Teaching Machine Learning

A. Gilad Kusne, aaron.kusne@nist.gov

National Institute of Standards & Technology

University of Maryland

ML Education

- NIST: Internships, Fellowships, Postdocs
- Bootcamp + Mini-Bootcamp
- Competition
- UMD Course + LEGOs!
- REMI, https://pages.nist.gov/remi/

Internships / Fellowships

- High School: Summer Internship (SHIP)
- Undergraduate: Summer Fellowship (SURF)
- Graduate: Host and Collaborations
- Postdoc (2 years): NSF NRC Fellowship
  - Primary recruitment tool.

Join Us!

- NRC – US Citizens only
  - ML-driven Autonomous Systems for Materials Discovery and Optimization
  - ML for Autonomous Genetic Engineering of Microbial Systems
  - ML for High Throughput Materials Discovery and Optimization Applications
- Non-US Citizens, **contact: aaron.kusne@nist.gov**

Annual Machine Learning for Materials Research Boot Camp and Workshop

Date: Aug 8-12

Location: Hybrid, UMD College Park

NIST National Institute of Standards and Technology U.S. Department of Commerce

University of Maryland 18 56

2020+1

Introduce researchers from industry, national labs, and academia to ML theory and tools for rapid data analysis.

- 4 days of lectures and **hands-on** exercises (e.g. noise reduction, unsupervised and supervised techniques, computer vision, etc.) includes ML for robot science!

- Focus on handling real data, both experimental and computational.

- Open-source, Python-based modules

- Symposium on Friday

https://www.nanocenter.umd.edu/events/mlmr/

Contact: aaron.kusne@nist.gov

Annual Machine Learning for Materials Research Boot Camp and Workshop UMD and NIST

https://www.nanocenter.umd.edu/events/mlmr/

MACHINE LEARNING *for* MATERIALS RESEARCH

BOOTCAMP & WORKSHOP

mlmr@umd.edu

UNIVERSITY OF MARYLAND 18 56

NIST National Institute of Standards and Technology U.S. Department of Commerce

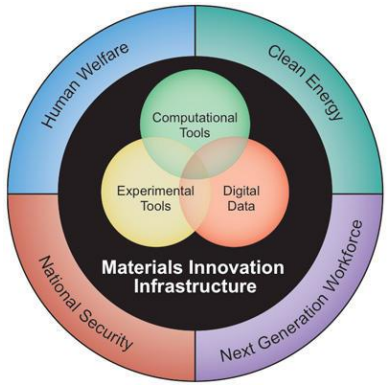
Organizers, Funding, Support

A. Gilad Kusne
National Institute of Standards & Technology
Materials Measurement Science Division

Ichiro Takeuchi
University of Maryland, College Park
Department of Materials Science & Engineering

Daniel Samarov
National Institute of Standards and Technology
Information Technology Laboratory

Alexei Belianinov
Oak Ridge National Laboratory
Center for Nanophase Materials Sciences



Jim Warren

NIST
MATERIALS GENOME INITIATIVE

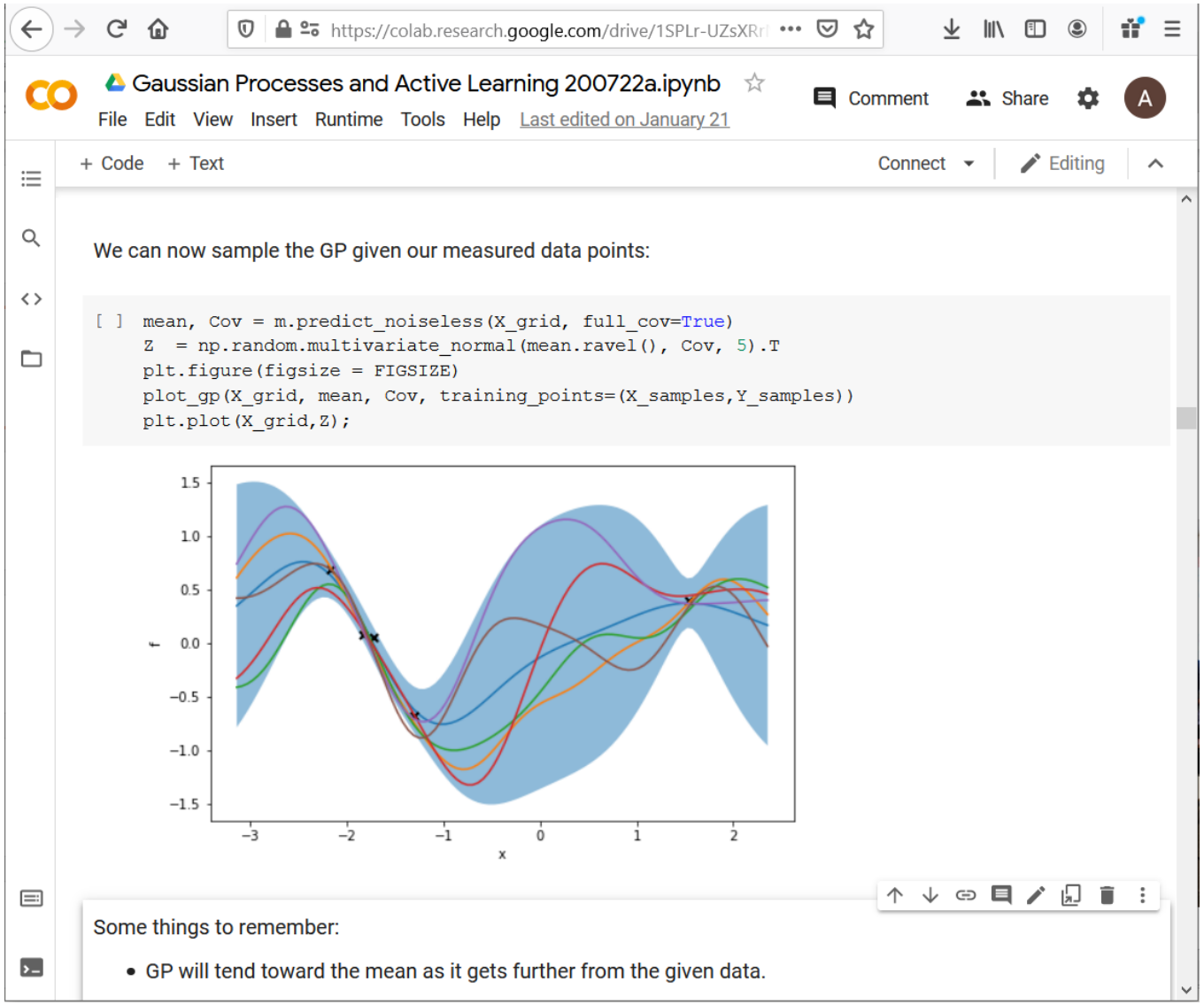
UNIVERSITY OF MARYLAND

For facilities

MRS® ENDORSED MEETING

Hands On

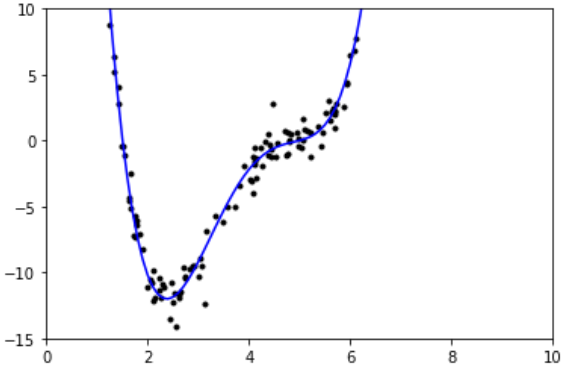
- Only open-source code (e.g. scikit learn)
  - No licensing issues
  - Free for budget-wary attendees
  - Minimum Programming
  - Matlab -> Anaconda -> Colab
- Colab – Online Platform
  - Saves from hours of installation
- Integration in Lectures
  - Alternate vs Split
- Attendees invited to share data to become exercises.

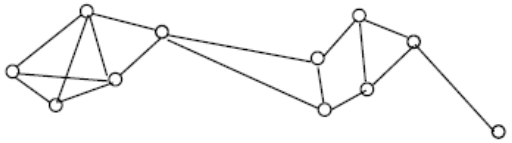


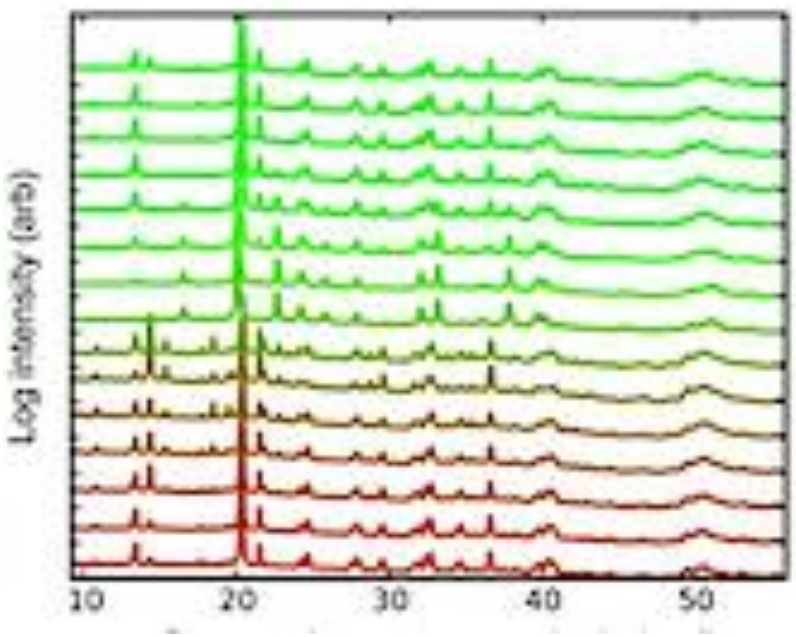
mlmr@umd.edu

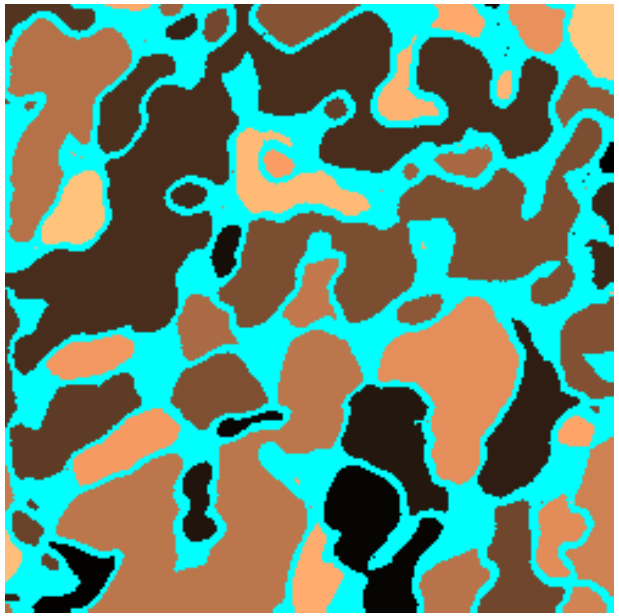
Hands On

- Diverse Data Types
  - Scalar, Spectra, Images, Hyperspectral Images, Graphs
  - Simulation & Experiment









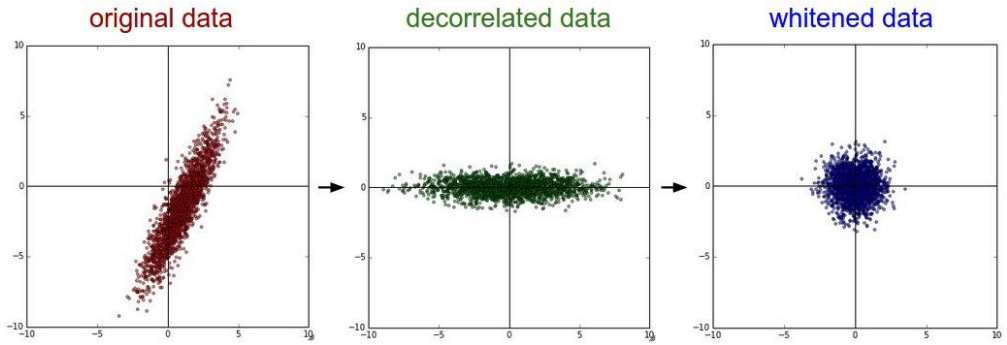
mlmr@umd.edu

Topics Covered

• Intro to Python	1/2 day
• Data Pre-processing	1/2 day
• Unsupervised Learning	1 day
• Supervised Learning	1 day
• Active Learning	1/2 day
• Recent pub work (hands on)	1/2 day
• Workshop	1 day

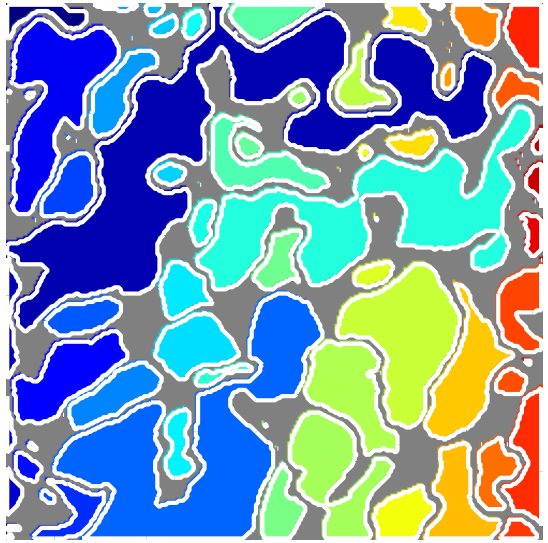
mlmr@umd.edu

Day 1: Intro to Python and Data Preprocessing



Feature Detection

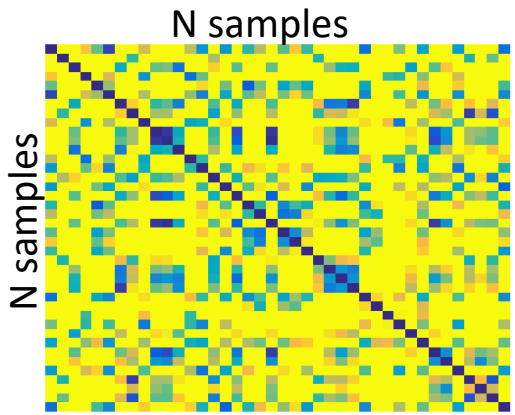
Domains and Boundaries



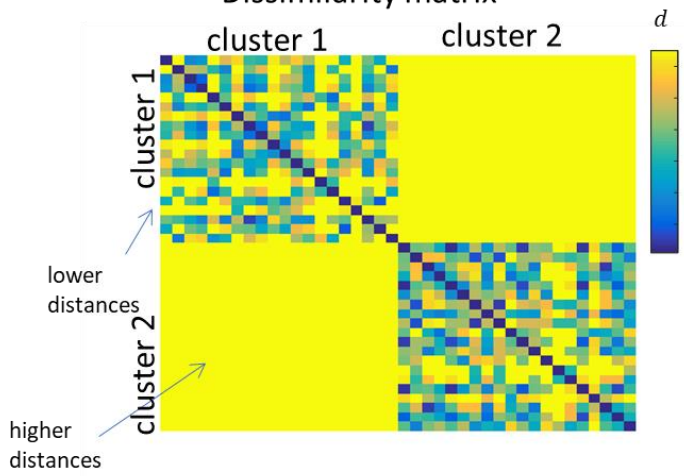
mlmr@umd.edu

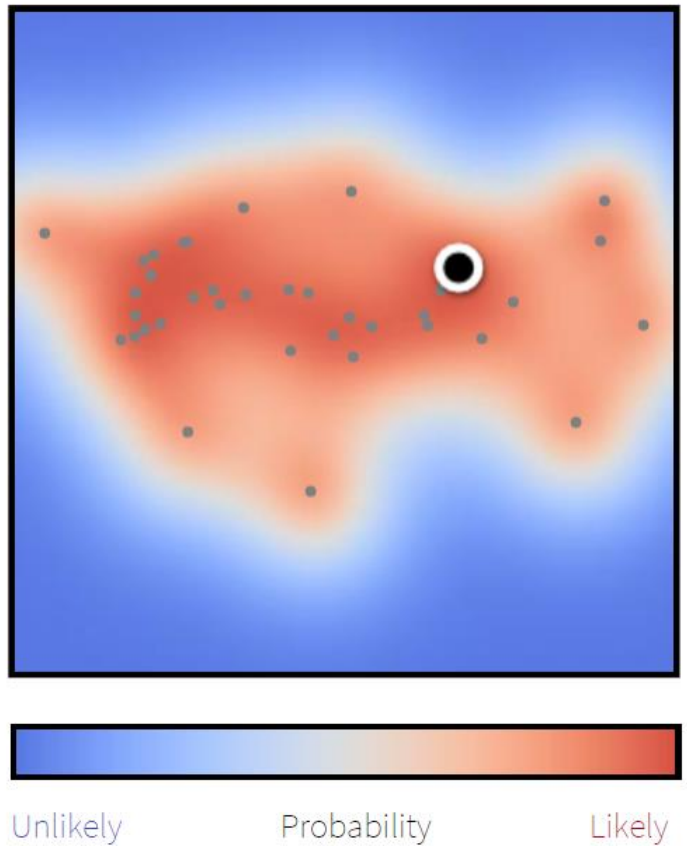
Day 2: Unsupervised Learning

Unordered Dissimilarity matrix $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$



