

Datalab:

Interoperable data management for fundamental battery research

Matthew Evans

*BEWARE Research Fellow
MODL-IMCN, UCLouvain & Matgenix*
<https://ml-evs.science>

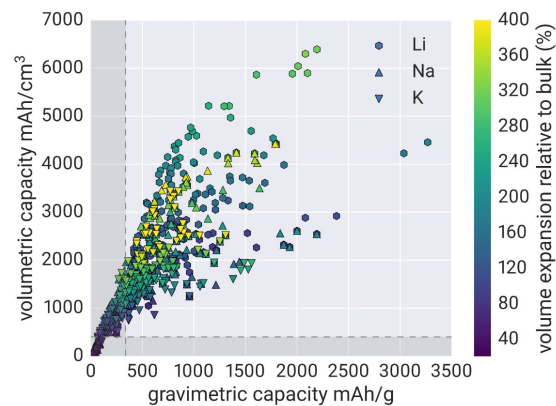
RSC 1st Annual Symposium for Advanced Battery Materials, 8th November 2023



High-throughput crystal structure prediction for beyond-Li anodes

PhD with (now) Prof Andrew Morris (University of Cambridge/Birmingham)

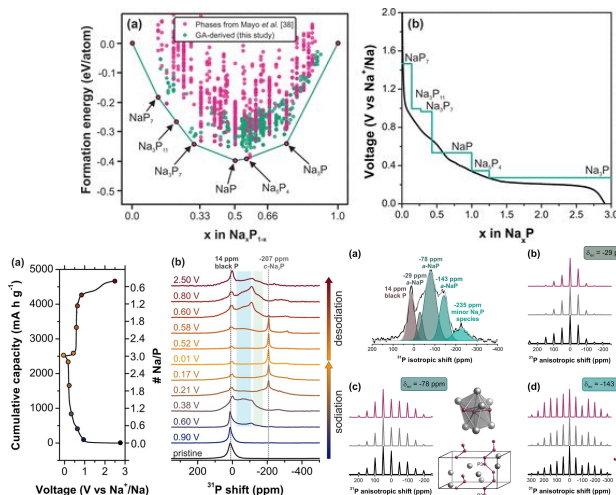
Conversion anode materials



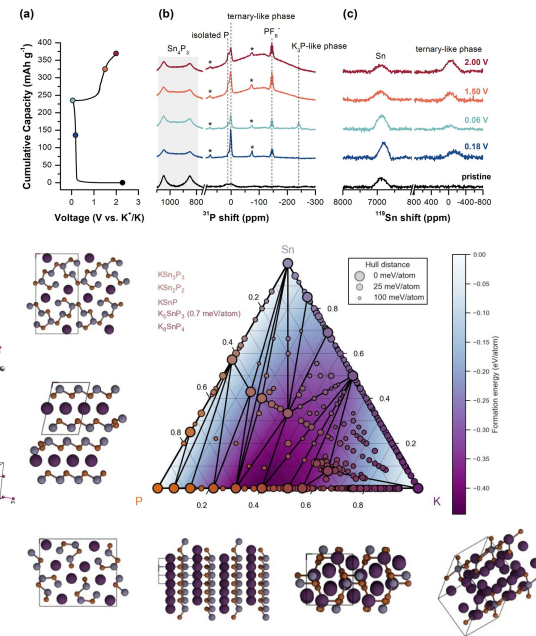
<https://github.com/ml-evs/matador>

Evans & Morris,
JOSS 2020
[10.21105/joss.02563](https://doi.org/10.21105/joss.02563)

Na-P



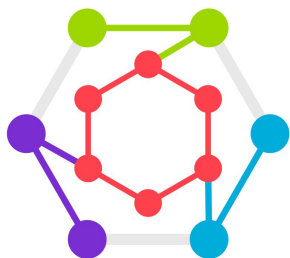
K-Sn-P



Marbella, Evans, et al,
JACS 2018
[10.1021/jacs.8b04183](https://doi.org/10.1021/jacs.8b04183)

Ells, Evans et al,
Chem. Mater. 2022
[10.1021/acs.chemmater.2c01570](https://doi.org/10.1021/acs.chemmater.2c01570)

Machine-learning accelerated materials d(iscovery/esign) - MLxMD

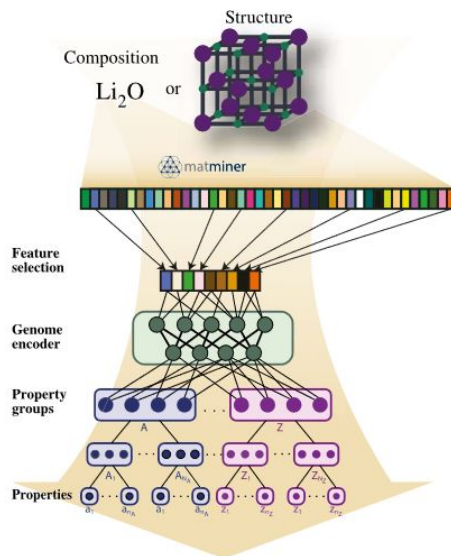


OPTIMADE

Open Databases Integration
for Materials Design

Andersen, C.W., et al.
Sci Data **8**, 217 (2021).

