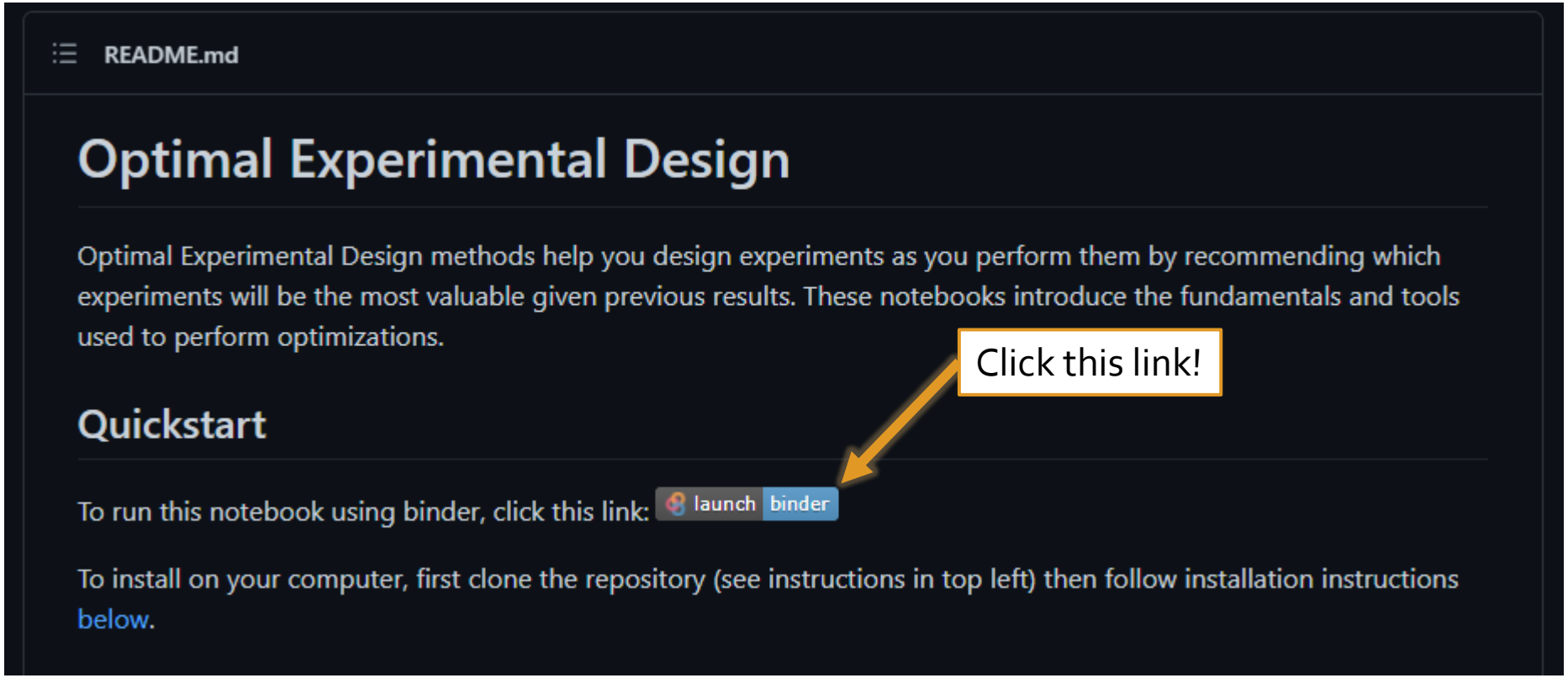
OPTIMAL EXPERIMENTAL DESIGN

Logan Ward
Asst. Computational Scientist
Data Science and Learning Division

Noah Paulson
Computational Materials Science
Applied Materials Division

Step 1: Launch Binder

We are using Binder for the practical exercises, which takes a few minutes to start.

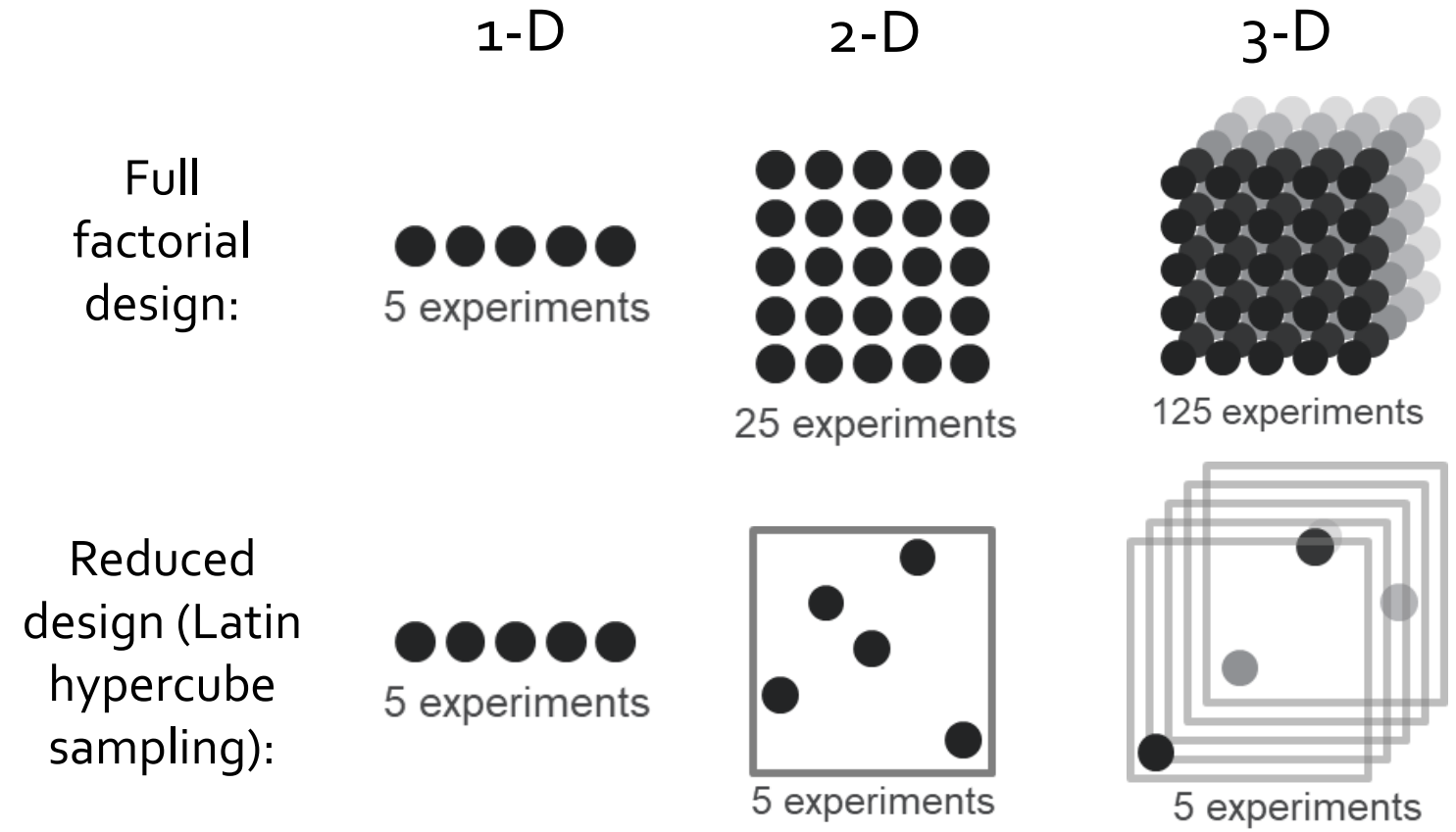


Challenges in Experimental Design

Design of Experiments: Traditional approaches are data hungry

Real campaigns of experiments or simulations have constraints! For example:

- Limited time: e.g. beamline access is 24 hrs, 1 hr per experiment
- Limited resources: e.g. 1M CPU hrs on HPC, 1000CPU hrs per simulation
- Limited budget: $10k allocated for materials, $1k per sample



Which designs are optimal for which scenarios?

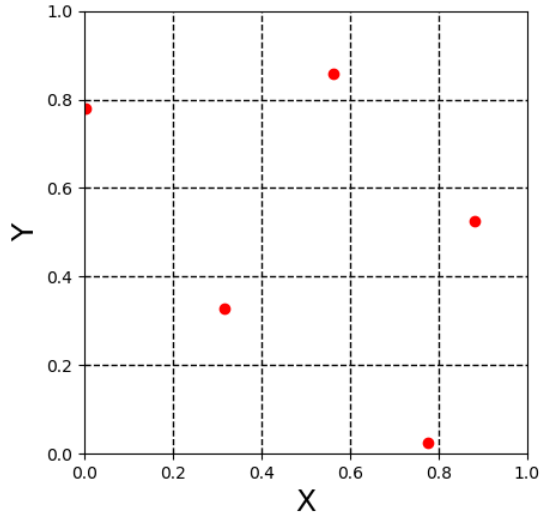
“Static” Experimental Design

Design of Experiments: How to choose experiments under a finite budget

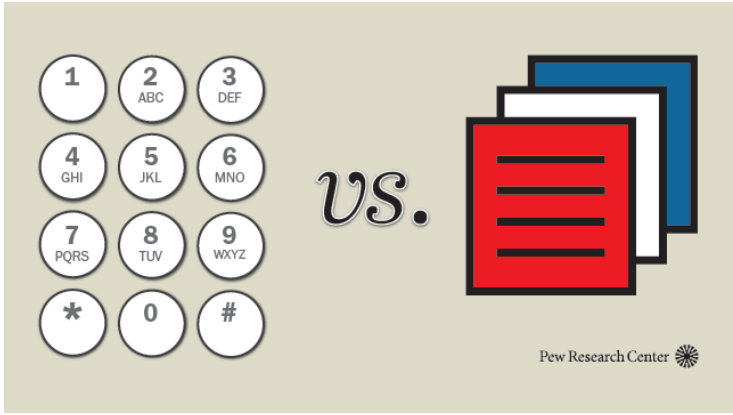
Treatment combinations for a 2^{5-2} design

Treatment combination	I	A	B	C	D = AB	E = AC
de	+	−	−	−	+	+
a	+	+	−	−	−	−
be	+	−	+	−	−	+
abd	+	+	+	−	+	−
cd	+	−	−	+	+	−
ace	+	+	−	+	−	+
bc	+	−	+	+	−	−
abcde	+	+	+	+	+	+

Source: Wikipedia



Source: ICME@MSE



Source: Pew Research

What if you can learn between experiments?

Key concept: "Active Learning"

Optimal Design: Select new experiments as you learn more

An idea that takes many forms and names…

- *Active learning*
- *Bayesian optimization*
- *Optimal experimental design*
- *Sequential learning*
- *Surrogate-based optimization*

Components of "optimal design":

- Initial design space and training data
- Machine learning model with uncertainty
- Policy for sampling

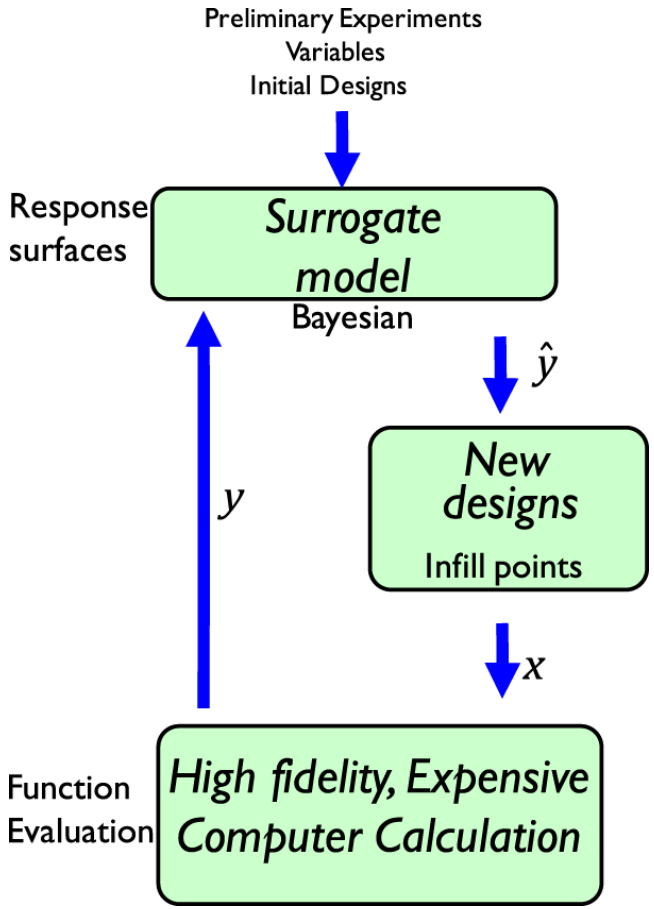


Figure: Lookman et al. npj Comp. Mat. (2019)

BUILDING MODELS WITH UNCERTAINTY ESTIMATES

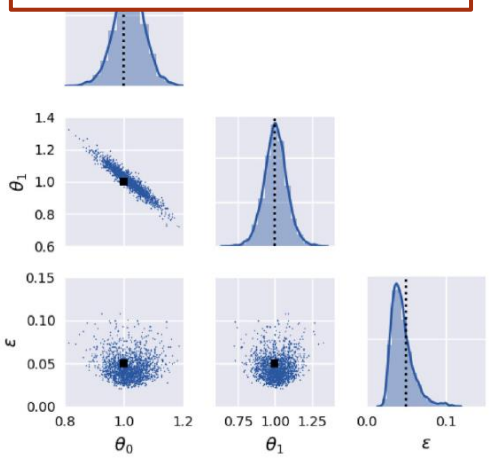
Not as hard as you might think

Two Key Ways for "ML with Uncertainty"

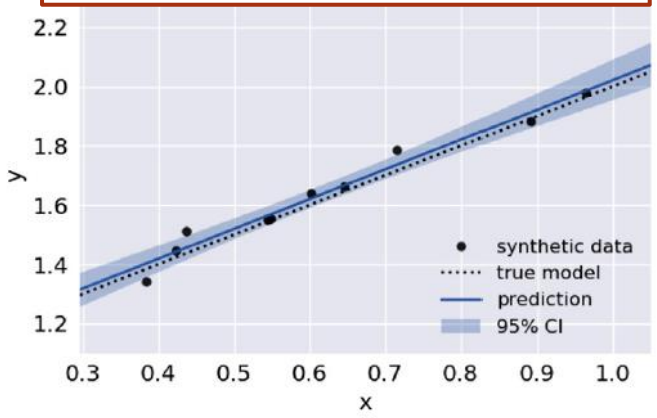
Bayesian Machine Learning

Concept: Estimate distribution of *parameters*

Range of Parameters



Predictions with Uncertainty



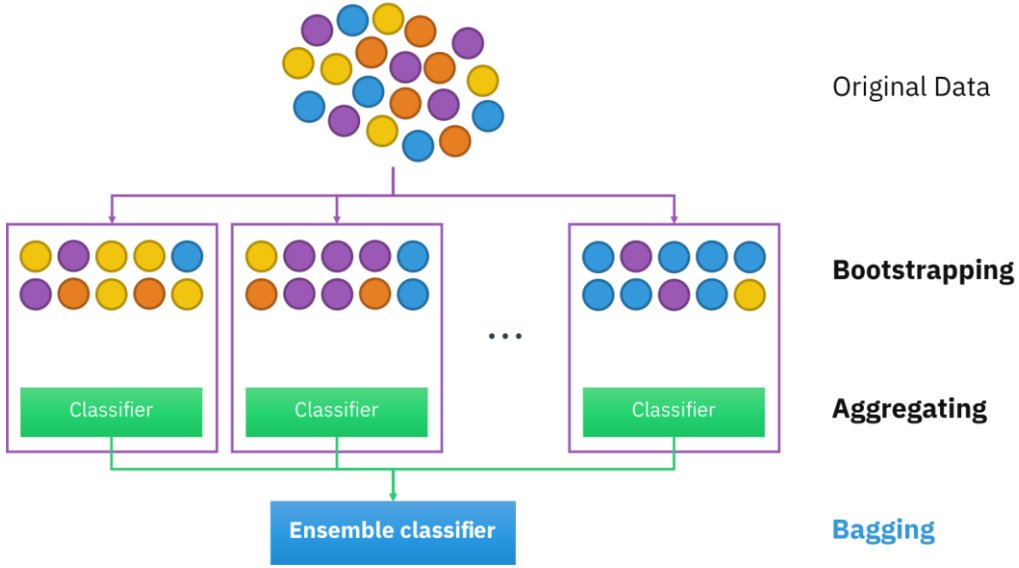
Advantage: Robust statistical basis

Disadvantage: Restricted model forms

Key Method: Gaussian Process Regression

Bootstrapped Ensembles

Concept: Create distribution of models



Advantage: Can use any model form

Disadvantage: High computational cost

Key Method: Random forest

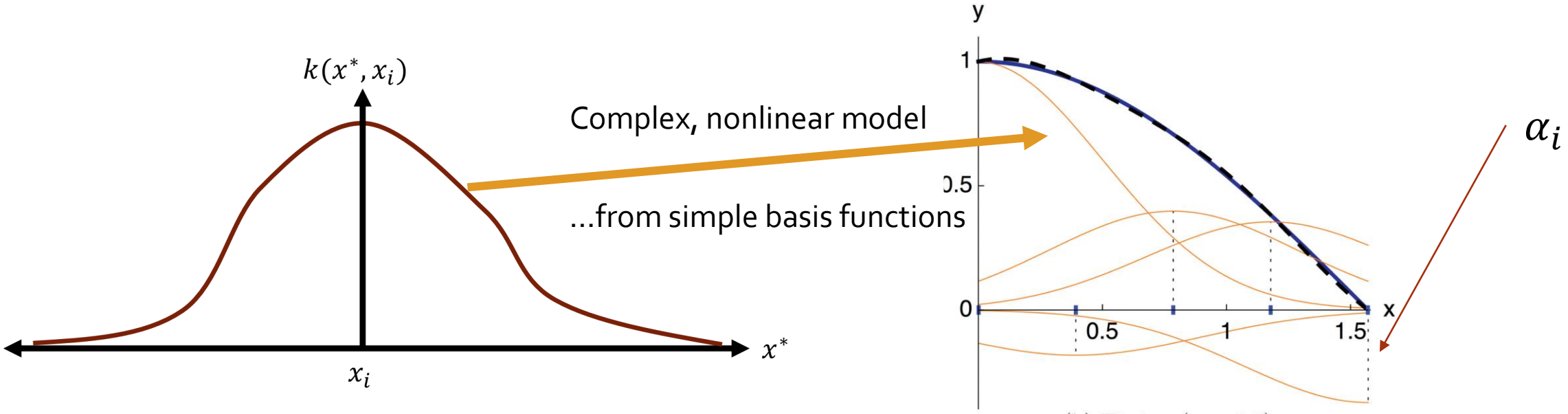
Understanding Gaussian Process Regression

Bayesian Learning with a "kernel trick"

(Simplified) Model Form: $f(x^*) = \sum_i \alpha_i \boldsymbol{k}(x^*, x_i)$

Some complex math gives an expression for $\sigma(x^*)$

Kernels (**k**) express the shape of your model, for example a "radial basis function"

