OPTIMAL EXPERIMENTAL DESIGN

Logan Ward
Asst. Computational Scientist
Argonne National Laboratory

9 February 2022

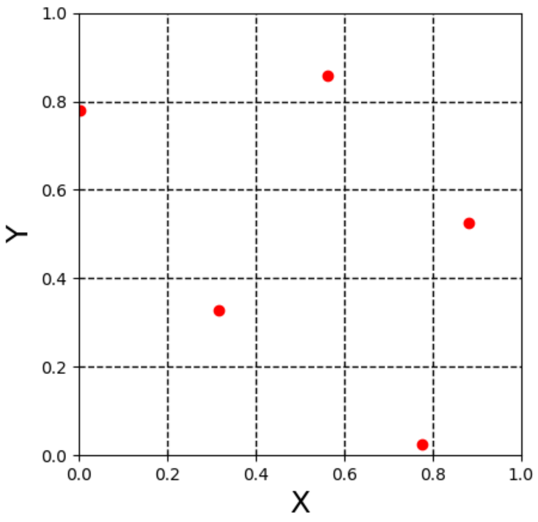
“Static” Experimental Design

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

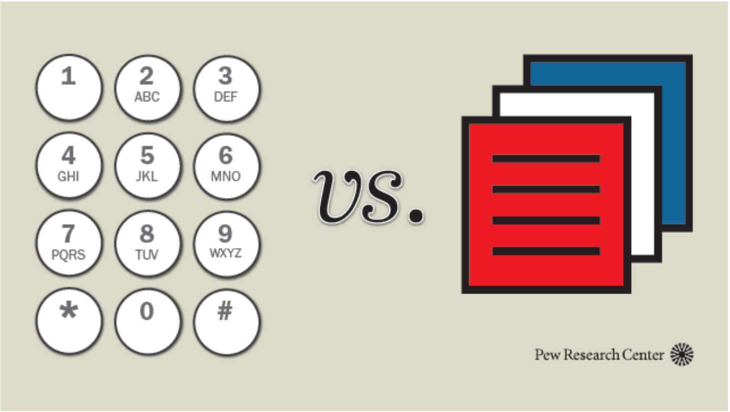
Treatment combinations for a 2^{5-2} design

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":
- Machine learning model <u>with uncertainty</u>
- Space of possible experiments
- **Policy for sampling**

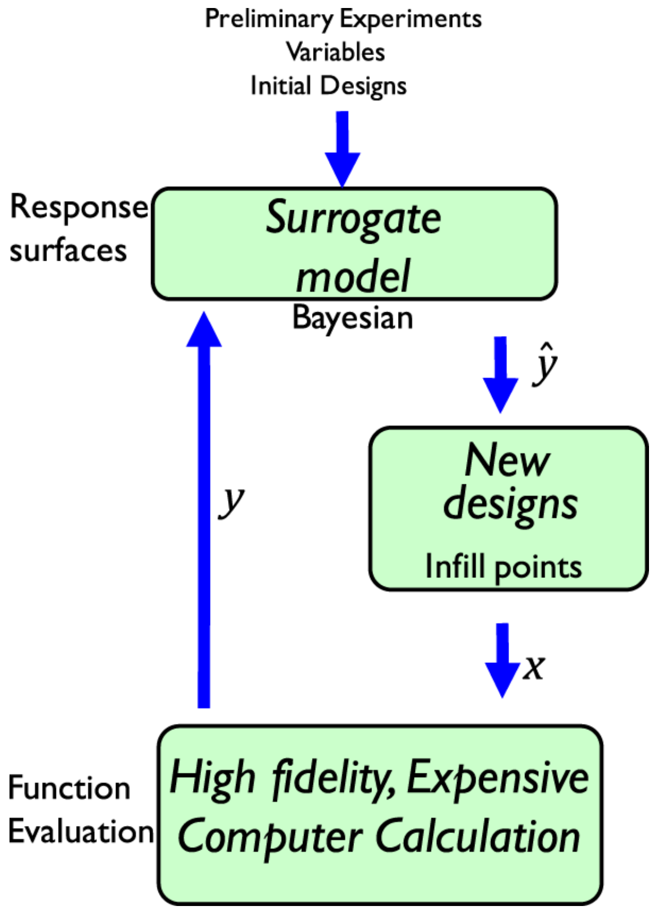


Figure: Lookman et al. npj Comp. Mat. (2019)

BUILDING MODELS WITH UNCERTAINTY ESTIMATES

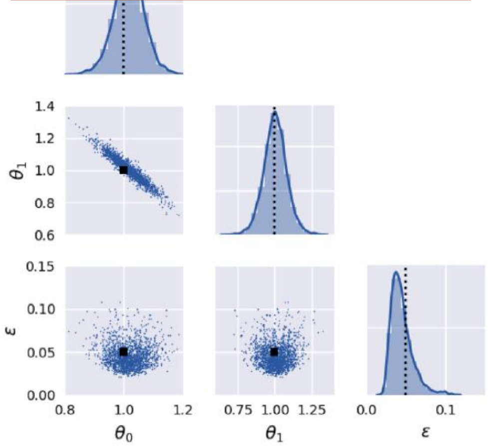
Not as hard as you might think

Two Key Ways for "ML with Uncertainty"

