Moving towards a more data-centric scientific literature

S.J.L. Billinge

Columbia University,

Brookhaven National Laboratory

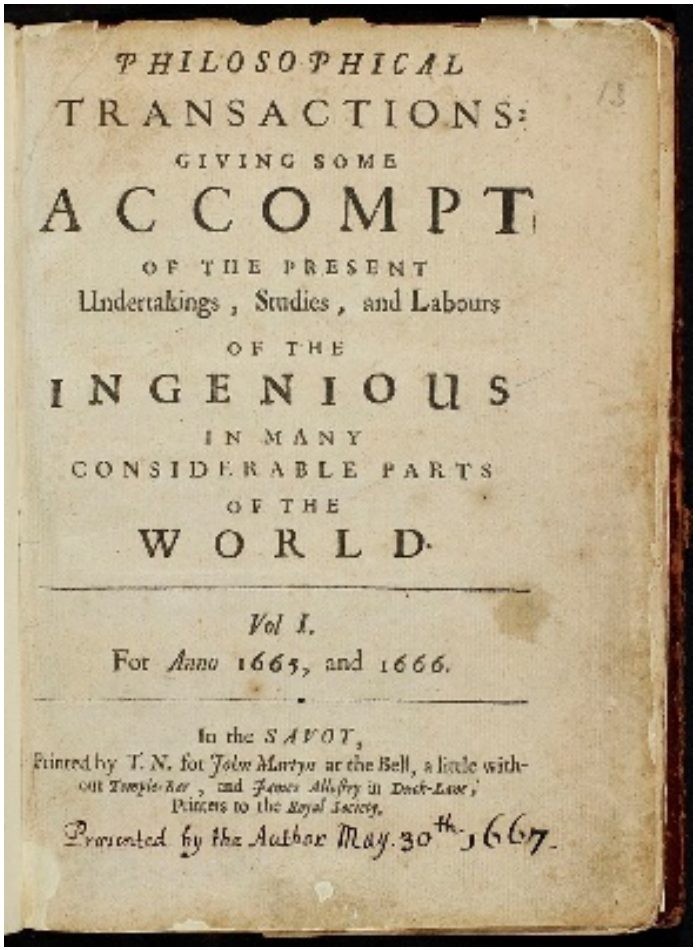
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com

Scientific literature in the service of mankind since 1665

- Transactions of the Royal Society, first published March 1665
- "giving some accompt of the present undertakings, studies and labours of the ingenious"

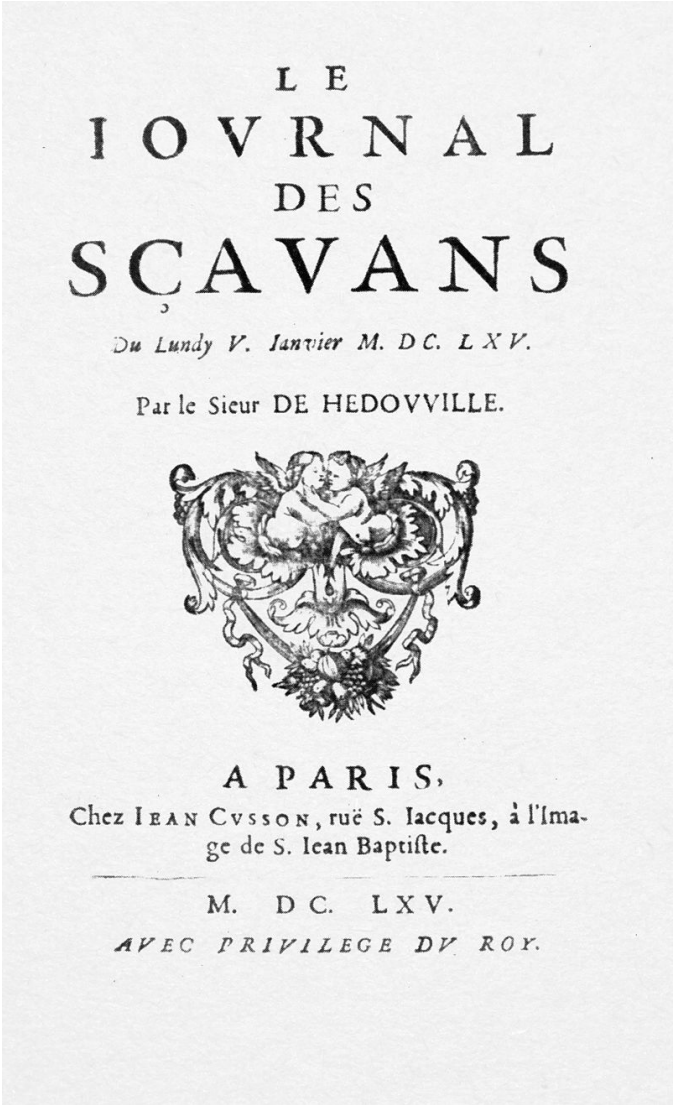


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane;
Printers to the Royal Society.
Presented by the Author May. 30th. 1667.

The first academic journal



LE
IOVRNAL
DES
SCAVANS

Du Lundy V. Ianvier M. DC. LXV.

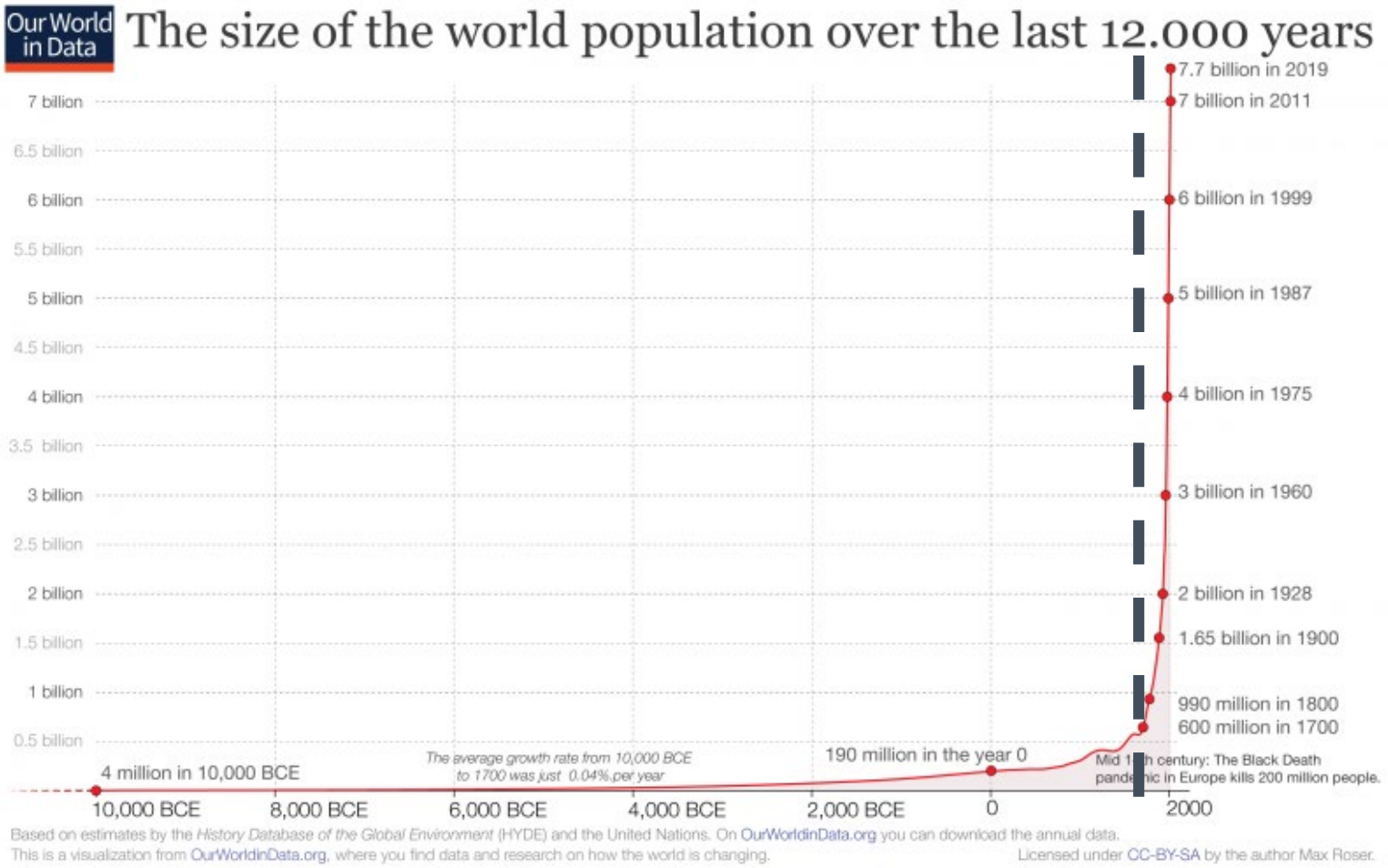
Par le Sieur DE HEDOVVILLE.

A PARIS,
Chez IEAN CVSSON, ruë S. Iacques, à l'Image de S. Iean Baptiſte.

M. DC. LXV.
AVEC PRIVILEGE DV ROY.

- January 1665!
- But not a scientific journal…
- later renamed Journal des savans and then Journal des savants, established by Denis de Sallo,
- Its content included obituaries of famous men, church history, and legal reports.

