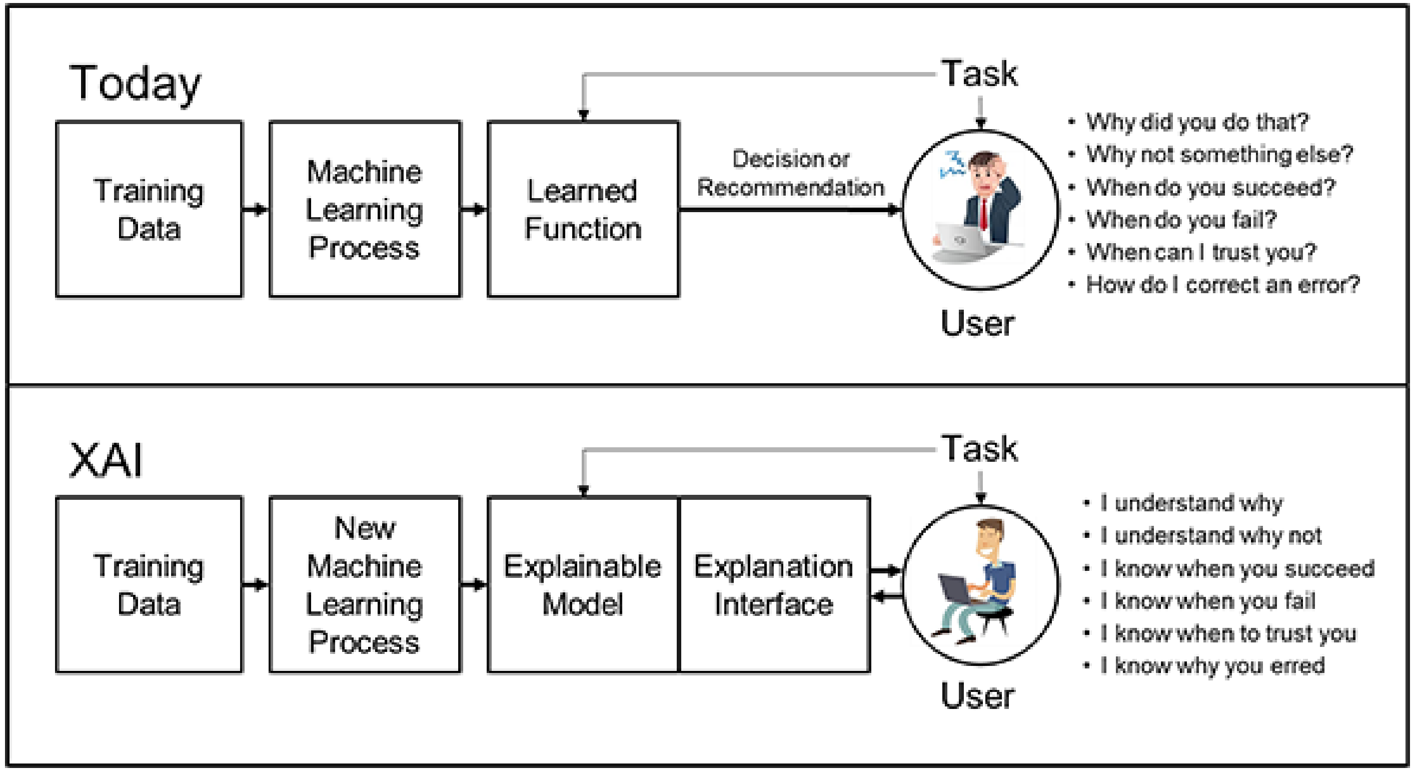
Techniques to learn more structured, interpretable, causal models

Model Induction

Techniques to infer an explainable model from any model as a black box

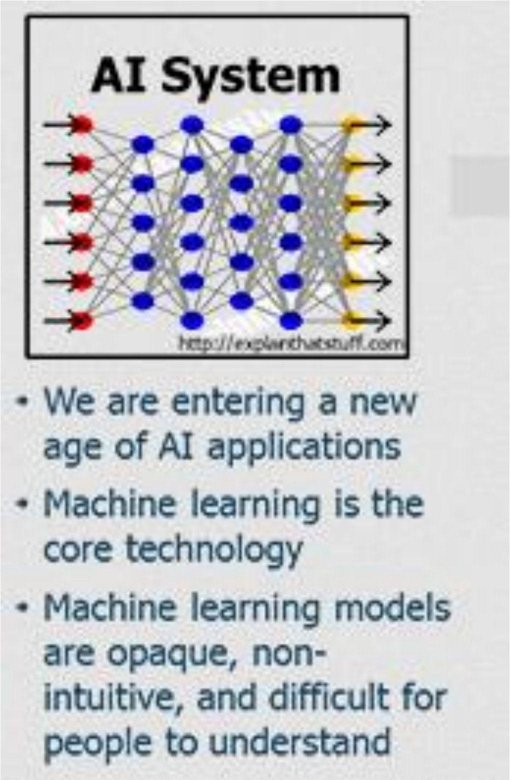
Gunning, 2017, Explainable AI https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

