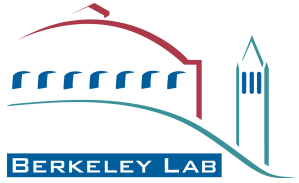
Community Accessible Datastore of High-Throughput Calculations: Experiences from the Materials Project



Dan Gunter, Shreyas Cholia, Anubhav Jain, Michael Kocher, Kristin Persson, Lavanya Ramakrishnan

Shyue Ping Ong, Gerbrand Ceder

BACKGROUND

Our energy future relies on the rapid development of novel functional materials.

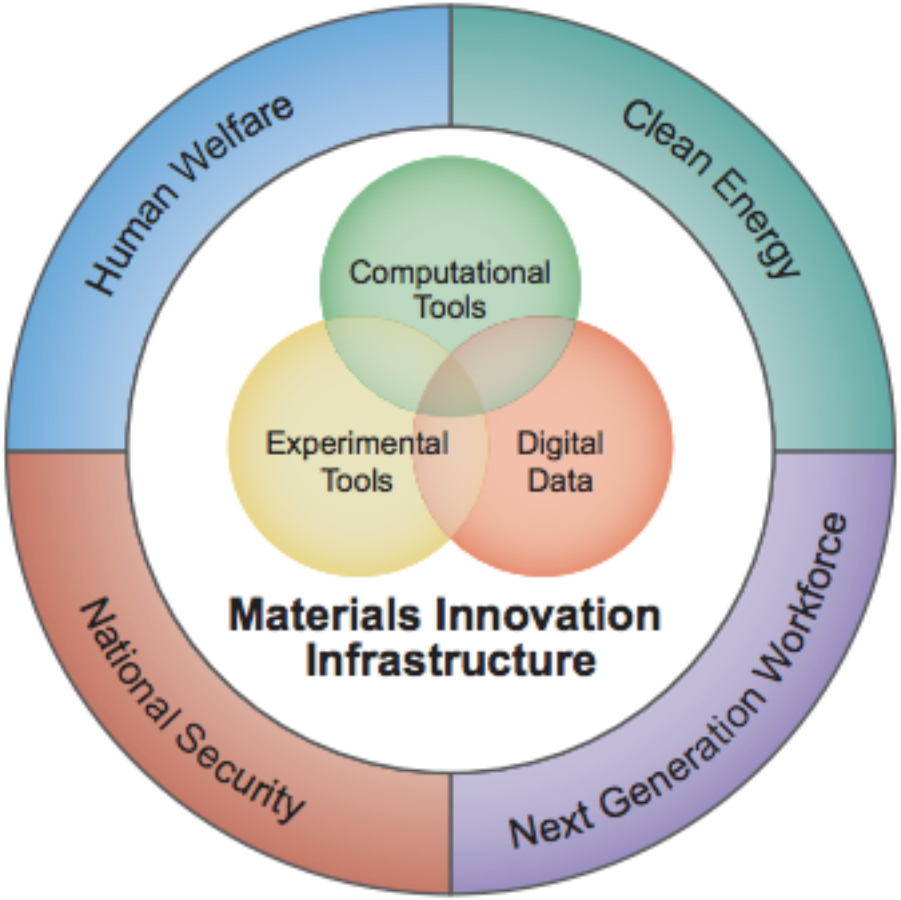
Solar cells, advanced batteries, TCOs, and fuel cells will all play a role in our energy future.

