

Beyond the dataset with *datalab*

Data capture as a bottleneck of AI4Materials

Dr Matthew Evans

*Leverhulme Trust Early Career Research Fellow
Department of Chemistry, University of Cambridge*

*Director
datalab industries ltd., King's Lynn, UK*

Materials 4.0: AI for Science Showcase, Manchester
19th March 2026

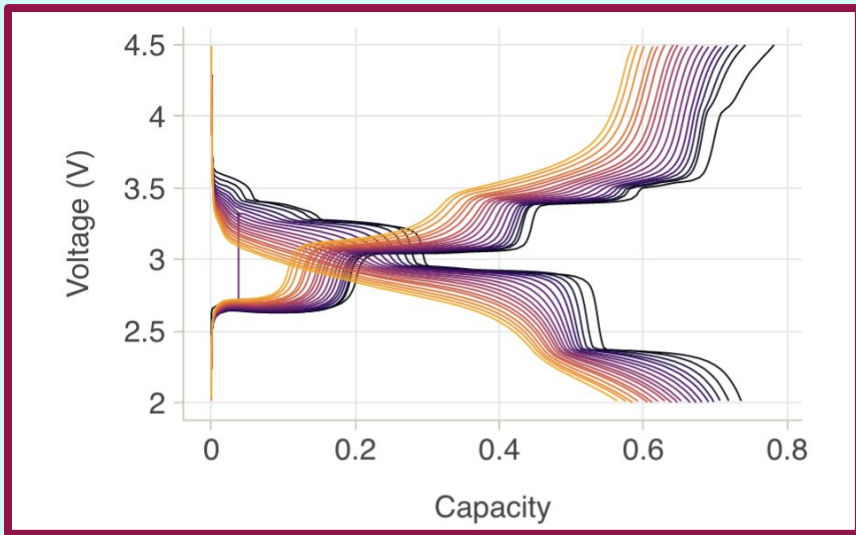
LEVERHULME
TRUST



UNIVERSITY OF
CAMBRIDGE



> managing experimental data

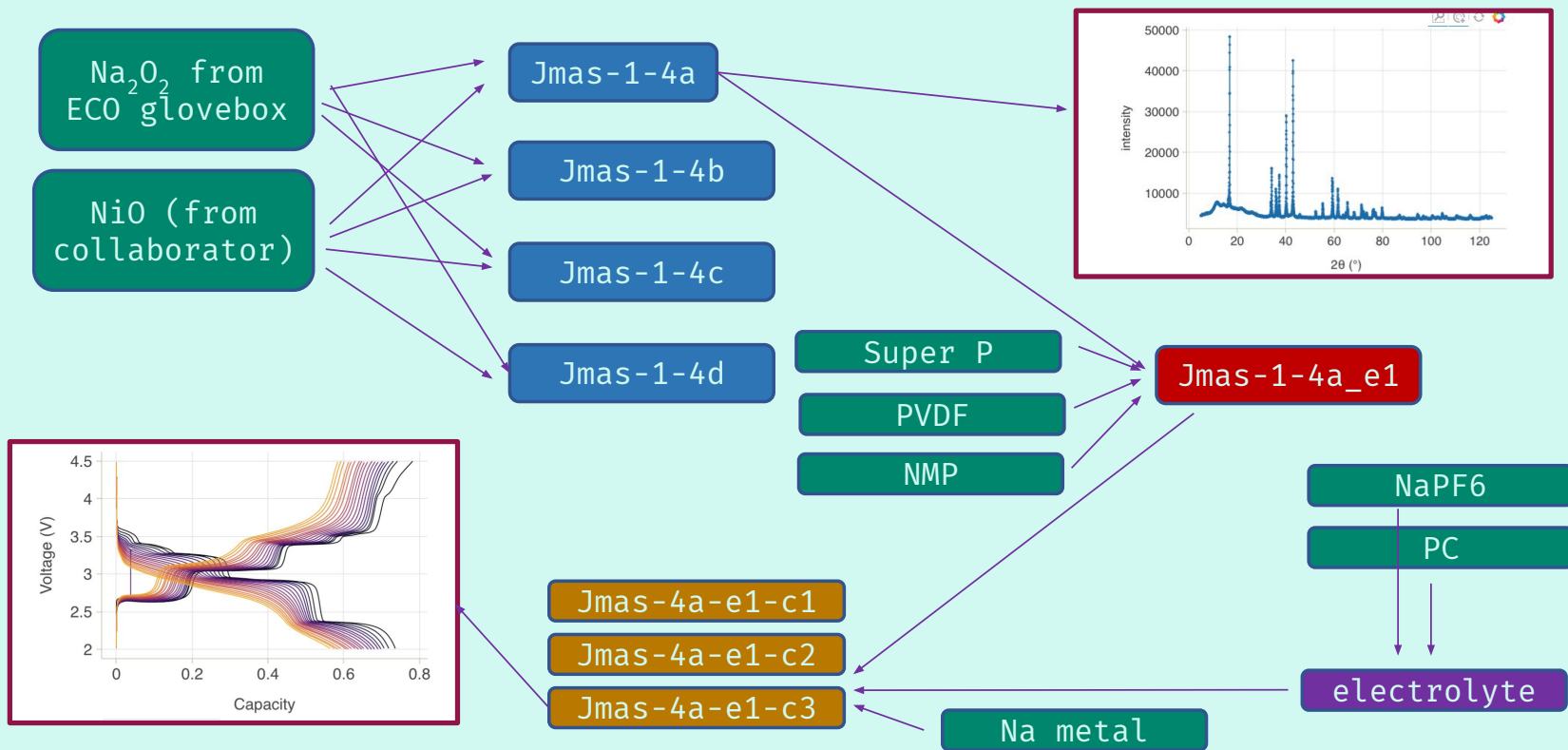


What information would need to be recorded to make this information useful after the student leaves?

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?

> experimental data is only useful with context



> 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?
- Someone accidentally caused an oxygen spike in a glovebox. How can I report this and notify everybody who has samples in there at that time?

> 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?
- Someone accidentally caused an oxygen spike in a glovebox. How can I report this and notify everybody who has samples in there at that time?

> 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?**
- Someone accidentally caused an oxygen spike in a glovebox. How can I report this and notify everybody who has samples in there at that time?

