MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

[answer1] The dataset is released to address the gap in publicly available Online Controlled Experiment (OCE) dataset that can support research on the end-to-end design and running of OCEs with adaptive stopping. For a detailed motivation of the problem, please refer to the "Datasets for Online Controlled Experiments" paper.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

[answer2] This dataset is created by the AI & Data Science Platform at ASOS.com, in collaboration with the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning at Imperial College London and University of Oxford (StatML.IO).

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

[answer3] The dataset, as part of the wider research work on surveying and building a taxonomy for OCE datasets, is part funded by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning at Imperial College London and University of Oxford (StatML.IO) and ASOS.com.

[question4] Any other comments?

[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

[answer5] The dataset contains one relational table, with each row containing a set of business performance measurements for a particular experiment, treatment, and decision metric, covering a specific time period of the said experiment. Please refer to the schema hosted on the Wiki page on the OSF project for further information.

[question6] How many instances are there in total (of each type, if appropriate)?

[answer6] There are 24,153 rows in the dataset, recording the results of 78 online controlled experiments. Each experiment has 2–5 variants (including one variant acting as the control), with a total of 99 treatments being experimented. Four decision metrics, specific to

the business unit which ran the experiments, are captured
with a daily or bi-daily frequency.

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] The dataset records the results of experiments run
by a specific business unit in ASOS.com, a global
online fashion retail platform mainly targeting 20-
somethings based in North America, Europe, and
Oceania. The experiments are run at various times
over a 2-year period (2019 to 2020). Within the said
scope the dataset contains all possible rows for all
valid experiments. Rows associated to experiments
that the team had to abandon due to incorrect setup
on the experimentation platformhave been removed
as they do not carry any learning value.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Each row consists of six entries:
• Number of users in the control group
• Sample mean of the responses (under the corresponding
decision metric) from users in the control
group
• Sample variance of the responses from users in the
control group
• Number of users in the treatment group
• Sample mean of the responses from users in the
treatment group
• Sample variance of the responses from users in the
treatment group
Please refer to the dataset schema for further information.
All responses are aggregated up to the control/
treatment group level via the use of summary
statistics. The dataset does not identify or record any
individual responses from users.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] The six entries mentioned above act as the "label/
target" for one to run statistical tests on.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] A small number of experiment/treatments (5 out of
99 treatments under Metric IDs 2, 3, and 4) are missing
the variance entries due to an issue in capturing
the underlying data.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] The rows are linked via the experiment ID, variant
ID, and (decision) metric ID. The experiment ID
uniquely identifies an experiment. The metric ID
uniquely identifies a metric. The combination of the
experiment ID and the variant ID uniquely identifies
a treatment—The same variant ID can be seen
in multiple experiments and hence can not be used
to identify a treatment

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] N/A for the running of OCEs as it is the norm to
apply a statistical test to all available data. We also
have no specific data split recommendations when it
comes to learning the statistical test hyperparameters
from the data.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] There are no duplicated rows in this dataset. However,
common pitfalls in the running and analysis of
online controlled experiments like sample ratio mismatch,
novelty effect and telemetry loss may lead to
unexpected data points. Some possible examples include:
• The treatment mean entry for Experiment
'162a38', Metric 1, taken 38 days since the start
of the experiment demonstrates a spike as compared
to that of the control mean and treatment in
surrounding time points, which could be a result
of telemetry loss.
• The treatment mean entries for Experiment
'54a85a' are increasing faster than the control
mean entries of the same experiment, indicating a
possible primacy effect in play.
Note we are unable to definitively tell whether the
said pitfall applies aswe are observing a real-life system
without a ground truth. We intentionally leave
these data in as we believe they are reflective of what
an OCE practicioner will face in their day job. We
hope the dataset will attract methods that address
the underlying problem(s).

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or

relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] The dataset is self-contained, as designed.

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15] No. All entries in the table that are considered business
sensitive have been anonymized (e.g. experiment
ID) or removed (e.g. name of experiment,
start date/time of experiment) to prevent third parties
linking the entry back to a particular experiment
being done on the digital platforms

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] No.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] No. The dataset does not identify any subpopulations
based on protected characteristics. The dataset
does record the aggregated activity attained by the
treatment and control groups, which users are randomly
assigned into

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18] No. It is impossible to identify an individual from
this data as each instance aggregates the activity of
at least a hundred thousand individuals.

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of

government identification, such as social security numbers; criminal history)? If so, please provide a description.

[answer19] No.

[question20] Any other comments?

[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] The dataset is aggregated from a raw clickstream data, which records the browsing behavior of individual users. The rawclickstream data is *not* released along with this dataset.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] The aggregation is done via a series of proprietary data processing pipelines within ASOS.com.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] No sampling was carried out when aggregating the numbers. This is to ensure the calculated decision metrics are accurate.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] The data collection is done via digital platforms built by ASOS.com and related third-parties. The plat forms are general-purpose and not built specifically to create this dataset.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] The dataset describes results from OCEs that were run in years 2019 and 2020. The results are cleaned and curated in May 2021.

