

# Future-Proofing Research Data Repositories: Keeping up with the *ML/AI* Revolution

Stephanie Labou  
Data Science Librarian  
UC San Diego Library

# Before we start – AI vs ML

Artificial intelligence (AI) and machine learning (ML) are often used interchangeably, but there are differences

To summarize from Google Cloud and Amazon Web services pages some relevant differences:

- ML is an application of AI where models use statistical methods to identify patterns and extract meaning from very large amounts of data
- ML relies on structured and semi-structured data whereas AI can use unstructured data

# There is no escaping ML

- ML is rapidly becoming popular across research domains
- Social interest in ML (and AI) and questions about transparency
- Various gov memos/directives about sharing data and research outputs

*ML-related repository deposits are inevitable, if not already happening - and will only increase*

# First things first

*We need to better understand how ML objects are being shared by practitioners, so we can identify common practices as well as gaps and barriers to findability and reusability.*



# A sampling of repositories with ML content

## Specialist

- UC Irvine Machine Learning Repository
- OpenML
- Kaggle

## Generalist

- Figshare
- Zenodo
- Dryad
- Harvard Dataverse
- UC San Diego Library Digital Collections

# There's a lot of ML content in repositories, and it comes in a lot of different formats

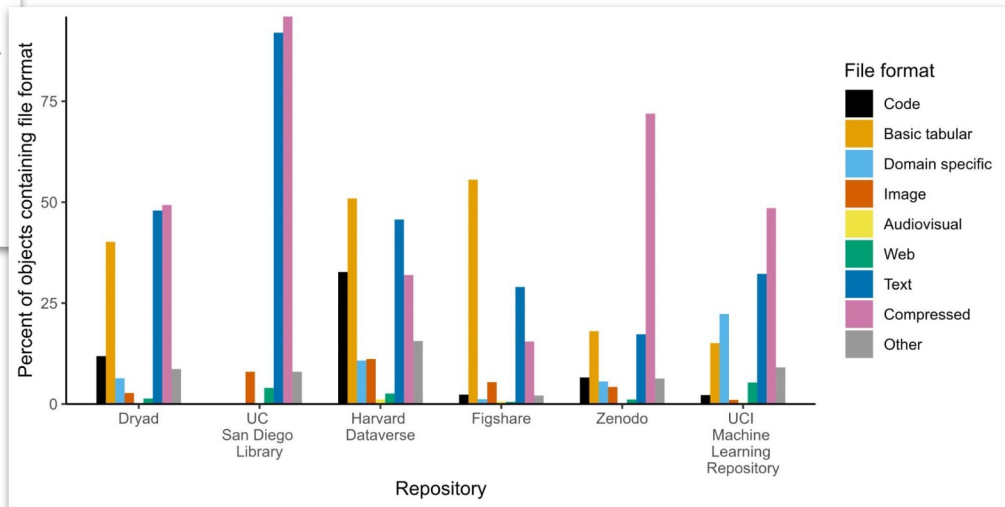
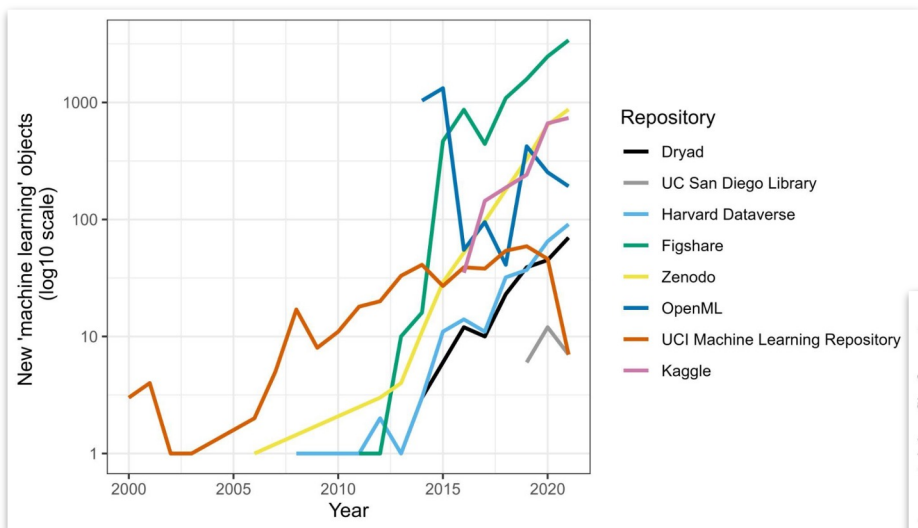


