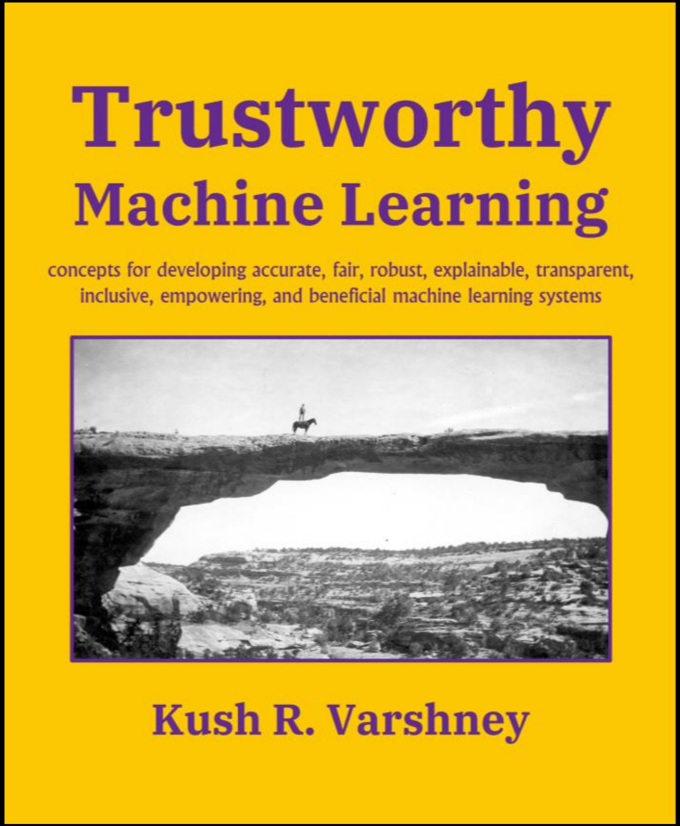
Interpretable Machine Learning for Safety and Teaming

—

Kush R. Varshney
Distinguished Research Staff Member and Manager
krvarshn@us.ibm.com | @krvarshney



http://www.trustworthymachinelearning.com

Research

Responsible AI has a few dimensions

AI Ethics

what should be done
principles, values, norms, laws, regulations

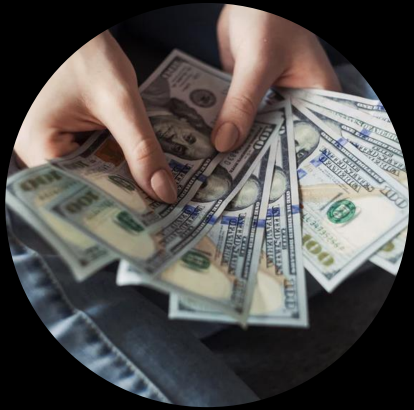
Trustworthy AI

how to instrument it
techniques, algorithms, software, best practices

AI Governance

how to operationalize it
mechanisms, systems, and processes to keep AI trustworthy

AI is powering critical workflows and trust is essential



loan processing

employment

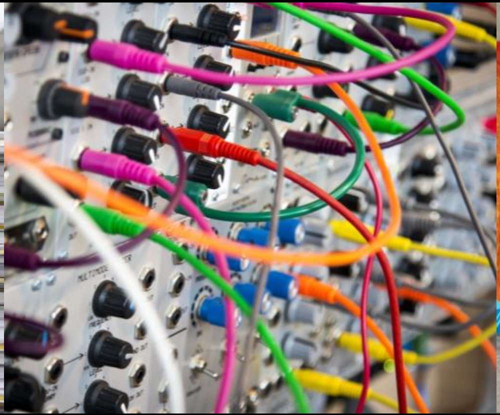
customer management

quality control

Multiple factors are placing trust in AI as a top business priority

brand reputation

increased regulation



complexity of AI deployments

focus on social justice

Attributes of trustworthiness

K. R. Varshney. "On Mismatched Detection and Safe, Trustworthy Machine Learning." *Conference on Information Sciences and Systems*, Mar. 2020.

	Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
trustworthy people	Mishra	competent	reliable	open	concerned
	Maister et al.	credibility	reliability	intimacy	low self-orientation
	Sucher and Gupta	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others' interests as well as its own
trustworthy AI	Toreini et al.	ability	integrity	predictability	benevolence
	Ashoori and Weisz	technical competence	reliability	understandability	personal attachment
		accuracy	distributional robustness; fairness; adversarial robustness	explainability; uncertainty communication; transparency; value alignment	social good; empowering

Attributes of trustworthiness

K. R. Varshney. "On Mismatched Detection and Safe, Trustworthy Machine Learning." *Conference on Information Sciences and Systems*, Mar. 2020.

safety

teaming

	Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
trustworthy people	Mishra	competent	reliable	open	concerned
	Maister et al.	credibility	reliability	intimacy	low self-orientation
	Sucher and Gupta	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others' interests as well as its own
trustworthy AI	Toreini et al.	ability	integrity	predictability	benevolence
	Ashoori and Weisz	technical competence	reliability	understandability	personal attachment
		accuracy	distributional robustness; fairness; adversarial robustness	explainability; uncertainty communication; transparency; value alignment	social good; empowering

1. Safety

Safety

K. R. Varshney and H. Alemzadeh. "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products." *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

- Commonly used term across engineering disciplines connoting the absence of failures or conditions that render a system dangerous (Ferrell, 2010)
  - Safe food and water, safe vehicles and roads, safe medical treatments, safe toys, safe neighborhoods, safe industrial plants, ...
- Each domain has specific design principles and regulations applicable only to it
- Few works attempt a precise definition applicable broadly
- Definition based on harm, risk, and epistemic uncertainty (Möller, 2012)

Harm

K. R. Varshney and H. Alemzadeh. “On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products.” *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

- A system yields an outcome based on its state and the inputs it receives
- The outcome event may be desired or undesired
- Outcomes have associated costs that can be measured and quantified by society
- An undesired outcome is a harm if its cost exceeds some threshold
- Unwanted events of small severity are not counted as safety issues

Risk

K. R. Varshney and H. Alemzadeh. “On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products.” *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

- We do not know what the outcome will be, but its distribution is known and we can calculate the expectation of its cost
- Risk is the expected value of the cost of harm

Epistemic uncertainty

K. R. Varshney and H. Alemzadeh. "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products." *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

- We still do not know what the outcome will be, but in contrast to risk, its probability distribution is also unknown
- Epistemic uncertainty, in contrast to aleatoric uncertainty, results from lack of knowledge that could be obtained in principle, but may be practically intractable to gather
- Some decision theorists argue that all uncertainty can be captured probabilistically, but we maintain the distinction between risk and uncertainty, following Möller (2012)

Safety

K. R. Varshney and H. Alemzadeh. "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products." *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

- Safety is the reduction or minimization of risk <u>and</u> epistemic uncertainty of harmful events
- Costs have to be sufficiently high in some human sense for events to be harmful
- Safety involves reducing both the probability of expected harms <u>and</u> the possibility of unexpected harms

Risk minimization in machine learning

K. R. Varshney and H. Alemzadeh. "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products." *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

– Risk minimization is the basis of statistical machine learning theory and practice

- Features $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$ with probability density $f_{X,Y}(x,y)$
- Function mapping $h \in \mathcal{H}: \mathcal{X} \rightarrow \mathcal{Y}$
- Loss function $L: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$
- Find h to minimize risk $R(h) = \mathbb{E}[L(h(X), Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(h(x), y) f_{X,Y}(x,y) dy dx$

– Given m i.i.d. training samples, not $f_{X,Y}(x,y)$

- Empirical risk minimization $R_m^{emp}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i)$
- R_m^{emp} converges to R uniformly for all h as m goes to infinity (Glivenko-Cantelli)