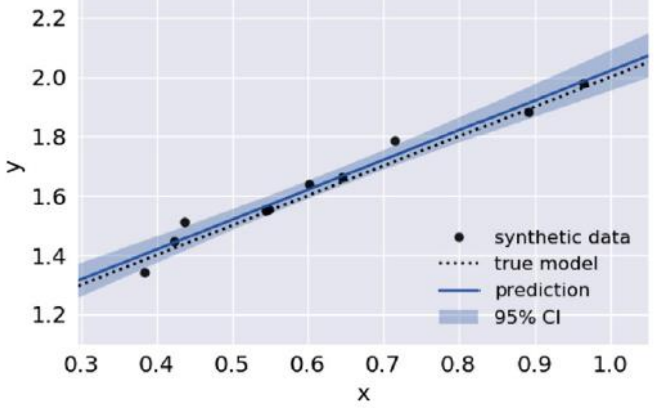
Bayesian Machine Learning

Concept: Estimate distribution of *parameters*

Range of Parameters



Predictions with Uncertainty



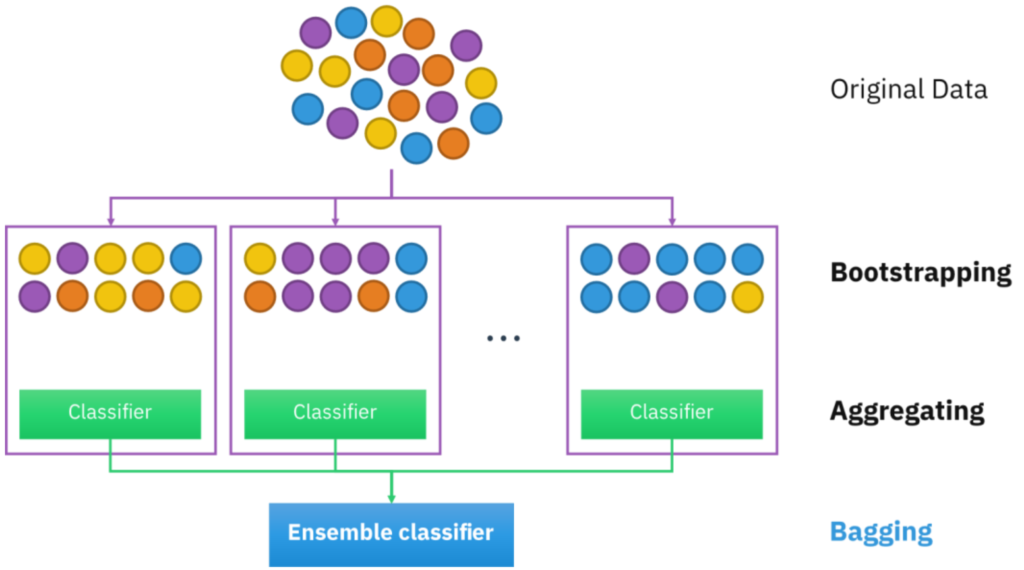
Advantage: Robust statistical basis

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

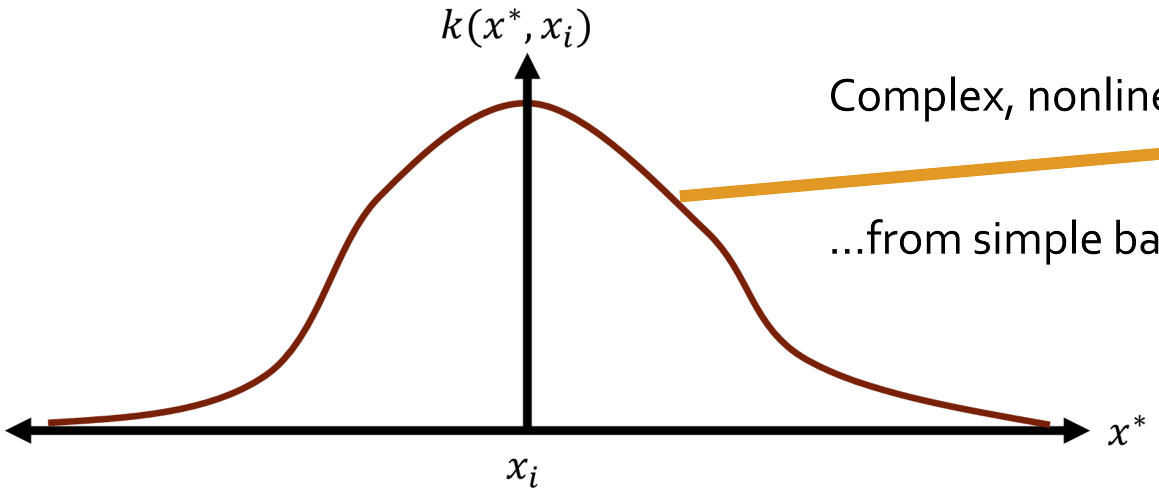
Understanding Gaussian Process Regression

Bayesian Learning with a "kernel trick"

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

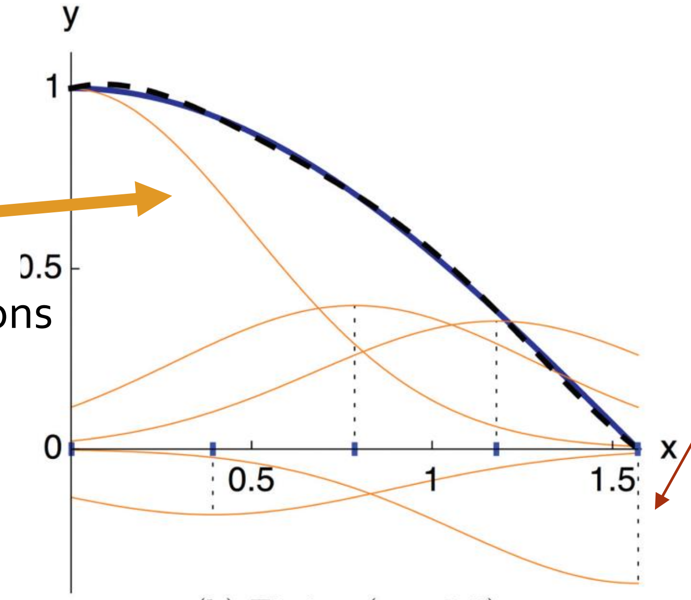
Some complex math gives an expression for $\sigma(x^*)$

