2018 한국소프트웨어종합학술대회 (Korea Software Congress 2018)
2018년 12월 21일 (금) 09:00-12:00

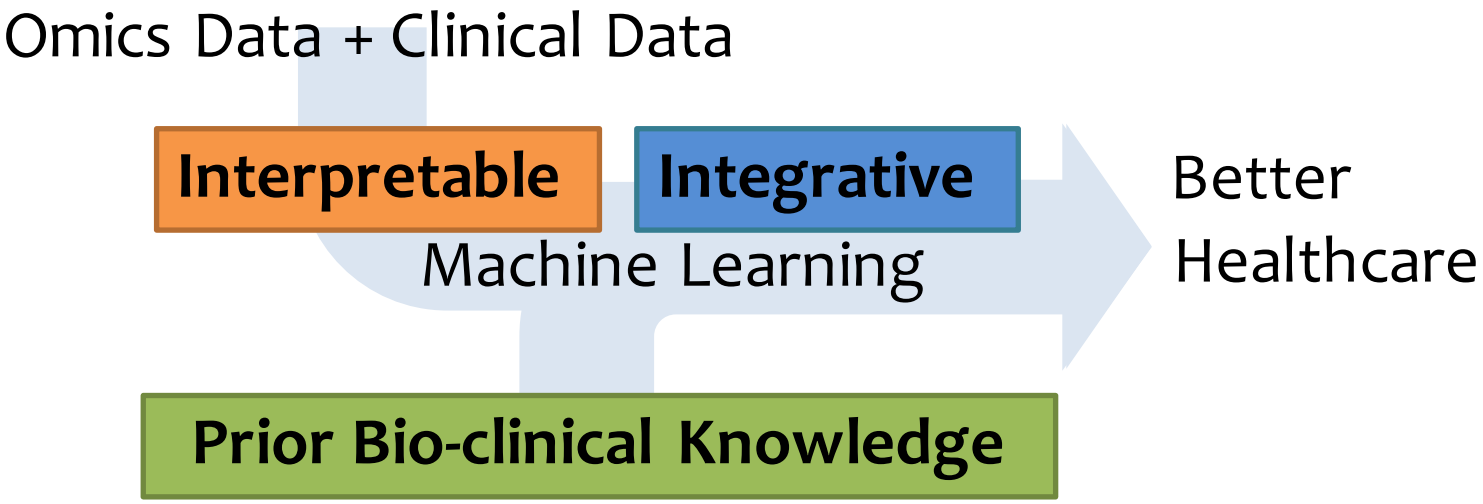
An Introduction to Interpretable Machine Learning

Sael Lee
BK Associate Professor
SNU Computer Science and Engineering
Seoul National University, Korea

My Research

Machine Learning for Bio-medical Data Analysis

- **Interpretable** Method
- **Integrative** Method
- Utilize **Prior Knowledge**



Overview

- Part 1: Introduction
  - What and why interpretable ML?
  - What is evaluated in interpretable ML?
  - How can we categorize interpretable ML?
- Part 2: Interpretable ML overview
  - Intrinsically interpretable models
  - Post-hoc interpretable methods
  - Transparent box models
- Part 3: Interpretable deep learning

Part 1: Introduction

- **What and why interpretable ML?**
- What is evaluated in interpretable ML?
- How can we categorize interpretable ML?

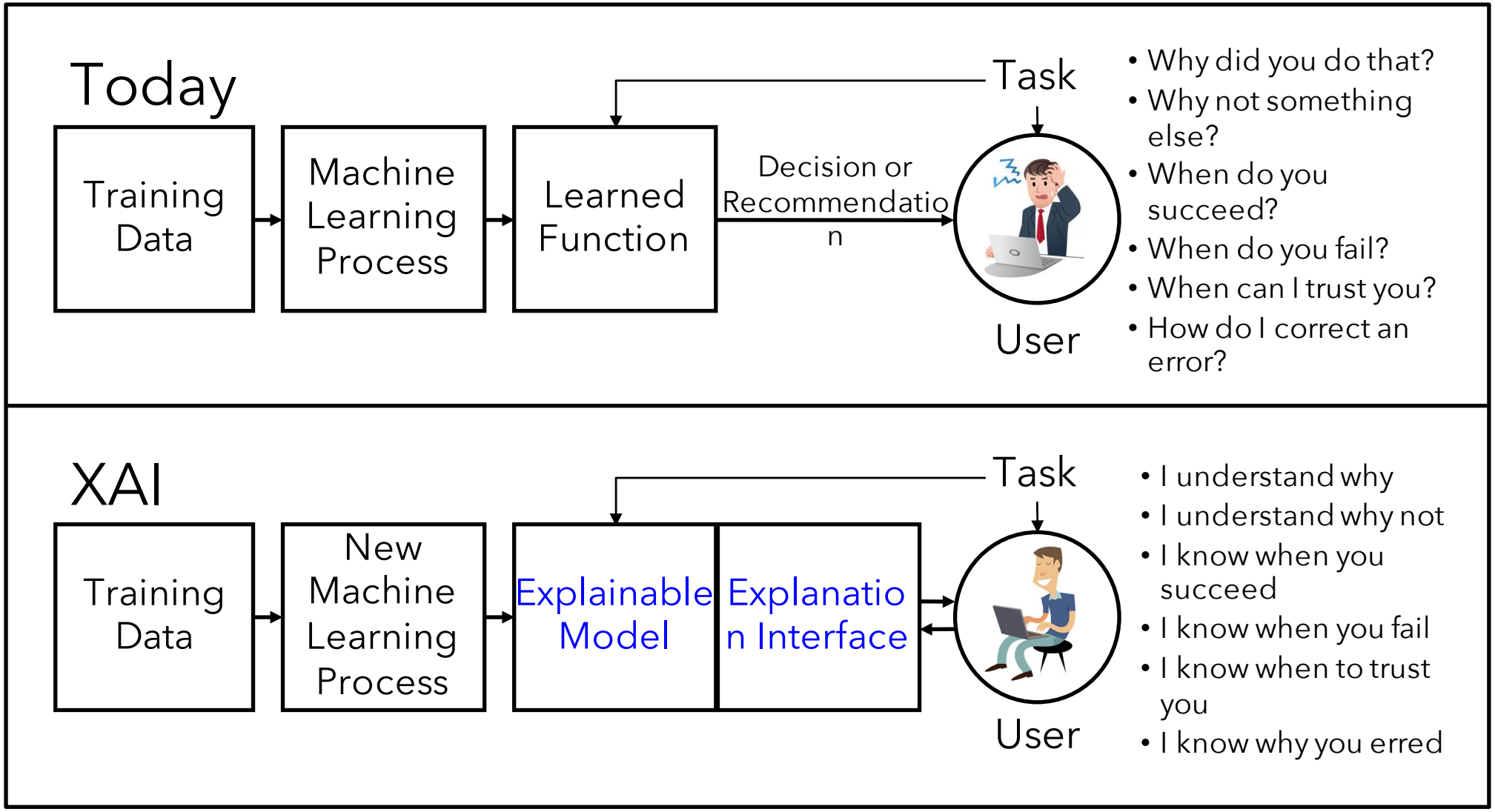
Most of the contents in Part 1 comes from

