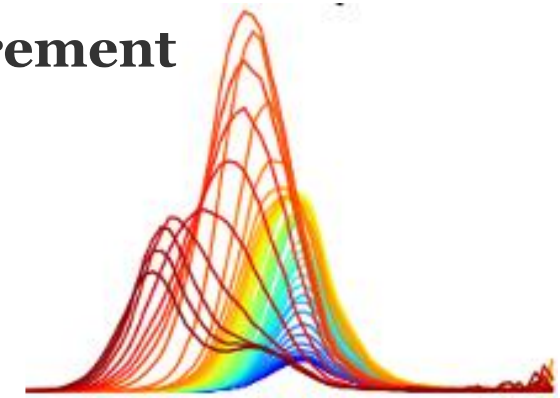
Causality and Dual BO

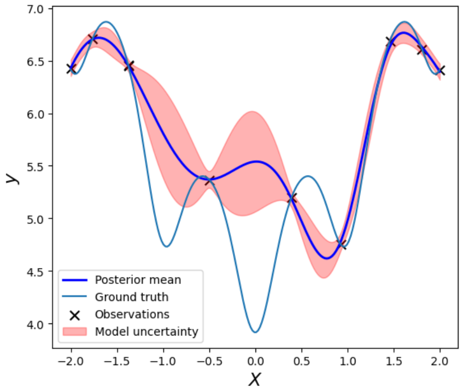
Sergei V. Kalinin

Classical Bayesian Optimization

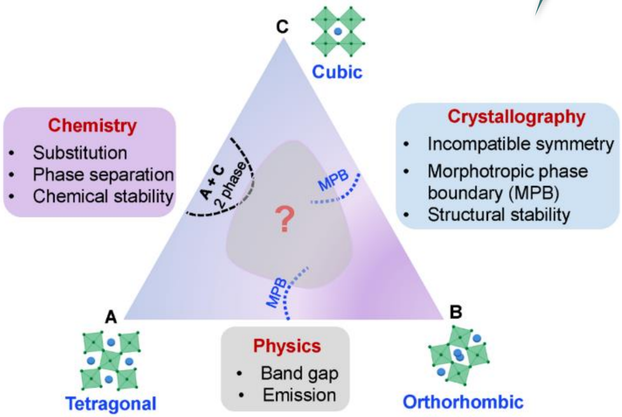
Measurement



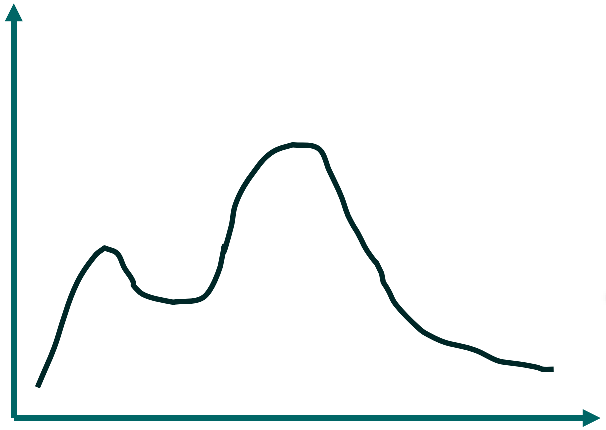
Surrogate model



Parameter space

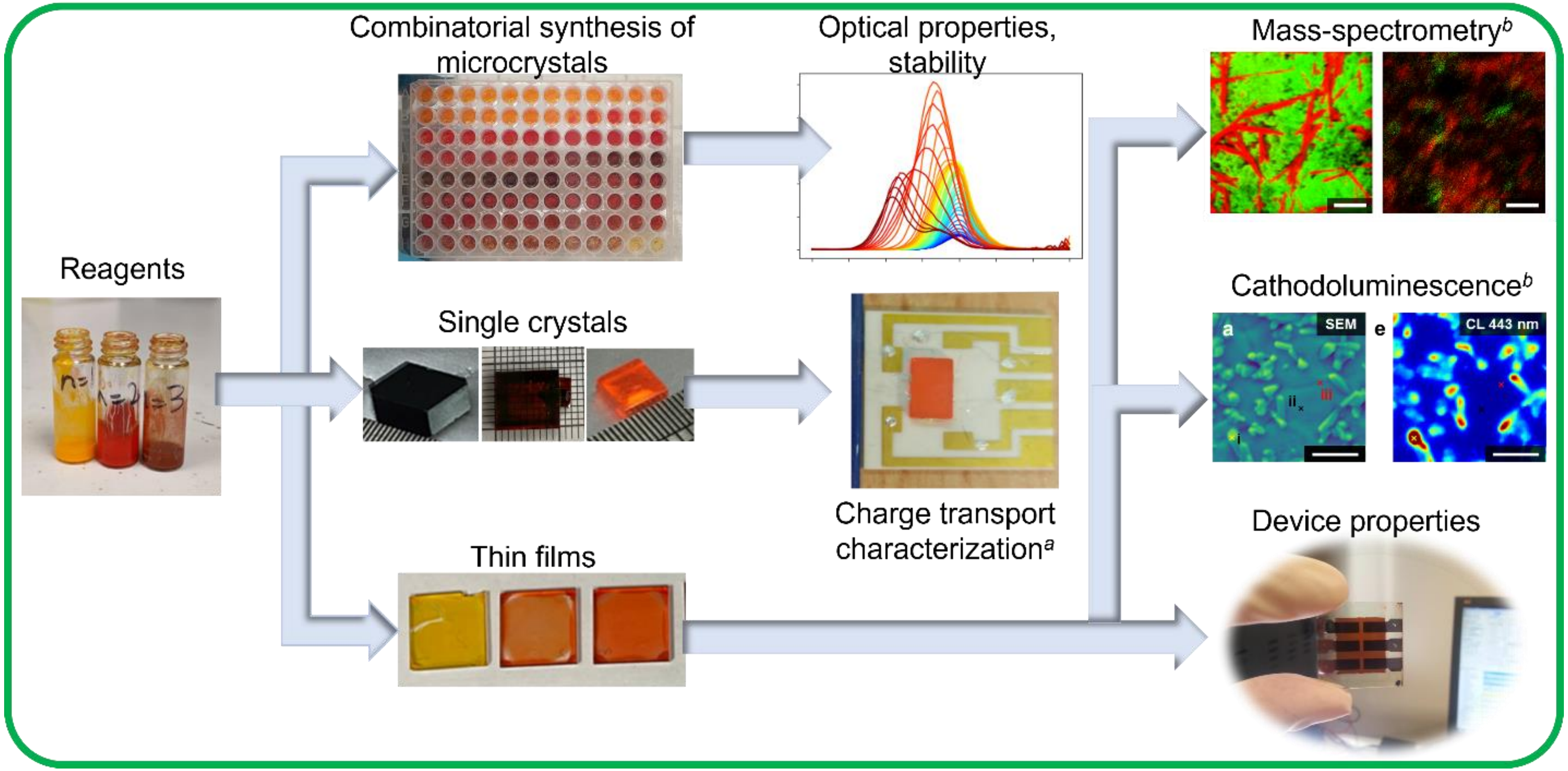


Acquisition function



Implicitly, we postulate that cost and gain are constant along the iterative cycle

What is A Workflow?



Workflow: ideation, orchestration, implementation

- Domain specific language
- Dynamic planning: latencies and costs
- Reward and value functions

- Are they optimal?
- Can we design them better?
- Can they be changed dynamically?

Building non-linear workflows

- Multi-Objective Bayesian Optimization

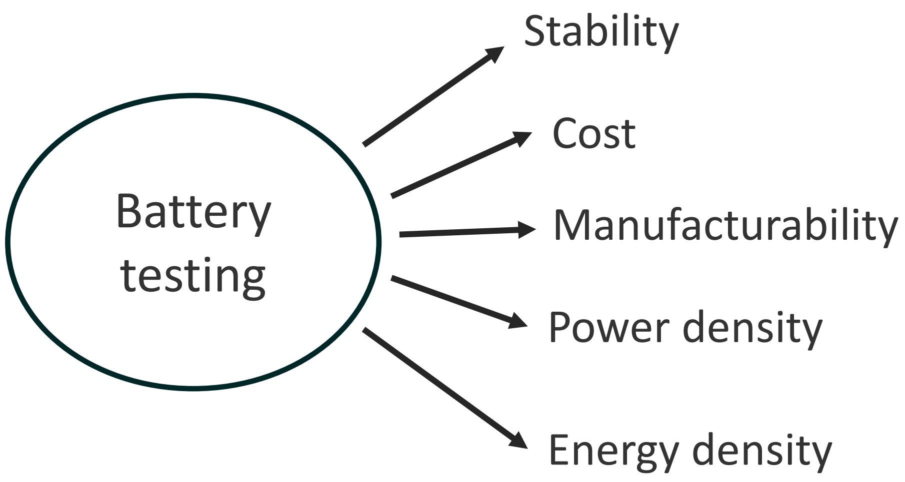
- Multifidelity Gaussian Processes
- Multitask Gaussian Processes

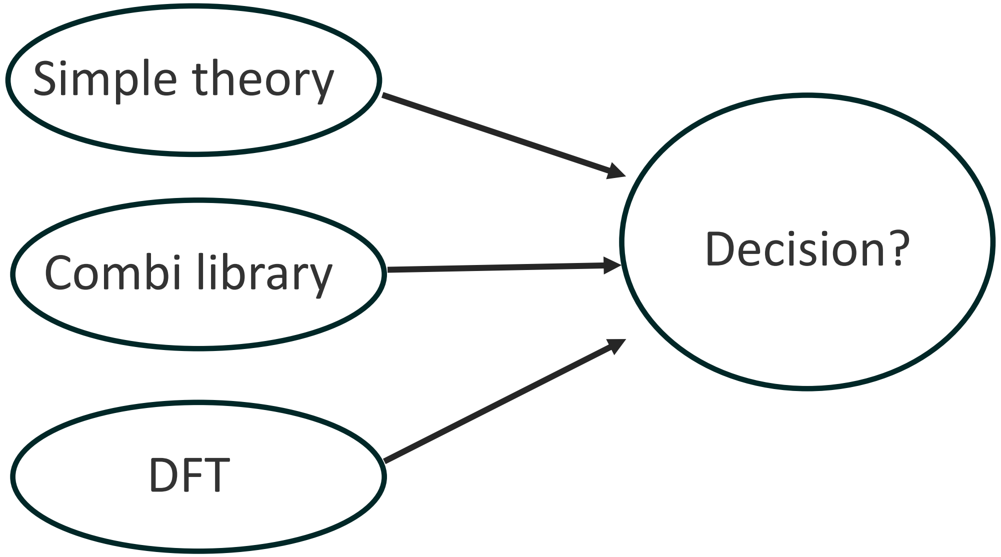
- Workflows with multidimensional data

- Co-orchestration multiple research tools
- Co-navigation of theory and experiment

- Perspectives

The real world is more complex!

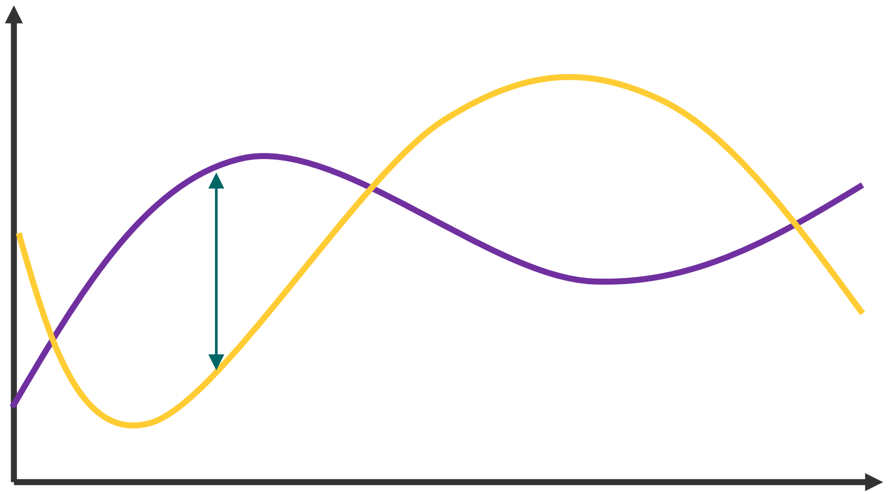




1. We need to balance multiple functionalities
2. Integrate multiple sources of data and make decisions considering costs, latencies, and beliefs

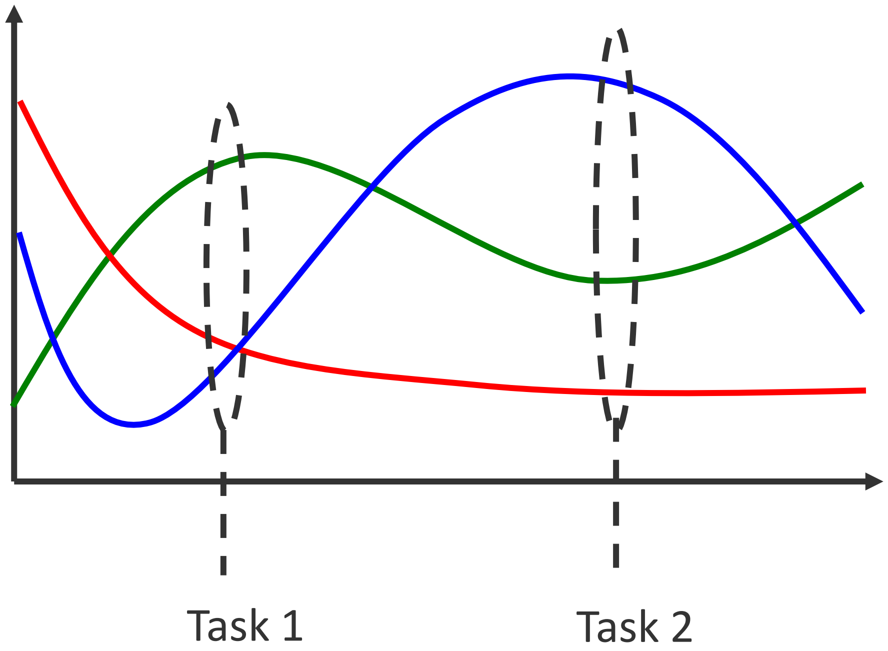
Multifidelity (structured) GP:

- We have the easy to evaluate function with probabilistic model and expensive to evaluate function
- The easy function is a proxy for expensive one and has some correlative relationship to it
- We create policy that balances evaluation costs

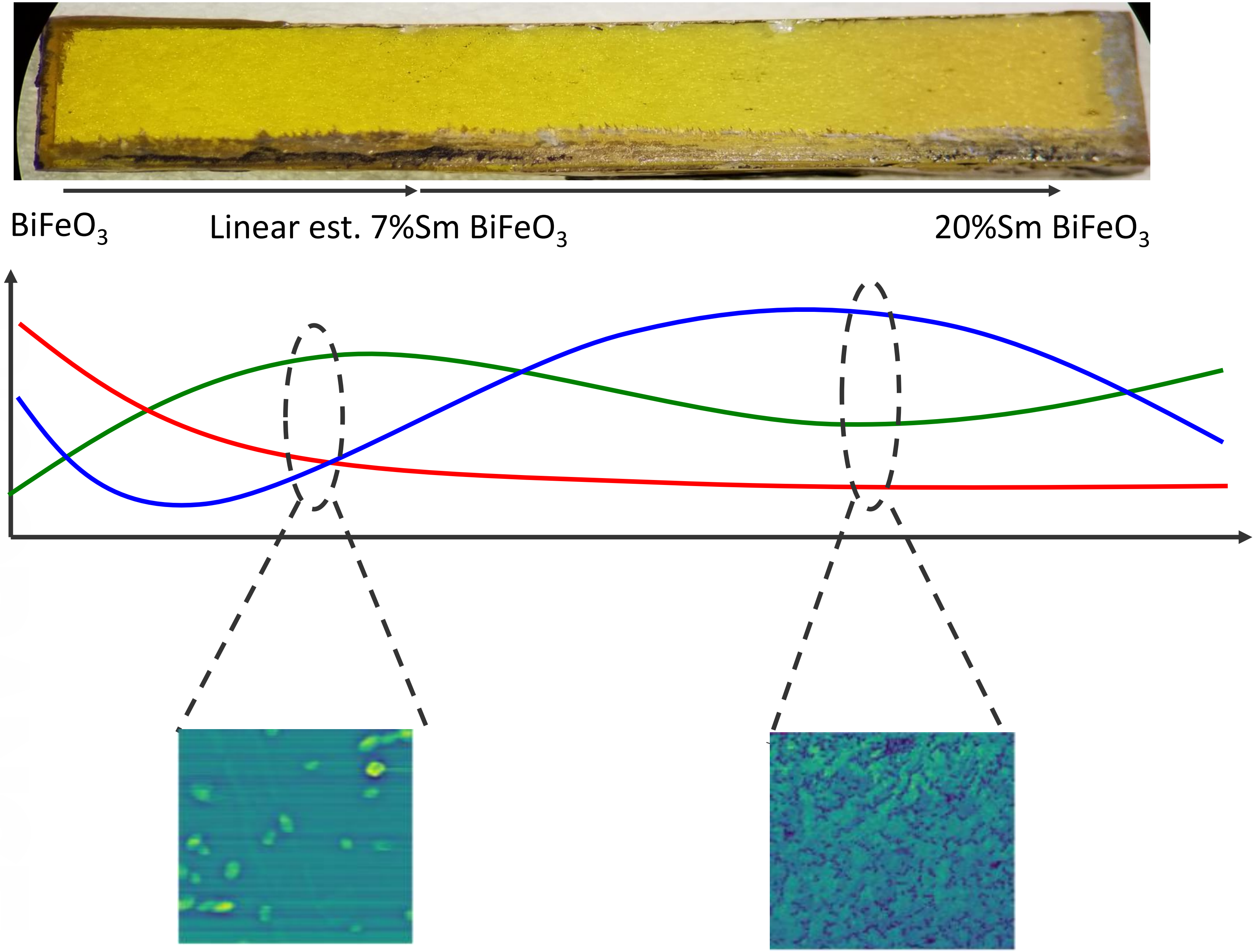


Multitask GP:

- We have multiple observables in different spaces
- And common **latent model** that emits them
- Can find minima in the expensive space suggested by cheap(er) function



General combinatorial library exploration



Compositional library (can be 1D or 2D, encoding from binary to quaternary diagrams). Composition c(x) or c(x,y) is assumed to be known

Non-observable latent variables that represent materials functionality. These form GP or sGP as a function of concentration (and via it, space)

The latent variables emit the observational data in the form of images or spectra (via second GP or decoder)

Experimental Instantiations

1. **Scenario I:** Data in full (microRaman across the combinatorial library, or grid measurement of topography or domains by PFM).
    a. Can use the simple VAE or GMM to find latents (or even PCA)
    b. However, VAE or GMM will not capture the spatial effects in sGP(c(x,y))

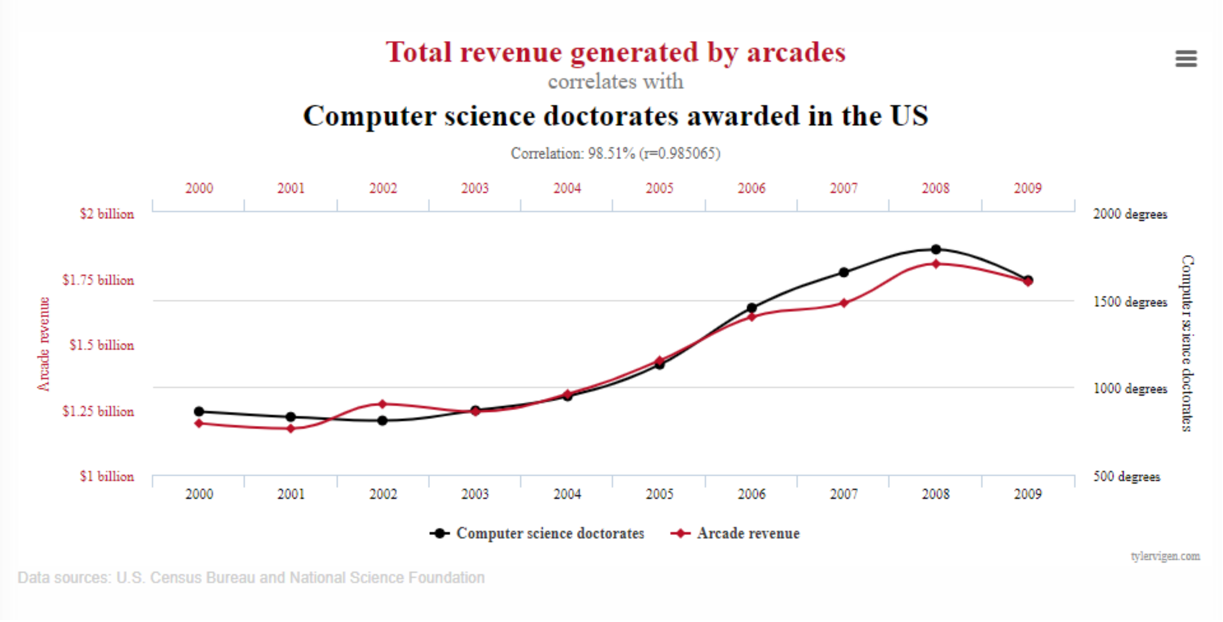
2. **Scenario II:** Active learning with one high dimensional imaging/spectroscopy method.
    a. Normal GP/sGP/HL if measured property is scalar (if we have good scalarizer for image/spectra)
    b. If it is active learning on images/spectra we do not have way to do it.

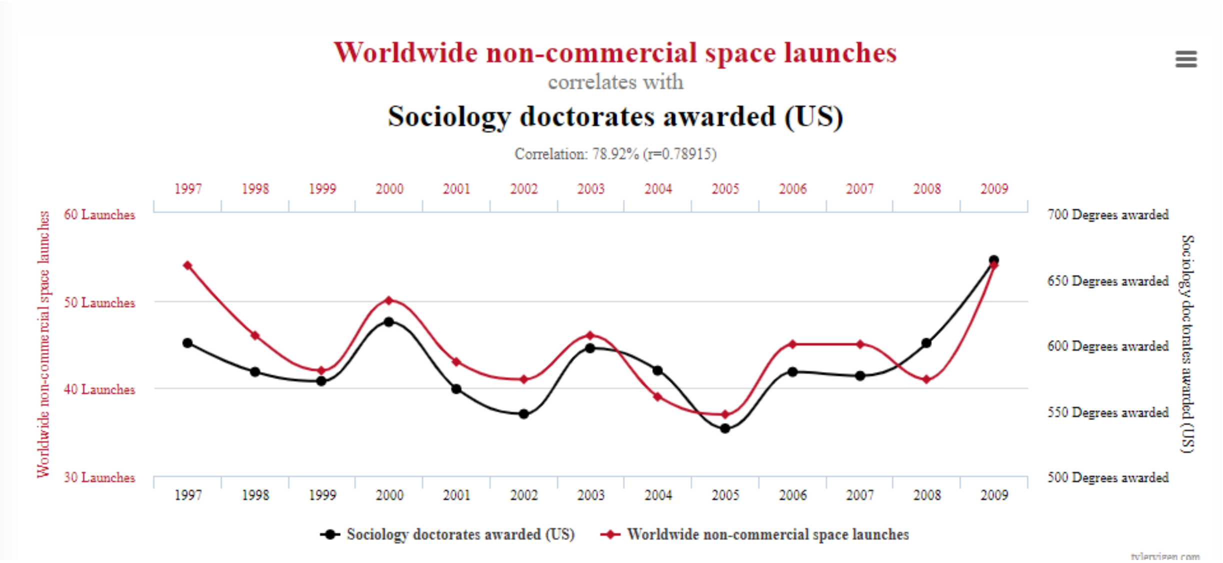
3. **Scenario III:** Active learning if we have full low dimensionality proxy data and active learning for low dimensional data. This is multifidelity GP and sGP

4. **Scenario IV:** Active learning when we have full high dimensional proxy data and use active learning for another high dimensional data (use Raman results to select places for STEM or PFM)

5. **Scenario V:** Co-navigation between 2 active high-dimensional data sets (meaning that measurements that emit from latents are different).

