Developing Machine Learning for Metadata Standardization with HTCondor and Snakemake

Throughput Computing 2023

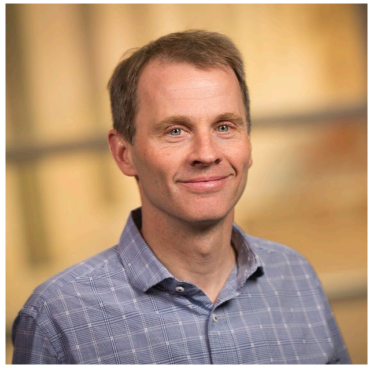
Department of Biostatistics and Medical Informatics
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Yuriy Sverchkov, Mark Craven, Colin Dewey

Funding and people

American Family Funding Initiative 2020 Award

Colin Dewey


Mark Craven

Yuriy Sverchkov

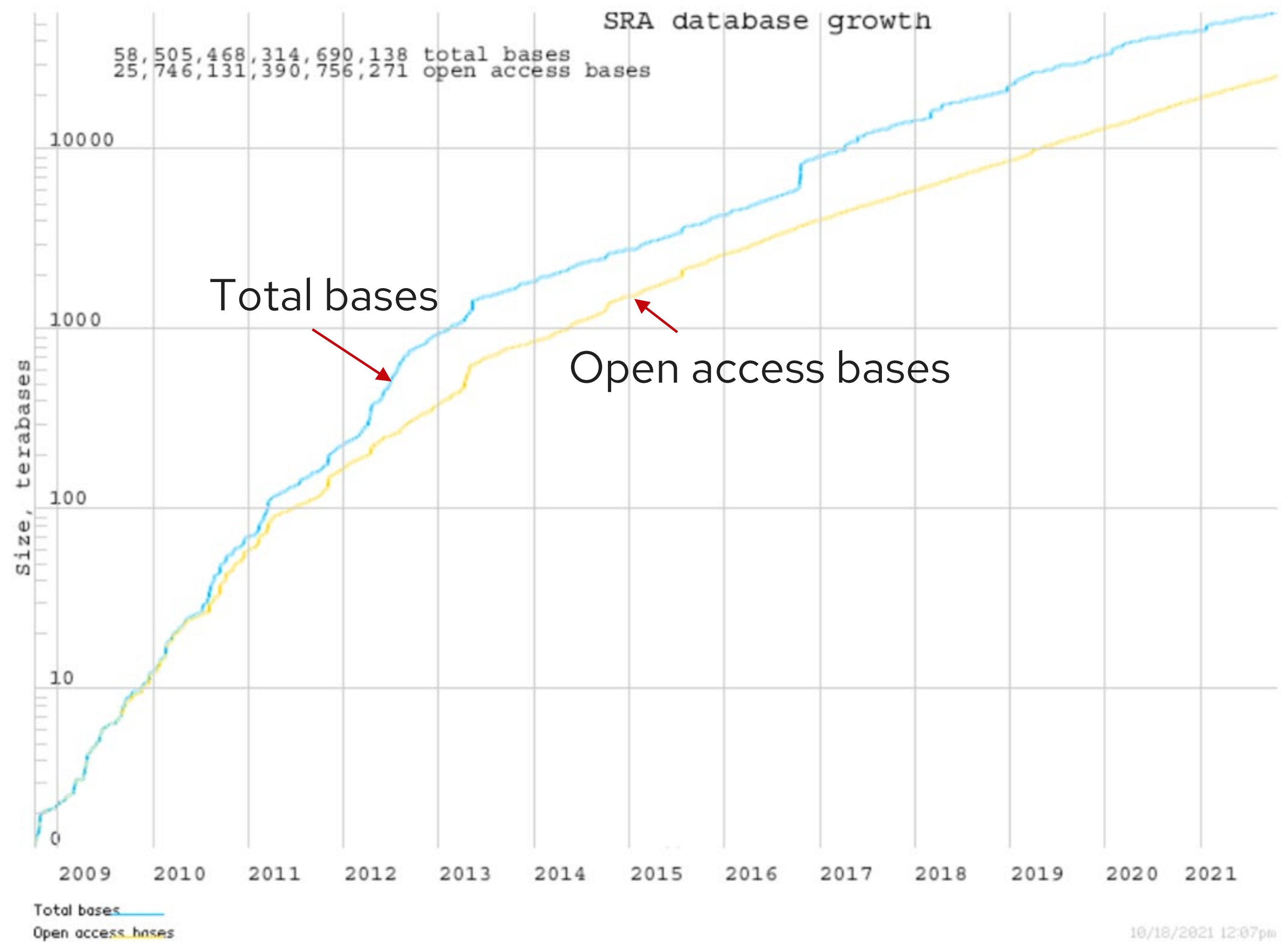
Jeremiah Yee

Abrar Majeedi

Tapanmitra Ravi

An abundance of (meta)data

- **The NCBI's Sequence Read Archive (SRA)**
- The largest publicly available repository of high throughput sequencing data
- Opportunity for aggregate analyses
- Metadata are free-text key-value pairs composed by individual submitters



Free-text metadata is hard to mass-process and query

Property	Property value
cell line	BJ
cell type	fibroblasts
harvest time	72 hours post seeding
lentiviral transgenes	no
passage number	passage 5
source_name	Human forskin fibroblast, BJ (ATCC, CRL-2522)

References to external sources

Key derives meaning from value

Spelling errors

Metadata standardization task

Free-text key-value pairs

Property	Property Value
Cell type	PBMC
Disease state	Lyme disease
Individual	Patient 01-20
Source_name	PBMCs
Time	Acute Lyme pre-treatment (V1)

Property	Property Value
Biomaterial_ provider	AllCells, 1301 Harbor Bay Parkway Suite 200 Alameda, CA 94502, USA
Bioproject_id	PRJDB3120
Cell_line	Peripheral blood mononuclear cells
Description	Healthy person 5
Sample_name	N5
Sample_title	Small cellular RNA

