Frederick National Laboratory
for Cancer Research

Big Data Big Theory

Eric Stahlberg, George Zaki

New York Scientific Data Summit 2018 - Brookhaven National Laboratory

August 6, 2018

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

Frederick National Laboratory for Cancer Research

FNLCR and Data Science

Overview of Frederick National Laboratory for Cancer Research (FNLCR)

Frederick National Laboratory for Cancer Research

- **FNLCR is the only Federally Funded Research and Development Center (FFRDC) dedicated exclusively to biomedical research**
  - Operated in the public interest by **Leidos Biomedical Research, Inc** (formerly SAIC-Frederick) on behalf of the National Cancer Institute
- **Main campus located on 70 acres at Ft. Detrick, MD**
  - Leidos Biomed employees co-located with NCI researchers and other contractors on the NCI Campus at Frederick
  - Additional Leidos Biomed scientists at Bethesda and Rockville sites

Mission

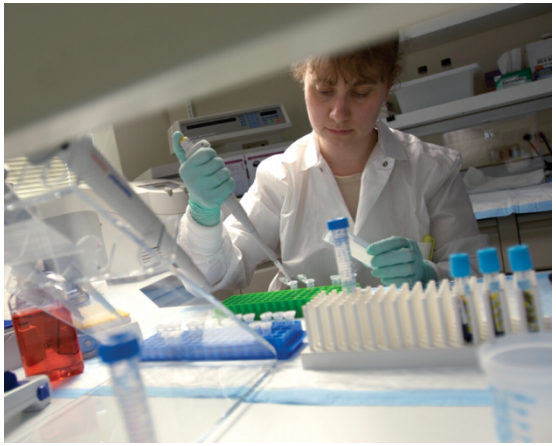
Provide a unique national resource for the development of new technologies and the translation of basic science discoveries into novel agents for the prevention, diagnosis and treatment of cancer and AIDS.

Research & Development at FNLCR

Frederick National Laboratory for Cancer Research

- **Research & Development**
  - **Basic Research**: New knowledge about AIDS and cancer
  - **Applied R&D**: New diagnostics and therapeutics
  - **Clinical Research:** Clinical trials and laboratory analysis
  - **cGMP manufacturing:** Biologicals and vaccine production

- **Specialties**
  - Genomics, proteomics, and metabolomics
  - Bioinformatics and imaging
  - Nanotechnology
  - Animal models
  - Tumor cell biology and virology
  - Immunology and inflammation



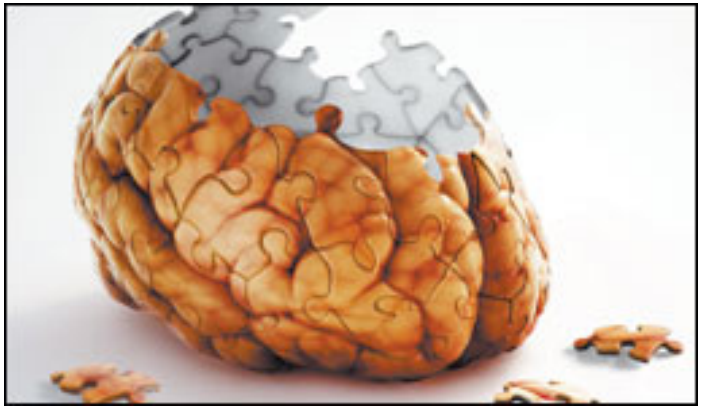
- **Data Science and Information Technology underpin and support all R&D activities and specialties**