– When m is small, minimizing R_m^{emp} may not yield an h with small R

- Restrict complexity of $\mathcal{H}$ based on some inductive bias (Vapnik, 1992)

Epistemic uncertainty in machine learning

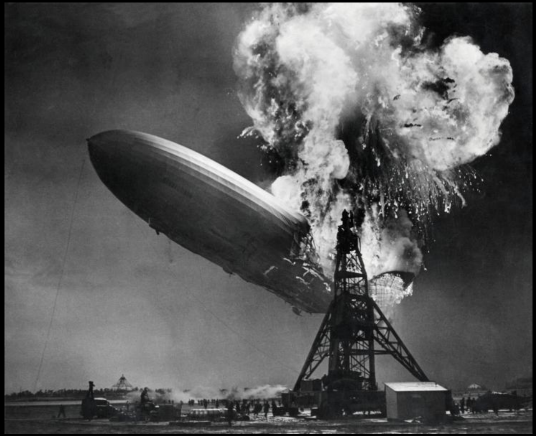
K. R. Varshney and H. Alemzadeh. "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products." *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

- Risk minimization has many strengths but does not capture epistemic uncertainty
- Not always the case that training samples are drawn from true underlying probability distribution of X, Y
  - The distribution the samples come from cannot always be known
  - Training on a data set from a different distribution can cause much harm
- Even when drawn from true distribution, training samples may be absent from large parts of $\mathcal{X} \times \mathcal{Y}$ due to small probability density there

How to achieve safety in engineering

K. R. Varshney and H. Alemzadeh. “On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products.” *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

– Inherently safe design: exclusion of a potential hazard from the system
  - Blimps filled with helium instead of hydrogen
– Safety margin: a system that is stronger than it needs to be for an intended load
  - Hurricane-resistant windows
– Safe fail: system remains safe when it fails in its intended operation
  - Dead man’s switches on trains
– Procedural safeguard: measures beyond ones designed into the core functionality of the system
  - Certifications and warning notices



How to achieve safety in AI

K. R. Varshney and H. Alemzadeh. "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products." *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

– Inherently safe design: exclusion of a potential hazard from the system
  • Blimps filled with helium instead of hydrogen

directly interpretable models and causal modeling

– Safety margin: a system that is stronger than it needs to be for an intended load
  • Hurricane-resistant windows
– Safe fail: system remains safe when it fails in its intended operation
  • Dead man's switches on trains

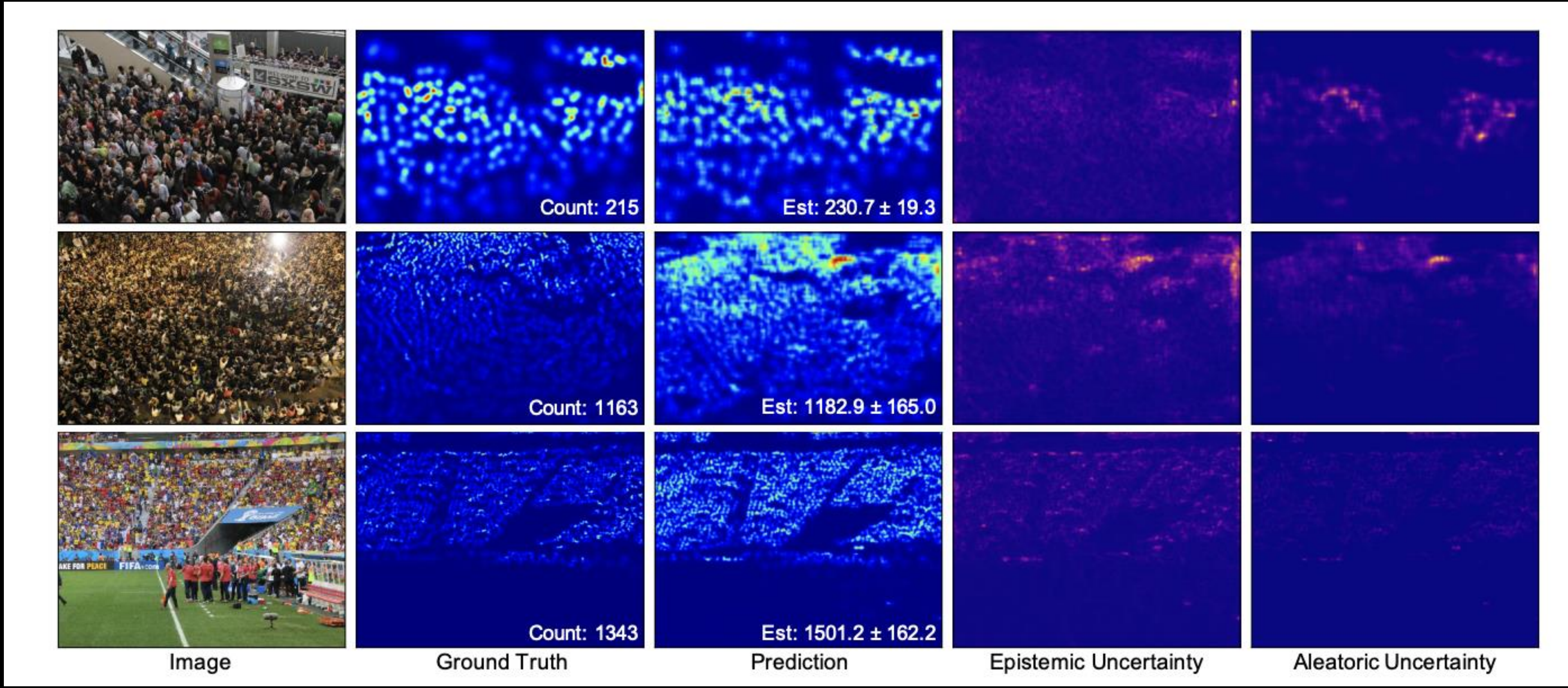
uncertainty quantification and selective classification

– Procedural safeguard: measures beyond ones designed into the core functionality of the system
  • Certifications and warning notices

transparency

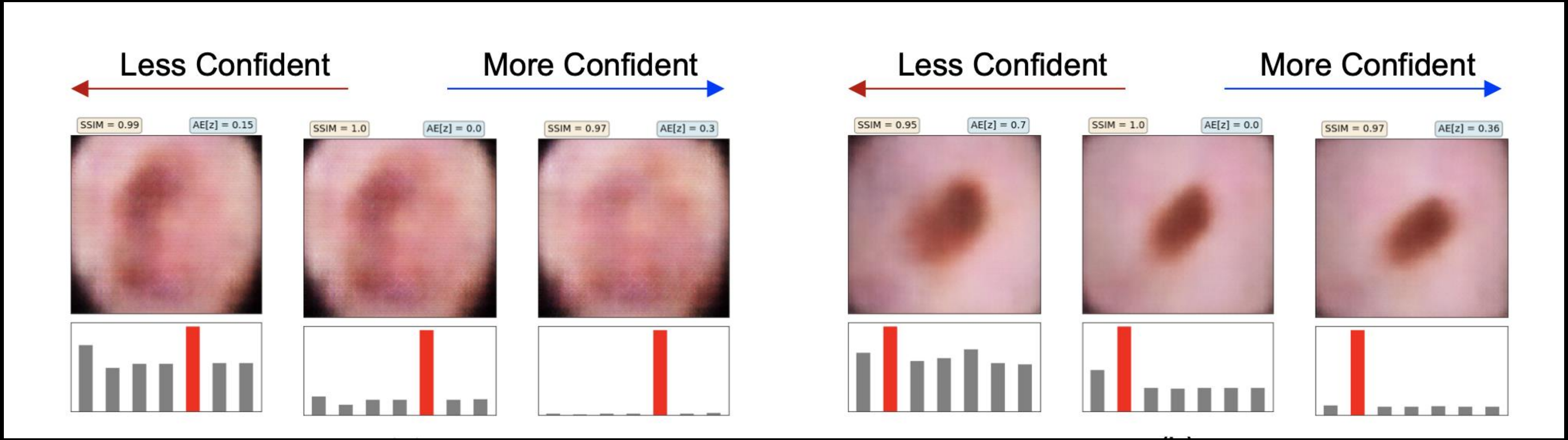
Uncertainty quantification in crowd counting
(safety margin – limiting attendance below venue capacity)

M. Oh, P. Olsen, and K. Natesan Ramamurthy. “Crowd Counting with Decomposed Uncertainty.” *AAAI Conference on Artificial Intelligence*, pp. 11799–11806, Feb. 2020.



