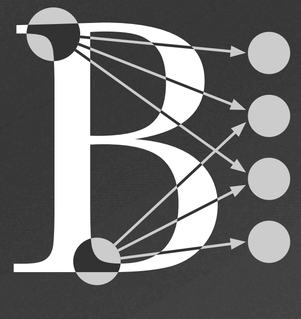
# Accelerating Data Engineering Pipelines

## BEAR | Baskerville | NVIDIA

Thursday 16th March 2023



# Why is data engineering important?

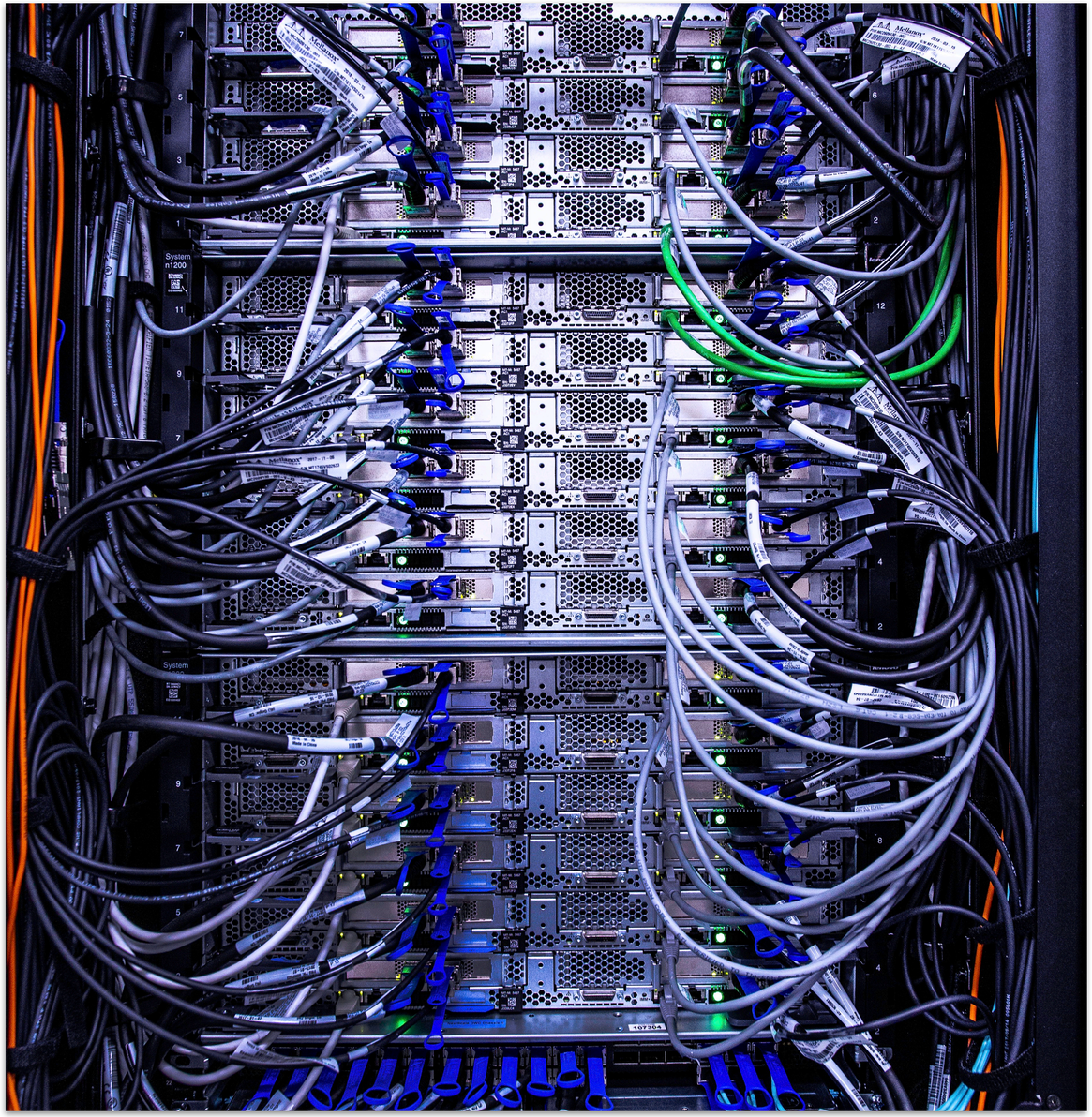
# 79 Zettabytes

**of data generated world wide in 2021**

(that's 79 x $10^{21}$ = 79 000 000 000 000 000 000 000 bytes!)

