Towards Networked, Automated Machine Learning

OpenML

Joaquin Vanschoren (TU/e) 2015
GALILEO GALILEI
DISCOVERS SATURN’S RINGS
1610

‘SMAISMIRMILMEPOETAL
EUMIBUNENUUGTTAUIRAS’
Research different.

Royal society: Take nobody’s word for it
Open sharing of findings and methods: scientific journal
Reputation-based economy
After 300 years, is the printing press still the best medium?

For Machine Learning?

- Code, data too complex (published separately)
- Experiment details scant
- Results unactionable, hard to reproduce, reuse
- Papers not updatable
- Slow, limited impact tracking
- Publication bias
Gaps in data-driven science

**Domain scientists:** doubts about latest/best techniques, run simple models on complex data, small (biased) collaborations

**Data scientists:** don’t speak the language, don’t know how to access scientific databases, run complex models on simple data

**Industry:** small companies delimitated by lack of access to data/expertise
LIES, DAMNED LIES, AND MEDICAL SCIENCE

By DAVID H. FREEDMAN

In 2001, rumors were circulating in Greek hospitals that surgery residents, eager to rack up scalpels time, were falsely diagnosing hapless Albanian immigrants with appendicitis. At the University of Ioannina medical school’s teaching hospital, a newly minted doctor named Athina Tatsioni was discussing the rumors with colleagues when a professor who had overheard asked her if she’d like to try to prove whether they were true—he seemed to be almost daring her. She accepted the challenge and, with the professor’s and other colleagues’ help, eventually produced a formal study showing that, for whatever reason, the appendices removed from patients with Albanian names in six Greek hospitals were more than three times as likely to be perfectly healthy as those removed from patients with Greek names. “It was hard to find a journal willing to publish it, but we did,” recalls Tatsioni. “I also discovered that I really liked research.” Good thing, because the study had actually been a sort of audition. The professor, it turned out, had been putting together a team of exceptionally brash and curious young clinicians and Ph.D.s to join him in tackling an unusual and controversial agenda.

Last spring, I sat in on one of the team’s weekly meetings on the medical school’s campus, which was planked crustily across a series of sharp hills. The building in which we met, like most at the school, had the look of a barracks and was festooned with political graffiti. But the group convened in a spacious conference room that would have been as home at a Silicon Valley start-up. Sprawled around a large table were Tatsioni and eight other youngish Greek researchers and physicians who, in contrast to the portly younger staff frequently seen in U.S. hospitals, looked like the casually glamorous cast of a television medical drama. The professor,
85% medical research resources are wasted: associations/effect are false, exaggerated, translation into applications is inefficient

Research findings less likely to be true if:
- Studies are smaller, few collaborators
- Effect sizes are smaller
- Flexibility in designs, definitions, outcomes, analytical modes
- Teams are chasing statistical significance

Increase credibility:
- Large-scale interdisciplinary collaboration
- Replication culture, reproducibility
- Registration/sharing of data
- Better statistical methods, models
- Better study design, training
LABELLING (CATALOGUING) OF EVENTS STILL DONE MANUALLY
Research different.

Polymaths: Solve math problems through massive collaboration (not competition)

Broadcast question, combine many minds to solve it

Solved hard problems in weeks

Many (joint) publications
Research different.

SDSS: Robotic telescope, data publicly online (SkyServer)
Broadcast data, allow many minds to ask the right questions
Thousands of papers, citations
+1 million distinct users vs. 10,000 astronomers

Next: Synoptic Telescope
Research different.

