SageBionetworks

Advancing cancer benchmarking and data sharing through crowd-sourced challenges

ITCR 2022

James Eddy @ Sage Bionetworks

Core Values @ Sage

- **Be intentional** – consider solutions from more than one perspective.
- Promote an **ecosystem of sharing** – with proper **attribution**.
- Solve specific problems with general solutions – make these available for **reuse and adaptation**.
- Do **trustworthy, impactful work** – prioritize **outcomes over ego**.
- **Be bold** and willing to experiment.

sagebionetworks.org/our-culture

Community Coordination @ Sage

Our Approach

Cutting edge research & data

Community challenges & benchmarking

Reproducible computing at scale

FAIR data sharing & discovery

Molecular, clinical, & imaging data integration

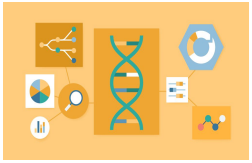
Standards & interoperability

What do we do?

At Sage we believe that by harnessing the power of open science, we help research communities develop reliable outcomes to advance our understanding of human health.

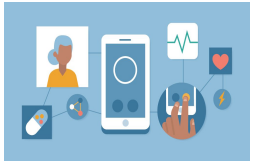
Responsible Data Sharing

Sage supports research collaborations by overseeing data coordination, visualization, and analytics across distributed teams. We manage grant- or project-based research consortia to share, evaluate and distribute data, methods, and insights.



Benchmarking Reliable Methods

Because we are all susceptible to the self-assessment bias, Sage has developed tools that help researchers to objectively benchmark the performance of computational methods, and to disseminate community-verified methods.



Understand Real-World Evidence

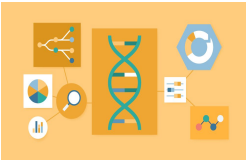
By applying our approach to digital health, Sage works with participants and researchers to understand how real-world environments impact our individual experience of health and disease.

What do we do?

At Sage we believe that by harnessing the power of open science, we help research communities develop reliable outcomes to advance our understanding of human health.

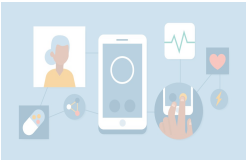
Responsible Data Sharing

Sage supports research collaborations by overseeing data coordination, visualization, and analytics across distributed teams. We manage grant- or project-based research consortia to share, evaluate and distribute data, methods, and insights.



Benchmarking Reliable Methods

Because we are all susceptible to the self-assessment bias, Sage has developed tools that help researchers to objectively benchmark the performance of computational methods, and to disseminate community-verified methods.

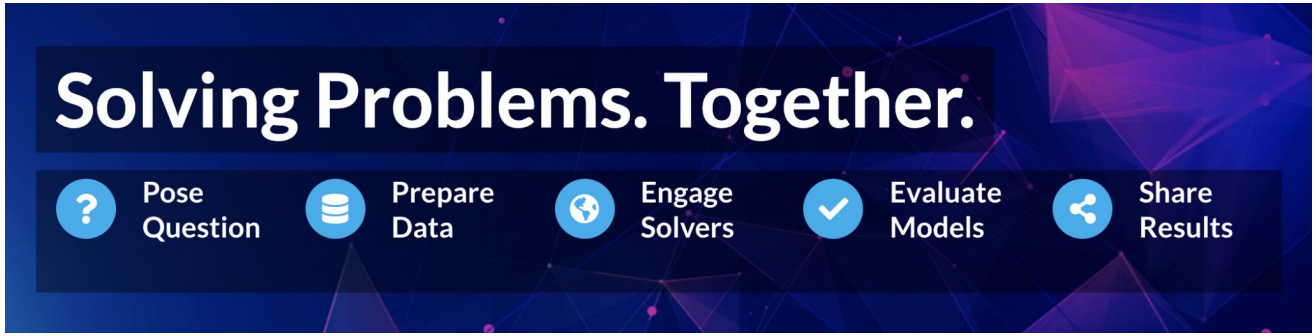


Understand Real-World Evidence

By applying our approach to digital health, Sage works with participants and researchers to understand how real-world environments impact our individual experience of health and disease.

Our mission is

- to contribute to the solution of important **biomedical** problems
- to foster **collaboration** between research groups
- to **democratize access** to data
- to **accelerate research**
- to **objectively assess** algorithms and their performance



DREAM Challenges use crowd-sourcing to solve complex biomedical research questions

dreamchallenges.org

60+

Crowd-sourced DREAM Challenges have benchmarked informatic algorithms in biomedicine

30,000

Cross-disciplinary participants from around the world have volunteered as solvers.

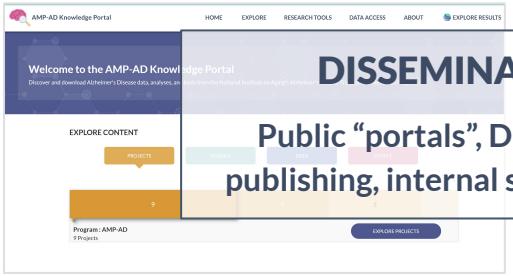
105+

Academic journal publications have resulted from DREAM Challenges covering a range of disease areas

A Few of Our Partners

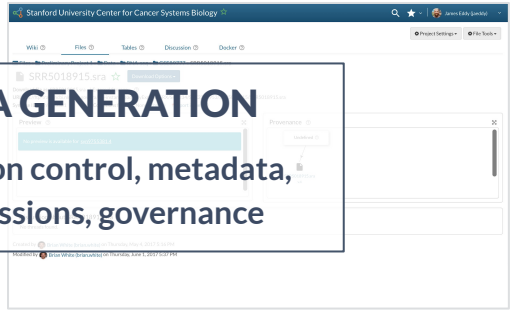
IBM Research
SageBionetworks
NIH National Institutes of Health
AstraZeneca
Celgene
Bill & Melinda Gates foundation
Bristol-Myers Squibb
Oregon Health & Science University
CD2H National Center for Data to Health
UCLA
Allen Institute
UW Medicine UW School of Medicine
Universität Heidelberg Zukunft seit 1386
Icahn School of Medicine at Mount Sinai
University of Colorado Anschutz Medical Campus
Columbia University Irving Medical Center
The Michael J. Fox Foundation for Parkinson's Research
Stanford University

Synapse: Cloud Platform for Challenges



DISSEMINATION

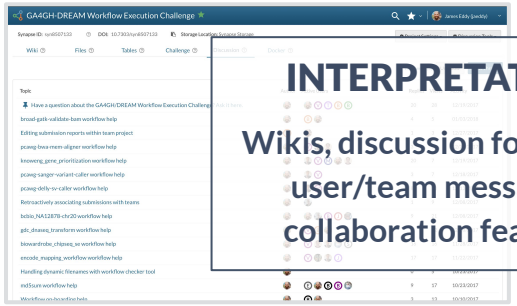
Public “portals”, DOIs for publishing, internal sharing



