Machine vision for X-ray Imaging, Intelligent Search and AutoML

Dani Ushizima

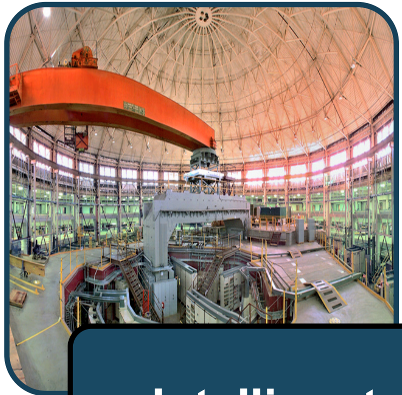
Staff Scientist, Affiliate Faculty

Computational Research Division, LBNL

Institute of Computational Health Sciences, UC San Francisco

Imaging Facilities

Machine vision



Intelligent Search

Auto-ML

DOE National User Facilities @LBNL

More than 10,000 researchers a year use these facilities.

ESnet

Molecular Foundry

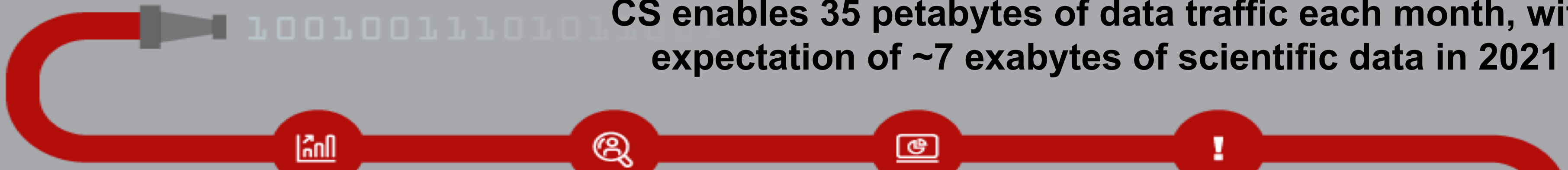
ALS

JGI

JOINT GENOME INSTITUTE

NERSC

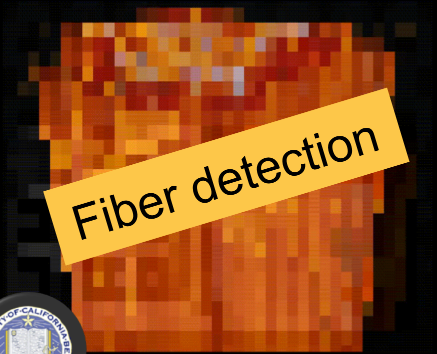
CS enables 35 petabytes of data traffic each month, with expectation of ~7 exabytes of scientific data in 2021



