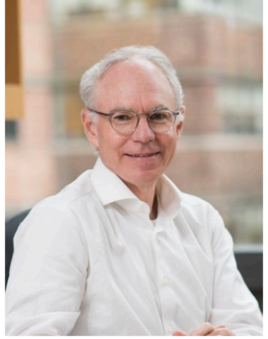
Data Science Opportunities for the NCI

Final Report

Mia Levy and Charles Sawyers
on behalf of the National Cancer Advisory Board
Ad hoc Working Group on Data Science

Joint Meeting of the Board of Scientific Advisors and National Cancer Advisory Board
June 10, 2019

Mia Levy
(co-chair)
Rush



Charles Sawyers
(co-chair)
MSKCC

Brian Alexander
Foundation Medicine

Regina Barzilay
MIT

John Carpten
USC

Amanda Haddock
Dragon Master
Foundation

George Hripcsak
Columbia

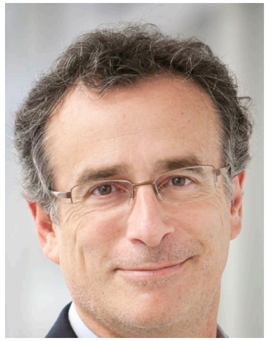
Mimi Huizinga
Novartis

Rebecca Jacobson
UPMC

Warren Kibbe
Duke

Michelle Le Beau
U Chicago

Anne-Marie Meyer
Roche



Vince Miller
(former member)
Foundation Medicine

Sylvia Plevritis
Stanford

Kim Sabelko
Komen

Sohrab Shah
(ad hoc member)
MSKCC

Lincoln Stein
OICR

Nick Wagle
Dana-Farber

Interim Recommendation Areas

1. Investments to leapfrog data sharing for high-value datasets
2. Harmonization of terminology between cancer research data and clinical care data
3. Support of data science training at the graduate level
4. Opportunities for funding challenges and prizes

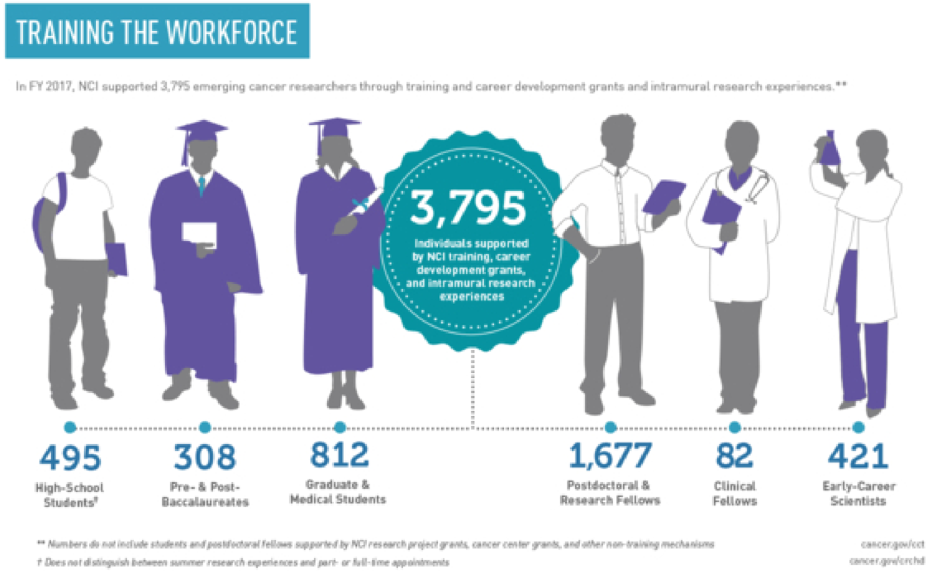
Accepted by NCAB August 2018

Additional Recommendation Areas

1. Additional areas of support for data science training and workforce development
2. Building machine learning infrastructure for cancer research
3. Facilitating the appropriate use of real-world data
4. Enabling the cultural shift toward data sharing

