Warning: slides degraded heavily to make this deck <791MB

# INTERPRETABILITY, HUMANS IN LOOPS, POLICIES AND POLITICS


FRANK RUDZICZ

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

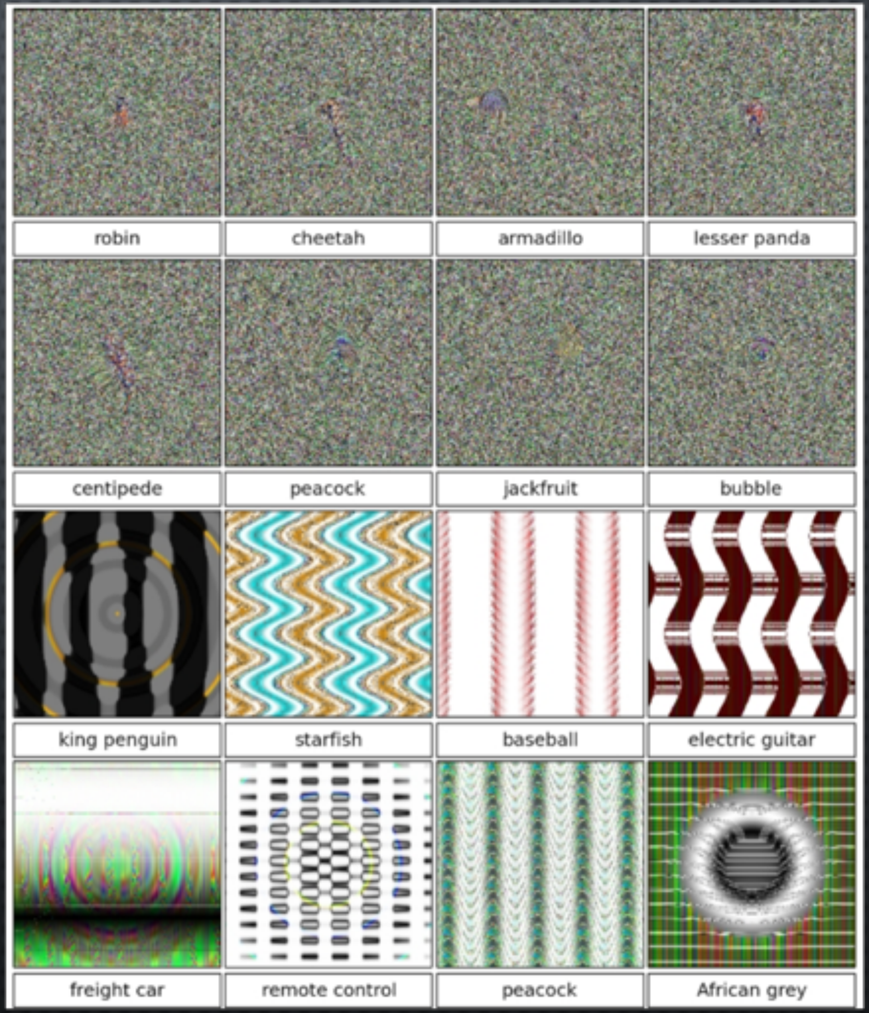
SST SURGICAL SAFETY TECHNOLOGIES

St. Michael's Inspired Care. Inspiring Science.

Standards Council of Canada
Conseil canadien des normes

# MAKING ALIEN MINDS

| | |
|---|---|
| Think rationally | Act rationally |
| Think like a human | Act like a human |




Labels with >99% confidence

Nguyen A, Yosinski J, Clune J.  (2015) Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proc. of IEEE CVPR*. 427–36.

# THE SAFETY OF AI

**1.** There *is* a risk that AI in the wrong hands, or in the hands of a select few, will perform tasks that may not be 'globally optimal'.

2. A ***bigger*** risk is that AI in the *right* hands will:
   1. lazily be given goals that are too abstract,
   2. find a 'trick' to achieve those goals that we don't understand, and
   3. result in unexpected, uninterpretable behaviour

*We need a means to explain model behaviour.*

# YOU GOT SOME 'SPLAININ TO DO

- What is actually meant by 'explainable'?
  - The wild, wild west is still working out its definitions…
- Here, we will try to stick to:
  - **explainable** *adj.* describes the model *in general*
  - **interpretable** *adj.* describes a specific decision.

"the term … holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi- mathematical way."

Lipton ZC. The Mythos of Model Interpretability. 2016. doi:10.1145/3233231, http://arxiv.org/abs/1606.03490

# DEFINITIONAL

IEEE Access

Received August 5, 2018, accepted September 4, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2870052*

## Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

**AMINA ADADI AND MOHAMMED BERRADA**

Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco

Corresponding author: Amina Adadi (amina.adadi@gmail.com)

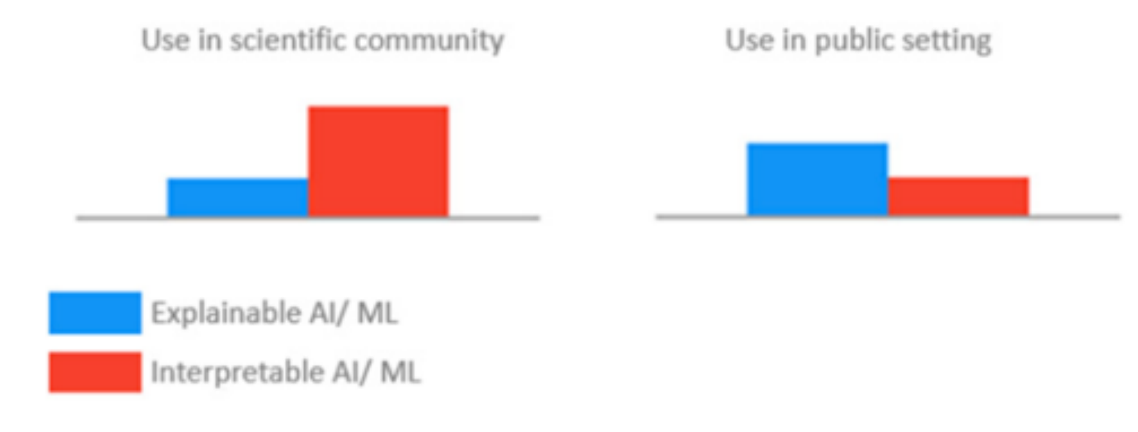


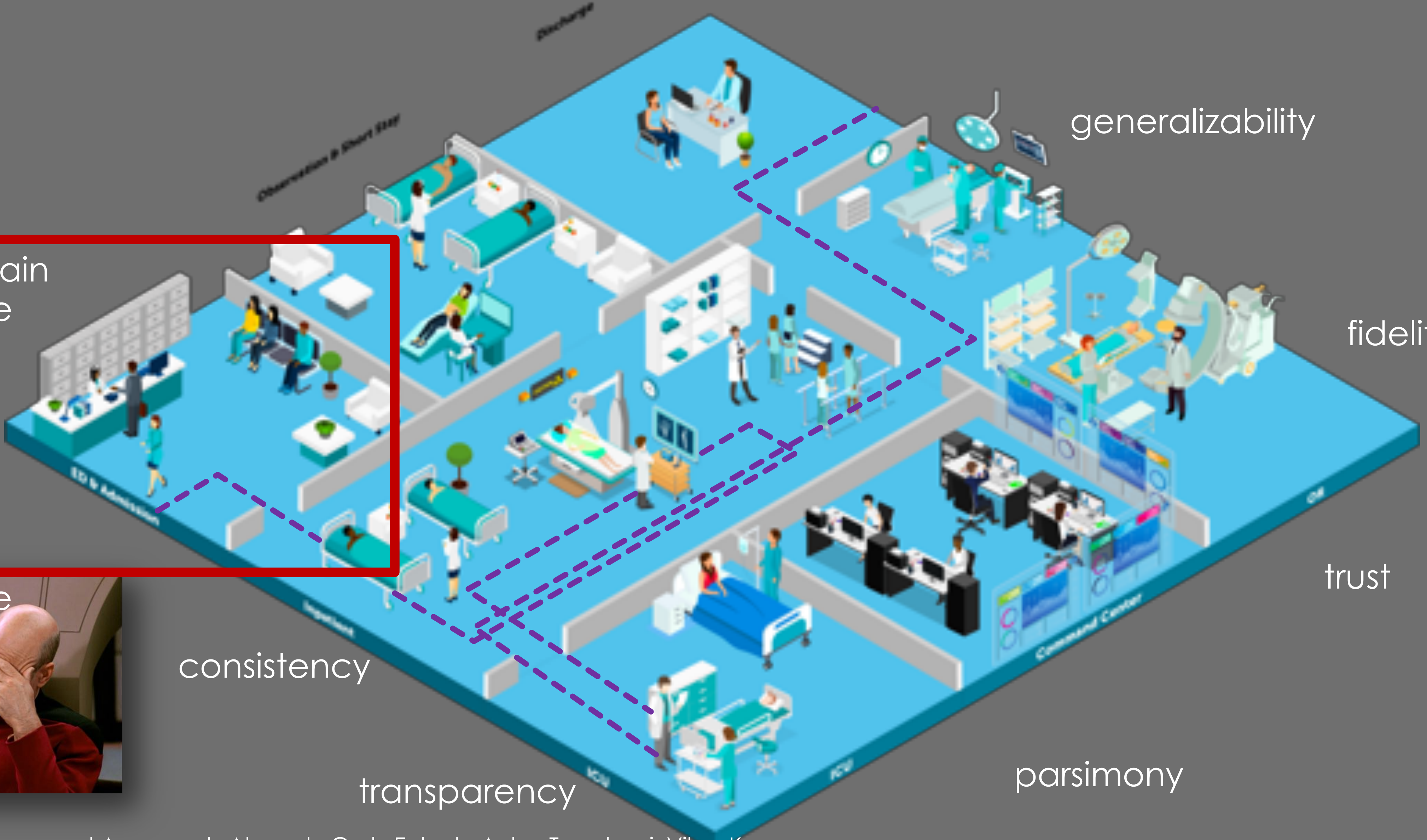

**FIGURE 3.** **Google trends result for comparing the use of "Explainable" and "Interpretable" according to the context.**

# YOU GOT SOME 'SPLAININ TO DO

- When do we ***want*** ML to be explainable?
  - We want to identify and remove bias to promote **safety**
  - We want to leverage **domain expertise**
  - We want to ensure **generalizability** and **consistency**
  - We want to **trust** the system

- When do we ***need*** ML to be explainable?
  - Regulatory approval process (e.g., FDA)
  - 'Right to explanation' (e.g., GDPA)

# JEAN-LUC'S PATH

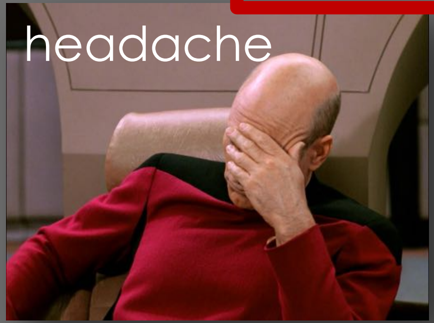


generalizability

domain sense

fidelity

trust




consistency

transparency

parsimony

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

# TRANSPARENCY

- Jean-Luc arrives at the ER.
- The nurse takes age, health history, vital signs, and inputs these into a ML model.
- Surprisingly, the model gives a $P(admission \mid JeanLuc) = 0.62$, which seems high.

- *Can we **audit** the system?*

# TRANSPARENCY

## The Mythos of Model Interpretability

Zachary C. Lipton [1]

**Abstract**

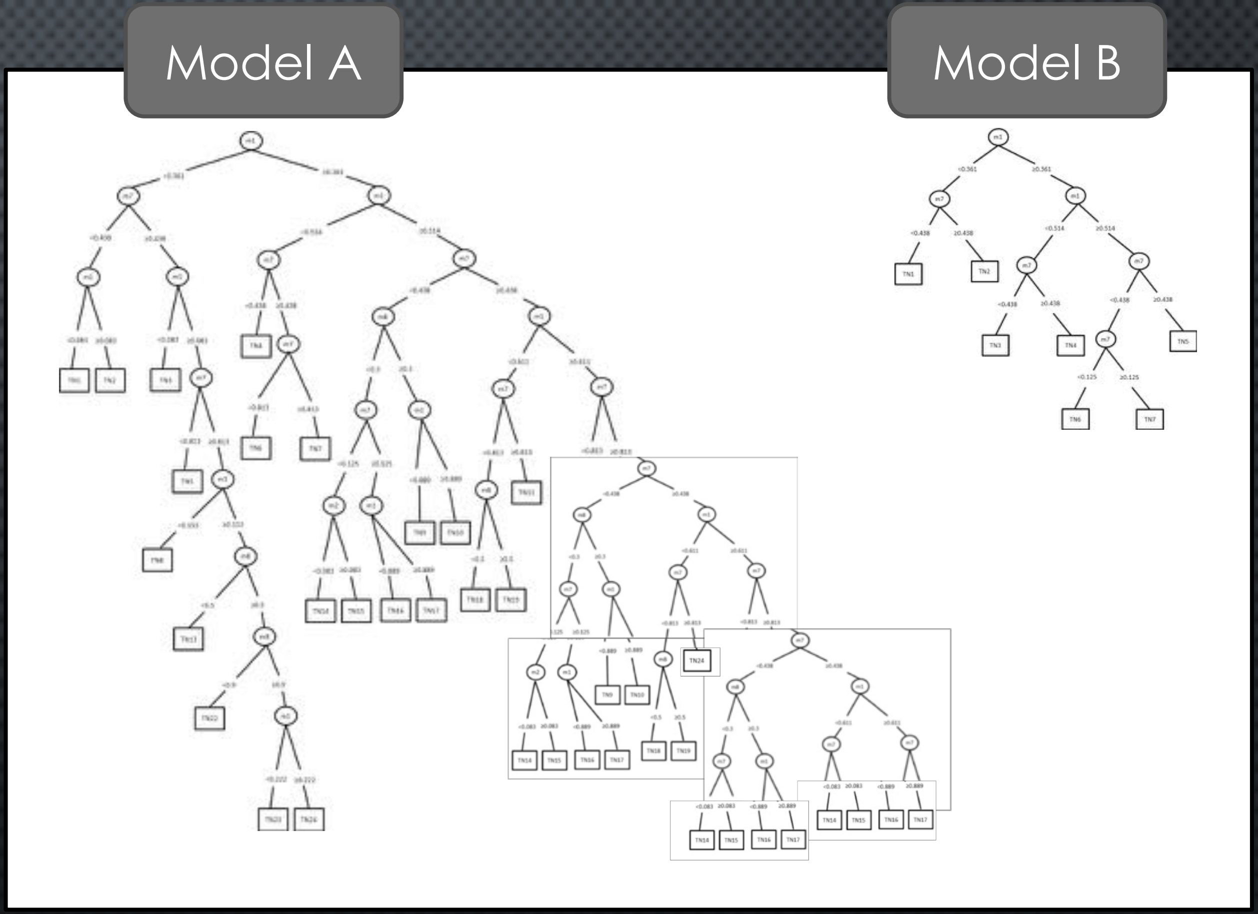
Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interno one has managed to set it in writing, or (ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept. In this paper, we seek to clarify both, suggesting that *interpretability* is not a monolithic concept, but in fact reflects several distinct ideas. We hope, through this critical analysis, to bring focus to the dialogue.

*Let's decompose interpretability into a few factors*

Lipton ZC. The Mythos of Model Interpretability. 2016. doi:10.1145/3233231, http://arxiv.org/abs/1606.03490

# TRANSPARENCY: SIMULTABILITY

- The entire model, or as much as possible, should be understood relatively holistically.
- Even basic decision trees can have thousands of nodes.



