# Joint Design of Advanced Computing for Cancer (JDACS4C)

## A Brief Overview and Perspective

Martin Berzins

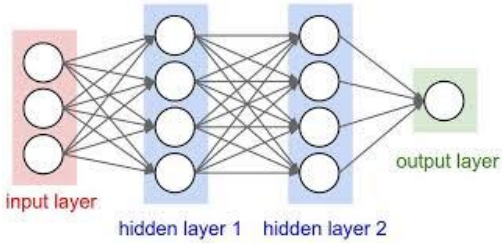
University of Utah

DOE-NC Collaborations Working Group Member

Advanced Scientific Computing Advisory Committee Member

# JDACS4C projects

Introduction




Pilot 1  Cellular Scale

Pilot  3 Population Scale

Cross Cutting Themes and UQ

Conclusions and Observations

**Acknowledgments : C. Lauzon,  A. Gryshuk,  F.  Streitz , Pilot Project Teams.**

# Deep Learning Successes and Challenges

Deep learning of aftershock patterns following large earthquakes

Such inputs distort how machine-learning-based systems are able to function in the world as it is.

BY IAN GOODFELLOW, PATRICK MCDANIEL, AND NICOLAS PAPERNOT

**Making Machine Learning Robust Against Adversarial Inputs**

Artificial intelligence tool 'as good as experts' at detecting eye problems

**Machine-learning system can identify more than 50 different eye diseases and could speed up diagnosis and treatment**

A.I. Comes Into Fashion

Thanks to artificial intelligence, machines can now encroach on high-skilled workers as well.

By NOAM SCHEIBER

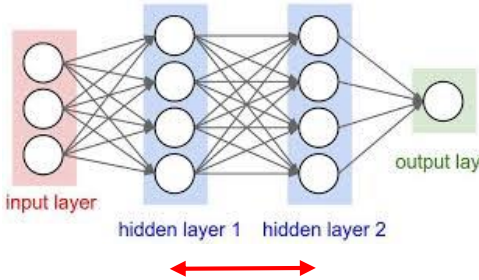
**One pixel attack for fooling deep neural networks**

CAT

DOG(78.2%)

IBM HAS A WATSON DILEMMA

Big Blue promised its AI platform would be a big step forward in treating cancer. But after pouring billions into the project, the diagnosis is gloomy.




**Not depth but depth**

# Integrated Precision Oncology

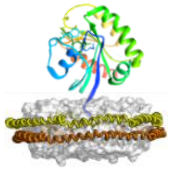
**Pilot 1**
**Pre-clinical Model Development**

