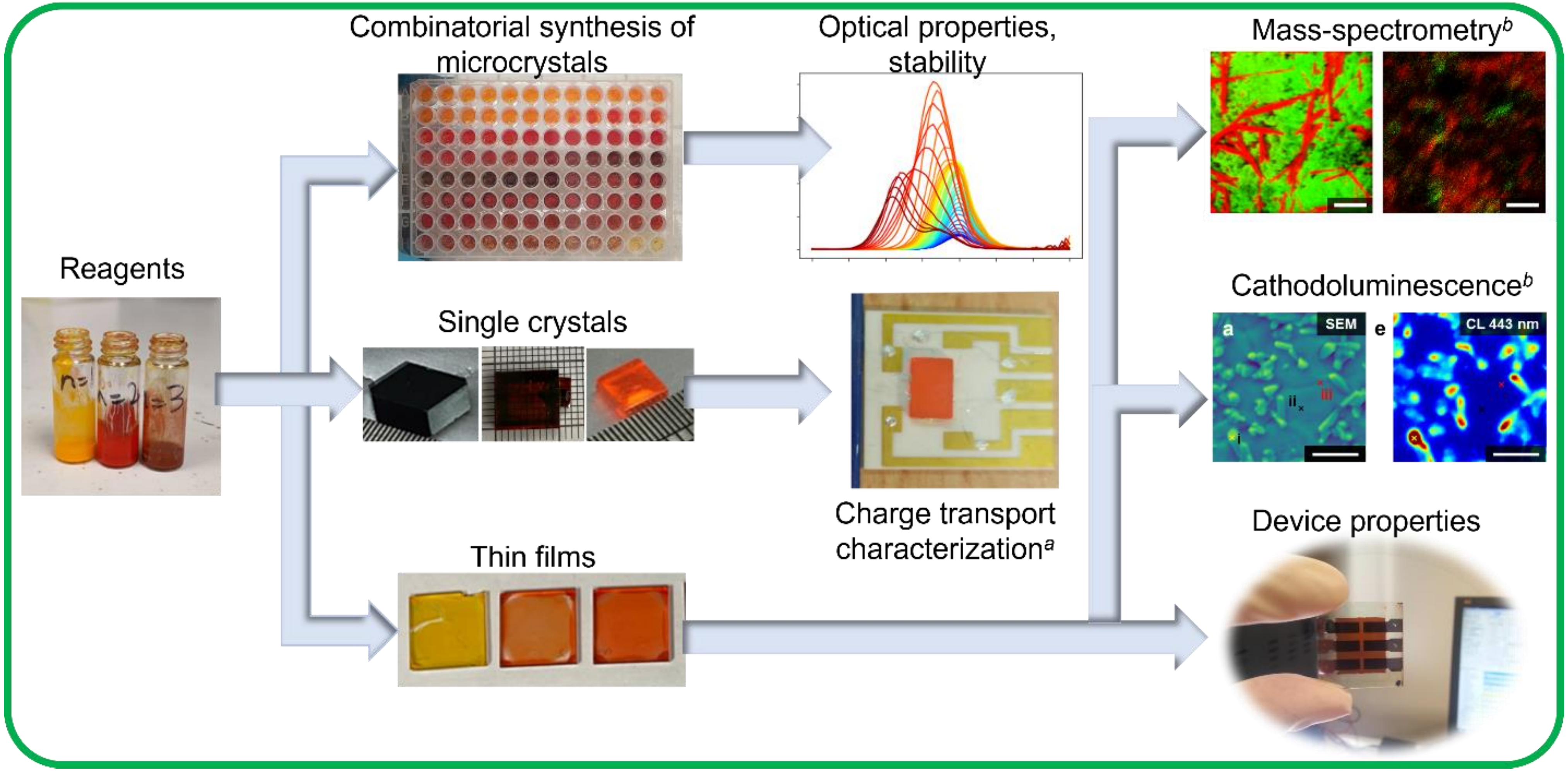
THE UNIVERSITY of TENNESSEE KNOXVILLE

Day 1: Why Should We Use Machine Learning?

Sergei V. Kalinin

Microscopy starts in the materials labs

What is A Workflow?



- **Workflow:**
- Ideation, orchestration, implementation
- Domain specific language
- Dynamic planning: latencies and costs
- Reward and value functions

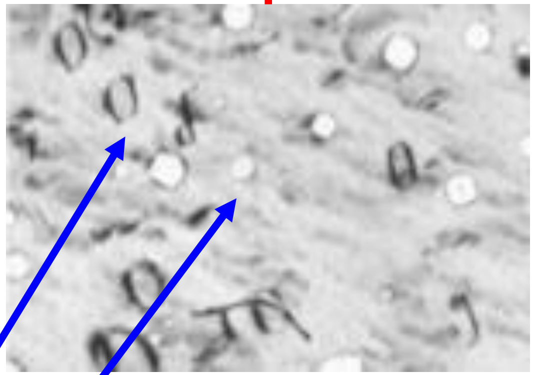
Designed in academia and adopted by industry

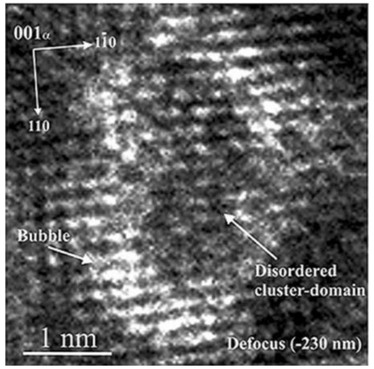
- Are they optimal?
- Can we design them better?
- Can they be changed dynamically?

