THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Mapping and understanding our Universe using HPC and AI

Cullan Howlett

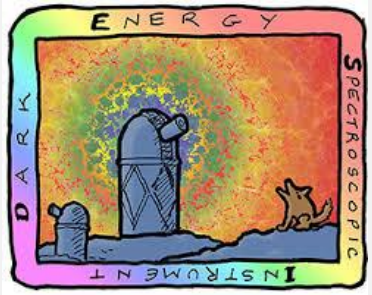
Motivation

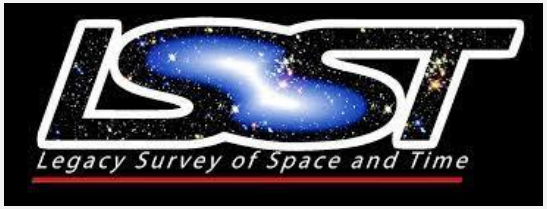
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Modern cosmology surveys aim to understand:

- What are the 'initial conditions' of the Universe
- What are its fundamental components and how have they evolved over the last 13,000,000,000 years.

Australia is involved in numerous large-scale surveys that are mapping the 3D positions of tens of millions of galaxies to answer these questions



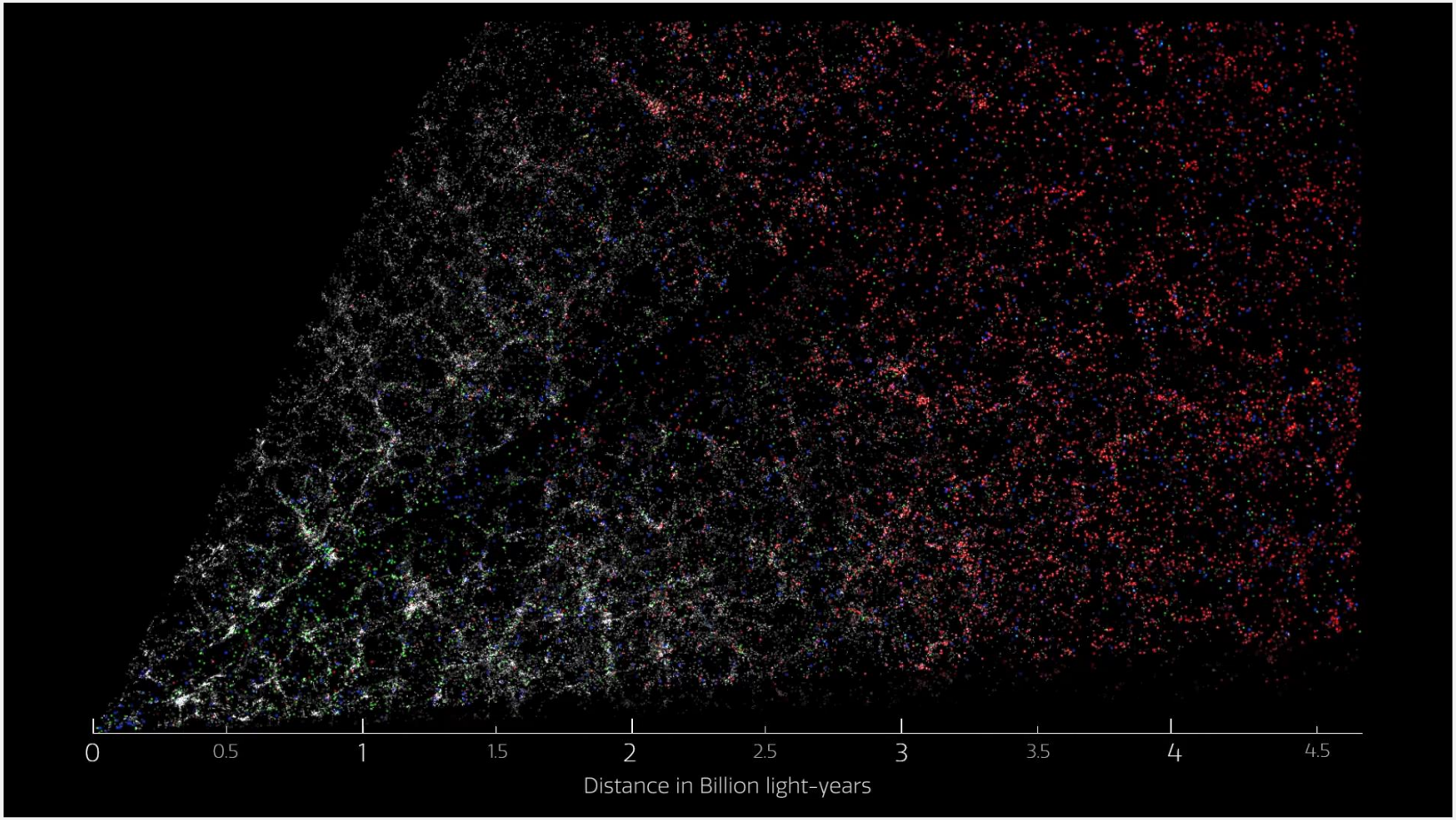


+ Many more...

Methods

The University Of Queensland
AUSTRALIA

The analysis of these maps requires statistical modelling of correlations between millions of objects, often applied to thousands of simulations



Methods

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Let's use DESI as an example of how a typical analysis goes…

1. Observe light of 40,000,000 galaxies and reduce to catalogue of 3D positions.
2. Produce ~10,000 simulations of these galaxies that look like the data.
3. For all 10,001 catalogues, compute the distance between each pair of points.