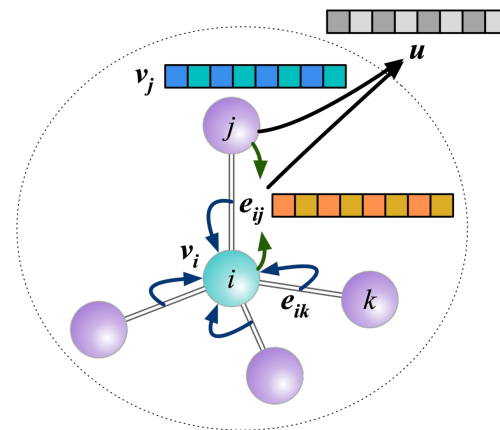
<https://optimade.org>



Pierre-Paul De Breuck et al
J. Phys.: Condens. Matter **33** 404002 (2021)



<https://github.com/modl-uclouvain/modnet>

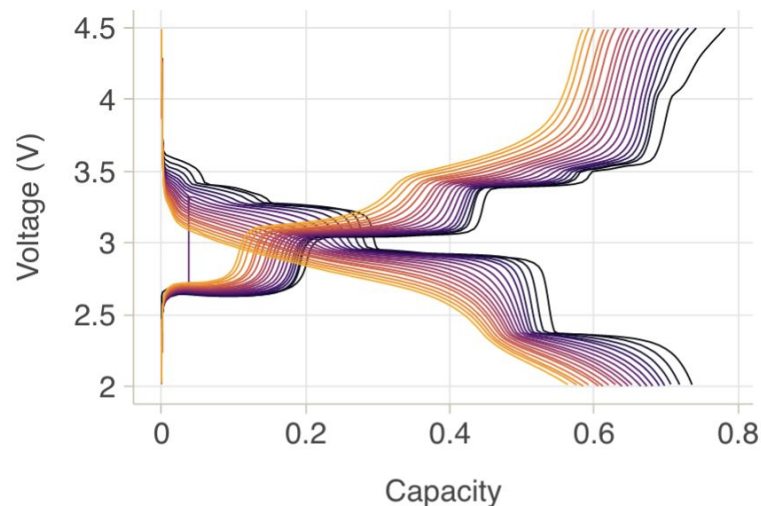


Reproduced from
<https://github.com/materialsvirtuallab/mat4l>

Software & workflow development for ML-accelerated materials discovery + design

Managing experimental lab data

"Data":



Typical battery cycling data

~700 GB stored in backup servers of the Grey group, Cambridge

What information would need to be recorded to make this information useful after the student leaves?

What information would need to be recorded to make this useful to a machine learning algorithm?

What type of cell is this? What are the electrodes? What batch are they from? Active mass? Where was it cycled? What instrument? Cycling parameters? Surface area? When was it made? Who made it? When? What batch of electrolyte? Was it made in a glovebox? Which one? Have other measurements been made on this cell? Has this result been repeated with other cells? Is there any characterization on the electrodes? On the active material? On the electrolyte? What was the temperature of the room? Were there any spikes? What does the dV/dQ look like? Who "owns" this data?

Motivation

- You are sitting in a conference talk and see some results. You seem to remember that an undergraduate in your lab made that compound 3 summers ago and took some measurements. Could you pull up their data to compare? How about share with the speaker?
- You suspect humidity may be affecting your synthesis. Of all the oxides prepared in your group over the past 10 years, do the ones with lower humidity show cleaner diffraction patterns?
- A new machine learning model is reported in the literature that allows for prediction of new multiferroics based on raw diffraction patterns. How quickly could you apply this model to all the diffraction patterns taken in your lab over the last 10 years?

Motivation

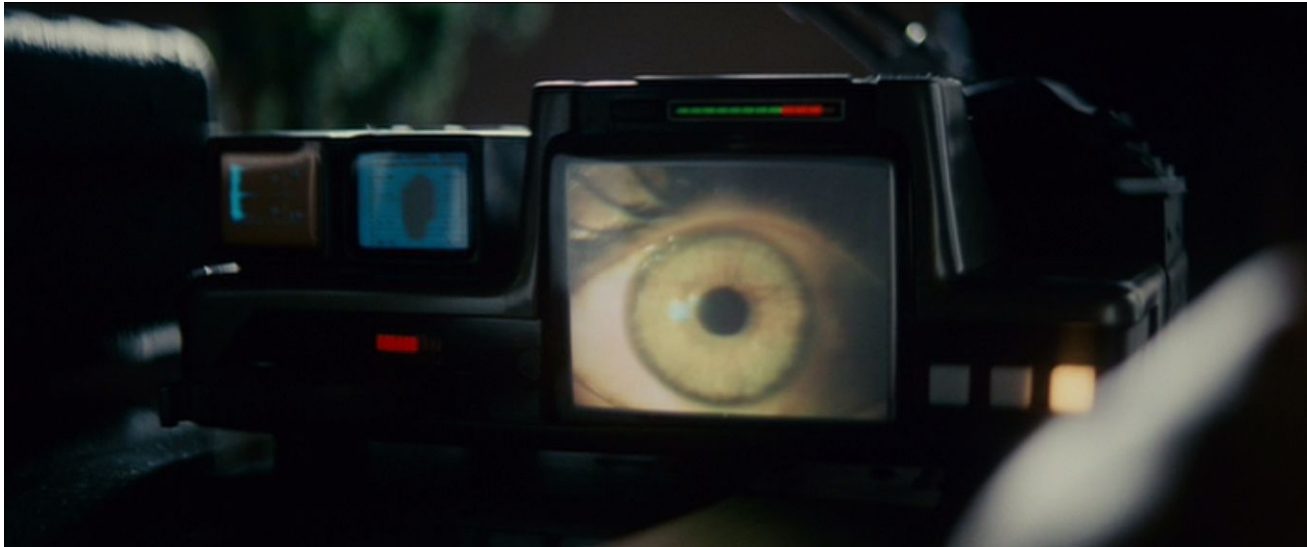
- You are sitting in a conference talk and see some results. You seem to remember that an undergraduate in your lab made that compound 3 summers ago and took some measurements. Could you pull up their data to compare? How about share with the speaker?
- You suspect humidity may be affecting your synthesis. Of all the oxides prepared in your group over the past 10 years, do the ones with lower humidity show cleaner diffraction patterns?
- A new machine learning model is reported in the literature that allows for prediction of new multiferroics based on raw diffraction patterns. How quickly could you apply this model to all the diffraction patterns taken in your lab over the last 10 years?

Motivation

- You are sitting in a conference talk and see some results. You seem to remember that an undergraduate in your lab made that compound 3 summers ago and took some measurements. Could you pull up their data to compare? How about share with the speaker?
- You suspect humidity may be affecting your synthesis. Of all the oxides prepared in your group over the past 10 years, do the ones with lower humidity show cleaner diffraction patterns?
- A new machine learning model is reported in the literature that allows for prediction of new multiferroics based on raw diffraction patterns. How quickly could you apply this model to all the diffraction patterns taken in your lab over the last 10 years?