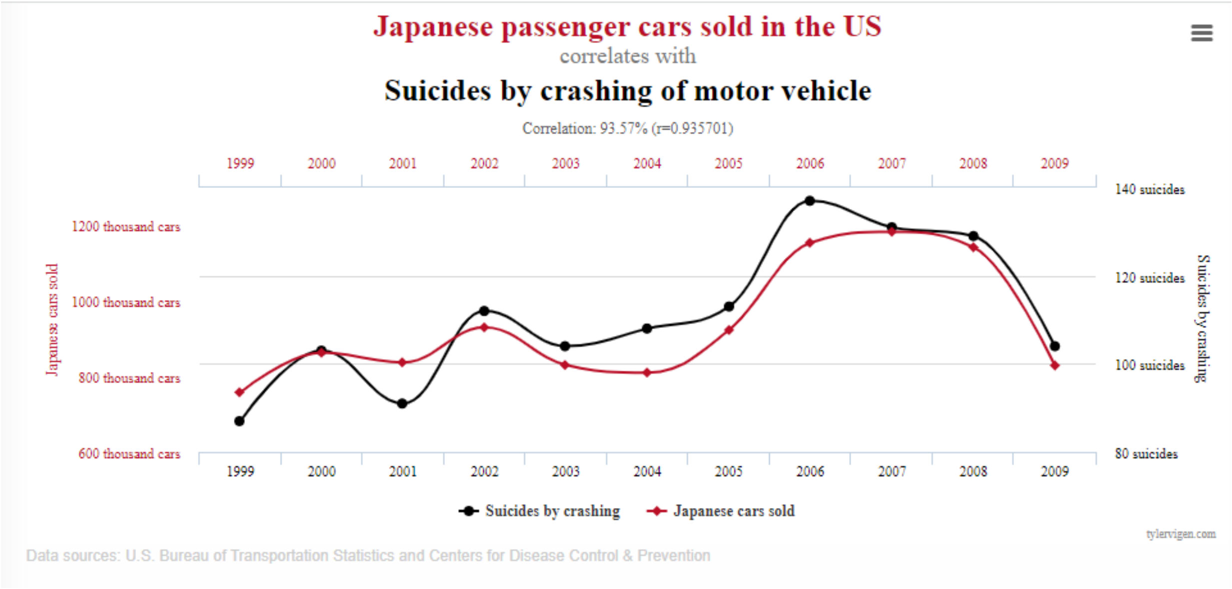
Correlation and causation

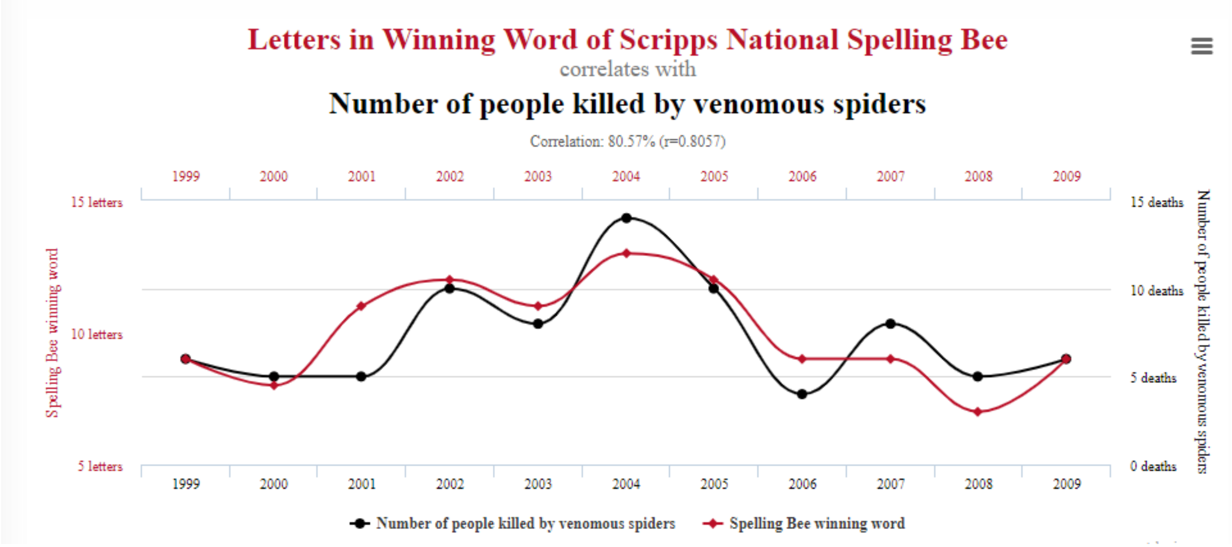




https://www.tylervigen.com/spurious-correlations

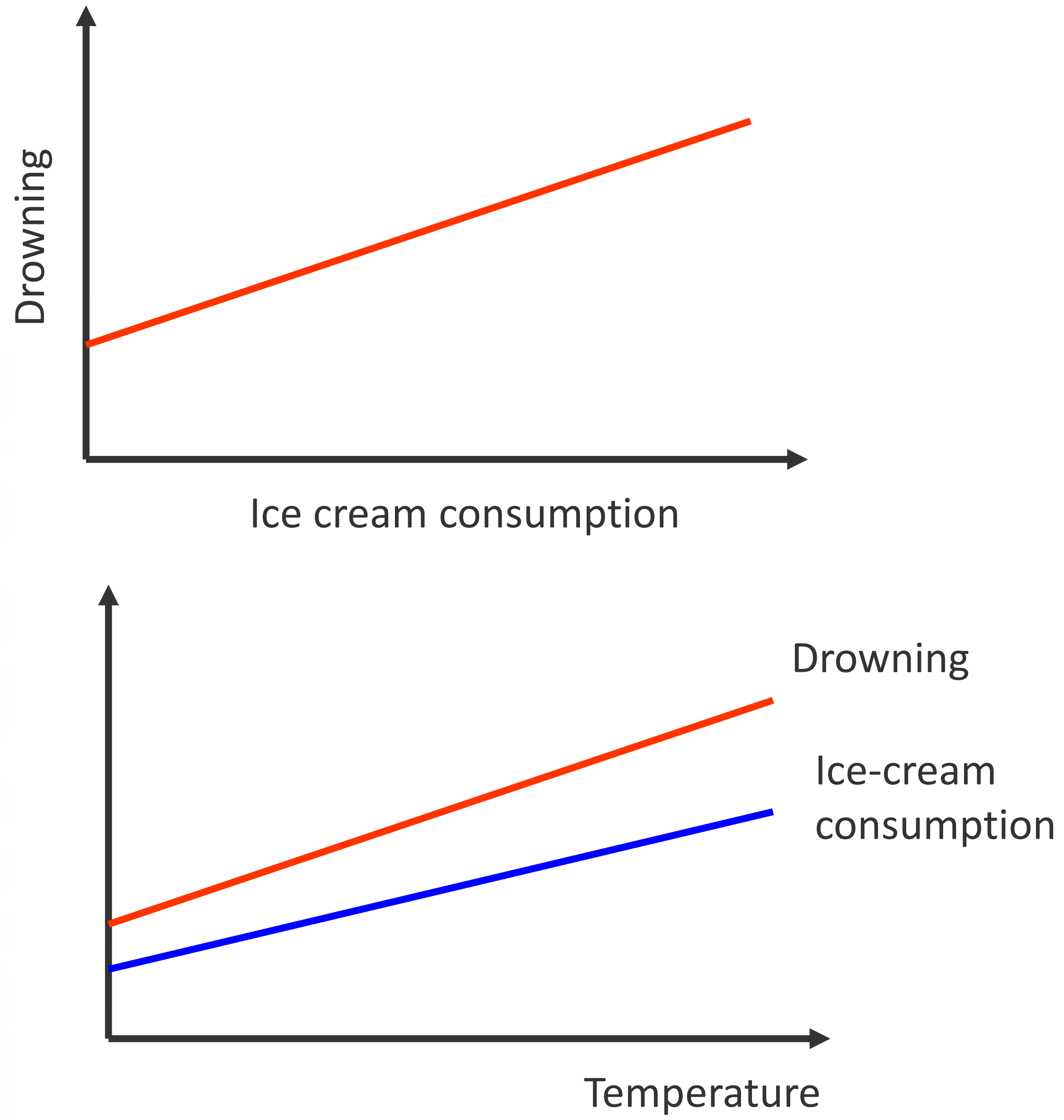
Correlation and causation

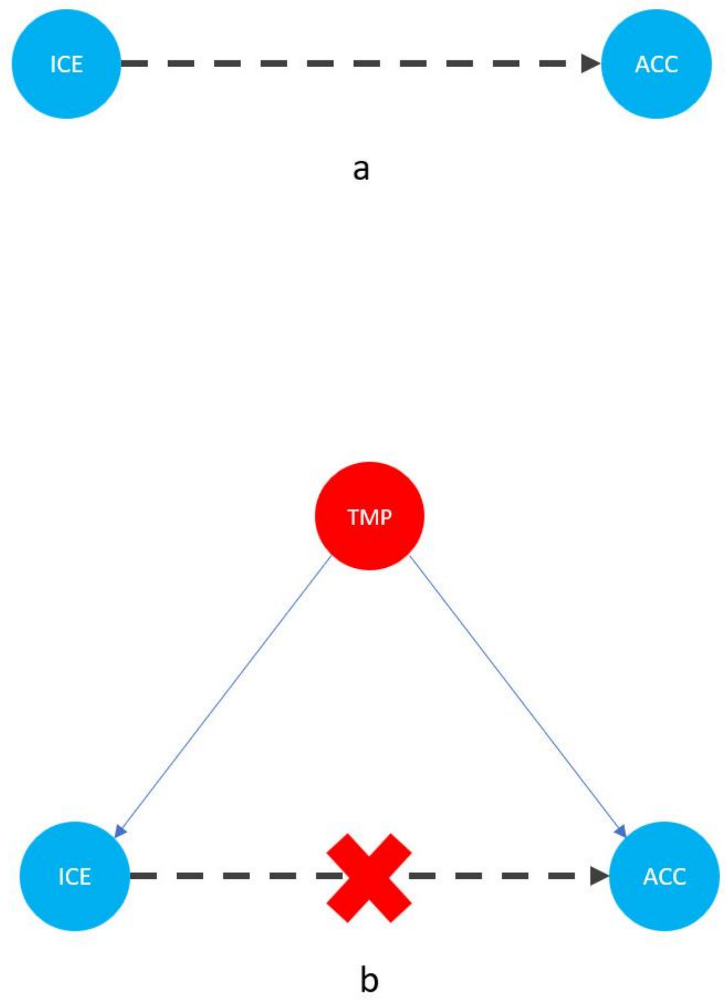




https://www.tylervigen.com/spurious-correlations

Ice-cream and drowning





From A. Molak, Causal Inference and Discovery in Python

Treatment effects

$$\tau_i = Y_i(1) - Y_i(0)$$

- τ_i is the treatment effect for person i
- $Y_i(1)$ is the outcome for person *i* when they received the treatment T
- $Y_i(0)$ is the outcome for the same person *i* given they did not receive the treatment

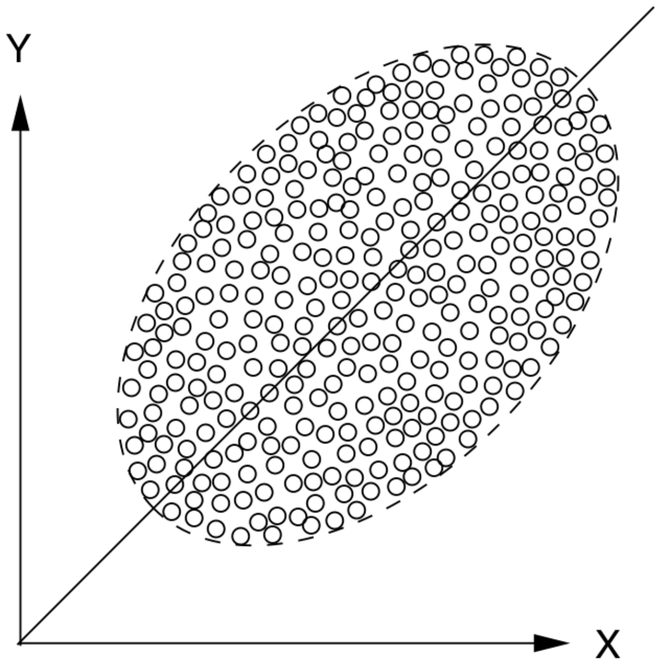
Very fundamental approach in:

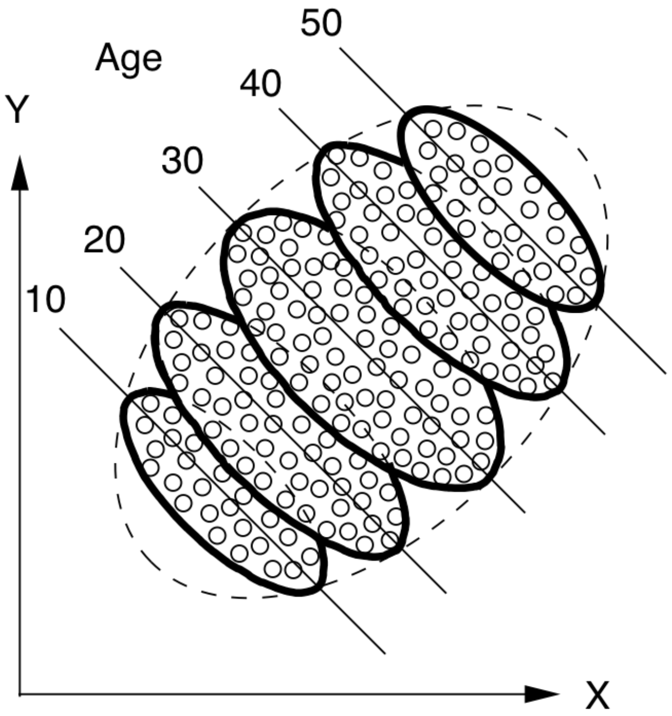
- Marketing
- Medicine
- And so on

But… How can the same person be treated and not treated?

From A. Molak, Causal Inference and Discovery in Python

Simpson paradox





Exercise is helpful in every age group but harmful for a typical person.

Is exercise helpful or not?

Simpson paradox

Drug	A		B	
Blood clot	Yes	No	Yes	No
Total	27	95	23	99
Percentage	22%	78%	19%	81%

Drug	A		B	
Blood clot	Yes	No	Yes	No
Female	24	56	17	25
Male	3	39	6	74
Total	27	95	23	99
Percentage	22%	78%	18%	82%
Percentage (F)	30%	70%	40%	60%
Percentage (M)	7%	93%	7.5%	92.5%

- Simpson's paradox appears when data partitioning (which we can achieve by controlling for the additional variable(s) in the regression setting) significantly changes the outcome of the analysis.
- In the real world, there are usually many ways to partition your data.
- You might ask: okay, so how do I know which partitioning is the *correct* one?

From A. Molak, Causal Inference and Discovery in Python

Berkeley discrimination lawsuit

In the early 1970s, the University of California, Berkeley was sued for gender discrimination over admission to graduate school. Of the 8,442 male applicants for the fall of 1973, 44 percent were admitted, but only 35 percent of the 4,351 female applicants were accepted

Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case

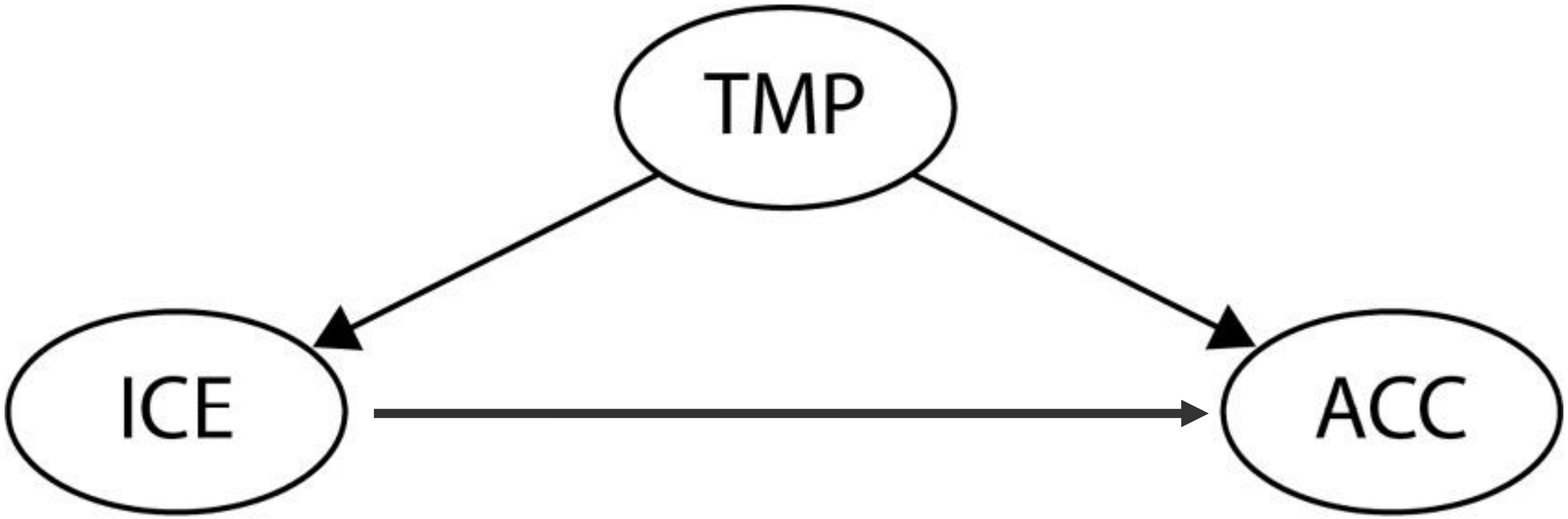
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Source: Bickel, Hammel, and O'Connell (1975); table accessed via Wikipedia at https://en.wikipedia.org/wiki/Simpson%27s_paradox

In the Berkeley case, the "paradox" occurred because women disproportionately applied to departments with low acceptance rates, as shown in the table above, while men disproportionately applied to departments with high acceptance rates.

https://www.brookings.edu/articles/when-average-isnt-good-enough-simpsons-paradox-in-education-and-earnings/

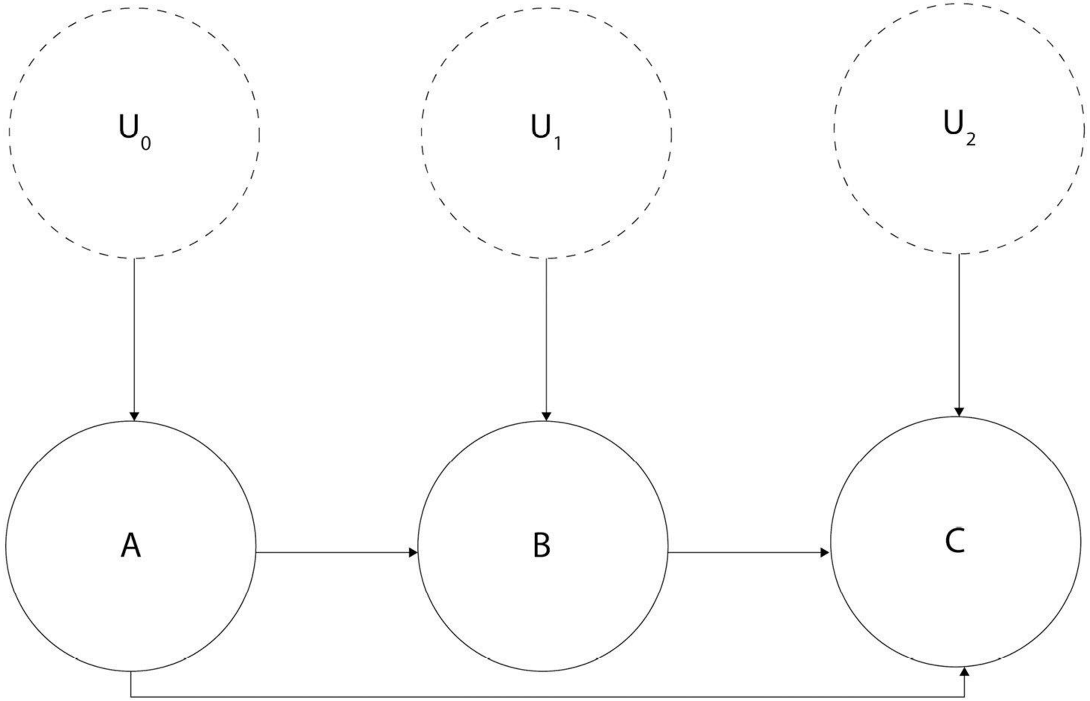
How can we even approach such problems?



- Observations give us correlations between temperature, ice cream consumption, and accident rate
- What we need to know is the causal links between these characteristics. Does change in ice cream consumption affect temperature or accident rate?
- But we cannot make an experiment!

From A. Molak, Causal Inference and Discovery in Python

Causal graphs



$$A := f_A(U_0)$$

$$B := f_B(A, U_1)$$

$$C := f_C(A, B, U_2)$$

- Here, := is an **assignment operator**, also known as a **walrus operator**. We use it to emphasize that the relationship that we're describing is *directional* (or asymmetric), as opposed to the regular equal sign that suggests a symmetric relation.

- And f_A, f_B, f_C represent arbitrary functions (they can be as simple as a summation or as complex as you want).

