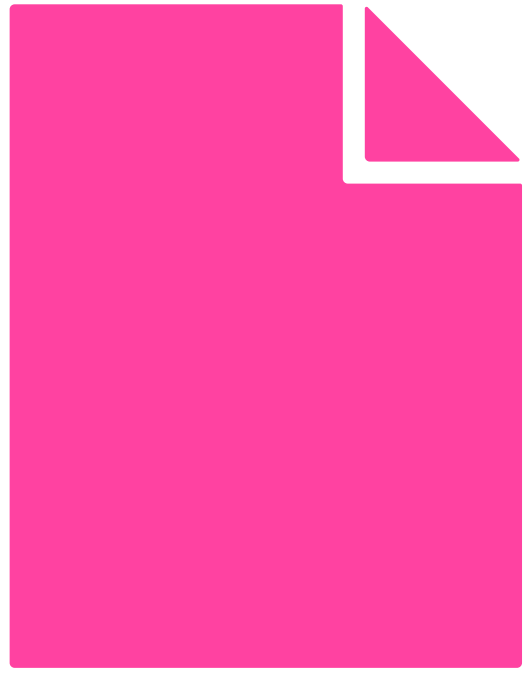
A Data Journey

Let’s be Concrete



Large HDF5 files,
abundance of useless data
80GB < … < 300GB

Feature selection
Feature engineering

Small Numpy files

Build a dataset

80 / 10 / 10

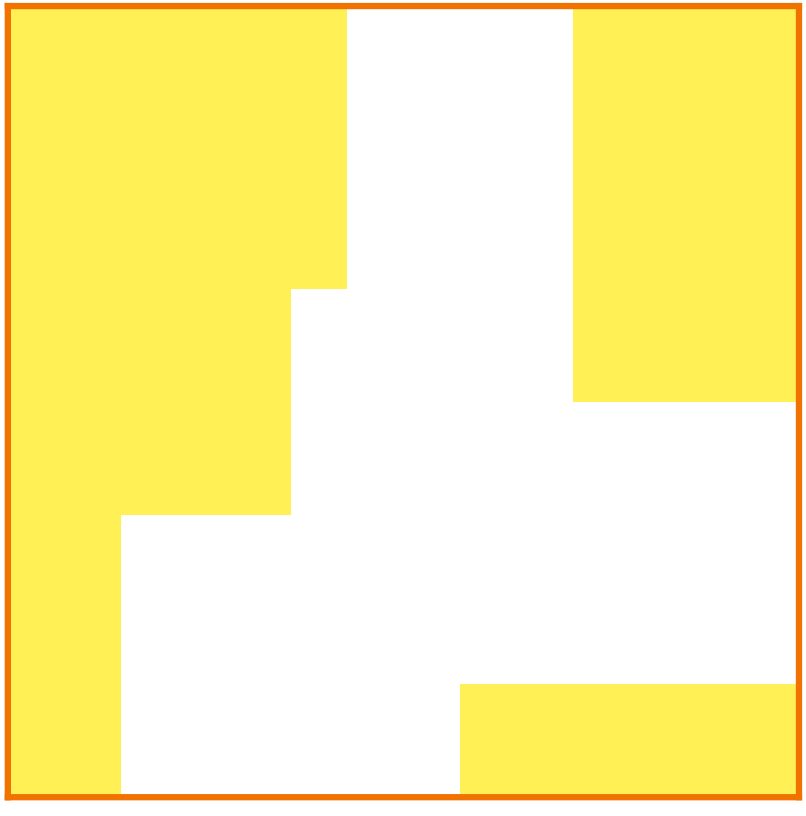
μ, σ

Compute basic Stats
or Characteristic Quantities
Possibly normalize the data

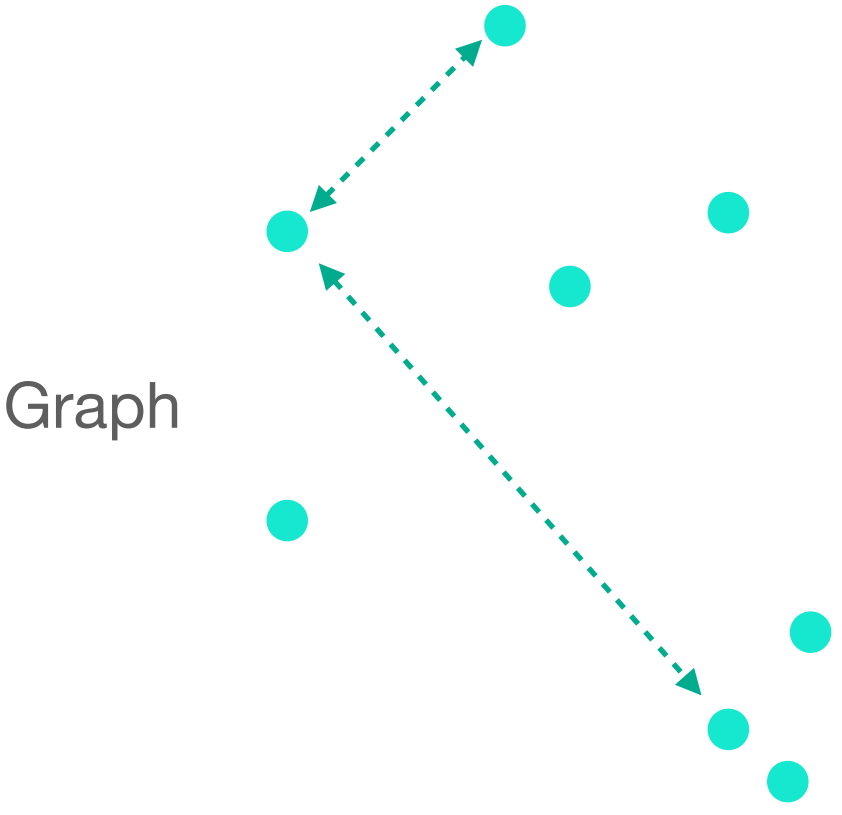
Dask

Reformatting
