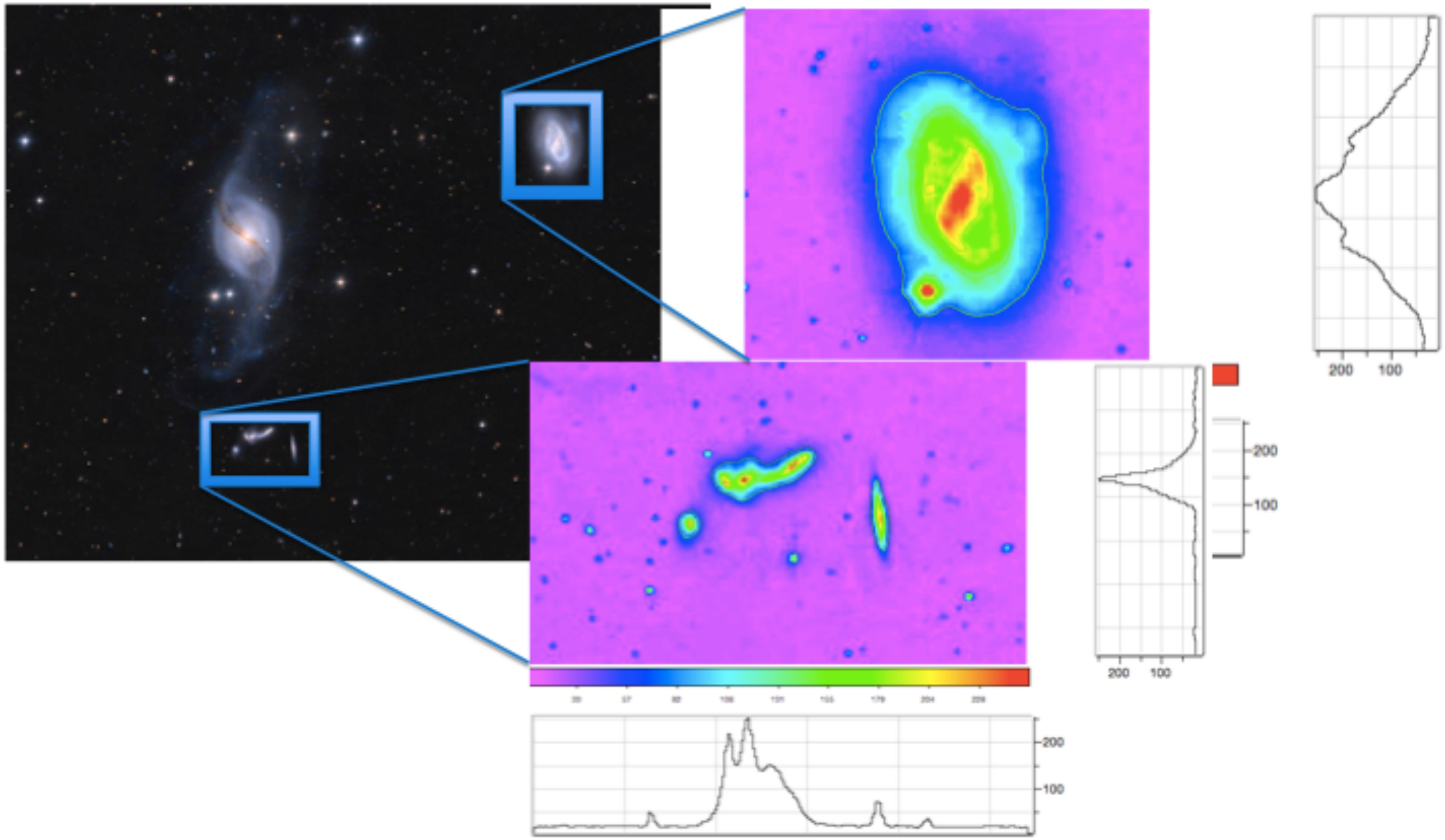
