Analysis of the AutoML Challenge series 2015-2018 Appendix

Isabelle Guyon and Lisheng Sun-Hosoya and Marc Boullé and Hugo Jair Escalante and Sergio Escalera and Zhengying Liu and Damir Jajetic and Bisakha Ray and Mehreen Saeed and Michèle Sebag and Alexander Statnikov and Wei-Wei Tu and Evelyne Viegas

A Meta Features

We define the information and statistics provided in the public or private "info" files

PUBLIC INFO:

- *task* = 'binary.classification', 'multiclass.classification', 'multilabel.classification', 'regression'
- *target_type* = 'Binary', 'Categorical','Numerical'
- $feat\_type$ = 'Binary', 'Categorical','Numerical'
- *metric* = 'bac', 'auc', 'f1', 'pac', 'abs', 'r2'
- $feat\_num$ = number of features
- $target\_num$ = number of columns of target file (one, except for multi-label problems)
- *label_num* = number of labels (number of unique values of the targets)
- $train\_num$ = number of trainign examples
- *valid_num* = number of validation examples (development test set)
- $test\_num$ = number of test examples (final test set)
- *has_categorical* = whether there are categorical variable (yes=1, no=0)
- *has_missing* = whether there are missing values (yes=1, no=0)
- is_sparse = whether the data are in sparse format (yes=1, no=0)

PRIVATE INFO:

- $real\_feat\_num$ = number of real features
- *probe_num* = number of fake features (probes)
- $frac\_probes$ = fraction of probes i.e. $probe_num/(probe\_num+real\_feat\_num)$
- $feat\_type\_freq$ = fraction of feature of each type 'Numerical', 'Categorical', or 'Binary'
- $train\_label\_freq$ = frequency of each label in trainign data
- *train_label_entropy* = entropy of labels in training data
- *train_sparsity* = sparsity of training data (fraction of occurence of zero values)
- $train\_frac\_missing$ = fraction of missing values in training data
- The last 4 statistics are also calculated for the validation set and the test set

- $train\_data\_aspect\_ratio$ = ratio of number of training examples over number of features

We define the meta features as implemented in [7, 6, 8][1]:

- ClassProbabilityMin = $min_{i=1...n}(p(Class_i)) = min_{i=1...n}(\frac{NumberOfInstances\_Class_i}{TotleNumberOfInstances})$
- ClassProbabilityMax = $max_{i=1...n}(p(Class_i)) = max_{i=1...n}(\frac{NumberOfInstances\_Class_i}{TotleNumberOfInstances})$
- ClassEntropy = $mean(-\sum_{i=1}^{n} p(Class_i)ln(p(Class_i)))$ where $p(Class_i)$ is the probability of having an instance of Class_i
- ClassOccurences = number of examples for each class
- ClassProbabilityMean = mean($\frac{ClassOcurrences}{NumberOfClasses}$)
- ClassProbabilitySTD = std($\frac{ClassOcurrences}{NumberOfClasses}$)
- DatasetRatio = $\frac{NumberOfFeatures}{NumberOfInstances}$
- InverseDatasetRatio = $\frac{NumberOfInstances}{NumberOfFeatures}$
- LogInverseDatasetRatio = log($DatasetRatio$)
- Landmark[Some_Model]: accuracy of [Some_Model] applied on dataset.
- LandmarkDecisionNodeLearner & LandmarkRandomNodeLearner: Both are decision tree with max_depth=1. 'DecisionNode' considers all features when looking for best split, and 'RandomNode' considers only 1 feature, where comes the term 'random'.
- Skewnesses: Skewness of each numerical features. Skewness measures the symmetry of a distribution. A skewness value > 0 means that there is more weight in the left tail of the distribution. Computed by scipy.stats.skew.
- SkewnessMax / SkewnessMin / SkewnessMean / SkewnessSTD: max / min / mean / std over skewness of all features.
- NumSymbols: Sizes of categorical features: for each categorical feature, compute its size (number of values in the category).
- SymbolsMax / SymbolsMin / SymbolsMean / SymbolsSTD / SymbolsSum = max / min / mean / std / sum over NumSymbols
- NumberOfCategoricalFeatures: Number of categorical features.
- NumberOfNumericFeatures: Number of numerical features
- RatioNumericalToNominal = $\frac{NumberOfNumericFeatures}{NumberOfCategoricalFeatures}$
- RatioNominalToNumerical = $\frac{NumberOfCategoricalFeatures}{NumberOfNumericFeatures}$
- Kurtosis = Fourth central moment divided by the square of the variance = $\frac{E[(x_i - E[x_i])^4]}{[E[(x_i - E[x_i])^4]]^2}$ where x_i is the i-th feature. Computed using scipy.stats.kurtosis.
- KurtosisMax / KurtosisMin / KurtosisMean / KurtosisSTD = max / min / mean / std of kurtosis over all features
- PCAKurtosis: Transform data by PCA, then compute the kurtosis
- NumberOfInstances = Number of examples
- NumberOfFeatures = Number of features
- NumberOfClasses = Number of classes
- LogNumberOfFeatures = log($NumberOfFeatures$)

[1] Kurtosis, Skewness, KurtosisPCA and SkewnessPCA are intermediate metafeatures used to calculate some other metafeatures

– LogNumberOfInstances = $\log(NumberOfInstances)$
– MissingValues: Boolean matrix of dim (NumberOfInstances , NumberOfFeatures), indicating if an element of is a missing value.
– NumberOfMissingValues: Total number of missing value
– NumberOfInstancesWithMissingValues: Number of examples containing missing values.
– NumberOfFeaturesWithMissingValues: Number of features containing missing values.
– PCA: PCA decomposition of data.
– PCAFractionOfComponentsFor95PercentVariance: Fraction of PCA components explaining 95% of variance of the data.
– PCAKurtosisFirstPC: Kurtosis of the first PCA component.
– PCASkewnessFirstPC: Skewness of the first PCA component.

B Datasets of the 2015/2016 AutoML challenge

ROUND 0

SET 0.1: ADULT

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multilabel	F1	3	1	0.16	0.011	1	0.5	9768	4884	34190	24	1424.58

This dataset was prepared by Isabelle Guyon from original data extracted by Barry Becker from the 1994 Census database. The data was donated to the UCI repository by Ron Kohavi: ”Adult data set” (`https://archive.ics.uci.edu/ml/datasets/Adult`).

Past Usage: The Adult data set is among the most used marketing-style datasets. The ADA dataset is a version of it that was used previously used in the Performance Prediction challenge, the Model Selection game, and the Agnostic Learning vs. Prior Knowledge (ALvsPK) challenge.

Description: The original prediction task was to determine whether a person makes over 50K a year from census data. The problem was transformed into a multilabel problem by adding sex and race in the target values (for race, separate white form others).

Preparation: A set of reasonably clean records was extracted using the following conditions: $((AAGE > 16)$ and $(AGI > 100)$ and $(AFNLWGT > 1)$ and $(HRSWK > 0))$.

Representation: The features include age, workclass, education, etc.

SET 0.2: CADATA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
regression	R2	0	NaN	0	0	0	0.5	10640	5000	5000	16	312.5

This dataset was prepared by Isabelle Guyon from original data provided by Kelley Pace and Ronald Barry: ”California houses” (`http://lib.stat.cmu.edu/datasets/`).

Past Usage: Part of the StatLib datasets. Pace, R. Kelley and Ronald Barry, Sparse Spatial Autoregressions, Statistics and Probability Letters, 33 (1997) 291-297. It was submitted by Kelley Pace (kpace@unix1.sncc.lsu.edu). [9/Nov/99].

Description: These spatial data contain 20,640 observations on housing prices with 9 economic covariates.

Preparation: The original authors collected information on the variables using all the block groups in California from the 1990 Census. In this sample a block group on average includes 1425.5 individuals living in a geographically compact area. Naturally, the geographical area included varies inversely with the population density. They computed distances among the centroids of each block group as measured in latitude and longitude. The final data contained 20,640 observations on 9 variables. The dependent variable is ln(median house value). For the purpose of the AutoML challenge, all samples were merged and the data were freshly randomly split in three sets: training, validation, and test. The order of the features was randomized, after adding a few distractor features (probes) that are permuted versions of real features.

Representation: Features.

SET 0.3: DIGITS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	BAC	10	1	0.42	0	0	0.5	35000	20000	15000	1568	9.57

This dataset was prepared by Isabelle Guyon from original data provided by Yann LeCun, Corinna Cortes, and Chris Burges: ”MNIST handwritten digit dataset” (`http://yann.lecun.com/exdb/mnist/`).

Past Usage: Many methods have been tried on the MNIST database, in its original data split (60,000 training examples, 10,000 test examples, 10 classes). This dataset was used in the NIPS 2003 Feature Selection Challenge under the name GISETTE and in the WCCI 2006 Performance Prediction Challenge and the IJCNN 2007 Agnostic Learning vs. Prior Knowledge Challenge under the name GINA, and in the ICML 2011 Unsupervised and Transfer Learning Challenge under the name ULE.

Description: This is a dataset of handwritten digits. It is a subset of a larger set made available from NIST. The digits in pixel representation have been size-normalized and centered in a fixed-size image by the authors. The data are quantized on 256 gray level values.

Preparation: For the purpose of the AutoML challenge, all samples were merged and the data were freshly randomly split in three sets: training, validation, and test. The order of the features (pixels) was also randomize, after adding a few distractor features (probes) that are permuted versions of real features.

Representation: Pixels.

SET 0.4: DOROTHEA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.46	0.99	0	0	0.5	800	350	800	100000	0.01

This dataset was prepared by Isabelle Guyon from original data provided by DuPont Pharmaceuticals: "Feature selection challenge data" (`http://www.cs.wisc.edu/~dpage/kddcup2001/`).

Past Usage: DOROTHEA was prepared for the NIPS 2003 variable and feature selection benchmark by Isabelle Guyon, 955 Creston Road, Berkeley, CA 94708, USA (isabelle@clopinet.com). The dataset with which DOROTHEA was created is one of the KDD (Knowledge Discovery in Data Mining) Cup 2001. The original dataset and papers of the winners of the competition are available at: `http://www.cs.wisc.edu/~dpage/kddcup2001/`. DuPont Pharmaceuticals graciously provided this data set for the KDD Cup 2001 competition. All publications referring to analysis of this data set should acknowledge DuPont Pharmaceuticals Research Laboratories and KDD Cup 2001.

Description: Synopsis of the original data: One binary attribute (active A or inactive I) must be predicted. Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.). The original training data set consisted of 1909 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting. The chemical structures of these compounds are not necessary for our analysis and were not included. Of the training compounds, 42 are active (bind well) and the others are inactive. To simulate the real-world drug design environment, the test set contained 634 additional compounds that were in fact generated based on the assay results recorded for the training set. Of the test compounds, 150 bind well and the others are inactive. The compounds in the test set were made after chemists saw the activity results for the training set, so the test set had a higher fraction of actives than did the training set in the original data split. Each compound is described by a single feature vector comprised of a class value (A for active, I for inactive) and 139,351 binary features, which describe three-dimensional properties of the molecule. The definitions of the individual bits are not included we only know that they were generated in an internally consistent manner for all 1909 compounds. Biological activity in general, and receptor binding affinity in particular, correlate with various structural and physical properties of small organic molecules. The task is to determine which of these properties are critical in this case and to learn to accurately predict the class value. In evaluating the accuracy, a differential cost model was used, so that the sum of the costs of the actives will be equal to the sum of the costs of the inactives.

Preparation: To prepare the data, we used information from the analysis of the KDD cup 2001 and the literature. There were 114 participants to the competition that turned in results. The winner of the competition is Jie Cheng (Canadian Imperial Bank of Commerce). His presentation is available at: `http://www.cs.wisc.edu/~dpage/kddcup2001/Hayashi.pdf`. The data was also studied by Weston and collaborators: J. Weston, F. Perez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff and B. Schoelkopf. "Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design". Bioinformatics. At lot of information is available from Jason Weston s web page, including valuable statistics about the data: `http://www.kyb.tuebingen.mpg.de/bs/people/weston/kdd/kdd.html`. To outperform these results, the paper of Weston et al., 2002, utilizes the combination of an efficient feature selection method and a classification strategy that capitalizes on the differences in the distribution of the training and the test set. First they select a small number of relevant features (less than 40) using an unbalanced correlation score that selects features that have non-zero entries only for positive examples. This score encodes the prior information that the data is unbalanced and that only positive correlations are likely to be useful. The score has an information theoretic motivation, see the paper for details.

Representation: The original data set was modified for the purpose of the feature selection challenge: The original training and test sets were merged. The features were sorted according to an unbalanced correlation critetion, computed using the original test set (which is richer is positive examples). Only the top ranking 100000 original features were kept. The all zero patterns were removed, except one that was given label ?1. For the second half lowest ranked features, the order of the patterns was individually randomly permuted (in order to create "random probes" or distrator features). The order of the patterns and the order of the features were globally randomly permuted to mix the original training and the test patterns and remove the feature order. The data was split into training, validation, and test set while respecting the same proportion of examples of the positive and negative class in each set.

