Argonne
NATIONAL LABORATORY

Overview of Machine Learning for Science

Bethany Lusch
Computer Scientist
Argonne Leadership Computing Facility
Argonne National Laboratory
blusch@anl.gov

ATPESC – Machine Learning track
August 11, 2023

Overview of Today

- Brief intro to machine learning
- Aspects more specific to applying to science
- Examples from Argonne

8:30 AM Intro: Overview of machine learning for science (talk)

9:00 AM transition time: splitting into groups (people new to deep learning vs. more experienced)

9:10 AM: Parallel Session, Part 1 (talk/hands on)

- Main room: Introduction to deep learning
- Breakout room: Building data pipelines for deep learning

10:15 AM Break

10:45 AM Parallel Session, Part 2 (talk/hands on)

- Main room: Introduction to convolutional neural networks
- Breakout room: Profiling deep learning

11:50 transition time: back to main room

Overview of Today

12 PM Research talk: AI Ethics and Responsible Data Science for Scientists

12:30 PM Lunch

1:30 PM Research talk: Transfer and Multi-Task Learning in Physics-Based Applications with Deep Neural Operators

2:00 PM Distributed Deep Learning (talk/hands on)

3:00 PM Break

3:30 PM AI Testbed (talk/hands on)

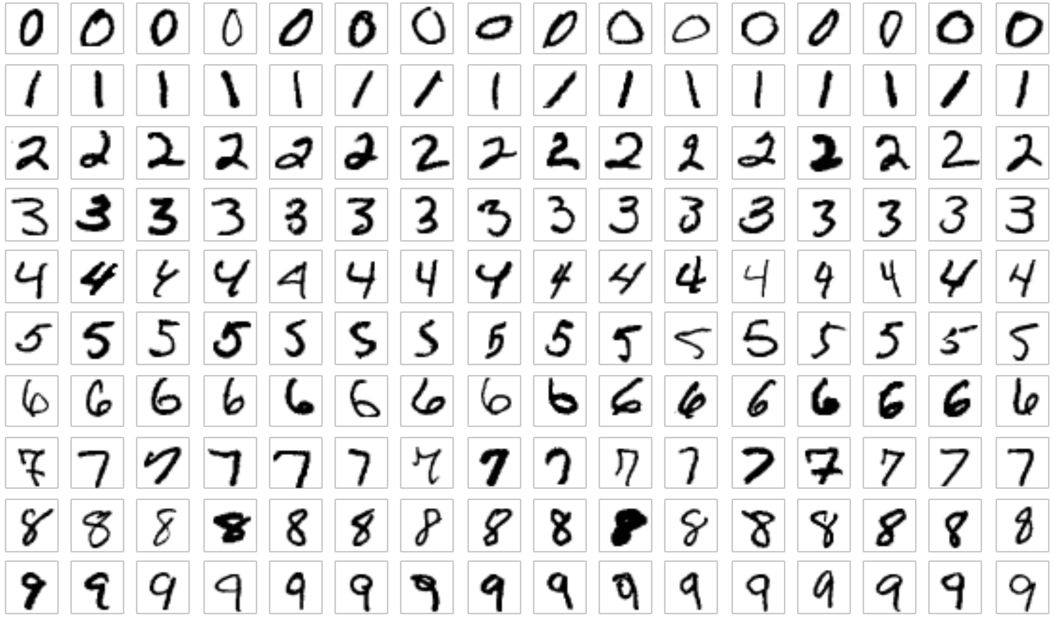
5 PM Close out/Exam

Argonne NATIONAL LABORATORY

Types of Machine Learning

What is machine learning?

field of study that gives computers the ability to learn without being explicitly programmed



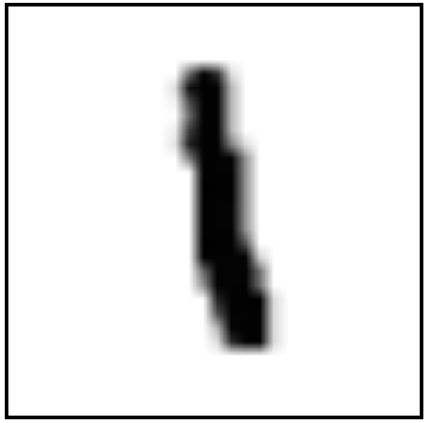
Example: post office wants machine to sort mail by zip code