Figure 1 and Figure 2 from Labou et al. 2024 (forthcoming)

# How to make repositories more amenable to ML (and eventually AI) content

Unambiguous citations

Emphasize related resources

Clear licensing

**Rich(er) metadata**

Labels for compressed files

**Access at scale**



# Findability: search strategy & use case are different

zenodo machine learning

Versions

View all versions

**Access status**

- Open 82,419
- Restricted
- Embargoed

**Resource types**

- Publication
- Dataset
- Software
- Presentation
- Other
- Image 1,221
- Poster

**File type**

- PDF 59,987
- ZIP 10,385
- DOCX 2,970
- CSV 1,666
- TXT 1,620
- JPG 1,352
- XLSX 1,322
- GZ 1,281
- BIN 1,257
- PNG 928

figshare machine learning

need help?

**Content Type**

- item (186,484)
- collection (43,364)
- project (634)

Select date ▾

**Item Type**

- dataset (51,767)
- journal contribution (50,3...)
- figure (39,119)
- thesis (11,781)
- conference contribution (...)

**Licence**

- CC BY 4.0 (110,461)
- In Copyright (16,544)
- CC BY-NC 4.0 (14,661)
- All Rights Reserved (13,450)
- CC BY + CC0 (7,467)

show more

**Source**

- Publisher (157,446)
- Institution (57,593)
- figshare.com (13,217)
- Preprint (1,230)
- Funder (598)

UC Irvine Machine Learning Repository

Keywords ▾

Data Type ▾

Subject Area ▾

**Task** ▾

- # Features ▾
- # Instances ▲

Less than 100

100 to 1000

More than 1000

**Feature Type** ▲

- Numerical
- Categorical
- Mixed

OpenML

- Datasets** 5.6k
- Tasks** 261.5k
- Flows** 17.8k
- Runs** 10.1M
- Collections** 198
- Benchmarks**
- Task Types** 8

**Task\_types**

Task types define a machine-readable schema for specific machine learning tasks created on specific datasets, and all ensuing runs can be evaluated against it.

- Subgroup Discovery**
- TBA
- 8 years ago

# ML-useful metadata



UC Irvine  
Machine Learning  
Repository

## Dataset Information

### Additional Information

The data is related with direct marketing campaigns of a Portuguese banking institution on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:

- 1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
  - 2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
  - 3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
  - 4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
- The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

SHOW LESS ^

### Has Missing Values?

No

## Variables Table

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
variance	Feature	Continuous		variance of Wavelet Transformed image		no
skewness	Feature	Continuous		skewness of Wavelet Transformed image		no
curtosis	Feature	Continuous		curtosis of Wavelet Transformed image		no
entropy	Feature	Continuous		entropy of image		no
class	Target	Integer				no

## Additional Variable Information

For Further information about the features see the features file in the data folder.

Additional metadata fields: number of instances, task type

Even just a structured README template or features file (since adding additional metadata fields can be a big ask)

Your system may vary, but any way to enable faceting/findability by *structure* of data (or code, or other)

Make repository  
content ML/AI-  
ready

(or at least, more  
ready than now)

First use case was how do we make ML-related objects in repositories more suited to how ML research is conducted in practice

Second use case is making content in repositories (which may or may not be explicitly ML) amenable to ML or AI methods

Libraries especially collect good data!

→ impactful, often publicly funded, well documented

# Ease of (manual) bulk downloads

Ease of access at the scale needed for ML/AI is paramount to surface and make content visible and high value

If it's too challenging to download manually, or web scrape, or otherwise access programmatically, end users may look elsewhere

Also worth considering as an explicit aspect of accessibility in a FAIR context

# Ease of (manual) bulk downloads

Search this collection...

