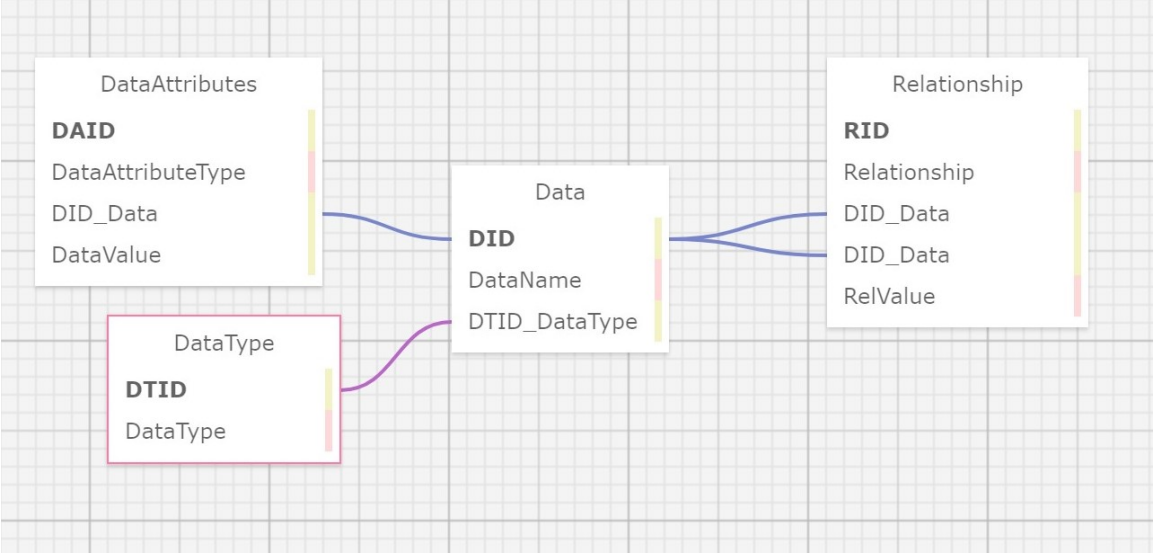
Biological Databases and Datamining
Midterm **DUE: 04/2/21 11:59pm**

Part 1) Create a database based on the schema below called netid_midterm.sqlite. It is similar to what we discussed in class.



In the figure each rectangle represents a separate table. The top row is the name of the table. The next row (in bold) is the primary key and the foreign keys are the fields where the name has a "_" followed by the name of the table it is coming from. For example DataID_Data means this is a foreign key from the table Data field name DataID.

Note In the Relationship table, the two foreign keys are from the same table Data and they should have different names, maybe DID_Data_1 and DID_Data_2.

The goal of this midterm is to integrate gene networks, RNA-seq expression values with gene annotations, including GO-terms so that you can ask questions such as:

"What is the expression value for all genes with the GO-term *nitrate assimilation* in each experiment?"

"How many interactions are there between genes that are associated with the same GO-term?"

To do this, we will load the following files into the database described above.:

1) Experiments (NextGenRaw.txt)
2) Gene annotations from BioMart (AthBiomart.txt) and
3) Gene network interactions (AthBIOGRID.txt)
    a. This file contains several columns, but what we are interested in are the 6th and 7th column (Systematic Name Interactor A and Systematic Name Interactor B, respectively) and the 12th column which is the name of the interaction and column 13, which is type of the interaction.

There are a total of 4 different Datatypes for this database:

Gene	=> Gene Stable ID	from Biomart
GOterm	=> GO term accession	from Biomart
Experiment	=> column names	from NextGenRaw

The Attributes for Gene are:

Gene name	from Biomart
Gene description	from Biomart

The Attributes for GOterm are:

Gene term name	from Biomart

The relationships are:

GOterm2Gene	from Biomart	does not have a value
Experiment2Gene	from NextGenRaw	has value
And the 12th column of BIOGRID		columns 13 (physical, genetic, etc)

Database creation and loading should be done in R.
***HINT** Use can use notes and code that we create in class.

Different strategies (pick one and go with it):

1) **Insert values:**
    a. Create empty tables in the database
    b. While reading in files, fill in the tables
2) **Write tables:**
    a. Use commands to restructure the input files into dataframes that look like the tables we need.
    b. Write the dataframes as tables into the database.

Both methods require you to come up with a way to store the primary keys so you can easily access them. Remember primary keys are unique to the row of that table. Having two primary keys with repeated information is not ok.

Part 2)

Choose one of the two following functions to write. You may choose to do both and the one with the best result will be counted.

1) **Write a function getReadCounts() where the input is a go-term and the output is the experiment values for each gene associated with the go-term**

So if I type getReadCounts("binding") I should get back a matrix (or data frame) with 4 columns (one for each each experiment) and one row for each gene associated with "binding".

The query does not have to be in one sql statement, just have a function that works.

The function **MUST** use the database you have created in part 1.

2) **Write a function getInteractions() where the input is a go-term and the output is an interaction between genes associated with the go-term.**

So if I type getInteractions("binding") I should get a list of interactions where both Interactors are associated with the go-term "binding".

The query does not have to be in one sql statement, just have a function that works.

The function **MUST** use the database you have created in part 1.