A TALE OF SCALES

MACHINE LEARNING FROM THE LAB BENCH TO THE SYNCHROTRON

SHEKAR VENKATESWARAN

- PhD in Computer Engineering, University of Massachusetts Dartmouth
- Computational Scientist at National Synchrotron Light Source II, Brookhaven National Lab (2022 – Present)
- Developer and machine learning researcher at Haverford college (2018 – 2022)
  - DARPA funded materials discovery project
- Small molecule drug screening ML models at Takeda Pharmaceuticals, Boston
- Pacific Northwest National Lab for dissertation work
  - Transportation network vulnerability assessment using modeling and simulation
- Software developer first, ML engineer second

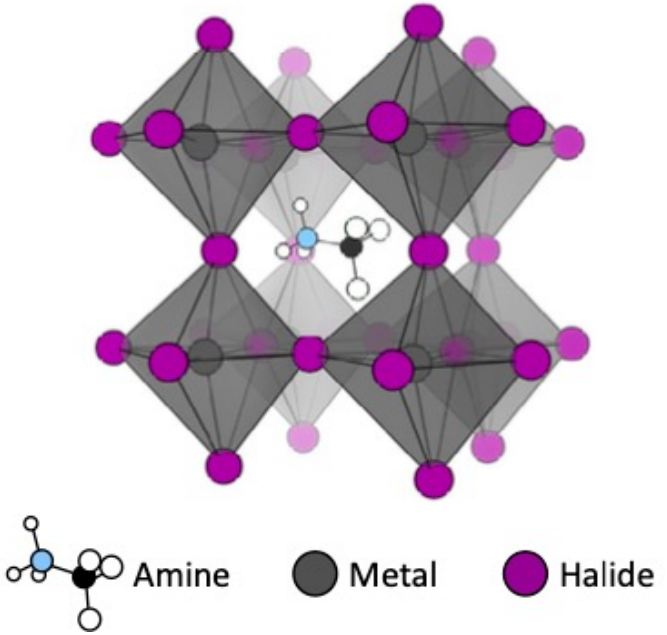
OUTLINE

- Small scale science at the lab bench
  - Applying ML techniques for high throughput materials discovery
  - Complexity of capturing and representing experimental data
  - Wrangling it for ML algorithms
  - The great ML model baking show!
- Large scale science at a Synchrotron light source
  - What is a Synchrotron
  - Data, data and more data!
  - Ways of managing and accessing
  - Opportunities to apply ML

BENCH SCALE SCIENCE

METAL HALIDE PEROVSKITES

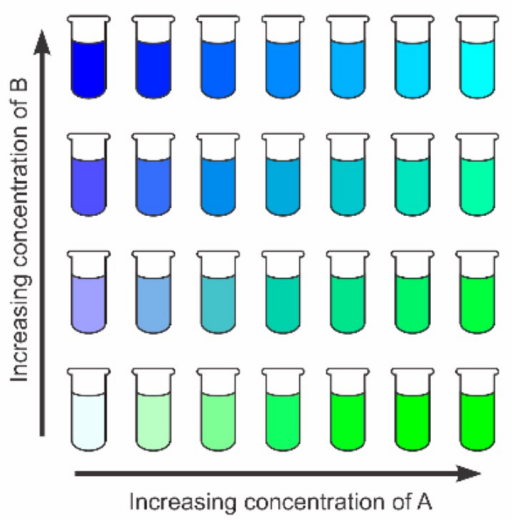
- Family of crystalline structures
- Emerging class of solar materials
- Reached efficiencies higher than Silicon photovoltaics (33% vs 20%)
- Cheaper to manufacture
- Can be synthesized as thin films
- What's the problem?
- Poor environmental stability/durability

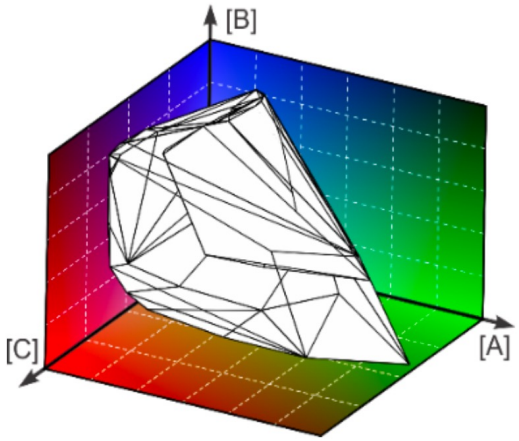


PIs: Dr. Sorelle Friedler, Dr. Joshua Schrier, Dr. Alex Norquist, Dr. Emory Chan (DARPA) under Contract No. HR001118C0036

HOW DO WE MAKE NEW PEROVSKITES?

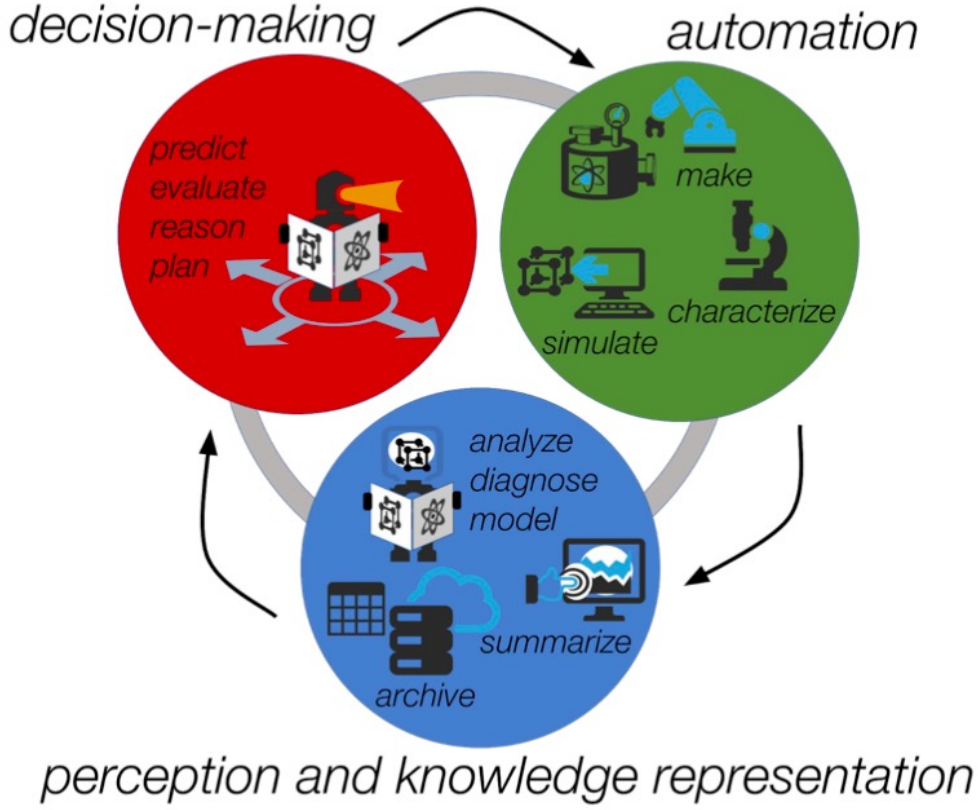
- Don't understand the physics to simulate
- 10^7 different experimental conditions
- To make good ML predictions one needs successes and failures
- Journals and scientists don't publish marginally successful or failed experiments
- "Dark data" improves models and gives insights into chemical processes