Data Science and Biomedical Informatics

Frederick National Laboratory for Cancer Research

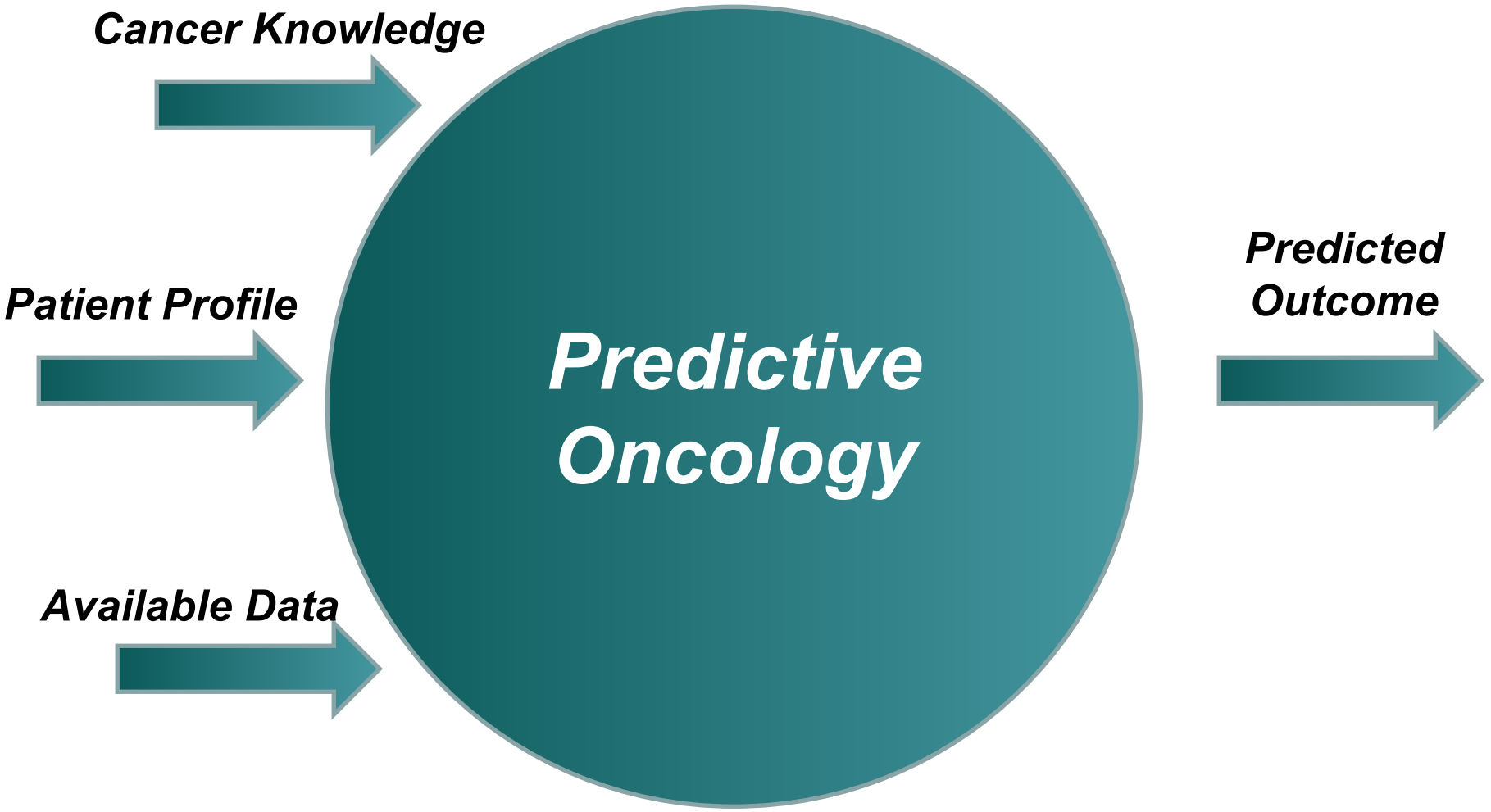
Leverage leading edge data science and information technology skills, tools, and capabilities to accelerate translation of biomedical data to scientific discoveries, medical treatments, diagnostic and prevention tools for cancer and AIDS patients.

Enabling physicians, scientists, and patients to make critical decisions based on knowledge gained from all and not only a fraction of data and information available to them.