> 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?
- **Someone accidentally caused an oxygen spike in a glovebox. How can I report this and notify everybody who has samples in there at that time?**

bottleneck: experimental data capture

- Non-machine readable data
- Missing connections and context between data points
- Missing negative/failed data of “dark reactions”
- Non-systematic experiment design

> ai4science

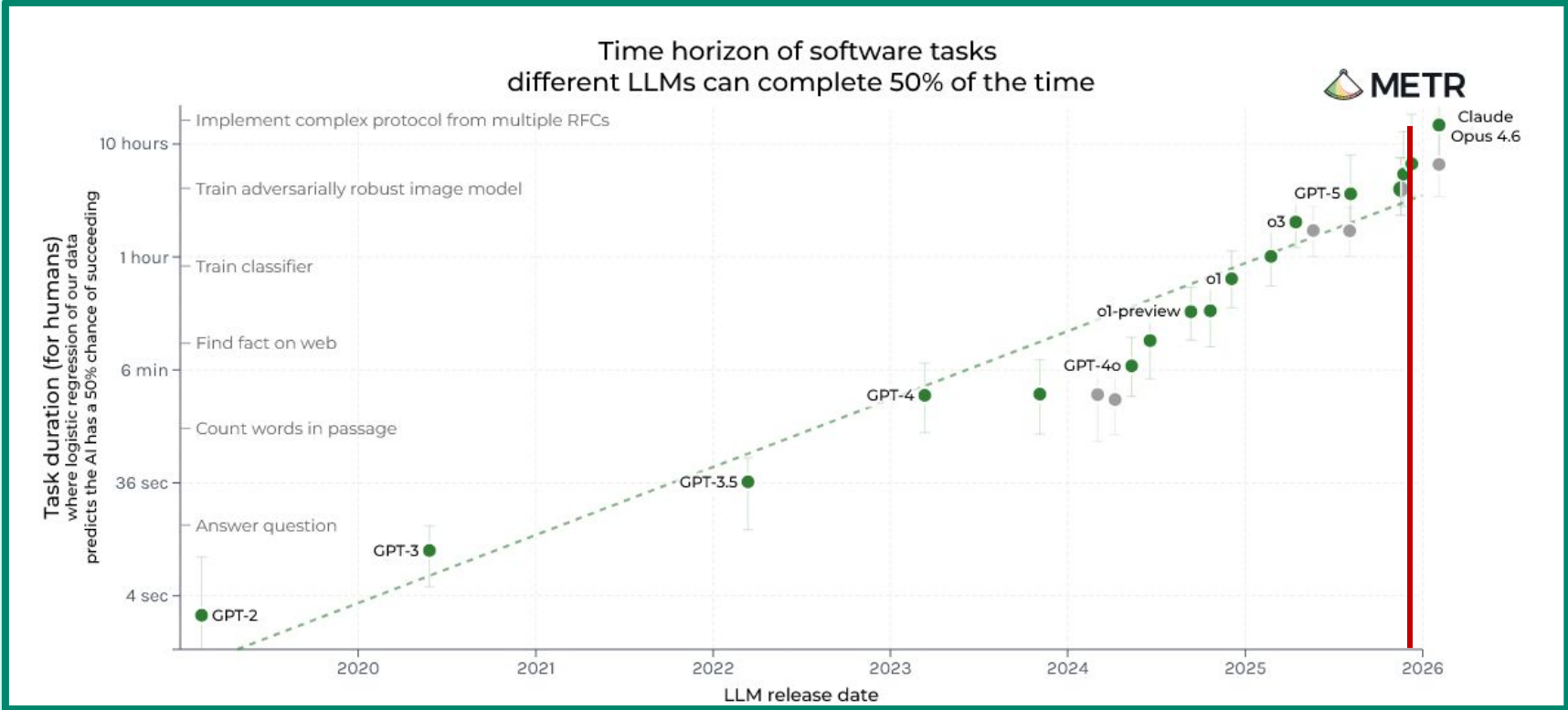
Universal, invertible simulators

- Universal MLIPs
- Property/spectra prediction
- Multiscale modelling
- Generative models
- Inverse design
- Conditional generation
- A “world model” for atoms

AI scientists

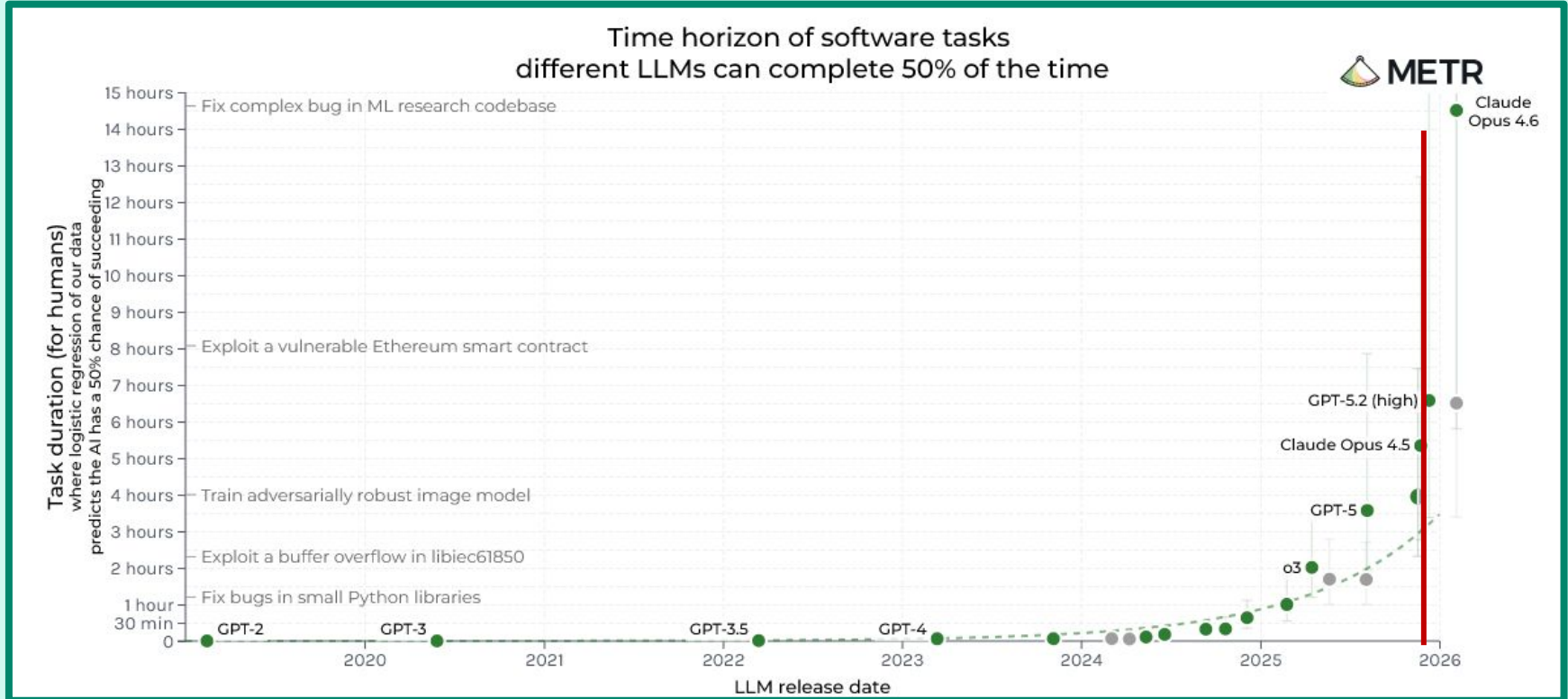
- Literature agents/assistants
- Hypothesis generation
- Experimental design
- Synthesis planning
- Multimodal reasoning
- Lab automation and control
- Code generation & steering simulations

LLMs & agentic AI




LLMs & agentic AI


Benchmark saturated?