DATA GENERATION

Version control, metadata, permissions, governance

Supporting Team Science

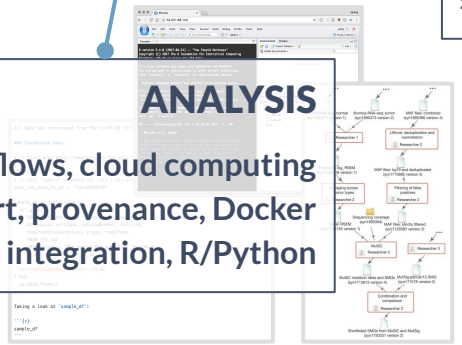


INTERPRETATION

Wikis, discussion forums, user/team messaging, collaboration features

CURATION

Data validation, annotation, custom metadata views, search & discovery



ANALYSIS

Workflows, cloud computing support, provenance, Docker integration, R/Python

SAGE BIONETWORKS

SYNAPSE

synapse.org

Project Overview

SageBionetworks

ITCR U24 for Advancing Method Benchmarking

Award Type	Award Year	PI Name(s)	Institution(s)	Project	Status
U24	2020	• James Eddy • Jacob Albrecht • Paul Christopher Boutros	• Sage Bionetworks	Advancing Method Benchmarking and Data Sharing Through Crowd-Sourced Competitions in Cancer Research	Active

- **AIM 1:** Develop a community hub and ***benchmarking toolkit*** for biomedical challenges.
- **AIM 2:** Develop ***portable software and services for distributed benchmarking*** on sensitive and protected data.
- **AIM 3:** Expand the biomedical challenge community through improvements in education, outreach, and empowering the organization of independent challenges and benchmarking projects.

Leadership Team

Justin Guinney
VP, Computational Oncology
Sage Bionetworks

TEMPUS

James Eddy
Director of Architecture & Operations
Data & Tooling @ Sage Bionetworks

Paul Boutros
Director, Cancer Data Science @ UCLA
Professor, Human Genetics @ UCLA

Paul Boutros
Still Paul!

Gustavo Stolovitzky
Director, Translational Systems Bio @ IBM
Founder, DREAM Challenges

sema4

Jake Albrecht
Director of Challenges and Benchmarking
Benchmarking @ Sage Bionetworks

Problem Statements

1. “Self-assessment” has impeded the validation and dissemination of bioinformatics tools and methods.

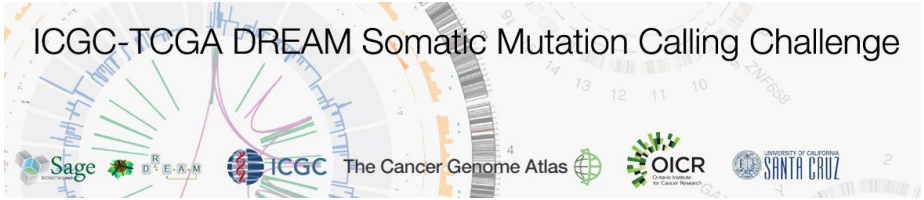
2. New paradigms for data sharing and community engagement are needed to unlock understanding around critical datasets.

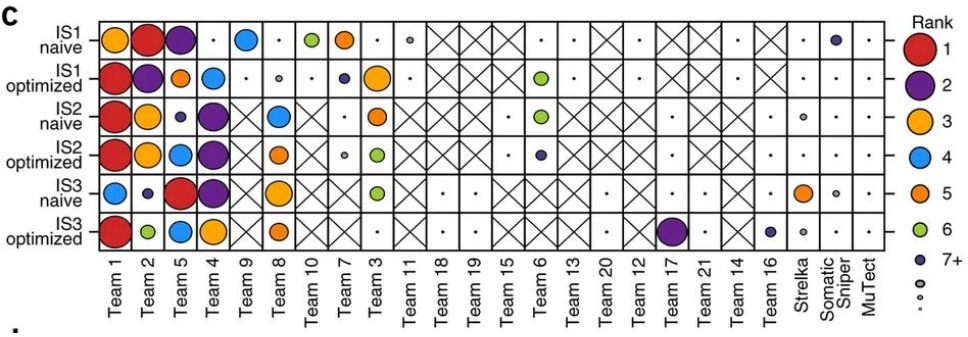
3. Translational and clinical tools require robust assessments of performance and generalizability on diverse patient cohorts.

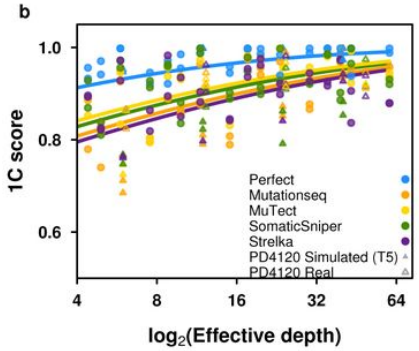
Problem #1: Fix the Self-Assessment Trap

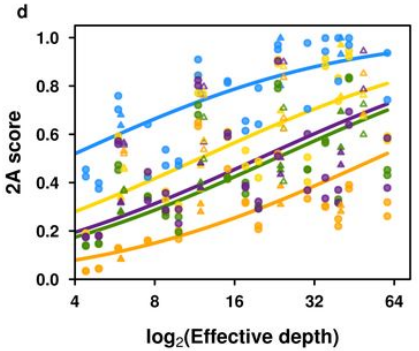
"Self-assessment" has impeded validation and dissemination of computational methods.

