THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 02: History of ML and Scientific Data

Instructor: Sergei V. Kalinin

How we make decisions:

human + AI

How we execute decisions:

human + (automated) instrumentation

Making it real:

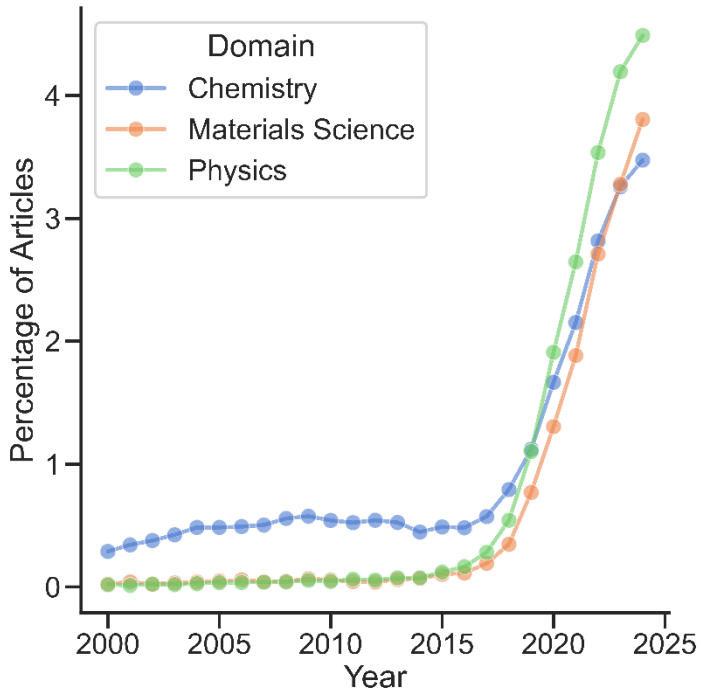
- What decisions can be executed?
- How fast can decisions be made?
- How fast can decisions be executed?
- How do we improve decisions?

Each hypothesis requires defining experimental action space

Curriculum and philosophy

ML Adoption in New Areas:

1. Solve extant problems that you can formulate (LLMs can help)
2. New ML opportunities: VAEs, game theory, decision theory, etc.
3. Building representations: there is no computation without representation
4. Building rewards, heuristics, and roll-outs: using ML for the real world problems



Analysis by B. Blaiszik, Argonne

Course Information

Faculty Contact Information:

Instructor: Prof. Sergei V. Kalinin,
Office: 314 IAMM
E-mail: sergei2@utk.edu
Teaching Assistant:

Instructor Availability:
Please don't hesitate to email me with updates, questions, or concerns. I will typically respond within 24 hours during the week and 48 hours on the weekend. I will notify you if I will be out of town and if connection issues may delay a response.

Meeting Time: 10:20 am - 11:10 am MWF, Ferris Hall 502

The lectures and materials will be posted on Canvas and at GitHub:
https://github.com/SergeiVKalinin/MSE_Fall2025

Office Hours:
Friday 1:30 - 3:00 PM are open for 1:1 meetings to discuss any course related item. Please set up time via email.

Prerequisites

To be successful in this course you will need a general background in materials science. Python or similar programming experience, while not essential, will be extremely useful. Students without any prior programming experience should expect to spend extra time outside of class learning basic skills.

This and that

Learning Environment:
The class will be delivered as in-person lectures. The Jupyter notebooks, code libraries, and videos provided. Weekly programming exercises will be assigned via Google Colabs and those students wishing to interact with the instructor in person should attend office hours.

Use of ChatGPT:
Strongly encouraged both for programming and written assignments. However, the students have to be aware of the limitations of the generative models.

Grading & Policies:

- Homework assignments 40%
- Mid-terms (2) 30%
- Final Project & Presentation 30%

Reference Materials

I will provide copies of lecture notes, presentations, and Colabs on GitHub and Canvas. There is no specific textbook for the course and we will take material from a variety of sources including:

- Andrew Bird et al, *Python Workshop – Second Edition*, https://subscription.packtpub.com/book/programming/9781804610619/1
- Sebastian Raschka, *Machine Learning with PyTorch and Scikit-Learn*, https://subscription.packtpub.com/book/data/9781801819312/1
- Rowel Atienza, *Advanced Deep Learning with TensorFlow 2 and Keras - Second Edition*, https://www.packtpub.com/product/advanced-deep-learning-with-tensorflow-2-and-keras-second-edition/9781838821654
- (Optional) Alaa Khamis, Optimization Algorithms: AI Techniques for Design, Planning, and Control Problems, https://www.manning.com/books/optimization-algorithms
- (Optional) Peter Norvig, Artificial Intelligence: A Modern Approach, Global Edition, https://www.amazon.com/Artificial-Intelligence-Modern-Approach-Global/dp/1292401133

Homework 1:

- Create new Colab, https://colab.google/
- Explore NotebookLM, https://notebooklm.google/
- Chapter 1-4, Python Workshop.

Zooming out on history

K. Pearson, 1901

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space. By* KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