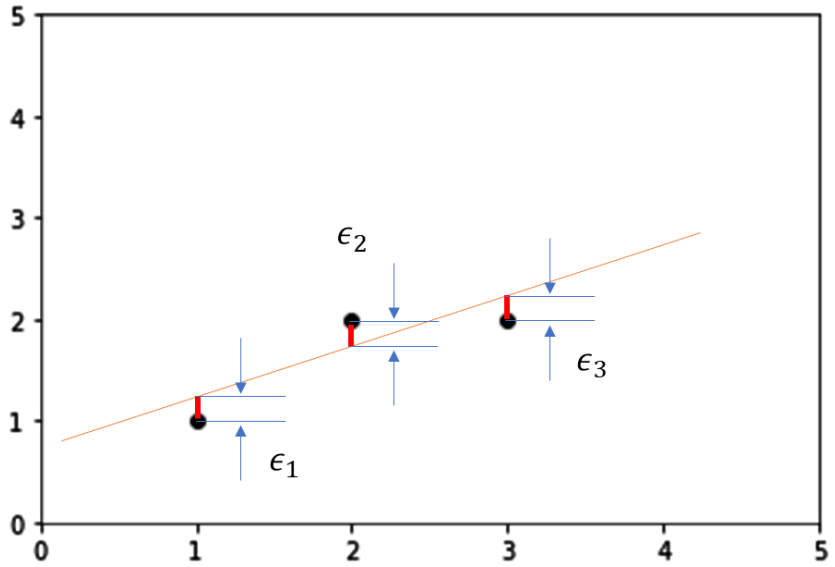
Cluster-ordered Dissimilarity matrix

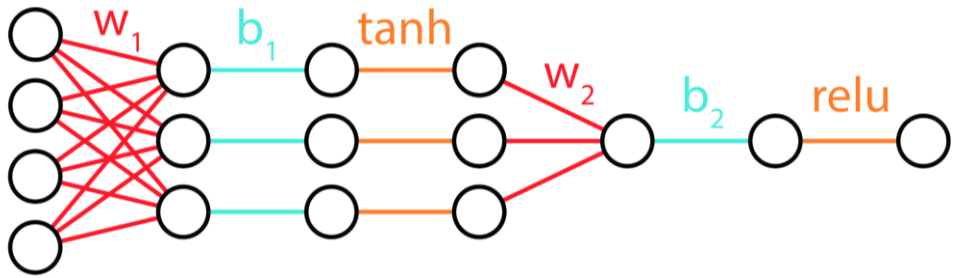




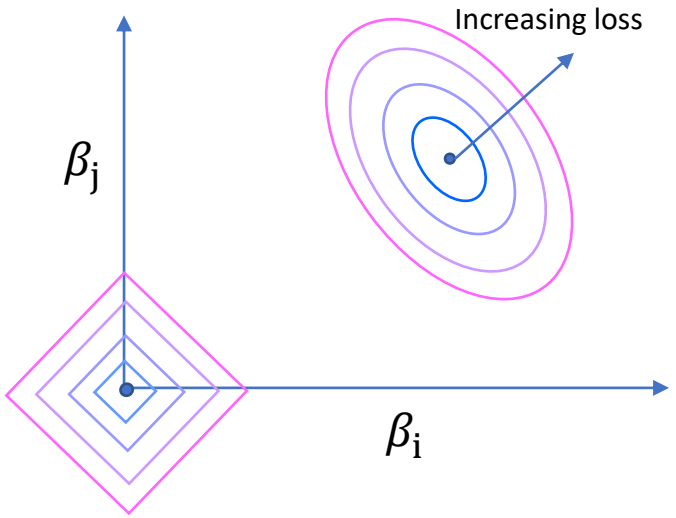
mlmr@umd.edu

Day 3: Supervised Learning





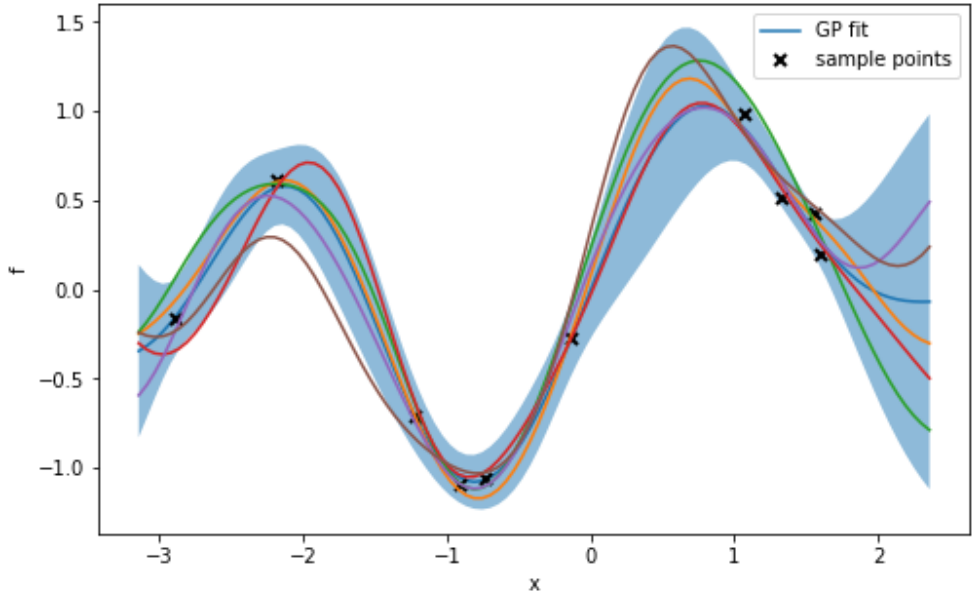
Loss

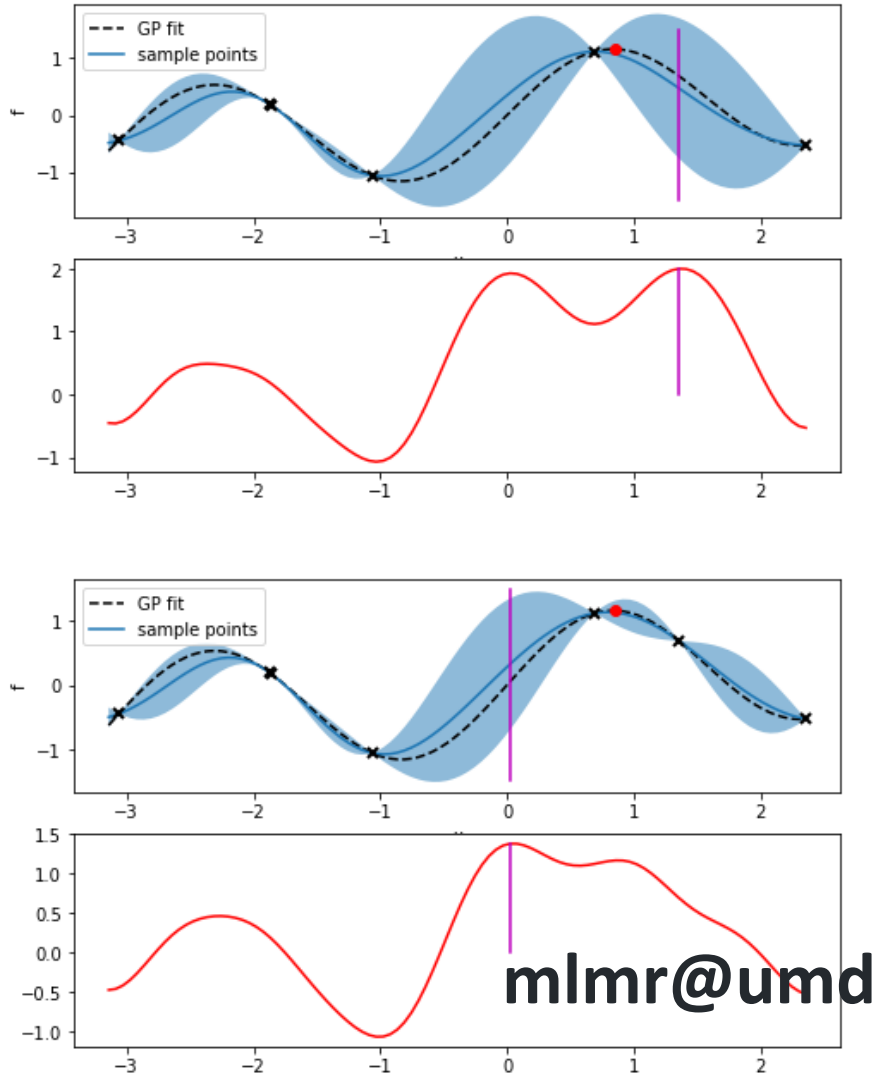


- From basic linear regression to the most advanced methods.

mlmr@umd.edu

Day 4a: GPs, Active Learning, Autonomous

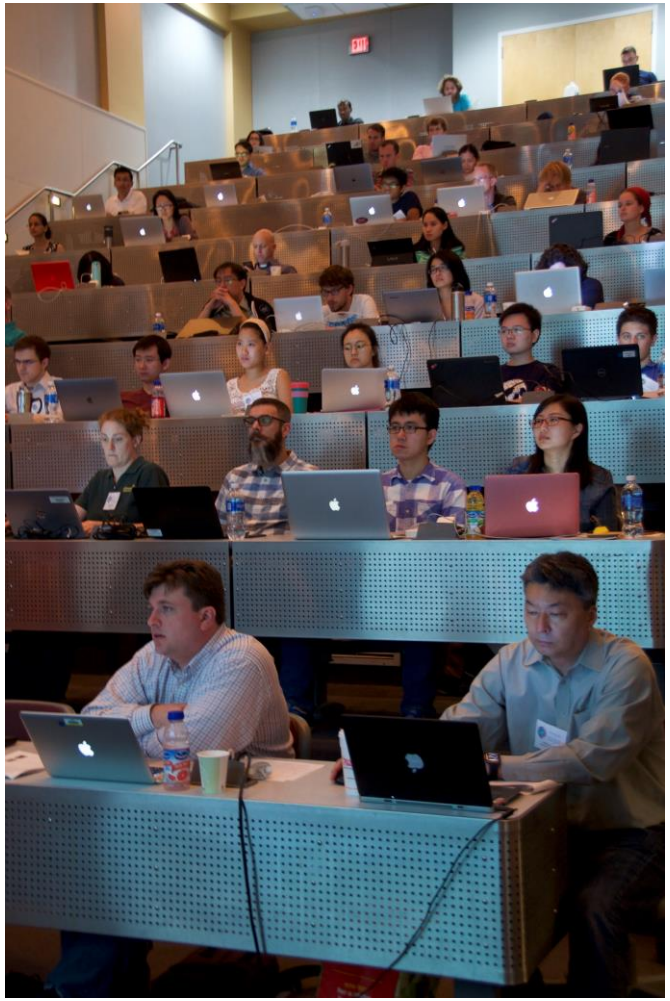




mlmr@umd.edu

Day 4b: Walkthrough of Recent Innovations

- Walkthrough by authors of recent high impact papers.
- Open data, open code.
- How to access and work with large DBs.
- Step-by-step in colab.
- Thought processes.



mlmr@umd.edu

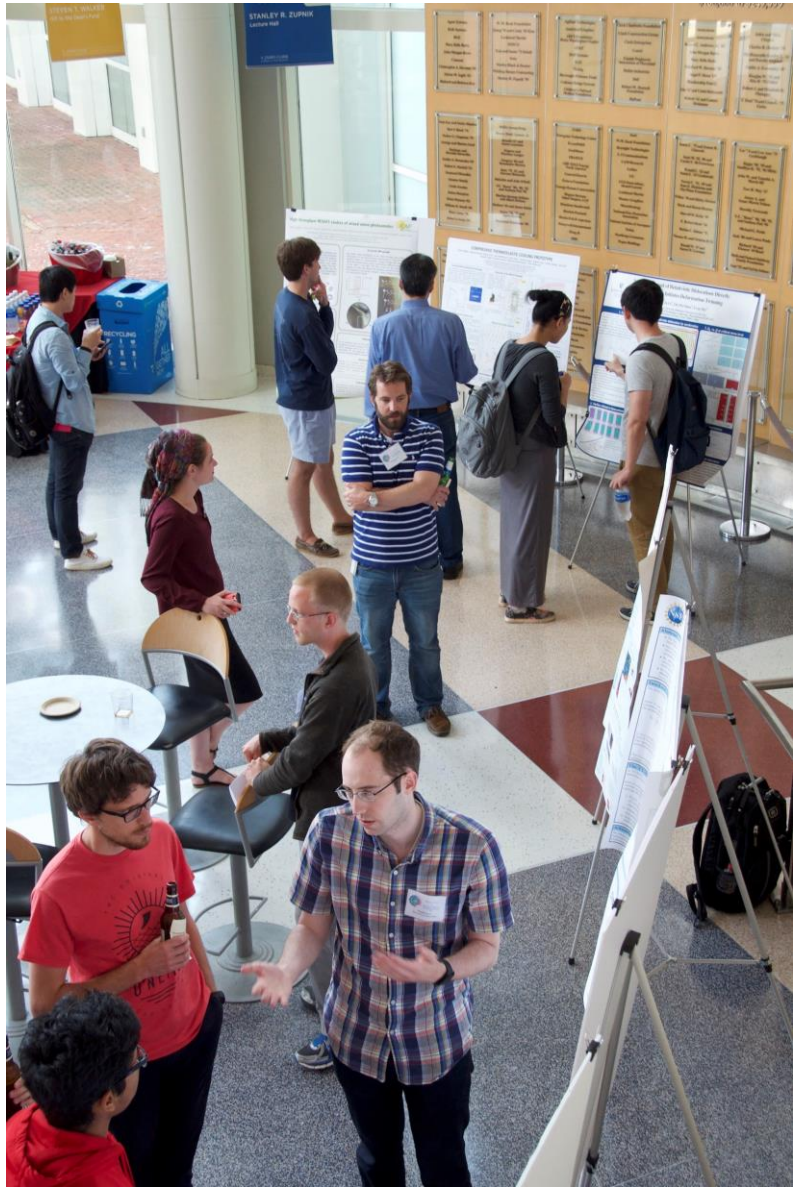
Focus

- Existing tools and their use
- Minimum Code for Maximum Productivity
- Best Practices
  - Many ML techniques exist, when/how to use them
  - Full ML Pipeline
- Small data / Big data
- Uncertainty Quantification & Propagation

mlmr@umd.edu

Poster Sessions

- Bring in poster on your data challenges.
- Brainstorm solutions with MLMR faculty & students over tacos.



