The KDD Cup 2019 Report

Wenjun Zhou
University of Tennessee
916 Volunteer Blvd.
Knoxville, TN 37996
wzhou4@utk.edu

Taposh Dutta Roy
Kaiser Permanente
2 San Pedro Pl
San Ramon, CA 94583
taposh.dr@gmail.com

Iryna Skrypnyk
Pfizer
37W 42nd St.
New York, NY 10017
iskrypn@gmail.com

ABSTRACT

The KDD Cup has been a data science competition affiliated with the ACM SIGKDD conference with more than 20 years' tradition. In 2019, we organized the KDD Cup by hosting 3 parallel tracks, each with tremendous innovation. The regular machine learning (ML) track was a context-aware travel mode recommendation problem, sponsored by Baidu.com. The automatic machine learning (AutoML) track, sponsored by 4Paradigm, was about finding cost-effective and transferable solutions for temporal relational data represented as multiple related tables. The humanity reinforcement learning (RL) track was sponsored by IBM Africa and Hexagon-ML to determine the best policy in distribution of control measures to eradicate Malaria. In 3 competitions collectively, we had more than 2,800 registered teams from over 39 countries and 230 academic and research institutions. Among the 1,200 most actively participating teams, over 5,000 individuals participated, and more than 17,000 submissions were made. A total exceeding 100 thousand U.S. dollars were awarded to the winning teams.

Keywords

SIGKDD, data science competition, competition organization, context-aware transportation mode recommendation, AutoML, humanity, reinforcement learning.

1. INTRODUCTION

Challenging a community of experts and enthusiasts to solve a scientific problem has been around for hundreds of years. One of the first recorded challenges like this was established in 1714 by British Board of Longitude. The prize was awarded to the person who could solve what was arguably the most important technological problem of the time: to determine the longitude of a ship at sea [11]. This prize was motivated after a disaster in 1707 at Scilly where 4 ships were drowned. The Board administered prizes for those who could demonstrate a working device or method. The max prize for a method that could determine longitude within 30 nautical miles (56 km; 35 mi) (£2,600,000 as of 2015) [12].
SIGKDD, an ACM Special Interest Group on Knowledge Discovery and Data Mining, has pioneered data science competition using large-scale datasets since 1998. The annual competition series, integrated with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), has been known as "The KDD Cup." This year's KDD Cup is celebrating 22 years. It has been an exciting journey and we have come a long way.
We felt extremely honored to serve as co-chairs of the KDD Cup 2019. After being tasked with its organization, we had countless meetings to prepare for the KDD Cup Day at the KDD 2019 conference. We did comprehensive survey and analysis of the industry's technology and data science innovation forefronts. Based on the analysis we set the criteria in the call for KDD's 2019 data science competition. Some of the important criteria we considered when selecting the competition problems included:

- novelty of the task and its attractiveness to the data science community;
- availability of a baseline solution; that is, how well the data have been studied internally;
- knowledge and description of specifics of the problem, such as cold-start in machine learning or concept drift;
- ability in providing full specifications of data (data structure, attributes, size, etc.), problem formulation, and evaluation metrics;
- proportions of engineering, data science, machine learning work to solve the problem;
- how easy it is to understand the domain and/or task to the participants;
- how well the domain might be accessible for all participants around the world;
- availability of resources and support, including computational resources, platform, human resources, and funds (some of these resources may be provided by an external sponsor when applicable) during and after the competition.

After seeing a record number of fantastic competition proposals, we decided to run 3 tracks in parallel. For one thing, since there were so many great proposals, we did not want to turn down so many. For another, we wanted to experiment with having a variety of types of competitions, especially the ones that are quickly picking up in industrial applications. Therefore, for KDD Cup 2019 we had the following 3 tracks:

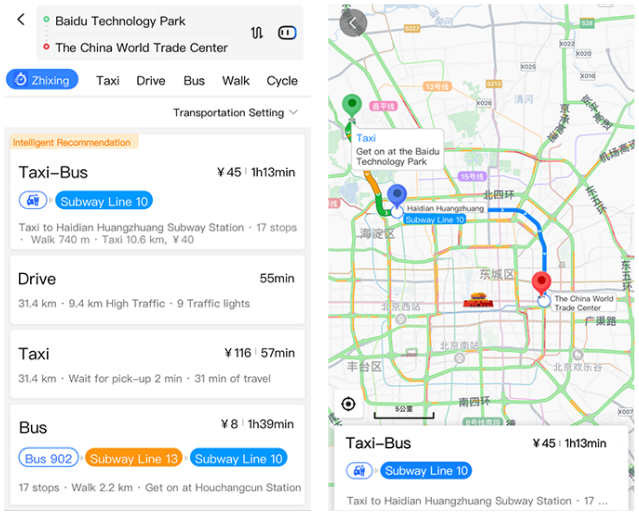
- The **Regular ML** track was a context-aware travel mode recommendation problem, sponsored by Baidu.com. Similar to prior years, it aims at solving

a novel real-world problem that seeks an efficient and effective ranking algorithm.

- The **AutoML** track, sponsored by 4Paradigm, was about automated discovery of useful information in temporal relational data streams represented as cross-linked tables leveraging commonalities in data from different sources.

- The **Humanity RL** track was sponsored by IBM Africa and Hexagon-ML to determine the best policy in distribution of control measures to eradicate Malaria.

In 3 competitions collectively, we had more than 2,800 registered teams from over 39 countries and 230 academic and research institutions. Among the 1,200 most actively participating teams, over 5,000 individuals participated and more than 17,000 submissions were made. A total exceeding 100K U.S. dollars were awarded to the winning teams.
In the rest of this report, we document information about each of the competitions and highlights of the KDD Cup Day that was held with the KDD 2009 conference at Anchorage, Alaska in August 2019.



(a) Recommendation list. (b) Details of the first route.

