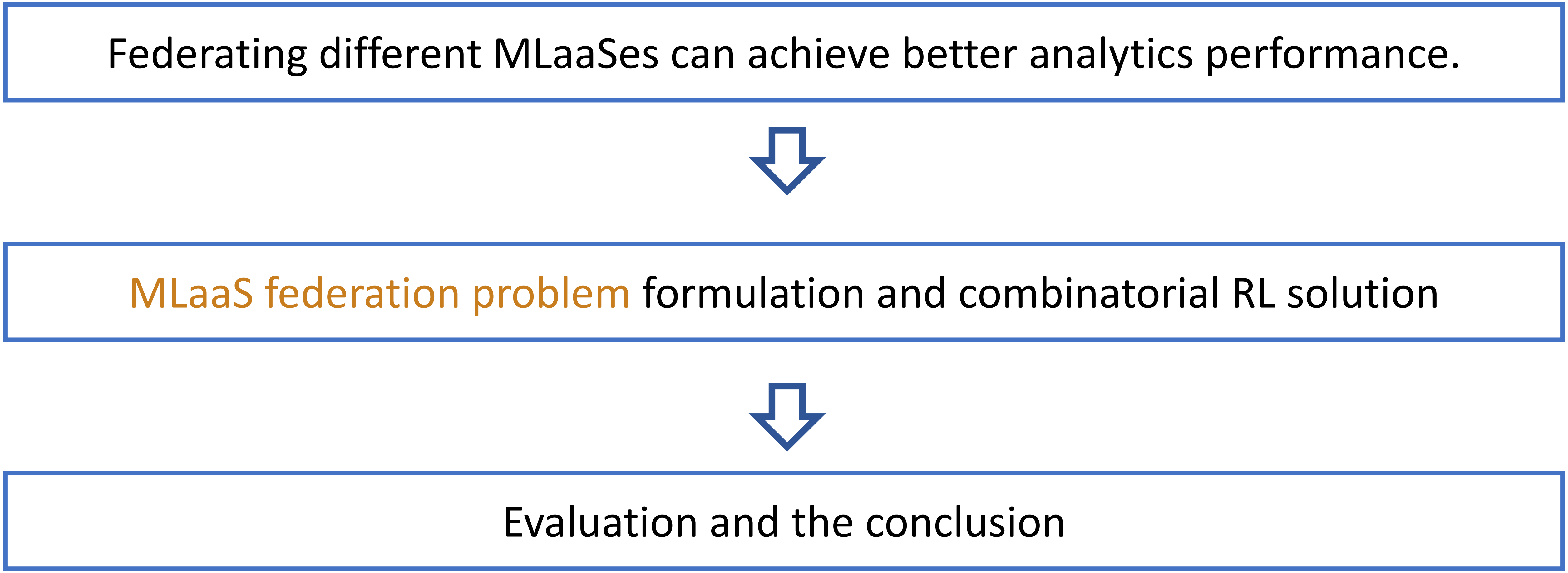
Cost Effective MLaaS Federation: A Combinatorial Reinforcement Learning Approach

Shuzhao Xie[🌲], Yuan Xue[🌲], Yifei Zhu[🌴], Zhi Wang[🌲]

[🌲] 清華大學 Tsinghua University

[🌴] 上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Overview



Federating different MLaaSes can achieve better analytics performance.

MLaaS federation problem formulation and combinatorial RL solution

Evaluation and the conclusion

Machine Learning as a Service (MLaaS)

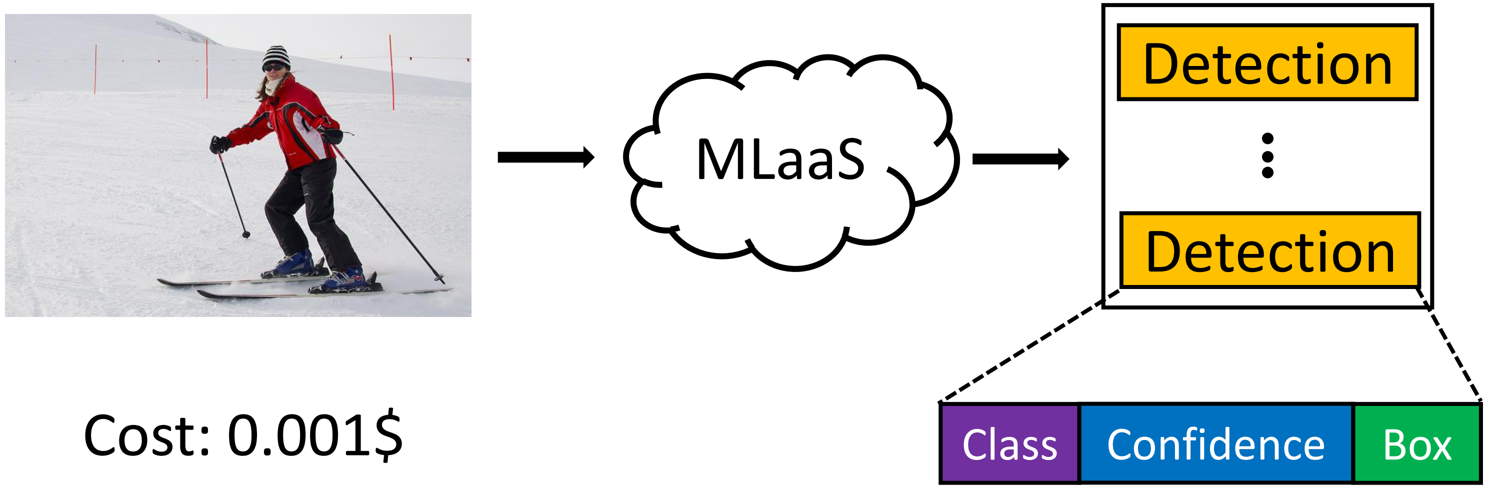
Major Providers

aws Azure

Niche Providers

bigml® ALGORITHMIA

Example: object detection service



Cost: 0.001$

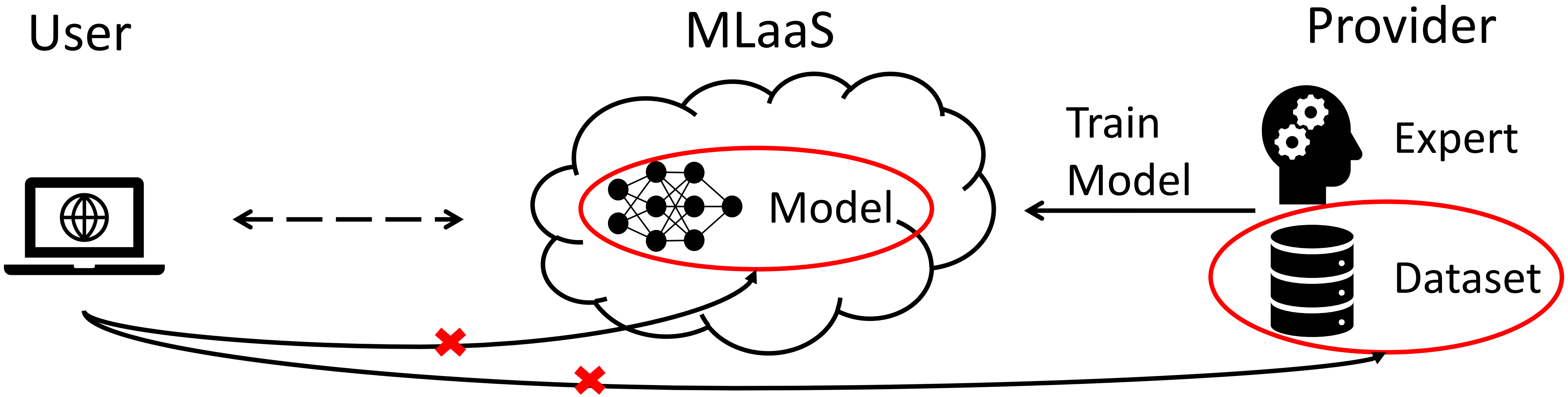
Usage

- Security
- Agriculture
- Online shopping

Strengths

- Well-defined interfaces
- Free maintenance burden
- Accessed from any where, at any time

MLaaS is a black box



Confidential Untouchable

Which MLaaS is the best?