- "Interpretable Machine Learning *A Guide for Making Black Box Models Explainable." by Christoph Molnar 2018.* (https://christophm.github.io/interpretable-ml-book/)
- CVPR'18 Tutorial by W. Samek, G. Montavon and K.R. Müller [CVPR'18 Tutorial]

What is Interpretable Machine Learning?

- **"Interpretability is the degree to which a human can understand the cause of a decision."** (Miller 2017)
- **"Interpretability is the degree to which a human can consistently predict the model's result."** *(Kim et al. 2016)*
- **"Interpretable Machine Learning** refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans." (C. Molnar 2018)
- "Making a machine learning interpretable can, but does not necessarily have to, imply providing a (human-style) explanation of a prediction."

DARPA XAI Project

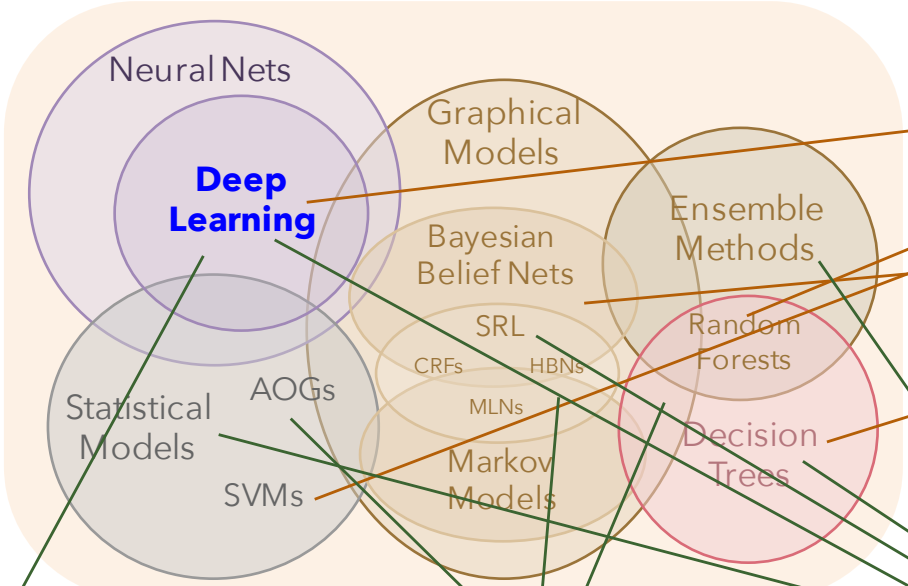


Explainable Models

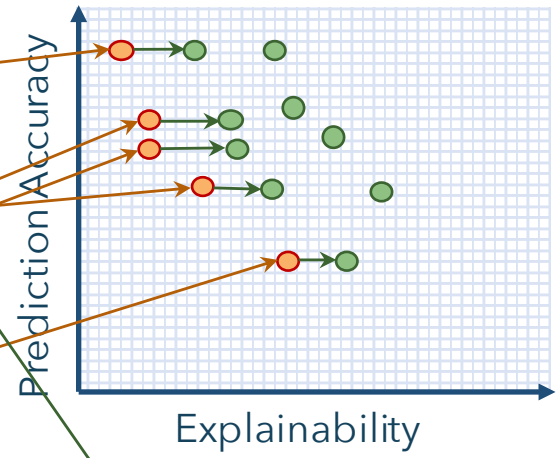
New Approach

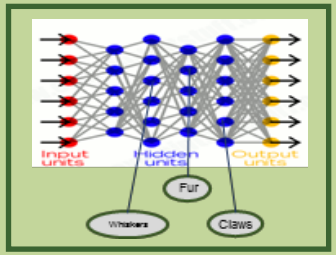
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)



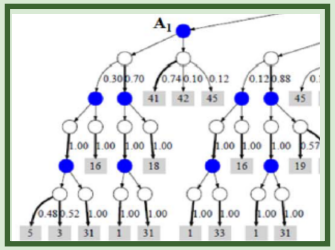
Explainability (notional)





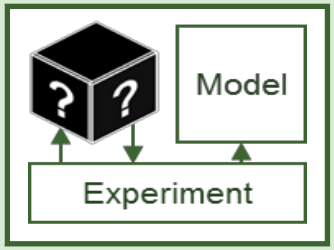
Deep Explanation

Modified deep learning techniques to **learn explainable features**



Interpretable Models

Techniques to learn more **structured, interpretable, causal models**

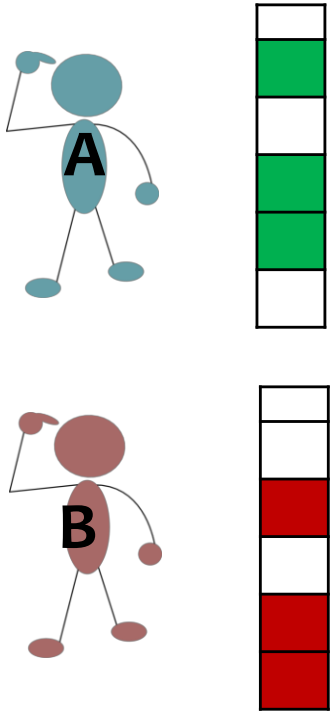


Model Induction

Techniques to **infer an explainable model** from any model as a black box

Why Interpretability?

1. Gain trust by **verifing** that classifier works as expected.



Omics feature

Interpretable
Machine
Learning

Suggested treatment of patient A is
because the model detected/learned
features in the patient.

Suggested treatment of patient B is
because the model detected/learned
features in the patient.

