Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Big Data
National Research Programme

Granger-causal Attentive Mixtures of Experts

Learning Important Features with Neural Networks

Patrick Schwab[1]

@schwabpa

Djordje Miladinovic[2] and Walter Karlen[1]

[1]Institute of Robotics and Intelligent Systems, ETH Zurich
[2] Department of Computer Science, ETH Zurich

Motivation

Motivation

Motivation

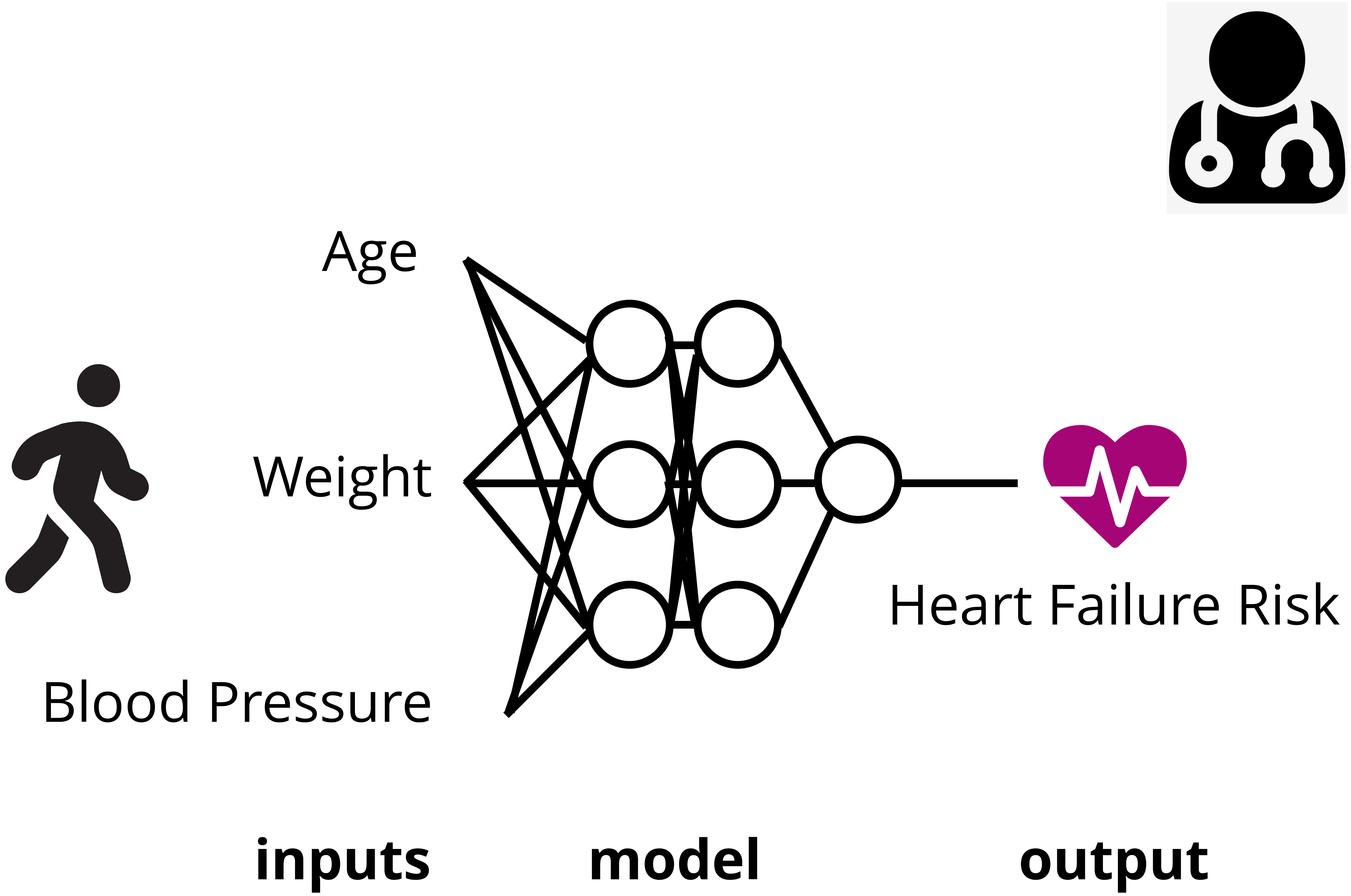
Age

Weight

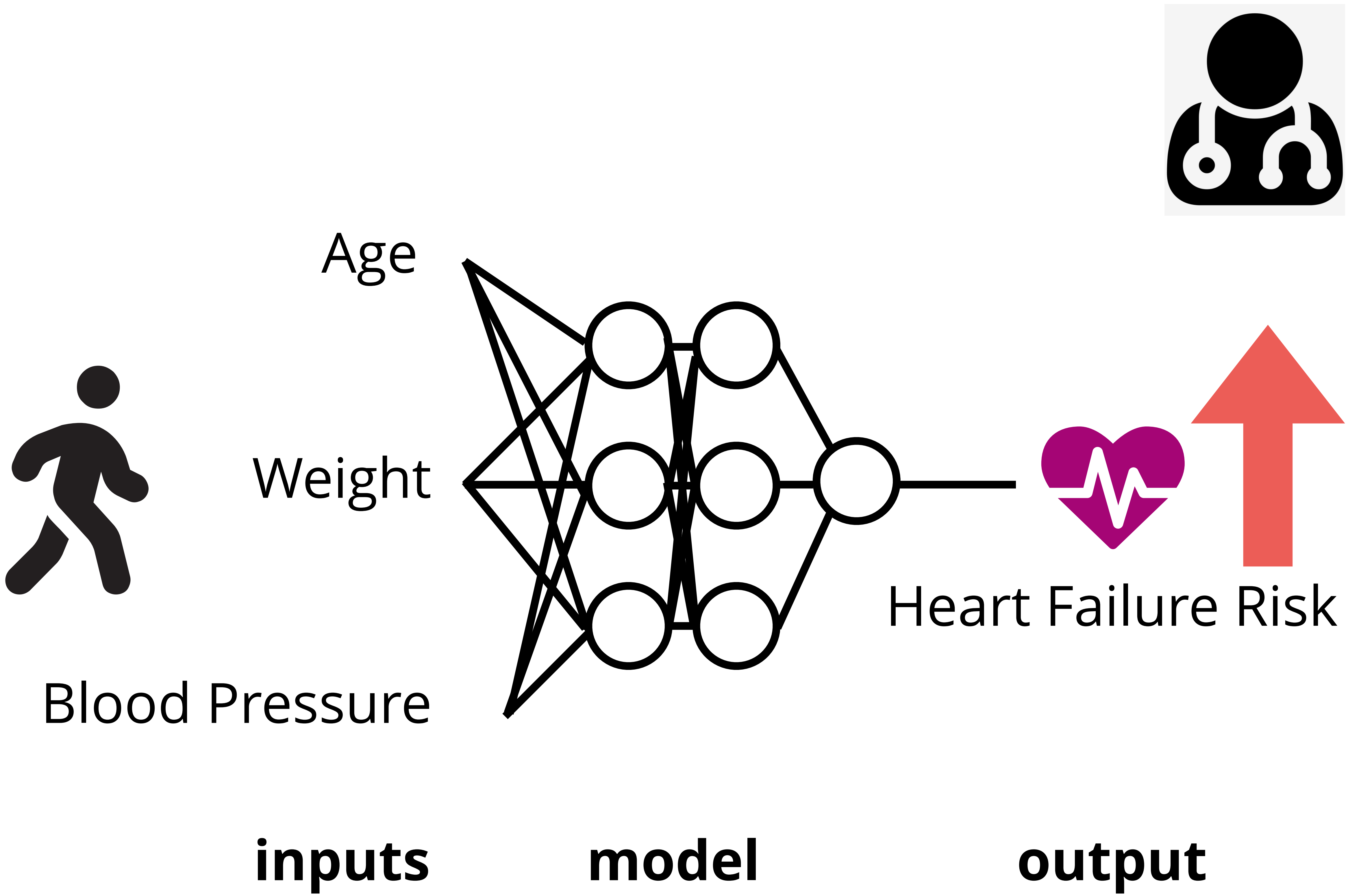
Blood Pressure

inputs

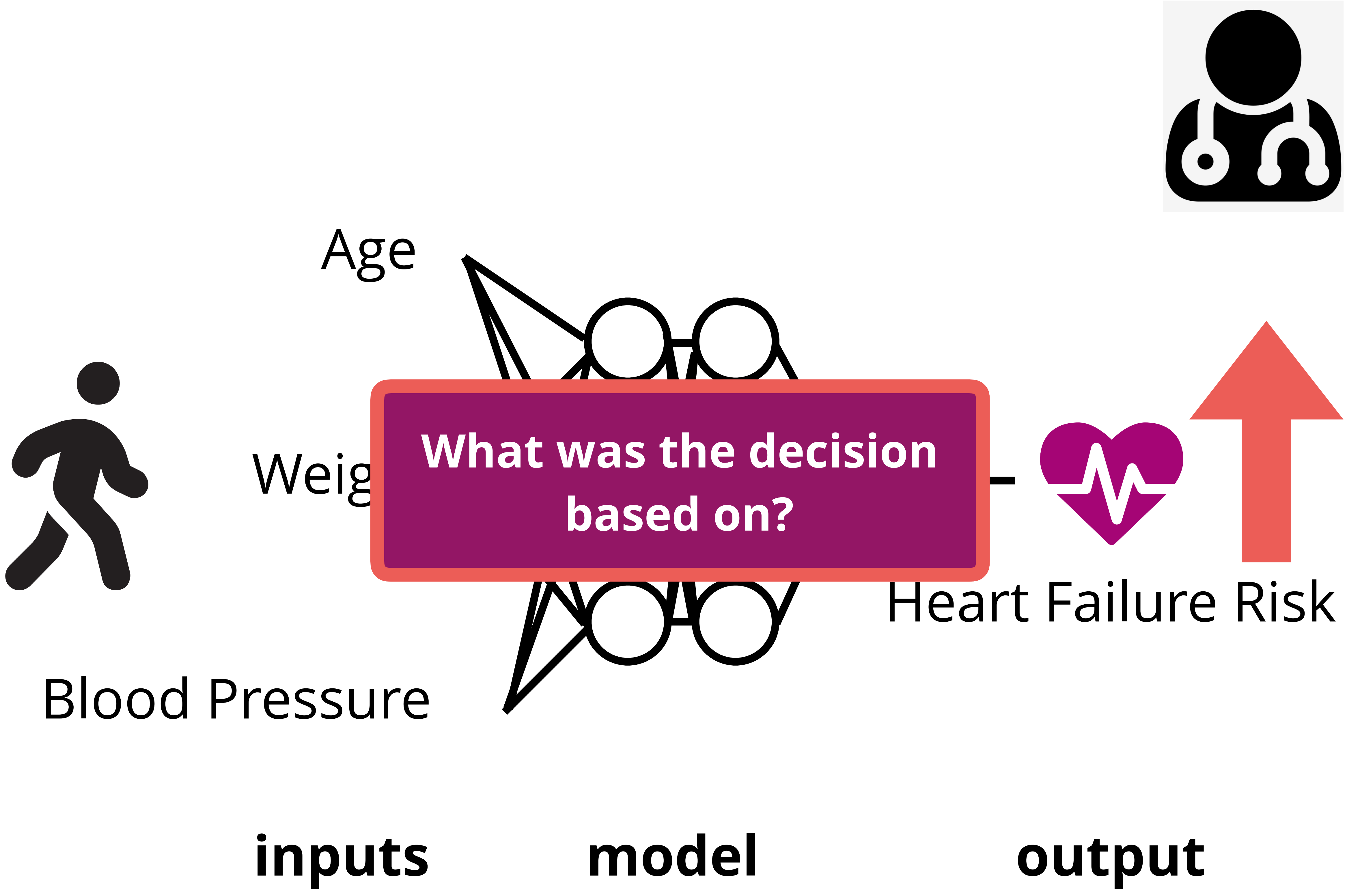
Motivation


Age
Weight
Blood Pressure
Heart Failure Risk
inputs
model
output

Motivation

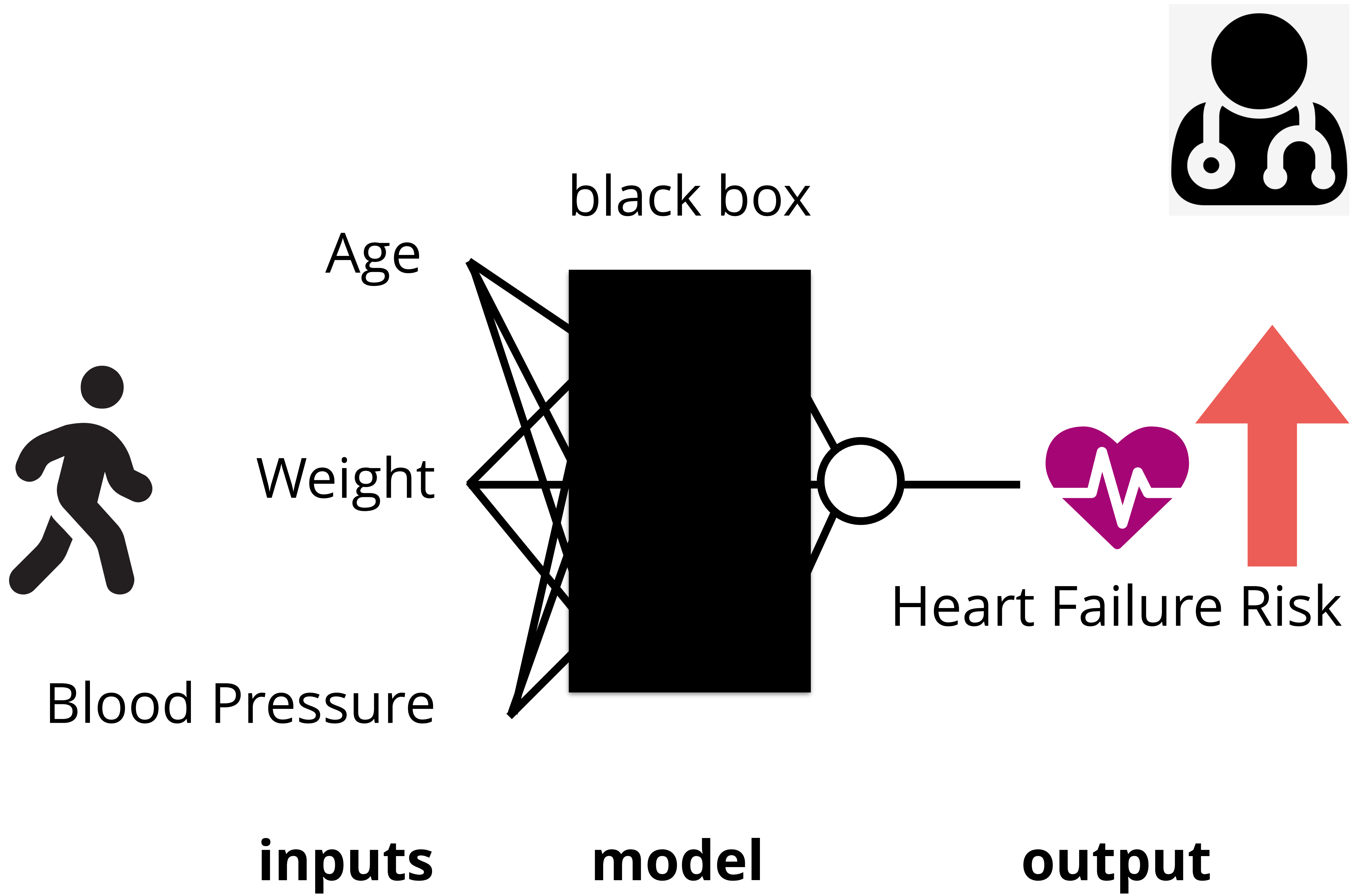

Age
Weight
Blood Pressure
Heart Failure Risk
