

# The 5<sup>th</sup> AutoML Challenge

## AutoML for Temporal Relational Data

Wei-Wei Tu, Isabelle Guyon, Hugo Jair Escalante, Xiawei Guo, Mengshuo Wang,  
Yuanfei Luo, Ling Yue, Hai Wang, Wenhao Li,  
Evelyne Viegas, Sergio Escalera, Yuqiang Chen, Wenyuan Dai and Qiang Yang

Thanks to KDD Cup Chairs: [Taposh Dutta-Roy](#), [Wenjun Zhou](#) and [Iryna Skrypnyk](#)



# Schedule

13:00 ~ 13:10

Competition Overview

13:10 ~ 13:15

*5th Place Solution Presentation*

13:15 ~ 13:25

*3rd Place Solution Presentation*

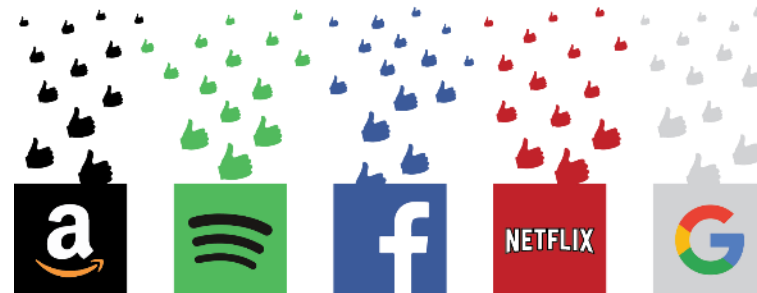
13:25 ~ 13:35

*2nd Place Solution Presentation*

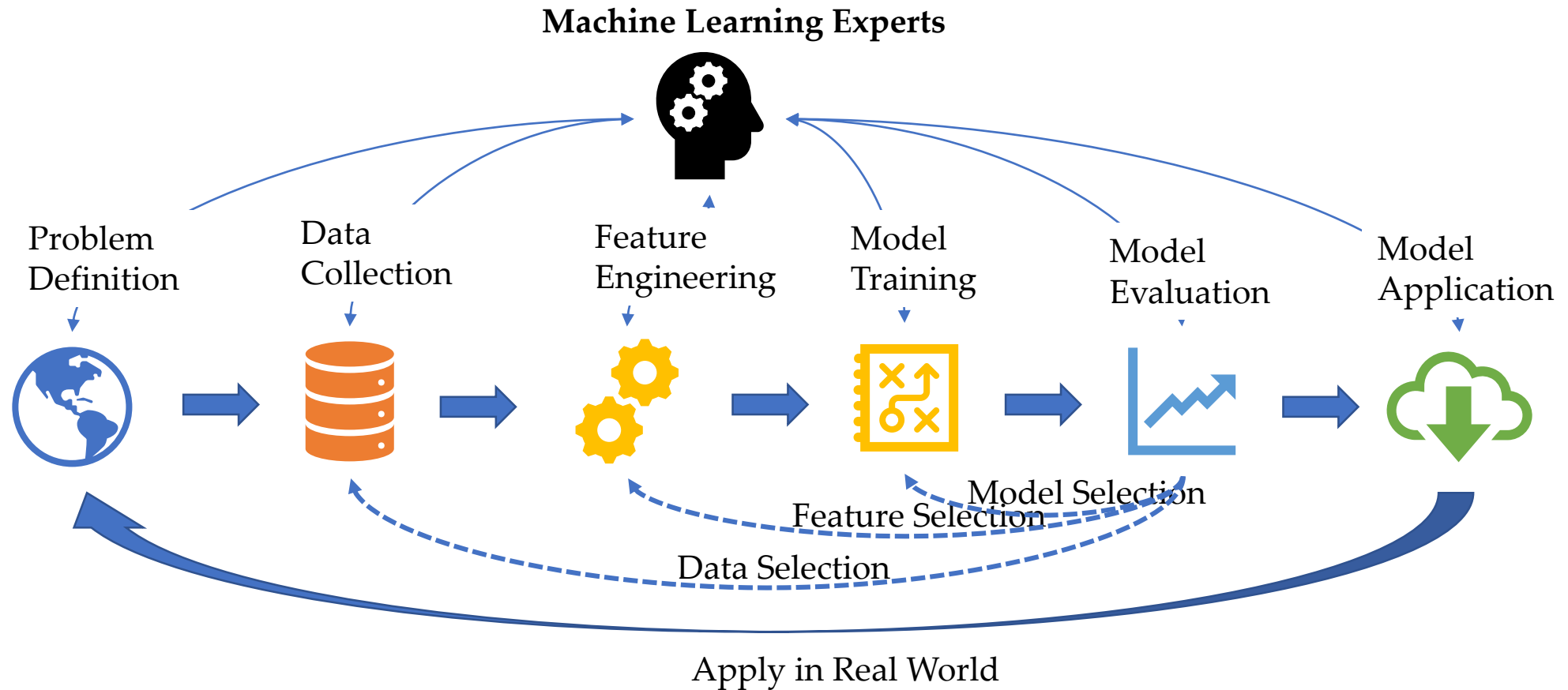
13:35 ~ 13:45

*1st Place Solution Presentation*

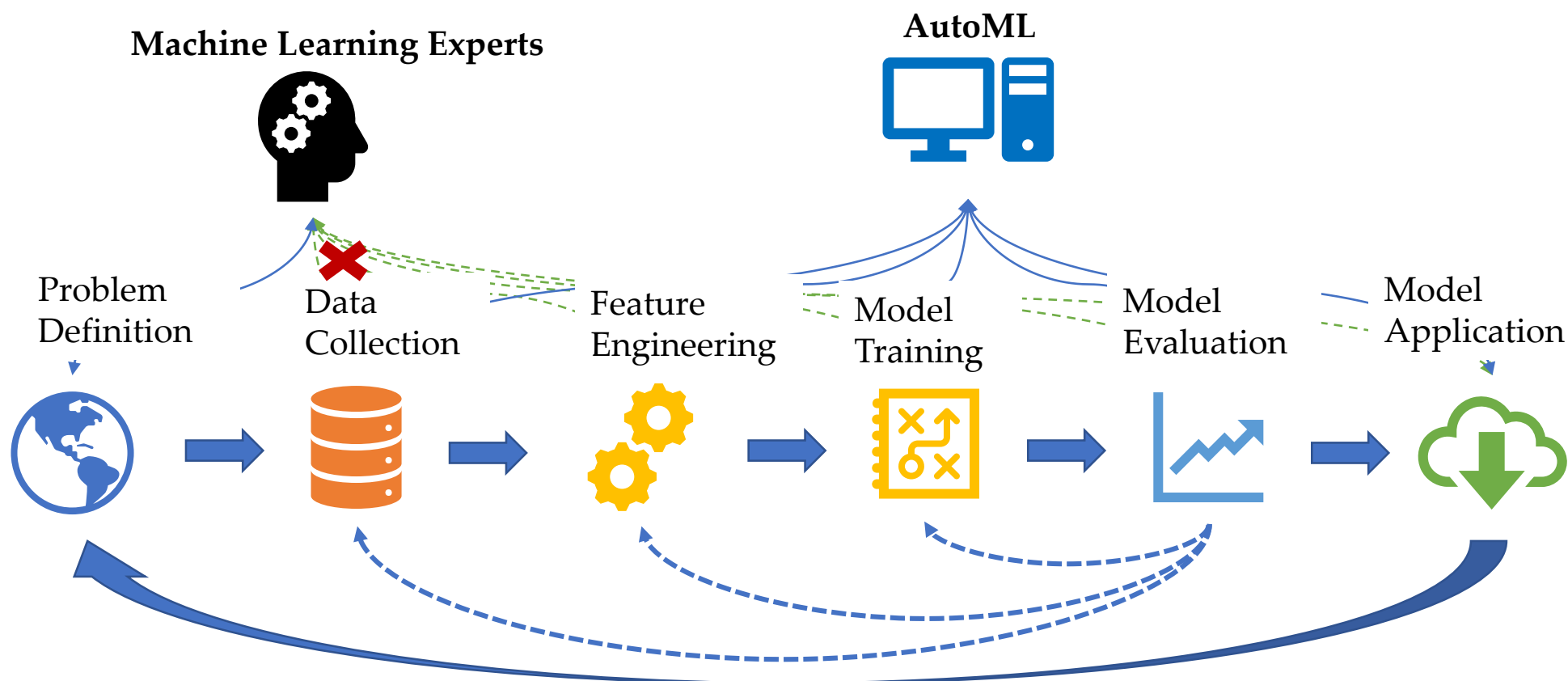
# Machine Learning Applications



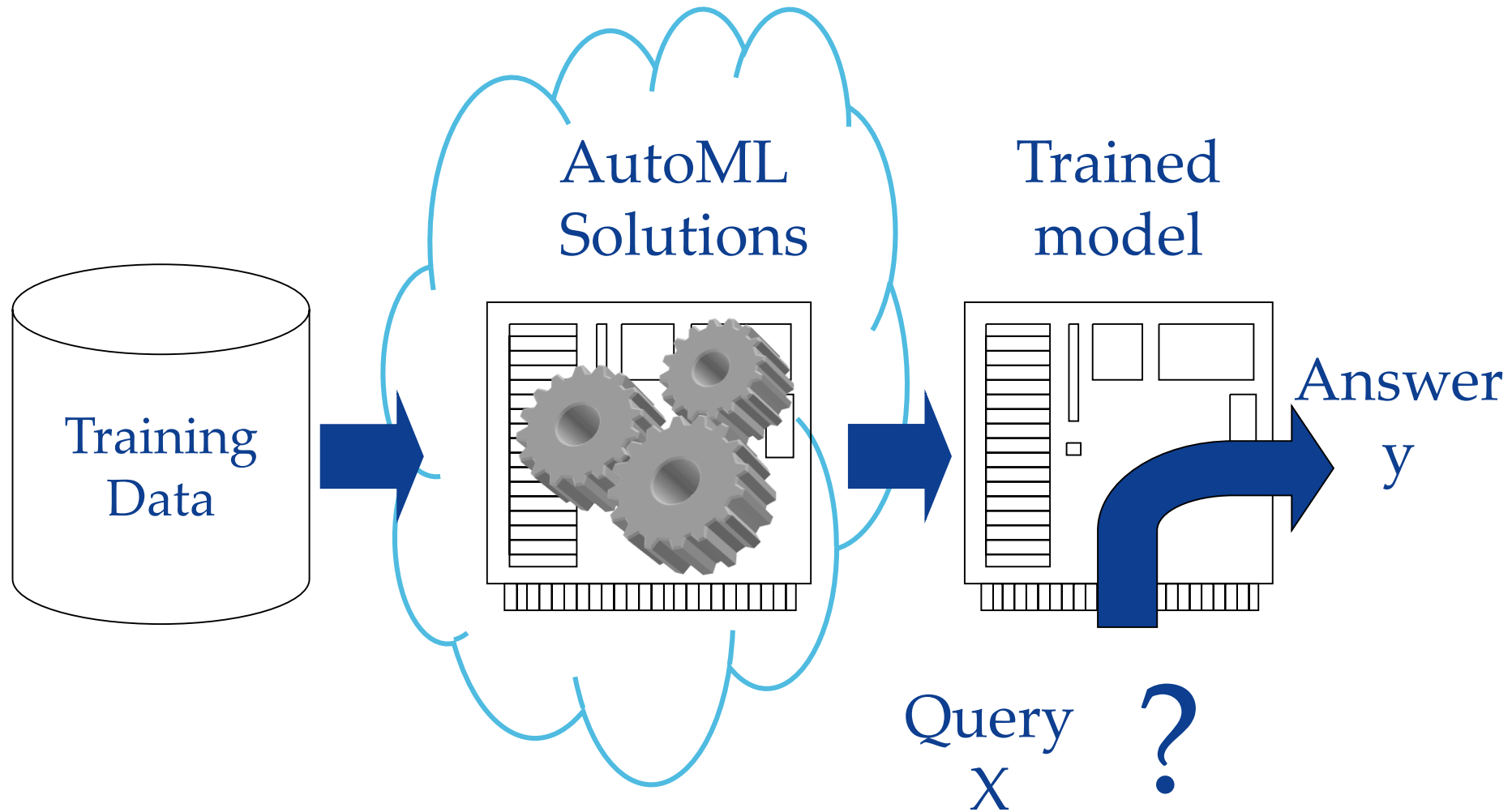