4. For these catalogues apply 'reconstruction' to move the galaxies back in time.
5. Compute the distance between each pair of points in each of the 10,001 'reconstructed' catalogues.
6. Use variance in 20,000 simulations as errors on the data.
7. Fit the data before and after reconstruction using ~10 different methods and ~5 different models of the Universe.
8. …Profit!!

Clearly a 'Big Data' challenge requiring HPC+AI, so let's explore how…

HPC Usage

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

DESI data reduction runs daily on-the-fly at NERSC-Perlmutter

Partition	# of nodes	CPU	GPU	NIC
GPU	1536	1x AMD EPYC 7763	4x NVIDIA A100 (40GB)	4x HPE Slingshot 11
	256	1x AMD EPYC 7763	4x NVIDIA A100 (80GB)	4x HPE Slingshot 11
CPU	3072	2x AMD EPYC 7763	-	1x HPE Slingshot 11

Partition	Type	Aggregate Peak FP64 (PFLOPS)	Aggregate Memory (TB)
GPU	CPU	3.9	440
GPU	GPU	59.9 tensor: 119.8	280
CPU	CPU	7.7	1536

Our 2022 requirement/allocation were effectively:

- 24M CPU only core hours
- 6.4M CPU core + 400k GPU hours
- 5.6 PB of storage

Measuring the correlations between galaxies

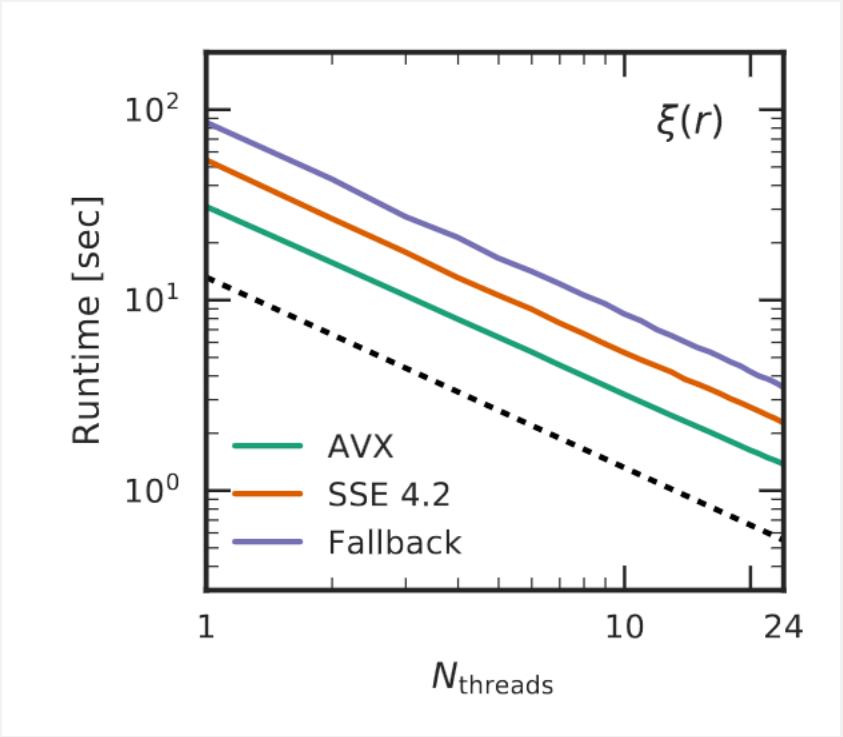
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