But it takes almost twenty years to develop new materials. **How can we do it *faster*?**

Materials Genome Initiative

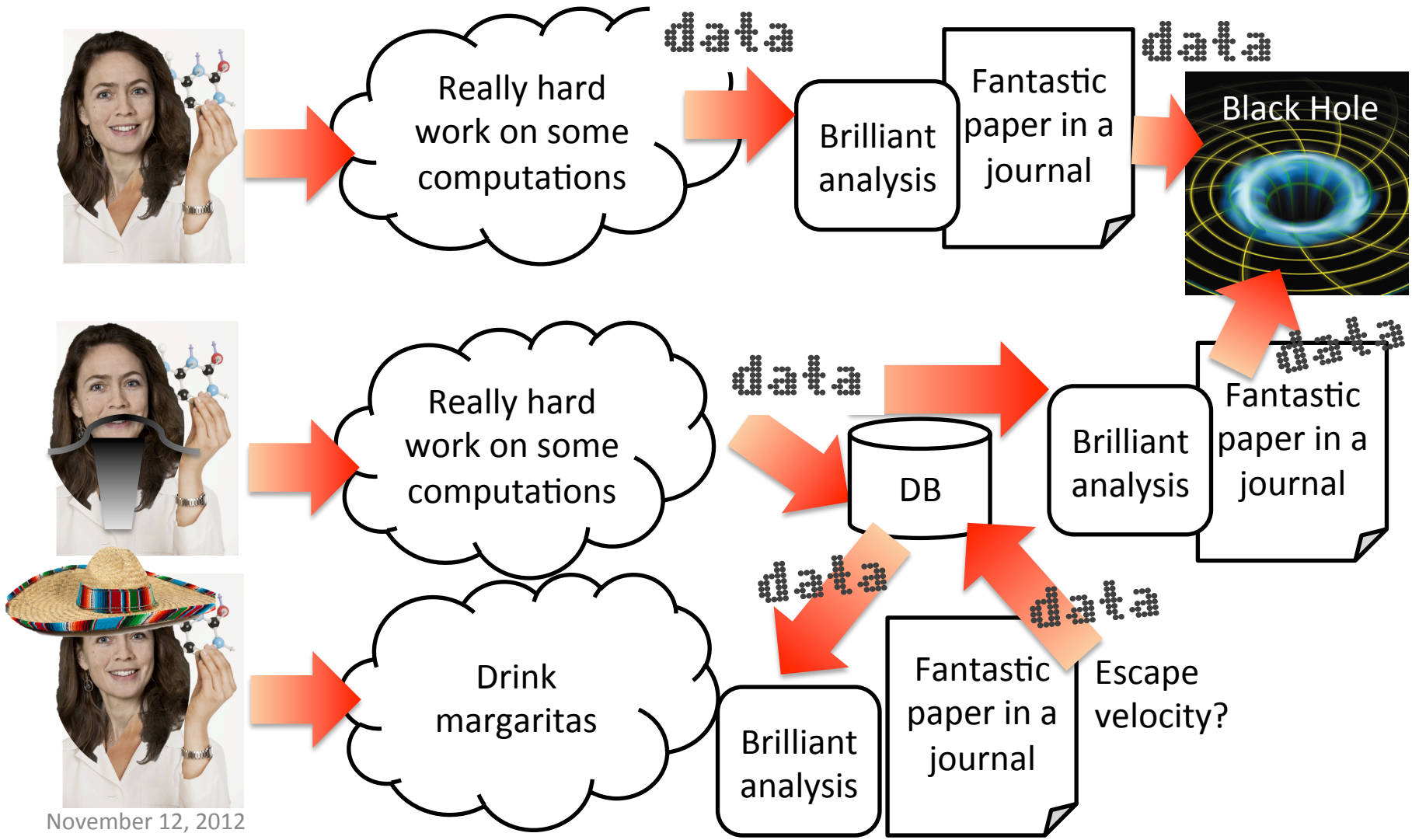
Materials Genome Initiative: A Renaissance of American Manufacturing

June 2011: **Materials Genome Initiative** which aims to "*fund **computational tools**, **software, new methods** for material characterization, and the development of open standards and databases that will make the process of **discovery and development of advanced materials faster, less expensive, and more predictable***"