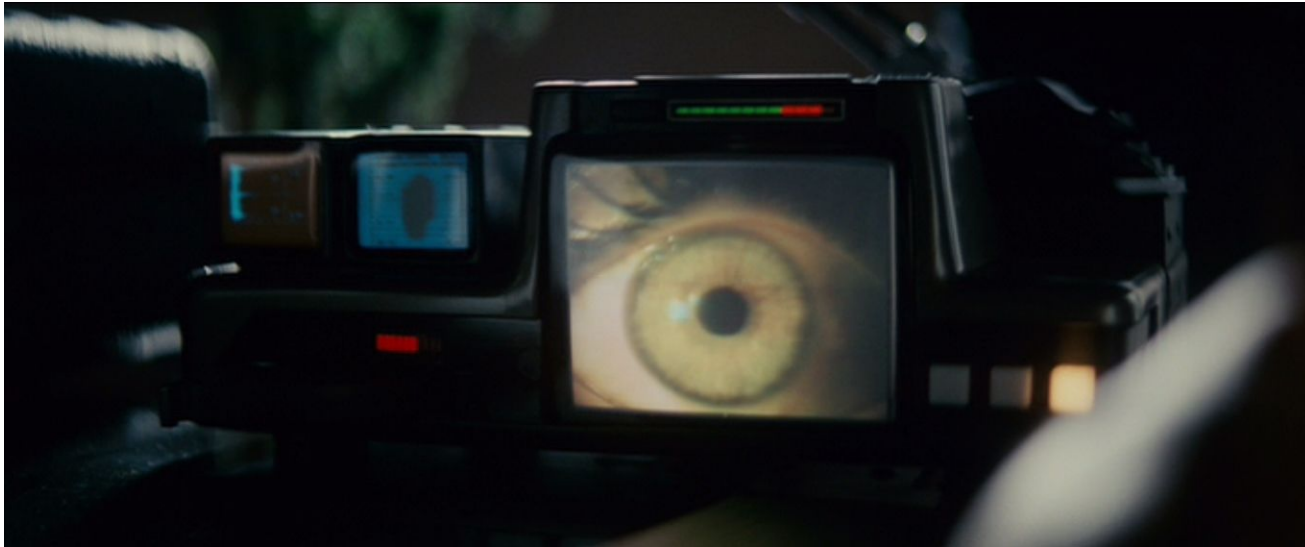
Motivation

- You're in a desert walking along in the sand when all of the sudden you look down, and you see an experimentalist, they are crawling toward you. You reach down, you flip the experimentalist over on its back. The experimentalist lays on its back, its belly baking in the hot sun, beating its legs trying to turn itself over, but it can't, not without your help. But you're not helping. Why is that?