or rebuilding masks

Dask, Numpy



Masked Array

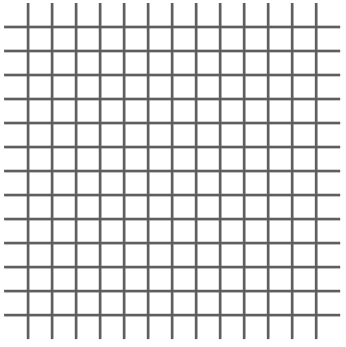


Build a Graph on-the-fly

PyG

Unstructured to Structured Interpolation

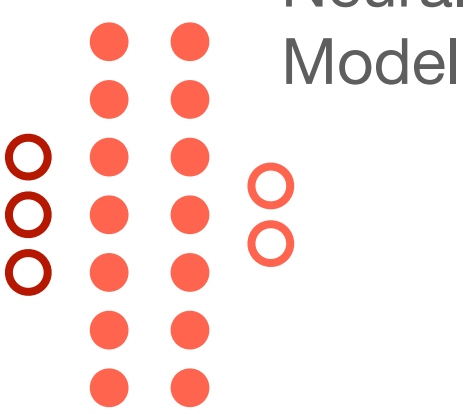
Coupling solution, SciPy



Cartesian grid

PyTorch, PyG
Ray
WandB, MLflow

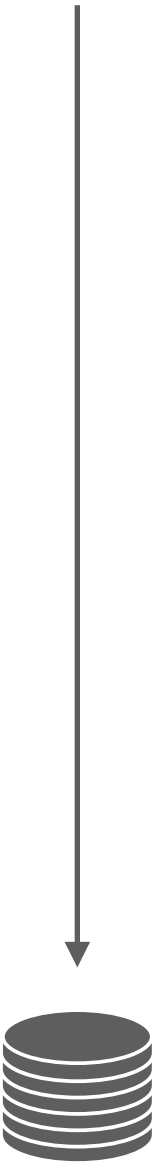
Train a GraphNet



Train a ConvNet

TensorFlow, PyTorch

14



A Data Journey

Let's be Concrete