Computational Science Initiative

Community Advisory Council

January 11, 2018

BNL is a Data Driven Science Laboratory

BNL provides Data-rich Experimental Facilities:

- **RHIC** - Relativistic Heavy Ion Collider - supporting over 1000 scientists world wide
- **NSLS II** - Newest and Brightest Synchrotron in the world opened in the world, supporting a multitude of scientific research in academia, industry and national security
- **CFN** - Center for Functional Nanomaterials, combines theory and experiment to probe materials

BNL supports large scale Experimental Facilities:

- **LHC ATLAS -** Largest Tier 1 Center outside CERN
- **ARM** - Atmospheric Radiation Measurement Program - Partner in multi-side facility, operating its external data Center
- **QCD -** Facilities for BNL, RIKEN & US QCD communities

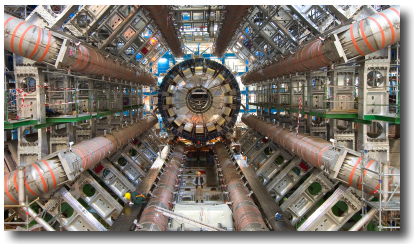
Science Today is Data Driven Discovery

RHIC

NSLS II

CFN

ATLAS



QCD

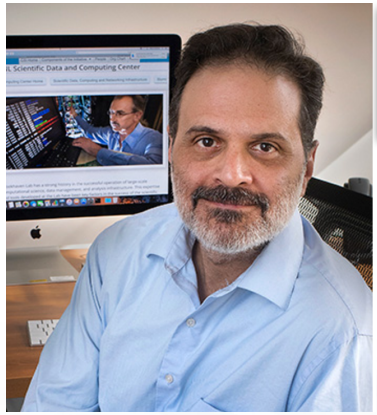
BNL Data Statistics

- **2017 Milestone: Over 100PB**, of catalogued data archived, long term, frequent reuse
- **2nd largest scientific archive in the US**, 4th largest in the world (ECMWF, UKMO, NOAA)
- **2016 - 400PB analyzed, 2017 -expected 500PB**
- **2016 37 PB exported**
- **BNL PanDA software enabled processing of ~1.6 EXBytes** of data worldwide in 2016

Computational Science Initiative

- Established in 2015
- An umbrella to bring together computing and data expertise across the lab
- Aims to foster cross-disciplinary collaborations in areas of computational sciences, applied math, computer science and data analytics.
- Aims to drive developments in programming models, advanced algorithms and novel computer architectures to advance scientific computing and data analysis.

Director:
Kerstin Kleese van Dam



Deputy Director:
Francis Alexander

Computational Science Initiative

Director:
Barbara Chapman

Computer Science & Mathematics

Director:
Eric Lancon

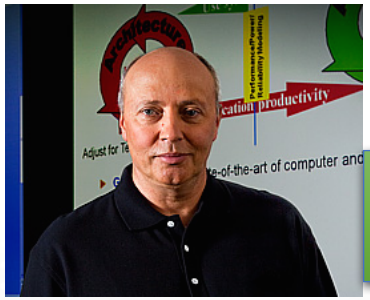
Scientific Data & Computing Center

Director:
Shantenu Jha

Center for Data-Driven Discovery

Director:
Nick D'Imperio

Computational Science Lab



Director:
Adolfy Hoisie

Computing for National Security

Key research initiatives: *Making Sense of Data at Exabyte-Scale and Beyond*

Real-Time Analysis of Ultra-High Throughput Data Streams

Integrated, extreme-scale machine learning and visual analytics methods for real time analysis and interpretation of streaming data.

New in situ and in operando experiments at at large scale facilities (e.g., NSLS-II, CFN and RHIC)

Data intensive science workflows possible in the Exascale Computing Project

Analysis on the Wire

Autonomous Optimal Experimental Design

Goal-driven capability that optimally leverages theory, modeling/simulation and experiments for the autonomous design and execution of experiments

Complex Modeling Infrastructure

Interactive Exploration of Multi-Petabyte Scientific Data Sets

Common in nuclear physics, high energy physics, computational biology and climate science

Integrated research into the required novel hardware, system software, programming models, analysis and visual analytics paradigms

CSI is planning to grow our National Security Computing Portfolio

Contribute to the development of a world class capability in computing the end state being ”BNL as a computing powerhouse”

Help solve challenges of societal importance in this arena: cybersecurity, intelligence applications, urban science, non-proliferation, technologies at the leading edge