Machine Vision and Metrology for Materials

Analyze microstructure of materials to advance manufacturing

Ceramic matrix composites



Robert Ritchie

Yan Gao

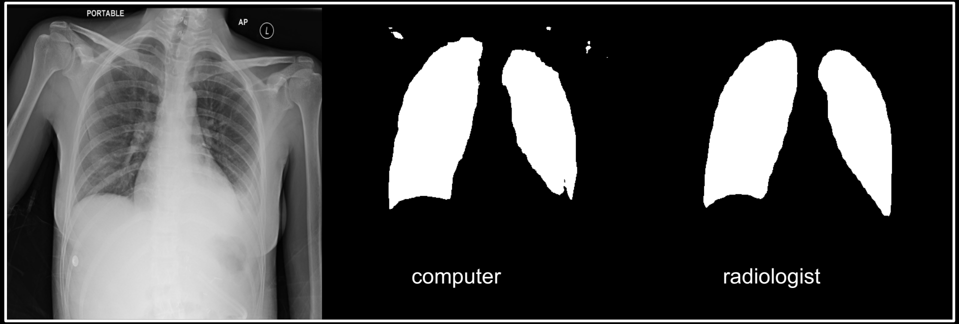
Carbon textiles

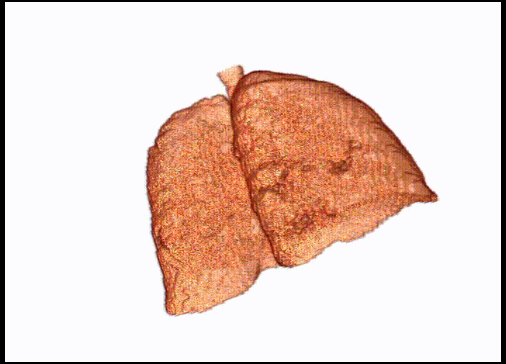
Francesco Panerai

Roman concrete

Paulo Monteiro

LDRD ACTS: lung scans of Covid-19 patients





Chest X-ray

2D

Broadly available

Projection of 3D

Bones muddle the image

Computed tomography

3D

Restricted availability

Details at sub-mm scale

Bones are “erased” digitally

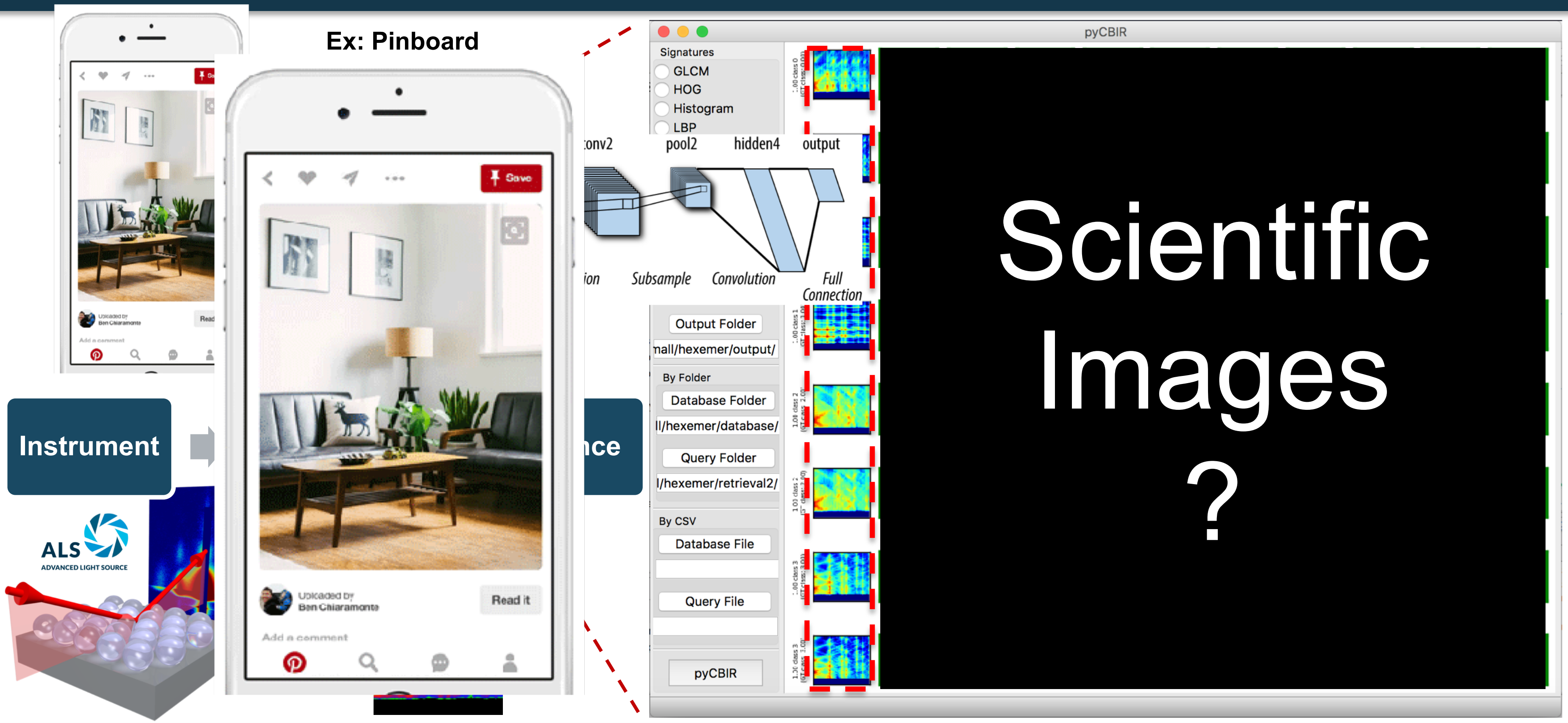
Imaging Facilities

Machine vision

Intelligent Search

Auto-ML

Deep learning to search patterns by similarity



Araujo, Silva, **Ushizima**, Medeiros, Hexemer, Parkinson, Carneiro, Bale, "Reverse Image Search for Scientific Data within and beyond the Visible Spectrum", **Expert Systems with Applications, 2018.**

GISAXS and CNN

- **Can we train a CNN to aid in the classification of new samples in real time?**
  - Fast classification and feedback at beamtime;
  - Disregard poor data if it doesn't hold useful information.
- **Focused on unit cell classification:**
  - 7 classes, ≠ resolutions and noise models;
  - Numerous ML architectures;
  - Simulated data.

Shuai Liu and Nathan Melton