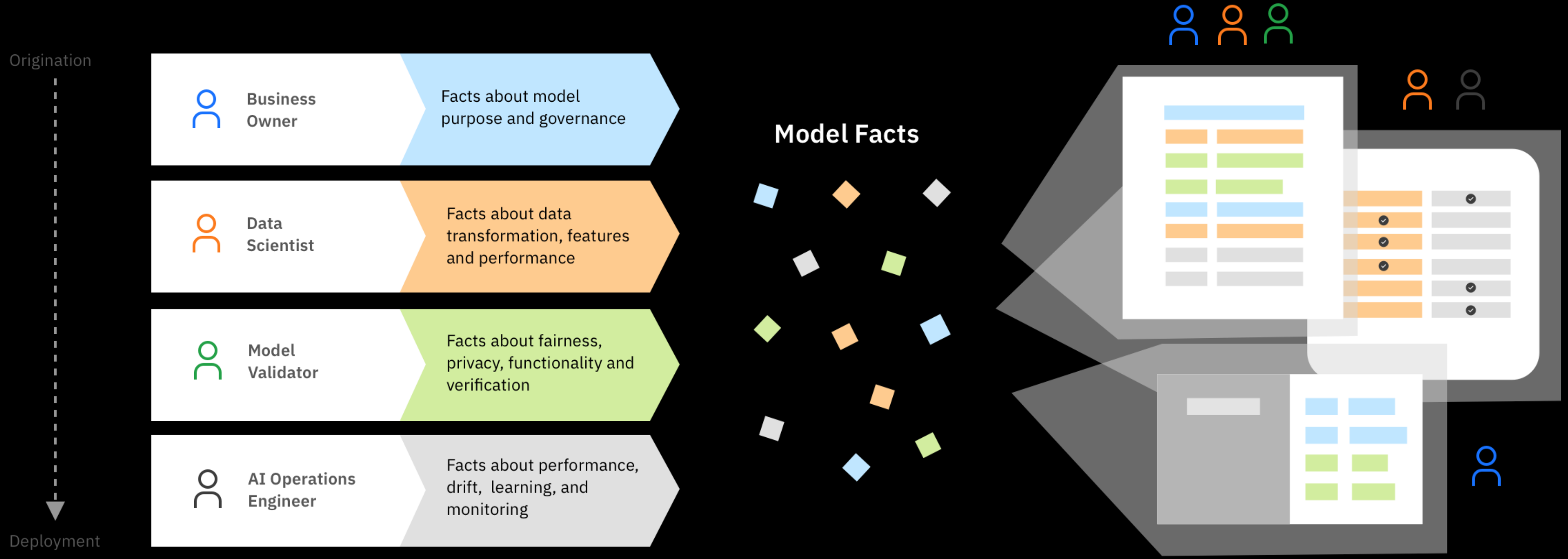
Uncertainty quantification in dermatology
(safe fail – dermatologist decides when machine has low confidence)

J. J. Thiagarajan, P. Sattigeri, D. Rajan, and B. Venkatesh. "Calibrating Healthcare AI: Towards Reliable and Interpretable Deep Predictive Models." arXiv:2004.14480, Apr. 2020.



Transparency via AI FactSheets
(procedural safeguard)

M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. “FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity.” *IBM Journal of Research and Development*, vol. 63, no. 4/5, p. 6, Jul./Sep. 2019.



Directly interpretable models
(inherently safe design)

S. Dash, O. Günlük, and D. Wei. "Boolean Decision Rules via Column Generation." *Advances in Neural Information Processing Systems*, pp. 4660–4670, Dec. 2018.

Home Equity Line of Credit:

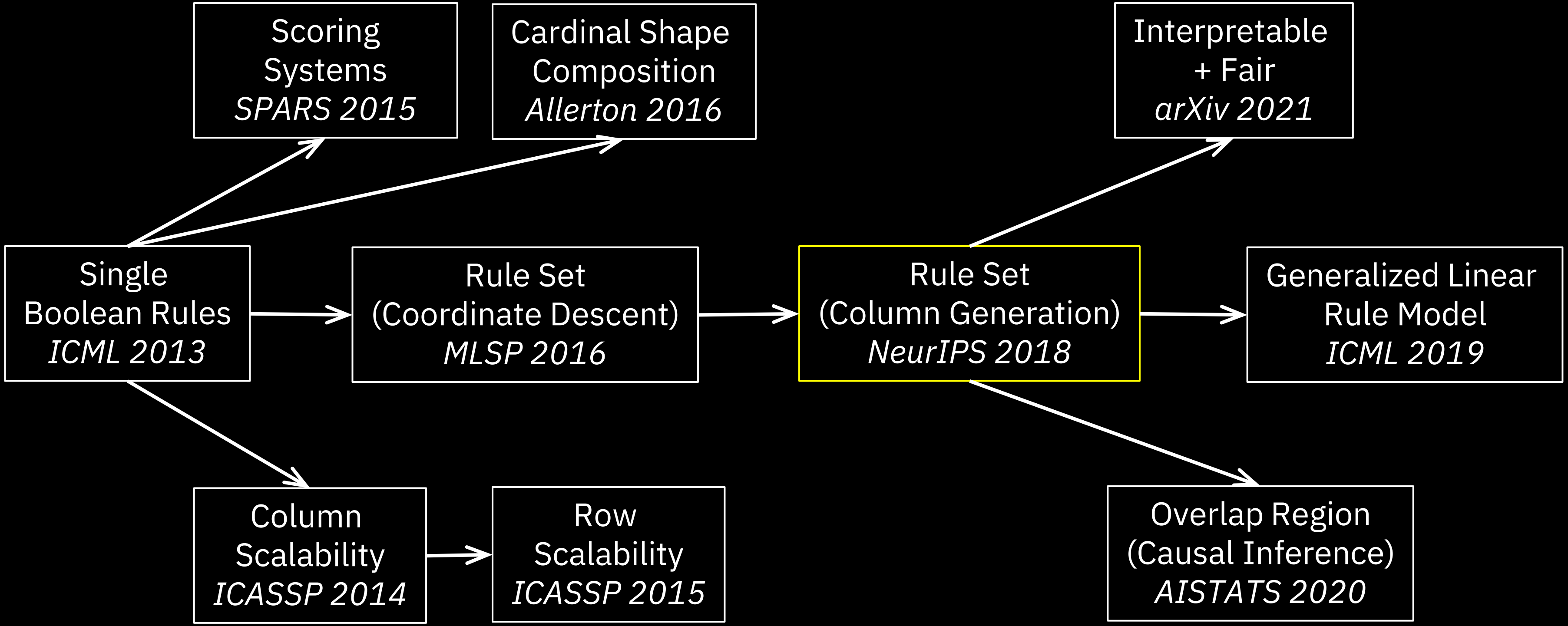
(NumSatTrades ≥ 23) AND (ExtRiskEstimate ≥ 70) AND (NetFracRevolvBurden ≤ 63)

OR

(NumSatTrades ≤ 22) AND (ExtRiskEstimate ≥ 76) AND (NetFracRevolvBurden ≤ 78)

1st Place Winner of FICO Explainable Machine Learning Challenge

Our rule learning agenda



Our rule learning agenda

A. Emad, K. R. Varshney, and D. M. Malioutov. “A Semiquantitative Group Testing Approach for Learning Interpretable Clinical Prediction Rules.” *Signal Processing with Adaptive Sparse Structured Representations Workshop*, Jul. 2015.

