Towards the autonomation of experiments: quasi-real time analysis of heterogeneous streaming data from scientific experiments

S.J.L. Billinge

Columbia University,

Brookhaven National Laboratory

Columbia University in the City of New York

Brookhaven National Laboratory

HTTP://thebillingegroup.com

High level goal

To discover/design better materials

- For sustainable energy
- For health
- For environmental remediation

Scientific Challenge

1. Predict materials with better properties
2. Make them

Scientific Challenge

1. Predict materials with better properties
   1. Use quantum mechanical calculations (density functional theory – DFT) ✅
   2. Speed it up with AI, ML and data-mining ✅
2. Make them
   1. Predictive synthesis

"Given this recipe, what dish do I get?" -> "Given this dish, what recipe will make it?"

– No good solution for materials which have never been made before (Predicted materials!)

Proposed approach

- High throughput, automated and autonomous *in situ* experiments

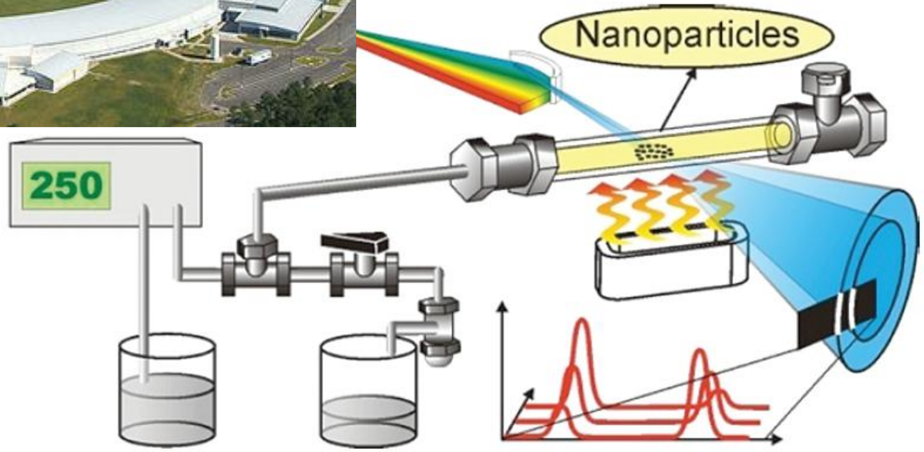
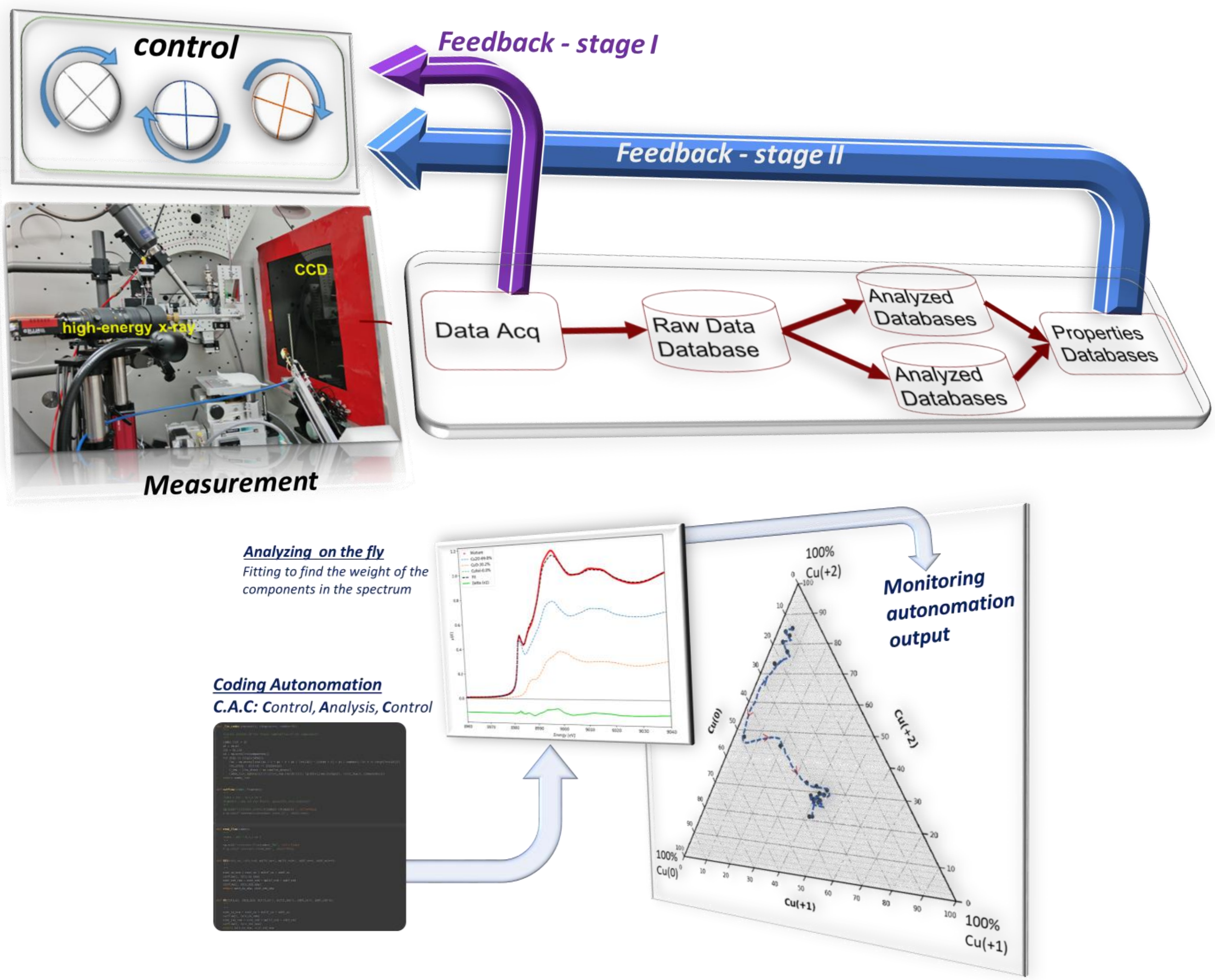


