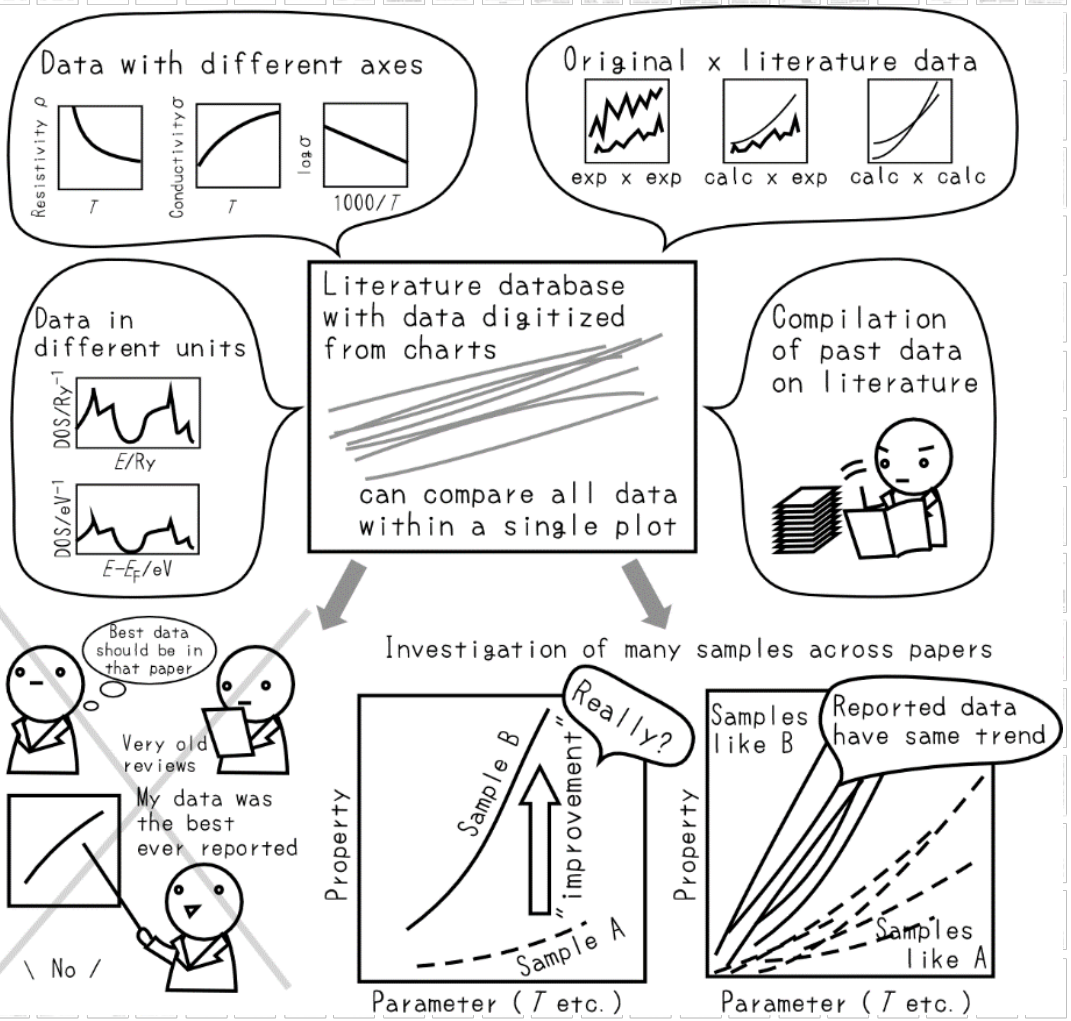
Starrydata2:

A plot mining web system for literature data collection

Yukari Katsura[1,2]*, Masaya Kumagai[3], Takushi Kodani[1,2], Hideyasu Ouchi[1,2], Sakiko Gunji[2], Yuki Ando[2], Yoji Imai[3], Kaoru Kimura[1] and Koji Tsuda[1,2,3]

[1]Univ. of Tokyo [2]MaDIS-MI²I, NIMS [3]RIKEN AIP

Key for the next MI: Experimental data from thousands of past papers



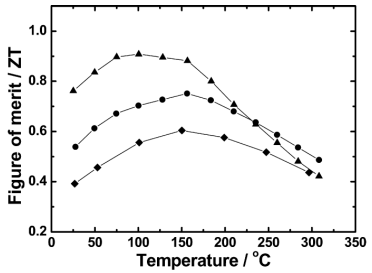
- Search engines for all researchers
- Data source for materials informatics

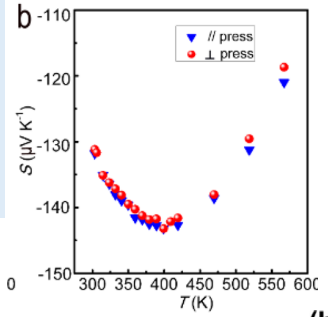
Recovery of experimental data from various plots

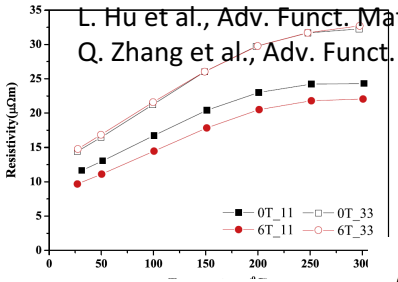
Tasks:

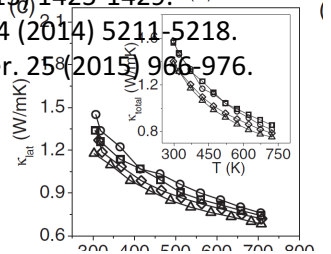
1. Plot digitization from images
2. Sample description from text (chemical composition etc.)

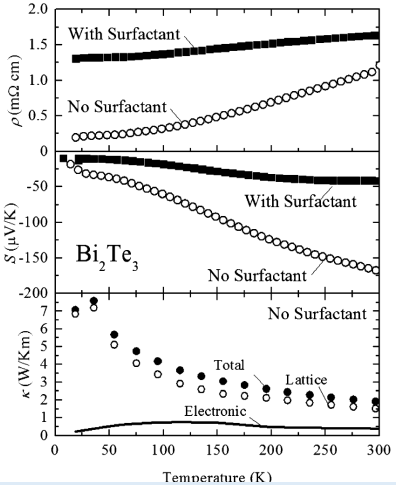
C. Kim et al., ACS Appl. Mater. Interfaces 4 (2012) 2949-2954.
V. Stavila et al., ACS Appl. Mater. Interfaces 5 (2013) 6678-6686.
L. Yang et al., ACS Appl. Mater. Interfaces 7 (2015) 23694-23699.
Y. Min et al., ACS Nano 9 (2015) 6843-6853.
D. Kim et al., Acta Materialia 59 (2011) 405-411.
D. Kim et al., Acta Materialia 59 (2011) 4957-4963.
L. Hu et al., Acta Materialia 60 (2012) 4431-4437.
K. Biswas et al., Adv. Energy Mater. 2 (2012) 634-638.
Y. Min et al., Adv. Mater. 25 (2013) 1425-1429.
L. Hu et al., Adv. Funct. Mater. 24 (2014) 5211-5218.
Q. Zhang et al., Adv. Funct. Mater. 25 (2015) 966-976.