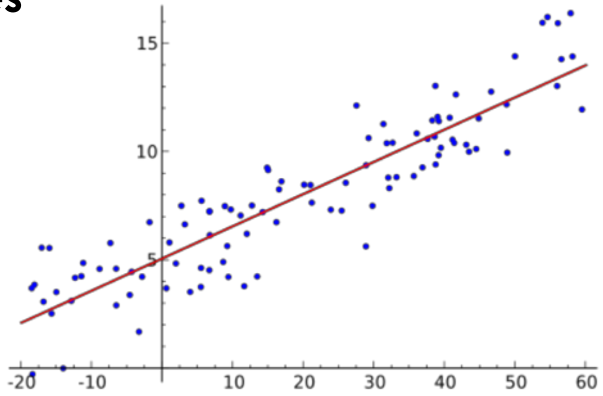

Lipton ZC. The Mythos of Model Interpretability. 2016. doi:10.1145/3233231, http://arxiv.org/abs/1606.03490

# TRANSPARENCY: DECOMPOSABILITY

- Each component should be **decomposable** into 'explainable' subcomponent.
  - E.g., noun-pronoun ratio vs variance of MFCC 14's $\delta\delta$

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

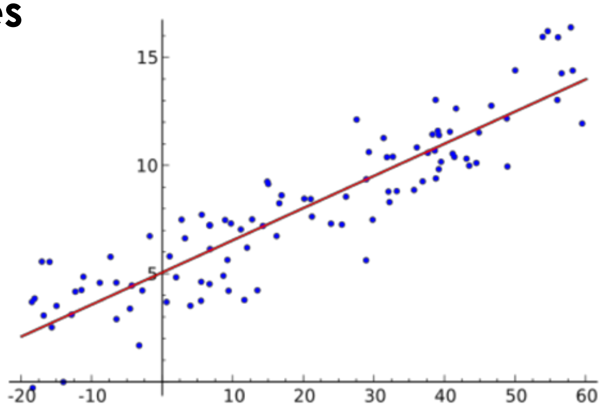
**Target Variable**
- Length of Stay




Model A

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

**Target Variable**
- Length of Stay

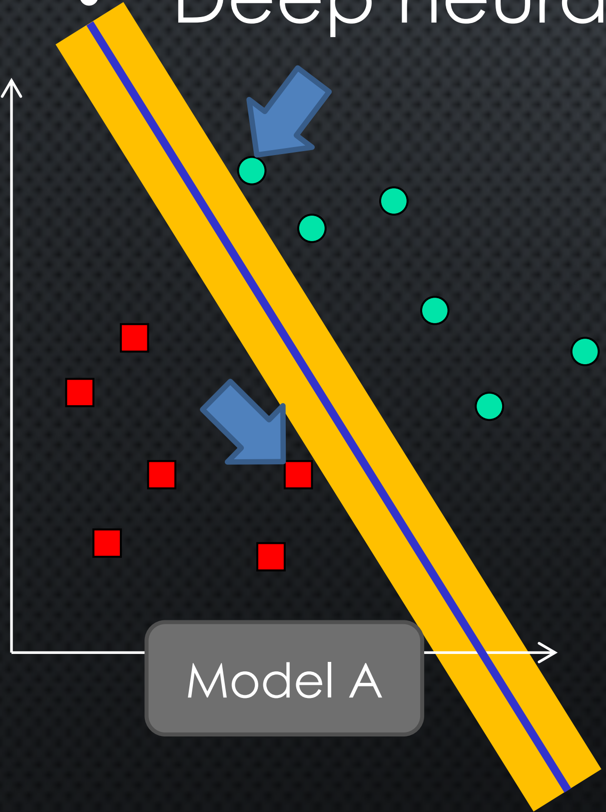



Model B

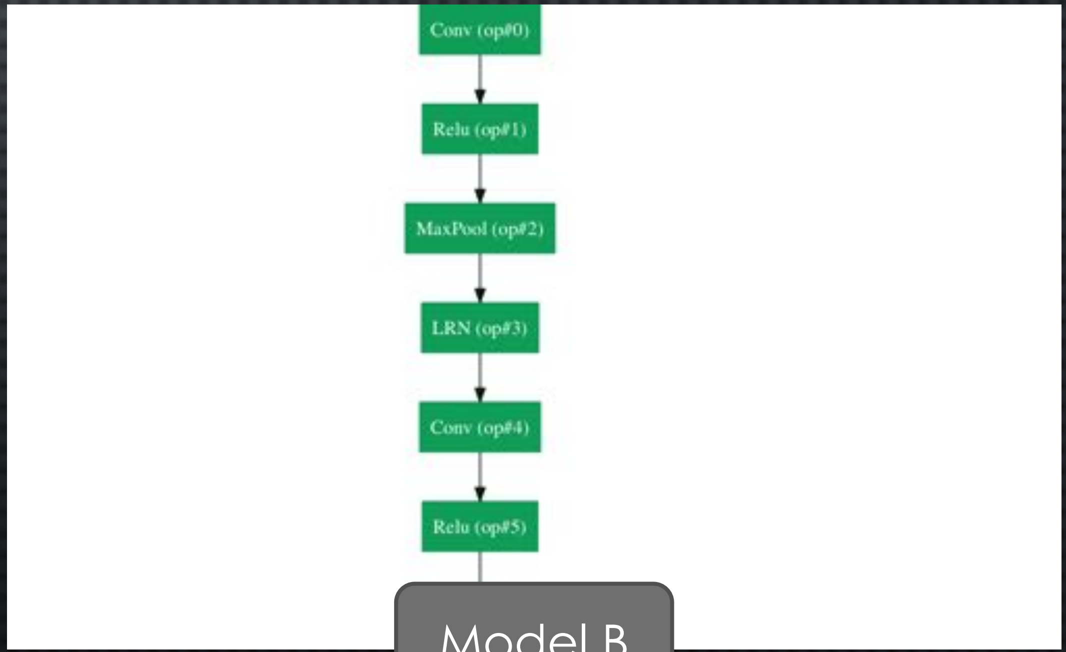
Lipton ZC. The Mythos of Model Interpretability. 2016. doi:10.1145/3233231, http://arxiv.org/abs/1606.03490

# TRANSPARENCY: ALGORITHMIC

- Is the shape of the solution understandable?
  Is convergence guaranteed?
  - Hill-climbing (MLE), margin maximizers (SVM), LR: **yes!**
  - Deep neural networks: ***not usually***



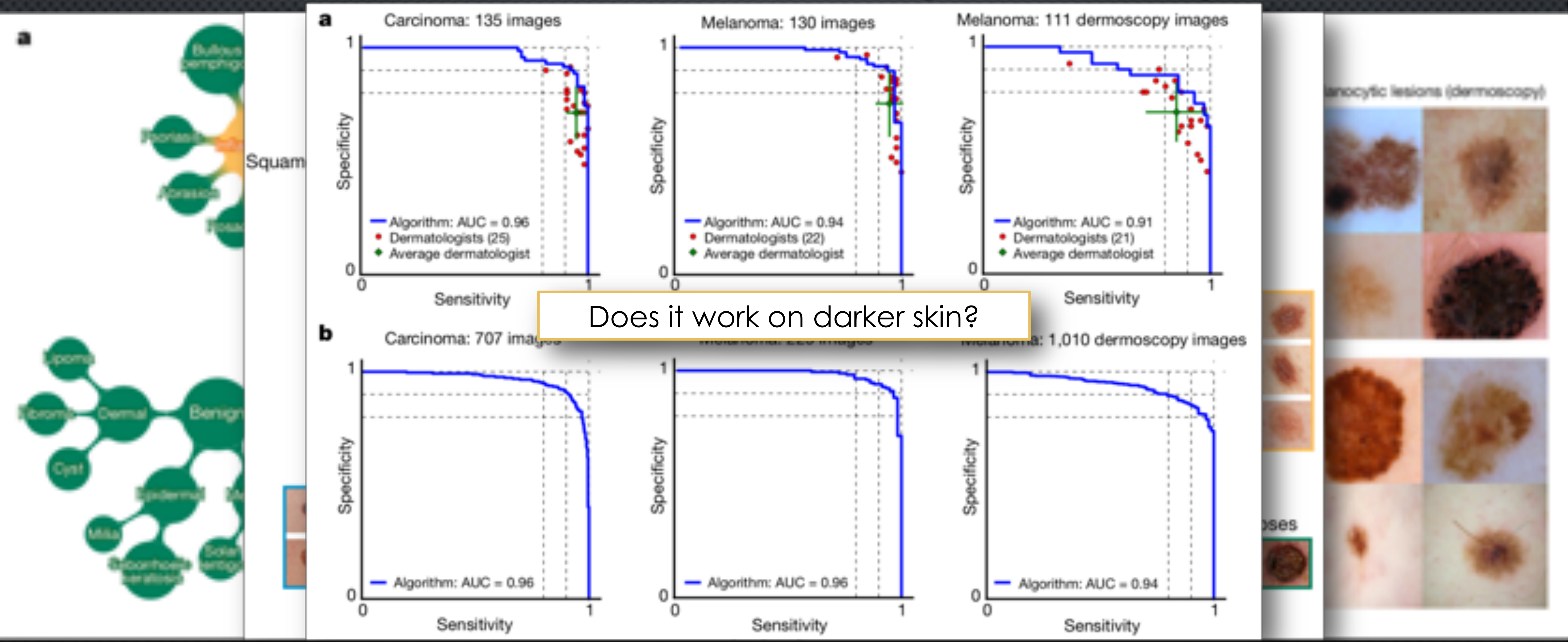
Model A




Model B

Lipton ZC. The Mythos of Model Interpretability. 2016. doi:10.1145/3233231, http://arxiv.org/abs/1606.03490

# TRANSPARENCY: VISUALIZATION (E.G., T-SNE)




Trained with 129,450 clinical images

Tested against 2 certified dermatologists.

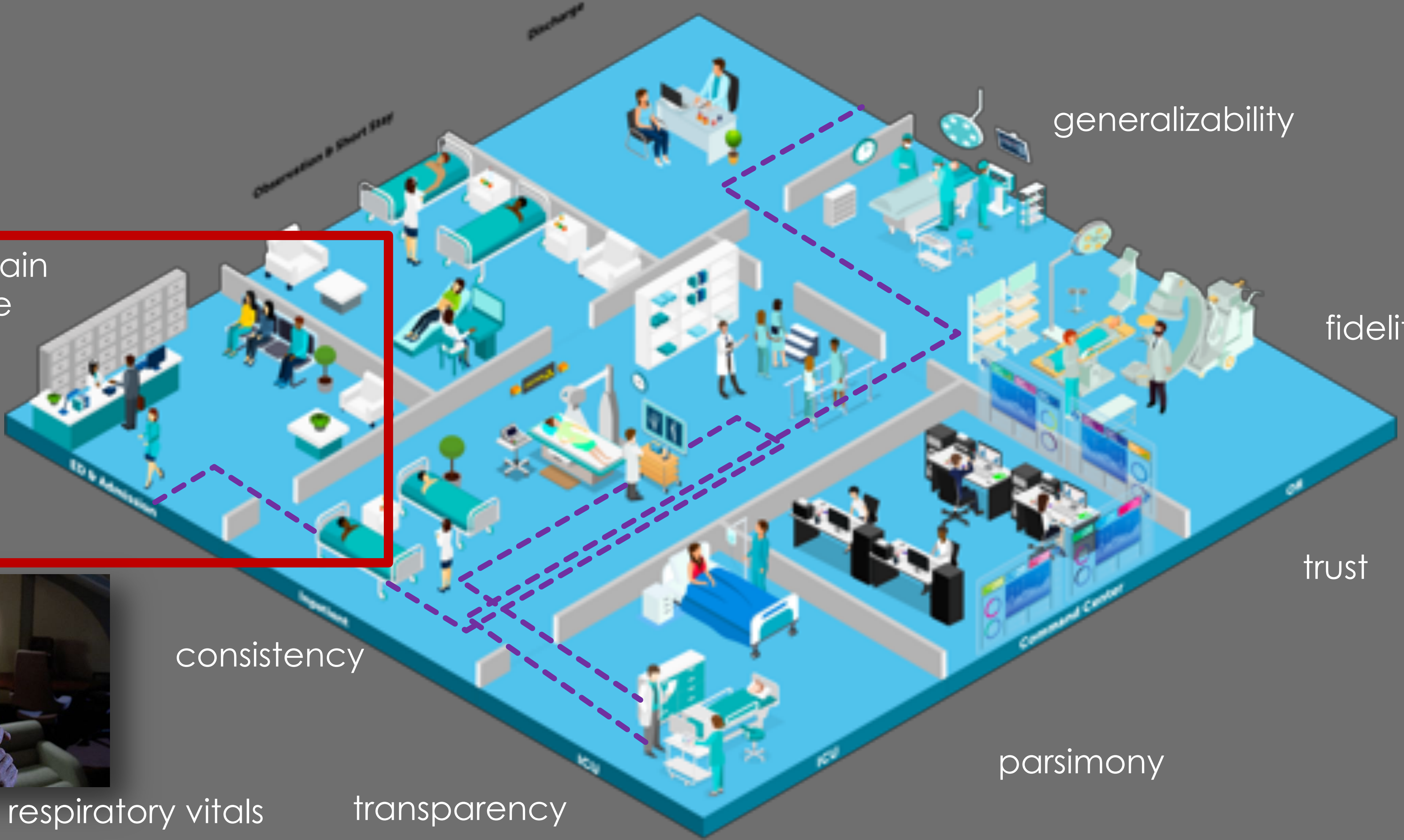
Van der Maaten L, Hinton G. Visualizing Data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605. doi:10.1007/s10479-011-0841-3
Esteva A, Kuprel B, Novoa RA, *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**:115-118

# POST-HOC INTERPRETABILITY

- "For all we know, the processes by which we humans make decisions and those by which we explain them may be distinct."
- "We caution against blindly embracing post-hoc notions of interpretability, especially when optimized to placate subjective demands. In such cases, one might - deliberately or not - optimize an algorithm to present misleading but plausible explanations."
- Correlation does not imply causation.

Lipton ZC. The Mythos of Model Interpretability. 2016. doi:10.1145/3233231, http://arxiv.org/abs/1606.03490

# JEAN-LUC'S PATH



generalizability

domain sense

fidelity

trust

consistency

parsimony

++ some respiratory vitals

transparency

# CASE STUDY: PNEUMONIA RISK

- 14,199 pneumonia patients
  - ICD-9-CM principal diagnosis of pneumonia at admission
  - 10.86% died. Bagging is used to ‘avoid overfitting’.
  - A single 😟 70/30 train/test split is used…
- 46 features extracted, e.g.,
  - Patient history: chronic lung disease (+/-), admitted to ER (+/-), age ($\mathbb{Z}$?)
  - Physical exam: heart rate ($\mathbb{R}$?), diastolic blood pressure ($\mathbb{R}$?)
  - Lab findings: potassium level ($\mathbb{R}$?), sodium level ($\mathbb{R}$?)
  - X-rays: pleural effusion, positive chest x-ray

Cooper GF, Aliferis CF, Ambrosinoa R, *et al.* An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997;**9**:107–38. doi:10.1016/s0933-3657(96)00367-3

Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

# GENERALIZED ADDITIVE MODELS (GAMS)

- Given a data set with $N$ instances, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_1^N$ , a standard GAM has the form

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j)$$

where $g(.)$ is the link function, and ~~"for each term $f_j$, $E[f_j] = 0$"~~.

  - ~~Logistic~~ regression is a special form of GAM where each $f_j$ is linear.

- To improve accuracy, pairwise interactions can be added:

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

# CASE STUDY: PNEUMONIA RISK

| Model | Pneumonia | Readmission |
|---|---|---|
| Logistic Regression | 0.8432 | 0.7523 |
| GAM | 0.8542 | 0.7795 |
| $GA^2M$ | 0.8576 | 0.7833 |
| Random Forests | 0.8460 | 0.7671 |
| LogitBoost | 0.8493 | 0.7835 |

**Table 2: AUC for different learning methods on the pneumonia and 30-day readmission tasks.**

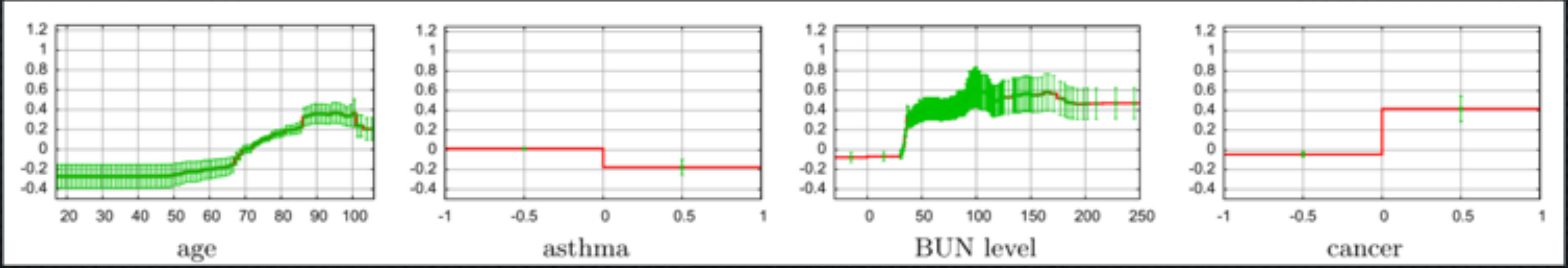
1.4% improvement

Cooper GF, Aliferis CF, Ambrosinoa R, *et al.* An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997;**9**:107–38. doi:10.1016/s0933-3657(96)00367-3

Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

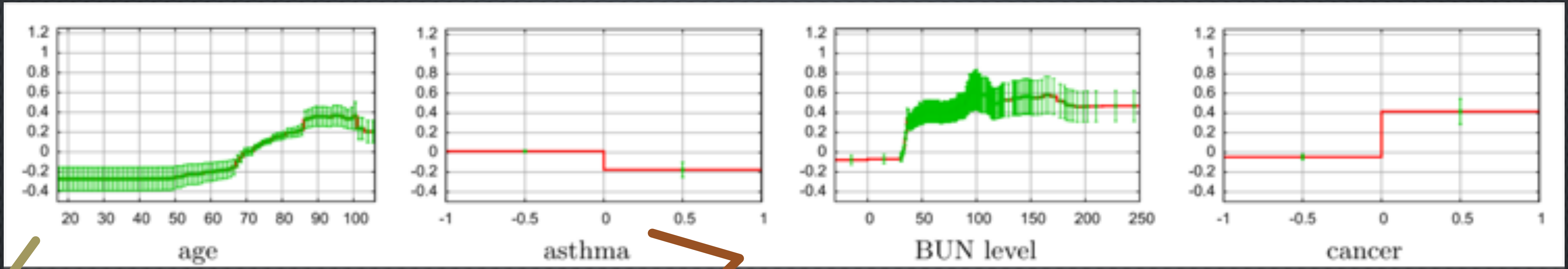
# CASE STUDY: PNEUMONIA RISK

- Sort features by 'importance'
  - Sec 5.3: ask someone fancy to rank them for you, or rank by "*drop in AUC when the term is removed*"
  - Better way (?): filter method, i.e., statistical tests of significance.
- Plot those features in terms of their ability to predict the outcome (risk score).
  - Green bars are ±1 standard deviation of the variation in the risk score ($y$-axis) measured by 100 rounds of bagging.




Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

# CASE STUDY: PNEUMONIA RISK




OK, good.
Risk of pneumonia increases with age.

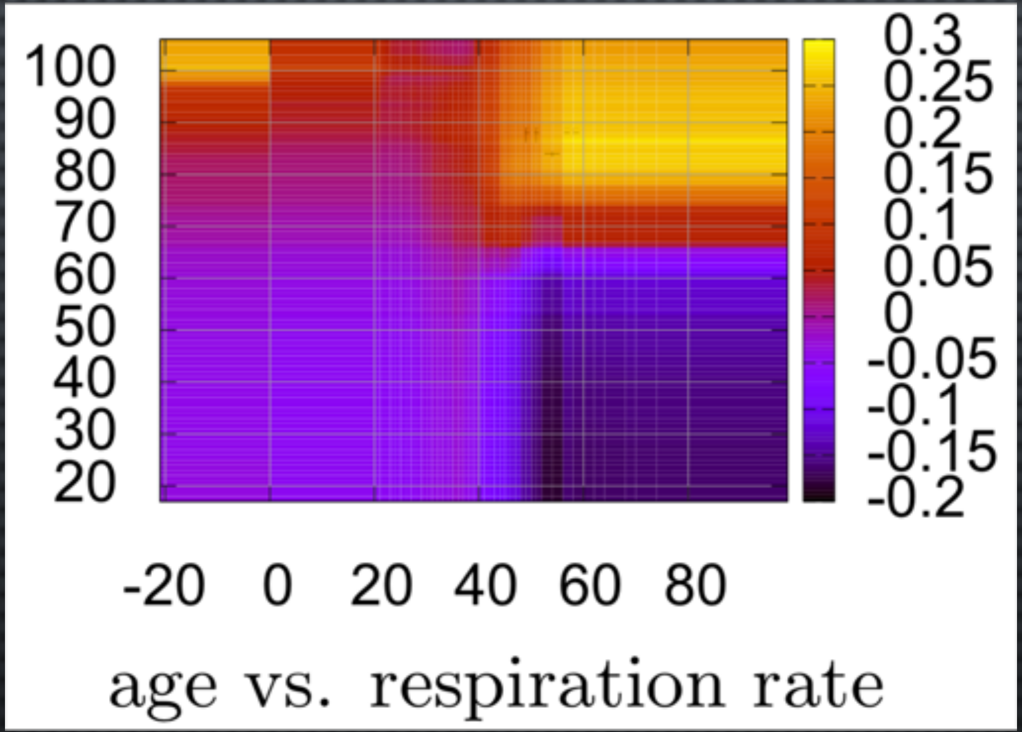
Uh oh, bad.
Risk of pneumonia decreases if you have asthma??

- It turns out, in the data, patients with a history of asthma who pre-sented with pneumonia usually were admitted not only to the hospital but directly to the ICU.

- Author's solution: remove the term, or ask a human to redraw the graph. This assumes the channel effect (or bias) is even recognized in the first place.

Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

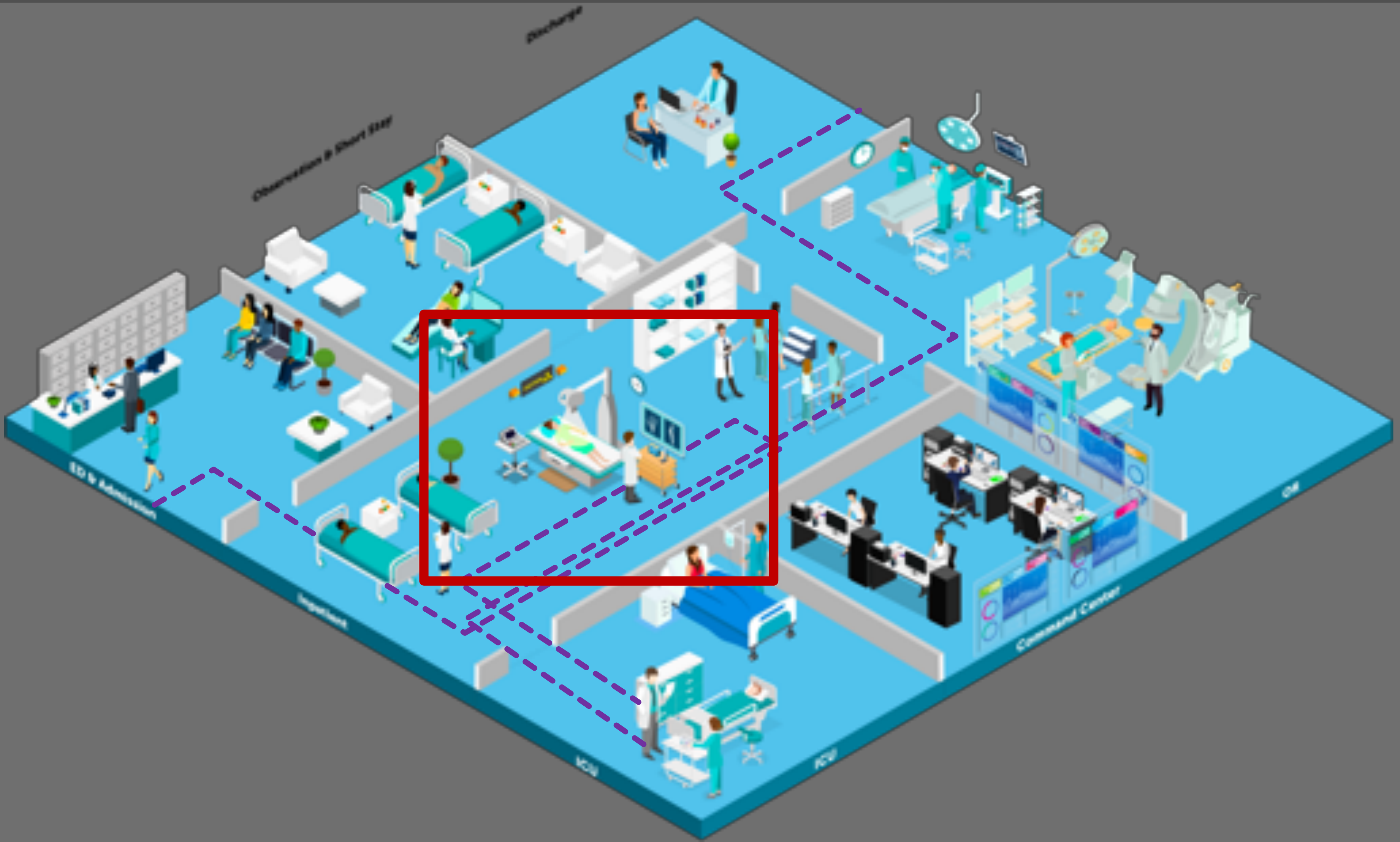
# CASE STUDY: PNEUMONIA RISK




- Sec 2.: “pairwise interactions are intelligible because they can be visualized as a heat map”

Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

# EXAMPLES AS EXPLANATIONS

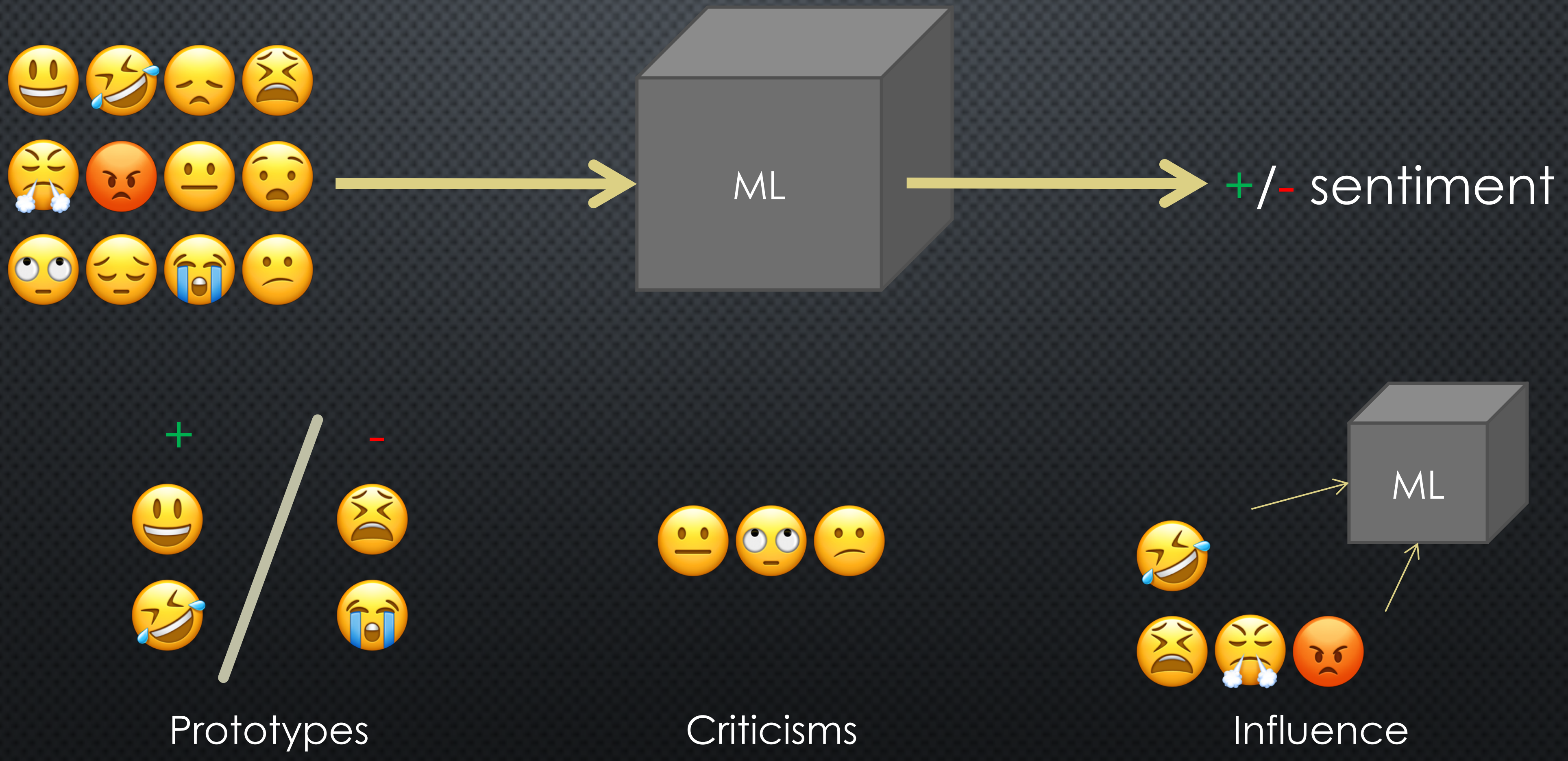


# EXAMPLES EXAMPLES EXAMPLES

- So, Jean-Luc has been admitted as an inpatient.
- The floor team now wants to **decide** whether he needs to go into the ICU.
- Like the legal system in many jurisdictions, this decision may be based on **precedent**.

- *Can we use prior **examples** to interpret decisions? To explain the model?*

# EXAMPLES AS EXPLANATIONS




Based on slide from Shalmali Joshi

# 1. PROTOTYPES BY LOCAL EXAMPLES

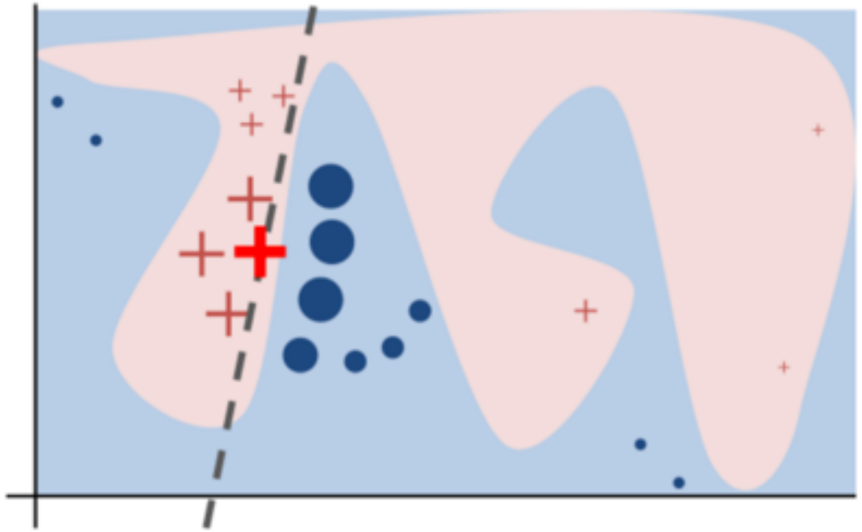


**Figure 3:** Toy example to present intuition for LIME. The black-box model's complex decision function $f$ (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using $f$, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

**Algorithm 1** Sparse Linear Explanations using LIME

**Require:** Classifier $f$, Number of samples $N$
**Require:** Instance $x$, and its interpretable version $x'$
**Require:** Similarity kernel $\pi_x$, Length of explanation $K$
  $\mathcal{Z} \leftarrow \{\}$
  **for** $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  **end for**
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with $z'_i$ as features, $f(z)$ as target
  **return** $w$

Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. 2016. doi:10.1145/1235

# 2. CRITICISMS FROM THE REAL DATA

### 3.2 Model Criticism

In addition to selecting prototype samples, MMD-critic characterizes the data points not well explained by the prototypes – which we call the model *criticism*. These data points are selected as the largest values of the witness function (5) i.e. where the similarity between the dataset and the prototypes deviate the most. Consider the cost function:

$$L(\mathsf{C}) = \sum_{l \in \mathsf{C}} \left| \frac{1}{n} \sum_{i \in [n]} k(x_i, x_l) - \frac{1}{m} \sum_{j \in \mathsf{S}} k(x_j, x_l) \right| . \qquad (9)$$


- May be most useful for explaining bias in a model, instead of a decision (?)

Kim B, Khanna R, Koyejo O. Examples are not enough, learn to criticize! criticism for interpretability.
Proc 30th Int Conf Neural Inf Process Syst 2016;:2288–96.

# 2. PSEUDO-CRITICISMS BY SYNTHESIZING DATA

**xGEMs: Generating Examplars to Explain Black-Box Models**

**Shalmali Joshi**
UT Austin
shalmali@utexas.edu

**Oluwasanmi Koyejo**
UIUC
sanmi@illinois.edu

**Been Kim**
Google Brain
beenkim@google.com

**Joydeep Ghosh**
UT Austin
jghosh@utexas.edu

**Abstract**

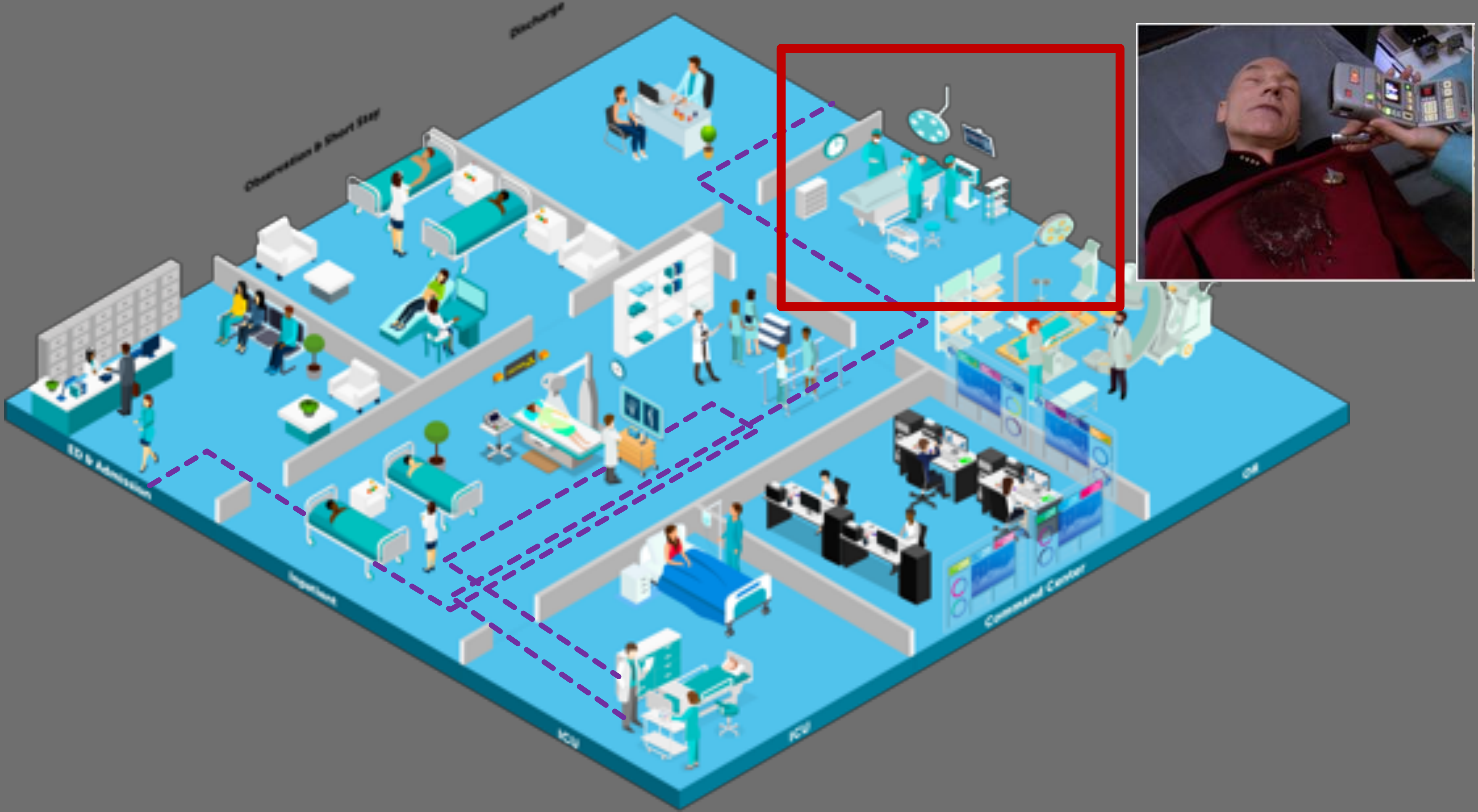
This work proposes **xGEMs** : or *manifold guided exemplars*, a framework to understand black-box classifier behavior by exploring the landscape of the underlying data manifold as data points cross decision boundaries. To do so, we train an

- Synthesize ***realistic*** data around decision boundaries.
  - Do this along a **manifold** that describes realistic data.
- May *also* be most useful for explaining bias in a model (?)

Joshi S, Koyejo O, Kim B, *et al.* xGEMs: Generating Examplars to Explain Black-Box Models. 2018;:1–12.


Figure 3: We test whether ResNet models $f_\phi^1$ and $f_\phi^2$, both trained to detect hair color but on different data distributions are confounded with gender. Two samples for classifiers $f_\phi^1$ (first sub row) and $f_\phi^2$ (second sub row) are shown. The leftmost image is the original figure, followed by its reconstruction from the encoder $F_\psi$. Reconstructions are plotted as Algorithm 1 (with $\lambda = 0.01$) progresses toward crossing the decision boundary. The red bar indicates change in hair color label indicated at the top of each image along with the confidence of prediction. The label at the bottom indicates gender as predicted by $\hat{g}$. For both samples, classifier $f_\phi^1$, trained on biased data changes the gender ($1^{st}$ and $3^{rd}$ rows) while crossing the decision boundary whereas the other black-box does not.

# LIVE, PIXEL-LEVEL ANNOTATIONS



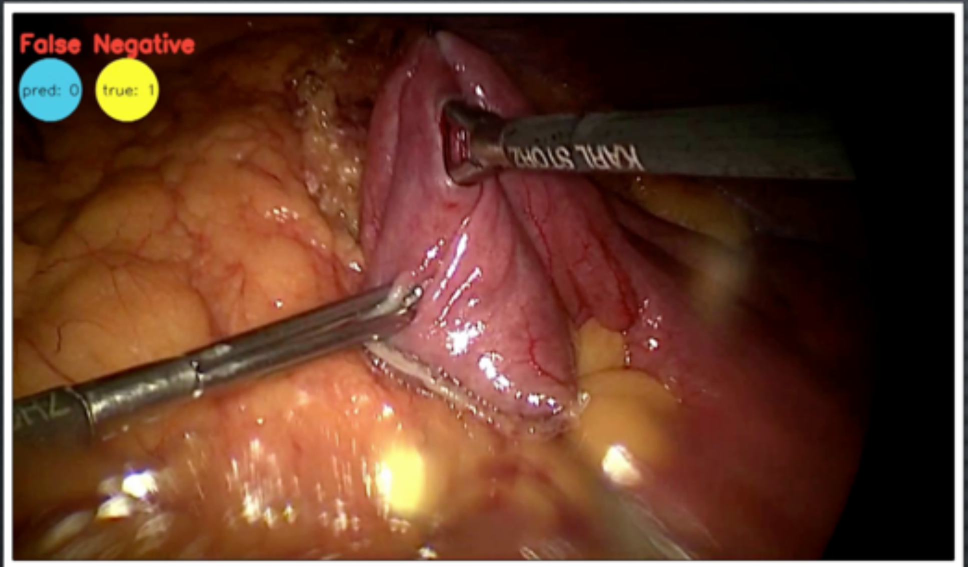
# MASKS AND HEATMAPS

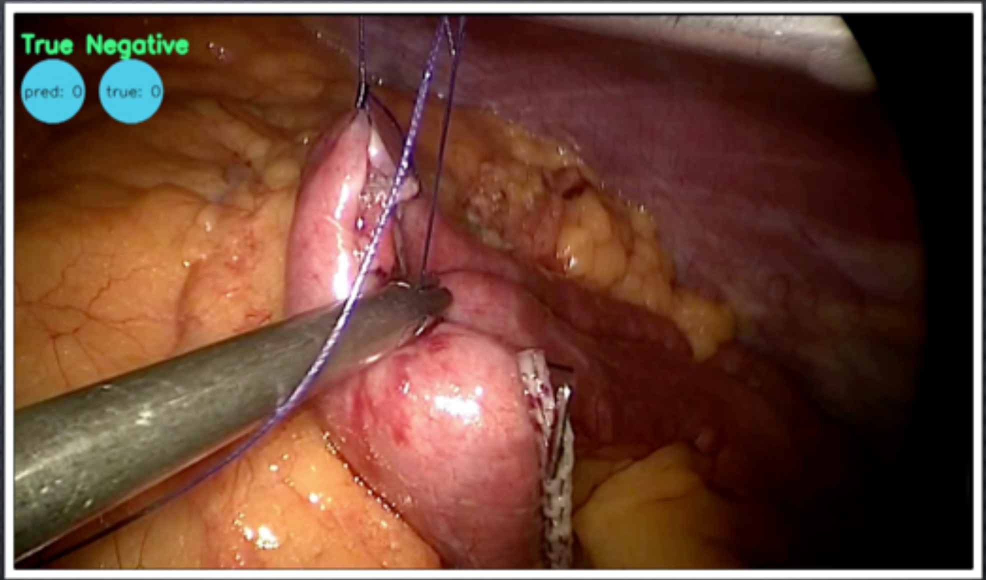
- So, while we got sidetracked using exemplars to explain the model itself, Jean-Luc was stabbed through the heart by a Nausicaan (or, more realistically, he took a turn for the worse).
- He needs an emergency **surgery**.

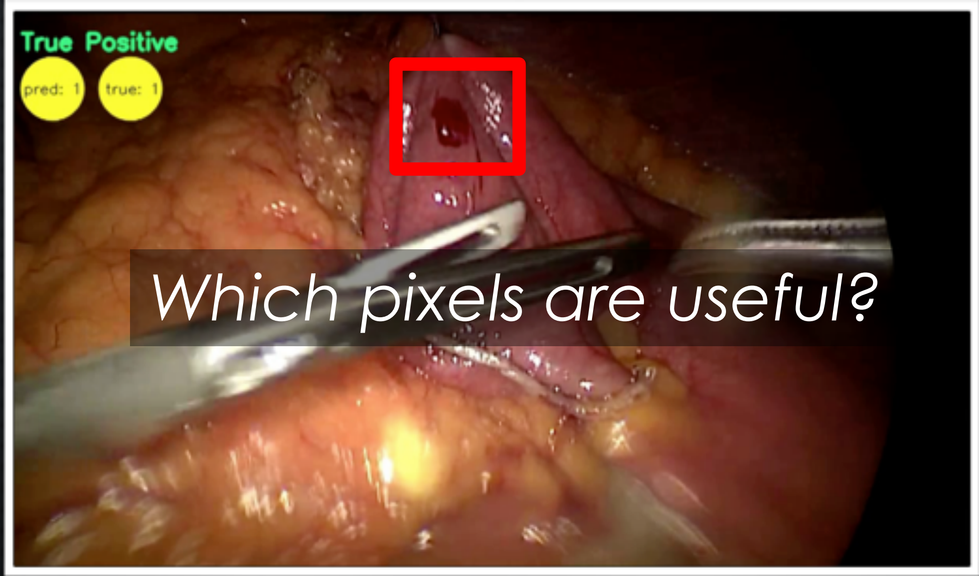
- *In surgery, we want to identify aspects* ***within*** *the live video.*

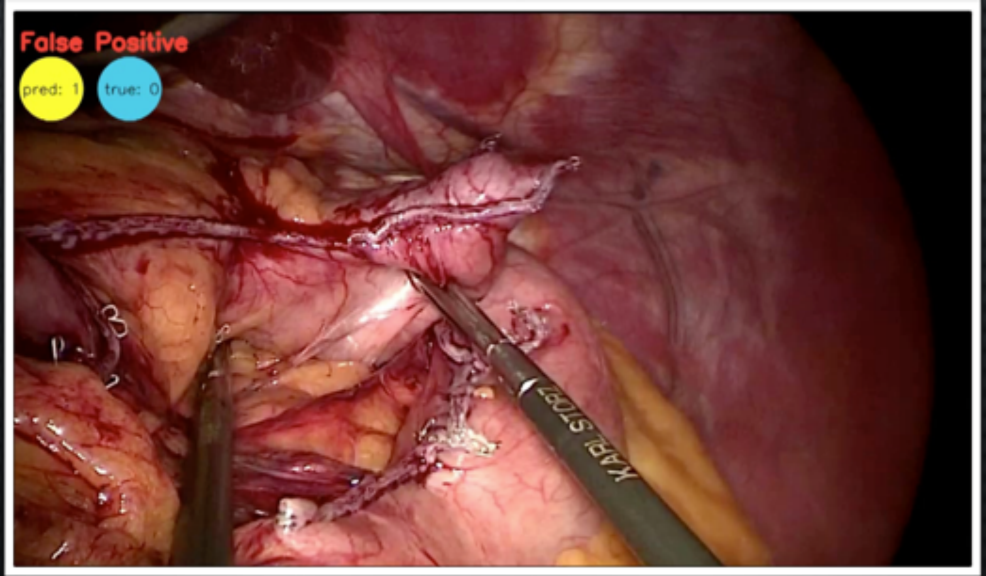
*Warning: blood on next slide!*

# BLEEDING DETECTION IN SURGERY

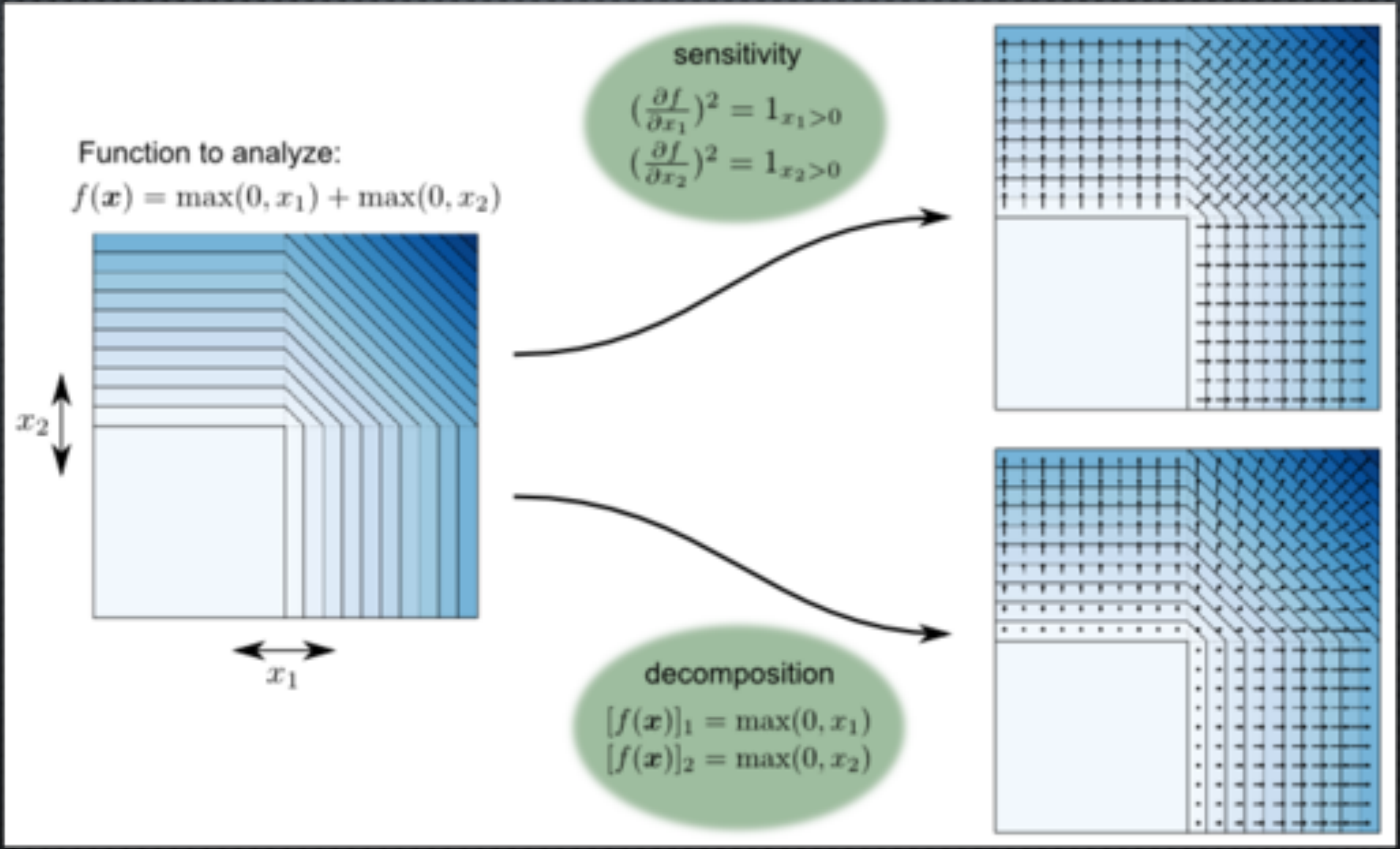


# DECOMPOSABILITY – MOTIVATING EXAMPLE




Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 2017;65:211–22.

# DECOMPOSABILITY

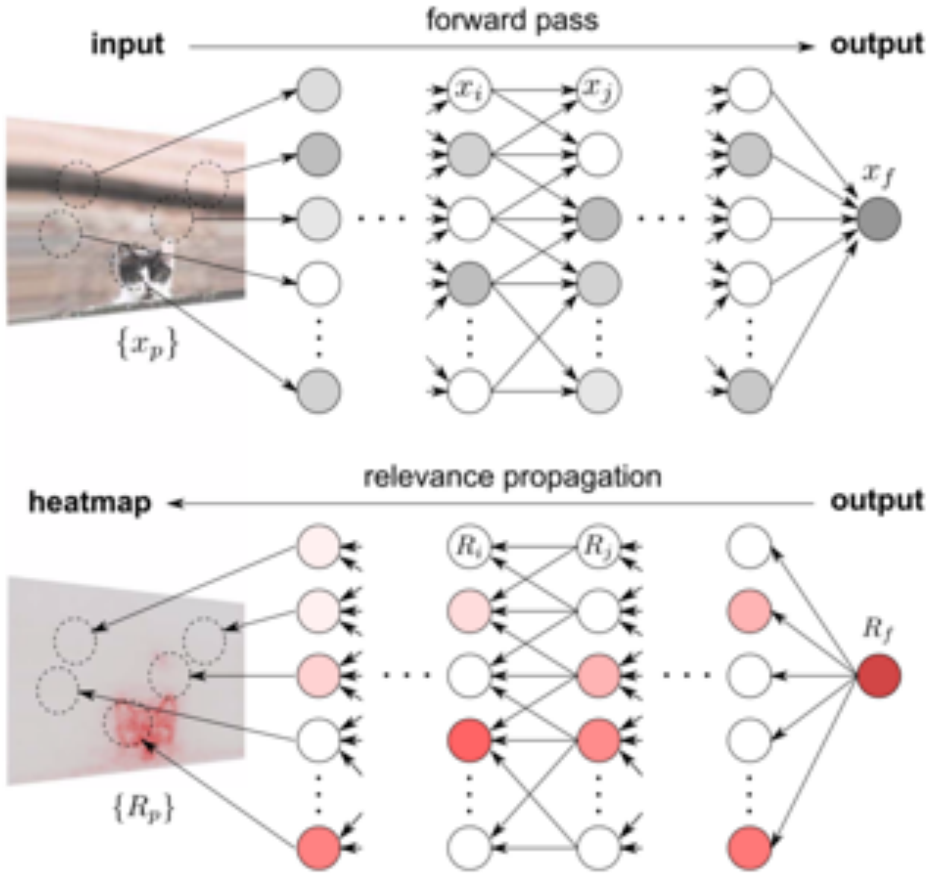



**Fig. 2.** Computational flow of deep Taylor decomposition. A prediction for the class "cat" is obtained by forward-propagation of the pixel values $\{x_p\}$, and is encoded by the output neuron $x_f$. The output neuron is assigned a relevance score $R_f = x_f$ representing the total evidence for the class "cat". Relevance is then backpropagated from the top layer down to the input, where $\{R_p\}$ denotes the pixel-wise relevance scores, that can be visualized as a heatmap.

First-order Taylor decomposition

$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \left(\frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}}\right)^{\top} \cdot (\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \varepsilon = 0 + \sum_p \underbrace{\frac{\partial f}{\partial x_p}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} \cdot (x_p - \widetilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon,$$

$$R_j = \left(\frac{\partial R_j}{\partial \{x_i\}}\Big|_{\{\widetilde{x}_i\}^{(j)}}\right)^{\top} \cdot (\{x_i\} - \{\widetilde{x}_i\}^{(j)}) + \varepsilon_j = \sum_i \underbrace{\frac{\partial R_j}{\partial x_i}\Big|_{\{\widetilde{x}_i\}^{(j)}} \cdot (x_i - \widetilde{x}_i^{(j)})}_{R_{ij}} + \varepsilon_j,$$

Deep Taylor decomposition of 'relevance' at neuron *j*

Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 2017;65:211–22.

# DECOMPOSABILITY

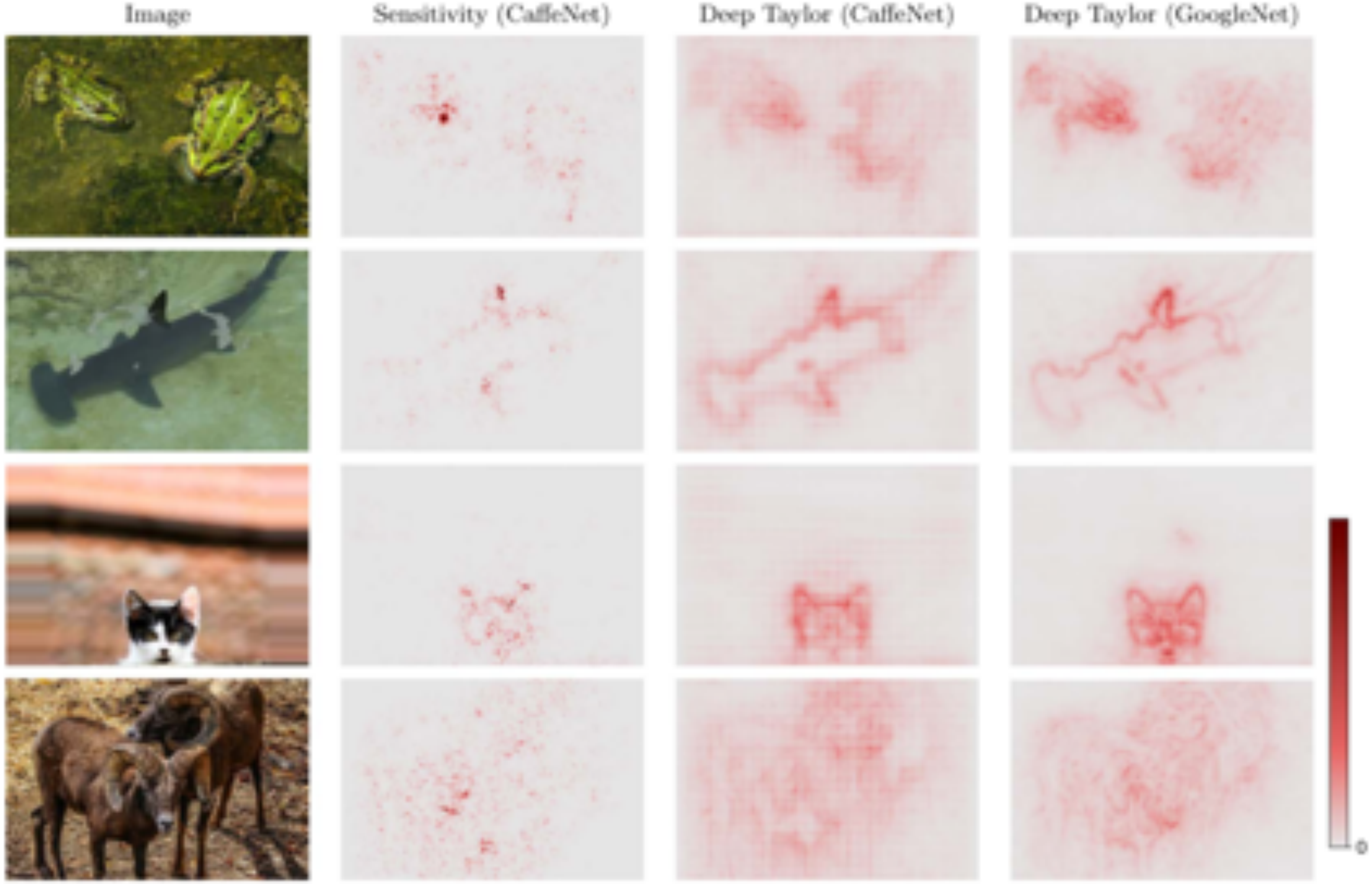



**Fig. 7.** Images of different ILSVRC classes ("frog", "shark", "cat", and "sheep") given as input to a deep network, and displayed next to the corresponding heatmaps. Heatmap scores are summed over all color channels of the image.

Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 2017;65:211–22.

# ANCHORS

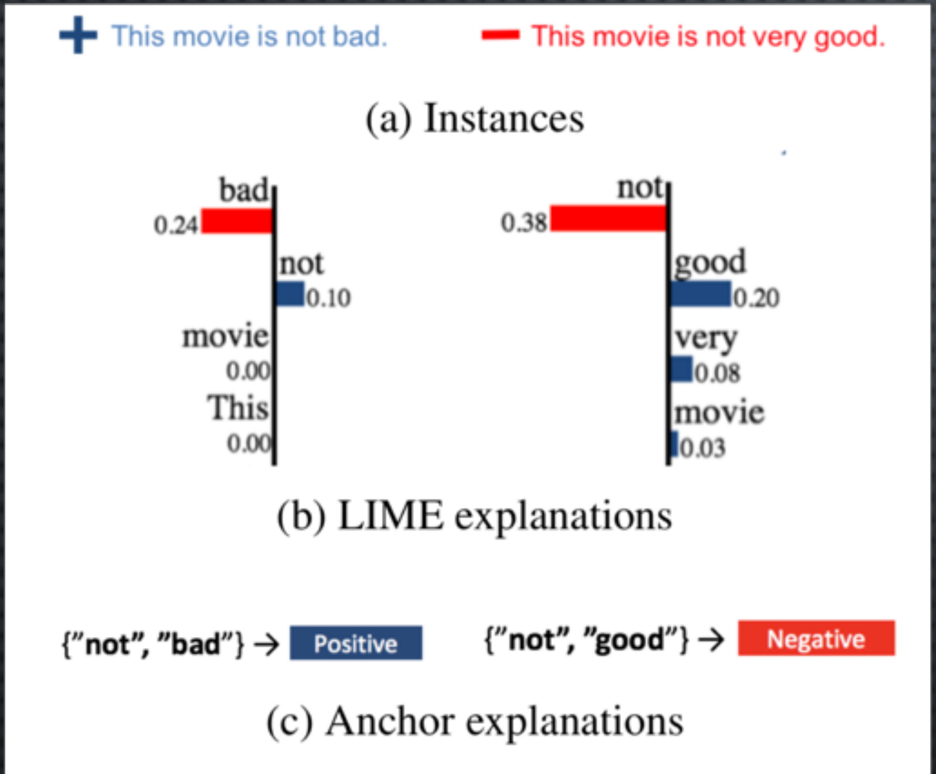



Figure 1: Sentiment predictions, LSTM

Let $A$ be a rule (set of predicates) acting on such an interpretable representation, such that $A(x)$ returns 1 if all its feature predicates are true for instance $x$. For example, in Figure 2a (top), $x =$ "This movie is not bad.", $f(x) = \textit{Positive}$, $A(x) = 1$ where $A = \{$"not", "bad"$\}$. Let $\mathcal{D}(\cdot|A)$ denote the conditional distribution when the rule $A$ applies (e.g. similar texts where "not" and "bad" are present, Figure 2a bottom). $A$ is an *anchor* if $A(x) = 1$ and $A$ is a sufficient condition for $f(x)$ with high probability — in our running example, if a sample $z$ from $\mathcal{D}(z|A)$ is likely predicted as *Positive* (i.e. $f(x) = f(z)$). Formally $A$ is an anchor if,

$$\mathbb{E}_{\mathcal{D}(z|A)}[\mathbb{1}_{f(x)=f(z)}] \geq \tau, A(x) = 1. \qquad (1)$$

Ribeiro MT, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. In: Proceedings of AAAI18. 2018.

# RELEVANCE MASKS

(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

| What animal is featured in this picture ? | dog |
|---|---|
| What floor is featured in this picture? | dog |
| What toenail is paired in this flowchart ? | dog |
| What animal is shown on this depiction ? | dog |

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

| | |
|---|---|
| **Where** is the **dog**? | on the floor |
| **What color** is the **wall?** | white |
| **When** was this picture taken? | during the day |
| **Why** is he lifting his paw? | to play |

(e) **VQA:** More example anchors (in bold)

Figure 3: Anchor Explanations for Image Classification and Visual Question Answering (VQA)

| | If | Predict |
|---|---|---|
| **adult** | No capital gain or loss, never married | $\leq$ 50K |
| | Country is US, married, work hours $> 45$ | $>$ 50K |
| **rcdv** | No priors, no prison violations and crime not against property | Not rearrested |
| | Male, black, 1 to 5 priors, not married, and crime not against property | Re-arrested |
| **lending** | FICO score $\leq 649$ | Bad Loan |
| | $649 \leq$ FICO score $\leq 699$ and $\$5,400 \leq$ loan amount $\leq \$10,000$ | Good Loan |

Table 3: Generated anchors for Tabular datasets

Ribeiro MT, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations. In: Proceedings of AAAI18. 2018.