# Challenges in Real World Applications



# Solution: AutoML



# The AutoML Competitions



# AutoML VS. Regular Competitions



|   | <b>Regular Competition</b>        | <b>AutoML Competition</b>          |
|---|-----------------------------------|------------------------------------|
| <b>Main Challenge</b>                       | Insight of Problem/Data           | General Solution                   |
| <b>Evaluation of Generalization Ability</b> | Single Dataset (almost)           | Multiple Datasets                  |
| <b>Competition Paradigm</b>                 | Training Data Visible             | Blind Test (No Human Intervention) |
| <b>Submission Types</b>                     | Result Submission                 | Code Submission                    |
| <b>Computation Resource Management</b>      | Efficiency Non-Sensitive (almost) | Efficiency Sensitive               |

# Brief History of AutoML Challenges

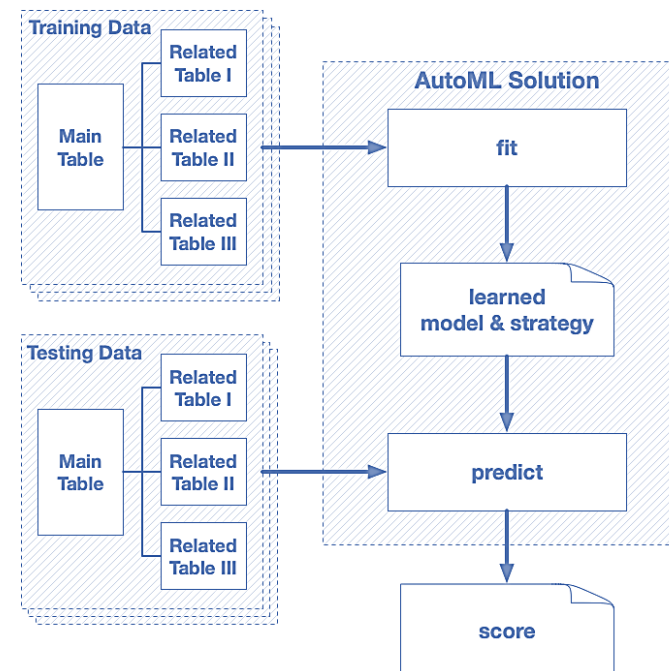
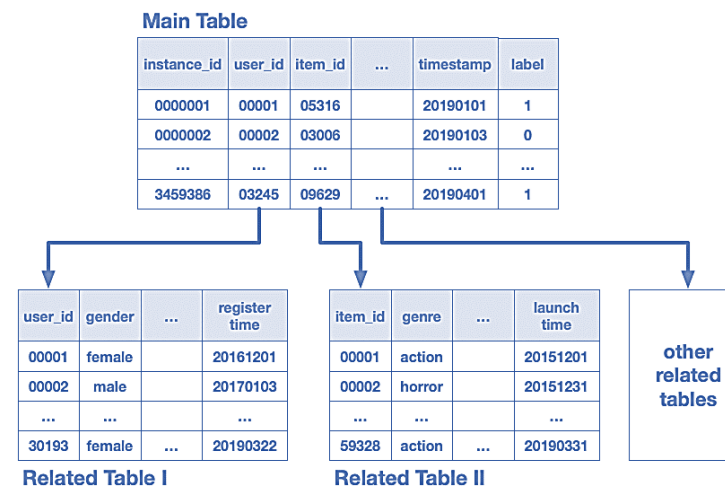


