SIGMETRICS
Measurements
for SelfDN

Unlocking the Black Box with Trustee: XAI for Networking Problems

Ronaldo A. Ferreira[3]

Joint work with Arthur Selle Jacobs[1], Roman Beltiukov[4] , Walter Willinger[2], Arpit Gupta[4], Lisandro Z. Granville[1]

June 19th, 2023

1 INSTITUTO DE INFORMÁTICA UFRGS
2 NIKSUN
3 UFMS
4

Traditional AI/ML Pipeline

- The “traditional AI/ML pipeline” consists of:
  - A training task characterized by a model specification
  - A (labelled) training dataset
  - An independent and identically distributed (IID) evaluation procedure
  - A (single) metric (e.g., F1-score) that measures the model’s expected predictive performance on data drawn from the training distribution

Traditional AI/ML Pipeline

- Main criticisms of the “traditional ML pipeline” include:
  - The pipeline’s output (i.e., best-performing ML model) is prone to suffer from the problem of underspecification
  - The evaluation of the pipeline’s output is agnostic to the particular inductive biases encoded by the trained model

Underspecification issues!

Shortcut Learning

Model takes shortcuts to classify data!

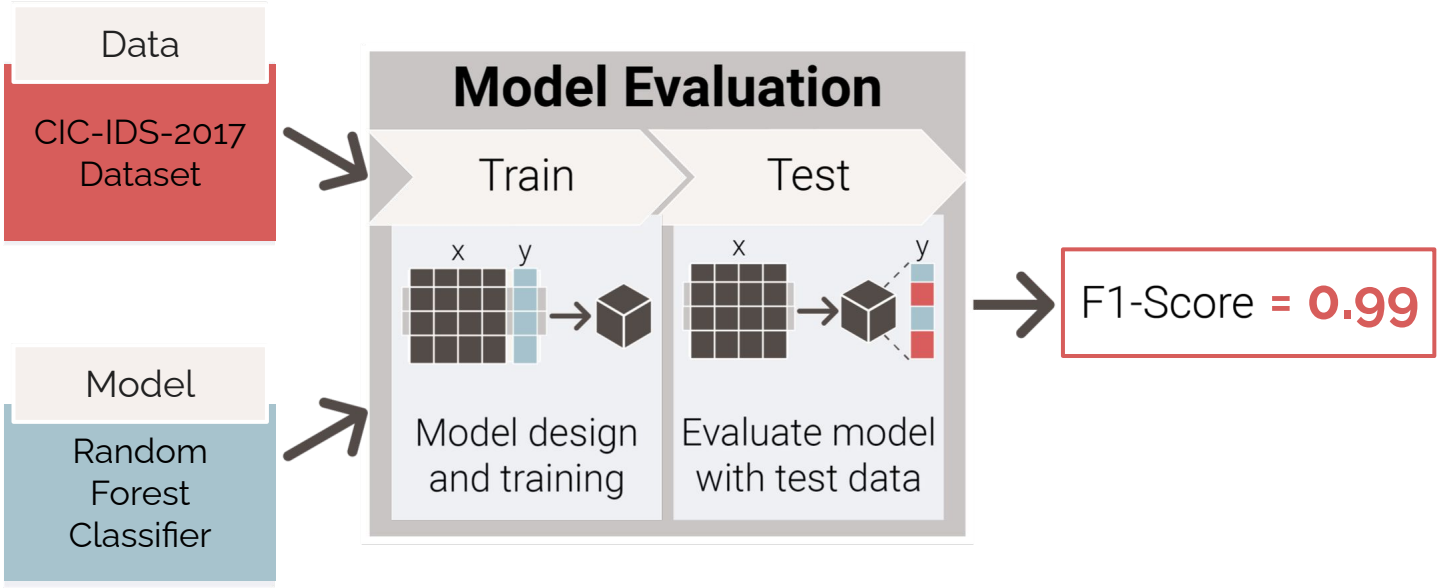
O.O.D. Samples

Model does not generalize!

Spurious Correlations

Model picks up wrong correlations in the data!

Consider this example…



Can you trust this model?

Can you trust this model?

Trust in AI/ML model

=

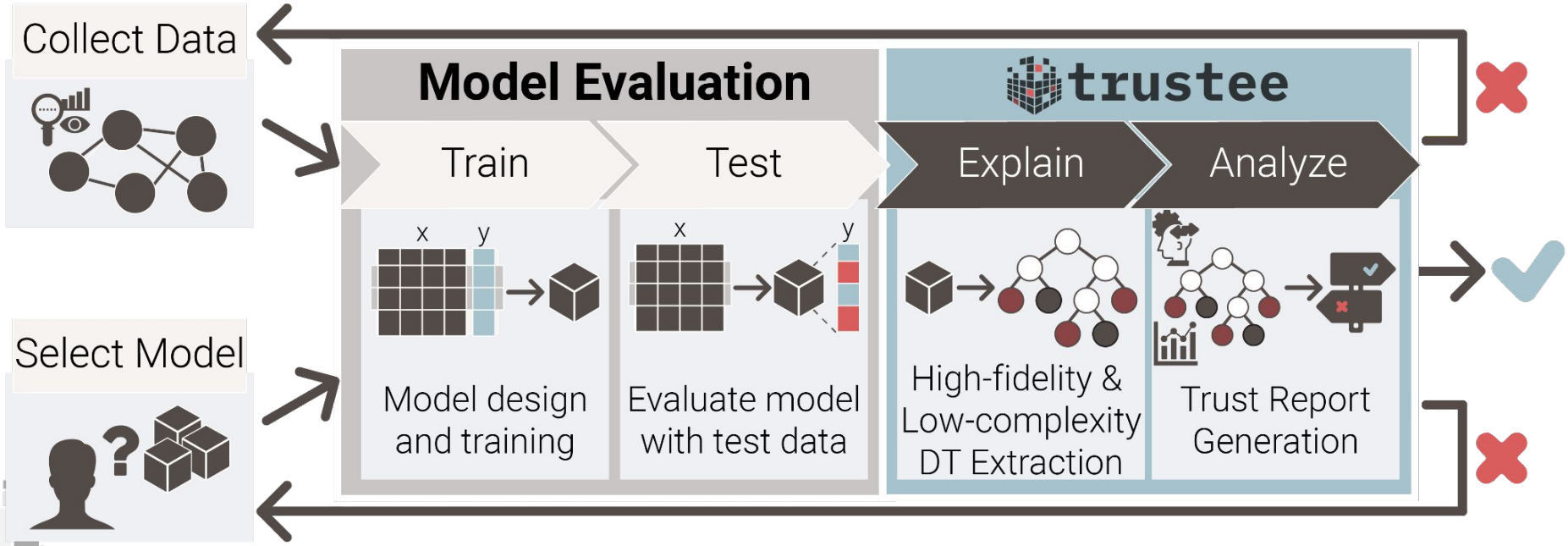
Hand over control to the AI/ML model

New AI/ML Pipeline

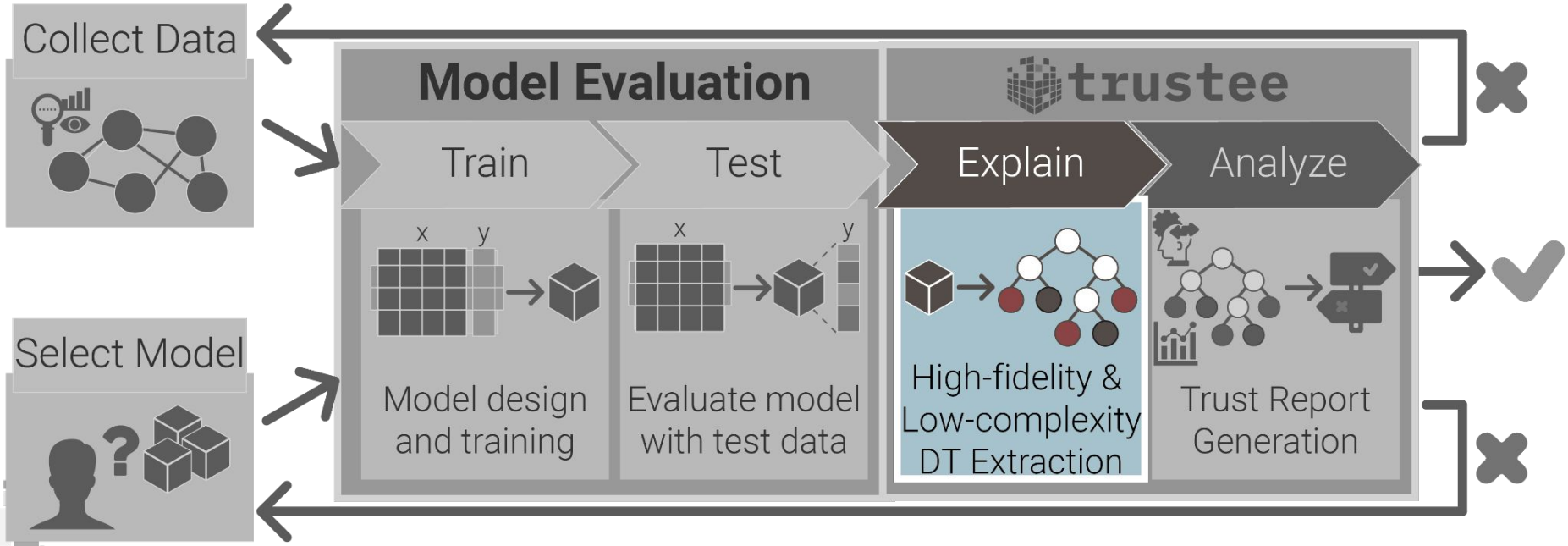
- AI/ML for networking is typically concerned with high-stakes decision-making
- ML models for networking should have the following properties:
  - It is less about understanding "effect":
    - showing the model works (e.g., has high F1-score) is insufficient
  - It is about understanding "cause"
    - "Why does the model make certain decisions (and not others)?
    - "When does the model not work?"
- We need a new AI/ML pipeline to fulfill these properties