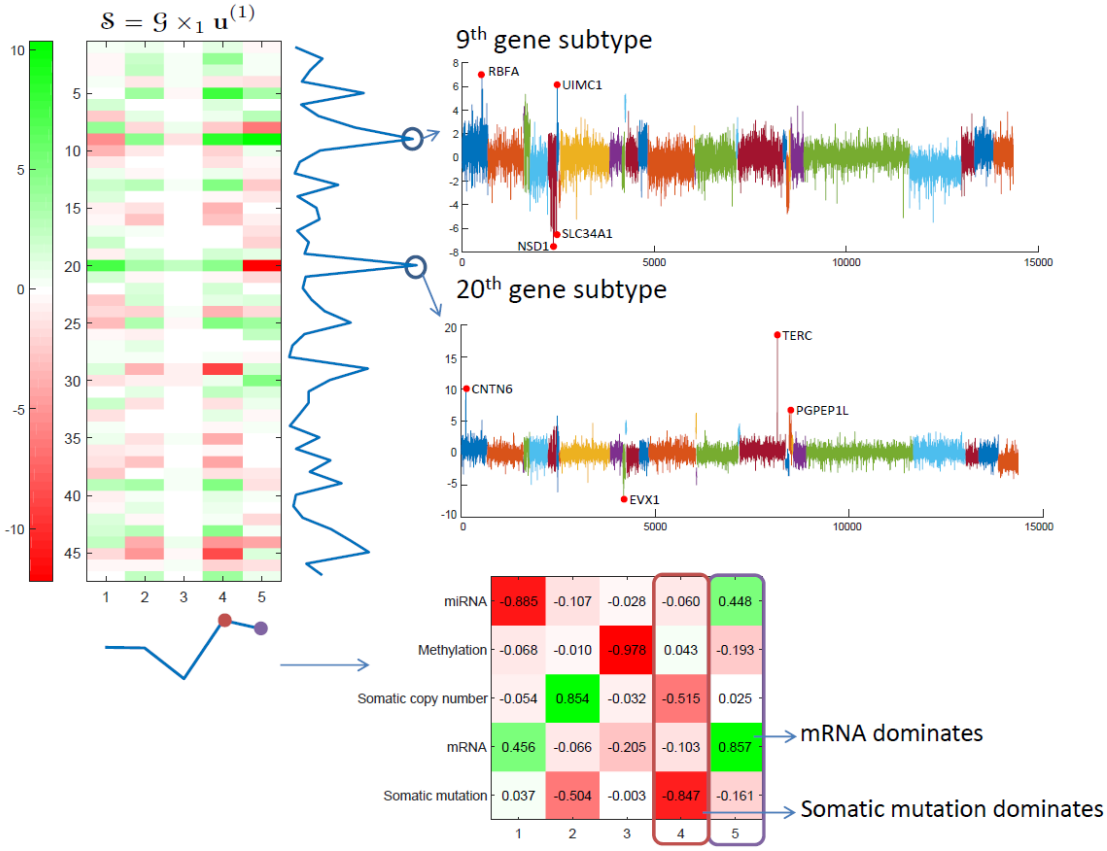
Why Interpretability?

2. **Compliance** to legislation

- EU GDPR's "**Right to Explanation**"
- **Retain human decision** in order to **assign responsibility!**
- **Protect privacy** by determining whether sensitive information in the data is revealed.
- **Unbiased** and **not discriminating** against protected groups

Why Interpretability?

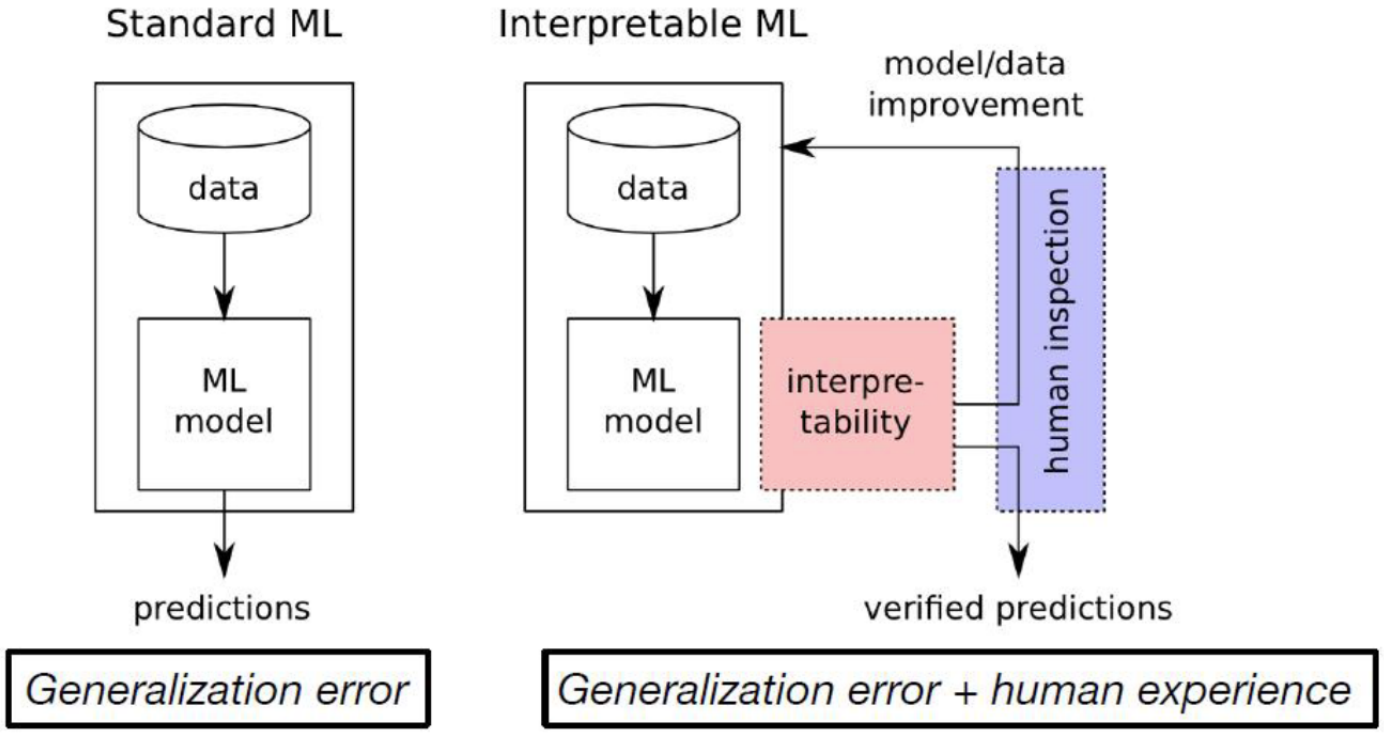
- 3. Gain **insights** for advance in science
  - Ex> Detecting causality, detecting significant features



[D. Choi & L. Sael (in submission)]

Why Interpretability?