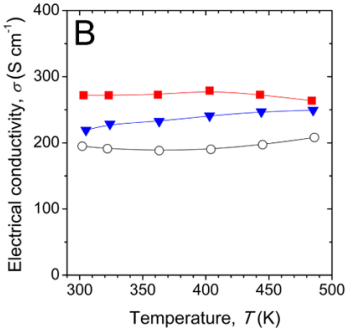


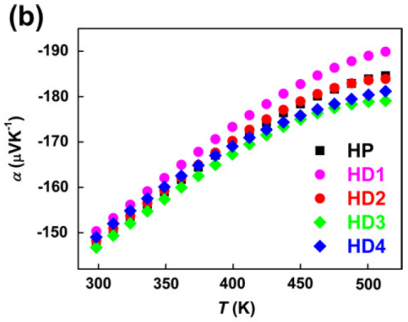


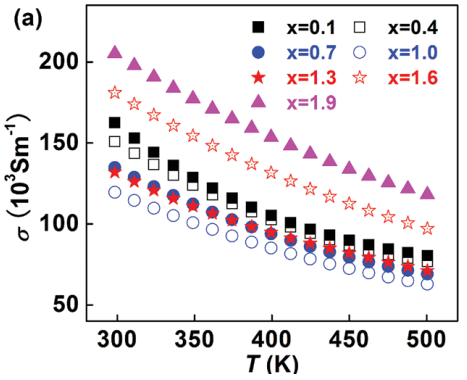


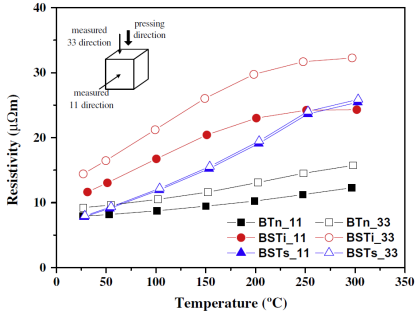


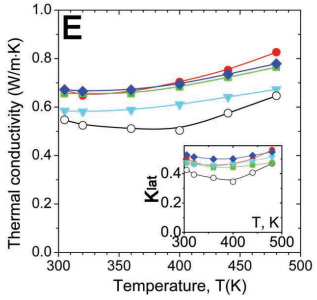


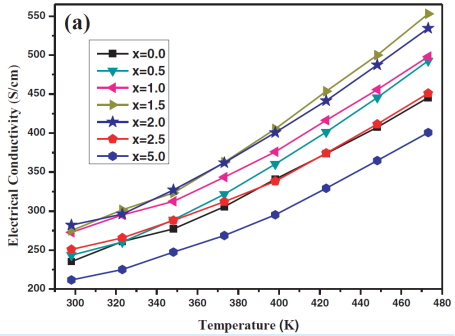






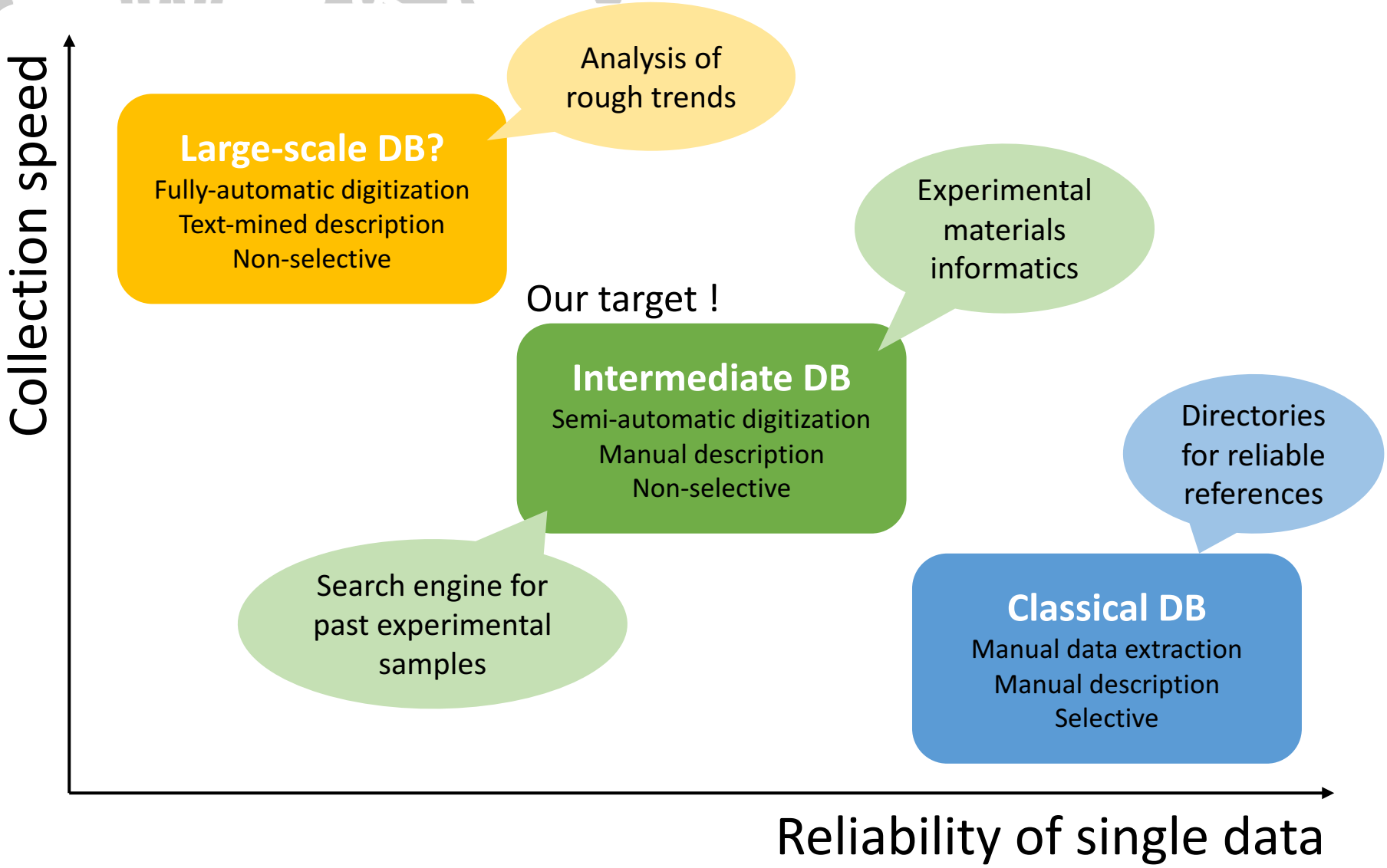






Time-consuming & boring for researchers → Developed a new web system

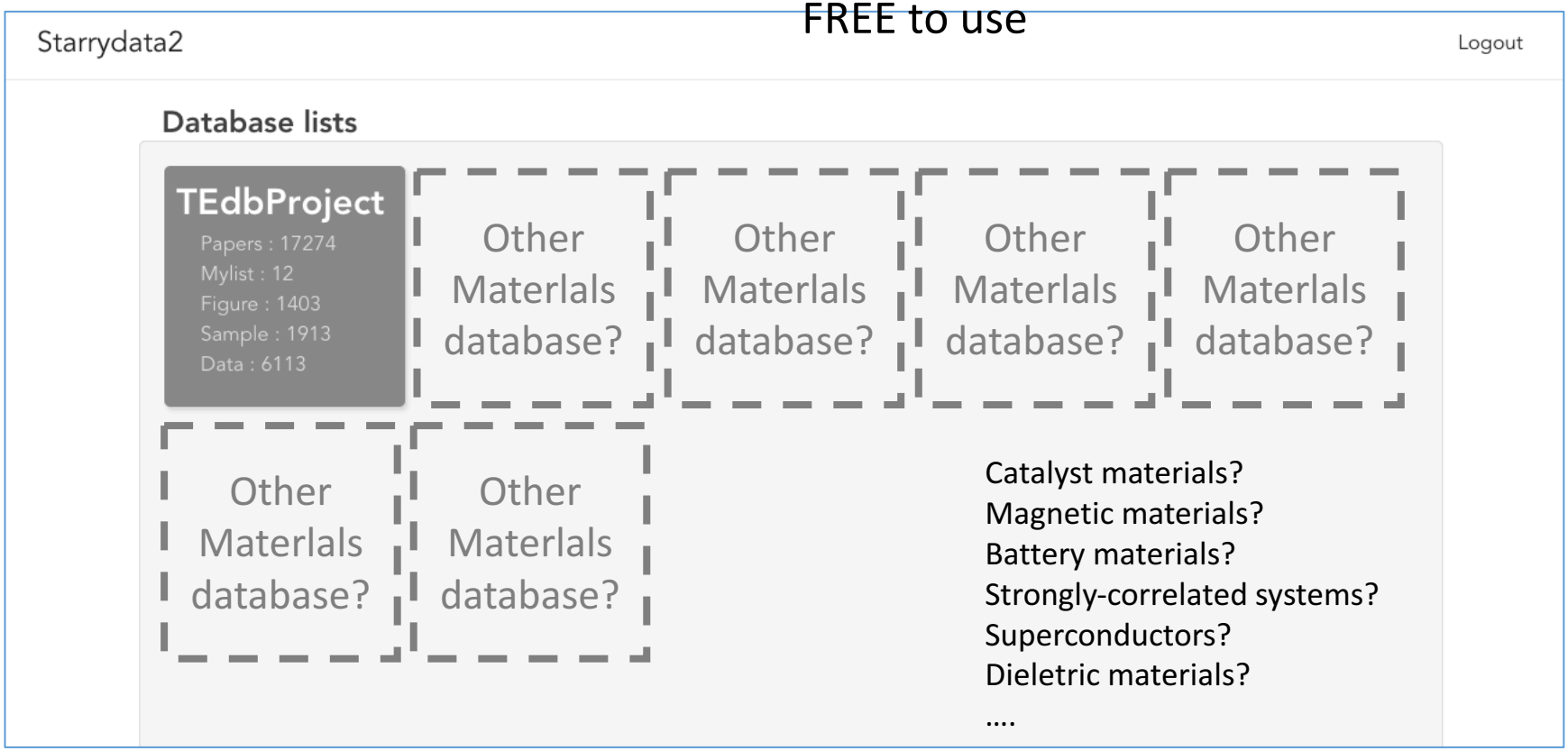
Target area of our experimental database



Starrydata2 web system: Online data-sharing platform for literature data

URL: http://starrydata2.org
Programming: Masaya Kumagai (RIKEN)
FREE to use

Database selection page



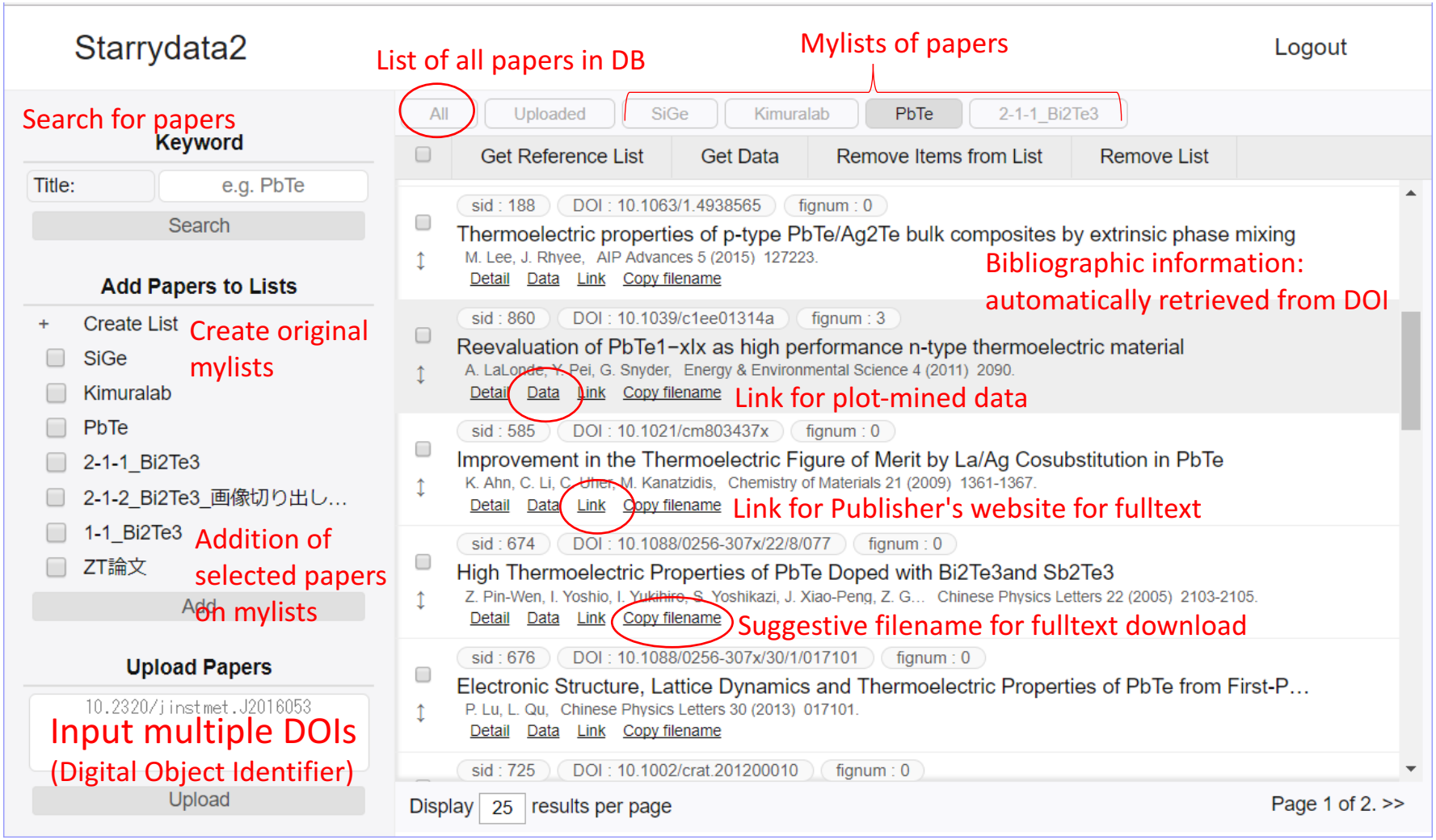
Prototype database: "TEdbProject"

A database of experimental thermoelectric properties

Starrydata2 web system : Index page

URL: http://starrydata2.org
Programming: Masaya Kumagai (RIKEN)

Designed as a tool for researchers to organize information about published papers



Starrydata2: Data-browsing page

All | Uploaded | PbTe系論文リスト | Aaa | Mg2Si | PbTe_

Get Reference List | Get Data | Get DOIList | Remove Item

sid : 6 | DOI : 10.1021/am405410e | fignum : 3

Exploration of Zn Resonance Levels and Thermoelectric Properties in I

P. Rawat, B. Paul, P. Banerji, ACS Applied Materials & Interfaces 6 (2014) 3995-4004.

Detail | Data | Link | Copy filename

①Click "Data"

sid : 31 | DOI : 10.1016/j.actamat.2008.10.005 | fignum : 0

Starrydata2 Logout

Browse / Add / Edit data

sid : 6 | DOI : 10.1021/am405410e | figure : 3

Exploration of Zn Resonance Levels and Thermoelectric Properties in I-Doped PbTe with ZnTe Nanostructures

P. Rawat, B. Paul, P. Banerji, ACS Applied Materials & Interfaces 6 (2014) 3995-4004.

Detail Link

Bibliographic information of original paper

Figures

No.	FigureNumber	Prop.X	Prop.Y
0	10(b)	Temperature	Thermal conductivity
1	3(a)	Temperature	Electrical resistivity
2	6(b)	Temperature	Seebeck coefficient

