# Quantitative Analysis of Job Trace Data

**Thomas MacDougall, UNC Charlotte**
**Dong Dai, Computer Science Department**

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

## Introduction

- Job traces created by HPC's are often considered proprietary data, and so are generally withheld from the public.
- Creating a way of artificially generating job traces may prove beneficial in research and development of HPC technologies.
- While a GAN capable of generating synthetic traces has been created already, there remains an obvious need for a better way to evaluate the quality of the traces it creates.
- Quantitative analysis of job traces is no simple task, and so great thought needs to be placed in how we go about evaluating our GAN's output.

## Background

### What are job traces?

Job traces provide us detailed information on the inner workings of High Performance Computing Centers (HPCCs).

### Why are most not public?

Most HPCCs consider traces as proprietary data, as sharing them would potentially leak sensitive and private information to competitors.

### Why do they matter?

Should we find a way to generate synthetic traces, HPCCs could use them for research and upscaling purposes, allowing them to improve their workloads and grow in size more efficiently, while still obfuscating sensitive data.

## Method

This study focused heavily on statistical distance formulas and measurements. Our process was so:

- Researching several distance methods and other formulas, such as:
  - Wasserstein distance
  - Mahalanobis distance
  - Hellinger distance
- Implementing each formula in Python, using the scipy library
- Pre-processing sample & synthetic job traces
- Studying the results of feeding the traces into each formula

Most formulas found to be impractical for our use, but we eventually settled on the usage of two formulas:

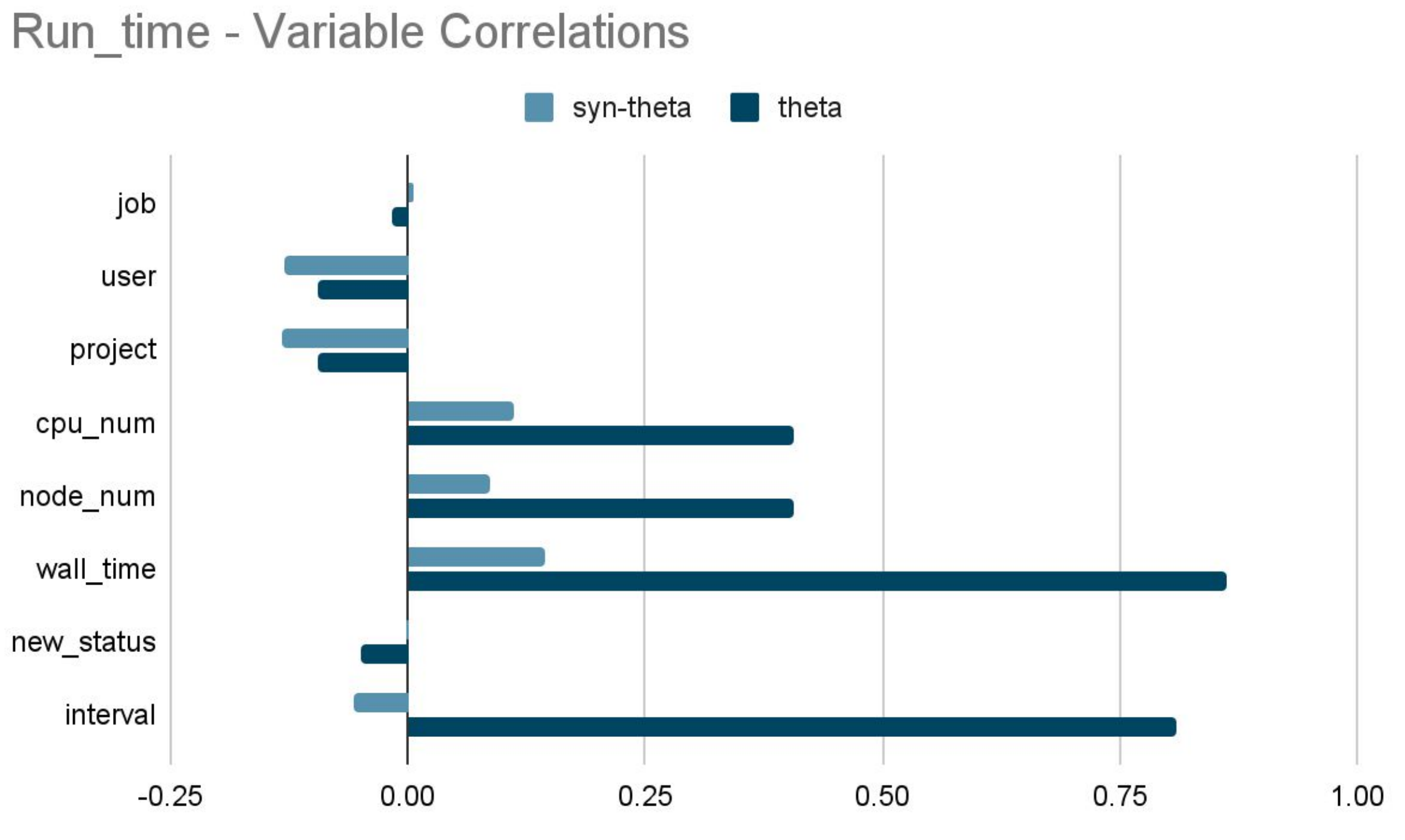
- **Kullback-Leibler Divergence**: The amount of information that would be lost/gained when transforming one dataset into another. Gives a number between 0 and infinity- the higher the number, the more dissimilar the two sets of information are.