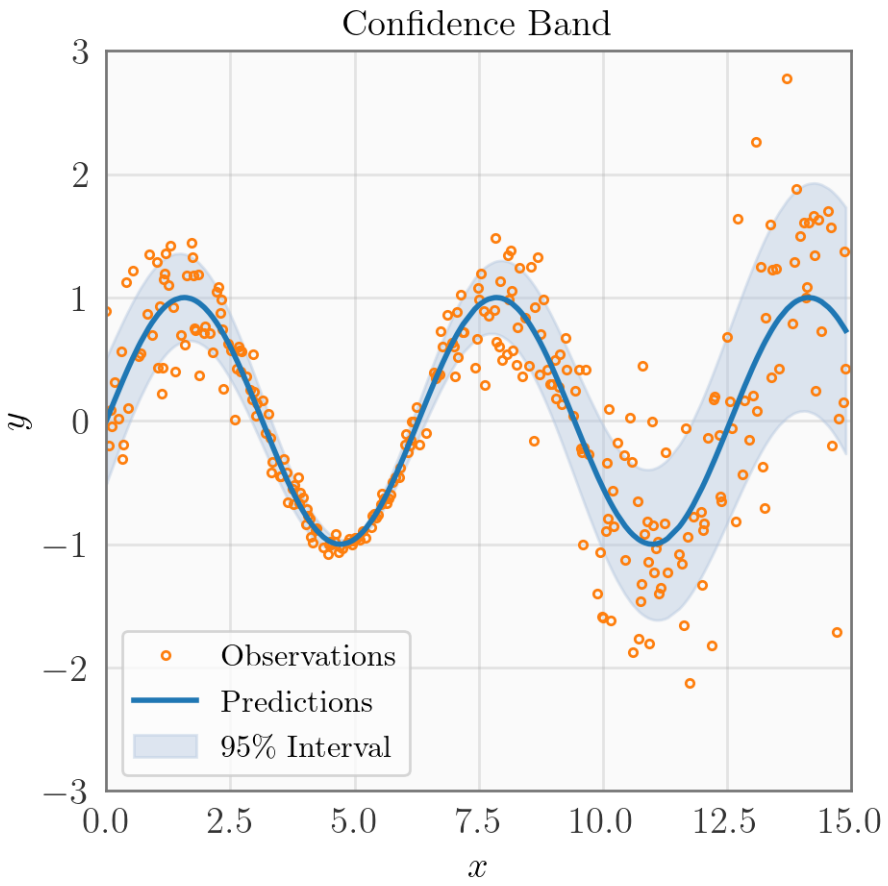


First Half of the Example Notebook

- Show to train a GPR model

github.com/AIScienceTutorial/intro-to-bayesian-optimization/explore-acquistion-functions.ipynb

A quick note: Uncertainty Intervals Are Not Perfect



Key points:

- Your uncertainties are still *estimates*
- They do not work "out of distribution"
- Not every "uncertainty" can be interpreted the same

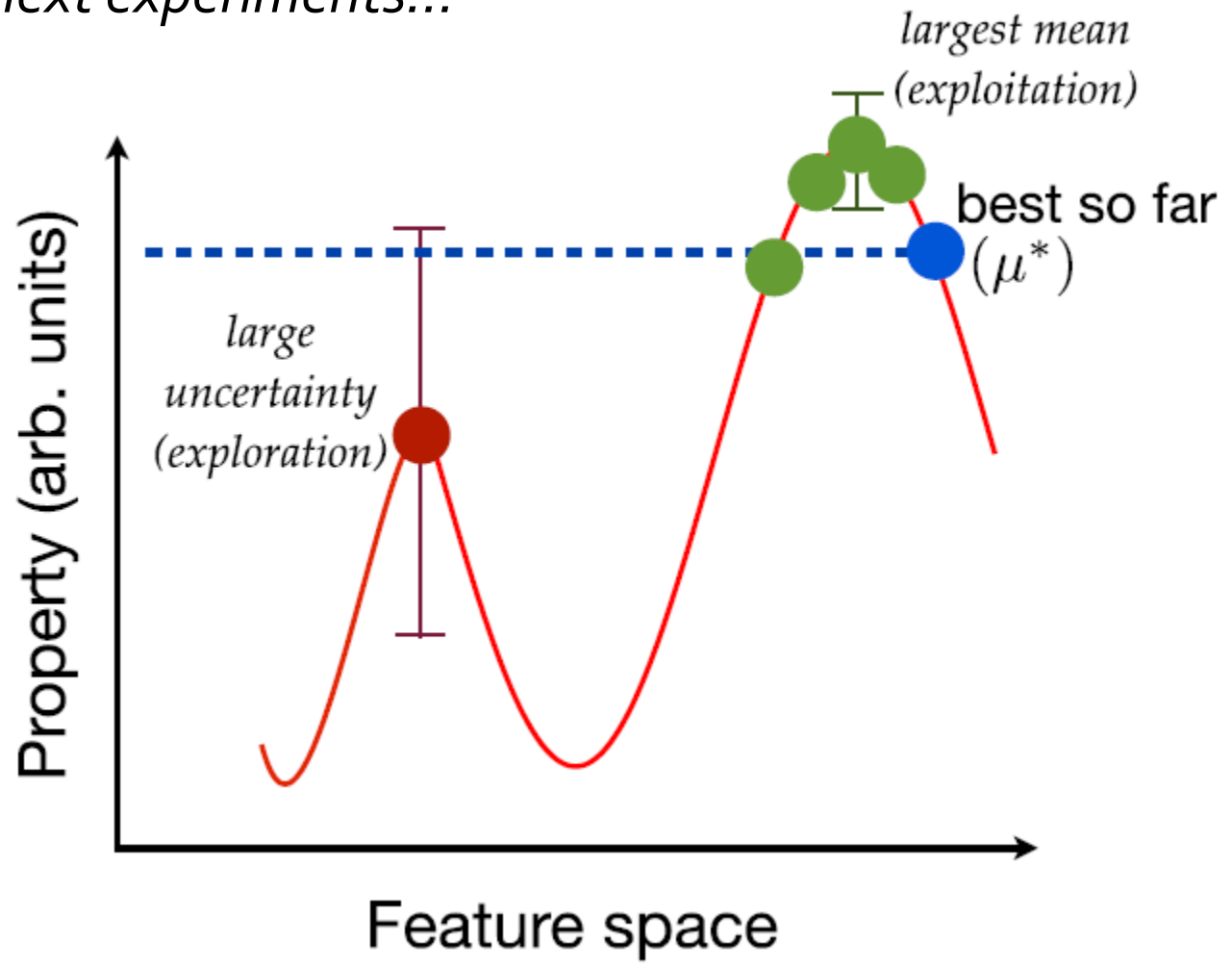
… but they can be good enough to guide experiments

