Trustworthy AI

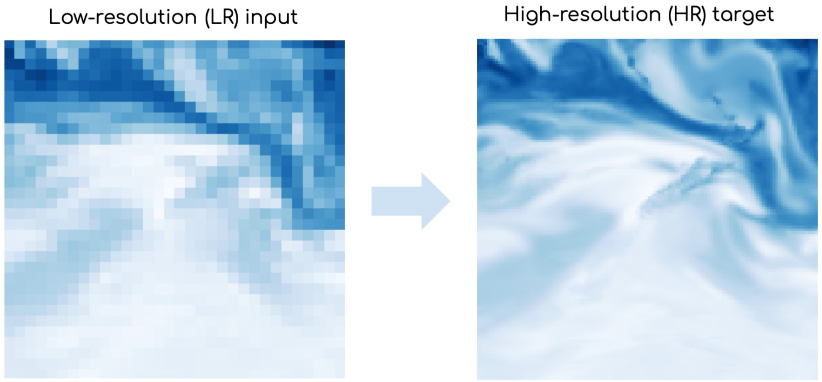
—

Prasanna Sattigeri

Principal Research Scientist
IBM Research

psattig@us.ibm.com

Example Uses of AI in Environmental, Climate and Ocean Sciences Applications

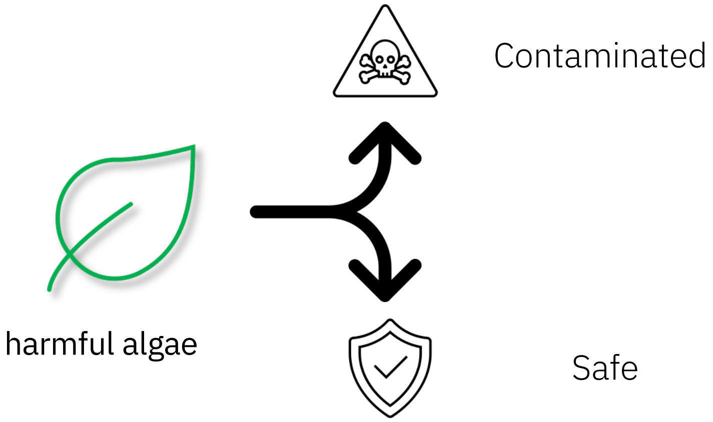


Low-resolution (LR) input

High-resolution (HR) target

Super-resolution

Harder, P., Ramesh, V., Hernandez-garcia, A., Yang, Q., Sattigeri, P., Szwarcman, D., ... & Rolnick, D. (2023, April). Physics-Constrained Deep Learning for Downscaling. In *EGU General Assembly*.



harmful algae

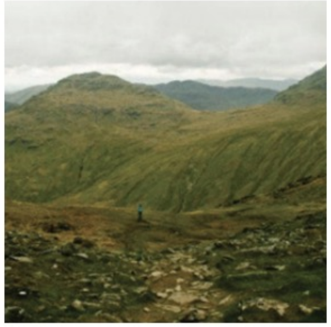
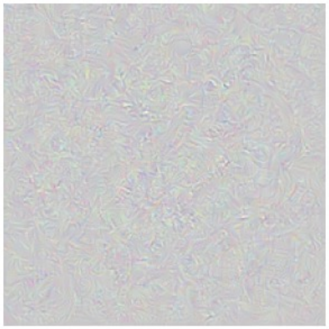
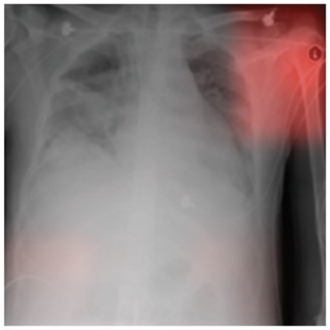
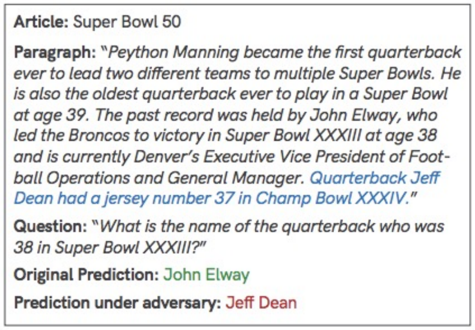
Contaminated

Safe

Classification

Cruz, Rafaela C., Pedro Reis Costa, Susana Vinga, Ludwig Krippahl, and Marta B. Lopes. "A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination." Journal of Marine Science and Engineering 9, no. 3 (2021): 283.

