BMeat Transformation Pipeline

- Read CSVs into dataframes by mapping strings to DataFrames (Scala)
  - Convert all columns to string columns
  - *flatMap* all items to **CELLS**
  - *groupBy* the cell values
  - *collect_set* the column names into sets per CSV
  - → **CACHE-BASED PREAGGREGATION**
- *reduce* (Scala) all DataFrames into one with *union* (Spark)
- *groupBy* value again and *collect_set* the column names again → **ATTRIBUTE_SETS**
- *flatMap* each row to the **INCLUSION LISTS**
- *reduceByKey*, intersecting the inclusion lists, effectively **partitioning** and **aggregating**
- *filter* by throwing away empty inclusion sets → **INDs**
- *cort*
- *collect*
- *print*