Challenges in Using ML for Networking Research: How to Label If You Must

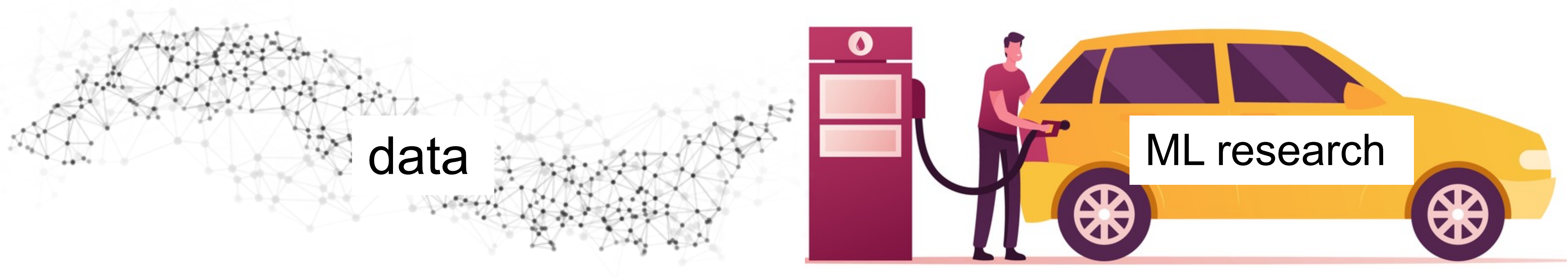
Yukhe Lavinia, University of Oregon

Ramakrishnan Durairajan, University of Oregon

Reza Rejaie, University of Oregon

Walter Willinger, NIKSUN, Inc.

Introduction



Fuel for Machine Learning (ML) research:

labeled data

Outline

Challenges

Contributions

Building blocks

Evaluation

Conclusion

Future work

Challenges in using ML in networking

Challenge 1: Lack of labeled networking data

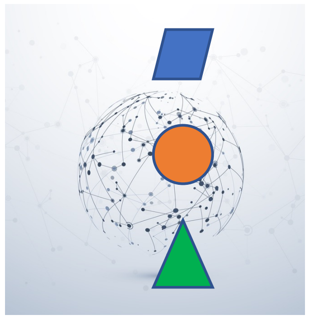
Difficulty in labeling at scale

Lack of agreement in community

Features of good data?

Features of bad data?

Networking Data



Label

?
?
?

High human cost of labeling

Limited number of experts

Large amount of data

Challenge 2: Privacy concern in network data

Safest: avoid a possibility of privacy leaks

Sharing raw or labeled data

Sharing learning models

Collaborate using ML in networking

Challenge 3: Hidden biases in data

Inherent in ML, made complicated by the nature of network data

Lack of representation of minority group, creating a model that does not generalize well

Contributions

EMERGE

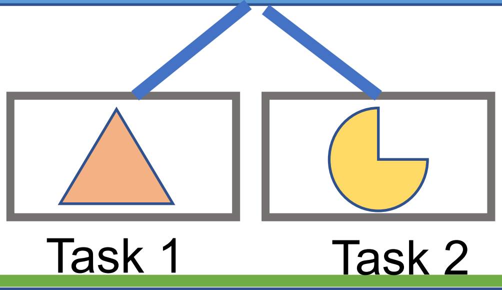
a framework to dEmocratize the use of ML for nEtwoRkinG rEsearch

Challenge

- Lack of labeled networking data
- Privacy concern in network data
- Hidden biases in data

Solution

- Create **high quality** labels **at scale** in a **programmable** fashion and at **low human labor cost**
- Share **only learning algorithms**
- Implement **multi-task learning** (MTL) (Future work)



→ More generalized data representation → Bias reduction

EMERGE

Create high quality networking data labels:

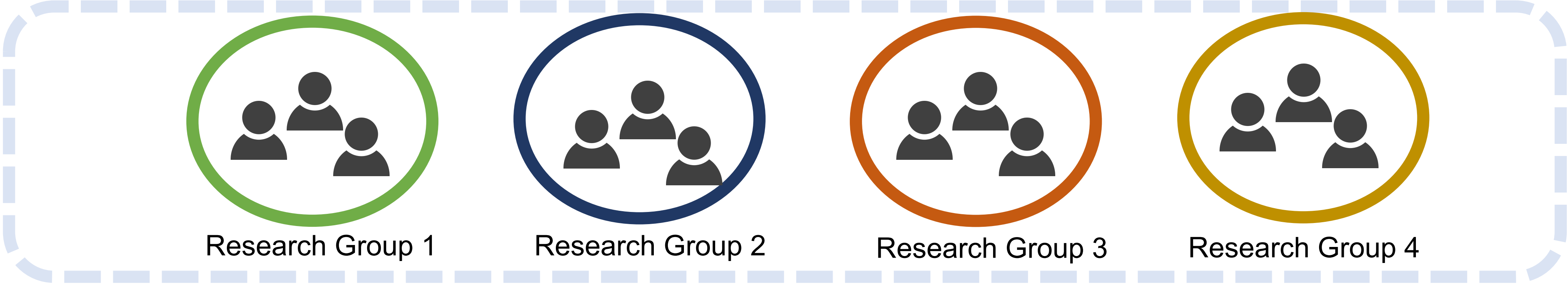
At scale

In programmable fashion

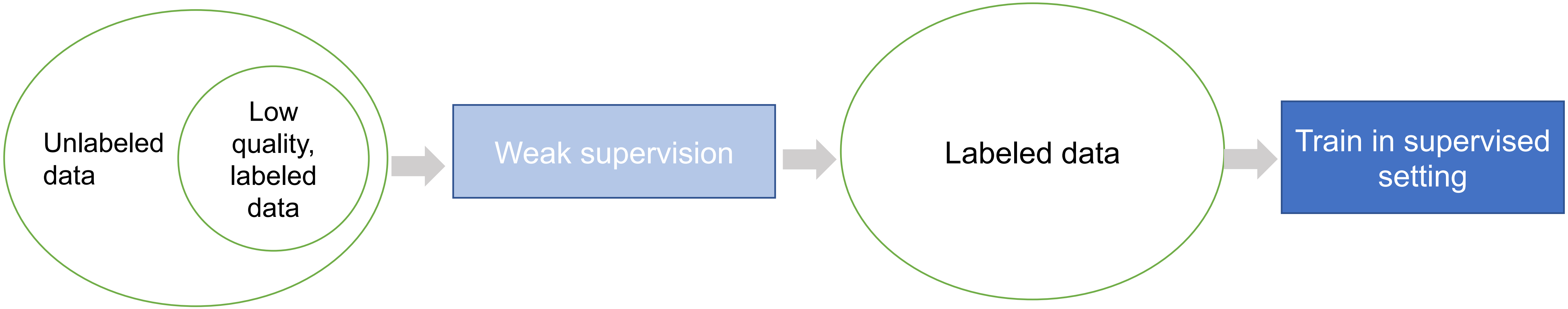
At low human labor cost

Promote:

Privacy-preserving collaboration



Building Blocks



Data Programming[1]

Domain Expert

```python
def free(word):
    if "free" in word:
        return 1
    else:
        return 0

def guarantee(word):
    if "guarantee" in word:
        return 1
    else:
        return 0
```

Labeling Functions

Probabilistic Labels

Data programming framework: Snorkel[2]

Limitations:

Not specific to networking

Scalability issue

Data amount, data diversity

Human labor cost

[1] Ratner et al., "Data programming: Creating large training sets, quickly", Advances in Neural Information Processing Systems (2016) .
[2] Ratner et al., "Snorkel: Rapid training data creation with weak supervision", VLDB Endowment (2017).

Building Blocks

Snuba[1]

Simple ML classifiers (e.g., logistic regressor, decision tree, nearest neighbor)