> *datalab*

 [datalab-org/datalab](https://github.com/datalab-org/datalab)

- Automating tedious data handling
- Designing for the (shifting) bottleneck
- Incrementally adoptable
- Enhanced dissemination
- Decentralized serendipity

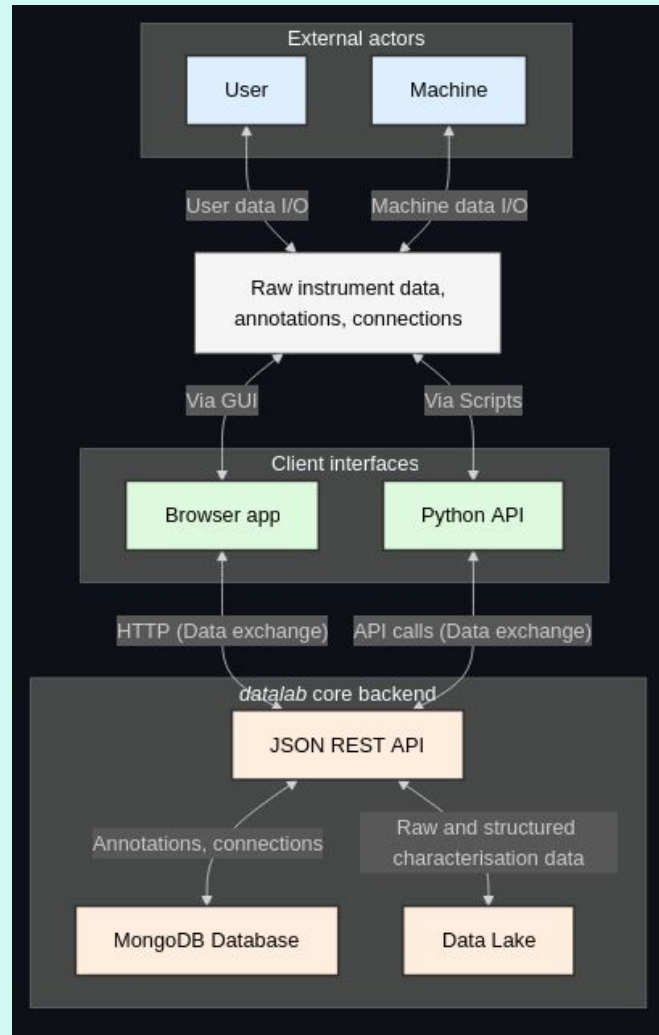
Prof Joshua
Bocarsly
(Cambridge 
UHouston)



Dame Prof
Clare Grey FRS
(Cambridge)



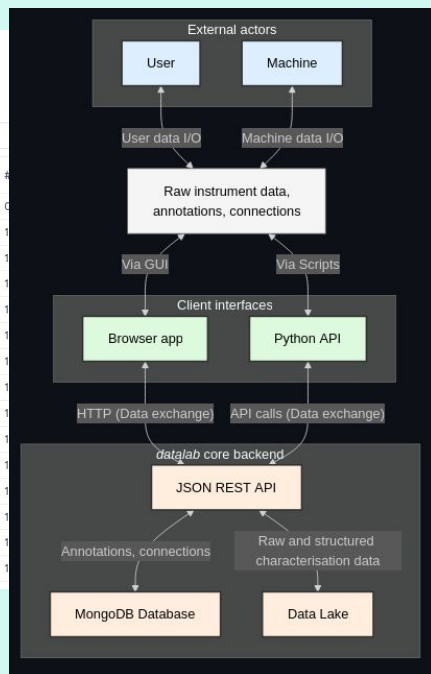
What is *datalab*?



What is *datalab*?

ID	Type	Sample name	Formula	Date	Collections	Creators
JUAWHS	samples			2025-02-26		Matthew Evans
FZCMFK	samples	250113 VJCIP36-2 T...		2025-01-13	Feb26-2025-import	Matthew Evans
ELCAVJ	samples	250113 TMAOH/H-F...		2025-01-13	Feb26-2025-import	Matthew Evans
UZMDW	samples	PFA500 OH (UF)		2025-01-10	Feb26-2025-import	Matthew Evans
BYXLJU	samples	D296-HPO42- EXC...		2025-01-09	Feb26-2025-import	Matthew Evans
LNKTY	samples	YKTH-102-HT STAB...		2025-01-08	Feb26-2025-import	Matthew Evans
BOEGFV	samples	D290(Y)-PO4 EXCH...		2025-01-07	Feb26-2025-import	Matthew Evans
BJBGMW	samples	D296-OH-RE		2025-01-06	Feb26-2025-import	Matthew Evans
IGYHFG	samples	PFA500 OH (UF)		2025-01-03	Feb26-2025-import	Matthew Evans
TJPFVO	samples	250102 PFA500-O...		2025-01-02	Feb26-2025-import	Matthew Evans
SSEWR	samples			2025-01-02	Feb26-2025-import	Matthew Evans
AZJHML	samples			2025-01-02	Feb26-2025-import	Matthew Evans
FYVBP	samples	250102 PFA500-O...		2025-01-02	Feb26-2025-import	Matthew Evans
PHOTOG	samples	250102 PFA500-O...		2025-01-02	Feb26-2025-import	Matthew Evans
GTHPLV	samples			2024-12-31	Feb26-2025-import	Matthew Evans

