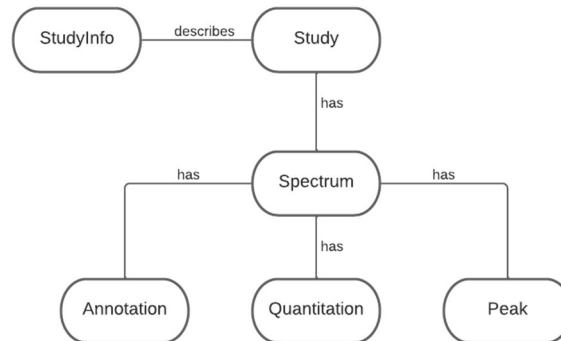# Lab 12

We continue working on our Metabolomics Database. As a reminder, this is the conceptual schema of the database.

This time, we'll focus on the information about each study stored in the JSON file:


```
"tags": {
    "ms instrument type": "GC-TOF",
    "ms instrument name": "Leco Pegasus III GC TOF",
    "chromatography system": "Leco Pegasus III GC",
    "column": "Agilent DB 5-MS (10 m x 0.18 mm x 0.18 um)",
    "subject type": "Human",
    "sample source": "Blood",
    "species (latin)": "Homo sapiens",
    "species (common)": "Human",
    "disease": "Necrotizing soft-tissue infections"
},
```

There are two approaches to design the table **study_info**:

1. Create a table with columns **ms_instrument_type**, **ms_instrument_name**, **chromatography_system**, etc. Then, we can save each value to the corresponding column.
2. Create a table with columns **key** and **value**, where the column **key** would store values "ms instrument type", "ms instrument name", etc., while the column **value** would store values "GC-TOF", "Leco Pegasus III GC TOF", etc.

The first approach is simpler to implement. However, it's good only if you know exactly what keys are used in the JSON file. If a new key-value pair shows up in the JSON file (say, "symptom" : "cought"), then you'd need to add a new column. So, this is not a good approach if the keys change often.

Instead, we'll follow the second approach, where you store data in a table like this:

Course: BINF_6211 LAB 12
Name: Drashti Mehta, SID: 801262877

Key	value
ms instrument type	GC-TOF
ms instrument name	Leco Pegasus III GC TOF
chromatography system	Leco Pegasus III GC
...	...

Task 1 (10pts)

Design and create a table to store the information about each study following the second approach. Which means you need to have columns **key** and **value** in your table. Think about what other columns you need.

Report the SQL for creating the table.

CREATE TABLE study_info (
info_id INTEGER PRIMARY KEY AUTOINCREMENT,
study_id TEXT,
[key] TEXT,
value TEXT
);

Task 2

Populate the new table with data from the four JSON files. You can either do it manually or write a script to convert each JSON file into CSV.

	info_id	study_id	key	value
1	1	ST001270	ms_instrument_type	GC-TOF
2	2	ST001270	ms_instrument_name	Leco Pegasus III GC TOF
3	3	ST001270	chromatography_system	Leco Pegasus III GC
4	4	ST001270	column	Agilent DB 5-MS (10 m x 0.18 mm ...
5	5	ST001270	subject_type	Human
6	6	ST001270	sample_source	Blood
7	7	ST001270	species_latin	Homo sapiens
8	8	ST001270	species_common	Human
9	9	ST001270	disease	Necrotizing soft-tissue infections
10	10	ST001369	ms_instrument_type	Single quadrupole
11	11	ST001369	ms_instrument_name	Agilent 5975C
12	12	ST001369	chromatography_system	NULL
13	13	ST001369	column	NULL
14	14	ST001369	subject_type	Human
15	15	ST001369	sample_source	Blood
16	16	ST001369	species_latin	Homo sapiens
17	17	ST001369	species_common	Human
18	18	ST001369	disease	Allergy

Task 3 (20pts)

Write SQL queries to perform the following tasks:

1. Find all humans studies;

select * from study_info where value = "Human" and key = "species_common"

	info_id	study_id	key	value
1	8	ST001270	species_common	Human
2	17	ST001369	species_common	Human

2. Find all studies where "species (common)" = "Human" and "ms instrument type" = "GC-TOF";

select study_id from study_info where value = "GC-TOF" and key = "ms_instrument_type"
intersect select study_id from study_info where value = "Human" and key = "species_common"

3. Find all studies where "species (common)" = "Human" or "ms instrument type" = "GC-TOF";

select study_id from study_info where value = "GC-TOF" and key = "ms_instrument_type"
union select study_id from study_info where value = "Human" and key = "species_common"

4. Find all studies where "disease" is missing or "disease" = "Allergy".

select study_id from study_info where value IS NULL and key = "disease"
union select study_id from study_info where value = "Allergy" and key = "disease"

Report the SQL queries and their outputs.

Task 4 (5pts)
Upload the database file with tables **study_info**, **studies**, **spectra**, **peaks**, and **annotations** onto Canvas.