Computing and Data in Precision Oncology

Warren A. Kibbe, Ph.D.
Professor, Biostats & Bioinformatics
Chief Data Officer, Duke Cancer Institute
warren.kibbe@duke.edu

@wakibbe

@supercomputing
#sc19
#DataSharing
#ComputationalPhenomics
#PrecisionOncology

Personal & Professional Background

- PhD in Chemistry at Caltech, Postdoc in molecular genetics of RAS
- Cancer research for 20+ years - cancer informatics, data science, healthcare
- Faculty in the Feinberg School of Medicine at Northwestern for 15+ years
- Director NCI CBIIT 2013-2017; NCI CIO 2013-2017; Acting NCI Deputy Director for Data Science 2016-2017
- Lost three grandparents to cancer, father to cancer in 2019

Take homes

- Biology, cancer is a grand challenge
- Computation, ML, High Dimensional Modeling is crucial for understanding cancer
- Advances in biology, human health, and implementation science are all needed
- Scaling is a hard problem

As of January 2019, there are an estimated

16,900,000

cancer survivors in the U. S.

From https://cancercontrol.cancer.gov/ocs/statistics/statistics.html , based on Bluethmann SM, Mariotto AB, Rowland, JH. Anticipating the ''Silver Tsunami'': Prevalence Trajectories and Comorbidity Burden among Older Cancer Survivors in the United States. Cancer Epidemiol Biomarkers Prev. (2016) 25:1029-1036

In 2030, there will be an estimated

22,000,000

cancer survivors in the U. S.

Survival, incidence, and all-cause mortality rates were assumed to be constant from 2016 through 2030.

From https://doi.org/10.3322/caac.21565 Miller KD, Nogueira L, et al, Cancer Treatment and Survivorship Statistics, 2019. CA Cancer J Clin (2019) 0:1-23

Cancer is a grand challenge

Courtesy of P. Kuhn (USC)

Requires:

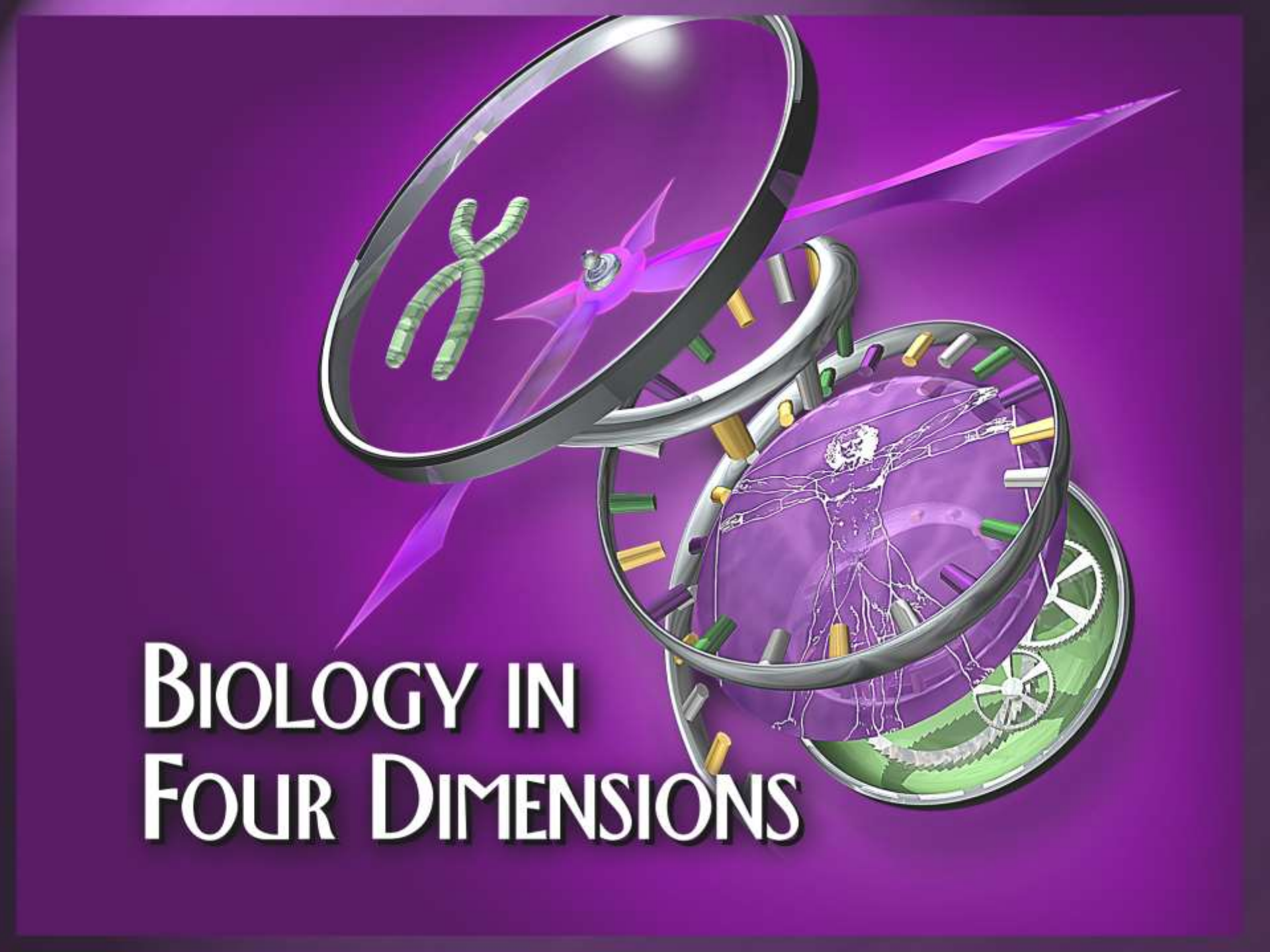
- Deep biological understanding
- Advances in scientific methods
- Advances in instrumentation
- Advances in technology
- Data and computation
- Mathematical models

*Cancer Research and Care generate detailed **data** that is critical to create a learning health system for cancer*

Understanding Cancer

- **Precision medicine** will lead to **fundamental understanding** of the complex interplay between genetics, epigenetics, nutrition, environment and clinical presentation and **direct effective, evidence-based prevention and treatment**.

Ramifications across many aspects of health care

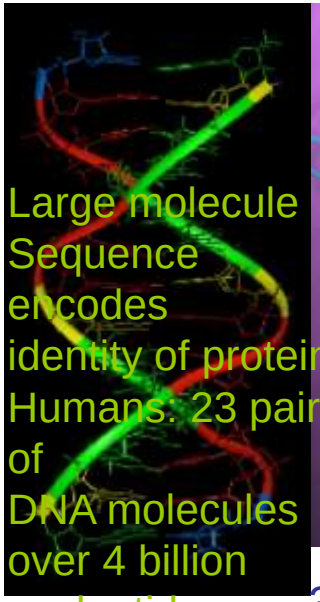

BIOLOGY IN
FOUR DIMENSIONS

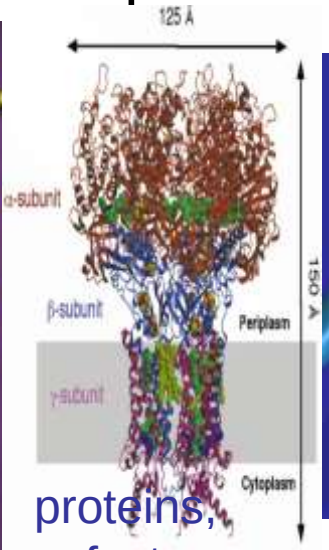
Biological Scales

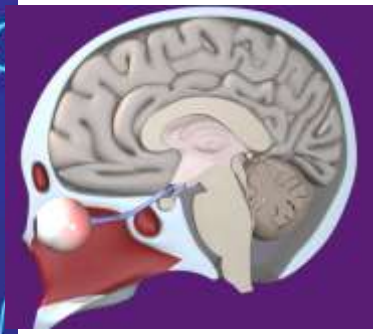
Molecular to Systems Biology

Size scale (meters)

10^{-9}- 10^{-4}	10^{-9} -10^{-8}	10^{-8} -10^{-7}	10^{-7} -10^{-5}	10^{-6} -10^{-2}	10^{-3} -1
DNA	Protein	Complexes	Organelles cells	Tissues	Organs
Large molecule Sequence encodes identity of protein Humans: 23 pairs of DNA molecules over 4 billion nucleotides per pair	30,000 proteins with variants	proteins, cofactors, metabolites, 2nd msgrs	compartments structures function	signaling, networks	emergent properties
10^{9}	10^{-12} -10^{-4}	10^{-6} -10^{3}	10^{-3} -10^{4}	1 -10^{8}	1 -10^{9}



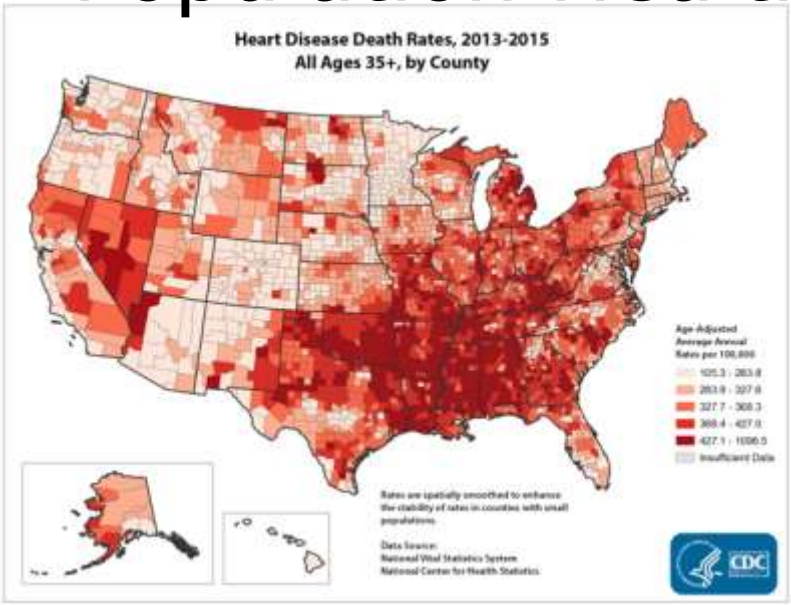


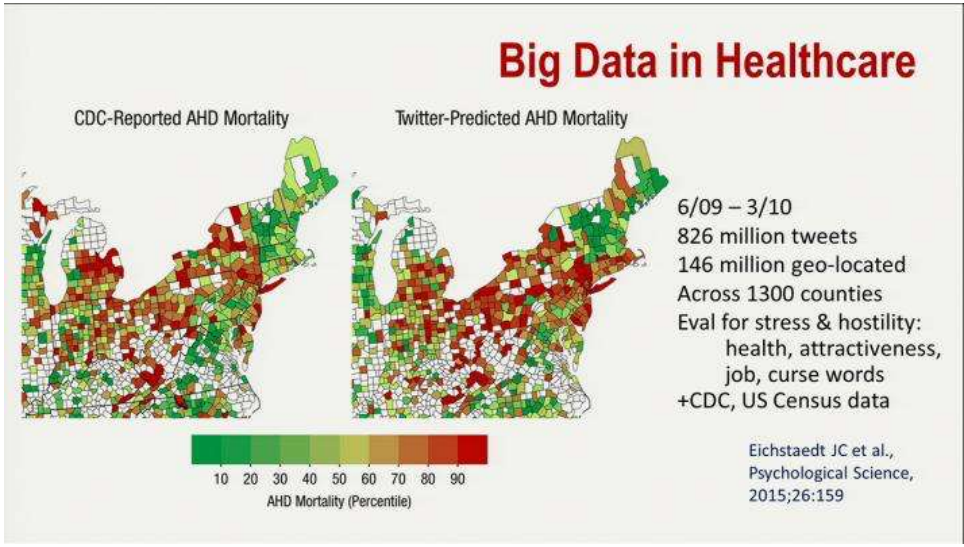


Time scale (seconds)

Health vs Disease

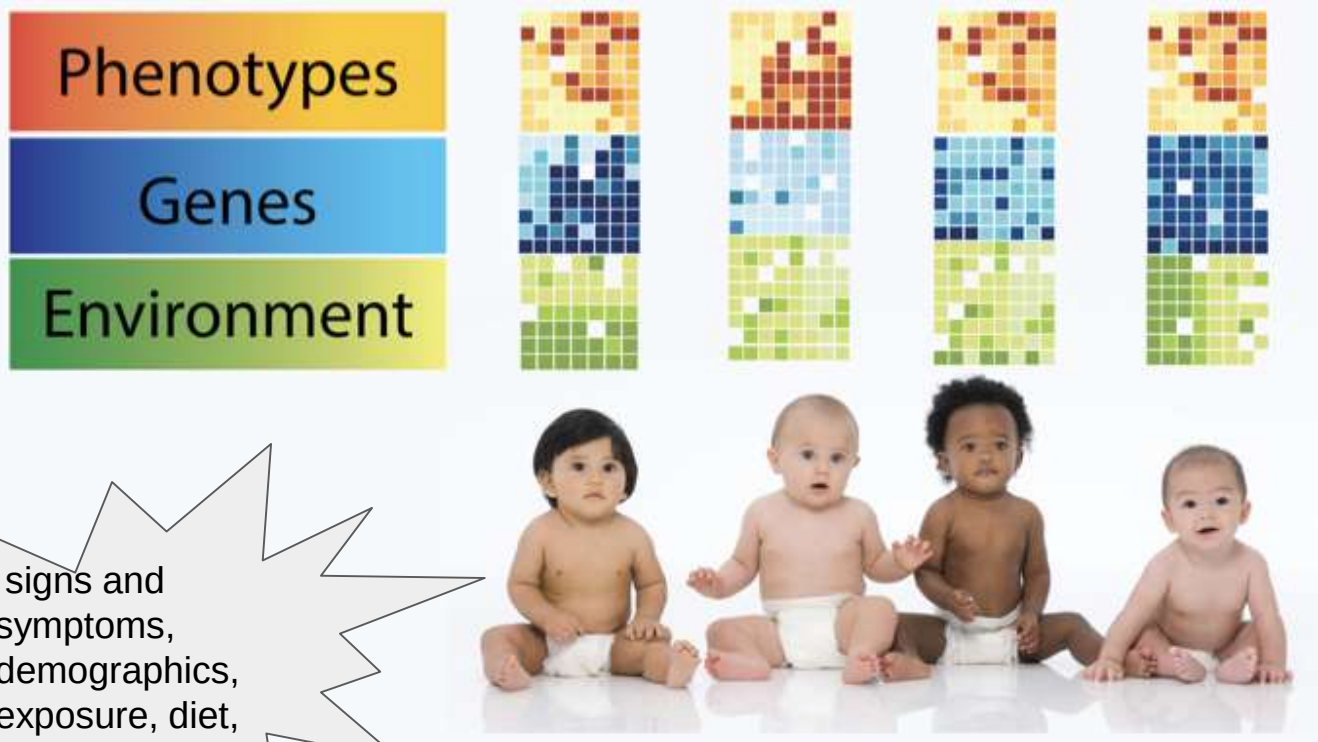
- What is 'normal'?
- Systematic and measurement error
- Biological heterogeneity
- Population Health





The promise of precision medicine:

How can we meaningfully group patients?



signs and symptoms, demographics, exposure, diet, traits, etc.

