U.S. DEPARTMENT OF ENERGY | Office of Science

Scientific Machine Learning
Basic Research Needs (BRN) Workshop

Workshop held January 30 - February 1, 2018
Workshop Chair: Nathan Baker, PNNL
https://www.orau.gov/ScientificML2018/

Report to Advanced Scientific Computing Research Advisory Committee (ASCAC)

Steven Lee (DOE ASCR)
April 17, 2018

Summary of Charge Letter for Scientific Machine Learning Workshop

Greater machine learning-based prediction & decision-support capabilities are needed to address & anticipate DOE mission challenges:

- **DOE scientific user facilities drive rapid growth in data** from experiments, observations, and simulations
- Increasingly powerful science technologies are driving the need for **algorithms & automation to facilitate the use of advanced technologies** for science breakthroughs

The charge for the workshop is:

- First consider the status, recent trends, and broad use of machine learning for scientific computing
- Examine the opportunities, barriers, & potential for high scientific impact through fundamental advances in the underlying research foundations
- ASCR grand challenges & resulting priority research directions should span several major machine learning categories & state-of-the-art modeling & algorithms research
- ***Identify the basic research needs & opportunities that can potentially enable machine learning-based approaches to transform the future of science and energy research.***

Working Definitions of Machine Learning

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

- Arthur Samuel, 1959

Machine Learning: A set of rules that allows systems to learn directly from examples, data and experience.

- Royal Society, 2017

“Learning” is the process of transforming information into expertise or knowledge; “Machine learning” is automated learning.

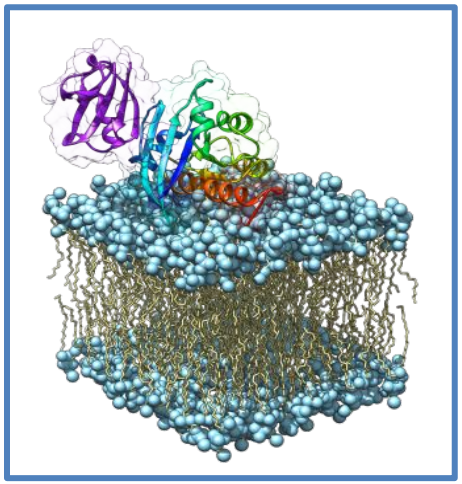
- Paraphrased from Jordan et al., 2015

Examples of Popular Machine Learning Methods

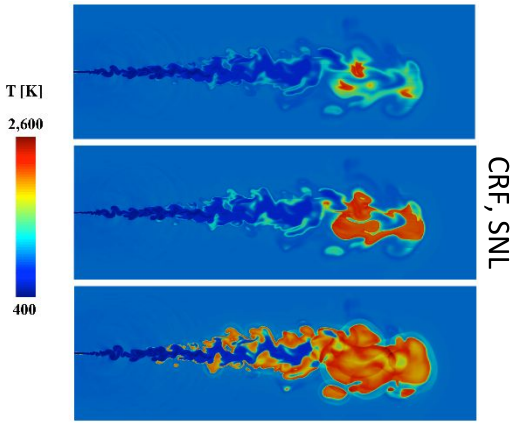
Deep Learning	Convolutional Neural Network, Deep Boltzmann Machine & Belief Networks, Stacked Auto-Encoders
Bayesian	Naive Bayes, Averaged One-Dependence Estimators, Bayesian Belief Networks, Gaussian Naive Bayes, Multinomial Naive Bayes, Bayesian Network
Ensemble	Random Forests, Gradient Boosting Machines, Boosting, Boostrapped Aggregation, AdaBoost, Stacked Generalization, Gradient Boosted Regression Trees
Decision Tree	Classification and Regression Tree, Iterative Dichotomizer 3, C4.5, C5.0, Chi-squared Automatic Interaction Detection, Decision Stump, Conditional Decision Trees, M5
Neural Networks	Radial Basis Function Network, Perceptron, Back-Propagation, Hopfield Network
Dimensionality Reduction	Principal Component Analysis & Regression, Partial Least Squares Regression, Multidimensional Scaling, Projection Pursuit; Partial Least Squares-, Mixture-, Quadratic-, & Linear Discriminants
Regularization	Least Absolute Shrinkage & Selection Operator (LASSO), Elastic Net, Least Angle Regression
Instance-Based	k-Nearest Neighbor, Learning Vector Quantization, Self-Organizing Map, Locally Weighted Learning
Clustering	k-Means, k-Medians, Expectation Maximization, Hierarchical Clustering
Regression	Linear-, Ordinary Least Squares-, Stepwise-, and Logistic Regression; Multivariate Adaptive Regression Splines, Locally Estimated Scatterplot Smoothing (LOESS)

Why: Define research challenges and directions for scientific machine learning

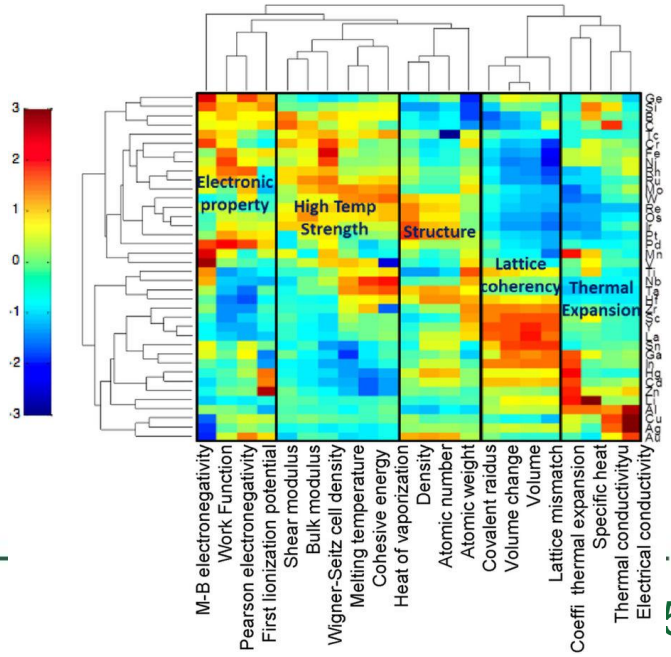
- Machine learning use is on the rise throughout science domains
- However, many popular ML methods lack mathematical approaches to understand robustness, reliability, etc.
- ASCR Applied Mathematics has a long track record for building mathematical foundations to critical computational tools
- **Workshop to help ASCR define the grand challenges and priority research directions for scientific machine learning**