At its basic, our statistic of interest just involves counting (many!) pairs.

Large benefit from using better algorithms (kd-trees), distributed computing (MPI), multithreading (OpenMP) and AVX

For current datasets, we can compute these statistics for millions of objects in a ~node-hour on the HPC

Doing this 1000s of times still presents a challenge:
(10M CPU hours in 2023 for DESI)

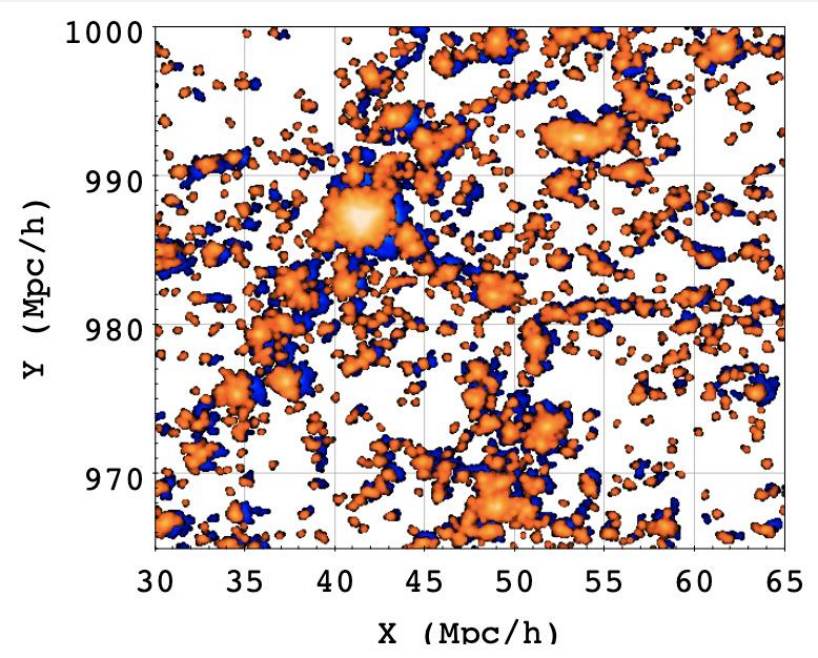


Sinha & Lehman, 2019

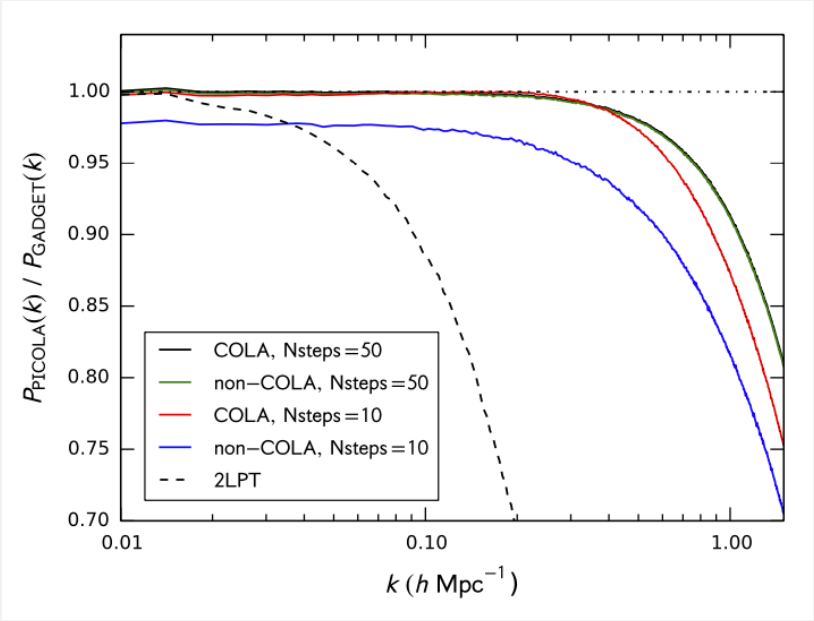
Fast simulation methods

Producing enough simulations to estimate the variance in our measurements requires fast approximate, methods