Pervious measurements on MLaaS

Type 1. White box

MLCommons

User-known models

- Accuracy
- Latency
- Quality

Type 2. Training Platform

AWS SageMaker

AI experts needed

- User control
- Complexity
- Accuracy

Type 3. Out-of-date MLaaS

Azure Machine Learning

Machine learning models

- SVM
- Neural networks
- Decision tree

Black box

Inference service

Deep learning models

Pervious measurements on MLaaS

Type 1. White box

MLCommons

User-known models

- Accuracy
- Latency
- Quality

Type 2. Training Platform

AWS SageMaker

AI experts needed

- User control
- Complexity
- Accuracy

Type 3. Out-of-date MLaaS

Azure Machine Learning

Machine learning models

- SVM
- Neural networks
- Decision tree

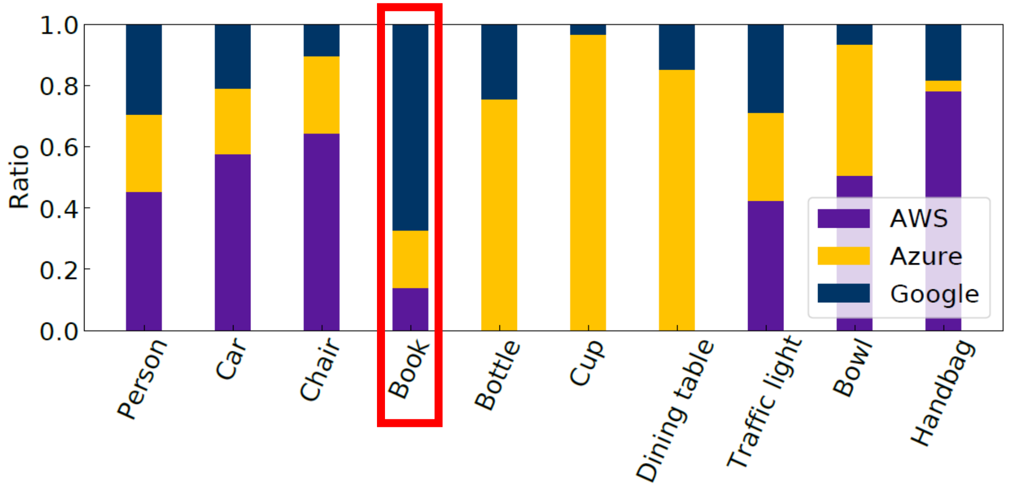
MLaaS in our work:

1. Black box, 2. Inference service, 3. Deep learning models

Which MLaaS is the best?

Provider	mAP	AP@50	AP@75
AWS	**18.81**	**28.88**	**20.84**
Azure	15.10	24.38	16.14
GCP	16.23	23.03	18.12

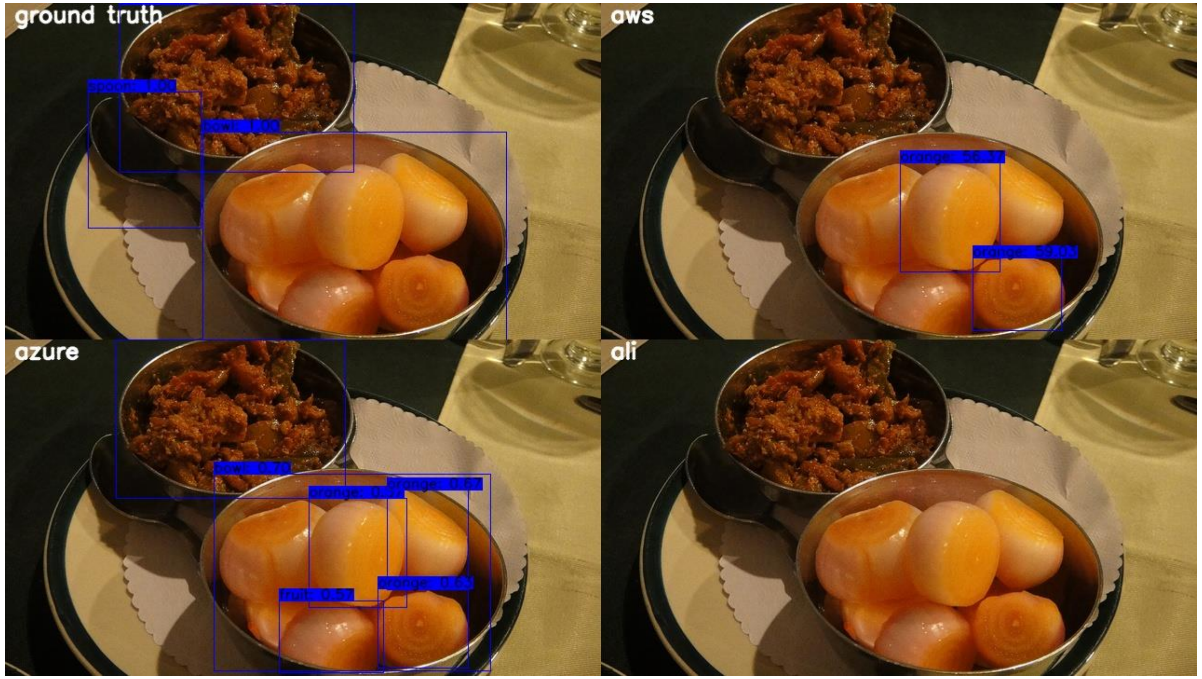
AWS is the best on average

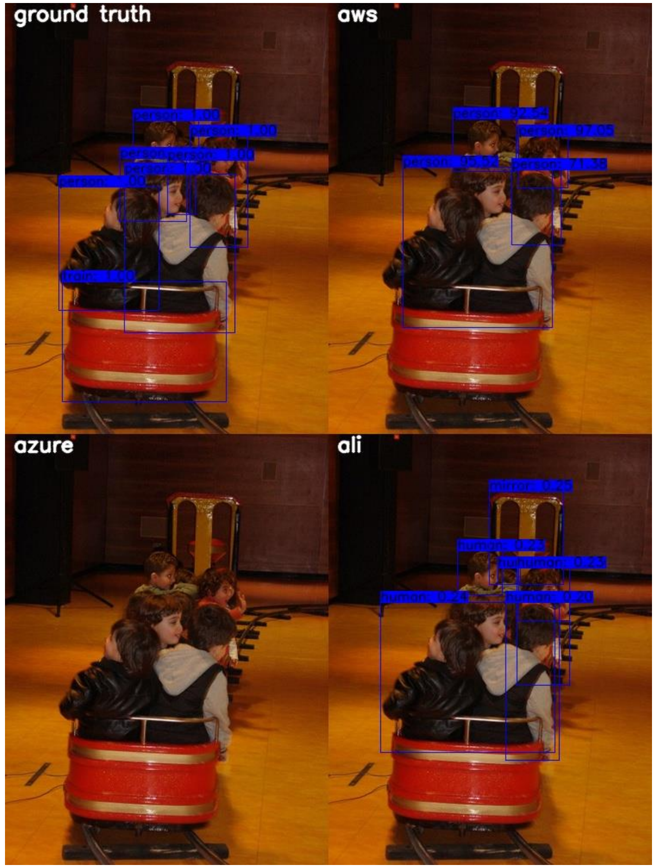


Google is the best for “Book”

Observation 1: For input with different features, the most appropriate MLaaS provider differs.