4. **Improve** the learning machine by
   - Finding the cause of low accuracy (**debug** and **audit**) or cause of reduction in **robustness**

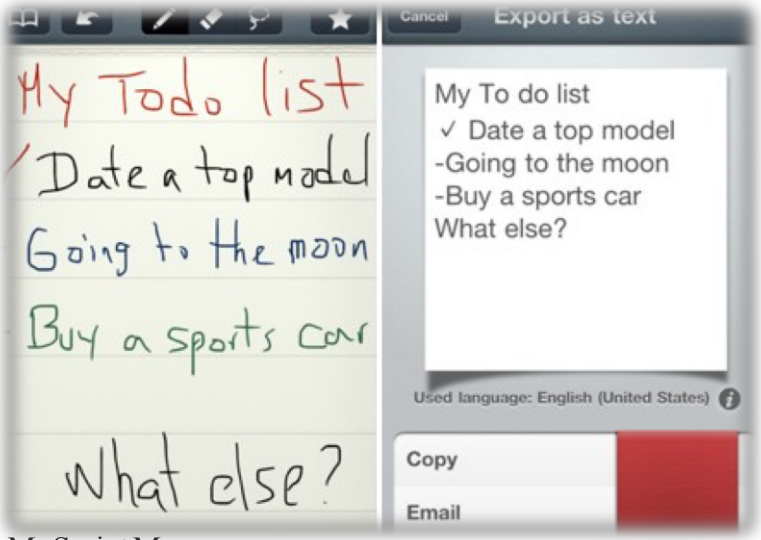


[Doshi-Velez and Kim 2017]

When Interpretability is Not Needed

- Model **has no significant impact**
  – for fun models

- When problem is **well-studied** and **validated.**



MyScript Memo

Part 1: Introduction

- What and why interpretable ML?
- **What is evaluated in interpretable ML?**
- How can we categorize interpretable ML?

[Doshi-Velez and Kim 2017]

Evaluating Interpretability

- Application level evaluation
  - Tested by domain experts
- Human level evaluation
  - Tested by non-experts
- **Function level evaluation**
  - Computed

Note: Currently there are no real consensus on what interpretability in machine learning is and how to measure it.

Application Level Evaluation

- Explanation is included in the software product and end user (domain expert) tests it.

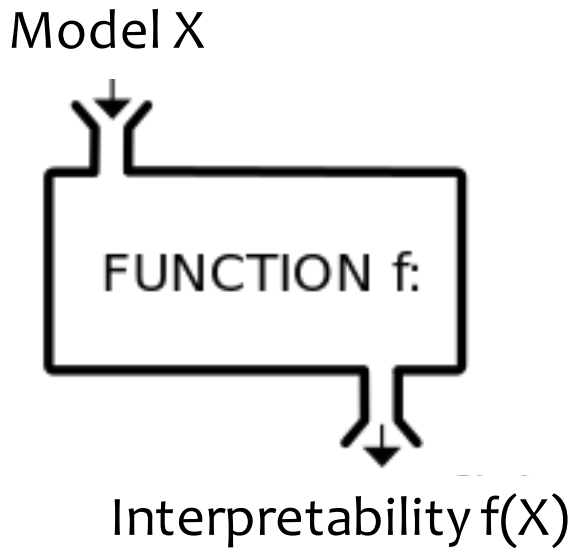
Human Level Evaluation

- A simplified application level evaluation where the evaluator is <u>not a domain expert</u>

Function Level Evaluation (FLE)

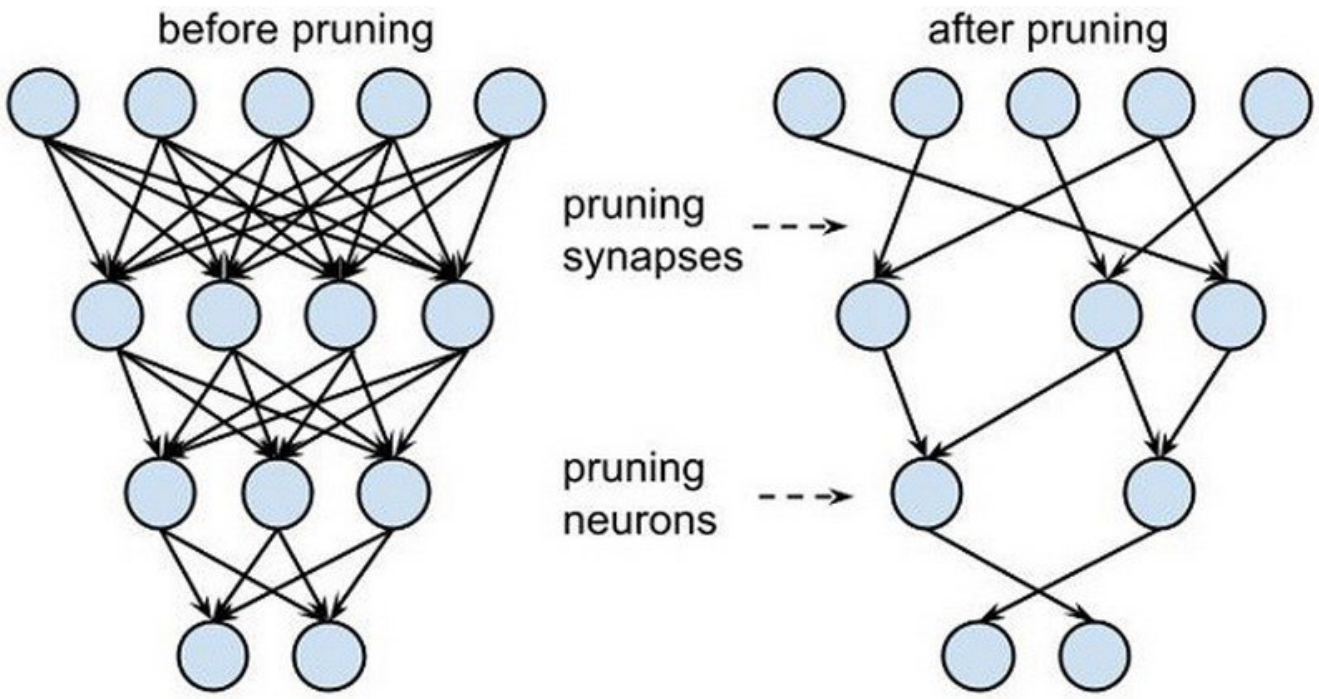
Function that has already been shown via human level evaluation to correlate with interpretability

- Example measures
  - Model sparsity
  - Conservation
  - Continuity (Monotonicity)
  - Selectivity
  - Positivity
  - Uncertainty
  - Cognitive processing time
  - Etc.