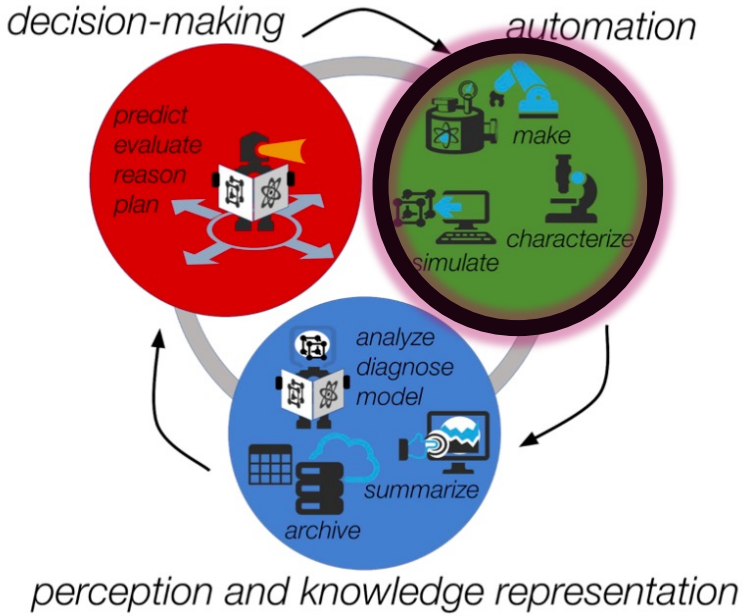
AUTONOMOUS RESEARCH LAB

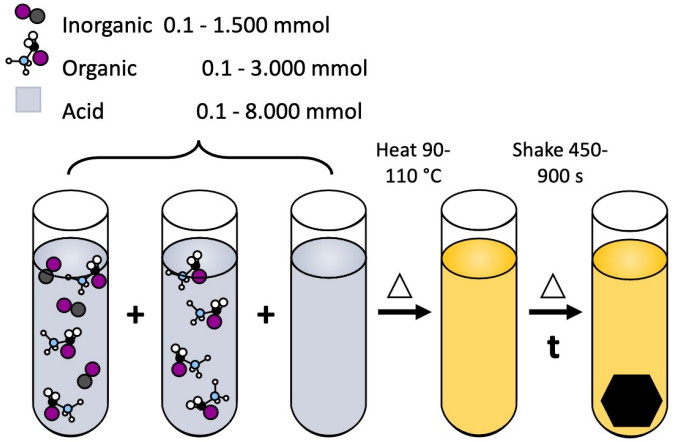
- Disclosure of the laboratory process
- Which leads to better replicability and reproducibility
- Complete record of successes and failures
- Can reduce human sampling biases
- Captures seemingly unimportant details that can lead to better insights

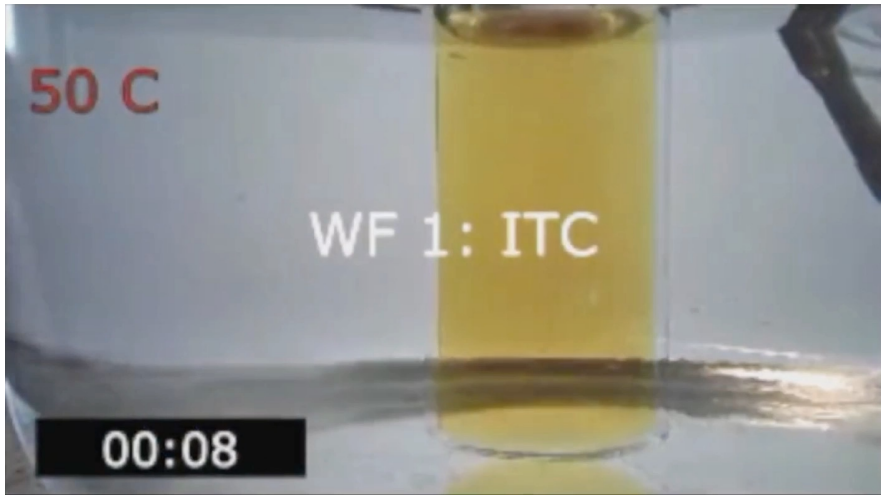


Perspective & Review article: *Matter* (2021) 10.1016/j.matt.2021.06.036

EXPERIMENTS IN THE LAB

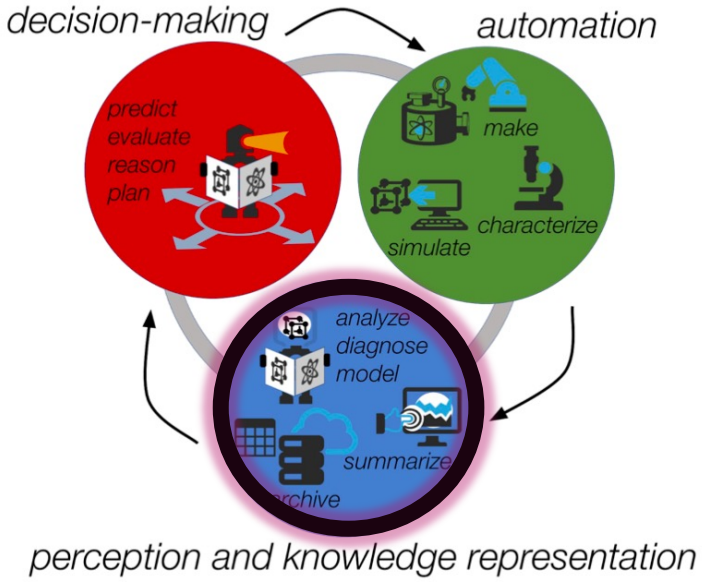






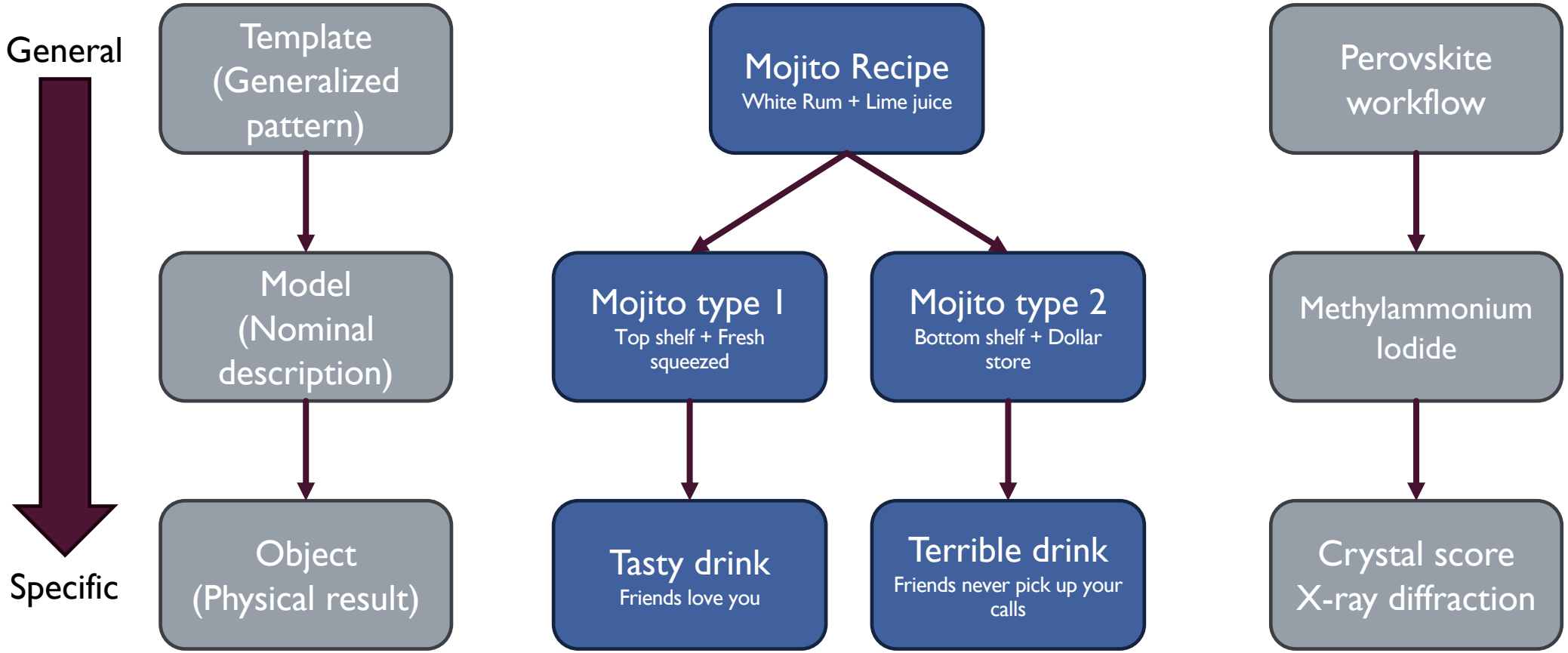
- Inverse temperature crystallization
- Liquid handling robots allowed high throughput experiments
- 96 experiments per 5 hours

CAPTURING DATA



- Experiment Specification Capture and Lab Automation Technology Environment (ESCALATE)
- Provide an API for humans or algorithms to specify new experiments
- Generate instructions for human operators and robots to conduct experiments.
- Archive experimental data and metadata
- Add interpretive layer (cheminformatics, stoichiometric calculations, etc.) to collected data
- Facilitate data reporting and export