Which MLaaS is the best?





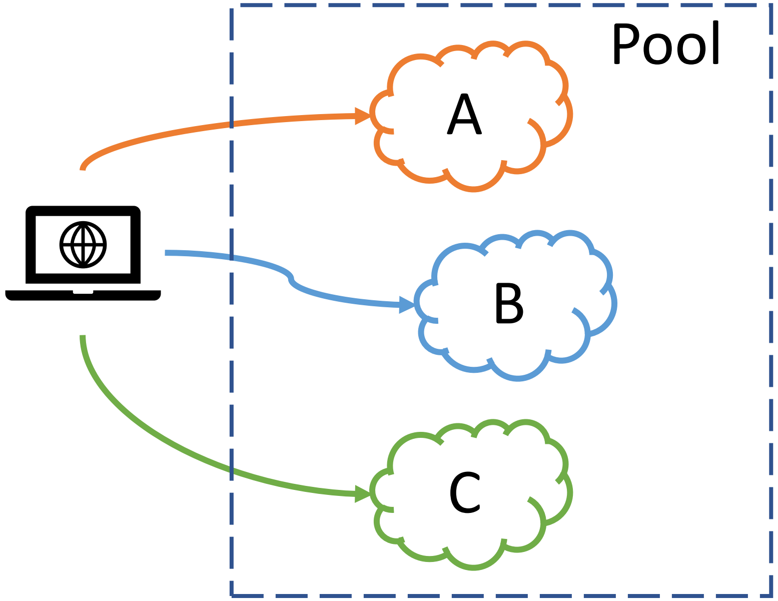
Observation 1: For input with different features, the most appropriate MLaaS provider differs.

Cloud federation

Previous cloud federation

System level metrics

- Latency
- Cost
- Scalability
- Stability



Distribute workloads to clouds from different providers

MLaaS federation

System level metrics

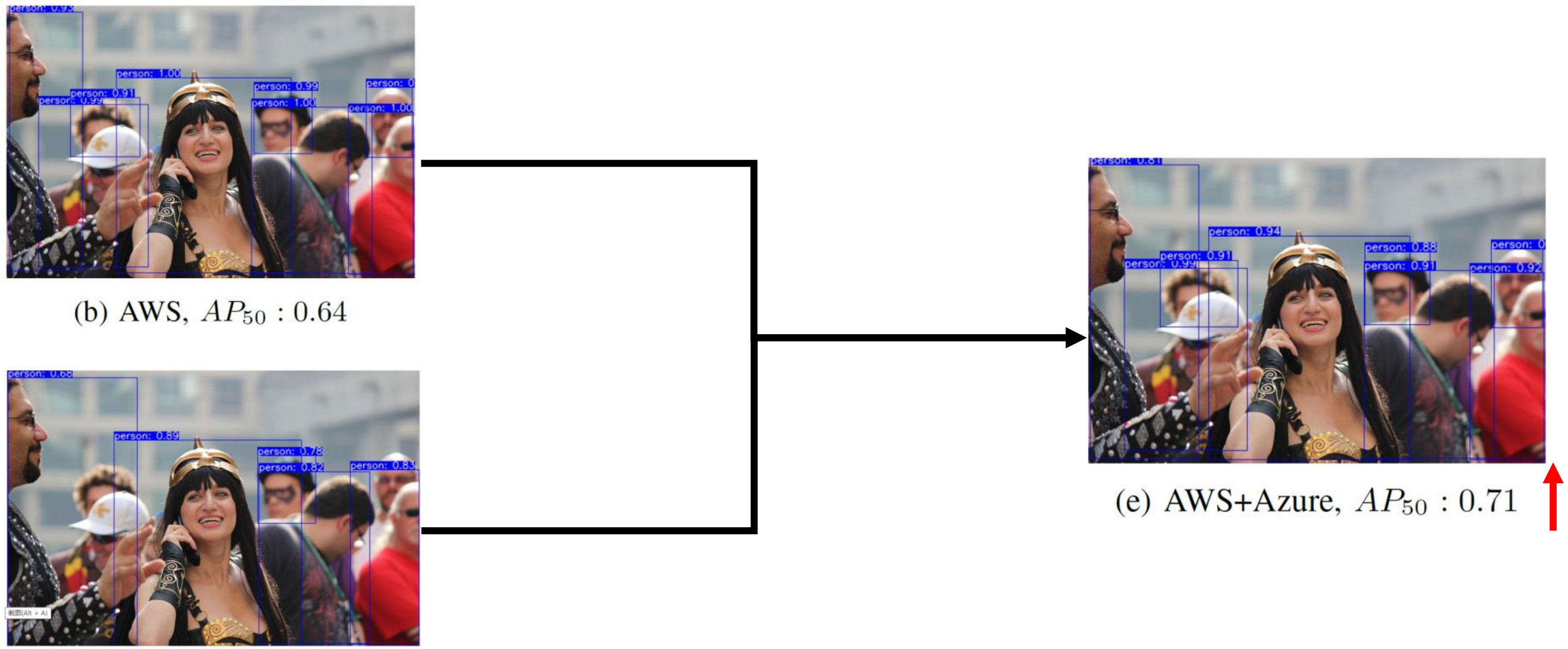
- Latency
- Cost
- …

Model level metric

- Accuracy

How about the performance of MLaaS federation?

The more MLaaSes, the higher accuracy?



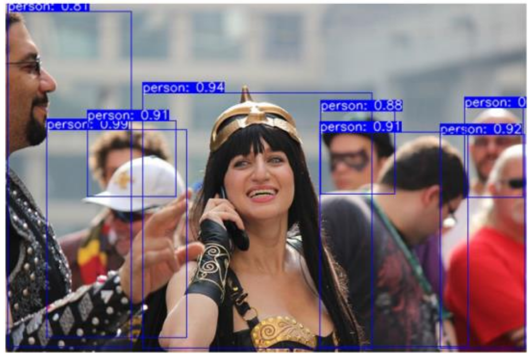
(b) AWS, AP_{50} : 0.64

(c) Azure, AP_{50} : 0.56

(e) AWS+Azure, AP_{50} : 0.71 ↑

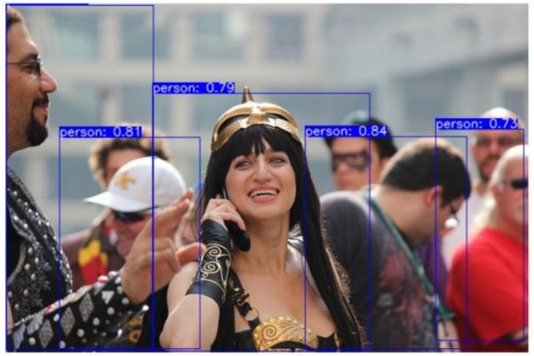
Observation 2: Federate MLaaSes can achieve higher accuracy.

The more MLaaSes, the higher accuracy?