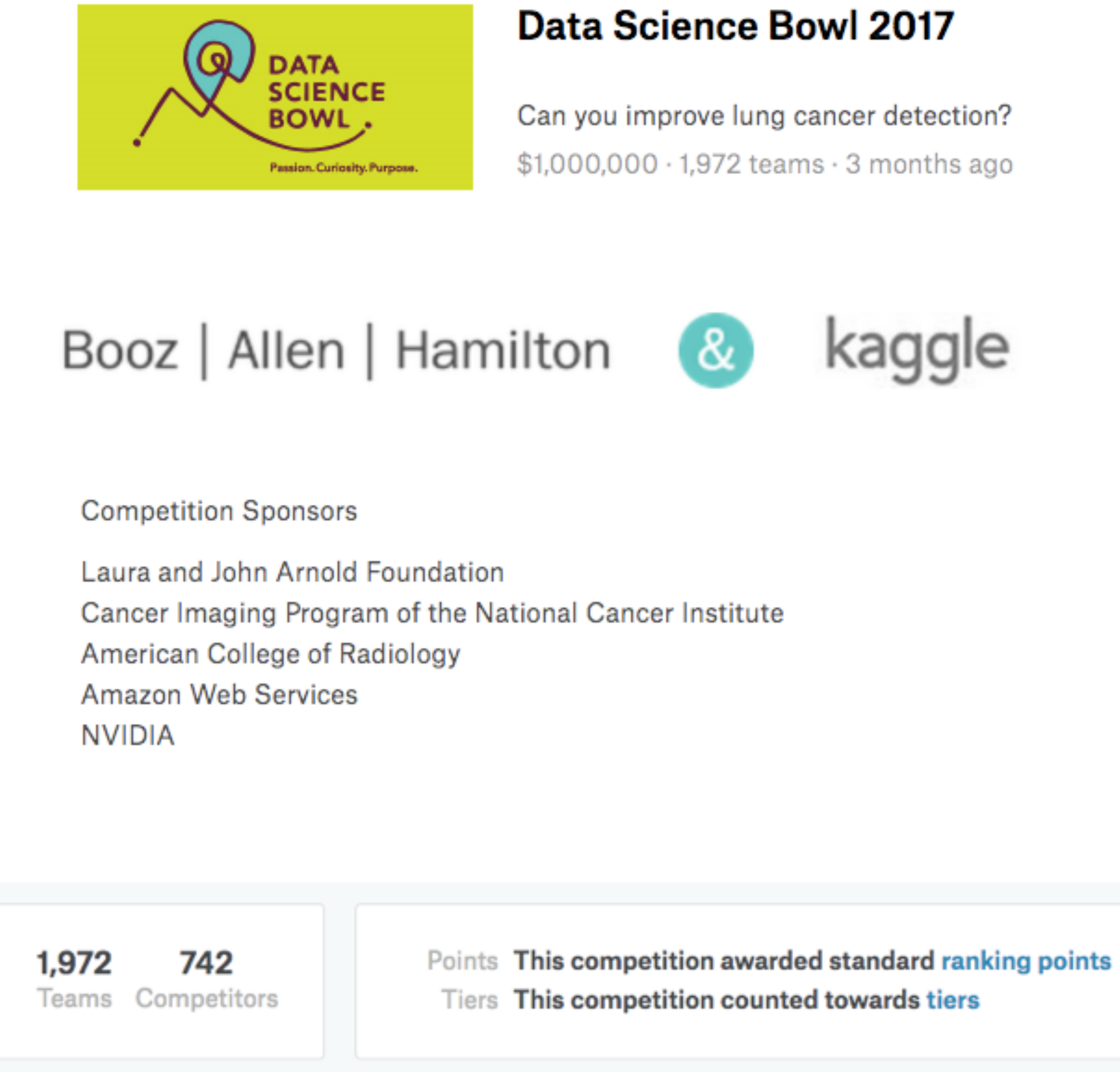
Recent NCI/NIH data challenges