inputs
model
output

Motivation

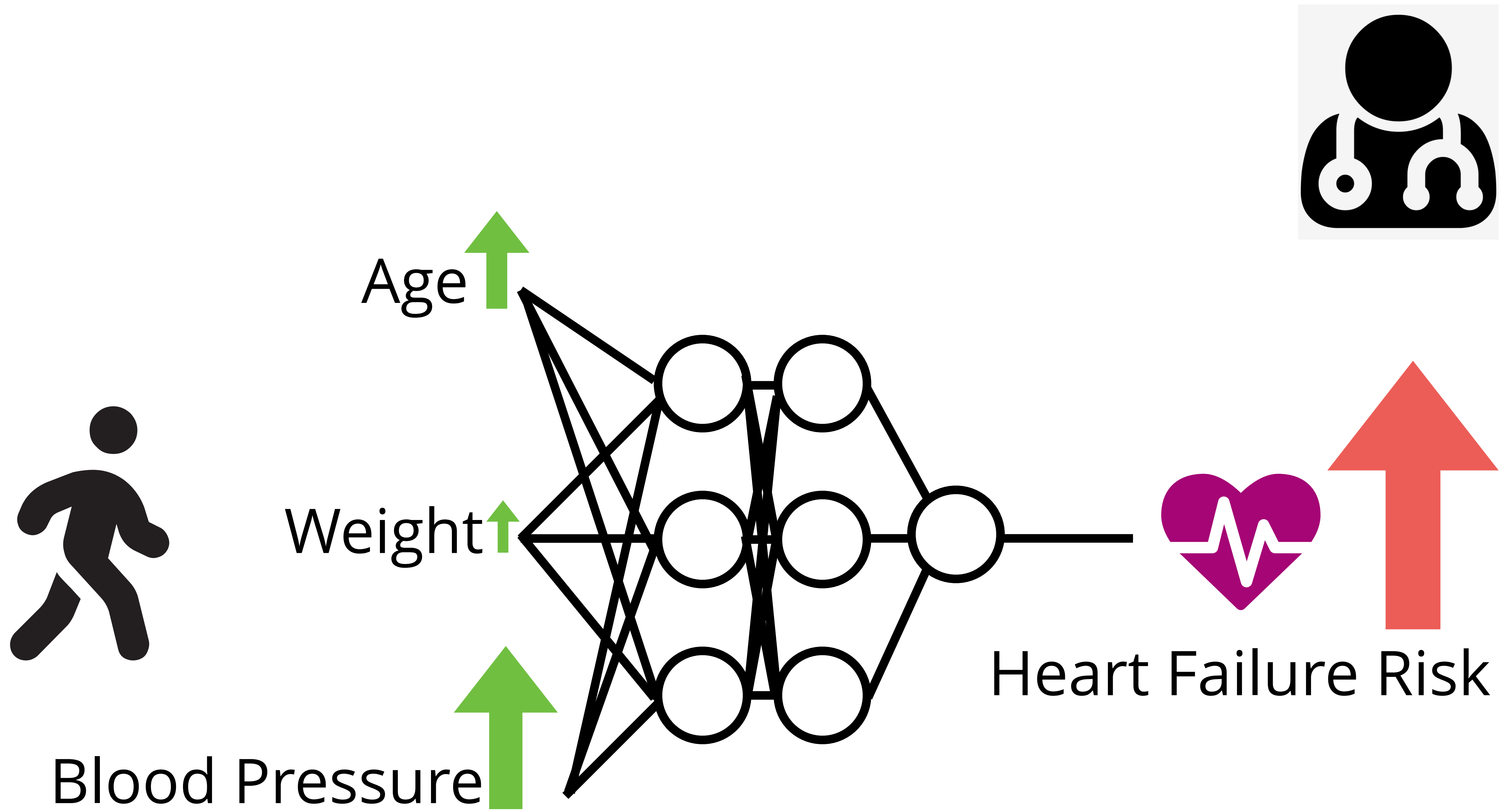


Age
Weig
What was the decision based on?
Heart Failure Risk
Blood Pressure
inputs
model
output

Motivation


black box
Age
Weight
Blood Pressure
Heart Failure Risk
inputs
model
output

Motivation



input **We desire explanation.** **output**

The Idea

Can we train a **neural network** to output both

(1) accurate **predictions**, and

(2) feature **importance scores**

?

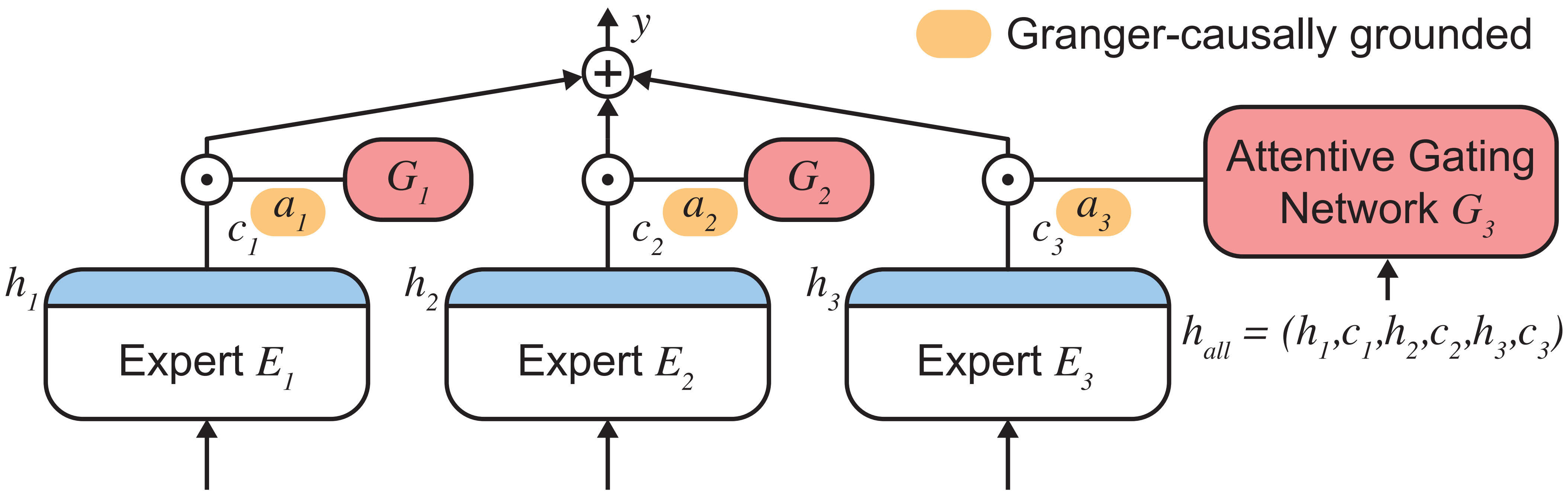
Use Cases

- **Model understanding**
  - Human-ML **cooperation - why** was this decision made?
  - Does this decision make **sense**?
  - Are my model's decisions **justifiable**?
  - What **patterns** has my model **discovered**?