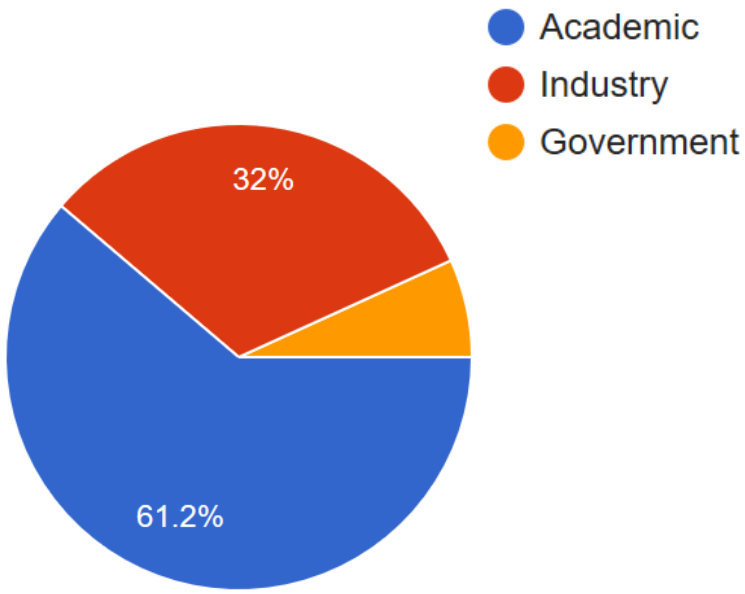
mlmr@umd.edu

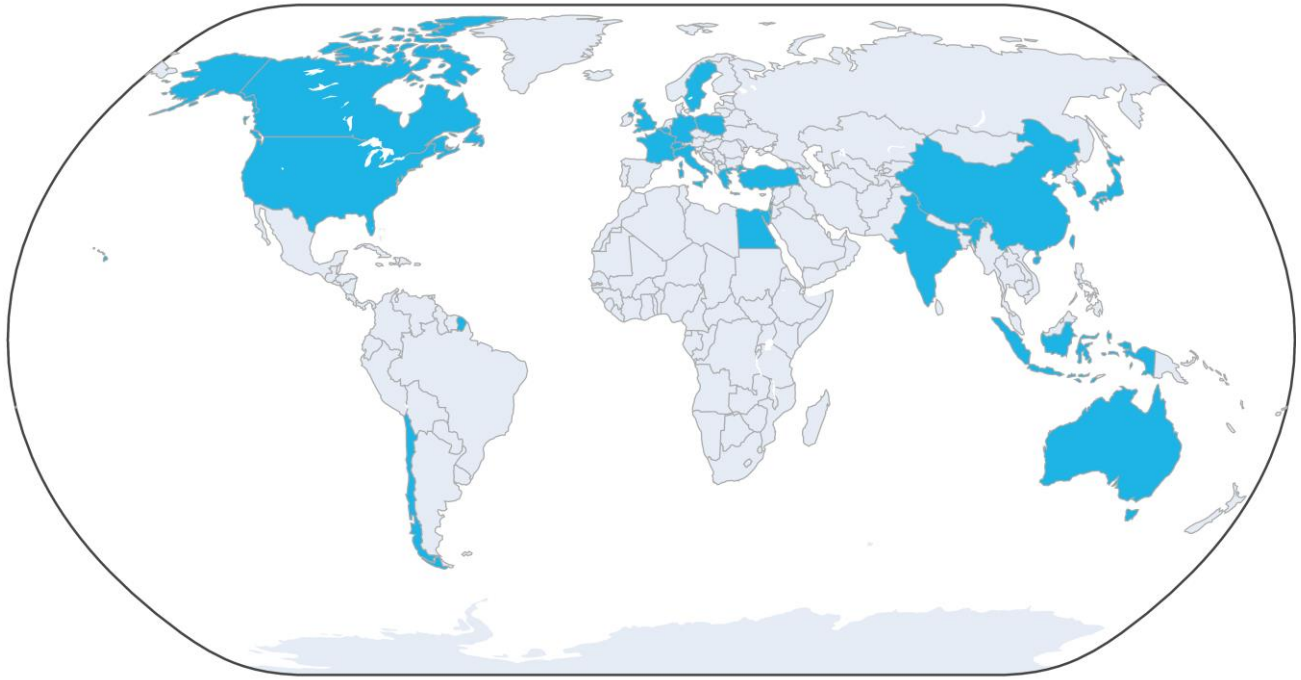
Hackathon

- For those that want to work through the evening.
- Day 1: Announce challenge
- Day 4: Walkthrough of solutions.

mlmr@umd.edu

Past Attendees: 24 Countries





Past attendees from around the world!

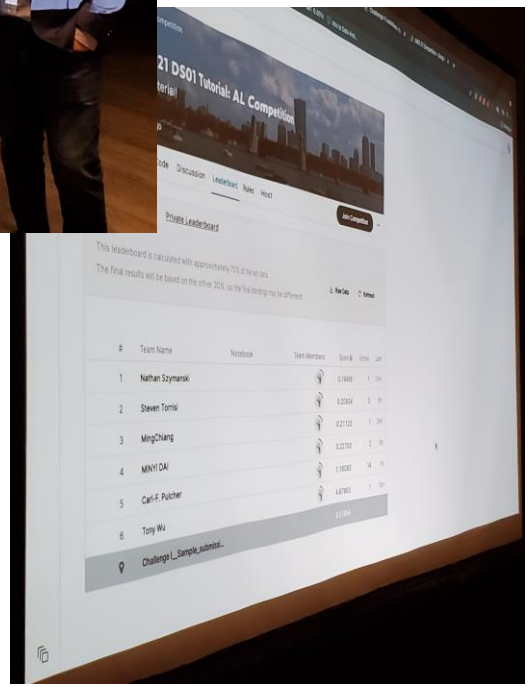
mlmr@umd.edu

Mini Bootcamps

- Format: 1.5 hours – 2 days
- Hands-on
- MRS, TMS, APS, NSF Workshops, MLSE, Etc.

mlmr@umd.edu

Competition – Big awards: Future on REMI



Matter of Opinion

Teaching machine learning to materials scientists: Lessons from hosting tutorials and competitions

Shijing Sun [1], Keith Brown [2], A. Gilad Kusne [3, 4]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.matt.2022.04.019

Get rights and content

The growing field of data-driven materials research poses a challenge to educators: teaching machine learning to materials scientists. We share our recent experiences and lessons learnt from organizing educational sessions at the fall 2021 meeting of the Materials Research Society.

- Next Competition: August

UMD Course: ML for Physical Scientists

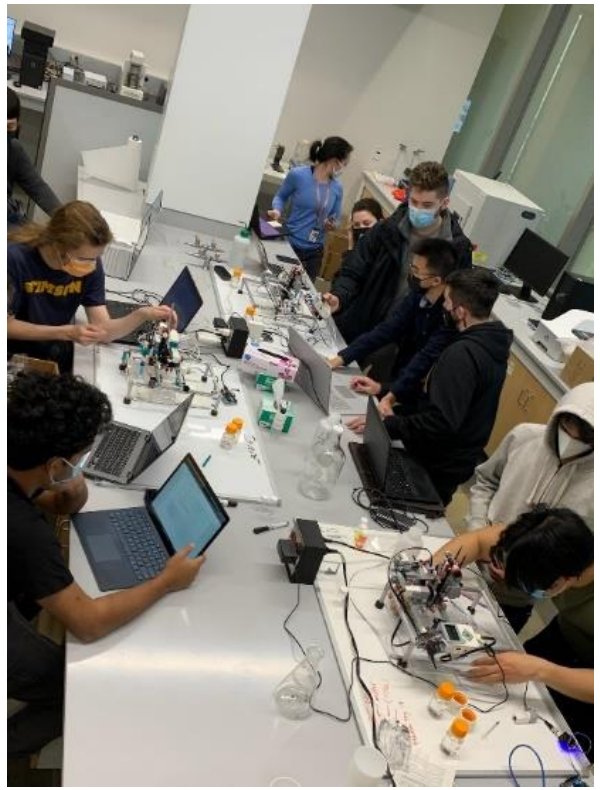
- From intro to Python through ML for autonomous physical science.
- Playing with LEGOs! low-cost LEGO platform for class projects

Why?

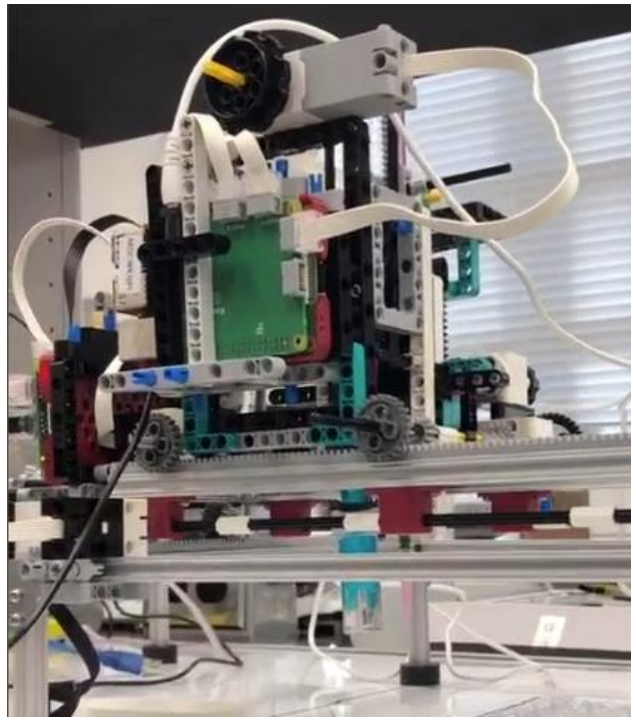
*** Teach Full Data Science Pipeline**

*** Students learn Consequences of Decisions!**

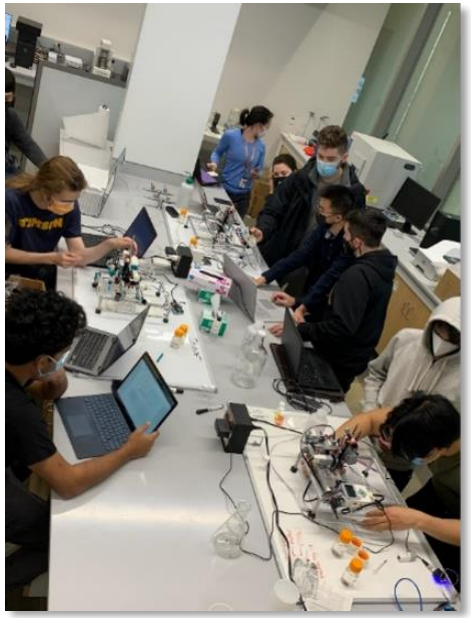
*** Teach controls level of complexity**

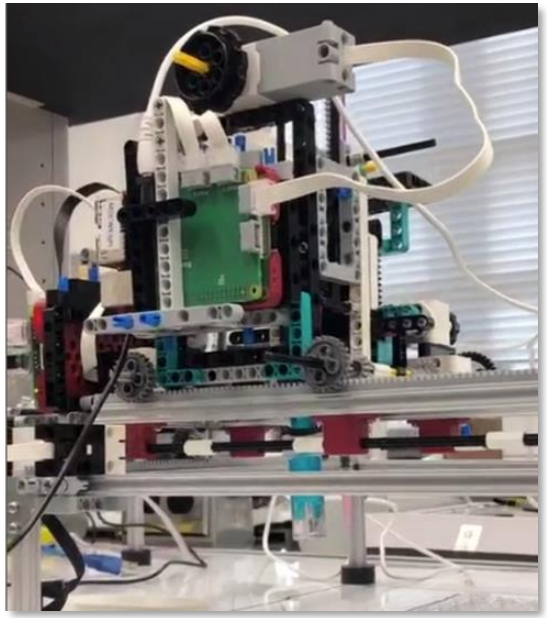


Household items: Vinegar + Milk of Magnesium



A Low-Cost Education Platform for Teaching Autonomous Physical Science





Logan Saar

*B.S., Materials Science and Engineering, UMD (**lsaar@umd.edu**)*

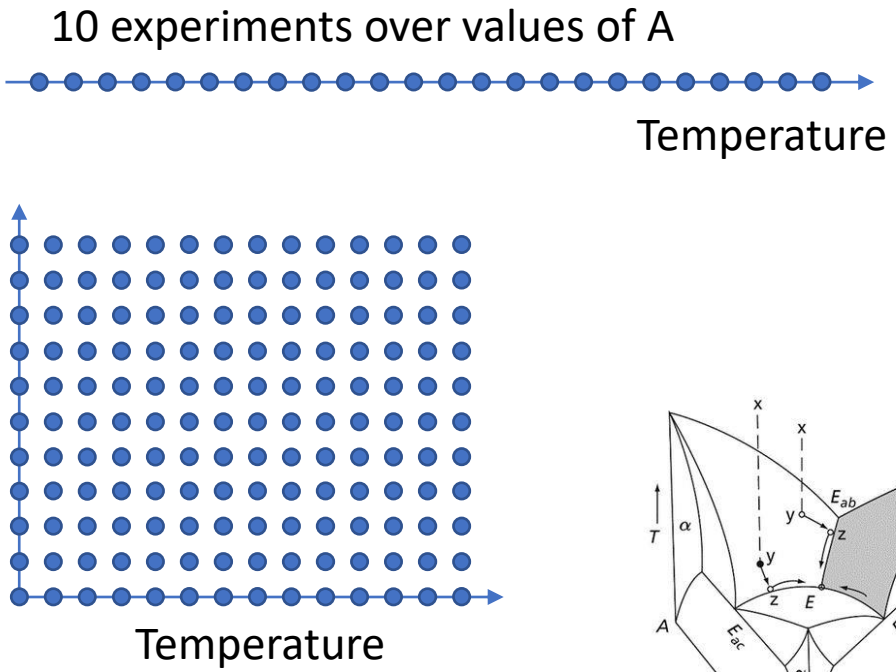
Gilad Kusne, Austin McDannald, Ichiro Takeuchi

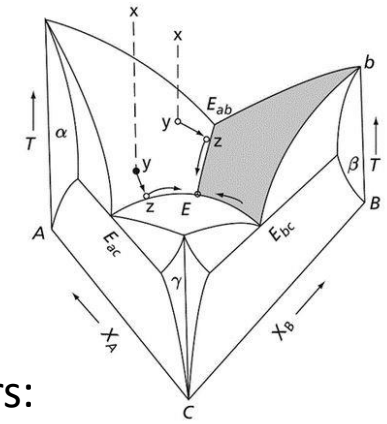
University of Maryland & NIST

The Challenge of Materials Exploration

Complex materials described by High dimensional space!

Exhaustive Search:





4 parameters:
3 Elements + Temperature

Assume: For each parameter, 10 experiments over range.

10^2 experiments

4 Parameters -> 10^4 experiments

For N parameters -> $10^{(N)}$ experiments!

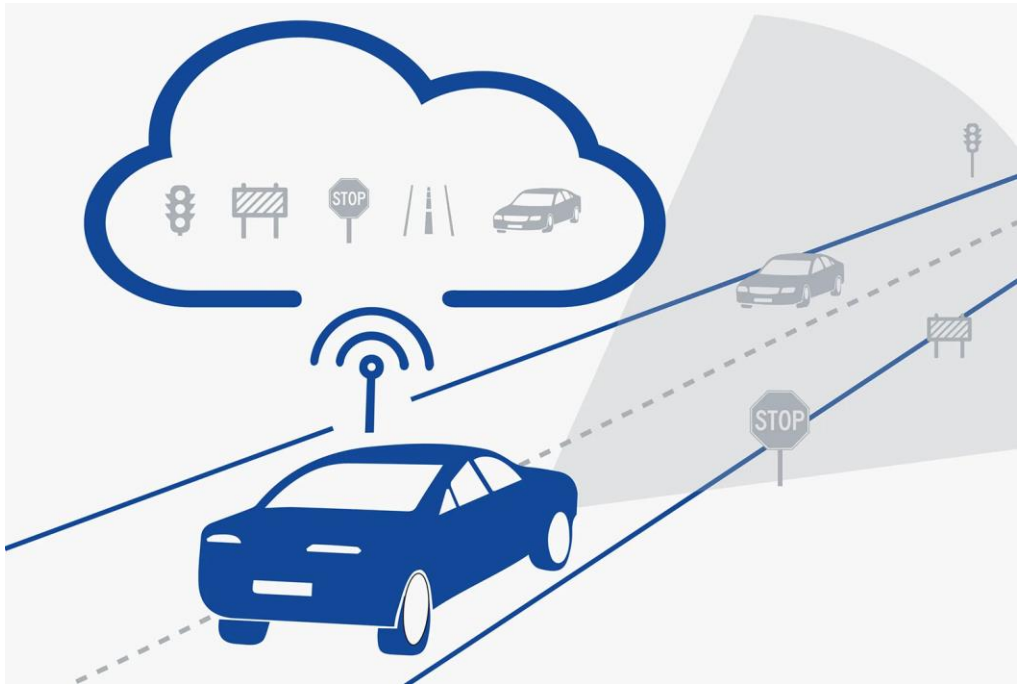
Complex materials and complex materials physics are out of reach!

