

Lecture #3: Data in materials science

Previously on

- What is materials informatics
 - Solving materials science problems with **data** science tools
 - HW0 announcement
- Python for materials modeling
 - Automation vs. Manual
 - Intro to ASE and Pymatgen

Goals

- Understand the role of data in materials science
- Provide sources of data
- Introduce the Materials Project database and its API

Agenda

- Experimental data
- Computational data
- Why use someone else's data?
- FAIR principles
- The Materials Project

What is data?

"Data is a representation of information stored in a systematic way for the purpose of inference, argument or decision making"

From [An Introduction to Data Analysis](#) by Michael Franke

As materials science researchers, we generate a lot of experimental and computational data

2-3D Images:

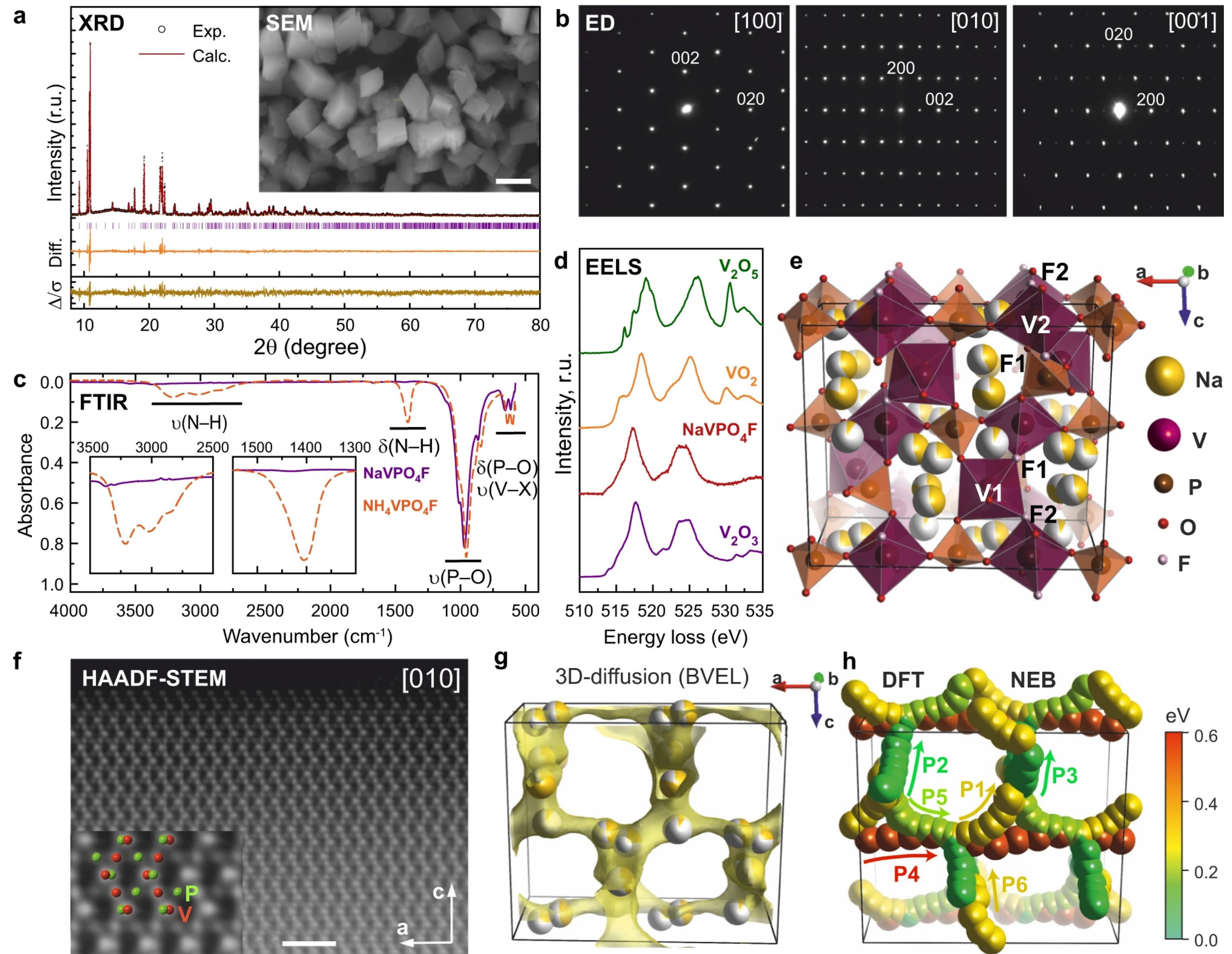
- Electron microscopy images (exp)
- Electron density (exp/comp)