How do you label a million galaxies?
Offer right tools so that anybody can contribute, instantly
Many novel discoveries by scientists and citizens

More? Read ‘Networked science’ by M. Nielsen
Why? Designed serendipity

Broadcasting data fosters spontaneous, unexpected discoveries

What’s hard for one scientist is easy for another: connect minds

How? Remove friction

Organized body of compatible scientific data (and tools) online

Micro-contributions: seconds, not days

Easy, organised communication

Track who did what, give credit
FRICTION-LESS ENVIRONMENT FOR MACHINE LEARNING RESEARCH

**Organized**: Experiments connected to data, code, people. Reproducible.

**Easy to use**: Automated download/upload within your ML environment

**Micro-contributions**: Upload single dataset, algorithm, experiment

**Easy communication**: Online discussions per dataset, algorithm, experiment

**Reputation**: Auto-tracking of downloads, reuse, likes.

**Real time**: Share and reuse instantly, openly or in circles of trusted people
Exploring machine learning better, together

1345 data sets
Find or add **data** to analyse

4830 tasks
Download or create scientific **tasks**

1400 flows
Find or add data analysis **flows**

452379 runs
Upload and explore all **results** online.
Data from various sources analysed and organised online for easy access

Scientists broadcast data by uploading or linking from existing repos. OpenML will automatically check and analyze the data, compute characteristics, annotate, version and index it for easy search.
Search by keywords or properties

Through website or API

Filters

Tagging

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>iris (1)</td>
<td>This is perhaps the best known dataset.</td>
</tr>
<tr>
<td>credit-a (1)</td>
<td>1. Title: Credit Approval 2. Source: Bank of America data.</td>
</tr>
<tr>
<td>anneal.ORIG (1)</td>
<td>1. Title of Database: Annealing.</td>
</tr>
<tr>
<td>diabetes (1)</td>
<td>1. Title: Pima Indians Diabetes.</td>
</tr>
<tr>
<td>colic (1)</td>
<td>Donor: Will Taylor (<a href="mailto:taylor@pluto.com">taylor@pluto.com</a>).</td>
</tr>
<tr>
<td>anneal (2)</td>
<td>This is a preprocessed version</td>
</tr>
<tr>
<td>mfeat-zernike (1)</td>
<td>The multi-feature digit dataset -</td>
</tr>
<tr>
<td>mfeat-morphological (1)</td>
<td>The multi-feature digit dataset -</td>
</tr>
<tr>
<td>solar-flare (2)</td>
<td>1. Title: Solar Flare database</td>
</tr>
</tbody>
</table>
• Search on keywords or properties
• Wiki-like descriptions
• Analysis and visualisation of features
• Auto-calculation of large range of meta-features
Scientific tasks that can be interpreted by tools and solved collaboratively.

**Tasks:** containers with all data, goals, procedures. **Machine-readable:** tools can automatically download data, use correct procedures, and upload results. Creates realtime, collaborative data mining challenges.
• **Example:** Classification on click prediction dataset, using 10-fold CV and AUC

• People submit results (e.g. predictions)

• Server-side evaluation (many measures)

• All results organized online, per algorithm, parameter setting

• Online visualizations: every dot is a run plotted by score
• Leaderboards visualize progress over time: who delivered breakthroughs when, who built on top of previous solutions
• Collaborative: all code and data available, learn from others, form teams
• Real-time: who submits first gets credit, others can improve immediately
Machine learning flows (code) that can solve tasks and report results.

**Flows**: wrappers that read **tasks**, return required **results**. Scientists upload code or link from existing repositories/libraries. Tool integrations allow automated **data download**, **flow upload** and **experiment logging and sharing**.
REST API + Java, R, Python APIs

- WEKA/MOA plugins: automatically load tasks, export results

- RapidMiner plugin: new operators to load tasks, export results and subworkflow

- R/Python interfaces: functions to down/upload data, code, results in few lines of code

```python
global import APICluster
from sklearn import preprocessing, ensemble
classifier = APICluster(username=username, password=password)
dataset = classifier.download_dataset(31)
X, y, categorical = dataset.get_pandas()
enc = preprocessing.OneHotEncoder(categorical_features=categorical)
X = enc.transform(X).todense()
clf = ensemble.RandomForestClassifier()
clf.fit(X, y)
```
• All results obtained with same flow organised online
• Results linked to data sets, parameter settings -> trends/comparisons
• Visualisations (dots are models, ranked by score, colored by parameters)
Experiments auto-uploaded, linked to data, flows and authors, and organised for easy reuse.

