CMPT 733 – Big Data Programming II

Automated Machine Learning (AutoML)

Instructor Steven Bergner

Course website https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages

Slides by: Lydia Zheng and Jiannan Wang

Motivation

1. Machine learning is very **successful**
2. To build a traditional ML pipeline:
   - Domain experts with longstanding experience
   - Specialized data preprocessing
   - Domain-driven meaningful feature engineering
   - Picking right models
   - Hyper-parameter tuning
   - ……

H20 Driverless AI Demo

https://www.youtube.com/watch?v=ZqCoFp3-rGc

1. Will AutoML software replace Data Scientists?
2. How to approach AutoML as a data scientist?

AutoML Vision

For Non-Experts

AutoML allows non-experts to make use of machine learning models and techniques without requiring to become an expert in this field first

https://en.wikipedia.org/wiki/Automated_machine_learning

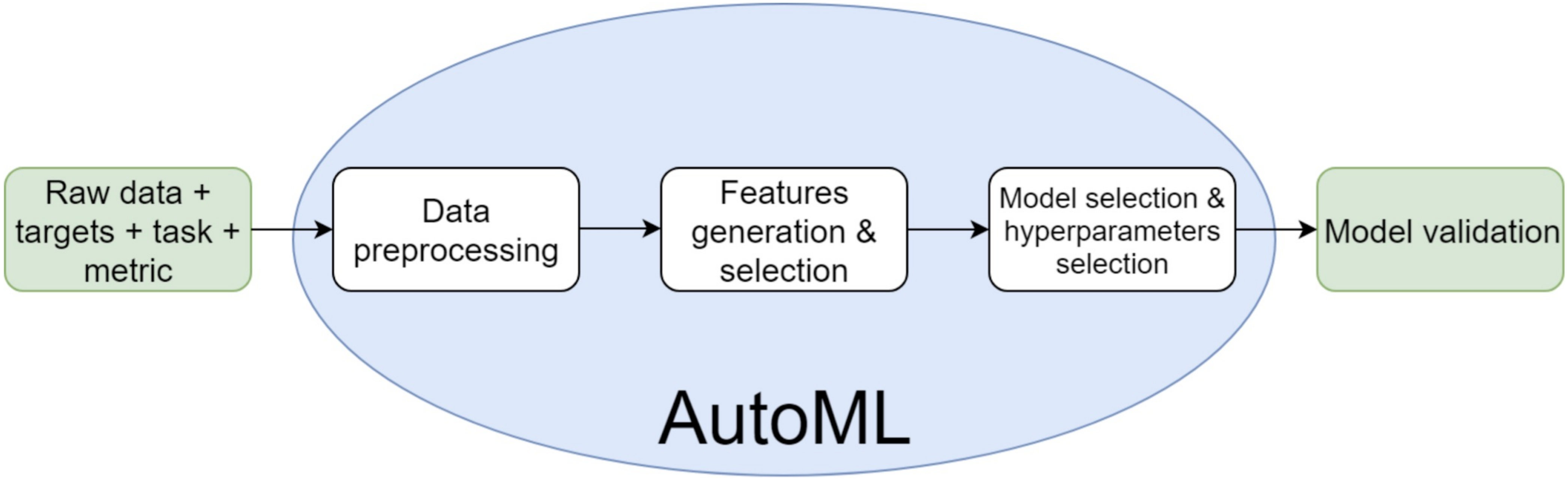
For Data Scientists

AutoML aims to augment, rather than automate, the work and work practices of heterogeneous teams that work in data science.

Wang, Dakuo, et al. "Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI." Proceedings of the ACM on Human-Computer Interaction 3.CSCW (2019): 1-24.

What is AutoML?

- Automate the process of applying machine learning to real-world problems



Outline

- Auto Feature Selection (Lecture 6)
- Auto Hyperparameter Tuning (Lecture 6)
- Auto Feature Generation (This Lecture) Neural Architecture Search (This Lecture)

Auto Feature Generation

Motivation

- The model performance is heavily dependent on quality of features in dataset
- It's time-consuming for domain experts to generate enough useful features

Feature Generation

- Unary operators (applied on a single feature)
  - Discretize numerical features
  - Apply rule-based expansions of dates
  - Mathematical operators (e.g., Log Function)
- Higher-order operators (applied on 2+ features)
  - Basic arithmetic operations (e.g., +, -, ×, ÷)
  - Group-by Aggregation (e.g., GroupByThenAvg, GroupByThenMax)

Katz, Gilad, Eui Chul Richard Shin, and Dawn Song. "Explorekit: Automatic feature generation and selection." 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016.

Featuretools

- An open source library for performing automated feature engineering
- Design to fast-forward feature generation across **multi-relational** tables

Concepts

- Entity is the relational tables
- An EntitySet is a collection of entities and the relationships between them
- Feature Primitives
  - Unary Operator: transformation (e.g., MONTH)
  - High-order Operator: Group-by Aggregation (e.g., GroupByThenSUM)

Entity sets

Customer

Customer_id	Birthdate	MONTH(Birthdate)	SUM(Product.Price)
1	1995-09-28	9	$500
2	1980-01-01	1	...
3	1999-02-02	2	...
...	...	...	...

Unary Operator:
MONTH

GroupBy
ThenSUM:

Product

Product_id	Customer_id	Name	Price
1	1	Banana	$100
2	1	Banana	$100
3	1	Orange	$300
4	2	Apple	$50
...	...	...	...

