FNLAC NCI-DOE Collaborations *ad hoc* Working Group Report

Dr. Piermaria J. Oddone and Dr. Melissa C. Smith
Co-chairs

Meeting of the Frederick National Laboratory Advisory Committee
May 8, 2018

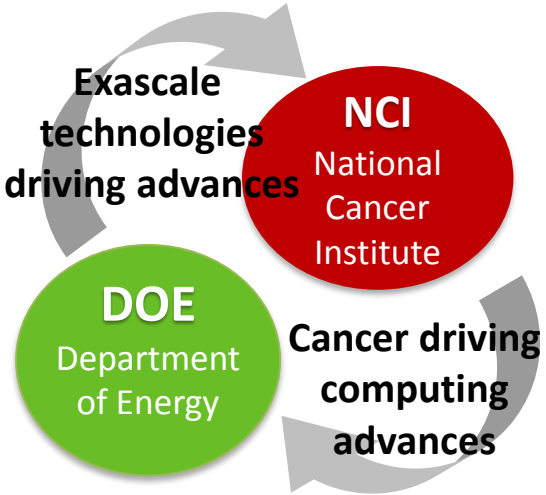
Mission of the FNLAC NCI-DOE Collaborations NCI Working Group

- Provide scientific evaluation of programs, projects and activities formed in support of or relevant to NCI-DOE collaborations
  - Explore the domains and activities in which collaborations between the NCI and DOE would be mutually beneficial and advance the missions of these entities
  - Optimize the functionality and output of the partnership and maximize impact on (i) the broader research community, (ii) the benefits of High Performance Computing (HPC) to systems biology and data science, and (iii) the acceleration of predictive modeling for cancer
- Activities will include
  - Technical evaluation of the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) pilot efforts and other collaboration projects (e.g. ATOM Consortium)
  - Guidance and insights on relevant partnerships with other entities
  - Extending the benefits and impacts of the partnership to the broader scientific community
- The Working Group will advise the FNLAC
  - In accordance with the NCI/DOE MOU, the DOE Secretary, DOE and DOE FACA committees may use the public products and public findings in furthering the DOE mission

NCI-DOE Collaboration

The NCI/DOE Collaborations were formed to jointly accelerate NCI and DOE federal missions in precision oncology and high-performance computing (HPC).

The partnership is designed to push the frontiers of high performance computing through application to NCI's mission to improve understanding of cancer biology and its application to more effective cancer therapies.



NIH NATIONAL CANCER INSTITUTE

Argonne NATIONAL LABORATORY

OAK RIDGE

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY EST.1943

Frederick National Lab for Cancer Research

Activities under the DOE-NCI Collaboration

Joint Design of Advanced of Computing Solutions for Cancer (JDACS4C)

- **Cellular Level Pilot 1:** Predictive Models for Pre-clinical Screening
- **Molecular Level Pilot 2:** RAS Biology in Membranes
- **Population Level Pilot 3:** Population Information Integration, Analysis, and Modeling
- **Uncertainty Quantification** Crosscut
- **CANDLE** (CANcer Distributed Learning Environment)**:** An Exascale Computing Project to develop Machine Learning framework for Cancer

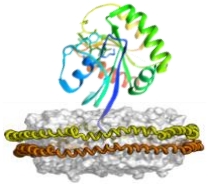
Accelerating Therapeutics for Opportunities in Medicine (ATOM)

Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

Pilot 1
Predictive Models for Pre-Clinical Screening

Aim 1: Develop reliable machine-learning-based predictive models of anti-cancer drug response
Aim 2: Integrate uncertainty quantification and optimal experimental design to assert quantitative limits on predictions
Aim 3: Develop hybrid predictive models that support the graded introduction of mechanistic models into the machine-learning framework

Pilot 2
RAS Biology in Membranes



Aim 1: Develop multiscale modeling capabilities to investigate RAS dynamics on cell membranes
Aim 2: Understand how RAS and extended RAS complexes are activated and simulate RAS-RAF interactions on realistic, lipid-bilayer membranes
Aim 3: Develop machine learning-enabled dynamic model validation approach to high-fidelity simulation