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



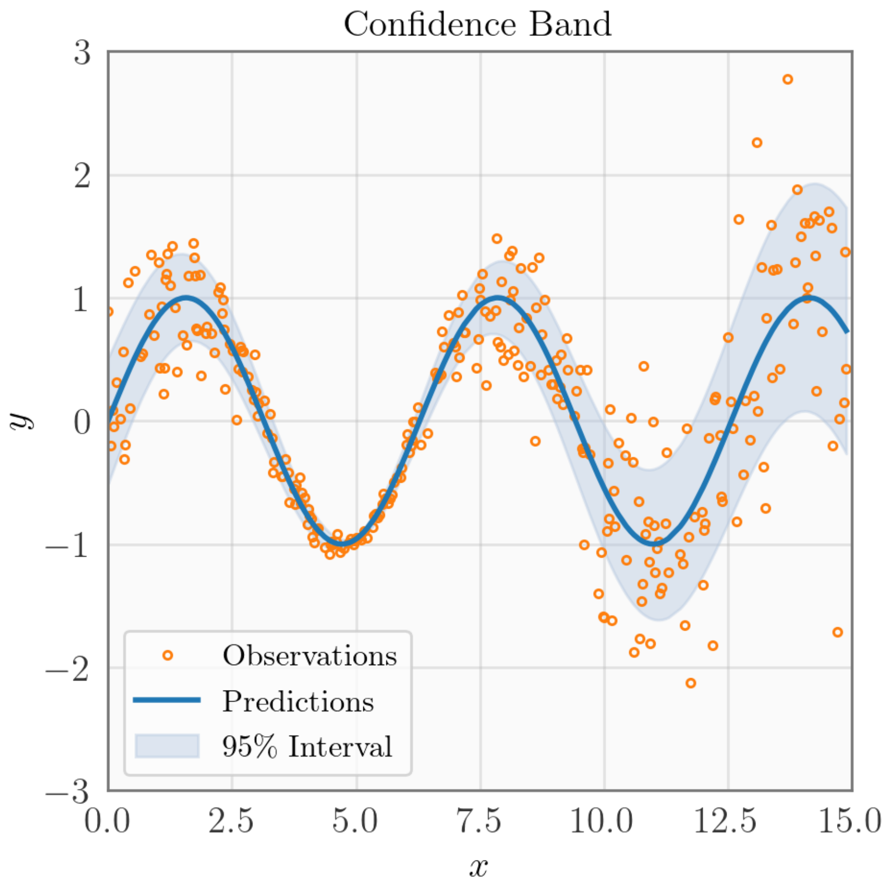
Complex, nonlinear model

...from simple basis functions



α_i

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

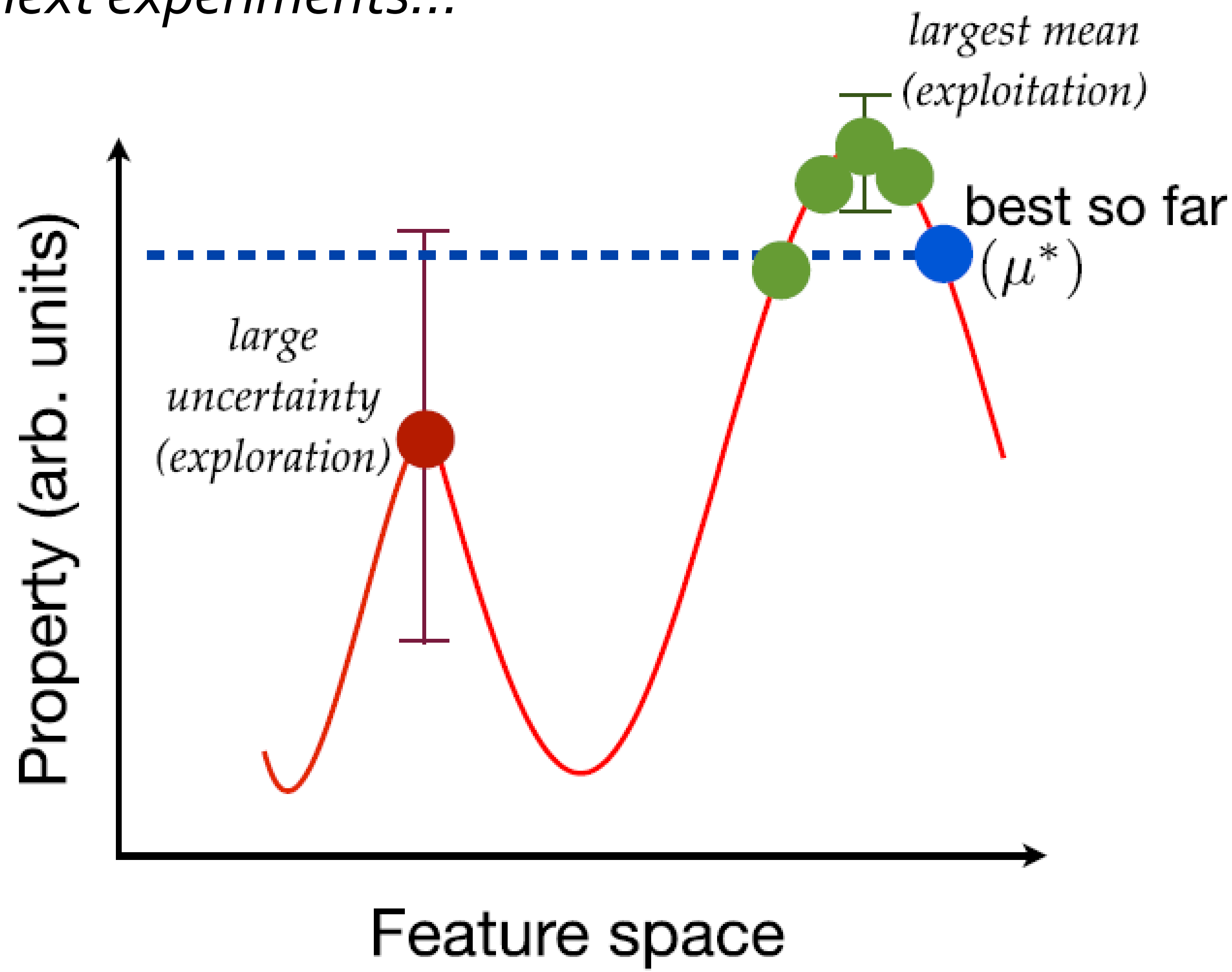
… but they can be good enough to guide experiments

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

SELECTING A SAMPLING POLICY

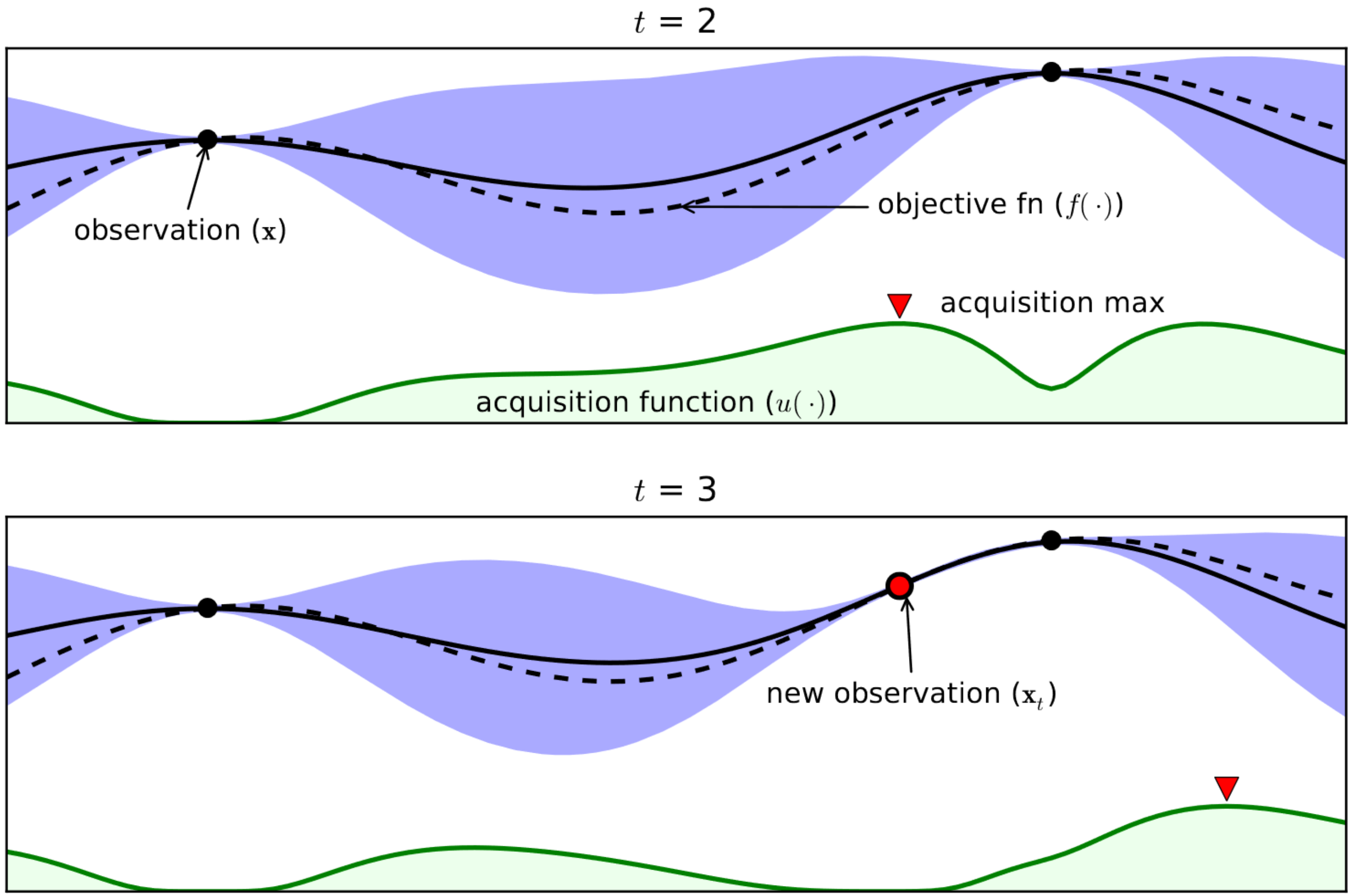
Sampling Policies: Exploration vs Exploitation