trustee

Augmented
AI/ML Development Pipeline



Augmented
AI/ML Development Pipeline



trustee

Explanation Requirements

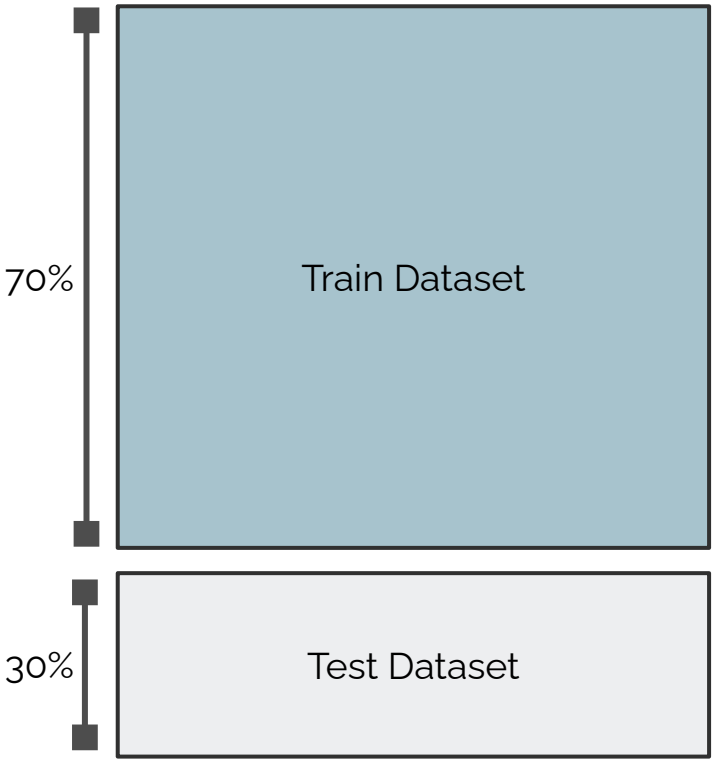
#1 Model Agnostic	#2 High Fidelity
#3 Low Complexity	#4 Stable