K. R. Varshney. “Interpretable Machine Learning via Convex Cardinal Shape Composition.” *Allerton Conference on Communication, Control, and Computing*, pp. 327–330, Sep. 2016.

C. Lawless, S. Dash, O. Günlük, and D. Wei. “Interpretable and Fair Boolean Rule Sets via Column Generation.” arXiv:2111.08466, Nov. 2021.

D. M. Malioutov and K. R. Varshney. “Exact Rule Learning via Boolean Compressed Sensing.” *International Conference on Machine Learning*, pp. 765–773, Jun. 2013.

G. Su, D. Wei, K. R. Varshney, and D. M. Malioutov. “Learning Sparse Two-Level Boolean Rules.” *IEEE Workshop on Machine Learning for Signal Processing*, Sep. 2016.

S. Dash, O. Günlük, and D. Wei. “Boolean Decision Rules via Column Generation.” *Advances in Neural Information Processing Systems*, pp. 4660–4670, Dec. 2018.

D. Wei, S. Dash, T. Gao, and O. Günlük. “Generalized Linear Rule Models.” *International Conference on Machine Learning*, pp. 6687–6696, Jun. 2019.

S. Dash, D. M. Malioutov and K. R. Varshney. “Screening for Learning Classification Rules via Boolean Compressed Sensing.” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3360–3364, May 2014.

S. Dash, D. M. Malioutov and K. R. Varshney. “Learning Interpretable Classification Rules Using Sequential Row Sampling.” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3337–3341, Apr. 2015.

M. Oberst, F. Johansson, D. Wei, T. Gao, G. Brat, D. Sontag, and K. R. Varshney. “Characterization of Overlap in Observational Studies.” *International Conference on Artificial Intelligence and Statistics*, pp. 788–798, Aug. 2020.

Challenges of rule learning

– Finding compact decision rules involving few Boolean terms that best approximate a given data set is an NP hard combinatorial optimization problem

– Old approaches maximize criteria such as information gain, support, confidence, lift, Gini impurity, etc.

- Decision trees, decision lists, RIPPER, SLIPPER, etc.
- Greedy heuristics with ad hoc pruning

– Renewed interest in rule learning driven by optimizing a principled objective, but which retains interpretability

Group testing problem

D. M. Malioutov and K. R. Varshney. "Exact Rule Learning via Boolean Compressed Sensing." *International Conference on Machine Learning*, pp. 765–773, Jun. 2013.

- Discover a sparse subset of faulty items in a large set of mostly good items using a few pooled tests
  - Blood screening of large groups of army recruits
  - Computational biology
  - Fault discovery in computer networks
- Mix together the blood of several recruits
  - If test is negative, none of the recruits are diseased
  - If test is positive, at least one of the recruits is diseased
  - Logical OR operation
- Construct the pools in an intelligent way to require a small number of tests with perfect recovery of diseased individuals

Rule learning as group testing

D. M. Malioutov and K. R. Varshney. "Exact Rule Learning via Boolean Compressed Sensing." *International Conference on Machine Learning*, pp. 765–773, Jun. 2013.

– Standard supervised binary classification problem

- $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\}$ with features $\boldsymbol{x}_i \in X$ and Boolean labels $y_i \in \{0,1\}$

– Construct individual Boolean clauses from features $a_j(\boldsymbol{x}) \in \{0,1\}$, for $j = 1, \dots, n$

- NumSatTrades ≥ 23
- CompensationPlan == 'quota-based'
- For continuous dimensions of X, make comparisons to set of thresholds

– Calculate the truth value of each Boolean term for each training sample to construct an $m \times n$ truth table matrix $\boldsymbol{A}$ with entries $a_{ij} = a_j(\boldsymbol{x}_i)$

Rule learning as group testing (continued)

D. M. Malioutov and K. R. Varshney. "Exact Rule Learning via Boolean Compressed Sensing." *International Conference on Machine Learning*, pp. 765–773, Jun. 2013.

- The positive training samples are now equivalent to diseased pools of army recruits
- Determine an $n \times 1$ Boolean coefficient vector $\boldsymbol{w}$ that specifies which Boolean terms a_j to OR together in a decision rule to recover the positive samples
- Learn $\boldsymbol{w}$ so that $\boldsymbol{y} \approx \boldsymbol{A} \vee \boldsymbol{w}$, where Boolean notation means:

$$y_i = \bigvee_{j=1}^{n} a_{ij} \wedge w_j$$

- Other papers go deeper into integer programming

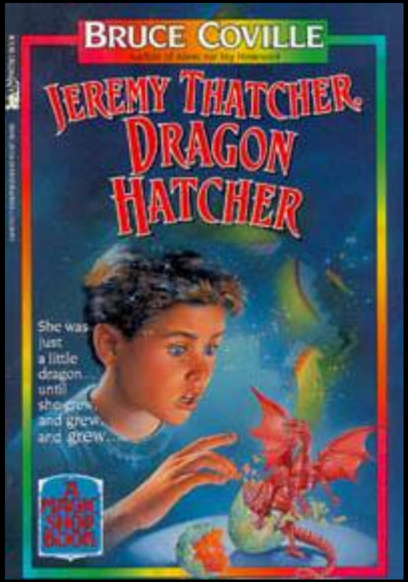
Newest work: Certifying safety

D. Wei, R. Nair, A. Dhurandhar, K. R. Varshney, E. M. Daly, and M. Singh. “On the Safety of Interpretable Machine Learning: A Maximum Deviation Approach.” Under review, 2022.

- Mathematical formulation for assessing the safety of supervised learning models based on their maximum deviation over a certification set
- For interpretable models including decision trees, rule lists, generalized linear and additive models, the maximum deviation can be computed exactly and efficiently
- Interpretability produces tighter bounds on the maximum deviation compared with black box functions