Timo Bremer, LLNL



CRF, SNL



Krishna Rajan, Buffalo

What: Deliverables and Products from this Workshop

- A BES Basic Research Needs-inspired process

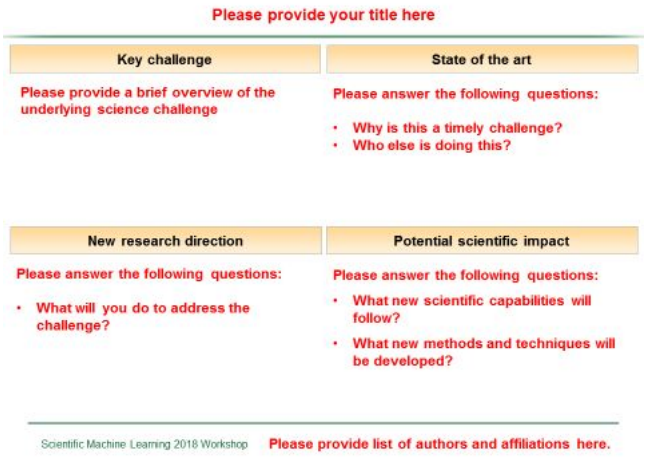
- Pre-workshop report
  - Factual status document describing the Scientific Machine Learning (SciML) landscape as it relates to ASCR

- Workshop deliverables
  - Articulation and refinement of grand challenges for SciML
  - Priority research directions for SciML

- Post-workshop report
  - Incorporate updated factual status document
  - Incorporate workshop deliverables

What is a "Priority Research Direction"?

- High-priority area of research for scientific machine learning
- Has following components (ala Heilmeier):
  - Clear statement of key challenge
  - Context in the current scientific landscape to establish timeliness and competition
  - Plausible research pathway(s)
  - Clear scientific impact
- *It is not*
  - A proposal for a specific project
  - Your favorite area of research without connection to SciML themes



How: Workshop Components

- A Basic Research Needs-inspired agenda

- Plenary talks
  - Highlight status of machine learning, challenges, open questions
- Panel discussions
  - Summarize pre-workshop report
  - Provide perspectives across DOE ASCR facilities, ECP, and other organizations
- Breakout sessions
  - Organized around ~140 submitted Position Papers presented as flash talks
  - The “work” in workshop: Crucible for new Priority Research Directions
  - Need high levels of interaction and input (long days…)
  - Brainstorming (Day 1), Refining (Day 2) & Presenting (Day 3) Priority Research Directions

Agenda Overview - Tuesday

- Note: **Observers** will be able to watch plenaries, panels, and breakout summaries via Zoom webinar
- **Tuesday**
  - Welcome
  - *Scientific Machine Learning: ASCR Facilities Perspective*
  - *Three Principles of Data Science: Predictability, Stability, and Computability*: **Bin Yu**
  - *Scientific Machine Learning across Federal Agencies*
  - Summary of Pre-Workshop Report & Themes
  - *Physics, Structure, and Uncertainty: Probabilistic Learning for Risk Mitigation:* **Roger Ghanem**
  - Parallel breakout sessions

Agenda Overview – Wednesday & Thursday

- **Wednesday**
  - *Machine Learning in the Wild:* **Jacob Shapiro**
  - Preliminary breakout reports and discussion
  - *Challenges & Scope of Empirical Modeling:* **Ronald Coifman**
  - Parallel breakout sessions

- **Thursday**
  - Final breakout reports and discussion
  - Summary of priority research directions

Breakout sessions

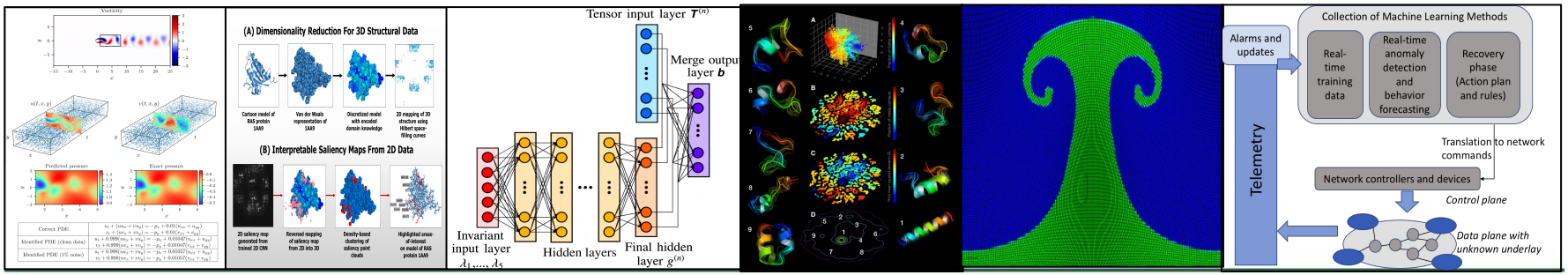
- ***Numerical Analysis for Machine Learning***
  Mark Ainsworth, James Sethian
- ***Machine Learning, Multifidelity, & Reduced-order Models***
  Karen Willcox, Abani Patra
- ***Machine Learning, Optimization, & Complexity***
  Stefan Wild, Manish Parashar
- ***Probabilistic Machine Learning***
  Habib Najm, Aric Hagberg
- ***Machine Learning Interpretability***
  Timo Bremer, Yannis Kevrekidis

Workshop Approach

- Each Breakout Session developed a list of critical research areas.
- Day 2: Research areas were evaluated & grouped into topics according to **5 Priority Research Directions.**
- The Session leads & members joined the relevant Priority Research Direction (PRD) group.
- The PRD teams met to formulate the research approaches and thrust areas.
- Day 3: Report out and writing of PRDs and Panel reports.