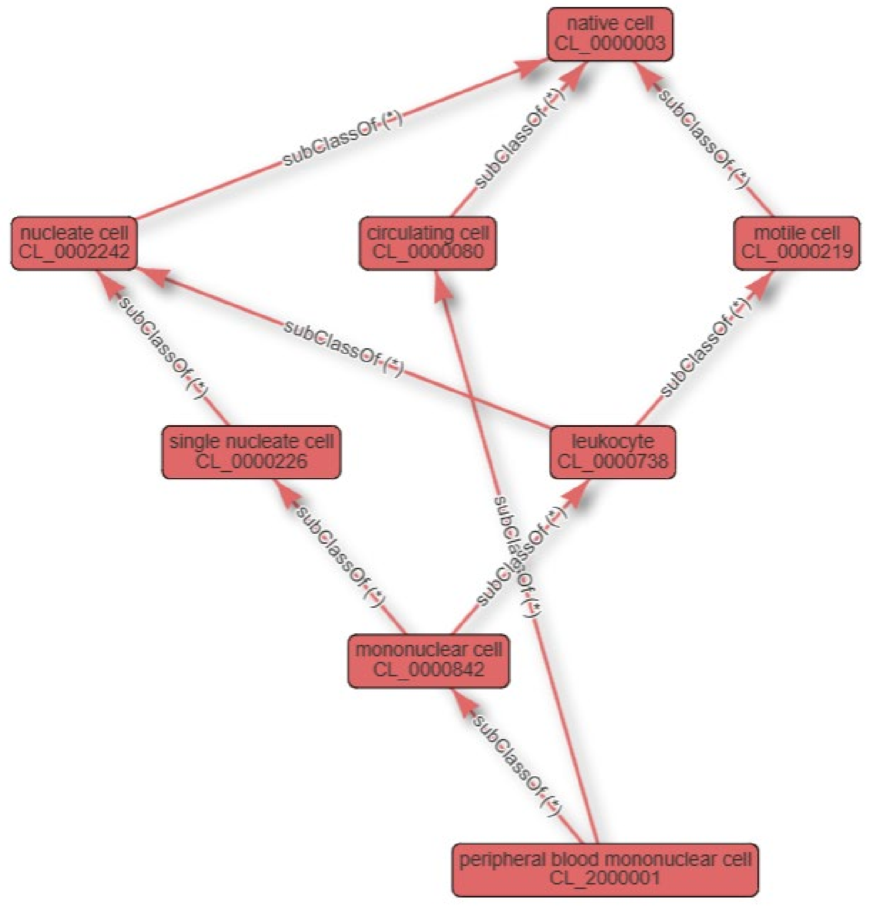
Relations to standardized terms

Relationship	Term ID	Term Name
Consists_of	CL:0000842	Mononuclear cell
Consists_of	CL:2000001	Peripheral blood mononuclear cell
Has_phenotype	DOID:11729	Lyme disease
Has_phenotype	EFO:0000408	Disease
Underwent	EFO:000727	Treatment

Relationship	Term ID	Term Name
Consists_of	CL:0000842	Mononuclear cell
Consists_of	CL:2000001	Peripheral blood mononuclear cell
Consists_of	EFO:0000322	Cell line
Part_of	UBERON:00001 87	Blood

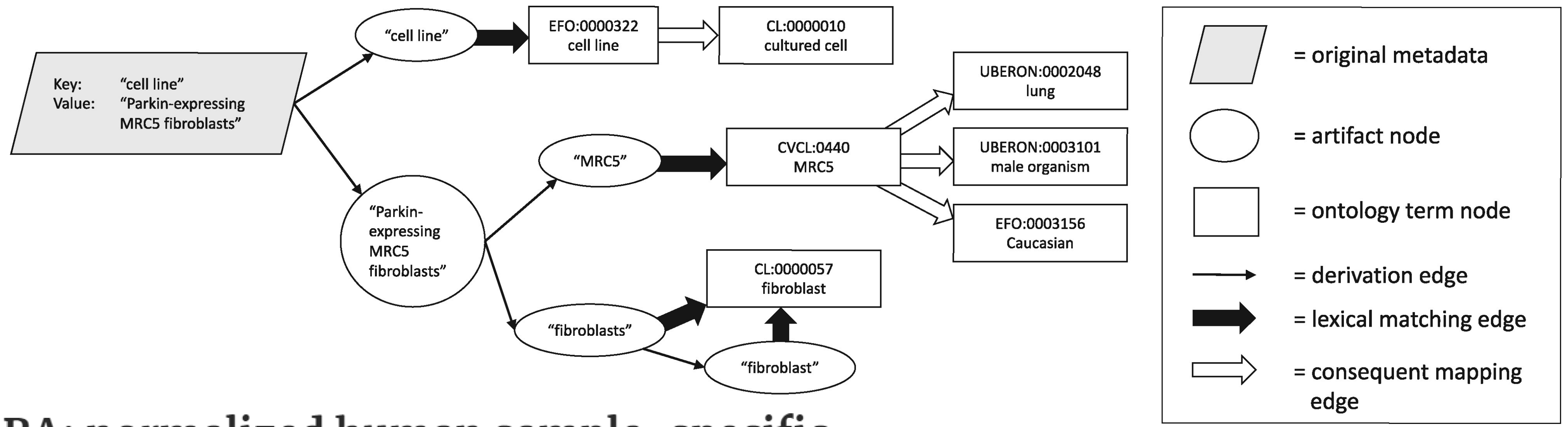
Ontologies provide a framework for standardization