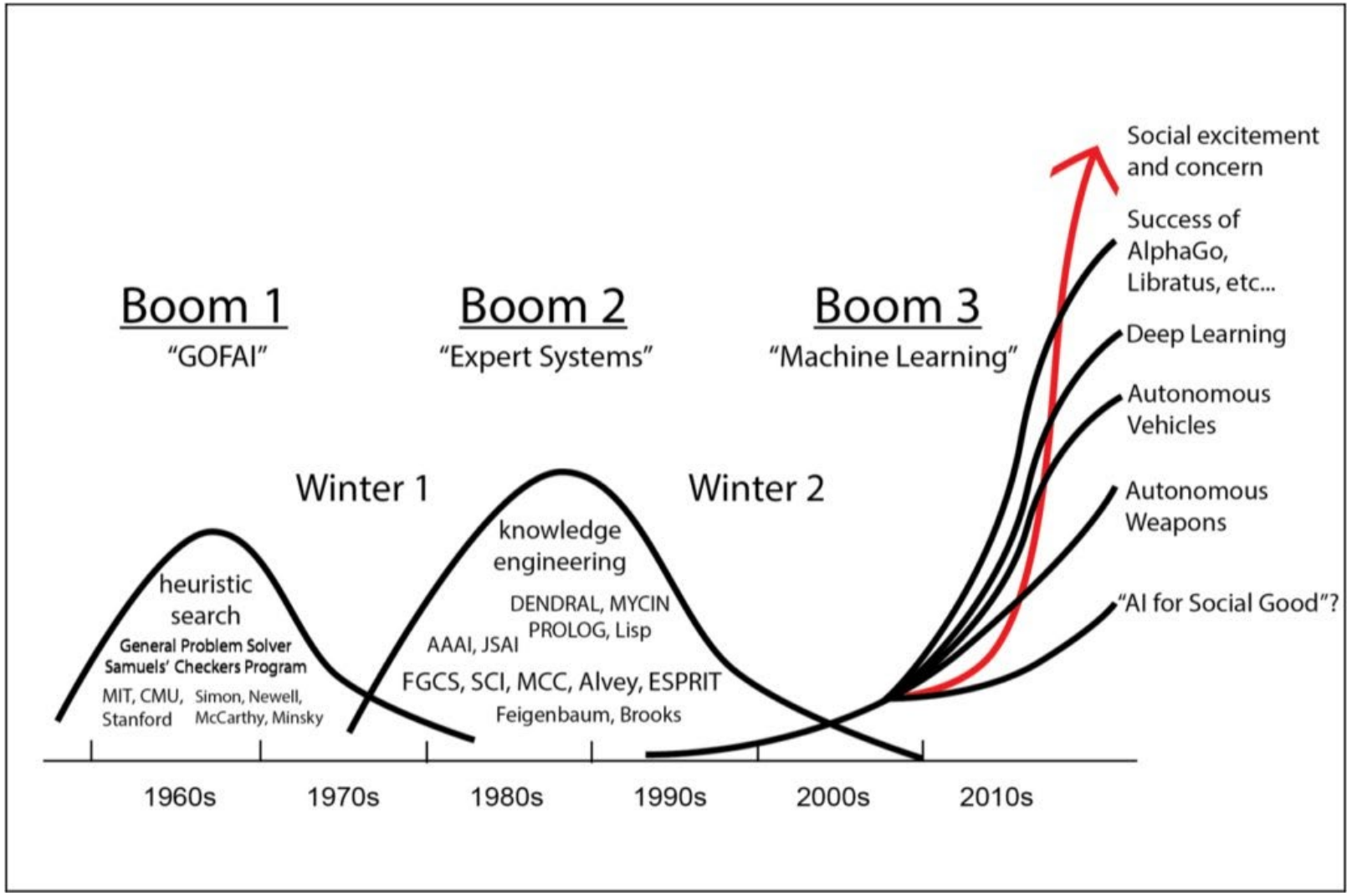
$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox about this; it is, in fact, an easily understood and most important feature of the theory of a system of correlated variables. The most probable value of y for a given value of x, say, is not given by the same relation as the most probable value of x for a given value of y. Or, to take a concrete example, the most probable stature of a man with a given length of leg l being s, the most probable length of leg for a man of stature s will not be l. The "best-fitting" lines and planes for the cases of z up to n variables for a correlated system are given in my memoir on regression †. They depend upon a determination of the means, standard-deviations, and correlation-coefficients of the system. In such cases the values of the independent variables are supposed to be accurately known, and the probable value of the dependent variable is ascertained.

(2) In many cases of physics and biology, however, the "independent" variable is subject to just as much deviation or error as the "dependent" variable. We do not, for example, know x accurately and then proceed to find y, but both x and y are found by experiment or observation. We observe x and y and seek for a unique functional relation between them. Men of given stature may have a variety

* Communicated by the Author.

† Phil. Trans. vol. clxxxvii. A, pp. 301 *et seq.*



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

Visions of the Future

AI Apocalypse: 80% of Projects Crash and Burn, Billions Wasted says RAND Report

August 19, 2024 Vernon Keenan Industry Analysis 0 Comments

A new RAND Corporation report reveals the sobering reality behind artificial intelligence (AI) projects: despite the hype, most of them fail. The study, based on interviews with 65 experienced data scientists and engineers, exposes the root causes of these failures and offers a roadmap for success.

Upcoming Events

AUG 22 8:00 am - 9:00 am
Transform Your Salesforce DevOps Tooling and Practice with AI-Driven OpsBridge Frameworks

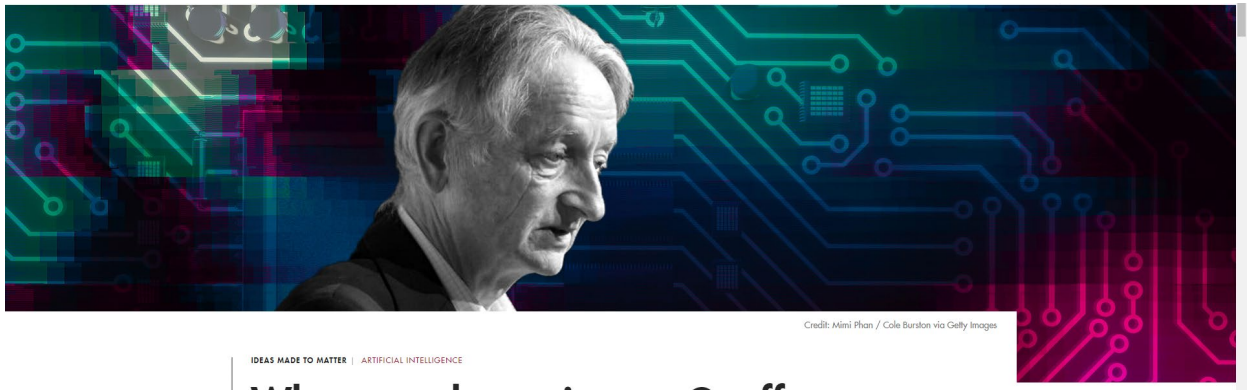
View Calendar

GET FREE EMAIL UPDATES

Taking the Human Out of the Loop: A Review of Bayesian Optimization

Citation
Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. 2016. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." Proc. IEEE 104 (1) (January): 148–175. doi:10.1109/jproc.2015.2494218.