List of figures with datasets

Samples

No.	Samplename	Composition
0	Pb1.0015Zn0.02Te1.02I0.003	Pb1.0015Zn0.02Te1.02I0.003
1	Pb1.00075Zn0.01Te1.01I0.0015	Pb1.00075Zn0.01Te1.01I0.0015
2	Pb1.00025Zn0.01Te1.01I0.0005	Pb1.00025Zn0.01Te1.01I0.0005
3	Pb1.00015Zn0.01Te1.01I0.0003	Pb1.00015Zn0.01Te1.01I0.0003
4	Pb1.0015Zn0.01Te1.01I0.003	Pb1.0015Zn0.01Te1.01I0.003
5	Pb1.00025Zn0.02Te1.02I0.0005	Pb1.00025Zn0.02Te1.02I0.0005
6	Pb1.00075Zn0.02Te1.02I0.0015	Pb1.00075Zn0.02Te1.02I0.0015
7	Pb1.0005Zn0.02Te1.02I0.001	Pb1.0005Zn0.02Te1.02I0.001
8	Pb1.00015Zn0.02Te1.02I0.0003	Pb1.00015Zn0.02Te1.02I0.0003
9	Pb1.0005Zn0.01Te1.01I0.001	Pb1.0005Zn0.01Te1.01I0.001

List of samples reported in the paper

② Figure selector

③Sample selector

Convert Unit

Add new data

Figure: 3(a)

Caption: unknown

SampleName: Pb1.0015Zn0.02Te1.02···

Composition: Pb1.0015Zn0.02Te1.02

Property(x): Temperature

Property(y): Electrical resistivity

Unit(x): 10^ 0 K

Unit(y): 10^ 0 ohm*m

Save

XY-value

```
3.00000e+2, 4.20359e-6
3.24976e+2, 4.63473e-6
3.50289e+2, 5.28144e-6
3.74928e+2, 6.08982e-6
3.99904e+2, 7.00599e-6
4.24879e+2, 7.92216e-6
4.50193e+2, 8.94611e-6
4.75169e+2, 1.01317e-5
5.00145e+2, 1.14790e-5
5.25121e+2, 1.27186e-5
5.50096e+2, 1.41198e-5
5.75072e+2, 1.54671e-5
6.00048e+2, 1.71916e-5
6.25024e+2, 1.90778e-5
6.50000e+2, 2.07485e-5
```

④Digital data in SI unit

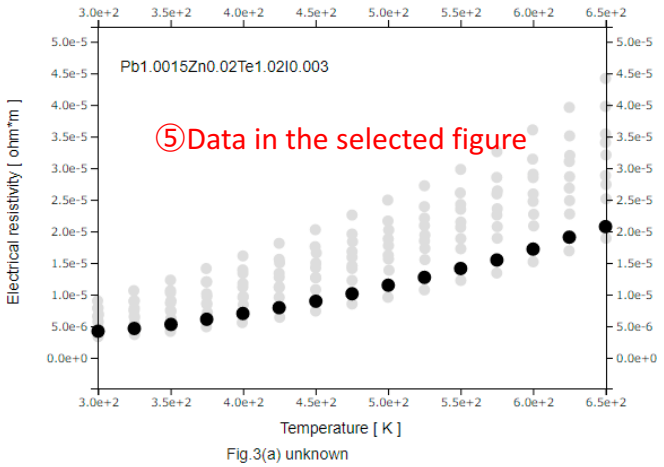


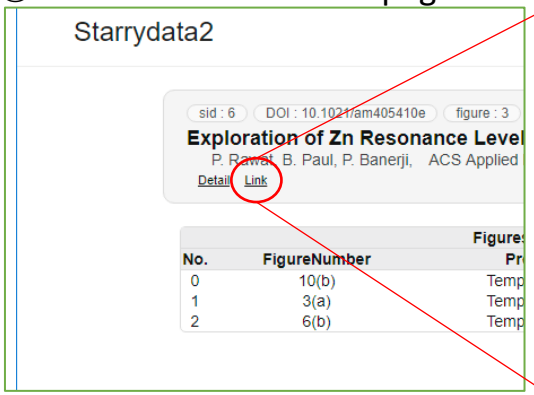
Fig.3(a) unknown

Automatic unit conversion
e.g. 10 mW/(cm*K)
→ 1 W/(m*K)

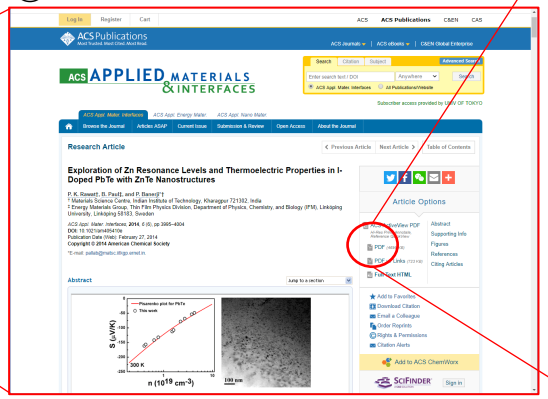
Plot-mining procedure:
(1)Get plot images

(Example in Google Chrome)

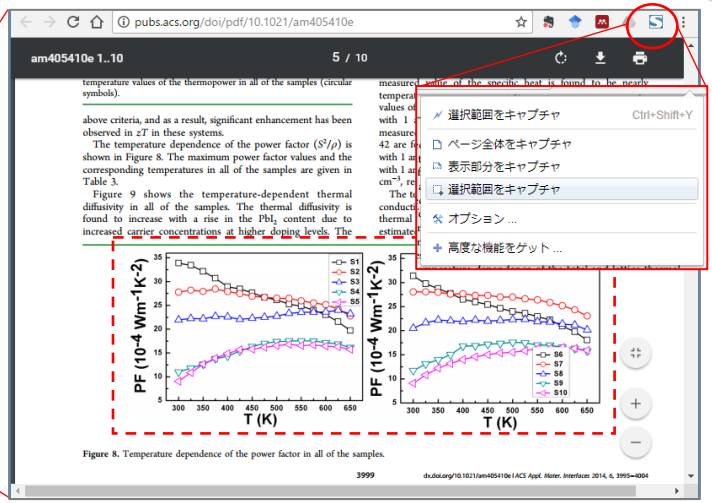
①Access Publisher's webpage



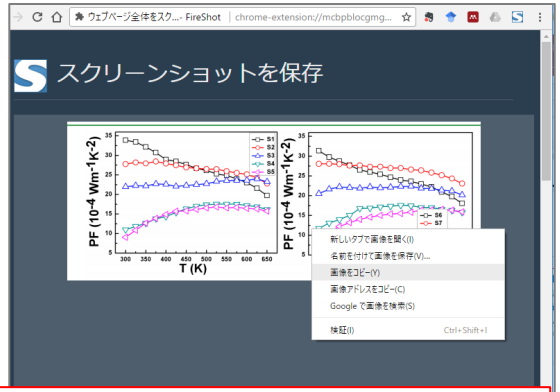
②Access Fulltext PDF



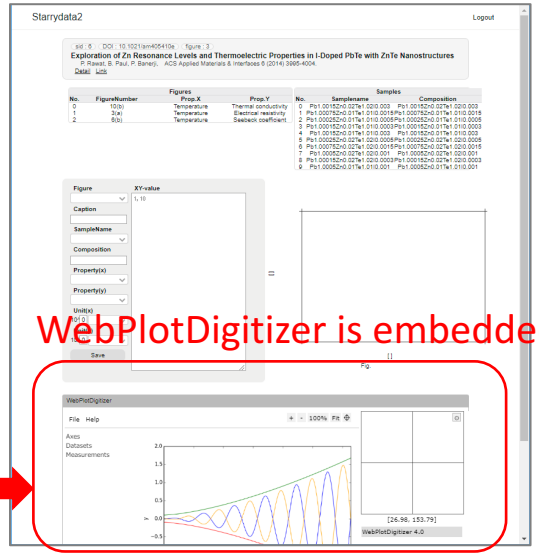
③Capture plot images
by Print Screen / FireShot / PDF reader etc.



