OAK RIDGE National Laboratory

Issues and Opportunities in the Application of Artificial Intelligence and Machine Learning

Smoky Mountain Mobility Conference

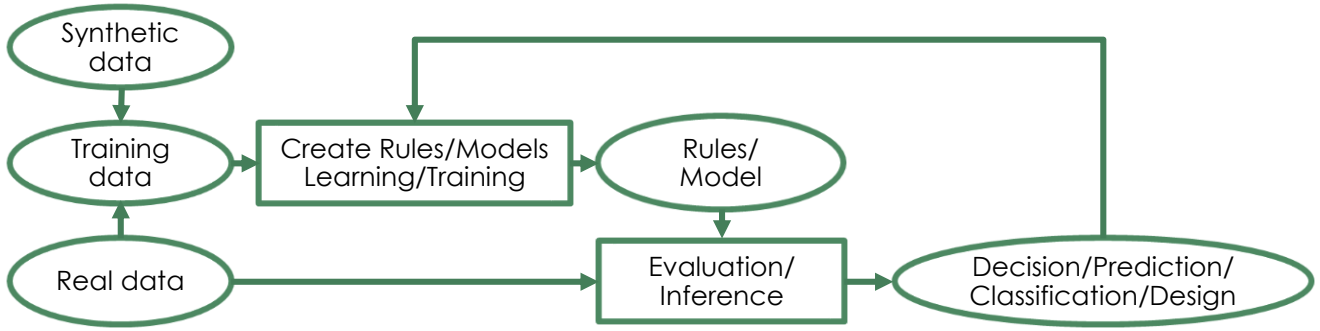
David E. Womble
Oak Ridge National Laboratory

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

What are Artificial Intelligence (AI) and Machine Learning (ML)

- Definition 1: The scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines. (AAAI)
- Definition 2: Computers trained to perform tasks that if performed by a human would be said to require intelligence
- Definition 3: A class of data analytics algorithms in which the rules and/or models are not known a priori and are learned as part of the process



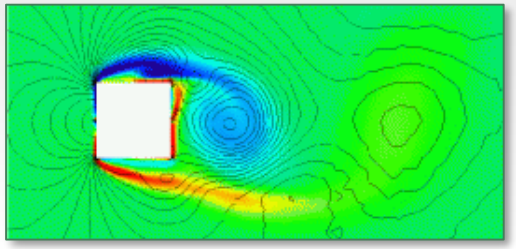
An AI Taxonomy

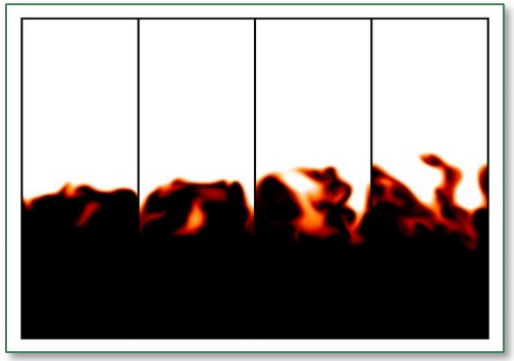
Classification and regression

Near Infrared (single band) WorldView-3 image

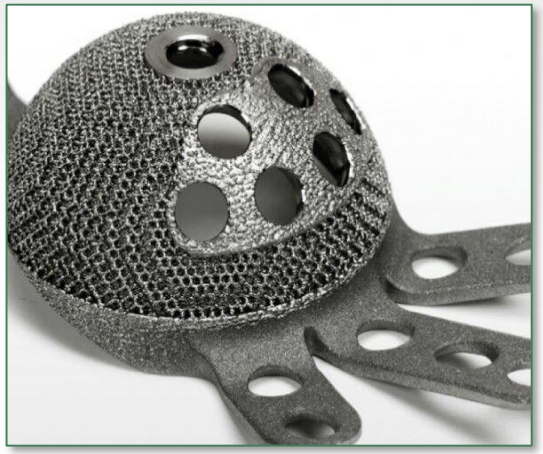
CODA cloud detection saliency map for image above