Published Version
doi:10.1109/JPROC.2015.2494218



Credit: Mimi Phan / Cole Burston via Getty Images

IDEAS MADE TO MATTER | ARTIFICIAL INTELLIGENCE

Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI

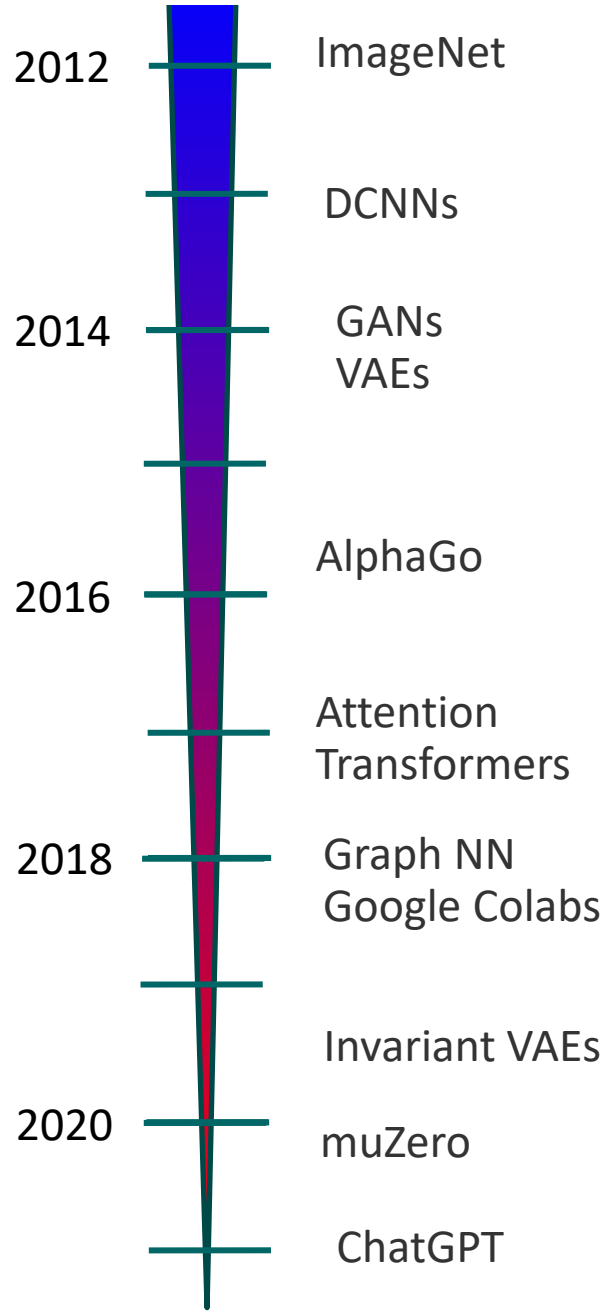
ML of the last decade

- 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….

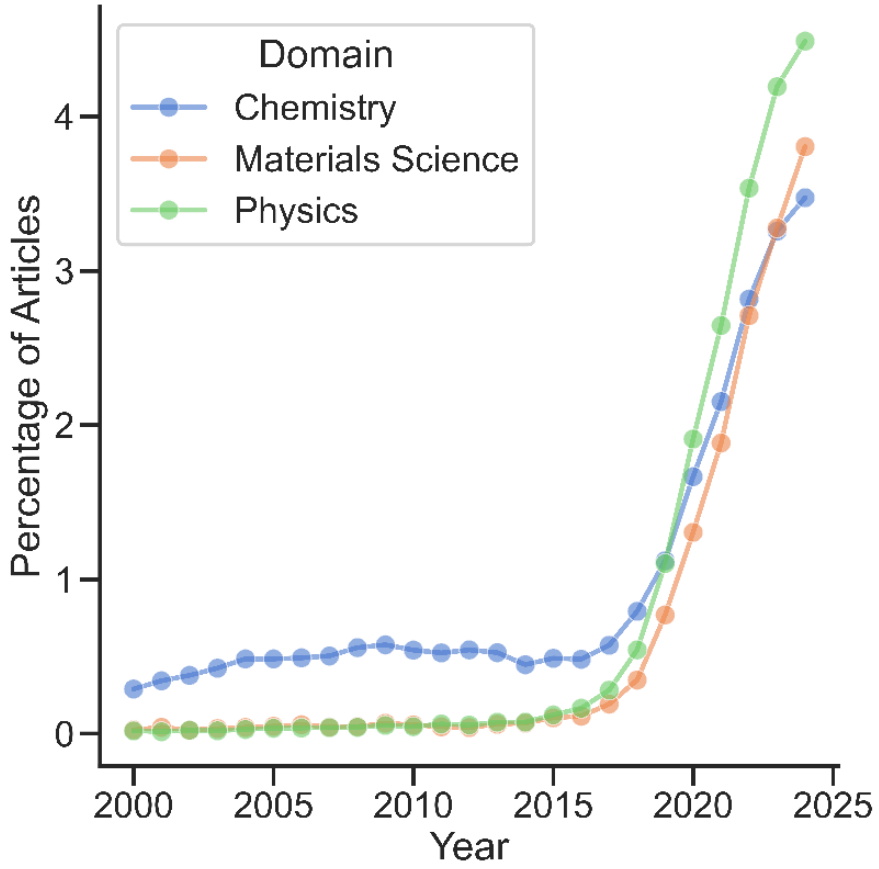
Why is it difficult?