Vue.js app



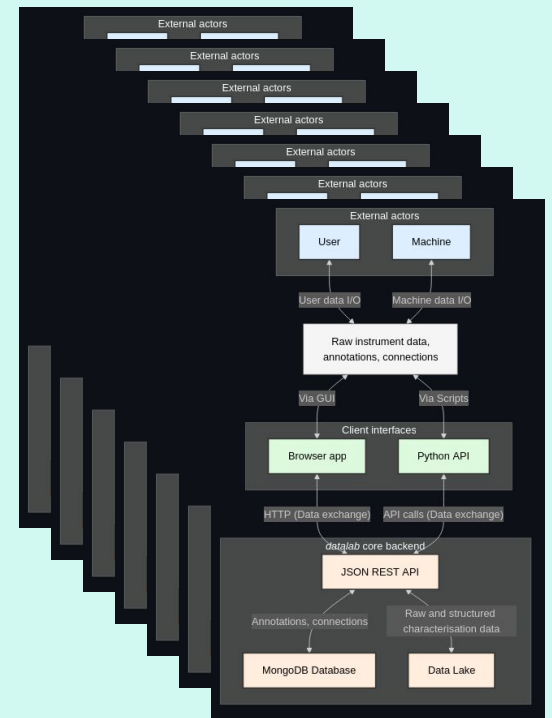
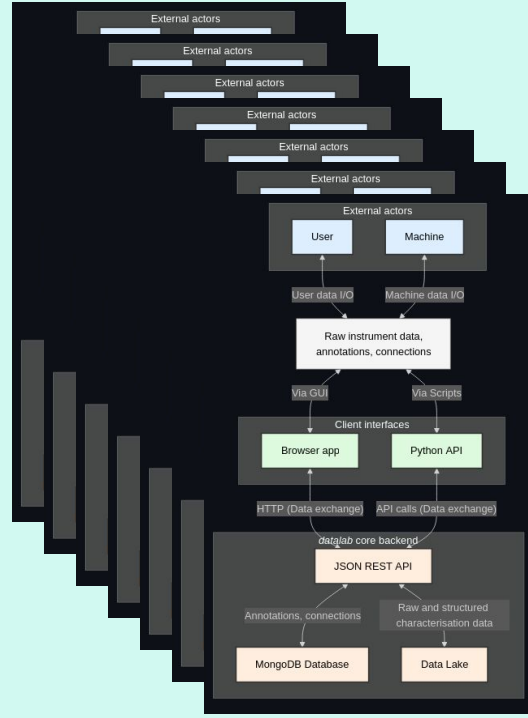
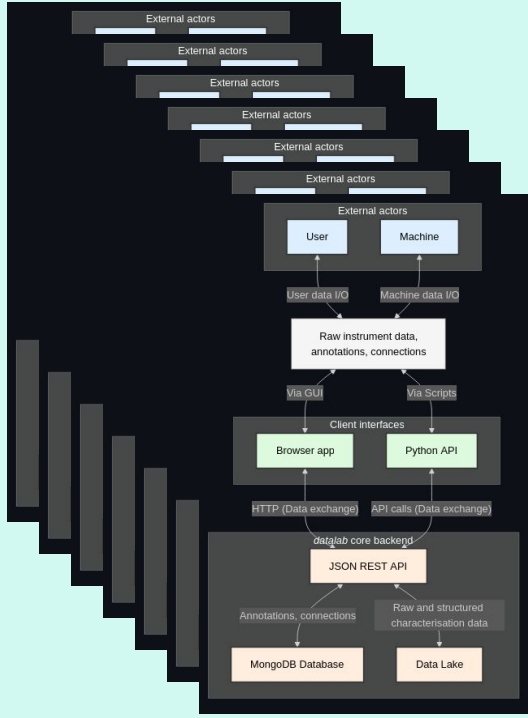
```
In [6]: items = client.search_items("test", display=True);
```

type	ID	refcode	name
S	test	pub:HOPZJO	test
S	my_test_sample	pub:IGCYOI	Test Sample
S	test-claude	pub:JMGOTA	
S	test_sample	pub:JTREWV	G1500
S	JPN01	pub:AIMCHP	Siloxane Reaction
S	jdb102	pub:WRIZSB	Sodium cobalt oxide
C	jdb2_e1_c1	pub:TBBADR	NaCo2 coin cell
S	jdb1	pub:VRSNCI	sodium cobalt oxide made by s...
S	jdb2	pub:HPPPKI	sodium cobalt oxide made by s...
C	NMC811_e1_c1	pub:FAJBJB	NMC-811 coin cell

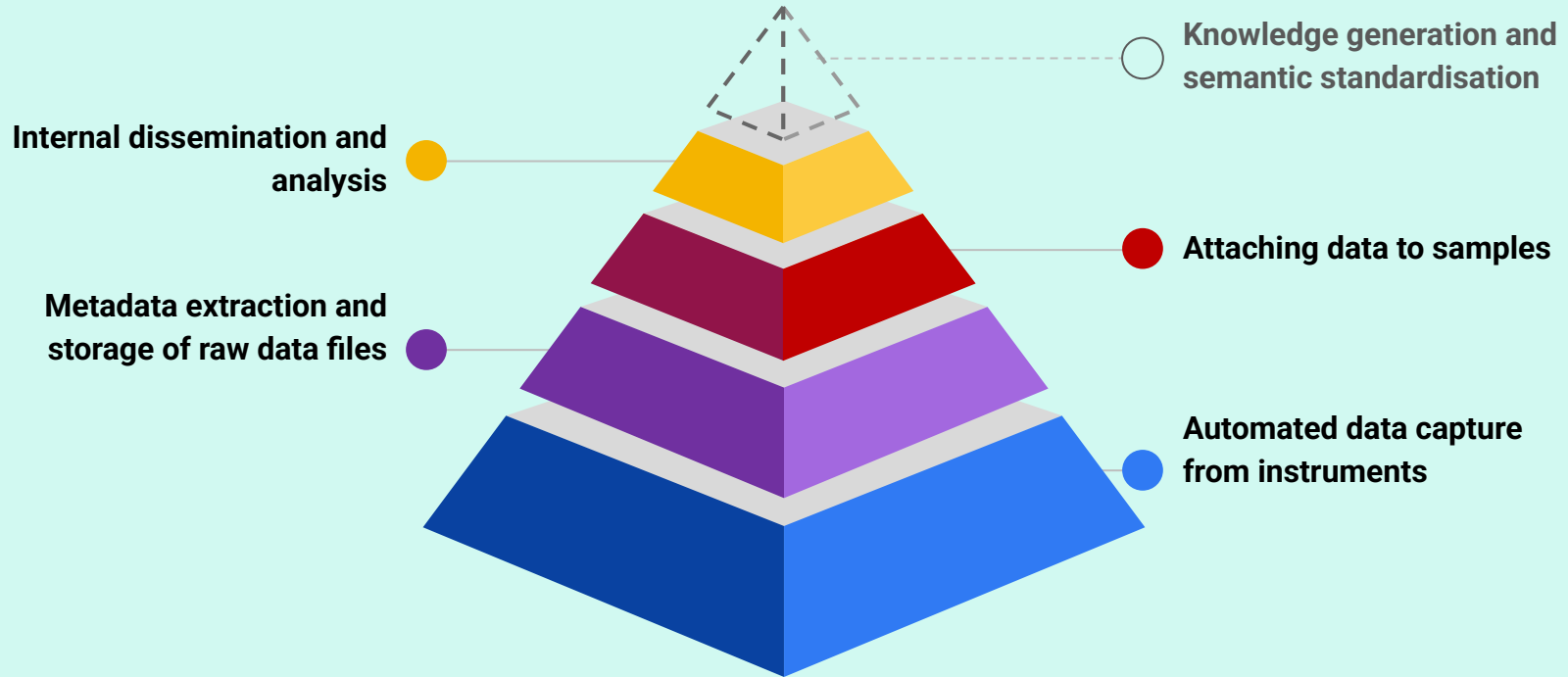
list all items of a specified type:

Python API package

What is *datalab*?



What is *datalab*?