Feature Primitives

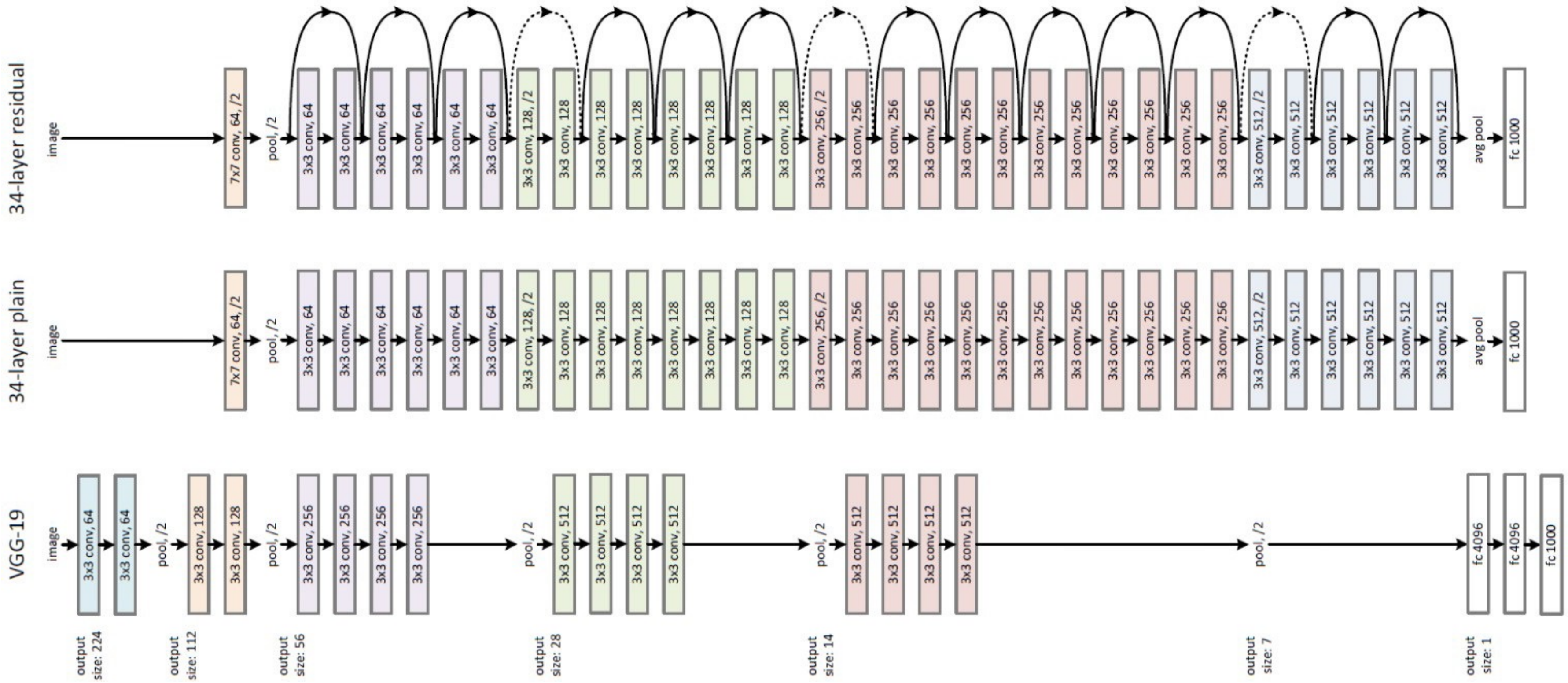
Outline

- Auto Feature Selection (Lecture 5)
- Auto Hyperparameter Tuning (Lecture 5)
- Auto Feature Generation (This Lecture) Neural Architecture Search (This Lecture)

Neural Architecture Search (NAS)

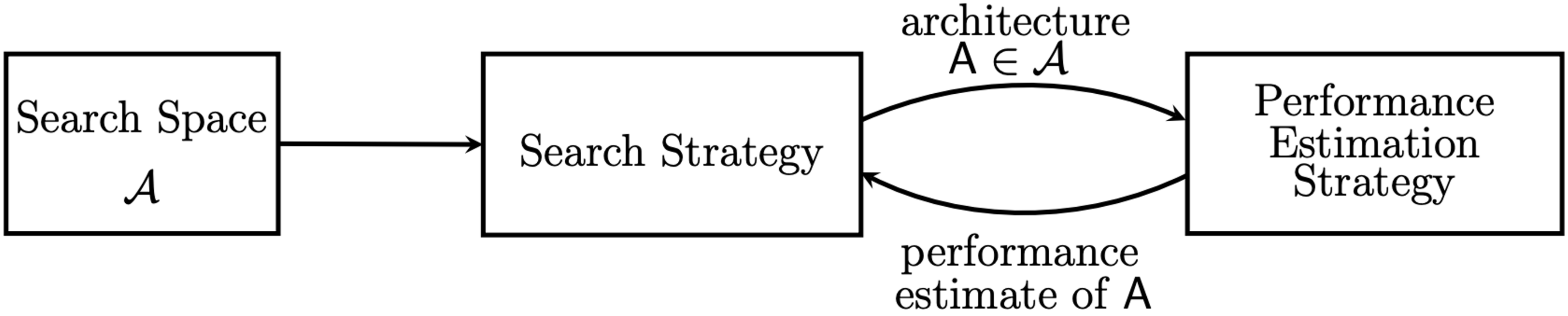
Motivation

How can someone come out with such an architecture?



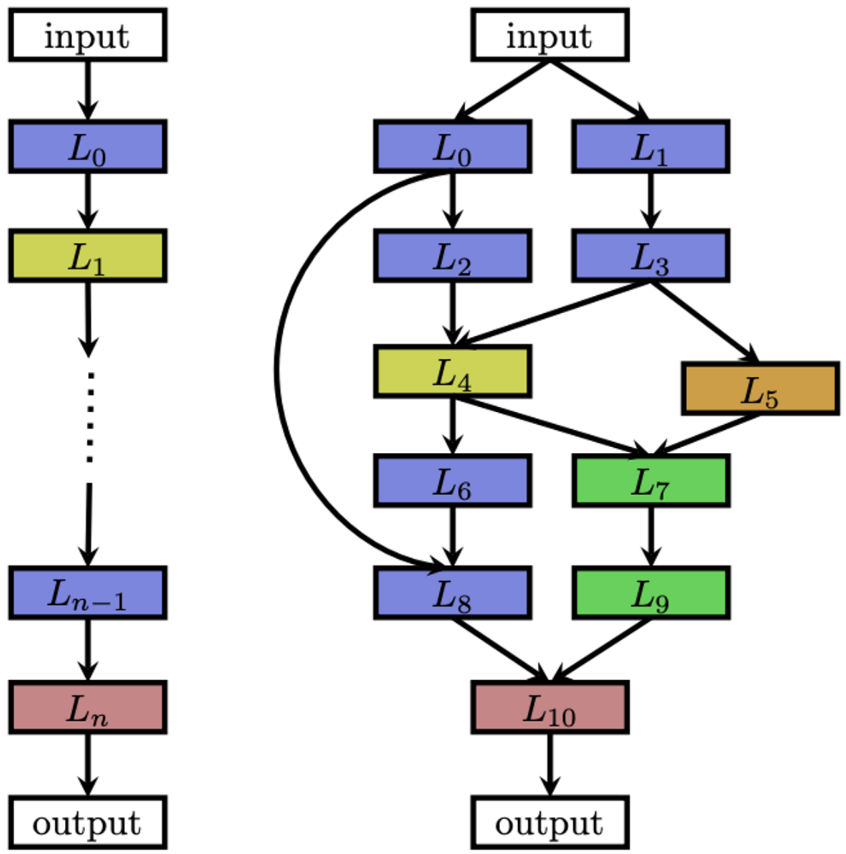
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." CVPR. 2016