Workflows for Nuclear Materials Design

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

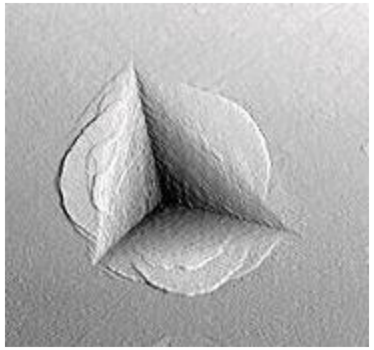
Materials selection





Prior knowledge

Operator curiosity

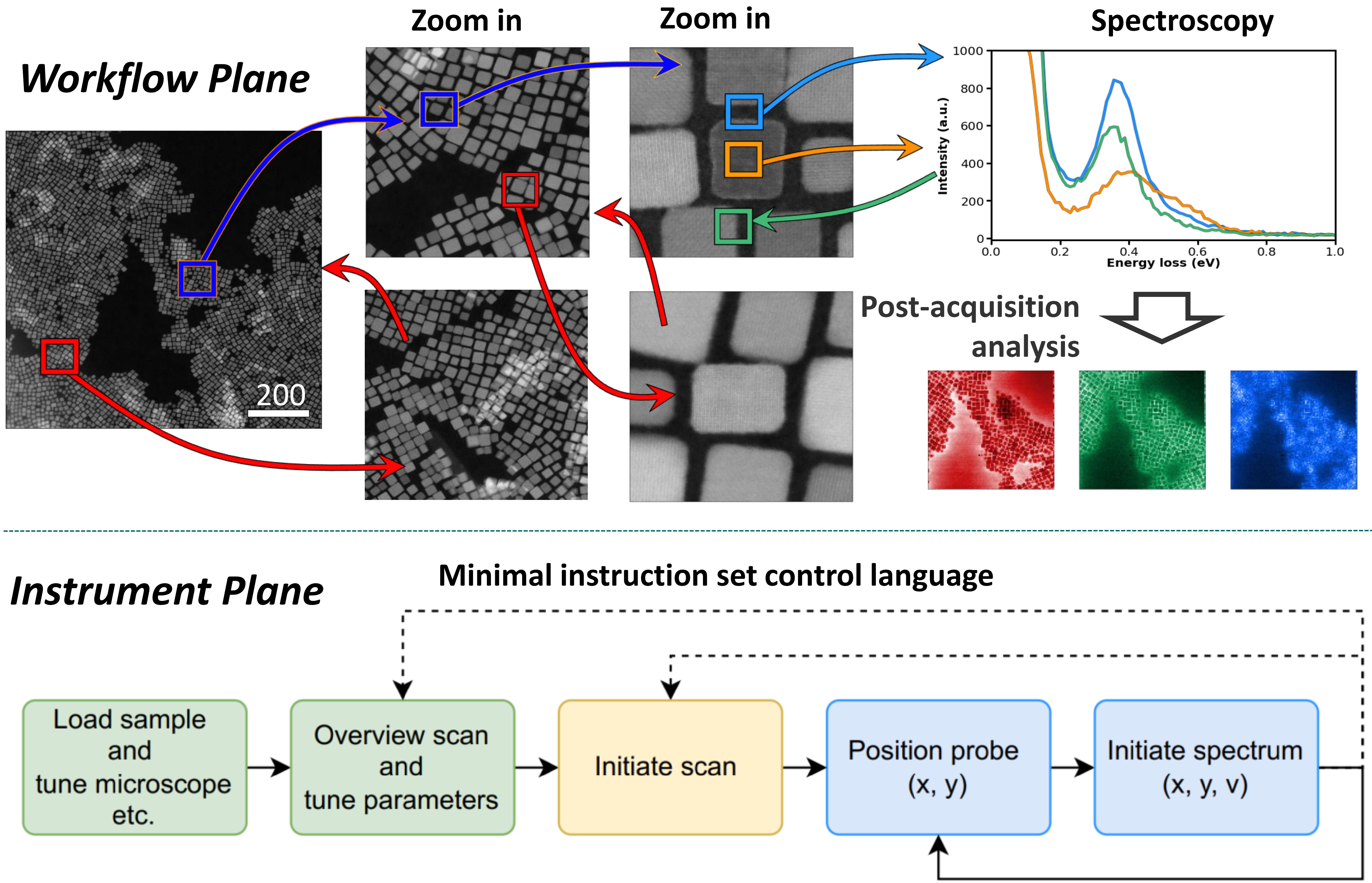


Value

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

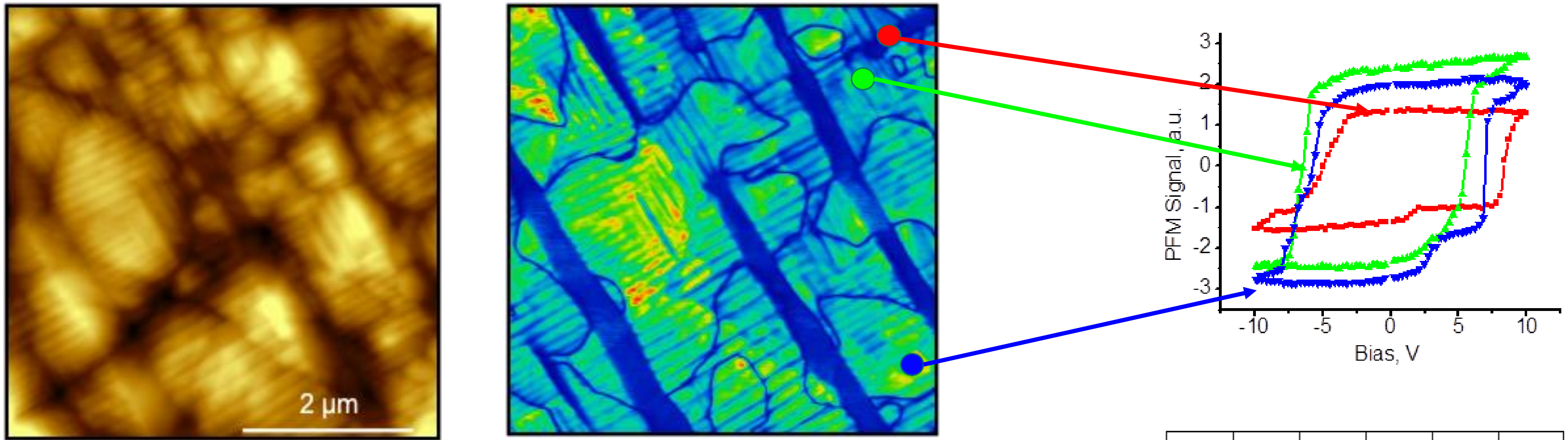
Workflows in STEM

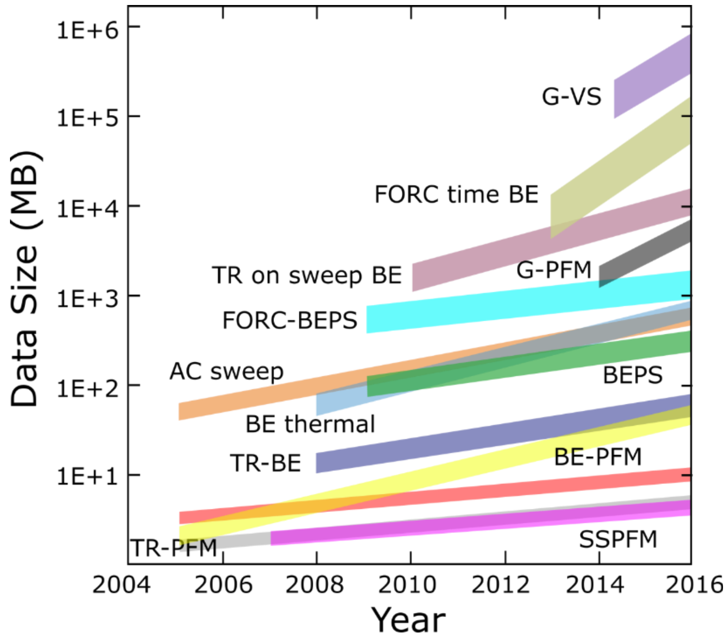
Prior Knoweldge



Objective and Reward

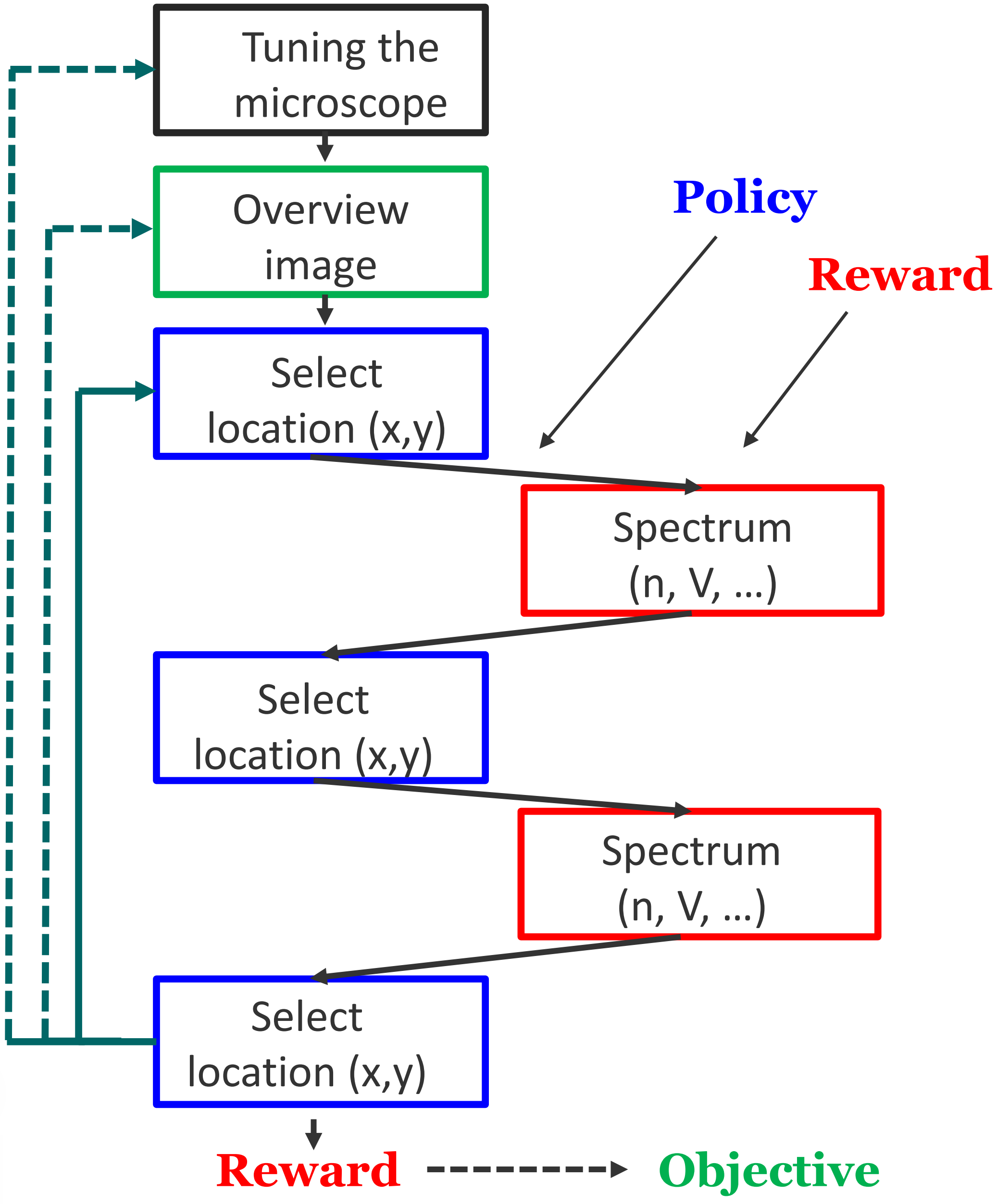
Decision Making in SPM





- Interesting functionalities are expected at the certain elements of domain structure
- We can guess some; we have to discover others

- **Experimental objectives → ML Rewards**
  - Microscope optimization
  - Properties of a priori known regions of interest
  - Discovery of regions with interesting properties
  - Physical theory falsification



To implement the ML workflows, we start from emulating the human operations:

- Well defined and explainable commands
- Extensive domain expertise
- Potentially available data from experiments

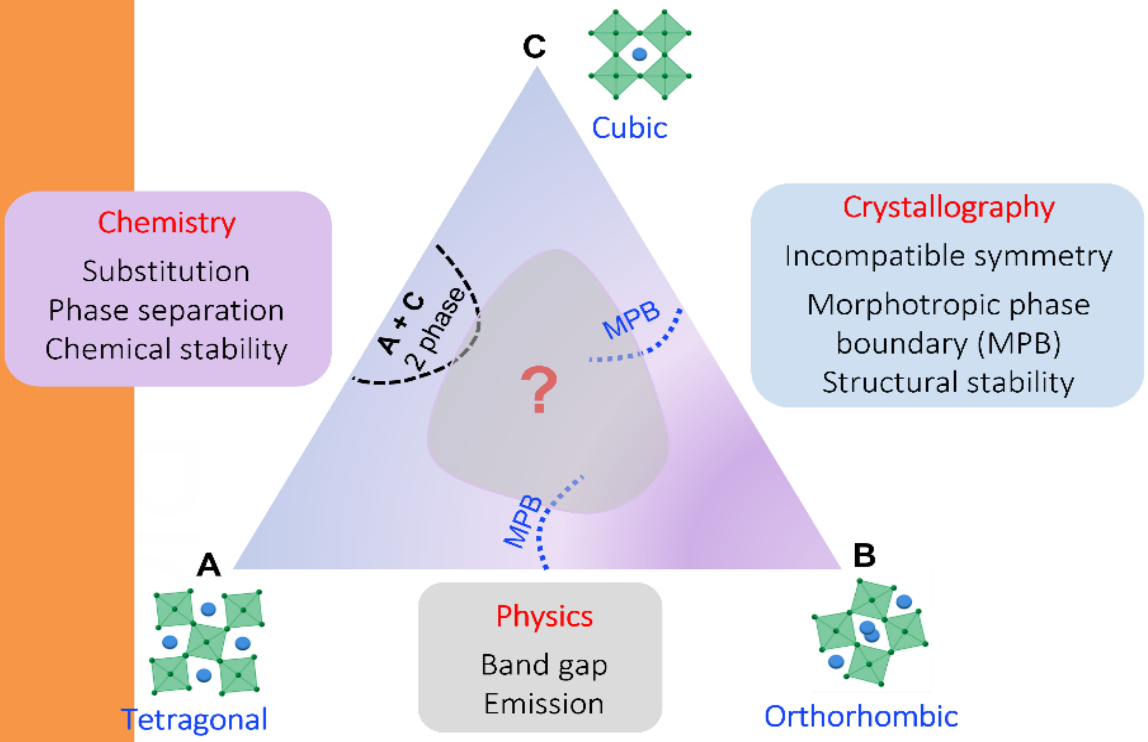
Development of ML workflows can give rise to more complex imaging modalities