DARPA XAI



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

eXplainable AI intended to address weaknesses of machine learning

https://www.darpa.mil/program/explainable-artificial-intelligence

Explainable AI

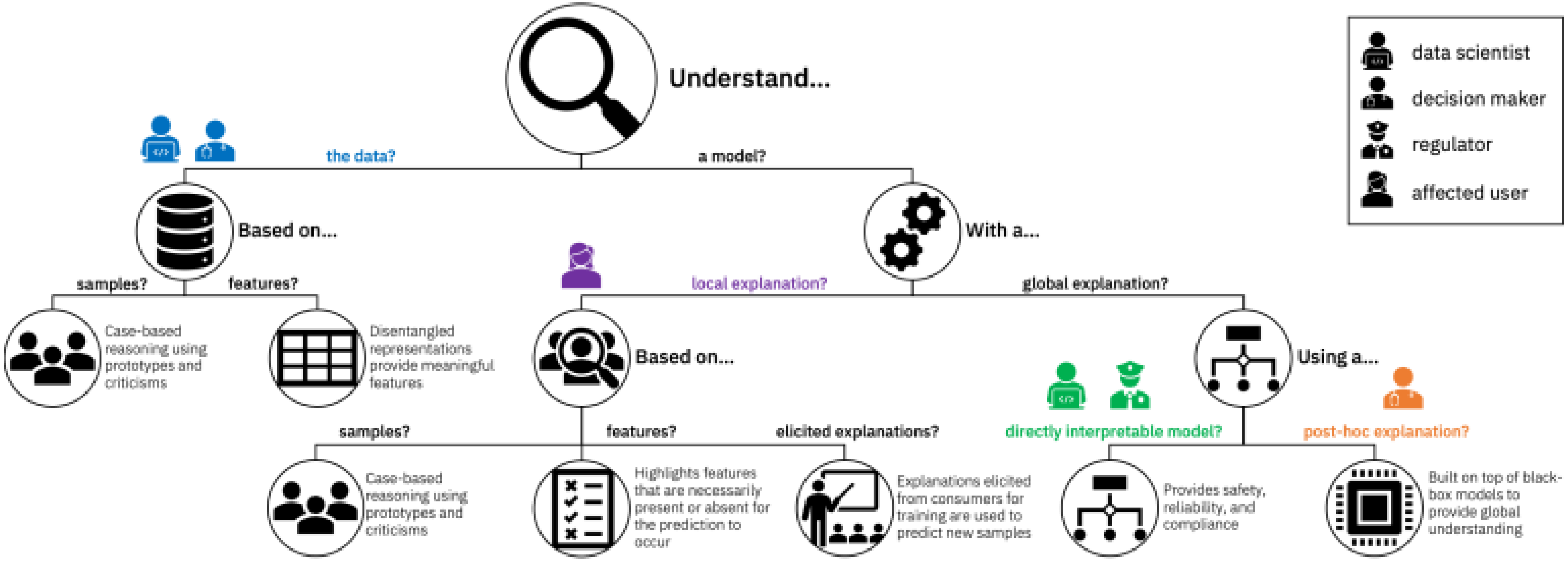
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

"Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms." It describes:

- The model
- The expected impact
- Potential biases

https://www.ibm.com/watson/explainable-ai
https://www.darpa.mil/program/explainable-artificial-intelligence

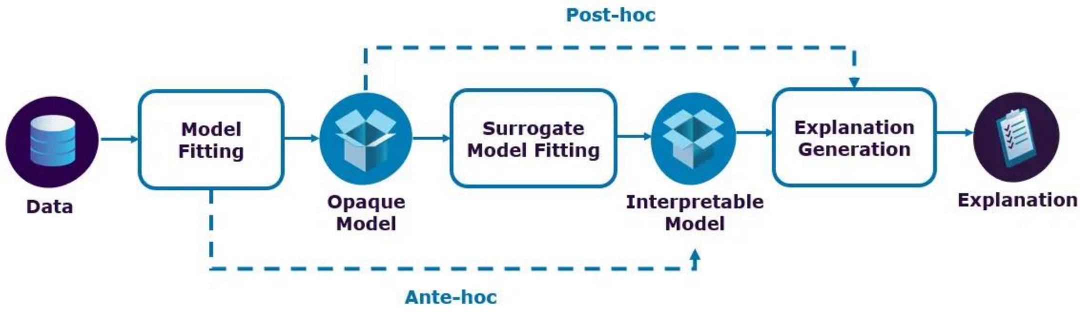
Process


Understand...
the data?
a model?
Based on...
samples?
features?
Case-based reasoning using prototypes and criticisms
Disentangled representations provide meaningful features
With a...
local explanation?
global explanation?
Based on...
samples?
features?
elicited explanations?
Case-based reasoning using prototypes and criticisms
Highlights features that are necessarily present or absent for the prediction to occur
Explanations elicited from consumers for training are used to predict new samples
Using a...
directly interpretable model?
post-hoc explanation?
Provides safety, reliability, and compliance
Built on top of black-box models to provide global understanding
data scientist
decision maker
regulator
affected user