Image credit
Christoffer Tyrsted

Adaptive feedback control: autonomation



COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

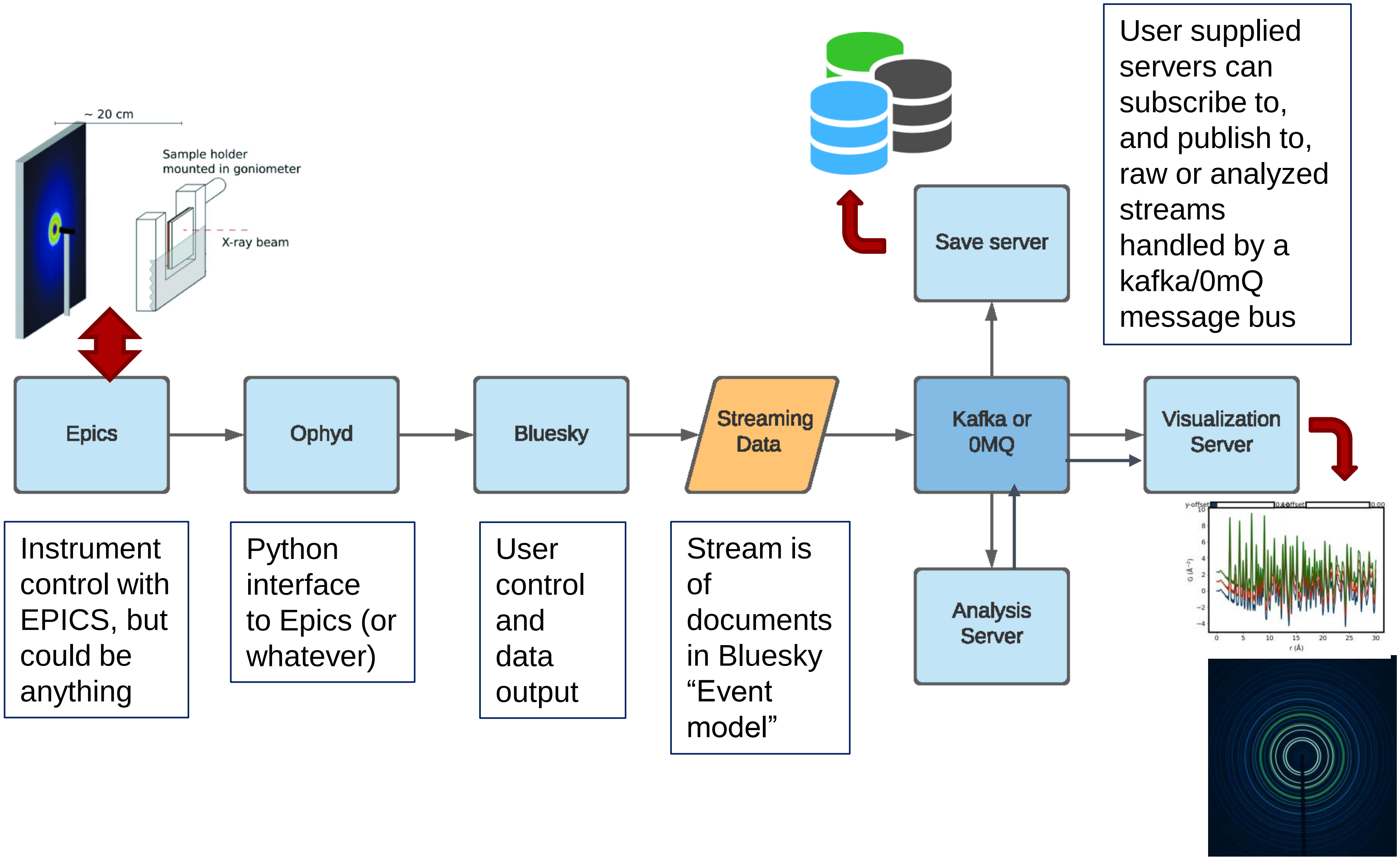
BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com

Computational challenges

1. Instrument control
2. In line rapid data analysis
3. Feedback
    1. Flexible to accommodate different decision policies and experiment designs

Current Stack

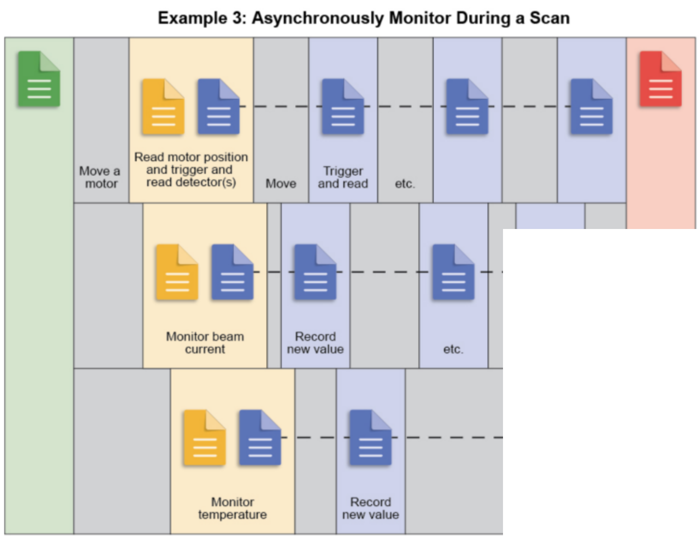


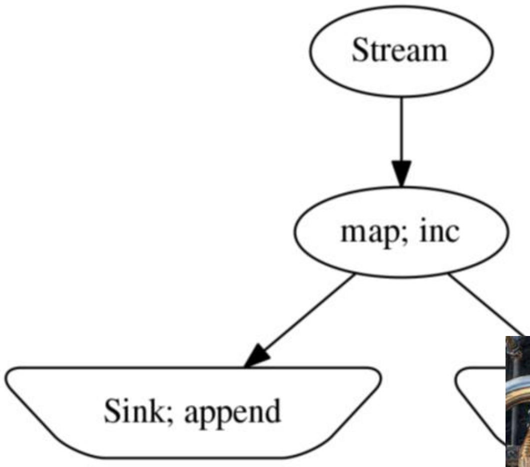
Columbia University in the City of New York