- Data volumes and dimensionalities above human level
- More complex modes of sampling
- "Guardian angel" modules

However, we always have to think about

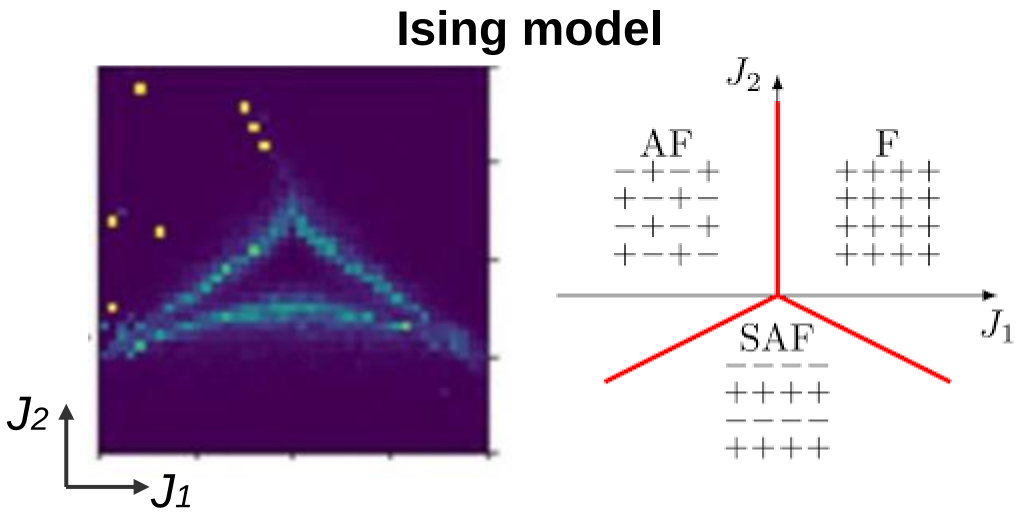
- Reward function(s) for imaging problem
- Reward functions for materials problem
- Overall objective

Why synthesis (or theory)?

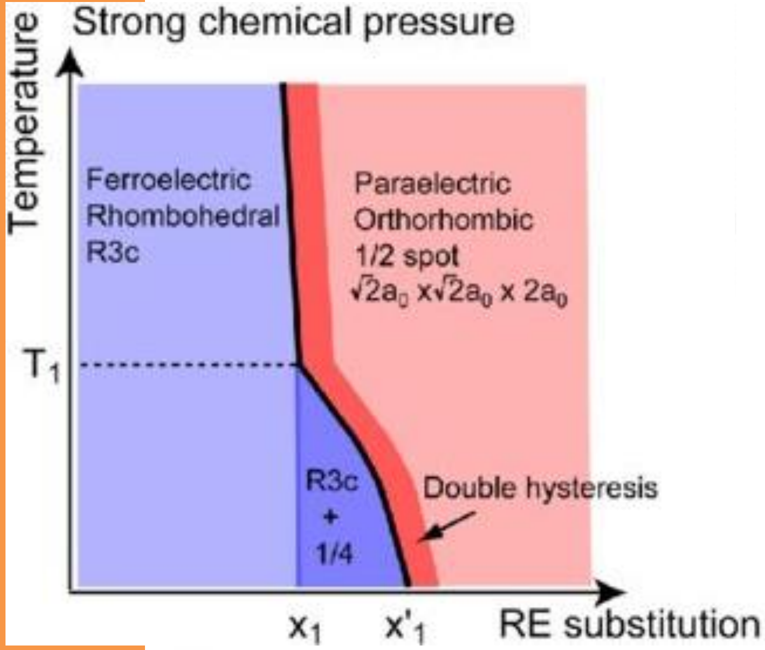


- Automated synthesis in its simplest form requires some way to navigate phase diagrams
- In more complex form, processing space.
- Ideally, incorporate physical knowledge

- Similar problem - theory

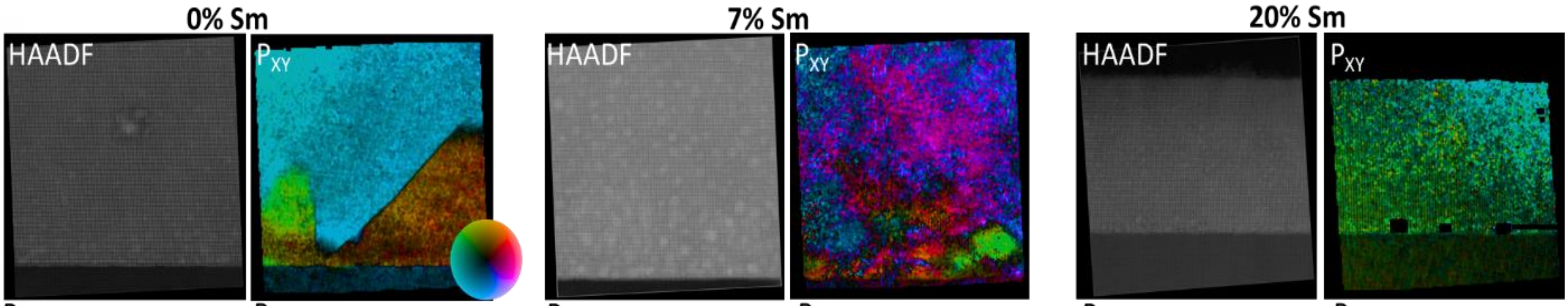


Combinatorial Synthesis



Sample by I. Takeuchi, UMD
Phase diagram by N. Valanoor et al.

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



arXiv:2004.11817

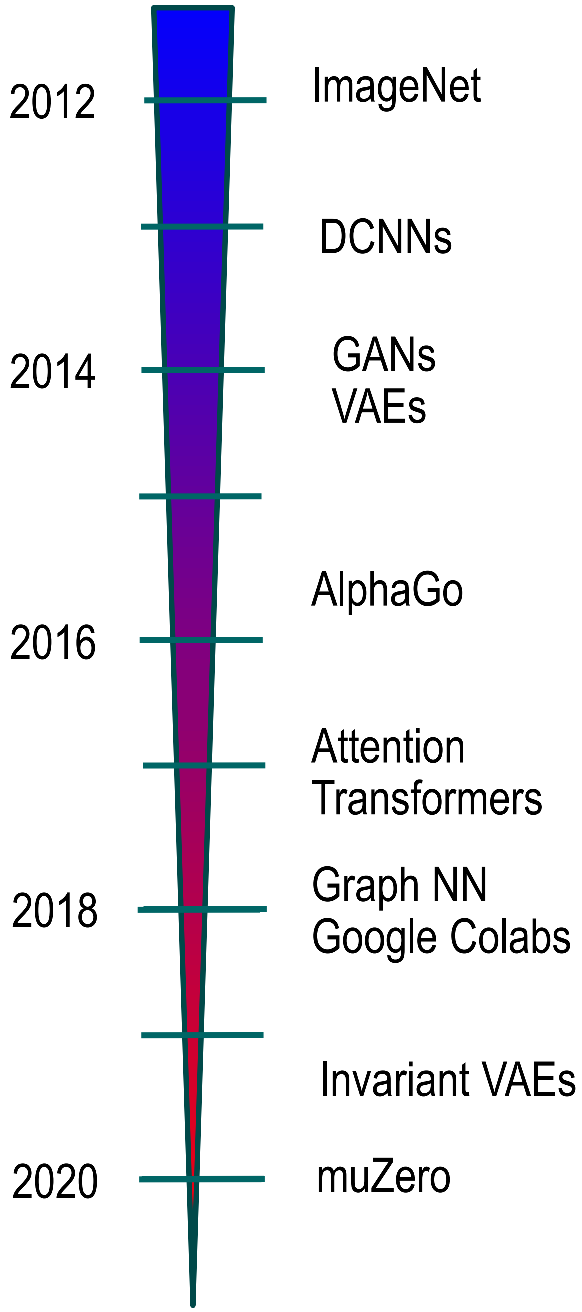
Why Machine Learning?

- Last decade has experienced an explosive growth of machine learning and artificial intelligence applications
- These developments have spanned areas from computer vision to medicine to autonomous systems and games
- However, the progress and impact as applied to experimental physical sciences has been minimal….

