Explainable AI in Healthcare

Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, Vikas Kumar

KenSci Inc

"Does your car have any idea why my car pulled it over?"

PAUL NOTH

Learning Objectives

- Why do we need explanations in AI and Machine Learning in Healthcare?
- What are Explanations in AI and Machine Learning in Healthcare?
- How does one choose between machine learning algorithms when explanations are needed?
- What are the different types of interpretable machine learning models?
- What does the future of explainable AI looks like in healthcare?

Doshi-Velez 2014

Terminology

Explanation vs. Interpretation vs. Understanding vs. Comprehension?

- **Explain:** Make (an idea or situation) clear to someone by describing it in more detail or revealing relevant facts.
- **Interpret:** Explain the meaning of (information or actions)
- **Understand:** Perceive the intended meaning of (words, a language, or a speaker)
- **Comprehend:** Grasp mentally; understand

[Oxford Dictionary]

Craik 1967, Doshi-Velez 2014

Explainable Machine Learning

- Interpretable machine learning refers to giving **explanations** of machine learning models to **humans** with **domain knowledge**
- Explanation: Why is the prediction being made?
- Explanation to Human: The explanation should be comprehensible to humans in (i) natural language (ii) easy to understand representations
- Domain Knowledge: The explanation should make sense to a domain expert

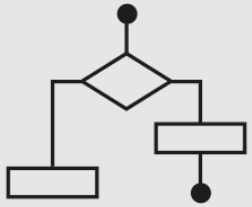
$$\mathcal{L}_{SM} = -\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc} f^{ade} g^b_\mu g^c_\nu g^d_\mu g^e_\nu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - igc_w(\partial_\nu Z^0_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - Z^0_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + Z^0_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - igs_w(\partial_\nu A_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + A_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - \tfrac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \tfrac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2 (Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + g^2 s_w^2 (A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w (A_\mu Z^0_\nu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu) - \tfrac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \tfrac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \beta_h \left(\tfrac{2M^2}{g^2} + \tfrac{2M}{g}H + \tfrac{1}{2}(H^2 + \phi^0\phi^0 + 2\phi^+\phi^-)\right) + \tfrac{2M^4}{g^2}\alpha_h - g\alpha_h M\left(H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-\right) - \tfrac{1}{8}g^2\alpha_h\left(H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2H^2\right) - gMW^+_\mu W^-_\mu H - \tfrac{1}{2}g\tfrac{M}{c_w^2}Z^0_\mu Z^0_\mu H - \tfrac{1}{2}ig\left(W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)\right) + \tfrac{1}{2}g\left(W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) + W^-_\mu(H\partial_\mu\phi^+ - \phi^+\partial_\mu H)\right) + \tfrac{1}{2}g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) + M\left(\tfrac{1}{c_w}Z^0_\mu\partial_\mu\phi^0 + W^+_\mu\partial_\mu\phi^- + W^-_\mu\partial_\mu\phi^+\right) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + igs_wMA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + igs_wA_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac{1}{4}g^2W^+_\mu W^-_\mu\left(H^2 + (\phi^0)^2 + 2\phi^+\phi^-\right) - \tfrac{1}{8}g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu\left(H^2 + (\phi^0)^2 + 2(s_w^2 - 1)^2\phi^+\phi^-\right) - \tfrac{1}{2}g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) - \tfrac{1}{2}ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac{1}{2}g^2s_wA_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) + \tfrac{1}{2}ig^2s_wA_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2 - 1)Z^0_\mu A_\mu\phi^+\phi^- - g^2s_w^2A_\mu A_\mu\phi^+\phi^- + \tfrac{1}{2}ig_s\lambda^a_{ij}(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu - \bar{e}^\lambda(\gamma\partial + m^\lambda_e)e^\lambda - \bar{\nu}^\lambda(\gamma\partial + m^\lambda_\nu)\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial + m^\lambda_u)u^\lambda_j - \bar{d}^\lambda_j(\gamma\partial + m^\lambda_d)d^\lambda_j + igs_wA_\mu\left(-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j\gamma^\mu d^\lambda_j)\right) + \tfrac{ig}{4c_w}Z^0_\mu\{(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{e}^\lambda\gamma^\mu(4s_w^2 - 1 - \gamma^5)e^\lambda) + (\bar{d}^\lambda_j\gamma^\mu(\tfrac{4}{3}s_w^2 - 1 - \gamma^5)d^\lambda_j) + (\bar{u}^\lambda_j\gamma^\mu(1 - \tfrac{8}{3}s_w^2 + \gamma^5)u^\lambda_j)\} + \tfrac{ig}{2\sqrt{2}}W^+_\mu\left((\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)U^{lep}_{\lambda\kappa}e^\kappa) + (\bar{u}^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)\right) + \tfrac{ig}{2\sqrt{2}}W^-_\mu\left((\bar{e}^\kappa U^{lep\dagger}_{\kappa\lambda}\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\kappa\lambda}\gamma^\mu(1+\gamma^5)u^\lambda_j)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_e(\bar{\nu}^\lambda U^{lep}_{\lambda\kappa}(1-\gamma^5)e^\kappa) + m^\lambda_\nu(\bar{\nu}^\lambda U^{lep}_{\lambda\kappa}(1+\gamma^5)e^\kappa)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_e(\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa}(1+\gamma^5)\nu^\kappa) - m^\kappa_\nu(\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa}(1-\gamma^5)\nu^\kappa)\right) - \tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{\nu}^\lambda\nu^\lambda) - \tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{e}^\lambda e^\lambda) + \tfrac{ig}{2}\tfrac{m^\lambda_\nu}{M}\phi^0(\bar{\nu}^\lambda\gamma^5\nu^\lambda) - \tfrac{ig}{2}\tfrac{m^\lambda_e}{M}\phi^0(\bar{e}^\lambda\gamma^5 e^\lambda) - \tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\tilde{\nu}_\kappa - \tfrac{1}{4}\overline{\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\tilde{\nu}_\kappa} + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_d(\bar{u}^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + m^\lambda_u(\bar{u}^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_d(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m^\kappa_u(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1-\gamma^5)u^\kappa_j)\right) - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{u}^\lambda_j u^\lambda_j) - \tfrac{g}{2}\tfrac{m^\lambda_d}{M}H(\bar{d}^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m^\lambda_u}{M}\phi^0(\bar{u}^\lambda_j\gamma^5 u^\lambda_j) - \tfrac{ig}{2}\tfrac{m^\lambda_d}{M}\phi^0(\bar{d}^\lambda_j\gamma^5 d^\lambda_j)$$

Standard Model Lagrangian

Explainable ML is more than models

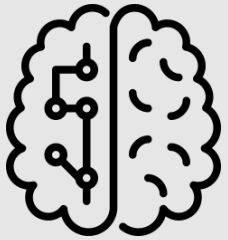
Machine Learning Solution

Features



Algorithm

Model Parameters

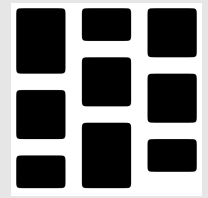


Model

Machine Learning User

Cognitive Capacity

Domain Knowledge



Explanation Granularity

Each element constituent of the machine learning solution process needs to be explainable for the solution to be truly explainable

Data type	Models/Tools	Applications
-EHR data -Insurance claims data	ML(logistic regression, XGBoost)	Predict outcomes (disease, death, readmission etc.)
-Clinical notes -Conversation text data	-Rule based approach(regular expression) -Deep learning approach	-Extract concepts from clinical notes -Knowledge graphs -Chat-bot -QA system
Medical image data (X-ray, CT, OCR image etc.)	CNN	-Detection: diagnosis of skin cancer lung nodule or diabetic reinopathy -Segmentation of tumor, histopathology
Time series data (EEG, ECG, vital sign data etc.)	HMM,RNN,CNN	-Heart disease -Sleep disorder(apnea) -ICU monitoring
Genomics data	GATK,QIIME	-Cancer mutation identification -Biomarker identification -Drug discovery
Other data (hospital operational data)	-ML(regression) -Queueing model	-Reduce operational cost -Improve patient experience -ER wait time and queueing

Need for Explanations in Machine Learning

When do we **need** explanations?

Al-Shedivat 2017B, Doshi-Velez 2014

When fairness is critical:

- Any context where humans are required to provide explanations so that people cannot hide behind machine learning models

When consequences are far-reaching:

- Predictions can have far reaching consequences e.g., recommend an operation, recommend sending a patient to hospice etc.

When the cost of a mistake is high:

- Ex: misclassification of a malignant tumor can be costly and dangerous

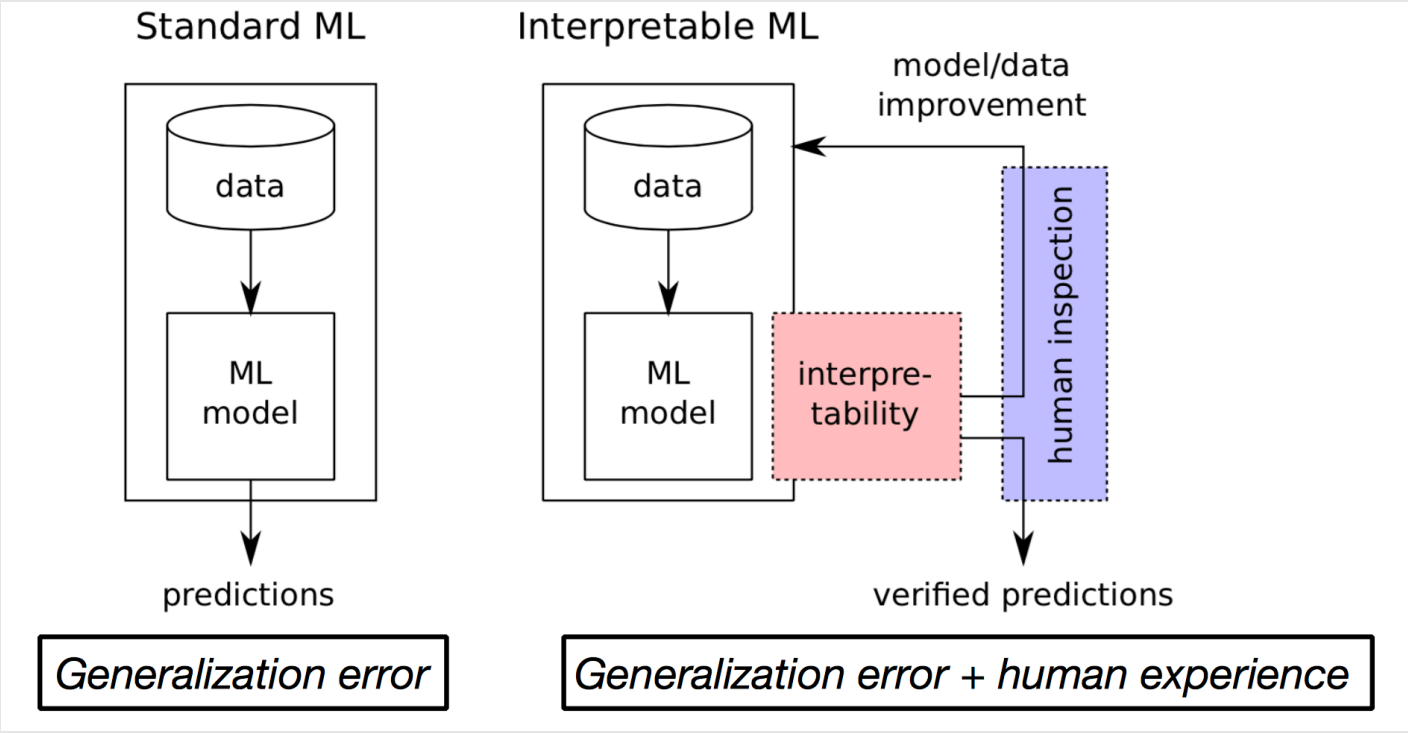
When a new/unknown hypothesis is drawn:

- *“It's not a human move. I've never seen a human play this move.” (Fan Hui)*
- Pneumonia patients with asthma had lower risk of dying (Caruana et al. 2015)

Doshi-Velez 2017

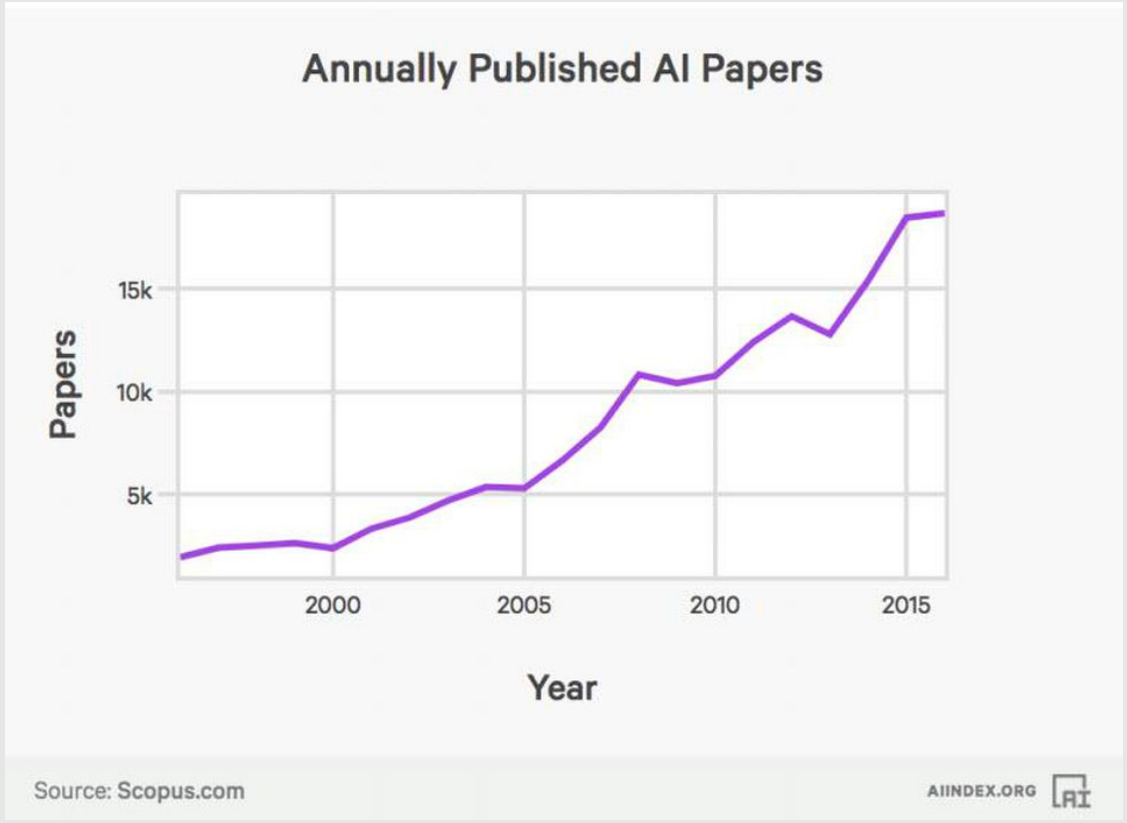
When do we need explanations?