Neural Architecture Search : Big Picture



Search Space

- Define which neural architectures a NAS approach might discover in principle
- May have human bias → prevent finding novel architectural building blocks



Chain-structured Multi-branch

Search Strategy

- **Basic Idea**
  - Explore search space (often exponentially large or even unbounded)
- **Methods**
  - Random Search
  - Bayesian Optimization [Bergstra et al., 2013]
  - Evolutionary Methods [Angeline et al., 1994]
  - Reinforcement Learning [Baker et al., 2017]
  - .....

Performance Estimation Strategy

- **Basic Idea**
  - The process of estimating predictive performance
- **Methods**
  - Simplest option: perform a training and validation of the architecture on data
  - Initialize weights of novel architecture based on weights of other architectures have been trained before
  - Using learning curve extrapolation [Swersky et al., 2014]
  - ……

Summary

What is AutoML and why we need it?

How AutoML works?

- Auto Feature Selection (Lecture 5)
- Auto Hyperparameter Tuning (Lecture 5)
- Auto Feature Generation (This Lecture)
- Neural Architecture Search (This Lecture)

CMPT 733 – Big Data Programming II

Explainable Machine Learning

Instructor	Steven Bergner
Course website	https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages
Slides by:	Xiaoying Wang and Jiannan Wang

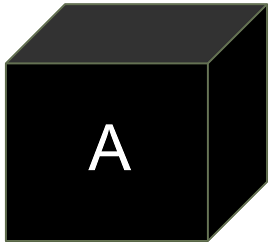
Outline

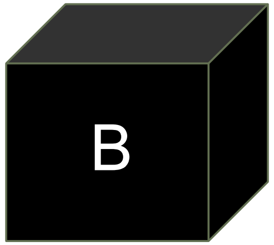
- Motivation: Why Explainable ML matters?
- Big Picture: Taxonomy State-of-the-art Techniques

Outline

- Motivation: Why Explainable ML matters?
- Big Picture: Taxonomy State-of-the-art Techniques

Evaluation





Bird: 99.0%

Bird: 99.9%

Which model are you going to choose?

Evaluation

A

Because it has wings and a beak

Bird: 99.0%

B

Because it is white and the background is blue

Bird: 99.9%

Which model are you going to choose?

Debugging

What's wrong?

Q: How symmetrical are the white bricks on either side of the building?

A: very

Q: How asymmetrical are the white bricks on either side of the building?

A: very

Q: How fast are the bricks speaking on either side of the building?

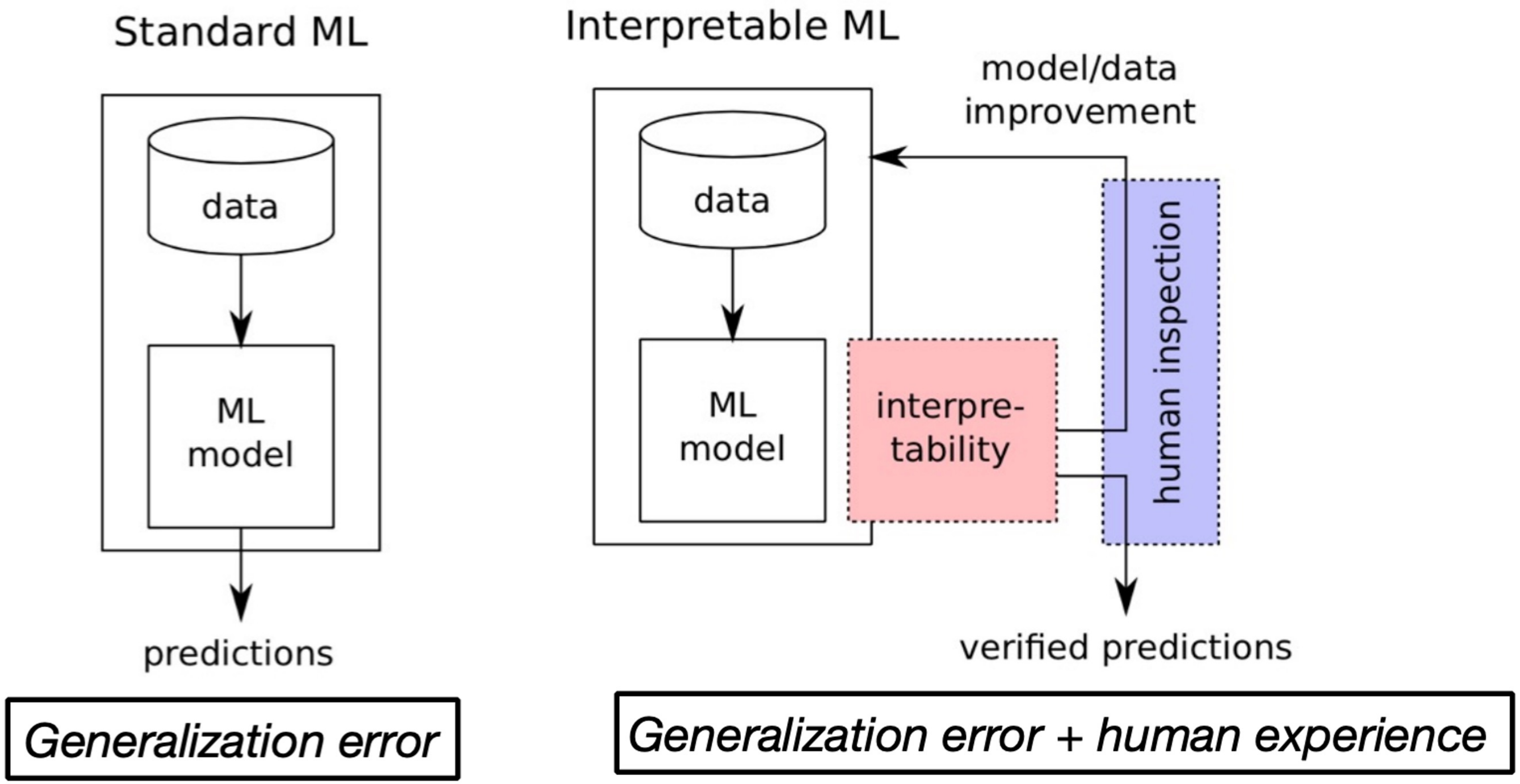
A: very

Debugging

How symmetrical are the white bricks on either side of the building?

red: high attribution
blue: negative attribution
gray: near-zero attribution

Improvement



Learning insights

"It's not a human move. I've never seen a human play this move"

"So beautiful."