Brookhaven National Laboratory

HTTP://thebillingegroup.com

Real time streaming data analysis

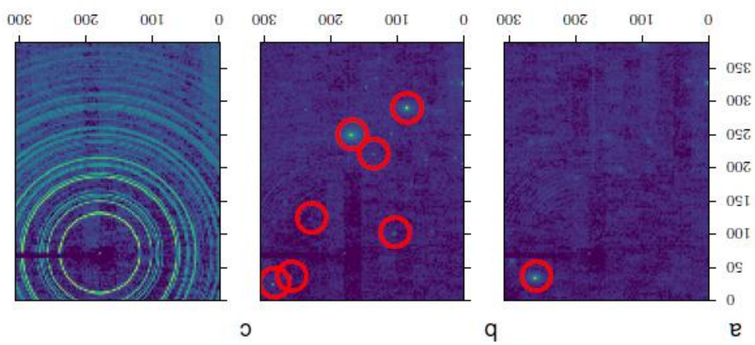


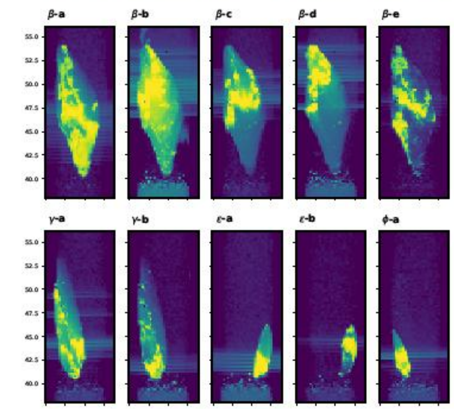


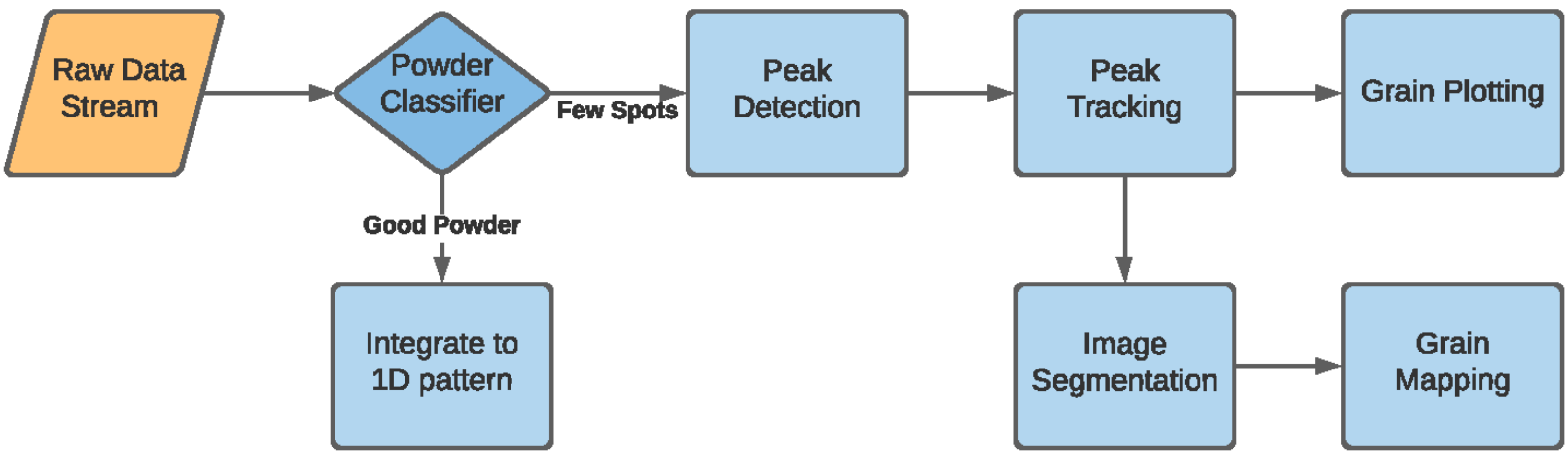
Columbia University in the City of New York