Here are some statistics

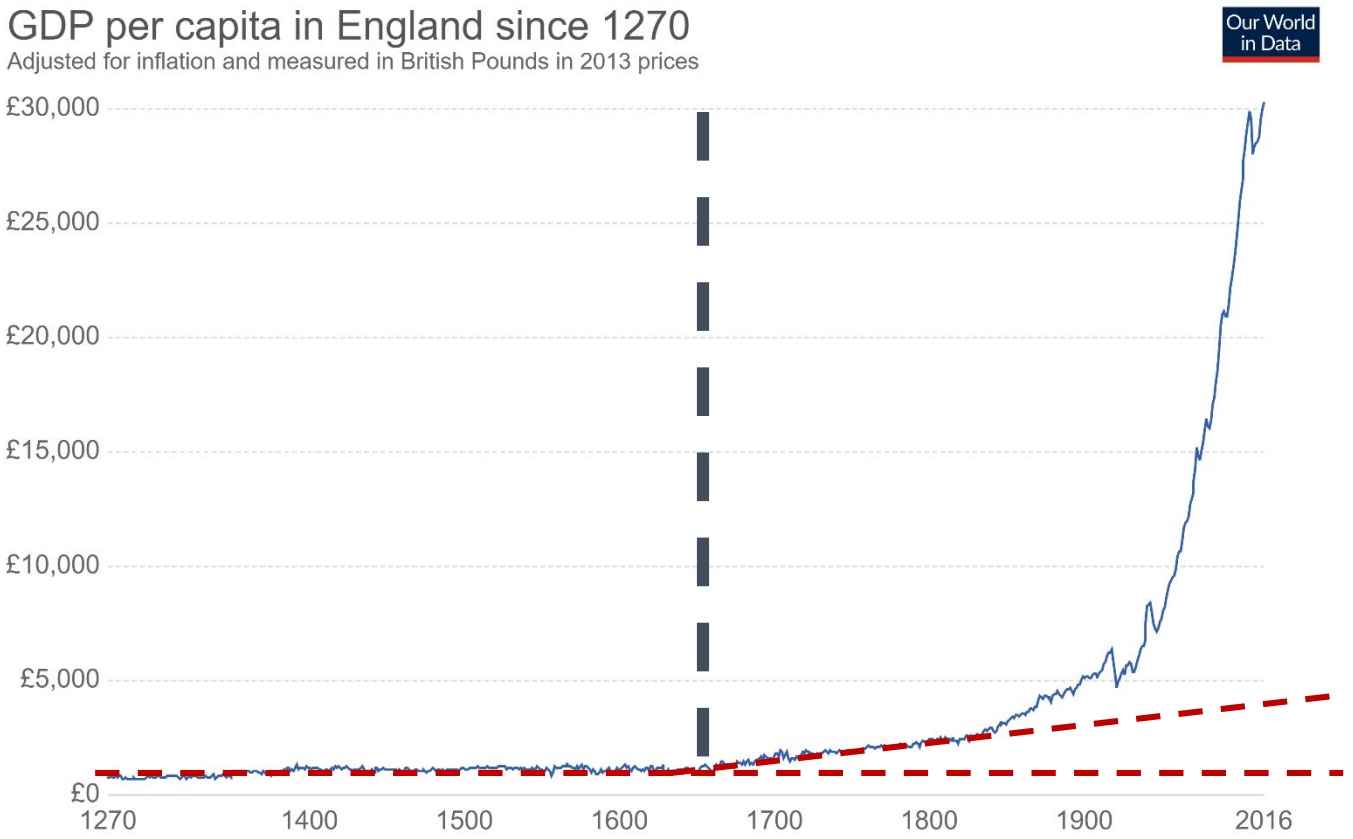


COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com

Some more statistics



Summary

- The scientific literature has caused all of human development
- We should be proud.
- Making life better for humans one paper at a time for 354 years!

How did it do it?

A personal perspective

(with apologies to professional librarians and Journalistas)

The success rests on 3 pillars

1. Disclosure
   - Creating an environment where people freely share (disclose) their ideas
     - Create incentives…
     - Citations
2. Qualitation
   - (needed an English word for the act of assuring quality, sorry)
   - Ensuring some level of quality of the work
     - peer review
     - reproducibility of work
3. Curation
   - Two aspects to this
     - Persistence of the information
     - Access to the information

The problem

- For 350 years, we have designed the scientific literature to be read by humans
- … and humans are so 20th Century!

Now we need a scientific literature that can be read by machines (and humans)

Side note

- People are building machines that can read the existing literature
- A process called Natural Language Processing (NLP)
- I am not talking about that

I want a literature that can be read by machine directly

Wait, did I just rediscover the database?

Well, yes, but I think that integrating complementary roles of databases and journals we can do much more.

What kinds of things can we do if we have a literature that can be read by machines

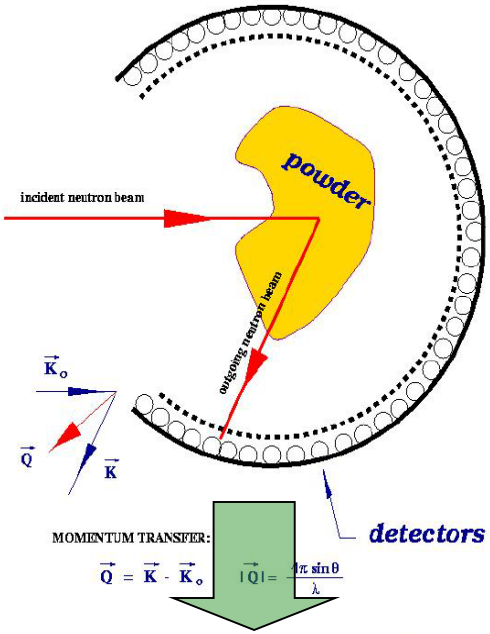
Two examples from my group:

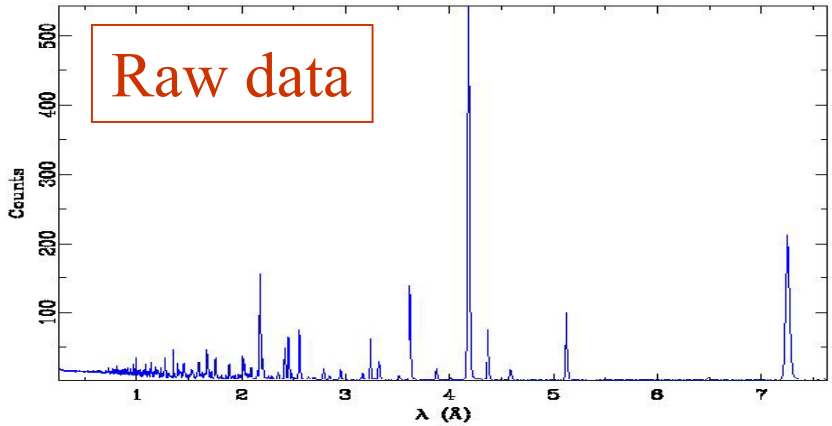
- Obtaining the space-group of a structure from its atomic pair distribution function (PDF) using Machine Learning

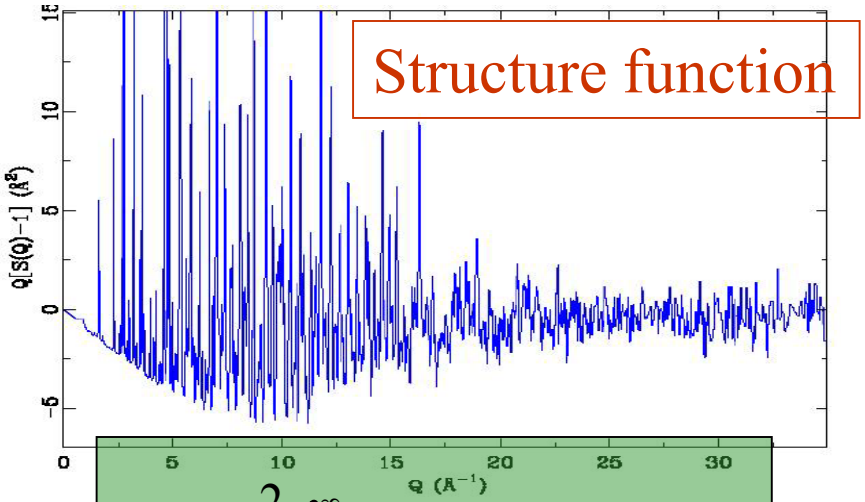
- Discovering candidate structures to fit to the PDF of an unknown structure

What is a PDF?

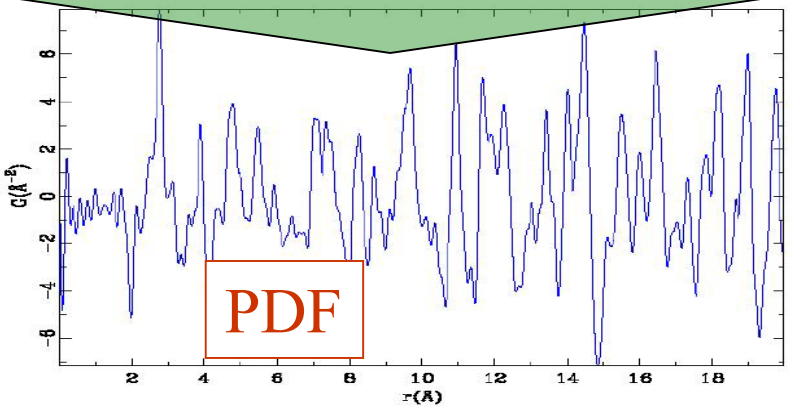
Local atomic structure from the Atomic Pair Distribution Function (PDF)







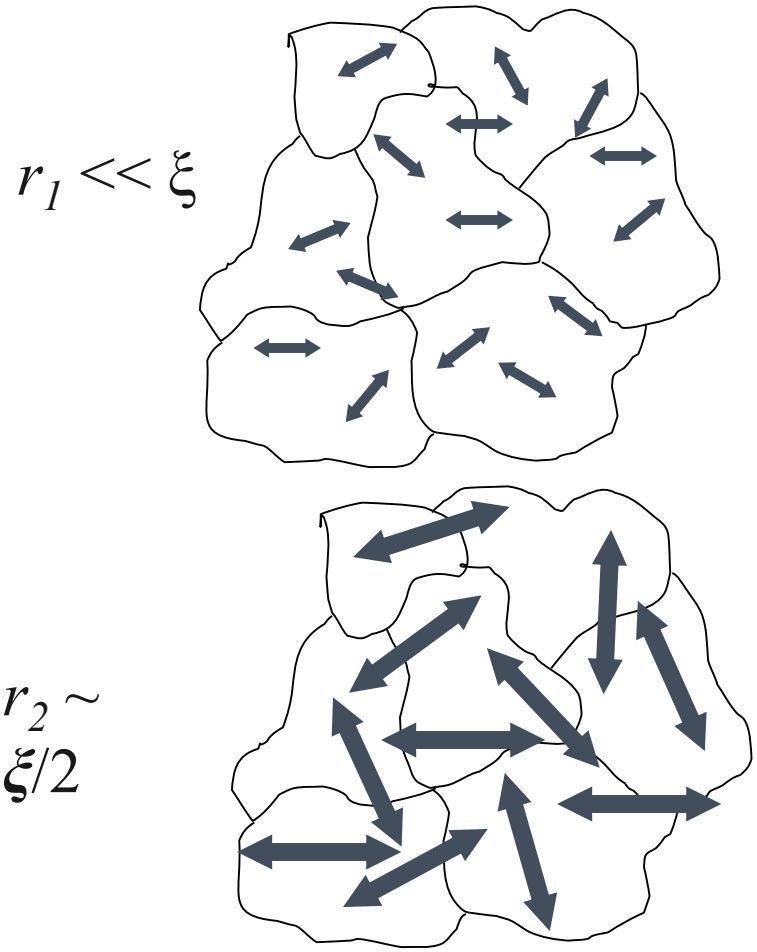
$$G(r) = \frac{2}{\pi}\int_0^\infty Q[S(Q)-1]\sin QrdQ$$

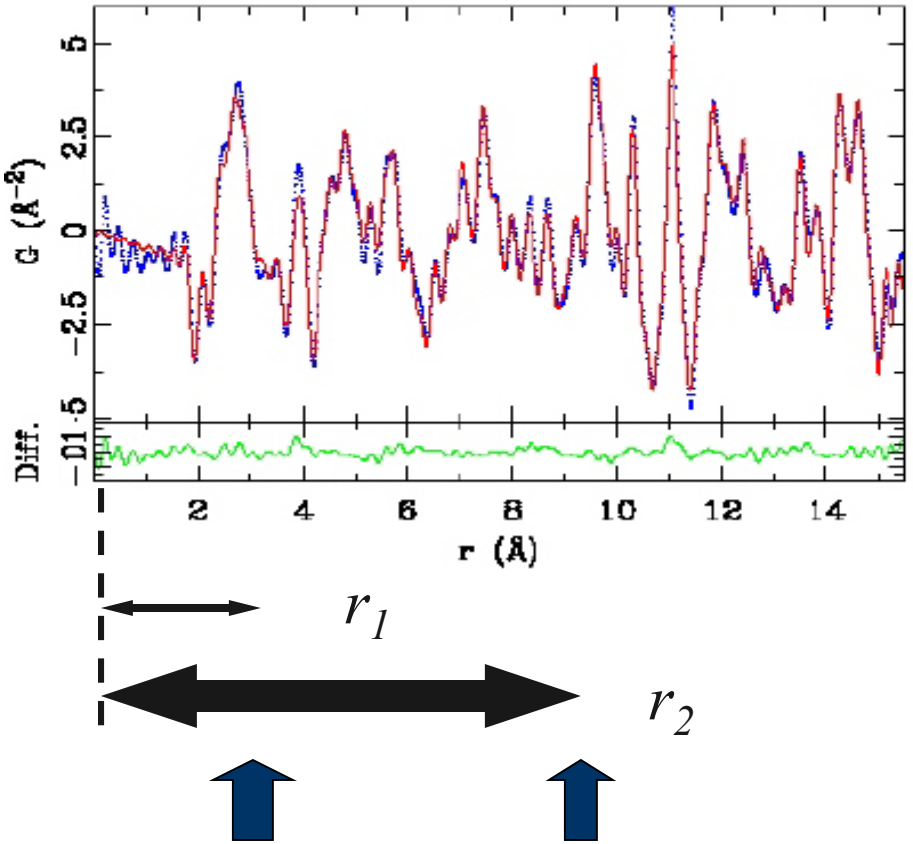


COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com


$r_1 << \xi$
$r_2 \sim \xi/2$


G (Å^{-2})
Diff.
r (Å)
r_1
r_2
Intra-domain structure
Inter-domain structure

A PDF is from a powder diffraction pattern. Why can't we index it?