https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

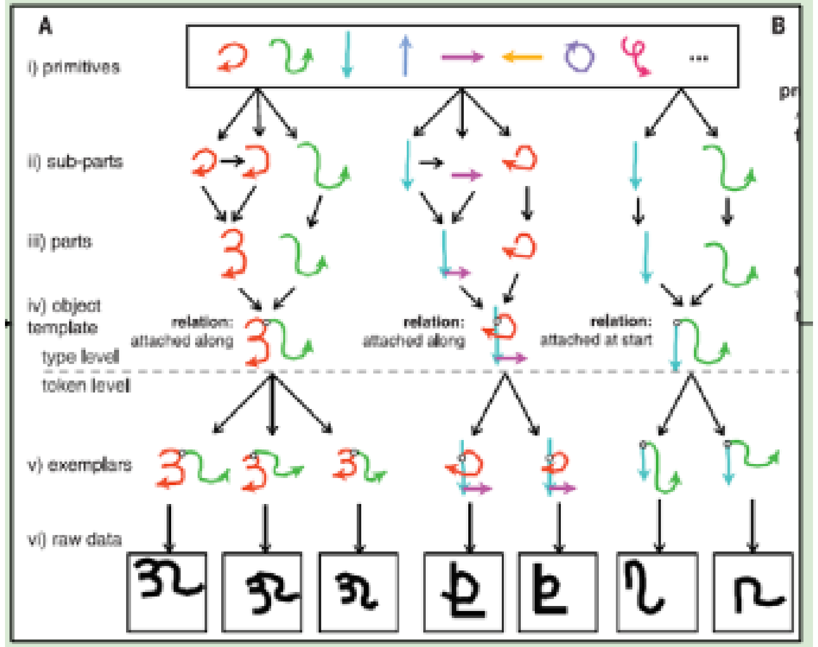
Post-Hoc Systems

- “provide local explanations for a specific decision and (make) it reproducible on demand”
    - E.g. LIME (Local Interpretable Model-Agnostic Explanations) developed by Ribeiro et al. (2016) based on “a class of potentially interpretable models, such as linear models, decision trees, or rule lists”
    - E.g. BETA (Black Box Explanations through Transparent Approximations) “optimizing for fidelity to the original model and in-terpretability of the explanation”



Holzinger, et.al., 2017 https://arxiv.org/pdf/1712.09923.pdf
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?

Ante-Hoc Systems



Generative Model
Recognizes characters by generating an explanation of how a new test character might be created (i.e., the most probable sequence of strokes that would create that character)

‘Glass-box’ approaches (Holzinger et al., 2017); “typical examples include linear regression, decision trees and fuzzy inference systems”

https://arxiv.org/pdf/1712.09923.pdf

Image: Gunning, 2017

Boolean Classification Rules

“Boolean Classification Rules via Column Generation, is an accurate and scalable method of directly interpretable machine learning” - IBM

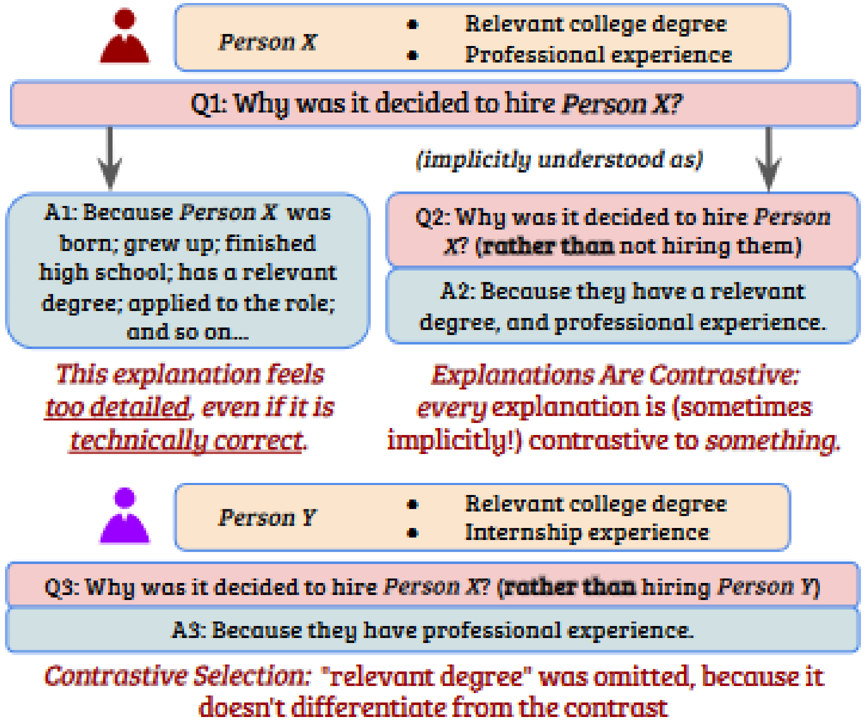
2ND KEY | CONTINUE PROTECTING NICHOLS IN POCKET
BODY AUTOBODY & GLASS | 3 KEYS TO THE GAME