Liu, Melton, Venkatakrishnam, Pandolfi, Freychet, Kumar, Tang, Hexemer, Ushizima, "Convolutional Neural Networks for Grazing Incidence X-ray Scattering Patterns: Thin Film Structure Identification", Materials Research Society - Special Issue on Artificial Intelligence, pp.1-7, 2019.

GISAXS, SAXS, WAXS, GIWAXS experimental data

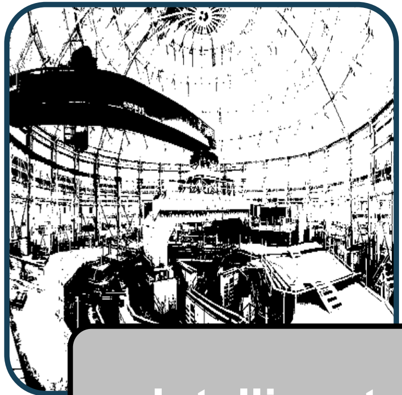
- **Can we recover metadata from experimental data?**
  - Choice of instrument mode is dynamic;
  - Experimental data is “harder” than simulation;
  - Moving million samples.
- **Focused on fast similarity search:**
  - Efficient similarity search and clustering of dense vectors;
  - KDTree, BallTree and MongoDB;
  - Auto-ML.

pyCBIR
image retrieval library in python

Imaging Facilities

Machine vision

Characterize



Intelligent Search

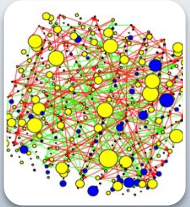
Link

Auto-ML

Steer

Auto-ML

Preprocessing
- cleaning
- encoding



Optimization
- feature extraction
- model selection

Deployment
- prediction
- interpretation

- Process of automating ML tasks to solve real-world problems;
- Covers the complete pipeline from the raw dataset to the deployable ML model;
- ML pipeline automation made of reusable parts = FAIR*.

*FAIR = findable, accessible, interpretable and reproducible

Who owns Auto-ML?