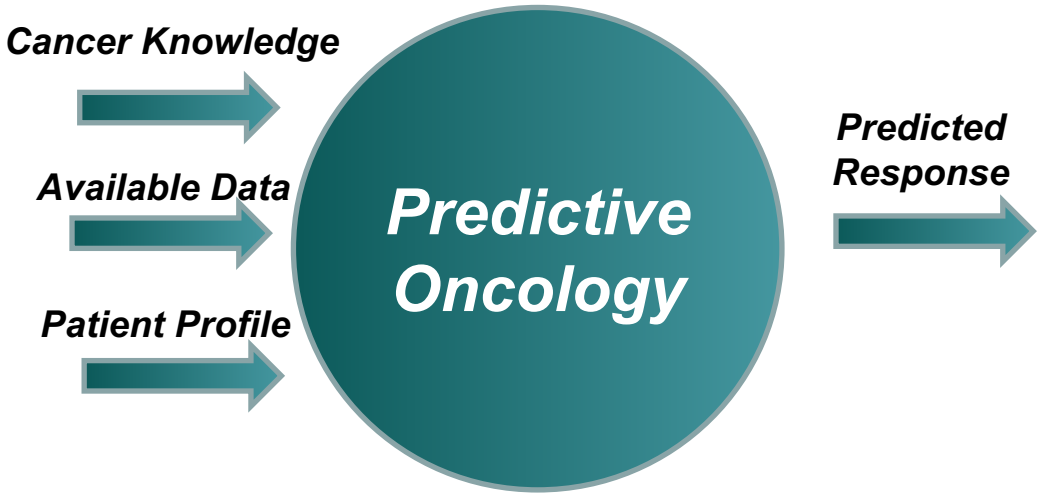
Predictive Oncology

Frederick National Laboratory for Cancer Research



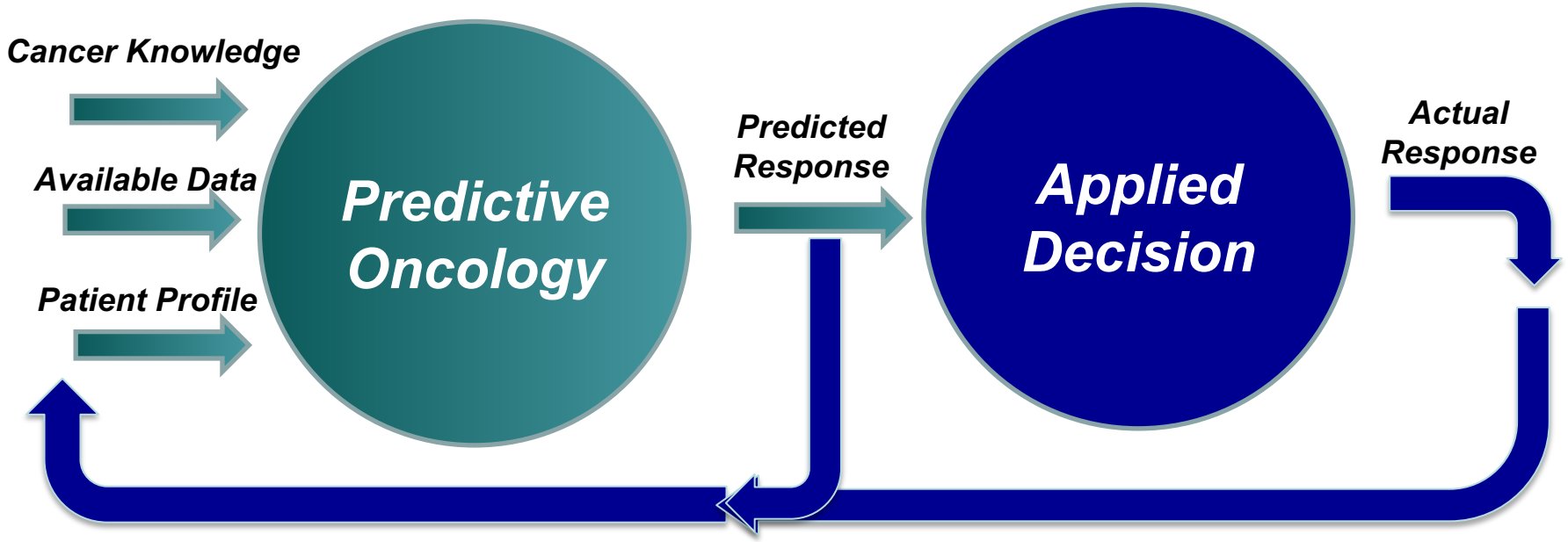
Predictive Oncology – Leveraging and Learning

Frederick National Laboratory for Cancer Research



Predictive Oncology Learning System

Frederick National Laboratory for Cancer Research



Predictive Oncology General Adoption Challenges

Frederick National Laboratory for Cancer Research

- **Potential criteria to achieve broad adoption of predictive oncology models**
  - Credible – scientifically substantiated
  - Reliable – performs as expected for specified conditions
  - Affordable – generally accessible in a cost-effective implementation
  - In patient's best interest – able to aid patient diagnosis or treatment in a timely manner

NCI Precision Oncology
Extending the Frontiers

Frederick National Laboratory for Cancer Research

- Identify promising new treatment options through the use of advanced computation to rapidly develop, test and validate predictive pre-clinical models for precision oncology.

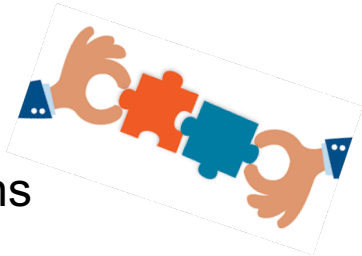
- Deepen understanding of cancer biology and identify new drugs through the integrated development and use of new simulations, predictive models and next-generation experimental data.

- Transform cancer care by applying advanced computational capabilities to population-based cancer data to understand the impact of new diagnostics, treatments and patient factors in real world patients.