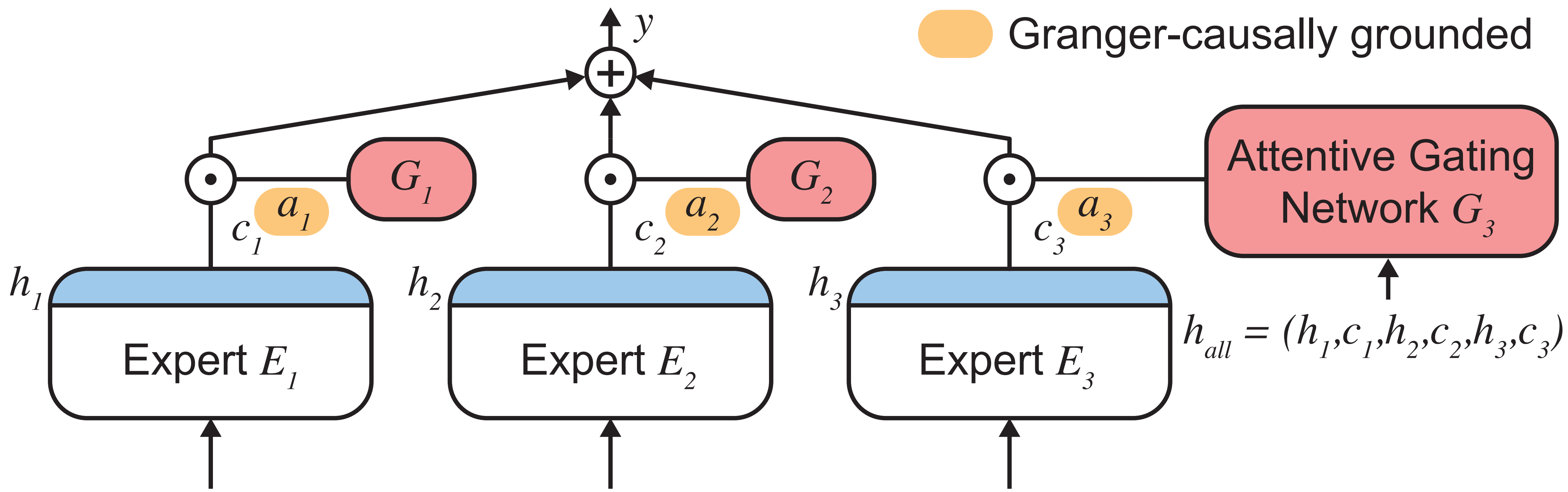
Schwab et al. Granger-causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks

Approach

Attentive Mixture of Experts (AME)



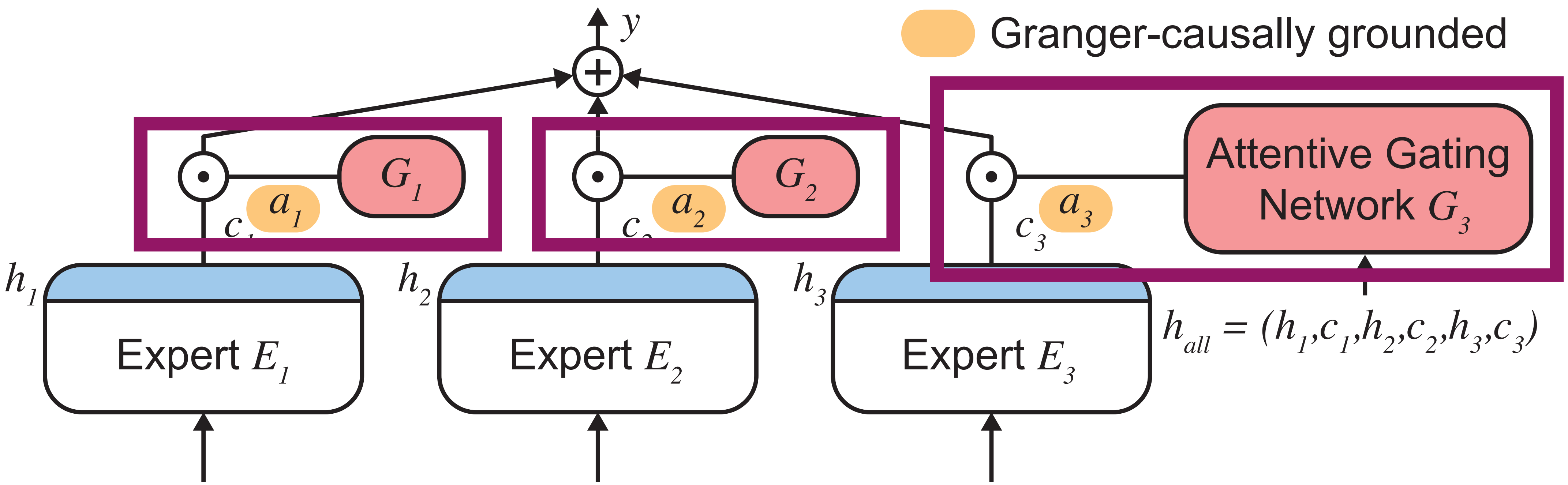
Attentive Mixture of Experts (AME)



One independent expert per feature / feature group

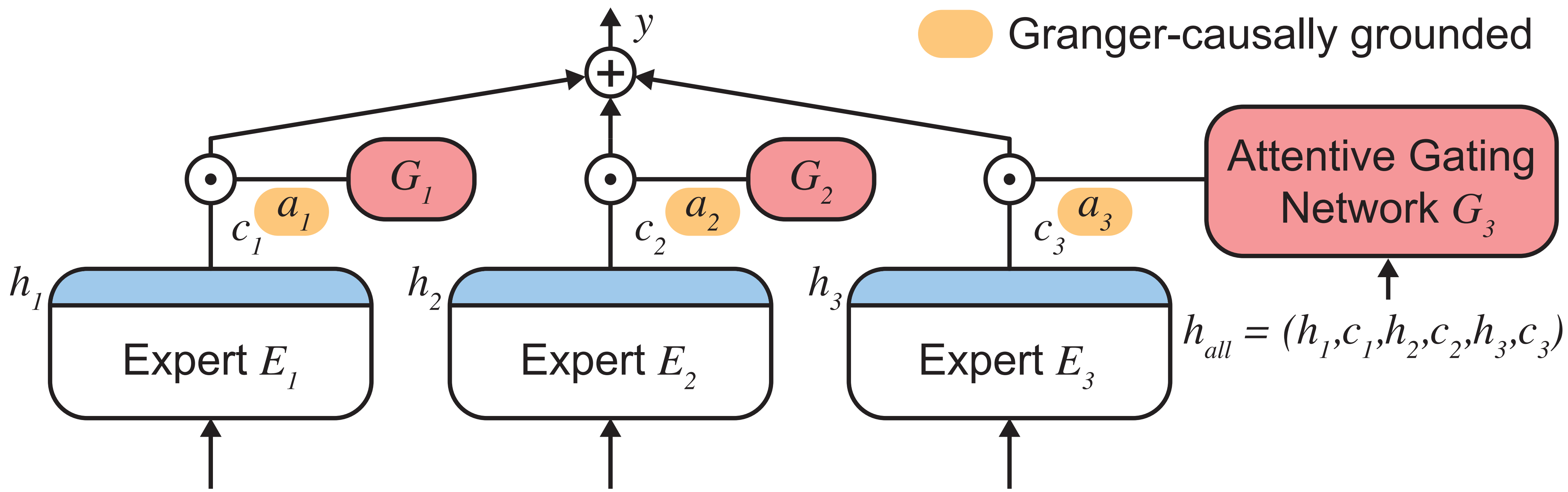
Attentive Mixture of Experts (AME)

Attentive gates control expert contributions



Attentive Mixture of Experts (AME)

Experts can only contribute to *y* after modulation by *a_i*



However, on its own this structure has the same issue as naive soft attention mechanisms:

- **No incentive** to learn to output **accurate** feature importance estimates [1].

- Often **collapses** to use only very few or a single expert early on during training [2, 3].

---

[1] Sundararajan, Taly, and Yan 2017; [2] Bengio et al. 2015; [3] Shazeer et al. 2017

Granger-causal Objective

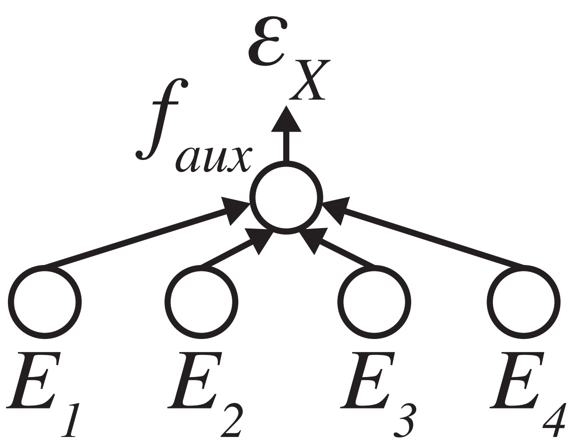
- Granger (1969) postulated **Granger-causality**
  - **declares** relationship X→Y if we are **better able to predict** Y using **all information** than if all information **apart from X** had been used*

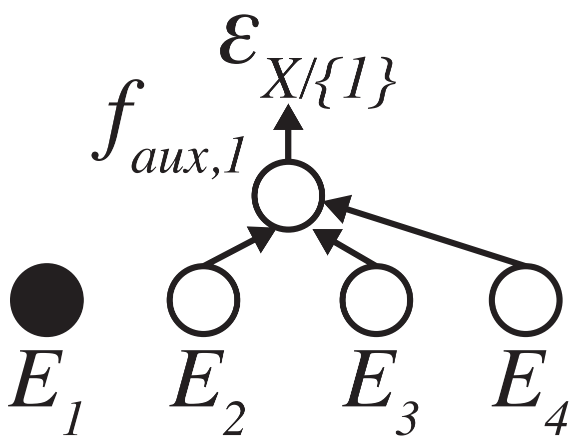
DEFINITION 1: *Causality*. If $\sigma^2(X|U) < \sigma^2(X|\overline{U - Y})$, we say that Y is causing X, denoted by $Y_t \Rightarrow X_t$. We say that Y_t is causing X_t if we are better able to predict X_t using all available information than if the information apart from Y_t had been used.

---

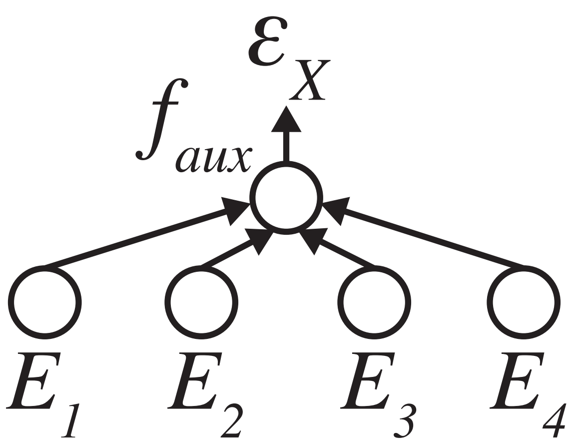
* Other assumptions apply that are not relevant in the presented setting.