Answer: challenges provide a rigorous framework for benchmarking tools









SageBionetworks

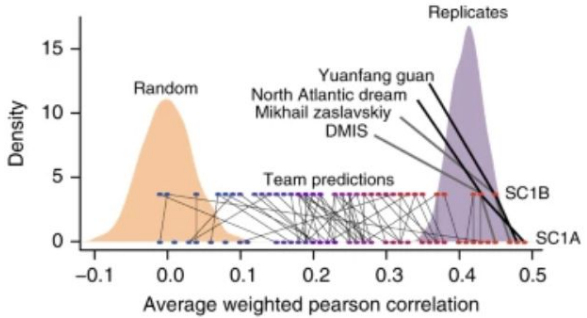
Nat Genetics, 2014; Nat Methods, 2015; Genome Bio, 2018

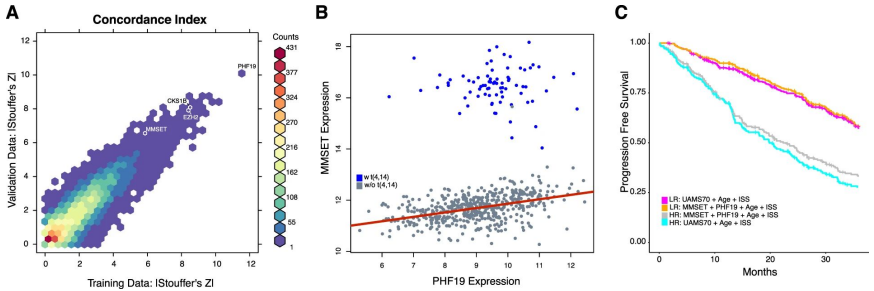
Nat Biotech, 2020; Nat Methods, in press

Problem #2: Unlock Hidden Data & Create Communities

New paradigms for data sharing and community engagement are needed to unlock understanding around critical datasets.

Answer: challenges provide innovative mechanism for data access & community participation





SageBionetworks

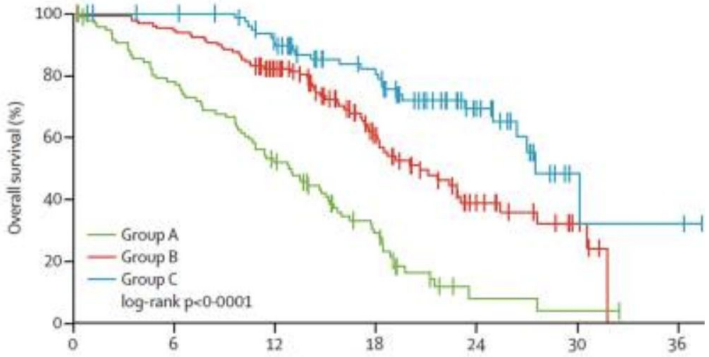
Nat Comm, 2019

Leukemia, 2020

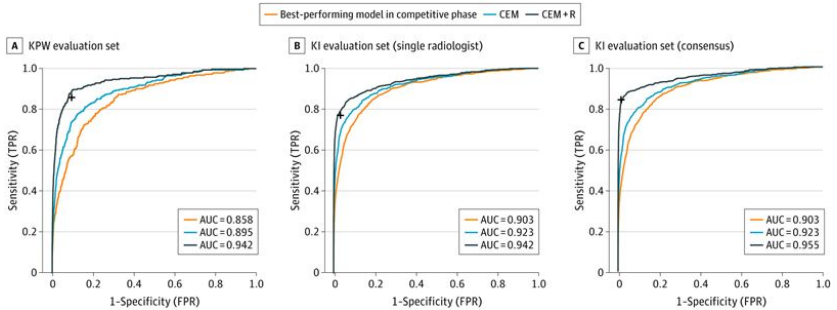
Problem #3: Robust Performance Assessment

Translational and clinical tools will require robust assessments of performance and generalizability on diverse patient cohorts.