# SPEAKING OF SURGERY…

nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-018-0304-0

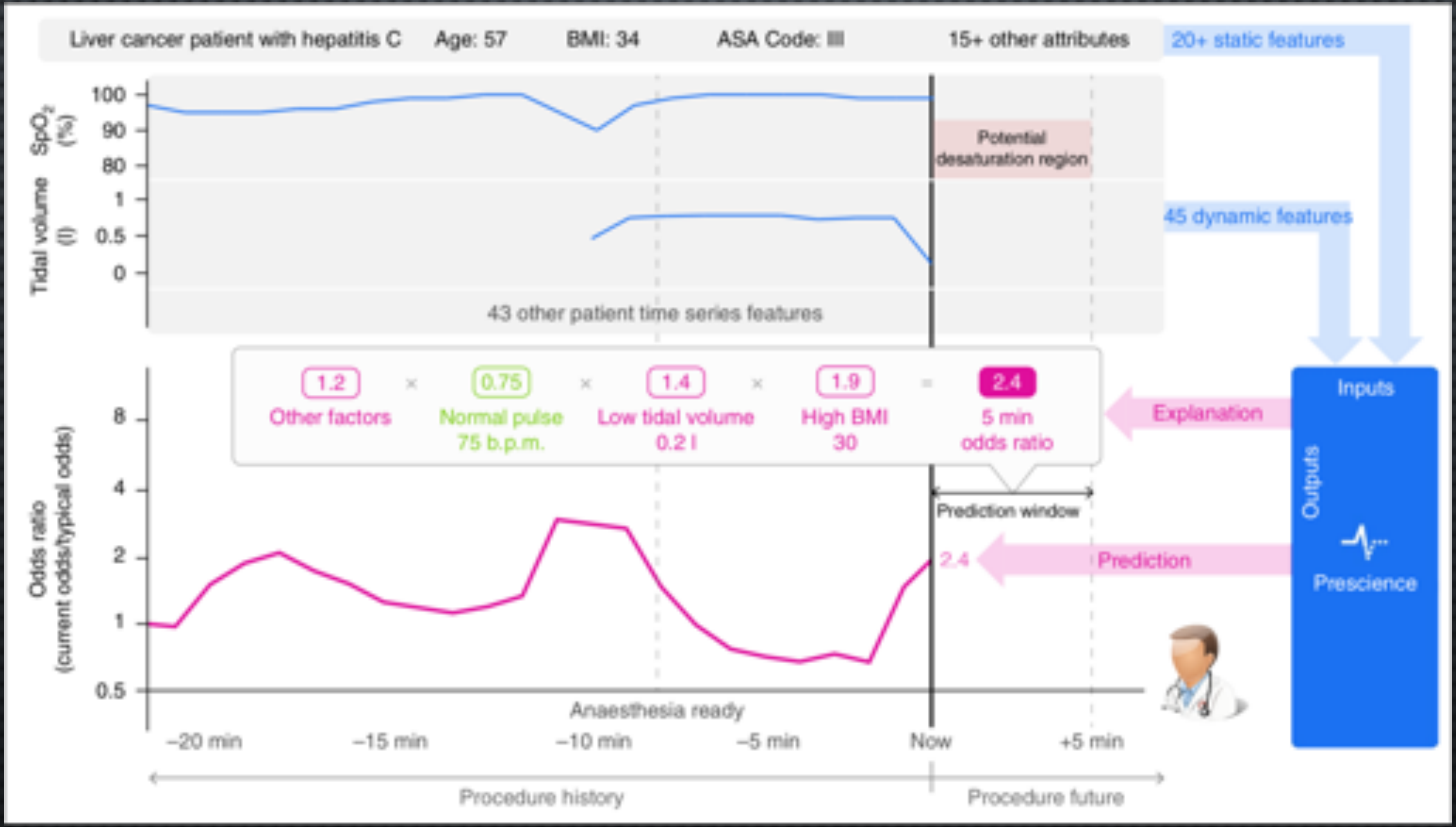
## Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

Scott M. Lundberg[1], Bala Nair[2,3,4], Monica S. Vavilala[2,3,4], Mayumi Horibe[5], Michael J. Eisses[2,6], Trevor Adams[2,6], David E. Liston[2,6], Daniel King-Wai Low[2,6], Shu-Fang Newman[2,3], Jerry Kim[2,6] and Su-In Lee[1*]

Although anaesthesiologists strive to avoid hypoxaemia during surgery, reliably predicting future intraoperative hypoxaemia is not possible at present. Here, we report the development and testing of a machine-learning-based system that predicts the risk of hypoxaemia and provides explanations of the risk factors in real time during general anaesthesia. The system, which was trained on minute-by-minute data from the electronic medical records of over 50,000 surgeries, improved the performance of anaesthesiologists by providing interpretable hypoxaemia risks and contributing factors. The explanations for the predictions are broadly consistent with the literature and with prior knowledge from anaesthesiologists. Our results suggest that if anaesthesiologists currently anticipate 15% of hypoxaemia events, with the assistance of this system they could anticipate 30%, a large portion of which may benefit from early intervention because they are associated with modifiable factors. The system can help improve the clinical understanding of hypoxaemia risk during anaesthesia care by providing general insights into the exact changes in risk induced by certain characteristics of the patient or procedure.

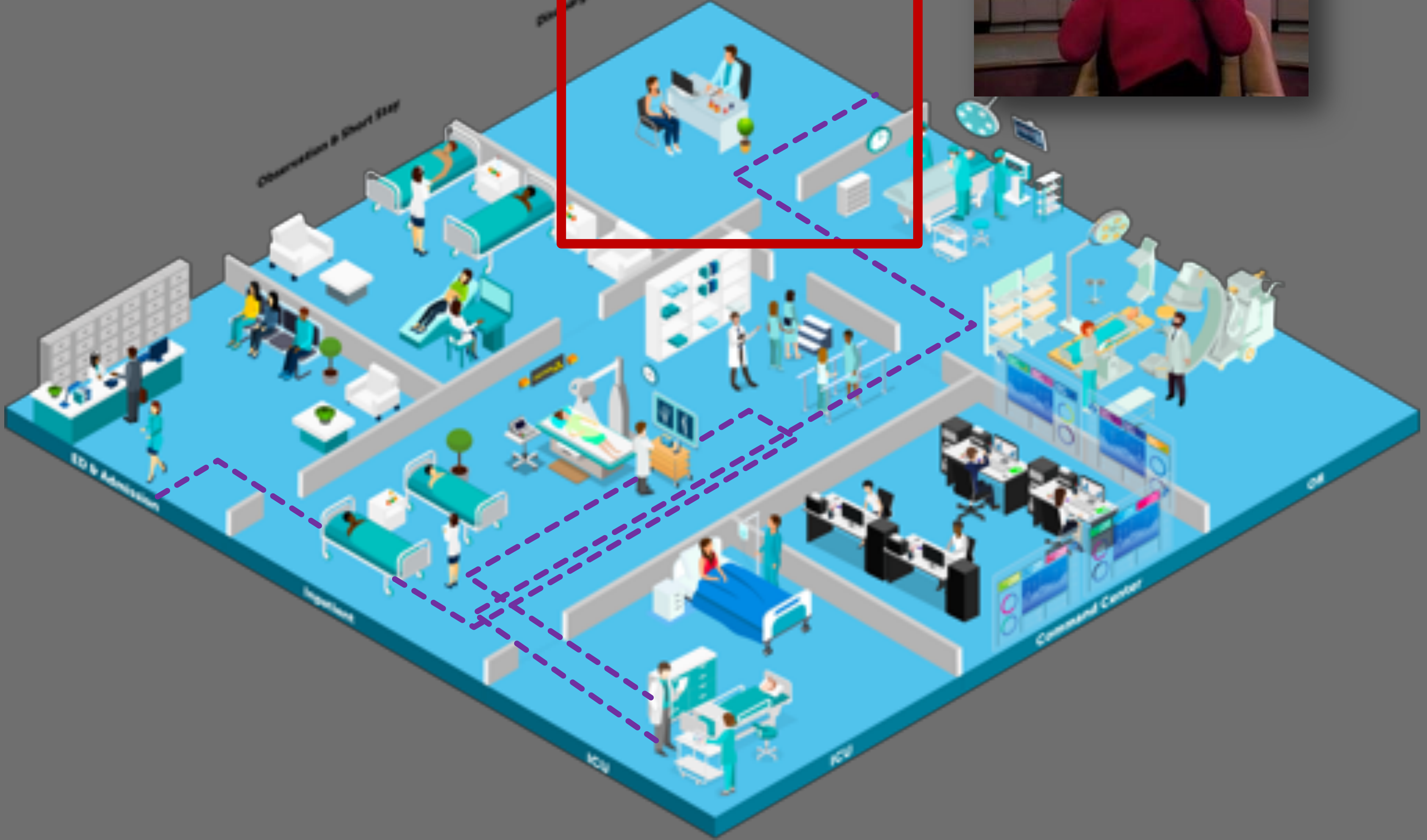
Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749–60. doi:10.1038/s41551-018-0304-0

# SPEAKING OF SURGERY…




Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749–60. doi:10.1038/s41551-018-0304-0

# DISCHARGE

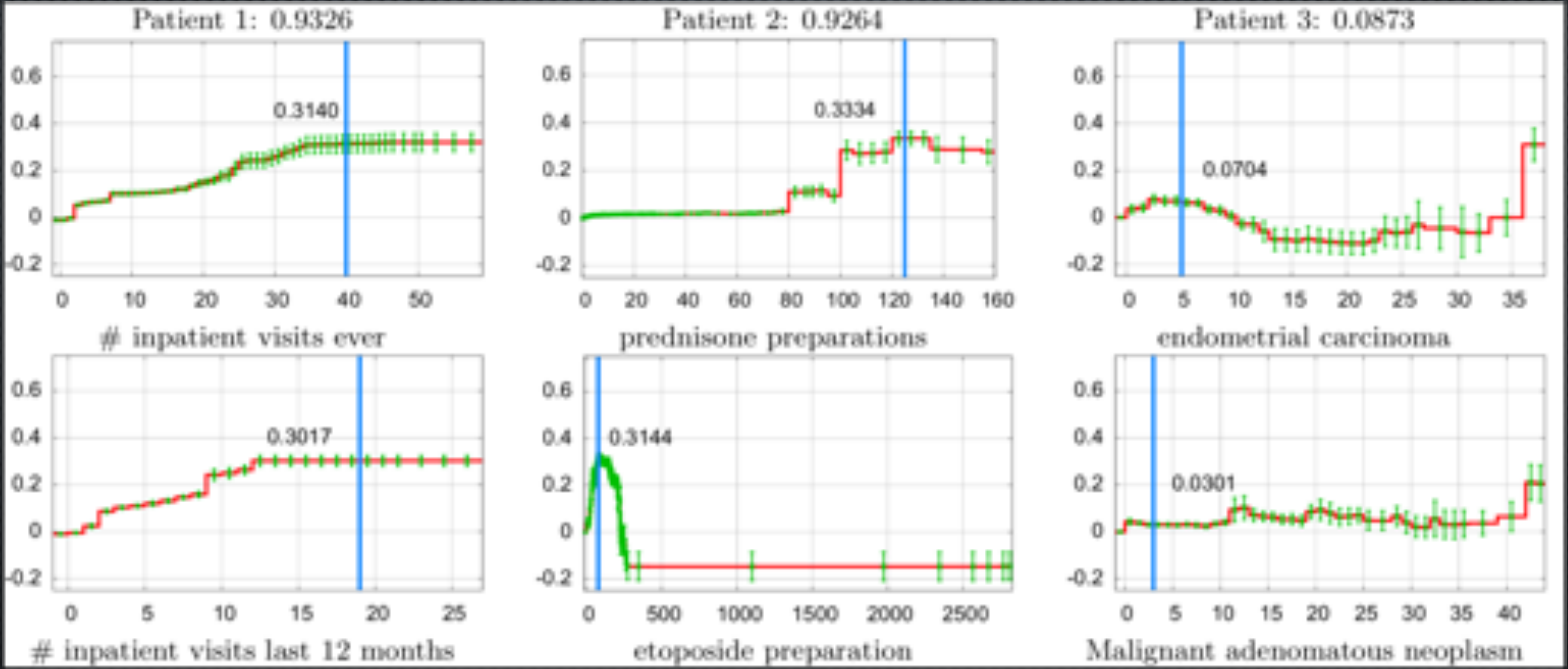


Observation & Short Stay
ED & Admission
Inpatient
ICU
ICU
Command Center
OR

# DISCHARGE, AND DEBRIEFING

- Jean-Luc had a successful heart surgery and wants to get back to his ship.
- He wants to know:
  - how likely he is to be **re-admitted**.
  - ~~is it part of a Romulan plot? A ploy to start a war?~~
  - anything at all about his experience.

# RE-ADMISSION RISK (CARUANA ET AL, SLIGHT RETURN)




Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for HealthCare. In: Proceedings of KDD. 2015. 1721–30. doi:10.1145/2783258.2788613

# MORE TEXT

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

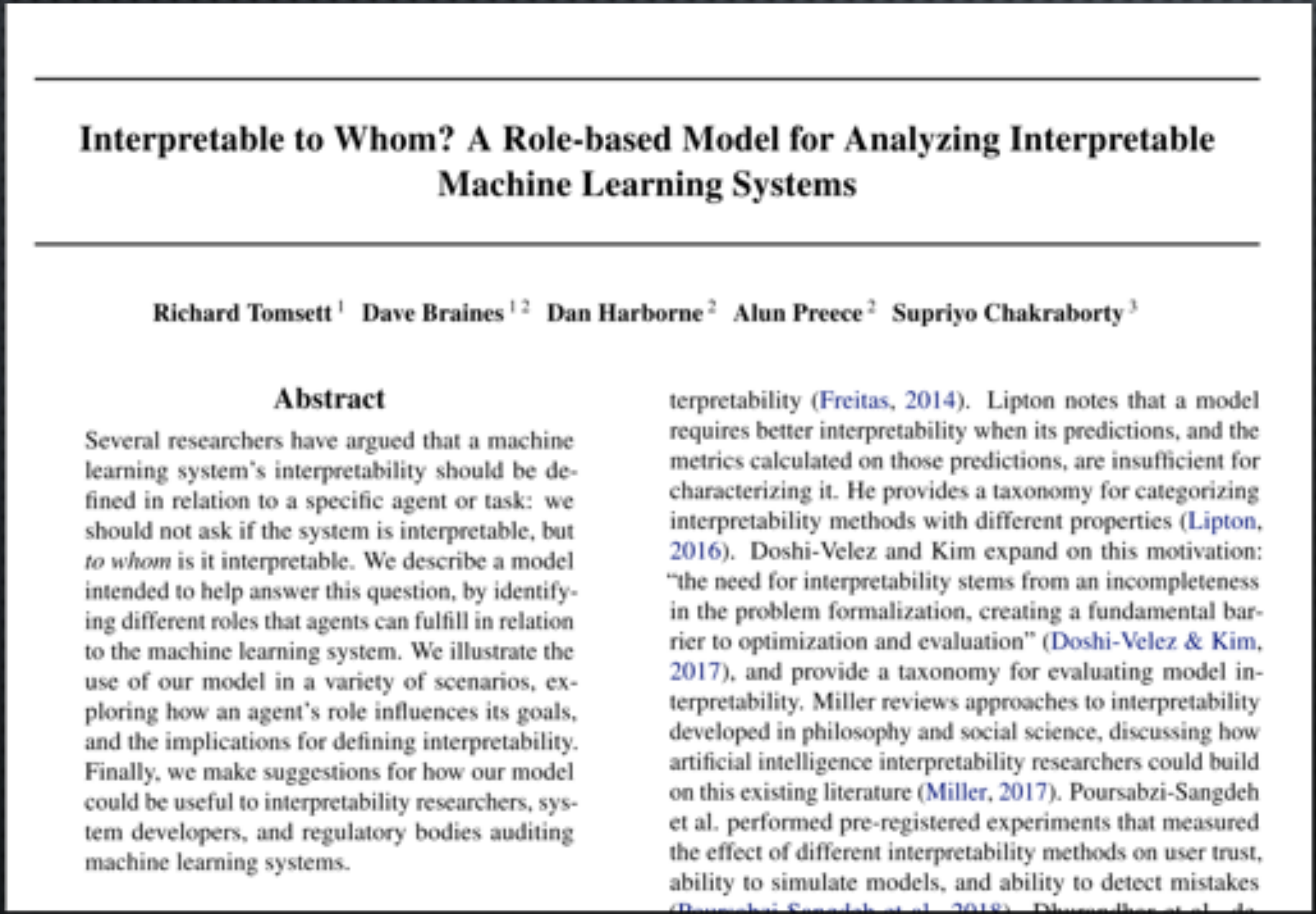
a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate ... ... m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation , not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

**Figure 3:** Examples of extracted rationales indicating the sentiments of various aspects. The extracted texts for appearance, smell and palate are shown in red, blue and green color respectively. The last example is shortened for space.