- Fan Hui

Legal Concerns

SR 11-7: Guidance on Model Risk Management

BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM
WASHINGTON, D.C. 20551

DIVISION OF BANKING
SUPERVISION AND REGULATION

SR 11-7
April 4, 2011

TO THE OFFICER IN CHARGE OF SUPERVISION AND APPROPRIATE SUPERVISORY AND EXAMINATION STAFF AT EACH FEDERAL RESERVE BANK

SUBJECT: Guidance on Model Risk Management

Art. 22 GDPR

Automated individual decision-making, including profiling

Outline

Motivation: Why Explainable ML matters?

Big Picture: Taxonomy

State-of-the-art Techniques

Taxonomy

Transparent Models — Linear Regression, Decision Tree, KNN, Bayesian Network…

Post-hoc Explanation

- **Global Model Explanation** — Permutations, Partial Dependence Plots, Global Surrogate …
- **Individual Prediction Explanation** — Attribution, Influential Instances, Local Surrogate …

Taxonomy

Transparent Models Linear Regression, Decision Tree, KNN, Bayesian Network…

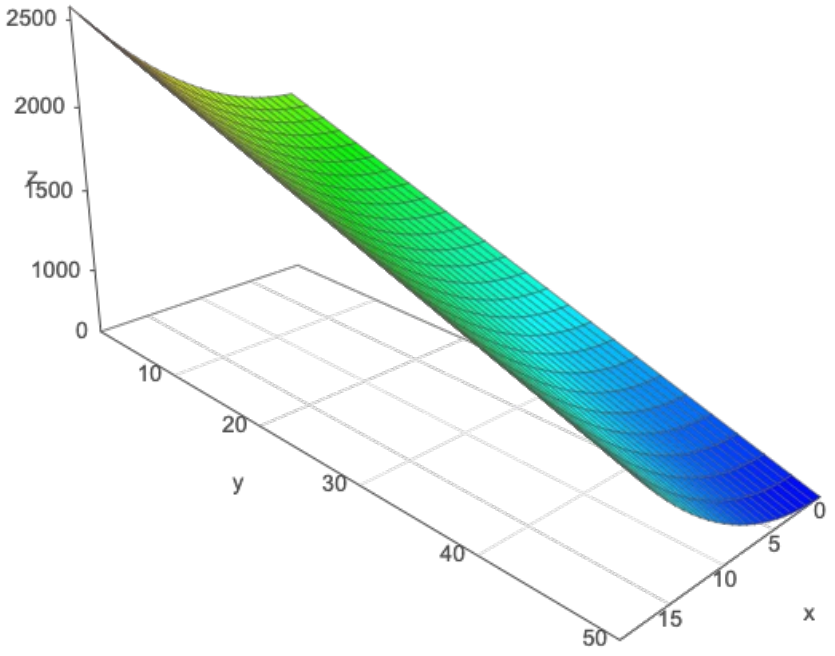
Post-hoc Explanation

- **Global Model Explanation** Permutations, Partial Dependence plots, Global Surrogate …
- **Individual Prediction Explanation** Attribution, Influential Instances, Local Surrogate …

Linear Regression

House rent (z) with respect to its area (x) and distance from SFU (y)