Figure 1: An example of the graphical user interfaces of the CAMMT recommendation service on Baidu Maps

2. THE REGULAR ML TRACK

The Regular ML track focused on developing intelligent transportation solutions taking into consideration travel contexts including the origin-destination pairs, the traveler demographics, and other contexts. Like most of the past KDD Cup competitions, this machine learning task was formulated as a prediction or ranking problem. A second task calling for research proposals was also new this year.

Sponsor: Baidu.com
Organizing Committee: Hui Xiong, Hao Liu, Hengshu Zhu, Shengwen Yang, Yuecheng Rong, and Zecheng Zhuo

2.1 Description

In early 2018, Baidu Maps introduced the context-aware multi-modal transportation (CAMMT) recommendation service. As illustrated in Figure 1, the left figure shows the travel plan, which includes a list of several different transportation modes. The ordering was determined by Baidu's recommendation model published in [15]. The right figure visualizes the travel plan that the user selected on the map. The first recommended plan, which is multimodal (i.e., first take taxi and then bus), is 26.3% faster than the third plan (bus only) and 61.2% cheaper than the second plan (taxi only). The user's choice would depend on their preference striking a tradeoff between convenience and time and monetary costs. Therefore, even for the same origin-destination pair, the travel plans may be ranked differently for different users, at different times of day, and even depend on weather.
In the year prior to this competition, Baidu Maps' CAMMT recommendation engine served more than one hundred million route planning requests from over ten million distinct users. As the competition sponsor, Baidu would like to challenge the contestants to further improve the accuracy and efficiency of the recommendation engine.
Despite the popularity of transportation recommendation on navigation Apps (e.g., Baidu Maps and Google Maps), existing transportation recommendation solutions only consider routes in one transportation mode. However, transport mode preferences may vary over different users and spatiotemporal contexts. Context-aware multi-modal transportation recommendation has the goal of crafting a travel plan that considers a mix of various transportation modes, such as walking, cycling, driving, and public transit, and their connections. Developing multi-modal transportation recommendations has a number of advantages, including but not limited to reducing transit time, balancing traffic flows, reducing traffic congestion, and ultimately, promoting the development of intelligent transportation systems.

2.1.1 Tasks and Evaluation

The Regular ML competition was organized into 3 phases. In Phase 1, a training set (covering a 2-month period from October 1st to November 30th, 2018) and a testing set (covering the week right after, i.e., December 1-7, 2018) were provided. Both datasets were collected from Beijing, one of the largest cities in the world. In Phase 2, data covering more cities in China were provided. As one would expect a good algorithm to work well on out-of-sample datasets, the data releases specifically indicated the time evolution nature and between-city variability. While the first two phases allowed participants to download data and play with them on their own computing hardware, Phase 3 required them to test their algorithms in an online testing environment.
For evaluation, a weighted F1 score was used as the evaluation metric in Phase 1. Additional efficiency metric and memory consumption cost were considered in Phase 3. A final documentation was required to ensure the interpretability of the solution.
Besides the main task (Task 1), which followed the classic format of a data science contest, the Regular ML track also has a Task 2, which was an open call for research proposals using the same dataset provided. A research or application proposal is required to explain: 1) the proposed research or application topic, including the aim of the research or application, the context and the justification; 2) the general

methodology and expected resources and conditions to complete the research or application; and 3) the importance of your research or application topic.
More details about this competition may be found on its official website [1].

2.1.2 Technology Highlights

The main competition was hosted on Baidu's Dianshi platform. A Linux Container was provided as the computational resources for each team with the following configuration:

- CPU: 12 Cores
- GPU: 1 NVIDIA Tesla P4
- Memory: 32 GB
- Disk: 100GB

The container had pre-installed Jupyter Lab machine learning suite, supporting both Python 2 and Python 3 (see Table 1).

Table 1: Deep learning frameworks available on Dianshi

Framework	Python 2.7	Python 3.6
PaddlePaddle	1.4.0.post97	1.4.0.post97
Torch	1.1.0	1.1.0
Mxnet	1.4.1.post0	1.4.1.post0
Cntk	2.7	2.7
Chainer	5.4.0	5.4.0
Caffe	1.0.0	1.0.0
Tensorflow	1.10.1	1.10.0
Keras	2.2.4	2.2.4

Additionally, free high-performance computing resources were provided for PaddlePaddle participants. Participants who use PaddlePaddle can log into the Baidu AI Studio, Baidu's one-stop deep learning development platform, to obtain a free computing resource as Tesla V100 for model training remotely. Each person can get up to 120 hours of computing resources. PaddlePaddle baseline code was shared on Github [2] to facilitate participants with a quick start interacting with the PaddlePaddle platform.

2.2 Winners

Task 1 honored the top 10 winning teams. The top 3 winning teams were:

- First Prize ($10,000): Shiwen Cui, Changhua Meng, Can Yi, Weiqiang Wang, Xing Zhao, and Long Guo from Ant Financial Services Group;
- Second Prize ($5,000): Hengda Bao from Shanghai Weimob Enterprise Development Co.Ltd., Jie Zhang from Trend Micro, Wenchao Xu from Didi Chuxing-Map Department, Qiang Wang from Beijing University of Posts and Telecommunications, Jiayuan Xie from South China University of Technology, He Wang from JD.COM, and Ceyuan Liang from JD.COM;
- Third Prize ($3,000): Hua Zhixiang and Sangyu from JIANGLI;

and the following teams, ranked from the 4th to the 10th, won the Honorable Prizes ($1,000 each):