**Aim 1: Predictive Models of Drug Response (signatures)**

**Aim 2: Uncertainty Quantification and Improved Experimental Design**

**Aim 3: Develop Hybrid Predictive Models**

**Pilot 2**
**RAS Therapeutic Targets**



**Aim 1: Adaptive time and length scaling in dynamic multi-scale simulations**

**Aim 2: Validated model for Extended RAS/RAS-complex interactions**

**Aim 3: Development of machine learning for dynamic model validation**

**Pilot 3**
**Precision Oncology Surveillance**

**Aim 1: Information Capture Using NLP and Deep Learning Algorithms**

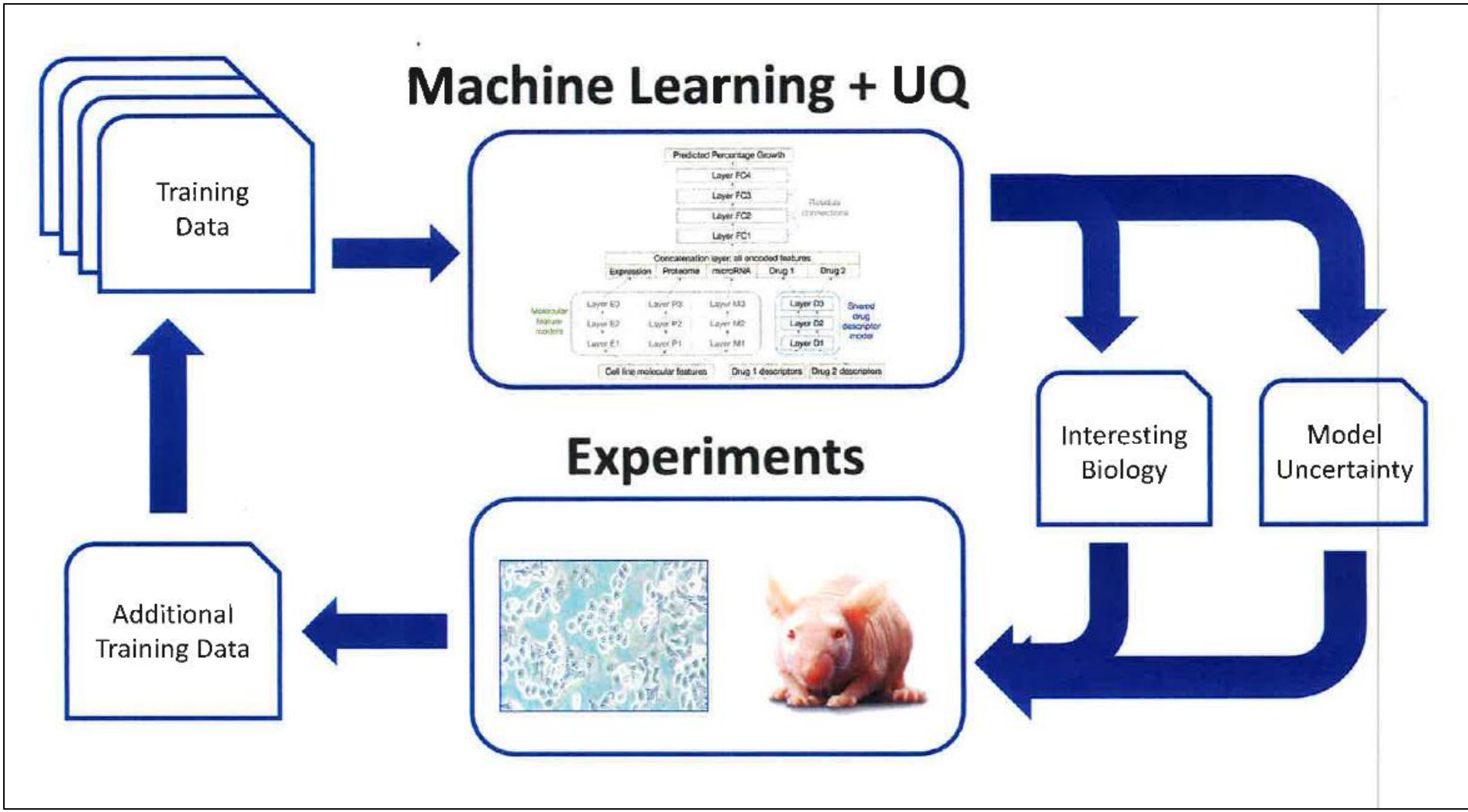
**Aim 2: Information Integration and Analysis for extreme scale heterogeneous data**

**Aim 3: Modeling for patient health trajectories**

***Crosscut: CANDLE exascale technologies, uncertainty quantification***

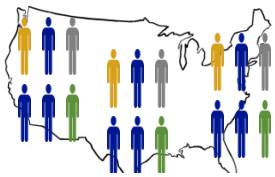
# Cellular Scale (Pilot 1)

**Scientific Co-Leads**
NCI: Yvonne Evrard – FNLCR
DOE: Rick Stevens – ANL



# Progress Examples - Cellular Scale (Pilot 1)

- Established database with integrated & multiple NCI data sources
- Established framework for machine learning applied to cancer
- Developed an integrated data collection of tumor characterization and drug response
- Developed a suite of ML models for single drugs and drug combinations
- Advanced use of ML and DL for QC on cancer data
- Developed an UQ formulation for DL
- Developed an integrated visualization method that combines the display of model predictions for drug synergy and response with those of model uncertainty



