Advancing the Applicability of Machine Learning Verification
Application DMP

Questionnaire

The questions in this section should only be answered if you are currently applying for FWO funding.
Are you preparing an application for funding?

- No

Advancing the Applicability of Machine Learning Verification
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

GDPR

Have you registered personal data processing activities for this project?

- No

Advancing the Applicability of Machine Learning Verification
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Covtype	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
FashionMnist	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Higgs	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Miniboone	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Mnist	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Prostate	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Roadsafety	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Sensorless	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Vehicle	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<100MB	\
Webspam	common ML benchmark	reuse existing data	digital	Compiled/aggregated data	.csv	<1GB	\
dtai-veritas	repository to represent and work with ensembles of decision trees	reuse existing data	digital	Software	Python and C++ scripts	<100MB	\

tree_compress	repository to compress ensembles of decision trees	reuse existing data	digital	Software	Python scripts	<100MB	\
Paper-specific repositories	For each published research paper, a dedicated repository will be made publicly available. These will typically contain computational scripts to reproduce the results of the paper.	generate new data	digital	Software	Python scripts	<100MB	\

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- covtype: https://www.openml.org/d/1596
- fmnist: https://www.openml.org/d/40996
- higgs: https://www.openml.org/d/42769
- miniboone: https://www.openml.org/d/44128
- mnist: https://www.openml.org/d/554
- prostate: https://www.openml.org/d/45672
- roadsafety: https://www.openml.org/d/45038
- sensorless: https://archive.ics.uci.edu/dataset/325
- vehicle: https://www.openml.org/d/357
- webspam: https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#webspam

- dtai-veritas: https://github.com/laudv/veritas
- tree_comrpess: https://github.com/laudv/tree_compress

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The benchmark datasets are commonly used in ML literature and are therefore already extensively documented.
The software dependencies are freely available and documented in GitHub, and can be installed from the Python Package Index (PyPI) using tools like `pip`.
The paper-specific repositories will be coded in Python and documented with a README.txt file, as is conventional, that explains the general structure of the repository, and that contains instructions on how to use it and how to exactly reproduce the results presented in the paper.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

3. Data storage & back-up during the research project

Where will the data be stored?

All the mentioned benchmark datasets are commonly used datasets in the ML research field, that are stored by other parties. Specific references for each dataset were given above.
The mentioned software dependencies are publicly available and maintained on GitHub.
Paper-specific repositories will also be hosted on GitHub.
In addition to this, all of the above will be also stored on a filesystem provided by the Department of Computer Science of KU Leuven.

How will the data be backed up?

KU Leuven’s Department of Computer Science provides various storage options that are backed up regularly.

Concretely, the Declarative Languages and Artificial Intelligence (DTAI) research group offers storage and back-up services which use limited redundancy (RAID-Z2) to its PhD students, in which snapshots of all text, source code, data and presentations are stored. This project will make use of this service for the purpose of backing up all research data, by using the service to store snapshots for all resulting publications.

Using these storage and back-up services, all research data will be retained for a period of at least 10 years after publication or the end date of the research project grant agreement.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

The service used for backup purposes has plenty of capacity available; much more than the data we plan to use and produce in this project (<100GB).

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

All the data we use are publicly available.

The paper-specific repositories will be made public at an appropriate time point (i.e., upon acceptance of the associated paper). Until then, they will be securely stored as private repositories on GitHub, as well as locally on my personal machine and on a fileserver at the Department of Computer Science.

Only user accounts created by the department can gain access to the departmental storage. The local filesystem also allows setting access control per file/directory.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

All storage and backup costs are covered by the DTAI research group.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data listed above will be retained for at least 5 years after the end of the project.

Where will these data be archived (stored and curated for the long-term)?

For every result in this project, a long-term snapshot will be created on the DTAI NetApp storage system (>40TB). This snapshot contains everything to be able to reproduce results with some minor effort. Moreover, it contains the final paper, presentations given as well as any result-specific DMP.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

All data preservation backup costs are covered by the DTAI research group.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

Data are publicly available.

If access is restricted, please specify who will be able to access the data and under what conditions.

\

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Links to datasets were provided above.
Code will be put on Github.

When will the data be made available?

Data are publicly available.
Paper-specific repositories will be made public at an appropriate time point (i.e., upon acceptance of the associated paper).

Which data usage licenses are you going to provide? If none, please explain why.

Data are publicly available under the CC-BY license.
Software will be available under the Apache License, Version 2.0.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

All data sharing costs are covered by the DTAI research group.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Lorenzo Cascioli

Who will manage data storage and backup during the research project?

Lorenzo Cascioli

Who will manage data preservation and sharing?

Lorenzo Cascioli

Who will update and implement this DMP?

Lorenzo Cascioli