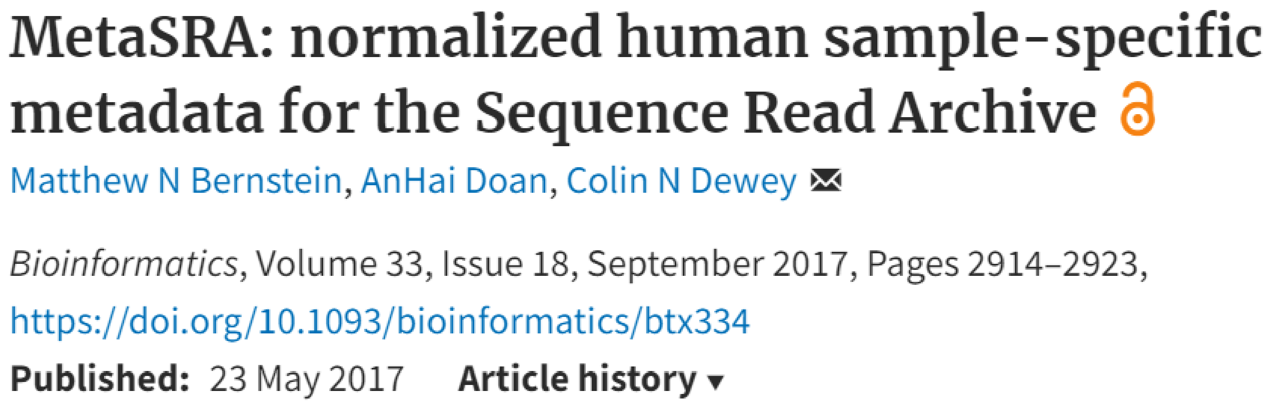
- An ontology formally represents entities, categories, and properties in a field of knowledge
- The terms in an ontology are uniquely identified
- Ontology relations relate specific terms to general terms that subsume them


native cell CL_0000003
nucleate cell CL_0002242
circulating cell CL_0000080
motile cell CL_0000219
single nucleate cell CL_0000226
leukocyte CL_0000738
mononuclear cell CL_0000842
peripheral blood mononuclear cell CL_2000001
subClassOf (*)

www.ebi.ac.uk/ols4

MetaSRA is a rule-based system for metadata standardization





MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive

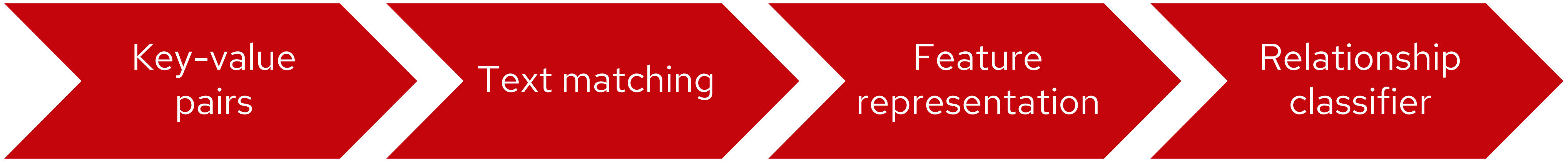
Matthew N Bernstein, AnHai Doan, Colin N Dewey

Bioinformatics, Volume 33, Issue 18, September 2017, Pages 2914–2923,

https://doi.org/10.1093/bioinformatics/btx334

Published: 23 May 2017 **Article history** ▾

ML4Meta: a machine learning approach



Text matching

- Based on MetaSRA
- Rule-based but domain-agnostic
- Uses ontologies and lexicons
- Produces a candidate term match and its derivation