FLE 1: Model Sparsity

- Small number of features are easier to explain



Deep Compression

[Montavon'17]

FLE2 : Conservation

- Total attribution on the input features should be proportional to the amount of (explainable) evidence at the output.

Input feature **X**

x_1
x_2
…
x_d

ML → f(**x**) = "cat"

Responsibility **R**

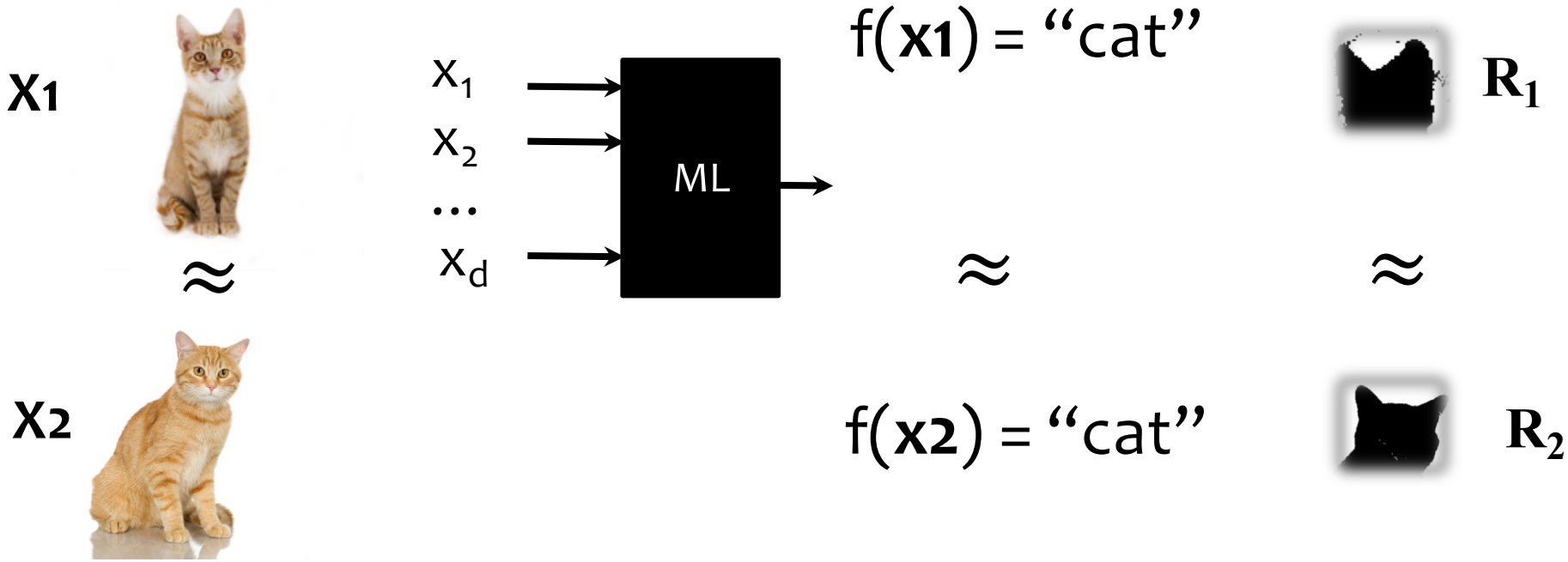
$R_1\ R_2 \ldots R_d$

$$\sum_{i=1}^{d} R_d \propto f(\boldsymbol{x})$$

[Montavon'18]

FLE 3: Continuity (Monotonicity)

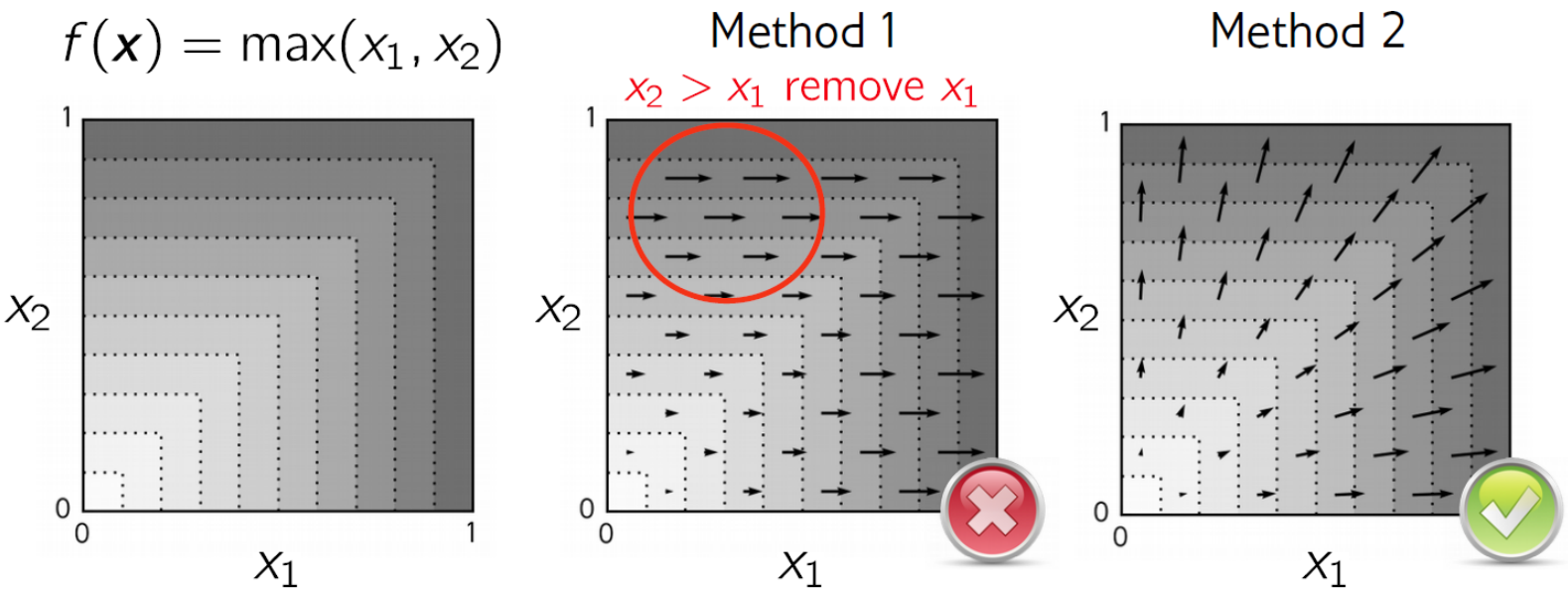
- If two inputs are similar and the prediction is similar, then the explanation should be similar.