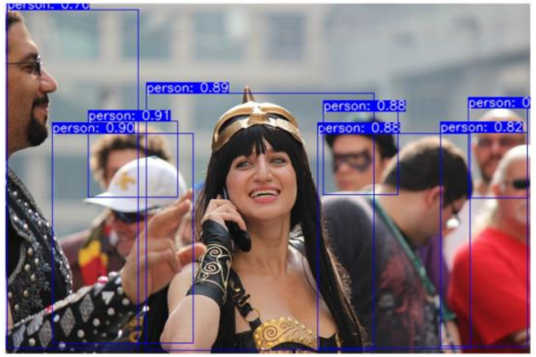
(e) AWS+Azure, AP_{50} : 0.71

+



(d) Google, AP_{50} : 0.56

=

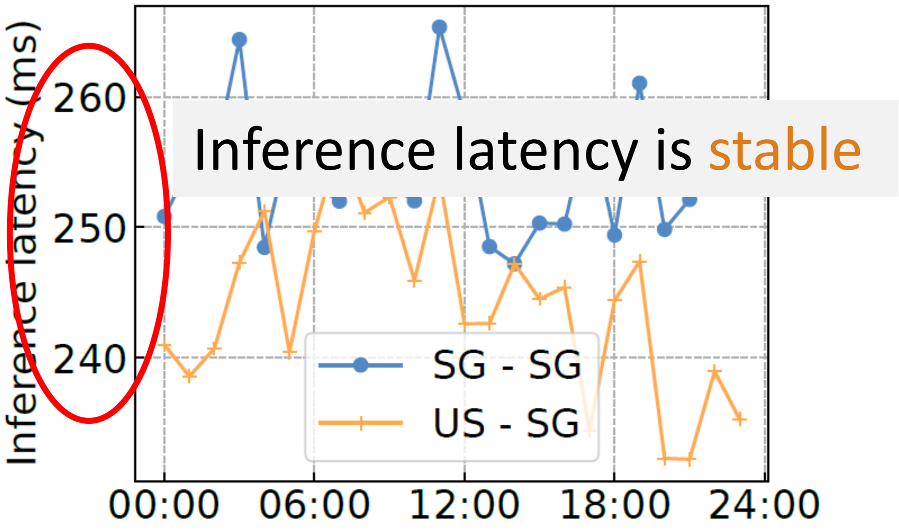


(h) Three providers, AP_{50} : 0.68 ↓

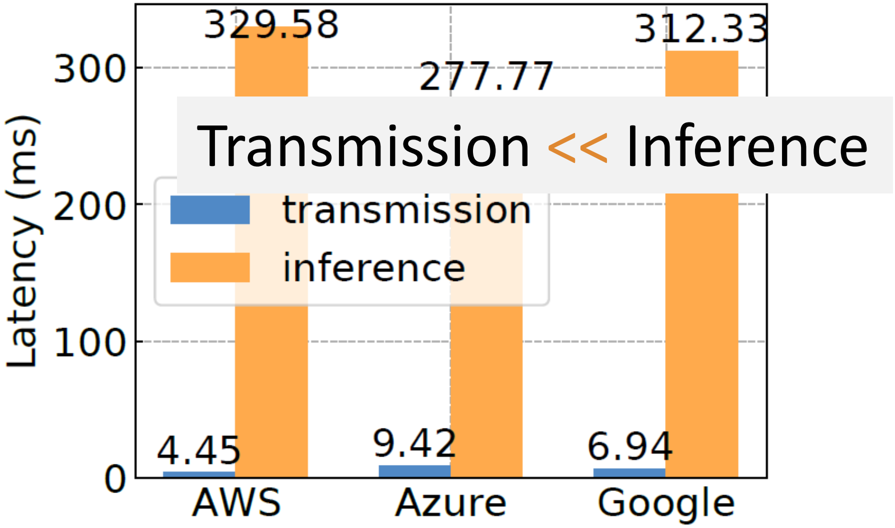
Observation 3: More MLaaSes (costs) do not imply higher accuracy.

Latency

Latency = Transmission latency + Inference latency



Inference latency is stable



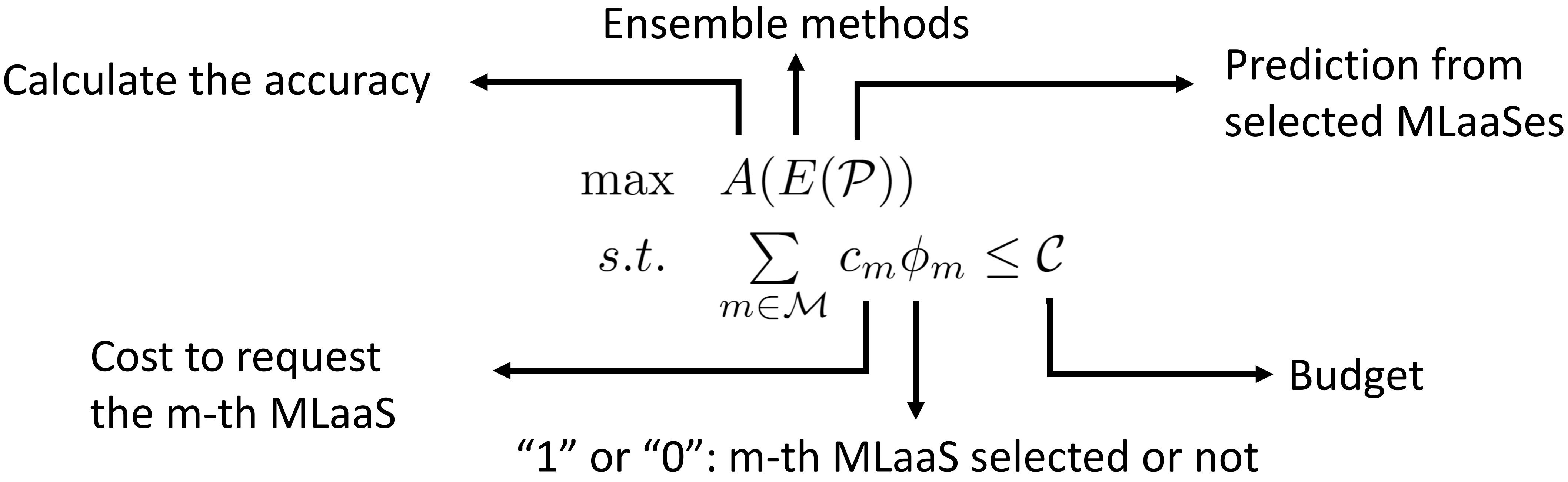
Transmission << Inference

Observation 4: Requesting multiple cloud services does not cause a significant increase in latency with efficient bandwidth.

Cost-effective MLaaS federation

For each input, how to adaptively select k MLaaSes from n available MLaaSes to achieve the highest accuracy while minimize the cost?

Formulation



Ensemble methods

Calculate the accuracy

Prediction from selected MLaaSes

$$\max \quad A(E(\mathcal{P}))$$

$$s.t. \quad \sum_{m \in \mathcal{M}} c_m \phi_m \leq \mathcal{C}$$

Cost to request the m-th MLaaS

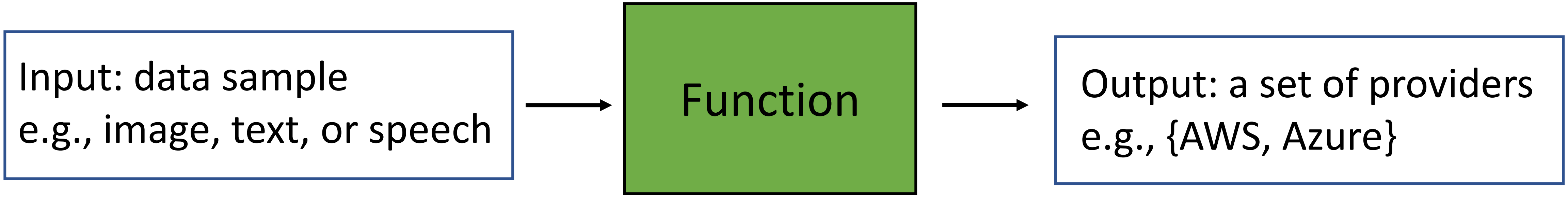
“1” or “0”: m-th MLaaS selected or not

Budget

N-power ($N \geq 2$) object binary knapsack problem

MLaaS federation problem is NP-Hard.