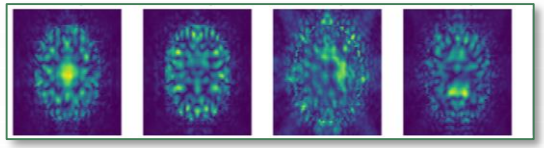
Surrogates





Inverse problems, design and optimization

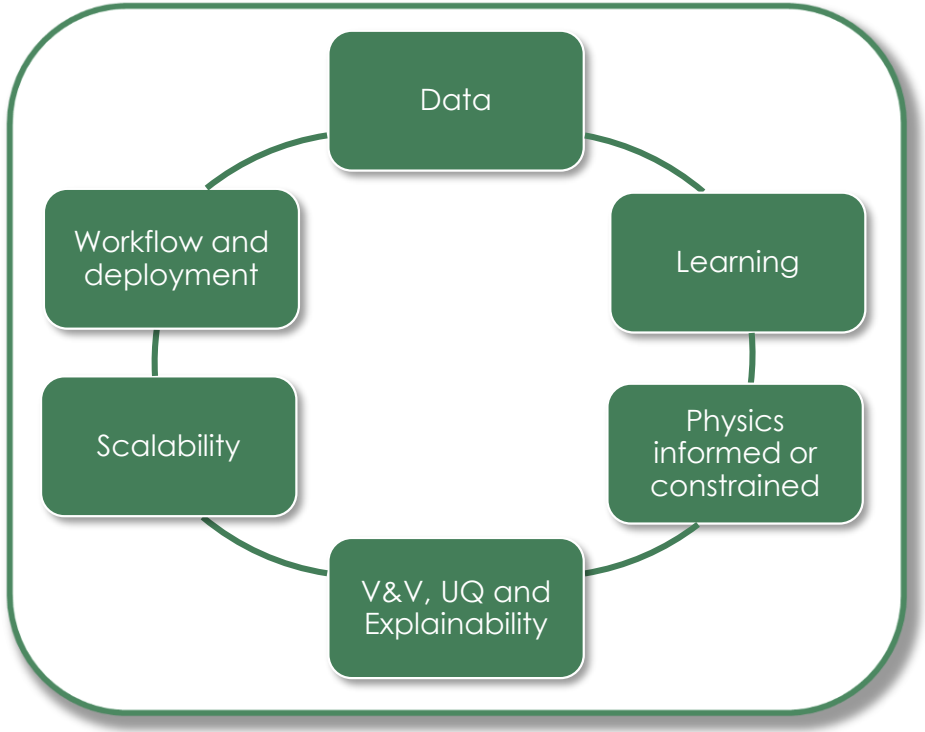




Control systems

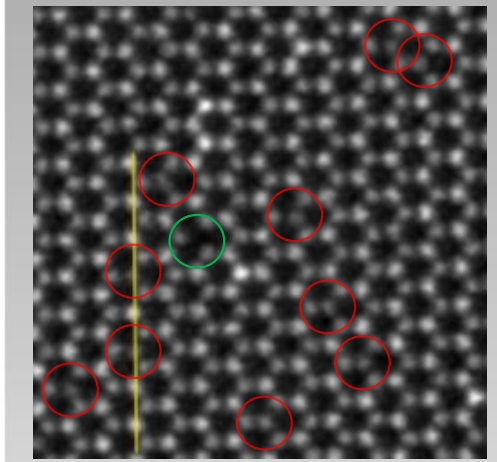
Six Research Areas Crosscut the Taxonomy

- Data quality and statistics
  - Even if we have enough data, it is not necessarily good data
  - Dealing with bias
- Machine learning
  - Needs to accelerate
  - Very model dependent
- Merging physics and AI
  - We can't violate the laws of physics
  - Characterizes ORNL data
- Verification, validation and explainability
  - Is the answer right, is the model appropriate, and can we understand it
  - What is the human-computer interface
- Computing
  - How do we use "big" computers
  - How do we use accelerated nodes
- Workflow and deployment
  - Computing at the edge
  - privacy, ethics and regulations