From A. Molak, Causal Inference and Discovery in Python

Do-operator

Conditioning: $P(X = x | Y = y)$

Intervention: $P(Y = 1 | do(X = 0))$

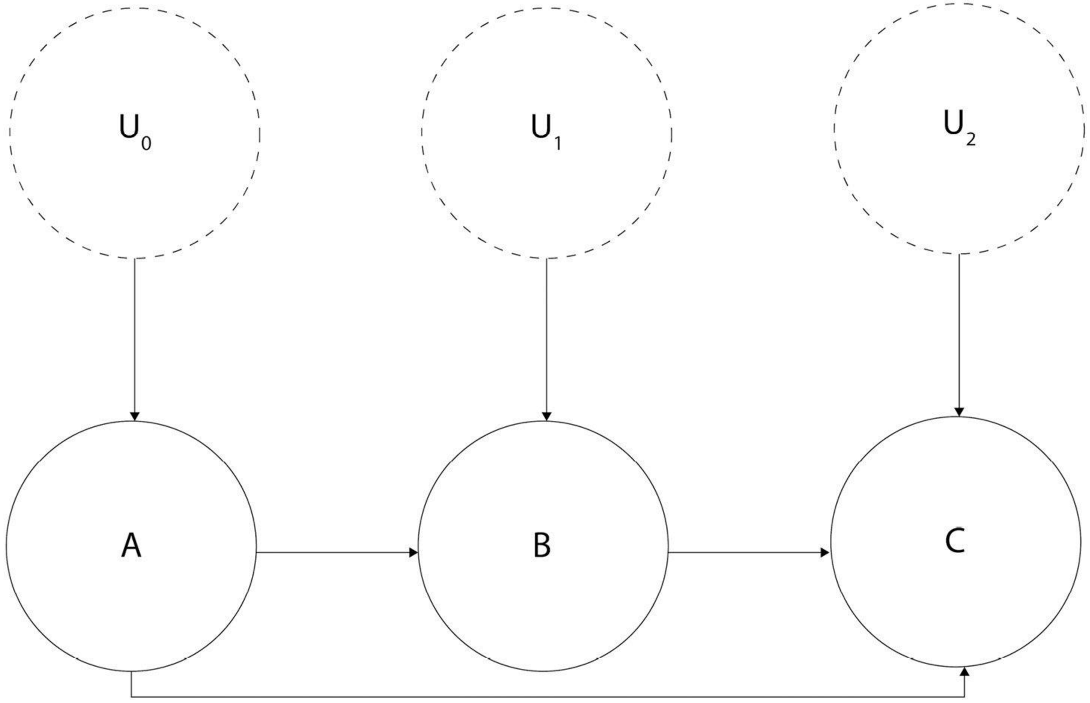
- Conditioning only modifies our *view* of the data, while intervening affects the distribution by *actively* setting one (or more) variable(s) to a *fixed value* (or a distribution).

- This is very important – intervention *changes* the system, but conditioning *does not*.

- You might ask, what does it mean that *intervention changes the system*? Great question!

From A. Molak, Causal Inference and Discovery in Python

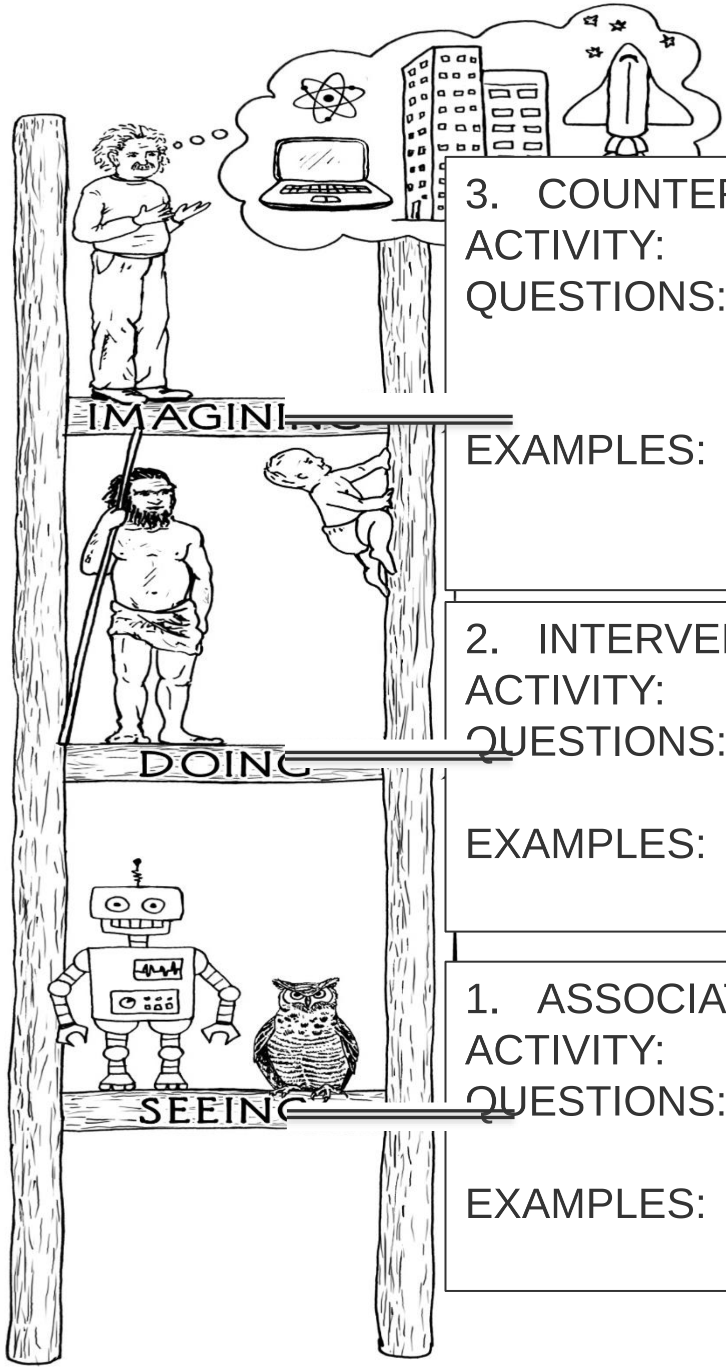
Properties of do - operator



- The change in B will influence the values of its descendants
- B will become independent of its ancestors

From A. Molak, Causal Inference and Discovery in Python

Ladder of causation



3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: *What if I had done . . . ? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache?
Would Kennedy be alive if Oswald had not killed him?
What if I had not smoked the last 2 years?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)
EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

From J. Pearl presentation, Why - 19

How can we learn causality

- **Causal discovery** and **causal structure learning** are umbrella terms for various kinds of methods used to uncover causal structure from observational or interventional data.

- **Expert knowledge** is a term covering various types of knowledge that can help define or disambiguate causal relations between two or more variables. Depending on the context, expert knowledge might refer to knowledge from randomized controlled trials, laws of physics, a broad scope of experiences in a given area, and more.

- **Combining causal discovery and expert knowledge:** Some causal discovery algorithms allow us to easily incorporate expert knowledge as a priority. This means that we can either *freeze* certain edges in the graph or *suggest* the existence or direction of these edges.

From A. Molak, Causal Inference and Discovery in Python

Independence and conditional independence

- Notation for independence involves the symbol, $\perp\!\!\!\perp$ (usually called *double up tack*), whose form visually encodes the notion of orthogonality.
- We can express the fact that X and Y are independent in the following way:

$$P(X, Y) = P(X)P(Y) \qquad X \perp\!\!\!\perp Y$$

- The concept of independence plays a vital role in statistics and causality.
- Its generalization – **conditional independence** – is even more important. We say that X and Y are conditionally independent given Z, when X does not give us any new information about Y assuming that we observed Z.
-

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \qquad X \perp\!\!\!\perp Y|Z$$

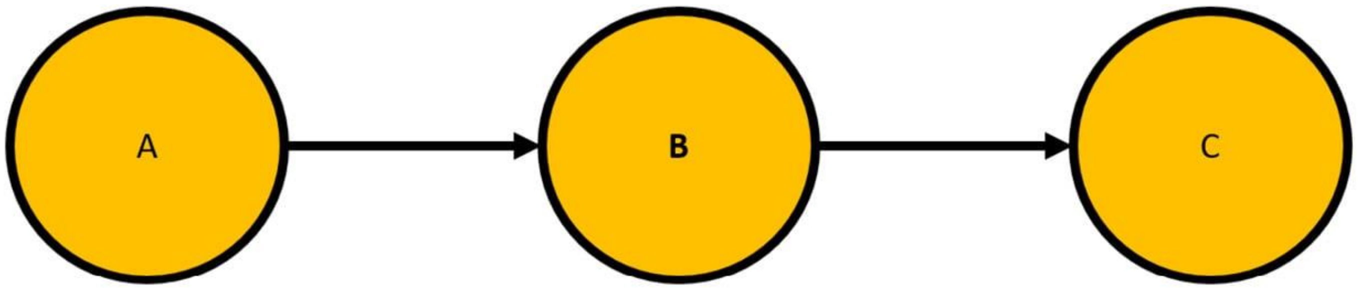
From A. Molak, Causal Inference and Discovery in Python

Conditional and unconditional independence

- We say that two nodes are *unconditionally* (or marginally) *independent* in the graph when there's *no open path* that connects them *directly* or *indirectly*.
- We say that two nodes, X and Y, are *conditionally independent* given (a set of) node(s) Z when Z blocks *all open paths* that connect X and Y.

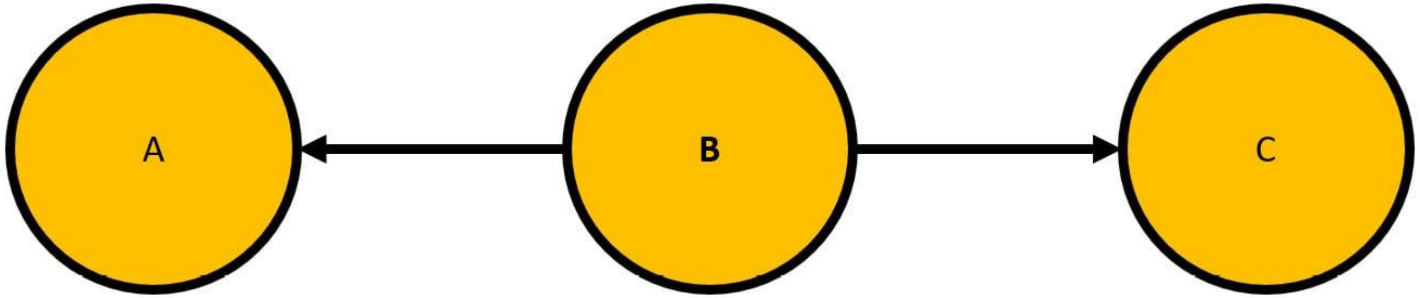
From A. Molak, Causal Inference and Discovery in Python

Chains and forks



$$A \perp\!\!\!\perp_G C | B$$

Chain 🔗

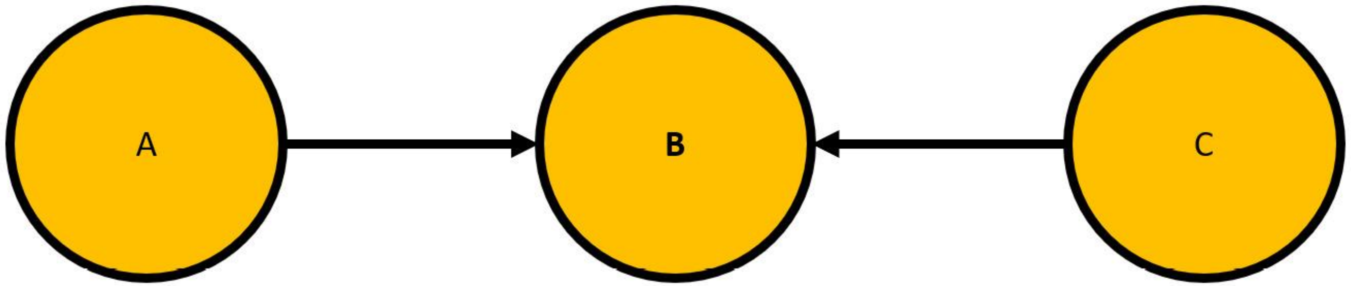


$$A \perp\!\!\!\perp_G C | B$$

Fork 🍴

From A. Molak, Causal Inference and Discovery in Python

Colliders



$$A \perp\!\!\!\perp C$$

$$A \not\!\perp\!\!\!\perp C \mid B$$

Collider 💥

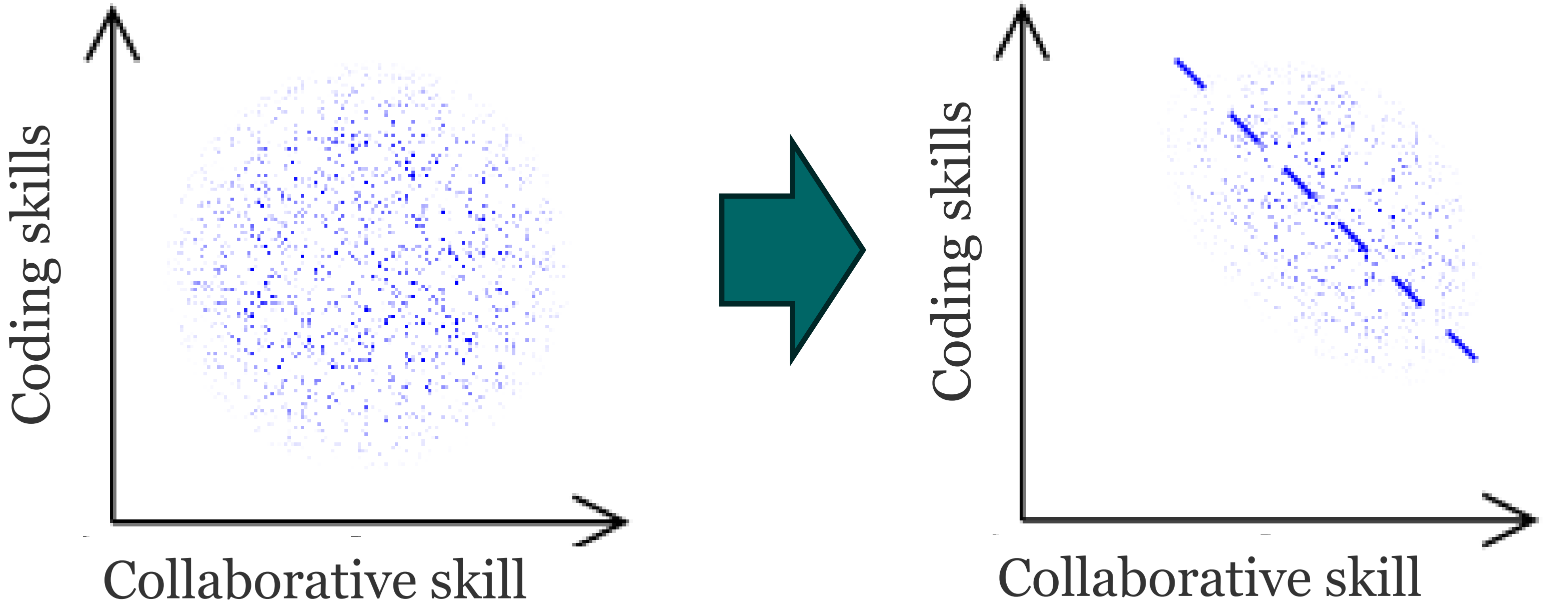
Imagine that both A and C randomly generate integers between *1* and *3*. Let's also say that B is a sum of A and C. Now, let's take a look at values of A and C when the value of B is 4. The following are the combinations of A and C that lead to B = 4:

- A = 1, C = 3
- A = 2, C = 2
- A = 3, C = 1

Although A and C are unconditionally independent (there's no correlation between them as they randomly and independently generate integers), they become correlated when we observe C !

From A. Molak, Causal Inference and Discovery in Python

Colliders and Berkson paradox



Many companies might hire people based on their skills and their personality traits. Imagine that company *X* quantifies a person's coding skills on a scale from one to five. They do the same for the candidate's ability to cooperate and hire everyone who gets a total score of at least seven. Assuming that coding skills and ability to cooperate are independent in the population (which doesn't have to be true in reality), you'll observe that in company *X*, people who are better coders are less likely to cooperate on average, and those who are more likely to cooperate have fewer coding skills. You could conclude that being non-cooperative is related to being a better coder, yet this conclusion would be incorrect in the general population.

https://en.wikipedia.org/wiki/Berkson%27s_paradox From A. Molak, Causal Inference and Discovery in Python

Estimator, estimate, and estimand

1. **Estimand**: The quantity or parameter that is intended to be estimated. It represents the true value of the parameter in the population. Suppose you're interested in the average height of adult men in a particular country. The actual average height of all adult men in that country is the estimand.

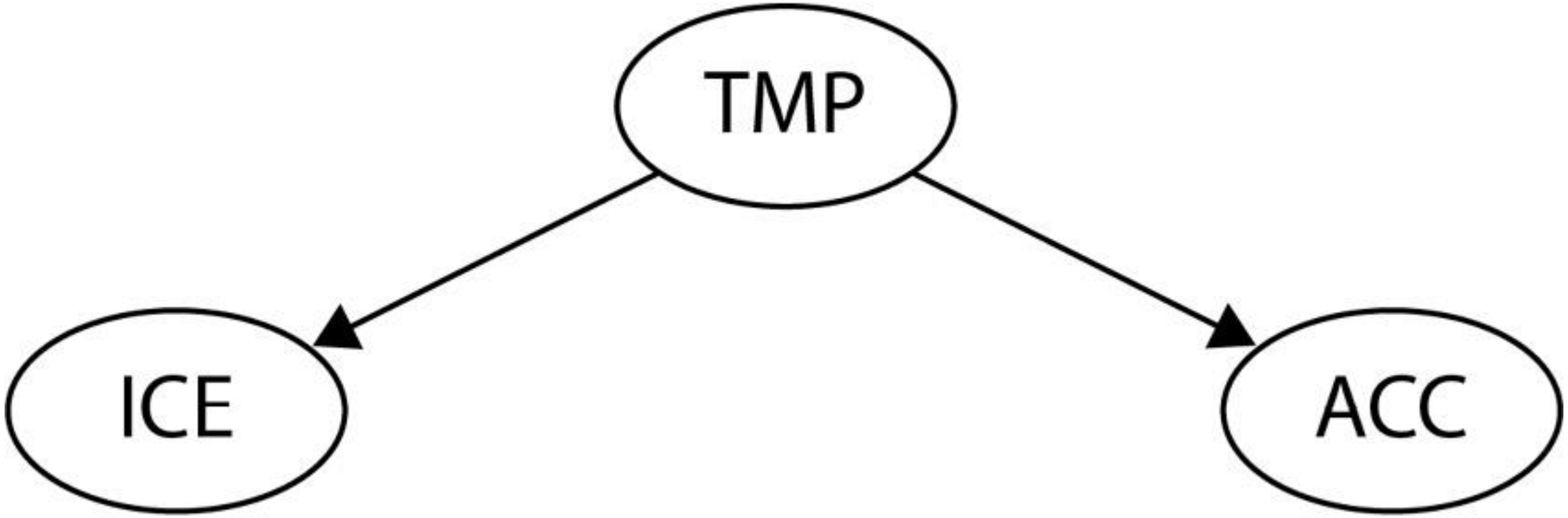
2. **Estimate**: The approximation or value obtained from the data to estimate the estimand. This is derived from a sample and used to infer information about the population. For example, from a sample of 1,000 adult men in the same country, you calculate an average height of 5 feet 9 inches. This value (5 feet 9 inches) is your estimate of the average height (the estimand).

3. **Estimator**: A rule, formula, or algorithm by which you derive the estimate from the data. It is a function of the sample data and is used to produce an estimate of the estimand. Here, the formula for calculating the mean (average) from a set of numbers is an estimator. When you apply this formula to your sample data, you obtain the estimate.

Estimator, estimate, and estimand

- **Estimand**: What you want to know (the actual, often unknown, value).
- **Estimate**: What you got from your sample data.
- **Estimator**: How you got it (the method or formula used).

Adjustment

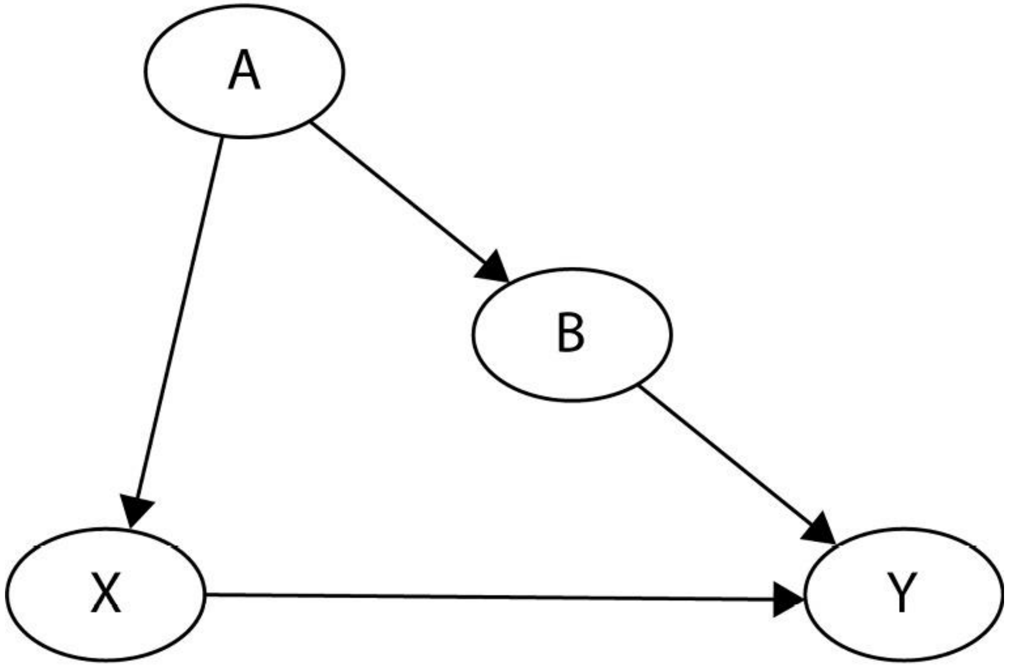


$$ACC \sim ICE + TMP$$

$$P(ACC|do(ICE)) = \sum_{tmp} P(ACC|ICE, TMP)P(TMP)$$

From A. Molak, Causal Inference and Discovery in Python

Back door criterion

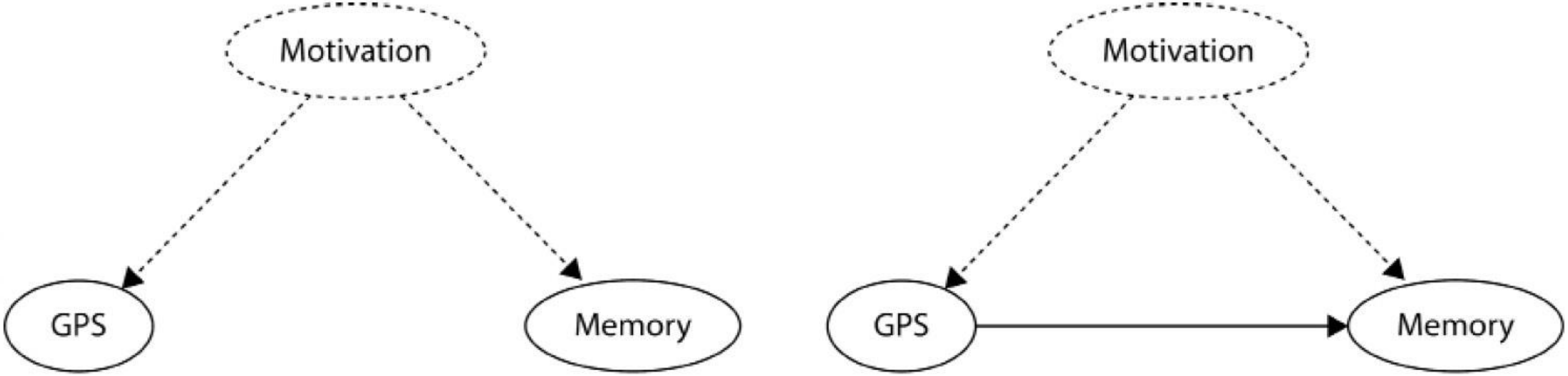


$$P(Y = y | do(X = x)) = \sum_a P(Y = y | X = x, A = a) P(A = a)$$
$$= \sum_b P(Y = y | X = x, B = b) P(B = b)$$

We can estimate effect even if one of A, B is unobserved!