[Bach'15, Samek'17]

FLE 4: Selectivity

- If input features are relevant to output decision, removing them should reduce evidence at the output.

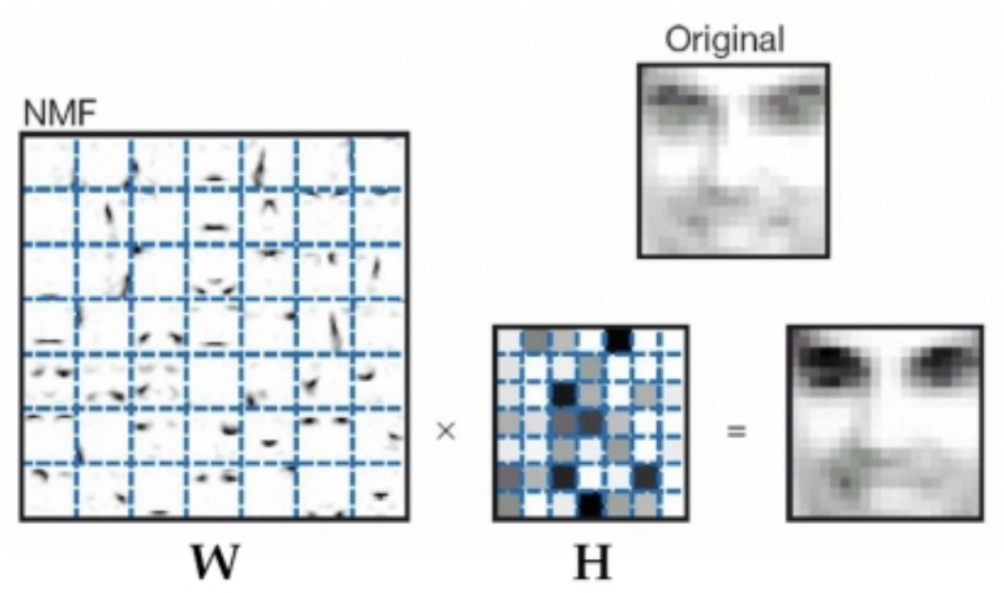


[CVPR'18 Tutorial]

FLE 5: Positivity

[Montavon'17, Sun'11, Landecker'13, Bach'15]

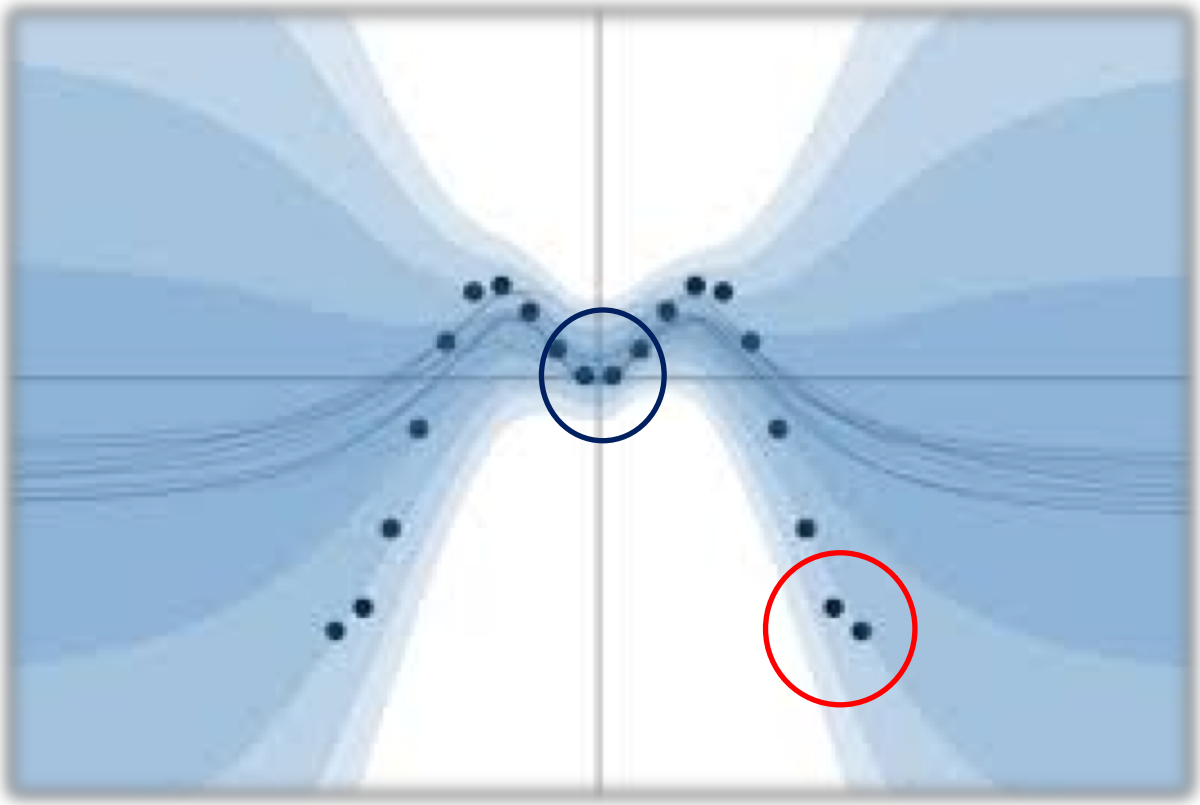
- If the model is certain about its prediction, input features are either relevant (positive) or irrelevant (zero)