Scientific Machine Learning: Priority Research Directions (PRDs)

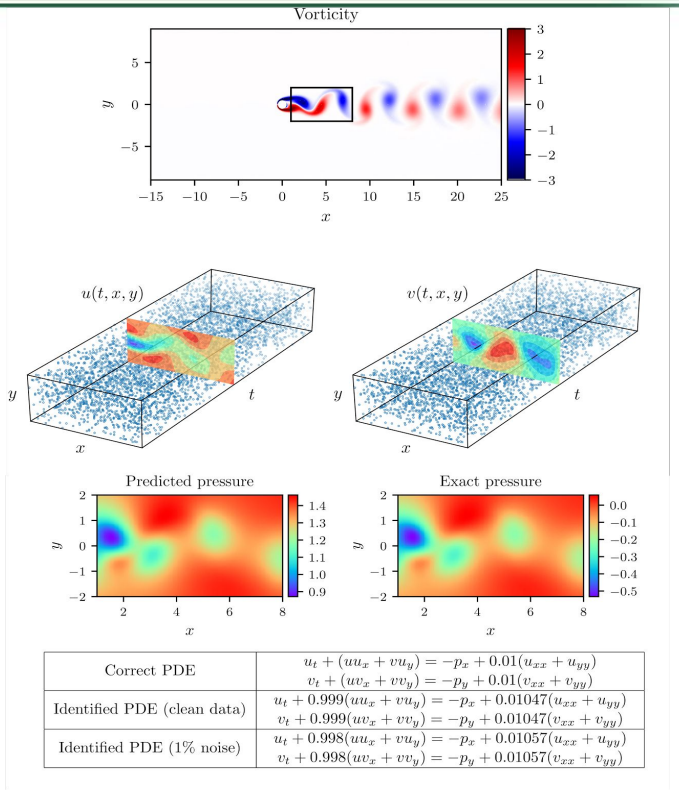
Foundational Themes	Capabilities Research
PRD1. Domain-Aware SciML Leveraging scientific domain knowledge	**PRD4. Data-Intensive SciML** Automated scientific inference & data analysis
PRD2. Interpretable SciML Explainable & understandable results	**PRD5. Inner-Loop SciML** Machine learning-embedded models & algorithms for better scientific computing tools
PRD3. Robust SciML Stable, well-posed & efficient formulations	**PRD6. Outer-Loop SciML** Automated decision-support, optimization, resilience, & control for complex systems & processes



PRD1: Domain-Aware Scientific Machine Learning
Leveraging Scientific Domain Knowledge

Key Points: How can domain knowledge be effectively incorporated into Scientific ML methods?

- Established domain models based on physical mechanism & scientific knowledge
- SciML offers significant opportunity to complement traditional domain models
- Domain knowledge: physical principles, symmetries, constraints, computational predictions, uncertainties, etc
- Potential to improve accuracy, interpretability, & defensibility while reducing data requirements & accelerating training process



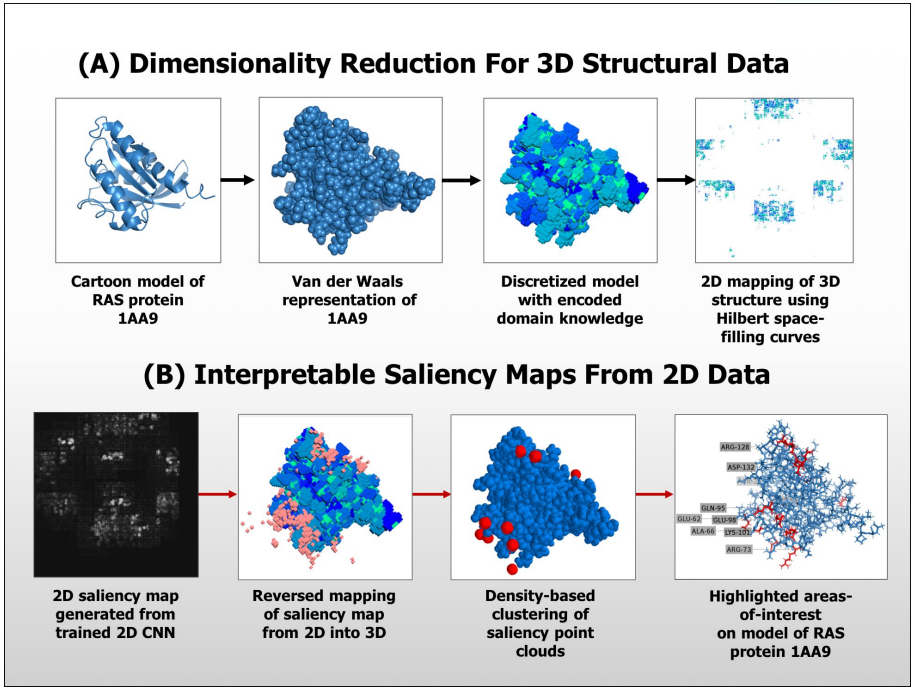
Correct PDE	$u_t + (uu_x + vu_y) = -p_x + 0.01(u_{xx} + u_{yy})$ $v_t + (uv_x + vv_y) = -p_y + 0.01(v_{xx} + v_{yy})$
Identified PDE (clean data)	$u_t + 0.999(uu_x + vu_y) = -p_x + 0.01047(u_{xx} + u_{yy})$ $v_t + 0.999(uv_x + vv_y) = -p_y + 0.01047(v_{xx} + v_{yy})$
Identified PDE (1% noise)	$u_t + 0.998(uu_x + vu_y) = -p_x + 0.01057(u_{xx} + u_{yy})$ $v_t + 0.998(uv_x + vv_y) = -p_y + 0.01057(v_{xx} + v_{yy})$

This example illustrates the capabilities obtained by incorporating domain knowledge into a deep neural network. Given scattered and noisy data components of an incompressible fluid flow in the wake of a cylinder, we can employ a physics-informed neural network that is constrained by the Navier-Stokes equation in order to identify unknown parameters, reconstruct a velocity field that is guaranteed to be incompressible and satisfy any boundary conditions, as well as recover the entire pressure field. Figure from: Raissi et al.