From A. Molak, Causal Inference and Discovery in Python

Front door criterion and mediation



Observation: People that use GPS more have less good memory

Hypothesis 1: Usage of GPS precludes memory development

Hypothesis 2: There is a common (unobserved) factor that affects both GPS usage and memory

From A. Molak, Causal Inference and Discovery in Python

Mediation

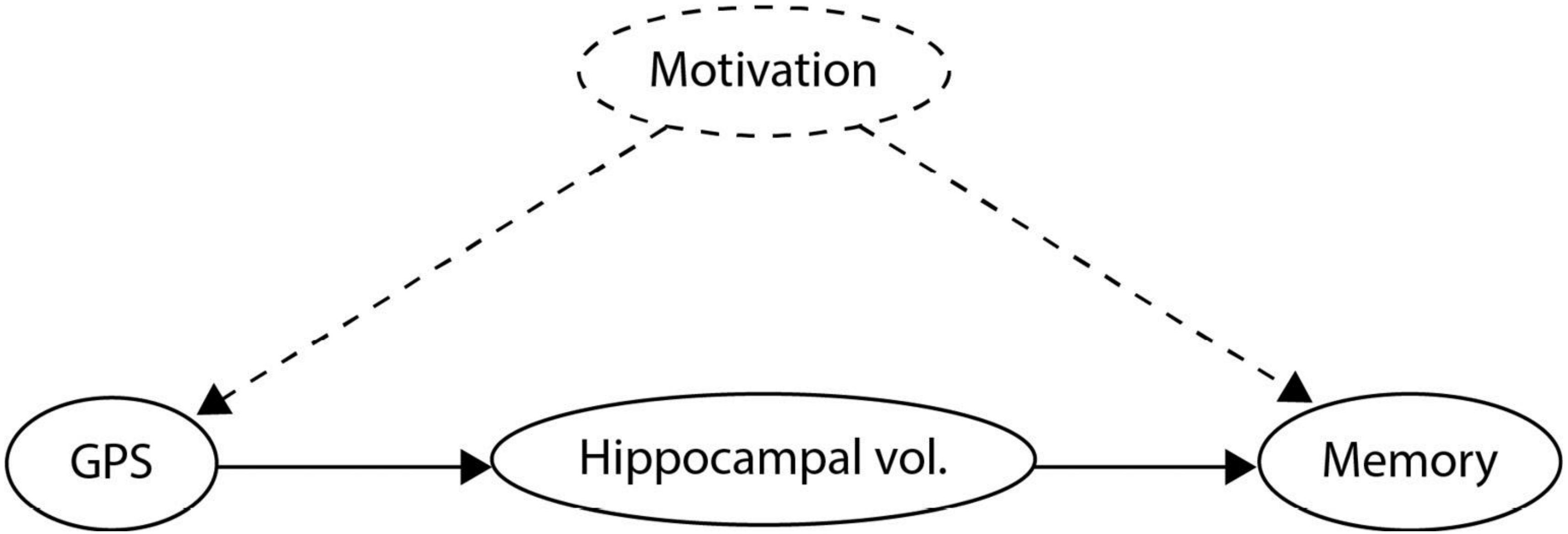
- The influence of one variable X on another Y is *mediated* by a third variable, Z (or a set of variables, **Z**), when at least one path from X to Y goes through Z.

- Z *fully mediates* the relationship between X and Y when the only path from X to Y goes through Z.

- If there are paths from X to Y that do not pass through Z, the mediation is *partial*.

From A. Molak, Causal Inference and Discovery in Python

Front door criterion

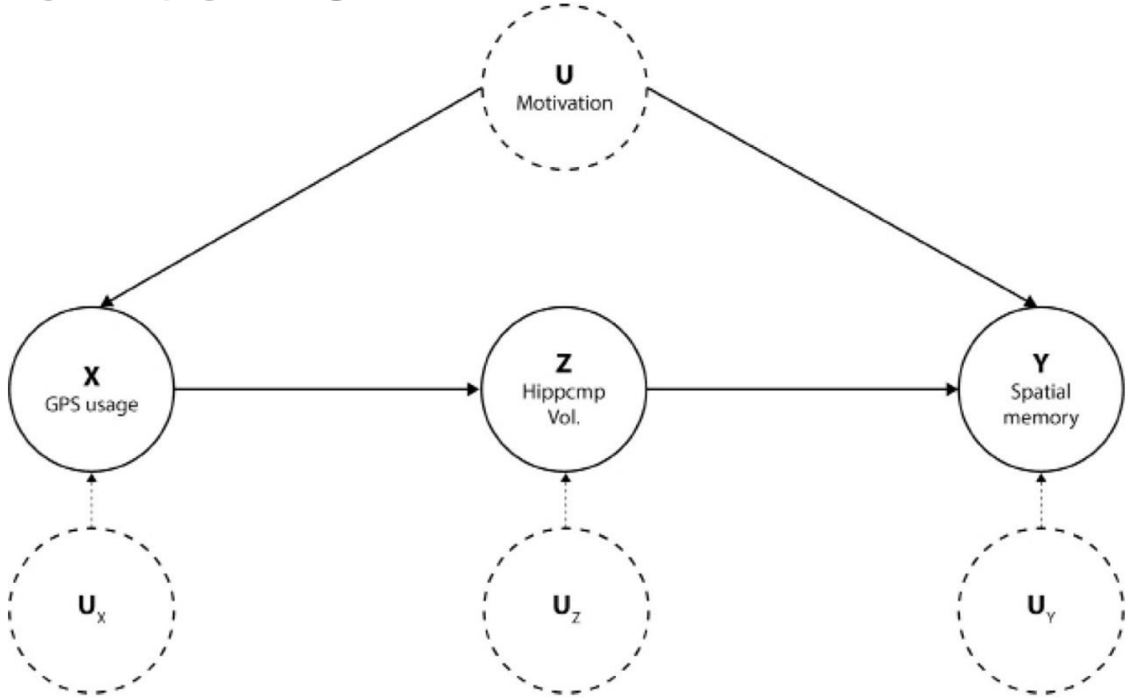


- We assume that hippocampal volume fully mediates the effects of GPS usage on a decline in spatial memory.
- The second important assumption we make is that motivation can only affect *hippocampal volume indirectly through GPS usage.*

If motivation would be able to influence hippocampal volume *directly*, front-door would be of no help. Luckily enough, the assumption that motivation cannot directly change the volume of the hippocampus seems reasonable (though perhaps you could argue against it!).

From A. Molak, Causal Inference and Discovery in Python

Front door criterion



$$P(Y = y | do(X = x)) = \sum_z P(Z = z | X = x) \sum_{x'} P(Y = y | X = x', Z = z) P(X = x')$$

- Fit a model, $Z \sim X$
- Fit a model, $Y \sim Z + X$
- Multiply the coefficients from model 1 and model 2

From A. Molak, Causal Inference and Discovery in Python

Do-calculus

- *Rule 1*: When an observation can be ignored:

$$P(Y = y|do(X = x), Z = z, W = w) = P(Y = y|do(X = x), W = w) \;\; if\, (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$$

- *Rule 2*: When intervention can be treated as an observation:

$$P(Y = y|do(X = x), do(Z = z), W = w) = P(Y = y|do(X = x), Z = z, W = w) \;\; if\, (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$$

- *Rule 3*: When intervention can be ignored:

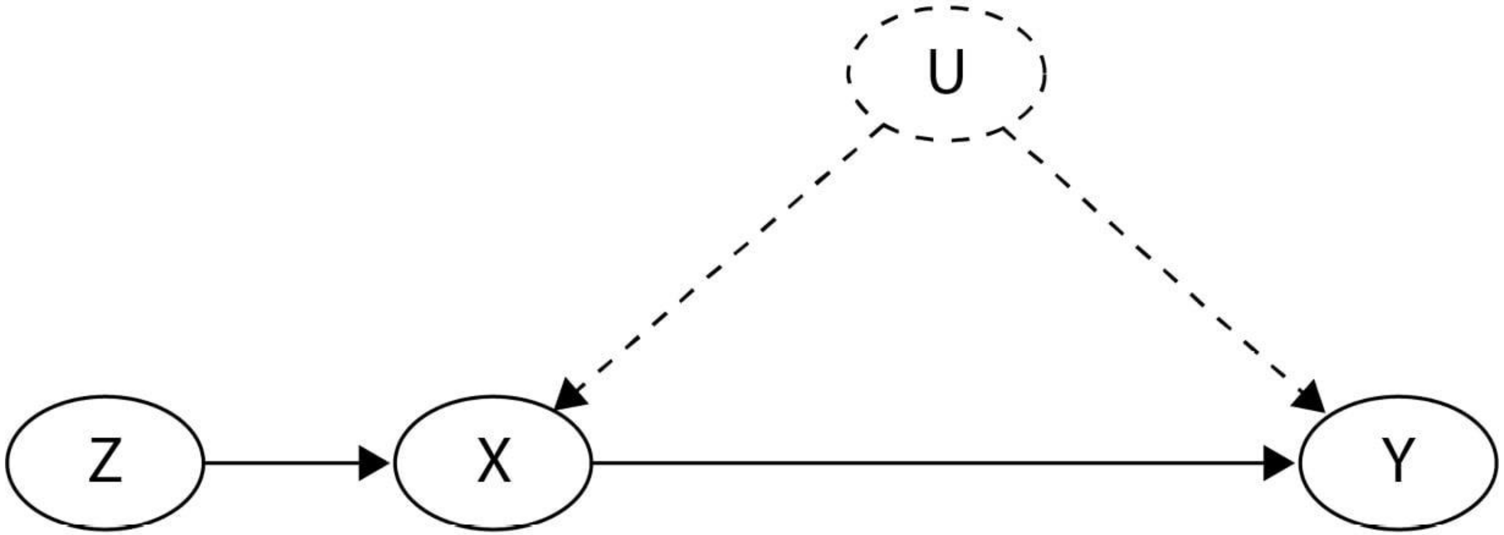
$$P(Y = y|do(X = x), do(Z = z), W = w) = P(Y = y|do(X = x), W = w) \; if\, (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X,Z(W)}}}$$

Given a DAG G, we can say that $G_{\overline{X}}$ is a modification of G, where we removed all the *incoming* edges to the node X. We will call $G_{\underline{X}}$ a modification of G, where we removed all the *outgoing* edges from the node X. For example, will denote a DAG, $G_{\overline{X}\underline{Z}}$, where we removed all the incoming edges to the node X and all the outgoing edges from the node Z.

- **Good news:** do-calculus exists and is complete
- **Not so good news:** it can be quite incomprehensible and takes a while to learn
- **Super good news:** now there are codes (DoWhy) that allow us to apply it

From A. Molak, Causal Inference and Discovery in Python

Instrumental Variables



We're interested in estimating the causal effect of X on Y .

- Cannot use the back-door criterion here because U is unobserved.
- Cannot use the front-door criterion because there's no mediator between X and Y .

Instrumental Variables: require a special variable called an *instrument*, Z, to be present in a graph. An *instrument* needs to meet the following three conditions:

- The instrument, Z, is associated with X
- The instrument, Z, doesn't affect Y in any way except through X
- There are no common causes of Z and Y

From A. Molak, Causal Inference and Discovery in Python

We want to study the effect of education (years of schooling) on earnings. However, the level of education might be influenced by many factors like family background, which also affect earnings. This correlation between the unobserved factors (like family background) and education can bias the results if you simply run a linear regression of earnings on education.

We need an instrument that is correlated with education but does not directly affect earnings except through education. Let's say we choose "proximity to college" as instrument.

Two-Stage Least Squares (2SLS) Regression:

1. Regress the potentially endogenous variable (education) on the instrument (proximity to college). This predicts the values of education that are not influenced by the unobserved confounders.

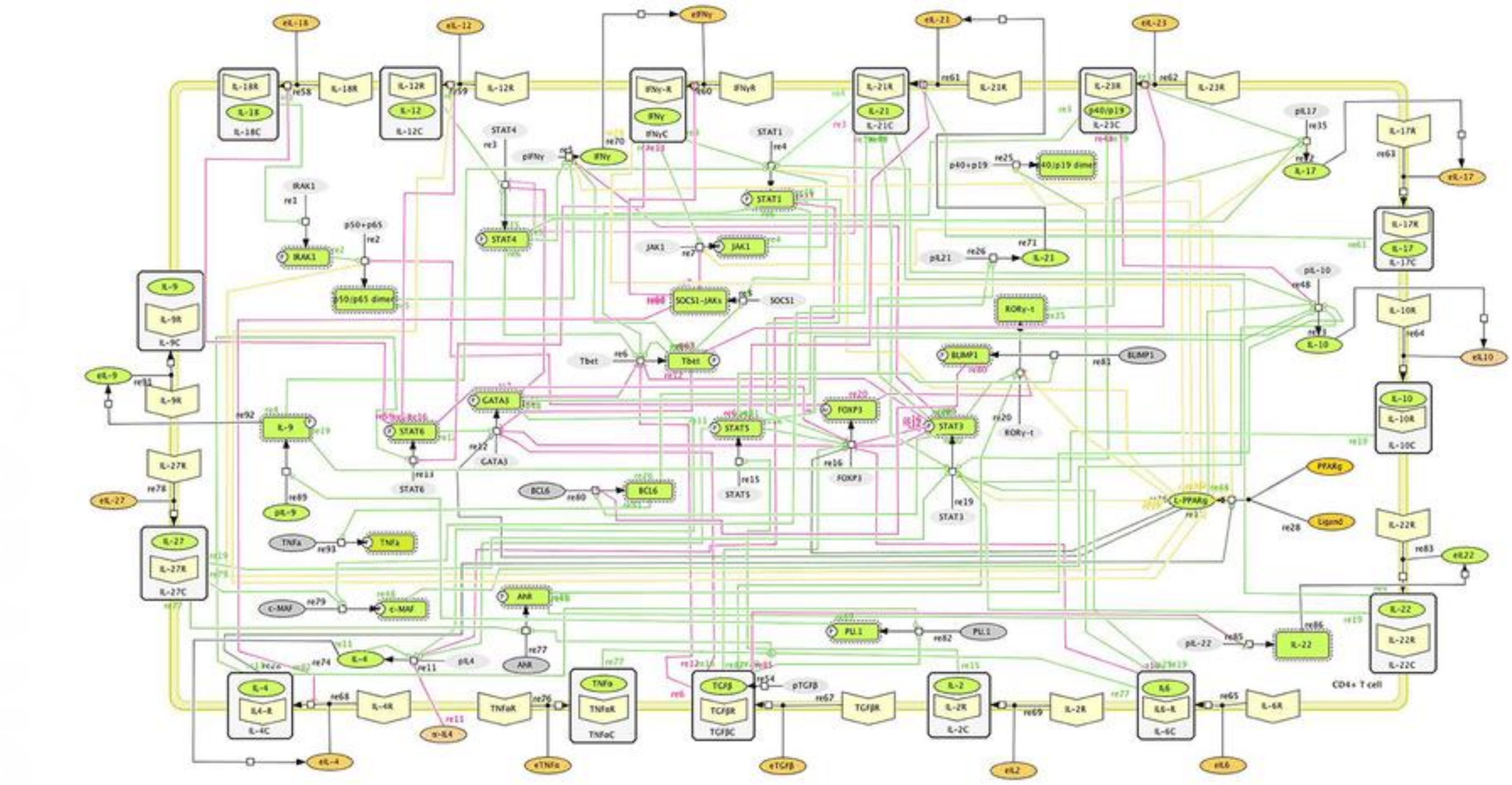
$$\text{Education} = \alpha + \beta * \text{ProximityToCollege} + \varepsilon$$

2. Regress the outcome (earnings) on the predicted values of education from the first stage

$$\text{Earnings} = \gamma + \delta * \text{PredictedEducation} + \zeta$$

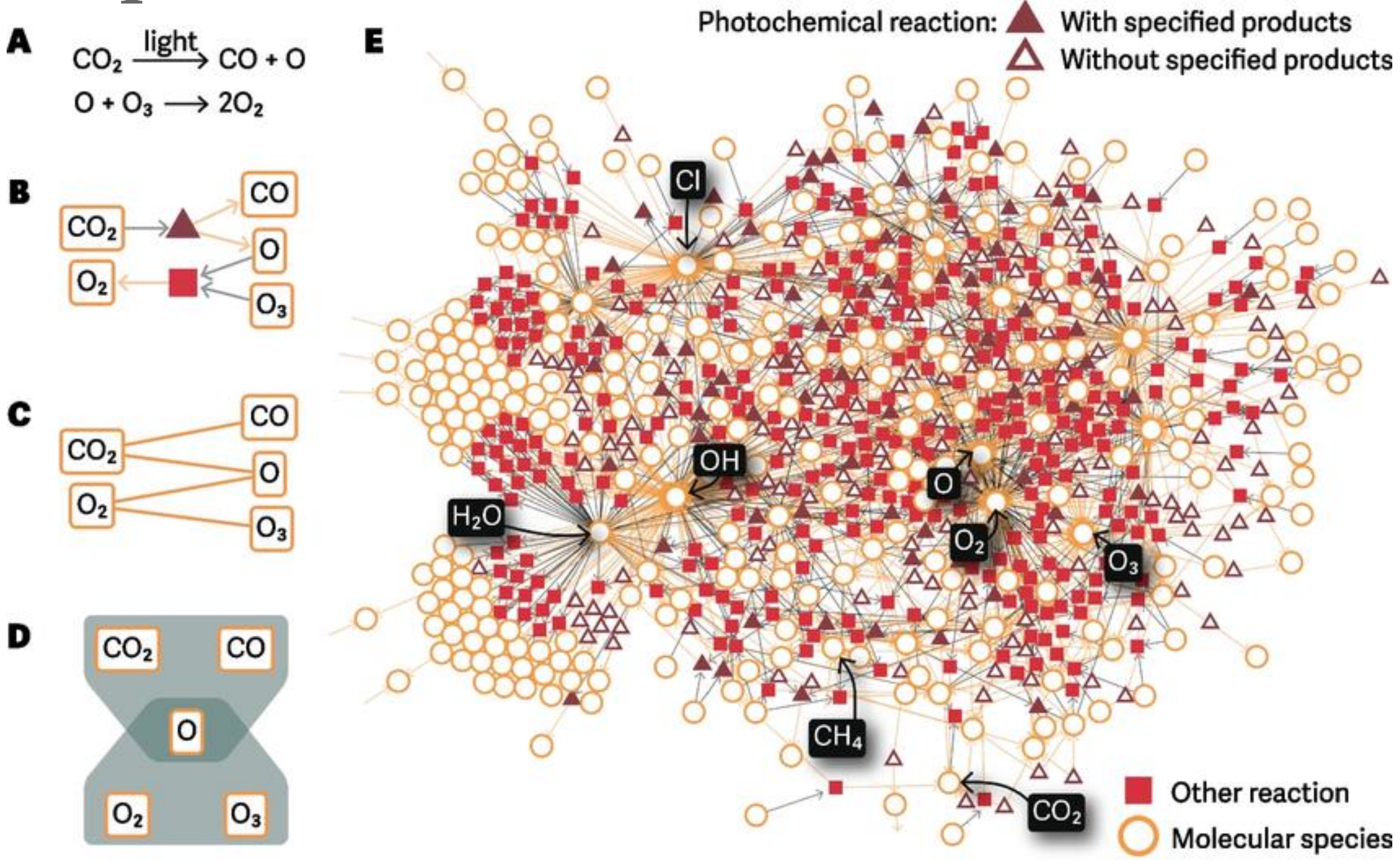
3. The coefficient δ on PredictedEducation in the second stage gives the estimated causal effect of education on earnings. This helps to isolate the variation in education that is independent of the unobserved confounders that also affect earnings.

Biochemical reaction networks



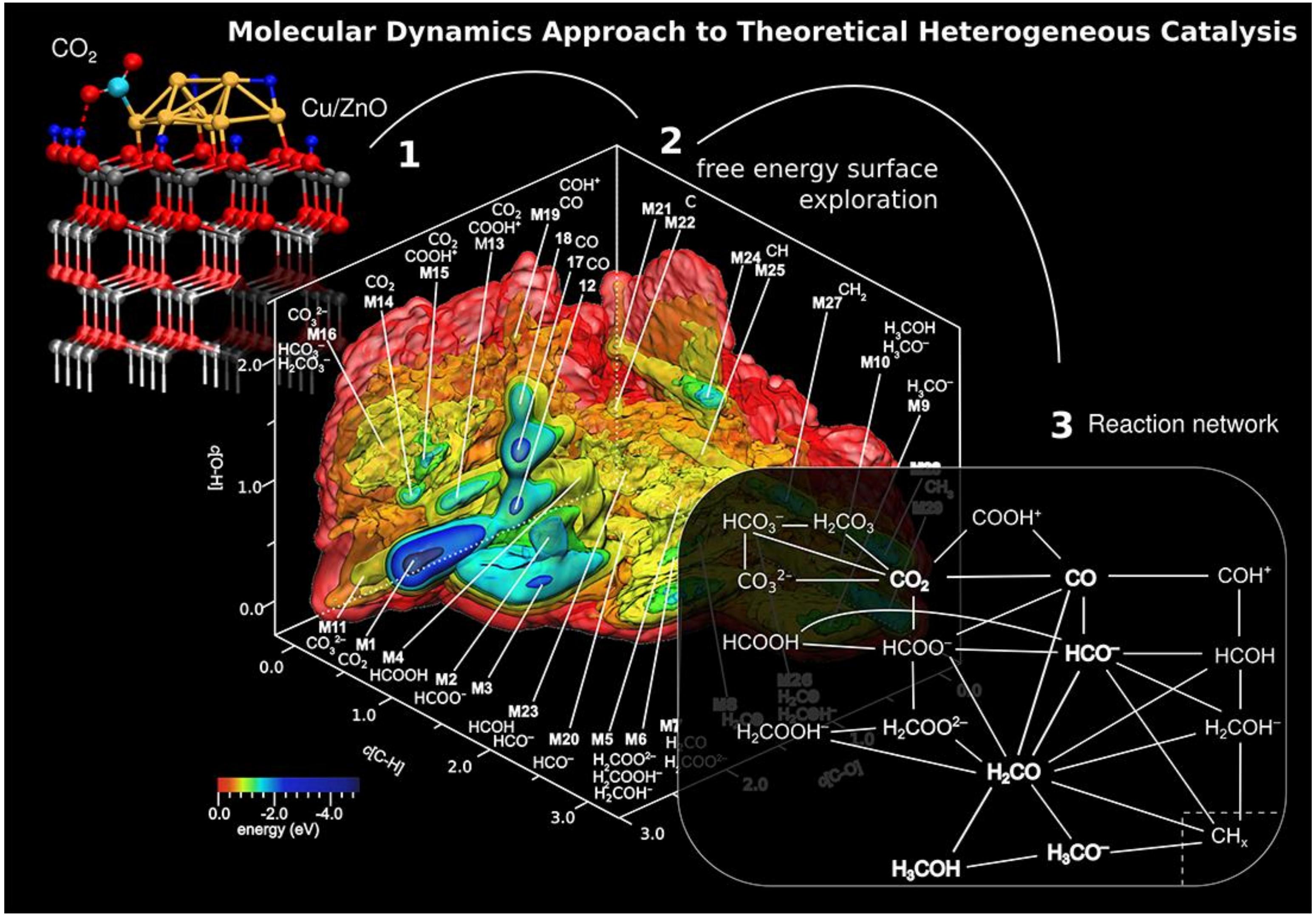
https://www.researchgate.net/figure/Main-intracellular-differentiation-pathways-of-a-single-CD4-T-cell-Systems-Biology_fig2_267753905

Atmospheric reaction networks



https://www.researchgate.net/figure/Network-structure-of-Earths-atmospheric-reaction-system-Panel-A-shows-a-minimal_fig4_368305699

Catalysis



https://www.gauss-centre.eu/results/materials-science-and-chemistry/theoretical-heterogeneous-catalysis-from-advanced-ab-initio-molecular-dynamics-simulations

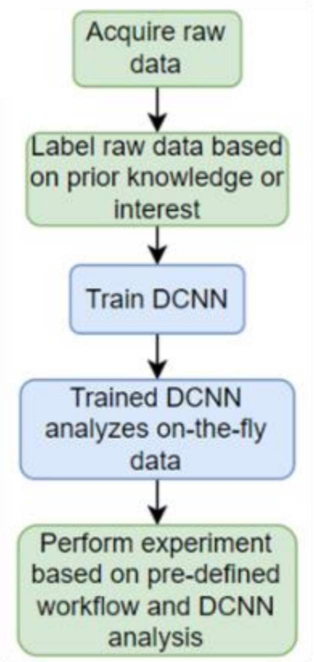
Why causal inference is difficult?