Automated

Robot **executes** tasks

Image Credit: https://blog.robotiq.com/a-brief-history-of-robots-in-manufacturing

Autonomous



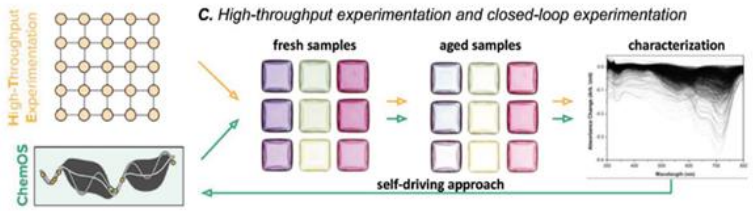
Robot **learns** & ..
Reacts to gathered data

ACTIVE LEARNING

Image C

A mobile robotic chemist

Burger et al., Nature **583**, 237 (2020)



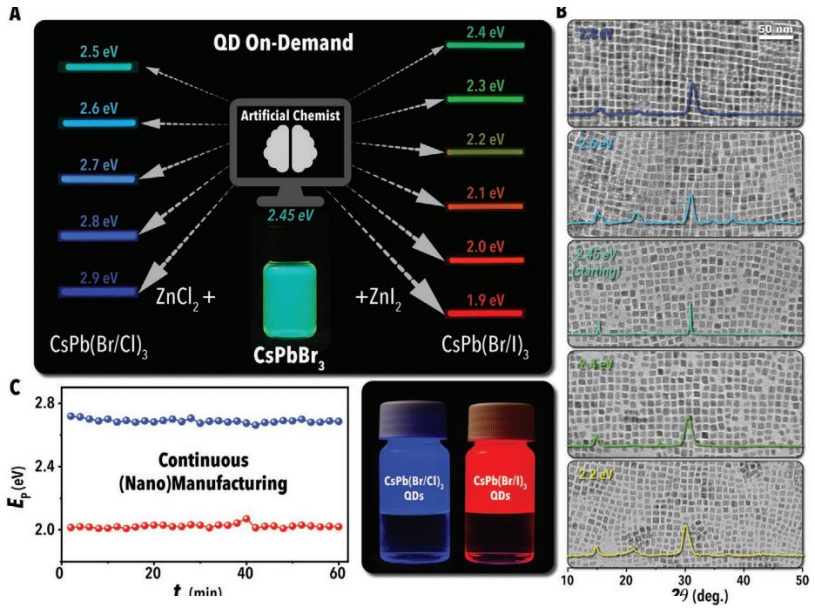
- Blending/mixing or polymers/organic molecules
- Number of experiments can be significantly reduced

Burger, B., Maffettone, P.M., Gusev, V.V. et al. A mobile robotic chemist. *Nature* 583, 237–241 (**2020**)

Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems

Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot

*Robert W. Epps, Michael S. Bowen, Amanda A. Volk, Kameel Abdel-Latif, Suyong Han, Kristofer G. Reyes, Aram Amassian, and Milad Abolhasani**



Adv. Mater. **2020**, *32*, 2001626

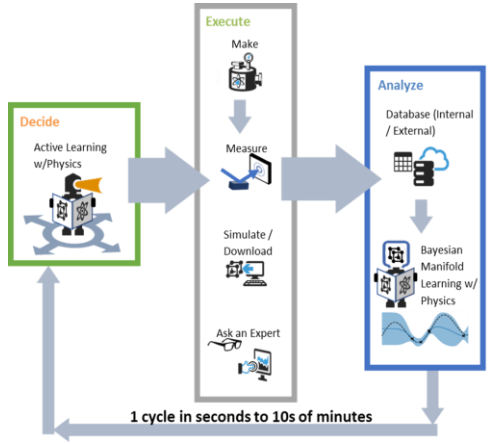
CAMEO: Closed-Loop Autonomous Materials Exploration and Optimization

Discovered: New best-in-class phase change memory material

ScientificAI: built in phase map and XRD physics

10x acceleration over off-the-shelf methods

Run at: SLAC

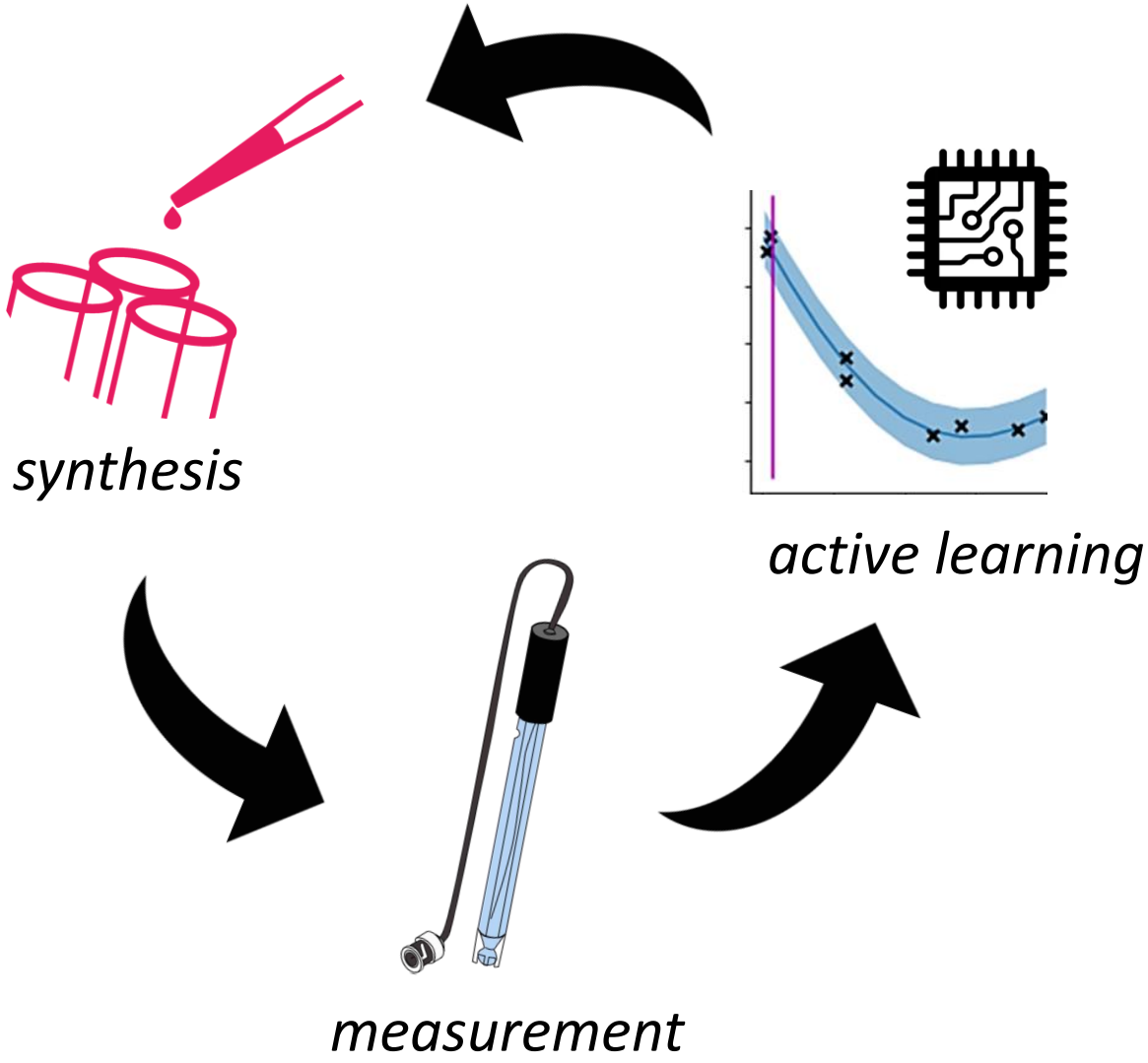


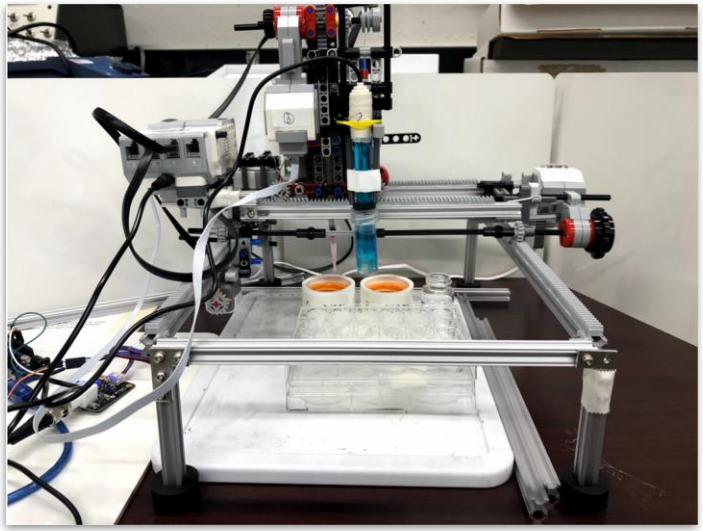
Kusne, et al. *Nature Communications* 11.1 (**2020**)

Other Works:

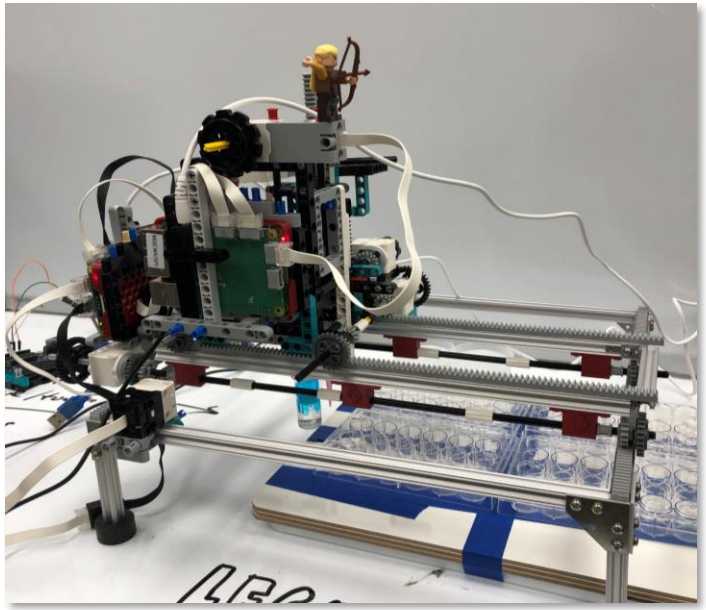
Stach, (**2021**). Autonomous experimentation systems for materials development: A community perspective. *Matter*, 4(9), 2702-2726.

Low Cost Autonomous Physical Science System





EV3



Raspberry Pi

Exploring pH of Buffer Solutions

Composition Space

Weak Acid - *Acetic Acid* - 1 M

Conjugate Base - *Sodium Acetate Solution* - 1 M

Goal

Recover Henderson-Hasselbalch Equation from data.

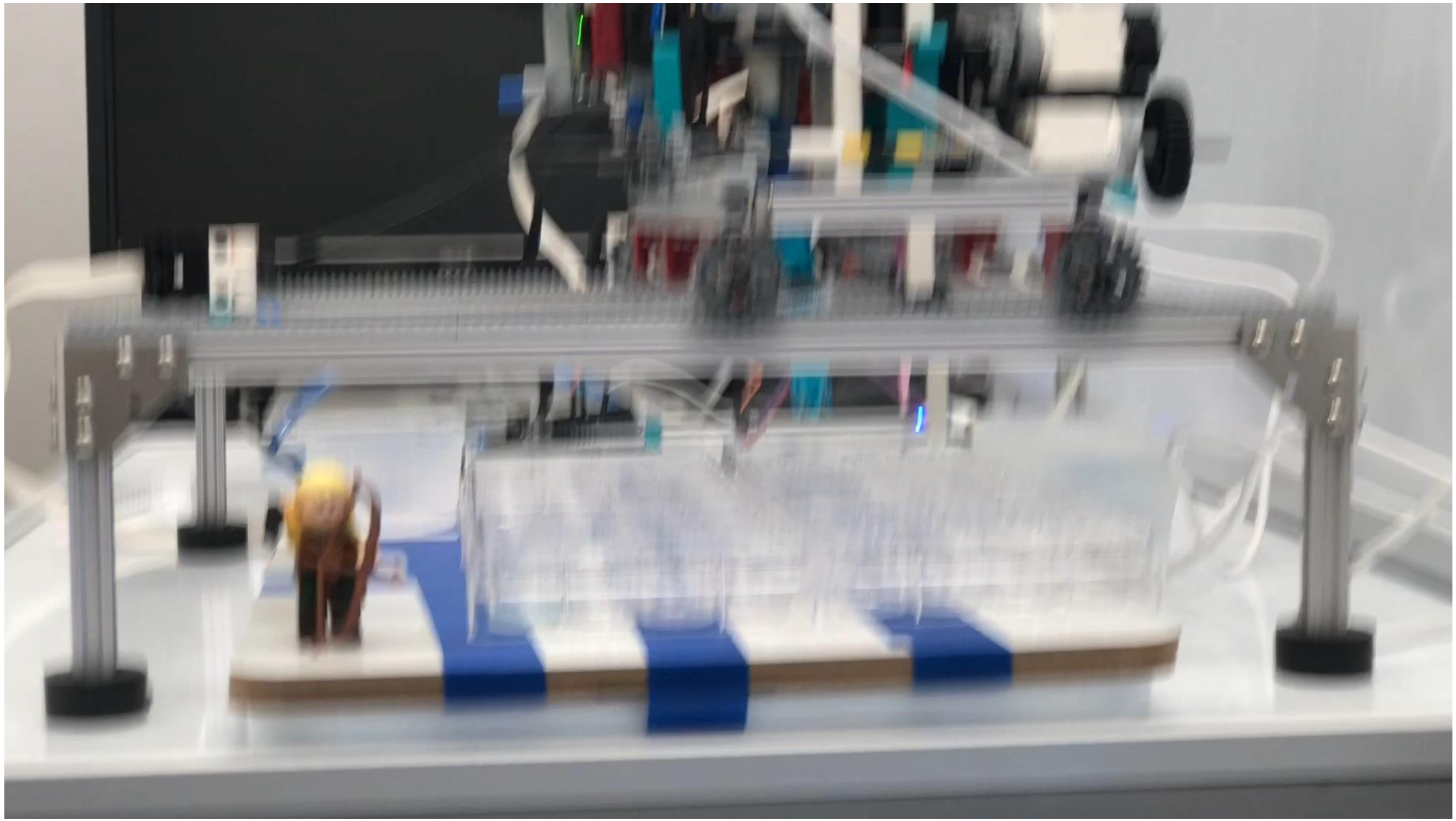
$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log_{10}\left(\frac{[\mathrm{Base}]}{[\mathrm{Acid}]}\right)$$

Our known parameters (for synthesis)

Response Variable (measured)

Dissociation Constant (unknown to robot)

Active Learning Closed Loop System – Rasp. Pi

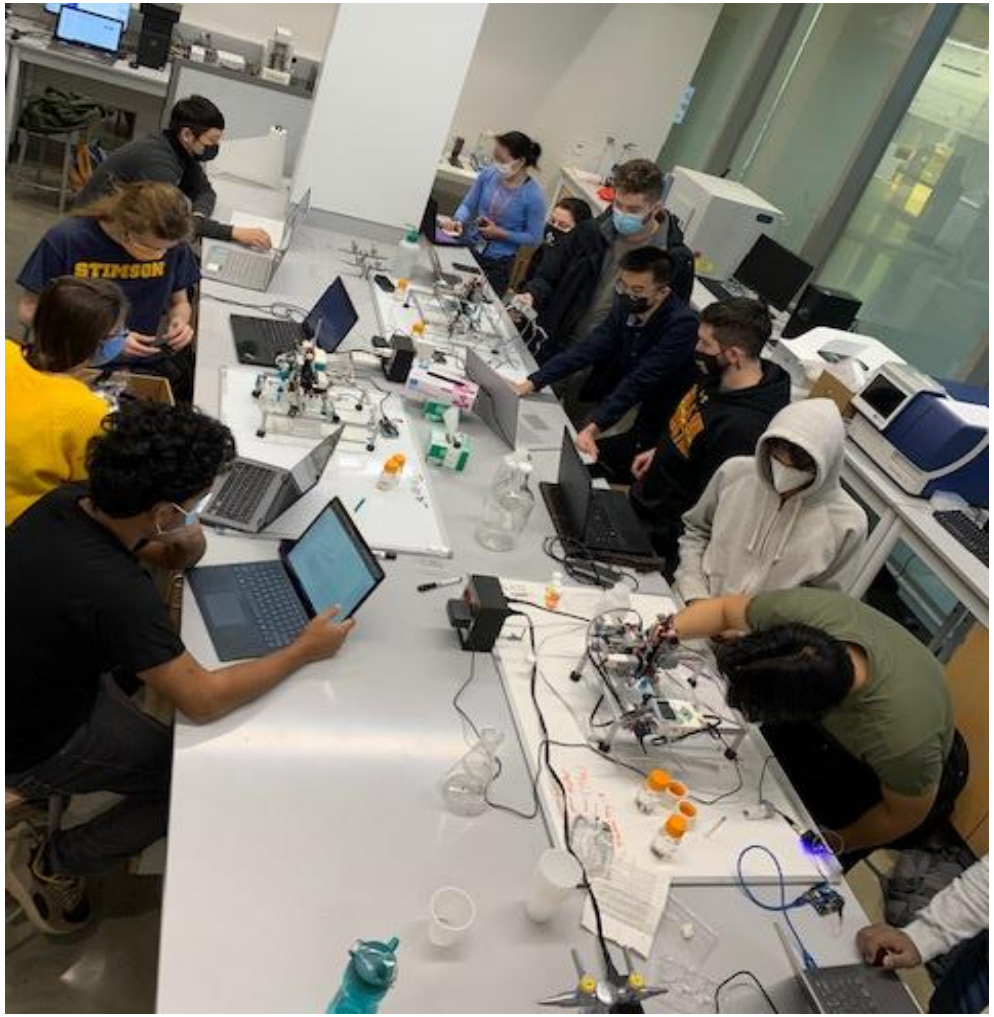


Educational Application (Fall 2021 ENMA 437/637)

UMD Machine Learning for Materials Science Course

Concepts and Challenges

- Acquisition Functions (Exploration/Exploitation)
- Gaussian Processes
- Hardware/Robotics
- Limitations (discretization of compositions, hardware, etc.)

