How far are we from true AutoML

Winning solutions and results of *AutoDL challenge*

7th ICML AutoML Workshop on AutoML

Presented by Z. Liu in the name of
the AutoDL challenge team

The AutoDL challenge team

Original lead organizers:

- Olivier Bousquet (Google, Switzerland)
- André Elisseeff (Google, Switzerland)
- Isabelle Guyon (U. Paris-Saclay; UPSud/INRIA, France and ChaLearn, USA)
- Zhengying Liu (U. Paris-Saclay; UPSud, France)
- Wei-Wei Tu (4paradigm, China)

Other core team members:

- Sergio Escalera (U. Barcelona, Spain, and ChaLearn, USA)
- Julio Jacques Jr. (U. Barcelona, Spain)
- Meysam Madani (U. Barcelona, Spain)
- Adrien Pavao (U. Paris-Saclay; INRIA, France and ChaLearn, USA)
- Sebastien Treger (La Pallaisse, France, and ChaLearn, USA)
- Zhen Xu (Ecole Polytechnique and U. Paris-Saclay; INRIA, France)

Other contributors to the organization, starting kit, and datasets, include:

- Stephane Ayache (AMU, France)
- Hubert Jacob Banville (INRIA, France)
- Mahsa Behzadi (Google, Switzerland)
- Kristin Bennett (RPI, New York, USA)
- Hugo Jair Escalante (IANOE, Mexico and ChaLearn, USA)
- Gavin Cawley (U. East Anglia, UK)
- Baiyu Chen (UC Berkeley, USA)
- Albert Clapes i Sintes (U. Barcelona, Spain)
- Bram van Ginneken (Radboud U. Nijmegen, The Netherlands)
- Alexandre Gramfort (U. Paris-Saclay; INRIA, France)
- Yi-Qi Hu (4paradigm, China)
- Tatiana Merkulova (Google, Switzerland)
- Shangeth Rajaa (BITS Pilani, India)
- Herilalaina Rakotoarison (U. Paris-Saclay, INRIA, France)
- Mehreen Saeed (FAST Nat. U. Lahore, Pakistan)
- Marc Schoenauer (U. Paris-Saclay, INRIA, France)
- Michele Sebag (U. Paris-Saclay; CNRS, France)
- Danny Silver (Acadia University, Canada)
- Lisheng Sun (U. Paris-Saclay; UPSud, France)
- Fengfu Li (4paradigm, China)
- Lichuan Xiang (4paradigm, China)
- Jun Wan (Chinese Academy of Sciences, China)
- Mengshuo Wang (4paradigm, China)
- Jingsong Wang (4paradigm, China)
- Ju Xu (4paradigm, China)

The challenge is running on the **Codalab platform**, administered by **Université Paris-Saclay** and maintained by CKCollab LLC, with primary developers:

- Eric Carmichael (CKCollab, USA)
- Tyler Thomas (CKCollab, USA)

Sponsors:

CHA LEARN

Google

4Paradigm 第四范式

Microsoft

Home institutions:

université PARIS-SACLAY

Inria

UNIVERSITAT DE BARCELONA

Conferences:

NEURAL INFORMATION PROCESSING SYSTEMS

WAIC

ACML 2019

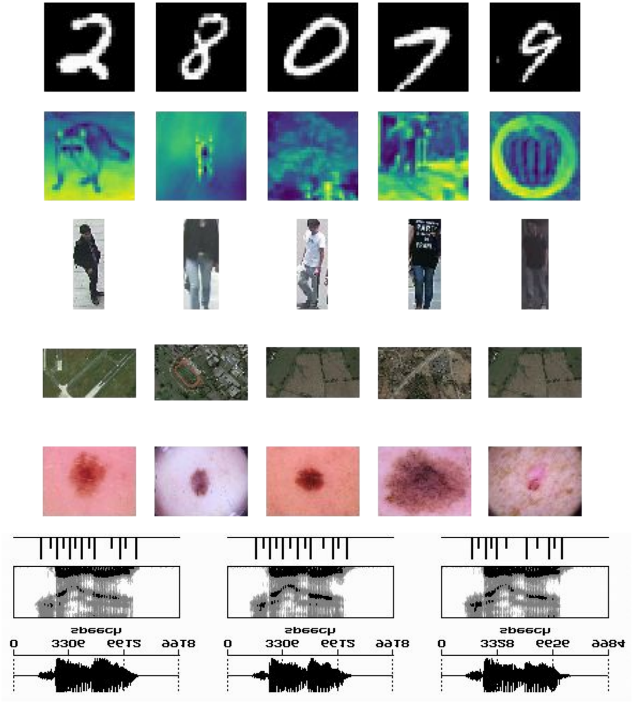
The International Joint Conference on Neural Networks