Initial Recommendation on Data Science Training: Increase the number of training programs and trainees in cancer data science

- Dedicate a specific T32 training program in cancer data science
- Contribute to existing NIH training programs
    - NLM T15 training programs
    - NIGMS Medical Scientist Training program
- Develop a short-term training program for clinicians and biological scientists


TRAINING THE WORKFORCE

In FY 2017, NCI supported 3,795 emerging cancer researchers through training and career development grants and intramural research experiences.**

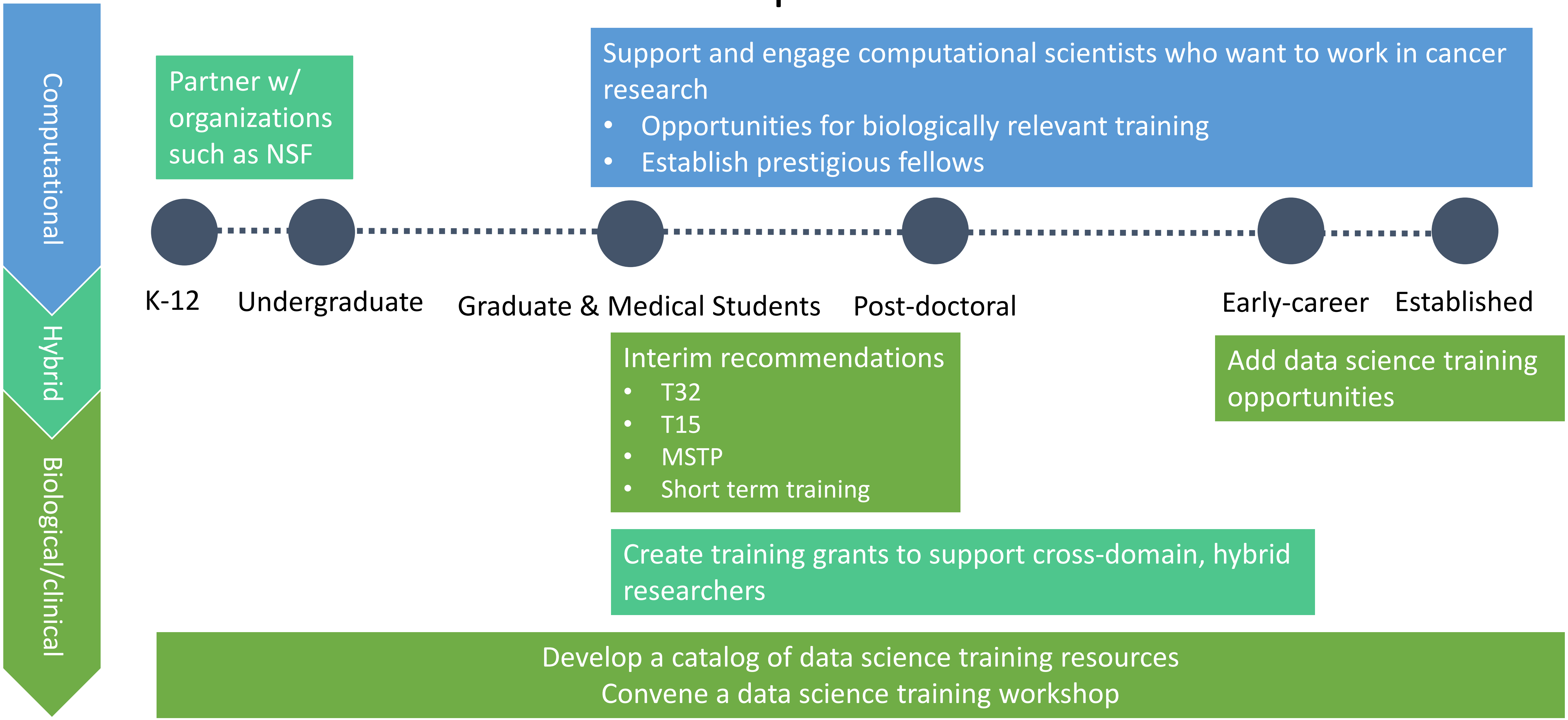
** Numbers do not include students and postdoctoral fellows supported by NCI research project grants, cancer center grants, and other non-training mechanisms