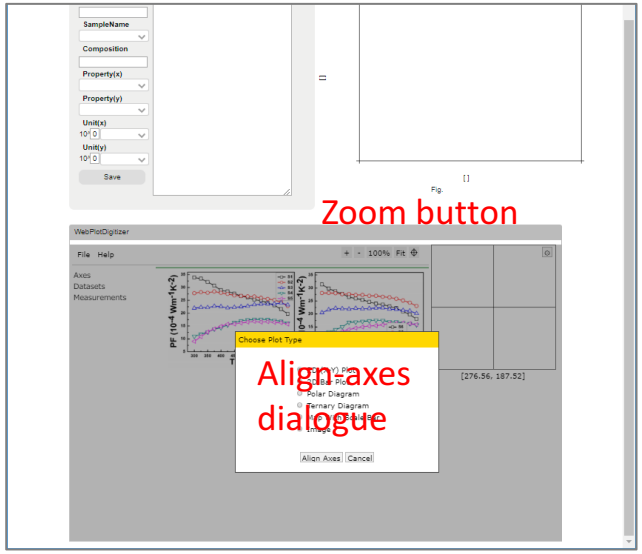
④Copy plot image



⑤Go to bottom of Starrydata 2



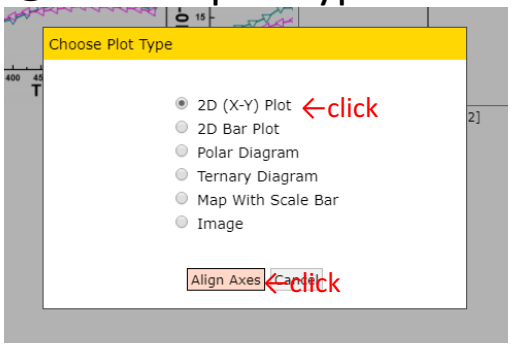
⑦Paste the plot image



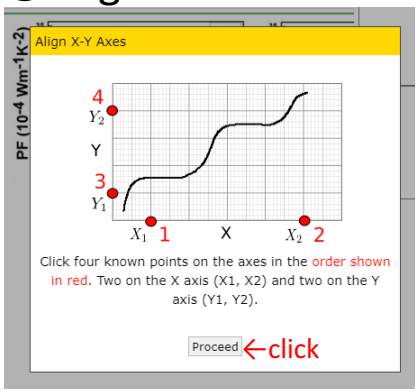
(2)Semi-automatic plot tracing by WebPlotDigitizer

WebPlotDigitizer
提供サイト: https://automeris.io Ankit Rohatgi
★★★★★ (127) 学術文献 ユーザー数: 18,941 人