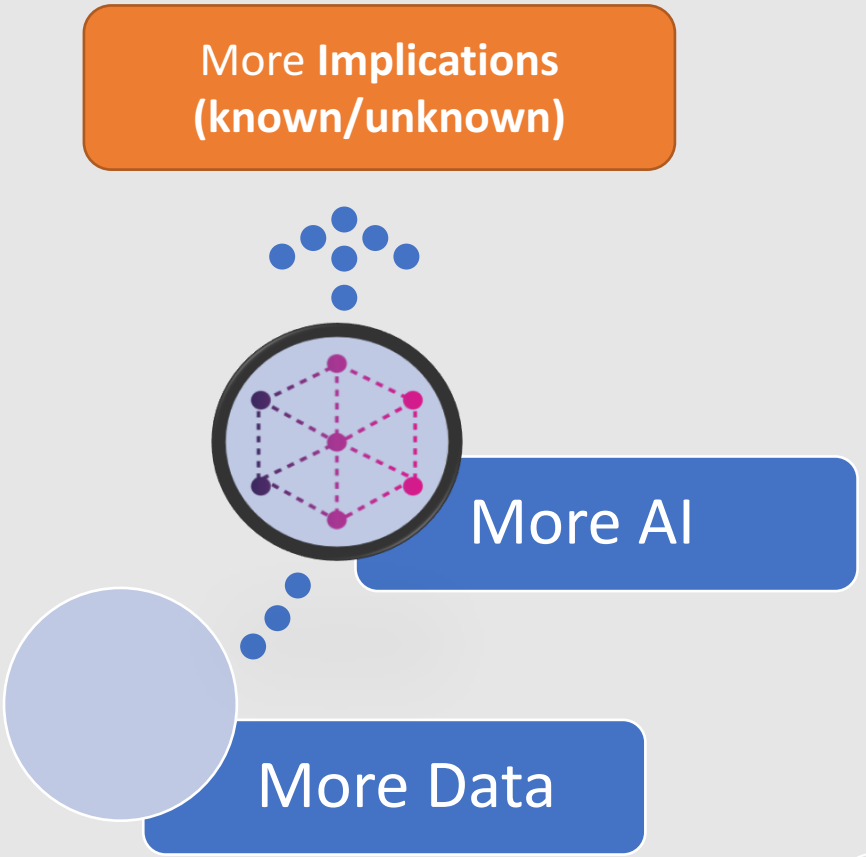
- **When performance is critical:**
- **When compliance is required:**
  - GDPR
  - Right to explanation
- **When trust is necessary:**
  - predictive performance is not enough



Why do we need explanations now?

Sampling of headlines about failures of AI in healthcare?

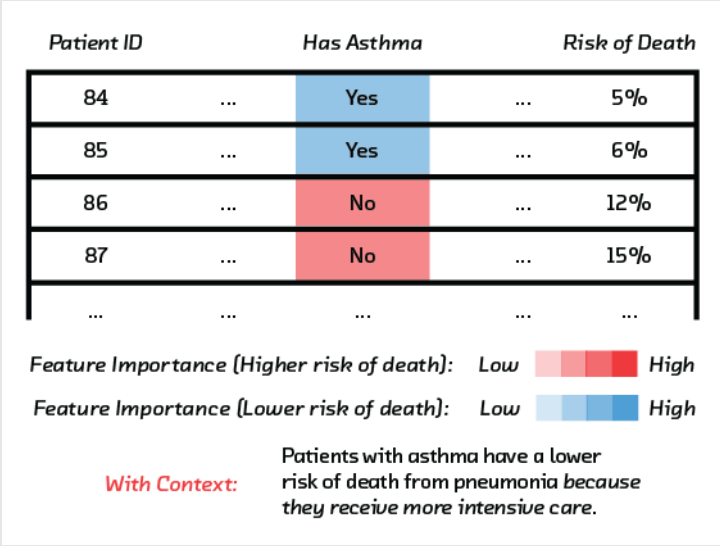




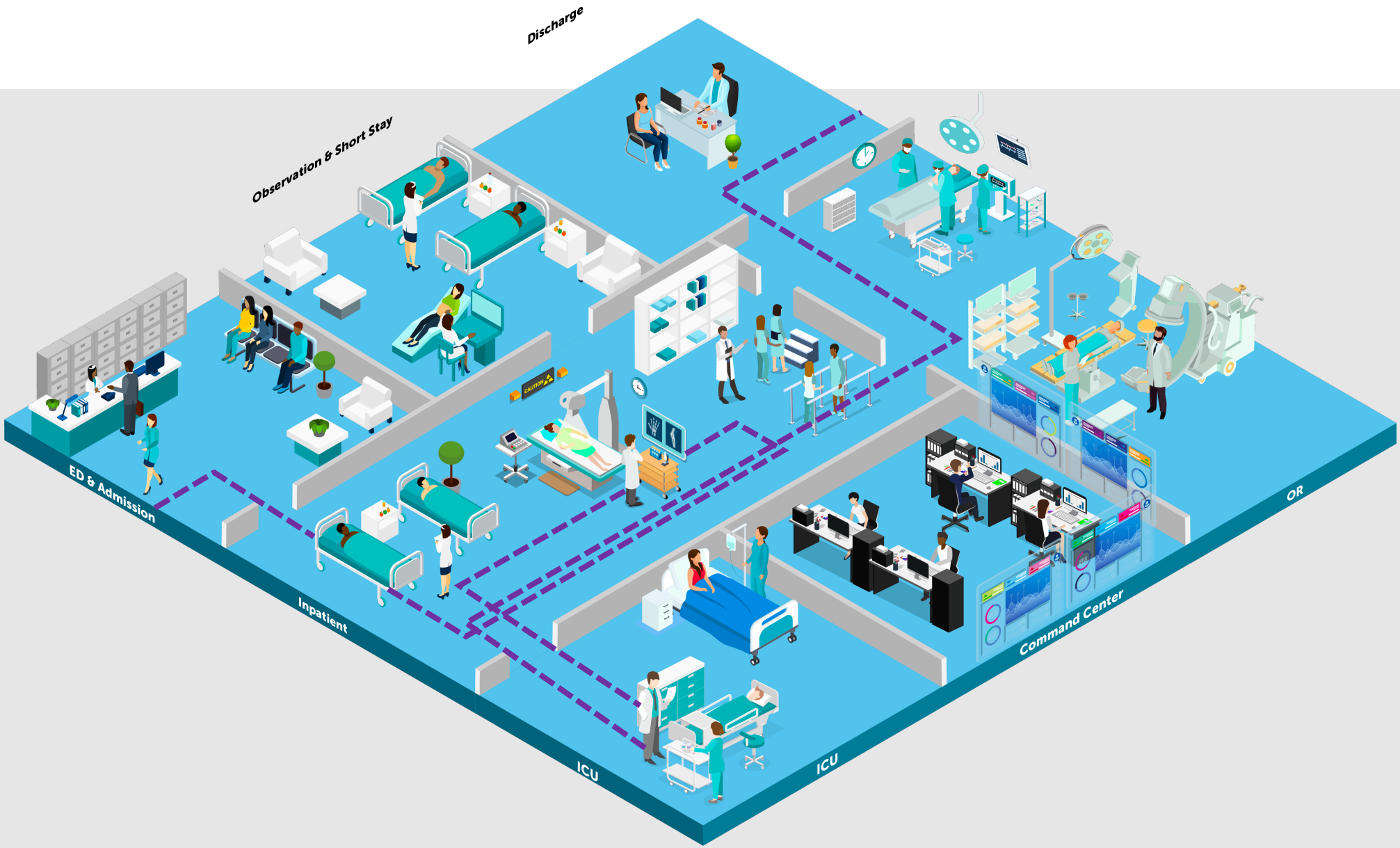
Caruna 2015, Caruna 2017

Need for ML Explanation in Healthcare

- Algorithms to predict which pneumonia patients should be admitted to hospital for treatment
- Neural nets were far more accurate than classical statistical methods
- The regression and the neural net inferred that asthma patients treated for pneumonia had a lower mortality risk, and therefore, should not be admitted
- In fact, due to their underlying lung condition, these patients were usually admitted directly to the ICU, treated aggressively, and survived