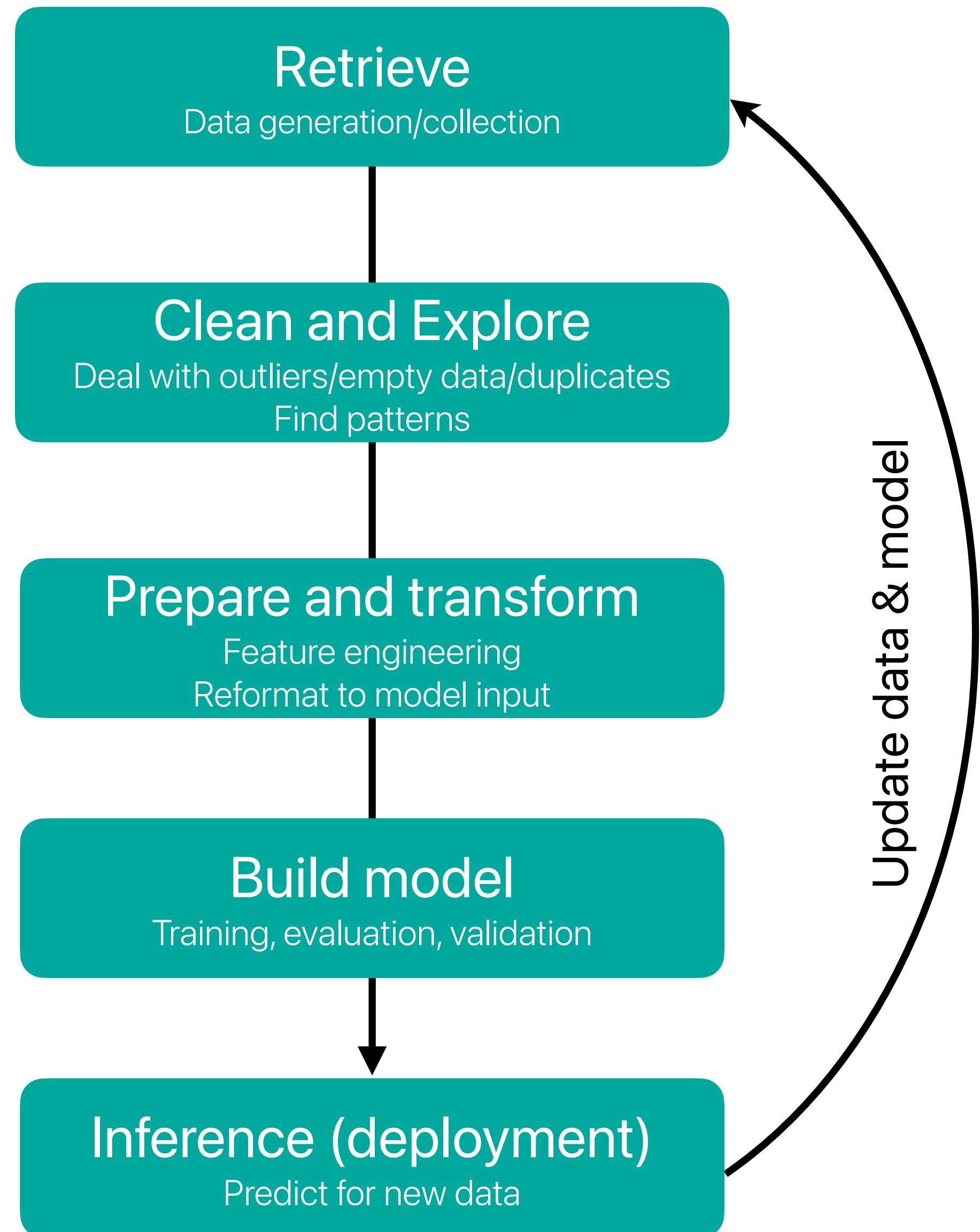
Tabular data

- E_g vs. composition (the last seminar)
- X-ray diffraction pattern (exp/comp)
- Particle size distribution (exp)
- Density of states (comp/exp)
- Energy profile (comp/exp)

Text

- Structure files (exp/comp)





Typical materials informatics workflow

Todays class

Retrieve
Data generation/collection

Where to get data?

