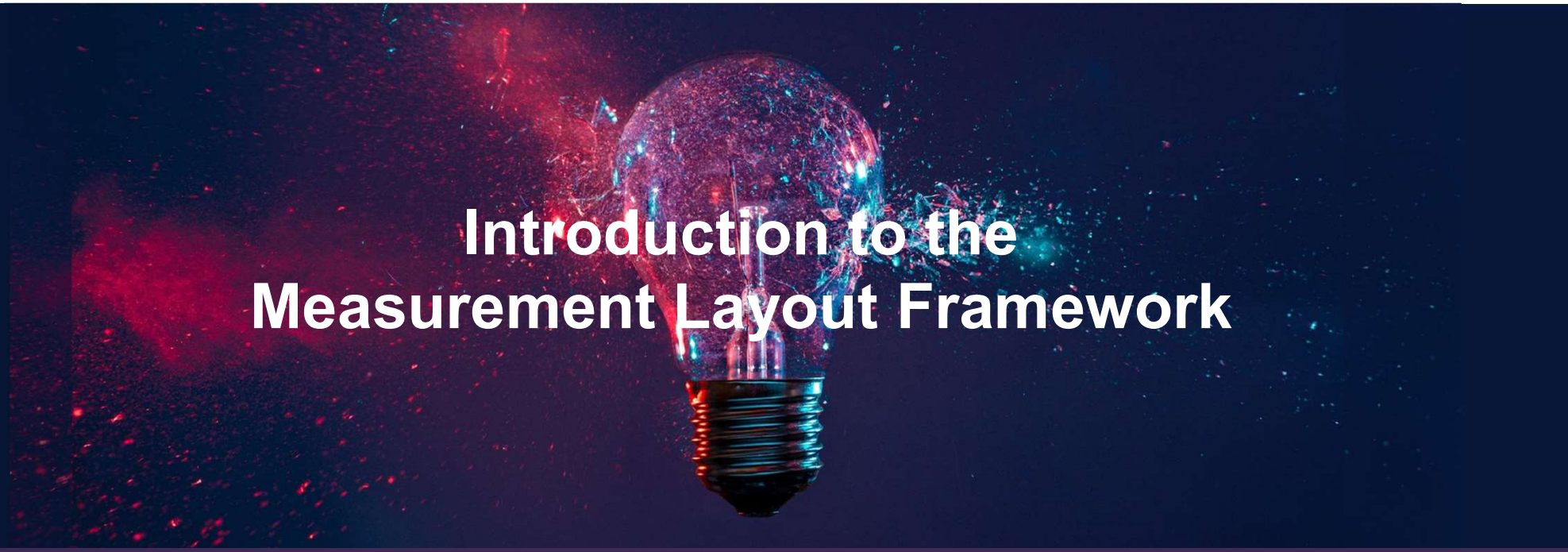
UNIVERSITY OF CAMBRIDGE



Introduction to the Measurement Layout Framework

John Burden

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Vancouver, 20 February, 2024

