Large-scale Open Dataset, Pipeline, and Benchmark for Bandit Algorithms

Yuta Saito[1,2] , Shunsuke Aihara[3],

Megumi Matsutani[3], and Yusuke Narita[1,4]

[1]Hanjuku-kaso Co, Ltd., [2]Tokyo Institute of Technology

[3]ZOZO Technologies, Inc. [4]Yale University.

REVEAL Workshop

https://sites.google.com/view/reveal2020/home?authuser=0

概要

- 推薦システムのバイアスの存在の指摘やその除去方法、バンディット・強化学習との関連に関するworkshop
- 2018年から3年連続で開催。今年はworkshopのなかで最多参加者数
- organizerやinvited talker, 参加者に有名な人が集まっており、口頭発表すると名を売ることができる
- **15本の採択論文のうち、4本のみに許される30分のlive talkを行ってきた**

Outline

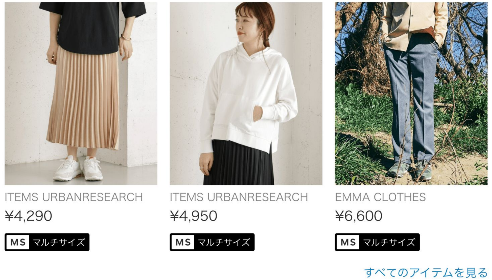
- overview of *off-policy evaluation*
- ***open bandit project*** (on-going)
  - open bandit dataset v1 (v2 will be released)
  - open bandit pipeline
- Q & A

Machine Learning for Decision Making (Bandit / RL)

We often use machine learning to make **decisions, not predictions**

decide which items to show

a coming user



often multiple items are recommended at the same time

observe reward (e.g., click)

Many Applications of “Machine Decision Making”

- news recommendation (by Yahoo)
- music/playlist recommendation (by Spotify)
- artwork personalization (by Netflix)
- ad allocation optimization (by Criteo)
- medicine
- education

OPEのモチベーション

We want to evaluate the performance of a *new decision making policy* using data generated by a *behavior, past policy*

Data Generating Process (contextual bandit setting)

Observes context vector *X* (e.g., a user visit)

A *policy* π selects an *action* *A* (e.g., a fashion item)

Observes reward *R* (e.g., a click indicator)

a ***policy*** interacts with the environment and produces the log data

本日の興味:**まだ見ぬ新たなpolicyの性能を評価すること**

Logged Bandit Feedback

We can use the *logged bandit feedback* collected by a *behavior (or past) policy* to estimate the policy value of a new policy

$$\mathcal{D} = \{(X_i, A_i, R_i)\}_{i=1}^{n}$$

$$A_i \sim \pi_b(a \mid X_i)$$

action choice by behavior policy

$$R_i \sim p(r \mid A_i, X_i)$$

observed reward

Estimation Target in Off-Policy Evaluation

In OPE, we aim to estimate the *policy value (policyの性能)*
of an *evaluation (or new)* policy

$$V(\pi_e) := \mathbb{E}_{p(x)\underline{\pi_e(a|x)}p(r|a,x)}[r]$$

expected reward obtained by running π_e on a real system

例えば、evaluation policyを仮にデプロイしたときの期待売り上げなど

Benefits of Off-Policy Evaluation

Policy value (policyの性能) を推定できると嬉しいことがたくさん

$$V(\pi_e) \approx \underline{\hat{V}(\pi_e; \mathcal{D})}$$

an estimated policy value of π_e using historical data $\mathcal{D}$

- avoid deploying poor performing policies
- identify promising new policies among many candidates

Direct Method (DM)