Answer: challenges can define and assess clinical benchmarks

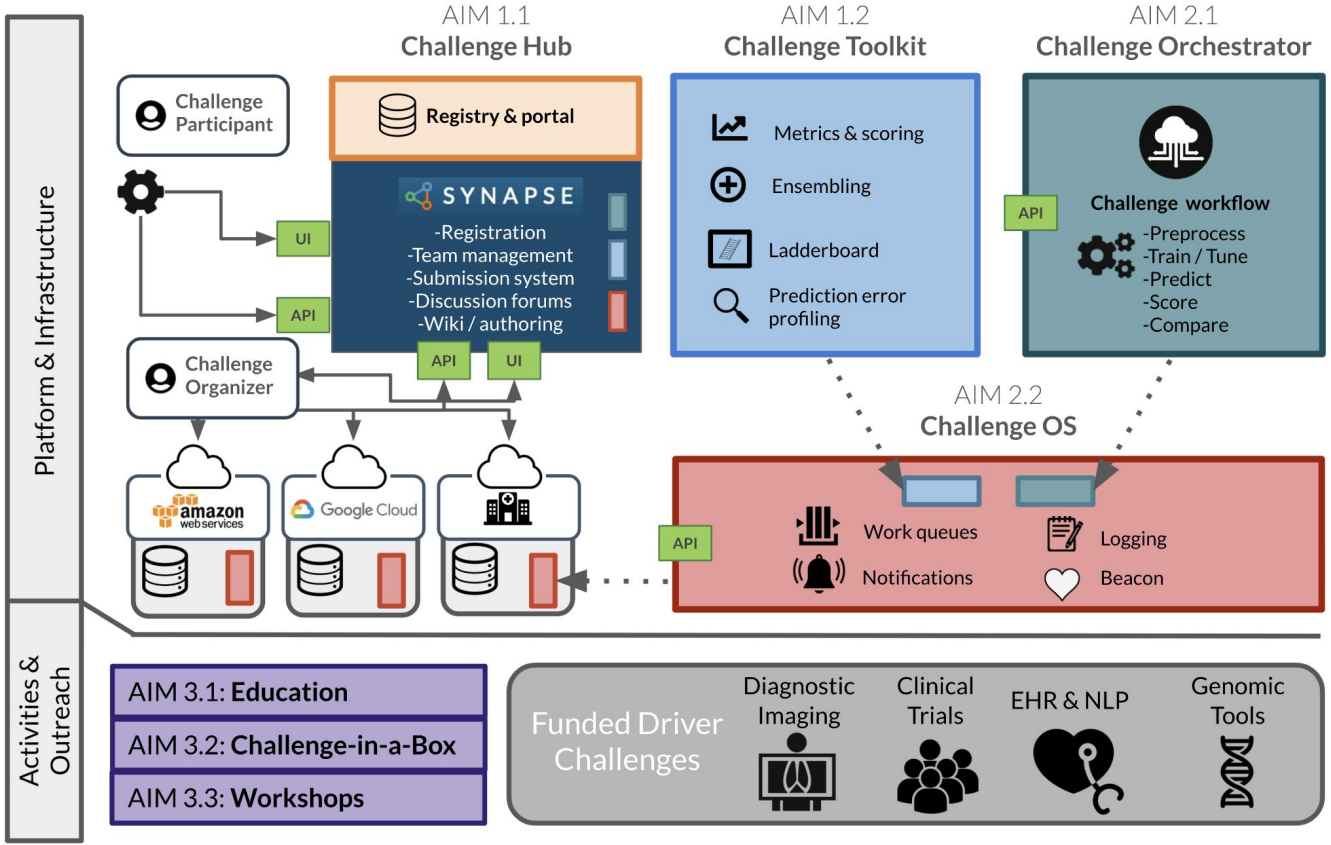


Lancet Oncology, 2017
JCO-Clinical Cancer Informatics 2017



JAMA Oncology, 2017
JAMA Network Open, 2020

Project at a Glance



SageBionetworks

Challenge Platform 2.0

SageBionetworks

Existing Limitations & Opportunities to Improve

- **Challenge user experience is… challenging**
  - Organizers need to be able to set up and launch new challenges with less overhead
  - Participants need easier to use interfaces and processes for submitting to a challenge (and getting help)
  - Evaluation infrastructure should be standardized and/or shared across challenges, so that less effort is needed to provision, configure, and maintain compute environments
- Discoverability of challenges and challenge results is minimal
  - Prospective participants (and organizers) need a better way to browse and search for challenges of interest, including upcoming, active, and closed challenges
  - We need to support discovery and insight that spans the boundaries of individual challenges
- Challenge solutions and artifacts are difficult to share and reuse
  - Modelers need a central view to identify challenge datasets that are open and available (to train or validate new methods)
  - Researchers need a way to access and use top performing methods and tools for their own projects
  - Organizers need a way to find and reuse validated metrics and evaluation routines

Existing Limitations & Opportunities to Improve

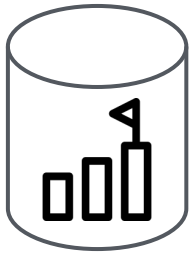
- Challenge user experience is… challenging
  - Organizers need to be able to set up and launch new challenges with less overhead
  - Participants need easier to use interfaces and processes for submitting to a challenge (and getting help)
  - Evaluation infrastructure should be standardized and/or shared across challenges, so that less effort is needed to provision, configure, and maintain compute environments
- **Discoverability of challenges and challenge results is minimal**
  - Prospective participants (and organizers) need a better way to browse and search for challenges of interest, including upcoming, active, and closed challenges
  - We need to support discovery and insight that spans the boundaries of individual challenges
- Challenge solutions and artifacts are difficult to share and reuse
  - Modelers need a central view to identify challenge datasets that are open and available (to train or validate new methods)
  - Researchers need a way to access and use top performing methods and tools for their own projects
  - Organizers need a way to find and reuse validated metrics and evaluation routines

Existing Limitations & Opportunities to Improve

- Challenge user experience is… challenging
  - Organizers need to be able to set up and launch new challenges with less overhead
  - Participants need easier to use interfaces and processes for submitting to a challenge (and getting help)
  - Evaluation infrastructure should be standardized and/or shared across challenges, so that less effort is needed to provision, configure, and maintain compute environments
- Discoverability of challenges and challenge results is minimal
  - Prospective participants (and organizers) need a better way to browse and search for challenges of interest, including upcoming, active, and closed challenges
  - We need to support discovery and insight that spans the boundaries of individual challenges
- **Challenge solutions and artifacts are difficult to share and reuse**
  - Modelers need a central view to identify challenge datasets that are open and available (to train or validate new methods)
  - Researchers need a way to access and use top performing methods and tools for their own projects
  - Organizers need a way to find and reuse validated metrics and evaluation routines

Key Elements of a Next-generation Challenge Platform

- **Lingua franca of challenges and benchmarking concepts**
- Purpose-built services and interfaces for challenge participation *and* administration
- Streamlined user interaction (with better docs) for submitting to a challenge, with accessible resources for improving performance
- **Standard libraries (or templates or SDKs) for challenge processes, data QC, evaluation metrics, and performance ranking**
- Information capture, dissemination, and reuse across challenges to accelerate the advancement of scientific knowledge



Challenge Silos?

Challenge Registry

SageBionetworks

Current Challenges (pun intended)

Motivations for the Challenge Registry

- Over 25 organizations that run biomedical challenges
  - Missing standardization
  - No central hub
  - Challenge searching can be overwhelming and time-consuming
- Limited exposure and participation

And more...

Our Solution

We aim to develop a portal to capture challenge metadata organized by different platforms.

1. Standardize the challenge annotations (via the MIAC schema)

Minimal Information About a Challenge

2. Apply MIAC to DREAM Challenges

3. Provide a web-based portal to access that data (Challenge Registry)

4. Engage with other challenge-running organizations to connect them to the Challenge Registry

Challenge Registry Team

Thomas Schaffter
Senior Research Software Engineer
Data & Tooling Group

Verena Chung
Bioinformatics Engineer
Benchmarking Group

Rong Chai
Bioinformatics Engineer
Benchmarking Group

Past contributors: Milen Nikolov, Thomas Yu, Michael Mason, Justin Guinney

Standardize and Restructure with MIAC

DREAM landscape snapshot

challenge	challengeYear	challengeKeywords	challengeSummary	ch
NIEHS-NCATS-UNC Toxicogenetics	2013	Toxicogenenetics	This challenge is designed to build pr	
Whole-Cell Parameter Estimation	2013	Whole-Cell Parame	The goal of this challenge is to exp	DF
HPN-DREAM Breast Cancer Network Inference	2013	Network Inference	The overall goal of the Heritage-DRE	
Rheumatoid Arthritis Responder	2014	Arthritis	The goal of this project is to use a	Ar
ICGC-TCGA DREAM Genomic Mutation Calling	2014	Mutation Calling	The ICGC-TCGA DREAM Genomic M	
Acute Myeloid Leukemia Outcome Prediction	2014	AML, DREAM	The AML Outcome Prediction Ch	M
Broad-DREAM Gene Essentiality Prediction	2014	Targeted Cancer Th	The goal of this project is to use a	CT
Alzheimer's Disease Big Data	2014	Alzheimers	The goal of the Alzheimer's Disea	Al
DREAM Olfaction Prediction	2015	Olfaction	The goal of the DREAM Olfaction	IFF
Prostate Cancer	2015	Prostate Cancer	This challenge will attempt to imp	Pr
ALS Stratification Prize4Life	2015	ALS	As illustrated by the overview figu	Pri
AstraZeneca-Sanger Drug Combination Prediction	2015	Drug synergy	To accelerate the understanding c	As
SMC-DNA Meta	2016	SMC-DNA	The goal of this Challenge is to ide	Or
SMC-Het	2016	Somatic Mutation C	The ICGC-TCGA DREAM Somatic	Pr
Respiratory Viral	2016	Respiratory Viral	Respiratory viruses are highly infe	DA
Disease Module Identification	2016	Module Identificatic	The Disease Module Identificatio	Sy
ENCODE	2016	Transcription Facto	*Transcription factors (TFs) are regulator*	
Idea	2016	Idea	The DREAM Idea Challenge is design	
SMC-RNA	2016	Somatic Mutation C	The ICGC-TCGA DREAM Somatic NS	
Digital Mammography	2017	Mammography	The Digital Mammography DREA	La
Multiple Myeloma	2017	Myeloma, DREAM,	Multiple myeloma (MM) is a cance	Ce

schema.org data models

schema.org

Thing

Thing

The most generic type of item.