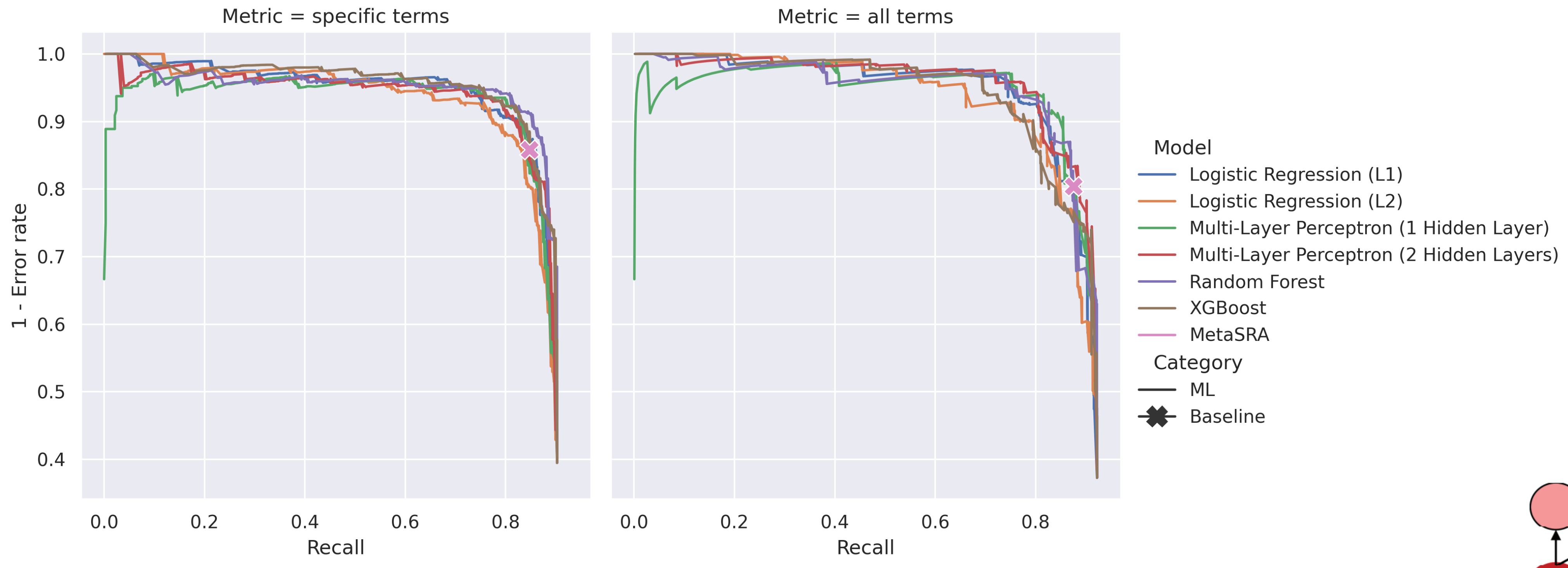
Feature representation

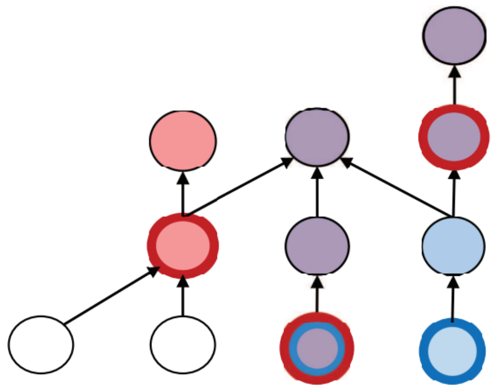
- Text features (bag-of-words)
- Derivation step features
- Ontology relation features
- Match goodness (edit distance and synonym level)