Slide from Melissa Haendel

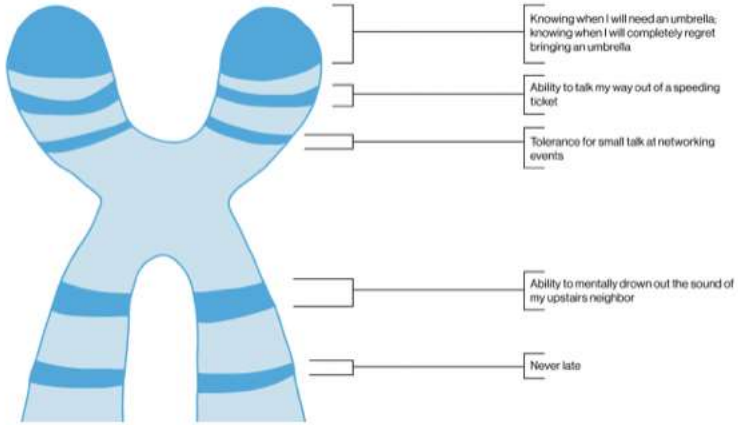
Biology and Medicine are now data intensive enterprises

Scale is rapidly changing

Technology, data, computing and IT are pervasive in the lab, the clinic, in the home, and across the population

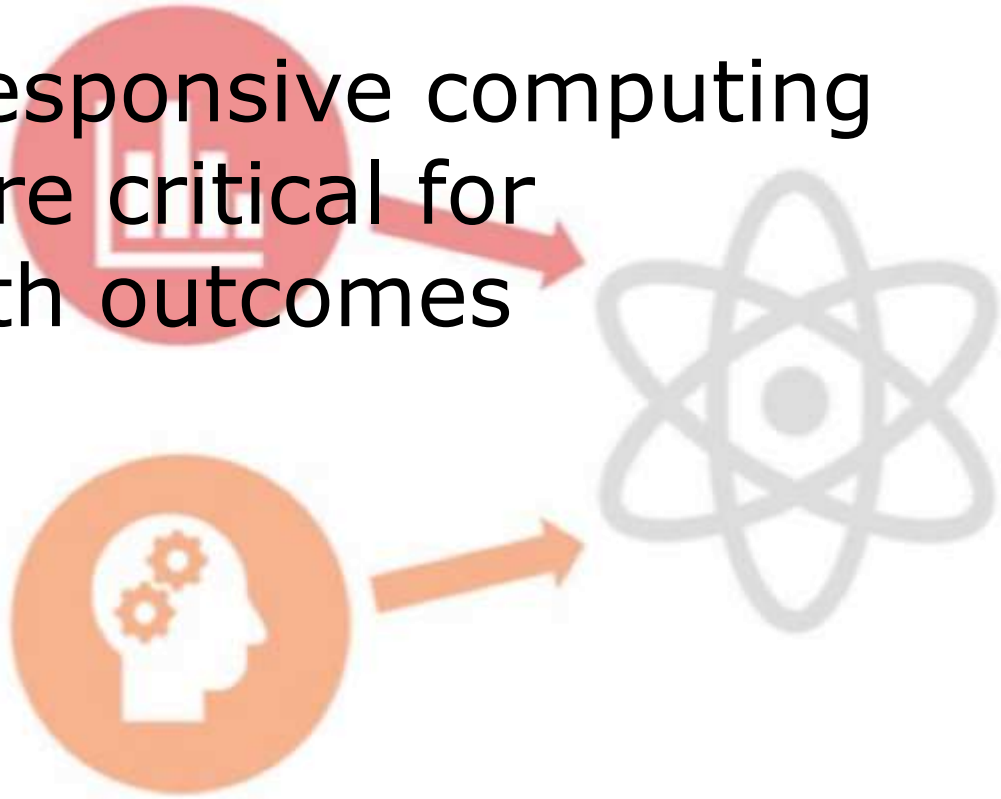
Healthcare

- Evidence is not consistently accessible and structured
- Outcomes are not connected to care
- Patient trajectories are not calculated or accessible



Healthcare

- More data is ‘digital first’ every day
- Decision aids are needed
- Good UX and responsive computing and analytics are critical for improving health outcomes

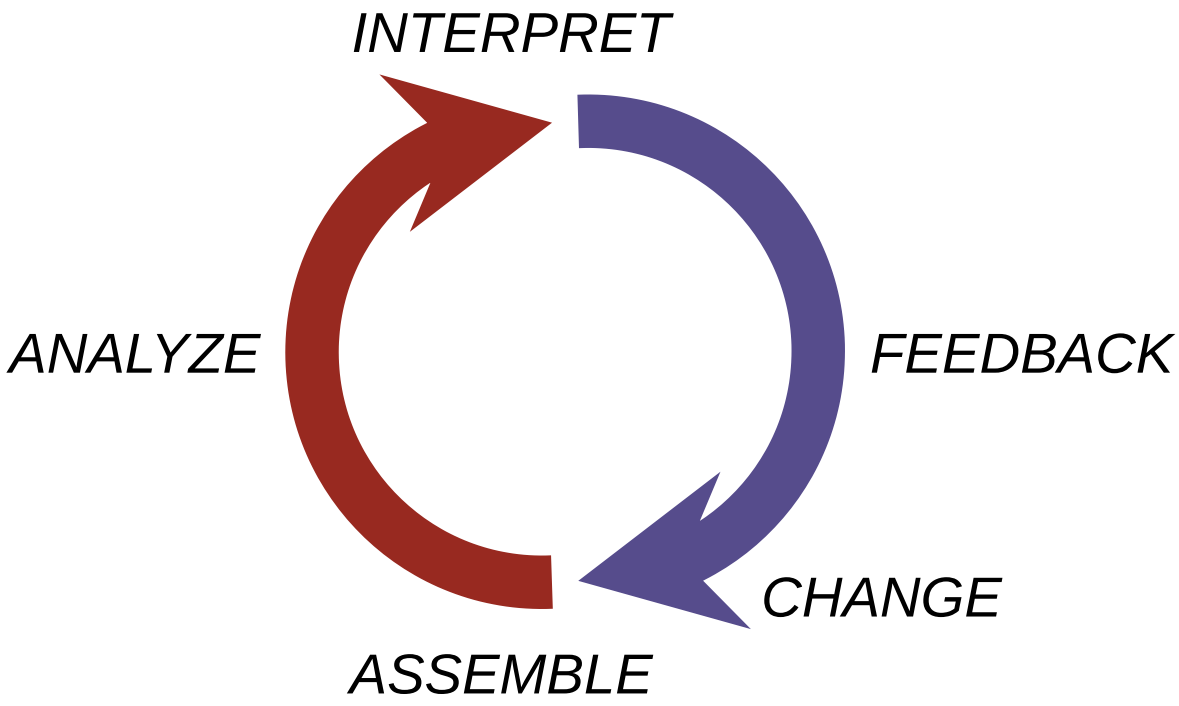


LEARNING HEALTH SYSTEMS

"Science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience."

—Institute of Medicine

LEARNING HEALTH SYSTEMS



“Science, informatics, incentives, and

View from 2006

'By the year 2020, ninety percent of clinical decisions will be supported by accurate, timely, and up-to-date clinical information, and will reflect the best available evidence."

Charter

IOM Roundtable on Value & Science-Driven Health Care

EHRs are now ubiquitous

'By the year 2020, ninety percent of clinical decisions will be supported by accurate, timely, and up-to-date clinical information, and will reflect the best available evidence."

Charter

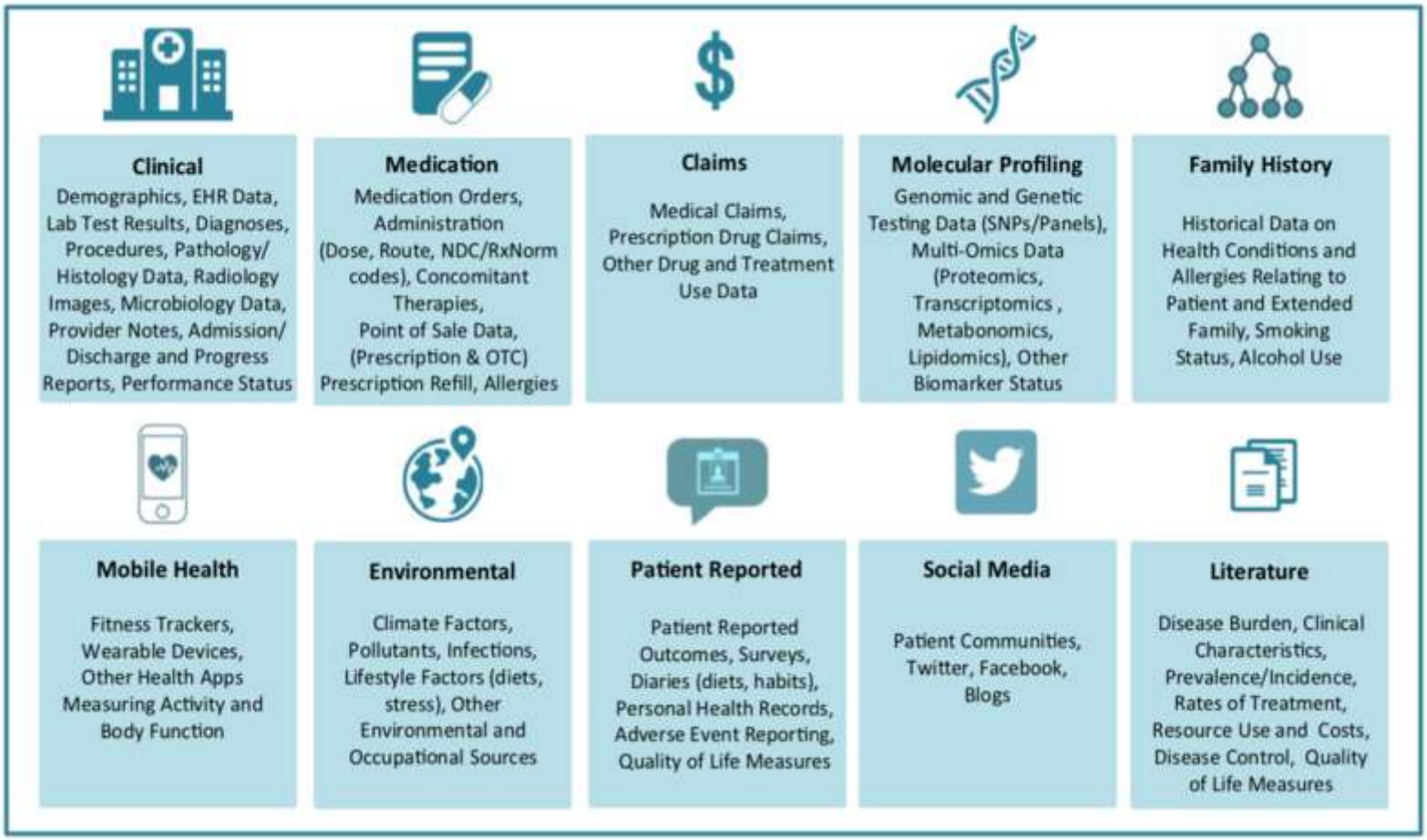
IOM Roundtable on Value & Science-Driven Health Care