DM first estimates the expected reward and uses it to estimate the policy value

$$\hat{V}_{DM}(\pi_e; \mathcal{D}) = \mathbb{E}_n\left[\sum_{a \in \mathcal{A}} \pi(a \mid X_i)\, \underline{\hat{q}(X_i, a)}\right]$$

estimated expected reward

- **High bias** when the model is mis-specified
- **Low variance**

$$\mathbb{E}[r \mid a, x] \approx \hat{q}(x, a)$$

Inverse Probability Weighting (IPW) Estimator

IPW re-weighs observed rewards by importance weights

$$\hat{V}_{IPW}\left(\pi_e; \mathcal{D}\right) = \mathbb{E}_n\left[\underbrace{\frac{\pi_e\left(A_i \mid X_i\right)}{\pi_b\left(A_i \mid X_i\right)}}_{\text{importance weight}} R_i\right]$$

- **Consistent**
- **High variance** when old and new policies are largely different

Doubly Robust (DR) Estimator

DR uses DM as a baseline and applies IPW to shifted rewards

$$\hat{V}_{DR}(\pi_e; \mathcal{D}) = \underbrace{\hat{V}_{DM}(\pi_e; \mathcal{D})}_{\text{baseline}} + \mathbb{E}_n \left[\underbrace{\frac{\pi_e(A_i|X_i)}{\pi_b(A_i|X_i)} (R_i - \hat{q}(X_i, A_i))}_{\text{weighted shifted reward}} \right]$$

$$\mathbb{E}[r \mid a, x] \approx \hat{q}(x, a)$$

- **Consistent**
- **Locally Efficient**

Theoretical/Methodological Advances in OPE

手法/定式化について
詳しくはブログ記事へ！

- Self-Normalized IPW [Swaminathan and Joachims 2015]
- Switch Doubly Robust Estimator [Wang+ 2017]
- More Robust Doubly Robust Estimator [Farajtabar+ 2018]
- Hirano-Imbence-Ridder Estimator [Narita+ 2019]
- REG and EMP [Kallus & Uehara 2019]
- Doubly Robust with Shrinkage [Su+ 2020]

It seems the OPE community have made great progress over the years!

There are many other estimators in the reinforcement learning setting

Theoretical/Methodological Advances in OPE

手法/定式化について
詳しくはブログ記事へ！

- Self-Normalized IPW [Swaminathan and Joachims 2015]
- Switch Doubly Robust Estimator [Wang+ 2017]
- More Robust Doubly Robust Estimator [Farajtabar+ 2018]
- Hirano-Imbence-Ridder Estimator [Narita+ 2019]
- REG and EMP [Kallus & Uehara 2019]
- Doubly Robust with Shrinkage [Su+ 2020]

機械的意思決定の
刷新サイクルの到来か

「policy導入」-> 「bandit feedback収集」

-> 「OPEによるpolicyの更新/改善」-> 「policy導入」...

Issues with the current experimental procedures

OPEに関する論文の実験は全て

非現実的

- Synthetic or classification data (unrealistic)

or

再現不可能

- (Real, but) Unpublished data (irreproduceble)

We need real-world data enabling the ***“evaluation of OPE”***

Project's Goal and Components

We enable *realistic* and *reproducible* experiments on

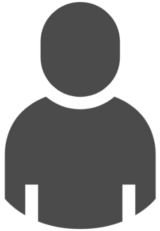
- Bandit Algorithms
- Off-Policy Evaluation (OPE)

"Open Bandit Dataset"