- 4th Place: Yang Liu from Southeast University, Fanyou Wu from Purdue University, Shan Zhang from Hebei University of Technology;
- 5th Place: Jianfei Huang, Peng Yan, Huan Chen, Xiaowei Shi, and Zhen Chen from Meituandianping;
- 6th Place: Zhipeng Luo, He Yan, and Chen Chen from DeepBlue Technology (Shanghai) Co., Ltd and Haibin Zhang;
- 7th Place: Jiangwei Luo, Shiji Qiao from SF-Technology, Xu Cheng from China Mobile, Zhimin Lin from Chongqing University of Posts and Telecommunications, Ruifeng Qian from Jiangnan University;
- 8th Place: Runxing Zhong from 4Paradigm Co. Ltd, Ziwen Ye from Beijing Forestry University, Yuanfei Luo from 4Paradigm Co. Ltd, and Mengjiao Bao from Beihang University;
- 9th Place: Zhangming Niu from Mind Rank AI & Aladdin Healthcare Technologies, Lao Li from Cisco Systems, Inc., Jiangshui Hong from Simula Research Laboratory, Norway, Binli Luo from Mind Rank Limited, Yinghui Jiang from Mind Rank AI & Hangzhou Ocean's Smart Boya Technology Co., Ltd, Ying Song from Sun Yat-sen University, An Xu from Tencent Computer System Co. Ltd., Qiang Li from Alibaba Group Holding Limited, Zhifeng Gao from Peking University, Wei Cao from Tsinghua University;
- 10th Place: Xin Chen from Netease Game, Changsheng Zhong from Guangzhou can-dao Technology Co., Ltd., Wenbin You from Guangzhou can-dao Technology Co., Ltd., Zhongjian Lv from Microsoft, Zhao Yin from Inspur Group.

Task 2 honored the top 3 winning teams, and a special award that utilizes the PaddlePaddle platform.

- First Prize ($5,000): Keiichi Ochiai, Tsukasa Demizu, Shin Ishiguro, Shohei Maruyama, and Akihiro Kawana from NTT DOCOMO, INC. Their research proposal is titled "*Simulating the Effects of Eco-Friendly Transportation Selections for Air Pollution Reduction.*"
- Second Prize ($3,000): Yang Liu, Cheng Lyu, and Zhiyuan Liu from School of Transportation, Southeast University. Their research proposal is titled "*Interdisciplinary Knowledge and Experience Fusion in Multi-Modal Transportation Recommendation System.*"
- Third Prize ($2,000): Xin Wei, Nanlin Liu, Yuan Chen, Xiaopei Liu, Tao Wang, Shijun Mu, Hongke Zhao, and Xi Zhang from College of Management and Economics, Tianjin University and College of Civil and Environmental Engineering, University of Alberta. Their research proposal is titled "*How to Build 'Age-friendly' Cities: Based on Big Data from Baidu Map.*"
- The PaddlePaddle Special Award ($4,000) winners were: Xianfeng Liang, Likang Wu, Joya Chen, Yang Liu, Runlong Yu, Min Hou, Han Wu, Yuyang Ye, Qi Liu, and Enhong Chen, a research team from the University of Science and Technology of China. Their research proposal is titled "*Long-term Joint Scheduling for Urban Traffic.*"

2.3 Impact and Outcome

The Regular ML track has attracted the participation of 1,696 teams that includes 2,403 individuals, who made around 5,000 submissions in total.

3. THE AUTOML TRACK

AutoML as a concept emerged a few years ago and quickly resulted in a number of commercial products as well as open source machine learning libraries (SkiLearn, H2O, Google Cloud AutoML, Microsoft Azure AutoML, to name a few). Nearly two decades ago, the process we called Knowledge Discovery from Databases [14] had somewhat similar components as it has today, and the goal was to design tools that automate as much of it as possible. Nowadays, the process has become much more complicated and automation is focusing on its counterparts: data pre-processing, feature engineering, algorithm selection, and hyperparameter optimization. Even with great help from AutoML, human experts are still needed.

AutoML started with automated parameter tuning for a given model/algorithm, quickly progressed into selection of the best performing models with those optimized parameters, especially beneficial for neural architectures, and lately made strides into data pre-processing, feature engineering, and model interpretation. Computational resources to implement AutoML approaches still have expected limits in runtime and memory consumption even in the presence of ever-increasing computing power, smart optimization and scalable machine learning algorithms. Therefore, data scientists need new skills to achieve best results with what AutoML has to offer.

AutoML-focused competitions recently have emerged as a new form of data science competitions, and KDD Cup pioneered this type of competition in 2019 in collaboration with 4Paradigm, Inc. 4Paradigm explored various applications of AutoML and ran a few AutoML competitions in collaboration with ChaLearn and CodaLab platform, including Automated Natural Language Processing (AutoNLP) and Automated Computer Vision (AutoCV). These competitions have drawn a lot of attention from both academic researchers and industrial practitioners. In 2019, the AutoML track aimed to emphasize new challenges and demonstrate how predictive problems can be solved with AutoML.

Sponsor: 4Paradigm, in collaboration with CodaLab, ChaLearn, and Microsoft.

Organizing Committee: Wei-Wei Tu (4Paradigm), Hugo Jair Escalante (ChaLearn), Sergio Escalera (University of Barcelona), Evelyne Viegas (Microsoft Research), Mengshuo Wang, Xiawei Guo, Ling Yue, Jian Liu, Hai Wang, Wenhao Li, Yuanfei Luo, Jingsong Wang, Runxing Zhong, Yadong Zhao, Feng Bin, Xiaojie Yu, Yuanmeng Huang, Shiwei Hu, Yuqiang Chen, and Wenyuan Dai (4Paradigm).

Advisors: Isabelle Guyon (Universt'e Paris-Saclay, France, ChaLearn), Qiang Yang (Hong Kong University of Science and Technology, Hong Kong).

3.1 Description

4Paradigm-sponsored team of competition organizers provided a task dealing with temporal relational data from different sources and application areas having in common only a few things: data represented as cross-linked tables and there is a timestamp component. Data for the different parts can be recorded with timestamps with errors, or at different periods, and otherwise differently, therefore, some inconsistencies may occur when merging data to create input for machine learning.