Same information as the powder diffraction pattern

Different encoding!

Can we get the s.g. just by looking at the PDF?

- I don't know an person who can look at a PDF and tell the s.g. (except perhaps if it is an fcc material or something)

Question:

- Can we train a Machine Learning model to recognize it?

Answer:

- Yes!

Liu, Chia-Hao, Tao, Yunzhe, Hsu, Daniel J., Du, Qiang, and Billinge, Simon J. L.. Acta Crystallogr. A . To be published. (2019)

HTTP://thebillingegroup.com

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

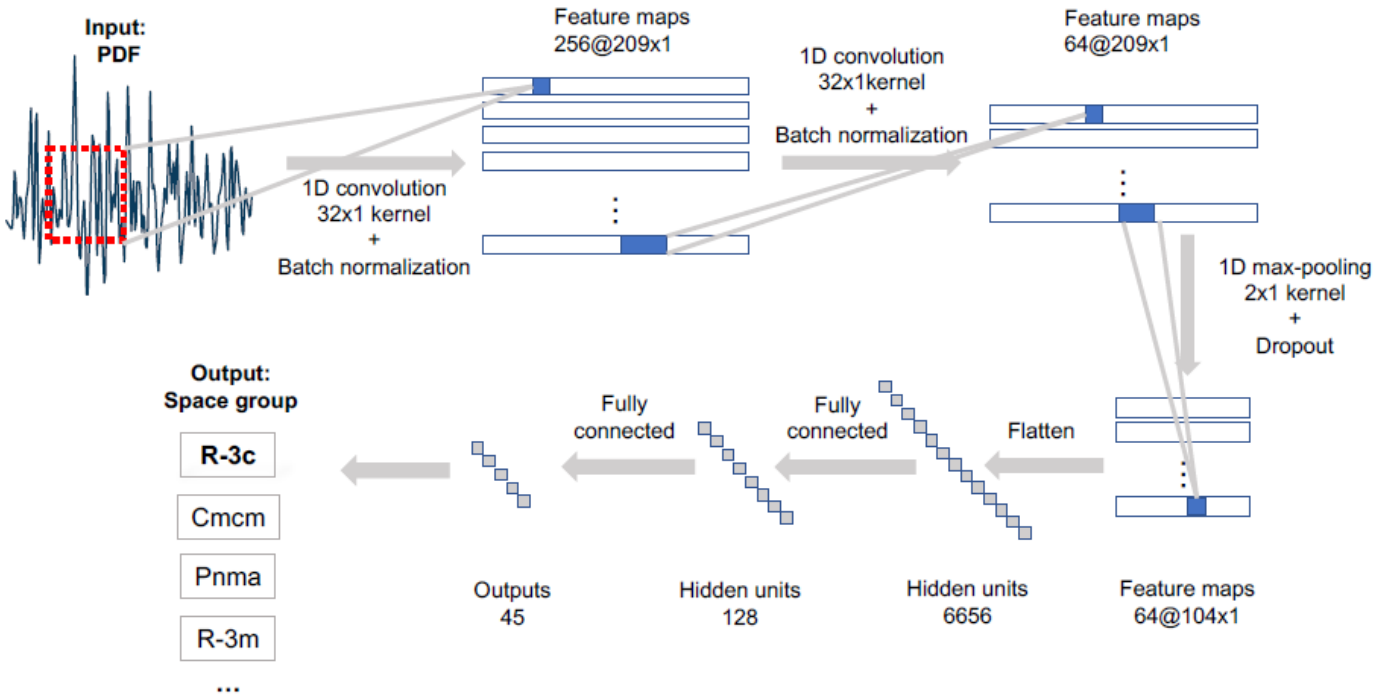
BROOKHAVEN NATIONAL LABORATORY

Teaching the machine

1. Have a database with 100,000 solved structures in it
2. Make that database machine readable (using CIF!!!!!!)
3. Divide the structures into a set of 80,000 for learning and 20,000 for testing
4. Show the ML algorithm the learning set: give it the PDF (a set of intensity-position pairs) and tell it the right answer. It will adjust its internal parameters so as to predict the right answer as often as possible
5. Check how well it did by showing it the 20,000 testing sets without telling it the right answer, but checking its answer
6. Tweak the learning till it is doing well on the test
7. The algo is ready….give it some real data and it will tell you the space-group

Details

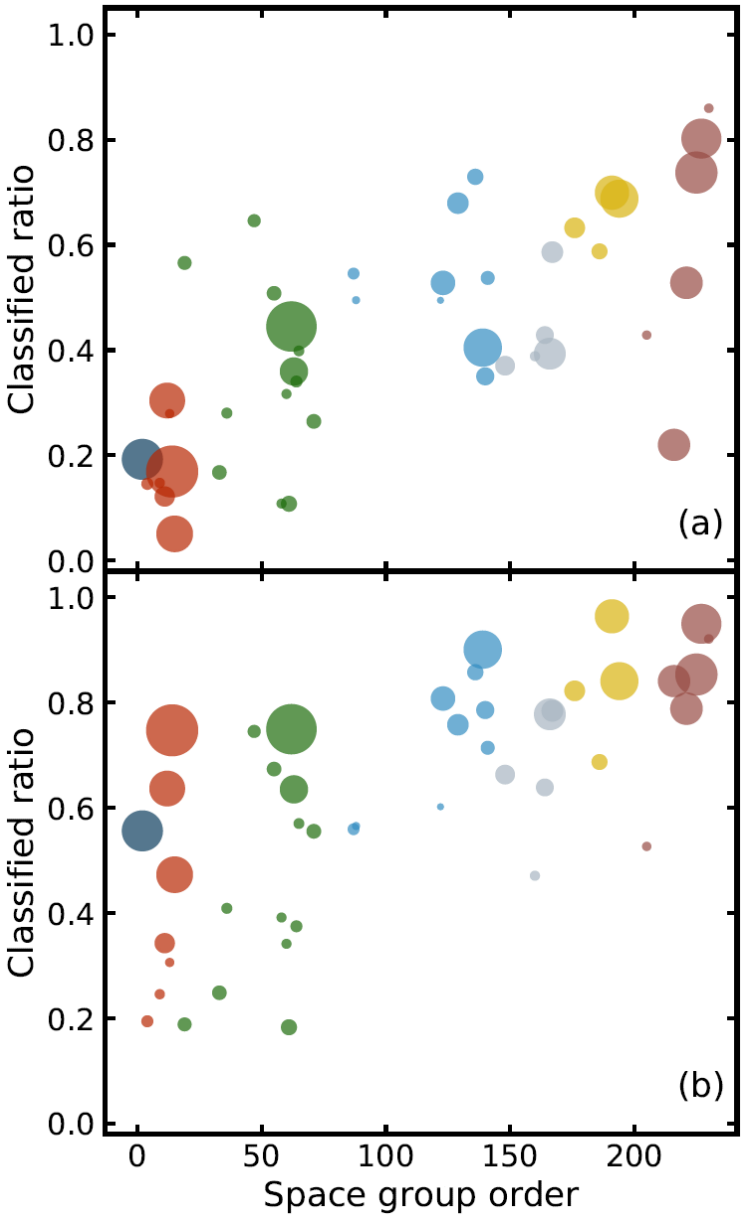
- Work of Yunzhe Tao and Chia-Hao Liu
- PDF itself is the input feature vector
- 101,802 structures in 45 space-groups
- Train on 80% of data, 20% retained for testing