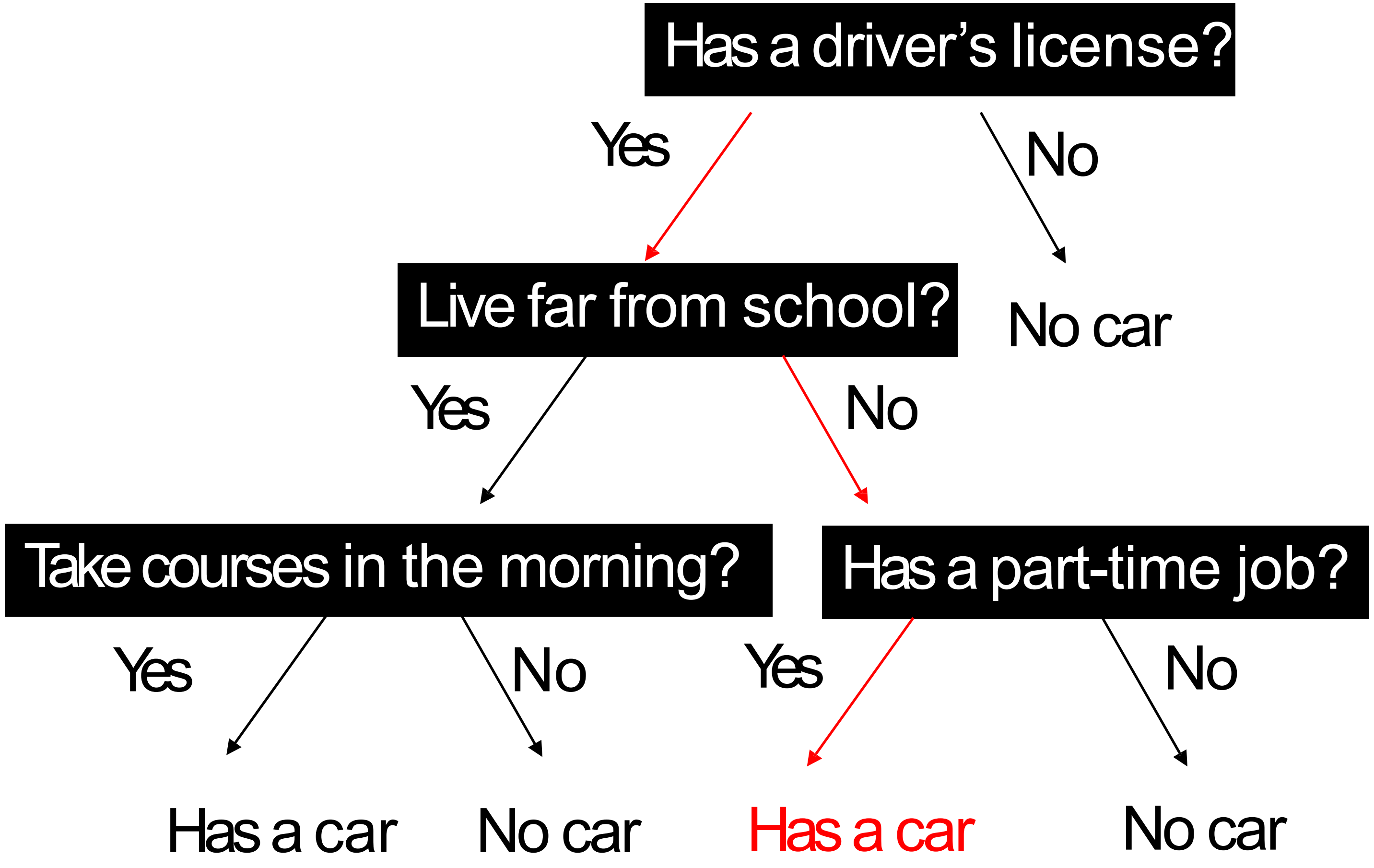
$$z = 2.1x - 2.4y + 1800$$



How do area and distance affect the house rent?

Decision Tree

Does a student own a car?



Why does the model predict student A has a car ?

Taxonomy

- **Transparent Models**
  - Linear Regression, Decision Tree, KNN, Bayesian Network…
- **Post-hoc Explanation**
  - **Global Model Explanation**
    - Permutations, Partial Dependence Plots, Global Surrogate …
  - **Individual Prediction Explanation**
    - Attribution, Influential Instances, Local Surrogate …

Permutations

Main idea: measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature

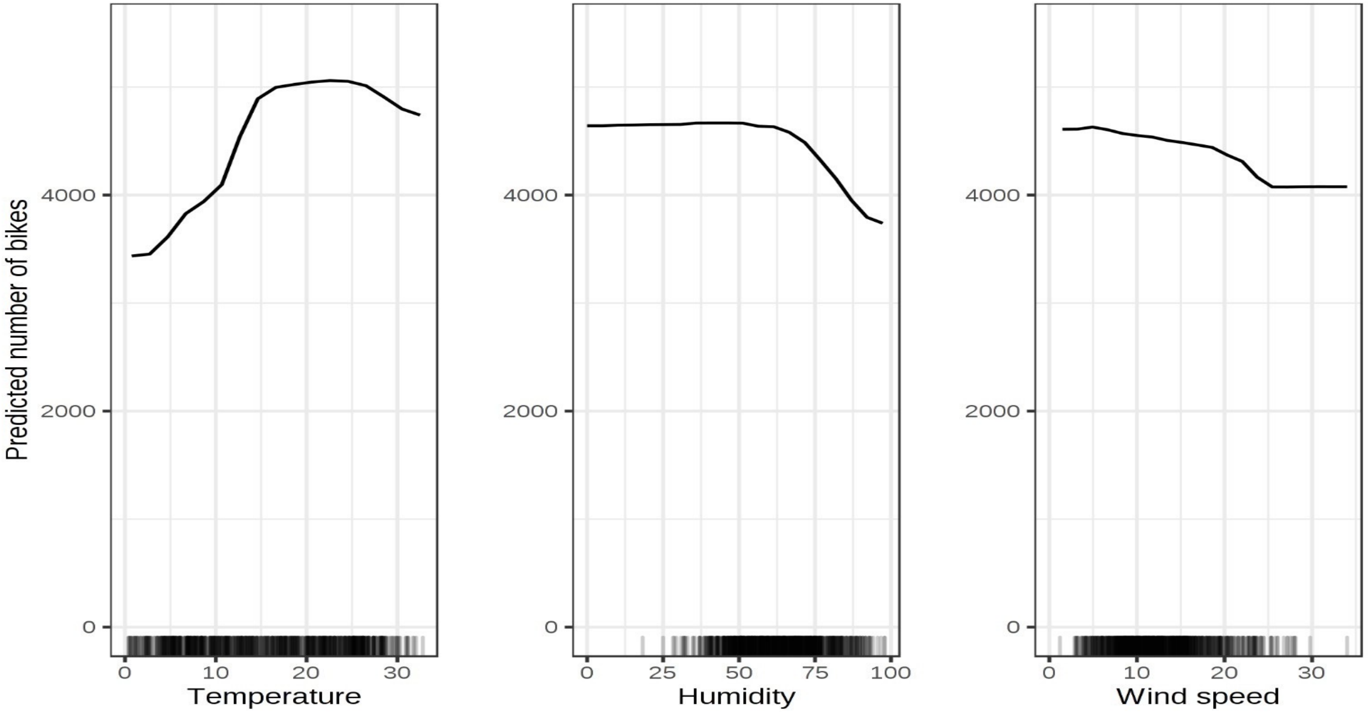
ID	Distance from SFU	# bathroom	Area	Closest bus stop	…
1	5.0km	1	$670ft^2$	0.30km	…
2	8.2km	2	$920ft^2$	0.12km	…
3	2.3km	2	$880ft^2$	1.20km	…
…	…	…	…	…	…
9999	10km	1	$680ft^2$	0.05km	…
10000	7.8km	1	$730ft^2$	0.23km	…

Permutations

- Input: trained model and labeled dataset for evaluation
- Output: relative importance for each feature
- Method:
  - Apply the model on original dataset and get an estimation error E
  - For each feature:
    - Permute feature and apply the model again on the permuted data to get a new estimation error E’
    - The feature importance can be measured by E’-E or E’/E

Partial Dependence Plots

Main idea: show the marginal effect one or two features have on the predicted outcome of a machine learning model



ID	Temperature	Humidity	Wind Speed	Rental#
1	20	30	20	3000
2	25	35	10	2500
3	22	25	15	3300
4	30	20	18	2000
..	..	…	…	…