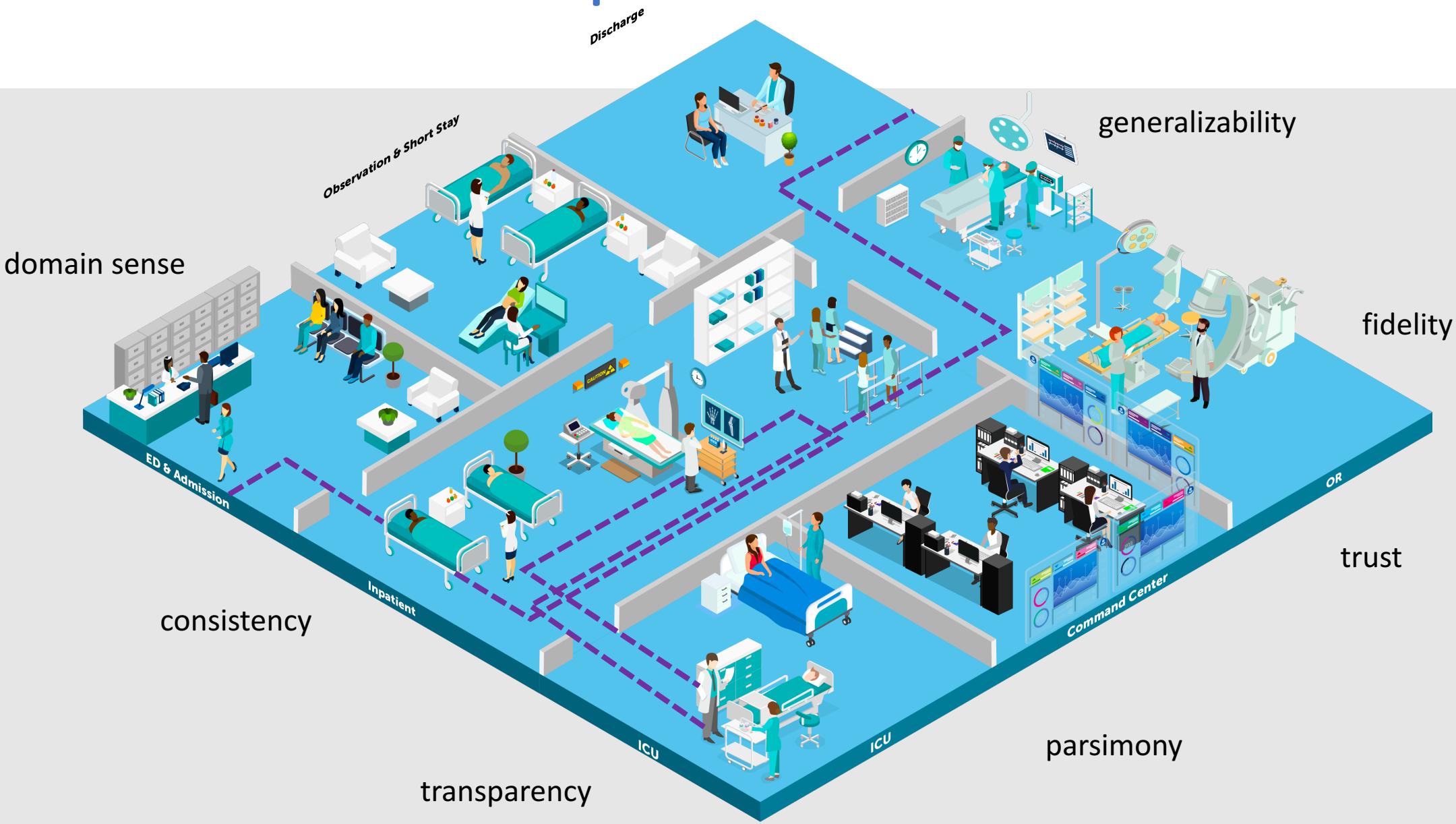
Problems in Patient Flow



Characteristics of Explainable AI in Healthcare

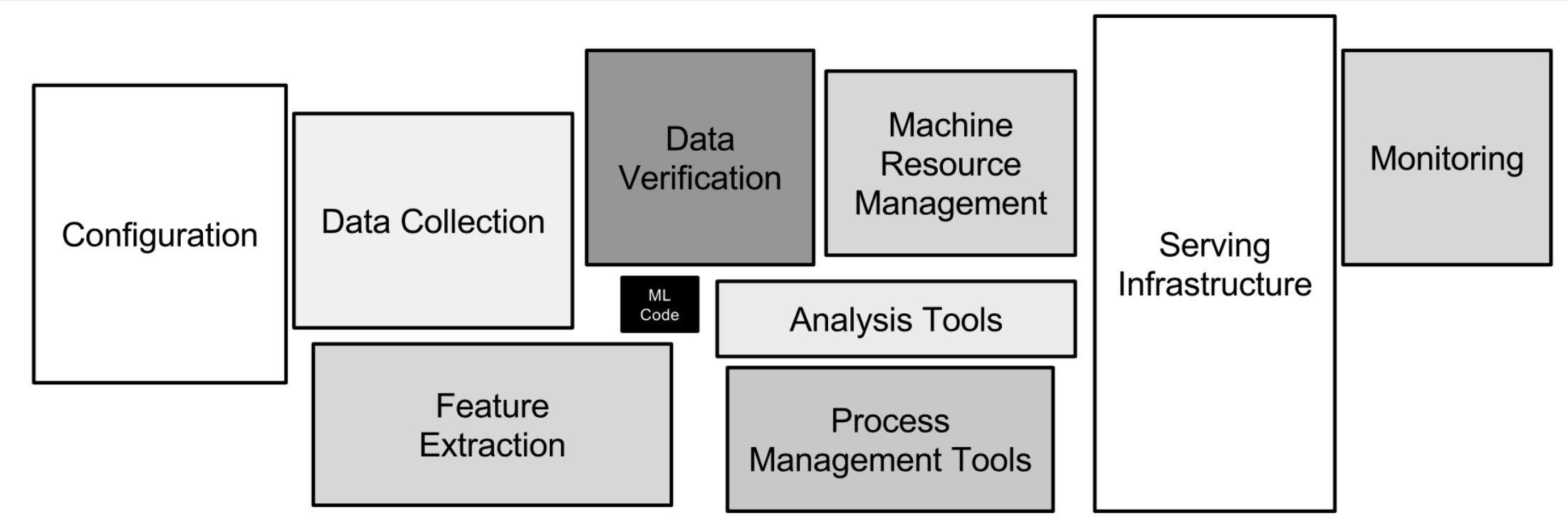
- Trust
- Transparency
- Fidelity
- Domain Sense
- Consistency
- Generalizability
- Parsimony

Characteristics of Explainable AI in Patient Flow



Sculley 2015

Operationalizing AI in Healthcare



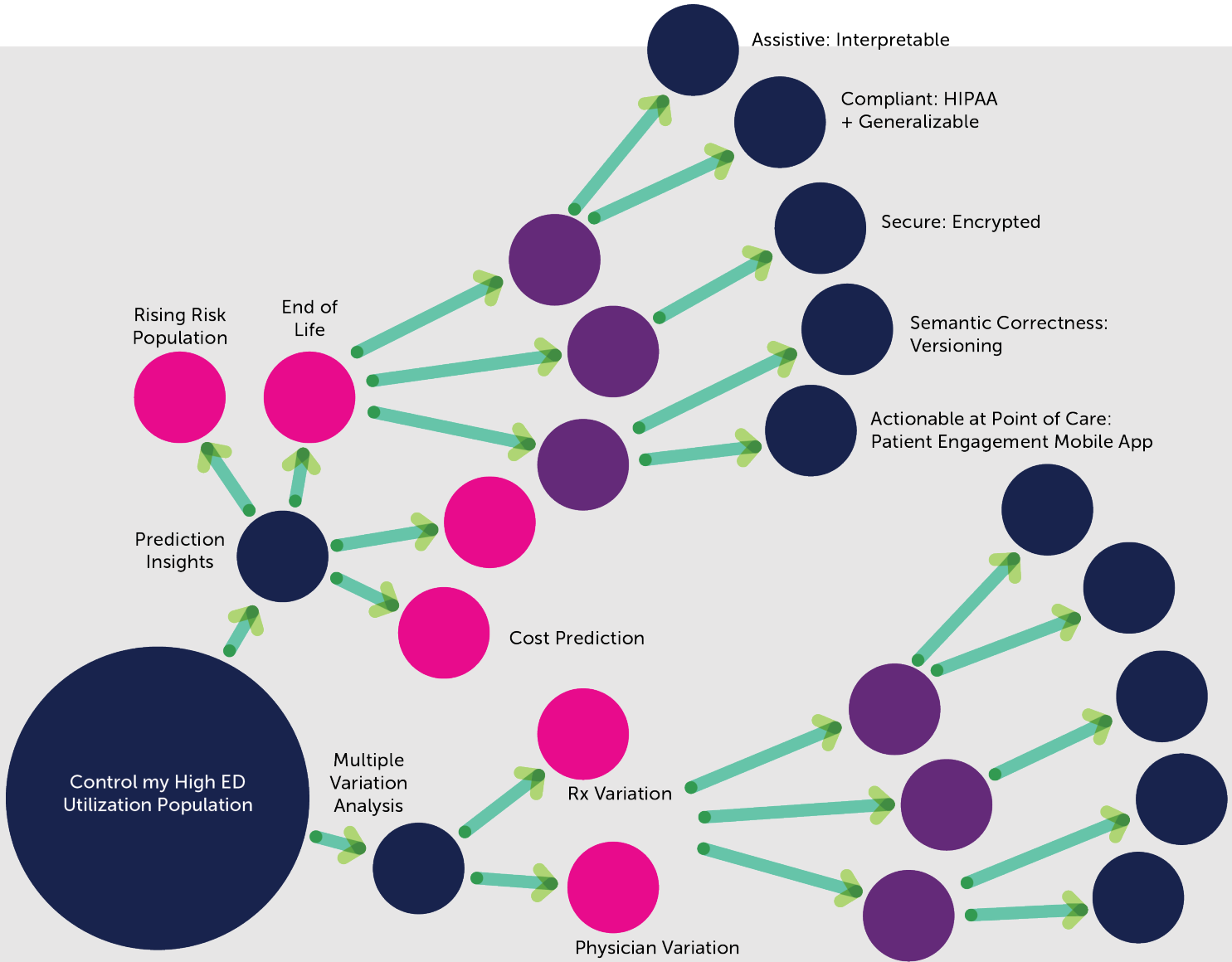
Only a small fraction of real-world machine learning systems actually constitutes machine learning code.

Data in Operationalized AI in Healthcare

- Syntactic Correctness
  - Is the data in the correct format
- Morphological Correctness
  - Is the data within the range of possible values e.g., a blood pressure of 500 does not make sense
- Semantic Correctness
  - Do the variables actually correspond to what the semantics that are being ascribed to them

Ahmad 2018

The Problem of Point Solutions



Anyone can do the math. An ML model alone doesn't solve a healthcare problem…

…and lots of models becomes a problem very quickly.

Admission Prediction

What is the likelihood of the patient being admitted to the hospital

Transparency

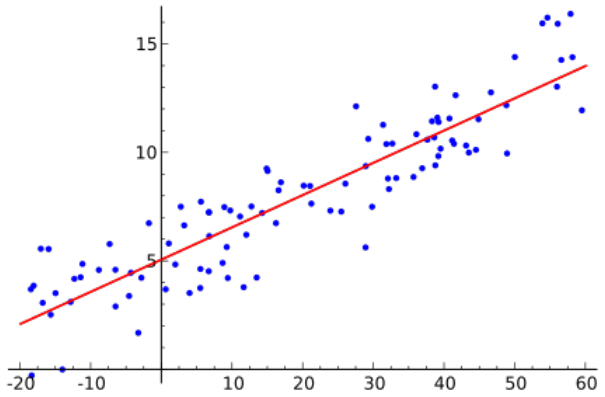
Ability of the machine learning algorithm, model and the features to be understand-able by the user of the system

Transparency

- The ML model for predicting Katherine's likelihood of admission gives her a high likelihood (0.62)
- Katherine's physician has noted her age, health history, and vital signs and is surprised by this elevated risk score
- The physician knows that the risk model is a deep learning model so he cannot understand how it is working
- But, he can examines the top factors associated with prediction

Transparent to whom?

- Transparency may mean different different things to different people
- Understanding Model Outputs:



- Understanding Algorithms:

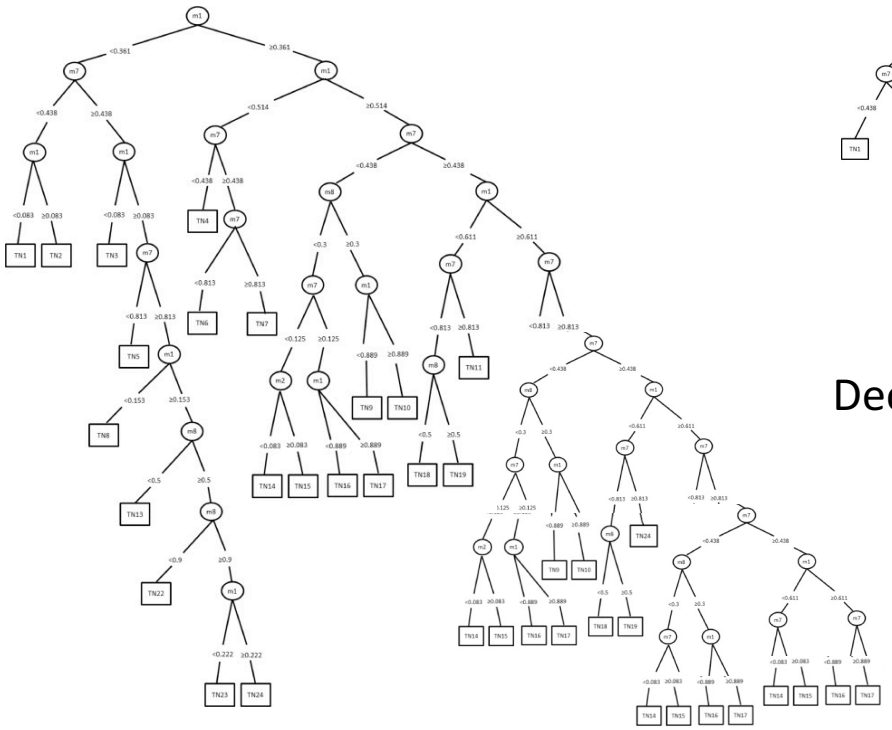
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, ..., n$$

- Understanding the algorithm may not mean be sufficient:

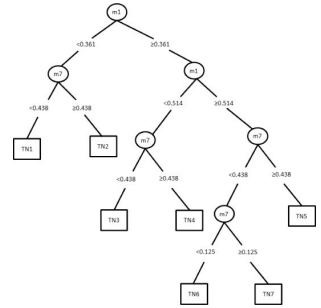
$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

Transparency: Simultability

- The whole model must be understandable simultaneously
- One must be able to look at the model and understand the whole model without too much cognizing about the model
- Example: While both Decision Trees are explainable, *Decision Tree B* has the property of Simultability but *Decision Tree A* does not



Decision Tree A



Decision Tree B

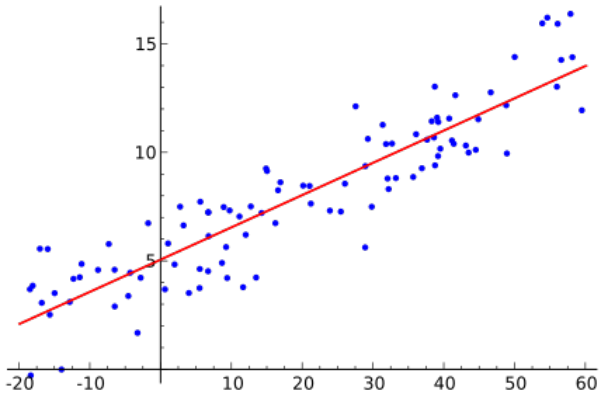
(Lipton 2016)

Transparency: Decomposability

- Each component should also admit to an easy/intuitive explanation
- A linear model with highly engineered features vs. a linear model with simple feature
- Example: Model A is decomposable but Model B is not

Regression Variables
- Age
- Gender
- Race
- Diabetic
- Smoker

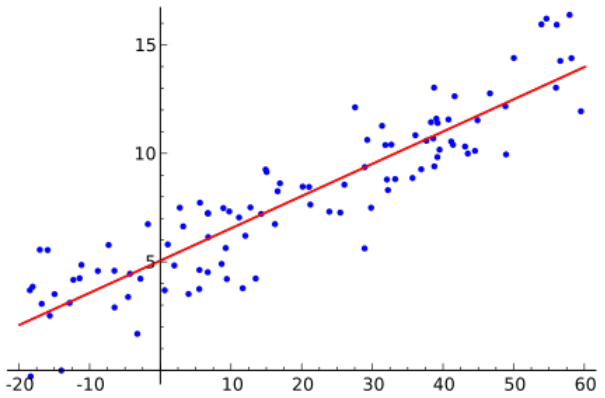
Target Variable
- Length of Stay



Regression Model A

Regression Variables
- $e^{\lambda t}\left(\cos(\mu t)+i\sin(\mu t)\right)$
- $\sum_{n=1}^{\infty} c_n (z-a)^n$
- $\frac{1}{z}+g'(z)+\sum_{n\neq 0}\left(\frac{1}{z-n}+\frac{1}{n}\right)$