Relationship classifier

- Assigns probabilities representing whether a term relates to a sample and how

ML4Meta outperforms MetaSRA on SRA metadata

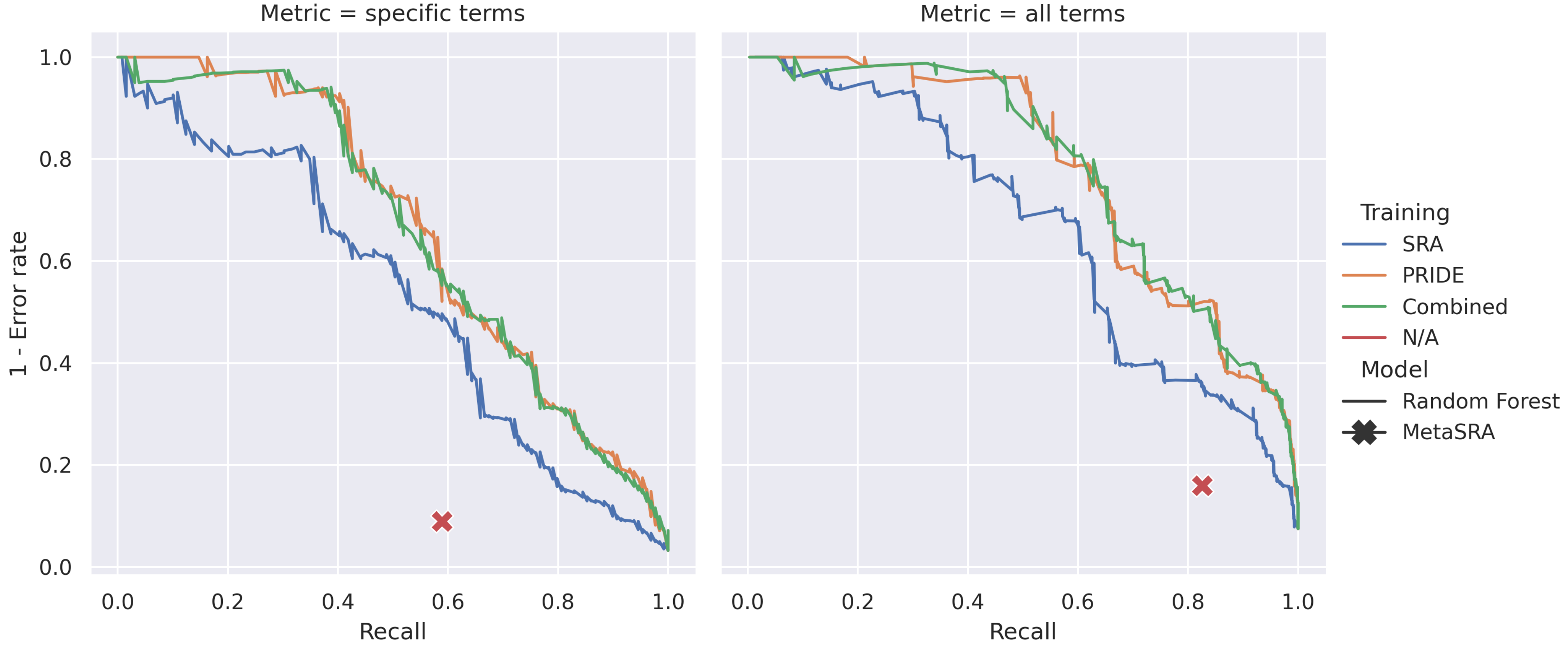




ML4Meta generalizes to other datasets

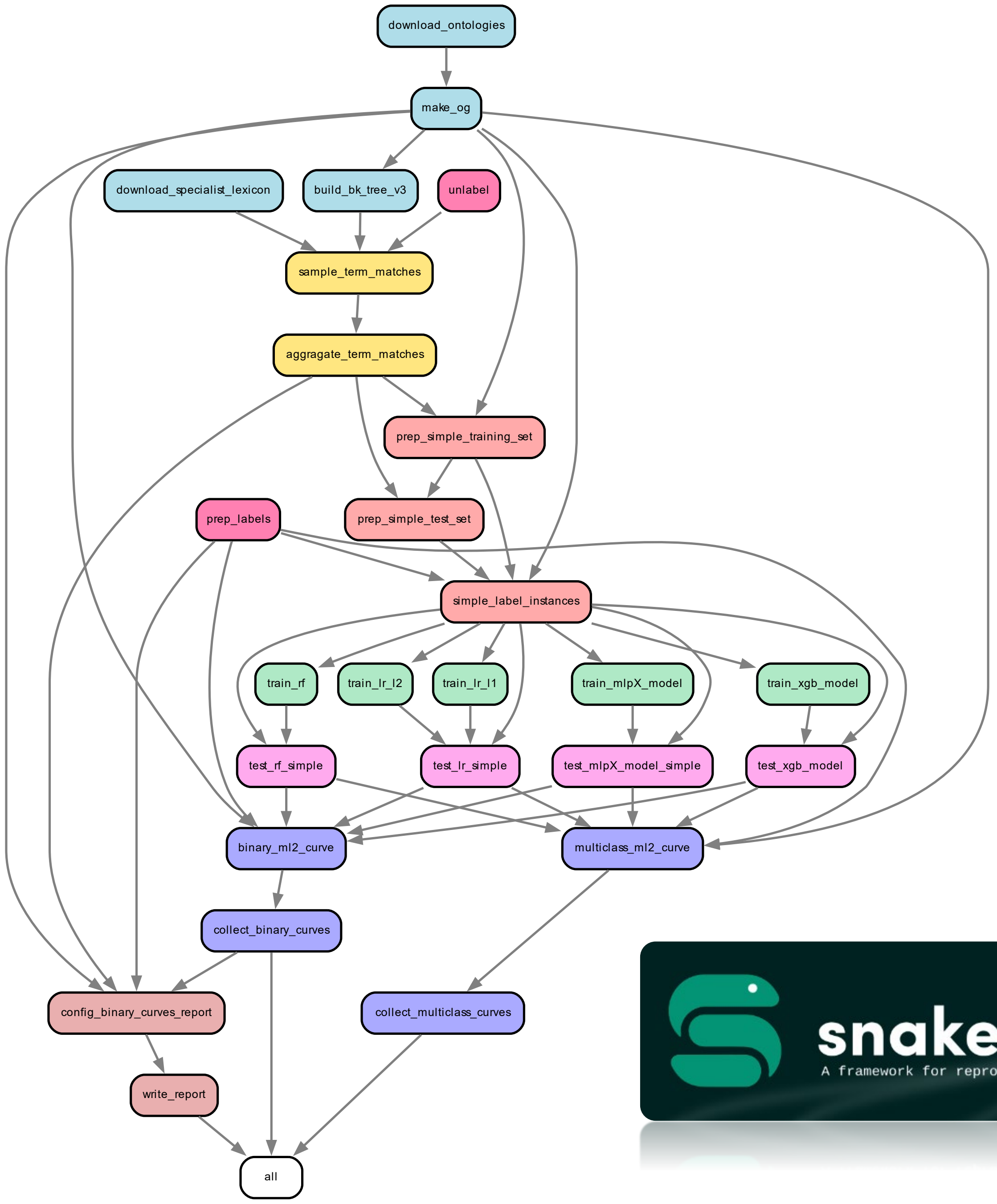
PRIDE

PRoteomics IDEntifications Database



Development Workflow

- Resource setup
- Data and labels
- Text matching
- Feature generation
- Model training
- Model testing
- Evaluation
- Report generation



snakemake
A framework for reproducible data analysis