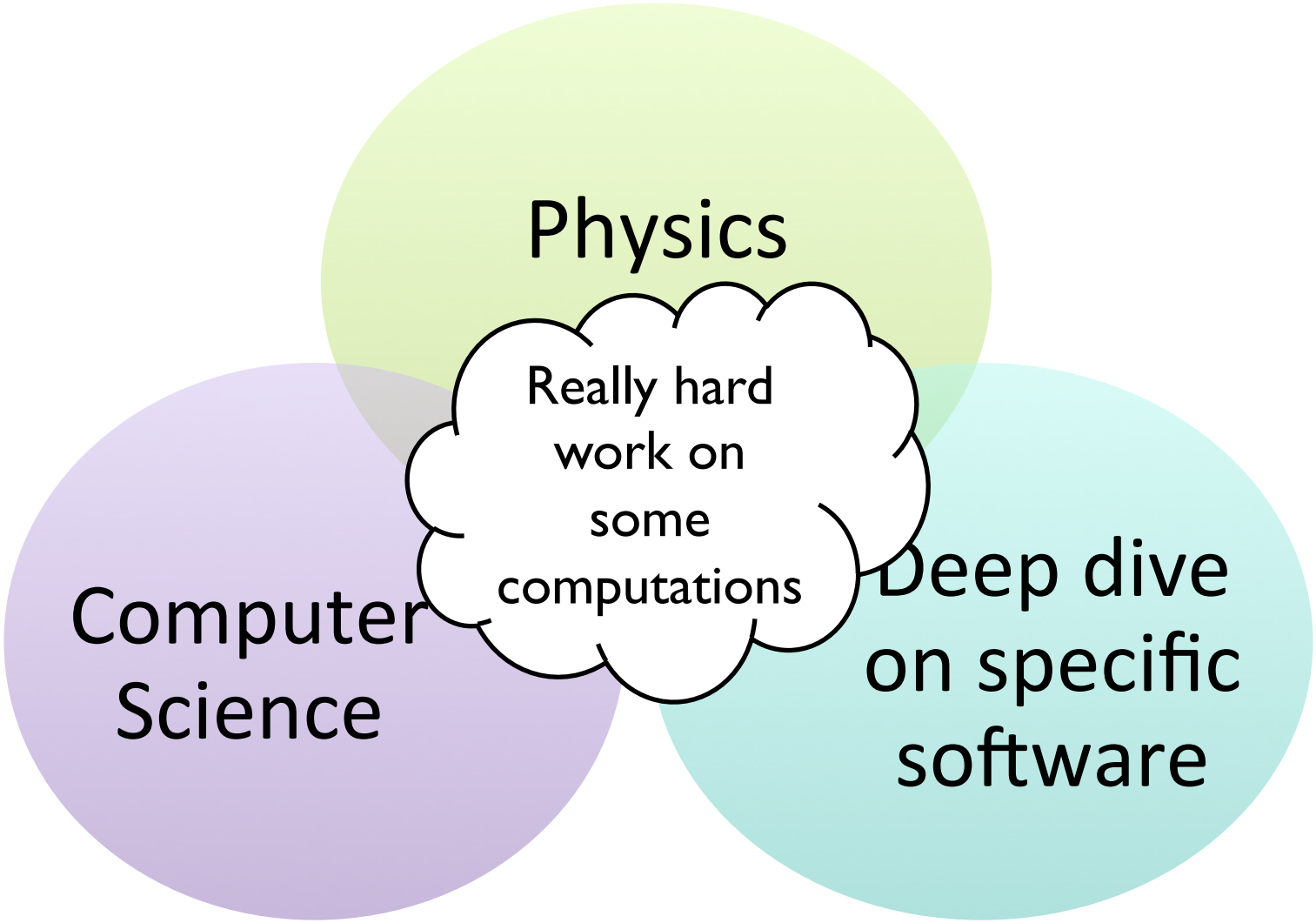


Source: "Materials Genome Initiative for Global Competitiveness"
http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf

It's the data , stupid!