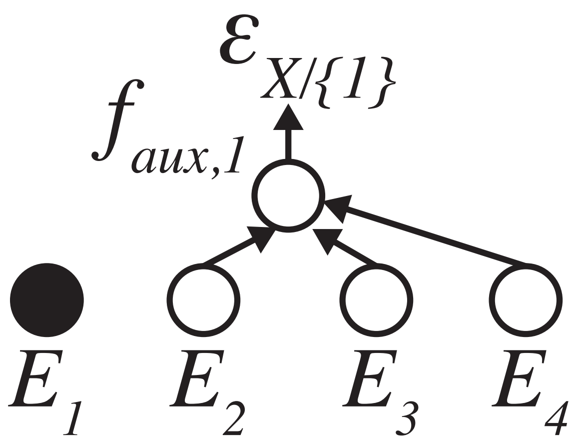
Granger-causal Objective





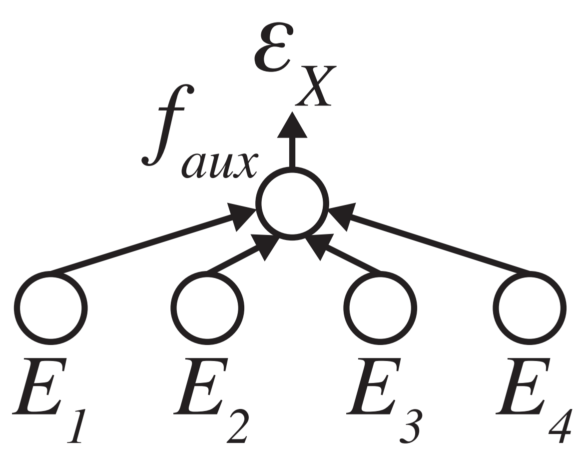
Granger-causal Objective

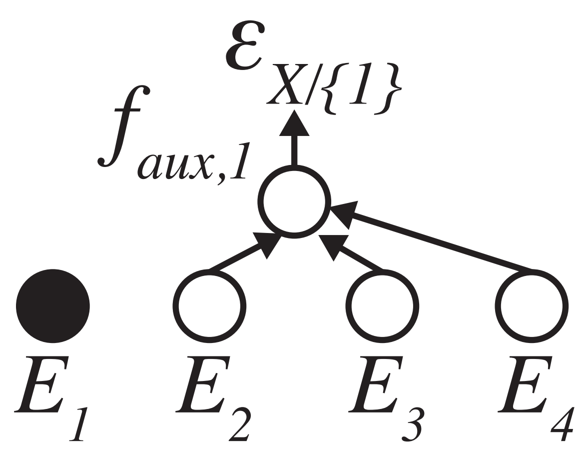




Error when considering all information

Granger-causal Objective

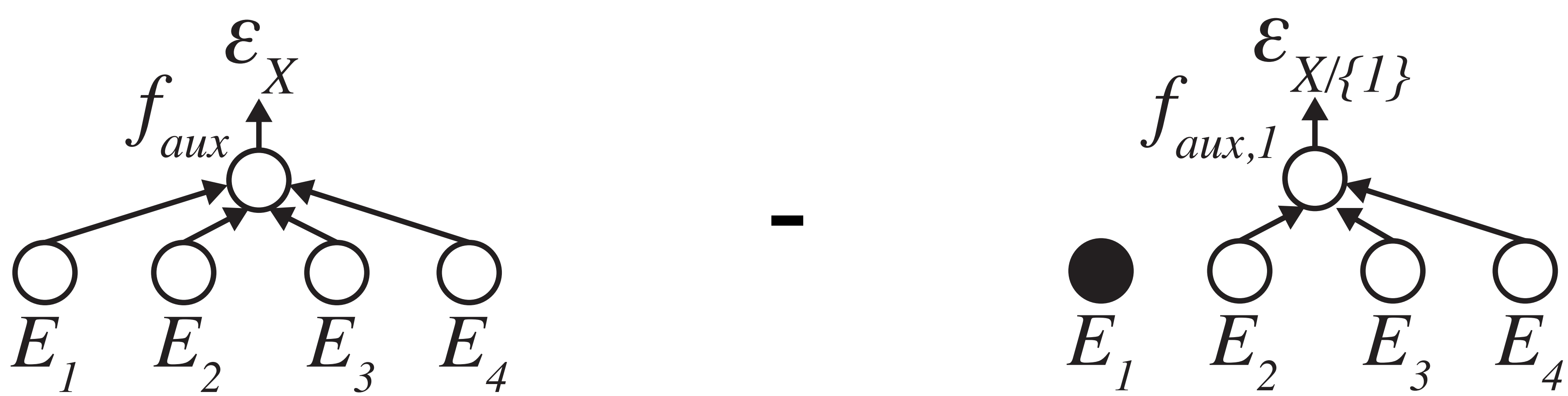




Error when considering all information

Error when considering information apart from E_1

Granger-causal Objective



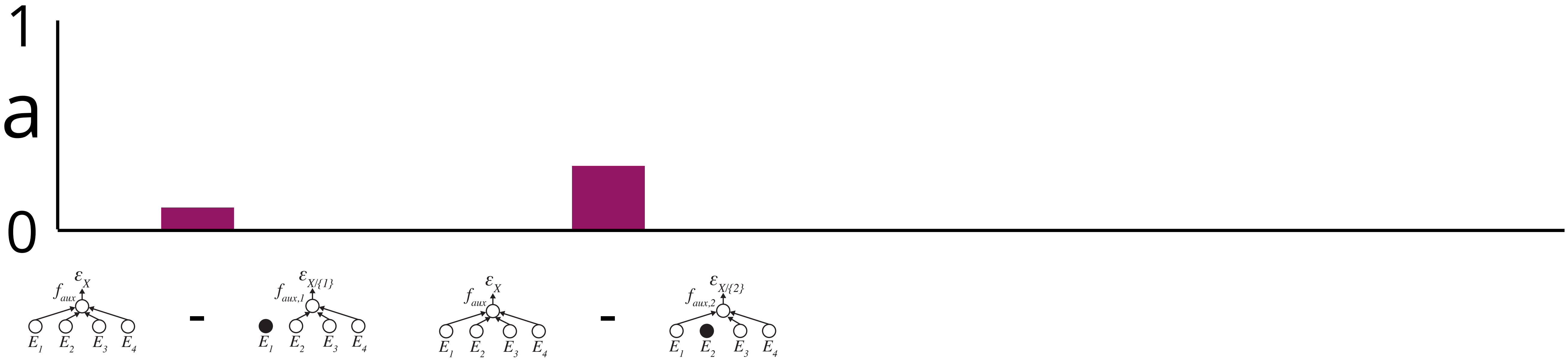
We define **feature importance** as the **reduction in prediction error** associated with adding that feature.

$$\Delta\varepsilon_{X,i} = \varepsilon_{X\setminus\{i\}} - \varepsilon_X$$

Granger-causal Objective

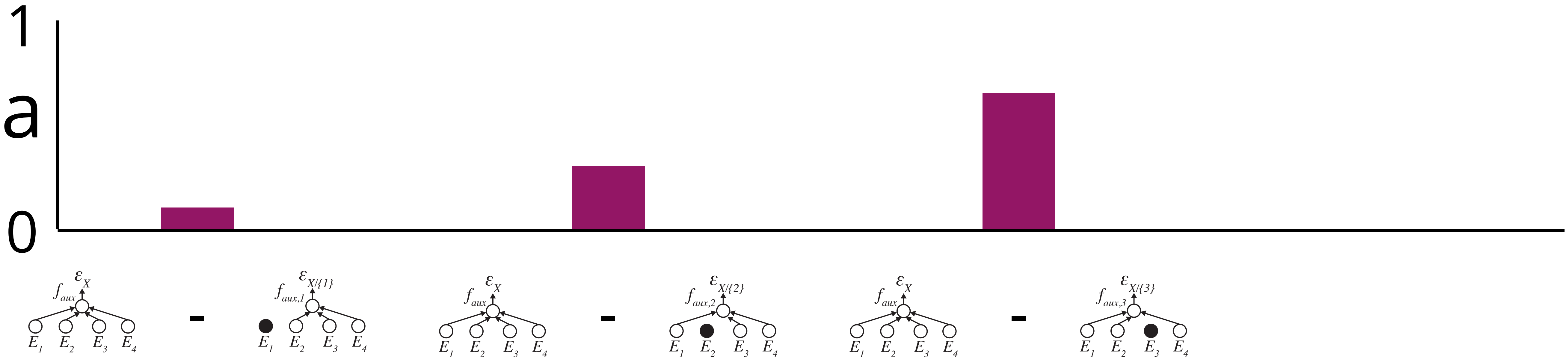
We define **feature importance** as the **reduction in prediction error** associated with adding that feature.

Granger-causal Objective



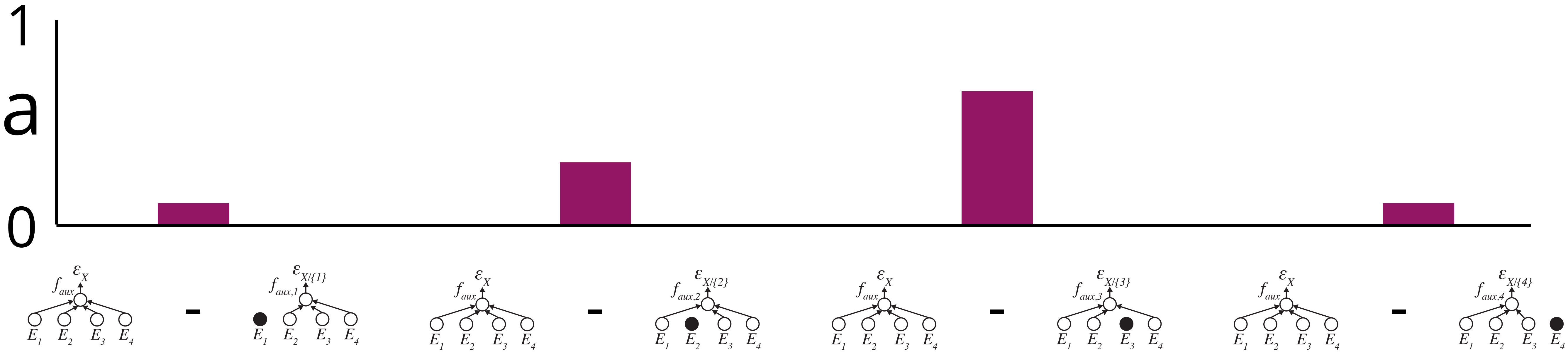
We define **feature importance** as the **reduction in prediction error** associated with adding that feature.

Granger-causal Objective



We define **feature importance** as the **reduction in prediction error** associated with adding that feature.