Property	Expected Type	Descrip
Properties from Thing		
additionalType	URL	An addi types fr relation syntax, – for m underst
alternateName	Text	An alias
description	Text	A descr

MIAC schematized data model

Attribute	Description	Valid Values	Requires	Properties	Required
Challenge Organizing Group	The name of the Challenge organizing group	BioCreAtivE, CAFA, CAGI, CAMDA, CAMEO, CAMI, C	Organizing Group URL, Organizing		TRUE
Organizing Group URL	URL to the Challenge organizing group homepage				FALSE
Organzing Group Summary	A summary of the Challenge organizing group (e.g., description, mission, focus areas)				FALSE
Organzing Group Class	Defines the Challenge organizing group as p	Academic, Industry			FALSE
Challenge	The name of the Challenge for which all other Challenge details will be tied. A Challenge can have multi		Challenge Keywc	Filename, File F	TRUE
Challenge Keywords	A non-restricted list of keywords in array format that will be used for searching the registry of Challenges.				TRUE
Challenge Summary	The high-level description of a Challenge aimed to address a key research topic.				TRUE
Challenge Sponsors	List if the Challenge sponsors as an array of strings				TRUE
Challenge URL	URL to the Challenge homepage				TRUE
Challenge Organizers	List of the Challenge organizers and their associated institutional affiliations. Names and institutions will be comma separated and lists will				TRUE
Challenge Participants	List of registered particpants				FALSE
Has Subchallenge	Indicates if this challenge has a subchalleng	Yes - has subchallenge, No			FALSE
Yes - has subchallenge	Challenge has a subchallenge		Subchallenge Status, Subchallenge Start Date, Subchallenge End Date, Su		
Subchallenge	The specific question that is being addressed. Each Subchallenge will have its own data and results.				TRUE
Subchallenge Status	The status of the Challenge, which can be o	Open, Closed, Preparing			TRUE
Subchallenge Start Date	The launch date of the subchallenge				TRUE
Subchallenge End Date	The completion date of the subchallenge				TRUE

[MIAC Schemas]

REGISTRY OF OPEN COMMUNITY CHALLENGES

Search...

Introduction

Challenge

POST Add a challenge

GET List all the challenges

GET Get a challenge

DEL Delete a challenge

Grant

Organization

Person

Tag

Documentation Powered by ReDoc

REQUEST BODY SCHEMA: application/json

- name (required) — string — The challenge name
- startDate (required) — string <date> — When the challenge started
- endDate (required) — string <date> — When the challenge ended
- url — string <uri> — The URL to the challenge website
- status (required) — string (ChallengeStatus) — Enum: "upcoming" "open" "closed" — The status of the challenge
- tags — Array of strings (TagId) — The tags associated to the challenge
- challengeResults — object (ChallengeResults) — The results of a challenge
- organizers — Array of strings (PersonId) — The organizers of the challenge

Responses

> 201 Success

Payload

Content type application/json

Copy Expand all Collapse all

```
{
  "name": "Awesome Challenge",
  "startDate": "2020-11-10",
  "endDate": "2020-12-31",
  "url": "https://synapse.org/awes
  "status": "open",
- "tags": [
    "awesome-tag1",
    "awesome-tag2"
  ],
- "challengeResults": {
    "nSubmissions": 0,
    "nFinalSubmissions": 0,
    "nRegisteredParticipants": 0
  },
- "organizers": [
    "507f1f77bcf86cd799439011"
  ]
}
```

Response samples

201 400 409 500

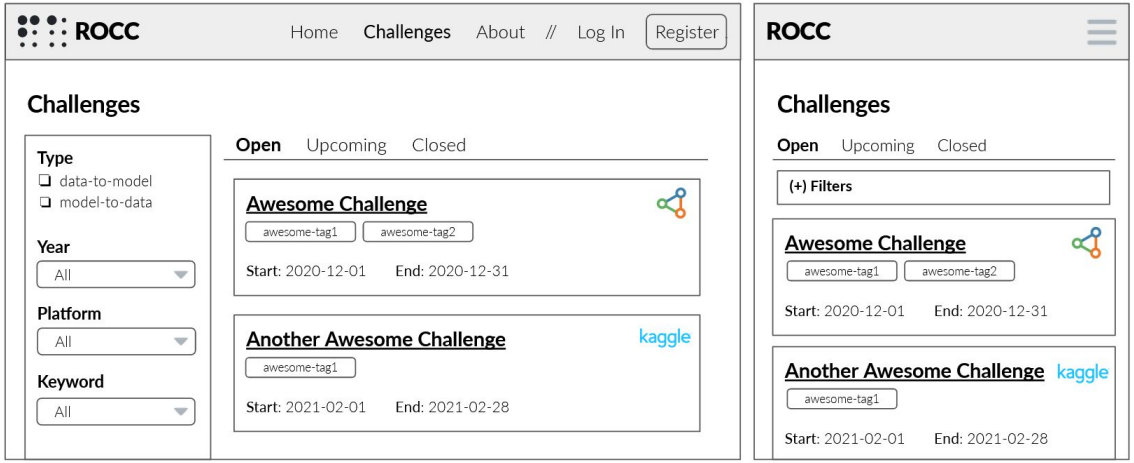
Content type application/json

Challenge Registry OpenAPI specification

[Sage-Bionetworks/challenge-registry]

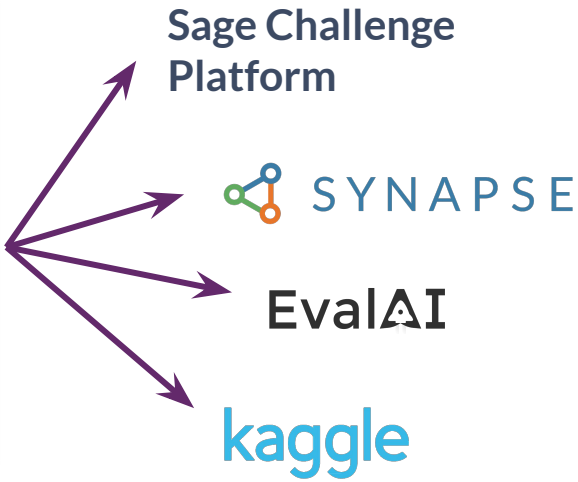
Challenge Ecosystem

Hub for Challenge Exploration and Promotion



Browser *Mobile*

Challenge Registry Wireframes



Challenge Platforms

Understanding Stakeholders

Challenge Participant
"I want to find interesting challenges"