**Runs** uploaded by **flows**, contain fully reproducible results for all **tasks**. OpenML evaluates and organizes all results online for discovery, comparison and reuse.
• Detailed run info
• Author, data, flow, parameter settings, result files, ...
• Evaluation details (e.g., results per sample)
OpenML Community

Used all over the world (and still in beta)
Great open source community, on GitHub
450+ active users, many more passive ones
1000s of datasets, flows, 450000+ runs
Things we’re working on

Circles
Create collaborations with trusted researchers
Share results within team prior to publication

Projects (e-papers)
- Online counterpart of a paper, linkable
- Merge data, code, experiments (new or old)
- Public or shared within circle

Altmetrics
- Measure real impact of your work
- Reuse, downloads, likes of data, code, projects,…
- Online reputation (more sharing)
**Distributed computing**
- Create jobs online, run anywhere you want
- Locally, clusters, clouds

**Algorithm selection, hyperparameter tuning**
- Upload dataset, system recommends techniques
- Model-based optimisation techniques
- Continuous improvement (learns from past)
Things we’re working on (please join)

**Data repository connections**
- Wonderful open data repo’s (e.g. rOpenSci)
- More data formats, data set analysis

**Algorithm/code connections**
- Improved API’s (R,Java,Python,CLI,...)
- Your favourite tool integrated

**Statistical analysis**
- Proper significance testing in comparisons
- Recommend evaluation techniques (e.g. CV)

**Online task creation**
- Definition of scientific tasks
- Freeform tasks or server-side support
Towards OpenML in education
Towards OpenML in education

Het bewijs dat ik studeer op zondag!
"@joavanschoren: #Machinelearning students on a #collaborative data mining "

Lauradorp, Landgraaf
Collaboratory: bring data scientists and domain scientists together online (and their data and tools)

Easy, large-scale collab: Extract actionable datasets, key tools. Scientists shares data and get help, DS can test technique on many current datasets.

Real-time collab: share experiments automatically, discuss online. Automate experimentation.
Bridging gaps in data-driven science

Collaboratory: bring data scientists and domain scientists together online (and their data and tools)

Easy, large-scale collab: Extract actionable datasets, key tools. Scientists share data and get help, DS can test technique on many current datasets.

Real-time collab: Share experiments automatically, discuss online. Automate experimentation.
Networked data(-driven) science

Data repositories

OpenML
meta-data, models, evaluations

Code repositories

API

data, code, experiment sharing, commenting

Human scientists
Automating machine learning

OpenML

meta-data, models, evaluations

API

Data cleaning
Algorithm Selection
Parameter Optimization
Workflow construction
Post processing

Human scientists
Automated processes
SEMI-AUTOMATED MACHINE LEARNING

Data Collection
Data Cleaning
Data Coding
Metric Selection
Algorithm Selection
Parameter Optimization
Post-processing
Online evaluation

Thanks to R. Caruana
AUTOMATED MACHINE LEARNING

Data Collection
Data Cleaning
Data Coding
Metric Selection
Algorithm Selection
Parameter Optimization
Post-processing
Online evaluation

Thanks to R. Caruana
AUTOMATED MACHINE LEARNING

Data Collection
Data Cleaning
Data Coding
Metric Selection
Algorithm Selection
Parameter Optimization
Post-processing
Online evaluation

Semi-Automated Machine Learning

Thanks to R. Caruana
MACHINE LEARNING PIPELINE

Data Collection  Data Cleaning  Data Coding  Metric Selection  Algorithm Selection  Parameter Optimization  Post-processing  Online evaluation

Manual heuristic search: surprisingly suboptimal
Grid search: only effective with very small number of parameters
Random search: better with larger number of parameters
Bayesian Optimization: better with very large number of parameters