† Does not distinguish between summer research experiences and part- or full-time appointments

cancer.gov/cct
cancer.gov/crchd

Accepted by NCAB August 2018

Recommendation 1: Data Science Training and Workforce Development



Computational

Hybrid

Biological/clinical

K-12

Undergraduate

Graduate & Medical Students

Post-doctoral

Early-career

Established

Partner w/ organizations such as NSF

Support and engage computational scientists who want to work in cancer research

- Opportunities for biologically relevant training
- Establish prestigious fellows

Interim recommendations

- T32
- T15
- MSTP
- Short term training

Add data science training opportunities

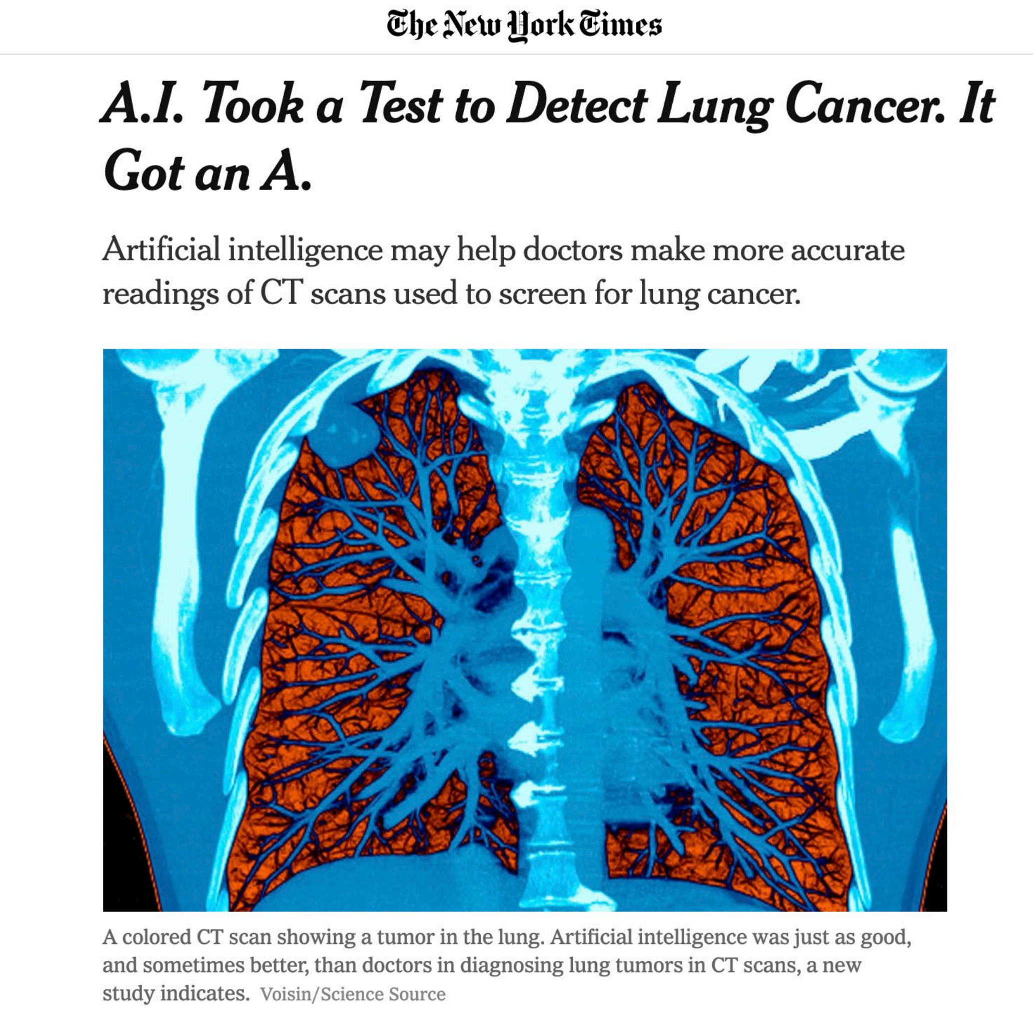
Create training grants to support cross-domain, hybrid researchers