Ashish Mahabal <aam@astro.caltech.edu>
LSST TVS Co-chair
Center for Data-Driven Discovery, Caltech
PLAsTiCC, LSST, Math for America, 2017-07-14



Data Science Bowl 2017

Can you improve lung cancer detection?

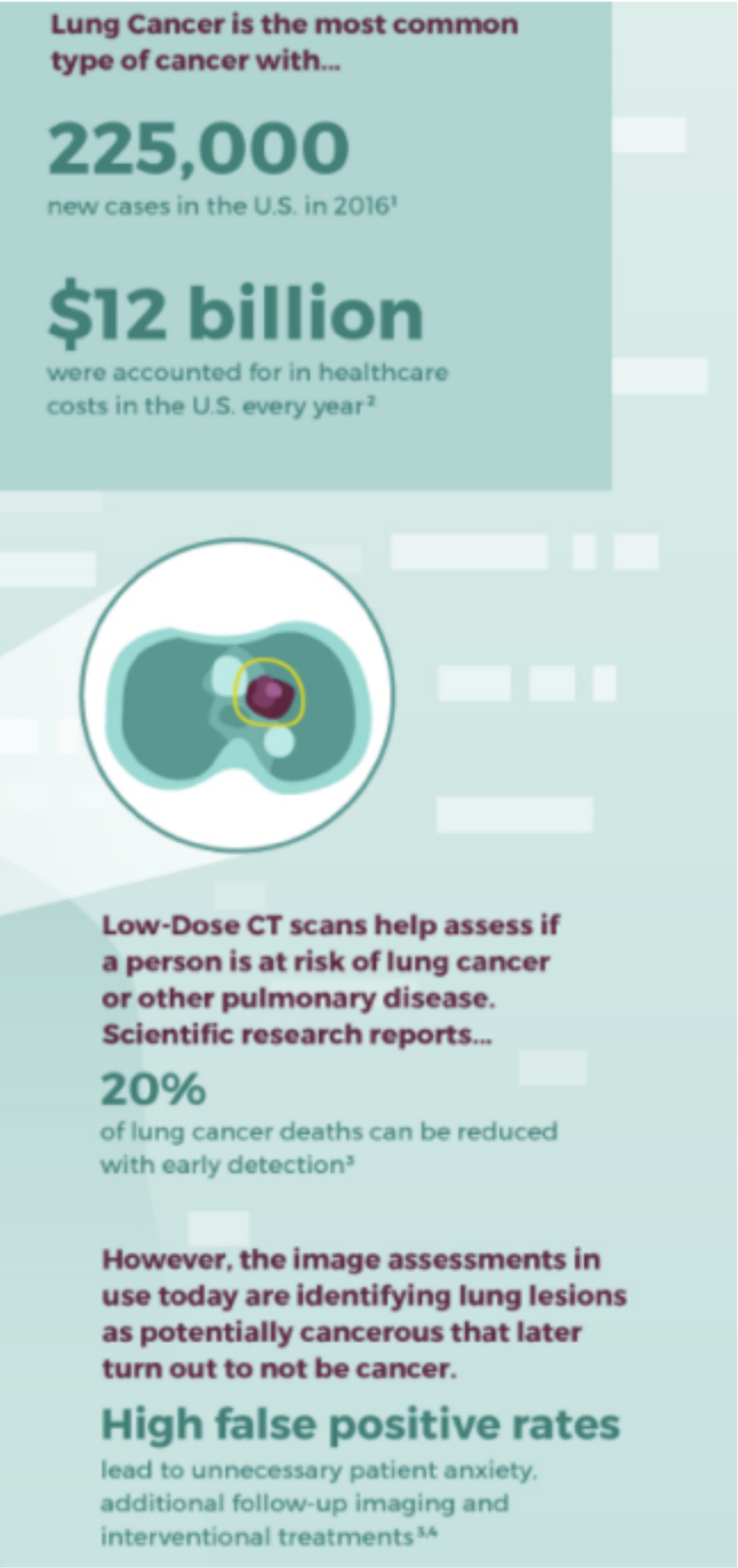
$1,000,000 · 1,972 teams · 3 months ago

Competition Sponsors

Laura and John Arnold Foundation
Cancer Imaging Program of the National Cancer Institute
American College of Radiology
Amazon Web Services
NVIDIA

1,972 Teams
742 Competitors

Points This competition awarded standard ranking points
Tiers This competition counted towards tiers



Lung Cancer is the most common type of cancer with...

225,000
new cases in the U.S. in 2016[1]

$12 billion
were accounted for in healthcare costs in the U.S. every year[2]