But evidence-driven decision support remains a future vision

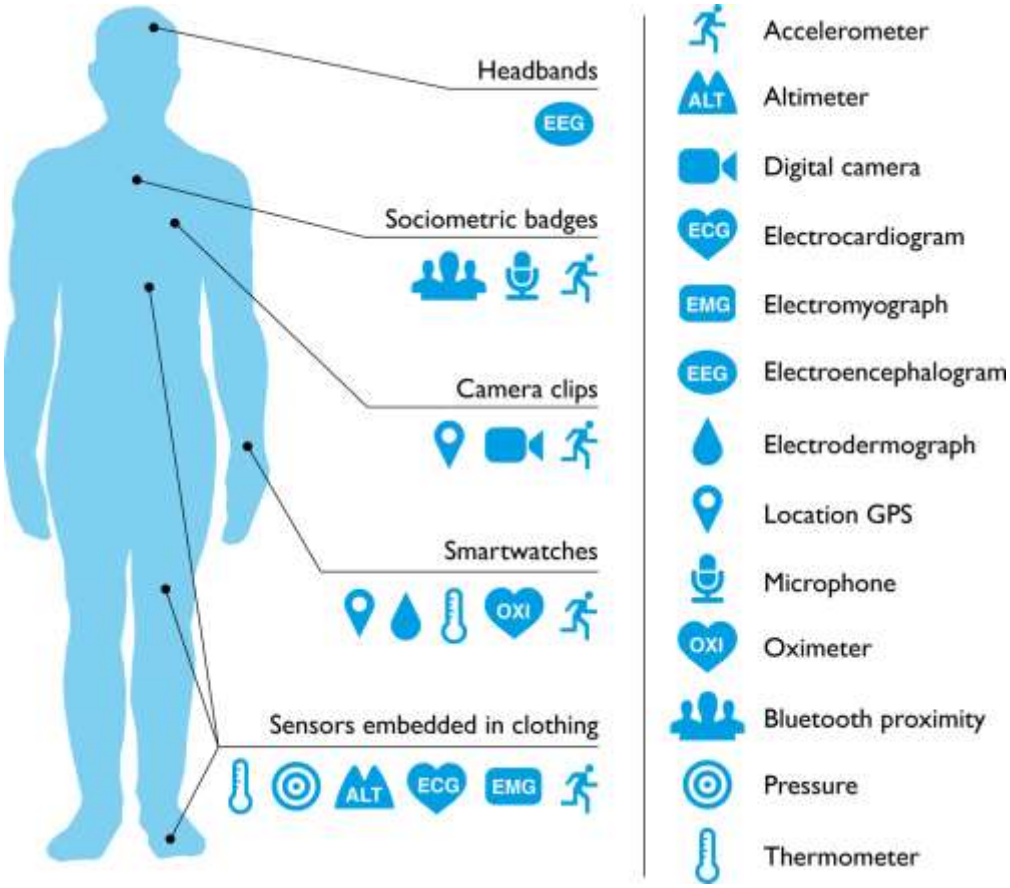
Real World Evidence

- Needs big data!
- Needs **population** representation
- Need **epidemiologists** and **statisticians** to understand the potential **biases** in representation
- **EHRs**, **NLP**, **Machine Learning** can power real world evidence learning
- Critical for a **Learning Health System**

Types of "Real-World Data"

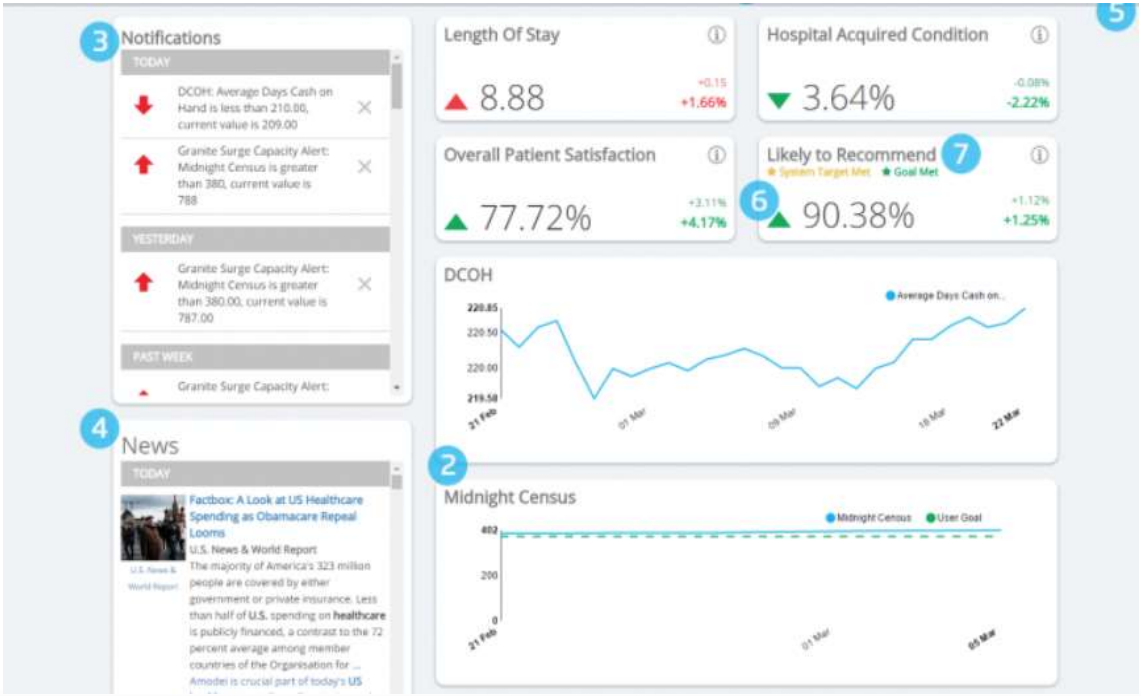


Mobile sensors to enhance monitoring of effects of new therapies

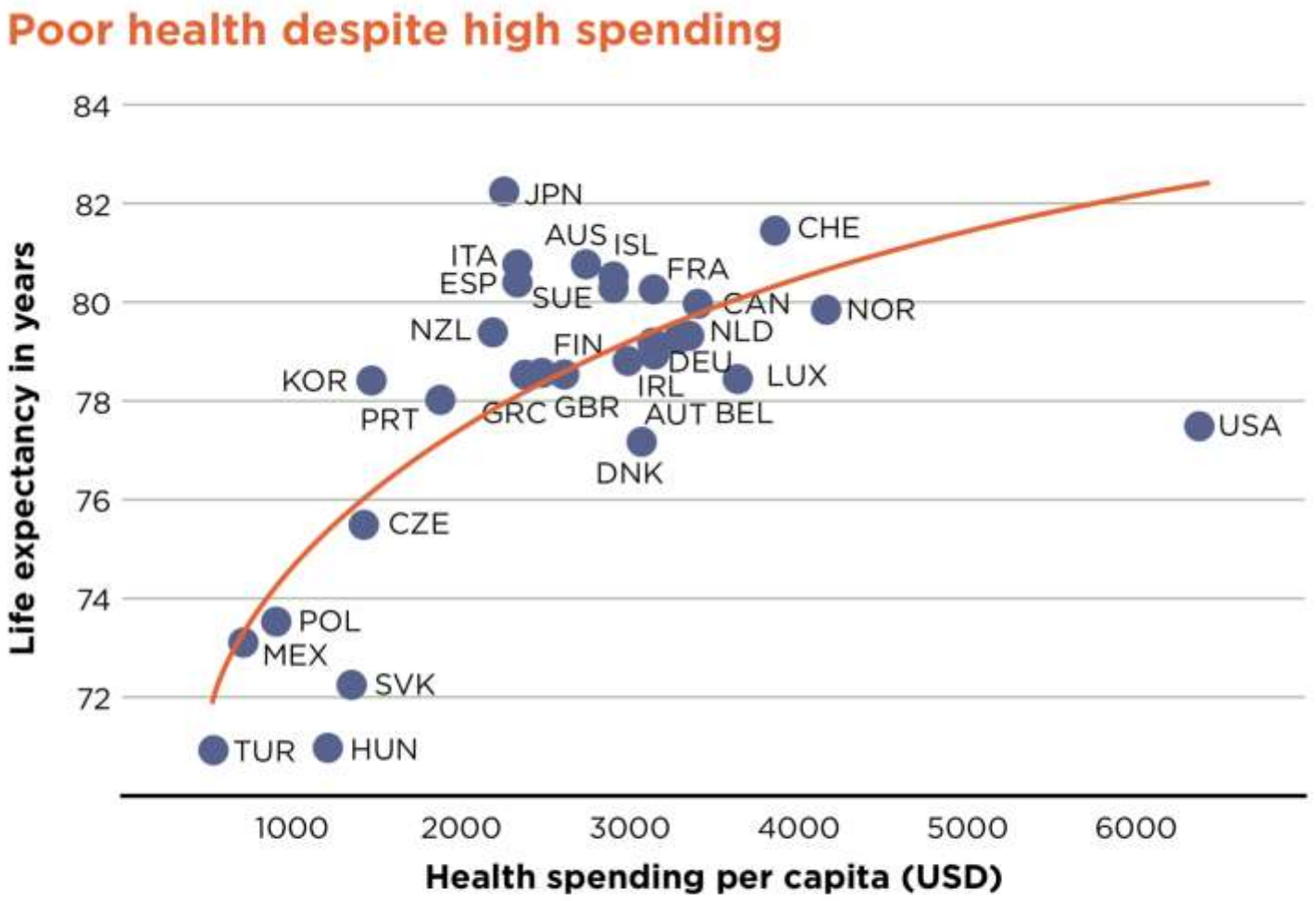


Emergence of “population health analytics” to measure quality

- Requires very similar infrastructure, tools, and data to outcomes/pragmatic research with RWD



Poor Health in spite of high expenditures



OECD Health Data, 2009. Life expectancy at birth in different countries versus per capita expenditures on health care in dollar terms, adjusted for purchasing power. The United States is a clear outlier on the curve, spending far more than any other country yet achieving less.

Team Science is critical

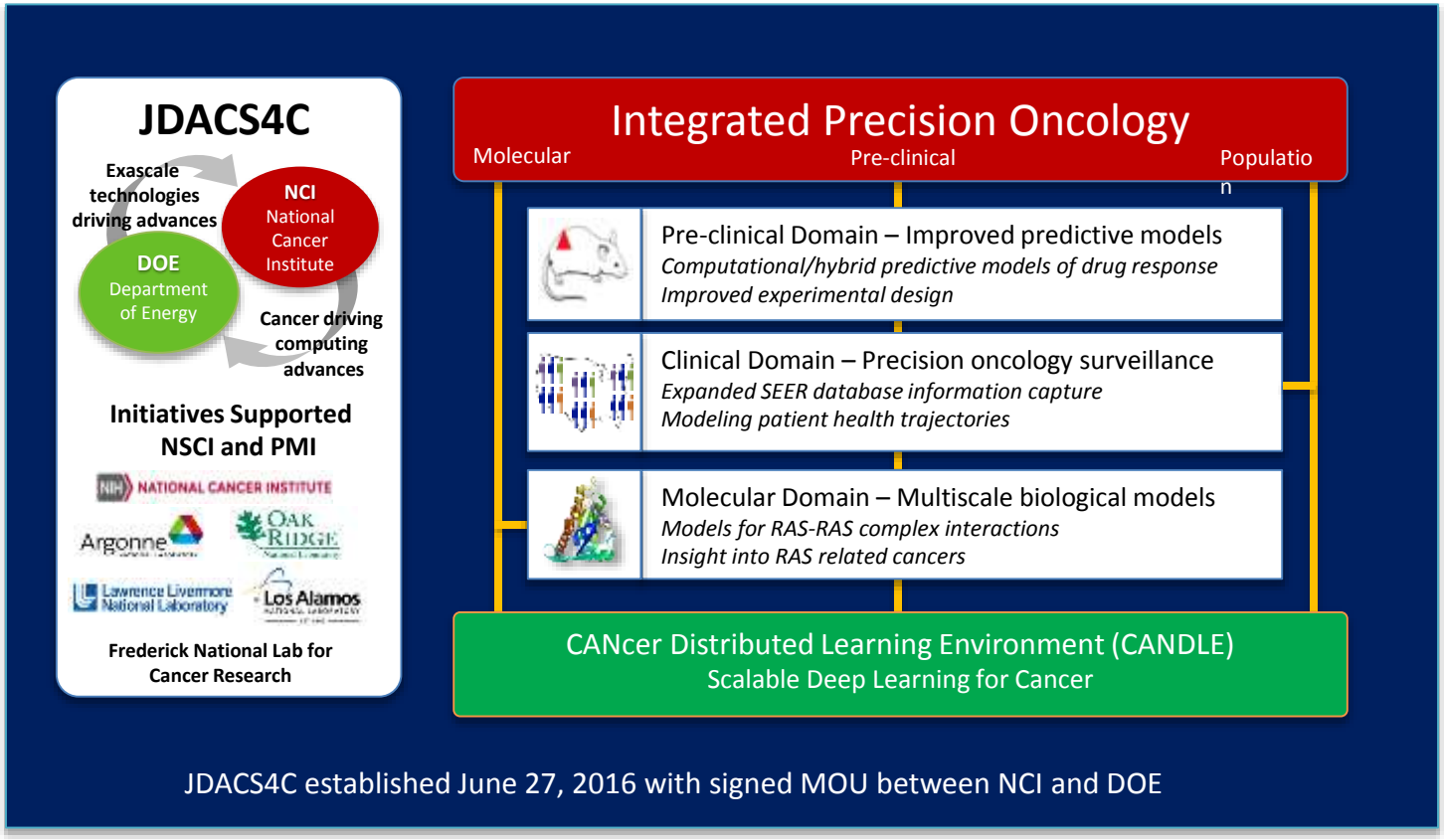
- Analytics and Visualization
- Bioinformatics
- Biostatists
- Clinical Care
- Clinical Research
- Clinical Trials
- Data Science, ML, BD, HPC
- EHRs, Imaging, Lab Systems

Open Data enhances collaboration and team science!

A brief history

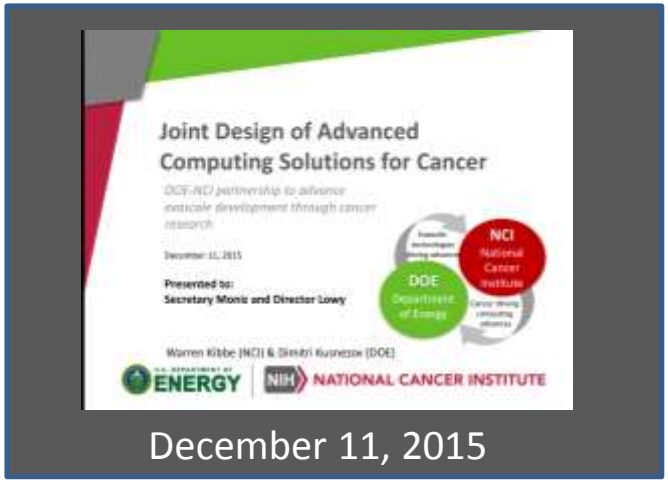
- **Eric Stahlberg** and **Jason Paragas** discussed the potential for a joint HPC-focused collaboration between NCI and DOE after SC14
- Winter – Spring of 2015 started fleshing out a potential set of pilots
- First public presentation to the NIH HPC Scientific Management workgroup in 8/2015
- Presented to FNLAC (NCI FACA) in 9/2015
- First Cancer Workshop at SC15
- Presented to NCI Director and DOE Secretary 12/2015

Joint Design of Advanced Computing Solutions for Cancer



Chronology of Significant Events

January 20, 2015



December 11, 2015

Executive Order, July 29, 2015

January 12, 2016

From the Frontiers of Predictive Oncology and Computing meeting, July 2016

National Strategic Computing Initiative (NSCI)

Executive Order, July 29, 2015

It is the policy of the United States to sustain and enhance its scientific, technological, and economic leadership position in HPC research, development, and deployment through a coordinated Federal strategy guided by four principles:

1) The United States must deploy and apply new HPC technologies broadly for economic competitiveness and scientific discovery.
2) The United States must foster public-private collaboration, relying on the respective strengths of government, industry, and academia to maximize the benefits of HPC.
3) The United States must adopt a "whole-of government" approach that draws upon the strengths of and seek cooperation among all Federal departments and agencies with significant expertise or equities in HPC in concert with industry.
4) The United States must develop a comprehensive technical and scientific approach to efficiently transition HPC research on hardware, system software, development tools, and applications into development and, ultimately, operations.