PRD2: **Interpretable** Scientific Machine Learning
Explainable and Understandable Results

Key Points: How to balance the use of increasingly complex ML models with the need for users to understand conclusions & derive insights?

- Physical understanding has been the bedrock of modeling
- User confidence linked to the conviction that model accounts for domain knowledge (variables, parameters, physical laws, etc.)
- Need exploration & visualization approaches for “debugging” complex machine learning models
- Need metrics to quantify model differences

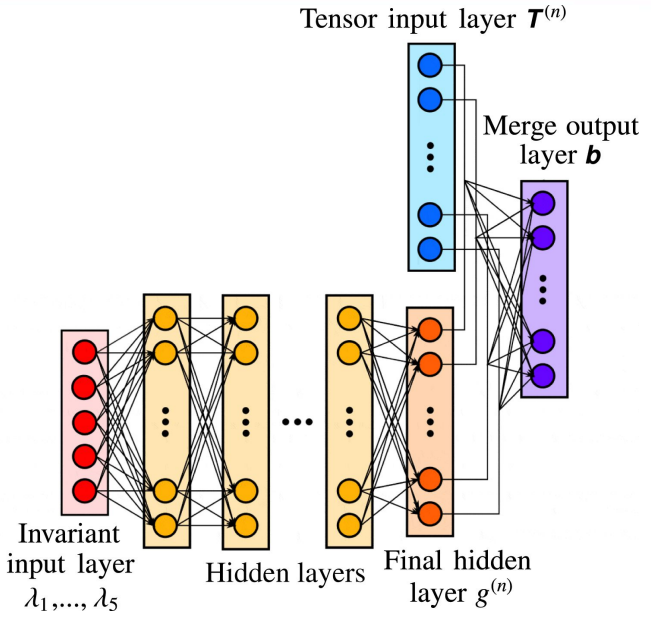


High-level data pipeline overview for dimensionality reduction of 3D protein structures (A) and interpretation of saliency maps from trained CNN model (B). Saliency maps generated from CNN models can then be clustered to identify areas along the 3D structure that are regions that highly influence the output of the CNN model. From these salient regions, specific residues can be identified that fall in close proximity to the salient regions.
Image credit: Rafael Zamora-Resendiz and Silvia Crivelli, LBNL.

PRD3: **Robust** Scientific Machine Learning
Stable, Well-Posed, and Efficient Formulations

Key Points: How can computationally efficient SciML methods be developed and implemented to ensure outcomes are not unduly sensitive to perturbations in training data and model selection?

- SciML methods need to establish the properties of robustness & reliability
- Integration of protocols for verification & validation are in their infancy
- Progress will require research proving that developed methods and implementations are stable and well-posed

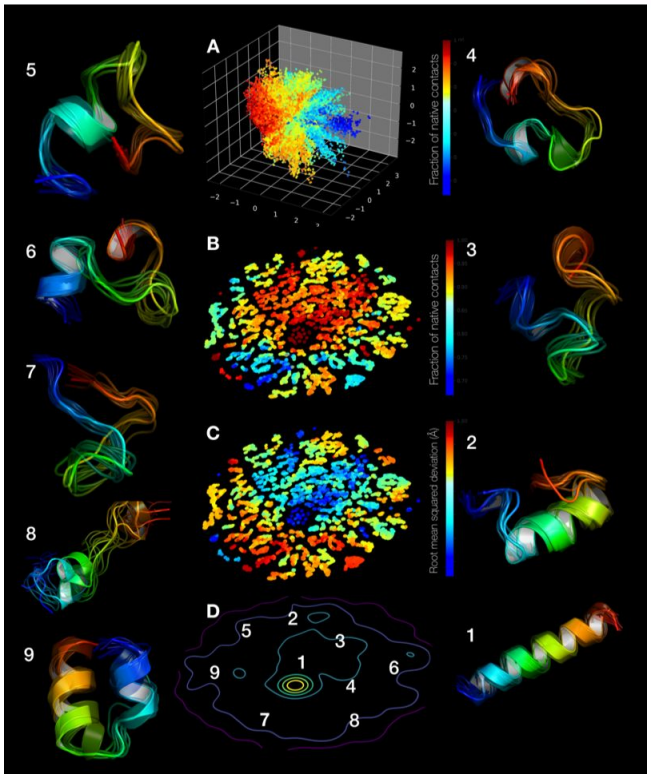


In the context of Reynolds averaged incompressible turbulence modeling, a neural network has been used in an eddy viscosity turbulence closure model. From physical arguments, the model needs to satisfy rotational invariance, ensuring that the physics of the flow is independent of the orientation of the coordinate frame of the observer. A special network architecture, a tensor basis neural network (TBNN), embeds rotational invariance by construction. Without this guarantee, the NN model evaluated on identical flows with the axes defined in different directions could yield different predictions.
Image credit: SNL.

PRD4: Data-Intensive Scientific Machine Learning
Automated Scientific Inference & Data Analysis

Key Points: What novel approaches can be developed for reliably finding signals, patterns or structure within high-dimensional, noisy, uncertain input data?