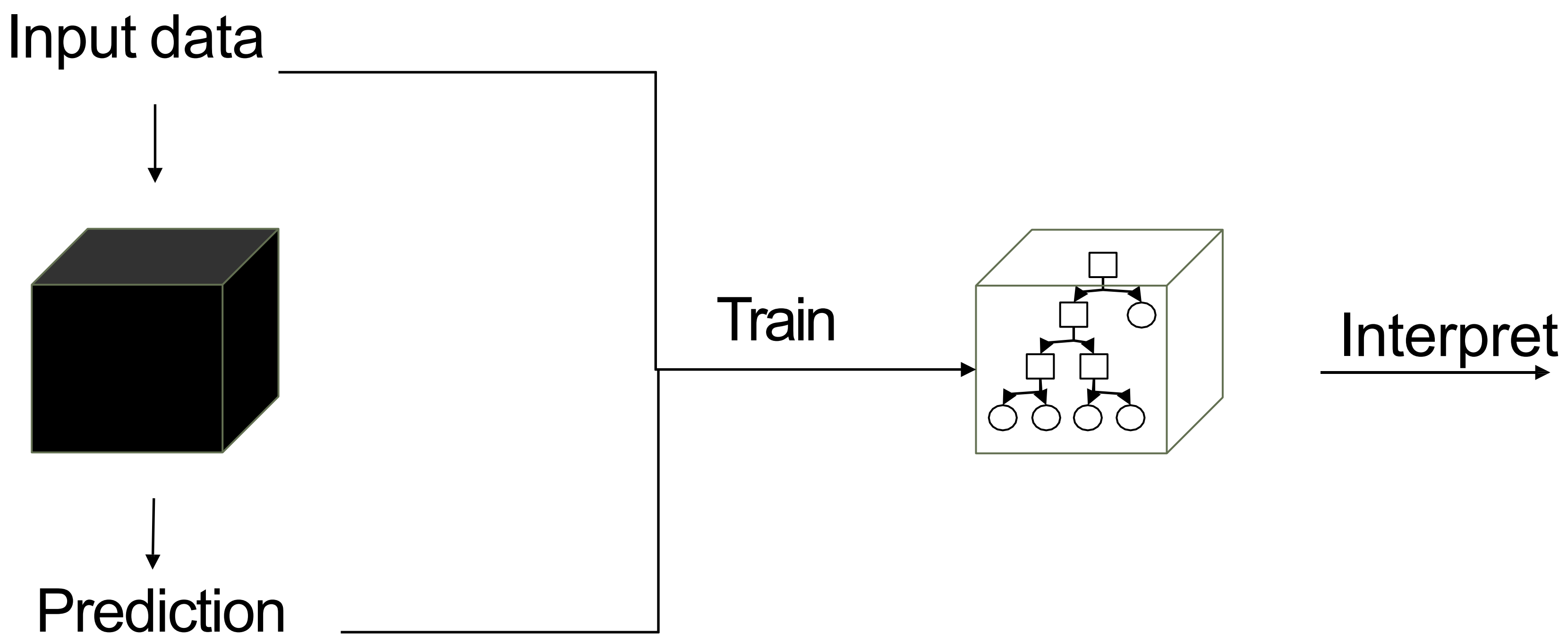
Partial Dependence Plots

Let x_s be the feature set ($|x_s| \in \{1,2\}$) we want to examine, and x_c be the rest of the features used in the model $\hat{f}$:

- Partial dependence function: $\hat{f}_{x_s}(x_c) = E_{x_c}\left[\hat{f}(x_s, x_c)\right] = \int \hat{f}(x_s, x_c)dP(x_c)$
- Can be estimated by: $\hat{f}_{x_s}(x_c) = \frac{1}{n}\sum_{i=1}^{n}\left(x_s, x_c^{(i)}\right)$

Global Surrogate

Main idea: train a transparent model to approximate the predictions of a black box model



Global Surrogate

Let $\hat{y}^{(i)}$ and $\hat{y}_*^{(i)}$ be the target model and surrogate model's prediction for the ith input data, we can use R-squared measure we can evaluate how good the surrogate model is in approximating the target model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(\hat{y}_*^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{n} \left(\hat{y}^{(i)} - \hat{y}_{avg}\right)^2}$$

Taxonomy

Transparent Models

Linear regression, Decision tree, KNN, Bayesian Network…

Global Model Explanation

Permutations, Partial dependence plots, Global surrogate …

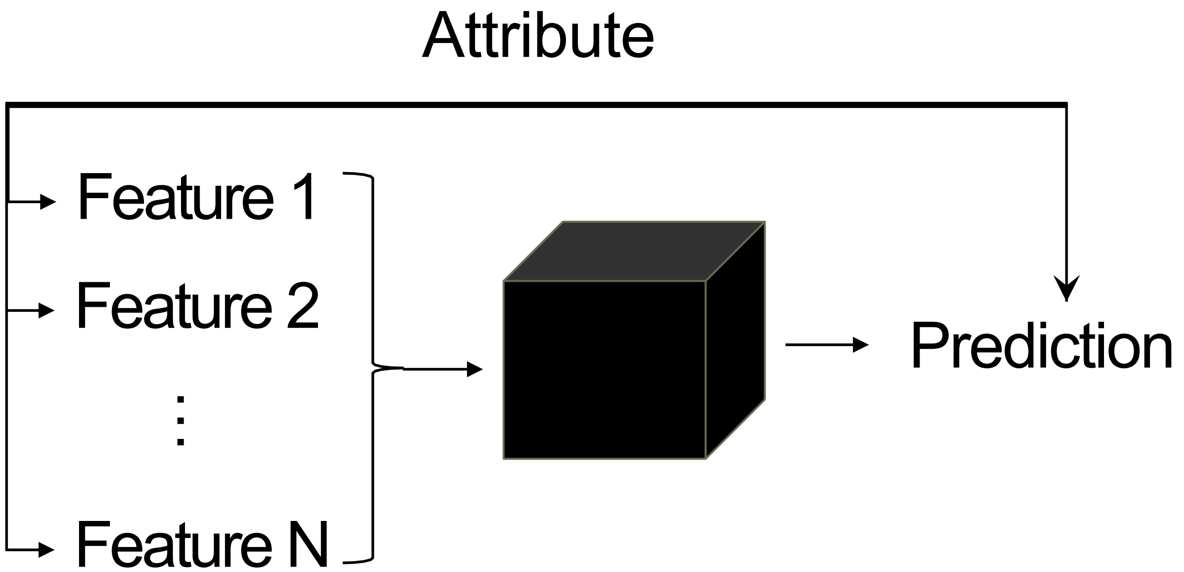
Post-hoc Explanation

Individual Prediction Explanation

Attribution, Influential instances, Local surrogate …

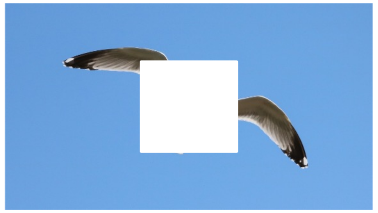
Attribution

- **Main idea:**
  - Attribute a model's prediction on a sample to its input features
- **Approaches:**
  - Ablation
  - Shapely value
  - …



