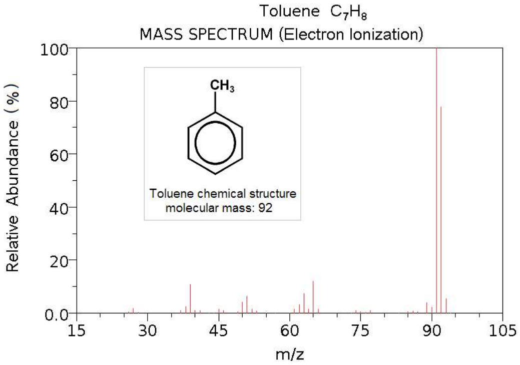
Course: BINF-6211 – Lab 9
Name: Drashti Mehta

Lab 9

During this and the next few labs, we'll create a database to store metabolomics mass-spectrometry data.

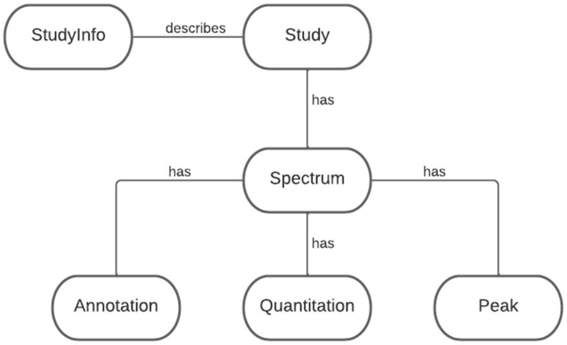
From Wikipedia:

- **Metabolomics** is the scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates and products of cell metabolism.
- **Mass spectrometry** (MS) is an analytical technique that is used to measure the mass-to-charge ratio of ions. In a typical MS procedure, a sample is ionized by bombarding it with a beam of electrons. This causes some of the sample's molecules to break up into positively or negatively charged fragments. These ions (fragments) are then separated according to their mass-to-charge ratio (m/z). The results are presented as a mass spectrum, a plot of intensity as a function of the mass-to-charge ratio (m/z). Resulting mass spectra are (mostly) unique for each molecule, and can be used for identifying/annotating unknown molecules.



NIST Chemistry WebBook (http://webbook.nist.gov/chemistry)

Conceptual schema of the database:



Lab 9 task:

In this lab, we'll focus on three entities: **Study**, **Spectrum**, and **Peak**. They are related as follows:

- Each study has multiple spectra. Each spectrum belongs to a single study.
- Each spectrum has multiple peaks. Each peak belongs to a single spectrum.

Task 1.

Download and unzip the data archive from Canvas. It contains 4 folders, each folder contains data from a single study. In each folder, you'll see the following files:

- **meta.json** contains information (title, summary, url) about each study,

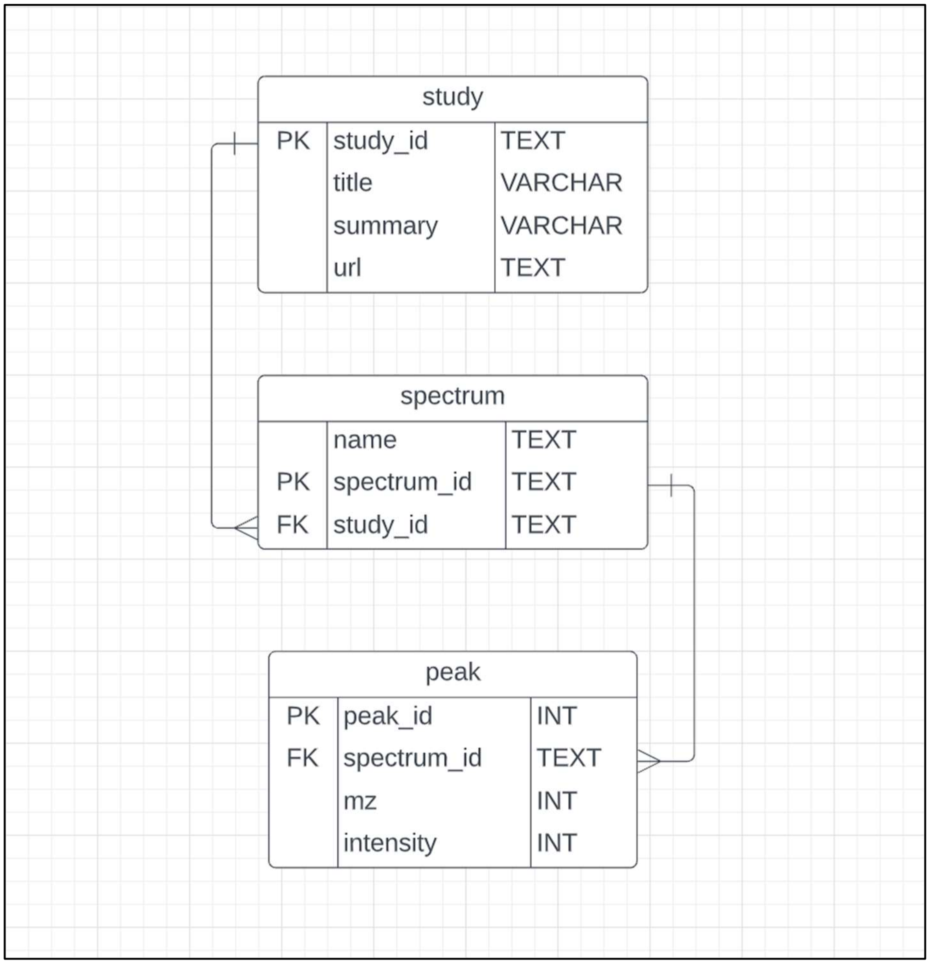
Course: BINF-6211 – Lab 9
Name: Drashti Mehta