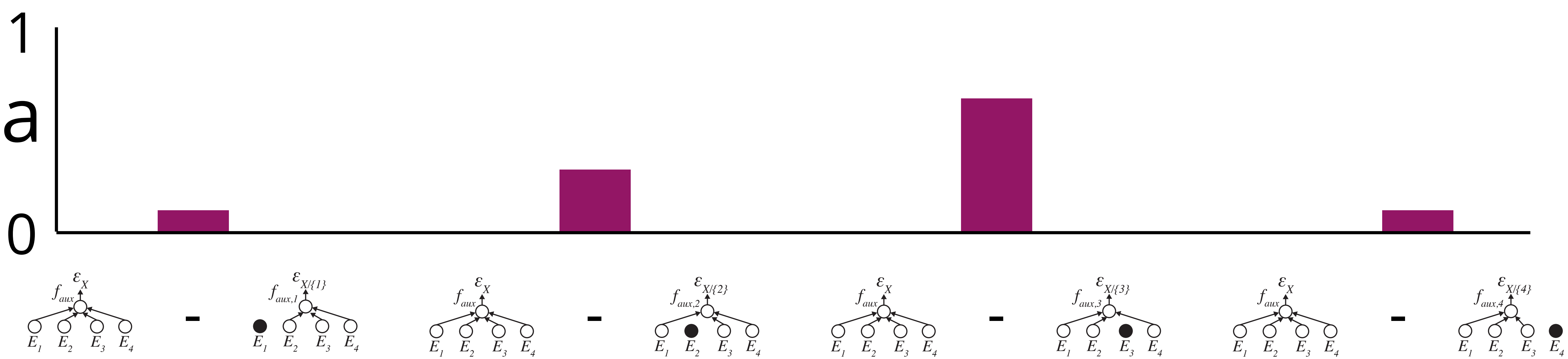
Granger-causal Objective



We define **feature importance** as the **reduction in prediction error** associated with adding that feature.

Granger-causal Objective

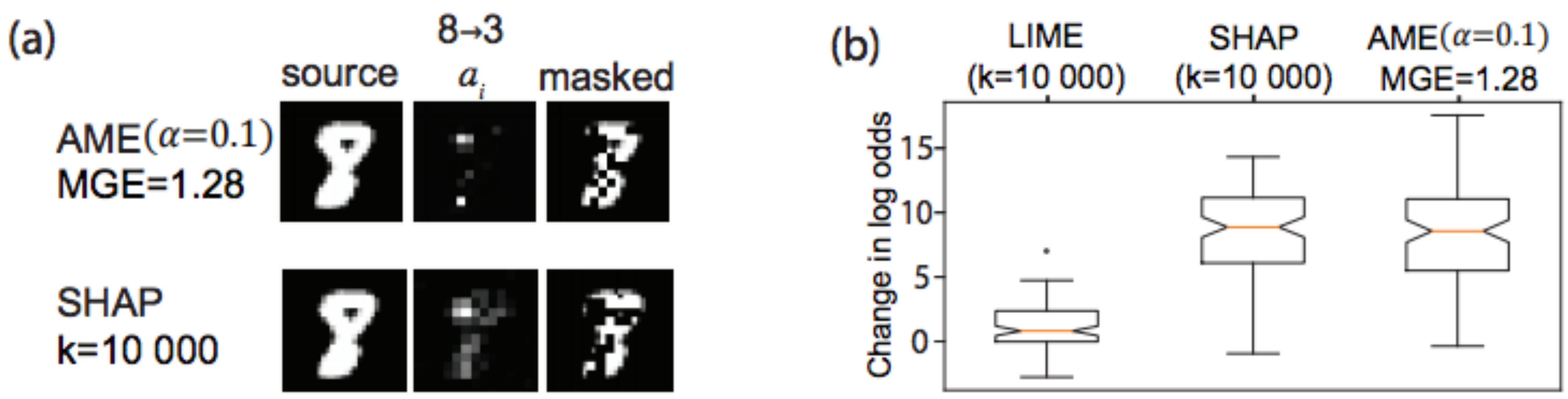
We now have a differentiable link between labels (prediction error) and feature importance.



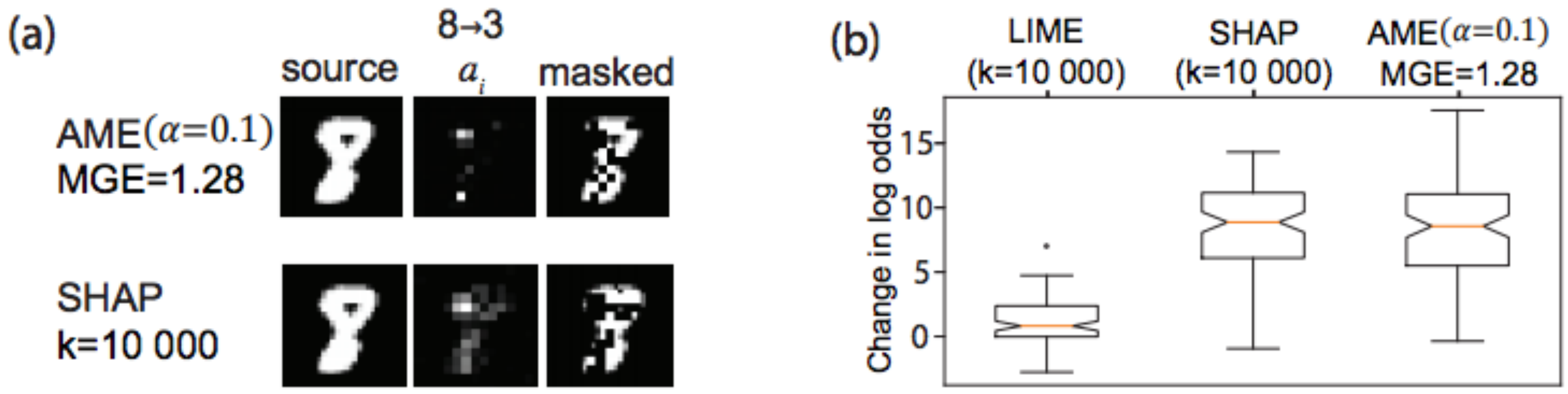
We define **feature importance** as the **reduction in prediction error** associated with adding that feature.

Evaluation

Important Features in Handwritten Digits

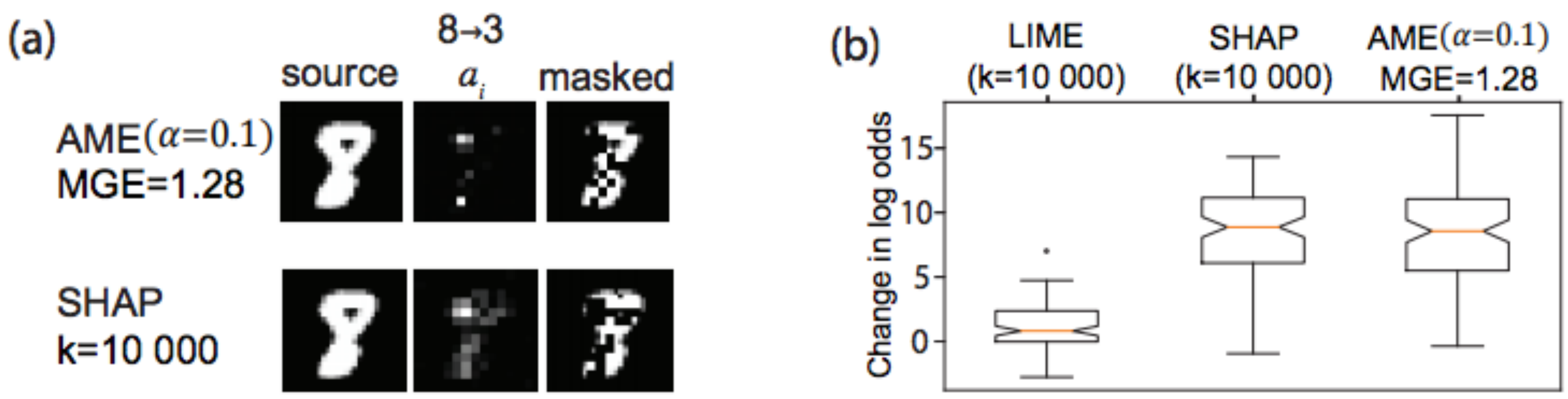


Important Features in Handwritten Digits



Estimation accuracy comparable to SHAP.

Important Features in Handwritten Digits

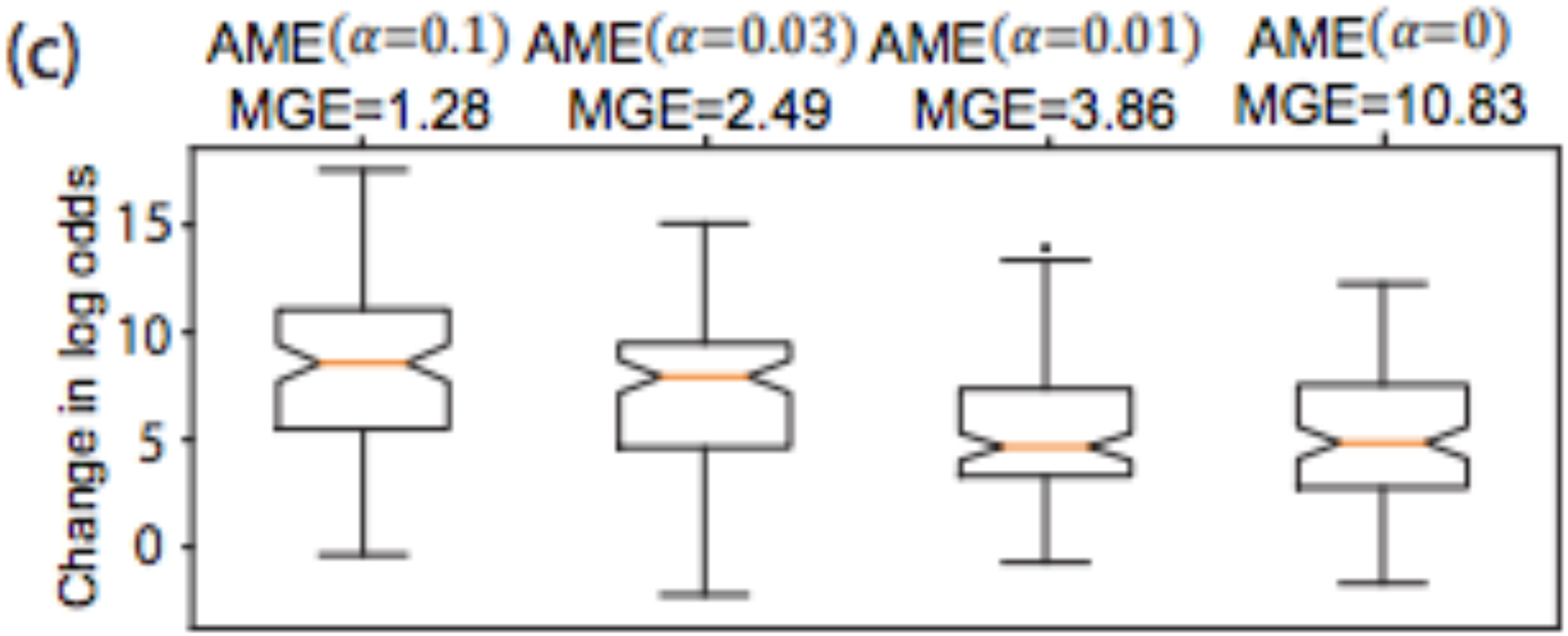


(d)

