Taming the Data Divide to Enable AI-Driven Networks

Alessandro Finamore

Principal engineer, AI4NET Datacom lab
Huawei Technologies Research Center Paris

alessandro.finamore@huawei.com
afinamore.io

ifip tc6 TMA CONFERENCE 2022

HUAWEI

The talk in 1 slide

1. "Our" challenges surrounding data and replicability
2. How other communities face those challenges

I have a lot of Qs … but not many As

HUAWEI

Part I

Introduction

Part II

Data-sharing & reproducibility

Part III

Easing the data-dependency

Part I

Introduction

Part II

Data-sharing & reproducibility

Part III

Easing the data-dependency

Rewinding time

COST – EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

IC0703 - Data Traffic Monitoring and Analysis:

theory, techniques, tools and applications for the future networks

(2008 – 2012)

Samos, Greece

Sep, 2008 – TMA meeting @ Samos

My "first contact" with the European traffic measurements research community

...since Samos'08

Cooperation with Italian ISP on traffic monitoring

Malware traffic analysis

Country-scale mobile network analytics

Customer satisfaction modeling via BigData and Machine learning

AI technologies applied to traffic monitoring

Telcos data as the common denominator

Huawei R&D in a nutshell

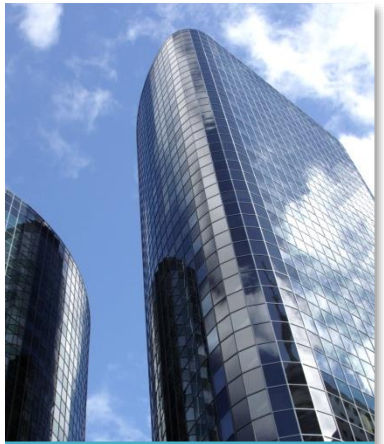
180,000
Employees

80,000
R&D employees

170+
Countries

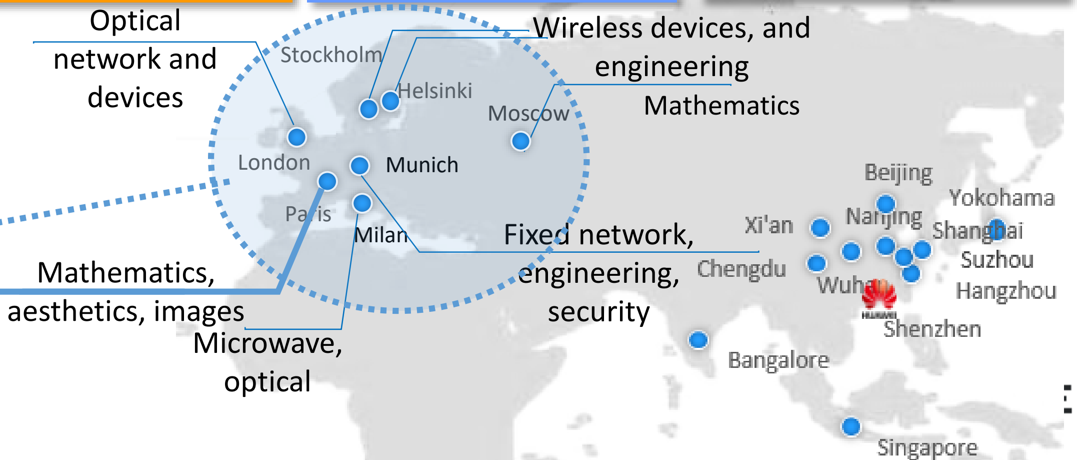
15+
Oversea R&D Centers

No. 70
Interbrand's Top100 Global Brands



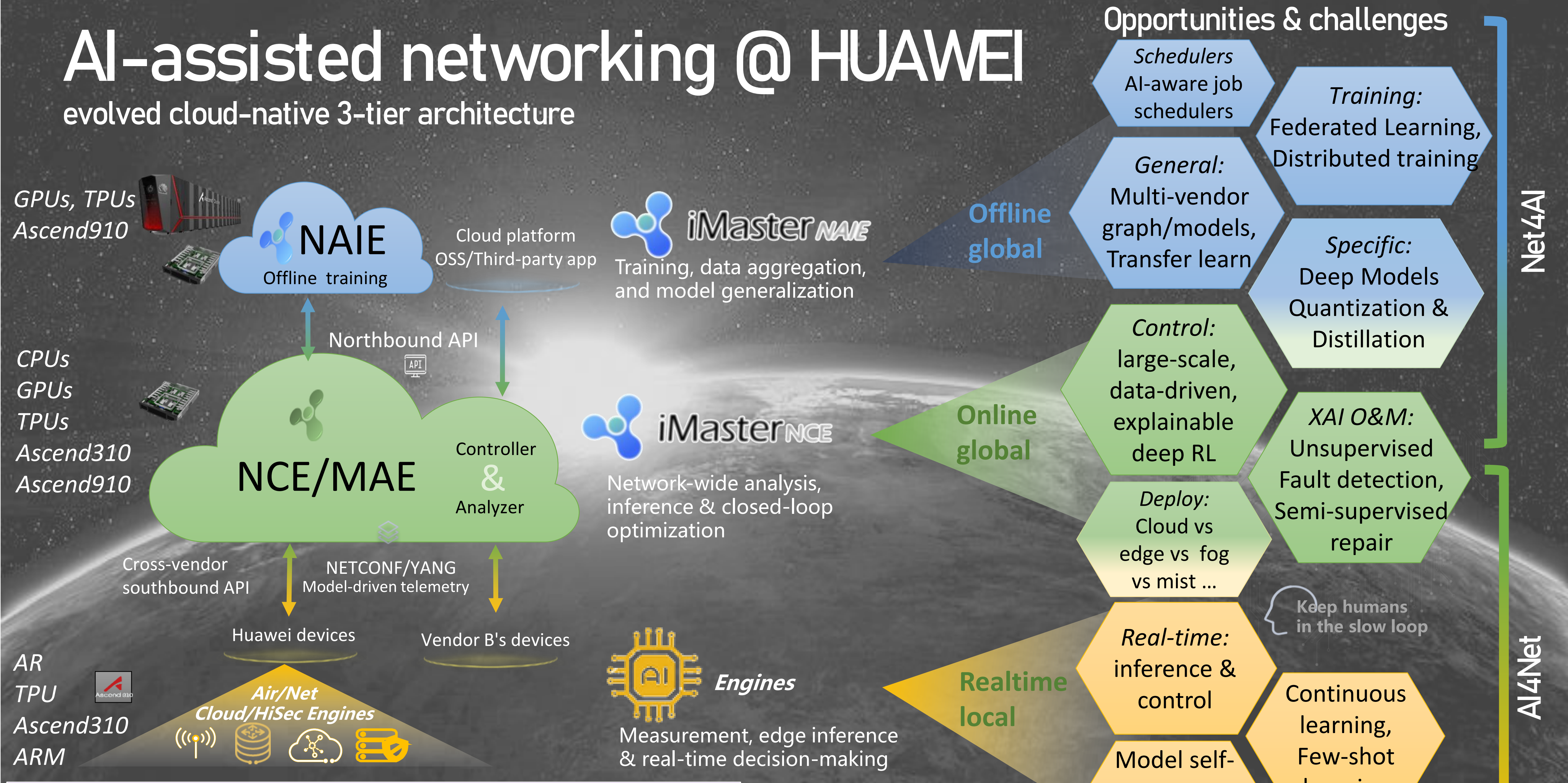
No. 83
Fortune Global 500

~15% of revenue re-invested in R&D
~30/70% long/short-term split
~2000 researchers in European Research Institute (ERI)
~150 researchers in Paris RC
~30 network researchers (aka DataCom Lab)

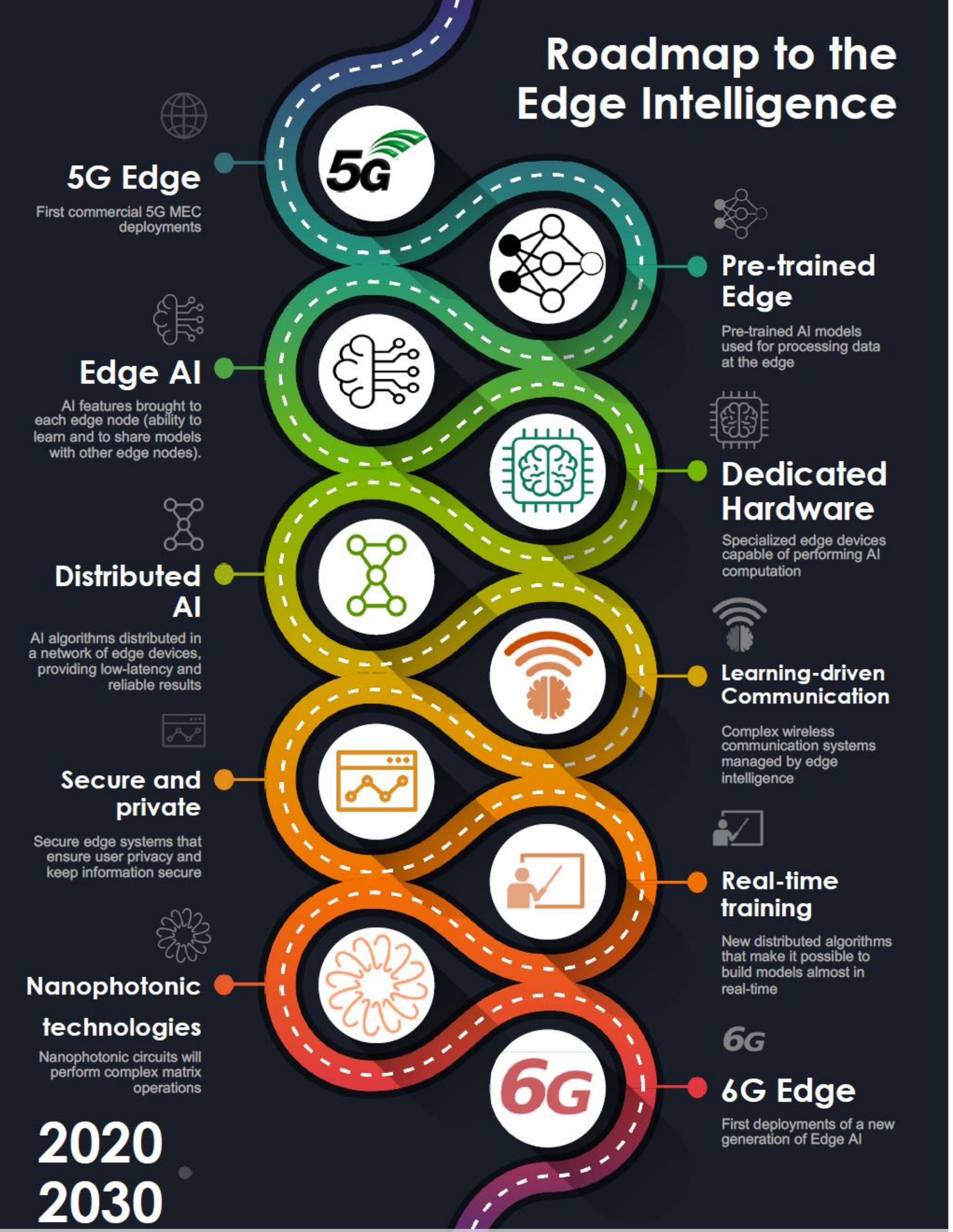


AI-assisted networking @ HUAWEI

evolved cloud-native 3-tier architecture



TNSM'22 : Landing AI on Networks: An equipment vendor viewpoint on Autonomous Driving Networks



Ella Peltonen and et al.. 6G white paper on edge intelligence. *6G RESEARCH VISIONS, NO. 8* (2020).

https://www.fortunebusinessinsights.com/infographics/artificial-intelligence-ai-chipsets-market-104500

https://www.hackerearth.com/blog/developers/applications-of-artificial-intelligence/

...but are we ready for such AI data-driven networks?

https://xkcd.com/1838/

HUAWEI

People smarter than me say…

MARCH 2021
WED 24
theNetworking Channel
https://networkingchannel.eu/
YouTube

The network will be programmed by many, operated by a few

Nick McKeown – Stanford University

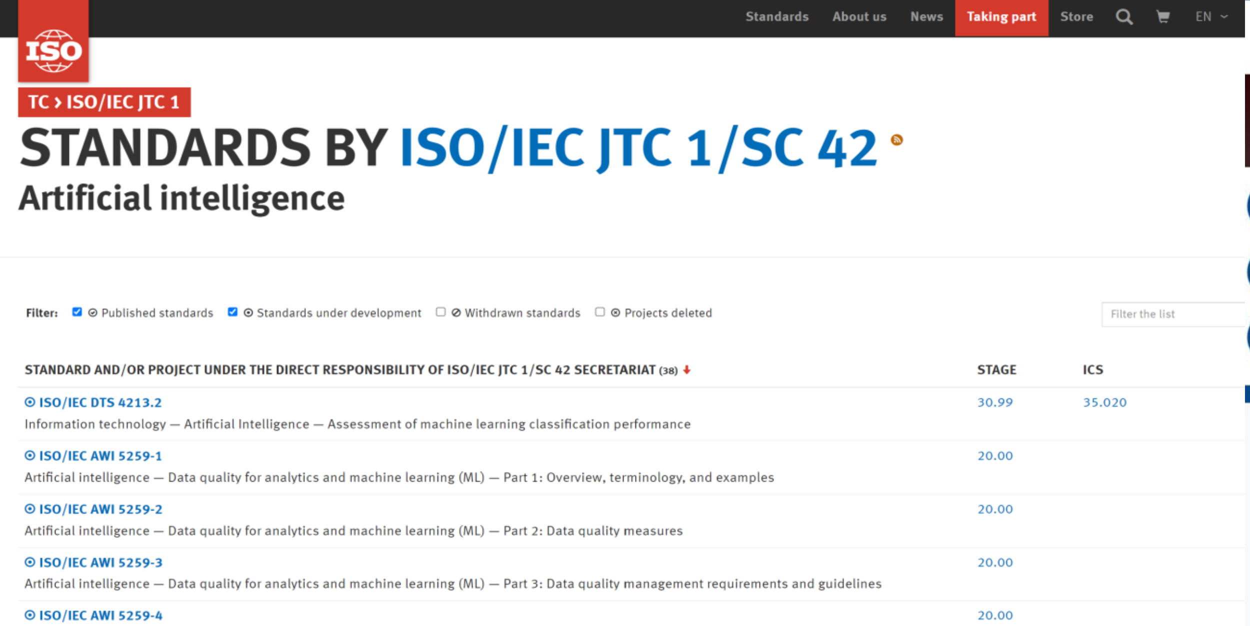
With a panel of graduate student discussants from around the world.

« *Machine learning is very good at understanding and predicting the behavior of systems we do not understand* […] *but networking is mostly about implementing something according to a “model” we already know* »

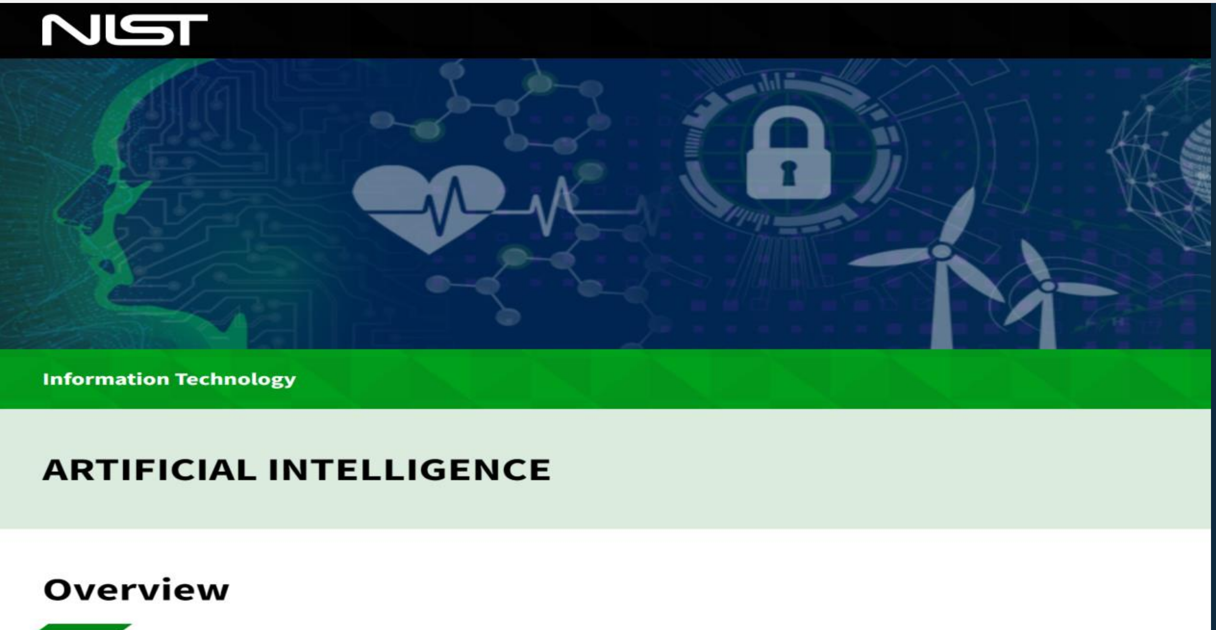
My take

AI is an opportunity BUT *if/what* we need to and *where/how* to integrate AI in networks is still largely an ongoing debate