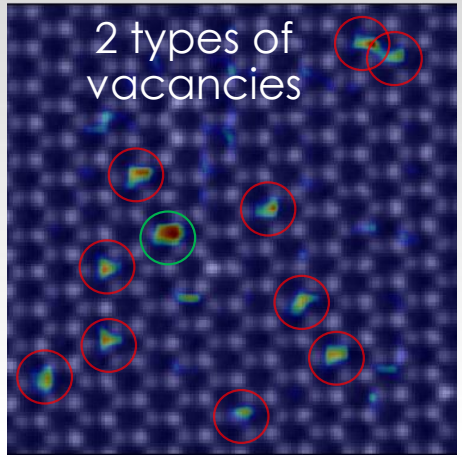


Use case: AI in materials science

Defect detection

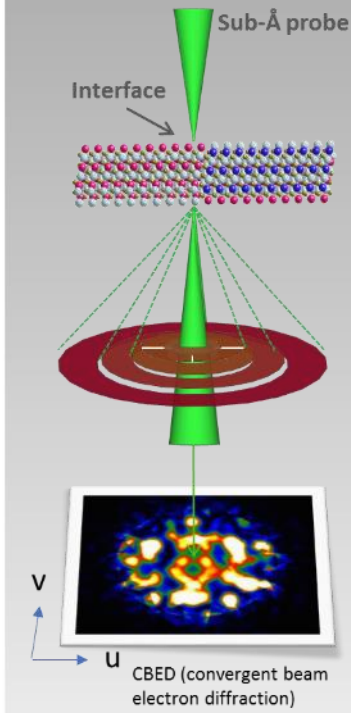


Experimental image

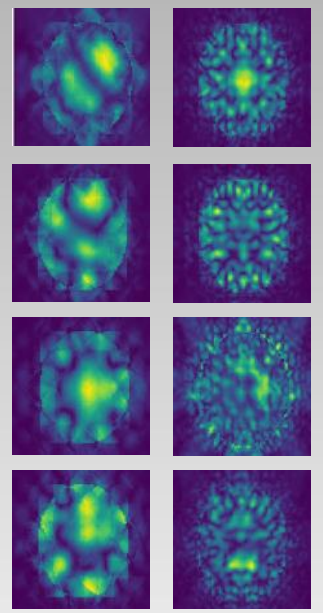


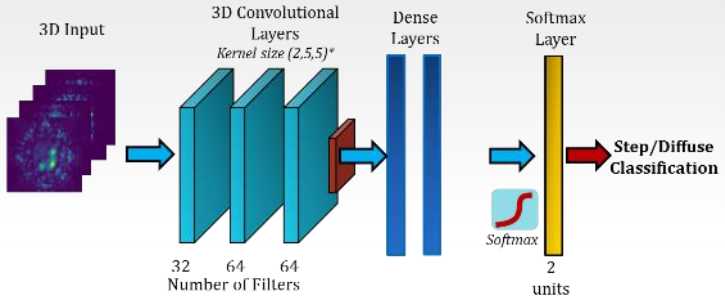
Model output

Inverse problems

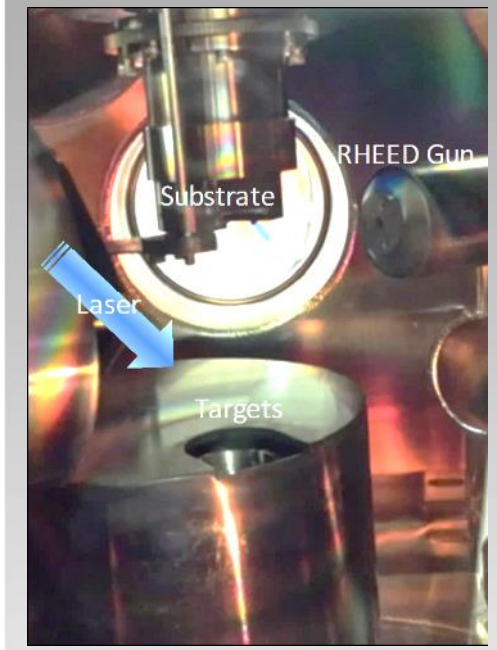


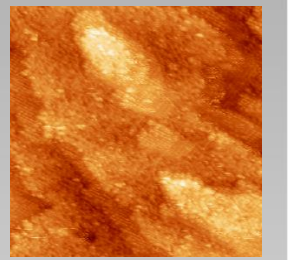
Interface structure classification

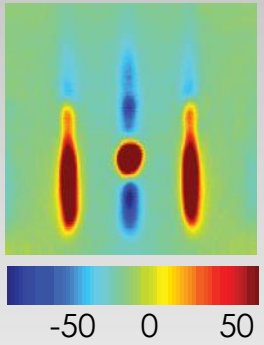




Process control



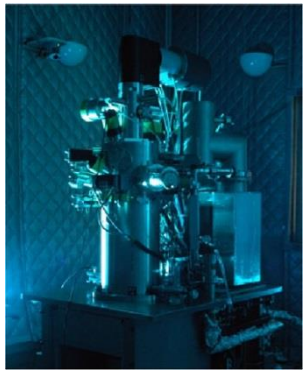


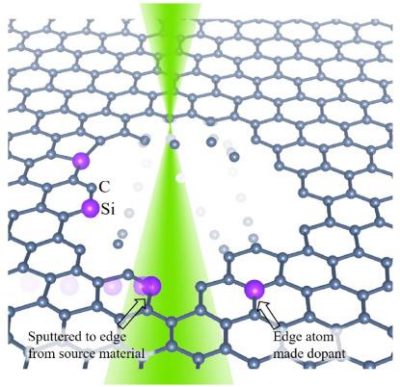


- Design and control of thin-film heterostructures
- Prediction of growth conditions and control of trajectory

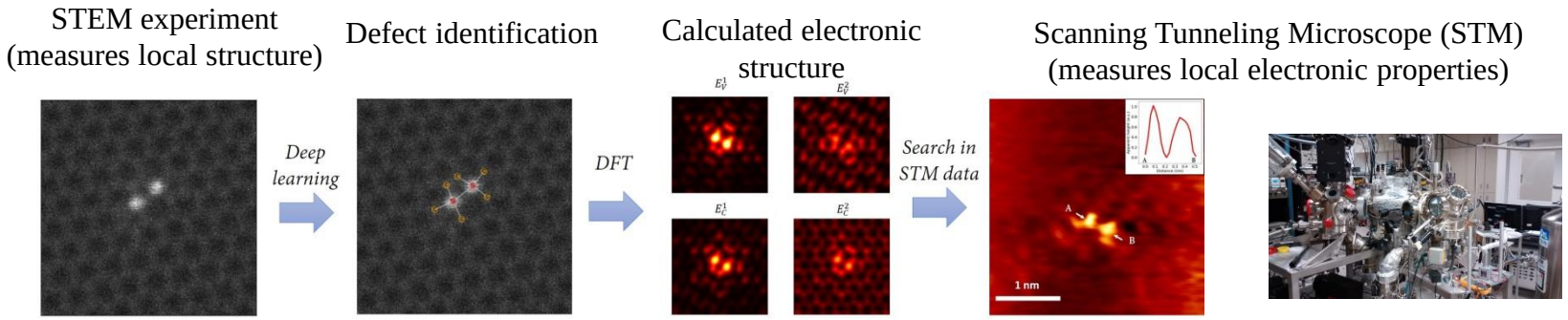
Building and exploring libraries of atomic defects in graphene

- STEM generates large amounts of data (GB to TB range per single experiment)
- Atomic positions are key for understanding atomic-scale processes in materials
- Current state-of-the-art approaches for atom/defect identification are slow and frequently fail for noisy data