DEFINING AN EXPERIMENT

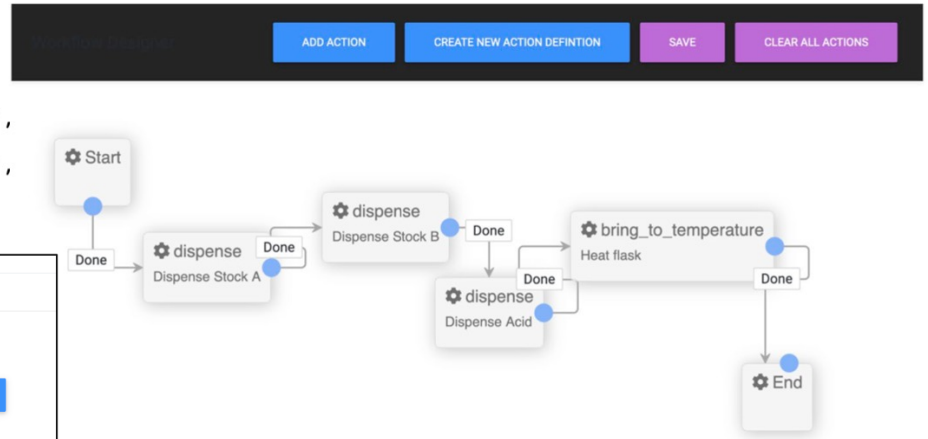


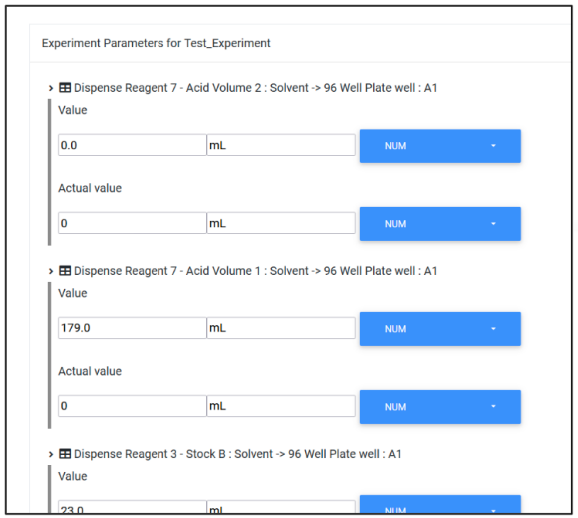
CAPTURING AND REPORTING EXPERIMENTAL DATA

```
perovskite_demo = get_data('experimenttemplate',
        {'description': 'perovskite_demo',
         'expand': 'workflow'})
[wf['description'] for wf in perovskite_demo['workflow']]
```

```
<Response [200]>
GET: OK
Found one resource, returning dict

['Perovskite Demo: Preheat Plate',
 'Perovskite Demo: Prepare Stock A',
 'Perovskite Demo: Prepare Stock B',
 'Perovskite Demo: Dispense Solvent',
 'Perovskite Demo: Dispense Stock A',
 'Perovskite Demo: Dispense Stock B',
 'Perovskite Demo: Dispense Acid Vol 1',
 'Perovskite Demo: Heat Stir 1',
 'Perovskite Demo: Dispense Acid Vol 2',
 'Perovskite Demo: Heat Stir 2',
 'Perovskite Demo: Heat']
```





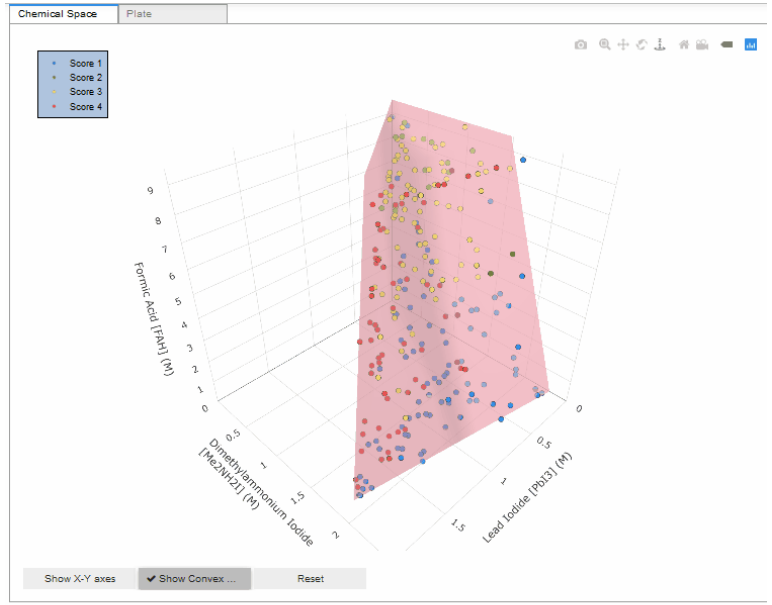
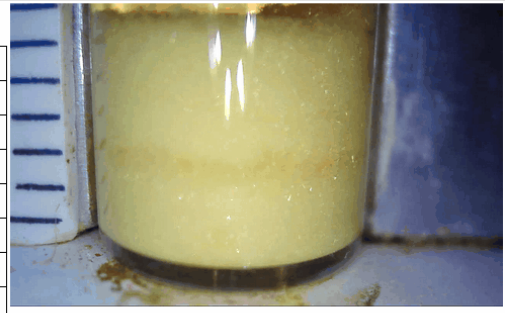
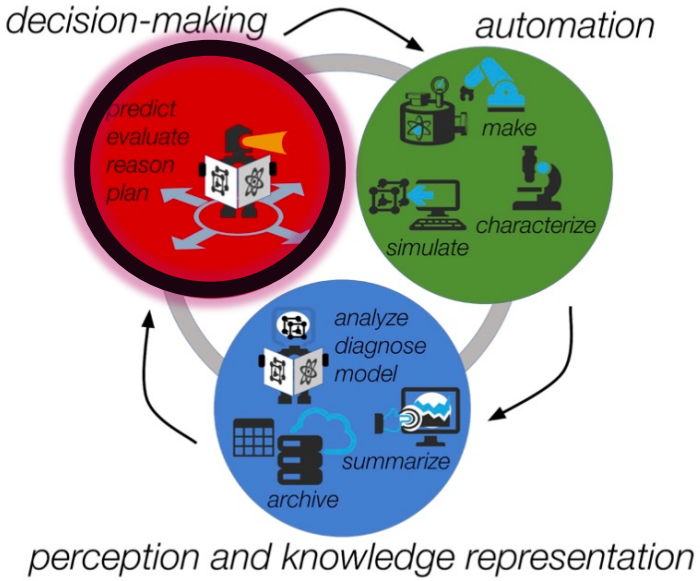


Plate ID:
2019-06-10T21_14_00.866379+00_00_LBL

Well ID	B9
Formic Acid [FAH]	5.619
Lead Iodide [PbI2]	0.9
Dimethylammonium Iodide [Me2NH2I]	1.602
Mixing Time Stage 1 (s)	900.0
Mixing Time Stage 2 (s)	1200.0
Reaction Time (s)	12600.0
Stir Rate (RPM)	750.0
Temperature (C)	95.0

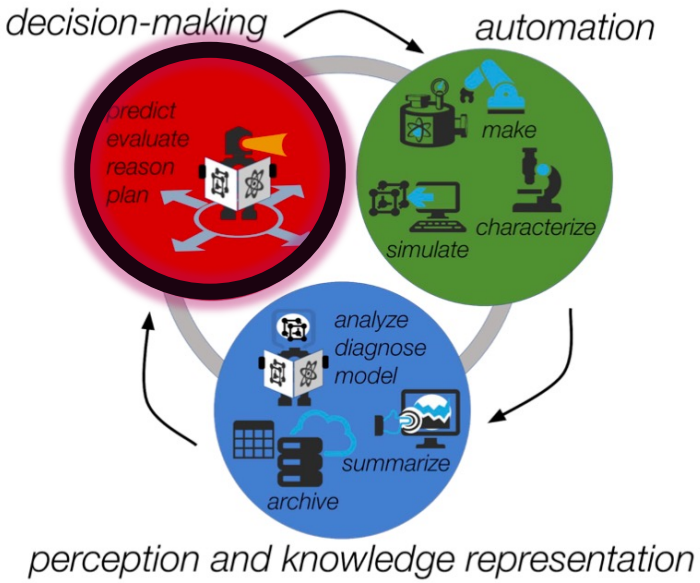


MAKING USE OF THIS DATA!



- Robot-Accelerated Perovskite Investigation and Discovery *Chem. Mater. (2020)* doi:10.1021/acs.chemmater.0c01153
- Can Machines "Learn" Halide Perovskite Crystal Formation without Accurate Physicochemical Features? *J. Phys. Chem C.* (2020)
- Active learning experiment selection for phase boundary mapping & control *Chem. Mater.* (2022) doi:10.1021/acs.chemmater.1c03564
- Active meta-learning for predicting and selecting perovskite crystallization experiments *J. Chem. Phys. (2022)* doi:10.1063/5.0076636
- Combine computer vision with simulation to infer virtual experiments from time-series observations *Chem Mater.* (2022)
- Serendipity based recommender system for perovskites material discovery: balancing exploration and exploitation across multiple models ChemRXiv (2022) doi:10.26434/chemrxiv-2022-l1wpf
- Identifying crystal growth additives with iterative machine learning + feature selection. *Cryst Growth & Design (2022)* doi:10.1021/acs.cgd.2c00522

MAKING USE OF THIS DATA!



