*Data science for Materials Science & Engineering*

# Active learning for materials discovery

**In this module**

- Introduction to active learning for materials science (this lecture)
- Hands on tutorial using nanoHUB: high Li+ conductivity oxides
- Homework assignment

Alejandro Strachan, Juan Carlos Verduzco
strachan@purdue.edu
School of Materials Engineering & Network for Computational Nanotechnology
Purdue University
West Lafayette, Indiana USA

# Learning objectives and prerequisites

After completing this lecture you will:

- Be able to use and modify active learning workflows
- Evaluate different information acquisition functions
- Use active learning to reduce the number of experiments in materials discovery or design

Pre-requisites:

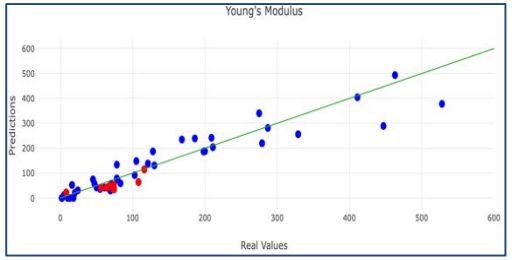
- Basic Jupyter and Python (https://nanohub.org/resources/33266)
- Querying materials repositories (in this series)

# Data Science & Machine Learning in Sci & Eng

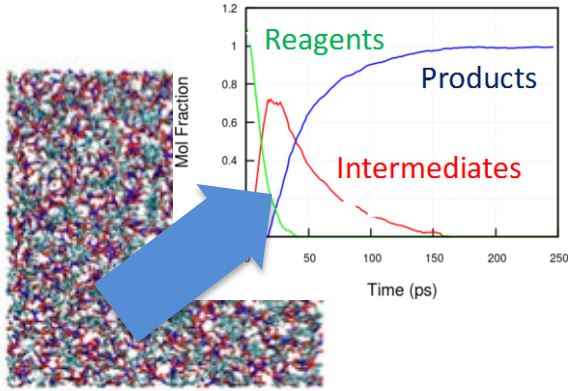
## Learning from data

### Acquiring and handling data

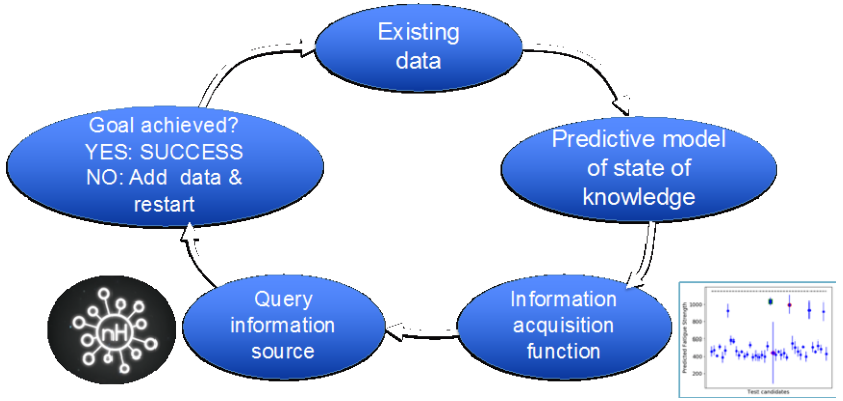
### Predictive models (supervised learning)