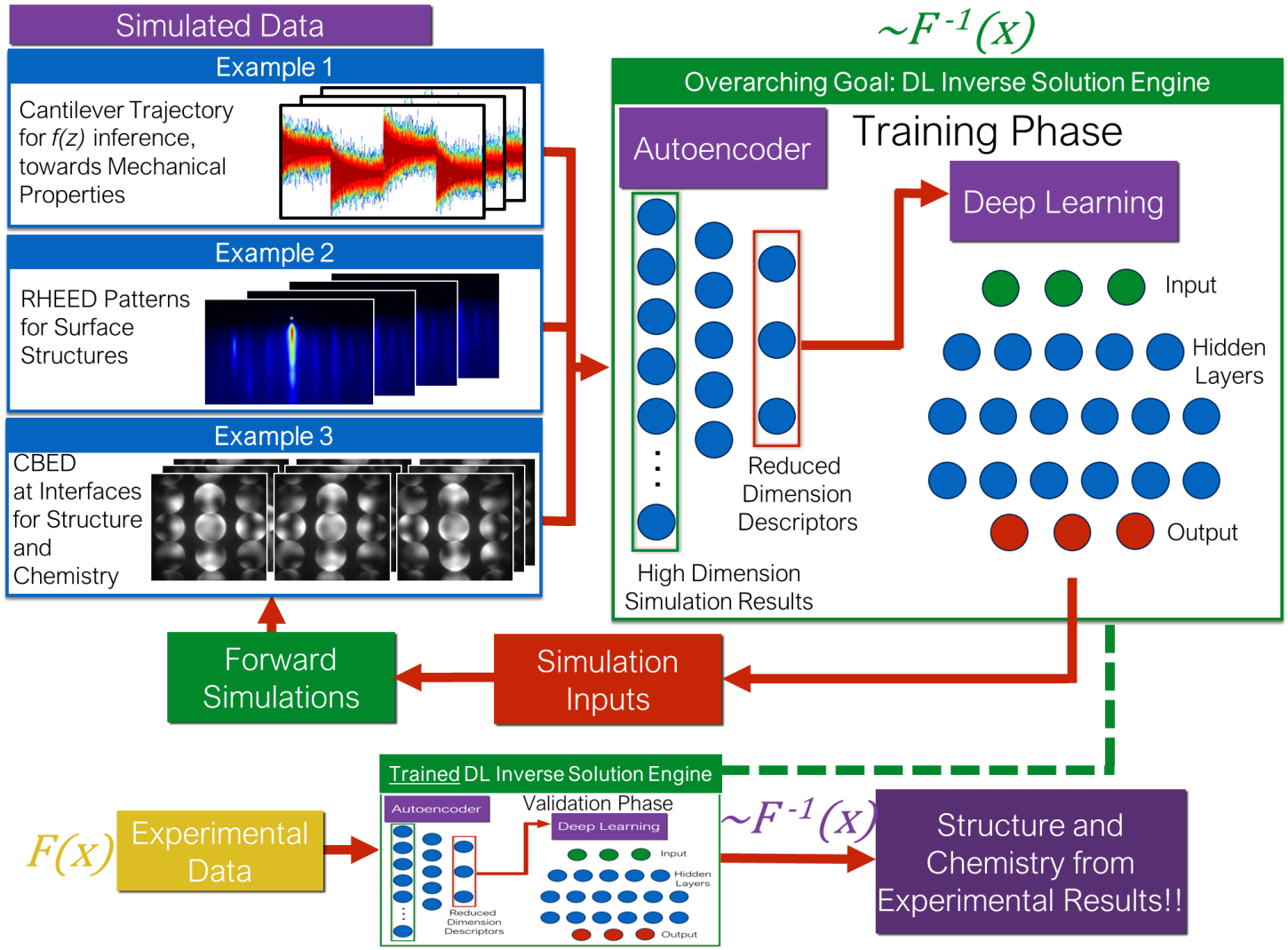


Building and exploring libraries of atomic defects in graphene



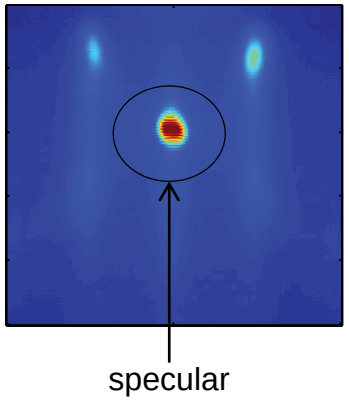
M. Ziatdinov, O. Dyck, B. G. Sumpter, S. Jesse, R. K. Vasudevan, S. V. Kalinin. *Building and exploring libraries of atomic defects in graphene: scanning transmission electron and scanning tunneling microscopy study*. ArXiv:1809.04256 (2018)

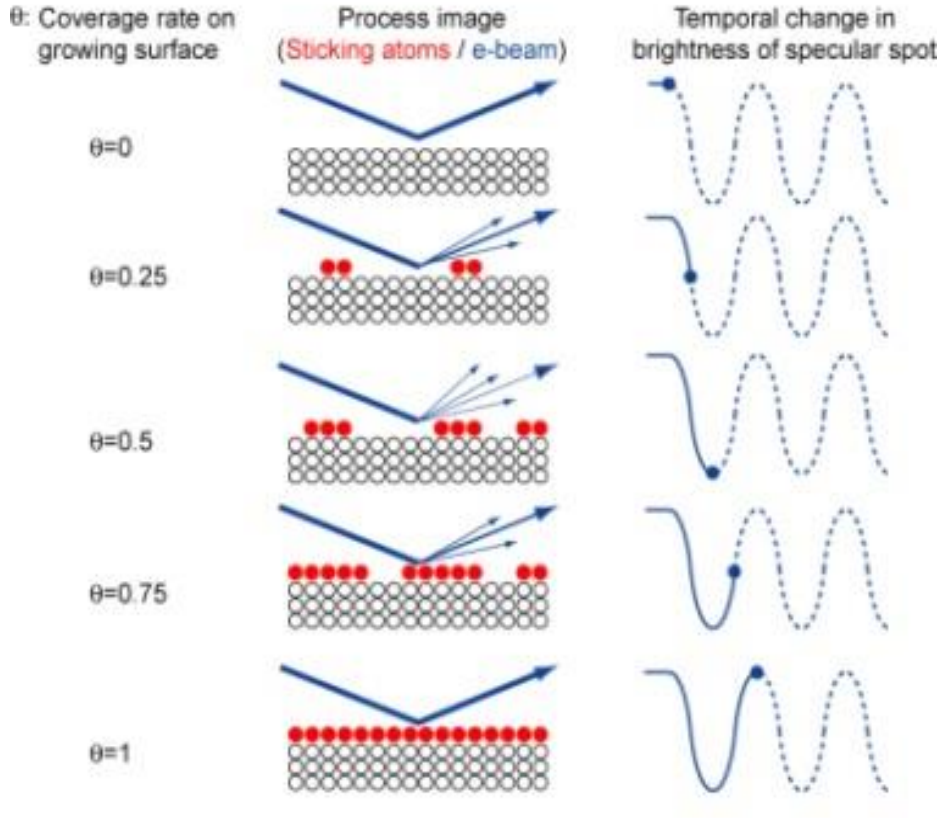
Physics-based Inverse Problems



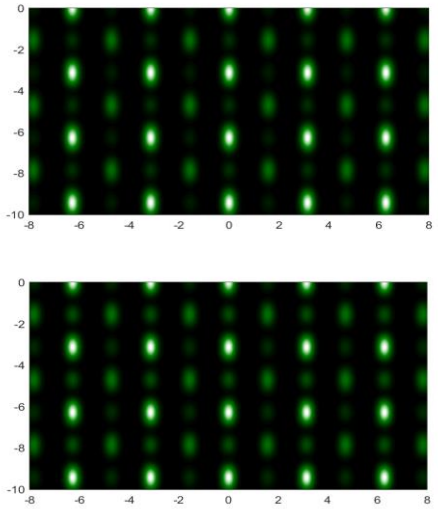
Reflection High Energy Electron Diffraction (RHEED)

Reflection High Energy Electron Diffraction (RHEED) is an indispensable technique to monitor film properties (thickness, surface ordering, etc.) during growth by pulsed laser deposition (PLD) or Molecular Beam Epitaxy (MBE)