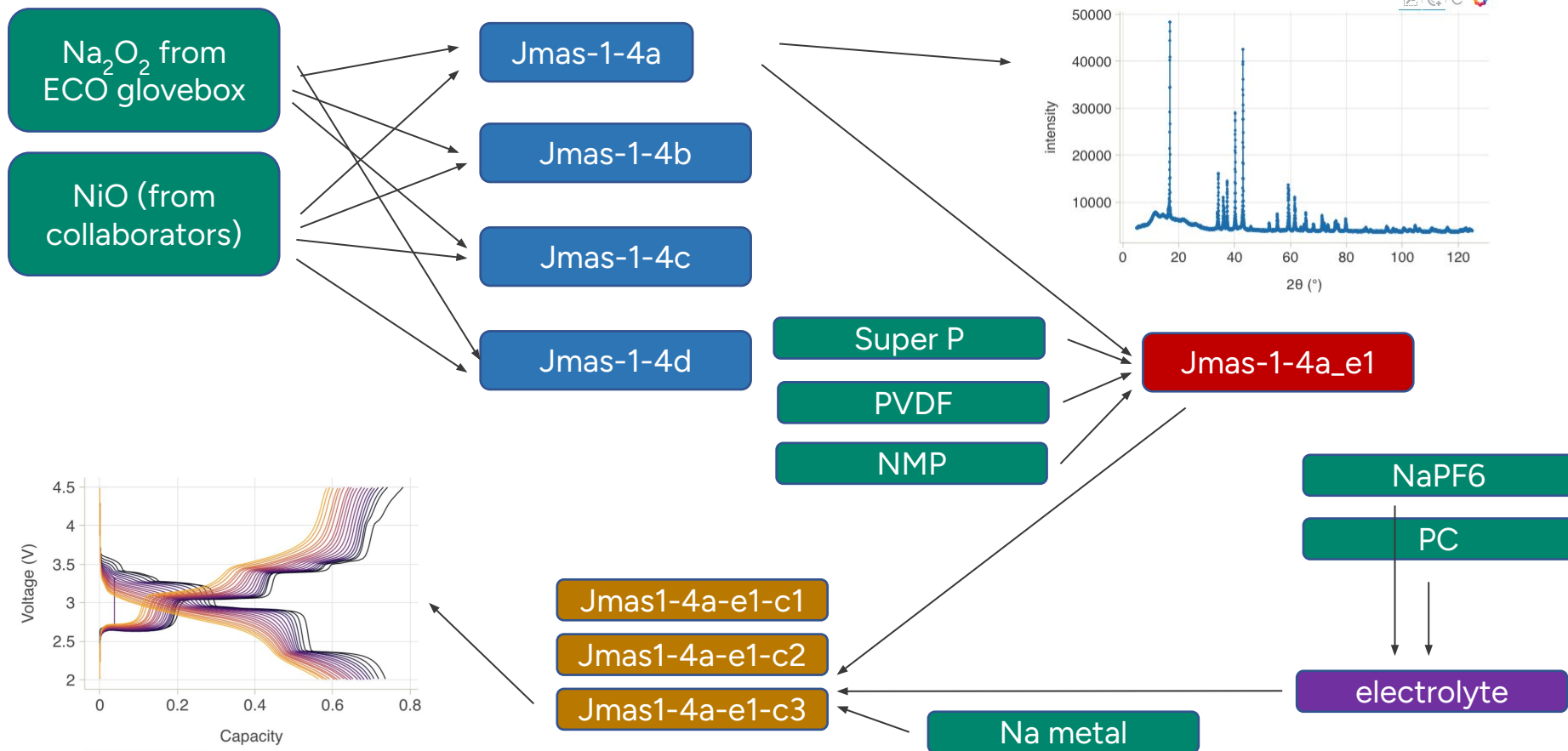


Motivation

- You're in a desert walking along in the sand when all of the sudden you look down, and you see an experimentalist, they are crawling toward you. You reach down, you flip the experimentalist over on its back. The experimentalist lays on its back, its belly baking in the hot sun, beating its legs trying to turn itself over, but it can't, not without your help. But you're not helping. **Why is that?**



Experimental data is only useful in its context!



Options for digital data management

Electronic lab notebook (ELN)

The screenshot shows the OneNote Online interface for a document titled "JM notebook1". The page is titled "Page B" and dated "29 January 2019 12:08". It contains a large block of placeholder text, a "Biography" bar chart, a "Histograms" bar chart, a "Line graph", and a "Pie chart". Below these charts is a table with 4 columns and 6 rows of data. At the bottom, there is a "Protocol" section with text and a button that says "I can move text boxes about".

	Biscuit	Tea	Coffee
Simon	Jamie Dodger	Butlers	occasionally
Aime		occasionally	Froth milk, 2 sugars
Felix	Leibkuchen	various	Caution: extremely fussy
Jo	any	Weak, black, no milk	
Laura	Missionaire's shortbread	milk	

- Flexible
- Human-friendly
- Data generally not normalized and exposed in machine-readable method

Laboratory information management system (LIMS)

The screenshot shows a LIMS interface titled "Lab Test Requests" with a table of test results. The table has columns for Order, Test Type, Date, Date, Patient, Doctor, State, and Urgent. The data is as follows:

Order	Test Type	Date	Date	Patient	Doctor	State	Urgent
13	COMPLETE BLOOD COUNT	10/24/2017	22:58:20	Betz, Ana Bikop	Cordara, Cameron	Ordered	
13	ENDOCRINOLOGY	10/24/2017	22:58:20	Betz, Ana Bikop	Cordara, Cameron	Ordered	
12	ENDOCRINOLOGY	10/24/2017	22:54:37	Carlos, Roberto	Cordara, Cameron	Ordered	
12	RENAL FUNCTION TEST	10/24/2017	22:54:37	Carlos, Roberto	Cordara, Cameron	Ordered	
11	COMPLETE BLOOD COUNT	10/24/2017	22:38:16	Betz, Ana Bikop	Cordara, Cameron	Draft	
11	ZIEHL-NEELSEN BACILLOSCOPY	10/24/2017	22:38:16	Betz, Ana Bikop	Cordara, Cameron	Draft	
10	COMPLETE BLOOD COUNT	10/24/2017	22:20:48	Betz, Ana Bikop	Cordara, Cameron	Draft	
10	HAEMATOLOGY	10/24/2017	22:20:48	Betz, Ana Bikop	Cordara, Cameron	Draft	
9	COMPLETE BLOOD COUNT	01/25/2016	07:57:02	Betz, Ana Bikop	Cordara, Cameron	Ordered	
8	COMPLETE BLOOD COUNT	01/25/2016	04:26:36	Betz, Ana Bikop	Cordara, Cameron	Ordered	
7	COMPLETE BLOOD COUNT	01/24/2016	16:38:15	Carlos, Roberto	Cordara, Cameron	Ordered	
6	LIVER FUNCTION TEST	01/23/2016	04:37:23	Betz, Ana Bikop	Cordara, Cameron	Ordered	
5	RENAL FUNCTION TEST	07/23/2014	04:59:02	Zenon Betz, Matt	Cordara, Cameron	Ordered	
3	RENAL FUNCTION TEST	02/26/2014	14:48:14	Zenon Betz, Matt	Cordara, Cameron	Ordered	
2	CHAGAS KINODONOSIS	02/26/2014	09:23:23	Zenon Betz, Matt	Cordara, Cameron	Draft	
2	DENGUE PRINT	02/26/2014	09:23:23	Zenon Betz, Matt	Cordara, Cameron	Draft	
2	STOOL EXAMINATION	02/26/2014	09:23:23	Zenon Betz, Matt	Cordara, Cameron	Ordered	
1	COMPLETE BLOOD COUNT	09/30/2013	16:46:34	Betz, Ana Bikop	Cordara, Cameron	Ordered	

- Rigid, normalized data model
- Data often machine-readable
- Not often used in academic labs



K:

CELLS. INTERLINKED.