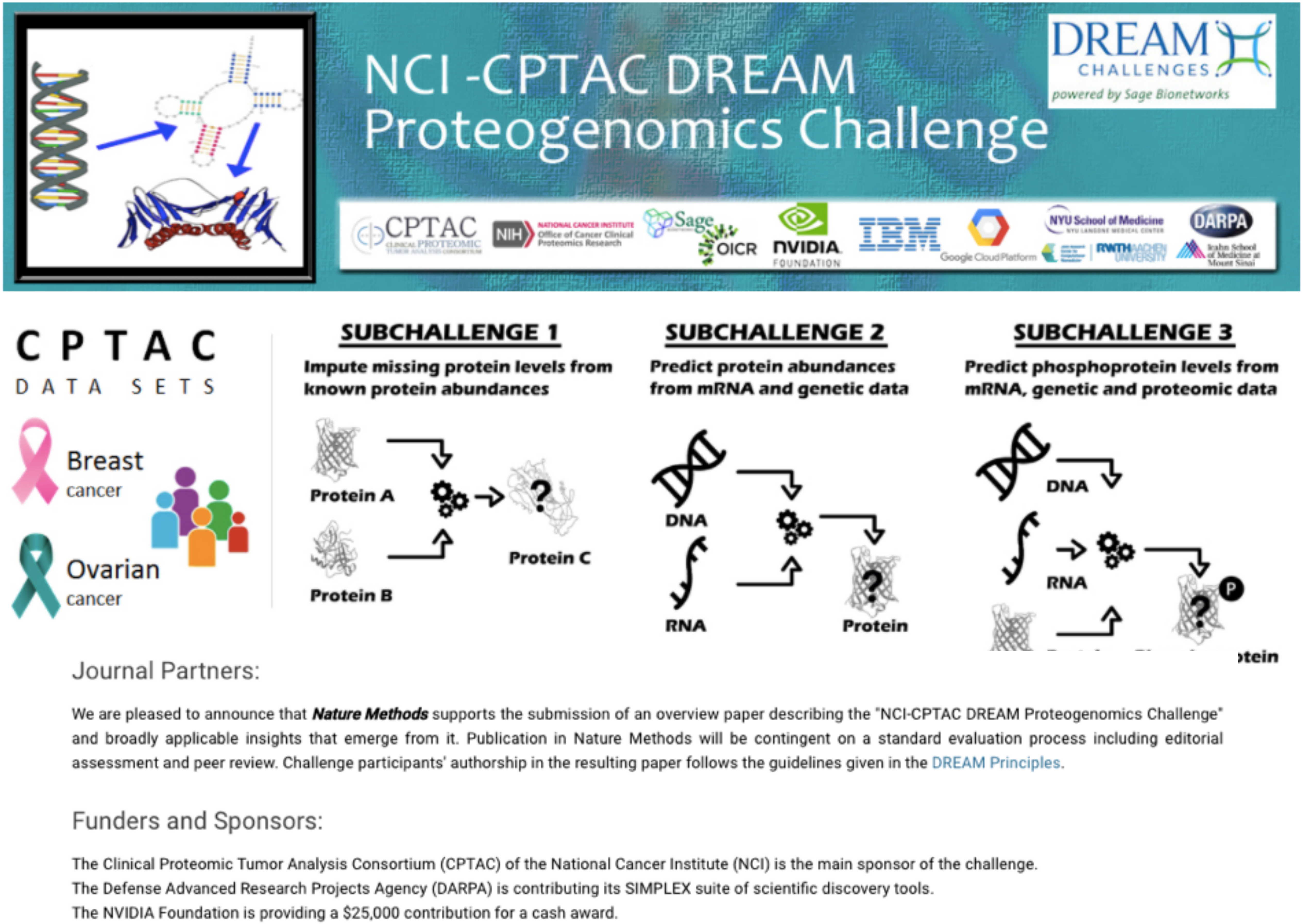
Low-Dose CT scans help assess if a person is at risk of lung cancer or other pulmonary disease. Scientific research reports...

20%
of lung cancer deaths can be reduced with early detection[3]

However, the image assessments in use today are identifying lung lesions as potentially cancerous that later turn out to not be cancer.

High false positive rates
lead to unnecessary patient anxiety, additional follow-up imaging and interventional treatments[3,4]

https://www.kaggle.com/c/data-science-bowl-2017 https://www.synapse.org/#!Synapse:syn8228304/wiki/413428



NCI -CPTAC DREAM Proteogenomics Challenge

DREAM CHALLENGES
powered by Sage Bionetworks
CPTAC CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM
NIH NATIONAL CANCER INSTITUTE Office of Cancer Clinical Proteomics Research
Sage
OICR
NVIDIA FOUNDATION
IBM
Google Cloud Platform
NYU School of Medicine NYU LANGONE MEDICAL CENTER
RWTH AACHEN UNIVERSITY
DARPA
Icahn School of Medicine at Mount Sinai