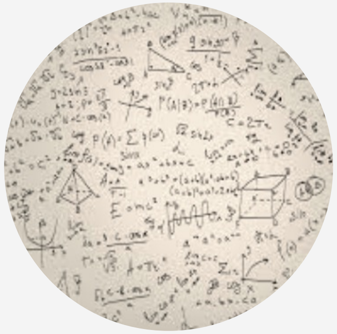
Risks of AI systems

				 **Article:** Super Bowl 50 **Paragraph:** "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*" **Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*" **Original Prediction:** John Elway **Prediction under adversary:** Jeff Dean
Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

What does it take to trust a decision made by a AI system?

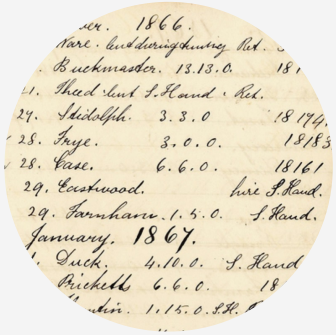
(Other than that it is 99% accurate)

Is it fair?



Is it easy to understand?

Did anyone tamper with it?



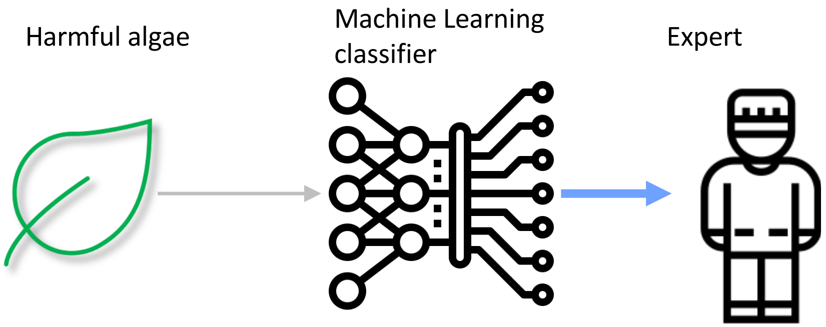
Does it know its limitations?

AI-assisted decision-making

one-way

AI makes recommendations

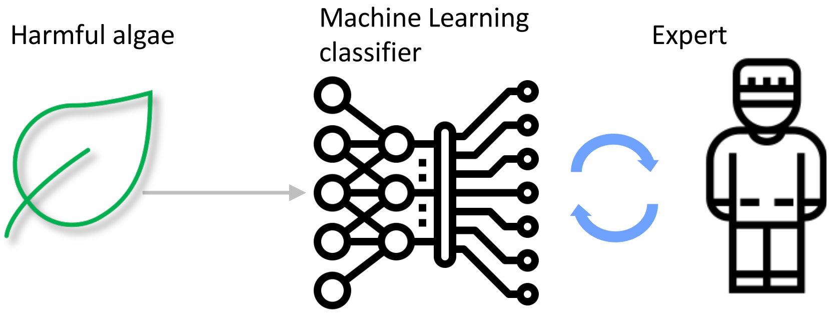
Human decision maker makes the final call