Many ways to pick next experiments…



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

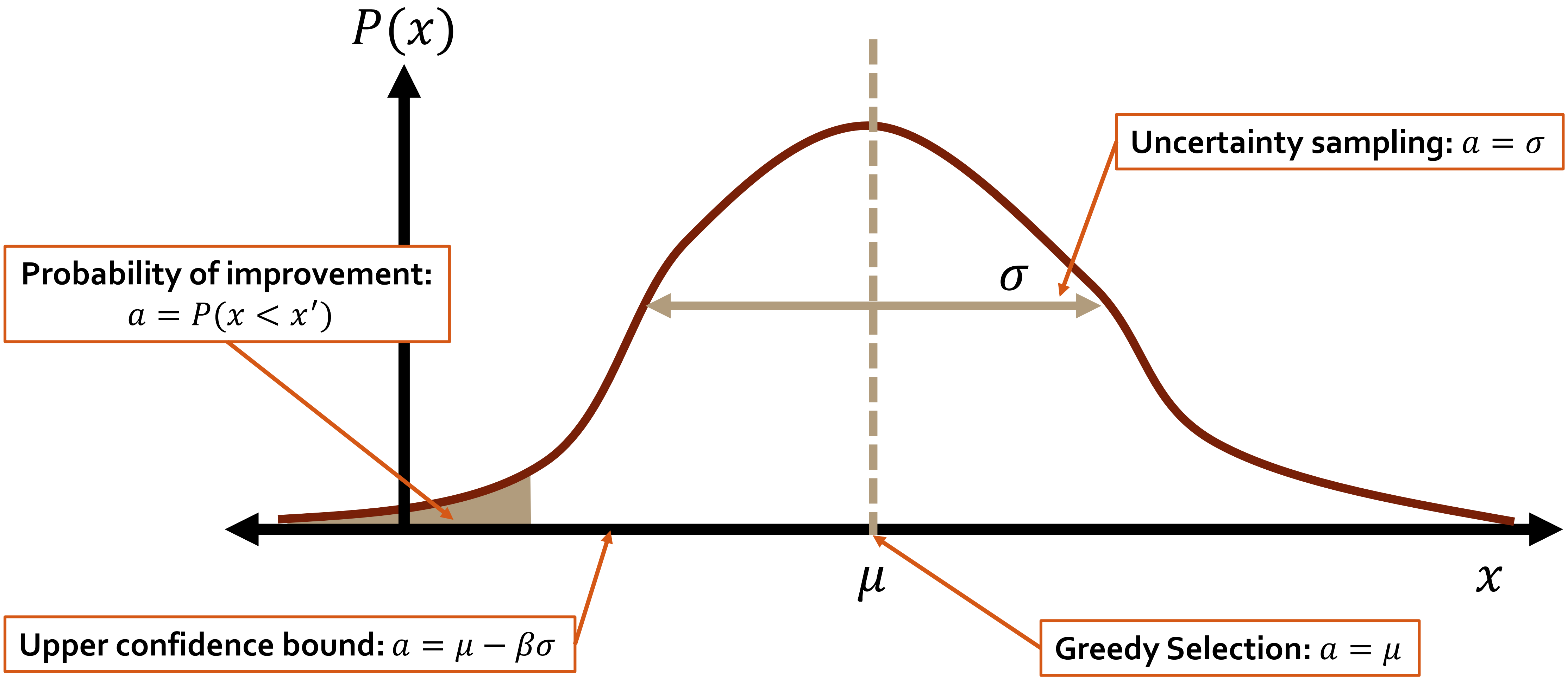
Bayesian Optimization: Quantifying value judgements



Source: [Towards Data Science]

Simple Acquisition Functions

Further variety in ways to "explore" or "exploit"



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!

Batch Active Learning via Coordinated Matching

Javad Azimi AZIMI@EECS.OREGONSTATE.EDU
Alan Fern AFERN@EECS.OREGONSTATE.EDU
Xiaoli Z. Fern XFERN@EECS.OREGONSTATE.EDU

Cost effective active search

Multi-Information Source Optimization

Matthias Poloczek Jialei Wang

Parallelised Bayesian Optimisation via Thompson Sampling

Kirthevasan Kandasamy, Carnegie Mellon University, kandasamy@cs.cmu.edu
Akshay Krishnamurthy, University of Massachusetts, Amherst, akshay@cs.umass.edu
Jeff Schneider, Barnabás Póczos, Carnegie Mellon University, {schneide, bapoczos}@cs.cmu.edu

Abstract

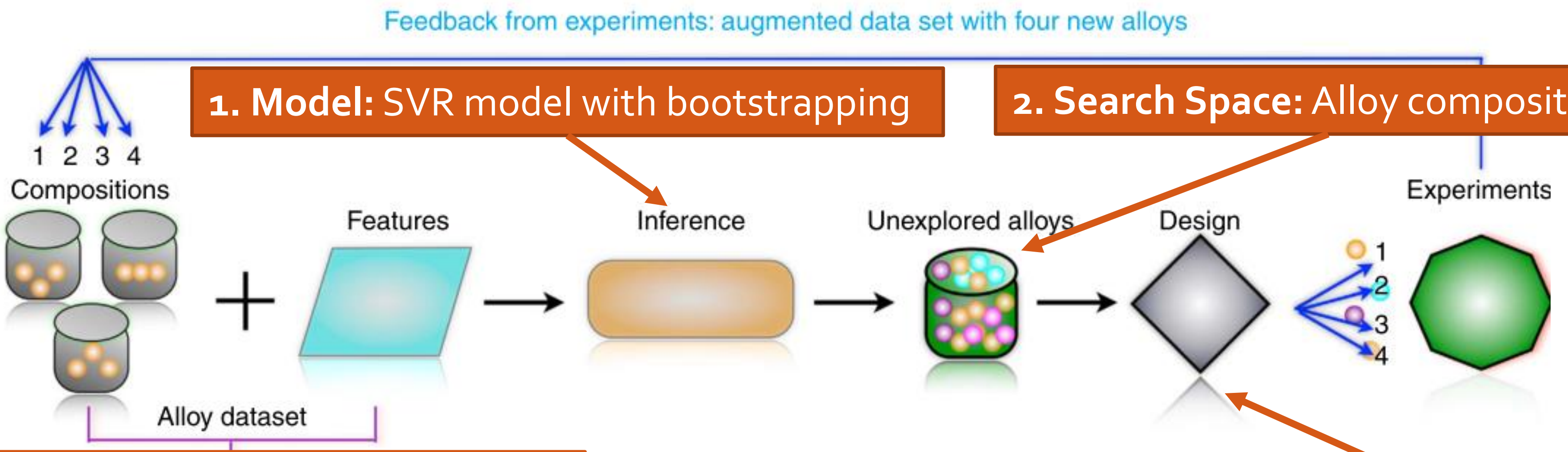
We design and analyse variations of the classical

x characterises a candidate drug and $f(x)$ measures various qualities such as the potency, specificity, and solubility of the drug via an expensive *in vitro* or *in vivo* test. Bayesian

EXAMPLES FROM MATERIALS SCIENCE

A relatively new idea, but catching on quickly

Example: Shape memory alloys with small ΔT

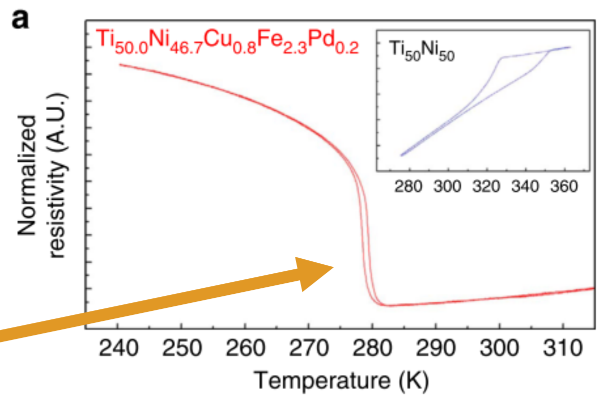


1. Model: SVR model with bootstrapping

2. Search Space: Alloy compositions

3. Policy: Knowledge Gradient

Of [the 36 compositions we tested], 14 had smaller ΔT than any of the 22 in the original data set.