- 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

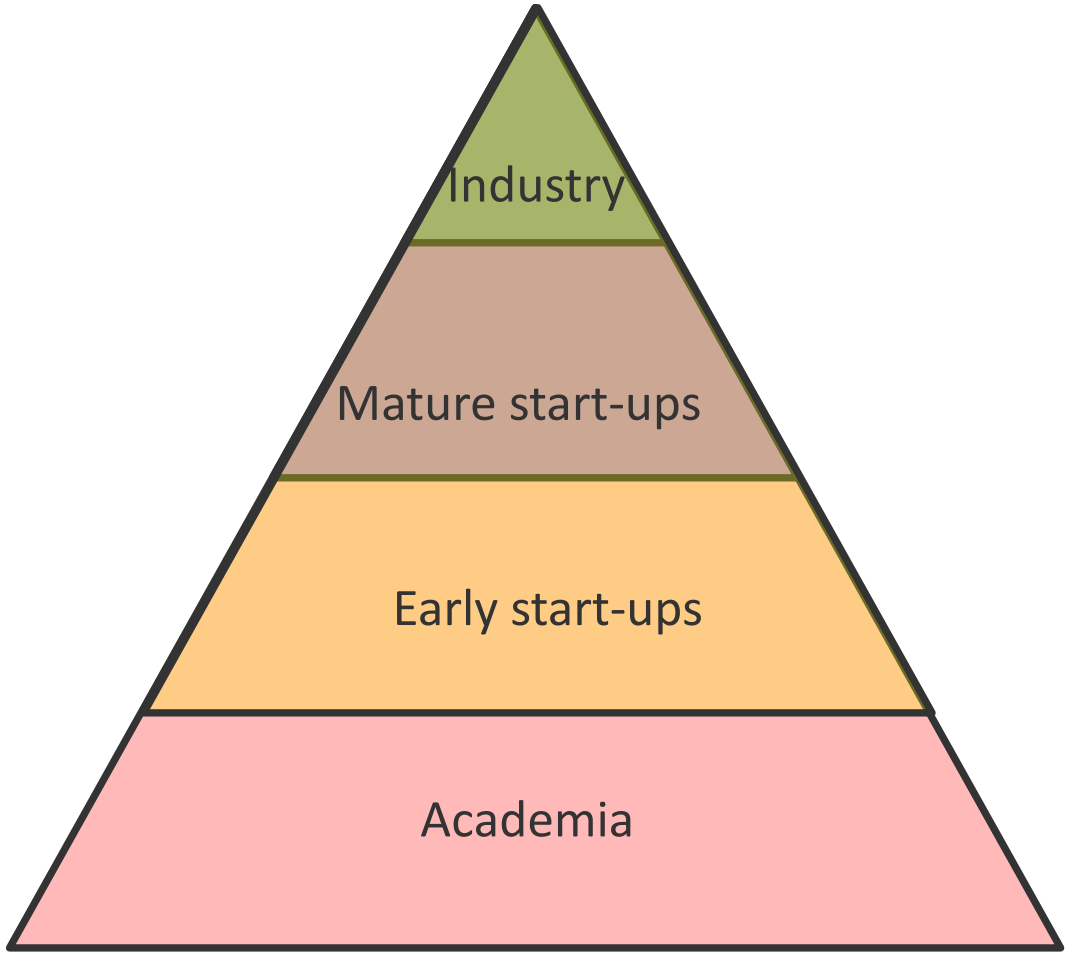
Microsoft: GitHub
Meta: Open Catalyst,
Meta: Papers with Code
Toyota: TRI
Google: AlphaFold
NVIDIA: protein folding



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, **human heuristics**, and strong set of inferential biases

ML for real-world applications is different!

Types of Machine Learning

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

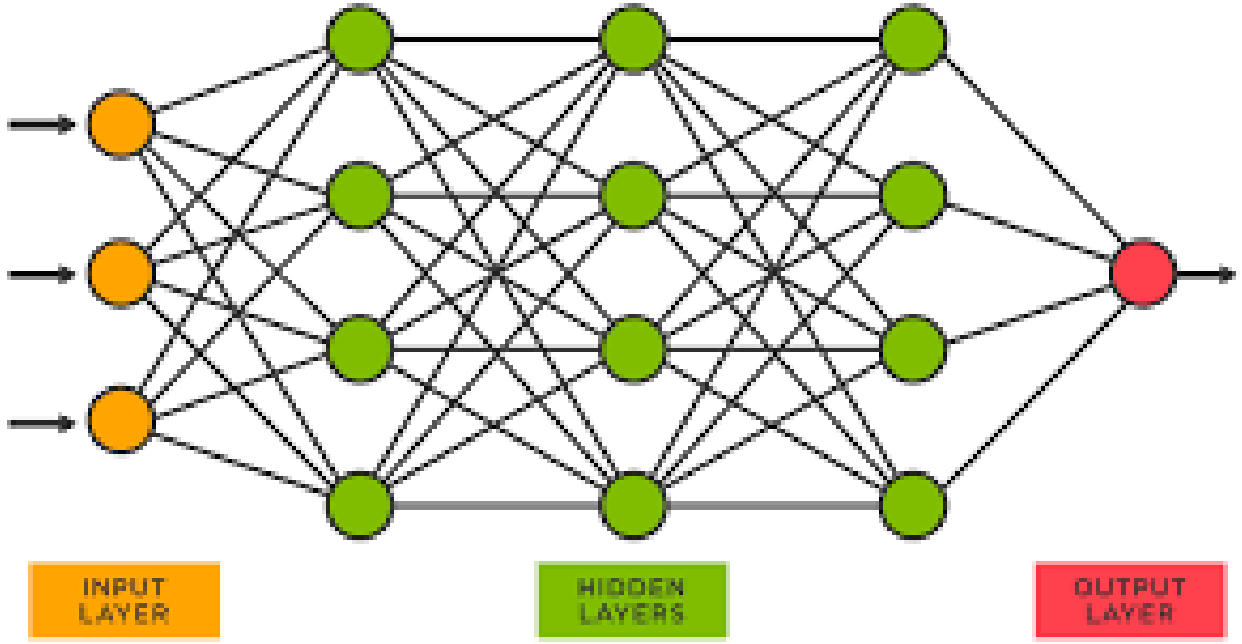
- **Unsupervised learning**
– Given: training data (without desired outputs)

