# Term Project Update #2

Wenliang Zhang 10/24/2021

1. What is the advanced database area you are focusing on related to this course?

- Learn and try the new emerging database technology, called TileDB, and compared to noSQL database, MongoDB, for storing and query multi-dimensional genetic data
- try and compared to Cloudera HDFS and query with Hive and Impala?
- Query from the TileDB/MongoDB and data analysis using PySpark

2. What is the proof of concept component in your term project?

- Being able to save multiple complicated genomic datasets in the json format into TileDB and MongoDB
- Being able to query data from TileDB and MongoDB
- pySpark and TileDB/MongoDB work together to do data analysis

3. What are some of the goals that you plan to learn in this project?

- Learn how to design TileDB array and an efficient noSQL database for storing and querying complicated genomic data
- working with pySpark and TileDB/MongoDB together for data analysis

4. What skills are you bringing from other courses, and what is the new element that you are learning in this class that's related to advanced database management?

- Learn and bring the new database, TileDB, and the techniques to store and access multi-dimensional datasets
- I will bring skills like Python and pySpark programming and analytics
- noSQL database like MongoDB and Spark that I am learning from this class will be useful

5. What data are you looking to use specifically?

Multiple Open Target genetics datasets, which can be downloaded from the website: ftp://ftp.ebi.ac.uk/pub/databases/opentargets/genetics.

1) Most of the data are either in the format of multiple JSON files or parquet files.
2) The total size of all datasets are about 1.2 Terabytes, size varying from 10M to 700G. I am probably going to pick only a few datasets for the demonstration of data modeling for this class. For examples,

- **Credible set files (JSON): ~ 2Gb**
  ftp://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/20022712/v2d_credset/
- **Variant record files (JSON): ~ 45G**
  ftp://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/20022712/lut/variant-index/
- **Study record files (JSON): ~10M**
  ftp://ftp.ebi.ac.uk/pub/databases/opentargets/genetics/20022712/lut/study-index/

3) The main business question I want to answer: what are the coding variants with high posterior probability for a list of EFOs / trait categories of interest in the study
4) Fields in the datasets that we can use to join to other datasets. For example,
   *studies. study_id = v2d_credset.studid;*
   *v2d_credset.(tag_chrom, tag_pos, tag_ref,tag_alle) = variant_index(chr_id, position, ref_allele, alt_allele)*

# Scalable Data Storage and Analysis with TileDB and pySpark

POTENTIAL SESSIONS:

1. Introduction to TileDB (week of 10/28)
    a. TileDB Array concept
    b. TileDB Data Structure
2. Build database with TileDB(weeks 11/4 ~ 11/25)
    a. Installation
    b. Build Arrays
    c. Data injection
3. Interact with TileDB using pySpark (weeks 11/25~ 12/2)
    a. Installation of pySpark
    b. Read and interact with TileDB Arrays using pySpark
    c. Perform analysis using pySpark
4. Compare to MongoDB and/or HDFS (depending on the progress)
    a. Installation of MongoDB and pyMongo
    b. Data injection
    c. Interact MongoDB with pySpark
    d. Perform data analysis using pySpark

POTENTIAL SOURCES:

1. TileDB documentation: https://docs.tiledb.com/main/
2. pySpark SQL documentation: http://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html
3. pyMongo: https://pymongo.readthedocs.io/en/stable/tutorial.html