pyCBIR
image retrieval library in python

Google AutoML

AutoWeka
An Automated Data Mining Software Based on Weka

aws

H_2O AutoML

Azure No-Code AutoML

Auto-Sklearn

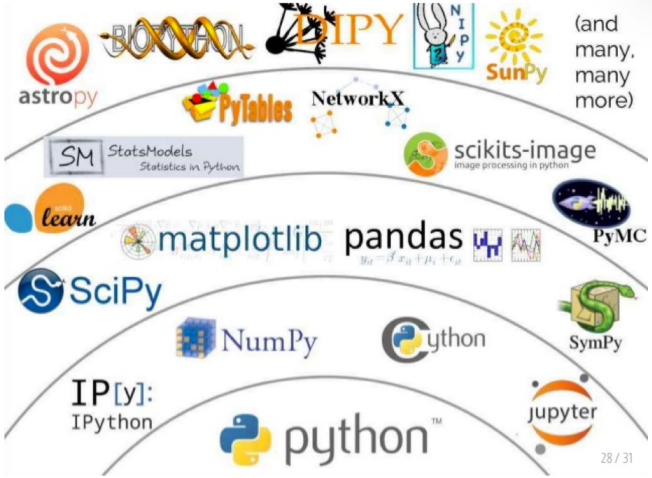
AutoKeras

BERKELEY LAB

*FAIR = findable, accessible, interpretable and reproducible

Hands-on with Python

1. Arrays with numpy
2. Picture with skimage
3. Volume with itkwidgets
4. ML with sklearn
5. Auto-ML



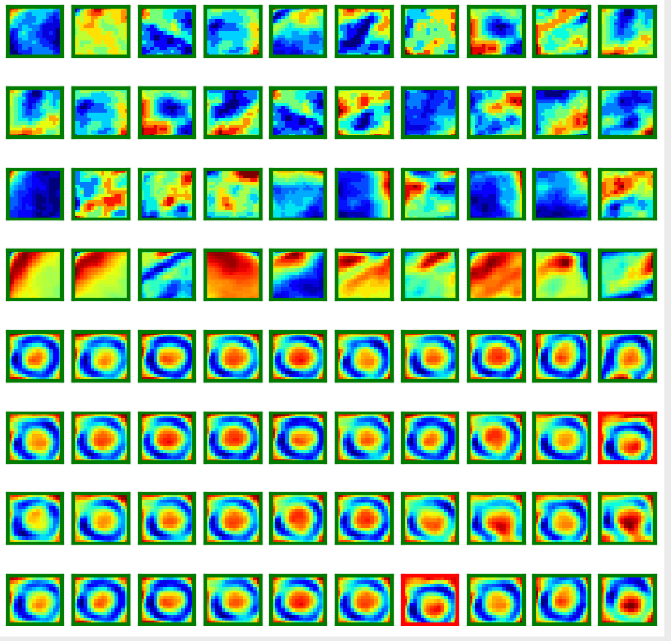
https://bit.ly/AI4ALS

AUTO-ML

pyCBIR

image retrieval library in python

CNN featurization



3. Compare Baseline

```
[*]: best_model = compare_models()
```

Processing:

Initiated	.... 14:28:09
Status	.... Fitting Fold 6 of 10
Estimator	.... CatBoost Classifier
ETC	.... 8.27 Minutes Remaining

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Quadratic Discriminant Analysis	0.9709	0.0	0.9667	0.9767	0.9696	0.9553	0.9592	0.0362
1	Random Forest Classifier	0.9527	0.0	0.9444	0.9646	0.9487	0.9271	0.9351	2.1587
2	K Neighbors Classifier	0.9518	0.0	0.9444	0.9620	0.9498	0.9262	0.9325	0.0166
3	Gradient Boosting Classifier	0.9518	0.0	0.9472	0.9644	0.9482	0.9261	0.9342	0.2904
4	Naive Bayes	0.9427	0.0	0.9417	0.9583	0.9407	0.9136	0.9225	0.0159
5	Decision Tree Classifier	0.9427	0.0	0.9361	0.9571	0.9387	0.9122	0.9215	0.0205
6	Logistic Regression	0.9336	0.0	0.9333	0.9474	0.9320	0.9000	0.9078	0.0704
7	Ada Boost Classifier	0.9327	0.0	0.9250	0.9446	0.9287	0.8968	0.9048	0.2755
8	Extra Trees Classifier	0.9327	0.0	0.9250	0.9446	0.9287	0.8968	0.9048	1.1128
9	Extreme Gradient Boosting	0.9327	0.0	0.9250	0.9496	0.9284	0.8970	0.9076	0.0670
10	Light Gradient Boosting Machine	0.9327	0.0	0.9250	0.9446	0.9287	0.8968	0.9048	0.0665
11	Linear Discriminant Analysis	0.8873	0.0	0.8861	0.9076	0.8852	0.8300	0.8413	0.0191
12	Ridge Classifier	0.8582	0.0	0.8583	0.8843	0.8558	0.7871	0.8014	0.0244
13	SVM - Linear Kernel	0.8264	0.0	0.8167	0.7710	0.7832	0.7260	0.7506	0.0157

