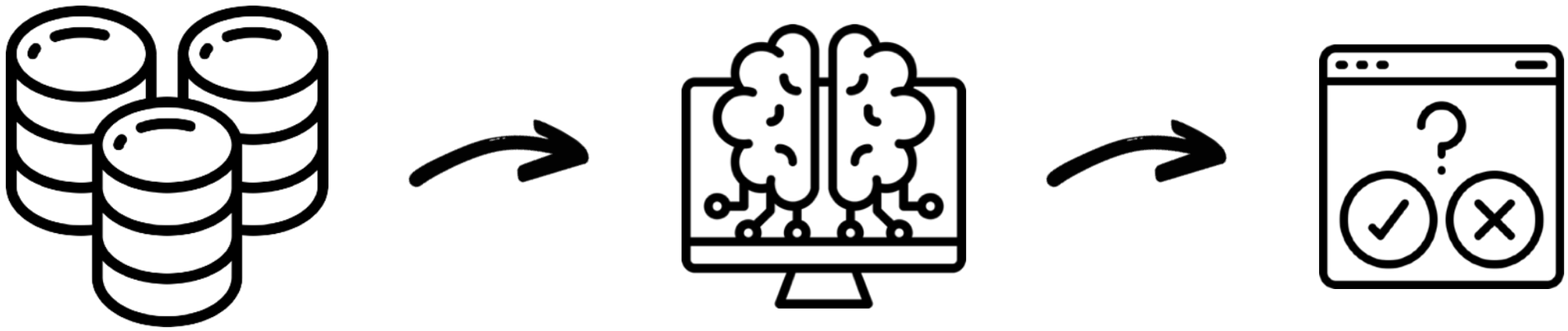
19: GP for explainable AI

- Explainable AI
- Explainability vs. interpretability
- GP for XAI/IML
- Example studies

Troubles with automated machine decision-making



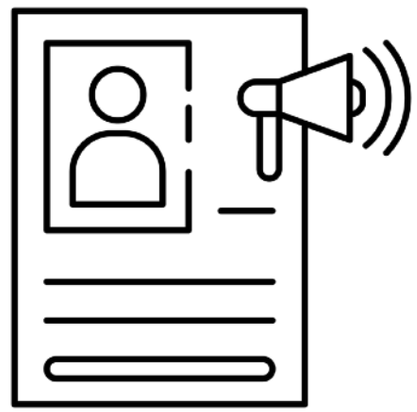
"When a Computer Program Keeps You in Jail"

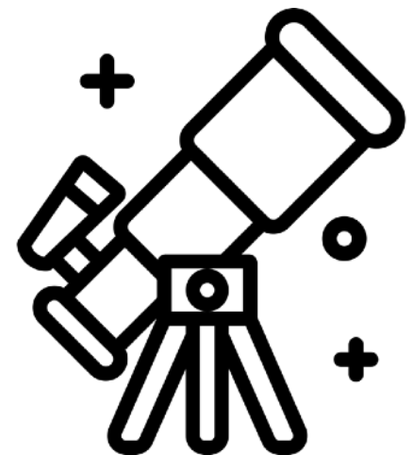
- *New York Times*, 2017

"Sacramento air quality: Internet results vary on health risk"

- The Sacramento Bee, 2018

When do we need to understand machine decisions?





Explainable

Interpretable

Transparent

Trustworthy

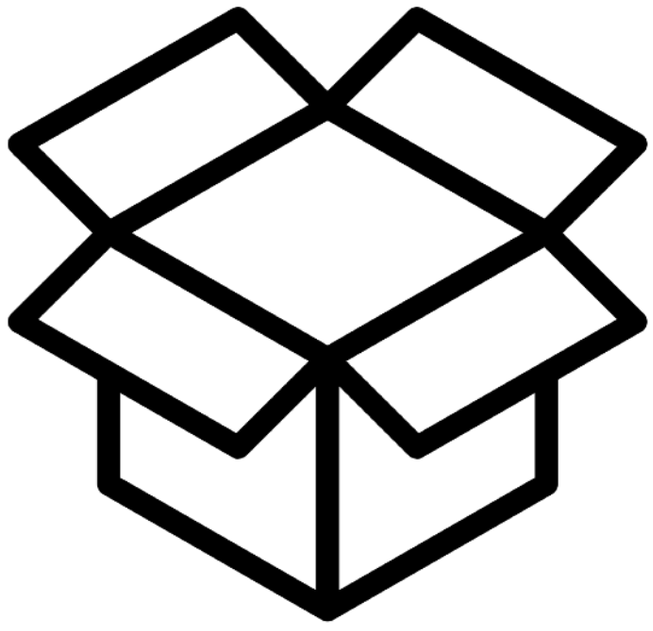
Reliable

Safe

Fair

Ethical

…



Explainability vs. Interpretability

Explainable: We can understand *why* a decision is made

- Influential features on the prediction
- Similar data points in the training set

Interpretable: We can understand *how* a decision is made

- Cause and effect
- Repeat the prediction by looking at the model and parameters

Lipton: The mythos of model interpretability. *ACM Communications*, 2016
Rudin: Stop explaining black box machine learning models and use interpretable models instead. *Nature Machine Intelligence,* 2019
Monar: Interpretable Machine Learning, 2020

Explain a model or a prediction

- Often post-hoc, using an additional model/technique
- Model-agnostic vs model-specific
- Global vs local