Howlett et. al., 2015

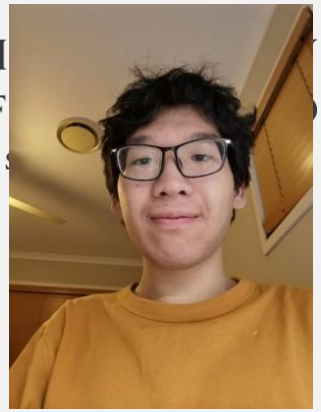


Again, large benefit from using better algorithms (COLA, FastPM), distributed computing and multithreading



These methods can produce large high fidelity gravity-only simulations in a few tens of CPU hours

Clever fitting methods: Compression



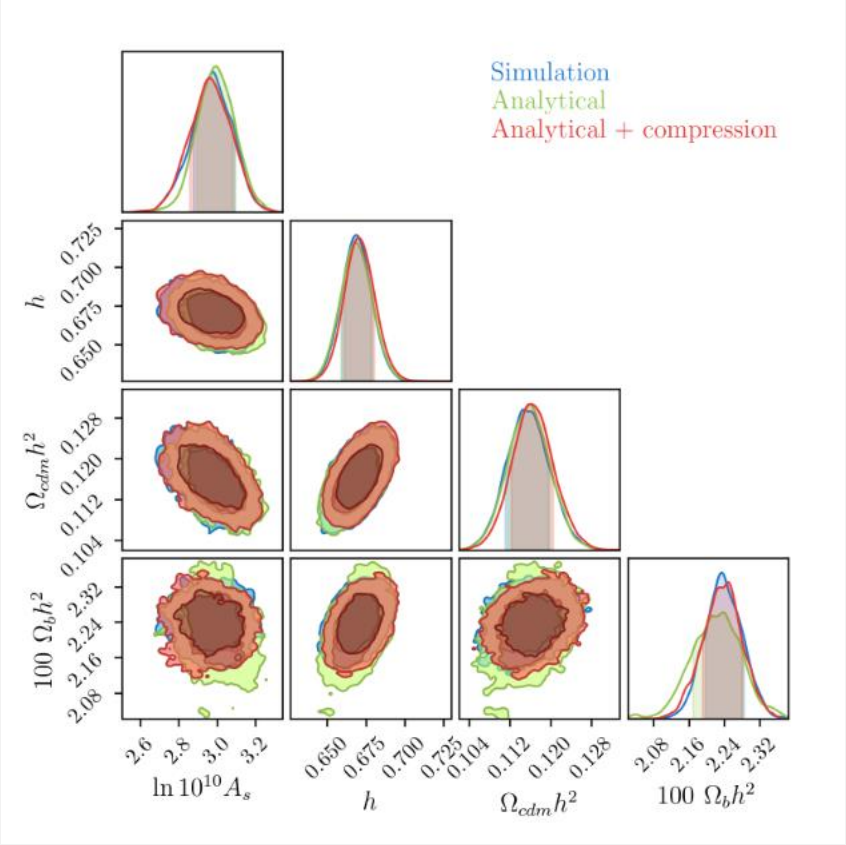
Once we have a statistic and its error from data+simulations, we need to fit cosmological models

Testing different models means this can take a lot of time (3M CPU hours in 2023)

Active area of research is replacing simulations with theory and using compression:

- w/ Theoretical modelling
- w/ Unsupervised learning

Yan Lai et. al., 2023



Clever fitting methods: Emulation

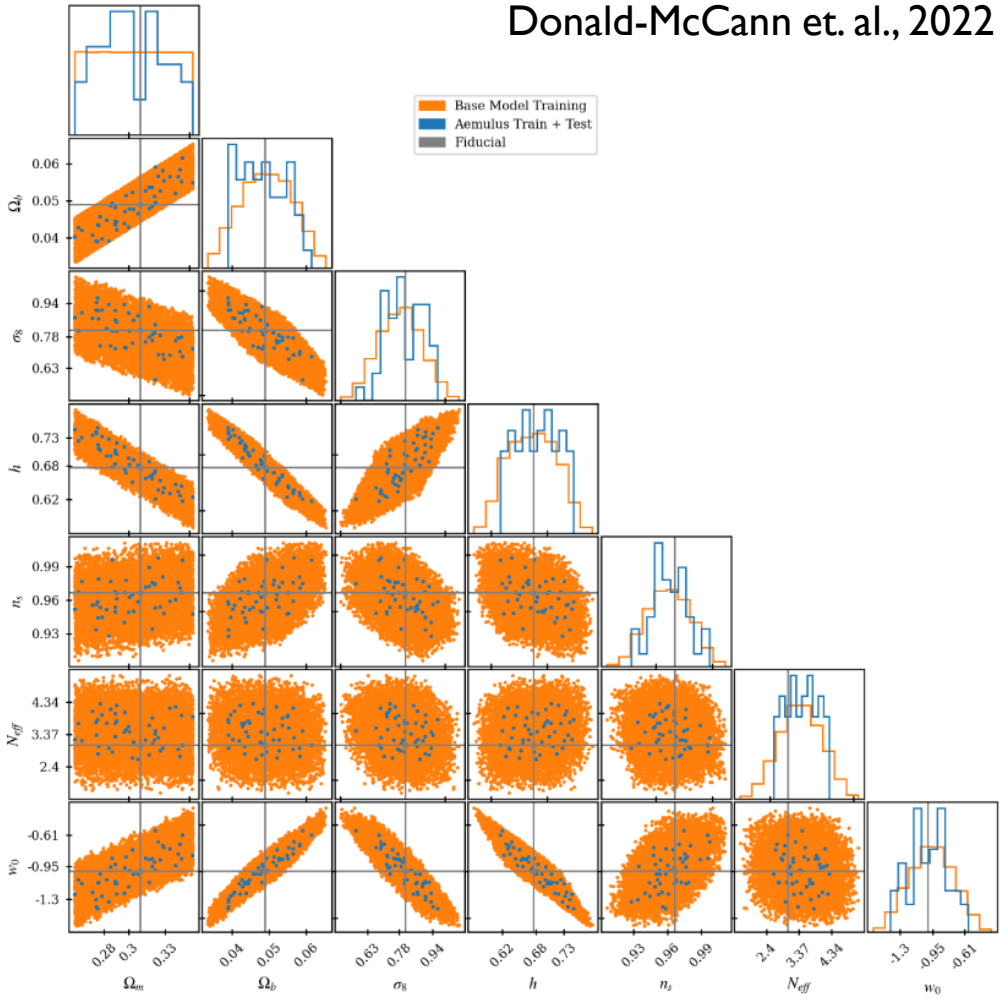
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Even with standard methods, we can speed up fitting by *learning* the model

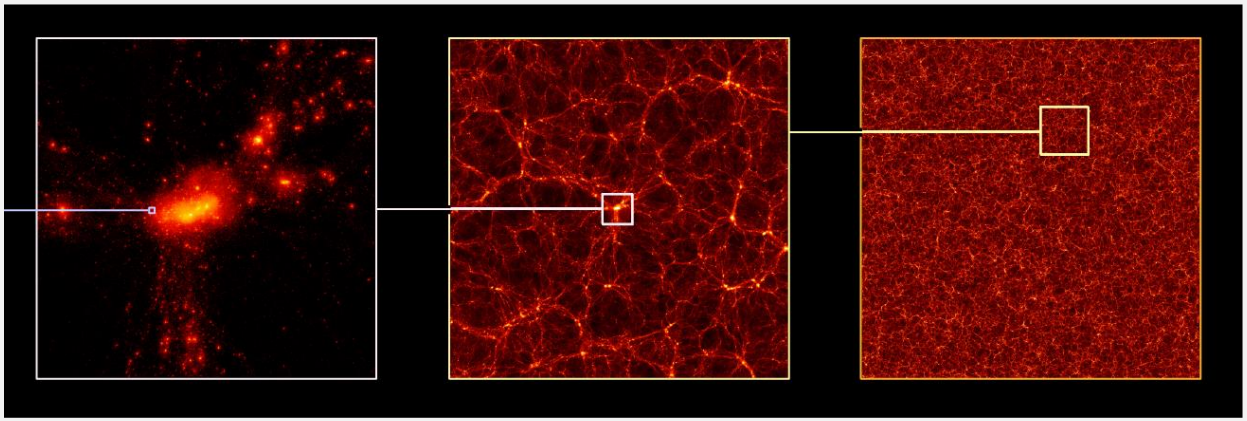
For our current state-of-the-art models of galaxy clustering, these accurate emulators are 1000 times faster.