Showing results for 1 - 20 of 48

100 Island Challenge Collection

< 1 2 3 > Go

Sort: date created 20 per page Advanced Search

ITEM

**Palau 2022-11 (Expedition) - Helen (Island) - Raw Images**

Part of: 100 Island Challenge Collection

Name: Hatohebel Community; OneReef; Sandin Lab, Scripps Institution of Oceanography, UC San Diego; Ministry of Agriculture, Fisheries, and the Environment, Republic of Palau

Date: 2022-11

Topic: Large-area imagery, Coral reef

Format: image; data

ITEM

**Palau 2022-11 (Expedition) - Sonsorol (Island) - Raw Image**

Part of: 100 Island Challenge Collection

Name: Sonsorol community, Sandin Lab, Scripps Institution of Oceanography, OneReef, Ministry of Agriculture, Fisheries, and the Environment, Republic of Palau

Date: 2022-11

Topic: Coral reef, Large-area imagery

Format: image; data

ITEM

**Palau 2022-01 (Expedition) - Palau North (Island) - Raw Image**

Part of: 100 Island Challenge Collection

Name: Waitt Institute, Ministry of Agriculture, Fisheries, and the Environment, Sandin Lab, Scripps Institution of Oceanography, UC San Diego; OneReef, Republic of Palau Department of Conservation and Law Enforcement

Date: 2022-03

Topic: Large-area imagery, Coral reef

Format: image; data

Palau 2022-11 (Expedition) - Helen (Island) - Raw Images

Component 6 of 12

**Raw Images**

File Size	134 GB
File Format	ZIP Format
Scope And Content	Raw images collected from the survey site.
Technical Details	This resource includes files in the following formats: nef and .txt

Download file View file

Collection

- 100 Island Challenge Collection

Cite This Work

100 Island Challenge Project (2024). Palau 2022-11 (Expedition) - Helen (Island) - Raw Images. In 100 Island Challenge Collection. UC San Diego Library Digital Collections. <https://doi.org/10.6075/J01Z44N3>

Description

In November of 2022, the 100 Island Challenge team visited Helen to collect large-area imagery as part of the Palau 2022-11 expedition. The imagery collected here resurveys of permanent sites established during the Palau 2017-06 expedition. For each site, a preview image and the raw digital images used to create composite large-area image products of the underwater sites have been made available. A general description of the methods used to acquire the images can be found in a README.txt file for each site. Site specific metadata can be found within a METADATA.txt file for each site.

DeFungi

Donated on 1/28/2023

DOWNLOAD

CITE

1 citations

16750 views

Citations/Acknowledgements

If you use this dataset, please cite:  
Please cite the below paper published in arXiv.org if you use the dataset.

Paper link:  
<https://arxiv.org/abs/2109.07322...>

Keywords

Dataset Characteristics

Image	Subject Area	Associated Tasks
Real	Computer Science	Classification

Feature Type

Real	# Instances	# Features
	9114	

Dataset Information

For what purpose was the dataset created?  
The dataset was created to develop a machine-learning algorithm for detecting and classifying Fungi images.

Who funded the creation of the dataset?  
No funder.

What do the instances in this dataset represent?  
Photos

C:\Users\slabou\Downloads\defungi.zip\H1\

Name	Size	Packed Size
H1_1a_1.jpg	15 549	15 157
H1_1a_2.jpg	15 354	14 961
H1_1a_3.jpg	12 738	12 185
H1_1a_4.jpg	13 254	12 759
H1_1a_5.jpg	15 923	15 524
H1_1a_6.jpg	16 037	15 642
H1_1a_7.jpg	14 730	14 331
H1_1a_8.jpg	15 067	14 656
H1_1a_9.jpg	13 991	13 548
H1_1a_10.jpg	12 345	11 787
H1_1a_11.jpg	16 114	15 730
H1_1a_12.jpg	18 114	17 759
H1_1a_13.jpg	14 853	14 464
H1_1a_14.jpg	13 796	13 347

# API (Application Programming Interface)

Changing unit of bundling objects, so users can more easily bulk download content of interest, is likely a big ask, requiring big back end changes

An API (in theory) would allow users to programmatically download the metadata & data for their items of interest

Caveat: an API is *also* a big ask, in terms of up front development, potential security concerns, and sustainability

# Decisions, decisions...

To be clear: Not every repository or each piece of repository content may be suitable! Repositories don't have to be all things to all end users

However, consider:

- Are collection items likely to be of interest to ML (or AI) practitioners?
- Is there a benefit to having repository items higher visibility/use in the area of ML/AI?
- Where and how (and if) to invest additional resources to change ingest process, metadata fields, and/or access methods?

## Acknowledgements

This project was funded by a Librarians Association of the University of California (LAUC) Research Grant and the UC San Diego Library Research Data Curation Program.

Forthcoming paper in *Journal of eScience Librarianship*

Co-authors:

Abby Pennington (Library)

Ho Jung Yoo (Library)

Michael Baluja (UC San Diego grad student)