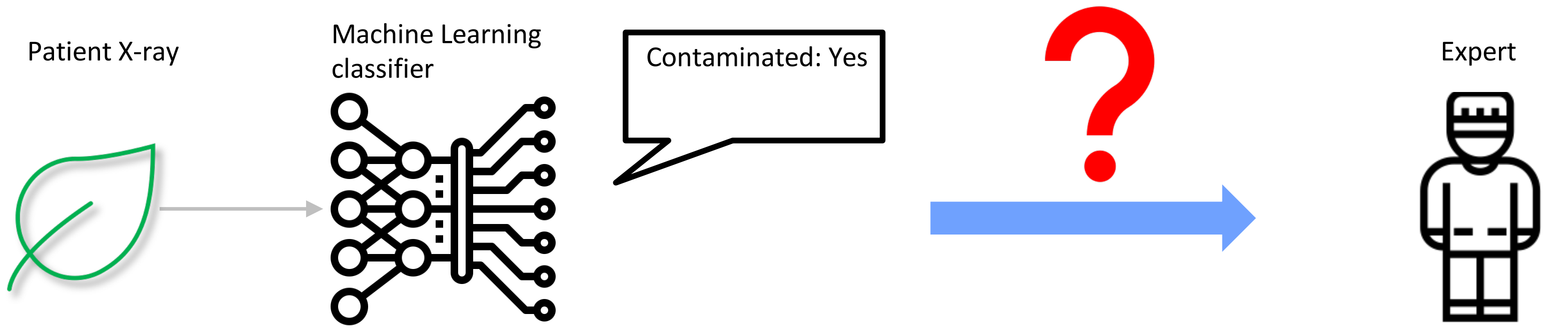
two-way

Human and AI “communicate” each others' strengths

Best “agent” decisions are accepted



Selective Prediction



Patient X-ray

Machine Learning classifier

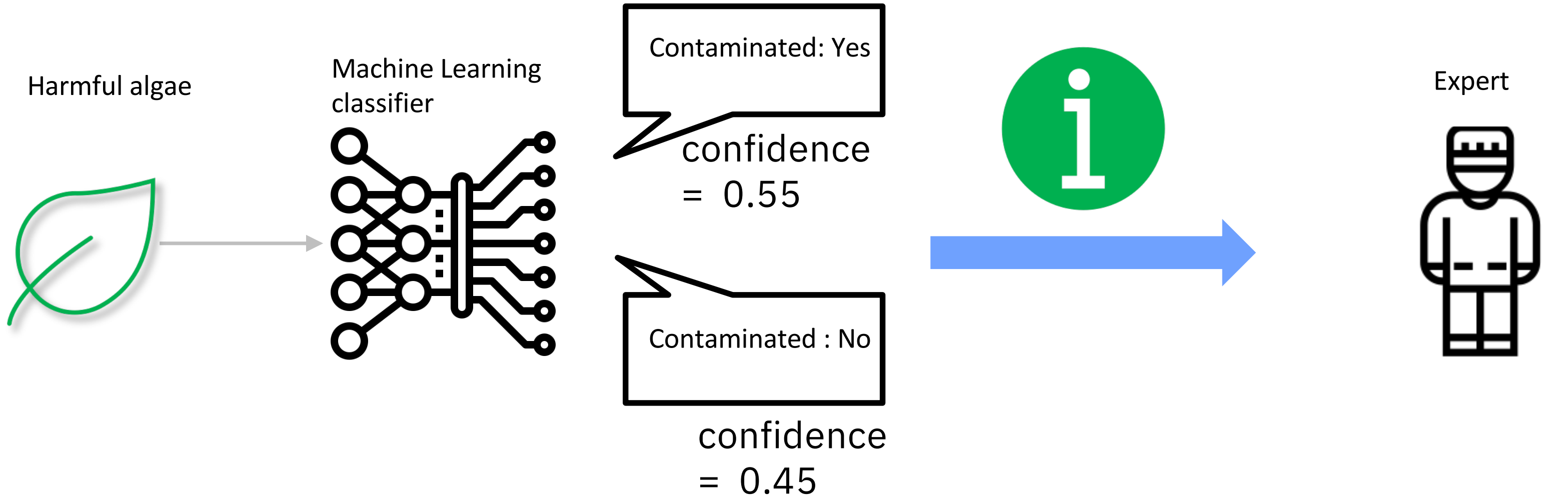
Contaminated: Yes

?

Expert

Uncertainty Quantification in AI

It is the ability of an AI model to convey the ***confidence*** in its predictions.