ECMLPKDD Würzburg 2019

AutoDL Challenge Design

(1) **Data:** diverse modalities/domains

We formatted >100 datasets, 66 or which ended up being used in challenges

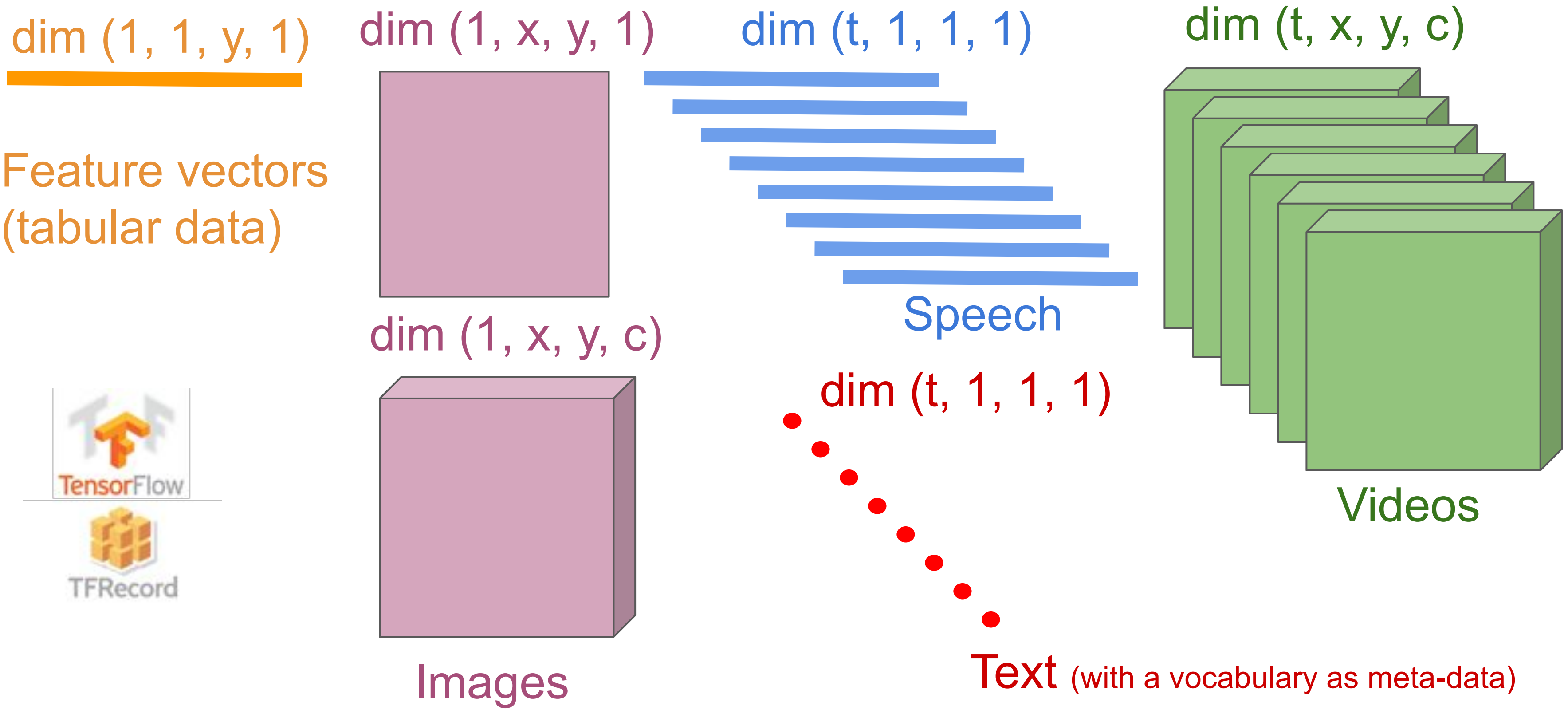


#	Dataset	Challenge(s)	Phase	Domain	Type	Class num.	Sample number train	Sample number test	Tensor dimension time	Tensor dimension row	Tensor dimension col	Tensor dimension chnl
1	Munster	AutoCV1	public	HWR	image	10	60000	10000	1	28	28	1
2	Chucky	AutoCV1	public	objects	image	100	48061	11939	1	32	32	3
3	Pedro	AutoCV1	public	people	image	26	80095	19905	1	var	var	3
4	Decal	AutoCV1	public	aerial	image	11	634	166	1	var	var	3
5	Hammer	AutoCV1	public	medical	image	7	8050	1965	1	600	450	3
6	Ukulele	AutoCV1	feedback	HWR	image	3	6979	1719	1	var	var	3
7	Caucase	AutoCV1	feedback	objects	image	257	24518	6089	1	var	var	3
8	Beatriz	AutoCV1	feedback	people	image	15	4406	1094	1	350	350	3
9	Saturn	AutoCV1	feedback	aerial	image	3	324000	81000	1	28	28	4
10	Hippocrate	AutoCV1	feedback	medical	image	2	175917	44108	1	96	96	3
11	Loukoum	AutoCV1 AutoCV2	final final	HWR	image	3	27938	6939	1	var	var	3
12	Tim	AutoCV1	final	objects	image	200	80000	20000	1	32	32	3
13	Apollon	AutoCV1 AutoCV2 **AutoDL**	final final feedback	people	image	100	6077	1514	1	var	var	3
14	Ideal	AutoCV1 AutoCV2	final feedback	aerial	image	45	25231	6269	1	256	256	3
15	Ray	AutoCV1 **AutoDL**	final final	medical	image	7	4492	1114	1	976	976	3
16	Kraut	AutoCV2	public	action	video	4	1528	863	var	120	160	1
17	Katze	AutoCV2	public	action	video	6	1528	863	var	120	160	1
18	Kreatur	AutoCV2	public	action	video	4	1528	863	var	60	80	1
19	Freddy	AutoCV2	feedback	HWR	image	2	546055	136371	1	var	var	3
20	Homer	AutoCV2	feedback	action	video	12	1354	353	var	var	var	3
21	Isaac2	AutoCV2	feedback	action	video	249	38372	9561	var	102	78	1
22	Formula	AutoCV2	feedback	misc.	video	4	32994	8203	var	80	80	3
23	Fiona	AutoCV2 **AutoDL**	final final	action	video	6	8038	1962	var	var	var	3
24	Monica1	AutoCV2 **AutoDL**	final feedback	action	video	20	10380	2565	var	168	168	3
25	Kitsune	AutoCV2	final	action	video	25	18602	4963	var	46	82	3
26	data01	AutoSpeech	public	speech	time	100	3000	3000	var	1	1	1
27	data02	AutoSpeech	public	speech	time	7	428	107	var	1	1	1
28	data03	AutoSpeech	public	speech	time	3	796	200	var	1	1	1
29	data04	AutoSpeech	public	speech	time	20	939	474	var	1	1	1
30	data05	AutoSpeech	public	speech	time	10	199	597	var	1	1	1
31	data11	AutoSpeech	feedback	speech	time	55	1300	2000	var	1	1	1
32	data12	AutoSpeech	feedback	speech	time	5	3120	346	var	1	1	1
33	data13	AutoSpeech	feedback	speech	time	3	5000	1330	var	1	1	1
34	data14	AutoSpeech	feedback	speech	time	8	767	191	var	1	1	1
35	data15	AutoSpeech	feedback	speech	time	76	2286	571	var	1	1	1
36	data21	AutoSpeech	final	speech	time	50	800	1200	var	1	1	1
37	data22	AutoSpeech	final	speech	time	4	2649	294	var	1	1	1
38	data23 Oreal	AutoSpeech **AutoDL**	final final	speech	time	3	2000	264	var	1	1	1
39	data24	AutoSpeech	final	speech	time	16	384	386	var	1	1	1
40	data25 Sahak	AutoSpeech **AutoDL**	final feedback	speech	time	100	3008	752	var	1	1	1
41	O1	AutoNLP	public	english	text	2	7792	1821	var	1	1	1
42	O2	AutoNLP	public	english	text	20	11314	7532	var	1	1	1
43	O3	AutoNLP	public	english	text	2	60000	40000	var	1	1	1
44	O4	AutoNLP	public	chinese	text	10	55000	10000	var	1	1	1
45	O5	AutoNLP	public	chinese	text	18	156000	72000	var	1	1	1
46	PU1	AutoNLP	feedback	english	text	9	2822	499	var	1	1	1
47	PU2	AutoNLP	feedback	english	text	5	132651	23409	var	1	1	1
48	PU3	AutoNLP	feedback	chinese	text	2	1110203	195919	var	1	1	1
49	PU4	AutoNLP	feedback	chinese	text	11	100000	50000	var	1	1	1
50	PU5	AutoNLP	feedback	chinese	text	31	600000	400000	var	1	1	1
51	PR1	AutoNLP	final	english	text	20	33807	5967	var	1	1	1
52	PR2 Tanak	AutoNLP **AutoDL**	final feedback	english	text	2	42500	7501	var	1	1	1
53	PR3	AutoNLP	final	english	text	4	90000	30000	var	1	1	1
54	PR4	AutoNLP	final	chinese	text	11	100000	50000	var	1	1	1
55	PR5 Tal	AutoNLP **AutoDL**	final final	chinese	text	15	250000	132688	var	1	1	1
56	Adult	**AutoDL**	public	categorical	tabular	5	39074	9768	1	1	24	1
57	Dilbert	**AutoDL**	public	objects	tabular	5	14860	9720	1	1	2000	1
58	Digits	**AutoDL**	public	HWR	tabular	10	35000	35000	1	1	1568	1
59	Madeline	**AutoDL**	public	artificial	tabular	2	4220	3240	1	1	259	1
60	Barak	**AutoDL**	feedback	CE pair	tabular	4	21869	2430	1	1	270	1
61	Bilal	**AutoDL**	final	audio	tabular	20	10931	2733	1	1	400	1

- IMAGE
- VIDEO
- SPEECH
- TEXT
- TABULAR
- Multi-label tasks

Liu Z, Xu Z, Rajaa S, Madadi M. Towards Automated Deep Learning: Analysis of the AutoDL challenge series 2019. To appear in *NeurIPSCD2019* in Proceedings of Machine Learning Research (PMLR) 2019:10.