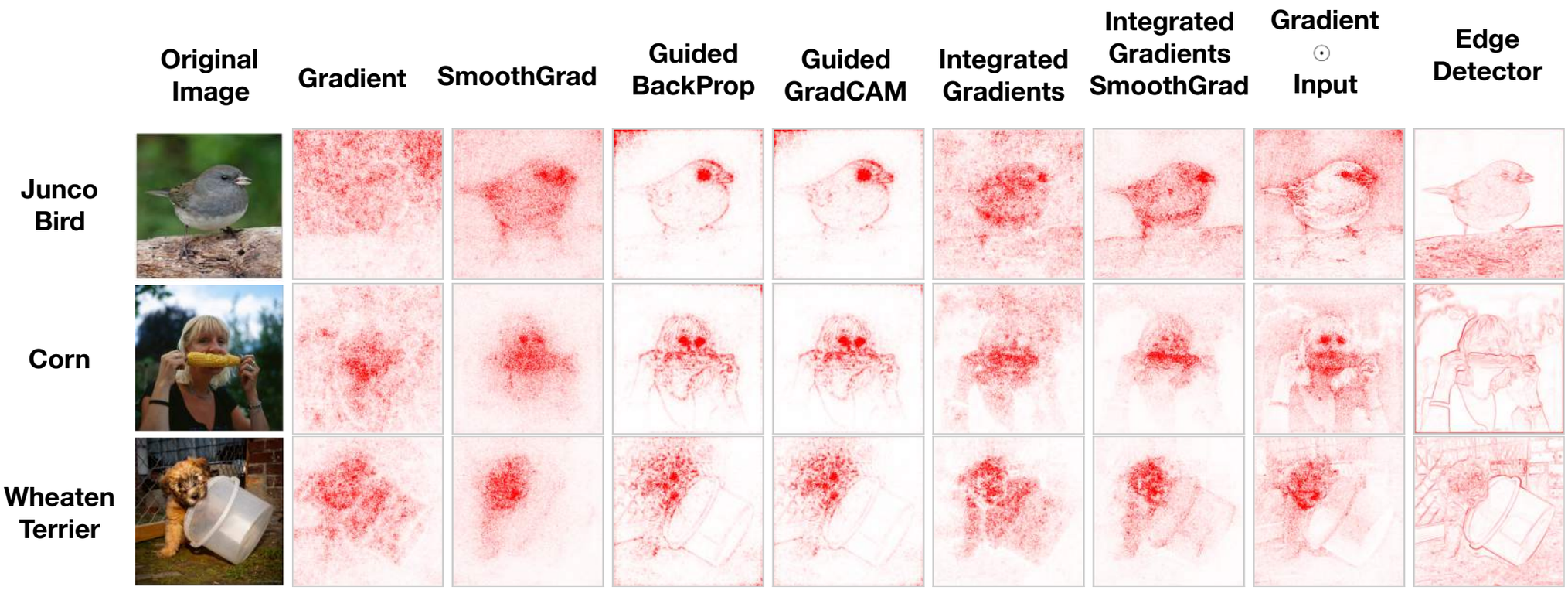


Figure 1: **Saliency maps for some common methods compared to an edge detector.** Saliency masks for 3 inputs for an Inception v3 model trained on ImageNet. We see that an edge detector produces outputs that are strikingly similar to the outputs of some saliency methods. In fact, edge detectors can also produce masks that highlight features which coincide with what appears to be relevant to a model's class prediction. We find that the methods most similar (see Appendix for SSIM metric) to an edge detector, i.e., Guided Backprop and its variants, show minimal sensitivity to our randomization tests.

