# Generalizable and Scalable Decision-Focused Learning

*A Data Management Plan created using DMPonline.be*

**Creator:** n.n. n.n.

**Affiliation:** KU Leuven (KUL)

**Funder:** Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

**Template:** FWO DMP (Flemish Standard DMP)

**Project Administrator:** n.n. n.n.

**Grant number / URL:** 11PQ024N

**ID:** 205736

**Start date:** 01-11-2023

**End date:** 31-10-2027

**Project abstract:**

Supervised machine learning methods have been deeply studied and widely used in practice. The predictive models involved are typically trained to minimize the prediction error with respect to a given set of historical data. However, in many applications, making accurate predictions is not the true end goal. Typically, the predictions play some role in a larger context; their purpose is to aid good decision making.

In recent years, there has been a growing interest in the research area of decision-focused learning, which develops ways to explicitly train models to make predictions that lead to good decisions. However, existing methods suffer from two large limitations. First, they do not generalize towards decision-making tasks that they have not been trained on. Second, they are not scalable, as they integrate the solving of difficult optimization problems into the predictive model's training loop.

This project will tackle these issues. We will start by investigating task generalization, by developing task representations, and building a new attention-based architecture to process these representations. Then, we will improve scalability in two different ways. First, we will develop a technique that only has to solve some of the optimization problems that occur during learning. Second, we will build a novel neural architecture that can be trained end-to-end to produce both predictions and decisions, without requiring the use of a conventional solver.

**Last modified:** 26-04-2024

## Generalizable and Scalable Decision-Focused Learning
## 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:*<br>• Generate new data<br>• Reuse existing data | *Please choose from the following options:*<br>• Digital<br>• Physical | *Please choose from the following options:*<br>• Observational<br>• Experimental<br>• Compiled/aggregated data<br>• Simulation data<br>• Software<br>• Other<br>• NA | *Please choose from the following options:*<br>• .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, …<br>• NA | *Please choose from the following options:*<br>• <100MB<br>• <1GB<br>• <100GB<br>• <1TB<br>• <5TB<br>• <10TB<br>• <50TB<br>• >50TB<br>• NA | |
| SPO dataset | Commonly used synthetic dataset in decision-focused learning, originating from the seminal Smart "Predict, then Optimize" paper | Reuse existing data | Digital | Other (synthetically generated data) | .csv | <100MB | / |
| Blackbox dataset | Commonly used synthetic dataset in decision-focused learning, originating from the seminal Differentiation of Blackbox Combinatorial Solvers paper | Reuse existing data | Digital | Compiled/aggregated data | .npy | <100GB | / |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CORA dataset | Dataset containing scientific publications and their associated citation network. Commonly used in a graph learning context | Reuse existing data | Digital | Compiled/aggregated data | .csv | <100MB | / |
| DFL repository | A repository of DFL methods and computational scripts developed throughout the project | Generate new data | Digital | Software | Python scripts | <100MB | / |
| Paper-specific repositories | For each published research paper, a dedicated repository will be made publicly available. These repositories will generally consist of the computational scripts used to obtain the results detailed in 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:**

**SPO DOI:** 10.1287/mnsc.2020.3922
**Blackbox DOI:** 10.17617/3.YJCQ5S
**CORA DOI:** 10.3886/E109167V2-11132

**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

Additional note - the pre-existing datasets we will use have the following licenses:

- SPO dataset: CC-BY-4.0
- Blackbox dataset: CC-BY-4.0
- CORA dataset: CC-BY-4.0

**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 SPO, Blackbox and CORA datasets are commonly used in DFL literature and are therefore already extensively documented.

The DFL repository that will be developed throughout the research project will consist of Python scripts that will follow the PEP 8 style standard, and that will be documented following standard software documentation best practices. The goal is to create a repository that practitioners can use to easily experiment with both existing and new DFL methods on several datasets from the literature. This repository will also come 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.

The paper-specific repositories will be coded and documented in a similar fashion to the DFL repository, but will additionally contain procedures, along with instructions on how to use them, to reproduce the exact experimental results included in the associated 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?**

The SPO, Blackbox and CORA datasets are commonly used datasets in the DFL research field that are stored by other parties.

More concretely,

The SPO dataset is not a fixed dataset, but rather an algorithmic description of how to generate the synthetic data they are using. The description is detailed in doi.org/10.1287/mnsc.2020.3922, and public implementations of the data generator are available at https://github.com/paulgrigas/SmartPredictThenOptimize and https://github.com/khalil-research/PyEPO. The exact data that will be produced in this project using these generators will be quite small (<100MB), and will hence be stored in GitHub repositories associated with the various research papers coming out of this project.

The Blackbox dataset is available at doi.org/10.17617/3.YJCQ5S

The CORA dataset is available at doi.org/10.3886/E109167V2-11132

The DFL repository will be hosted on GitHub

The paper-specific repositories will also be hosted on GitHub

In addition to this, all of the above will be stored on the locally, on my own machine, as well as on a filesystem provided by the Department of Computer Science.

**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 a NetApp storage service 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

Both the Department of Computer Science and the DTAI research group offer various storage and backup options. The local filesystem at the department does not impose a quota on its users. Similarly, the NetApp 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?**

The SPO, Blackbox and CORA datasets are publicly available datasets.

The DFL repository and paper-specific repositories will be made public at an appropriate time point (i.e., when it is ready in case of the DFL repository, and upon acceptance of the associated paper, in case of the paper-specific repositories). 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.

**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 10 years after the end of the project.

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

All data will be archived on the NetApp storage service offered by the Department of Computer Science in the form of snapshots of all text, source code, data and presentations.

**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

The SPO, Blackbox and CORA datasets are publicly available datasets.
The DFL repository and paper-specific repositories will be made public at an appropriate time point (i.e., when it is ready in case of the DFL repository, and upon acceptance of the associated paper, in case of the paper-specific repositories).

**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.**

The SPO, Blackbox and CORA datasets are publicly available datasets, to which links are provided above.
The DFL repository and paper-specific repositories will be made available on GitHub.

**When will the data be made available?**

The SPO, Blackbox and CORA datasets are publicly available datasets.
The DFL repository will be made publicly available when it is ready for general use (i.e., when some state-of-the-art methods and existing datasets are implemented, and the intended use of the repository is properly documented). The paper-specific repositories will be made public upon acceptance of the associated paper.

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

Our newly generated data will be the DFL and paper-specific repositories, for which we will use the MIT License.

**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?**

Senne Berden

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

Senne Berden

**Who will manage data preservation and sharing?**

Tias Guns

**Who will update and implement this DMP?**

Senne Berden