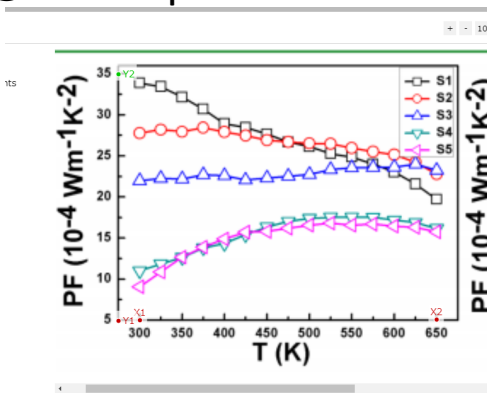
①Choose plot type



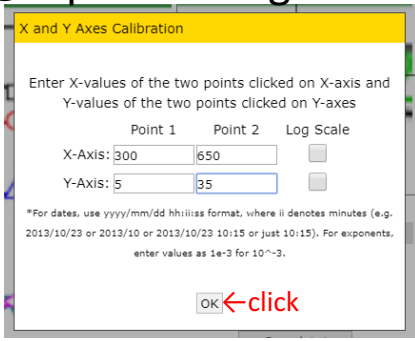
②Align X-Y Axes



③Click 4 points on axes

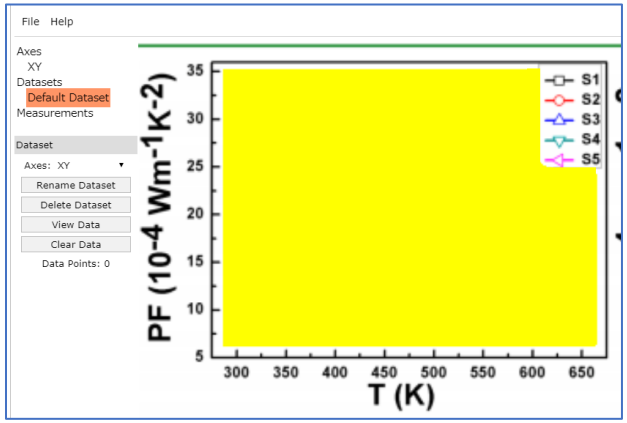


④Input readings



⑤Select tracing regions

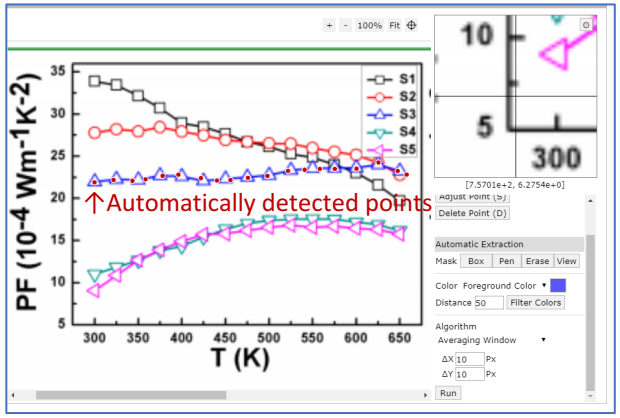
Box, Pen, Erase tools are available



Optional

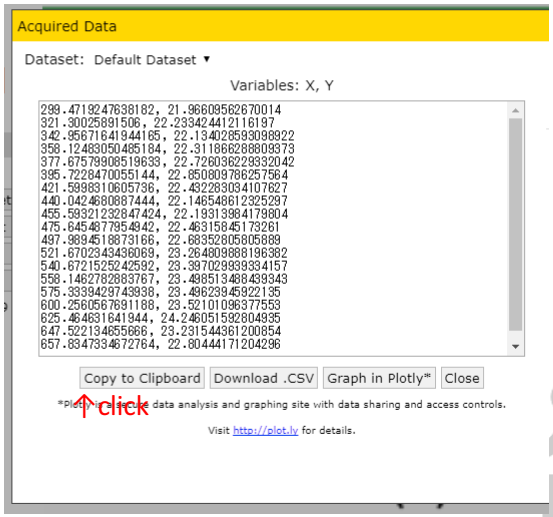
⑥Select color and Run

Select algorithm and tune parameters



Tracing by manual clicking is also possible

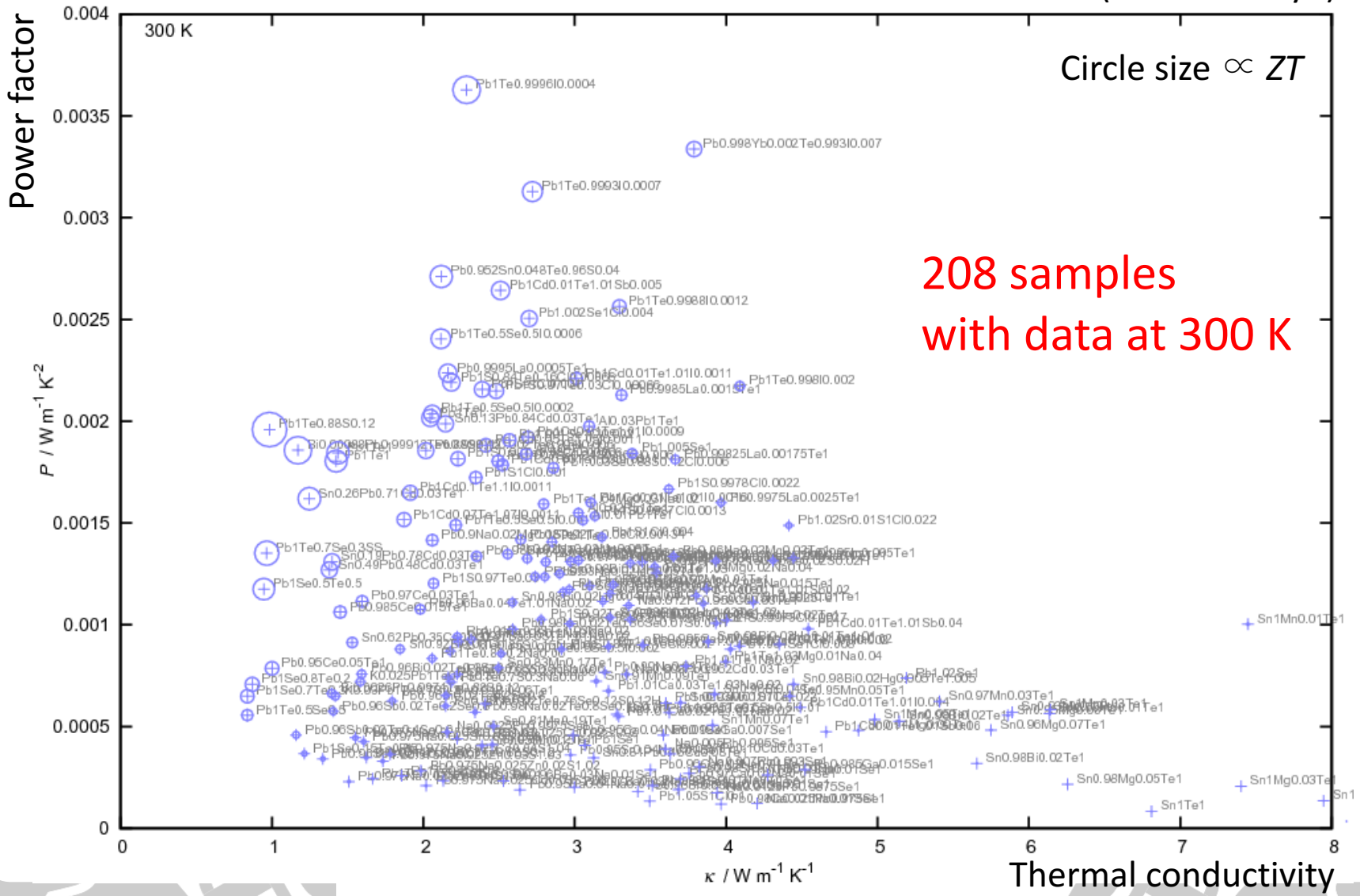
⑦View data and copy the digitized data



⑧Paste the digitized data to Starrydata2→Save

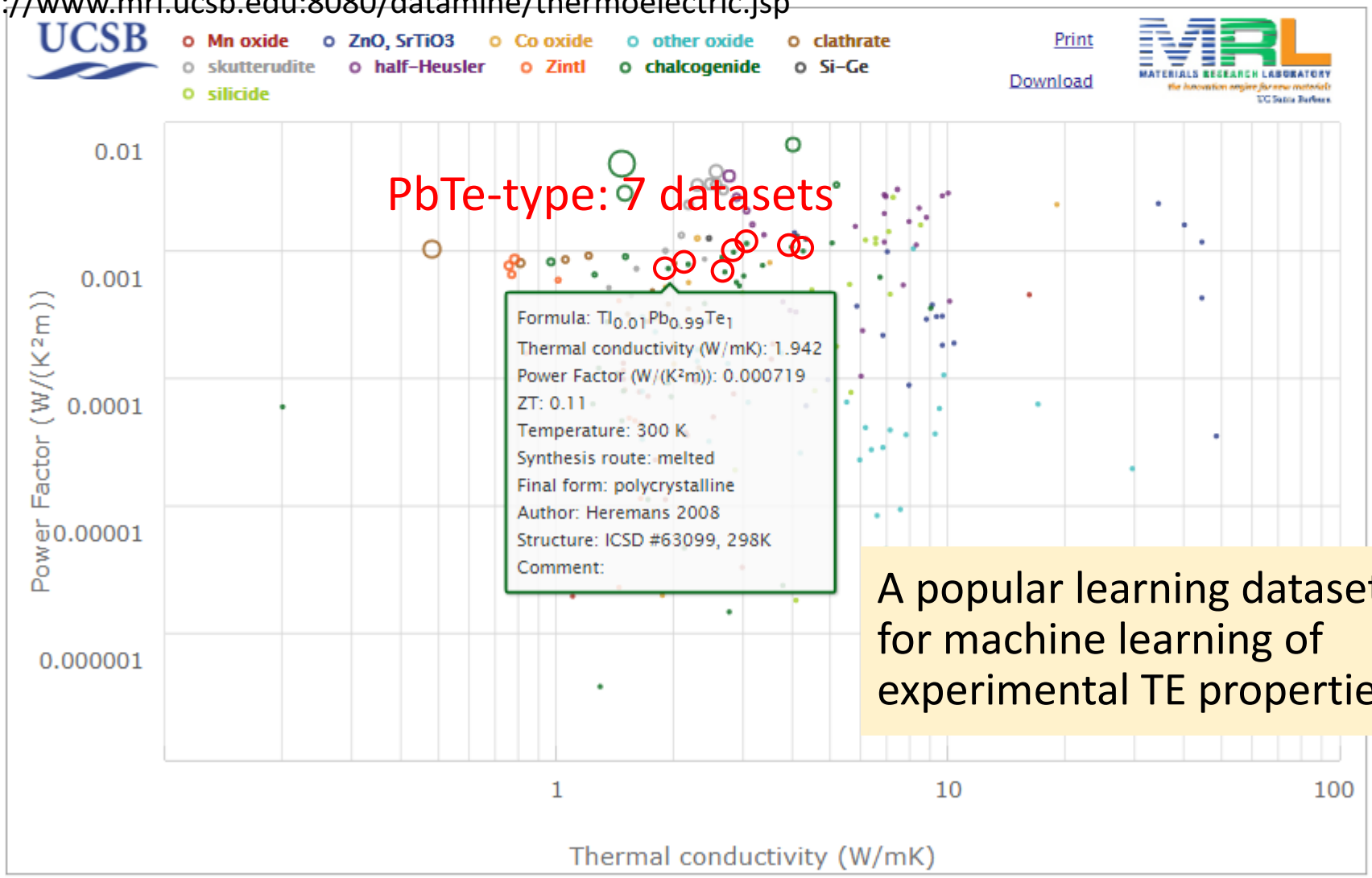
Recorded samples of PbTe-type thermoelectric materials

Data collection: Takushi Kodani (Univ. of Tokyo)