- **spectra.msp** contains some information about each spectrum and its peaks (m/z-intensity pairs)

You can open these files in any text editor, but I recommend Visual Studio Code (https://code.visualstudio.com/). Take a look at these files and make sure you understand their content.

Task 2 (15pts).

Design tables for the entities **Study**, **Spectrum**, and **Peak**. Draw an Entity-Relationship diagram (ERD). In ERD, specify name of each table, list of attributes, primary and foreign keys, and relationships between the tables.



study		
PK	study_id	TEXT
	title	VARCHAR
	summary	VARCHAR
	url	TEXT

spectrum		
	name	TEXT
PK	spectrum_id	TEXT
FK	study_id	TEXT

peak		
PK	peak_id	INT
FK	spectrum_id	TEXT
	mz	INT
	intensity	INT

Fig: ERD for study, spectrum and peak

It is up to you to decide what attributes you want to have in each table.

- At the minimum, you need the attributes to be able to execute the SQL queries from Task 4,
- I'd also suggest to add attributes for study title, study summary, study ID, and spectrum name,
- The **Peak** entity should have attributes **mz** and **intensity**.

Report the Entity-Relationship diagram.

Here, I'm creating a temporary table for importing unique data in the Spectrum and Peak tables. This "temp_table" has information from all the spectra files (created as shown in part 3) and later data is imported in the tables. The way it works is that the data from every file is imported into this temp_table one by one and it is taken into Spectrum and Peak table. Study_id can be added manually as default as and when each file is imported.

```
CREATE TABLE Study (
study_id CHAR PRIMARY KEY,
title VARCHAR,
summary VARCHAR,
url TEXT
);

CREATE TABLE Spectrum (
name TEXT,
spectrum_id TEXT PRIMARY KEY,
study_id TEXT REFERENCES Study (study_id)
);

CREATE TABLE temp_table (
name TEXT,
id TEXT PRIMARY KEY,
num_peaks INTEGER,
mz INTEGER,
intensity INTEGER
);

CREATE TABLE Peak (
peak_id INTEGER PRIMARY KEY AUTOINCREMENT,
spectrum_id TEXT REFERENCES Spectrum (spectrum_id),
mz INTEGER,
intensity INTEGER
);
```

Task 3 (15pts).

Fill out the tables that you created with data from the archive. In order to read the .msp files, I've attached scripts both in Python and R to convert those files into a csv.

You can run the Python script in the terminal (you may need to install Pandas):

```
> python3 read_msp.py PATH_TO_MSP
```

Course: BINF-6211 – Lab 9
Name: Drashti Mehta

You can run the R script in RStudio or in the terminal (you need to edit paths to the files inside the R script):
> Rscript read_msp.py

Upload the database to Canvas.

ANSWER:
The files are attached to the submission section with the database.

Task 4 (15pts).

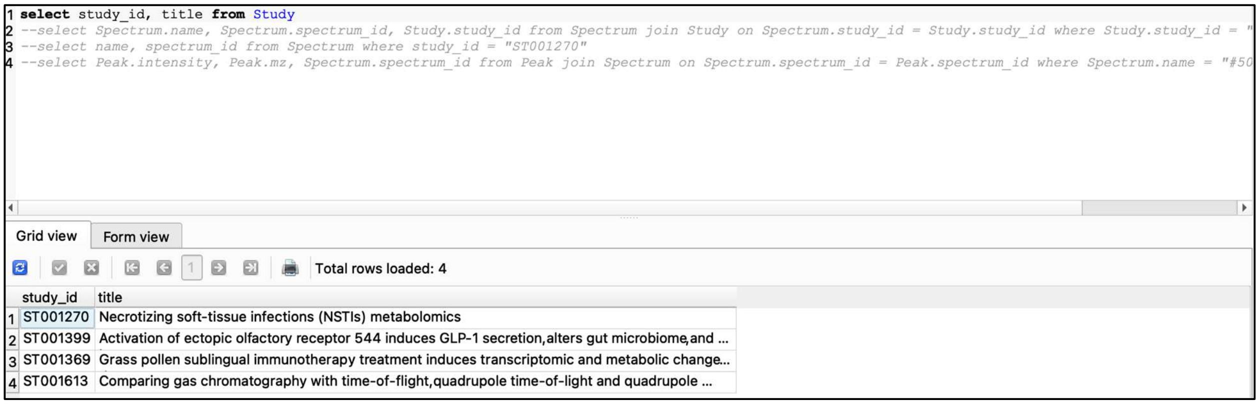
Try to execute the following SQL queries:

- List all studies in the database
- List all spectra for the study "ST001270"
- List all peaks for the spectrum "#50 75.0000@6.0753 MS1+"

Report the output of these queries. You can take a screenshot of the first few rows of the output.

ANSWER:

1. select study_id, title from Study



```
select study_id, title from Study
--select Spectrum.name, Spectrum.spectrum_id, Study.study_id from Spectrum join Study on Spectrum.study_id = Study.study_id where Study.study_id = "
--select name, spectrum_id from Spectrum where study_id = "ST001270"
--select Peak.intensity, Peak.mz, Spectrum.spectrum_id from Peak join Spectrum on Spectrum.spectrum_id = Peak.spectrum_id where Spectrum.name = "#50
```

Total rows loaded: 4

	study_id	title
1	ST001270	Necrotizing soft-tissue infections (NSTIs) metabolomics
2	ST001399	Activation of ectopic olfactory receptor 544 induces GLP-1 secretion,alters gut microbiome,and ...
3	ST001369	Grass pollen sublingual immunotherapy treatment induces transcriptomic and metabolic change...
4	ST001613	Comparing gas chromatography with time-of-flight,quadrupole time-of-light and quadrupole ...

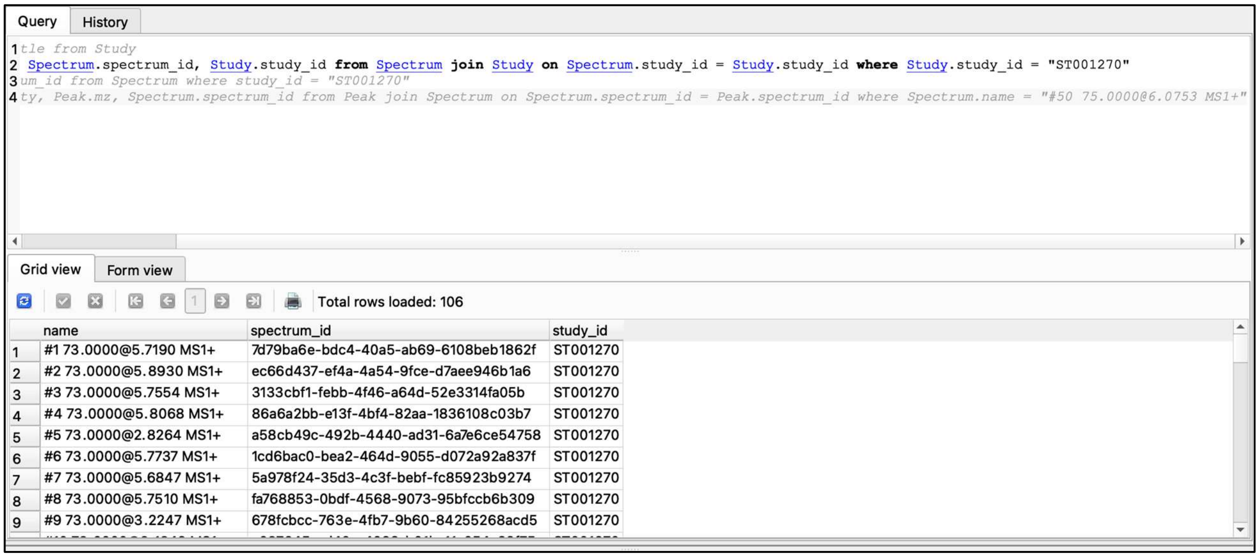
2. This can be done in two ways but way 1 is preferred.