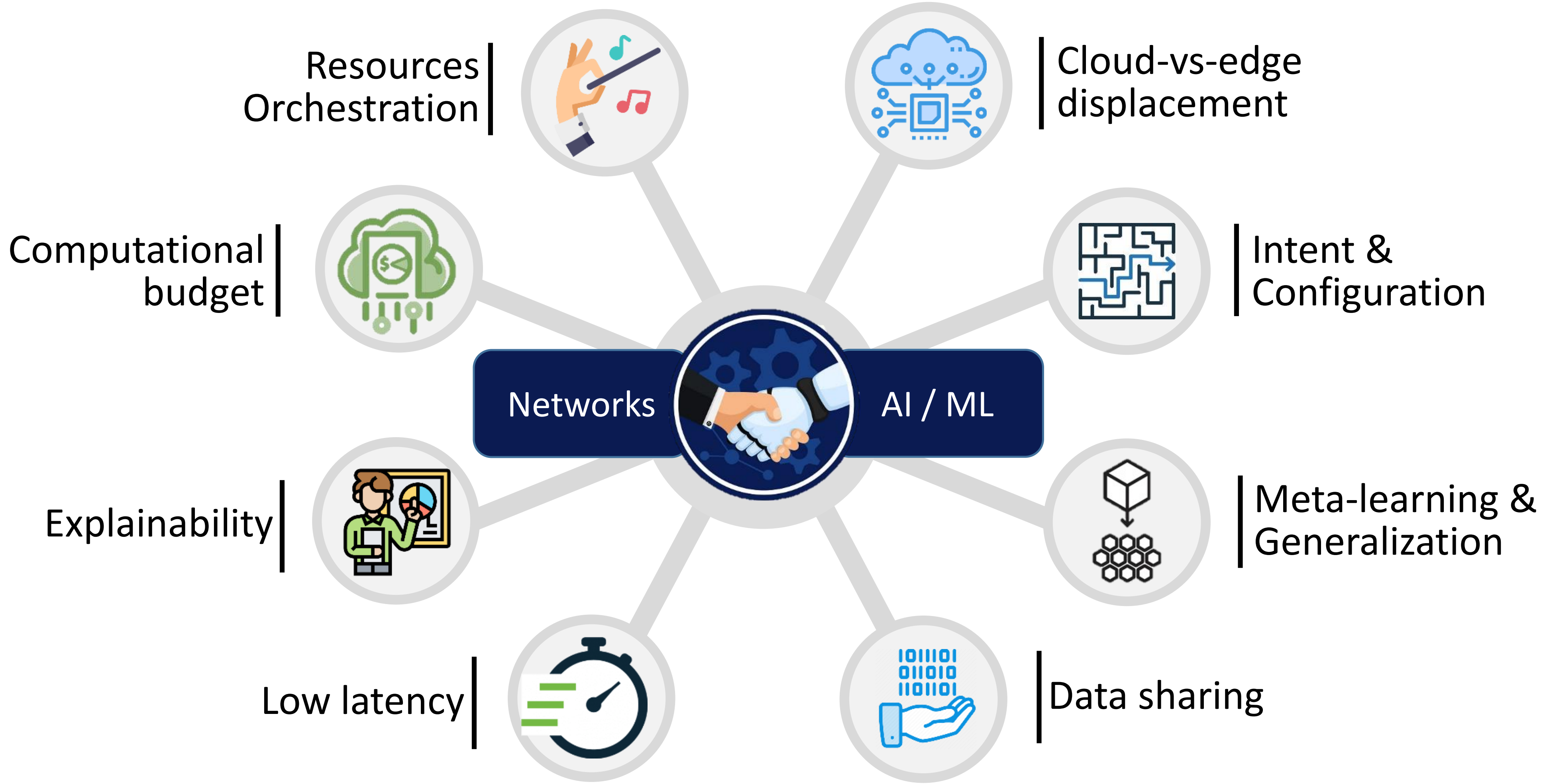
HUAWEI



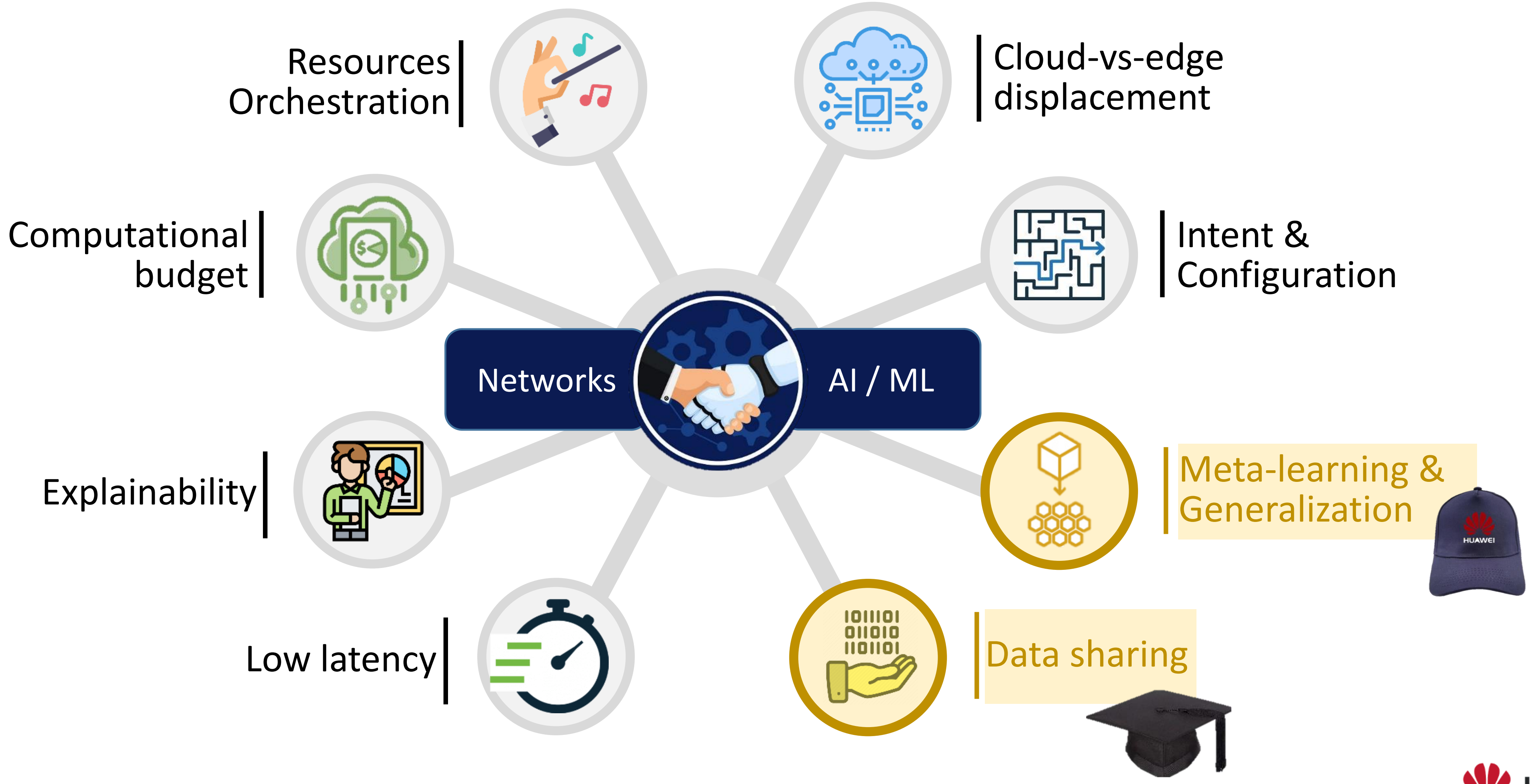
Standards are in their infancy



AI and networks: a multi-faced relationship



AI and networks: a multi-faced relationship



HUAWEI

Part I

Introduction

Part II

Data-sharing & reproducibility

Part III

Easing the data-dependency

Part I

Introduction

Part II

Data-sharing & reproducibility

In a broad sense, but in particular **fine-grained logs** *across planes/layers*

Part III

Easing the data-dependency

"We" know how to build platforms

RIPE NCC
RIPE Atlas

MLAB

BGPSTREAM

ARCHIPELAGO

EdgeNet

PLANETLAB
An open platform for developing, deploying, and accessing planetary-scale services

emulab

FIT IOT-LAB

geni
Exploring Networks of the Future

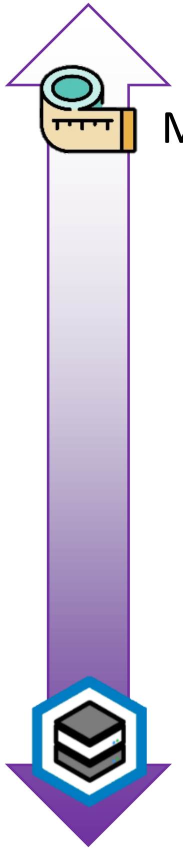
OneLab
FUTURE INTERNET TESTBEDS

FED4FIRE
FEDERATION FOR FIRE PLUS

Chameleon

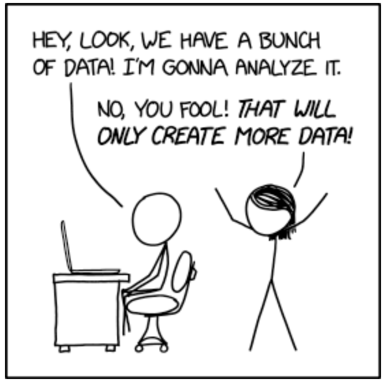
CloudLab

Grid'5000



The top-2 (known) problems in our community

Public data-access & **Measurements longevity**



https://xkcd.com/2582/

…and AI is just renewing this (old) divide

HUAWEI

The increasing data divide

Hotnets'19

An Effort to Democratize Networking Research in the Era of AI/ML

Arpit Gupta, UC Santa Barbara, Santa Barbara, CA
Chris Mac-Stoker, NIKSUN Inc., Princeton, NJ
Walter Willinger, NIKSUN Inc., Princeton, NJ

[…] *publicly available network measurements in support of network automation tasks are rare, not necessarily representative, often a by-product of some other measurement activities* […]

Recommendation: treat university campus networks are real production environments

CCR'22

Data-driven Networking Research: models for academic collaboration with industry (a Google point of view)

Jeffrey C. Mogul
Priya Mahadevan
Christophe Diot
Google
network-data-sharing@google.com

John Wilkes
Phillipa Gill
Amin Vahdat
Google

[…] *We encourage academic researchers to focus less on "can we obtain network-related data from Google?" and more on "how can we do more collaborative, data-driven networking research with Google?"* […]

Recommendation: setup ad-hoc collaboration between Companies (Google) and researchers

CCR'21

Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021) Final Report

kc claffy, UCSD/CAIDA, kc@caida.org
David Clark, MIT/CSAIL, ddc@csail.mit.edu
Fabián E. Bustamante, Northwestern, fabian@cs.northwestern.edu
John Heidemann, USC/ISI, johnh@isi.edu
Mattijs Jonker, University of Twente, m.jonker@utwente.nl
Aaron Schulman, UC San Diego, schulman@cs.ucsd.edu
Ellen Zegura, Georgia Tech, ewz@cc.gatech.edu

[…] *data sets require longitudinal (long-term, ongoing) data collection and sharing, support for which is more challenging for Internet research than other fields* […] *But often an employee of the associated company is an author on the paper, which triggers concerns regarding scientific objectivity* […]

Recommendations:

- New model for cross-sectors collaborations
- Fund longitudinal measurements platforms
- Annual state-of-the-Internet report/conference
- Data sharing code of conduct →

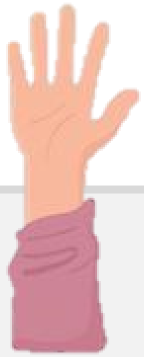
- Data is made available in curated repositories, or otherwise provided in ways that allows adequate access for legitimate scientific research
- Access requires registration with data source and legitimate research need
- Standard anonymization methods are used where needed
- Recipients agree to not repost corpus
- Recipients agree that they will not deanonymize data
- Recipients can publish analysis and data examples necessary to review research
- Recipients agree to use accepted protocols when revealing sensitive data, such as security vulnerabilities or data on human subjects
- Recipients agree to cite the repository and provide publications back to repository
- Repository can curate enriched products developed by researchers

Table 1: *Codes of conduct have been developed that enable responsible sharing of data in ways that protect stakeholders while allowing research [21, 22].*

Deeper roots than the mere lack of incentives

1. Hard constraints (privacy/business concerns)
2. Soft constraints (scale, operational risk, staff time)
3. Anonymization is not a panacea
4. Data/code cleansing for public release is time-consuming

…no, but wait a sec



1. We do have **Best Dataset awards** @ TMA, IMC, PAM

   Yes, and that's Awesome

   …but very opportunistic and tied to specific studies

2. We also have **badges**

   Yes, so let's talk about those

HUAWEI

Are we improving at replicability ?

Simple

Complex

Functional
The artifacts associated with the research are found to be documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation.

Reusable
The artifacts associated with the paper are of a quality that significantly exceeds minimal functionality.

Available
Author-created artifacts relevant to this paper have been placed on a publically accessible archival repository. […] they need not be complete in the sense described above.

Reproduced
results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author.

Replicated
Results of the paper have been independently obtained in a subsequent study by a person or team other than the authors, without the use of author-supplied artifacts.

Venue	2018	2021
IMC	16% (7/43)	0% *
CoNEXT	38% (12/31)	38% (14/36)
SIGCOMM	17% (7/40)	48% (28/58)

Numbers collected from programs
ACM Digital library

* Community award implied a dataset release (but apparently no badge was assigned)

However badges...

1. Do not imply (data) longevity
   (single snapshots of specific moment in time)

2. Do not imply (data) generalization
   (focus on specific problems)

Going beyond "badges"

Fabian E. Bustamante
@bustamantefe

All CS conferences, and for sure all the experimental ones @ACMSIGCOMM @usenix , should have a "replication" track to build confidence in the scientific merit of our results.

5:05 PM · Jun 9, 2022 · Twitter for iPhone

Oliver Hohlfeld @ohohlfeld · Jun 9
Replying to @bustamantefe @ACMSIGCOMM and @usenix
CS largely lacks behind other disciplines. E.g., meta-analysis plays a fundamental role in medicine-i.e. statistically combining data from multiple studies on a particular topic. Single efficacy studies are often too small to reliably assess risks. CS is not yet at this point.

2

HUAWEI

Going beyond “badges”

Fabian E. Bustamante
@bustamantefe

All CS conferences, and for sure all the experimental ones @ACMSIGCOMM @usenix , should have a “replication” track to build confidence in the scientific merit of our results.

5:05 PM · Jun 9, 2022 · Twitter for iPhone

Oliver Hohlfeld @ohohlfeld · Jun 9
Replying to @bustamantefe @ACMSIGCOMM and @usenix
CS largely lacks behind other disciplines. E.g., meta-analysis plays a fundamental role in medicine-i.e. statistically combining data from multiple studies on a particular topic. Single efficacy studies are often too small to reliably assess risks. CS is not yet at this point.

2

My take

- We already do replicability when doing the state-of-the-art comparison
- Yet, we lack venues to foster those discussions

HUAWEI

Meanwhile, in other communities

Long tradition for artifacts (SIGMOD since 2008, VLDB since 2012)

Authors submits artifacts evaluated by a committee

Yearly reproducibility challenge

“Crowdsourced”: select a paper, reproduce it, and submit a report

Papers With Code

Data-sharing is a core value for earth & space-related sciences

Sharing for the benefit of the research community and humanity

Sara Issaoun PyCon22 keynote : *“Imaging a black hole with the event horizon telescope”*

Netflix documentary -- *Black holes The edge of all we know*

How AI communities foster debate around data

NeurIPS 2021 Datasets and Benchmarks Track

The pre-proceedings are now available! See the **NeurIPS Online Proceedings page**.

Quickly find papers in the **virtual conference**: **click on the paper in the Accepted Paper List**.

We are immensely grateful for the tremendous contributions of the **33 area chairs** and **548 reviewers** to make this new endeavor a success.

The **Datasets and Benchmarks track** serves as a novel venue for high-quality publications, talks, and posters on highly valuable machine learning datasets and benchmarks, as well as a forum for discussions on how to improve dataset development. Datasets and benchmarks are crucial for the development of machine learning methods, but also require their own publishing and reviewing guidelines. For instance, datasets can often not be reviewed in a double-blind fashion, and hence full anonymization will not be required. On the other hand, they do require additional specific checks, such as a proper description of how the data was collected, whether they show intrinsic bias, and whether they will remain accessible.

CRITERIA. We are aiming for an **equally stringent review** as the main conference, yet better suited to datasets and benchmarks. Submissions to this track will be **reviewed according to a set of criteria and best practices specifically designed for datasets and benchmarks**, as described below. A key criterion is accessibility: datasets should be available and accessible, i.e. the data can be found and obtained without a personal request to the PI, and any required code should be open source. Next to a scientific paper, authors should also submit supplementary materials such as detail on how the data was collected and organized, what kind of information it contains, how it should be used ethically and responsibly, as well as how it will be made available and maintained.

RELATIONSHIP TO NEURIPS. Submissions to the track will be **part of the main NeurIPS conference**, presented alongside the main conference papers. Accepted papers will be **officially published in associated proceedings** clearly linked to, yet separate from, the NeurIPS proceedings. The proceedings will be called ***Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*** and they will be **hosted on the NeurIPS website** next to the main NeurIPS proceedings. We will maintain a page on the NeurIPS website with all accepted datasets and additional information.

174 papers in the program!!!

SIGCOMM’21	58
INFOCOM’21	251
NeurIPS’21	2,300

HUAWEI

If I had three wishes for the genie

1. More “code challenges”
   They can be occasion to release data and put focus on specific problems

2. Create one permanent replicability track/workshop
   - *Decouple study state-of-the-art from promoting new ideas*
   - *Foster data/code sharing for the benefit of the community*

3. Federate universities/research centers for data access/sharing
   Break the barrier of 1-to-1 cooperations

The data divide affects the whole measurements community
AI-driven measurement methods is just exacerbating it

...anything else to mitigate the data divide?

What if we **reduce the dependency from data** ?