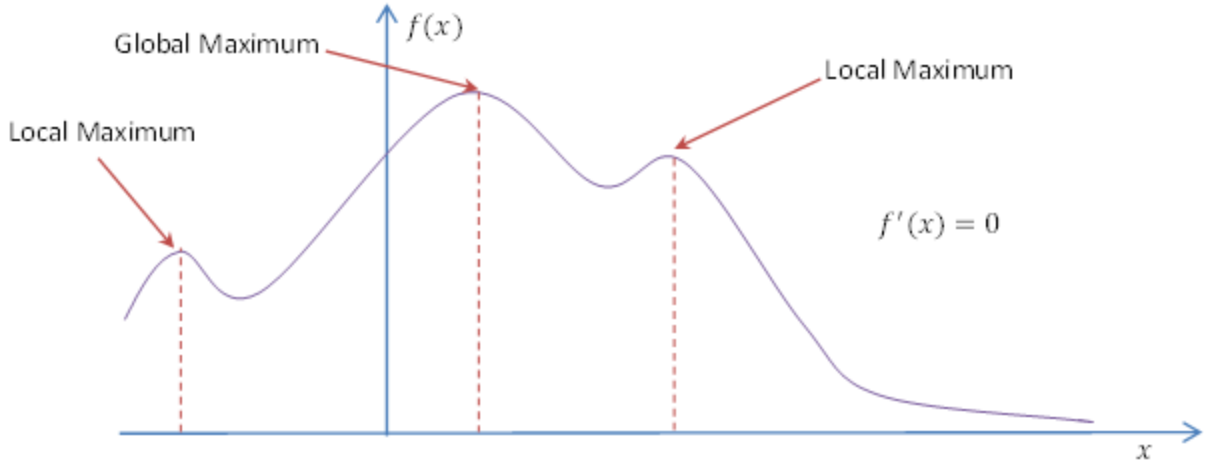
Target Variable
- Length of Stay



Regression Model B

Transparency: Algorithmic

- Guarantee that a model will converge
- Models like Regression Models, SVM etc. have this property
- Deep Learning does not have this property



Transparency: Feedback

- Feedback Transparency refers to how change in the model will affect the model prediction
- How do multi-objective optimization models affect each other e.g., optimizing reduction of risk of readmission and reduction of length of stay
- Not an absolute requirement but rather nice to have property
- Especially applicable to **Scrutable** machine learning systems

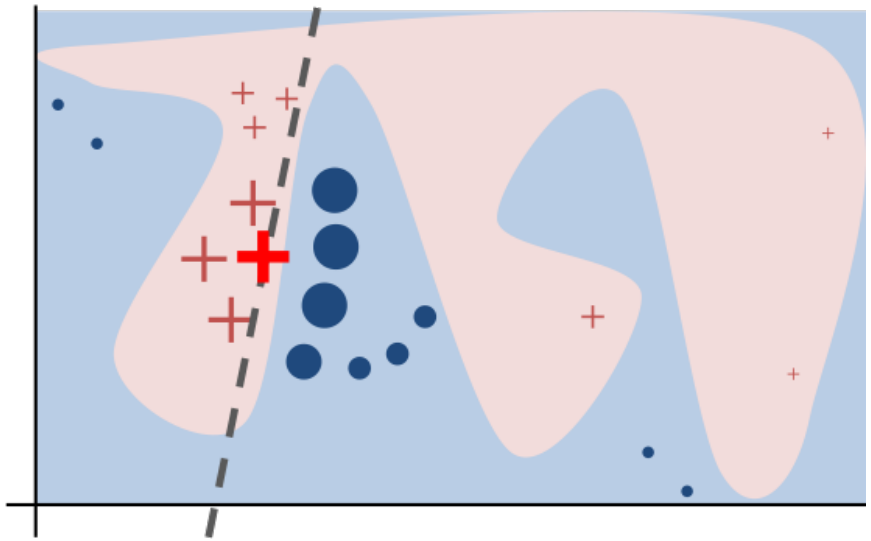
Transparency: Examples

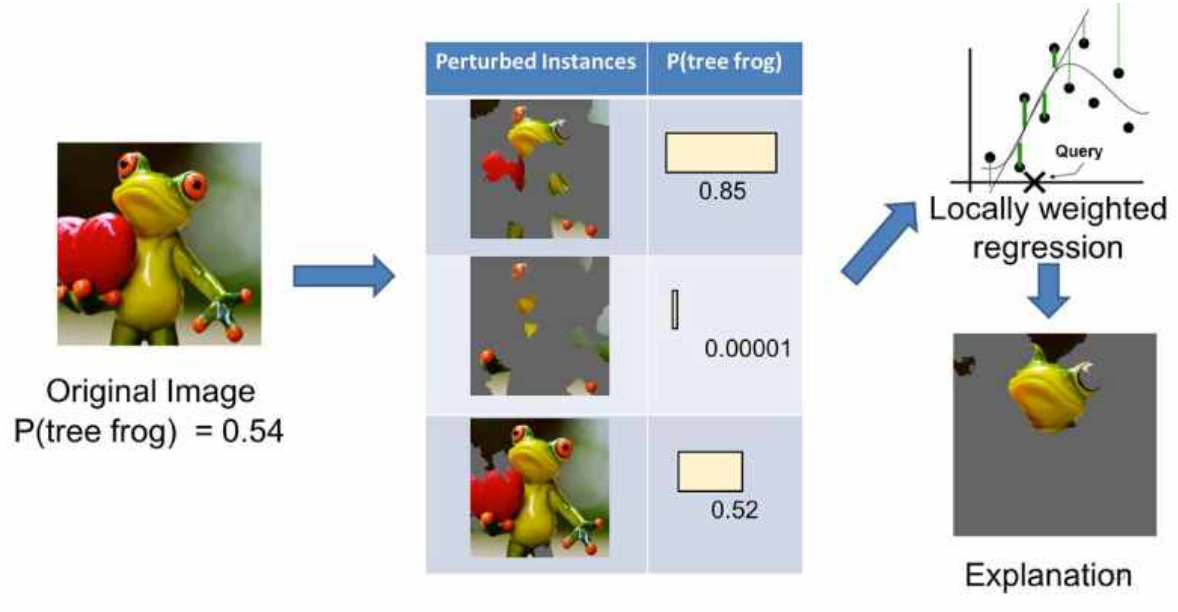
- **Transparent**
  - GAM
  - GA2M
  - Naïve Bayes
  - Regression Models
  - Falling Rule Lists
  - SLIM
- **Semi-Transparent**
  - Shallow Ensembles
- **Non-Transparent**
  - Deep Learning
  - SVM
  - Gradient Boosting Models

Locally Interpretable Model Explanations

(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 network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

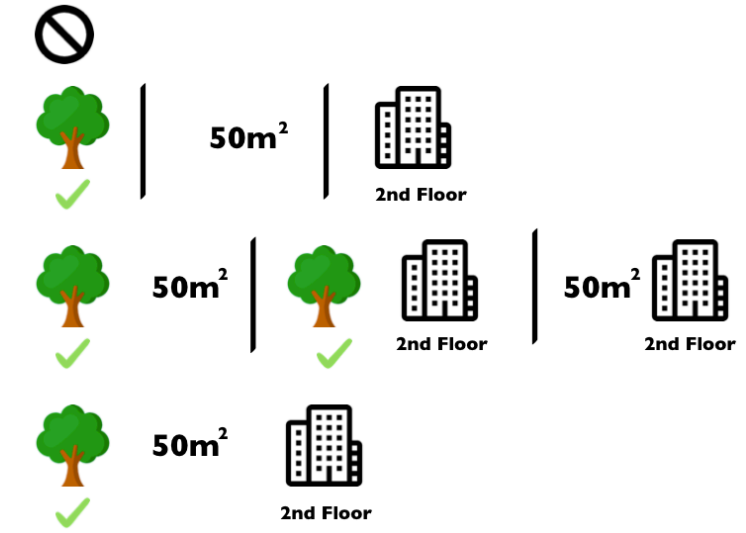




Shapley Values

- Game Theoretic Method for determining feature contributions
- Each feature is a ‘player’ in a game where the prediction is the payout
- The Shapley value - a method from coalitional game theory - tells us how to fairly distribute the ‘payout’ among the features.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].$$



Domain Sense

The explanation should make sense in the domain of application and to the user of the system

ED Census Prediction

Predict the number of patients in the ED at a given time

Domain Sense

Katherine and her physician discuss the problem of ED census prediction i.e., predicting the number of patients expected in the emergency department in a given time frame
The top features are temporal features (day of the week, month) which are not really helpful
The explanations are reconfigured to surface only factors that are modifiable
Whende, a nurse sees the new dashboard but the modifiable factors are not really helpful to her. Explanation dashboard is customized for Whende
In the aggregate the top reason for high ED census corresponds to people getting drunk because of a college football game
Proper staffing and proper stocking of supplies can be done with relevant explanations

Doshi-Velez 2014, Druzdzel 1996

Domain Sense: Explanations are role-based

- A physician requires different explanations as compared to a staffing planner
- Explanations need to be in the right language and also in the right context
- If actionability is required then the factors for explanations should reflect that

Taxonomy of Explainable Factors

Mutability	Interveniability	Actionability	Example
Immutable			Age, Sex, Ethnicity
Mutable	Non-Interveniable		Intrinsic Heart rate variability Marital Status
	Interveniable	Signal	Temperature (in Apendectomy)
		Intervention	Appendicitis
	Post-Interveniable		Immunization

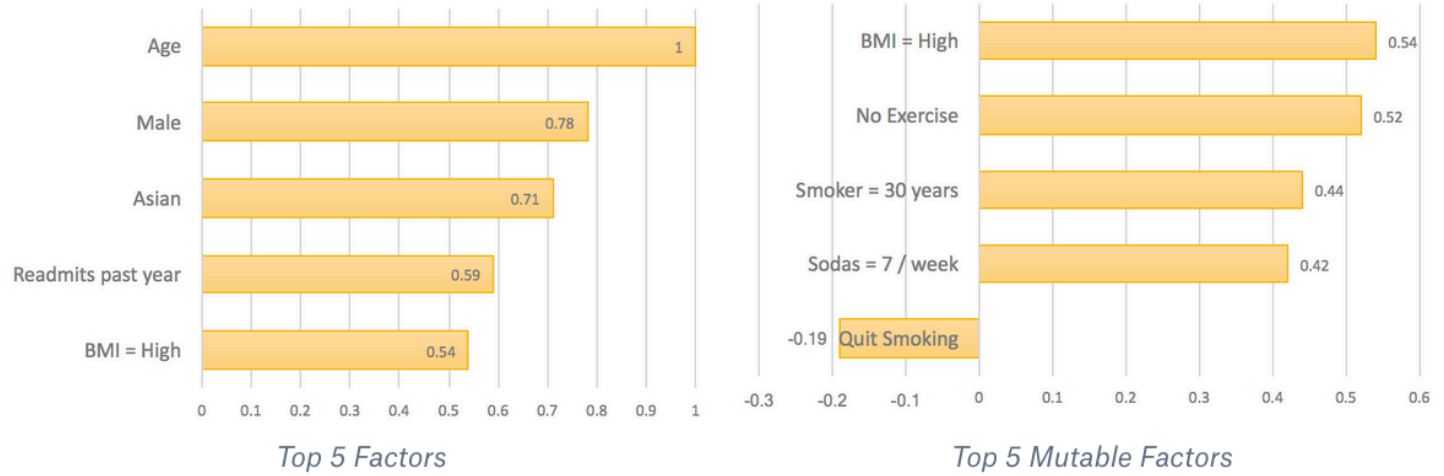


Figure 1: Factors for Predicting Risk of Readmission

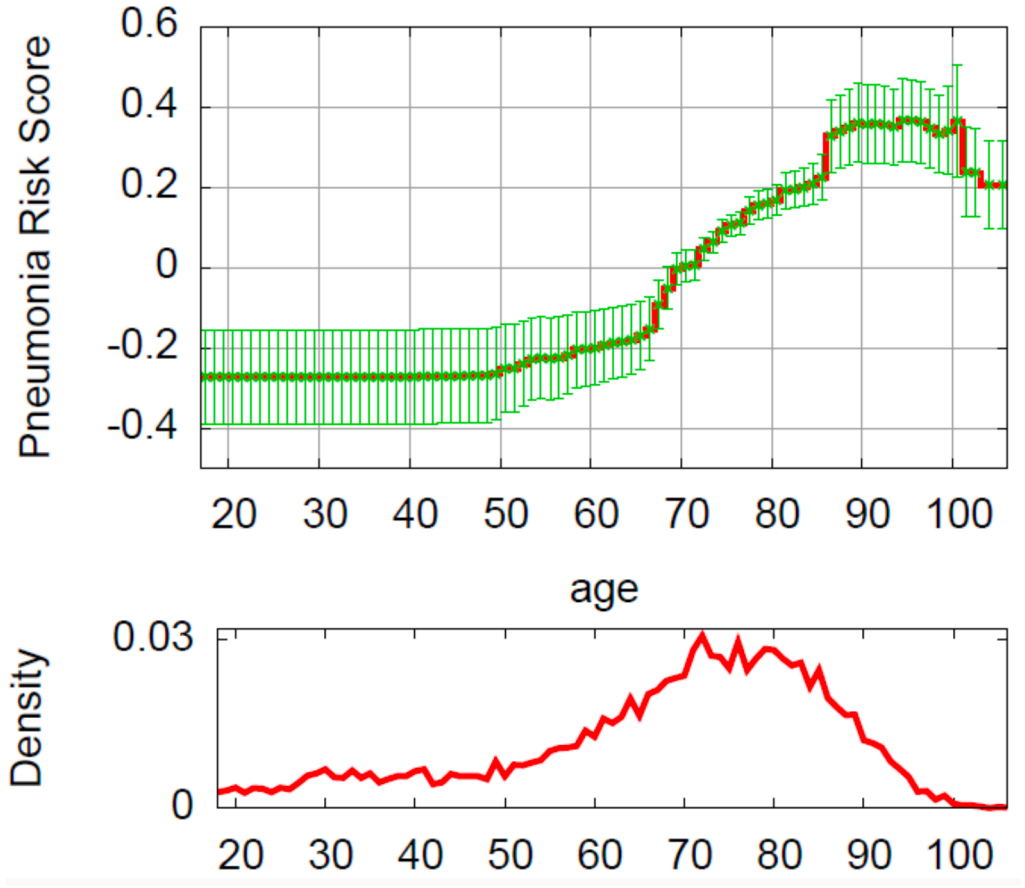
Actionability of Explanations in Healthcare