Best alloy was 42% lower!

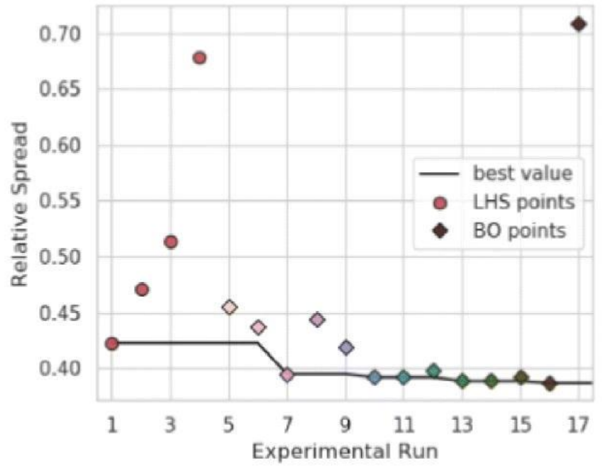
Ref: Xue *et al.* Nat Comm. (2016), 11241.

Faster optimization of industrial processes

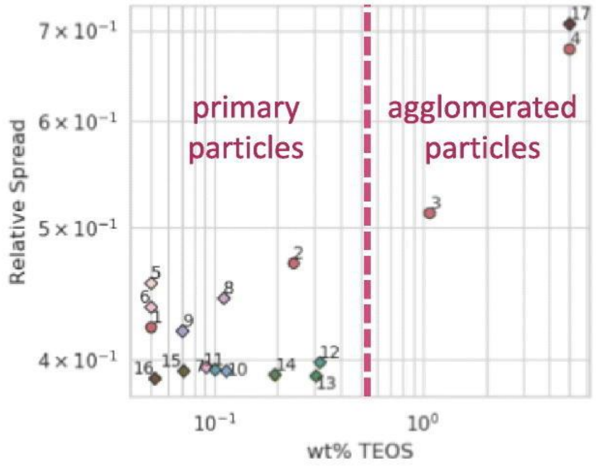
flame spray pyrolysis processing space

	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

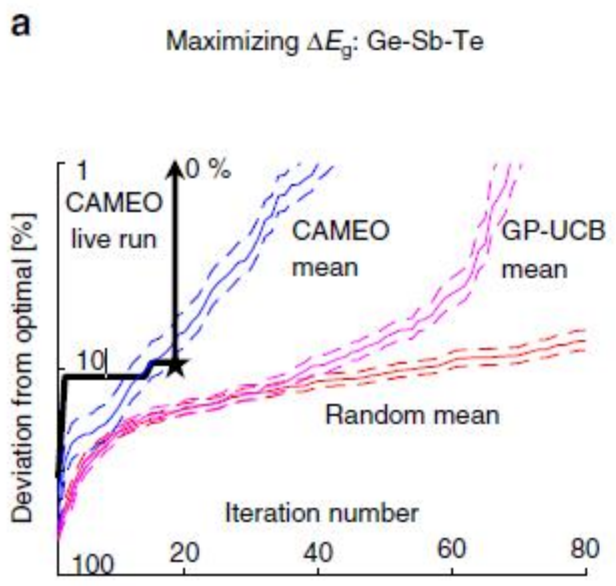


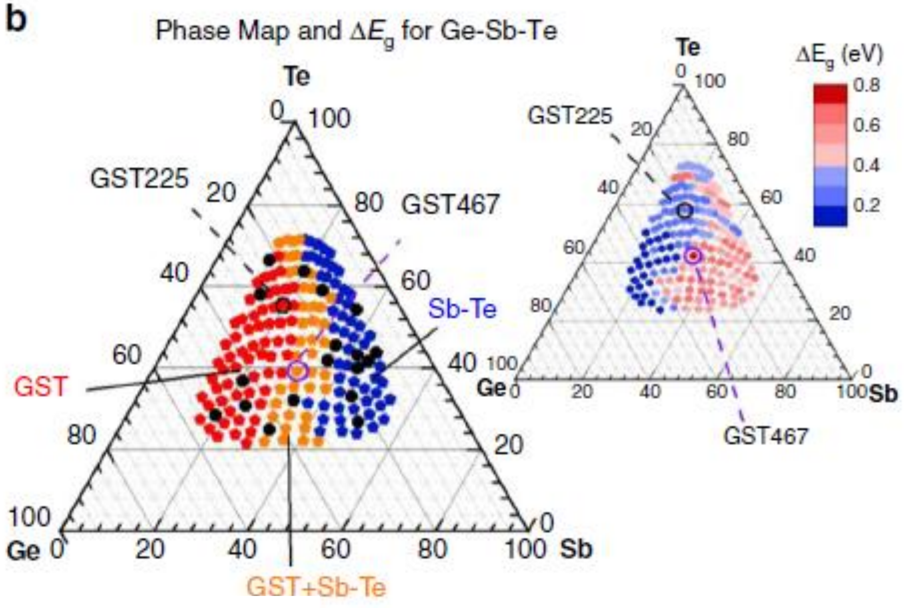
physics understanding

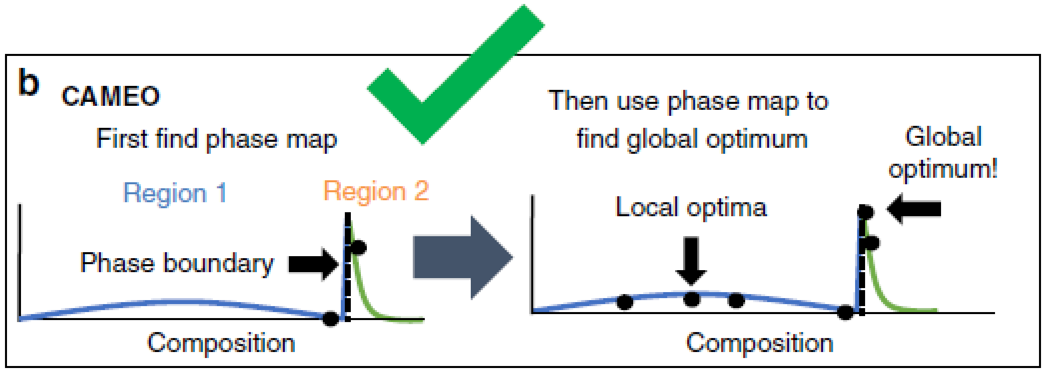
1. **Model:** Gaussian Process with RBF Kernel
2. **Search Space:** 6-D process parameters
3. **Policy:** Expected Improvement