YAML configuration

```yaml
experiments:
  workspace/2023-05-11_sra-to-sra-with-chebi-experiment:
    run: true
    results:
      - binary curves
      - multiclass curves report
    models:
      lr-l1: Logistic Regression (L1)
      lr-l2: Logistic Regression (L2)
      mlp1: Multi Layer Perceptron (1 hidden layer)
      mlp2: Multi Layer Perceptron (2 hidden layers)
      rf: Random Forest
      xgb: XGBoosted Trees
    evaluation_ontology: workspace/ontology_graphs/v11/ontology_graph.pickle
    training: workspace/sra-training-set_2023-05-11
    testing: workspace/sra-to-sra-testing-set_2023-05-11
```

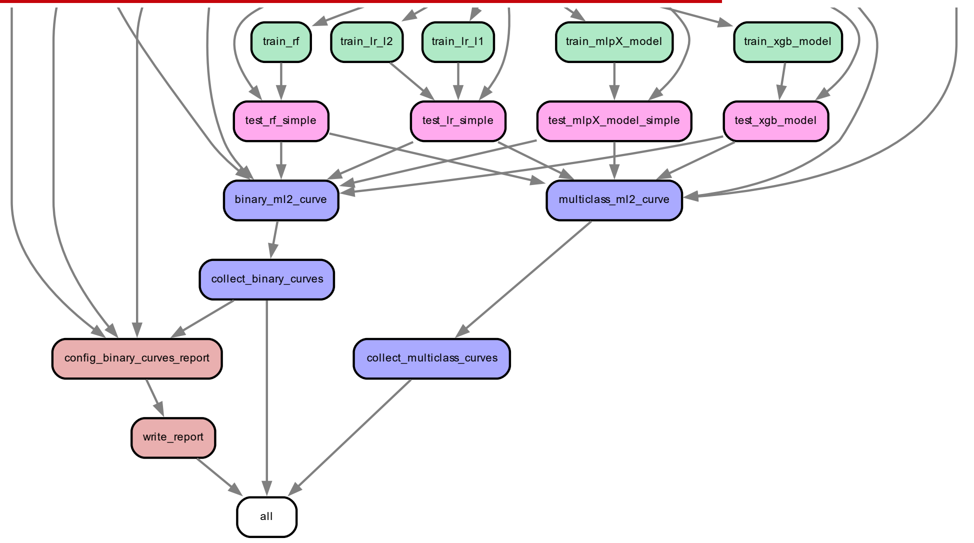
Snakemake rule definitions

```python
rule collect_binary_curves:
    output:
        result = '{experiment_dir}/binary_curves.csv'
    input:
        input_to_collect_binary_curves
    run:
        import pandas as pd
        pd.concat([
            pd.read_csv(f, index_col=['model', 'threshold'])
            for f in input
        ]).to_csv(output.result)
```