- Actionability is context dependent and role dependent
- Actionability is NOT causality
- There may be circumstances were actionability is not possible

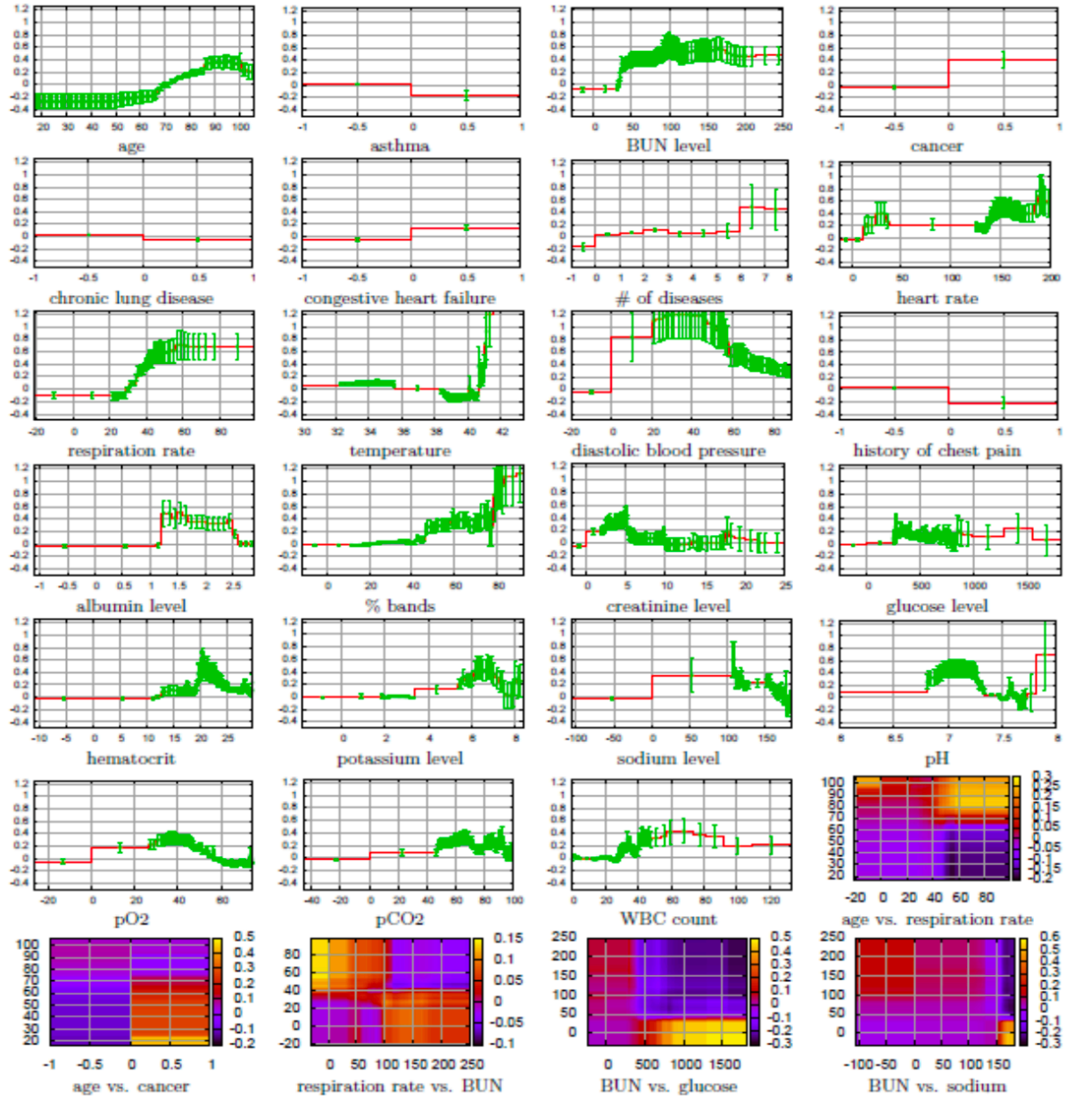
Domain Sense: GA2M

Caruna 2015, Caruna 2017

- Generalized Additive Models with pairwise interactions
- Accounts for interactions between the target variable and the feature space
- Possible to visualize the relationship between the target variable to gain insights into why the prediction is being made



(Carauna 2017)


age
asthma
BUN level
cancer
chronic lung disease
congestive heart failure
# of diseases
heart rate
respiration rate
temperature
diastolic blood pressure
history of chest pain
albumin level
% bands
creatinine level
glucose level
hematocrit
potassium level
sodium level
pH
pO2
pCO2
WBC count
age vs. respiration rate
age vs. cancer
respiration rate vs. BUN
BUN vs. glucose
BUN vs. sodium

(Caruna 2017)

Consistency

The explanation should be consistent across different models and across different runs of the model

LWBS

Left without being seen refers to a patient leaving the facility without being seen by a physician

Consistency

- Dr. Marcos is examining the reasons for a patient who is predicted to leave without being seen and notices that the explanation for them leaving are different from what observed 4 hours ago
- Upon investigation Katherine determines that the LIME model is being used for prediction which is non-deterministic, hence differences in explanations
- Having different explanations for the same instance can be confusing for users, hence the need for consistency

Evaluating Consistency

- Kendall's Tau

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{n(n-1)/2}$$

W	Interpretation
$W \leqslant 0.3$	Weak agreement
$0.3 < W \leqslant 0.5$	Moderate agreement
$0.5 < W \leqslant 0.7$	Good agreement
$W > 0.7$	Strong agreement

- Kendall' W

Judges	A	B	C	D	E	F	G	H
1	7	8	6	4	5	3	2	1
2	8	7	2	6	4.5	4.5	3	1
3	8	5.5	7	5.5	4	3	2	1
4	8	6.5	4	5	3	6.5	1	2
5	6	4.5	2	7	3	8	1	4.5
6	7.5	6	7.5	2.5	5	4	1	2.5
7	6	7	4	8	4	4	2	1
	50.5	44.5	32.5	38	28.5	33	12	13

Model Multiplicity vs Explanation Consistency

- Given the same dataset, multiple machine learning algorithms can be constructed with similar performance
- The explanations that are produced by multiple explainable algorithms should be very similar if not the same
- Wide divergence in explanations is a sign of problem with explanations or with the algorithm(s)

Variable	Model A	Model B	Model C
Age	1	1	5
Gender	2	4	6
Diabetic	3	5	1
Race	4	6	4
Smoker	5	2	3
Alcoholic	6	3	2

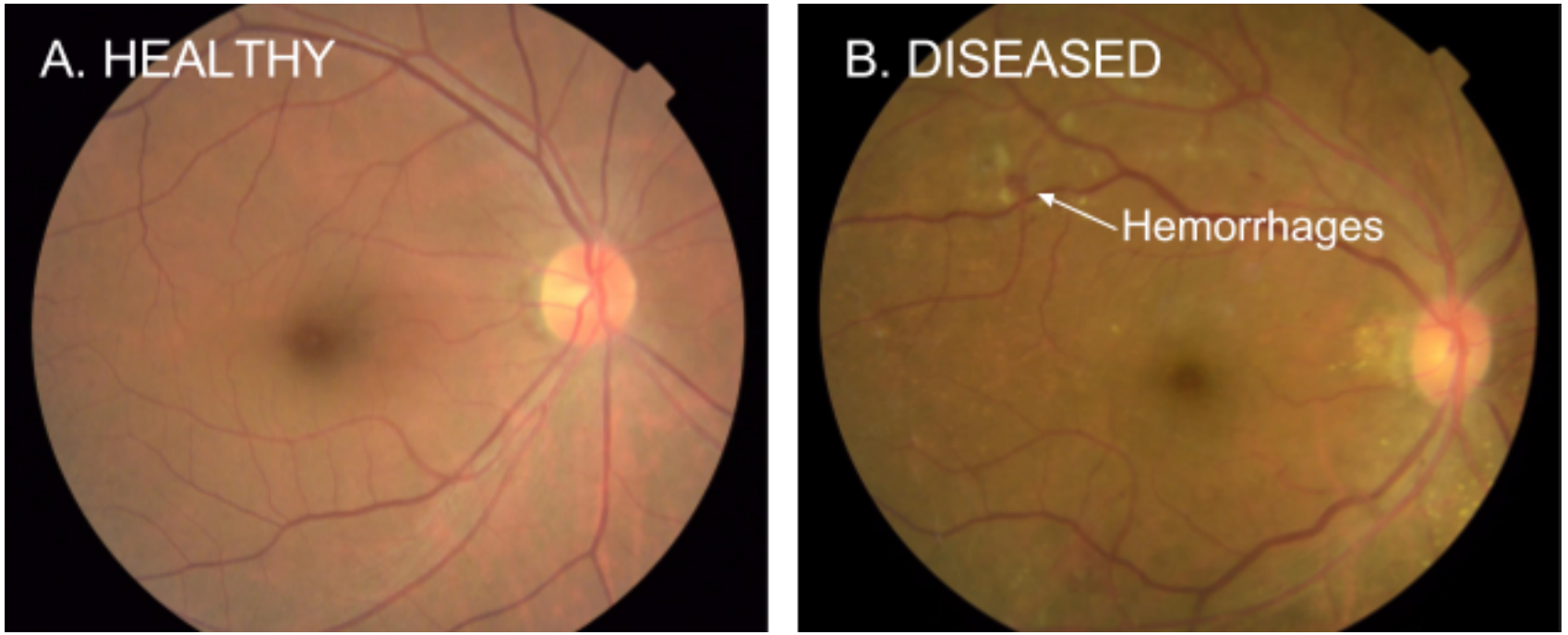
Variables for Length of Stay Prediction Ranked

County ED | ED Arrivals Prediction

Rebuttal – explanation not always needed
Examples of when explanation is not needed

- Whende and Katherine are looking at ED arrivals prediction
- Kate shows the explanation factors to Whende but the performance of the corresponding model is relatively low (Precision = 0.34, Recall = 0.46)
- Whende states that getting explanations for ED arrivals prediction is not very important, performance takes precedence

Exceptions to Explanations



Limits of Explanations

"You can ask a human, but, you know, what cognitive psychologists have discovered is that when you ask a human you're not really getting at the decision process. They make a decision first, and then you ask, and then they generate an explanation and that may not be the true explanation."

- Peter Norvig

Explainability and Cognitive Limitations

- Machine Learning is used in problems where the size of the data and/or the number of variables is too large for humans to analyze
- What if the most parsimonious model is indeed too complex for humans to analyze or comprehend?
- Ante-Hoc explanations may be impossible and post-hoc explanations would be incorrect
- Explanations may not be possible in some cases

Parsimony

The explanation should be as simple as possible

Admission Disposition

Where in the hospital the patient should go once they are admitted

Parsimony

- MDL (Minimum Description Length) and Occam's Razor
- Occam's Razor in Machine Learning
  - Occam's First Razor
  - Occam's Second Razor
- Occam's Razor in Interpretable Machine Learning
- The simplest explanation is not always the best one

Hickman's Dictum and Chattam's Anti-Razor

- Applying Occam's Razor can be counter productive
- Occam's Razor is context free
- Healthcare in Practice
  - Multiple competing hypothesis
- Heliocentric vs. Geocentric Models
- Hiccam's Dictim
- Chattam's Anti-Razor

Decision Lists

- Falling Rule Lists (FRL) [1]
- Bayesian Rule List (BRL) [2]
- Interpretable Decision Sets (IDS) [3]

	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density $\geq$ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age $\geq$ 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

TABLE 1

Decision list for mammographic mass dataset.

