Explainable AI

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University

Explainable AI through Argumentation

What is Argumentation?

- Evaluate “possible conclusions” by considering reasons for and against
  - Constructing pros and cons arguments
  - Evaluating arguments accordingly
- Resolve conflicts (within or across “agents”)
- Often studied and applied in
  - Disciplines: philosophy, logic, law, artificial intelligence, computer science, etc.
  - Applications: decision-making, dispute resolution, negotiation, security, bioinformatics, etc.

Argumentation for Decision Making: How

1. Represent the decision problem: decision, goals, attributes, contexts, and their relationships…

↓

2. Construct arguments and attacks following an argumentation framework

↓

3. Evaluate the acceptability of arguments w.r.t chosen criteria (typically certain argumentation semantics)

↓

4. Derive argumentative explanations from the decision making process

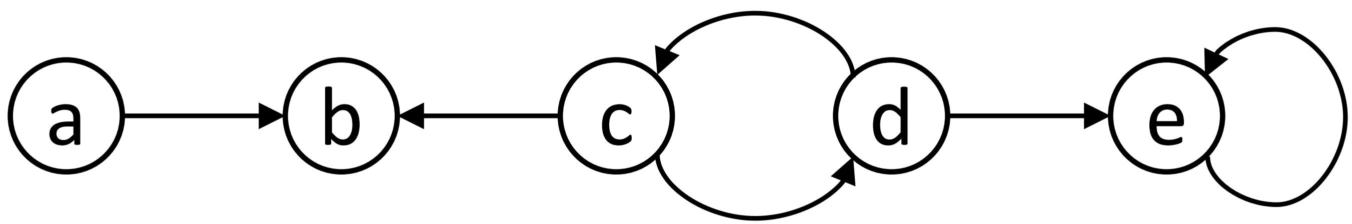
Argumentation: A Simple Example

- Abstract Argumentation
  - Arguments are "atomic"
  - Formalize relations ("attacks") between arguments

- An **abstract argumentation framework** (AF) is a pair (A, R) where
  - A is a set of arguments
  - $R \subseteq A \times A$ is a relation representing "attacks"

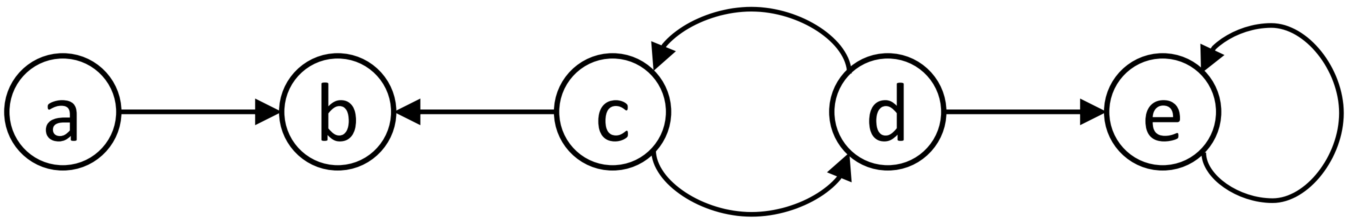
Argumentation: A Simple Example

- $A = \{a, b, c, d, e\}$
- $R = \{(a, b), (c, b), (c, d), (d, c), (d, e), (e, e)\}$



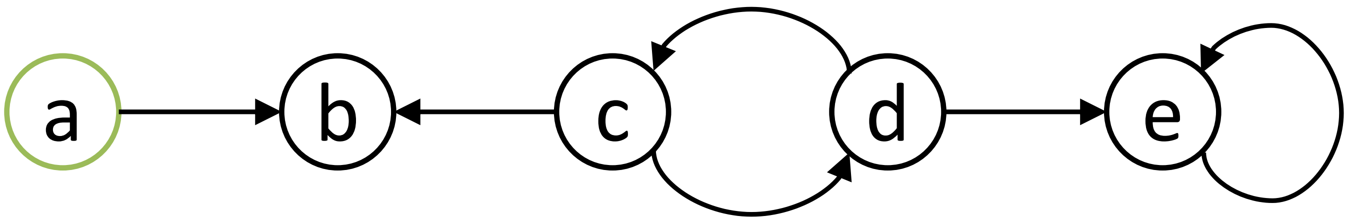
Argumentation: A Simple Example

- Conflict Free Set:
  - Given an AF $F = (A, R)$. A set $S \subseteq A$ is conflict-free (*cf*) in F, if, for each $a, b \in S, (a, b) \notin R$.



Argumentation: A Simple Example

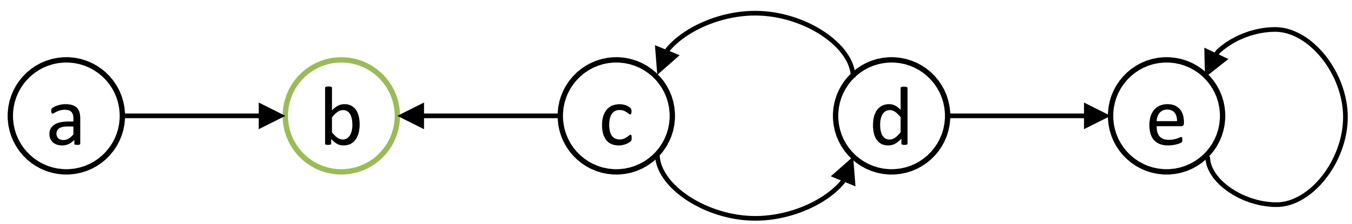
- Conflict Free Set:



- $cf(F) = \{\{a\},$

Argumentation: A Simple Example

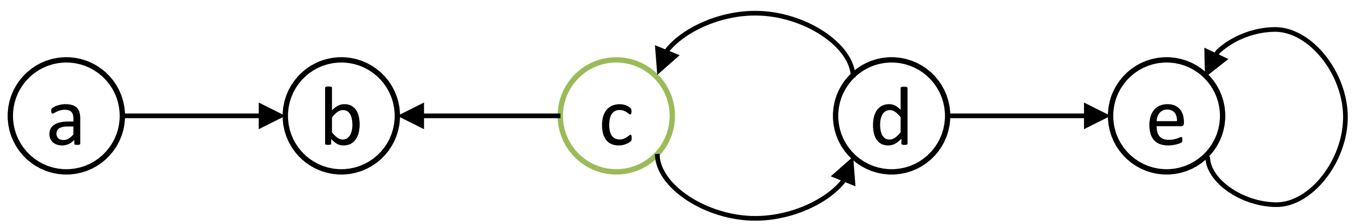
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}$

Argumentation: A Simple Example

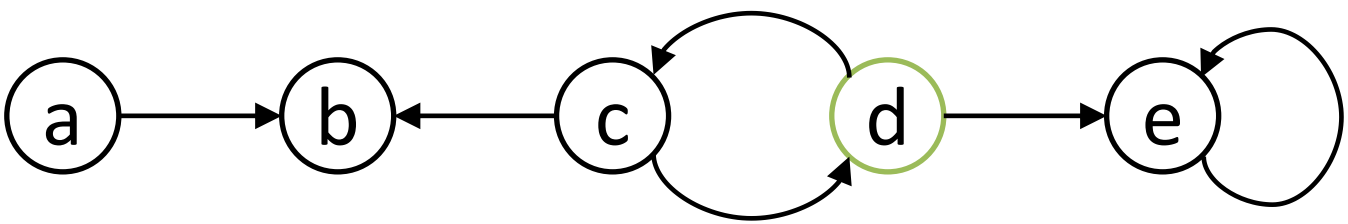
- Conflict Free Set:



- $cf(F) = \{\{a\}, \{b\}, \{c\}$

Argumentation: A Simple Example

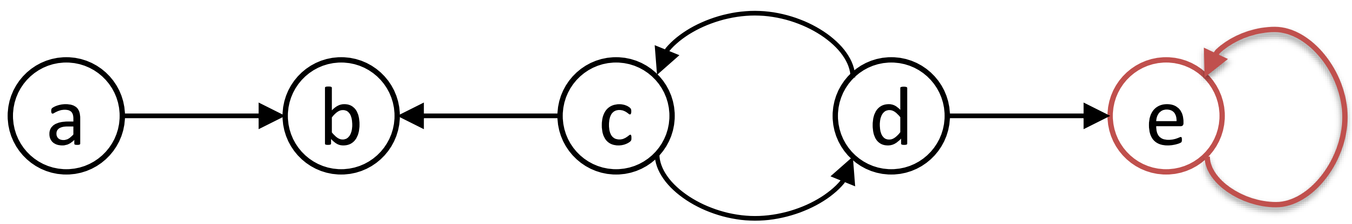
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}$

Argumentation: A Simple Example

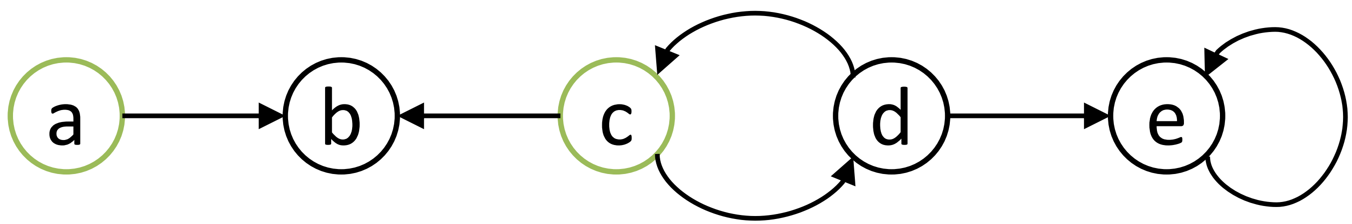
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}$

Argumentation: A Simple Example

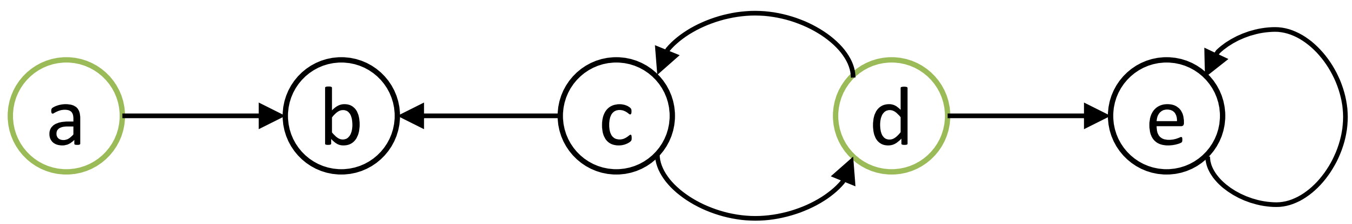
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}\}$

Argumentation: A Simple Example

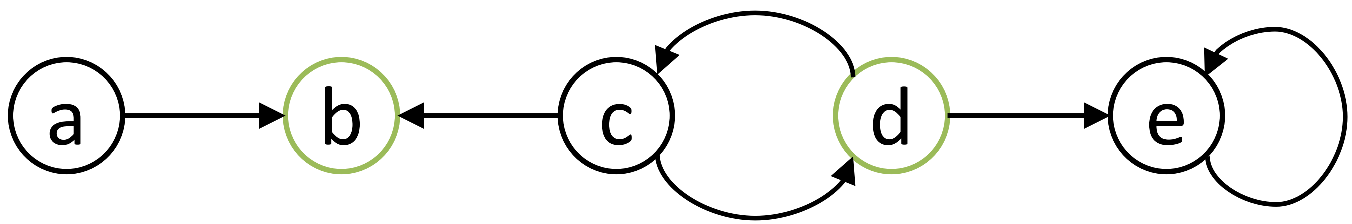
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}\}$