# Data Engineering

- Storing, analysing and visualising large volumes of data is not fast enough using traditional methods (SQL, CPUs)
- Essential to accelerate and parallelise processes using multiple GPUs
- This workshop will guide you through the tools to manipulate large datasets and visualise results using
  - cuDF
  - Dask
  - Plotly



Baskerville - a national accelerated compute system funded by EPSRC

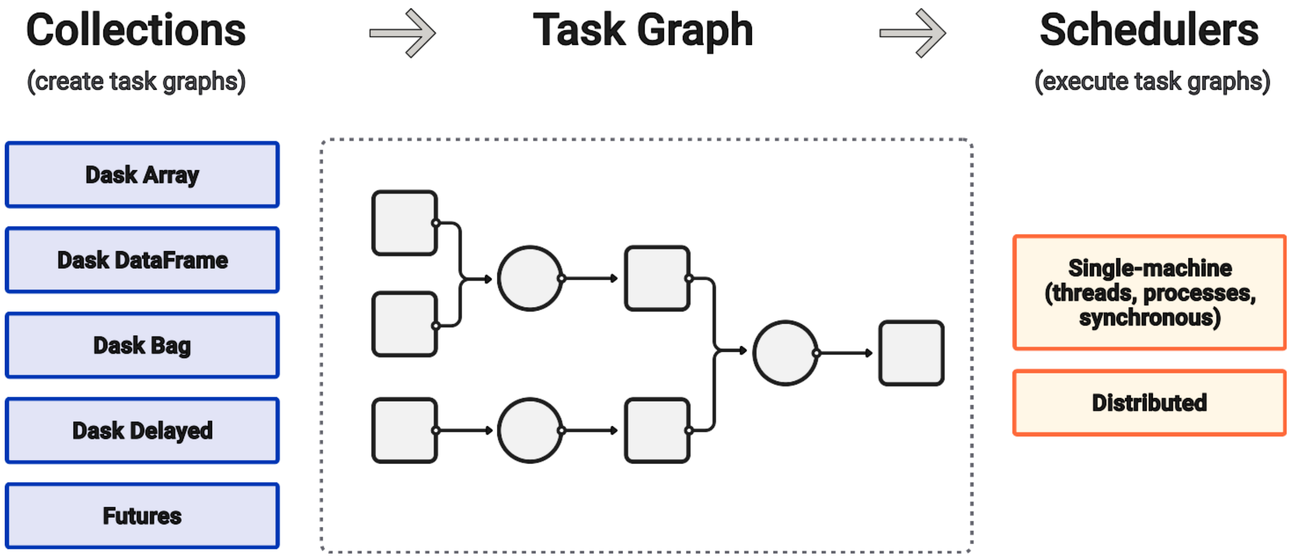
# cuDF

## GPU dataframe library

- Pandas-like API
- Built on the Apache Arrow columnar memory format
- Uses CUDA under the hood so you don’t have to learn C/C++/Fortran
- For workflows on a single GPU or if your data fits in memory on a single GPU
- Multi-GPU support with Dask

# Flexible parallel computing

- Dask for CPUs and Dask-cuDF for GPUs
- Stages of computation
  - “Lazy” = calculation computed only when needed
  - Operations on dataframe are “queued-up” and built into task graphs
  - Run with `.compute()` or `.persist()`