Develop a catalog of data science training resources

Convene a data science training workshop

Recommendation 2: Machine Learning (ML)

- Develop targeted machine learning (ML) methodology for cancer research
  - Artificial Intelligence (AI) ethics
  - ML infrastructure for drug discovery
  - Automation of data curation
  - Effective translation of ML methodologies into clinical care
- Compile large, diverse datasets for training and ML algorithms



The New York Times

A.I. Took a Test to Detect Lung Cancer. It Got an A.

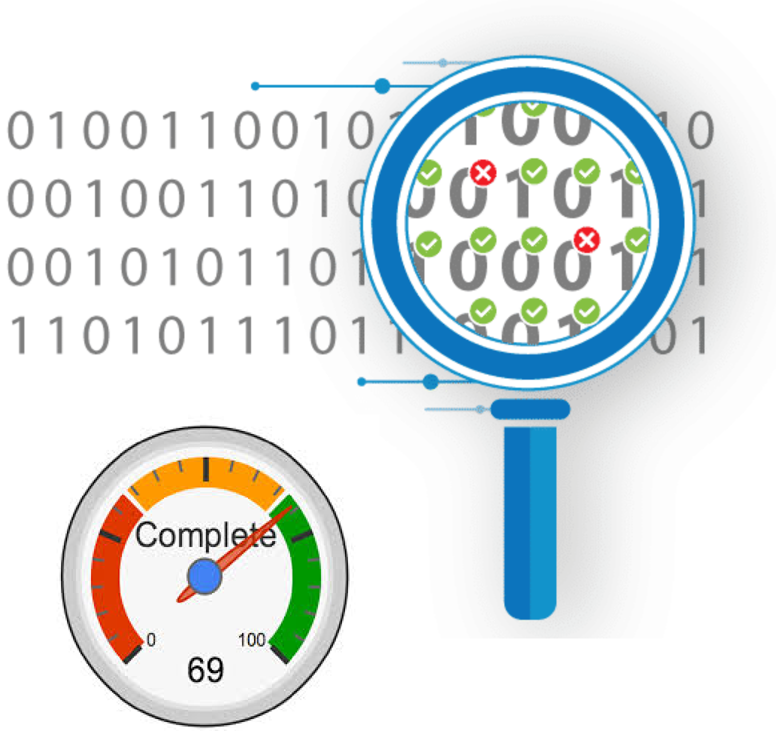
Artificial intelligence may help doctors make more accurate readings of CT scans used to screen for lung cancer.

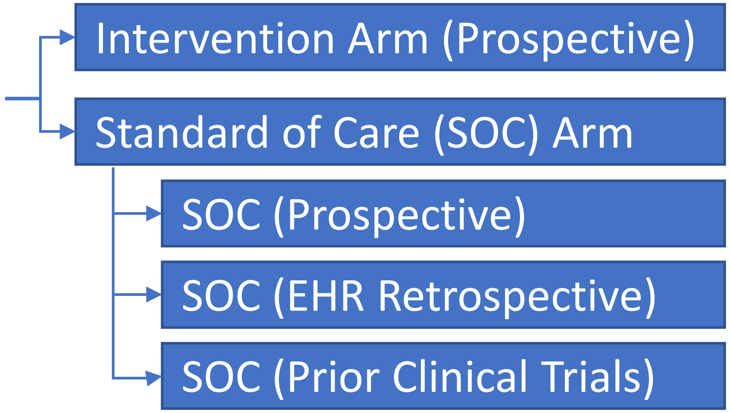
A colored CT scan showing a tumor in the lung. Artificial intelligence was just as good, and sometimes better, than doctors in diagnosing lung tumors in CT scans, a new study indicates. Voisin/Science Source