Challenge Areas for Predictive Oncology

Frederick National Laboratory for Cancer Research

- **Challenges for cancer**
  - Insufficient data for describing all possibilities
    - Over 250,000 unique cancer characterizations
    - Observation gaps – absence of specific confirming data
    - Bridging molecular with preclinical and preclinical to clinical domains
  - Data fusion and scientific credibility
    - Achieving coherence across scales
    - Achieving coherence and quality across organizations
  - Achieving reliability
    - Consistency of response for characterized conditions
    - Accounting for uncertainty of unknown factors
    - Similarity of behavior across similar models



Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) Developing Key Capabilities

Frederick National Laboratory for Cancer Research

Integrated Precision Oncology

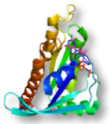
Molecular | Pre-clinical | Population

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

250,000 cancer types

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

1000s of drugs, millions of combinations



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

4 Billions core hours per simulation

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

NIH NATIONAL CANCER INSTITUTE

Argonne NATIONAL LABORATORY

OAK RIDGE National Laboratory

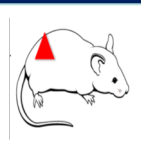
Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY EST. 1943

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

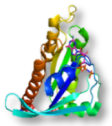
Integrated Precision Oncology

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

Integration for Predictive Oncology

Frederick National Laboratory for Cancer Research

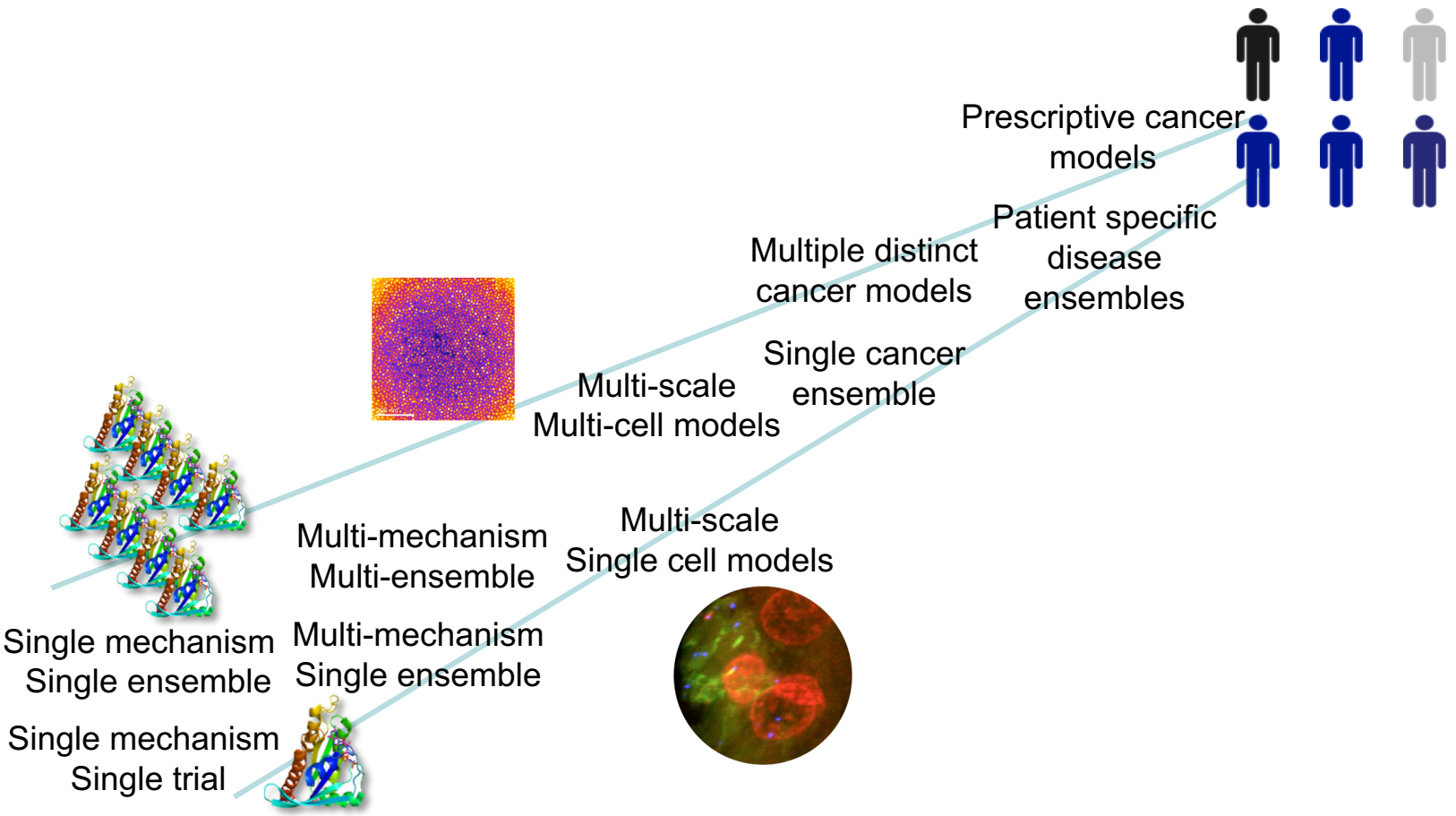
- **Timely, accurate, adoptable and affordable precision oncology predictions to inform clinical decisions**
  - Spanning domain of available information from molecular insights to population factors
- **Requires integrating new capabilities across multiple observational domains**
- **Integrate JDACS4C pilots across the domain of scales**
- **How to develop predictive models that are efficient while maintaining experimental, scientific and computational credibility?**