The predictive decision rule for Federer defeating Murray in the 2013 Australian Open was:
- Win more than 59% of 4 to 9 shot rallies; and
- Win more than 78% of points when serving at 30-30 or Deuce; and
- Serve less than 20% of serves into the body.

https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/
Dmitry M. Malioutov & Kush R. Varshney , 2013, Exact Rule Learning via Boolean Compressed Sensing
http://proceedings.mlr.press/v28/malioutov13.pdf

Contrastive Explanations Method

“Addresses the most important consideration of explainable AI that has been overlooked by researchers and practitioners: explaining why an event happened not in isolation, but why it happened instead of some other event.” - IBM



Person X

- Relevant college degree
- Professional experience

Q1: Why was it decided to hire *Person X?*

(implicitly understood as)

A1: Because *Person X* was born; grew up; finished high school; has a relevant degree; applied to the role; and so on...

This explanation feels too detailed, even if it is technically correct.

Q2: Why was it decided to hire *Person X?* (**rather than** not hiring them)

A2: Because they have a relevant degree, and professional experience.

Explanations Are Contrastive: every explanation is (sometimes implicitly!) contrastive to *something.*

Person Y

- Relevant college degree
- Internship experience

Q3: Why was it decided to hire *Person X?* (**rather than** hiring *Person Y*)

A3: Because they have professional experience.

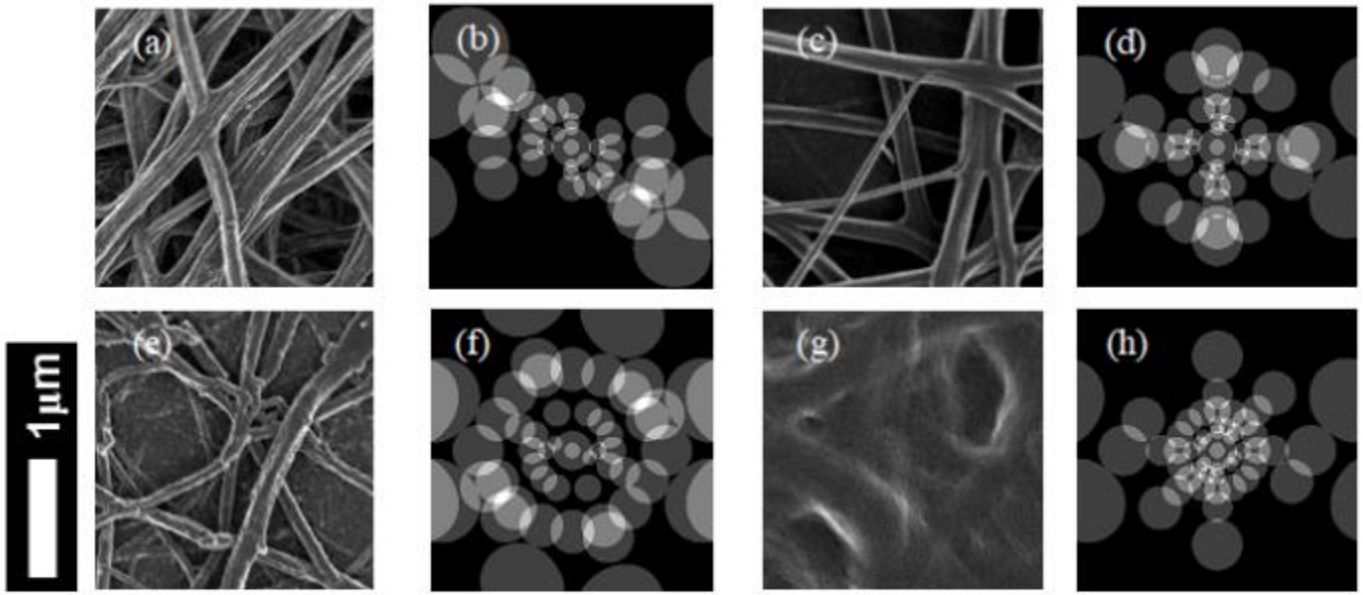
Contrastive Selection: "relevant degree" was omitted, because it doesn't differentiate from the contrast

https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

Jacovi, et.al., 2021 https://arxiv.org/pdf/2103.01378.pdf

Interpreting Deep Neural Networks

- Annotated data is extremely difficult to obtain
- Montavon et al. (2017 based on interpreting the input layer

