Offene und Reproduzierbare Datenanalyse mit OpenML

Bernd Bischl and Giuseppe Casalicchio

Sommerklausur 2015, Holzhausen

Section 1

What is OpenML?

What is OpenML?

- Main idea: Make ML experiments reproducible and most parts computer-readable
- Share everything
- Enrich with meta-information
- Later: Mine the results, meta-learn on it

1 minute intro

Data from
various sources
analysed and
organised online
for easy access

Scientists can **broadcast data,** explaining the challenge that needs to be addressed. OpenML will (for known data formats) **automatically analyze the data**, compute data characteristics, **annotate and index it for easy search**

1 minute intro

Scientific tasks that can be **interpreted by tools**, and **solved collaboratively**

Tasks are **realtime (collaborative) data mining challenges**, allowing anyone to build on previous results. OpenML creates **machine-readable descriptions** so that tools can **automatically download data**, use the correct procedures, and **upload all results online**.

1 minute intro

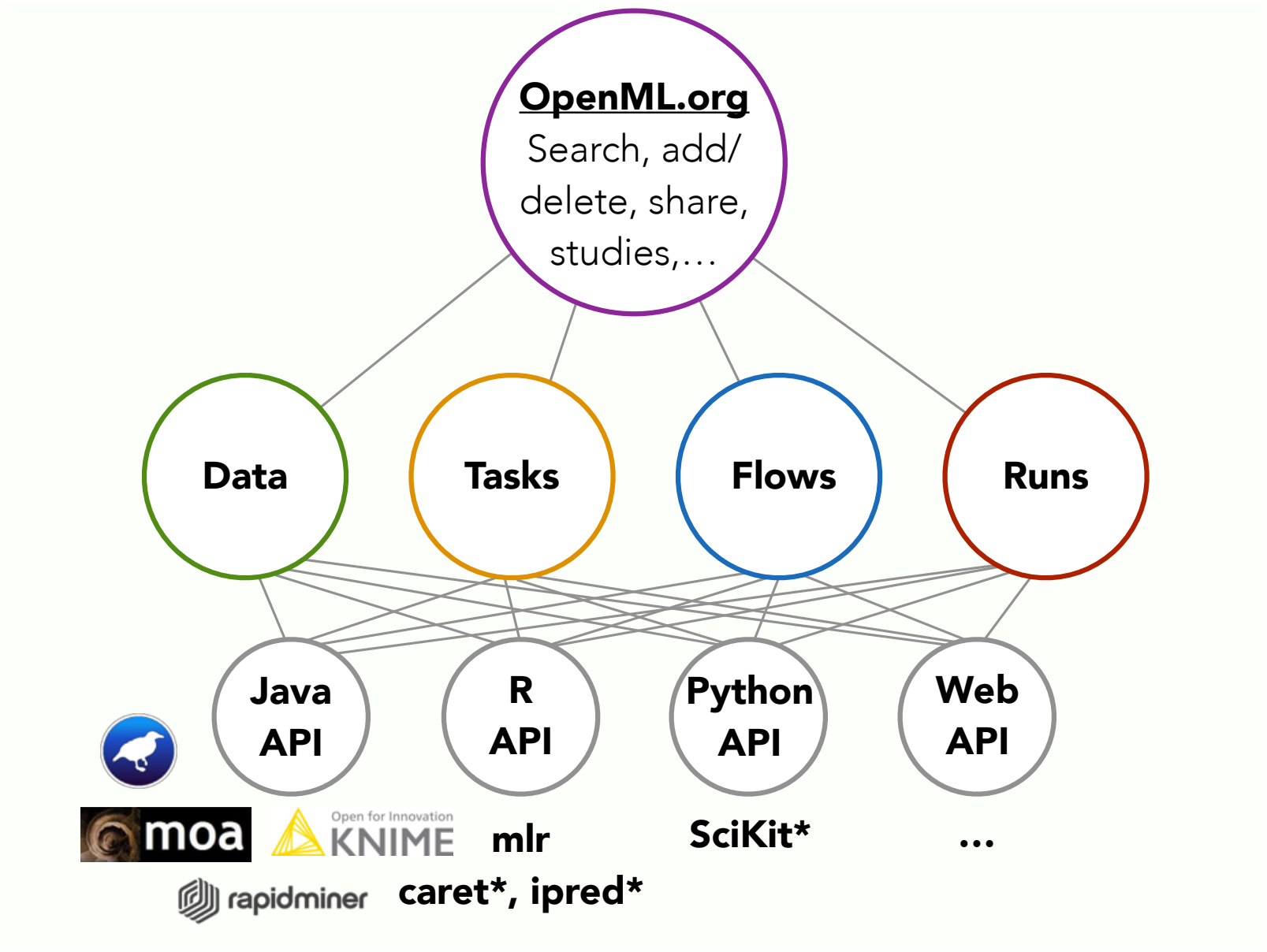
Tool plugins
for automated
data download,
workflow upload and
experiment logging
and sharing