Part I

Introduction

Part II

Data-sharing & reproducibility

Part III

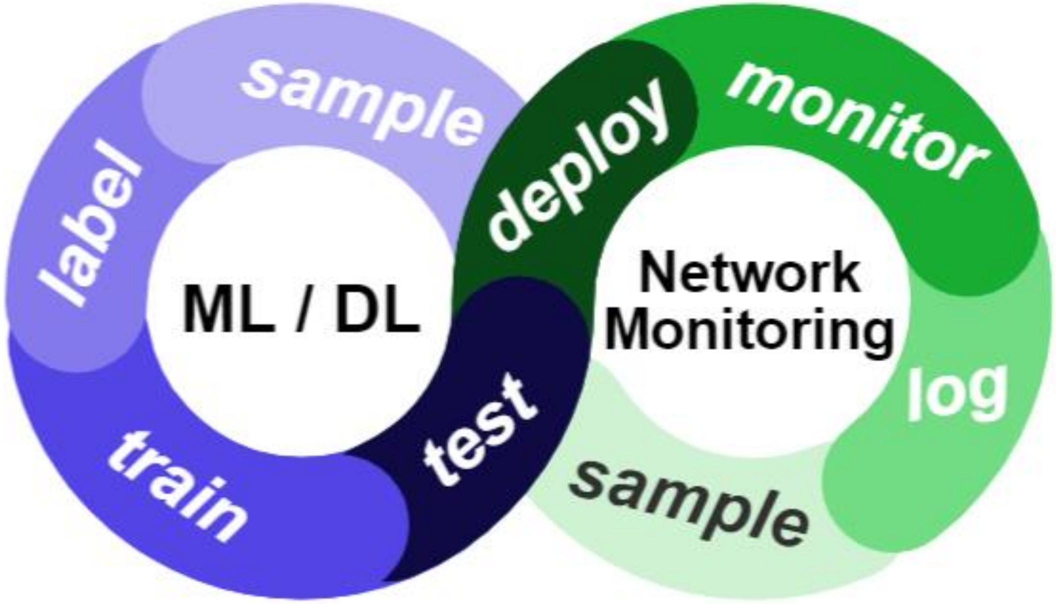
Easing the data-dependency

The infinite data loop in (AI-based) monitoring

How to reduce data-dependency

Continuous learning, Few shot learning

Self-supervised learning, repr. learning



How to reduce on-device costs

Real-time inference

Distributed agents coord.

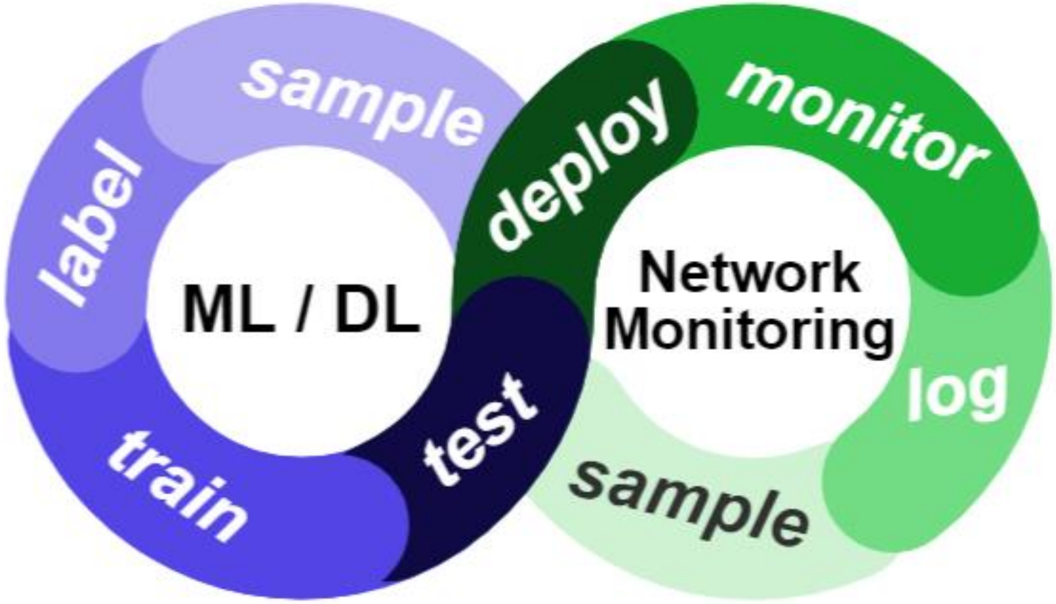
HUAWEI

The infinite data loop in (AI-based) monitoring

How to reduce data-dependency

Continuous learning, **Few shot learning**

Self-supervised learning, repr. learning



How to reduce on-device costs

Real-time inference

Distributed agents coord.

HUAWEI

Few-Shots Learning (FSL): A definition

Popular Classes
(many samples)

Rare classes
(few samples)

“Learning from a limited number of examples with supervised information”

Generalizing from a Few Examples: A Survey on Few-shot Learning (ACM Computing Surveys‘20)

ELI5 : Use the accumulated knowledge to solve a new problem using few examples as reference

The need for FSL

Biology

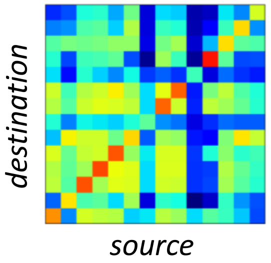
"People learning often generalize successfully from just a single example
…
people learn richer representations than machines do, using them for a wider range of functions, including creating new abstract categories of objects based on existing categories"

Joshua Tenenbaum
Cognitive Scientist
MacArthur Fellow (2019)

Empirical evidence

Massive (manual) *labeling is hard*
"Even the most well-known hand labeled datasets [ImageNet] have label error rates of at least 5%"

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks (arxiv'21)
[*https://labelerrors.com/*](https://labelerrors.com/)



Network *traffic is imbalanced* by nature
"Traffic is neither rack-local nor all-to-all; locality depends upon the service"

Inside the Social Network's (Datacenter) Network (SIGCOMM'15)

Let's consider a practical use-case

IEEE TNSM'21

Deep Learning and Zero-Day Traffic Classification:
Lessons learned from a commercial-grade dataset

Lixuan Yang, Alessandro Finamore, Feng Jun, Dario Rossi
Huawei Technologies, France

popular

A traffic classifier covering 200 classes (4x the literature)

…BTW, we are working internally so to release an anonymized version of the dataset to the community

unsupervised techniques such as clustering, received less coverage by the traffic classification literature which focuses deriving DL models via supervised techniques. More combination of supervised and unsupervised tech

in the Post-Snowden era, this second wave of research is ularly relevant to industry and telco vendors actively ploying statistical classification approaches — as pointed out in [20], until this point traffic classification

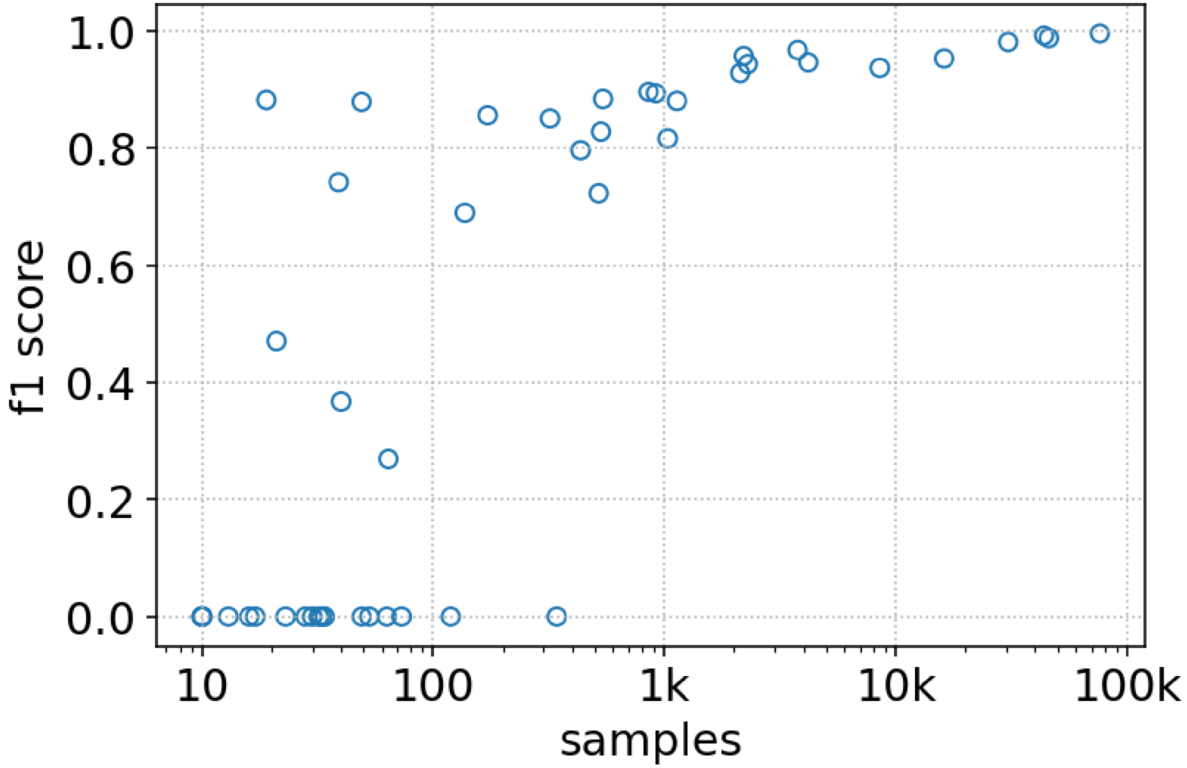
Long-term goal: A classifier covering O(1,000) classes
This talk: A toy-case example

HUAWEI

Modeling Toy-case via a CNN ( monolithic approach)

Dataset *(pkt-size + direction of first 10 pkts)*

- 45 classes with >= 10 samples
- Only 10 classes with > 10k samples

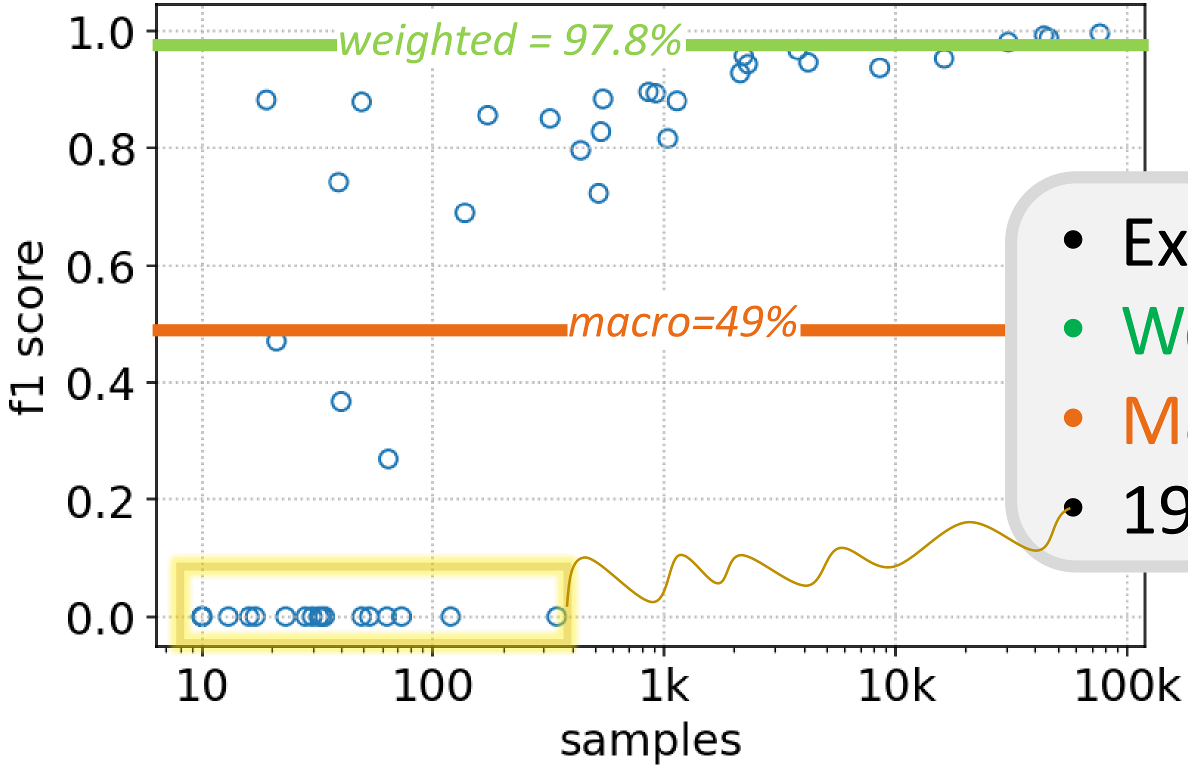


HUAWEI

Modeling Toy-case via a CNN (monolithic approach)

Dataset *(pkt-size + direction of first 10 pkts)*

- 45 classes with >= 10 samples
- Only 10 classes with > 10k samples



- Extreme imbalance : ρ = 7,000
- Weighted f1 score : 97.8%
- Macro f1 score : 49%
- 19 classes (42%) with F1 score = 0

Why learning from a few examples is hard?

FSL and Empirical risk

Why learning from a few examples is hard?

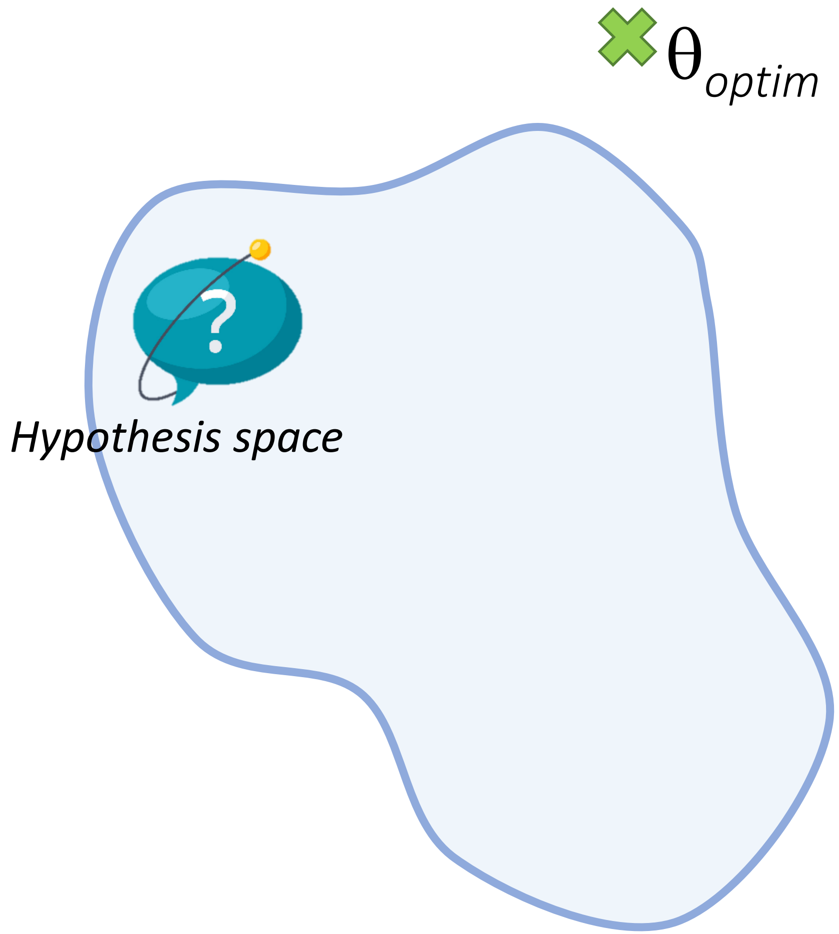
FSL and Empirical risk

✖ θ_{optim}

- A problem has an optimal solution

Why learning from a few examples is hard?

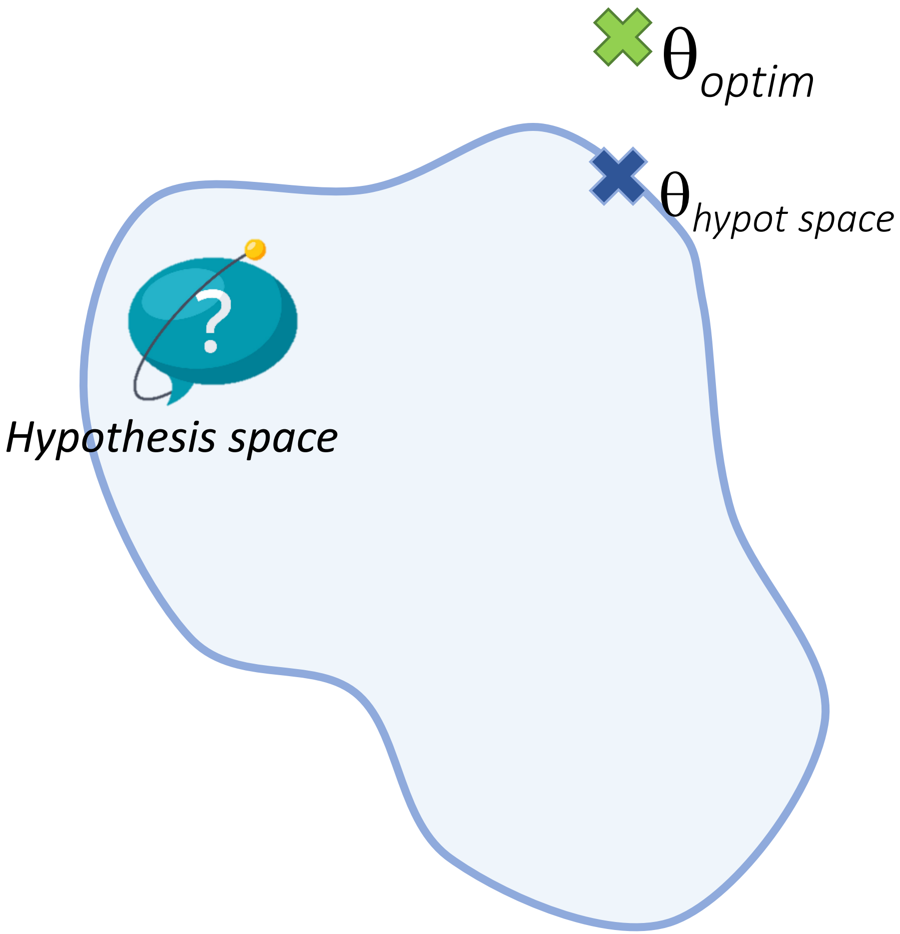
FSL and Empirical risk



- A problem has an optimal solution
- The hypothesis space constrains the best solution that can be found

Why learning from a few examples is hard?