([1] Wang and Rudin, 2015, [2] Letham 2015, [3] Lakkaraju, Bach & Leskovec, 2016)

Bayesian Rule Lists

- BRLs are decision lists--a series of if-then statements
- BRLs discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements.
- Experiments show that BRLs have predictive accuracy on par with the current top ML algorithms (approx. 85- 90% as effective) but with models that are much more interpretable

- **if** hemiplegia and age > 60
  - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
  - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
  - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
  - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
  - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
  - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)

(Letham 2015)

Generalizability

The explanation should be generalizable

Length of Stay

How long is the patient going to stay in the facility

Model Generalizability

- Dr. Marcos is looking at a patient's info for length of stay prediction. She observes that many of the instances have similar explanations
- However the explanations do not appear to be generalizable to the whole population

Explanation Generalizability

- **Local Models:** Models that give explanations at the level of an instance
- *Examples:* LIME, Shapley Values etc.
- **Global Models:** Models that give explanations
- *Examples:* Decision Trees, Rule Based Models etc.
- **Cohort Level Models:** A type of global models where the explanations are generated at the level of cohort
- *Examples:* Same as global models

Algorithm Generalizability

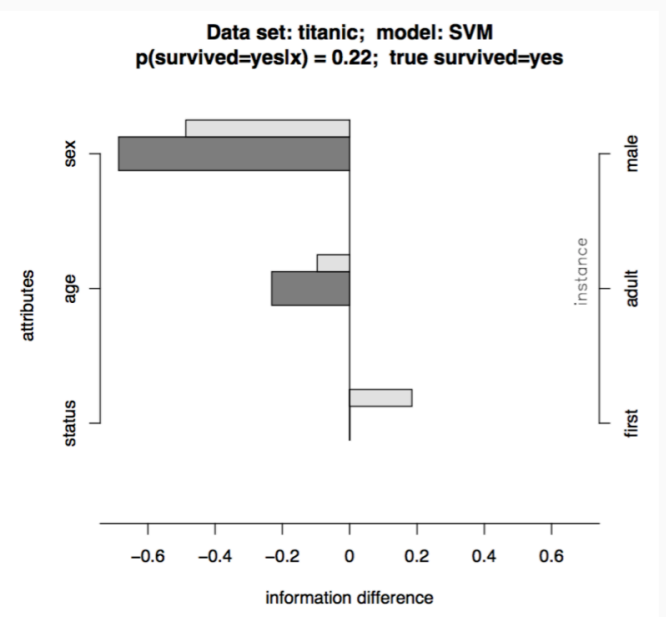
- Are the interpretable/explainable algorithms generalizable to all predictive algorithm, a particular class of algorithms or tied to a particular algorithm
- Model Agnostic Explanations:
  - Examples: LIME, Shapley Values etc.
- Model Class Specific Explanations:
  - Examples: Tree Explainers
- Model Specific Explanations:
  - Examples: CENs, Decision Trees, SLIM etc.

Dimensions of Explanations

- Data
  - What variables or features are most relevant for the prediction of length of stay?
- Prediction
  - Explain why certain patients are being predicted to have long length of stays
- Model
  - What are the patterns belonging to a particular category (long length of stay) typically look like?

Prediction Decomposition

- Explain the model prediction for one instance by measuring the difference between the original prediction and the one made with omitting a set of features

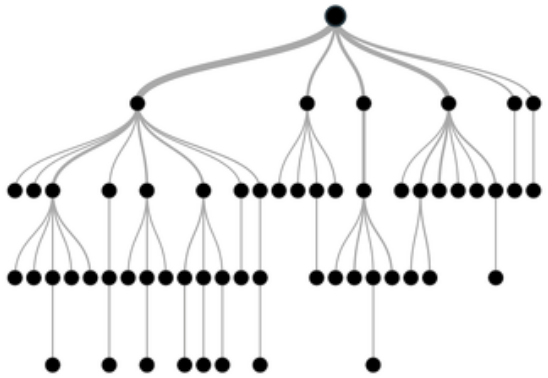


- Problem: Imputation may be needed in case of missing values

Robnik-Sikonja and Kononenko (2008)

Model Outputs

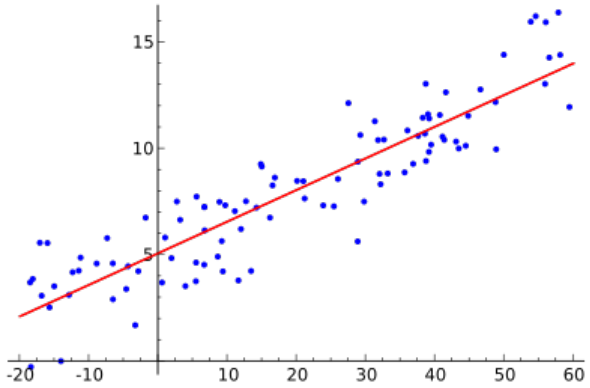
Rule Based Models



Rules for Length of Stay Prediction

IF age > 55 AND gender = male AND condition = 'COPD' AND complication = 'YES'
THEN
Length of stay = long (> 7 days)

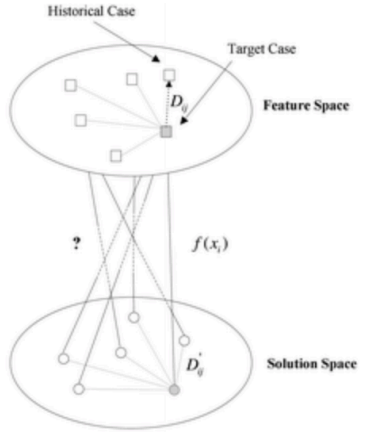
Relative Variable Importance



Top 5 Variables for Length of Stay Prediction

Variable	*Importance*
Age	0.45
Gender	0.37
Diabetic	0.32
Race	0.21
Smoker	0.14

Case Based Models



Example cases for Length of Stay Prediction

Patient X is predicted to have a length of stay of 20 days because he is most similar to these 5 patients who on average had length of stay of 5 days

Trust / Performance

The expectation that the corresponding predictive algorithm for explanations should have a certain performance

ICU Transfer Prediction

Predict if the patient will be transferred to the ICU

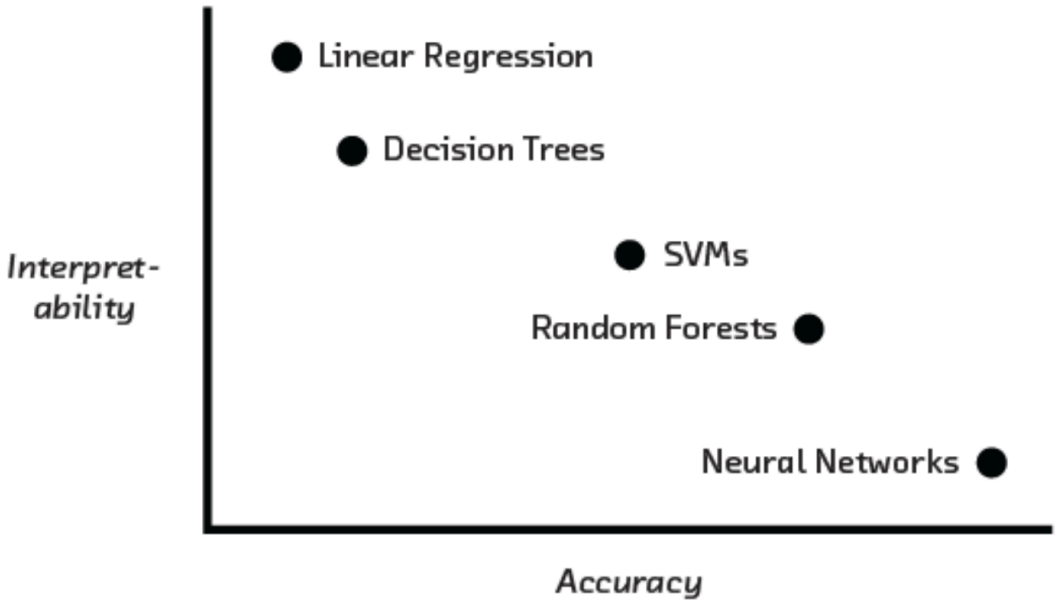
Trust in Human Performance Parity

- The model has at least parity with the performance of human practitioners
- Example: Model B has human performance parity

Model	Precision	Recall	F-Score	Accuracy
Physician's Prediction	0.73	0.71	0.72	0.60
Model A	0.82	0.68	0.74	0.65
Model B	0.83	0.81	0.82	0.89

Results for Length of Stay Prediction (Long vs. Short Stays)

Performance vs. Explainability



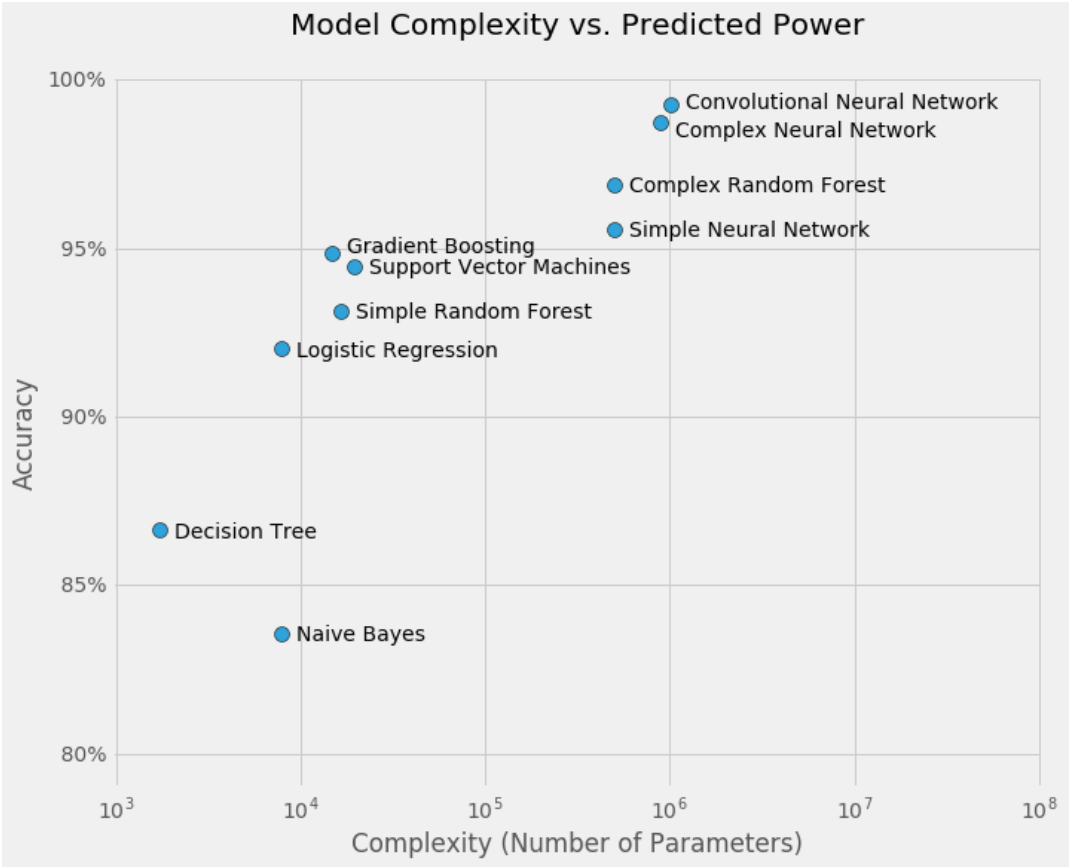
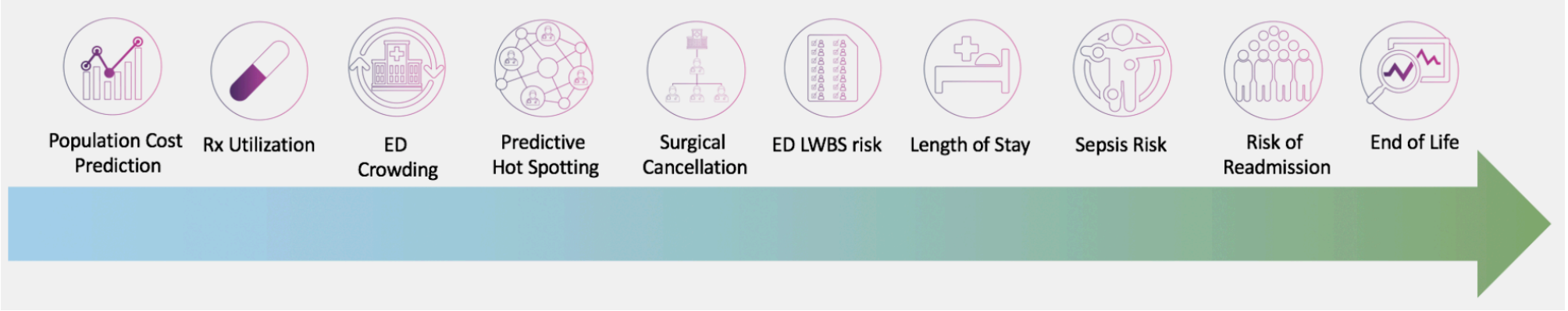