| Competition           | Year        | Lasts           | Collocated Events   | #Participants     | Prizes            | Providers & Sponsors                             |
|-----------------------|-------------|-----------------|---------------------|-------------------|-------------------|--|
| AutoML1               | 2015-2016   | 2 Years         | NIPS, ICML, IJCNN   | 600+              | 30,000 USD        | Microsoft and ChaLearn                           |
| AutoML2               | 2018        | 4 Months        | PAKDD 2018          | 250+              | 10,000 USD        | 4Paradigm and ChaLearn                           |
| AutoML3               | 2018        | 3 Months        | NeurIPS 2018        | 300+              | 15,000 USD        | 4Paradigm, Microsoft and ChaLearn                |
| AutoML4 (AutoML3+)    | 2019        | 3 Months        | PAKDD 2019          | 130+ Teams        | 6,500 USD         | 4Paradigm, ChaLearn, Microsoft and Amazon        |
| <b>AutoML5</b>        | <b>2019</b> | <b>3 Months</b> | <b>KDD Cup 2019</b> | <b>860+ Teams</b> | <b>33,500 USD</b> | <b>4Paradigm, ChaLearn, Microsoft and Amazon</b> |
| AutoDL Pre1 (AutoCV)  | 2019        | 2 Months        | IJCNN 2019          | 100 Teams         | 4,000 USD         | ChaLearn, 4Paradigm and Google                   |
| AutoDL Pre2 (AutoCV2) | 2019        | 2 Months        | ECML PKDD 2019      | Ongoing           | 4,000 USD         | ChaLearn, 4Paradigm and Google                   |
| AutoDL Pre3 (AutoNLP) | 2019        | 1 Month         | WAIC 2019           | Ongoing           | 7,000 USD         | ChaLearn, 4Paradigm and Google                   |
| AutoDL                | 2019        | Coming Soon     | NeurIPS 2019        | Coming Soon       | 15,000 USD        | ChaLearn, 4Paradigm and Google                   |



# AutoML5 at KDD Cup 2019

- **Hardest** AutoML Competition
- AutoML challenged by:
  - **Temporal Relational Data.**
    - *Temporal*
      - Dependency between instances, concept changing through time
      - Appropriate temporal feature generation
    - *Relational*
      - **First** AutoML competition dealing with **multiple-tables**
  - **AutoML.**
    - Efficiency and generalization ability of solutions



# AutoML5 at KDD Cup 2019



- 3 Phases:

- *Feedback* Phase

- ~ 2 Months 26 Days
- 5 public datasets
- Realtime leaderboard

- *Check* Phase

- ~ 2 Weeks
- Only One Chance to fix bug

- *AutoML* Phase

- ~ 1 Week
- Blind Testing on 5 fresh private datasets

## 10 Datasets from different Real-World Applications

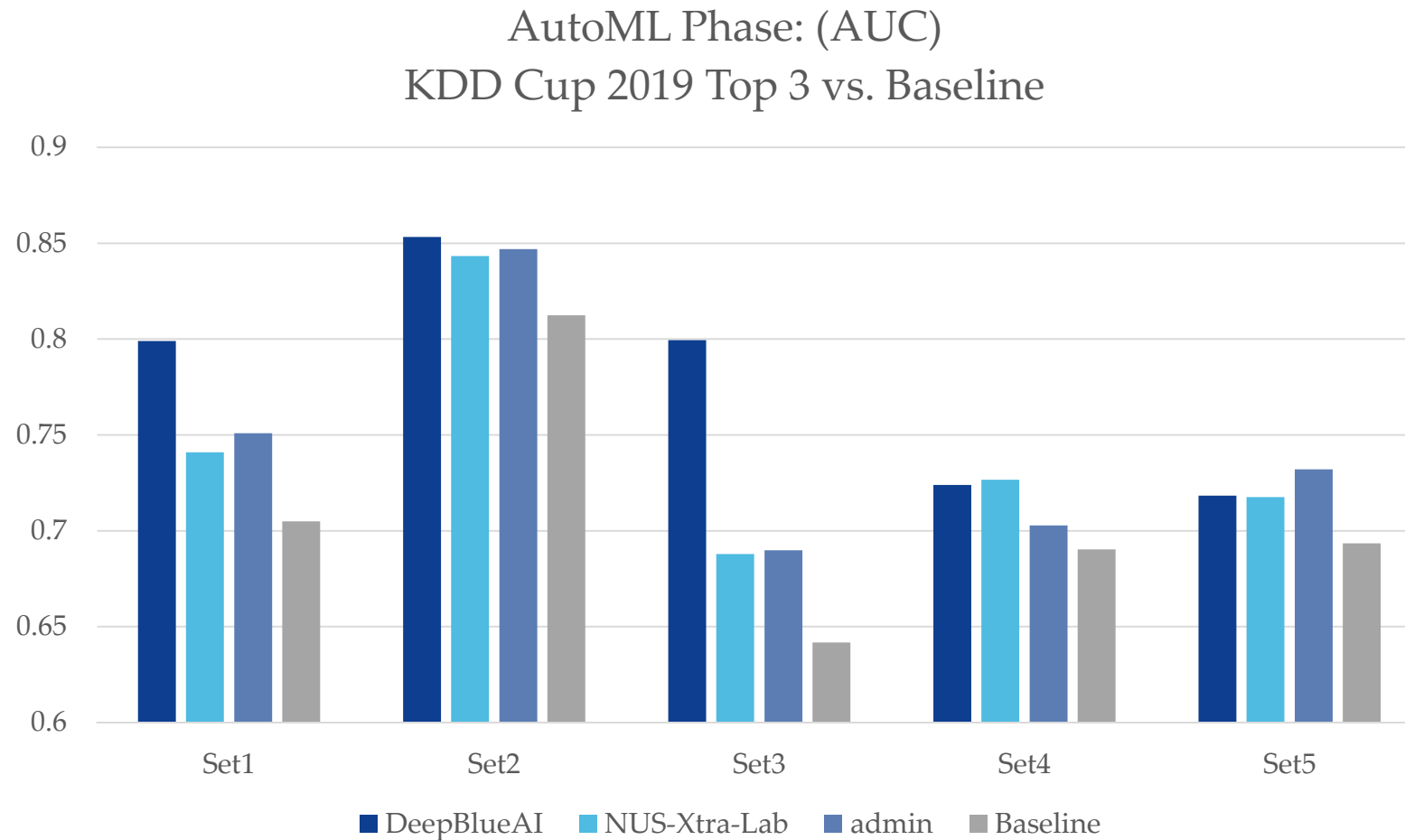
(Online Advertising, Recommender System, Retail Marketing, ...)