- Robot-Accelerated Perovskite Investigation and Discovery *Chem. Mater. (2020)* doi:10.1021/acs.chemmater.0c01153
- Can Machines "Learn" Halide Perovskite Crystal Formation without Accurate Physicochemical Features? *J. Phys. Chem C.* (2020)
- Active learning experiment selection for phase boundary mapping & control *Chem. Mater.* (2022) doi:10.1021/acs.chemmater.1c03564
- Active meta-learning for predicting and selecting perovskite crystallization experiments *J. Chem. Phys. (2022)* doi:10.1063/5.0076636
- Combine computer vision with simulation to infer virtual experiments from time-series observations *Chem Mater.* (2022)
- Serendipity based recommender system for perovskites material discovery: balancing exploration and exploitation across multiple models ChemRXiv (2022) doi:10.26434/chemrxiv-2022-l1wpf
- Identifying crystal growth additives with iterative machine learning + feature selection. *Cryst Growth & Design (2022)* doi:10.1021/acs.cgd.2c00522

THE GREAT PEROVSKITE BAKING SHOW

- Can models predict successful experiments in a previously unseen chemical system with just a few random initial data points?
- External teams used algorithms to (remotely) play this game
- Evaluate **11 Bayesian/active/meta-learning algorithms** in a standard process
- Everyone starts with the same 10 randomly selected experiments
- Request 10 additional experiments
- Make 10 predictions of best outcome

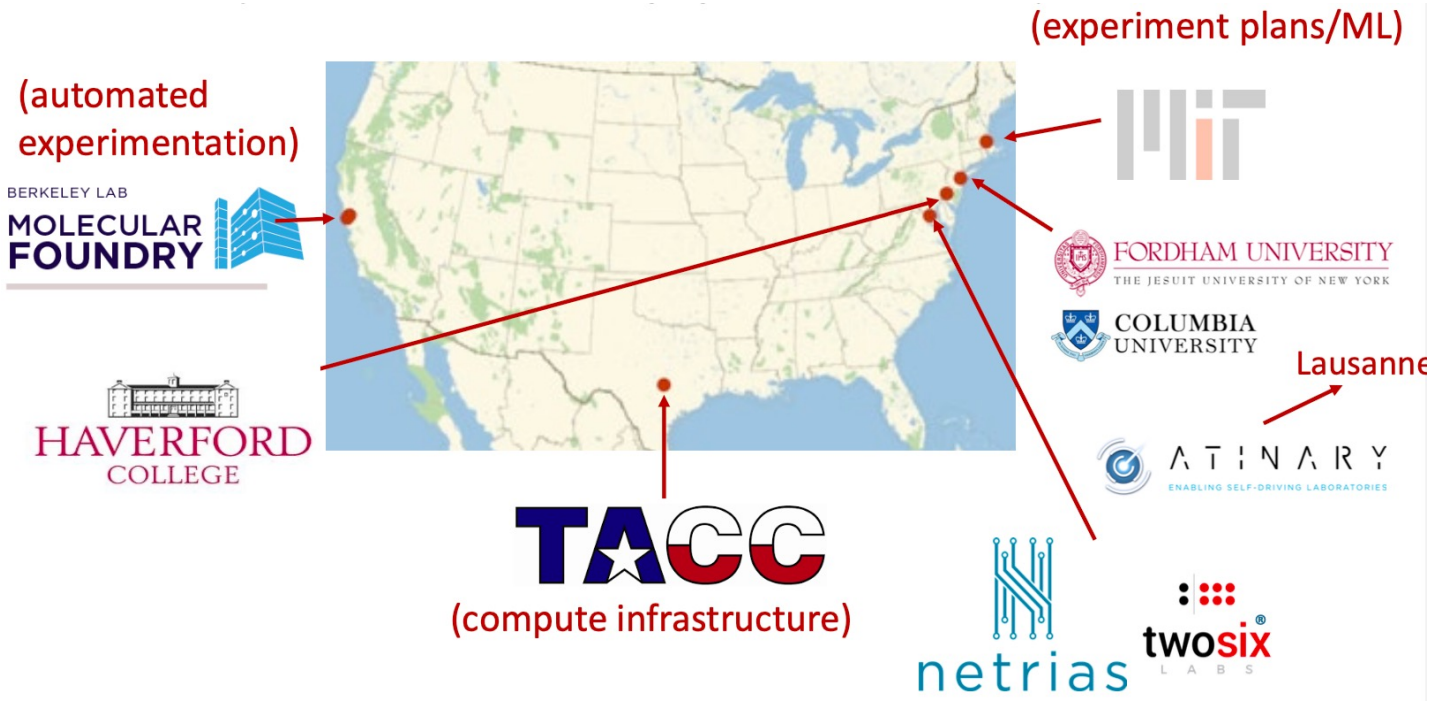
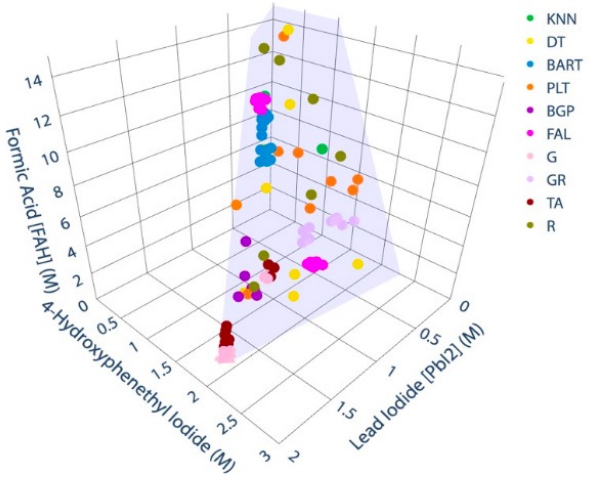
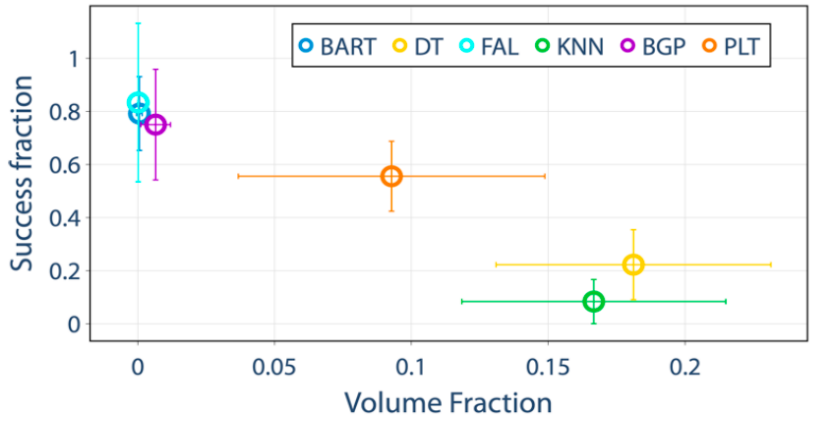


Image from "Creating Complex Scientific Workflows that Reach Into the Real World" Joshua Schrier, Institute for Pure & Applied Mathematics (IPAM) 2023

RESULTS

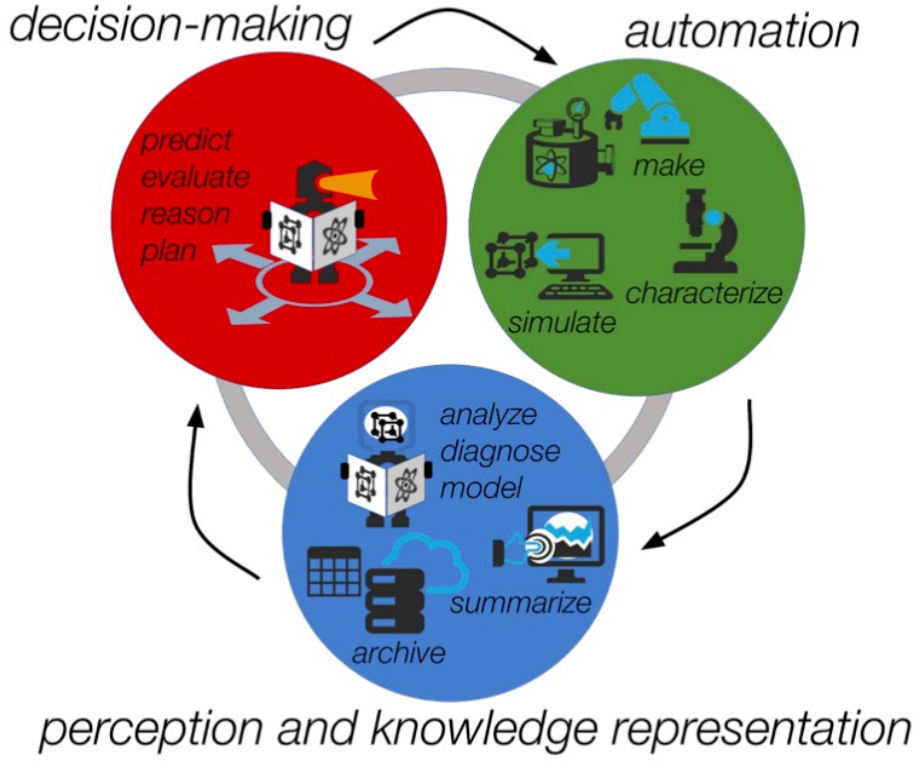
- Best performance by Gaussian Process-type methods & Bayesian Additive Regression Trees.
- Addition of chemical properties & historical data is useful for training
- Many active algorithms get trapped in local minima depending on the cold-start
- All algorithms tend to “clump” in regions of initial success—propose some solutions to force recommendations away in a model agnostic way