On your own:

- Manual collection from your experiments/simulations
- High-throughput experiment or calculations
- Extract from review articles: [WebPlotDigitizer](#)
- Text mining from research articles using LLMs

From available datasets:

- Data purchase
- Open materials databases

There is a huge amount of materials science data

... scattered across various papers, datasets, databases and websites

Github:

- <https://github.com/blaiszik/Materials-Databases>

Curated databases

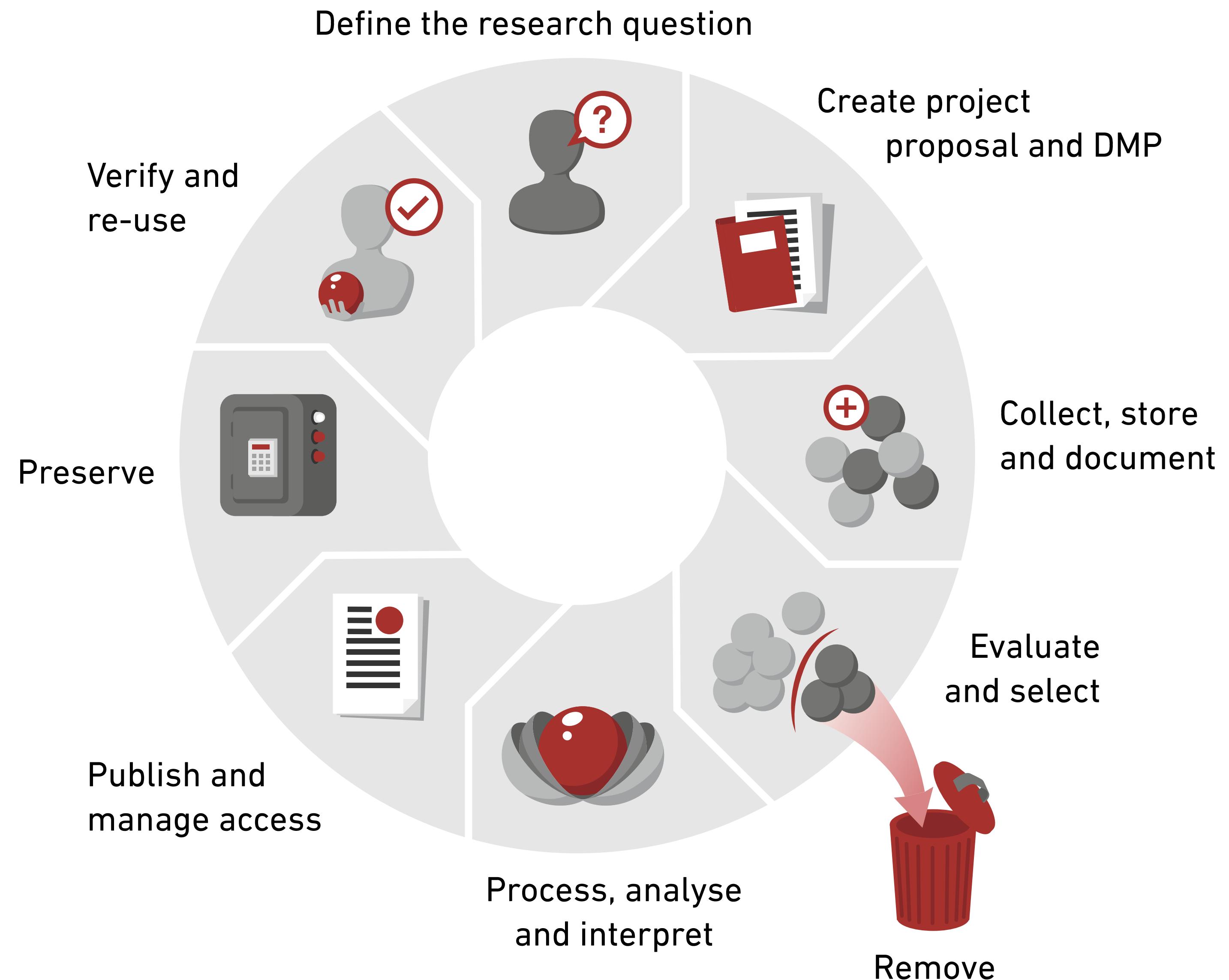
- ICSD: <https://icsd.products.fiz-karlsruhe.de/>
- COD: <https://www.crystallography.net/cod/>
- CCDC: <https://www.ccdc.cam.ac.uk/>
- Materials project: <https://next-gen.materialsproject.org/>
- The Open Quantum Materials Database: <http://oqmd.org/>
- AFLOW: <http://www.aflowlib.org/>
- matbench: <https://matbench.materialsproject.org/>

...and more

Why use someone else's data?