### Finding patterns (unsupervised learning)




### Design of experiments




### Cyber-infrastructure

nanoHUB

CITRINE INFORMATICS

The Materials Project

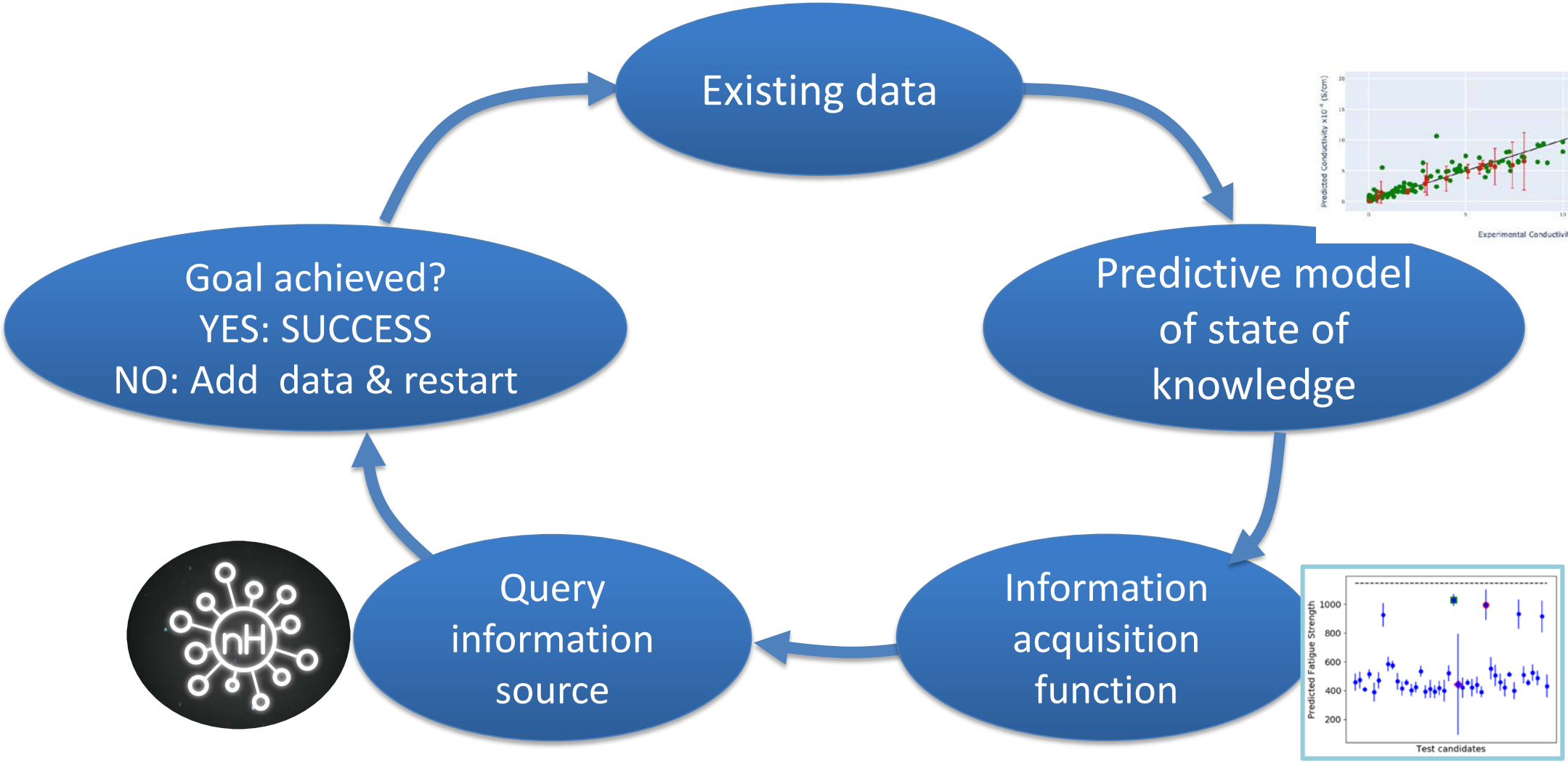
WolframAlpha

KIM

MATERIALS DATA FACILITY

# Data science for design of experiments

Can ML help reduce the number of experiments to achieve a desired goal?