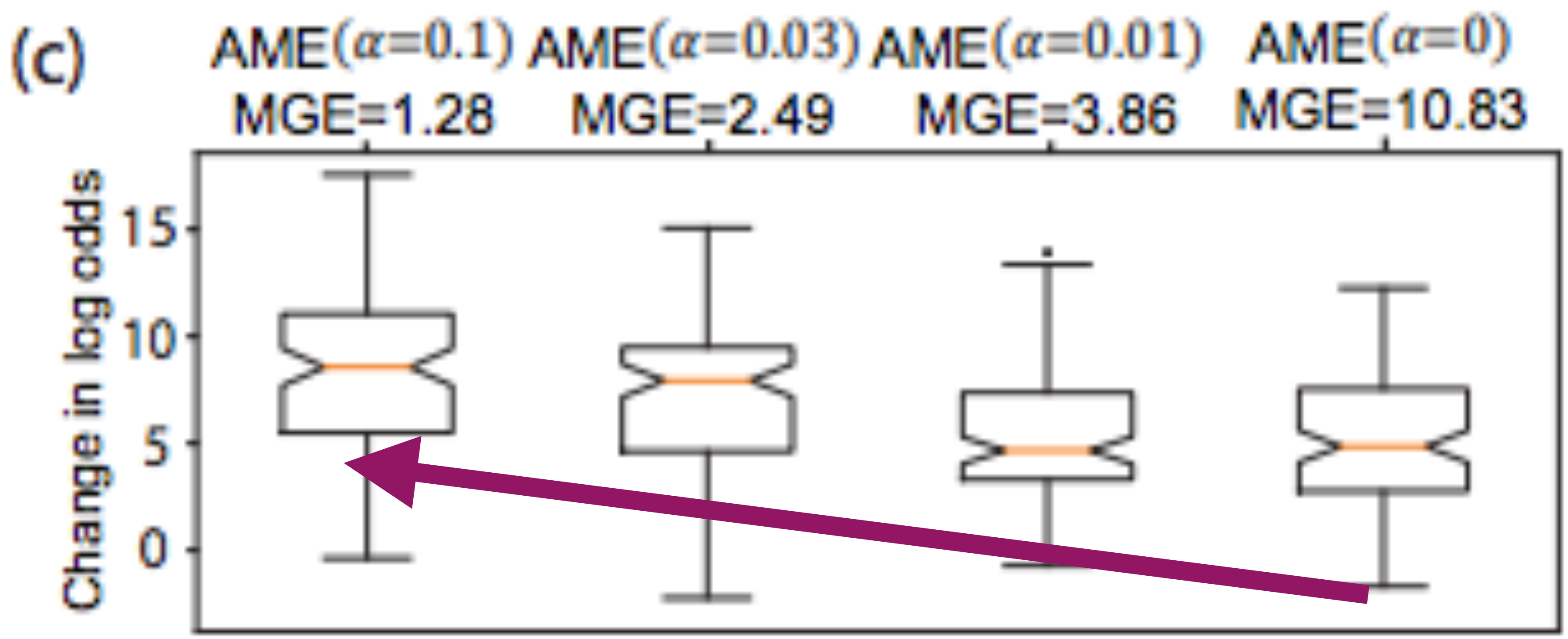
Method	CPU(s)
AME(α=0.1)	**3**
SHAP	982
LIME	2063

Orders of magnitude faster at importance estimation

Important Features in Handwritten Digits

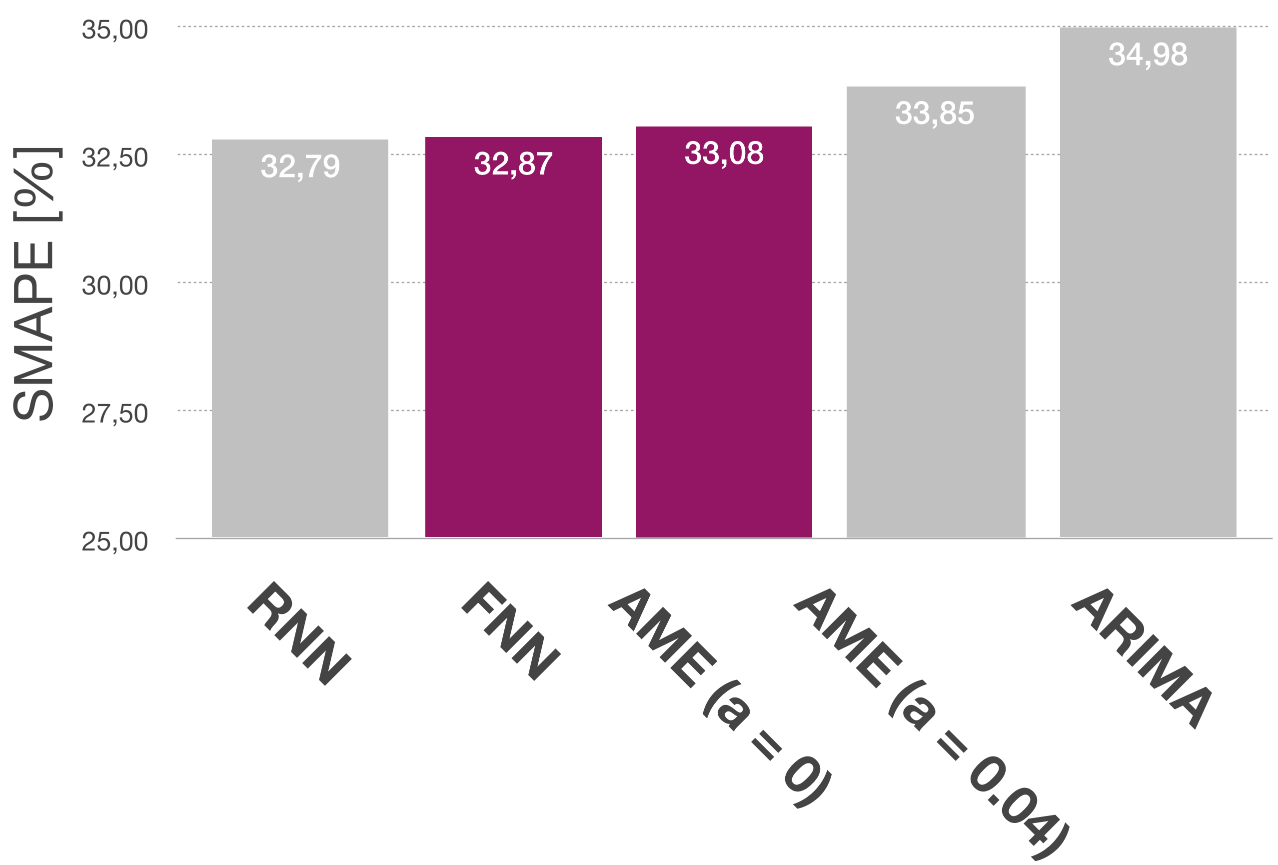


Important Features in Handwritten Digits



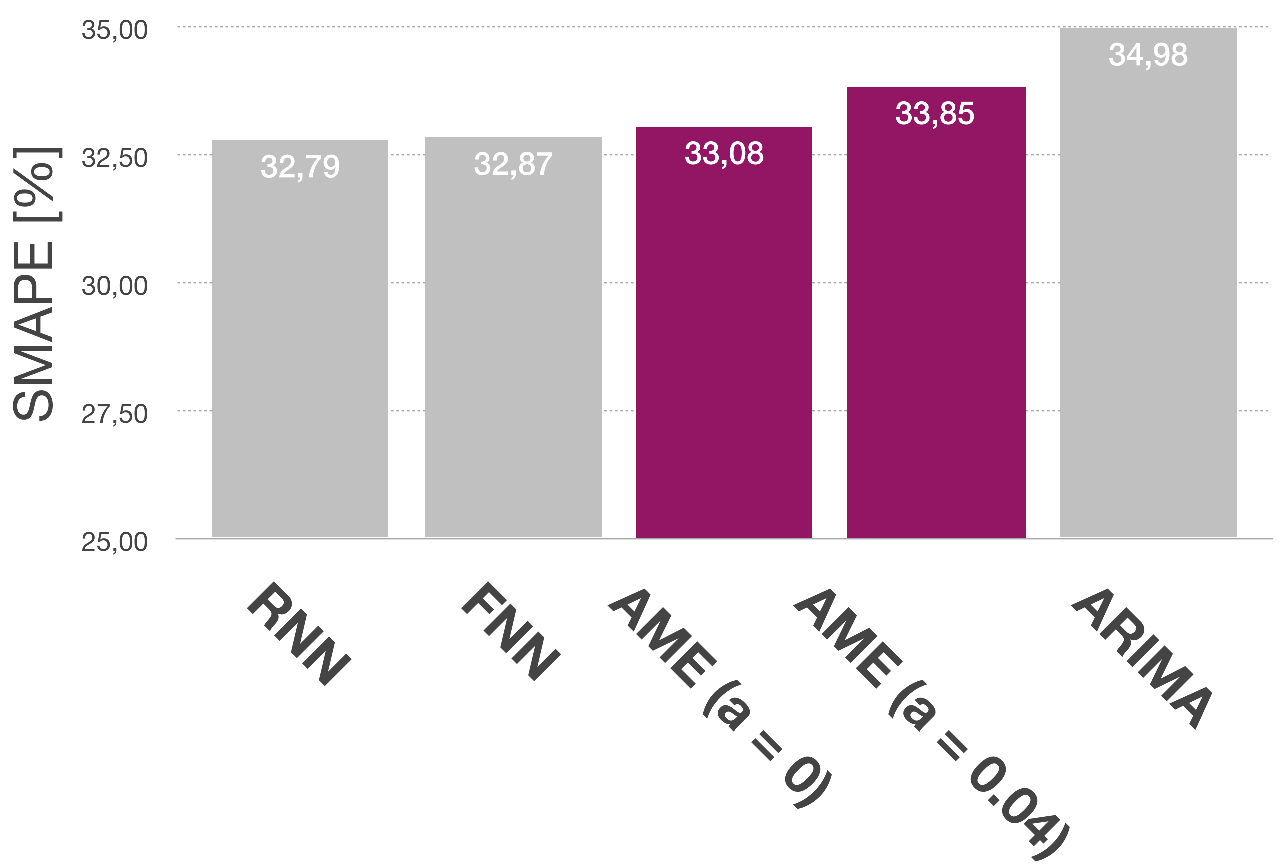
Lower MGE correlates with better feature importance estimates.

