Interpretability in classifiers

Hybrid techniques based on black-box classifiers and interpretable control modules

Luis Galárraga

Workshop on *Advances in Interpretable Machine Learning and Artificial Intelligence* (AIMLAI@EGC)

Metz, 22/01/2019

Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

Agenda

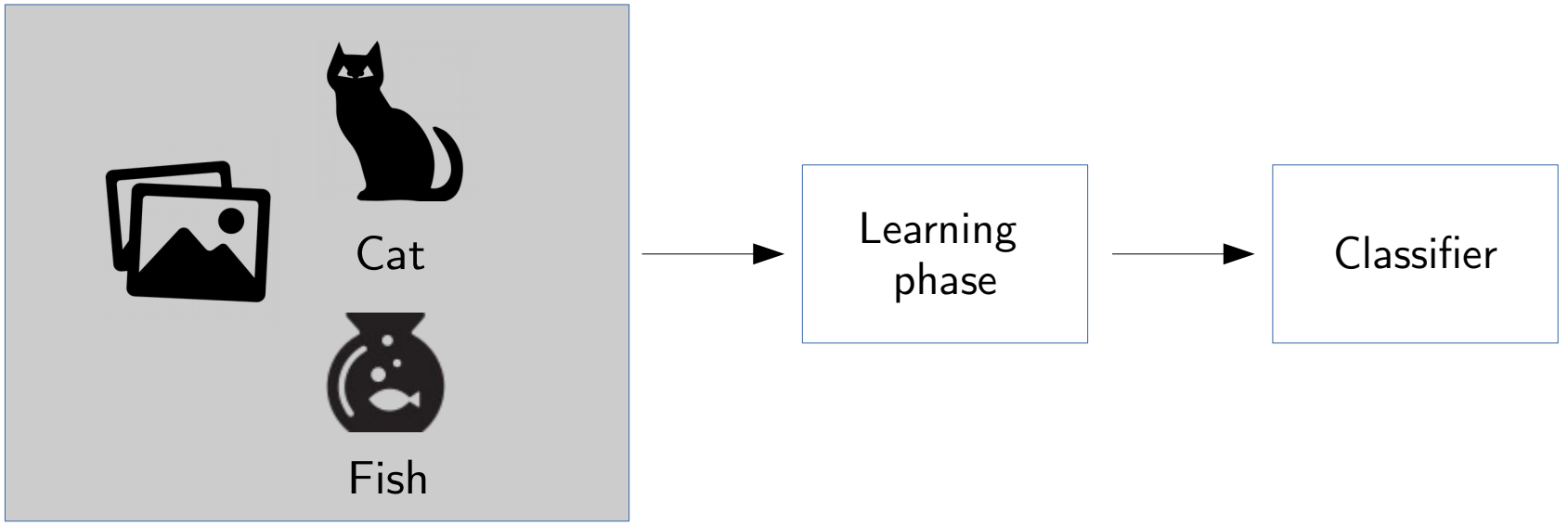
- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

Agenda

- Interpretability in **classifiers**: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

Machine Learning Classifiers

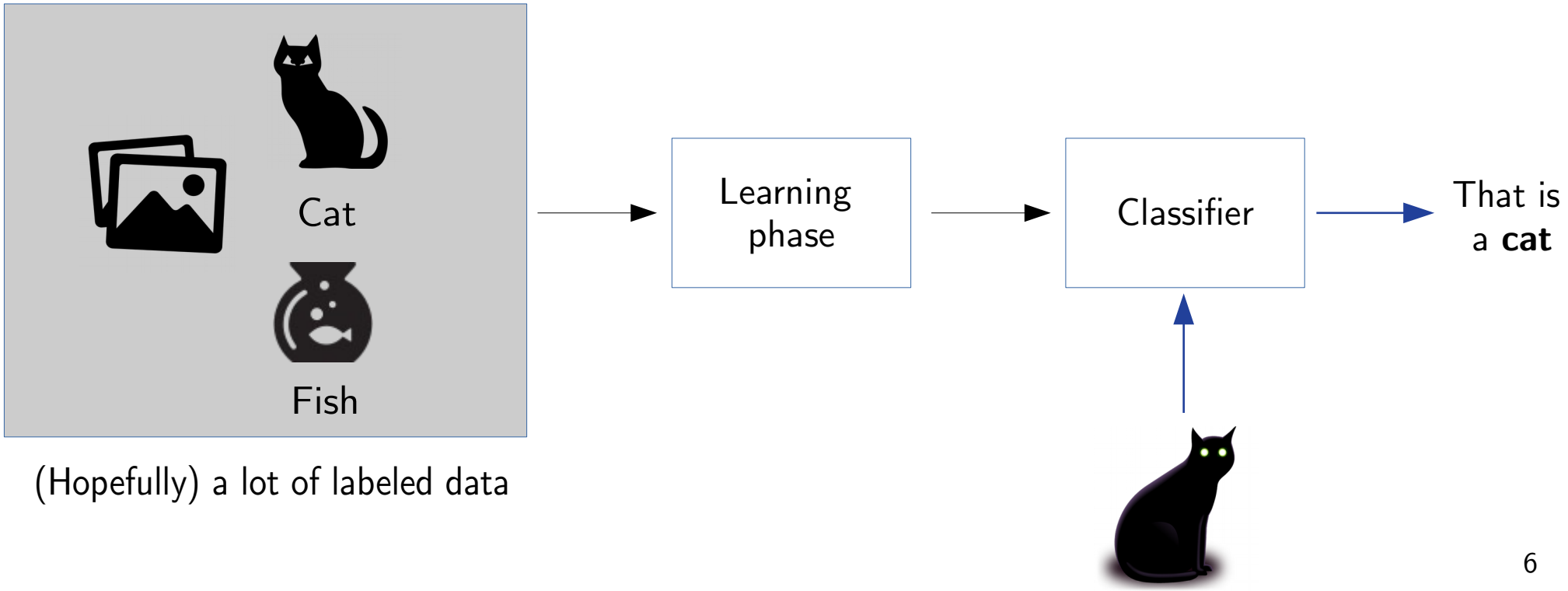
- Models that assign classes to instances
  - The model is learned and trained from labeled data
  - Labels are predefined: supervised learning



(Hopefully) a lot of labeled data

Machine Learning Classifiers

- Models that assign classes to instances
  - The model is learned and trained from labeled data
  - Labels are predefined: supervised learning



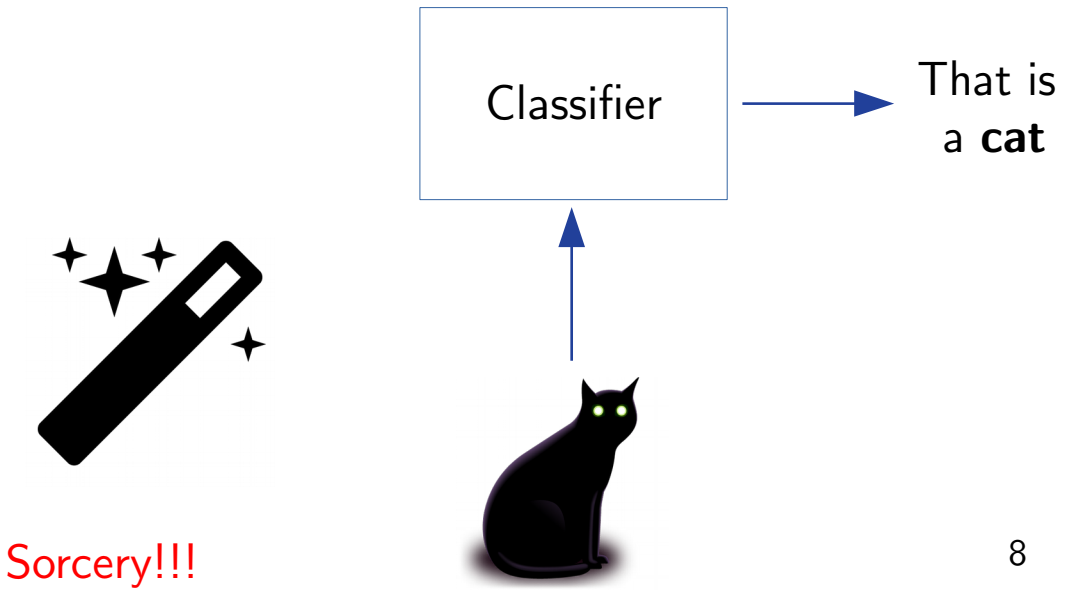
(Hopefully) a lot of labeled data

Agenda

- **Interpretability** in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

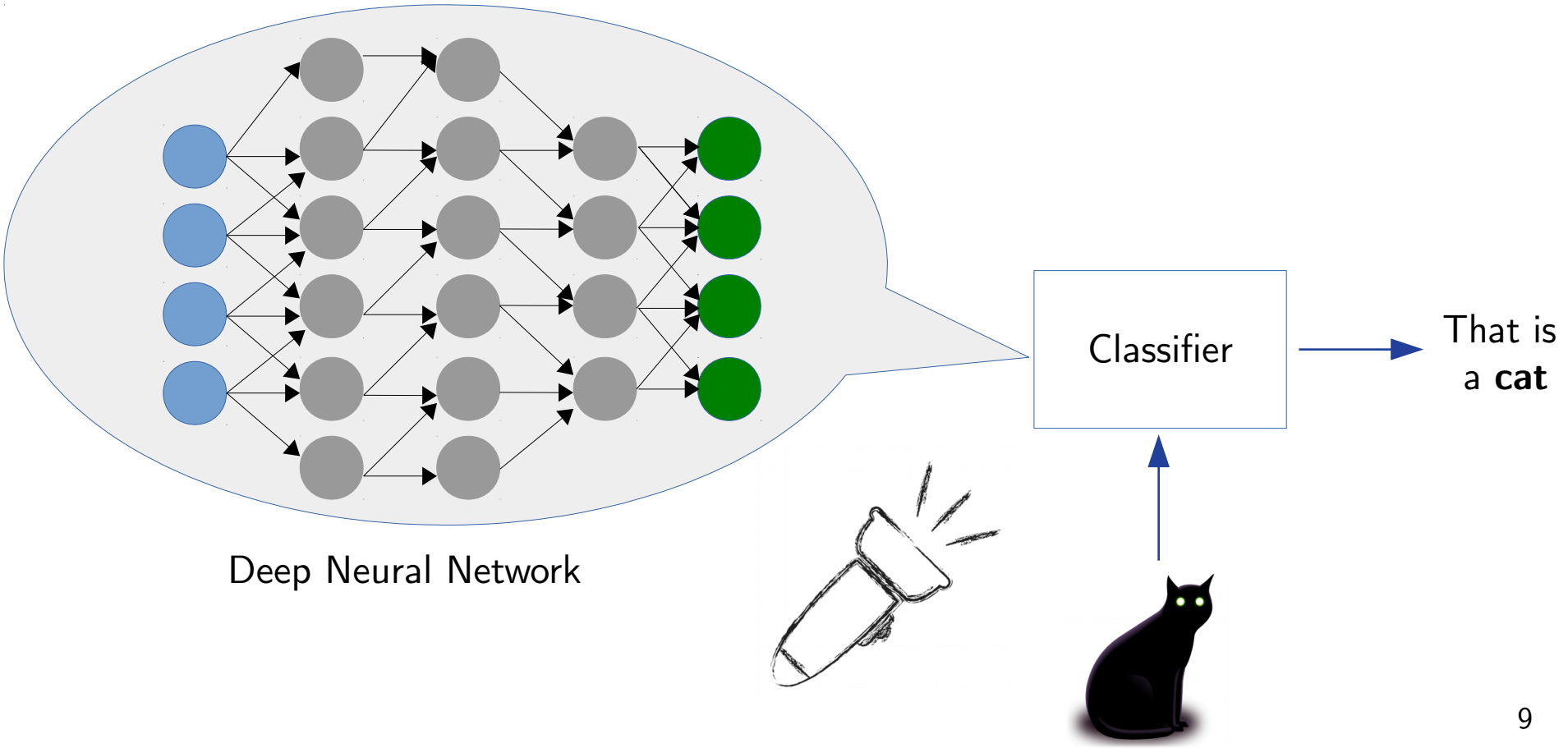
Interpretability in classifiers

Some ML classifiers can be really complex



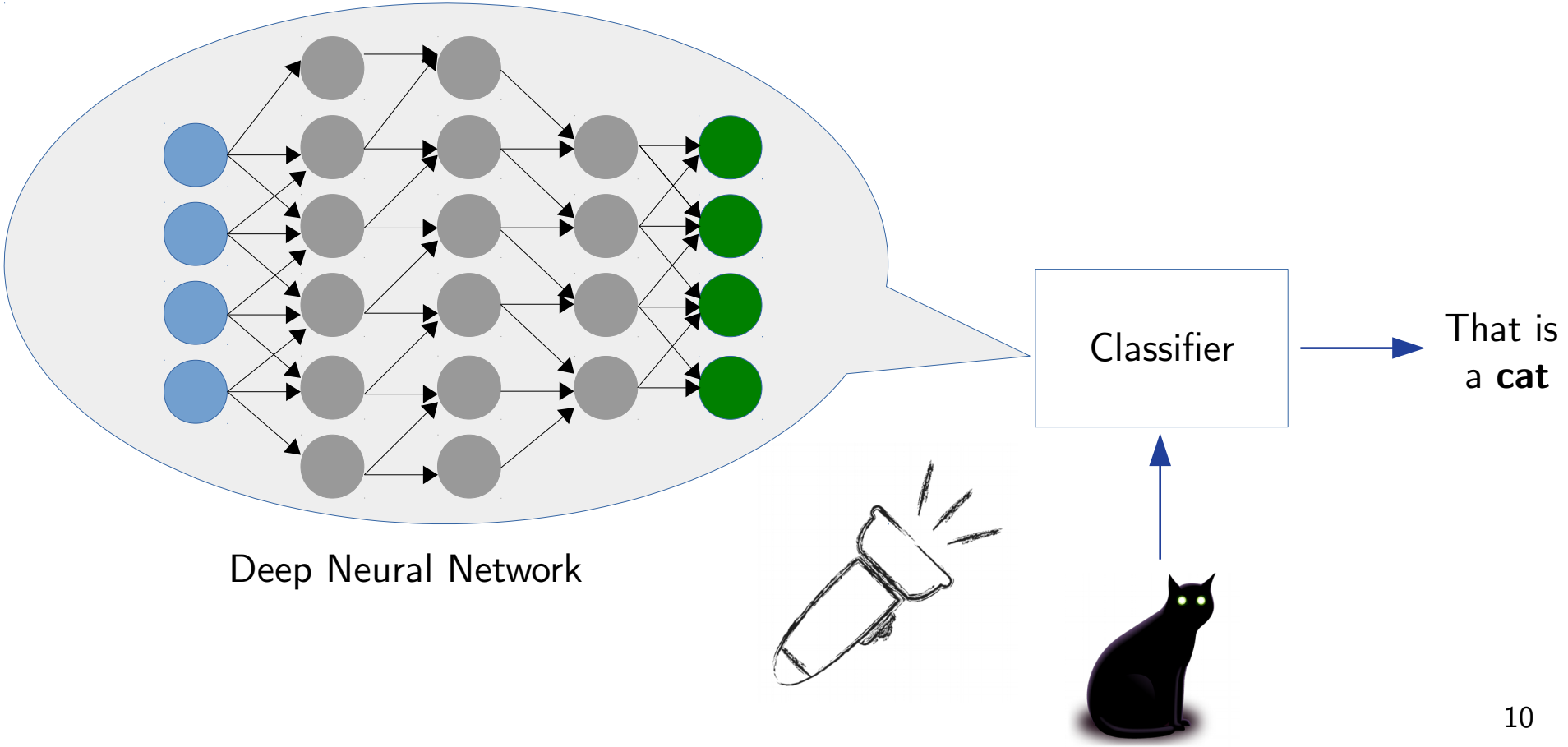
Interpretability in classifiers

Some ML classifiers can be really complex



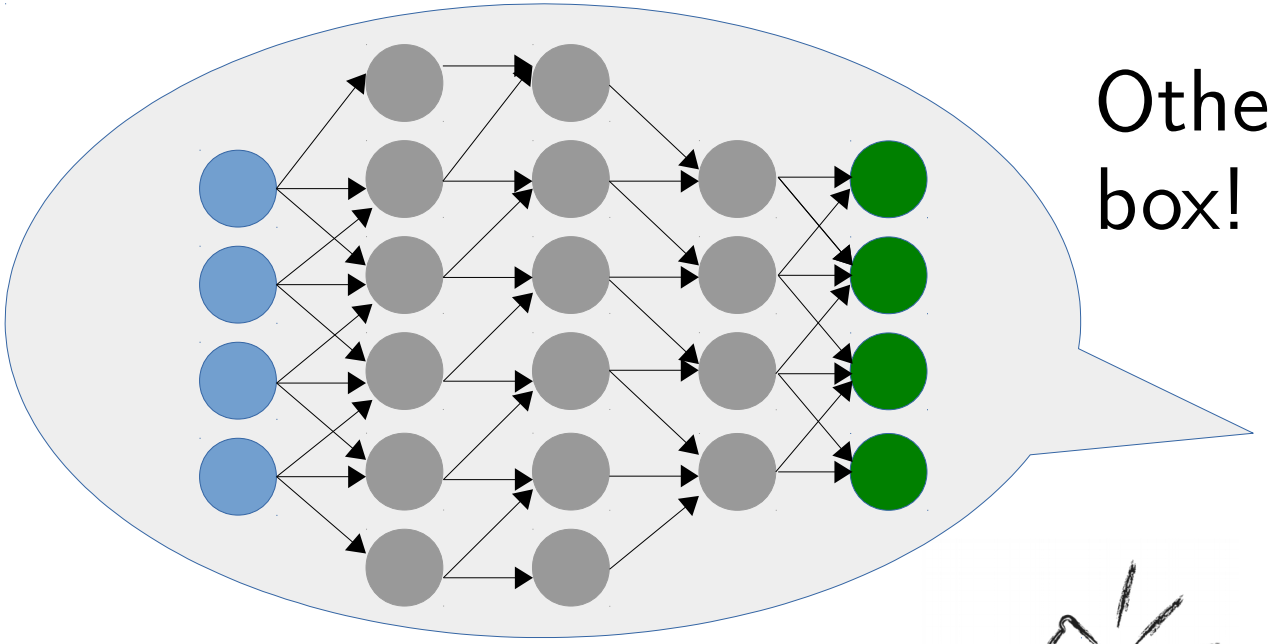
Interpretability in classifiers: What?

A classifier is *interpretable* if the rationale behind its answers can be easily *explained*



Interpretability in classifiers: What?

A classifier is *interpretable* if the rationale behind its answers can be easily *explained*

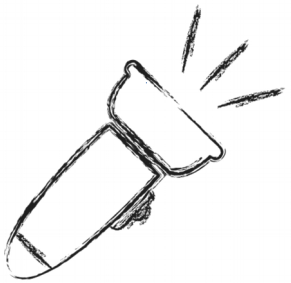


Deep Neural Network

Otherwise it is a black box!

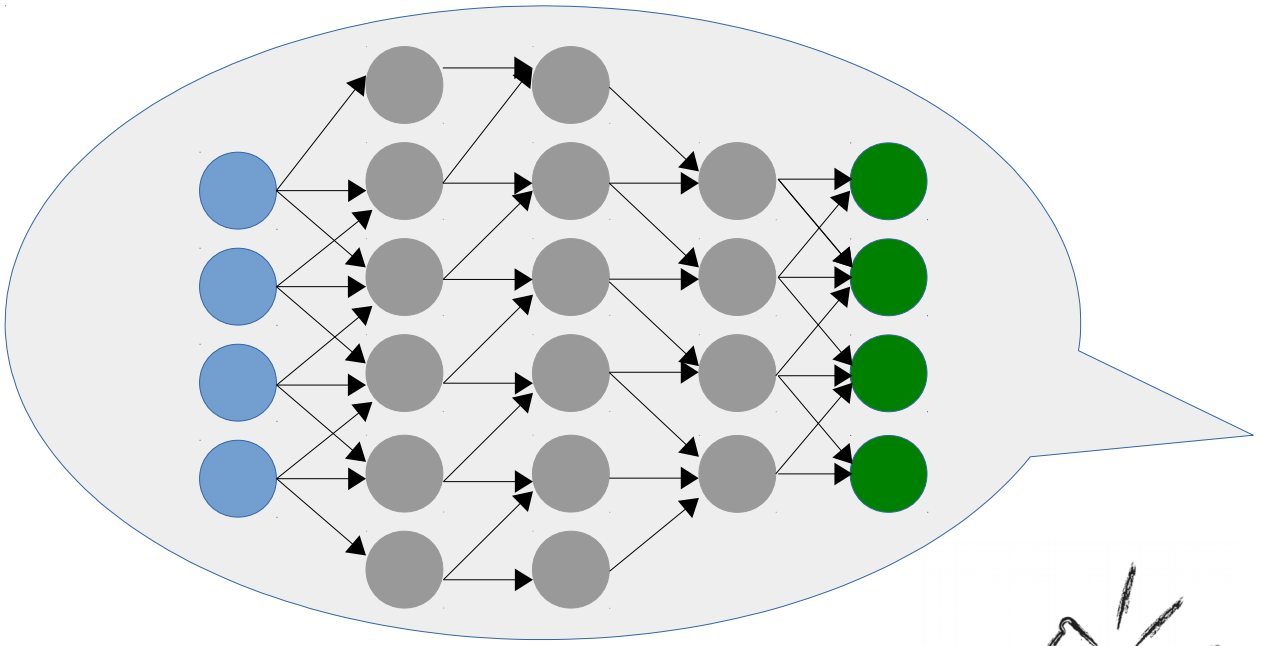
Classifier

That is a **cat**



Interpretability in classifiers: What?

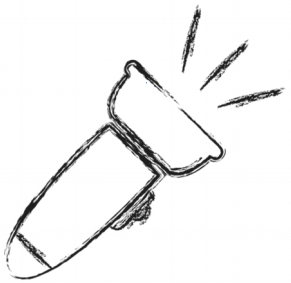
interpretability ≅ explainability ≅ comprenhensibility



Deep Neural Network

Classifier

That is a **cat**



Interpretability in classifiers: Why?

- Classifiers are used to make critical decisions

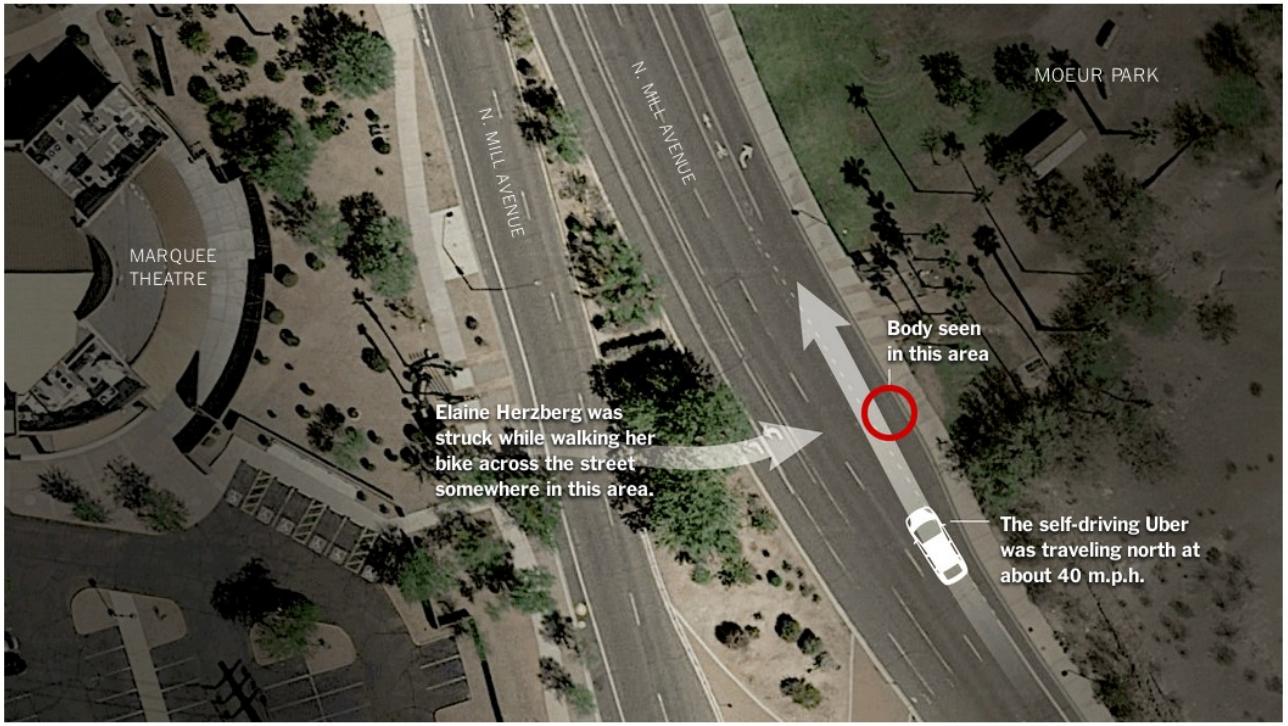
Interpretability in classifiers: Why?

How a Self-Driving Uber Killed a Pedestrian in Arizona

By **TROY GRIGGS** and **DAISUKE WAKABAYASHI** UPDATED MARCH 21, 2018

A woman was struck and killed on Sunday night by an autonomous car operated by Uber in Tempe, Ariz. It was believed to be the first pedestrian death associated with self-driving technology.

What We Know About the Accident



https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

Interpretability in classifiers: Why?

- Classifiers are used to make critical decisions
- Need to know the rationale behind an answer
  - For debugging purposes
    - To tune the classifier
    - To spot biases in the data
  - For legal and ethical reasons
    - General Data Protection Regulation (GDPR)
    - To understand the source of the classifier's decision bias
    - To generate trust

Interpretability in classifiers: Why?

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

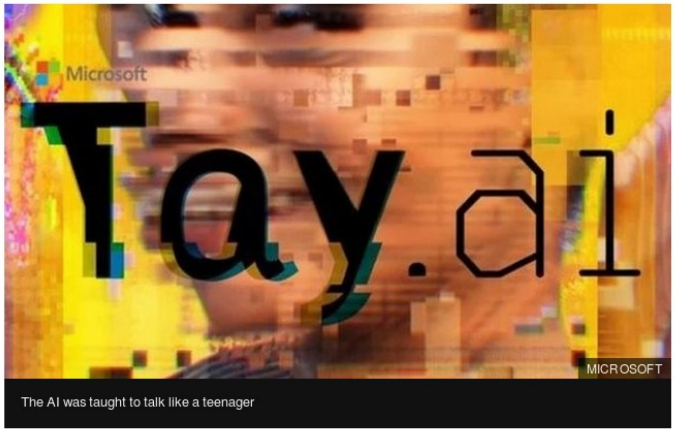
Interpretability in classifiers: Why?

https://www.bbc.com/news/technology-35902104

Tay: Microsoft issues apology over racist chatbot fiasco

Dave Lee
North America technology reporter

25 March 2016 | Share



Microsoft has apologised for creating an artificially intelligent chatbot that quickly turned into a holocaust-denying racist.

But in doing so made it clear Tay's views were a result of nurture, not nature. Tay confirmed what we already knew: people on the internet can be cruel.