Snakemake input functions

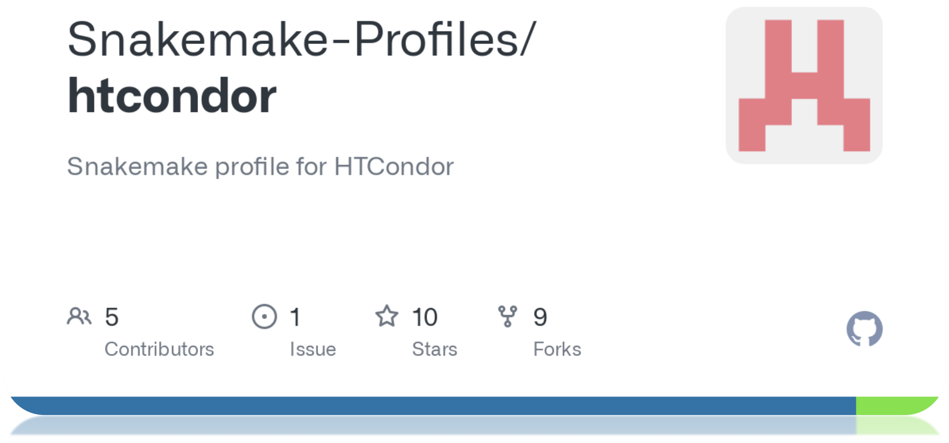
```python
def input_to_collect_binary_curves(wildcards):
    return expand(
        f'{wildcards.experiment_dir}/binary_curves/{{model}}.csv',
        model = config['experiments'][wildcards.experiment_dir]['models'].keys()
    )
```

Workflow execution



The HTCondor profile brings the power of HTCondor to Snakemake

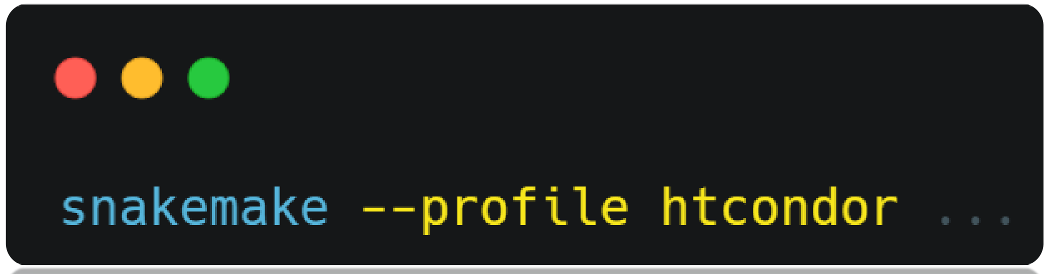
- Simple configuration and usage
- Implemented using the HTCondor Python API
- Each Snakemake rule becomes a Condor job



Snakemake-Profiles/
htcondor

Snakemake profile for HTCondor

5 Contributors · 1 Issue · 10 Stars · 9 Forks



```
snakemake --profile htcondor ...
```

ML development with Snakemake and HTCondor: Strengths and weaknesses

- ▲ Snakemake is backend-agnostic and portable
- ▲ Avoids re-generating dependencies in subsequent executions
- ▲ Python syntax
- ▲ Input functions are powerful
- ▲ Data-dependent workflow definition
- ▲ Defining workflows with configuration files makes iterative development easy

- ▼ Snakemake is file-based
- ▼ The Snakemake HTCondor profile is harder to configure without a network share
- ▼ Snakemake process needs to stay running on submit node throughout workflow execution

THANK YOU!