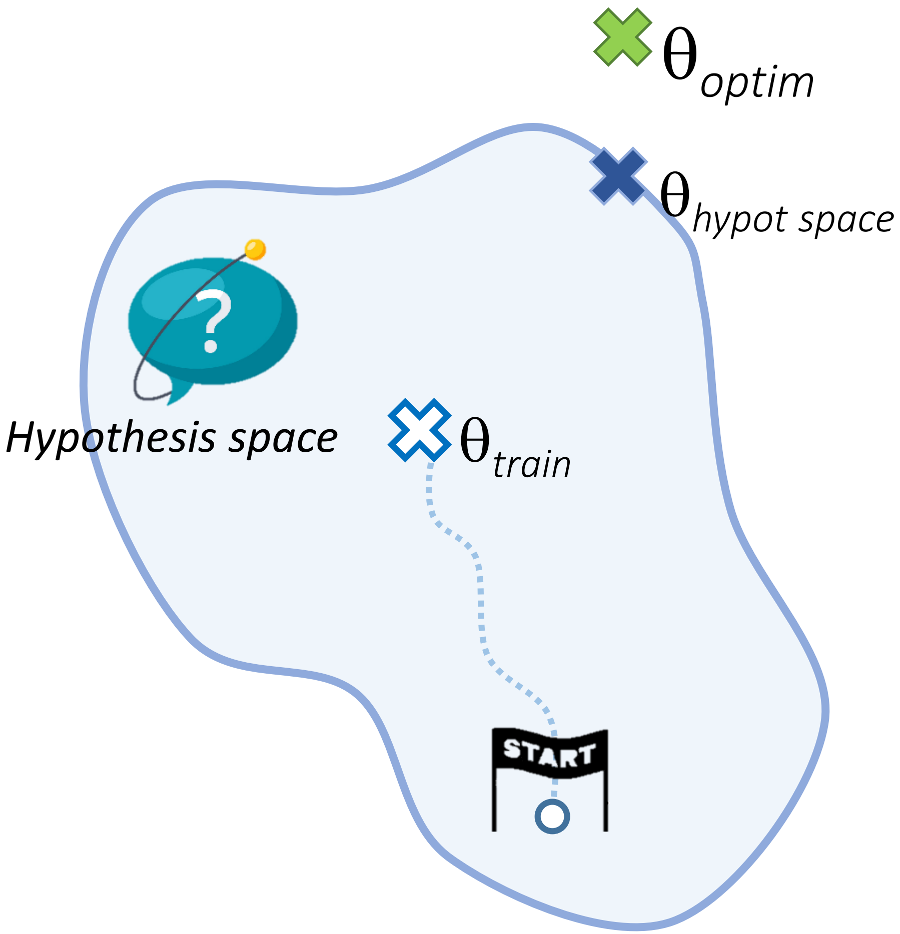
FSL and Empirical risk



- A problem has an optimal solution
- The hypothesis space constrains the best solution that can be found

HUAWEI

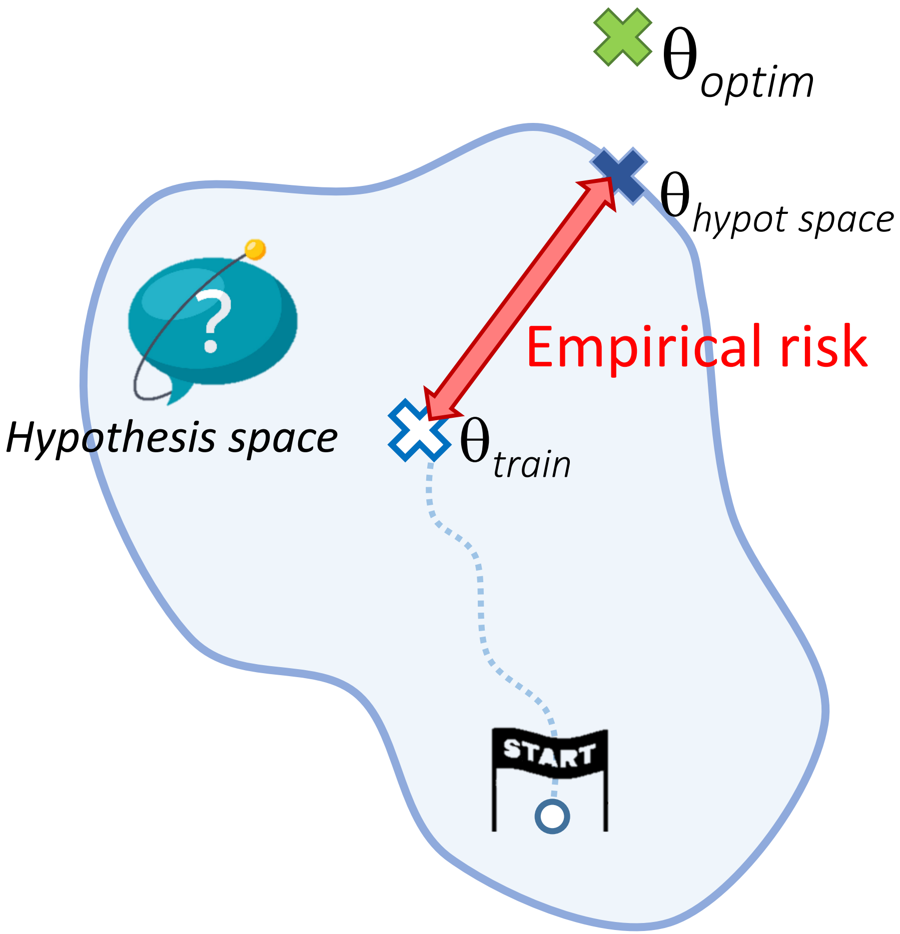
Why learning from a few examples is hard?
FSL and Empirical risk



- A problem has an optimal solution
- The hypothesis space constrains the best solution that can be found
- The search is further constrained by the data available (the more the data, the better)

Why learning from a few examples is hard?

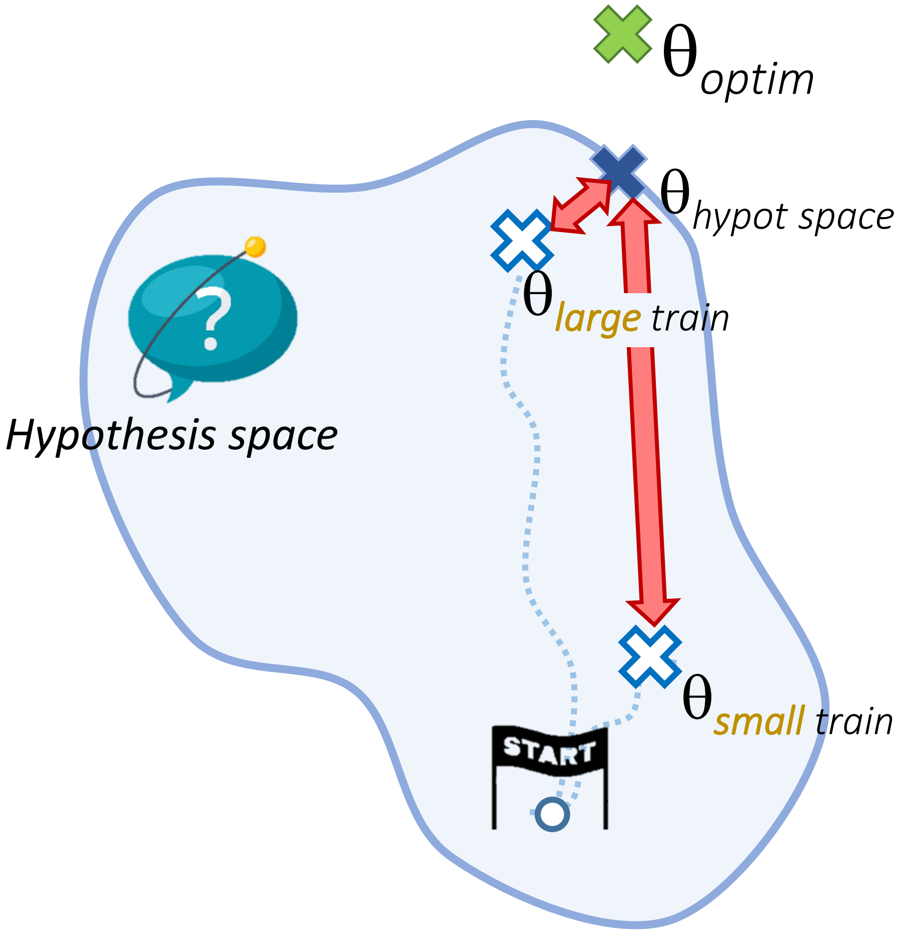
FSL and Empirical risk



- A problem has an optimal solution
- The hypothesis space constrains the best solution that can be found
- The search is further constrained by the data available (the more the data, the better)

Why learning from a few examples is hard?

FSL and Empirical risk



- A problem has an optimal solution
- The hypothesis space constrains the best solution that can be found
- The search is further constrained by the data available (the more the data, the better)
- FSL problems have high empirical risk

FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

"Hallucinate" the train set by introducing new (synthetic) data

Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models

FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

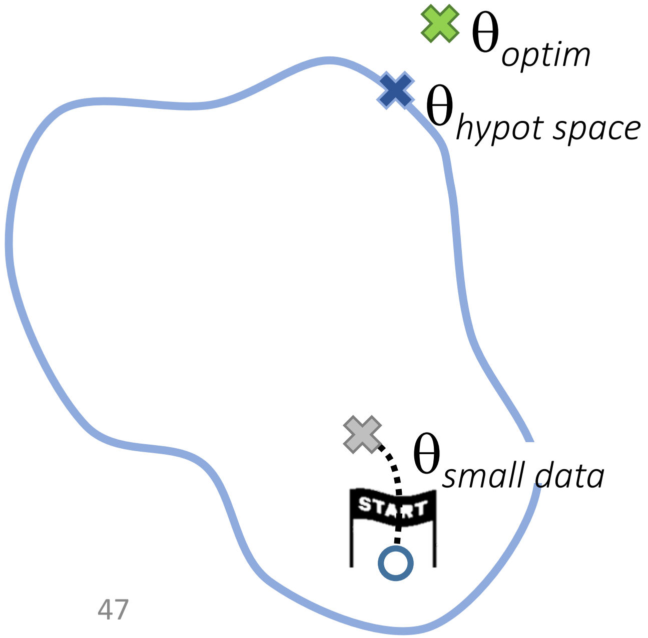
"Hallucinate" the train set by introducing new (synthetic) data

Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models



FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

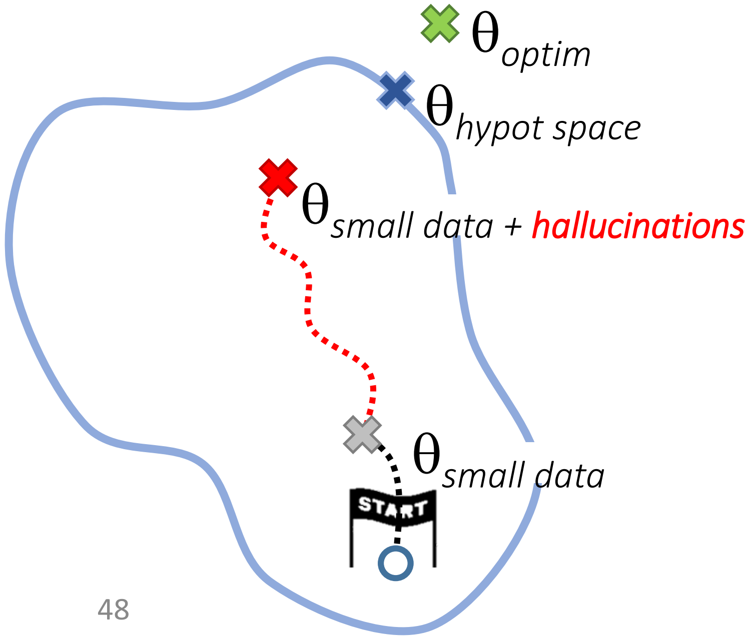
"Hallucinate" the train set by introducing new (synthetic) data

Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models



FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

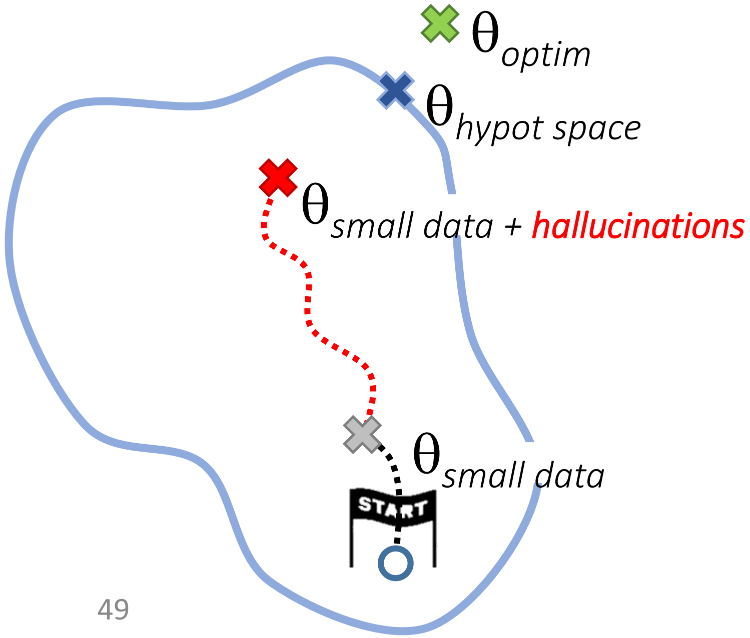
"Hallucinate" the train set by introducing new (synthetic) data

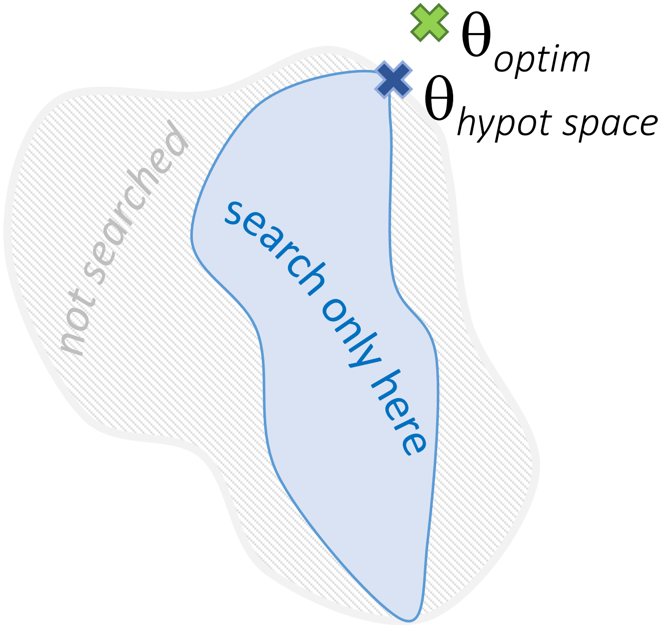
Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models





HUAWEI

FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

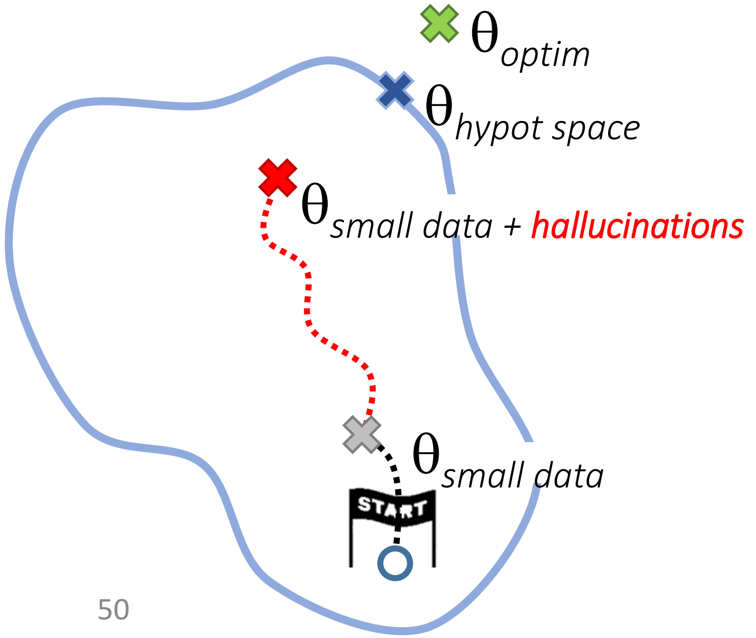
"Hallucinate" the train set by introducing new (synthetic) data

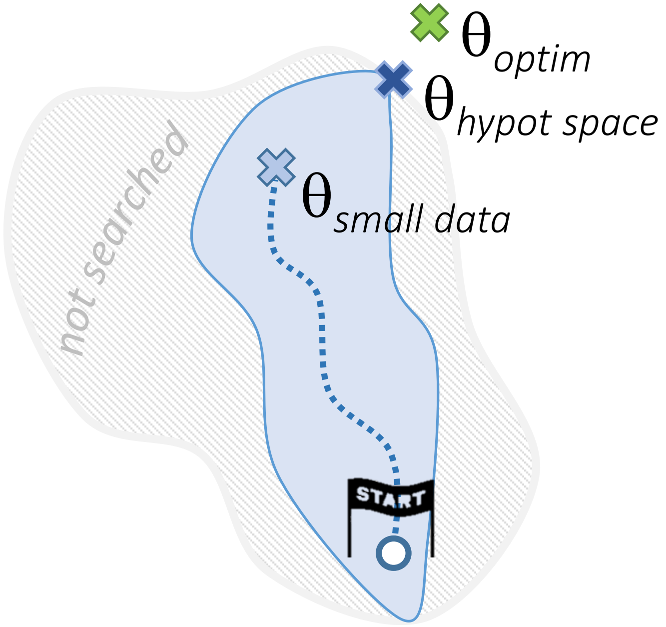
Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models