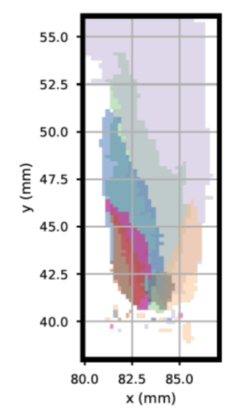
Brookhaven National Laboratory

Current Stack: Crystal mapping, dark field microscopy









COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

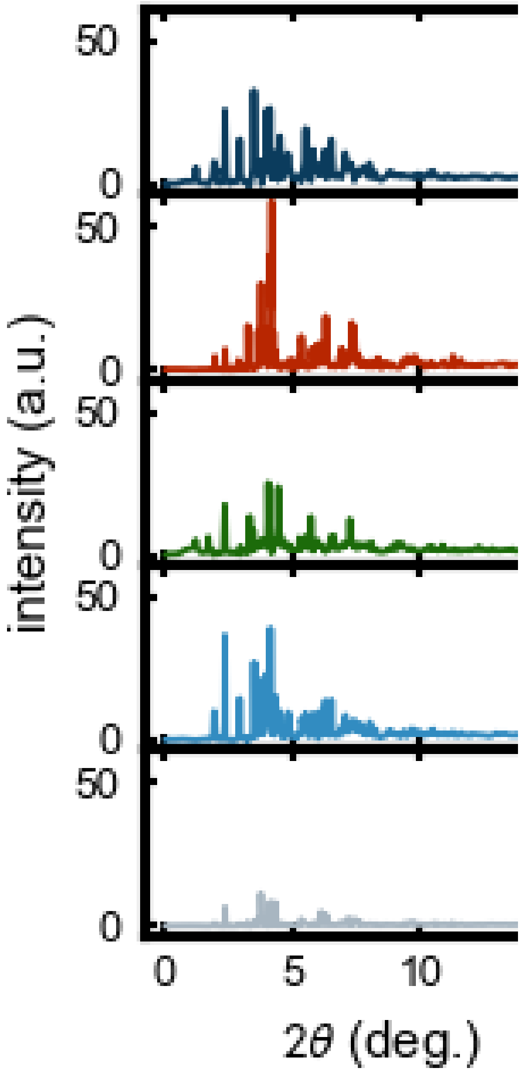
HTTP://thebillingegroup.com

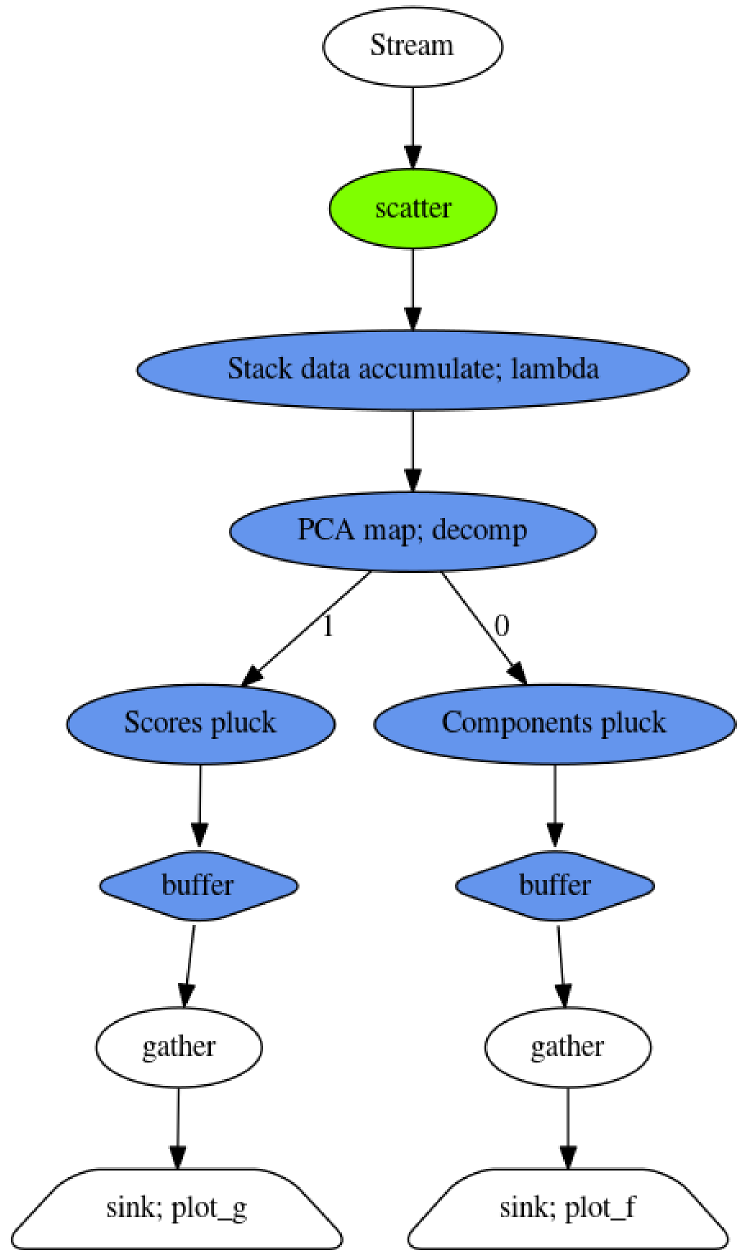
Streaming data analysis

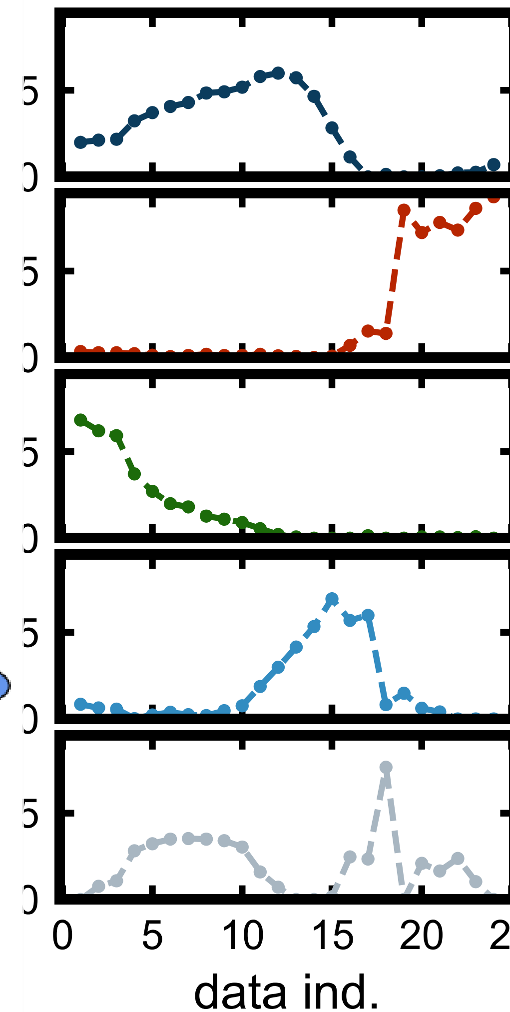
Python based

- Bluesky event model
- Streamz (native python in a stream)
- SHED (Streaming heterogeneous event data) (Bluesky events in a stream)
  - Also builds in provenance and replay
- Visualization using, for example, Matplotlib
- Raw and analyzed metadata in a MongoDB using Intake for cataloging
- Various object store options for the (large) raw data files themselves