Probabilistic Labels

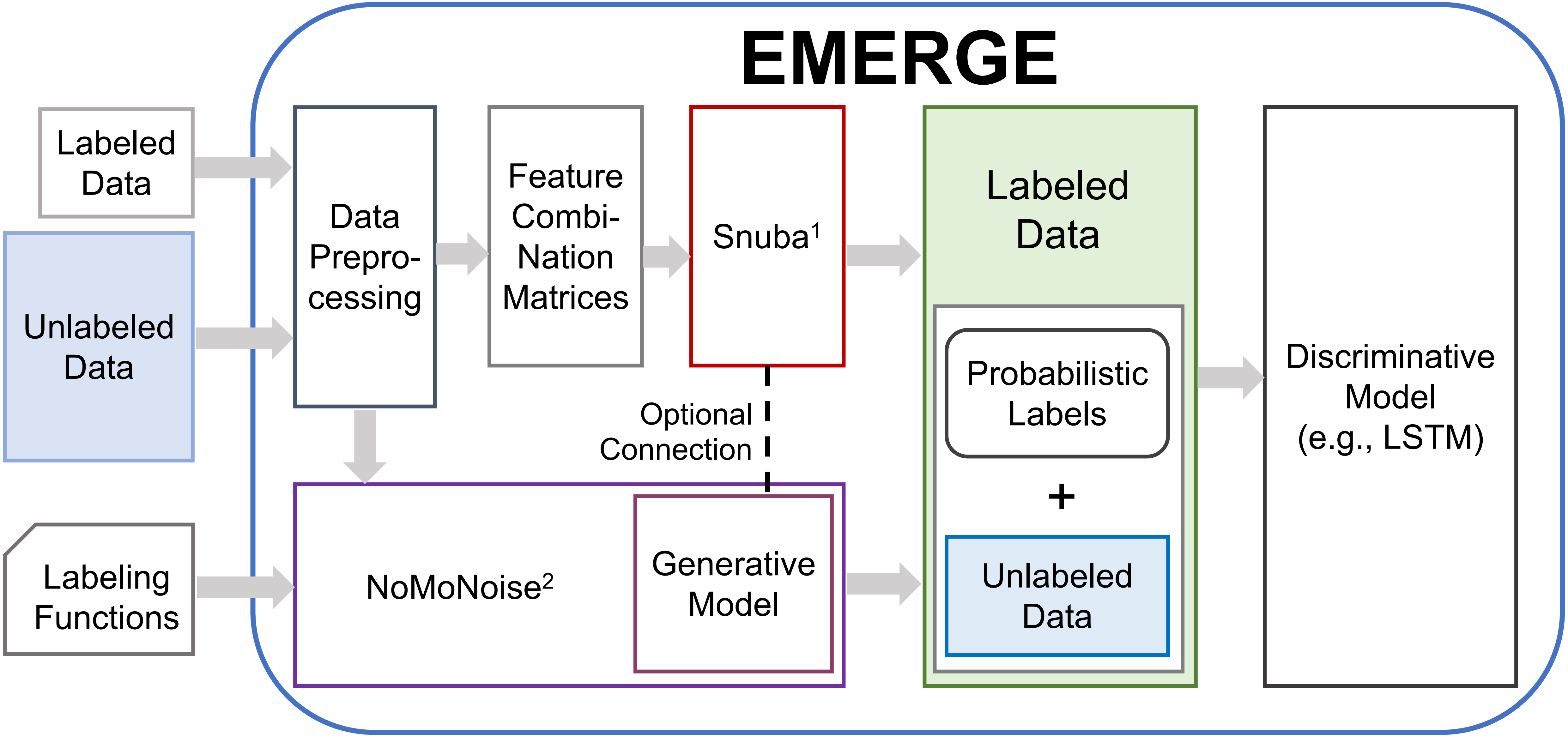
Limitation: Not specific to networking

NoMoNoise[2]

Solve networking problem: remove noise in latency measurements

Limitation: Scalability issue

[1] Varma et al., “Snuba: Automating weak supervision to label training data”, Proc. VLDB Endow 2018.
[2] Muthukumar et al., “Denoising internet delay measurements using weak supervision”, ICMLA 2019.



Goals: Create high quality labels at scale and at low cost   Promote privacy-preserving collaboration

[1] Varma et al., “Snuba: Automating weak supervision to label training data”, Proc. VLDB Endow 2018.
[2] Muthukumar et al., “Denoising internet delay measurements using weak supervision”, ICMLA 2019.