Supervised Learning (×)



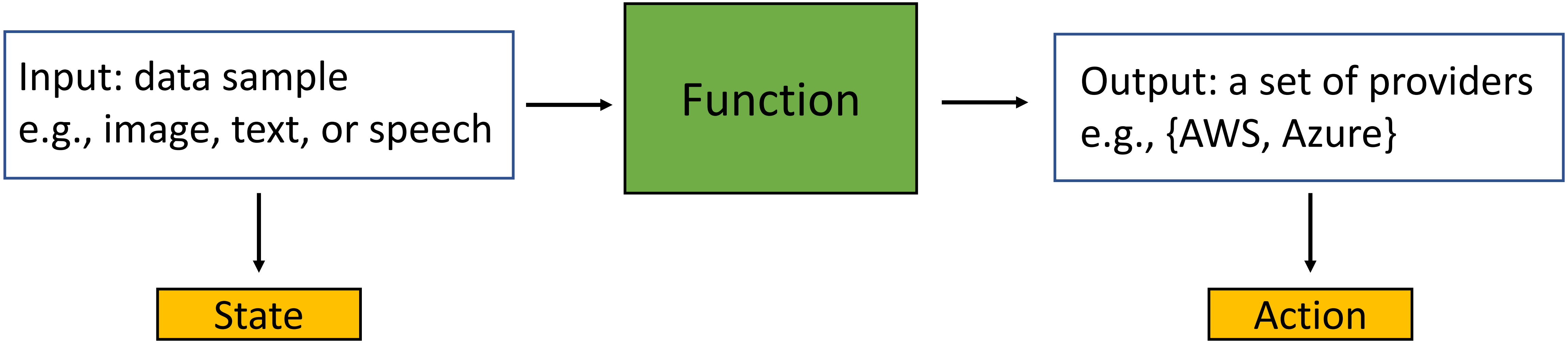
Input: data sample
e.g., image, text, or speech

Function

Output: a set of providers
e.g., {AWS, Azure}

Complexity to generate the training set is exponential to the number of available providers. ($n \sim O(2^n)$)

Reinforcement Learning (√)



Input: data sample
e.g., image, text, or speech

Function

Output: a set of providers
e.g., {AWS, Azure}

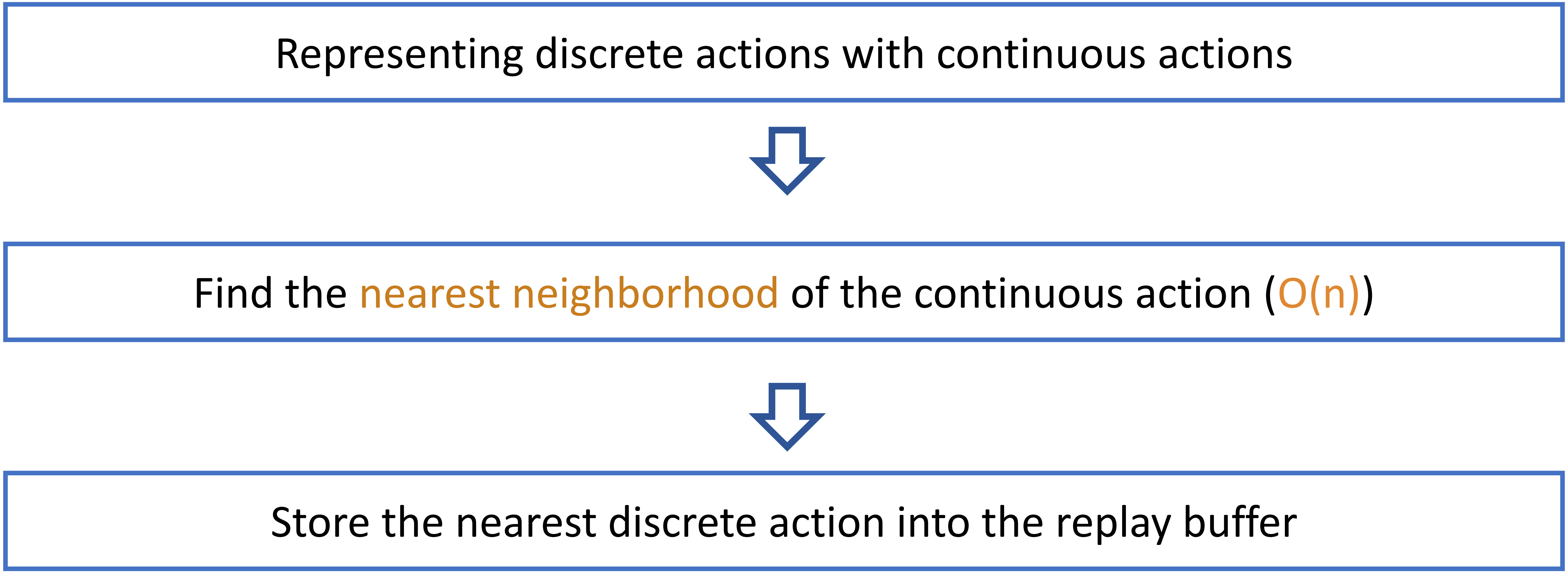
State

Action

If a set has “n” elements, then the number of proper subsets of the given subset is given by 2^n-1.

How to handle $2^n - 1$ discrete actions?

How to handle combinatorial action space?

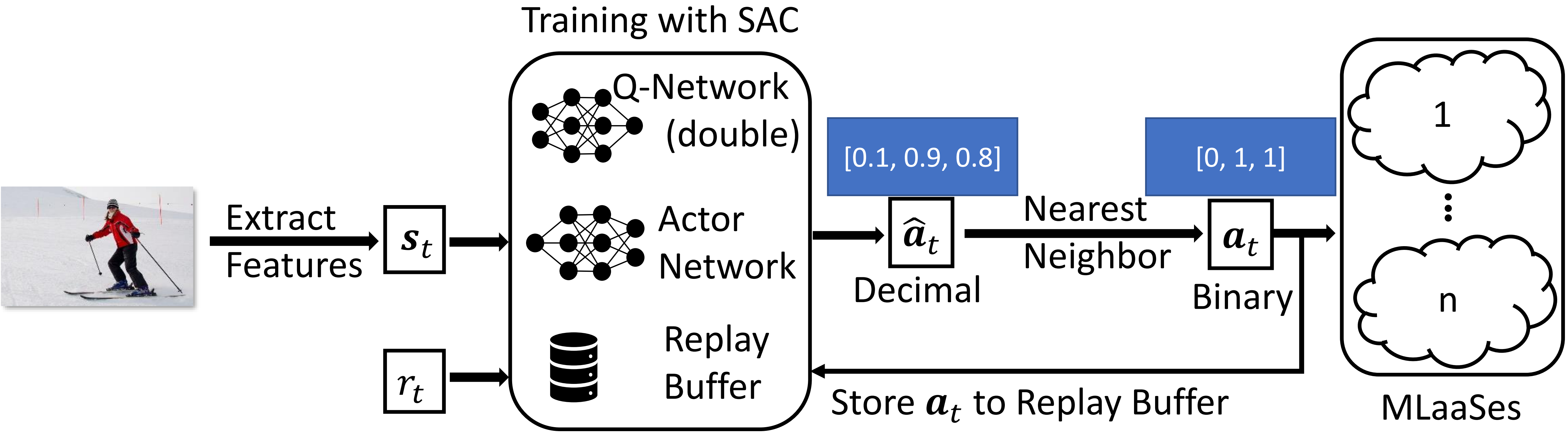


Representing discrete actions with continuous actions

Find the nearest neighborhood of the continuous action (O(n))

Store the nearest discrete action into the replay buffer

Combinatorial RL-based provider selection



Training with SAC

Extract Features

s_t

r_t

Q-Network (double)

Actor Network

Replay Buffer

[0.1, 0.9, 0.8]

$\hat{a}_t$

Decimal

Nearest Neighbor

[0, 1, 1]

a_t

Binary

Store a_t to Replay Buffer

1

⋮

n

MLaaSes

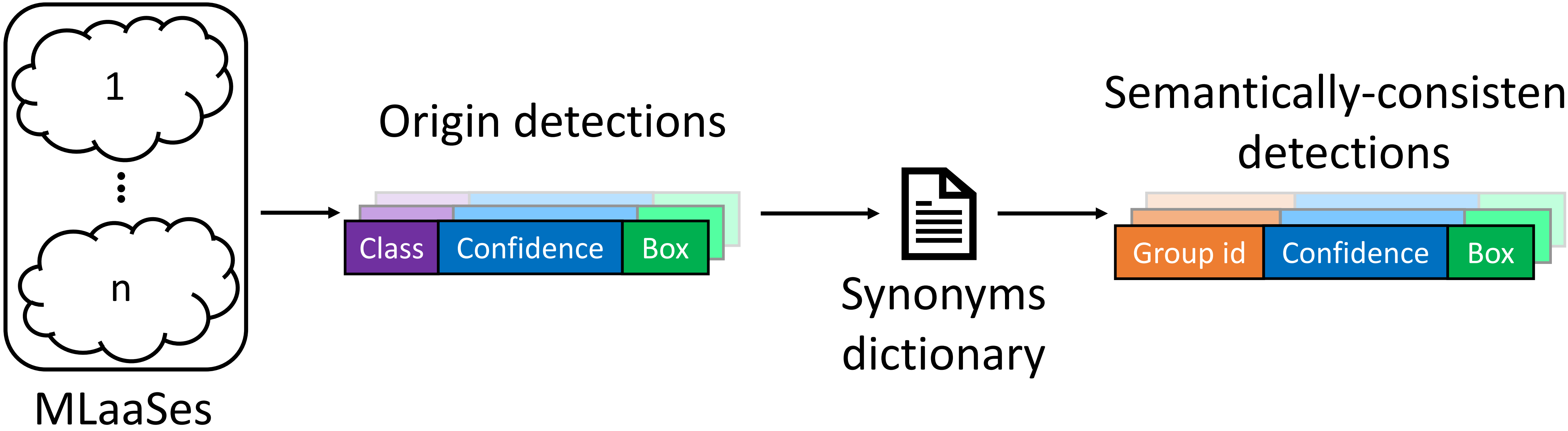
How to aggregate the predictions from multiple MLaaSes?

Group synonym labels into same category

Cloud A → "Motorbike"

Cloud B → "Motorcycle"

→ Same meaning



1

⋮

n

MLaaSes

Origin detections

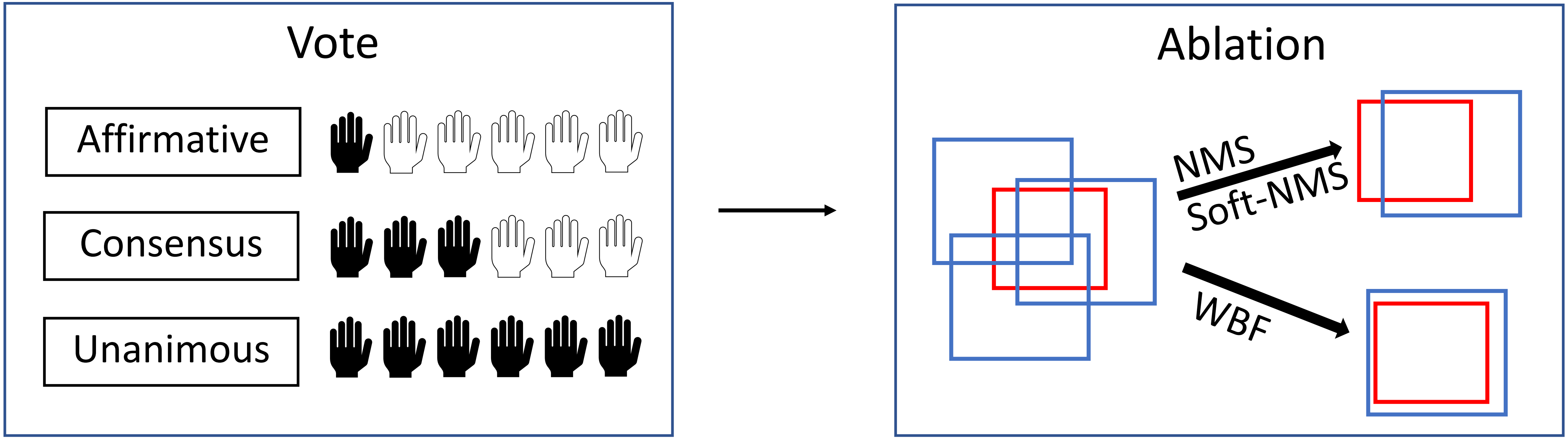
Class	Confidence	Box

Synonyms dictionary

Semantically-consistent detections

Group id	Confidence	Box

Ensemble predictions



Vote

Affirmative

Consensus

Unanimous

Ablation

NMS

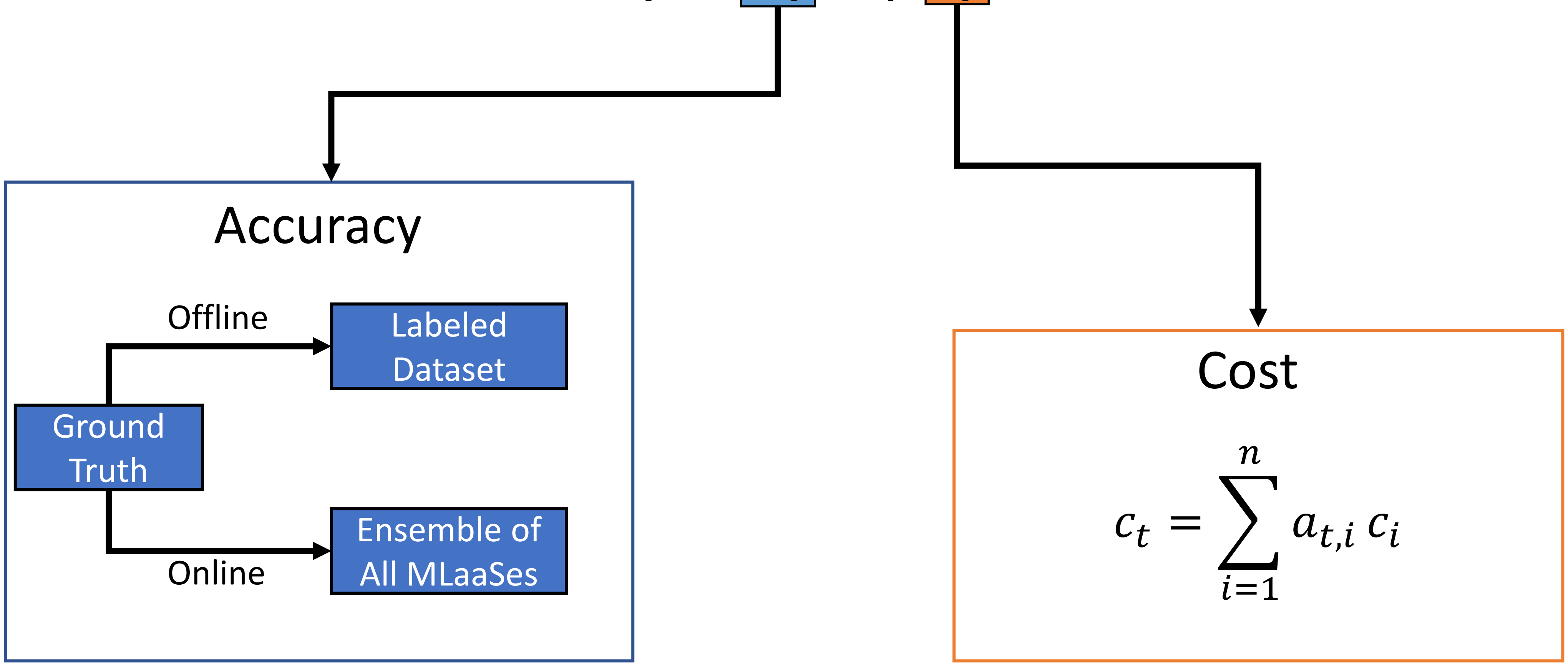
Soft-NMS

WBF

We choose "Affirmative" and "WBF" strategies.