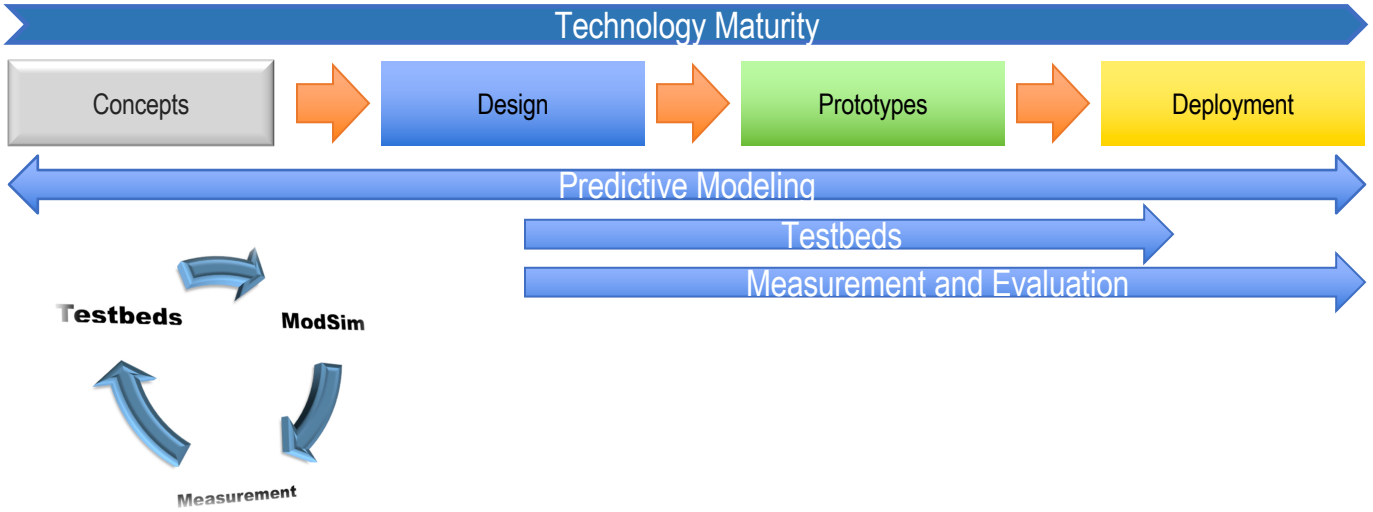
Build on existing capabilities and significant synergies with computing in science: data sciences, architectures, algorithms

Develop a world-class workforce of computer scientists and applied mathematicians, domain scientists

Contribute to solving computational challenges of multiple agencies, such as intelligence community, multiple DOD departments including DARPA and I-ARPA, DOE/NNSA, DHS, to just name a few

A holistic approach to solution along multiple dimensions

A Computing Space/Time/Methodology View

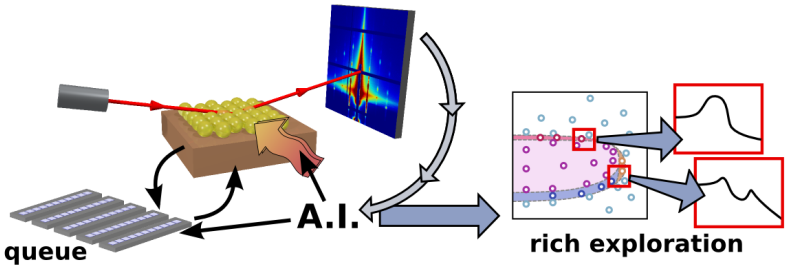


Sub-systems	Processing	Memory	Network	Storage and I/O
Emerging Paradigms	Approximate	Quantum	Neuromorphic	Superconductive
Workloads	Numeric	Machine Learning	Data Analytics	Graph Analytics

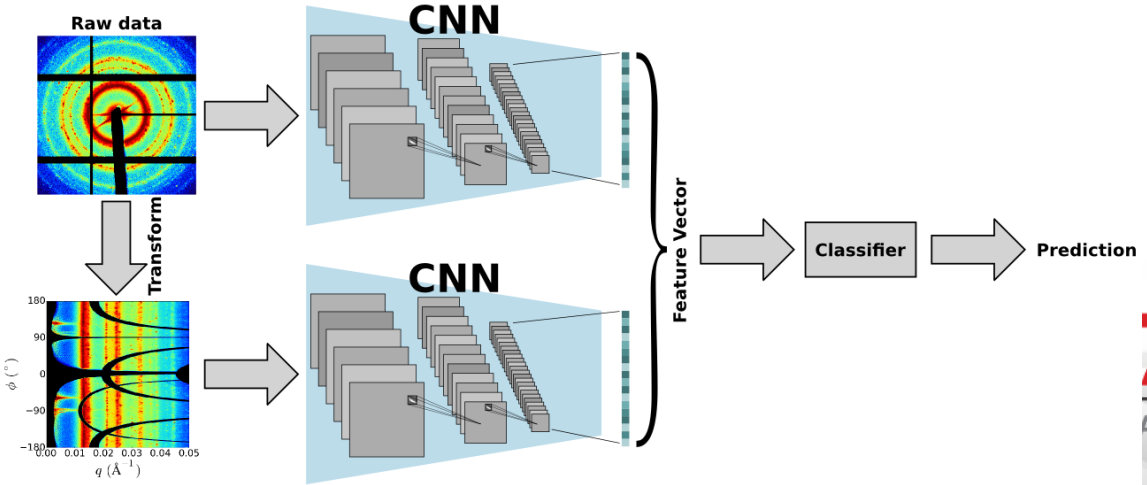
CSI is Enabling Scientific Discovery

Autonomous X-ray Scattering Beamline

- PIs: Kevin Yager (CFN) and Dantong Yu (CSI and NJIT)



- Data analysis streaming library developed that automatically and immediately analyzes each new image
- Pipeline triggers hand-crafted analysis modules and machine-learning classifiers
- Prototype deployed and running on CMS beamline (11-BM at NSLS-II), with close involvement of beamline staff (lead scientist I