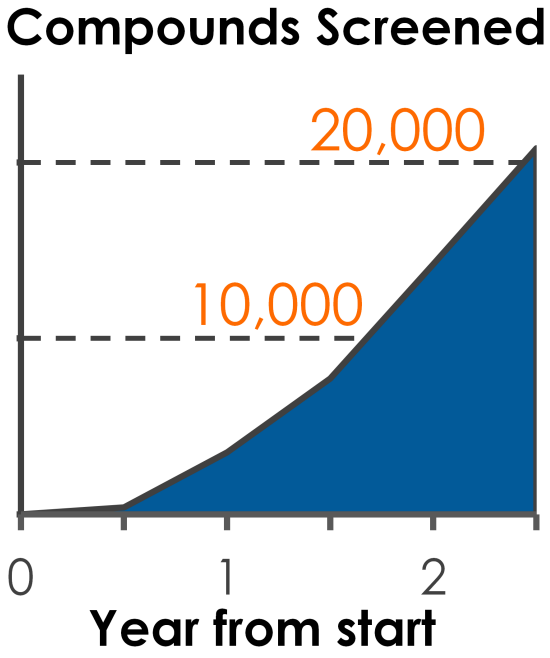
Very specialized skill-set

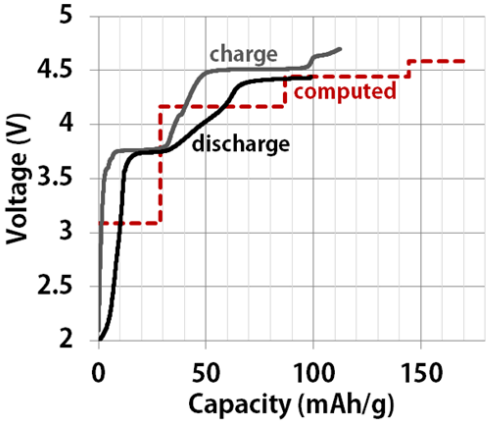


Example

The Materials Project used quantum chemistry calculations to screen over 20,000 materials as potential cathodes for Li ion batteries.

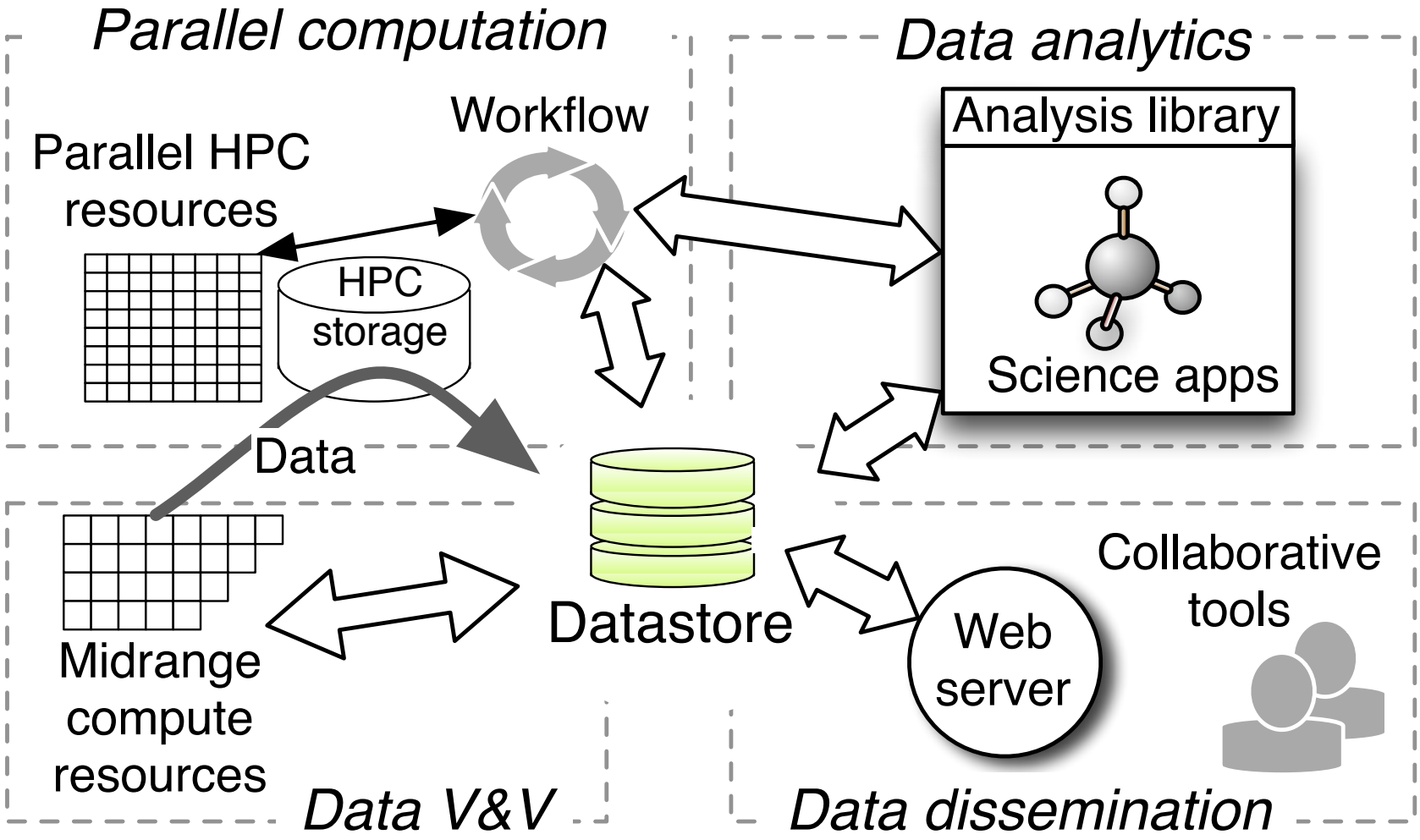
From the results, three new materials were identified, tested, and currently have patents pending.