This order establishes the NSCI to implement this whole-of-government strategy, in collaboration with industry and academia, for HPC research, development, and deployment.

DOE is a Lead Agency for NSCI
NIH/NCI is a Broad Deployment Agency for NSCI

Precision Medicine Initiative (PMI)

January 30, 2015

Objectives of the Precision Medicine Initiative:

1. **More and better treatments for cancer** *NCI will accelerate the design and testing of effective, tailored treatments for cancer by expanding genetically based clinical cancer trials, exploring fundamental aspects of cancer biology, and establishing a national "cancer knowledge network" that will generate and share new knowledge to fuel scientific discovery and guide treatment* decisions.
2. Creation of a voluntary national research cohort
3. Commitment to protecting privacy
4. Regulatory modernization
5. **Public-private partnerships** *The Obama Administration will forge strong partnerships with existing research cohorts, patient groups, and the private sector to develop the infrastructure that will be needed to expand cancer genomics, and to launch a voluntary million-person cohort. The Administration will call on academic medical centers, researchers, foundations, privacy experts, medical ethicists, and medical product innovators to lay the foundation for this effort, including developing new approaches to patient participation and empowerment. The Administration will carefully consider and develop an approach to precision medicine, including appropriate regulatory frameworks, that ensures consumers have access to their own health data – and to the applications and services that can safely and accurately analyze it – so that in addition to treating disease, we can empower individuals and families to invest in and manage their health.*

NCI Precision Oncology – Extending the Frontiers

- 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.

DOE Exascale Computing – Extending the Frontiers

- Broaden CORAL functionality through co-design of highly scalable machine learning tools able to exploit node coherence.

- Explore how deep learning can define dynamic multi-scale validation, uncertainty quantification and optimally guide experiments and accelerate time-to-solution.

- Shape the design of architectures for exascale simultaneously optimized for big data, machine learning and large-scale simulation.

Cancer Moonshot

- **Precision Medicine Initiative (PMI)**
- **National Strategic Computing Initiative (NSCI)**

- **Making data available: Genomic Data Commons**
- **Using the cloud: NCI Cloud Pilots**
- **Computation and data: DOE-NCI Pilots**

- **Audacious yet possible!**
- **Investigate, explore, predict using real-world data!**

SuperComputing 2016

Cancer Moonshot is using MLDA

CANCER MOONSHOT

National Cancer Moonshot Summit
Washington, DC

June 29, 2016

CANCER MOONSHOT

IMPLEMENTATION

FUNDING OPPORTUNITIES

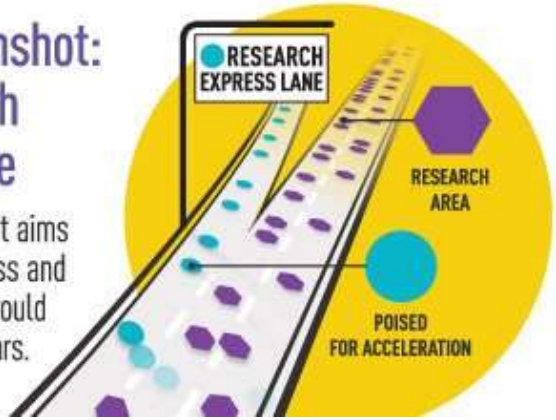
BLUE RIBBON PANEL RECOMMENDATIONS

CANCER MOONSHOT

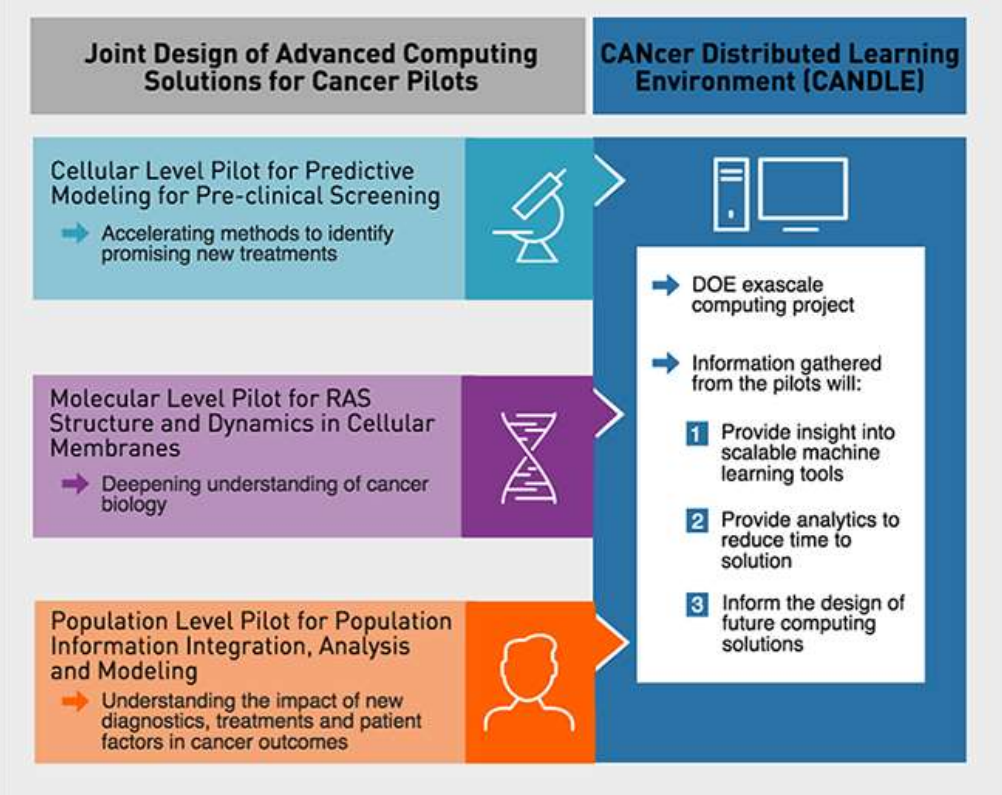
cancer.gov/brp

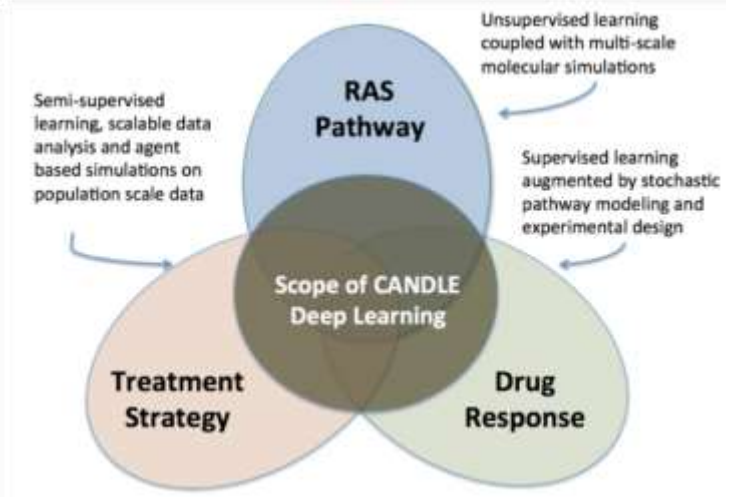
Cancer Moonshot: The Research Express Lane

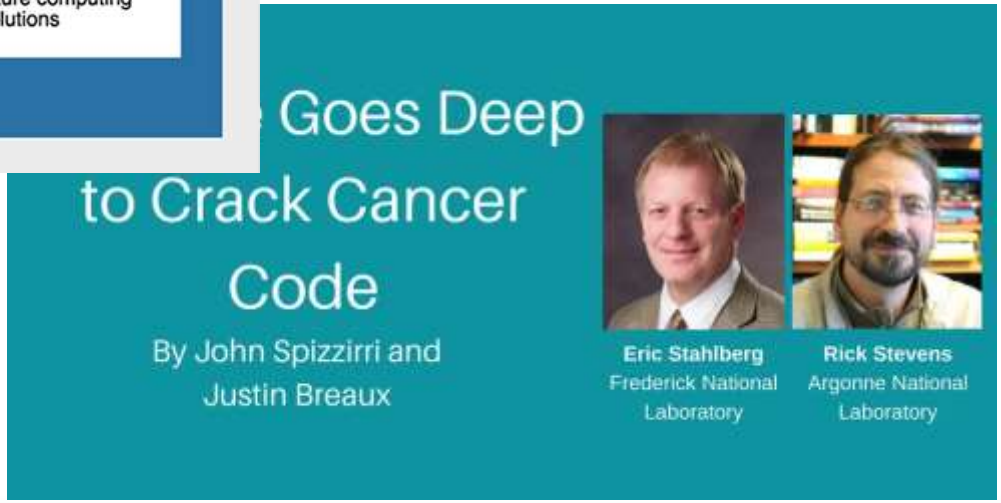
The Cancer Moonshot aims to accelerate progress and do in 5 years what would otherwise take 10 years.



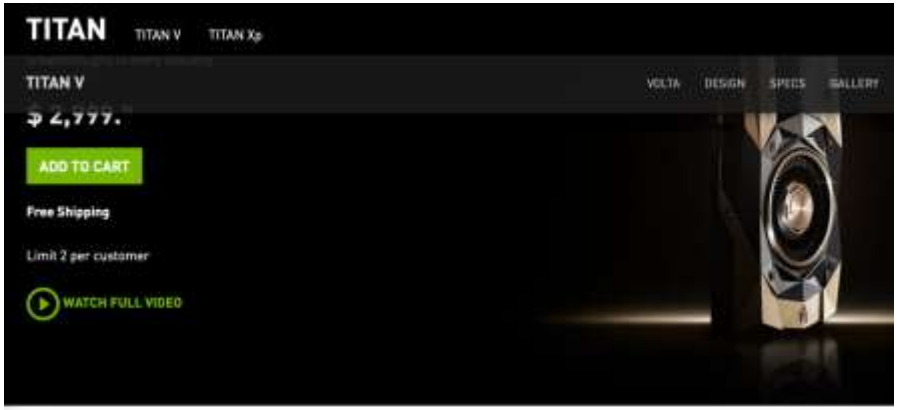
NCI – DOE pilots







Machine Learning and Data Analytics requires computing horsepower

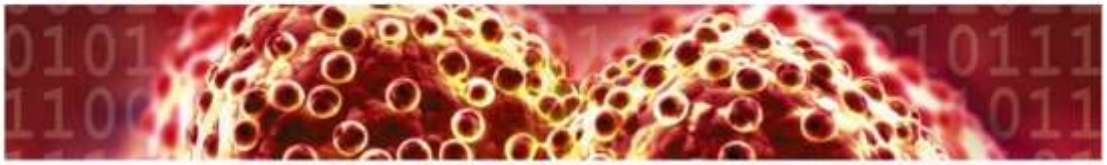


POWERED BY NVIDIA VOLTA—THE WORLD'S MOST ADVANCED GPU ARCHITECTURE

Volta GPU architecture pairs NVIDIA ® CUDA ® and Tensor Cores to deliver new levels of performance in a desktop PC GPU.

110	3D	21
DEEP LEARNING TeraFLOPS	STACKED MEMORY	BILLION TRANSISTORS

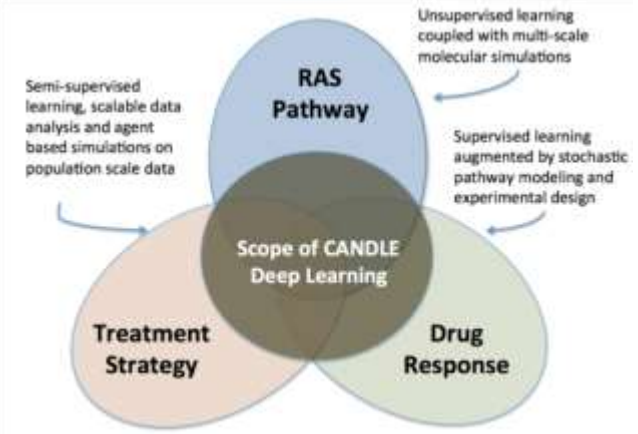
CANDLE



CANDLE

Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer

The nation has recently embarked on an all government approach to the problem of cancer, codified in the 'Cancer Moonshot' initiative of the Obama administration led by Vice President Biden. Cancer is an extremely complex disease, which disrupts basic biological processes at a fundamental level, leading to renegade cells threatening the health of an individual. To accelerate the capabilities needed to realize the promise envisioned for the Cancer Moonshot and to establish a new paradigm for cancer research for years to come, the Department of Energy (DOE) entered into a partnership with the National Cancer Institute (NCI) of the National Institutes of Health (NIH). This partnership identified three key challenges that the combined resources of DOE and NCI can accelerate: to provide better understanding of the disease, to make effective use of the ever-growing volumes and diversity of cancer related data to build predictive models, and, ultimately, to provide guidance and support decisions on anticipated effective treatments for individual patients.



ANALYTICS IN ACTION NEWS

VA, DOE Launch Healthcare Big Data, Machine Learning Project

Researchers will combine the VA's vast healthcare big data assets with the Department of Energy's machine learning prowess to improve diagnosis and treatment of certain diseases.

Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

DOE-NCI partnership to advance exascale development through cancer research

March 28, 2017

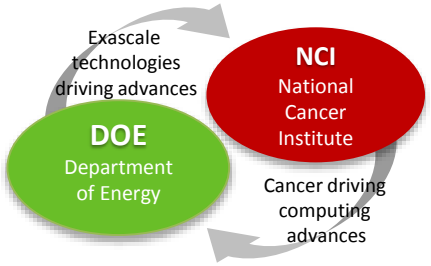
Presented to:
DOE-NCI Governance Review Committee

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

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. LLNL-MI-727782.

JDACS4C – NCI-DOE Collaboration Overview

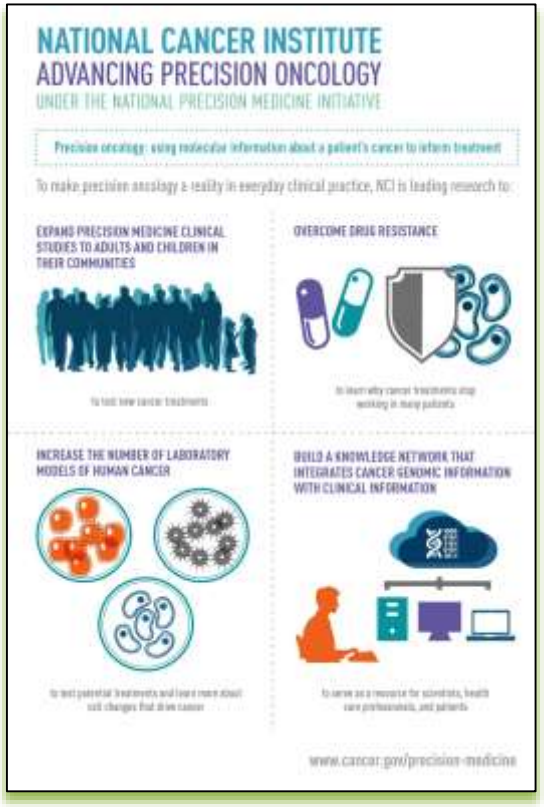
- **Shared Interests**
  - Cancer scientific challenges driving advances in computing
  - Exascale technologies driving cancer advances
- **Supports Two Primary Executive Office Initiatives**
  - Precision Medicine Initiative (Jan 2015)
  - National Strategic Computing Initiative (July 2015)
- **Three 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

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



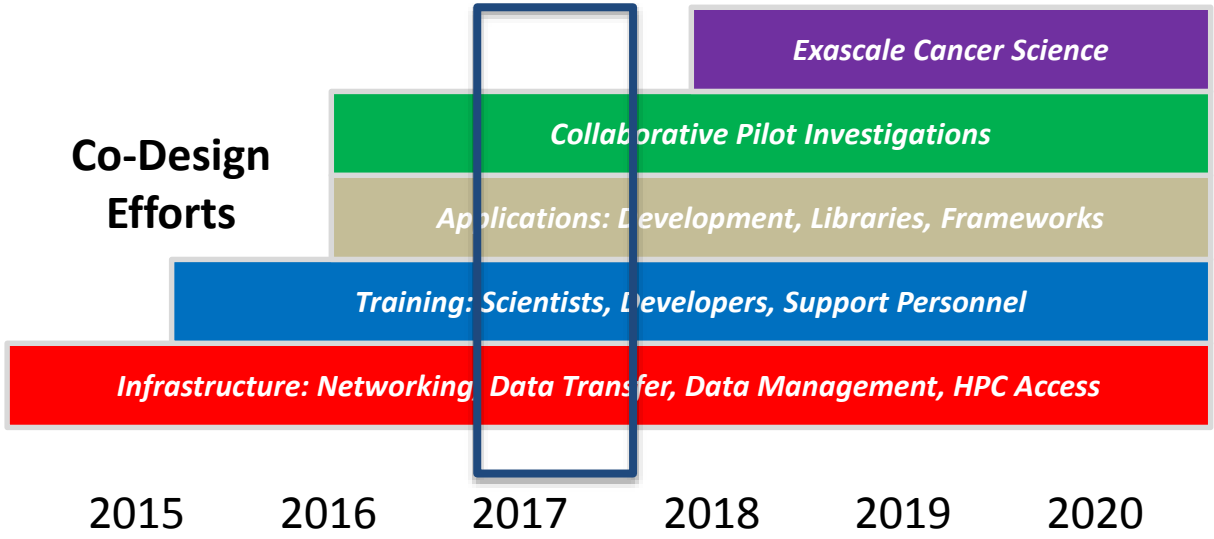
Pilot 2:
Biological Models
Multi-scale computational biological models

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

Accelerating Computational and Data-driven Cancer Research

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



Crosscut: Integrated Precision and Predictive 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: Uncertainty Quantification (UQ) and CANDLE exascale technologies

NCI - DOE Pilots

Crosscut: Integrated Precision and Predictive 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: Uncertainty Quantification (UQ) and CANDLE exascale technologies

Eric Stahlberg, PhD

NCI RAS Initiative + NCI-DOE Joint Initiative

Oncogenic KRAS is responsible for many human cancers

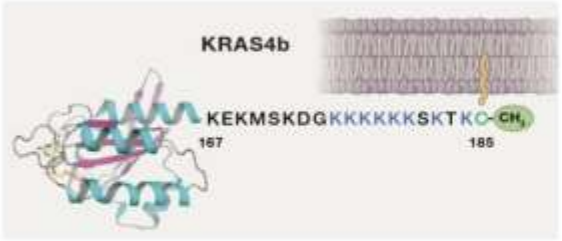
93% of all pancreatic

42% of all colorectal

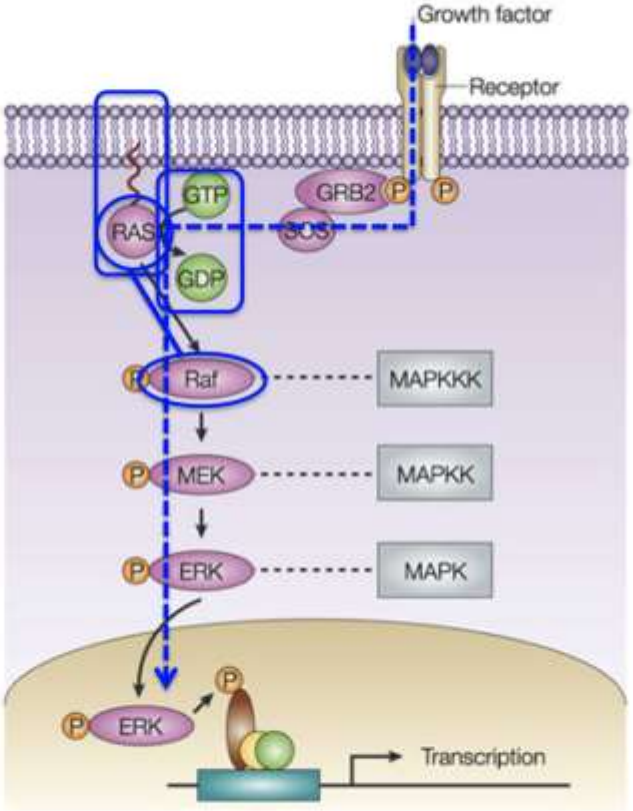
33% of all lung cancers

1 million deaths/year world-wide

No effective inhibitors



Simanshu,Cell 170, 2017



Nature Reviews | Molecular Cell Biology

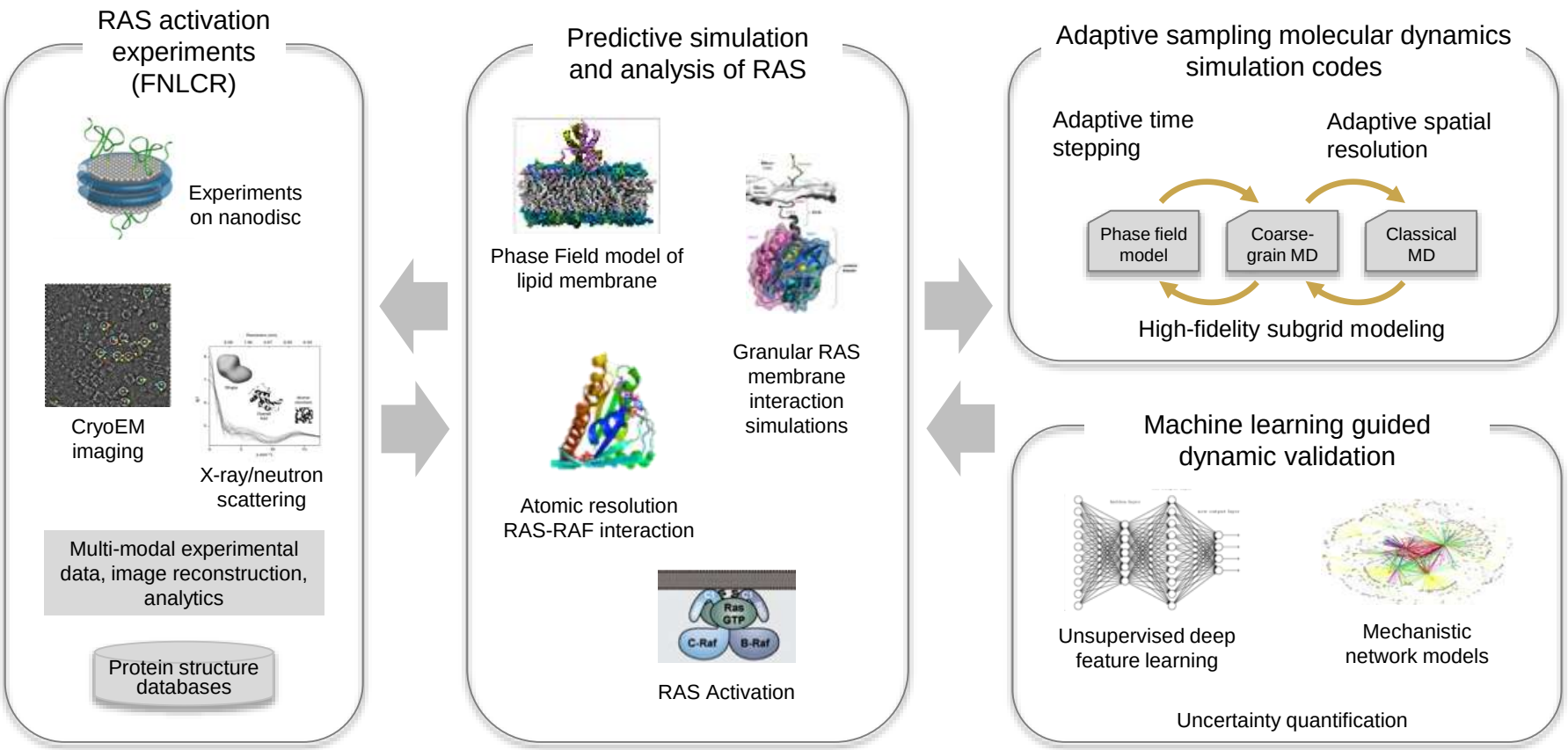
Pathway transmits signals

RAS is a switch
oncogenic RAS is "on"

RAS localizes to the plasma membrane

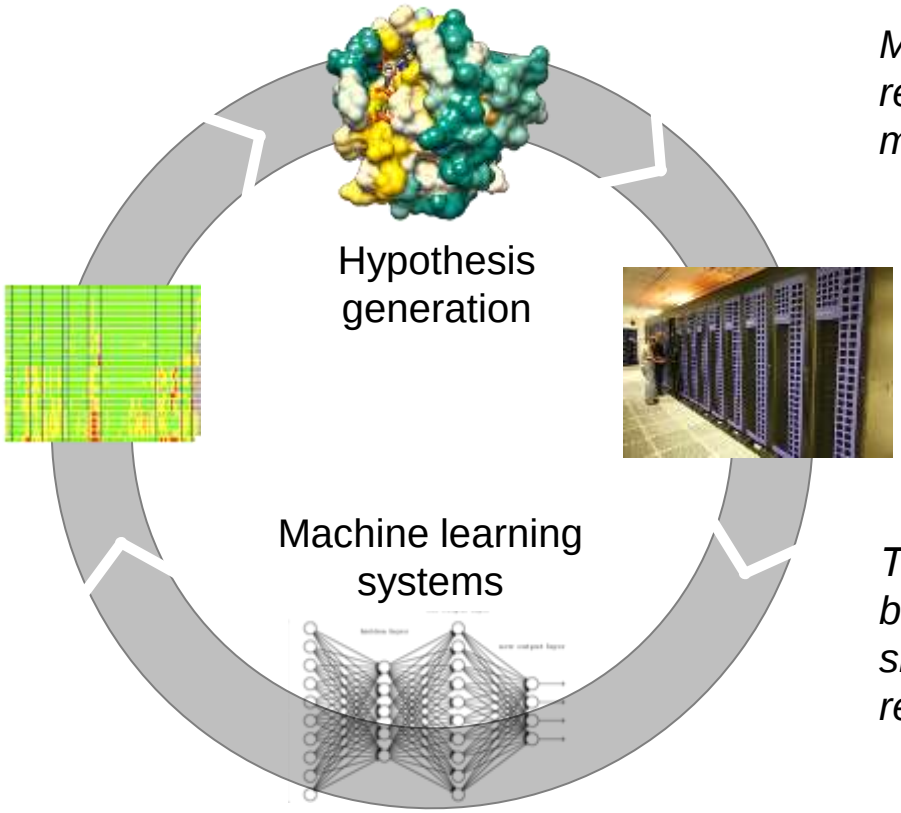
RAS binds effectors (RAF) to activate growth

Cancer Moonshot Pilot 2



Dr. Fred Streitz, Lawrence Livermore National Lab

Machine learning enables a new dynamic validation approach to high-fidelity simulation