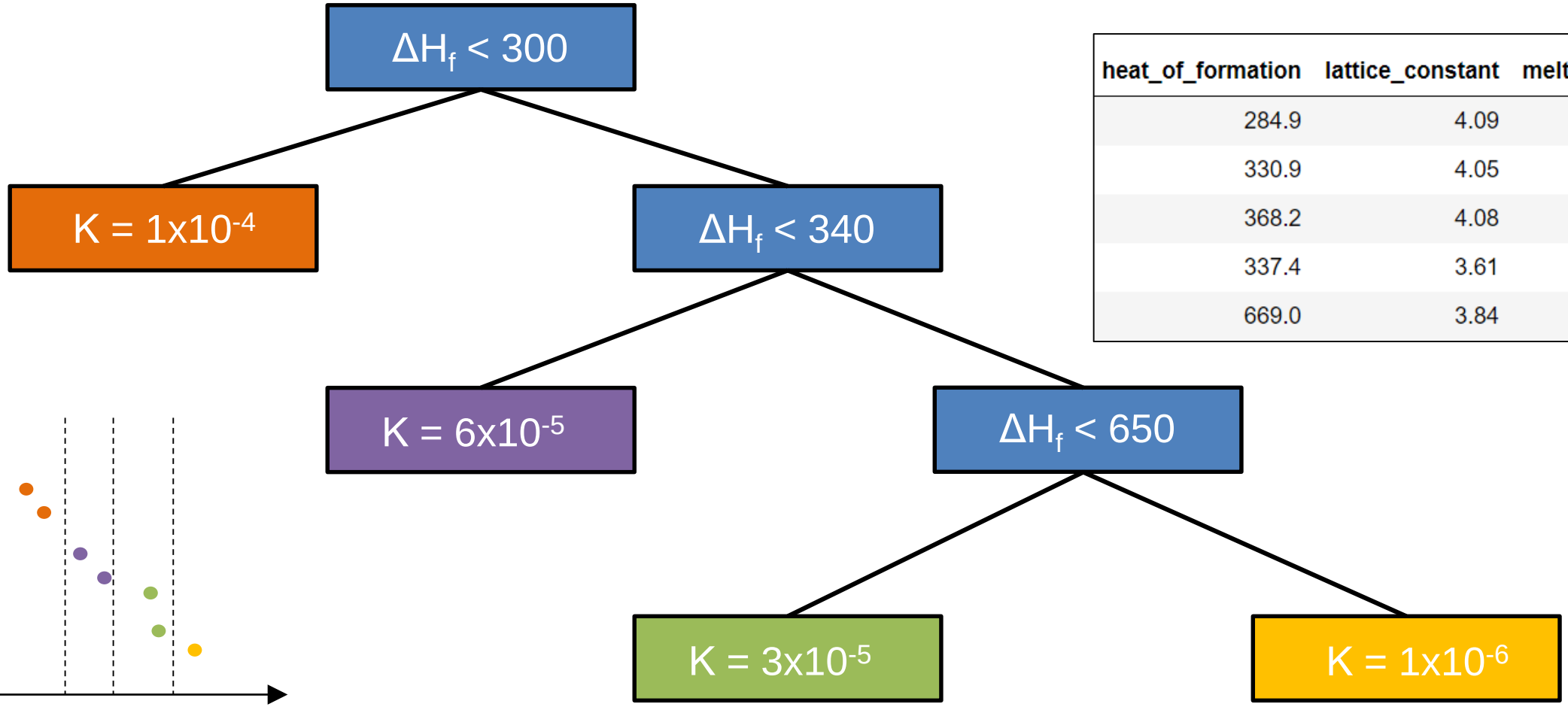




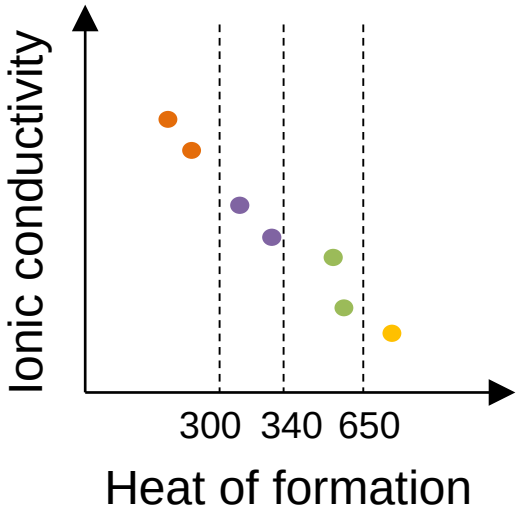
Pandita P, Bilionis I, Panchal J. Journal of Mechanical Design. 2016 138, 111412.

Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. Int. Mat. and Manuf. Inn.. 2017 Sep 1;6(3):207-17.

# Predictive models: Decision trees & random forests



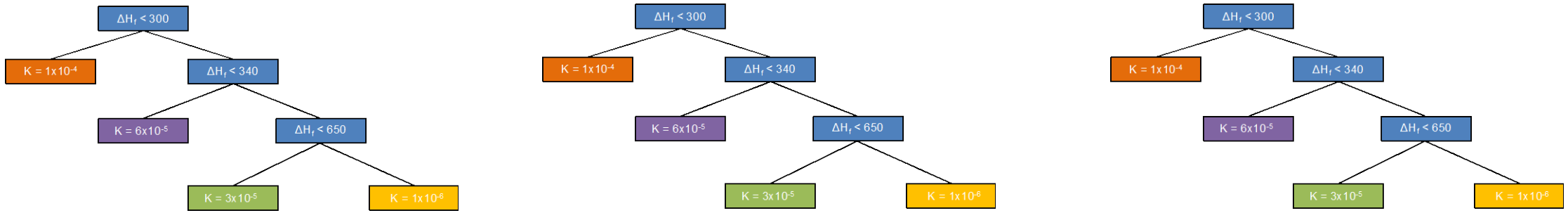

| heat_of_formation | lattice_constant | melting_point |
|---|---|---|
| 284.9 | 4.09 | 1235.10 |
| 330.9 | 4.05 | 933.50 |
| 368.2 | 4.08 | 1337.58 |
| 337.4 | 3.61 | 1356.60 |
| 669.0 | 3.84 | 2683.00 |



# Random forests

| heat_of_formation | lattice_constant | melting_point | specific_heat | atomic_mass | atomic_radius | electrical_resistivity |
|---|---|---|---|---|---|---|
| 284.9 | 4.09 | 1235.10 | 0.237 | 107.868200 | 1.60 | 1.630000e-08 |
| 330.9 | 4.05 | 933.50 | 0.900 | 26.981539 | 1.25 | 2.700000e-08 |
| 368.2 | 4.08 | 1337.58 | 0.129 | 196.966569 | 1.35 | 2.200000e-08 |
| 337.4 | 3.61 | 1356.60 | 0.385 | 63.546000 | 1.35 | 1.720000e-08 |
| 669.0 | 3.84 | 2683.00 | 0.133 | 192.217000 | 1.35 | 4.700000e-08 |

Random forests are ensembles of decision trees




Choose training data for each tree randomly

Choose features for each tree randomly

# Random forest: expected value and uncertainties

Expected value: mean over all decision trees

Number of trees

$$E[M(x)] = \frac{1}{N}\sum_{J}^{N_T} T_J(x)$$

*J* Tree prediction

Tree index

Uncertainty estimate at test point *x*

Test point

$$\sigma[M(x)] = \sqrt{\sum_{i=1}^{S} \max\left[\sigma_i^2(x), \omega\right] + \tilde{\sigma}^2(x)}$$

Training points, *i*

Explicit bias

Variance at point *x* due to training point *i*

Threshold

$$\omega = \min_i \sigma^2[M(x_i)]$$

$$\sigma_i^2(x) = \mathrm{Cov}_J\left[n_{i,J}, T_J(x)\right]^2 + \left[\bar{T}_{-i}(x) - \bar{T}(x)\right]^2 - \frac{ev}{N_T}$$

