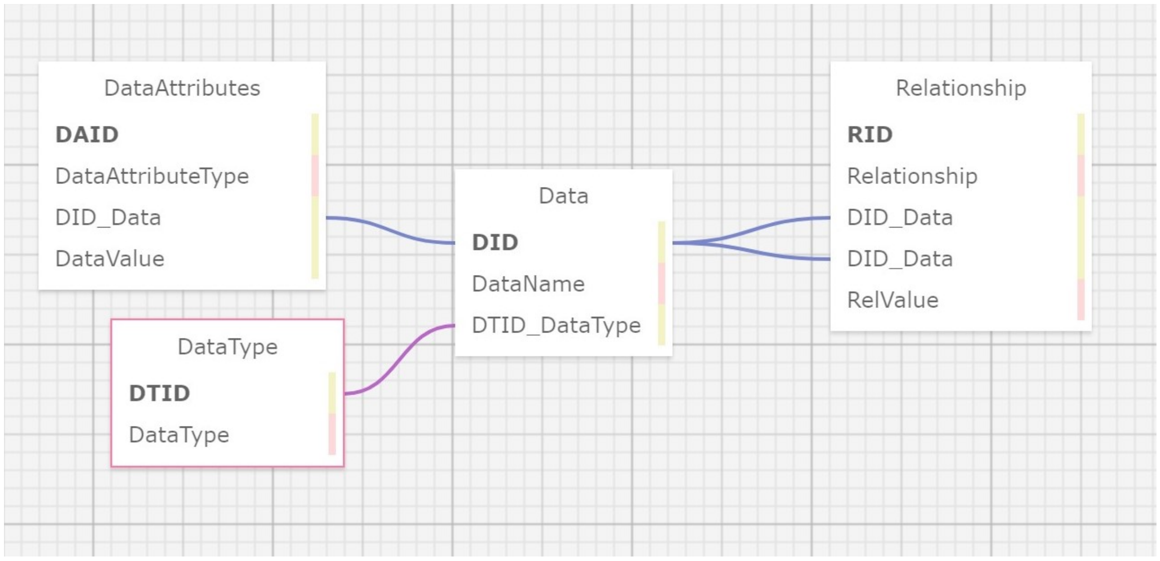
Biological Databases and Datamining
Midterm

DUE: 03/29/20 11:59pm

1) Create a database based on the schema below called netid_midterm.sqlite. It is similar to hw2 but can support many types of data and keep the tables normalized.



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 of the table followed by the other fields. The type of field designates the color of the row. Yellow is any type of numeric and Pink is any type string. The lines show a primary key being used as a foreign key. The colors don't mean anything in this case. The foreign key 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.

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?"

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) This file contains several columns, but what we are interested in are the 8th and 9th column (Official Symbol Interactor A and Official Symbol Interactor B, respectively).

There are a total of 4 different Datatypes for this database:
Gene => Gene Stable ID
Probe => Affymetrix array Arabidopsis ATH1 121501 ID
GOterm => GO term accession
Experiment => column names of NextGenRaw.txt

The relationships are :
Probe -> Gene
GOterm -> Gene
Experiment -> Gene

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

2) **Write a function getReadCounts() where the input is a go-term and the output is the read count for each gene in each experiment.**

So if I type getReadCounts("binding") I should get back a matrix (or data frame) with 4 columns 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.

3) **Speeding up our query.**

Take a time stamp in R to calculate how long the query in part 2 takes.
Use proc.time() and save the result in a vector just before you run the query and run it one more time after the function call and save it as a different variable.
Simply subtract the two variables to see how long your query took.

```
A = proc.time()
getReadCounts("binding")
B = proc.time()
Time_it_took = B-A
```

b. Create indexes on your tables to try to speed up the query. Explain the reason for creating the indexes and how long your query takes now?