- SciML methods require the development of improved methods for statistical learning in high-dimensional SciML systems with noisy and complex data
- Need approaches required to identify structure in complex high-dimensional data
- SciML requires efficient sampling in high-dimensional parametric and model spaces

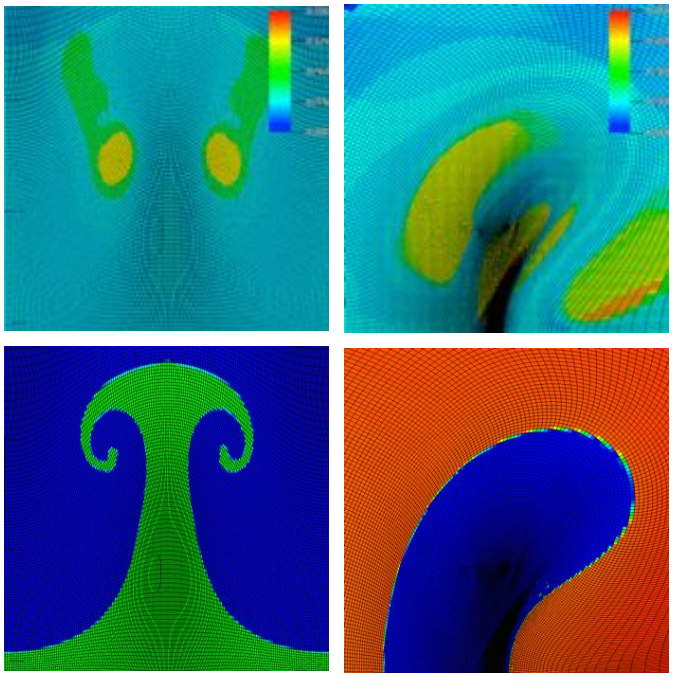


ML techniques reveal Fs-peptide folding events from long time-scale molecular dynamics simulations. A low dimensional embedding of the simulation events reveal transitions from fully unfolded states (blue) to fully folded states (red). A two dimensional embedding using t-test stochastic neighborhood embedding shows the presence of near native states (labeled state 1) versus partially unfolded (2-7) and fully unfolded states (8-9) in the picture.
Image Credit: Arvind Ramanathan, ORNL.

PRD5: **Inner-Loop** Scientific Machine Learning
Hybrid Machine Learning, Models, & Algorithms

Key Points: What is the role and potential advantages of ML-embedded approaches in computational model and algorithm development?

- Combination of scientific computing with learned adaptivity for more efficient simulations
- ML for in-situ parameter tuning
- ML for sub-grid physics models
- Progress will require the development of new methods to quantify tradeoffs and optimally manage the interplay between traditional and ML models and implementations

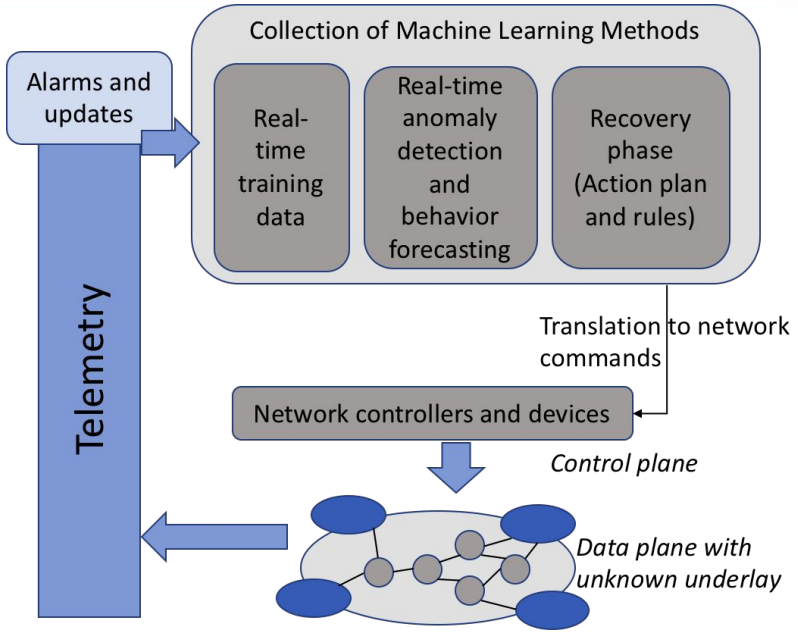


The arbitrary Lagrangian-Eulerian (ALE) method is used in a variety of engineering and scientific applications for enabling multi-physics simulations. Unfortunately, the ALE method can suffer from simulation failures, such as mesh tangling, that require users to adjust parameters throughout a simulation just to reach completion. A supervised ML framework for predicting conditions leading to ALE simulation failures was developed and integrated into a production ALE code for modeling high energy density physics.
Image credit: M. Jiang, LLNL.

PRD6: Outer-Loop Scientific Machine Learning
Automated Decision Support, Optimization, Resilience, & Control

Key Points: What are the challenges in managing the interplay between automation & human decision-making?

- Outer-Loop applications include optimization, uncertainty quantification, inverse problems, data assimilation, & control.
- New mathematically & scientifically justified methods to guide data acquisition and ensure data quality and adequacy.
- SciML methods for improving system resilience or responsiveness.



Exascale applications are exponentially raising demands from underlying DOE networks such as traffic management, operation scale and reliability constraints. Networks are the backbone to complex science workflows ensuring data is delivered securely and on-time for important compute to happen. In order to intelligently manage multiple network paths, various tasks such as pre-computation and prediction are needed to be done in near-real-time. ML provides a collection of algorithms that can add autonomy and assist in decision making to support key facility goals, without increased device costs and inefficiency. In particular, ML can be used to predict potential anomalies in current traffic patterns and raise alerts before network faults develop. Image credit: Prabhat, LBNL.

More Scientific Machine Learning Examples

Scientific Machine Learning has widespread Science & Energy uses

Three Capability PRDs (and combinations) seem to cover most examples

- Data-Intensive SciML
- Inner-Loop SciML
- Outer-Loop SciML

Compelling Big Science use cases include:

- Improved operational capabilities of scientific user facilities
- Better computational models from data-compute convergence
- Automation & adaptivity within scientific method (systems, processes)
- Many more ...