Spatio-Temporal Learning for Solar Energy Forecasting

Scientific Achievement

Our newly developed spatio-temporal learning algorithm (LSTnet) significantly improves solar energy forecasting

Significance and Impact

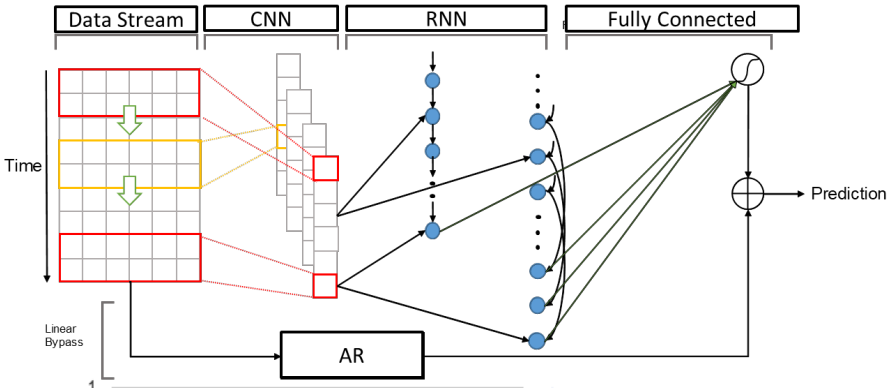
Utility scale spatio-temporal solar irradiance forecasting algorithms can improve operation and integration of solar farms with electric power grid and they can be used for any spatio-temporal data

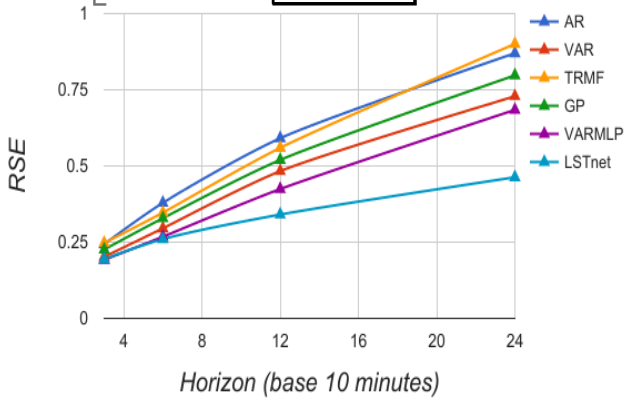
Research Details

Jointly modeling short and long-term time dependency by combining Deep Learning (DL) and Autoregressive (AR) model

CNN (Convolutional Neural Network) captures local dependency patterns whereas RNN (Recurrent Neural Network) captures long-term dependency patterns

Automatically adapting the mixtures weights of the AR and CNN/RNN components based on input





Top shows LISF (Long Island Solar Farm) and middle one shows our developed LSTNet algorithm. Left shows that LSTnet outperforms other models significantly (shown up to 4 hours)

Healing X-ray Scattering Images

Scientific Achievement

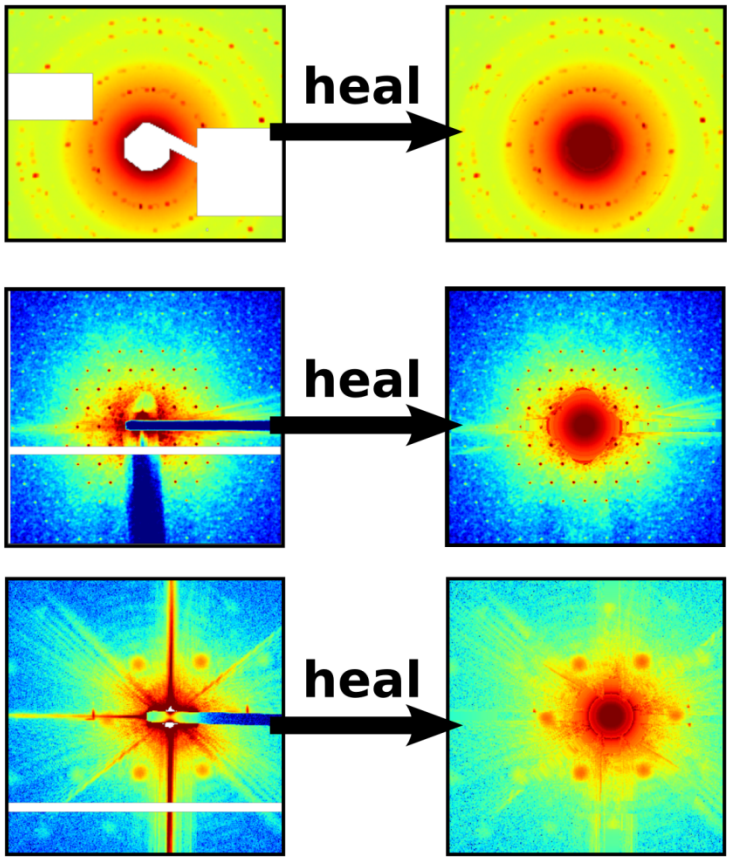
A "physics-aware" algorithm was developed to heal defects in x-ray scattering datasets

Significance and Impact

Healing x-ray data allows rapid automated analysis, and is a useful pre-processing step for machine-learning

Research Details

- Modern x-ray synchrotrons generate data at a massive rate; however, defects in the data (gaps, artifacts) complicate automated analysis
- Traditional 'inpainting' fails on scientific data
- By exploiting the known physics of x-ray scattering, a tuned healing algorithm was developed; for instance, symmetry is recognized and exploited
- Healed images can be more easily analyzed using existing data pipelines, including modern machine-learning methods
- The image healing algorithm also outputs physically-useful information (symmetry, type of ordering, etc.)