SET 0.5: NEWSGROUPS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	PAC	20	1	1	0	0	0	3755	1877	13142	61188	0.21

This dataset was prepared by Hugo Jair Escalante from original data provided by Ken Lang. The version we used was obtained from Deng Cai.: "TNW - 20 Newsgroups data set" (`http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html`).

Past Usage: The 20 NewsGroups data set is among the most used data sets for text categorization. It has been used to evaluate standard text categorization and recently it has been also widely used for the evaluation of cross domain text categorization.

Description: In this version of the data set the training and test documents were mixed in a single matrix, then plit into training, validation, and test set for the needs of the challenge.

Preparation: The data is organized into 20 different newsgroups (each newsgroup corresponds to a class), each corresponding to a different topic (see `http://qwone.com/~jason/20Newsgroups/`). Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). .

Representation: Documents are represented by their bag-of-words using a term-frequency weighting scheme.

ROUND 1

SET 1.1: CHRISTINE

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	BAC	2	1	0.071	0	0	0.5	2084	834	5418	1636	3.31

This dataset was prepared by Isabelle Guyon from original data provided by Curt Breneman, Charles Bergeron, and Kristin Bennett: ”Activation of pyruvate kynase” (`http://www.causality.inf.ethz.ch/activelearning.php?page=datasets`).

Past Usage: Active learning challenge, C dataset, see `http://www.causality.inf.ethz.ch/activelearning.php`.

Description: The task is to predict chemical activity of molecules. This is a two-class classification problems. The variables represent properties of the molecule inferred from its structure. The problem is therefore to relate structure to activity (a QSAR=quantitative structure-activity relationship problem) to screen new compounds before actually testing them (a HTS=high-throughput screening problem). The problem is to predict the activation of pyruvate kynase, a well characterized enzyme, which regenerates ATP in glycolysis by catalyzing phosphoryl transfer from phosphoenol pyruvate to ADP to yield pyruvate and ATP.

Preparation: We modified the original data split and added probes.

Representation: Features/Attributes representing properties of molecules.

SET 1.2: JASMINE

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	BAC	2	1	0.78	0	0	0.5	1756	526	2984	144	20.72

This dataset was prepared by Isabelle Guyon from original data provided by Reza Farrahi Moghaddam, Mathias Adankon, Kostyantyn Filonenko, Robert Wisnovsky, and Mohamed Cheriet: ”Arabic manuscripts” (`http://www.causality.inf.ethz.ch/activelearning.php?page=datasets`).

Past Usage: Active learning challenge, A dataset, see `http://www.causality.inf.ethz.ch/activelearning.php`.

Description: The task is to classify cursive script subwords from data in a feature representation extracted from Arabic Historical Manuscripts..

Preparation: We modified the original data split and added probes..

Representation: Features/Attributes.

SET 1.3: MADELINE

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	BAC	2	1	1.2e-06	0	0	0.92	3240	1080	3140	259	12.12

This dataset was prepared by Isabelle Guyon from original data provided by Isabelle Guypn: "Feature selection challenge data" (http://www.nipsfsc.ecs.soton.ac.uk/datasets/;https://archive.ics.uci.edu/ml/datasets/Madelon).

Past Usage: NIPS 2003 feature selection challenge. See Result analysis of the NIPS 2003 feature selection challenge, Isabelle Guyon, Steve R. Gunn, Asa Ben-Hur, Gideon Dror, 2004.

Description: MADELON is an artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five dimensional hypercube and randomly labeled $+1$ or -1. The five dimensions constitute 5 informative features. 15 linear combinations of those features were added to form a set of 20 (redundant) informative features. Based on those 20 features one must separate the examples into the 2 classes (corresponding to the ± 1 labels). We added a number of distractor feature called *probes* having no predictive power. The order of the features and patterns were randomized. See http://www.nipsfsc.ecs.soton.ac.uk/papers/NIPS2003-Datasets.pdf.

Preparation: To draw random data, the program takes the following steps: (1) Each class is composed of a number of Gaussian clusters. N(0,1) is used to draw for each cluster num_useful_feat examples of independent features. (2) Some covariance is added by multiplying by a random matrix A, with uniformly distributed random numbers between -1 and 1. (3) The clusters are then placed at random on the vertices of a hypercube in a num_useful_feat dimensional space. The hypercube vertices are placed at values ± 1 *class_sep*. (4) Redundant features are added. They are obtained by multiplying the useful features by a random matrix B, with uniformly distributed random numbers between -1 and 1. (5) Some of the previously drawn features are repeated by drawing randomly from useful and redundant features. Useless features (random probes) are added using N(0,1). (6)- All the features are then shifted and rescaled randomly to span 3 orders of magnitude. (7) Random noise is then added to the features according to N(0,0.1). (8) A fraction $flip\_y$ of labels are randomly exchanged.

Representation: Continuous valued features.

SET 1.4: PHILIPPINE

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	BAC	2	1	0.0012	0	0	0.5	4664	1166	5832	308	18.94

This dataset was prepared by Isabelle Guyon from original data provided by Emmanuel Faure, Thierry Savy, Louise Duloquin, Miguel Luengo Oroz, Benoit Lombardot, Camilo Melani, Paul Bourgine, and Nadine Peyrieras: "Mitosis classification" (http://www.causality.inf.ethz.ch/activelearning.php?page=datasets).

Past Usage: Active learning challenge, E dataset, see http://www.causality.inf.ethz.ch/activelearning.php.

Description: A feature representation of cells of zebrafish embryo to determine whether they are in division (meiosis) or not. All the examples are manually annotated.

Preparation: We modified the original data split and added probes.

Representation: Features extracted from video data.

SET 1.5: SYLVINE

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	BAC	2	1	0.01	0	0	0.5	10244	5124	5124	20	256.2

This dataset was prepared by Isabelle Guyon from original data provided by Department of Forest Sciences, Colorado: ”Forest cover types” (`https://archive.ics.uci.edu/ml/datasets/Covertype`).

Past Usage: Active learning challenge, F dataset, see `http://www.causality.inf.ethz.ch/activelearning.php`.

Description: The tasks is to classify forest cover types. The original multi-class problem is brought back to Krummholz vs. other classes of trees.

Preparation: We modified the original data split and added probes.

Representation: Features/Attributes

ROUND 2

SET 2.1: ALBERT

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	F1	2	1	0.049	0.14	1	0.5	51048	25526	425240	78	5451.79

This dataset was prepared by Hugo Jair Escalante from original data provided by Olivier Chapelle (CRITEO): ”Criteos Delayed Feedback in Display Advertising” (`https://www.kaggle.com/c/criteo-display-ad-challenge/details/about-criteo?`).

Past Usage: The data set used for the AutoML challenge was taken from the training-set partition of Criteos-Kaggle challenge. The challenge is runing and there are about 350 teams registered. A closely related data set is described in: O. Chapelle. Modeling delayed feedback in display advertising. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 2014.

Description: The data set is a small subset of the training set provided for the above mentioned challenge. The data set has been balanced (originally the class imbalance ratio was 70/30). .

Preparation: For the purpose of the AutoML challenge, missing values are denoted with NaN, the meaning of the variables has not been described yet (it contains sensitive data). Variables 1-13 are numeric, variables 14-39 are categorical.

Representation: Features related to click prediction, the semantics of the features has not been described elsewhere.

SET 2.2: DILBERT

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	PAC	5	1	0	0	0	0.16	9720	4860	10000	2000	5

This dataset was prepared by Hugo Jair Escalante from original data provided by Fu Jie Huang, Yann LeCun, Leon Bottou: ”NORB data set (2 feature maps)” (`http://cs.nyu.edu/~ylclab/data/norb-v1.0/`).

Past Usage: This data set has been widely used for the evaluation of 3D object cassifcation, it has been very popular recently for deep learning computer vision, the paper introducing the data set is: Yann LeCun, Fu Jie Huang, Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting, CVPR, 2004. (google scholar reports 380 citations).

Description: The data set has 48600 images comprising 5 categories, images come from 50 toys belonging to 5 categories. The objects were imaged by two cameras under 6 lighting conditions, 9 elevations (30 to 70 degrees every 5 degrees), and 18 azimuths (0 to 340 every 20 degrees). Images have been represented with features derived with a convolutional neural network. (TCNN with random weights).

Preparation: For the purpose of the AutoML challenge, all samples were merged (the standard procedure uses half of the images for training and half for testing). Images are represented with the upper layer of a (1-layer) Tiled Convolutional Neural Network (TCNN), no pretraining was performed, random weights were used (see A. Saxe code: `http://web.stanford.edu/~asaxe/random_weights.html`).

Representation: Features learned by a TCNN, we used 2 maps to erduce the dimensionality of the representation, the inputs to the TCNN are the pixels from the two stereo images, a 4x4 window wans considered.

SET 2.3: FABERT

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	PAC	7	0.96	0.99	0	0	0.5	2354	1177	8237	800	10.3

This dataset was prepared by Sergio Escalera from original data provided by Sergio Escalera, Xavier Baro, Jordi Gonzalez, Miguel A. Bautista, Meysam Madadi, Miguel Reyes, Victor Ponce: ”LAP2014 Gesture Recognition Data set using Skeleton features” (`http://sunai.uoc.edu/chalearn/`).

Past Usage: The data from which the LAPSD was generated have been used by several people in two challenges (Multimodal Gesture Recognition and Looking at People Challenges), the number of registered participants exceeded 200 hundred (al least 20 people participated throughout the final stages and developed highly competitive methods). More information can be found in: Sergio Escalera et al. Multi-modal Gesture Recognition Challenge 2013: Dataset and Results. Proc. of ICMI 2013, pp. 445-452, 2013, and in Sergio Escalera et al. ChaLearn Looking at People Challenge 2014: Dataset and results, ECCV-Chalearn workshop 2014.

Description: This is a dataset of gesture recognition. It comprises all of the samples (training+validation+test) of the original data set, a total of 13845 samples. Skeleton information was used to represent gestures (BOW formulation). The original data set has 20 gesture classes, for this data set, the 20 gestures are grouped into 10 different classes (0. Perfetto and frieganiente, 1. Prendere, ok and noncenepiu, 2. Bounissimo, furbo, seipazo and cosatifarei, 3. Chevoui, daccordo and combinato, 4. Sonostufo and messidaccordo, 5. Vattene and Vieniqui, 6. Basta, 7. Fame, 8. Tantotempofa, 9. Cheduepalle.

Preparation: For the purpose of the AutoML challenge, all samples were merged. Skeleton frames were first described by the difference of world-coordinates of joint points and the head joint, and then clustered o generate a 400-words vocabulary, which was used to represent the videos.

Representation: Bag-of-Visual-Words using Skeleton coordinates, vocabulary of 400 codewords was considered.

SET 2.4: ROBERT

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	BAC	10	1	0.01	0	0	0	5000	2000	10000	7200	1.39

This dataset was prepared by Isabelle Guyon from original data provided by Antonio Torralba, Rob Fergus, and William T. Freeman, collected and made available publicly the 80 million tiny image dataset. Vinod Nair and Geoffrey Hinton collected and made available publicly the CIFAR datasets.: "Image classification (from Unsupervised and Transfer Learning Challenge)" (`http://www.cs.toronto.edu/?kriz/cifar.html,http://groups.csail.mit.edu/vision/TinyImages/`).

Past Usage: The data were used in the Unsupervised and Transfer Learning challenge: `http://www.causality.inf.ethz.ch/unsupervised-learning.php`.

Description: These are small pictures of objects, animals. etc. We merged the CIFAR-10 and the CIFAR-100 datasets. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. The CIFAR-100 dataset is similar to the CIFAR-10, except that it has 100 classes containing 600 images each. The 100 classes in the CIFAR-100 are grouped into 20 superclasses..

Preparation: The raw data came as 32x32 tiny images coded with 8-bit RGB colors (i.e. 3x32 features with 256 possible values). We converted RGB to HSV and quantized the results as 8-bit integers. This yielded 30x30x3 = 900 x 3 features. We then preprocessed the gray level image to extract edges. This yielded 30 x 30 features (1 border pixel was removed). We then cut the images into patches of 10x10 pixels and ran kmeans clustering (an on-line version) to create 144 cluster centers. We used these cluster centers as a dictionary to create features corresponding to the presence of one the 144 shapes at one of 25 positions on a grid. This created another 144 x 25 = 3600 features. See `http://jmlr.org/proceedings/papers/v27/supplemental/datasetsutl12a.pdf` for details..

Representation: Bag of word features.

SET 2.5: VOLKERT

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	PAC	10	0.89	0.34	0	0	0	7000	3500	58310	180	323.94

This dataset was prepared by Hugo Jair Escalante from original data provided by J. Vogel and B. Schiele.: "VOGEL data set - image classification" (`http://ccc.inaoep.mx/~hugojair/ebm/ebm_code_and_data.zip`).

Past Usage: This data set has been used in a few publications for the evaluation of region labeling and image retrieval techniques. The data set was introduced in: J. Vogel, B. Schiele. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. Journal of Computer Vision, Vol. 72(2):133–157, 2007, this paper has ben cited around 250 times according to google scholar.

