NeurIPS Privacy Challenge

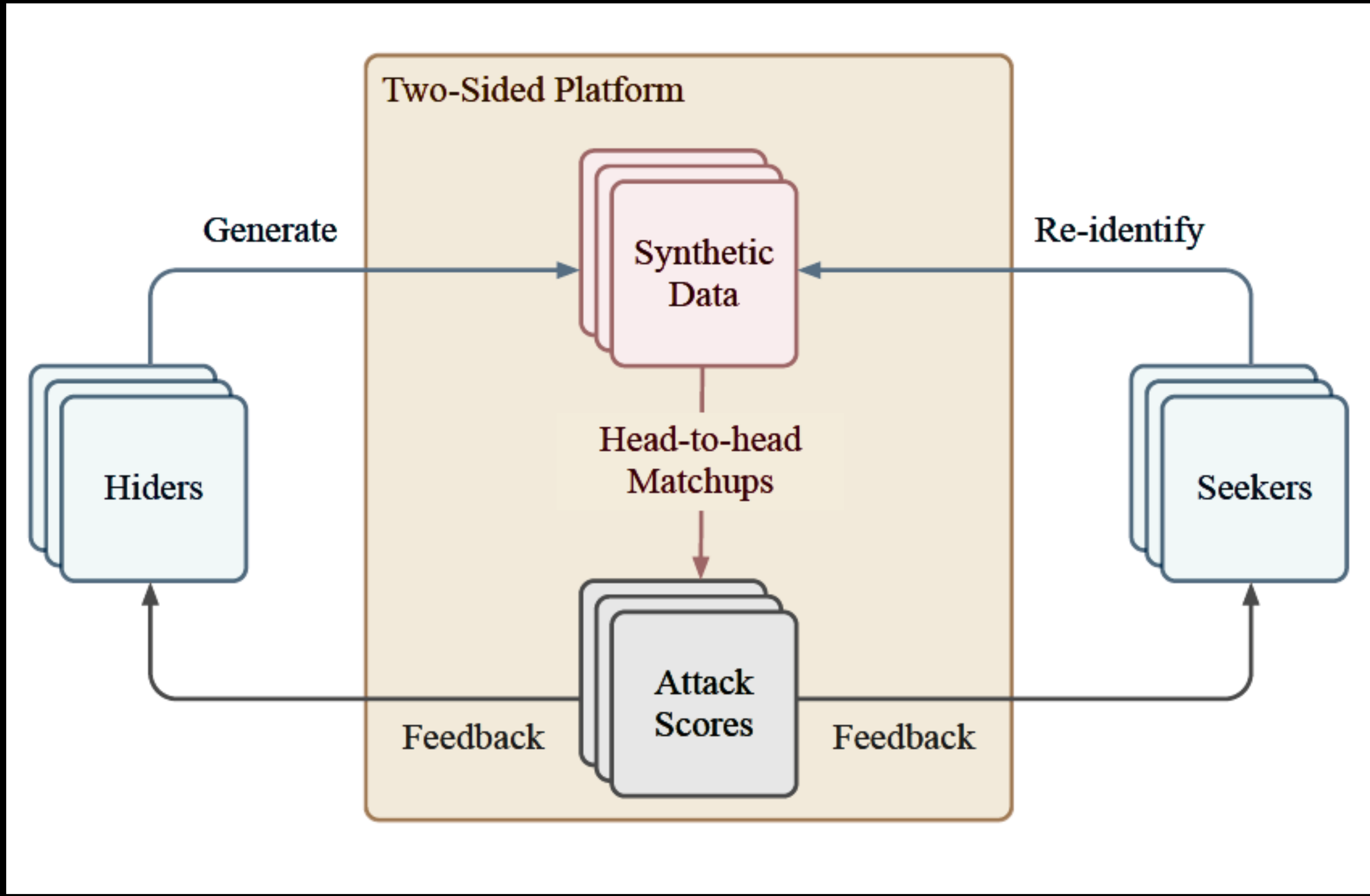
https://www.vanderschaar-lab.com/privacy-challenge/

…as an objective for adversarial testing UoE for iCAIRD and as a means of establishing a collaboration.

Marija Jegorova PostDoc@UoE: m.jegorova@ed.ac.uk
Sotirios Tsaftaris Professor@UoE: S.Tsaftaris@ed.ac.uk

Quick summary:

- **Time:** 1st of July to 1st of October
- **Dataset** (AmsterdamUMCdb) is available online - 1bn ICU data samples from over 20k patients. https://github.com/AmsterdamUMC/AmsterdamUMCdb/wiki
- **Novelty:** universal membership inference attacks/defences (patients re-id) on generative models.
- **Two tracks:** hiders and seekers



Scoring criteria:

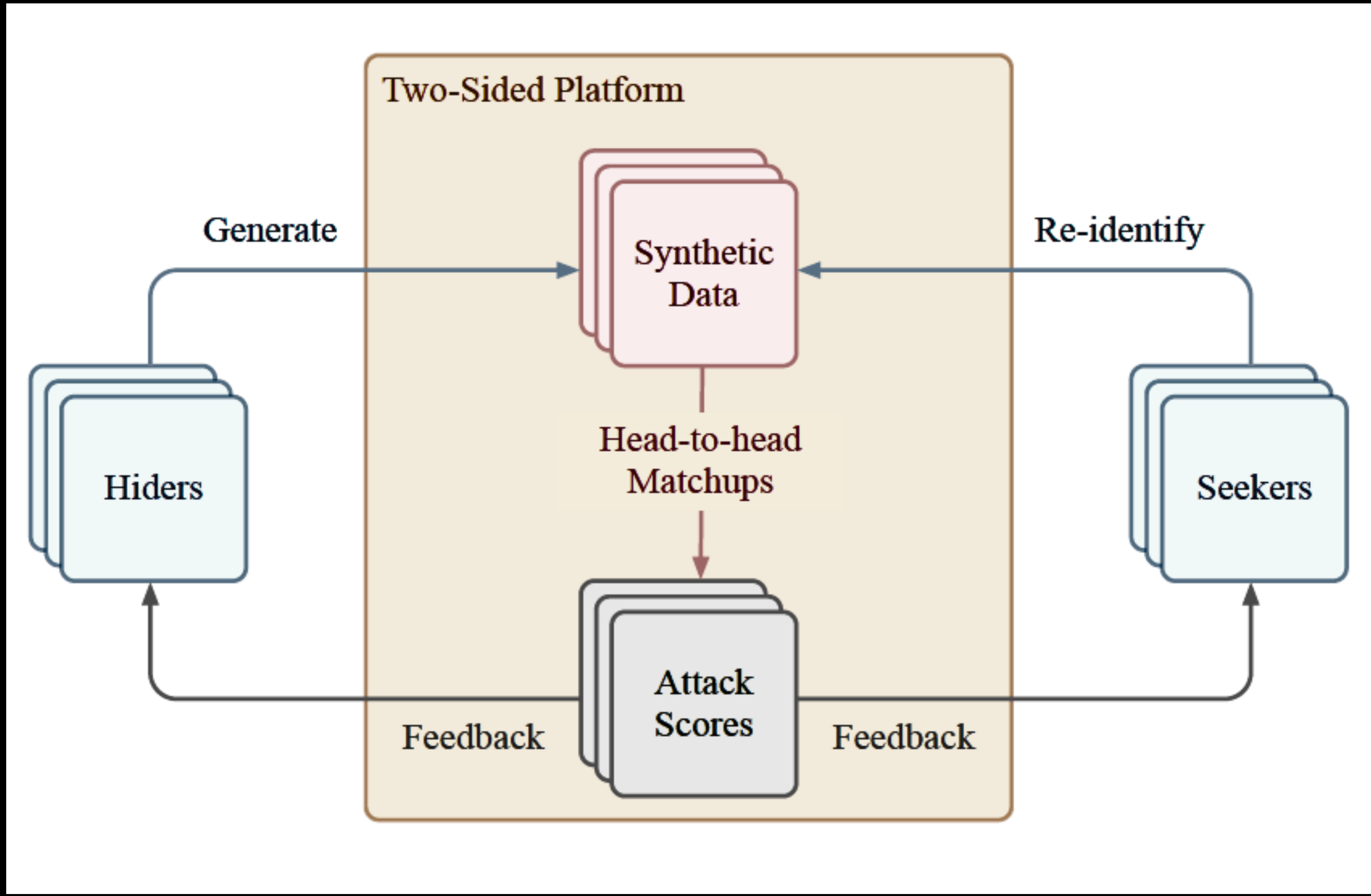
- **Hiders:** score on how well their generation algorithms hold up to membership inference attacks.

  + adequately *capture the feature and temporal correlations* in the original data;

  + *pass a minimum quality bar* (in terms of fidelity and predictivity)

- **Seekers:** scored on their accuracy at the membership inference task over each hider submission

  (in correctly identifying whether a given instance was employed in the process of generating of a given synthetic dataset)



Benefit: What is the benefit to iCAIRD and the current work-package?

Globally:	**For iCAIRD:**
An opportunity to *compete against and learn about the state-of-the-art attacks/defences*, without having to implement all of them (competitors will).	A perfect platform for data collection on *the state-of-the-art attacks/defences* for generative models, which is in the agenda of the current WP.
<u>The minimum outcome:</u> *aggregated knowledge* about how different state-of-the-art attacks/defences compare in an identical setting.	<u>The minimum outcome:</u> a complete report for WP UoE I, on attacks/defences on generative models, in the medical data synthesis setting. Information on generative model defences for WP UofG II.
<u>The ideal outcome:</u> a universal attack against generative models, that would flush out the data used for training. And potentially a metric assessing remembered vs. synthesised by the model.	<u>The ideal outcome:</u> a universal attack for data synthesis models, determining data leakages. Potentially, a universal metric providing quantitative assessment of data leakage, to help the client to improve their model with respect to such metric.