Harmful algae

Machine Learning classifier

Contaminated: Yes

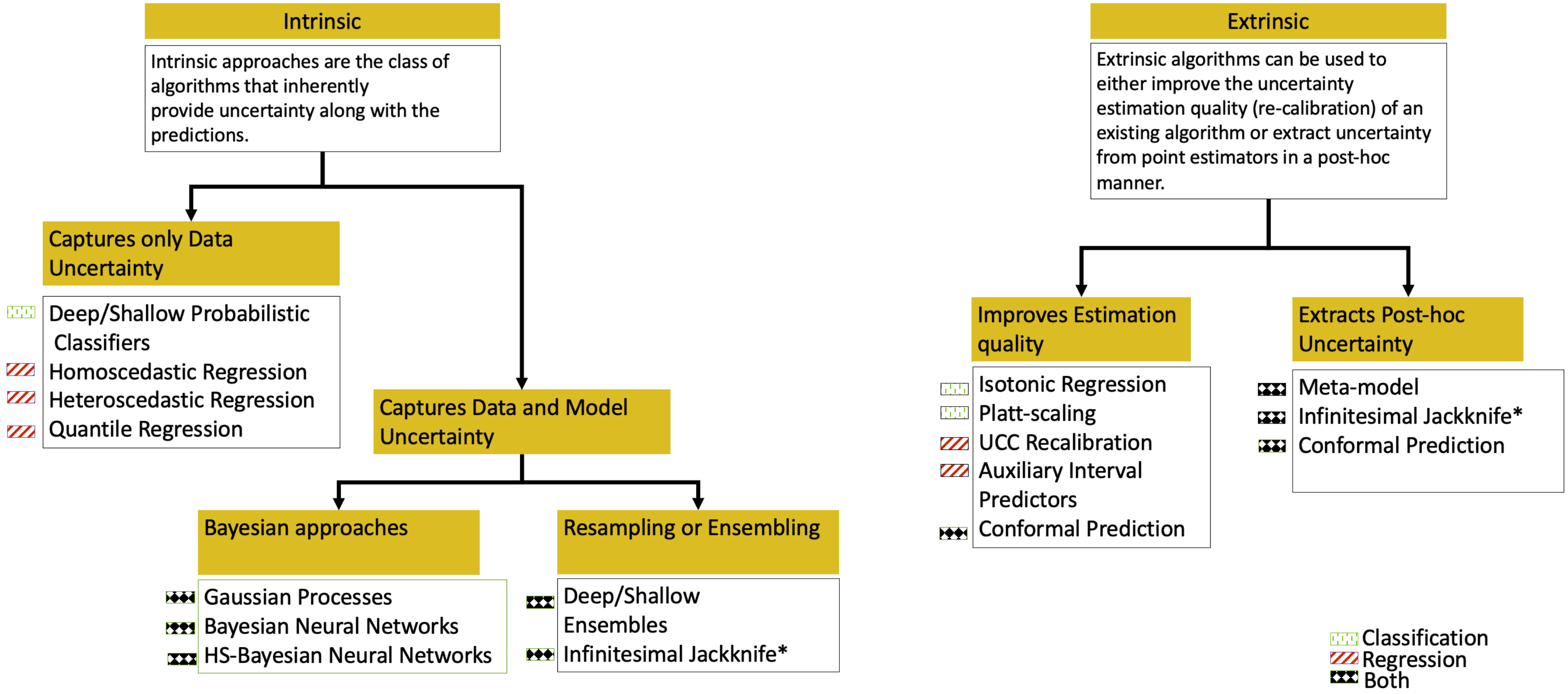
confidence = 0.55

Contaminated : No

confidence = 0.45

Expert

Ways to get uncertainty scores



Intrinsic

Intrinsic approaches are the class of algorithms that inherently provide uncertainty along with the predictions.

Captures only Data Uncertainty

- 🟩 Deep/Shallow Probabilistic Classifiers
- 🟥 Homoscedastic Regression
- 🟥 Heteroscedastic Regression
- 🟥 Quantile Regression

Captures Data and Model Uncertainty

Bayesian approaches

- ⬛ Gaussian Processes
- ⬛ Bayesian Neural Networks
- ⬛ HS-Bayesian Neural Networks

Resampling or Ensembling

- ⬛ Deep/Shallow Ensembles
- ⬛ Infinitesimal Jackknife*

Extrinsic

Extrinsic algorithms can be used to either improve the uncertainty estimation quality (re-calibration) of an existing algorithm or extract uncertainty from point estimators in a post-hoc manner.