HUAWEI

FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

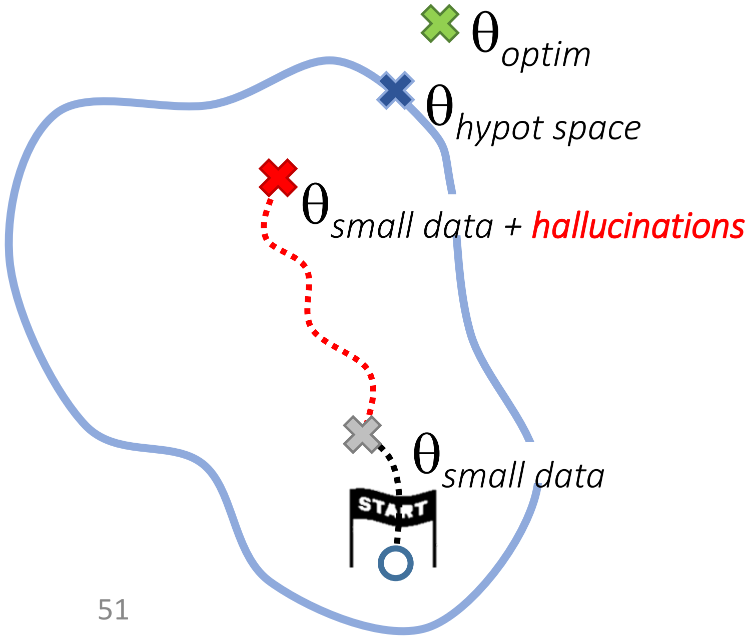
"Hallucinate" the train set by introducing new (synthetic) data

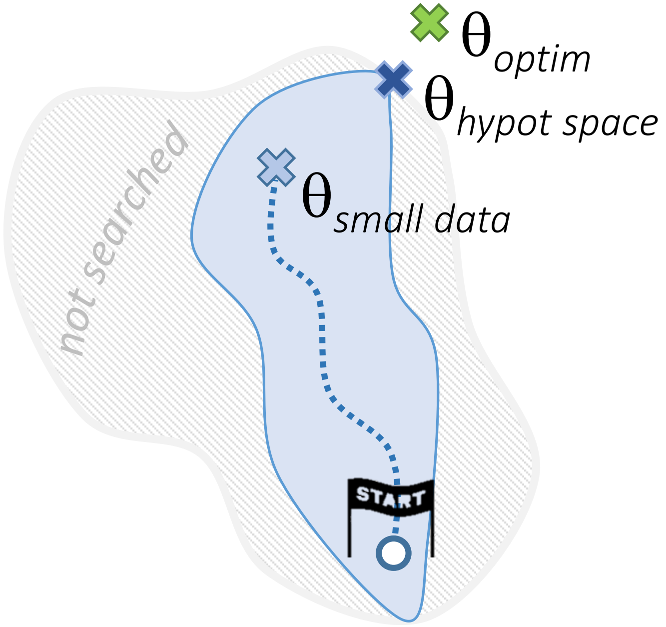
Model: learn to *compare*

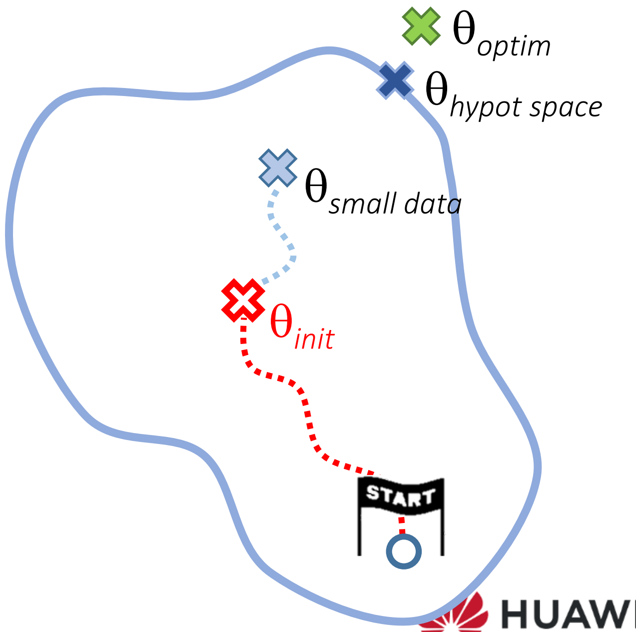
FSL models derive from baseline models (new classes "compare" with baseline ones)

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models







FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

"Hallucinate" the train set by introducing new (synthetic) data

Extra knowledge = Generative model

Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Extra knowledge = a pre-trained model

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models

Extra knowledge = a pre-trained model

The need for supplementary knowledge

Unexperienced

+

Very wise

=

Happy

FSL methods taxonomy

How to handle empirical risk

FSL

Data: learn to *augment*

"Hallucinate" the train set by introducing new (synthetic) data

Extra knowledge = Generative model

Model: learn to *compare*

FSL models derive from baseline models (new classes "compare" with baseline ones)

Extra knowledge = a pre-trained model

Algorithm: learn to *initialize*

Learn a generalized model from which is easy to derive FSL models

Extra knowledge = a pre-trained model

The need for supplementary knowledge

Meta-Learning methods extract such supplementary knowledge

Meta-learning: *“learning to learn”*

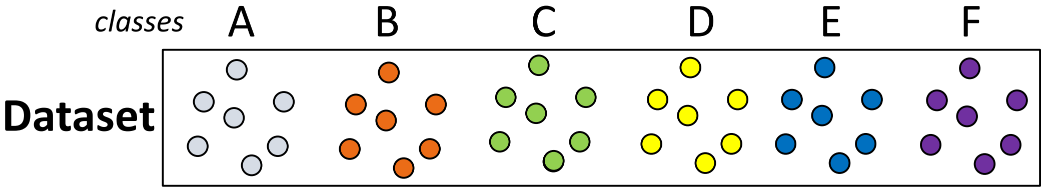
“The goal of the trained model is to quickly learn a new task from a small amount of new data, and the model is trained by the meta-learner to be able to learn on a large number of different tasks.”

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks (ICML’17)

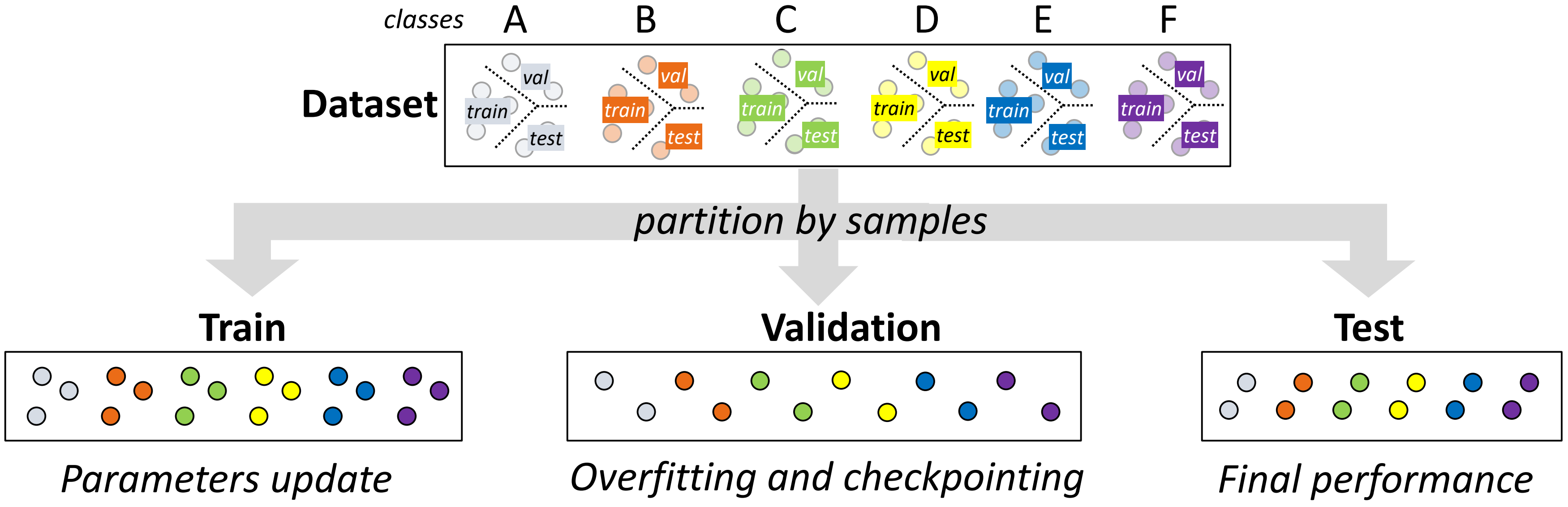
The “unit of learning” are tasks NOT samples