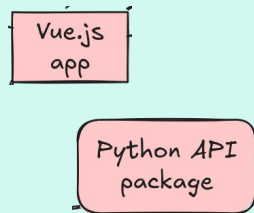


> architecture

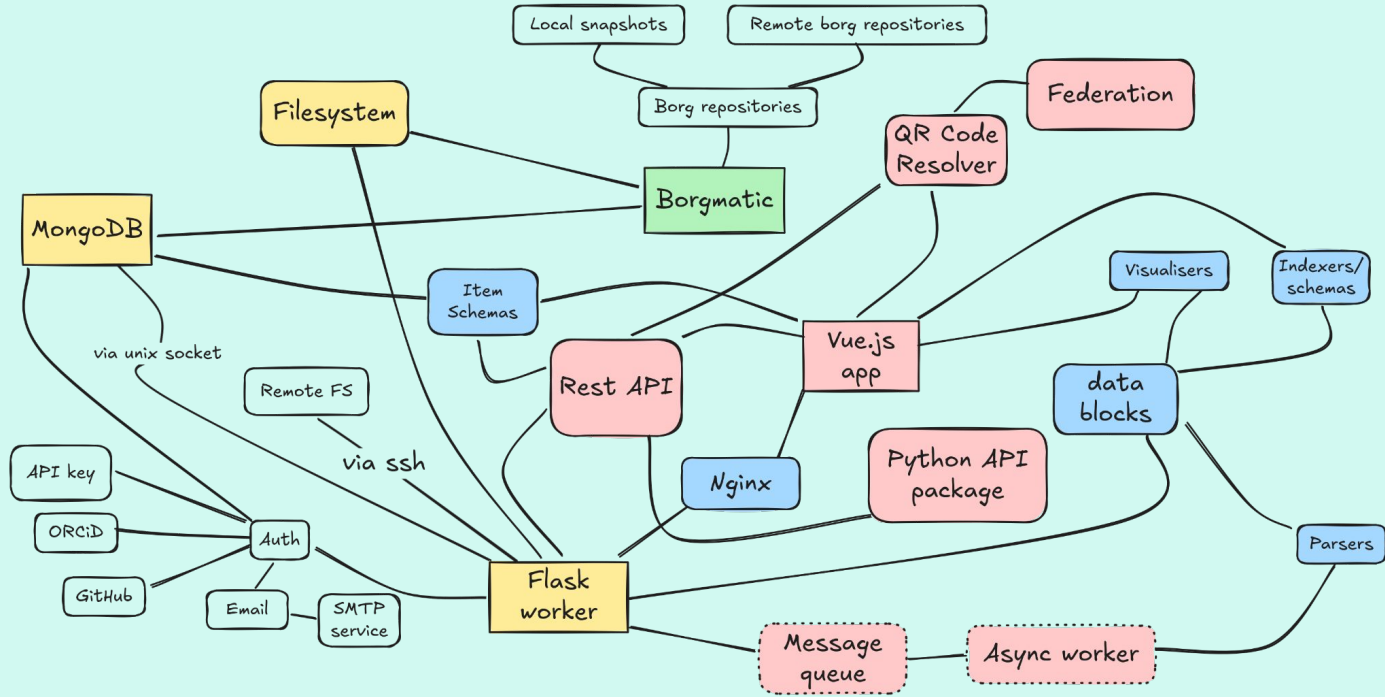


Vue.js
app

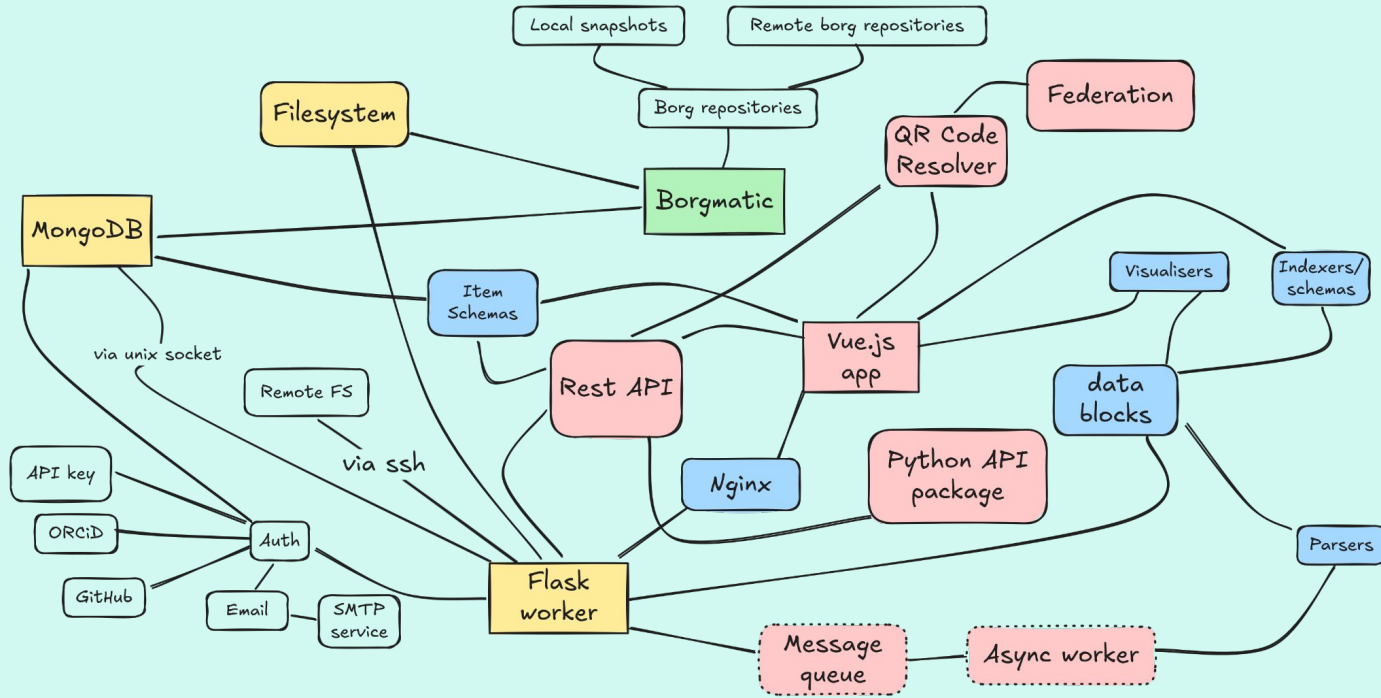
> architecture



> ~~architecture~~ spaghetti



> ~~architecture spaghetti~~ “infrastructure”



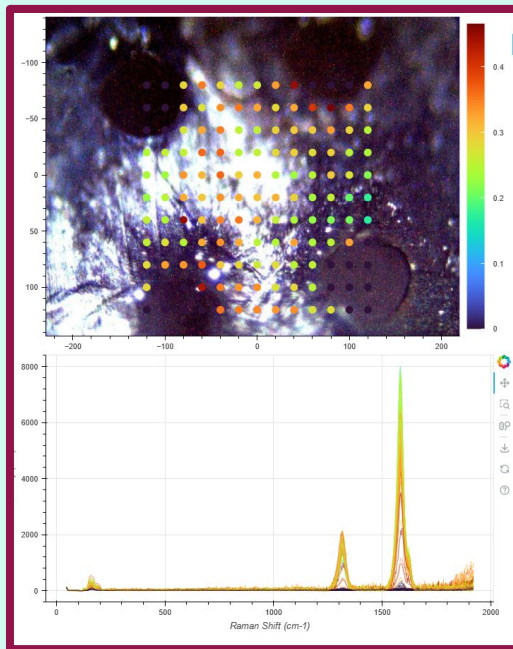
> data blocks

Modular support for 15+ characterisation techniques (XRD, NMR, imaging, electrochemistry, etc.)