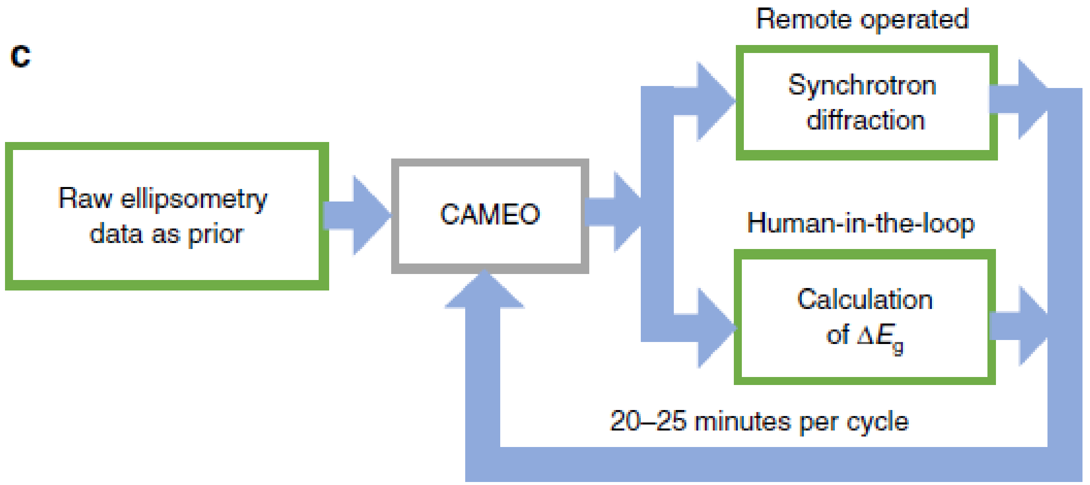
Ref: Paulson et al. Materials & Design. (2020)

Characterization with Fewer Measurements





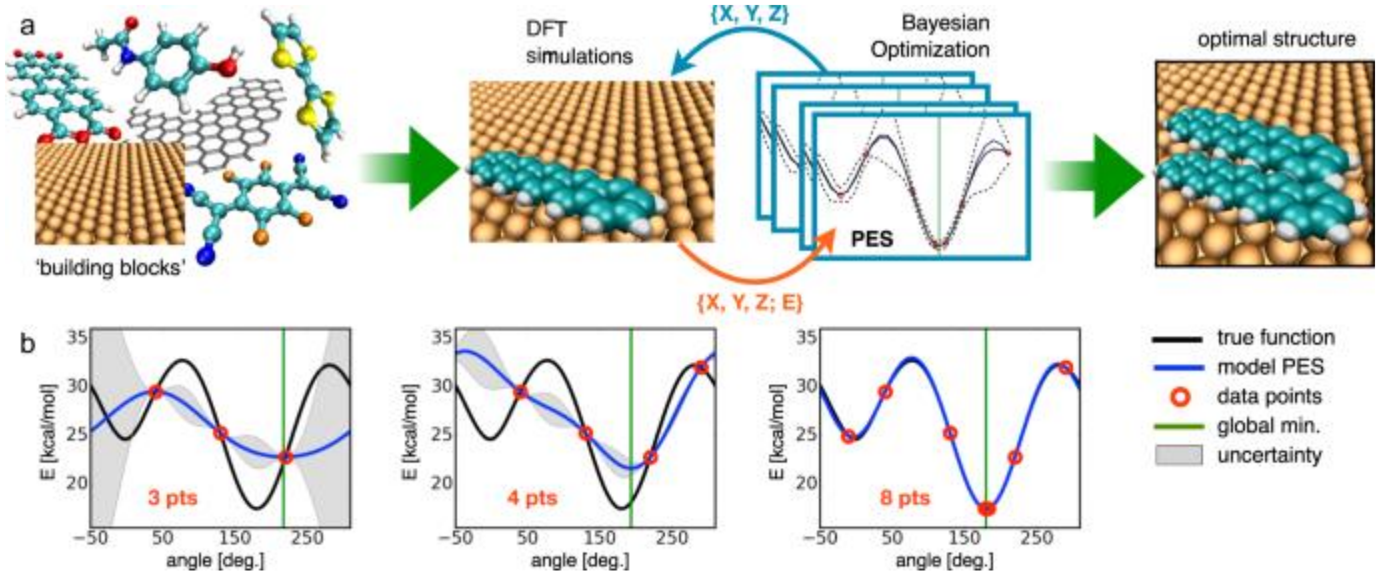




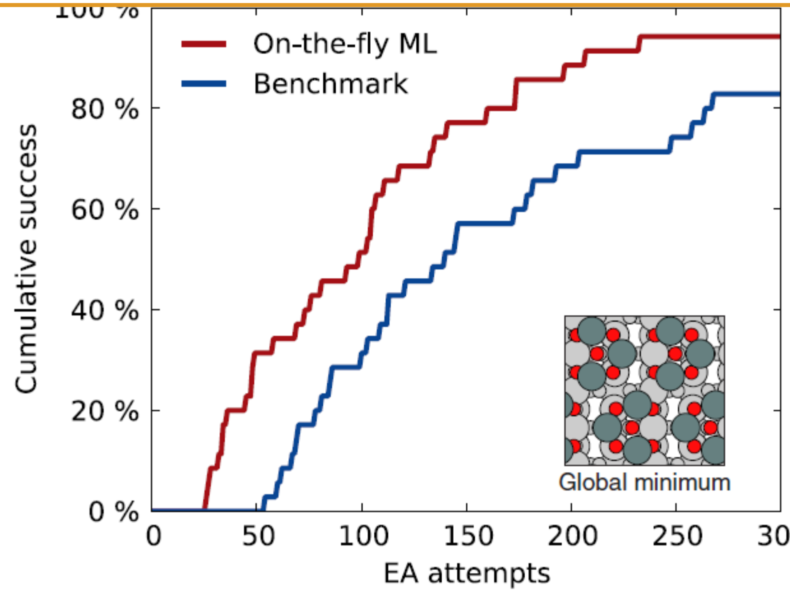
Ref: Kusne *et al.* Nat. Comm. (2020)

Structure Optimization via Bayesian Optimization

BOSS: Todorović et al. npj Comp Mat (2019)