https://github.com/uncertainty-toolbox/uncertainty-toolbox

SELECTING A SAMPLING POLICY

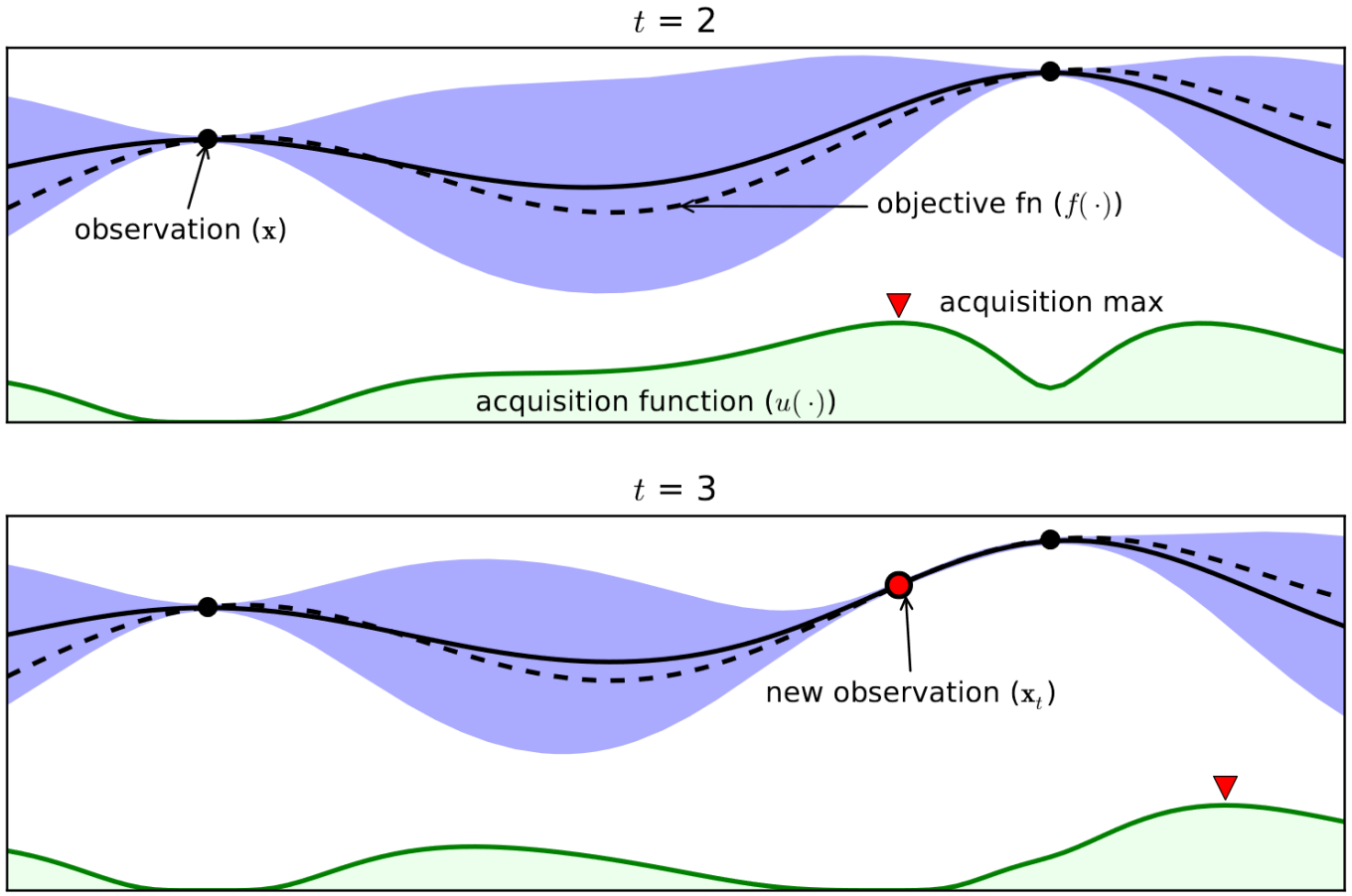
Sampling Policies: Exploration vs Exploitation

Many ways to pick next experiments…



Source: Balachandran et al.. Sci. Rep. (2016)

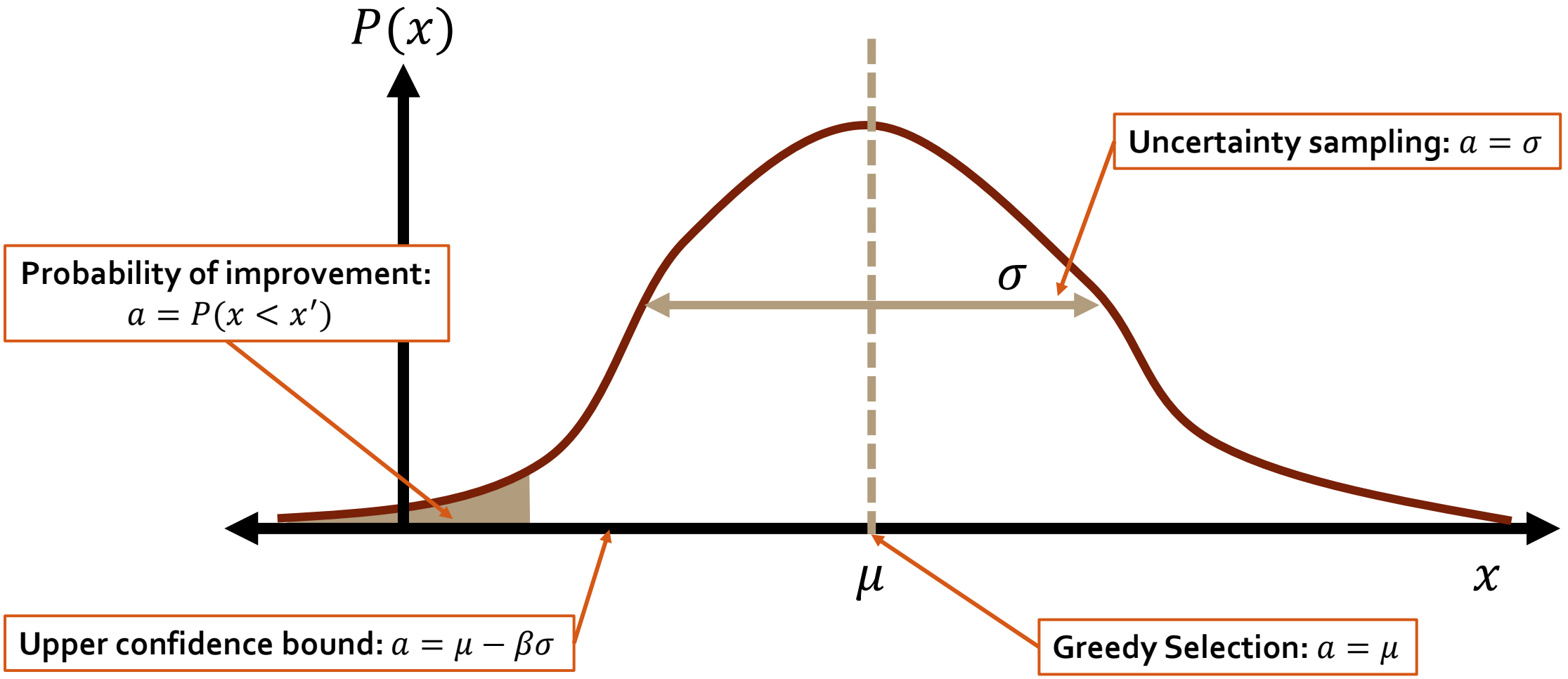
Bayesian Optimization: Quantifying value judgements