trustee

Dataset

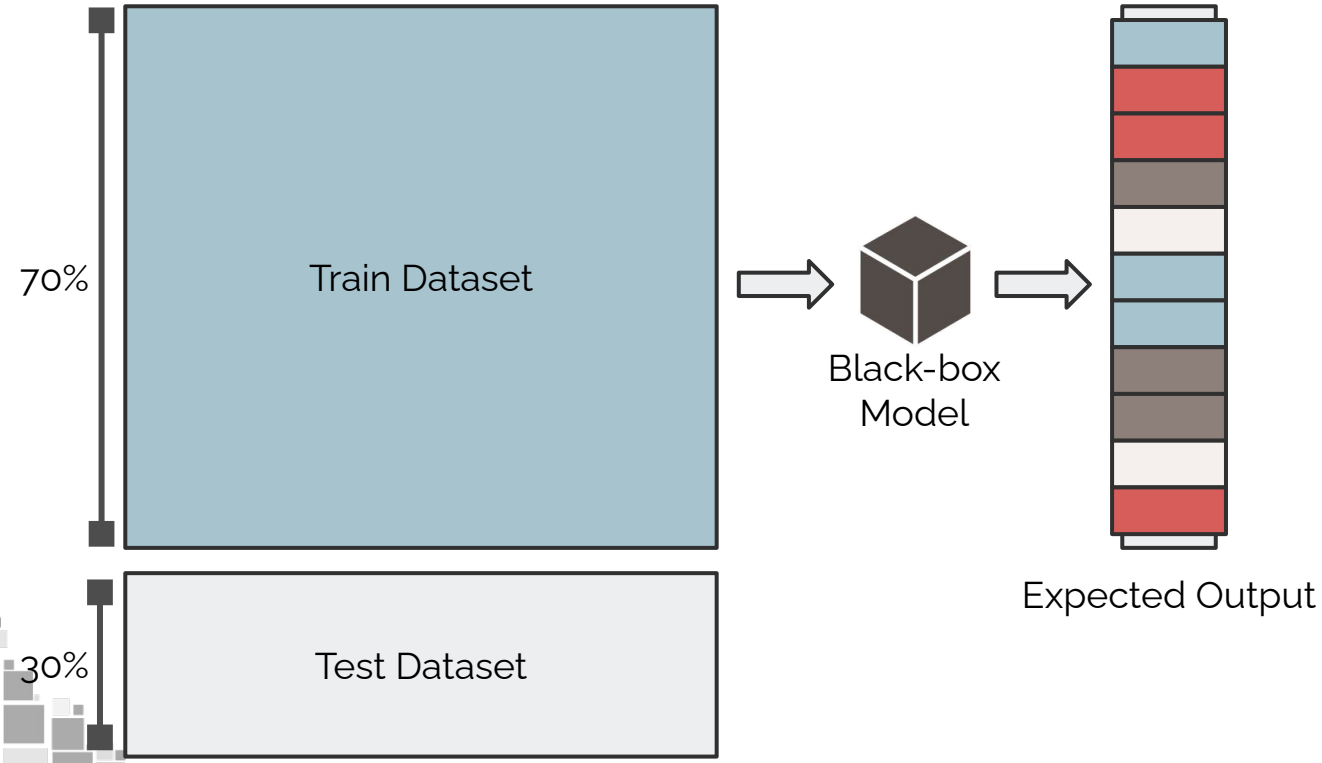
Black-box
Model

trustee



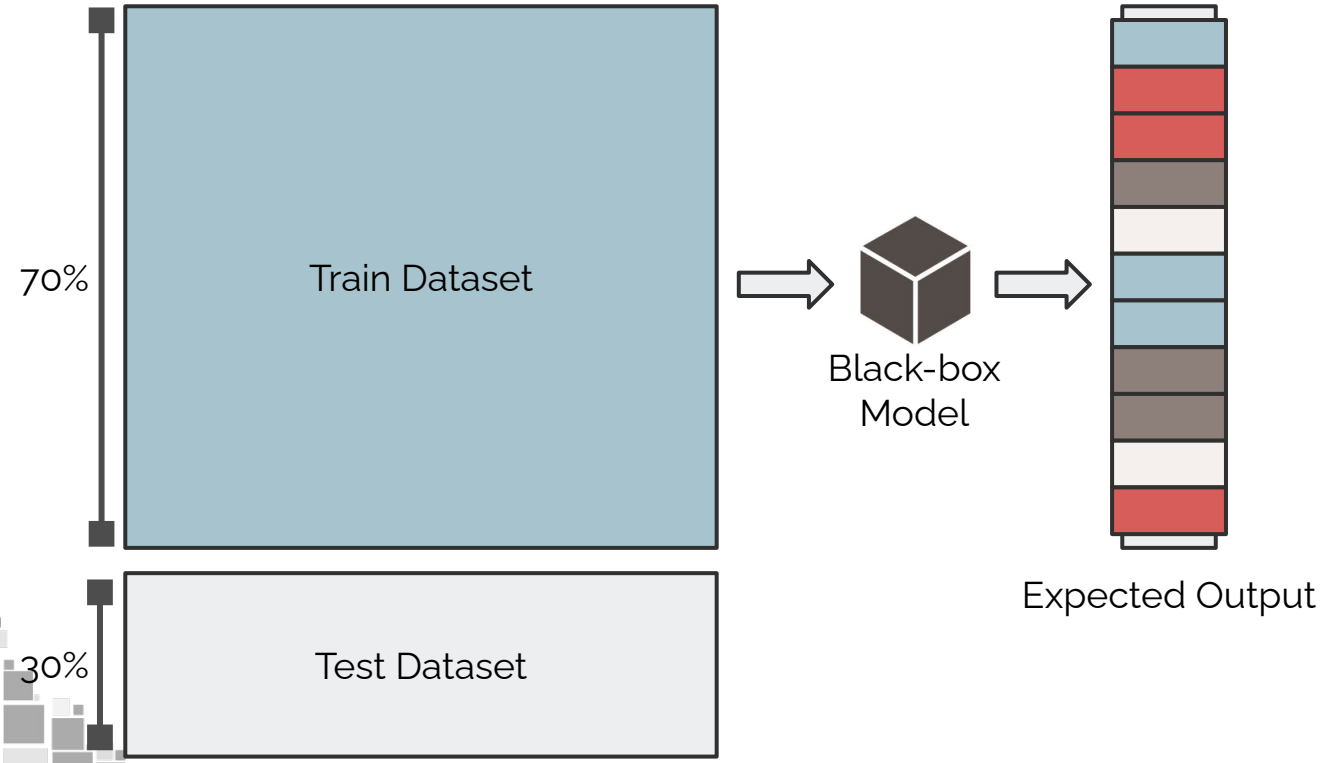
Black-box
Model

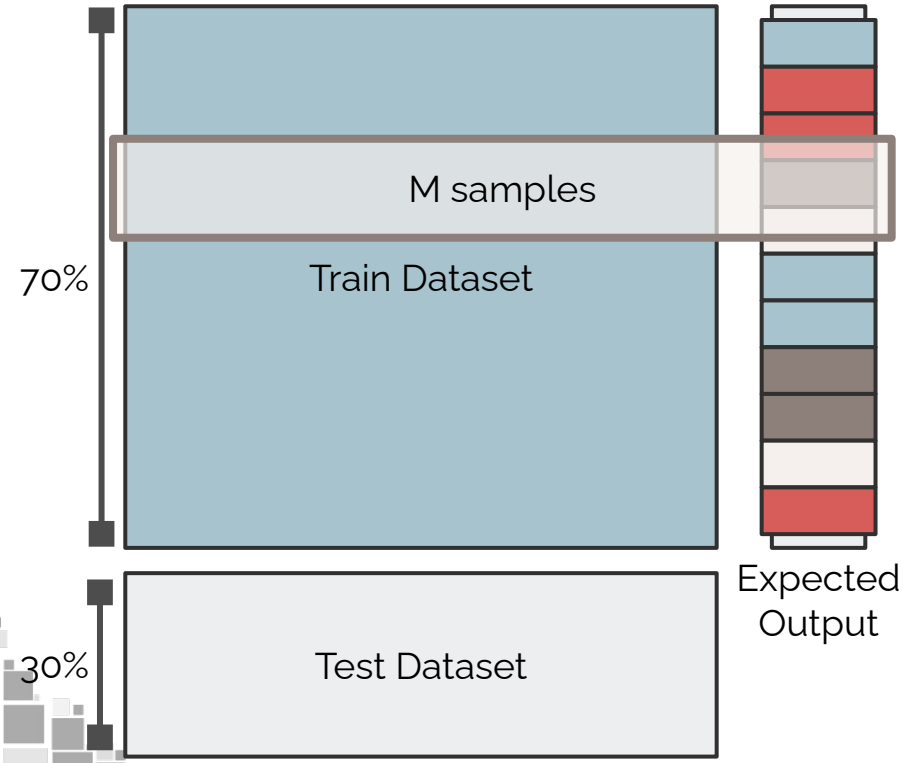
trustee


70%
Train Dataset
30%
Test Dataset
Black-box
Model
Expected Output

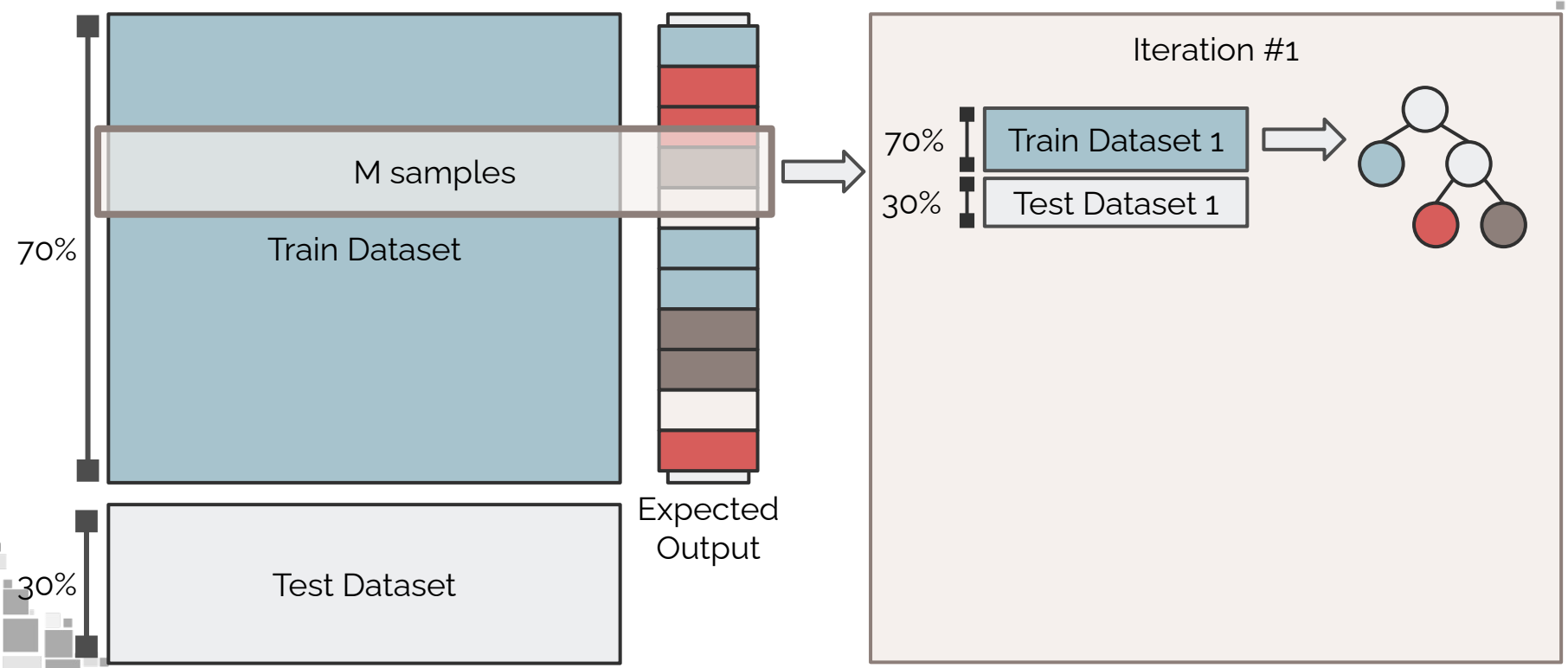
#1
Model
Agnostic

trustee

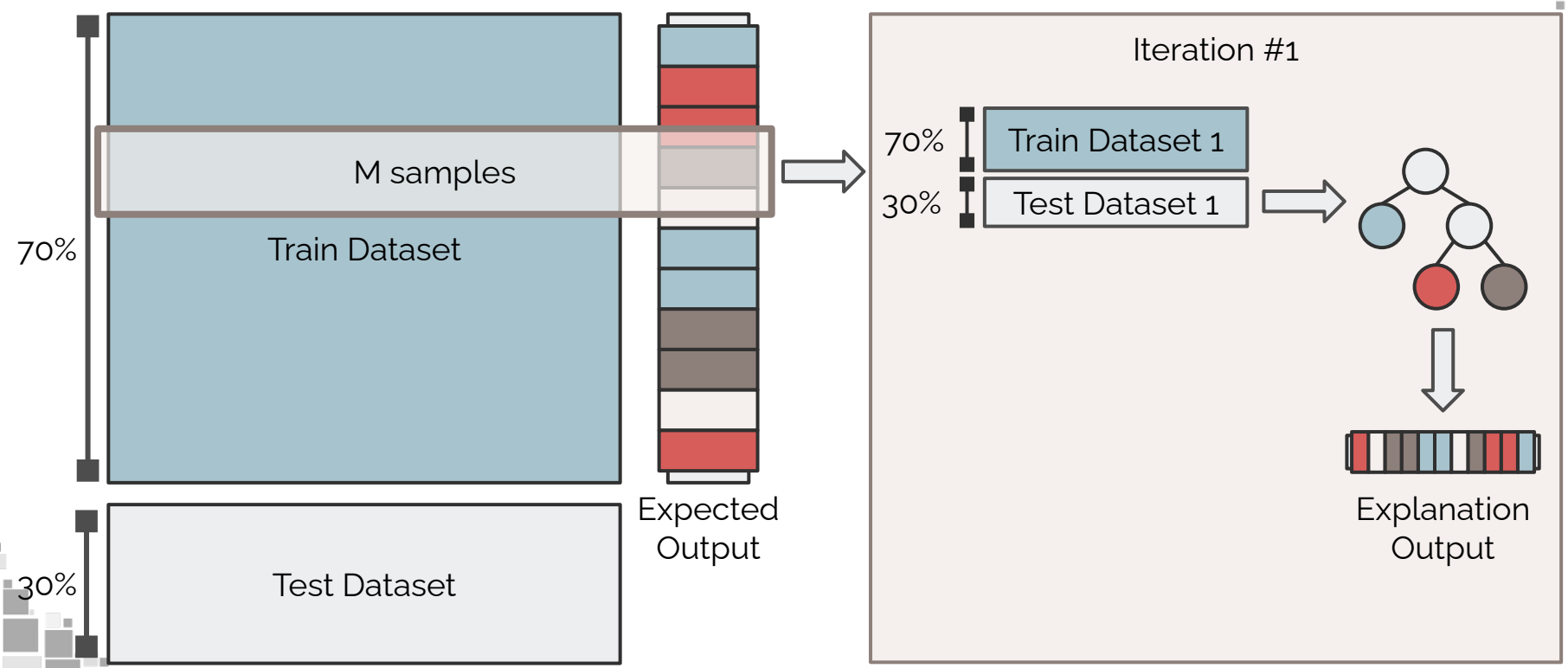



M samples
70%
Train Dataset
Expected
Output
30%
Test Dataset

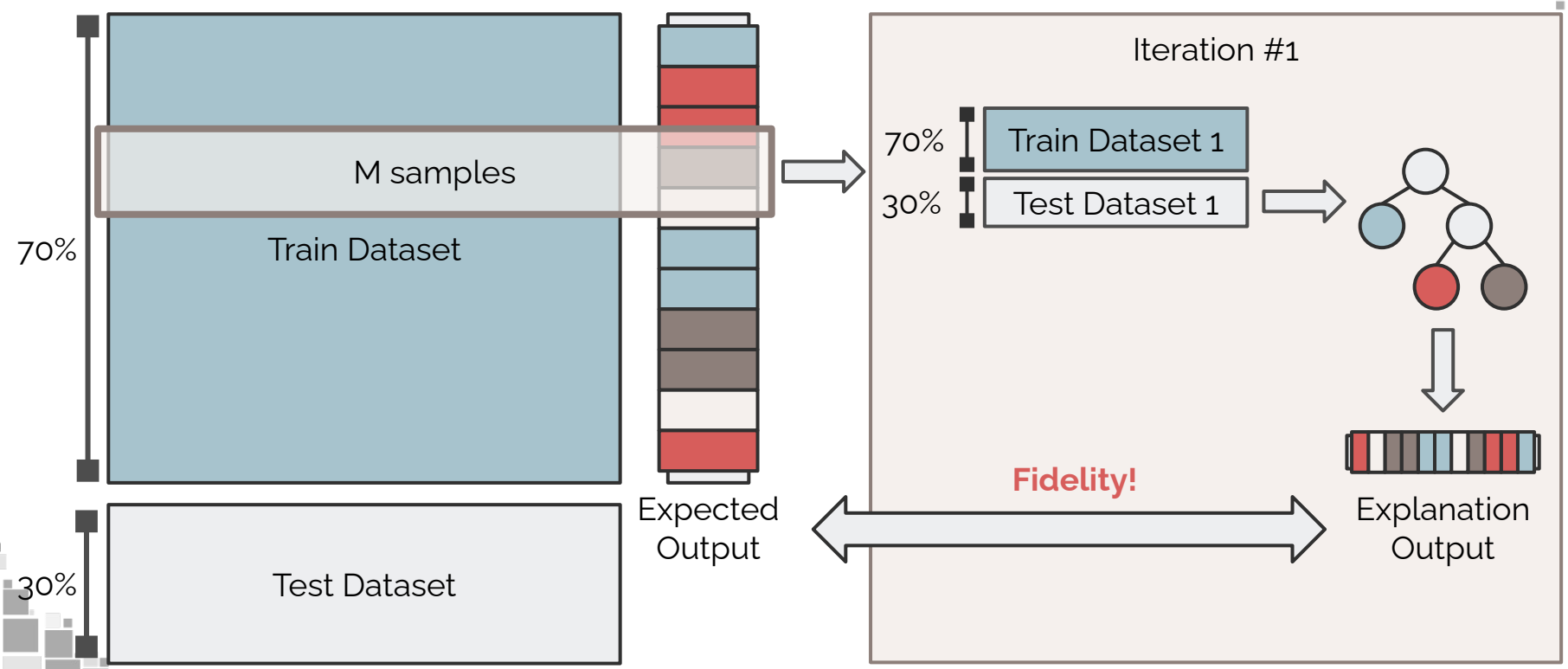
trustee


M samples
70%
Train Dataset
30%
Test Dataset
Expected Output
Iteration #1
70%
Train Dataset 1
30%
Test Dataset 1