1942 Alaskan Highway Construction Analogy

- *Pathfinders to survey and find ways in each area*
- *Deployed seven engineering regiments concurrently with unique assignments – each about 350 miles to complete*
- *Common goals to connect with the other by a set time*
- *Within each regiment, two teams leapfrogged each other*
- *Technical impediments (like permafrost) required novel engineering, shared solutions across the effort*

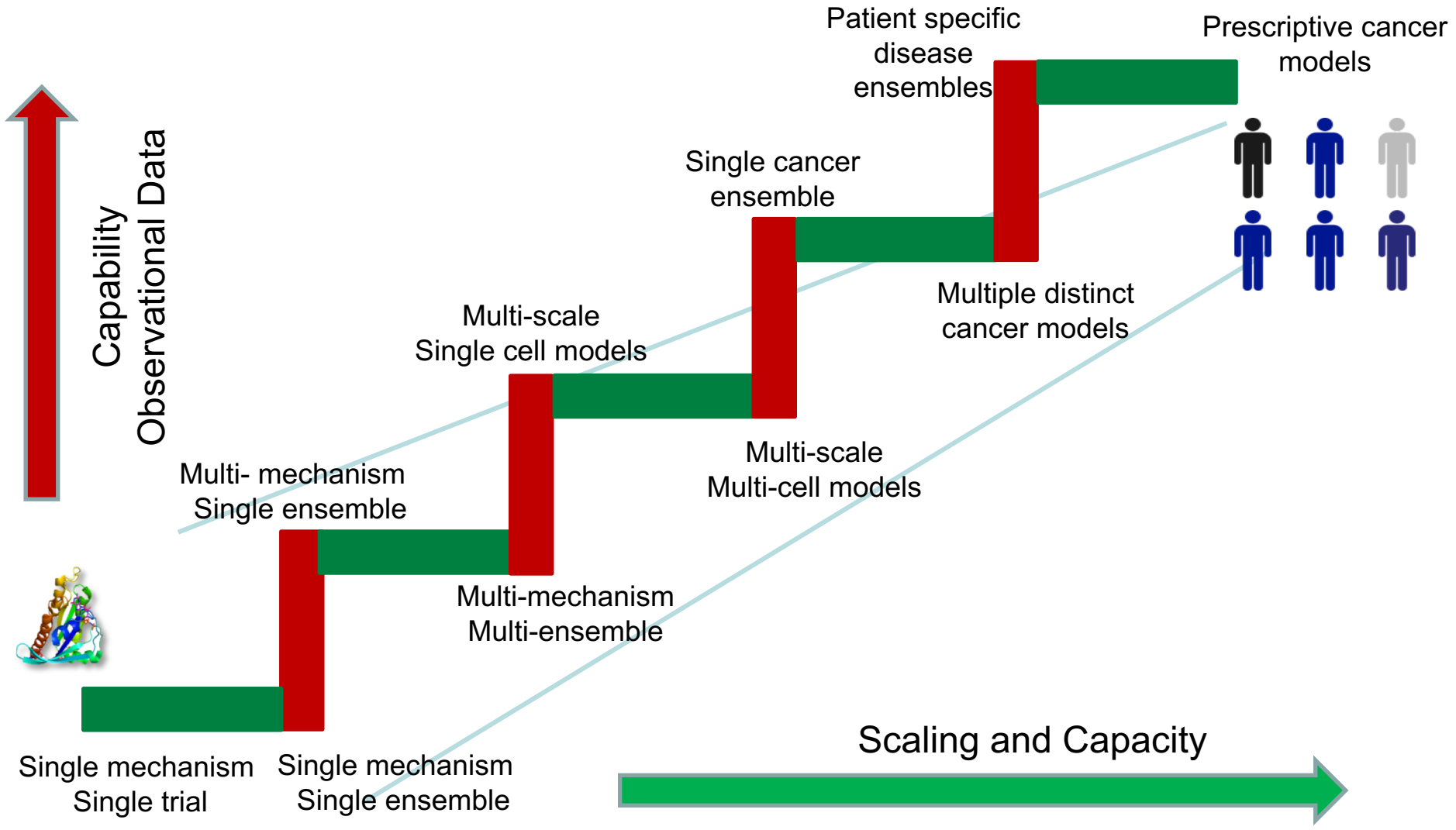
Multi-domain computing for integrated prescriptive cancer predictions

