Baskerville - Accelerate your research with GPUs

Accelerating Data Engineering Pipelines

Tuesday 7th February 2023

Why is data engineering important?

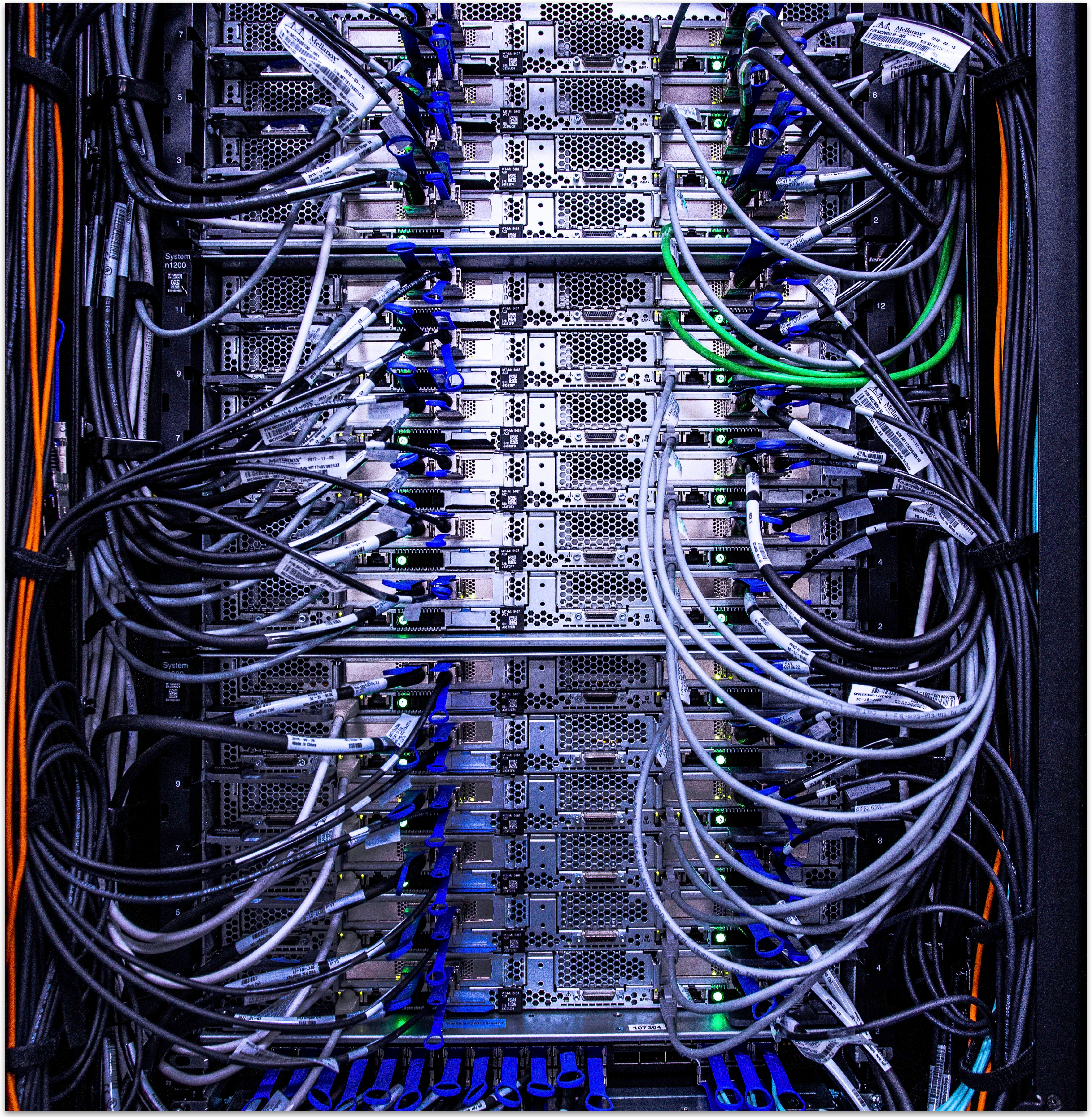
79 Zettabytes

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

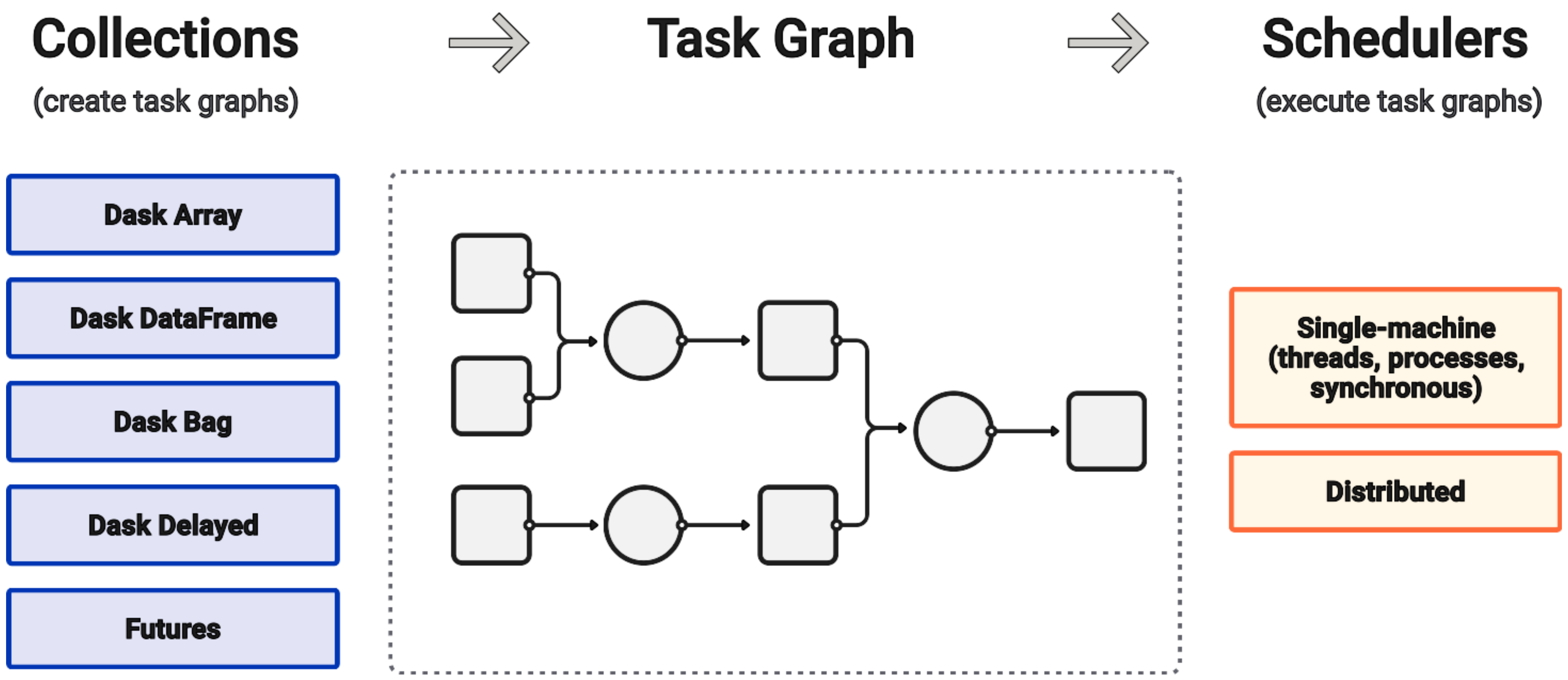
GPU dataframe library

- 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

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()`



Collections

(create task graphs)

→

Task Graph

→

Schedulers

(execute task graphs)

Dask Array

Dask DataFrame

Dask Bag

Dask Delayed

Futures

Single-machine (threads, processes, synchronous)