Tay, aimed at 18-24-year-olds on social media, was targeted by a "coordinated attack by a subset of people" after being launched earlier this week.

Within 24 hours Tay had been deactivated so the team could make "adjustments".

Taken from (7)

(a) Husky classified as wolf (b) Explanation

(7) M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

Black-box vs. interpretable classifiers

- Black-box
  - Neural Networks (DNN, RNN, CNN)
  - Ensemble methods
    - Random Forests
  - Support Vector Machines

- Interpretable
  - Decision Trees
  - Classification Rules
    - If-then rules
    - *m*-of-*n* rules
    - Lists of rules
    - Falling rule lists
    - Decision sets
  - Prototype-based methods

Black-box vs. interpretable classifiers

- Black-box
  - Neural Networks (DNN, RNN, CNN)
  - Ensemble methods
    - Random Forests
  - Support Vector Machines

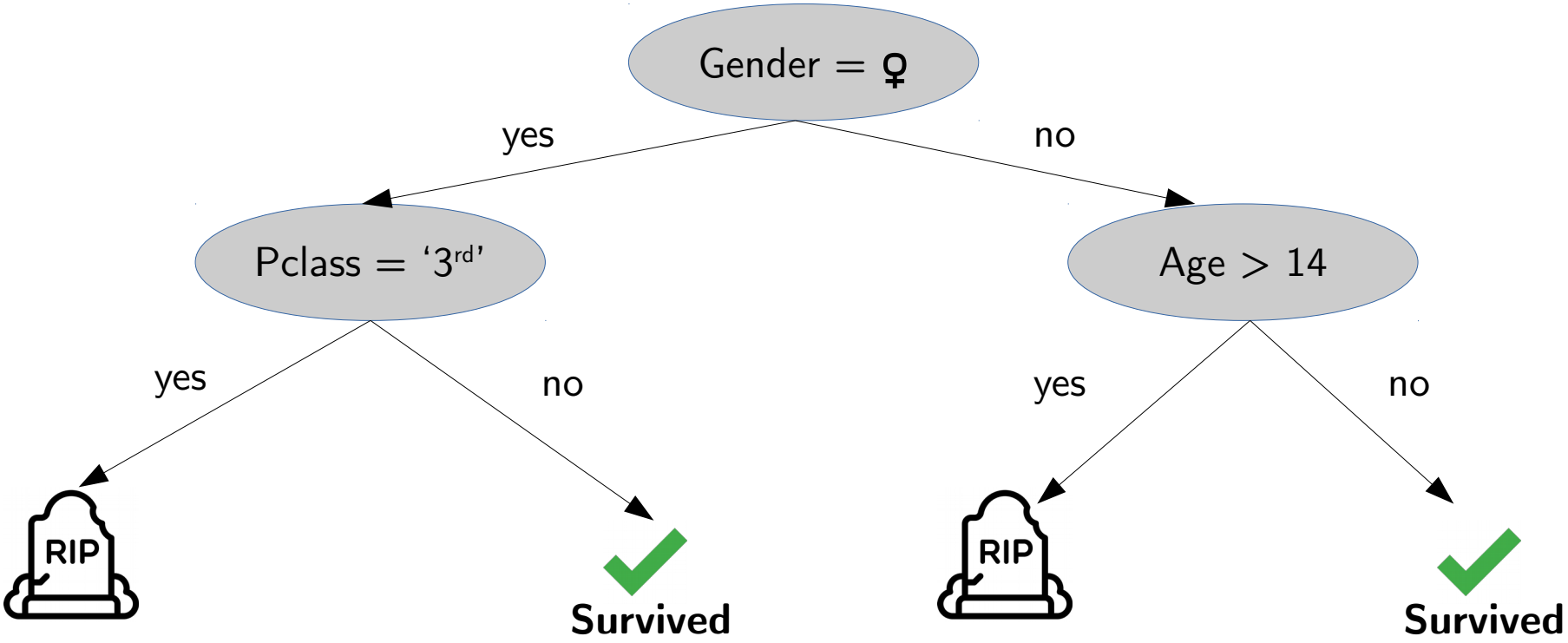
Accurate but not intepretable

- Interpretable
  - Decision Trees
  - Classification Rules
    - If-then rules
    - *m*-of-*n* rules
    - Lists of rules
    - Falling rule lists
    - Decision sets
  - Prototype-based methods

Simpler but less accurate

Black-box vs. **interpretable** classifiers

Decision Trees (CART, ID3, C4.5)



Black-box vs. **interpretable** classifiers

- If-then Rules (OneR)
  - Past-Depression $\wedge$ Melancholy $\Rightarrow$ Depressed
- m-of-n rules
  - Predict a class if at least *m* out of *n* attributes are present
    - If 2-of-{Past-Depression, $\neg$Melancholy, $\neg$Insomnia} $\Rightarrow$ Healthy
- Decision Lists (CPAR, RIPPER, Bayesian RL)
  - CPAR: Select the top-k rules for each class, and predict the class with the rule set of highest expected accuracy
  - Bayesian RL: Learn rules, select a list of rules with maximal posterior probability

Black-box vs. **interpretable** classifiers

- Falling rule lists

Falling Rule Lists

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

- Decision sets

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

Black-box vs. **interpretable** classifiers

- Prototype-based methods
  - Predict a class and provide a prototypical instance labeled with the same class
  - Challenge: pick a set of prototypes per class such that
    - The set is of minimal size
    - It provides full coverage, i.e., every instance should have a close prototype
    - They are far from instances of other classes
  - (a) formulates prototype selection as an optimization problem and uses it to classify images of handwritten digits

(a) J. Bien and R. Tibshirani. Prototype selection for interpretable classication. The Annals of Applied Statistics, pages 2403-2424, 2011.

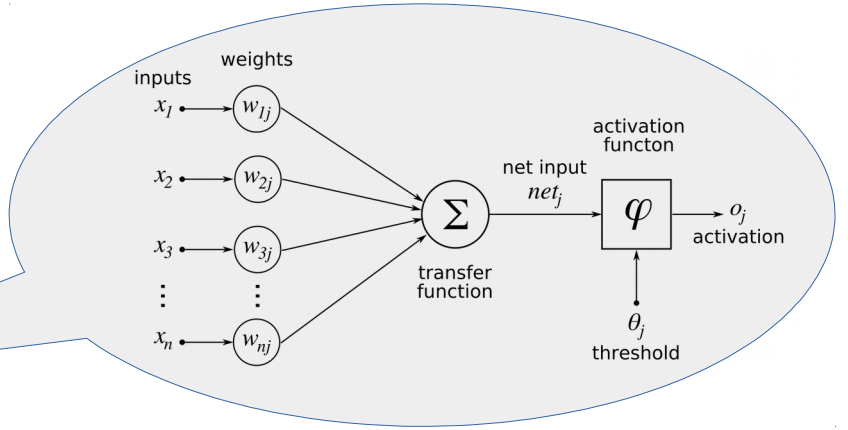
Black-box vs. **interpretable** classifiers

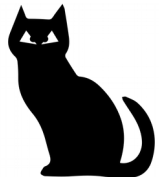
- Prototype-based methods

FIG. 7. *The first 88 prototypes (out of 3,372) of the greedy solution. We perform MDS (R function* sammon*) on the tangent distances to visualize the prototypes in two dimensions. The size of each prototype is proportional to the log of the number of correct-class training images covered by this prototype.*