Frederick National Laboratory for Cancer Research



Complex System Modeling Drives Demand for Computing

Frederick National Laboratory for Cancer Research



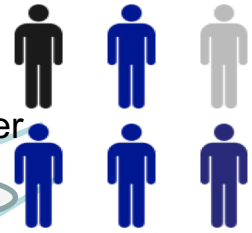
JDACS4C Pilots Pioneering Computational Capabilities in New Areas

Frederick National Laboratory for Cancer Research

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

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

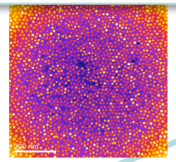
Molecular Domain – Multiscale biological models
Models for RAS-RAS complex interactions
Insight into RAS related cancers



Prescriptive cancer models

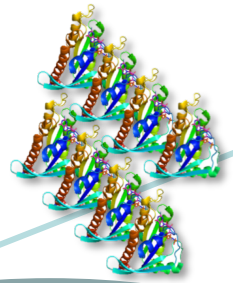
Patient specific disease ensembles

ltiple distinct cancer models



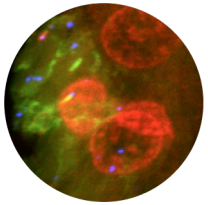
Multi-scale Multi-cell models

Single cancer ensemble



Multi-mechanism Multi-ensemble

Multi-scale Single cell models



Single mechanism Single ensemble

Multi-mechanism Single ensemble

Single mechanism Single trial

Complex System Modeling Drives Demand for Computing

Frederick National Laboratory for Cancer Research

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

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

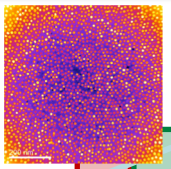
Molecular Domain – Multiscale biological models
Models for RAS-RAS complex interactions
Insight into RAS related cancers