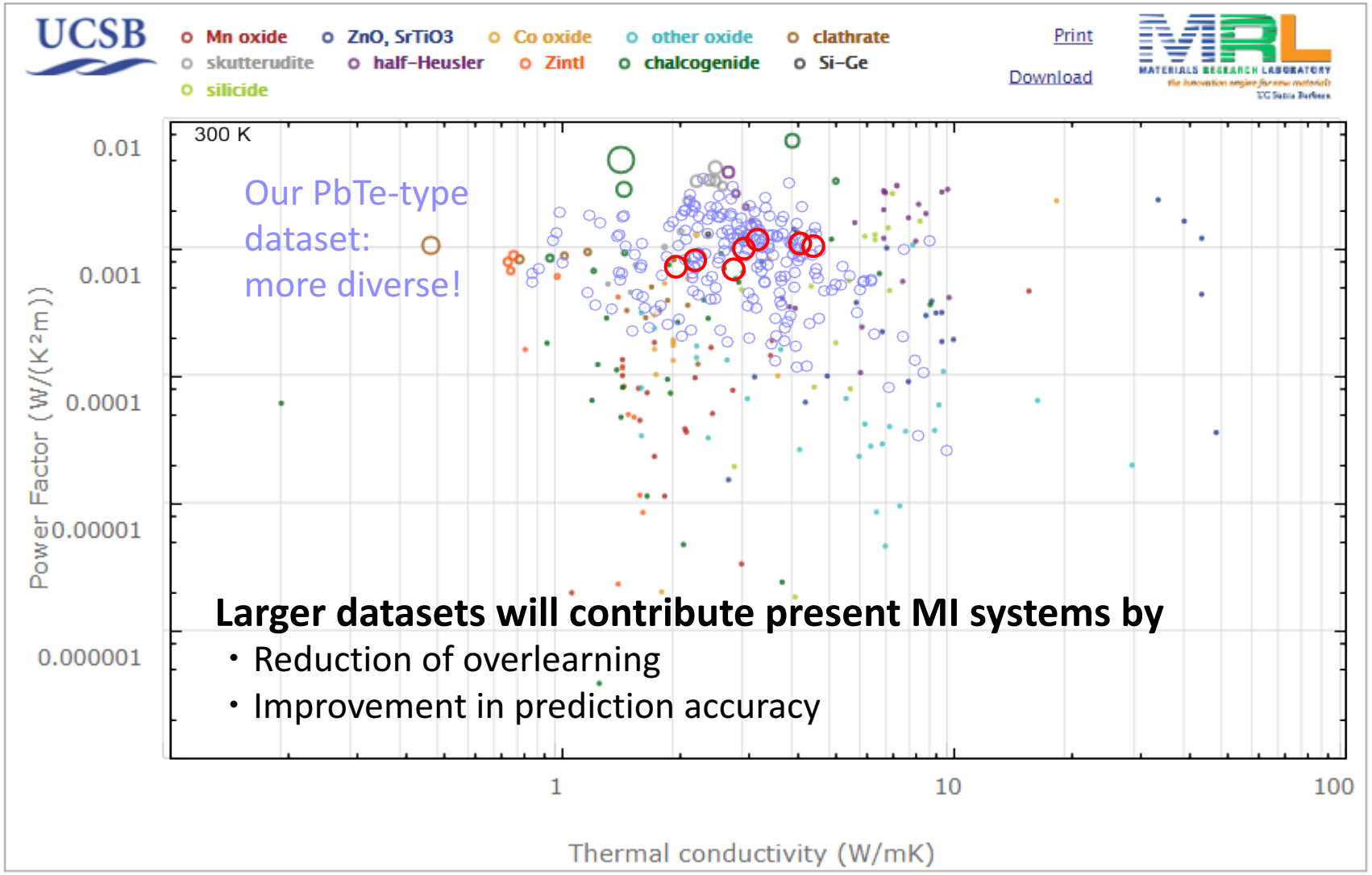
Comparison to present DB: UCSB-MRL datamining chart

M.W. Gaultois et al., Chemistry of Materials, 25 (2013) 2911.
http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp

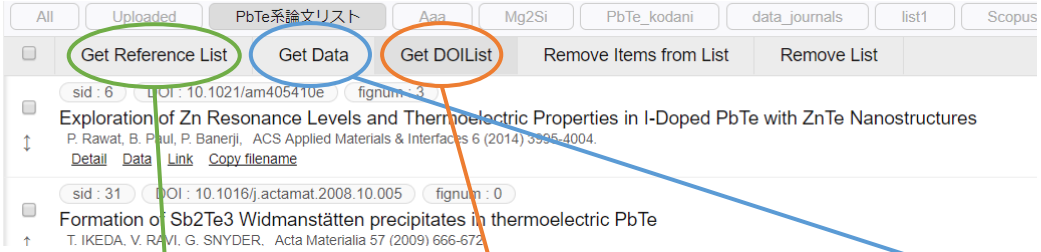


A popular learning dataset for machine learning of experimental TE properties

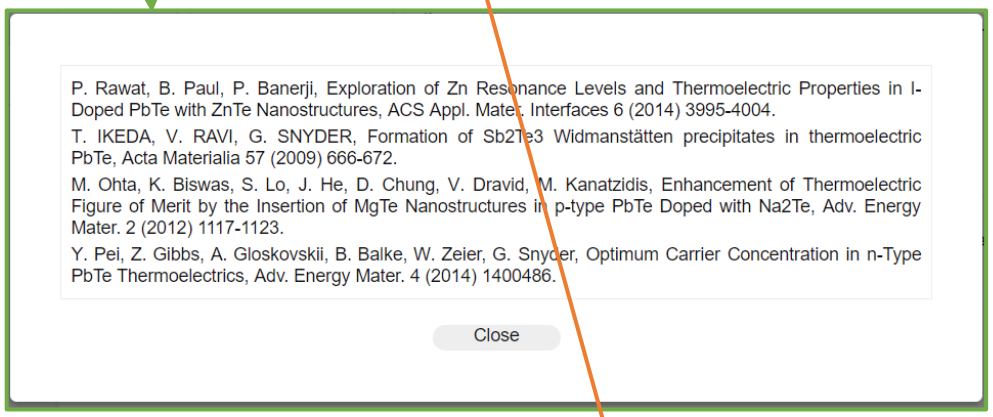
Comparison of our dataset with UCSB-MRL datamining chart



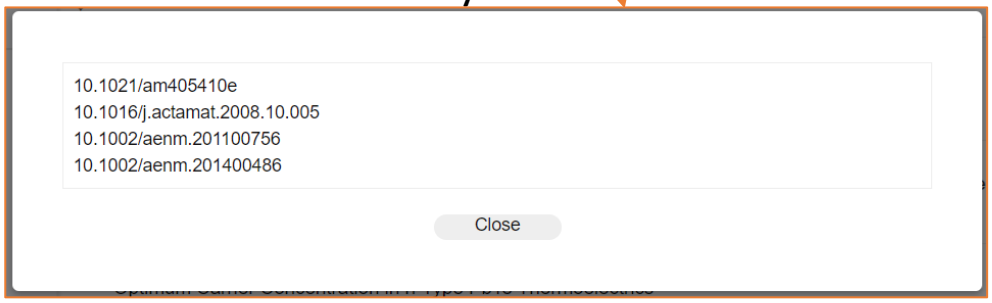
Starrydata 2 web system: Data export functions



Reference list to write papers



DOI list to share the mylist



A JSON file containing all datasets in the mylist

...
{'figureid': 14, 'paperid': 2, 'propertyid_x': 0, 'propertyid_y': 1, 'sampleid': 17, 'x': 1003.2921810699587, 'y': -7.749077490774911e-05}, {'figureid': 14, 'paperid': 2, 'propertyid_x': 0, 'propertyid_y': 1, 'sampleid': 18, 'x': 300.4115226337448, 'y': -0.0001188191881918819 5}, ...],
'sample': [
{'composition': 'Ba24Ga0Ag0Ge100', 'paperid': 0, 'sampleid': 0, 'samplename': 'x=0,y=0'},
{'composition': 'Ba24Ga4Ag0Ge96', 'paperid': 0, 'sampleid': 1, 'samplename': 'x=4,y=0'},
...