Evaluation

Datasets

Methodology (2 experiments)

Experimental results

Future work

Datasets

CAIDA Ark traceroute data

28 source-destination (SD) pairs

75,359 RTT measurements

Methodology: Experiment 1

Challenge

Lack of labeled networking data

Goal

Demonstrate that EMERGE can create **high quality** labels **at scale** in a **programmable** fashion and at **low human labor cost**

Statistical heuristics, outlier detection heuristic, anomaly detection heuristic

Task: Differentiate good data vs. noise

Naïve

Data preprocessing

```
...
[1]
[-1]
[1]
[1]
[-1]
[-1]
[1]
[1]
[-1]
...
```

labels

LSTM models

F1 score

compare

EMERGE

Data pre-processing

Feature com-bination

```
...
0.999611776506926
0.00038822349307393296
0.999611776506926
0.999611776506926
0.00038822349307393296
0.00038822349307393296
0.999611776506926
0.00038822349307393296
0.999611776506926
0.999611776506926
0.999611776506926
0.999611776506926
0.999611776506926
```

prob. labels

LSTM models

F1 score

Methodology: Experiment 1

Data Preprocessing

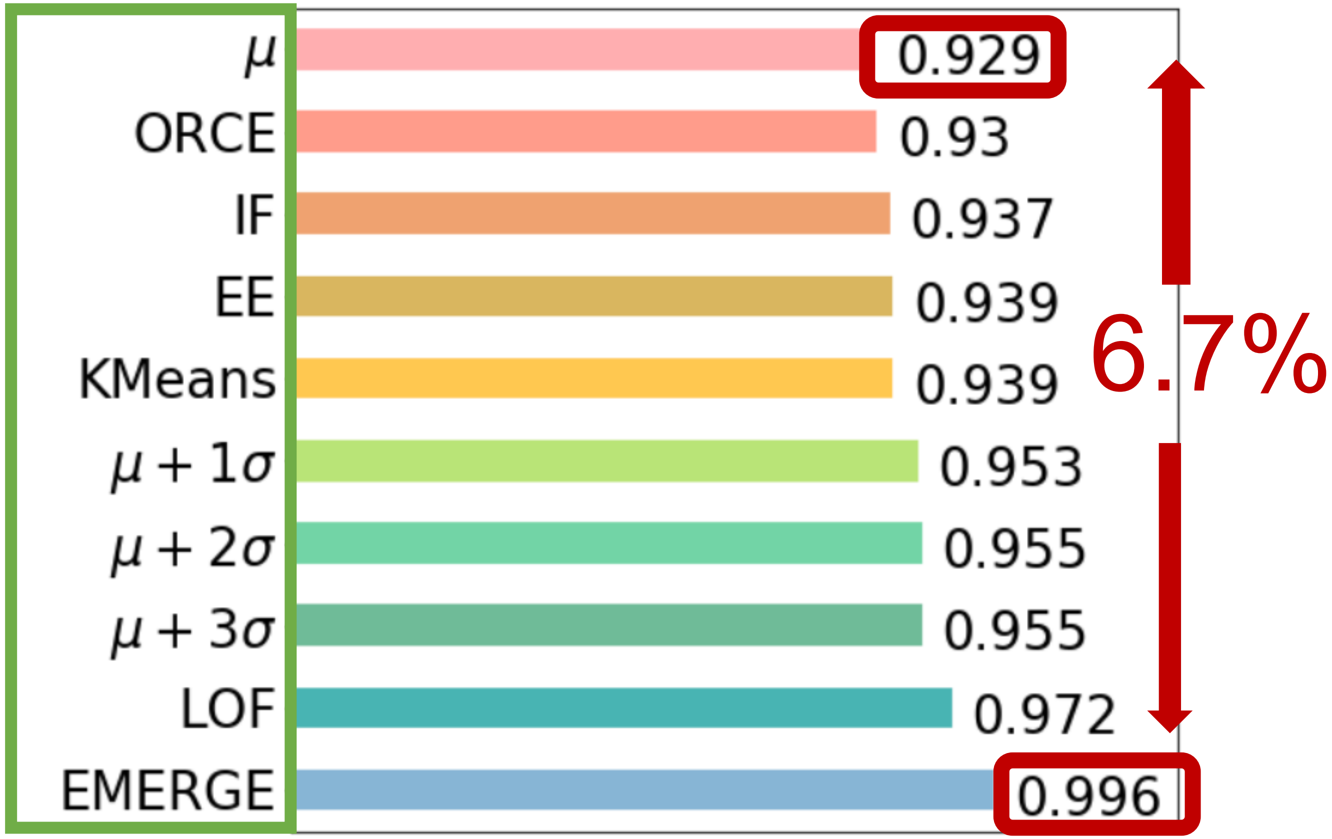
1. Determine threshold
2. Record threshold values for the naïve methods
3. Oversample noise data
4. Divide data into test, validation, training sets
5. Create ground truth labels for validation and test data

