Frederick National Laboratory for Cancer Research

JDACS4C – Joint Design of Advanced Computing Solutions for Cancer

Frederick National Laboratory Advisory Committee Meeting

Eric Stahlberg, PhD

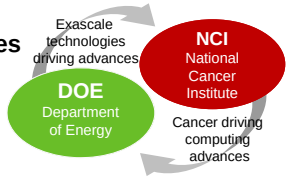
May 8, 2017

The Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute
DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

JDACS4C – NCI-DOE Collaboration Overview

Frederick National Laboratory for Cancer Research

- **Shared Interests**
  - Cancer scientific challenges driving advances in computing
  - Exascale technologies driving cancer advances
- **Supports Two Primary Executive Office Initiatives**
  - Precision Medicine Initiative (Jan 2015)
  - National Strategic Computing Initiative (July 2015)
- **Three Pilot Efforts**
  - Molecular domain pilot
    - NCI Lead: **Frank McCormick** (FNL/UCSF), **Dwight Nissley** (FNL)
    - Lead DOE Leads: **Fred Streitz** (LLNL), **Felice Lightstone** (LLNL)
  - Pre-clinical domain pilot
    - NCI Lead: **Jim Doroshow** (DCTD) and **Yvonne Evrard** (FNL)
    - Lead DOE Leads: **Rick Stevens** (ANL)
  - Population/clinical domain pilot
    - NCI Lead: **Lynne Penberthy** and **Paul Fearn** (DCCPS)
    - Lead DOE Labs: **Gina Tourassi** (ORNL), **Gil Weigand** (ORNL)

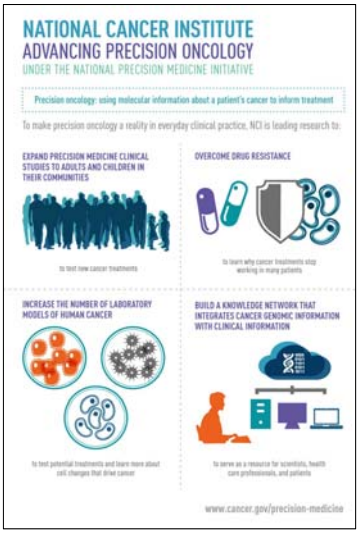


JDACS4C Collaboration Pilots: Capabilities to Accelerate Precision Oncology

Frederick National Laboratory for Cancer Research

NCI Mission Impact:
Accelerating development of new treatment options for precision cohorts

Pilot 1:
Pre-clinical Models
Predictive patient drug response models with advanced computing



Pilot 2:
Biological Models
Multi-scale computational biological models

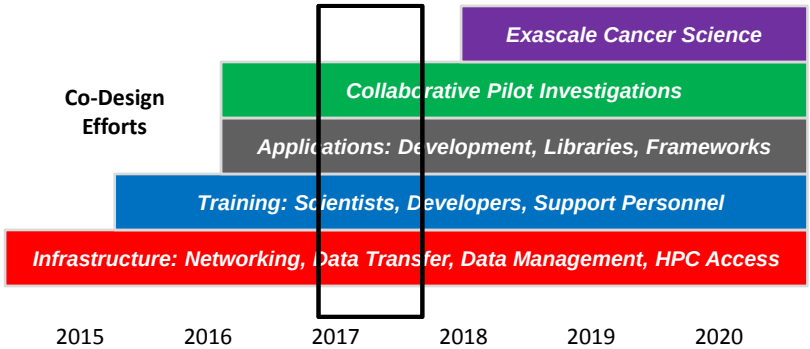
Pilot 3:
Cancer Surveillance
Computational insight into factors impacting clinical response

Accelerating Computational and Data-driven Cancer Research

Frederick National Laboratory for Cancer Research

Exascale in a nutshell:

- Millions of CPU cores contributing to a single task
- Nearly 1000 times faster than fastest computer today
- Focus of DOE Advanced Strategic Computing



Joint Design of Advanced Computing Solutions for Cancer

Frederick National Laboratory for Cancer Research

JDACS4C

Exascale technologies driving advances

NCI National Cancer Institute

DOE Department of Energy

Cancer driving computing advances

Initiatives Supported NSCI and PMI

NATIONAL CANCER INSTITUTE

Argonne · Oak Ridge · Lawrence Livermore National Laboratory · Los Alamos

Frederick National Lab for Cancer Research

Integrated Precision Oncology

Molecular — Pre-clinical — Population

Pre-clinical Domain – Improved predictive models
Computational/hybrid predictive models of drug response
Improved experimental design