CPTAC DATA SETS

Breast cancer
Ovarian cancer

SUBCHALLENGE 1

Impute missing protein levels from known protein abundances

Protein A
Protein B
?
Protein C

SUBCHALLENGE 2

Predict protein abundances from mRNA and genetic data

SUBCHALLENGE 3

Predict phosphoprotein levels from mRNA, genetic and proteomic data

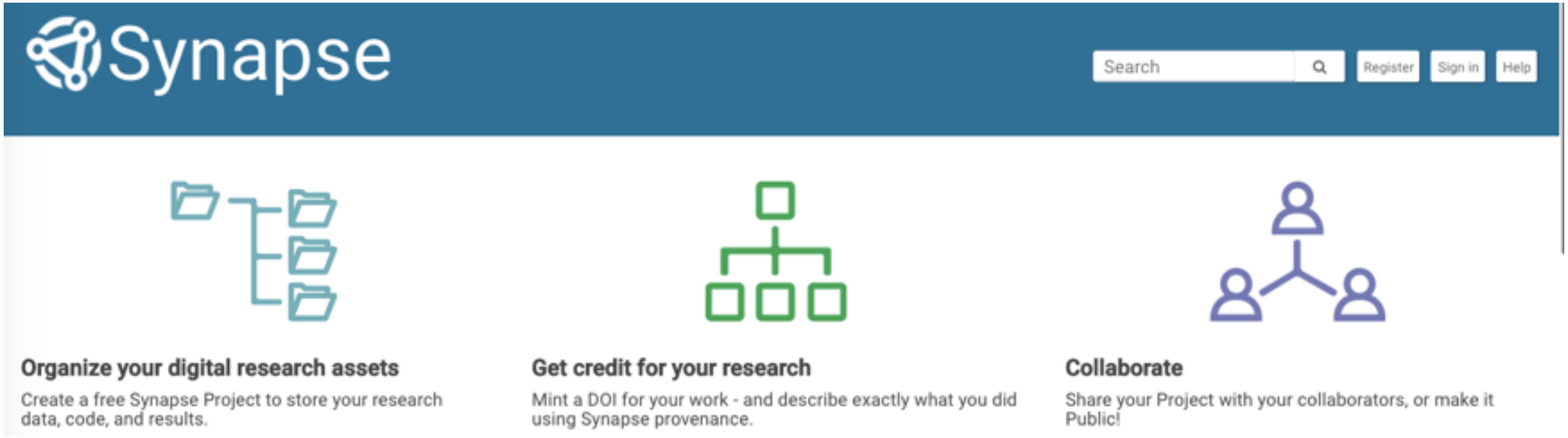
Journal Partners:

We are pleased to announce that ***Nature Methods*** supports the submission of an overview paper describing the "NCI-CPTAC DREAM Proteogenomics Challenge" and broadly applicable insights that emerge from it. Publication in Nature Methods will be contingent on a standard evaluation process including editorial assessment and peer review. Challenge participants' authorship in the resulting paper follows the guidelines given in the DREAM Principles.

Funders and Sponsors:

The Clinical Proteomic Tumor Analysis Consortium (CPTAC) of the National Cancer Institute (NCI) is the main sponsor of the challenge.
The Defense Advanced Research Projects Agency (DARPA) is contributing its SIMPLEX suite of scientific discovery tools.
The NVIDIA Foundation is providing a $25,000 contribution for a cash award.

synapse < https://www.synapse.org/>



Organize your digital research assets

Create a free Synapse Project to store your research data, code, and results.

Get credit for your research

Mint a DOI for your work - and describe exactly what you did using Synapse provenance.

Collaborate

Share your Project with your collaborators, or make it Public!

DREAM CHALLENGES

dream challenges < http://dreamchallenges.org/>

GA4GH/DREAM Workflow Execution Challenge

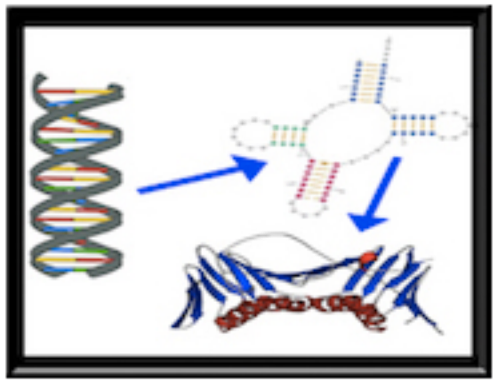
Launches July 5 (Pre-registration is open)

The goal of this challenge is to evaluate systems and platforms for executing portable analysis workflows in the interest of developing common standards and best practices. Participants will run high quality genomics workflows in their system of choice to assess portability and reproducibility in a concrete way.

Parkinson's Disease Digital Biomarker DREAM Challenge

Launching July 6, 2017 (Pre-registration is open)

Using data collected through two Parkinson's Disease mobile research studies, the goal of this challenge is to identify the best methods for processing mobile sensor data in order to distinguish gait and motor differences between Parkinson's Disease patients and controls.