Drivers of Medical Prescription Demand



Slightly lower prediction accuracy when using AME architecture

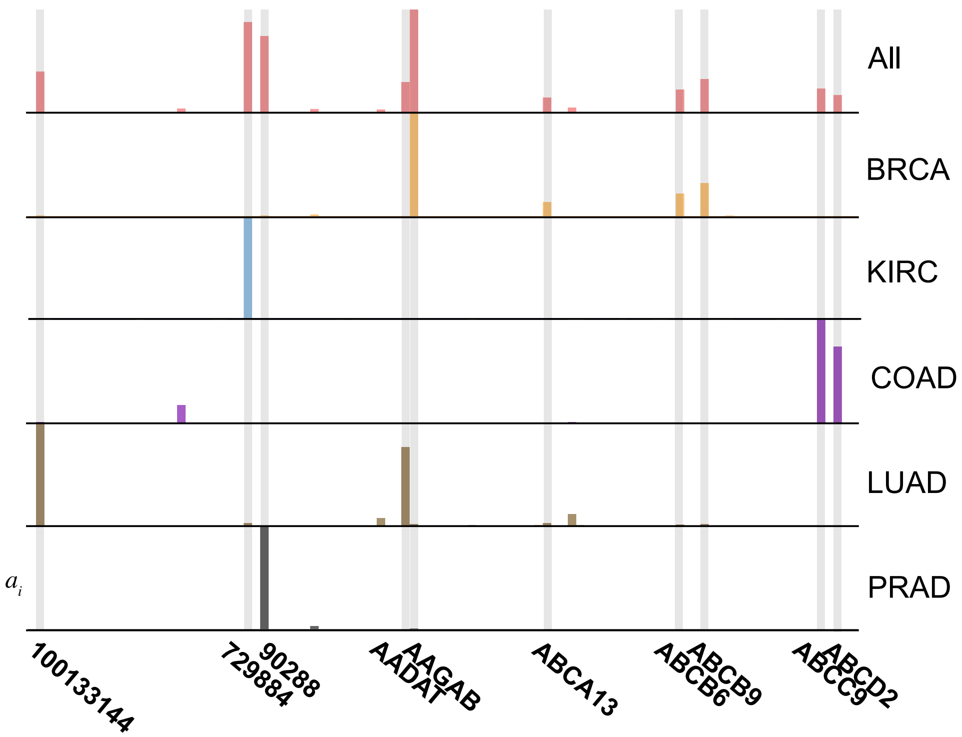
Drivers of Medical Prescription Demand



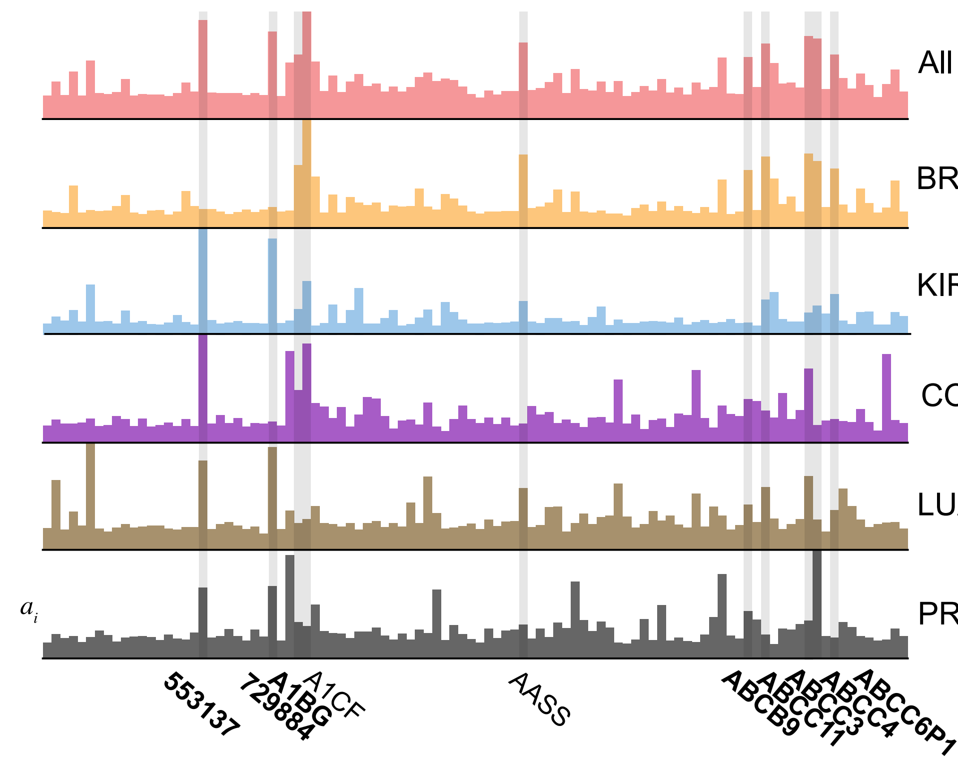
Slightly lower prediction accuracy when using Granger-causal objective