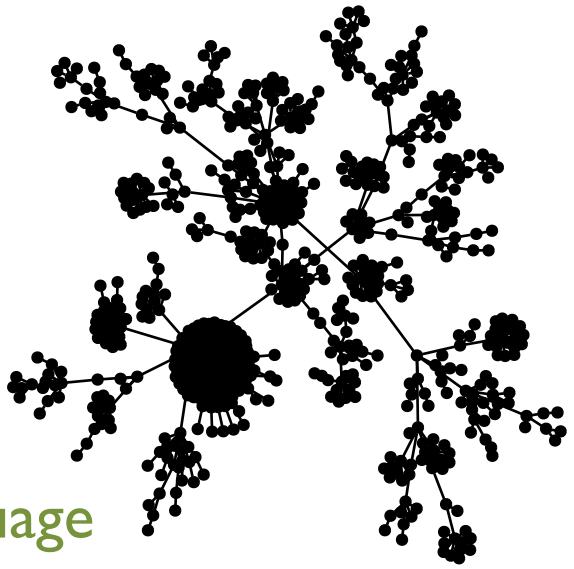


Predicted and measured performance of of $Li_9V_3(P_2O_7)_3(PO_4)_2$ during cell cycling.

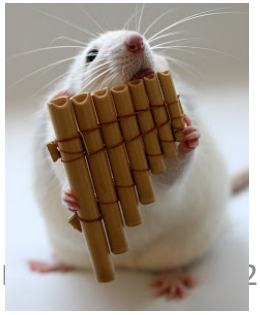
COMPONENTS


Parallel computation
Data analytics
Workflow
Parallel HPC resources
Analysis library
HPC storage
Science apps
Data
Datastore
Collaborative tools
Web server
Midrange compute resources
Data V&V
Data dissemination