Number of times *i* was used to train *J*

Tree index

Average over trees not using *i*

Average over all trees

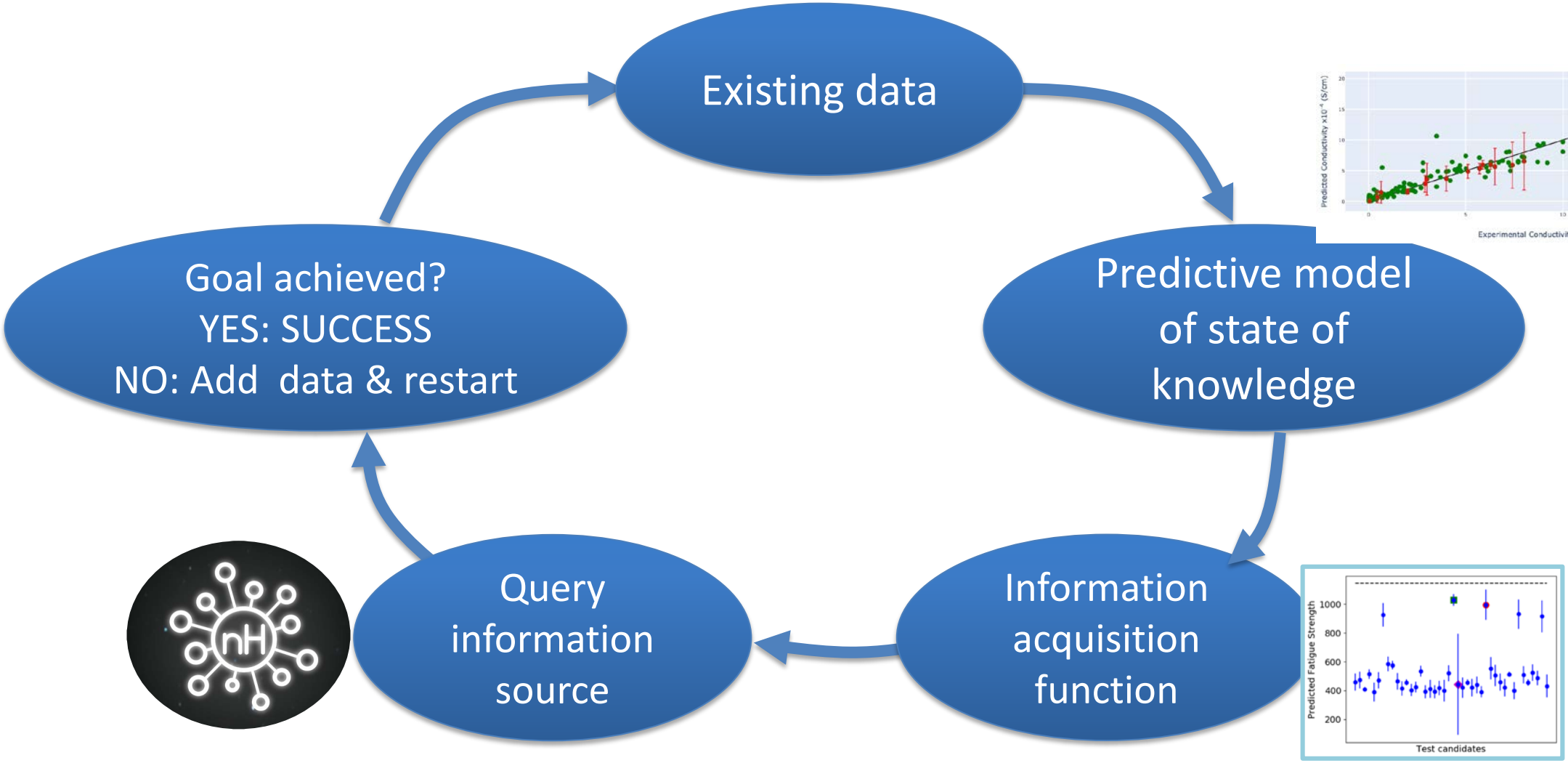
Euler number

Variance overall all trees

Wager S, Hastie T, Efron B (2014) J Mach Learn Res 15:1625–1651.
Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. Int. Mat. and Manuf. Inn.. 2017 Sep 1;6(3):207-17.

# Data science for design of experiments

Can ML help reduce the number of experiments to achieve a desired goal?

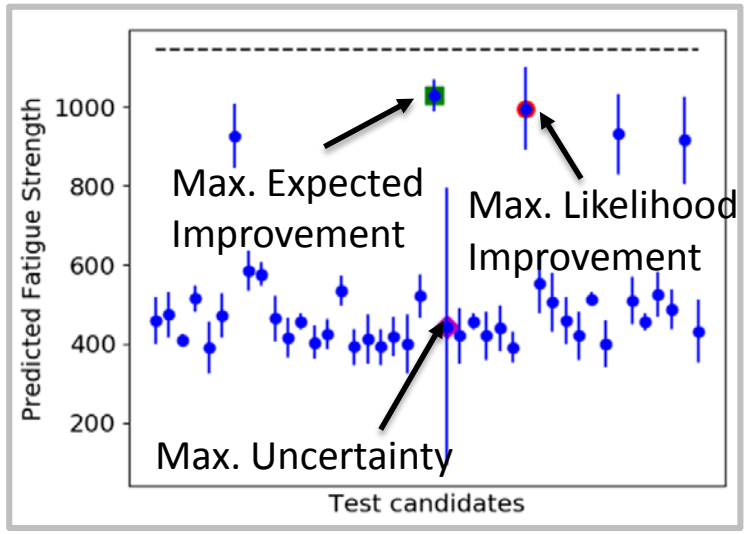




Pandita P, Bilionis I, Panchal J. Journal of Mechanical Design. 2016 138, 111412.

Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. Int. Mat. and Manuf. Inn.. 2017 Sep 1;6(3):207-17.

# Information acquisition functions




What experiment should I carry out next?

## Maximum expected improvement

$$x^* = \arg_max\ E\ [M(x_i)]$$

Expected value

Possible experiments

Model (in this case RF)