Flows are implementations of algorithms, workflows, or scripts **solving OpenML tasks**. OpenML keeps track of **flow details and versioning**, **organizes all their results** for easy comparison, even across tools.

1 minute intro

Experiments auto-uploaded, linked to **data, flows** and **authors**, and organised for easy reuse

Runs contain the results that **flows** obtained on specific tasks. Runs are **fully reproducible**, linked to the underlying data, tasks, flows and authors. OpenML **organizes all results online** for **discovery, comparison and reuse**


OpenML.org
Search, add/ delete, share, studies,…
Data
Tasks
Flows
Runs
Java API
R API
Python API
Web API
moa
Open for Innovation
KNIME
mlr
SciKit*
…
rapidminer
caret*, ipred*

Section 2

OpenML with mlr

Motivation

The good news

- CRAN serves hundreds of packages for machine learning (cf. CRAN task view machine learning)
- Many packages are compliant to the unwritten interface definition:

```
> model = fit(target ~ ., data = train.data, ...)
> predictions = predict(model, newdata = test.data, ...)
```

Motivation

The bad news

- Some packages do not support the formula interface or their API is "just different"
- Functionality is always package or model-dependent, even though the procedure might be general
- No meta-information available or buried in docs (sometimes not documented at all)
- Many packages require the user to "guess" good hyperparameters
- Larger experiments lead to lengthy, tedious and error-prone code

Our goal: A domain-specific language for many machine learning concepts!

Motivation: mlr

https://github.com/berndbischl/mlr

- Unified interface for the basic building blocks: tasks, learners, resampling, hyperparameters, ...
- Reflections: nearly all objects are queryable (i.e. you can ask them for their properties and program on them)
- Possibility to fit, predict, evaluate and resample models
- Easy extension mechanism through S3 inheritance
- Abstract description of learners and tasks by properties
- Different visualizations for e.g. ROC curves and predictions
- Benchmarking of learners for muliple data sets
- Variable selection with filters and wrappers
- Parallelization is built-in
- ...

Section 3

OpenML Website

OpenML Website

Let's visit the website

Section 4

OpenML R-Package

OpenML R-Package

`https://github.com/openml/r`

Current API in R

- Explore data and tasks
- Download data and tasks
- Register learners
- Upload runs
- Explore your own and other people's results

Already nicely connected to `mlr`!

OpenML R-Package

Let's have a look at the R-Package

Thanks!

References I

Van Rijn, J. N., Bischl, B., Torgo, L., Gao, B., Umaashankar, V., Fischer, S., Winter, P., Wiswedel, B., Berthold, M. R., and Vanschoren, J. (2013).
Openml: A collaborative science platform.
In *Machine learning and knowledge discovery in databases*, pages 645–649. Springer.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2014).
Openml: networked science in machine learning.
ACM SIGKDD Explorations Newsletter, 15(2):49–60.