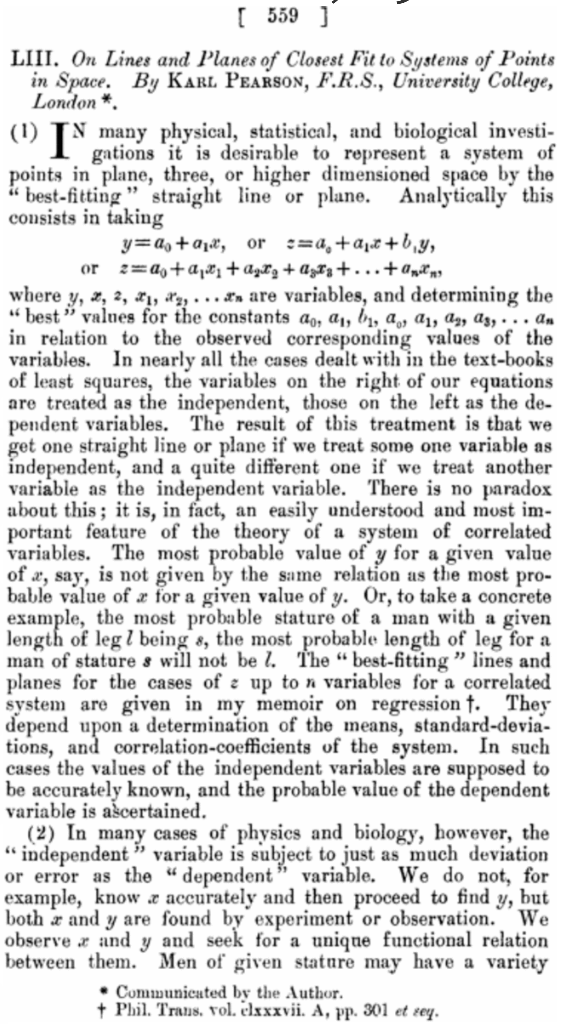
THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 02: History of ML and Scientific Data

Instructor: Sergei V. Kalinin

Zooming out on history

K. Pearson, 1901



[559]

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

$$y=a_0+a_1x, \quad \text{or} \quad z=a_0+a_1x+b_1y,$$

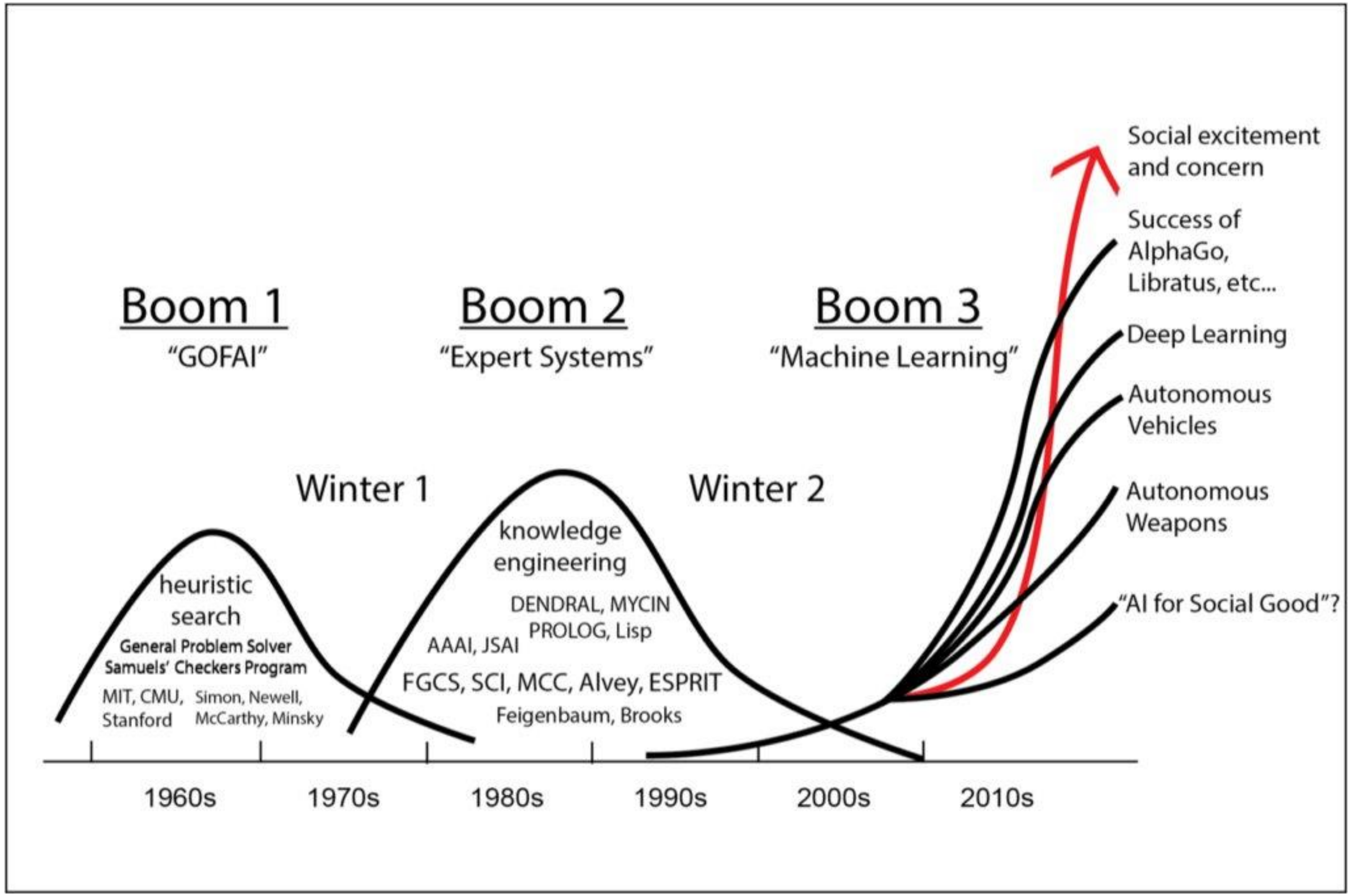
$$\text{or} \quad z=a_0+a_1x_1+a_2x_2+a_3x_3+\ldots+a_nx_n,$$

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

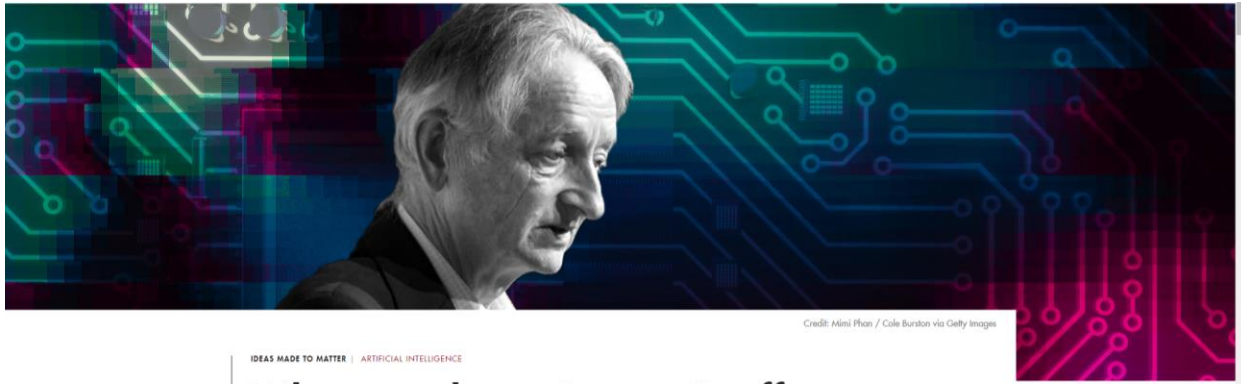
Taking the Human Out of the Loop: A Review of Bayesian Optimization

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

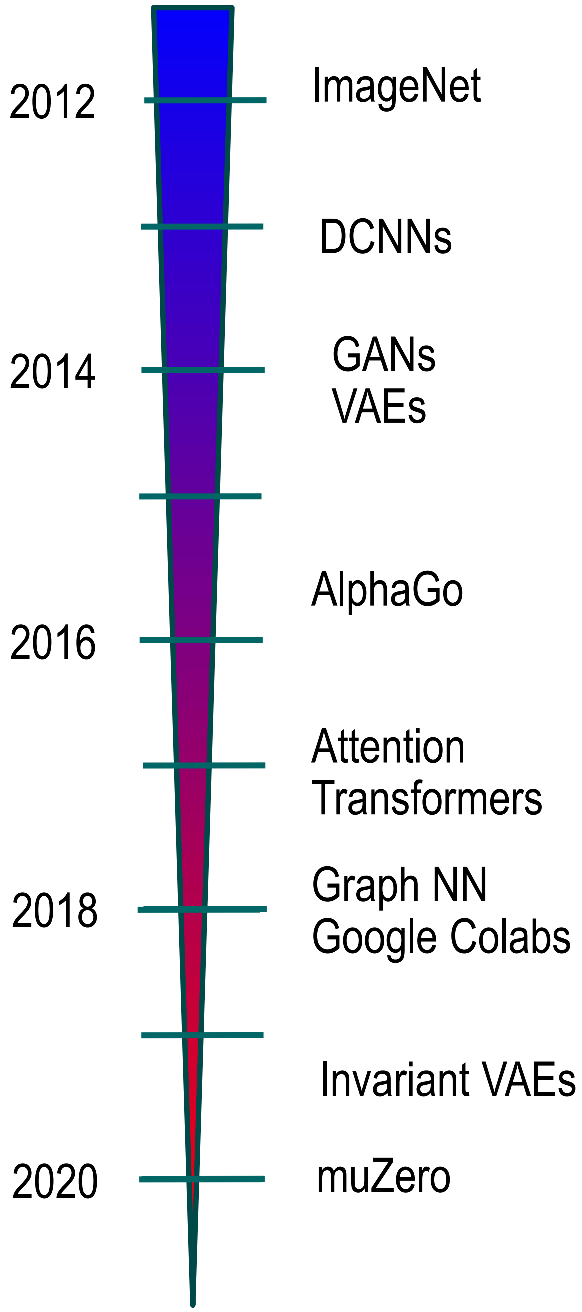
ML of the last decade

- 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….

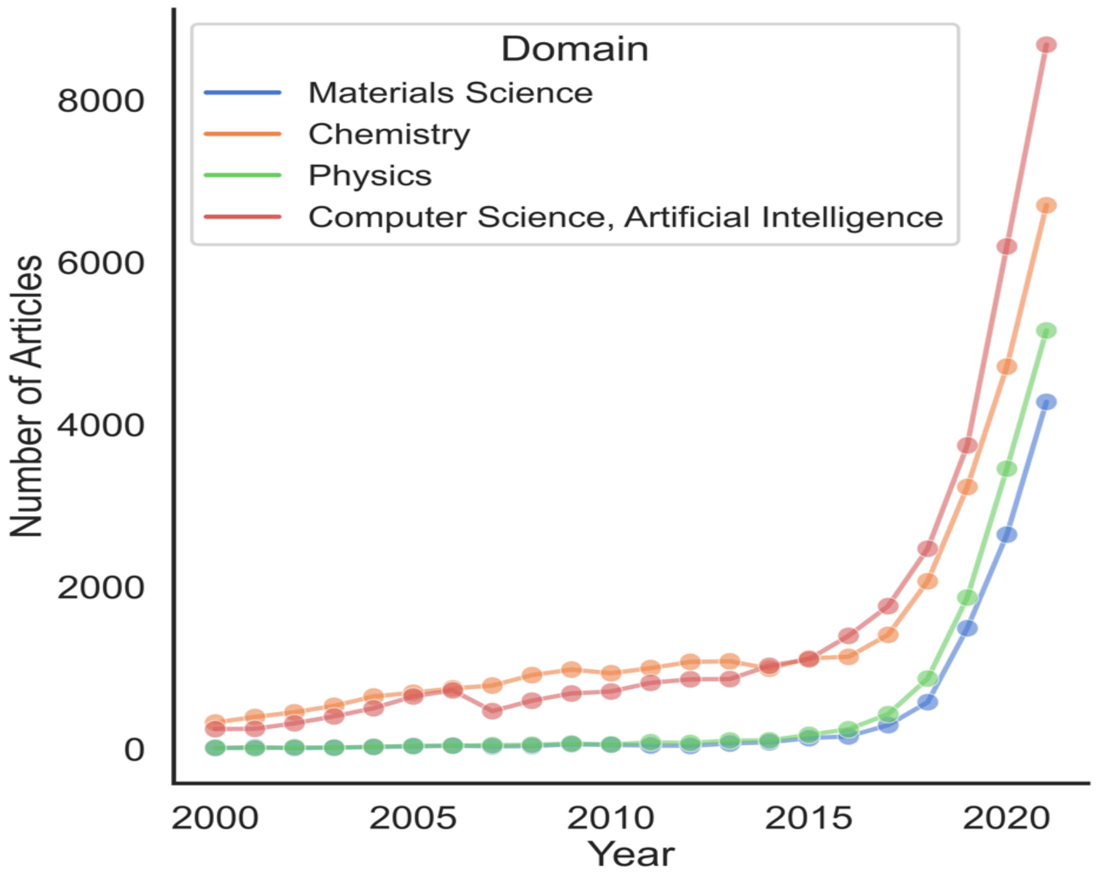
Why is it difficult?

- 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

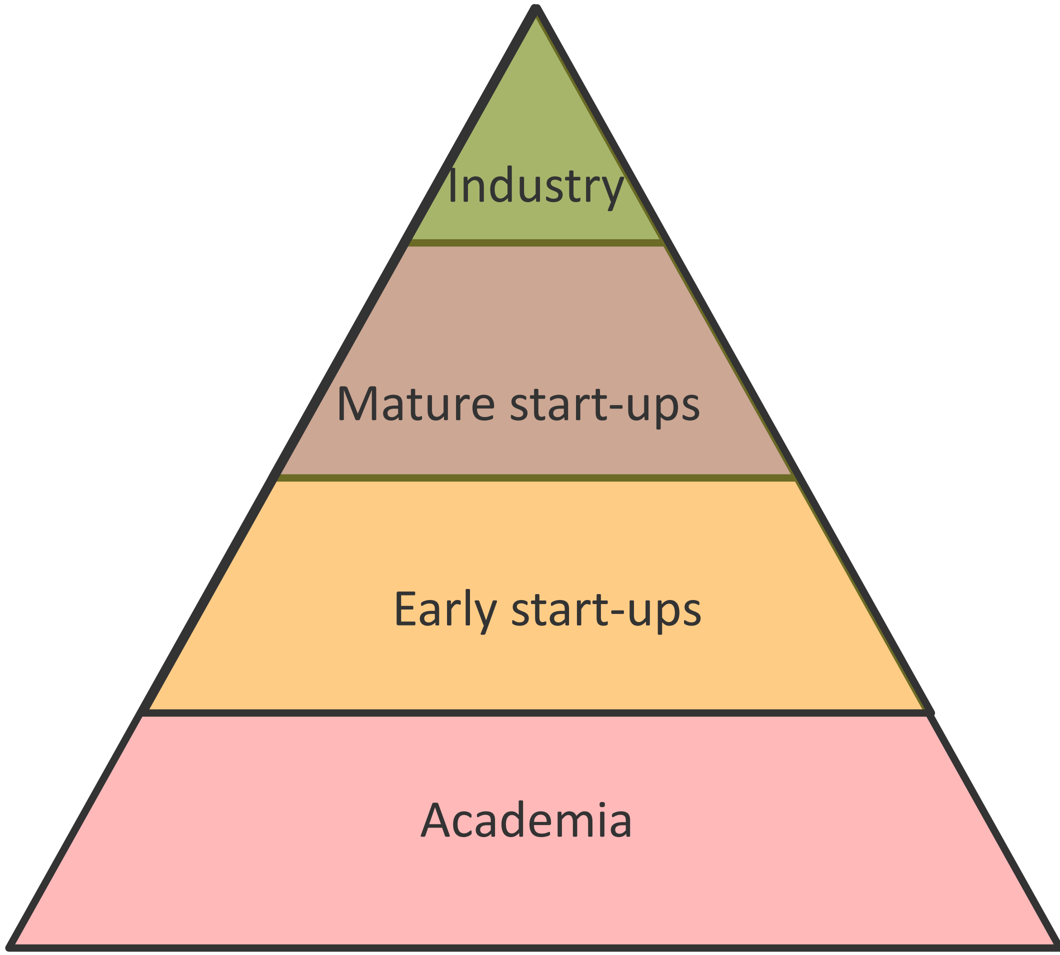
Microsoft: GitHub
Meta: Open Catalyst,
Meta: Papers with Code
Toyota: TRI
Google: AlphaFold
NVIDIA: protein folding



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)

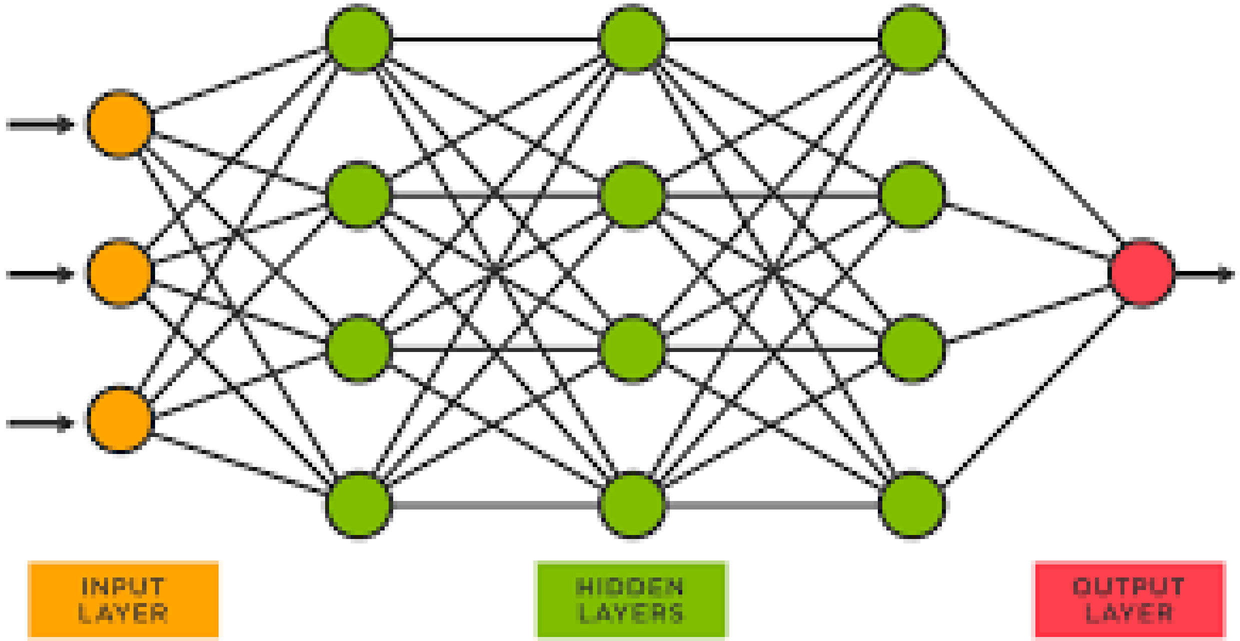
- **Unsupervised learning**
– Given: training data (without desired outputs)

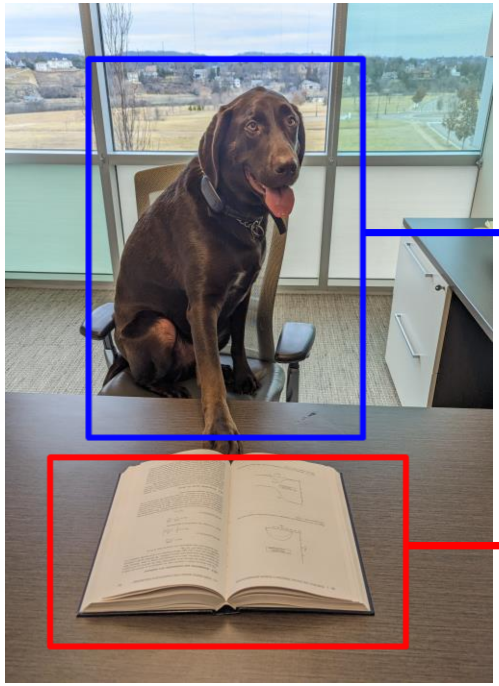
- **Semi-supervised learning**
– Given: training data + a few desired outputs

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

Supervised Machine Learning

- Regression
- Classification
- Semantic segmentation
- Instance segmentation
- ...





Dog

Book

Classification

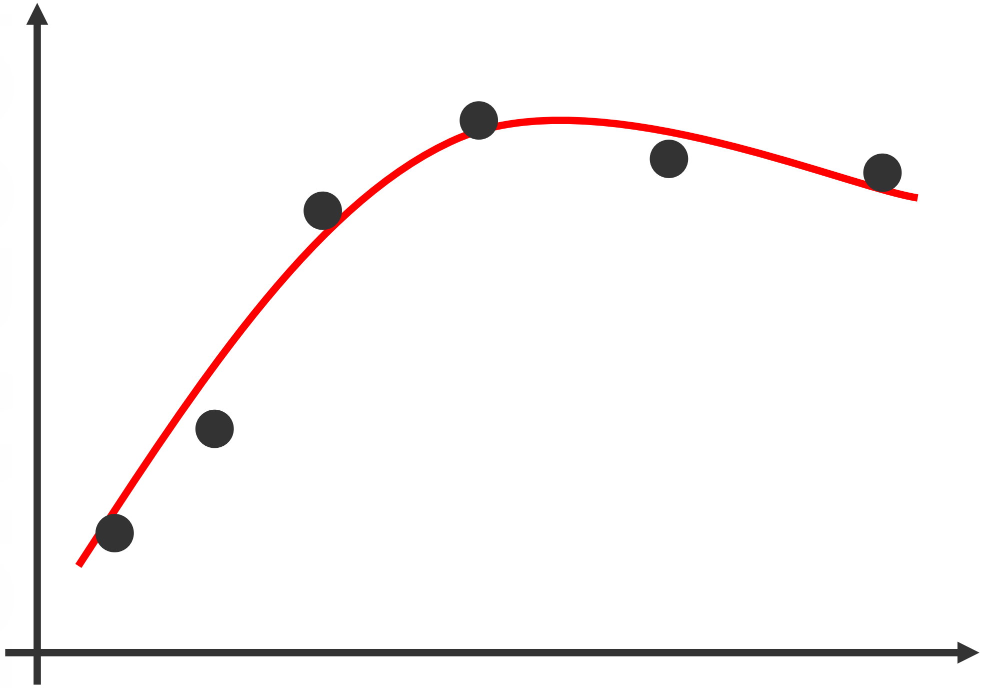
- Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
- If y is categorical == classification

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

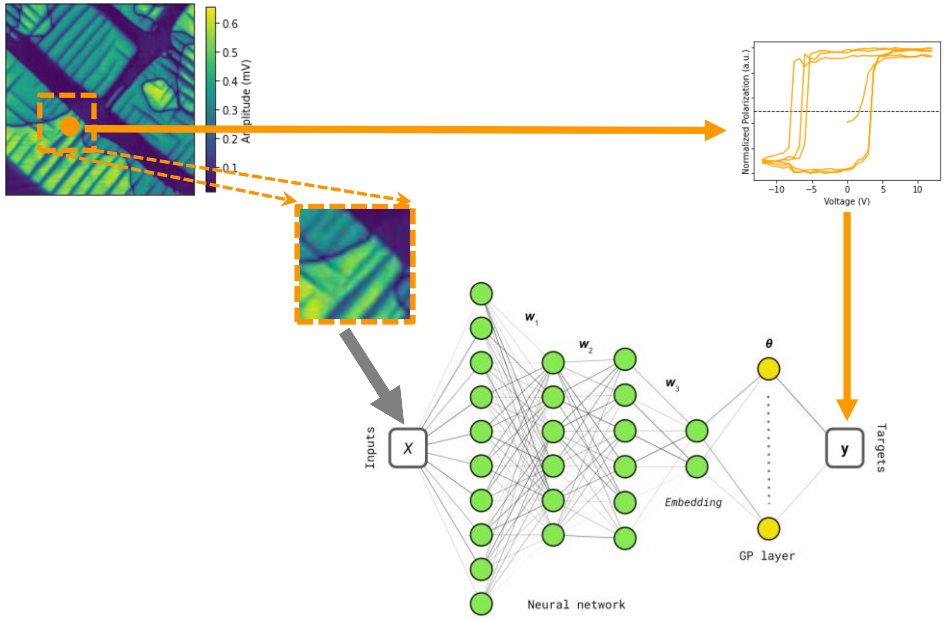
Regression

- 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: $R^1 \rightarrow R^1$

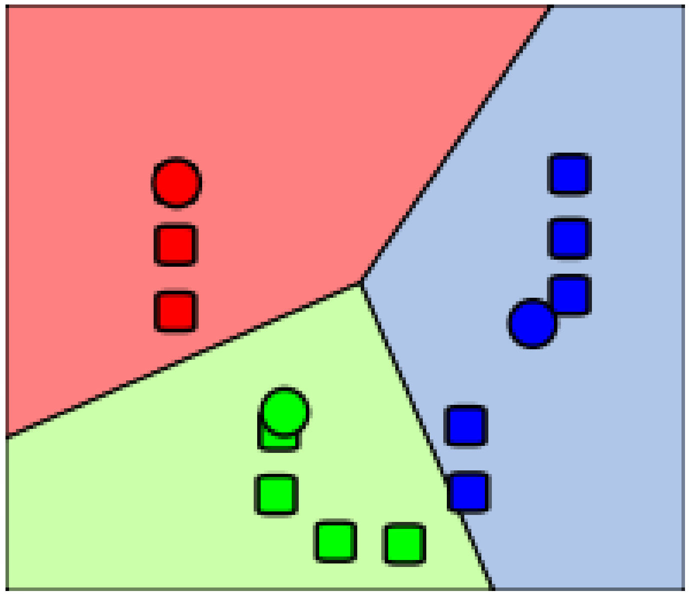


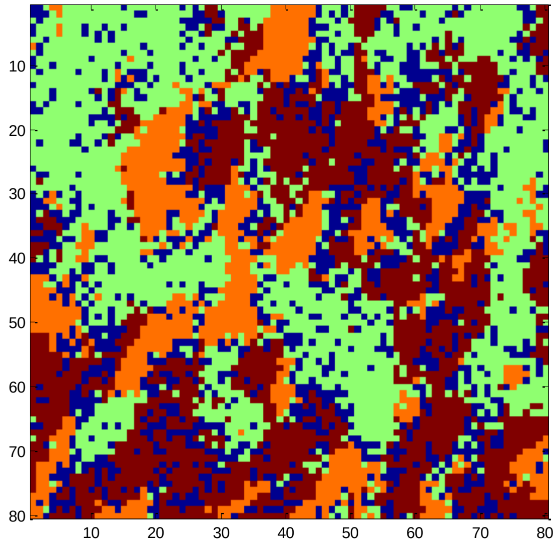
Not so simple regression

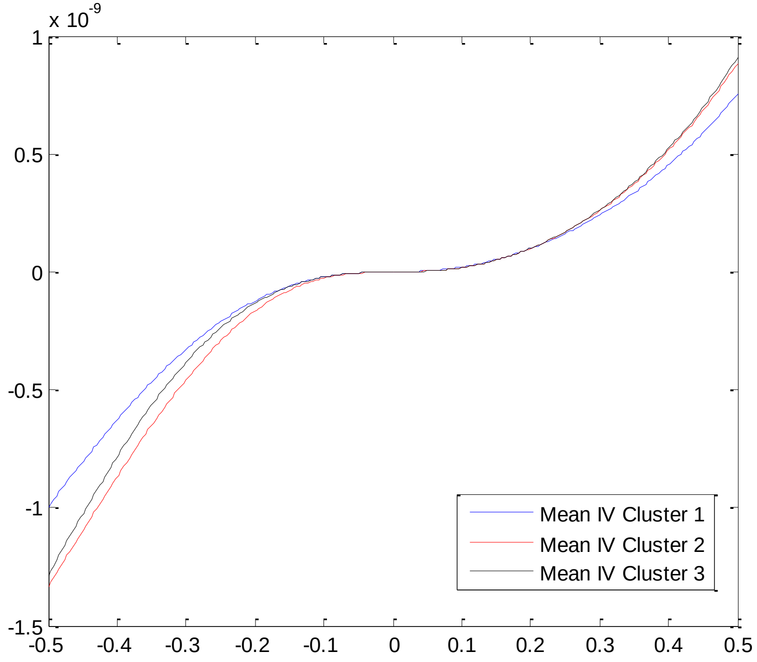


Unsupervised Learning

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



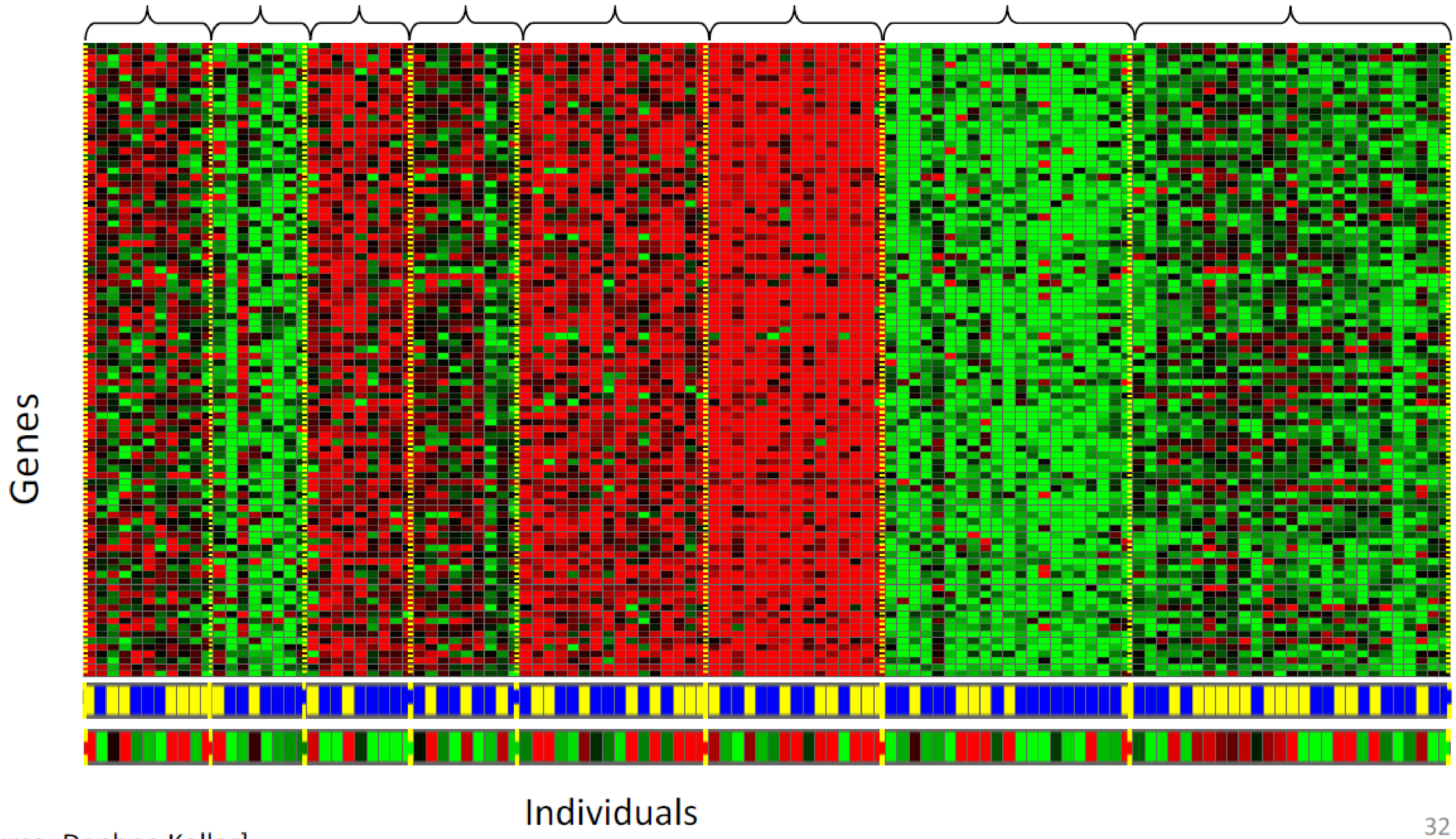




M. Ziatdinov, A. Maksov, L. Li, A. Sefat, P. Maksymovych, and S.V. Kalinin, *Deep data mining in a real space: Separation of intertwined electronic responses in a lightly-doped* $BaFe_2As_2$, Nanotechnology **27**, 475706 (2016).

Unsupervised Learning

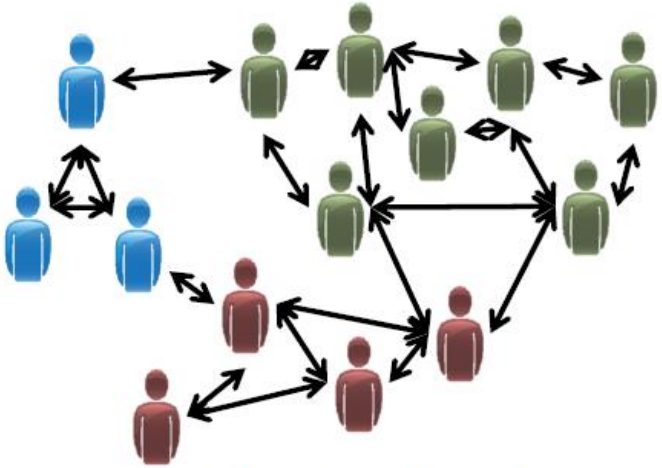
Genomics application: group individuals by genetic similarity



[Source: Daphne Koller]

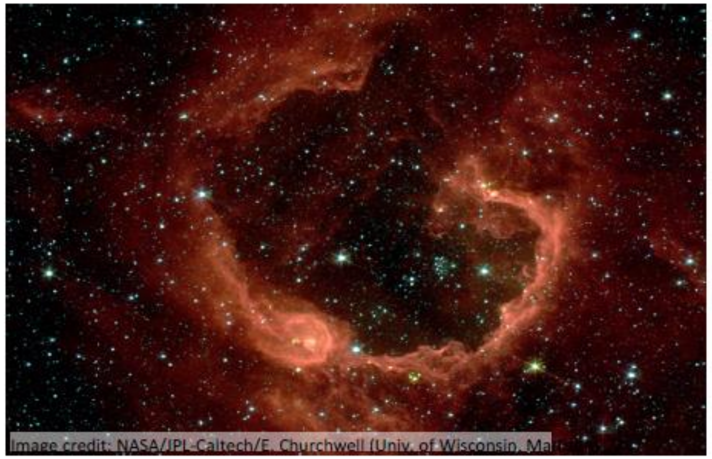
Unsupervised Learning

Organize computing clusters



Social network analysis

Market segmentation

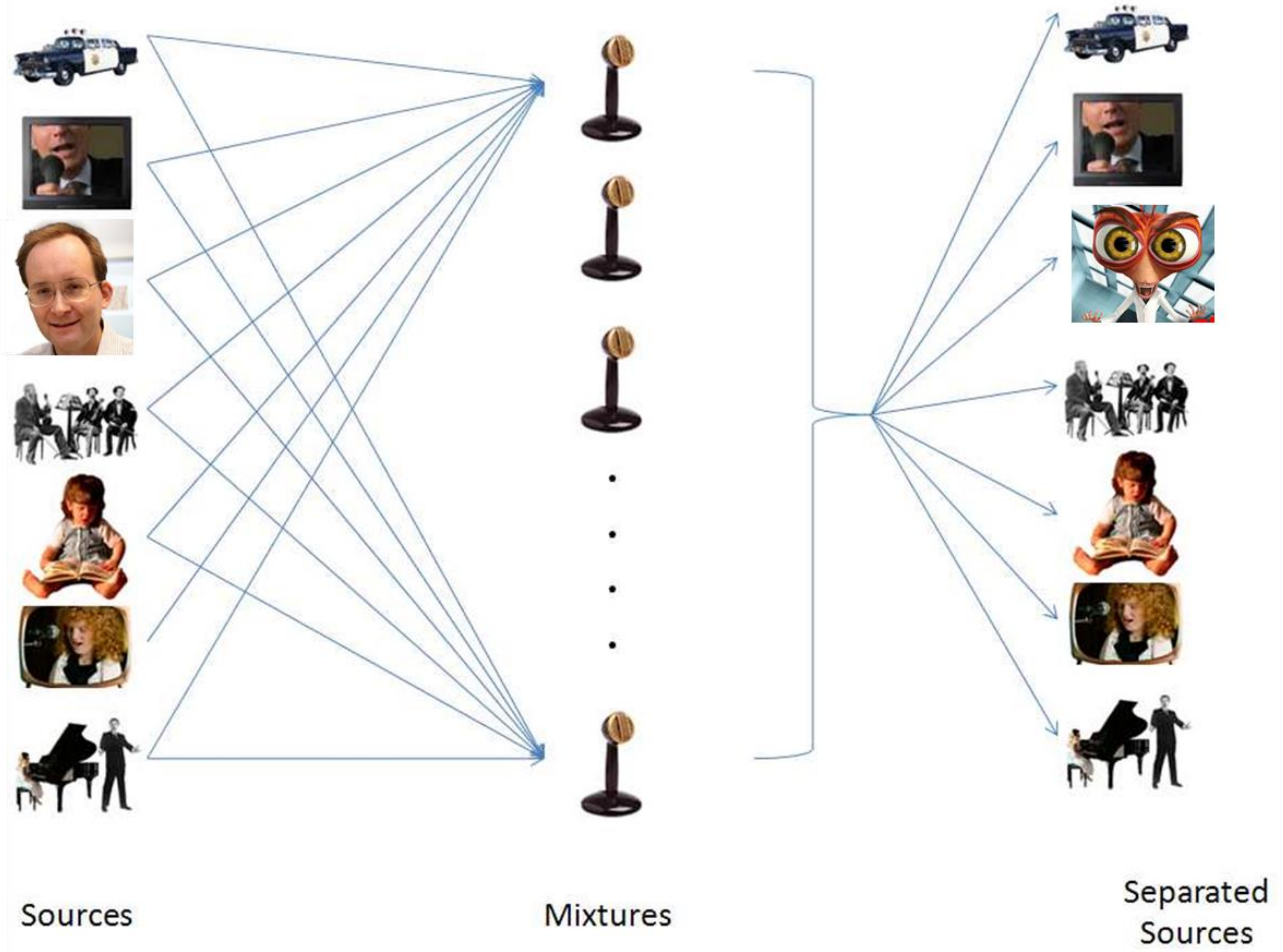


Astronomical data analysis

Slide credit: Andrew Ng

Unsupervised Learning

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

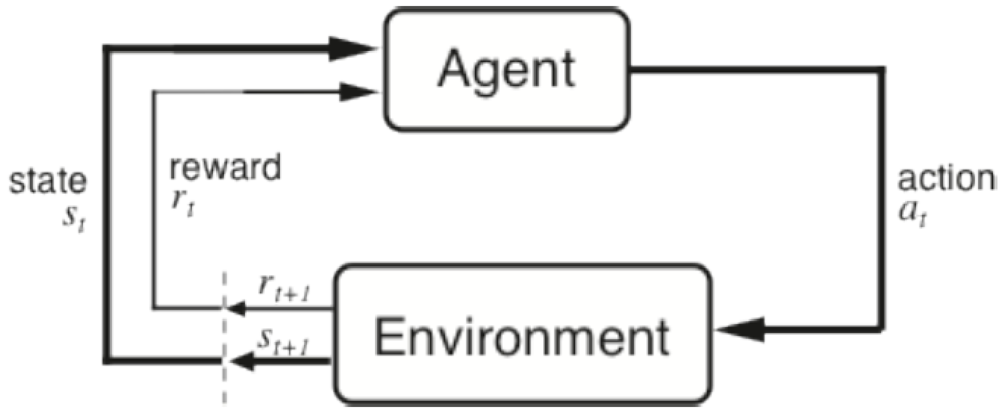


Reinforcement Learning

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



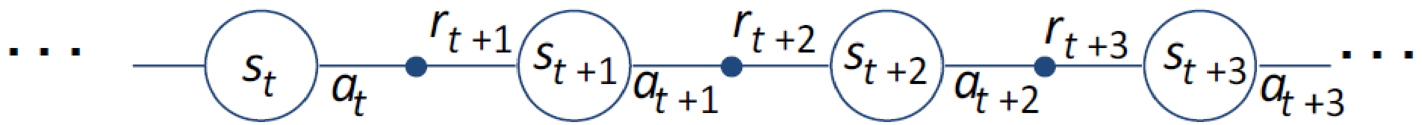
Agent and environment interact at discrete time steps : $t = 0, 1, 2, K$

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

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

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

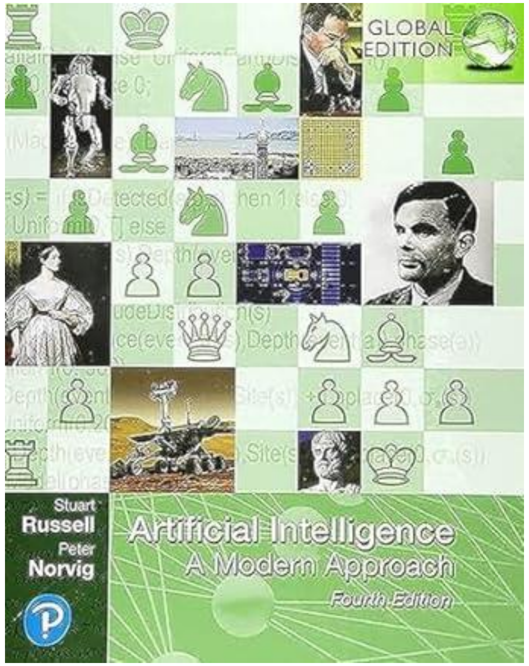
and resulting next state : s_{t+1}



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

Approach

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:

Build the case for machine learning

- Explore the business/scientific problem
- Build workflow (operations, costs, latencies)
- Identify bottlenecks
- Chart the solution
- Prototype the solution
- Test and iterate
- Deploy
- Support
- Upgrade
- Sunset

Homework assignment 2 (part 1)

Identify the type of problem

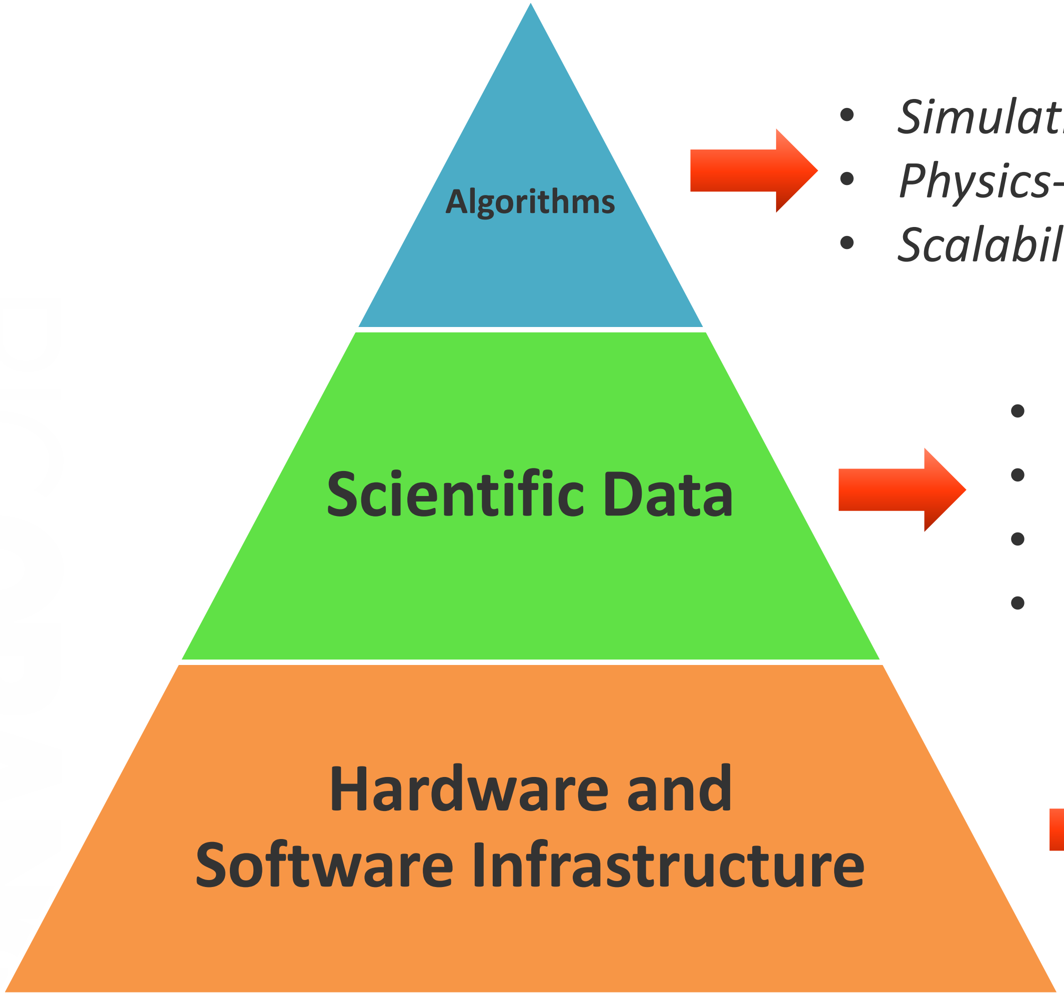
Supervised (inductive) learning
– Given: training data + desired outputs (labels)

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

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

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

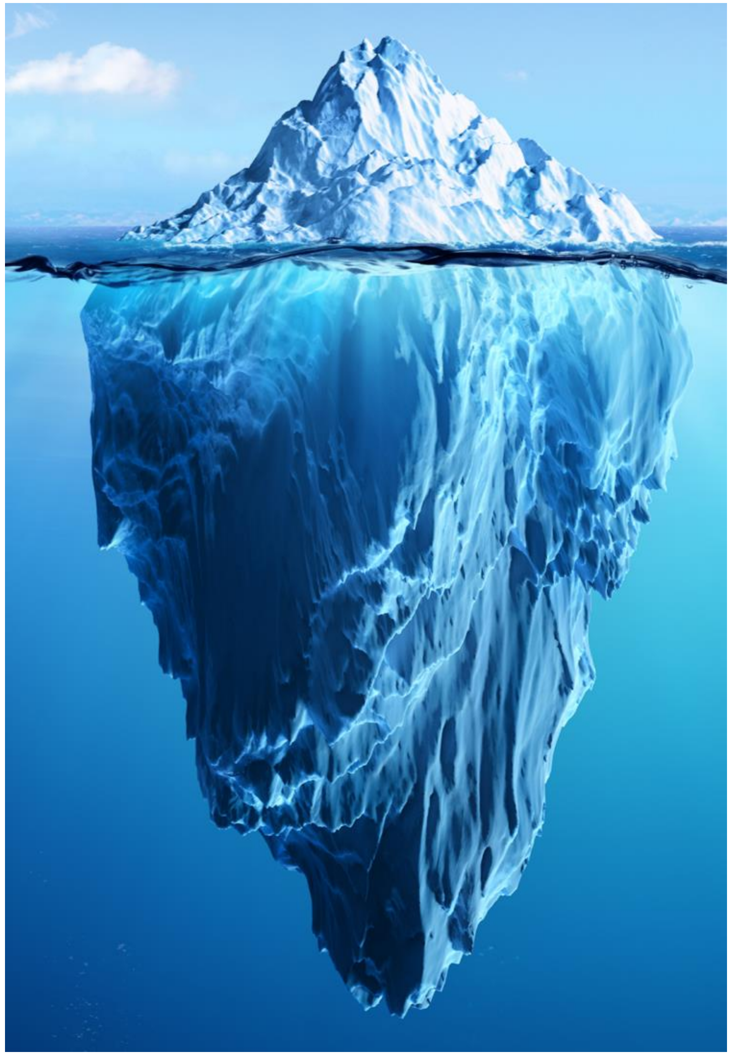
The pyramid of machine learning



- *Simulations*
- *Physics-informed methods*
- *Scalability*

- *Curated*
- *Searchable*
- *Standardized*
- *Best Practices*

- *Codes*
- *Databases*
- *Computing*



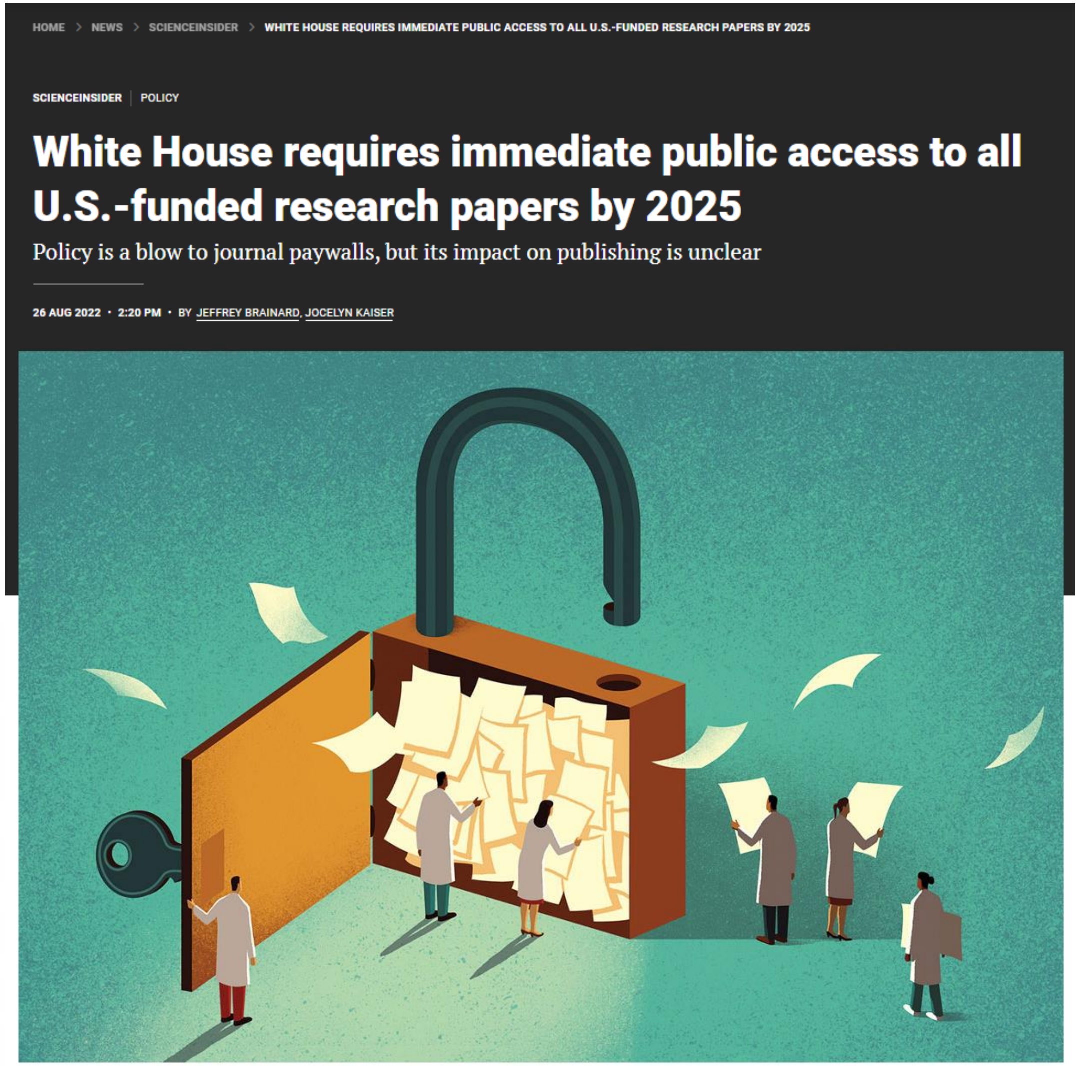
Slide courtesy of R. Vasudevan

Why bother with infrastructure?

- Almost all of machine learning relies extensively on having access to good quality data
- In most laboratories, this data is acquired via multiple instruments in different formats, and not findable or accessible, and often lacks necessary metadata for ML labeling
- As such, in many cases, ML in science is impossible especially in the experimental domains, without the necessary investments in data standardization and storage
- Similarly, reproducibility of workflows relies on strongly tested codebases, not one-off scripts.

Slide courtesy of R. Vasudevan

Soon to be a requirement!

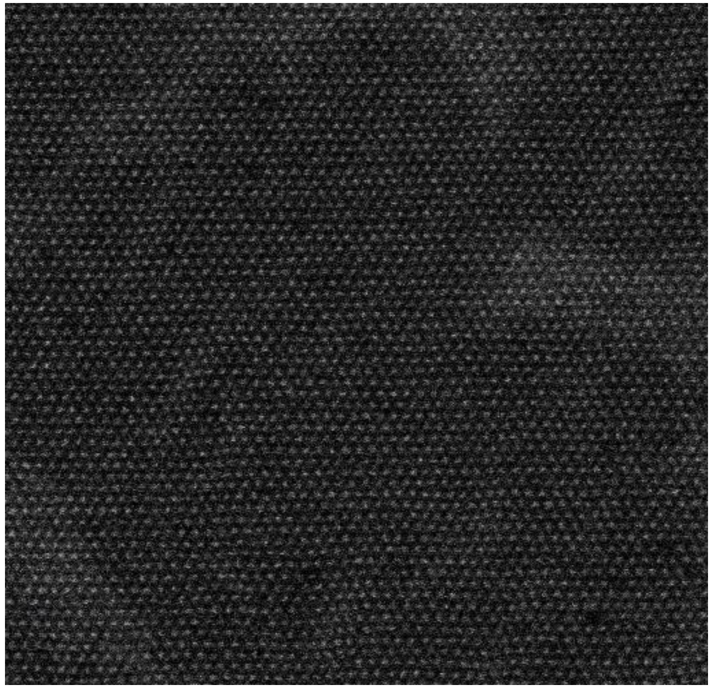


- Data management plans in proposals
- Repositories of data and codes associated with publications
- Good to be ready!

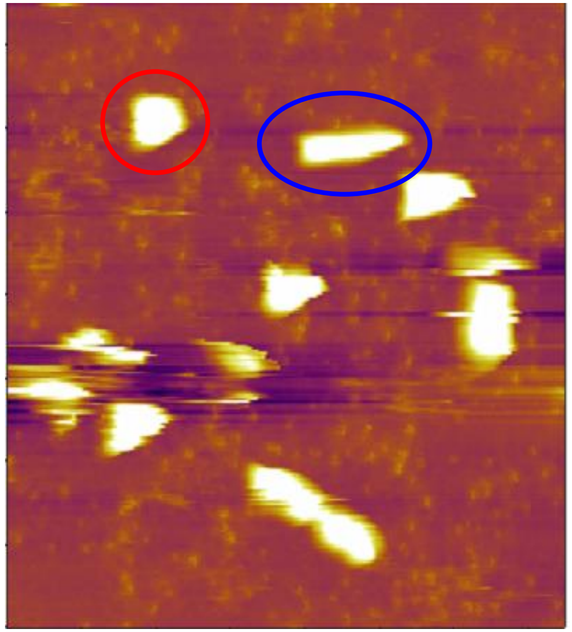
https://www.science.org/content/article/white-house-requires-immediate-public-access-all-u-s--funded-research-papers-2025

Get data – from scientific tools

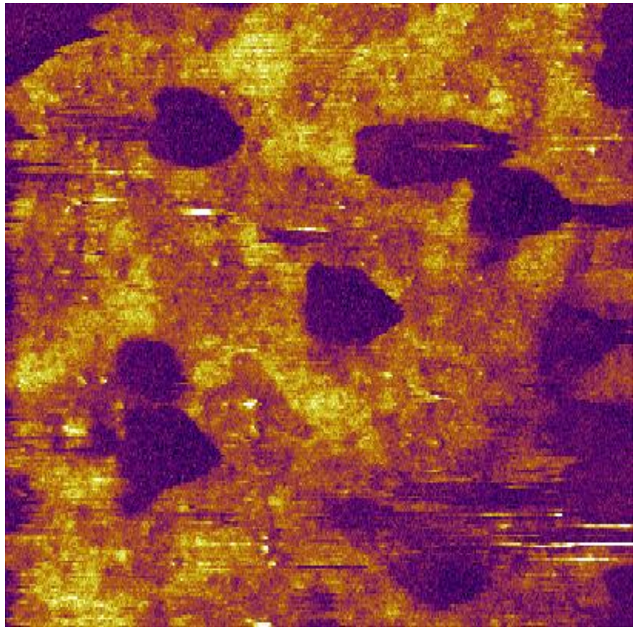
- Spectra
- (Multimodal) Images
- Hyperspectral images
- Videos
- Time traces
- …

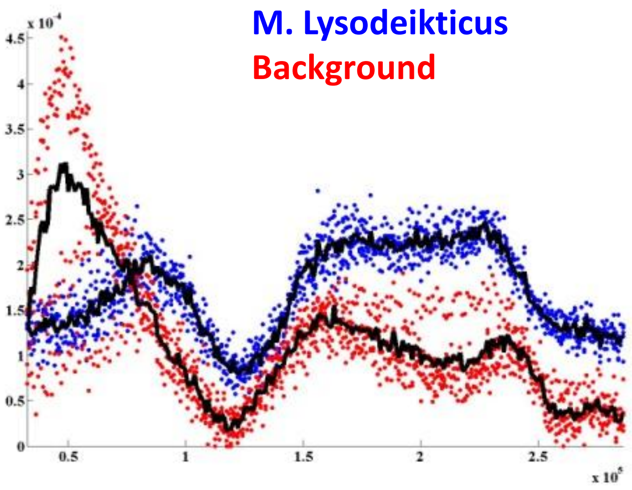


Topography



PFM Amplitude





As scientists, we rarely have to deal with the classical ELT (Extract-Load Transform, aka Data Wrangling) problems. But....

Standardization of Microscope Data

Micro Raman Microscope

Atomic Force Microscope (AFM)

AFM with Infrared spectroscopy (AFM-IR)

Scanning Tunneling Microscope (STM)

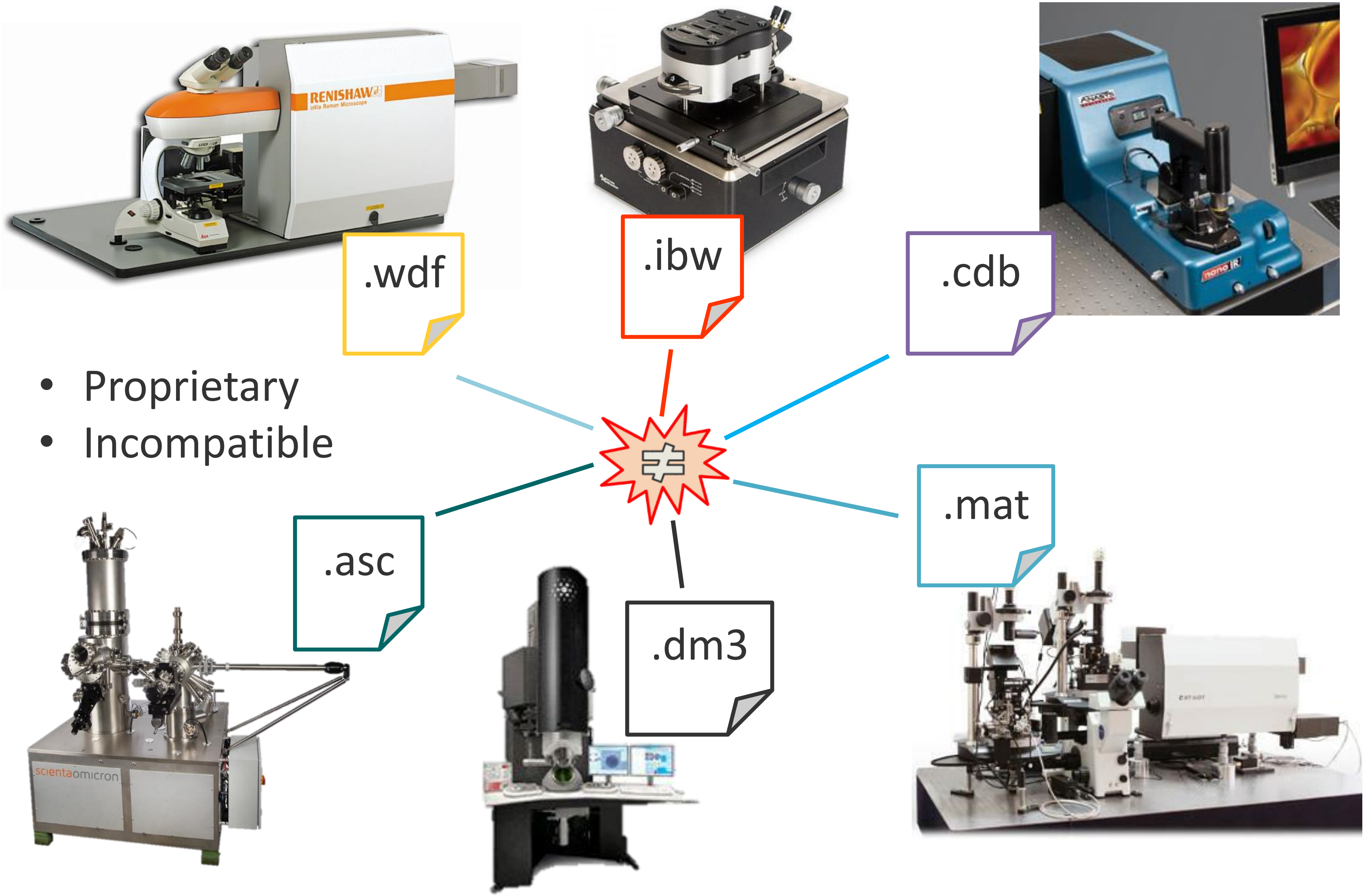
Scanning Transmission Electron Microscope (STEM)

AFM with Raman spectroscopy

Slide by S. Somnath

Multitude of File Formats

- Proprietary
- Incompatible



Disjointed communities....

- Clustering
- Fit spectra …

- Filter Image
- Register Image …

- Fit Spectra
- SVD Filtering …

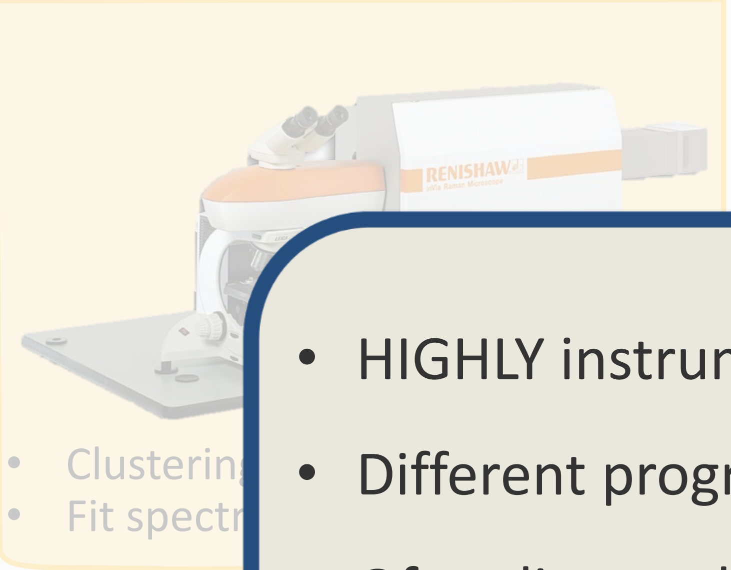
- FFT Filtering
- SVD Filtering …

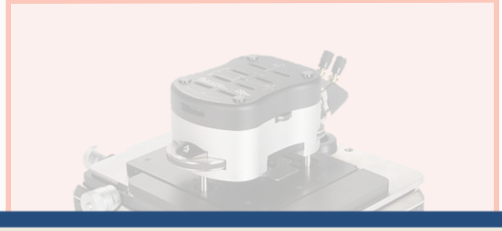
- FFT Filtering
- Classify Images …

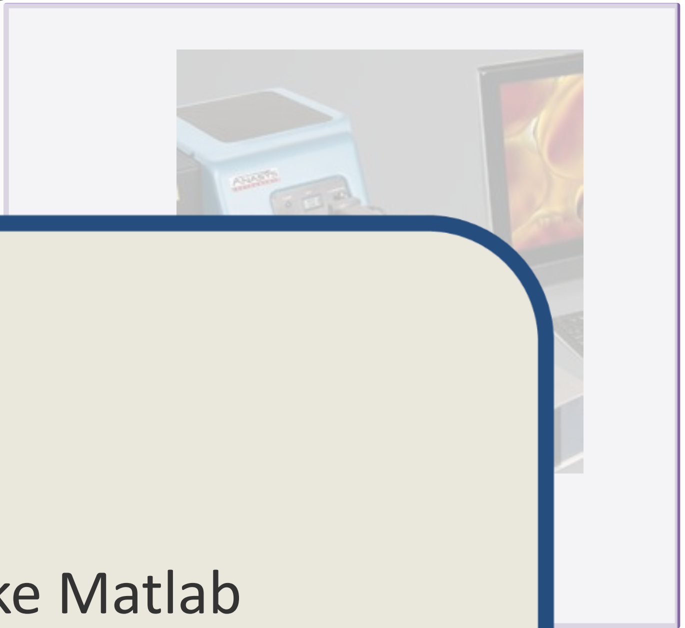
- Register Images
- Clustering

Slide by S. Somnath

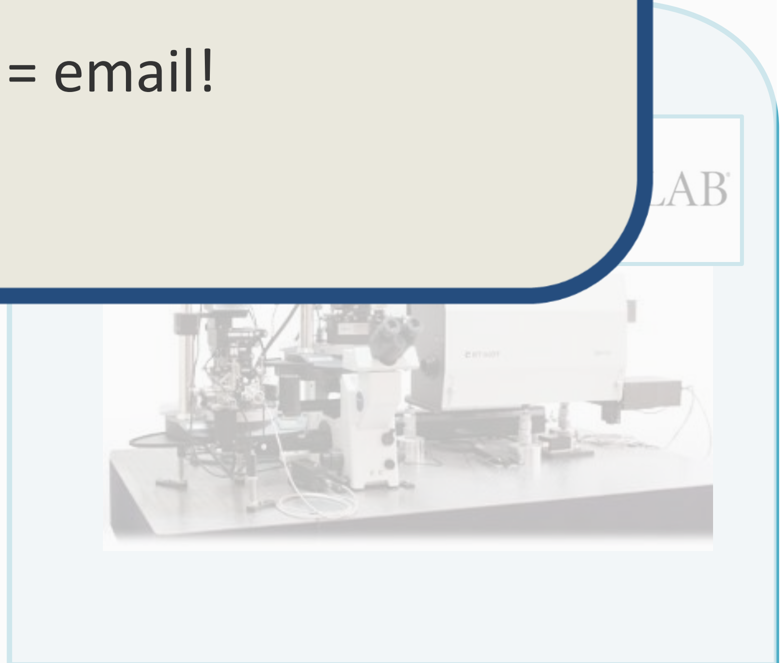
… cannot share code efficiently





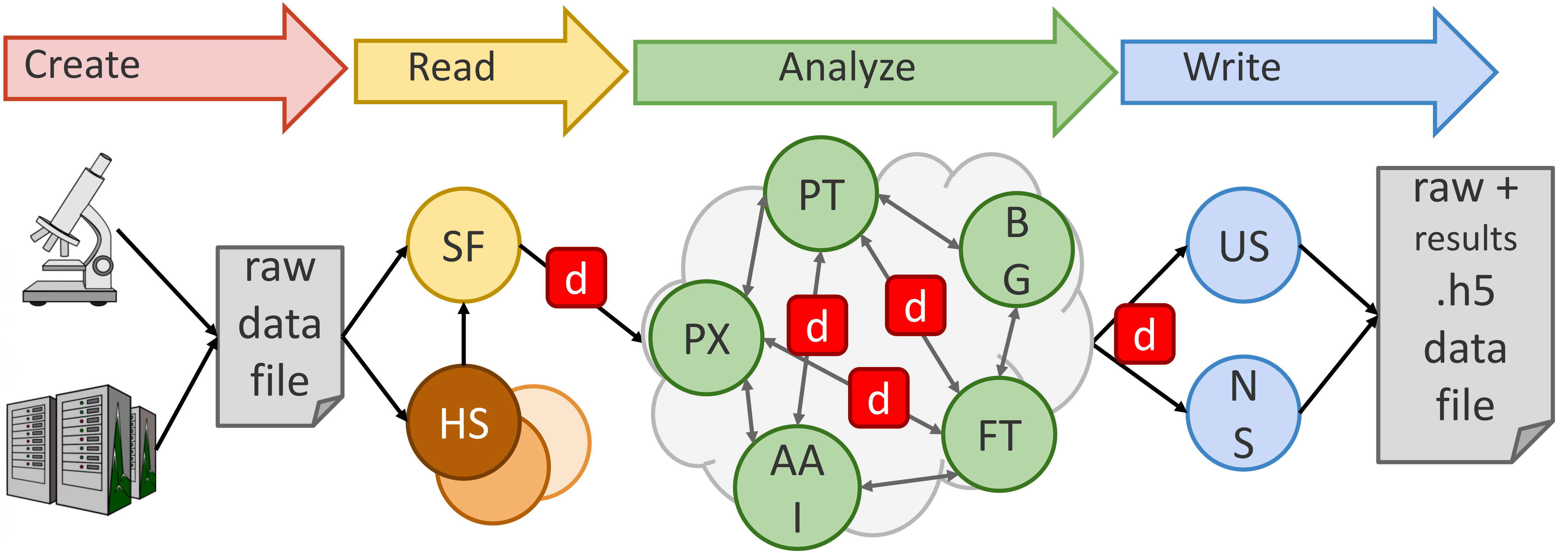


- HIGHLY instrument-specific code
- Different programming languages
- Often licensed / costly software like Matlab
- Most popular sharing method = email!
- No centralized repository



Slide by S. Somnath

Solutions: Integrated Ecosystems

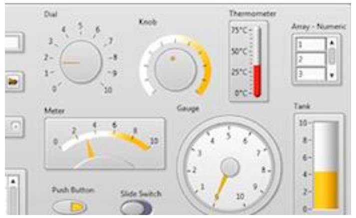


Data from measurements or simulations are read into **`sidpy.Dataset`** (d) objects directly by **`SciFiReaders`** (SF). Data are processed using multiple science packages in the Pycroscopy ecosystem that interoperate via **`Dataset`** objects. **`Dataset`** objects are written to HDF5 files via **`pyUSID`** (US) or **`pyNSID`** (NS).

Slide courtesy of R. Vasudevan

Solutions: Integrated Ecosystems

Instrument Tier



Automated, standardized, modularized data acquisition

Instrument-agnostic, self-describing, model in an open friendly file format

Centralized repository for data processing, analysis

Interactive visualization + analysis + storage on the cloud

Slide courtesy of R. Vasudevan