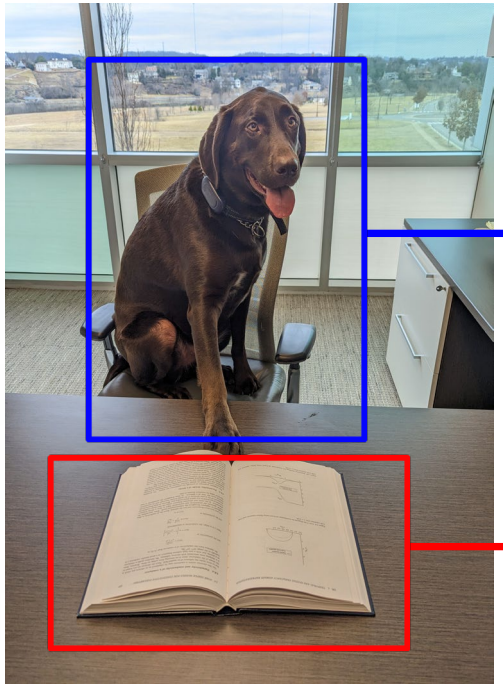
- **Semi-supervised learning**
– Given: training data + a few desired outputs

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

Supervised Machine Learning

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





Dog

Book

Classification

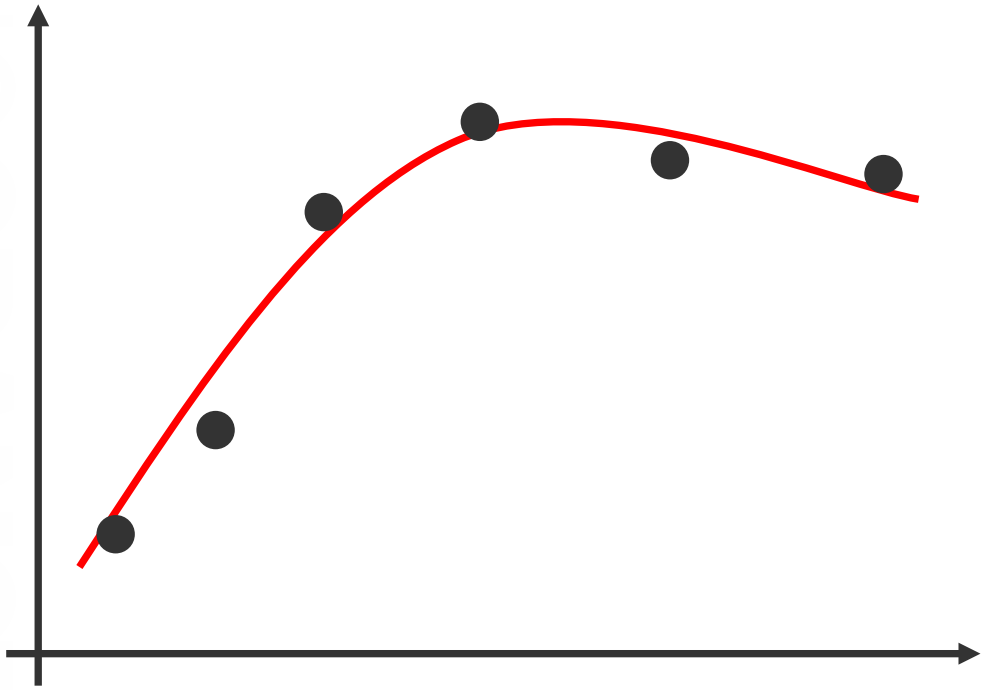
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
- If y is categorical == classification