Discriminatory Genes across Cancer Types

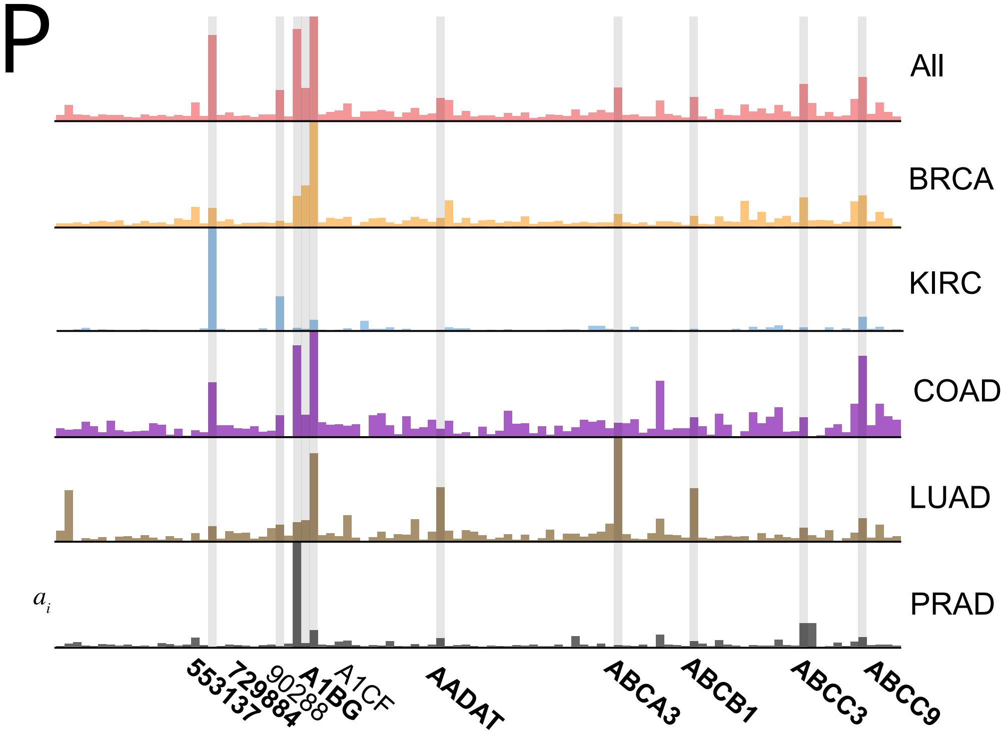
AME



LIME

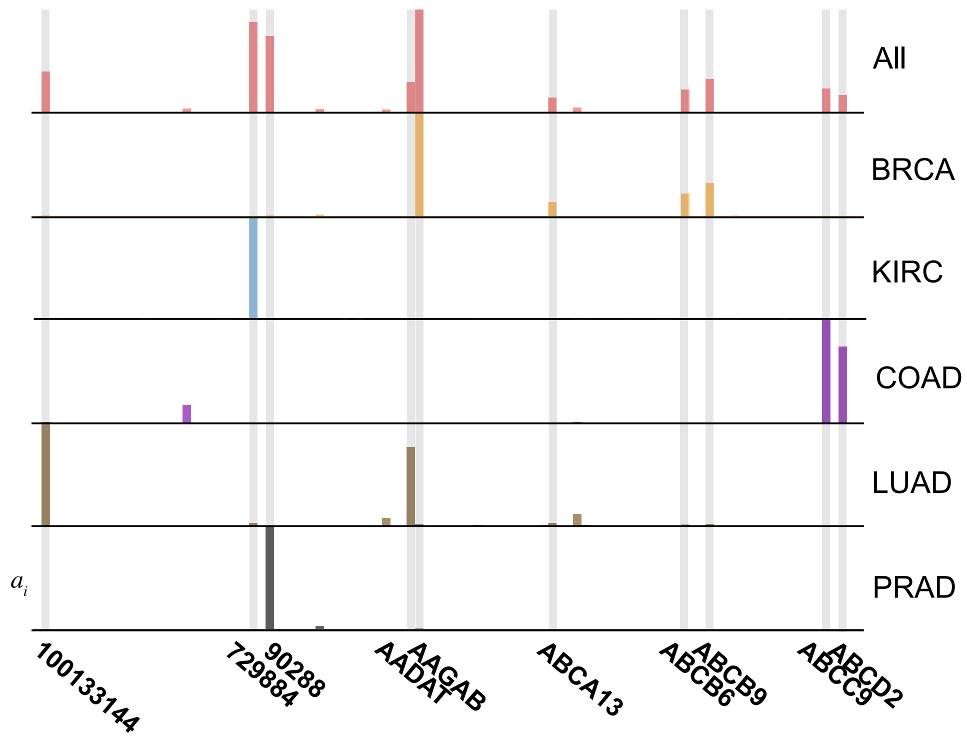


SHAP

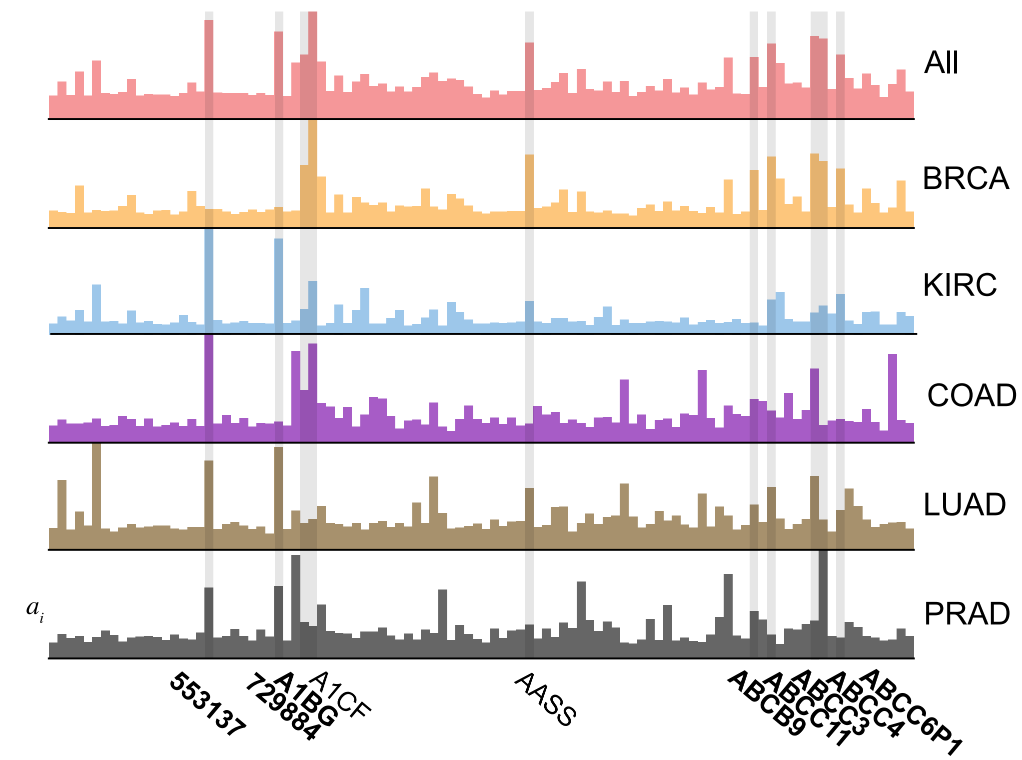


Discriminatory Genes across Cancer Types

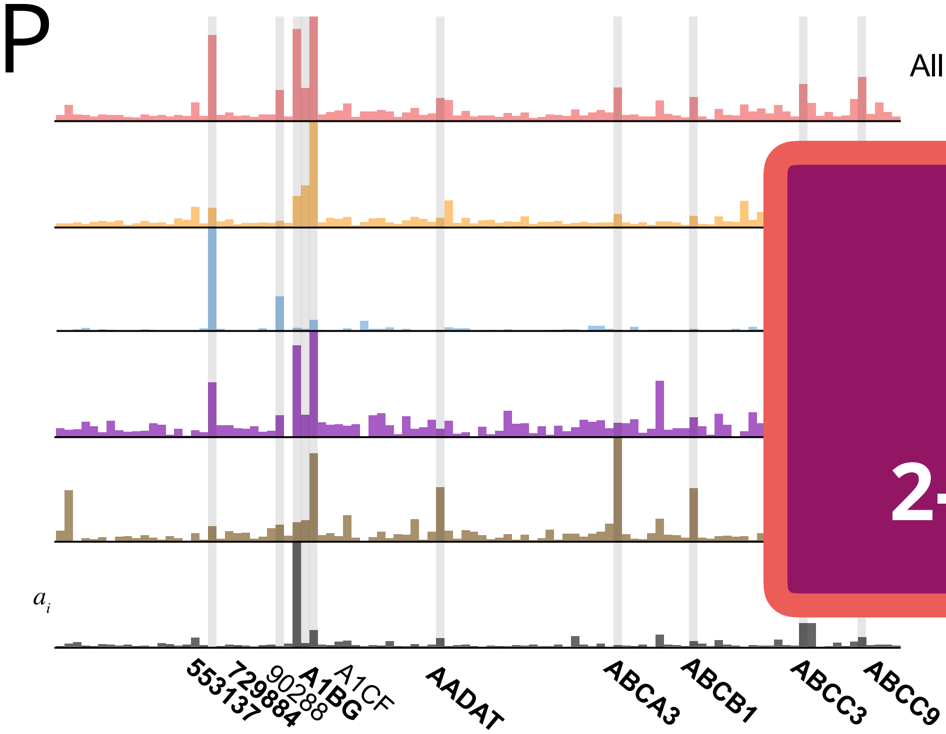
AME



LIME

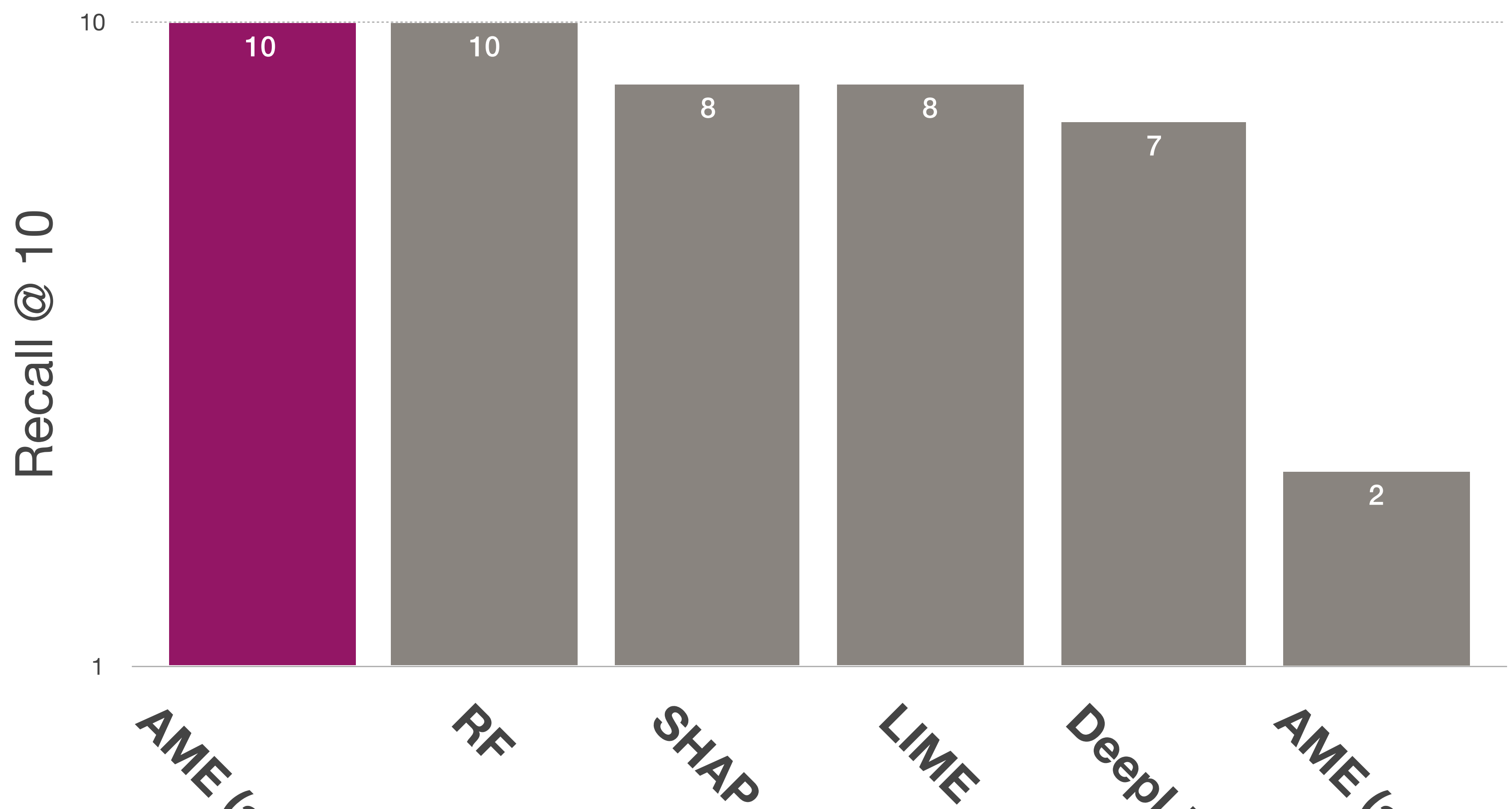


SHAP



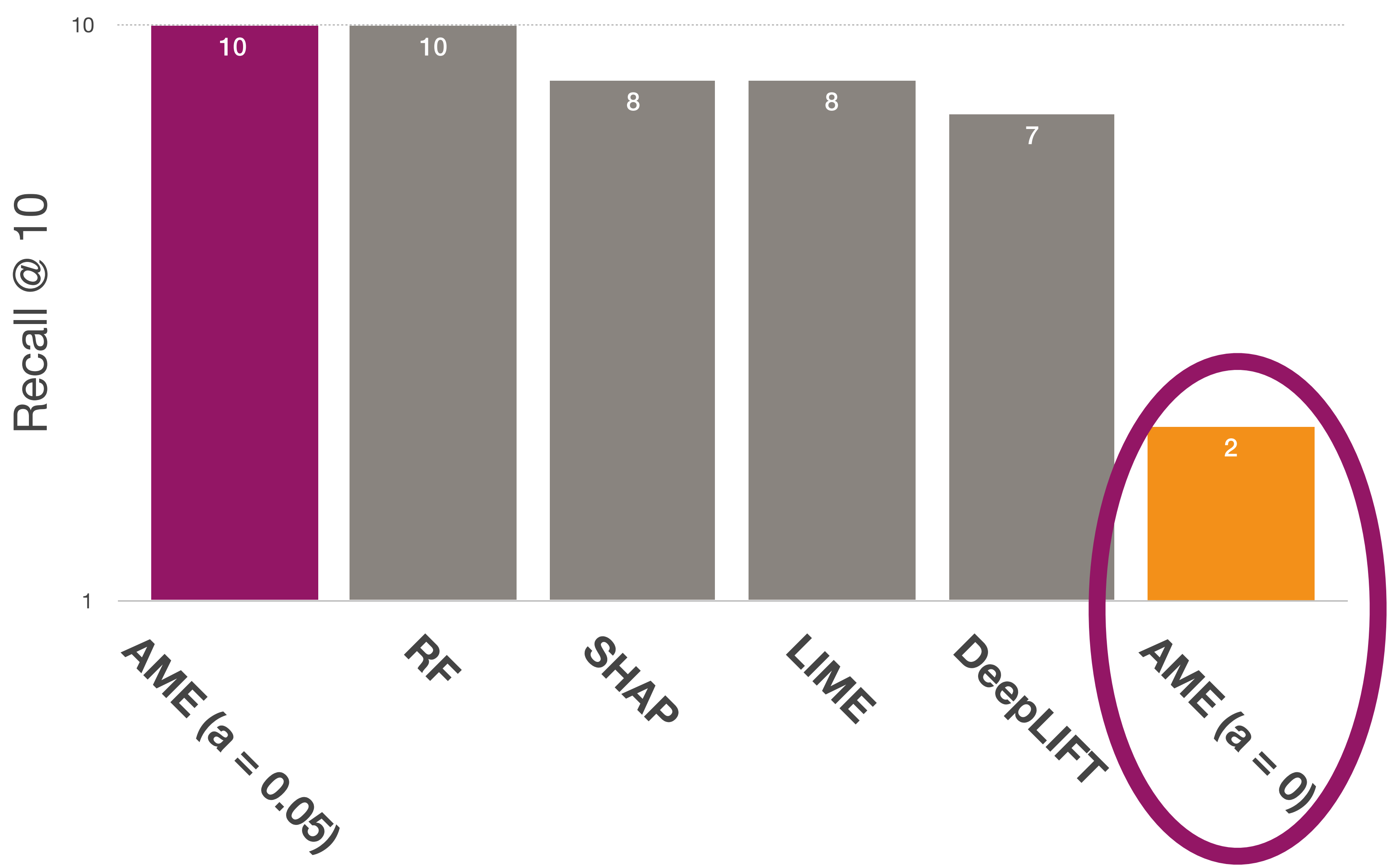
AME discriminates well between
1- cancer types, and
2- important and unimportant genes

Discriminatory Genes across Cancer Types



Associations discovered by AMEs are consistent with those reported by domain experts.

Discriminatory Genes across Cancer Types



Granger-causal objective is crucial for estimation accuracy.

Limitations

- No information about **direction** of importance, i.e. negative evidence
- Large numbers of experts (>200) can become **slow at training time**
  - Workaround: Feature grouping
- Requires **specific model architecture**

Conclusion

Conclusion

- We present a feature importance estimation approach that

✔ **learns to estimate importance** from labelled data

✔ produces accurate **predictions and importance** scores **in a single model**

✔ is **orders of magnitude faster** at estimating importance than perturbation-based approaches

✔ is **consistent with** associations reported by **domain experts**

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

75 NRP
Big Data
National Research Programme

Questions?

Patrick Schwab

@schwabpa

patrick.schwab@hest.ethz.ch
Institute for Robotics and Intelligent Systems
ETH Zurich

Schwab, Patrick, Miladinovic, Djordje, and Karlen, Walter.
Granger-causal Attentive Mixtures of Experts:
Learning Important Features with Neural Networks.
AAAI 2019
Source Code: github.com/d909b/AME