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26] Each OCE run within the business unit is subject to a
review process, which among many things, include
ethical considerations.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27] The data is ultimately based upon users' browsing
behavior, which is recorded on digital platforms built
by ASOS.com and related third-parties

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

[answer28] Yes. Users are notified of the data collection policy[1]
and privacy policy via a banner upon visiting the
website. The privacy policy describes how the information
will be used under the section "How we use
your information":
 "We also anonymise and aggregate personal information
(so that it does not identify you) and use it
for purposes including testing our IT systems, research,
data analysis, improving our site and app,
and developing new products and services."
"We use the data we collect to help us provide you
with the best service, the best shopping experience
and to show you the latest and greatest products
and services that we think you will love."

[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

[answer29] Yes. Users have to consent to the data collection in
order to be able to use the service.

[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

[answer30] Yes. Users are made aware of ASOS.com's Cookies
Notice and Privacy Policy.The latter provides a
mechanism for users to revoke their consent in the

future / for certain use subject to the relevant Data
Protection laws.

[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer31] The Data Protection Officer at ASOS.com has reviewed the dataset and its schema, and given each measurement is reflective of the collective activity of hundreds of thousands to millions of users, determined there are no data protection risk from the publication of this dataset.

[question32] Any other comments?

[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

[answer33] The measurements contained in this dataset is aggregated from a raw clickstream dataset via a series of proprietary data pipelines. Some manual data cleaning is carried out to remove invalid experiments that do not carry any learning value.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

[answer34] The raw clickstream data that this dataset is derived from is proprietary to ASOS.com. The data is stored subject to applicable laws and regulations but is not (and will not be) publicly available. **[question35] Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

[answer35] No. The data processing pipelines that curate this dataset are proprietary to ASOS.com.

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] We demonstrated in the "Datasets for Online Controlled Experiments" paper that the dataset can support
the development in OCE methods via the following:
1. Running of meta-analyses
2. Design and running of experiments with adaptive
stopping
3. Acting as a quasi-benchmark for adaptive stopping
methods
For details, please refer to Section 5.1 of the "Datasets
for Online Controlled Experiments" paper.

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

[answer38] No. Although we invite individuals who have benefited
from the use of this dataset to cite either the
dataset or the broader work which also provides
a survey and taxonomy on publicly available OCE
datasets. The resultant citation network may provide
a glimpse on the dataset's impact

[question39] What (other) tasks could the dataset be used for?

[answer39] Being both a multi-experiment and time series
dataset, we believe the dataset can also be useful for:
    1.   Bias (of the estimator) detection across time; and
2. Learning of correlation structure across experiments,
metrics, variants and time
See Section 5.1 of the "Datasets for Online Controlled Experiments" paper for a more detailed discussion

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] We urge potential users of this dataset to exercise
caution when attempting to generalize the learnings.
Examples of generalizing the learnings include the
use of this dataset as a full performance benchmark
and directly applying the value of hyperparameter(
s) obtained while training a model on this dataset
to another dataset.
Please refer to Section 5.2 of the "Datasets for Online
Controlled Experiments" paper for a detailed discussion
on this matter.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] The dataset is made available with the intent to support
development in the statistical methods required
to run OCEs. The experiment results shown in the
dataset is not representative of ASOS.com's overall

business operations, product development, or experimentation
program operations, and no conclusion
of such should be drawn from this dataset.

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] Yes. The dataset is open-sourced under a CC-By Attribution
4.0 International (CC BY 4.0) License.

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] The dataset can be downloaded from Open Science
Framework (OSF) via the link https:// osf.io/64jsb/. The DOI for this dataset is
10.17605/OSF.IO/64JSB

[question45] When will the dataset be distributed?

[answer45] The dataset is currently available to the public.

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] Yes, the dataset is distributed under a CC-By Attribution
4.0 International (CC BY 4.0) License.

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47] No.

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48] No.

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The primary contributor, C. H. Bryan Liu, will be the
main contact for support, hosting, and maintenance
of the dataset.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] The primary contributor, C. H. Bryan Liu, can be
contacted via the email address bryan.liu12 (at)
imperial.ac.uk.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] No. In the case where one is necessary, we will update
the datasheet to reflect the change.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] We commit to update the dataset in the case where
major error(s) in the measurements are found. We
have no plans to add new rows describing the result
of more recent experiments in the short term.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

[answer54] N/A. The rows concern business measurements and
does not record the activity of an individual directly.

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

[answer55] We commit to supporting older versions in the case
where we are in the position to add new instances
corresponding to measurements arising from new
experiments. If minor error(s) in the measurements
were to be found, we will update the existing dataset
(i.e. new version of the data file under the same file
name on OSF) and datasheet. If major error(s) in
the measurements were to be found, we will release
a new dataset (i.e. upload the data file under a different
file name on OSF) and add an errata to the
publication associated with the dataset. The old, erroneous
version would be discontinued (i.e. hidden
from view on OSF).

[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not?

Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] There are no mechanisms/process developed for this purpose yet. We welcome suggestions from any potential contributors.

[question57] Any other comments?

[answer57] No.