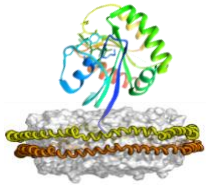
Pilot 3
Precision Oncology Surveillance

Aim 1: Information capture of unstructured clinical text using Natural Language Processing (NLP) and Deep Learning algorithms
Aim 2: Information integration and analysis to understand drivers in patterns of cancer outcomes and predict clinical endpoints
Aim 3: Data-driven modeling of patient-specific and population level health trajectories

Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

Pilot 1
Predictive Models for Pre-Clinical Screening

Pilot 2
RAS Biology in Membranes

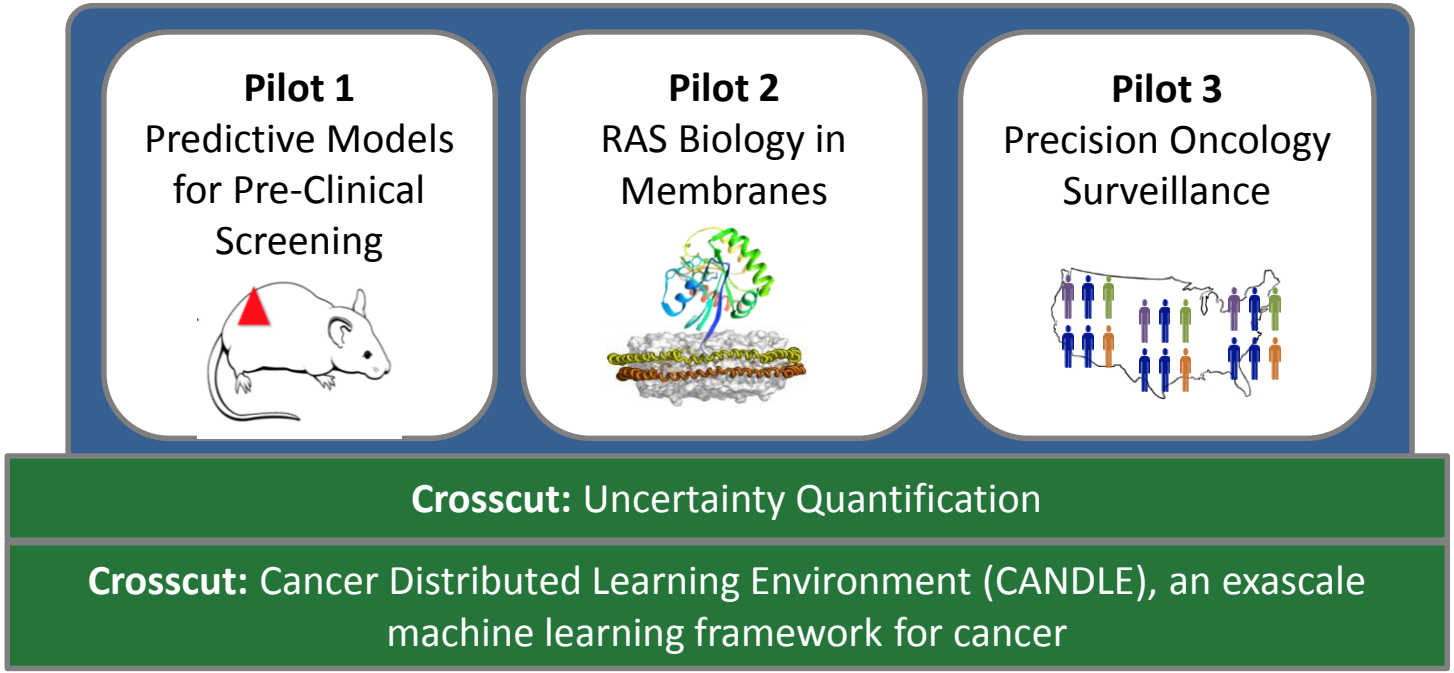


Pilot 3
Precision Oncology Surveillance

Crosscut: Uncertainty Quantification

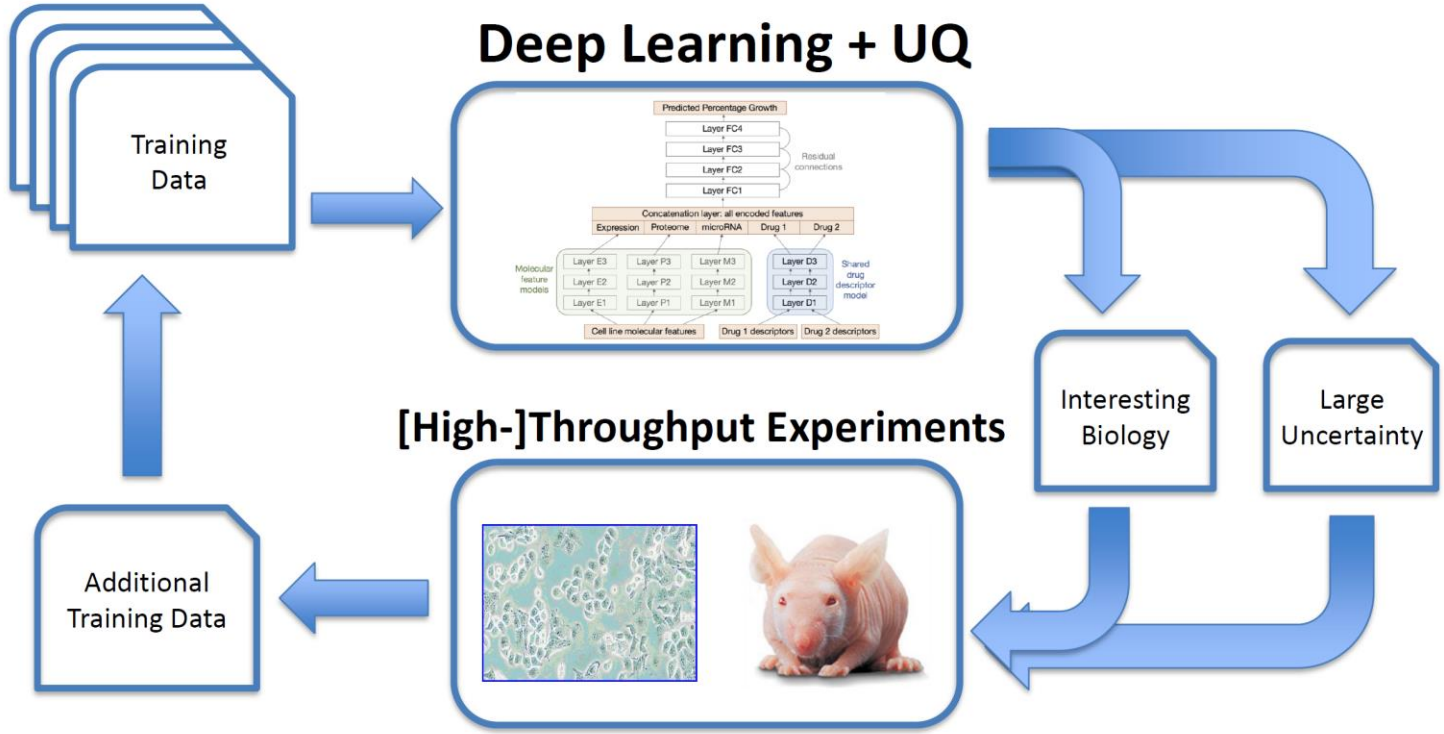
Crosscut: Cancer Distributed Learning Environment (CANDLE), an exascale machine learning framework for cancer

Other Activities Under NCI-DOE Collaboration



- Accelerating Therapeutics for Opportunities in Medicine (ATOM)

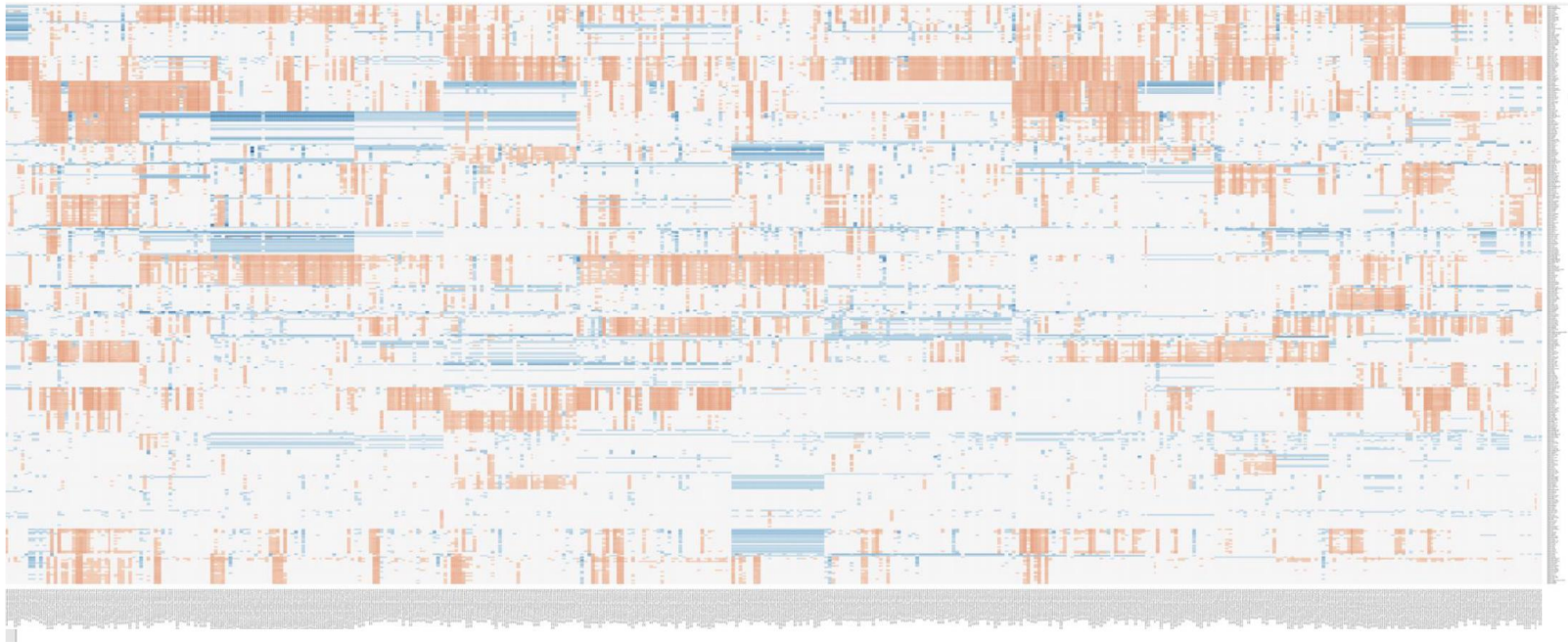
Pilot 1: Predictive Models for Pre-clinical Screening



Pilot 1: Predictive Models for Pre-clinical Screening

Combined Synergy and Uncertainty Map

Red == Predicted Synergy with High or Medium Confidence , White == Not Interesting, Blue == Low Model Confidence



Pilot 1: Predictive Models for Pre-clinical Screening

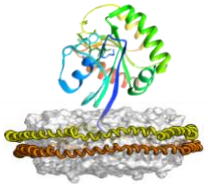
Summary:

- Very ambitious project, well defined milestones, with already remarkable results
- Landscape of drug pair interactions on cancer cell lines
- Including error quantification
- Already biologically interesting observations:
  - The per-drug fraction of synergistic interactions appears to be predictable from drug features
  - The prediction performance in breast cancer is significantly better than for other tissue types

Pilot 1: Predictive Models for Pre-clinical Screening

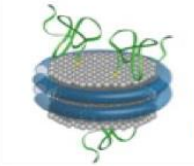
Proposed Next Steps:

- Assess the bottleneck of dataflow to quicken the experimental feedback loop.
- Develop potential mechanisms to engage the extramural community to generate collaborations.
- Develop potential mechanisms to engage extramural community to gather data that will allow a focus on human samples and PDX models as opposed to cell lines.
- For future studies, consider primary human tissue-derived data sets that might be available or developing from patients entering early and late phase NCTN-sponsored clinical trials, which would be highly annotated and where response/resistance to specific agents was known.
- The working group will consider working with FNLCR to identify an expert to provide additional scientific oversight.
- Down the line, consider convening an *ad hoc* Working Group specifically for pilot 1 and pilot 3 with expertise in Deep Learning applications.
- NCI should consider additional scientific staff at FNLCR to better interface with and support the scientific projects within NIH that could take advantage of exascale computing.

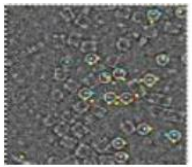


Pilot 2: RAS Biology in Membranes

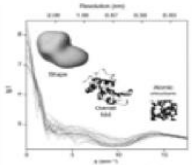
Ras activation experiments



Experiments on nanodisc



CryoEM imaging

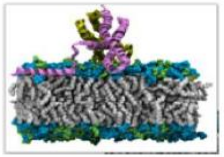


X-ray/neutron scattering

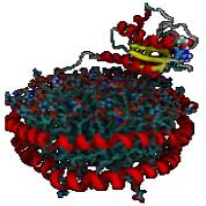
Multi-modal experimental data, image reconstruction, analytics

Protein structure databases

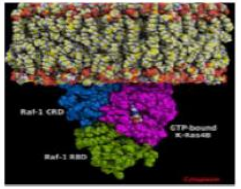
Predictive simulation and analysis of Ras



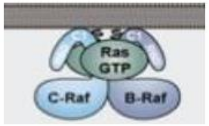
Phase Field model of lipid membrane



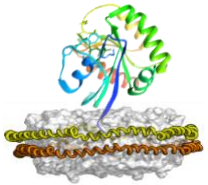
Granular Ras membrane interaction simulations



Atomic resolution Ras-RAF interaction

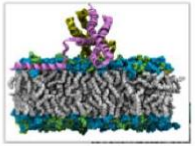


Ras Activation

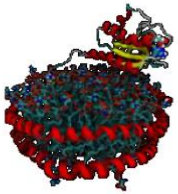


Pilot 2: RAS Biology in Membranes

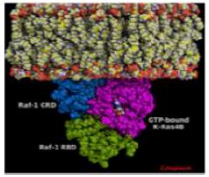
Predictive simulation and analysis of Ras



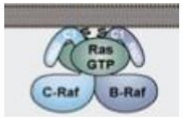
Phase Field model of lipid membrane



Granular Ras membrane interaction simulations



Atomic resolution Ras-RAF interaction

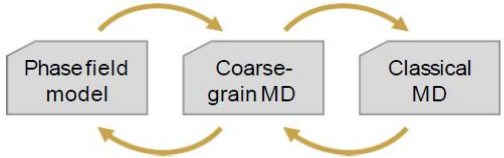


Ras Activation

Adaptive sampling molecular dynamics simulation codes

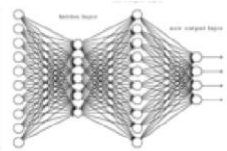
Adaptive time stepping

Adaptive spatial resolution



High-fidelity subgrid modeling

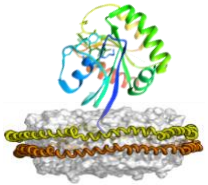
Machine learning guided dynamic validation



Unsupervised deep feature learning

Mechanistic network models

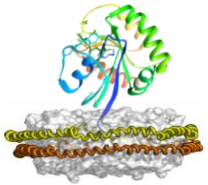
Uncertainty quantification



Pilot 2: RAS Biology in Membranes

Summary:

- Strong team in place, imbedded in the larger RAS initiative
- Many tools developed to connect dynamics from micro-to-macro scales
- Simulations validated with multiple experiments with testable hypothesis derived from initial simulations
- Change in dynamics of KRAS and RAF on the lipid membrane with mutated hypervariable regions observed
- By incorporating RAF kinase, membrane association of isolated RAF-cysteine-rich domains was observed in the simulations