Leverhulme Centre for the Future of Intelligence,
University of Cambridge,

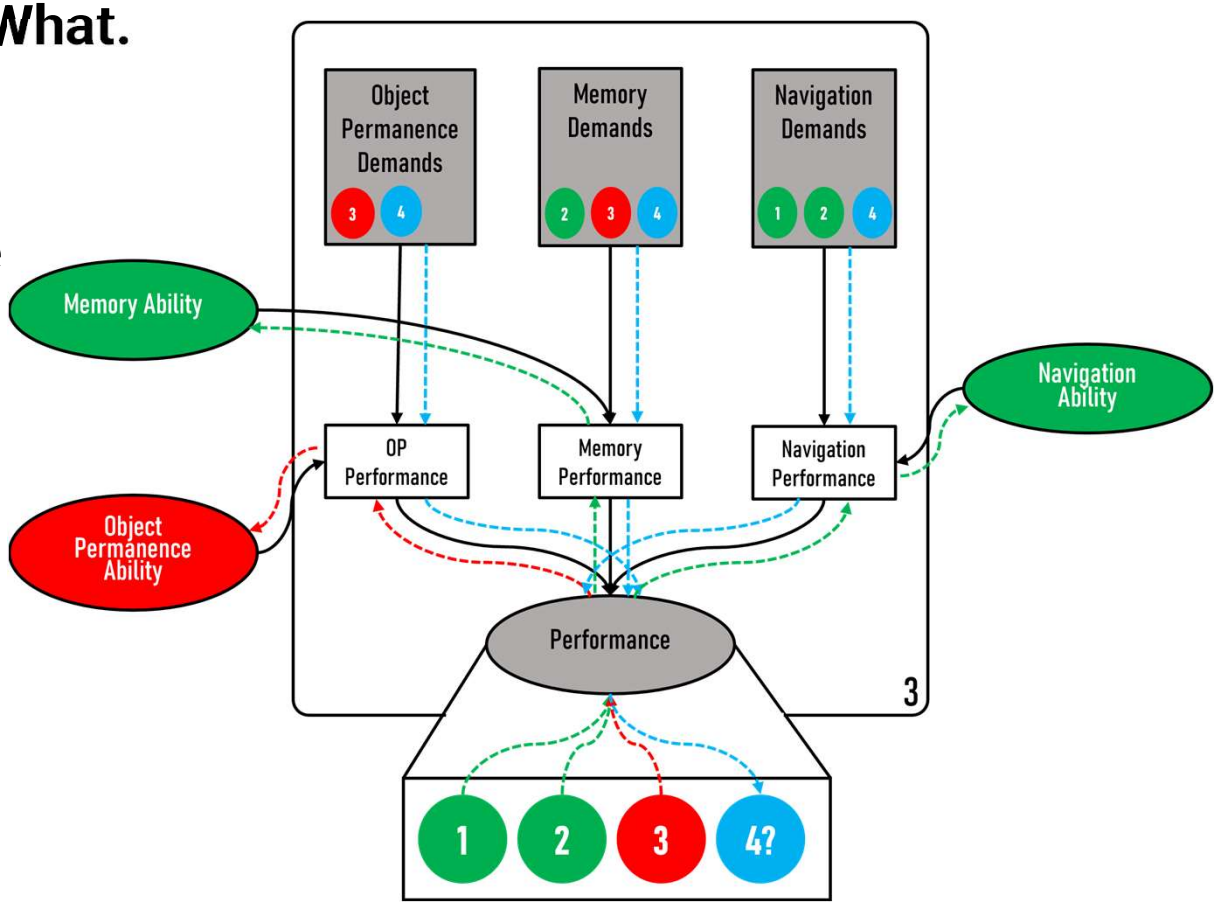
This Session

- Introducing the Measurement Layout Framework:
  - What are they?
  - How are they used?
  - Why do we think they can provide capability-oriented evaluation?

- Hands on experience learning to construct them.

Measurement Layouts – The What.

- If we know the demands of a set of tasks or task instances and how they are related, we can build a Bayesian model that connects system capabilities and task demands.
- From a collection of instances, we can infer capabilities.
- Then we can predict future performance for new instances.

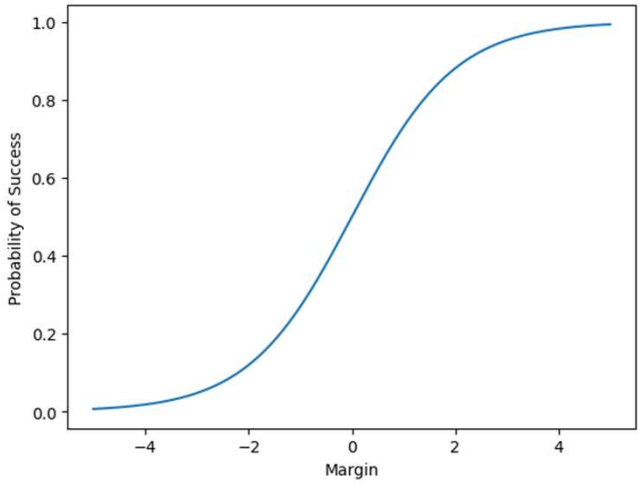


Measurement Layouts – The What

- Increased capabilities make success more likely, while increased demands decrease likelihood of success.
- For a single demand and capability:

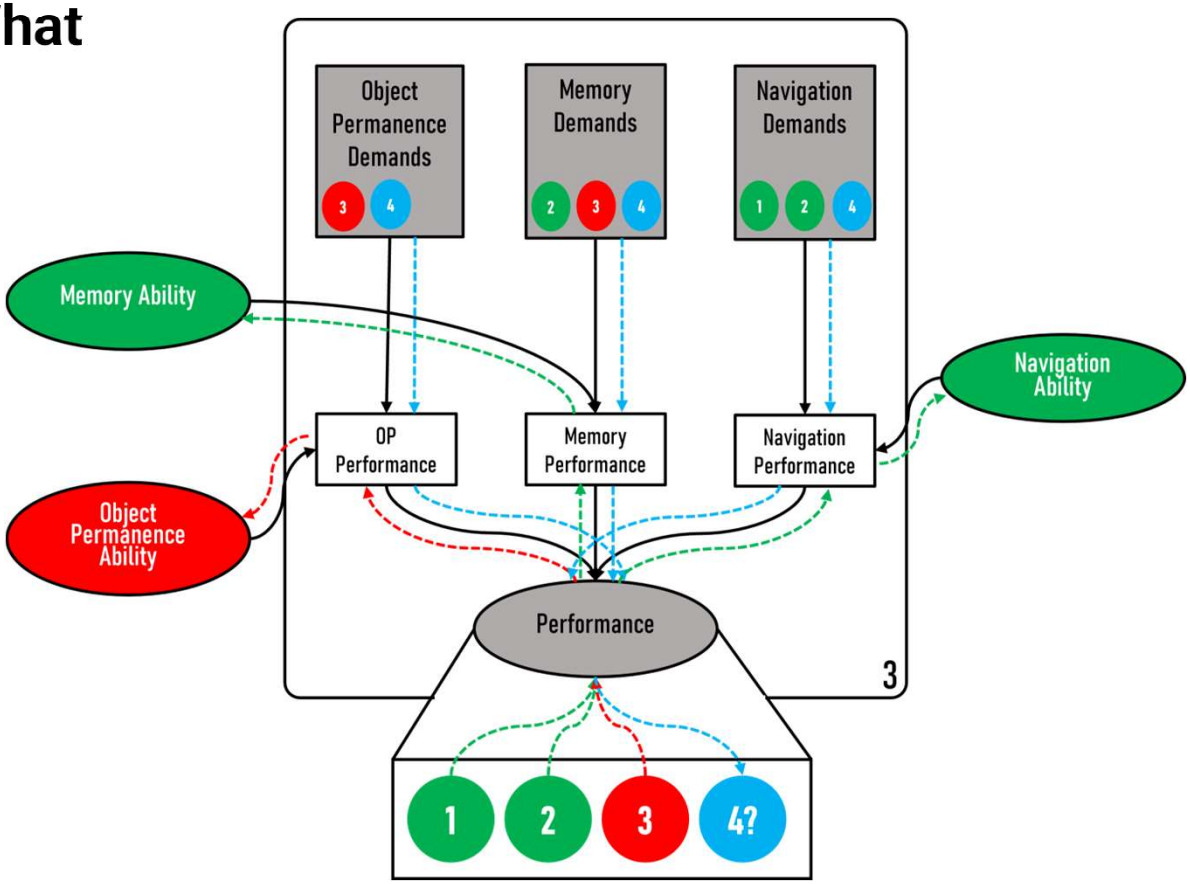
$$P(\mathrm{s}|\mathrm{capability}, \mathrm{demand}) = \sigma(\mathrm{capability} - \mathrm{demand})$$

- The difference between capability and demand is referred to as the *margin*.



Measurement Layouts – The What

- More than a single ability and demand.
- Introduce *partial performance* nodes.
- Final performance will depend on some combination of these.



Measurement Layouts – The What

- The connective structure between nodes can encode knowledge about the task.
- Complex connective structure can arise from the need to accurately distinguish between success, failure, confounding behaviour.
- We refer to the function describing these relationships as *Linking Functions*.
- LFs inform how partial performance is assessed.

Measurement Layouts – The What

- LFs also inform how partial performances are combined to give final performance.
- Some tasks will be non-compensatory. High capabilities in one area will not compensate for lacking in others.
- Some tasks compensatory. One high capability can compensate for low levels in others

Measurement Layouts – The How

- So, we have the structure of a measurement layout in place. What do we do with it?
- Need a data-set with performance outcomes labelled along with demands for each instance.
- Bayesian model needs priors
  - Either encode system knowledge or uninformative.
- We can then leverage a probabilistic programming implementation of the network using PyMC.

Measurement Layouts – The Why

- Why do we expect Measurement Layouts to help provide capability-oriented evaluation.
  - Capabilities are independent of the distribution of tasks instances
  - Allows us to predict future performance at the instance-level. The Measurement Layout models reflect domain expertise
    - Higher internal validity.
    - Enables predictions to be valid even out of distribution.

Exploring Measurement Layouts in PyMC

- https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial