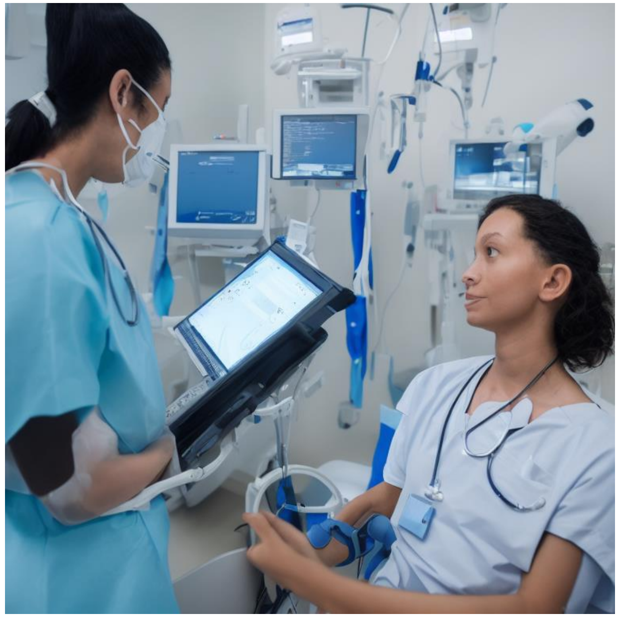
Human-AI Collaboration in Healthcare

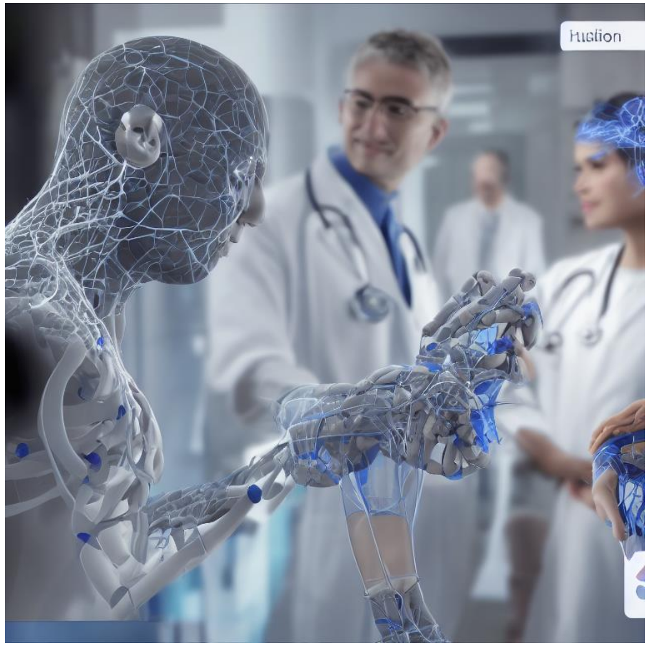


"clinician interacting with machine learning model in hospital 4k"

Hussein Mozannar

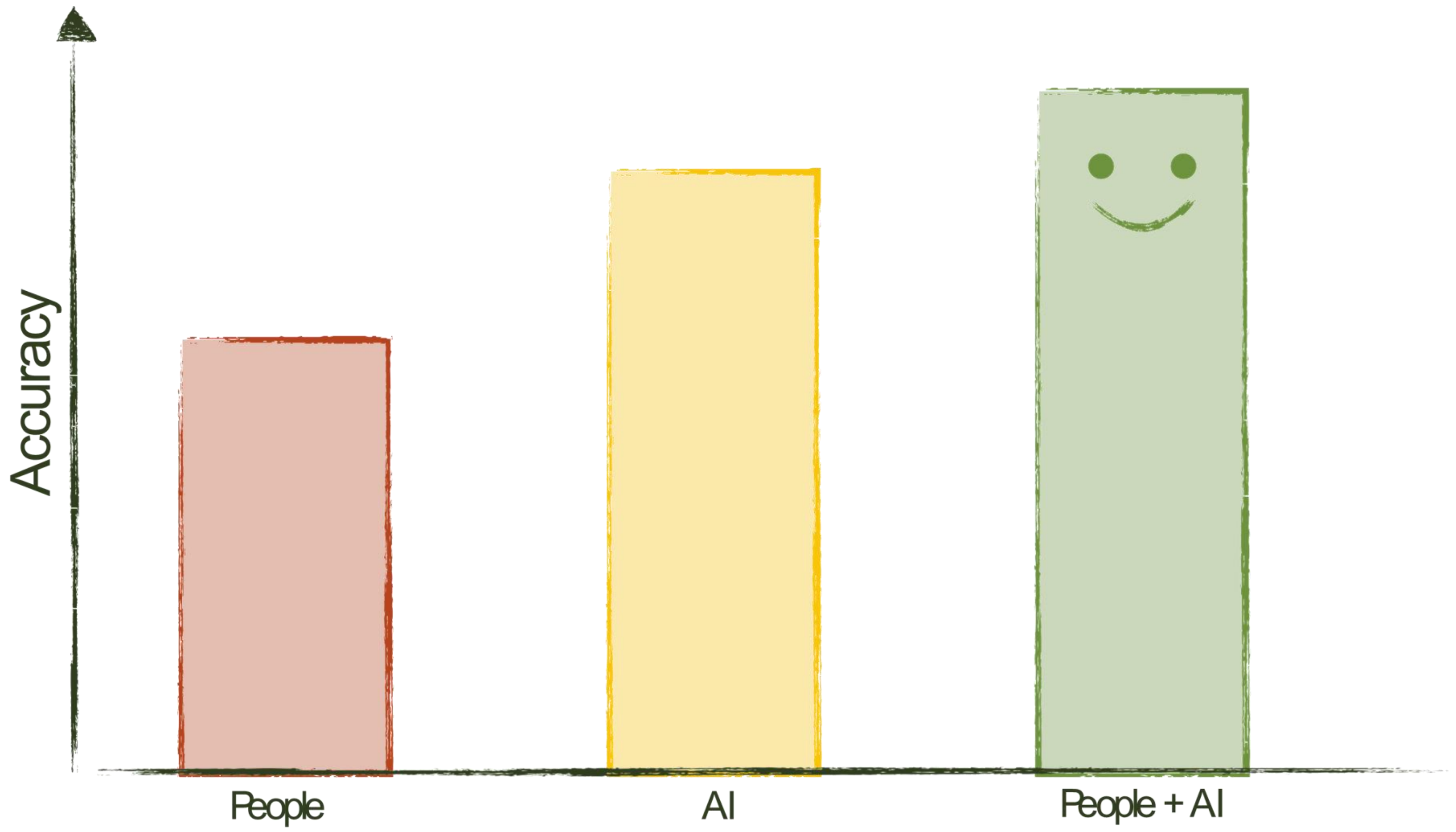
6.793/HST.956
March 16, 2023

https://replicate.com/stability-ai/stable-diffusion



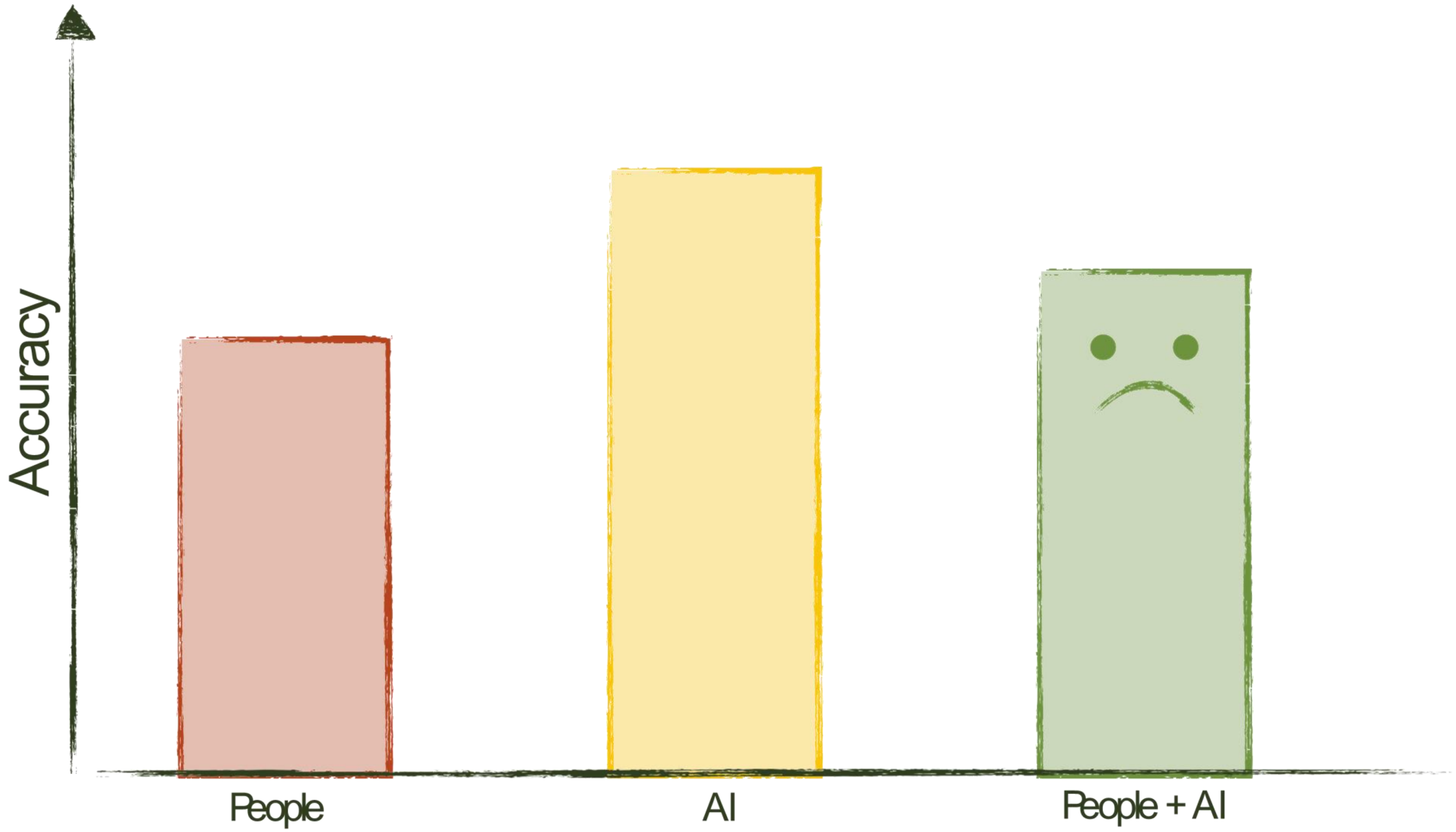
human-ai collaboration in healthcare realistic HD"

Hope of AI-Assisted Decision Making



From Krzysztof Gajos https://www.dropbox.com/s/ht2cmjulebk9lev/2022.10.28%20-%20Chicago%20-%20Human%20Cognitive%20(Dis)Engagement%20during%20AI-Assisted%20Decision-Making.pdf?dl=0

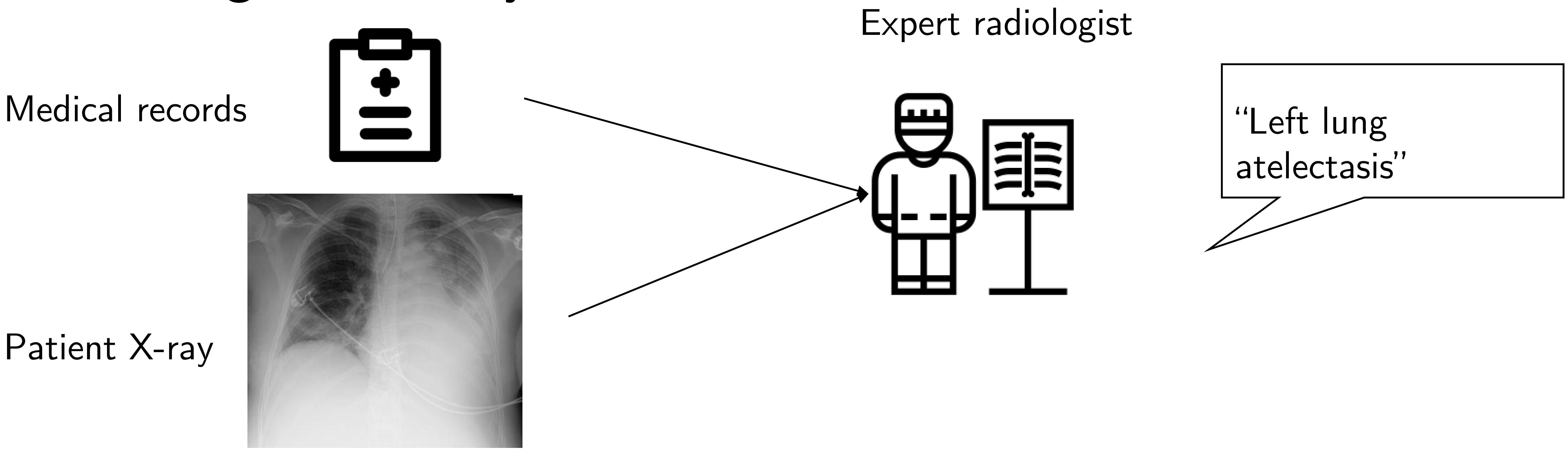
Reality of AI-Assisted Decision Making



From Krzysztof Gajos https://www.dropbox.com/s/ht2cmjulebk9lev/2022.10.28%20-%20Chicago%20-%20Human%20Cognitive%20(Dis)Engagement%20during%20AI-Assisted%20Decision-Making.pdf?dl=0

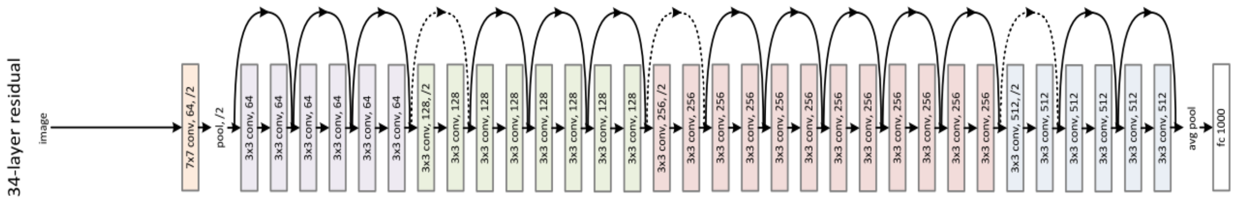
Detecting Atelectasis From Chest X-rays

- Atelectasis: the collapse of part or all of a lung.
- Can be caused by mucus, foreign objects or tumors blocking the airway.



Detecting Atelectasis From Chest X-rays

- A student from class decided to build an ML model for detecting Atelectasis instead.
- They use CheXpert [1] dataset of >200k chest x-rays with annotations
- They train a ResNet-34 model [2]



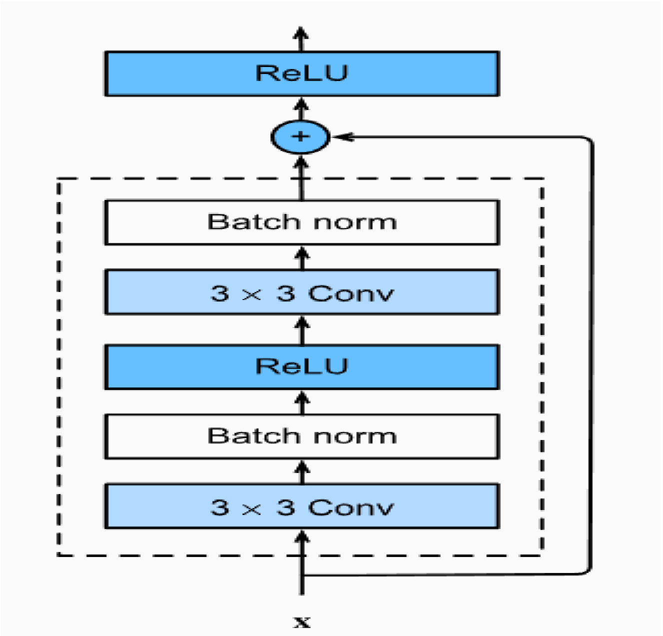
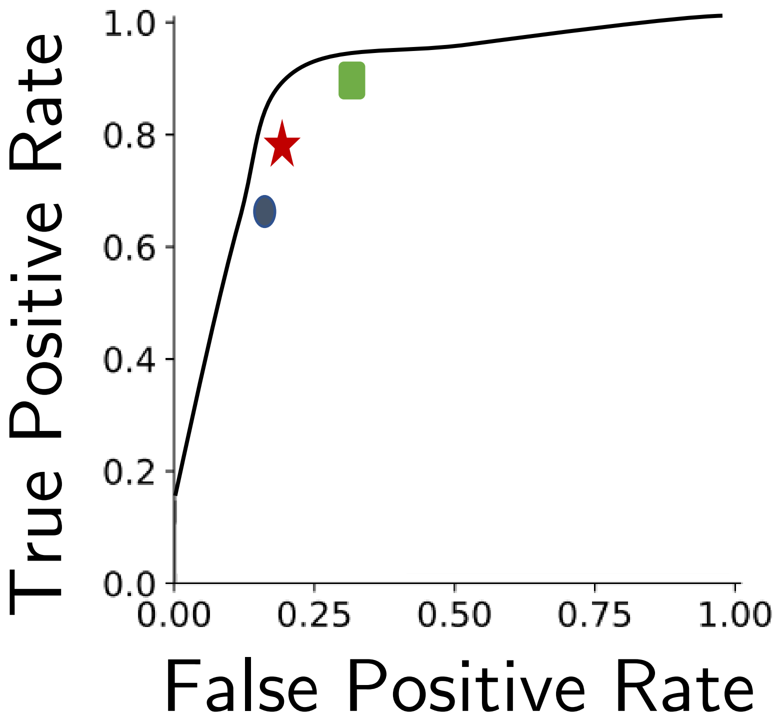


Figure 2. Residual learning: a building block.

[1]: Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. 2019. [2]: He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

AI vs Human performance

- Test set: 500 x-rays annotated each by 5 radiologists, ground truth is their majority vote. 3 other radiologists to compare to.



– Model (AUC = 0.91)
★ Rad1 (0.21,0.80)
● Rad2 (0.18,0.71)
■ Rad3 (0.31,0.92)

Model outperforms all 3 radiologists

How do we integrate the AI into the current pipeline?

Outline

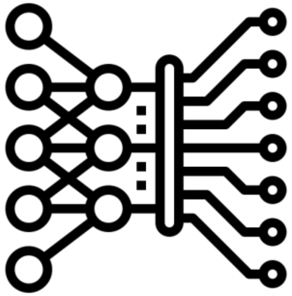
- **Modes of Human-AI Interaction**
- Mental Models
- Onboarding
- Over-reliance on AI and fixes

Deploying the AI to replace the radiologist

- **Model in isolation:** after X-ray is taken, the model makes its prediction, then referring physician can give treatment

Patient X-ray

Model



Radiology Report

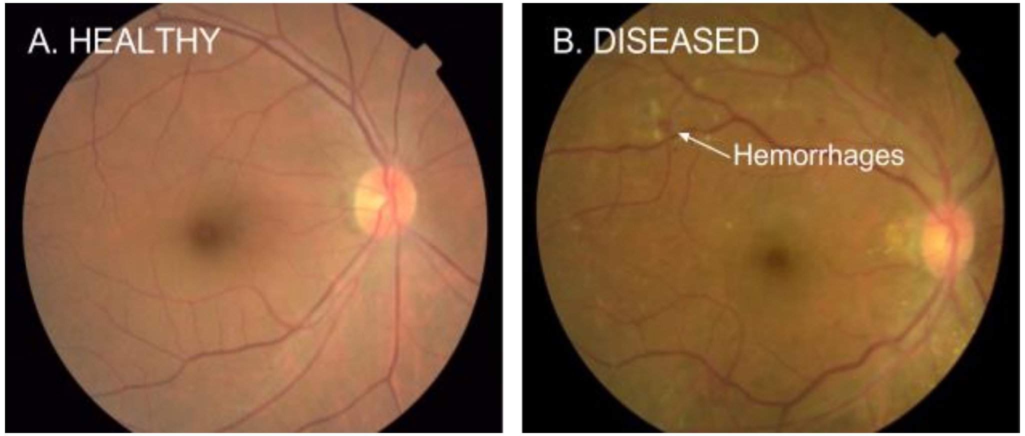
Heart size upper normal but stable. Mediastinal contours within normal limits. Minimal right middle lobe atelectasis. No focal airspace consolidation, pleural effusion, or pneumothorax. Degenerative endplate changes of the spine. [1]

Physician

[1]: Buendía, Félix, Joaquín Gayoso-Cabada, and José-Luis Sierra. "An Annotation Approach for Radiology Reports Linking Clinical Text and Medical Images with Instructional Purposes." Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality. 2020.

Model in isolation: Diabetic Retinopathy

- **Diabetic Retinopathy:** diabetes complication affecting the eye
- **Why we need AI:** access to care is a huge problem, especially in places like India (70mil diabetics, 2 months to get results, need to travel)



- **Model:** Dataset from Thailand, model reduces FNR by 23% but increases FPR by 2% [1]

[1]: Ruamviboonsuk, Paisan, et al. "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program." *NPJ digital medicine* 2.1 (2019): 1-9.

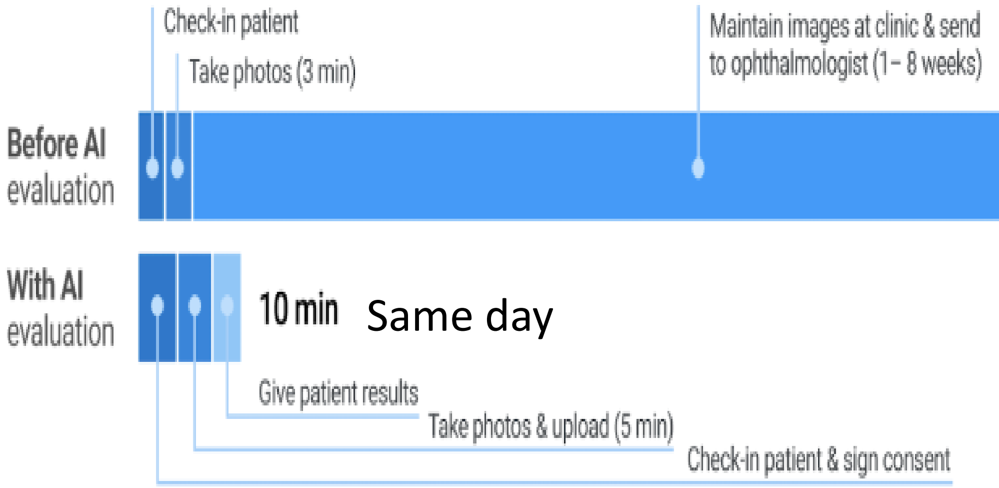
Deployment details

- Model deployed in 8 sites in Thailand, 1.5-year study, 7600 patients
- 200 patients/day, 5 hours wait, 90sec eye exam

[1]:Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Deployment details

- Prospective study after deployment with the nurses taking the images [1]



Same day

[1]:Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Results after deployment

- Model refused to predict on 20% of images, images were unreadable to the model
  - Imperfect lighting conditions
  - Old cameras
  - Limited time to align patients
- Nurse's observations:

"Some images are blurry, and I can still read it, but the system can't", "it's good but I think it's not as accurate. If [the eye] is a little obscured, it can't grade it"

- **Those ungraded, now needed to travel to see an ophthalmologist instead of just waiting for image to be read.**

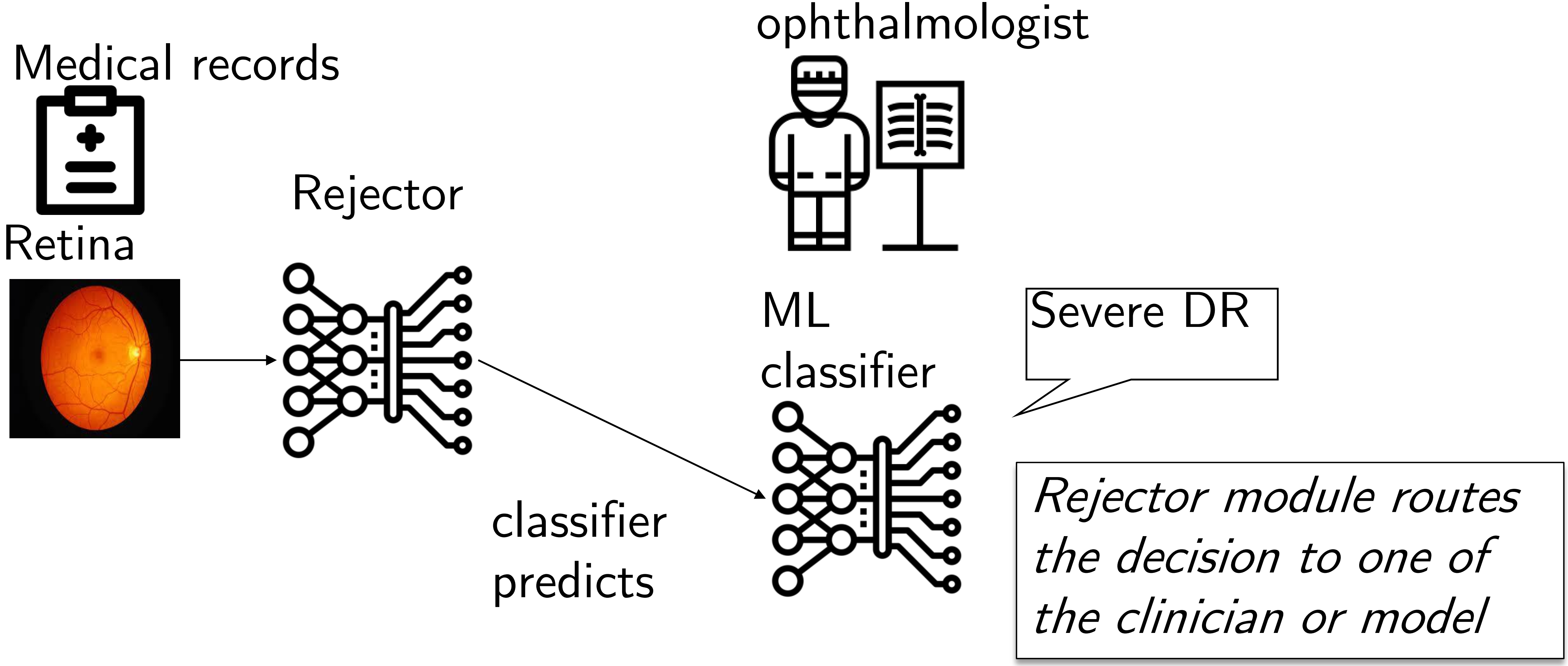
[1]:Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Takeaways from deployment

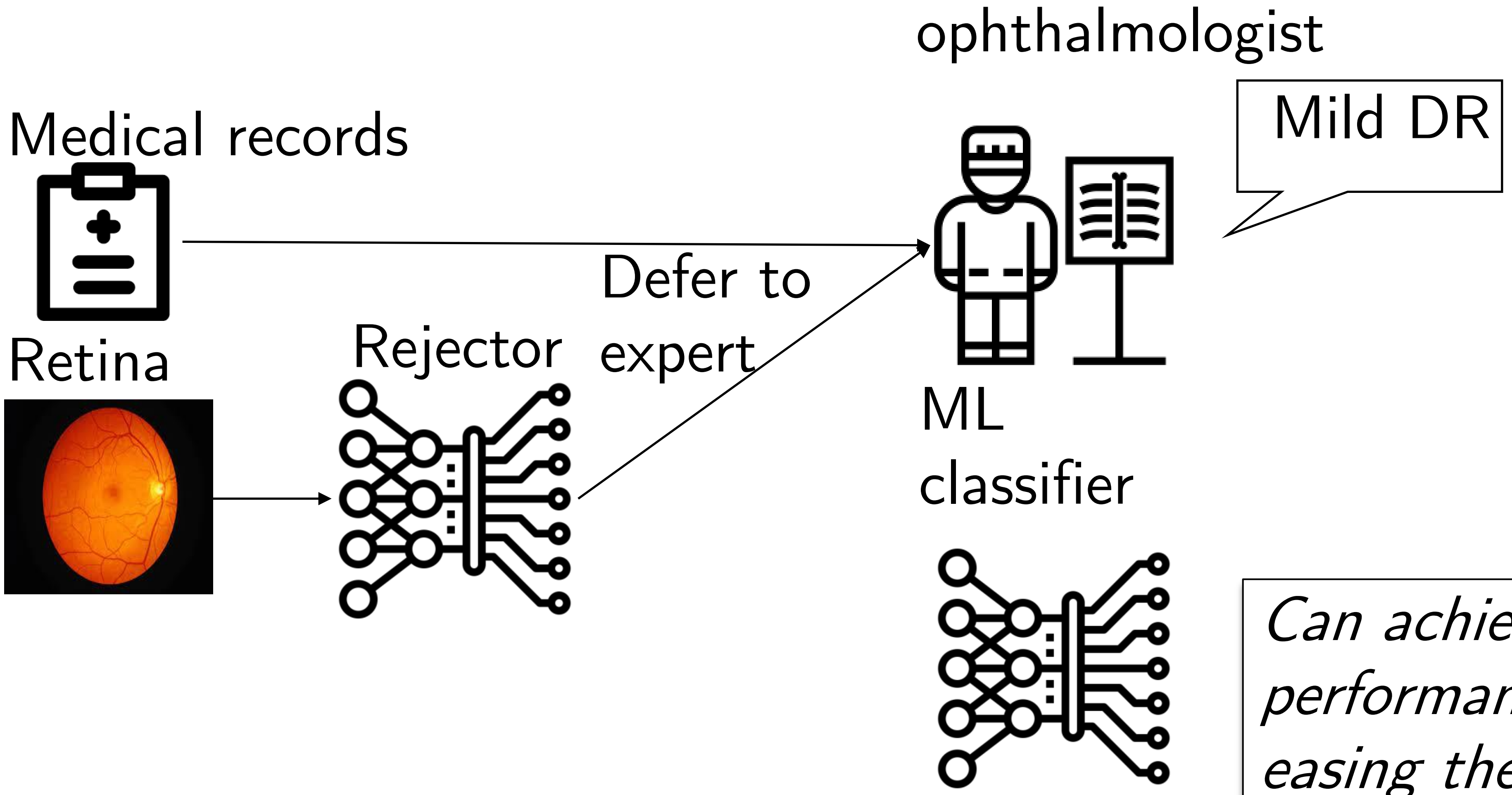
1. Protocols around use of model are crucial to its success
2. Human centered evaluation is crucial to be able to understand issues required for effective deployment

- Eliminating the ophthalmologists from the system removes safety checks against model failure (e.g., distribution shift) and input issues
- Can do better by combining model and ophthalmologists then each alone!

Model + Human: Algorithmic Triage



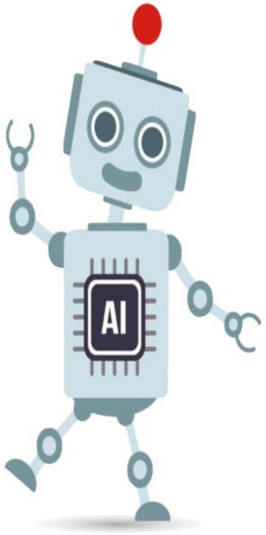
Algorithmic Triage

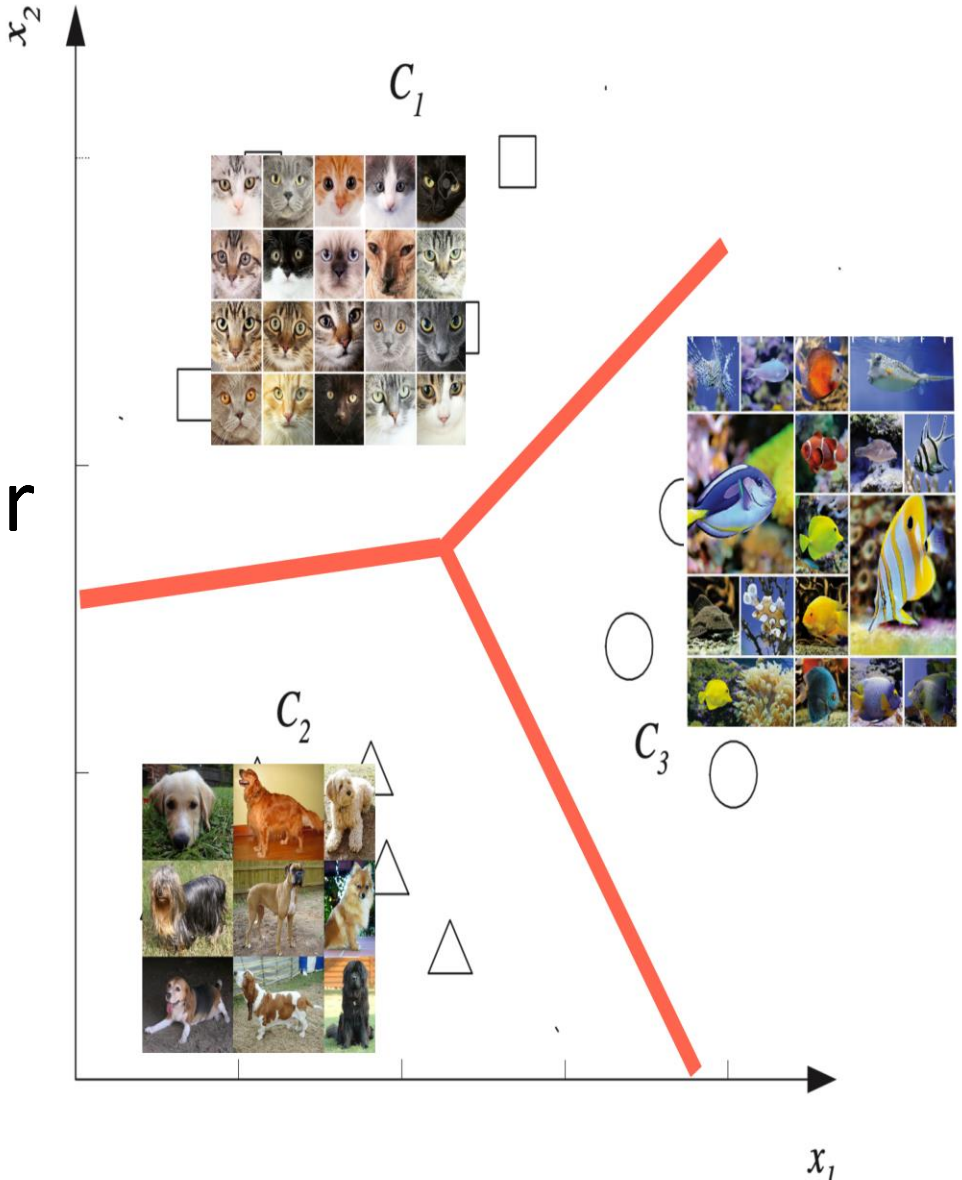


Can achieve better performance while still easing the burden on the ophthalmologist

Goal: Given input **x**, predict membership in one **K** classes

Classifier

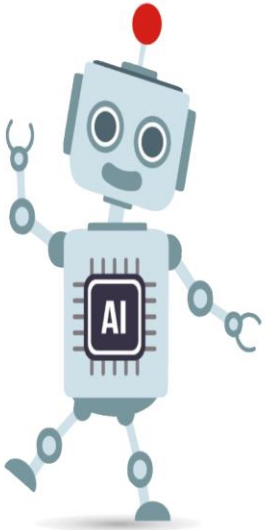


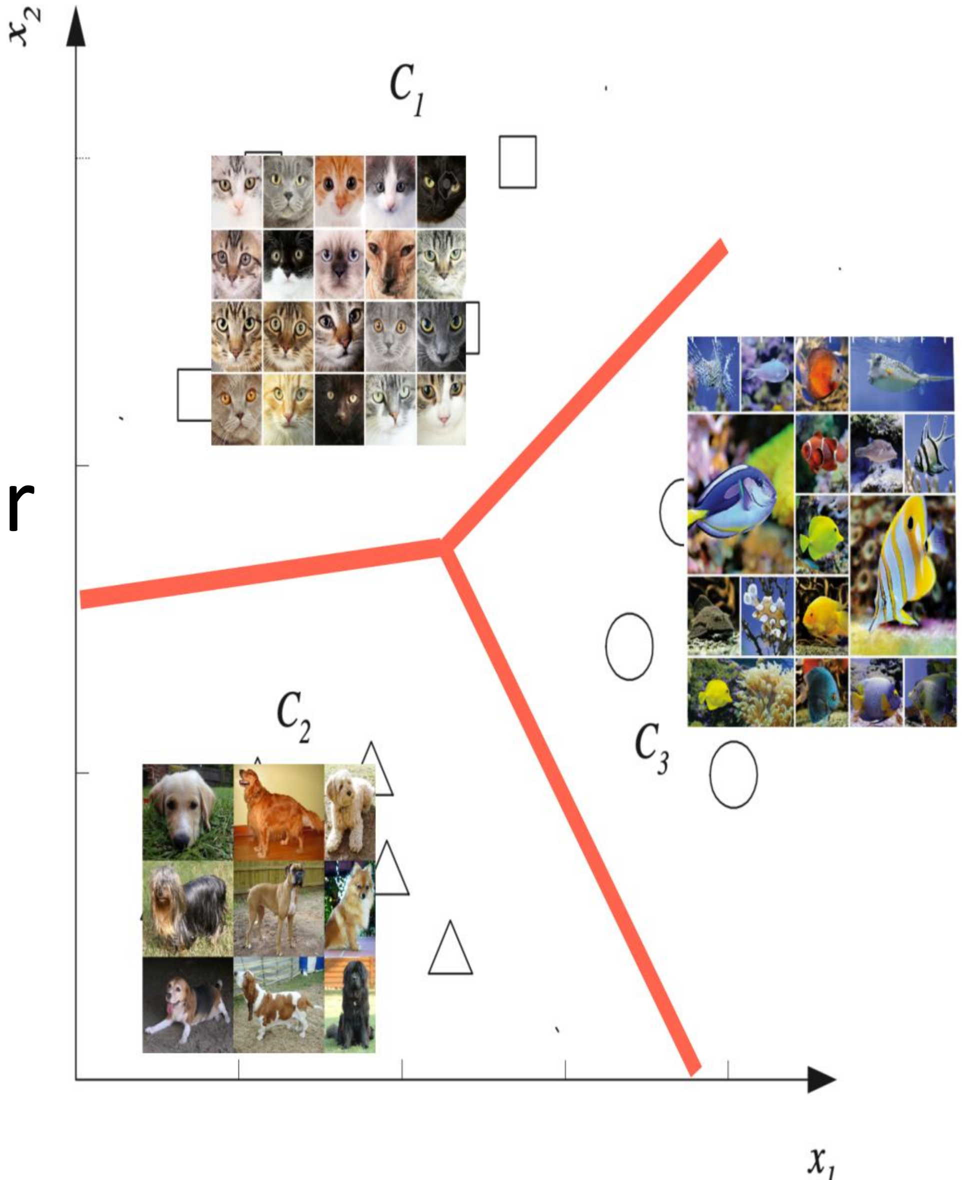


*Slides adapted from Eric Nalisnick
https://enalisnick.github.io/Calibrated_L2D_talk.pdf

Goal: Given input **x**, predict membership in one **K** classes

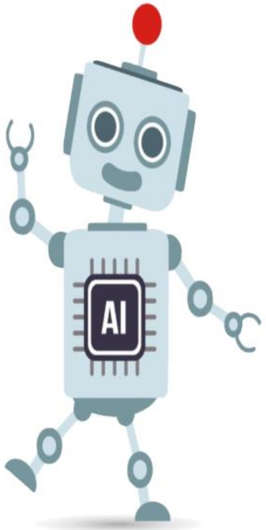
Classifier

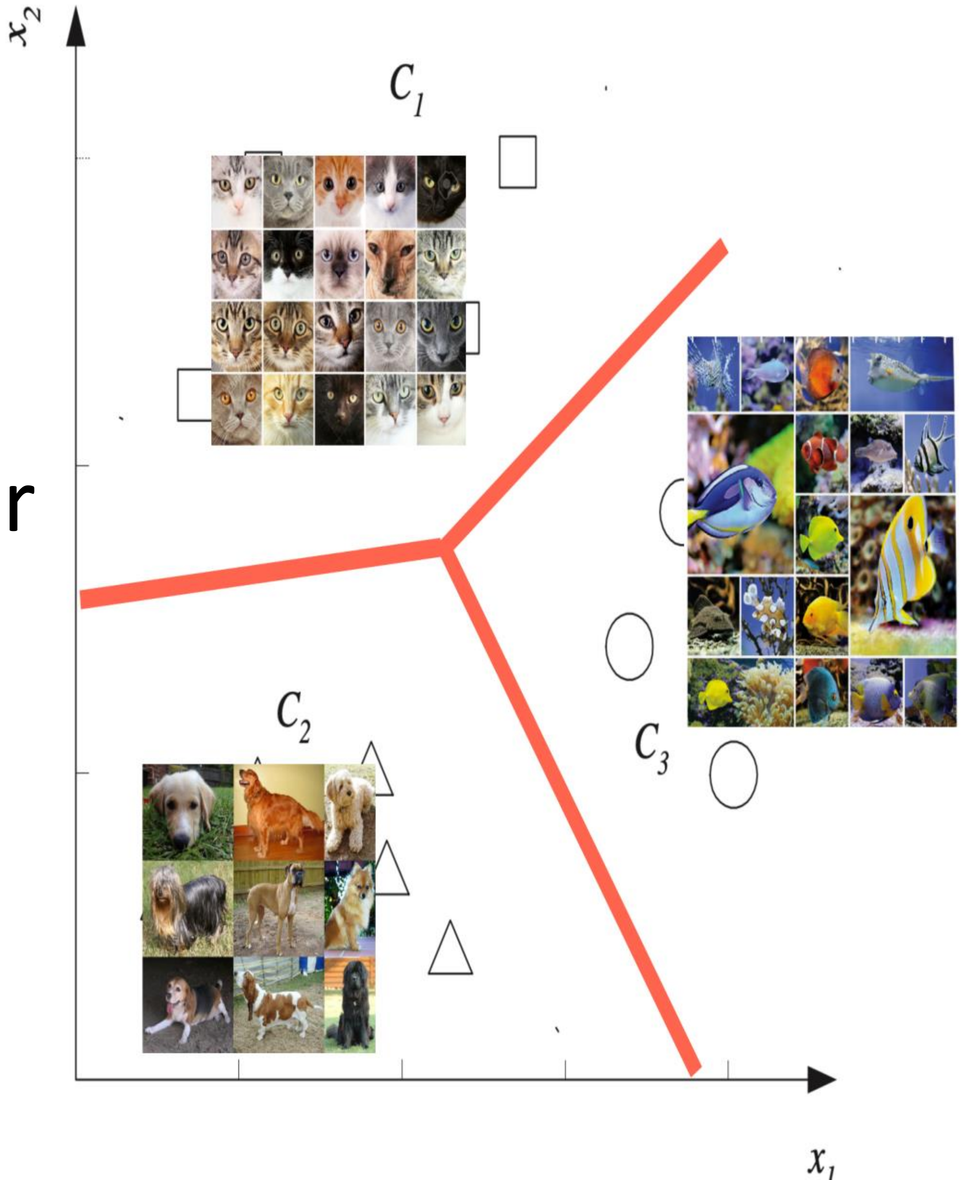


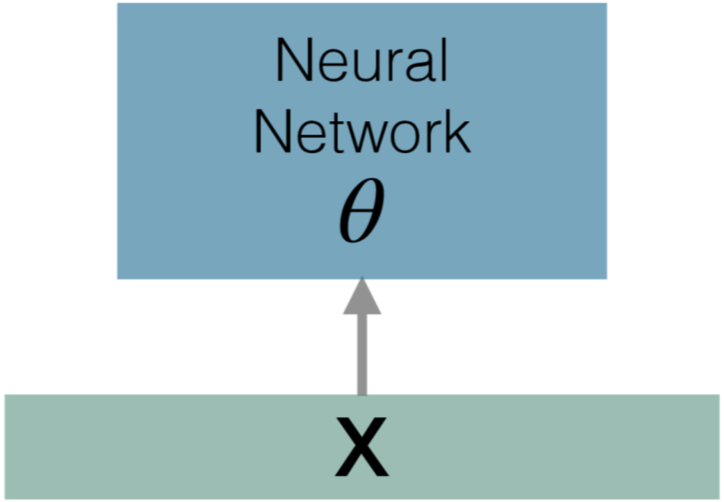


Goal: Given input **x**, predict membership in one **K** classes

Classifier

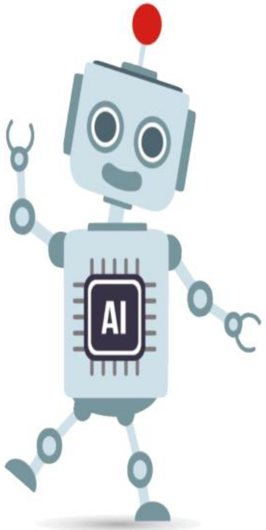


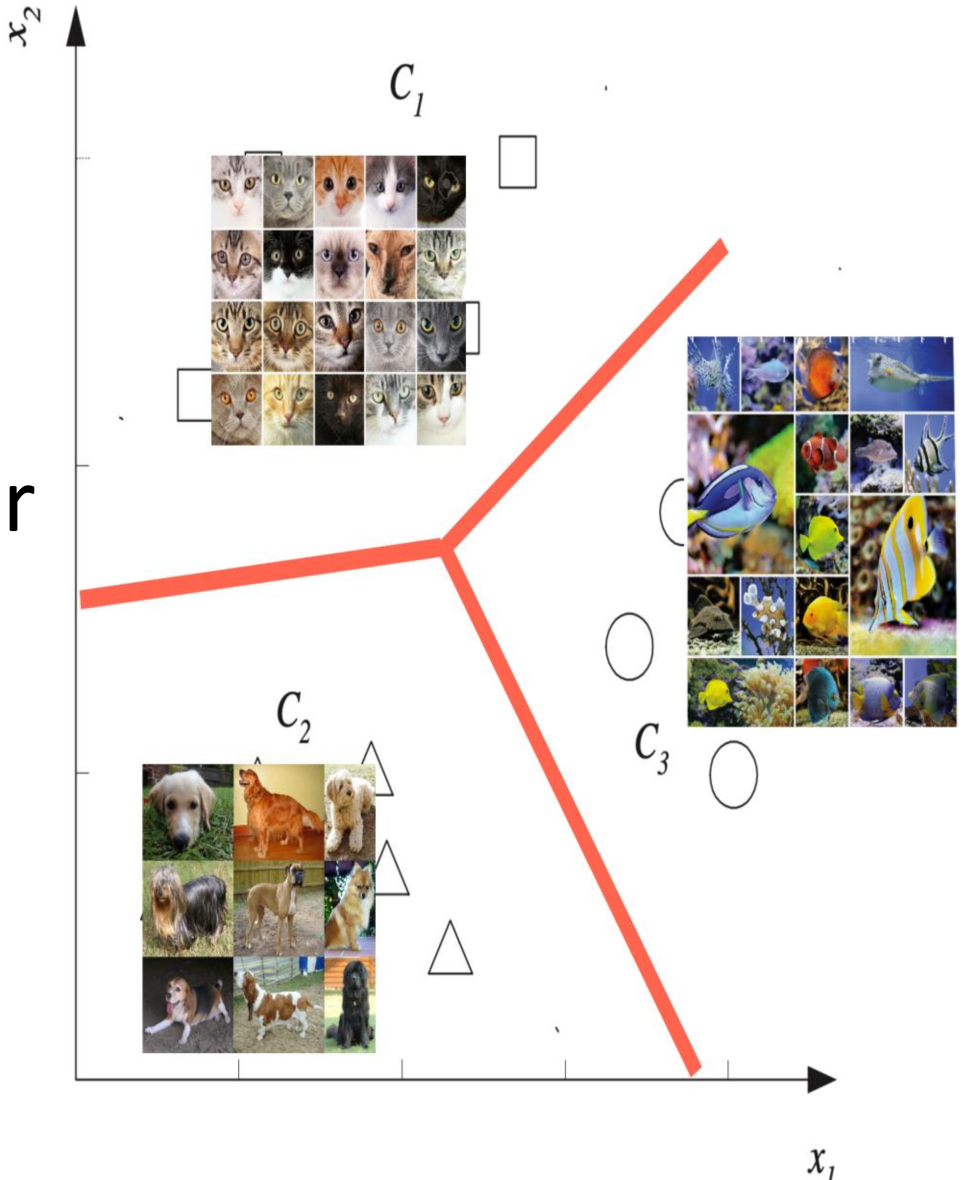




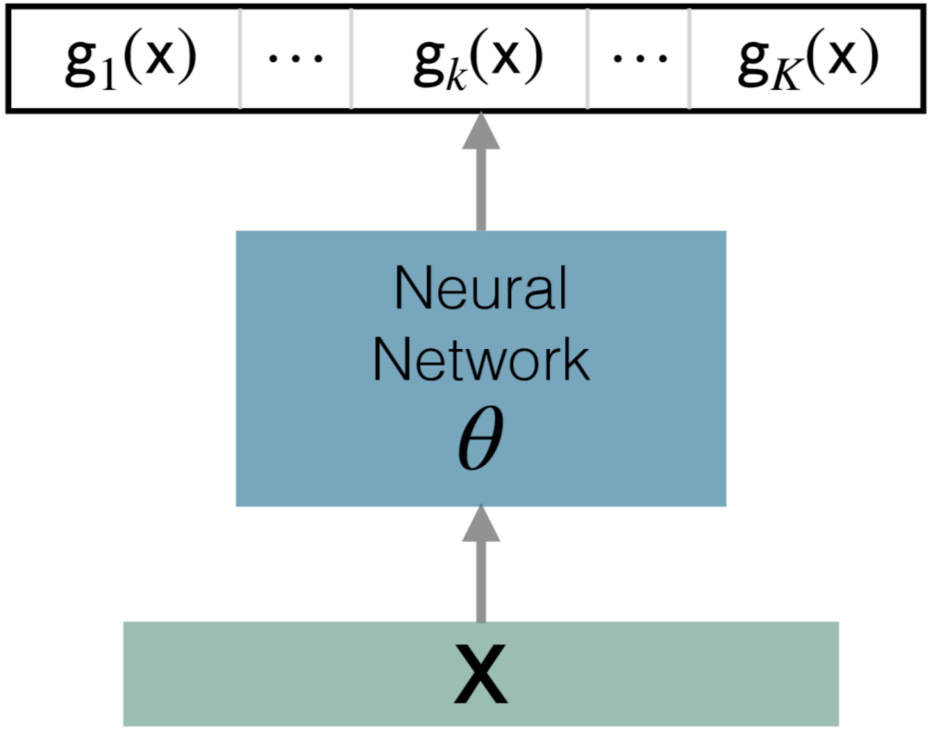
Goal: Given input **x**, predict membership in one **K** classes

Classifier



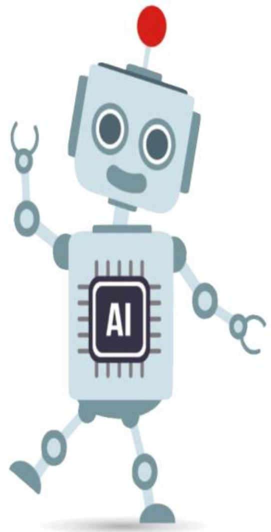


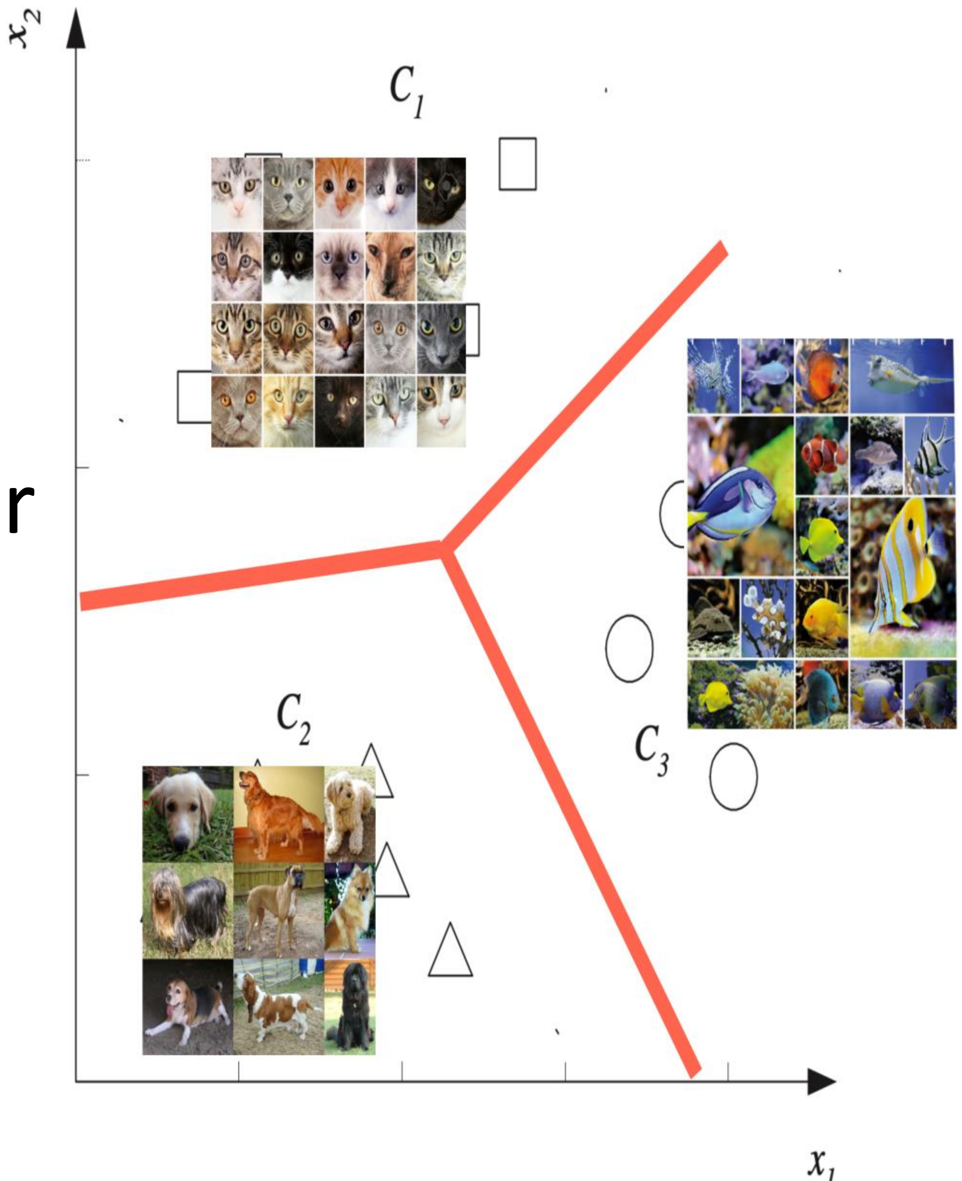
$$g_k(\mathbf{x}) \in \mathbb{R}$$



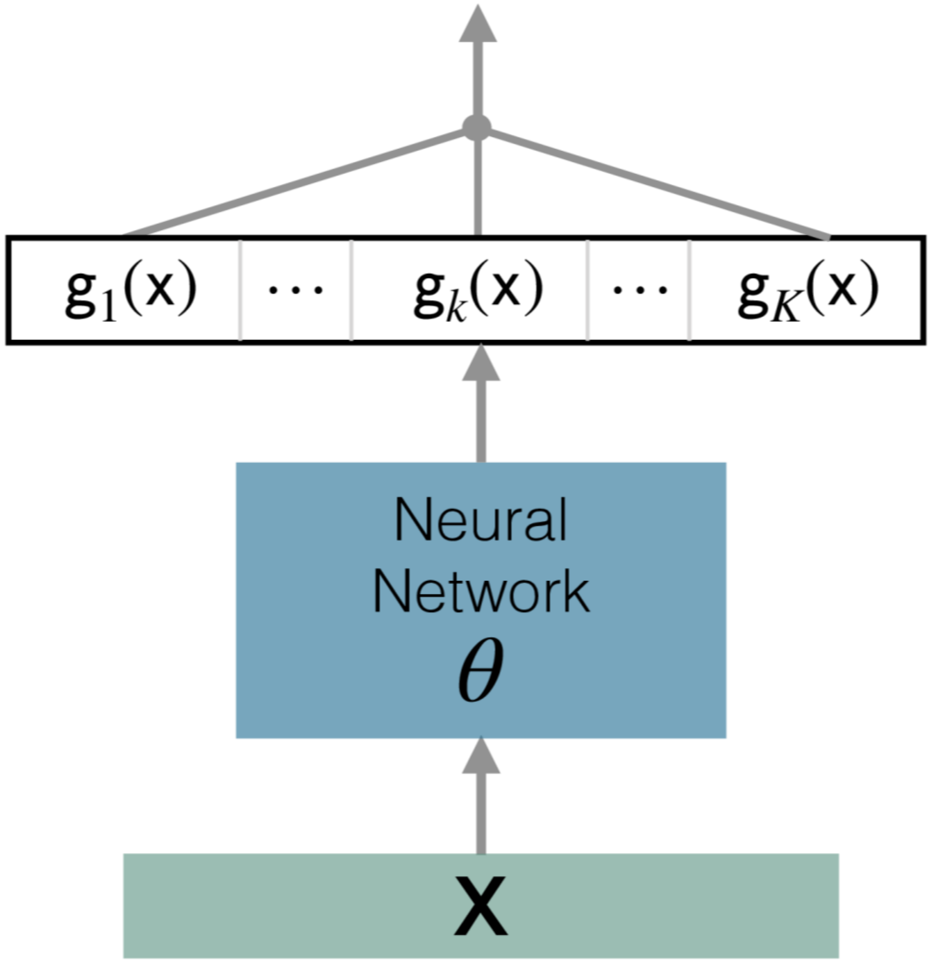
Goal: Given input **x**, predict membership in one **K** classes

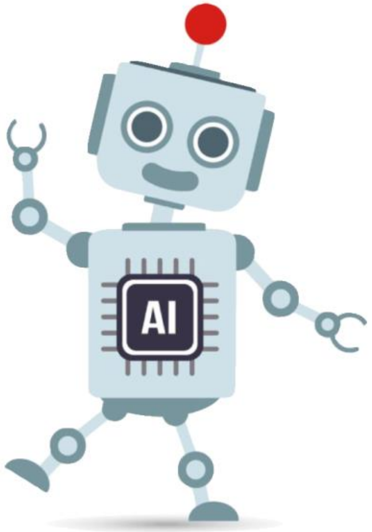
Classifier





$$P(y \mid x) = \frac{\exp\{g_y(x)\}}{\sum_{k=1}^{K} \exp\{g_k(x)\}}$$





?

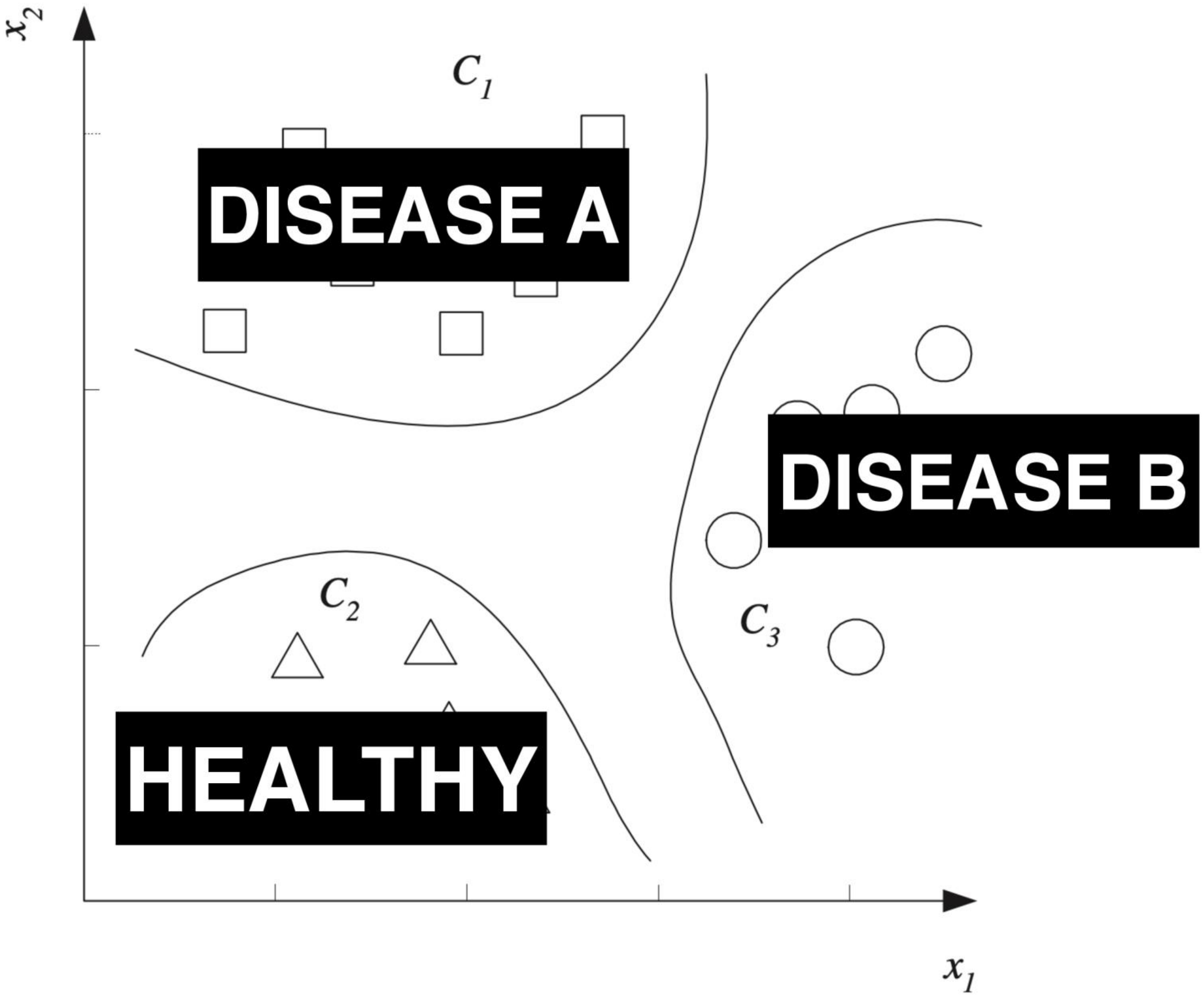
+

Classifier

Human Expert

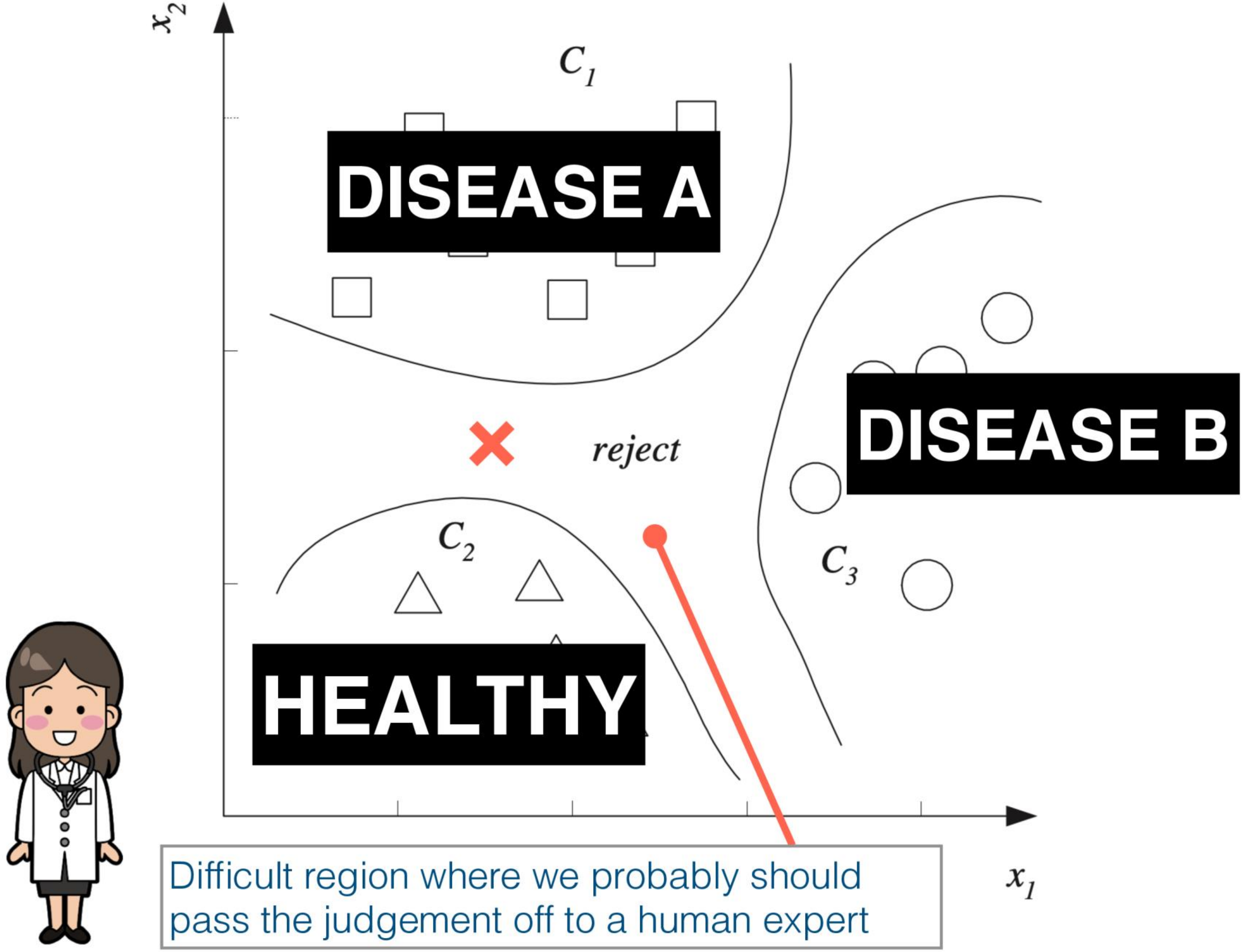
Warm Up

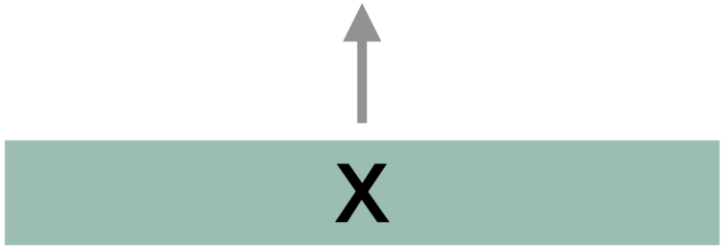
Classification with a Rejection Option

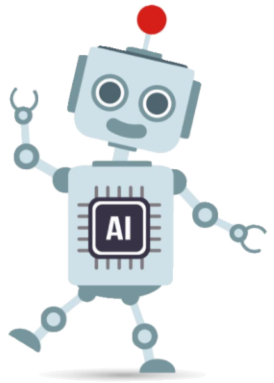


x_2
C_1
DISEASE A
DISEASE B
C_2
C_3
HEALTHY
x_1

x_2
C_1
DISEASE A
DISEASE B
C_2
C_3
HEALTHY
x_1

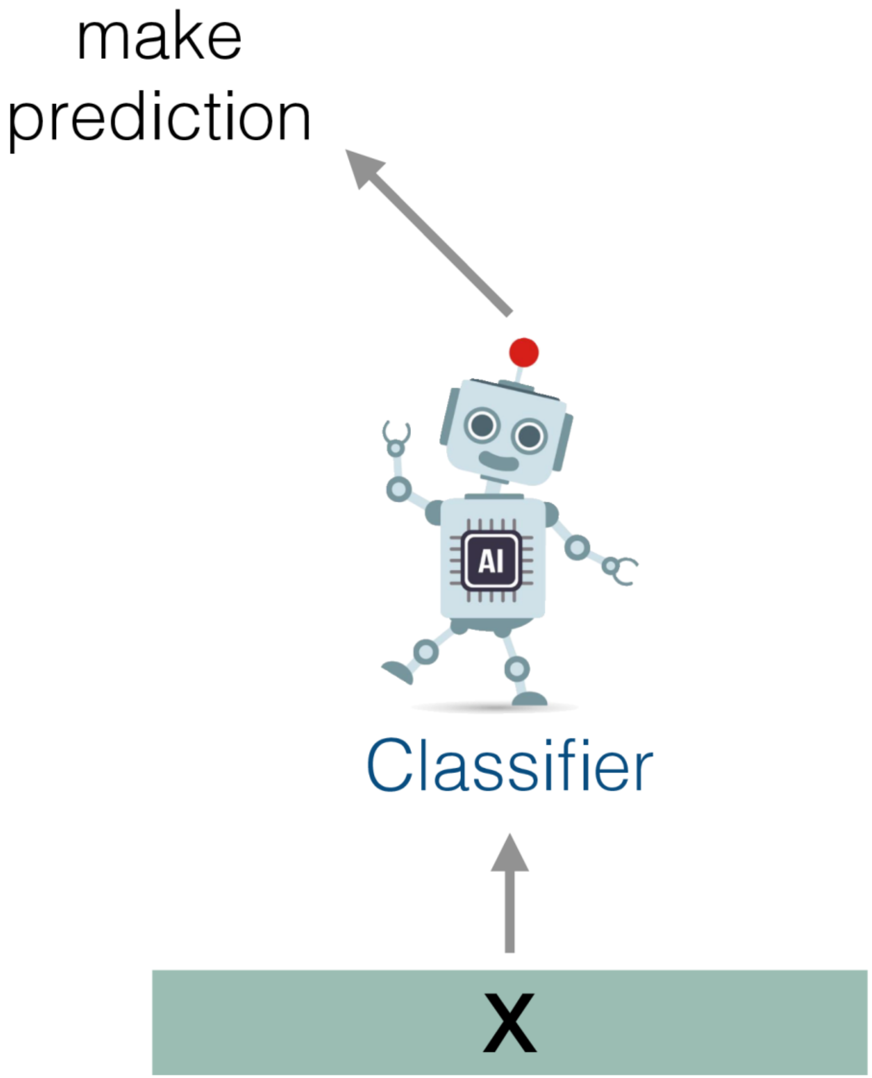

x_2
C_1
DISEASE A
DISEASE B
reject
C_2
C_3
HEALTHY
x_1
Difficult region where we probably should pass the judgement off to a human expert

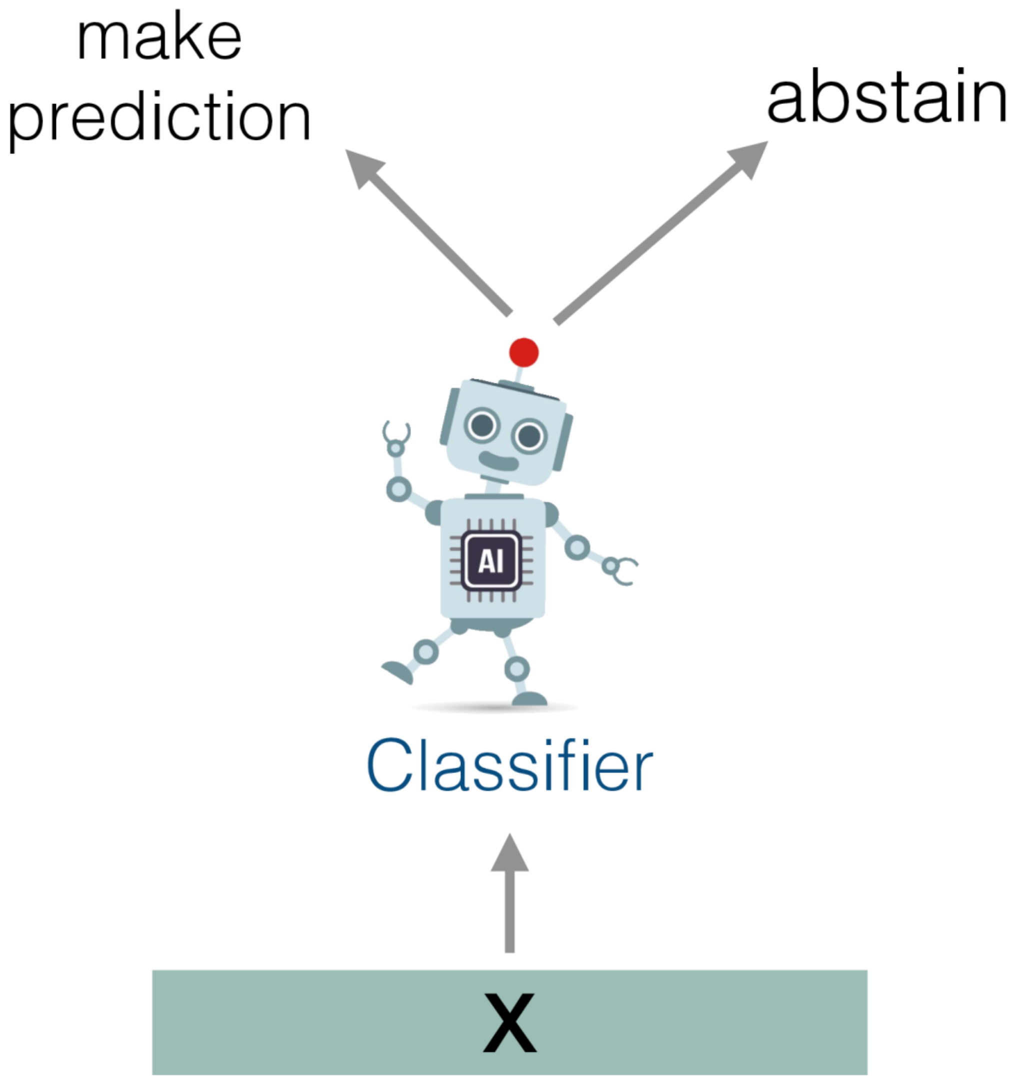

X

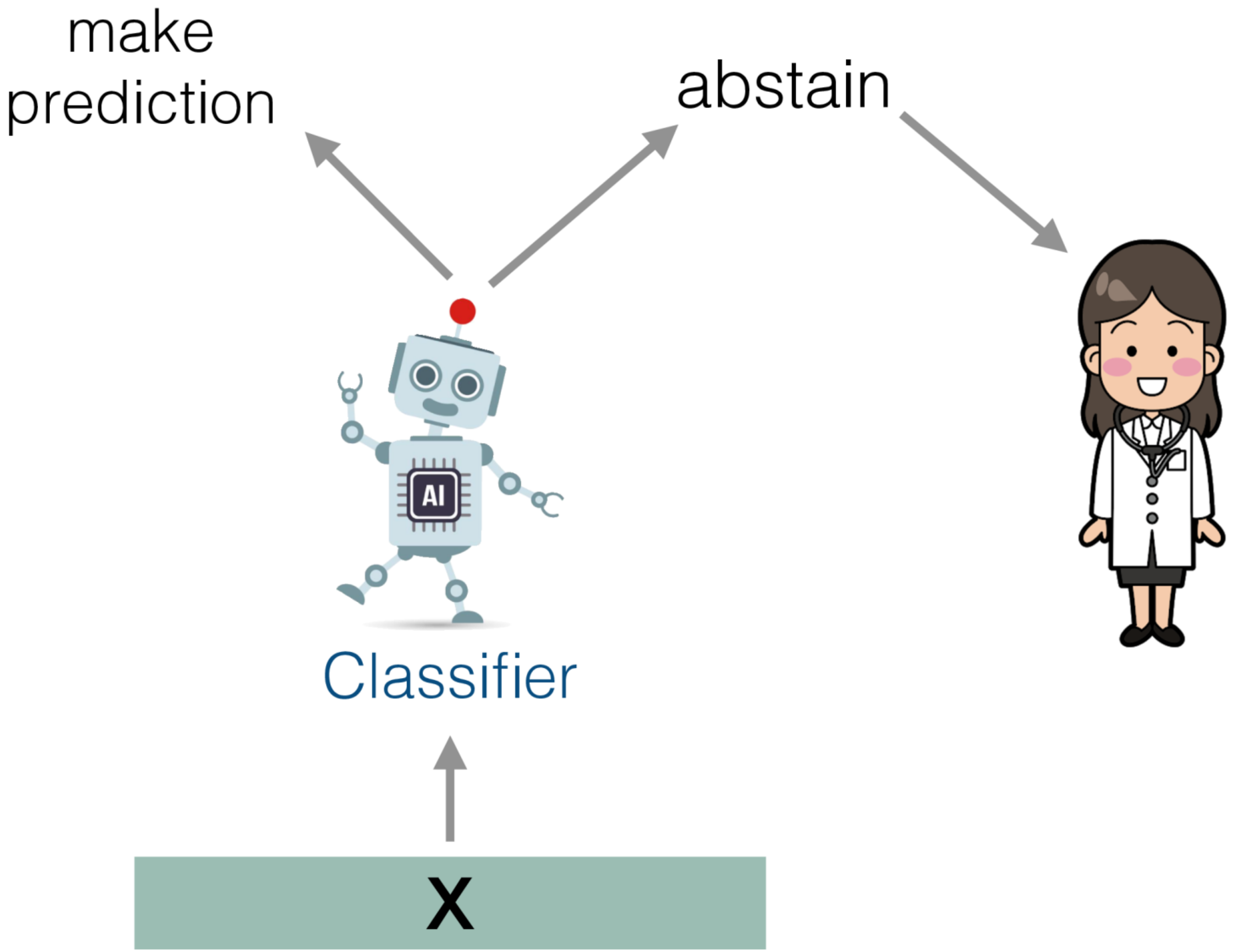


Classifier

X

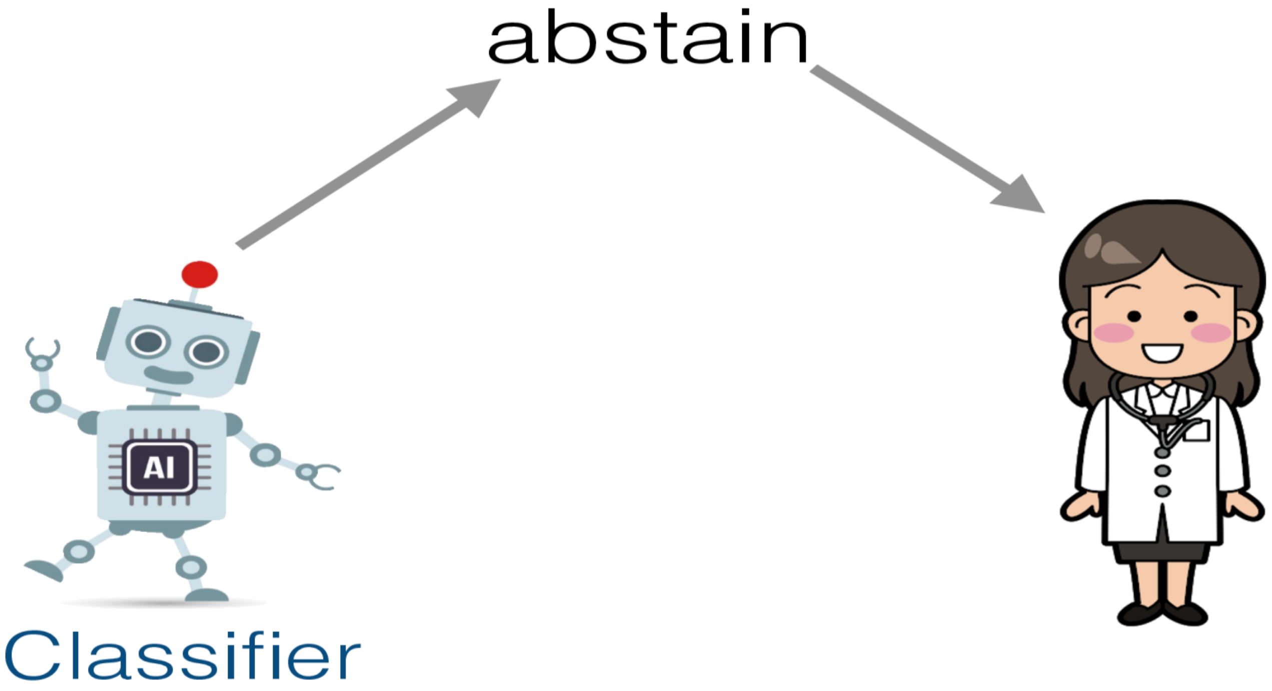

make
prediction
Classifier
X


make
prediction
abstain
AI
Classifier
X


make
prediction
abstain
AI
Classifier
X

Score-Based Rejection: Abstain if the model is unconfident in its prediction:

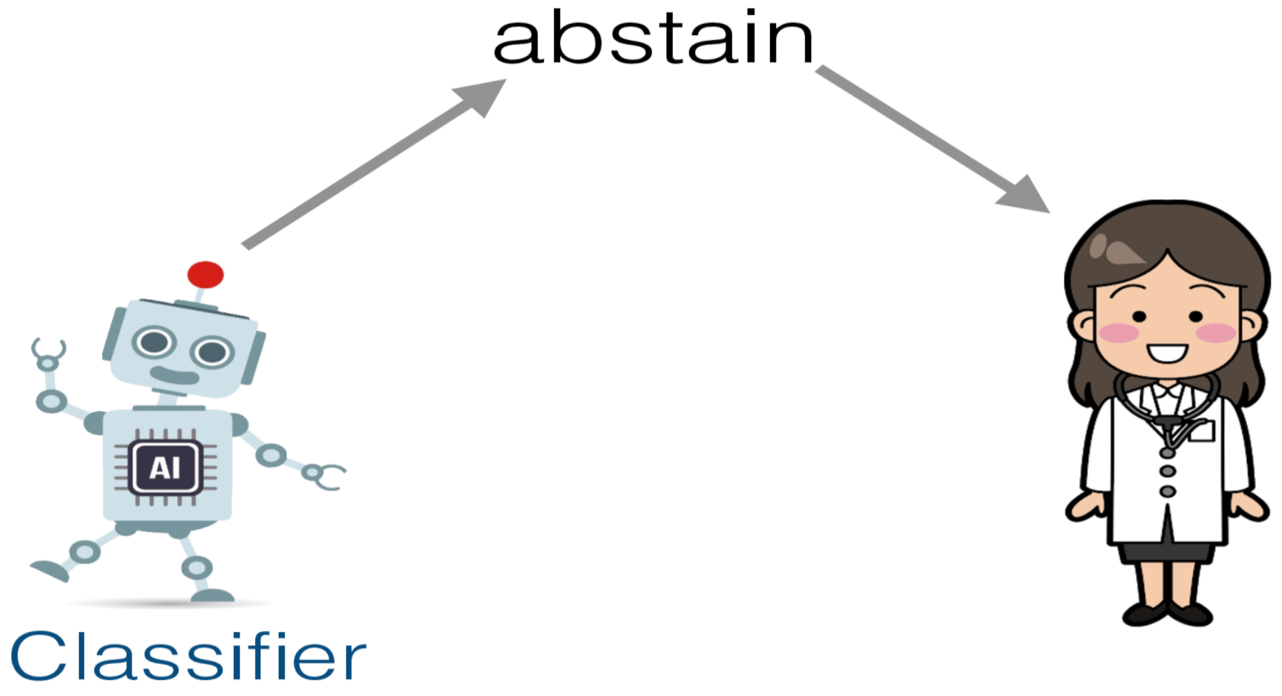
$$\max_{y} \; \mathrm{P}(\mathrm{y} \,|\, \mathrm{x}_n) \; < \; \tau$$



Score-Based Rejection: Abstain if the model is unconfident in its prediction:

$$\max_{y} \; \mathrm{P}(\mathrm{y} \mid \mathrm{x}_n) \; < \; \tau$$

Human behavior is not modeled!



Challenge: how can we model the human?

If they are a true expert, modeling their decision making—$\mathbb{P}_h(\mathbf{y} \mid \mathbf{x})$—is assumed to be impossible.

Better Formulation

Model what the human knows,
so we can enable *collaboration*

Better Formulation

Model what the human knows,
so we can enable *collaboration*

Data: $\mathfrak{D} = \left\{ \mathrm{x}_n, \mathrm{y}_n, \mathrm{m}_n \right\}_{n=1}^{N}$

expert predictions

Better Formulation

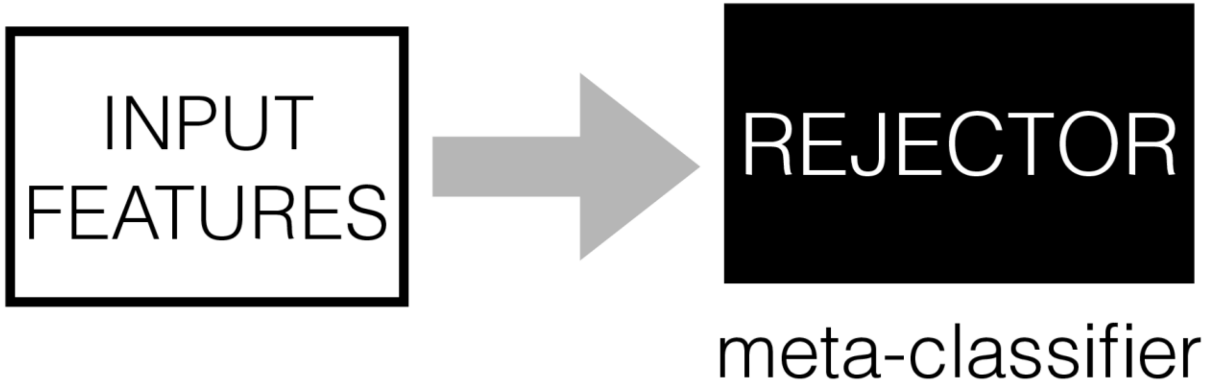
Model what the human knows,
so we can enable *collaboration*

Data: $\mathfrak{D} = \{\mathbf{x}_n, \mathbf{y}_n, \mathbf{m}_n\}_{n=1}^{N}$

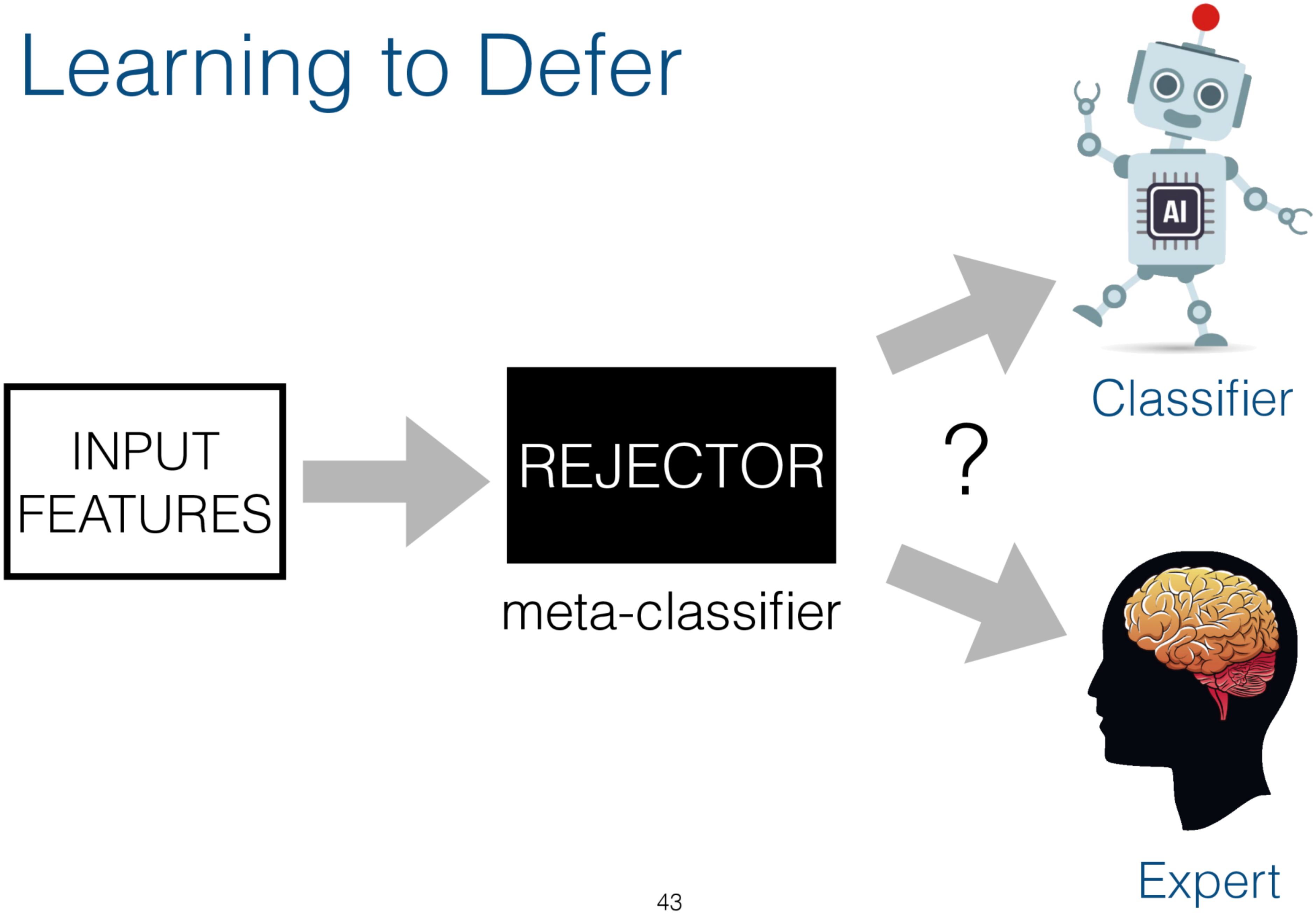
expert predictions

Models: $\mathbf{r}(\mathbf{x})$ Rejector $\mathbf{h}(\mathbf{x})$ Classifier

Learning to Defer



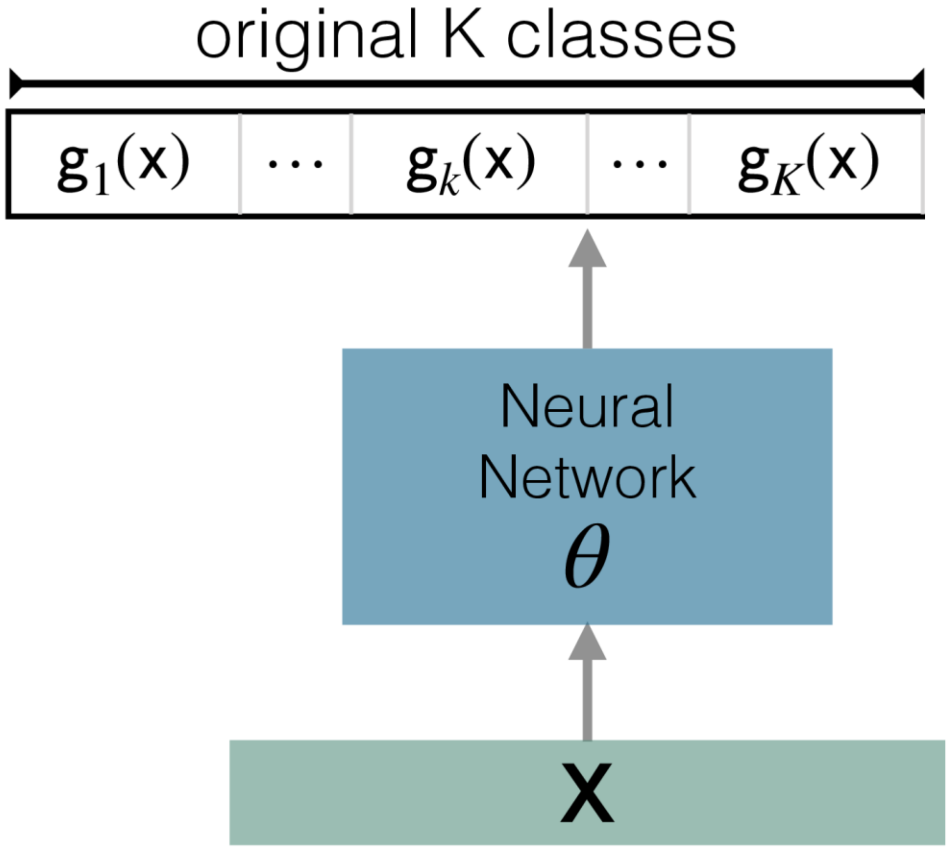
Learning to Defer



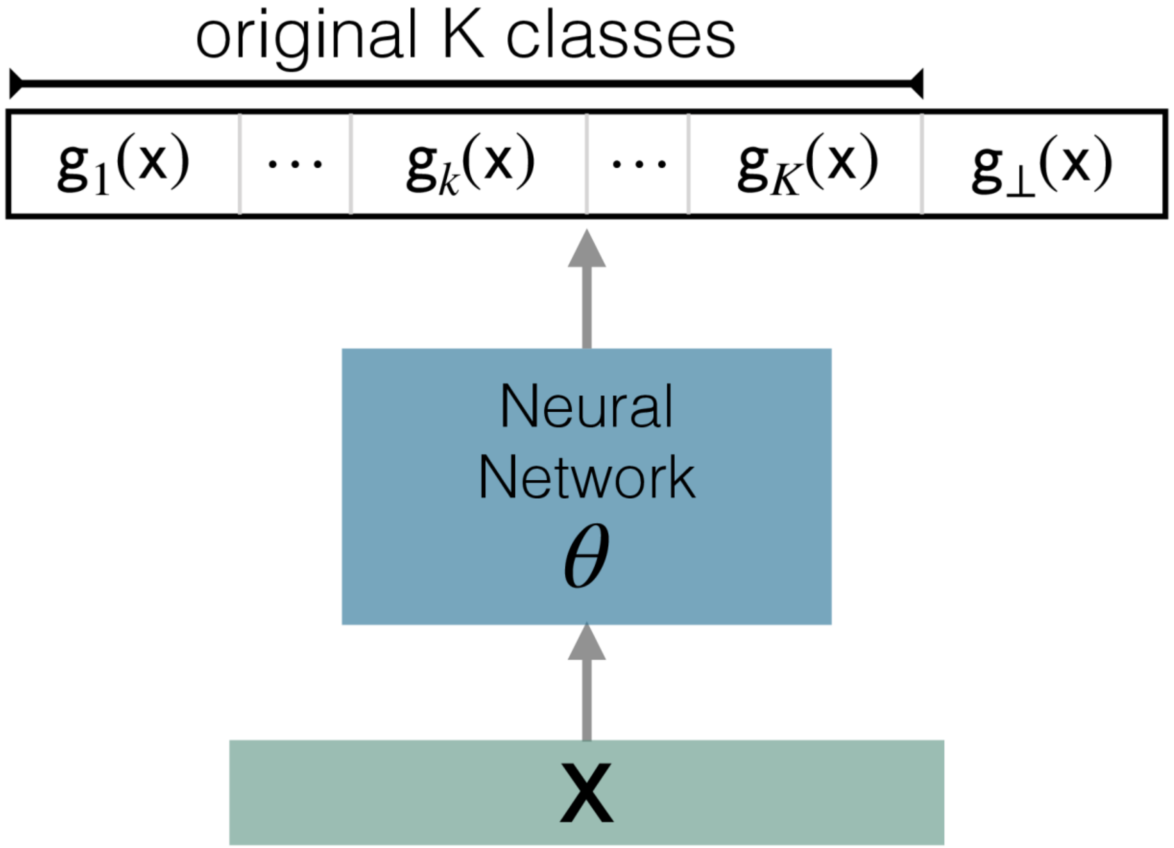
Softmax Approach [Mozannar & Sontag, ICML 2020]

Mozannar, Hussein, and David Sontag. "Consistent estimators for learning to defer to an expert." *International Conference on Machine Learning*. PMLR, 2020.

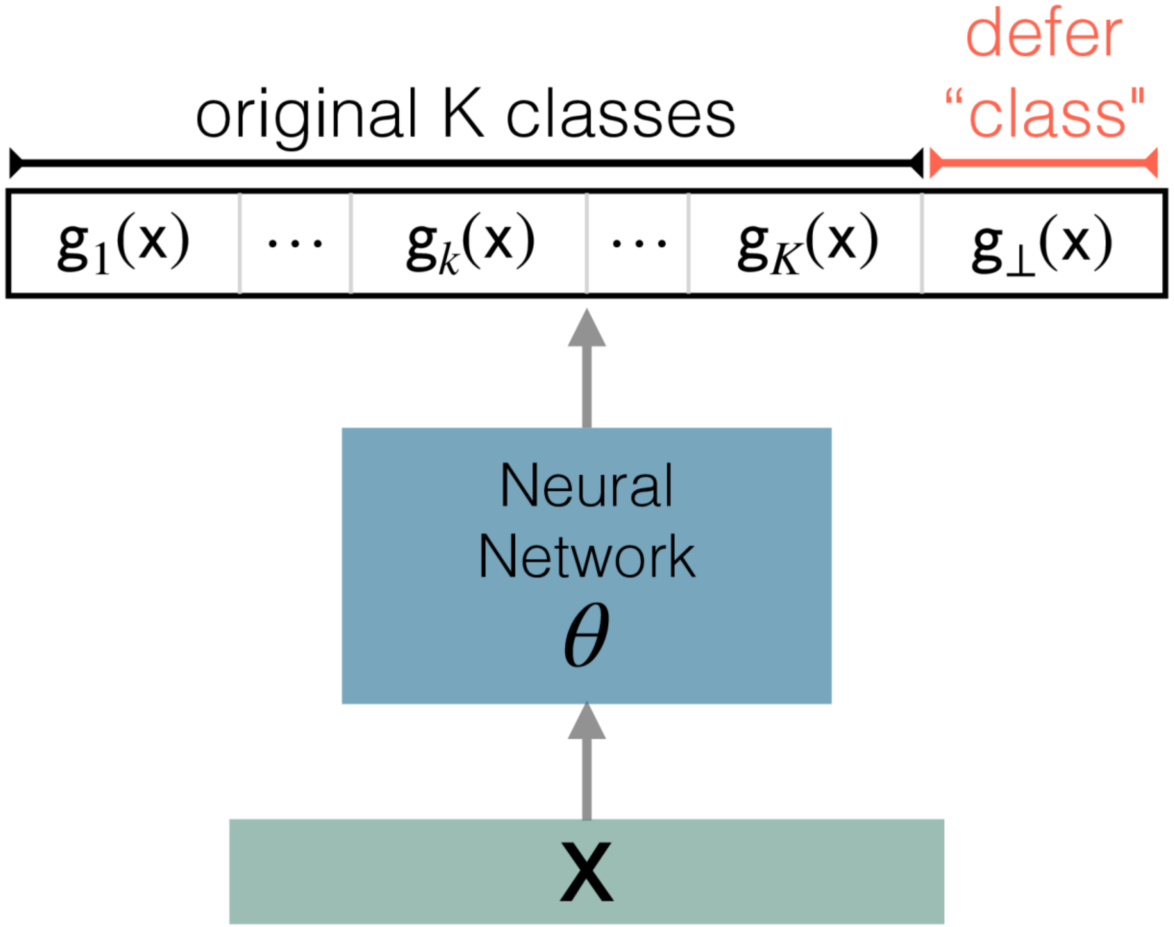
Softmax Approach [Mozannar & Sontag, ICML 2020]



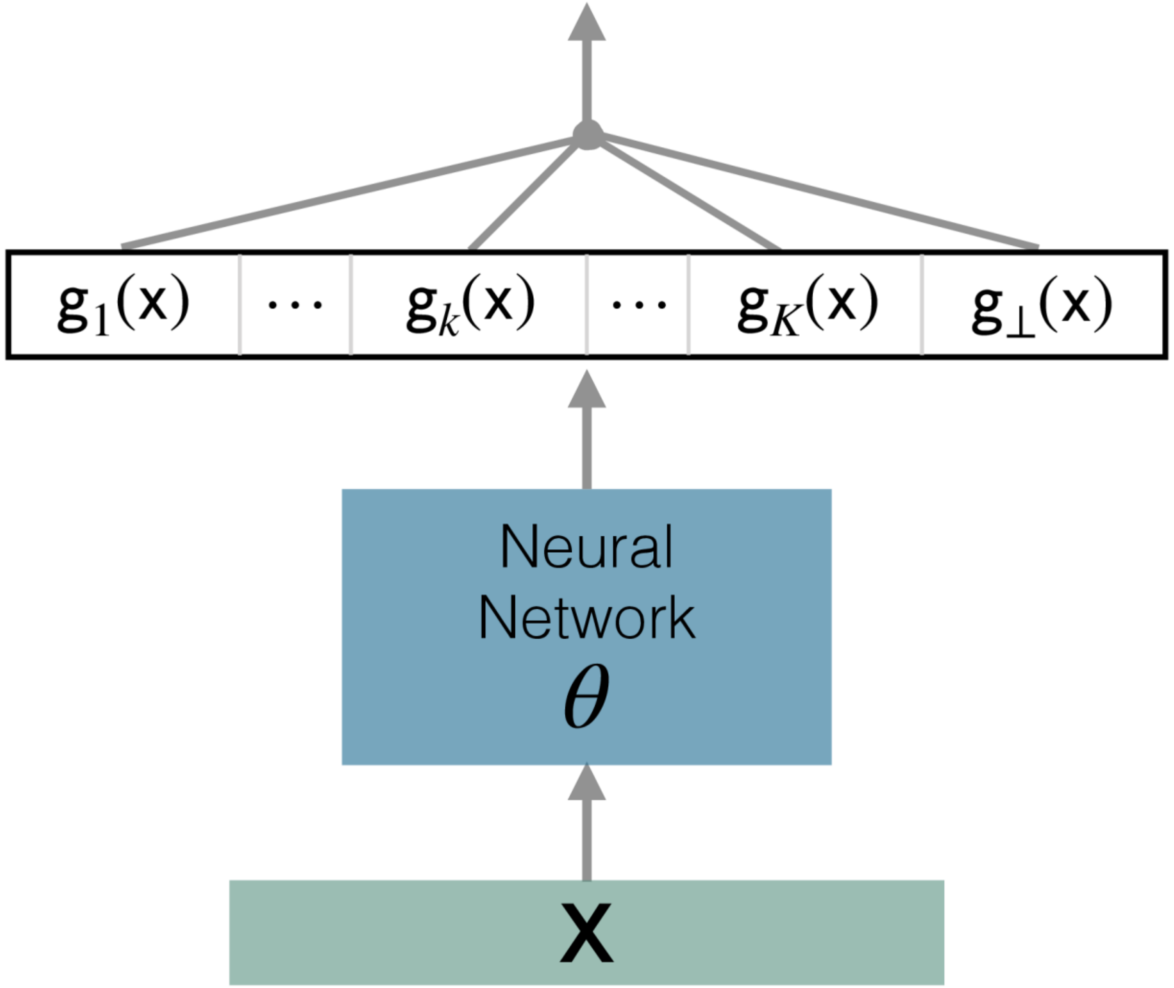
Softmax Approach [Mozannar & Sontag, ICML 2020]



Softmax Approach [Mozannar & Sontag, ICML 2020]

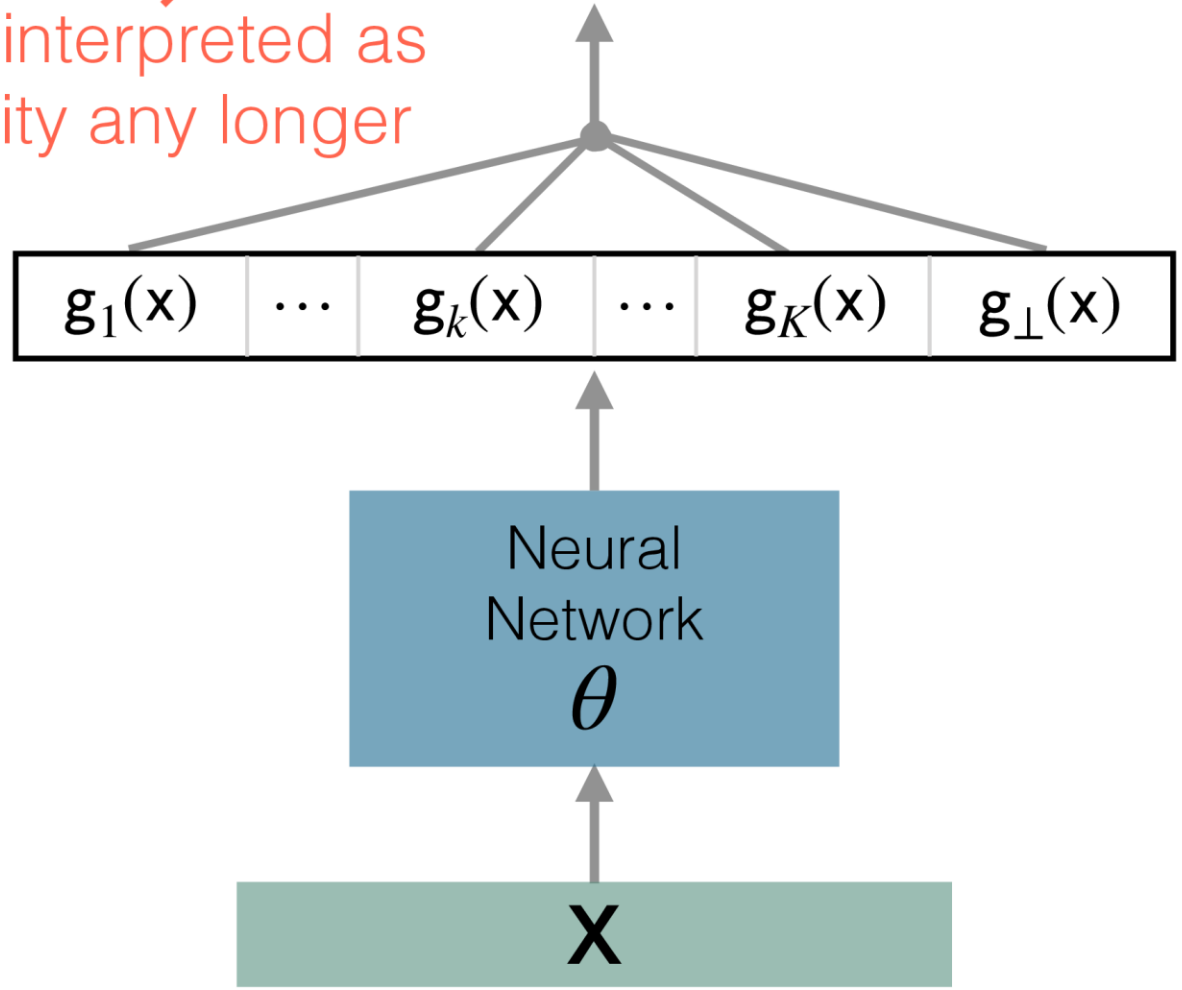


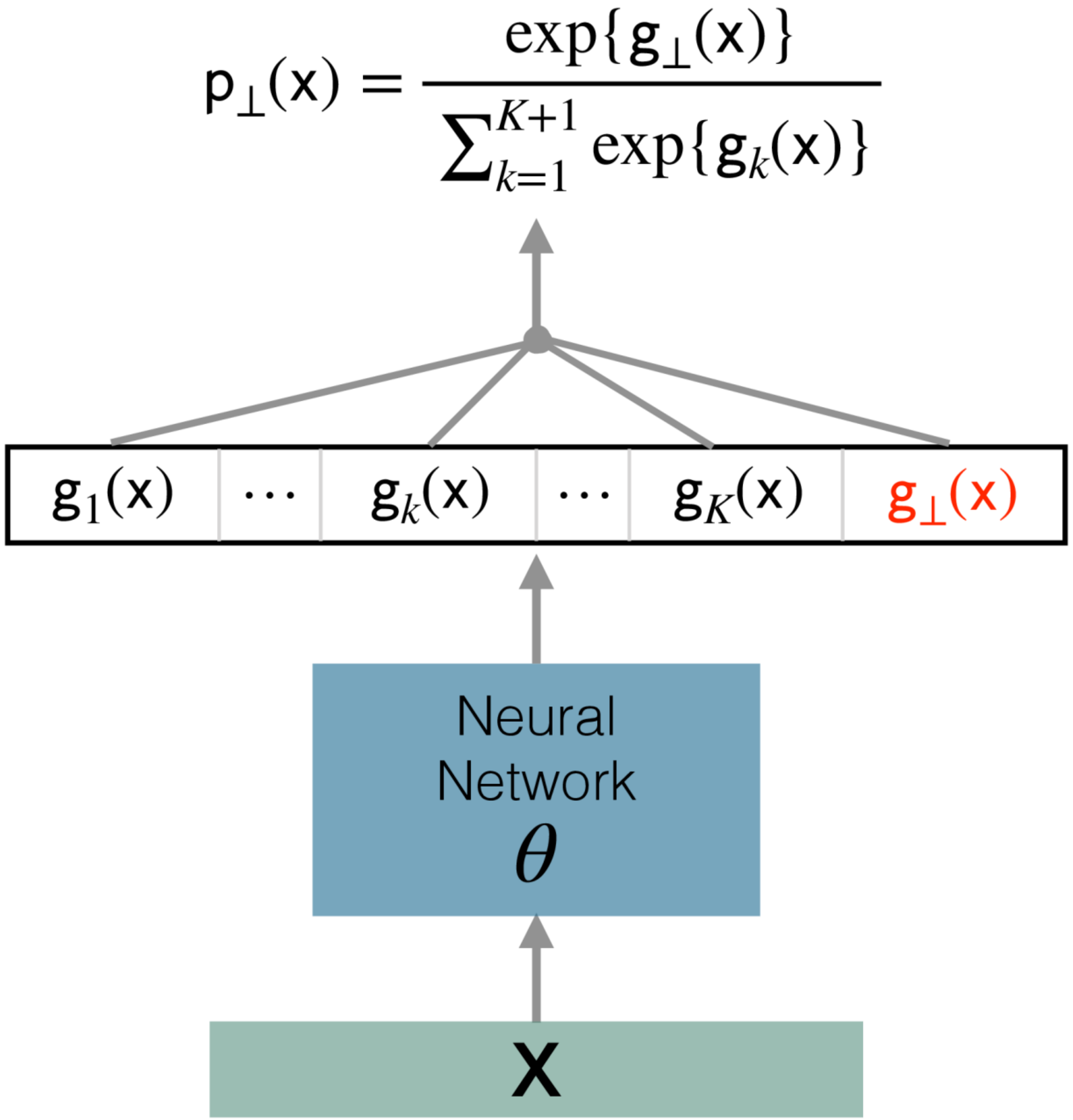
$$p_i(x) = \frac{\exp\{g_i(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$$



$$\mathrm{p}_i(\mathrm{x}) = \frac{\exp\{\mathrm{g}_i(\mathrm{x})\}}{\sum_{k=1}^{K+1} \exp\{\mathrm{g}_k(\mathrm{x})\}}$$

cannot be interpreted as a probability any longer





$p_\perp(x) = \frac{\exp\{g_\perp(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$
$g_1(x)$ $\cdots$ $g_k(x)$ $\cdots$ $g_K(x)$ $g_\perp(x)$
Neural Network
θ
x

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_n \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \ + \ \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \ \log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_n \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \;+\; \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \; \log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_{n}\left(\underbrace{\log \mathsf{p}_{y_n}(\mathsf{x}_n)}_{\text{classifier loss}} + \underbrace{\mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \; \log \mathsf{p}_{\perp}(\mathsf{x}_n)}_{\text{rejector loss}}\right)$$

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_{n} \left(\underbrace{\log \mathsf{p}_{y_n}(\mathsf{x}_n)}_{\text{classifier loss}} + \underbrace{\mathbb{I}[\mathsf{y}_n = \mathsf{m}_n]}_{\text{only if expert is correct}} \underbrace{\log \mathsf{p}_{\perp}(\mathsf{x}_n)}_{\text{rejector loss}} \right)$$

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

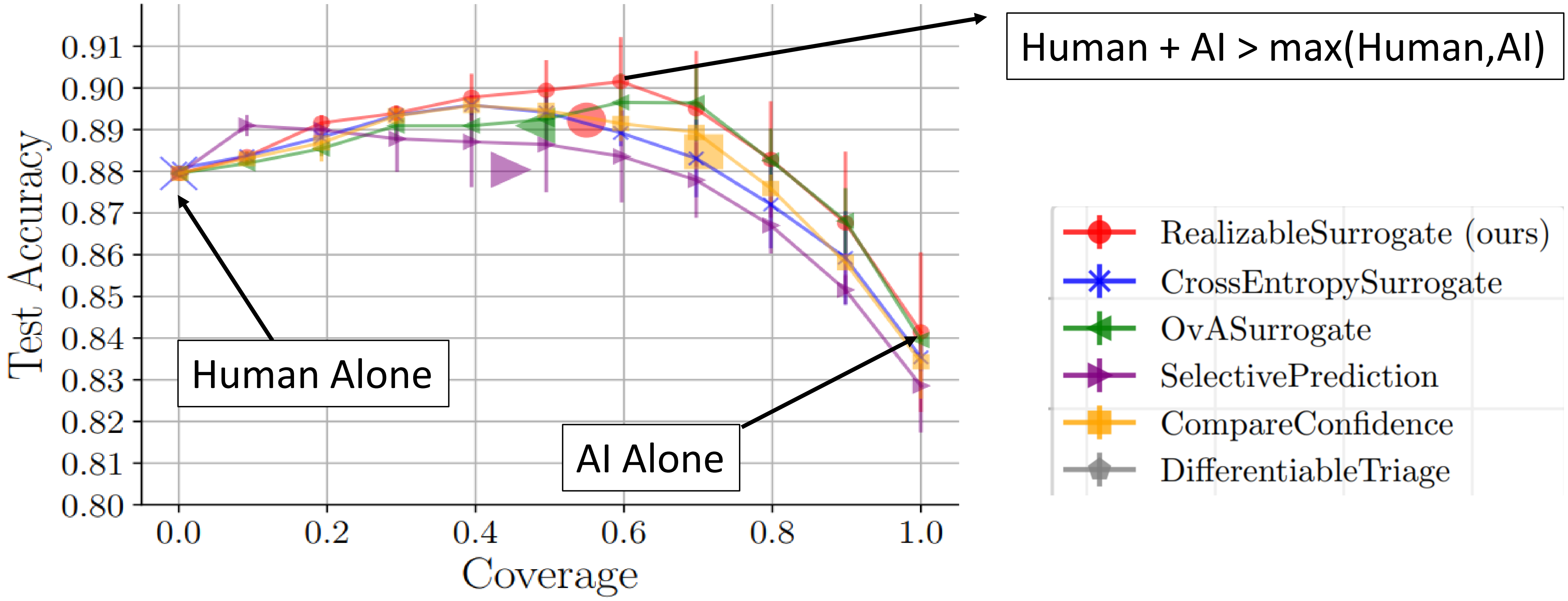
$$-\sum_n \left(\log \mathrm{p}_{y_n}(\mathrm{x}_n) \; + \; \underbrace{\mathbb{I}[\mathrm{y}_n = \mathrm{m}_n]}_{\text{only if expert is correct}} \; \log \mathrm{p}_{\perp}(\mathrm{x}_n) \right)$$

classifier loss

rejector loss

Consistency: The minimizrs (w.r.t. g) correspond to the Bayes optimal classifier and rejector

Chest Xray (NIH dataset) Results



(e) Chest X-ray - Airspace Opacity

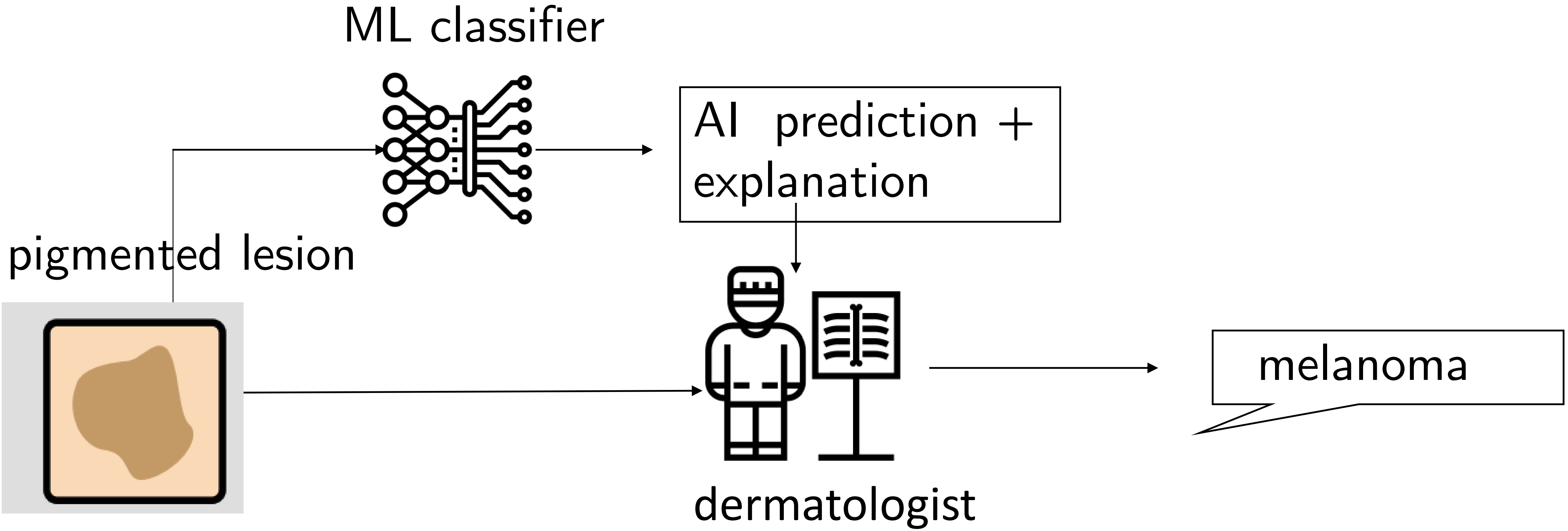
Mozannar, Hussein, et al. "Who Should Predict? Exact Algorithms For Learning to Defer to Humans." AISTATS 2023.

Triage can help towards automation

- The last iteration of the diabetic retinopathy project implemented this deferral setup with ungradable images being graded by an ophthalmologist.
- The human-AI team satisfies the constraints of the clinic, and if the rejector is chosen appropriately, can improve performance of the team
- However, when clinician time is less scarce, we can allow for more explicit interaction between human-AI

Model as a second opinion

Classify lesion into one of 7 categories: melanoma, ..., vascular lesions [1]



[1]:Tschandl, Philipp, et al. "Human–computer collaboration for skin cancer recognition." *Nature Medicine* 26.8 (2020): 1229-1234.

AI second opinion for skin cancer recognition

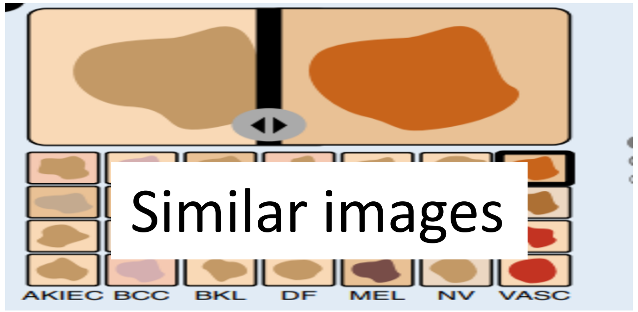
- 155 raters classified each 28 random images, and their performance (time and accuracy) was first measured (1) without AI and then (2) with AI predictions and explanations.

- Performance can vary based on two factors: 1) the AI explanations and 2) the specific dermatologist

[1]:Tschandl, Philipp, et al. "Human–computer collaboration for skin cancer recognition." *Nature Medicine* 26.8 (2020): 1229-1234.

Form of AI explanations has a big effect

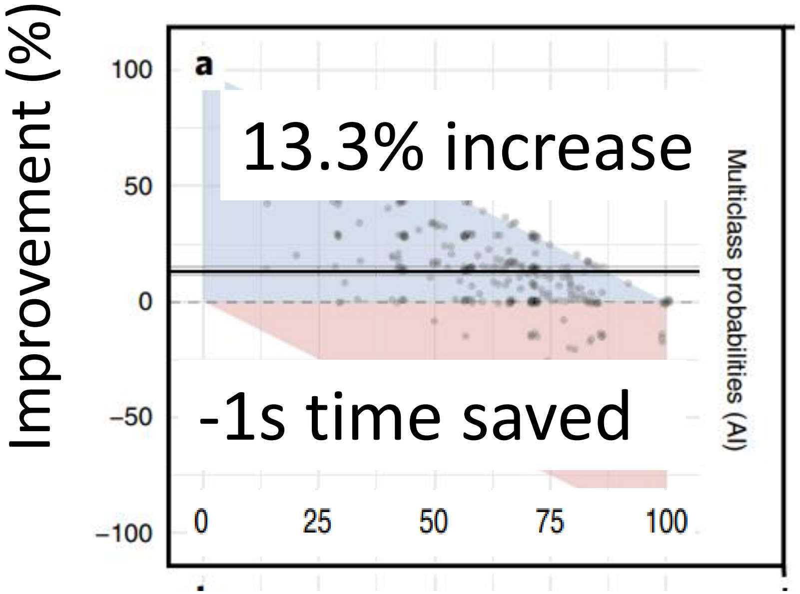
Multiclass probabilities



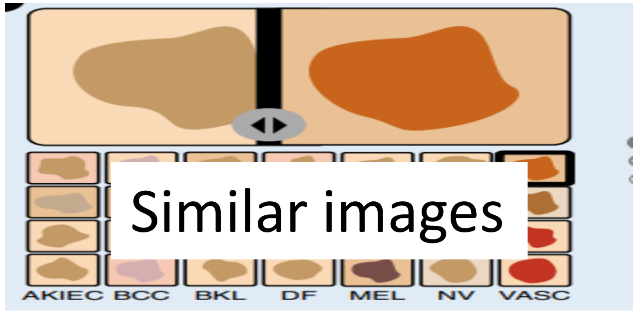
Which Explanation will clinicians benefit more from?

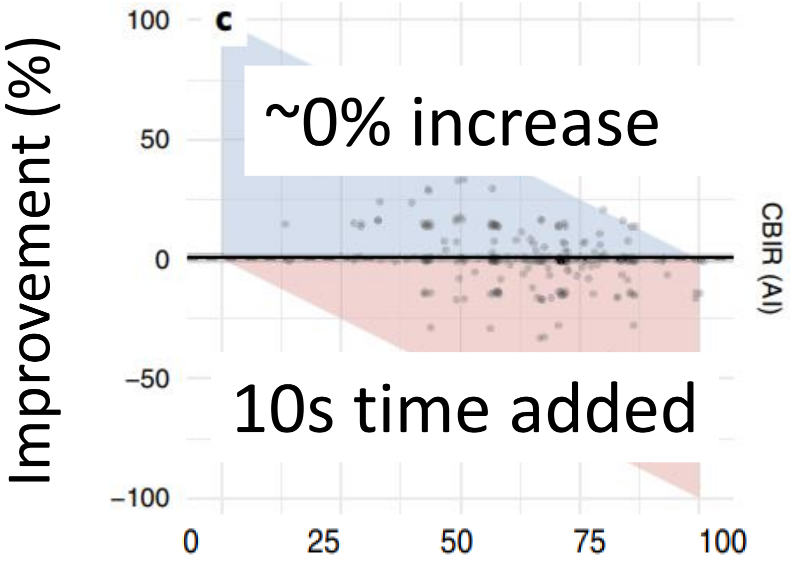
Form of AI explanations has a big effect

Multiclass probabilities



Accuracy before interaction





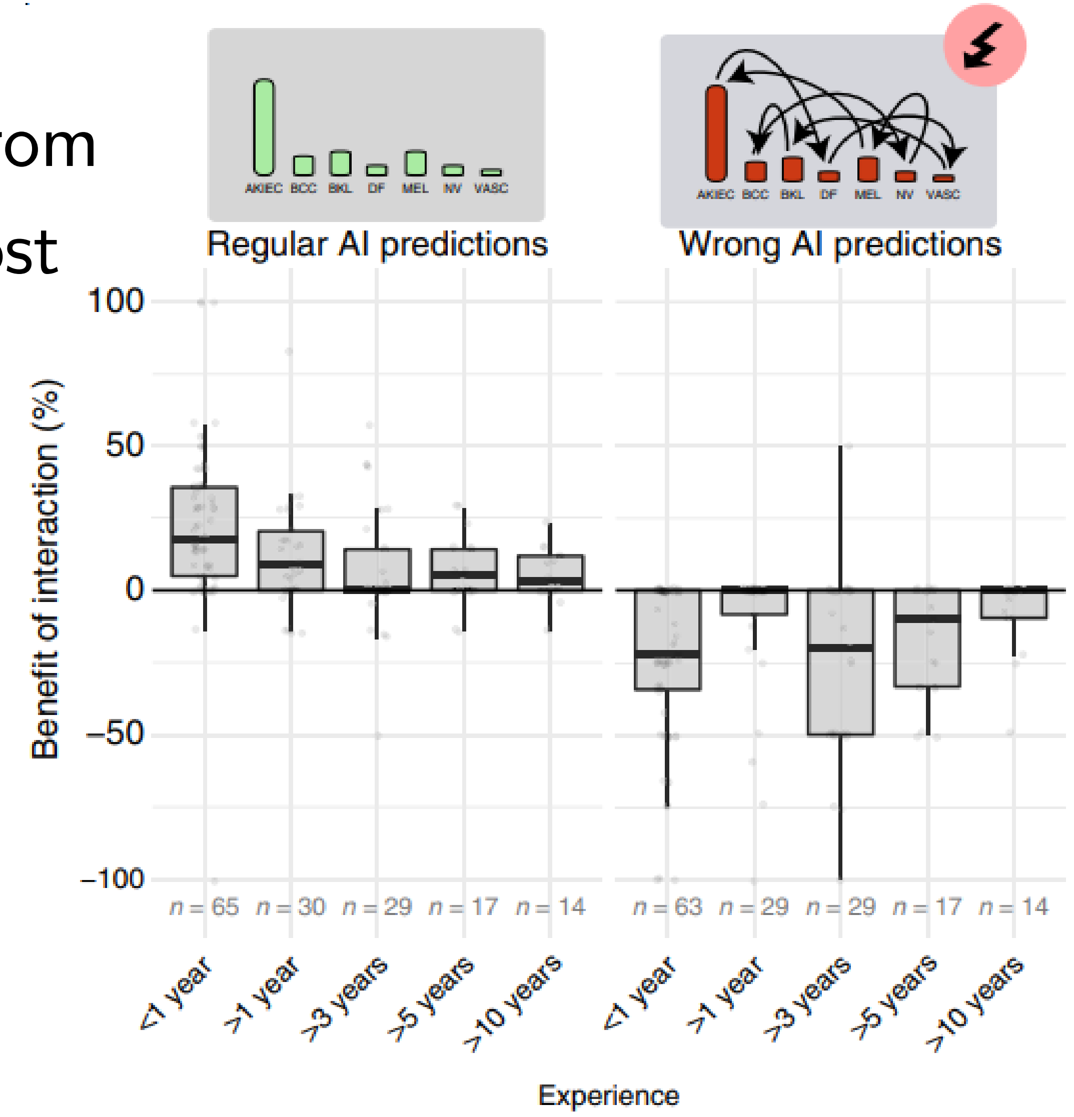
Accuracy before interaction

Possible benefit Possible loss

Clinician Experience and Confidence affects interactions

Clinician Experience and Confidence affects interactions

- Inexperienced raters benefit hugely from the regular AI, but are harmed the most from a bad AI model
- Experienced rater benefit the least from regular AI, and are harmed the Least by a bad AI model
- The difference is how sound their mental model of the AI is

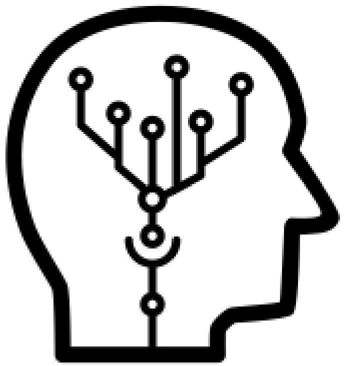


Outline

- Modes of Human-AI Interaction
- **Mental Models**
- Onboarding
- Over-reliance on AI and fixes

Mental Models

- **Mental model:** a person's understanding of how something works and how their actions affect it.
  - based on beliefs, flexible, limited and filters information.
  - sets expectation about what something can and cannot do and value can be gained from it
- What is special about **mental models of AI?**
  - Our priors are often wrong
  - It is hard to experiment with the AI model
  - AI's are evolving



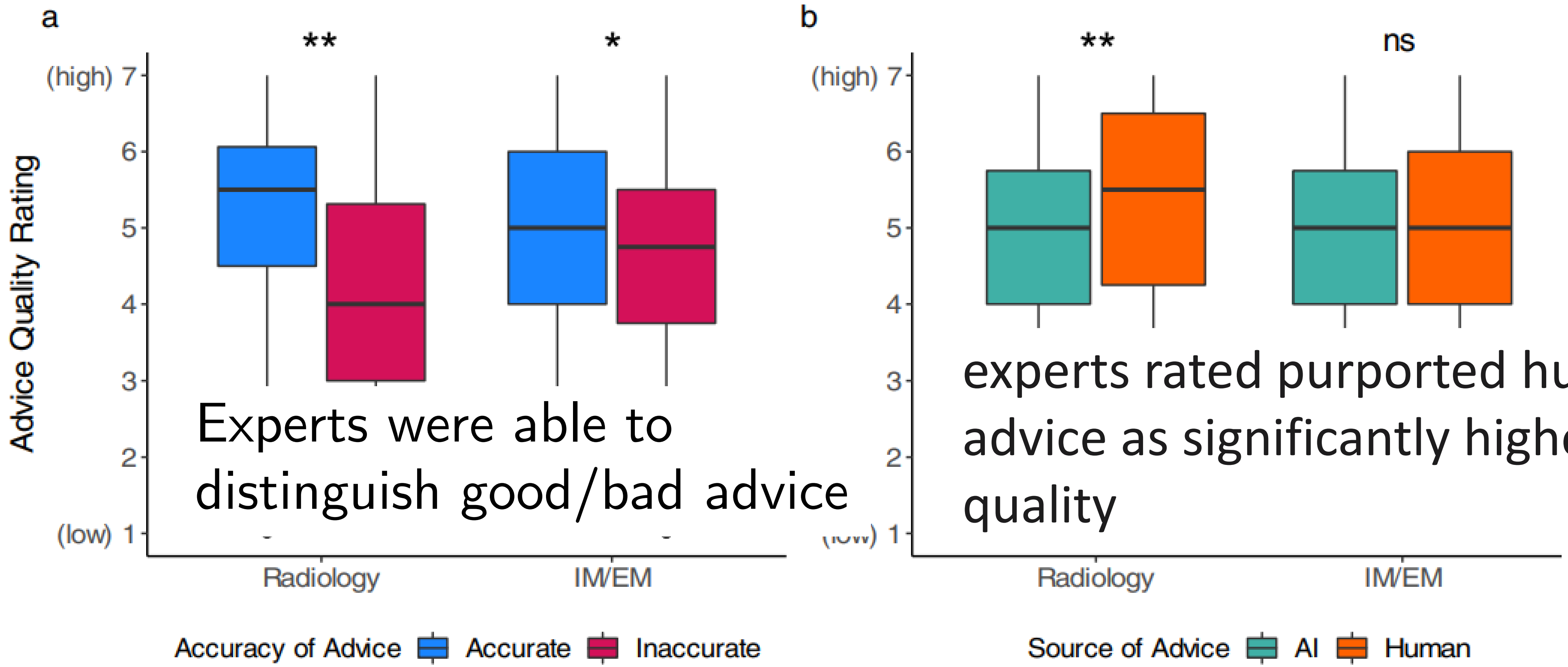
Mental Models Experiment

- Radiologists and physicians were presented with 8 cases: told the advice they get is from a human or an AI, and then are asked to rate advice quality.
- Trick is that all the advice is from a human and only on 4 cases is it correct

[1]:Gaube, Susanne, et al. "Do as AI say: susceptibility in deployment of clinical decision-aids." *NPJ digital medicine* 4.1 (2021): 1-8.

1) Will advice said to be given by an AI be rated lower or higher than that by a human?
2) Will this vary by the radiologist's expertise?

Human advice is rated higher than AI



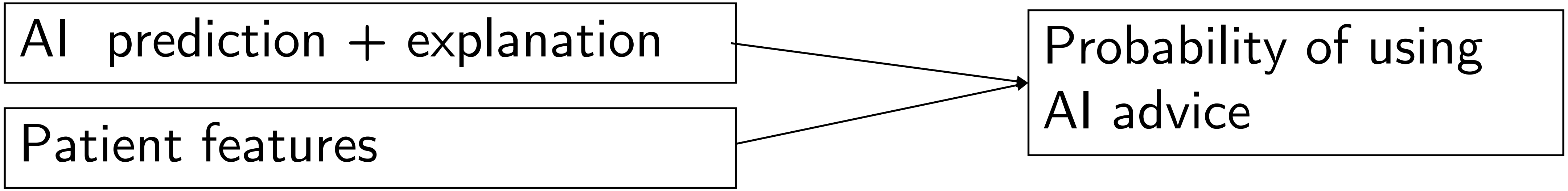
Experts were able to distinguish good/bad advice

experts rated purported human advice as significantly higher quality

[1]:Gaube, Susanne, et al. "Do as AI say: susceptibility in deployment of clinical decision-aids." *NPJ digital medicine* 4.1 (2021): 1-8.

Mental Model of AI

- **Mental model definition**: internal human map



- **How to measure it:**
  - Compute Trust: how often AI prediction and human decision agree
  - Stratify human accuracy by AI predictions being correct/incorrect
  - Questionnaires that try to elicit human's understanding of the AI (often what they say is not how they behave) [1]

[1]:Buçinca, Zana, et al. "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems." Proceedings of the 25th international conference on intelligent user interfaces. 2020..

Factors affecting the Mental Model

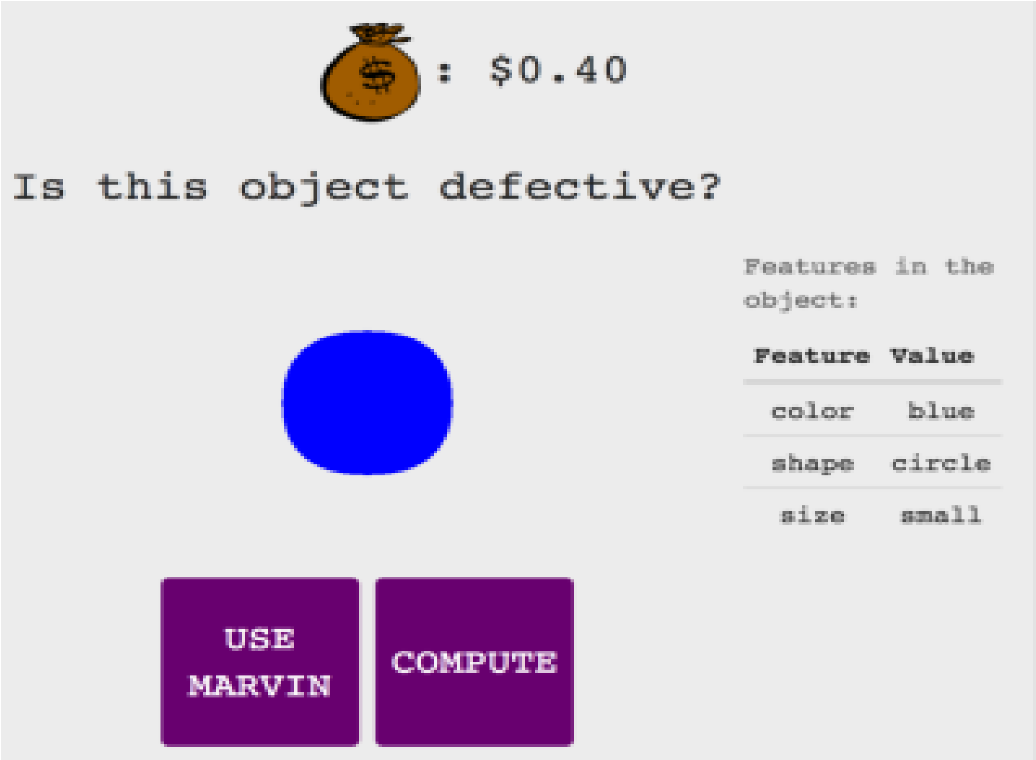
- Experimental setup [1,2],
- Payoff Matrix

	Marvin Correct	Marvin Wrong
Use Marvin	\$0.04	-\$0.16
Compute	0	0

Get Feedback immediately

- AI "Marvin" is 80% correct depending on condition on object: example

$F = \text{blue} \cap \text{square}$ and $P(\text{error}|F)$



[1]:Bansal, Gagan, et al. "Beyond accuracy: The role of mental models in human-AI team performance." Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. Vol. 7. 2019. [2]: Bansal, Gagan, et al. "Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

Stochasticity and AI Complexity

1. As error boundary is more **stochastic**, it becomes harder for users to know when to use AI

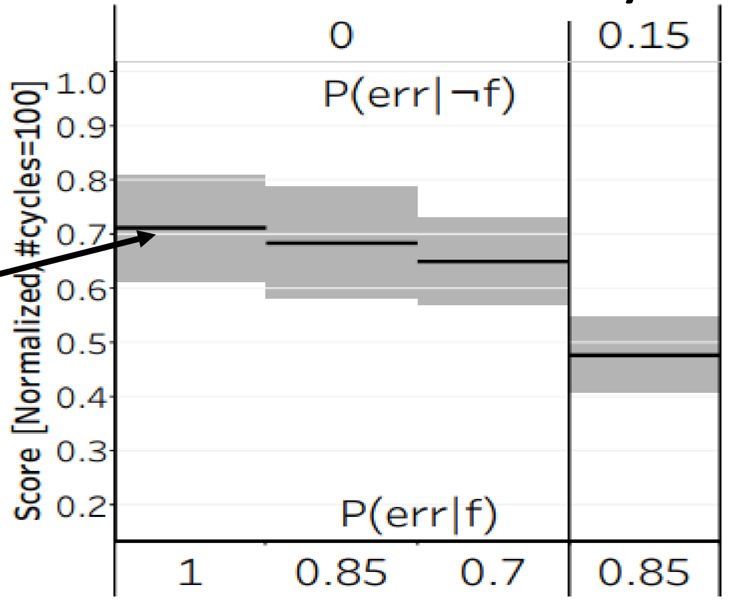
Change P(err|F) from deterministic error, to more stochastic

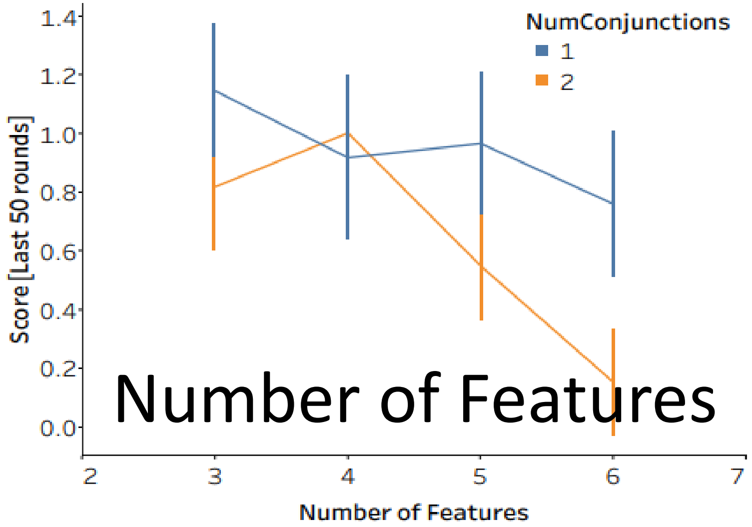
2. As AI error boundary becomes more **complex,** harder to detect error.

i.e. F1 = = blue ∩ square (1 conjunction, 2 features) vs F2 = (blue ∩ square) ∪ (red ∩ circle) (2 conjunctions, 2 features), F3= blue ∩ square ∩ small

F2 more complex than F1, F3 more complex than F1

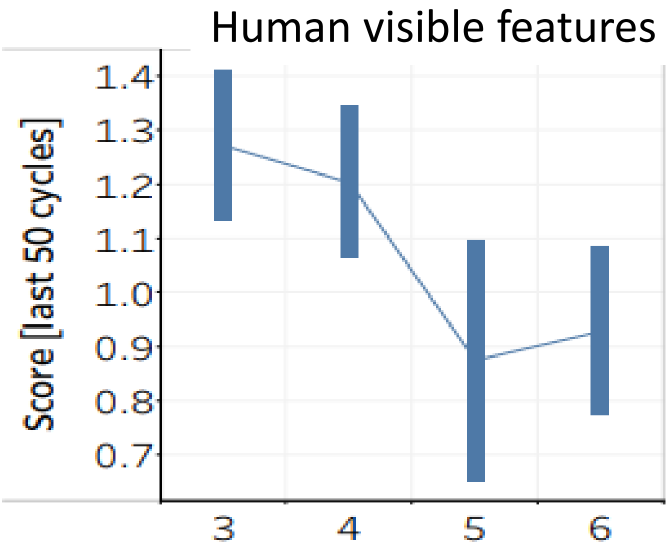
AI Stochasticity



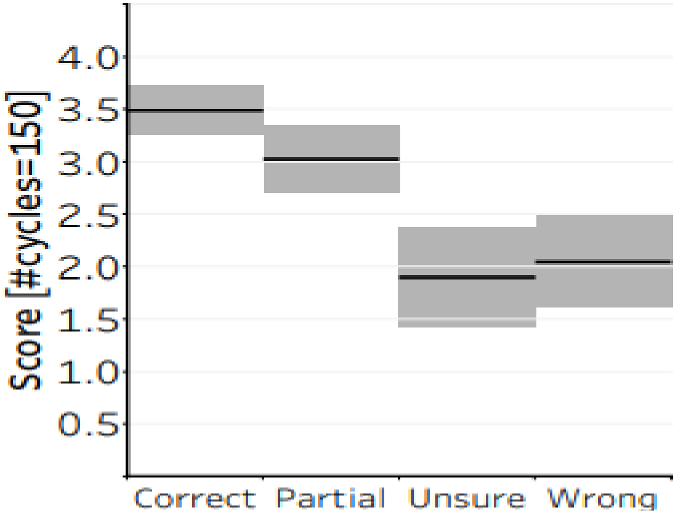


Observable Features

3. As human observes more **features about the object**, becomes harder to detect AI error boundary



Better mental models (i.e., knowing the AI error boundary) -> better score. Measured by letting participants describe the AI



Mental Model Accuracy

Takeaways of Mental Models

- Humans rely on their mental model of the AI to know when to use it
- Accurate mental models of AI's error boundary -> better task performance, and influenced by the following factors:
  1. **Stochasticity of AI:** how predictable are the errors
  2. **Complexity of AI:** size of the error boundary description
  3. **Human observable features:** amount of information available to humans
- **Unresolved question:** How can we allow humans to understand the AI error boundary better?

Outline

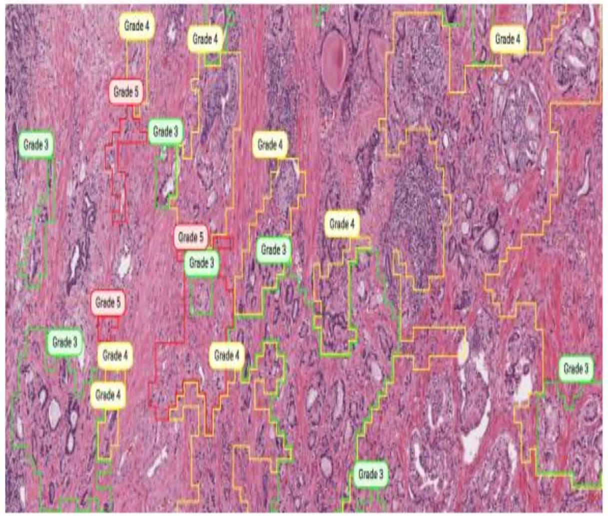
- Modes of Human-AI Interaction
- Mental Models
- **Onboarding**
- Over-reliance and under-reliance on AI

Mental Model Formation

- Recap: How do humans know when to use the AI
  - Rely on their mental model which is a function of the AI's explanations (e.g., confidence score) and their knowledge and experience with the AI (through interacting with it)
- In almost all research mentioned, the AI was initially described to the users.
- How to onboard users on the AI and what information should we share?

Study of Onboarding in Pathology

- 21 pathologists on task to understand prostate cancer risk [1]
- **Pre-Probe:** What types of information would you need to know about an AI assistant before using it?
- **Probe:** Diagnose a case with AI assistant
- **Post-probe:** What other information would you need to know about an AI assistant to work with it effectively?

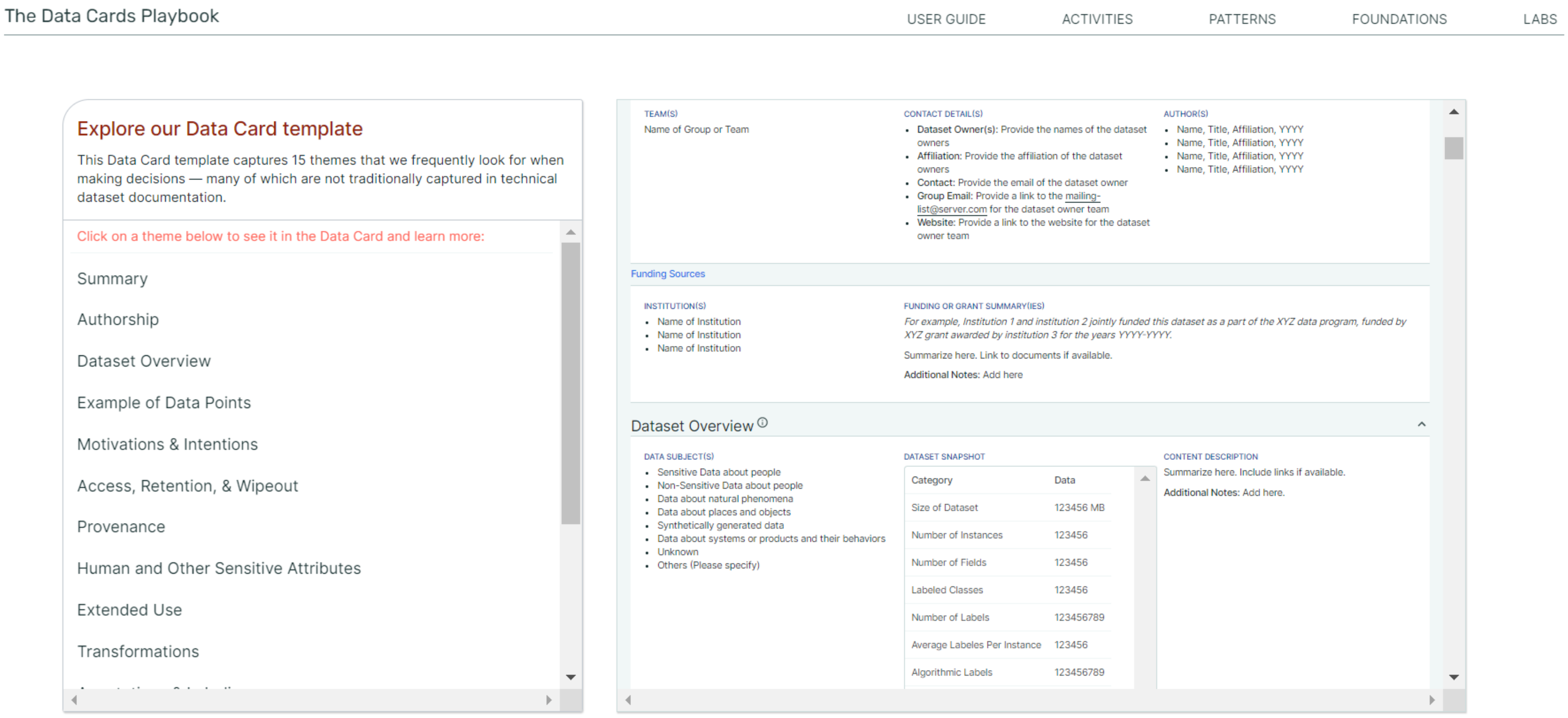


[1]:Cai, Carrie J., et al. "" Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making." Proceedings of the ACM on Human-computer Interaction 3.CSCW (2019): 1-24..

Training and Inference

- **Describe the scale of the training data.**
  - Some suggested that the number of data points should be on par with the volume of cases pathologists are typically trained on…
- **Describe the diversity of the training data.**
  - “More variation is better… Covering from community hospital to small groups, to academic medical centers”
- **Enumerate the data modalities that are accessible to the algorithm.**
  - “Does the AI assistant have access to information that I don’t have? Does it have access to any ancillary studies?”
  - “I want to know if the AI is being generated off of one image of if it’s being generated on sequential images… Sequential I would trust more.

Enable this with Data Cards



https://sites.research.google/datacardsplaybook/

Training and Inference

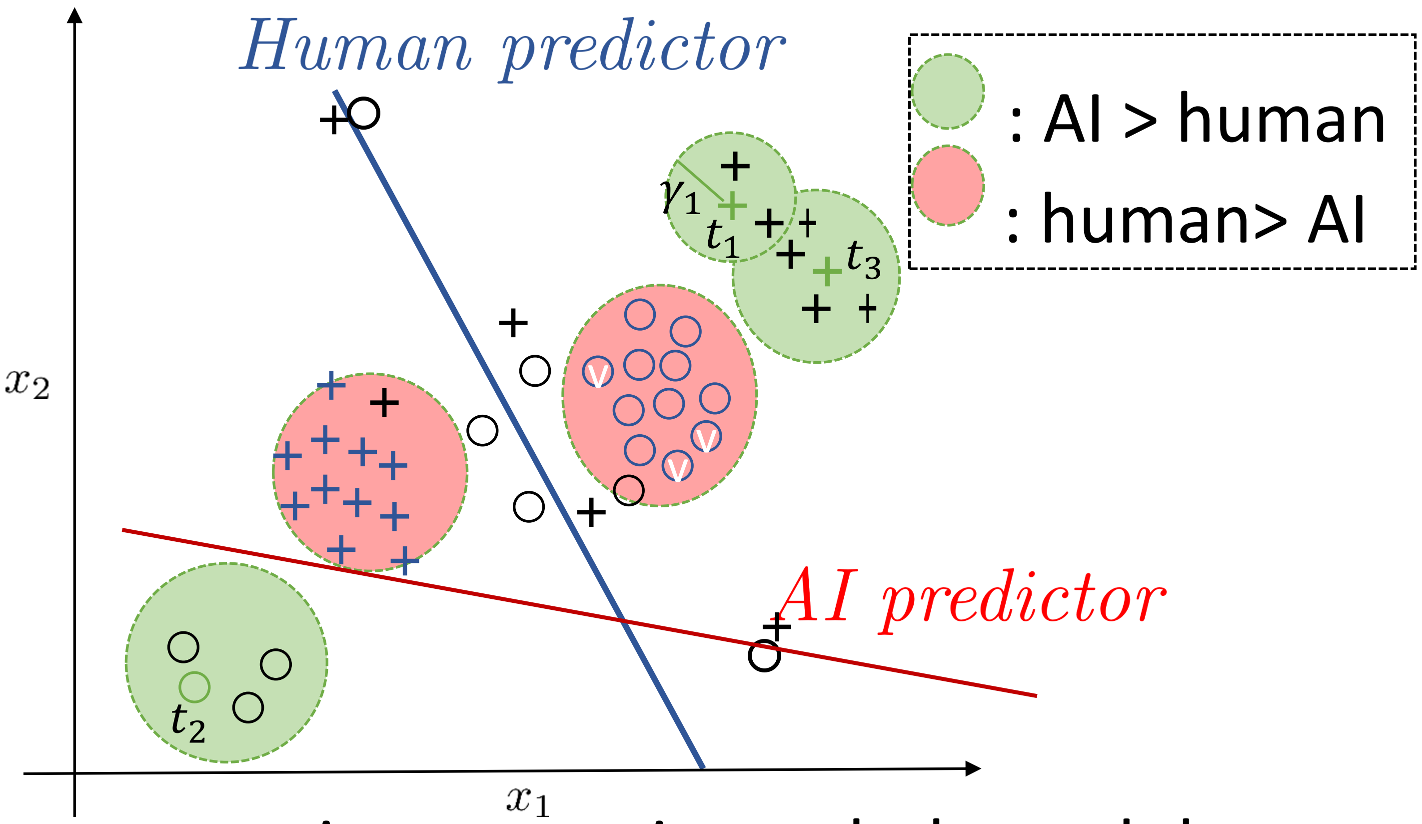
- **Specify the main steps of how the AI analyzes its inputs**
  - Some guessed it could only learn visual patterns derived from basic visual elements ("Maybe light and dark? Maybe colors? Maybe shapes, lines?")
  - "Does it take into account the relationship between gland and stroma? Nuclear relationship?"
- **Specify where the algorithm received its source of ground truth.**
  - Participants asked whether the algorithm had learned from diagnoses made by general pathologists, GU pathologists, or an entire panel…
  - A few participants asked if the AI was based on an even more objective source of truth than GU pathologists, such as patient prognosis or immunostatins.

Calibration / “Point-of-View”

- **Demonstrate the subjective thresholds of the model using borderline cases.**
  - “I know what my friend... Will call… what would AI call it?... I’m treating it as a peer.”
- **Include a human-AI calibration phase.**
  - Pathologists envisioned assembling a set of cases with ground truth and comparing their diagnoses and the AI’s diagnoses with the ground truth in a calibration phase.
  - Work we’ve done in this area “Teaching Humans When To Defer to a Classifier via Exemplars” Mozannar et al., AAAI 2022 [1]

[1]:https://arxiv.org/abs/2111.11297

Calibration / "Point-of-View": Human-AI calibration phase



- User study on question answering task showed that teaching was successful 50% of the time and provided 10% improvement when effective

Calibration / “Point-of-View”

- **Make explicit the AI’s intended utility over the status quo**
- **Make transparent how the AI accounts for differential costs of errors**

[1]:https://arxiv.org/abs/2111.11297

Accuracy and Performance

- **Define accuracy precisely.**
  - Although participants were told that the Assistant predicts Gleason grades, many assumed that accuracy referred to the binary classification of benign versus cancer.
- **Provide human-relatable benchmarks for performance metrics**
  - Many were not sure what should constitute a reasonable performance threshold
- **Report AI performance on sub-categories of known human pitfalls**
  - “Maybe it has really good accuracy except for perineural invasion. If you see perineural invasion… Don’t fall for that.”

Enable this with Model Cards

Face Detection

Model Card v0 Cloud Vision API

Overview

Limitations

Trade-offs

Performance

Test your own images

Provide feedback

Explore

Object Detection

About Model Cards

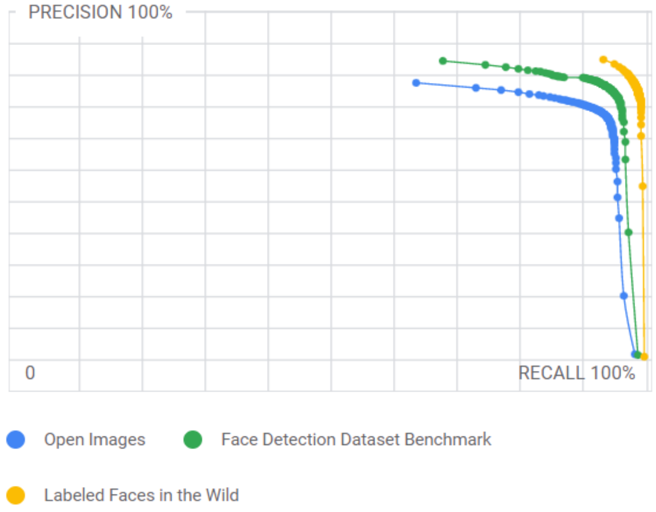
Performance

Here you can dig into the model's performance on a selection of evaluation datasets drawn from different data sources than the training data. You can assess model performance across variables such as face size and facial orientation, as well as human-perceived skin tone, gender presentation, and age. Annotations for demographic variables were made by humans and used purely for testing; the model cannot detect them.

SUMMARY

- Area under the P-R curve (PR-AUC) is 0.84 (Open Images subset), 0.92 (Face Detection Dataset and Benchmark), and 0.94 (Labeled Faces in the Wild).
- Face size, facial orientation, and degree of occlusion all have a significant impact on model performance, with the model performing least well on faces that appear large (>25% of the image area), are looking to the left or right, and/or obstructed in some way.
- Disparities in recall are relatively small (< 3% gap) for all human-annotated demographic variables evaluated (perceived skin tone, gender presentation, age).

P-R CURVES



https://modelcards.withgoogle.com/face-detection and https://huggingface.co/blog/model-cards

What can happen if people have inaccurate mental models?