- unmeasured confounders
- measurement error, or discretization of data
- mixtures of different causal structures in the sample
- feedback
- reversibility
- the existence of a number of models that fit the data equally well
- an enormous search space
- low power of tests of independence conditional on large sets of variables
- selection bias
- missing values
- sampling error
- complicated and dense causal relations among sets of variables,
- complicated probability distributions

Orchestrating Multiple Experiments

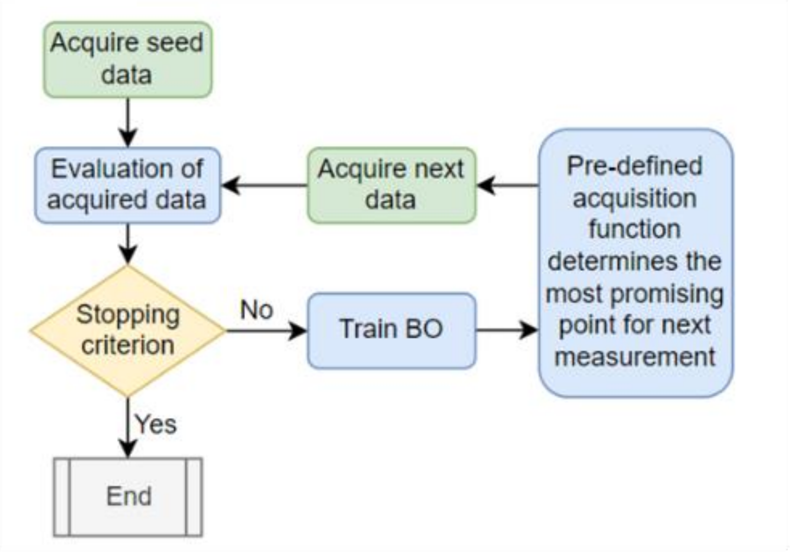
Types of automated experiment

Direct



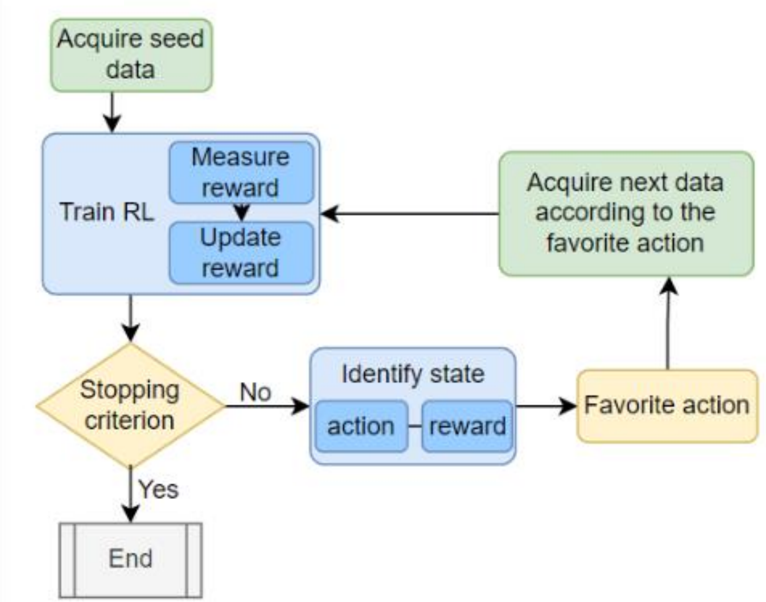
- Fixed policies
- Need DCNNs stable wrt. out of distribution shift

Myopic discovery



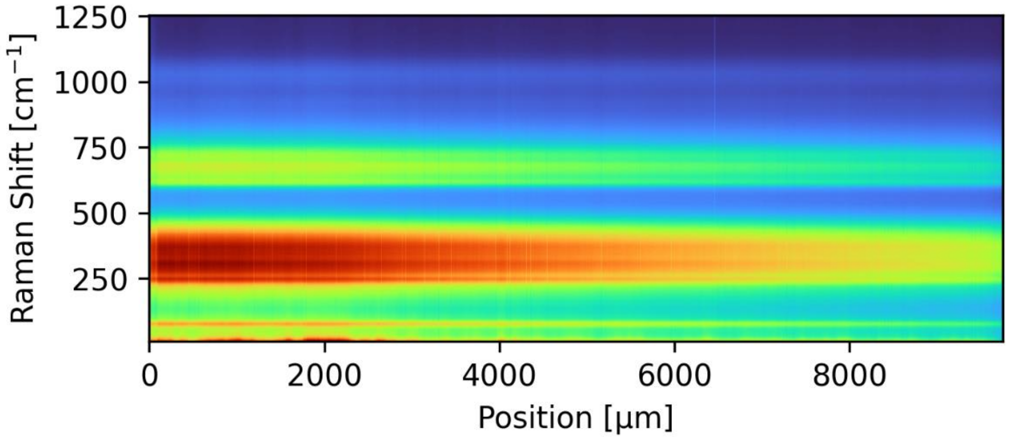
- Adjustable policies
- One step planning
- Can be implemented via Bayesian workflows
- Can be human in the loop

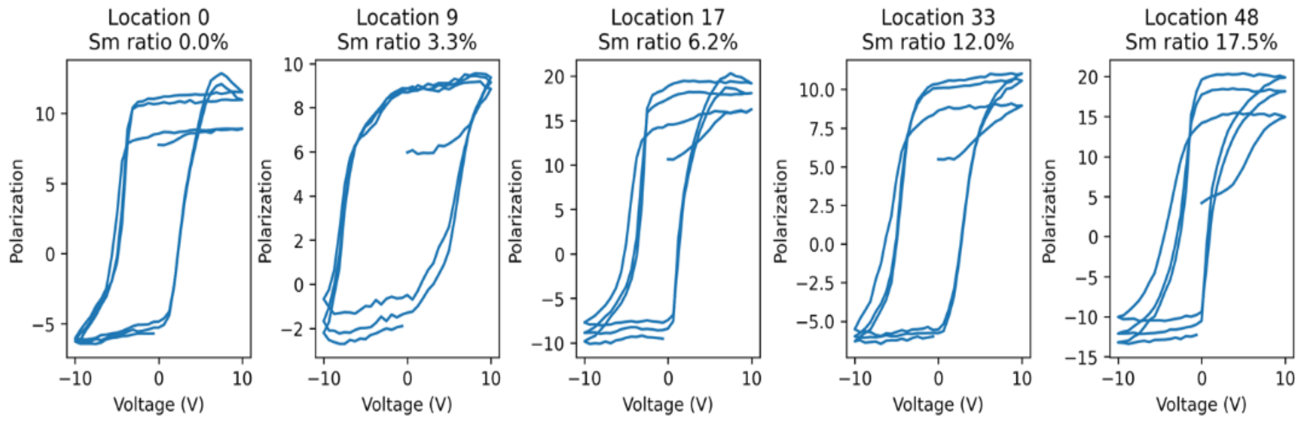
Multistage discovery



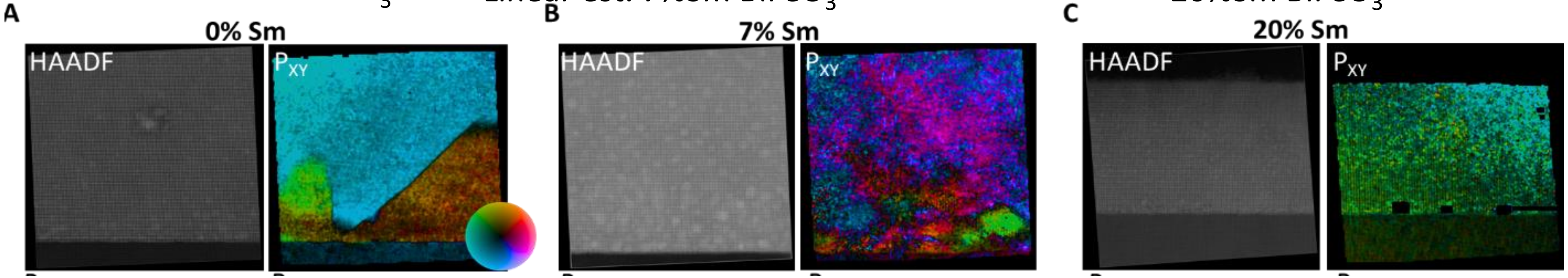
- Adjustable policies
- Multi-step planning
- Requires heuristic to start
- Requires **reward function**

Active Multimodal Experiment



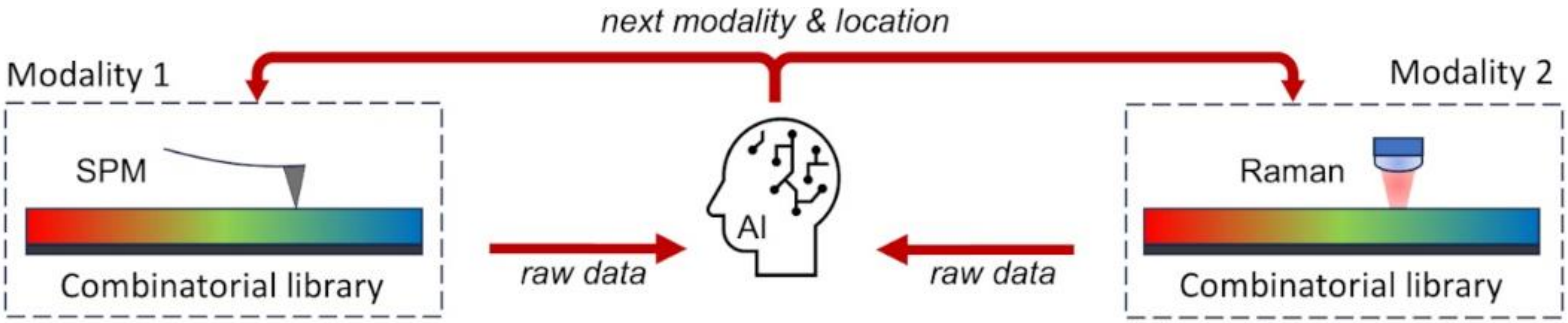


$BiFeO_3$ Linear est. 7%Sm $BiFeO_3$ 20%Sm $BiFeO_3$

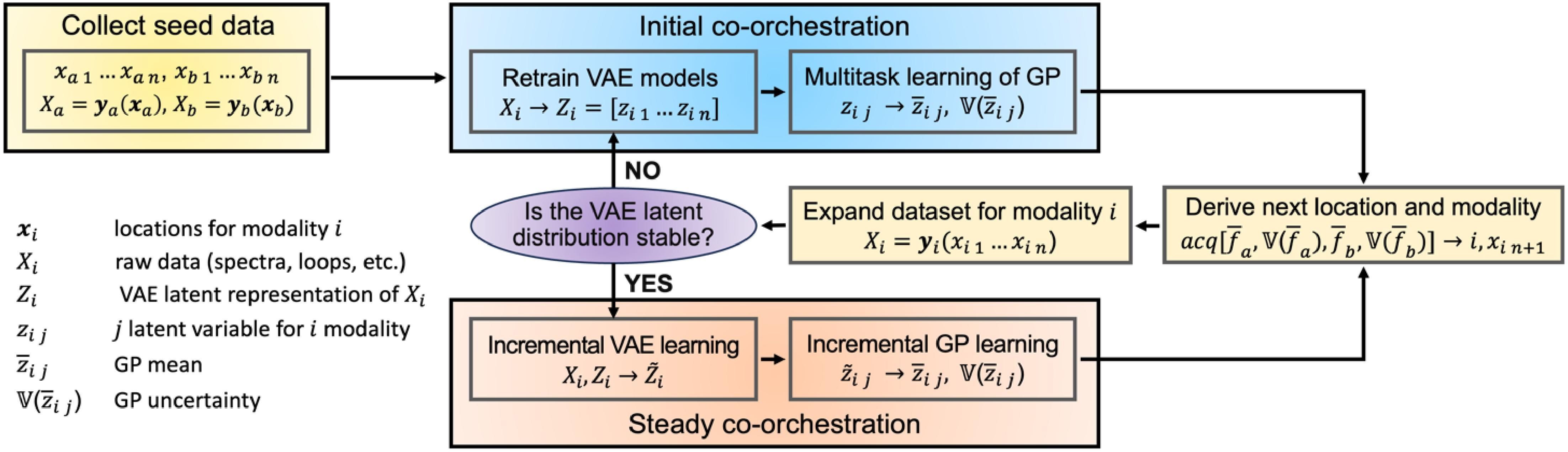


- **Scenario I:** using known proxies to guide active learning of one parameter
- **Scenario II:** orchestrating 2 experiments with complementary information

Co-orchestrating complex observables

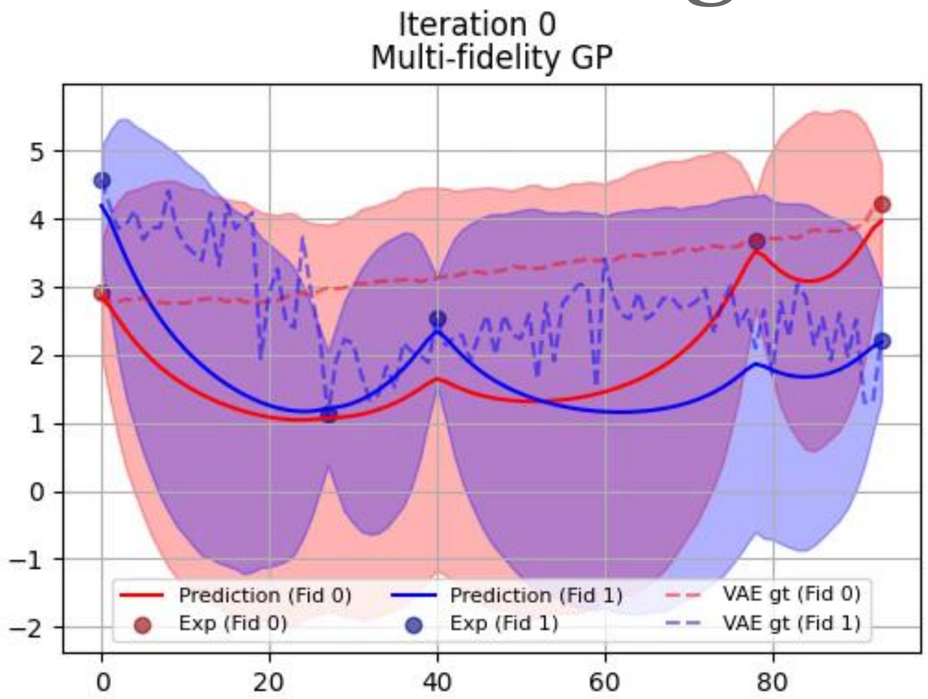


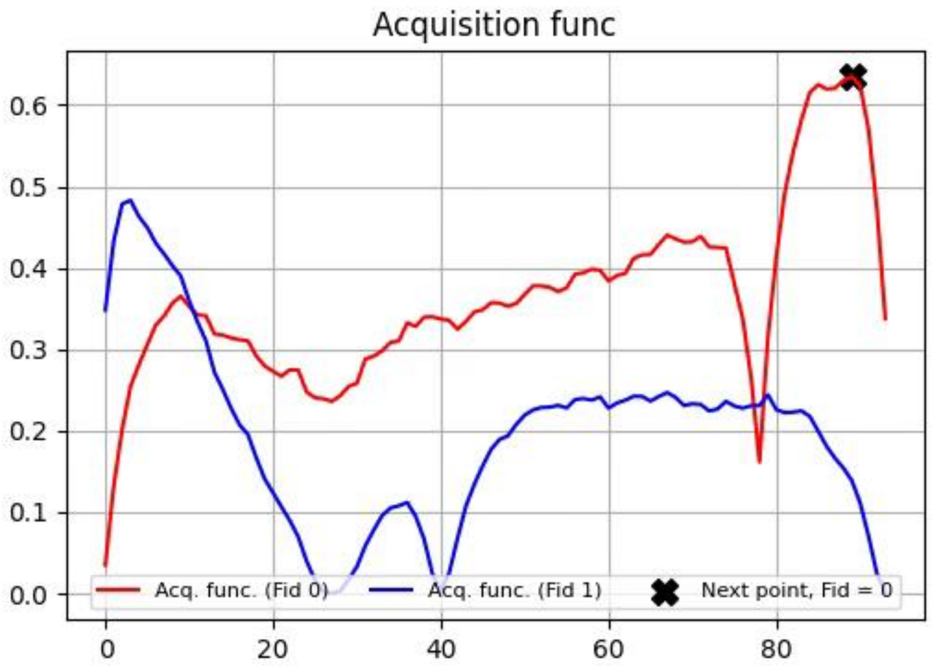
Co-orchestrating complex observables

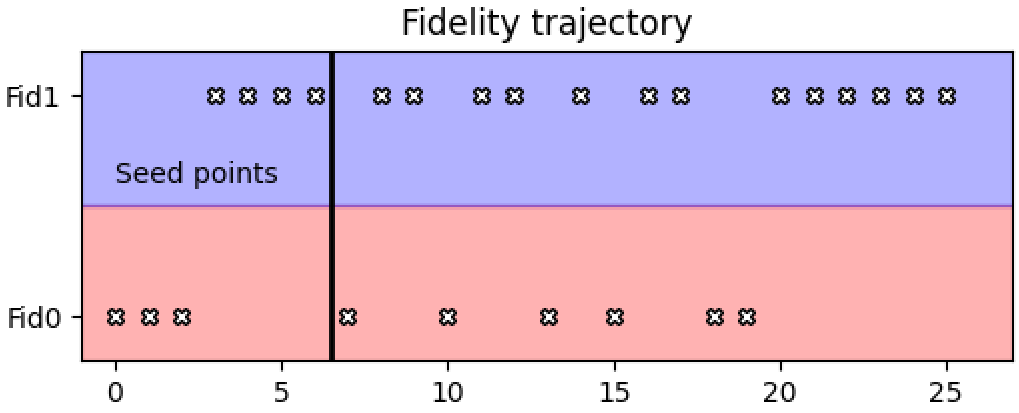


- Workflow combining dynamic VAE embeddings with multitask GP or sGP over latent space
- Representation learning control to bias latent variables and ensure VAE convergence
- Here, one channel is **Raman** and another is **PFM hysteresis loops**

Co-orchestrating complex observables





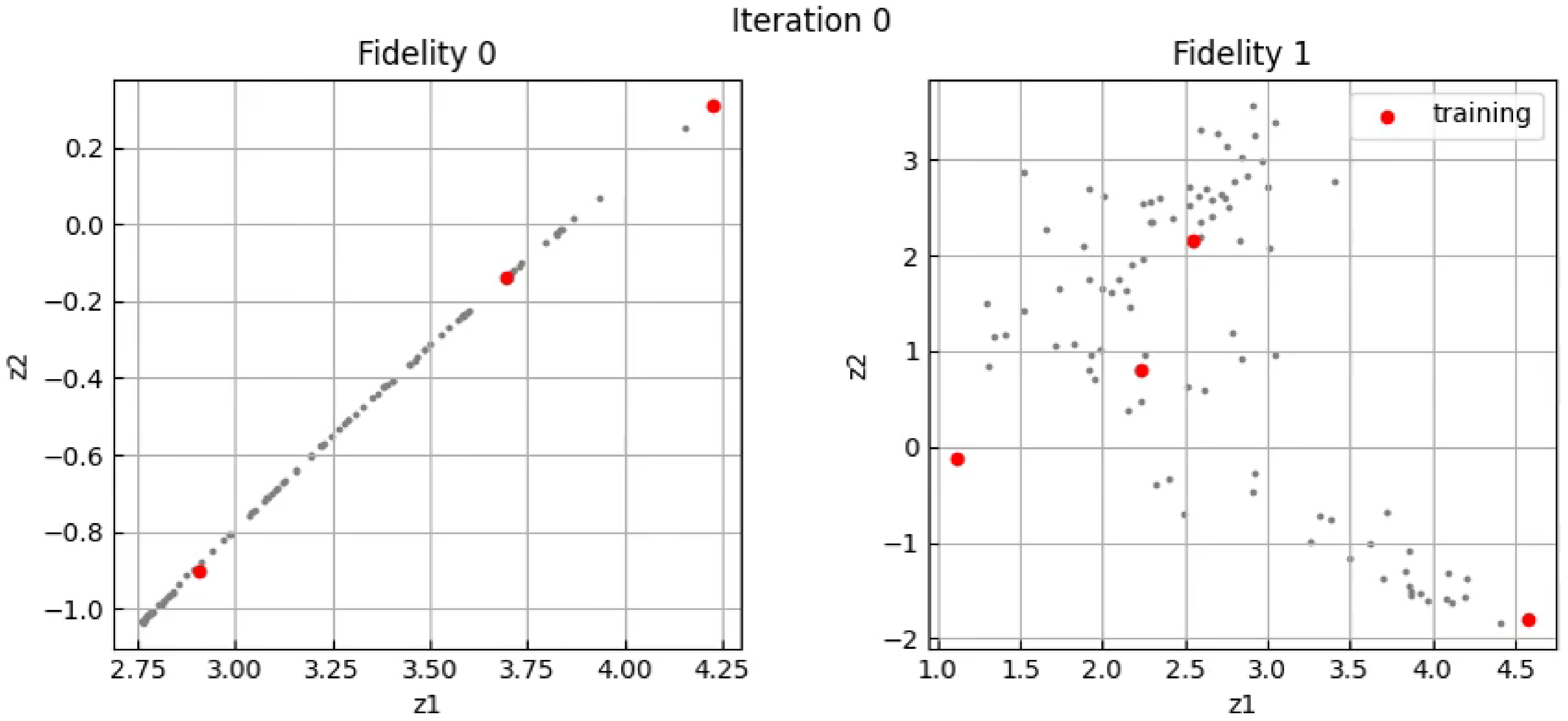


- Co-orchestration of the Raman and PFM experiments on the pre-acquired data

- Learned are:
  - **decoding functions**
  - spatial distribution of **latent variables**,
  - (symmetric) **correlation kernel**