Prescriptive cancer models

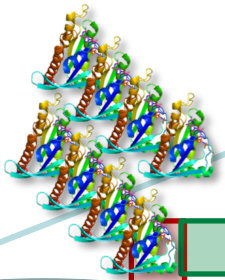
Patient specific disease ensembles

Multiple distinct cancer models

Single cancer ensemble

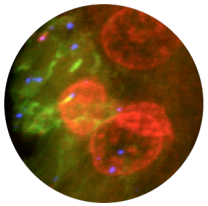


Multi-scale Multi-cell models



Multi-mechanism Multi-ensemble

Multi-scale Single cell models



Single mechanism Single ensemble

Multi-mechanism Single ensemble

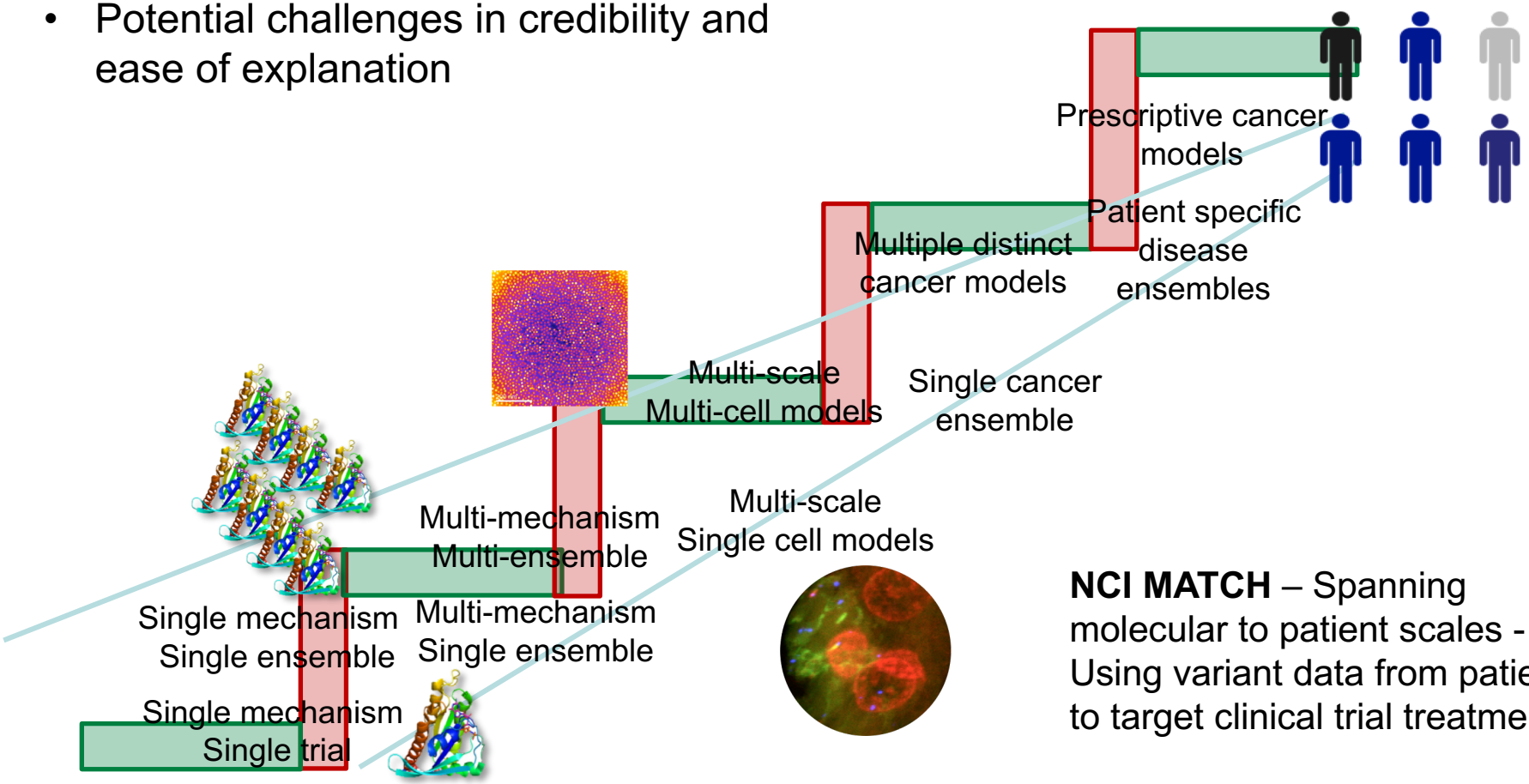
Single mechanism Single trial

Complex System Modeling Drives Demand for Computing

Frederick National Laboratory for Cancer Research

Ensembles of models –

- Promises for affordability and customizability
- Potential challenges in credibility and ease of explanation



NCI MATCH – Spanning molecular to patient scales - Using variant data from patients to target clinical trial treatments

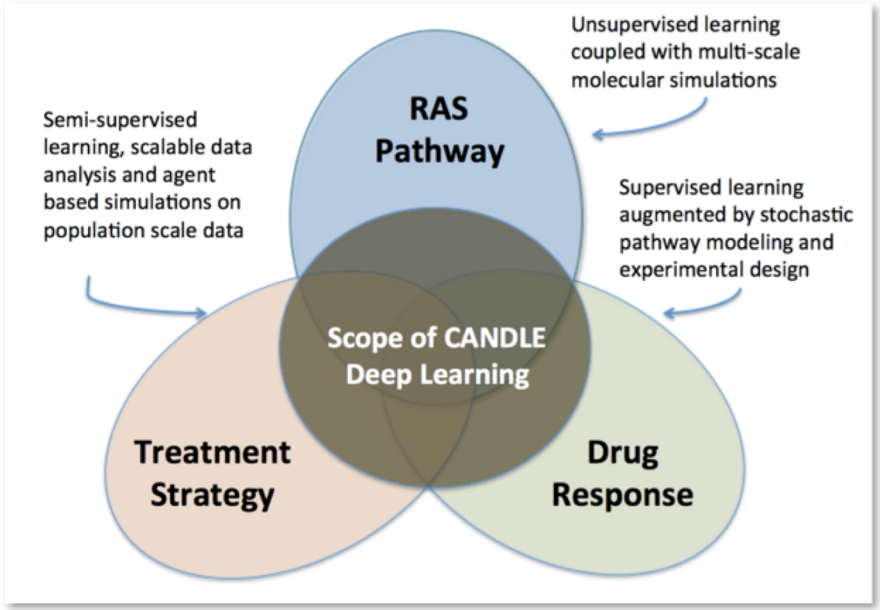
From ensemble to predictive model

- Can ensemble data lead to a reduced representation using a predictive model? What is the cost?
  - Deep learning can be an avenue to get to that point without having to do the computational investment in the ensemble.
  - How to build a predictive model for that corpus of data?
    - Data size
    - Model complexity
    - Hyperparameters
    - Available compute
- CANDLE is a tool to scale the training of these DL models on next generation of Exascale systems.