## Maximum likelihood of improvement

$$x^* = \arg_max \frac{E[M(x_i)] - E[M(x_{best})]}{\sigma[M(x_i)]}$$

Uncertainty estimate

Best current case

## Maximum uncertainty

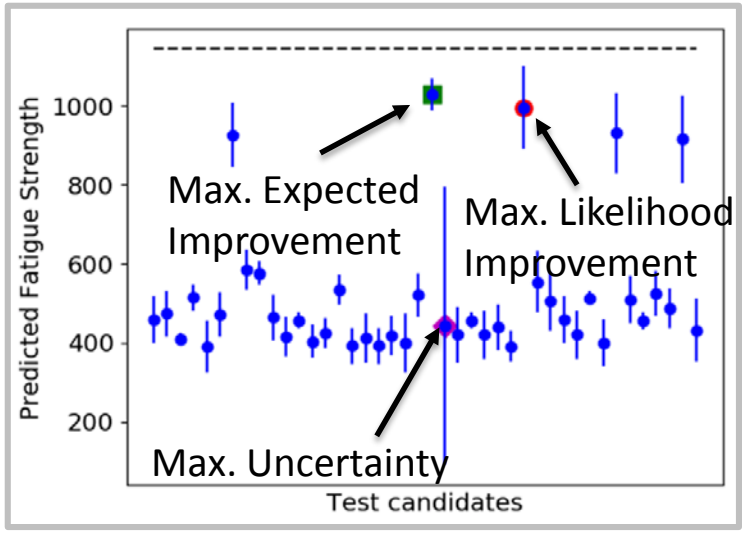
$$x^* = \arg_max\ \sigma\ [M(x_i)]$$

Uncertainty estimate

Possible experiments

Model (in this case RF)

# Information acquisition functions




Maximum expected improvement

$$x^* = \arg_max\, E\,[M(x_i)]$$

Maximum uncertainty

$$x^* = \arg_max\, \sigma\,[M(x_i)]$$

Upper Confidence Bound

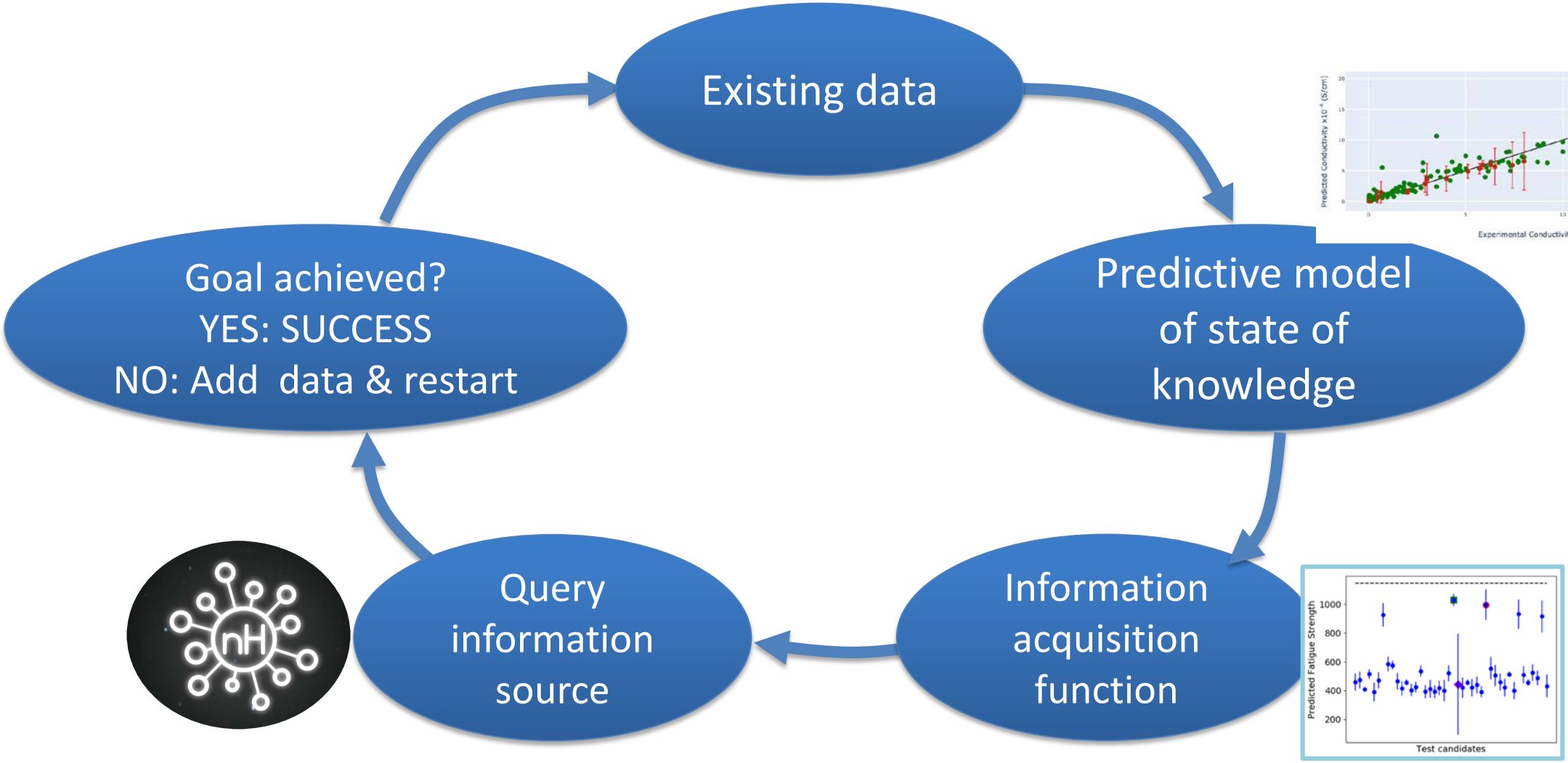
$$x^* = \arg_max\big[E[M(x_i)] + \sigma[M(x_i)]\big]$$

Maximum likelihood of improvement

$$x^* = \arg_max\, \frac{E[M(x_i)] - E[M(x_{best})]}{\sigma[M(x_i)]}$$

# Data science for design of experiments

Can ML help reduce the number of experiments to achieve a desired goal?





Pandita P, Bilionis I, Panchal J. Journal of Mechanical Design. 2016 138, 111412.

Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. Int. Mat. and Manuf. Inn.. 2017 Sep 1;6(3):207-17.