Description: Images are natural scenes from 6 different categories (coasts / rivers-lakes / forests / mountains / plains / sky-clouds). Each image has been divided in regions of 10x10 pixels each (grid segmentation), 100 regions per image were extracted. The goal of the task is to classify the regions. Regions are represented by a set of visual descriptors, and regions are labeled with one of 17 labels, associated to the scene categories.

Preparation: There are 70000 regions to be labeled with one of 17 labels, where every 100 regions (in the actual order of the X file) were extracted from the same image (each image corresponds to a single natural scene category). In the past 10 fold CV has been used for evaluation.

Representation: Images are represented by their edge and HSI-color histograms, as well as by texture features extracted from the co-occurrence matrix.

ROUND 3

SET 3.1: ALEXIS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multilabel	AUC	18	0.92	0.98	0	0	0	15569	7784	54491	5000	10.9

This dataset was prepared by Isabelle Guyon from original data provided by The dataset was constructed from the KTH human action recognition dataset of Ivan Laptev and Barbara Caputo and the Hollywood 2 dataset of human actions and scenes of Marcin Marszalek, Ivan Laptev, and Cordelia Schmidt.: ”Action recognition (from Unsupervised and Transfer Learning Challenge)” (`http://www.nada.kth.se/cvap/actions/,http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`).

Past Usage: The data were used in the Unsupervised and Transfer Learning challenge: `http://www.causality.inf.ethz.ch/unsupervised-learning.php`.

Description: The data include video clips of people performing actions. The identification and recognition of gestures, postures and human behaviors has gained importance in applications such as video surveillance, gaming, marketing, computer interfaces and interpretation of sign languages for the deaf.

Preparation: The data were preprocessed into STIP features using the code of Ivan Laptev: `http://www.irisa.fr/vista/Equipe/People/Laptev/download/stip-1.0-winlinux.zip`. The final representation is a ?bag of STIP features?. Details are found in the report `http://jmlr.org/proceedings/papers/v27/supplemental/datasetsutl12a.pdf`.

Representation: Bag of word features.

SET 3.2: DIONIS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	BAC	355	1	0.11	0	0	0	12000	6000	416188	60	6936.47

This dataset was prepared by Mehreen Saeed from original data provided by Sarmad Hussain and Qurat ul Ain Akram: ”Urdu OCR dataset” (`http://www.cle.org.pk/clestore/imagecorpora.htm`).

Past Usage: `http://www.cle.org.pk/Publication/papers/2013/Binarization%20and%20its%20Evaluation%20for%20Urdu%20Nastalique%20Document%20Images%208-3-1.pdf,http://www.cle.org.pk/Publication/papers/2014/AdaptingTesseract%20for%20Complex%20Scripts-%20an%20Example%20for%20Nastalique%203.10.pdf,www.UrduOCR.netandwww.cle.org.pk/clestore/imagecorpora.htm`.

Description: This is a dataset of Urdu printed ligatures shapes with diacritics stripped off. The dataset has been derived from an original dataset found at : `http://www.cle.org.pk/clestore/imagecorpora.htm` by generating new images from existing ones. A subset of shapes is included in this dataset.

Preparation: For the purpose of the AutoML challenge, new shape images were created using elastic deformations, rotations, shear and scaling. Features were then extracted from the generated images.

Representation: DCT transform of image contours, spatial density computed by dividing each image in a 3x3 grid and computing the density for each cell, eigen values and eigen vector of (x,y) coordinates of foreground shape pixels.

SET 3.3: GRIGORIS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multilabel	AUC	91	0.87	1	0	0	0	9920	6486	45400	301561	0.15

This dataset was prepared by Grigorios Tsoumakas et al. from original data provided by Grigorios Tsoumakas et al.: ”WISE 2014 - Greek Media Monitoring Multilabel Classification” (`https://www.kaggle.com/c/wise-2014`).

Past Usage: The data set is being used in the WISE 2014 - Greek Media Monitoring Multilabel Classification, the challenge is being managed in the Kaggle platform, at the moment of writing this file 121 teams have registered for the competition, these are teams that have made at least one submission.

Description: This is a multi-label classification competition for articles coming from Greek printed media. Raw data comes from the scanning of print media, article segmentation, and optical character segmentation, and therefore is quite noisy. Data was collected by scanning a number of Greek print media from May 2013 to September 2013. There are 301561 numerical attributes corresponding to the tokens encountered inside the text of the collected articles. Articles were manually annotated with one or more out of 203 labels (in this version, there are considered 200 labels only).

Preparation: For the purpose of the AutoML challenge, only the training subset of documents has been considered, a total of 200 labels are considered where each has at least one example.

Representation: Text are represented by their bag-of-words with a tfidf weighting scheme.

SET 3.4: JANNIS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	PAC	4	0.8	7.3e-05	0	0	0.5	9851	4926	83733	54	1550.61

This dataset was prepared by Hugo Jair Escalante from original data provided by Hugo Jair Escalante, Michael Grubinger: "SAIAPR TC12 benchmark - main-branches classification" (`http://imageclef.org/SIAPRdata`).

Past Usage: Several methods for image annotation have been evaluated in the SAIAPR-TC12 collection (see `http://scholar.google.com/scholar?oi=bibs&hl=en&cites=8812357429744542982`); including region-level (this data) and image-level methods. A previous version of this collection (IAPR-TC12) has been widely used to benchmark multimodal image retrieval techniques in the CLEF forum. The data set is described in detail in the following publication: H. J. Escalante, et al. The Segmented and Annotated IAPR-TC12 Benchmark. Computer Vision and Image Understanding Journal, 114(4):419-428, 2010.

Description: In this version of the SAIAPR-TC 12 data set the goal is to classify image regions into one of the 4-most populated branches (Animals, Man-made objects, Persons, Landscape) of a hierarchy of concepts. Each instance is associated to a region of an image. Regions in images have been segmented manually, each region is described by a 27-dimensional verctor comprising the following visual-content attributes: area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in the RGB and CIE-Lab color spaces. In the past, 10-fold cross validation has been used for evaluation.

Preparation: For the purpose of the AutoML challenge, all regions are labeled by the first-level branch of the original labels.

Representation: Region area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in the RGB and CIE-Lab color spaces.

SET 3.5: WALLIS

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	BAC	11	0.91	1	0	0	0	8196	4098	10000	193731	0.05

This dataset was prepared by Hugo Jair Escalante from original data provided by Ana Cardoso Cachopo.: "C12 - the CADE 12 data set" (`http://web.ist.utl.pt/~acardoso/datasets/`).

Past Usage: This data set has been used to evaluate standard (single label) text categorization. There are no too much references using this data set, most work has been reported from Portuguese and Brazilian colleagues.

Description: The documents in the Cade12 correspond to a subset of web pages extracted from the CADE Web Directory, which points to Brazilian web pages classified by human experts.

Preparation: The data is organized into 12 classes each corresponding to a different webpage category (see `http://web.ist.utl.pt/~acardoso/datasets/`).

Representation: Documents are represented by their bag-of-words using a term-frequency weighting scheme.

ROUND 4

SET 4.1: EVITA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.21	0.91	0	0	0.46	14000	8000	20000	3000	6.67

This dataset was prepared by Isabelle Guyon from original data provided by National Cancer Institute (NCI)DTP AIDS Antiviral Screen program: "HIV" (`http://dtp.nci.nih.gov/docs/aids/aids_data.html`).

Past Usage: This data set has been previous use in several challenges including the Performance Prediction Challenge under the name HIVA and the Causation and Prediction Challenge under the name SIDO.

Description: This is a problem of drug activity classification. The data contains descriptors of molecules, which have been tested against the AIDS HIV virus. The target values indicate the molecular activity (+1 active, -1 inactive).

Preparation: The features were reshuffled and a fresh data split was made.

Representation: The molecular descriptors were generated programmatically from the three dimensional description of the molecule, with several programs used by pharmaceutical companies for QSAR studies (Quantitative Structure-Activity Relationship). For example, a descriptor may be the number of carbon molecules, the presence of an aliphatic cycle.

SET 4.2: FLORA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
regression	ABS	0	NaN	0.99	0	0	0.25	2000	2000	15000	200000	0.08

This dataset was prepared by C J Lin from original data provided by These data were collected primarily by Bryan Routledge, Shimon Kogan, Jacob Sagi, and Noah Smith. This version was obtained from C. J. Lin.: "E2006-tfidf 10-K Corpus" (`https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html`).

Past Usage: https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD.

Description: Prediction of the release year of a song from audio features. Songs are mostly western, commercial tracks ranging from 1922 to 2011, with a peak in the year 2000s.

Preparation: The data were obtained by C J Lin. They respect the original representation..

Representation: Features: 12 = timbre average, 78 = timbre covariance.

SET 4.3: HELENA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	BAC	100	0.9	6e-05	0	0	0	18628	9314	65196	27	2414.67

This dataset was prepared by Hugo Jair Escalante from original data provided by Hugo Jair Escalante, Michael Grubinger: "SAIAPR TC12 benchmark - top-100 frequent labels" (`http://imageclef.org/SIAPRdata`).

Past Usage: Several methods for image annotation have been evaluated in the SAIAPR-TC12 collection (see `http://scholar.google.com/scholar?oi=bibs&hl=en&cites=8812357429744542982`); including region-level (this data) and image-level methods. A previous version of this collection (IAPR-TC12) has been widely used to benchmark multimodal image retrieval techniques in the CLEF forum. The data set is described in detail in the following publication: H. J. Escalante, et al. The Segmented and Annotated IAPR-TC12 Benchmark. Computer Vision and Image Understanding Journal, 114(4):419-428, 2010.

Description: In this version of the SAIAPR-TC 12 data set the goal is to classify image regions into one of 100 labels (the top-100 more frequent ones). The original data set has about 276 labels, organized into a hierarchy of concepts, in this version of the data set the goal is to classify the leaf-labels of the hierarchy. Each instance is associated to a region of an image. Regions in images have been segmented manually, each region is described by a 27-dimensional verctor comprising the following visual-content attributes: area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in the RGB and CIE-Lab color spaces. In the past, 10-fold cross validation has been used for evaluation.

Preparation: For the purpose of the AutoML challenge, all regions are labeled by their leaf-label in the hierarchy of concepts.

Representation: Region area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in the RGB and CIE-Lab color spaces.

SET 4.4: TANIA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multilabel	PAC	95	0.79	1	0	0	0	44635	22514	157599	47236	3.34

This dataset was prepared by Isabelle Guyon from original data provided by The original data were donated by Reuters and downloaded from: Lewis, D. D. RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (12-Apr- 2004 Version).: "Text classification (from REUTERS data)" (`http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README`).

Past Usage: The data were used in the Unsupervised and Transfer Learning challenge: `http://www.causality.inf.ethz.ch/unsupervised-learning.php`.

Description: We used a subset of the 800,000 documents of the RCV1-v2 data collection.

Preparation: The data were formatted in a bag-of-words representation. The representation uses 47,236 unique stemmed tokens, see `http://jmlr.org/proceedings/papers/v27/supplemental/datasetsutl12a.pdf` for details. We considered all levels of the hierarchy to select the most promising categories.

Representation: Bag-of-word features.

SET 4.5: YOLANDA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
regression	R2	0	NaN	1e-07	0	0	0.1	30000	30000	400000	100	4000

This dataset was prepared by T. Bertin-Mahieux from original data provided by This data is a subset of the Million Song Dataset: `http://labrosa.ee.columbia.edu/millionsong/` a collaboration between LabROSA (Columbia

University) and The Echo Nest. Prepared by T. Bertin-Mahieux. This version was obtained from C. J. Lin.: ”YearPredictionMSD Data Set ” (`https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD`).

Past Usage: The Million Song Dataset. Thierry Bertin-Mahieux, Daniel P.W. Ellis and Brian Whitman, Paul Lamere. `https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD`; `http://ismir2011.ismir.net/papers/OS6-1.pdf`.

Description: Prediction of the release year of a song from audio features. Songs are mostly western, commercial tracks ranging from 1922 to 2011, with a peak in the year 2000s.

Preparation: The data were obtained by C J Lin. They respect the original representation.

Representation: Features: 12 = timbre average, 78 = timbre covariance.

ROUND 5

SET 5.1: ARTURO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	F1	20	1	0.82	0	0	0.5	2733	1366	9565	400	23.91

This dataset was prepared by Sergio Escalera from original data provided by Sergio Escalera, Xavier Baro, Jordi Gonzalez, Miguel A. Bautista, Meysam Madadi, Miguel Reyes, V?ctor Ponce: ”Multimodal Gesture Recognition Data set using audio features” (`http://sunai.uoc.edu/chalearn/`).

Past Usage: The data from which the ABGR was generated have been used by several people in two challenges (Multimodal Gesture Recognition and Looking at People Challenges), the number of registered participants exceeded 200 hundred (al least 20 people participated throughout the final stages and developed highly competitive methods). More information can be found in: Sergio Escalera et al. Multi-modal Gesture Recognition Challenge 2013: Dataset and Results. Proc. of ICMI 2013, pp. 445-452, 2013, and in Sergio Escalera et al. ChaLearn Looking at People Challenge 2014: Dataset and results, ECCV-Chalearn workshop 2014.