CLOSING THE LOOP

- Attempted to "close the loop" on autonomous labs
- Workflow to quickly run many experiments at once
- Built a software model to capture details of experiments in a machine consumable format
- Provided mechanisms to define new experiments by humans and machines for the next round



LARGE SCALE SCIENCE

NATIONAL SYNCHROTRON LIGHT SOURCE II

- Located 60 miles east of NYC in Long Island, NY
- Sponsored by the US Dept. of Energy Office of Science
- Half mile long circular building that can encompass Yankee stadium
- 1700 researchers/year
- Free science driven access policy

Source: Google Earth
https://earth.google.com

WHAT DOES IT DO?

- NSLS-II creates light beams 10 billion times brighter than the sun
- Directs light towards specialized experimental stations called beamlines
- Can reveal the electronic, chemical, and atomic structure using a broad spectrum of light beams, from infrared to hard x-rays
- Chemical changes within batteries
- Study motion of nanomaterials in plants
- Figure out structures of proteins

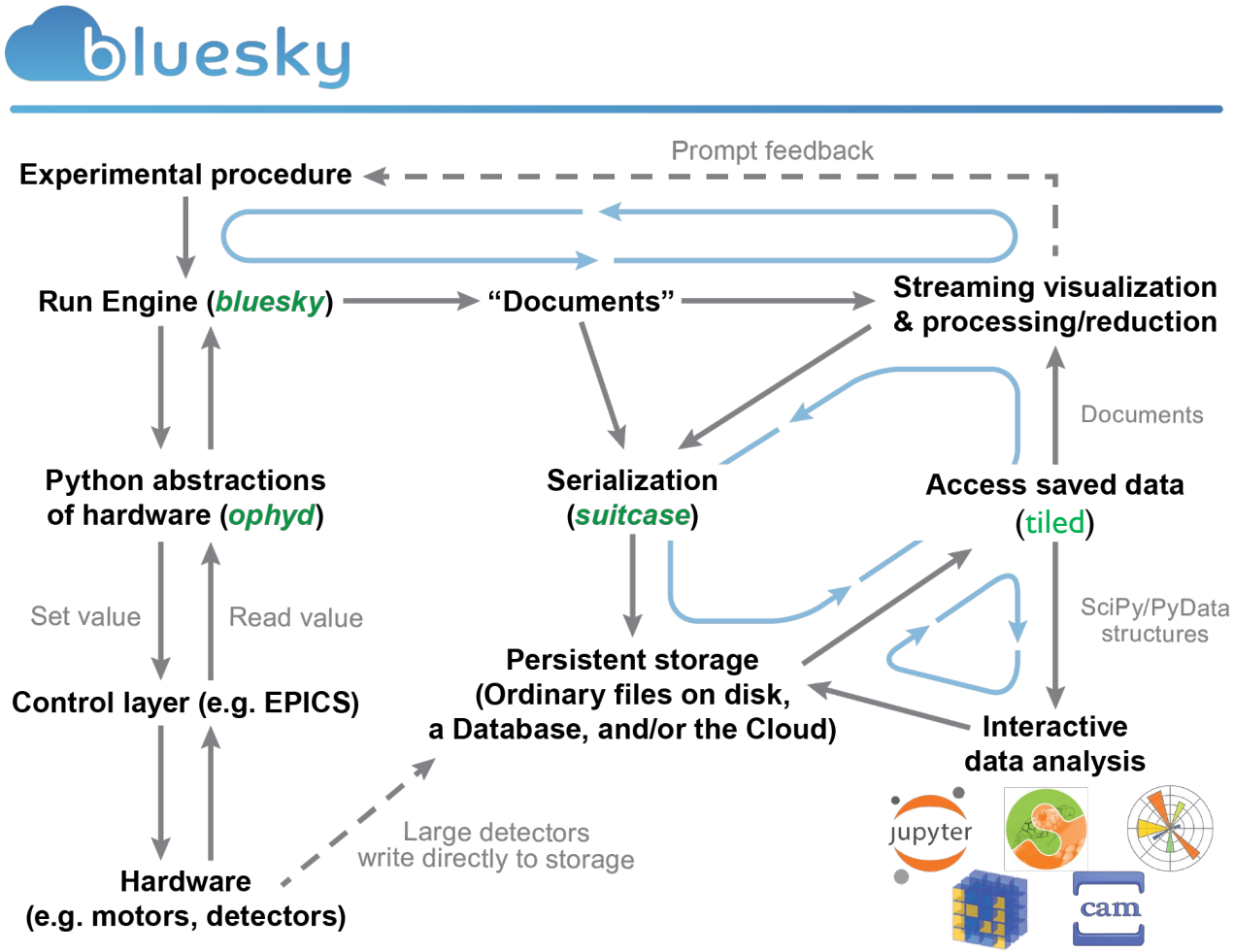
HOW DOES IT WORK?

Courtesy: Diamond Light Source
https://www.youtube.com/watch?v=SF4kr_Zjza4

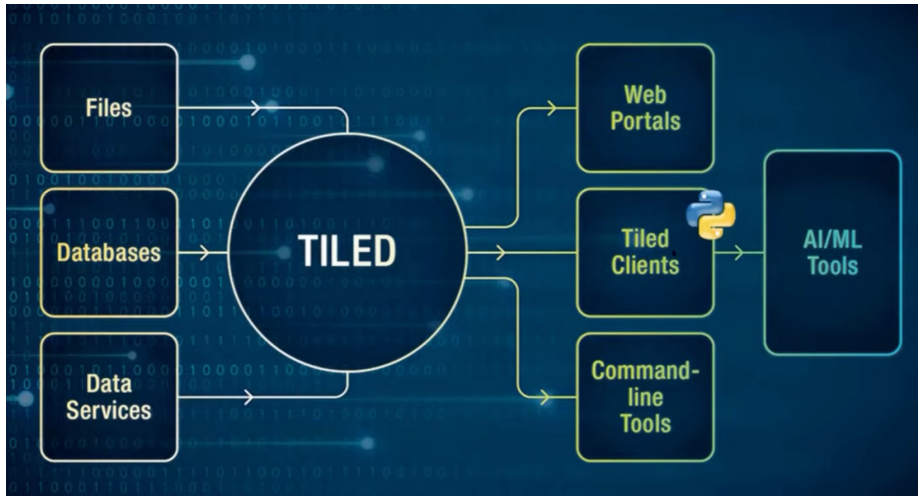
DATA SCIENCE AND SYSTEMS INTEGRATION

- Experimental requirements change all the time
  - Scientists dream up creative ways to use instruments
- 1000 networked devices in beamlines (2300 in accelerator)
- Detectors and instruments generate a large volume of data at high velocities
  - Wide range of data structures and access patterns
- Challenge: How do we manage the upstream hardware co-ordination and downstream data access/analysis?