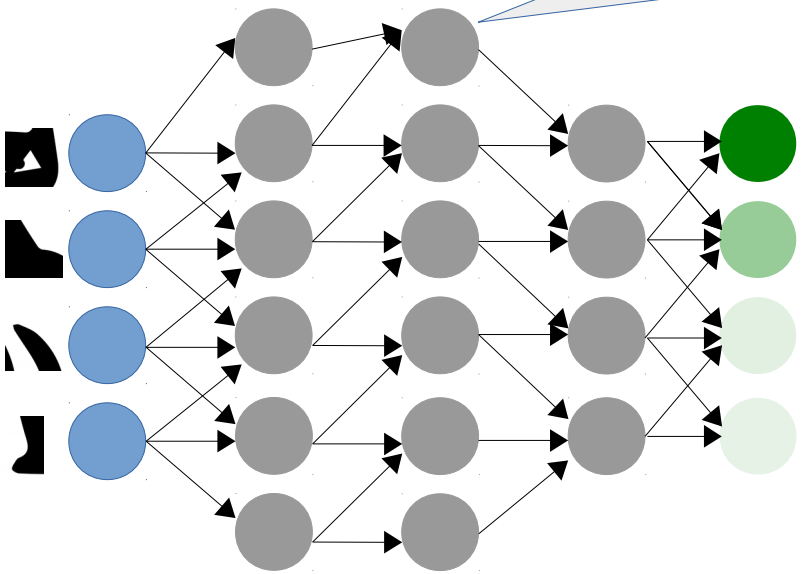
Black-box vs. interpretable classifiers

Neural Networks





Fish



Input Layer

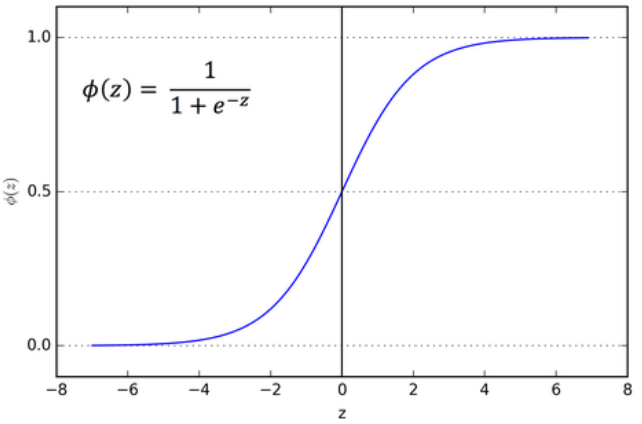
Output Layer

Cat

Dog

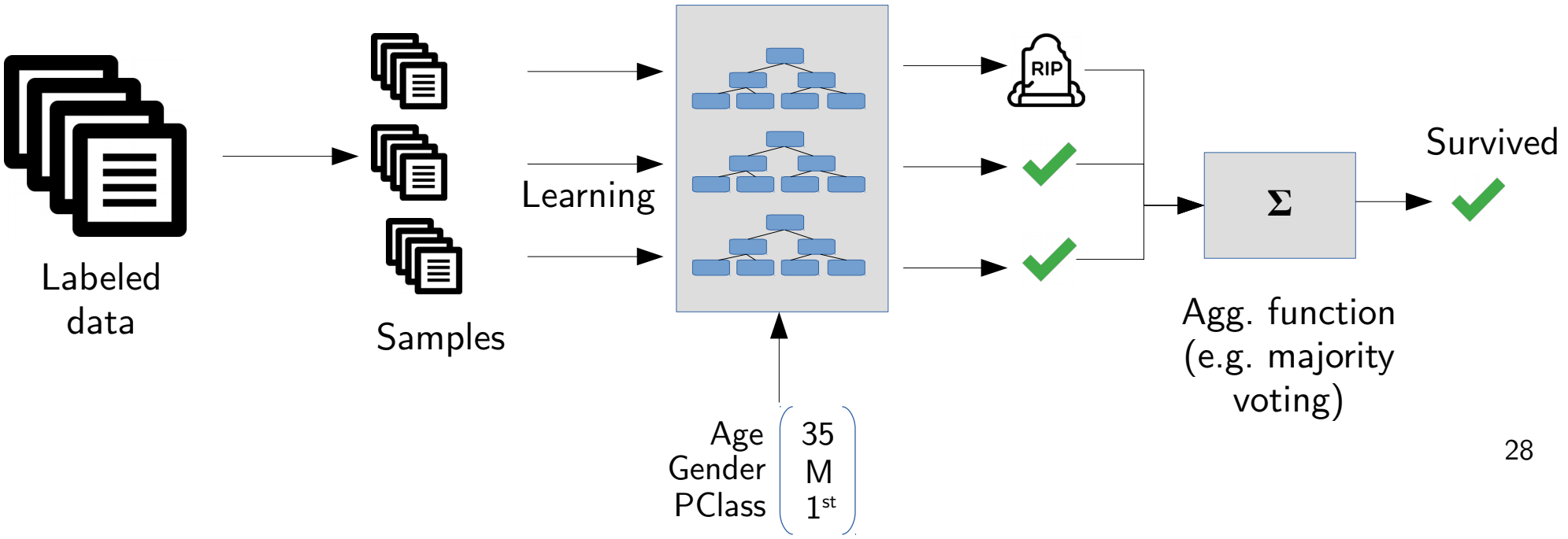
Fish

Koala



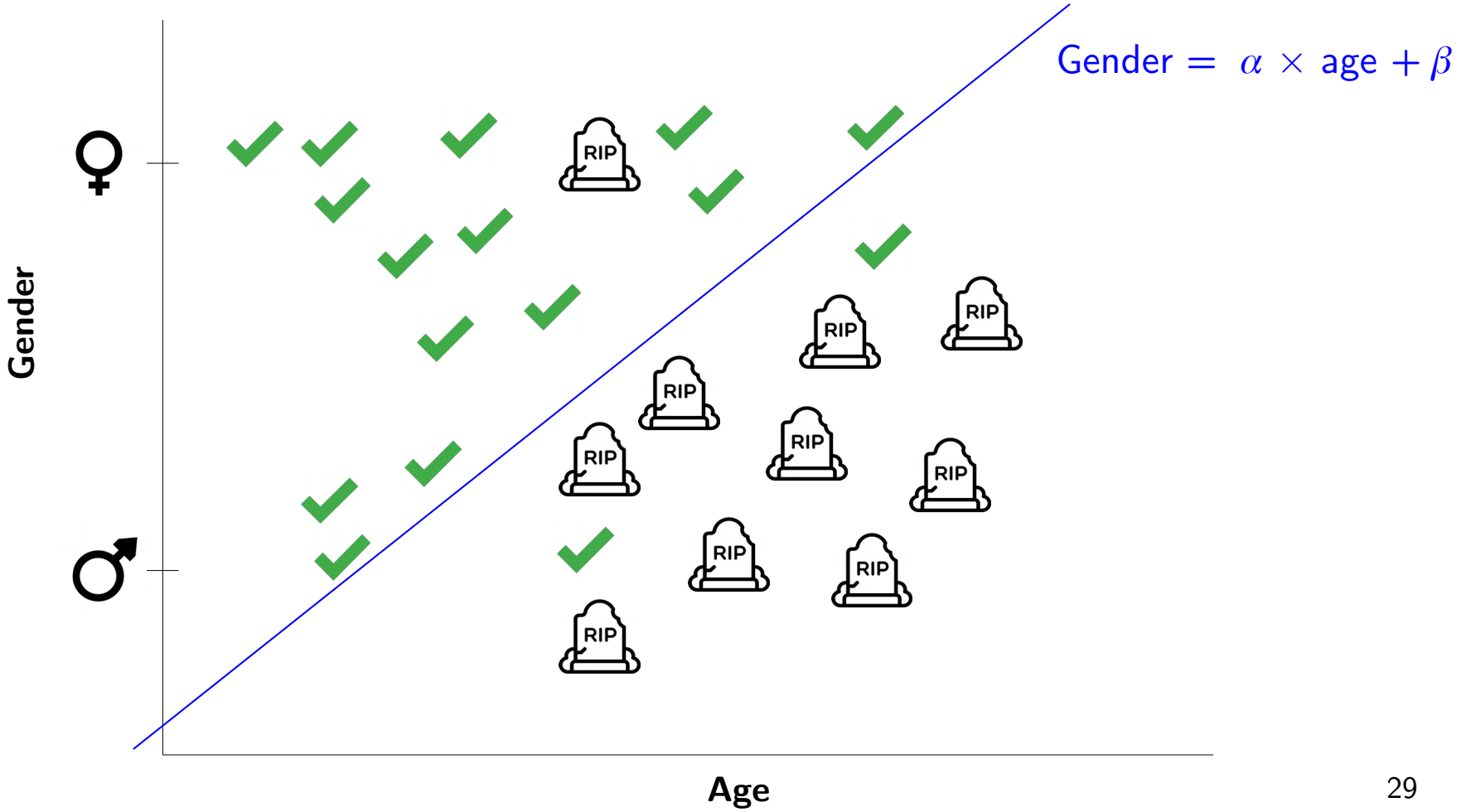
Black-box vs. interpretable classifiers

- Random Forests
  - Bagging: select *n* random samples (with replacement) and fit n decision trees.
  - Prediction: aggregate the decisions of the different trees to make a prediction



Black-box vs. interpretable classifiers

Support Vector Machines

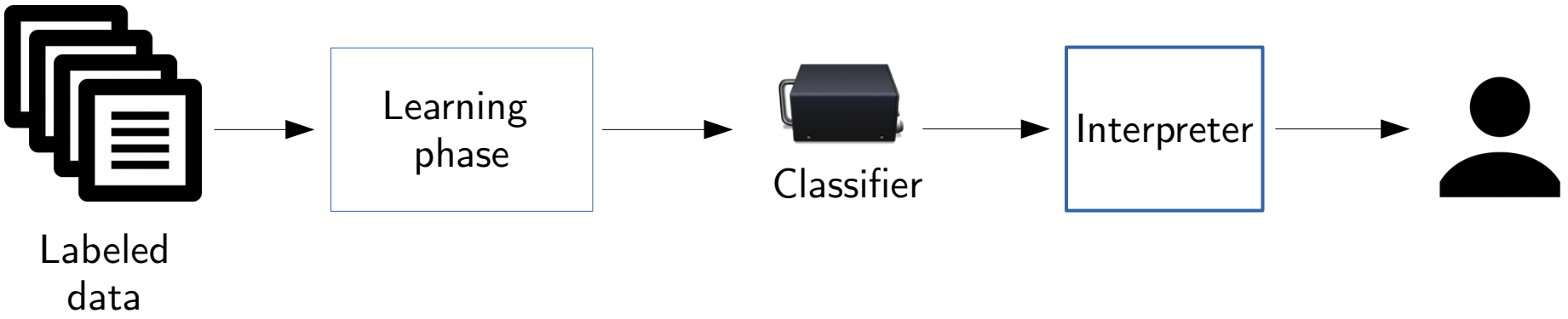


Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

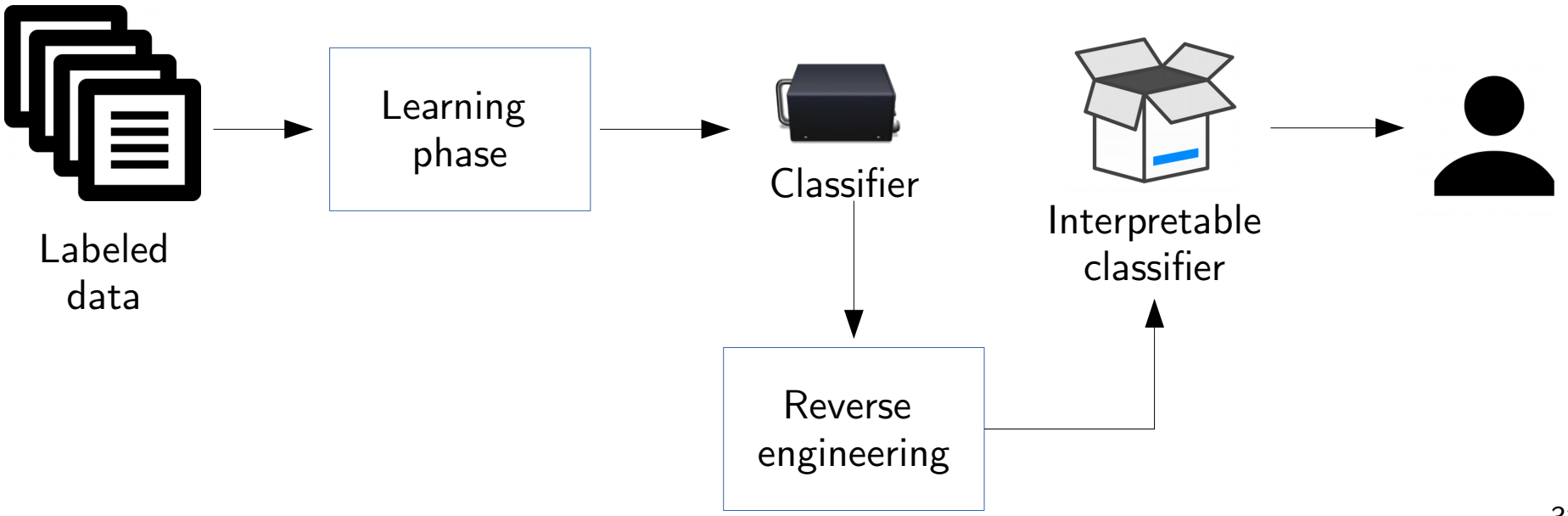
Explaining the black-box

Design an interpretation layer between the classifier and the human user



Explaining the black-box

Design an interpretation layer between the classifier and the human user



Explaining the black-box

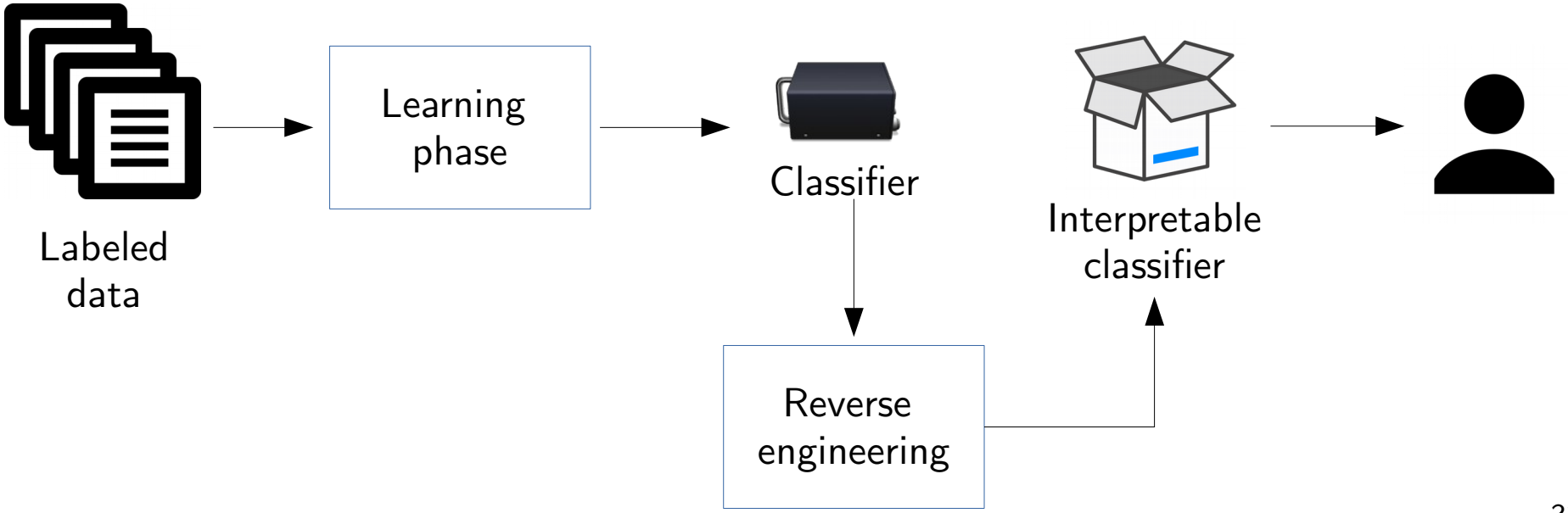
Design an interpretation layer between the classifier and the human user

Evaluation

$$accuracy(classifier, truth) = \frac{\#\text{ examples such that } classifier = truth}{\#\text{ all examples}}$$

$$fidelity = accuracy(interpretable\ classifier, classifier)$$

$$complexity = f(interpretable\ classifier)$$



Explaining the black-box

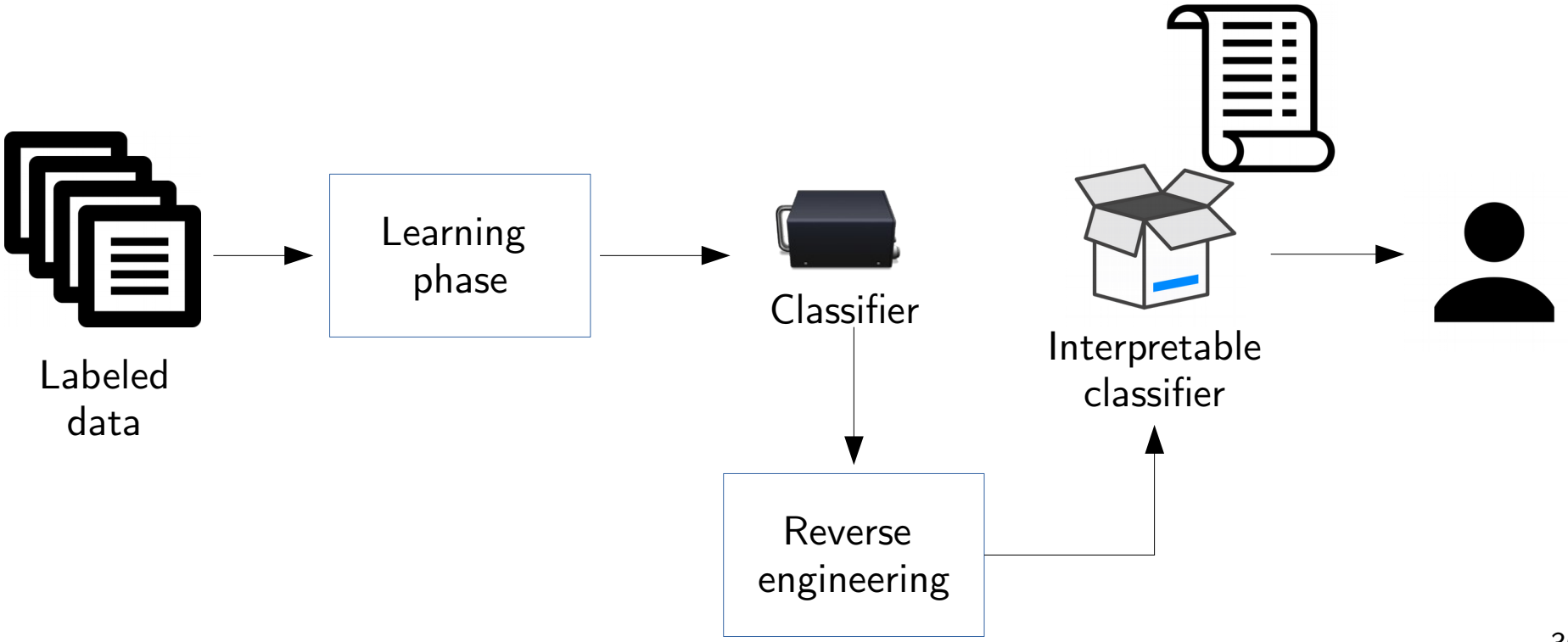
- Methods can be classified into three categories:
  - Methods for global explainability
  - Methods for local (outcome) explainability
  - Methods for classifier inspection
- Methods can be black-box dependent or black-box agnostic

Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
  - Global explainability
  - Local explainability
  - Classifier inspection
- Conclusion & open research questions