Why use someone else's data?

- guide to your research objective
- reference
- baseline
- insight
- explanation
- enrichment
- time



What do we expect from the data source?

The FAIR Guiding principles

... for scientific **data management and stewardship**

- **F** is about discovery and permanent reference.

It answers "Does this data exist?"



Findable

- **A** is about the retrieval process.

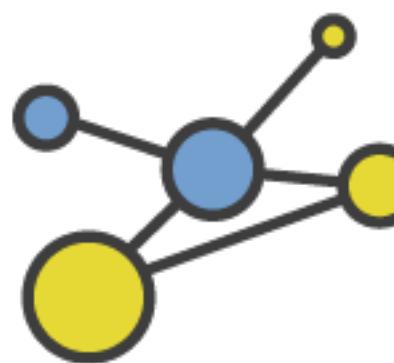
It answers "What are the rules and route for getting the data?"



Accessible

- **I** is about speaking a common language.

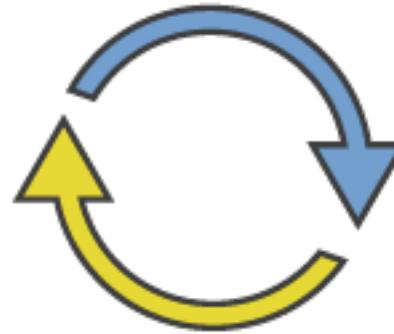
It answers "Can I merge this data with another dataset without a huge manual cleanup effort?"



Interoperable

- **R** is about trust and context for future use.

It answers "Do I have everything I need to use this data correctly and ethically in my own work?"



Reusable

These principles should guide your projects (and future research)

Bad (toy) example - SuperDuperConductorsDB

- a dataset of superconductivity temperatures of 5,000,000 novel stable crystal structures
- with only transition temperatures and chemical compositions shared
- no crystal structures, no methods provided

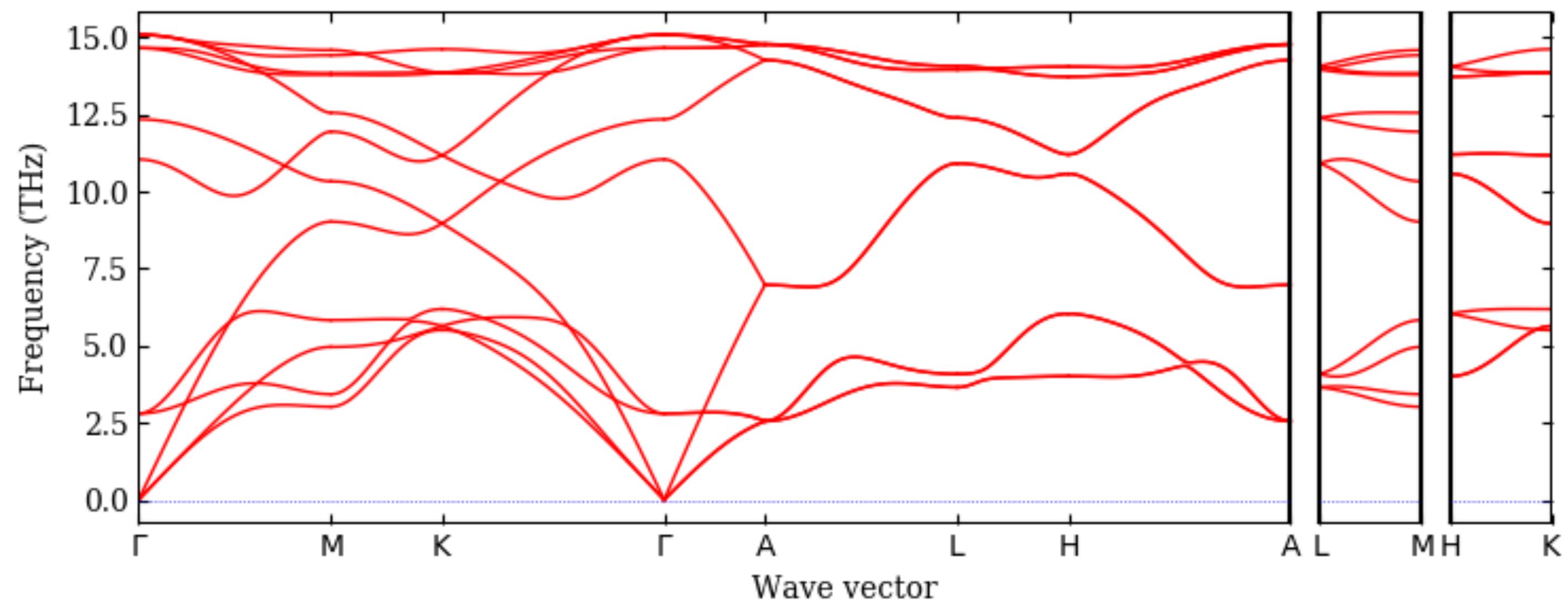
website: <https://idontsharemydata.com/>