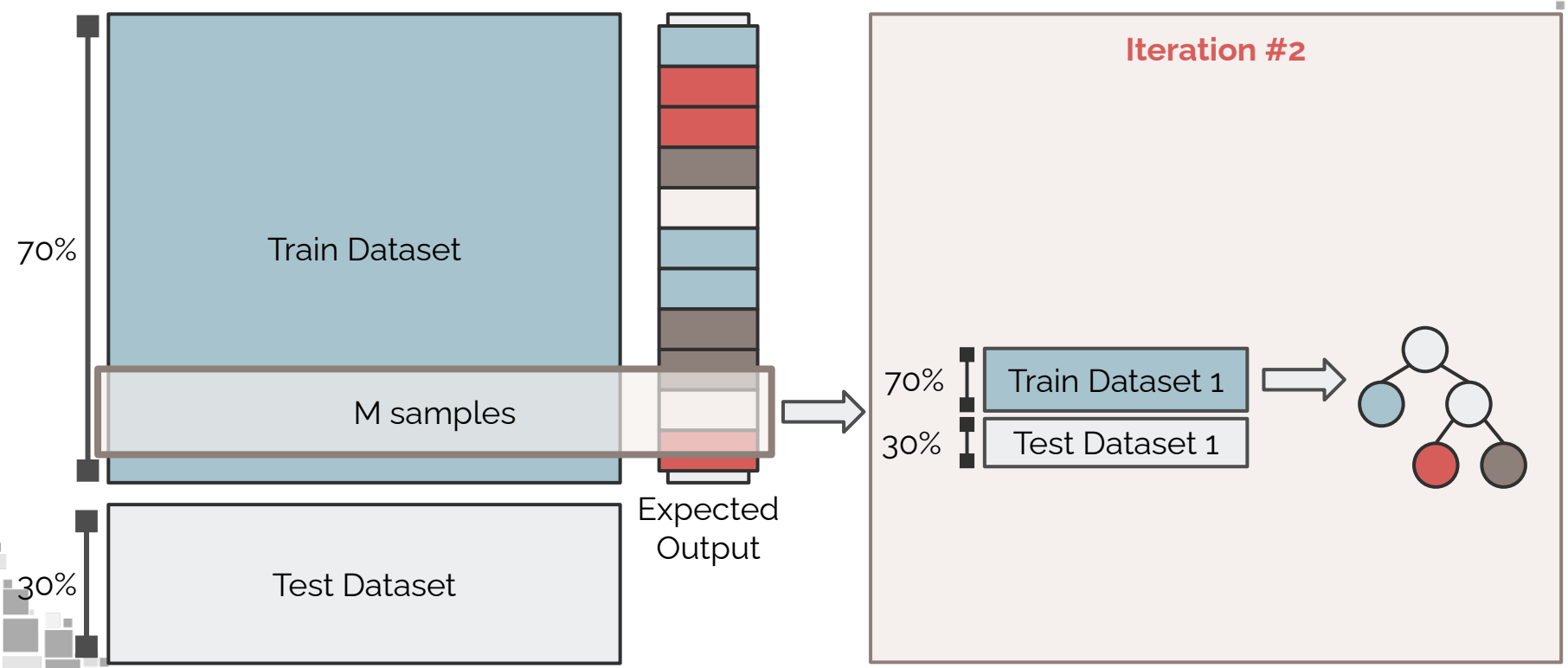
trustee


70%
M samples
Train Dataset
30%
Test Dataset
Expected Output
Iteration #1
70%
Train Dataset 1
30%
Test Dataset 1
Explanation Output

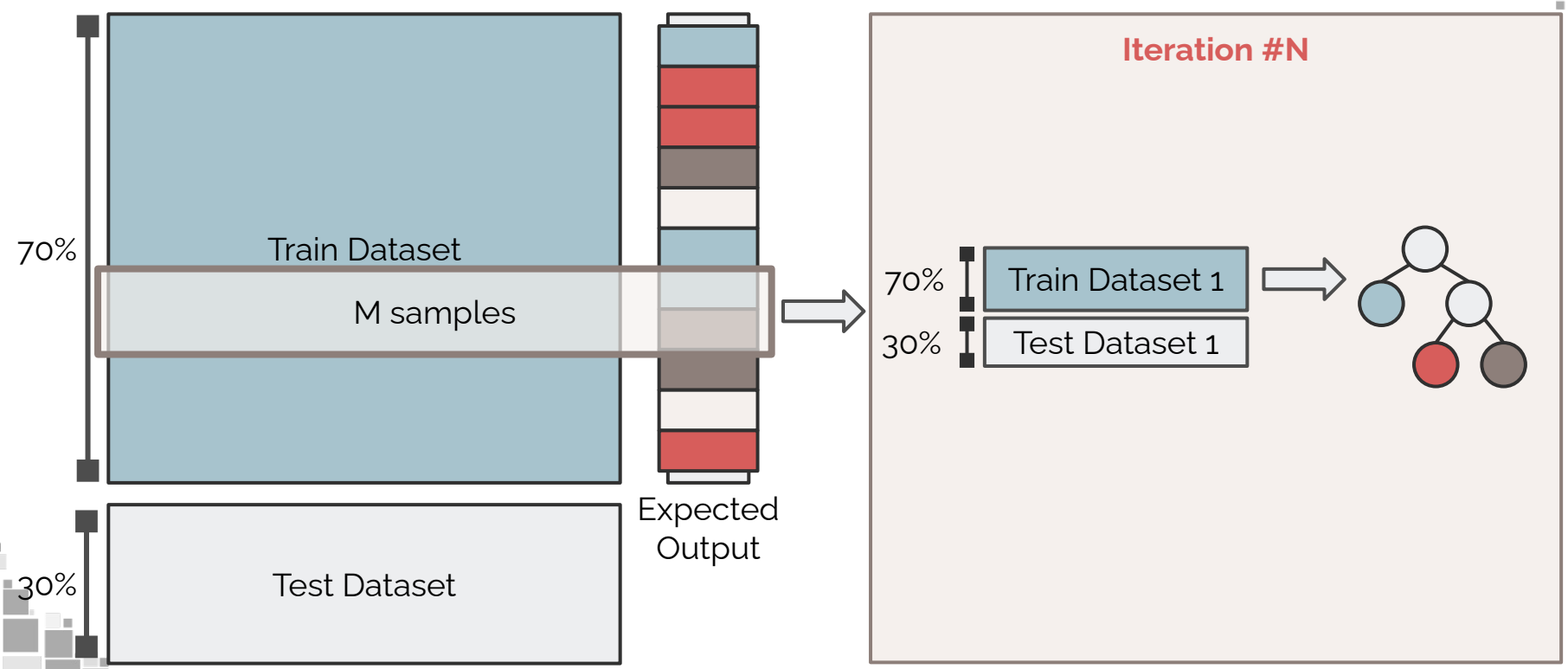
trustee


70%
M samples
Train Dataset
30%
Test Dataset
Expected Output
Iteration #1
70%
Train Dataset 1
30%
Test Dataset 1
Fidelity!
Explanation Output