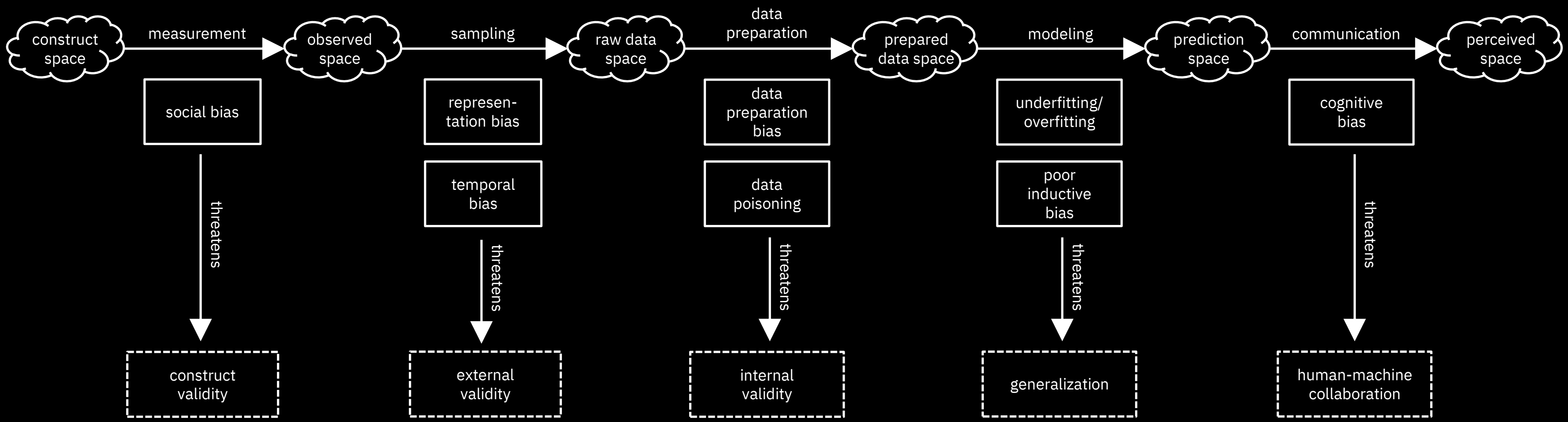
2. Teaming

When you create a Human+AI team, the hard part isn't the 'AI'. It isn't even the 'Human'. It's the '+'. (Case, 2018)



Big picture of trustworthy machine learning

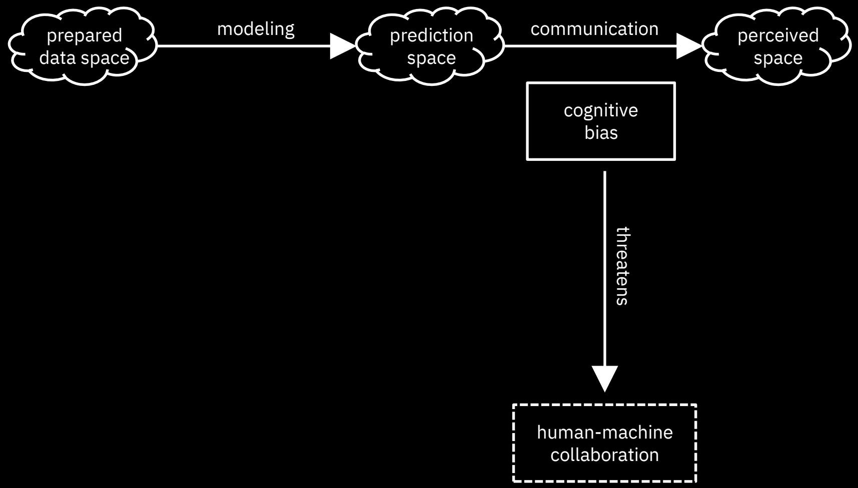
K. R. Varshney. *Trustworthy Machine Learning*. Independently Published, 2022.



Collaboration requires communication

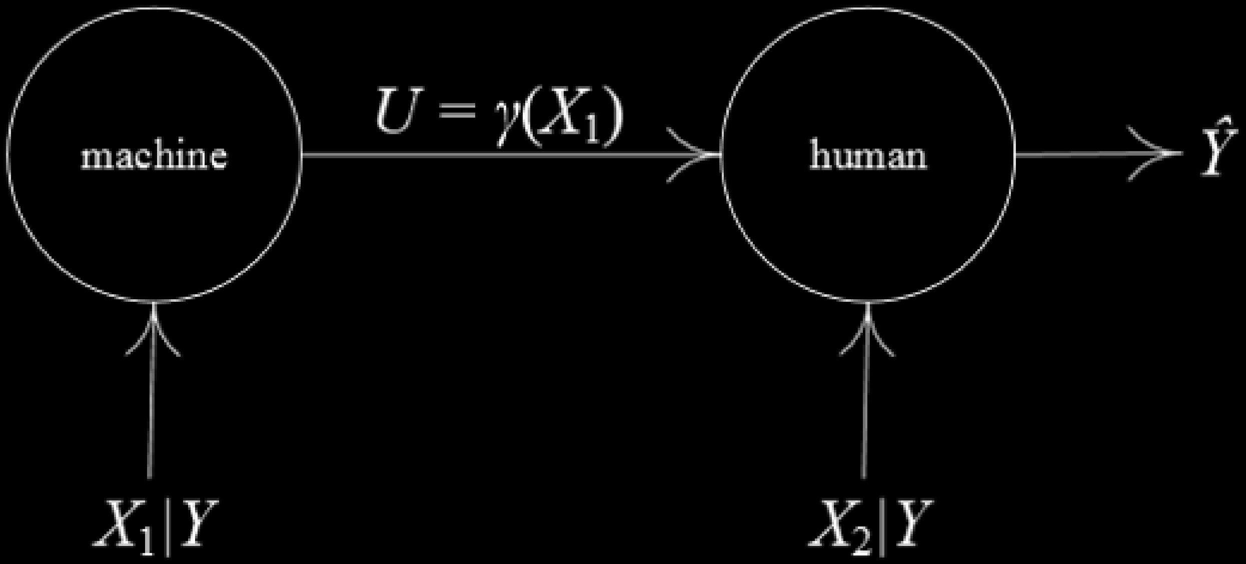
K. R. Varshney. *Trustworthy Machine Learning*. Independently Published, 2022.

- Interaction is mainly a communication problem
  - Last mile problem
- The end consumer of model predictions is a person with their own local observations and cognitive biases
- Model explainability is a problem of communicating a quantized variable that the human consumer fuses with their own information to make a final decision