Healing X-ray Images: X-ray diffraction/scattering images typically have defects, including missing data and artifacts, which complicate subsequent analysis. However, traditional image healing algorithms do not interpolate scientific data in a physically-meaningful way. A "physics-aware" inpainting method was developed, allowing x-ray scattering images to be healed in a physically-rigorous way.

Jiliang Liu, et al., *IUCrJ* 4, 455 (2017).

Work is being performed at Brookhaven National Laboratory

Acceleration of Radar Simulator Code for Cloud Research

Scientific Achievement

Accelerated Cloud resolving model radar simulator code (CR-SIM) from 18 hours to 6 minute execution time.

Significance and Impact

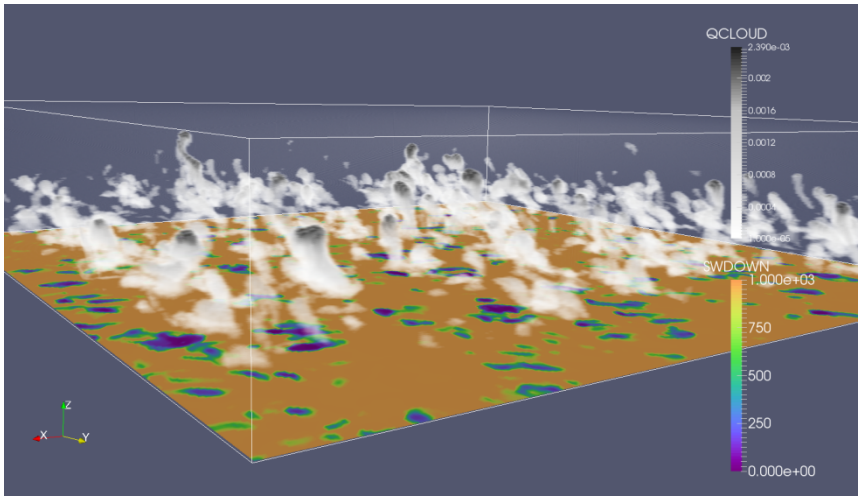
Optimization allows now for effective comparison of CRM and LES models to real observations.

Research Details

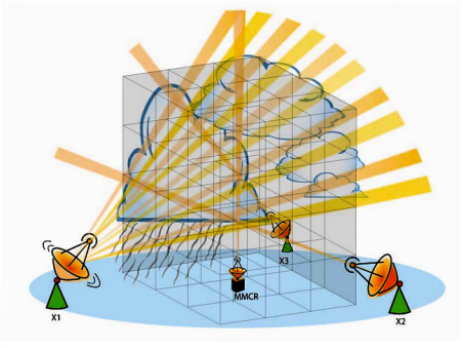
Creates Virtual Cloud Observations accounting for various sensor limitations

Using Data Intensive Programming Model Features for code Optimization

Users: BER ARM LASSO to test Large Eddy Simulations (LES), Climate Model Development and Validation (CMDV) teams. Now also international interest.



LES is commonly used to simulate clouds and the planetary boundary layer (lowest part of atmosphere). Shown here are the cloud water content (QCLOUD) and the resulting shadows that impact the sunlight reaching the ground. Credit: ARM Climate Research Facility.



Work is being performed at Brookhaven National Laboratory

BNL Hosted GPU Hackathon June, 2017

Event Description

As a member of the OpenACC standards committee and the NVIDIA GPU Research Center, BNL is committed to bringing more GPU expertise to the lab-wide community and playing a more visible role in GPU computing.

This 5-day GPU Hackathon brought together application teams both inside and outside of BNL who need expert assistance with porting their codes to GPUs or improving the performance of their existing GPU codes.

Each team was assigned 1-2 mentors who gave hands-on guidance throughout the Hackathon.

Jointly organized by

- Brookhaven National Laboratory
- University of Delaware
- Oak Ridge National Laboratory
- Stony Brook University

Image Source: https://www.olcf.ornl.gov/training-event/2017-gpu-hackathons/

NVIDIA GPU RESEARCH CENTER

Image Source: www.nvidia.com

OpenACC Directives for Accelerators

Image Source: www.openacc.org

A Unified Mathematical Framework for Integrating Data and Models: Bayesian Inversion & Optimal Experimental Design for Complex Large-Scale Systems

The scientific method: systematic acquisition of knowledge about our world via the continuous interplay of theory and experiment

HPC: radically transformed our ability to model complex multiscale systems, analyze complex multimodal data

A principled, rigorous, and scalable mathematical framework to optimally guide the interplay between complex models and complex data and to account for uncertainty in the process is thoroughly lacking.