Ok example - phonondb

- phonon dispersion curves calculated for ~10,000 crystal structures
- methods, metadata, structure files provided
- hard to handle the data

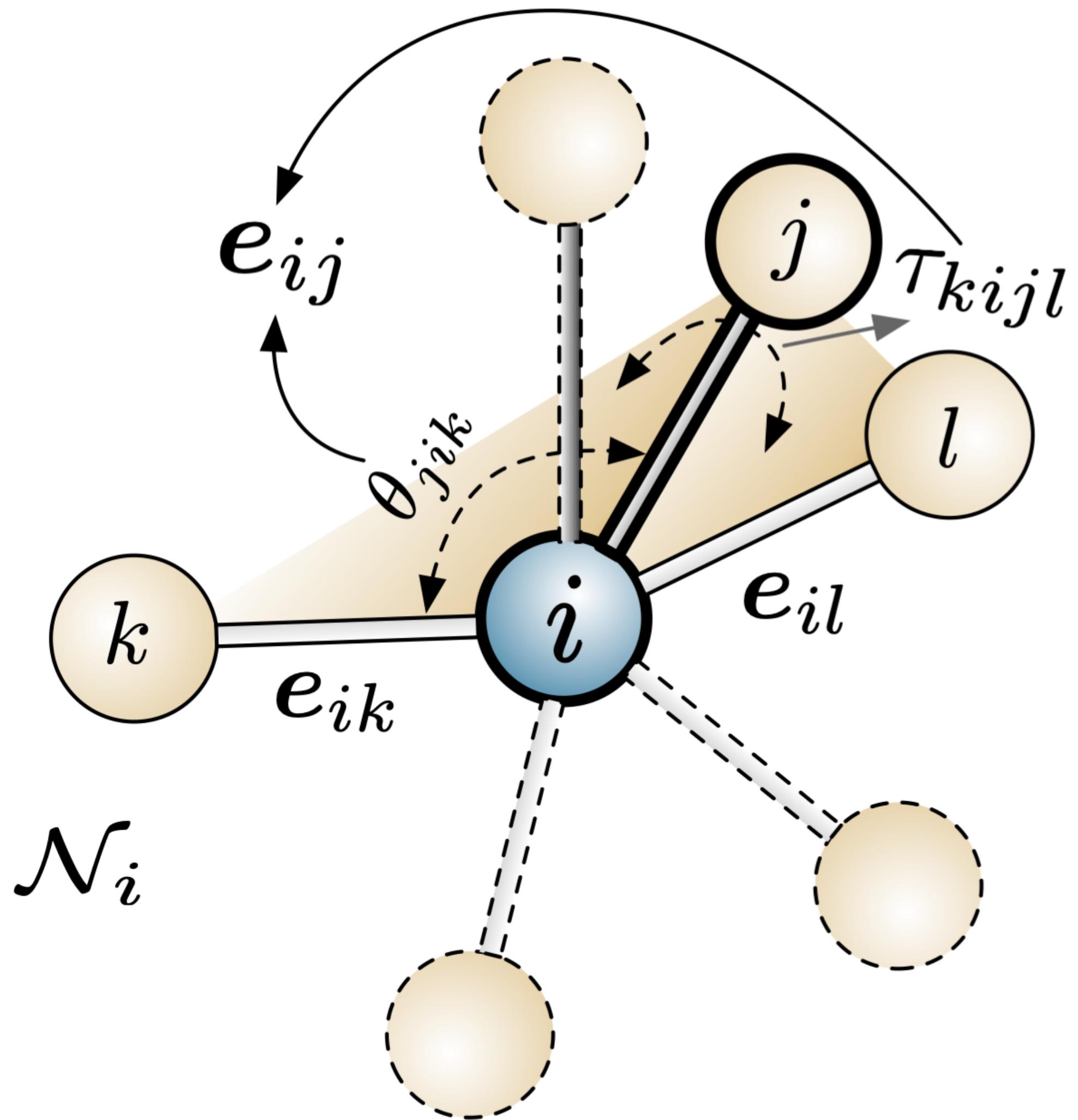
website: <https://github.com/atztogo/phonondb>