# Maximizing Li+ conductivity in solid oxides

CITRINE INFORMATICS

nanoHUB

## Materials Informatics resources fr

### Authorship and credits

**Based on work published on:** Julia Ling et al.
**Original notebooks from:** *Logan Ward*, Argonne National Laborator
**nanoHUB tools by:** *Juan Carlos Verduzco* and *Alejandro Strachan*,

## Machine Learning Algorithms for Ionic Conductivity of LLZO-type Garnets

### Authorship and credits

**Based on work published on:** Julia Ling et al.
**Original notebooks from:** *Logan Ward*, Argonne National Laboratory, and *Max Hutchinson*, Citrine Informatics. Their Github implementation can be found here.
**nanoHUB tools by:** *Juan Carlos Verduzco* and *Alejandro Strachan*, Materials Engineering, Purdue University
**Database curated by:** *Juan Carlos Verduzco*, Materials Engineering, Purdue University

### Overview

In this notebook, I'll introduce different regression mach

the fa

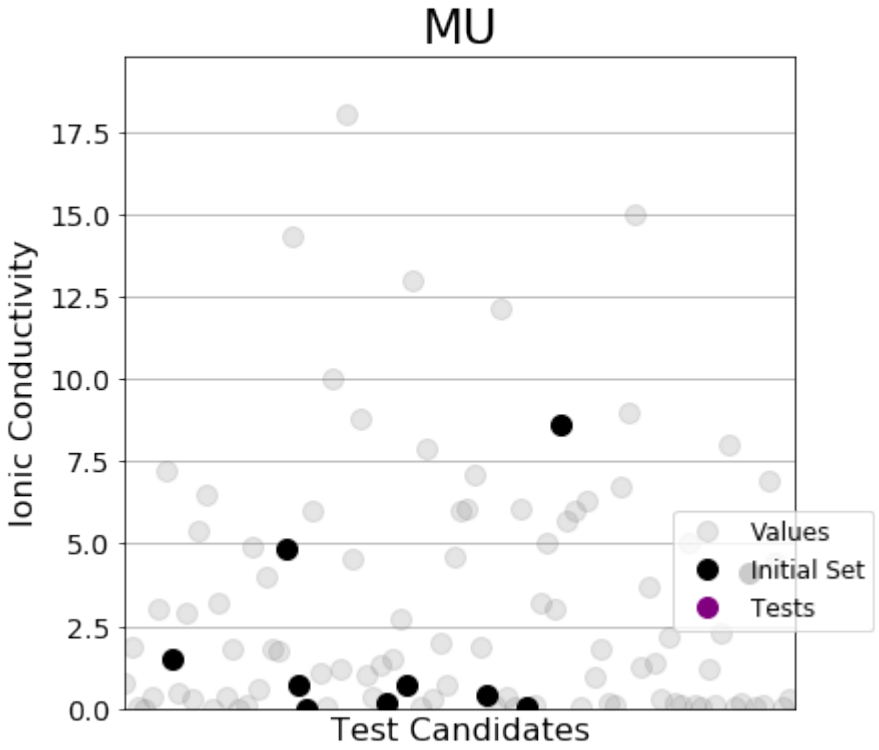
In [3]:
```
cdr = CitrineDataRetrieval('[CITRINE_API_KEY]') # Citrine Key

data = cdr.get_dataframe(criteria={'data_set_id': 184812}, print_properties_options=False) # LLZO Database
property_interest = 'Ionic Conductivity' # Property to be queried

display(data.head(n=10))
```

```
  0%|          | 0/199 [00:00<?, ?it/s]/apps/share64/debian7/anaconda/anaconda-6/lib/python3.7/site-packages/pandas/core/frame.py:
:

Sorting because non-concatenation axis is not aligned. A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.


100%|██████████| 199/199 [00:04<00:00, 46.71it/s]
```