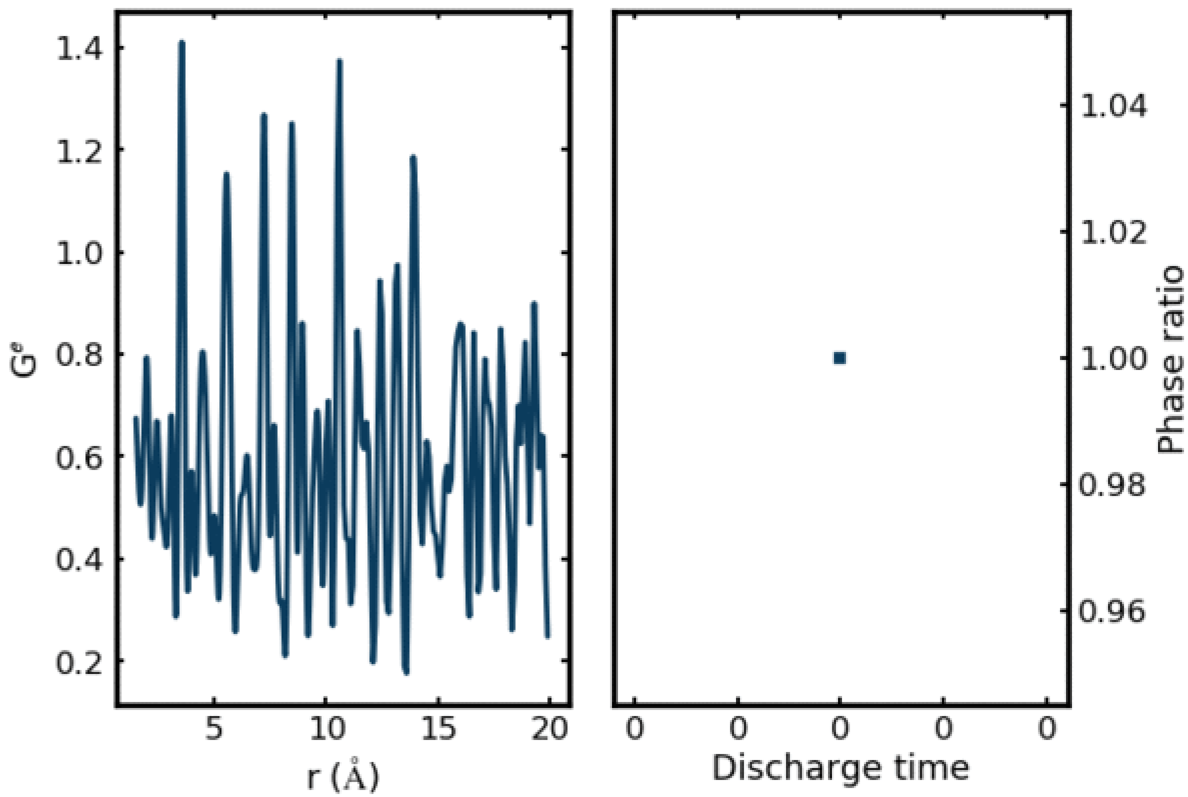
Unsupervised machine learning, e.g., NMF

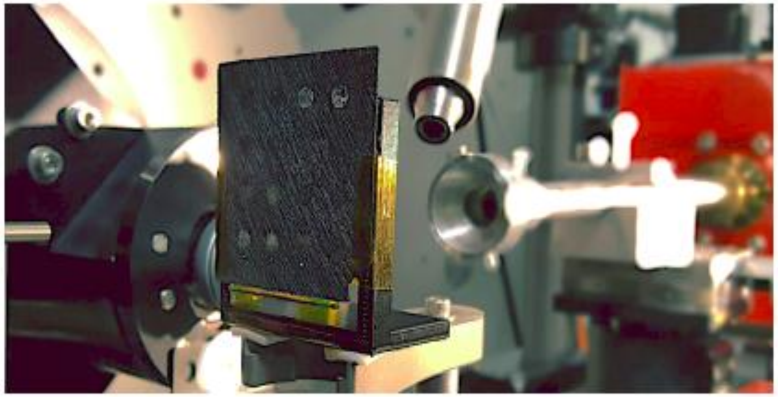






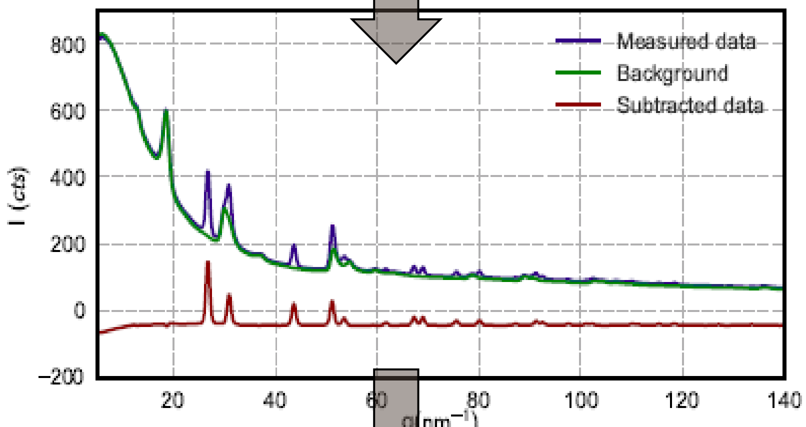
NMF in a stream example: Chemical components in a discharging battery