Feature Combination

8 statistical features:

- Length
- Mean
- Median
- Variance
- Standard deviation
- Minimum value
- Maximum value
- Sum

Results: Experiment 1



F1 score

Unique characteristics in data

More accurate labels

Goal:

Demonstrate that EMERGE can create **high quality** labels **at scale** in a **programmable** fashion and at **low human labor cost**

Methodology: Experiment 2

Challenge

Privacy concern in network data

Goal

Demonstrate that EMERGE supports privacy-preserving collaboration to advance ML and networking research by sharing **only learning algorithms**

Task: Show how researchers from different groups can use EMERGE to collaborate

LFs: μ
EE
LOF
μ + 2𝝈

Combine LFs

prob. labels

LSTM models

F1 score

improve label quality ?

Results: Experiment 2

1 LF

μ	EE	LOF	$\mu + 2\sigma$
0.637	0.759	0.836	0.721

2 LFs

μ & EE	μ & LOF	EE & LOF	$\mu + 2\sigma$ & EE	$\mu + 2\sigma$ & LOF
0.720	0.821	0.890	0.826	0.810

3+ LFs

μ & EE & LOF	$\mu + 2\sigma$ & EE & LOF	μ & $\mu + 2\sigma$ & EE & LOF
0.733	0.838	0.914

↑ Number of LFs

↑ Learning opportunity

↑ Label quality

Goal: Demonstrate that EMERGE supports privacy-preserving collaboration to advance ML and networking research by sharing **only learning algorithms**

Hyperparameter Setup

Different datasets can have different hyperparameter values

Hyperparameter	Values
Batch size	16, 32, 64, 128, or 256
Learning rate	Between 1e-5 and 1e-2
Number of epochs	5, 10, 20, 25, or 30
Number of LSTM units	32, 64, or 128
L2 regularization	Between 0.0 and 0.6
Dropout	0.0, 0.2, or 0.4

Conclusion

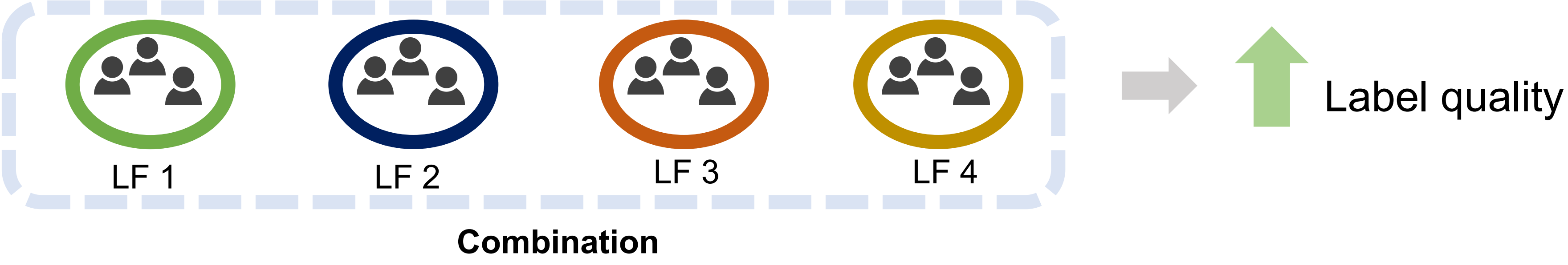
Proposed solutions to address the lack of labeled network data, the privacy concern in network data, and the hidden biases in data

Demonstrated:

- Create high quality labels at scale, and at low human labor cost

EMERGE F1 score > Naïve F1 scores

- Promote privacy-preserving collaboration that advances ML and networking research



Proposed multi-task learning to reduce bias

Future Work

Address hidden bias in data using Multi-Task Learning (MTL)

Use other networking data types to assess the versatility of EMERGE

Use different events of interest for EMERGE to detect

Thank you!

Code available at https://gitlab.com/onrg/emerge

We thank NSF for funding this project