Argumentation: A Simple Example

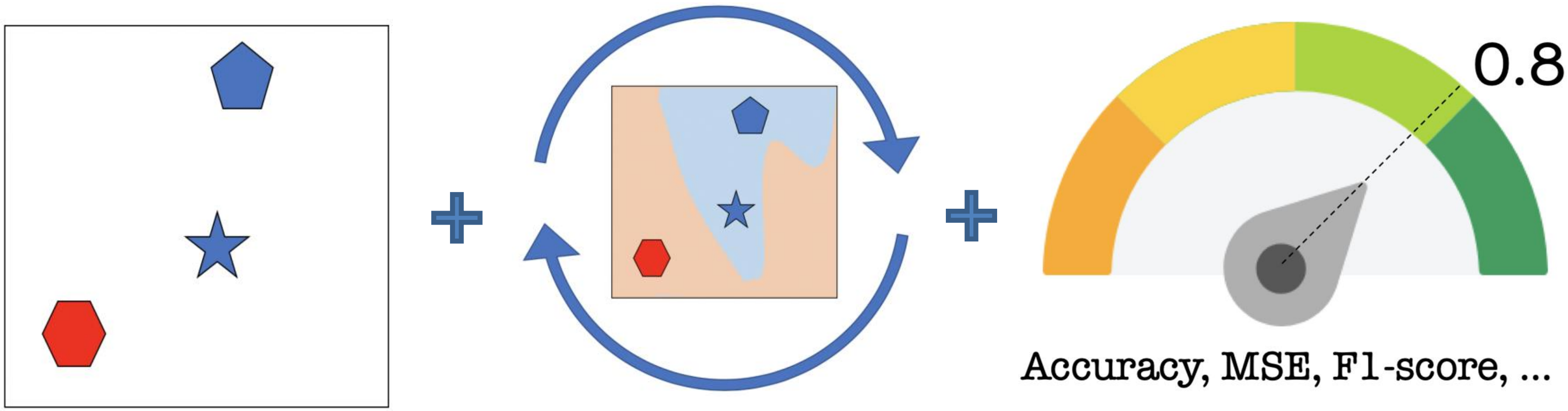
- Conflict Free Set:



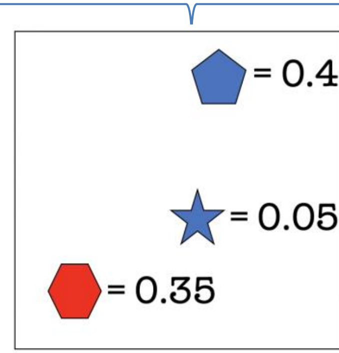
- $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}, \{b, d\}, \emptyset\}$

Explainable Deep Learning through Data Relevance Analysis

AI and Data Contribution



Train Data | Learning Algorithm | Performance Evaluation

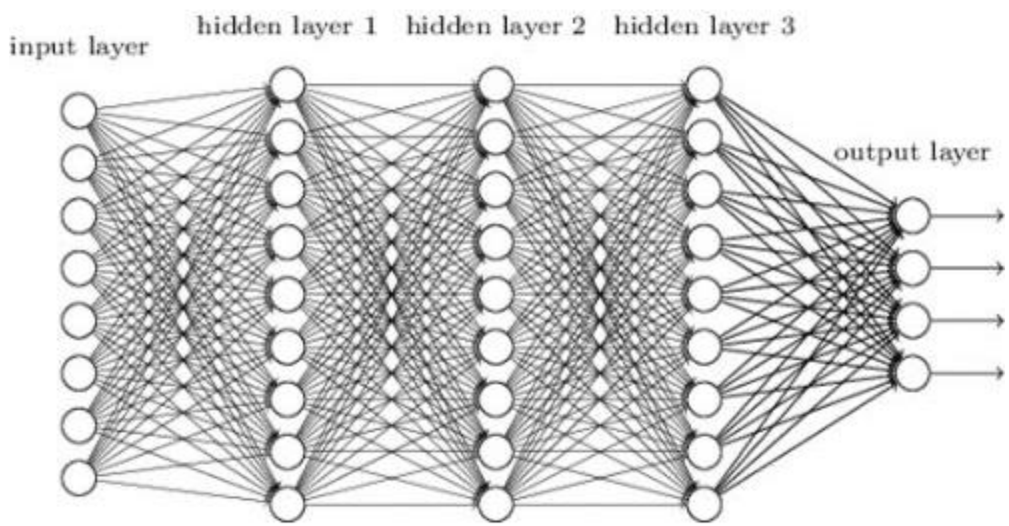


Value

How to quantify?

Interpreting Neural Networks

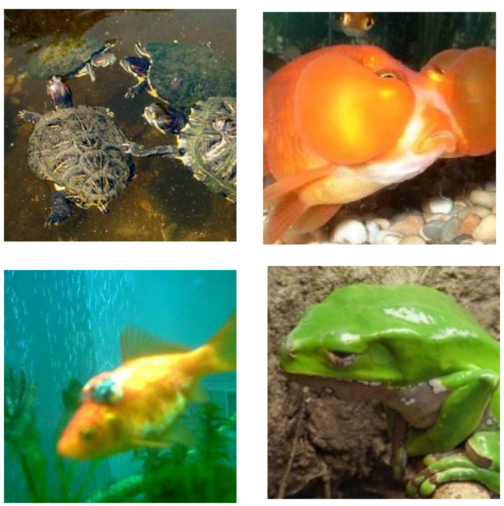
- Interpreting black-box neural networks helps training, auditing, and debugging
- Trust is gained when you can explain why you make certain decisions

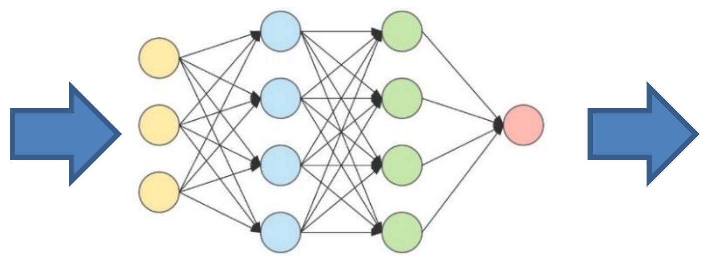


Data-based Model Interpretability

- How do the training data contribute to the model performance on the test sample?
- Does each training sample contribute positively or negatively?