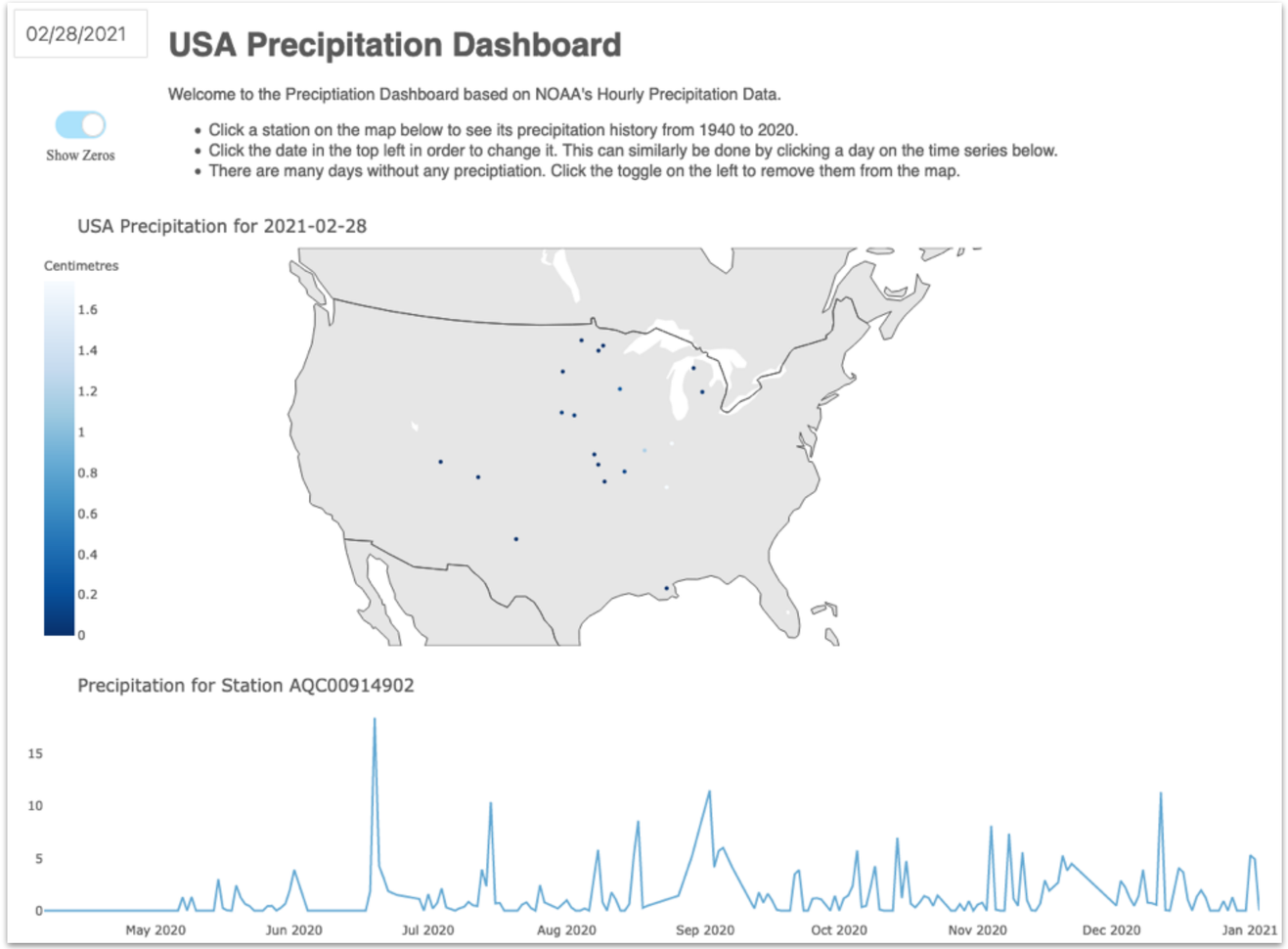
# plotly

## Interactive plotting library

- Build interactive web-based visualisations
- Rendered using JavaScript under the hood
- Export static images with Kaleido for non-web plots
- Integrate into Dash applications

# Workshop Outline

- You will visualise precipitation data from US NOAA
- You will accelerate and parallelise a “colleague’s” unfinished notebook
- You will use Jupyter Notebooks on Baskerville Portal
- Each account is limited to 2 GPUs each



Baskerville OnDemand Files Jobs Clusters Interactive Apps

The Baskerville portal provides web-based access to the Baskerville Tier 2 system

This service is operated by Advanced Research Computing at the University of Birmingham and is funded by EPSRC Grant EP/T022221/1