Data types inferred automatically

Handles imputation

Time it for you

Tune hyperparams

Individual ML

Ensemble, Blend, Stack

Conclusion

Machine vision

Characterize

Intelligent Search

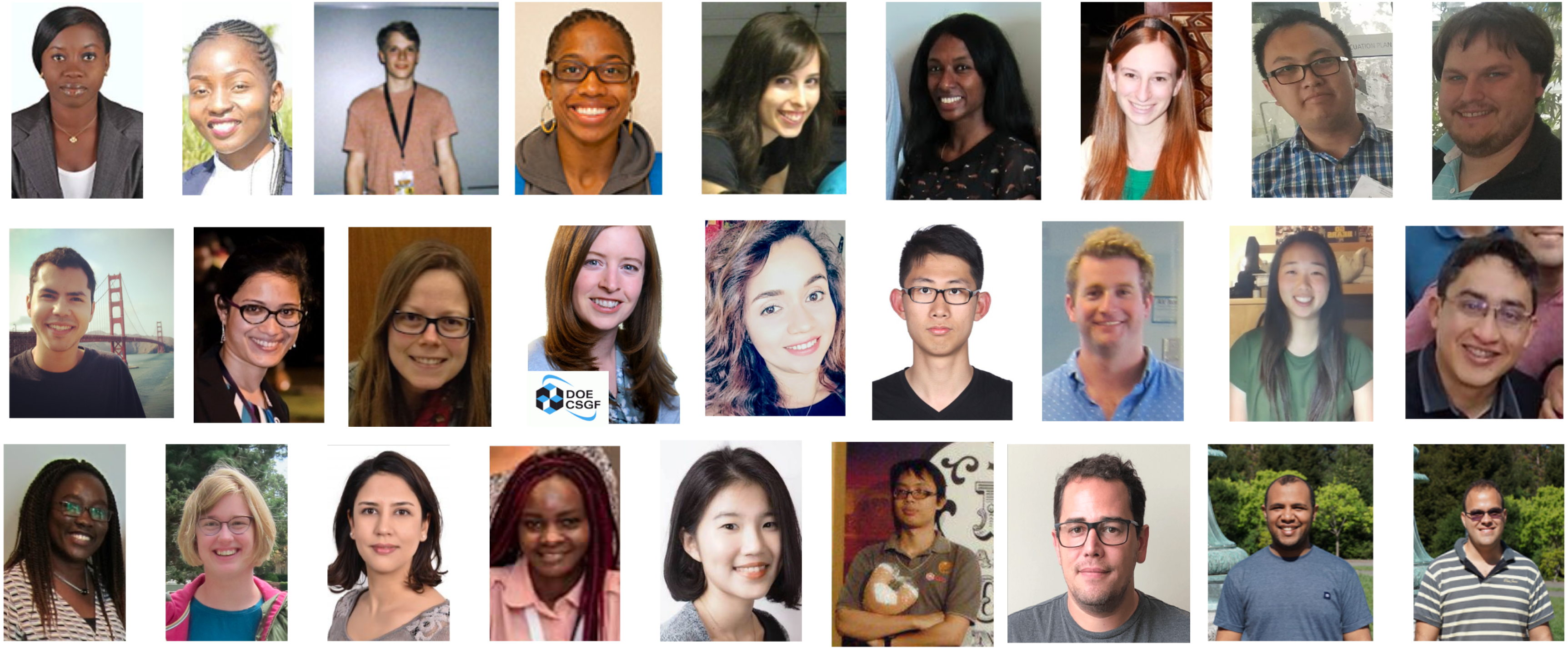
Link

Auto-ML

Steer

Cultivating emerging scientists

Research and development through collaboration with and support to the next generation of scientists and engineers around the world