and ***"Open Bandit Pipeline"***

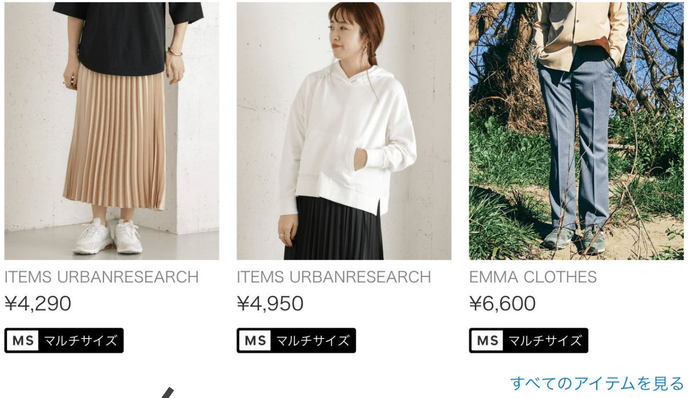
Overview of Open Bandit Dataset

user



Behavior Policy

ZOZOTOWN recommendation



$$\mathcal{D} = \{(X_i, A_i, R_i)\}_{i=1}^{n}$$

Schema of Open Bandit Dataset

	A		$\pi_b(\cdot \mid X)$	R	X	
timestamp	**item_id**	**position**	**action prob**	**click indicator**	**features**	**...**
2019-11-xx	25	1	0.0002	0	e2500f3f	...
2019-11-xx	32	2	0.043	1	7c414ef7	...
2019-11-xx	11	3	0.167	0	60bd4df9	...
2019-11-xx	40	1	0.0011	0	7c20d9b5	...
...	...	...	...	...	...	...

Essential Features of Open Bandit Dataset

- over 25M records collected by online experiments of bandit algorithms on a large-scale fashion e-commerce (*ZOZOTOWN*)

- **logged bandit feedback collected by *multiple* bandit policies**
    - *Uniform Random* (fixed)
    - *Bernoulli Thompson Sampling* (pre-trained)

最重要の特徴

➡ **enabling realistic experiments on OPE for the first time**

Comparison with Existing Real-World Bandit Datasets

Table 2: Comparison of Currently Available Large-scale Bandit Datasets

	Criteo Data (Lefortier et al. 2016)	**Yahoo! Data** (Li et al. 2010)	**Open Bandit Dataset** (ours)
Domain	Display Advertising	News Recommendation	Fashion E-Commerce
#Data	>= 103M	>= 40M	>= 26M (will increase)
#Behavior Policies	1	1	2 (will increase)
Random A/B Test Data	✗	✓	✓
Behavior Policy Code	✗	✗	✓
Evaluation of Bandit Algorithms	✓	✓	✓
Evaluation of OPE	✗	✗	✓
Pipeline Implementation	✗	✗	✓

Our Open Bandit Dataset

- contains ***multiple*** behavior policies
- enables the ***evaluation of OPE*** for the first time
- *comes with the pipeline implementations* (Open Bandit Pipeline)

既存データセットではOPEの
正確さの評価を行うことが不可能

Protocol for the Evaluation of OPE with Open Bandit Dataset

1. Prepare logged bandit feedback data collected by ***two different policies***

$$\mathcal{D}^{(1)} = \left\{ \left(X_i^{(1)}, A_i^{(1)}, R_i^{(1)} \right) \right\}_{i=1}^{n}$$

$$\sim p(x) \underline{\pi^{(1)}(a \mid x)} p(r \mid x, a)$$

$$\mathcal{D}^{(2)} = \left\{ \left(X_i^{(2)}, A_i^{(2)}, R_i^{(2)} \right) \right\}_{i=1}^{n}$$

$$\sim p(x) \underline{\pi^{(2)}(a \mid x)} p(r \mid x, a)$$

Protocol for the Evaluation of OPE with Open Bandit Dataset

2. Regard one policy as an ***evaluation policy*** and the other as a ***behavior policy***. Then, estimate the performance of the evaluation policy by OPE

$$V(\pi^{(1)}) \approx \hat{V}(\pi^{(1)}; \mathcal{D}^{(2)})$$

$\pi^{(1)}$: **evaluation policy**

$\pi^{(2)}$: **behavior policy**

- **The task here is to evaluate the estimation accuracy of** $\hat{V}$

Protocol for the Evaluation of OPE with Open Bandit Dataset

3. Regard the ***on-policy estimation*** of the policy value of the evaluation policy as the ground-truth policy value

$$V(\pi^{(1)}) = \mathbb{E}_{n(1)}[R^{(1)}]$$

we can do this on-policy estimation because we have $\mathcal{D}^{(1)}$ in our data