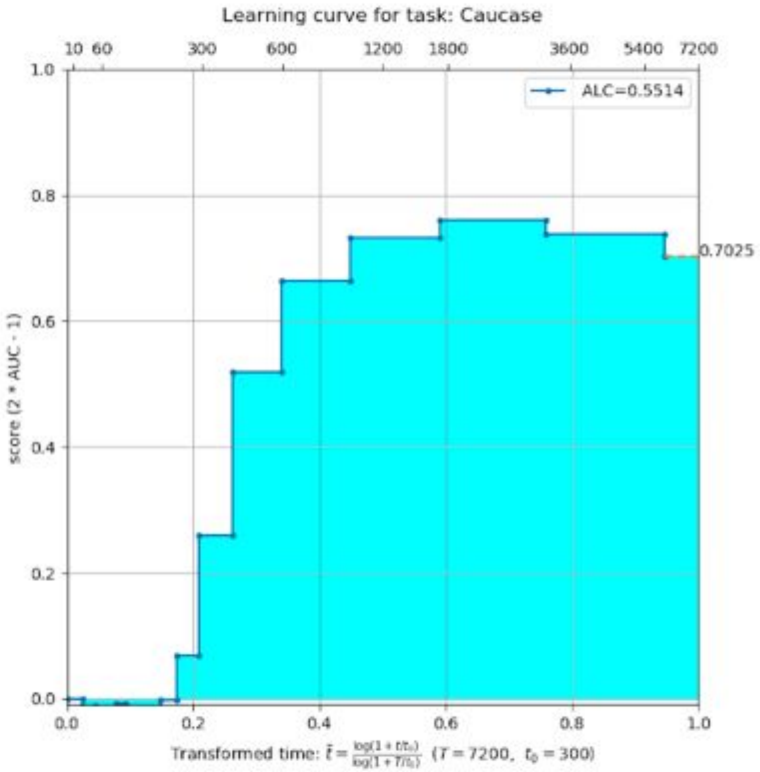
(1) **Data:** RAW data in a Tensor Format



(2) **Evaluation:** Fixed max T + Any-time Learning

Time rescaling: $$\tilde{t}(t) = \frac{\log(1+t/t_0)}{\log(1+T/t_0)}$$

$$\begin{aligned} ALC &= \int_0^1 s(t) d\tilde{t}(t) \\ &= \int_0^T s(t)\tilde{t}'(t)dt \\ &= \frac{1}{\log(1+T/t_0)} \int_0^T \frac{s(t)}{t+t_0} dt \end{aligned}$$



(2) **Evaluation:** Blind testing

Two phases:

- **Feed-back phase**: 5 datasets, 5 submissions/day for 3-4 months.
- **Final test phase**: 10 OTHER unseen datasets, **ONE** single submission.

BOTH phases, code TRAINED and TESTED on the platform => **datasets invisible**.

Additional **“public” datasets** => **META-LEARNING**.

(3) **Starting Kit** and **Baselines**

Baseline 0: **Constant** predictions (for debug purposes)

domain-agnostic

Baseline 1: **Linear model**

Baseline 2: **Multi-dimensional CNN** (auto-rescaling)

domain-dependent

Baseline 3: Combination of winning solutions from previous challenges:

Image & Video: Kakaobrain, ResNet (He et al, 2016) and Fast Auto Augment (Cubuk et al. (2018); Lim et al. (2019))

Speech: PASA_NJU, Spectral transform, logistic reg., lightGBM, CNN, ResNet, VggVox, LSTM, etc.

Text: Upwind_flys, LinearSVC, LSTM, BERT, etc., selected with meta-controller

Tabular (new): Fully connected network.

Challenge Design Recap

(1) Raw data from **5 domains**: Image, Video, Speech, Text, Tabular.

(2) Fixed time budget. Any-time learning (ALC metric). **Blind testing**.

(3) Starting kit, sample "public" data and **baselines** provided.

(4) Fixed computational resources.

(5) Using **Deep Learning** was **NOT imposed.**

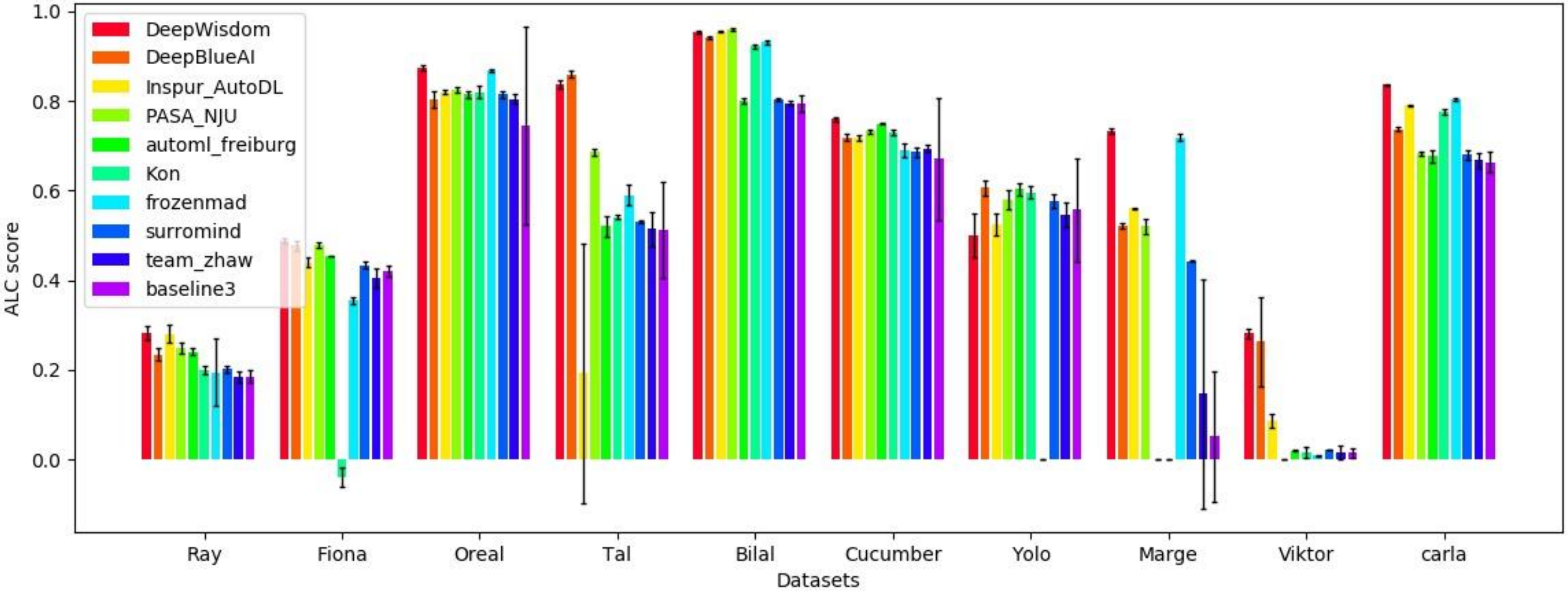
Challenge Results

WINNERS

(code open-sourced at autodl.chalearn.org)