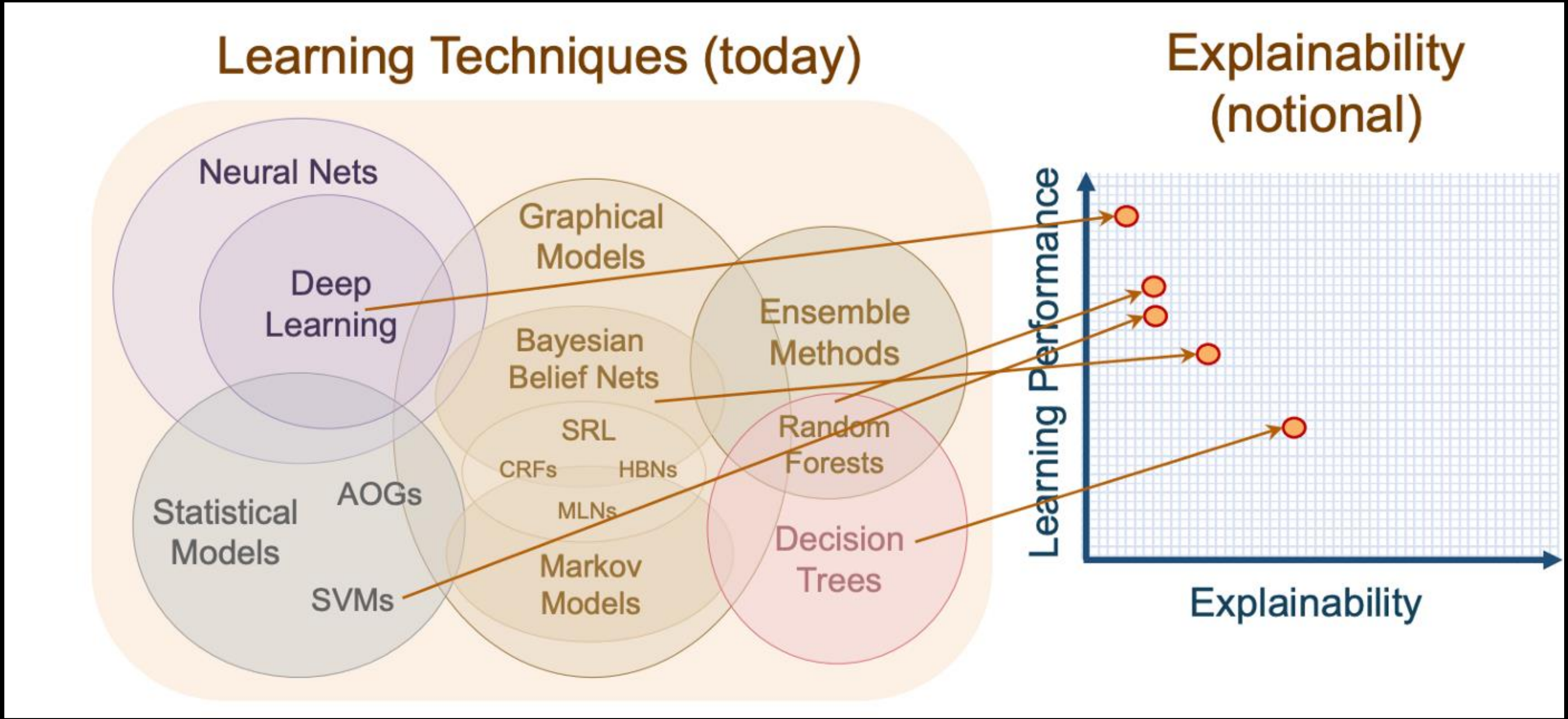
Human and machine collaboration

K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. “Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory.” *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.



Is this tradeoff true or false?

K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. "Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory." *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.



Let's use information theory

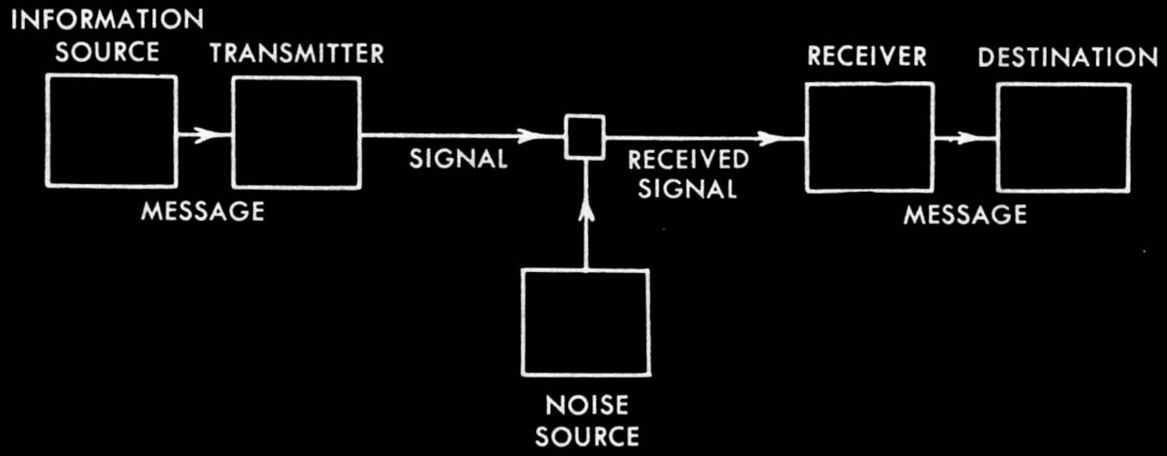


Fig. 1. — Schematic diagram of a general communication system.

Plan: Treat as a distributed detection problem

K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. "Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory." *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.

Model the model output as a multilevel quantizer

- 2 levels (1 bit) is a black box model
- More than 2 levels (but not too many) is an interpretable model

Analyze the overall accuracy of the human and machine collaboration, not just the machine in isolation

Prove that the system with more than 2 levels has higher Chernoff information and thus higher accuracy

Distributed detection theory

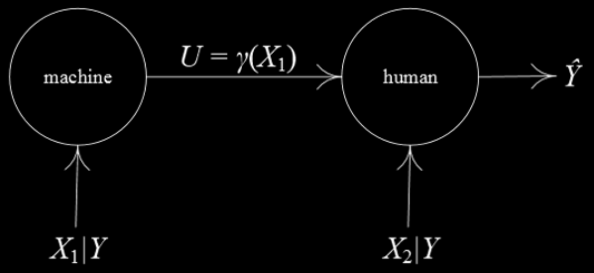
K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. "Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory." *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.

Bayes-optimal decision rules

Classical detection theory assumes that complete observations are available at a central processor for decision-making

Distributed detection: observations are processed in a distributed manner and decisions are made at the distributed processors, or processed data (compressed observations) are conveyed to a fusion center that makes the global decision

Setup



Binary classification problem with labels Y

Features X_1 observed by machine and X_2 observed by human

Independent conditioned on Y

U is an optimally-quantized version of an optimal classification based on X_1 to K levels

$\hat{Y}$ is the final classification based on U and X_2

Theorems

K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. "Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory." *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.

Consider two learnable two-node networks as described with different numbers of quantizer levels K and K' with $K' > K$ and corresponding quantized transmissions U and U'. Then, the following relationship among Chernoff informations holds:

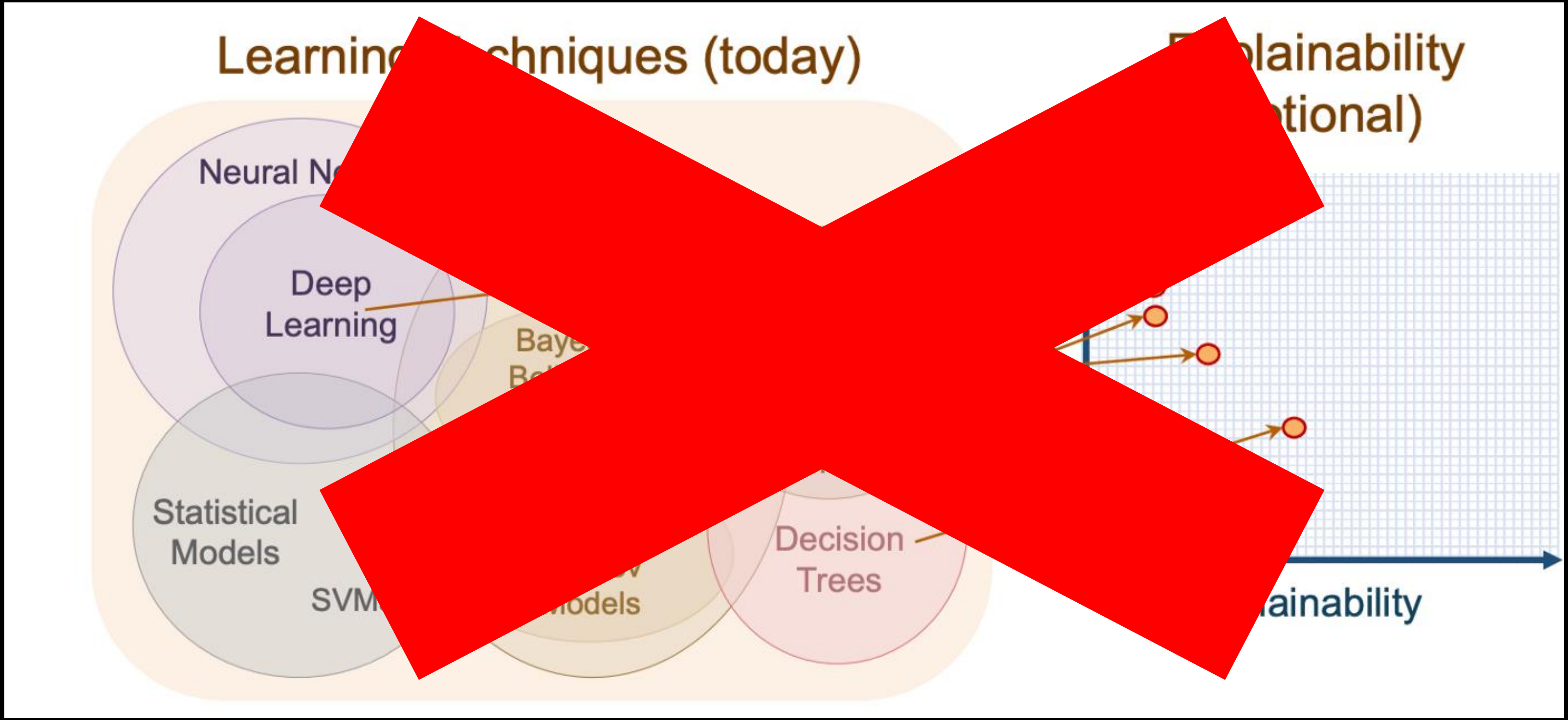
$$C\left(f_{U',X_2|Y}(u',x_2|y=1)||f_{U',X_2|Y}(u',x_2|y=0)\right) > C\left(f_{U,X_2|Y}(u,x_2|y=1)||f_{U,X_2|Y}(u,x_2|y=0)\right)$$