Protocol for the Evaluation of OPE with Open Bandit Dataset

4. Compare the estimated policy value with the ground-truth to evaluate the OPE estimator, for example, using the *relative estimation error*

$$\textit{relative estimation error of } \hat{V} = \left| \frac{\hat{V}\left(\pi^{(1)}; \mathcal{D}^{(2)}\right) - V\left(\pi^{(1)}\right)}{V\left(\pi^{(1)}\right)} \right|$$

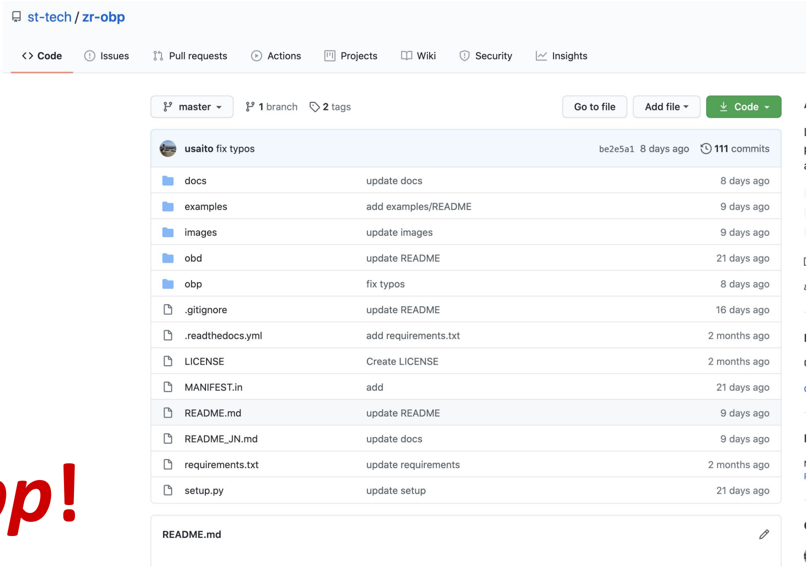
By applying this procedure to several estimators, we can do the "evaluation of OPE"

Open Bandit Pipeline (OBP)