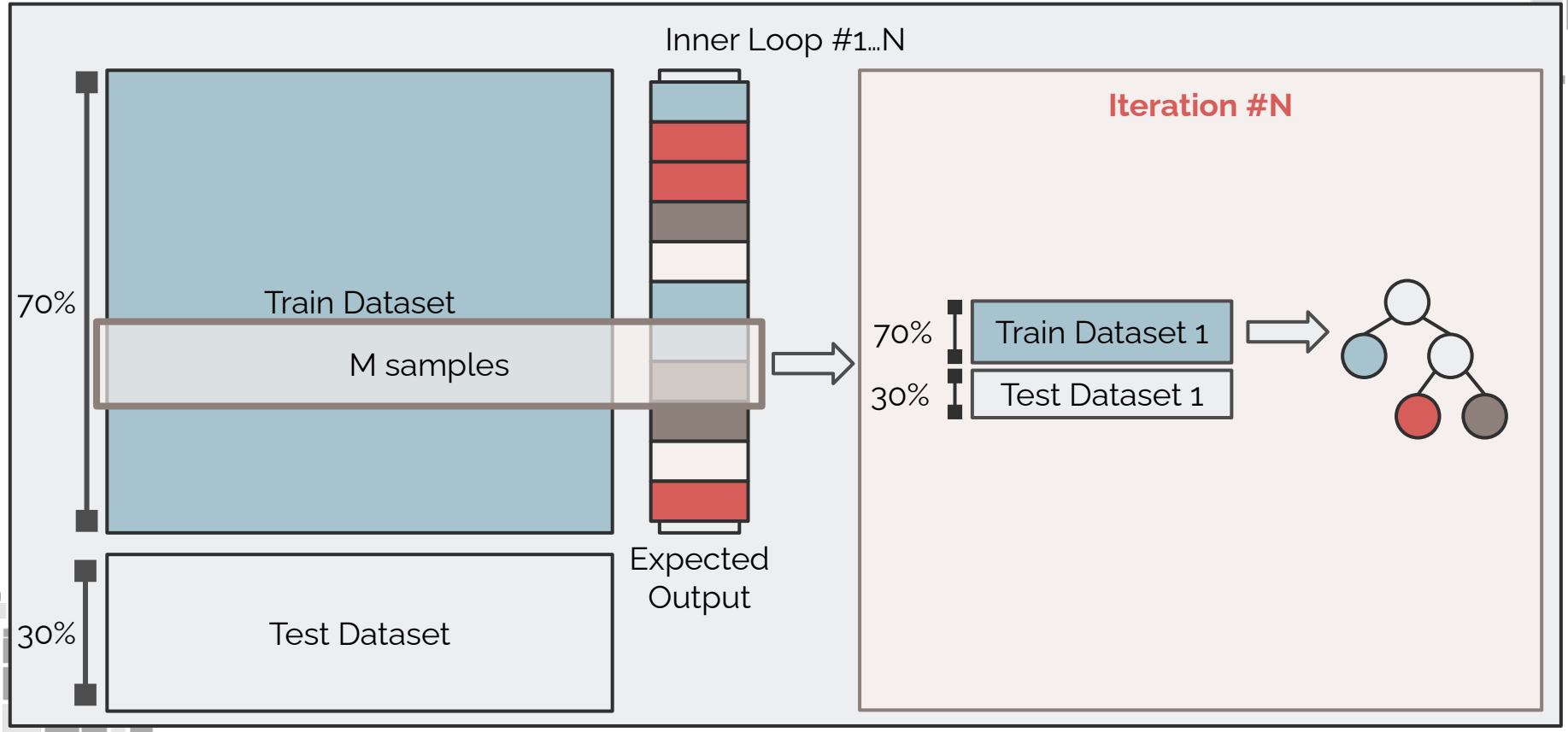
trustee


70%
Train Dataset
M samples
30%
Test Dataset
Expected Output
Iteration #2
70%
Train Dataset 1
30%
Test Dataset 1

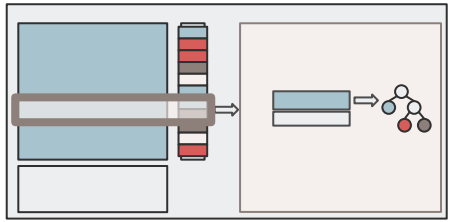
trustee

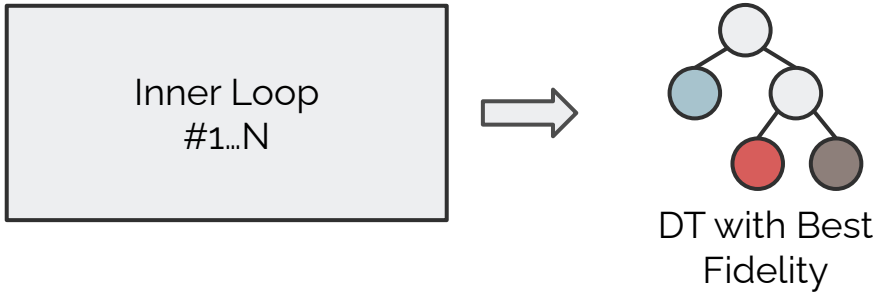

70%
Train Dataset
M samples
30%
Test Dataset
Expected Output
Iteration #N
70%
Train Dataset 1
30%
Test Dataset 1

trustee


Inner Loop #1…N
70%
Train Dataset
M samples
30%
Test Dataset
Expected Output
Iteration #N
70%
Train Dataset 1
30%
Test Dataset 1

trustee

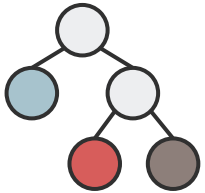



Inner Loop
#1…N
DT with Best
Fidelity

#2
High
Fidelity

trustee

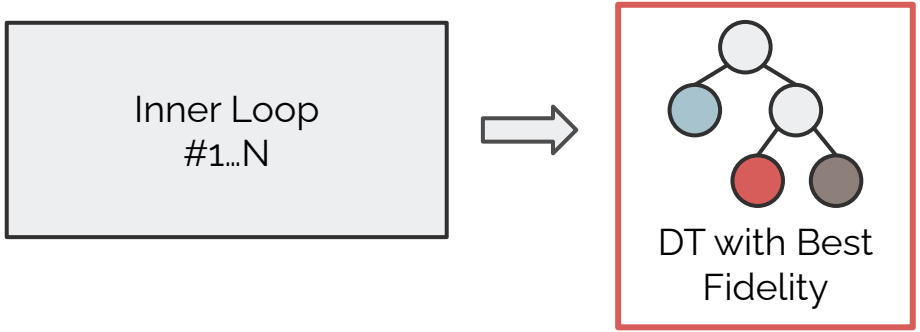
Inner Loop
#1…N



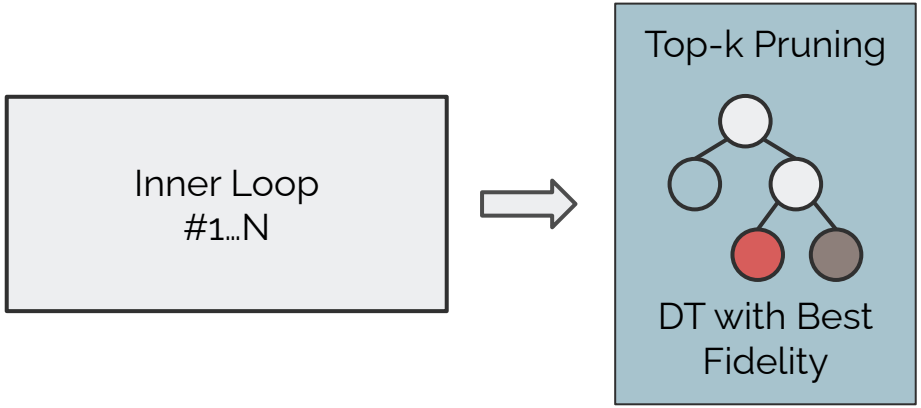
DT with Best
Fidelity

trustee

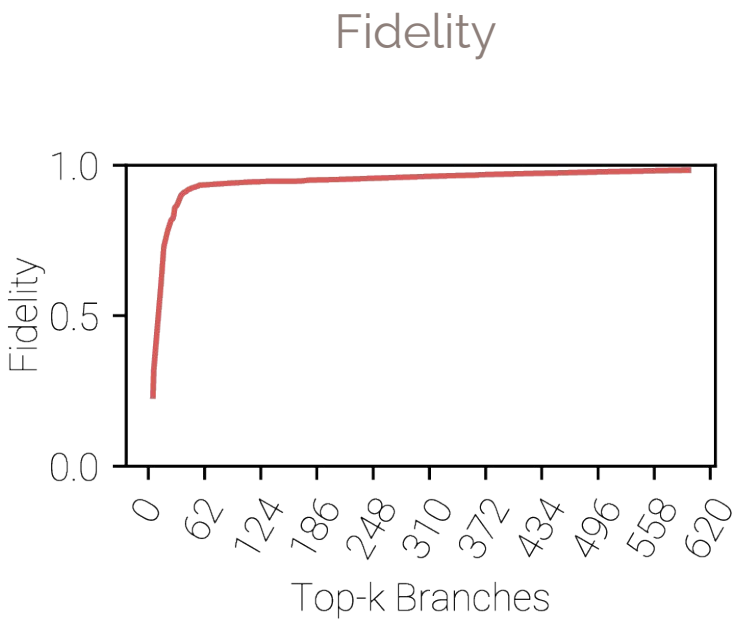
Size matters!