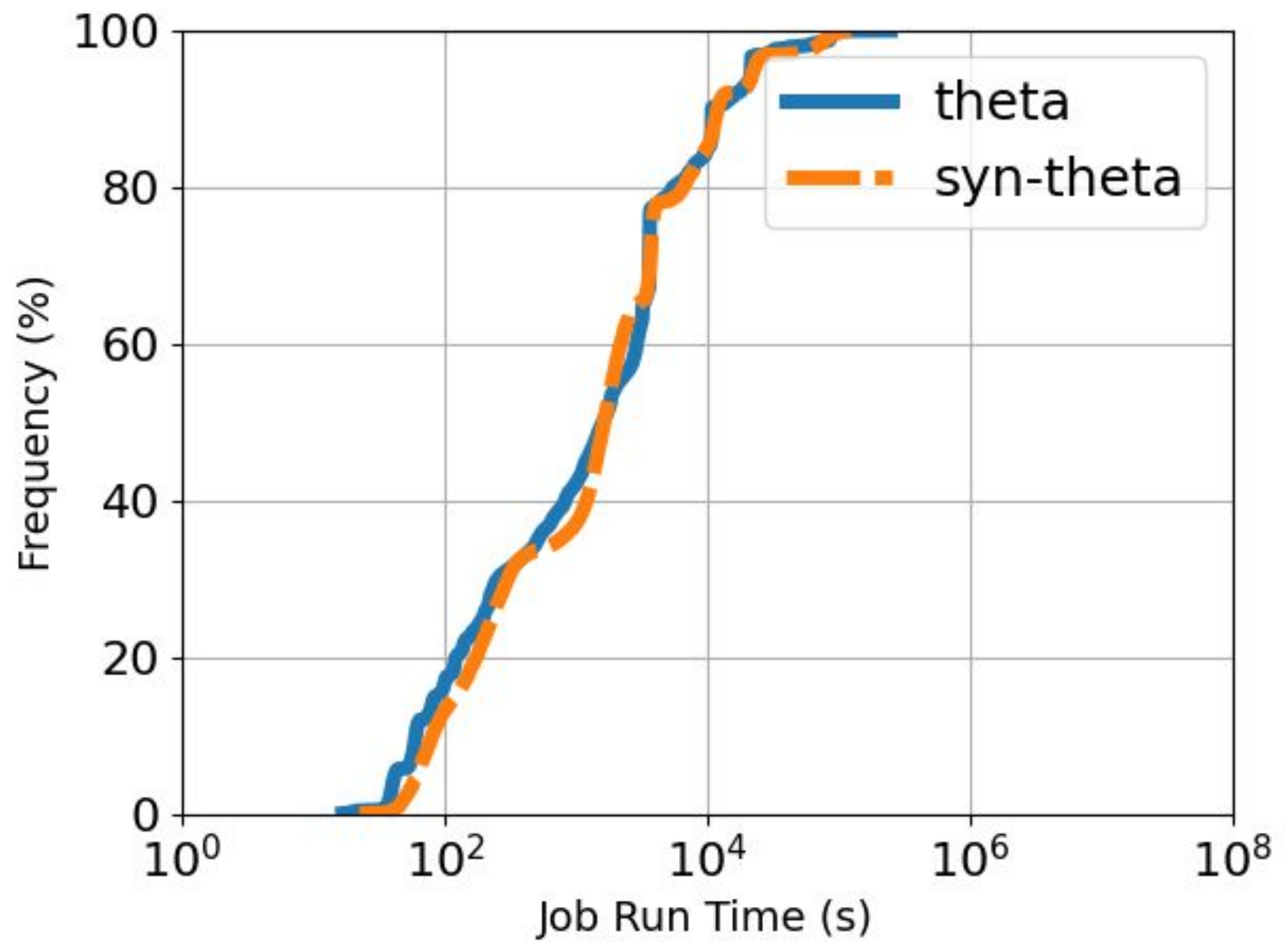
$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

- **Correlation Coefficient**: The measure of how correlated two variables are- 1 implying a strong positive correlation, and -1 implying the opposite.

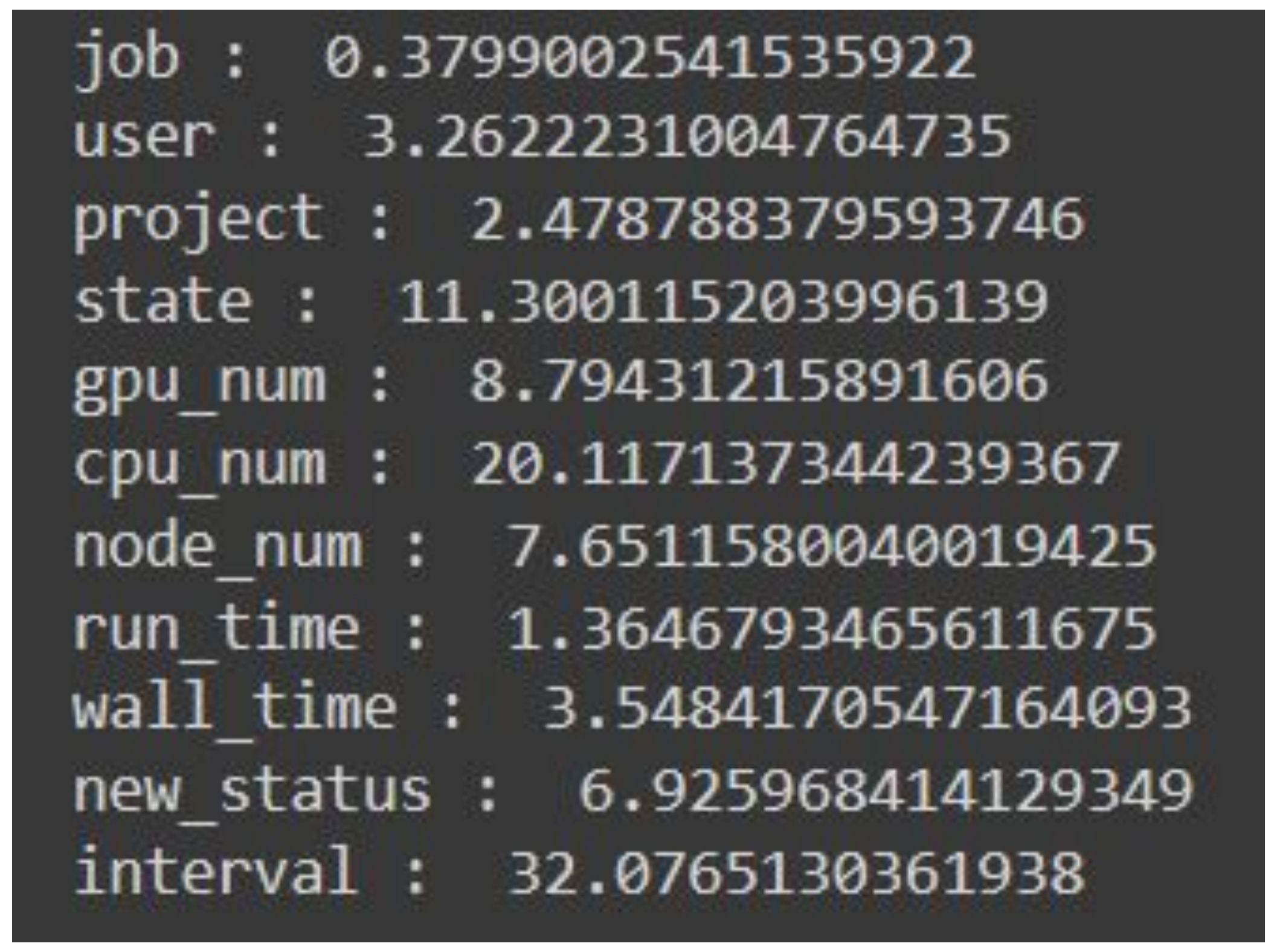
$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

## Results





*(Below) A list of the KL divergences between a job trace and a synthetic job trace generated based on it*



```
job :  0.3799002541535922
user :  3.2622231004764735
project :  2.478788379593746
state :  11.300115203996139
gpu_num :  8.79431215891606
cpu_num :  20.117137344239367
node_num :  7.6511580040019425
run_time :  1.3646793465611675
wall_time :  3.5484170547164093
new_status :  6.925968414129349
interval :  32.0765130361938
```

These divergences tell us a few different things about the quality of this job trace, such as that the two dataset's job fields contain similar information, while the interval field contains (comparatively) vastly dissimilar information.

## Results (cont.)

*(Top Left) A graph describing the correlations between the run time and other job trace variables for two job traces- one real, and one synthetic*
*(Middle Left) A graph showing the frequencies of different values of the run time variable in the two traces above*

As evidenced to the left, just because a variable in the synthetic trace is, when taken on its own, similar to the sample trace variable, doesn't mean that the synthetic trace is realistic or cohesive.

Realism would likely require them to show similar correlations to the sample job trace. Generating each variable without consideration for the others within the trace leads to an unrealistic trace.

## Conclusions

Using a combination of KL divergences and correlation coefficients as the quality metric within our GAN will likely provide us with higher quality synthetic job traces than we have now.

- KL divergence allows us to control the amount of information our synthetic job trace would contain that would differ from the sample trace.
- The correlation coefficient would ensure that our synthetic job traces show similar correlation between variables to real traces, ideally making the synthetic traces more realistic.