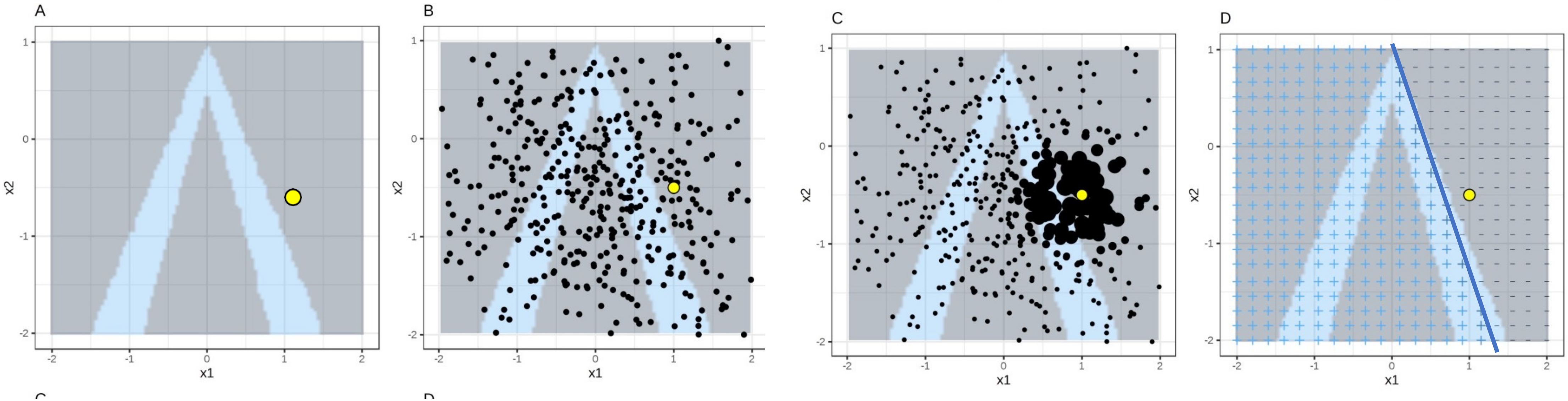
Attribution (Ablation)

Ablation: drop each feature and attribute the change in prediction to the feature

	A	B
	Bird (99%)	Bird (99%)
	Bird (20%)	Bird (98%)
	Bird (96%)	Bird (35%)

Local Surrogate (LIME)

Main idea: Test what happens to the prediction when give variations of data into the machine learning model



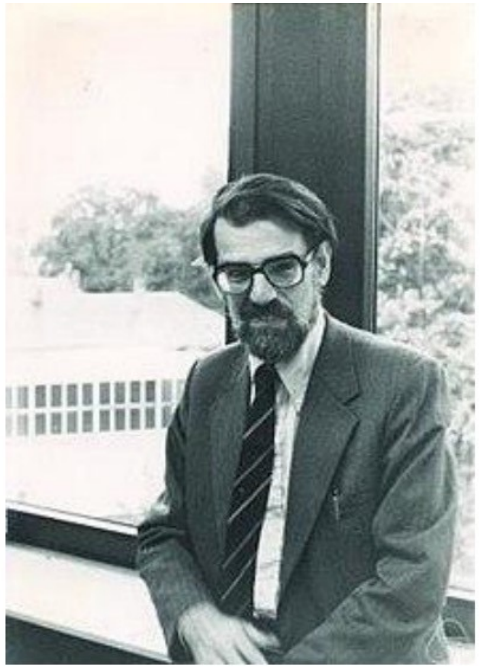
RIBEIRO, MARCO TULIO, SAMEER SINGH, AND CARLOS GUESTRIN. "WHY SHOULD I TRUST YOU?: EXPLAINING THE PREDICTIONS OF ANY CLASSIFIER." PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. ACM (2016)

Local Surrogate (LIME)

- The local surrogate model is obtained by: $argmin_{g \in G}\ L(f, g, \pi_x) + \Omega(g)$
  - f: target model, g: surrogate model, G: family of all possible g, π_x: neighborhood of target sample
  - L: measure fidelity, how the surrogate model approximate the target model
  - Ω: measure complexity of the surrogate model
- Get variation of data:
  - Text and image: turn single word or super-pixels on and off
  - Tabular data: create new samples by perturbing each feature individually

Shapley Value

- Classic result in game theory on distributing the total gain from a **cooperative game**
- Introduced by **Lloyd Shapley** in 1953 , who won the **Nobel Prize in Economics** in 2012
- Popular tool in studying cost-sharing, market analytics, voting power, and most recently **explaining ML models**



Lloyd Shapley in 1980

"A Value for n-person Games". Contributions to the Theory of Games 2.28 (1953): 307-317

Attribution (Shapely Value)

- Shapely value: derive from game theory on distributing gain in a coalition game
- Coalition game: players collaborating to generate some gain, function $val(S)$ represents the gain for any subset S of players
    - Game: prediction task
    - Players: input features
    - Gain: marginalized actual prediction minus average prediction $val_x(S) =$

$$\int \hat{f}(x_1, x_2, \ldots, x_p) dP_{x \notin S} - E(\hat{f}(X))$$

- Marginal contribution of a feature i to a subset of other features: $val_x(S \cup \{x_i\}) - val_x(S)$

Attribution (Shapely Value)

- Shapely value of a feature i on sample x: weighted aggregation of its marginal contribution over all possible combinations of subsets of other features

$$\sum_{S \subseteq \{x_1, x_2, \ldots, x_p\} \setminus \{x_i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} (val_x(S \cup \{x_i\}) - val_x(S))$$

- Intuition: The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them.

Example

- A company with two employees **Alice** and **Bob**
  - No employees, **0** profit
  - Alice alone makes **20** units of profit
  - Bob alone makes **10** units of profit
  - Alice and Bob make total **50** units of of profit
- What should be the bonuses be?