Space Groups from PDFs

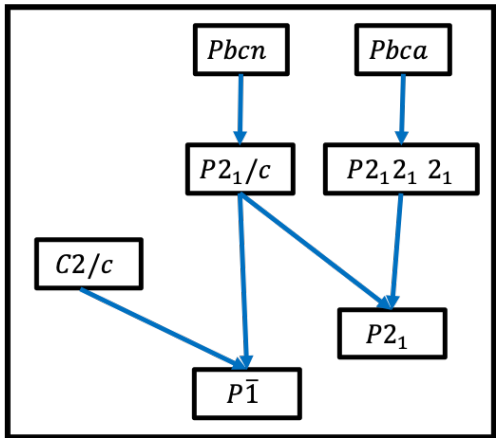
- First try, Logistic Regression

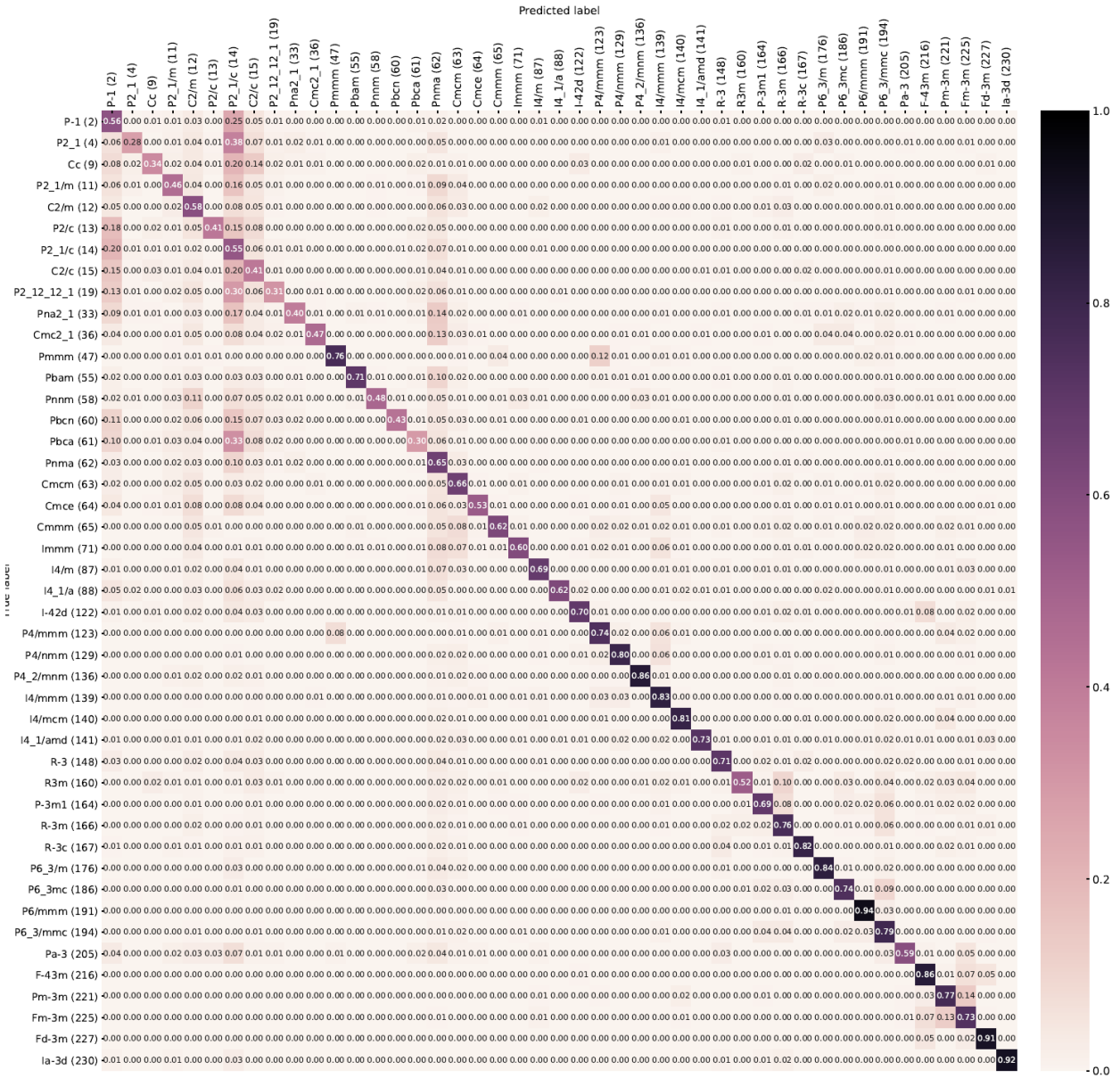
- Second try, Convolution Neural Net
  - Classification ratio's are for finding the right s.g. in the top 6



Confusion Matrix

- Rows: true label
- Columns: predicted label



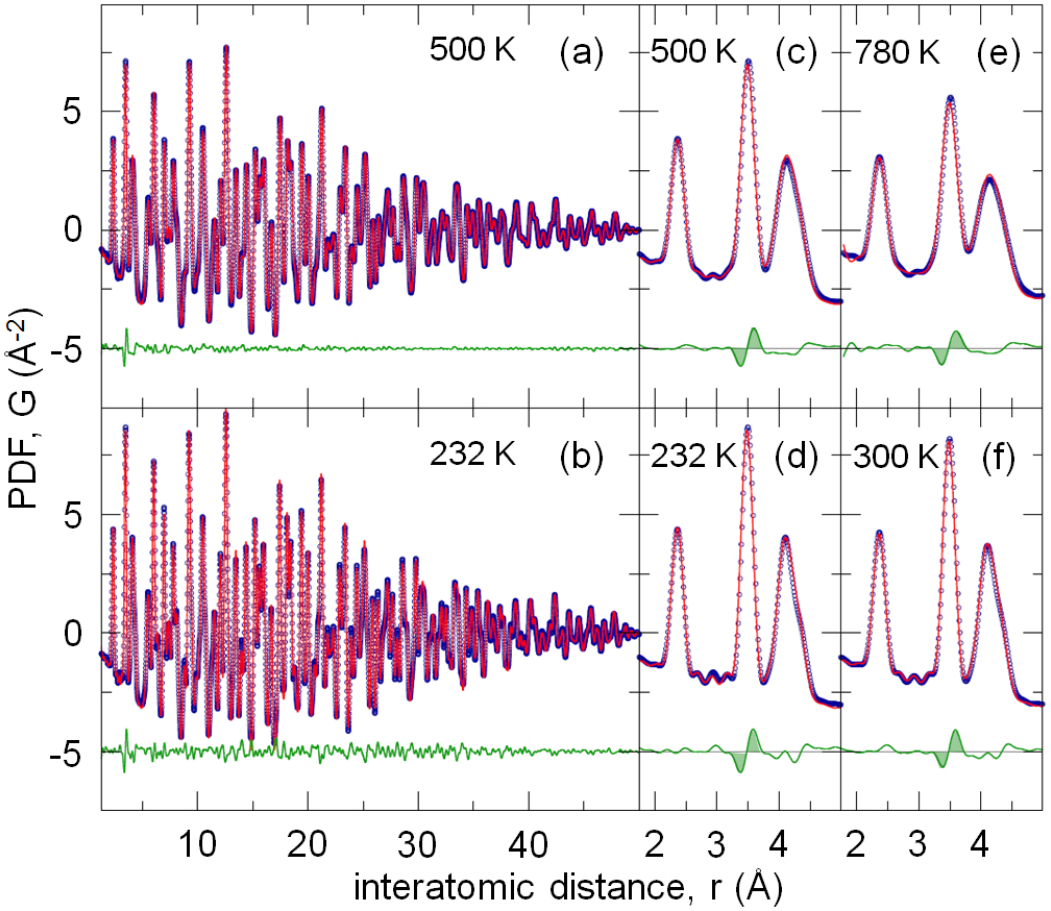


Automating modeling

- Can we think about how to make modeling more automated, and possibly introducing it into the pipeline?

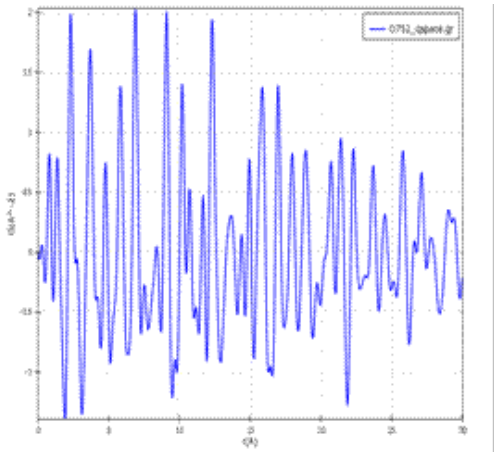
Standard approach

- 1 PDF
- 1 model
- 20 parameters
- Vary the parameters until the model agrees as well as possible with the data.
- Emphasis on parameter estimation, not on model selection
- Challenge is finding the right model



New approach: Structure-Mining

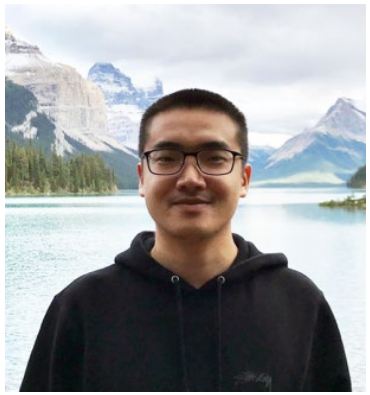
- 1 PDF
- many models
- few parameters
- Iterate over large numbers of models
- Emphasis on model selection
- Advantage: find multiple nearby models!
- Challenge: structure must be in the structure-mine



Structure-mining: An automated screening of large numbers of candidate structures to the atomic pair distribution function (PDF)

Long Yang, Pavol Juhas, Maxwell W. Terban, Simon J. L. Billinge

Acta Crystallographica A (2019) in review



COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com

Introduction

Crystal structural database

Materials Project Database
Crystallography Open Database

PDF auto-refinement

DiffPy-CMI

R_w sorted structure-mining result

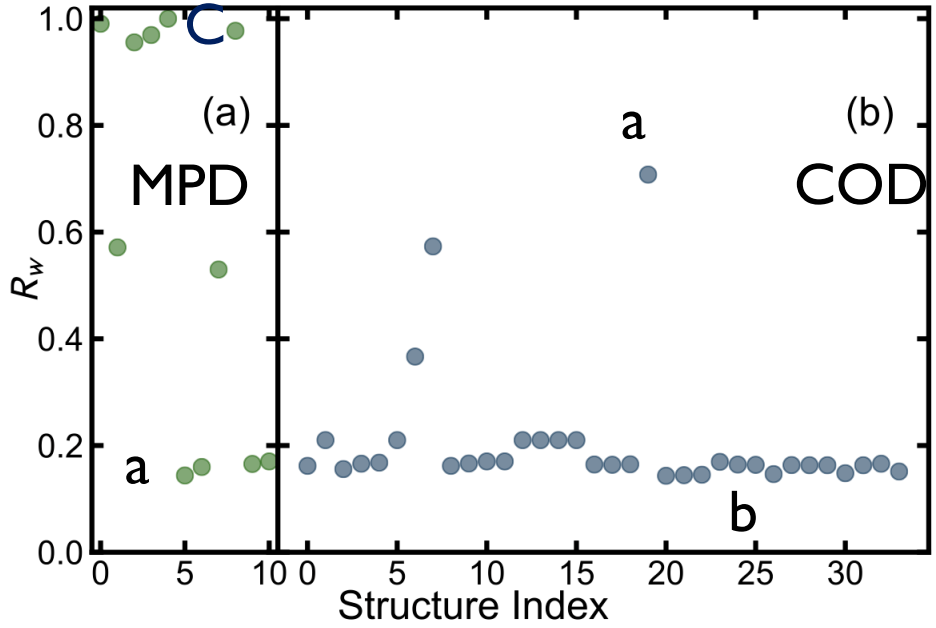
No.	R_w	formula	space_group	Ba_U_{iso}_r	O_U_{iso}_r	Ti_U_{iso}_r
20	0.143308	Ba2 O6 Ti2	Amm2	0.002085	0.018062	0.008006
21	0.144425	Ba2 O6 Ti2	Amm2	0.002027	0.017803	0.008684
22	0.145316	Ba2 O6 Ti2	Amm2	0.001999	0.017465	0.009089
26	0.146433	Ba2 O6 Ti2	Amm2	0.001973	0.016756	0.009420
30	0.148121	Ba2 O6 Ti2	Amm2	0.001957	0.017870	0.009978
33	0.151343	Ba2 O6 Ti2	Amm2	0.002535	0.006989	0.008263
2	0.155847	Ba1 O3 Ti1	P4mm	0.002735	0.011571	0.006421
8	0.161942	Ba1 O3 Ti1	P4mm	0.002634	0.016226	0.008579
0	0.161942	Ba1 O3 Ti1	P4mm	0.002634	0.016226	0.008579
31	0.163028	Ba1 O3 Ti1	R3m	0.002704	0.016278	0.007618
29	0.163086	Ba2 O6 Ti2	Amm2	0.002827	0.002921	0.005437
28	0.163120	Ba1 O3 Ti1	R3m	0.002699	0.016083	0.007675
27	0.163405	Ba1 O3 Ti1	R3m	0.002681	0.015975	0.007887
25	0.163844	Ba1 O3 Ti1	R3m	0.002663	0.015968	0.008149
17	0.163934	Ba1 O3 Ti1	R3m	0.002649	0.015838	0.008251
24	0.164318	Ba1 O3 Ti1	R3m	0.002630	0.015253	0.008431
16	0.164382	Ba1 O3 Ti1	R3m	0.002639	0.015667	0.008449
18	0.164820	Ba1 O3 Ti1	R3m	0.002614	0.015019	0.008652
32	0.165866	Ba1 O3 Ti1	R3m	0.002599	0.014932	0.009055
3	0.165928	Ba1 O3 Ti1	P4mm	0.002610	0.015074	0.009552
9	0.166332	Ba1 O3 Ti1	P4mm	0.002616	0.015758	0.009675
4	0.167933	Ba1 O3 Ti1	Pmm2	0.002485	0.015098	0.009475
23	0.169423	Ba1 O3 Ti1	P4mm	0.002625	0.016343	0.010339
11	0.170190	Ba1 O3 Ti1	P4/mmm	0.002643	0.017426	0.010542
10	0.170190	Ba1 O3 Ti1	P4/mmm	0.002643	0.017426	0.010542
15	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
1	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
13	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
12	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
5	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
14	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
6	0.366657	Ba1 O3 Ti1	Pm-3m	0.005790	0.079941	0.012610
7	0.573221	Ba6 O18 Ti6	P6_3/mmc	0.006970	0.046913	0.004137
19	0.707676	Ba1 O3 Ti1	P4mm	0.004189	0.047858	0.248950