Core Research Agenda for Scientific Machine Learning

Scientific Machine Learning: Priority Research Directions (PRDs)	
Foundational Themes	**Capabilities Research**
PRD1. Domain-Aware SciML Leveraging scientific domain knowledge	**PRD4. Data-Intensive SciML** Automated scientific inference & data analysis
PRD2. Interpretable SciML Explainable & understandable results	**PRD5. Inner-Loop SciML** Machine learning-embedded models & algorithms for better scientific computing tools
PRD3. Robust SciML Stable, well-posed & efficient formulations	**PRD6. Outer-Loop SciML** Automated decision-support, optimization, resilience, & control for complex systems & processes

Machine Learning for Big Science

- **Lens of Scientific Computing & Applied Mathematics → SciML**
- **Capabilities Research: "Taxonomy" & PRDs for major use cases**
- **Foundational Themes: Basic research is essential**

History: DOE Applied Math Base Program & Research Initiatives are Key Foundations for Scientific Machine Learning

DOE Applied Math Base Program ⇒ Foundational Themes in SciML

Fundamental research in robust & stable formulations, Data-intensive analysis, Multi-physics & multi-scale models, Scalable linear algebra & solvers, Optimization under uncertainty, UQ, etc

DOE Applied Math Research Initiatives

Scientific Inference & Data Analysis ⇒ Data-Intensive SciML

- **2009** - Mathematics for Analysis of Petascale Data
- **2013** - DOE Data-Centric Science at Scale

Multiscale Models & Algorithms ⇒ Models, Algorithms, & **Inner-Loop SciML**

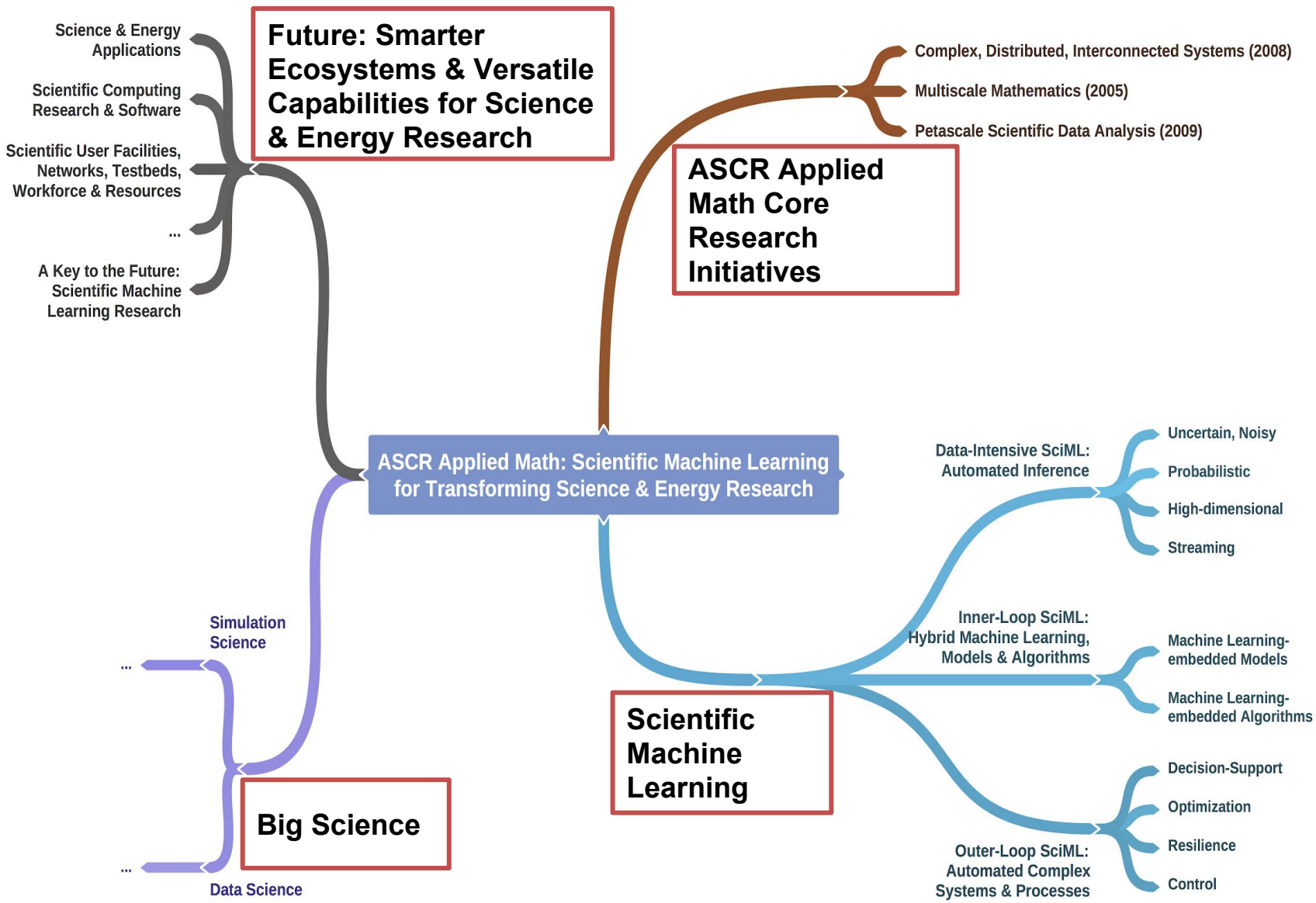
- **2005** - Multiscale Mathematics Research and Education
- **2008** - Multiscale Mathematics for Complex Systems (also MMICCs in **2012 & 2017**)

Integrated Capabilities for Complex Systems ⇒ Outer-Loop SciML

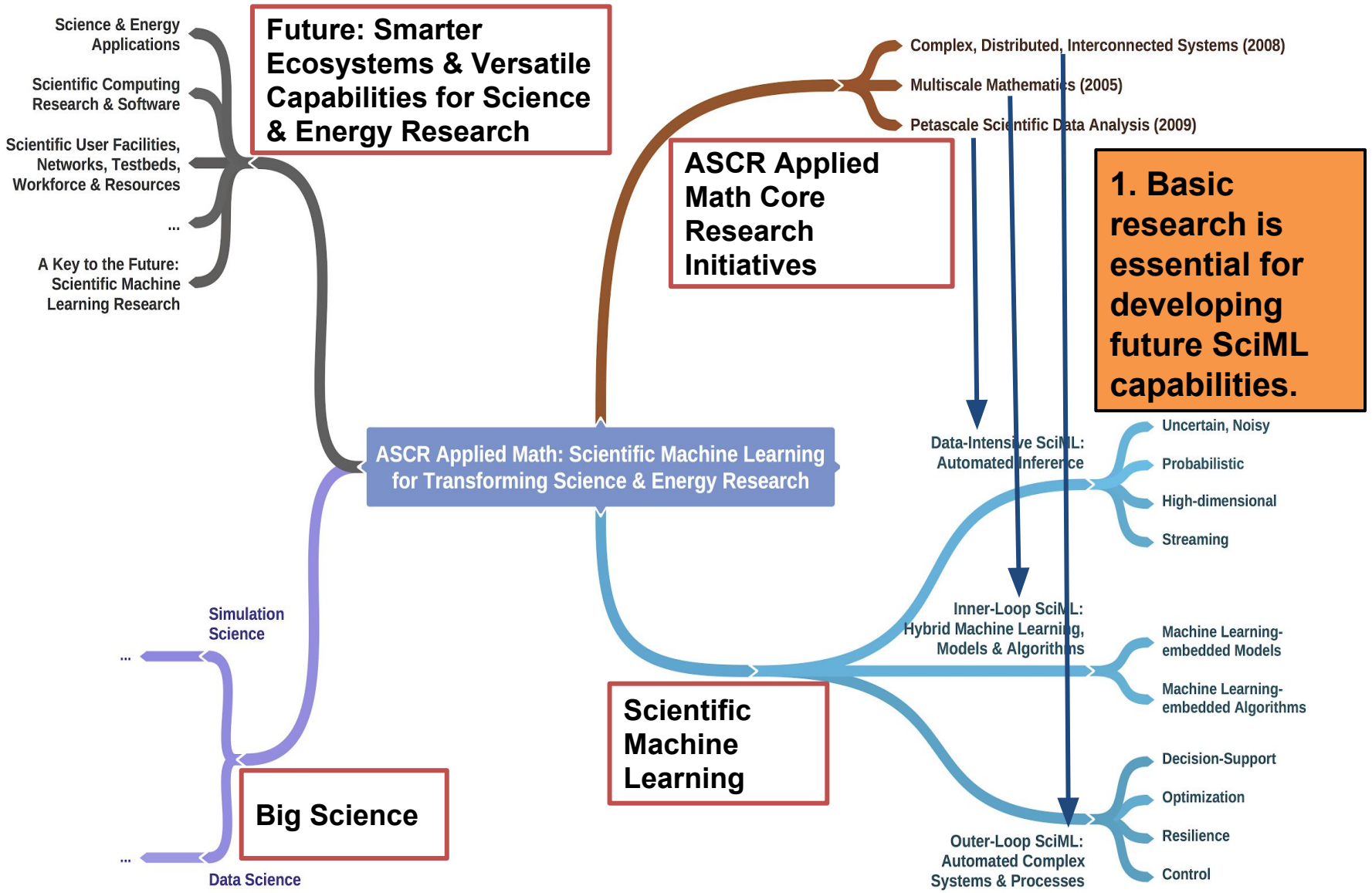
- **2009** - Mathematics for Complex, Distributed, Interconnected Systems
- **2010 & 2013** - Uncertainty Quantification for Complex Systems; UQ for Extreme-Scale Science
- **2012 & 2017** - Mathematical Multifaceted Integrated Capability Centers

Scientific Machine Learning will leverage basic research investments, widespread Science & Energy use cases, & DOE workforce expertise.

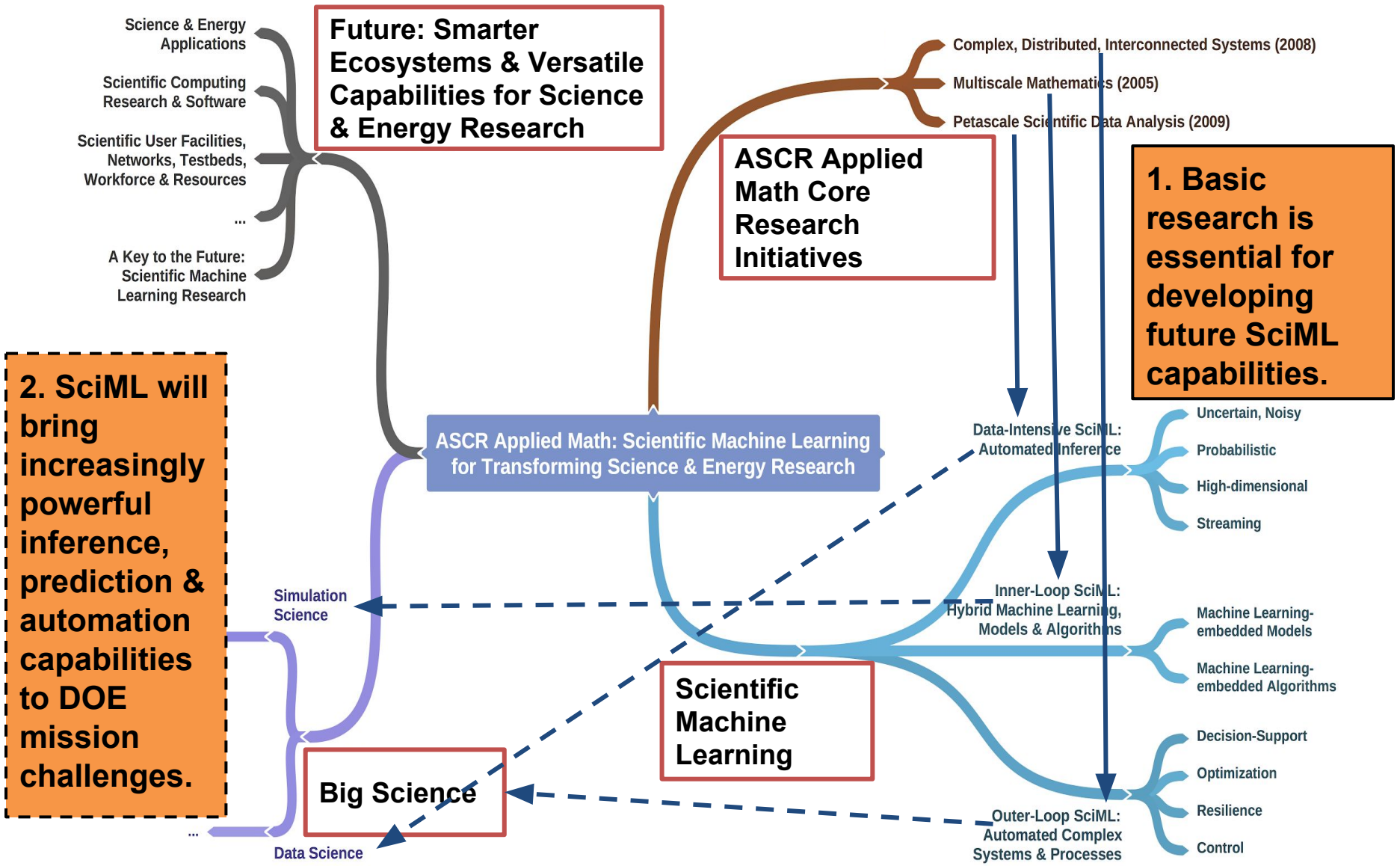
Workshop Charge Letter: Scientific Machine Learning (SciML) for transforming the Future of Science & Energy research



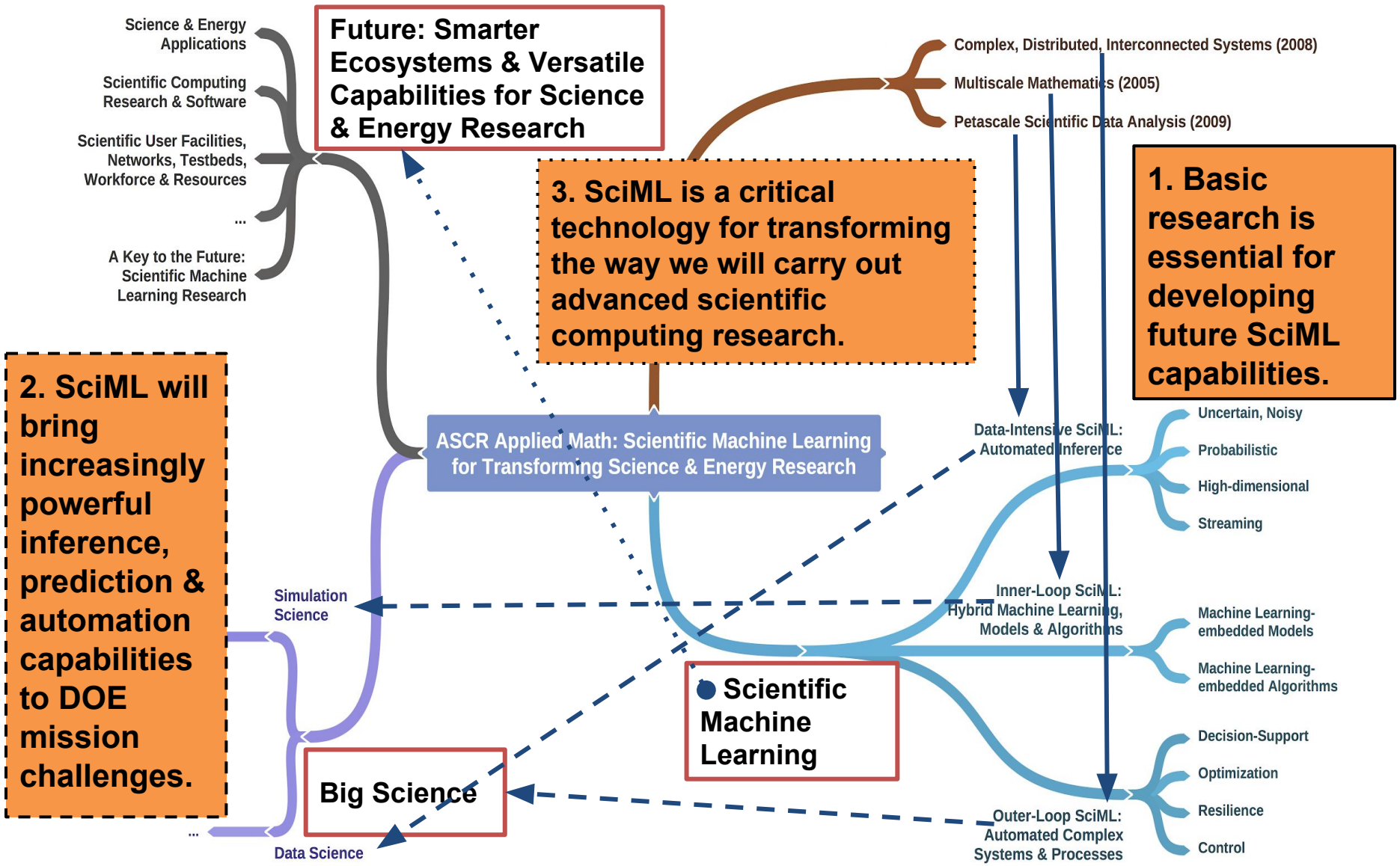
Workshop Charge Letter: Scientific Machine Learning (SciML) for transforming the Future of Science & Energy research #1



Workshop Charge Letter: Scientific Machine Learning (SciML) for transforming the Future of Science & Energy research #2



Workshop Charge Letter: Scientific Machine Learning (SciML) for transforming the Future of Science & Energy research #3



SUMMARY & OUTLOOK

Machine Learning is a powerful scientific enabling technology

- More than Data. Also for Modeling, Complex Systems, Science
- Basic research in scientific computing & mathematical foundations is essential
- Fast moving area → Need roadmap, blueprint, strategy
- Compelling: Re-visit ML, Re-think scientific computing uses

Pump is Primed for DOE leadership

- Roots from previous decade(s) of Applied Math basic research
- Ready: Researchers & expertise, Professional communities, etc

Future of Science & Energy Research

- Advanced technologies: More complex, more heterogeneous
- Greater Automation & Adaptivity for research breakthroughs
- Scientific Machine Learning Priority Research Directions are a basis for a cross-cutting research initiative toward this future