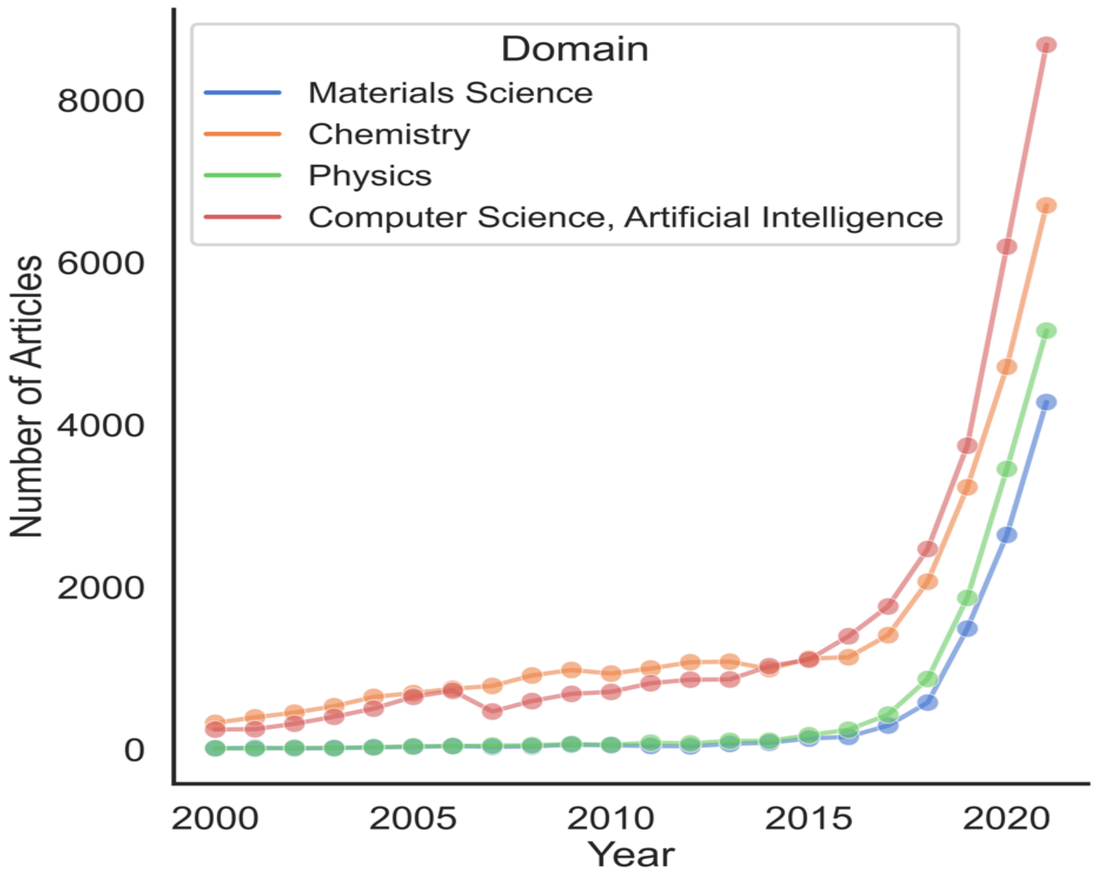
Why is it difficult?

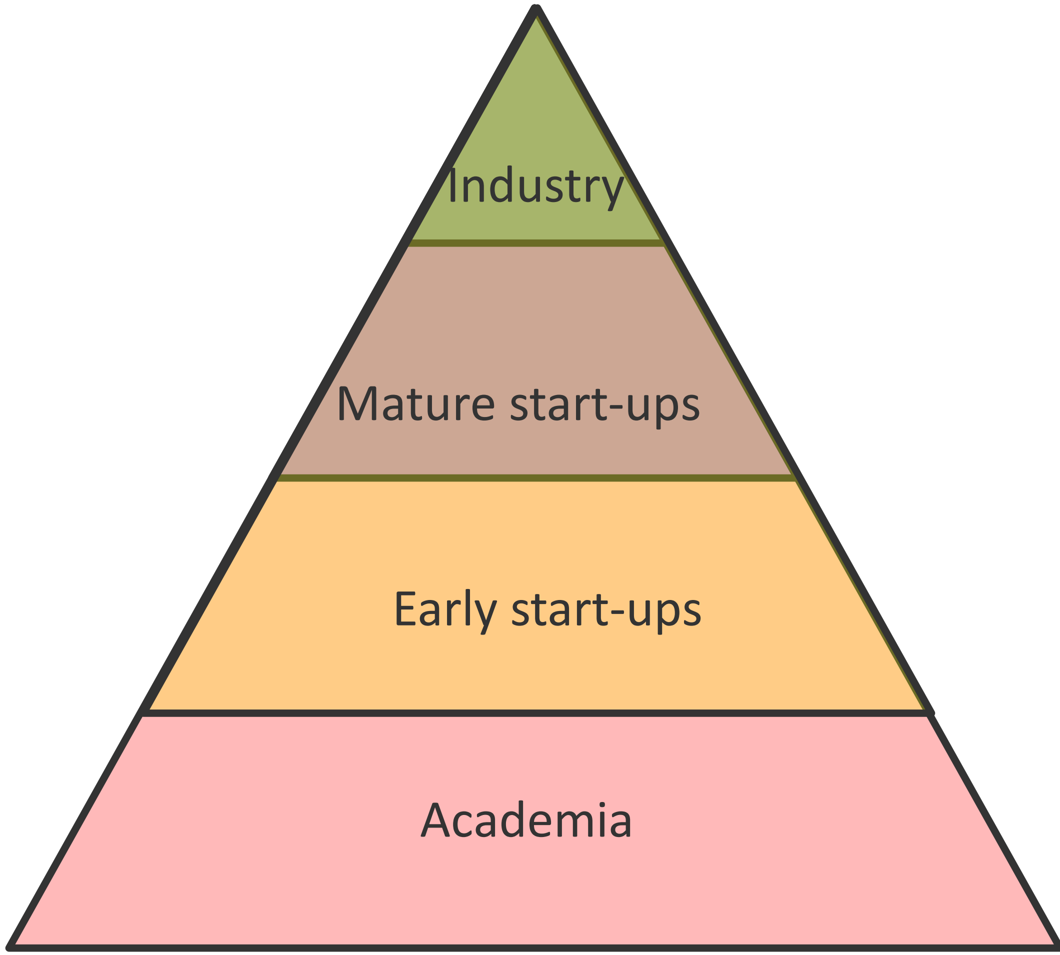
- Requires domain expertise and domain-specific goals
- Deeply causal and hypothesis drive nature of domain sciences
- No single answer: culture, not a method
- Infrastructure, open code, open data
- **Most important:** active nature of scientific process

Microsoft: GitHub
Meta: Open Catalyst,
Meta: Papers with Code
Toyota: TRI
Google: AlphaFold
NVIDIA: protein folding



ML in Domain Sciences





Analysis by B. Blaiszik, Argonne

- The rapid adoption of ML in domain sciences and industrial R&D is a very recent trend
- Technologies and workforce emerge from academia into industry
- We can estimate potential growth rates comparing to cloud computing 15 - 20 years ago

“Eras” of ML in Industry

- **Before 2000:** It’s all about IT (dotcoms, Amazon, etc)
- **2000 - 2010:** It’s all about collecting and searching data (Facebook, Google, Uber)
- **2010 – 2020:** What do we learn from data (correlative era)
- **2020 – now:** Physics is the new data

- Classical machine learning is underpinned by the existence of the large static data sets – from MNIST to emerging medical, bio, faces, etc.
- Real world problems are associated with the large distribution shifts, often small data sets, and presence of uncontrollable exogenous factors
- Also, real world problems are often active learning: we interrogate the data generation process and provide feedback, not deal with static data sets
- However, we often have extensive prior knowledge of past data, physical laws generalizing them, and strong set of inferential biases

ML for real-world applications is different!

Approach

THE UNIVERSITY of TENNESSEE KNOXVILLE

0. Getting big data: making imaging tools a part of data infrastructure

Physics: Why something happens

1. Big data:
How does
it happen?

- Unsupervised learning, clustering, and visualization

- **Biggest hurdle:** Language/ elementary tools

2. Deep data:
How can
we understand?

- Physics informed data analytics/ supervised methods

- **Biggest hurdles:** Mathematical framework, scalability of computational tools

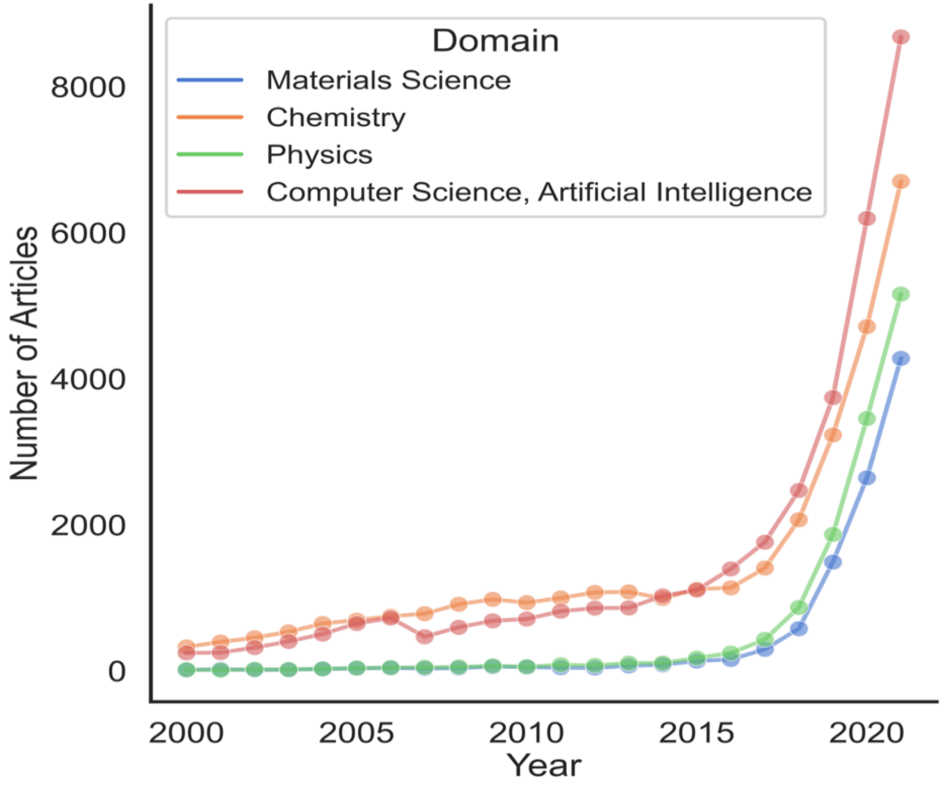
3. Smart data: How can we do better?