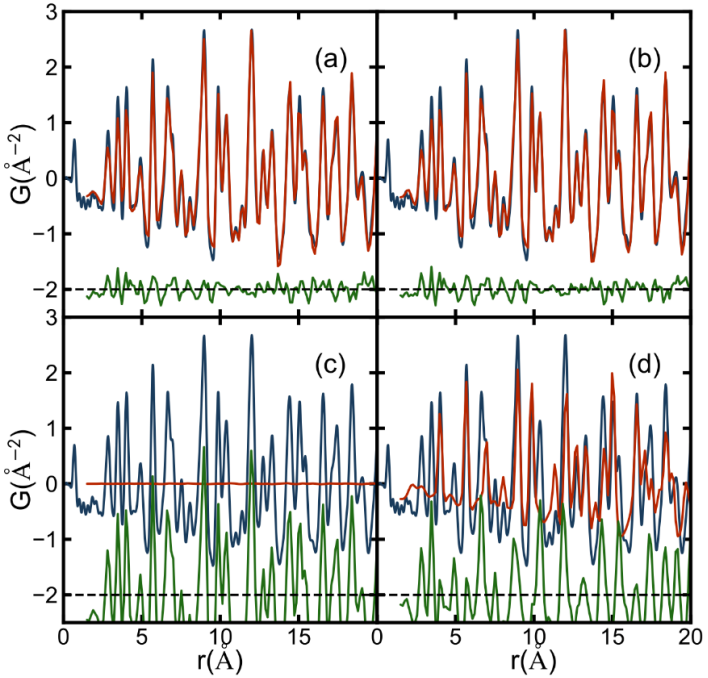
Example result of BaTiO3 nanoparticle heuristic- I

$BaTiO_3$ nanoparticle

- Heuristic 1: $BaTiO_3$
- Heuristic 2: Ba-Ti-O
- Heuristic 3: Ba-Ti-O-*
- Heuristic 4: Ba-*-* or even *-*-*

Heuristic-1

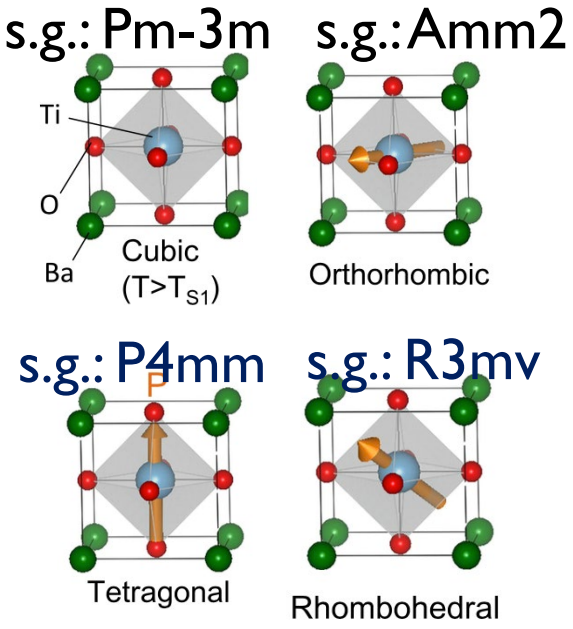




PDFs from representative satisfactory and unsatisfactory structures from the MPD (a, c) and the COD (b, d)

Heuristic-I (search for $BaTiO_3$)

- Candidate structures



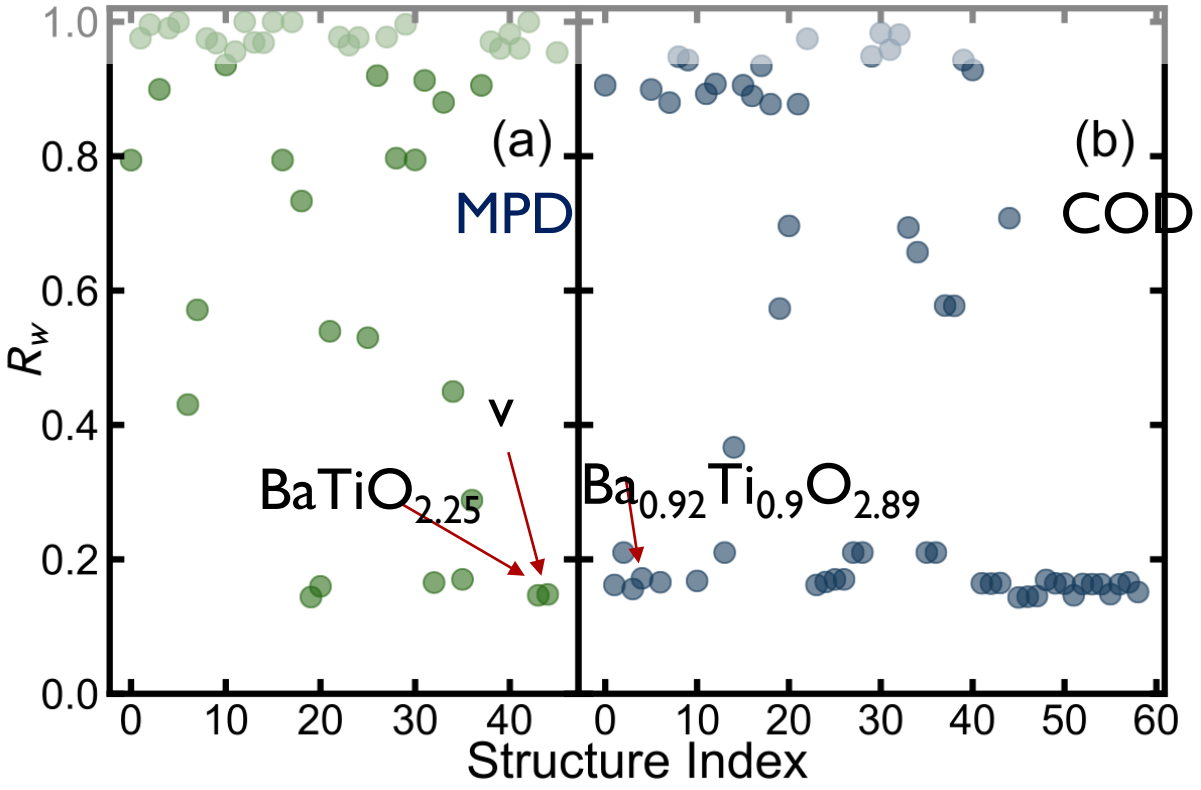
Figures are from: Fujioka, J., et al. Scientific reports 5 (2015).

Lombardi, J., Yang, L., Pearsall, F. A., Farahmand, N., Gai, Z., Billinge, S. J., & O'Brien, S. P. (2019). Stoichiometric control over ferroic behavior in $Ba(Ti_{1-x}Fe_x)O_3$ nanocrystals. Chemistry of Materials.

No.	R_w	formula	space_group	Ba_U_{iso}_r	O_U_{iso}_r	Ti_U_{iso}_r
20	0.143308	Ba2 O6 Ti2	Amm2	0.002085	0.018062	0.008006
21	0.144425	Ba2 O6 Ti2	Amm2	0.002027	0.017803	0.008684
22	0.145316	Ba2 O6 Ti2	Amm2	0.001999	0.017465	0.009089
26	0.146433	Ba2 O6 Ti2	Amm2	0.001973	0.016756	0.009420
30	0.148121	Ba2 O6 Ti2	Amm2	0.001957	0.017870	0.009978
33	0.151343	Ba2 O6 Ti2	Amm2	0.002535	0.006989	0.008263
2	0.155847	Ba1 O3 Ti1	P4mm	0.002735	0.011571	0.006421
8	0.161942	Ba1 O3 Ti1	P4mm	0.002634	0.016226	0.008579
0	0.161942	Ba1 O3 Ti1	P4mm	0.002634	0.016226	0.008579
31	0.163028	Ba1 O3 Ti1	R3m	0.002704	0.016278	0.007618
29	0.163086	Ba2 O6 Ti2	Amm2	0.002827	0.002921	0.005437
28	0.163120	Ba1 O3 Ti1	R3m	0.002699	0.016083	0.007675
27	0.163405	Ba1 O3 Ti1	R3m	0.002681	0.015975	0.007887
25	0.163844	Ba1 O3 Ti1	R3m	0.002663	0.015968	0.008149
17	0.163934	Ba1 O3 Ti1	R3m	0.002649	0.015838	0.008251
24	0.164318	Ba1 O3 Ti1	R3m	0.002630	0.015253	0.008431
16	0.164382	Ba1 O3 Ti1	R3m	0.002639	0.015667	0.008449
18	0.164820	Ba1 O3 Ti1	R3m	0.002614	0.015019	0.008652
32	0.165866	Ba1 O3 Ti1	R3m	0.002599	0.014932	0.009055
3	0.165928	Ba1 O3 Ti1	P4mm	0.002610	0.015074	0.009552
9	0.166332	Ba1 O3 Ti1	P4mm	0.002616	0.015758	0.009675
4	0.167933	Ba1 O3 Ti1	Pmm2	0.002485	0.015098	0.009475
23	0.169423	Ba1 O3 Ti1	P4mm	0.002625	0.016343	0.010339
11	0.170190	Ba1 O3 Ti1	P4/mmm	0.002643	0.017426	0.010542
10	0.170190	Ba1 O3 Ti1	P4/mmm	0.002643	0.017426	0.010542
15	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
1	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
13	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
12	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
5	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
14	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
6	0.366657	Ba1 O3 Ti1	Pm-3m	0.005790	0.079941	0.012610
7	0.573221	Ba6 O18 Ti6	P6_3/mmc	0.006970	0.046913	0.004137
19	0.707676	Ba1 O3 Ti1	P4mm	0.004189	0.047858	0.248950