The best achievable exponent in the Bayesian probability of error in a binary classification problem with class labels Y and features X is $C\left(f_{X|Y}(x|y=1)||f_{X|Y}(x|y=0)\right)$

The probability of error in the two-node network as described with $K = 2$ quantizer levels is larger than the network with $K' > 2$ quantizer levels

This tradeoff is false for team performance

K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. “Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory.” *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.



Limitations

K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. “Why Interpretability in Machine Learning? An Answer Using Distributed Detection Theory.” *ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, Jul. 2018.

We do not intend to imply that more quantization levels leads to more interpretability

Assumes conditionally independent observations between human and machine

Population setting implies all models have the same optimal accuracy

This stylized abstraction does not differentiate between a truly interpretable model (e.g. decision list) and the quantization of a score function of a black box with probabilistic outputs

- A call for human-centered explainability

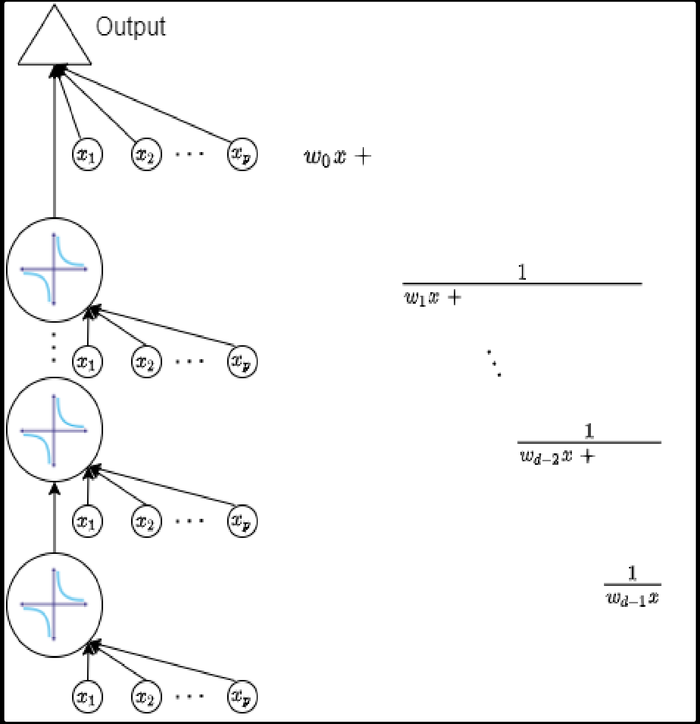
Human-centered explainability

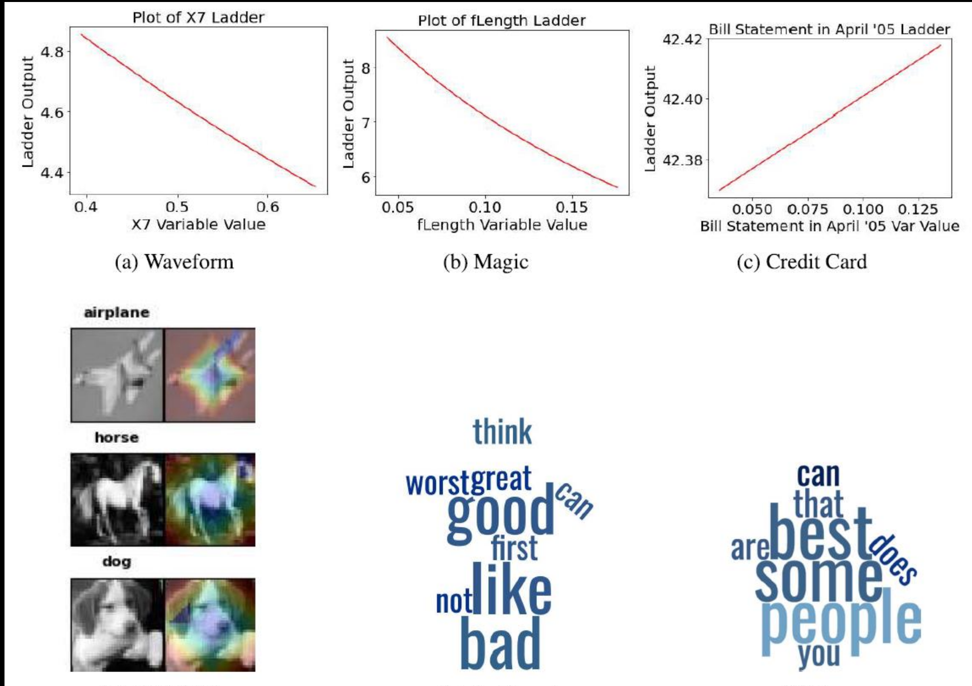
Q. Vera Liao and K. R. Varshney. "Human-Centered Explainable AI (XAI): From Algorithms to User Experiences." arXiv:2110.10790, Oct. 2021.

Question	Ways to explain	Example XAI methods
How (global model-wide)	• Describe the general model logic as feature impact*, rules† or decision-trees‡ • If user is only interested in a high-level view, describe what are the top features or rules considered	ProfWeight*†‡ [28], Global feature importance* [71, 105], Global feature inspection plots* (e.g. PDP [49]), Tree surrogates‡ [25]
Why (a given prediction)	• Describe how features of the instance, or what key features, determine the model's prediction of it* • Or describe rules that the instance fits to guarantee the prediction† • Or show similar examples with the same predicted outcome to justify the model's prediction‡	LIME* [89], SHAP* [72], LOCO* [63], Anchors† [90], ProtoDash‡ [47]
Why Not (a different prediction)	• Describe what features of the instance determine the current prediction and/or with what changes the instance would get the alternative prediction* • Or show prototypical examples that have the alternative outcome†	CEM* [27], Counterfactuals* [69], ProtoDash† (on alternative prediction) [47]
How to Be That (a different prediction)	• Highlight feature(s) that if changed (increased, decreased, absent, or present) could alter the prediction to the alternative outcome, with minimum effort required* • Or show examples with minimum differences but had the alternative outcome†	CEM* [27], Counterfactuals* [69], Counterfactual instances† [100], DiCE† [78]
How to Still Be This (the current prediction)	• Describe features/feature ranges* or rules† that could guarantee the same prediction • Or show examples that are different from the instance but still had the same outcome	CEM* [27], Anchors† [90]
What if	• Show how the prediction changes corresponding to the inquired change of input	PDP [49], ALE [10], ICE [44]

How

I. Puri, A. Dhurandhar, T. Pedapati, K. Shanmugam, D. Wei, and K. R. Varshney. "CoFrNets: Interpretable Neural Architecture Inspired by Continued Fractions." *Advances in Neural Information Processing Systems*, Dec. 2021.