Challenge	1st Prize ($2000)	2rd Prize ($1500)	3nd Prize ($500)
AutoCV	**kakaobrain** (Kakao Brain)	**DKKimHCLee** (Hana. Tech. Inst.)	**base_1** (Hanyang University)
AutoCV2	**kakaobrain** (Kakao Brain)	**tanglang** (Xiamen University)	**kvr** (-)
AutoNLP	**DeepBlueAI** (DeepBlue Technology)	**upwind_flys** (Lenovo)	**txta** (gsdata.cn)
AutoSpeech	**PASA_NJU** (Nanjing University)	**DeepWisdom** (fuzhi.ai)	**Kon** (NS Solutions Corporation)
AutoDL	**DeepWisdom** (fuzhi.ai)	**DeepBlueAI** (DeepBlue Technology)	**Inspur_AutoDL & PASA_NJU**

AutoDL final phase results



Did we get answers to our questions?

(1) Unified approach for ALL 5 domains?
(Image, Video, Speech, Text, Tabular.)

(2) Time budget sufficient? **Any-time learning** possible?

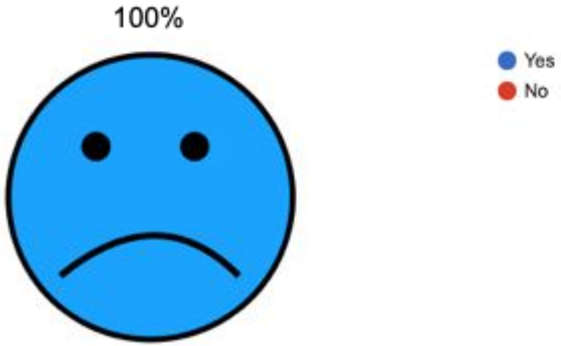
(3) Was sample **"public" data** used for **meta-learning**?

(1) Unified approach for ALL 5 domains?

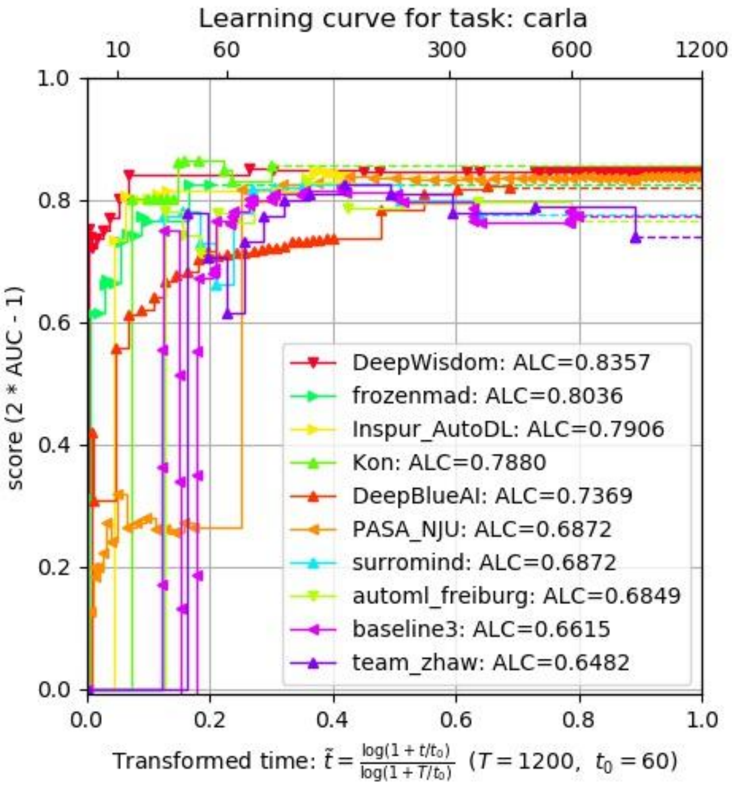
Participant survey:

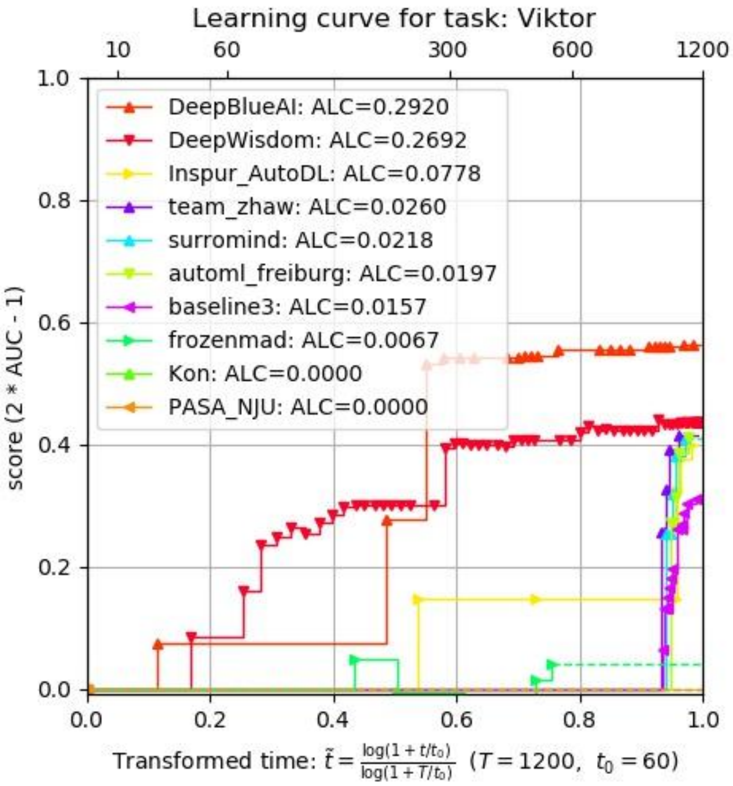
[DOMAIN-DEPENDENCY] Do you use a separate approach for each dataset/domain (rather than a single unified approach)

10 responses



(2) Time budget sufficient? Any-time learning possible?