Description: This is a dataset of gesture recognition. It comprises all of the samples (training+validation+test) of the original data set, a total of 13664 samples. 20 classes of gestures were considered and only audio-based features are used to represent clips.

Preparation: For the purpose of the AutoML challenge, all samples were merged. Frames of the clip were first described by the 13 MEL coefficients extracted from the audio signal, and then clustered o generate a 200-words vocabulary, which was used to represent the videos.

Representation: Bag-of-Visual-Words using Mel coefficients coordinates, vocabulary of 200 codewords was considered.

SET 5.2: CARLO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	PAC	2	0.097	0.0027	0	0	0.5	10000	10000	50000	1070	46.73

This dataset was prepared by Bisakha Ray from original data provided by Bisakha Ray, Javier Orlandi, Olav Stetter, Isabelle Guyon: ”Connectomics-features-normal-1” (http://www.kaggle.com/c/connectomics/leaderboard).

Past Usage: Used for Connectomics challenge at http://www.kaggle.com/c/connectomics/leaderboard.

Description: This is a dataset of Connectomics Challenge. The outcome considered is presence or absence of connection.

Preparation: For the purpose of the AutoML challenge, all samples were merged and the data were freshly randomly split in three sets: training, validation, and test. The order of the features (pixels) was also randomize, after adding a few distractor features (probes) that are permuted versions of real features.

Representation: neuronal connection.

SET 5.3: MARCO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multilabel	AUC	180	0.76	0.99	0	0	0	20482	20482	163860	15299	10.71

This dataset was prepared by Yin Aphinyanaphongs from original data provided by William Hersh: ”Ohsumed TEXT dataset” (http://ir.ohsu.edu/ohsumed/ohsumed.html).

Past Usage: Many studies have used the ohsumed corpora for information retrieval research in the biomedical literature. See http://scholar.google.com/scholar?es_sm=91&um=1&ie=UTF-8&lr=&cites=13802943827211985373 for a listing of papers that cite this work.

Description: See the dataset url for more information. To summarize, these are biomedical articles from 1987 to 1991 from over 270 medical journals from the primary literature that contain titles, abstracts, human-assigned MeSH terms, publication types, authors, and source.

Preparation: The original dataset contains 348,566 references from MEDLINE. , the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. We applied the following steps in order: (1) Filter references to contain an abstract, contain a title, and is of type ”journal article.” (2) Concatonate title (.T), abstract (.W), Author (.A), and Source (.S). (3) Replace all punctuation with blanks. (4) Remove stopwords defined in nltk.corpus. (5) Set minimum token occurence to 50. (6) Apply tf-idf to the resulting corpus. The classification targets are determined by ranking the top 200 mesh terms assigned to all the documents and building independent classification tasks for each MeSH term. See Google sheet at https://docs.google.com/spreadsheets/d/1Kihqtds6mYVWTwV4l5qRCNuUCZOnXkpM9zYbbtHT3ts/edit#gid=1366784086 for initial performance estimates on the various classification tasks.

Representation: Bag-of-word features.

SET 5.4: PABLO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
regression	ABS	0	NaN	0.11	0	0	0.5	23565	23565	188524	120	1571.03

This dataset was prepared by Bisakha Ray from original data provided by Jaume Bacardit and Natalio Krasnogor: "The ICOS PSP benchmarks repository" (`http://icos.cs.nott.ac.uk/datasets/psp_benchmark.html`).

Past Usage: M. Stout, J. Bacardit, J.D. Hirst, N. Krasnogor Prediction of recursive convex hull class assignments for protein residues in Bioinformatics, 24(7):916-923, April 2008.

Description: This is a dataset of PSP benchmark repository. The outcome considered is protein structure prediction. It consists of 60 real-valued features for regression. The fold considered is TrainFold09w1.

Preparation: For the purpose of the AutoML challenge, all samples were merged and the data were freshly randomly split in three sets: training, validation, and test. The order of the features (pixels) was also randomize, after adding a few distractor features (probes) that are permuted versions of real features.

Representation: Protein structure features.

SET 5.5: WALDO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
multiclass	BAC	4	1	0.029	0	1	0.5	2430	2430	19439	270	72

This dataset was prepared by Isabelle Guyon from original data prepared from various sources, all in the feature representation designed by Jose Fonollosa: "Cause-Effect Pairs challenge data (in Jarfo representation)" (`http://www.causality.inf.ethz.ch/cause-effect.php?page=data`).

Past Usage: The data were used in the cause-effet pairs challenge in their raw representation.

Description: We provided hundreds of pairs of real variables with known causal relationships from domains as diverse as chemistry, climatology, ecology, economy, engineering, epidemiology, genomics, medicine, physics. and sociology. Those were intermixed with controls (pairs of independent variables and pairs of variables that are dependent but not causally related) and semi-artificial cause-effect pairs (real variables mixed in various ways to produce a given outcome). The goal is to classify the pairs in one of 4 classes "A causes B", B causes A", "A and B are independent" or "A and B are dependent but not causally related".

Preparation: One of the participant extracted features of the joint distribution of the variable pairs. Those feature (which we provide), include information theoretic features such as conditional entropy and results of independence tests.

Representation: Features.

C Datasets of the 2018 AutoML challenge

C.1 PHASE 1: development

SET 1.1: ADA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	1	0.33	0	0	0	41471	415	4147	48	86.39

This dataset is a version of the Adult data set used in round 0 of the 2015/2016 AutoML challenge. It was prepared by Isabelle Guyon from original data extracted by Barry Becker from the 1994 Census database. The data was donated to the UCI repository by Ron Kohavi: "Adult data set" (`https://archive.ics.uci.edu/ml/datasets/Adult`).

Past Usage: It was used previously used in the Performance Prediction challenge, the Model Selection game, and the Agnostic Learning vs. Prior Knowledge (ALvsPK) challenge. Adult, a version of ADA was used in round 0 of the 2015/2016 AutoML challenge.

Description: The task of ADA is to discover high revenue people from census data. This is a two-class classification problem. The raw data from the census bureau is known as the Adult database in the UCI machine-learning repository. The 14 original attributes (features) include age, workclass, education, education, marital status, occupation, native country, etc. Categorical features were eliminated and the original numerical features were preprocessed to obtain 48 attributes.

Preparation: A set of reasonably clean records was extracted using the following conditions: (($AAGE > 16$) and ($AGI > 100$) and ($AFNLWGT > 1$) and ($HRSWK > 0$)).

Representation: The features include age, workclass, education, etc.

SET 1.2: ARCENE

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.78	0.54	0	0	0	700	100	100	10000	0.01

This dataset was made available by Isabelle Guyon. The tasks consist in distinguishing cancer versus normal patterns from mass-spectrometric data. This is a two-class classification problem with continuous input variables. More information on the dataset is available from this link: `https://archive.ics.uci.edu/ml/datasets/Arcene`

Past Usage: The Arcene dataset has been used previously in the NIPS 2003 feature selection challenge.

Description: The data were obtained from two sources: The National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS). All the data consist of mass-spectra obtained with the SELDI technique. The samples include patients with cancer (ovarian or prostate cancer), and healthy or control patients. Ovarian cancer samples comprise 253 spectra, including 91 controls and 162 cancer spectra. Regarding the prostate cancer, there are 253 normal samples and 69 disease samples. The number of original features is 15154.

Preparation: The samples were prepared as described in `http://clopinet.com/isabelle/Projects/NIPS2003/Slides/NIPS2003-Datasets.pdf`. After preprocessing, 3000 informative features and 7000 probes were included in the data set.

Representation: See `http://clopinet.com/isabelle/Projects/NIPS2003/Slides/NIPS2003-Datasets.pdf`.

SET 1.3: GINA

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	1	0.97	0.31	0	0	31532	315	3153	970	3.25

This dataset was prepared by Isabelle Guyon. The associated task is handwritten digit recognition. Specifically, the problem of separating the odd numbers from even numbers. This is a twoclass classification problem with sparse continuous input variables, in which each class is composed of several clusters. It is a problems with heterogeneous classes.

Past Usage: It was used previously used in the Performance Prediction challenge, the Model Selection game, and the Agnostic Learning vs. Prior Knowledge (ALvsPK) challenge.

Description: The dataset was formed with instances from the MNIST dataset that is made available by Yann LeCun at `http://yann.lecun.com/exdb/mnist/`.

Preparation: The following process was followed for preparing the data: Pixels that were 99% of the time white were removed. This reduced the original feature set of 784 pixels to 485. The original resolution (256 gray levels) was kept. The feature names are the (i,j) matrix coordinates of the pixels (in a 28x28 matrix). Two digit numbers were generated by dividing the datasets into to parts and pairing the digits at random. The task is to separate odd from even numbers. The digit of the tens being not informative, the features of that digit act as distracters.

Representation: Pixels from the images were used as features. More information on the dataset can be found in the following link: `http://clopinet.com/isabelle/Projects/agnostic/Dataset.pdf`

SET 1.4: GUILLERMO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.67	0.53	0	0	0	5000	5000	20000	4296	4.65

This data set was prepared by Luis Pellegrin and Hugo Jair Escalante. It comprises preprocessed image-text pairs. Original data was obtained from the SAIAPR TC12 benchmark, provided and prepared by Michael Grubinger and Hugo Jair Escalante (`http://imageclef.org/SIAPRdata`).

Past Usage: The GUILLERMO data set was previously used in the RICATIM - Text Image Matching challenge.

Description: The prediction task consists of determining whether a pair of image - text is related. A word (text) is relevant to an image (and vice versa) if the word was used as label for the image in the original SAIAPR TC12 benchmark. Thus, the image labeling problem is casted as one of binary classification. Images and words are encoded via learned representations as described below, both representations are concatenated to generate the input space of instances. Negative pairs were generated by sampling irrelevant labels.

Preparation: The data set was generated by sampling around 3,000 labeled images from the SAIAPR TC12 data set (formed by 20,000 images). The data set is almost balanced.

Representation: Images were represented by the response of a pretrained CNN (penultimate layer of VGG-16). Words were represented by their Word2Vec representation. An embedding of 200 dimensions was considered, the embedding was trained with the Wikipedia collection.

SET 1.5: RL

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.10	0.99	0.11	1	0	24803	0	31406	22	1427.5

This is a confidential dataset provided by the 4paradigm company, hence we cannot disclose confidential information about it. Although this dataset is publicly available as it was used for the feedback phase of the 2018 AutoML challenge.

Past Usage: This data set was specifically generated for the 2018 AutoML challenge.

Description: The RL data set is associated to a real-world recommendation task involving real users. Items can be: video, audio and activities recommendations, and labels are generated by clicks from users. Instances in this dataset are chronologically ordered, real recommendations and clicks of users from a small time period were considered.

Preparation: A small sample from real recommendations-clicks was taken for preparing this data set. The class imbalance ratio for this dataset was determined to resemble the actual imbalance ratio observed in practice in the associated recommendation task.

Representation: Processed numerical and categorical features encoding descriptive information were made available with this data set.

C.2 PHASE 2: final AutoML testing

SET 2.1: PM

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.01	1	0.11	1	0	20000	0	29964	89	224.71

This is a confidential dataset provided by the 4paradigm company, hence we cannot disclose confidential information about it.

Past Usage: This data set was specifically generated for the 2018 AutoML challenge.

Description: The PM data set is associated to a real-world click prediction task involving real users. More specifically a search-result-click through rate-prediction problem is considered. Instances in this dataset are chronologically ordered, real clicks of users from a small time period were considered.

Preparation: A small sample from real search-results-clicks was taken for preparing this data set. The class imbalance ratio for this dataset was determined to resemble the actual imbalance ratio observed in practice in the associated recommendation task.

Representation: Processed categorical features encoding descriptive information were made available with this data set.

SET 2.2: RH

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.04	0.59	0	1	0	28544	0	31498	76	414.44

This is a confidential dataset provided by the 4paradigm company, hence we cannot disclose confidential information about it.

Past Usage: This data set was specifically generated for the 2018 AutoML challenge.

Description: The RH data set is associated to a real-world recommendation task involving real users. Items can be: video, audio and activities recommendations, and labels are generated by clicks from users. Instances in this dataset are chronologically ordered, real recommendations and clicks of users from a small time period were considered.

Preparation: A small sample from real recommendations-clicks was taken for preparing this data set. The class imbalance ratio for this dataset was determined to resemble the actual imbalance ratio observed in practice in the associated recommendation task.

Representation: Processed numerical and categorical features encoding descriptive information were made available with this data set.

SET 2.3: RI

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.02	0.91	0.26	1	0	26744	0	30562	113	270.46

This is a confidential dataset provided by the 4paradigm company, hence we cannot disclose confidential information about it.

Past Usage: This data set was specifically generated for the 2018 AutoML challenge.

Description: The RI data set is associated to a real-world recommendation task involving real users. Items can be: video, audio and activities recommendations, and labels are generated by clicks from users. Instances in this dataset are chronologically ordered, real recommendations and clicks of users from a small time period were considered.

Preparation: A small sample from real recommendations-clicks was taken for preparing this data set. The class imbalance ratio for this dataset was determined to resemble the actual imbalance ratio observed in practice in the associated recommendation task.

Representation: Processed numerical and categorical features encoding descriptive information were made available with this data set.