Clinical Domain – Precision oncology surveillance
Expanded SEER database information capture
Modeling patient health trajectories

Molecular Domain – Multiscale biological models
Models for RAS-RAS complex interactions
Insight into RAS related cancers

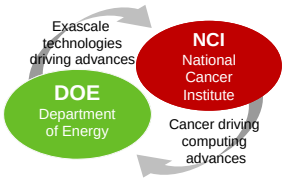
CANcer Distributed Learning Environment (CANDLE)
Scalable Deep Learning for Cancer

JDACS4C established June 27, 2016 with signed MOU between NCI and DOE

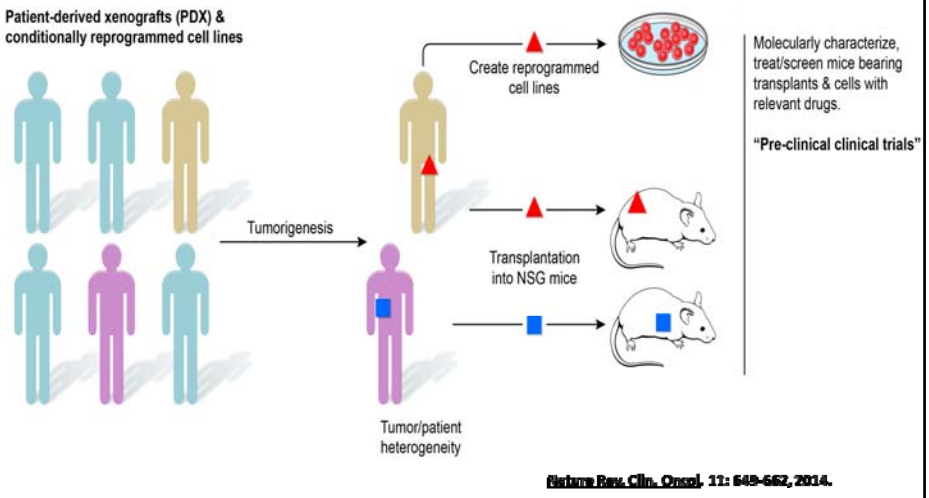
Frederick National Laboratory for Cancer Research

JDACS4C Pilot 1 Highlights

Predictive Modeling for Pre-Clinical Screening



Pilot 1 : Pre-clinical Domain



Pilot 1: Overarching Goal, Aims and Impacts

Goal: Improve predictive capabilities for pre-clinical screening, through advances in machine learning and integrated modeling

Aim 1:

Predictive Models of Drug Response

- Cell Line ⟹ Cell Line
- Cell line ⟹ CLX
- CLX ⟹ CLX
- Cell Line + CLX ⟹ PDX
- PDX ⟹ PDX
- ALL ⟹ ANY

Aim 2:

UQ and Improved Experimental Design

- More cell line types or more drugs?
- More assay types or more replicates?
- Which assay types are most predictive?

Aim 3:

Develop Hybrid Predictive Models

- Combining mechanistic models and machine learning
- Large-scale simulation vs training vs accuracy
- Hypothesis formation

Impact DOE: Co-design of architectures integrating learning systems and simulation
Impact NCI: Expand breadth of treatments for cancer precision medicine

Pilot 1: Cross Laboratory Team

Frederick National Laboratory for Cancer Research

ANL: **Rick Stevens**, Frank Alexander, Jillian Aurisano, Prasanna Balaprakash, Tom Brettin, Jim Davis, Emily Dietrich, Nicoli Dryden, Hal Finkel, Ian Foster, Monisha Ghosh, Ushma Kriplani, Ravi Madduri, Sergei Maslov, Bob Olson, Dan Olson, Mike Papka, Lorenzo Pesce, John Santerre, Maulik Shukla, Venkat Vishwanath, Fangfang Xia

LANL: **Marian Anghel**, Tanmoy Bhattacharya, Judith Cohn, Paul Dotson, Will Fischer, Kumkum Ganguly, Jason Gans, Cristina Garcia-Cardona, Nick Hengartner, William Hlavacek, John Hogden, Patrick Kelly, Miranda Lynch, Ben Mcmahon