Extracting raw data to performing basic analysis and indexing metadata

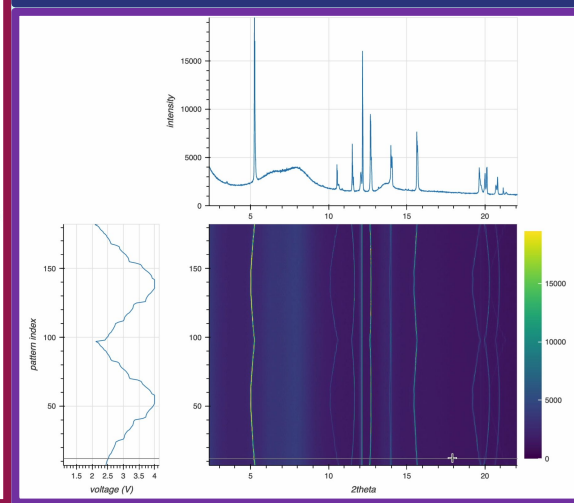
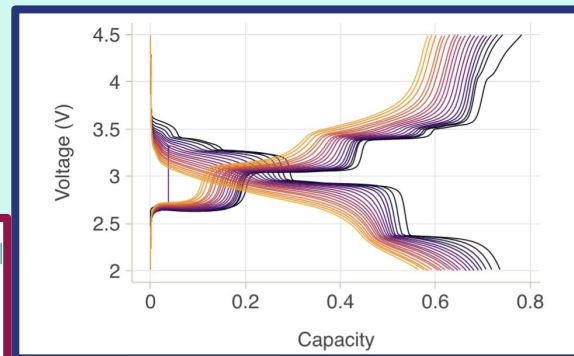
Purposefully simple to add new blocks:

- File loader
- Interactive plot
- Provide out of band information (e.g., wavelength)
- Baseline corrections, background subtraction, search, QC, comparative analysis...



Raman microscopy

Electrochemical cycling



in situ Synchrotron XRD

> whinchat (2023)



Chat interface to individual *datalab* entries or collections

- Stuff structured data from a collection of samples into context window of an LLM, then interrogate “in-context learning”

Jablonka *et al*, *Digital Discovery*, [10.1039/D3DD000113J](https://doi.org/10.1039/D3DD000113J)

Hello! I'm whinchat 🐦, your virtual data management assistant. It looks like you've provided a comprehensive JSON dataset for a series of sodium cobalt oxide (NaCoO_2) coin cells and related samples, all made to test their electrochemical performance.

raw | formatted

In these experiments, is there a relationship between the battery performance and the synthesis or processing conditions?

raw | formatted

whinchat 🐦 here. Let me analyze the relationships between synthesis and performance:

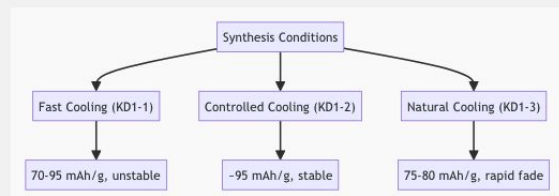
The samples (KD1-1, KD1-2, KD1-3) were all synthesized similarly at 850°C under O_2 , but with some key differences:

- KD1-1: 20 hours dwell, fast cooling (4 hours)
- KD1-2: 24 hours dwell, controlled cooling ($5^\circ\text{C}/\text{min}$)
- KD1-3: 24 hours dwell, natural cooling (15 hours)

Looking at cell performance:

- KD1-1 cells: ~70-95 mAh/g, significant degradation
- KD1-2 cells: Best performer with cell KD1-2_e1_c3 showing ~95 mAh/g with “reasonably stable cycling”
- KD1-3 cells: 75-80 mAh/g, all showing rapid capacity fade

The trend suggests that the controlled cooling rate (KD1-2) produced the most stable material, while both fast cooling (KD1-1) and very slow cooling (KD1-3) resulted in materials with poorer cycling stability. This could be related to the formation of optimal crystal structure or sodium content during the cooling process.



raw | formatted

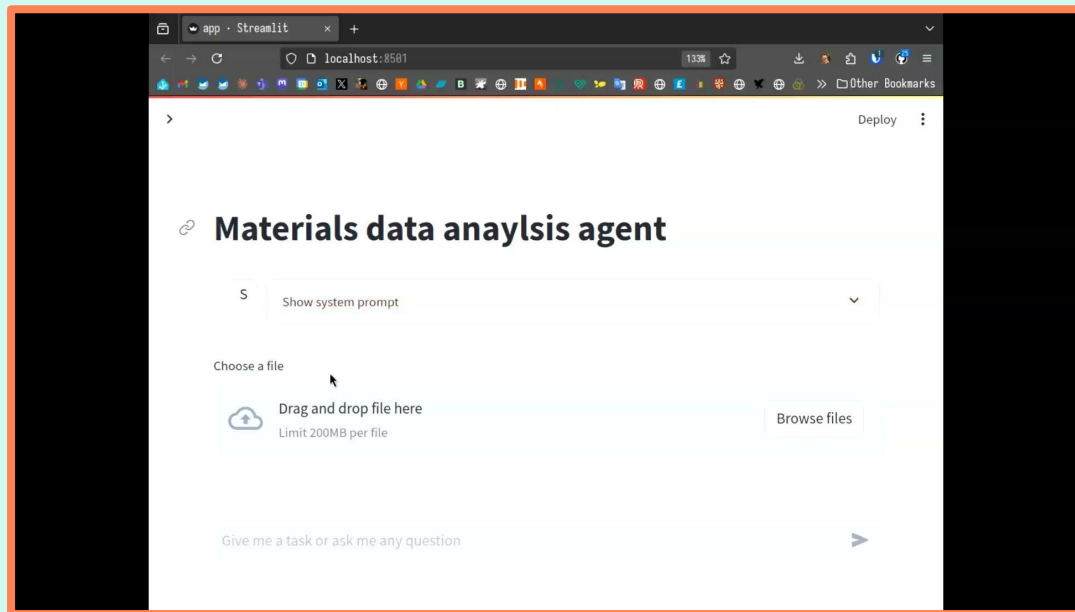
> yeLLowhaMMer (2024)



datalab-org/yellowhammer (with Josh Bocarsly & Yue Wu)

Agentic control of *datalab* API; code generation from natural language for end user programming with added guard rails