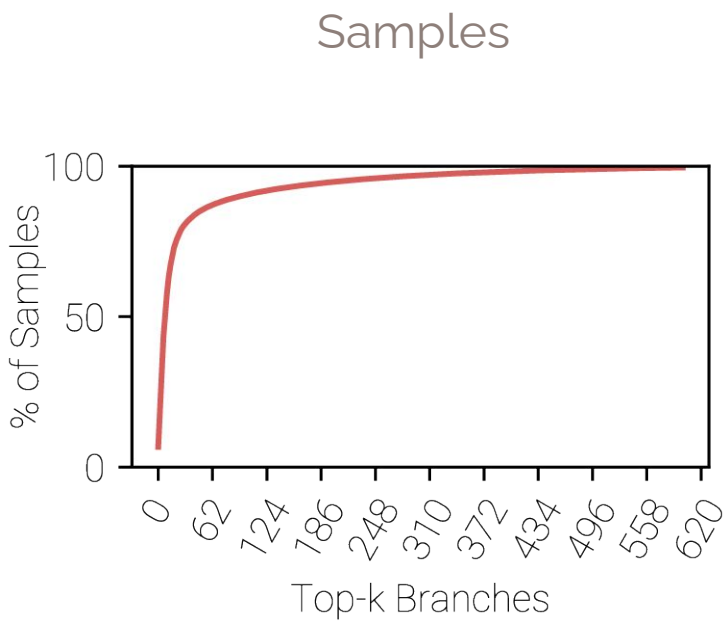


trustee

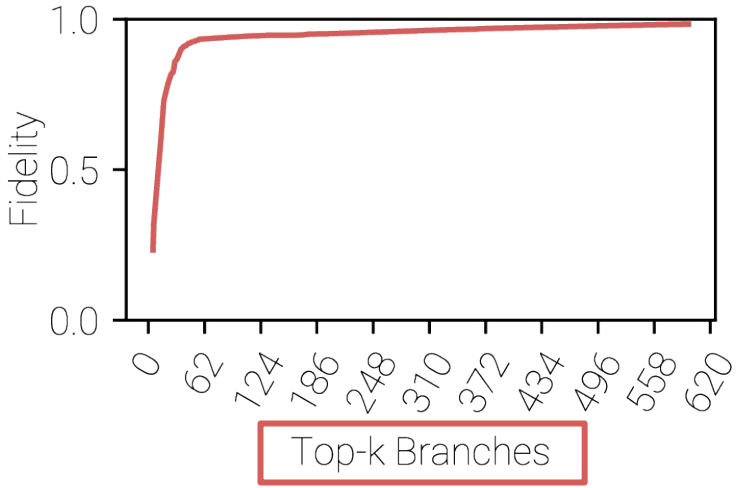


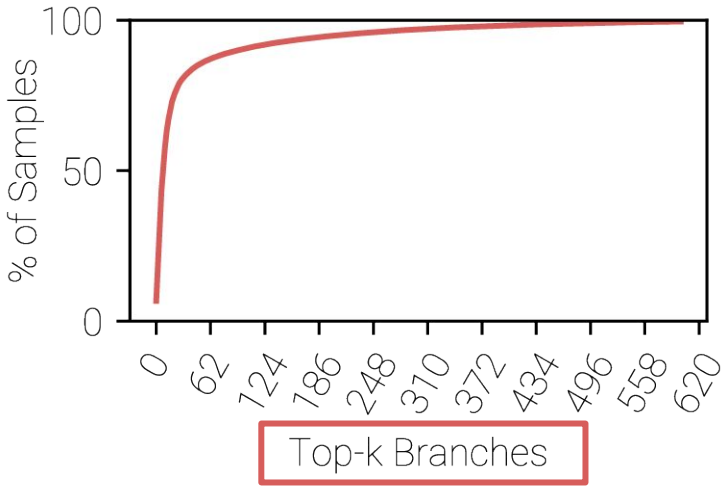
Top-k Pruning





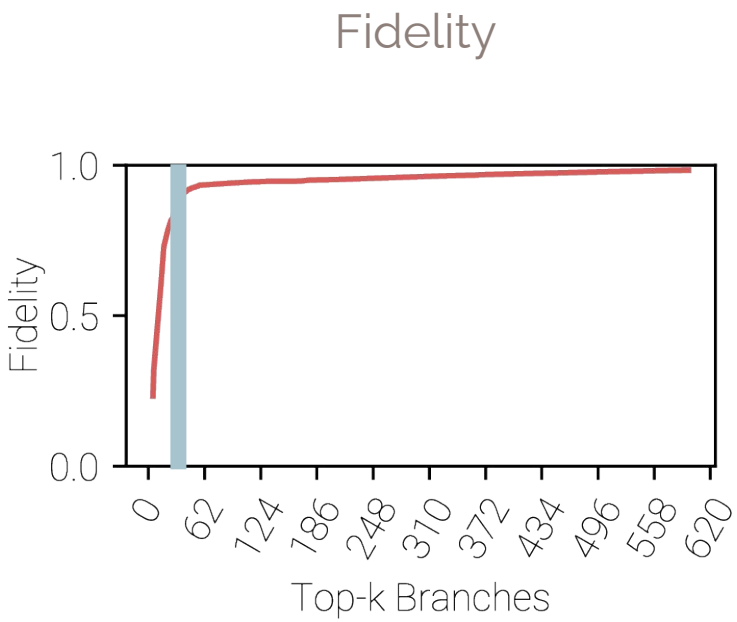
Top-k Pruning

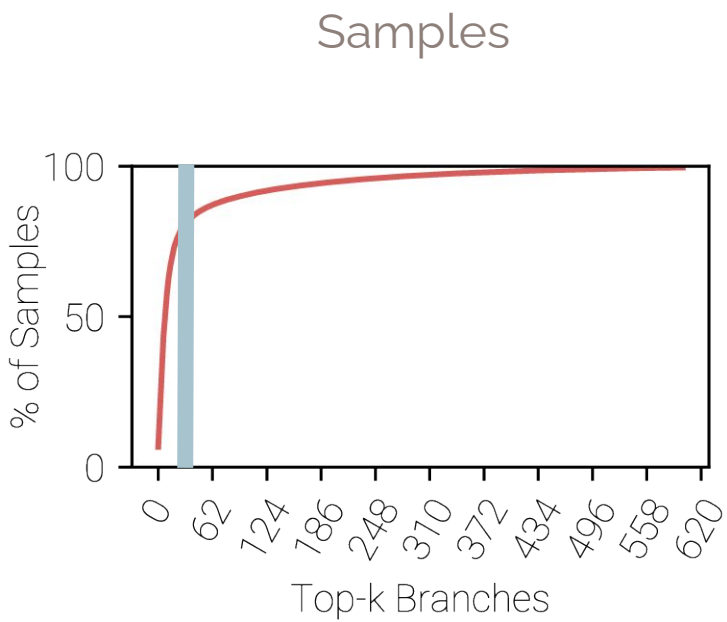




Diminishing returns!

Top-k Pruning



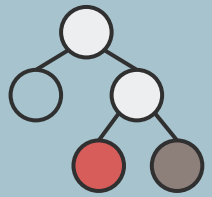


#3
Low
Complexity

trustee

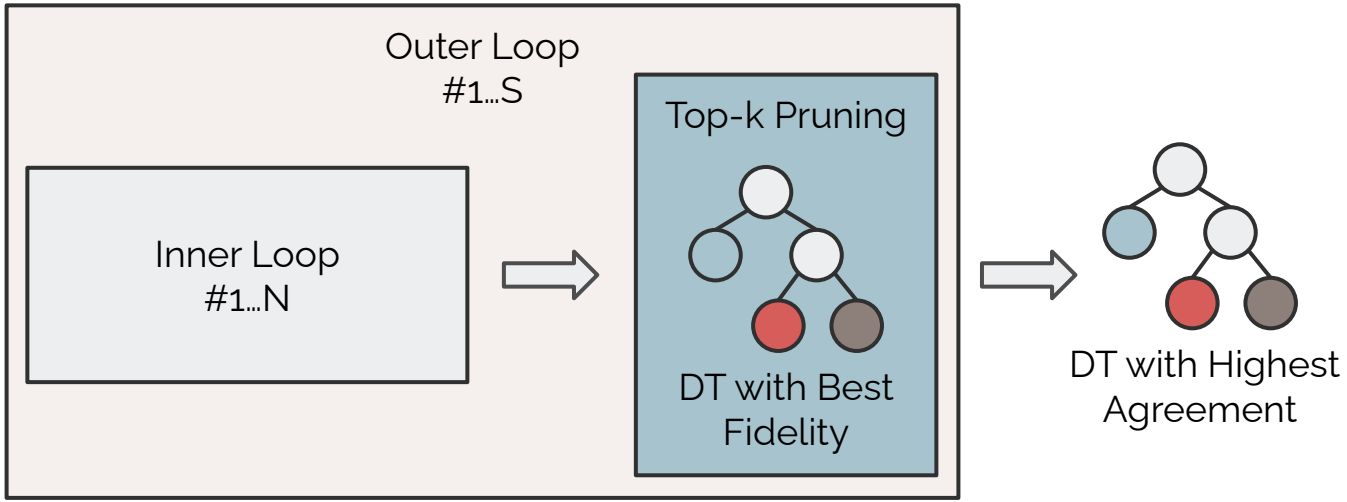
Inner Loop
#1…N

Top-k Pruning



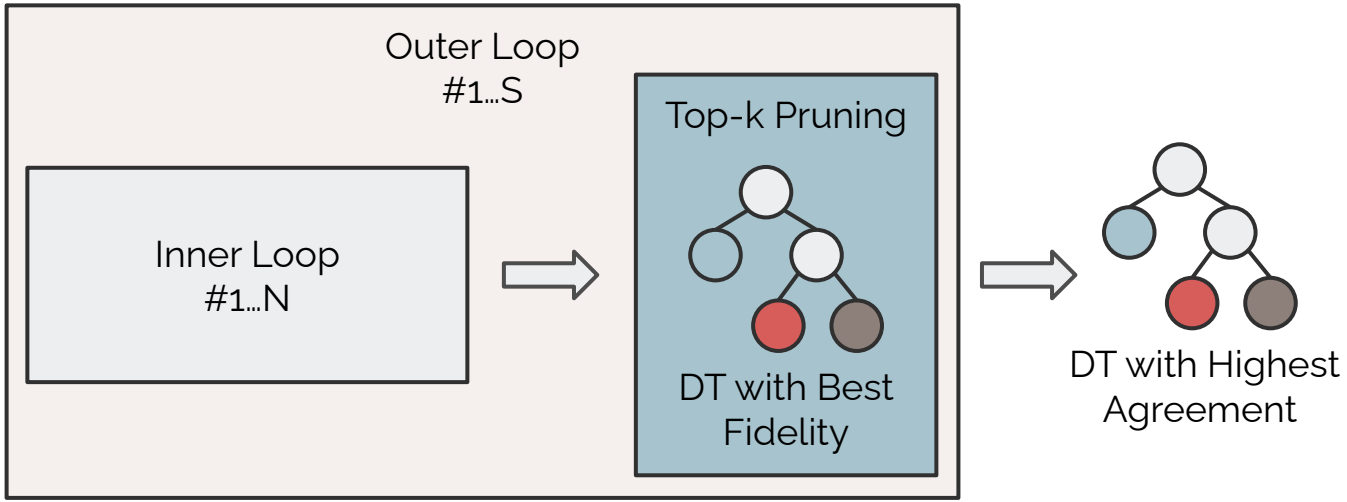
DT with Best
Fidelity

trustee


Outer Loop
#1…S
Inner Loop
#1…N
Top-k Pruning
DT with Best
Fidelity
DT with Highest
Agreement