We have implemented ***Open Bandit Pipeline (OBP)***

to streamline and standardize experiments on OPE

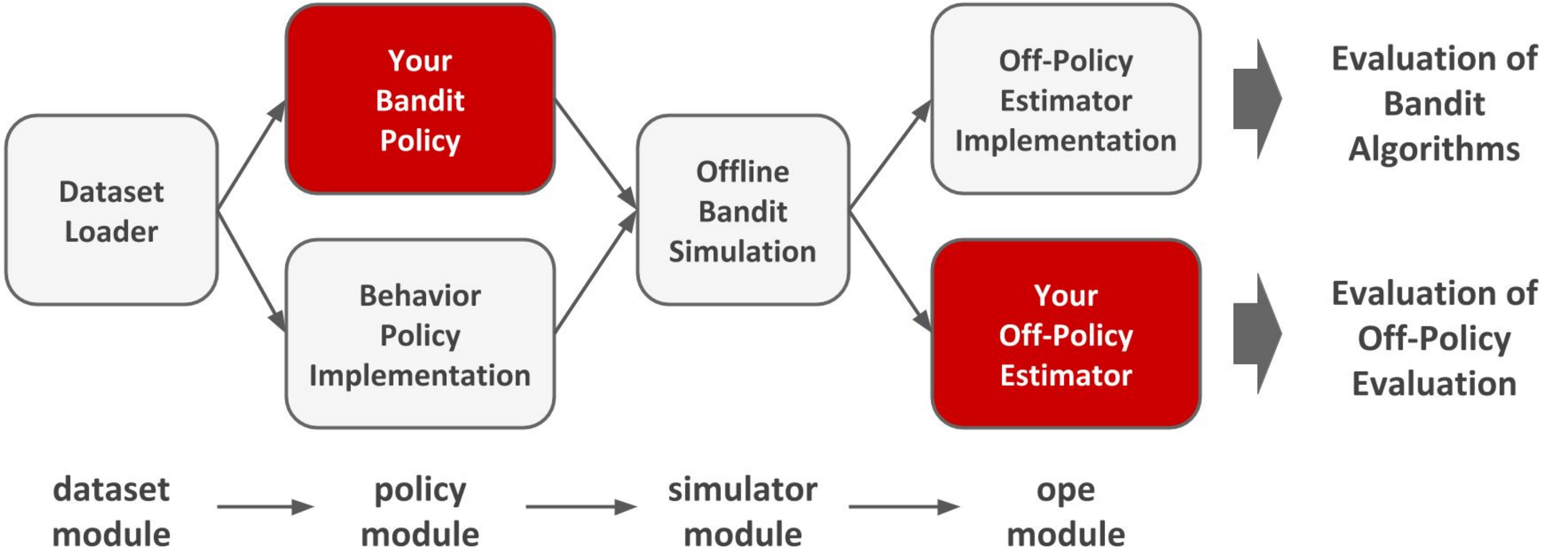
OPEN
BANDIT
PIPELINE ™



find out *zr-obp*!

Structure of Open Bandit Pipeline

OBP consists of **four main modules** (dataset, policy, simulator, and ope)