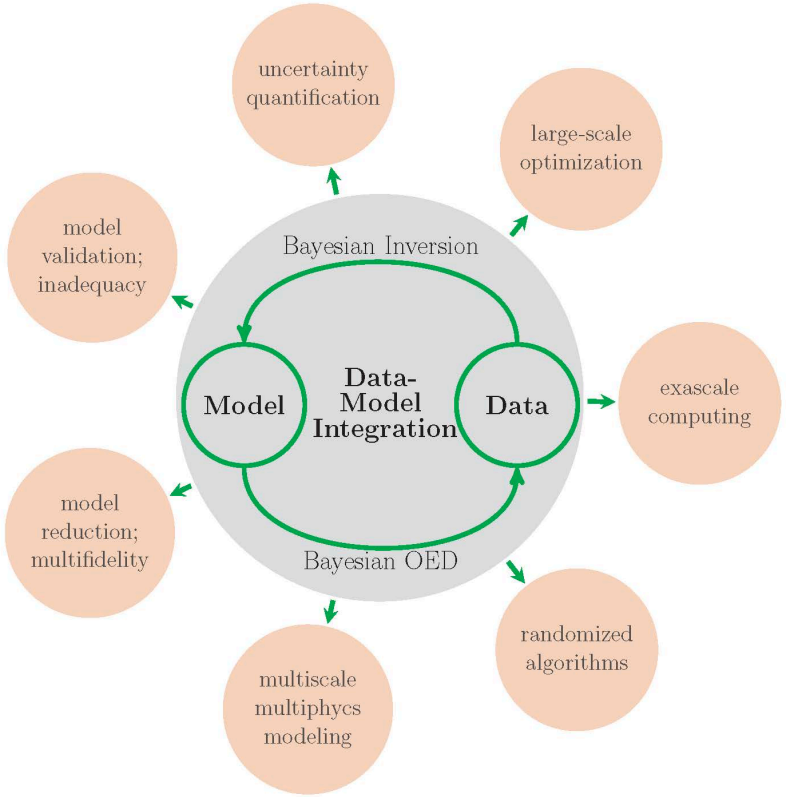
The methodologies by which experiments inform theory, and theory guides experiments, remain ad hoc, particularly when the physical systems under study are multiscale and complex.

Problems that span wider ranges of scales, represent richer interacting “physics”, and inform decisions of greater societal consequence (fusion, climate, precision medicine, advanced manufacturing, environmental contamination, combustion, …) are in critical need of a formal and rigorous framework for integrating data and models

Bayesian inverse theory & Bayesian optimal experimental design provide a rational and systematic framework for principled integration of data and models under uncertainty in both

- **Data → Models: Bayesian inverse problem**
  - The problem of how models are best informed by data is fundamentally an inverse problem.
  - Recent advances in Bayesian inverse theory & algorithms have made solution of inverse problems under uncertainty tractable for certain classes of problems.
  - Bayesian solution of the IP provides a probability distribution reflecting the probability of any particular model, given uncertainty in the data & the model. This provides a basis for OED.

- **Models → Data: Bayesian optimal experimental design problem**
  - How, where, when, and from which source to acquire experimental, observational, or simulation data to best inform models is fundamentally an OED problem.
  - Bayesian OED optimizes the data acquisition system to reduce uncertainties in the Bayesian inverse solution.
  - OED subsumes Bayesian inversion, leading to a stochastic optimization problem constrained by an inner Bayesian inverse problem.



New Projects in Precision Medicine

DOE/VA Pilot Projects benefit both VA and DOE

HPC, modeling/simulation and large scale data analysis to VA data to improve healthcare for our Veterans
Develop scalable machine learning algorithms for challenging classification and data imputation problems in DOE

***Enhanced prediction and diagnosis of Cardiovascular Disease* (CVD)**

Develop methods to inform individualized drug therapies to prevent, pre-empt and treat CVD.

Enhance prediction, diagnosis and management of major CVD subtypes in Veterans

Precision discrimination of lethal from non-lethal Prostate Cancer

Build improved classifiers to distinguish lethal from non-lethal prostate cancers.

Reduce unnecessary treatments /provide an increased quality of life for patients

Patient-specific analysis for Suicide Prevention

Provide tailored and dynamic suicide risk scores for each Veteran at risk.

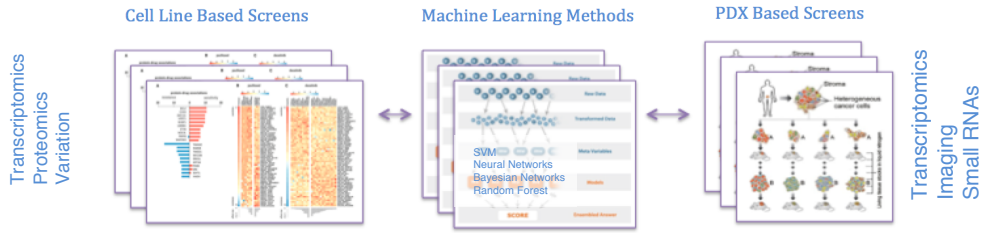
Create a clinical decision support system that assists VA clinicians in suicide prevention efforts, and helps to evaluate effectiveness of various prevention strategies.