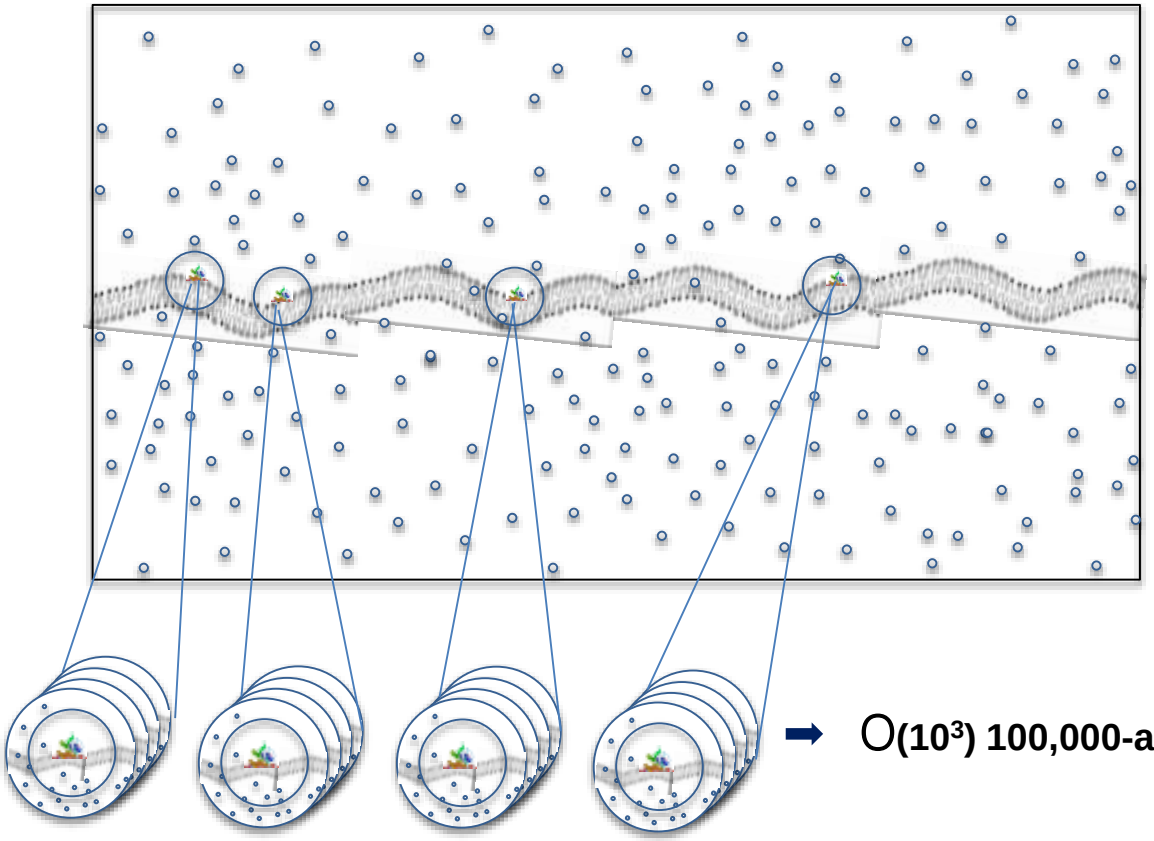


Machine learning systems uncover relationships between computational models and experimental data

This "active learning" approach can be used to optimize solutions with significant reduction in compute requirements

Dr. Fred Streitz, Lawrence Livermore National Lab

Simulation of full system will incorporate a large number of smaller simulations



10-100 μm lipid patches
Dynamic membrane
Hundreds of Ras proteins
- Mutant and wild-type
- Many conformations
- Many environments

Investigate diffusion and aggregation in of Ras in context of specific membrane properties

➡ **O(10^3) 100,000-atom simulations**

Dr. Fred Streitz, Lawrence Livermore National Lab

Multi-disciplinary team from FNLCR, LLNL, LANL, ORNL and ANL

FNLCR / NCI Team: Frantz Jean-Francois, **Frank McCormick**, Dhirendra Simanshu, Eric Stahlberg, **Andy Stephen**, Tommy Turbyville, Debanjan Goswami

Argonne National Lab: *Prasanna Balaprakash,* ***Tom Brettin****, FangFang Xia*

Oak Ridge National Lab: , Pratul K. Agarwal, Debsindhu Bhowmik, ***Arvind Ramanathan****,* Blake A. Wilson, Christopher B. Stanley

Los Alamos National Lab: Angel Garcia, Christoph Jungans, Cesar Lopez, Chris Neale, Danny Perez, Sandrasegaram Gnanakaran, Tim Travers, Art Voter

Lawrence Livermore National Lab: Harsh Bhatia, Barry Belmont, Tim Carpenter, Francesco Di Natale, **Jim Glosli**, Helgi Ingolfsson, Piyush Karande, **Felice Lightstone**, Tomas Oppelstrup, Liam Stanton, Michael Surh, *Sachin Talathi, Brian Van Essen,* Yue Yang, Xiaohua Zhang

Dr. Fred Streitz, Lawrence Livermore National Lab

NCI-DOE collaboration: CANDLE

- **CAN**cer **D**istributed **L**earning **E**nvironment
- Expanding collaboration between NCI and DOE
- CANDLE's goal is to use deep/machine learning to accelerate cancer research

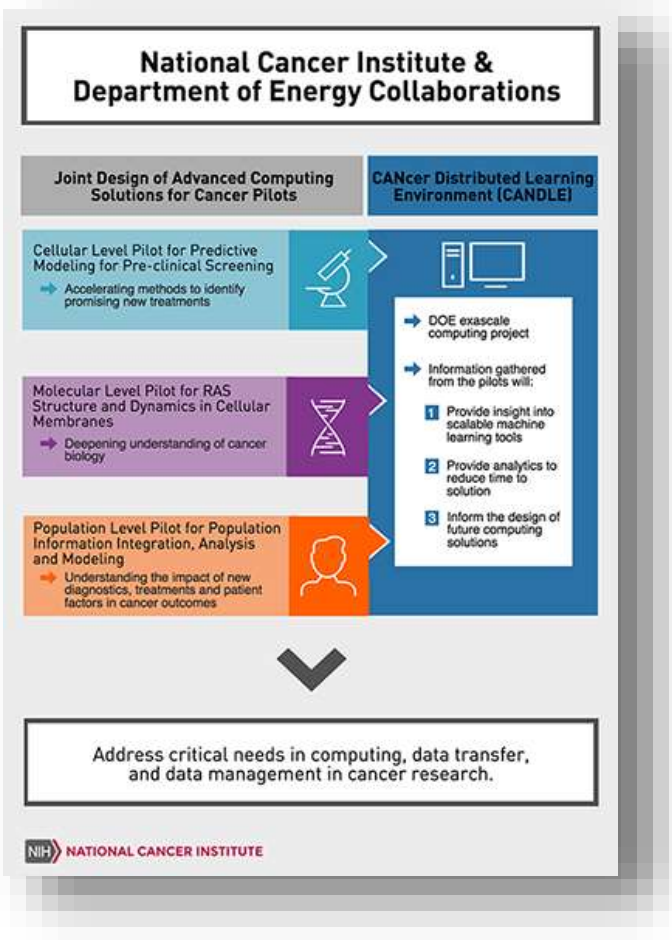
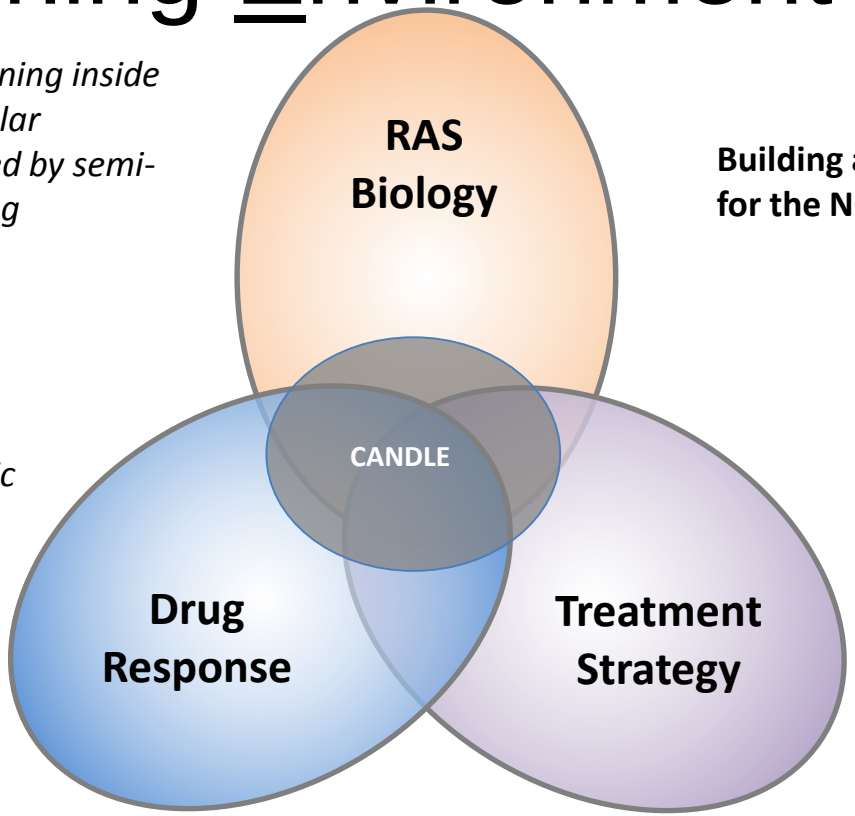


image credit: www.globus.org

CANDLE: Cancer Distributed Learning Environment

Unsupervised learning inside multiscale molecular simulations steered by semi-supervised learning



Building a machine learning framework for the NCI-DOE pilots

Supervised learning augmented by stochastic pathway modeling and experimental design

Semi-supervised learning, scalable data analysis and agent based simulations on population scale data

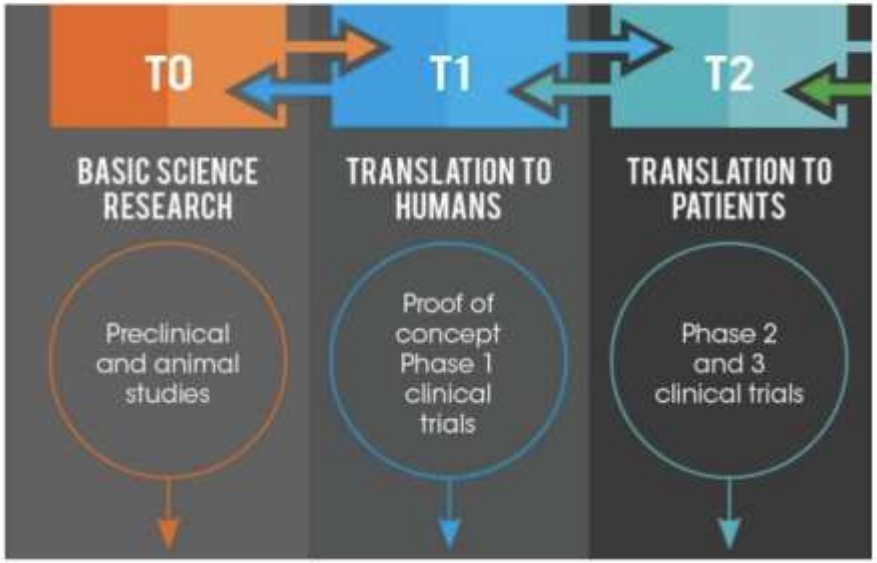
Cancer Moonshot Data & Technology Team

Co-Chairs: **Dimitri Kusnezov (DOE), DJ Patil (OSTP), and Jerry Lee (OVP)**

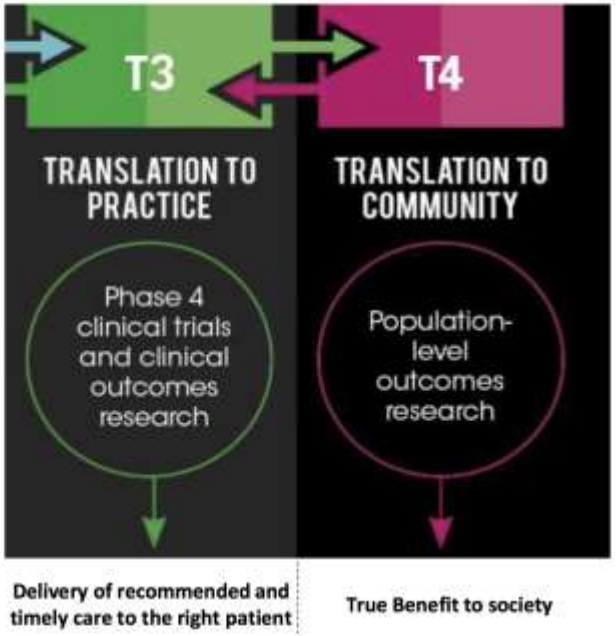
Members:

- John Scott **(DoD)**
- Craig Shriver **(DoD)**
- Cheryll Thomas **(CDC)**
- Frances Babcock **(CDC)**
- Teeb Al-Samarrai **(DOE)**
- Sean Khozin **(FDA)**
- Alexandra Pelletier **(PIF)**
- Maya Mechenbier **(OMB)**
- Henry Rodriguez **(NCI)**
- Karen Cone **(NSF)**
- Michael Kelley **(VA)**
- Louis Fiore **(VA)**
- Warren Kibbe **(NCI)**
- Betsy Hsu **(NCI)**
- Niall Brennan **(CMS)**
- Thomas Beach **(USPTO)**
- Claudia Williams **(OSTP)**
- Vikrum Aiyer **(USPTO)**
- Tom Kalil **(OSTP)**
- Kathy Hudson **(NIH)**
- Dina Paltoo **(NIH)**
- Al Bonnema **(DoD)**
- Michael Balint **(PIF)**
- Kara DeFrias **(OVP)**
- Greg Pappas **(FDA)**
- Erin Szulman **(OSTP)**
- Paula Jacobs **(NCI)**

Translational from basic science to human studies



Translational of new data into the clinic and health decision making



NCI CPTAC Program

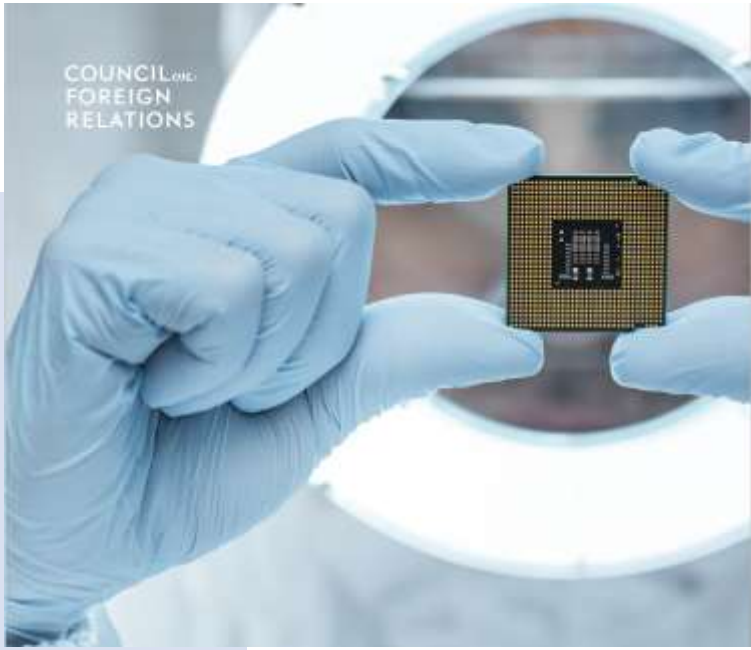
US BloodPAC

CMS OCM

DOE/VA MVP-CHAMPION & NCI/VA BD-STEP

NCI/DoD/VA APOLLO

Is the US able to keep up?