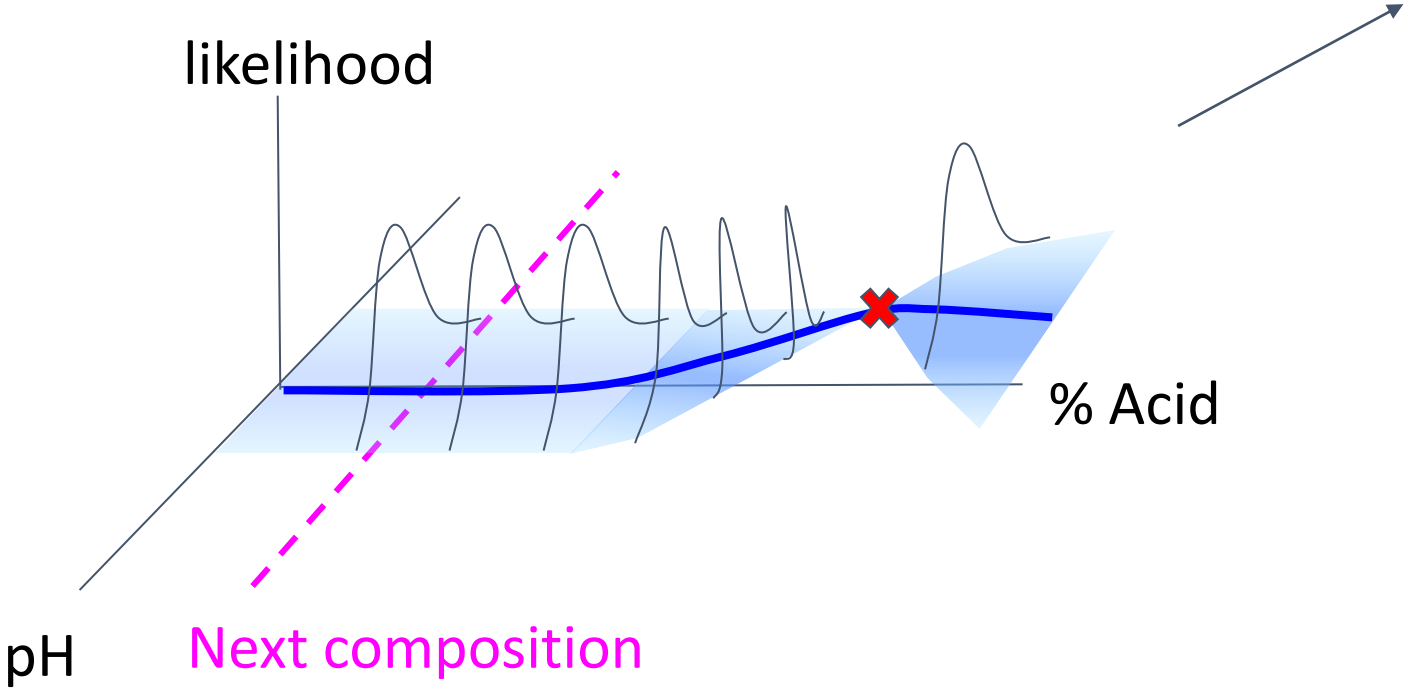


Exploration & Exploitation

Can the Robot Explore the Relationship between Composition & Properties?

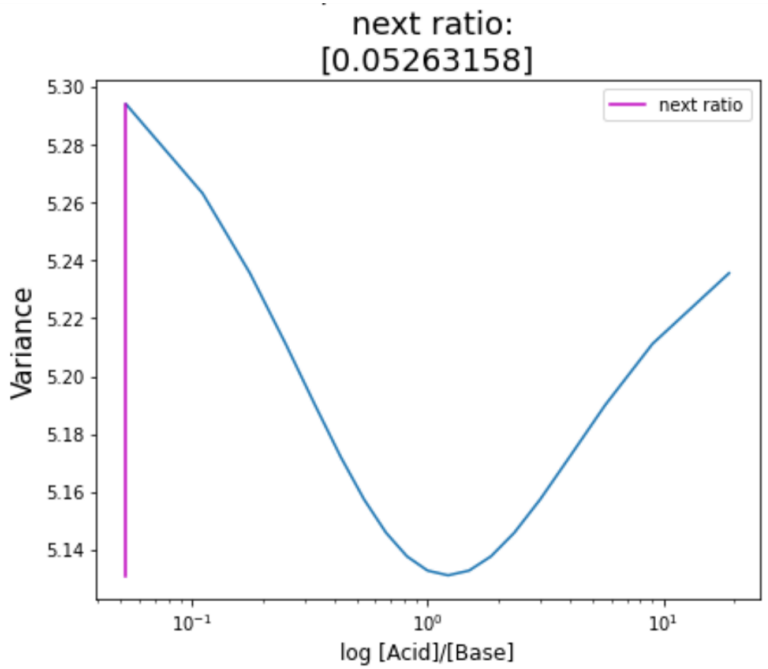
Can it use that understanding to prepare a sample with a particular properties?

Exploration Initiative - (Gaussian Process)



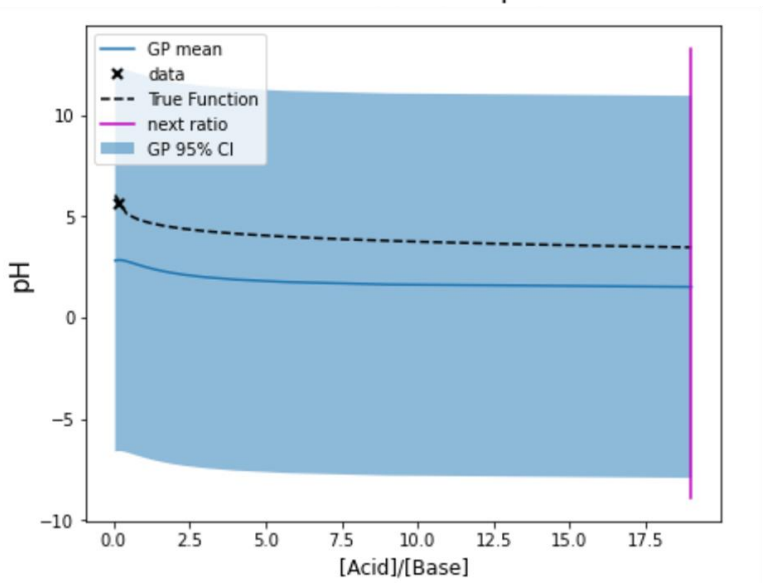
Active Learning:

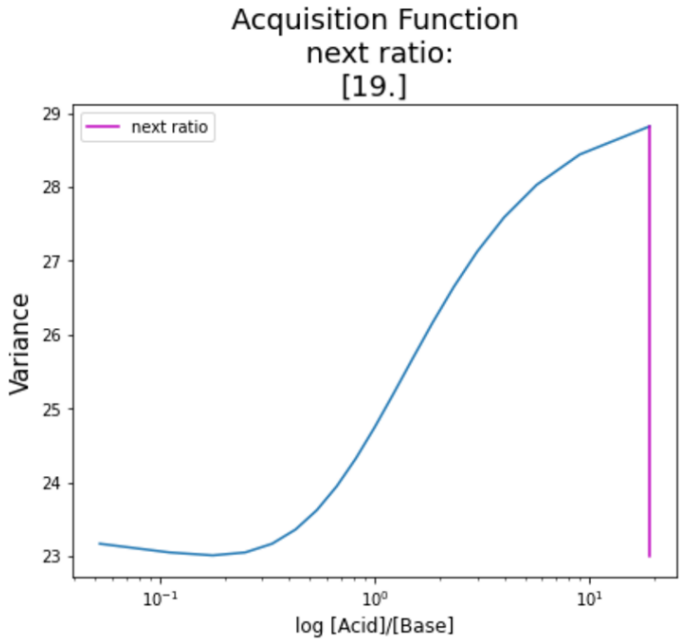
- ➔ Acquisition Function
- ➔ *argmax (variance)*



Autonomous Results - (Gaussian Process)

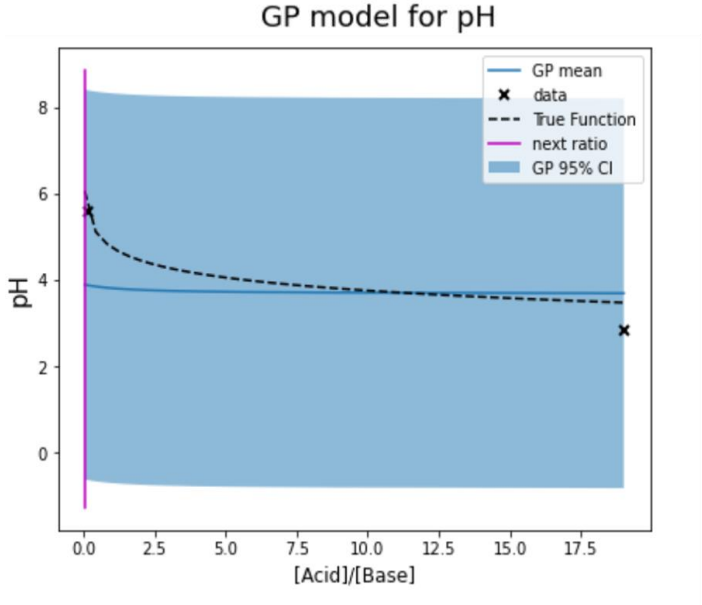
1 data point

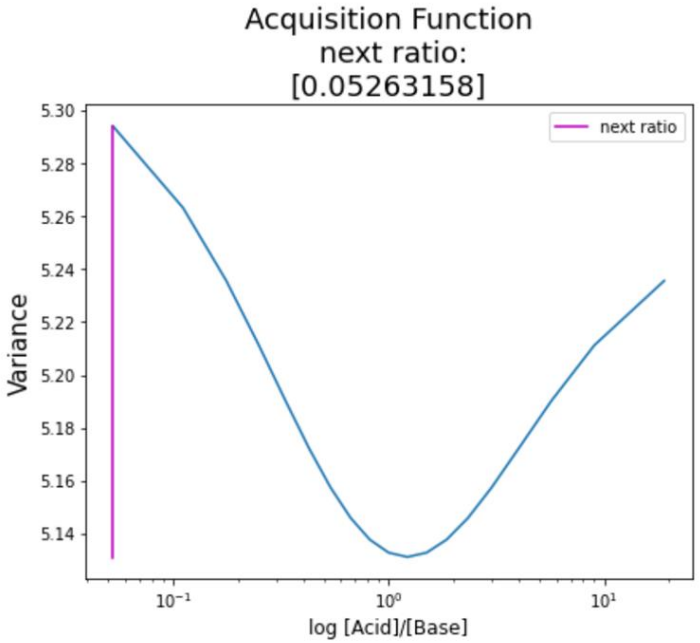




Autonomous Results - (Gaussian Process)

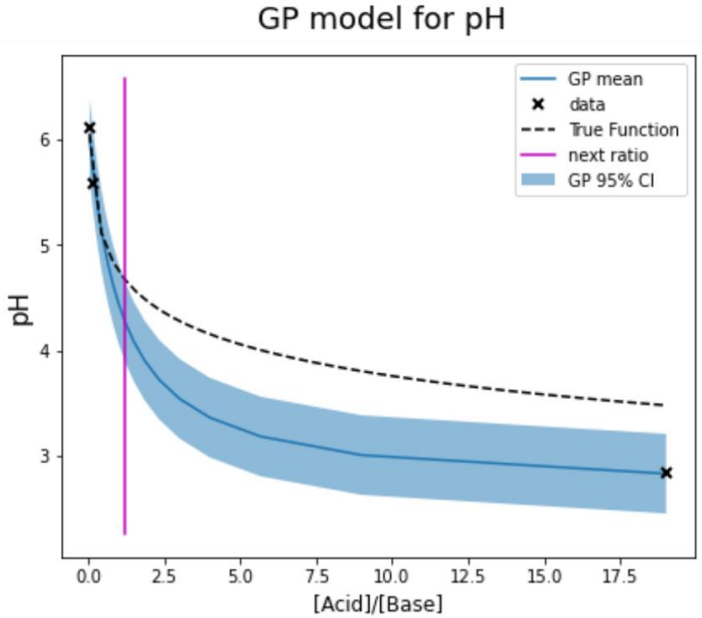
2 data points

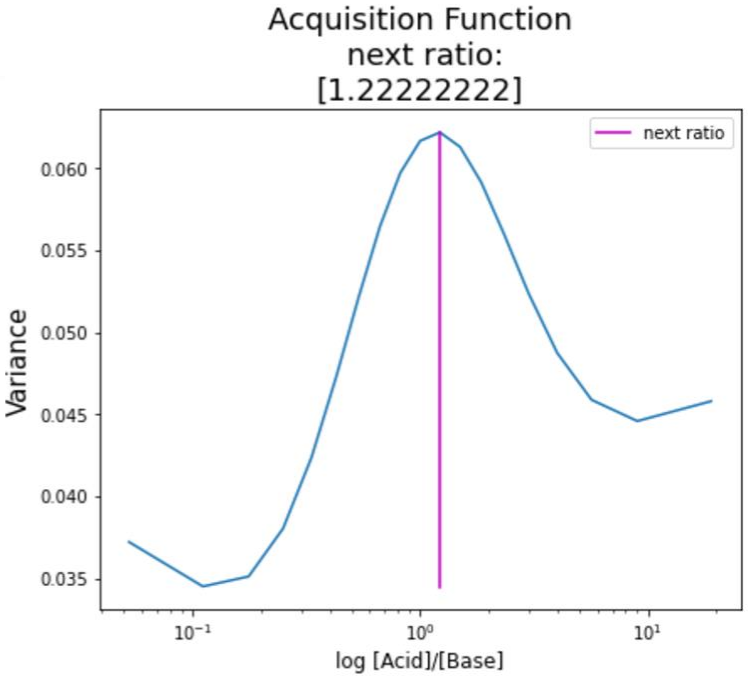




Autonomous Results - (Gaussian Process)

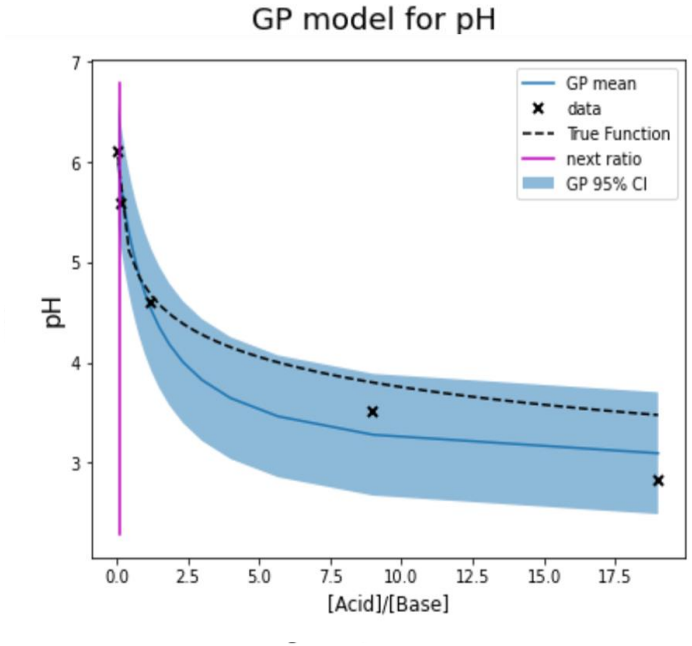
3 data points

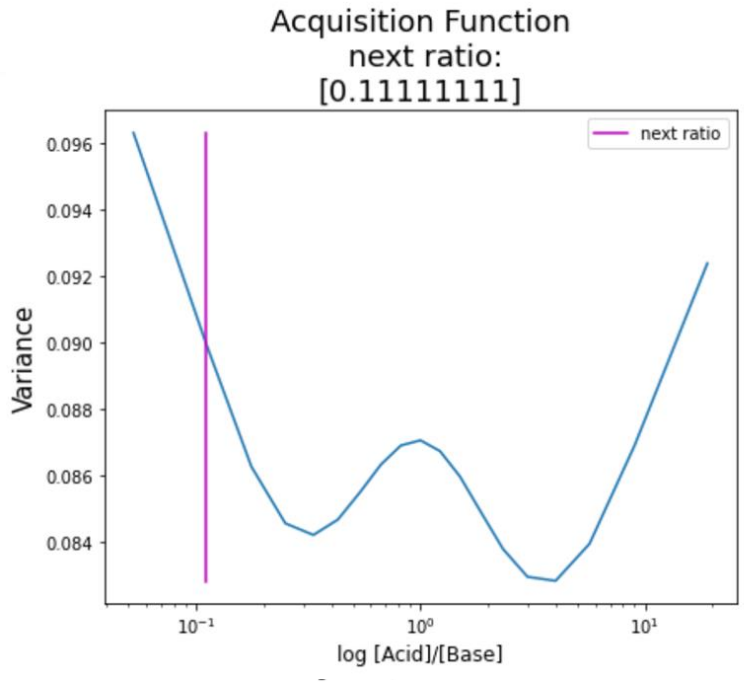




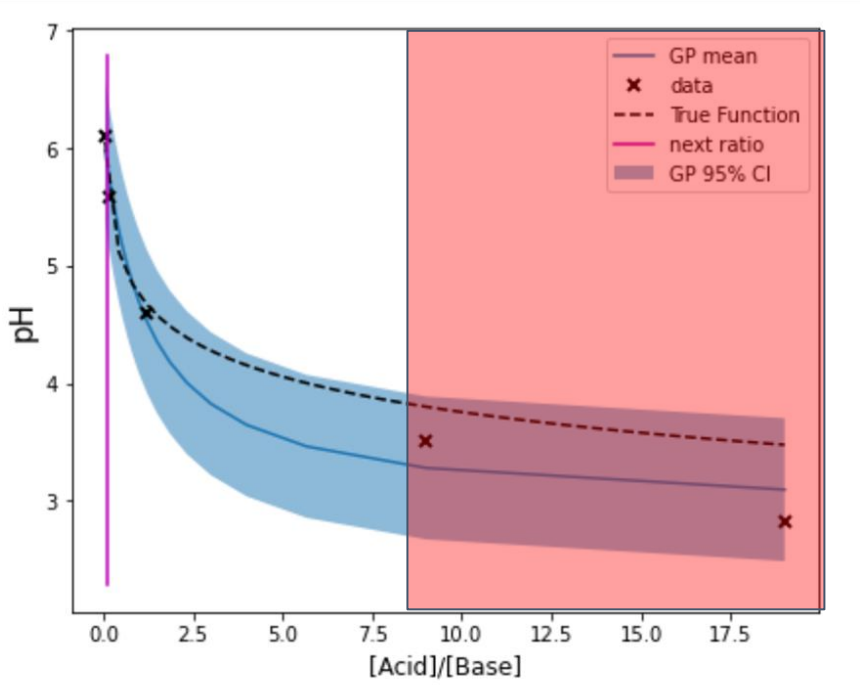
Autonomous Results - (Gaussian Process)

5 data points





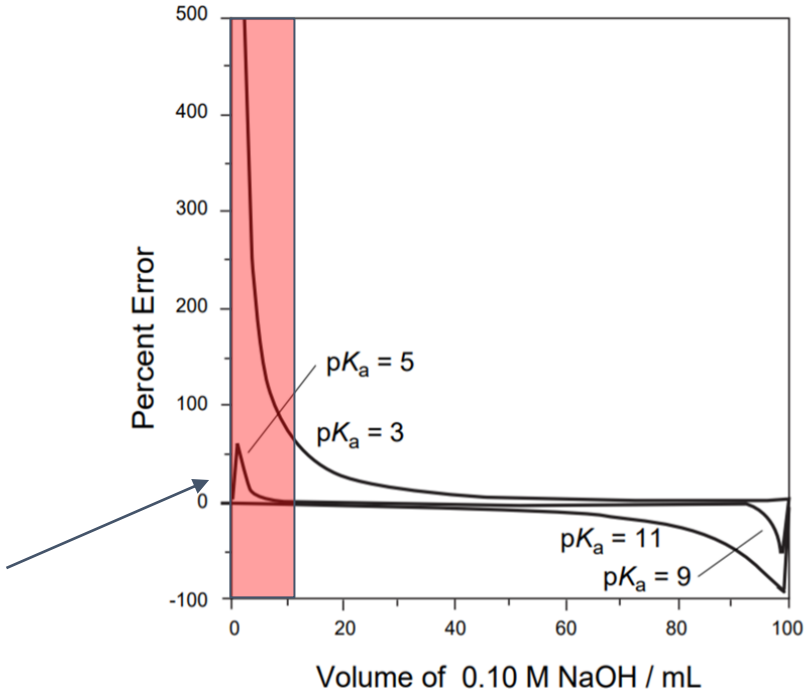
GP flexibility - (Gaussian Process)



↑↑ % error in HH simplification ...

HH equation relies on assumptions

- ➔ No self-ionization of water
- ➔ Valid only in certain composition range
- ➔ pKa ~ 4.7



Parameter Determination

If we start with a hypothetical model, can the robot determine the best parameters?

Brief Overview - Bayesian Inference

Probabilistic interpretation … quantifying **uncertainty** (how confident are we?)

Bayes Theorem

$$\mathbf{P(model \mid data)} = \frac{P(\text{data} \mid \text{model})\,\mathbf{P(model)}}{P(\text{data})}$$

"Posterior"

Our confidence in this model being "correct" given the data (what we want to know)

"Prior"

Our confidence in this model being "correct" before getting data (assumption)

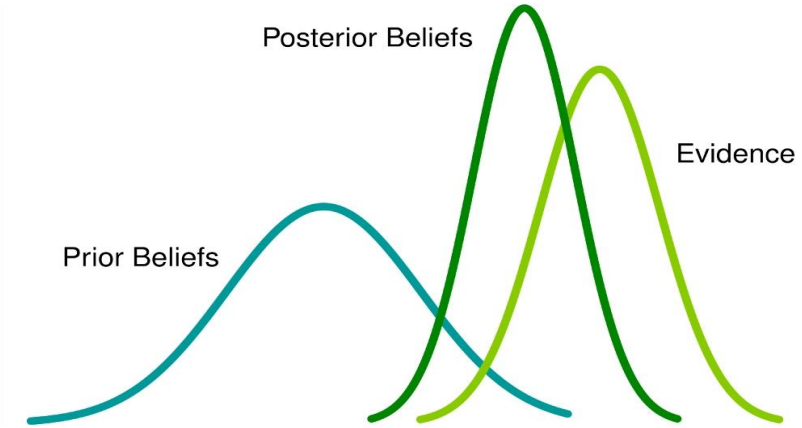
Parameter Refinement - (Bayesian Inference)

Prior: Assume model has logarithmic form (pH = A + B*log(x))

➔ A and B are our model parameters

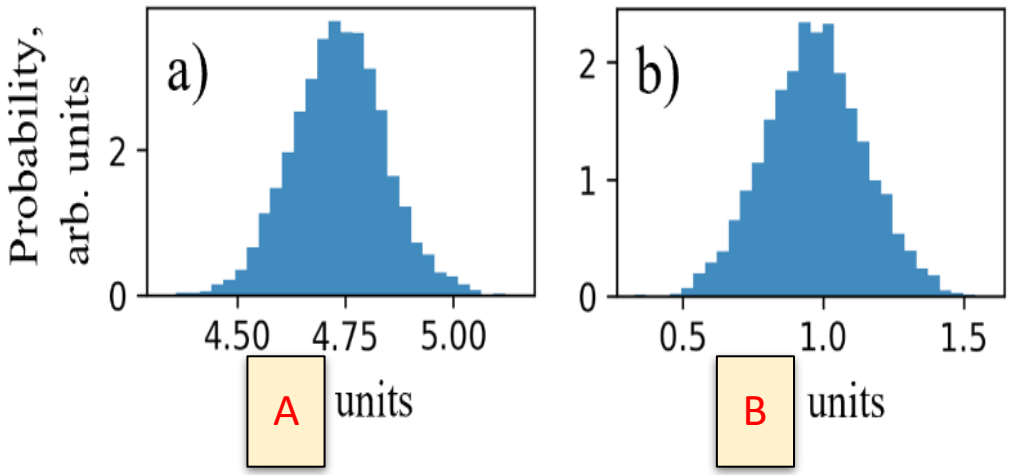
Posterior: Probability of this model and its model parameters given the data

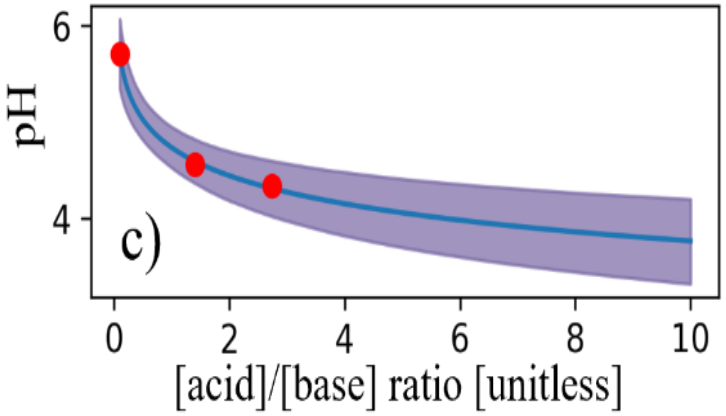
New data alters our prior beliefs → posterior beliefs



Autonomous Results - (**Bayesian Inference**)

→ Solve for Posterior using MCMC
→ Confidence interval based on MCMC sampling





pH = A + B*log(x)

→ Can use parameter uncertainty to decide which composition to measure next

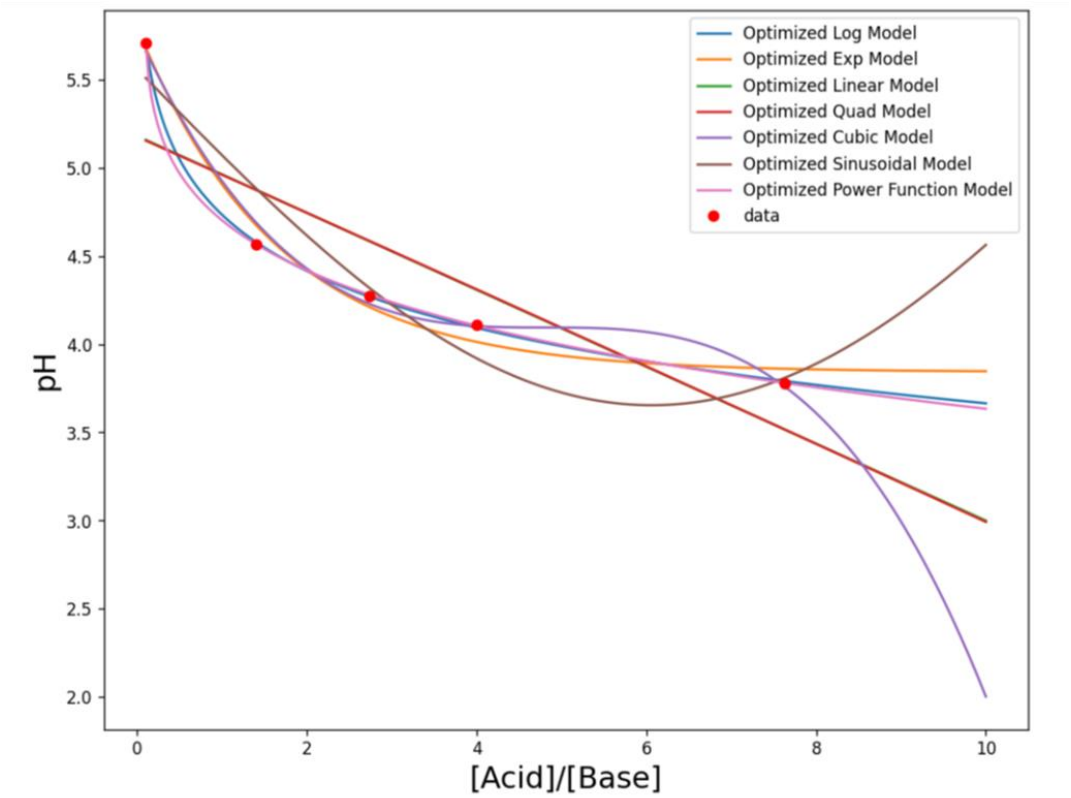
ANDiE: the Autonomous Neutron Diffraction Explorer. *McDannald, et. al, 2021*

Model Determination

Can the Robot Determine the Physical Law by Itself?

Can the Robot Discover the Physical Law?

1. **Fit multiple functional forms to the data ("Candidates")**
- (sinusoidal, power function, logarithmic, exponential, quadratic, etc.)
- Non-linear least squares regression



x = [Acid]/[Base]

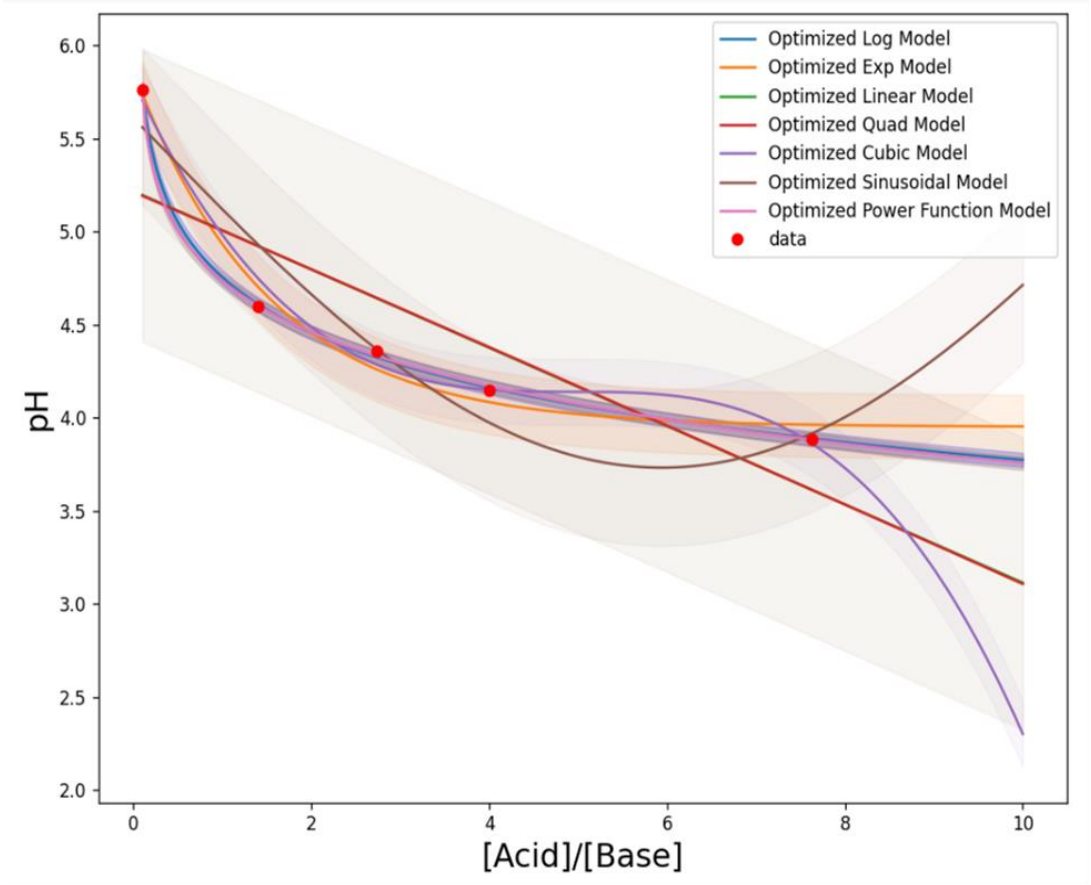
What is the correct form?

pH = A + B * **log** [C * (x-D)] ?
pH = A + B * **sin** [C * (x-D)] ?
pH = A + B * **exp**[C * (x-D)] ?
…
Alter parameters to get best fit

Can the Robot Discover the Physical Law?