- Feedback and expert/AI systems

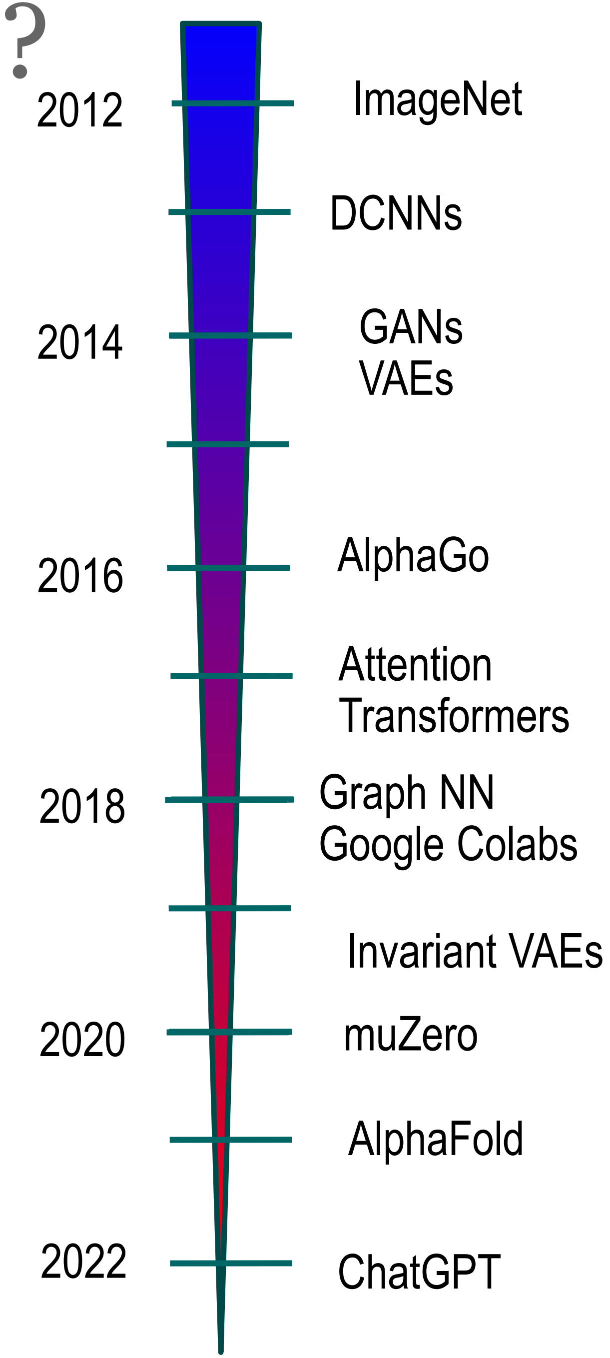
- **Biggest hurdles:** With LLMs, it is possible

How it feels most of the time:

Why machine learning in materials ?

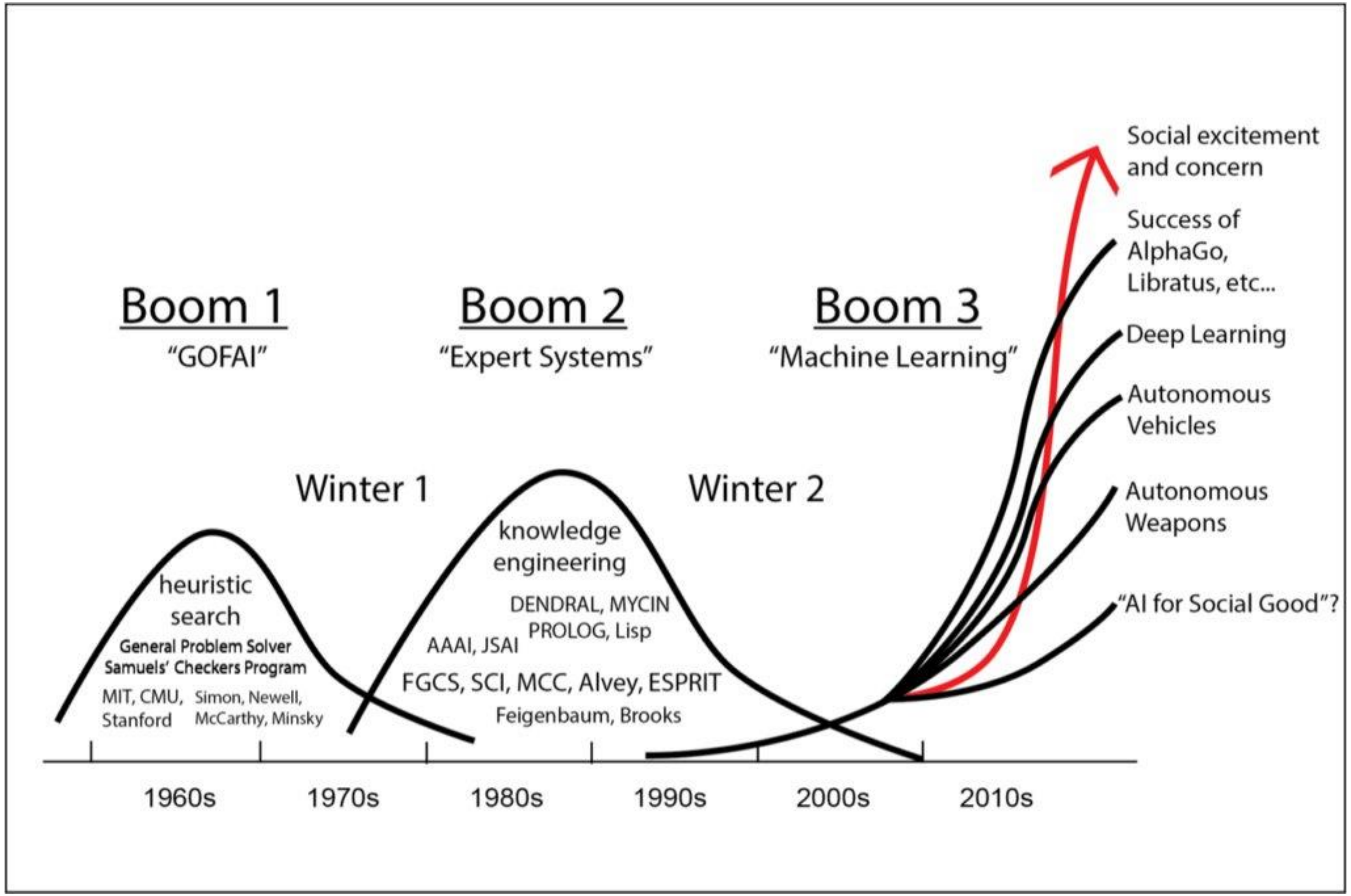


Analysis by B. Blaiszik, Argonne



- Last decade has experienced an explosive growth of machine learning and artificial intelligence applications
- These developments have spanned areas from computer vision to medicine to autonomous systems and games
- However, the progress and impact as applied to experimental physical sciences has been minimal….

Zooming out on history



https://jaylatta.net/history-of-ai-from-winter-to-winter/

History of Machine Learning - 1

- 1950s
  – Samuel's checker player
  – Selfridge's Pandemonium

- 1960s:
  – Neural networks: Perceptron
  – Pattern recognition
  – Learning in the limit theory
  – Minsky and Papert prove limitations of Perceptron

- 1970s:
  – Symbolic concept induction
  – Expert systems and the knowledge acquisition bottleneck
  – Michalski's AQ and soybean diagnosis
  – Scientific discovery with BACON
  – Mathematical discovery with AM

History of Machine Learning - 2

- 1980s:
  - Advanced decision tree and rule learning
  - Explanation-based Learning (EBL)
  - Learning and planning and problem solving
  - Cognitive architectures
  - Resurgence of neural networks (connectionism, backpropagation)

- 1990s
  - Data mining
  - Adaptive software agents and web applications
  - Text learning
  - Reinforcement learning (RL)
  - Inductive Logic Programming (ILP)
  - Ensembles: Bagging, Boosting, and Stacking
  - Bayes Net learning

History of Machine Learning - 3

- 2000s
  - Support vector machines
  - Kernel methods
  - Graphical models
  - Statistical relational learning
  - Transfer learning
  - Sequence labeling
  - Collective classification and structured outputs
  - Computer Systems Applications
    - Compilers
    - Debugging
    - Graphics
    - Security (intrusion, virus, and worm detection)
  - E mail management
  - Personalized assistants that learn
  - Learning in robotics and vision

Types of Machine Learning

Supervised (inductive) learning
– Given: training data + desired outputs (labels)