Nonnegative Matrix Factorization (NMF) of facial images [Lee and Seung (1999)]

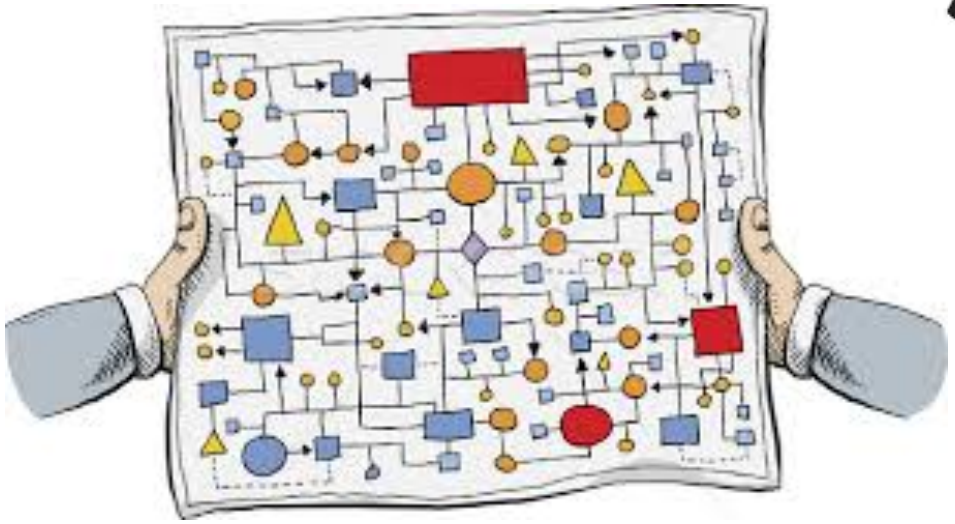
FLE 6: Uncertainty

- How uncertain the model is about the output.



FLE 7: Cognitive Processing Time

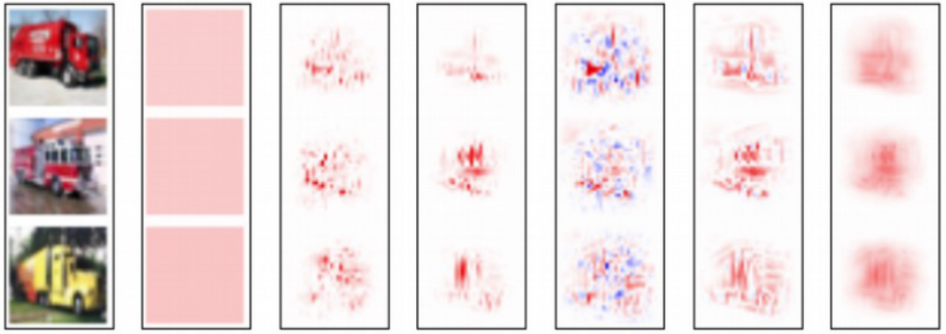
- How long does it take to understand the explanation?



e.g., Functional evaluation of intrinsically interpretable models

Algorithm	Linear	Monotone	Interaction	Task
Linear models	Yes	Yes	No	Regr.
Logistic regression	No	Yes	No	Class.
Decision trees	No	No	Yes	Class. + Regr.
RuleFit	Yes	No	Yes	Class. + Regr.
Naive Bayes	Yes	Yes	No	Class.n
k-nearest neighbours	No	No	No	Class. + Regr.

e.g., Functional Evaluation of DNN



Explanation techniques / Properties	Uniform	$(\text{Gradient})^2$	$(\text{Guided BP})^2$	Gradient x Input	Guided BP x Input	LRP-$\alpha_1\beta_0$	...
1. Conservation	✔			✓	✓	✔	
2. Positivity	✔	✔	✔		✓	✔	
3. Continuity	✔		✓		✓	✔	
4. Selectivity		✓	✔	✓	✔	✔	
...							

Part 1: Introduction

- What and why interpretable ML?
- What is evaluated in interpretable ML?
- **How can we categorize interpretable ML?**

[Zachary C Lipton. 2016]

Criteria for Interpretability 1

Intrinsic vs post hoc

- **Intrinsically** interpretable models
  - Decision trees
  - Linear models
  - Logistic regression
  - Decision rules
  - RuleFit
  - Other interpretability integrated machine learning

- **Post hoc** interpretable models
  - Selecting and training a black box model and applying interpretability methods after the training

[C. Molnar 2018]

Criteria for Interpretability 2

Model-specific vs model-agnostic

- Model-specific interpretation tools
  - Limited to specific model classes
  - E.g. interpretation of regression weights in a linear model
  - *Interpreting neural networks is also model-specific
- Model-agnostic tools
  - Independent of machine learning model and are usually post hoc
  - Assumes that model internals, e.g., weights or structural information, are not accessible

Note: Model-Agnostic Methods (MAM)

Pros of MAM [Ribeiro et al. 2016]

- **Model flexibility**
  - Not tied to any models
- **Explanation flexibility**
  - Not tied to a form of explanation
- **Representation flexibility**
  - Need not use the same feature representation as the model being explained