Improves Estimation quality

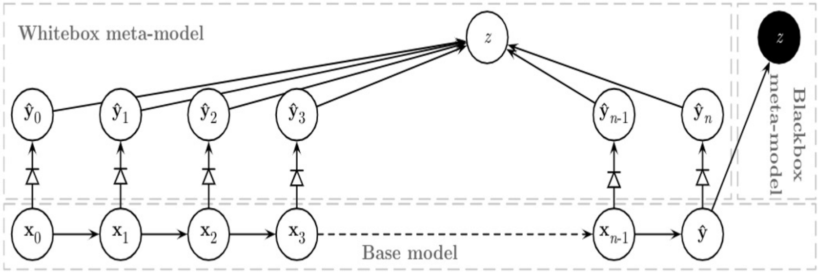
- 🟩 Isotonic Regression
- 🟩 Platt-scaling
- 🟥 UCC Recalibration
- 🟥 Auxiliary Interval Predictors
- ⬛ Conformal Prediction

Extracts Post-hoc Uncertainty

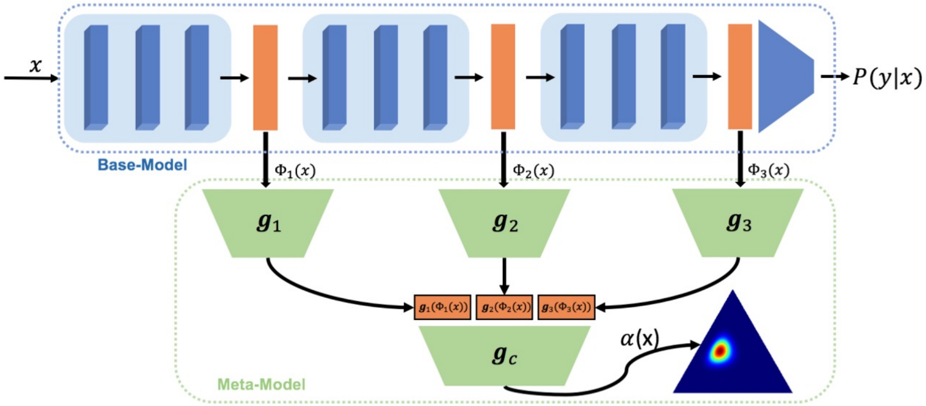
- ⬛ Meta-model
- ⬛ Infinitesimal Jackknife*
- ⬛ Conformal Prediction

🟩 Classification
🟥 Regression
⬛ Both

Meta Models



Confidence scoring using whitebox meta-models with linear classifier probes, AISTATS 2019.



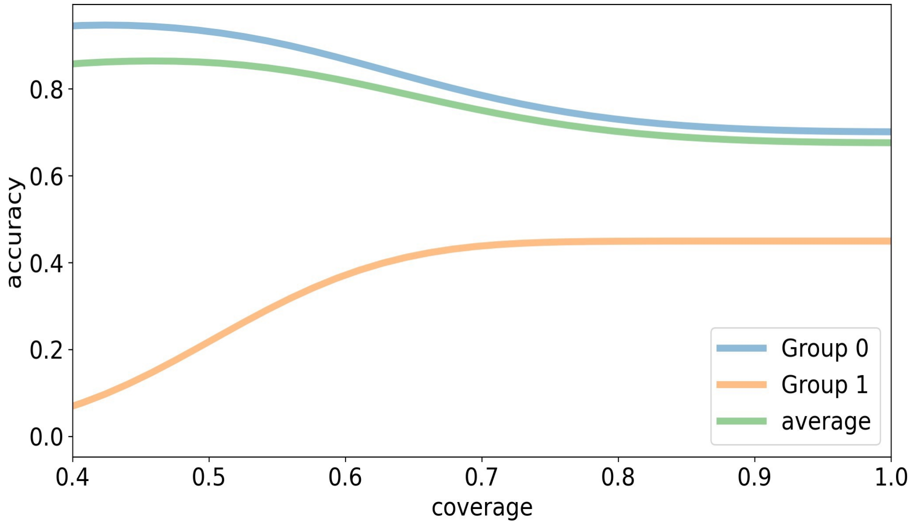
Post-hoc Uncertainty Learning using a Dirichlet Meta-Model, AAAI 2023

Mutual Information (MI): Epistemic uncertainty

$$\mathcal{I}\left(y,\boldsymbol{\pi}|\Phi(\boldsymbol{x}^*)\right) = \mathcal{H}\left(P(y|\Phi(\boldsymbol{x}^*);\boldsymbol{w}_g)\right) - \mathbb{E}_{Q(\boldsymbol{\pi}|\Phi(\boldsymbol{x}^*);\boldsymbol{w}_g)}[\mathcal{H}\left(P(y|\boldsymbol{\pi})\right)]$$