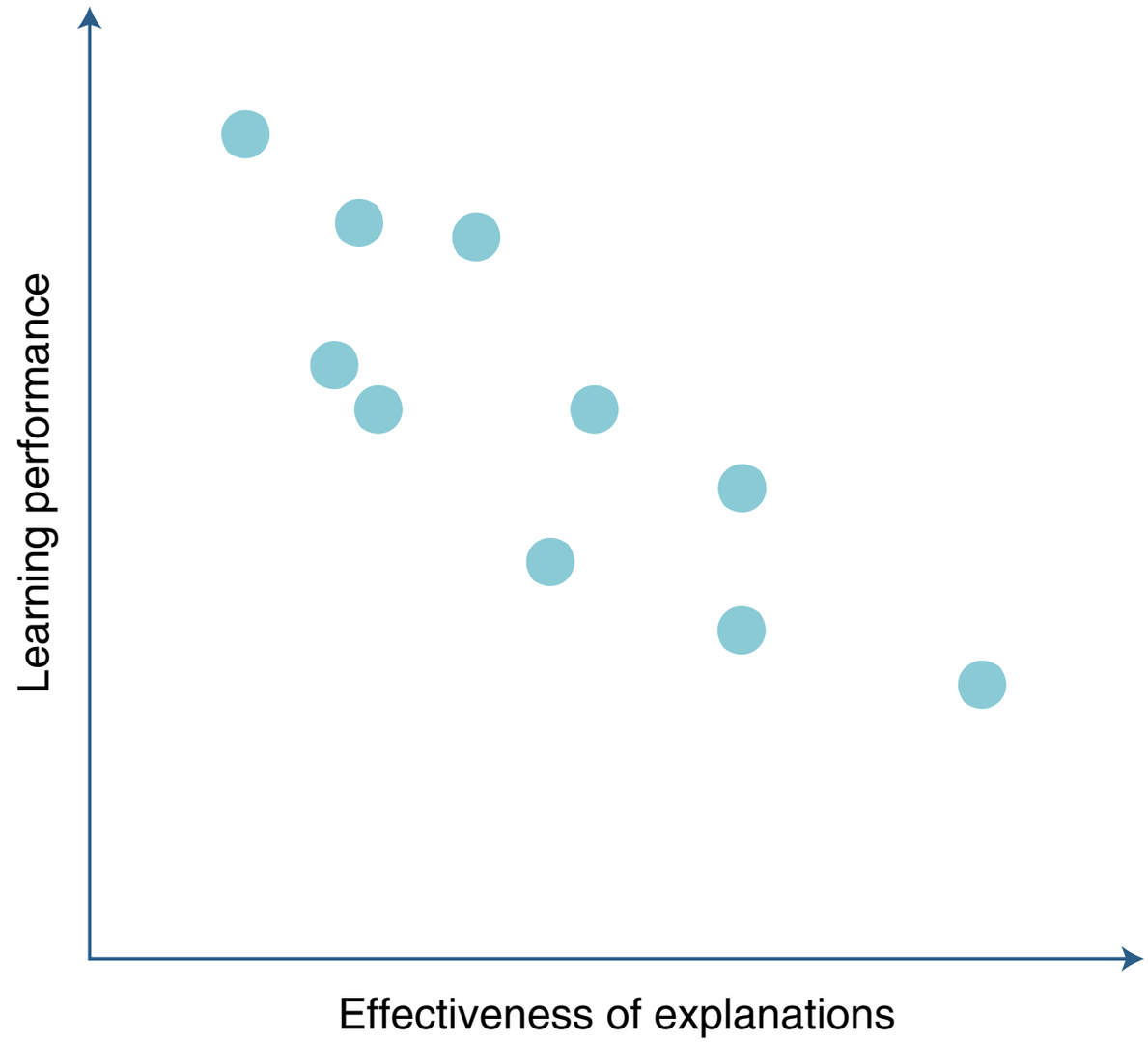
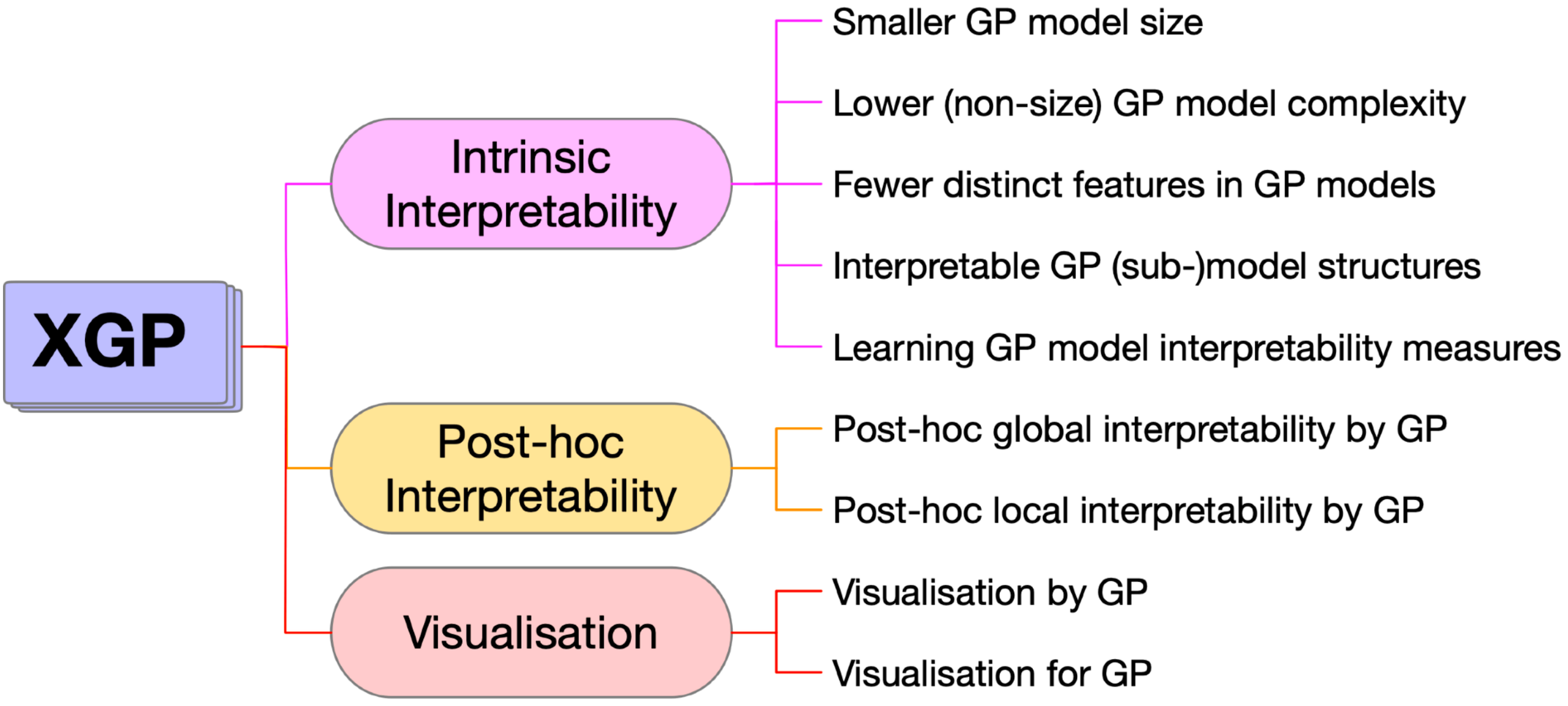


Fig. 1 | A fictional depiction of the accuracy–interpretability trade-off. Adapted from ref. [18], DARPA.

Interpretable models

- Models intrinsically understandable
- Shallow decision lists or decision trees
- Optimized compact/sparse models
- Interpretability defined for specific domains