All Possible Orders	Marginal for Alice	Marginal for Bob
Alice, Bob		
Bob, Alice		
Shapley Value		

Example

- A company with two employees **Alice** and **Bob**
  - No employees, **0** profit
  - Alice alone makes **20** units of profit
  - Bob alone makes **10** units of profit
  - Alice and Bob make total **50** units of of profit
- What should be the bonuses be?

All Possible Orders	Marginal for Alice	Marginal for Bob
Alice, Bob	20	30
Bob, Alice	40	10
Shapley Value	30	20

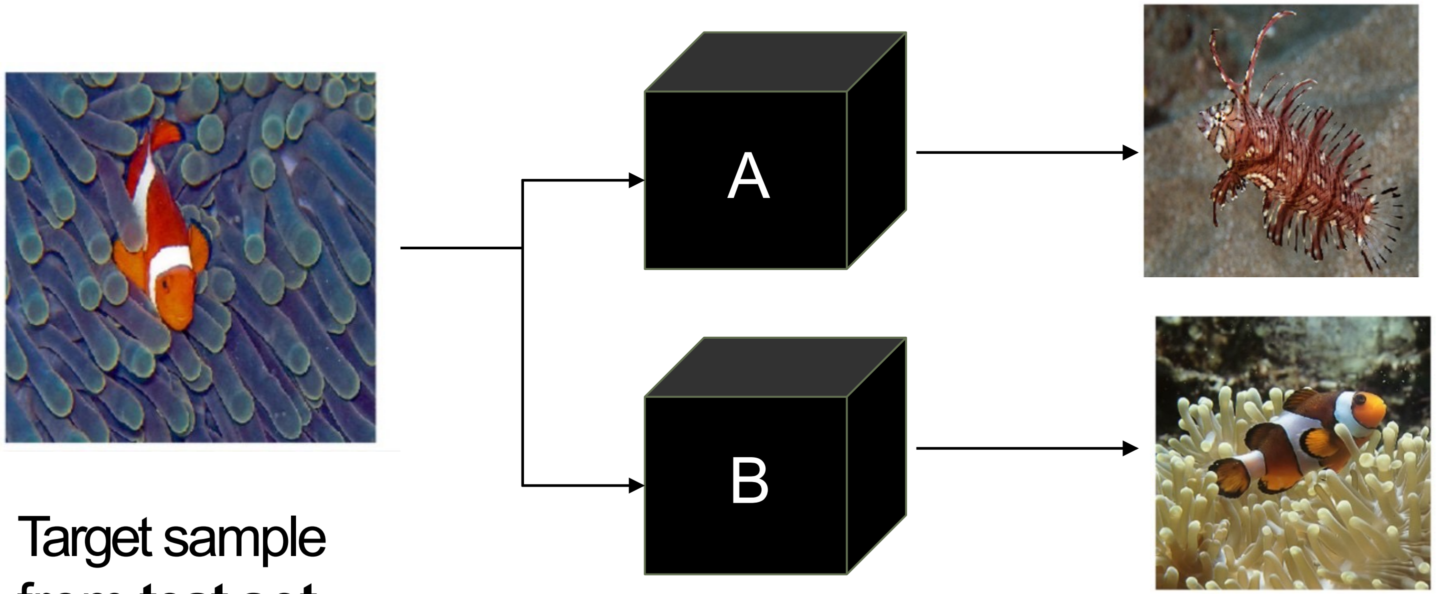
Attribution (Shapely Value)

- Two challenges when computing shapely value:
  - Exponential time since the permutation
  - Cannot inference on models when some features are not provided
- SHAP (SHapley Additive exPlanations) provide solutions for these two challenges:
  - KernelSHAP: an approximation solution for all models:
    - Sample a subset of feature orders
    - Filling missing features with background dataset provided by user

LUNDBERG, SCOTT M., AND SU-IN LEE. "A UNIFIED APPROACH TO INTERPRETING MODEL PREDICTIONS." ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. 2017.

Influential Instances

Main idea: debug machine learning model by identifying influential training instances (a training instance is influential when its deletion from training data considerably changes the model's prediction)



Target sample from **test set**

The most influential **training** sample for each model

KOH, PANG WEI, AND PERCY LIANG. "UNDERSTANDING BLACK-BOX PREDICTIONS VIA INFLUENCE FUNCTIONS." PROCEEDINGS OF THE 34TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING-VOLUME 70. JMLR. ORG, 2017

Influential Instances

- Naïve approach: deletion diagnostics
  - Train a model on all data instances, predict on test data and choose a target sample, for example: an incorrectly predicted sample with high confidence
  - For each training data, remove the data and retrain a model, predict on target sample and calculate the differences between the prediction and original prediction
  - Get the most influential top K instances (very likely to be mislabeled in this scenario)
  - Train a transparent model to find out what distinguishes the influential instances from the non-influential instances by analyzing their features (optional, for better understand the model)

Evaluation

- Human review: which method that human can get more insight of the model?
- Fidelity: how well does the method approximate the black box model?
- Stability: how much does an explanation differ for similar instances?
- Complexity: computational complexity of the method
- Coverage: the types of models that the method can explain
- ...

Available Tools

- LIME https://github.com/ankurtaly/Integrated-Gradients
- SHAP implementation in Python https://github.com/slundberg/shap
- Captum: PyTorch model interpretability toolhttps://github.com/pytorch/captum
- Skater: a Python Library for Model Interpretation/Explanations
  https://oracle.github.io/Skater/overview.html
- ELI5: a library for debugging/inspecting machine learning classifiers and explaining their predictions
  https://eli5.readthedocs.io/en/latest/
- Influence function implementation in Python https://github.com/kohpangwei/influence-release

References

- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2021. https://christophm.github.io/interpretable-ml-book/.
- Anon. KDD'19 Explainable AI Tutorial. Retrieved September 13, 2019 from https://sites.google.com/view/kdd19-explainable-ai-tutorial
- Anon. ICCV'19 Tutorial on Interpretable Machine Learning in Computer Vision. Retrieved September 20, 2019 from https://interpretablevision.github.io/

Summary

Transparent Models — Linear Regression, Decision Tree, KNN, Bayesian Network…

Post-hoc Explanation

- **Global Model Explanation** — Permutations, Partial Dependence Plots, Global Surrogate …
- **Individual Prediction Explanation** — Attribution, Influential Instances, Local Surrogate …