|         | Dataset | Train Positive Ratio | Test Positive Ratio | # Train Main Instance | # Test Main Instance | # Tables | # Columns |
|---------|---------|----------------------|---------------------|-----------------------|----------------------|----------|-----------|
| Public  | A       | 7.11%                | 7.03%               | 808953                | 779656               | 3        | 69        |
|         | B       | 2.25%                | 2.05%               | 1111133               | 1000160              | 4        | 81        |
|         | C       | 0.06%                | 0.05%               | 1003286               | 981609               | 2        | 63        |
|         | D       | 89.50%               | 89.11%              | 1888366               | 1119778              | 3        | 33        |
|         | E       | 1.66%                | 2.33%               | 226091                | 34867                | 3        | 35        |
| Private | F       | 7.78%                | 7.39%               | 1392374               | 716549               | 3        | 68        |
|         | G       | 1.79%                | 1.65%               | 999925                | 1052087              | 4        | 79        |
|         | H       | 4.63%                | 4.50%               | 946566                | 323192               | 3        | 74        |
|         | I       | 52.29%               | 50.32%              | 1111112               | 889152               | 5        | 18        |
|         | J       | 0.18%                | 0.15%               | 435329                | 31076                | 2        | 40        |

# An Overview

- Lasts 3 months
- **Largest** AutoML Competition!
  - About **860 teams**, over **5000 code submissions**!
- May be the first KDD Cup competition with all Top 10 Solutions Open-Source!
- **Much Better** Performance over baseline methods!
- Unfortunately there was an accident at CodaLab server!
  - Fortunately we recovered it very soon and got all final submissions from all top teams.

# Great Improvement vs. Baseline!



# Winners!



- **Champion: DeepBlueAI**
  - Zhipeng Luo, Jianqiang Huang, Mingjian Chen, Bohang Zheng
  - DeepBlue Technology (Shanghai) Co., Ltd, Peking University
- **Second Place: NUS-Xtra-Lab**
  - Chengxi Xue, Shu Yao, Zeyi Wen; Bingsheng He
  - National University of Singapore
- **Third Place: admin**
  - Suiyuan Zhang, Jinnian Zhang, Zhanhao Liu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Shaojian He, Xu Chu, Xin Li, Jingren Zhou
  - Alibaba Group, Georgia Institute of Technology
- **4<sup>th</sup> Place: ts302\_team**
  - Jian Sun, Hao Zhang, Chunmeng Zhong, Zaiyu Pang, Hongyu Jia, Xiao Huang, Bin Lin
  - Tsinghua University, Nanjing University, Dalian University of Technology
- **5<sup>th</sup> Place: yoshikawa**
  - Masashi Yoshikawa, Takeru Ohta
  - Preferred Networks
- **6<sup>th</sup> Place: Alpha**
  - Suiqian Luo
  - Guazi
- **7<sup>th</sup> Place: teews**
  - Yu Luo, Qianzhen Yao
  - Hikvision Research Institute
- **8<sup>th</sup> Place: PASA\_NJU**
  - Guanghui Zhu, Xu Guo, Xin Fang, Zhuoer Xu
  - Nanjing University
- **9<sup>th</sup> Place: \_\_CHAOS\_\_**
  - Mengjiao Bao, Hui Xue, Huan Chen, Peng Yan
  - Beihang University, Microsoft Research Asia, Meituandianping
- **10<sup>th</sup> Place: IIIS\_FLY**
  - Jin Xu, Hantao Shu, Jian Li
  - Tsinghua University

# Open-Source: All winner solutions!

- <https://www.4paradigm.com/competition/kddcup2019>

| Rank | Team Name    | github URL  |
|------|--------------|---|
| 1    | DeepBlueAI   | <a href="https://github.com/DeepBlueAI/AutoSmart">https://github.com/DeepBlueAI/AutoSmart</a>                                 |
| 2    | NUS-Xtra-Lab | <a href="https://github.com/shuyao95/kddcup2019-automl.git">https://github.com/shuyao95/kddcup2019-automl.git</a>             |
| 3    | admin        | <a href="https://github.com/DominickZhang/KDDCup2019_admin">https://github.com/DominickZhang/KDDCup2019_admin</a>             |
| 4    | ts302_team   | <a href="https://github.com/ts302team/KDDCUP2019-AutoMLTrack">https://github.com/ts302team/KDDCUP2019-AutoMLTrack</a>         |
| 5    | yoshikawa    | <a href="https://github.com/pfnet-research/KDD-Cup-AutoML-5">https://github.com/pfnet-research/KDD-Cup-AutoML-5</a>           |
| 6    | Alpha        | <a href="https://github.com/luosuiqian/sample_code_submission">https://github.com/luosuiqian/sample_code_submission</a>       |
| 7    | teews        | <a href="https://github.com/teews-ly/Kdd-Cup-2019-automl">https://github.com/teews-ly/Kdd-Cup-2019-automl</a>                 |
| 8    | PASA_NJU     | <a href="https://github.com/guojixu/KDD-Cup-2019-AutoML-Solution">https://github.com/guojixu/KDD-Cup-2019-AutoML-Solution</a> |
| 9    | _CHAOS_      | <a href="https://github.com/SpongeBob/AutoTableGBDT">https://github.com/SpongeBob/AutoTableGBDT</a>                           |
| 10   | IIIS_FLY     | <a href="https://github.com/shuhantao/KDD2019-Challenge">https://github.com/shuhantao/KDD2019-Challenge</a>                   |

# Statistics about KDD Cup 2019 AutoML



- Fact Sheets

- 45 Questions
- Types of questions:
  - Solution Detail: Preprocessing, Feature Engineering, Model Training
  - User Experience at this kind of problem and AutoML competition
  - Framework and tools
  - Suggestions