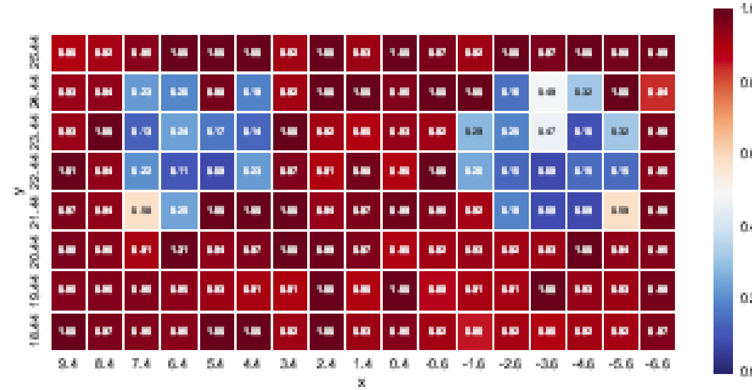


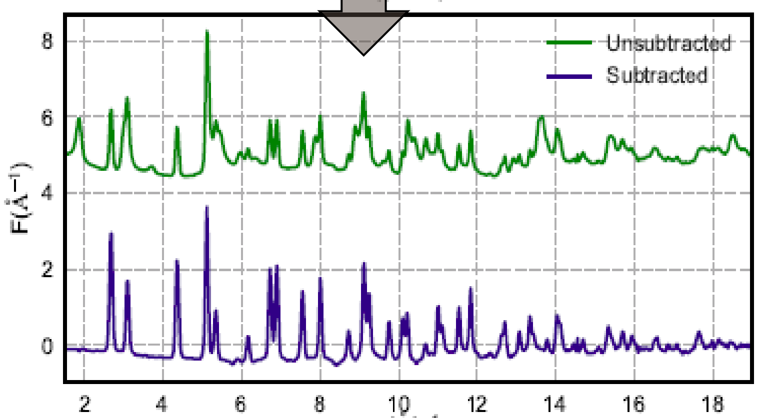


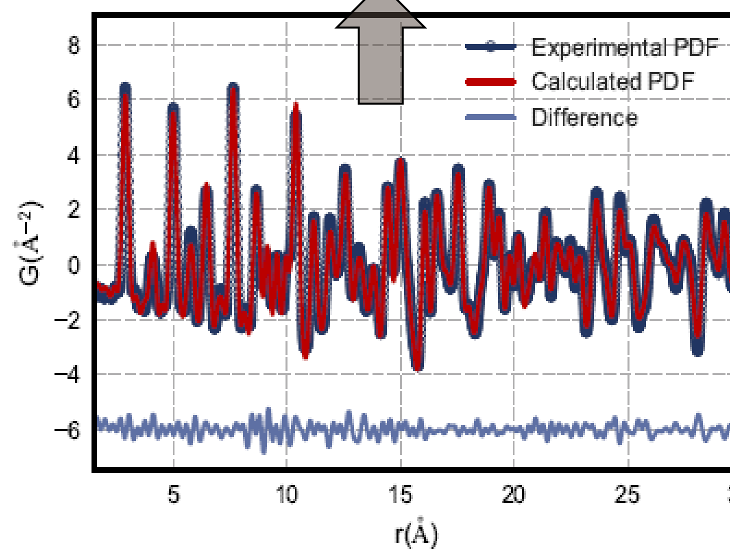
=> Spatially Resolved PDFs

- Anton Kovyakh, Soham Banerjee, Chia Hao Liu







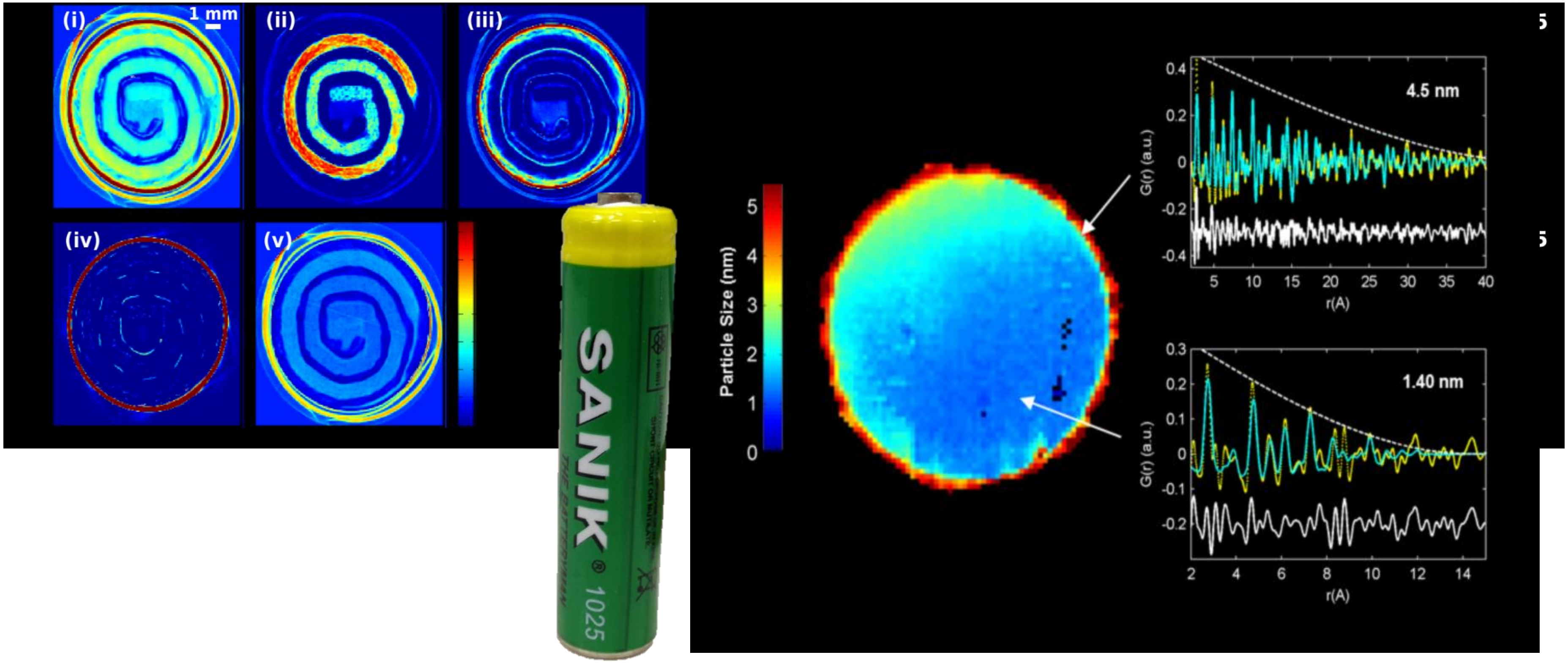


COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

BROOKHAVEN
NATIONAL LABORATORY

http://thebillingegroup.com

=> Combine PDF and tomography (ctPDF)