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

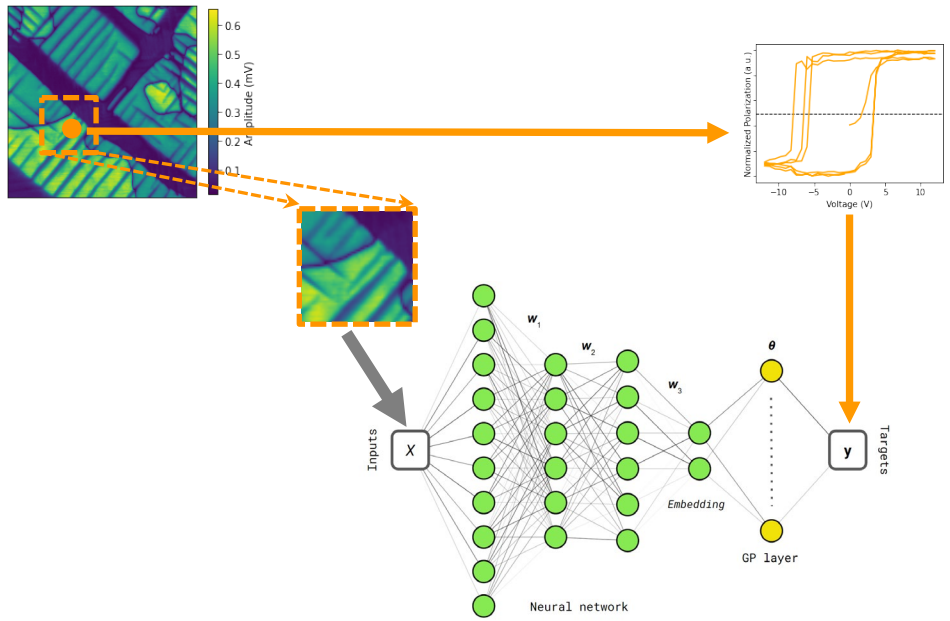
Regression

- 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

Simple regression: $R^1 \rightarrow R^1$

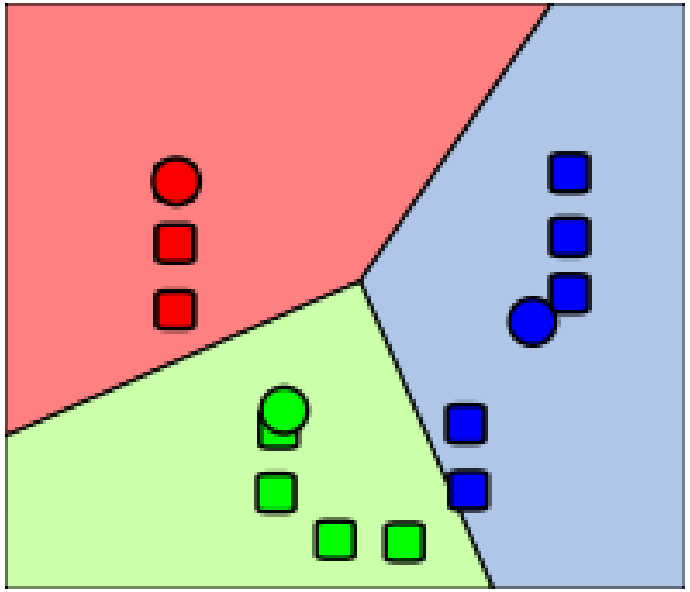


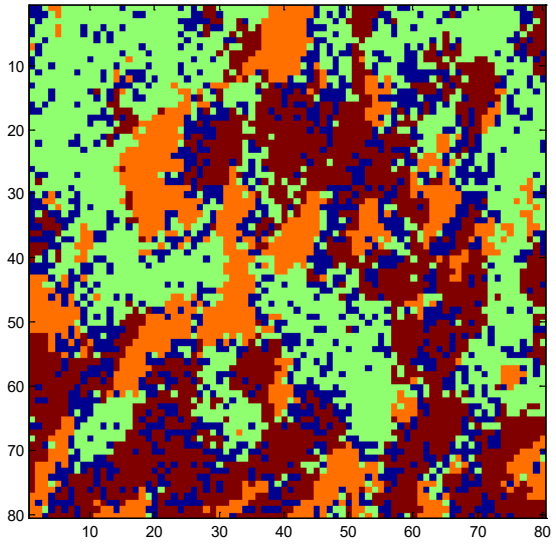
Not so simple regression

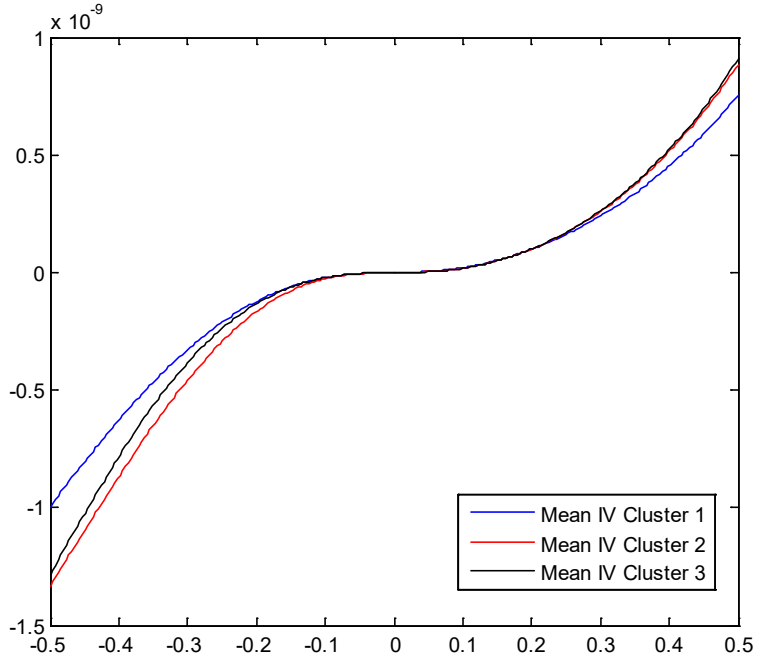


Unsupervised Learning

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