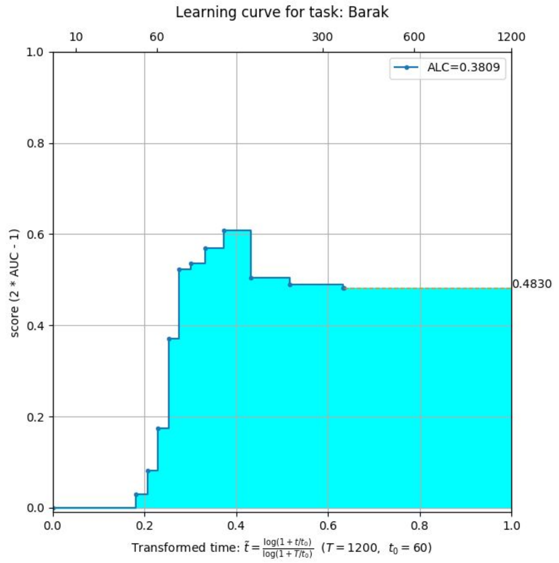




Impact of t0

$$\tilde{t}(t) = \frac{\log(1 + t/t_0)}{\log(1 + T/t_0)}$$

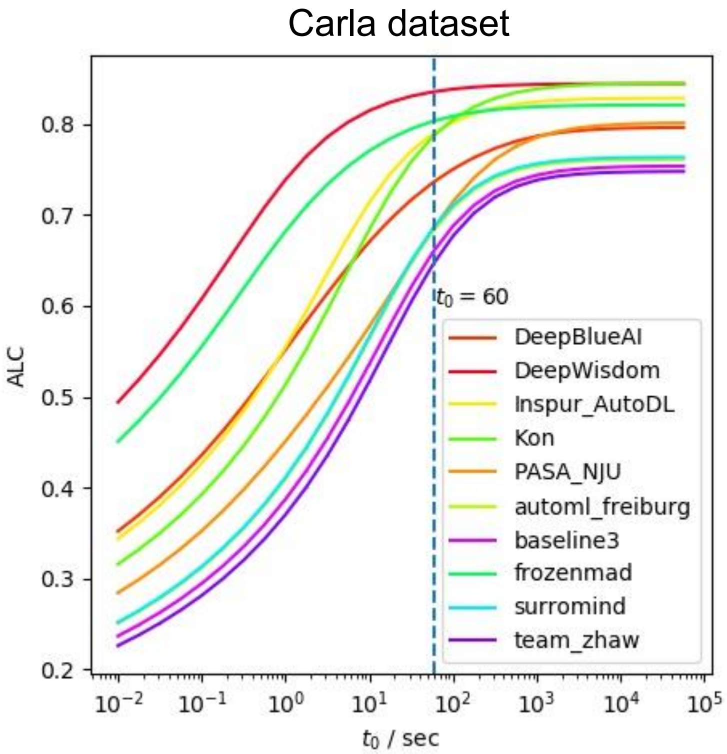
$$\begin{aligned} ALC &= \int_0^1 s(t) d\tilde{t}(t) \\ &= \int_0^T s(t)\tilde{t}'(t)dt \\ &= \frac{1}{\log(1 + T/t_0)} \int_0^T \frac{s(t)}{t + t_0} dt \end{aligned}$$



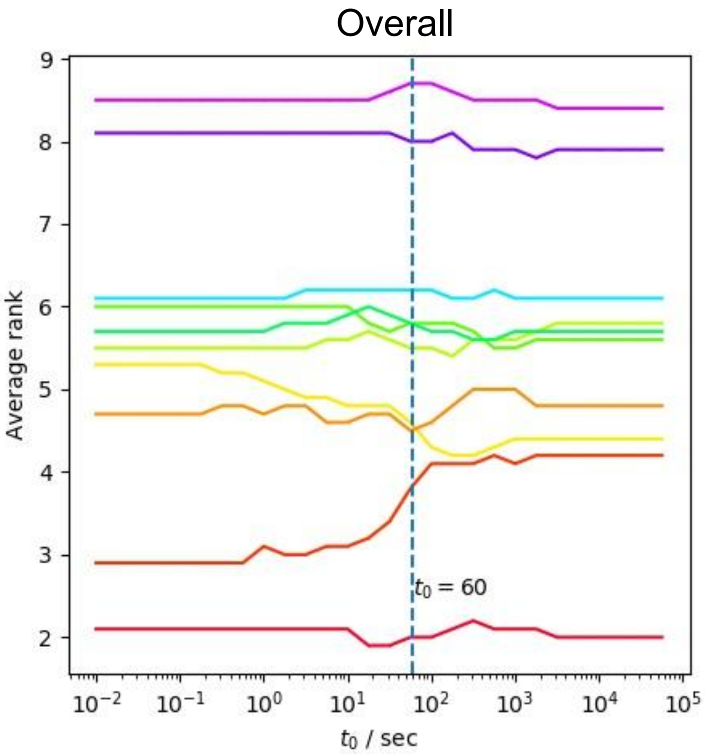
$$\lim_{t_0 \to 0^+} ALC(t_0) = s(0)$$

$$\lim_{t_0 \to +\infty} ALC(t_0) = \frac{1}{T} \int_0^T s(t)dt$$

Impact of t0



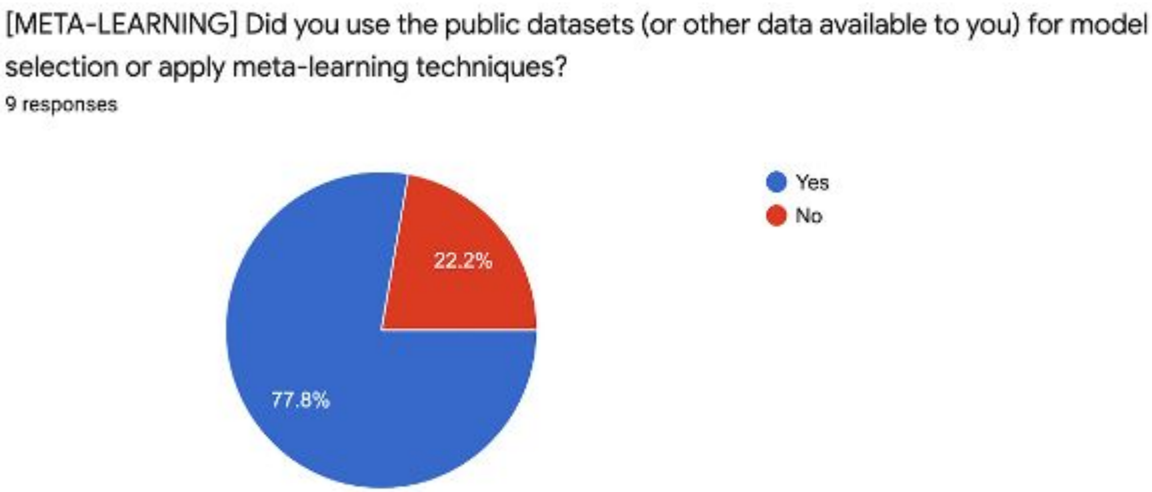
Carla dataset
ALC
t_0 / sec
$t_0 = 60$
DeepBlueAI
DeepWisdom
Inspur_AutoDL
Kon
PASA_NJU
automl_freiburg
baseline3
frozenmad
surromind
team_zhaw



Overall
Average rank
t_0 / sec
$t_0 = 60$

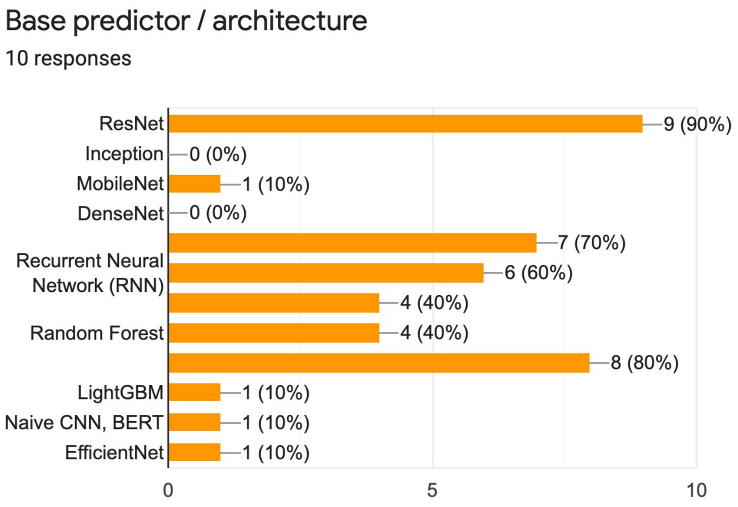
(3) Was **“public” data** used for **meta-learning**?