Aim for a higher level of abstraction/generalization

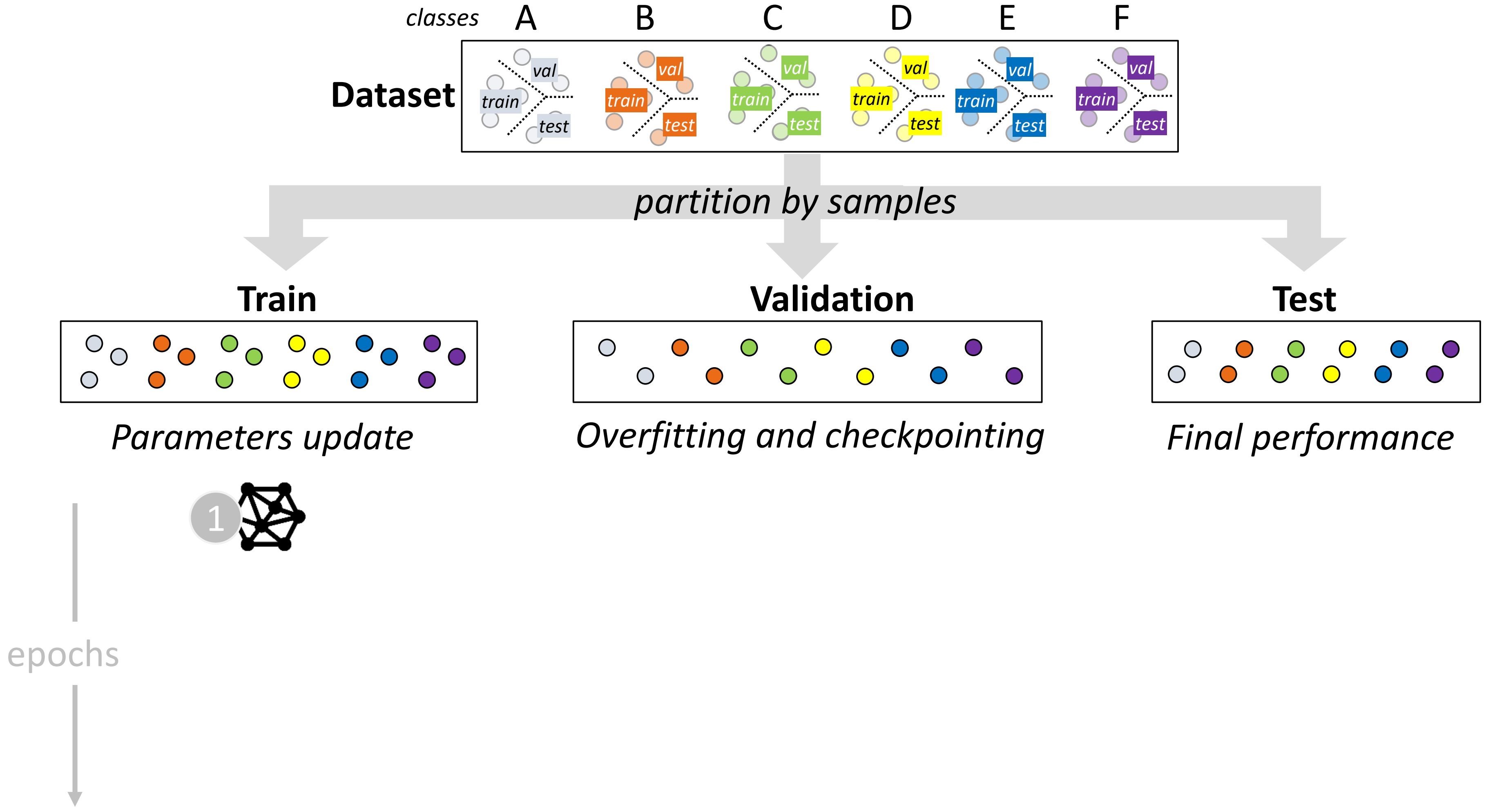
Traditional model training



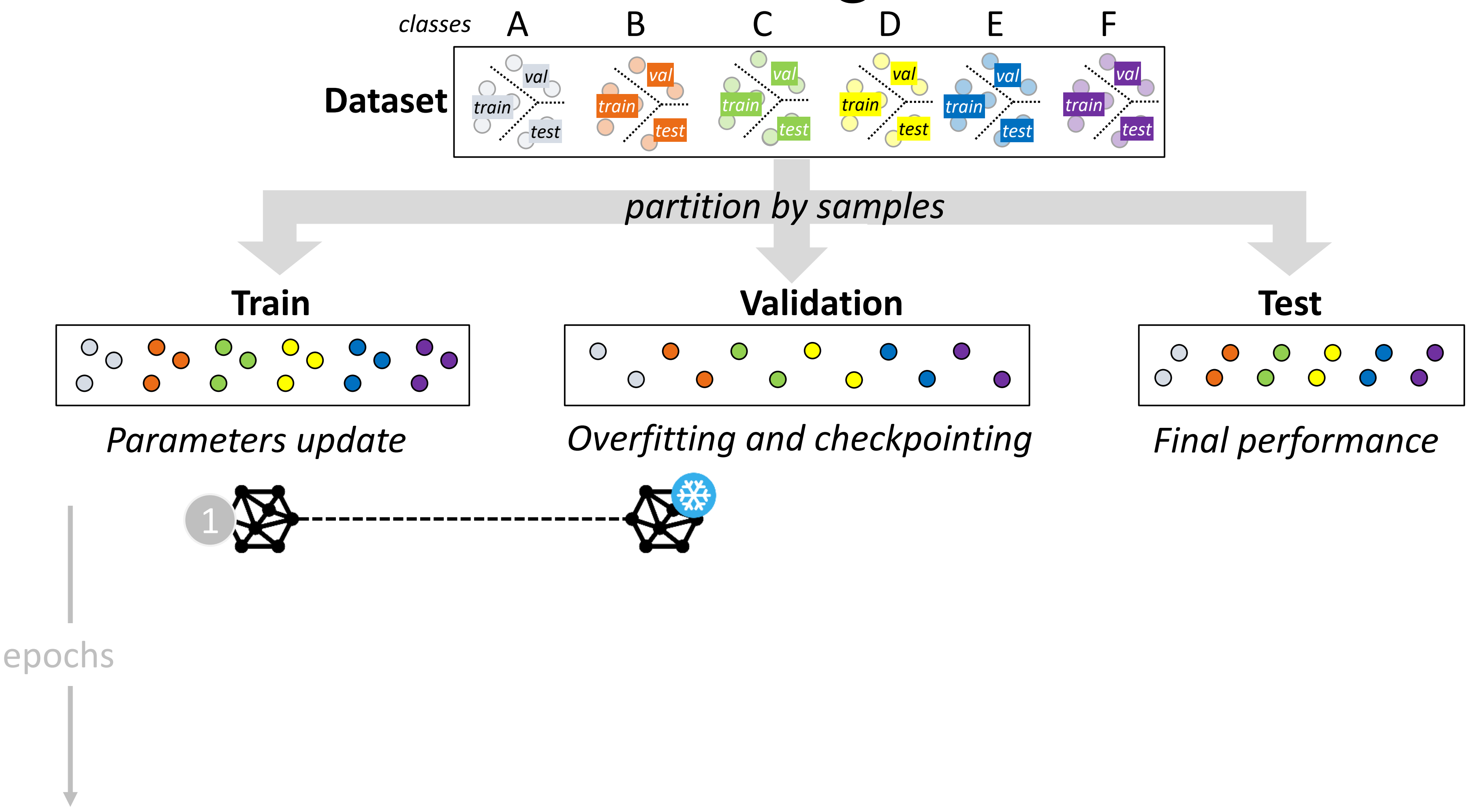
Traditional model training



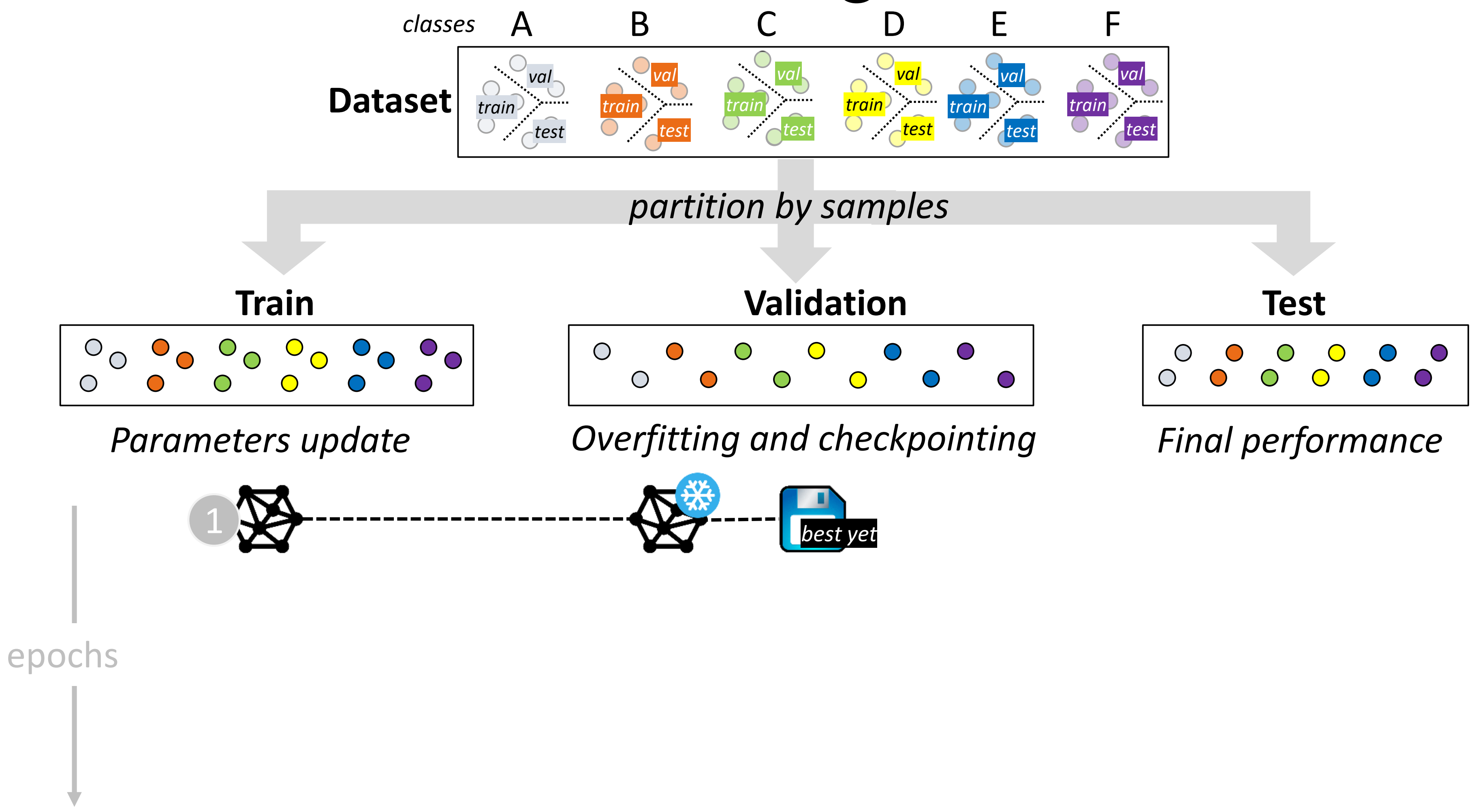
Traditional model training



Traditional model training

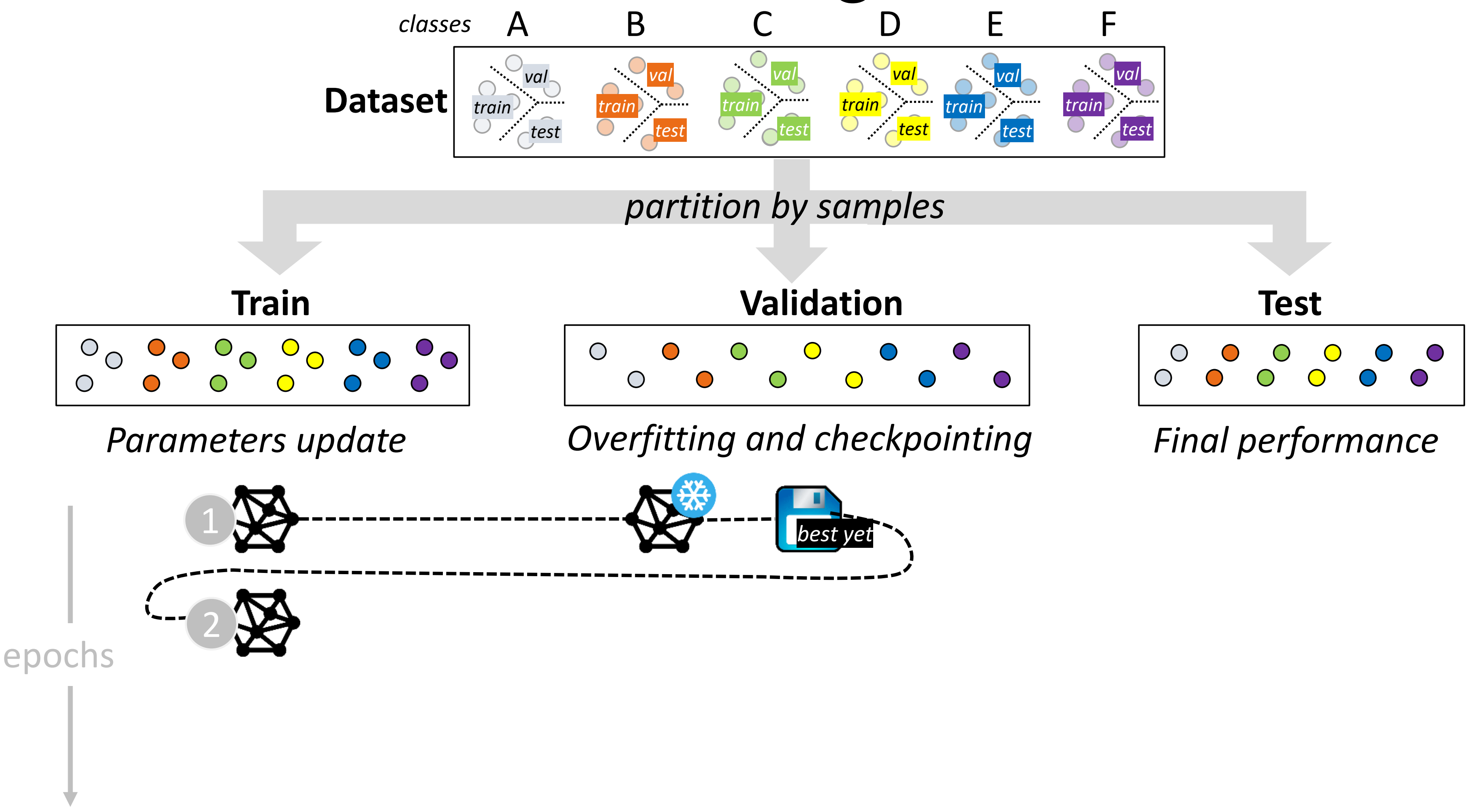


Traditional model training



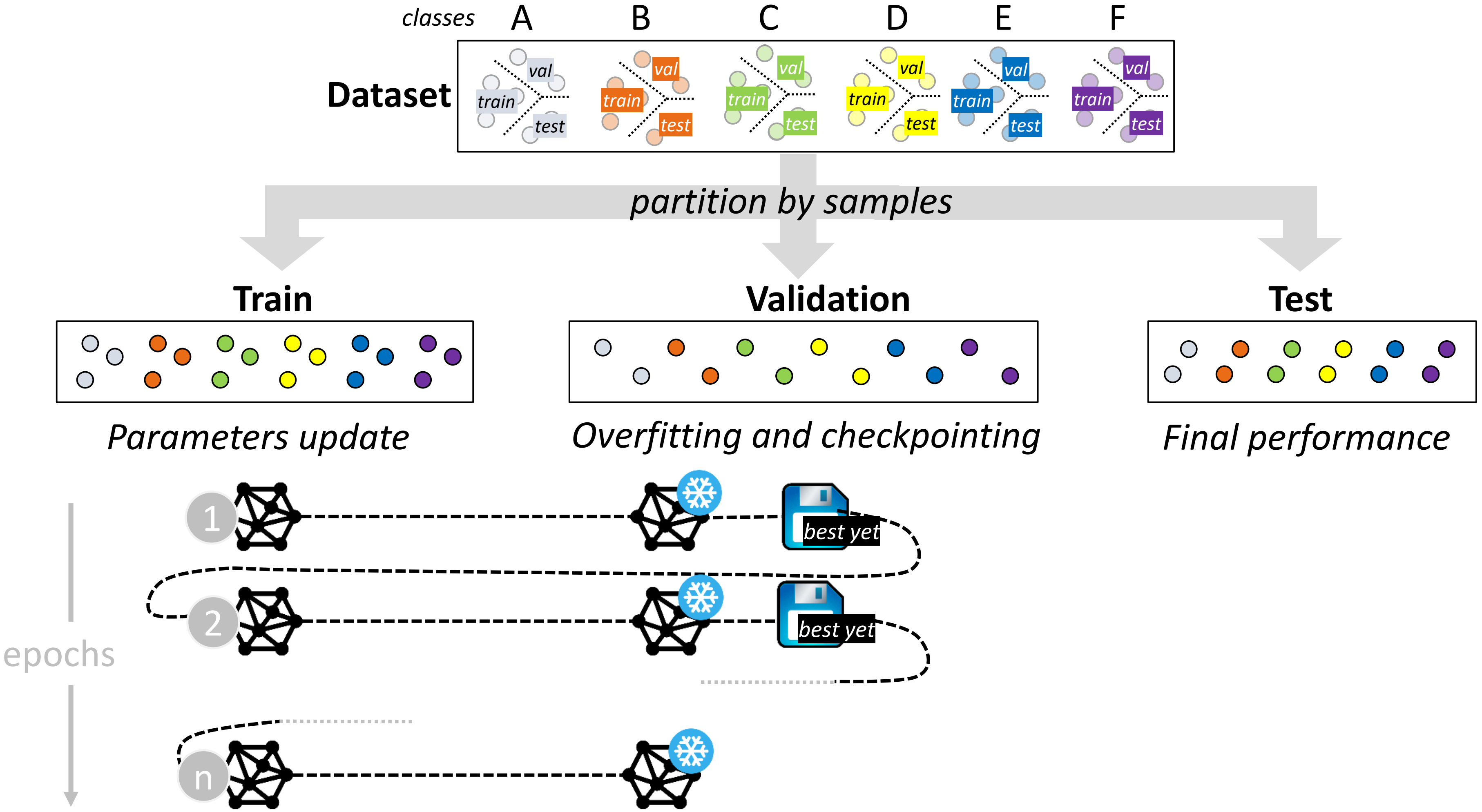
HUAWEI

Traditional model training

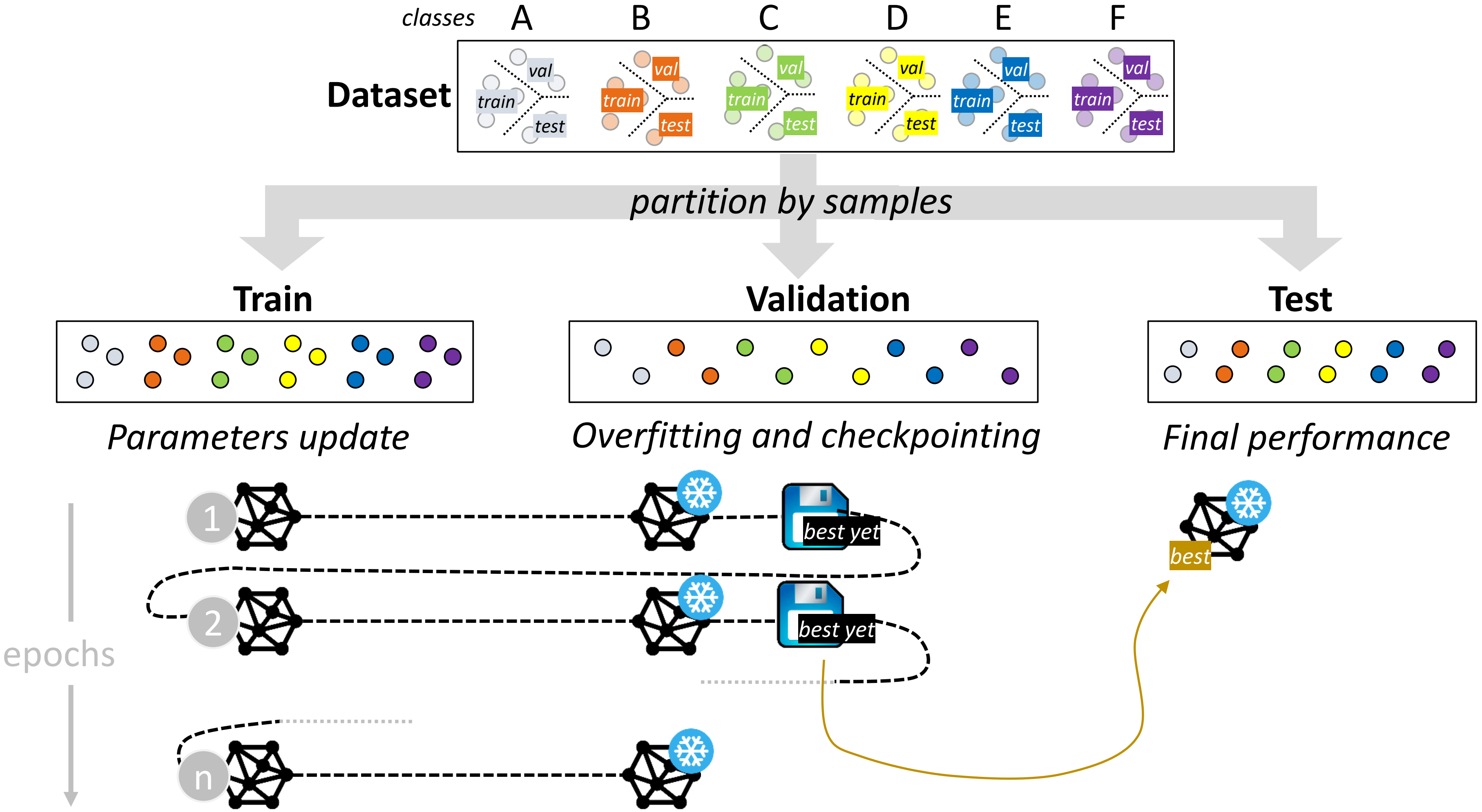


HUAWEI

Traditional model training



Traditional model training



HUAWEI

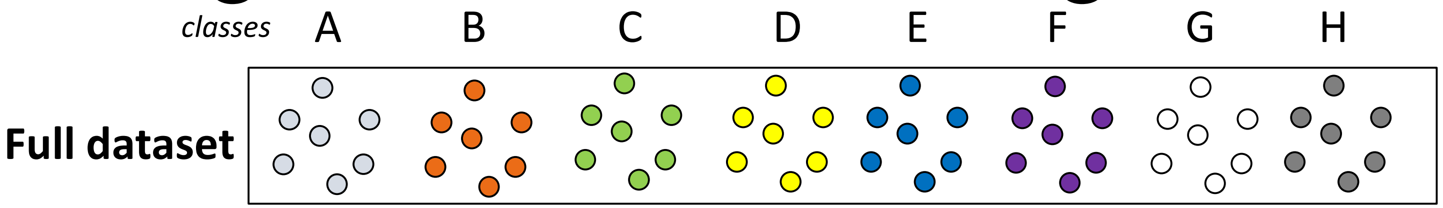
Traditional model training

Core principles

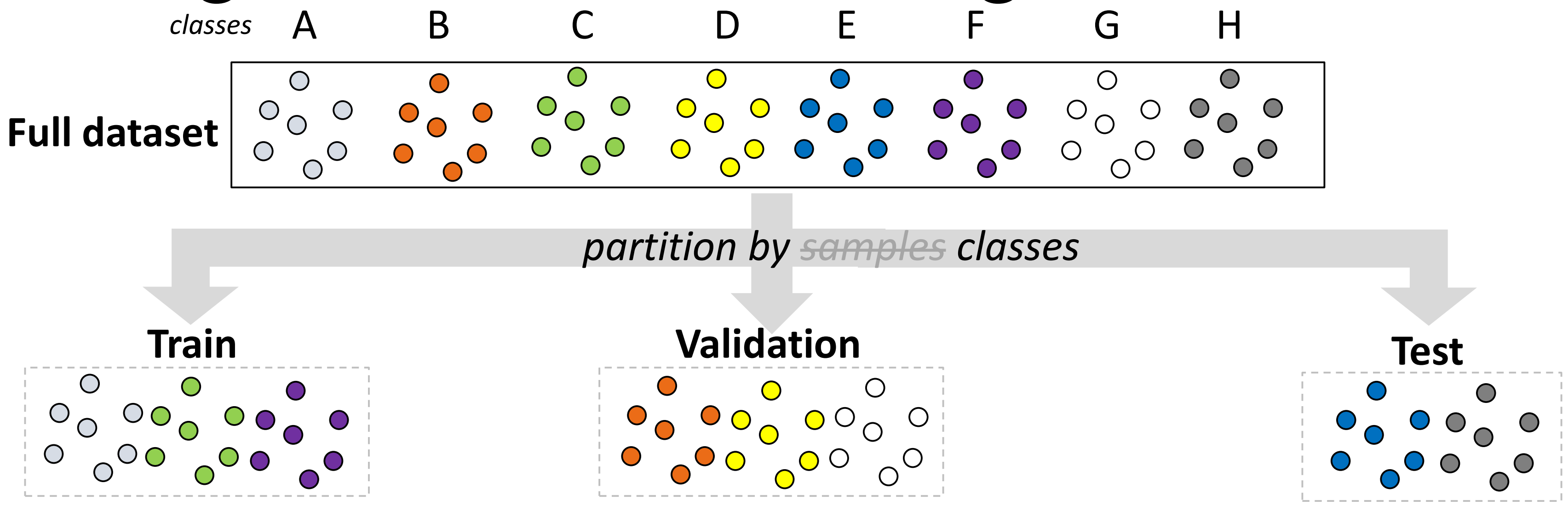
1. **One task**: *the model learns from subset of samples from ALL classes*
2. **Multiple scans**: *the model learns by iterating across ALL samples*