Good example - matterverse

- ~30,000,000 crystal structures optimized with deep learning potential
- methods, metadata, structure files provided
- has the platform
- REST API is not that good

website: <https://matterverse.ai/about>



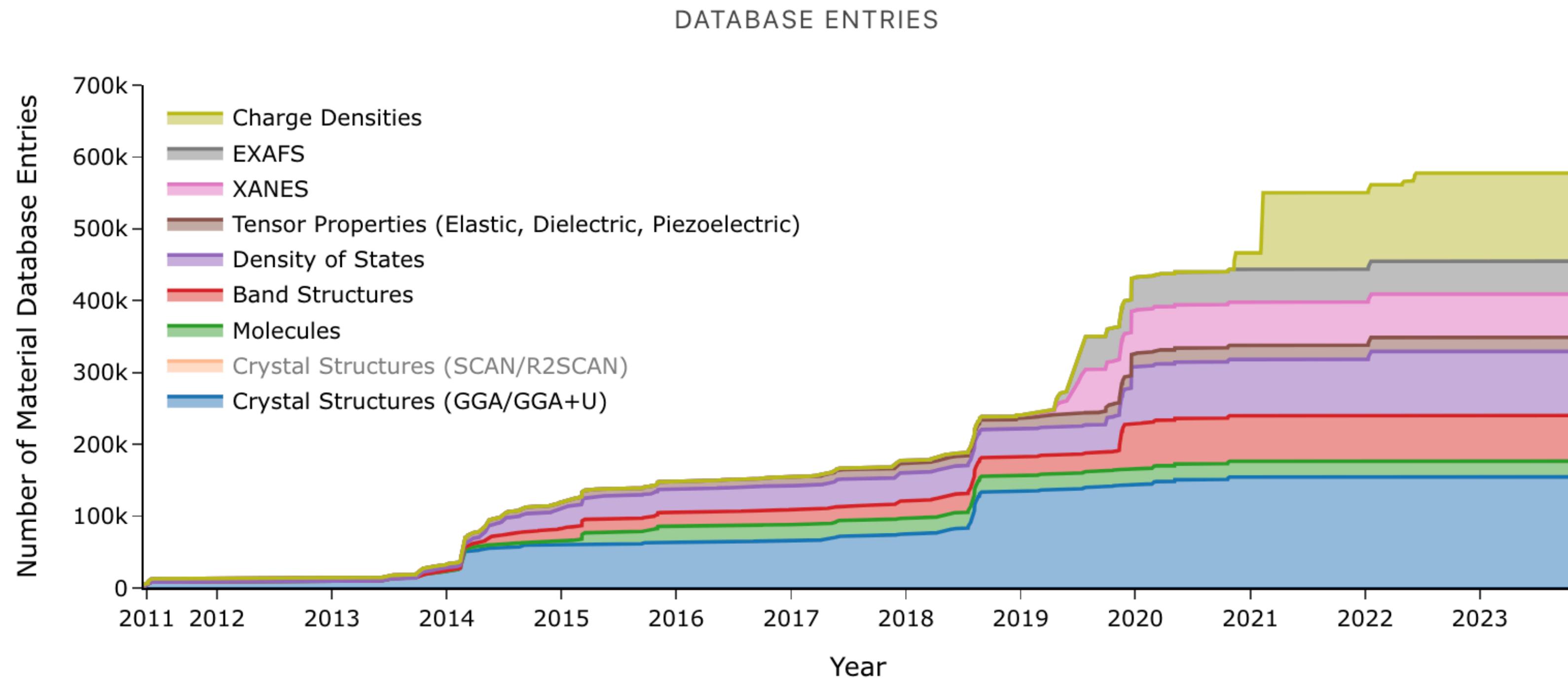
Brilliant example - The Materials Project

(MP) database

- ~200,000 crystal structures
- high quality density functional theory calculations
- “provides one of the largest publicly available data set of computed materials properties”
- methods, metadata, structure files provided
- handy platform
- good REST API