Participant survey:



Winning solutions

Neural architectures used in the winning approaches



Architecture name	# Parameters	Domains	Teams
ResNet-18, ResNet-9 (He et al 2015)	11.4M, 5.7M	image, video	Kakaobrain, DeepWisdom, automl_freiburg
MC3 (Du Tran et al CVPR 2018)	**32.8M**	video	DeepWisdom
EfficientNet-(b0, b1, b2) (M. Tan and Q. Le. 2019)	5.3M, 7.8M, 9.2M	image, video	DeepWisdom, automl_freiburg
MobileNetV2 (M. Sandler et al 2019)	3.4M	image, video	team_zhaw, DeepBlueAI
TextCNN	variable	text	Upwind_flys, DeepWisdom
Fast RCNN (Ross Girshick)		text	DeepWisdom
LSTM, BiLSTM (Hochreiter, Schmidhuber 1997)	0.2M-1M	text, speech	frozenmad, PASA_NJU
GRU, BiGRU, (Kyunghyun Cho et al 2014) GRU with Attention	0.1M-1M	text, speech	DeepBlueAI, DeepWisdom
BERT-like (Tiny-BERT(X.Jiao et al))	<110M	text	frozenmad, upwind_flys
DNN	<1M	tabular	DeepWisdom

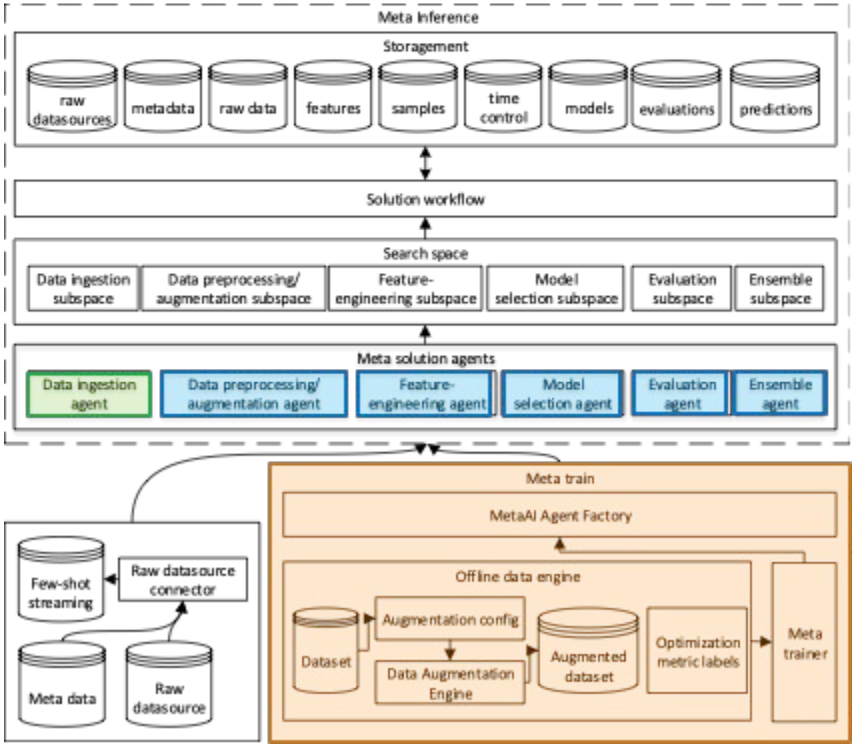
AutoML techniques vs domains

Approach	image	video	speech	text	tabular
Meta-learning	Offline meta-training transferred with AutoFolio [25] based on meta-features (*automl freiburg*) Offline meta-training generating solution agents, searching for optimal sub-operators in predefined sub-spaces, based on dataset meta-data. (*DeepWisdom*) MAML-like method [17] (*team zhaw*)				
Preprocessing	image cropping and data augmentation (*PASANJU*), fast autoaugment (*DeepBlueAI*)	Sub-sampling keeping 1/6 frames and adaptive image size (*DeepBlueAI*) Adaptive image size	MFCC, Mel Spectrogram, STFT	root features extractions with stemmer, meaningless words filtering (*DeepBlueAI*)	Numerical and Categorical data detection and encoding
Hyperparameter Optimization	Offline with BOHB [26] (Bayesian Optimization and Multi-armed Bandit) (*automl freiburg*) Sequential Model-Based Optimization for General Algorithm Configuration (SMAC) (*automl freiburg*)				Baysien Optimization (*PASANJU*) HyperOpt [27] (*Inspur AutoDL*)
Transfer learning	Pre-trained on ImageNet [28] (all teams except *Kon*)	Pre-trained on ImageNet [28] (all top-8 teams except *Kon*) MC3 model pretrained on Kinetics (*DeepWisdom*)	ThinResnet34 pre-trained on VoxCeleb2 (*DeepWisdom*)	FastText pre-trained on Common Crawl (*frozenmad*)	
Ensemble learning	Adaptive Ensemble Learning (ensemble latest 2 to 5 predictions) (*DeepBlueAI*)	Ensemble Selection [29] (top 5 validation predictions are fused) (*DeepBlueAI*); Ensemble models sampling 3, 10, 12 frames (*DeepBlueA*)	last best predictions ensemble strategy (*DeepWisdom*) averaging 5 best overall and best of each model: LR, CNN, CNN+GRU (*DeepBlueA*)	Weighted Ensemble over 20 best models [29] (*DeepWisdom*)	LightGBM ensemble with bagging method [30] (*DeepBlueAI*), Stacking and blending (*DeepWisdom*)