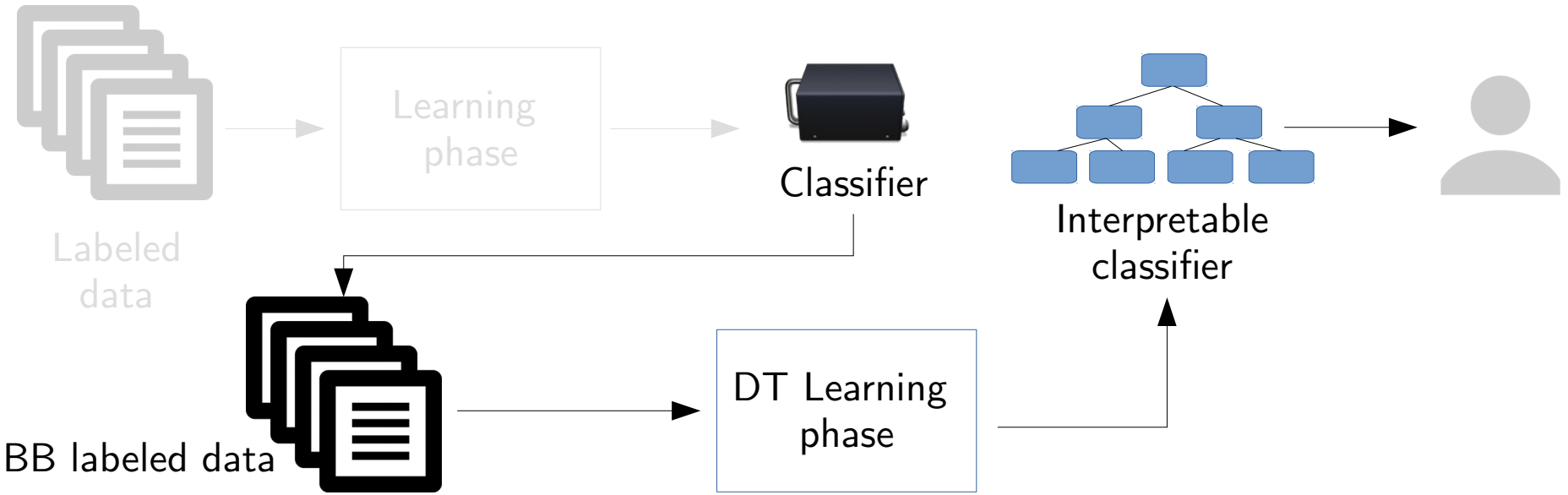
Global explainability

The interpretable approximation is a classifier that provides explanations for all possible outcomes



Global explainability

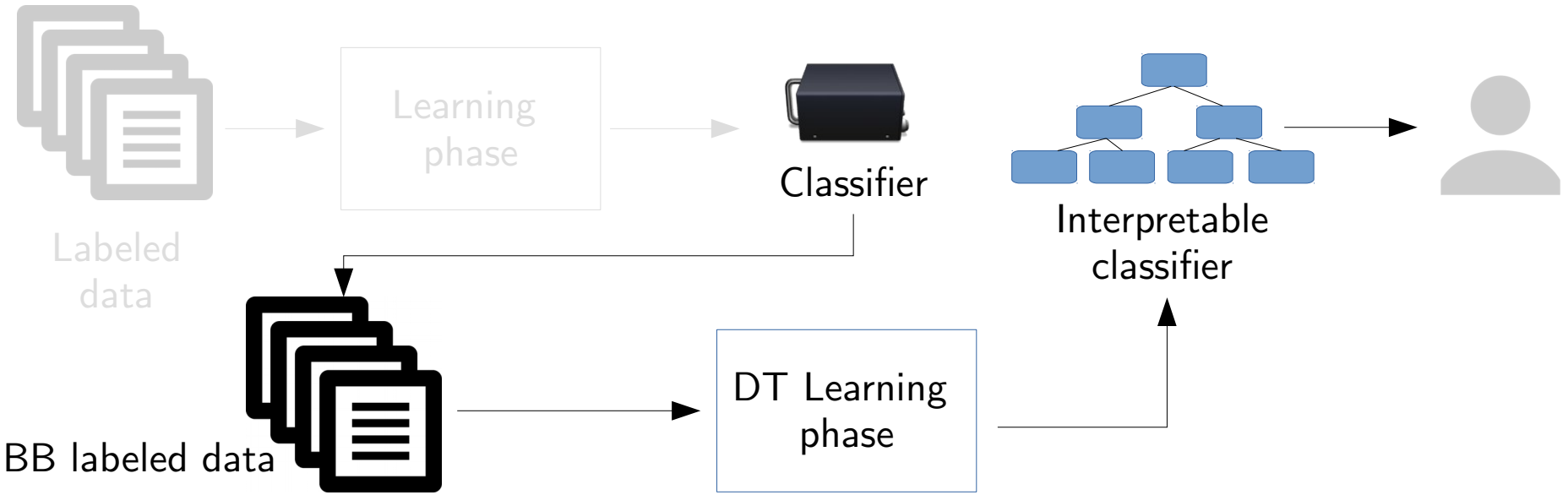
- Global explanations for NNs date back to the 90s
  - Trepan(1) is a black-box agnostic method that induces decision trees by querying the black box



(1) M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In Advances in neural information processing systems, pages 24-30, 1996.

Global explainability

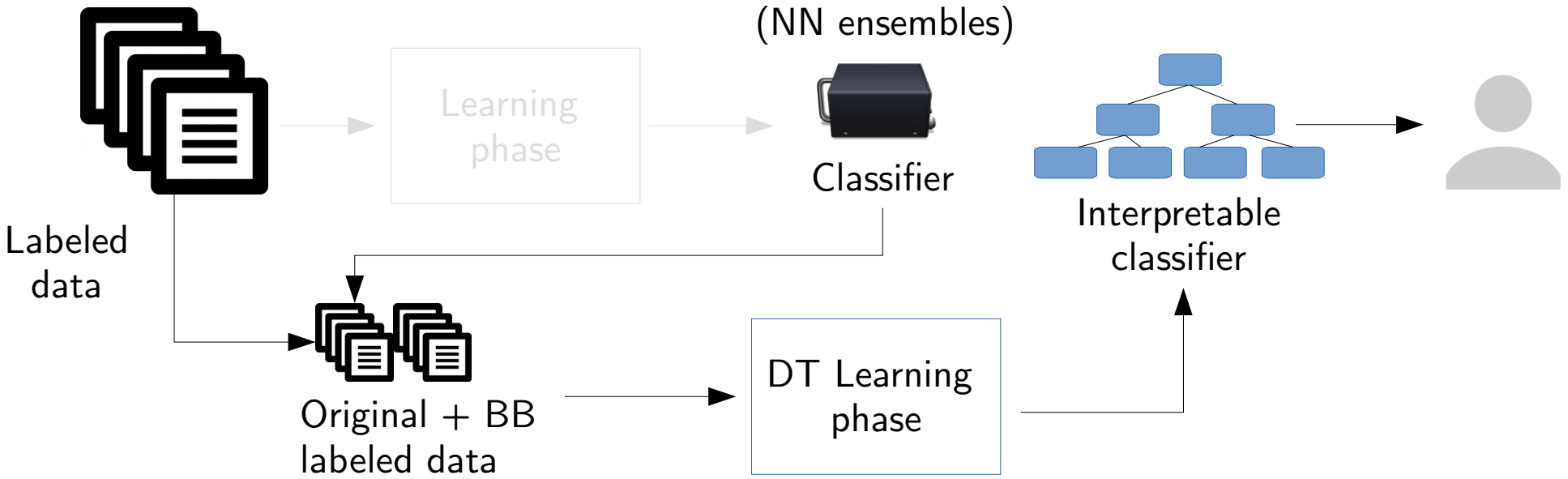
- Global explanations for NNs date back to the 90s
  - Trepan(1) is a black-box agnostic method that induces decision trees by querying the black box
  - Trepan's split criterion depends on entropy and fidelity



(1) M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In Advances in neural information processing systems, pages 24-30, 1996.

Global explainability

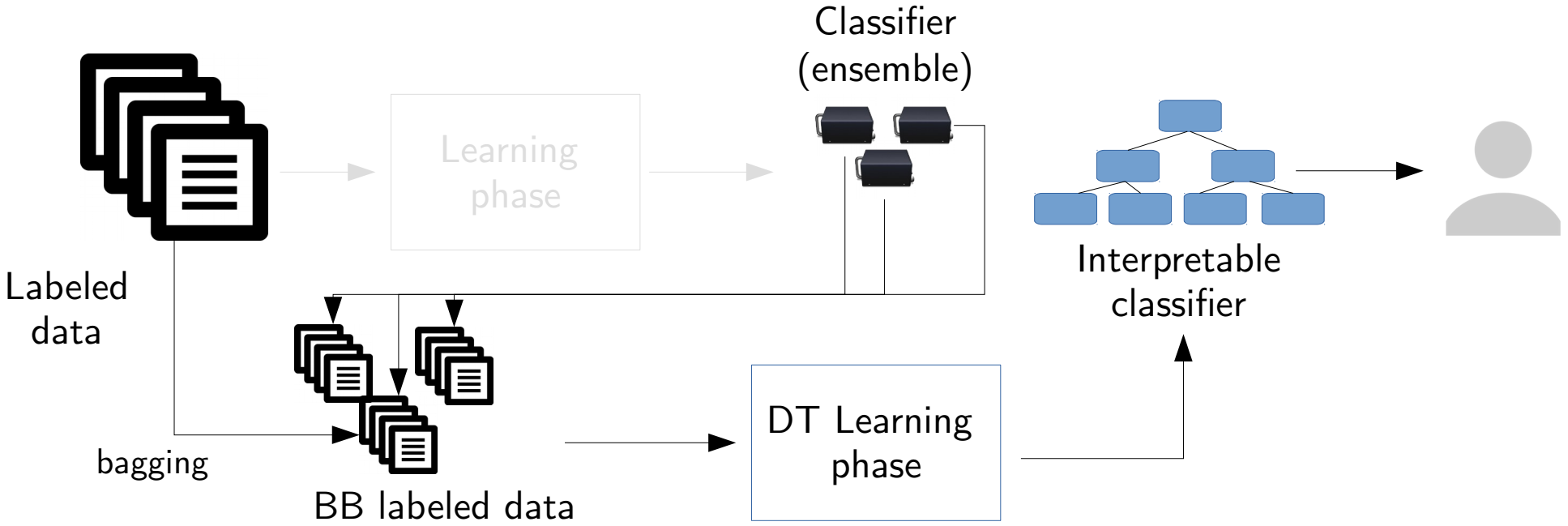
- (2) uses genetic programming to enhance explanations
  - By modifying the tree
  - By mixing the original and the BB labeled data



(2) U. Johansson and L. Niklasson. Evolving decision trees using oracle guides. In Computational Intelligence and Data Mining, 2009. CIDM'09., pages 238-244. IEEE, 2009.

Global explainability

- (3) uses notions of ensemble methods (*bagging*) to improve accuracy
  - Use BB-labeled data from different models



(3) P. Domingos. Knowledge discovery via multiple models. Intelligent Data Analysis, 2(1-4):187-202, 1998.

Global explainability

- Other methods generate sets of rules as explanations
  - (4) learns m-of-n rules from the original data plus BB labeled data (NNs)

(4) M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In ICML, pages 37-45, 1994.
(5) H. Lakkaraju and E. Kamar and R. Caruana and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. https://arxiv.org/pdf/1707.01154.pdf

Global explainability

- Other methods generate sets of rules as explanations
  - (4) learns m-of-n rules from the original data plus BB labeled data (NNs)
  - BETA(5) applies reverse engineering on the BB and then itemset mining to extract if-then rules.
    - Rules are restricted to two levels
    - If two contradictory rules apply to an example, the one with higher fidelity wins

If Age > 50 and Gender = Male Then
If Past-Depression = Yes and Insomnia = No and Melancholy = No ⇒ Healthy
If Past-Depression = Yes and Insomnia = No and Melancholy = Yes ⇒ Depressed

(4) M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In ICML, pages 37-45, 1994.

(5) H. Lakkaraju and E. Kamar and R. Caruana and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. https://arxiv.org/pdf/1707.01154.pdf

Global explainability

- BETA(5) applies reverse engineering on the BB and then itemset mining to extract if-then rules
  - Conditions (gender=♀) obtained via pattern mining
  - Rule selection formulated as an optimization problem

$$\arg\max_{\mathcal{R}\subseteq\mathcal{ND}\times\mathcal{DL}\times\mathcal{C}} \sum_{i=1}^{5} \lambda_i f_i(\mathcal{R}) \quad (1)$$

$$\text{s.t. } size(\mathcal{R}) \leq \epsilon_1, maxwidth(\mathcal{R}) \leq \epsilon_2, numdsets(\mathcal{R}) \leq \epsilon_3 \quad (2)$$

$$f_1(\mathcal{R}) = \mathcal{P}_{max} - numpreds(\mathcal{R}), \text{where } \mathcal{P}_{max} = \mathcal{P}_{max} = 2 * \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_2(\mathcal{R}) = \mathcal{O}_{max} - featureoverlap(\mathcal{R}), \text{where } \mathcal{O}_{max} = \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_3(\mathcal{R}) = \mathcal{O'}_{max} - ruleoverlap(\mathcal{R}), \text{where } \mathcal{O'}_{max} = N \times (|\mathcal{ND}| * |\mathcal{DL}|)^2$$

$$f_4(\mathcal{R}) = cover(\mathcal{R})$$

$$f_5(\mathcal{R}) = \mathcal{F}_{max} - disagreement(\mathcal{R}), \text{where } \mathcal{F}_{max} = N \times |\mathcal{ND}| * |\mathcal{DL}|$$

(4) M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In ICML, pages 37-45, 1994.