Challenge Organizer
"I want to maximize the visibility of my Challenge"

Organization Manager
"I want consistent communications and Challenges analytics data"

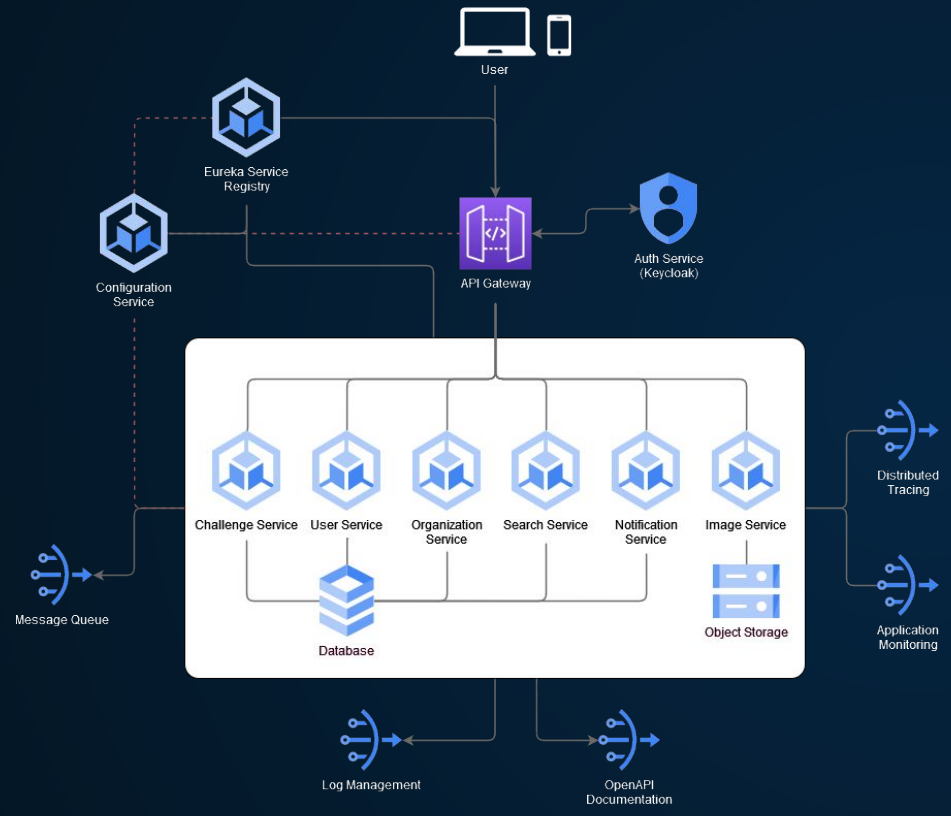
Examples from BraTS 2021

Fereshteh K.
Masters of Medical Radiation
Shiraz University

Ujjwal B.
Postdoctoral Researcher
University of Pennsylvania

Spyros B.
Assistant Professor
University of Pennsylvania

Spyros B.
Organizing Committee
MICCAI



Challenge Registry Architecture

The Challenge Registry is the entrypoint to Sage Challenge Ecosystem where Organizers and prospective Participants connect.

The Registry adopts a microservice and Micro-Frontend architecture.

Thomas Schaffter
Architecture

Challenge Technology Stack

Backend Servers & Microservices (Java - Spring)

Web Apps (TypeScript - Angular)

Portable Apps & Services (Docker)

Uniform Dev Environment (VS Code)

UI Mockups (Figma - teleportHQ)

OpenAPI for Code Generation & Documentation

Data Storage (SQL / Mongo / Elastic)

IAM (Keycloak)
Log Management (ELK Stack)
Monorepo (Nx)
Message Queue (RabbitMQ)
App Monitoring (Prometheus & Graphana)
Distributed Tracing (Zipkin?)

Our development workflow is built upon modern best practices (e.g. using code generators) to **accelerate development and improve collaboration**.

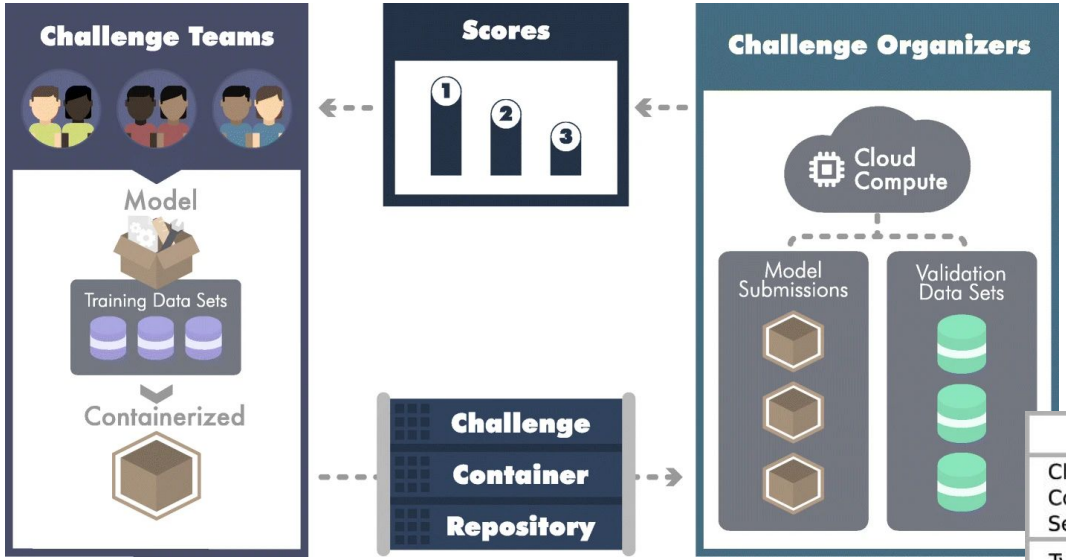
FAIR (Findable, Accessible, Interoperable, Reusable) • Standardization • Open source • Portable