SET 2.4: RICCARDO

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.33	0.51	0	0	0	5000	5000	20000	4296	4.65

This data set was prepared by Luis Pellegrin and Hugo Jair Escalante. It comprises preprocessed image-text pairs. Original data was obtained from the common objects in context collection (`http://cocodataset.org/`).

Past Usage: This data set was specifically generated for the 2018 AutoML challenge. It was built following a similar methodology as with the GUILLERMO data set above.

Description: The prediction task consists of determining whether a pair of image - text is related. A text (text could be either a word or the caption accompanying an image) is relevant to an image (and vice versa) if the text was used as caption (or word in the caption) for the image in the original MS COCO benchmark. Thus, the image captioning/labeling problem is casted as one of binary classification. Images and texts are encoded via learned representations as described below, both representations are concatenated to generate the input space of instances. Negative pairs were generated by sampling irrelevant labels.

Preparation: This data set was generated by sampling labeled images from the MS COCO data set. Texts were generated by either captions or words appearing in the captions. The data set is almost balanced.

Representation: Images were represented by the response of a pretrained CNN (penultimate layer of VGG-16). Texts were represented by their Word2Vec representation. An embedding of 200 dimensions was considered, the embedding was trained with the Wikipedia collection. For words, the direct embedding was used. For captions, the average embedding (over words appearing in the caption) was considered.

SET 2.5: RM

Task	Metric	C	Cbal	Sparse	Miss	Cat	Irr	Pte	Pva	Ptr	N	Ptr/N
binary	AUC	2	0.001	1	0.11	1	0	26961	0	28278	89	317.73

This is a confidential dataset provided by the 4paradigm company, hence we cannot disclose confidential information about it.

Past Usage: This data set was specifically generated for the 2018 AutoML challenge.

Description: The RM data set is associated to a real-world click prediction task involving real users. More specifically a search-result-click through rate-prediction problem is considered. Instances in this dataset are chronologically ordered, real clicks of users from a small time period were considered.

Preparation: A small sample from real search-results-clicks was taken for preparing this data set. The class imbalance ratio for this dataset was determined to resemble the actual imbalance ratio observed in practice in the associated recommendation task.

Representation: Processed categorical features encoding descriptive information were made available with this data set.

D Methods of the 2015/2016 AutoML challenge

In this appendix, we first present the results of a survey we conduccted after the challenge, then briefly summarize the best methods based on fact sheets and papers presented at the ICML 2016 workshop where the winners presented their results.

D.1 Survey Analysis

Twenty-eight teams responded to a survey we conducted on methods used in the challenge. **Preprocessing.** Preprocessing consisted in normalization, feature extraction, and dimensionality reduction. About one half of the respondents performed classical preprocessing steps, including feature standardization, sample normalization, and replacement of missing values. This is consistent with the frequent use of ensembles of decision trees based on decision thresholds, which do not require complex preprocessing. Other preprocessing steps included grouping modalities for categorical variables (20%) and discretization (4%). Few participants also reported having used non-linear transforms such as log. Most participants did not perform any feature engineering, which can largely be explained by the fact that they did not know the application domain of the data sets. Those who reported using feature extraction either relied on the (embedded) feature learning of their algorithm (21%) or applied random functions (36%). More than 2/3 of the participants used dimensionality reduction, linear manifold transformations (e.g., PCA, ICA) being the most popular (43%). About 1/3 used feature selection alone. Other methods included non-linear dimensionality reduction (e.g., KPCA, MDS, LLE, Laplacian Eigenmaps) and clustering (e.g., K-means).

Predictor. The methods most frequently used involved (ensembles of) decision trees; 75% of the participants reported having used them, alone or in combination with other methods. The challenge setting lent itself well to such methods because each individual base learner trains rapidly and performance improves by increasing the number of learners, making such methods ideal any-time learning machines. Almost 1/2 of the participants used linear methods and about 1/3 used at least one of the following methods: Neural Nets, Nearest Neighbor, and Naive Bayes. The logistic loss was frequently used (75%). This may be due to the fact that producing probability-like scores is the most versatile when it comes to being able to be judged with a variety of loss functions. About 2/3 of the participants reported having used knowingly some form of regularization; two-norm regularization was slightly more popular than one-norm regularization.

Model selection and ensembling. About 2/3 of the respondents used one form of cross-validation for model selection; the rest used just the leaderboard. This may be due to the fact that the validation sets were not small for the most part. While K-fold cross-validation and leave-one-out remain the most popular, 20% of the respondents used the out-of-bag estimator of bagging methods and 10% used bi-level optimization methods. 4% reported transferring knowledge from phase to phase. However, such a strategy may be worth considering since both winners of phase AutoML5 used it. Only 18% of the respondents did not choose ensemble methods. For those who did, boosting and bagging were the most common—60% reported having used one of the two.

Implementation. Most respondents could not reliably evaluate how their methods scaled computationally. We are at least assured that they delivered results in less than 20 minutes on every data set, because this was the time limit for the execution. Most respondents claimed to have developed a simple method, easy to implement and parallelize (75% used multi-processor machines, 32% used algorithms run in parallel on different machines), but few claimed that their method was original or principled, and most relied on third-party libraries; scikit-learn, which was used in the starting kit, was frequently used. Luckily, this also resulted in code that was made available as open source—with only 10% exceptions. Python was used by 82% of the respondents. This is also explained by the fact that the starting kit was in Python. Although Codalab allows users to submit any Linux executable, the organizers provided no support for this. Even then, 25% used at least one of the following languages: C/C++, Java, or R, sometimes in combination with Python. The fact that the Codalab backend ran on Linux may also explain that 86% of the respondents ran on Linux; others used Windows or MacOS. Memory consumption was generally high (more than half of the respondents used between 8 and 32 GB, and 18% used more that 32 GB). Indeed, when we introduced sparse data in Round 3, the sample code was memory demanding and we had to increase the memory on the server up to 56 GB. Unfortunately, this remained a problem until the end of the challenge—which we traced to an inefficient implementation of the data reader and of Random Forest for sparse matrices.

D.2 Fact Sheets

The methods of top ranking participants of the 2015/2016 challenge are briefly summarized.

ideal.intel.analytics and amsl.intel.com The proprietary solution of the Intel team was presented by Eugene Tuv at the CiML workshop at NIPS, Montreal, December 2015 [2]. It is a fast implementation of tree-based methods in C/C++, which was developed to drive acceleration of yield learning in semiconductor process development. Using this software, the Intel team consistently has ranked high in ChaLearn challenges since 2003. The method is based on gradient boosting of trees built on a random subspace dynamically adjusted to reflect learned features relevance. A Huber loss function is used. No pre-processing was done, except for feature selection [21]. The classification method called Stochastic Gradient Tree and Feature Boosting selects a small sample of features at every step of the ensemble construction. The sampling distribution is modified at every iteration to promote more relevant features. The SGTFB complexity is of the order of $N_{tree}N_{tr}logN_{tr}logN_{feat}$, where N_{tree} is the number of trees, N_{tr} the number of training examples, and N_{feat} the number of features.

aad_freiburg The open-source solution of AAD Freiburg uses a heterogeneous ensemble of learning machines (auto-sklearn [7, 8]) combining the machine learning library scikit-learn [15] with the state-of-the-art SMBO method SMAC to find suitable machine learning pipelines for a data set at hand. This is essentially a reimplementation of Auto-WEKA. To speed up the optimization process they employed a meta-learning technique [6] which starts SMAC from promising configurations of scikit-learn. Furthermore, they used the outputs of all models

[2] http://ciml.chalearn.org/home

and combined these into an ensemble using ensemble selection. Their latest version uses a python reimplementation of SMAC [10] of Bayesian Optimization with Random Forests applied to a flexible configuration space describing scikit-learn. For the GPU version [14], they used the Java version of SMAC to tune auto-sklearn and deep neural networks implemented in Lasagne/Theano [4, 20].

jrl44, backstreet.bayes, and lise_sun Freeze Thaw Ensemble Construction [13] of J. Lloyd (a.k.a. jrl44 and backstreet.bayes) is a modified version of the Freeze Thaw Bayesian optimization algorithm [17] for ensemble construction. The strategy is to keep training the most promising members of an ensemble, while freezing the least promising ones, which may be thawed later. Probabilistic models based on Gaussian processes and decision trees are used to predict which ensemble member should be trained further. Joining late in the challenge, L. Sun made an entry in AutoML5 that ranked third using a similar approach [16].

abhishek4 AutoCompete of [19] is an automated machine learning framework for tackling Machine Learning competitions. This solution performed well in late rounds of the AutoML challenge and won the GPU track [18]. The pipeline includes (1) stratified data splitting, (2) building features, (3) feature selection, (4) performing model and hyper-parameter selection (Random Forests, Logistic Regression, Ridge Regression, Lasso, SVM, Naive Bayes, and Nearest Neighbors), and (5) ensembling solutions. Search space is specified with prior knowledge on similar data sets (a form of meta-learning). Thakur found that this strategy is faster and yields comparable results to hyperopt.The underlying implementation is based purely on Python and scikit-learn with some modules in Cython. Their GPU solution is an advanced version of the AutoCompete solution, which uses Neural Networks built with Keras [3].

djajetic Djajetic [11] is based on heterogeneous ensembles of models obtained by searching through model-space and adjusting hyper-parameters (HP) without any communication between models. Jajetic believes that this makes search more effective in non-convex search spaces. This strategy lends itself well to efficient and simple parallelization. The search space and ensembling properties for each individual model is defined in a separate Python script. Each model is trained and explores its own parameter space and only communicates its training error and best prediction results to the outside. The ensembling module operates in a hierarchical manner. It uses only the N best HP settings from each model, based on the training error, and only M best models from each model group. For the GPU track, Jajetic used a Neural Network [12] based on the Lasagne and Theano libraries.

marc.boulle Orange, the main French telecommunication operator, has developed the Khiops, which they made avaiable for licensing. The software was designed to address the needs of Orange to analyze their data accross a wide range of cases, without hyper-parameter tuning, and provide solutions that are robust and understandable with modest computational resources. Khiops exploits regularized methods for variable preprocessing, variable selection, variable construction for multi-table data mining, correlation analysis via k-coclustering, model averaging of selective naive Bayes classifiers and regressors. The classifier called Selective Naive Bayes (SNB) [1, 2] extends the Naive Bayes classifier using an optimal estimation of the class conditional probabilities, a Bayesian variable selection and a Compression-based Model Averaging. The same framework was extended to regression in [9]. The Khiops tool was used throughout the challenge, using python scripts to be compliant to the challenge settings. Beyond the necessary but easy adaptation to the input/output requirements, the python scripts also had to manage the sparse format, the any-time learning settings and the scoring metrics, which were specific to the AutoML challenge and not supported by Khiops.

E Methods of the 2018 AutoML challenge

E.1 Survey Analysis

Eleven teams responded to a survey we conducted on methods used in the 2018 challenge. The answers to this survey were consistent with the one reported in Appendix D.1. In the following we briefly summarize the main findings.
Preprocessing. 73% of teams applied feature standardization, 54% of teams applied a preprocessing to replace missing values, and 37% applied data normalization. Interestingly, the winning team applied data discretization and scaling in addition to the other preprocessing procedures. Regarding feature extraction, most teams adopted either trained feature extractors or random functions in the same proportion. More than half of the surveyed teams performed linear transformations of the input space, a third of teams performed feature selection.
Predictor. Decision trees was the predictive model adopted by most participants (9 out of 11) that is 81%, the rest of teams used linear models. Hinge loss with 1 or 2 norm regularization was adopted in by most of the teams. **Model selection and ensembling.** As model selection criterion, the usual $k-$fold cross validation and the feedback obtained from the leader board were adopted by 50% of the teams each. Interestingly, all teams that filled in the survey adopted an ensemble methodology for generating the final predictor (mostly boosting-based ensembles). This is consistent with the answers observed in the previous edition of the challenge. **Implementation.** Python was used by all participants and about 20% of teams reported using the scikit-learn library (we believe that most, if not all, participants relied on this library, though).