clathrate.json

149 papers / 623 figures / 890 samples

73281 data points

File size: 10 MB / Download time: ~3 min

Reading Starrydata2 data on Python

clathrate.json (10 MB)

```
import json
import pandas as pd
```

```
f=open('clathrate.json','r')
dict_all=json.load(f)

df_rawdata=pd.DataFrame(dict_all["rawdata"])
df_paper=pd.DataFrame(dict_all["paper"])
df_figure=pd.DataFrame(dict_all["figure"])
df_sample=pd.DataFrame(dict_all["sample"])
df_property=pd.DataFrame(dict_all["property"])
```

rawdata (all numeric)

	figureid	paperid	propertyid_x	propertyid_y	sampleid	x	y
73272	623	145	0	3	904	820.619060	1.257546
73273	623	145	0	3	904	837.799285	1.250562
73274	623	145	0	3	904	869.048740	1.250039
73275	623	145	0	3	904	894.034353	1.236228
73276	623	145	0	3	904	915.936873	1.262647
73277	623	145	0	3	904	933.138024	1.275752
73278						6259	1.295658
73279						1913	1.315485
73280						0148	1.335391
73281						0747	1.355166

sample

	composition	paperid	sampleid	samplename
884	Ba8Cu6Si16Ge24	140	884	4GPa
885	Ba8Cu6Si16Ge24	140	885	5GPa
886	Ba8Cu6Si40	141	886	Ba8Cu6Si40
887	Ba8Cu6Ge20Si20	141	887	Ba8Cu6Ge20Si20
888	Ba8Cu6Ge8Si32	142	888	x=0
889	Ba8Cu6Ge			
890	Ba8Cu6Ge1			

paper

	author	author_full	doi	issue	journal	journal_full	year
145	A. Saramat, G. S...	A. Saramat, G. S...	10.1063/1.2163979	2	Journal of Appli...	Journal of Appli...	
146	J. Martin, S. Er...	J. Martin, S. Er...	10.1063/1.2171775	4	Journal of Appli...	Journal of Appli...	
147	Chr						
148							
149	A.						

figure

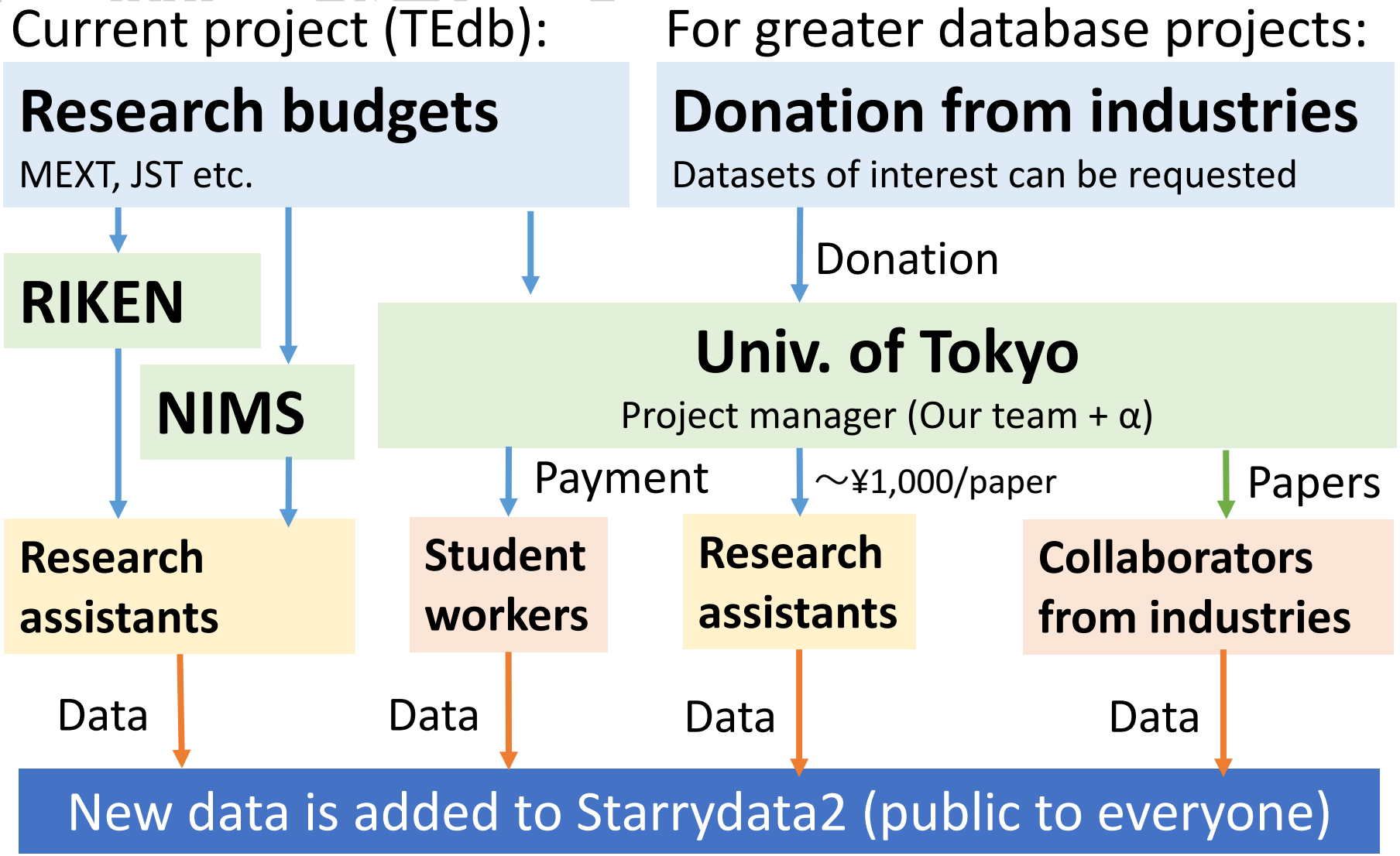
	caption	figureid	figurename	paperid
619	Total thermal conductivity of purified Ba8Ga16...	619	5	144
620	Figure of merit ZT of purified Ba8Ga16Ge30 sam...	620	6	144
621	a Seebeck coefficient S as a function of t...	621	2(a)	145
622	2. a Seebeck coefficient S as a function o...	622	2(b)	145
623	a Seebeck coefficient S as a function of t...	623	2(c)	145

property

	propertyid	propertyname	unit
0	0	Temperature	K
1	2	Electrical conductivity	ohm^(-1)*m^(-1)
2	1	Seebeck coefficient	V*K^(-1)
3	5	Power factor	W*m^(-1)*K^(-2)
4	3	Thermal conductivity	W*m^(-1)*K^(-1)
5	8	Dimensinoless figure of merit	-
6	4	Electrical resistivity	ohm*m
7	7	Inverse temperature	K**(-1)

Possible collaborations

To use the electronic journal systems, the project has to be **non-commercial.**



Summary

Starrydata2 web system (FREE)

- For efficient collection and sharing of data from plot images on literature
- Includes data export system for materials informatics

Prototype database: TEdbProject (FREE)

- Sample-based database of experimental thermoelectric properties
- 200～500 samples per material family (Largest experiental dataset)

URL: http://starrydata2.org

for first use: http://starrydata2.org/signup

Manual page: https://sites.google.com/site/yukarisearch/starrydata

E-mail: katsura@phys.mm.t.u-tokyo.ac.jp