Source: Towards Data Science

Simple Acquisition Functions

Further variety in ways to capture experimental objectives



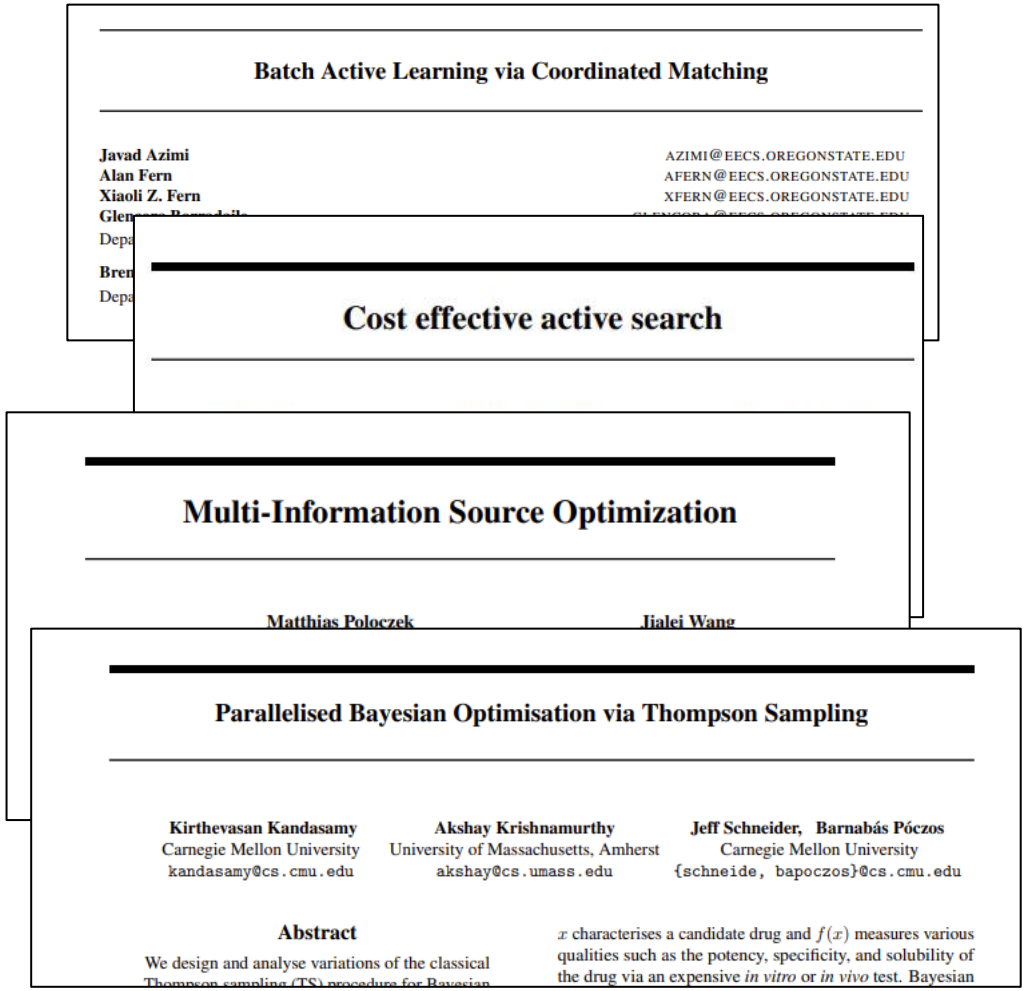
Nice reference: Roman Garnett's Course Materials

It can get *very* complicated…

Many different complicating factors (or opportunities to be clever!):

- Performing experiments in parallel vs sequential?
- Different properties of learning algorithms?
- More than one objective?
- Different ways to access your experiments?
- Experiments are different costs?
- Do experiments take the same amount of time?
- Is retraining your model expensive?
- ….!

My view: Make a friend in applied mathematics!



Second Half of the Notebook

- Let's go back to our example notebook

Challenges with adapting active learning to science

General issue: Many experiments are hard to express as a list of coordinates

My design space is not Cartesian!

If they were, I could just…

- call `np.meshgrid` to generate samples
- use `scipy.optimize` to locate extrema

Since it isn't, you could…

- enumerate a set of designs beforehand
- create a genetic algorithm to optimize

I cannot train a machine learning model!

