

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

_____ *Sean Mooney* _____

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



CAGI Challenge Format

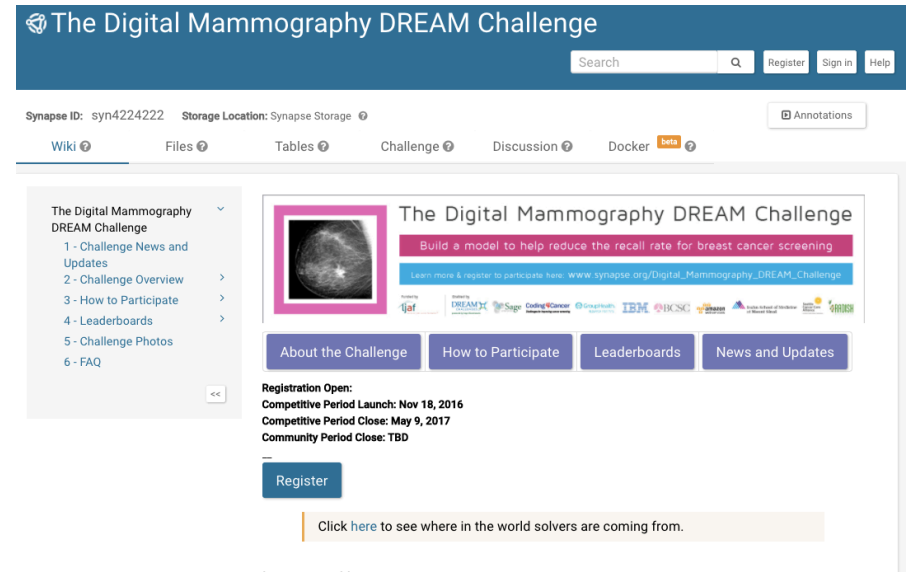
- CAGI challenges have so far involved releasing data when an answer is known but not shared until after assessment
 - These challenges are ‘retrospective’ since the answers are usually already collected
 - These challenges are ‘data to method’ since the data is provided to the method developers for prediction and training

Prospective challenges have been successful elsewhere

- For example, the Critical Assessment of Functional Annotation (CAFA) is a prospective challenge.
- Predictions on entire databases are collected and then assessed based on accumulated discoveries over time

‘Method to Data’ challenges enable assessment on private datasets

- Sage Bionetworks has been a leader in establishing the ‘method to data’ challenge
- Ex. Digital Mammography Challenge



A challenge is proposed based on ‘secret’ curated data

Challenge submitters submit solutions to the problem in the form of methods

The challenge assessors or providers apply the methods to the ‘secret’ data

Best solutions may or may not be made available for use

Pros and Cons

- There are many pros to this model:
 - Some data that may be unavailable could be used for a challenge
 - Truly unbiased over the long term (some biases early)
 - Has benefits over over time and not a 'one off'
- There are some cons
 - Prospective challenges suffer from the 'open world' problem in that interrogated data maybe biased
 - Method to Data challenges may lack training data or access to good training data

A quick plug --> The Center for Data To Health is building prospective, method to data challenges with Sage BioNetworks on EHR datasets



NATIONAL CENTER
FOR DATA TO HEALTH

dbNSFP Challenge

- CAGI has proposed a challenge that was prospective, and was basically a 'method to data' approach where all possible answers are provided to assessors, but the exact target set of data is unknown to predictors
- My understanding is 5 teams submitted, but assessment would happen between CAGI 5 and 6, presumably
- We can discuss this!

Possible data sets that could be used for assessment

- ClinVar pathogenic variants over time
- Clinical human genomes/exomes where underlying data has been discovered
- Biobank and EDW datasets
- Disease specific datasets
- Other specific genetic studies that are otherwise unshare-able

Discussion

- Panelists: introduce themselves and their submission and their interest in this
- What are barriers to this challenge?
- What are some datasets that could be used initially for assessment?
- What does CAGI community want to see with such a challenge?
- How often should assessment occur?