GP for explainable/interpretable learning

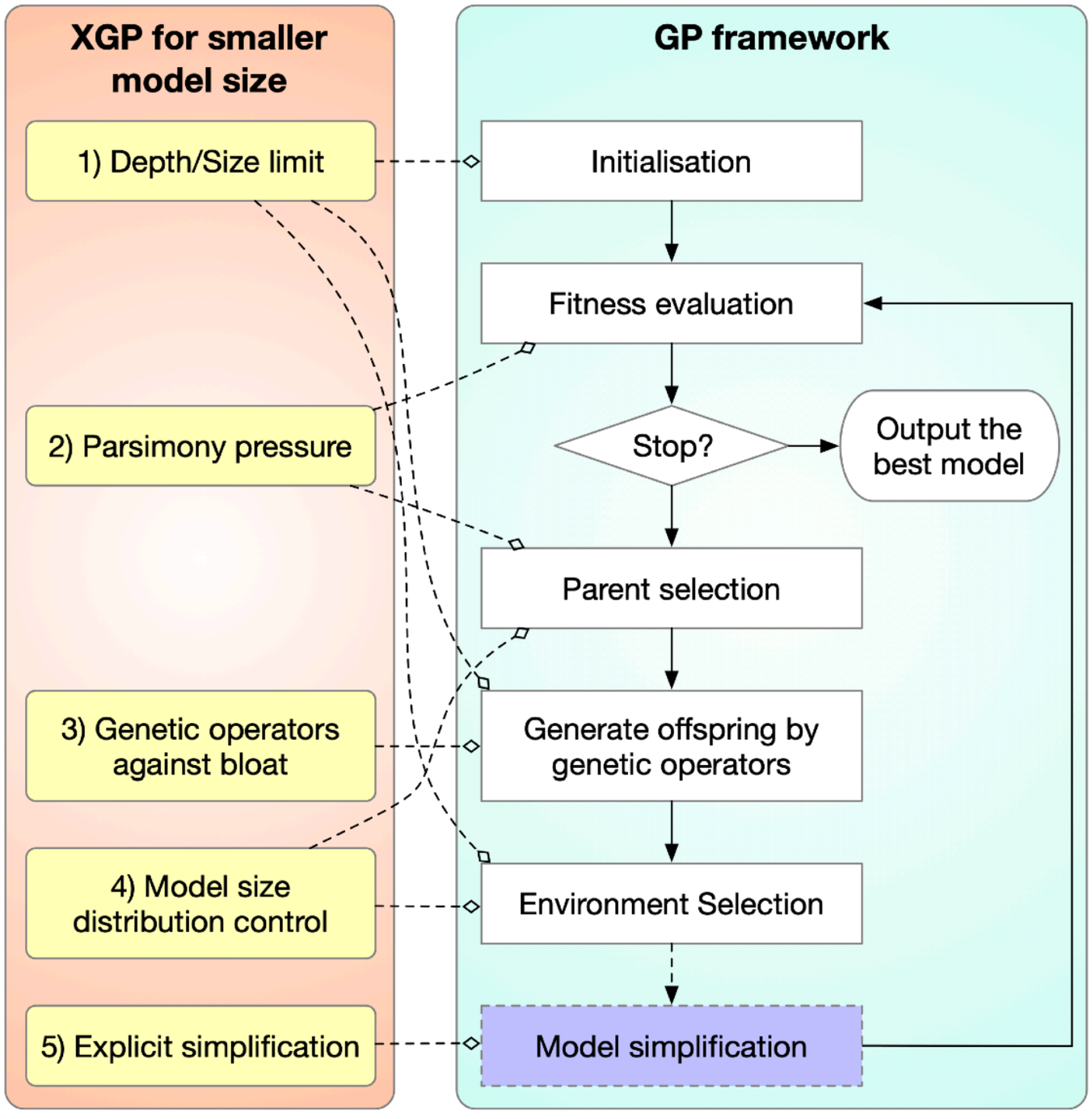


Mei et al.: Explainable artificial intelligence by genetic programming: A survey . *IEEE Transactions on Evolutionary Computation,* 2022

Example studies

- Optimize decision trees for black-box classifiers (Evans et al., *GECCO*, 2019)
- Fit a local explanation model for a given input example (Ferreira et al., *CEC*, 2020)
- Evolve symbolic expression trees for reinforcement learning tasks (Hein et al., *Engineering Applications of Artificial Intelligence*, 2018)
- Evolve compact, explicit rules for Parkinson disease handwriting classification (Partial et al., *Artificial Intelligence in Medicine*, 2021)

Explicitly reduce GP model size



Mei et al.: Explainable artificial intelligence by genetic programming: A survey . *IEEE Transactions on Evolutionary Computation,* 2022

Sha *et al. BMC Bioinformatics* (2021) 22:284
https://doi.org/10.1186/s12859-021-04209-1

BMC Bioinformatics

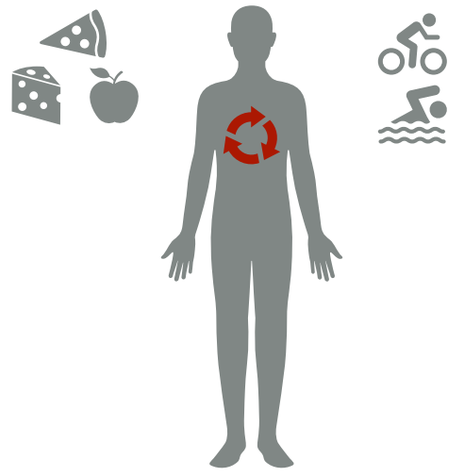
RESEARCH Open Access

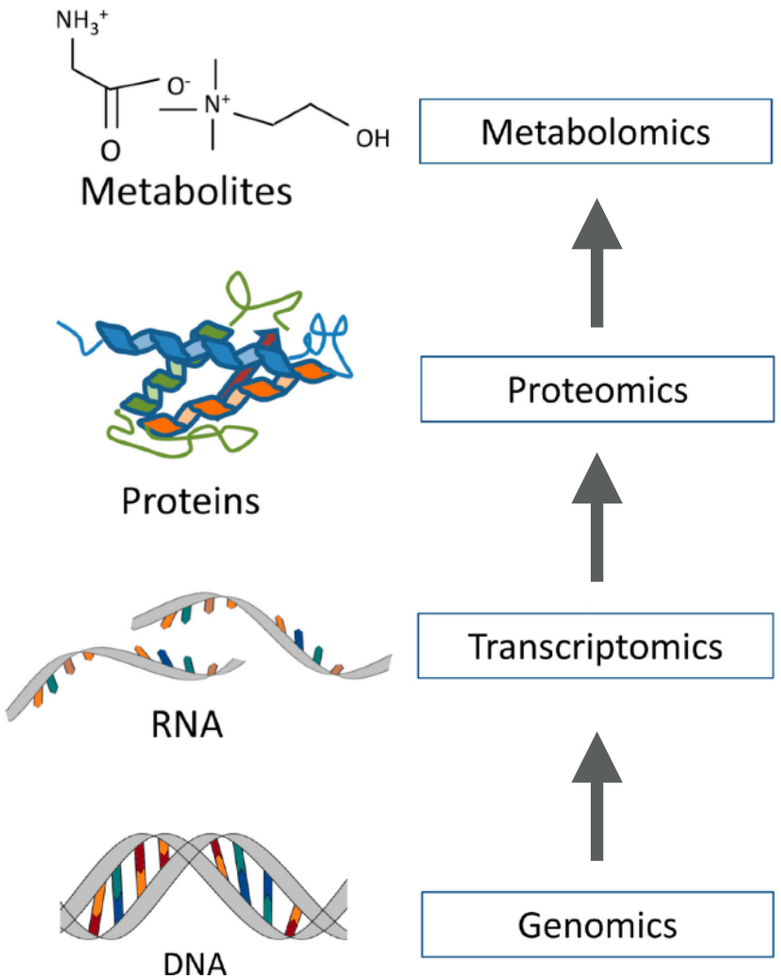
SMILE: systems metabolomics using interpretable learning and evolution