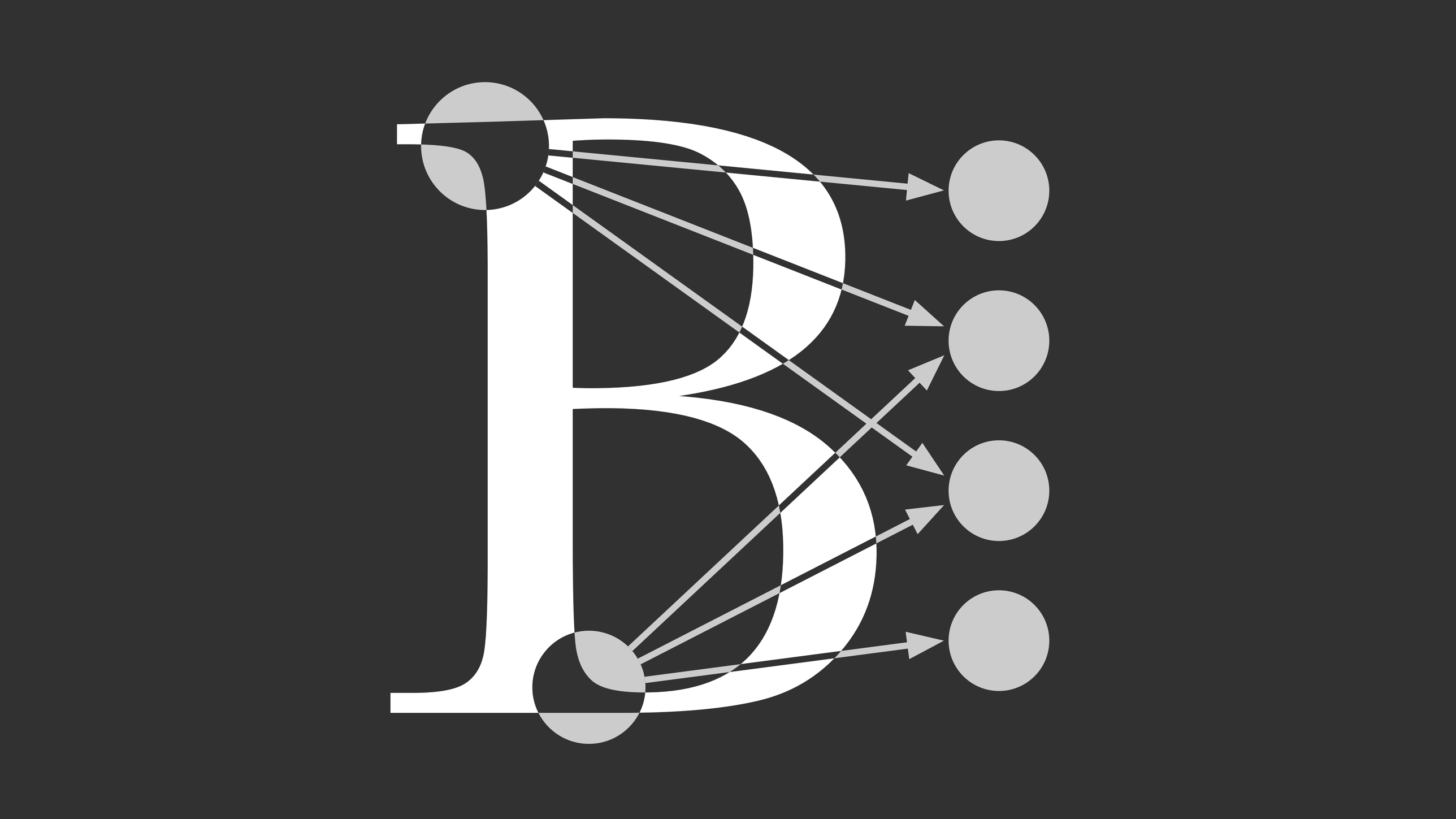
**https://portal.baskerville.ac.uk/**

# Workshop Setup

1) Create a symlink from the project to your home folder with ***ln /bask/projects/w/wongj-bham-training ~***
2) Create your user folder with

   ***cd ~/wongj-bham-training/users && mkdir $USER***
3) Setup your environment with ***source ../create_participant.sh*** (takes a while)
4) Close and re-launch the JupyterLab server (make sure ‘Show Conda Environments’ is ticked)
5) Work through ***users/$USER/info_data_engineering/challenge_instructions.ipynb***
6) Challenge yourself with ***users/$USER/info_data_engineering/challenge_notebook.ipynb***
7) Results are collected at **16:30**

**Will your notebook feature in the Top 5?**



# Collecting results…