- Number of fact sheets received

- 38

- Verification

- Expert Check
- 2 of them are disqualified

12. How many times have you participated in AutoML challenges? \*

- 0
- 1
- 2
- 3
- > 3

13. Do you have experience dealing with time series data in your life or work? \*

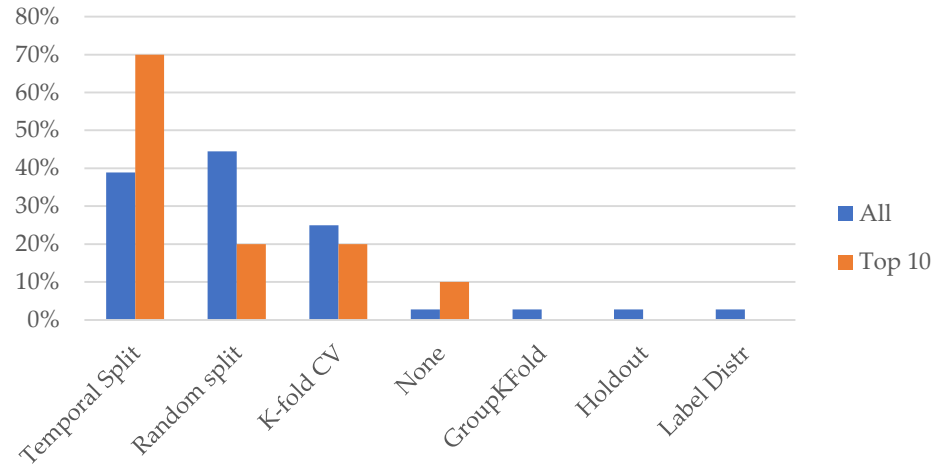
- Yes
- No

14. Do you think it's challenging to deal with concept drift in temporal data? \*

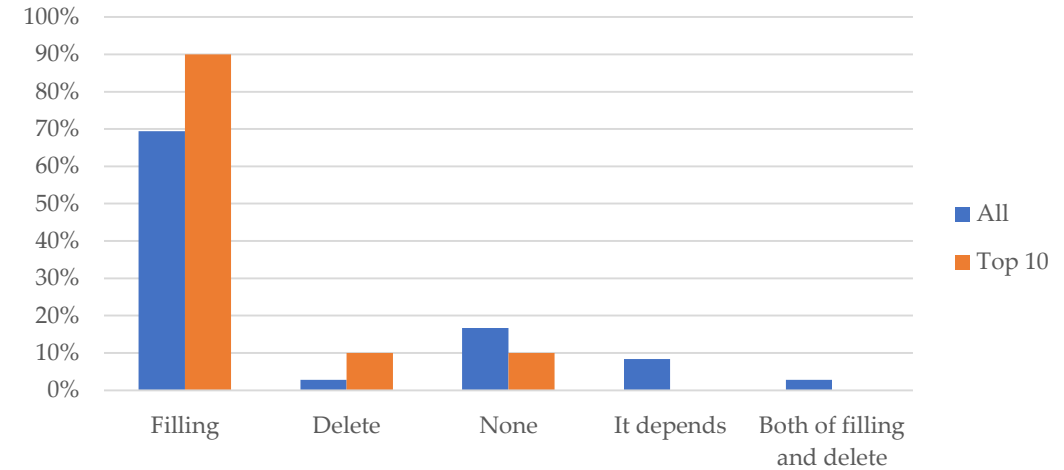
- Yes
- No

# Statistics – Data Preprocessing

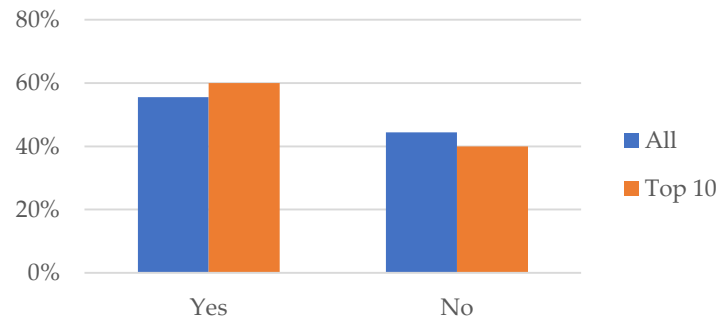
## Validation Split



## Missing Value



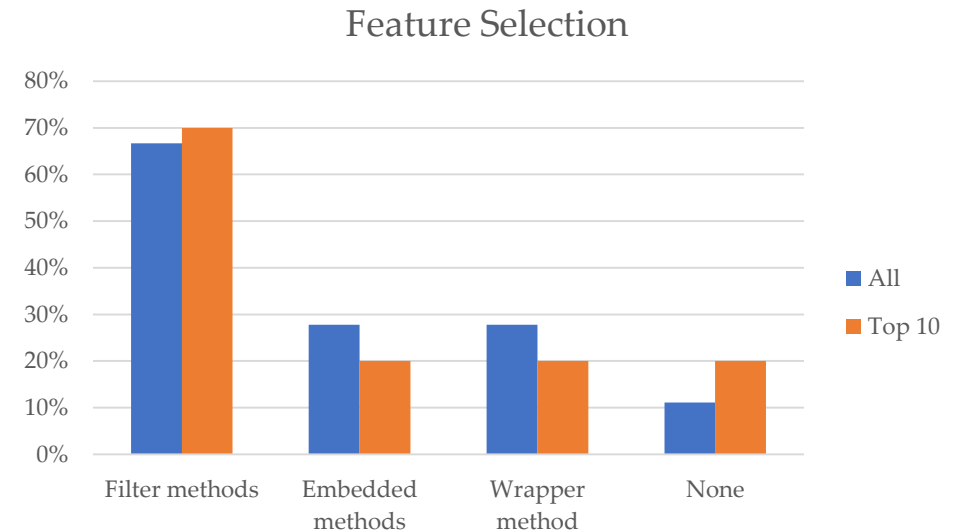
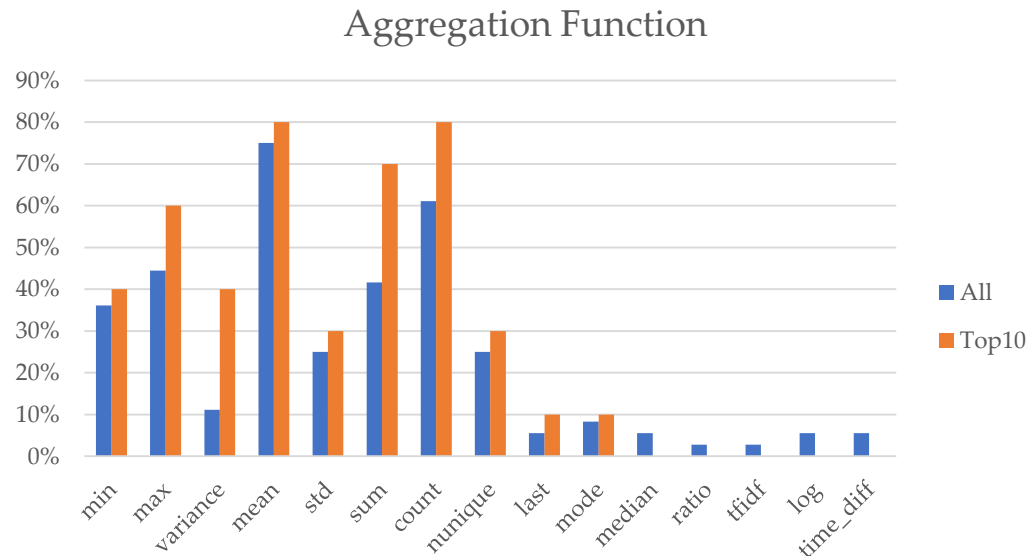
## Temporal Join?



- **Temporal Split** was more used by winners
- **Filling** was mostly used to deal with missing value
- About half teams used **temporal join**.