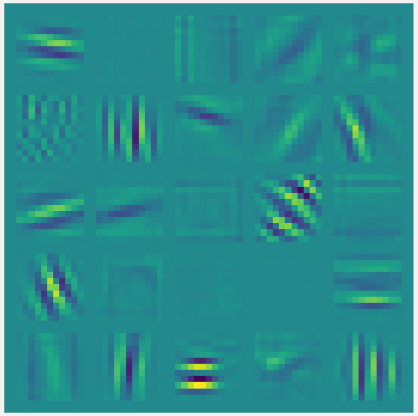
This has significant impact on our ability to fit different models and test systematics.

Donald-McCann et. al., 2022

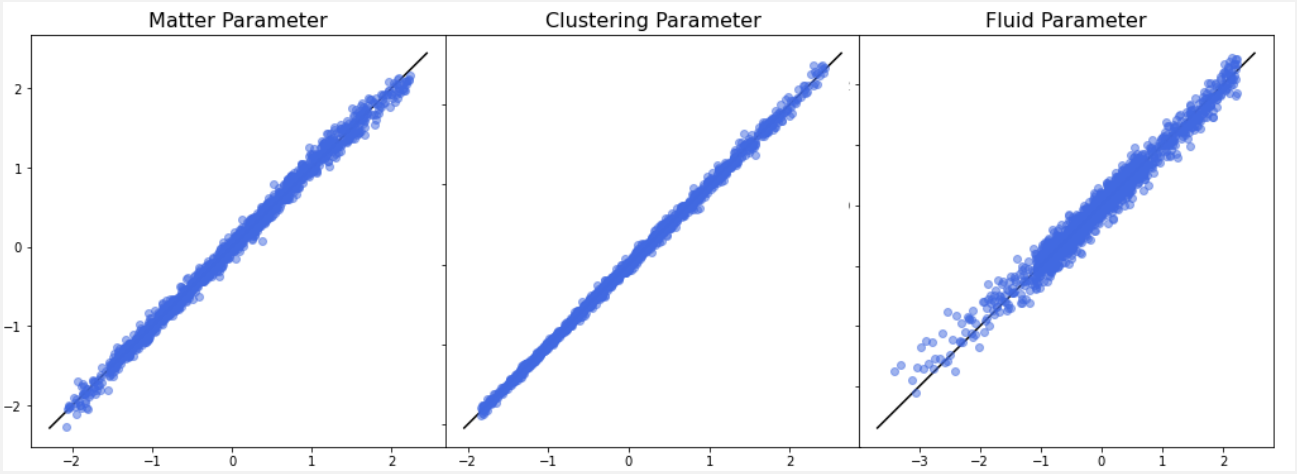


Clever fitting methods: Neural Networks





Replacing standard pair-counting with other *learned* statistics is also very popular



Matt Craigie et. al., in prep.

Clever fitting methods: Emulation

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

HPC and AI use are absolutely essential to dealing with the volume of data from current cosmological galaxy surveys.

- Surveys in Australia and overseas have/will map 3D positions for tens of millions of galaxies
- These provide essential constraints on the origin and evolution of the Universe.
- Analysing this data requires brute-force statistics on thousands of simulations
- To make this tractable, even on the largest HPCs in the world means:
  - SIMD, distributed, and multi-threaded programming
  - Clever methods for simulating the Universe
  - Statistical compression techniques
  - Model emulation and analysis of the data with both supervised and unsupervised learning.

These challenges are not going to go away, so interaction/input/ideas from the HPC/AI communities is essential for cosmology!