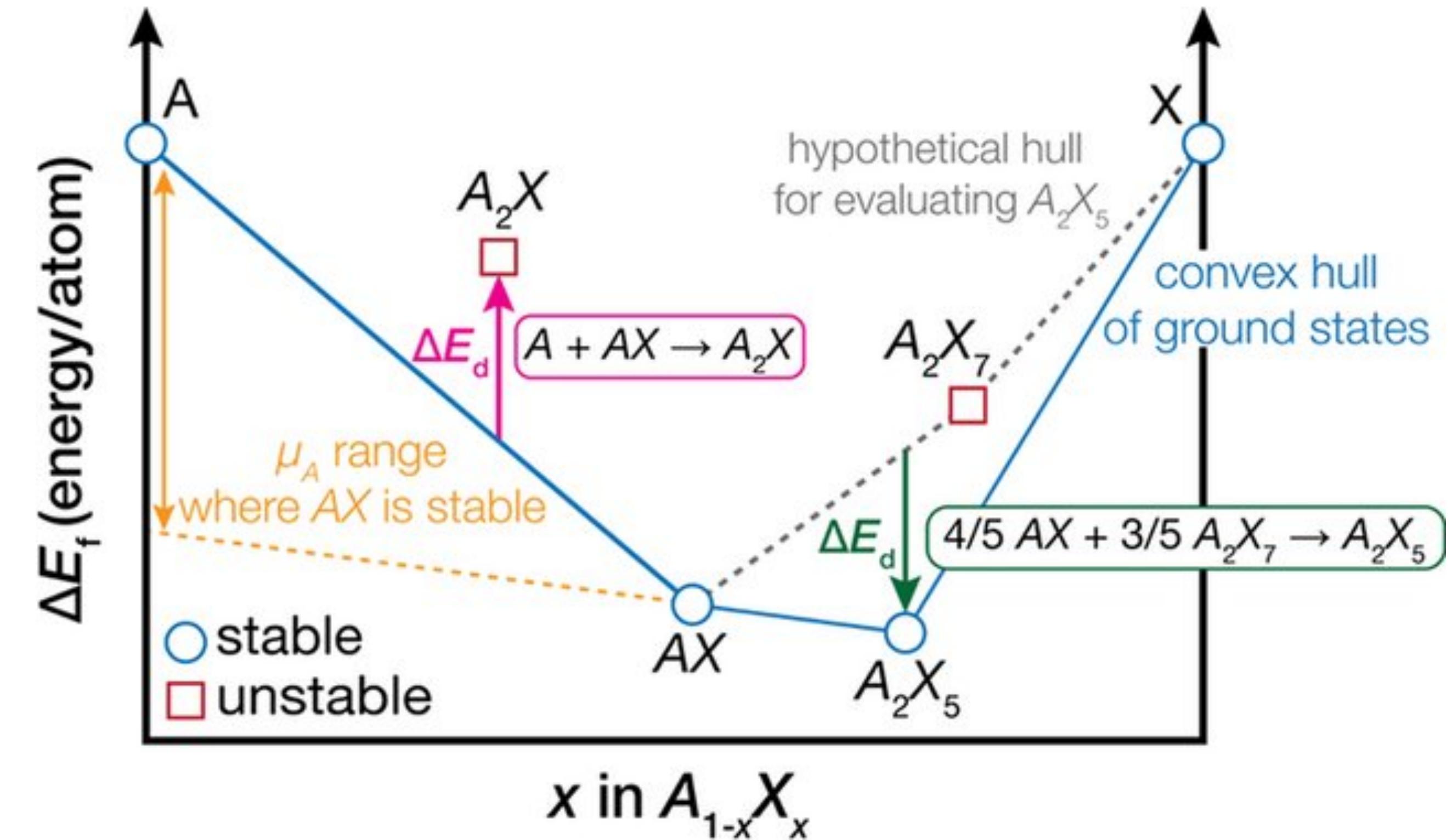
We will learn how to use it during today's seminar

website: <https://next-gen.materialsproject.org/>



What can we do with this data?

- Structure files
 - Input for your calculations
- Calculations results
 - Correlation analysis
 - Comparison
 - Extract more data
 - Fit a surrogate model (seminar #5)
 - Screening
 - Thermodynamics
 - Formation energies
 - Phase diagrams
 - Stability