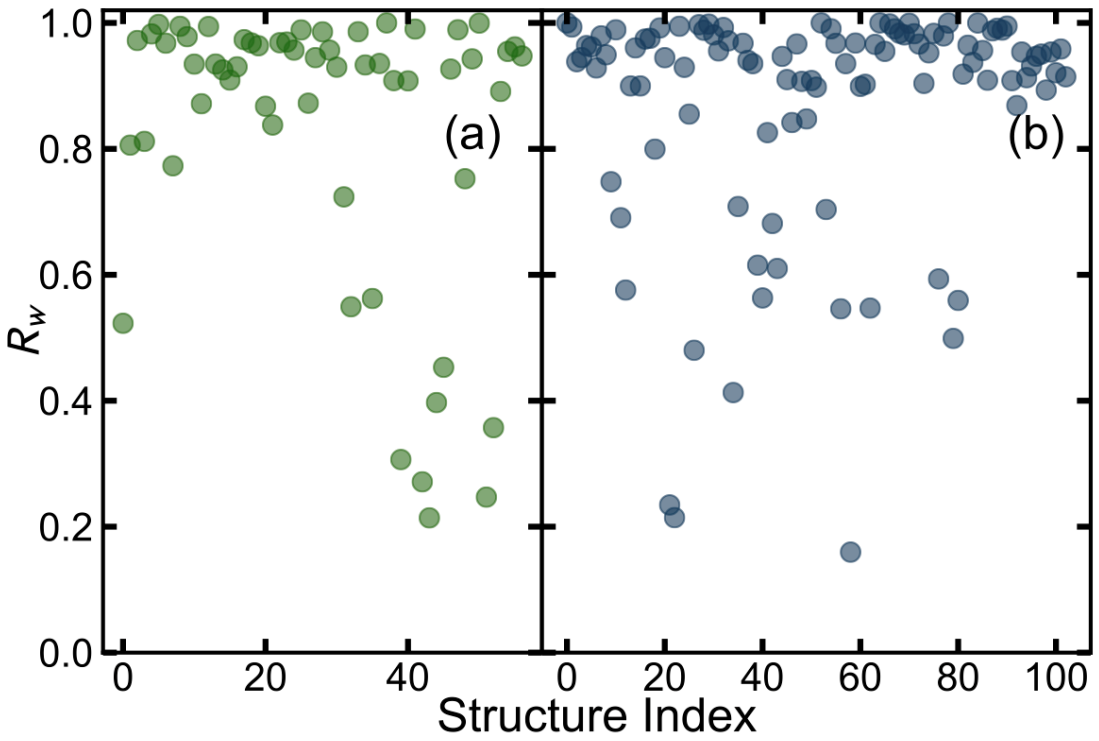
structure-mining finds them all.

Heuristic-2 (Ba-Ti-O)



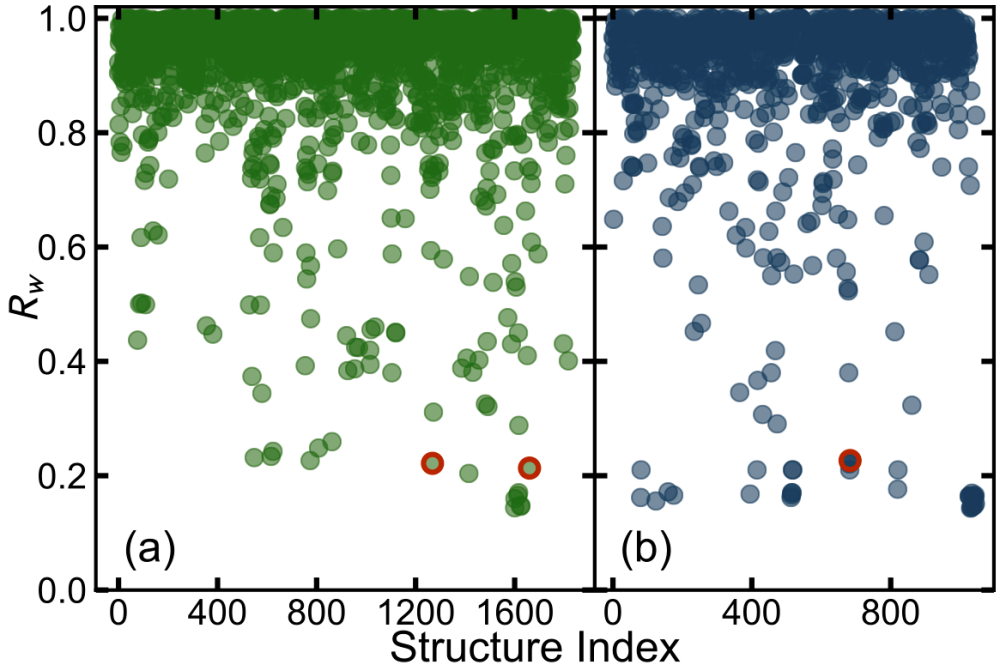
- H-2 found all the structures that were found with H-1, as expected
- It additionally returned some oxygen deficient models
  - MPD No. 43 ($BaTiO_{2.25}$)
  - MPD No. 44 ($BaTiO_{2.67}$)
  - COD No. 4 ($Ba_{0.92}Ti_{0.9}O_{2.89}$)
  - oxygen is most likely deficient in the nanoparticle samples, which was missed in the original structure refinements (Lombardi et al., 2019), but is suggested by the structure-mining .
- Models with stoichiometry far away from 1:1:3 results in poorer fit.

Heuristic-3 (Ba-Ti-O-*)



- Most of the best fit structures in H-3 have slightly worse Rw (~ 0.2) than those in heuristic-1 and 2 (~ 0.14).
- The new structures pulled are mostly Ba or Ti site doped by another element and they also have an approximate stoichiometry 113.

Heuristic-4 (Ba-*-*)



- The normal BaTiO3 perovskite structures are still ranked at the top.
- Following that, it additionally returns some perovskite structures that have Ti replaced with other species with similar x-ray scattering power to Ti, such as MPD No. 1660 ($BaVO_3$), MPD No. 1268 ($BaMnO_3$), and COD No. 683 ($BaFeO_3$). These gave agreements of R_w ~ 0.2 to 0.1433 for the best-fit structures ($BaTiO_3$).
- So the structure-mining is able to distinguish these nearby but incorrect structures from the ones with correct atom species.

Nanowire $NaFeSi_2O_6$

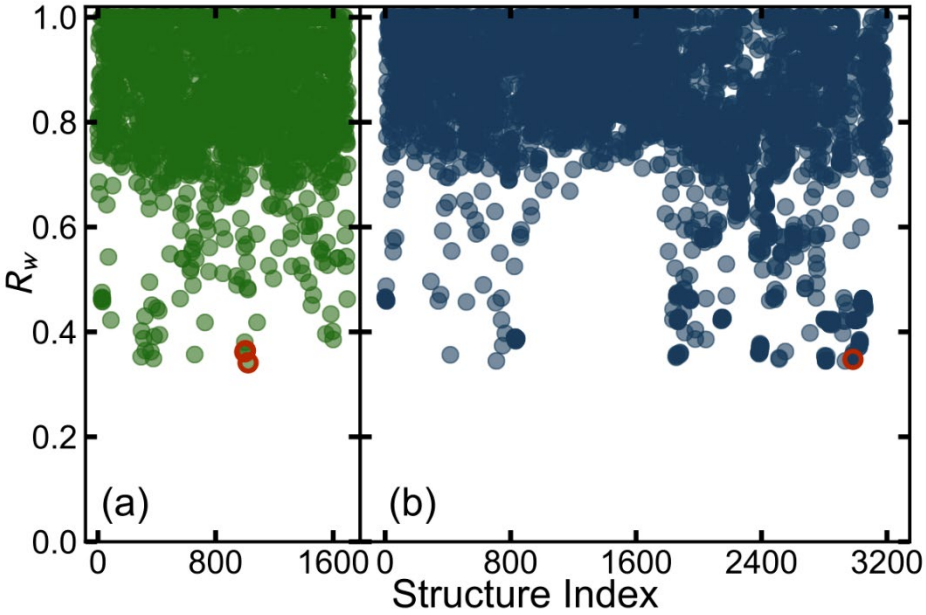
- A large amount of labor and time were spent to **manually** determine the structure of the pyroxene-type silicates with a generic composition of XYSi2O6 (wherein both 'X' = mono- or divalent metals and 'Y' = diva- lent or trivalent metals).
- The prior work found that the structure is most likely $NaFeSi_2O_6$ (s.g.: C 2/c).

Table S1: Information for all structures tested against the PDF data is tabulated. The results of the Pearson correlation calculation between model and experimental PDFs are shown. The most highly correlated model results are highlighted (red > orange).

			Composition	Name	Space group	Pearson
	Database	#				DM154_1
1	AMCSD	0009492	Na Sc Si2 O6	Jervisite	C 1 2/c 1	0.67
2	COD	9000393	Co2 (Si O4)	Co-olivine	I m m a	0.19
3	COD	9005438	Na Fe Si2 O6	Aegirine	C 1 2/c 1	0.92
4	COD	9007046	Fe2 O4 Si	Fayalite	P n m a	0.23
5	ICSD	844	Fe2 (Si O4)	Spinel-Al2MgO4	F d -3 m S	0.17
6	ICSD	845	Co2 (Si O4)	Spinel-Al2MgO4	F d -3 m S	0.20
7	ICSD	4353	Fe2 Si O4	Olivine-Mg2SiO4	P b n m	0.21
8	ICSD	4369	Na6 Fe (Al4 Si8 O26)	Naujakasite	C 1 2/m 1	-0.06
9	ICSD	6248	Co2 (Si O4)	Olivine-Mg2SiO4	P b n m	0.23
10	ICSD	8132	Co2 Si O4	Mn2GeO4	I m m a	0.19
11	ICSD	9362	Co3 O4	Spinel-Al2MgO4, reduced symmetry	F -4 3 m	0.19
12	ICSD	9671	Na Fe Si2 O6	Pyroxene-CaMg(SiO3)2	C 1 2/c 1	0.92
13	ICSD	10480	Li Al5 O8	Spinel-LiFe5O8	P 43 3 2	-0.06
14	ICSD	15388	Na2 (Si O3)	Na2SiO3	C m c 21	-0.15
15	ICSD	17054	Co (Si O3)	Pyroxene-MgSiO3 (cobaltian)	P b c a	0.68
16	ICSD	34669	Na2 (Si2 O5)	Li2Si2O5	P c n b	-0.02
17	ICSD	34688	Na2 (Si2 O5)	Na2Si2O5 (Natrosilite)	P 1 21/a 1	-0.04
18	ICSD	34950	Na2 Li Fe Si6 O15	Tuhualite (Emeleusite)	A c a m	0.21
19	ICSD	41257	(Fe2.38 Co0.62 O4)	Spinel-Al2MgO4	F d -3 m S	0.21
20	ICSD	62594	Na4 (Si O4)	Na4SiO4	P -1	0.31
21	ICSD	64802	Fe1.58 (Si O4)	Laihunite 3M substructure	P 21/b 1 1	0.09
22	ICSD	66759	(Co Fe2) O4	CoFe2O4	R -3 m H	0.21
23	ICSD	68984	Na (Fe O2)	NaFeO2(beta)	P n 21 a	0.17
24	ICSD	68985	Na0.975 (Fe0.975 Si0.025) O2	NaFeO2(beta)	P n 21 a	0.24
25	ICSD	68986	Na0.95 (Fe0.95 Si0.05) O2	NaFeO2(beta)	P n 21 a	0.17
26	ICSD	68987	Na0.925 (Fe0.925 Si0.075) O2	NaFeO2(beta)	P n 21 a	0.17
27	ICSD	68988	Na0.9 (Fe0.9 Si0.1) O2	NaFeO2(beta)	P n 21 a	0.16
28	ICSD	68989	K (Ga O2)	KAlO2	P b c a	0.29
29	ICSD	68990	Na0.895 (Fe0.905 Si0.095) O2	(NaFe)1-xSixO2	P b c a	0.42
30	ICSD	74640	Na2 (Si O3)	Na2SiO3	C m c 21	-0.13
31	ICSD	80378	Na2 (Si2 O5)	Epsilon sodium silicate	P b c 21	0.18
32	ICSD	81134	Na2 (Si3 O7)	Na2Si3O7(mS48)	C 1 2/c 1	0.14
33	ICSD	82410	Na2 Si (Si3 O9)		P 1 21/n 1	0.25
34	ICSD	84819	Na2 Co (Si4 O10)	KNaFeS14O10	P -1	0.41
35	ICSD	85551	Na8 Si (Si6 O18)		R -3 R	-0.04
36	ICSD	98551	(Co0.465 Fe0.535) (Co1.535 Fe0.465) O4	Spinel-Al2MgO4	F d -3 m Z	0.22
37	ICSD	98564	Na2 (Si2 O5)	Na2Si2O5(oP36)	P n 21 a	0.23
38	ICSD	98710	Na2 (Si2 O5)		I 41/a Z	-0.08
39	ICSD	109044	Co Fe2 O4	Spinel-Al2MgO4	F d -3 m S	0.20
40	ICSD	154114	Na6 (Si6 O19)		P 1 21/c 1	0.08
41	ICSD	166645	Na0.628 ((Co0.97 Fe0.03) O2)	NaxCoO2	P 63/m m c	-0.17
42	ICSD	200805	K2 Na2 Fe2 (Si4 O10)2	KNaFeS14O10	P -1	0.02
43	ICSD	240870	K2 Na2 Fe2 (Si4 O10)2	KNaFeS14O10	P -1	0.53
44	ICSD	250222	K2 Na2 Fe2 (Si4 O10)2	KNaFeS14O10	P -1	0.21