BERKELEY LAB

IDEAL

Acknowledgements

LBNL	Wes Bethel , James Sethian, Chao Yang, Ann Almgren (CRD)
	Hari Krishnan, Alexander Hexemer, Dula Parkinson (ALS)
	Brent Helms (Molecular Foundry), Peter Ercius (NCEM), Nick Sauter (MBIB)
UCB	Fernando Perez, Stefan van der Walt, Paulo Monteiro, Hans Wenk
UCSF	Lea Grinberg, Tonya Pierges, Alexander Ehrenberg
UCSB	Peter Zok, Natalie Larson
ANL	Nick Schwarz, Doga Gursoy
ORNL	Singanallur Venkatakrishnan
SLAC	Xiaobiao Huang, Chris Tassone, Ryan Coffee
BNL	Kevin Yager, Tom Caswell
LANL	Aric Hagberg
VA	Duygu Tosun
GE:	Yan Gao, Anjali Singhal
IBM:	Alexandre Andreopoulos
Intel:	David J. Michalak
Microsoft:	Vani Mandava
NASA:	Nagi Mansour, Francesco Panerai (now UIUC)

DAV

Thank you

Thank you

https://bit.ly/AI4ALS

Machine vision for images across domains

	Instrument	Materials sample	Science discovery	New Analytics/ML
X-ray	**microCT** 3D - μm	CMC, carbon fiber, concrete, rocks, soil, archeological assets	Material deformation, search of experiments by microCT similarity	Detection, segmentation, classification
X-ray	**GISAXS** 2D - nm	Thin films - applications to gaseous sensors and piezoelectric devices	Categorization of million-scale databases	Classification
X-ray	**Crystalography** 2D - μm	Protein structure	Screening of diffraction patters containing Bragg peaks	Detection, Classification
X-ray	**CT** 3D - μm	Human brain	Evaluation of biomarkers (*locus ceruleous)* - correlation to Alzheimer's	Detection, segmentation, characterization
	MRI 3D - mm	Human brain	Data fusion with CT and histology for enhancement of clinical data	Detection, segmentation, classification
Electron	**SEM** 2D - nm	Nanocrystal frameworks for film design	Quantitative tools to drive architecture of colloidal nanocrystal films	Segmentation, classification
Electron	**STEM** 3D - nm	Microelectronics, concrete design	Lowest ever dielectric constant for PMO, new structure on cement shrinkage (foil)	Segmentation, characterization, classification

Recent Accomplishments

	Instrument	Selected Publications
X-ray	**microCT** 3D - µm	MacNeil, Ushizima, Panerai, Masour, Parkinson, *Interactive Volumetric Segmentation for Textile Microtomography Data using Wavelets and Non-local Means*, **Journal of Statistical Analysis and Data Mining** 2019.
X-ray	**GISAXS** 2D - nm	. Araujo, Silva, Ushizima, Parkinson, Hexemer, Carneiro, Medeiros, *Reverse Image Search for Scientific Data within and beyond the Visible Spectrum*, **Expert Systems with Applications**, 2018 . Ushizima, Araujo, Romuere, “Searchable datasets in Python: images across domains, experiments, algorithms and learning – pyCBIR”, pyData San Francisco 2016.
X-ray	**Crystalography** 2D - µm	Ke, Brewster, Yu, Yang, Ushizima, Sauter, *A Convolutional Neural Network-Based Screening Tool for X-ray Serial Crystallography*, **Journal of Synchrotron Radiation** 2018.
X-ray	**CT** 3D - µm	Alegro, Theofilas, Nguy, Castruita, Seeley, Ushizima, Grinberg, *Automating Cell Detection and Classification in Human Brain Fluorescent Microscopy Images Using Dictionary Learning and Sparse Coding*, **Journal of Neuroscience Methods**, 2017.
	MRI 3D - mm	
Electron	**SEM** 2D - nm	Williams, Ushizima, Zhu, Anders, Milliron, Helms, *Nearest-Neighbour Nanocrystal Bonding Dictates Framework Stability or Collapse in Colloidal Nanocrystal Frameworks*, **Chemical Communications**, Royal Society of Chemistry, 2017.
Electron	**STEM** 3D - nm	Ushizima, Bale, Bethel, Ercius, Helms, Krishnam, Grinberg, Haranczyk, Macdowell, Odziomek, Perciano, Parkinson, Ritchie, Yang. *IDEAL: Images across Domains, Experiments, Algorithms and Learning*, **Journal of Minerals, Metals and Materials**, 2016.