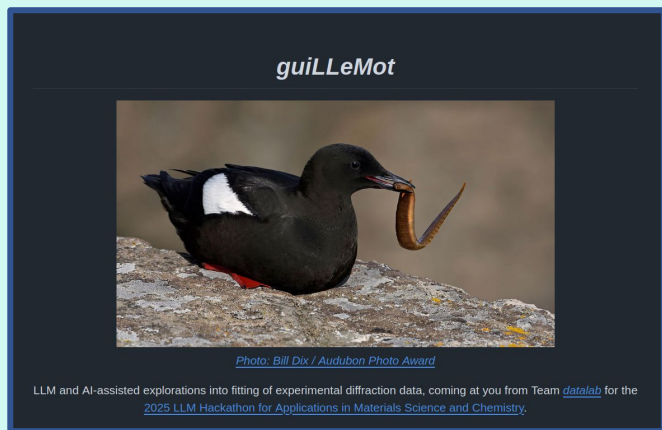
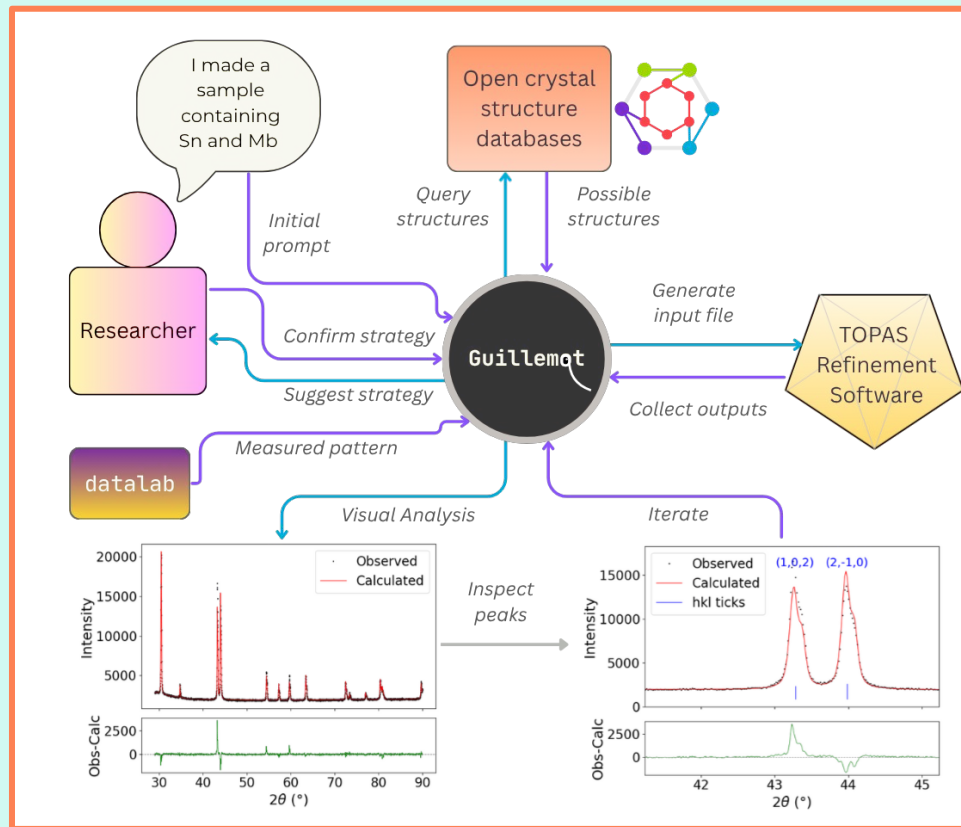
- Can automate complicated (not complex) data analysis tasks, e.g.
 - Search/RAG
 - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Jupyter “magic” `%llm` interface reduces complexity
- Saves us writing UI workflows for complicated features!



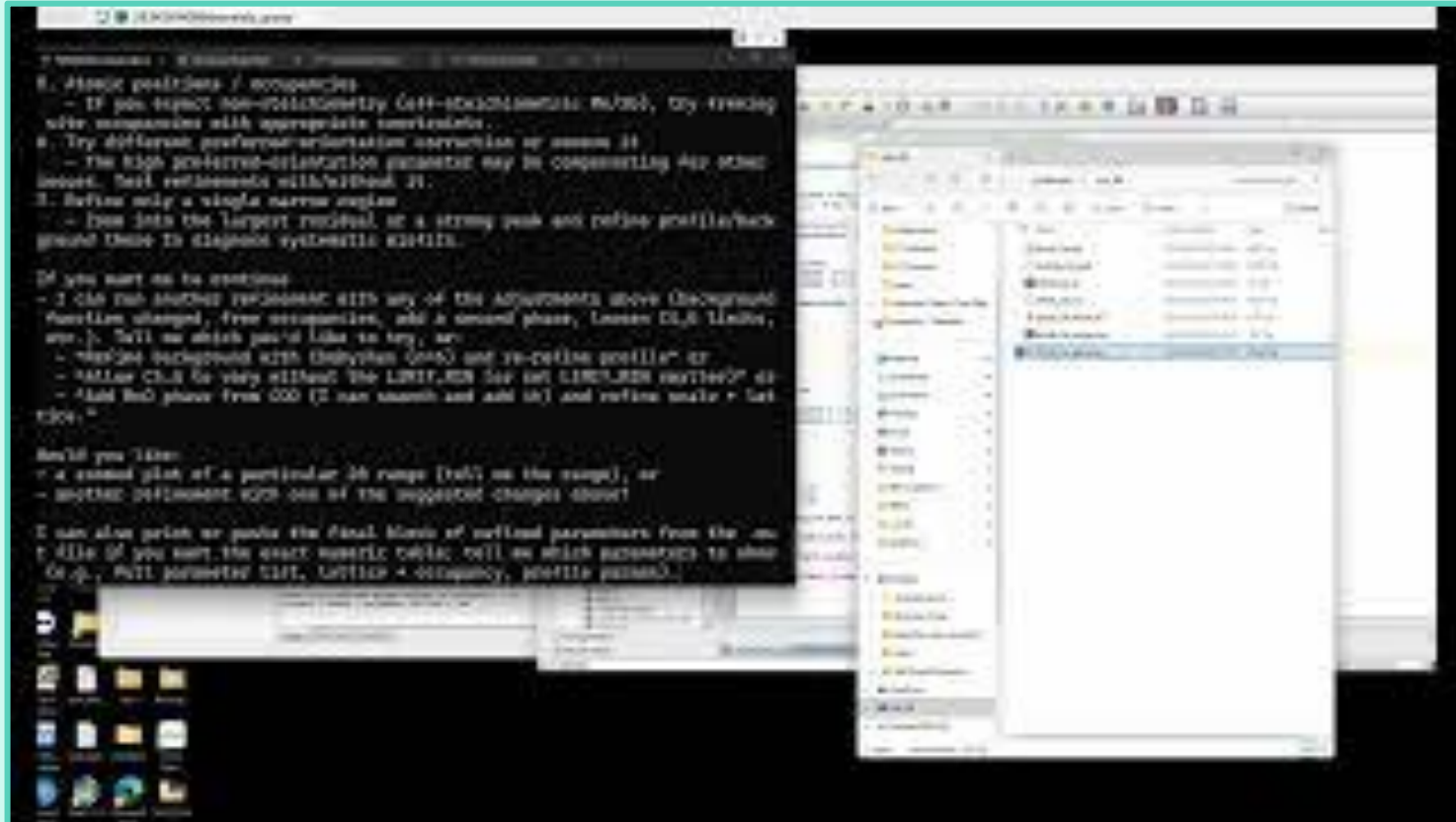
Zimmerman *et al*, *Mach. Learn.: Sci. Tech.*
[10.1088/2632-2153/ae011a](https://doi.org/10.1088/2632-2153/ae011a) (2025)

> guiLLeMot (2025)

Wrapping “legacy” tools use for semi-autonomous analysis tasks, XRD refinement from sample data using *datalab*



datalab-org/guilLeMot



> (AI + dataLab)