E.2 Fact Sheets

The methods of the top ranking participants of the 2018 challenge are briefly summarized.

aad_freiburg PoSH Auto-sklearn (*Portfolio Successive Halving* combined with Auto-sklearn) is the solution of the aad_freiburg team, which obtained the best performance in the 2018 challenge. PoSH Auto-sklearn uses a fixed portfolio of machine learning pipeline configurations on which it performs successive halving. If there is time left, it uses the outcome of these runs to warmstart a combination of Bayesian optimization and successive halving. Greedy submodular function maximization was used on a large performance matrix of ≈421 configurations run on ≈421 datasets to obtain a portfolio of configurations that performs well on a diverse set of datasets. To obtain the matrix, aad_freiburg used SMAC [10] to search the space of configurations offline, separately for each of the ≈421 datasets. The configuration space was a subspace of the Auto-sklearn configuration space: dataset preprocessing (feature scaling, imputation of missing value, treatment of categorical values), but no feature preprocessing (this constraint due to the short time limits / resources in the competition), and one of SVM, Random Forest, Linear Classification (via SGD) or XGBoost. The combination of Bayesian optimization and successive halving is an adaptation of a newly developed method dubbed BO-HB (Bayesian Optimization HyperBand) [5]. The solution was further designed to yield robust results within the short time limits as follows: the number of iterations was used as a budget, except for the SVM, where the dataset size was the budget. If the dataset had less than 1000 data points, they reverted to simple cross-validation instead of successive halving. If a dataset had more than 500 features, they used univariate feature selection to reduce the number of features to 500. Lastly, for datasets with more than 45,000 data points, they capped the number of training points to retain decent computational complexity.

narnars0 The narnars0 team proposed an **Automated Machine Learning System for Voting Classifier with Various Tree-Based Classifiers.** This team based their solution in a voting ensemble formed with the following tree-based classifiers: gradient boosting, random forests, and extra-trees classifiers. They optimized the hyperparameters of tree-based classifiers by means of Bayesian optimization. Several machine learning models in scikit-learn were used to implement this system, including narnars0's own Bayesian optimization package, bayeso (`https://github.com/jungtaekkim/bayeso`), which was used to optimize a selection of hyperparameters of classifiers.

wlWangl An AutoML solution resembling **Q-Learning** in reinforcement learning was proposed by the wlWangl team. This team considers the machine learning design pipeline as composed of three phases: data preprocessing, feature selection, and classification. Each phase associated to a set of methods. They view the candidate methods in each phase as the states of Q-Learning. The classification performance of the pipeline representing the reward. This team used Q-Learning to find the pipeline with the maximum reward. To further improve efficiency and robustness of the proposed method, they integrated meta learning and the ensemble learning into the method. Meta learning was used first to initialize the values of Q-Table for Q-Learning. Then, after the Q-Learning, the good discovered pipelines where ensembled with a stacking method.

thanhdng The solution bt thanhdng was based on the ensemble solution provided as **starting kit** for the competition. Basically, this team adjusted the parameters of the ensemble (increasing the number of learning cycles and estimators).

F Result tables of all 30 dataset of the 2015/2016 challenge

In this appendix, we provide result tables on which several graphs are based. In Table 1, we reran the code of the participants who made it available to us on all the datasets of the 2015/2016 challenge (the last version of code submitted to the challenge platform). In Figure 14 , we reran again these codes to compute their error bars with bootstrapping. In Tables 2 and 3, we ran four "basic models" with default hyper-parameter settings and with hyper-parameter optimization on all the datasets of the 2015/2016 challenge.

Table 1: **Systematic study of participants' methods:** The team abbreviations are the same as in the previous table. The colors indicate the rounds.

Datasets	aad	abhi	djaj	ideal	jrl44	lisheng	marc	ref
ADULT	0.82	0.82	0.81	0.83	0.81	0.8	0.81	0.82
CADATA	0.8	0.79	0.78	0.81	0.09	0.79	0.64	0.76
DIGITS	0.95	0.94	0.83	0.96	0.73	0.95	0.86	0.87
DOROTHEA	0.66	0.87	0.82	0.89	0.82	0.84	0.79	0.7
NEWSGROUPS	0.48	0.46	0.64	0.59	0.33	0.05	0.38	0.56
CHRISTINE	0.49	0.46	0.48	0.55	0.48	0.46	0.45	0.42
JASMINE	0.63	0.61	0.62	0.65	0.62	0.61	0.56	0.56
MADELINE	0.82	0.59	0.64	0.81	0.57	0.58	0.18	0.53
PHILIPPINE	0.66	0.53	0.52	0.72	0.52	0.52	0.45	0.51
SYLVINE	0.9	0.87	0.89	0.93	0.89	0.87	0.83	0.89
ALBERT	0.38	0.32	0.36	0.37	0.32	0.34	0.35	0.32
DILBERT	0.94	0.79	0.75	0.98	0.21	0.24	0.46	0.79
FABERT	0.36	0.19	0.33	0.35	0.03	0.18	0.21	0.24
ROBERT	0.46	0.33	0.33	0.51	0.21	0.4	0.37	0.36
VOLKERT	0.33	0.26	0.28	0.37	0.11	0.15	0.14	0.25
ALEXIS	0.75	0.65	0.67	0.76	0.62	0.68	0.62	0.64
DIONIS	0.9	0.32	0.75	0.93	0.02	0.87	0.81	0.31
GRIGORIS	0.73	0.76	0.8	0.97	0.54	0.88	0.96	0.75
JANNIS	0.55	0.38	0.41	0.42	0.24	0.36	0.39	0.4
WALLIS	0.71	0.63	0.74	0.71	0.12	0.23	0.58	0.62
EVITA	0.59	0.59	0.58	0.61	0.59	0.59	0.52	0.41
FLORA	0.5	0.51	0.5	0.53	0.02	0.42	0.51	0.37
HELENA	0.22	0.23	0.15	0.25	0.06	0.2	0.19	0.08
TANIA	0.47	0.76	0.39	0.73	0.53	0.6	0.66	0.54
YOLANDA	0.32	0.37	0.29	0.39	0.02	0.24	0.19	0.26
ARTURO	0.75	0.8	0.78	0.77	0.3	0.72	0.7	0.77
CARLO	0.45	0.37	0.43	0.18	0.36	0.4	0.37	0.14
MARCO	0.55	0.71	0.69	0.54	0.66	0.54	0.68	0.25
PABLO	0.3	0.29	0.31	0.27	0.03	0.29	0.25	0.28
WALDO	0.59	0.56	0.57	0.61	0.56	0.56	0.46	0.56

Table 2: Performances (original task metrics) of basic models using their **scikit-learn default HP setting**. All negative scores and NaN (due to the fact that algorithm didn't succeed in generating predictions within time limit) are brought to zero.

Rnd	DATASET	KNN	NAIVE BAYES	RANDOMFOREST	SGD(LINEAR)
0	ADULT	0.66±0.01	0.72±0.01	0.786±0.009	0.74±0.01
0	CADATA	0.08±0.03	0.62±0.03	0.73±0.02	0.0±0.0
0	DIGITS	0.661±0.007	0.252±0.007	0.924±0.004	0.758±0.007
0	DOROTHEA	0.01±0.04	0.02±0.06	0.4±0.2	0.5±0.2
0	NEWSGROUPS	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
1	CHRISTINE	0.39±0.07	0.36±0.06	0.39±0.06	0.17±0.04
1	JASMINE	0.54±0.05	0.32±0.06	0.58±0.06	0.54±0.07
1	MADELINE	0.57±0.05	0.17±0.05	0.41±0.05	0.0±0.0
1	PHILIPPINE	0.23±0.04	0.36±0.04	0.46±0.05	0.23±0.03
1	SYLVINE	0.52±0.03	0.78±0.02	0.86±0.01	0.55±0.02
2	ALBERT	0.11±0.02	0.0±0.0	0.19±0.02	0.0±0.0
2	DILBERT	0.0±0.0	0.0±0.0	0.01±0.04	0.0±0.0
2	FABERT	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
2	ROBERT	0.1±0.02	0.16±0.02	0.29±0.02	0.22±0.02
2	VOLKERT	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
3	ALEXIS	0.002±0.001	0.38±0.01	0.001±0.001	0.42±0.01
3	DIONIS	0.02±0.01	0.017±0.009	0.033±0.009	0.0±0.01
3	GRIGORIS	0.04±0.02	0.0±0.0	0.0±0.02	0.62±0.03
3	JANNIS	0.13±0.02	0.29±0.04	0.32±0.01	0.22±0.01
3	WALLIS	0.21±0.02	0.04±0.01	0.34±0.02	0.39±0.02
4	EVITA	0.32±0.06	0.35±0.07	0.18±0.05	0.28±0.07
4	FLORA	0.42±0.04	0.43±0.04	0.29±0.04	0.0±0.0
4	HELENA	0.082±0.009	0.14±0.01	0.15±0.01	0.034±0.006
4	TANIA	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
4	YOLANDA	0.0±0.0	0.24±0.01	0.0±0.0	0.0±0.0
5	ARTURO	0.03±0.02	0.35±0.03	0.49±0.03	0.68±0.03
5	CARLO	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
5	MARCO	0.0±0.004	0.007±0.002	0.0006±0.0003	0.04±0.01
5	PABLO	0.09±0.01	0.25±0.01	0.25±0.01	0.15±0.01
5	WALDO	0.03±0.03	0.23±0.03	0.49±0.04	0.03±0.03

G Learning Curve of all 30 datasets of the 2015/2016 challenge

In this appendix we show learning curves on all 30 datasets for two top ranking methods: **auto-sklearn** (aad_freiburg), as a representative of a Bayesian search method and **abhishek** as a representative of a heuristic method. In all figures (Figures 1-6), we represent in yellow the learning curve of auto-sklearn within the time budget of the challenge; they are prolongated in green beyond the time budget. We represent in blue the learning curves of abhishek (it was not trivial for us to modify the code of abhishek to extend the learning curves beyond the

Table 3: Performances (original task metrics) of basic models using **auto-sklearn-tuned HP setting**. The time limit has been respected for this tuning. All negative scores and NaN (due to the fact that algorithm didn't succeed in generating predictions within time limit) are brought to zero.

Rnd	DATASET	KNN	NAIVE BAYES	RANDOMFOREST	SGD(LINEAR)
0	ADULT	0.748±0.009	0.74±0.01	0.808±0.007	0.777±0.009
0	CADATA	0.48±0.03	0.0±0.0	0.48±0.03	0.52±0.02
0	DIGITS	0.0±0.0	0.59±0.009	0.933±0.004	0.802±0.007
0	DOROTHEA	0.4±0.2	0.4±0.2	0.0±0.0	0.5±0.2
0	NEWSGROUPS	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
1	CHRISTINE	0.47±0.05	0.41±0.06	0.49±0.07	0.5±0.05
1	JASMINE	0.6±0.05	0.52±0.05	0.63±0.06	0.58±0.05
1	MADELINE	0.81±0.03	0.22±0.05	0.76±0.03	0.21±0.05
1	PHILIPPINE	0.55±0.04	0.39±0.04	0.58±0.03	0.45±0.04
1	SYLVINE	0.91±0.01	0.81±0.02	0.89±0.01	0.85±0.01
2	ALBERT	0.0±0.0	0.0±0.0	0.0±0.0	0.25±0.02
2	DILBERT	0.34±0.09	0.0±0.0	0.29±0.09	0.0±0.0
2	FABERT	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
2	ROBERT	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
2	VOLKERT	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
3	ALEXIS	0.0±0.0	0.42±0.01	0.11±0.01	0.267±0.008
3	DIONIS	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
3	GRIGORIS	0.0±0.02	0.55±0.03	0.0±0.02	0.0±0.0
3	JANNIS	0.34±0.02	0.24±0.02	0.33±0.01	0.32±0.04
3	WALLIS	0.26±0.02	0.26±0.02	0.34±0.02	0.18±0.01
4	EVITA	0.2±0.05	0.0±0.0	0.27±0.05	0.15±0.05
4	FLORA	0.0±0.002	0.0±0.0	0.0±0.0	0.0±0.001
4	HELENA	0.14±0.01	0.17±0.01	0.0±0.006	0.0±0.0
4	TANIA	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
4	YOLANDA	0.24±0.01	0.0±0.0	0.24±0.01	0.24±0.01
5	ARTURO	0.66±0.03	0.65±0.03	0.75±0.03	0.51±0.03
5	CARLO	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
5	MARCO	0.0±0.0	0.3±0.04	0.0±0.0	0.0±0.0
5	PABLO	0.25±0.01	0.0±0.0	0.25±0.01	0.25±0.01
5	WALDO	0.45±0.03	0.27±0.03	0.55±0.04	0.35±0.03

time budget). The scores are computed using the task-specific metrics of the challenge.

We noticed that in about 2/3 of the cases, abhishek's learning curves start quite high but do not improve very much over time, they even sometimes go down, which may be an indication of overfitting. In about 80% of the cases, aad_freiburg's learning curves start lower that the learning curves of abhishek. Hence, in spite of their use of meta-learning, aad_freiburg did not come up with as good heuristic startign points. However, their hyper-parameter search is more efficient: in about 1/2 of the cases, they end up higher at the end of the learning curve, within the time budget; in about 80% they end up higher if let run longer (green part of the curve).

These learning curves show that there is still a large margin for improvement in terms of combining techniques.

We also show in Figures 7-13 all learning curves of a given round ovelaid for the same two high ranking participants ('aad_freiburg' (solid-dots) and 'abhishek' (solid-empty square)). This representation shows that the two optimization strategies differ in their management of time. The 'aad_freiburg' made use of parallelism. Since 4 cores were available on the computers used for the challenge, they started working on 4 (out of 5) datasets simultaneously and started on the fifth one by interrupting working on one of the other datasets, or interleaving work. In contrast, 'abhishek' processed one dataset after the other. For ease of visualisation, we connect the learning curves of 'abhishek' on the various datasets with a dashed line. The error bars were estimated by bootstrapping.