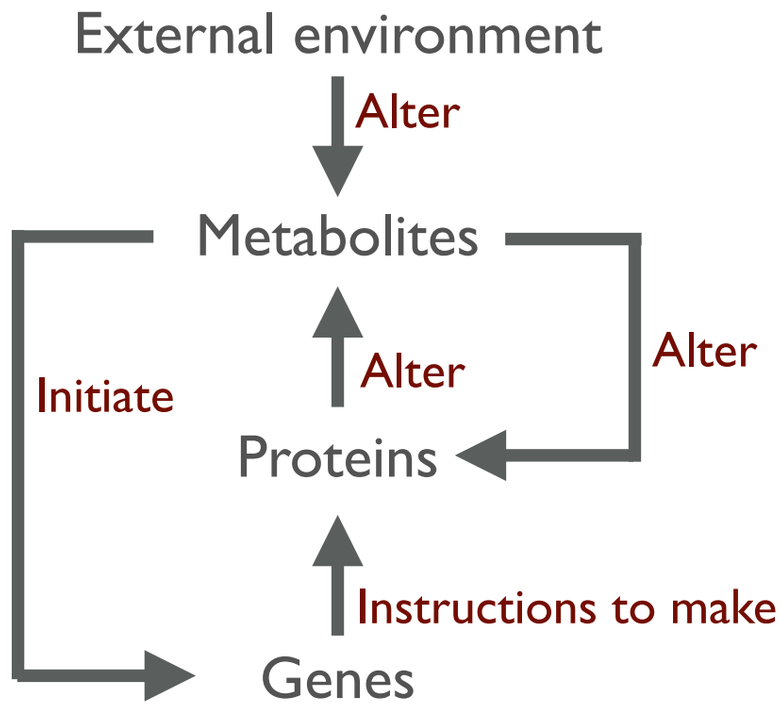
Chengyuan Sha[1], Miroslava Cuperlovic-Culf[2] and Ting Hu[1*]

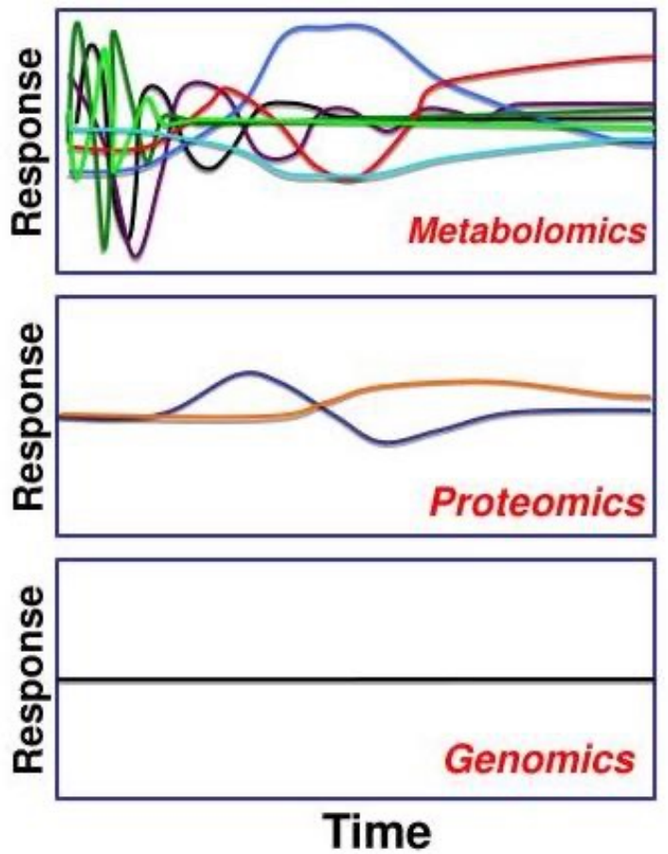
- A learning framework tailored to the application problem
- Linear GP for Alzheimer's Disease classification
- Training on metabolomics data
- Collection of best programs out of 1000 runs
- Interactive interface for results interpretation and visualization