- Train an extractive summarizer (‘**generator**’) and an **encoder** *simultaneously*

Lei T, Barzilay R, Jaakkola T. Rationalizing Neural Predictions. In: Proceedings of EMNLP: 2016. doi:10.1177/1087057107312127

# INTERPRETABLE TO WHOM?



## Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems

Richard Tomsett [1] Dave Braines [1 2] Dan Harborne [2] Alun Preece [2] Supriyo Chakraborty [3]

### Abstract

Several researchers have argued that a machine learning system's interpretability should be defined in relation to a specific agent or task: we should not ask if the system is interpretable, but *to whom* is it interpretable. We describe a model intended to help answer this question, by identifying different roles that agents can fulfill in relation to the machine learning system. We illustrate the use of our model in a variety of scenarios, exploring how an agent's role influences its goals, and the implications for defining interpretability. Finally, we make suggestions for how our model could be useful to interpretability researchers, system developers, and regulatory bodies auditing machine learning systems.

terpretability (Freitas, 2014). Lipton notes that a model requires better interpretability when its predictions, and the metrics calculated on those predictions, are insufficient for characterizing it. He provides a taxonomy for categorizing interpretability methods with different properties (Lipton, 2016). Doshi-Velez and Kim expand on this motivation: "the need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation" (Doshi-Velez & Kim, 2017), and provide a taxonomy for evaluating model interpretability. Miller reviews approaches to interpretability developed in philosophy and social science, discussing how artificial intelligence interpretability researchers could build on this existing literature (Miller, 2017). Poursabzi-Sangdeh et al. performed pre-registered experiments that measured the effect of different interpretability methods on user trust, ability to simulate models, and ability to detect mistakes

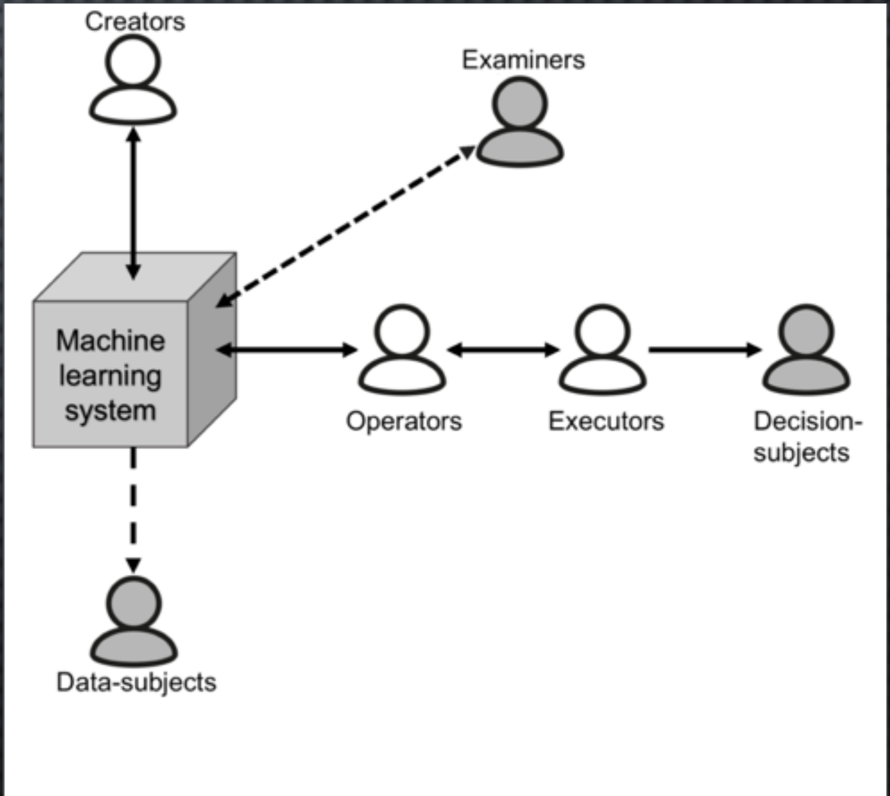



*Figure 1.* Illustration of a machine learning ecosystem. Direction of arrow indicates direction of interaction (e.g., data-subjects do not interact with the system, but the system has their data)

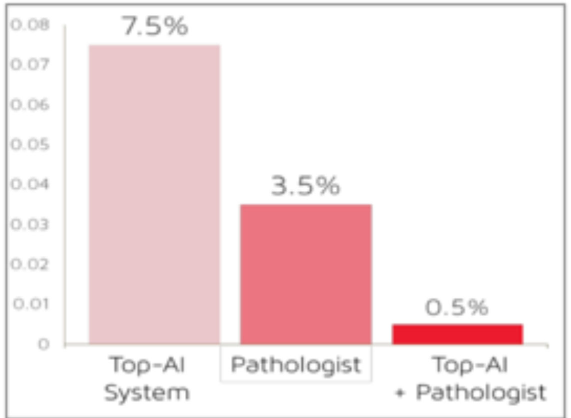
Tomsett R, Braines D, Harborne D, et al. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems.; 2018

# ACTIVE LEARNING

National Institutes of Health (NIH) grants-supported research

ARTIFICIAL INTELLIGENCE FOR COMPUTATIONAL PATHOLOGY

*Image interpretation plays a central role in the pathologic diagnosis of cancer. Since the late 19th century, the primary tool used by pathologists to make definitive cancer diagnoses is the microscope. Pathologists diagnose cancer by manually examining stained sections of cancer tissues to determine the cancer subtype. Pathologic diagnosis using conventional methods is labor-intensive with poor reproducibility and quality concerns. New approaches use fundamental AI research to build tools to make pathologic analysis more efficient, accurate, and predictive. In the 2016 Camelyon Grand Challenge for metastatic cancer detection,[69] the top-performing entry in the competition was an AI-based computational system that achieved an error rate of 7.5%.[70] A pathologist reviewing the same set of evaluation images achieved an error rate of 3.5%. Combining the predictions of the AI system with the pathologist lowered the error rate to down to 0.5%, representing an 85% reduction in error (see image).[71] This example illustrates how fundamental research in AI can drive the development of high performing computational systems that offer great potential for making pathological diagnoses more efficient and more accurate.*




AI significantly reduces pathologist error rate in the identification of metastatic breast cancer from sentinel lymph node biopsies.

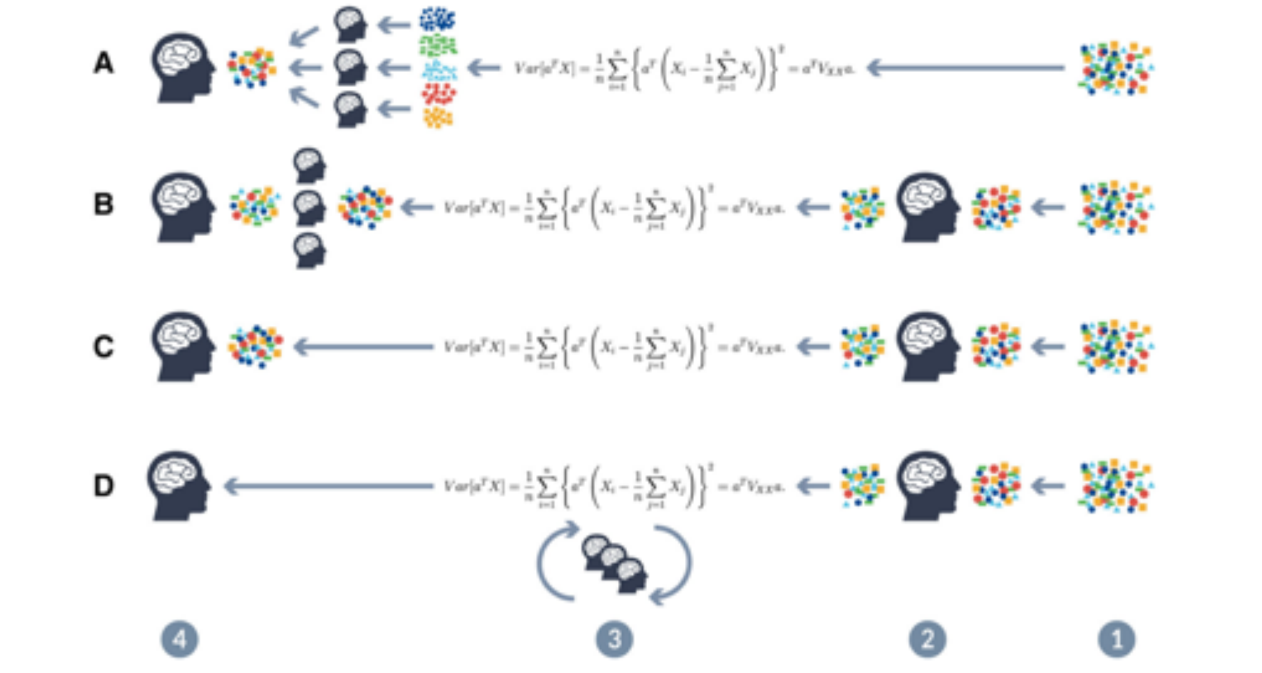



**Fig. 1** Four different ML-pipelines: *A* unsupervised, *B* supervised—e.g., humans are providing labels for training data sets and/or select features, *C* semi-supervised, *D* shows the iML human-in-the-loop approach: the important issue is that humans are not only involved in pre-processing, by selecting data or features, but actually during the learning phase, directly interacting with the algorithm, thus shifting away the black-box problem to a wished glass-box, *1* input data, *2* pre-processing phase, *3* human agent(s) interacting with the computational agent(s), allowing for crowdsourcing or gamification approaches, *4* final check done by the human expert


Holzinger, A. (2016). Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? *Springer Brain Informatics*, *3*(1), in print. http://doi.org/10.1007/s40708-016-0042-6

# SUMMARY

- By following Jean-Luc through a hospital, we've also visited the three main general approaches to XAI:
  - Explanations by **influence** of its input features
  - Explanations by **examples** (both actual and synthetic)
  - Explanations by **heatmaps** or **masks**

- *How will (or must?!) XAI be used in practice?*

# REGULATION AND THE LAW

# OBLIGATORY CARTOON AND AWKWARD PAUSE

# TO ERR IS HUMAN. DOUBLE STANDARDS

- **Humans** are notoriously **bad** with information.
  - Patients **misread** or **miscommunicate** their own symptoms.
    - Nearly **half** of American adults have difficulty understanding and acting upon health information (IOM, 2004).
  - Faulty memory; skill obsolescence;  cognitive biases; cognitive/time limitations; **recency biases**; other human biases.
    - *Diagnoses* correlate with advertising and media exposure.

- Winters *et al.* (2012) showed that ~40,500 patients die in ICU, in the USA, each year due to misdiagnosis.


http://www.nap.edu/openbook.php?record_id=10883&page=1
Winters *et al.* (2012) Diagnostic errors in the intensive care unit: a systematic review of autopsy studies.
*BMJ Qual Saf* 2012;**21**:894-902

# TO ERR IS HUMAN. DOUBLE STANDARDS

- Graber et al. (2005) studied one hundred cases of **diagnostic error** involving internists …
  - **Cognitive factors** contributed to 74% of cases.
    - Most common cause: ‘premature closure’.

- Eddy (1990) showed top surgeons descriptions of surgical problems and asked: *Should the patient have surgery?*
  - 50% said **Yes**, 50% said **No**.
  - 40% gave conflicting answers upon retesting.


Graber et al. (2005) Diagnostic Error in Internal Medicine. Arch Intern Med., 165(13):1493-1499

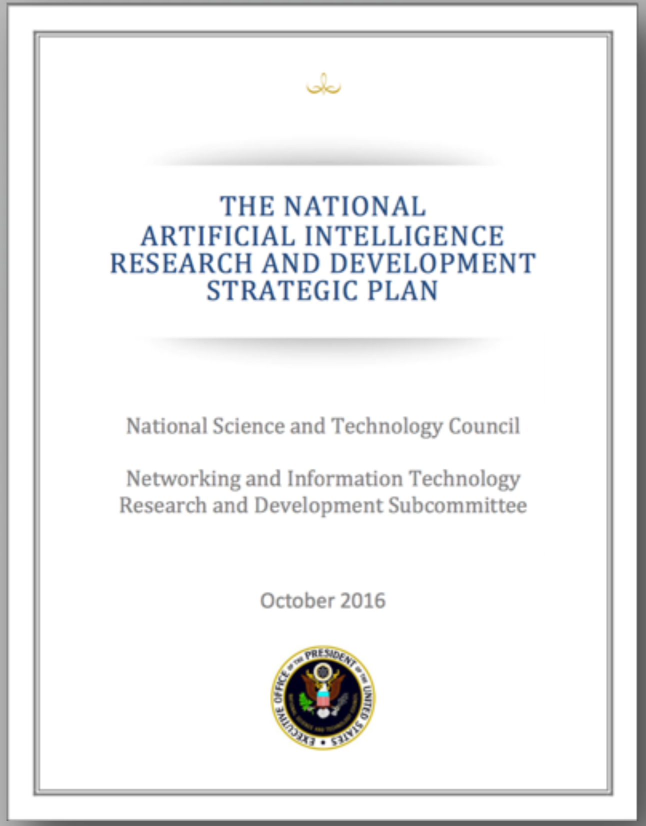
Eddy (1990) The Challenge. JAMA, 263(2):287-290. http://jama.jamanetwork.com/article.aspx?articleid=380215

# REGULATION FROM THE 1990s

- The standards that HealthCanada and the FDA used to assess software in diagnostic (Class I/Class II) devices don't make sense anymore.
- *As soon as the AI makes an observation, its behaviour can change*.

# STRATEGIES

- The Affordable Care Act shifted from a fee-for-service towards a pay-for-performance model[1]
  - Health IT is *rewarded*.
- Despite prohibitions in the Genetic Information Non-discrimination Act (2008), there is growing interest in using risk information for insurance stratification[2].
  - Differential pricing has become one of the standard practices for data analytics vendors, introducing new avenues to perpetuate inequality.
- **The (previous!) White House viewed AI as providing "increased medical efficacy, patient comfort, and less waste"[3].**




[1] David Blumenthal, Melinda Abrams, and Rachel Nuzum (2015) "The Affordable Care Act at 5 Years," *NEJM* **372**(25): 2453
[2] Yann Joly *et al* (2014) "Life Insurance: Genomic Stratification and Risk Classification," *European J of Human Genetics* **22**(5): 575–79.
[3] Bryan Biegel, & Kurose, J. F. (2016). *The National Artificial Intelligence Research and Development Strategic Plan*.

# H.R.6 – 114TH CONGRESS – 21ST CENTURY CURES ACT 1

- The 21st Century Cures Act passed House of Representatives (344-77) on 13 July 2015.
  - Received in the Senate, read twice, and referred to the Committee on Health, Education, Labor, and Pensions.