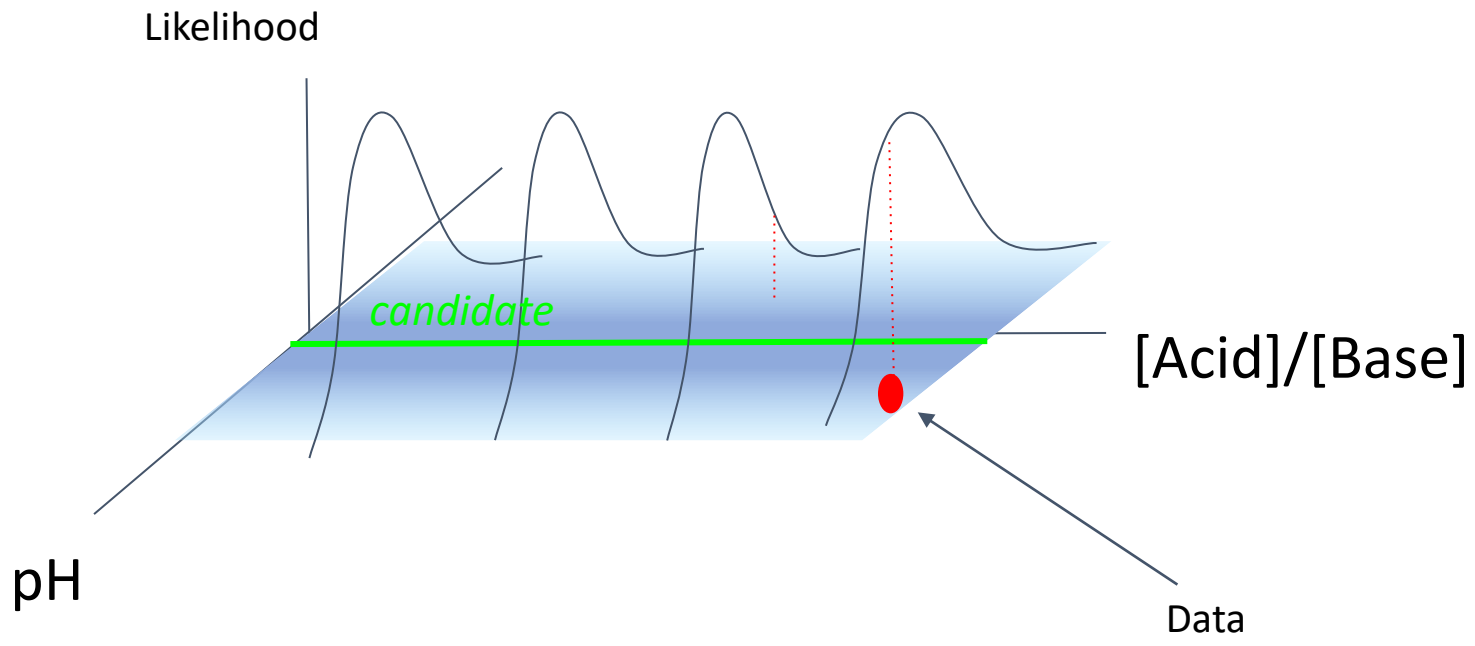
2. Create PDF for each candidate at every composition

- (std. of PDF given by std. of residuals)
- Better models have narrow distributions, Worse are broad



Can the Robot Discover the Physical Law?

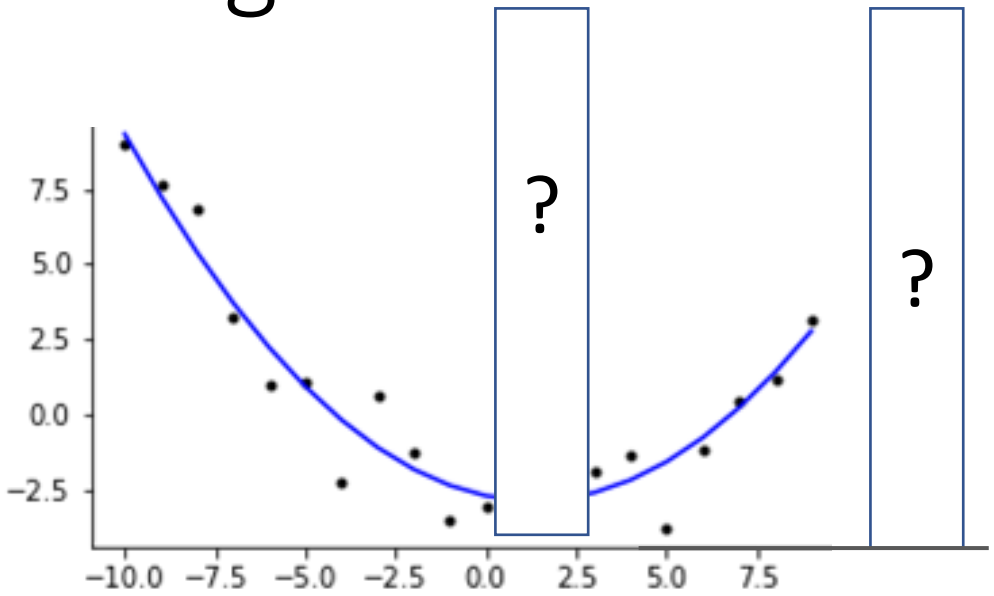
3. Rank the total likelihood that each candidate model produced the data

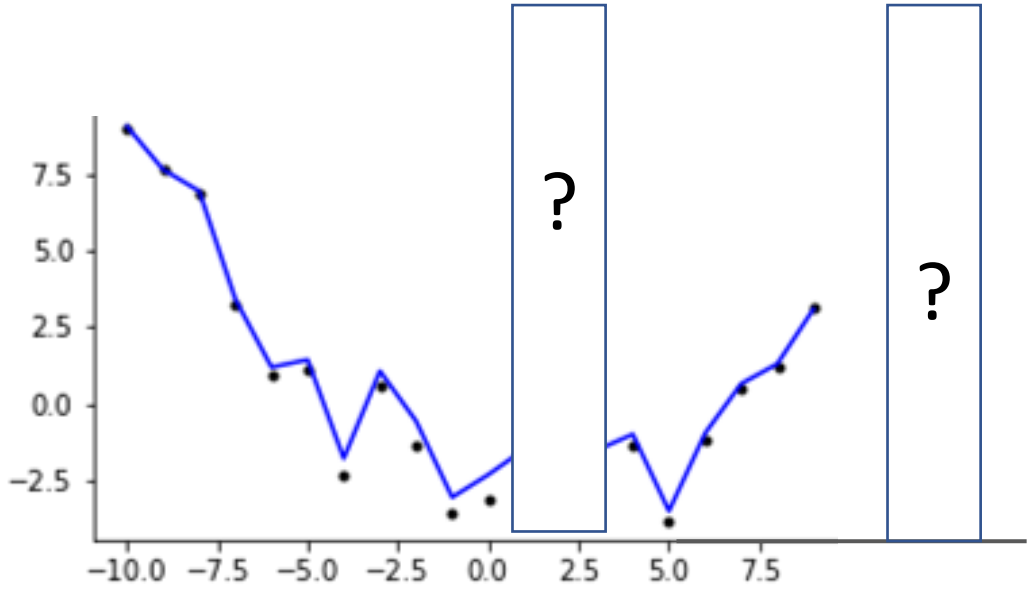


Performance Metric for each candidate is the total likelihood [sum of log(likelihood)] along every collected data point

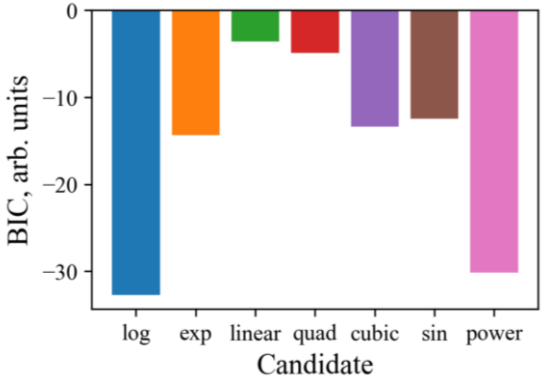
Candidates with least certainty will have lower total likelihood

Overfitting





→ Prior method only considers goodness of fit
→ Max(sum of log likelihoods)
→ Power function emerges as best model

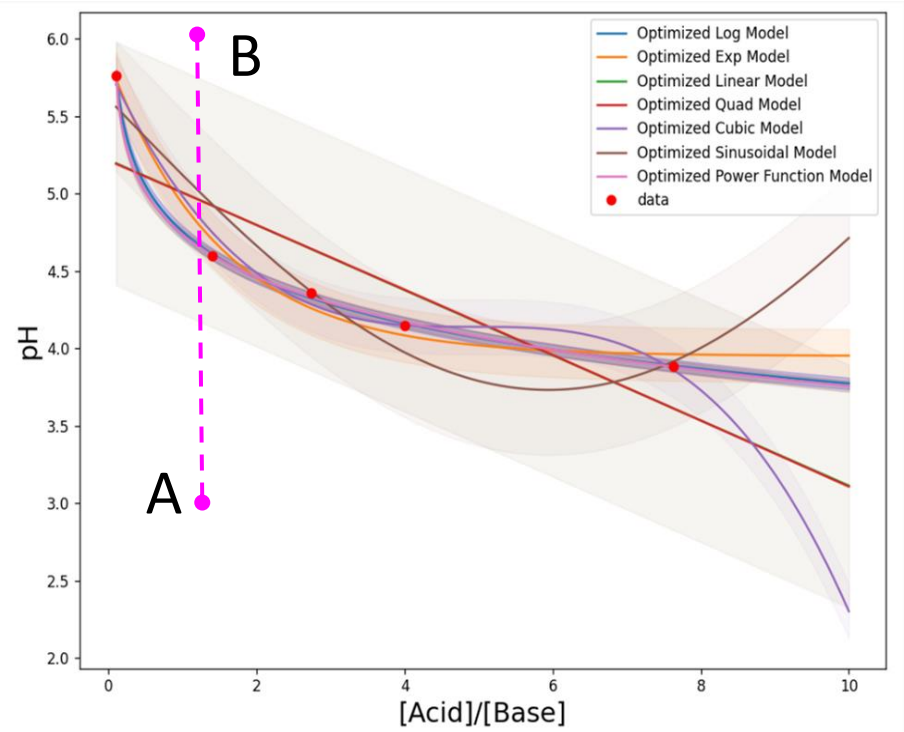


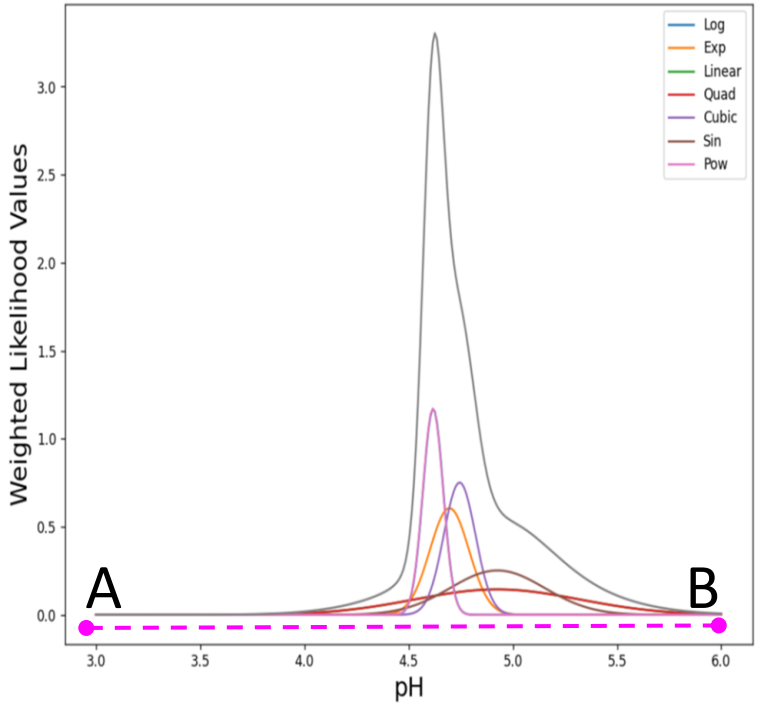
$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

→ BIC considers # of model parameters (n)
→ Min(BIC), log function is best model

Can the Robot Discover the Physical Law?

4. Create a cumulative distribution of all PDFs at each composition





Yes!

6. Determine which composition to measure next

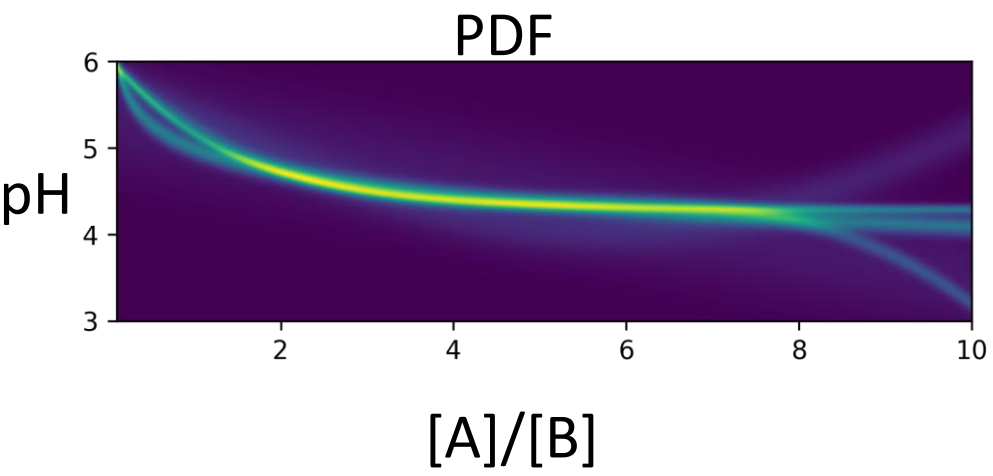
- Look where candidate differ the most
- Better candidates weighted more

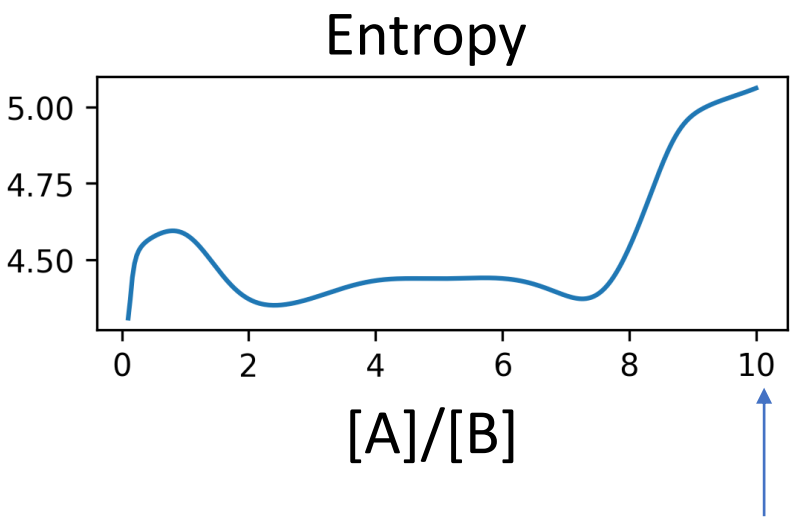
After 5 measurements:

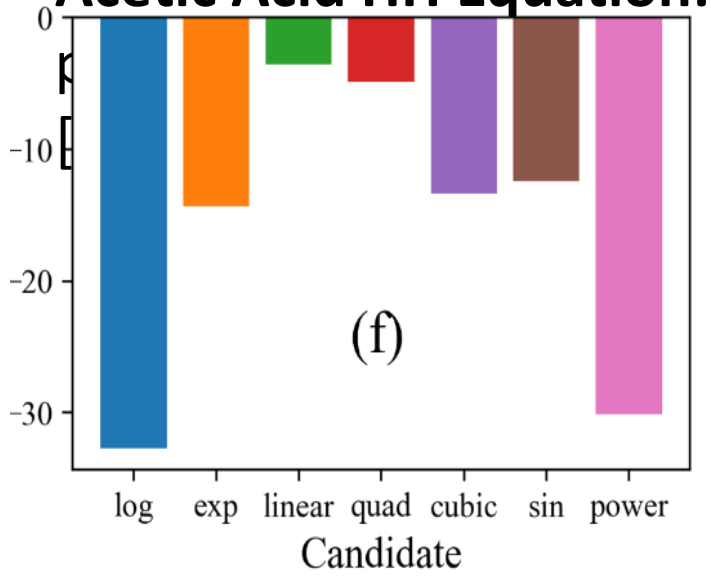
Top Ranked Model

pH = 4.753 + 1.02 * **log** [A/B]

Acetic Acid HH Equation:





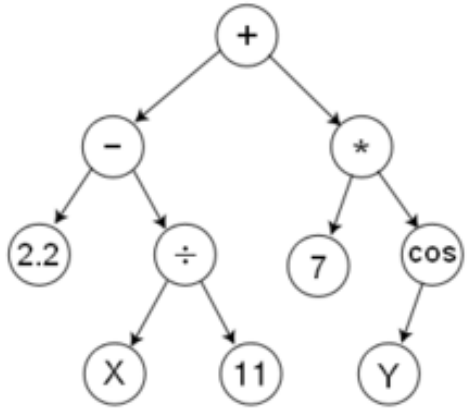


Model Generation (Symbolic Regression)

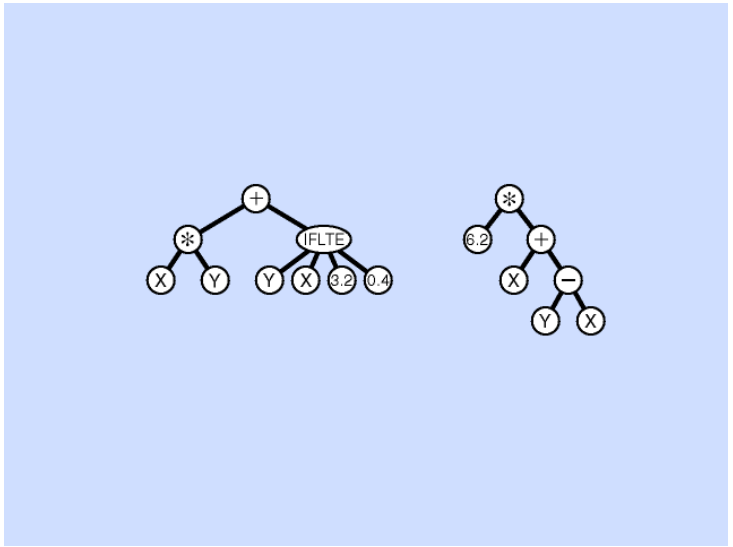
Can the Robot also self-generate explanatory models?

Symbolic Regression Overview

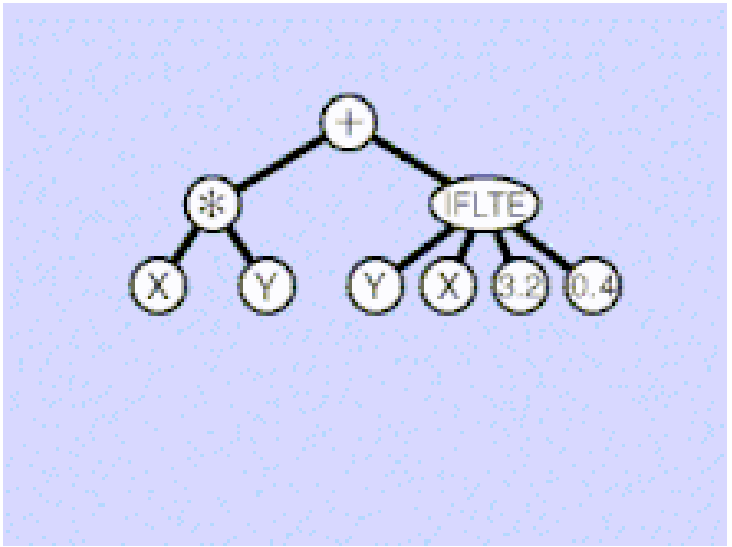
- Genetic programming
- “mates” best functions during fitting
- “mutations” possible as well
- Generates potential explanatory functions



$$\left(2.2-\left(\frac{X}{11}\right)\right)+\left(7*\cos(Y)\right)$$



mating



mutating

Application to Autonomous Physical Science

Set of “candidates” generated after each data point measurement

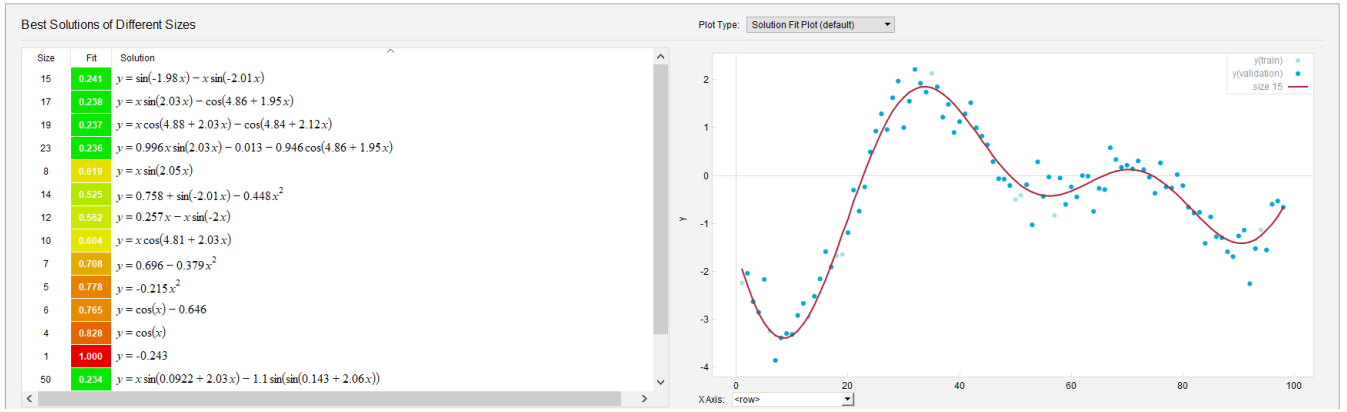
Penalize complexity

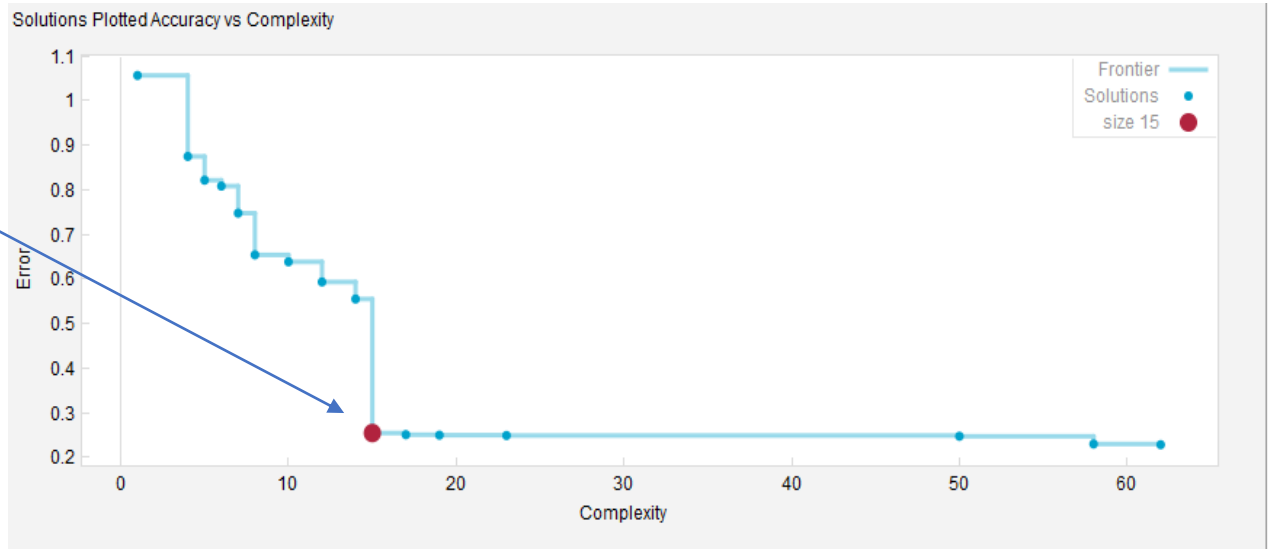
→ Occam’s Razor

→ Prevent overfitting

Best candidates lie along Pareto Front (Error vs. Complexity)

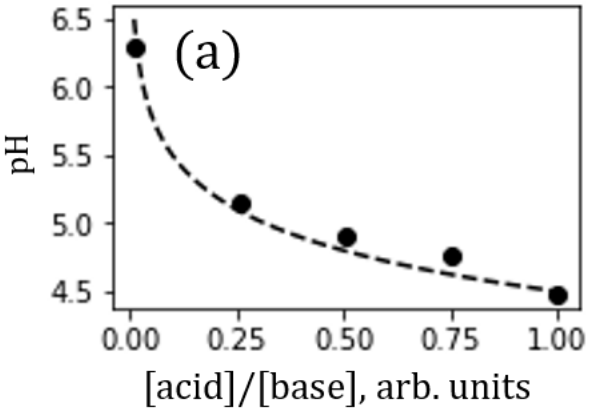
Can choose Acquisition Function





Eureqa Free Trial

Results



b)

Complexity	MSE	Score	Equation
1	0.518	0.000	5.076
4	0.361	0.121	4.253+cos(r)
5	0.106	1.222	6.001-1.833r
6	0.007	2.772	4.483-0.984*log(r)
11	0.004	0.093	2.886+cos(r)+exp[exp[-6.387*sin(r)]]

Figure 5. Symbolic regression combined with active learning for probabilistic model determination. a) example data, b) output from symbolic regression with 5 models. The model with the highest score matches the HH equation with a slight deviation of parameters.

Generated candidate of same functional form as HH equation with one input variable

→ Logarithmic candidate had the highest score

→ Candidate with lowest error penalized due to complexity (4 internal functions)

Publications

DOI: 10.48550/arXiv.2204.04187 • Corpus ID: 248069191

A Low-Cost Robot Science Kit for Education with Symbolic Regression for Hypothesis Discovery and Validation

Logan Saar, Haotong Liang, +4 authors A. Kusne • Published 8 April 2022 • Education • ArXiv

The next generation of physical science involves robot scientists – autonomous physical science systems capable of experimental design, execution, and analysis in a closed loop. Such systems have shown real-world success for scientific exploration and discovery, including the first discovery of a best-in-class material. To build and use these systems, the next generation workforce requires expertise in diverse areas including ML, control systems, measurement science, materials synthesis… Expand

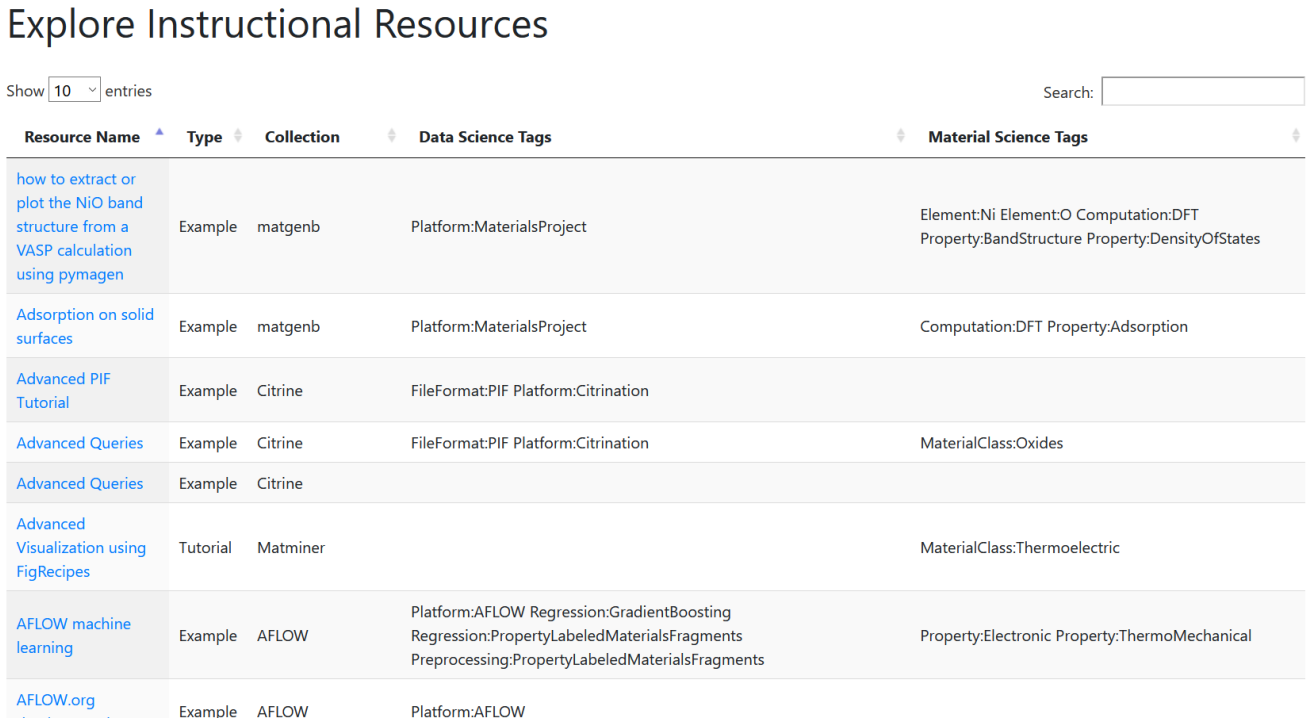
→ MRS Bulletin August Edition
→ Available now on *Arxiv*

REMI: REsource for Materials Informatics

- Code in many different platforms, languages, etc.
- Centralize in a curated, searchable list.
- REMI is open source.

- Please Submit!!

pages.nist.gov/remi



Explore Instructional Resources

Show 10 entries

Search:

Resource Name	Type	Collection	Data Science Tags	Material Science Tags
how to extract or plot the NiO band structure from a VASP calculation using pymagen	Example	matgenb	Platform:MaterialsProject	Element:Ni Element:O Computation:DFT Property:BandStructure Property:DensityOfStates
Adsorption on solid surfaces	Example	matgenb	Platform:MaterialsProject	Computation:DFT Property:Adsorption
Advanced PIF Tutorial	Example	Citrine	FileFormat:PIF Platform:Citrination	
Advanced Queries	Example	Citrine	FileFormat:PIF Platform:Citrination	MaterialClass:Oxides
Advanced Queries	Example	Citrine		
Advanced Visualization using FigRecipes	Tutorial	Matminer		MaterialClass:Thermoelectric
AFLOW machine learning	Example	AFLOW	Platform:AFLOW Regression:GradientBoosting Regression:PropertyLabeledMaterialsFragments Preprocessing:PropertyLabeledMaterialsFragments	Property:Electronic Property:ThermoMechanical
AFLOW.org	Example	AFLOW	Platform:AFLOW	

Questions