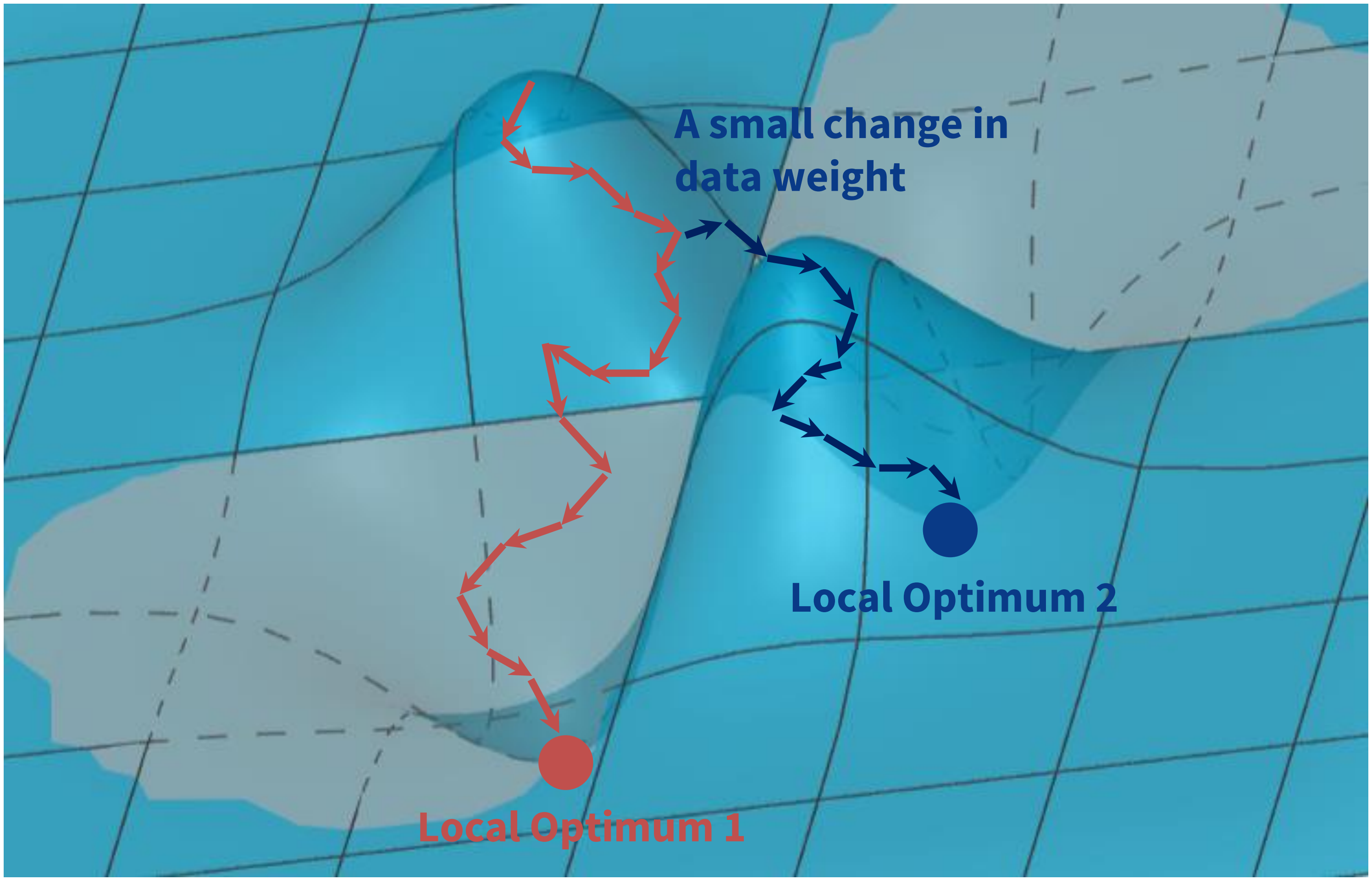
Training Data





Prediction: Goldfish

Training Data Influence the Entire Optimization Trajectory



Influence Function

- What is the influence of a training sample on the model (or on the loss of a test sample)?

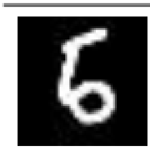
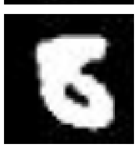
Optimal model param. : $\hat{\theta} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta)$

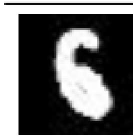
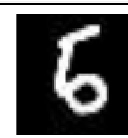
Model param. by training w/o z : $\hat{\theta}_{-z} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \sum_{z_i \neq z} L(z_i, \theta)$

Model param. by upweighting z : $\hat{\theta}_{\epsilon, z} \stackrel{\text{def}}{=} \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z, \theta)$

Influence Function

- With the influence of upweighting a sample *x* on the parameters, we can **linearly approximate** the parameter changes due to removing *x* without retraining the model.