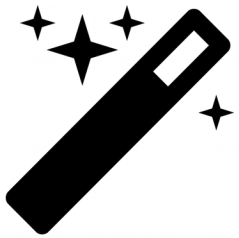
(5) H. Lakkaraju and E. Kamar and R. Caruana and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. https://arxiv.org/pdf/1707.01154.pdf

Global explainability

- BETA(5) applies reverse engineering on the BB and then itemset mining to extract if-then rules
  - Conditions (gender=♀) obtained via pattern mining
  - Rule selection formulated as an optimization problem

$$\arg\max_{\mathcal{R}\subseteq\mathcal{ND}\times\mathcal{DL}\times\mathcal{C}} \sum_{i=1}^{5} \lambda_i f_i(\mathcal{R}) \quad (1)$$

$$\text{s.t. } size(\mathcal{R}) \leq \epsilon_1, maxwidth(\mathcal{R}) \leq \epsilon_2, numdsets(\mathcal{R}) \leq \epsilon_3 \quad (2)$$



This is sorcery!!!

$$f_1(\mathcal{R}) = \mathcal{P}_{max} - numpreds(\mathcal{R}), \text{where } \mathcal{P}_{max} = \mathcal{P}_{max} = 2 * \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_2(\mathcal{R}) = \mathcal{O}_{max} - featureoverlap(\mathcal{R}), \text{where } \mathcal{O}_{max} = \mathcal{W}_{max} * |\mathcal{ND}| * |\mathcal{DL}|$$

$$f_3(\mathcal{R}) = \mathcal{O'}_{max} - ruleoverlap(\mathcal{R}), \text{where } \mathcal{O'}_{max} = N \times (|\mathcal{ND}| * |\mathcal{DL}|)^2$$

$$f_4(\mathcal{R}) = cover(\mathcal{R})$$

$$f_5(\mathcal{R}) = \mathcal{F}_{max} - disagreement(\mathcal{R}), \text{where } \mathcal{F}_{max} = N \times |\mathcal{ND}| * |\mathcal{DL}|$$

(4) M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In ICML, pages 37-45, 1994.

(5) H. Lakkaraju and E. Kamar and R. Caruana and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. https://arxiv.org/pdf/1707.01154.pdf

Global explainability

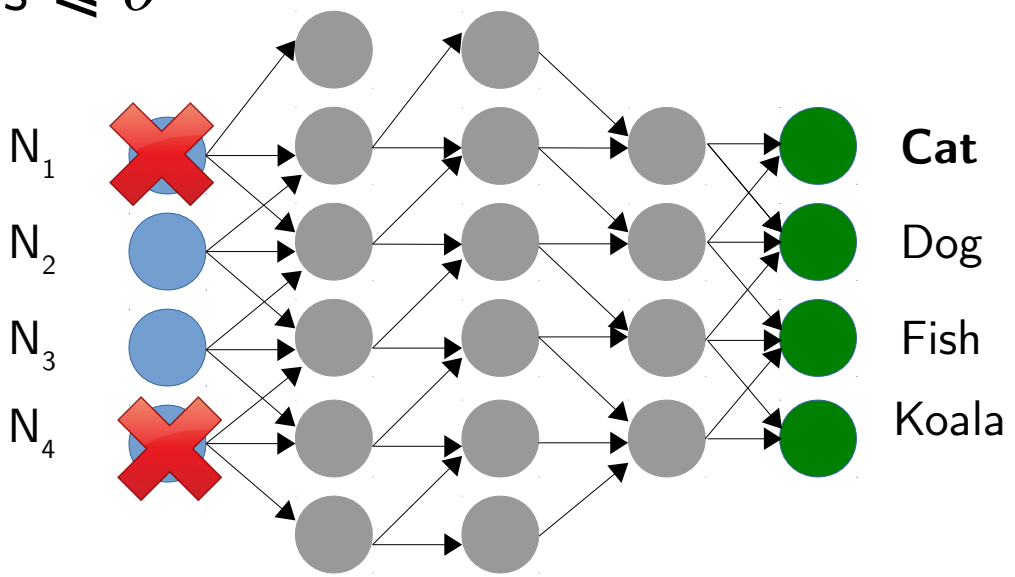
- BETA(5) applies reverse engineering on the BB and then itemset mining to extract if-then rules
  - Conditions (gender=♀) obtained via pattern mining
  - Rule selection formulated as an optimization problem that
    - Maximizes fidelity and coverage
    - Minimizes rule, feature overlap, and complexity
    - Constrained by number of rules, maximum width, and number of first level conditions

(4) M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In ICML, pages 37-45, 1994.

(5) H. Lakkaraju and E. Kamar and R. Caruana and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. https://arxiv.org/pdf/1707.01154.pdf

Global explainability

- RxREN(6) learns rule-based explanations for NNs
  - First, it iteratively prunes *insignificant* input neurons while the accuracy loss is less than 1%
    - Store #errors caused by the removal of each neuron
    - #errors $\leqslant \theta$



(6) M. G. Augasta and T. Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. Neural processing letters, 35(2):131-150, 2012.

Global explainability

- RxREN(6) learns rule-based explanations for NNs
  - Second, build a matrix with [min, max] of the values of the remaining neurons when predicting a class.

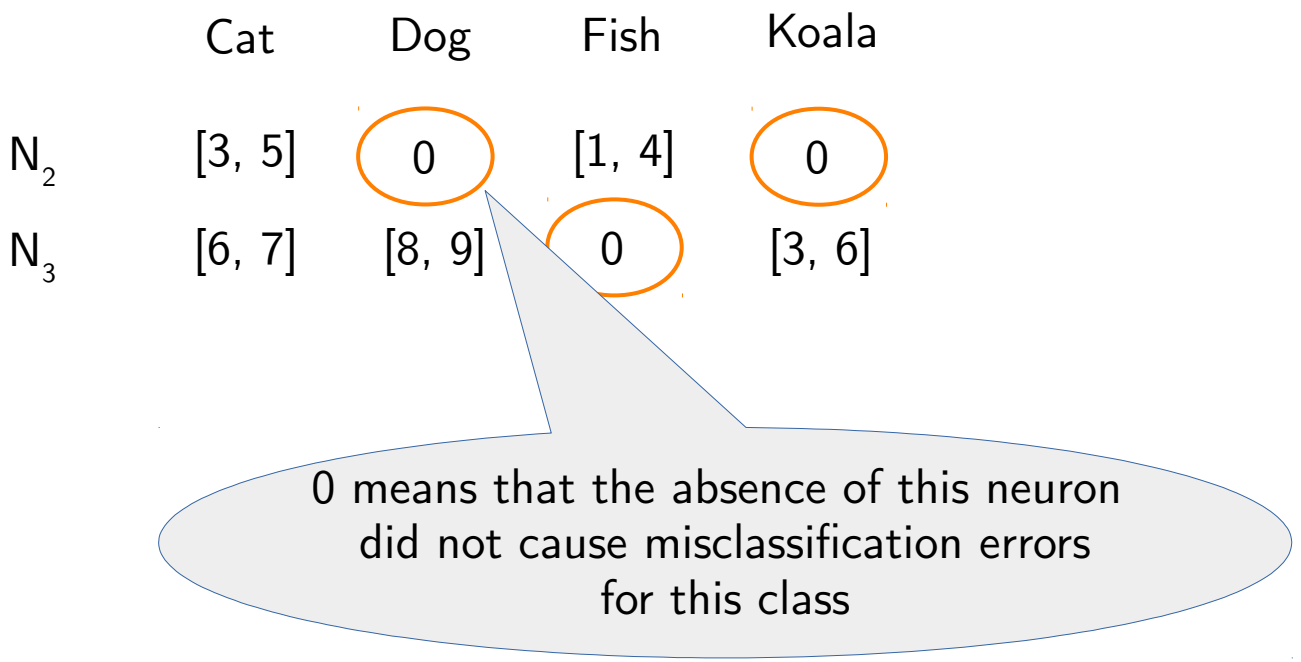
	Cat	Dog	Fish	Koala
N_2	[3, 5]	0	[1, 4]	0
N_3	[6, 7]	[8, 9]	0	[3, 6]

$$M[N_i][C_j] = \begin{cases} [\min(N_i \mid C_j), \max(N_i \mid C_j)] \text{ if } \#\text{errors} > \gamma \\ 0 \text{ otherwise} \end{cases}$$

(6) M. G. Augasta and T. Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. Neural processing letters, 35(2):131-150, 2012.

Global explainability

- RxREN(6) learns rule-based explanations for NNs
  - Second, build a matrix with [min, max] of the values of the remaining neurons when predicting a class.



(6) M. G. Augasta and T. Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. Neural processing letters, 35(2):131-150, 2012.

Global explainability

- RxREN(6) learns rule-based explanations for NNs
  - Third, learn rules from the matrix
  - Sort classes by significance (#non-zero entries)

	Cat	Dog	Fish	Koala
N_2	[3, 5]	0	[1, 4]	0
N_3	[6, 7]	[8, 9]	0	[3, 6]

If $N_2 \in [3, 5] \wedge N_3 \in [6, 7] \Rightarrow$ Cat
Else If $N_2 \in [8, 9] \Rightarrow$ Dog
Else ....

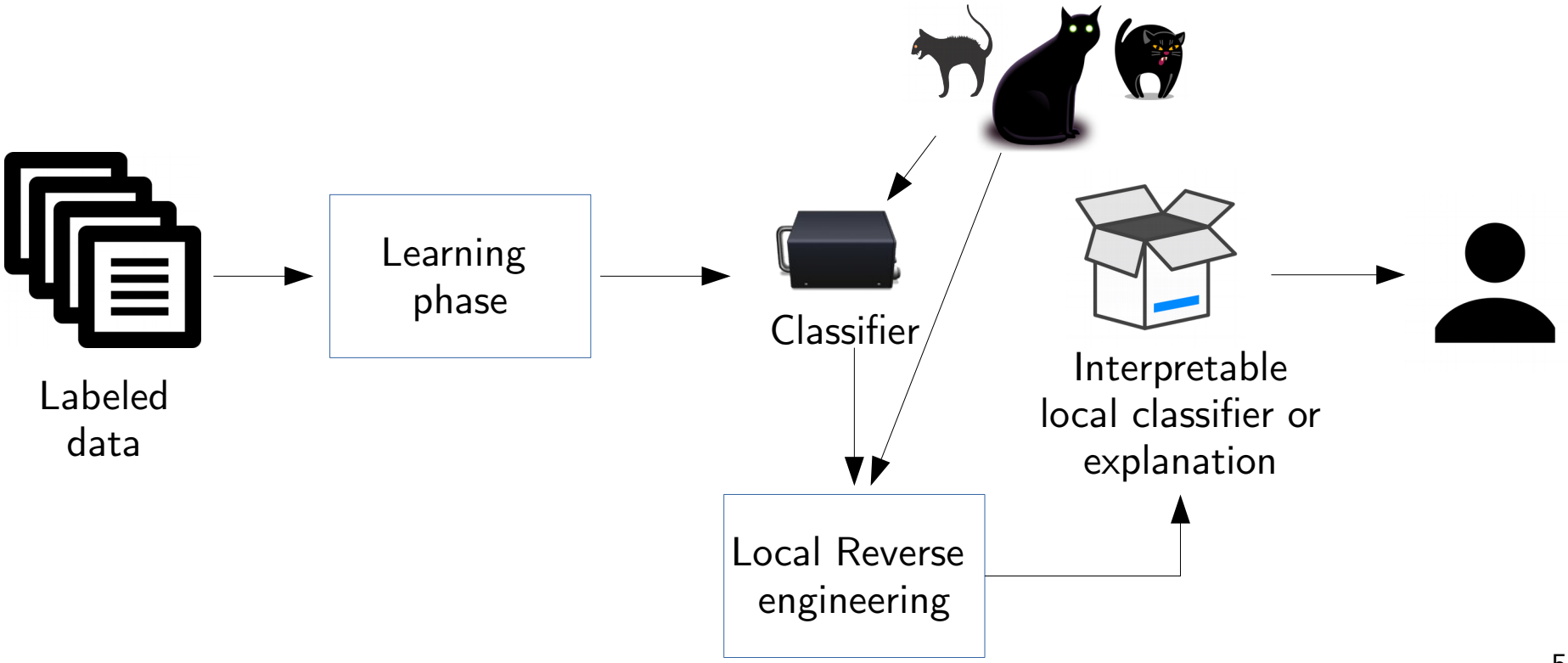
(6) M. G. Augasta and T. Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. Neural processing letters, 35(2):131-150, 2012.

Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
  - Global explainability
  - Local explainability
  - Classifier inspection
- Conclusion & open research questions

Local explainability

The interpretable approximation is a classifier that provides explanations for the answers of the black box in the vicinity of an individual instance.