NCI-CPTAC DREAM Proteogenomics Challenge

Launches June 26 (Registration is open)

This challenge will use public and novel proteogenomic data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) to benchmark an understanding of the interfaces between different layers of information in a population of

Overview of training data set:

In silico data (subchallenge 1):

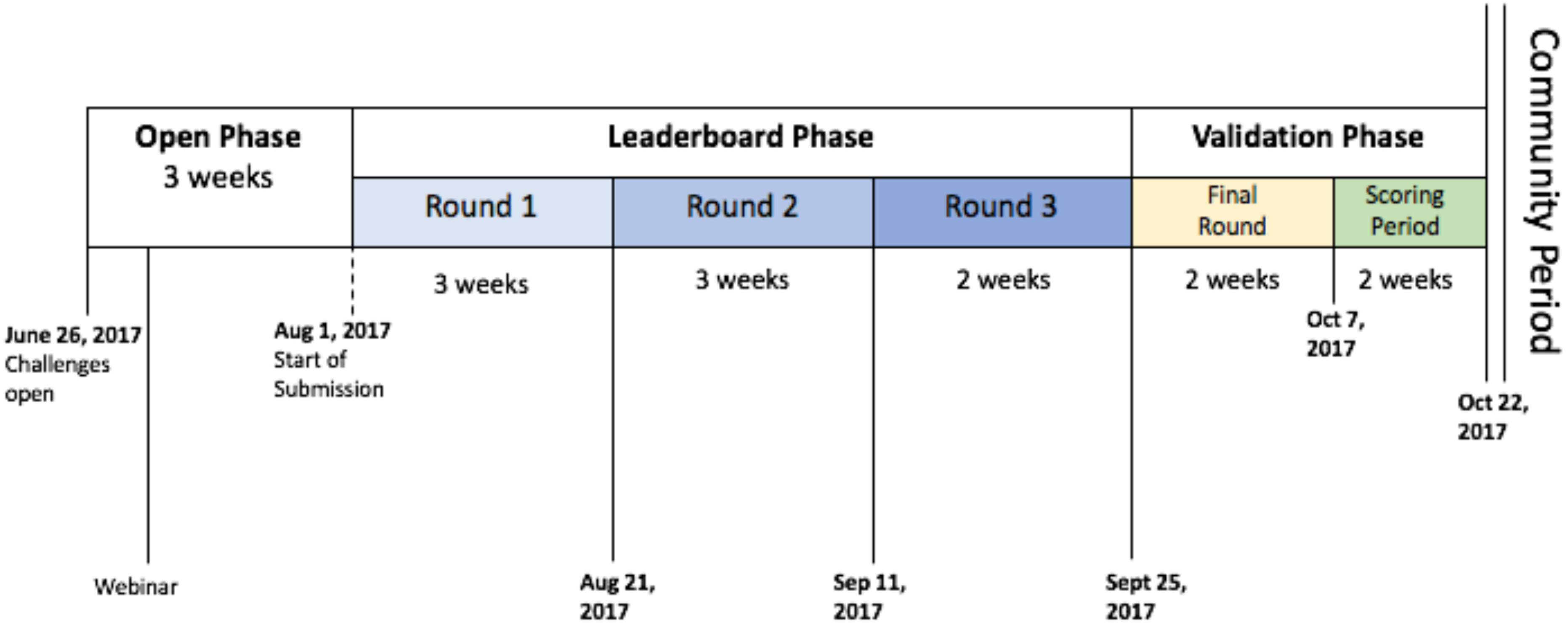
- Complete data without missing values: 80 by 7927.
- 10 training matrices with missing values: 80 by 7927.

Breast cancer (subchallenge 2 and 3):

- Proteomics: 10597 proteins for 105 patients
- Phosphoproteomics: 54994 phospho-sites for 105 patients
- CNA: 16884 genes for 77 patients
- mRNA: 15115 genes for 77 patients

Ovarian cancer (subchallenge 2 and 3):

- Proteomics: 7163 proteins for 174 patients
- Phosphoproteomics: 12473 phosphosites for 69 patients
- CNA: 11859 genes for 559 patients
- mRNA(Array): 15632 genes for 569 patients
- mRNA(RNA-seq): 15632 genes for 294 patients



Desirable stipulations

- Publication embargo: Don’t publish your methods until data challenge results are published
- External data: OK to use if it is public
- Reproducibility: entry should not contain proprietary software or databases