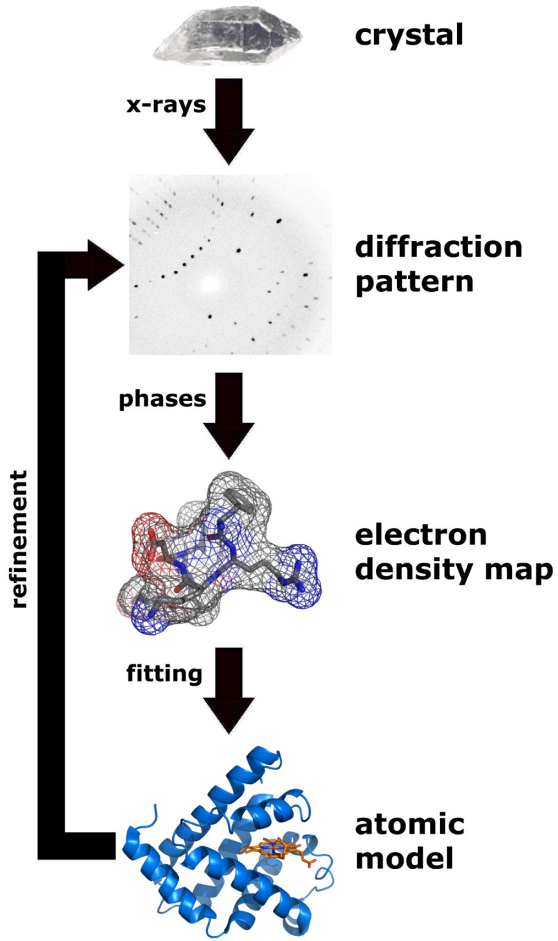
ABSTRACTIONS AND MORE ABSTRACTIONS!

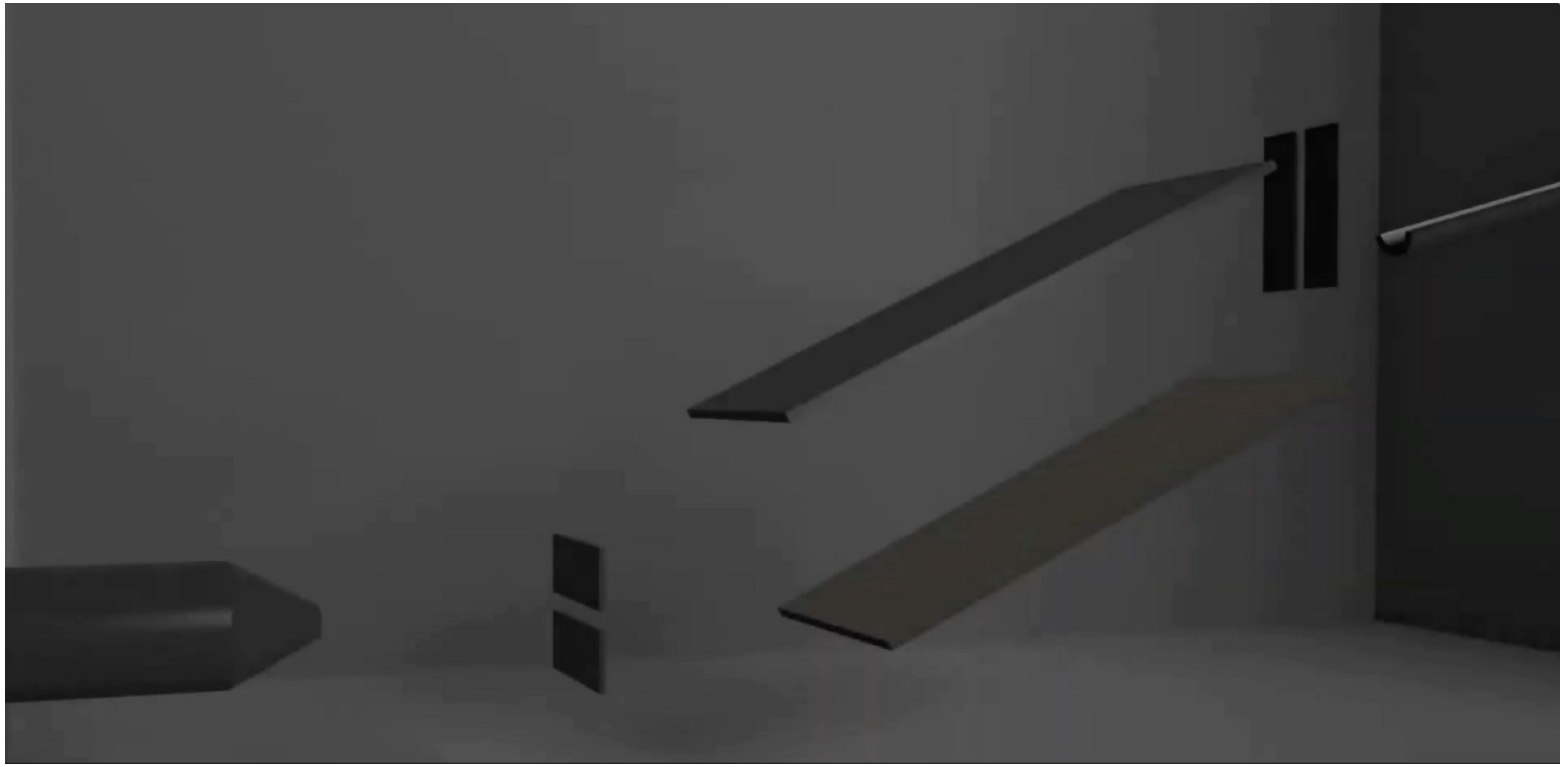


- https://blueskyproject.io/
- https://blueskyproject.io/ophyd/
- https://blueskyproject.io/tiled/



PROTEIN CRYSTALLOGRAPHY

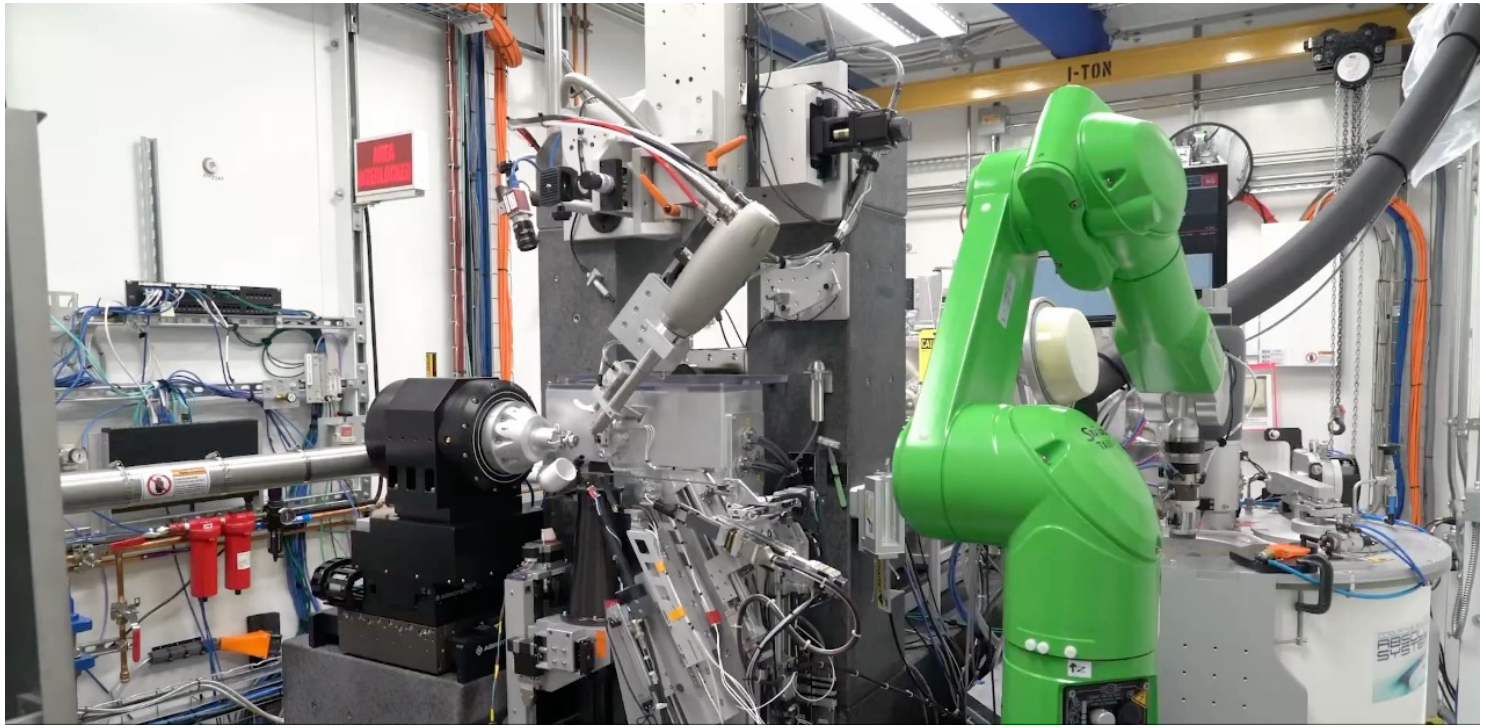




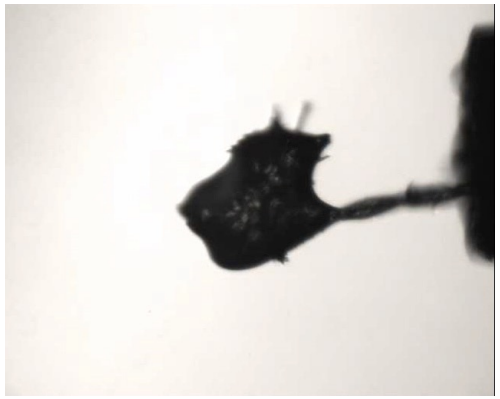
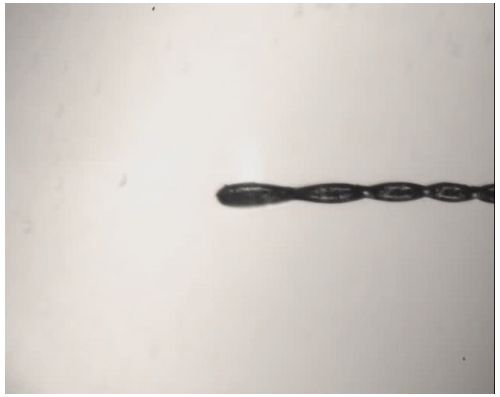
Courtesy Diamond Light source:
https://www.youtube.com/watch?v=hphmTRuV5nc