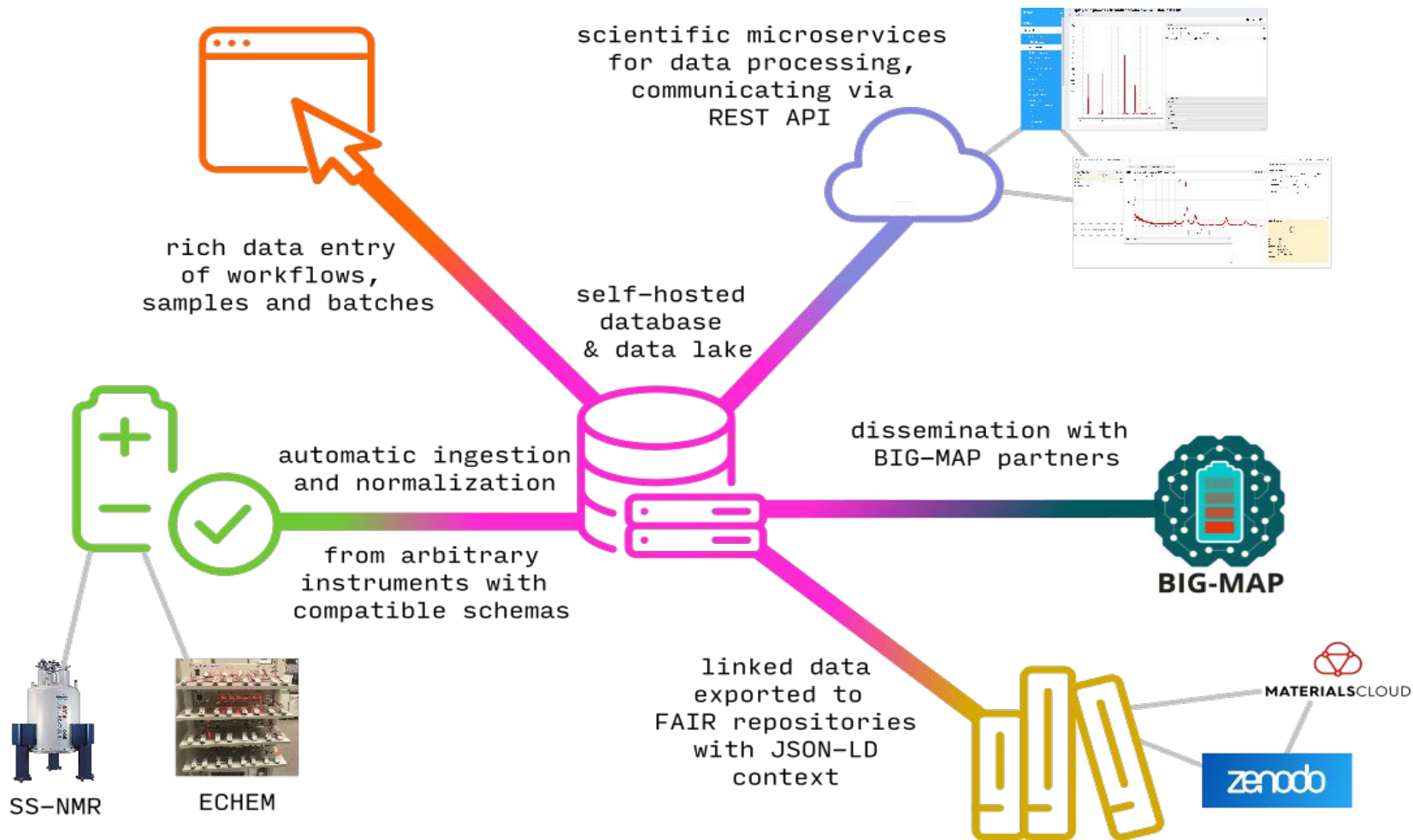


Datalab



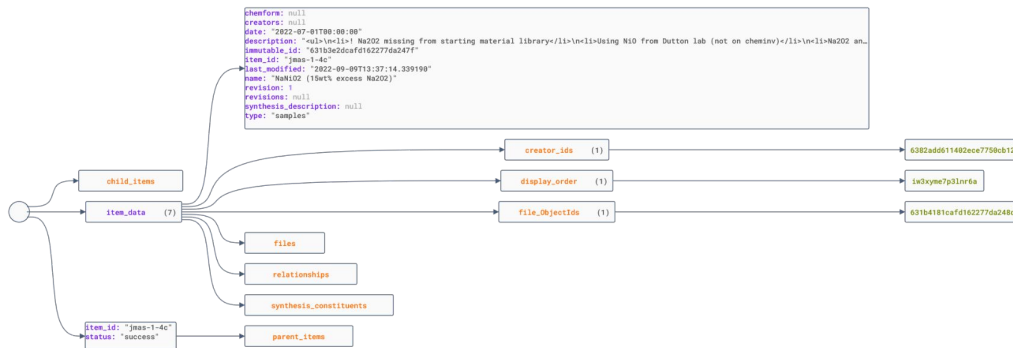
the-grey-group/datalab

- Automating tedious processes
- Reproducible, robust science
- Enhanced dissemination
- Decentralized data unification
- Putting the FUN in data management



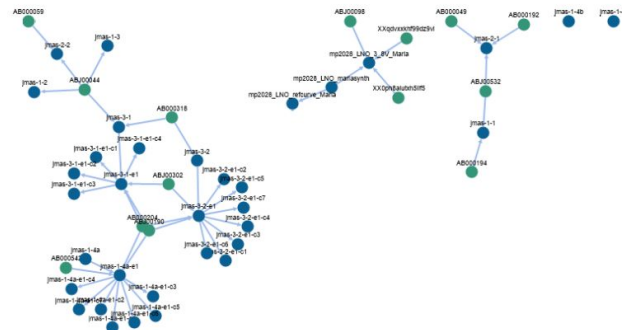
Server

- A REST API for chemical data, analysis and their connections.
- Real-time data streaming and syncing with remote data sources (e.g., instrumentation, archives and file stores).
- OAuth2-based user authentication via GitHub or ORCID and simple user role management.
- RabbitMQ + Celery asynchronous data extraction



UI

- A simple, intuitive UI for recording sample metadata and relationships with other samples (batches, offshoots), alongside synthesis parameters and raw data.
- Basic analysis and plotting of live data attached to a sample, e.g., XRD, NMR, EChem and photos/videos.
- Interactive network visualisation of the connections between samples, cells and starting materials.





Greymon (bot)



Logout

[About](#) | [Samples](#) | [Collections](#) | [Inventory](#) | [Graph View](#)

datalab is a place to store experimental data and the connections between them.

datalab is open source (MIT license) and development occurs on GitHub at [the-grey-group/datalab](#) with documentation available on [ReadTheDocs](#).

Deployment stats:

24	191	204
Active Users	Samples	Cells

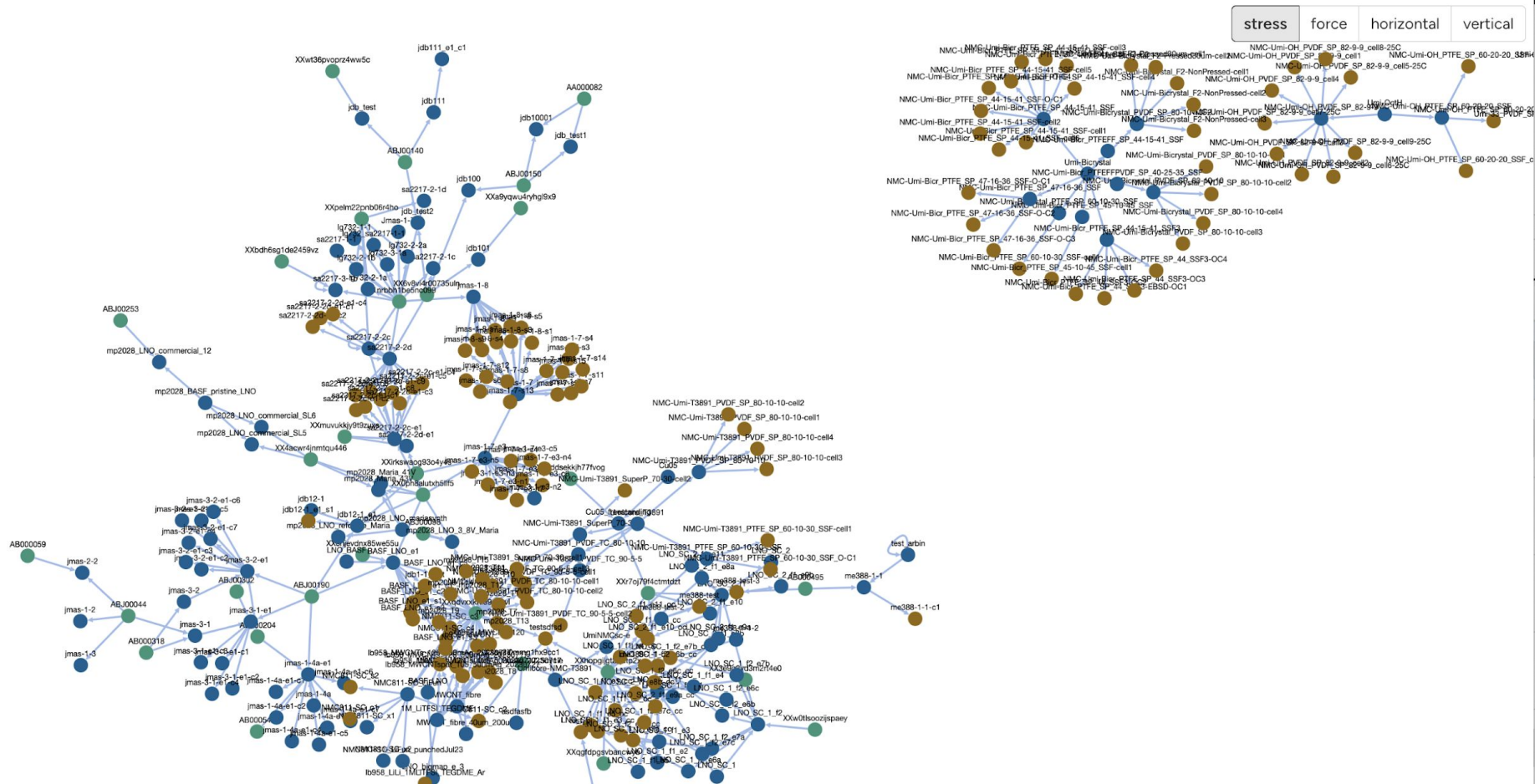
datalab was primarily developed by:

- [Joshua Bocarsly](#) (Department of Chemistry, University of Cambridge)
- [Matthew Evans](#) (MODL-IMCN, UCLouvain & Matgenix)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 957189 (DOI: [10.3030/957189](#)), the [Battery Interface Genome - Materials Acceleration Platform \(BIG-MAP\)](#), as an external stakeholder project.



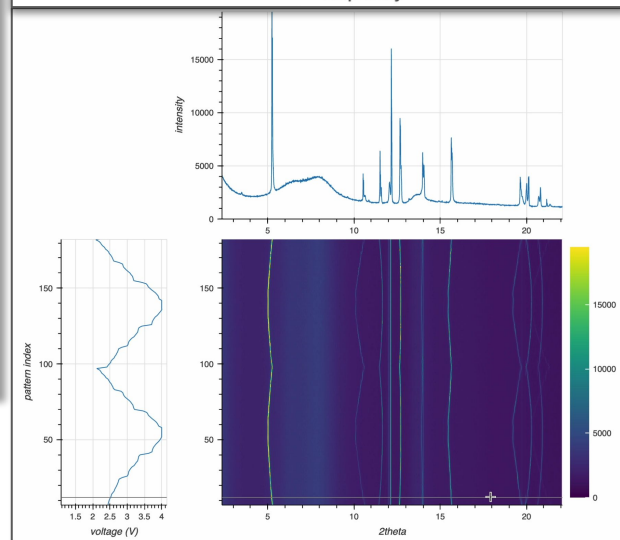
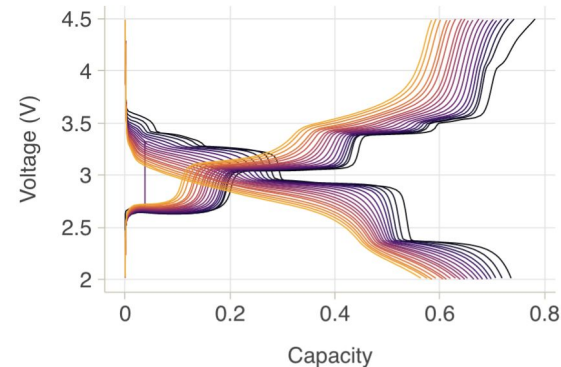
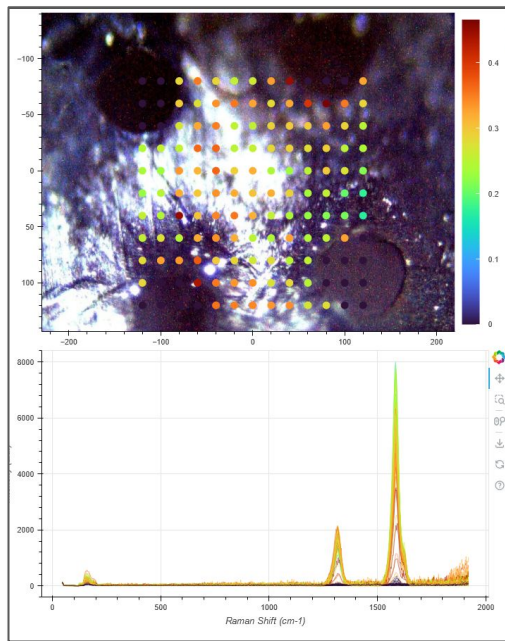
Datalab pilot: connected, contextual data from ~5 users in ~3 months