Training Sample	True vs. Predicted Labels	Model Conf.	Contribution to Test Data
	5 / 5	0.66	-1.0×10^{-4}
	8 / 8	0.73	4.8×10^{-5}

Test Sample	Influencer	Contrib.	True / Predicted Label	Model Conf.
		-0.31	6 / 5	0.78
		0.091	8 / 3	0.80

Useful Applications

- Understanding DL model behaviours
- Debugging DL models
- Fixing wrongly labelled training data samples

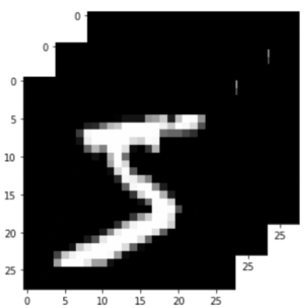
Key Challenge

- Computational complexity is high (involving computing large Hessian-vector products (HVPs))
- Current solution:
  - Approximating HVPs with less computationally expensive means, possibly at the cost of some performance loss (https://arxiv.org/abs/2102.02515)

Explainable Federated Learning through Shapley Value

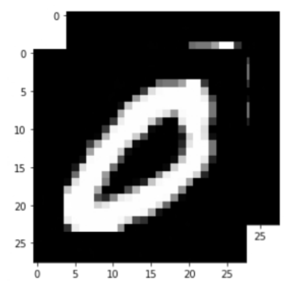
Data Valuation Metrics

- Quantity, Quality, Label Quality



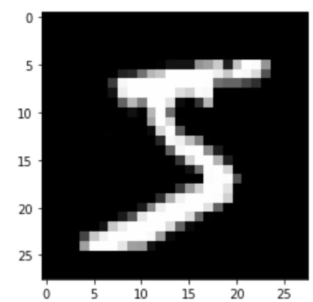
300 images

V.S.



200 images

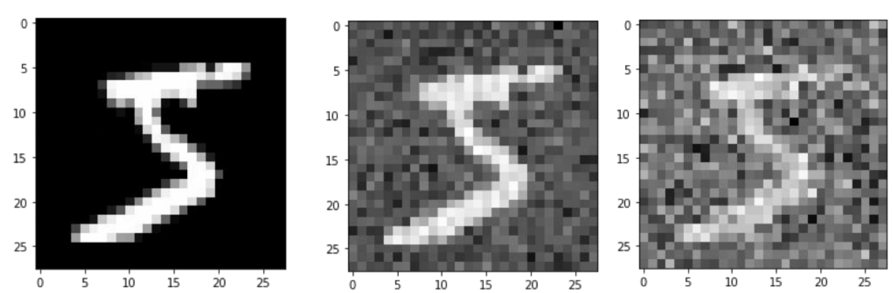
“Quantity”



wrong label “1”

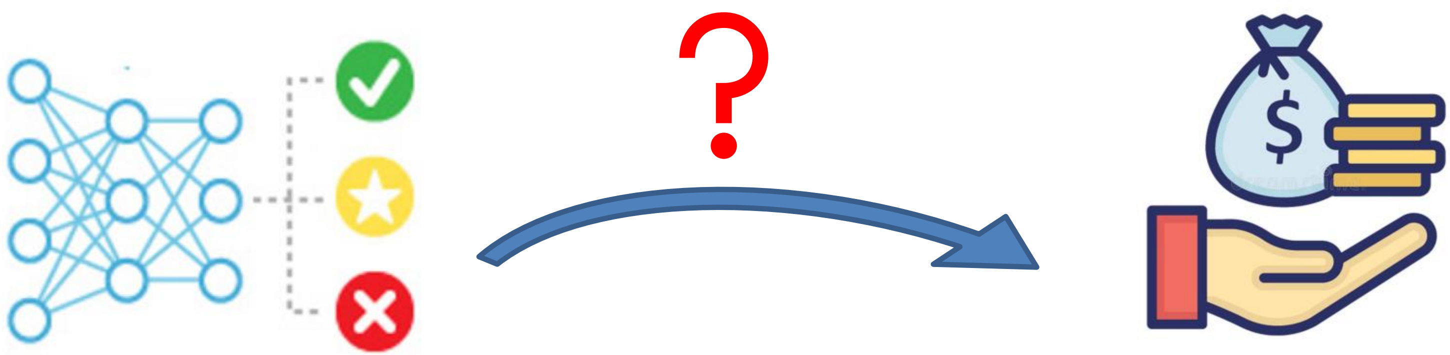
“Label Quality”

“Quality”



Data Valuation Obstacle in FL

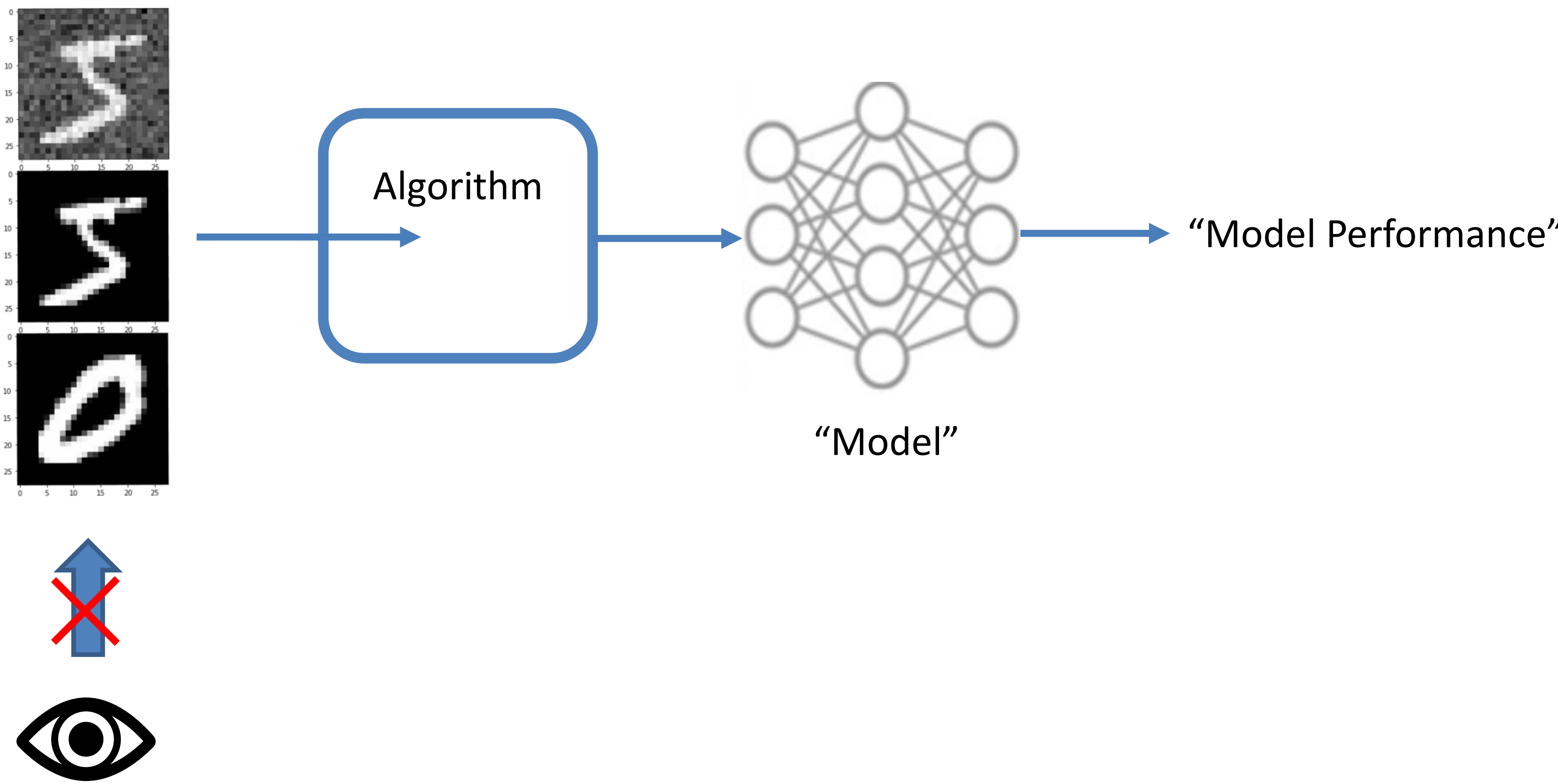
- Evaluate contributions without seeing actual dataset?