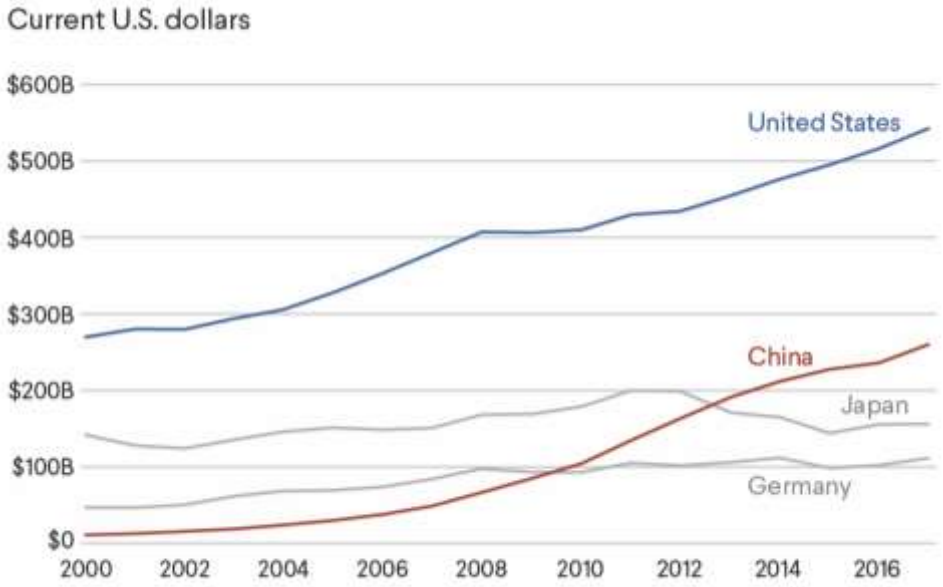
Independent Task Force Report No. 77

Innovation and National Security

Keeping Our Edge

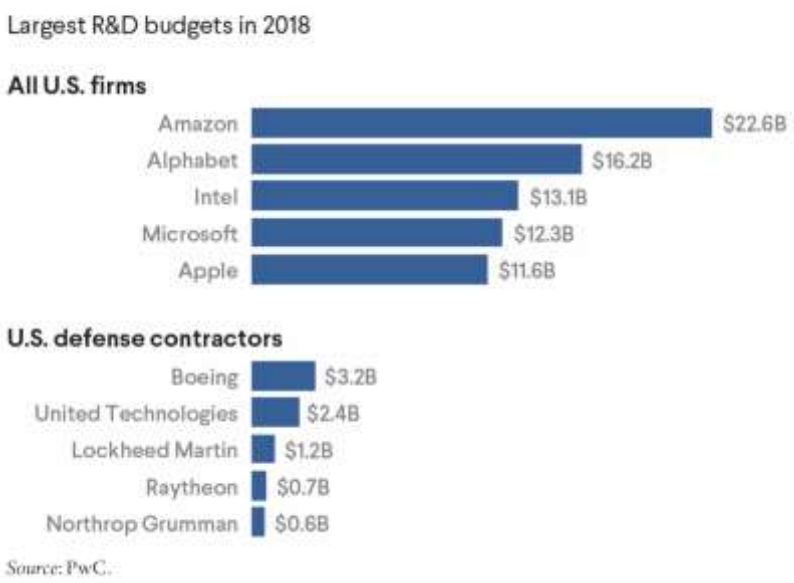
James Manyika and William H. McRaven, *Chairs*
Adam Segal, *Project Director*

R&D By Country



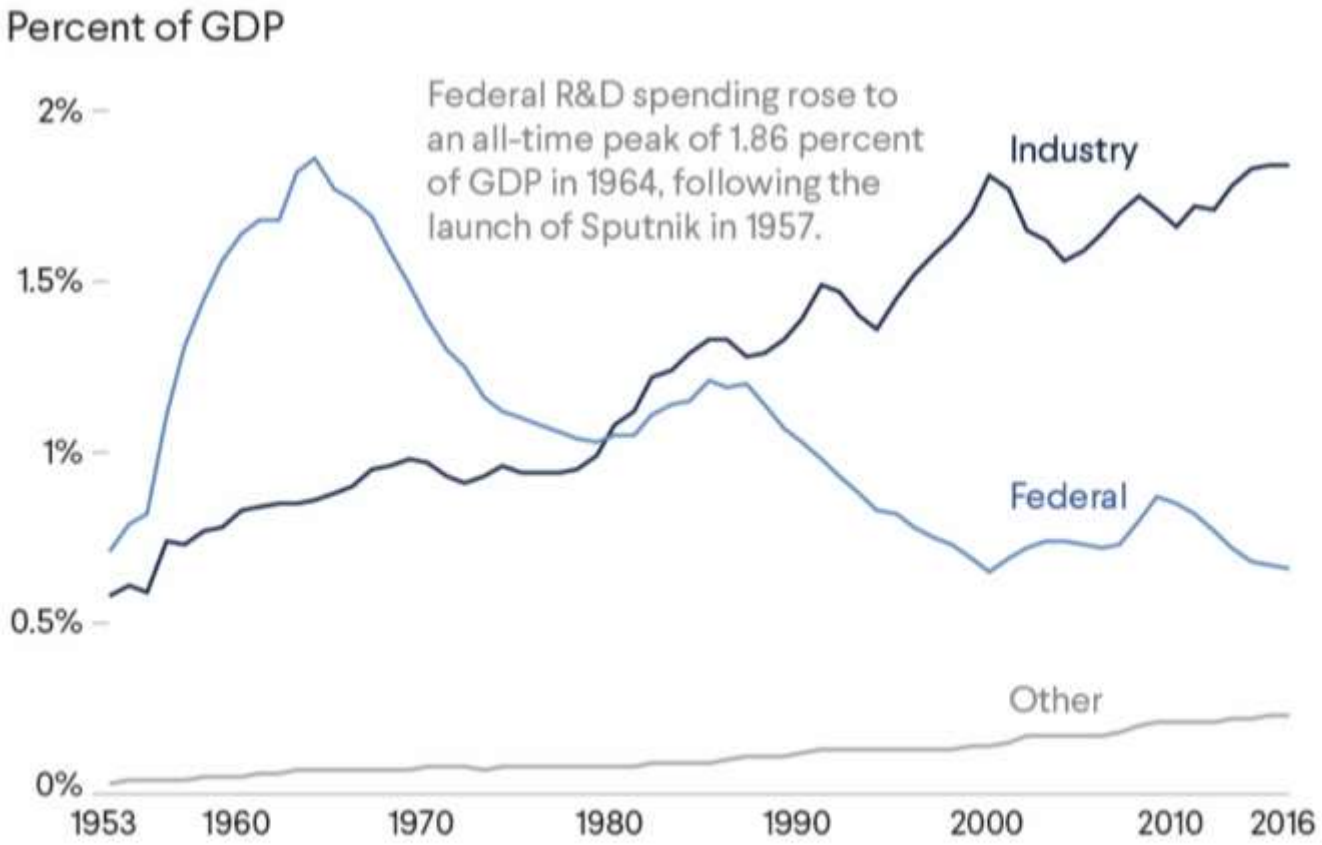
Source: Organization for Economic Cooperation and Development.

Figure 6. U.S. TECH COMPANIES OUTSPEND DEFENSE CONTRACTORS ON R&D



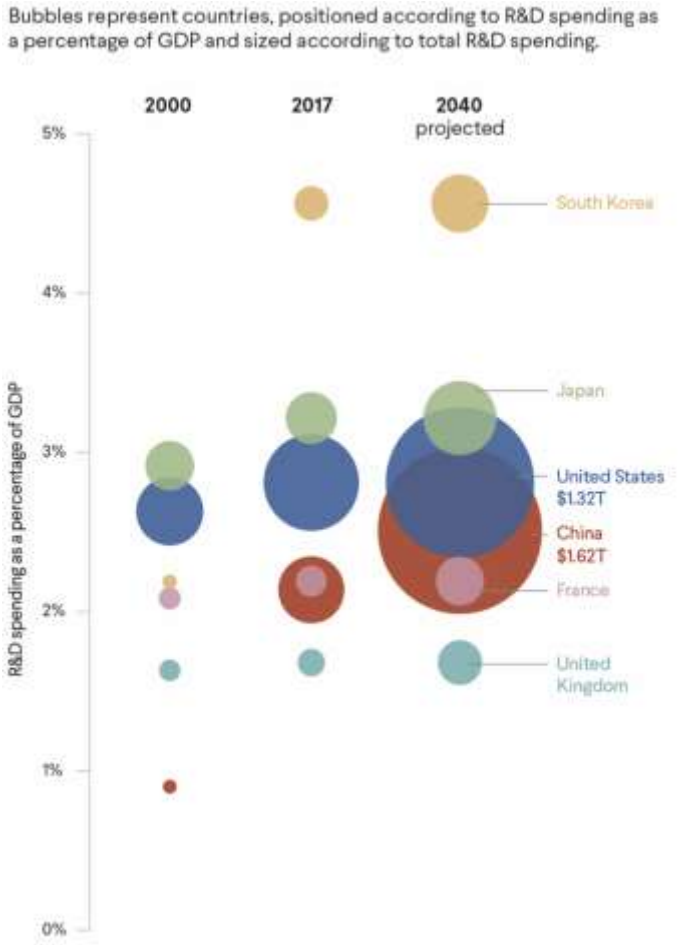
Source: PwC.

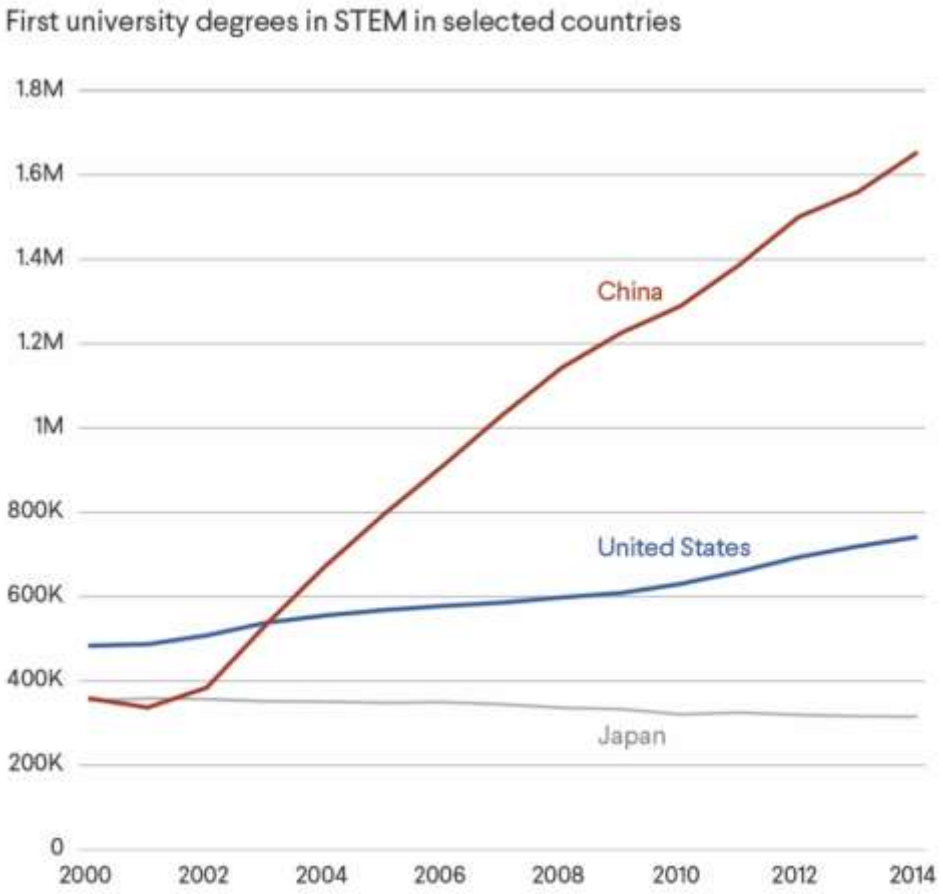
US R&D Funding as share of GDP



Source: National Science Foundation.