NoSQL Datastore

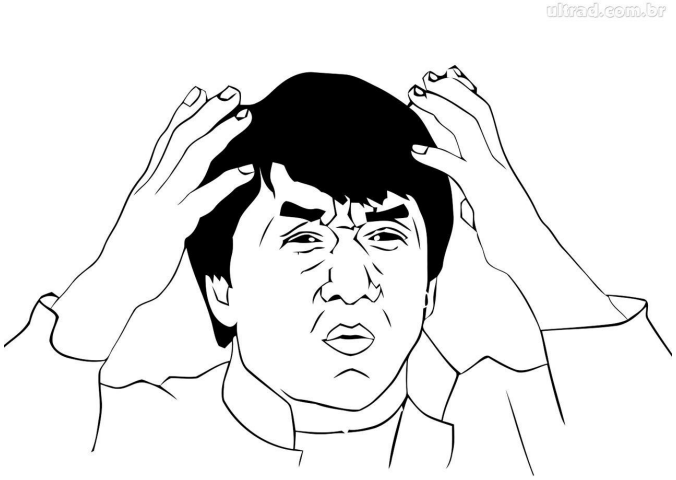


Powerful but simple query language
Ease of administration
Good performance on read-heavy workloads where most of the data can fit into memory.



Poor performance at huge scale
Bad for write-heavy workloads

FireWorks workflow engine



Why?!

Programmability. Scripting, not GUIs and DSL's.

Administration overhead. No extra servers.

Flexibility. DB support, reconfiguring running workflows.