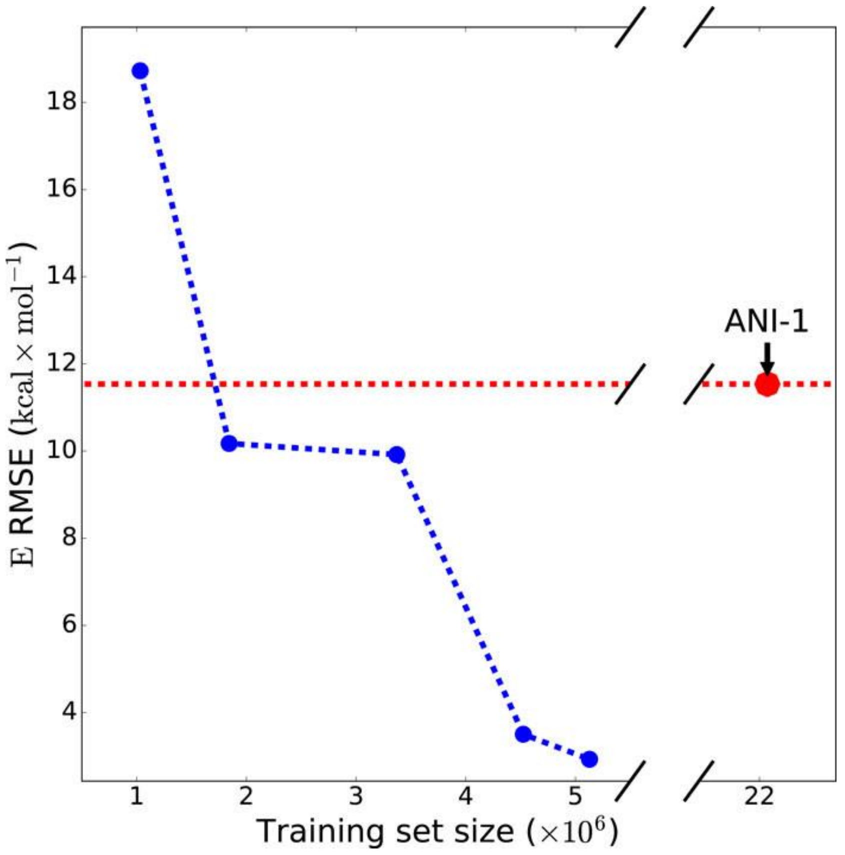


Surface Reconstruction: Jacobsen. PRL (2018)



Fitting Better Models: Fitting Interatomic Potentials

"Better model with 10% of the data"
- J. Smith et al. JCP (2018)



Cool innovation: Accounting for clustering within data

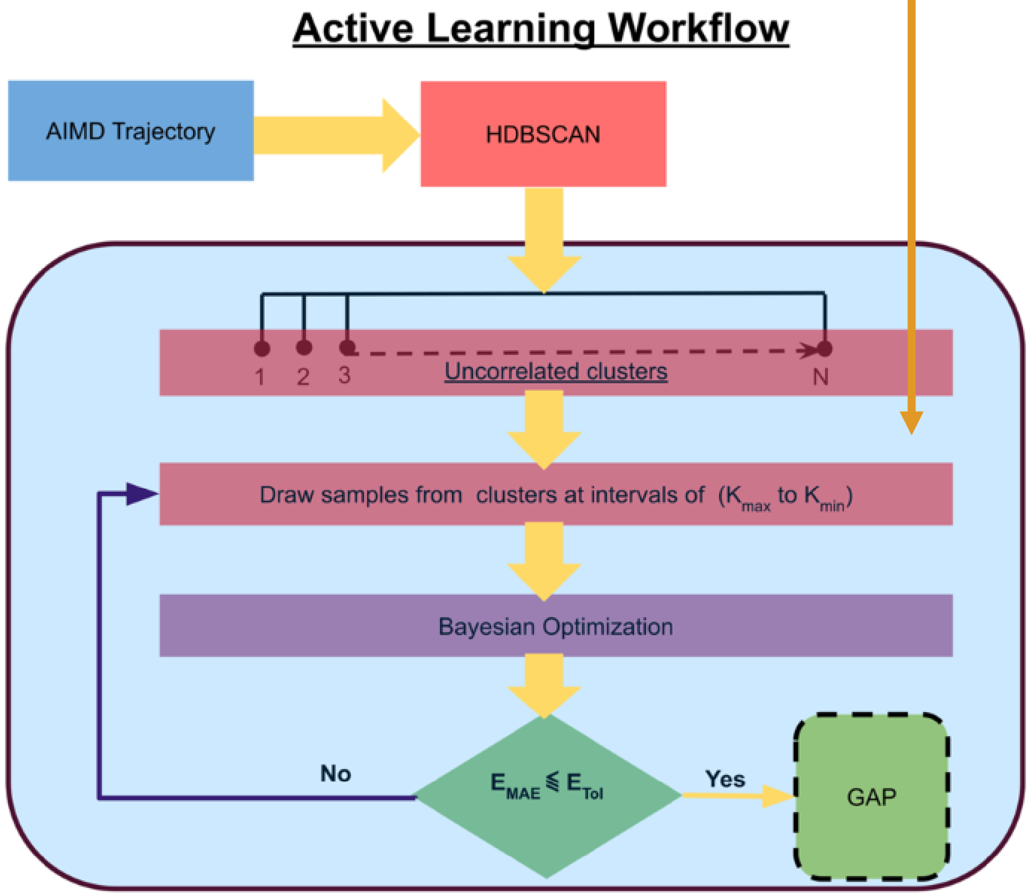
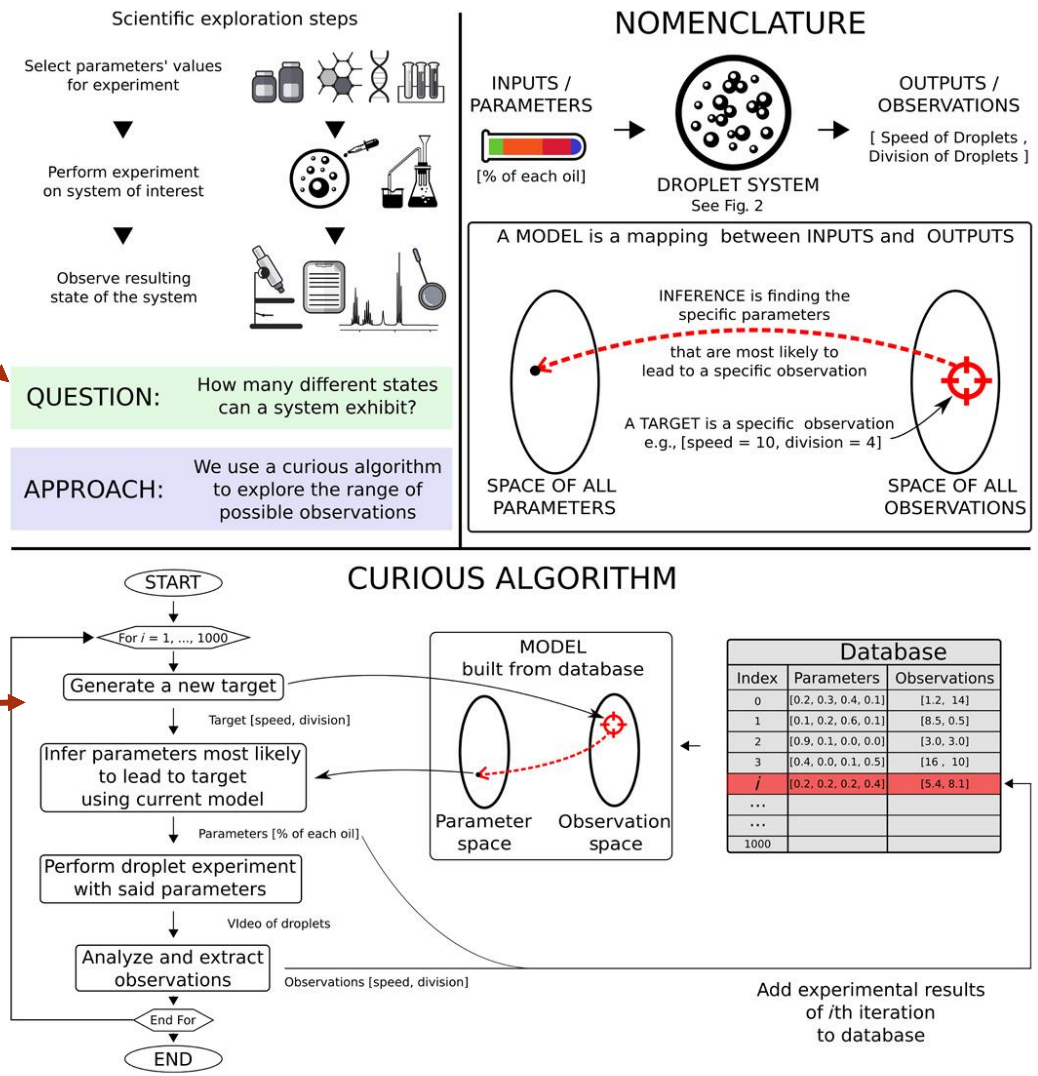