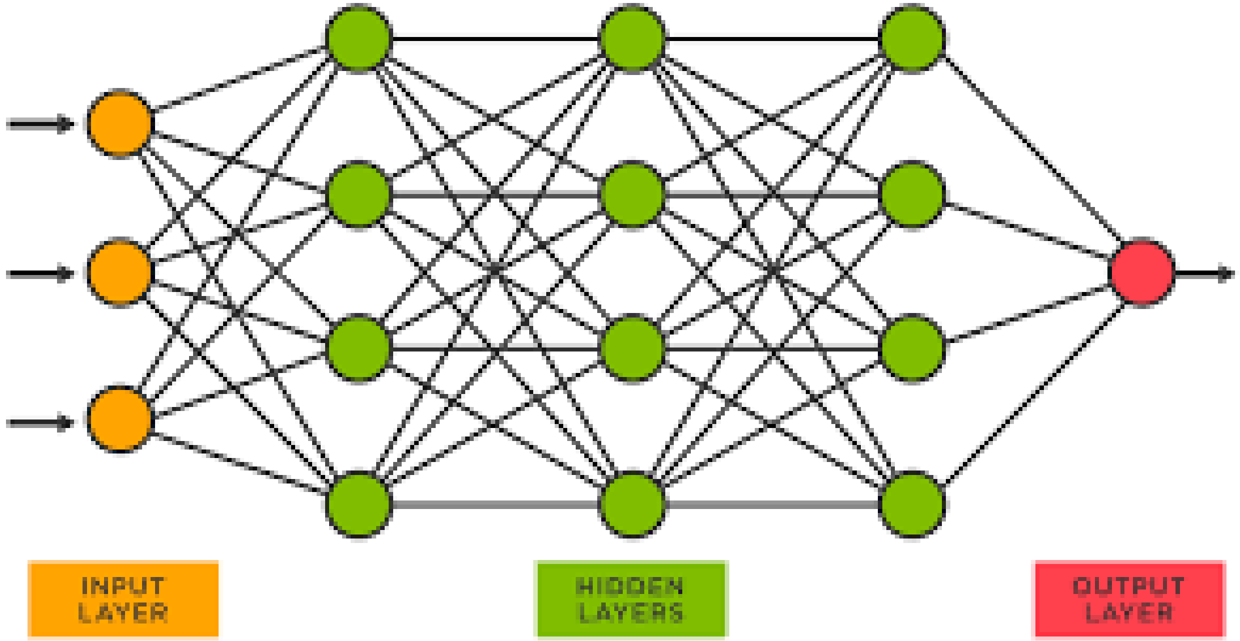
- **Unsupervised learning**
– Given: training data (without desired outputs)

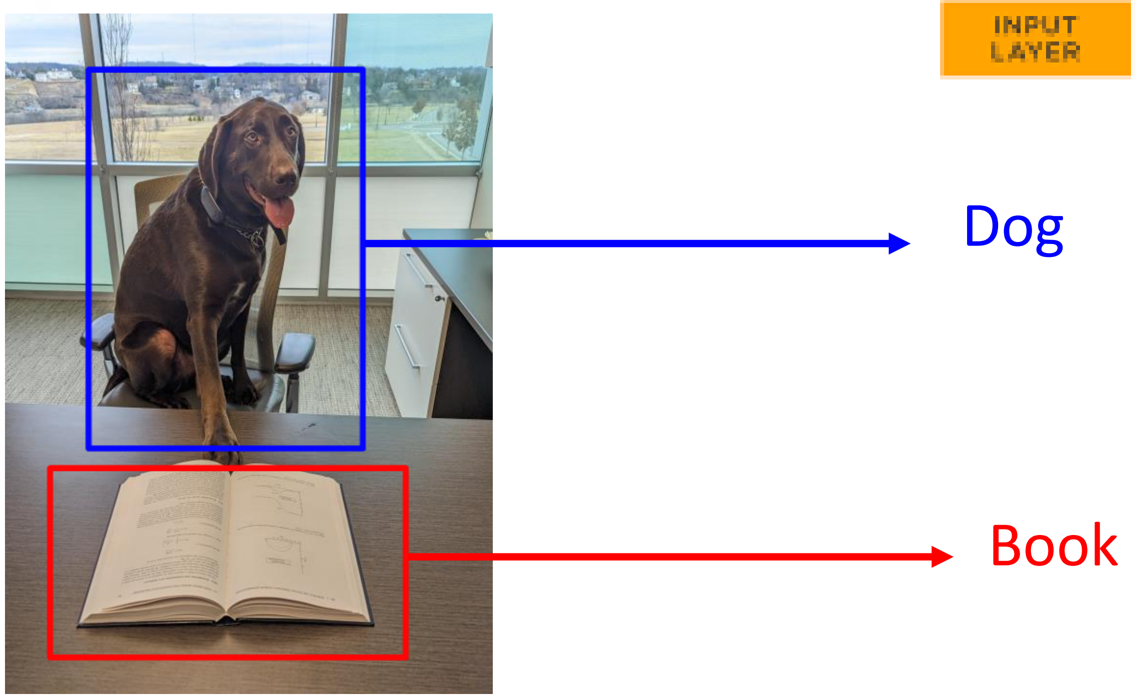
- **Semi-supervised learning**
– Given: training data + a few desired outputs

- **Reinforcement learning**
– Rewards from sequence of actions

Supervised Machine Learning

- Classification
- Regression
- Semantic segmentation
- Instance segmentation
- …





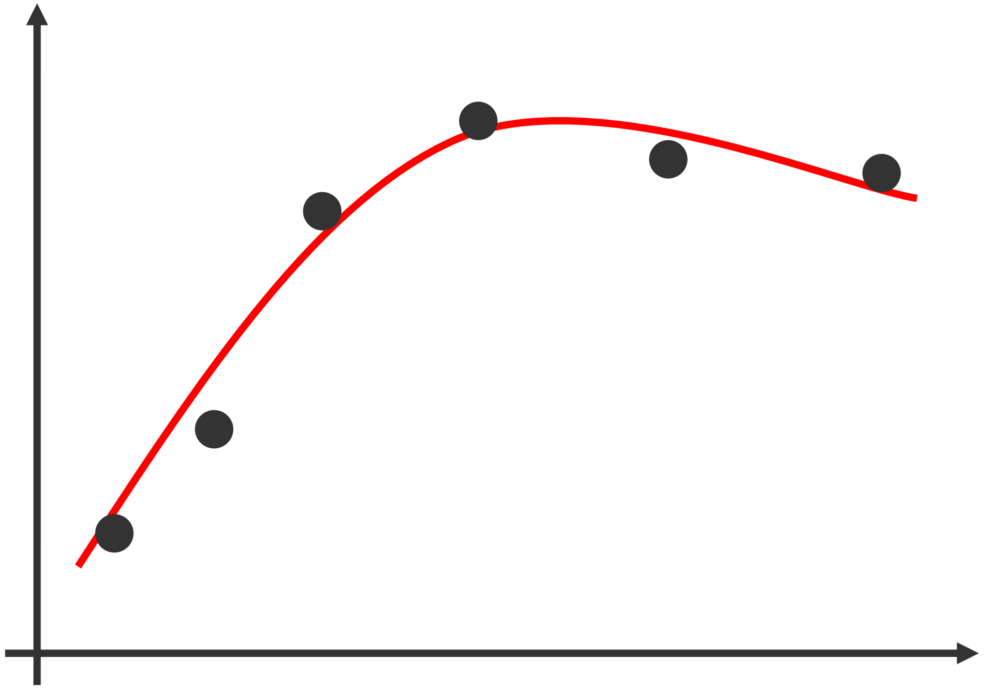
Classification

- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function f(x) to predict y given x
- If y is categorical == classification

Application	Input Data	Classification
Medical Diagnosis	Noninvasive tests	Results from invasive measurements
Optical Character Recognition	Scanned bitmaps	Letter A-Z and digits 0-9
Protein Folding	Amino acid sequence	Protein shape (helices, loops, sheets)
Materials Discovery	Composition	Metal/Semiconducotr
Research Paper Acceptance	Words in paper title	Paper accepted or rejected

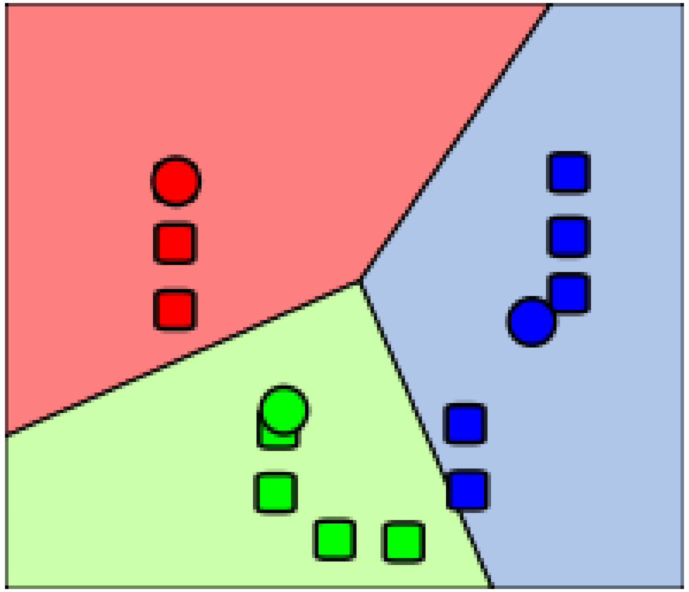
Regression

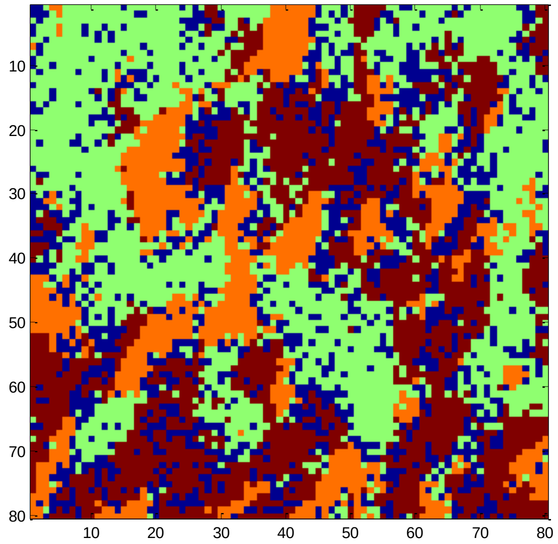
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function f(x) to predict y given x
- y is real-valued == regression