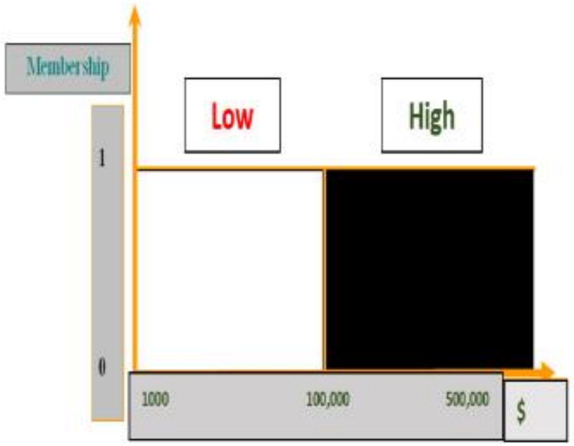


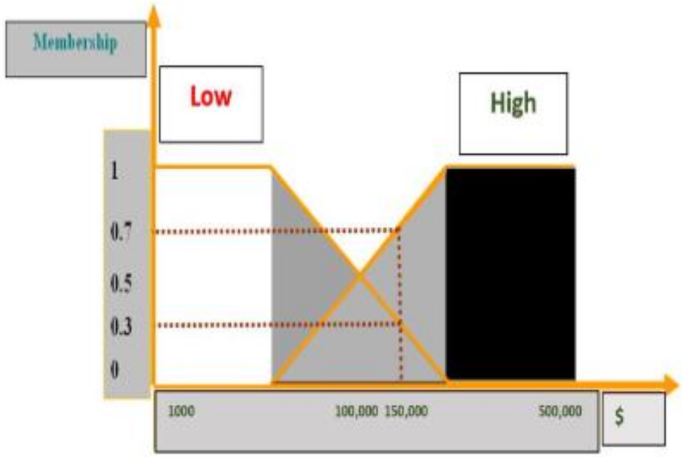
Gr´egoire Montavon, Wojciech Samek, and Klaus-Robert M¨uller. Methods for interpreting and understanding deep neural networks. arXiv:1706.07979, 2017
Holzinger, et.al., 2017 https://arxiv.org/pdf/1712.09923.pdf

Fuzzy Logic Approaches

The focus is on how humans think in an approximate rather than an precise way.

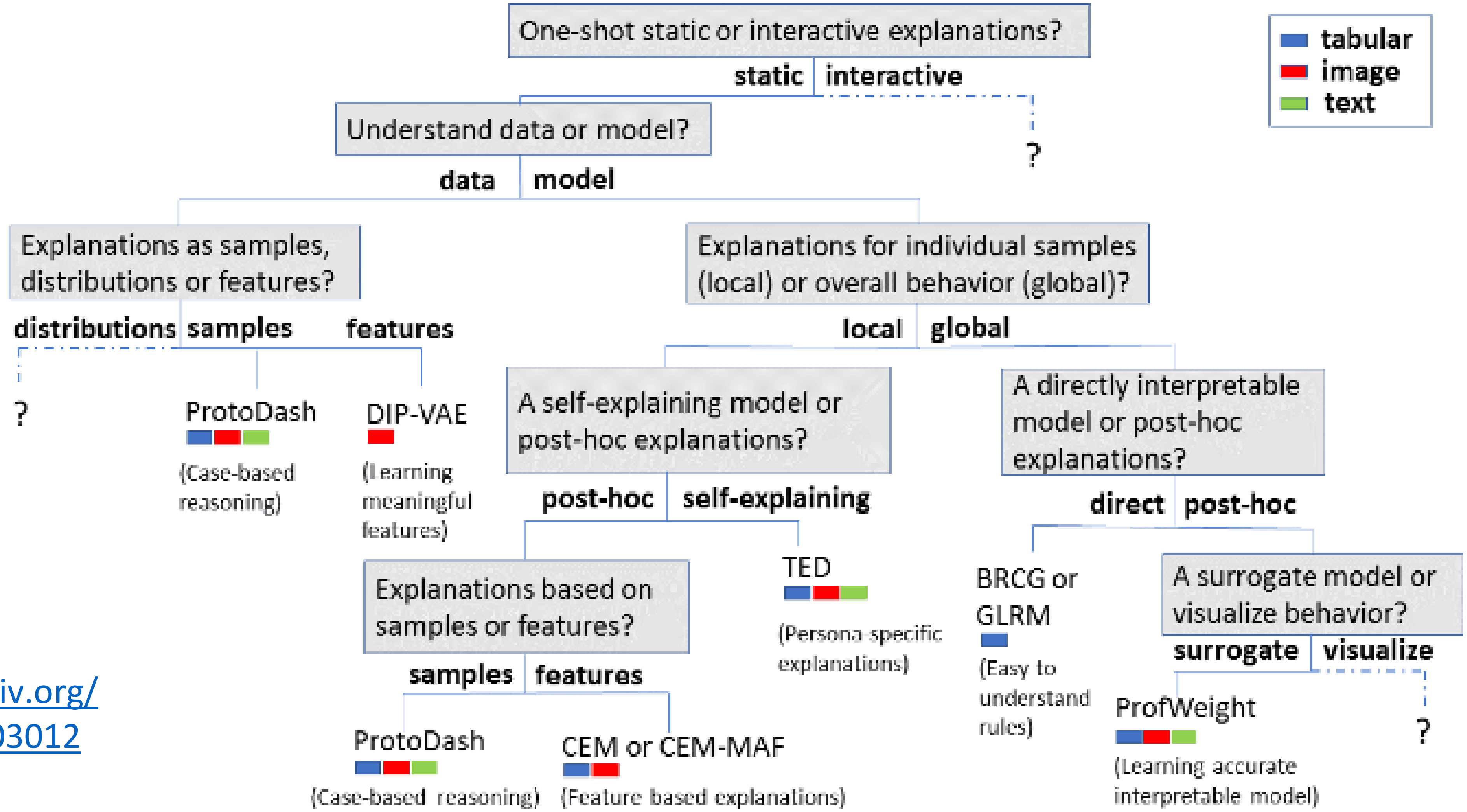
- E.g. “if the distance to the car ahead is low and the road is slightly slippery Then slow down. The numerical meanings of ‘low’, ‘close’ and ‘slow down’ will differ between drivers.”