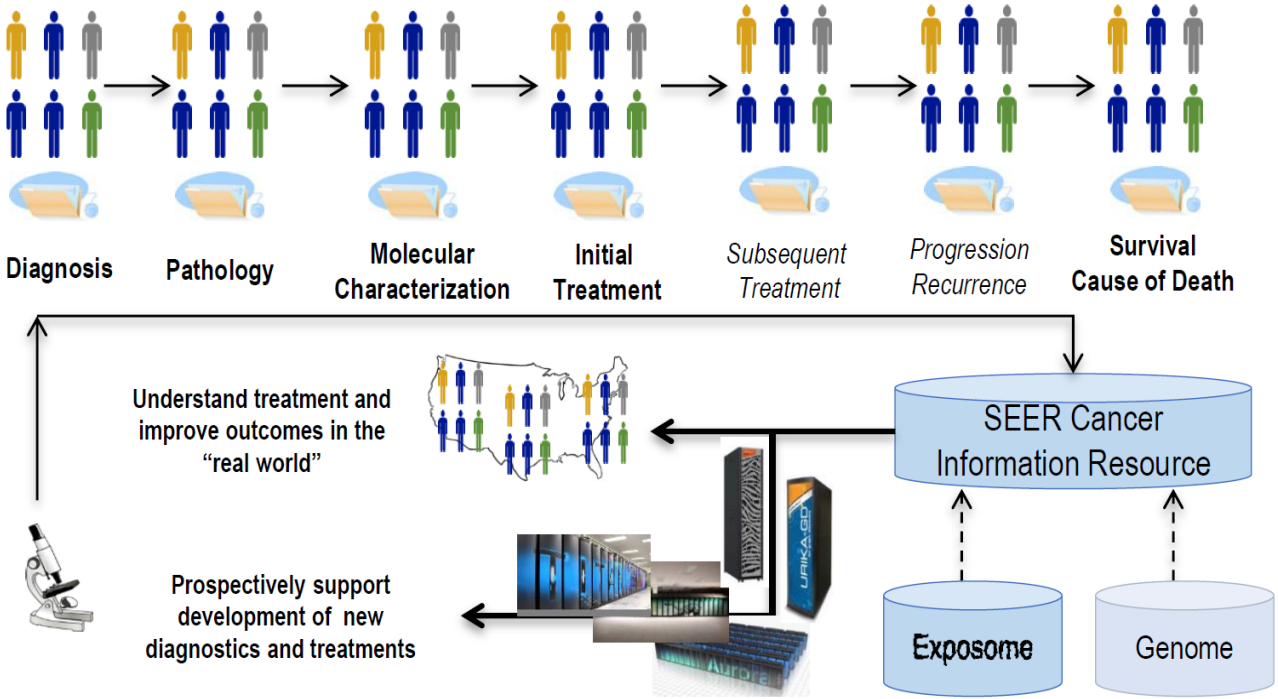
Pilot 2: RAS Biology in Membranes

Proposed Next Steps:

- Determine the feasibility of engaging the private sector for data acquisition, collaboration, and project acceleration.
- Identify potential mechanisms to access additional data sets.
- Down the line, consider convening an *ad hoc* Working Group specifically for pilot 2 with expertise in molecular dynamics simulations.
- Determine if there is a need for NCI (and DOE) to provide more management oversight to ensure that the best capabilities of each agency are brought to the collaboration.

Pilot 3: Population Information Integration, Analysis, and Modeling

Improve the effectiveness of cancer treatment in the "real world" through computing



Pilot 3: Population Information Integration, Analysis, and Modeling

Pilot 3 goals

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

Pilot 3: Population Information Integration, Analysis, and Modeling

Summary:

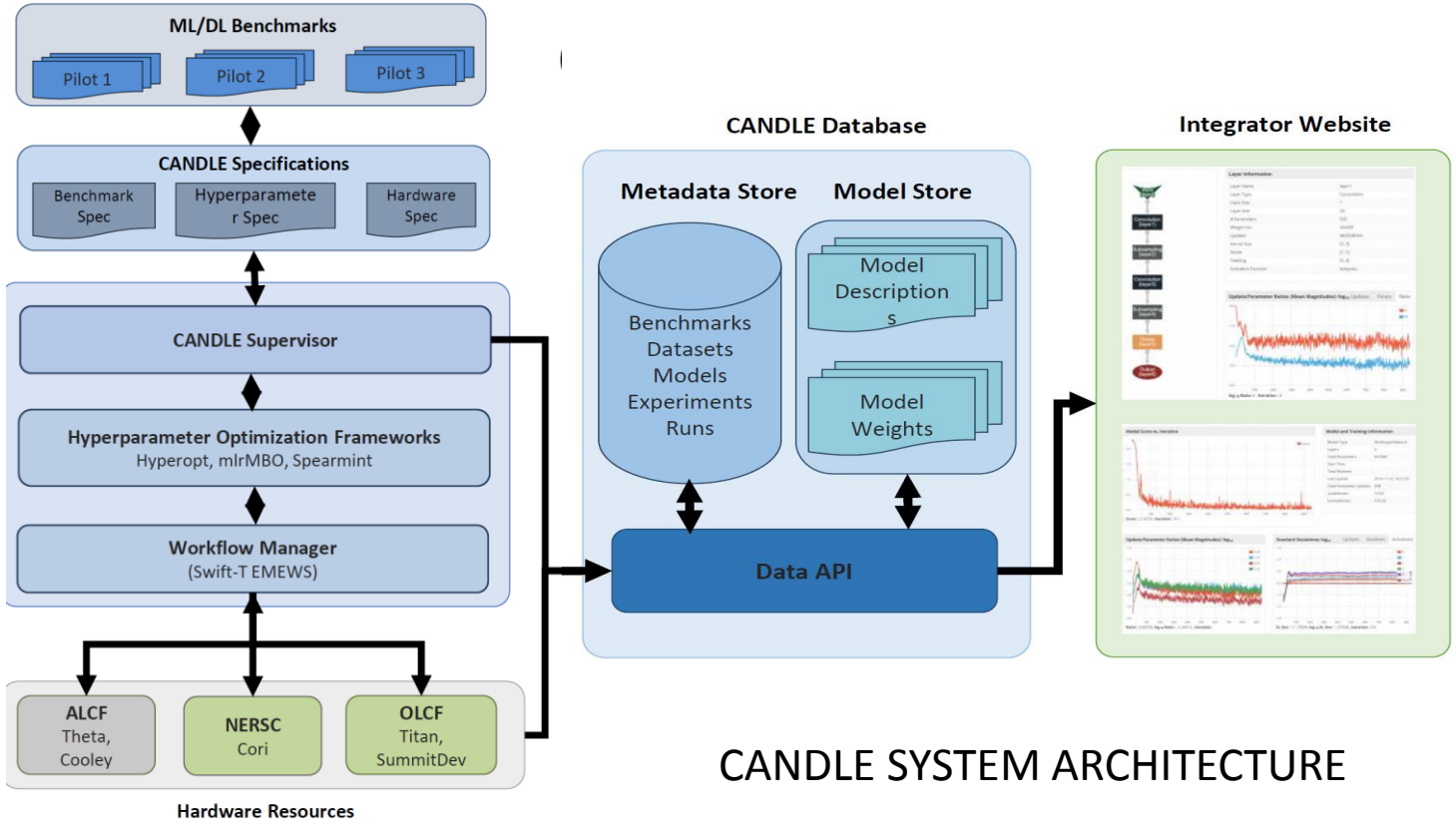
- Very good progress and interactions with the community
- Relatively slow start on accessing registries
- Developed, deployed, and refined annotation pipeline
- Partnering with Information Management Services (IMS), tested Natural Language Processing (NLP) tool delivery via Application Programming Interfaces (IPS)
- Developed, deployed and tested NLP tools for automated identification of primary site, laterality, histology, grade, and behavior
- Developed breast cancer schema for biomarkers and recurrence data elements
- Developed and benchmarked four supervised DL architectures
- Packaged deep learning model into software product to return to cancer community working with CANDLE
- Developed and benchmarked 2 UQ strategies

Pilot 3: Population Information Integration, Analysis, and Modeling

Proposed Next Steps:

- Determine feasibility of incorporating different types of patient reports, including whole slide images and radiology reports.
- Determine how SEER data could be more relevant to precision medicine and consider stated project goals. While retrospective analyses of SEER data sets that captured more comprehensive information would be highly desirable, and may well generate novel NLP approaches and methods, it is not clear how the current project is relevant to real time clinical decision making.
- Down the line, consider convening an *ad hoc* Working Group for pilot 1 and pilot 3 with expertise in Deep Learning applications.
- Determine if there is a need for NCI (and DOE) to provide more management oversight to ensure that the best capabilities of each agency are brought to the collaboration.
- Develop potential mechanisms to engage extramural community to ensure that the best capabilities of each agency are brought to provide more data for deep learning applied to NLP and other laudable (but yet addressed) goals of this important pilot.

Crosscut: CANcer Distributed Learning Environment (CANDLE)



CANDLE SYSTEM ARCHITECTURE

Crosscut: CANcer Distributed Learning Environment (CANDLE)

Highlights:

- The functional goals of CANDLE include
  - (1) using deep learning to help others increase productivity;
  - (2) supporting established deep-learning frameworks (e.g., Google) to run on DOE supercomputers; and
  - (3) managing CANDLE training data.
- CANDLE will enlarge the community that uses exascale computing to answer scientific questions by making products and training readily available
- CANDLE contributions
  - Created a prototype deep neural network (DNN) for information extraction from clinical reports for Pilot 3
  - Provided the first version of Combo in CANDLE, delivered to Pilot 1
  - Created a prototype DNN that performs unsupervised feature learning for Pilot 2
- Clearly defined future milestones