LLNL: Jonathan Allen, Ya Ju Fan, Adam Zemla

ORNL: **Mike Lueze**, Arvind Ramanathan

NCI: James Doroshow, Yvonne Evrard, Susan Holbeck, Eric Stalhberg, George Zaki

P1 Hackathon 2

CANDLE Hackathon 1

P1 Hackathon 3

Pilot 1: Recent Key Results

Frederick National Laboratory for Cancer Research

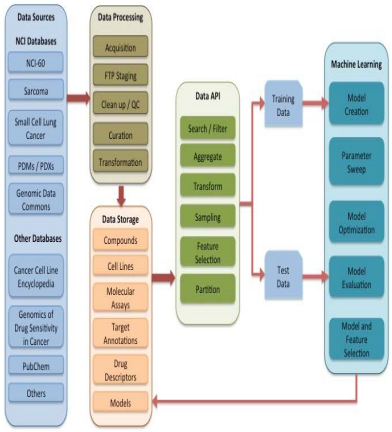
- **Data environment in place**$_{ANL,NCI,LANL}$
- **PDX model production, NCI-60 RNAseq**$_{NCI}$
- **Aim1 : Predictive Models of Drug Response**$_{ANL, UC, LLNL,LANL}$
  - Shallow models (feature selection and drug/tumor specific signatures)
  - Deep models to utilize larger training datasets (single drug, drug pairs)
  - Demonstration of convolution success* in P1B3 (CANDLE benchmark)
  - Initial transfer learning models and autoencoders (expression, drugs)
- **Aim2 : UQ and Improved Experimental Design**$_{LANL,ANL}$
  - QA, clustering and mapping studies of NCI60, CCLE, GDSC, GDC
  - Initial tradeoff studies (sample size vs error)
  - Model transfer problem formulation and initial development
  - Feature analysis {transcriptome > proteome > kinome}, {exp >> SNPs}, {RNAseq ≧ microarray}
- **Aim3 : Develop Hybrid Predictive Models**$_{ANL, UIUC}$
  - Initial molecular network based convolution experiments*
- **PILOT1 Benchmarks for CANDLE (three model problems)**
  - NCI workshop in April 18-19th
  - NVIDIA workshop May 9th

10

Pilot 1: Data Sources and Integration

Frederick National Laboratory for Cancer Research

- **NCI-60 Cell Lines**
  - 61 cell lines, 92,691 compounds, 40 assay types
- **NCI Sarcoma Project**
  - 74 cell lines, 445 compounds
- **NCI Small Cell Lung Cancer Project (SCLC)**
  - 76 cell lines, 525 compounds
- **NCI Patient-derived Models and Xenografts (PDM**
  - 274 PDM/PDX Models
- **Broad-Novartis Cancer Cell Line Encyclopedia (C**
  - 504 cell lines, 24 compounds
- **Genomics of Drug Sensitivity in Cancer (GDSC)**
  - 1,074 cell lines, 265 compounds
- **Genomic Data Commons (GDC)**
  - 14,531 samples, 29 primary sites, 38 cancer types



Cancer Data Processing, Storage and Machine Learning Workflow

U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

Pilot 1: Early Insights

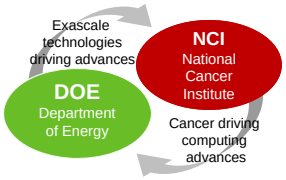
Frederick National Laboratory for Cancer Research

- Effort is not limited by model representations – many model types have predictive capability on response problems
- Efforts are constrained by available data that capture the distributions expected in clinically relevant cases
- Convolutional neural networks have shown promise on expression and drug descriptors
- Shallow models can learn responses for single drugs/tumor types and be used for feature selection
- Transfer learning can be used to learn features in one problem and reapplied in another problem