Generate reward

$$r_t = \boxed{v_t} + \beta\boxed{c_t}$$

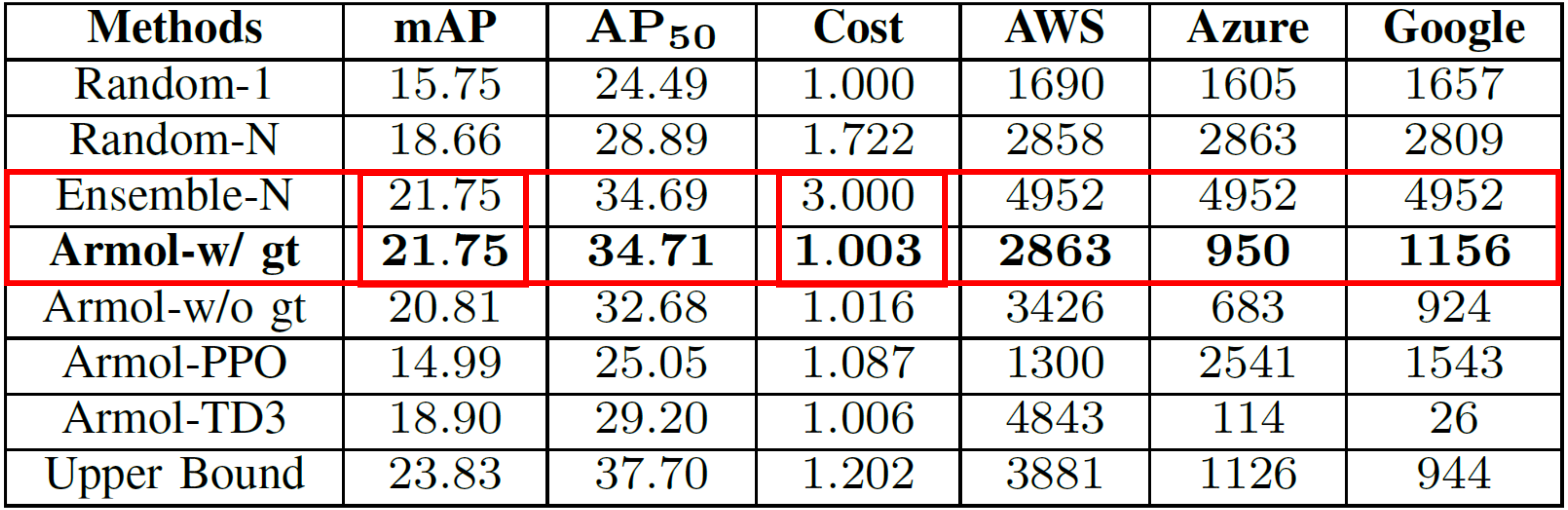


$$c_t = \sum_{i=1}^{n} a_{t,i}\, c_i$$

Performance metrics

- AP@50:
  - Average precision of predictions with a 50% IoU threshold.
- Cost:
  - Average cost in a test episode, in unit of 10^{-3} USD.

Comparison with other baselines

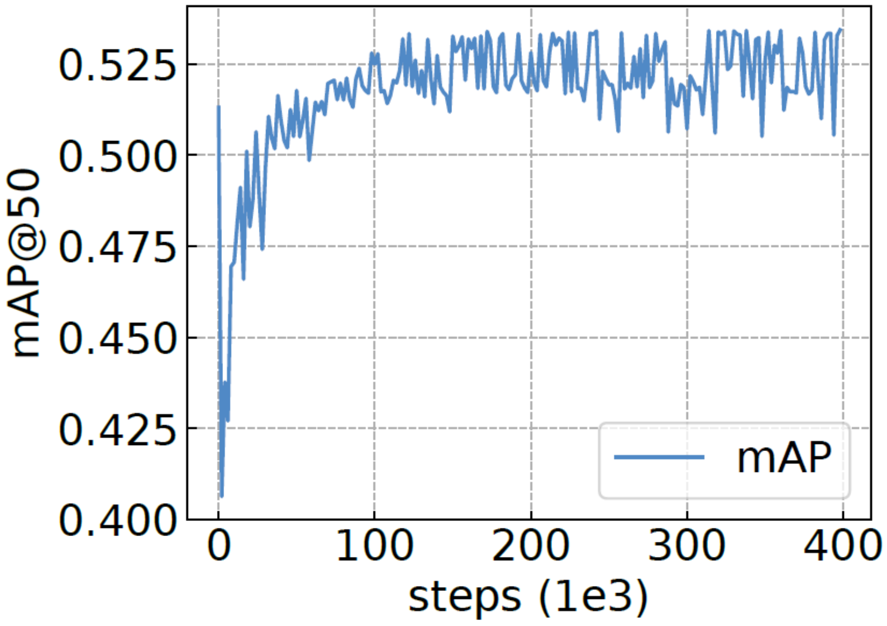


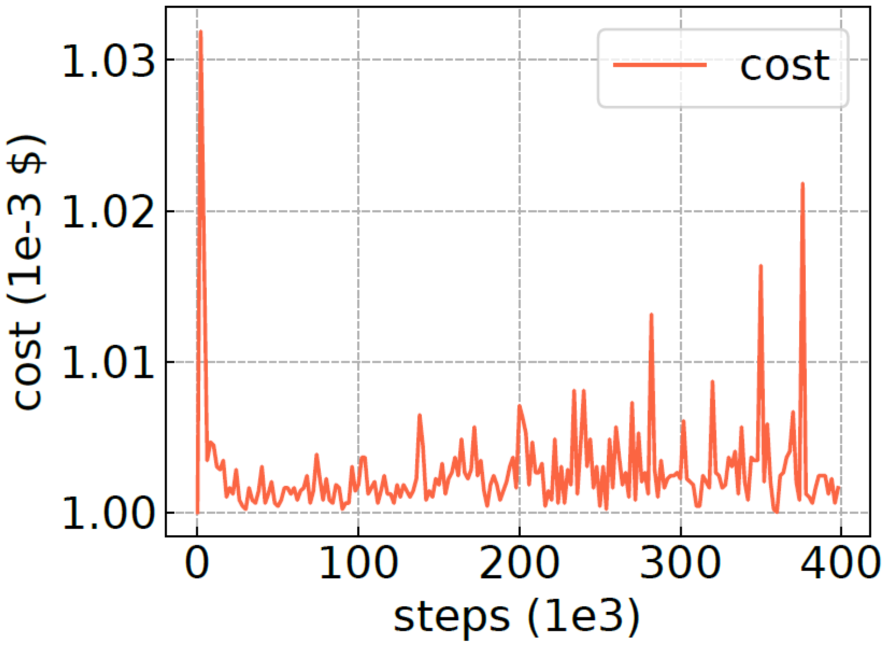
Methods	mAP	AP_{50}	Cost	AWS	Azure	Google
Random-1	15.75	24.49	1.000	1690	1605	1657
Random-N	18.66	28.89	1.722	2858	2863	2809
Ensemble-N	21.75	34.69	3.000	4952	4952	4952
Armol-w/ gt	**21.75**	**34.71**	**1.003**	**2863**	**950**	**1156**
Armol-w/o gt	20.81	32.68	1.016	3426	683	924
Armol-PPO	14.99	25.05	1.087	1300	2541	1543
Armol-TD3	18.90	29.20	1.006	4843	114	26
Upper Bound	23.83	37.70	1.202	3881	1126	944

Compared to “Ensemble-N”, our approach reduces the cost by 66%.