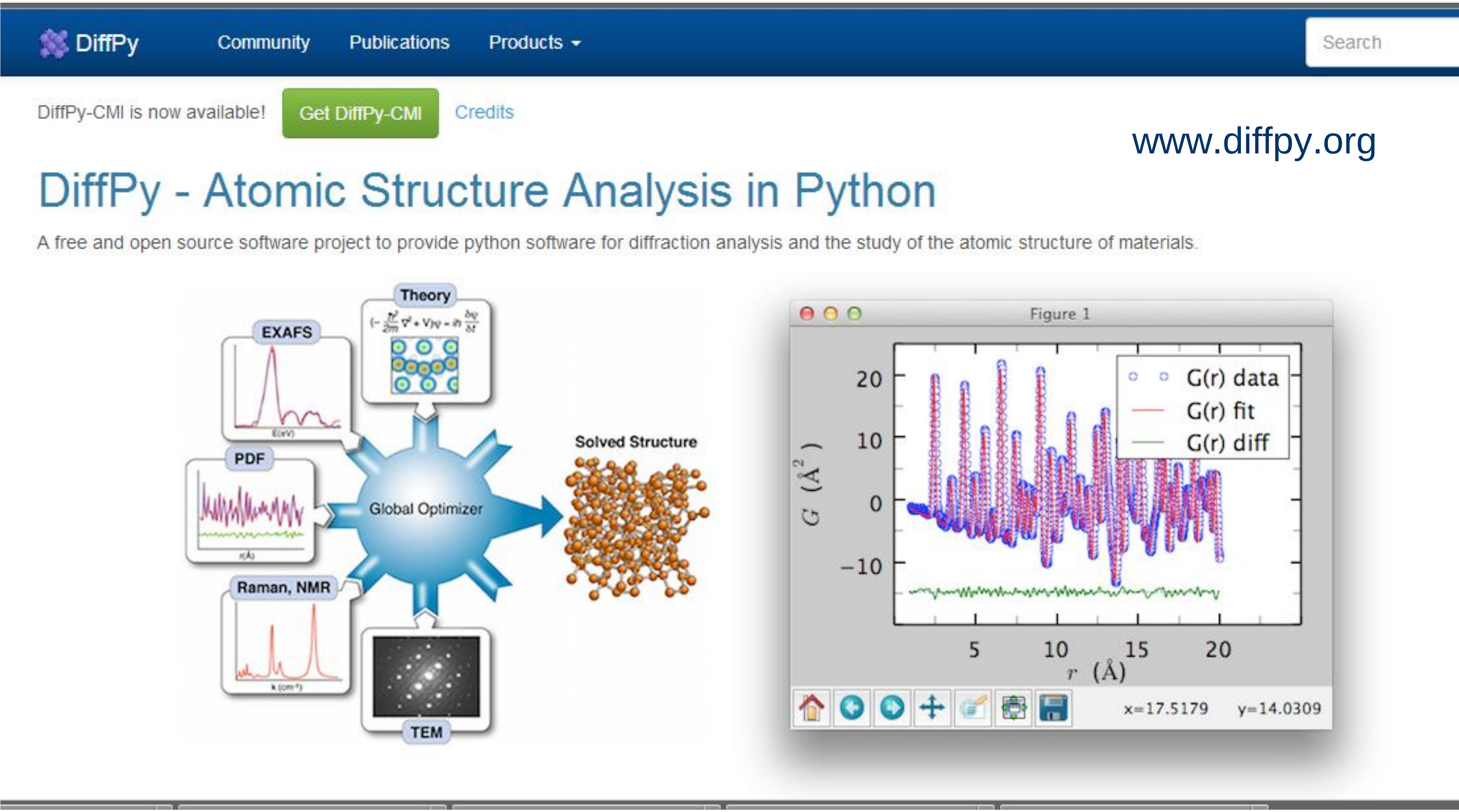
ctPDF developed with Simon Jacques, Marco DiMichiel, Andy Beal and Bob Cernik

10,000 2D datasets per image, 30 mins per image ~10Tb/day

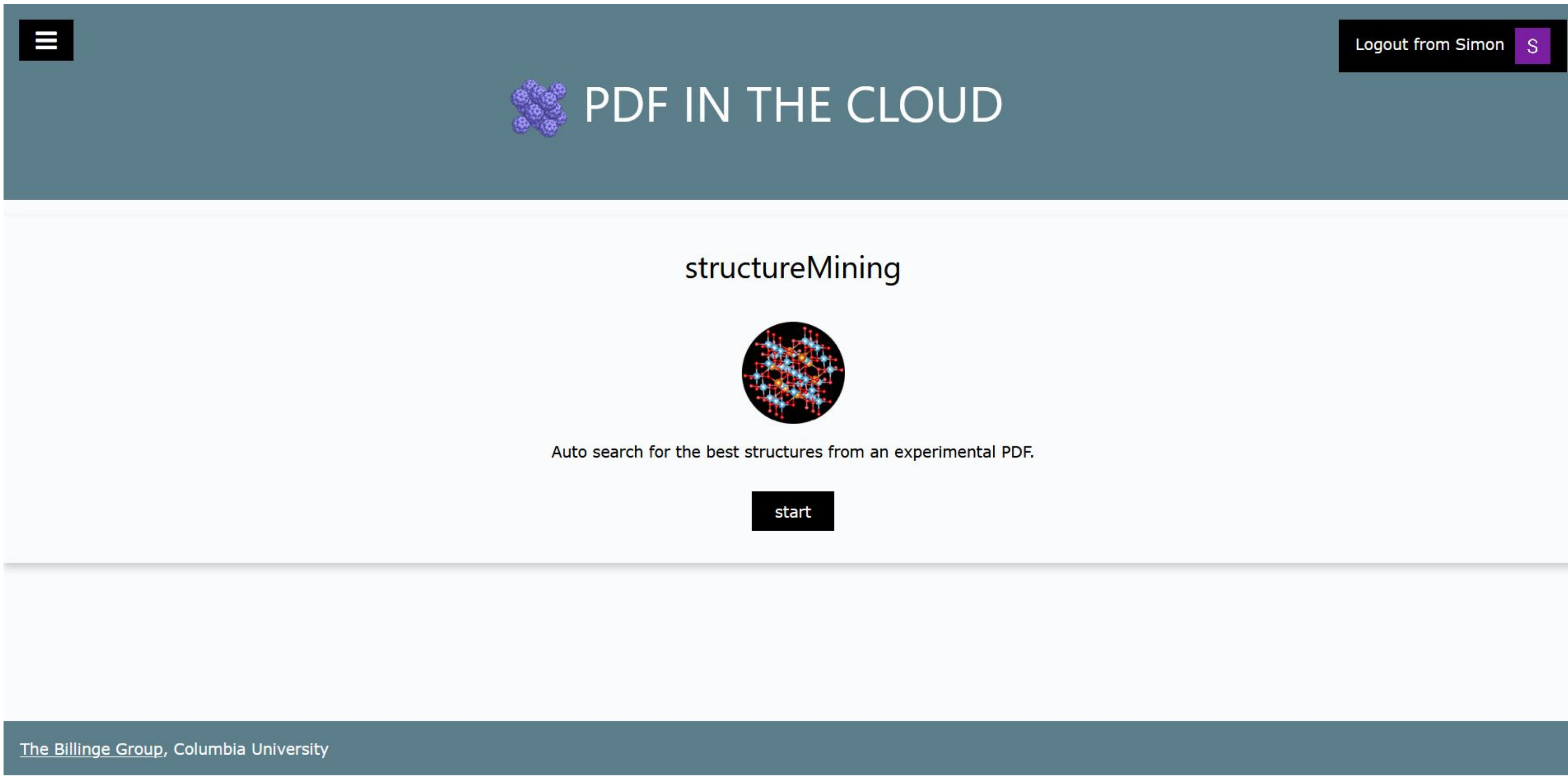
Simon D. M. Jacques, Marco Di Michiel, Simon A. J. Kimber, Xiaohao Yang, Robert J. Cernik, Andrew M. Beale and Simon J. L. Billinge, *Nat. Commun.* **4**, 2536 (2013).