Modular “Data block” approach: bottom-up interoperability

Currently, we support:

- Electrochemical cycling
- EIS
- Materials synthesis
- Cell preparation
- Images (common + raw)
- Videos
- Powder X-ray diffraction
- Raman (thanks to Larry Brazel)
- NMR
- Mass spectrometry/TGA
- Your method?



Purposefully simple to add new “blocks”:

- File loader
- Interactive plot
- Baseline corrections, background subtraction, schemas, models, validation, search, comparative analysis...

Whinchat , a chat-based interface to materials data

One benefit of digitally recording data, is that it is available for **future** uses in artificial intelligence

For example have found that large language models (e.g. chatGPT) are able to serve as “virtual lab data assistants” when provided with interconnected data. **whinchat**, a GPT-based virtual assistant, is integrated into datalab.

Great potential for:

- dealing with structured data
- formatting (e.g, generate a table in RSC journal style...)
- explorative analysis
- fine-tuning on relevant knowledge bases

LLM Chat Block with contextual data (powered by GPT-3.5-turbo) ↻ ↗ ✕

Hello! I am your virtual assistant and I am here to help you manage your experimental data. What can I help you with today? raw | formatted

Please make me a mermaid flowchart showing how this cell was made from other items in the database raw | formatted

Certainly! Here is a Mermaid flowchart that shows how the cell with item_id jdb1-1_e1_c1 was made from other items in the database:

```
graph LR; A[AL00140 Sodium carbonate] -- "was used to create" --> B[jdb1-1 NaCoO2 made by solid state reaction]; C[Xb02072pyridine Cobalt (II) Oxide] -- "was used to create" --> B; B -- "was used to create" --> D[jdb1-1_e1 NaCoO2 electrode]; E[OxideInorganicPoly Super P] -- "was used to create" --> D; F[AB00042 Pyran 100 900 proof monomer] -- "was used to create" --> D; G[Xidm0200m0200f N-methyl-2-pyrrolidone] -- "was used to create" --> D; D -- "was used to create" --> H[jdb1-1_e1_c1 NaCoO2 coin cell]
```

In this flowchart, each rectangular box represents an item in the database, and the arrows between the boxes represent the relationships between the items. For example, SM1 and SM2 were used to create sample JDB1-1, which was then used to create sample E1, which was finally used to create the cell C1.

raw | formatted

K.M. Jablonka, ..., J. Bocarsly, ..., **M. L. Evans**, *et al.*, 14 Examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon, *Digital Discovery* **2** (2023) 1233-1250.



Scaling



the-grey-group/datalab

Vertical

- Simple plugin ecosystem of “data blocks” for new techniques
- Integration with MaRDA extractors working group: communalize the effort of parsing new files
- Support development of open core of *datalab*

Horizontal

- Public deployments
- Automated cloud deployments (paid?)
- Federation of instances, c.f. OPTIMADE
 - Opt-in registry of instances with own namespace
- Customisable schemas



Metadata extractor interoperability for materials science and chemistry

www.marda-alliance.org

Matthew Evans (UCLouvain)
Peter Kraus (TU Berlin)
David Elbert (Johns Hopkins)



[marda-alliance/metadata_extractors](https://github.com/marda-alliance/metadata_extractors)

<https://marda-registry.fly.dev/docs>

Aims & Motivations

- I. Enable infrastructure, archive or ELN developers robustly parse new file types
- II. Improving the quality and discoverability of parsers in the community with schemas
- III. Indexing over relevant domain data and metadata rather than using generic archives

Design Goals

- I. **A lightweight metadata schema for extractors**
- II. **A common API specification for executing extractor code**
- III. **A searchable registry of extractors and file types**



Roadmap - near future

- Robust feedback! **Avoid Lilliput and Potemkin**
- Comparative plots and analysis documents: combining arbitrary sets of samples/devices on single plots
- Distributed development team
- Enhanced admin tools