Re-runs

Detours

Duplicates

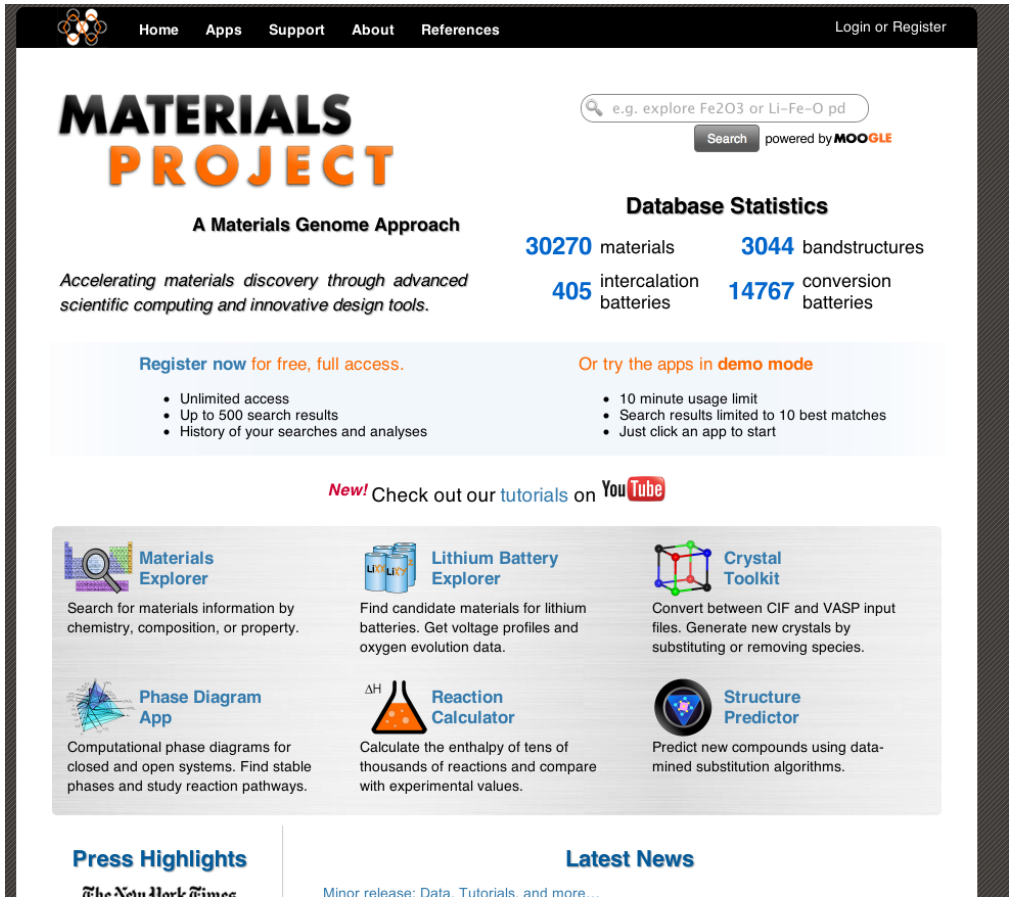
Iteration

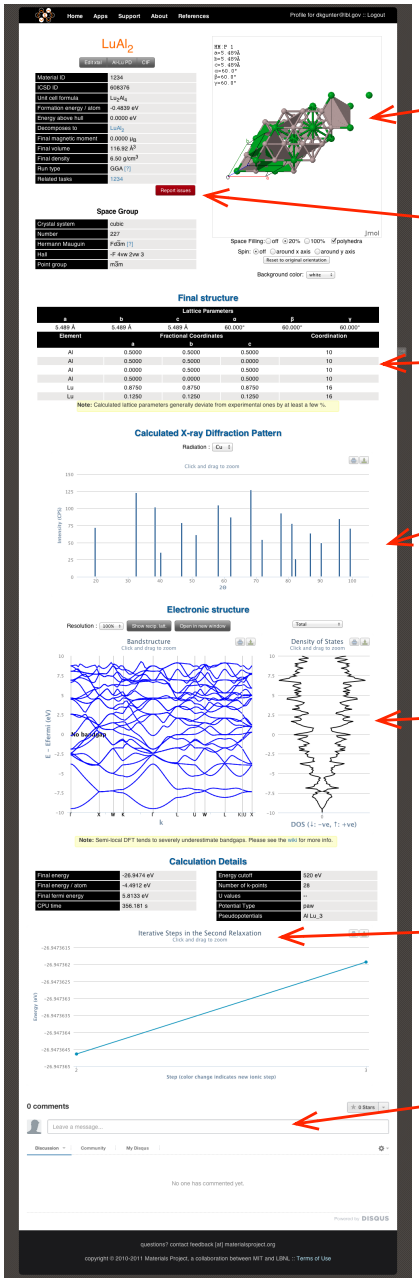
Dissemination with REST

Preamble *Version* *Application I.D.* *Datatype* *Property*

https://www.materialsproject.org/rest/v1/materials/Fe2O3/vasp/energy

Web UI


Home
Apps
Support
About
References
Login or Register
MATERIALS PROJECT
A Materials Genome Approach
Accelerating materials discovery through advanced scientific computing and innovative design tools.
e.g. explore Fe2O3 or Li-Fe-O pd
Search
powered by MOOGLE
Database Statistics
30270 materials
3044 bandstructures
405 intercalation batteries
14767 conversion batteries
Register now for free, full access.
Unlimited access
Up to 500 search results
History of your searches and analyses
Or try the apps in demo mode
10 minute usage limit
Search results limited to 10 best matches
Just click an app to start
New! Check out our tutorials on YouTube
Materials Explorer
Search for materials information by chemistry, composition, or property.
Lithium Battery Explorer
Find candidate materials for lithium batteries. Get voltage profiles and oxygen evolution data.
Crystal Toolkit
Convert between CIF and VASP input files. Generate new crystals by substituting or removing species.
Phase Diagram App
Computational phase diagrams for closed and open systems. Find stable phases and study reaction pathways.
Reaction Calculator
Calculate the enthalpy of tens of thousands of reactions and compare with experimental values.
Structure Predictor
Predict new compounds using data-mined substitution algorithms.
Press Highlights
Latest News
Minor release: Data, Tutorials, and more...


3-D model of unit cell
Disqus comment button
Detailed structure
X-ray diffraction pattern (interactive)
Bandstructure and Density of states (interactive)
Calculation iterations
Comments

WE'RE DOING IT WRONG?

