DMP title

Project Name My plan (FWO DMP) - DMP title

Project Identifier u0133704

Grant Title 1166222N

Principal Investigator / Researcher Lorenzo Perini

Description My research is mainly theoretical and aims at developing machine learning models able to deal with specific tasks. I work on mathematical proofs as well as on programming code. Benchmark datasets available online are used to validate the proposed models.

Institution KU Leuven

1. General Information

Name applicant

Lorenzo Perini

FWO Project Number & Title

FWO Project number: 1166222N

FWO Project title: Measuring and Exploiting the Uncertainty in Anomaly Detection

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Source code uses the following:

- Python: numpy, pandas, scikit-learn, ipython
- Jupyter, Anaconda;
- Cython;
- Python frameworks: PyOD, anomatools.

No datasets are collected, only existing datasets are re-used.

Type of data	Format	Volume	How created
Source code	Python, Cython	0.4 MiB	Self-written

We use the following freely available benchmark datasets:

1. Outlier detection datasets from the DAMI library (format arff, volume less than 5Mb): https://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/
2. IoT datasets for anomaly detection from UCI (format csv, volume between 10Mb and 15Mb): https://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT\#
3. Three Wind Turbines datasets for anomaly detection (format csv, volume 70Mb, 146 Mb, and 72Mb): http://www.industrial-bigdata.com/Home
4. Artificially generated datasets from https://scikit-learn.org/stable/modules/outlier_detection.html

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for

Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

Privacy Registry Reference:
Short description of the kind of personal data that will be used:

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- Yes

The source code is authored by Lorenzo Perini, and is available on GitHub under the Apache 2.0 License.

- https://github.com/Lorenzo-Perini/Confidence_AD
- https://github.com/Lorenzo-Perini/Active_PU_Learning
- https://github.com/Lorenzo-Perini/StabilityRankings_AD
- https://github.com/Lorenzo-Perini/pyod

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

Public benchmark daatasets listed in Sec.2 are released with no licese

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

Documentation for the source code is provided as part of the code repository.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

NA

5. Data storage and backup during the FWO project

Where will the data be stored?

All the data is freely available to anyone online.
Code is released on GitHub (https://github.com/Lorenzo-Perini)

How is backup of the data provided?

The source code is available on GitHub, and is backed-up internally in the form of a long-term snapshot.

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

Code (Jupyter notebook and python .py file) has little volume (max 3-4 Mb) with respect to the amount available.

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

DTAI has its own large-scale (TB), long-term, NetApp storage system with backup. Costs are covered by the DTAI section and potentially the bench fee can contribute.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

DTAI's NetApp system only allows access for team members (connection via VPN/SSH). Files can be made read-only.
DTAI suggests a hash data integrity policy. No data nor code is stored on a private laptop.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

After every project is fulfilled, a snapshot of all text, source, code, data, presentations is collected by the first ML-KU Leuven author (we zip everything together). This is to allow for reproducible research.

Where will the data be archived (= stored for the longer term)?

The data will be stored on the university's central servers (with automatic backup procedures) for at least 10 years, conform the KU Leuven RDM policy

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

DTAI has its own large-scale (TB), long-term, NetApp storage system with backup. Costs are covered by the DTAI section. Moreover, DTAI demands 10 years of data preservation.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

Data is freely available online, while the source code is released on GitHub.

Which data will be made available after the end of the project?

Data is freely available online, while the source code is released on GitHub.

Where/how will the data be made available for reuse?

- In an Open Access repository

The source code will be released on GitHub.

When will the data be made available?

- Immediately after the end of the project

The full material will be uploaded immediately after the end of the project.

Who will be able to access the data and under what conditions?

Data is freely available online, while the source code is released on GitHub. No restrictions are applied.

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

No costs are expected, as GitHub offers a free online service. In case of unexpected additional costs, the FWO bench fee may be used.

8. Responsibilities

Who will be responsible for data documentation & metadata?
I will be responsible for data documentation and metadata.

Who will be responsible for data storage & back up during the project?
I will be responsible for data storage and backup.

Who will be responsible for ensuring data preservation and reuse ?
I will be responsible for ensuring data preservation and reuse.

Who bears the end responsibility for updating & implementing this DMP?
The PI bears the end responsibility of updating & implementing this DMP.