Local explainability

- LIME[7] is BB-agnostic and optimizes for local fidelity
  - First, write examples in an *interpretable* way

Original Image

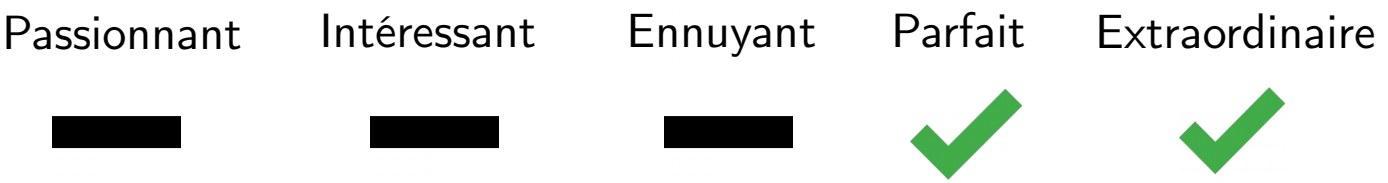
Interpretable Components

(7) M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

Local explainability

- LIME[7] is BB-agnostic and optimizes for local fidelity
  - First, write examples in an *interpretable* way

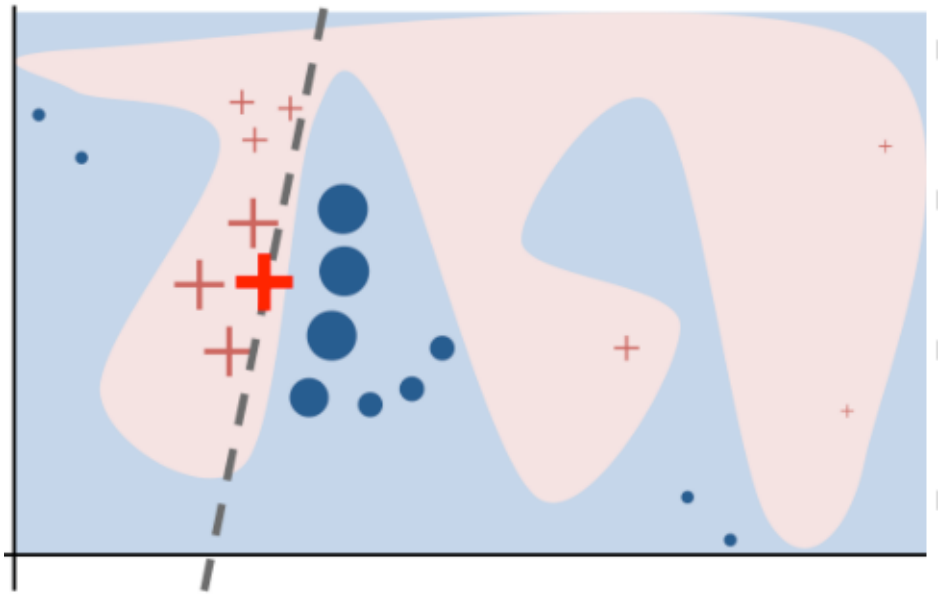
Délai de livraison parfait très bon état du livre en ce qui concerne le bouquin en lui-même c'est extraordinaire . un point de vue sur l'histoire de l'humanité qui fait voir les choses sous un nouvel angle



(7) M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

Local explainability

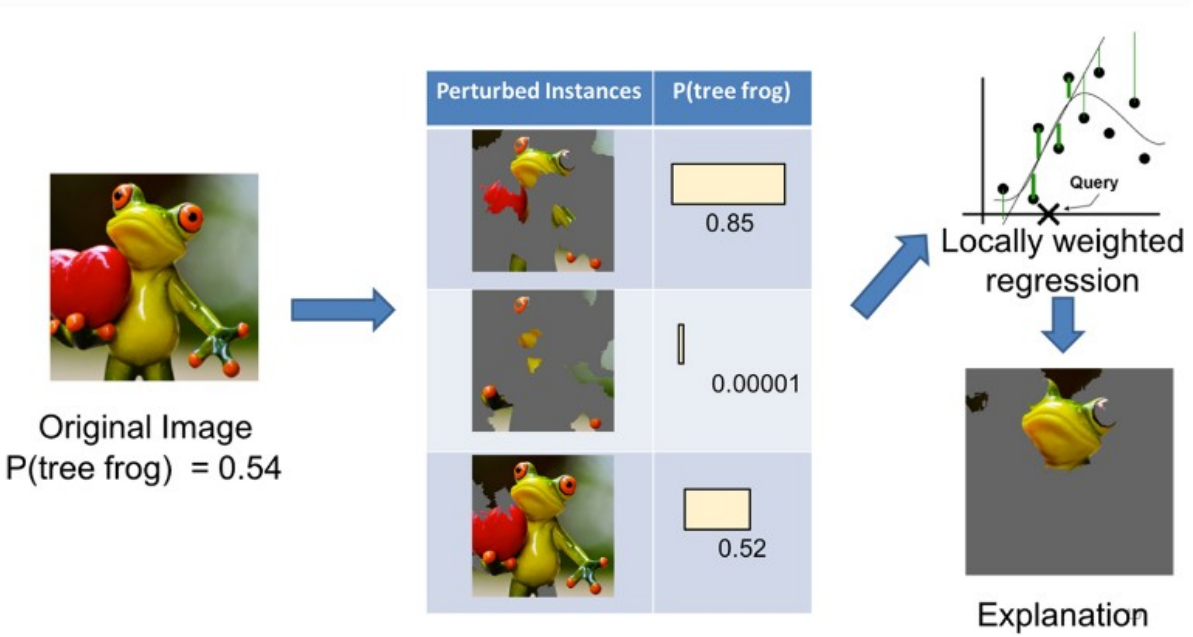
- LIME[7] is BB-agnostic and optimizes for local fidelity
  - Then, learn a linear model from the interpretable examples + their BB labels in the vicinity of the given instance.



(7) M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

Local explainability

- LIME[7] is BB-agnostic and optimizes for local fidelity
  - Then, learn a linear model from the interpretable examples + their BB labels in the vicinity of the given instance.



(7) M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

Local explainability

- SHAP(9) is BB-agnostic and it uses additive feature attribution to quantify feature importance

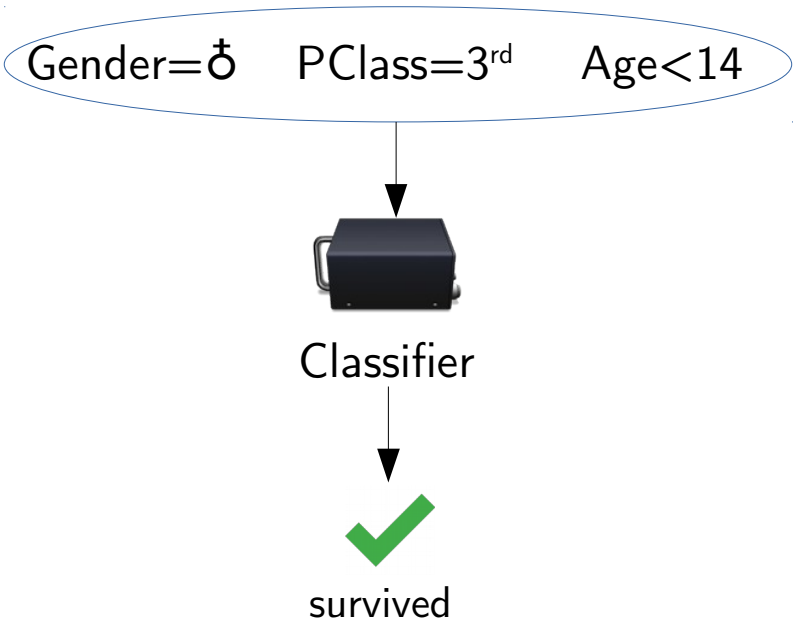
(9) Lundberg, Scott M. and Su-In Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017.

Local explainability

- SHAP(9) is BB-agnostic and it uses additive feature attribution to quantify feature importance
  - *Shapley values:* averages of feature influence on **all possible feature coalitions (reduced models)**

Influence of Gender=♂

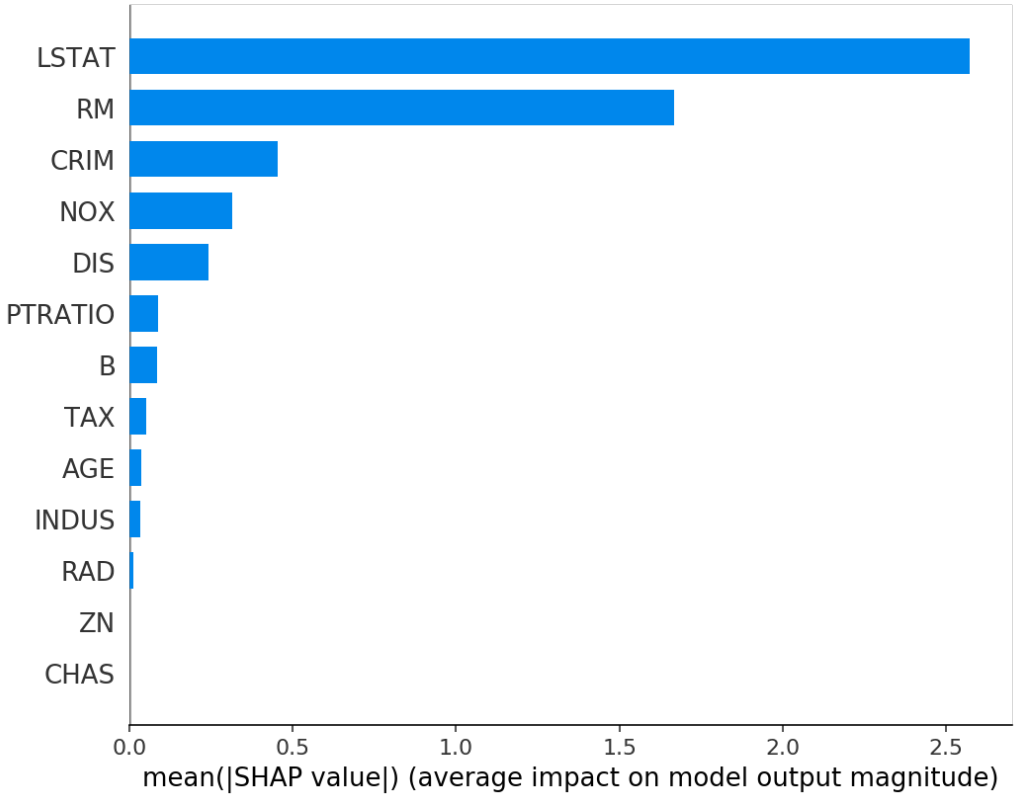
				Difference
G=♂	✓	∅	RIP	1
PClass=3rd + G=♂	RIP	PClass=3rd	RIP	0
Age<14 + G=♂	✓	Age<14	✓	0
Age< 14 + PClass=3rd+ G=♂	✓	PClass=3rd + Age<14	RIP	1



(9) Lundberg, Scott M. and Su-In Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017.

Local explainability

- SHAP(9) is BB-agnostic and it uses additive feature attribution to quantify feature importance



(9) Lundberg, Scott M. and Su-In Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017.

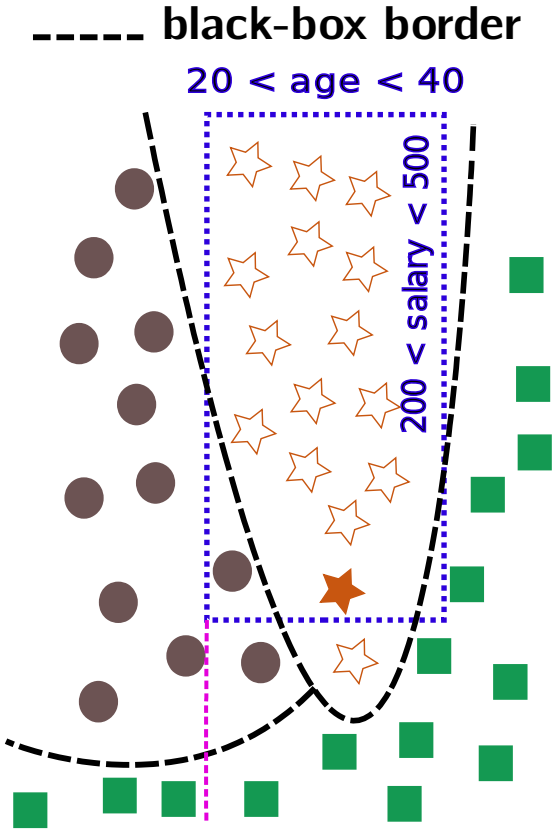
Local explainability

- SHAP[9] feature attribution model guarantees *local accuracy*, *missingness* and *consistency*
  - Generalization of LIME
  - Shapley values can be calculated via sampling (e.g., Montecarlo Sampling)
  - It offers some model-specific extensions such as DeepShap and TreeShap
  - It is written in Python and available at https://github.com/slundberg/shap

(9) Lundberg, Scott M. and Su-In Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017.

Local explainability

- Anchors[10] are regions of the feature space where a classifier behaves as with an instance of interest.



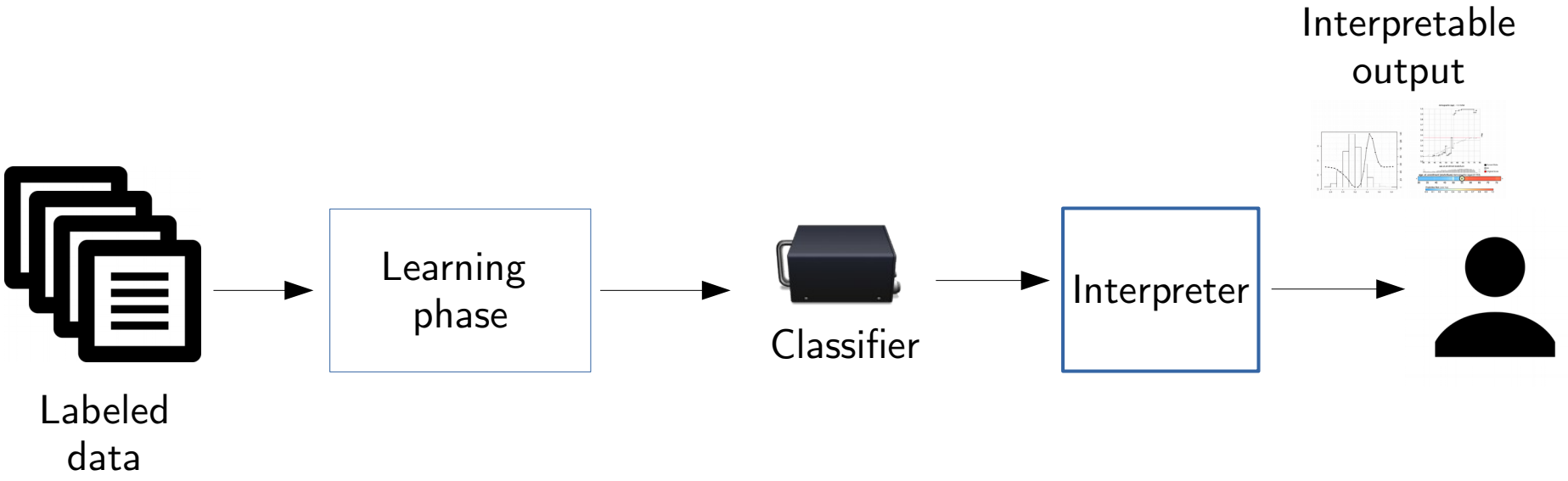
(10) Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018.

Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
  - Global explainability
  - Local explainability
  - Classifier inspection
- Conclusion & open research questions

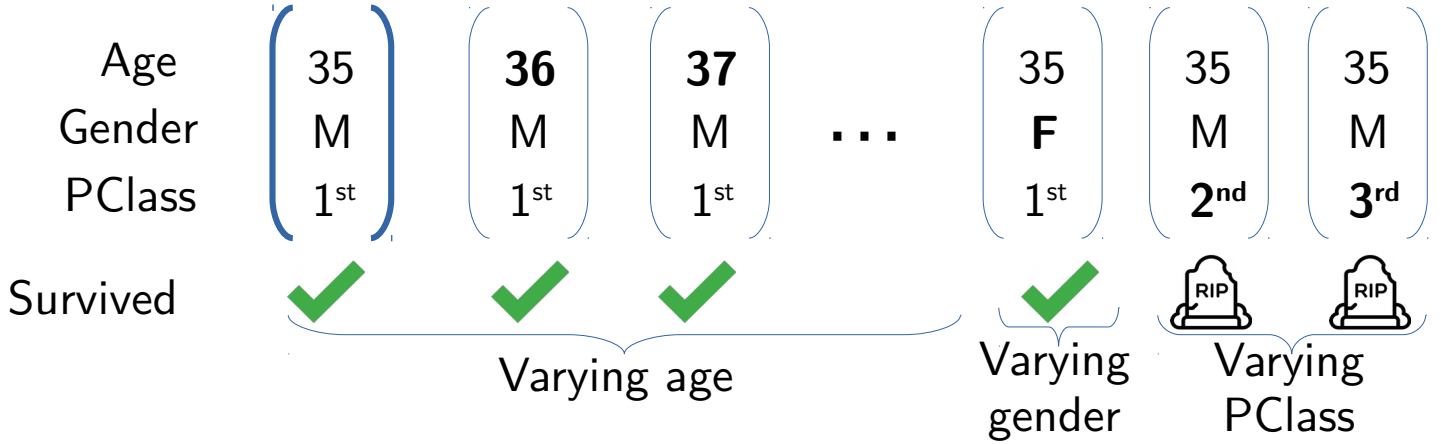
Inspecting the black box

The goal is to *plot* the *correlations* between the input features and the output classes



Inspecting the black box

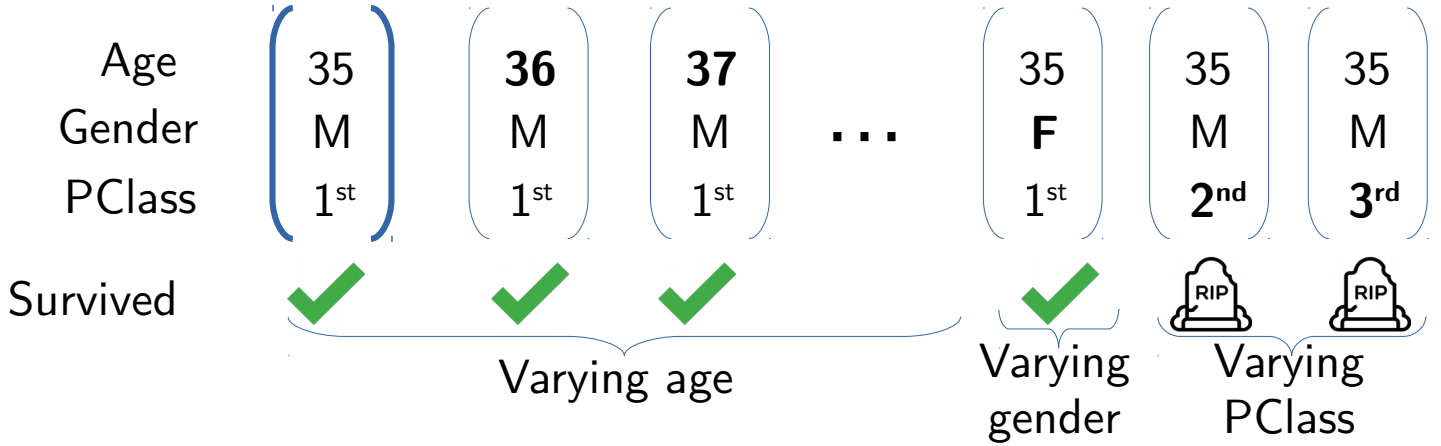
- Sensitivity analysis(11) explains the influence of the inputs on the classifier's output for each class
  - Build a prototype vector with the average/median/mode of the input attributes
  - Vary each attribute value, apply the classifier



(11) P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, pages 341-348. IEEE, 2011.

Inspecting the black box

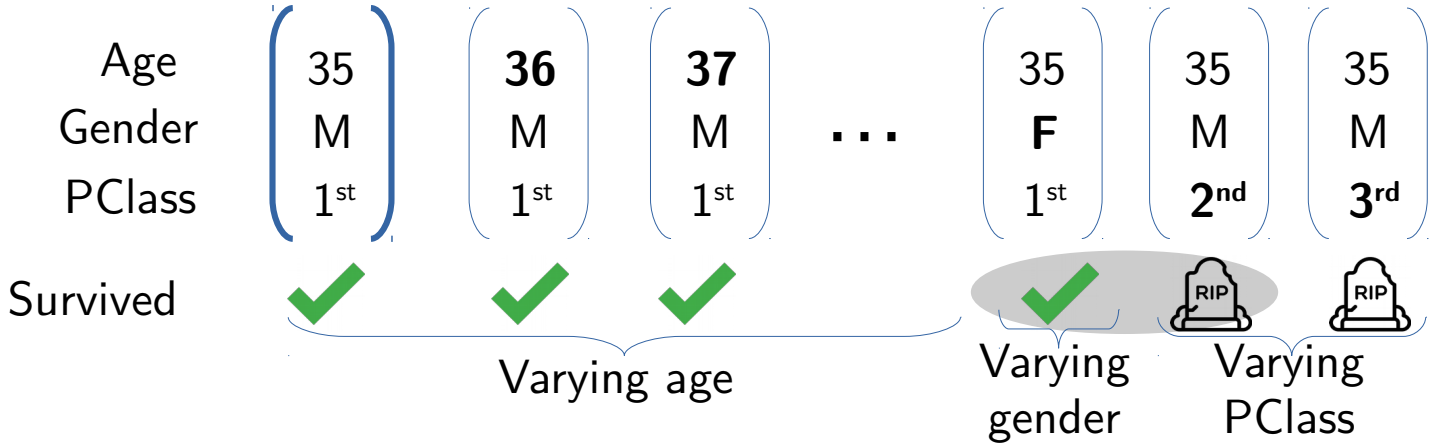
- Sensitivity analysis[11] explains the influence of the inputs on the classifier's output for each class
  - Express each output class as a binary variable
  - Compute metrics for each attribute: range, gradient, variance, importance



(11) P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, pages 341-348. IEEE, 2011.

Inspecting the black box

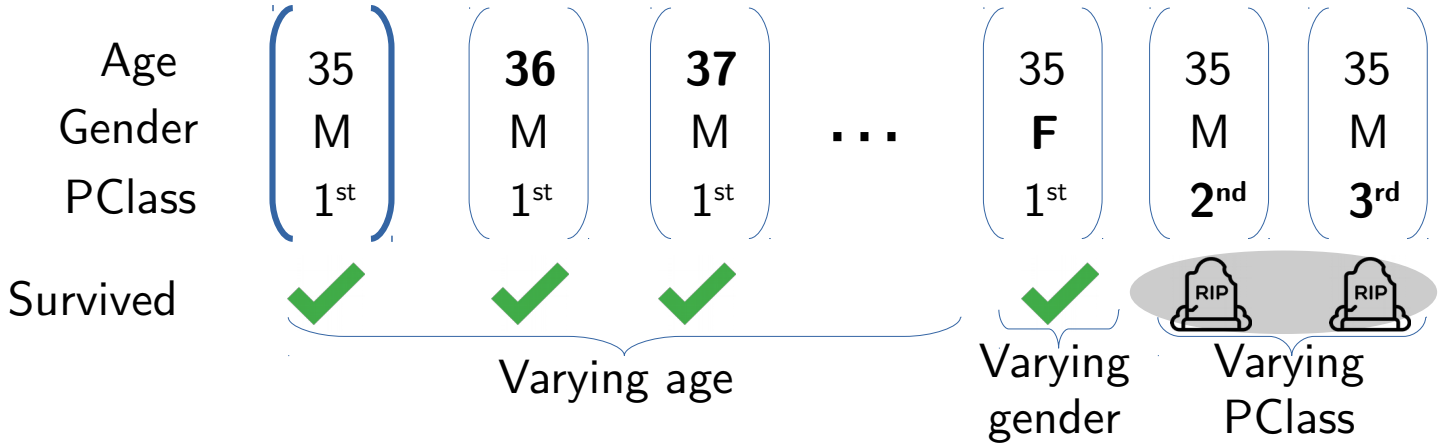
- Sensitivity analysis[11] explains the influence of the inputs on the classifier's output for each class
  - $\text{range}_{\text{class}=\checkmark}(\text{Age}) = 0$, $\text{range}_{\text{class}=\checkmark}(\text{PClass}) = 1$
  - $\text{gradient}_{\text{class}=\checkmark}(\text{PClass}) = (\mathbf{1} + 0)/2 = 0.5$



(11) P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, pages 341-348. IEEE, 2011.

Inspecting the black box

- Sensitivity analysis[11] explains the influence of the inputs on the classifier's output for each class
  - $\text{range}_{\text{class}=\checkmark}(\text{Age}) = 0$, $\text{range}_{\text{class}=\checkmark}(\text{PClass}) = 1$
  - $\text{gradient}_{\text{class}=\checkmark}(\text{PClass}) = (1 + \mathbf{0})/2 = 0.5$

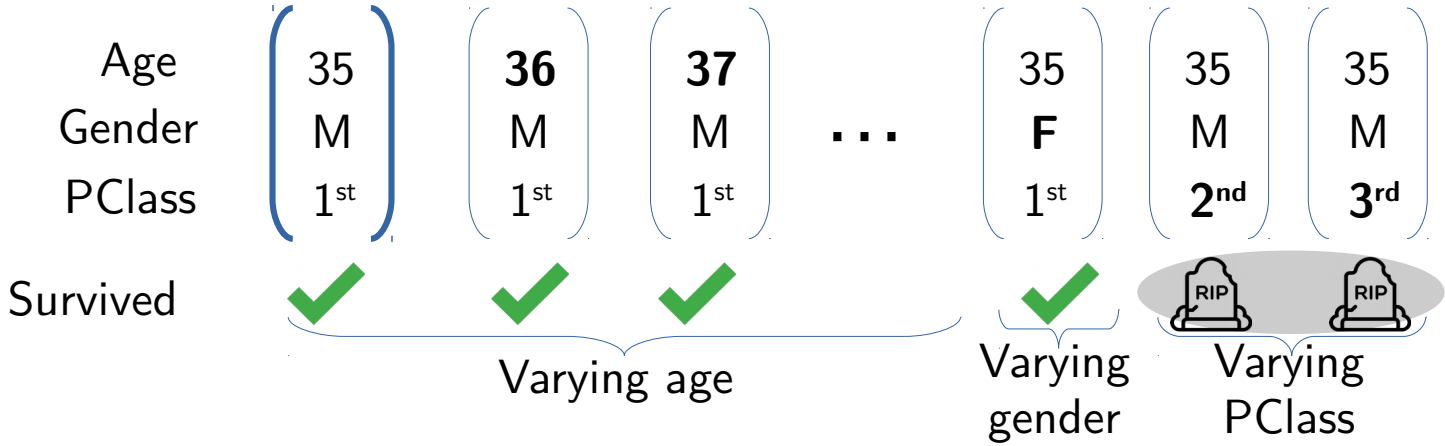


(11) P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, pages 341-348. IEEE, 2011.

Inspecting the black box

- Sensitivity analysis[11] explains the influence of the inputs on the classifier's output
  - Importance of an input feature *a* according to metric *s*

$$R_a = s_a / \sum_{i=1}^{I} s_i \times 100\,(\%).$$



(11) P. Cortez and M. J. Embrechts. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, pages 341-348. IEEE, 2011.

Agenda

- Interpretability in classifiers: What and Why?
- Black-box vs. interpretable classifiers
- Explaining the black-box
- Conclusion & open research questions

Conclusion

- Interpretability in ML classifier matters
  - For human, ethical, legal, and technical reasons
- Interpretability has two dimensions: global & local
  - Global interpretability has been more studied in the past
  - The trend is moving towards local BB-agnostic explanations
- The key of opening the black box is *reverse engineering*

Open research questions

- Can we talk about *automatic interpretability*?
  - Are linear attribution models more interpretable than decision trees or rule lists?
  - How to account for users' background in the explanations?
- Interpretable simplified spaces: how to give semantics to them?
  - Objects instead of superpixels
  - Entity mentions, verb/adverbial phrases instead of words

Other sources

- A Survey Of Methods For Explaining Black Box Models, https://arxiv.org/pdf/1802.01933.pdf
- Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/