Machine Learning for Science

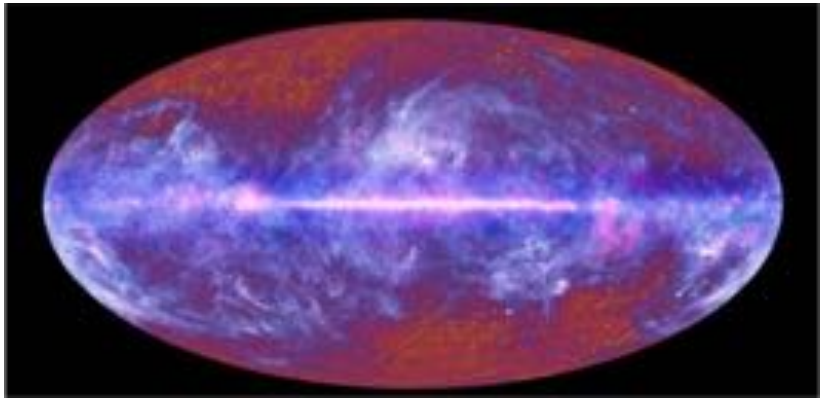
Kathy Yelick
Associate Laboratory Director for Computing Sciences
Lawrence Berkeley National Laboratory

Professor of Electrical Engineering and Computer Sciences
UC Berkeley

Three ingredients for Machine Learning

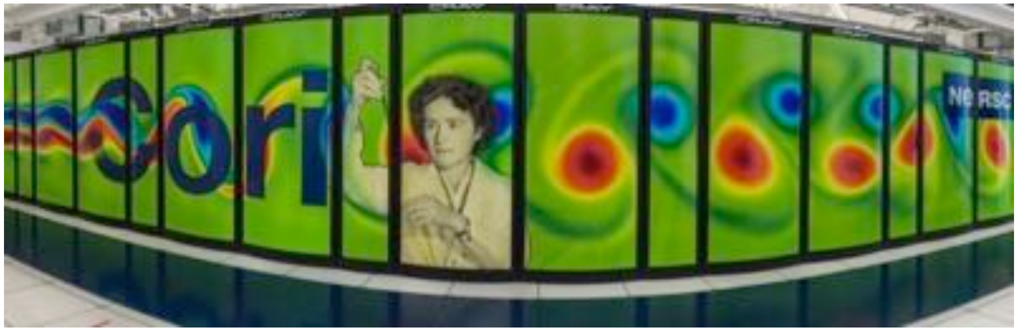
Data

Complexity



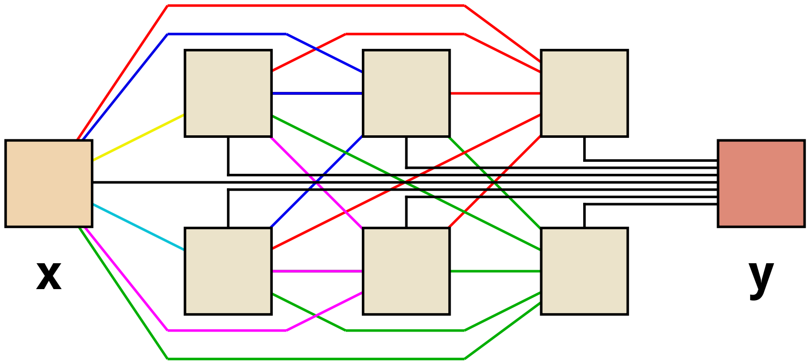
Machines

Scale



Algorithms

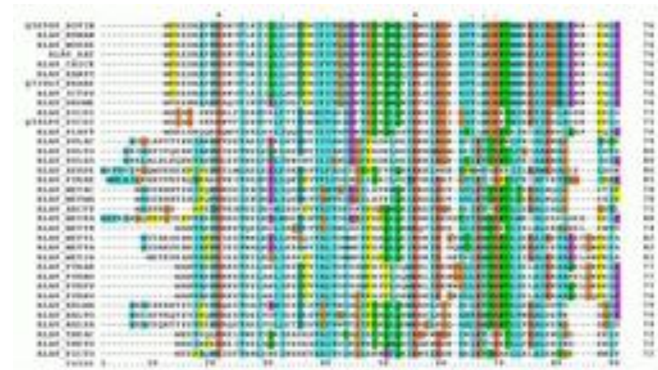
Interpretability



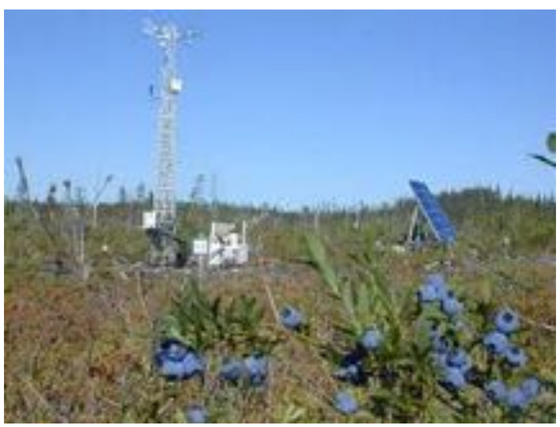
Experimental, Observational, and Simulation Data in Science

Image / Video Processing

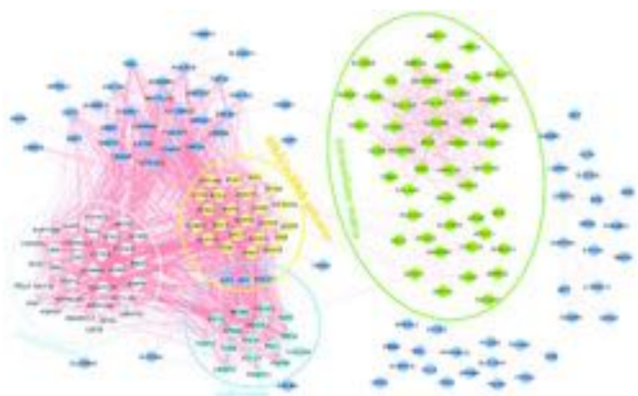
Text



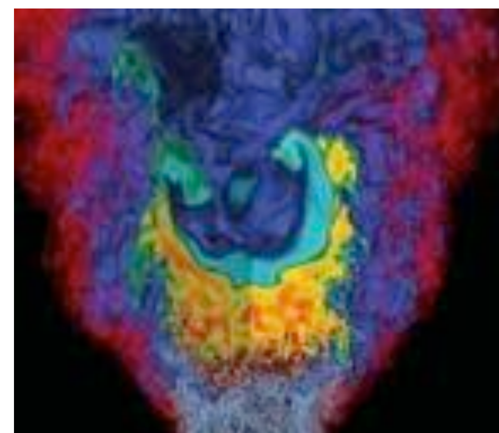
Genomics



Signal Processing

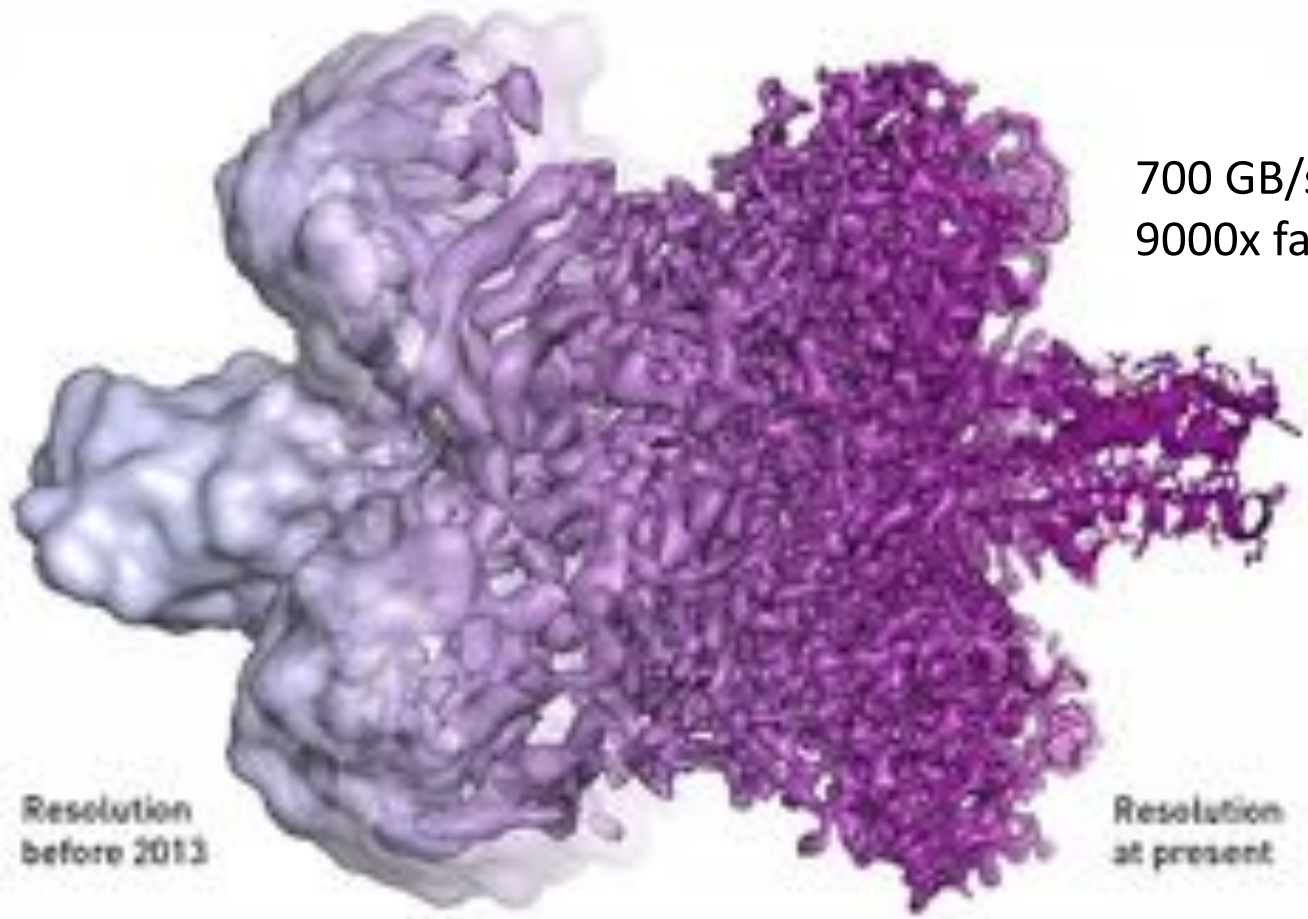


Graphs (Relationships)



Simulation Analytics

Superhuman "sensors" for science



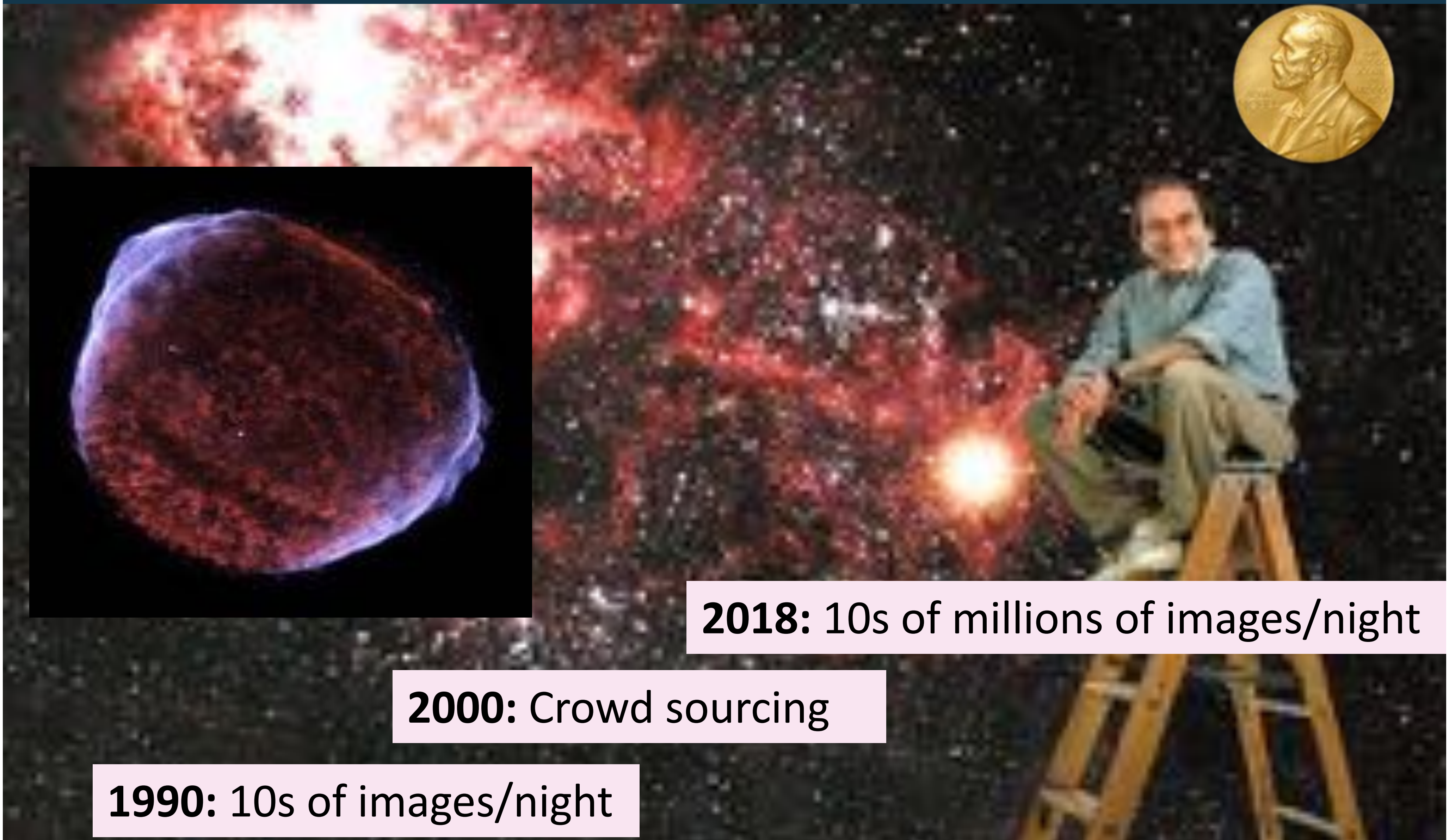
700 GB/sec in 2018
9000x faster than 2008

Berkeley Lab advances detector technology for many fields of science, including (above CryoEM) biology, cosmology, material science, physics, and more.

Machine Learning in Science

Cosmology, Climate, Cats, Catalysts and Carrots

Cosmology: Finding Features in Images



2018: 10s of millions of images/night

2000: Crowd sourcing

1990: 10s of images/night

Understanding from Observation + Simulation

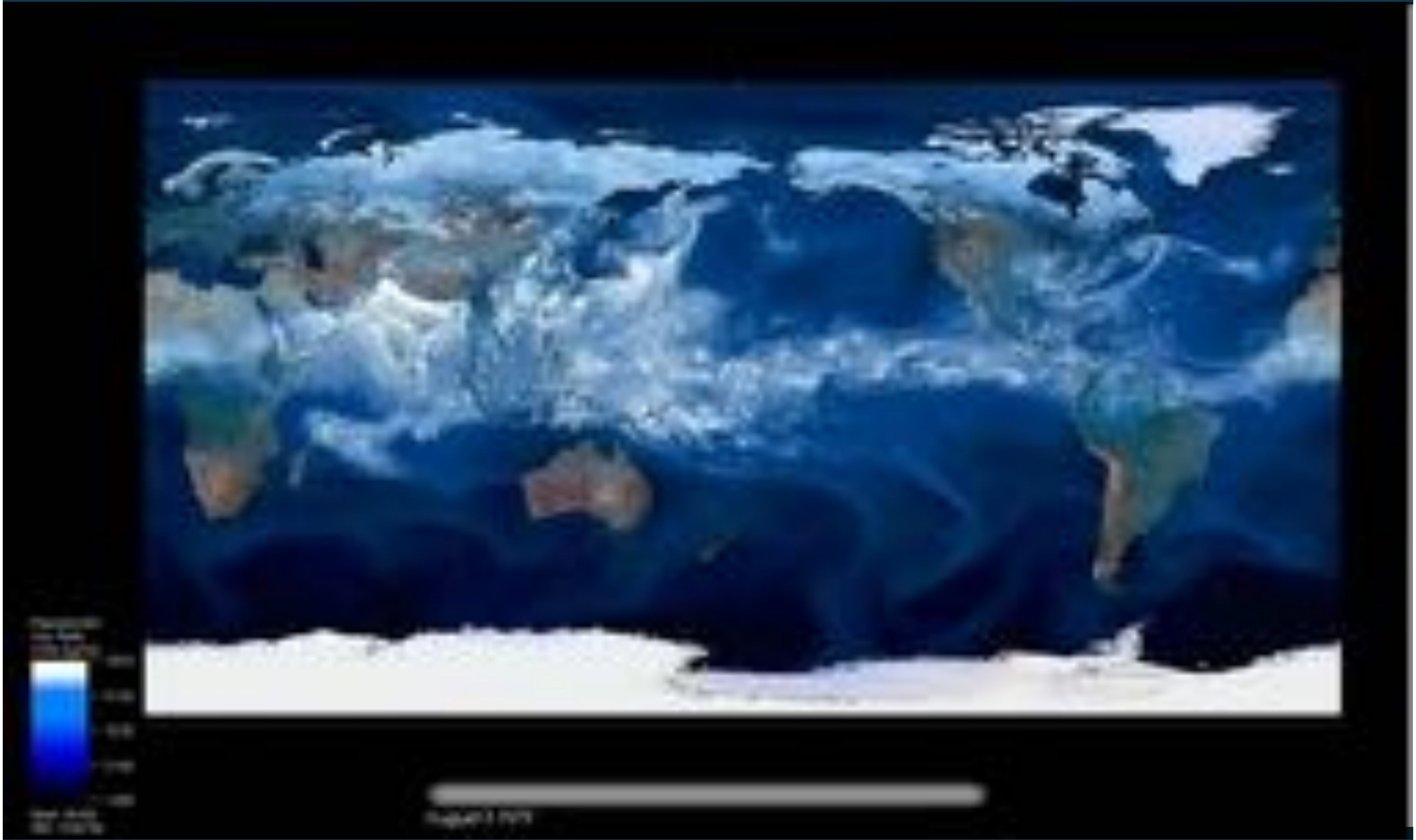
Science is about understanding

- Use simulations to interpret observations
- ML (reduced order models) to accelerate simulation “campaign”
- Using DL to improve cosmological constants from simulations

CosmoFlow on TensorFlow: Trained on 8K nodes, 10 min

Shirley Ho (Physics), Debbie Bard (NERSC)

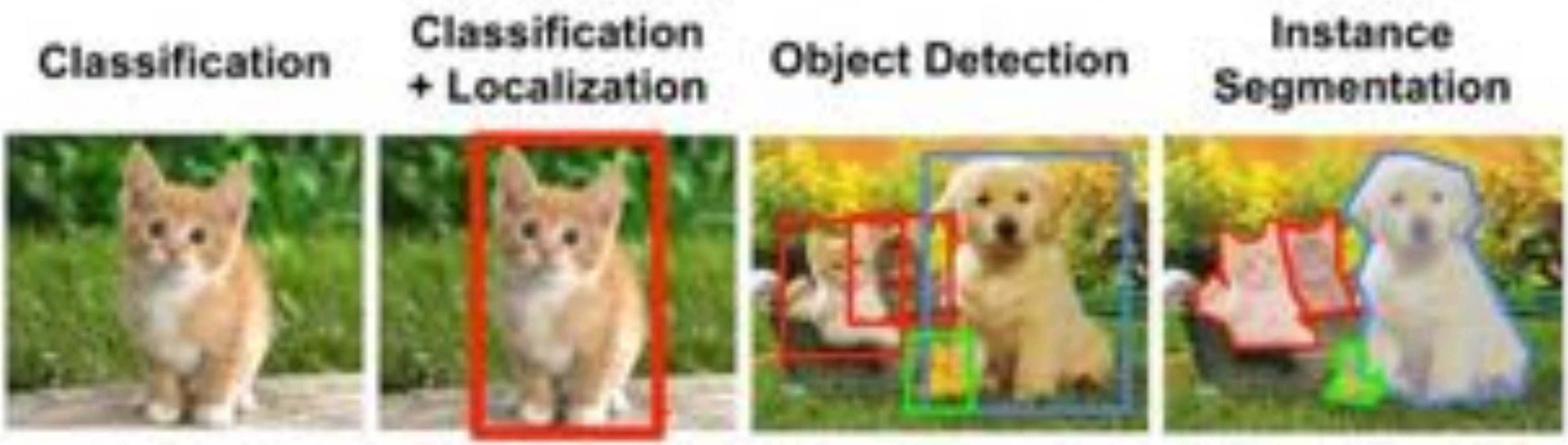
Features in Simulation: 3D, 4D, Adaptive, Unstructured

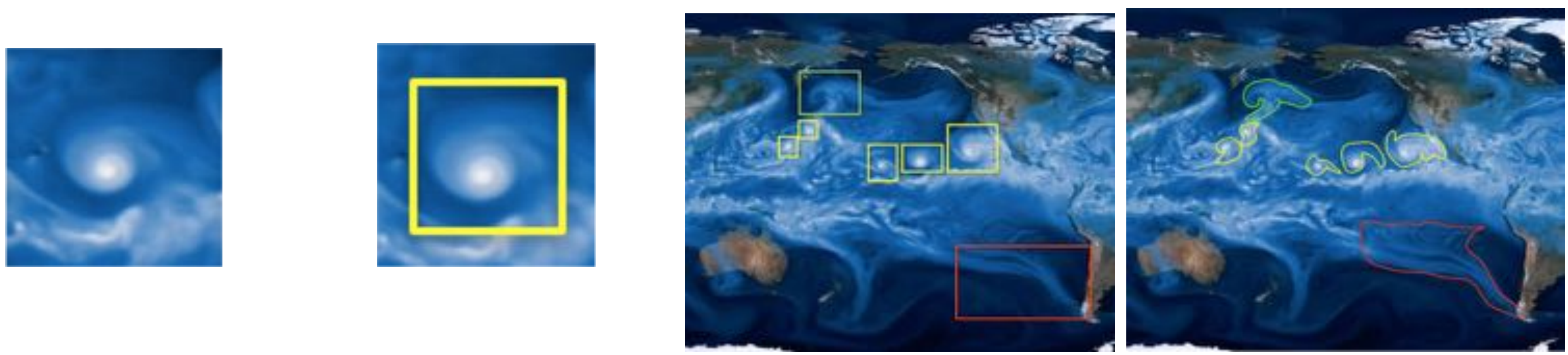


BERKELEY LAB U.S. DEPARTMENT OF ENERGY Office of Science

Prabhat, Michael Wehner

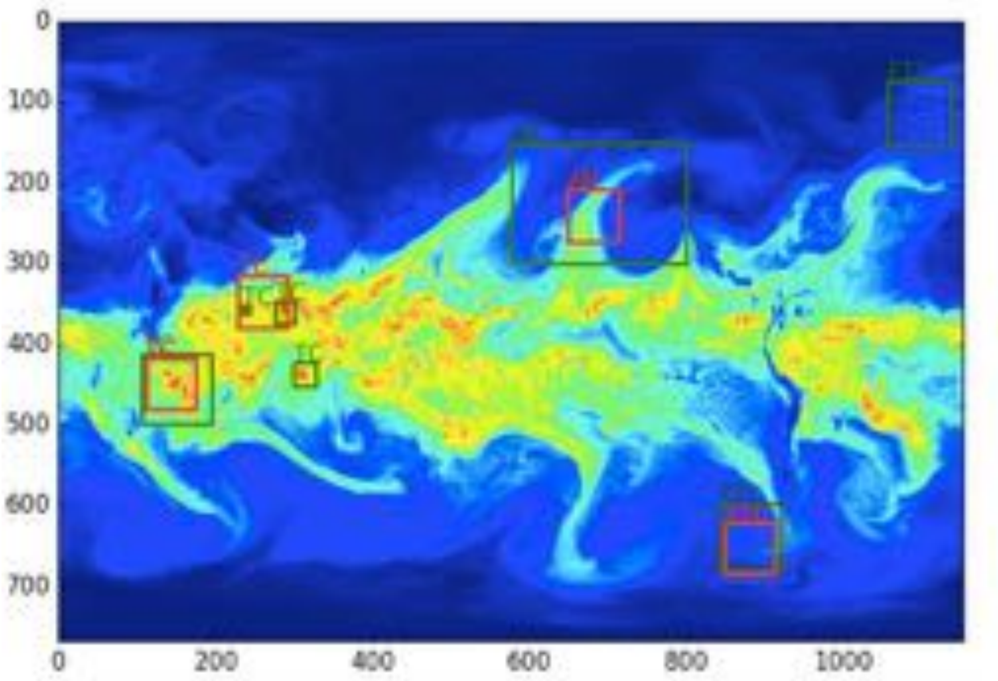
Machine Learning in Climate Data



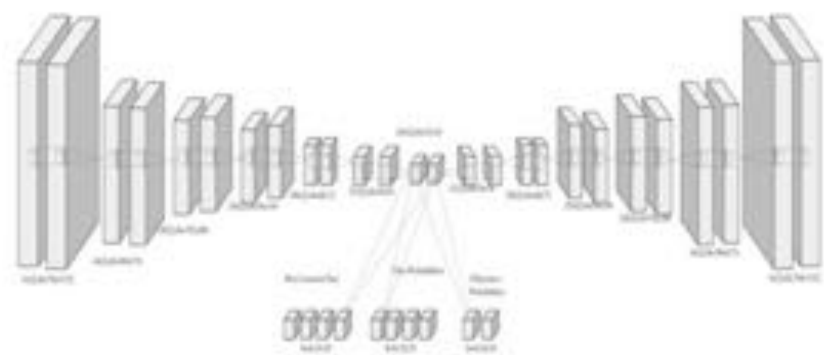


Contributors: Prabhat, Thorsten Kurth, Jian Yang, Ioannis Mitliagkas, Chris Pal, Nadathur Satish, Narayanan Sundaram, Amir Khosrowshahi, Michael Wehner, Bill Collins.

Deep Learning at 250 PF for Extreme Weather Events



Ground Truth vs Prediction

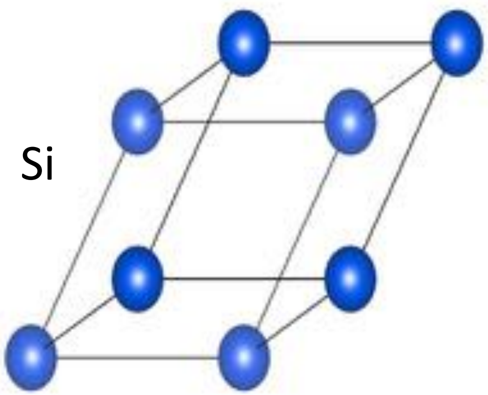


Use of deep learning (CNNs)

- Supervised and semi-supervised learning on CAM5 data
- 85-99% accuracy at identifying extreme climate events
- Scaled to 250PF on Summit at ORNL; trained in 100 minutes

Material design with computation

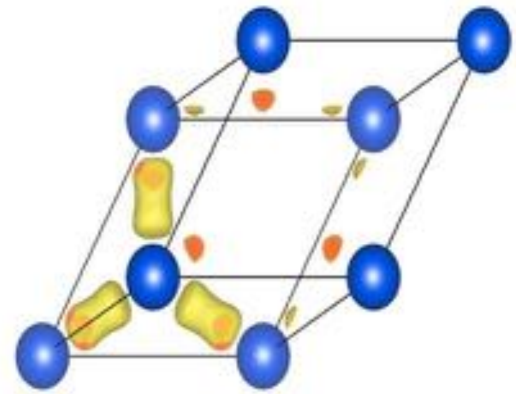
Given an atomic structure,



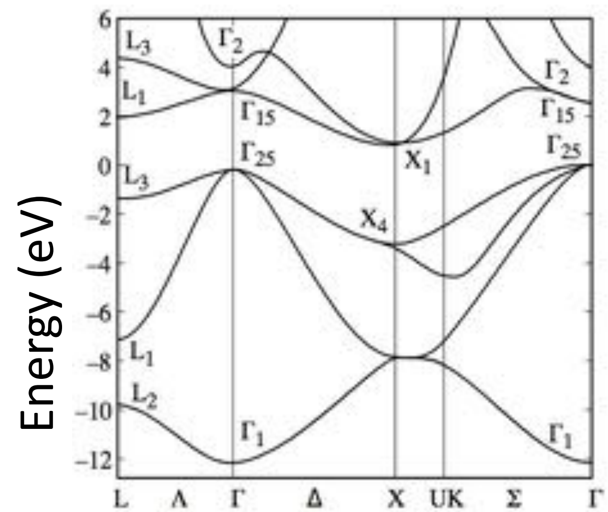
...use quantum theory and supercomputers to determine...

$$\hat{H}\left|\psi\right\rangle = E\left|\psi\right\rangle$$

...where the electrons are...



...and what the electrons are doing.

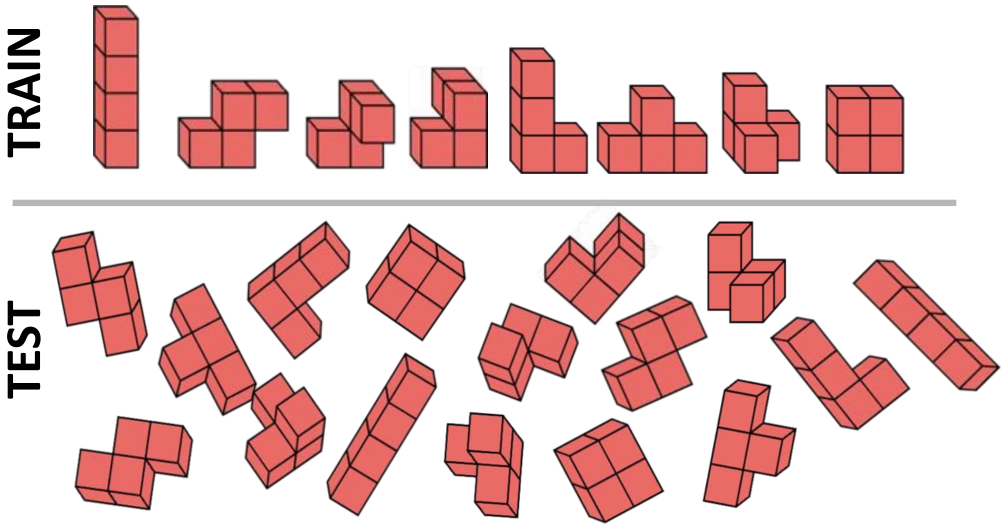


Reduce, reuse and recycle data: Materials Project has >40,000 users

Recognizing Motifs in 3D Materials Structures

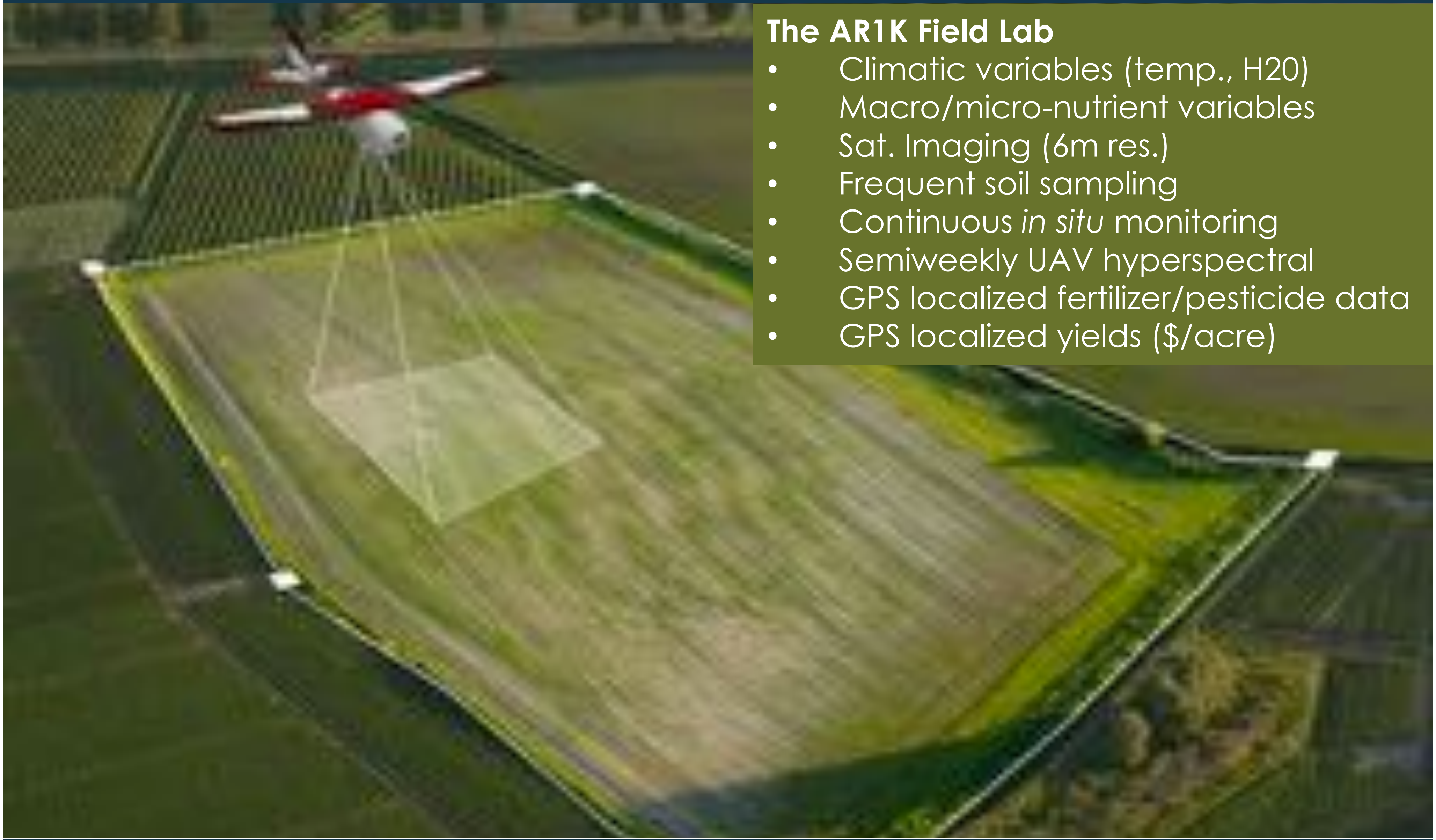
Rotated image

CNN filter output



A network with 3D translation- and 3D rotation-equivariance

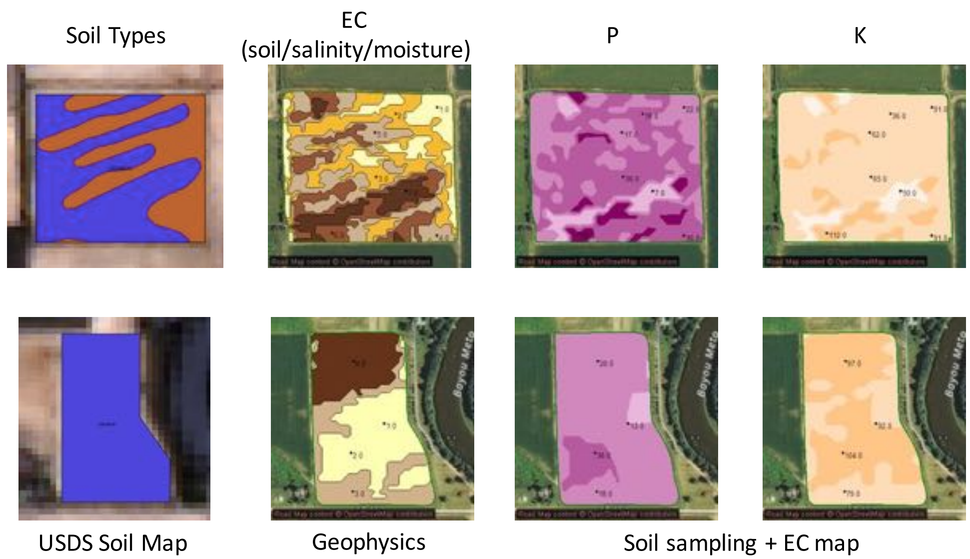
Multimodal data in agriculture



The AR1K Field Lab

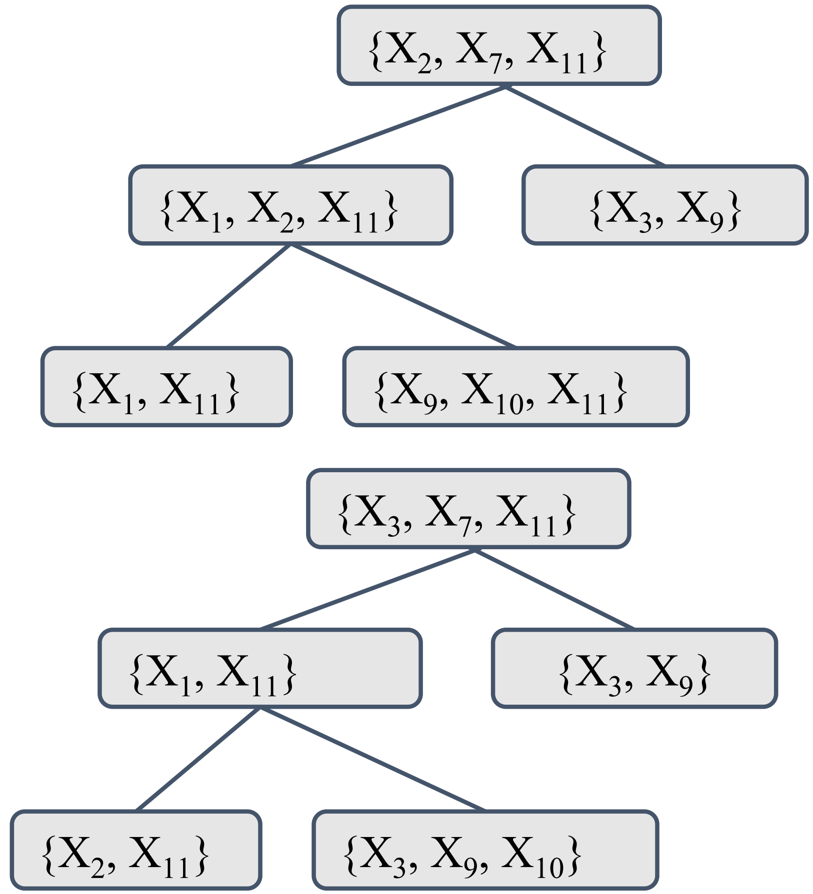
- Climatic variables (temp., H20)
- Macro/micro-nutrient variables
- Sat. Imaging (6m res.)
- Frequent soil sampling
- Continuous *in situ* monitoring
- Semiweekly UAV hyperspectral
- GPS localized fertilizer/pesticide data
- GPS localized yields ($/acre)

Learning Mechanistic Models



Iterative Random Forest

Breaking dimensionality curse



- **Construct a 4D Virtual Farmland**
- **Feature selection**
  - Hyperspectral phenotypes
  - Microbes/metabolites impacts
- **Design microbial amendments**

Basu *et al.* 2018. PNAS.

Large-scale microbiome genomic analysis

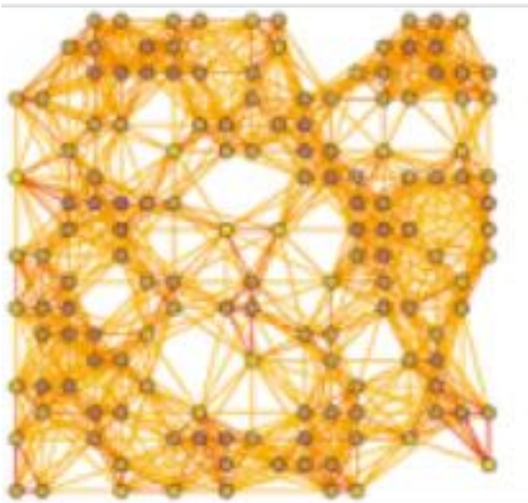
Metagenome Assembly

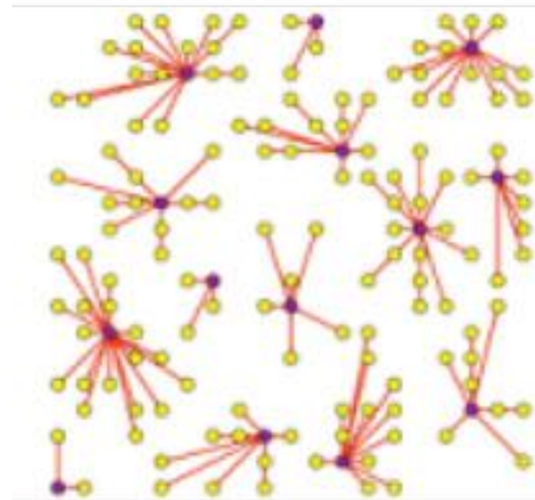
- Thousands of species mixed, with errors
- No reference
- HPC MetaHipMer assembly puts the pieces together
- 2.8 TB Twitchell Wetlands -- largest of its kind?

Cluster gene/protein families at scale

Input: pairwise similarities between proteins (sparse)

Output: clusters of similar proteins





- **Desired scale: 10s of billions of genes/proteins, trillions of nonzero pairwise similarities ("all metagenomes")**
- **Today: 282M genes in 3 hours on 2K nodes**

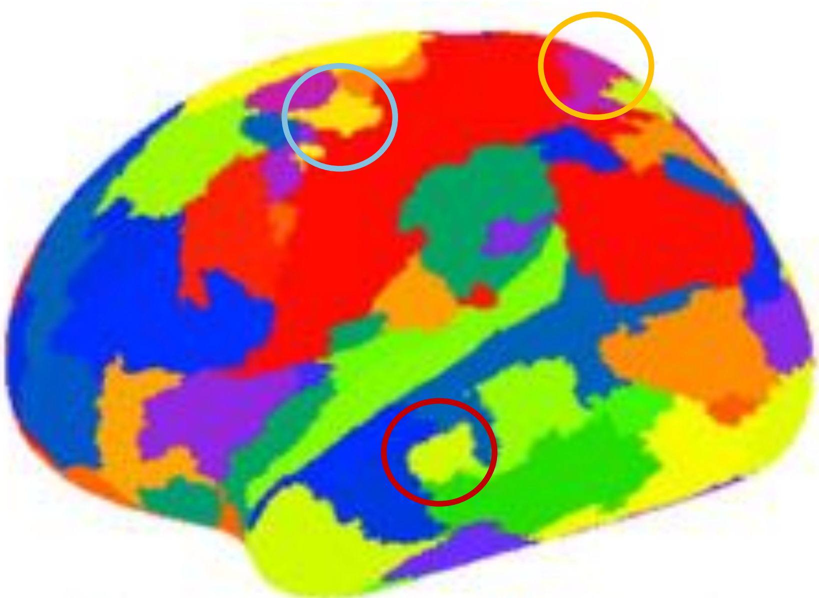
HipMCL work by Aydın Buluç (ECRP) and Ariful Azad

Learn the relationship between features with Graphical Model Estimator

Source: https://media4.s-nbcnews.com/i/newscms/2017_25/958456/150401-dna-strand-mn-1645_9d74198e59853eb79be3124a876ad4fd.jpg

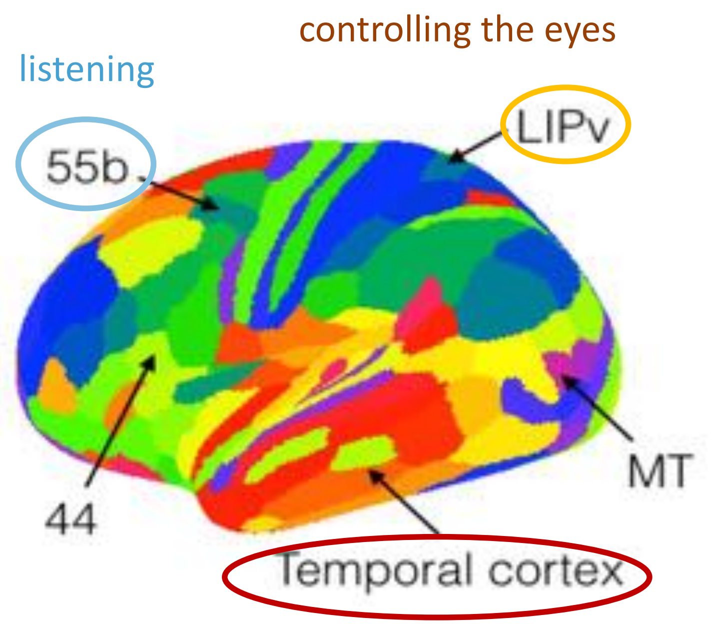
HPC Graphical Model Estimator Discovers Regions and Co-regions

Automatic parcellation from fMRI data alone



$\lambda_1 = 0.48, \ \lambda_2 = 0.39, \ \epsilon = 3,$
$\%$ of best score $= 100$

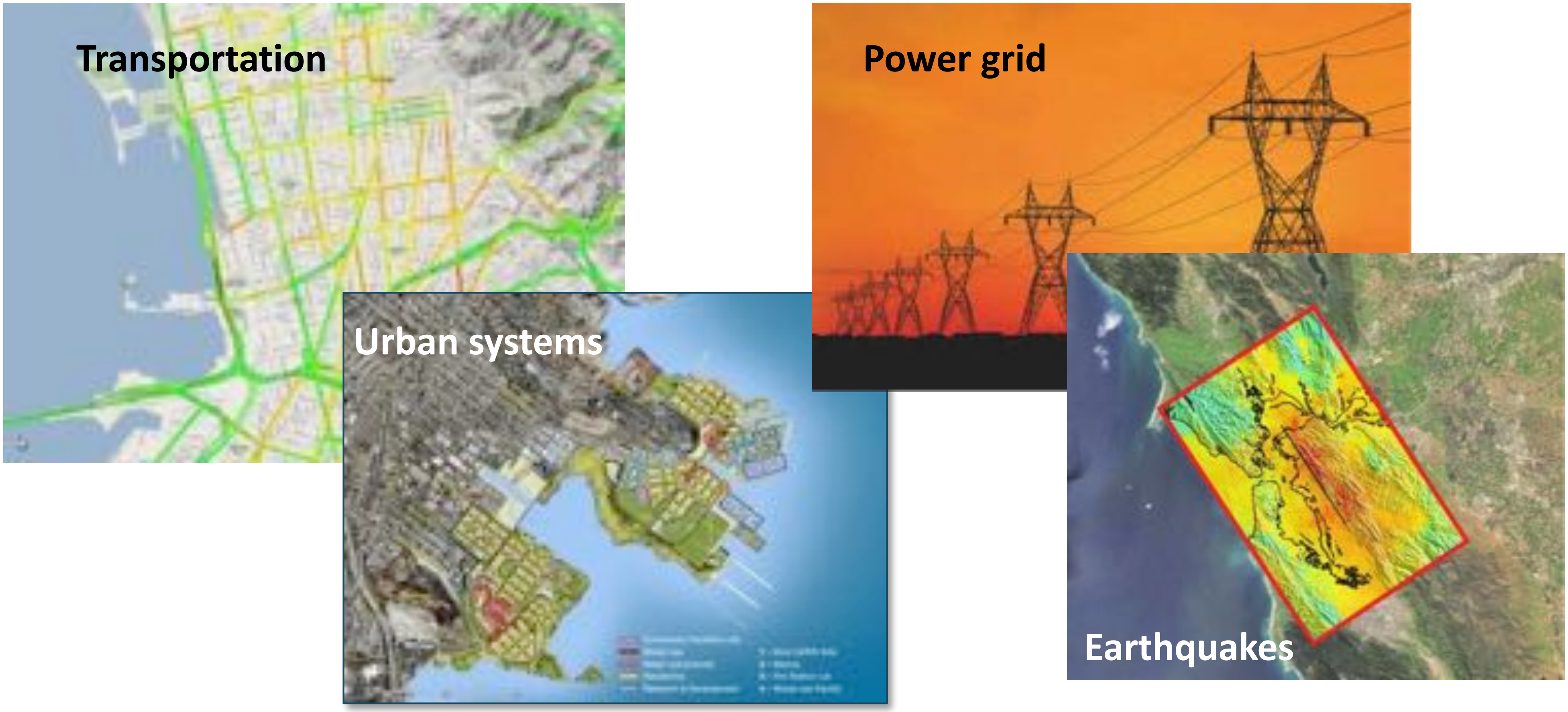
Baseline parcellation from Glasser
[Glasser et al. 2016]



First of kind analysis at this scale using new algorithm and high performance computing at LBNL

Koanantakool, Oh, Buluc, Morozov, Oliker, Yelick, AISTAT 2018.

Energy science from embedded sensors



Use physics-based simulations, augmented with precise, localized data-driven models

Tempered Enthusiasm for Machine Learning (Especially Deep Learning) in Science

ML Explainability is not the same as Performance

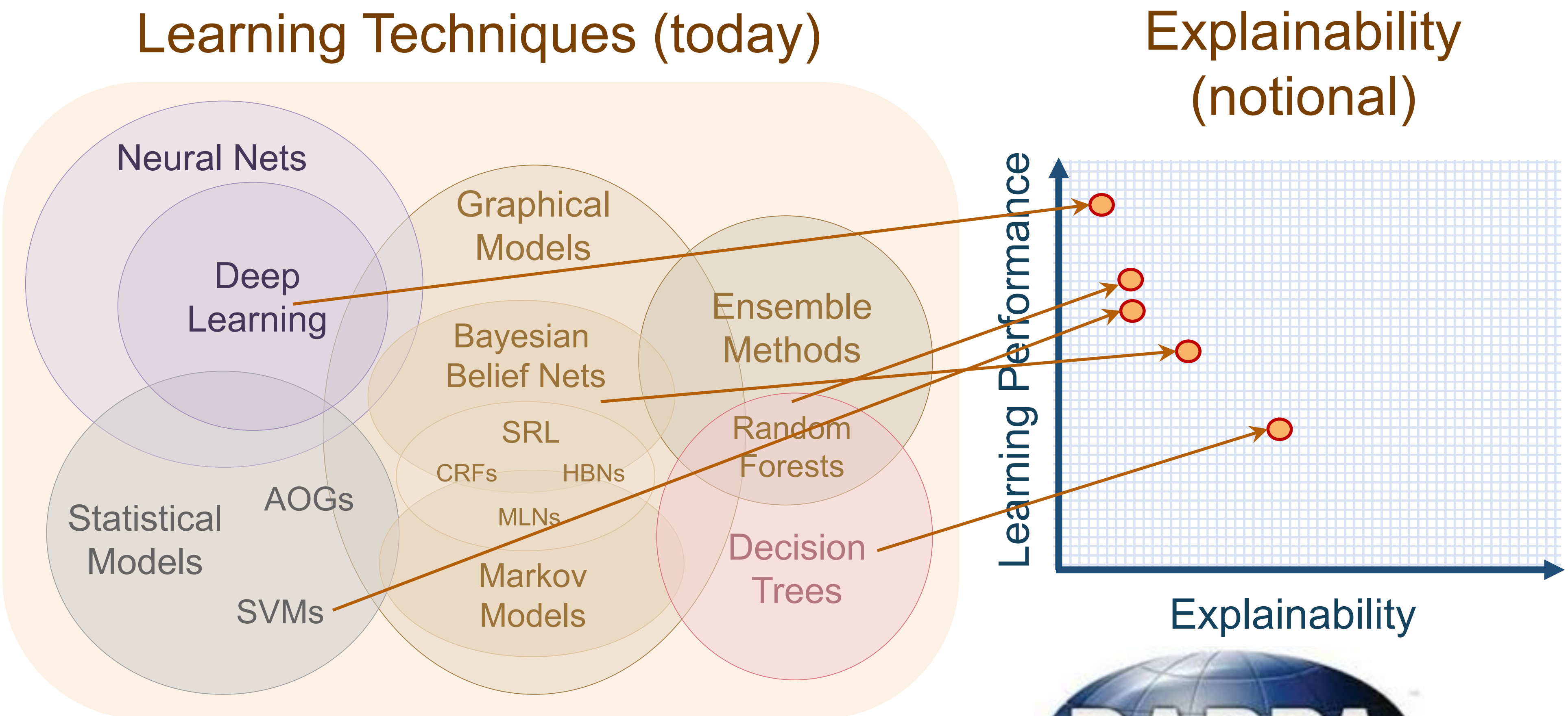
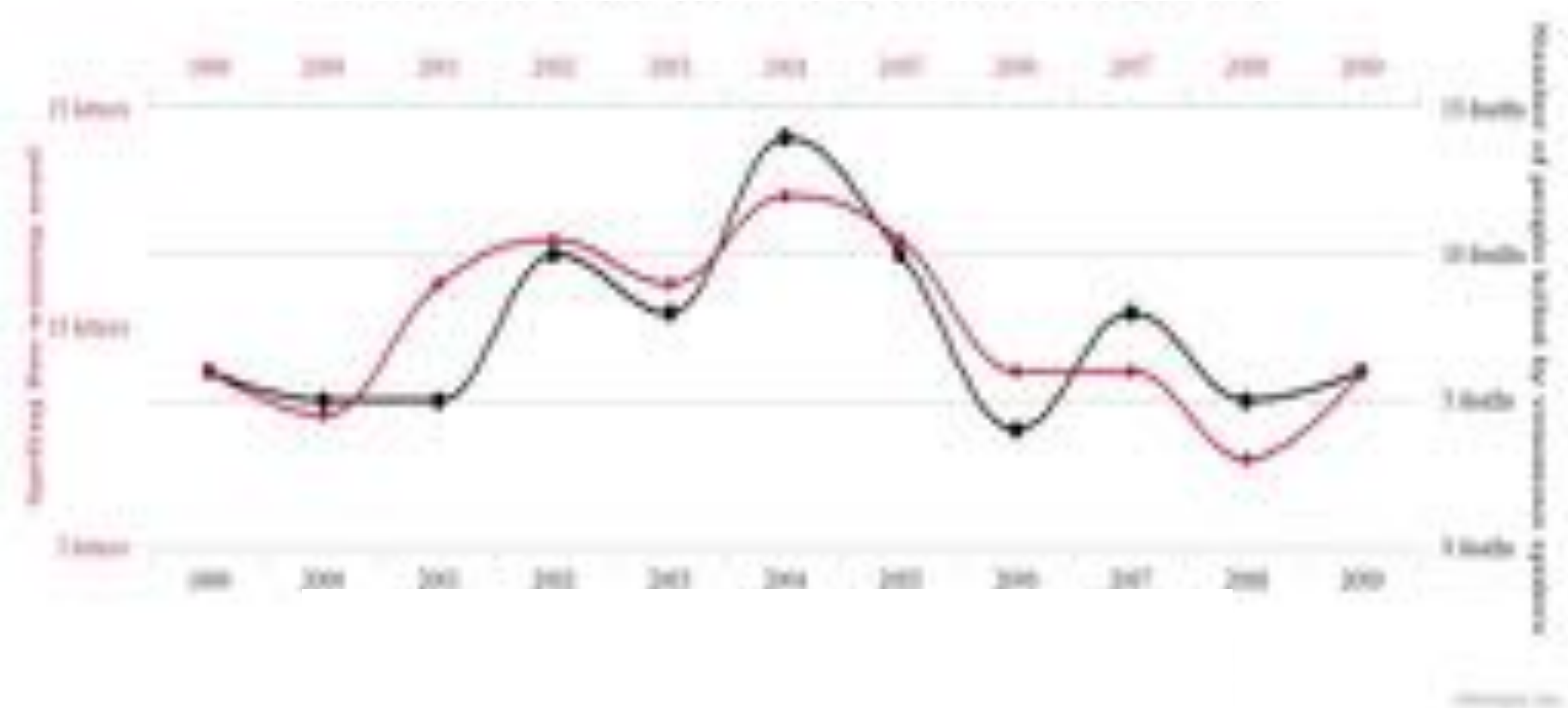


Image from DARPA's XAI Program, David Gunning

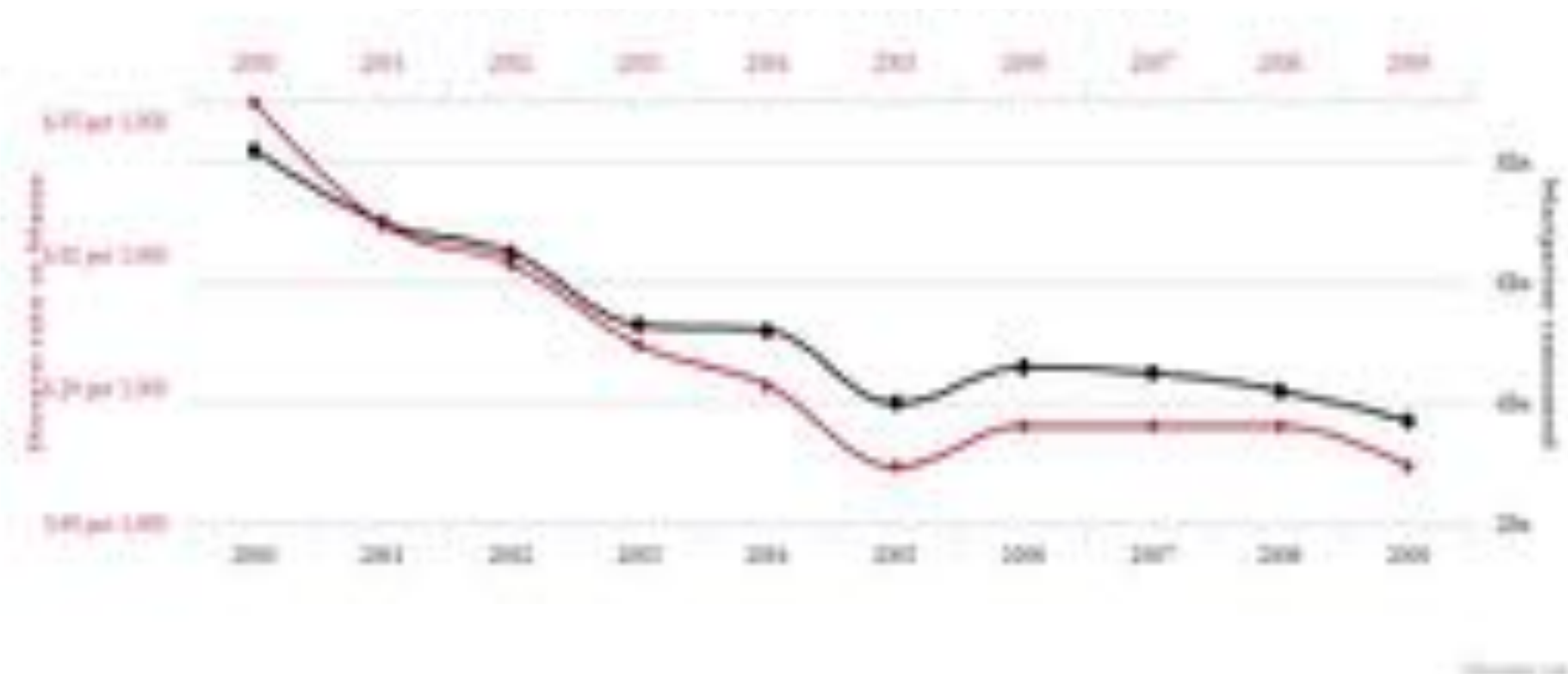
Non-actionable correlation

***r* ~ 0.81**



Correlation is not Causation

***r* ~ 0.99**



Filtering, De-Noise and Curating Data

AmeriFlux & FLUXNET: 750 users access carbon sensor data from 960 carbon flux data years

Arno Penzias and Robert Wilson discover Cosmic Microwave Background in 1965

Machine Learning in Science

Excitement over many uses of ML for:

- Feature extractions from observations, experiments, and simulations
- Clustering and regression
- Dimensionality reduction for complex data
- Surrogate models to approximate expensive simulations or experiments
- Designing and controlling experiments
- Filling in missing models in simulations

A robust peer review process in science domains and great training opportunities on open science data