The task has been based on premises that multiple related tables might contain useful information in form of inter-table interaction that can be exploited to improve machine learning performance and there is a useful temporal information to be extracted. The AutoML challenge greatly emphasized the engineering aspects, focusing on cost and efficiency as well as generalization of the processing scripts onto new sources of data with a temporal relational component. It required a code submission rather than submission of a file with predictions. Submitted solutions were required to perform automated data assemble, transformation, feature engineering and prediction. Limitations were imposed on the memory and runtime to test for efficiency. These settings have been largely driven by practical business applications, for example, to plug data from multiple data partnerships into a commercial analytical platform, while data providers and vendors may come in hundreds and change monthly. In such cases, there is a need for quick and efficient integration, where some parts can be automated. Data provided for competition are typical in online advertising, recommender systems, financial market analysis, medical treatment, fraud detection, and other application areas.

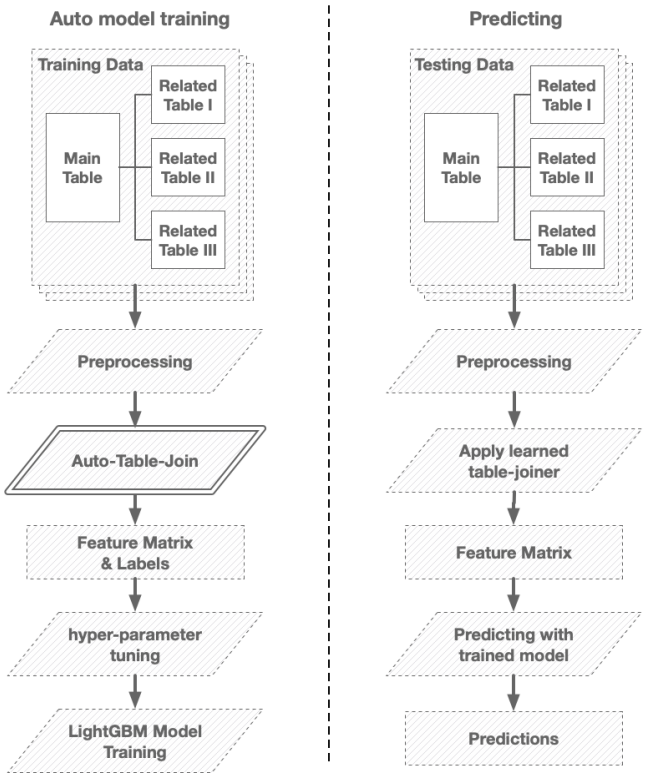


Figure 2: Baseline solution for the AutoML track: process steps

The predictive task took a form of a binary classification problem. Multiple data sources were exemplified by 10 dif-

ferent datasets, each in the form of multiple related tables with a timestamp field. Each dataset was split into training and test sets in a chronological order. Domain context was completely removed from data and substituted by generic names.

The competition encompassed several phases. In the Feedback phase that lasted 2 months and 26 days, participants could build their solutions code, including all necessary data manipulations, parameter tuning, training and testing. This phase let participants make sure their code will run under given restrictions. Each team could perform several submissions of their code per day to see if they pass tests and obtain performance scores on the holdout sets from 5 public datasets shared with participants.

In the following Check phase, which lasted 2 weeks, the code was tested against 5 private, unseen datasets. Participants were given a chance to correct their code if it failed by exceeding time or memory limits with only one chance to resubmit. Model performance was not revealed.

Finally, performance of the models was tested during the last, AutoML phase that lasted for 1 week, on 5 private datasets withheld from participants without any human intervention. The results of these five datasets determined the final ranking.

The competition organizers evaluated submitted solutions, incorporating their baseline solution into the AUC-based score.

Organizers provided the baseline solution to the participants. The baseline solution was shared with the participants at the competition website [3] and CodaLab platform [4], where the competition took place.

This solution included the Auto-Table-Join (ATJ) step (Figure 2), where temporal information and inter-table interactions are extracted from the given data.

ATJ first constructs a tree-structure based on the relations among tables, where the root is the main table and the edges indicate the relation types, as shown in Figure 3.

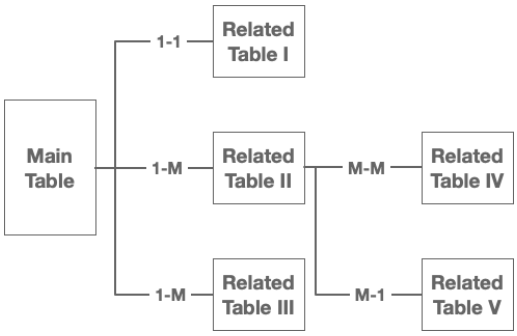


Figure 3: Baseline solution for the AutoML track: Auto-Table-Join (ATJ) step details

Afterwards, ATJ traverses the tree for table joining that includes two cases: non-temporal, where one of the tables does not have a timestamp column, and temporal, where both tables u and v have a timestamp column. The latter case needs data leak prevention - for each record in u, only records in v that occurs prior to it can be utilized during joining in the baseline solution. A sliding time-window (covering the current record and last several ones) is then used to traverse the intermediate table w where records are sorted in chronological order. Once a record extracted from u is visited, records from v with the same key and in the window are used in joining. To be more specific, groupby operations are applied on these records and the results are appended to i_u. The baseline implementation involves depth first search (dfs) for tree traversal in a recursive function. A recursive function $dfs(u)$ is defined to: 1) apply $dfs()$ on all the children of u, and 2) join these children to u. When joining a table v to a table u, ATJ ensures that the resulting table has the same row number as u.

Gradient Boosting Model (Light GBM) was used in the baseline solution for the machine learning part to the binary classification problem. The base solution was distributed in a Docker container provided by CodaLab.