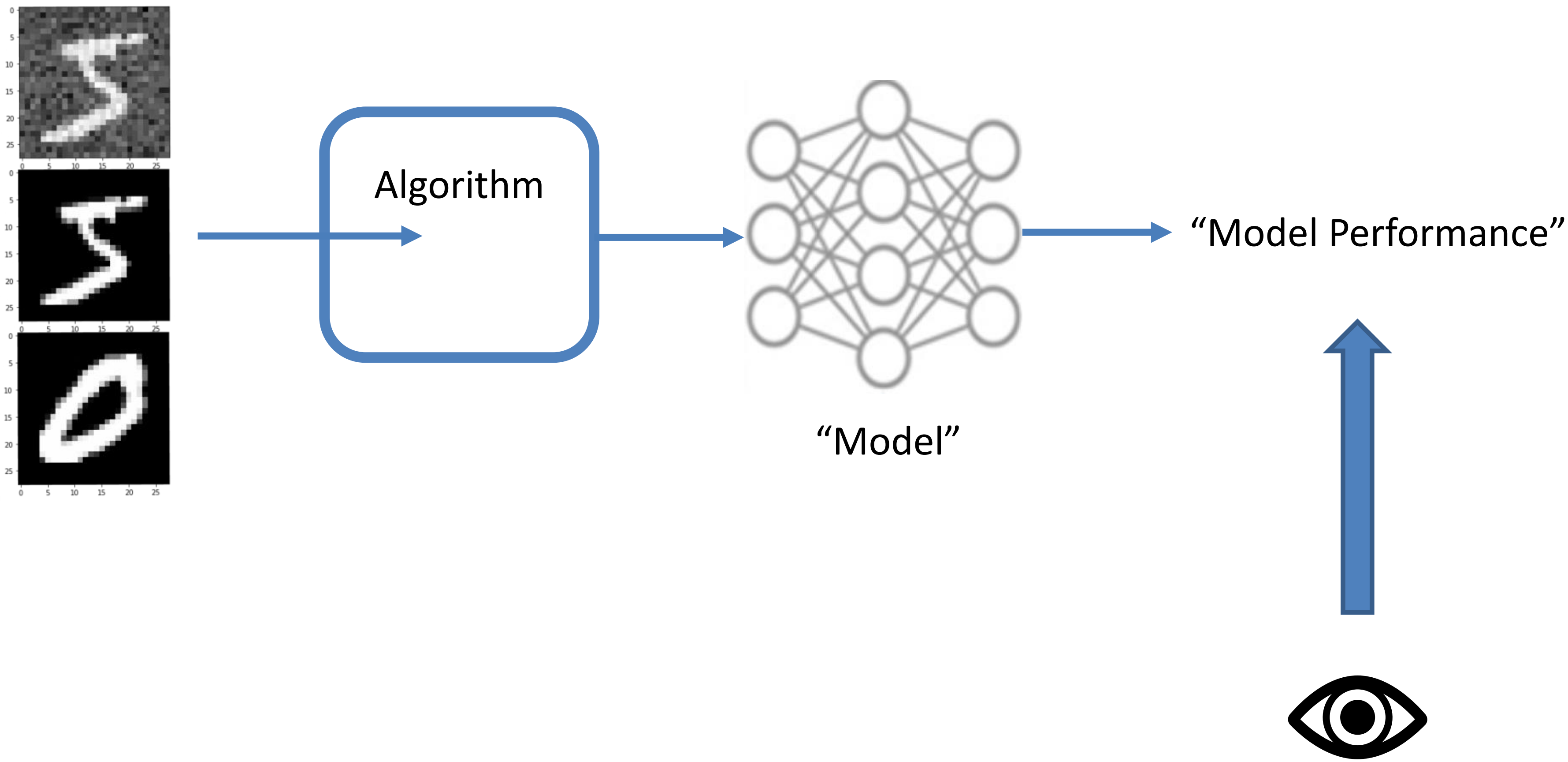
Model Performance

Individual Contribution

Data Valuation Obstacle in FL



Data Valuation Obstacle in FL

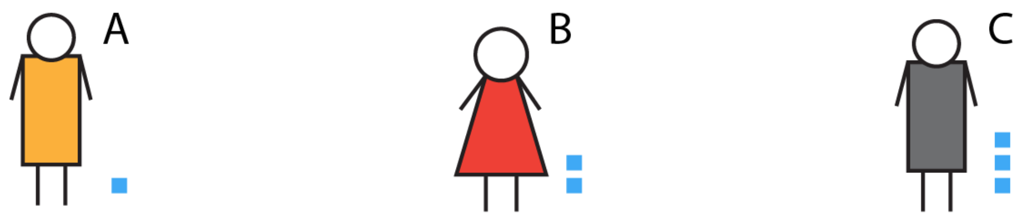


Data Valuation Principles

- Efficiency: all contributions add up to 100%.
- Symmetry: Two contributors' contributions should show the same value if they join FL in different orders.
- Free-rider: Outcomes of any grouping won't change regardless of whether the contributor joins or not.
- Linearity: If divided into two parts, one's contribution is the sum of contributions in two parts.