- **AI-accelerated digital transformation for labs of all sizes**
 - Onboarding and ingestion of historic data without our help
 - Uses existing models with an annotated API and structured tool-calling

> (AI + datalab)

- **AI-accelerated digital transformation for labs of all sizes**

- Onboarding and ingestion of historic data without our help
- Uses existing models with an annotated API and structured tool-calling

- **Agent-driven data analysis**

- Generic agents that can use datalab tools as part of a broader workflow including e.g., literature data, simulation and other packaged tools
- *datalab* provides the substrate to expose your data for interrogation

> (AI + datalab)

- **AI-accelerated digital transformation for labs of all sizes**

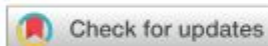
- Onboarding and ingestion of historic data without our help
- Uses existing models with an annotated API and structured tool-calling

- **Agent-driven data analysis**

- Generic agents that can use datalab tools as part of a broader workflow including e.g., literature data, simulation and other packaged tools
- *datalab* provides the substrate to expose your data for interrogation










- **Moonshot: “ambient” science**

- Semi-autonomous agents operating internal but accessible data across the *datalab* federation in an exploratory manner, transparently crediting data providers
- Constructing and disseminating traceable “synthetic” datasets
- Access across other platforms, ELNs, autonomous labs etc.
- Now technically feasible, but limited fit with current scientific problems



Cite this: DOI: 10.1039/d5cs00146c

Cross-disciplinary perspectives on the potential for artificial intelligence across chemistry

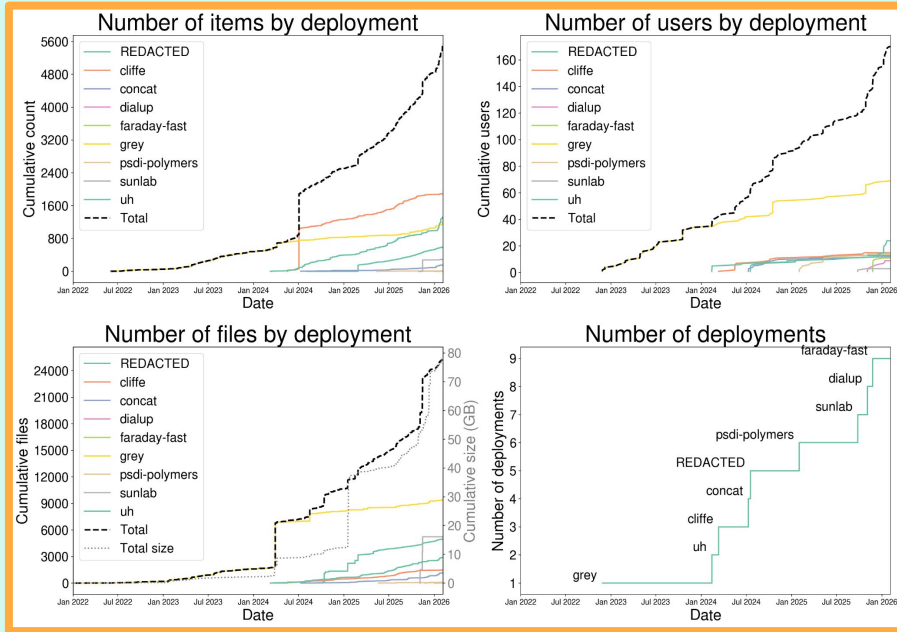
Austin M. Mroz, ^{ab} Annabel R. Basford,^a Friedrich Hastedt, ^c
Isuru Shavindra Jayasekera,^d Irea Mosquera-Lois, ^e Ruby Sedgwick,^f
Pedro J. Ballester, ^g Joshua D. Bocarsly,^h Ehecatl Antonio del Rio Chanona,^c
Matthew L. Evans, ^{ijk} Jarvist M. Frost,^a Alex M. Ganose, ^a
Rebecca L. Greenaway, ^a King Kuok (Mimi) Hii, ^a Yingzhen Li,^f Ruth Misener,^f
Aron Walsh, ^e Dandan Zhang^{dg} and Kim F. Jelfs ^{★a}

Mroz *et al.*, Chem Soc Rev (2025) [10.1039/D5CS00146C](https://doi.org/10.1039/D5CS00146C)

> future

- More AI-driven interfaces for *datalab*
 - **AI4Scientists**
- Drive scientific use cases and close the loop:
 - Robotic labs
 - Automated/offline analysis
- More expressive UIs for collaboration (human and agent)
 - Voice, video for future labs
- *Decentralised serendipity*
 - Grey group (Cambridge)
 - Bocarsly group (UHouston)
 - Cliffe group (Cambridge)
 - Sun lab (Cambridge)
 - ConCat lab (TUBerlin)
 - PSDI Polymers (UK)
 - Centre for Additive Manufacturing (Nottingham)
 - 20+ person R&D startup (USA)
 - 10+ person R&D startup (UK)
 - Cumby group (Edinburgh)
 - Materials Research Lab (UCSB/MPI Stuttgart)
 - ➔ **Faraday Institution FAST Project (UK)**
 - ➔ **DIGIBAT (Imperial)**

> future



- Grey group (Cambridge)
- Bocarsly group (UHouston)
- Cliffe group (Cambridge)
- Sun lab (Cambridge)
- ConCat lab (TUBerlin)
- PSDI Polymers (UK)
- Centre for Additive Manufacturing (Nottingham)
- 20+ person R&D startup (USA)
- 10+ person R&D startup (UK)
- Cumby group (Edinburgh)
- Materials Research Lab (UCSB/MPI Stuttgart)
- **Faraday Institution FAST Project (UK)**
- **DIGIBAT (Imperial)**

We are a company: *datalab* industries ltd. (-> cbc.)

<https://datalab.industries>

- We will keep the servers running
- We will store and backup your research data
- We will provide training materials
- We will help with user administration tasks
 - ◆ Creating user groups and accounts
 - ◆ Enforcing default data permissions
- We create robust deployment rules to make it easier to self-host
- We will continue to tune and improve *datalab* for this specific project, whilst feeding back into the core “product”
- Treat us as a (limited, consumable) resource!



Dr Matthew Evans



Dr Ben Smith

> acknowledgements

<https://ml-evs.science/talks>

datalab team:



Prof Josh Bocarsly
(UHouston)



Dame Prof Clare Grey FRS
(Cambridge)



Dr Ben Smith



Ben Charmes



LEVERHULME
TRUST



datalab
industries
ltd

Current funding:

- Prof Clare Grey & the Grey Group
- Leverhulme Trust, Isaac Newton Trust

Previous funding:

- SPWallonie, Horizon Europe
- Prof Gian-Marco Rignanese (UCLouvain)
- Dr David Waroquiers (Matgenix)

Plus close collaborators:

- Prof Matt Cliffe (Cambridge)
- Prof Shijing Sun (Cambridge)
- Dr Yue Wu
- Dr Peter Kraus (TUBerlin)
- *datalab* federation
- PSDI
- Faraday Institution (CATMAT, FAST)
- DIGIBAT