The MP data access - simple

- you need your API key
- and python

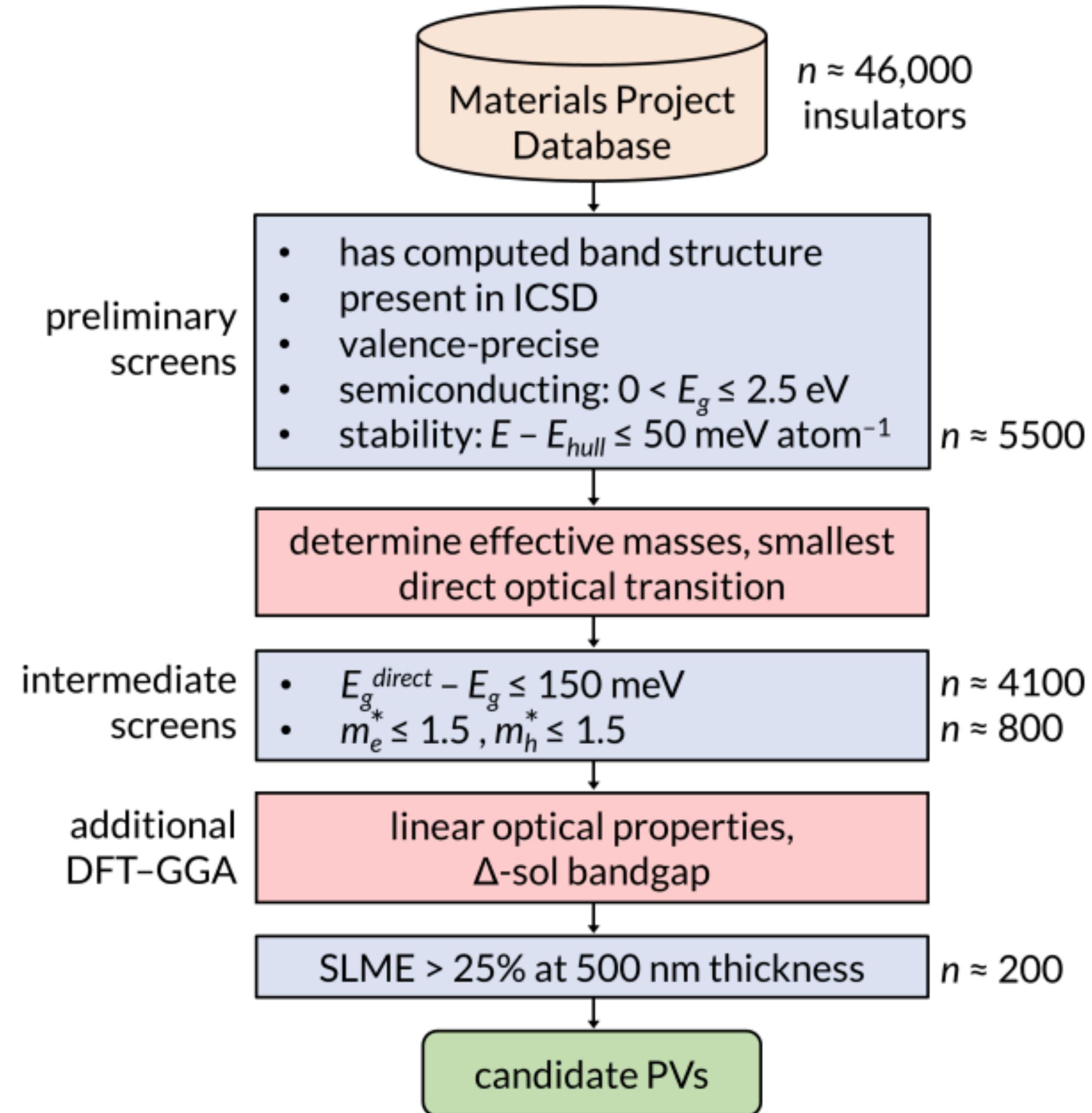
Each material in the database has an identifier (mp_id).

```
from mp_api.client import MPRester

with MPRester(api = "your_key") as mpr:
    # retrieve SummaryDocs for a list of materials
    docs = mpr.summary.search(material_ids=["mp-149", "mp-13"])
```

The MP usage example

- Screened database
- Identified candidates
- Calculated properties of interest for ~800 compounds
- Shared the data



A few words about Density Functional Theory (DFT) before the MP database seminar

- Quantum-Mechanical approach to calculate the electronic structure (ground state electron density) of materials
- The most popular method in materials modelling
- The main (most important) data generator in computational materials science
- Has high predictive power, but:
 - scales as $O(N^3)$, N - number of electrons in a system
 - requires a lot of compute
- Out of the scope of this course
- Consider Computational Chemistry and Materials Modeling course (MA060008) for learning fundamentals of DFT

Take home message

- There are many (open) sources of materials science related data
- Using someone else's data can help guide or improve your research
- Consider the FAIR principles for sharing your own data

Resources

<https://github.com/sedaoturak/data-resources-for-materials-science?tab=readme-ov-file>

https://github.com/sp8rks/MaterialsInformatics/blob/main/course_notes/5.%20Materials%20Data%20Repositories.pdf

<https://matbench.materialsproject.org/>