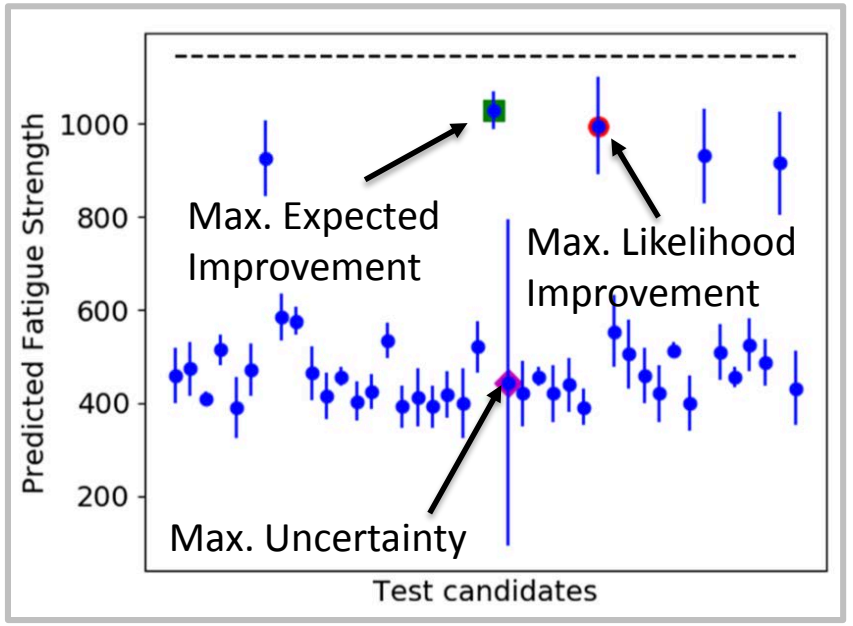
| | chemicalFormula | preparation | references | Crystallographic Structure | Ionic Conductivity | Ionic Conductivity-units |
|---|---|---|---|---|---|---|
| 1 | Li6.25La3Zr1.25Bi0.75O12 | [{'name': 'SINTERING', 'details': [{'name': 'S... | [{'citation': 'http://arxiv.org/abs/1902.06831'}] | Cubic | 2 | 10^-4 S/cm |
| 2 | Li6La2Nb1.5Y0.5O12 | [{'name': 'SINTERING', 'details': [{'name': 'S... | [{'citation': '10.1021/jp304737x', 'doi': '10.... | Cubic | 1.85 | 10^-4 S/cm |
| 3 | Li6.25La3Zr2Ga0.25O12 | [{'name': 'SINTERING', 'details': [{'name': 'S... | [{'citation': '10.1149/2.0171903jes', 'doi': '... | Cubic | 14 | 10^-4 S/cm |
| 4 | Li6La3Ta1.5Y0.5O12 | [{'name': 'SINTERING', 'details': [{'name': 'S... | [{'citation': '10.1039/C4CP00418C', 'doi': '10... | Cubic | 1.26 | 10^-4 S/cm |
| 5 | Li6.55La3Zr2Ga0.15O12 | [{'name': 'SINTERING', 'details': [{'name': 'S... | [{'citation': '10.1021/acsami.6b13902', 'doi':... | Mixed Cubic / Tetragonal | 0.85 | 10^-4 S/cm |
| 6 | Li7.4La3Zr1.76W0.24O12 | [{'name': 'SINTERING', 'details': [{'name': 'S... | [{'citation': '10.1016/j.electacta.2015.08.046... | Cubic | 7.7 | 10^-4 S/cm |
| 7 | Li6.4La3Zr2Ga0.2O12 | [{'name': 'SINTERING', 'details': | [{'citation': | Cubic | 13.2 | 10^-4 S/cm |

https://nanohub.org/tools/citrinetools

# Best conductor with the fewest experiments




- Start with 10 data points (black) out of 100
- Train a ML model and evaluate on possible experiments
- Use acquisition functions to decide what experiment to reveal next
- Add result and iterate

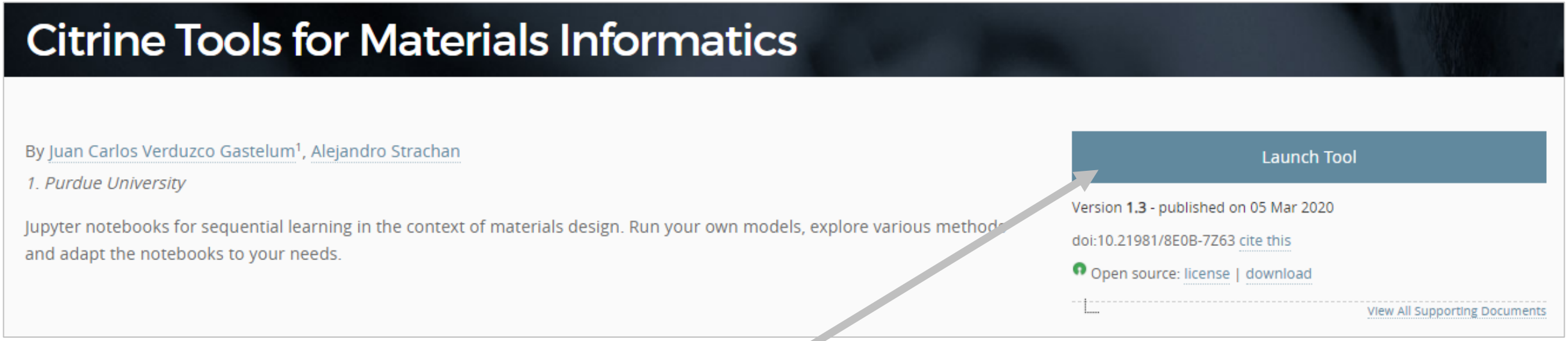



Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. Int. Mat. and Manuf. Inn.. 2017 Sep 1;6(3):207-17.