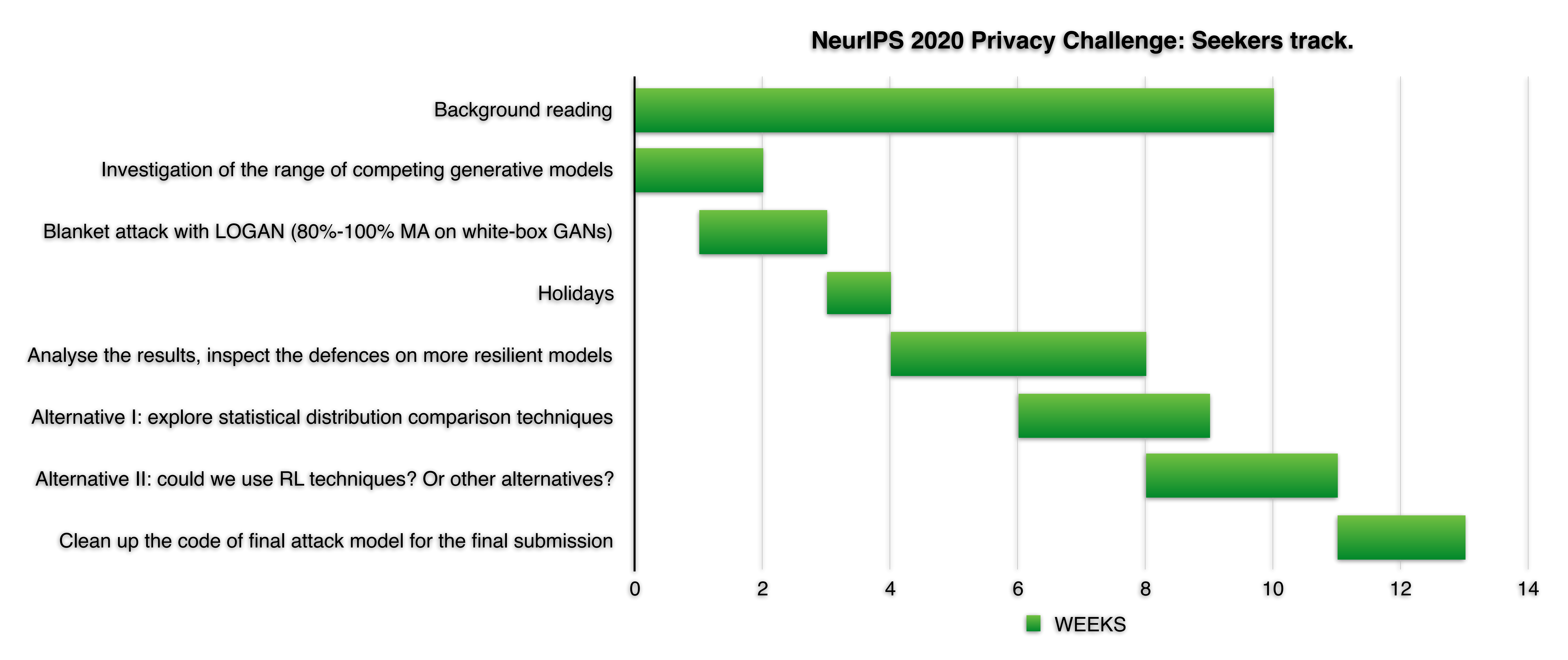
Preliminary Assessment:

- The attack task seems to be **white-/grey- box**.
- Literature review suggests that:
  - there is **a range of existing attacks** for each colour of the box.
  - we would have to come up with **a universal attack / defence** to be successful.
  - we might have to come up with the new attacks for new defences.
- **Starting point:** we can start with LOGAN/GAN-Leaks, and carry on to investigating statistical comparisons of the output distributions.
- (**NB!** No code available in OpenAccess, so would need to re-implement)
- Not all the generative models will be GANs - so statistical approach is better, as more universal for the problem.

	Latent code	Gen-erator	Dis-criminator
[38] full black-box	×	■	×
[41] full black-box	×	■	×
Our full black-box (Sec. 4.2)	×	■	×
Our partial black-box (Sec. 4.3)	✓	■	×
Our white-box (Sec. 4.4)	✓	□	×
[38] accessible discriminator	×	×	✓

Table 1: Taxonomy of attack settings against GANs over the previous work and ours. (×: without access; ✓: with access; ■: black-box; □: white-box). The settings are more and more knowledgeable to attackers from top to bottom.

Approximate Timeline:



Collaboration / competition for this challenge:

- It is an important and non-trivial task on both tracks.
- We think it might be a great way of giving a kick-start to a collaboration between UoE and UofG.
- It is an important direction to be able to assess the risks of patients re-identification in generative models, and this challenge is a great opportunity to do it in a time-efficient manner.
- https://www.vanderschaar-lab.com/privacy-challenge/