Critical with modern (low bias) algorithms: boosting, deep learning,…

Thanks to R. Caruana
Bayesian Optimization

1. Initial sample

Thanks to Matthew Hoffmann
Bayesian Optimization

1. Initial sample

Thanks to Matthew Hoffmann
Bayesian Optimization

1. Initial sample
2. Posterior model

Thanks to Matthew Hoffmann
Bayesian Optimization

1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)

Thanks to Matthew Hoffmann
Bayesian Optimization

1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)
4. Optimize it

Thanks to Matthew Hoffmann
Bayesian Optimization

1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)
4. Optimize it
5. Sample new data, update model

Thanks to Matthew Hoffmann
Bayesian Optimization

1. Initial sample
2. Posterior model
3. Exploration strategy (acquisition function)
4. Optimize it
5. Sample new data, update model
6. Repeat

Thanks to Matthew Hoffmann
MACHINE LEARNING PIPELINE

Data Collection  Data Cleaning  Data Coding  Metric Selection  Algorithm Selection  Parameter Optimization  Post-processing  Online evaluation

Meta-learning: Learn link between data properties & algorithm performance
Active testing: Sequentially predict algorithm, learn from evaluation score
Recommender systems: Evaluations are ‘ratings’ of datasets by algorithms

Data properties are crucial, depend on data domain (streams, graphs,…) Best combined with parameter optimisation
OpenML in drug discovery

ChEMBL database

Molecular properties (e.g. MW, LogP)

Fingerprints (e.g. FP2, FP3, FP4, MACSS)

<table>
<thead>
<tr>
<th>MW</th>
<th>LogP</th>
<th>TPSA</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
<th>b5</th>
<th>b6</th>
<th>b7</th>
<th>b8</th>
<th>b9</th>
</tr>
</thead>
<tbody>
<tr>
<td>377.435</td>
<td>3.883</td>
<td>77.85</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>341.361</td>
<td>3.411</td>
<td>74.73</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>197.188</td>
<td>-2.089</td>
<td>103.78</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>346.813</td>
<td>4.705</td>
<td>50.70</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

10.000+ regression datasets
OpenML in drug discovery

Predict best algorithm with meta-models

I. Olier et al. MetaSel@ECMLPKDD 2015
Fast algorithm selection

By evaluating algorithm on smaller data samples, multi-objective evaluation
MACHINE LEARNING PIPELINE

Data Collection
Data Cleaning
Data Coding
Metric Selection
Algorithm Selection
Parameter Optimization
Post-processing
Online evaluation

Detect variable types, missing values, anomalies, …
Auto-coding: Different learning algorithms require different coding
Detect changes in data (that affect model): sensors break, human error,…
Feedback loop: implementing a model changes the data it was trained on
Leakage: model is trained on data it should not see

Thanks to R. Caruana
Towards an AI for data-driven research

**Symbiotic AI:** learns from thousands of human data scientists how to analyse data. Learns which workflows work well on which data.

**Collaborates with scientists:** take care of tedious, error-prone tasks: clean data, find similar data, try state-of-the-art, algorithm selection, configuration,…

**Couple human expertise and machine learning:** offer the right information at the right time to the right person, so that she can make informed decisions based on clear evidence.
Towards an AI for data-driven research

**AI** (novice) - what?
Upload dataset, give 1 day to return best possible model

**Man-Computer Symbiosis** - why?

**Expert level** (expert) - how?
Expert asks system how it did something, system explains all details, show code, data
Look at my new data set :)  
I analysed it, here’s a full report. 
Can you remove the outliers? 
I removed them, does this look ok? 
Yes, now I want to predict X 
OK, let me build you some optimized classification models. 
Are there similar data sets? 
Here’s a ranked list, and some papers. 
What’s the state-of-the-art? 
I’m running all state-of-the-art techniques, 
I’ll send a report soon. 
What’s the difference with algo A and B? 
Here’s a comparison. B looks most promising on high-dimensional data.
THANK YOU

Nenad Tomašev
Luis Torgo
Jan van Rijn
Giuseppe Casalicchio
Joaquin Vanschoren
Michel Lang
Bernd Bischl
Matthias Feurer
You?