- **Next challenge: causal structure in kernels**

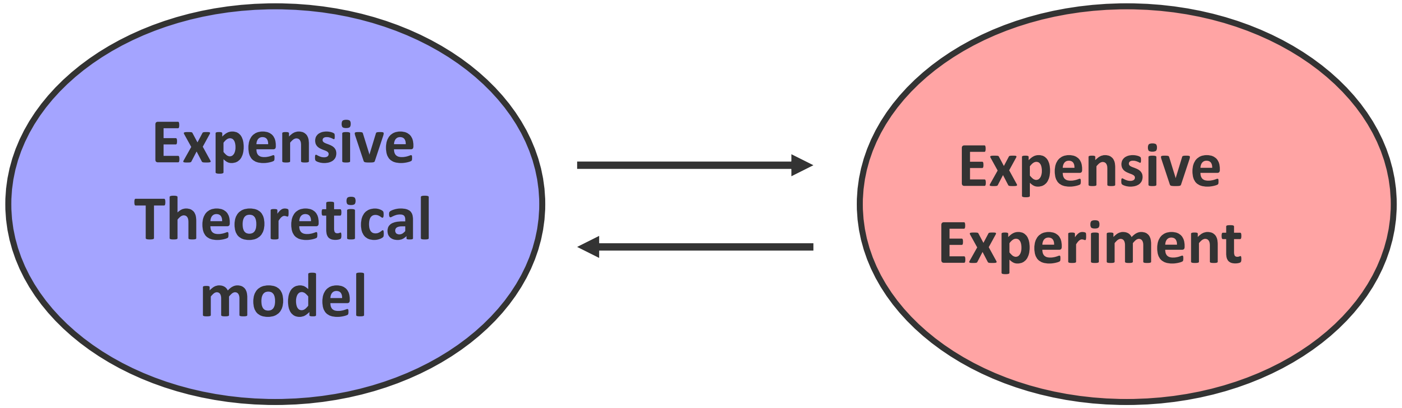
Co-orchestrating complex observables



- Workflow combining dynamic VAE embeddings with multitask GP or sGP over latent space
- Representation learning control to bias latent variables and ensure VAE convergence
- Here, one channel is **Raman** and another is **PFM hysteresis loops**

Theory in the Loop workflows: Bayesian co-Navigation

Next step: co-navigation theory and experiment space via human supervised policies

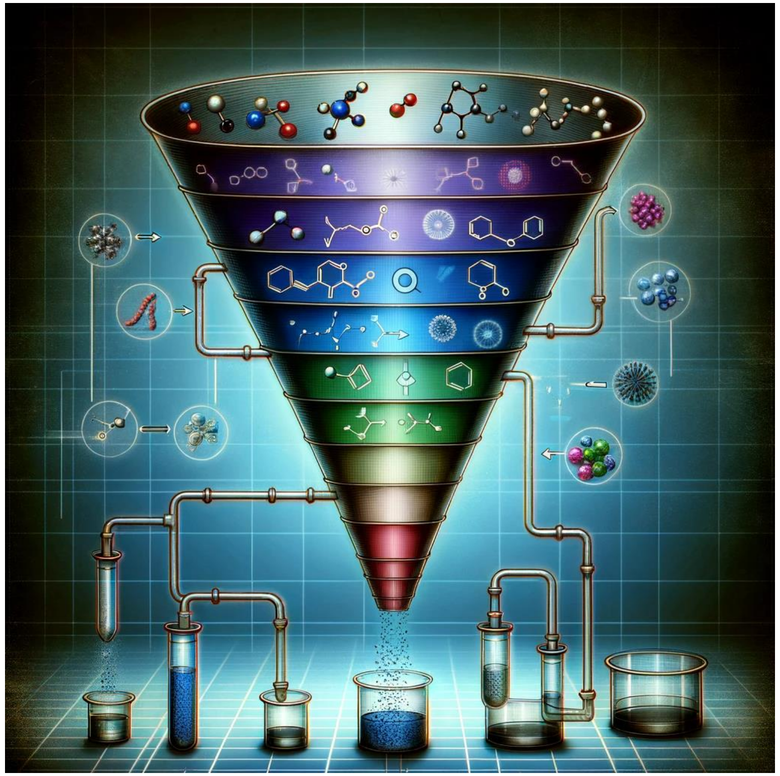


1. Human-in-the-loop automated experiment
2. Policy tuning:
   - Exploration-exploitation balance
   - Fidelity of theory
   - Local physical model

Theory-experiment matching

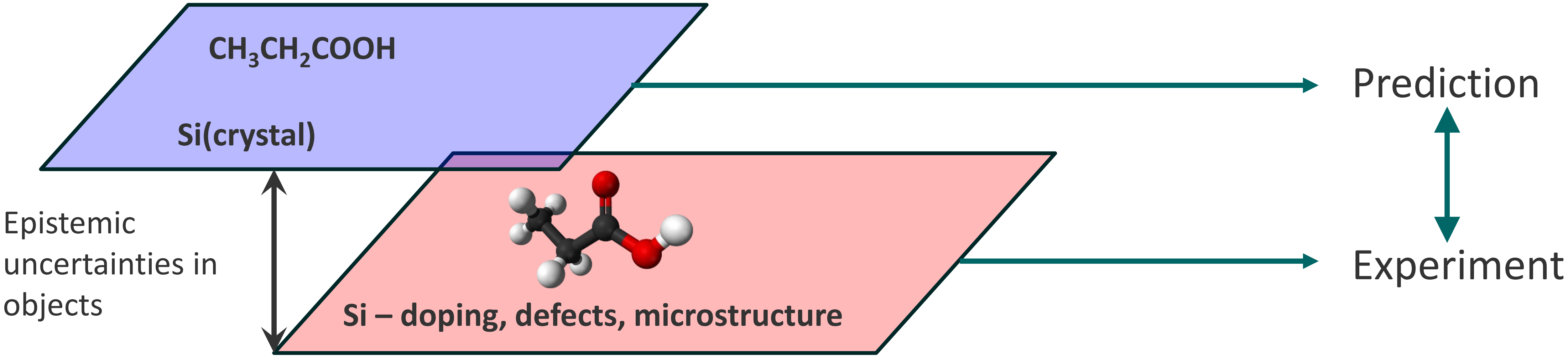
Theory in the loop in SDL requires feedback to theory!

- Local fast models
- Potentially informed by immutable oracle (e.g. DFT)
- Long term: multiple models can be used to update oracle



Building theory in the loop requires identifying:

- Model structure and parameter space
- Object spaces in theory and experiment: range and epistemic mismatch
- Measures of theory-experiment matching



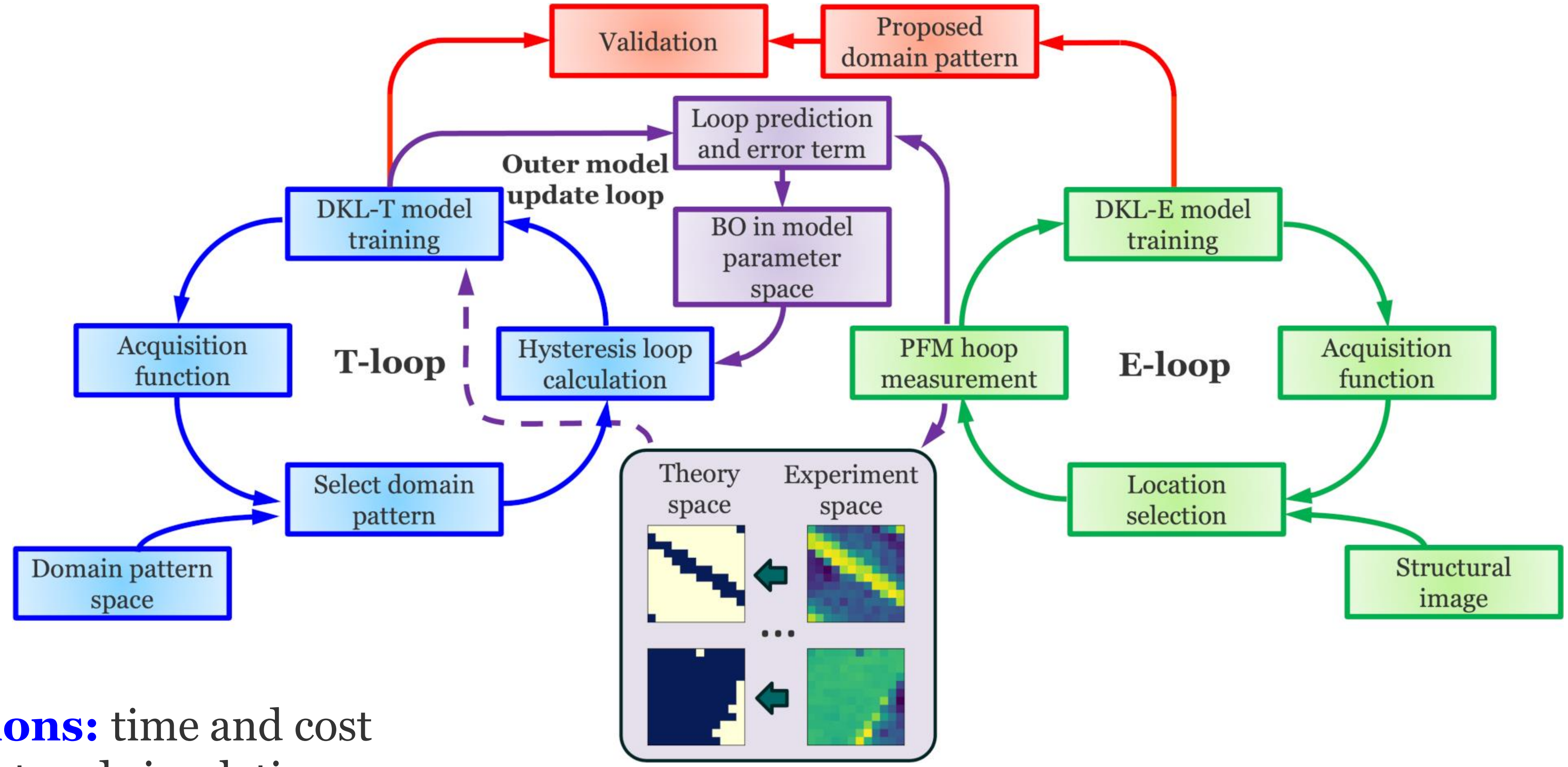
Co-navigation for theory-experiment matching

Problem: adopting a theoretical model to reduce the disparity with observed experimental outcomes can demand significant resources and time.

Idea: integrate theoretical and experiment discovery into the active learning workflow to *in-situ* adjust the theoretical model based on feedback from the experiment to describe the specific experimental system

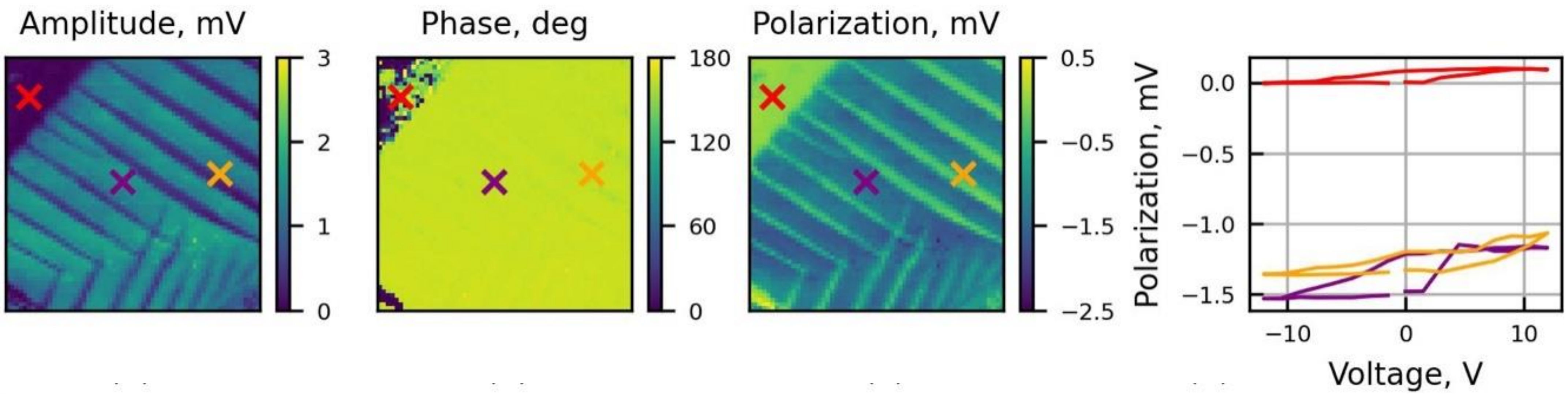
Co-navigation for theory-experiment matching

- Continuous training of the surrogate model for theory (T-loop)
- Co-training surrogate model for experiment (E-loop)
- Theory update based on error function over support of interest (Outer theory update loop)



- **Cost functions:** time and cost of experiment and simulation
- **Gain:** predictive uncertainty

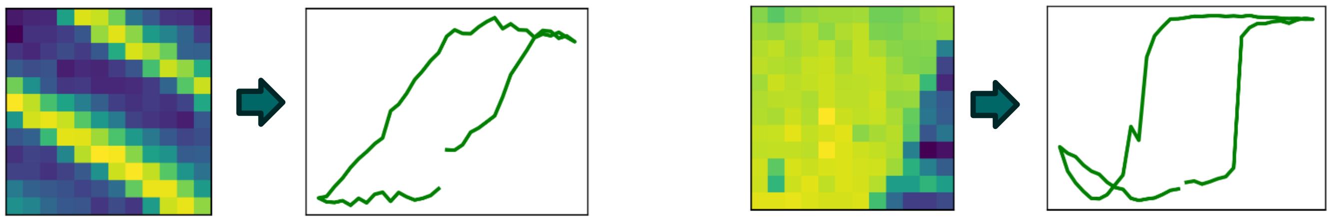
Experiment: BEPS hysteresis loops



Sample
- Chemical vapor deposition
- **PbTiO3** film on KTaO3
- Conductive SrRuO3 sublayer

Domain structure
- In-plane *a*-domains
- Out-of-plane *c*-domains
- Ferroelastic domain walls

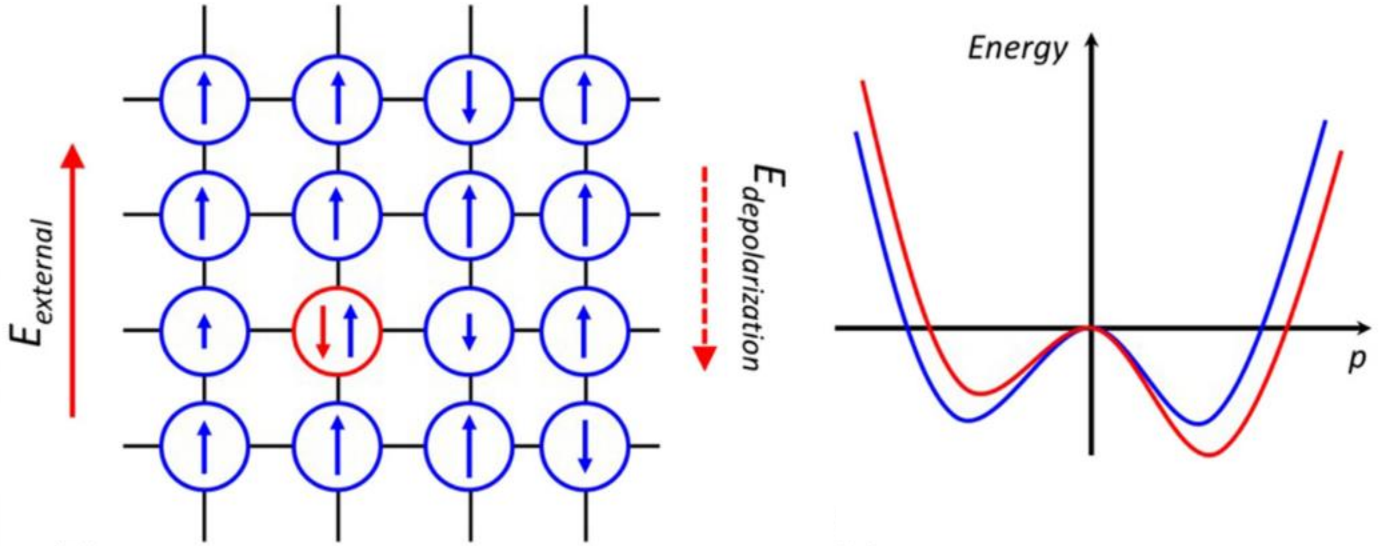
- Experimental object space consists of the local domain arrangement pattern from the structural image



- Surrogate DKL model is training to predict the loop by the domain arrangement in vicinity of the tip

Theory: FerroSim spin-lattice model

Ferroelectric material is represented as a 2D lattice comprising local polarizations p at each site.



*Kalinin, S. V., et al., Journal of Applied Physics, 128(2), 2020.

Total free energy:

$$F = \sum_{ij}\left[F_{ij} + K\sum_{k,l}\left(p_{i,j} - p_{i+k,j+l}\right)^2\right]$$

Dynamics of spins:

$$\gamma\frac{dp_{ij}}{dt} = -\frac{\partial F}{p_{ij}}$$

γ relates to the domain wall motion rates

Local free energy in GLD form:

$$F_{ij} = \frac{\alpha}{2}p_{ij}^2 + \frac{\beta}{4}p_{ij}^4 - E_{loc}(i,j)\,p_{ij}$$

Local electric field:

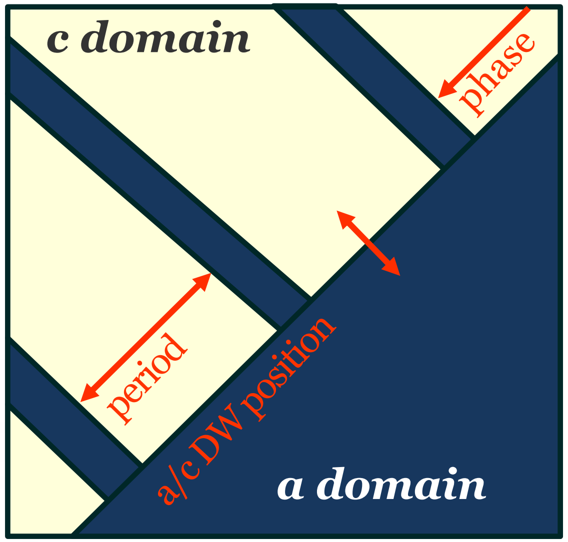
$$E_{loc}(i,j) = E_{ext} + E_{dep} + E_d(i,j)$$

- Applied external field: E_{ext}
- Depolarization: $E_{dep} = -\alpha_{dep} < p >$
- Defect field: $E_d(i,j)$

Model Hyperparameters:

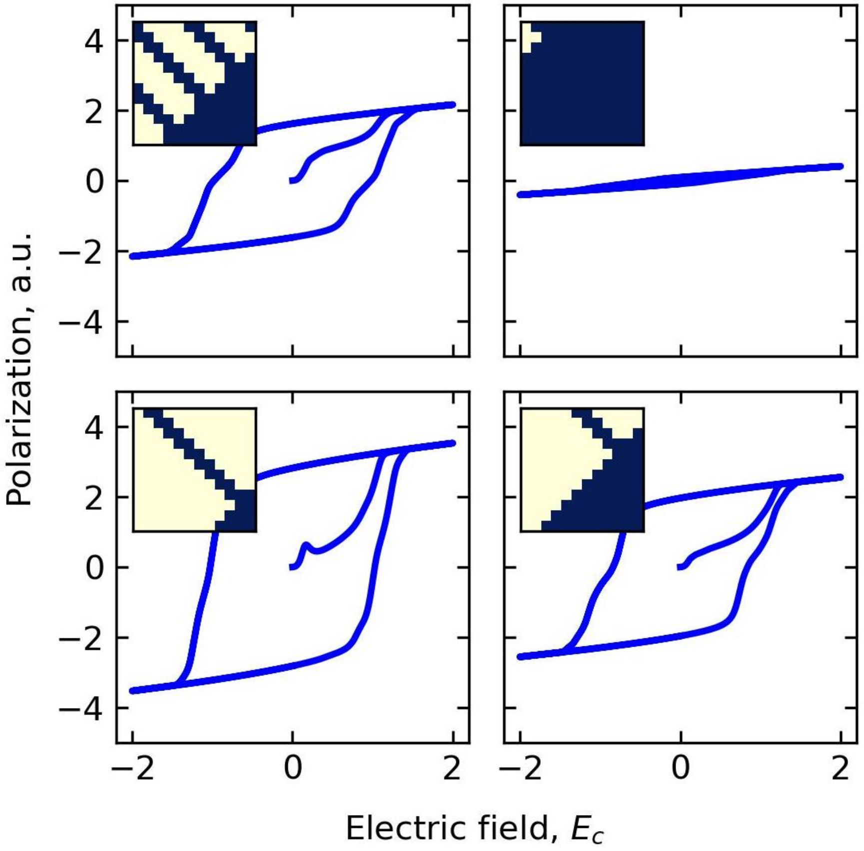
- K – nearest neighbor coupling coefficient
- α_d - depolarization constant

Theory: Theoretical object space



	c-domain	a-domain
E_d	(0,0)	$(0, 30E_c)$
K	hyperparameter	0
E_{ext}	$(2E_c, 0)$	
α_d	hyperparameter	
γ	100	