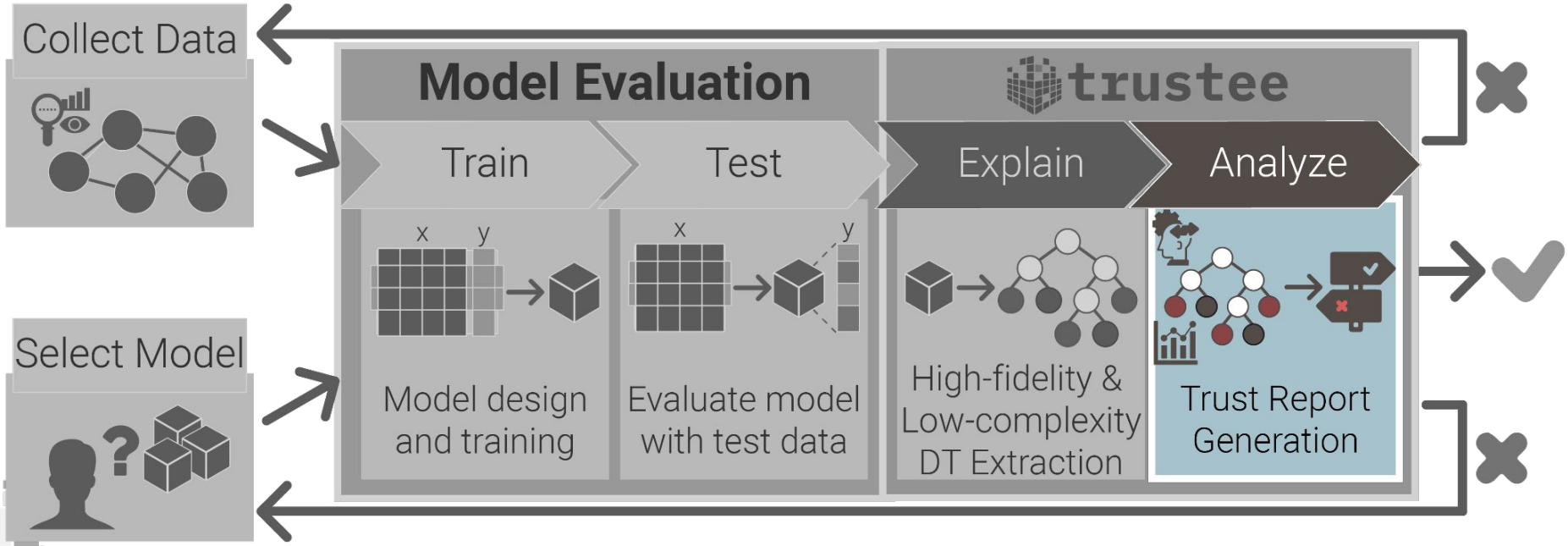
#4
Stable

trustee

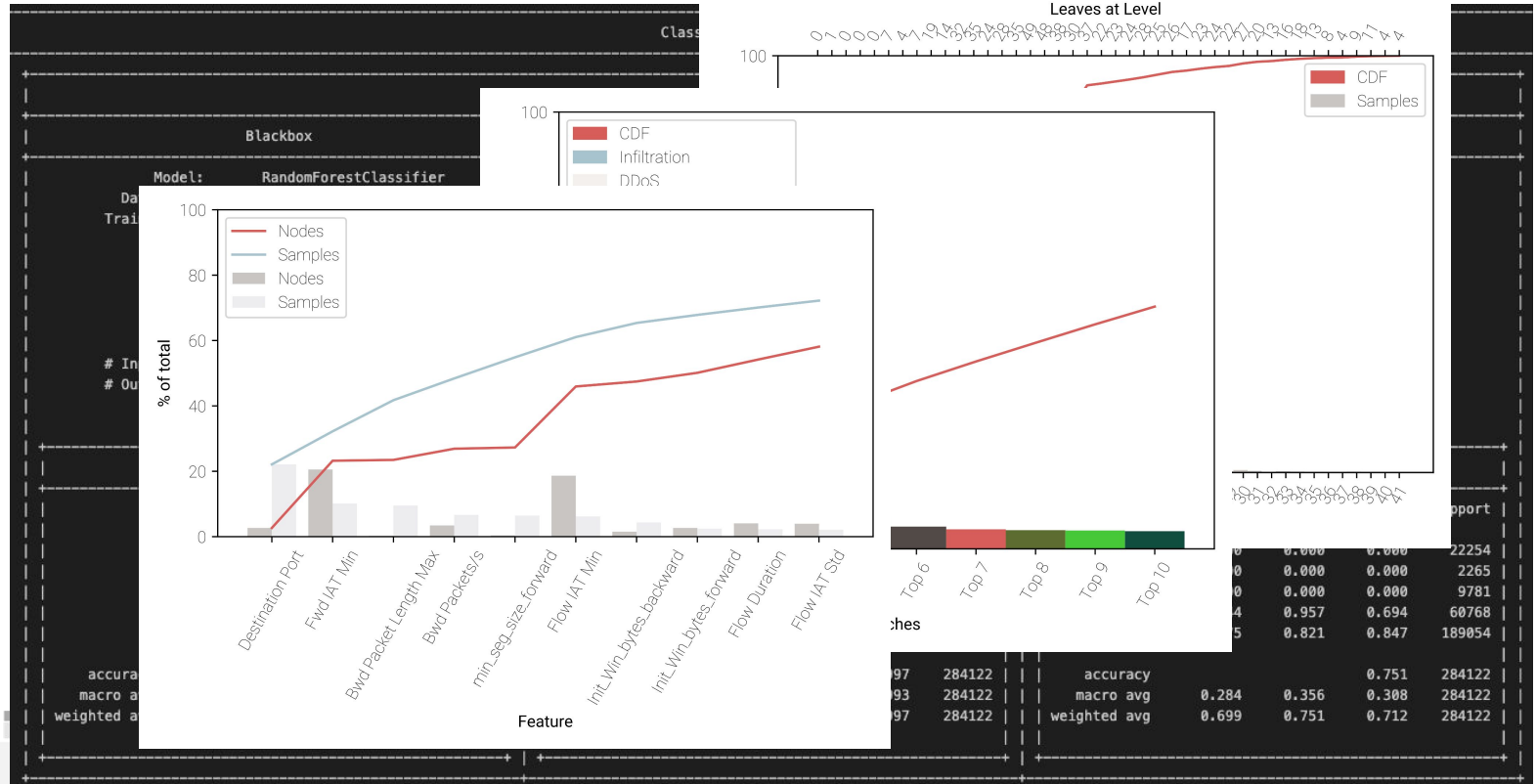


Dataset
Black-box Model
trustee
DT with Highest Agreement

Augmented
AI/ML Development Pipeline



Generating Trust Reports

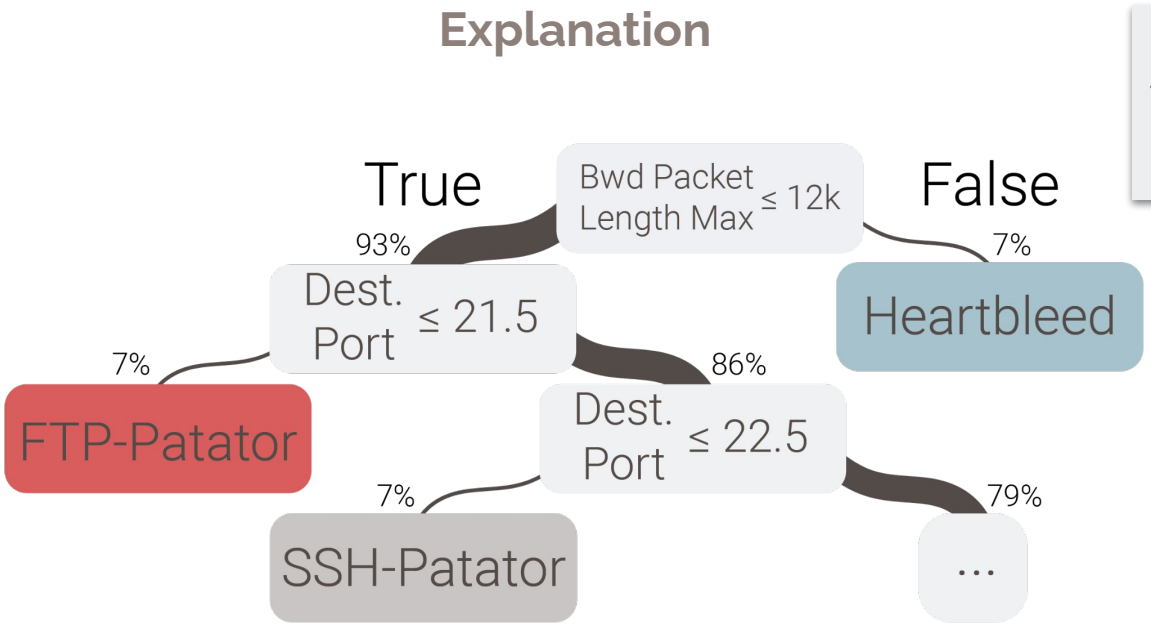


Detecting Heartbleed Traffic

Problem Setup

- Selected publications:
  - *Many papers that rely on the CIC-IDS-2017 dataset*
  - *"Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization" — Sharafaldin et al., 2018*
- Proposal:
  - **Model**: Random Forest to classify traffic between benign traffic and 13 different attacks (e.g. PortScan, DDoS, Heartbleed)
  - **Features**: 78 pre-computed features, from flow statistics (e.g. flow duration, mean IAT)
  - **Dataset**: CIC-IDS-2017 [https://www.unb.ca/cic/datasets/ids-2017.html]
- Results:
  - Reported F1-score: 0.99
  - Reproduced F1-score: 0.99