- ***Scalable Algorithms for***
  - ***Binary and Multiclass classification***
  - ***Data Imputation (for missing data)***
  - ***Imbalanced data problems with constrained resources***
- ***Integrating large multimodal data sets (>20M patients)***
  - ***Images***
  - ***Mechanistic biological models***
  - ***Full Genomes***
  - ***Longitudinal Data***
- ***UQ and error analysis (skill assessment)***
- ***Potential Applications to BER programs in***
  - ***Genomics***
  - ***Climate state assessment***
  - ***Prediction of Complex Systems Behavior***

DOE/NCI Predictive Models for Preclinical Screening

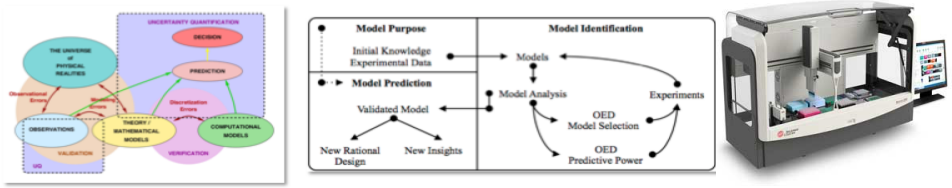
- Terabytes — Cell Line Compound Screens
- Terabytes — Compound Databases
- Terabytes — Cell Line Molecular Assays
- Terabytes — PDX Compound Screens
- Petabytes — PDX Features (DNA, RNA, Images)
- Petabytes — Protein-Protein Interactions
- Terabytes — Cancer Pathway Databases
- Petabytes — Cancer Genome Atlas
- Petabytes — Gene Expression Omnibus
- Petabytes — Cancer Data Commons

Machine Learning Based Predictive Models



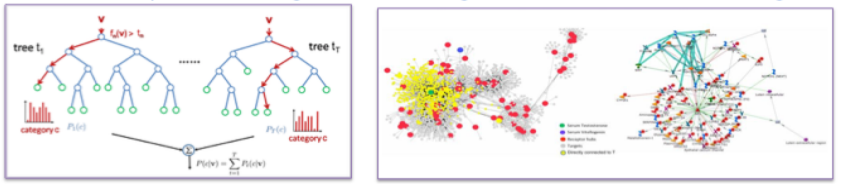
Feature Engineering, Cross Validation, Scalable Compute on CORAL

Uncertainty and Optimal Experiment Design



UQ Analysis, Model Selection, Model Improvement, Proposed Experiments

Hypotheses Formation and Mixed Modeling



Integration of Mechanistic, Statistical and Inferential Modeling

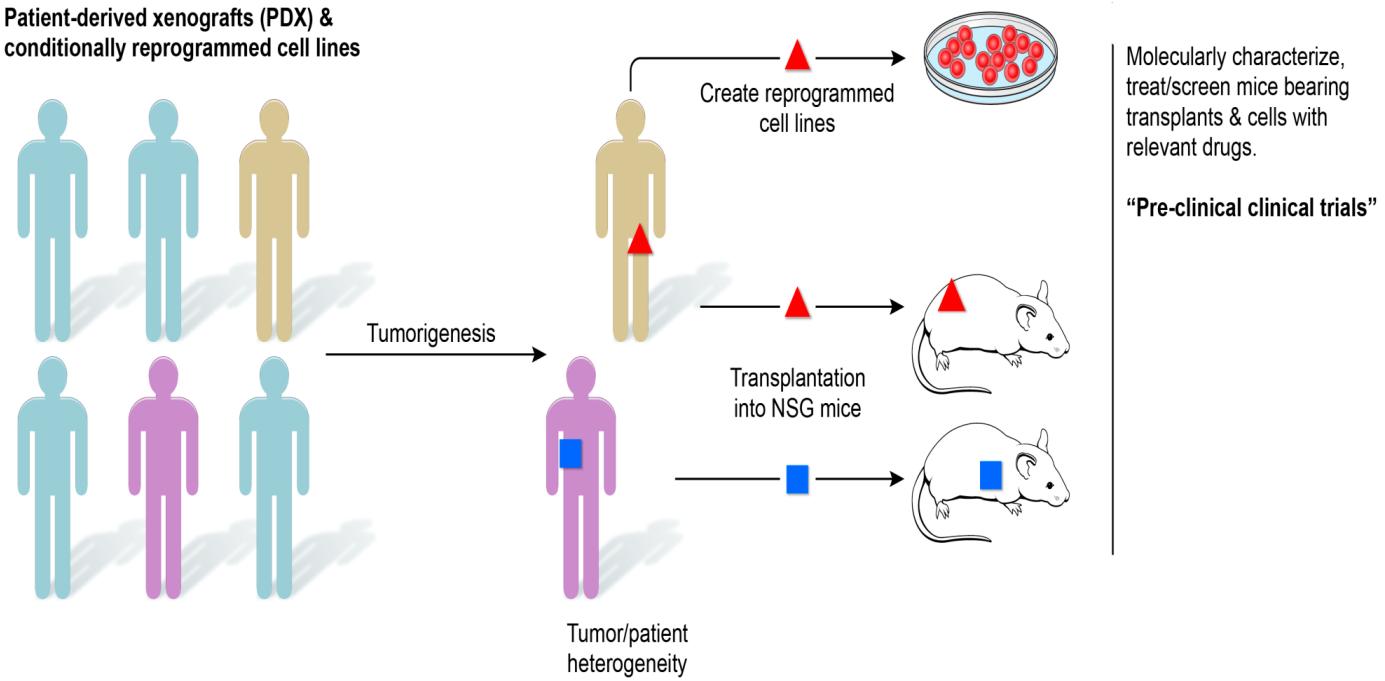
Aims for Preclinical Screening Pilot

Reliable machine learning based predictive models of drug response that enable the projection of screening results from and between cell-lines and PDX models