Figure: Sivaraman et al. npj Comp Mat. (2020)

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)

Pathway to Fully-Autonomous Laboratories

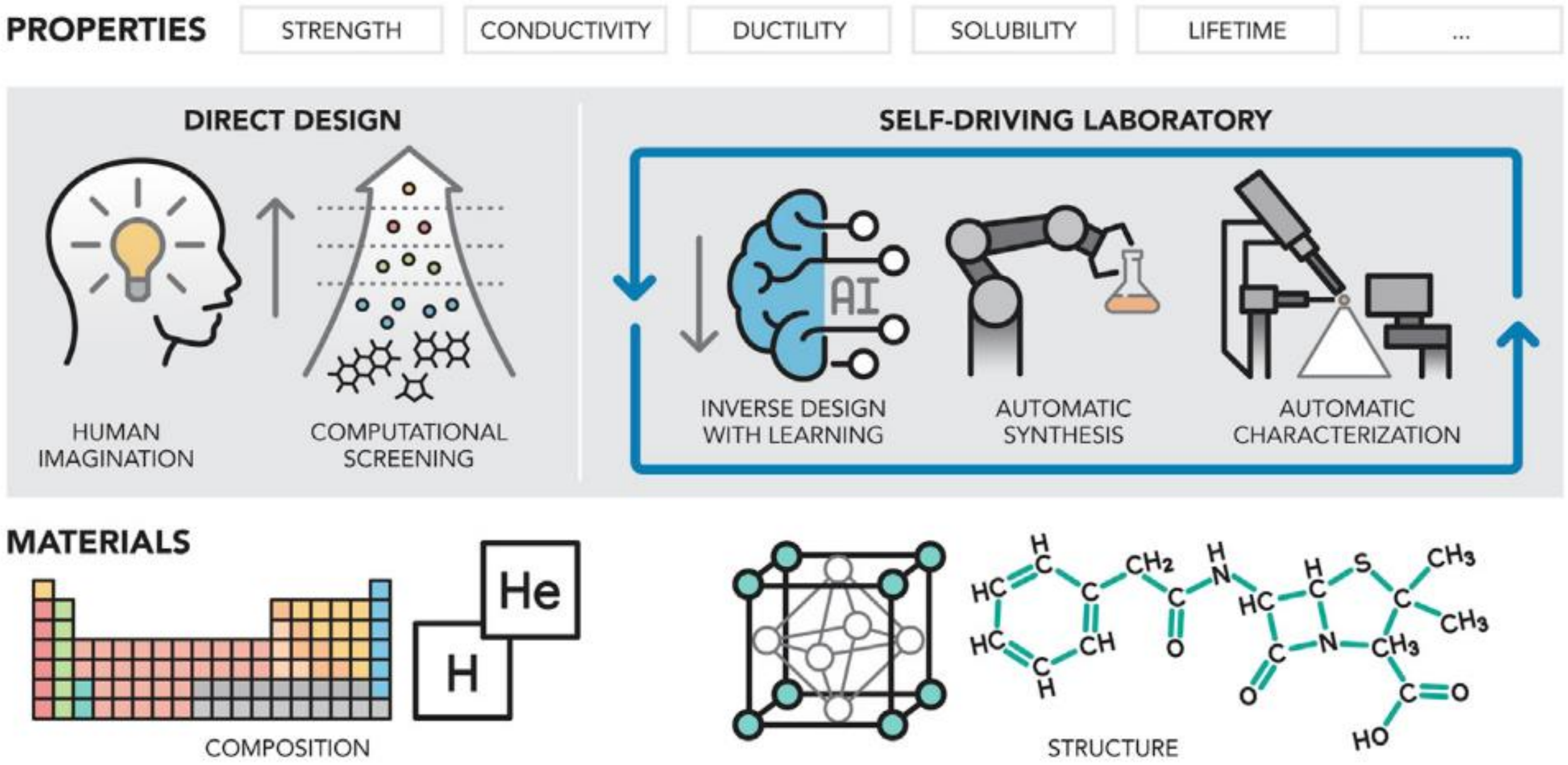


Figure 1. The Evolution of Materials Discovery Paradigms

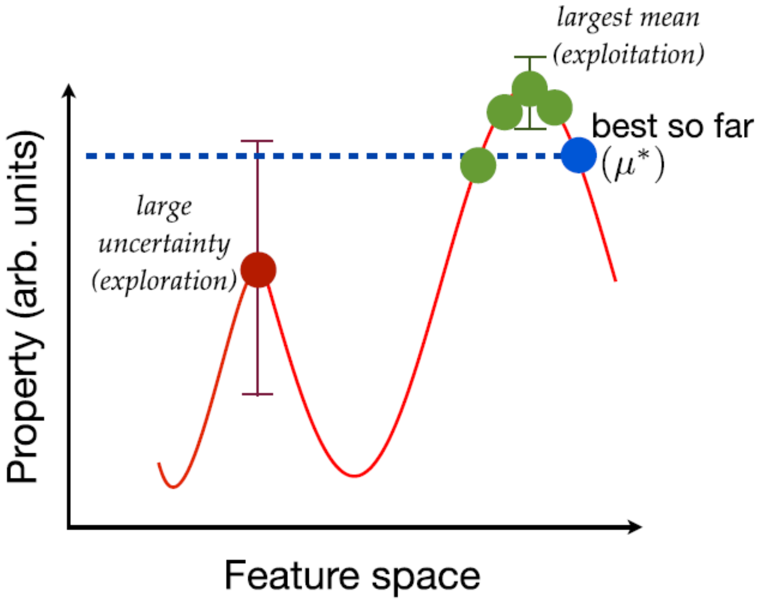
Ref: Crabtree. Joule (2020)

Take-Away Points

<u>"Optimal design" = "Learning while doing"</u>

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

Ex: "explore" vs "exploit"

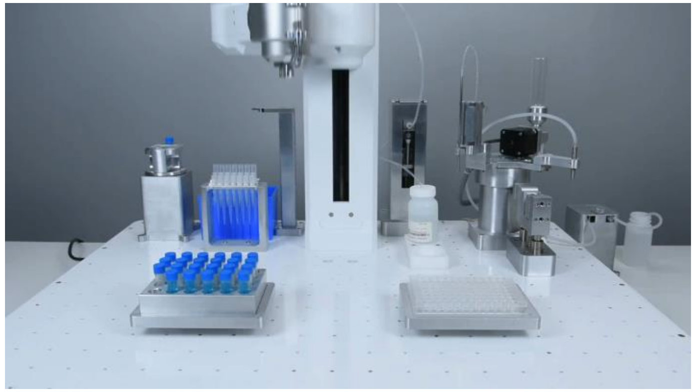


Source: [Balachandran et al.. Sci. Rep. (2016)](#)

Many ways to use active learning!

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

Next step ->



Source: Curtis Berlinguette (UBC)