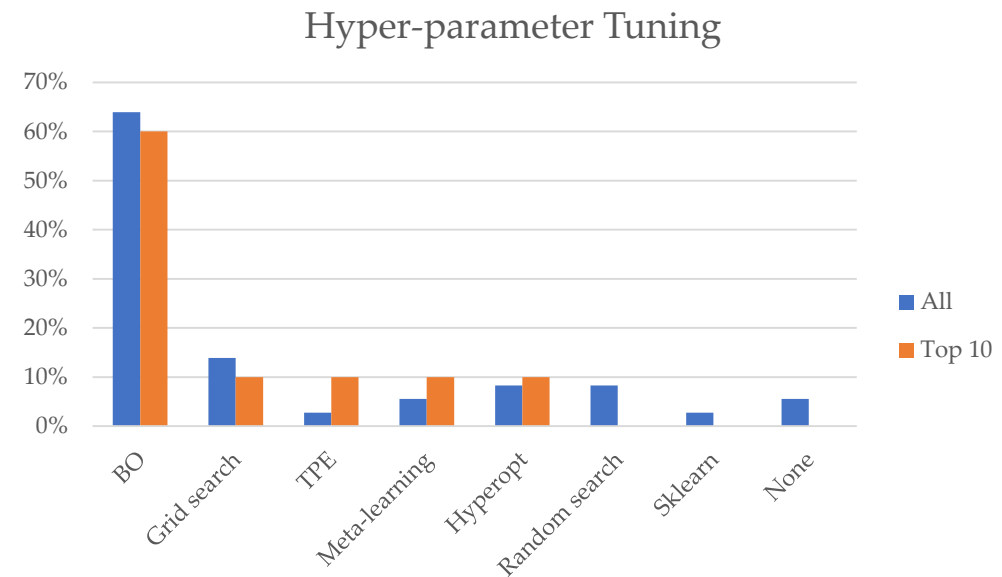
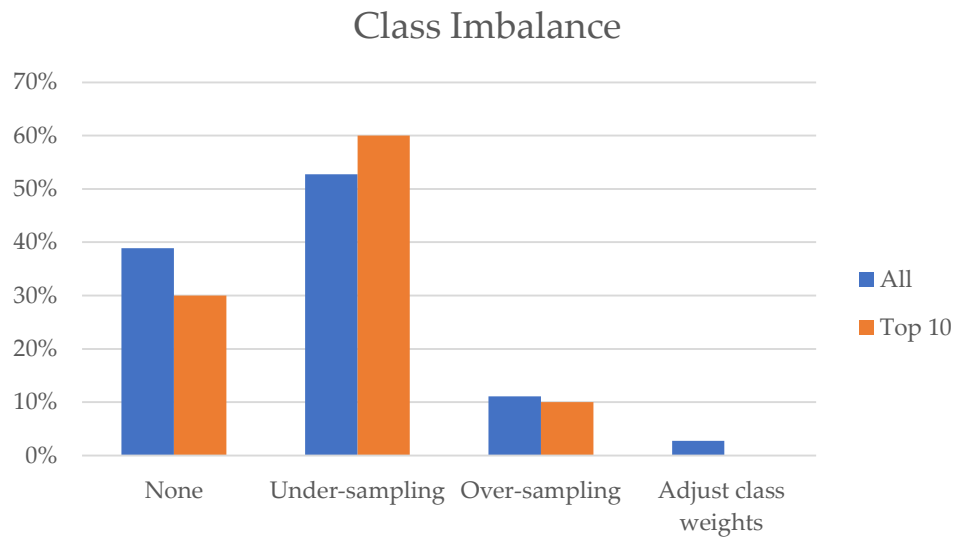


# Statistics – Feature Engineering



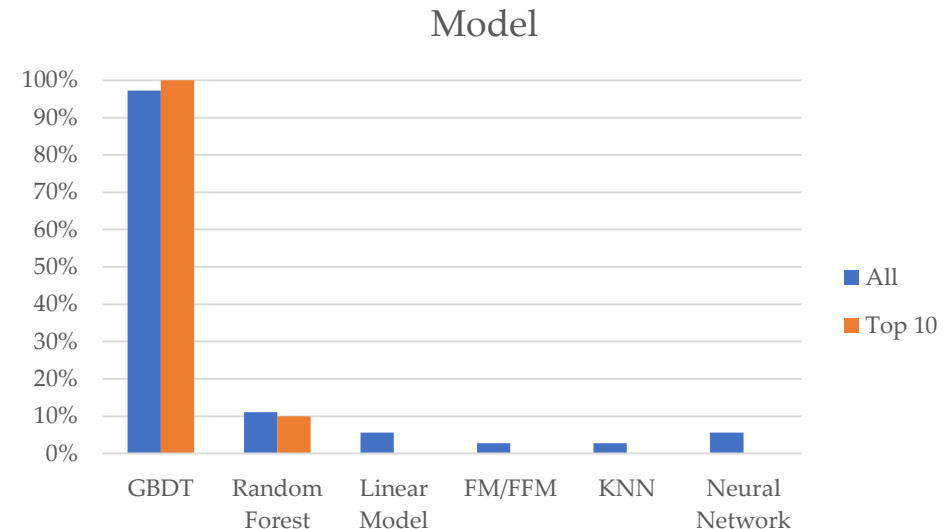
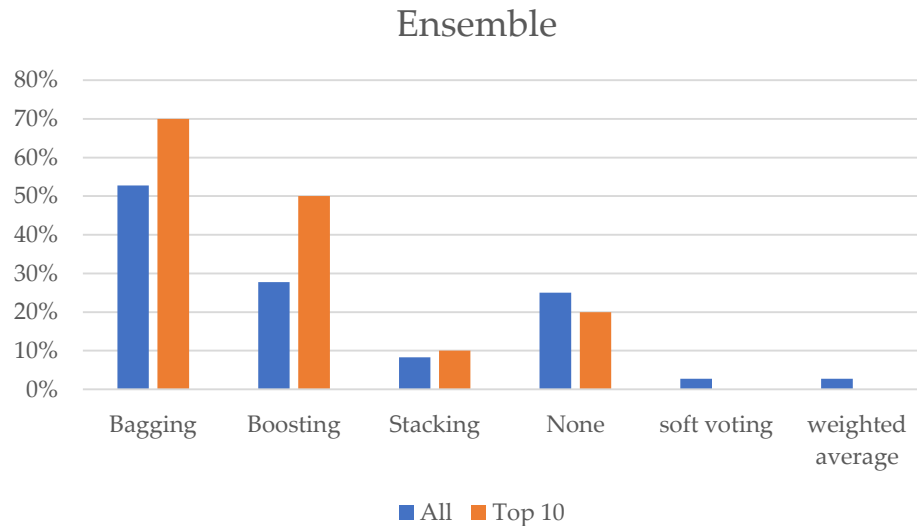
- Aggregation Function Top 5: **Count, mean, max, sum, min**
- **Filter Methods** were mostly used in feature selection

# Statistics – Model Training



- **Under-Sampling** was mostly used to deal with class imbalance problem
- **Bayesian Optimization** was mostly used to tune the hyper parameters

# Statistics – Model Training



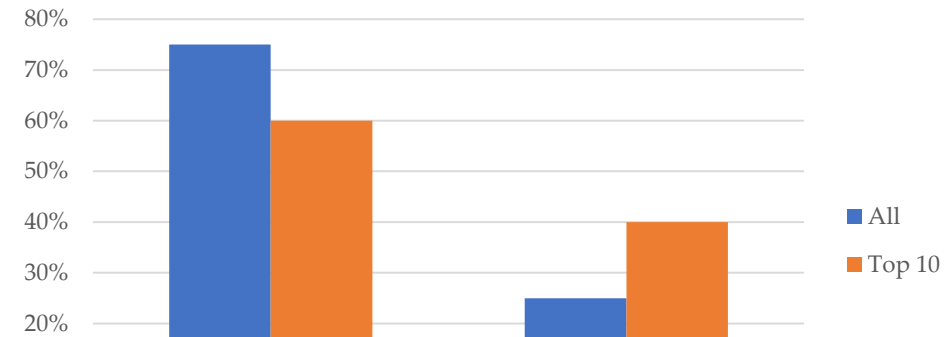
- Most teams used ensemble methods, mostly used were **Bagging and Boosting**
- **GBDT** was the most popular model! Dominated!