# Population Scale (Pilot 3)

**Scientific Co-Leads**
NCI: Lynne Penberthy– NCI
DOE: Gina Tourassi–ORNL

**Deep NLP for information capture**

Advanced machine learning for scalable patient Information capture from unstructured clinical reports to semi-automate the SEER program

**Novel data analytic techniques for patient information integration**

Scalable graph and visual analytics to understand the association between patient trajectories and patient outcomes

**Data-driven integrated modeling and simulation for precision oncology**

Precision modeling of patient trajectories

In silico clinical trials

# Why DOE-NIH partnership is essential for cancer surveillance

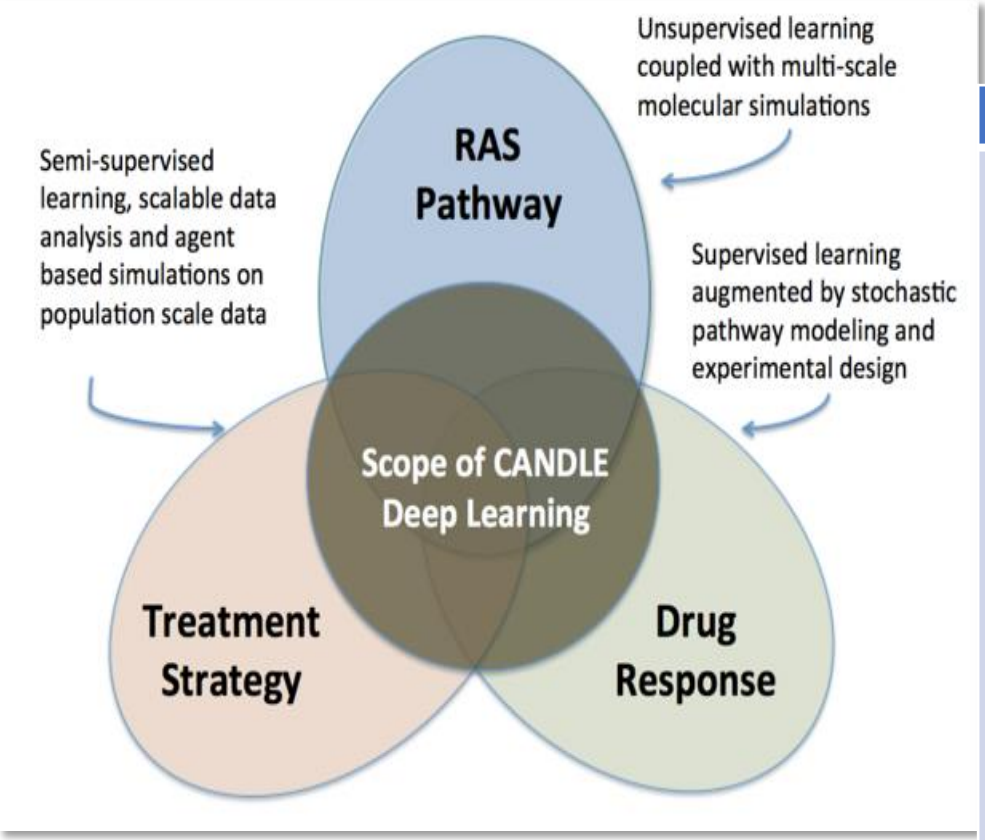
- Commercial entities have not been able to scale, are using manual extraction, are expensive or unwilling to adopt open source.
- The broad variety and varied structure of pathology reports is a huge challenge
- Academics have solutions at limited scale and applicability (populations, hospitals or cancers)
- While NLP and deep learning from pathology reports is not new, the scale is-the goal is to be able to analyze reports from 360 labs 10K+ pathologists and many data sources
- The reduction in effort for the cancer surveillance community from this open source effort is considerable