- Develop new funding opportunities for ML research to attract a broad ML community to cancer research

Recommendation 3: Real World Data (RWD)

- Convene stakeholders around a RWD metadata model to describe the completeness and quality of RWD
- Create a RWD framework and criteria for evaluating and populating key concepts from EHRs and other RWD sources
    - Build off of existing frameworks such as PRISSMM, mCODE
- Demonstrate the utility of RWD in a series of Learning Healthcare Systems reference implementations
    - NCI clinical trials leveraging RWD to:
        - Design eligibility criteria
        - Supplement recruitment to a standard of care trial arm
    - EHR implementation of RWD framework and demonstration of utility in driving use cases





Intervention Arm (Prospective)

Standard of Care (SOC) Arm

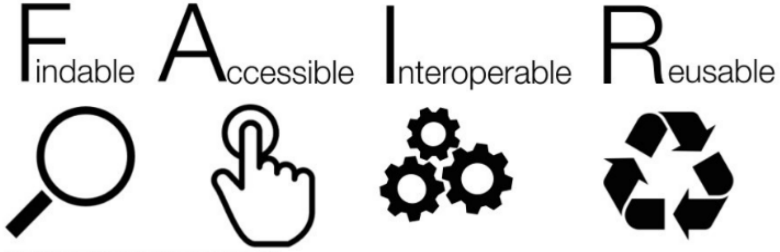
SOC (Prospective)

SOC (EHR Retrospective)

SOC (Prior Clinical Trials)

Recommendation 4: Enabling the cultural shift toward data sharing

- Develop best practices for consent/common consent language
- Streamline data sharing policy and requirements, including access to data
- Provide appropriate funding and resources to support data sharing
- Develop training for data management processes and policies
- Create systems to attribute and credit investigators for sharing data