Jensen, Corr, Di Michiel, SJLB et al., *J. Electrochem. Soc.* (2015)

Diffpy project
Complex Modeling infrastructure: Diffpy-CMI



To the cloud!!!!!



COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com

Log

PDF IN THE CLOUD

structureMining

Given a measured (or calculated) PDF, structureMining will search databases to find the best structures to fit it.

Please see the strucureMining paper for more information. Please cite the structureMining paper if this helps you get a publishable result.

Upload Data

PDF file: Browse... No file selected.

(•) X-ray ( ) Neutron

Composition:

Optional Parameter:

Submit

The Billinge Group, Columbia University

Columbia University in the City of New York

Brookhaven National Laboratory

HTTP://thebillingegroup.com

PDF IN THE CLOUD

structureMining

Results

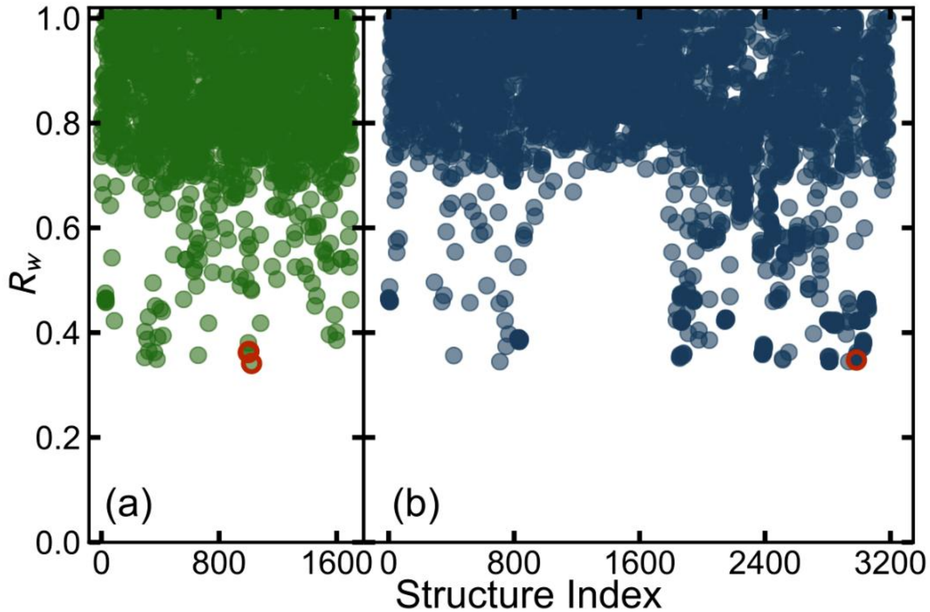
SM found total 3 structures and 3 structures with weighted profile agreement factor, Rw < 0.5. Download Full Table

Expand Table

	rw	formula	space_group	db	db_id	ref	download
0	0.140570	As4 Fe4 Sr2	I4/mmm	COD	9016374	Tegel. Journal of Physics: Condensed Matter. 20 (2008) 452201_1--452201_5	
1	0.140673	As8 Fe8 Sr4	Fmmm	COD	9016523	Tegel. Journal of Physics: Condensed Matter. 20 (2008) 452201_1--452201_5	
2	0.325028	As2 Fe2 Sr1	I4/mmm	MPD	mp-4488	Saha et al. Journal of Physics: Conference Series. 273 (2011) 012104/1-012104/4	

The Billinge Group, Columbia University

Heuristic-4 (*-*-Si-O)



- Structure-mining found the same model as in prior work, MPD No. 1003 ($NaFeSi_2O_6$) and COD No. 2983 ($NaFeSi_2O_6$), s.g.: C 2/c.
- It also returns some structures with space group C 2, such as MPD No. 998 ($Na_{0.83}FeSi_2O_6$), which may be viewed as a very similar structure but with a lowered symmetry and deficient atoms at some sites
- It also returns some structures substituting at Na or Fe sites by other elements. For example, MPD No. 1021 ($NaGaSi_2O_6$).

Acknowledgements

- A special thank you to all my current and former students and post-docs
- Also my many wonderful collaborators, mentioned during the talk
- Facilities:
  - APS, CHESS, NSLS, NSLS-II (and people therein)
  - MLNSC, ISIS, SNS (and people therein)
- Funding: DOE-BES and NSF-DMR

HTTP://thebillingegroup.com