Hagras, 2018 https://www.researchgate.net/publication/328088140_Toward_Human-Understandable_Explainable_AI

AI Explainability Taxonomy



🟦 tabular
🟥 image
🟩 text

One-shot static or interactive explanations?

static | interactive

?

Understand data or model?

data | model

Explanations as samples, distributions or features?

distributions | samples | features

?

ProtoDash 🟦🟥🟩 (Case-based reasoning)

DIP-VAE 🟥 (Learning meaningful features)

Explanations for individual samples (local) or overall behavior (global)?

local | global

A self-explaining model or post-hoc explanations?

post-hoc | self-explaining

Explanations based on samples or features?

samples | features

ProtoDash 🟦🟥🟩 (Case-based reasoning)

CEM or CEM-MAF 🟦🟥 (Feature based explanations)

TED 🟦🟥🟩 (Persona specific explanations)

A directly interpretable model or post-hoc explanations?

direct | post-hoc

BRCG or GLRM 🟦 (Easy to understand rules)

A surrogate model or visualize behavior?

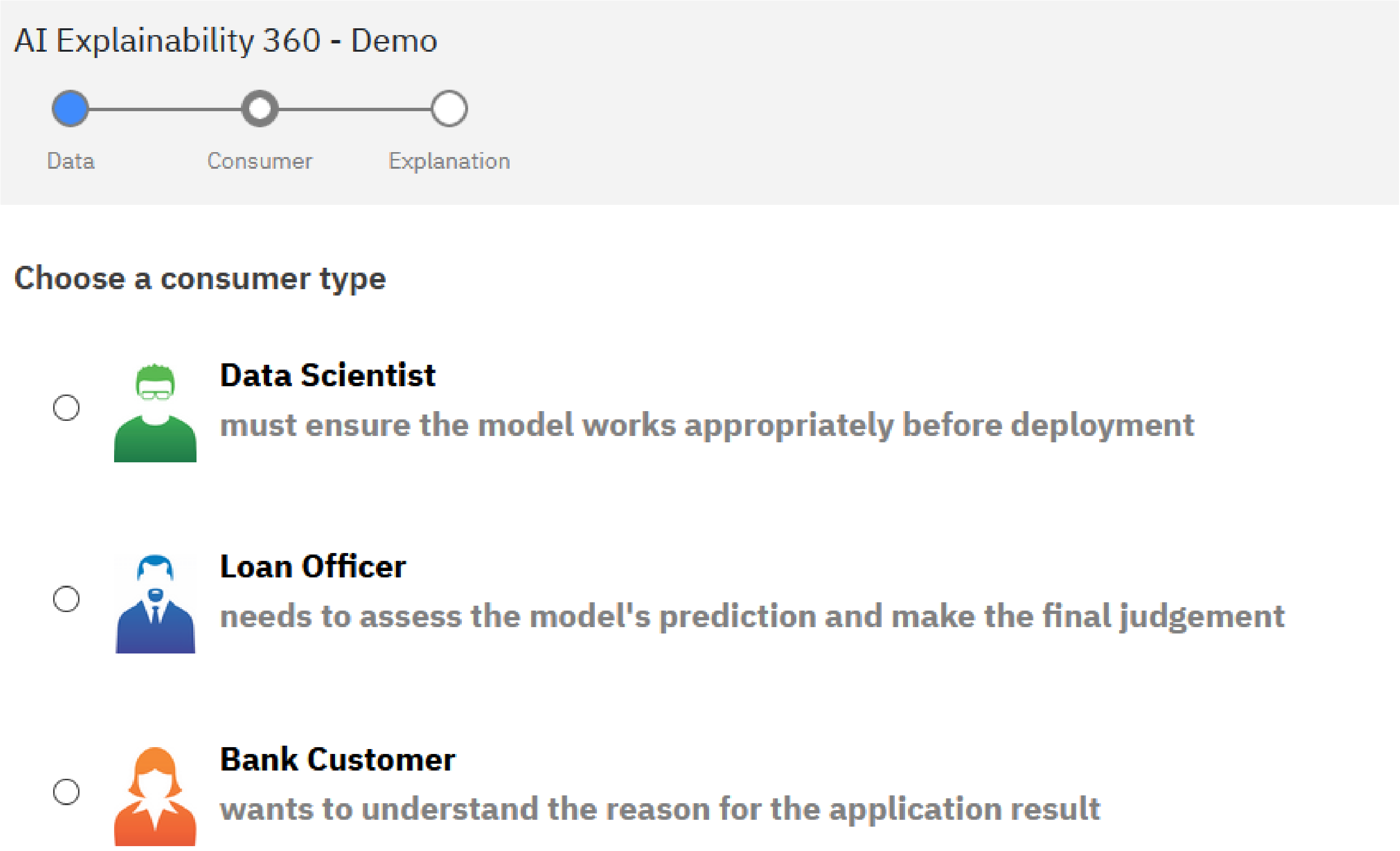
surrogate | visualize

ProfWeight 🟦🟥🟩 (Learning accurate interpretable model)