Distributed

plotly

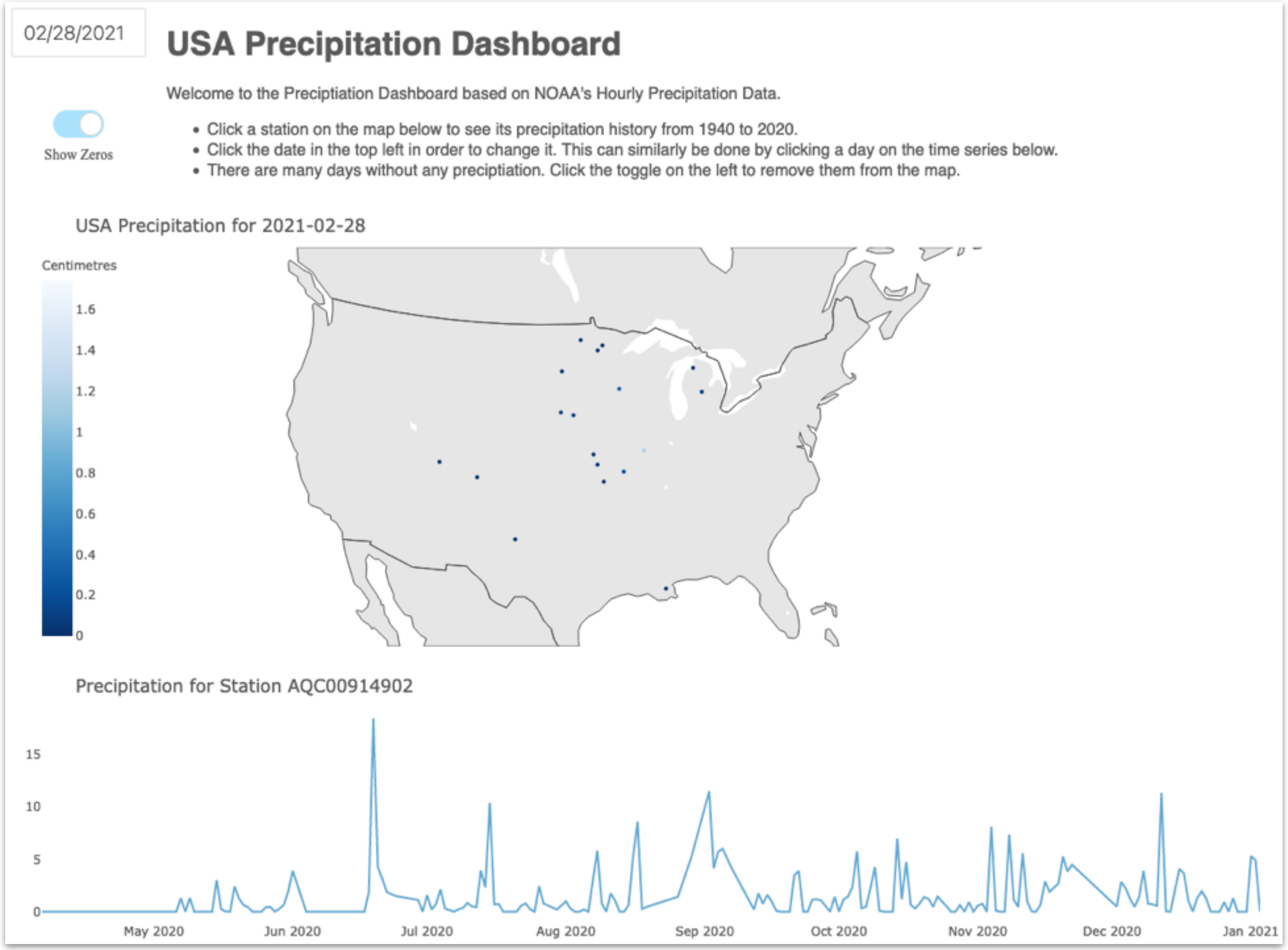
Interactive plotting library

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

(−0.4845348, 0.3549026)
plotly
Co
Cobalt
Atomic Mass: 58.933195
As
Rh Pd Ag Cd In Sn Sb
Ir Pt Au Hg Tl Pb Bi
Mt Ds Rg Cn Uut Fl Uup
Eu Gd Tb Dy Ho Er Tm
Am Cm Bk Cf Es Fm Md