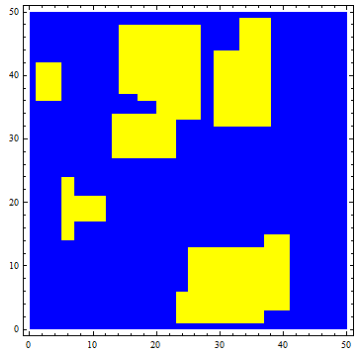




www.pascal-co-ltd.co.jp



Example of two different island morphologies

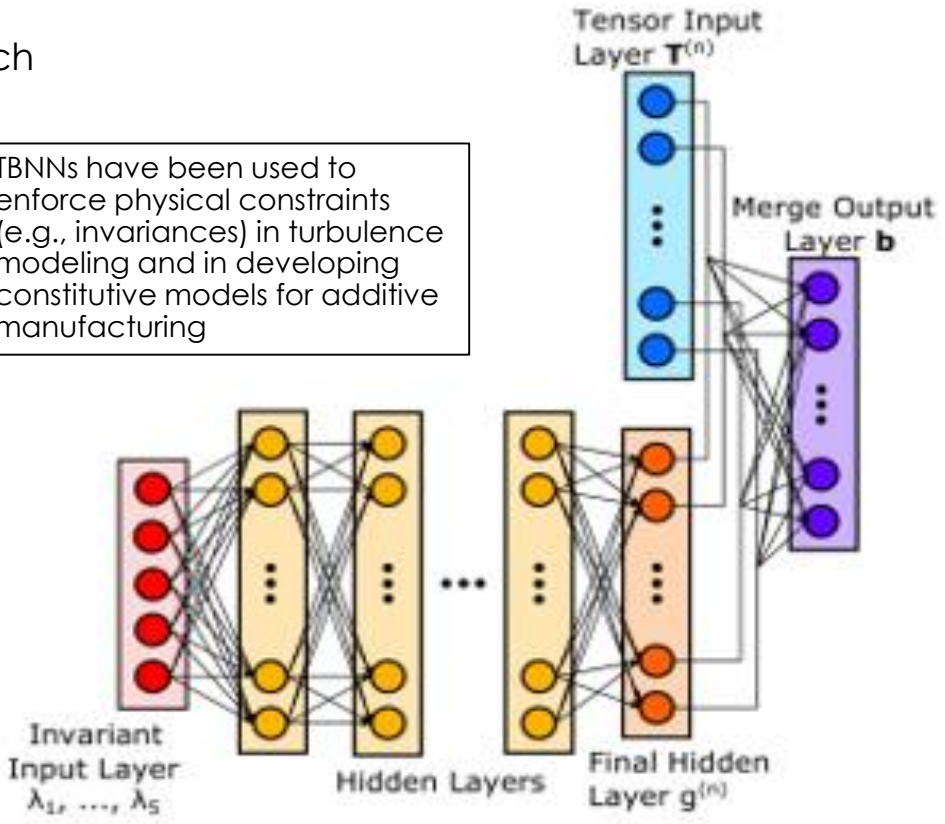


Example image of one island cell

Surrogates

- Surrogates
  - When traditional mod-sim is too expensive (e.g., when many function evaluations are needed, such as in ensembles)
  - When working across scales (e.g., turbulence closure)
  - When we don't have physics-based models (e.g., some bio-science problems, or fracture prediction)
  - To steer computations
- Some examples where surrogates have been useful
  - Constitutive models, e.g., for additive manufacturing
  - Turbulence models
- Surrogates can often be trained with synthetic data. But they are still models that need to be validated. Uncertainties need to be computed.

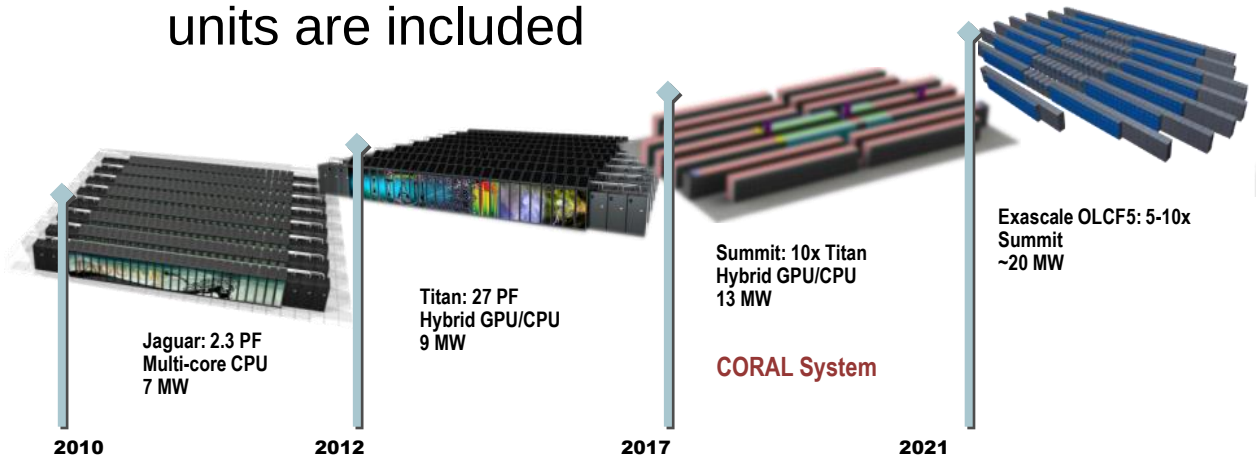
TBNNs have been used to enforce physical constraints (e.g., invariances) in turbulence modeling and in developing constitutive models for additive manufacturing



$$\boldsymbol{b} = \sum g^{(n)}(\lambda_1, \ldots, \lambda_5)\, \boldsymbol{T}^{(n)}$$

Enabled by HPC

- We have the ability to collect and store large amounts of data
- Computational power continued to increase, with architectural improvements that are amenable to neural networks
  - For example, GPU became practical for accelerated computations.
  - Reduced-precision tensor core units are included



OAK RIDGE National Laboratory

Summit Overview

Summit includes

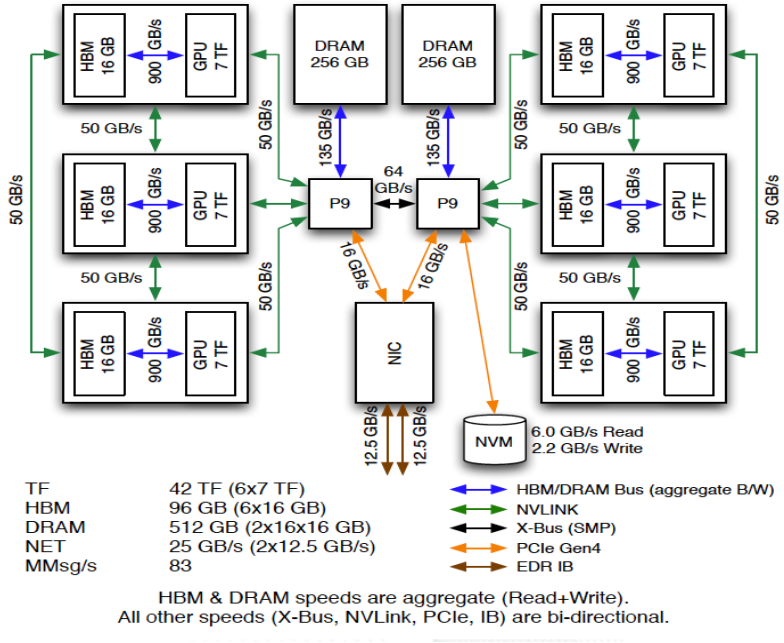
- 4608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM Spectrum Scale file system transferring data at 2.5 TB/s