?

https://arxiv.org/abs/1909.03012

Example From IBM



AI Explainability 360 - Demo

Choose a consumer type

○ **Data Scientist**
must ensure the model works appropriately before deployment

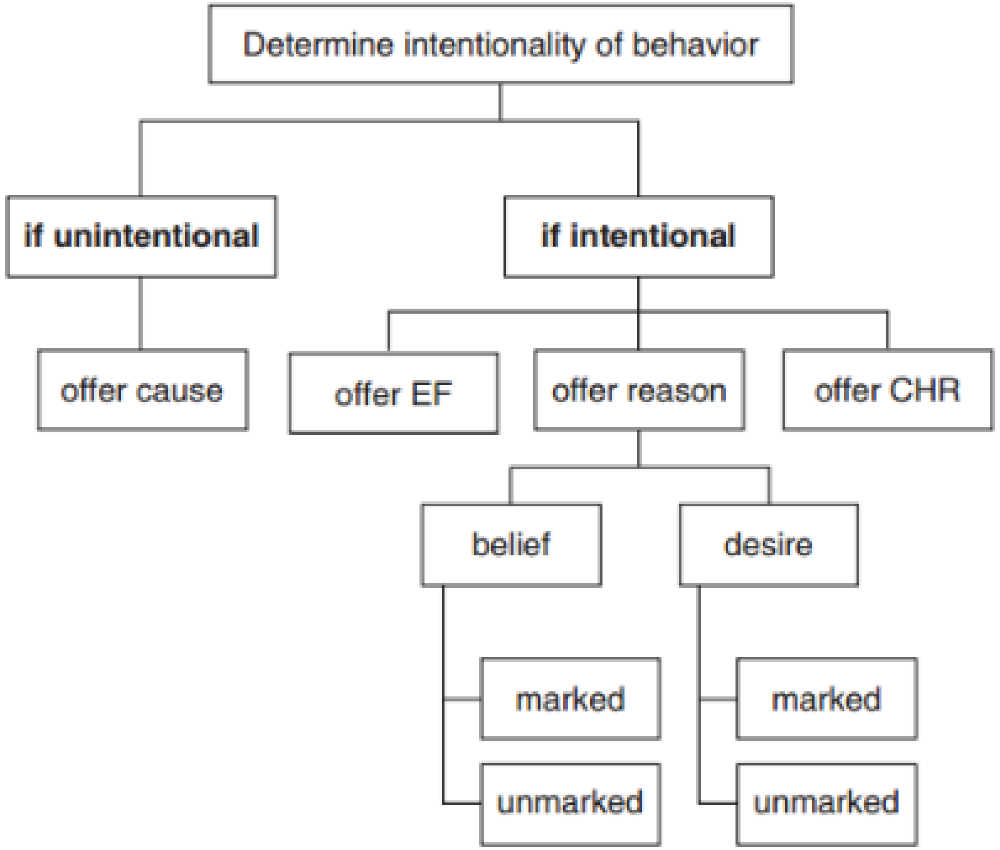
○ **Loan Officer**
needs to assess the model's prediction and make the final judgement

○ **Bank Customer**
wants to understand the reason for the application result

https://aix360.mybluemix.net/consumer

Insights from the Social Sciences

- How do we explain behaviours?
  - Intentions and intentionality – "person perception"
  - Folk psychology – beliefs, desires, etc.
- Reason expectations
  - Norms and morals
  - Collective intelligence
  - Malle's Models
    - Information requirements, information access, pragmatics goals, functional capacities