Lewis, Crystal S., et al. "Synthesis, characterization, and growth mechanism of motifs of ultrathin cobalt-substituted NaFeSi2O6 nanowires." CrystEngComm 20. 2 (2018): 223-236.

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 ($NaGaSi_2O_6$).

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

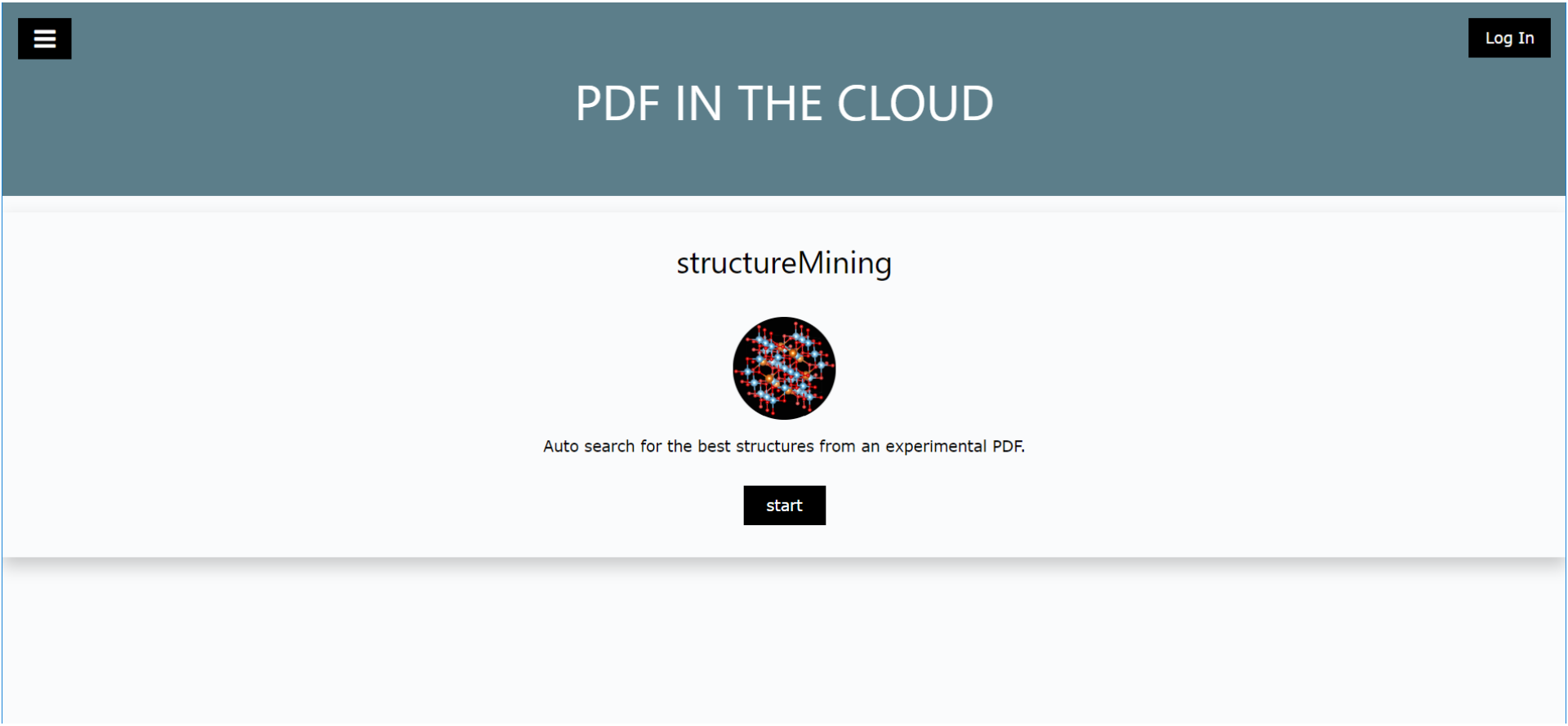
MPD

No.	R_w	formula	space_group
1021	0.341109	Ga2 Na2 O12 Si4	C2/c
377	0.349254	Ca1 Ni2 O12 Si4	C2
294	0.352749	Ca1 Co2 O12 Si4	C2
658	0.356962	Cr2 Na2 O12 Si4	C2/c
998	0.361981	Fe6 Na5 O36 Si12	C2
360	0.362239	Ca1 Mn2 O12 Si4	C2
1003	0.364362	Fe2 Na2 O12 Si4	C2/c
322	0.364513	Ca1 Fe2 O12 Si4	C2
999	0.380528	Fe2 Na1 O12 Si4	C2
1600	0.385671	Na2 O12 Si4 V2	C2/c
.....			

COD

No.	R_w	formula	space_group
709	0.344690	Ga4 Na4 O24 Si8	C2/c
2935	0.344690	Ga4 Na4 O24 Si8	C2/c
2809	0.344713	Ga4 Na4 O24 Si8	C2/c
2810	0.345004	Ga4 Na4 O24 Si8	C2/c
2811	0.345617	Ga4 Na4 O24 Si8	C2/c
2813	0.345751	Ga4 Na4 O24 Si8	C2/c
2812	0.345756	Ga4 Na4 O24 Si8	C2/c
2983	0.347616	Fe4 Na4 O24 Si8	C2/c
2513	0.348362	Fe4 Na4 O24 Si8	C2/c
2512	0.350588	Fe4 Na4 O24 Si8	C2/c
.....			

PDF in the cloud



COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

BROOKHAVEN NATIONAL LABORATORY

HTTP://thebillingegroup.com

What would a machine-readable journal look like?

- It would look like a database!
- Why do we need it, don't we already have databases?
- We do, but they are limited (and despite their limitations, they are so powerful!)
- Imagine if every table in every paper in the literature could be read (and understood) by a computer
- And if every plot could be viewed (in .pdf or .png format) by a human, but the data and plot metadata (so it knows what it is a plot of) could be read by a machine
- We could do so much more science!

Proposal

- IUCr has a tradition of taking the lead in scientific computing and scientific information science (CIF!)
- Let's do it again!
- Prototype a machine-readable journal in the IUCr family
- There are many hurdles, including some we don't know about yet, but if we don't start the journey we will never get there.

Random thoughts I:
How do we get there

1. Database infrastructure
2. Schema that will grow, mechanism to grow it
3. REST-API
4. Browser access
5. Load all the IUCr back-catalog of cifs in there, or partner with databases to pull from them?
6. Infrastructure to deliver cif data as documents (e.g., json)
7. Infrastructure to capture usage

Vision: Raw data, tables and figures available through the API

- Create an infrastructure that
  - Makes it easy to upload crystallographic structure factors and measured intensity data
  - Creates tables from data in different forms (excel spreadsheets, web-forms, python pandas dataframes or numpy arrays)
  - Similarly for figures
  - Metadata schema to capture as much information about the experiments/calculations as possible
  - Access to these through the API

Finally, let's return to the 3 pillars

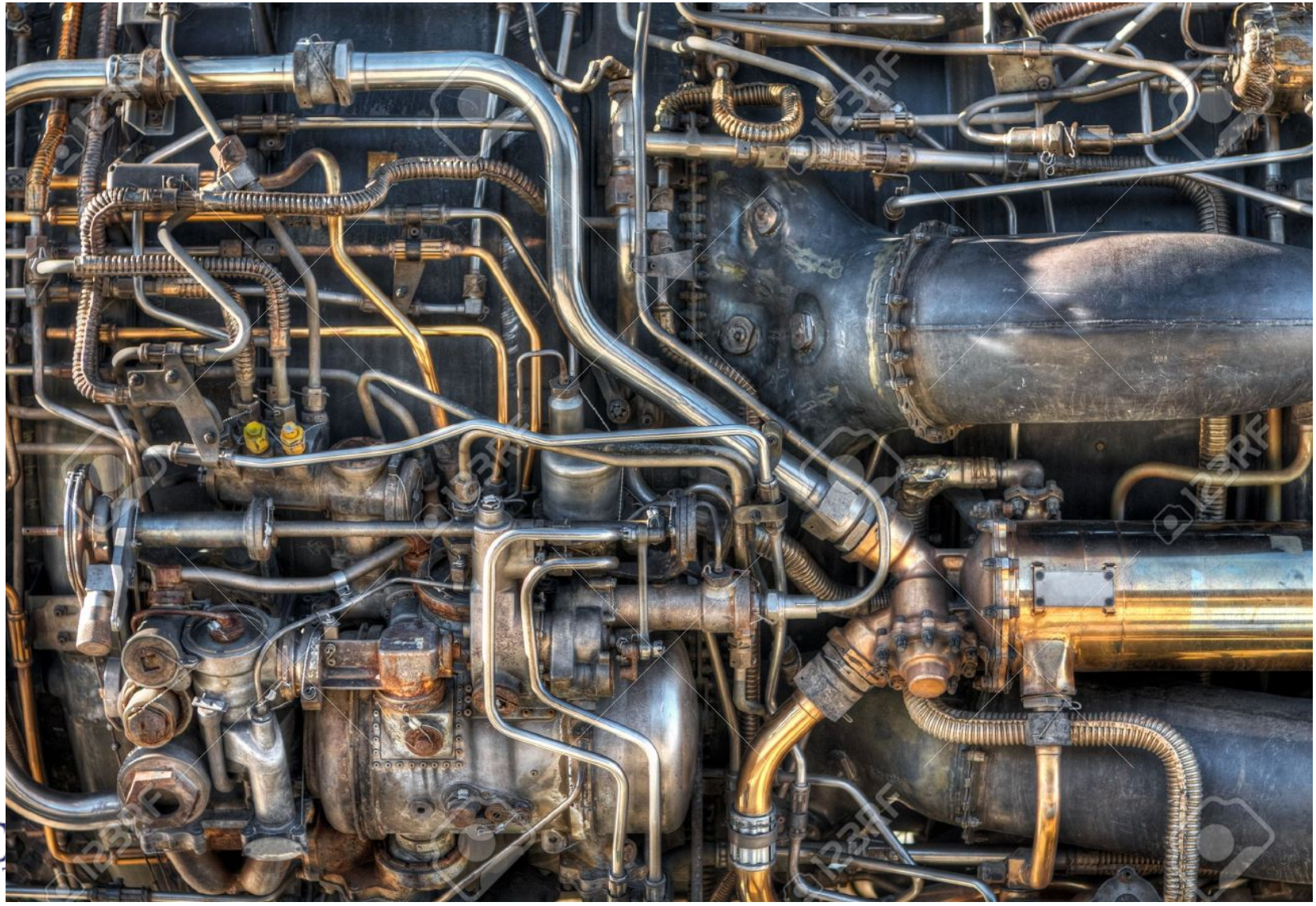
1. Disclosure
   - Creating an environment where people freely share (disclose) their ideas
     - Citations
2. Qualitation
   - (needed an English word for the act of assuring quality, sorry)
   - Ensuring some level of quality of the
     - peer review
     - reproducibility of work)
3. Curation
   - Two aspects to this
     - Persistence of the information
     - Access to the information

How would HMR literature impact these?