Metabolomics



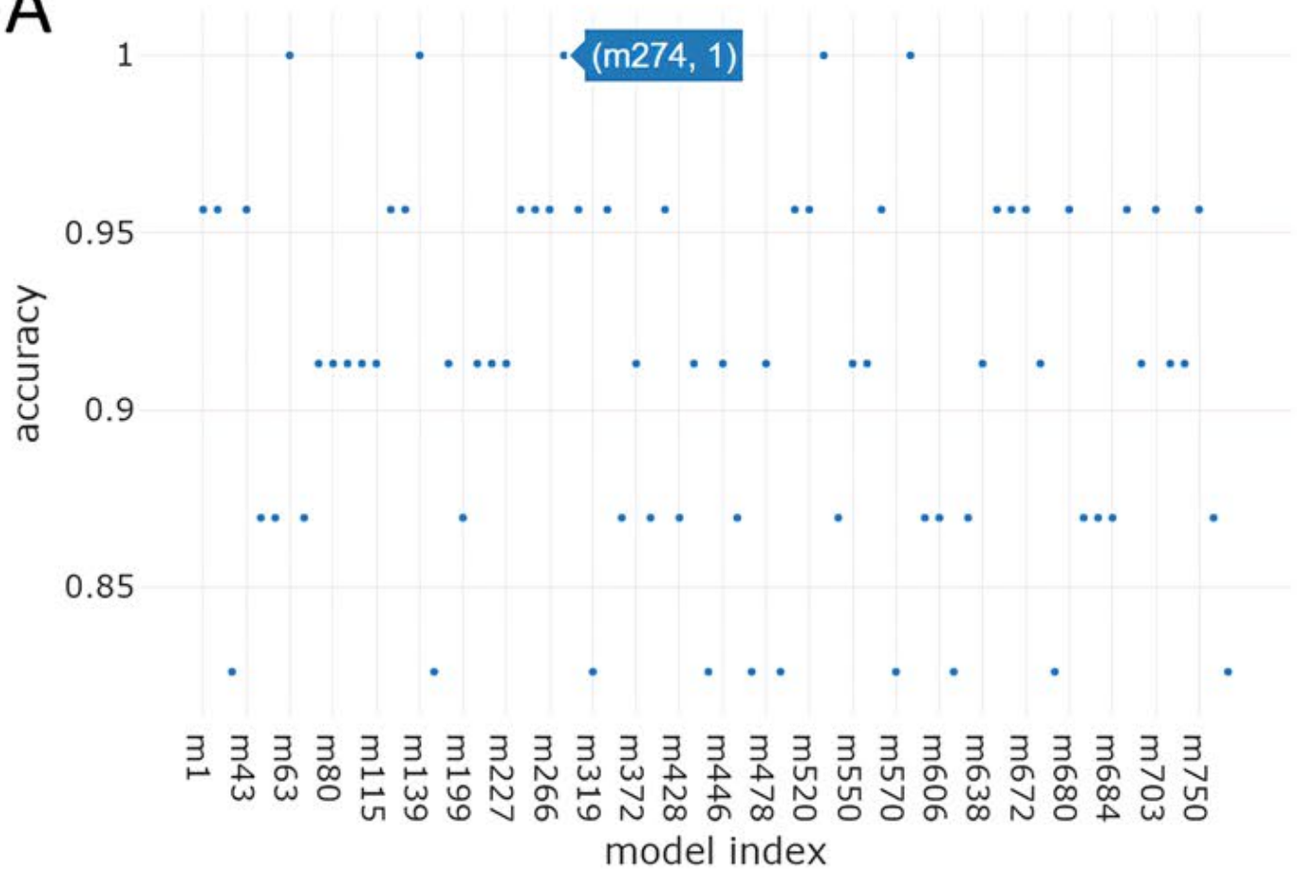






David Wishart, *Bioinformatics*, 2016
Haukaas, Euceda, Giskeodegard, and Bathen, *Metabolites*, 2017

