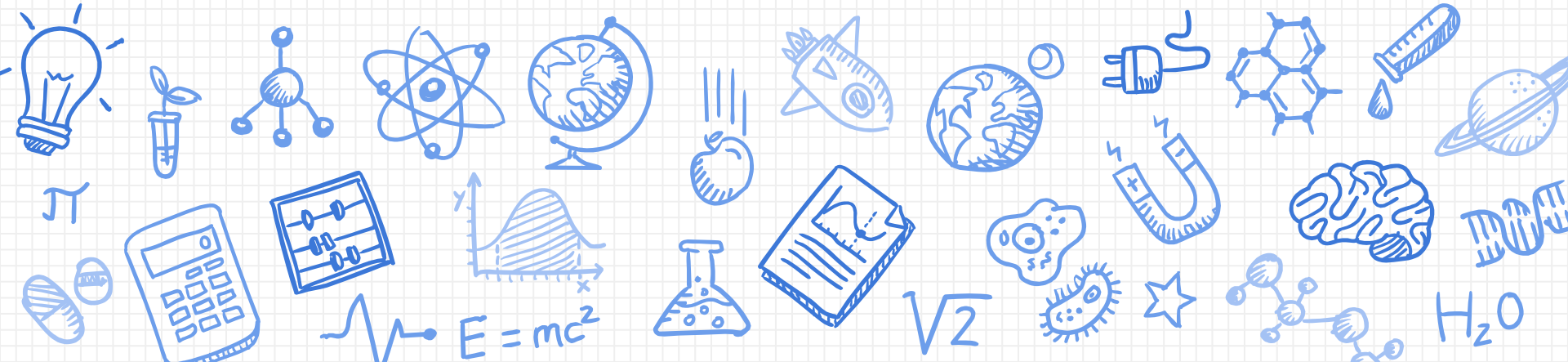
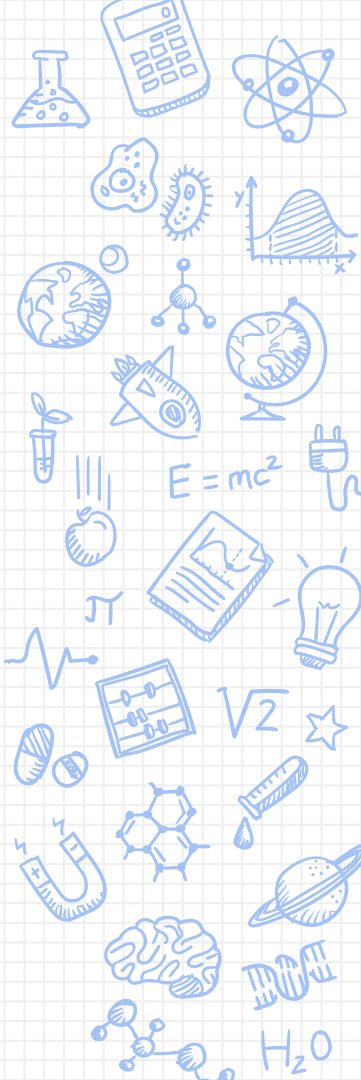
Rechunking



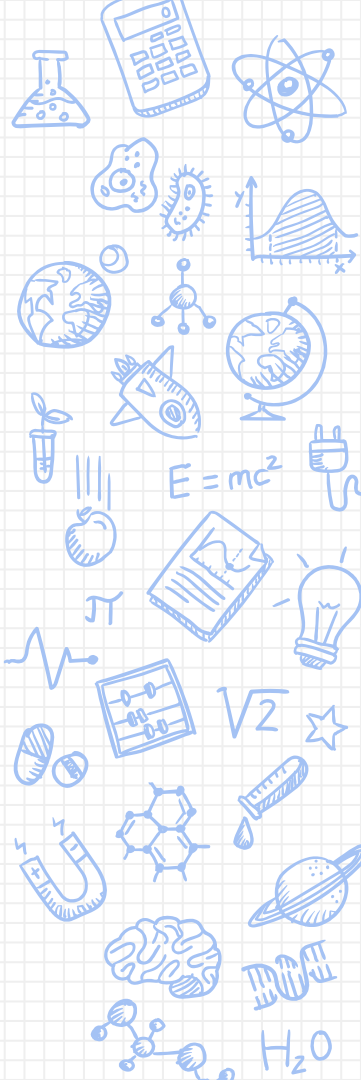
DataFrames in Memory

- ✘ Polars may store a DataFrame's data (or even a single column's data) in multiple locations in memory.
- ✘ A **chunk** is a location/group in memory.
- ✘ The CPU can read values in the same chunk faster because there is a smaller distance to travel.
- ✘ The word **contiguous** means "next to each other in memory".



Rechunking

- ✘ **Rechunking** refers to the process of merging multiple chunks together into one chunk.
- ✘ **Rechunking** has an upfront cost/penalty. Polars has to gather/move the data to a single chunk.
- ✘ *Future* operations on the data become faster.



Analogy: Warehouse

- ✘ Imagine your business sells bicycles and stores them in multiple warehouses (chunks).
- ✘ An employee gathering bicycles to ship has to travel to multiple warehouses to collect all the units.
- ✘ **Rechunking** is equivalent to investing the time to move all the bicycles to one warehouse.
- ✘ Future shipments are faster because the bicycles are closer together.