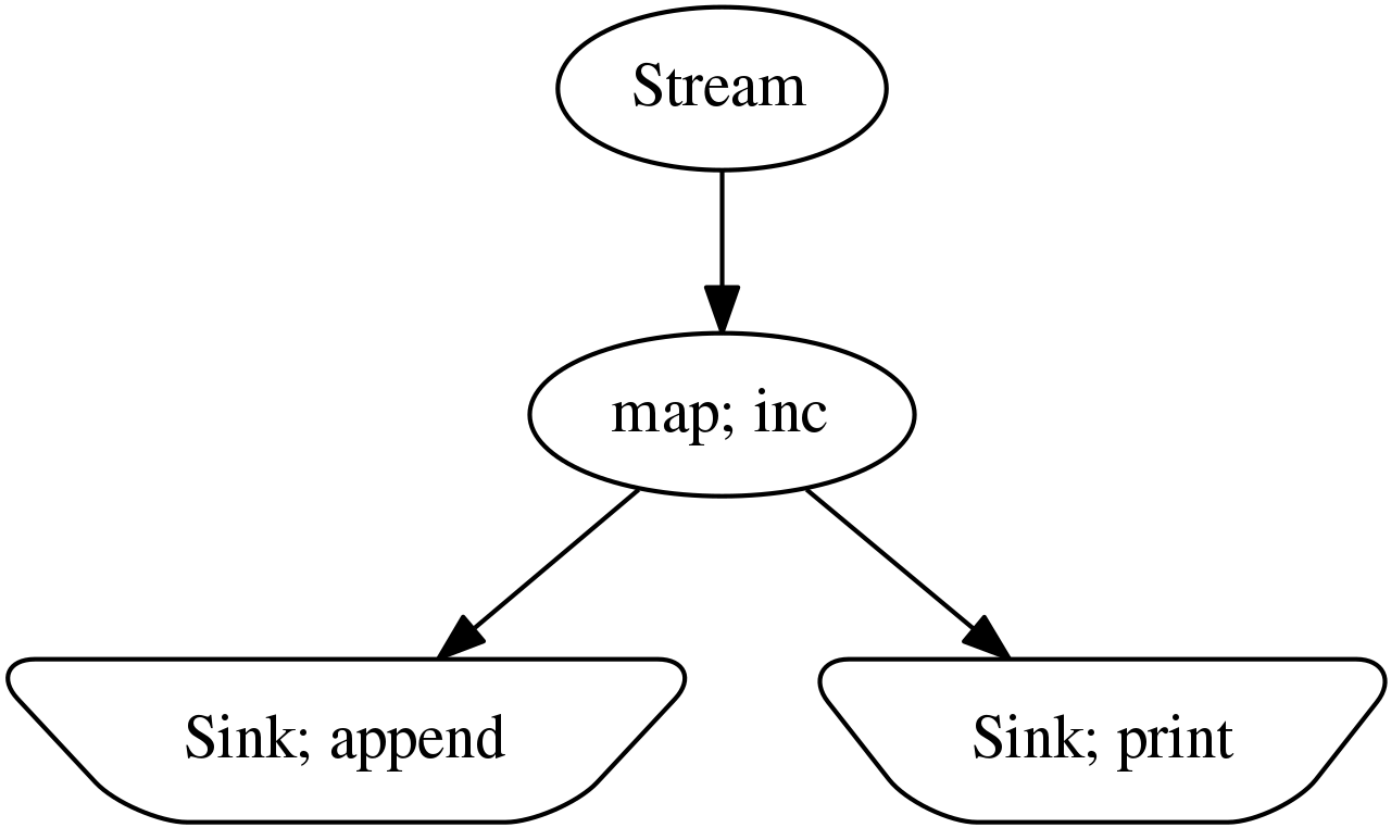
It will be disruptive, can we make it positively disruptive?

- Disclosure
  - Incentives to share data:
    - Citecoin? Get citecoin when someone pulls yo
  - Incentives to take the effort to put data into standard form:
    - Tools that make it easier (lower the cost)
- Qualitation
  - Much more validation on data because it is much easier to do on standardized data!
  - Easier to detect plagiarism and fraud? Use ML to flag suspicious data?
  - Capture the raw data, analyzed data, and the analysis

Data analysis/modeling as pipes

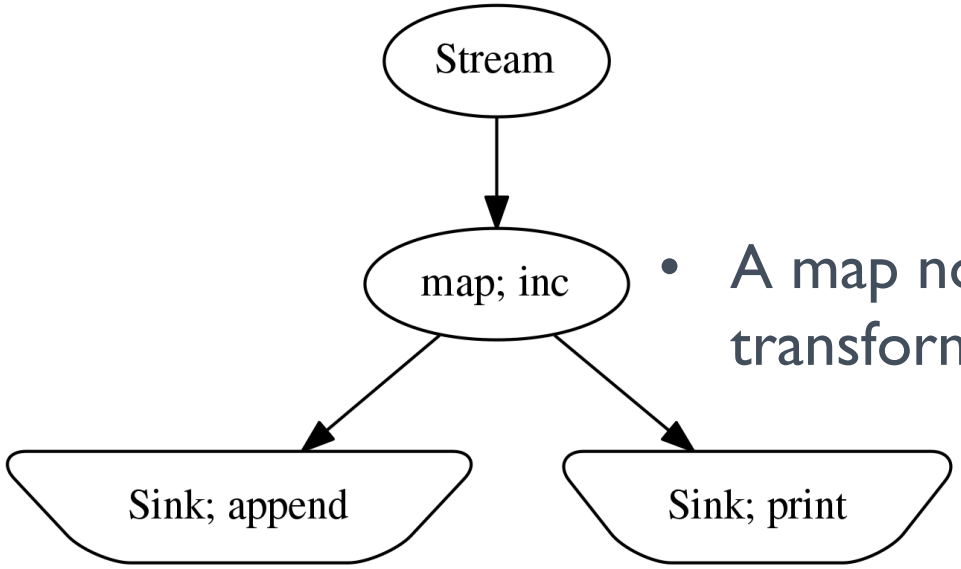


Data as streams through those pipes



Take inspiration from the Open Source Software community

- They have largely solved the problem of maintaining integrity of codes that are developed over time by large groups of people
  - Versioning software now very sophisticated (Git)
  - Workflow software and workflows now sophisticated (GitHub, release management, continuous integration)



Save the complete analysis!

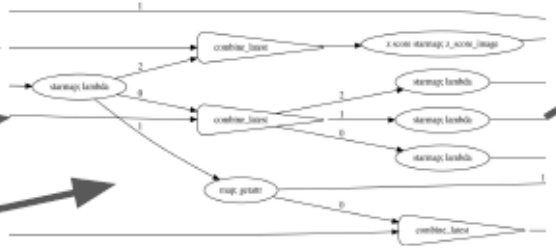
- A map node carries out a transformation on the data

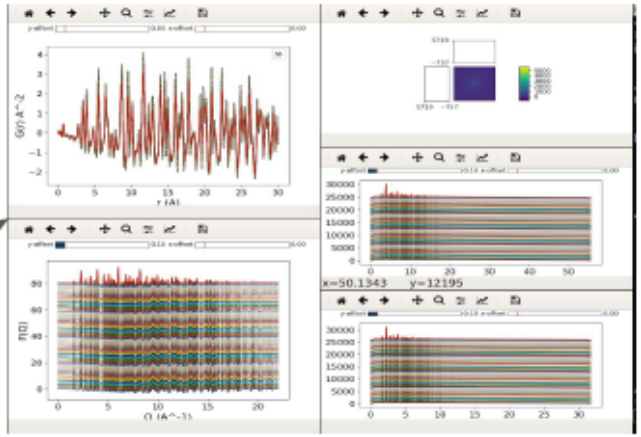
- Save pipeline in a database with a uuid
- Users can search for this and pull it from the db and rebuild it using properties of CI
- Users can adapt it then resave it to the database (with a NEW uuid)

- Analyses can be "published" by linking a doi to the uuid
- (uuid is universal unique ID)

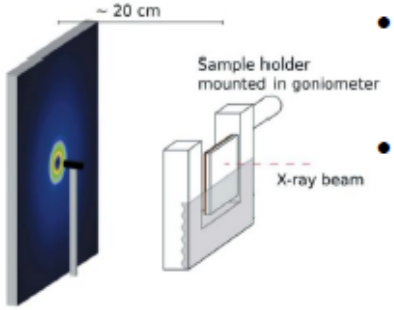
xpdAn/xpdtools

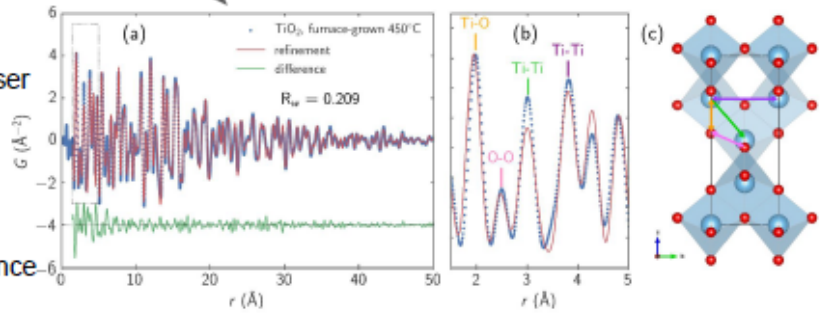
Data and metadata capture, combined with streaming data analysis pipelines, produces live data visualization and structure refinement.





- High quality data and metadata, are required for data mining, machine learning and automation
- The software stack is designed to be
  - Easy to use and automated
  - Capture rich and accurate metadata with minimal user overhead
- Software stack (GitHub and conda) includes:
  - **xpdAcq** for data/metadata acquisition
  - **xpdAn** for live data analysis
  - **xpdtools** for analysis tools
  - **SHED** for streaming the data analysis with provenance





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