Uncertainty quantification and optimal experimental design to assert quantitative limits on predictions and to recommend experiments that will improve predictions

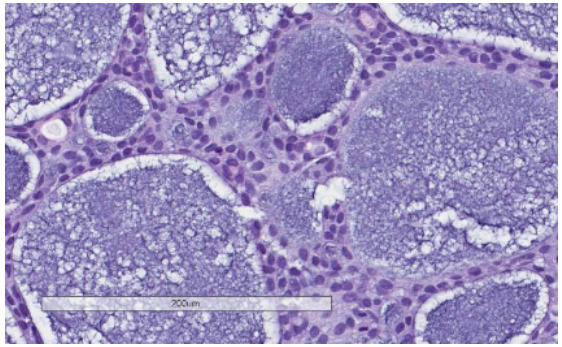
Improved modeling paradigms that support the graded introduction of mechanistic models into the machine learning framework and to rigorously assess the potential modeling improvements obtained thereof

Patient Derived Xenograft Models



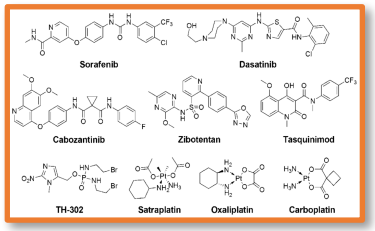
Nature Rev. Clin. Oncol. 11: 649-662, 2014.

Problem: Modeling Drug Response

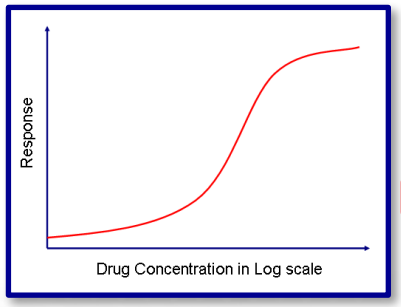


Drug (s)