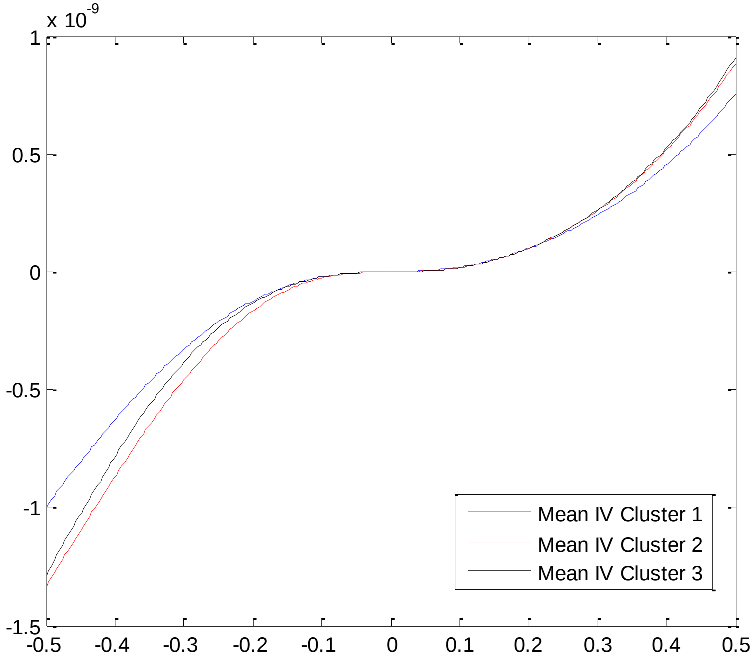


Unsupervised Learning

- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
- E.g., clustering



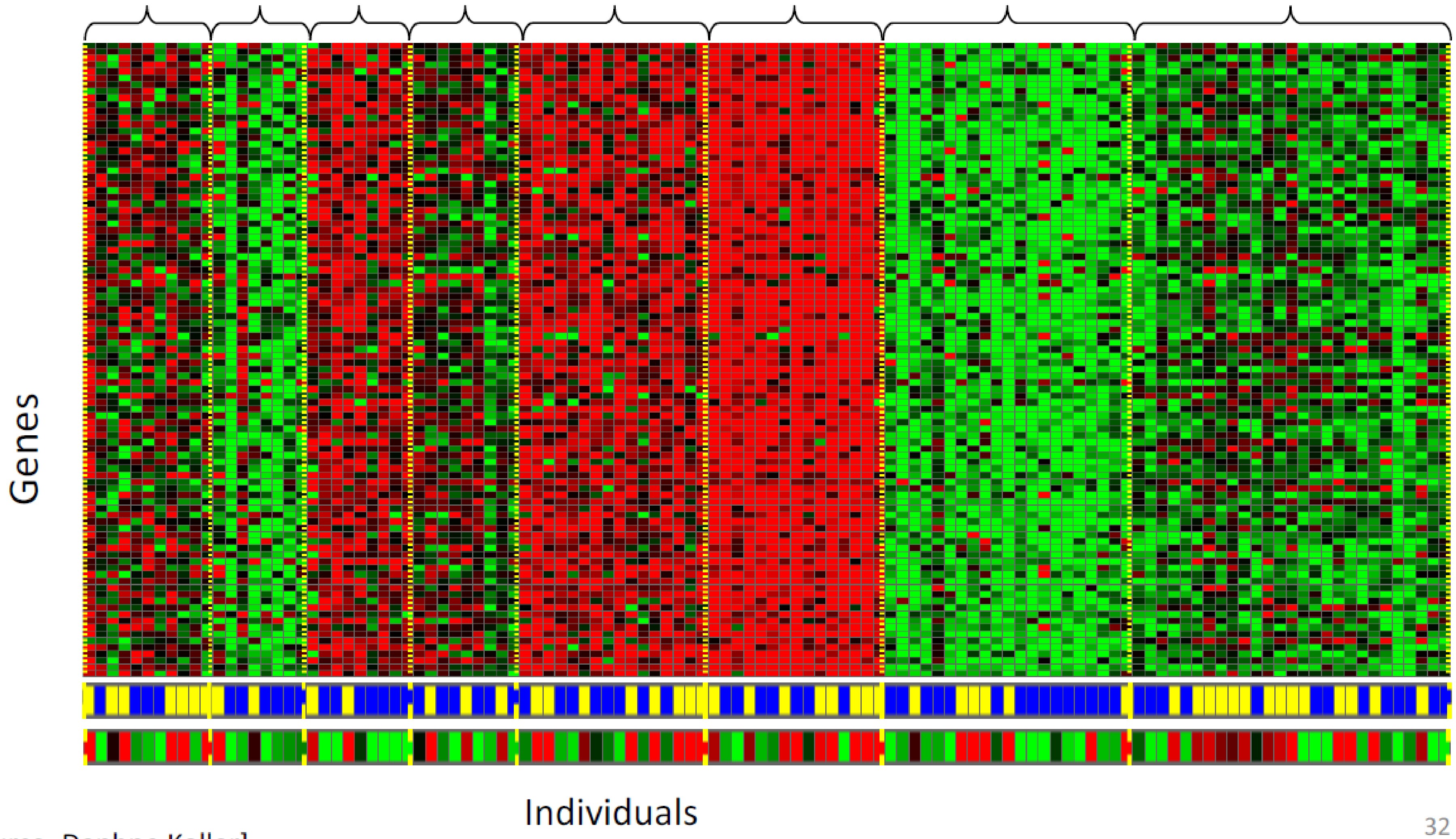




M. Ziatdinov, A. Maksov, L. Li, A. Sefat, P. Maksymovych, and S.V. Kalinin, *Deep data mining in a real space: Separation of intertwined electronic responses in a lightly-doped* $BaFe_2As_2$, Nanotechnology **27**, 475706 (2016).

Unsupervised Learning

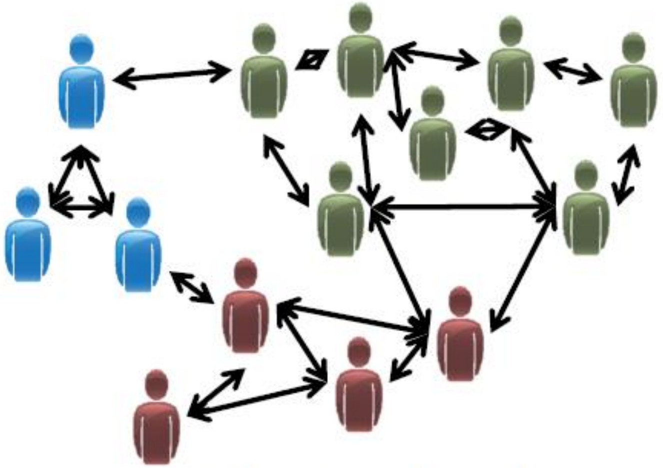
Genomics application: group individuals by genetic similarity



[Source: Daphne Koller]