Frederick National Laboratory for Cancer Research

JDACS4C Pilot 3 Highlights

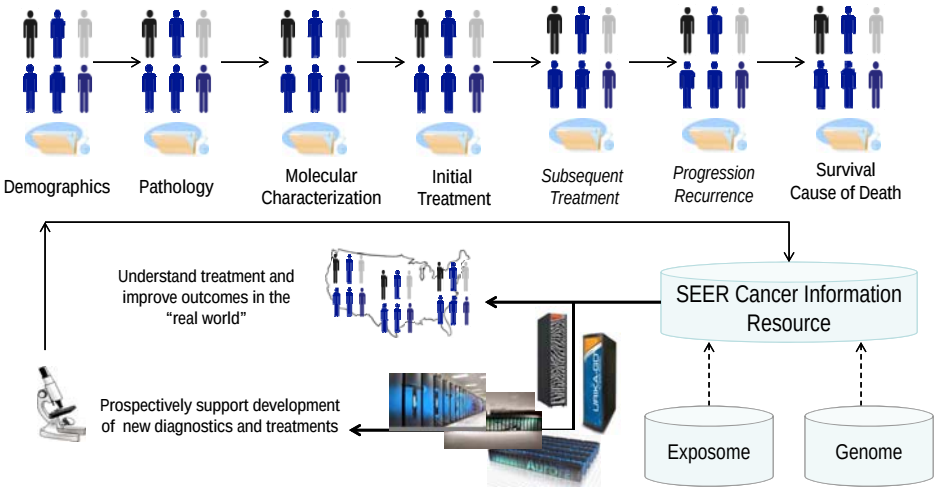
Population Information Integration, Analysis, and Modeling



Frederick National Laboratory for Cancer Research

Pilot 3: Population Domain

Surveillance data captured on each cancer patient for the entire population

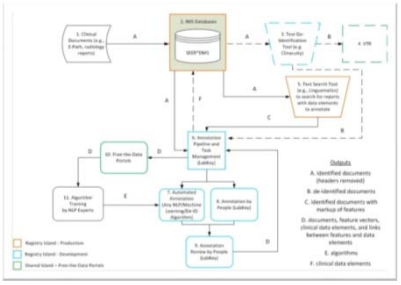


Pilot 3: Community Engagement

- **NCI Team**
- **DOE Labs- Oak Ridge, Los Alamos, Lawrence Livermore, Argonne**
- **4 SEER registries- Kentucky, Louisiana, Georgia, Seattle**
- **Contracts/Sub-Contracts**
  - IMS as honest broker and agent for registries, hosting and sub-contracts
  - Software and support from LabKey, Linguamatics for Clinical Document Annotation & Processing (CDAP) Pipeline
  - Annotation Services
- **Data acquisition and linkages (e.g. claims, pharmacy, radiation oncology)**
- **Clinical experts to lead use cases for breast, colorectal, lung, and prostate cancer**
- **NLP Workshop with NCI, DOE, CDC, FDA and academic partners**

Pilot 3: AIM 1 - Progress Update

- **Annotation Framework:**
  - Utilizing clinical document annotation pipeline to annotate ALK, EGFR biomarkers in e-path reports to send to DOE for algorithm development
  - Developing schema for annotation of recurrence, progression data elements in path reports for breast & colorectal cancer
  - Plans to scale up pipeline to annotate up to 10,000 documents per month
    - Add biomarkers from breast, colorectal, lung, prostate and pathology elements from CAP protocols

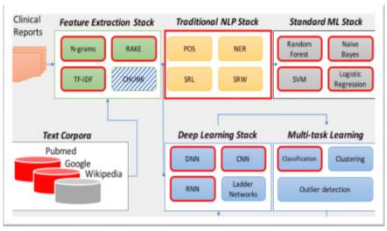


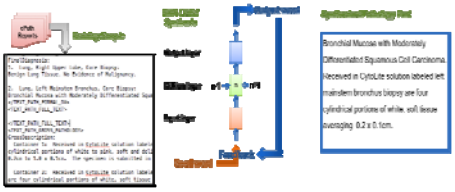
16

Pilot 3: AIM 1 - Progress Update

Frederick National Laboratory for Cancer Research

- **Text Comprehension:**
  - Developed and benchmarked rule-based, conventional ML-based, and DL-based NLP tools for e-paths for three information extraction tasks: (i) primary cancer site, (ii) histological grade, (iii) behavior
    - DL tools: CNN, HAN, MT-DNN
    - DL interpretability
    - Cross-registry robustness validation
  - Established reproducible experimental design pipeline for future studies with new data and different tasks
- **Text Synthesis:**
  - Generative models for e-path text synthesis under development
  - Character and word-based LSTM-RNNs e-path text synthesis under development
  - Initial validation in progress