- **Guidance I**, "**general wellness products**": Include "audio recordings, video games, software programs and other products that are commonly … available from retail establishments."
  - The FDA will *not* regulate such products as medical devices, as long as they meet two factors, specifically they:
    i) are intended for only general wellness; and ii) present low risk to users.
  - These products' value derives from *information*, rather than doing something directly to the body.

John Graham, Artificial Intelligence, Machine Learning, And The FDA, 19 Aug 2016, Forbes

# CURRENTLY APPROVED

| Company | FDA Approval | Indication |
|---|---|---|
| Apple | September 2018 | Atrial fibrillation detection |
| Aidoc | August 2018 | CT brain bleed diagnosis |
| iCAD | August 2018 | Breast density via mammography |
| Zebra Medical | July 2018 | Coronary calcium scoring |
| Bay Labs | June 2018 | Echocardiogram EF determination |
| Neural Analytics | May 2018 | Device for paramedic stroke diagnosis |
| IDx | April 2018 | Diabetic retinopathy diagnosis |
| Icometrix | April 2018 | MRI brain interpretation |
| Imagen | March 2018 | X-ray wrist fracture diagnosis |
| Viz.ai | February 2018 | CT stroke diagnosis |
| Arterys | February 2018 | Liver and lung cancer (MRI, CT) diagnosis |
| MaxQ-AI | January 2018 | CT brain bleed diagnosis |
| Alivecor | November 2017 | Atrial fibrillation detection via Apple Watch |
| Arterys | January 2017 | MRI heart interpretation |

U.S. FOOD & DRUG
ADMINISTRATION

September 11, 2018

subject to the following special controls:

1. Clinical [testing] under anticipated conditions of use must demonstrate…:
    1. The ability to obtain an ECG of sufficient quality for display and analysis; and
    2. **The performance characteristics of the detection algorithm as reported by sensitivity and either specificity or positive predictive value**.
2. **Software verification, validation, and hazard analysis must be performed**. Documentation must include a characterization of the technical specifications of the software, including the **detection algorithm and its inputs and outputs**.
3. Non-clinical performance testing must **validate** detection algorithm performance **using a previously adjudicated data set**.
4. Human factors and usability testing must demonstrate the following:
    1. The user can correctly use the device based solely on reading the device labeling; and
    2. The user can correctly **interpret the device output and understand when to seek medical care**.
5. …

FDA identifies this generic type of device as:

U.S. Food & Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993
www.fda.gov

# THE QUANTIFIED SELF VS THE MEDICAL RECORD

- Many apps serve to **shift** the **responsibility** for care and monitoring from healthcare professionals to patients themselves.
  - This may disadvantage patients who do not have the time, resources, or access to technology.
  - **What kinds of patients are favored in this new dynamic**, and might patients not well-equipped to manage and maintain their own data receive substandard care?
  - What new roles and responsibilities do the *developers* of such apps take on, and how do the ethical responsibilities of medical professionals get integrated into these differing contexts?.

- How to combine *models* in different AIs? There's no EDI in HIPAA for *models*.

VS



Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., & Ferryman, K. (2016). *The AI Now Report*.

# Medical-record software companies are selling your health data

By Sheryl Spithoff Special to the Star
*Wed., Feb. 20, 2019*

STAR INVESTIGATION

There's a booming business in patient medical records and up to five million Ontarians are part of that boom, whether they know it or not.

Perspectives

# Privacy versus artificial intelligence in medicine

Taryn J Rohringer (BMSc)[1]; Akshay Budhkar (BASc)[2,5]; Frank Rudzicz (PhD)[2,3,4,5]

[1]Faculty of Medicine, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, ON, Canada, M5S 1A8.

The same month that GDPR came into effect, Canada issued new guidance for the Personal Information Protection and Electronic Documents Act (PIPEDA) … subsection 5(3) of PIPEDA states that "**An organization may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances**." Given that consensus has not been widely achieved with regards to the details of surveillance of this type (e.g., what risks to personal information are necessary, given the technology, to achieve some perceived benefit to the person involved), **it is not yet clear what a "reasonable person would consider appropriate**."

companies overstepping their bounds in the pursuit of patient data to train their systems, and new regulations around privacy of those data, this discussion is especially pertinent. Here, we suggest that a common good can be achieved in which data can be kept private while also useful for artificial intelligence in the practice of medicine.

## Introduction

Recent advances in artificial intelligence (AI) have accelerated their use in healthcare, from remote monitoring and wearables to clinical decision support.[1]

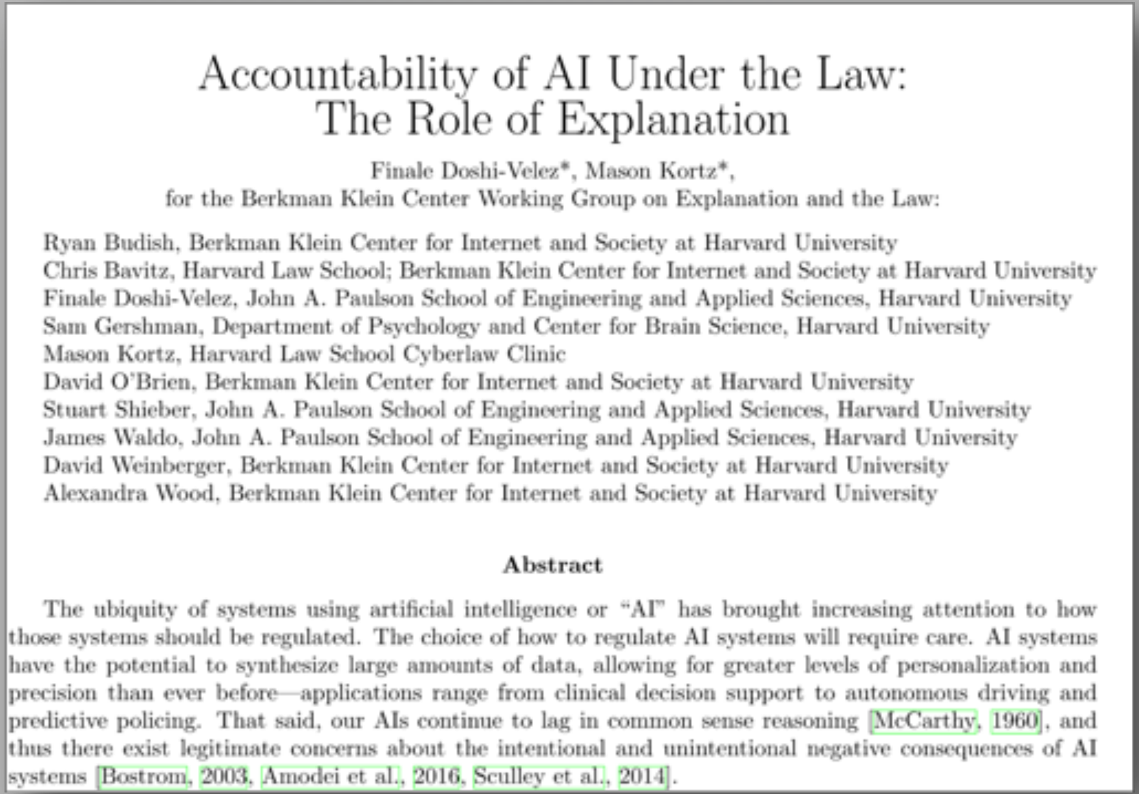
may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances." Given that consensus has not been widely achieved with regards to the details of surveillance of this type (e.g., what risks to personal information are necessary, given the technology, to achieve some perceived benefit to the person involved), it is not yet clear what a "reasonable person would consider appropriate."

As AI is increasingly integrated into clinical practice, various challenges will persist (e.g. data acquisition, reporting, and re-identification) and these emphasize a potential struggle between patient privacy and the promise of these systems.

## Challenges to Data Acquisition

Personal health data is extremely valuable; for example, the $6 billion acquisition of Medco Containment Services by Merck was

# PRINCIPLES AND SOCIETAL NORMS



Accountability of AI Under the Law:
The Role of Explanation

Finale Doshi-Velez*, Mason Kortz*,
for the Berkman Klein Center Working Group on Explanation and the Law:

Ryan Budish, Berkman Klein Center for Internet and Society at Harvard University
Chris Bavitz, Harvard Law School; Berkman Klein Center for Internet and Society at Harvard University
Finale Doshi-Velez, John A. Paulson School of Engineering and Applied Sciences, Harvard University
Sam Gershman, Department of Psychology and Center for Brain Science, Harvard University
Mason Kortz, Harvard Law School Cyberlaw Clinic
David O'Brien, Berkman Klein Center for Internet and Society at Harvard University
Stuart Shieber, John A. Paulson School of Engineering and Applied Sciences, Harvard University
James Waldo, John A. Paulson School of Engineering and Applied Sciences, Harvard University
David Weinberger, Berkman Klein Center for Internet and Society at Harvard University
Alexandra Wood, Berkman Klein Center for Internet and Society at Harvard University

Abstract

The ubiquity of systems using artificial intelligence or "AI" has brought increasing attention to how those systems should be regulated. The choice of how to regulate AI systems will require care. AI systems have the potential to synthesize large amounts of data, allowing for greater levels of personalization and precision than ever before—applications range from clinical decision support to autonomous driving and predictive policing. That said, our AIs continue to lag in common sense reasoning [McCarthy, 1960], and thus there exist legitimate concerns about the intentional and unintentional negative consequences of AI systems [Bostrom, 2003, Amodei et al., 2016, Sculley et al., 2014].

- When do we expect an explanation?
  - **Impact**. Does the action affect a 3rd party?
  - **Value**. Can something be done if we know the action was erroneous?
  - **Error**. Do we expect error?
    - **Unreliable inputs**
    - **Inexplicable outcomes**
    - **Distrust** in **system integrity**
- A few precedents are listed in US law.
  - Strict liability, divorce, discrimination

Doshi-Velez F, Kortz M, Budish R, *et al.* Accountability of AI Under the Law: The Role of Explanation. 2017;:1–15. doi:10.2139/ssrn.3064761

# LAW AND EXPLANATIONS

- EU General Data Protection Regulation (enacted 2016), extends the automated decision-making rights in the **1995 Data Protection Directive** to provide a right to an explanation, in Recital 71:

> **The data subject should have the right** not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based **solely** on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.
> ...
> [S]uch processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view**, to obtain an explanation of the decision** reached after such assessment and to challenge the decision.

- Note: recitals are *not* binding (indeed, explainability was removed from the binding Article during the legislative process.
- ***Solely?!***

# THE WAY FORWARD

# THE DAWN OF AI STANDARDS

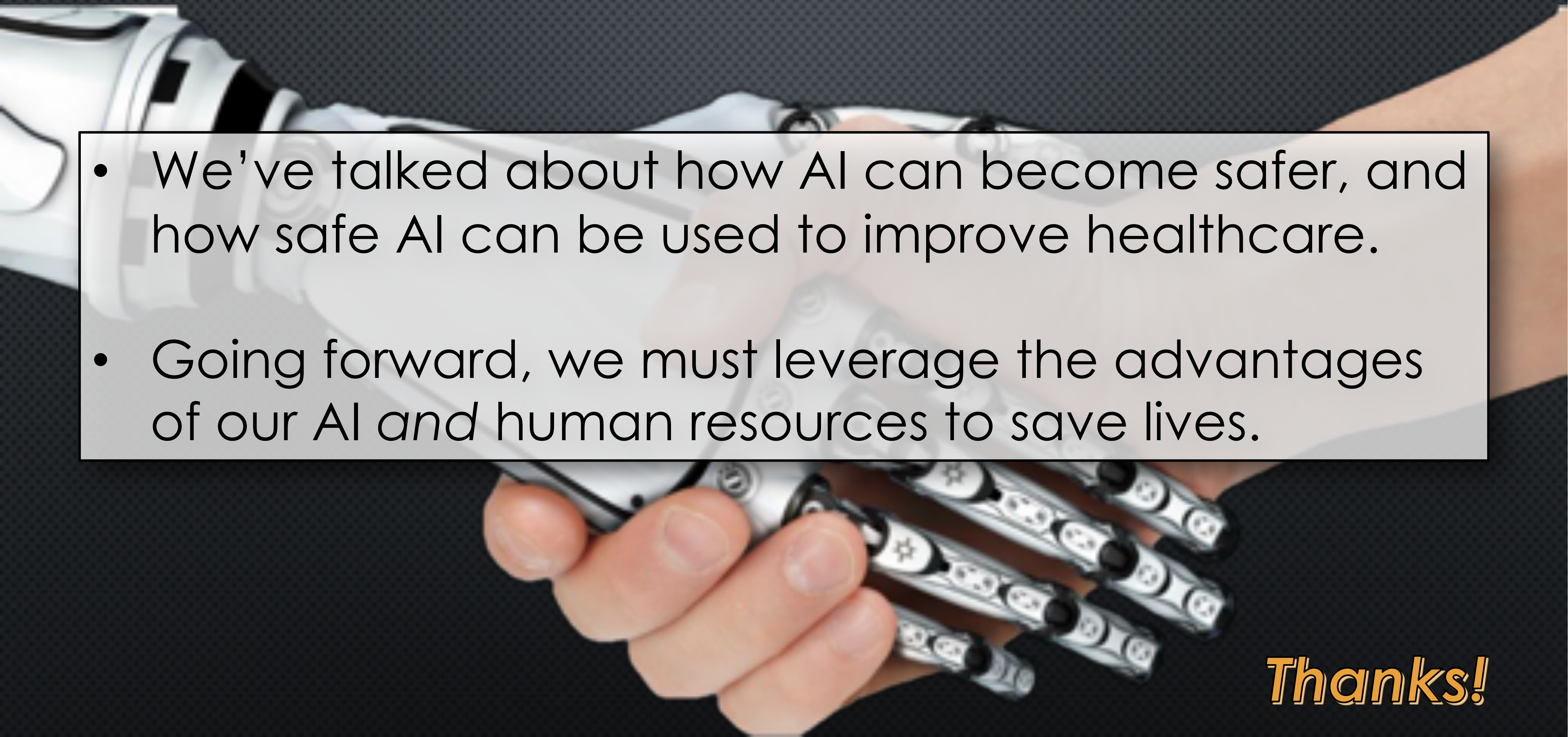
- Three study groups were formed within ISO/JTC1 SC 42 in 2018:
  - **Computational approaches and characteristics** includes specialized AI systems (e.g., NLP or computer vision), their underlying computational approaches, architectures, and characteristics.
  - **Trustworthiness** concerns approaches to establish trust in AI systems, e.g., through **transparency**, **verifiability**, **explainability**, **controllability**. Typical threats and risks, their mitigation techniques, and approaches to robustness, accuracy, privacy, and safety will also be investigated.
  - **Use cases and applications** focuses on application domains for AI (e.g., social networks and embedded systems) and the different context of their use (e.g., health care, smart homes, autonomous cars).

# STANDARDS FOR EVALUATING ML MODELS

- When comparing the performance of two or more models, several aspects must be carefully controlled and reported:
  - **Implementation** E.g., if an algorithm can be accelerated in such a way that can affect outcomes, then this must be made explicit.
  - **Hyper-parameter optimization** should not favor one model over another.
  - **Preprocessing** will not unjustly favour one model over another. E.g., removing outliers, incomplete data, or noise should not unfairly affect performance.
  - **Training and testing data** should be ecologically valid, statistically indistinct, or otherwise similar to data expected to be observed in deployment.
  - **Appropriate baselines** Any classifier should be compared against ≥1 representative, appropriate baseline. Trivial baselines should not be considered.
  - **Limiting channel effects** incl. characteristics of the manner in which data were recorded, in addition to the nature of the data themselves. Some strategies explicitly factor out channel effects.

- Appropriate statistical tests of significance must be undertaken, when possible.

# TRANSPARENCY, TRUSTWORTHINESS, EFFECTIVENESS



- We've talked about how AI can become safer, and how safe AI can be used to improve healthcare.
- Going forward, we must leverage the advantages of our AI *and* human resources to save lives.

***Thanks!***