Each node has

- 2 IBM POWER9 processors
- 6 NVIDIA Tesla V100 GPUs
- 608 GB of fast memory
- 1.6 TB of NVMe memory

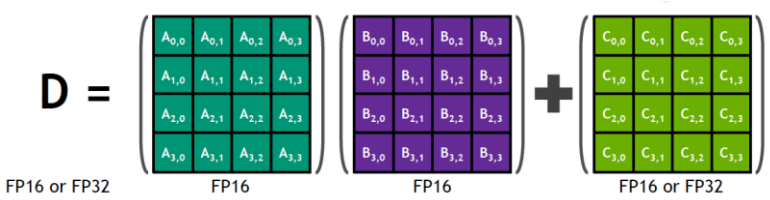
System Performance

- Peak performance of 200 petaflops for modeling & simulation
- Peak of 3.3 ExaOps for data analytics and artificial intelligence



NVIDIA's tesla v100

- 5,120 CUDA cores (64 on each of 80 SMs)
- 640 NEW Tensor cores (8 on each of 80 SMs)
- 20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
- 300 GB/s NVLink
- 7.5 FP64 TFLOPS | 15 FP32 TFLOPS | 120 Tensor TFLOPS
- ~57 times faster in 64-bit peak floating point performance than the CM-5 we worked on 25 years ago
- >27K of these coming on ORNL's Summit system!
- Mixed precision matrix math 4x4 matrices



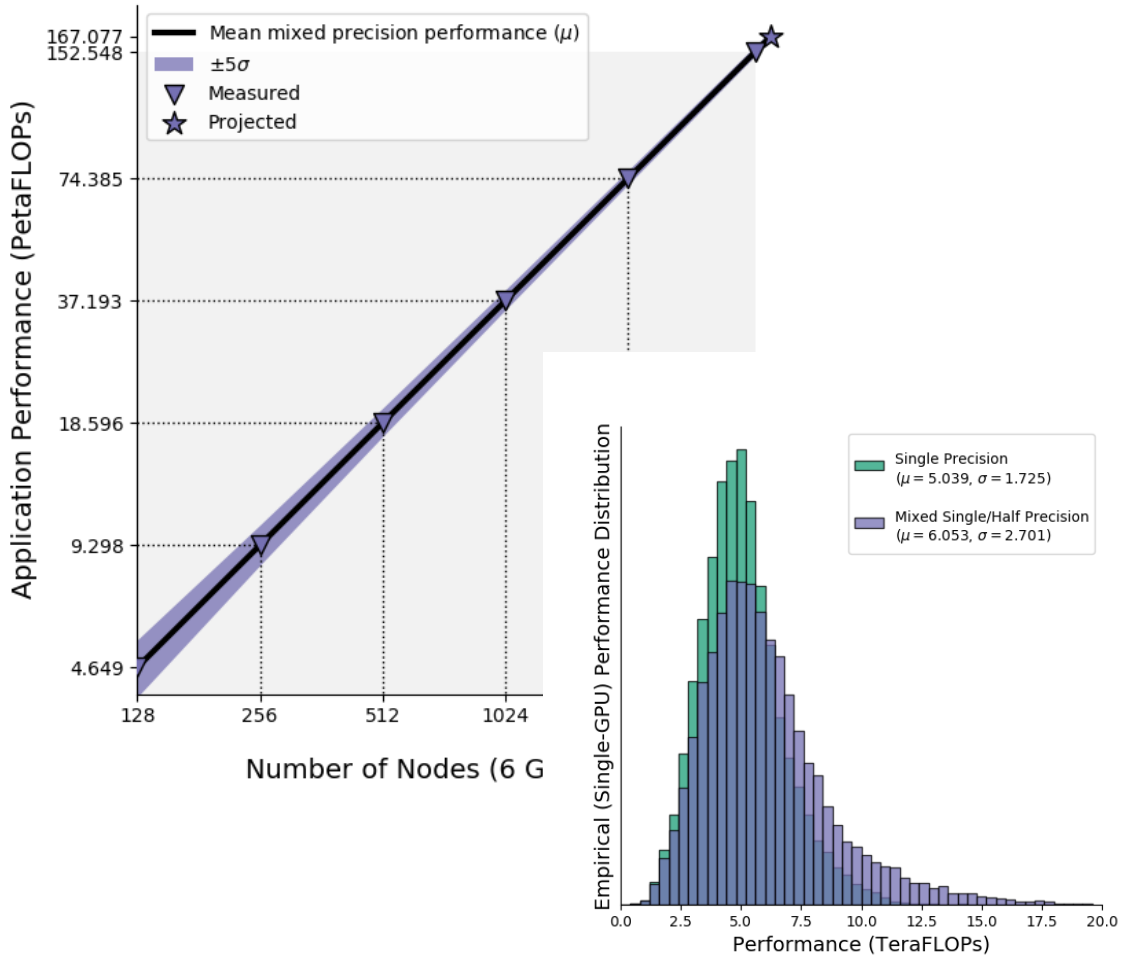
D = AB + C

Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- The M&S community must figure how out to "cheat" and utilize mixed / reduced precisions
- Ex: Jack Dongarra shows he can get 4x FP64 peak for 64bit LU on V100 with iterative mixed precision (using GMRES!)