Outline

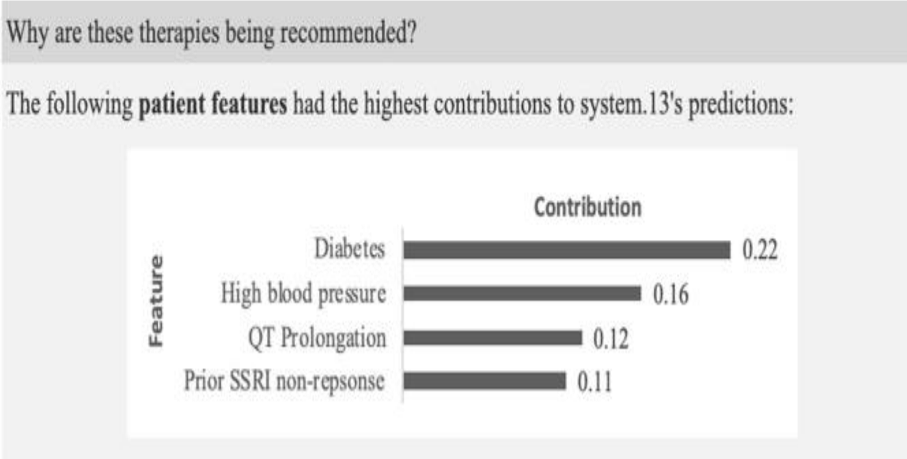
- Modes of Human-AI Interaction
- Mental Models
- Onboarding
- **Over-reliance and under-reliance on AI**

Over-reliance on AI

- Suppose the clinician was told the AI assistant sometimes performs better than humans
- There is an incentive to rely on the AI, however, we often observe over-reliance on the AI:
  - **Over-reliance = using incorrect AI recommendations**
- One contributing reason is misleading explanations – among those are things like Lime and saliency maps

Over-reliance on AI: Explanations

- In a study for recommending antidepressants [1], participants performance was worse with explanations (observed elsewhere)



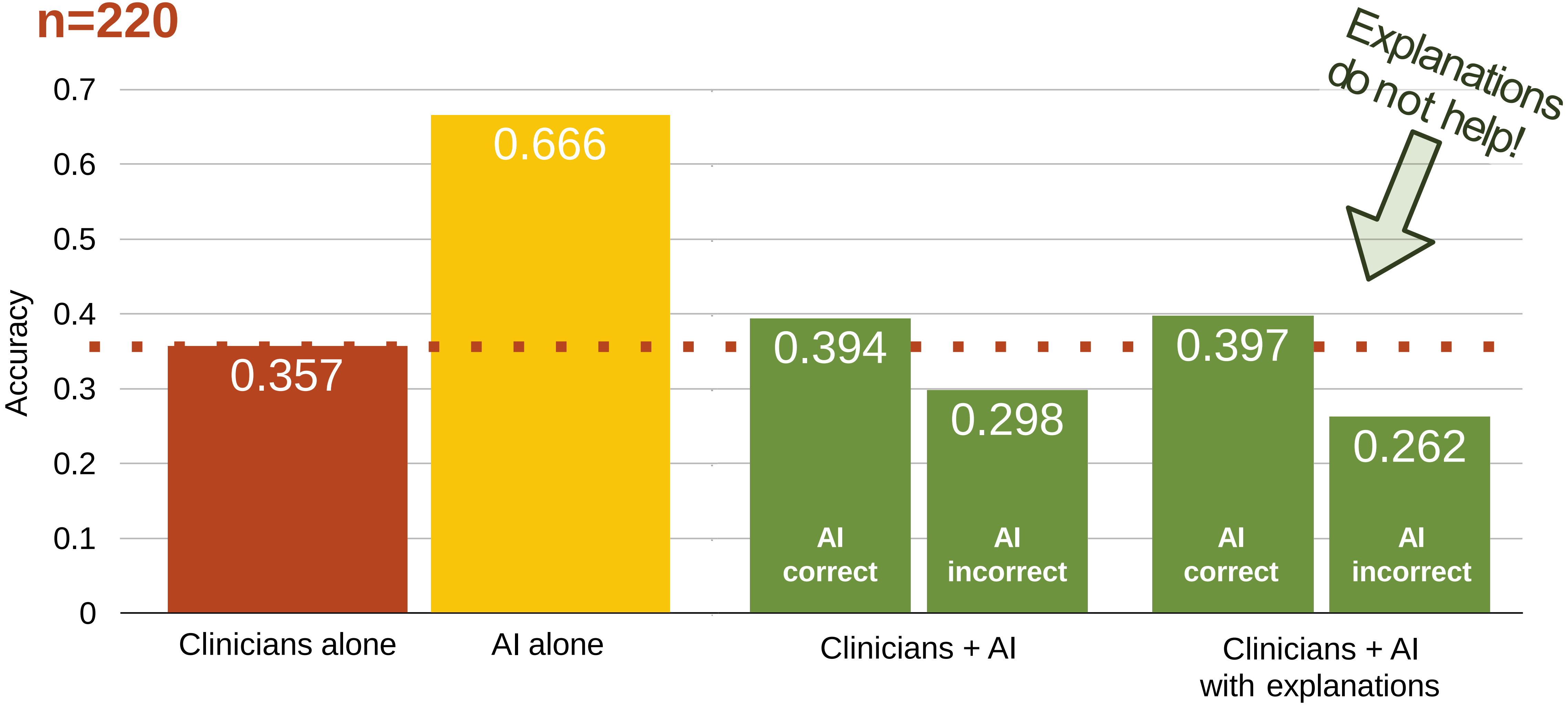
- When AI predicted incorrectly:

Type	No AI	Prediction only	Prediction + Explanation
Accuracy on correct AI	0.357	0.394	0.397
Accuracy on incorrect AI	0.357	0.298	**0.262**

[1]:Jacobs, Maia, et al. "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection." Translational psychiatry 11.1 (2021): 1-9.

AI-Assisted Antidepressant Selection

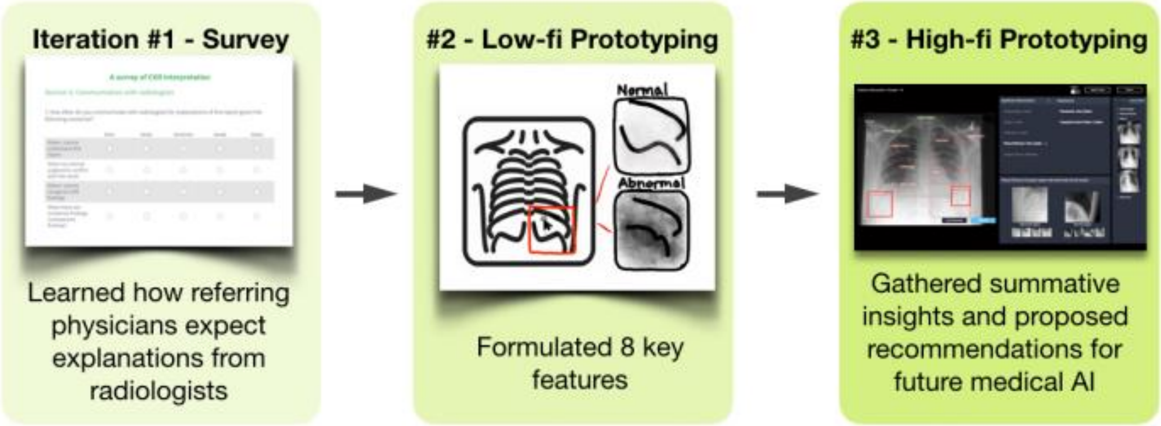
n=220



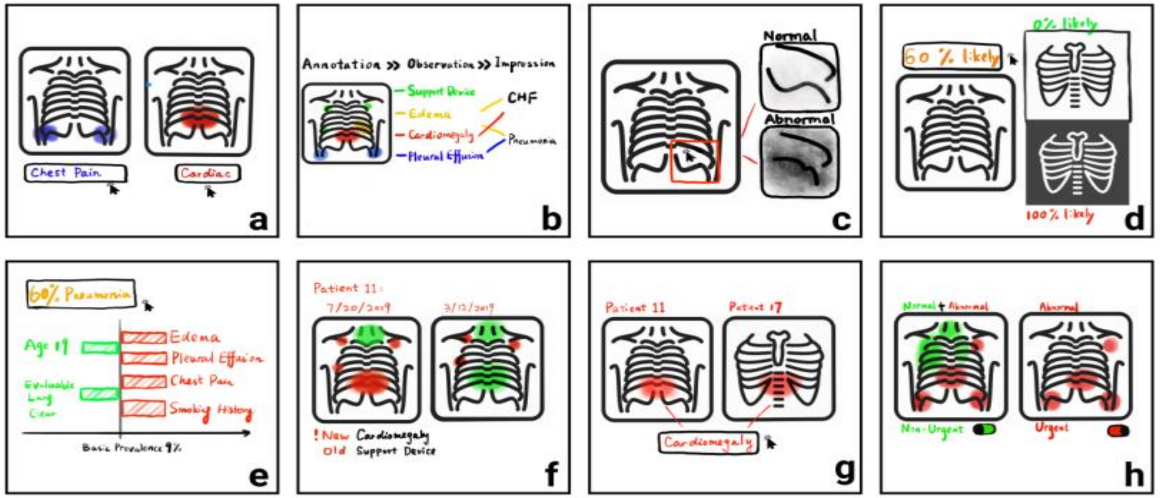
Jacobs, et al. **How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection**. *Translational Psychiatry*, 11, 2021.

Design Explanations (and UI) with feedback from Clinicians

- CheXplain [1]: asking when and what kind of explanations are needed

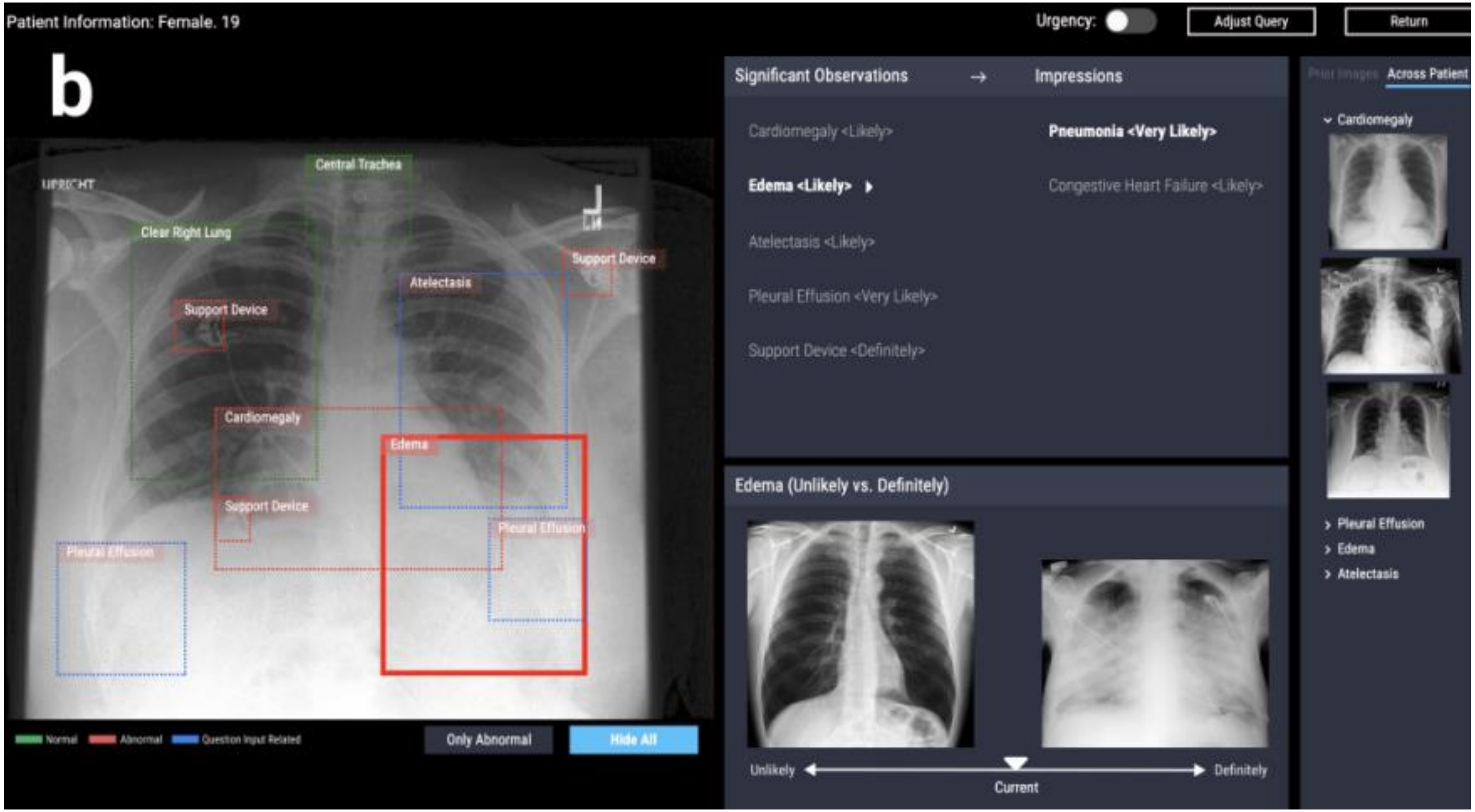


- Designing sketches: 1) allow for questions, 2) hierarchical explanations 3) contrastive examples, 4) probabilities, 6) across time



[1]:Xie, Yao, et al. "CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.

Design Explanations (and UI) with feedback from Clinicians



Saliency maps are not enough

- There is a growing body of evidence that shows that are insufficient form of explanation (to say they don't add more than a confidence score)

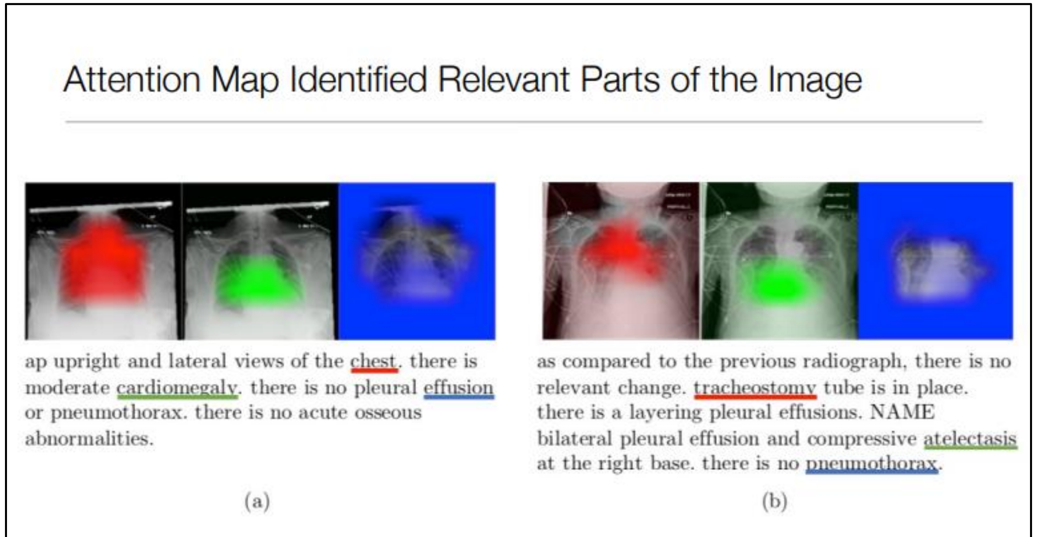


Figure 3: **Visualization of the generated report and image attention maps.** Different words are underlined with its corresponding attention map shown in the same color.

[1]:Arun, Nishanth, et al. "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging." Radiology: Artificial Intelligence 3.6 (2021): e200267.

Under-reliance

- Setting: Clinical decision support tools that gives alerts in electronic medical record

Alert type	Total alerts		Alert overrides		
Patient alle[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Drug–drug [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Duplicate [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Drug–class interaction	19 593	12.4	4782	24.4	Transitioning from one drug to the other
Class–clas[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]on long term therapy with combination
Age-based suggestion	10 501	6.7	8297	79.0	Patient has tolerated this drug in the past
Renal suggestion	3890	2.5	3035	78.0	Patient has tolerated this drug in the past
Formulary substitution	15 945	10.1	13 554	85.0	Intolerance/failure of suggested substitution
Total	157 483	100.0	82 899	$\boxed{52.6}$	

Half of alerts were overridden (other studies estimate 90% override)

Half of overrides were appropriate (estimated)

Cause can be alert fatigue

Alert type	Override appropriate
[illegible]	[illegible]
Drug–drug interaction†	12
Duplicate drug‡	82
Drug–class interaction‡	88
Class–class interaction‡	69
Age-based suggestion†	39
Renal suggestion†	12
Formulary substitution†	57
Average	53

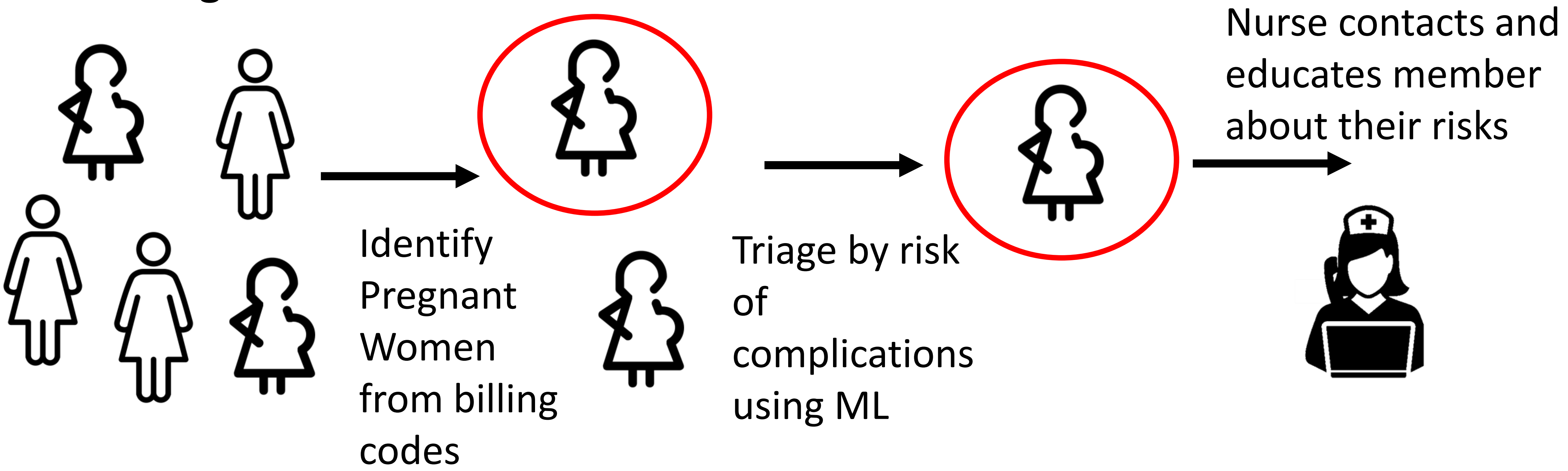
[1]:Nanji, Karen C., et al. "Overrides of medication-related clinical decision support alerts in outpatients." Journal of the American Medical Informatics Association 21.3 (2014): 487-491.

Under-reliance fixes

1. Make it easy to dismiss the CDS when needed
2. When override dismissed, let the system know why
3. Personalize the alerts by the attending physician and allow for alert rate to change depending on override rates
4. Update model given corrections by user
5. Inform user about model updates to allow their mental model to also update

Human-Centered Design Methodology

- **Case study:** algorithmic support for high-risk pregnancy care management team



https://online.hbs.edu/blog/post/what-is-human-centered-design

Human-Centered Design Methodology

1) Needs Assessment
   - Interviews about their needs
   - Mockup calls of nurses with members
   - Shadowing nurse process

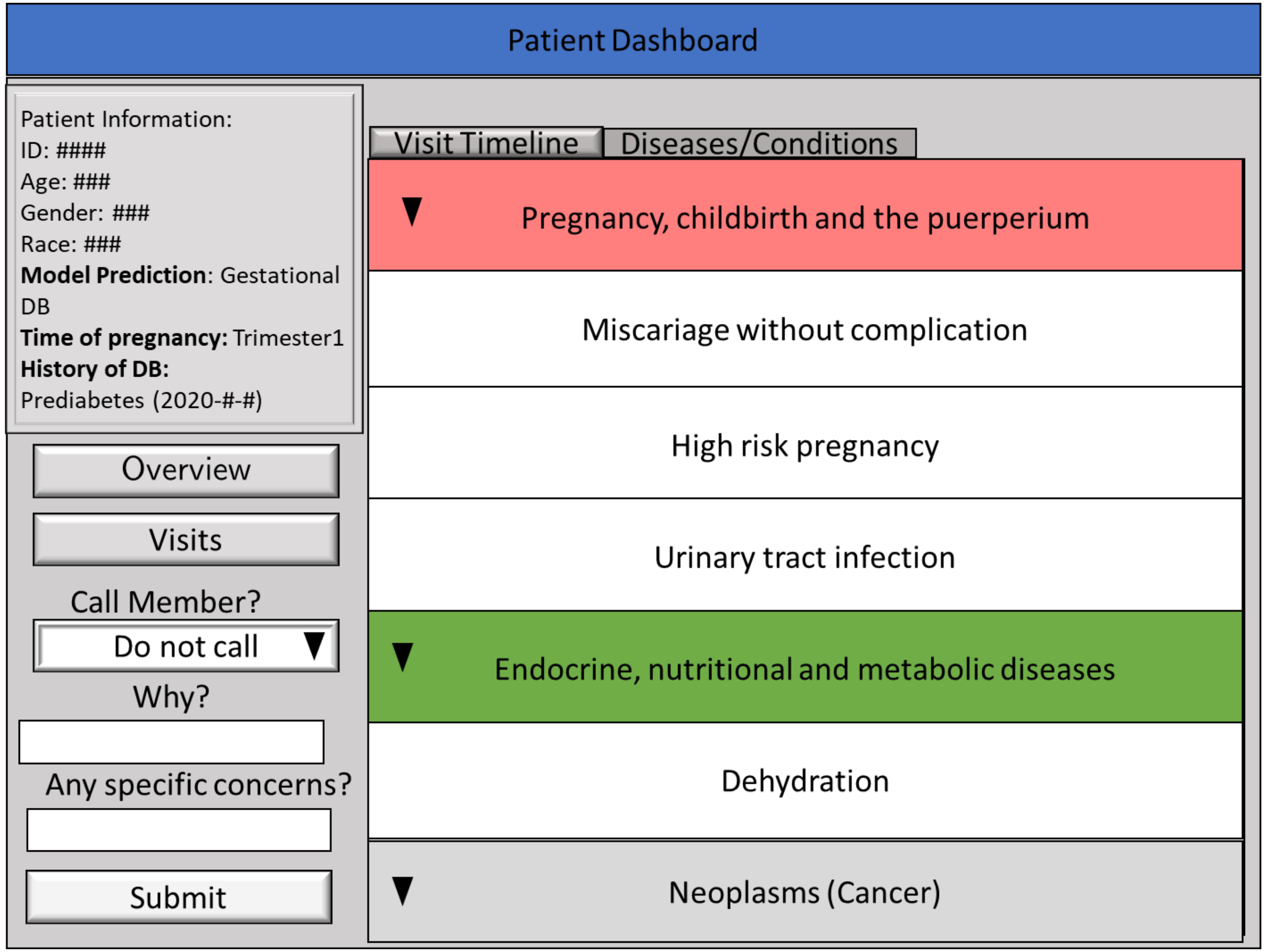
   -> members often surfaced after they're pregnant, members risk determination is not calibrated, no explanation surfaced for risk

- 2) Ideate
  - Build Algorithm to predict pregnancy, improve risk calibration and provide explanations
- 3) Implement & Evaluate Using Retrospective Data
- 4) Test (then go back to step 1) – Using User Studies in-situ

https://online.hbs.edu/blog/post/what-is-human-centered-design

Human-Centered Design Methodology

- Iterative design of user interface after pilot studies
- Explanations Integrated into dashboard with colors
- Final user studies confirm nurses prefer new interface over status quo and can make risk predictions faster (~20s) with same accuracy



Guidelines for Human AI Interaction

Learn more: https://aka.ms/aiguidelines

INITIALLY

1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

DURING INTERACTION

3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.

WHEN WRONG

7. Support efficient invocation.
8. Support efficient dismissal.
9. Support efficient correction.
10. Scope services when in doubt.
11. Make clear why the system did what it did.

OVER TIME

12. Remember recent interactions.
13. Learn from user behavior.
14. Update and adapt cautiously.
15. Encourage granular feedback.
16. Convey the consequences of user actions.
17. Provide global controls.
18. Notify users about changes.

Takeaways

- Figure out what mode of Human-AI interaction is appropriate for your problem
- Human's mental model of the AI determines the success of the system
- Design onboarding stages to allow the human to form an accurate mental model of the AI

Takeaways

- Design AI and AI explanations with human in mind to avoid over-reliance
- Allow for updates over time to interface and model to avoid under-reliance