Proof of Concept Demo with Our Data and Pipeline

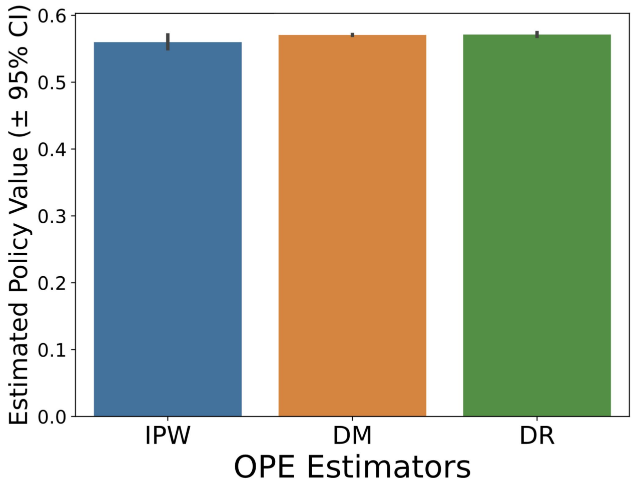
OPEN
BANDIT
PIPELINE™

Let me now run a quickstart example of OBP

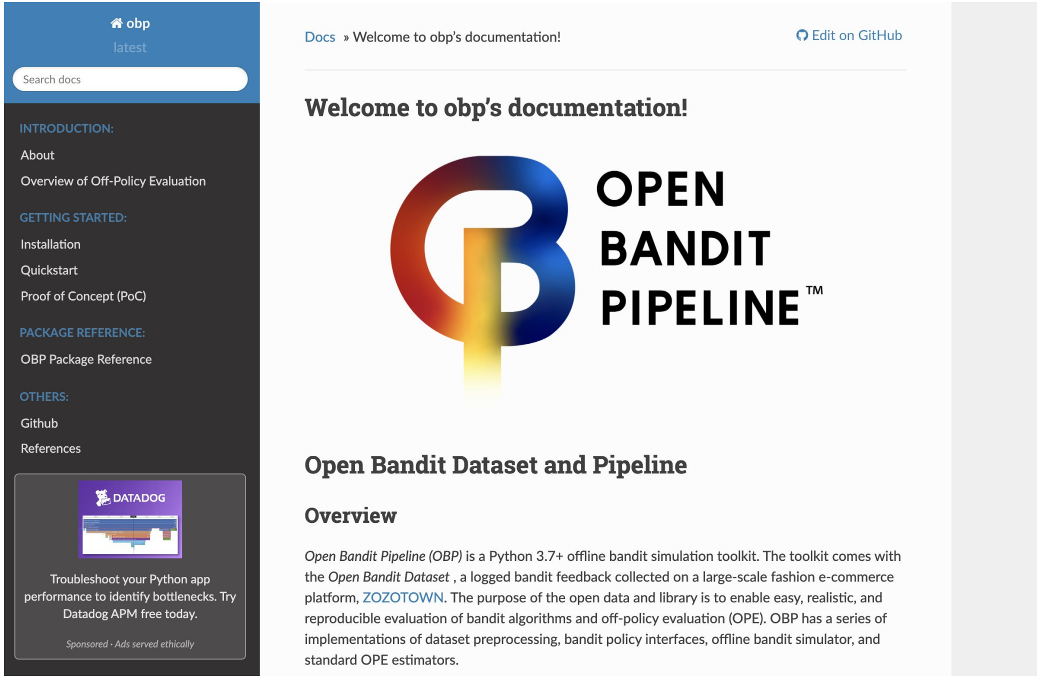
Other Nice Features

We can easily implement experiments on OPE or OPE itself with our OBP

```python
# a case for implementing OPE of the BernoulliTS policy using log data generated by the Random policy
from obp.dataset import OpenBanditDataset
from obp.policy import BernoulliTS
from obp.simulator import run_bandit_simulation
from obp.ope import OffPolicyEvaluation, ReplayMethod

# (1) Data loading and preprocessing
dataset = OpenBanditDataset(behavior_policy='random', campaign='women')
bandit_feedback = dataset.obtain_batch_bandit_feedback()

# (2) Offline Bandit Simulation
counterfactual_policy = BernoulliTS(n_actions=dataset.n_actions, len_list=dataset.len_list)
selected_actions = run_bandit_simulation(bandit_feedback=bandit_feedback, policy=counterfactual_policy)

# (3) Off-Policy Evaluation
ope = OffPolicyEvaluation(bandit_feedback=bandit_feedback, ope_estimators=[ReplayMethod()])
estimated_policy_value = ope.estimate_policy_values(selected_actions=selected_actions)

# estimated performance of BernoulliTS relative to the ground-truth performance of Random
relative_policy_value_of_bernoulli_ts = estimated_policy_value['rm'] / bandit_feedback['reward'].mean()
print(relative_policy_value_of_bernoulli_ts) # 1.120574...
```



Other Nice Features



We built a detailed documentation of Open Bandit Pipeline

Comparison with Existing Bandit Packages

Table 3: Comparison of Currently Available Packages of Bandit Algorithms

	contextualbandits	**RecoGym** (Rohde et al. 2018)	**Open Bandit Pipeline** (ours)
Synthetic Data Generator	✗	✓	✓
Support for Real-World Data	✗	✗	✓
Implementation of Bandit Algorithms	✓	✓	✓
Implementation of Basic Off-Policy Estimators	✓	✗	✓
Implementation of Advanced Off-Policy Estimators	✗	✗	✓
Evaluation of OPE	✗	✗	✓

Our Open Bandit Pipeline

- can *handle real-world bandit data* (including ours)
- implements *advanced OPE estimators* (SNIPW, Switch, MRDR, and DML)
- streamline *the evaluation of OPE*

Lots of Positive Reactions!

Atendra Gautam から 皆様 :
will the notebook be available ?

Faith Too から 皆様 :
https://github.com/st-tech/zr-obp
got this from his slides. i hope that is helpful to you

Maddie Shang から 皆様 :
Nice work!

Matt Corkum から 皆様 :
very nice indeed

Qiang Chen から 皆様 :
Very nice

Brian Regan から 皆様 :
Really nice work

Flavian Vasile から 皆様 :
Great work!