# Statistics – Participation Experience

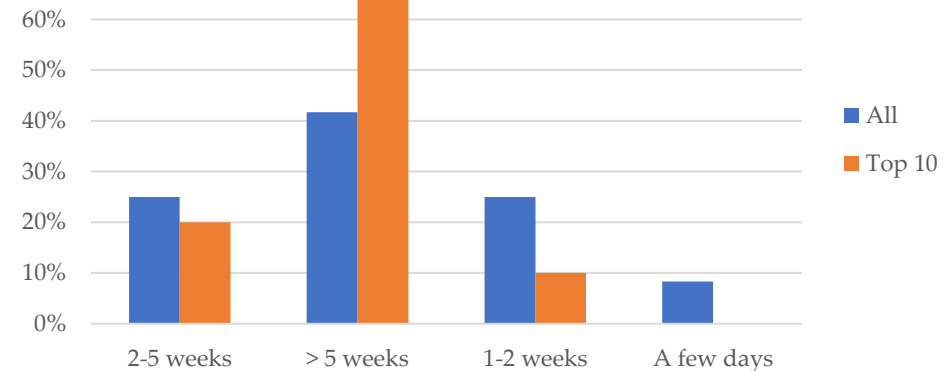
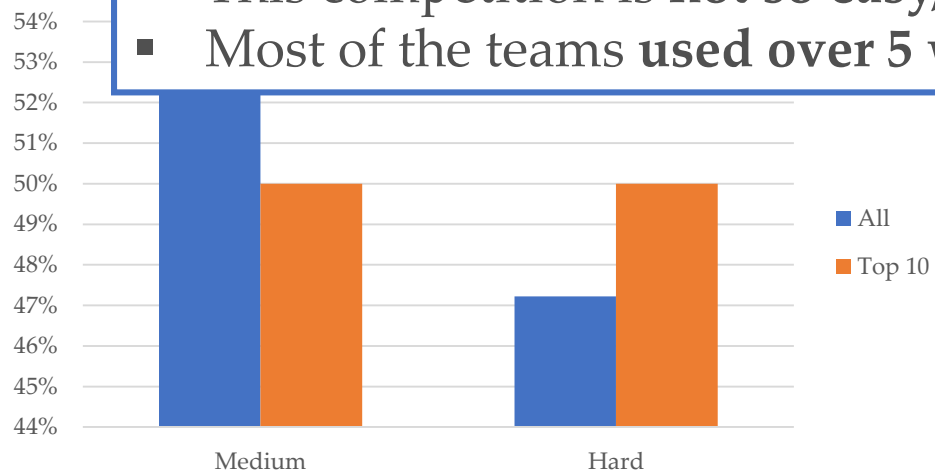
# Participation in AutoML competition



Temporal Data Experience

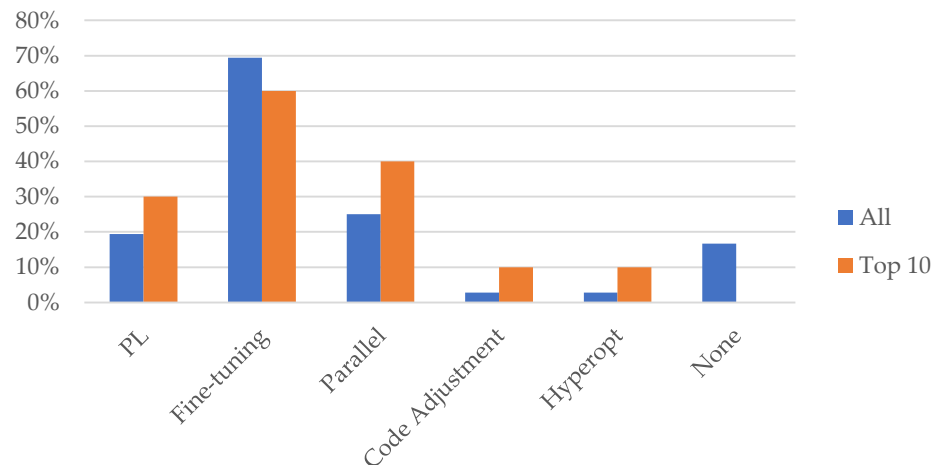


- Most of the participants are **freshers**.
- Most of the participants have experience dealing with **temporal data**.
- This competition is **not so easy, it's hard**.
- Most of the teams **used over 5 weeks** to build AutoML solutions.