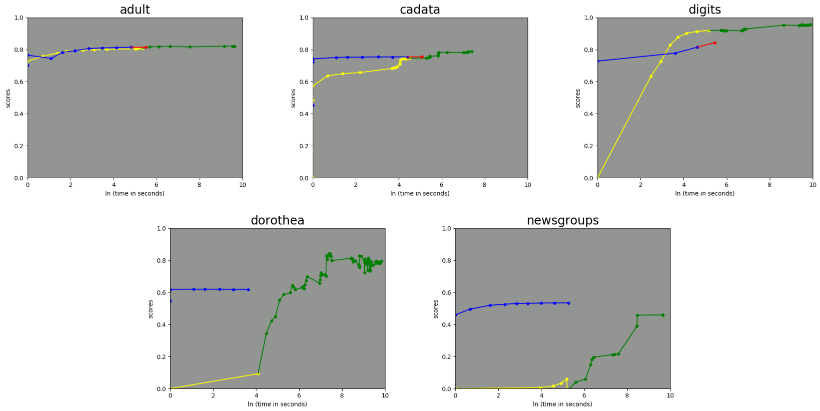

Fig. 1: **Learning Curve of 'aad_freiburg' (yellow+green) and 'abhishek' (blue+red) for Round 0.**

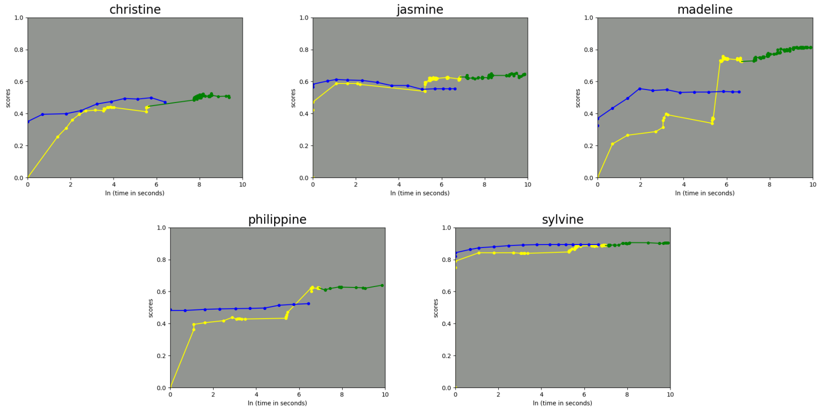

Fig. 2: **Learning Curve of 'aad_freiburg' (yellow+green) and 'abhishek' (blue) for Round 1.**

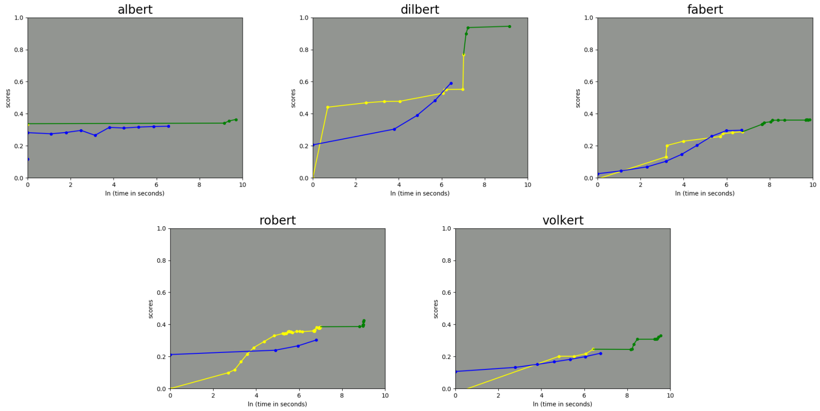

Fig. 3: **Learning Curve of 'aad_freiburg' (yellow+green) and 'abhishek' (blue) for Round 2.**

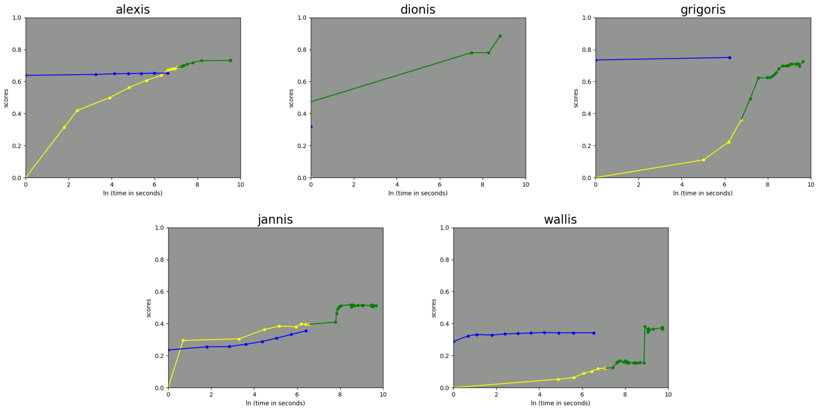

Fig. 4: **Learning Curve of 'aad_freiburg' (yellow+green) and 'abhishek' (blue) for Round 3.**

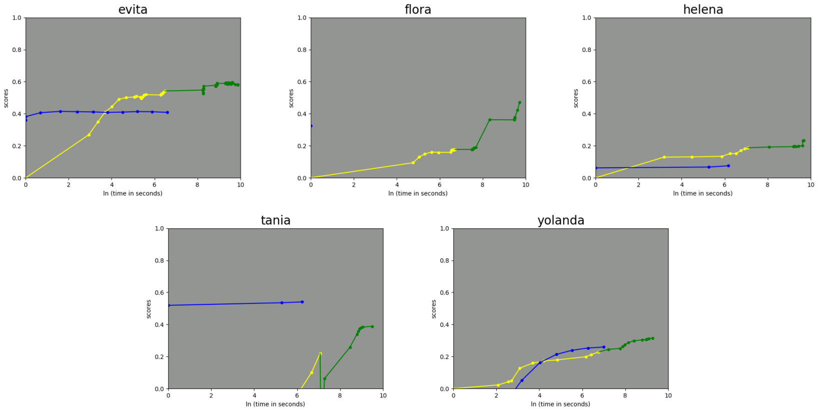

Fig. 5: **Learning Curve of 'aad_freiburg' (yellow+green) and 'abhishek' (blue) for Round 4.**

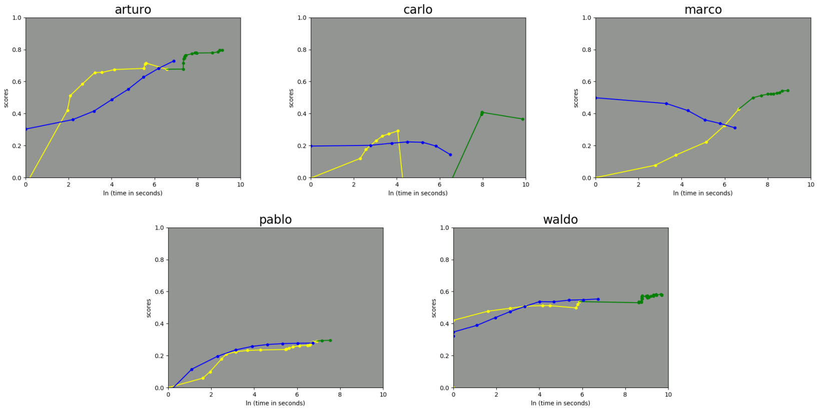

Fig. 6: **Learning Curve of ‘aad_freiburg’ (yellow+green) and ‘abhishek’ (blue) for Round 5.**

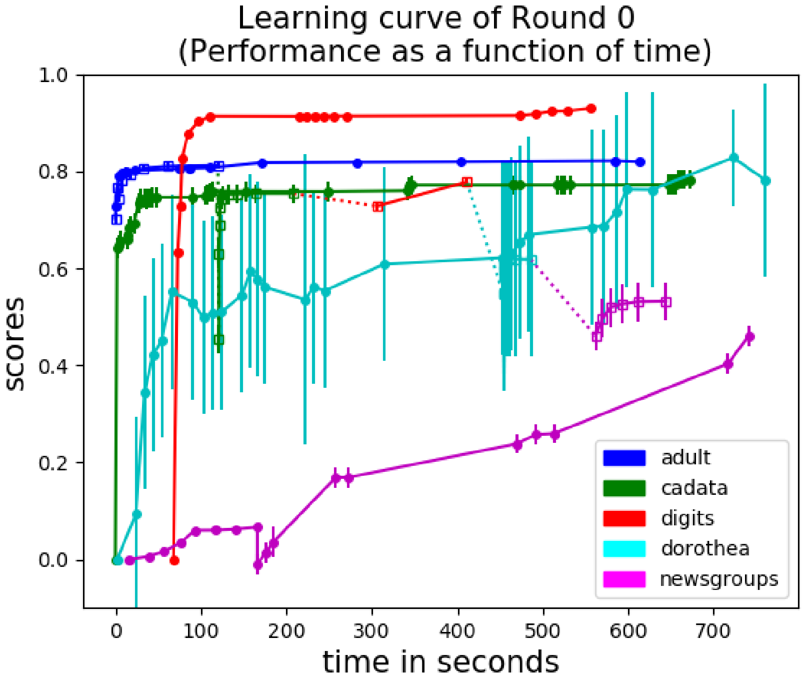

Fig. 7: **Learning Curve of ‘aad_freiburg’ (solid-dots) and ‘abhishek’ (solid-empty square) for Round 0.**

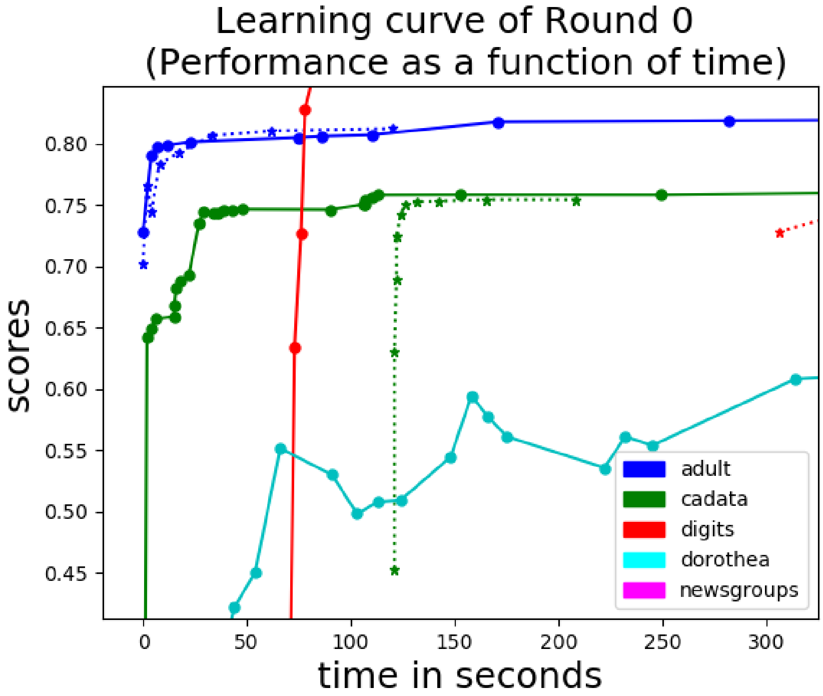

Fig. 8: **Partial magnification of Figure 7**

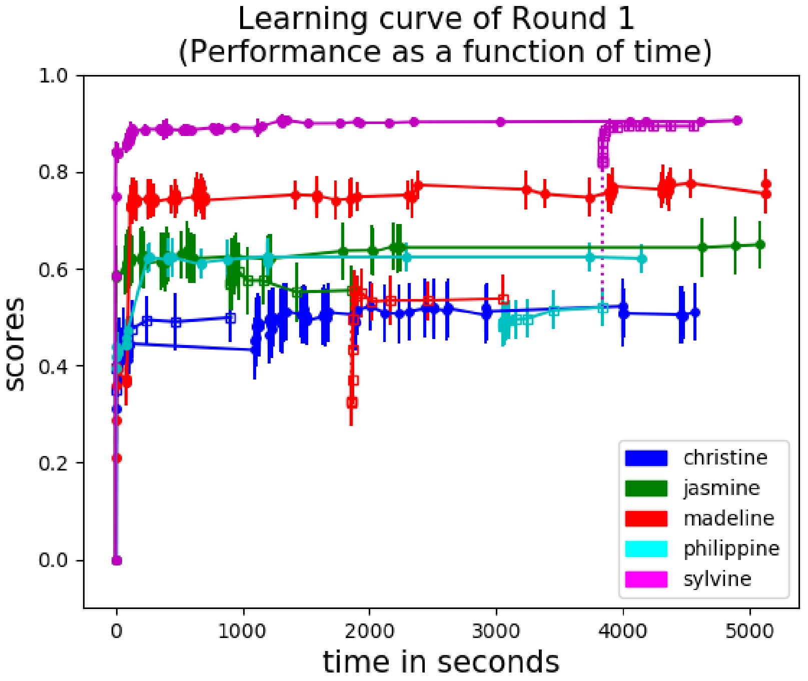

Fig. 9: **Learning Curve of 'aad_freiburg' (solid-dots) and 'abhishek' (solid-empty square) for Round 1.**