Thomas Schaffter
Lead Software Developer

Evaluation Orchestration

SageBionetworks

Model-to-Data for Benchmarking



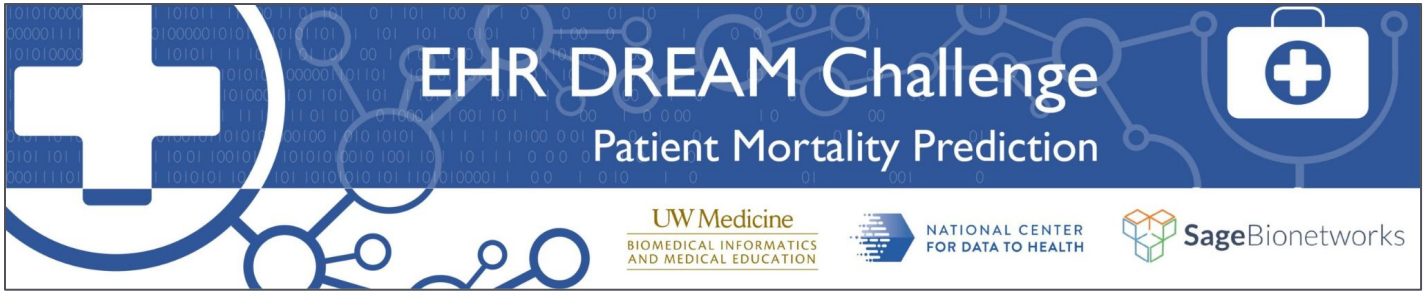
- Permits use of sensitive or proprietary data
- Preserves integrity of gold-standard validation data
- Algorithm reproducibility and re-usability
- Prospective assessment

	SMC-Het	SMC-RNA	MM	DM	Proteo
Cloud Compute Service	Openstack, Google Cloud Platform	Google Cloud Platform	Amazon Web Services	IBM Cloud, Amazon Web Services	Amazon Web Services
Type of Compute	CPU	CPU	CPU	GPU	CPU
Data Type	Simulated, Real	Simulated, Real	Real	Real	Simulated, Real
Data	Genomics	Genomics	Clinical, Genomics	Clinical, Imaging	Genomics
Model Form	Galaxy, Docker	Common Workflow Language, Docker	Docker	Docker	Docker

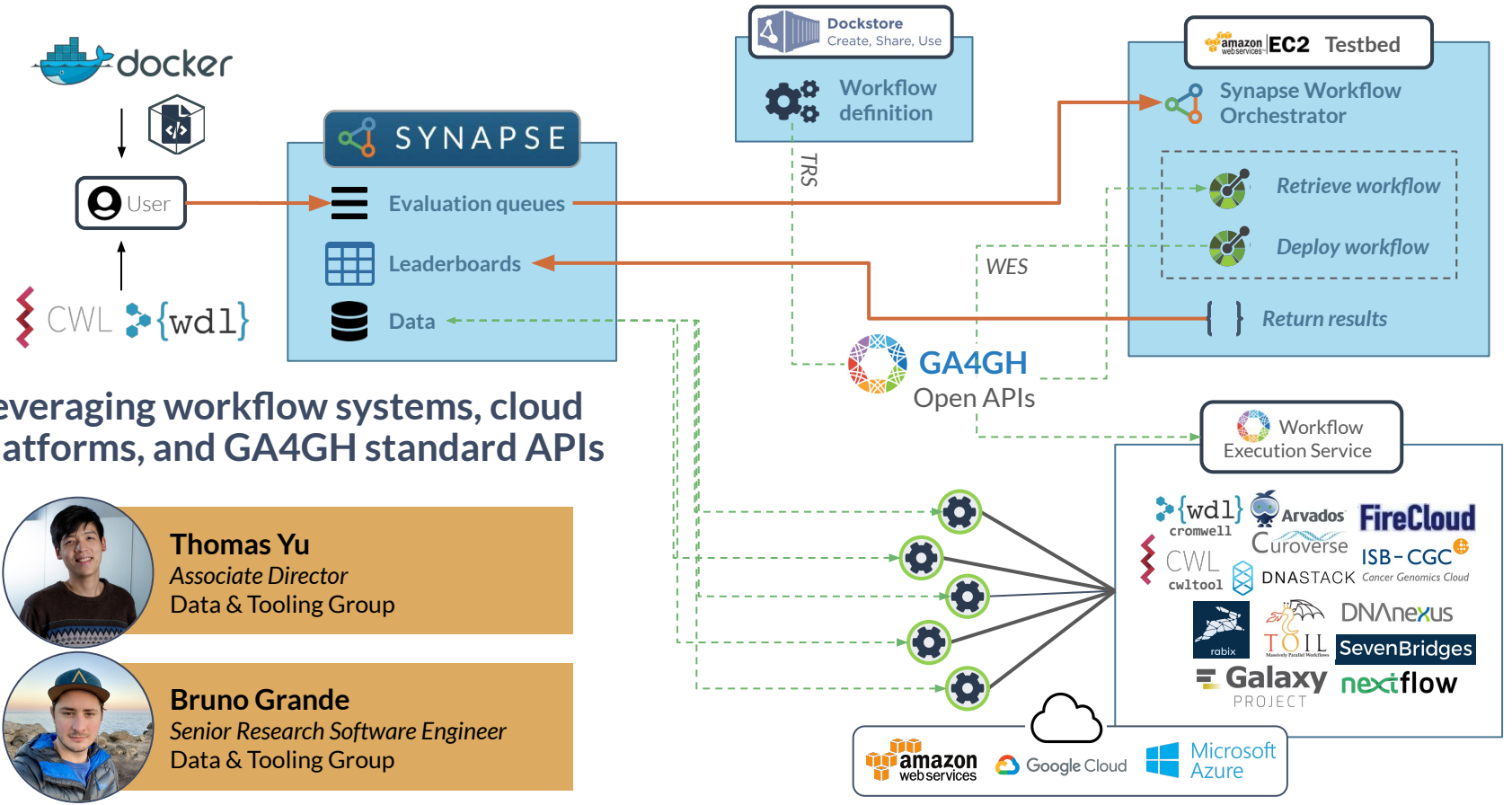
Images from Ellrott, et al., Genome Biol (2019):
[Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges]

SageBionetworks

Federated Model Evaluation in Clinical Contexts



Model-to-Data: Why not Workflows?



Leveraging workflow systems, cloud platforms, and GA4GH standard APIs

Thomas Yu
Associate Director
Data & Tooling Group

Bruno Grande
Senior Research Software Engineer
Data & Tooling Group

Next Steps

SageBionetworks

Ongoing Challenges & Collaborations

Roadmap for a Challenge Platform 2.0

- **Moving from design to implementation:** developing prioritized components and services for Challenge Platform user stories
- **Building out a Challenge Toolkit:** QC approaches, evaluation metrics, bootstrapping procedures, etc. — standardizing and packaging for reuse
- **Enabling scalable model evaluation:** deployment and hardening of workflow orchestration infrastructure for challenges
- **Leveraging cloud workbench platforms:** connecting to third party systems (e.g., Terra, Seven Bridges / Cavatica, DNAnexus) for workflow execution
- **Improving robustness for federated training and evaluation:** incorporating tools like ***FeTS.ai*** and others into Challenge Toolkit

Thanks!