- descriptors
- fingerprints
- structures
- SMILES
- dose



$$\mathcal{R} = f(\mathcal{T}, \mathcal{D})$$

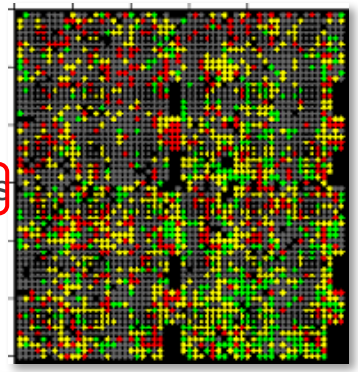


IC50
GI50
% growth
Z-score

Response

gene expression levels
SNPs
protein abundance
microRNA
methylation

Tumor



Hybrid Models in Cancer

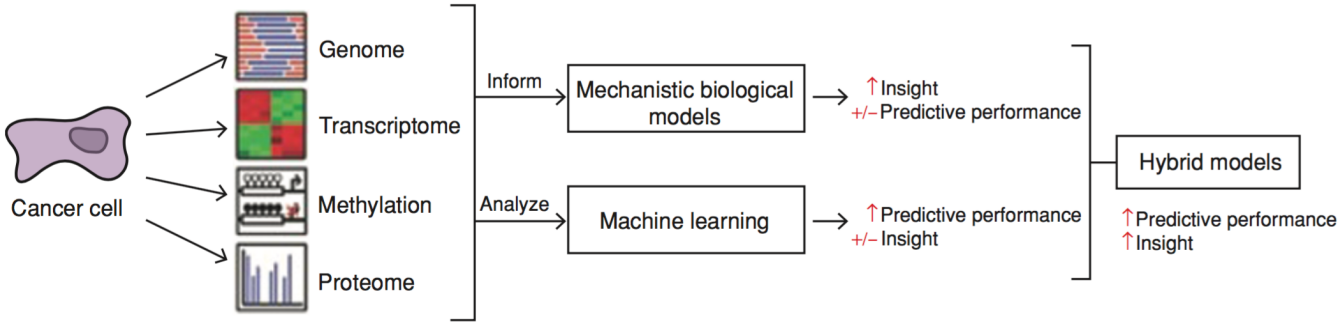


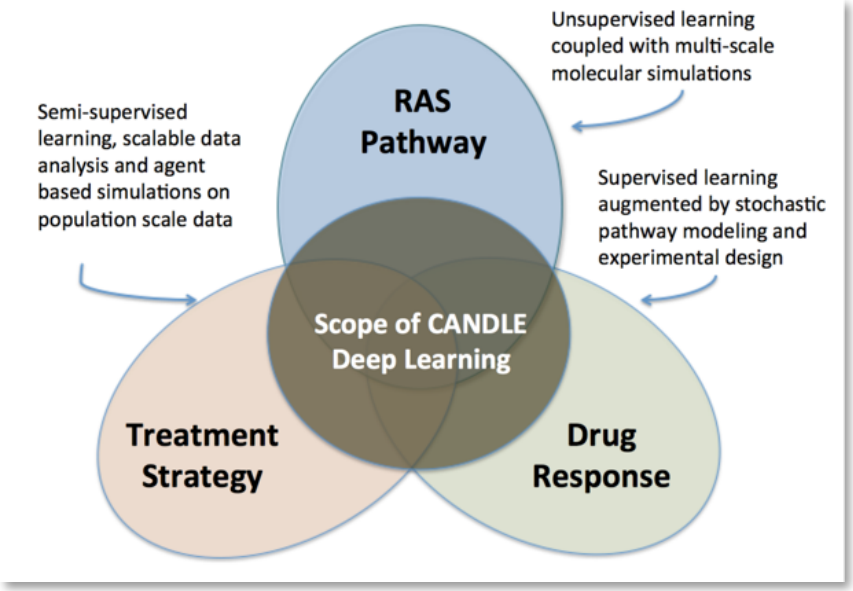
Figure 1. In two DREAM challenges, high throughput data characterizing cancer cells are used to build predictive models. Mechanistic models provide insight into the underlying biology, but do not take full advantage of the information within the data to achieve high performance. Machine learning methods are associative and extract maximum predictive value from the data, but do not always provide insight about mechanism. The future may bring hybrid models that combine the best of both approaches.

Predicting Cancer Drug Response: Advancing the DREAM

Russ B. Altman

Summary: The DREAM challenge is a community effort to assess current capabilities in systems biology. Two recent challenges focus on cancer cell drug sensitivity and drug synergism, and highlight strengths and weaknesses of current approaches. *Cancer Discov; 5(3); 237–8. ©2015 AACR.*

ECP-CANDLE : CANcer Distributed Learning Environment



CANDLE Goals

Develop an exscale deep learning environment for cancer

Building on open source Deep learning frameworks

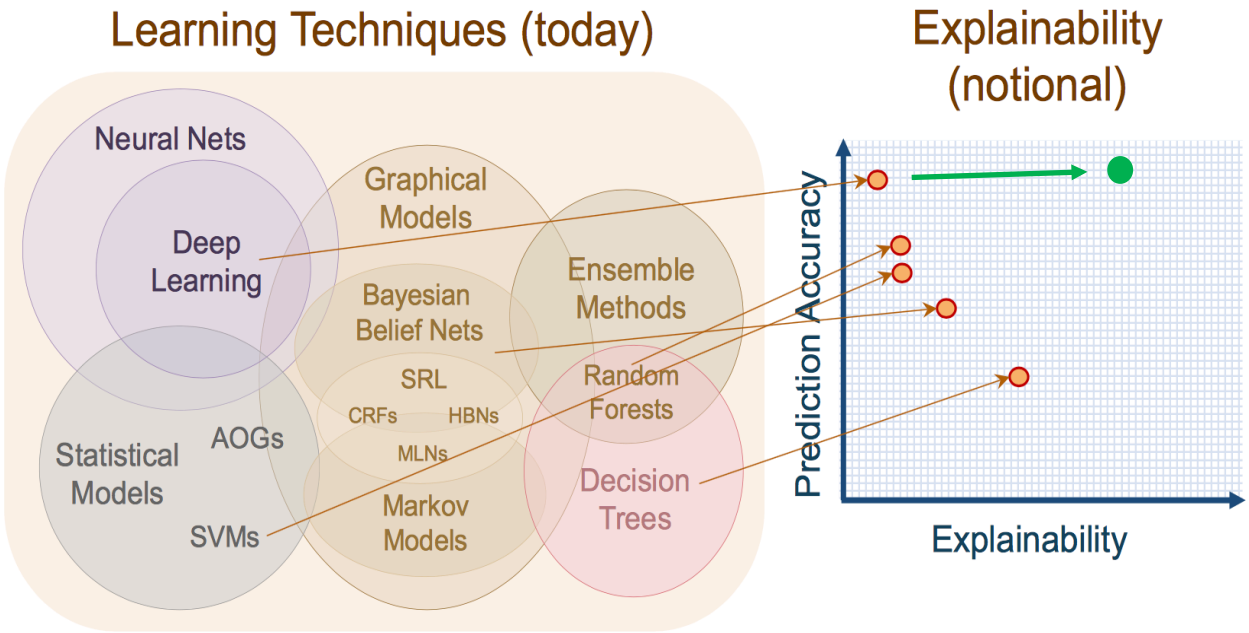
Optimization for CORAL and exascale platforms

Support all three pilot project needs for deep dearning

Collaborate with DOE computing centers, HPC vendors and ECP co-design and software technology projects

ECP EXASCALE COMPUTING PROJECT

Machine Understanding




BNL
Brookhaven
National Laboratory
OPERATED BY BROOKHAVEN SCIENCE ASSOCIATES UNDER CONTRACT WITH THE UNITED STATES DEPARTMENT OF ENERGY
Thank You!