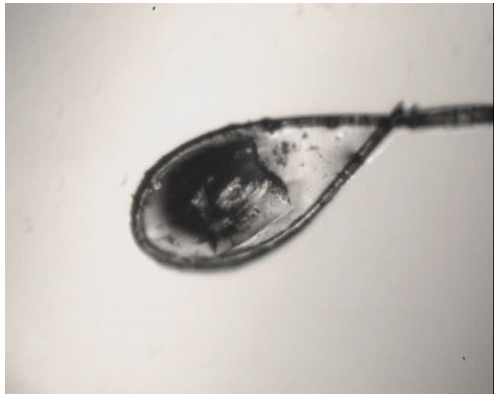
DATA COLLECTION

- Goal: Align, scan and collect best sample information as quickly as possible

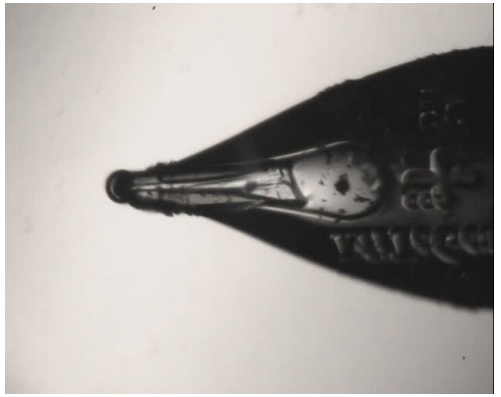


TYPES OF LOOPS

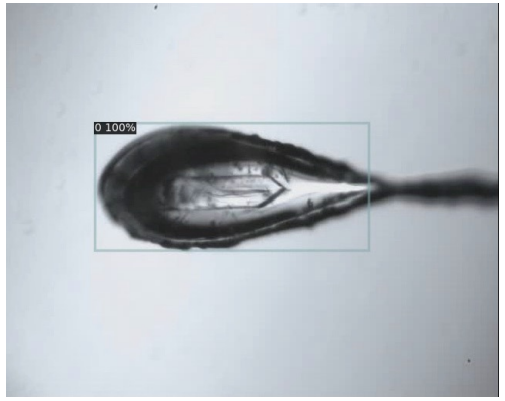


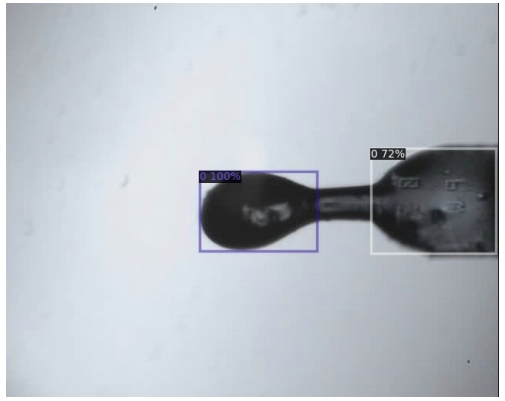


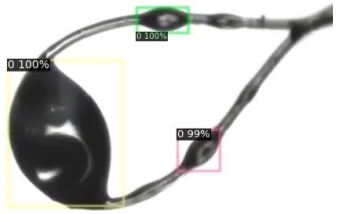


SAMPLE DETECTION

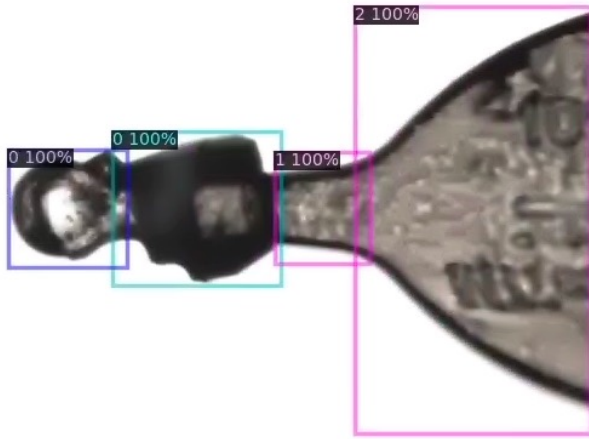
- Annotated samples using COCOs format
- Pre-trained Faster RCNN model (Detectron2)