Teams vs domains

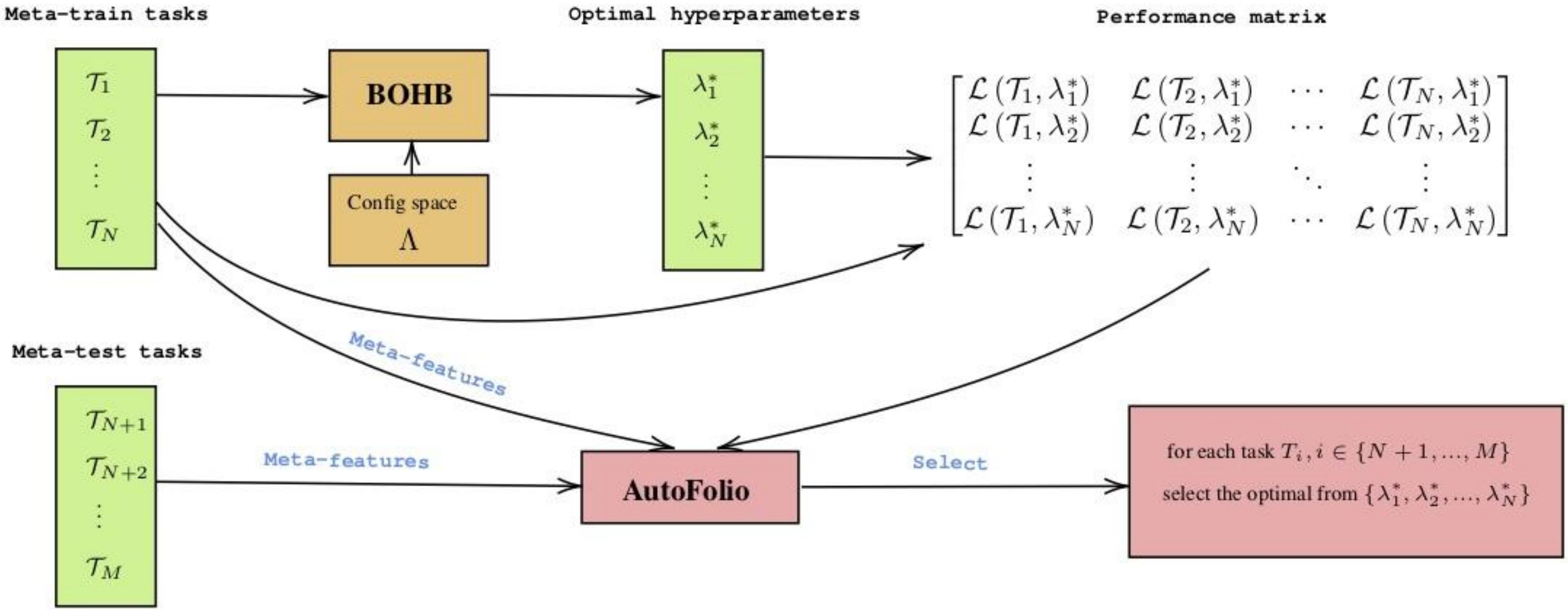
Team	image	video	speech	text	tabular
DeepWisdom	[**ResNet-18** and ResNet-9 models] [pretrained on ImageNet]	[MC3 model] [pretrained on Kinetics]	[fewshot learning] [LR, ThinRestnet34 models] [pretrained on VoxCeleb2]	[fewshot learning] [task difficulty and similarity evaluation for model selection] [SVM, **TextCNN**,[fewshot learning] RCNN, GRU, GRU with Attention]	[**LightGBM**, Xgboost, Catboost, DNN models] [no pretrained]
DeepBlueAI	[data augmentation with Fast AutoAugment] [**ResNet-18** model]	[subsampling keeping 1/6 frames] [Fusion of 2 best models]	[iterative data loader (7, 28, 66, 90%)] [MFCC and Mel Spectrogram preprocessing] [LR, CNN, CNN+GRU models]	[Samples truncation and meaningless words filtering] [Fasttext, **TextCNN**, BiGRU models] [Ensemble with restrictive linear model]	[3 **LightGBM** models] [Ensemble with Bagging]
PASA NJU	**ResNet-18** and SeResnext50; preprocessing: shape standardization and image flip (data augmentation)	**ResNet-18** and SeResnext50; preprocessing: shape standardization and image flip (data augmentation)	[data truncation(2.5s to 22.5s)][LSTM, **VggVox** ResNet with pretrained weights of DeepWis- dom(AutoSpeech2019) ThinRestnet34?]	[data truncation(300 to 1600 words)][TF-IDF and word embedding]	[iterative data loading] [Non Neural Nets models] [models complexity increasing over time] [Baysien Optimization of hyperparameters]
frozenmad	[images resized under 128x128] [progressive data loading increasing over time and epochs] [**ResNet-18** model] [pretrained on ImageNet]	[Successive frames difference as input of the model] [pretrained **ResNet-18** with RNN models]	[progressive data loading in 3 steps 0.01, 0.4, 0.7] [time length adjustment with repeating and clipping] [STFT and Mel Spectrogram preprocessing] [LR, LightGBM, **VggVox** models]	[TF-IDF and BERT tokenizers] [SVM, RandomForest , CNN, tinyBERT]	[progressive data loading] [no preprocessing] [Vanilla Decision Tree, RandomForest, Gradient Boosting models applied sequentially over time]

DeepWisdom



Meta Inference
Storagement
raw datasources
metadata
raw data
features
samples
time control
models
evaluations
predictions
Solution workflow
Search space
Data ingestion subspace
Data preprocessing/ augmentation subspace
Feature-engineering subspace
Model selection subspace
Evaluation subspace
Ensemble subspace
Meta solution agents
Data ingestion agent
Data preprocessing/ augmentation agent
Feature-engineering agent
Model selection agent
Evaluation agent
Ensemble agent
Few-shot streaming
Raw datasource connector
Meta data
Raw datasource
Meta train
MetaAI Agent Factory
Offline data engine
Dataset
Augmentation config
Data Augmentation Engine
Augmented dataset
Optimization metric labels
Meta trainer

automl_freiburg



Winning Solutions Recap

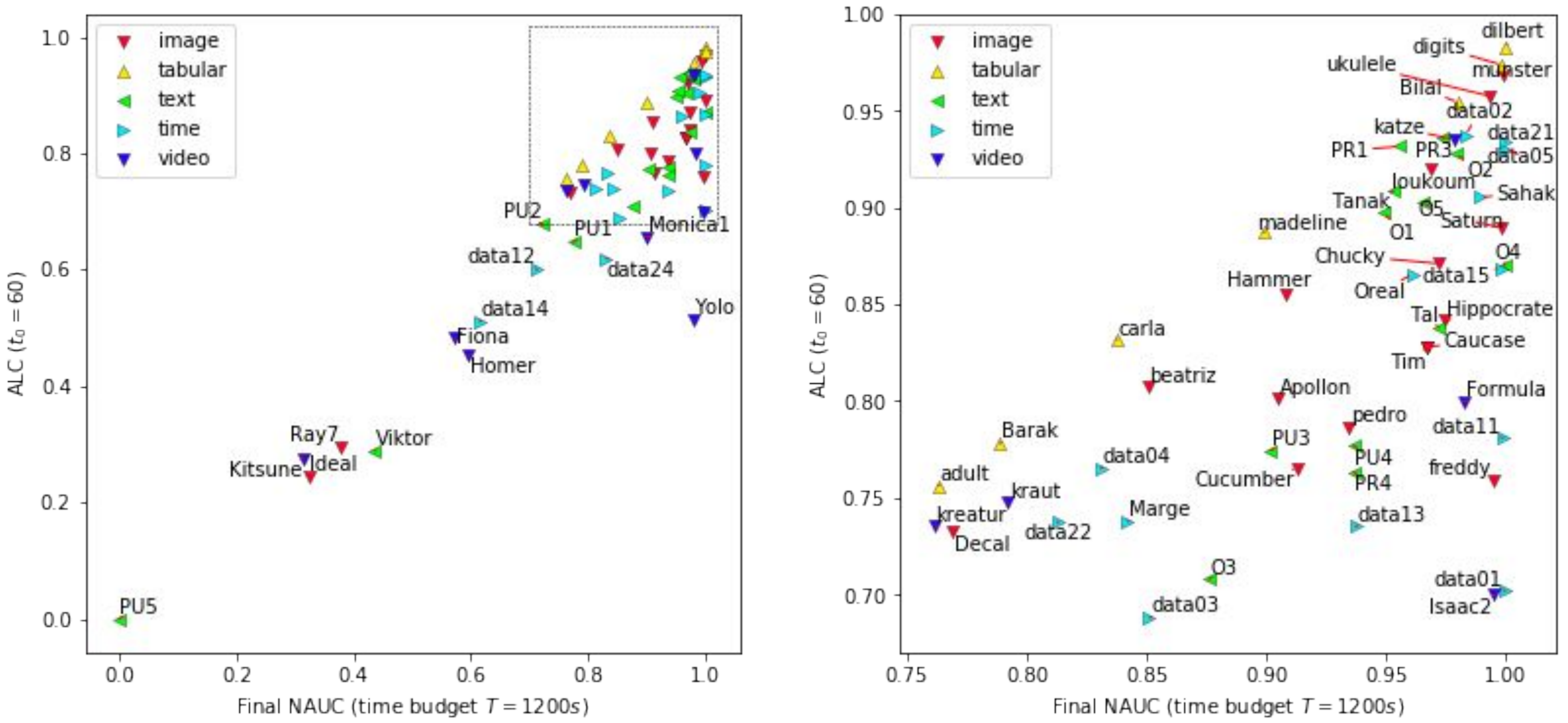
(1) **Deep learning** is still dominant;