E_c – coercitive field



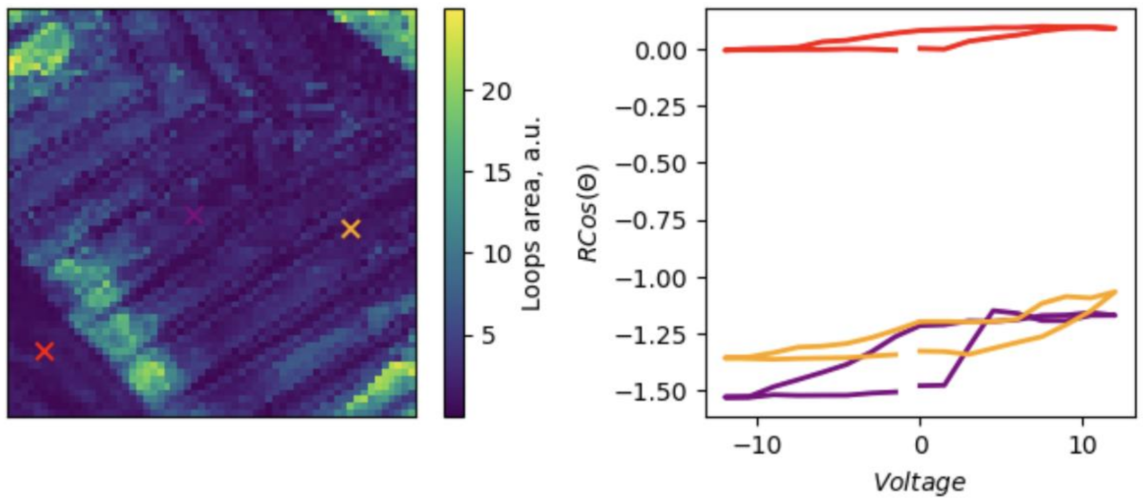
- **Lattice:** 12x12
- **Parametrization:**
    - DW position
    - Period of striped structure
    - Phase of striped structure

Outer theory update loop

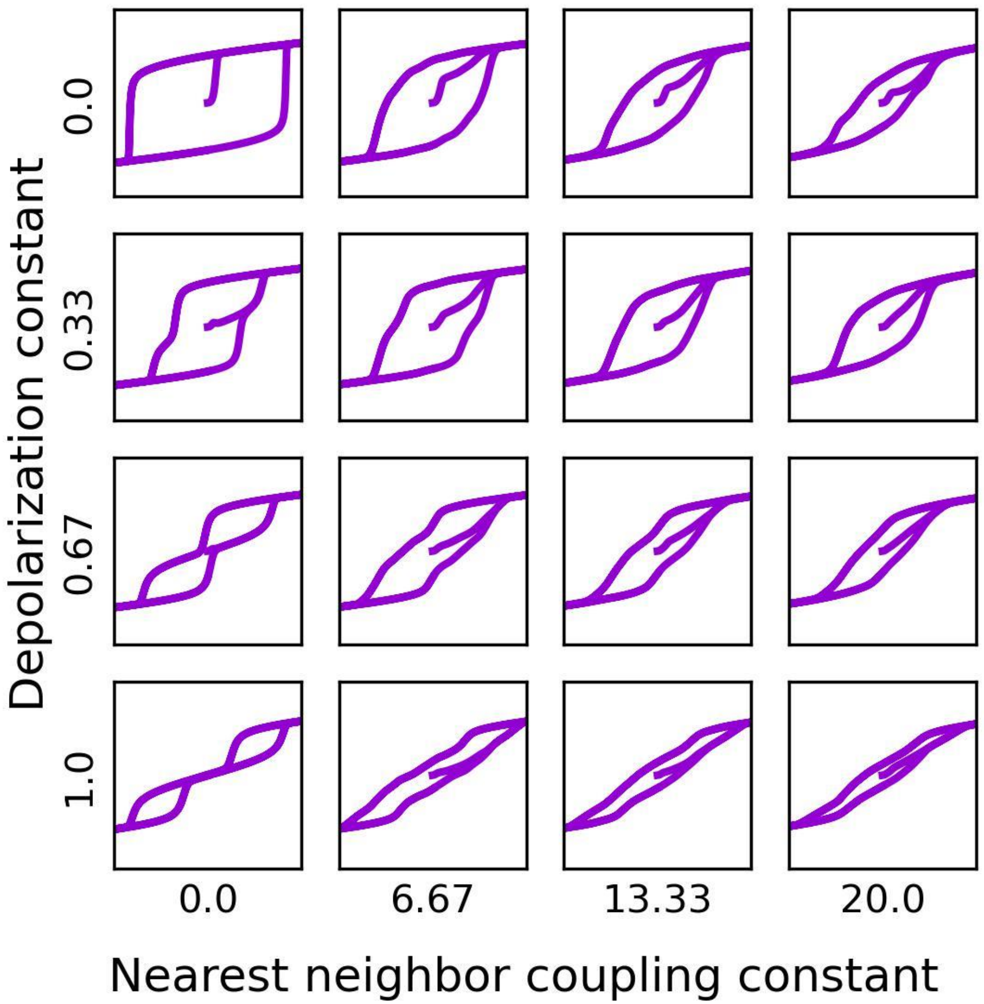
Strives to converge towards optimal hyperparameter configuration to minimize the disparity between experimental and theoretical outcomes.

- Minimization of the MSE between areas of measured and predicted loops
- UCB acquisition function

Experimental ground truth



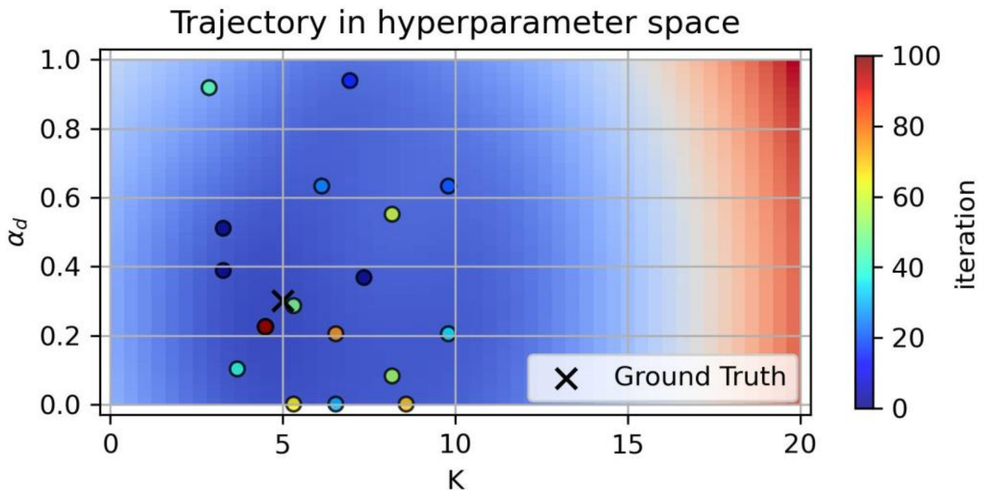
Object space

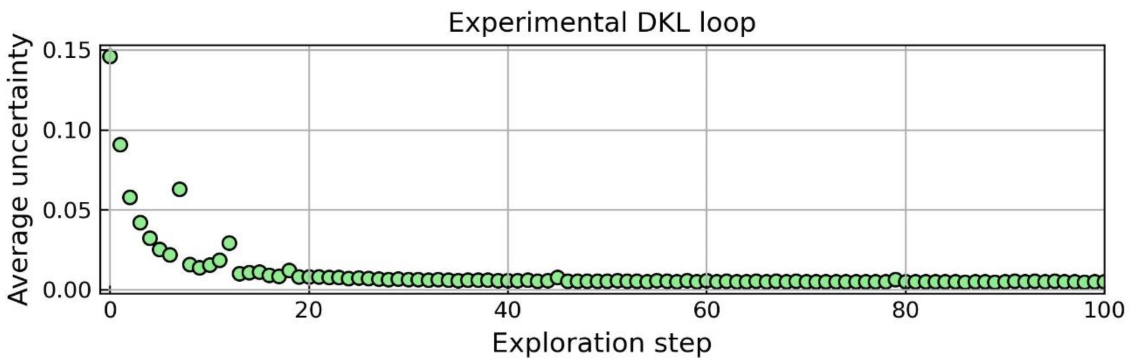


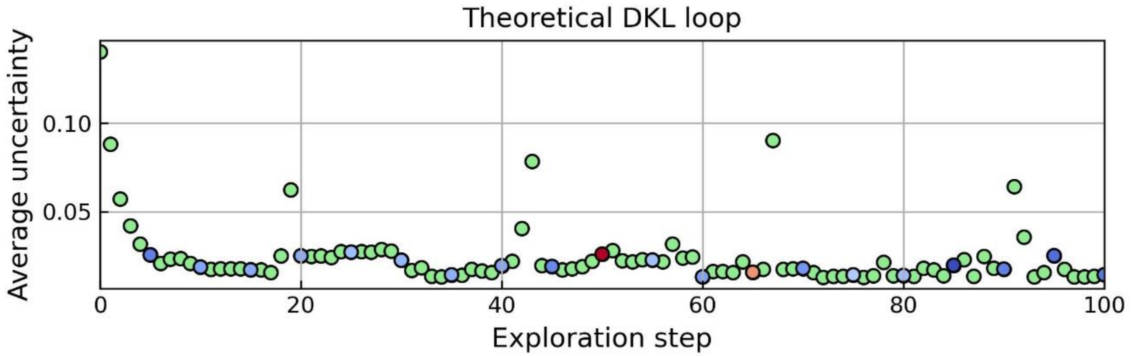
Theory-theory

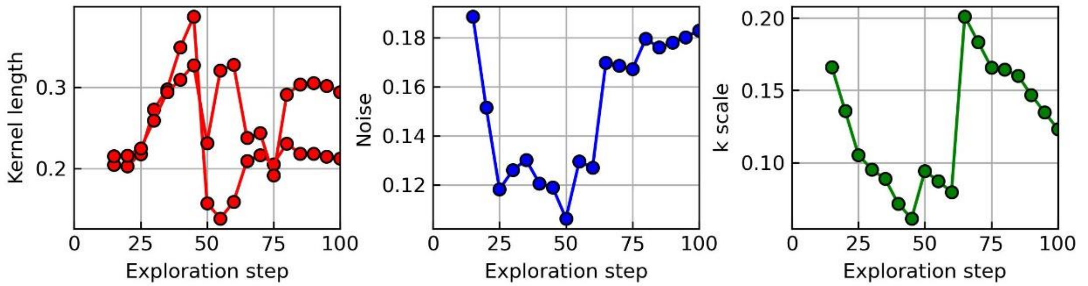
Priors:

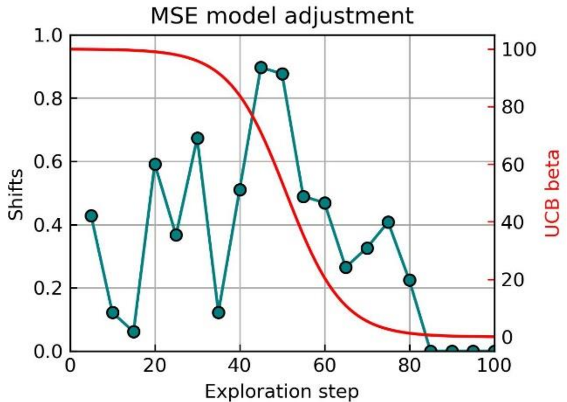
- N iterations = 101
- N_{th} seeds = 5
- N_{exp} seeds = 10
- DKL theory tail =15
- Exp/Theory ratio = 1
- Mean function = max(MSE)
- UCB kernel length = Gamma(2,5)
- UCB beta decay = 100 - 0.01 (sigmoid)
- Mean function $= \alpha(x - x_0) + \beta(y - y_0)$





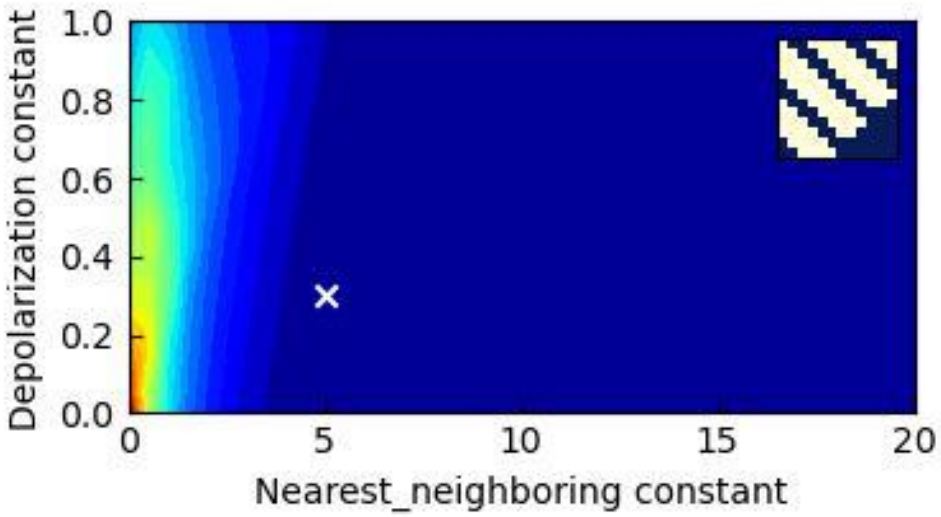


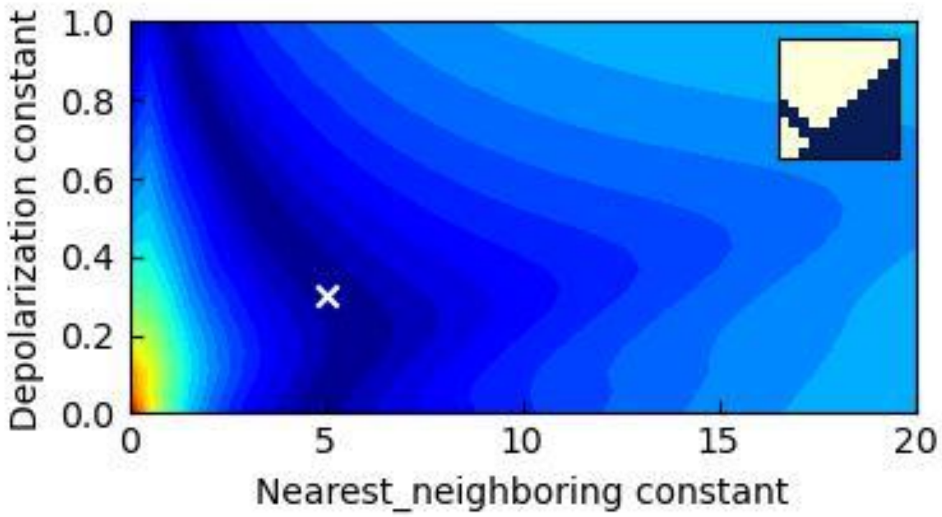




Ground truth

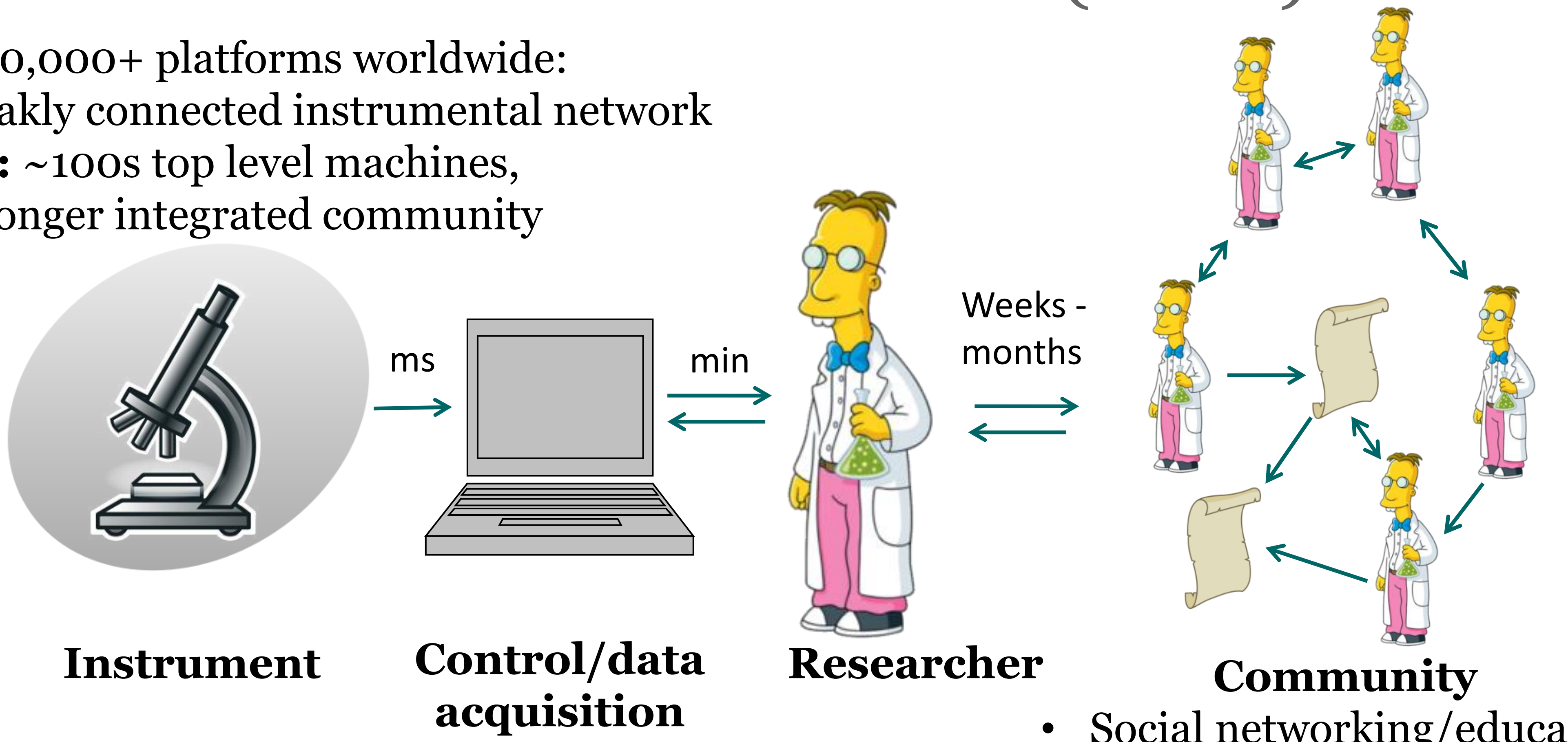
MAE for different domain patterns





Classical Instrumental Research (2016)

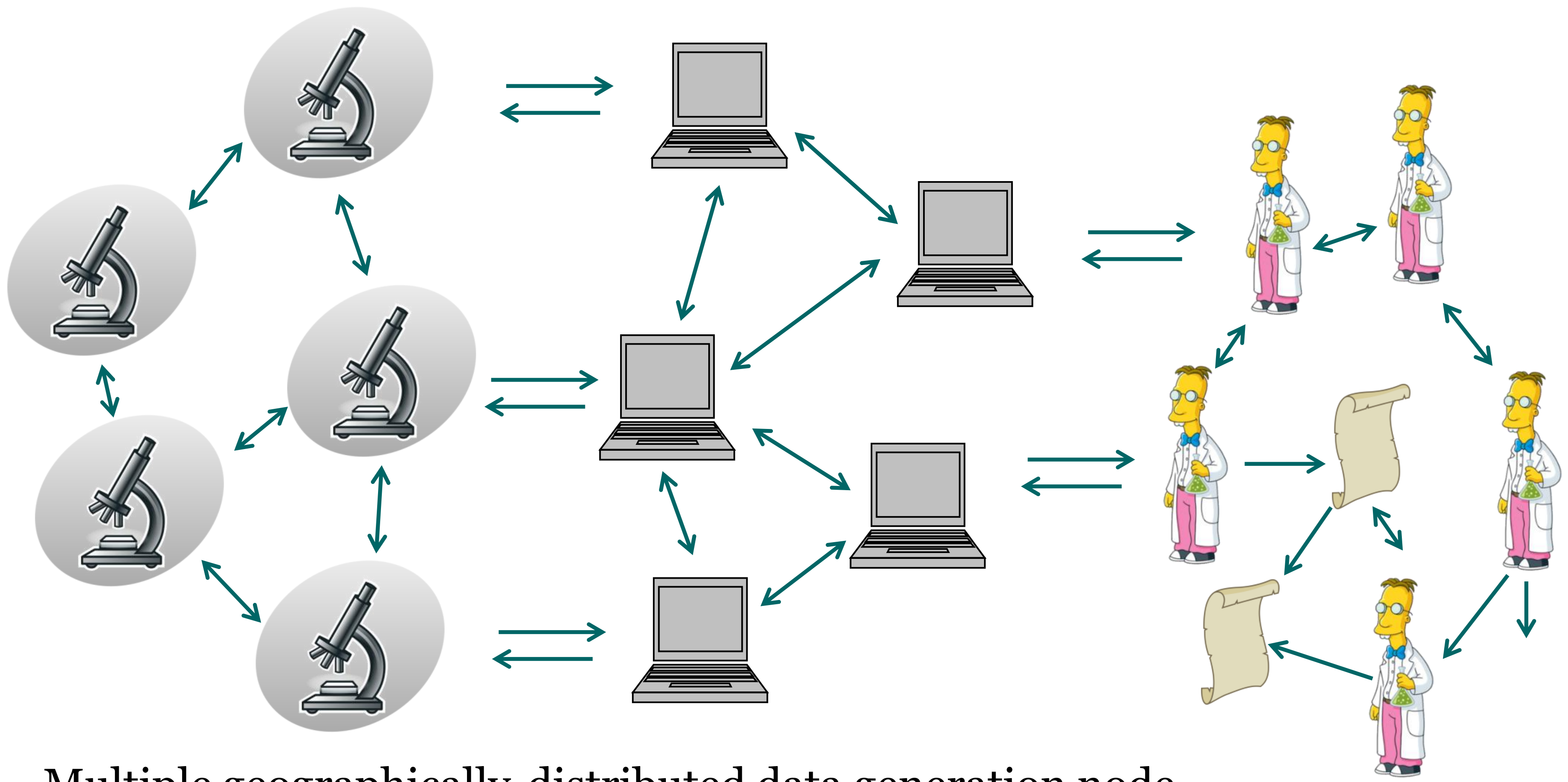
SPM: 100,000+ platforms worldwide:
Large weakly connected instrumental network
(S)TEM: ~100s top level machines,
much stronger integrated community