If they were, I could just…

- use `sklearn` out of the box

Since it isn't, you could…

- compute features that are good coordinates*
- develop a neural network architecture

*See Arun's tutorial from last year

EXAMPLES FROM MATERIALS SCIENCE

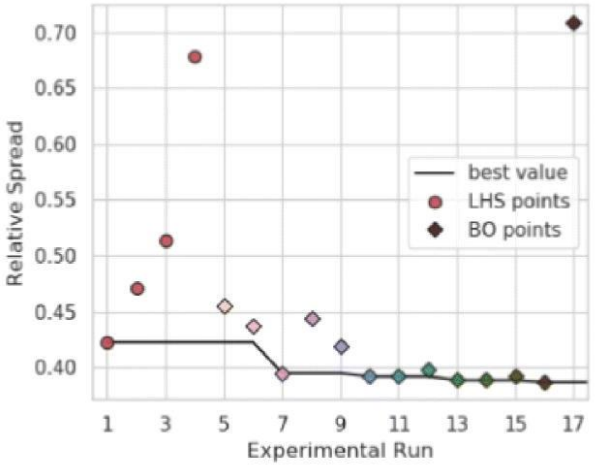
Because Noah and I are materials engineers,
not because active learning is only good for materials science

Faster optimization of industrial processes

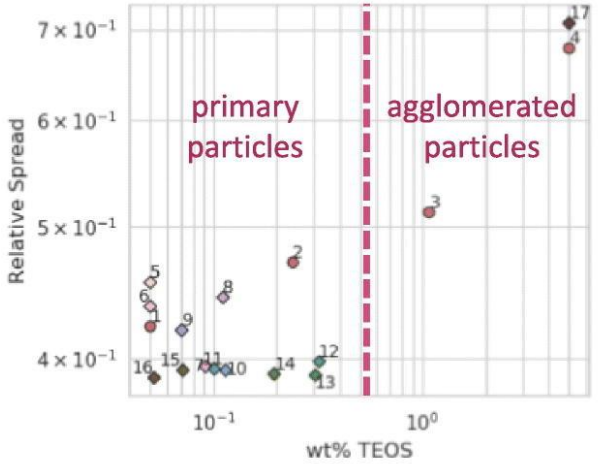
flame spray pyrolysis processing space

	lower	upper
TEOS concentration (wt%)	0.05	5
liquid flow rate (mL/min)	4	10
atomization O_2 flow rate (L/min)	6	12
pilot CH_4 flow rate (L/min)	2	4
pilot O_2 flow rate (L/min)	3	6
sheath O_2 flow rate (L/min)	15	25

design of experiments
1) Latin hypercube sampling
2) Bayesian Optimization



optimized particle morphology



physics understanding

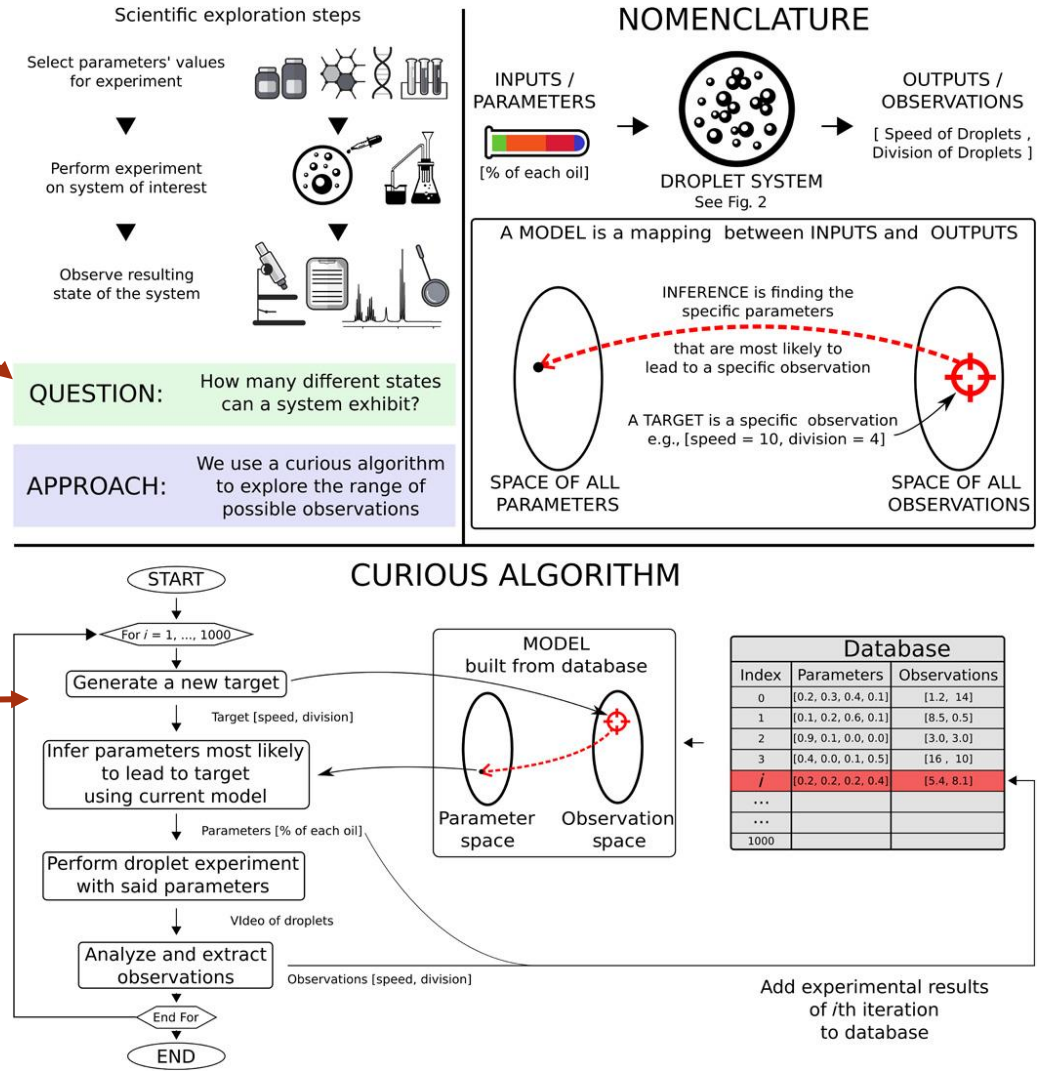
1. **Model:** Gaussian Process with RBF Kernel
2. **Search Space:** 6-D process parameters
3. **Policy:** Expected Improvement

Ref: Paulson et al. Materials & Design. (2020)

Curiosity Driven Active Learning

The goal of your experimental design can be "to discover"