A

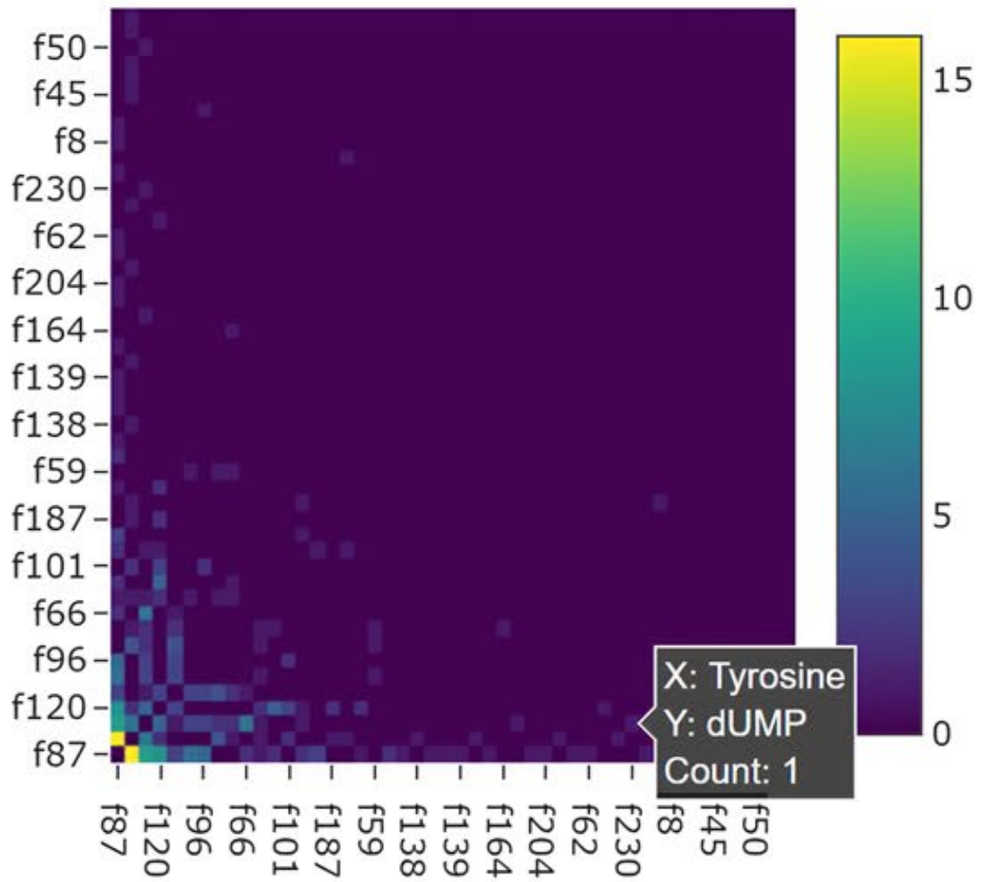


B

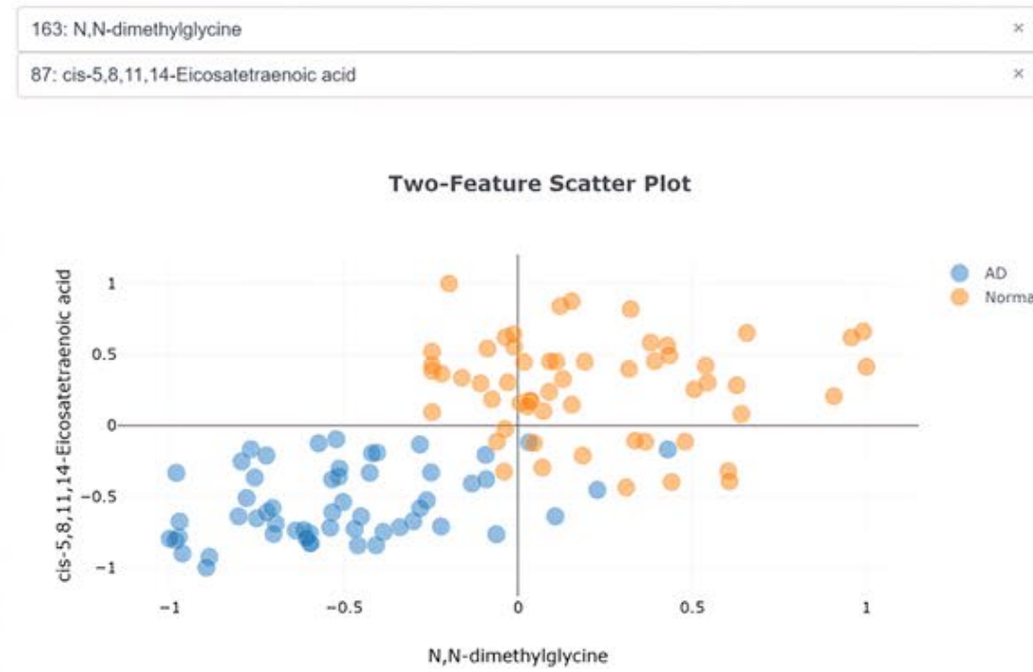
Detailed Model Info:

```
The default value of r0: 0.91
Models:
I0: r28 = cis-5,8,11,14-Eicosatetraenoic acid + 0.19
I1: r0 = N,N-dimethylglycine + r28
Output register r[0] will then go through sigmoid transformation S
if S(r[0]) is less or equal than 0.5:
  this sample will be classified by this model as class 0, i.e. diseased.
else:
  class 1, i.e. healthy
```