Full details describing the competition are available via CodaLab platform and 4Paradigm Web-site and supposed to be available for at least one year since the competition ends.

3.2 Winners

The AutoML track honored the top 10 winning teams. The Top 3 winners were:

- First Prize ($15,000): Zhipeng Luo from DeepBlue Technology (Shanghai) Co., Ltd, Jianqiang Huang from Peking University, Mingjian Chen and Bohang Zheng from DeepBlue Technology (Shanghai) Co., Ltd;
- Second Prize ($10,000): Chengxi Xue, Shu Yao, Zeyi Wen, and Bingsheng He from National University of Singapore;
- Third Prize ($5,000): Suiyuan Zhang and Jinnian Zhang from Alibaba Group, Zhanhao Liu from Georgia Institute of Technology, Zhiqiang Tao, Yaliang Li, Bolin Ding, and Shaojian He from Alibaba Group, Xu Chu from Georgia Institute of Technology, Xin Li and Jingren Zhou from Alibaba Group;

and the following teams, ranked from the 4th to the 10th, won the Honorable Prizes ($500 each):

- 4th Place: Jian Sun, Hao Zhang, Chunmeng Zhong, and Zaiyu Pang from Tsinghua University, Hongyu Jia from Dalian University of Technology, Xiao Huang from Nanjing University, Bin Lin from Nanjing University of Posts and Telecommunications, and Xibin Zhao and Hai Wan from Tsinghua University;
- 5th Place: Masashi Yoshikawa and Takeru Ohta from Preferred Networks, Inc.;
- 6th Place: Suiqian Luo from Guazi;
- 7th Place: Yu Luo and Qizhen Yao from Hikvision Research Institute;
- 8th Place: Guanghui Zhu, Xu Guo, Xin Fang, and Zhuoer Xu from Nanjing University;
- 9th Place: Mengjiao Bao from Beihang University, Hui Xue from Microsoft Research Asia, Huan Chen and Peng Yan from Meituandianping;
- 10th Place: Jin Xu, Hantao Shu, and Jian Li from Tsinghua University.

3.3 Impact and Outcome

Solutions offered by top 10 winning teams have introduced automation scripts following good coding practices and observing possible sources of failure on 5 private datasets.

For data cleaning, traditional steps included correct parsing, conversion, de-duplicaton of columns after merging, removing constantly valued columns, missing value substitution (used by the majority of the winning teams), etc. Specific to given data origin, there were a user id and session id attributes identification out of all other attributes; these columns played an important role in the subsequent feature engineering part.

For data transformation (table merging), aggregation operations were performed on both tables to implement M to M keys joining operations. Half of the teams used the temporal join in their solutions. The winning team DeepBlueAI used a temporal split.

Feature engineering, when automated, can produce too many features causing issues in time and memory consumption. To take control over this process multiple solutions were suggested: iterative step-by-step feature generation followed by feature selection step to reduce dimensionality - from feature transformation to feature combinations of higher order.

Different feature kinds played a role in feature engineering, sometimes, by analogy with recommender systems: categorical and ordinal features, date and time slice features, entity and id features, and features produced out of original features in combinatorial ways. Filter feature selection was used by many teams. As a part of feature engineering, features extracted to introduce information from table merging were based on aggregation functions, in particular, mean, sum, count, min, max.

For hyper parameter tuning, solutions were to reduce the search space, in particular, implementing wrapper-like approach on data samples with introduction of a prior knowledge that was specific to a parameter of the selected machine learning model (DeepBlueAI, 1st place winner). Most teams used Bayesian Optimization.

To deal with class imbalance during modeling, the most used technique was under sampling. Gradient Boosting Decision Tree (GBDT) model dominated in the winners' choice for the machine learning part. Ensemble solutions were mostly bagging and boosting.

Winning implementations notably utilized multithreading. Python was a language of choice for the vast majority of participants. Most popular open-source machine learning libraries were Pandas, Numpy, SkLearn and LightGBM.

AutoML organizers requested opening code of the winning solutions to the public. Winning teams shared their code on GitHub.

The AutoML track has attracted the participation of 860 teams who made around 4,955 submissions in total. This was the largest AutoML competition to the date. Winning solution from DeepBlueAI outperformed the baseline solution significantly (see Figure 4).

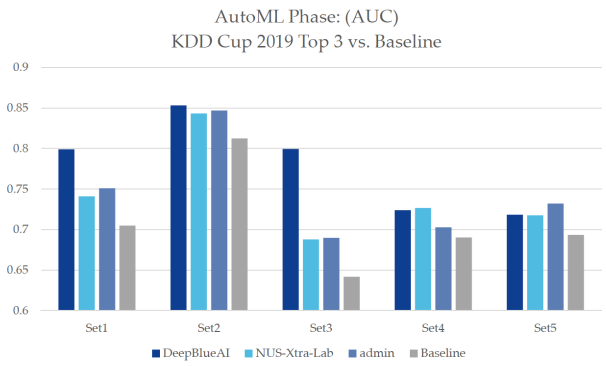


Figure 4: AutoML track: top 3 winning solution compared to baseline (provided by 4Paradigm)

4. THE HUMANITY RL TRACK

The Humanity RL track competition was developed with a noble cause - to solve complex problems facing humanity. The focus of this track for 2019 was to develop RL competition for the greater good of humanity. Unlike the past KDD cup competitions, which focused on mainly supervised learning, this was focused on RL. Further, it was solving a humanitarian cause helping Malaria control in Africa.

Sponsor: IBM Research, Hexagon-ML.com, SIGKDD
Organizing Committee: Sekou Remy, Oliver Bent, Ankur Agrawal, Uday Kumar Naik, Charles Wachira, and Nelson Bore.