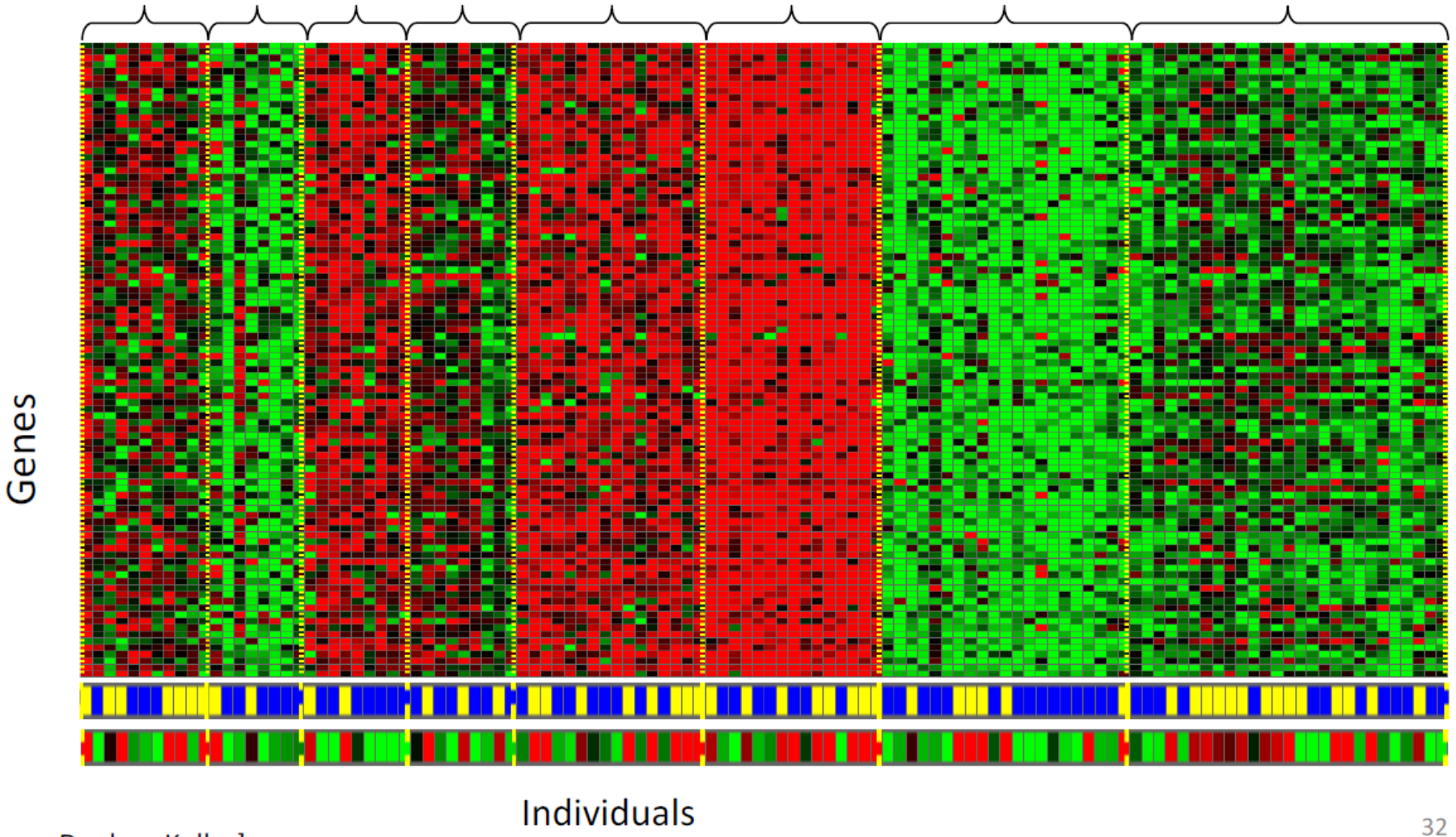




M. ZIATDINOV, A. MAKSOV, L. LI, A. SEFAT, P. MAKSYMOVYCH, and S.V. KALININ, *Deep data mining in a real space: Separation of intertwined electronic responses in a lightly-doped* $BaFe_2As_2$, Nanotechnology **27**, 475706 (2016).

Unsupervised Learning

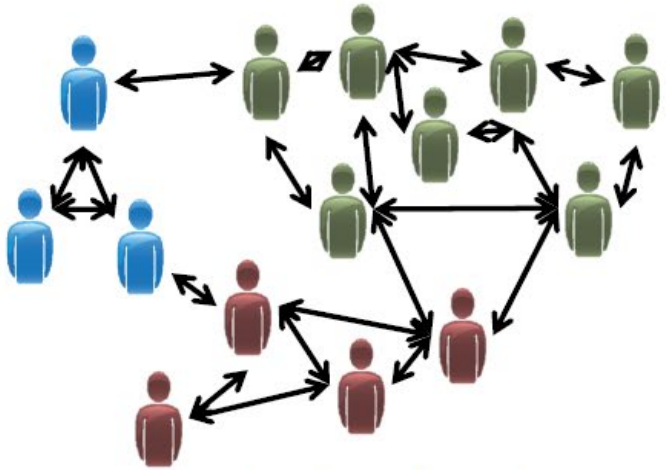
Genomics application: group individuals by genetic similarity



[Source: Daphne Koller]

Unsupervised Learning

Organize computing clusters



Social network analysis

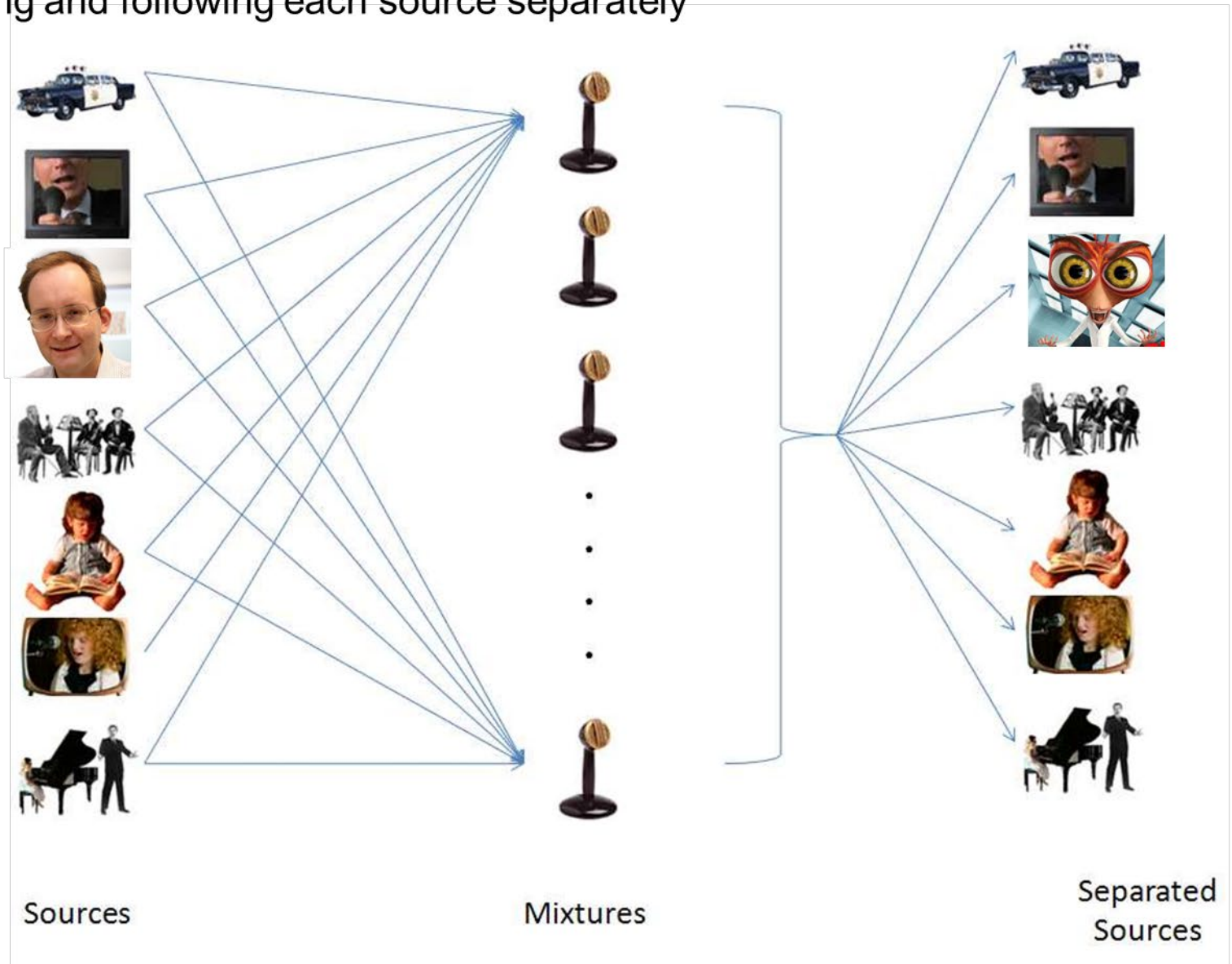
Market segmentation

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

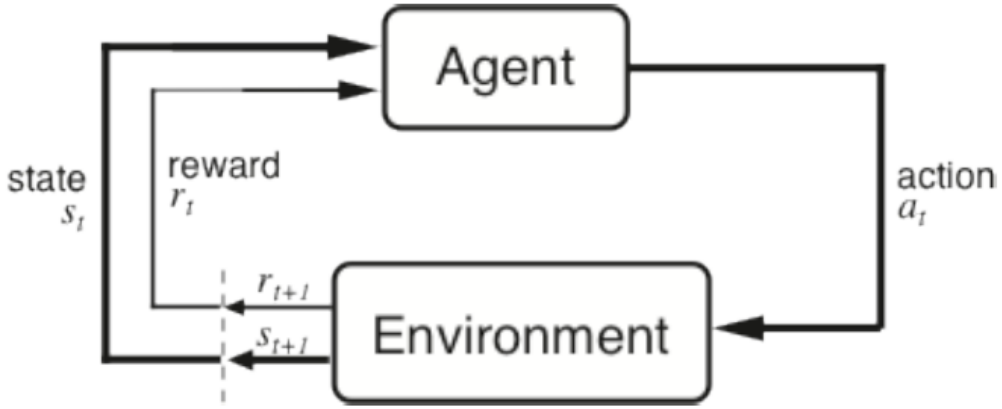


Reinforcement Learning

Given a sequence of **states** and **actions** with (delayed) **rewards**, output a policy, i.e. 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



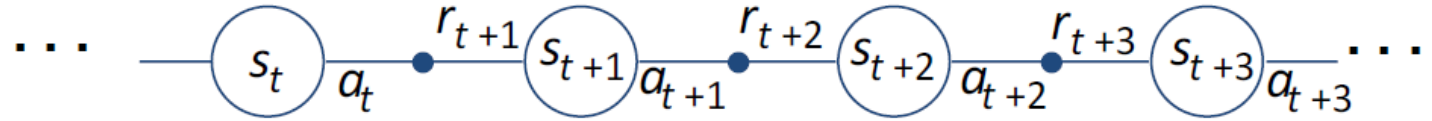
Agent and environment interact at discrete time steps : $t = 0, 1, 2, \mathrm{K}$

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

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

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

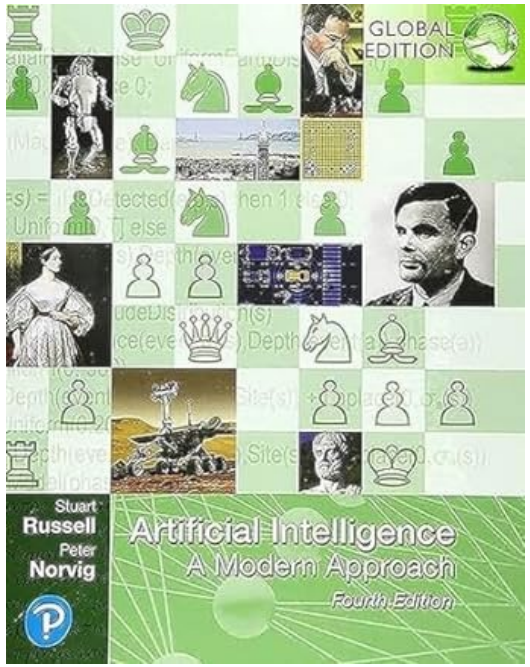
and resulting next state : s_{t+1}



Slide credit: Sutton & Barto

Reinforcement Learning in Action

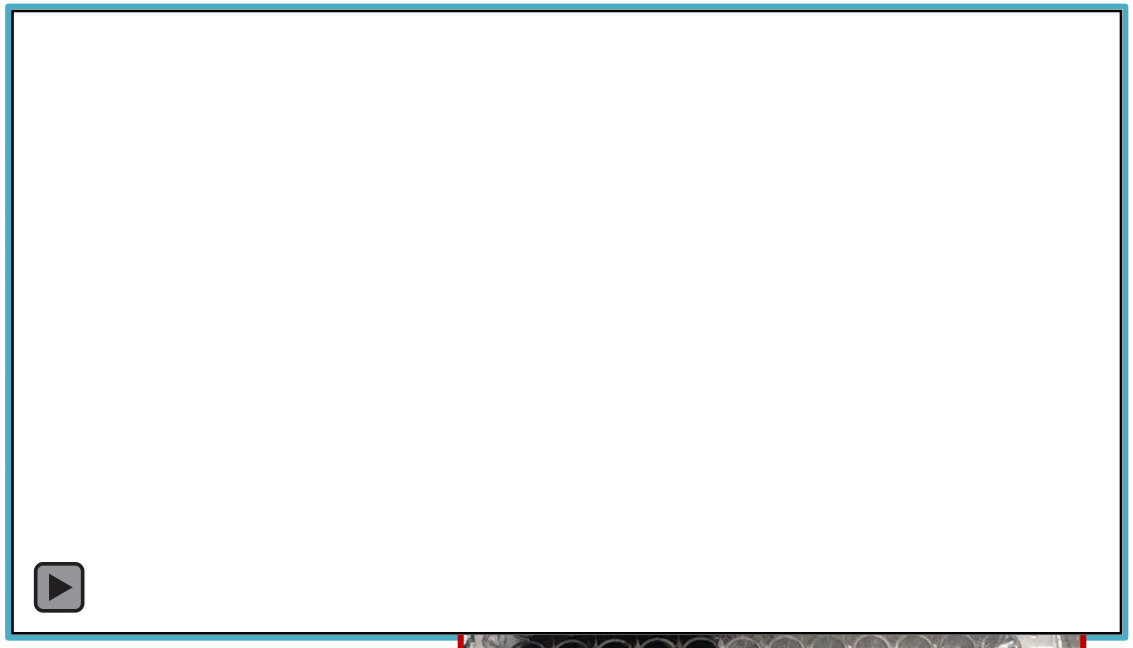
https://www.youtube.com/watch?v=GtYlVxv0py8



Somewhat remarkably, almost all AI research until very recently has assumed that the performance measure can be exactly and correctly specified in the form of utility or reward function

Reinforcement Learning Applications

Chemical Synthesis and Drug Discovery



M. Ahmadi lab
UTK MSE

Cloud Laboratories

Emerald Cloud Lab,
SF and CMU