Data science for Materials Science & Engineering

Data Querying, Organization, Filtering

In this module

• Introduction to data repositories and APIs (this lecture)
• Hands on tutorial using nanoHUB: data querying for oxides
• Hands on tutorial using nanoHUB: performing your own tailored queries and data filtering
• Homework assignments

Zachary McClure and Alejandro Strachan
zmcclure@purdue.edu || strachan@purdue.edu
School of Materials Engineering & Birck Nanotechnology Center
Purdue University
West Lafayette, Indiana USA
Learning objectives and prerequisites

After completing this lecture you will:

• Learn about and querying and data repositories
• Manage data through efficient Pandas dataframes
  • Parsing
  • Logic operations
  • Data visualization and selection
  • Type management
• Perform tailored queries specific to user application

Pre-requisites:
• Basic Python programming
Data Science & Machine Learning in Science & Engineering

Acquiring and handling data

Finding patterns (unsupervised learning)

Learning from data

Predictive models (supervised learning)

Design of experiments

Cyber-infrastructure

**Acquiring and handling data**

**Finding patterns (unsupervised learning)**

**Learning from data**

**Predictive models (supervised learning)**

**Design of experiments**
Data acquisition

How do we get data?
• Individual researcher
• Collaboration
• Extracting results from publications

Can we accelerate innovation by publishing and consuming data?
• Open repositories with community results
• Researchers can share all their results
  • Not just those that make it to a publication
• Cyberinfrastructure to query and analyze large datasets
• Use machine learning to extract knowledge from community results
Data acquisition

How to interact with data repositories?
Option 1: Web front ends
Web front end example

Example case: searching for MgO

- Pros: easy to use
- Cons: labor intensive if you need lots of data
Application programming interfaces (APIs)

Databases

API

Web Application

Advanced User

Beginner User

CITRINE INFORMATICS

WolframAlpha

OQMD
The Open Quantum Materials Database

The Materials Project
materialsproject.org

FACE CAMERA
API example

Example case: searching for MgO
API example

Example case: searching for MgO

Query all oxides in MP database
### Storing and managing data

- **CSV files** – simple, good for storage
- **Numpy arrays** – good for math operations
- **MATLAB datafiles**
- **Pandas**
  - Sorting, parsing, editing, and storage
  - Think of it as excel in python

---

<table>
<thead>
<tr>
<th>Elements</th>
<th>Formula Dict</th>
<th>Volume</th>
<th>IPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>[M, B, C]</td>
<td>(\text{V}^1, 1.5, \text{B}^1, 2.0, \text{C}^3, 6.0)</td>
<td>94.04/1024</td>
<td>0.552616</td>
</tr>
<tr>
<td>[Na, Y, O]</td>
<td>(\text{Na}^+, 1.0, \text{Y}^1, 1.0, \text{O}^2- 4.0)</td>
<td>124.03/32</td>
<td>0.485259</td>
</tr>
<tr>
<td>[Ca, Al, C]</td>
<td>(\text{Ca}^{2+}, 1.0, \text{Al}^3+, 2.0, \text{C}^6, 8.0)</td>
<td>392.12/547</td>
<td>0.307014</td>
</tr>
<tr>
<td>[Er, P, C]</td>
<td>(\text{Er}^3+, 1.0, \text{P}^5+, 2.0, \text{C}^8, 8.0)</td>
<td>143.77/2253</td>
<td>0.539514</td>
</tr>
<tr>
<td>[La, Si, C]</td>
<td>(\text{La}^{3+}, 1.0, \text{Si}^4+, 2.0, \text{C}^8, 5.0)</td>
<td>479.32/3535</td>
<td>0.407970</td>
</tr>
<tr>
<td>[Ti, B, Br, C]</td>
<td>(\text{Ti}^{4+}, 1.0, \text{B}^3+, 4.0, \text{Br}^-, 1.0, \text{C}^8, 4.0)</td>
<td>140.13/6334</td>
<td>0.356335</td>
</tr>
<tr>
<td>[Sc, Si, C]</td>
<td>(\text{Sc}^{3+}, 1.0, \text{Si}^4+, 2.0, \text{C}^8, 2.0)</td>
<td>137.62/592</td>
<td>0.394036</td>
</tr>
<tr>
<td>[Cr, Dy, O]</td>
<td>(\text{Cr}^{3+}, 1.0, \text{Dy}^{3+}, 2.0, \text{O}^2- 4.0)</td>
<td>151.01/5774</td>
<td>0.630675</td>
</tr>
</tbody>
</table>

---

https://pandas.pydata.org/
Pandas – operation examples

- Dictionary to dataframe
- Removing data based on logic operations
- Removing data without entries
- Plotting based on logic operations
- Appending existing data
Summary

- Easily queryable databases offer researches access to data that can be useful for initial material searches, or for creating rapid sets of descriptors for machine learning models.

- Management of these queried datasets can be easily done using pandas dataframes.

- The code in https://nanohub.org/tools/matdatarepo can be easily modified to meet needs in a wide range of problems.
Next steps

**Hands-on tutorial**: run the [https://nanohub.org/tools/matdatarepo](https://nanohub.org/tools/matdatarepo) example and learn how to query for data and manage it with Pandas

**Homework assignment**: to reinforce concepts and help students modify the workflow and adapt it for their needs