Want to label each image as a digit 0…9

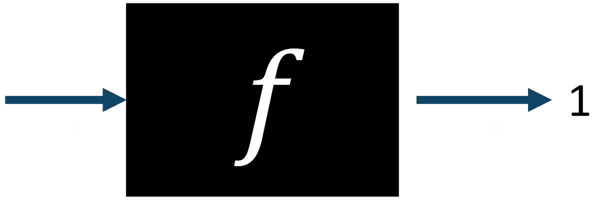
Explicit programming: IF 80% of black pixels are in middle 30% of image, THEN label as 1.

Reading Zip Codes

Machine learning: field of study that gives computers the ability to learn without being explicitly programmed



$$\simeq \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .6 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .7 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .5 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & .7 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .9 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & 1 & .1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$



fit function
f(image) = label

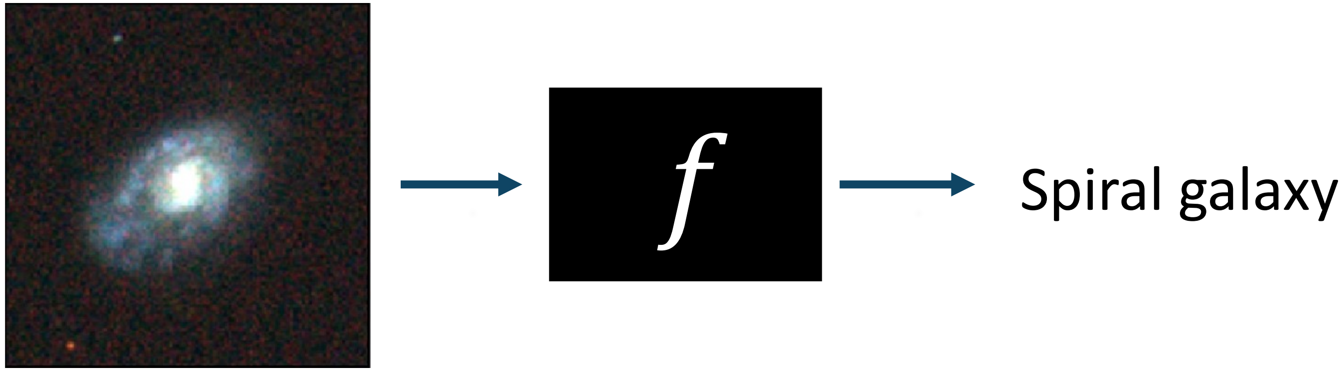
by considering many image & label pairs
“learns” as sees more examples

Classification

Have a category label for each data point, learn to categorize

Example: Label galaxies from sky surveys

- Why? Citizen science campaigns have classified thousands, but large surveys observe millions of galaxies



ALCF project! You can ask Huihuo about this.

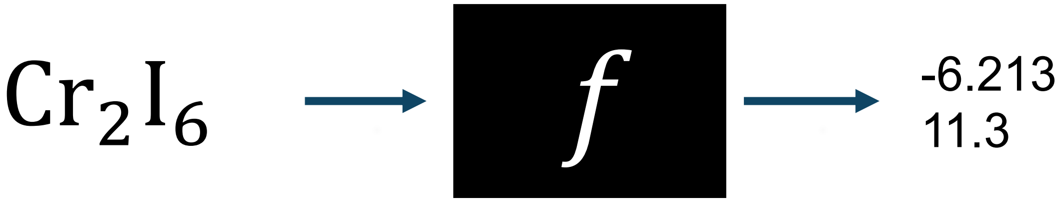
Khan, et al. “Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey.” *Physics Letters B* (2019)

Regression

Have a numeric label for each data point, learn to predict number

Learn how to predict the formation energy and magnetic moment of a material (after seeing a database of DFT calculations)

- Why? Trying to discover new chemically stable two-dimensional magnetic materials
- Potential applications in data storage and spintronics



$\mathrm{Cr_2I_6}$

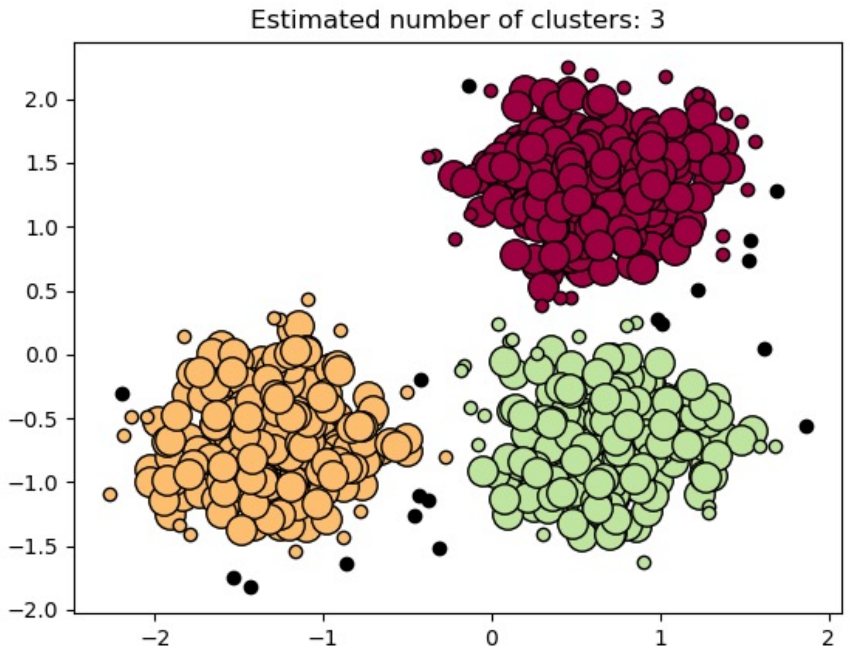
ALCF project! You can ask me about this.

Rhone, et al. "Artificial Intelligence Guided Studies of van der Waals Magnets." *Advanced Theory and Simulations* (2023)

Clustering

Have an unlabeled dataset, find groups of similar data points

- Find subtypes of breast cancer (after seeing data from a bunch of patients)
- Find communities in a social network (after seeing Twitter data)
- Cluster pixels of a cosmology simulation to find halos
- Cluster snapshots of a fluids simulation to find regimes in time*



* Barwey, et al., “Data-Driven Reduction and Decomposition via Time-Axis Clustering” AIAA SciTech Forum (2020)

Reinforcement Learning

An agent explores an environment and learns how to get rewarded

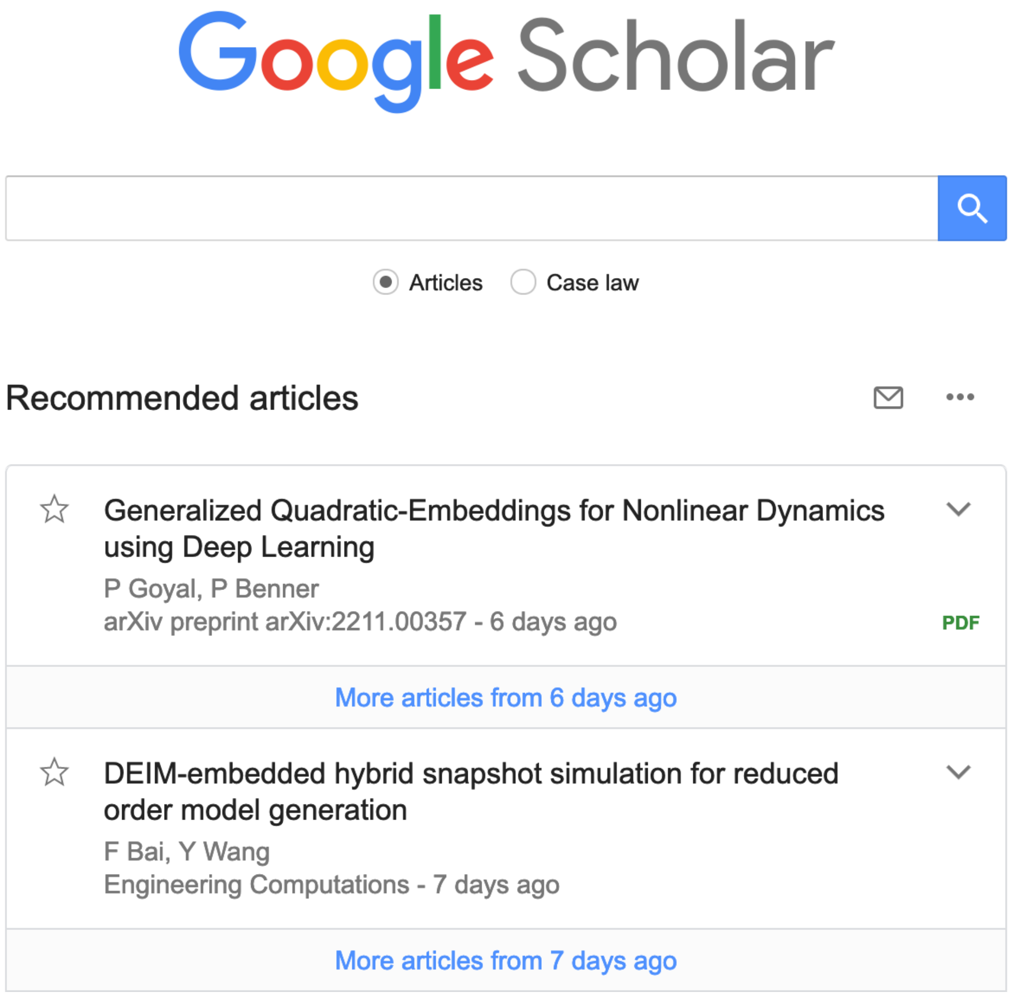
- Learn to play Frogger by playing the game and receiving a reward (score)
- Learn to suggest useful chemical reactions
  - Reward: highest yield (from simulations and experiments)*
- Learn to optimize the shape of an airfoil
  - Reward: negative drag coefficient from numerical simulation**

* Zhou, et al. "Optimizing Chemical Reactions with Deep Reinforcement Learning" *ACS Cent. Sci.* (2017)
** Bhola, et al. "Multi-fidelity reinforcement learning framework for shape optimization" *Journal of Computational Physics* (2023)

Recommendation Systems

Generate personalized suggestions

- Recommend a movie to watch on Netflix
- Recommend a scientific paper to read



Google Scholar

● Articles ○ Case law

Recommended articles

☆ Generalized Quadratic-Embeddings for Nonlinear Dynamics using Deep Learning
P Goyal, P Benner
arXiv preprint arXiv:2211.00357 - 6 days ago
PDF

More articles from 6 days ago

☆ DEIM-embedded hybrid snapshot simulation for reduced order model generation
F Bai, Y Wang
Engineering Computations - 7 days ago

More articles from 7 days ago

Generative Modeling

Create new examples from a probability distribution

- Generate an image based on a text description
- Generate an essay based on a prompt
- Super-resolution for physics simulation
- Learn an easier way to sample a difficult probability distribution
    - Particle physics: LatticeQCD, need to generate samples
    - Problem: sampling gets stuck in small region of space
    - Model can learn mapping between simple distribution and complicated one → better sampling, guarantees*

"a painting of a fox sitting in a field at sunrise in the style of Claude Monet"
DALL-E 2 (openai.com)

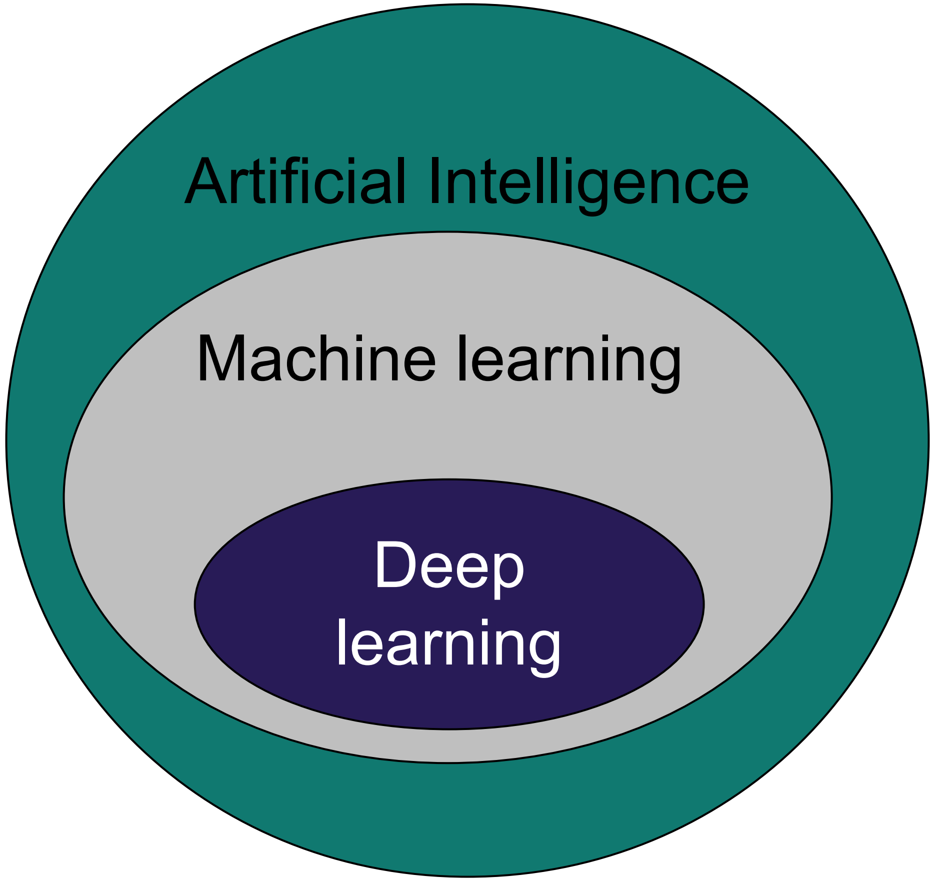
* Abbott, et al. "Gauge-equivariant flow models for sampling in lattice field theories with pseudofermions" *Physical Review D* (2022)

Early Science Project for Aurora!

Supervision

- Supervised learning: have labeled data (input and output pairs) Examples: classification & regression
- Unsupervised learning: have unlabeled data and want to find patterns, structure, etc. Example: clustering
- Semi-supervised learning: mix of labeled & unlabeled data

Focus of today: Deep learning



Artificial Intelligence

Machine learning

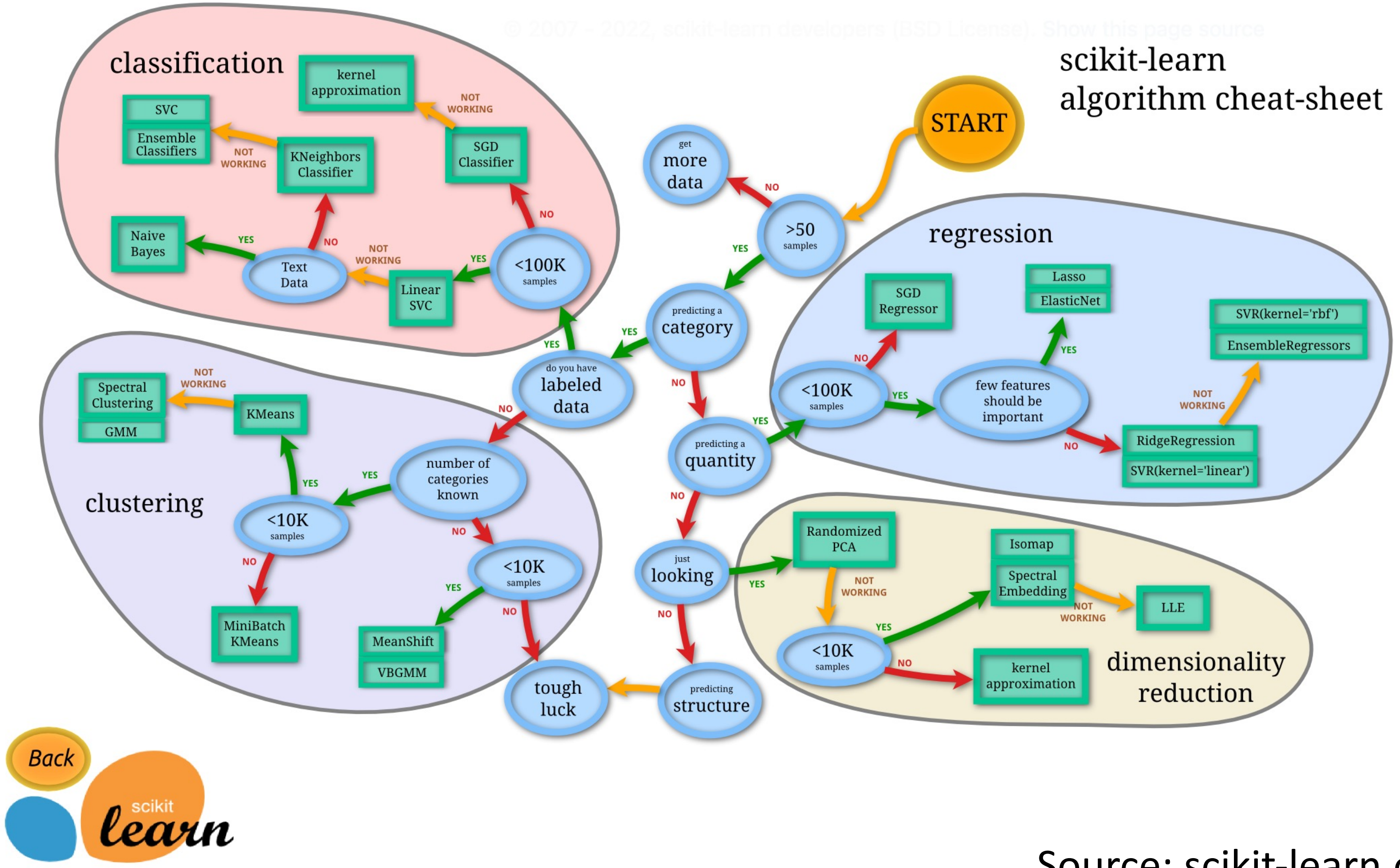
Deep learning

Examples of Other AI Goals

- Intelligently **searching** through many solutions, such as planning how to a robot should complete a task
- Automated **reasoning**, such as theorem proving
- Natural **language** processing, such as writing an essay based on a prompt
- **Perception**, such as speech recognition and computer vision
- **Dimensionality reduction,** such as finding the primary modes within a fluids simulation
- **Anomaly detection**, such as flagging suspicious behavior in computer networks

“Classical” Machine Learning

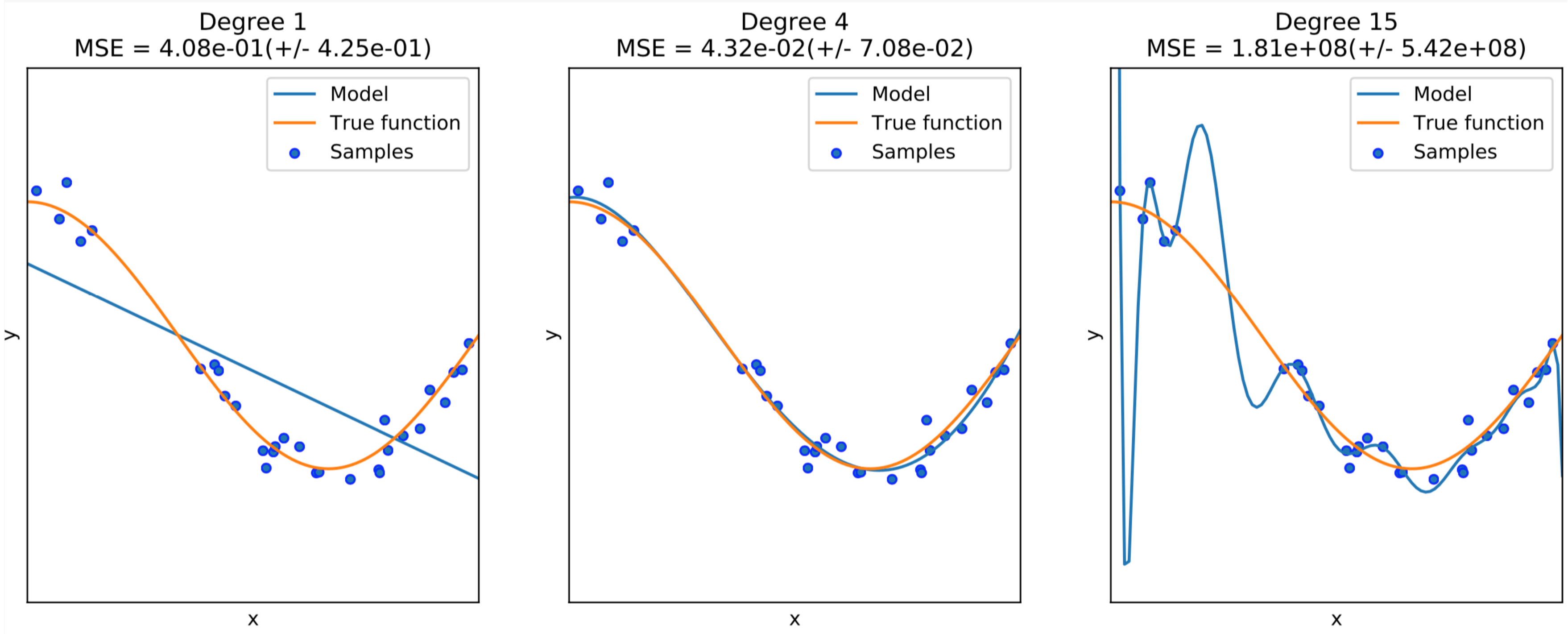
- Non-deep machine learning
- Many methods are in the Python package scikit-learn



Source: scikit-learn.org

Comparing ML model options: Bias vs. Variance

underfitting balance overfitting



High bias Low bias, low variance High variance

A major theme of machine learning!

Pictures from Kyle Felker, produced from code in scikit-learn documentation

To Check for Overfitting vs. Underfitting

“test” data

MUST hold out some data and check error *at very end*

Common:

- Randomly split data 70% training, 20% validation, 10% test
- Use training data to fit the model
- Use validation data to compare options (different algorithms or different hyperparameters)
- Report test error at **end of project**

If you peek, not really reporting generalization error!

To Check for Overfitting vs. Underfitting

Monitor training and validation error…

If training error too high → underfitting

If training error << validation error → overfitting

Augmenting Techniques

Automated Machine Learning

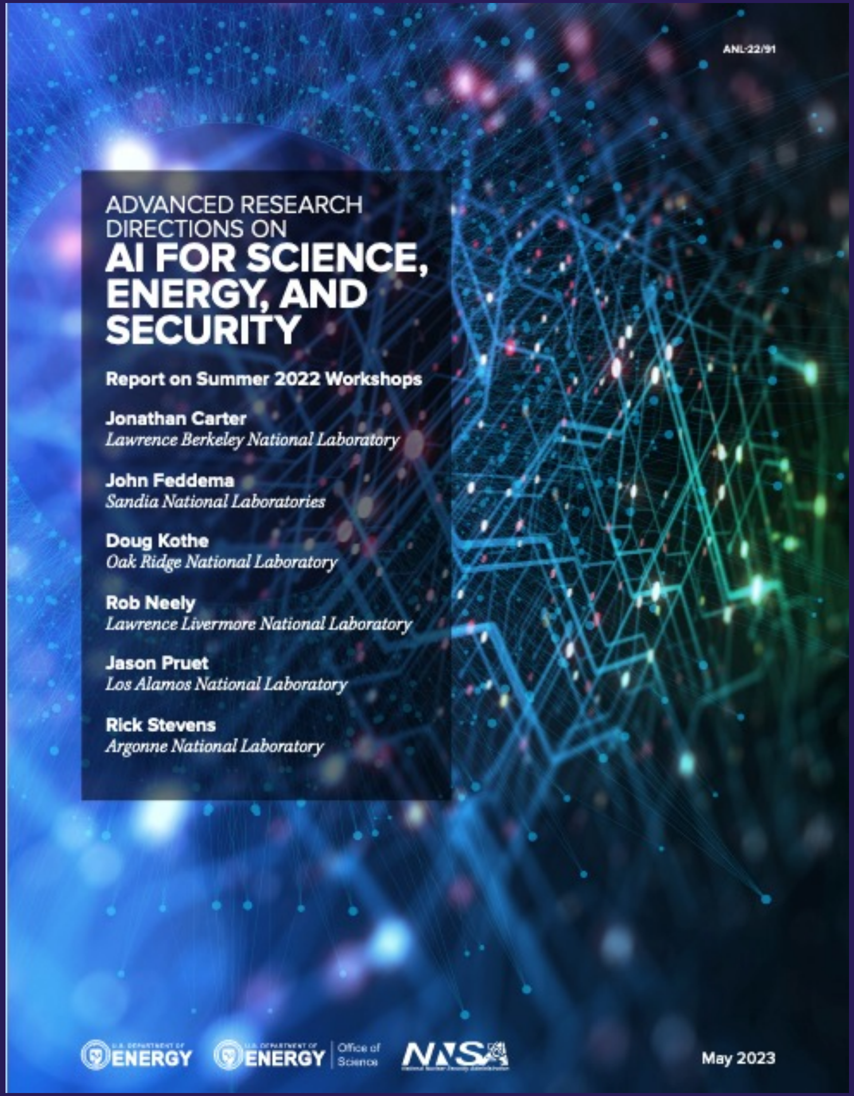
- Hyperparameter optimization: search for good hyperparameters of algorithms
- Neural Architecture Search: search for a good neural network architecture, even a very unusual one
- Even data cleaning, choose a good ML method, etc.

Improving Datasets

- Active learning: choose new training data to label that will improve the ML model
- Data augmentation: make dataset larger by, for example, including all rotations of existing images
- Subset selection: find a subset of your data examples that is sufficient for good accuracy
- Feature selection: reduce the dimensionality of your examples by removing useless features

Argonne
NATIONAL LABORATORY

Open Problems: AI + Science



ANL-22/91

ADVANCED RESEARCH DIRECTIONS ON
AI FOR SCIENCE, ENERGY, AND SECURITY

Report on Summer 2022 Workshops

Jonathan Carter
Lawrence Berkeley National Laboratory

John Feddema
Sandia National Laboratories

Doug Kothe
Oak Ridge National Laboratory

Rob Neely
Lawrence Livermore National Laboratory

Jason Pruet
Los Alamos National Laboratory

Rick Stevens
Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY   U.S. DEPARTMENT OF ENERGY | Office of Science   NNSA

May 2023

Sampling of AI Approaches and Open Challenges

- Learning surrogate models for scientific computing
  - Hard to generalize
  - Data-hungry
  - Infrastructure/workflows for coupling simulations and AI non-trivial
  - Challenging to quantify uncertainty
- Foundation models for scientific knowledge discovery, integration, and synthesis
  - Difficult to make broadly accurate
  - Difficult to interface with traditional workloads
- Learning to predict/infer properties and inverse design
  - Want to incorporate constraints from known physics
  - To be trustworthy, prefer explainable/interpretable methods
  - Inverse design should take into account full cycle, such as cost constraints

Cross-cutting challenges

Carter, et al. “Advanced Research Directions on AI for Science, Energy, and Security Report on Summer 2022 Workshops” (2023)

Sampling of AI Approaches and Open Challenges

- Design, prediction, and control of complex engineered systems
  - Operational use requires robustness, reliability, etc.
  - Need infrastructure for coordinating data securely across entities
  - Hardware ”at the edge”
- AI and robotics for autonomous discovery
  - Real-time learning as data is collected by robots
  - Data volumes are growing but require curation, sharing infrastructure, etc.
- Assisting with programming and software engineering
  - Science and engineering codes are not well-represented in large repositories
  - Need to check that the code is correct

Cross-cutting challenges

Carter, et al. “Advanced Research Directions on AI for Science, Energy, and Security Report on Summer 2022 Workshops” (2023)

Tips on Machine Learning for Science

- Brainstorm:
  - Is this a problem that you can solve in your head? What information are you using?
  - What domain knowledge do you have that your ML model might be missing?
- What do you want to generalize across?
- Interpolation is easier than extrapolation
  - Choose a good input representation
  - Carefully consider representative/diverse training data
- Take time to understand the science and end-users
- How can you validate your model? What are the shortcomings or assumptions?
- Can you check your method where you know the answer?
- Many systems have lower-dimensional structure

Conclusions

- Applications of ML to scientific fields are “non-traditional” but of growing interest
- Today we will discuss:
  - Fundamentals of deep learning
  - How to make deep learning training faster
  - Scaling deep learning across multiple GPUs
  - Applications of deep learning to science
  - Responsible use of AI in the sciences

Main room:

- Introduction to deep learning
- Introduction to convolutional neural networks

Breakout room:

- Building data pipelines for deep learning
- Profiling deep learning

Hands-on,
TensorFlow

Questions later? Bethany Lusch, blusch@anl.gov