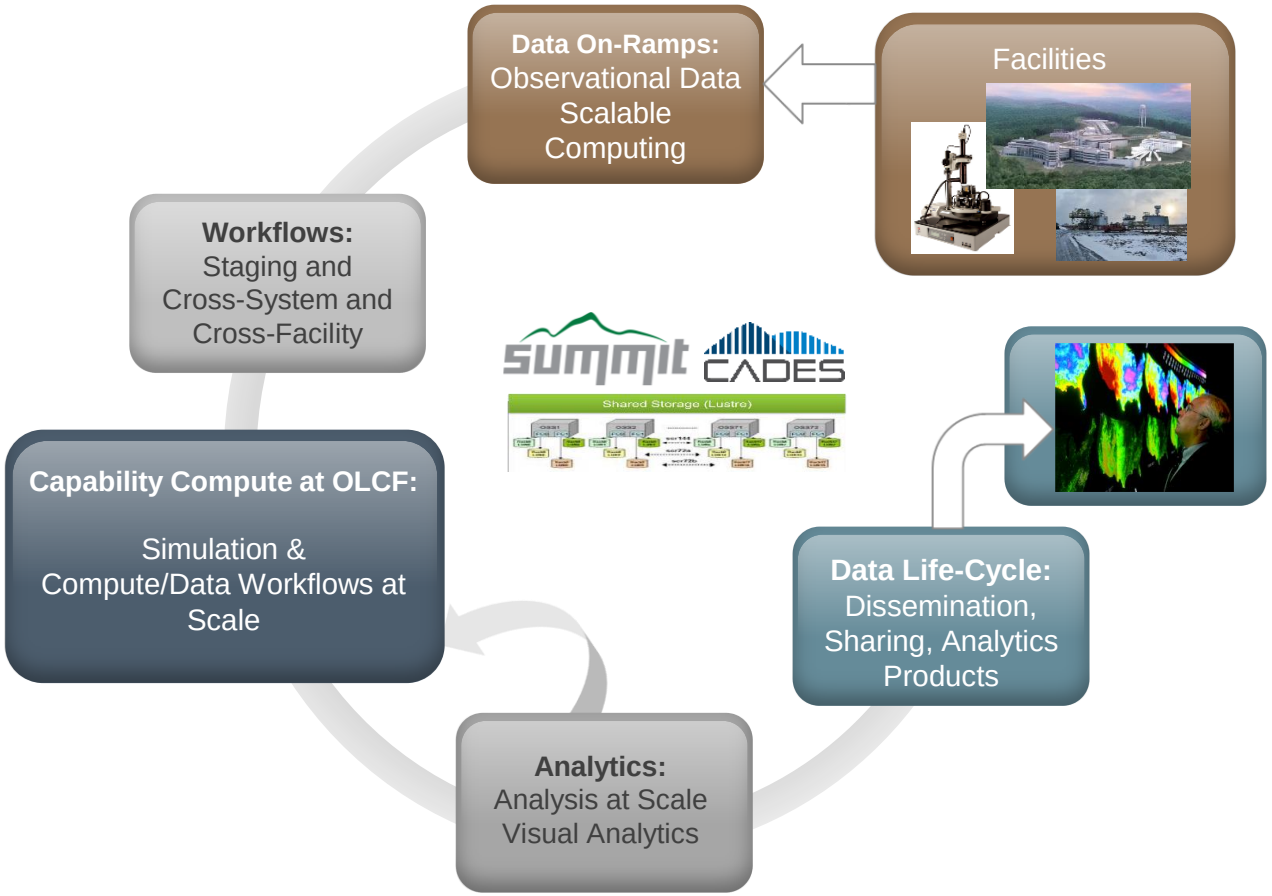
Using Tensor Cores is key to high performance

- MENNDL is the kernel of one of our Gordon Bell finalists
  - Determines optimal hyperparameters for a DNN
  - Relatively easy to parallelize
  - Effectively demonstrates the power of the Tensor core units (as well as the challenge of using them)
- Mixed precision presents algorithmic challenges
  - What accuracy is actually needed for simulations
  - Performance must now be correlated to accuracy



Travis Johnston, ORNL

A Good Infrastructure Is Required To Manage Data



Issue: "Syntactic" Space vs. "Semantic" Space

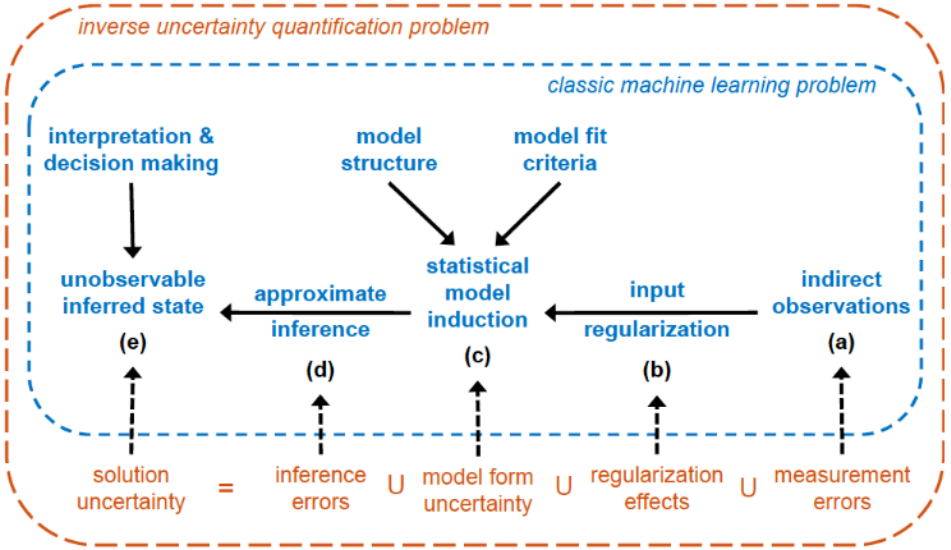
- Humans tend to think in semantic space, i.e., in terms of the meaning.

  And metrics in semantic space are fundamentally different from those in syntactic space

- Implications
  - Easy to spoof classification systems
  - Transfer learning doesn't map well. (Humans tend to transfer learning in semantic space, e.g., transfer what I learned about human behavior in kindergarten to how I drive. Most AI approaches transfer in syntactic space or transfer parts of the model (a sort of "gene transfer").

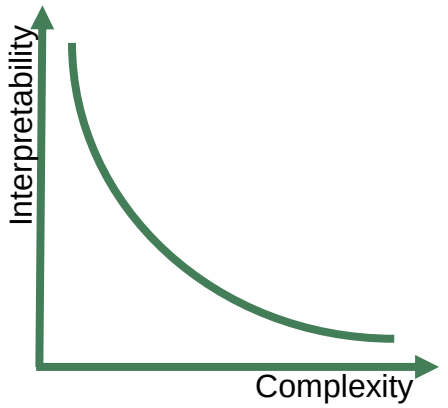
Issue: Uncertainty Quantification

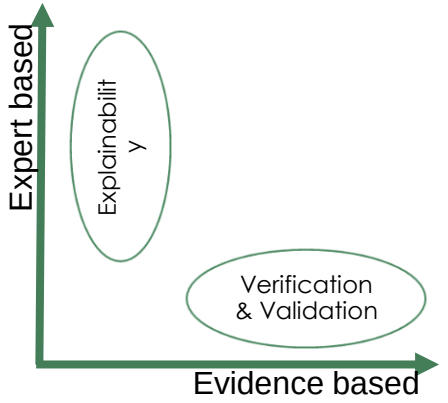
- An AI is simply a model and has uncertainty. UQ can also enhance explainability and support decision making.
- Types of UQ
  - Uncertainty propagation
  - Calibration uncertainty
- Bayesian methods/variational inference common, but are computationally expensive.



Issue: Verification, Validation, Explainability and Interpretability

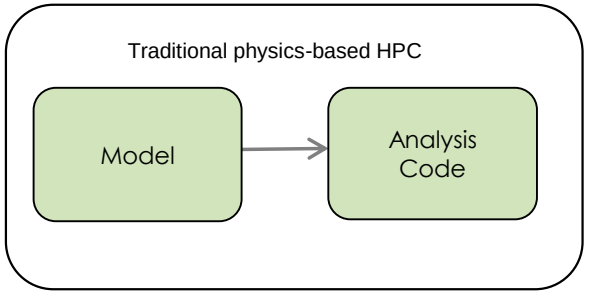
- Interpretibility
  - Can a human understand the model? For example, do the basis vectors in a dimension reduction algorithm have a physical meaning?
- Explainability
  - Can the model present a sequence of steps that can justify the answer to an expert?
  - Expert based
- Reproducibility
  - Does the same experiment lead to the same conclusion?
  - Can we run different experiment and not contradict our conclusion?
  - If we create a new model with the same data, do we get the same conclusions?
  - Required for good science
-





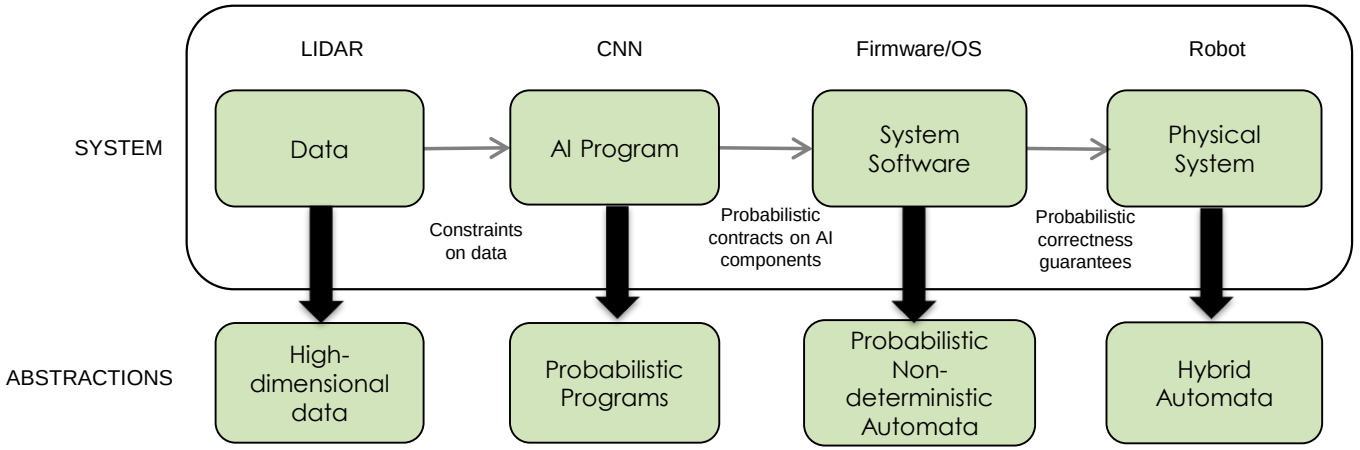
Issue: Verification, Validation, Explainability and Interpretability

- Verification
  - Is the model implemented correctly?
- Validation
  - Is the model (including training data) appropriate for the decisions being made?
  - Must be evidence based
  - Requires some form of UQ, robustness guarantees and bounds on “distortion”



Issue: Verification, Validation, Explainability and Interpretability

- Verification
  - Is the model implemented correctly?
- Validation
  - Is the model (including training data) appropriate for the decisions being made?
  - Must be evidence based
  - Requires some form of UQ, robustness guarantees and bounds on “distortion”



Issue: Data Is A Major Problem

- Need more data than was imagined just a few years ago
  - We are looking for complex correlations
  - Using primarily statistical methods
- Labelled data is a problem
  - Generating labels is expensive and labor intensive (e.g., Mechanical Turk)
  - Need to move toward reinforcement learning
- Synthetic data and simulated environments are partial solutions
  - But an AI can learn the flaws in these systems

Issue: AI Is An Art

- Choosing the model form and hyper parameters is often ad-hoc and requires experience and insight
- AI models must be tuned
- Neural networks design is difficult and often requires tuning
- Interpreting the results requires expertise

> "Machine learning methods are often described in papers at an abstract level, for maximum generality. However, a good choice of hyperparameters is usually necessary to make them work well on real-world problems, and tricks are often used to make most efficient use of these methods and extend their capabilities."
>
> G. Montrevan, et.al., "Methods for Interpreting and Understanding Deep Neural Networks."

Summary: Observations and Issues

- AI is effective for narrowly defined tasks and only identifies correlations in (complex) data
- Access to and availability of “good” and “labelled” data is one of the biggest challenges for AI
- We need a sustainable data and compute infrastructure
- While we have big machines, we don’t have scalable algorithms
- Vulnerability threats for AI (hacking, intentional manipulation) are a huge concern for deployment
- We don’t know how to do validation
- HCI is an important component of the workflow, including explainability and interpretability
- Deployment of AIs introduces a whole new set of challenges
- Need to understand the ethics and human impact