(a) Waveform (b) Magic (c) Credit Card

(d) CIFAR10 (e) Sentiment (f) Quora

Why

K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal. "Efficient Data Representation by Selecting Prototypes with Importance Weights." *IEEE International Conference on Data Mining*, pp. 260–269, Nov. 2019.

AI Explainability 360 - Demo

Not secure | aix360.mybluemix.net/explanation_emp#

IBM Research Trusted AI

Home Demo Resources Events Videos Community

Alice
Approved

Robert
Denied

Based on these similar applicants, the Loan Officer can more confidently decide to approve Alice's loan. For example, the Loan Officer sees that although Alice did have a delinquency, similar to Mia and Kate it was less than 60 days late (seen e.g. from features NumTrades60Ever2DerogPubRec and NumTrades90Ever2DerogPubRec), and similar to Mia it was 26 months ago (MSinceMostRecentDelq). Both Mia and Kate repaid on time. The Loan Officer also sees that Alice has not opened any accounts in the last 12 months (NumTradesOpeninLast12M), like Mia and Cala, or had any recent inquiries to her credit file (NumInqLast6M), like Kate and Cala. These suggest that Alice is not desperate for credit due to financial difficulty.

Click on the features listed in the chart below to learn more.

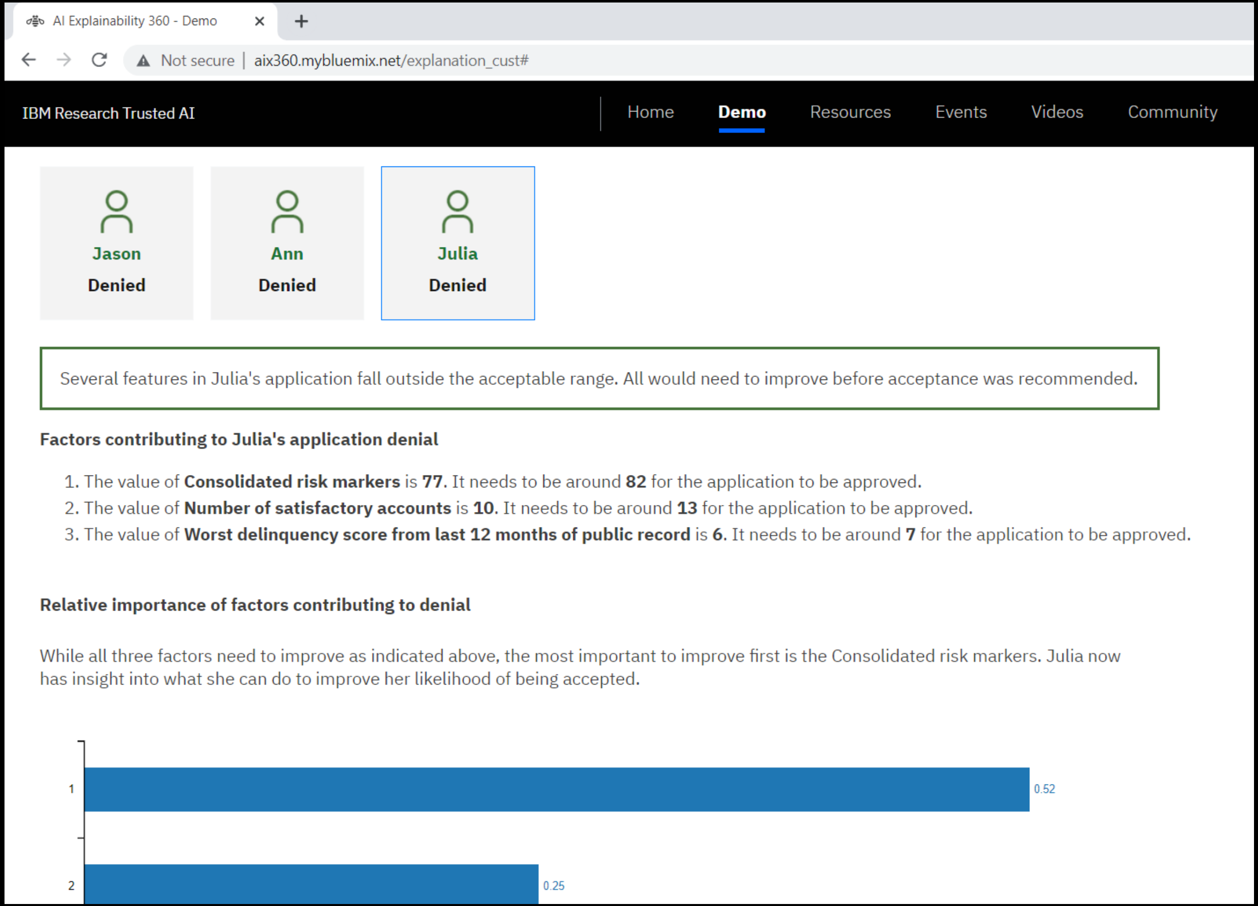
Customers similar to Alice and their repayment outcome.

Highlighted feature values match Alice's.

	Alice	Mia	Kate	Cala
Outcome	-	Paid	Paid	Paid
Similarity to Alice (from 0 to 1)	-	0.765	0.081	0.065
ExternalRiskEstimate	82	85	80	89
MSinceOldestTradeOpen	280	223	382	379
MSinceMostRecentTradeOpen	13	13	4	156

How to be that

A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. "Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives." *Advances in Neural Information Processing Systesms*, pp .590–601, Dec. 2018.



Other considerations for teaming

– Ask the user population to provide training explanations in their own language

– Give humans more time to overcome cognitive biases such as anchoring

– Play to the complementary strengths of humans and machines

- Task definition
- Input
- Internal processing
- Output

M. Hind, D. Wei, M. Campbell, N. C. F. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, and K. R. Varshney. "TED: Teaching AI to Explain Its Decisions." *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123–129, Jan. 2019.

C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett. "Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making." *ACM Conference on Computer-Supported Collaborative Work and Social Computing*, Nov. 2022.

C. Rastogi, Liu L., K. Holstein, and H. Heidari. "A Unifying Framework for Combining Complementary Strengths of Humans and ML Toward Better Predictive Decision-Making." arXiv:2204.10806, Apr. 2022.

Attributes of trustworthiness

K. R. Varshney. “On Mismatched Detection and Safe, Trustworthy Machine Learning.” *Conference on Information Sciences and Systems*, Mar. 2020.

		safety		teaming	
	Source	Attribute 1	Attribute 2	Attribute 3	Attribute 4
trustworthy people	Mishra	competent	reliable	open	concerned
	Maister et al.	credibility	reliability	intimacy	low self-orientation
	Sucher and Gupta	competent	use fair means to achieve its goals	take responsibility for all its impact	motivated to serve others’ interests as well as its own
trustworthy AI	Toreini et al.	ability	integrity	predictability	benevolence
	Ashoori and Weisz	technical competence	reliability	understandability	personal attachment
		accuracy	distributional robustness; fairness; adversarial robustness	explainability; uncertainty communication; transparency; value alignment	social good; empowering

Thank you

Kush R. Varshney
Distinguished Research Staff Member and Manager
—
krvarshn@us.ibm.com | @krvarshney

http://aix360.mybluemix.net
http://uq360.mybluemix.net
https://ci360.mybluemix.net
http://aifs360.mybluemix.net