We will focus more on the Agnostic Methods

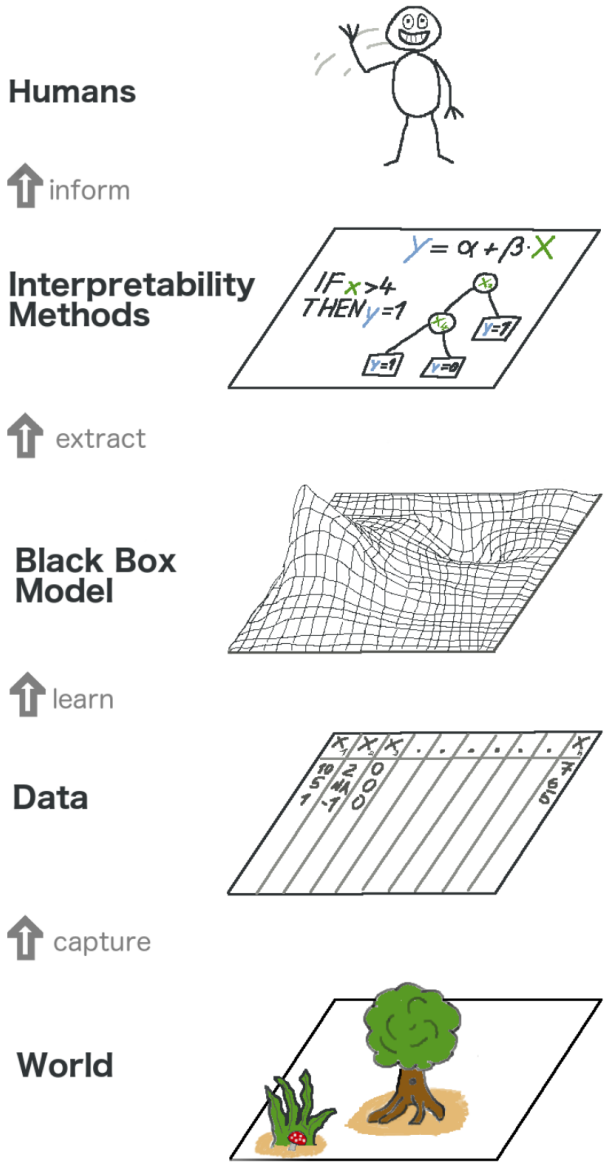


Figure from [C. Molnar, 2018]

[ICASSP 2017 Tutorial]
[C. Molnar 2018]

Criteria for Interpretability 3

focus on model (global) ⟷ *focus on data (local)*

1. Interpreting Learned Models	2. Explaining Decisions

Model

"What **pattern** describes a **class** according to the model?"

"What **prototypes** does the model learn?"

"How do **parts of the model** influence predictions?"

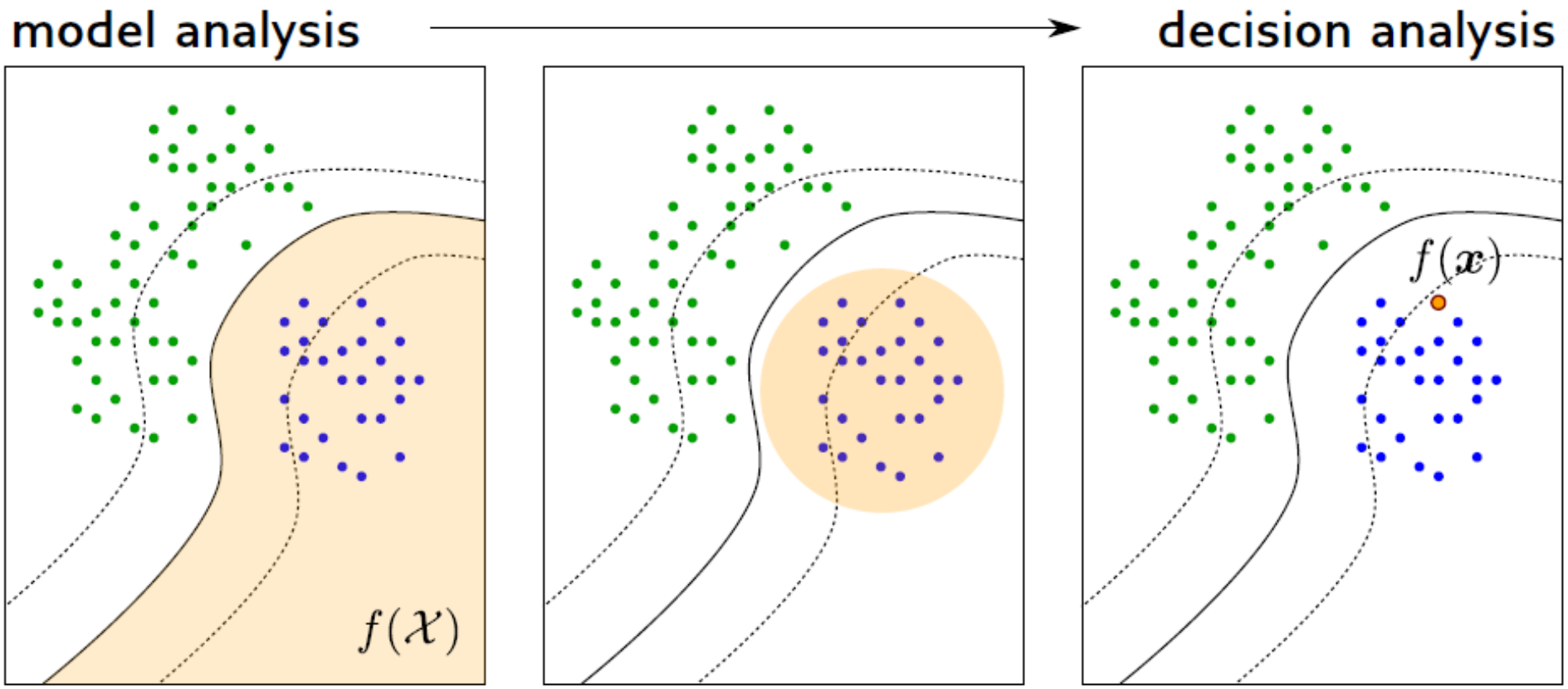
Prediction

"Explain **why** a instance (set of instances) has been classify in a certain way f(x)?"

Data

"Which part of the instance are **relevant** or **predictive** of the task?"

From Model Analysis to Decision Analysis



- Discriminative models
- p(C|X)

- Generative models
- p(X,C)=p(C)p(X|C)
- p(C|X) = p(C)p(X|C)/p(X)

[C. Molnar 2018]

Criteria for Interpretability 4

Outcome of the interpretable methods

- Intrinsically interpretable model
  - **Approximate** the black box model via interpretable model
  - E.g. model compression, global surrogate methods
- Model internals
  - *Applicable for intrinsically interpretable models
  - E.g. learned weights in linear model
- Data points
  - E.g. identification of **prototypes** of predicted classes
- Features
  - **Feature** summary statistics or visualization

Reference

- Christoph Molnar, *Interpretable Machine Learning:* A Guide for Making Black Box Models Explainable. (Interpretable Machine Learning A Guide for Making Black Box Models Explainable)
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Model-Agnostic Interpretability of Machine Learning." ICML Workshop on Human Interpretability in Machine Learning, no
- Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269
- Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning," no. Ml: 1–13.
- Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." ICML Workshop on Human Interpretability in Machine Learning
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. 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 (pp. 1135-1144). ACM.
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2018. "Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the 'Rashomon' Perspective.