Scalability

We simulated 10 MLaaS providers.





Our framework converges at about 150,000 steps even with 10 available providers (1023 actions)

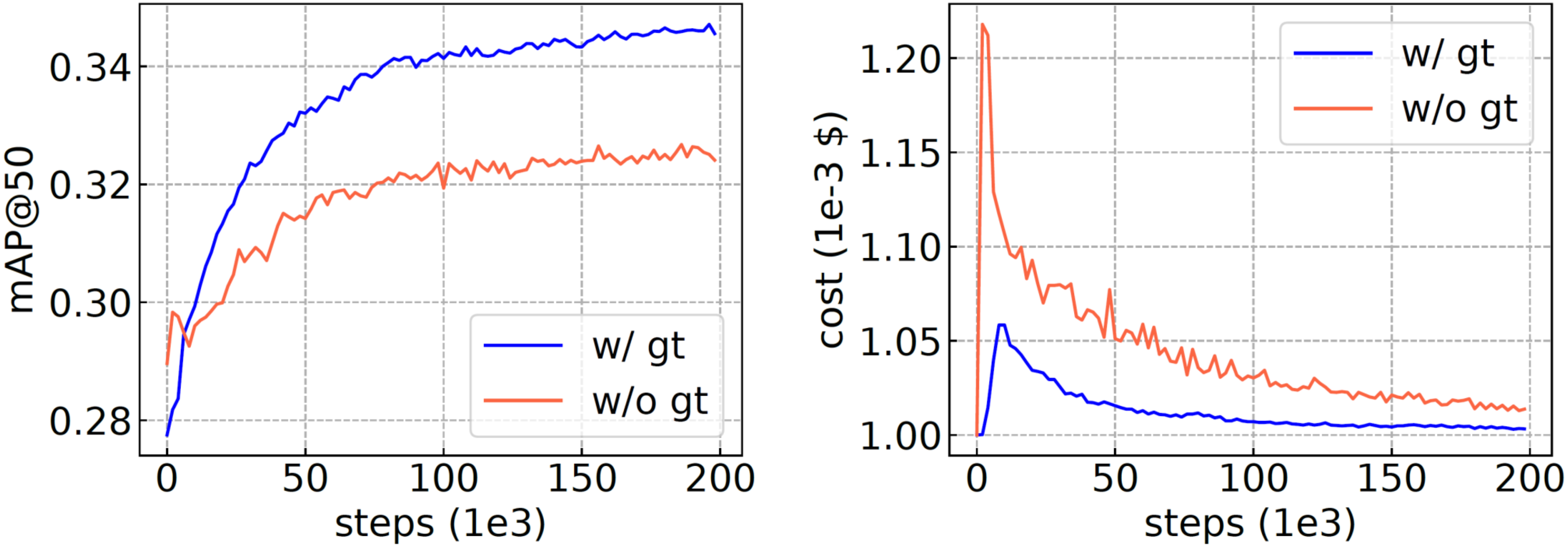
Conclusion

- Our contribution:
  - Measurement studies on major cloud providers reveal the varying differences among existing MLaaS offerings and the great potential in MLaaS federation to improve analytic performance.
  - We formulate the MLaaS federation problem as a combinatorial provider selection problem and propose a combinatorial reinforcement learning-based approach to maximize accuracy.
  - Efficient ensemble and grouping strategies are proposed to unify the vocabulary of different providers and aggregate the eventual results.

- More resources: https://github.com/ShuzhaoXie/Armol

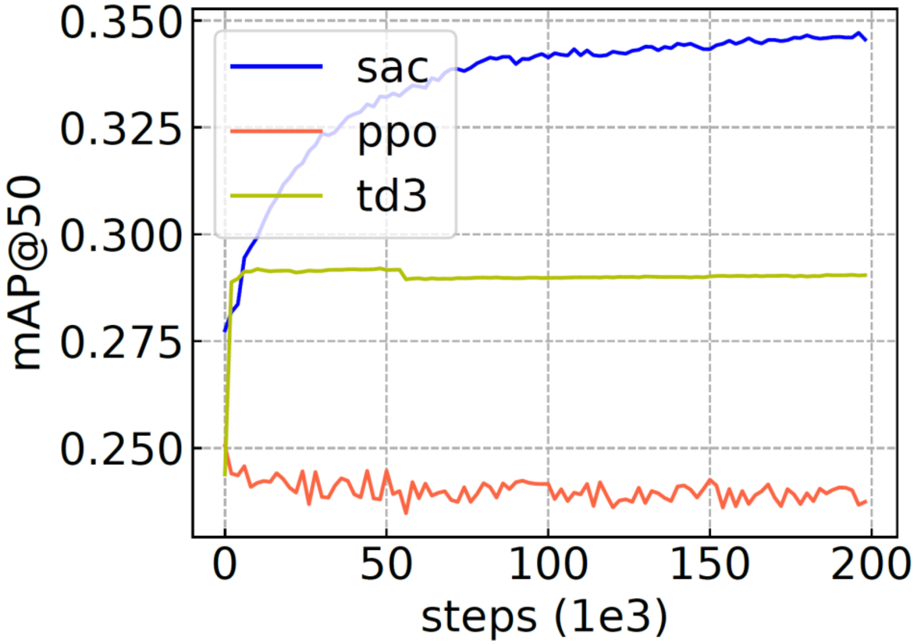
Thank you!

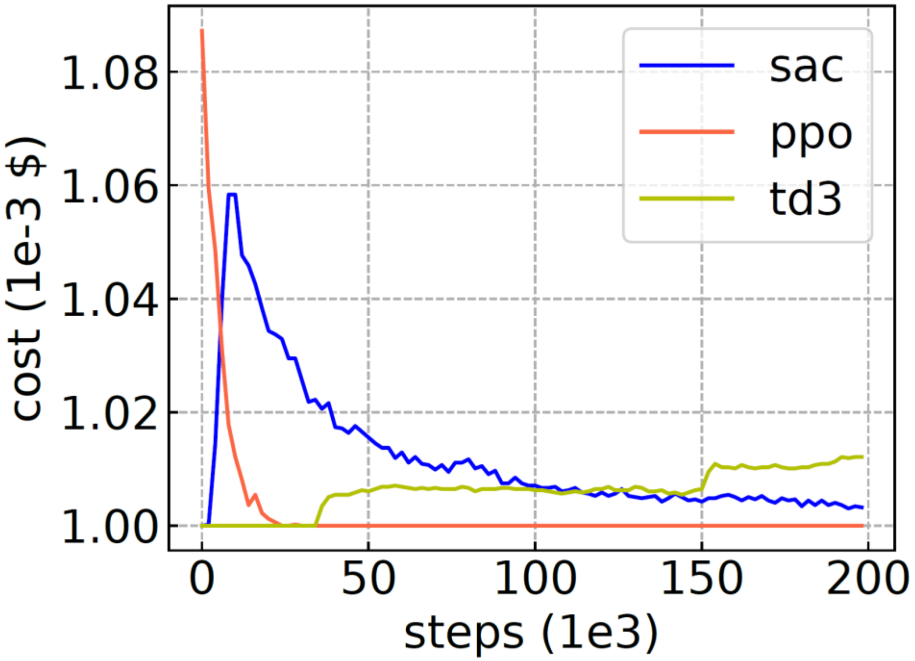
Offline vs. Online



Without ground truth, our method still achieves higher accuracy with less cost.

Comparison with other training algorithms





SAC is better than PPO and TD3 in both accuracy and cost during training.