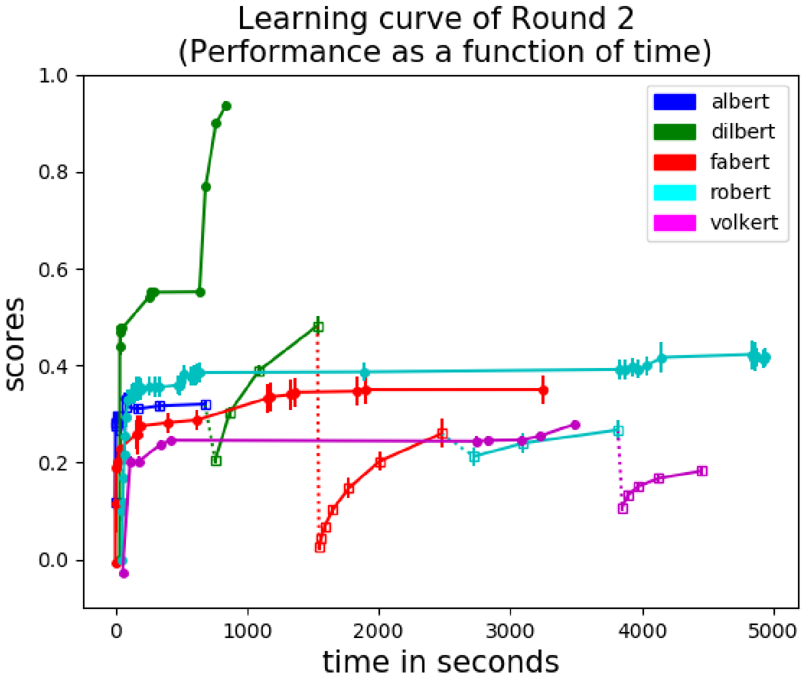

Fig. 10: **Learning Curve of 'aad_freiburg' (solid-dots) and 'abhishek' (solid-empty square) for Round 2.**

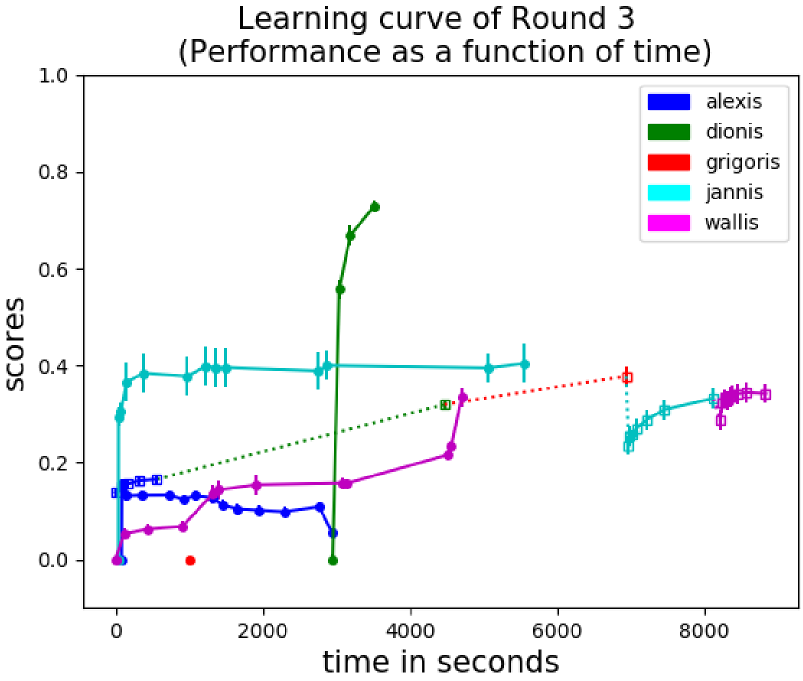

Fig. 11: **Learning Curve of 'aad_freiburg' (solid-dots) and 'abhishek' (solid-empty squares) for Round 3.**

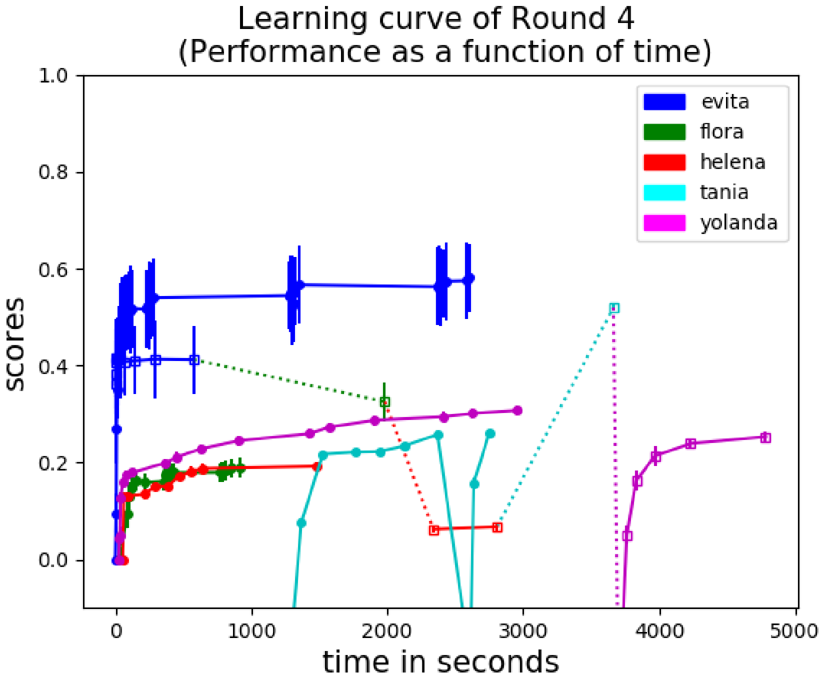

Fig. 12: **Learning Curve of 'aad_freiburg' (solid-dots) and 'abhishek' (solid-empty square) for Round 4.**

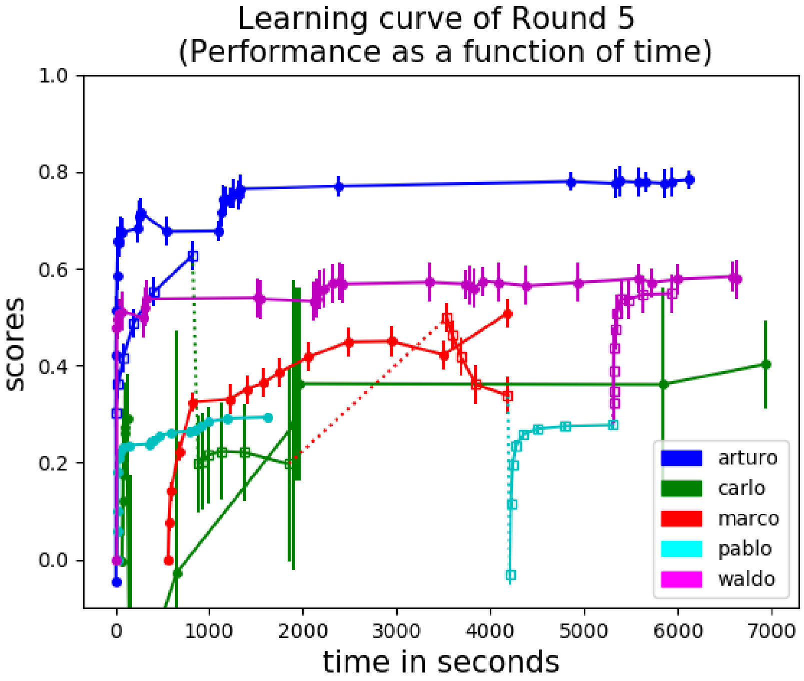

Fig. 13: **Learning Curve of 'aad_freiburg' (solid-dots) and 'abhishek' (solid-empty squares) for Round 5.**

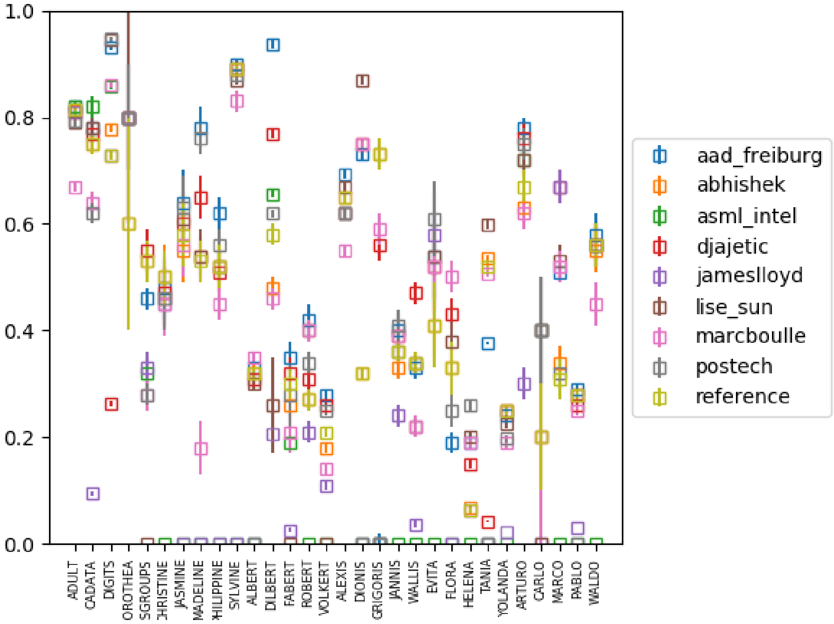

Fig. 14: **Scores of participants' methods with error bars.**

Bibliography

[1] Boullé, M.: Compression-based averaging of selective naive bayes classifiers. Journal of Machine Learning Research 8, 1659–1685 (2007), http://dl.acm.org/citation.cfm?id=1314554

[2] Boullé, M.: A parameter-free classification method for large scale learning. Journal of Machine Learning Research 10, 1367–1385 (2009), http://doi.acm.org/10.1145/1577069.1755829

[3] Chollet, F.: Keras. https://github.com/fchollet/keras (2015)

[4] Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J.D., Heilman, M., diogo149, McFee, B., Weideman, H., takacsg84, peterderivaz, Jon, instagibbs, Rasul, D.K., CongLiu, Britefury, Degrave, J.: Lasagne: First release. http://dx.doi.org/10.5281/zenodo.27878 (August 2015)

[5] Falkner, S., Klein, A., Hutter, F.: BOHB: Robust and efficient hyperparameter optimization at scale. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1437–1446. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018), http://proceedings.mlr.press/v80/falkner18a.html

[6] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Proceedings of the Neural Information Processing Systems, pp. 2962–2970 (2015), https://github.com/automl/auto-sklearn

[7] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Methods for improving bayesian optimization for automl. In: Proceedings of the International Conference on Machine Learning 2015, Workshop on Automatic Machine Learning (2015)

[8] Feurer, M., Springenberg, J., Hutter, F.: Initializing bayesian hyperparameter optimization via meta-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1128–1135 (2015)

[9] Hue, C., Boullé, M.: A new probabilistic approach in rank regression with optimal bayesian partitioning. Journal of Machine Learning Research 8, 2727–2754 (2007), http://dl.acm.org/citation.cfm?id=1390332

[10] Hutter, F., Hoos, H., Murphy, K., Ramage, S.: Sequential Model-based Algorithm Configuration (SMAC). http://www.cs.ubc.ca/labs/beta/Projects/SMAC/ (2018)

[11] Jajetic, D.: Djajetic Implementation. https://github.com/djajetic/AutoML5 (2016)

[12] Jajetic, D.: GPU_djajetic Implementation. https://github.com/djajetic/GPU_djajetic (2016)

[13] Lloyd, J.: Freeze Thaw Ensemble Construction. https://github.com/jamesrobertlloyd/automl-phase-2 (2016)

[14] Mendoza, H., Klein, A., Feurer, M., Springenberg, J.T., Hutter, F.: Towards automatically-tuned neural networks. In: ICML 2016 workshop on AutoML (June 2016), https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxhdXRvbWwyMDE2fGd4OjMzYjQ4OWNhNTFhNzlhNGE

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[16] Sun-Hosoya, L.: Automl challenge: System description of lisheng sun. In: ICML 2016 workshop on AutoML (June 2016), http://dx.doi.org/10.5281/zenodo.27878

[17] Swersky, K., Snoek, J., Adams, R.P.: Freeze-thaw bayesian optimization. arXiv preprint arXiv:1406.3896 (2014)

[18] Thakur, A.: AutoML challenge: Rules for selecting neural network architectures for automl-gpu challenge. In: ICML 2016 workshop on AutoML (June 2016), https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxhdXRvbWwyMDE2fGd4OjNmY2M0N2JhZGViZWY3ZDY

[19] Thakur, A., Krohn-Grimberghe, A.: AutoCompete: A framework for machine learning competitions. In: Proceedings of the International Conference on Machine Learning 2015, Workshop on Automatic Machine Learning (2015), https://docs.google.com/a/chalearn.org/viewer?a=v&pid=sites&srcid=Y2hhbGVhcm4ub3JnfGF1dG9tbHxneDo3YThhNmNiNDA0M2Q2NjM5

[20] Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688 (2016)

[21] Tuv, E., Borisov, A., Runger, G., Torkkola, K.: Feature selection with ensembles, artificial variables, and redundancy elimination. Journal of Machine Learning Research 10, 1341–1366 (January 2009)