   1) select Spectrum.name, Spectrum.spectrum_id, Study.study_id from Spectrum join Study on Spectrum.study_id = Study.study_id where Study.study_id = "ST001270"

   OR

   2) select name, spectrum_id from Spectrum where study_id = "ST001270"

Course: BINF-6211 – Lab 9
Name: Drashti Mehta

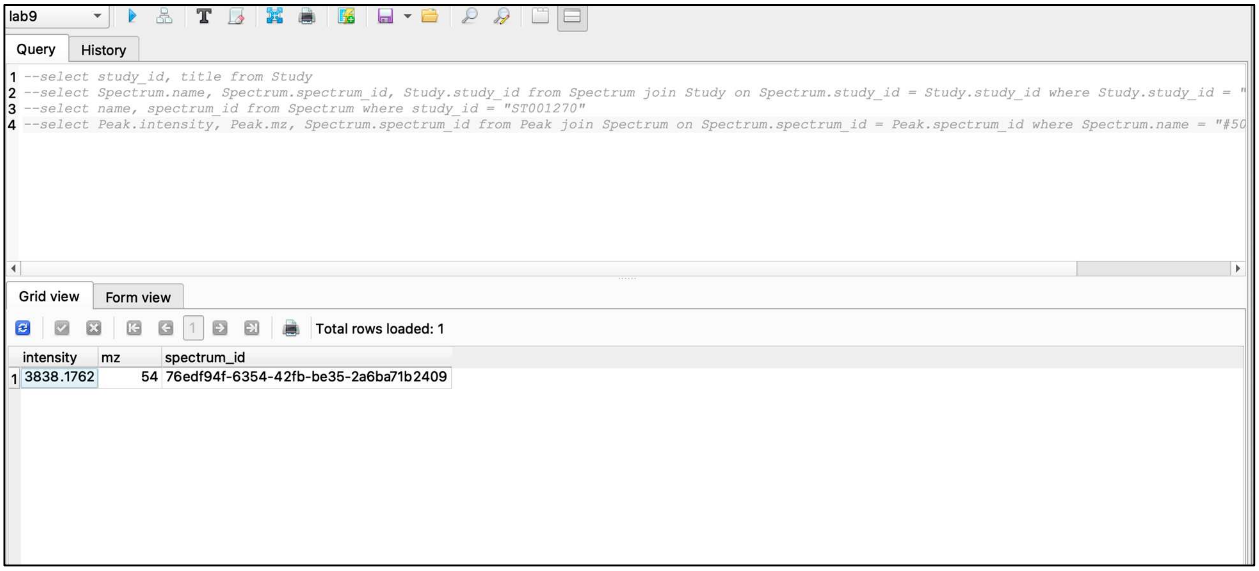


```
tle from Study
 Spectrum.spectrum_id, Study.study_id from Spectrum join Study on Spectrum.study_id = Study.study_id where Study.study_id = "ST001270"
um_id from Spectrum where study_id = "ST001270"
ty, Peak.mz, Spectrum.spectrum_id from Peak join Spectrum on Spectrum.spectrum_id = Peak.spectrum_id where Spectrum.name = "#50 75.0000@6.0753 MS1+"
```

Total rows loaded: 106

	name	spectrum_id	study_id
1	#1 73.0000@5.7190 MS1+	7d79ba6e-bdc4-40a5-ab69-6108beb1862f	ST001270
2	#2 73.0000@5.8930 MS1+	ec66d437-ef4a-4a54-9fce-d7aee946b1a6	ST001270
3	#3 73.0000@5.7554 MS1+	3133cbf1-febb-4f46-a64d-52e3314fa05b	ST001270
4	#4 73.0000@5.8068 MS1+	86a6a2bb-e13f-4bf4-82aa-1836108c03b7	ST001270
5	#5 73.0000@2.8264 MS1+	a58cb49c-492b-4440-ad31-6a7e6ce54758	ST001270
6	#6 73.0000@5.7737 MS1+	1cd6bac0-bea2-464d-9055-d072a92a837f	ST001270
7	#7 73.0000@5.6847 MS1+	5a978f24-35d3-4c3f-bebf-fc85923b9274	ST001270
8	#8 73.0000@5.7510 MS1+	fa768853-0bdf-4568-9073-95bfccb6b309	ST001270
9	#9 73.0000@3.2247 MS1+	678fcbcc-763e-4fb7-9b60-84255268acd5	ST001270

3. select Peak.intensity, Peak.mz, Spectrum.spectrum_id from Peak join Spectrum on Spectrum.spectrum_id = Peak.spectrum_id where Spectrum.name = "#50 75.0000@6.0753 MS1+"



```
--select study_id, title from Study
--select Spectrum.name, Spectrum.spectrum_id, Study.study_id from Spectrum join Study on Spectrum.study_id = Study.study_id where Study.study_id = "
--select name, spectrum_id from Spectrum where study_id = "ST001270"
--select Peak.intensity, Peak.mz, Spectrum.spectrum_id from Peak join Spectrum on Spectrum.spectrum_id = Peak.spectrum_id where Spectrum.name = "#50
```

Total rows loaded: 1

	intensity	mz	spectrum_id
1	3838.1762	54	76edf94f-6354-42fb-be35-2a6ba71b2409