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



02/28/2021

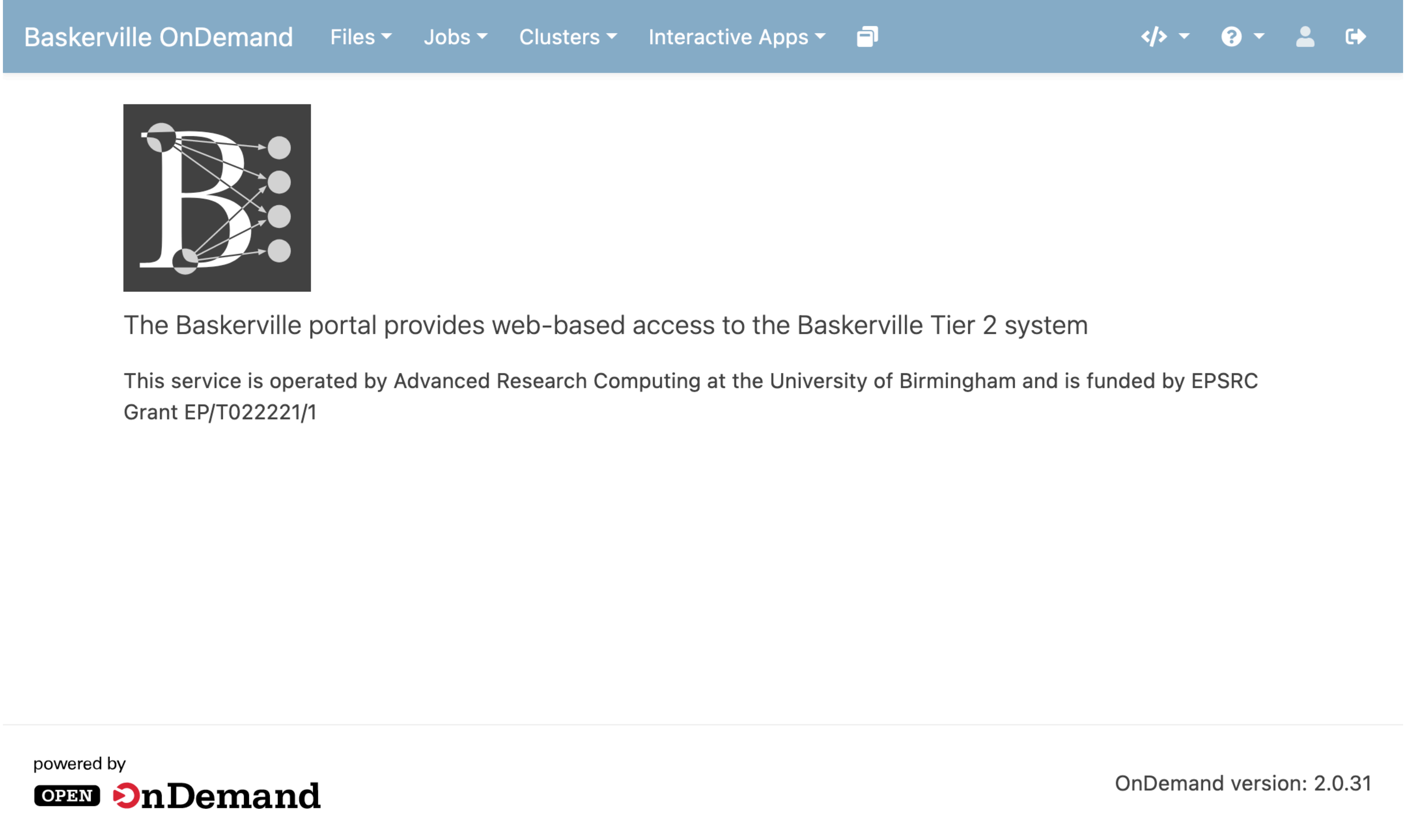
USA Precipitation Dashboard

Welcome to the Preciptiation Dashboard based on NOAA's Hourly Precipitation Data.

- Click a station on the map below to see its precipitation history from 1940 to 2020.
- Click the date in the top left in order to change it. This can similarly be done by clicking a day on the time series below.
- There are many days without any preciptiation. Click the toggle on the left to remove them from the map.