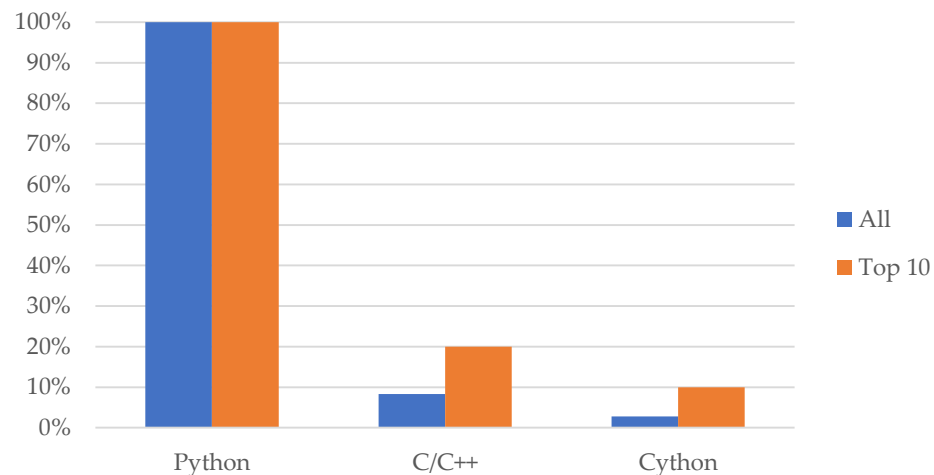


# Statistics – Framework

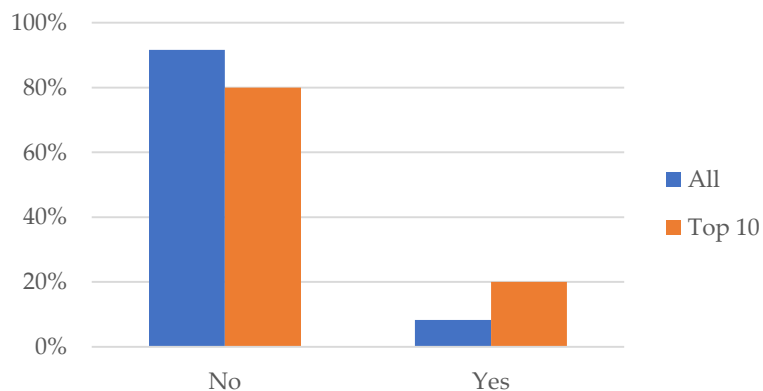
### Efficiency



### Programing Language

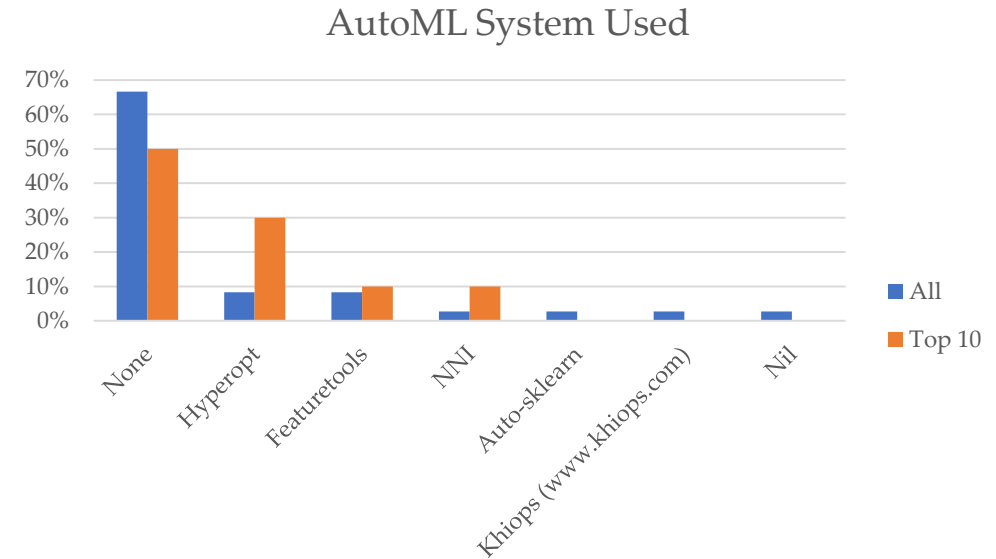
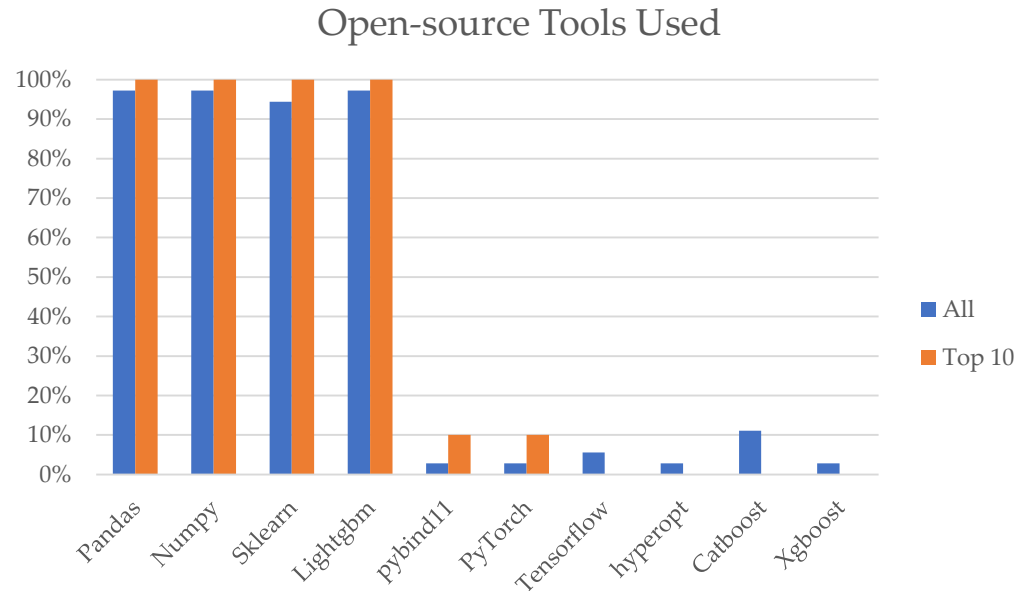


### Meta-Learning Used



- Most teams carefully **tuned** their codes and winning teams used more **parallel** tricks.
- **Python** was the most popular language! Dominated!
- **Very few meta-learning** methods was used.

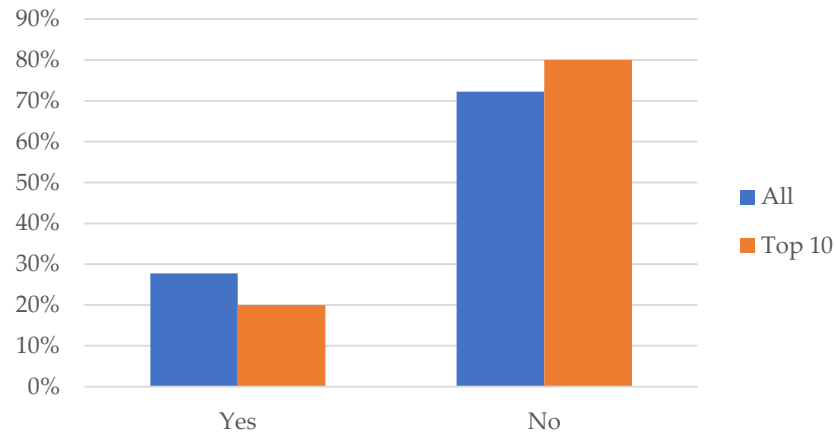
# Statistics – Tools



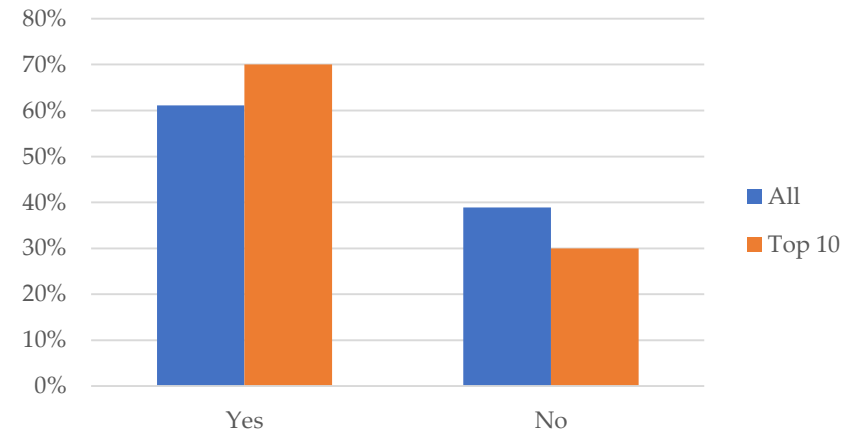
- Open-Source tools top 4 are: **Pandas, NumPy, SkLearn and LightGBM.**
- **Most of the teams** chose build their own AutoML systems.

# Statistics – User Experience at CodaLab

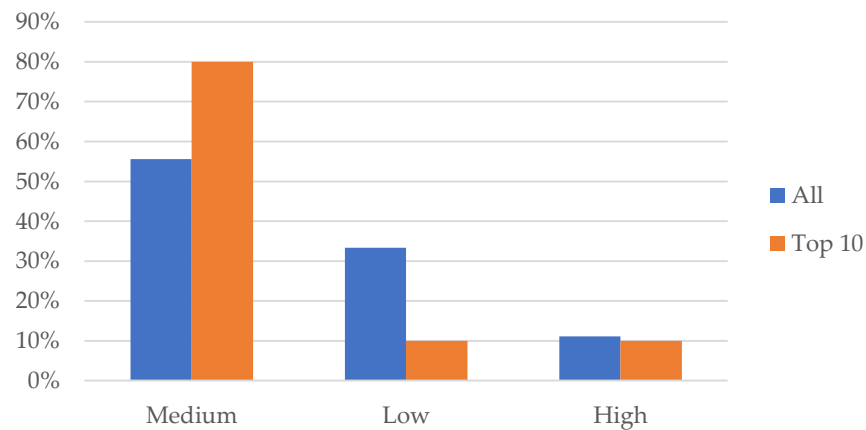
### Satisfaction of Stability



### Satisfaction of Response Speed



### Satisfaction of CodaLab



- User experience at CodaLab is just so so.
- CodaLab needs improvement especially the **stability** problem.

# Following Activities



- AutoCV2 at ECML PKDD 2019



- AutoNLP at WAIC 2019



- AutoML Special Issue at TPAMI

Call for papers - Special Issue  
Automation in AI and Machine Learning



- AutoDL Competition at NeurIPS 2019



- CiML(Challenges in Machine Learning) Workshop at NeurIPS 2019



# Sponsors



- Thanks to our sponsors!



**Let's welcome the Top Winners!**