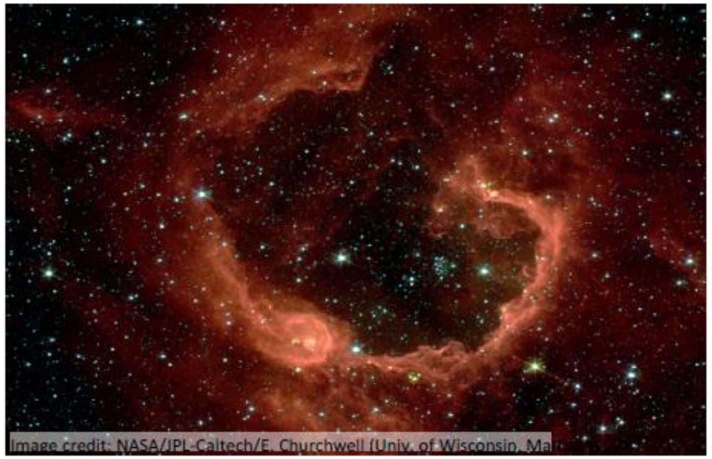
Unsupervised Learning

Organize computing clusters



Social network analysis

Market segmentation

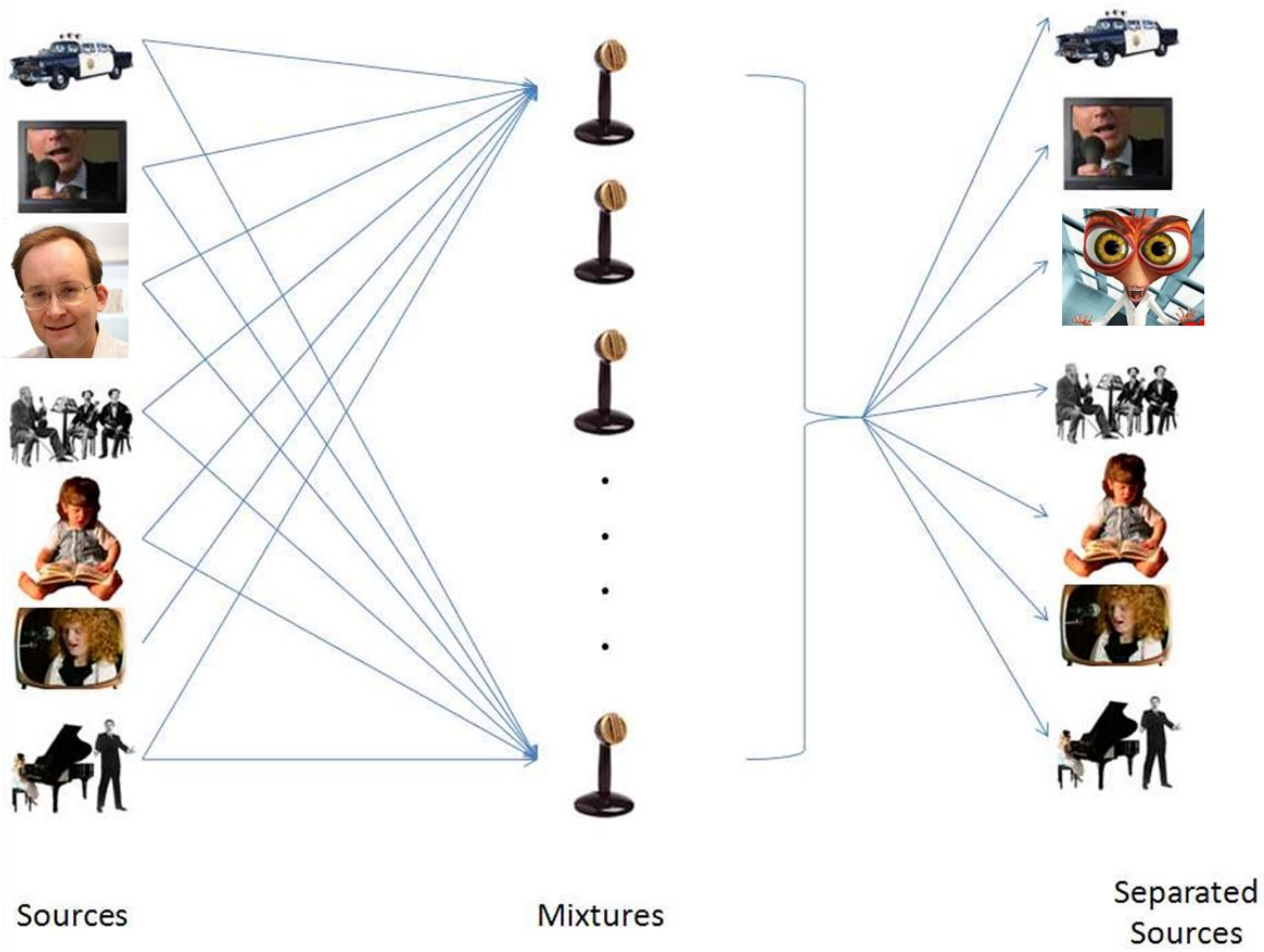


Astronomical data analysis

Slide credit: Andrew Ng

Unsupervised Learning

Number of signals are being produced simultaneously; with the objective of separating and following each source separately

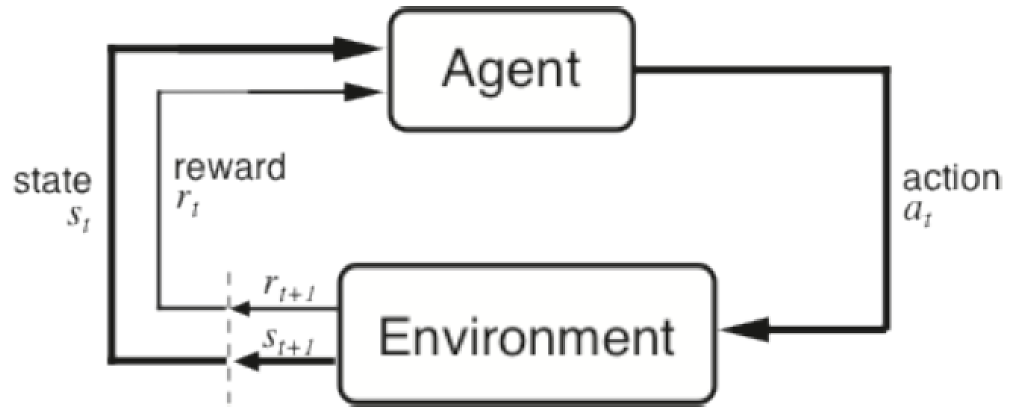


Reinforcement Learning

Given a sequence of states and actions with (delayed) rewards, output a policy
– Policy is a mapping from states to actions that tells you what to do in a given state

- Examples:
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand

RL: Agent and Environment



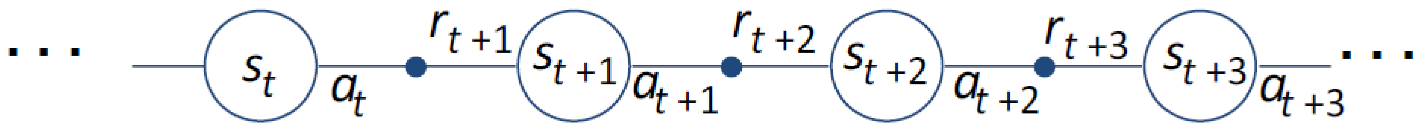
Agent and environment interact at discrete time steps : $t = 0, 1, 2, K$

Agent observes state at step t: $s_t \in S$

produces action at step t: $a_t \in A(s_t)$

gets resulting reward : $r_{t+1} \in \Re$

and resulting next state : s_{t+1}



Slide credit: Sutton & Barto

Reinforcement Learning in Action

https://www.youtube.com/watch?v=GtYlVxv0py8

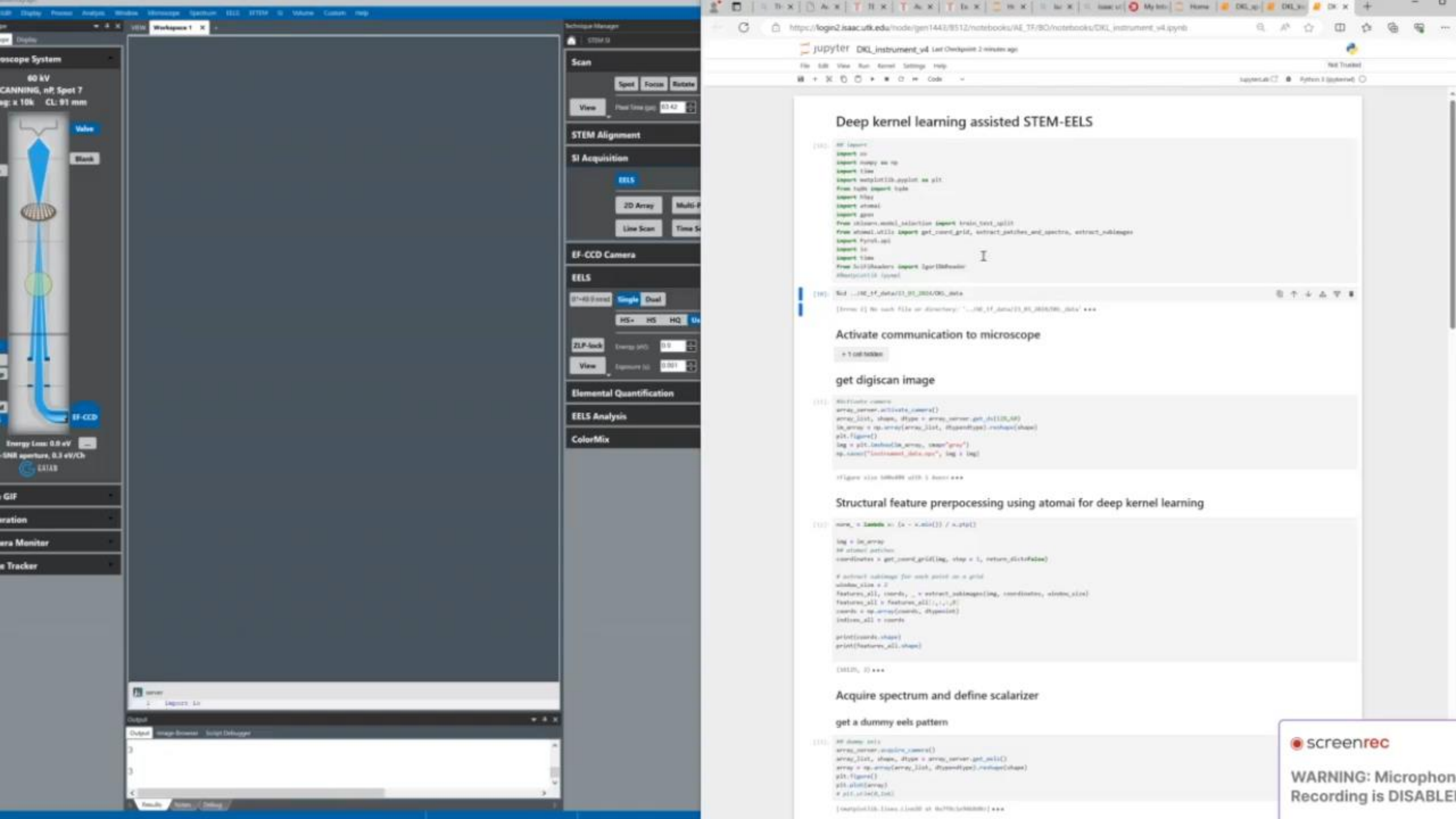
Reinforcement Learning Applications

Chemical Synthesis and Drug Discovery

M. Ahmadi lab
UTK MSE

Cloud Laboratories

Emerald Cloud Lab,
SF and CMU



Scan
STEM Alignment
SI Acquisition
EELS
2D Array
Line Scan
EF-CCD Camera
EELS
Single
Dual
ZLP-lock
View
Elemental Quantification
EELS Analysis
ColorMix
60 kV
Energy Loss: 0.0 eV
jupyter DKL_instrument_v4
Deep kernel learning assisted STEM-EELS
Activate communication to microscope
get digiscan image
Structural feature perprocessing using atomai for deep kernel learning
Acquire spectrum and define scalarizer
get a dummy eels pattern
screenrec
WARNING: Microphon
Recording is DISABLE