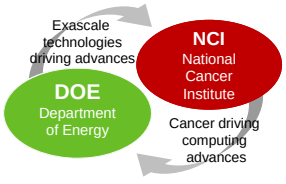




17

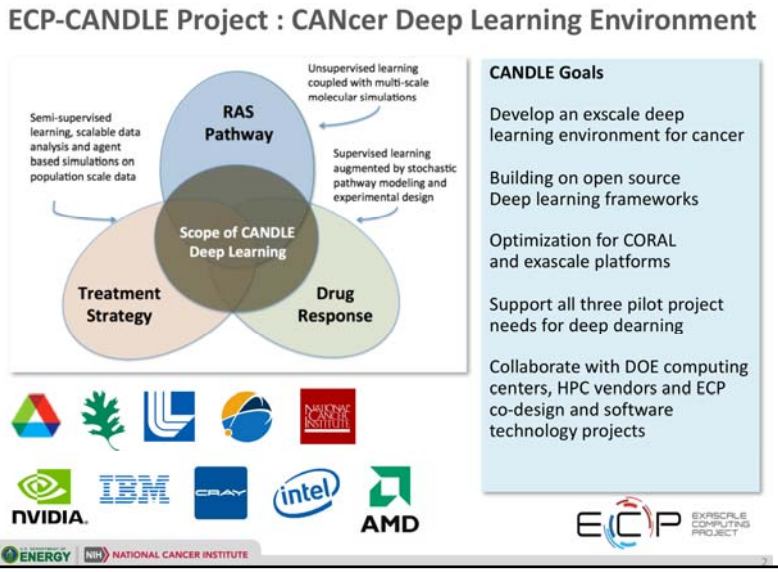
CANDLE Highlights

CANcer Distributed Learning Environment



CANDLE: Deep Learning Across JDACS4C

Frederick National Laboratory for Cancer Research



CANDLE: CANcer Distributed Learning Environment

Frederick National Laboratory for Cancer Research

- **CANDLE is DOE supported contribution to JDACS4C**
- **Four year multi-lab project commencing in September 2016**
- **Focuses on creating scalable, open and portable Deep Learning framework**
- **Supports Deep Learning needs for all JDACS4C pilots**
  - DOE scientific leads bring pilot-specific deep learning challenges
- **Open Source software release**
- **Reference benchmarks released in January 2017**
- **Workshops and early community building**
  - CANDLE @ NIH – April 18-19, 2017 – More than 60 attendees involving more than 12 NIH institutes/centers
  - CANDLE @ GTC – May 9, 2017

JDACS4C Summary

- **Building, motivating and expanding interdisciplinary collaborations and partnerships**
- **Piloting cross-disciplinary activities to enable integrated precision oncology**
  - Identifying key linkages among and gaps between domains
- **Delivering scientific insight together with new capabilities**
  - Scalable frameworks, environments, conventions and standards
  - Cutting-edge predictive models for cancer across multiple domains
  - Advanced machine learning accounting for uncertainty in data
- **Enabling open and team science to more rapidly achieve goals in precision oncology**

Acknowledgements

- **NCI CBIIT**
  - Warren Kibbe, Carl McCabe, Betsy Hsu and the CBIIT team
- **NCI DCCPS**
  - Lynne Pemberthy, Paul Fearn, Jessica Boten and DCCPS team
  - Louisiana, Seattle, Georgia and Kentucky cancer registries
- **NCI DCTD**
  - Jim Doroshow, Susan Holbeck
- **FNLCR**
  - Data Science and IT Program
  - RAS Initiative - Dwight Nissley, Frank McCormick and RAS team
  - PDX - Yvonne Evrard and PDX team
- **Department of Energy**
  - Dimitri Kusnesov, Steve Binkley, Doug Wade, Carolyn Lauzon
- **Argonne National Lab**
  - Rick Stevens, Tom Brettin, Fangfang Xia, Emily Dietrich, and the ANL team
- **Lawrence Livermore**
  - Jason Paragas, Amy Gryshuk, Fred Streitz, Felice Lightstone, Brian Van Essen, Dave Rakestraw and the LLNL team
- **Los Alamos National Lab**
  - Marian Anghel and LANL team
- **Oak Ridge National Lab**
  - Gina Tourassi, Gil Weigand, Arvind Ramanathan, Joe Lake, and the ORNL team

And many, many more supporting and working with JDACS4C !

22