Image Source: Easy Solutions

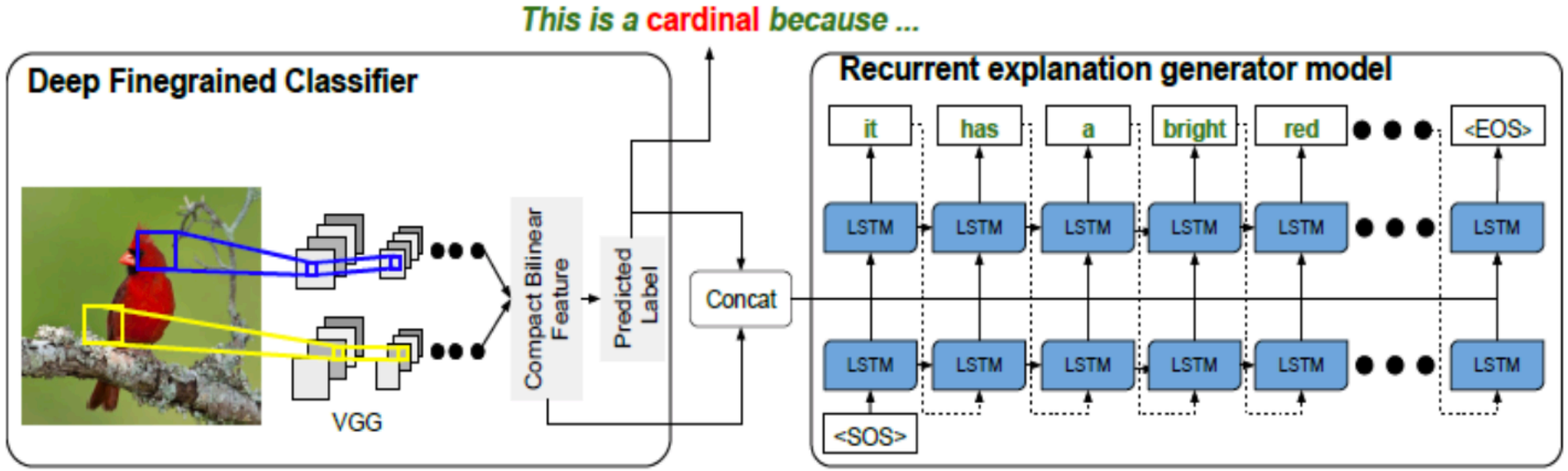
The Tripartite Trade-Off

- Performance vs. Explanation vs. Risk



Explanations vs Associations

- A CNN is trained to recognize objects in images
- A language generating RNN is trained to translate features of the CNN into words and captions



Explanations vs Assocations

- Hendricks et al created a system to generate explanations of bird classifications. The system learns to:
- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)
- Limitations
  - Limited (indirect at best) explanation of internal logic
  - Limited utility for understanding classification errors

Explainability and Adversarial ML

ADD SLIDE ABOUT ACTIVATIONS

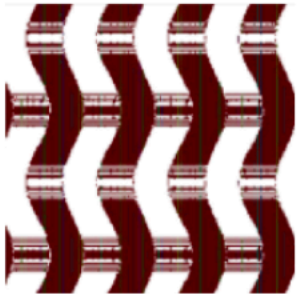
(a)

(b)

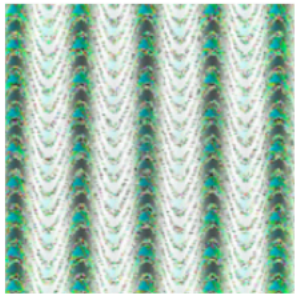
(c)

(d)

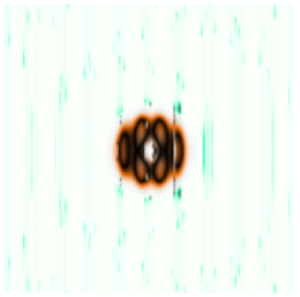
(e) School bus



(f) Guitar



(g) Peacock



(h) Pekinese

(Szegedy et al., 2013, Ghorbani 2017)

Fidelity

The expectation that the explanation and the predictive model align well with one another

Risk of Readmission

Predict if the patient will be readmitted within a particular span in time

Fidelity

- Dr. Marcos examines some explanations for risk of readmission prediction and discovers that the explanations do not appear to be correct
- After ruling out non-determinism and thus lack of consistency as an explanation, Katherine examines the data

Data Provenance

- The Explanation is going to be as good as the data
- Caruana's Mortality Prediction Example
- Low Quality data
  - Instrumentation problems
  - Censoring
- Incorrect explanations may be given because of problems in the data

Model Fidelity

- Fidelity with the underlying phenomenon
- Readmission model which uses lunar cycles as a feature, the model may have good predictive power and the feature may even be quite helpful but the model does not have fidelity with the underlying phenomenon

Explanation Fidelity

- An explanation is **Sound** if it adheres to how the model actually works
- An Explanation is **Complete** if it encompasses the complete extent of the model
- Soundness and Completeness are relative
- The soundness of an explanation can also vary depending upon on the constraints on the explanations
- Ante-Hoc models are perfectly sound by definition, Post-Hoc models can have varying leveling of soundness

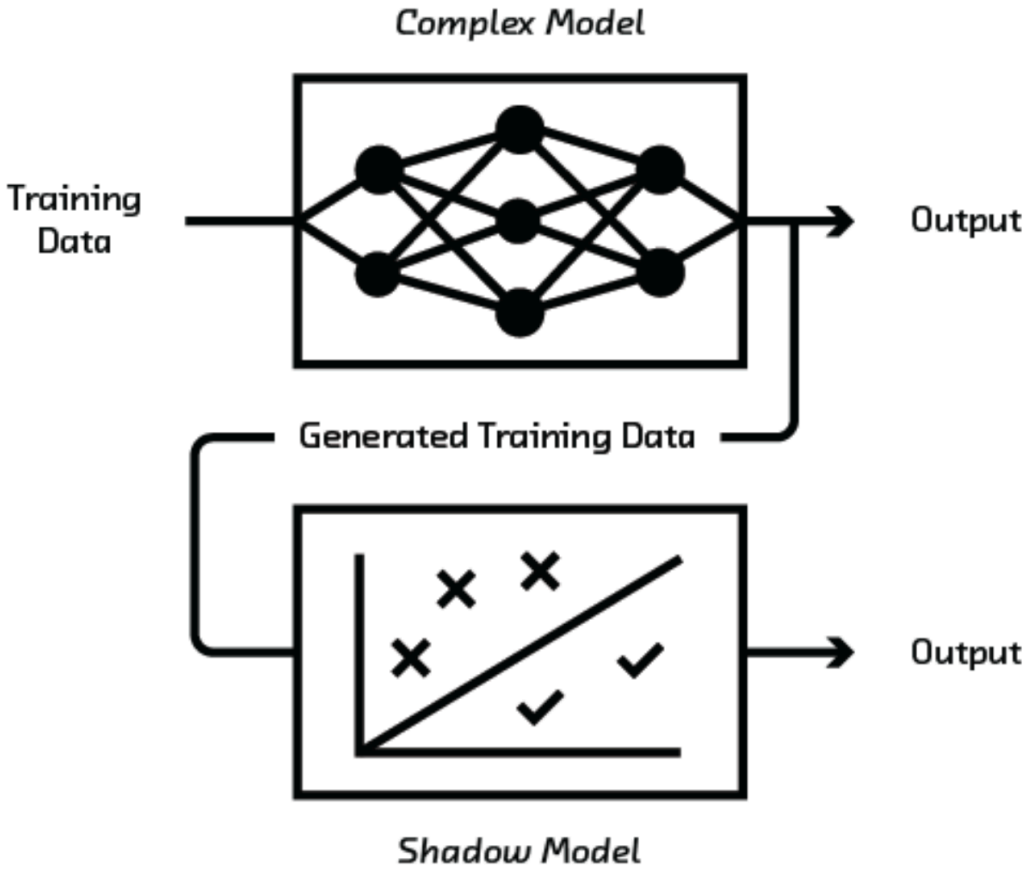
Explanation Fidelity

- Fidelity with the prediction model i.e., explanations should align as close to the predictive model as possible (Soundness)
- Ante-Hoc: Models where the predictive model and the explanation model is the same
- Post-Hoc Models: : Models where the predictive model and the explanation model are different
- Special Case: Mimic Models

Mimic Models

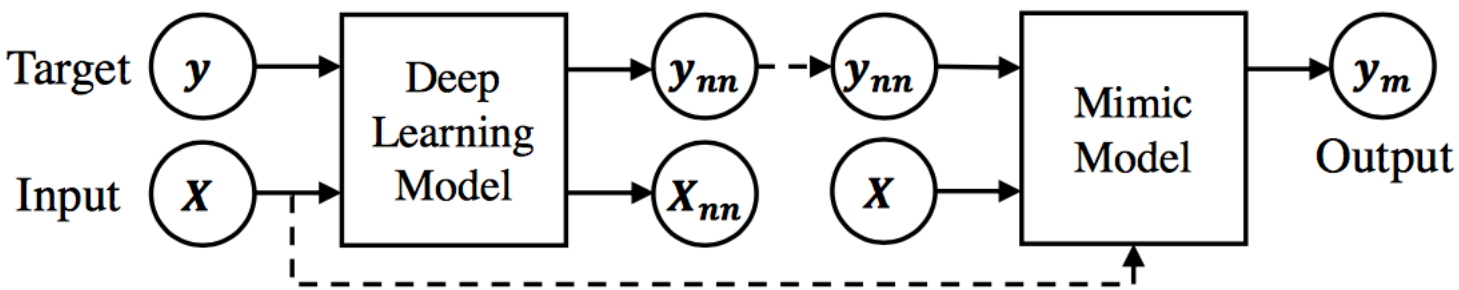
Bucilua 2006

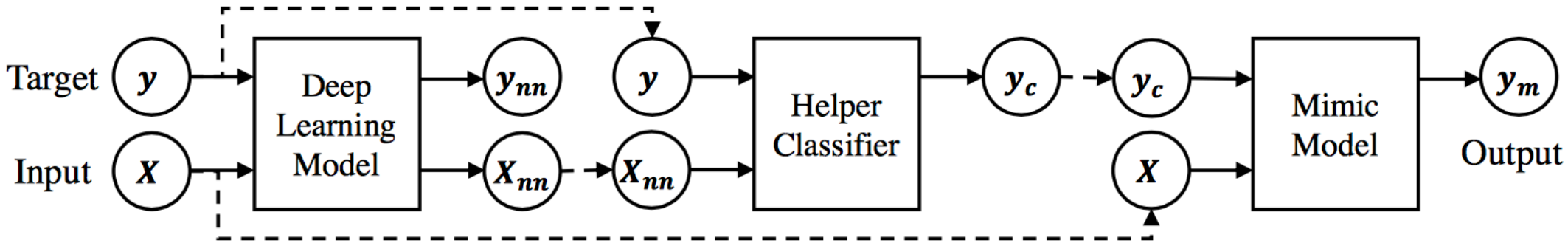
- Also known as Shadow, Surrogate or Student Models
- Use the output (instead of the true labels) from the complex model and the training data to train an model which is explainable
- The performance of the student model is usually quite good
- Example: Given a highly accurate SVM, train a decision tree on the predicted label of the SVM and the original data



(Tan 2017)

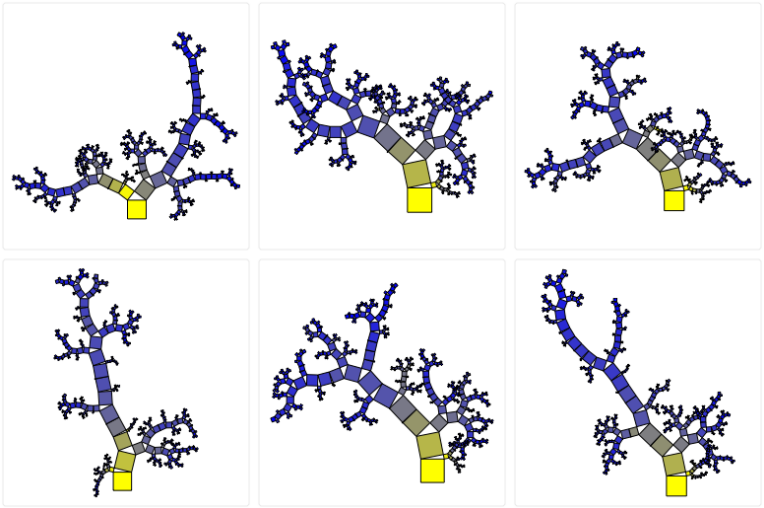
Mimic Models for Deep Learning





Interpreting Random Forests

- For each decision that a tree (or a forest) makes there is a path (or paths) from the root of the tree to the leaf
- The path consists of a series of decisions, corresponding to a particular feature, each of which contribute to the final predictions
- Variable importance can be computed as the contribution of each node for a particular decision for all the decision trees in the random forest