Sungya Pundir, Wikimedia commons, CC By SA 4.0

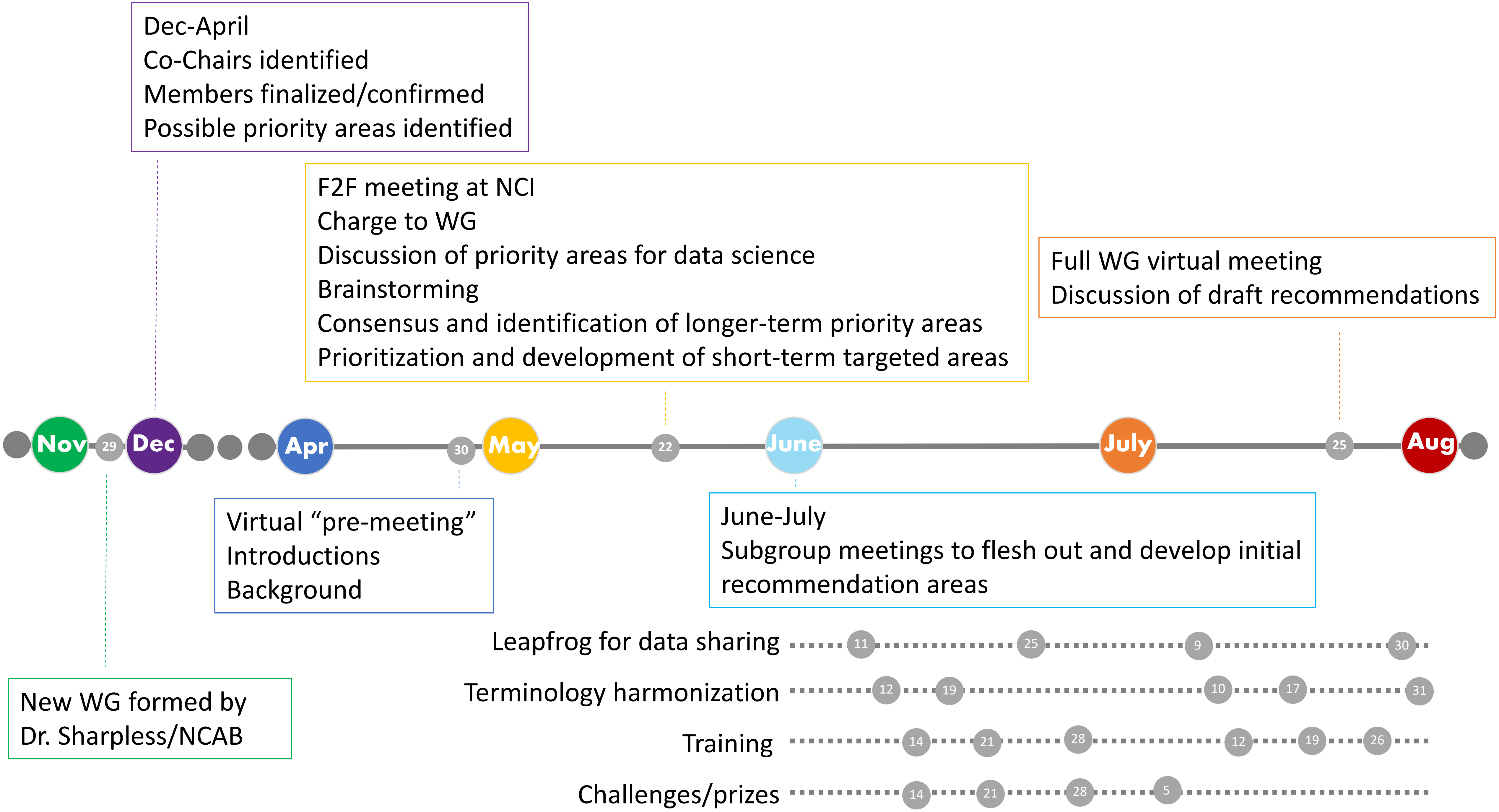
Data Science Opportunities for the NCI

Interim Recommendations

National Cancer Advisory Board

Ad hoc Working Group on Data Science

August 14, 2018



Dec-April
Co-Chairs identified
Members finalized/confirmed
Possible priority areas identified

F2F meeting at NCI
Charge to WG
Discussion of priority areas for data science
Brainstorming
Consensus and identification of longer-term priority areas
Prioritization and development of short-term targeted areas

Full WG virtual meeting
Discussion of draft recommendations

Virtual “pre-meeting”
Introductions
Background

June-July
Subgroup meetings to flesh out and develop initial recommendation areas

New WG formed by
Dr. Sharpless/NCAB

Initial Recommendation Areas

1. Investments to leapfrog data sharing for high-value datasets
2. Harmonization of terminology between cancer research data and clinical care data
3. Support of data science training at the graduate level
4. Opportunities for funding challenges and prizes

Recommendation 1: Investments to leapfrog data sharing for high-value datasets

- Resources to support
  - Identification
  - Enrichment
  - Curation
  - Harmonization
  - Annotation
  - Publishing