Acknowledgements

Sage Bionetworks

- James Eddy, *Data & Tooling*
- Jake Albrecht, *Challenges & Benchmarking*
- Thomas Schaffter, *Data & Tooling*
- Thomas Yu, *Data & Tooling*
- Verena Chung, *Challenges & Benchmarking*
- Rong Chai, *Challenges & Benchmarking*
- Timothy Bergquist, *Challenges & Benchmarking*
- Aarthi Krishnan, *Challenges & Benchmarking*
- Amy Heiser, *Challenges & Benchmarking*
- Kevin Boske, *Platform Product Management*
- Ljubo Bradic, *Platform Design*

UCLA

- Paul Boutros
- The Boutros Lab

Collaborators

- Spyros Bakas @ UPenn
- Ujjwal Baid @ UPenn

Project Scientist

- Keyvan Farahani @ NCI

NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

U24-CA248265

SageBionetworks

UCLA David Geffen School of Medicine

Extras

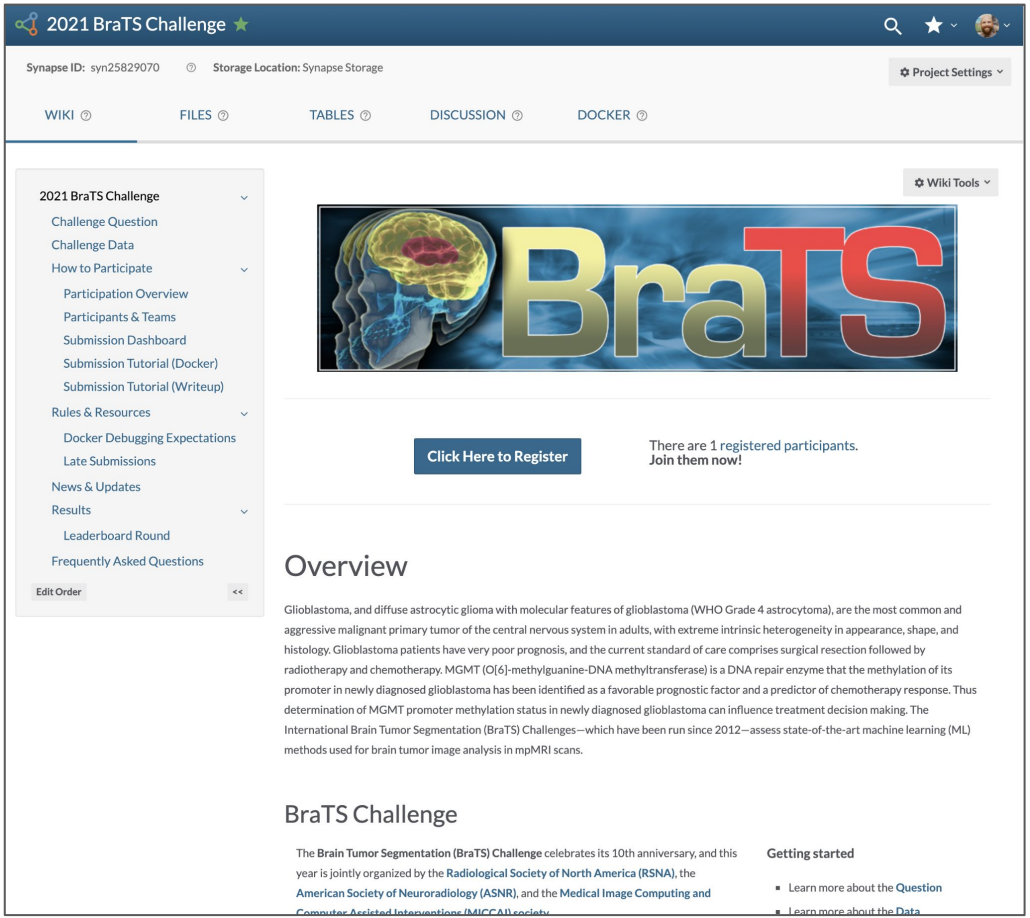
RSNA-ASNR-MICCAI BraTS Challenge on Synapse

SYNAPSE

Your go-to source for all things **BraTS Challenge**:

- Registration
- Documentation & instructions
- Challenge data files & tables
- Background info & references
- Discussion forums
- Leaderboards

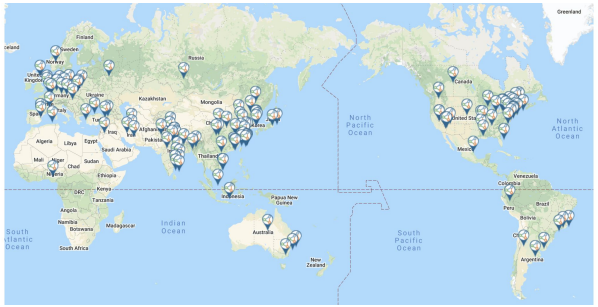
synapse.org/BraTS2021



SageBionetworks

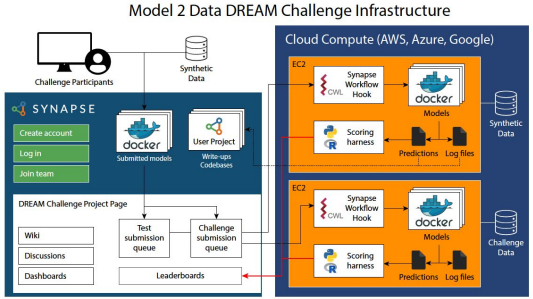
2021 BraTS at a Glance

Community



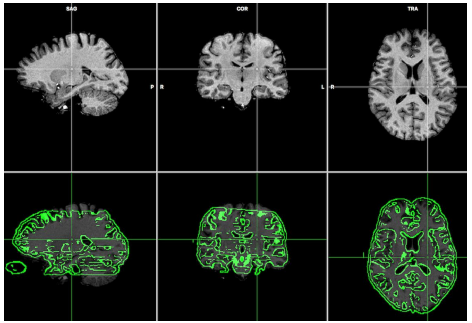
- 900+ participants across 5 continents
- 1400+ submissions scored
- 83 short papers submitted

Infrastructure



- **GPU-supported evaluation infrastructure**
- Validation phase used traditional challenge framework
- Final Docker submission phase uses model-to-data framework

Data



- *Nii.gz imaging files
- Nearly 2000 cases!
  - 1251 for Training
  - 219 for Validation
  - 570 for Testing

SageBionetworks