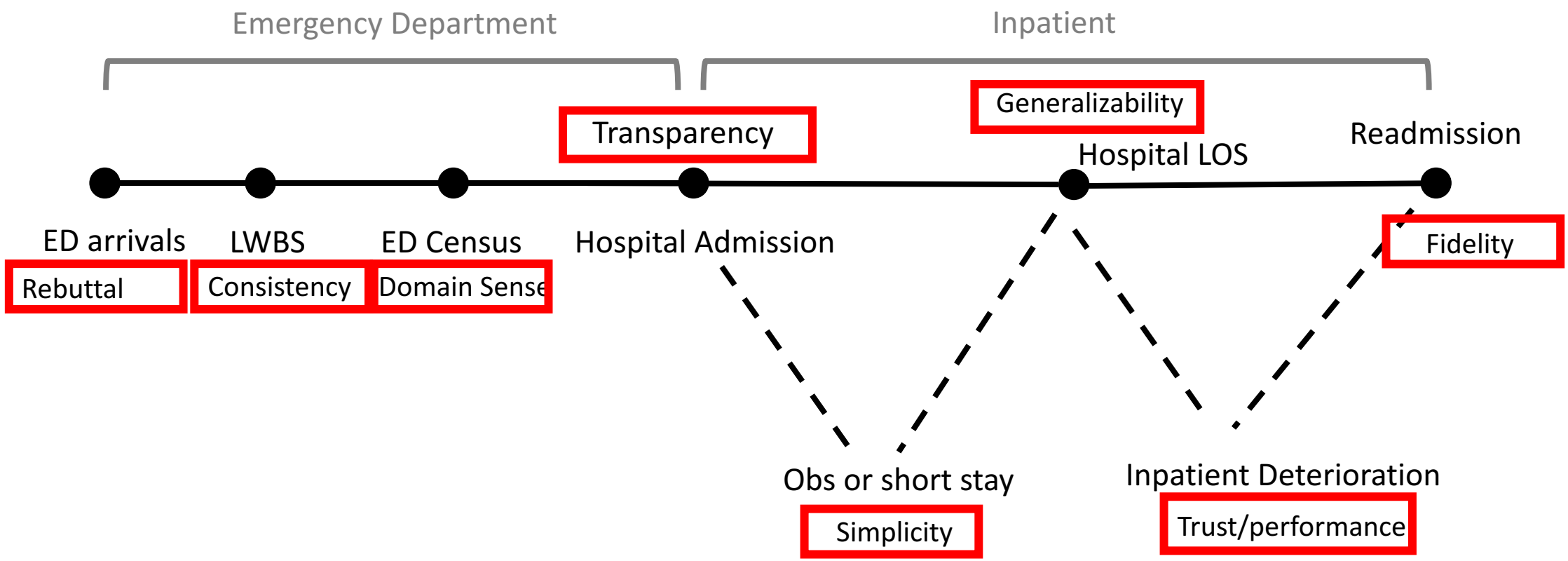
(Saabas 2014)

BETA

- Black Box Explanation through Transparent Approximations
- BETA learns a compact two-level decision set in which each rule explains part of the model behavior unambiguously
- Novel objective function so that the learning process is optimized for:
- **High Fidelity** (high agreement between explanation and the model)
- **Low Unambiguity** (little overlaps between decision rules in the explanation)
- **High Interpretability** (the explanation decision set is lightweight and small)

(Lakkaraju, Bach & Leskovec, 2016)

Technique	Composition	Performance*	Model Fidelity	Model Specificity	Explanation Type	Scalable**	Scope
Bayesian Rule List	Ante-Hoc	H	Yes	Self	Relative Importance	Yes	Global
BETA^^	Ante-Hoc	H	Yes	Self	Relative Importance	Yes	Global
Decision Trees	Ante-Hoc	M	Yes	Self	Rules	Yes	Global
Falling Rule Lists	Ante-Hoc	H	Yes	Self	Relative Importance	Yes	Global
GAM	Ante-Hoc	L	Yes	Self	Relative Importance	Yes	Global
GA2M	Ante-Hoc	M	Yes	Self	Graphs	Yes	Global
ICE Plots	Post-Hoc	N/A	No	Agnostic	Graphs	Yes	Global
Interpretable Decision Sets	Ante-Hoc	H	Yes	Self	Relative Importance	Yes	Global
k-LIME	Post-Hoc	N/A	No	Agnostic	Relative Importance	Yes	Local
LIME^^^	Post-Hoc	N/A	No	Agnostic	Relative Importance	Data size	Local
Logistic Regression	Ante-Hoc	M	Yes	Self	Relative Importance	Yes	Global
Model Distillation	Post-Hoc	M-H	Yes	Agnostic	*Any*	Yes	Global
Partial Dependence Plots	Post-Hoc	N/A	No	Agnostic	Graphs	Yes	Global
RF Explainer	Post-Hoc	H	No	Random Forest	Relative Importance	Yes	Local
Relative Baseline Contributions***	Post-Hoc	N/A	No	Agnostic	Relative Importance	Yes	Local
Right for the Right Reasons ^	Post-Hoc	H	No	Agnostic	Relative Importance	Yes	Local
Shapley Values	Post-Hoc	N/A	No	Agnostic	Graphs	Number of Features	Local
SLIM	Ante-Hoc	H	Yes	Self	Relative Importance	Yes	Global
XGB Explainer	Post-Hoc	H	No	XG Boost	Relative Importance	Yes	Local


Emergency Department
Inpatient
Transparency
Generalizability
Hospital LOS
Readmission
ED arrivals
LWBS
ED Census
Hospital Admission
Fidelity
Rebuttal
Consistency
Domain Sense
Obs or short stay
Simplicity
Inpatient Deterioration
Trust/performance

References

Ahmad 2018	Muhammad Aurangzeb Ahmad, Carly Eckert, Greg McKelvey, Kiyana Zolfagar, Anam Zahid, Ankur Teredesai. Death vs. Data Science: Predicting End of Life IAAI February 2-6, 2018
Al-Shedivat 2017A	Al-Shedivat, Maruan, Avinava Dubey, and Eric P. Xing. ”Contextual Explanation Networks.” arXiv preprint arXiv:1705.10301 (2017).
Al-Shedivat 2017B	Al-Shedivat, Maruan, Avinava Dubey, and Eric P. Xing. "The Intriguing Properties of Model Explanations."
Bach 2015	Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140
Bayes 1763	Bayes, Thomas, Richard Price, and John Canton. "An essay towards solving a problem in the doctrine of chances." (1763): 370-418.
Biran 2014	Biran, Or, and Kathleen McKeown. "Justification narratives for individual classifications." In Proceedings of the AutoML workshop at ICML, vol. 2014. 2014.
Biran 2017A	Biran, Or, and Kathleen R. McKeown. "Human-Centric Justification of Machine Learning Predictions." In IJCAI, pp. 1461-1467. 2017.
Biran 2017B	Biran, Or, and Courtenay Cotton. "Explanation and justification in machine learning: A survey." In IJCAI-17 Workshop on Explainable AI (XAI), p. 8. 2017.
Bucilua 2006	Buciluǎ, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535-541. ACM, 2006.
Caruna 2015	Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "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, pp. 1721-1730. ACM, 2015.
Caruna 2017	Caruana, Rich, Sarah Tan, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad et al. "Interactive Machine Learning via Transparent Modeling: Putting Human Experts in the Driver’s Seat.” IDEA 2017
Chang 2009	Chang, Jonathan, et al. “Reading tea leaves: How humans interpret topic models.” Advances in Neural Information Processing Systems. 2009
Craik 1967	Craik, Kenneth James Williams. The nature of explanation. Vol. 445. CUP Archive, 1967.
Craven 1996	Mark W. Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. Advances in Neural Information Processing Systems, 1996.
Datta 2016	Datta, Anupam, Shayak Sen, and Yair Zick. "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems." Security and Privacy (SP), 2016 IEEE Symposium on. IEEE, 2016
Doshi-Velez 2014	Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
Doshi-Velez 2017	Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. "Accountability of AI under the law: The role of explanation." arXiv preprint arXiv:1711.01134 (2017).
Druzdzel 1996	Druzdzel, Marek J. "Qualitiative verbal explanations in bayesian belief networks." AISB QUARTERLY (1996): 43-54.
Farah 2014	Farah, M.J. Brain images, babies, and bathwater: Critiquing critiques of functional neuroimaging. Interpreting Neuroimages: An Introduction to the Technology and Its Limits 45, S19-S30 (2014)
Freitas 2014	Freitas, Alex A. "Comprehensible classification models: a position paper." ACM SIGKDD explorations newsletter 15, no. 1 (2014): 1-10.

References

Friedman 2001	Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. New York: Springer series in statistics, 2001.
Friedman 2008	Friedman, Jerome H., and Bogdan E. Popescu. "Predictive learning via rule ensembles." The Annals of Applied Statistics2, no. 3 (2008): 916-954.
Goldstein 2014	Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." Journal of Computational and Graphical Statistics 24, no. 1 (2015): 44-65.
Guidotti 2018	Guidotti, Riccardo, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. "A survey of methods for explaining black box models." arXiv preprint arXiv:1802.01933(2018).
Gunning 2016	David Gunning Explainable Artificial Intelligence (XAI) DARPA/I2O 2016
Gorbunov 2011	Gorbunov, K. Yu, and Vassily A. Lyubetsky. "The tree nearest on average to a given set of trees." Problems of Information Transmission 47, no. 3 (2011): 274.
Hall 2018	Hall, P., Gill, N., Kurka, M., Phan, W. (May 2018). Machine Learning Interpretability with H2O Driverless AI. http://docs.h2o.ai.
Hardt 2016	Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." In Advances in neural information processing systems, pp. 3315-3323. 2016.
Hastie 1990	T. Hastie and R. Tibshirani. Generalized additive models . Chapman and Hall/CRC, 1990.
Herlocker 2000	Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. "Explaining collaborative filtering recommendations." In Proceedings of the 2000 ACM conference on Computer supported cooperative work, pp. 241-250. ACM, 2000.
Hoffman 2017	Hoffman, Robert R., Shane T. Mueller, and Gary Klein. "Explaining Explanation, Part 2: Empirical Foundations." IEEE Intelligent Systems 32, no. 4 (2017): 78-86.
Hutson 2018	Hutson, Matthew. "Has artificial intelligence become alchemy?." (2018): 478-478.
Koh 2017	Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." arXiv preprint arXiv:1703.04730 (2017).Harvard
Kulesza 2014	Kulesza, Todd. "Personalizing machine learning systems with explanatory debugging." (2014).
Lakkaraju 2016	Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. "Interpretable decision sets: A joint framework for description and prediction." Proc. 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM, 2016.
Lei 2017	Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. Journal of the American Statistical Association, 2017
Lipovetsky 2001	Lipovetsky, Stan, and Michael Conklin. "Analysis of regression in game theory approach." Applied Stochastic Models in Business and Industry 17.4 (2001): 319-330
Lipton 2016	Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).
Lipton 2017	Lipton, Zachary C. "The Doctor Just Won't Accept That!." arXiv preprint arXiv:1711.08037 (2017).
Lou 2013	Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate intelligible models with pairwise interactions." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 623-631. ACM, 2013.
Lundberg 2017	Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." In Advances in Neural Information Processing Systems, pp. 4765-4774. 2017.
Miller 2017A	Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv preprint arXiv:1706.07269 (2017).
Miller 2017B	Miller, Tim, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of inmates running the asylum." In IJCAI-17 Workshop on Explainable AI (XAI), vol. 36. 2017.
Montavon 2017	Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. "Explaining nonlinear classification decisions with deep taylor decomposition." Pattern Recognition 65 (2017): 211-222.

References

Moosavi-Dezfooli 2016	Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574-2582. 2016.
Morstatter 2016	Fred Morstatter and Huan Liu Measuring Topic Interpretability with Crowdsourcing KDD Nuggets November 2016
Nott 2017	Nott, George "Google's research chief questions value of 'Explainable AI'" Computer World 23 June, 2017
Olah 2018	Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. "The building blocks of interpretability." Distill 3, no. 3 (2018): e10.
Perez 2004	Pérez, Jesus Maria, Javier Muguerza, Olatz Arbelaitz, and Ibai Gurrutxaga. "A new algorithm to build consolidated trees: study of the error rate and steadiness." In Intelligent Information Processing and Web Mining, pp. 79-88. Springer, Berlin, Heidelberg, 2004.
	Quinlan, J. Ross. "Some elements of machine learning." In International Conference on Inductive Logic Programming, pp. 15-18. Springer, Berlin, Heidelberg, 1999.
Ras 2018	Ras, Gabrielle, Pim Haselager, and Marcel van Gerven. "Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges." arXiv preprint arXiv:1803.07517(2018).
Ribeiro 2016a	Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "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, pp. 1135-1144. ACM, 2016.
Ribeiro 2016b	Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin Introduction to Local Interpretable Model-Agnostic Explanations (LIME) August 12, 2016 Oriely Media
Ross 2018	Ross, Casey IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show Stat News July 25, 2018
Ross 2017	Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez. "Right for the right reasons: Training differentiable models by constraining their explanations." arXiv preprint arXiv:1703.03717 (2017).
Saabas 2014	Saabas, Ando. Interpreting random forests Data Dive Blog 2014
Shrikumar 2017	Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." arXiv preprint arXiv:1704.02685 (2017)
Strumbelj 2014	Strumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.
Tan 2017	Tan, Sarah, Rich Caruana, Giles Hooker, and Yin Lou. "Detecting Bias in Black-Box Models Using Transparent Model Distillation." arXiv preprint arXiv:1710.06169 (2017).
Turing 1950	Machinery, Computing. "Computing machinery and intelligence-AM Turing." Mind 59, no. 236 (1950): 433.
Ustun 2016	Ustun, Berk, and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems." Machine Learning 102, no. 3 (2016):349-391.
Wang 2015	Wang, Fulton, and Cynthia Rudin. "Falling rule lists." In Artificial Intelligence and Statistics, pp. 1013-1022. 2015.
Weld 2018	Weld, Daniel S., and Gagan Bansal. "Intelligible Artificial Intelligence." arXiv preprint arXiv:1803.04263 (2018).
Weller 2017	Weller, Adrian. "Challenges for transparency." arXiv preprint arXiv:1708.01870 (2017).
Wick 1992	M. R. Wick and W. B. Thompson. Reconstructive expert system explanation. Artificial Intelligence, 54(1- 2):33–70, 1992
Yang 2016	Yang, Hongyu, Cynthia Rudin, and Margo Seltzer. "Scalable Bayesian rule lists." arXiv preprint arXiv:1602.08610 (2016).