# Examples of Technology Development

- Develop deep learning natural language processing methods for unstructured text; pathology reports.
- Multi-Tasking-Convolution Neural Net MT-CNN clinical performance exceeded that of traditional ML and single-task CNNs as well as literature benchmarks
- Evaluate performance for scaling DL networks. **Model has been scaled to train efficiently ~80M documents and 100s of NLP tasks.**

- Advance and investigate UQ applications
- Refined pipeline to create and review large annotated datasets for NLP development and automated data abstraction. MT-CNN can be integrated with imaging workflows
  - Model has been extended for image-based phenotyping using the NIH Chestx-ray8 data

# JDACS4C – Cross-cutting

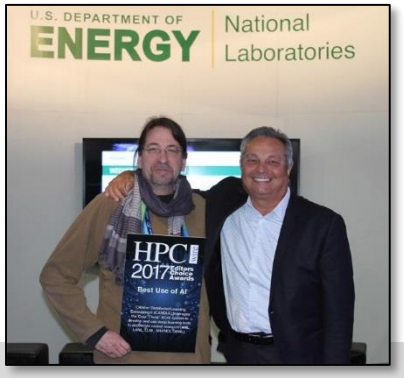



**Select Accomplishments**

- Delivered CANDLE benchmarks and being used for A21 Acceptance
- CANDLE is running at ANL, ORNL, LBL and NIH facilities
- Demonstrate prototype DNN for each pilot project use case
- Optimizing millions of parameters per model – applications beyond deep learning
- Integrating Uncertainty Quantification into each pilot

**CANDLE Lead**
Rick Stevens
Argonne National Lab

**Uncertainty Quantification Lead**
Tanmoy Bhattacharya
Los Alamos National Lab




**CANDLE in 2017 HPCwire Readers' and Editors' Choice Awards**

# Pilot Contributions CORAL and Exascale

- Pilot1 has contributed two benchmarks for the A21 acceptance tests
- Pilot3 has contributed an A21 benchmark
- Pilot1/CANDLE influenced systems requirements for machine learning applications on A21 and CORAL2
- Pilot1 selected as one of the ALCF Data and Machine Learning Early Science Projects
- Pilot 3 ran large Multi Task –Convolutional Neural Nets on Summit training almost 27K MT-CNNs Simultaneously

# Why Uncertainty Quantification?

## Benefits to DOE

- DOE is using Deep Learning broadly
- Many problems have a high penalty for overconfident incorrect predictions - "high-regret" applications

- DL modeling uncertainties are difficult to accurately quantify e.g (out-of-sample data) and adversarial situations especially problematic
- Developing new methods for rigorous UQ for DL will benefit other DOE problems

## Benefits to NCI

- DL is a powerful "black box," but
  - Provides "answers" without explanation
  - Finds unexpected correlations in data
  - Interpolates and extrapolates data trends across many variables
- Need reliable way of deciding how to interpret DL output in context of precision medicine
  - Extremely confident, repeatable answers may be incorrect
  - Interpolation/extrapolation difference unclear
  - Not clear which answers trustworthy Without UQ, good only for hypothesis generation
- Precision medicine requires building trustworthy predictive tools that know when they fail

# Conclusions and Observations

Current research in CS is making deep learning more robust and perhaps integrating it with model-based approaches

Pearl (2011 ACM Turing award) *discusses model-based vs observation-based (deep learning). Babylonians vs Greeks

" Learning machines need the guidance of a model of reality"

**The challenge then is to**

(i) Use DL to see patterns that are otherwise "invisible"

(ii) Integrate deep learning models with other models and

(iii) apply uncertainty quantification and ideas on data robustness to deep learning,

(iv) use robust DL in the context of multi-disciplinary approaches to solve the most challenging "deep regret" applications.

This initiative is addressing these issues in the context of one of the most challenging illnesses that (unfortunately) many of us will face.

*https://arxiv/abs/1801.04016