Dongzhenhua d00584022 から 皆様 :
Great great work!

Morten Arngren から 皆様 :
good stuff

Anuradha Uduwage から 皆様 :
Fantastic work

Tao Ye
@taoyeah

Really nice work presented at REVEAL workshop on "A Large Scale Open Dataset for Bandit Algorithms" by Yuta Saito. This includes evaluation tools for bandit algorithms and OPE. Impressive contribution. Looking forward to more contextual Data. github.com/st-tech/zr-obp #recsys2020

ツイートを翻訳

ZOZO Technologies
st-tech/zr-obp
Open Bandit Pipeline: a
and off-policy evaluati
github.com

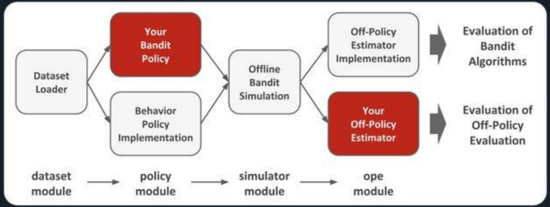
Karl Higley
@karlhigley

The Open Bandit Dataset and Open Bar
really impressive and look super-useful.
resource for the community:

github.com/st-tech/zr-obp

#RecSys2020 #recsys

ツイートを翻訳



usaito
@usait0

> There was also the release of Open Bandit Pipeline – a python library for bandit algorithms and off-policy evaluation that was considered as one of the highlights of this workshop.

ツイートを翻訳

RecSys 2020: Highlights of a Special Conference - inovex Blog
Read my take on the highlights of the 14th ACM Conference on Recommender Systems, such as the winners of the best long and short paper awards as well...
inovex.de

第五回統計・機械学習若手シンポジウムで招待講演

今日はお話しできなかった詳細・進捗について話す予定です

ホーム　プログラム　懇親

第5回 統計・機械学習若手シンポジウム

2020年12月3日（木）- 5日（土）

オンライン開催

開催概要

開催日時：2020年12月3日（木）～5日（土）

12月3日（木）：13:00 - 18:30

12月4日（金）：10:30 - 16:40

12月5日（土）：10:00 - 16:20

会場：**オンライン開催**

12月3-5日（zoom開催）

開催要旨

近年，いわゆる「人工知能」技術が様々な分野に垣根を超えて広がっていく様相を呈しています．本シンポジウムは，統計学および機械学習という現在の「人工知能」技術の基盤を支える分野の若手研究者を中心に，活発な議論・交流を目的として企画しました．本シンポジウムが，知見の共有にとどまらず，研究テーマの発見，共同研究の立ち上げといった新たな研究の発展に寄与することを期待します．

特別講演

福水健次 氏（統計数理研究所 教授）

講演タイトル「TBA」

企画講演：海外で活躍する若手研究者に迫る

Koichiro Shiba 氏（Harvard T.H. Chan School of Public Health Research Fellow）

講演タイトル「疫学研究における因果推論と機械学習：応用研究者の立場から事例紹介」

招待講演（敬称略）

岩澤有祐（東京大）

黒木祐子（東京大）

齋藤優太（東工大/半熟仮想（株））

篠崎智大（東京理科大）

菅澤翔之助（東京大）

竹野思温（名工大）

寺田吉壱（大阪大）

野沢健人（東京大）

幡谷龍一郎（東京大）

林直輝（NTTデータ数理システム/東工大）

林祐輔（Japan Digital Design（株））

横田達也（名工大）

https://sites.google.com/view/statsmlsymposium20/

Thank you!

github: https://github.com/st-tech/zr-obp

google group: https://groups.google.com/g/open-bandit-project/members

dataset: https://research.zozo.com/data.html

blog: https://techblog.zozo.com/entry/openbanditproject

press: https://corp.zozo.com/news/20200818-11223/