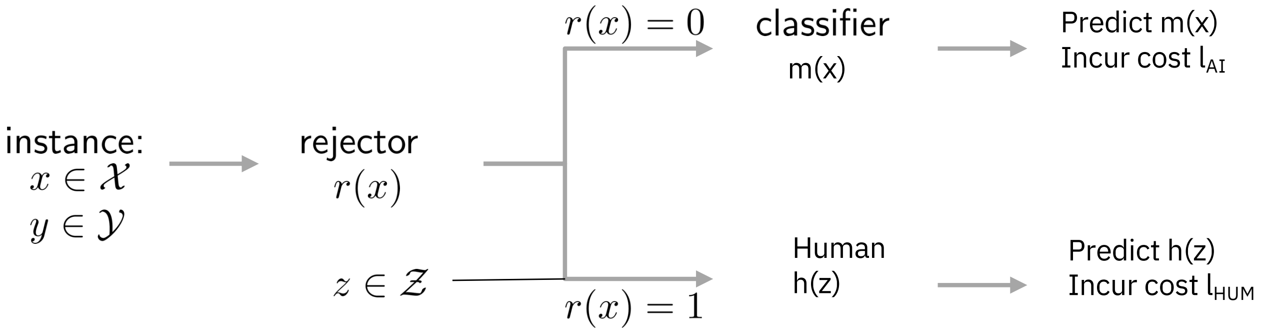
Bias in Selective Prediction

Predictors can have ***good average selective prediction*** performance but perform poorly on certain groups, where reducing uncertainty threshold may result in a ***decrease in performance for under-represented group.***



Learning to Defer

Problem Formulation



instance:
$x \in \mathcal{X}$
$y \in \mathcal{Y}$

rejector
$r(x)$

$z \in \mathcal{Z}$

$r(x) = 0$

classifier
m(x)

Predict m(x)
Incur cost l_{AI}

$r(x) = 1$

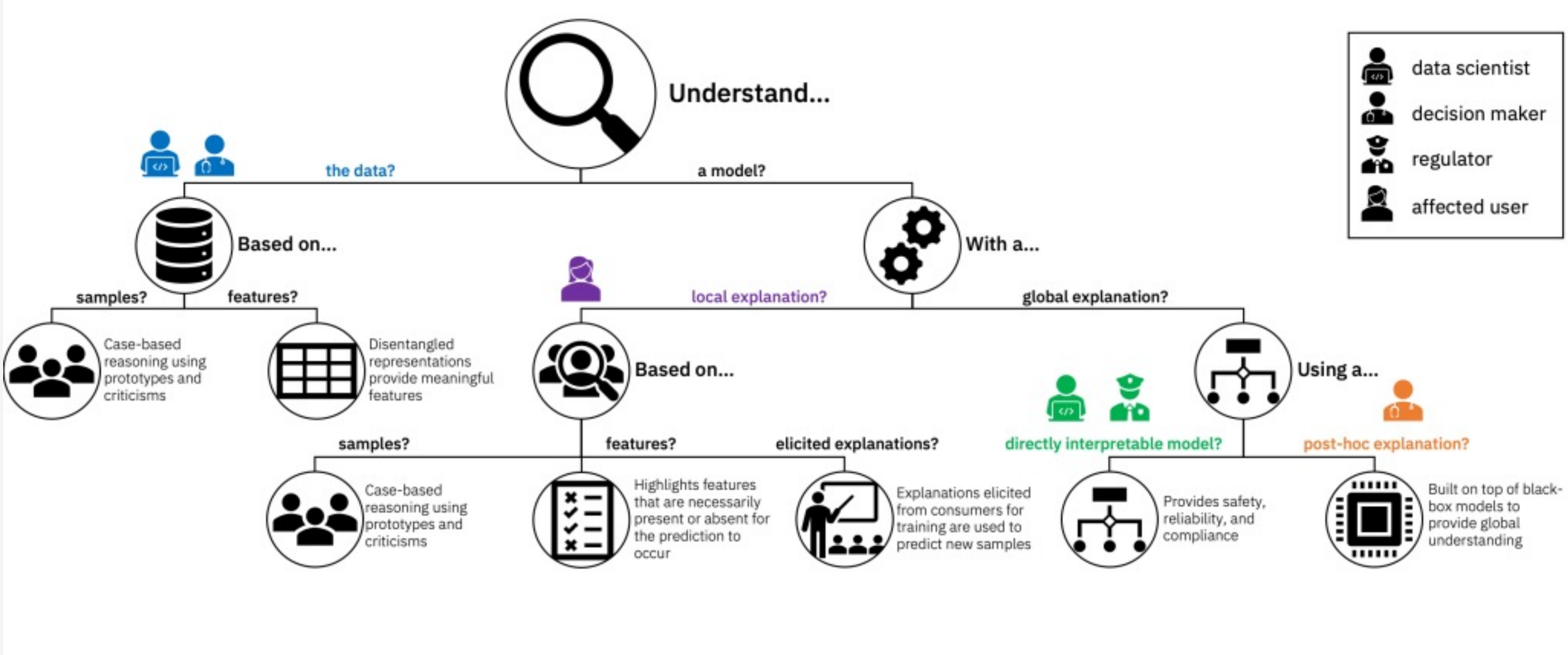
Human
h(z)