# Sequential Learning tool in nanoHUB

Citrine Tools for Materials Informatics

From your browser go to link: https://nanohub.org/tools/citrinetools

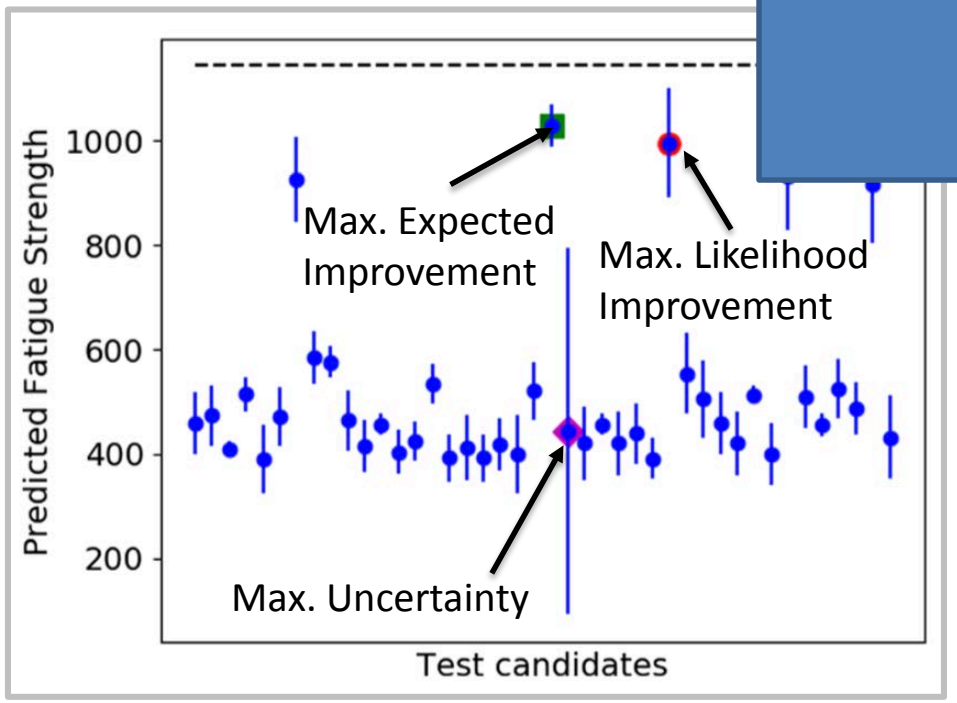



Click on Launch Tool to begin

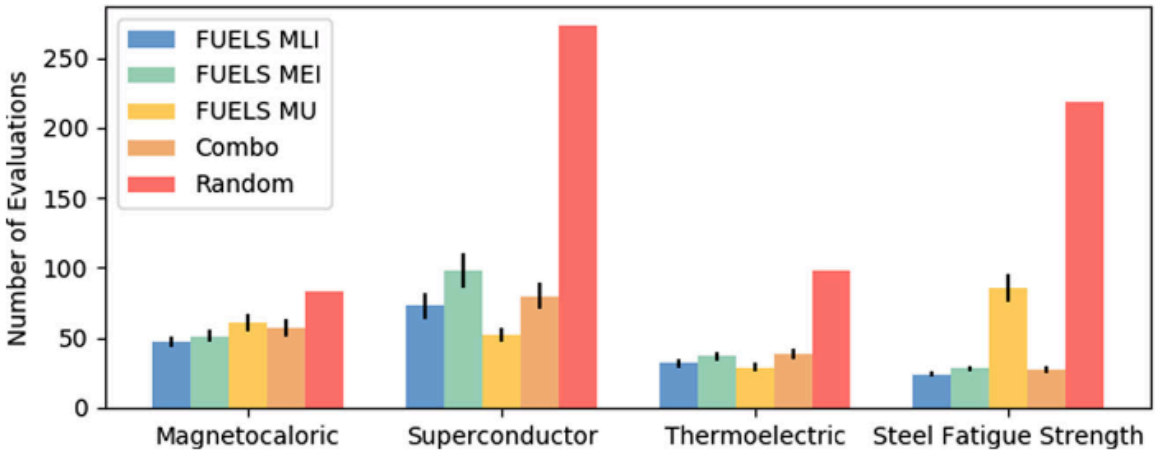
# Finding best conductor with the fewest experiments

- Run the notebook online
- Adapt it to your specific problem

https://nanohub.org/tools/citrinetools

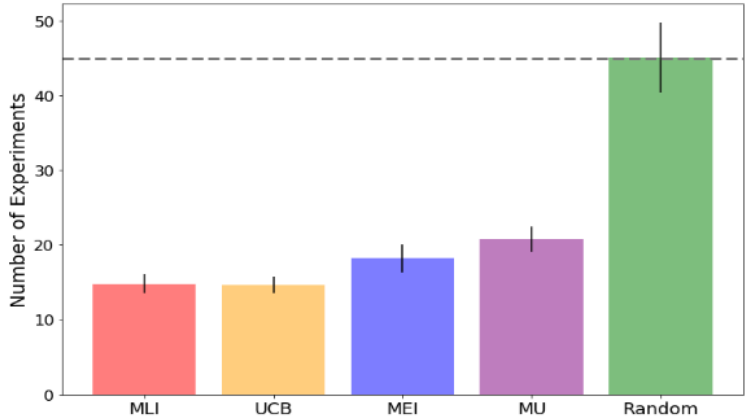


# Results

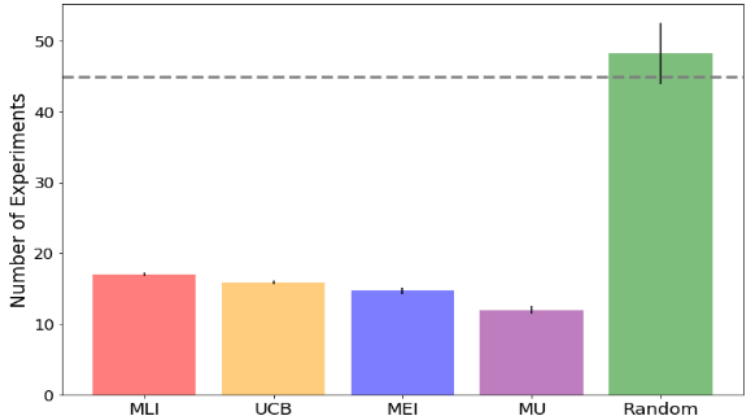



Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. Int. Mat. and Manuf. Inn.. 2017 Sep 1;6(3):207-17.

Random initial set




Chronological selection of initial set




Verduzco, Marinero, Strachan (in preparation)

# Summary

- Predictive models can be used to reduce the number of experiments or simulations in materials discovery or optimization processes
- Knowledge of uncertainties in the model is highly desirable for decision making
- The code in https://nanohub.org/tools/citrinetools can be easily modified to meet needs in a wide range of problems

# Next steps

**Hands-on tutorial**: walks students over all the steps required to run the https://nanohub.org/tools/citrinetools example

**Homework assignment**: to reinforce concepts and help students modify the workflow and adapt it for their needs