Running on HPC

- Batch queues and large numbers of jobs with unpredictable runtimes
- Talking to the database

Data analytics

- Scaling community contributions to code
- Scaling analytic functions

Data V&V

- Loading new data into a production resource
- Constant validation and verification

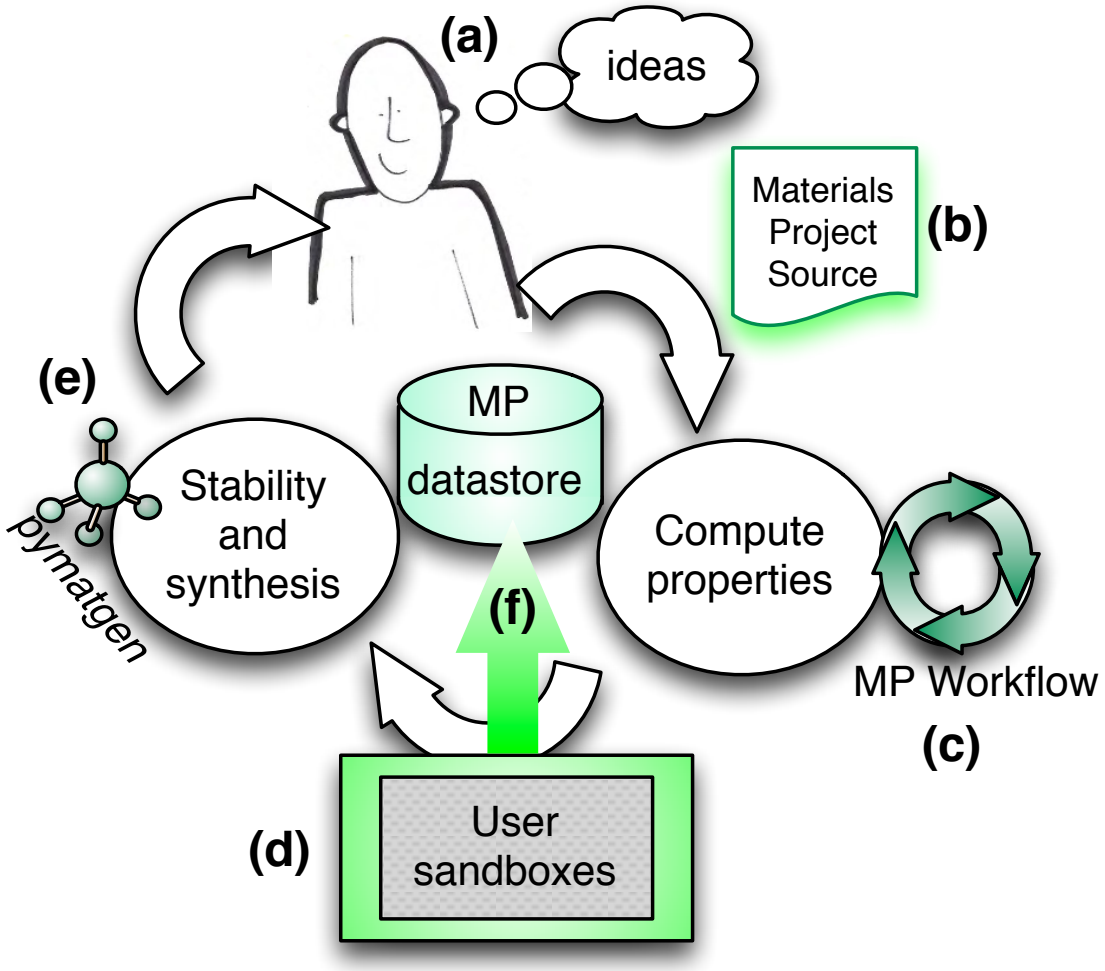
Data dissemination

- Security and privacy
- Query performance

FUTURE WORK

Opening up data access

Towards materials ***design***



Questions?