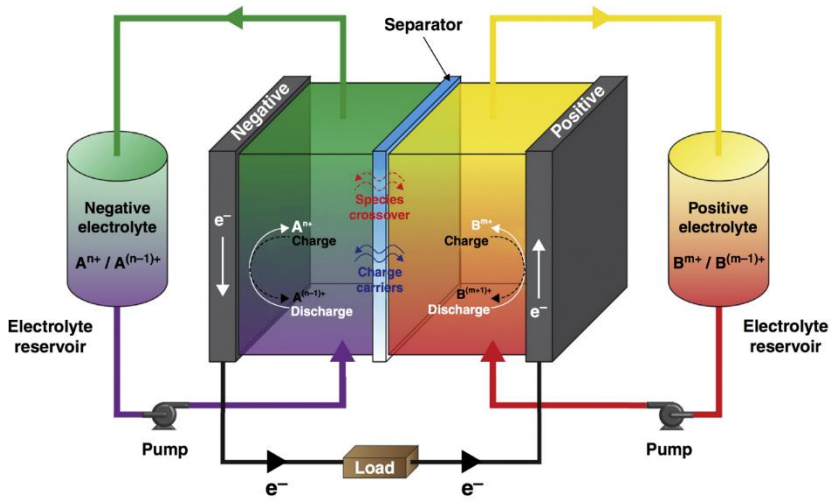
It is a matter of defining a "curious algorithm"



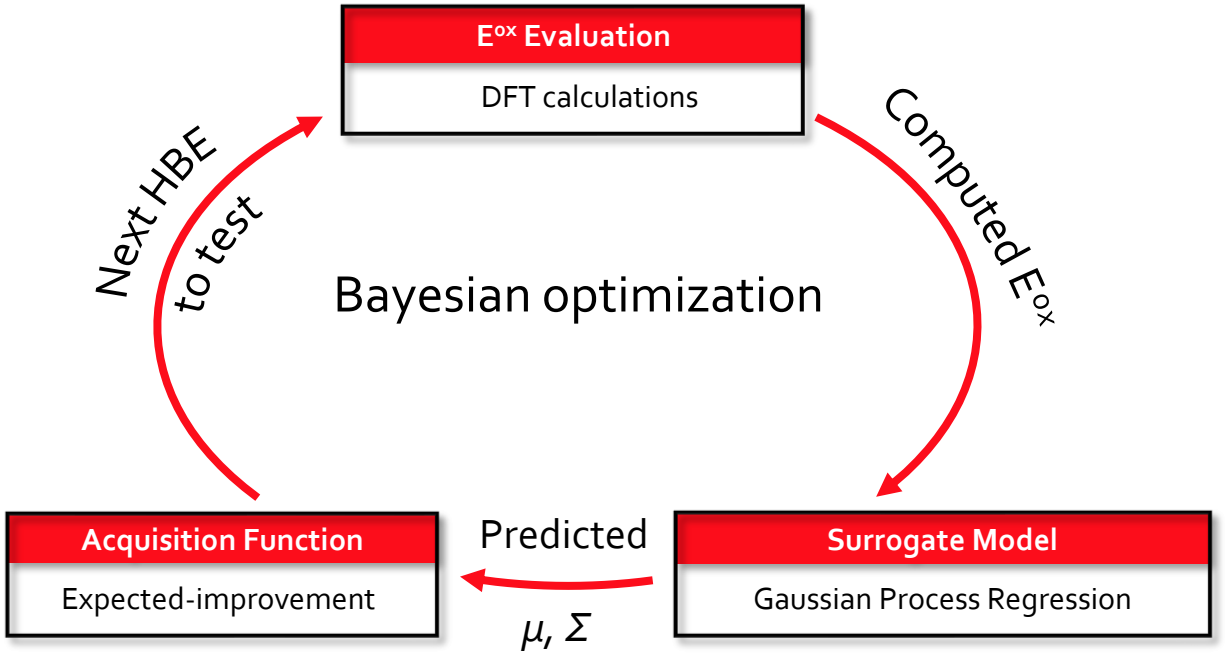
Ref: Grizou *et al.* Science Advances (2020)

Stay Tuned for Next Week!

Case study: Active Learning via Bayesian Optimization for Discovery of Energy Storage Materials



Kowalski, Su, Milshtein, Brushett (2016) *Current Opinion in Chemical Engineering*



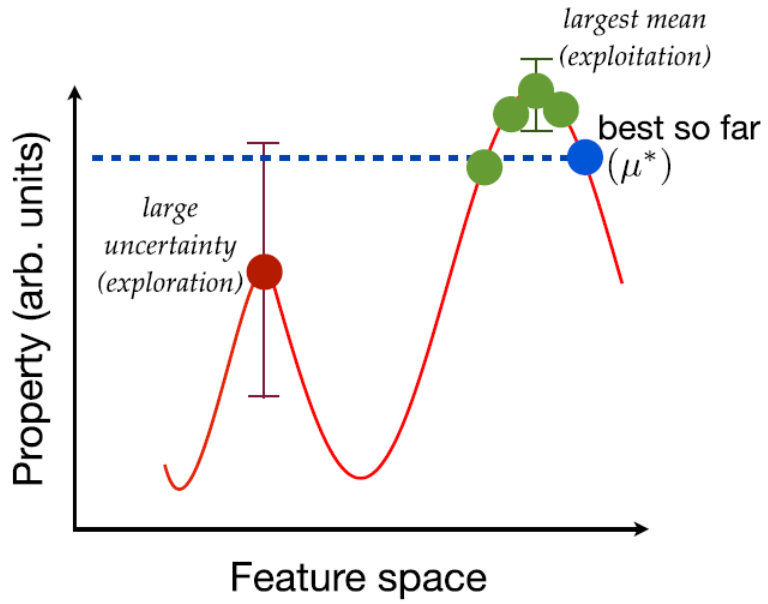
Source: [Doan et al. Chem. Mater. 2020, 32, 15, 6338–6346]

Take-Away Points

"Optimal design" = "Learning while doing"

Main challenge is to find a good way to pick the next experiments

Ex: "explore" vs "exploit"

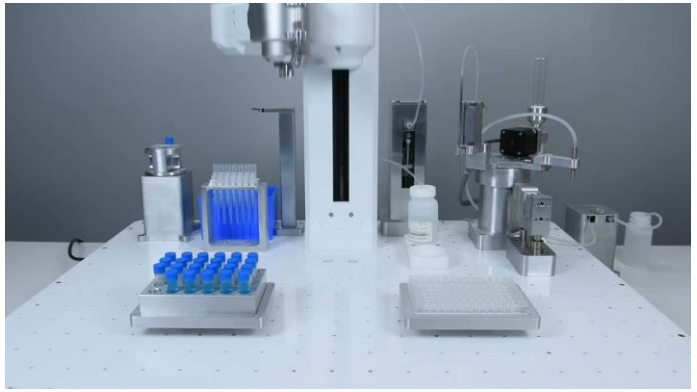


Source: Balachandran et al.. Sci. Rep. (2016)

Many ways to use active learning!

- Material design
- Model fitting
- Guiding characterization
- Solving structures
- Just for curiosity

Next step ->



Source: Curtis Berlinguette (UBC)