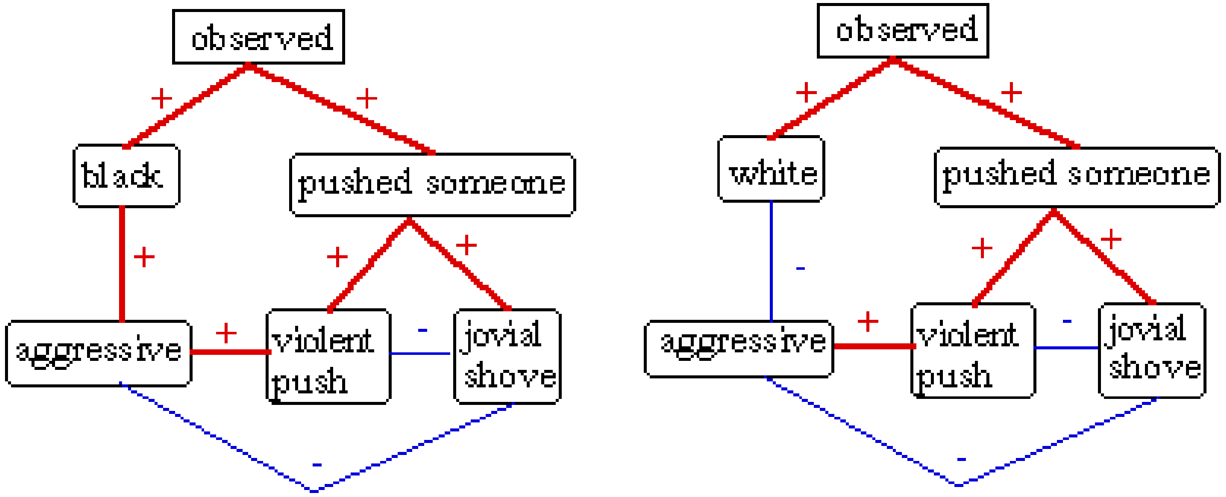


Miller, 2018 https://arxiv.org/pdf/1706.07269.pdf

B. F. Malle, How the mind explains behavior: Folk explanations, meaning, and social interaction, MIT Press, 2004.

Factors in Explanation

- Abnormality
- Temporality
- Controllability and intent
- Social norms
- Facts and Foils
- Responsibility
- Coherence, simplicity and generality
- Truth and probability



Paul Thagard and Ziva Kunda, 1997 Making sense of people: coherence mechanisms http://cogsci.uwaterloo.ca/Articles/Pages/Making.Sense.html

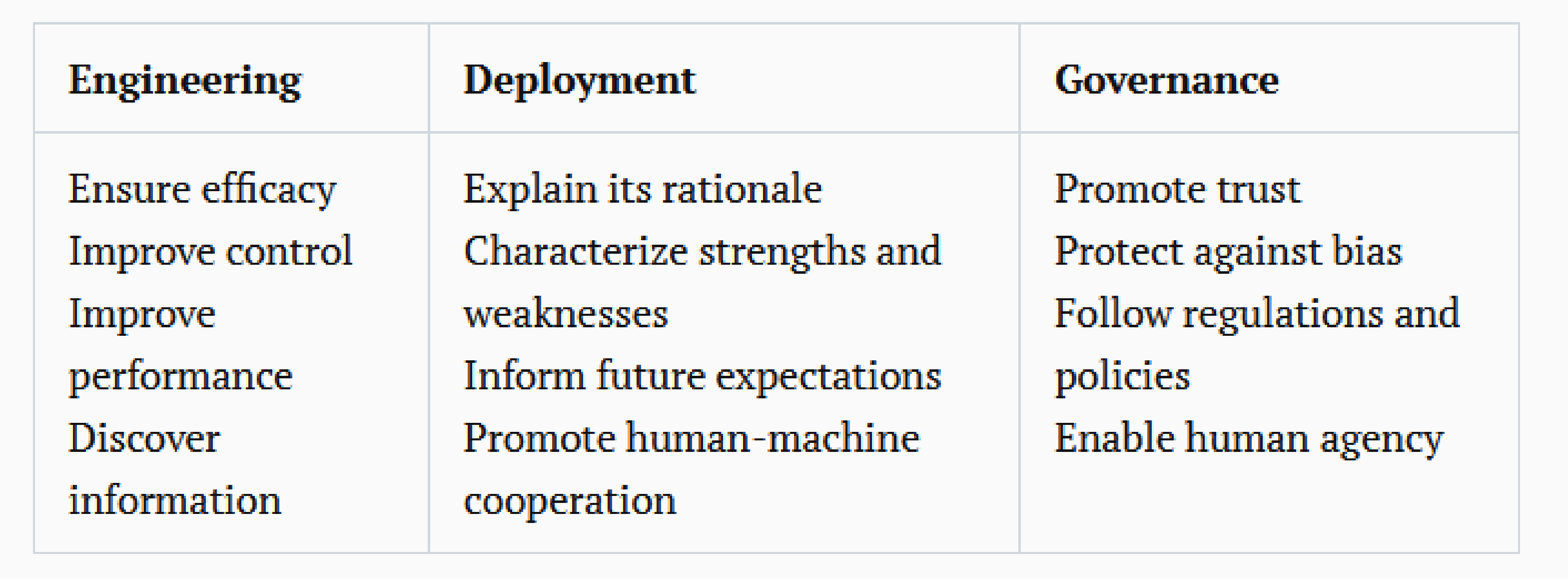
Explanation as Conversation

- Two stages:
  - the diagnosis of causality in which the explainer determines why an action/event occurred; and
  - the explanation, which is the social process of conveying this to someone.
- The problem is then to "resolve a puzzle in the explainee's mind about why the event happened by closing a gap in his or her knowledge"

Miller, 2018 https://arxiv.org/pdf/1706.07269.pdf

D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin 107 (1) (1990) 65–81.

Domains of Discourse



Engineering	Deployment	Governance
Ensure efficacy Improve control Improve performance Discover information	Explain its rationale Characterize strengths and weaknesses Inform future expectations Promote human-machine cooperation	Promote trust Protect against bias Follow regulations and policies Enable human agency

Ultimately, what will count as an explanation will depend more on who is doing the asking and explaining than on any inherent property of the system

https://www.brookings.edu/techstream/explainability-wont-save-ai/