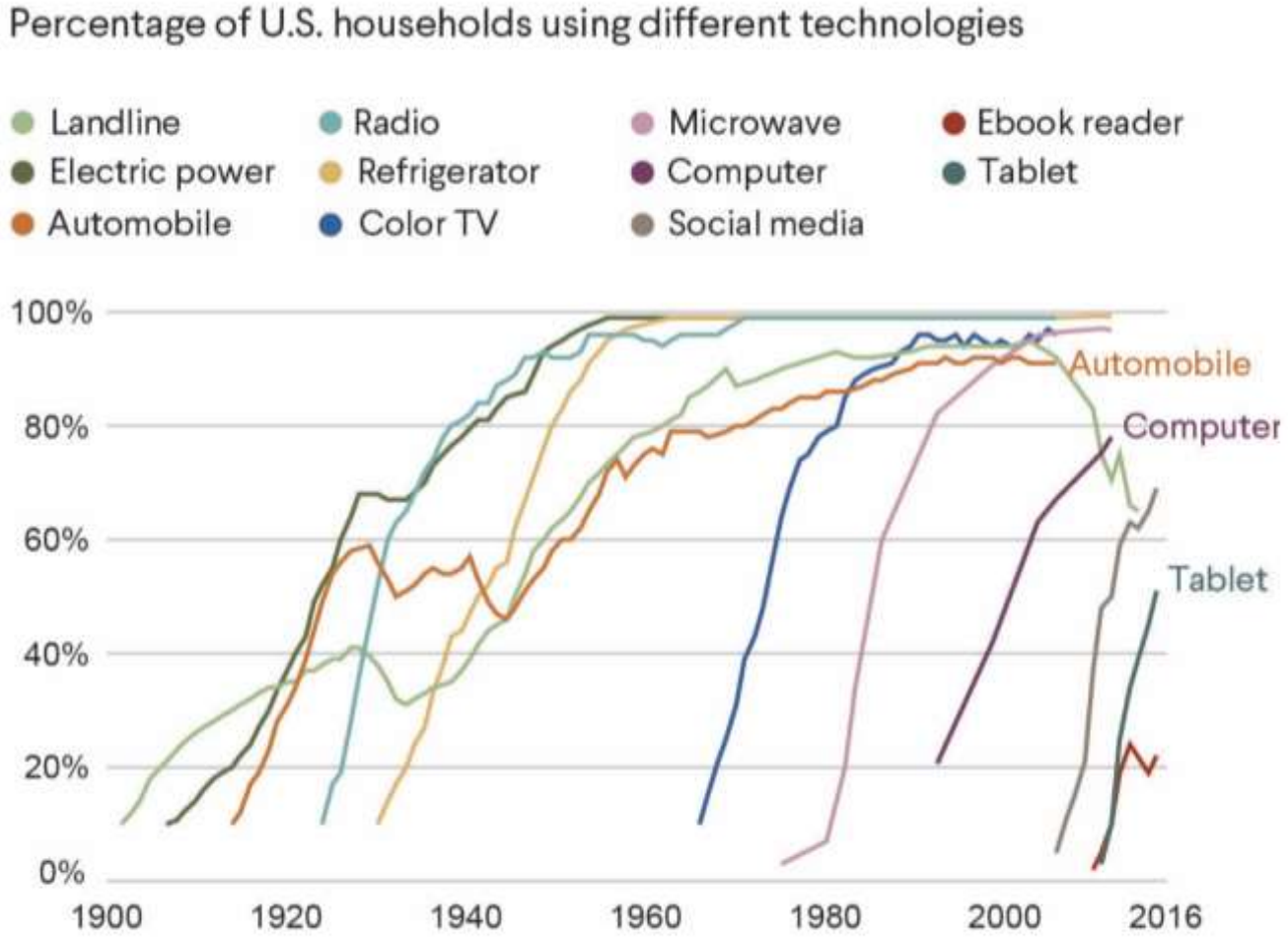
R&D spending / STEM





Source: National Science Board.

Pace of Technology Adoption



Source: Our World in Data.

Fundamental Changes

- Data generation is not the bottleneck
- Most data are now **'digital first'**
- Old statistical models assuming variable independence are inadequate – **systems and pathways are not independent!**
- **Project management** is critical in scaling population science

Well-defined experiments are still key

Changes in the Society

- **Perceptions of privacy**
- **GA4GH – right to benefit from research**
- **Social media**
- **Familiarity with IT**
- **Tobacco control**
- **Health disparities**
- **Affordable Care Act**

Cancer Moonshot

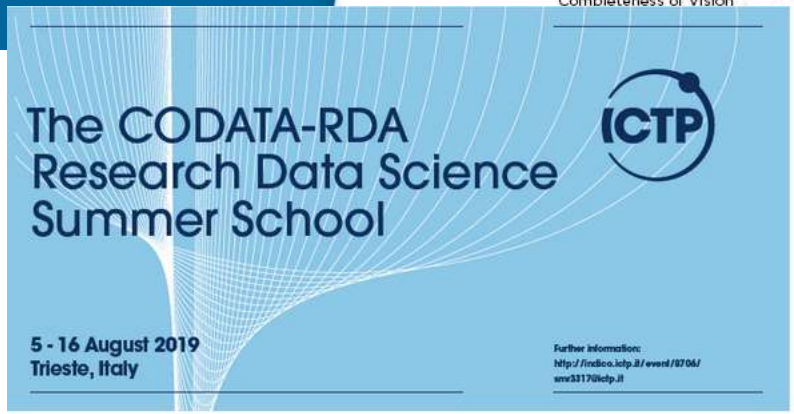
"…it is of critical national importance that we …**double the *rate* of progress** in the **fight against cancer**- and ***put ourselves on a path*** to achieve in just **5 years** research and treatment **gains** that **otherwise might take a decade** or more…"

WHAT IS “DATA SCIENCE”?

Knowledge among computer scientists about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited. A merger of the knowledge bases would produce a powerful force for innovation.

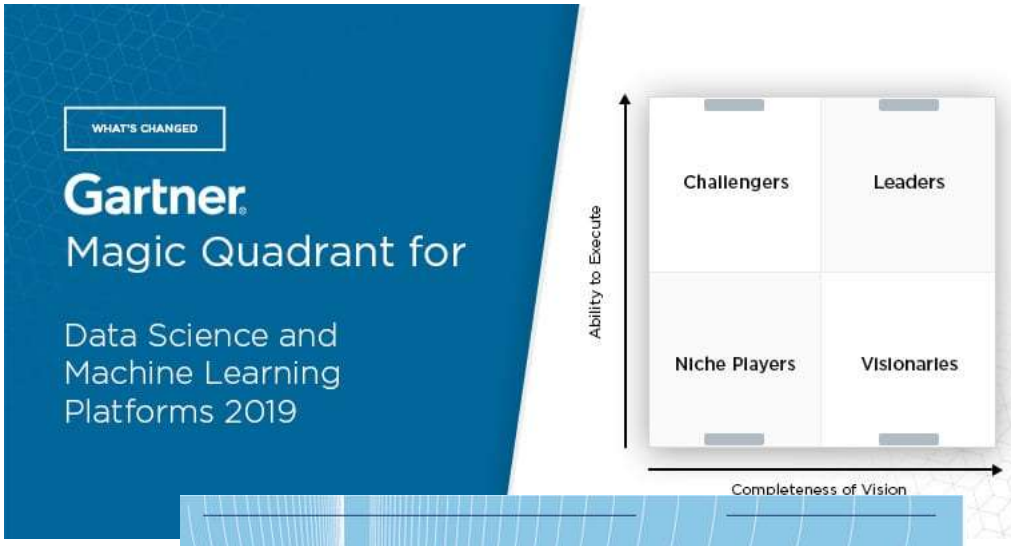
—Bill Cleveland on Data Science (2001)

Data Science Hype



Data is the new oil - NOT

Data Science Hype



Using data

increases its value

Cancer Genomics

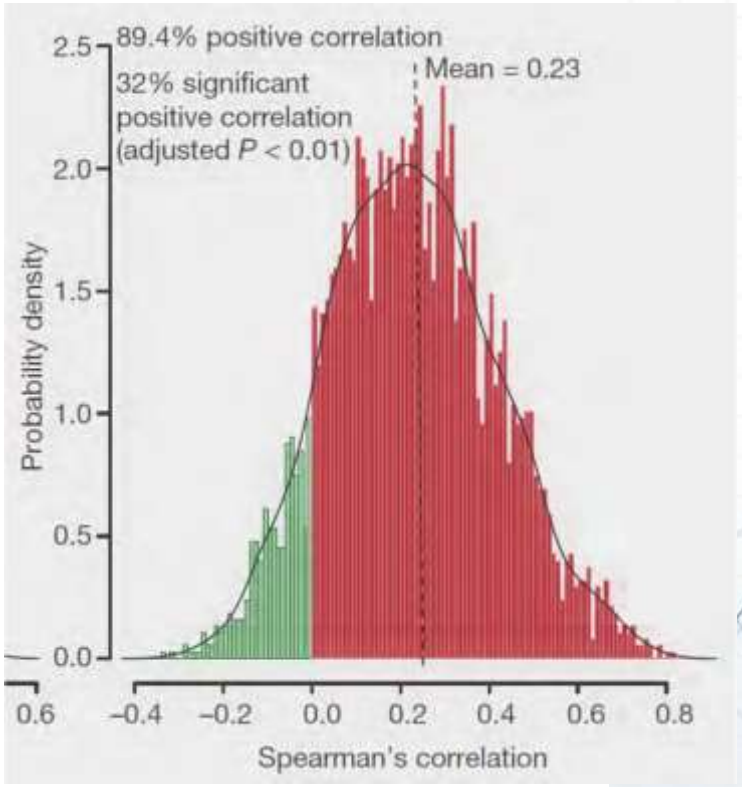
- Several distinct molecular forms of cancer at each organ site
- The genomic abnormalities of each cancer are unique
- The same molecular abnormalities are found in cancers that arise in different organs

Our understanding of biology, cancer, and intervention is changing!

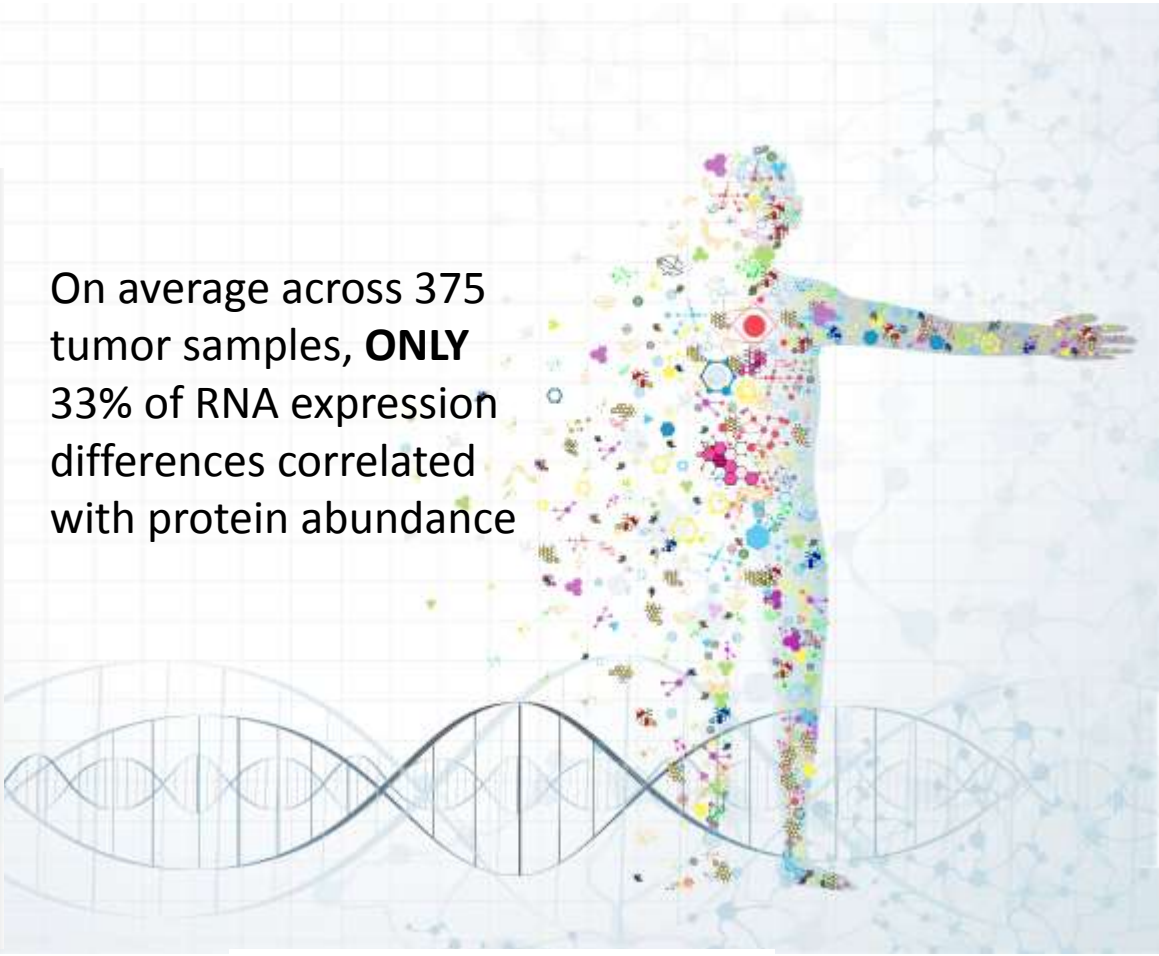
Genomics, **Data resources**, and **clinical trials** are changing

- NCI Cancer Genomic Data Commons
- Umbrella and basket clinical trials such as NCI MATCH and Pediatric MATCH

Keeping in mind cellular dynamics



On average across 375 tumor samples, **ONLY** 33% of RNA expression differences correlated with protein abundance



Zhang B et al, Proteogenomic characterization of human colon and rectal cancer, Nature, 2014, July 20.

Genomics is the beginning of precision medicine, not the end!

Questions?

Warren Kibbe, Ph.D.

warren.kibbe@duke.edu

@wakibbe