(2) **Fixed domain-dependent** and/or **pre-trained** neural architectures are heavily used

(3) **Neural architecture search** (NAS) hasn't been employed due to its huge computational cost

(4) **Meta-learning** and **data loading/ingestion** strategies are used (and are useful)

Benchmarking: DeepWisdom on all 66 AutoDL datasets



More info at: autodl.chalearn.org

References

[1] Liu Z, Bousquet O, Elisseeff A, et al. AutoDL Challenge Design and Beta Tests-Towards automatic deep learning. In: *MetaLearn Workshop @ NeurIPS2018*. Montreal, Canada; 2018. https://hal.archives-ouvertes.fr/hal-01906226. Accessed October 2, 2019.

[2] Liu Z, Guyon I, Junior JJ, et al. AutoCV Challenge Design and Baseline Results. In: *CAp 2019 - Conférence Sur l'Apprentissage Automatique*. Toulouse, France; 2019. https://hal.archives-ouvertes.fr/hal-02265053. Accessed November 5, 2019.

[3] Liu Z, Xu Z, Escalera S, et al. Towards Automated Computer Vision: Analysis of the AutoCV Challenges 2019. To appear in *Pattern Recognition Letters* of Elsevier. 2020. https://hal.archives-ouvertes.fr/hal-02386805. Accessed December 6, 2019.

[4] Liu Z, Xu Z, Rajaa S, Madadi M. Towards Automated Deep Learning: Analysis of the AutoDL challenge series 2019. To appear in *NeurIPSCD2019* in Proceedings of Machine Learning Research (PMLR) 2019:10.

[5] Liu Z, et al, Post-challenge analysis of AutoDL challenges 2019, submitted to TPAMI.

Takeaways

Lessons learned

(1) The winning methods are capable of generalizing on new unseen datasets => Potential universal AutoML solutions

(2) Domain-dependent approaches are dominant
=> No universal workflows, mostly hand-tuned meta-learning

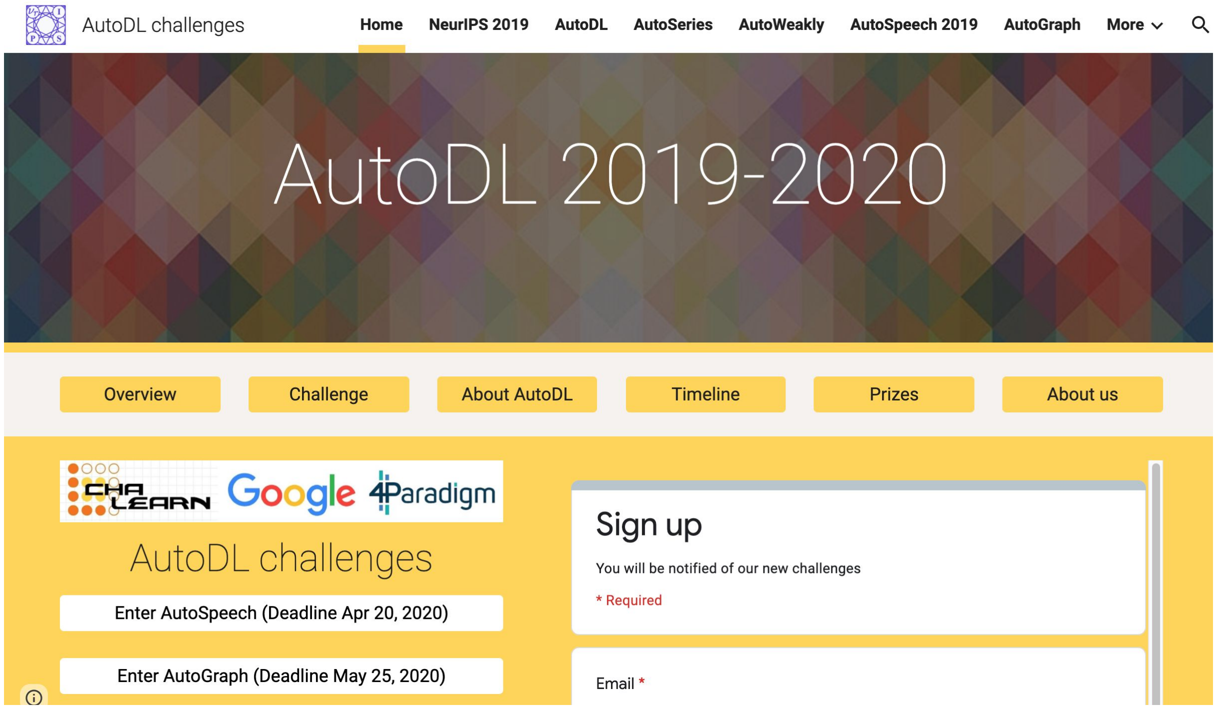
(3) We cannot afford to run expensive NAS for every new task
=> Need transferability of learned architectures

(4) Beating Baseline 3 by using “true” meta-learning is hard
=> Need more meta-train datasets (public datasets)

To achieve true AutoML, we need...

(1) Constructive and efficient representation of meta-knowledges: domain/modality related, pixel correlation, etc
(2) Constructive and efficient representation of learning algorithms: architecture encoding, code-based, etc
(3) Transferable neural architecture search (NAS) to learn a fast algorithm/function: meta-knowledges -> architecture
(4) Lifelong learning systems and/or world models that can learn ONCE but continuously

Thank you! Questions?



JOIN THE CHALLENGES!

autodl.chalearn.org