FOUNDATION MEDICINE

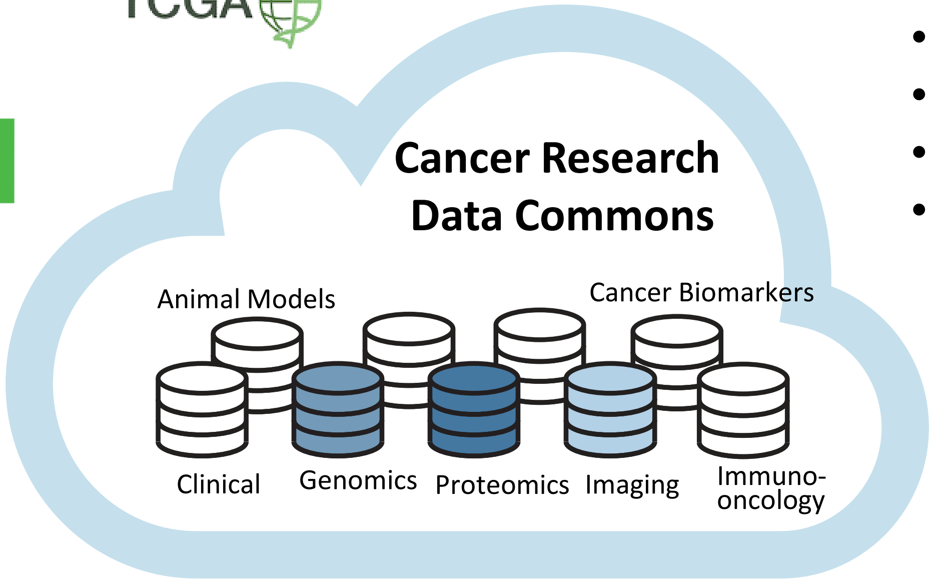
MM RF MULTIPLE MYELOMA Research Foundation

CANCER IMAGING ARCHIVE

TCGA

AACR American Association for Cancer Research FINDING CURES TOGETHER

PROJECTGENIE Genomics Evidence Neoplasia Information Exchange



Subgroup members:

- John Carpten
- Warren Kibbe
- Mia Levy
- Vince Miller
- Charles Sawyers
- Nick Wagle

- Examples of high-value datasets
  - Those fully collected and annotated but not yet shared in a public repository
  - Datasets that would be enhanced by additional data generation and/or collection (e.g., genomic datasets needing additional clinical annotation)

Recommendation 2: Harmonize terminologies between cancer research and clinical care

- Augment EHR data standards to further bridge clinical care and cancer research
- Fund research related to achieving near clinical trial grade data within traditional clinical care settings
- Identify and prioritize existing standards bodies and activities

RXNORM

SNOMED CT

The global
language of
healthcare

LOINC®

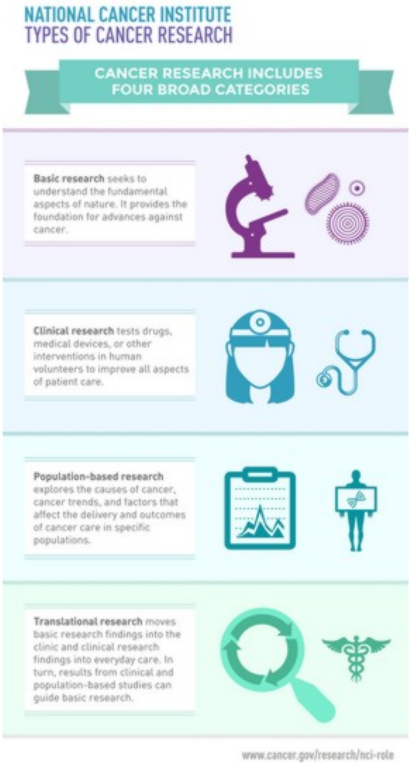
from Regenstrief

Subgroup members:

- George Hripcsak
- Mimi Huizinga
- Warren Kibbe
- Michelle Le Beau

Benefits of harmonized terminologies

- Increase the utility and ease of incorporation/integration of clinical care data from EHRs into cancer research
- Enable more efficient research, better patient care, and better real-world evidence generation
- Enhance integration of the cancer and non-cancer research communities



NATIONAL CANCER INSTITUTE
TYPES OF CANCER RESEARCH

CANCER RESEARCH INCLUDES FOUR BROAD CATEGORIES

Basic research seeks to understand the fundamental aspects of nature. It provides the foundation for advances against cancer.

Clinical research tests drugs, medical devices, or other interventions in human volunteers to improve all aspects of patient care.