CANDLE – Deep Learning Across JDACS4C

Frederick National Laboratory for Cancer Research

ECP-CANDLE Project : CANcer Deep 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

NVIDIA IBM CRAY intel AMD

ECP EXASCALE COMPUTING PROJECT

CANDLE – CANcer Distributed Learning Environment

Frederick National Laboratory for Cancer Research

- **CANDLE is DOE Funded contribution to JDACS4C**
- **Four year project.**
- **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**
- **FNL brings NCI connection to CANDLE**
  - Translating computational environment to broader cancer research community
  - Portability and standardization of model representations
  - Conventions and methods for model validation and evaluation in cancer

CANDLE Software Stack

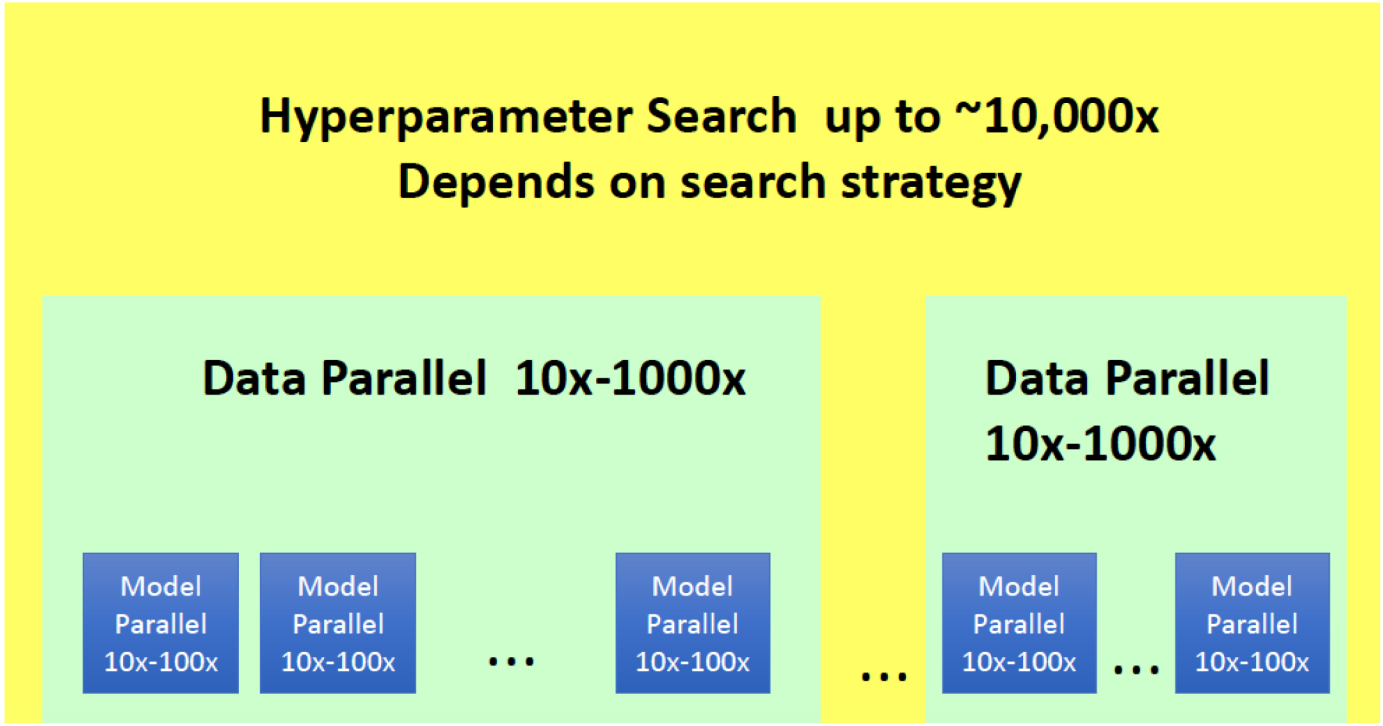
Frederick National Laboratory for Cancer Research

Hyperparameter Sweeps, Data Management (e.g. DIGITS, Swift, etc.)	Workflow
Network description, Execution scripting API (e.g. Keras, Mocha)	Scripting
Tensor/Graph Execution Engine (e.g. Theano, TensorFlow, LBANN-LL, etc.)	Engine
Architecture Specific Optimization Layer (e.g. cuDNN, MKL, etc.)	Optimization

CANDLE - Multi-level Parallelism

Parallelism Targets in CANDLE

10,000 x 10-1000 x 10-100 = 1M – 1000M "cores"



Summary

- Predictive oncology faces big data and big theory challenges to achieve wide adoption.
- Learning is an important aspect of improving models.
- DOE and NCI have embarked on pilot efforts to explore challenges and opportunities in different domains.
- Achieving coherence from the molecular domain to the clinical domain presents computational, engineering and scientific challenges.
- Deep learning and CANDLE point to potential paths forward.

Wrap-up

Frederick National Laboratory for Cancer Research

Thank you!

Questions?

George Zaki (george.zaki@nih.gov)