Meta-Learning violates both of those principles

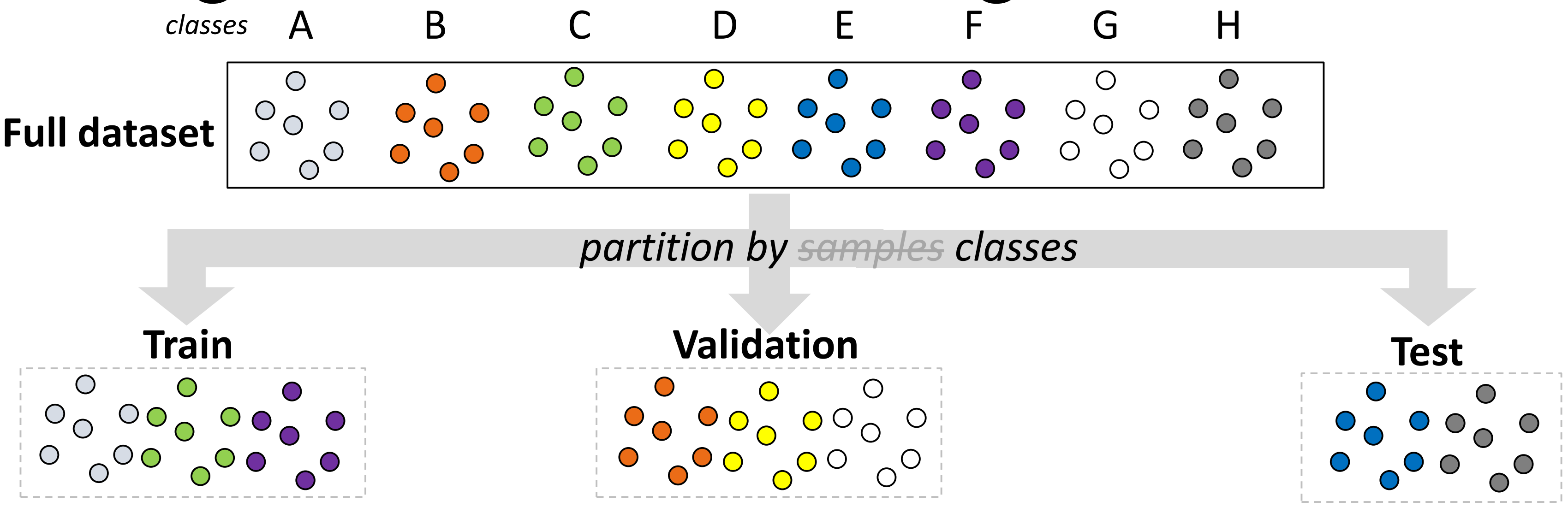
Model training via meta-learning



Model training via meta-learning

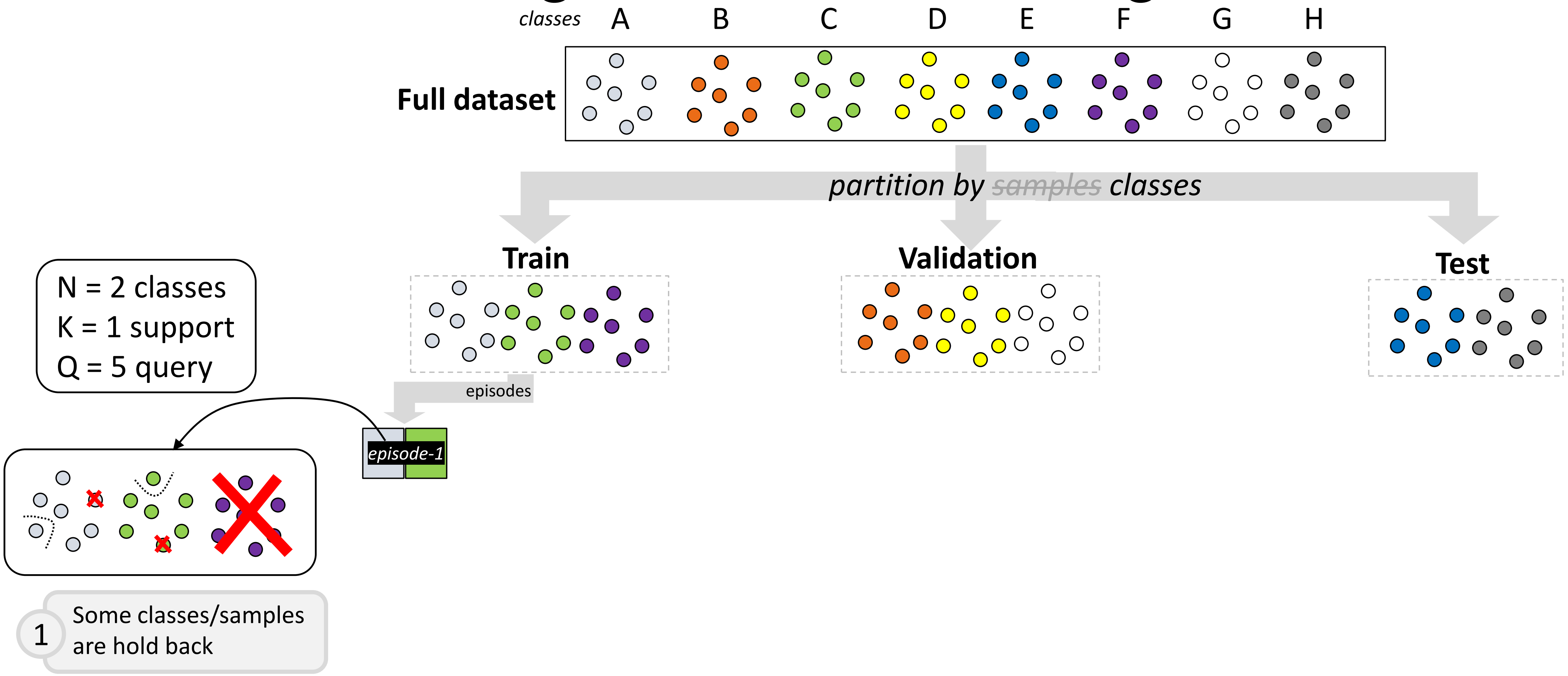


Model training via meta-learning

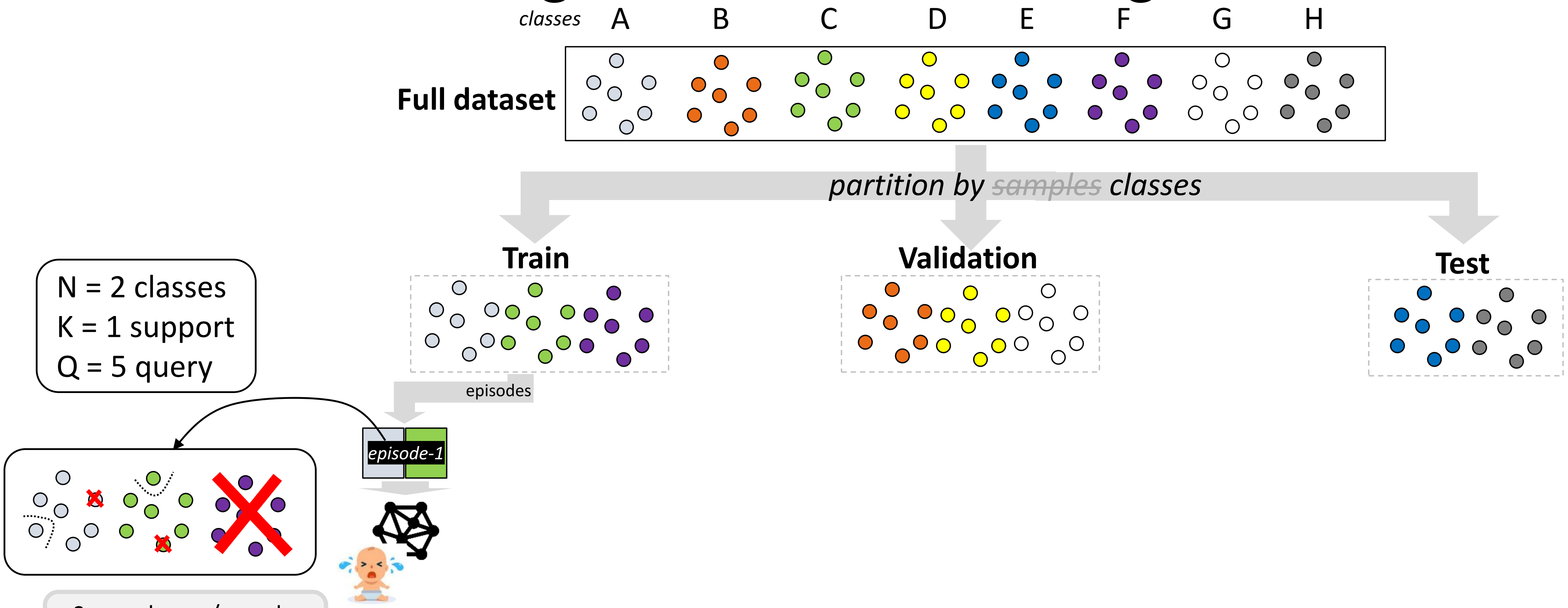


N = 2 classes
K = 1 support
Q = 5 query

Model training via meta-learning



Model training via meta-learning



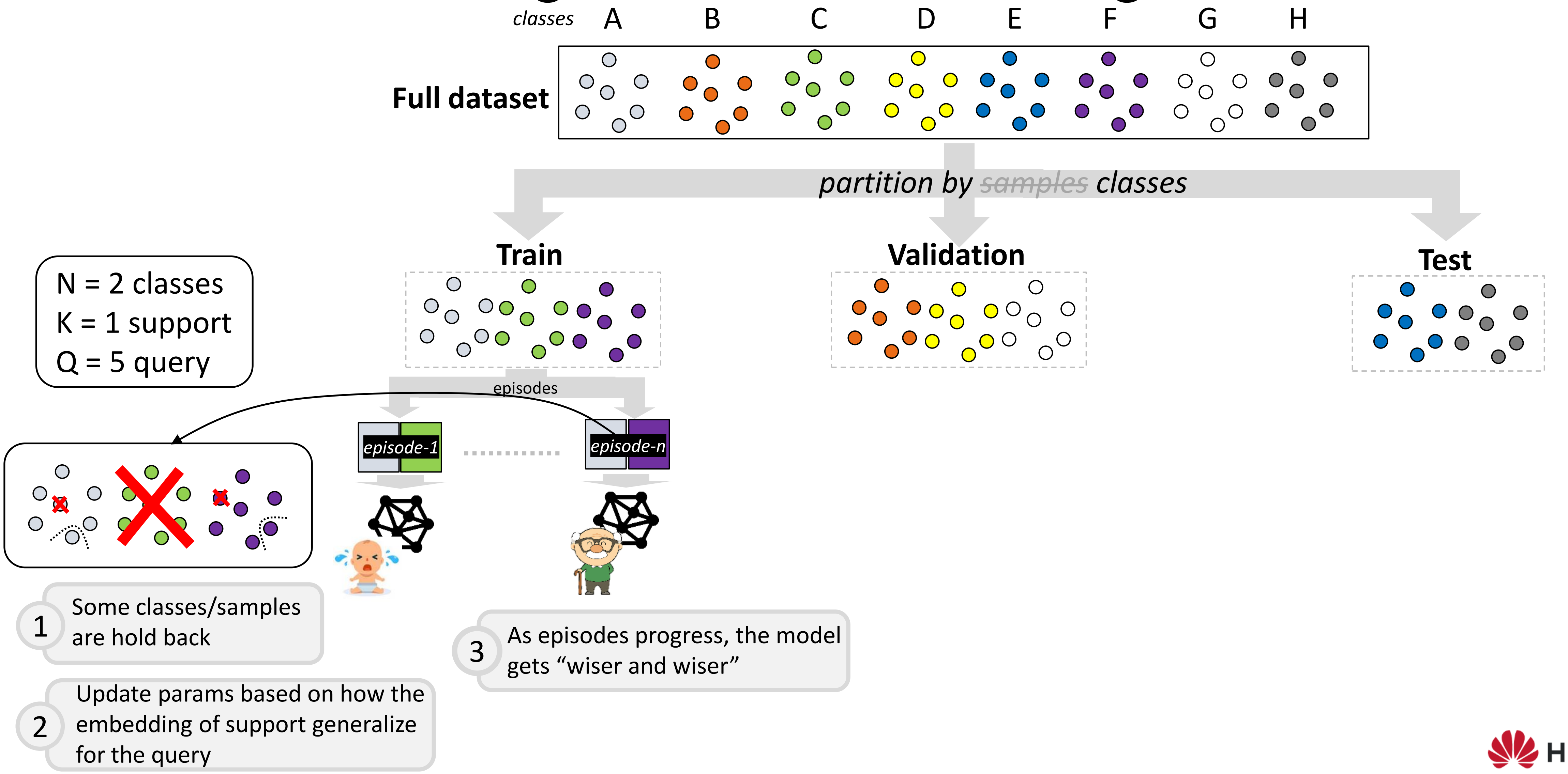
N = 2 classes
K = 1 support
Q = 5 query

1 Some classes/samples are hold back

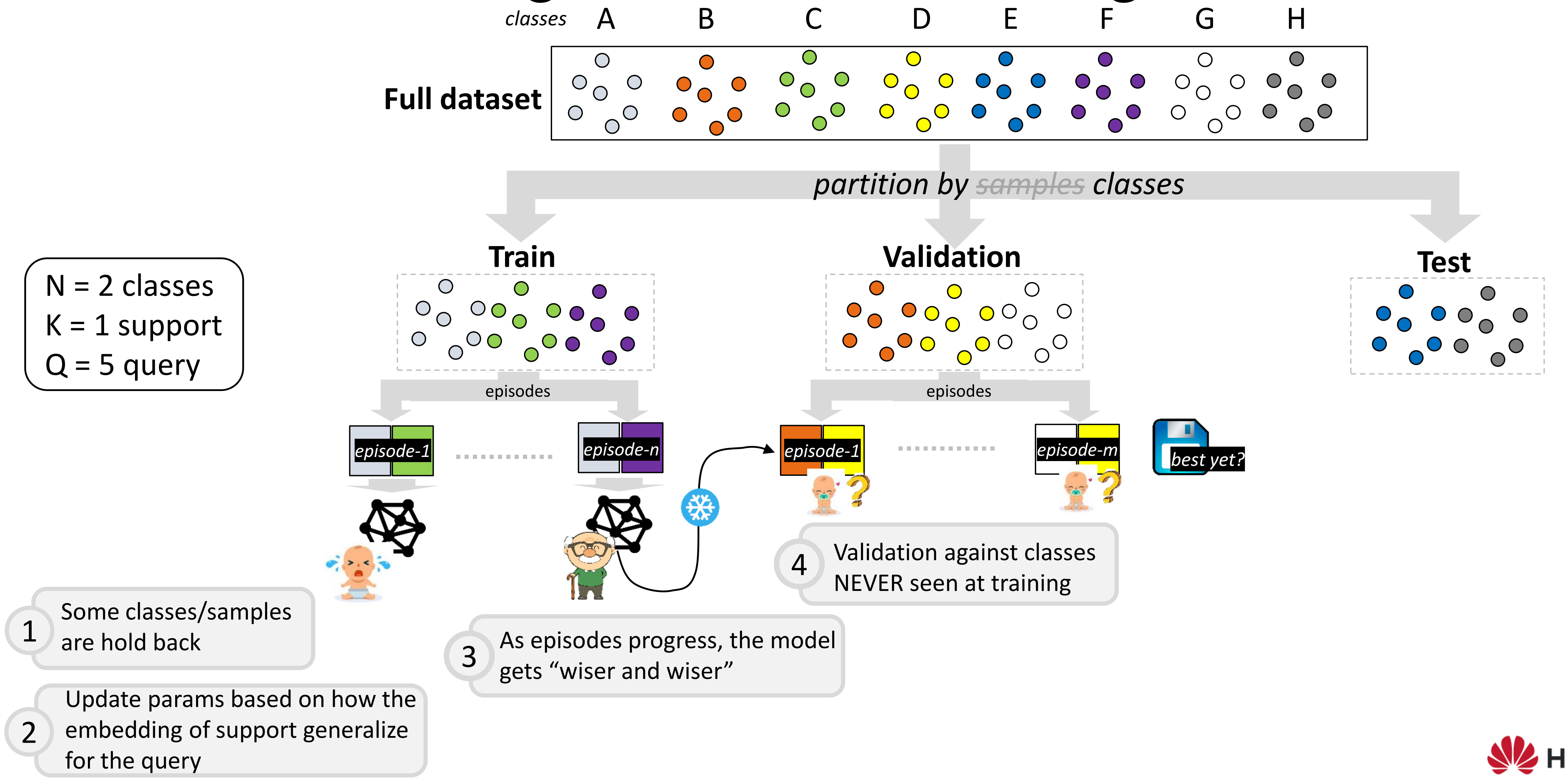
2 Update params based on how the embedding of support generalize for the query