Predict h(z)
Incur cost l_{HUM}

Jointly learn a classifier *m(x)* and rejector *r(x)* to minimize system loss:

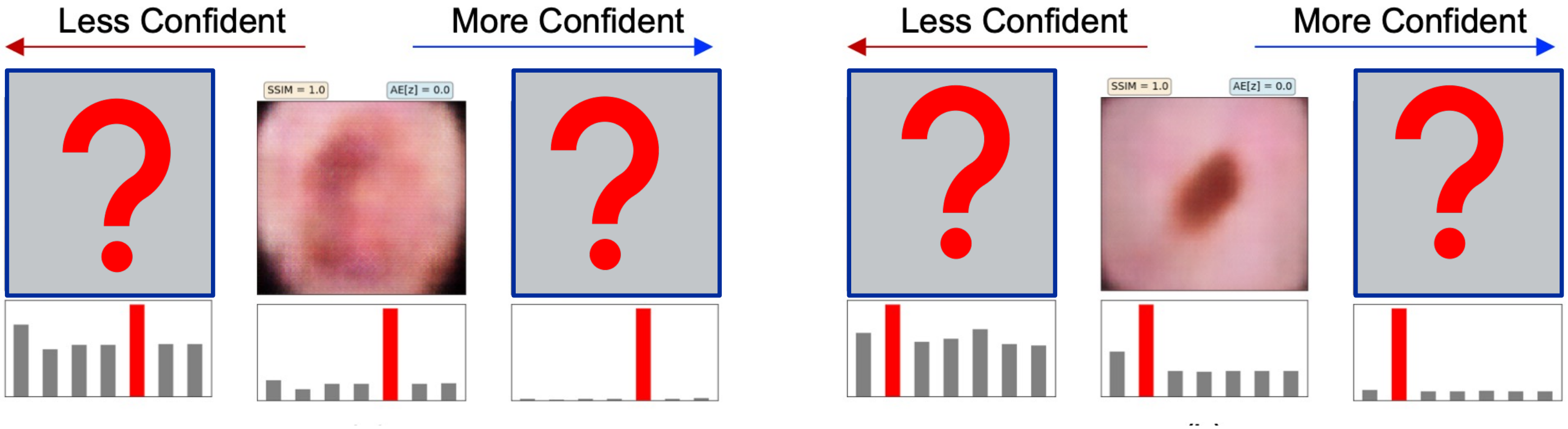
$$L^{0-1}_{\text{def}}(m,r) = \mathbb{E}_{X,Y,Z}\ [\ \ell_{\text{AI}}(X,Y,m(X)) \cdot \mathbb{I}_{r(X)=0} + \ell_{\text{HUM}}(X,Y,h(Z)) \cdot \mathbb{I}_{r(X)=1}\ ].$$

Explainability


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

Uncertainty based Introspection

Decision makers want to know what makes the model confident and vice versa?



Trustworthy AI toolkits

AI Fairness 360 http://aif360.mybluemix.net/

AI Explainability 360 http://aix360.mybluemix.net/

Adversarial Robustness 360 http://art360.mybluemix.net/

Uncertainty Quantification 360 http://uq360.mybluemix.net/

AI Privacy 360 http://aip360.mybluemix.net/

Causal Inference 360 http://ci360.mybluemix.net/

AI FactSheets 360 http://aifs360.mybluemix.net/

Thank you

Prasanna Sattigeri
Principal Research Scientist
—
psattig@us.ibm.com