4.1 Description

Malaria is a mosquito borne disease that continues to pose a significant global health burden. As of 2018, there are 228 Million cases of Malaria, causing 405,000 deaths worldwide [4]. About 50% of the world's population is at risk of malaria infection. However, some regions are more vulnerable than others. Sub Saharan Africa is most affected, with 90% of all cases originating there. One of the popular methods of controlling malaria are by using insecticide-treated nets (ITNs) and Indoor Residual Spraying (IRS). Recently, the combination of ITNs and IRS have become the primary method of malaria prevention, because the anopheles mosquito only bites after nine o'clock at night, when most kids are in bed. Once a mosquito lands on the net, it dies because of the insecticide, which disrupts the reproductive cycle [5].

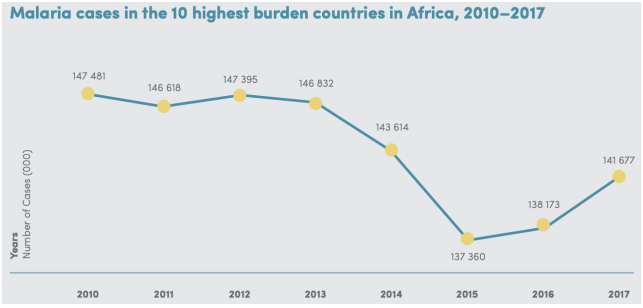


Figure 5: Malaria cases in Africa 2010–2017 [6].

Using the above Malaria control measures there has been about 29% reduction in mortality rates since 2010. However, the goal is more audacious - "Eradicate Malaria". In order to do this, a collaborative effort between IBM Research and University of Oxford, led to development of decision making tools using machine learning (ML) and artificial intelli-

gence (AI) may provide some potential answers. The work was summarized in the paper - 'Novel Exploration Techniques (NETs) for Malaria Policy Interventions' [17] aimed at augmenting the decision-making abilities of officials, and exploring more effective malaria policy interventions.
Through this KDD Cup 2019 Humanity RL competition we are looking for participants to apply machine learning tools to determine novel solutions which could impact malaria policy in Sub Saharan Africa. Specifically, how should combinations of interventions which control the transmission, prevalence and health outcomes of malaria infection, be distributed in a simulated human population. This challenge has been framed as a RL problem. The participants are expected to submit high performing solutions to the sequential decision making task. For this competition, actions receive stochastic and delayed rewards, which are resource constrained by both the monetary and computational costs associated with implementing and observing the impact of an action.

4.1.1 The Reinforcement Learning (RL) Problem

RL is a specialized field under machine learning. In RL, an agent is programmed using various algorithms and strategies, to learn in an interactive environment by trial and error. This enables the agents to leverage feedback from its own actions and experiences. Thus, RL is generally formulated and solved as the Markov decision problem (MDP) and has the following components - Environment, Agent, States of the environment, Actions, Transition probabilities, Transition rewards, Policy, and Performance metric. A typical RL system is shown below.

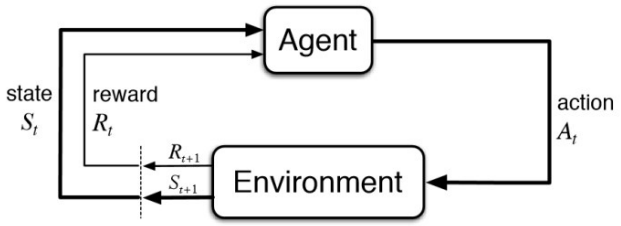


Figure 6: Typical RL System [17].

For the KDD 2019 competition, the environment was developed by IBM Research team, the state values were years 1 to 5, Actions were defined by either ITN or IRS distributions and rewards were a stochastic reward over a policy.
Mathematically, the different components of the KDD 2019 RL system can be noted as follows:
State: $S \in \{1,\ 2,\ 3,\ 4,\ 5\}$
Action: $a^{ITN} \in [0, 1]$ and $a^{IRS} \in [0, 1]$
Action-Value: $As = [a^{ITN}, a^{IRS}]$
Rewards: $R\pi \in (-\infty, \infty)$
Policy: π for this challenge consists of a temporal sequence of actions, as illustrated in Figure 8.

4.1.2 Technology Highlights

There were 2 main components of technology: the environment API and Hexagon-ML platform. The environment was designed as an open source project and deployed as an API that could be accessed via python. The API had the ability to provide load balancers to enable multiple participants as well as keep track of states and rewards. A max of 100 iterations were allowed per day.

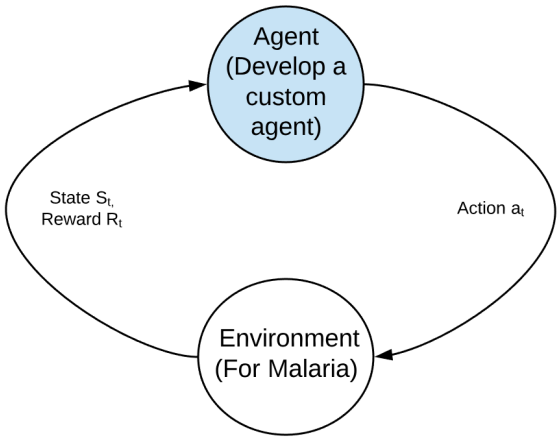


Figure 7: KDD 2019 Basic Humanity RL System [7].

The Hexagon-ml platform [8] provided the necessary infrastructure to compute the scores as total rewards, as well as a holistic competition experience. This experience included a concept of submission file analogous to Kaggle [9], a discussion forum, a profile page and a couple of webinars.

4.2 Winners

The Humanity RL track honored the top 10 winning teams.

- First Place ($5,000): Zi-Kuan Huang, Jing-Jing Xiao, and Hung-Yu Kao from National Cheng Kung University
- Second Place ($4,000): Lixin Zou from Tsinghua University, Long Xia from JD.com, Zhuo Zhang from Beihang University, and Dawei Yin from JD.com
- Third Place ($3,000): Suiqian Luo from Guazi
- Fourth Place ($3,000): Vladislav Shakh-Nazarov from Yandex School of Data Analysis and National Research University Higher School of Economics
- Fifth Place ($3,000): Etsuro Minami from NS Solutions Corporation & Financial Engineering Group, Inc., Akira Okada from NS Solutions USA Corporation, Michihiro Nakamura from Financial Engineering Group, Inc., and Masayuki Ishikawa from Financial Engineering Group, Inc.
- Sixth Place ($2,000): Xiaolan Jiang from Department of Informatics, the Graduate University for Advanced Studies (Sokendai) / National Institute of Informatics, Japan.
- Seventh Place ($2,000): Van Bach Nguyen, Bao Long Vu, and Mohamed Karim Belaid from the University of Passau, Germany
- Eighth Place ($1,000): Quanjun Chen and Haoran Zhang from the University of Tokyo, Japan, Mina He from IMT Atlantique, France, and Zhaonan Wang from National Institute of Advanced Industrial Science and Technology, Japan

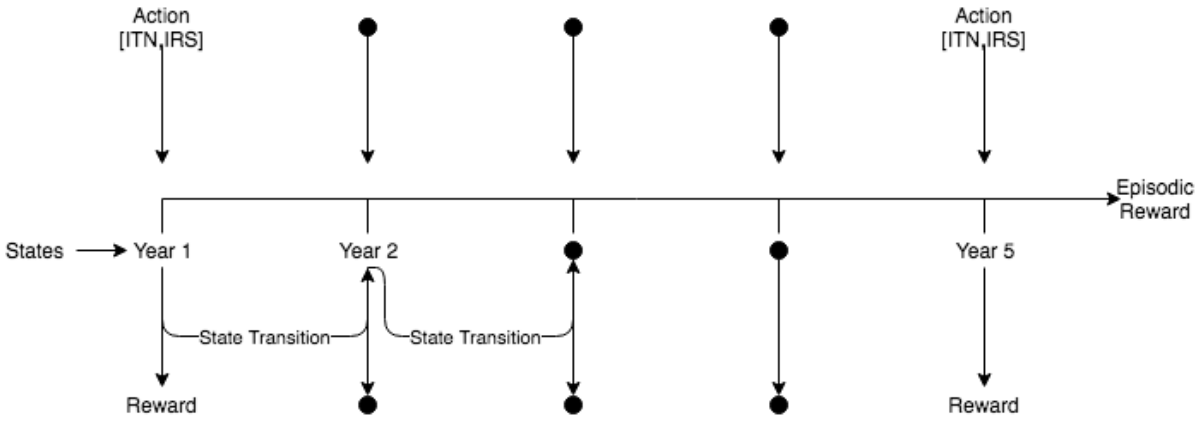


Figure 8: KDD 2019 - Policy Actions [13].

- Ninth Place ($1,000): Wei Xin from Wuxi Xuelang Industrial Intelligence Technology Co. Ltd.

- Tenth Place ($1,000): Anand Rajasekar from Indian Institute of Technology, Madras.

4.3 Impact and Outcome

Machine learning areas such as supervised and unsupervised learning have received the right focus and thus have seen industrial maturity in tools and skills. RL is the next frontier. We had a total of 244 teams, including 292 individual participants being part of the competition. A total number of 735 agents were submitted during this process. Hexagon-ML was provided an Innovation Award for developing a portal capable of RL competition and helping solve a huge humanitarian cause of helping to eradicate Malaria.
In addition, the KDD Cup 2019 Humanity RL competition had impacts in the following areas.
Community Development. Hosting a competition environment such as IBM-Malaria environment and having a competition platform such as Hexagon-Ml's to do RL has had a huge impact. It brought together and created a community of skilled professions, motivated students and other interested personnel.
Skill Enhancement. This competition has enabled a lot of people to learn new strategies in RL. Through code sharing they have learnt and developed their skills in this new area in machine learning.
Publications & Posters. This competition encouraged the community to develop new methods and strategies. As a result of this, there were several posters presented during the conference and papers written. One such paper "Policy Learning for Malaria Control" [16], developed by the seventh place holder talks about the challenges in sequential design with limited observations.

5. THE KDD CUP DAY

On the KDD Cup Day, we announced the winners, presented them their award certificates, and learned about the winning solutions both via their oral presentations and the poster sessions. We also had an invited talk by RL Guru, Dr. Balaraman Ravindran from IIT Madras, and hosted two panel discussions.

5.1 Panel 1: "How should companies use competition platforms?"

Companies have used data science competition as a strategy to bring cultural change or even crowd source their problems to external teams. Netflix in our recent past was one example, where they pioneered this practice by crowdsourcing their recommendation algorithm. Further, data science competition companies, such as Kaggle, Hexagon-ML and others, host competitions either sponsored by companies on their platforms or hosted in the companies itself. In this panel we discussed how corporate companies should use data science competition platforms with some of the industry leaders.

Moderator: Taposh Dutta-Roy
Panelists:

- Claudia Perlich, Senior Data Scientist, Two Sigma;
- Jason Jones, Chief Data Scientist, Health Catalyst;
- Lin Wang, Senior Data Scientist, Vesta Corporation.

5.2 Panel 2: "How will AutoML change the future of data science?"

AutoML, as a concept and as a product, gained traction several years ago, increasing in popularity and complexity ever since. Originally designed to automate certain steps that are beyond the abilities of non-experts, it makes data scientists more productive, inevitably shifting perspective, focus, and calling for different skills. During this panel we are hoping to collect opinions of people who invent, create, and use AutoML. In particular, we are interested to discuss non-trivial cases and applications of AutoML, current limitations, variety of existing products and how they are meeting new demands, arising applications, overall progress in the area over a few years, and debate on how data science jobs will change influenced by AutoML.

Moderator: Iryna Skrypnyk
Panelists:

- Ashwin Aravindakshan, Associate Professor, Graduate School of Management and Director of the M.S. in Business Analytics program, UC Davis;

- Dmitry Larko, Senior Data Scientist, H2O.ai;
- Ganesh Thondikulam, Executive Director of Analytics Digital Foundation, Kaiser Permanente;
- Wei-Wei Tu, Principle Machine Learning Architect, 4Paradigm, Inc.

5.3 KDD Cup Innovation Award

In order to foster innovation in data science competitions and encourage the community, this year we have established a special KDD Cup Innovation Award. In 2019, KDD Cup Innovation Award was given to Hexagon-ML, our competition platform sponsor for the Reinforcement Learning competition. Hexagon-ML obtains this award for:

- Pioneering a new type of competition, the Reinforcement Learning Competition, and implementing it under a novel computational environment for KDD Cup 2019.
- Contribution into solving a Malaria Problem for Humanity in successful collaboration with IBM Research Africa. In 2019, re-emergence of infectious diseases is among the top 10 humanitarian crises in the world.
- Efforts in gathering, developing, and growing a data science community in Reinforcement Learning.

6. SUMMARY

Three tracks of competitions were hosted at KDD Cup 2019, each with a unique set of offerings. Some of the highlights of the KDD 2019 included: 1. Total awards of more than 100K U.S. dollars for winners; 2. First time AutoML competition to encourage development in AutoML capabilities; 3. First time humanity track focusing on Reinforcement Learning Competition; and 4. Innovation award for key contributions to Humanity Track. We worked hard and also learned a lot while organizing these events. For example, lessons were learned about spreading the word [10], managing the compounding workload when running multiple, simultaneous tracks, and allowing for enough travel planning time that allows winning teams to attend the KDD Cup Day. To carry our experience forward, one of us was appointed to continue helping with KDD Cup 2020.

7. ACKNOWLEDGEMENTS

We would like to thank the individuals and organizations that submitted KDD Cup proposals for consideration. Although we were not able to host all of them, we saw great ideas and value in those proposals. We greatly appreciate the contest sponsors for their generosity in providing interesting problems, data, monetary awards, computing resources and staff support. We also extend our thanks to the SIGKDD leadership and KDD 2019 conference chairs and staff members for their guidance and strong support. Last but not the least, we would like to thank the contest participants, speakers, panelists, and conference attendees for their enthusiastic interactions.

8. REFERENCES

[1] https://dianshi.bce.baidu.com/competition/29/rule (last accessed on April 5, 2020).

[2] https://github.com/yaoxuefeng6/Paddle_baseline_KDD2019 (last accessed on March 31, 2020).

[3] https://www.4paradigm.com/competition/kddcup2019 (last accessed on March 31, 2020).

[4] https://codalab.lri.fr/competitions/559 (last accessed on March 31, 2020).

[5] https://www.who.int/publications-detail/strategic-advisory-group-malaria-eradication-executive-summary (last accessed on March 31, 2020).

[6] https://apps.who.int/iris/bitstream/handle/10665/275868/WHO-CDS-GMP-2018.25-eng.pdf (last accessed on March 31, 2020).

[7] https://medium.com/@taposhdr/reinforcement-learning-to-eradicate-malaria-with-ai-49e9e4016665 (last accessed on March 31, 2020).

[8] http://compete.hexagon-ml.com/ (last accessed on March 31, 2020).

[9] https://www.kaggle.com/ (last accessed on March 31, 2020).

[10] https://www.kdnuggets.com/2019/02/word-from-kdd-cup-2019-organizers.html (last accessed on March 31, 2020).

[11] Board of longitude. https://en.wikipedia.org/wiki/Board_of_Longitude (last accessed on March 31, 2020).

[12] Fueling enterprise innovation through competition. https://www.linkedin.com/pulse/fueling-enterprise-innovation-through-competition-taposh-dutta-roy/ (last accessed on March 31, 2020).

[13] O. Bent, S. L. Remy, S. Roberts, and A. Walcott-Bryant. Novel exploration techniques (NETs) for malaria policy interventions. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Publications.

[14] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–36, Menlo Park, CA, 1996. AAAI/MIT Press.

[15] H. Liu, T. Li, R. Hu, Y. Fu, J. Gu, and H. Xiong. Joint representation learning for multi-modal transportation recommendation. In *AAAI 2019*.

[16] V. B. Nguyen, B. M. Karim, B. L. Vu, J. Schlötterer, and M. Granitzer. Policy learning for malaria control. In *CoRR abs/1910.08926*. 2019.

[17] R. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, USA, 1998.

About the authors:

Iryna Skrypnyk is a Head of AI at IDEA Innovation Lab at Pfizer. She is introducing technological innovations to Real World Evidence studies across multiple therapeutic areas. Previously, he worked at several New York startup companies as a Head of Data Science. She earned a Ph.D. from the Computer Science Department, University of Jyväskylä, Finland, part-time performing her research at IBM T.J. Watson Research Center and Bell Laboratories. (https://www.linkedin.com/in/iryna-skrypnyk-9934b39/)

Taposh Dutta Roy leads the innovation team of decision support group at Kaiser Permanente. His work focuses on journey analytics, deep learning, data science architecture and strategy. He has masters in Electrical Engineering and Computer Engineering from Illinois Institute of technology and MBA from UC Davis. While in Ph.D. program his research was focused on simulation and modeling of device physics. (https://www.linkedin.com/in/taposh/)

Wenjun Zhou is an Associate Professor at the University of Tennessee Knoxville (UTK), where she teaches and performs research on data and text mining. She earned a Ph.D. from Rutgers University and a M.S. from the University of Michigan. She is currently a George and Margaret Melton Faculty Fellow and Roy & Audrey Fancher Faculty Research Fellow at UTK's Haslam Business School. (http://web.utk.edu/~wzhou4/)