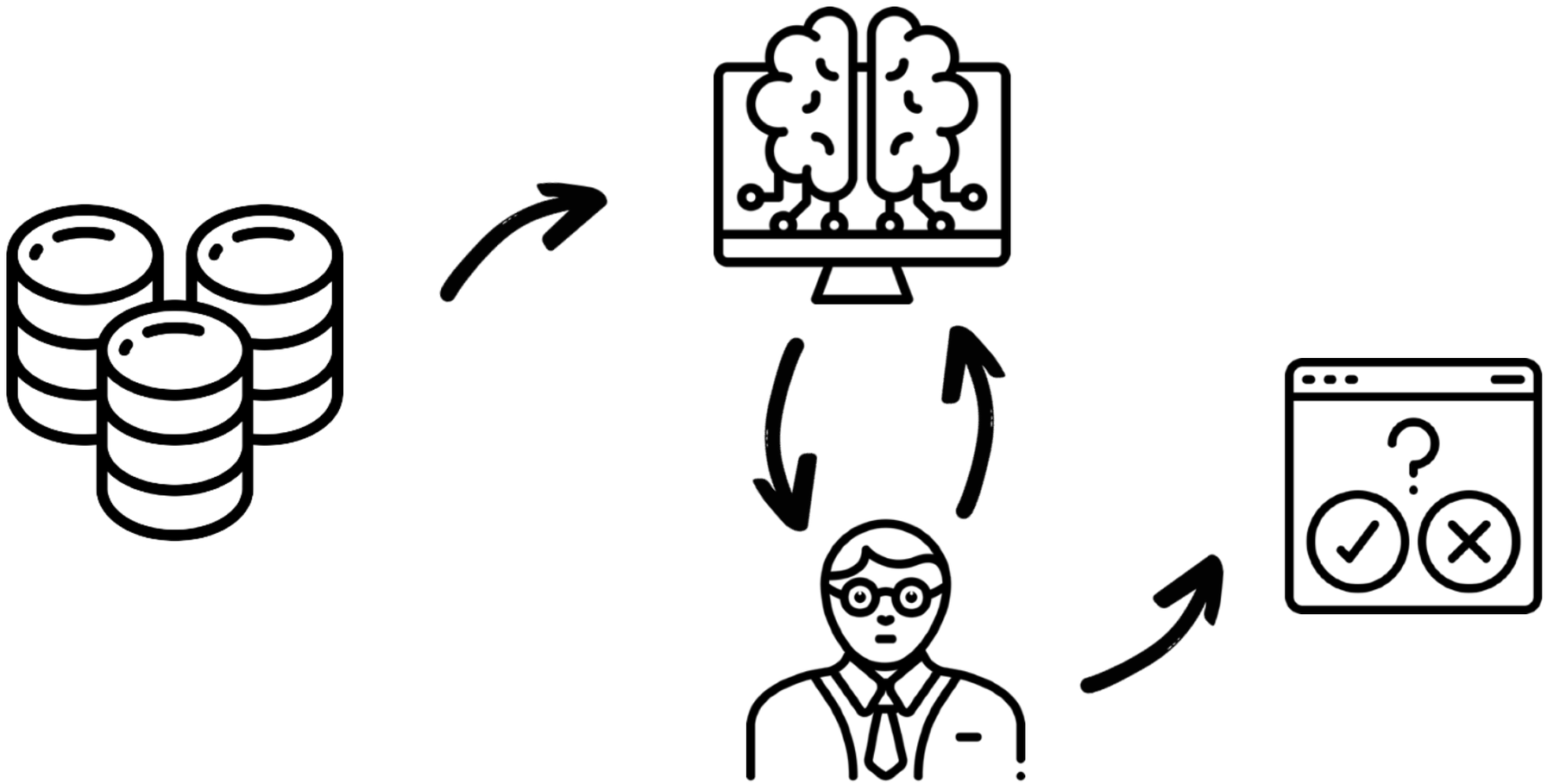
C



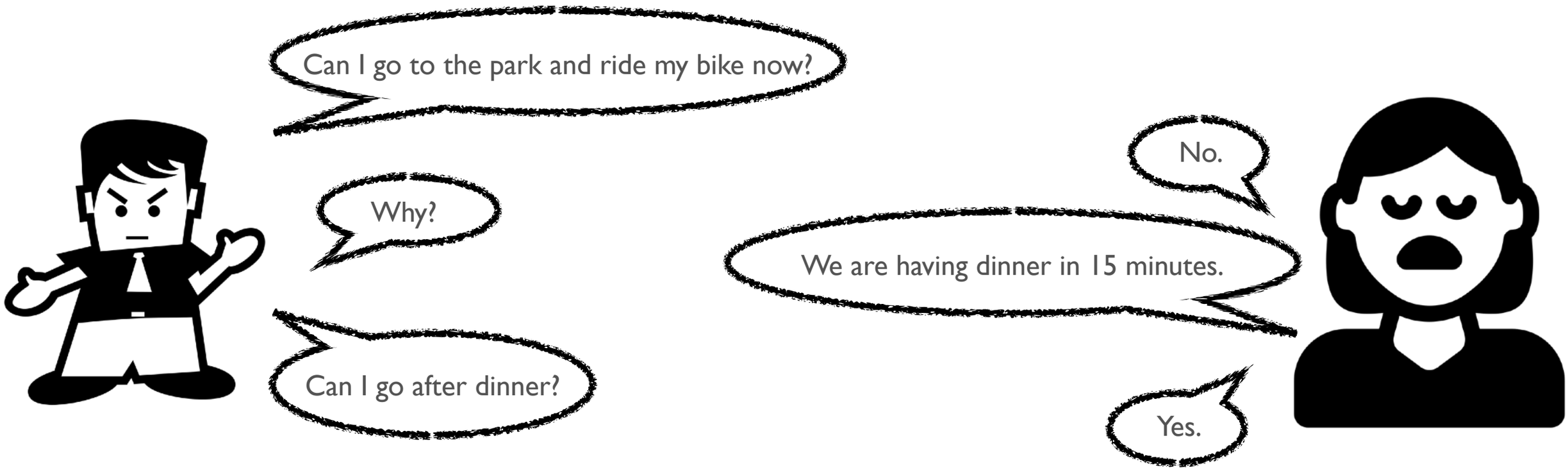
D



Bring human to the decision-making process



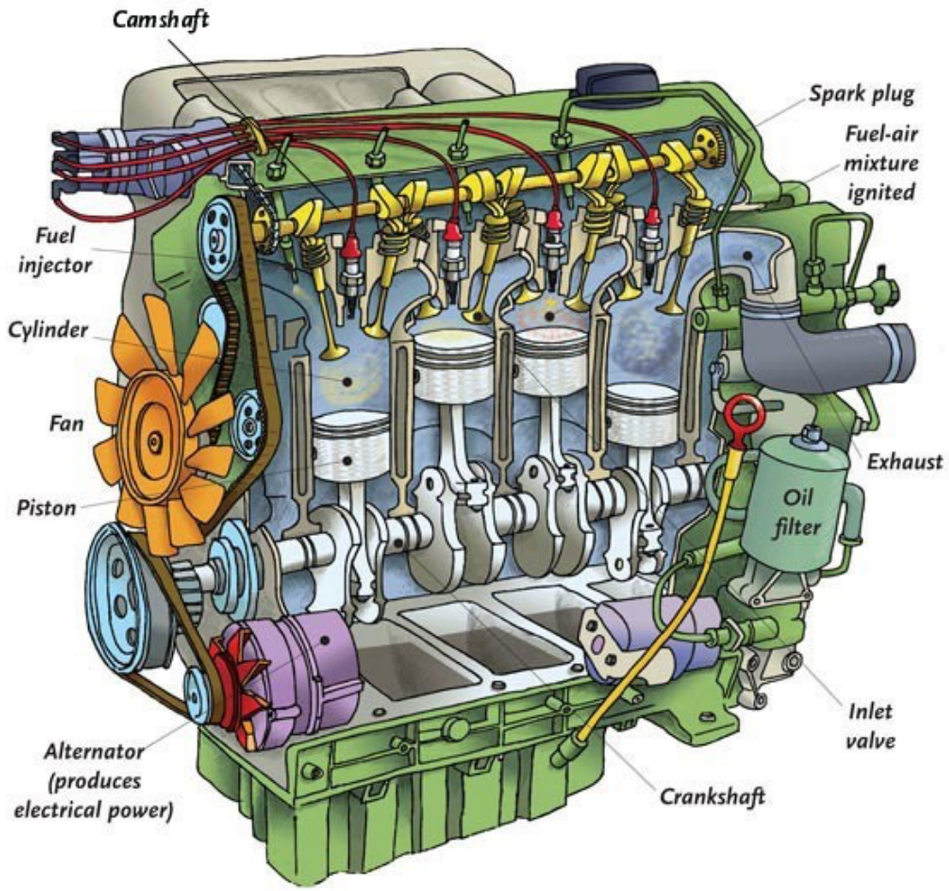
What is an effective explanation system?



Factor → Outcome

Alt. factor → Alt. outcome

What is an interpretable system?



a

$R_0 = 0$

$R_{15} = 0.0625$

$R_0 = R_{15} + R_0$

Skip next if Acetylornithine > C18

R_0 = Ornithine - Arginine

if $R_0 > 0$: predict high risk

else: predict low risk

b