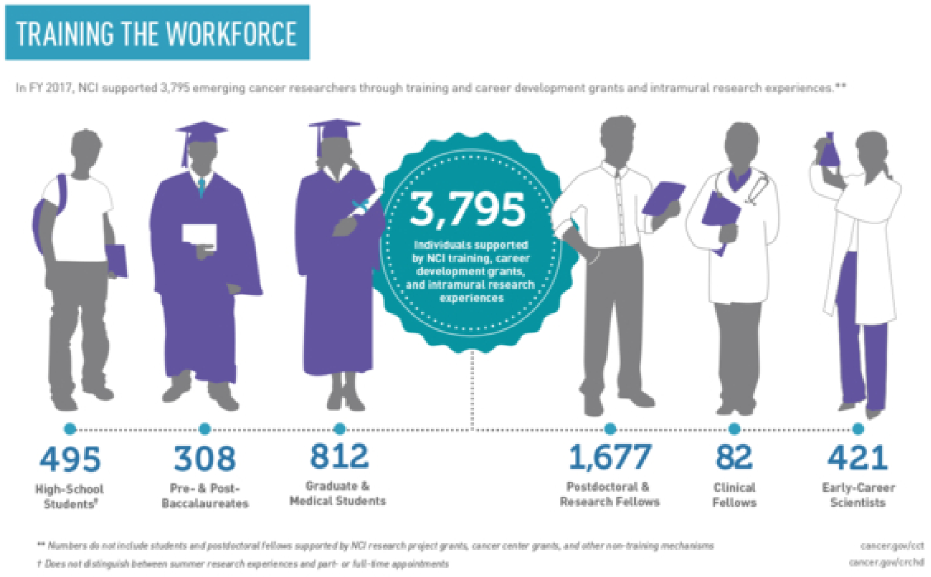
Population-based research explores the causes of cancer, cancer trends, and factors that affect the delivery and outcomes of cancer care in specific populations.

Translational research moves basic research findings into the clinic and clinical research findings into everyday care. In turn, results from clinical and population-based studies can guide basic research.

www.cancer.gov/research/nci-role

Recommendation 3: Increase the number of training programs and trainees in cancer data science

- Dedicate a specific T32 training program in cancer data science
- Contribute to existing NIH training programs
  - NLM T15 training programs
  - NIGMS Medical Scientist Training program
- Develop a short-term training program for clinicians and biological scientists



In FY 2017, NCI supported 3,795 emerging cancer researchers through training and career development grants and intramural research experiences.**

** Numbers do not include students and postdoctoral fellows supported by NCI research project grants, cancer center grants, and other non-training mechanisms

† Does not distinguish between summer research experiences and part- or full-time appointments

cancer.gov/cct
cancer.gov/crchd

Subgroup members:

- Regina Barzilay
- Amanda Haddock
- Rebecca Jacobson
- Anne-Marie Meyer
- Sylvia Plevritis
- Kim Sabelko

Recommendation 4: Sponsor a series of data science challenges

Subgroup members:

- Regina Barzilay
- Amanda Haddock
- Michelle Le Beau
- Lincoln Stein

- Potential challenge topics (~4-8 per year)
  - Drug response prediction
  - Discovery of multi-omic prognostic biomarkers
  - De-convolution of heterogenous tumors
  - Cancer diagnosis, grading, and staging
  - Facility of data access and integration from the ethical, legal, and social implications standpoint
- Consider beginning with an "idea challenge" to identify the appropriate challenge topic/task/question

The Open Science Prize

precisionFDA Challenges
Engage and improve NGS software through community challenges

DREAM CHALLENGES
powered by Sage Bionetworks

Benefits of data science challenges

- Spur research in computational cancer biology and increase the availability of advanced analytic software to the broader research community
- Attract new talent to cancer research
- Validation and dissemination of state-of-the-art tools and technologies
- Demonstrates the inter-relationship between all the recommendations. Challenges require:
  - Openly shared datasets
  - Ability to work across harmonized datasets
  - Participants with appropriate skillsets and expertise