HUAWEI

Model training via meta-learning



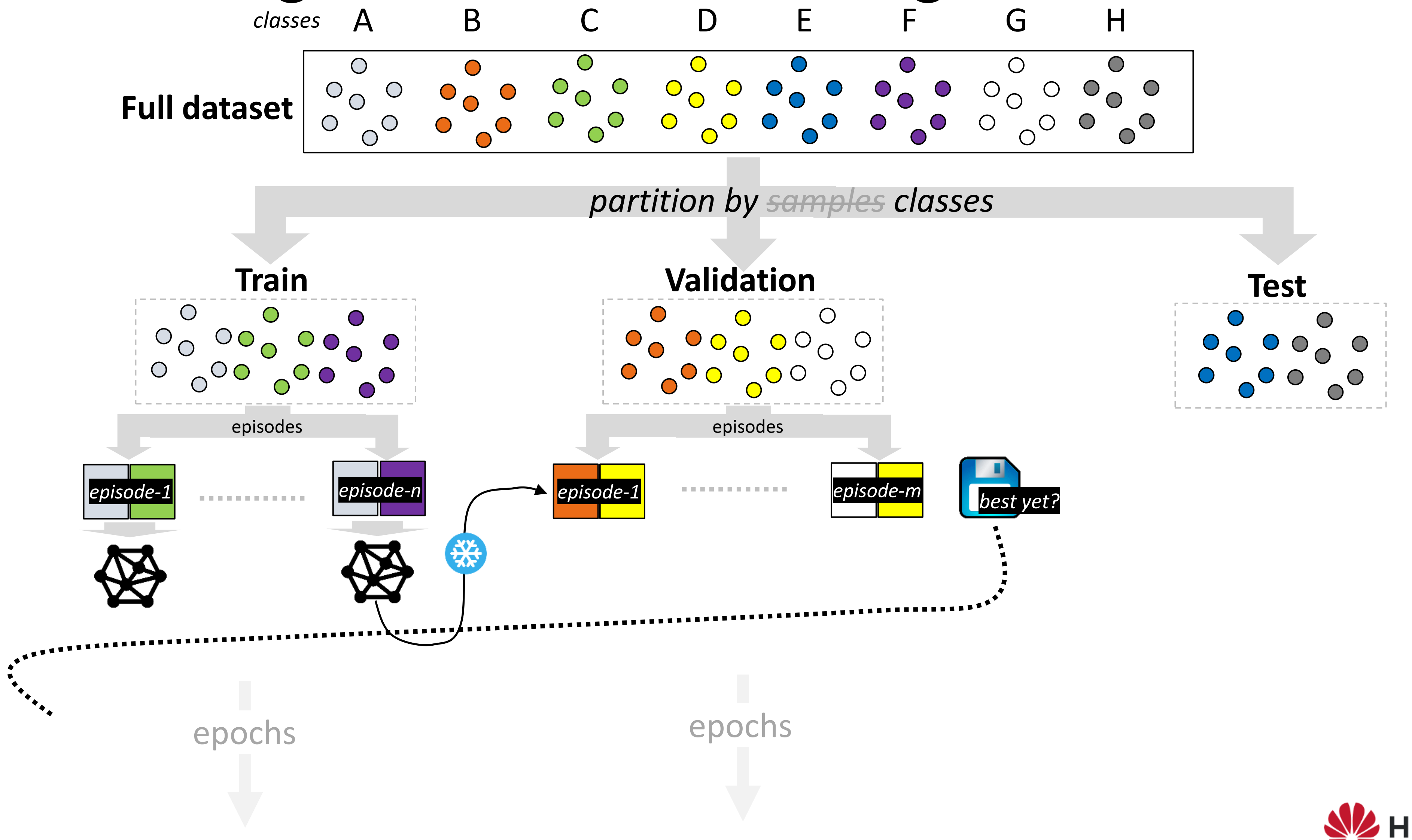
Model training via meta-learning



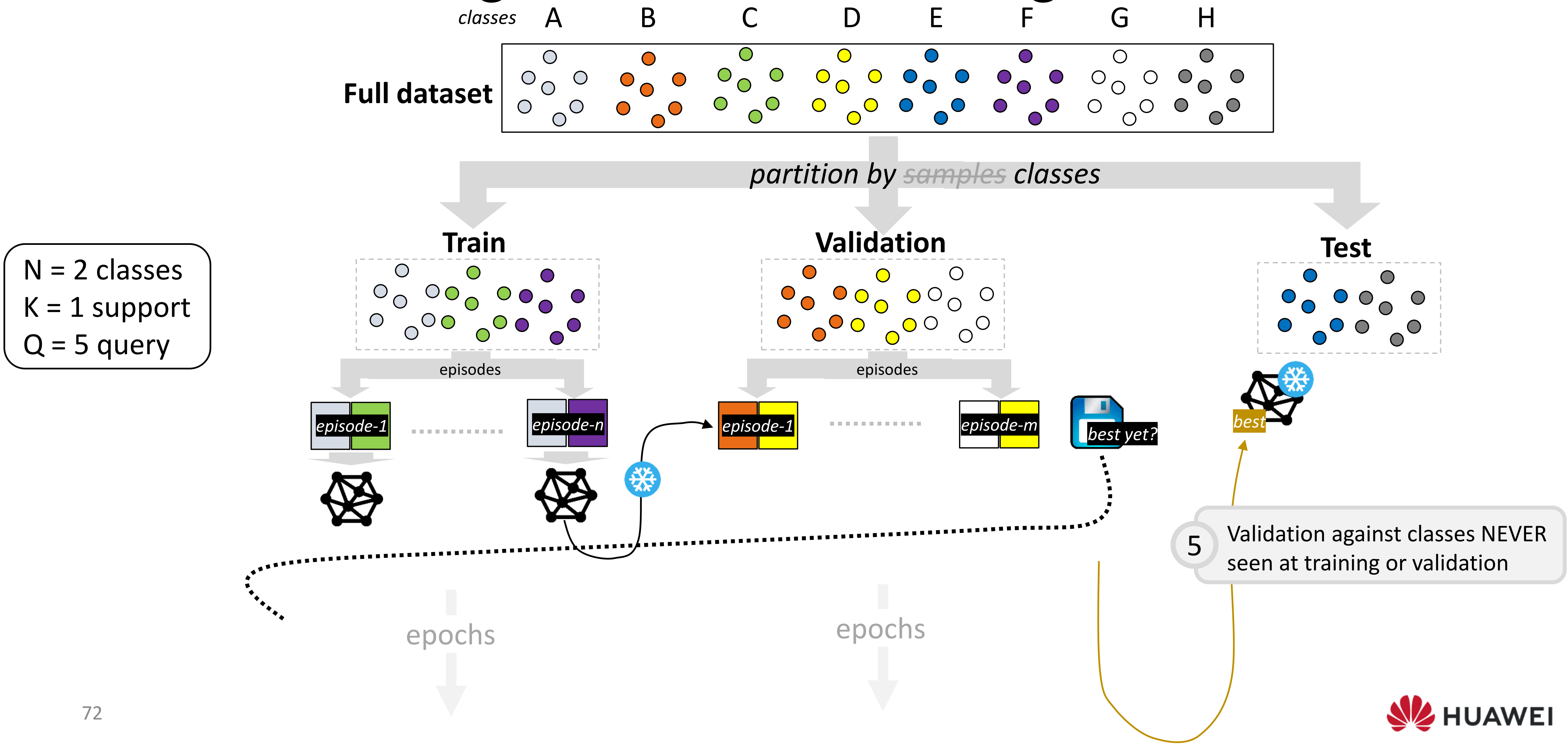
HUAWEI

Model training via meta-learning

N = 2 classes
K = 1 support
Q = 5 query



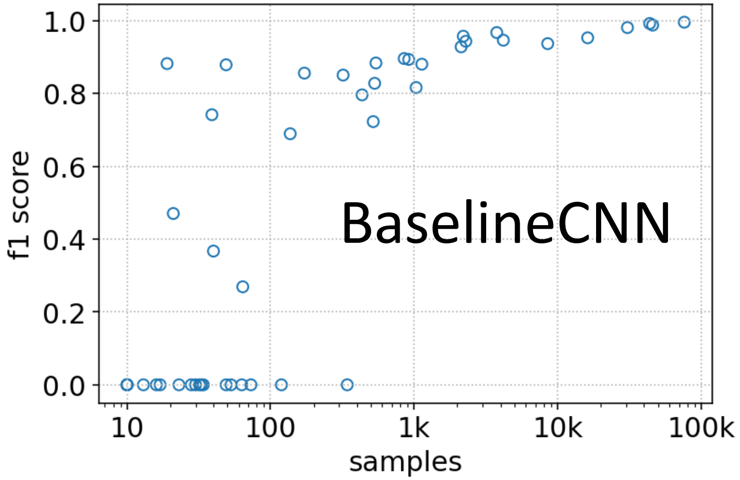
Model training via meta-learning

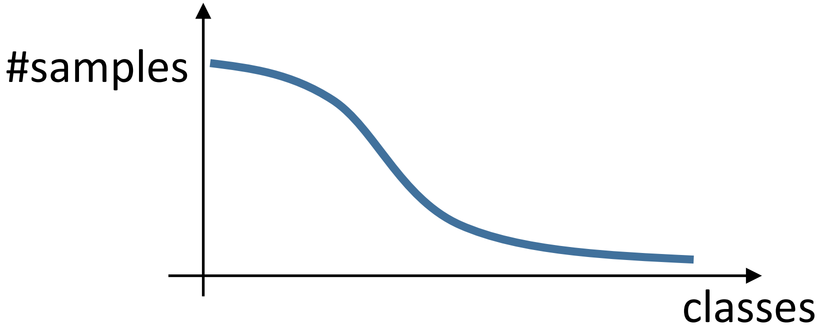


FSL in action (1/2)

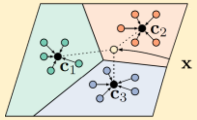
Monolithic approach

Model everything at once

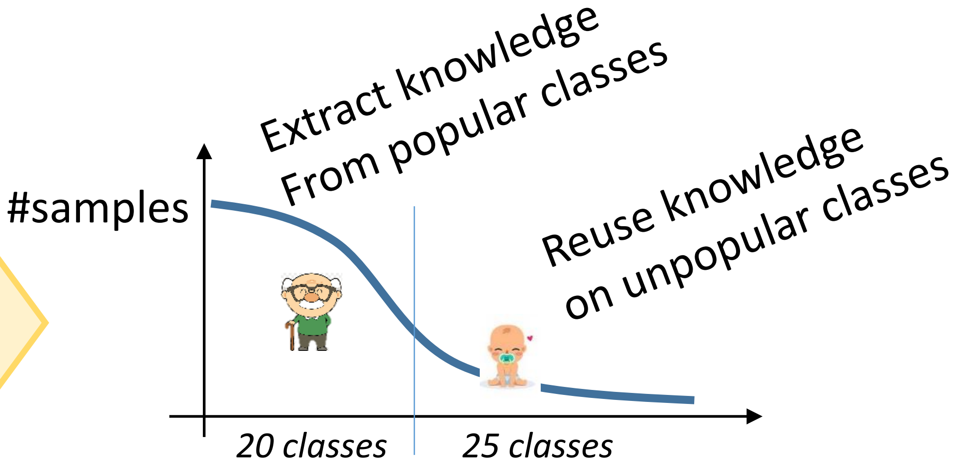




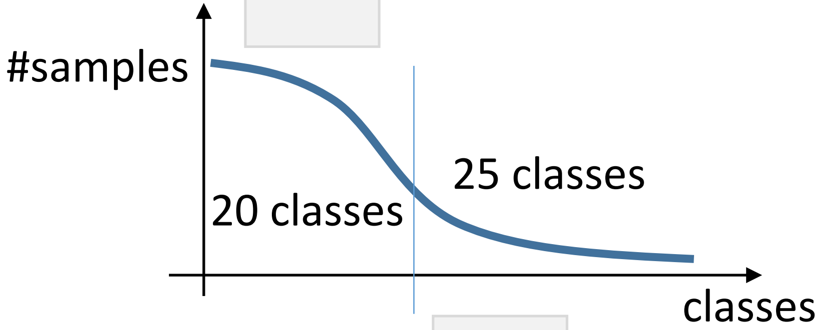
FSL Protonet (NIPS'17)



5-shots



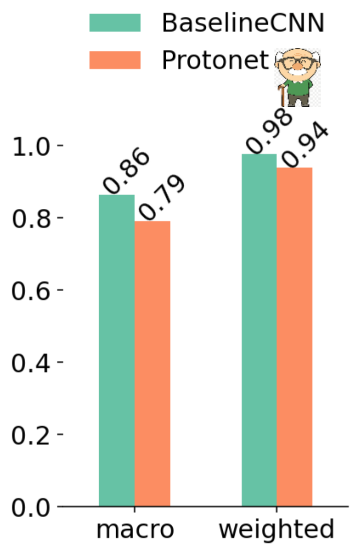
FSL in action (2/2)



Performance over popular classes

Performance over unpopular classes

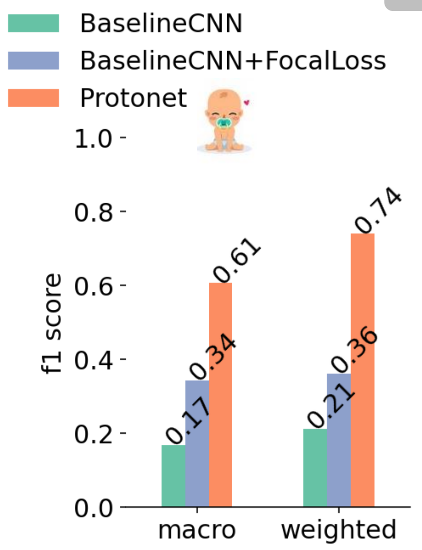
Top-20 classes



Not too bad performance considering we use 5-shots: good extrapolation power

Protonet model is 50x smaller than BaselineCNN

Bottom-25 classes



FSL significantly better than alternatives

…yet arguably far from being production ready

Conclusions

1 Many open challenges surrounding the data divide

Face them together as a community

2 Opportunities not fully explored offered by AI

Few-shot learning, Continuous learning, etc.

Datacom AI4NET Lab – Recent pointers

[TNSM-22] D.Rossi, L. Zhang
Landing AI on Networks: An equipment vendor viewpoint on Autonomous Driving Networks

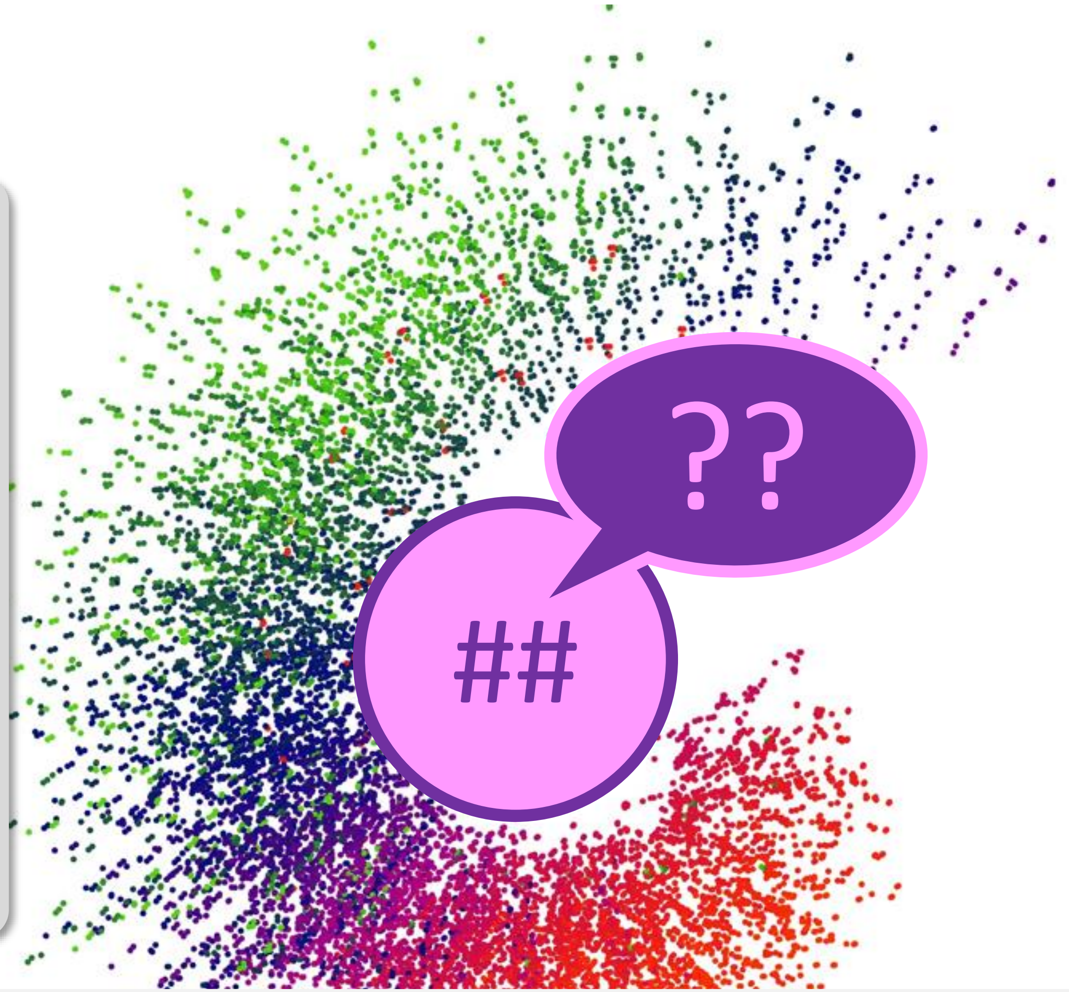
[INFOCOM-21] A. Finamore, J. Roberts, M. Gallo, D. Rossi
Accelerating Deep Learning Classification with Error-controlled Approximate-key Caching

[IEEE Network-21] L. Yang, D. Rossi
Quality monitoring and assessment of deployed Deep Learning models for Network AIOps

[ICML-UDL-21] L. Yang, D. Rossi,
Thinkback: Task Specific Out-of-Distribution Detection

[IEEE SEC-21] M. Gallo, A. Finamore, G. Simon, D. Rossi
FENXI: Fast In-Network Analytics

[TMA-21] G. Bovenzi, L. Lixuan, A. Finamore , G. Aceto, D. Ciuonzo, A. Pescape, D. Rossi
A First Look at Class Incremental Learning in Deep Learning Mobile Traffic



Alessandro Finamore
Principal engineer, AI4NET Datacom lab
Huawei Technologies Research Center Paris
alessandro.finamore@huawei.com
afinamore.io

ifip tc6 TMA CONFERENCE 2022

HUAWEI