KDD Cup 2019 Call for Proposals

This Call for Proposals invites industrial or academic institutions to submit their proposals for organizing the 2019 KDD Cup competition. Since 1997, KDD Cup has been the premier annual Data Mining competition held in conjunction with the ACM SIGKDD conference on Knowledge Discovery and Data Mining.

SIGKDD-2019 will take place in Anchorage, Alaska, US from August 3rd-7th 2019. The KDD Cup competition is anticipated to last for 2-4 months, and the winners will be notified by mid-July 2019. The winners will be honored at the KDD conference opening ceremony and will present their solutions at the KDD Cup workshop during the conference. The winners are expected to be monetarily rewarded, with the first prize being in the ballpark of ten thousand dollars.

We are looking for strong proposals that meet the following requirements: a novel and motivated goal, a broad business impact, a rigid and fair setup, a challenging yet manageable task, and domain accessibility to the general public.

1. **A novel and motivated goal.** Of a particular interest are tasks that imply machine learning solutions different from the traditional KDD Cup setting (an ensemble of classifiers is learned on a given training set to obtain a high-quality classification result on a held-out test set). Examples of non-traditional setups would be incrementally arriving data and evaluation on the accumulated error; prediction given a limited amount of resources; learning with mostly unlabeled data; addressing cold-start issues in learning; learning over multiple types of data; applications of deep learning models, etc.

2. **A broad business impact.** We encourage organizers to ponder on a practical challenge that has a potential to be deployed in a real-world application and get appreciated by millions of customers.

3. **A rigid and fair setup**. The organizers should guarantee the availability of the data and the confidentiality of the test set (to prevent information leakages at any cost). The evaluation metrics should be both meaningful for the application in-hand and statistically sound for the objective comparison. The baseline should be established to show that non-trivial results can be achieved. An estimate of what constitutes a significant difference in the performance will be much appreciated.

4. **A challenging yet manageable task**. The task should be challenging in the sense that there is enough room for improvement from the basic solutions, and novel ideas are required to succeed in the competition. The task should be manageable in about 3 months' time, and the underlying infrastructure should be supported by the organizers such that the competitors could mainly focus on the core challenge.

5. **Domain accessibility**. The notions presented in the competition description should be accessible to the majority of machine learning and data mining practitioners who might not have an excessive domain knowledge or access to a powerful computational infrastructure.

In addition to meeting the requirements above, good proposals might aim to address the following concerns raised over previous KDD Cup competitions:

1. **Model complexity**. Although increasing the complexity of the solutions (usually via an ensemble of multiple models) can improve the accuracy, it makes it harder to interpret or deploy the proposed solutions.
2. **Static test data** - Evaluating the models on static test data motivates participants to squeeze the maximum out of the test data (sometimes by exploiting data leaks), and the solutions might be an overfit.

For KDD Cup 2019, we encourage proposals taking either a regular setting (see previous years' contests for example) or an auto-ml setting (see FAQ for more details). If the auto-ml setting is chosen, the organizers should be able to set-up and validate the ability to do automatic machine learning without human intervention.

We suggest the proposals to answer the following questions:

1. How does the proposed challenge meet the five requirements?
2. How does the proposed challenge address the two concerns?
3. Which format do you plan to use, a regular competition or an auto-ml competition?
4. Which competition infrastructure do you plan to use (e.g., Kaggle, or on your own)? Is the competition platform you chose equally accessible to participants all over the world?
5. What resources (including people, time, and award money) do you plan to invest?
6. What is your time schedule for the competition?
7. Is there any concern of the privacy about the released data? Have you obtained the rights to release the data for the competition from your legal counsels?
8. Do you require the winners to submit the source code of their winning solutions?
9. How would you handle Q&A and possible revisions during the competition?
10. What is your baseline solution?

and also include:

11. Names, affiliations, email addresses, phone numbers, and short biographies of the organizers.
12. An endorsement letter from the executive-level management of your organization.

Please keep the proposal concise and strictly confidential. Please send your proposals in the PDF format to kddcup2019@kdd.org by January 10th, 2019.

Important dates:
December 4th, 2018 - CFP release
January 10th, 2019 - Proposal submission deadline
February 10th, 2019 - Decision notification
April 2nd, 2019 – Tentative start of the competition
July 20th, 2019 - Announcement of the KDD Cup Winner

Appendix | Auto-ml FAQ | KDD Cup 2019 Call for Proposals

What is new in KDD Cup 2019?

In this year's KDD Cup we added auto-ml as a new format for the competition.

What is auto-ml?

Auto-ml stands for automatic machine learning, which aims at automating the end-to-end process of machine learning, including feature extraction, algorithm selection, parameter tuning and model interpretability.

How does an auto-ml competition work?

Auto-ml competition will have 2 phases: feedback phase and auto-ml phase.

Feedback Phase

During the feedback phase organizers will provide competitors with 5 different problem data-sets (each with Train and Test data) and solutions will be scored on all the data-sets. The scoreboard will rank solutions based on the average or median ranking of scores achieved on the 5 data-sets. Also, the competitors will upload their workable code in Python, R, or other programming language/tool of the organizers' choice.

#	User	Entries	Date of Last Entry	Team Name	<Rank> ▲	Set 1 ▲	Set 2 ▲	Set 3 ▲	Set 4 ▲	Set 5 ▲
1	deepsmart	38	11/04/18	DeepSmart	1.2000	0.5614 (1)	0.3489 (2)	0.6216 (1)	0.6027 (1)	0.8112 (1)
2	HANLAB	60	11/04/18	MIT HAN LAB	3.0000	0.5344 (5)	0.3372 (4)	0.5815 (2)	0.5676 (2)	0.7848 (2)
3	Fong	50	11/04/18	Nanjing University PASA Lab	4.2000	0.5370 (4)	0.3356 (5)	0.5806 (3)	0.5561 (5)	0.7795 (4)
4	MI-Intelligence	3	11/04/18	MI-Intelligence	4.4000	0.5456 (2)	0.3539 (1)	0.4874 (10)	0.5443 (6)	0.7829 (3)

Auto-ml Phase

During this phase the organizer will run the code submitted by competitors to score a different data-set. Solutions capable of interpreting model results will be given preference in case there is a tie in final score.

Competitors of the top solutions will get an opportunity to submit a write-up on their strategy of auto-ml. They can also submit a video tutorial that could be leveraged to improve the overall learning in this area.

KDD Competition 2019

	Regular Competition	**Auto-ml Competition**		
Number of Problems	One problem	Five different problems	Phase 1 (Feedback Phase)	April 1st - June 1st
Data-Sets	Train and Test	5 different Train and Test		
Submission Files	one	one		
Public Leaderboard	Single leaderboard based on the code submitted.	Single leaderboard with median of 5 scores submitted.		
Code submission	Not needed	Mandatory for Auto-ml (Python and R only)		
Auto-ml Computation		Auto-ml code run on a final dataset by the competition hosts. This will done based on the code submitted by competitiors	Phase 2 (Auto-ml Blind)	June 15th - July 15th
Final Leaderboard	Leaderboard based on the truth dataset for the same problem	Leader board based on code run on the new problem	Results	July 20th

Are there examples of auto-ml competitions?

Yes. See for example: https://competitions.codalab.org/competitions/20203#phases

Instead of auto-ml, can we still adopt a regular competition format?

Yes. Regular competitions will follow the same approach as prior years with competitors provided with Train and Test data for an use case. Organizers would use a metric such as AUC to determine the winner. Code submission will be needed only for the winners.