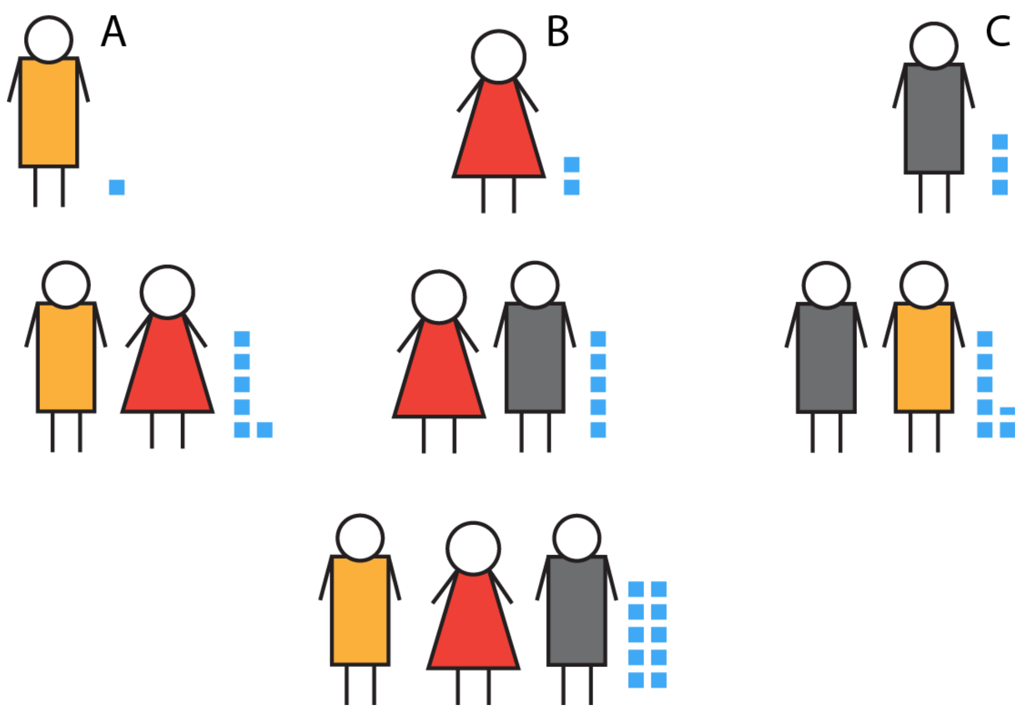
Shapley Value – An Example

- Example: A, B and C work together in a project worth 100 points.
- How many points should each of them get?



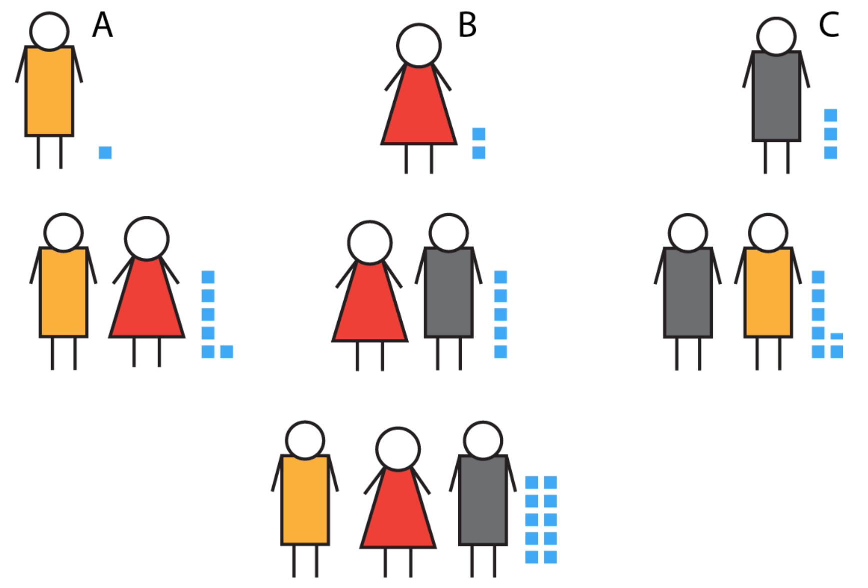
Shapley Value – An Example

- V(A)=10, V(B)=20, V(C)=30
- V(AB)=60, V(BC)=50, V(AC)=65, V(ABC)=100



Shapley Value – An Example

- B-C-A: (A,B,C)=(50,20,30)
- C-A-B: (A,B,C)=(35,35,30)
- A-C-B: (A,B,C)=(10,35,55)
- C-B-A: (A,B,C)=(50,20,30)
- B-A-C: (A,B,C)=(40,20,40)