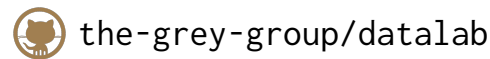


the-grey-group/datalab

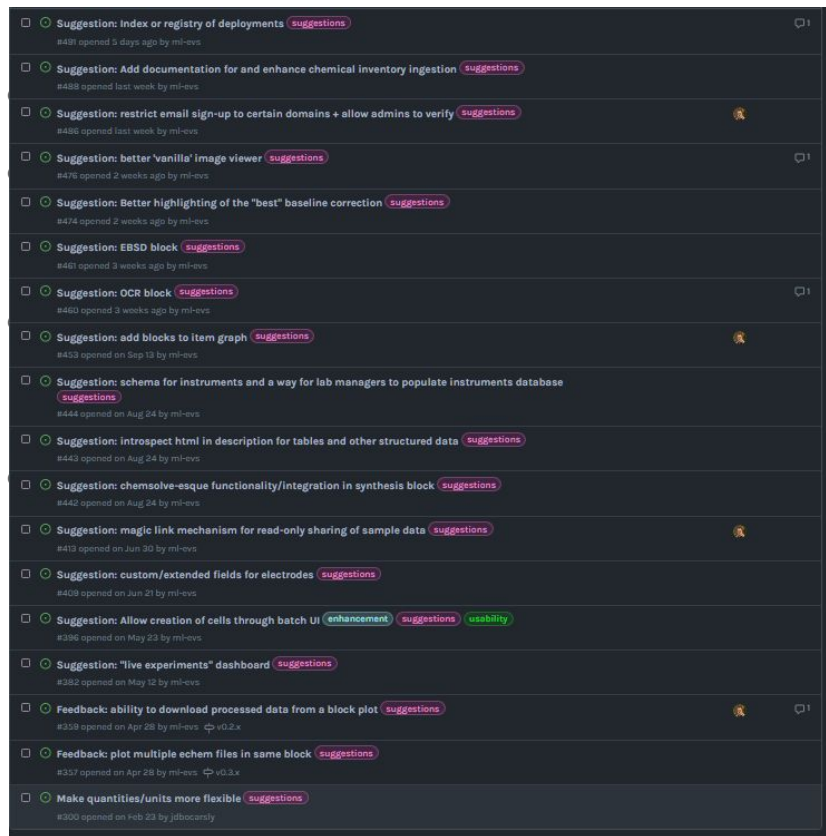
- Enumeration of instruments and locations:
 - Manager of instrument can issue alerts, e.g., calibration needed, glovebox contaminated
 - Automatically capture instrument state at time of file creation/measurement



Roadmap - medium future



- Python/R API for interacting with data in e.g., Jupyter notebooks
- Self-hosted Jupyter “Hub”
- Deployment-level schema customisation
- Secure some sustainable funding
- Proper/automatic exports to archive servers
- Dashboards of live experiments
- “Offline” use of plugins



Datalab is open-source software (MIT)

Collaborative development is performed on GitHub using modern best practices for open source software development.

~23000 lines of code

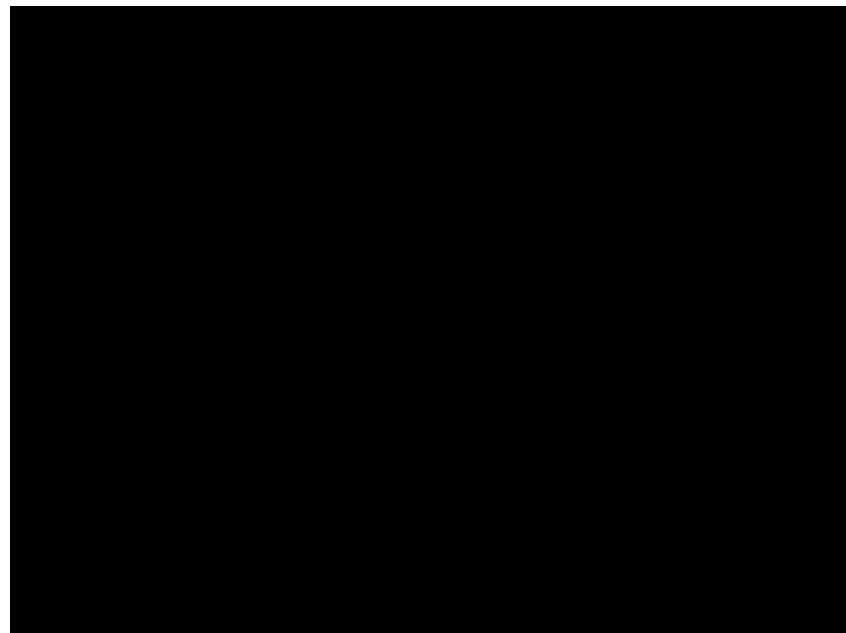
152 merged PRs

Automated CI with testing of server and GUI

Two components:

- pydatalab – a Flask-based Python web server
- A Vue-based JavaScript UI

Simplified containerized deployment via Docker (instructions online)



github.com/the-grey-group/datalab



Get involved!

Expression of interest



matthew.evans@uclouvain.be
jdbocars@central.uh.edu




the-grey-group/datalab

A screenshot of the datalab web interface. At the top, there is a navigation bar with links for "About", "Samples", and "Graph View". A "Login/Register" dropdown menu is open, showing options for "Login via GitHub", "Login via ORCID", and "Login via email". Below the navigation bar, the text reads: "datalab is a place to store experimental data and the connections between them." Further down, a paragraph states: "This is a public deployment of datalab, primarily used for testing and demonstrations. Anyone is free to make an account by either connecting their GitHub account, or by registering via their email (click sign-in/register above)." A yellow box contains a disclaimer: "Any data captured here will not be visible to other users except the admins of the deployment, who may use it solely for debugging purposes. We additionally provide no assurances about availability or permanent data retention on this system, at this time." A light blue box contains a call to action: "If you would like to register your interest and give feedback, please fill in the survey. Specific feedback can also be provided on the issue tracker on GitHub." At the bottom, it says: "datalab is open source (MIT license) and development occurs on GitHub at the-grey-group/datalab with documentation available on ReadTheDocs."

public.datalab.odbx.science

Thank you for listening!



Prof Joshua Bocarsly
(Cambridge 
UHouston)



Dr Peter Kraus
(TUBerlin)

- Prof Gian-Marco Rignanes & Dr David Waroquiers
- Dame Prof Clare Grey FRS
- Testers & developers in the Grey Group
 - Ben Smith
 - Larry Brazel
 - James Steele
 - Dr. Veronika Sedajova
 - Megan Penrod
 - Sarang Balan
 - + many more

 **UCLouvain**



Useful links

Source code repository + issue tracker:

<https://github.com/the-grey-group/datalab>

Documentation:

<https://the-datalab.readthedocs.io>

Public demo deployment:

<https://public.datalab.odbx.science>

Give feedback & register interest

<https://bit.ly/datalab-survey>

MaRDA Extractors WG

<https://bit.ly/marda-extractors>