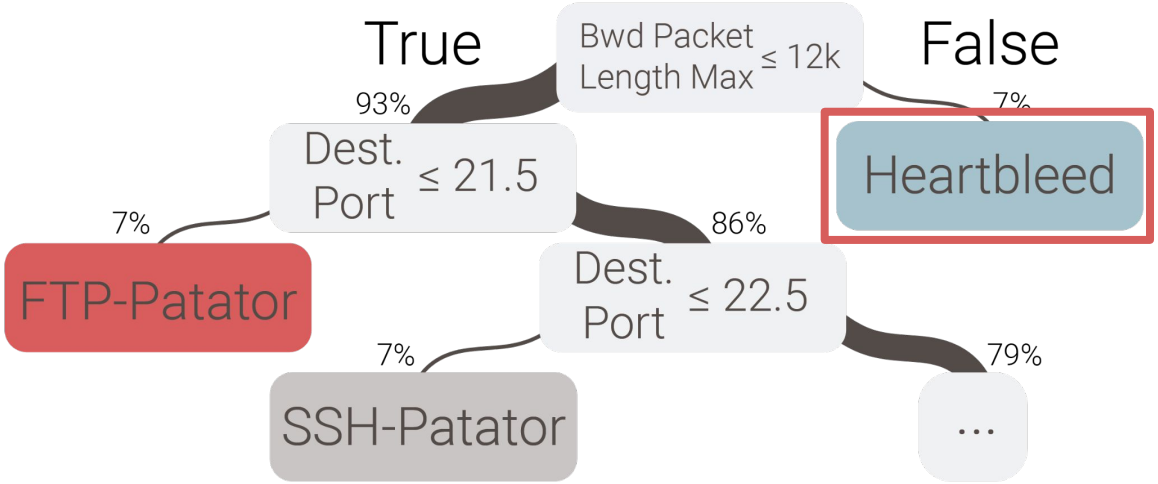
Detecting Heartbleed Traffic



Fidelity: 0.99
Top-3 pruning
6 nodes

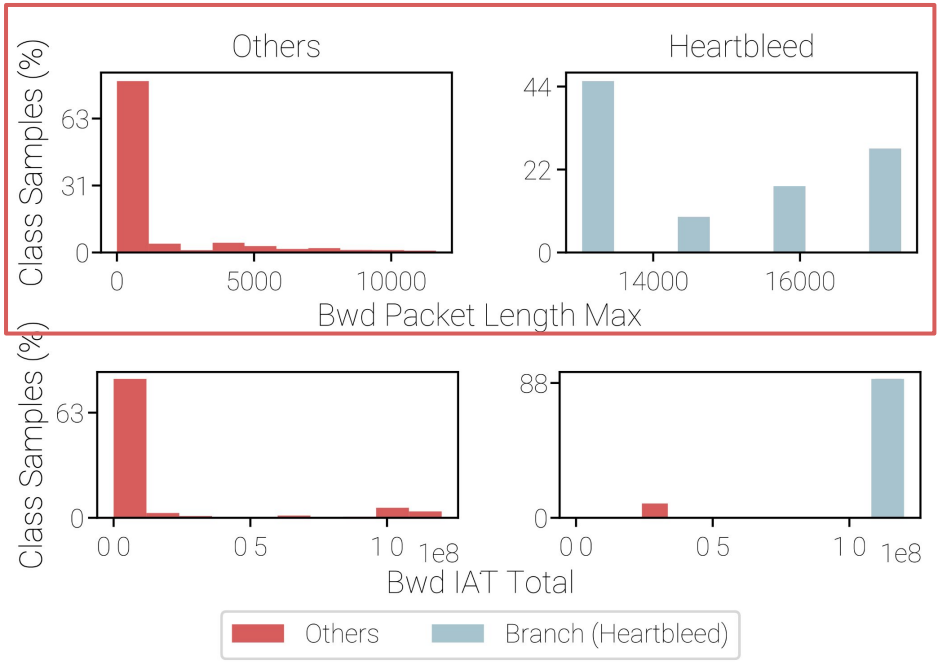
Detecting Heartbleed Traffic

Explanation



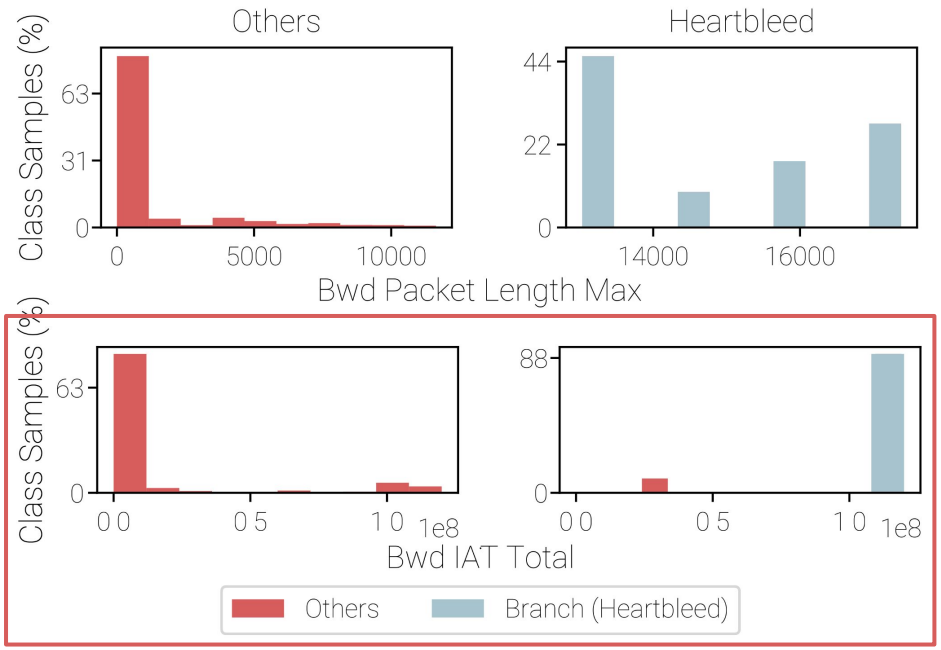
Detecting Heartbleed Traffic

Explanation



Detecting Heartbleed Traffic

Explanation



Detecting Heartbleed Traffic

- Heartbleed attack:
  - An attacker sends an HTTPS heartbeat message with a value in the size field bigger than the message
    - e.g., 1k bytes packet with 16k bytes size value
  - A vulnerable server responds with a message with the size equal to the value specified in the size field and reveals information stored locally in its memory
    - e.g. server returns 16k bytes (1k from the packet and 15k from the server's memory)

- In the CIC-IDS-2017 dataset:
  - HTTPS connection was never closed during the duration of the attack
    - Huge number of backward bytes and very high IAT in the flow!

Detecting Heartbleed Traffic

Validation

- Validation dataset:
  - 1000 new heartbleed flows closing connection after every heartbeat
  - Backward bytes and IAT similar to benign traffic

Class	Precision	Recall	F1
Heartbleed (i.i.d.)	1.000	1.000	1.000
Heartbleed (o.o.d)	0.000	0.000	0.000

Detecting Heartbleed Traffic

Validation

- Validation dataset:
  - 1000 new heartbleed flows closing connection after every heartbeat
  - Backward bytes and IAT similar to benign traffic

Class	Precision	Recall	F1
Heartbleed (i.i.d.)	1.000	1.000	1.000
Heartbleed (o.o.d)	0.000	0.000	0.000

Takeaway: the model is overfitted to training data and fails to identify o.o.d. samples!

Other Use Cases

Problem	Model(s)	Dataset(s)	Trustee Fidelity	Inductive Bias
Detect VPN traffic (Wang *et al.*, ISI'17)	1-D CNN	ISCX VPN-nonVPN	1.00	Shortcut learning
Detect Heartbleed traffic (Sharafaldin *et al.*, ICISSP'18)	RFC	CIC-IDS-2017	0.99	O.O.D.
Detect Malicious traffic (IDS) (Holland *et al.*, CCS'21)	nPrintML	CIC-IDS-2017	0.99	Spurious Correlation
Anomaly Detection (Mirsky *et al.*, NDSS'18)	Kitsune	Mirai dataset	0.99	O.O.D
OS Fingerprinting (Holland *et al.*, CCS'21)	nPrintML	CIC-IDS-2017	0.99	O.O.D
IoT Device Fingerprinting (Xiong *et al.*, HotNets'19)	Iisy	UNSW-IoT	0.99	Shortcut learning
Adaptive Bit-rate (Mao *et al.*, SIGCOMM'17)	Pensieve	HSDPA Norway	0.99	O.O.D

AI/ML for Network Security: The Emperor has no Clothes. ACM CCS 2022.

Trustee Python package

Search projects

Help Sponsors Log in Register

trustee 1.1.1

Latest version

```
pip install trustee
```

Released: Aug 28, 2022

This package implements the Trustee framework to extract decision tree explanation from black-box ML models.

Navigation

- Project description
- Release history
- Download files

Project links

- Homepage
- Repository

Statistics

GitHub statistics:

- Stars: 2
- Forks: 0
- Open issues/PRs: 0

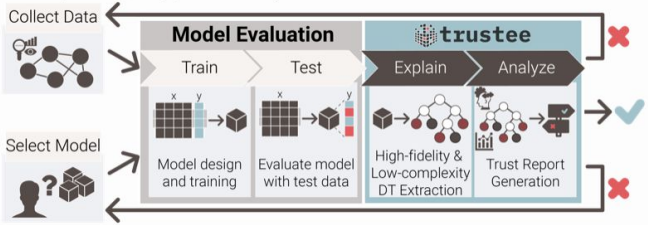
View statistics for this project via

Project description

trustee

This package implements the `trustee` framework to extract decision tree explanation from black-box ML models. For more information, please visit the documentation website.

Standard AI/ML development pipeline extended by Trustee.



Downloads 73 Weekly

Downloads 270 Monthly

Downloads 6k Total

Trustee 1.1.1 documentation

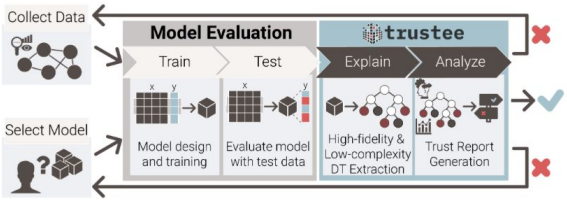
Search

- API
- Examples
  - TrustReport for Classification
  - TrustReport for Regression
  - RegressionTrustee
  - ClassificationTrustee

Github Repo Use Cases Tech Report

Overview

Trustee is a framework to extract decision tree explanation from black-box ML models.



Standard AI/ML development pipeline extended by Trustee.

Getting Started

This section contains basic information and instructions to get started with Trustee.

Python Version

Trustee supports Python >=3.7.

Install Trustee

Use the following command to install Trustee:

```
$ pip install trustee
```

Sample Code

```
from sklearn import datasets
```

Conclusions

1. ML in high-stakes requires trust
2. Trustee improves trust!
3. Trustee is ready to be used!
   - Just download our Python package

Thank you!

trustee

https://trusteeml.github.io

Trustee Python package

- https://pypi.org/project/trustee/

Trustee Repository

- https://github.com/TrusteeML/trustee

Use Cases Repository

- https://github.com/TrusteeML/emperor

.inf INSTITUTO DE INFORMÁTICA UFRGS

NIKSUN

UFMS

Financial Support

CNPq

FAPESP

CAPES

INCT
InterSCity