Shapley Value – An Example

- A=(10+50+35+10+50+40)/6=195/6=32.5
- B=(50+20+35+35+20+20)/6=180/6=30
- C=(40+30+30+55+30+40)/6=225/6=37.5

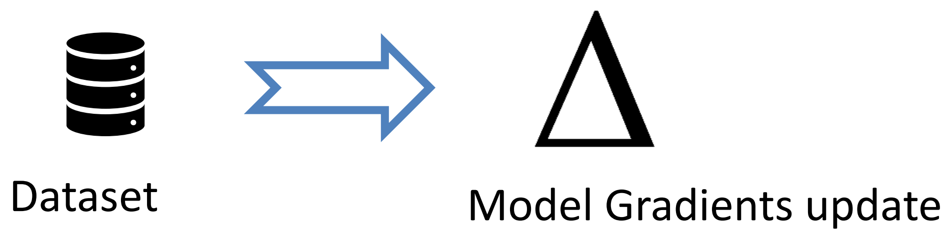
Limitations of Shapley Value in FL

- Fair but Inefficient
  - For a coalition with N participants, Shapley needs to train at least 2^N models to evaluate.

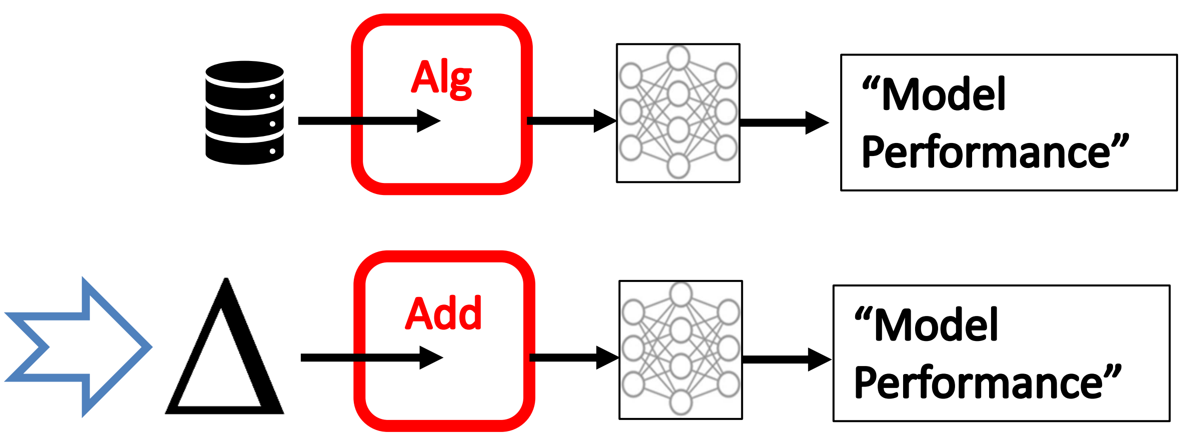
- If the FL model is a highly complex neural network, a single training session is already computationally expensive.

Guided Truncation Gradient (GTG)-Shapley

- **Transforming** evaluation process from training to gradient-based FL model reconstruction.

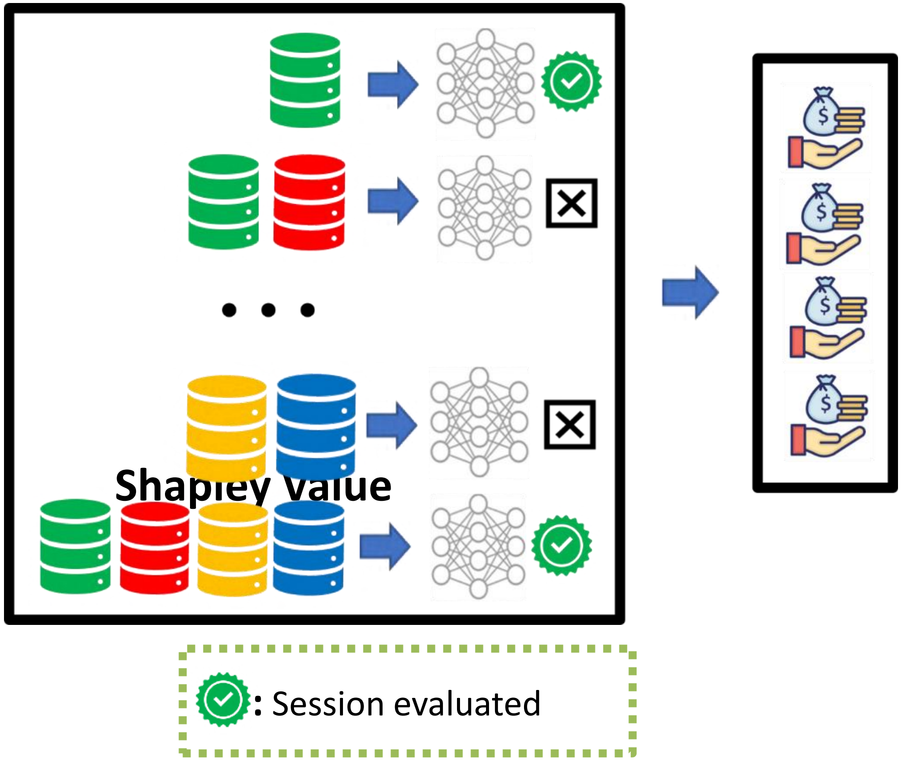


- Process



Guided Truncation Gradient (GTG)-Shapley

- Monte-Carlo Approximation with Truncation on unnecessary sub-model evaluations.
- Only those bringing performance change larger than threshold will be retained.



Resource: https://arxiv.org/pdf/2109.02053.pdf

Interesting Reading

Alejandro Barredo Arrieta *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, vol. 58, pp. 82-115 (2020)

Final Note

- Explainability is not a static concept.
- There is a spectrum of explanations w.r.t. AI.
- The definition of explainability can change based on the industry, the competitive. landscape, the regulatory environment, and the customer base.
- Decisions to adopt certain explainable AI techniques must be made by AI solution designers on a case-by-case basis.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Explainable AI

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University