USA Precipitation for 2021-02-28

Precipitation for Station AQC00914902



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

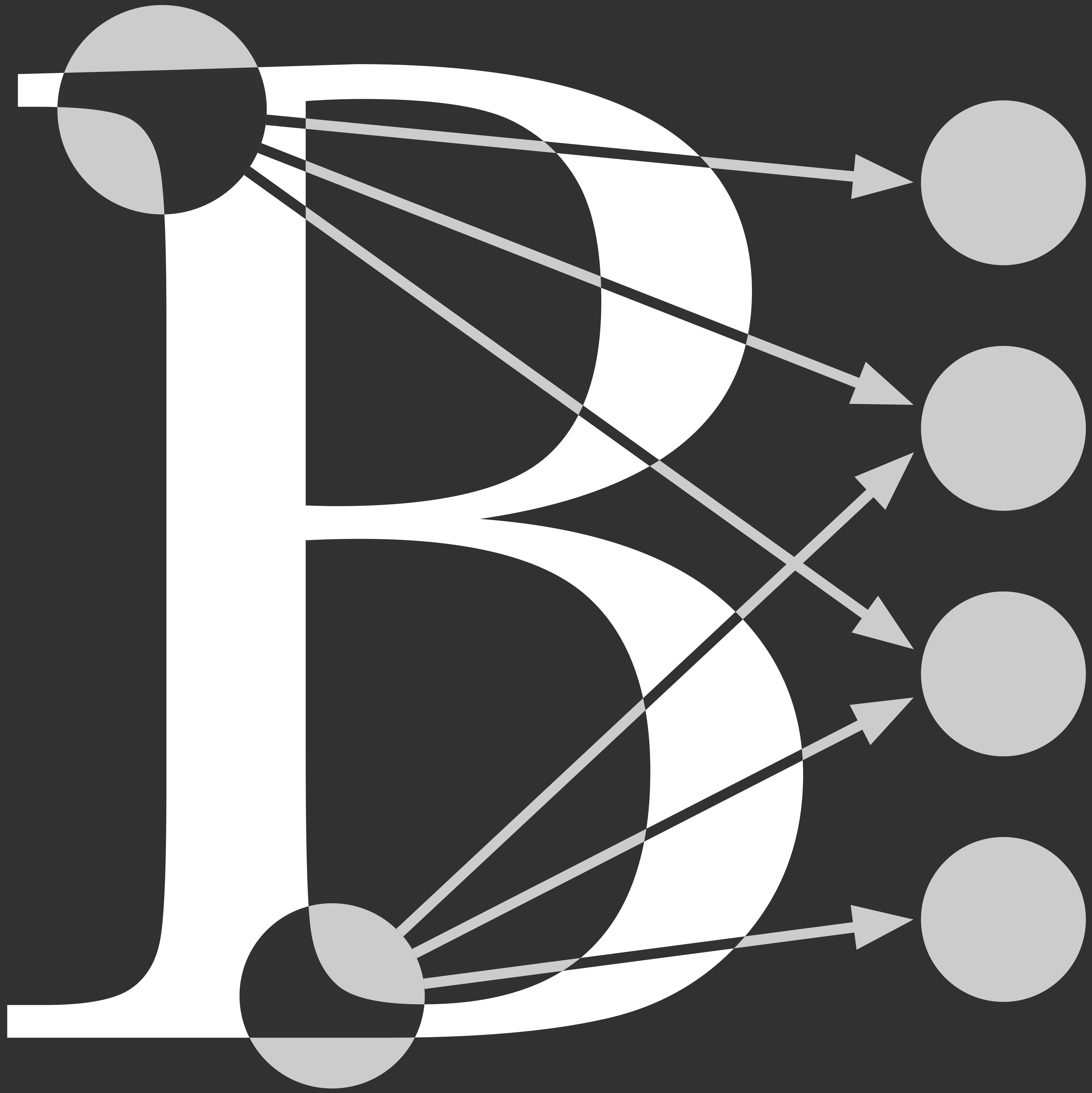
OnDemand version: 2.0.31

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?


B

Collecting results…