Crosscut: Uncertainty Quantification (UQ)

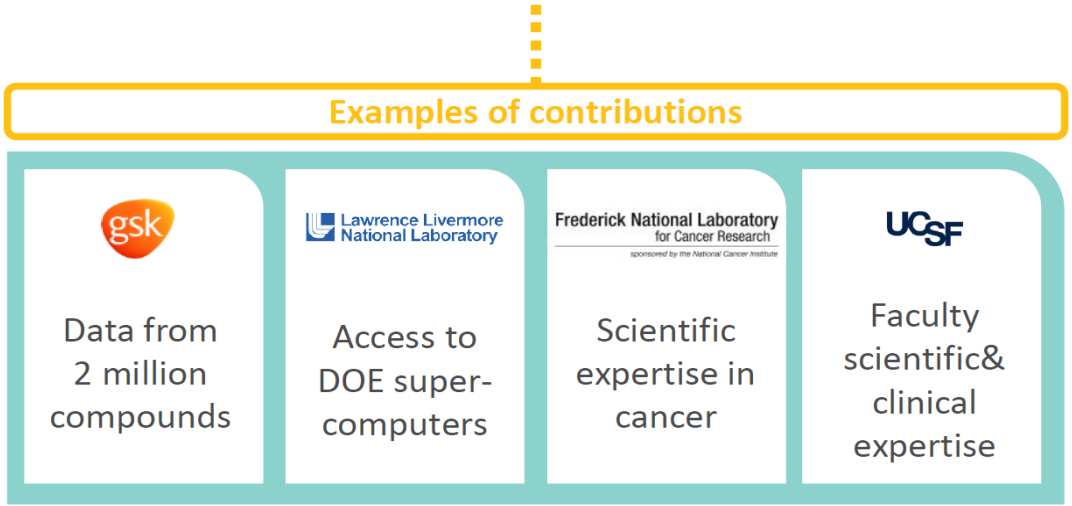
Highlights:

- Theory of Uncertainty Quantification is central to all scientific results and to all pilots; Many techniques developed
- Highly technical subject
- UQ collaboration cuts across all the pilots: each pilot may require different approaches at estimating uncertainty
- Certain tasks related to UQ, are common tasks and are implemented in CANDLE
    - Cross-validation
    - Rademacher bounds
    - Learning path robustness
    - Random weight robustness
    - Certainty distillation

ATOM: Accelerating Therapeutics for Opportunities in Medicine

Highlights:

- Strong public-private partnership
- Builds on both NCI and DOE strengths



ATOM: Accelerating Therapeutics for Opportunities in Medicine

Patient-specific data and samples input to workflow to develop new therapeutics

ATOM Workflow

design

synthesize

Empirical

simulate

In Silico

active learning

assay

PK

Safety

Efficacy

Released to public after a 1-year member benefit

Open source models and generated data

Members use workflow for internal drug discovery

Models of drug behavior in humans

Therapeutic Candidates

Commercialization by members for patient benefit

CANDLE, UQ, ATOM

Proposed Next Steps:

- CANDLE
  - Consider improving the application of CANDLE by partnering with NCI's ITCR program.
- UQ
  - Determine feasibility of specifying sources of uncertainty.
  - Determine feasibility of addressing validation in relation to UQ.
- ATOM
  - Consider the feasibility of establishing a nonprofit entity.
  - Consider expanding partnership to gain more relevant datasets.

Conclusions

- Overall, the three pilots are moving well and forcefully in their domains
- The cross cutting elements CANDLE and Uncertainty Quantification are essential components to all projects and are also developing well
- Through CANDLE there exists the foundation for strong participation of the wider research community
- ATOM, possibly organized as a not-for-profit corporation, could achieve major improvements in the time to develop new therapeutics

Overall Proposed Next Steps

- Consider strengthening the hub at FNLCR to better connect the DOE efforts to the large number of NCI supported programs that could support and/or profit from the collaboration
- While our working group can provide a broad evaluation of the pilots and cross-cutting efforts, many of the projects, such as Deep Learning, Uncertainty Quantification or Multi-scale Molecular Dynamics are highly specialized
  - At some time in the future, consider organizing *ad hoc* Working Groups in these subjects
- Start to plan for proposal driven scientific research using the tools and resources devoted to this program
  - For example, additional pilots or projects to answer scientific questions using all the machinery developed by the collaboration