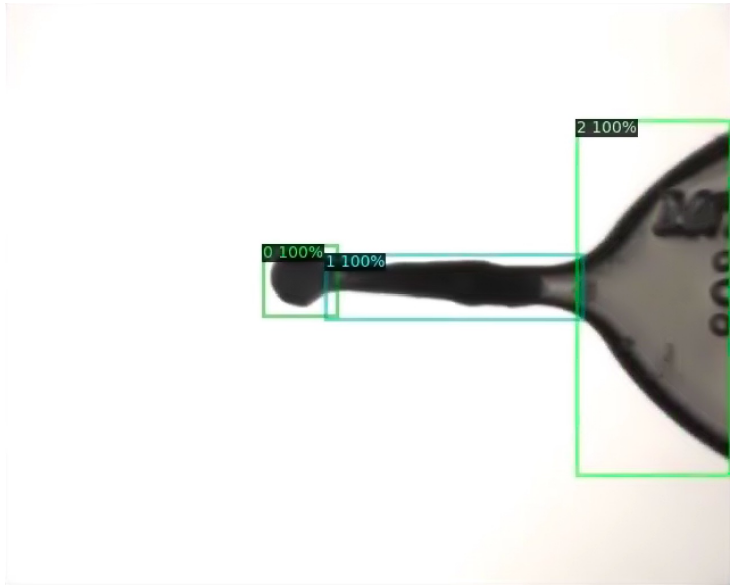


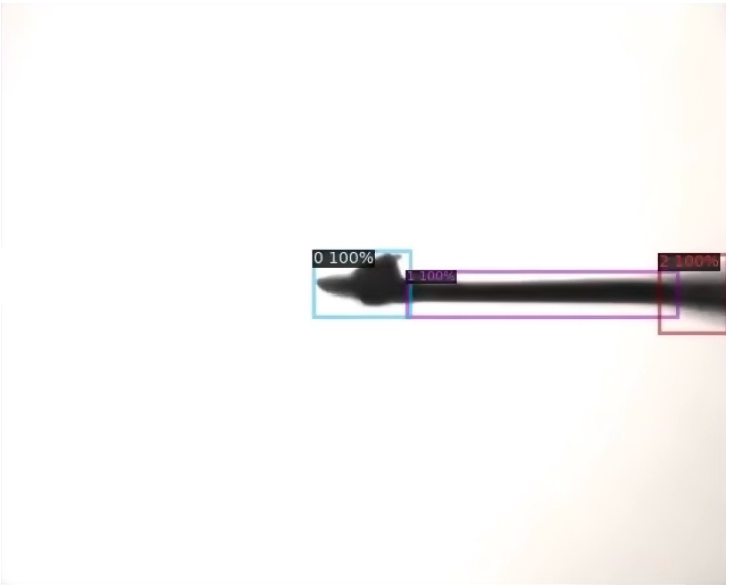




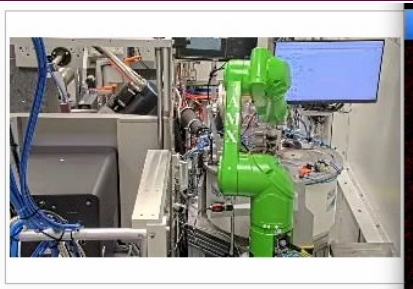
MORE ANNOTATIONS IMPROVED ACCURACY

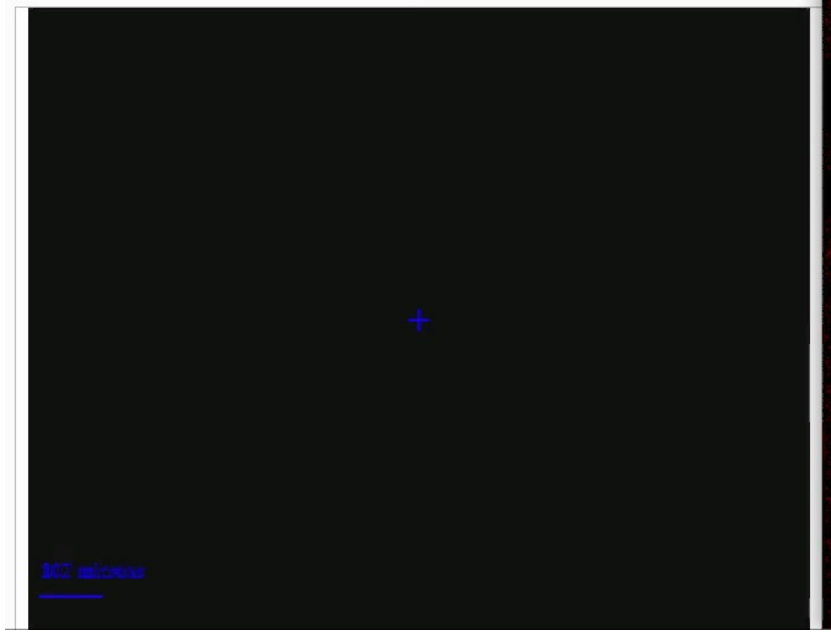


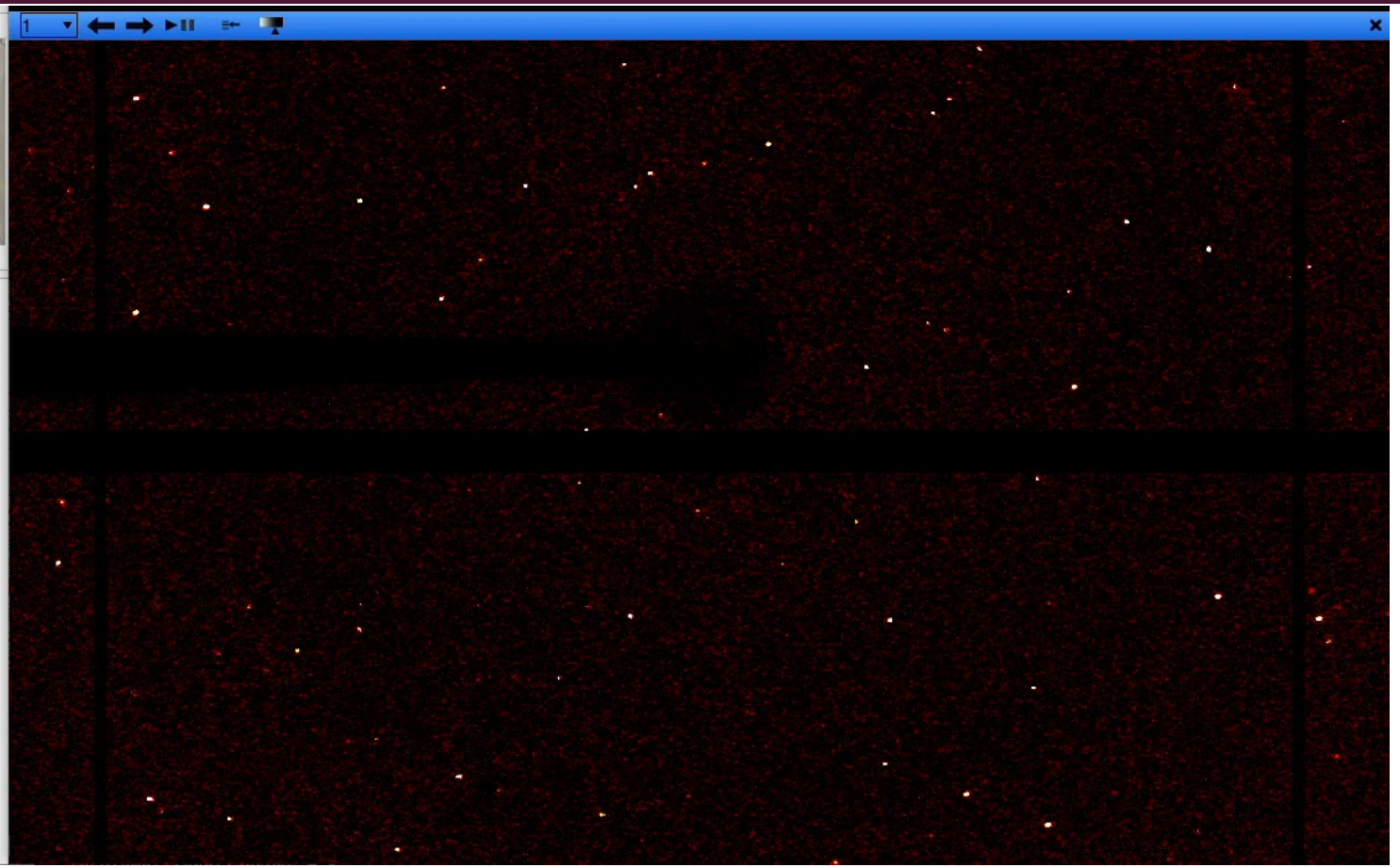




DATA ACQUISITION



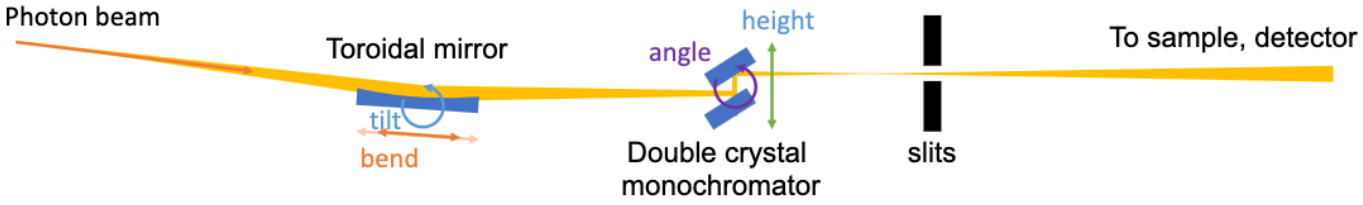




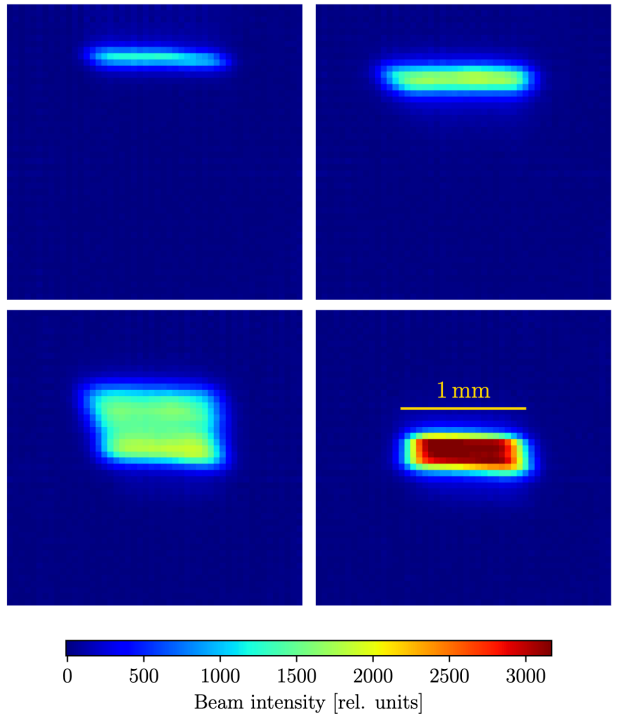
WHAT'S NEXT?

- More annotations to locate crystals if they are outside the loop
- Use a single raster diffraction image to predict quality of crystal
- Use models to identify useful diffraction spots
- Identify multiple regions within the same crystal for higher quality data

ML TO ALIGN BEAM



- Bayesian optimization strategy
- A General Bayesian Algorithm for the Autonomous Alignment of Beamlines, T.W. Morris et. al.



CLOSING THE (ELECTRON) LOOP

- Synchrotrons are complex machines
- Complex hardware co-ordination
- Generates large amounts of data at high velocities
- High throughput automation gives us multiple opportunities to use different kinds of ML algorithms

QUESTIONS?

THANK YOU!