- Social networking/education
- Publications/citations

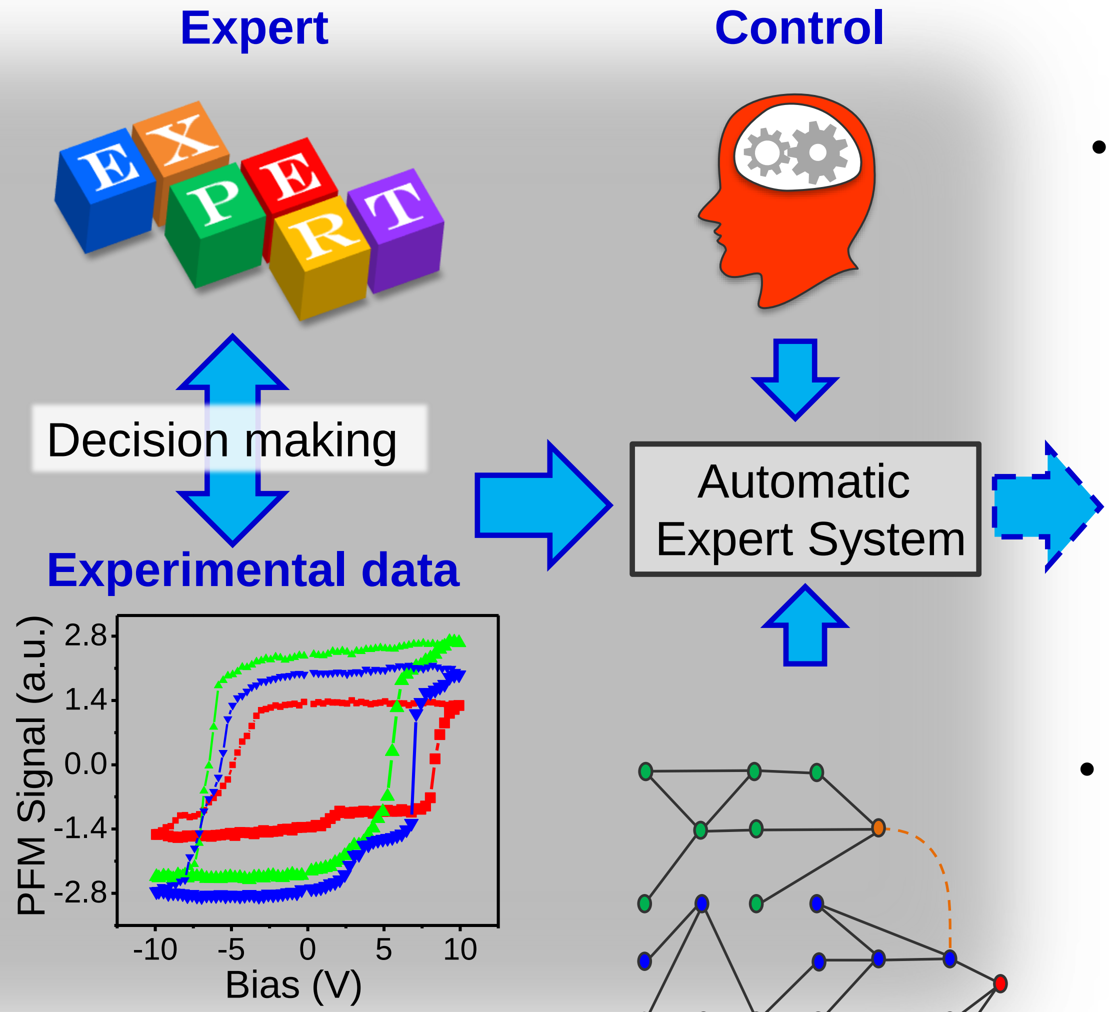
1. Only small fraction of data stream from the instrumentation is captured
2. Only small fraction of captured data is analyzed, interpreted, and put in the context
3. Human-machine interaction during acquisition is often slow and can be non-optimal
4. Human interpretation of data is limited: bias and ignoring serendipity
5. Information propagation and concept evolution in scientific community is slow

Step 1: Cloud Integration (2016)

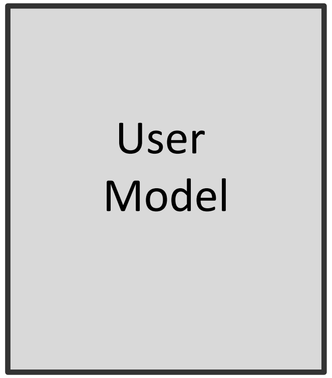


1. Multiple geographically-distributed data generation node
2. Full capture of instrumental data stream /compression/curation
3. Coordination of protocols and data/metadata across the cloud
4. Cloud-based processing and dimensionality reduction
5. Community-wide analytics

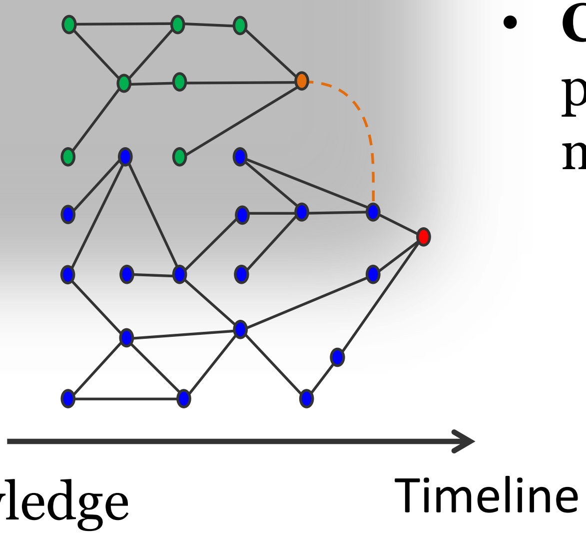
Step 2: Cloud Analytics (2016)



- **Synthesis of expertise:** factor in human expert knowledge



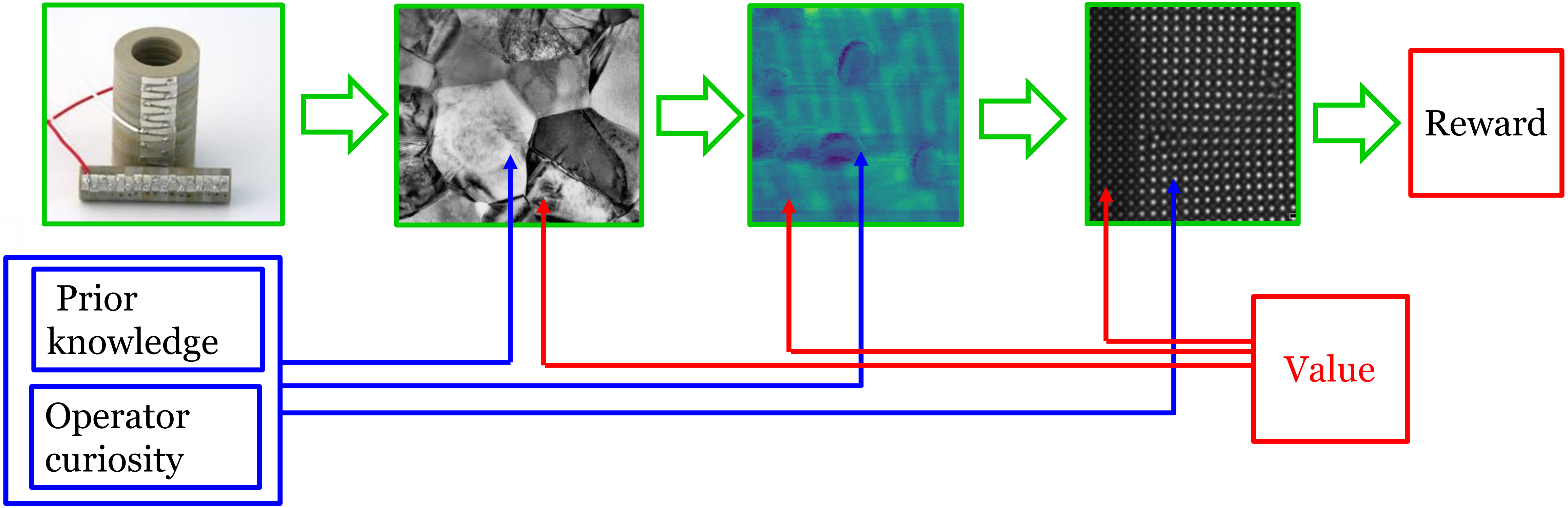
- **Context search:** published results data mining/social networks



Future:

- Automated analysis of routine data
- Identification of anomalies
- Initial training of new practitioners
- Data centers: information based on knowledge

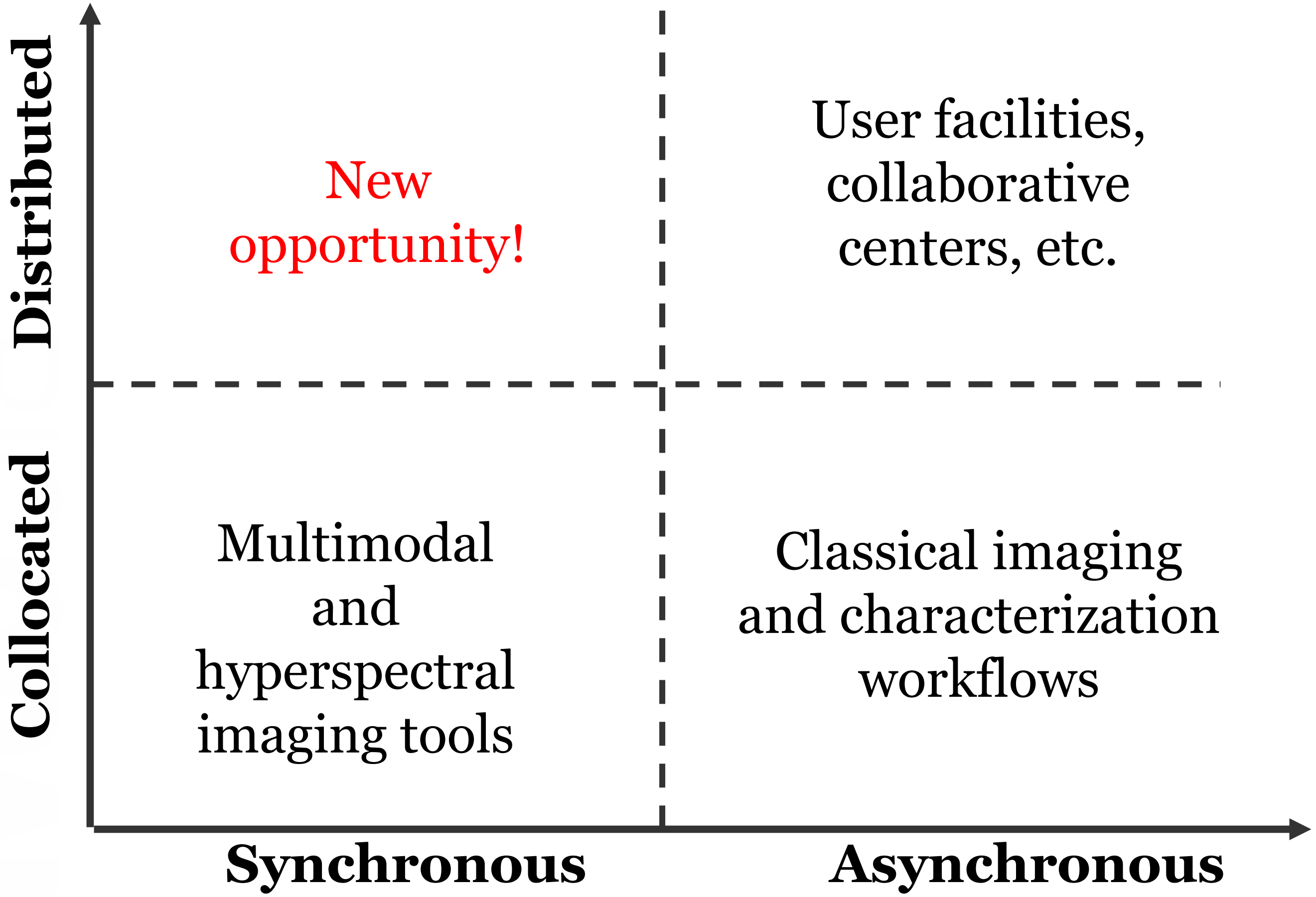
Step 3: Workflow Design (2022)



Traditional experiment:
1. Always based on workflows
2. Ideated, orchestrated, and implemented by humans
3. The “gain of value” during the workflow implementation is uncertain

Value of the step is key element:
- Either based on prior knowledge
- Or defined in a sense of the reinforcement learning Q-function

Cloud Labs: Facilities of the Future



Emerald Cloud Lab, SF and CMU

1. Combined human-machine workflow implementation
2. Computer orchestrating agent
3. How would beyond human workflows be ideated?

Reward Driven Workflow Design

1. Development of the labs capable of orchestrating predefined workflows based on human and robotic agents.

2. Workflow design based on AI and human decision making, meaning specific series of synthesis and characterization steps described via executable hyperlanguage.

3. Defining domain-specific reward functions. Why are we running experiments? Ultimately, we need to quantify (in the style of Bell's equation) what is the benefit of the specific step in the workflow, and how does it accomplish or affects exploration and exploitation goals.

4. Integration of reward functions from dissimilar domains. For example, how does better microscope help us learn physics of specific material? Why would the specific DFT calculation help us understand experimental data?

5. Creating experimentally falsifiable hypothesis from the domain specific body of knowledge that can be incorporated in the exploratory part of automated workflows.

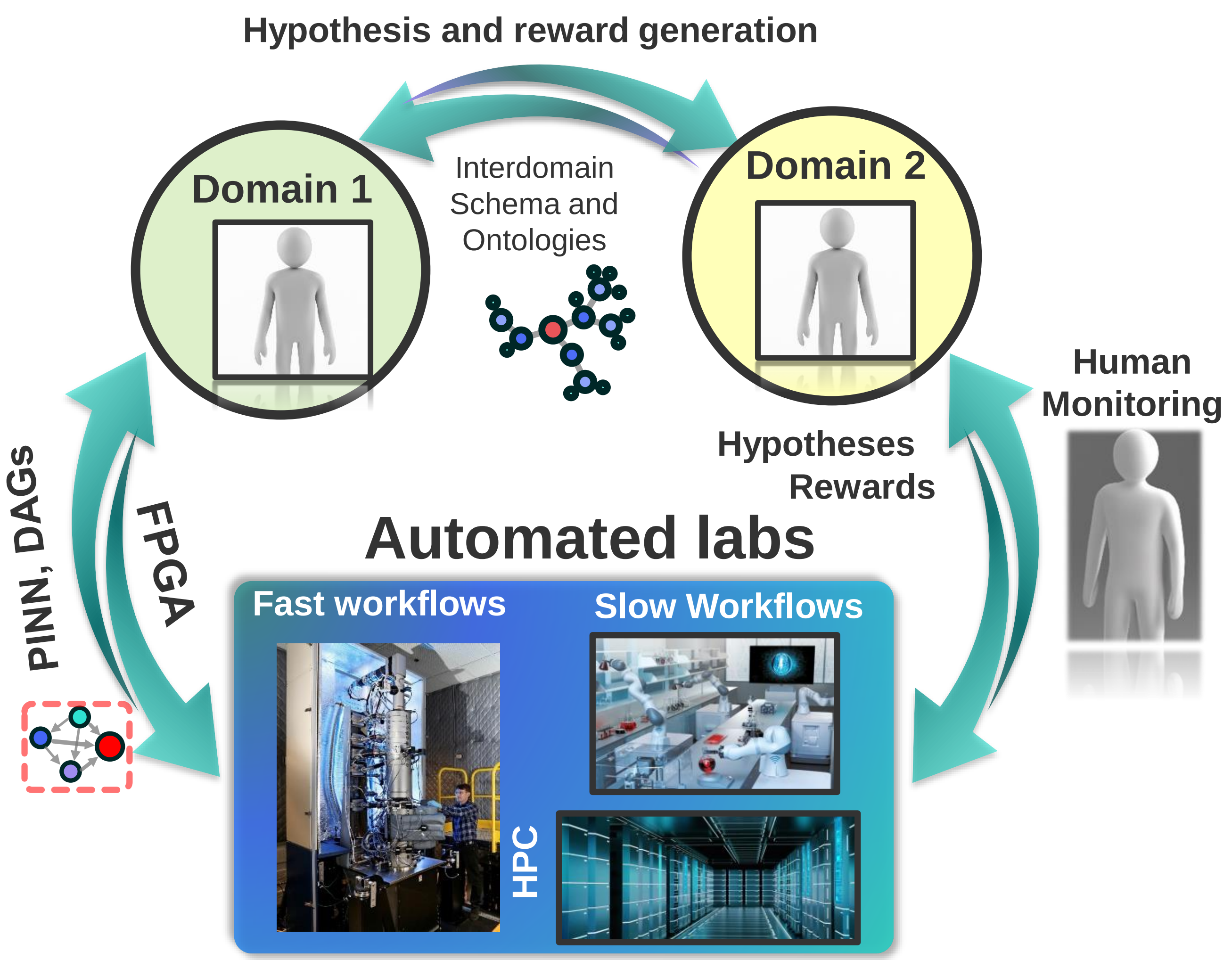
6. Hypothesis generation beyond human (an AGI question).

Integration across human domains

Simple GP-BO, structured GP with probabilistic physical models, and multifidelity GP can be used for fully autonomous workflows and account for real times/costs of experiments.

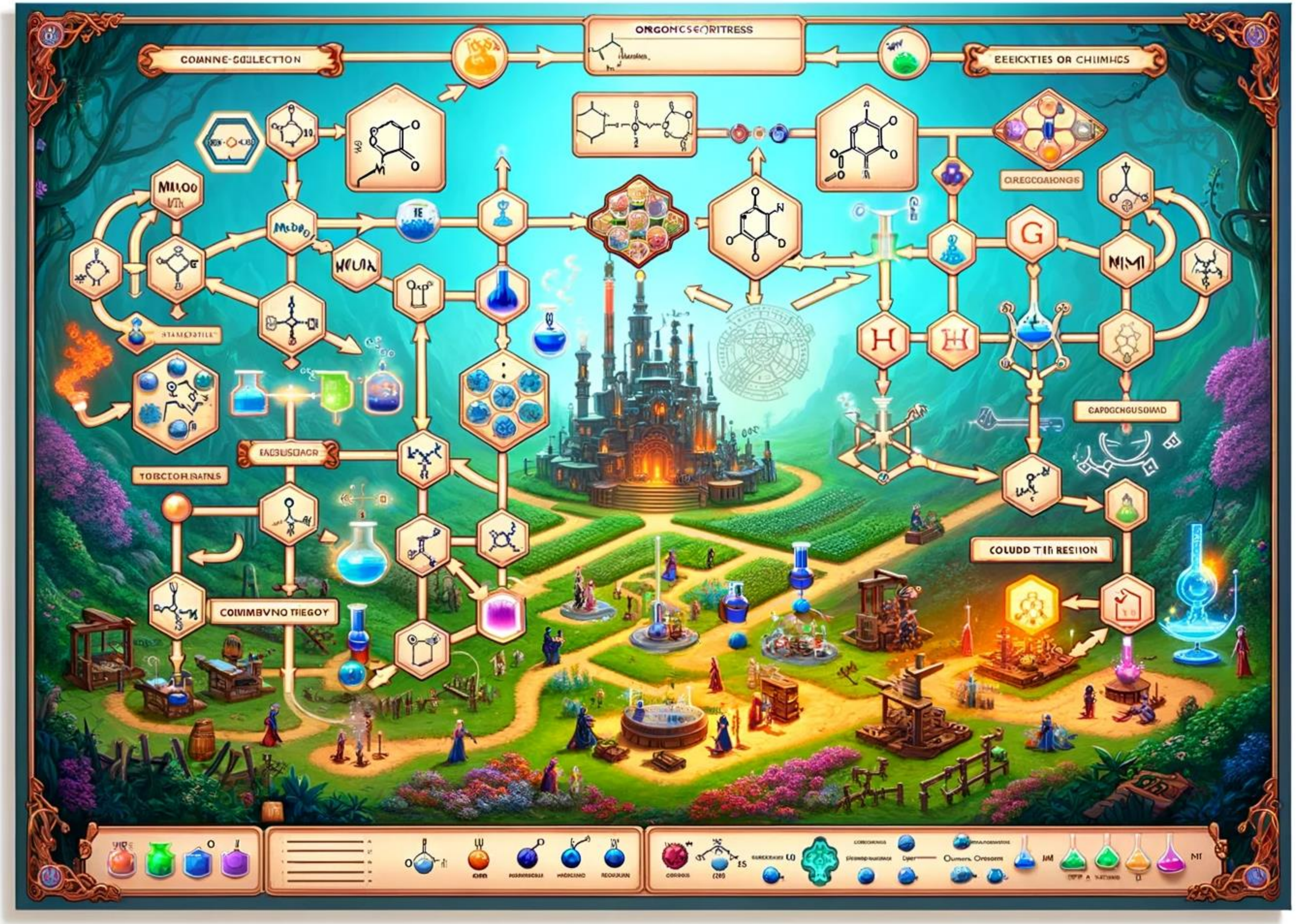
Multi-task GP-BO, multifidelity sGP-BO, and more complex methods require human intervention:

- **All metods:** reward and objective functions
- **Interactive BO:** balanceof exploration-exploitation and strength of fidelity- and physics priors
- **Asymmetric MF BO:** belief in parameters, causal structure between observables
- **A* and Decision trees:** human heuristics for roll-outs
- **Experiment planning:** hypotheses making



Homo Ludens: Gamification of SDLs

- Collaborative Environment between humans and NPCs
- Role Specialization
- Quests and Objectives
- Dynamic Interaction and Adaptation:
- Skill and Experience Growth
- Real-time Decision
- Interactive World
- Automated Tasks and Challenges
- Resource Management
- Progress Tracking and Rewards



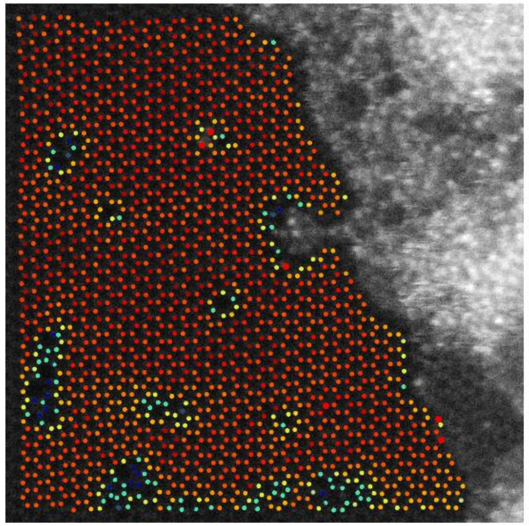
And technology

"New directions in science are launched by new tools much more often than by new concepts. The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained."

Freeman Dyson

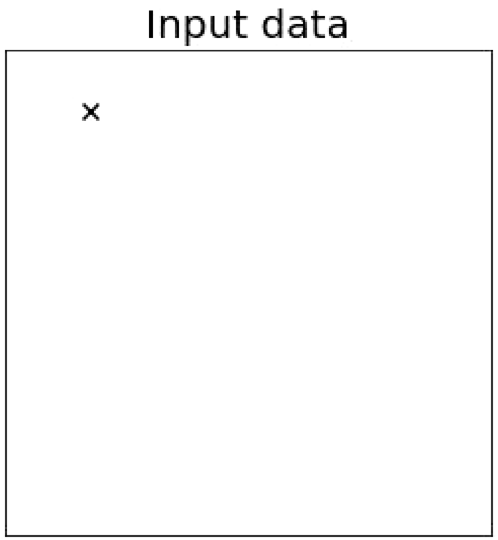
Open source, open science

AtomAI: general deep learning applications to experimental imaging data
https://github.com/pycroscopy/atomai



GPax: physics-informed Gaussian processes and deep kernel learning (JAX)
https://github.com/ziatdinovmax/gpax

pyroVED: invariant variational autoencoders for image and spectral analysis (PyTorch)
https://github.com/ziatdinovmax/pyroved



Concluding:

- **Machine learning is great, but**
  - Requires domain expertise
  - Ease of use for deployment
  - Some ML knowledge

- **ML transition to domain areas:**
  - Reward functions
  - Workflow design
  - Hyperlanguage

- **Human in the loop – hAE**
  - Explainability
  - Interventions
  - Alignment

Sergei V. Kalinin
sergei2@utk.edu

Connect!

YouTube

Twitter: Sergei_imaging