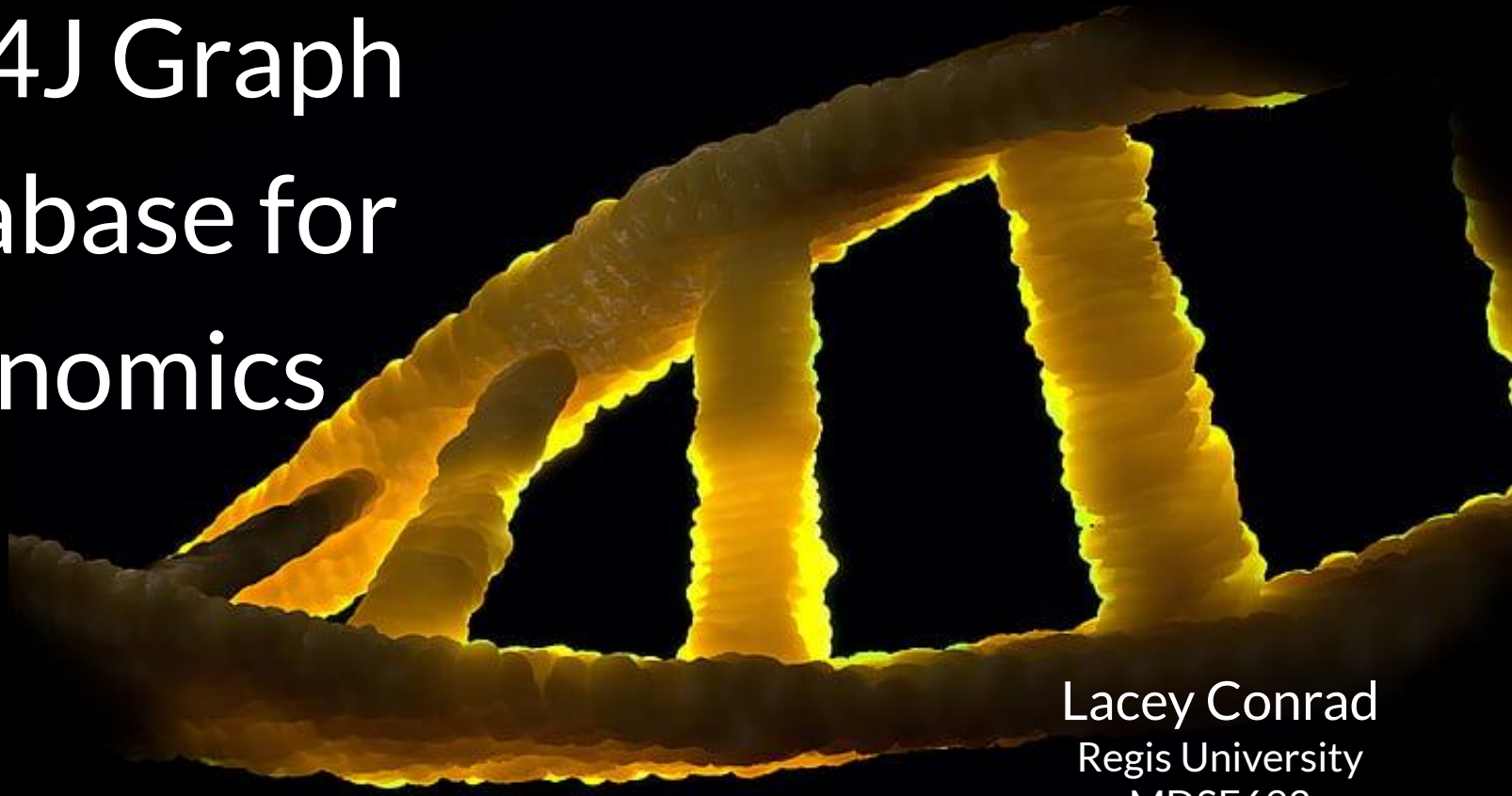
Neo4J Graph Database for Genomics



Lacey Conrad
Regis University
MDSE692
Data Engineering Practicum - 1

Outline

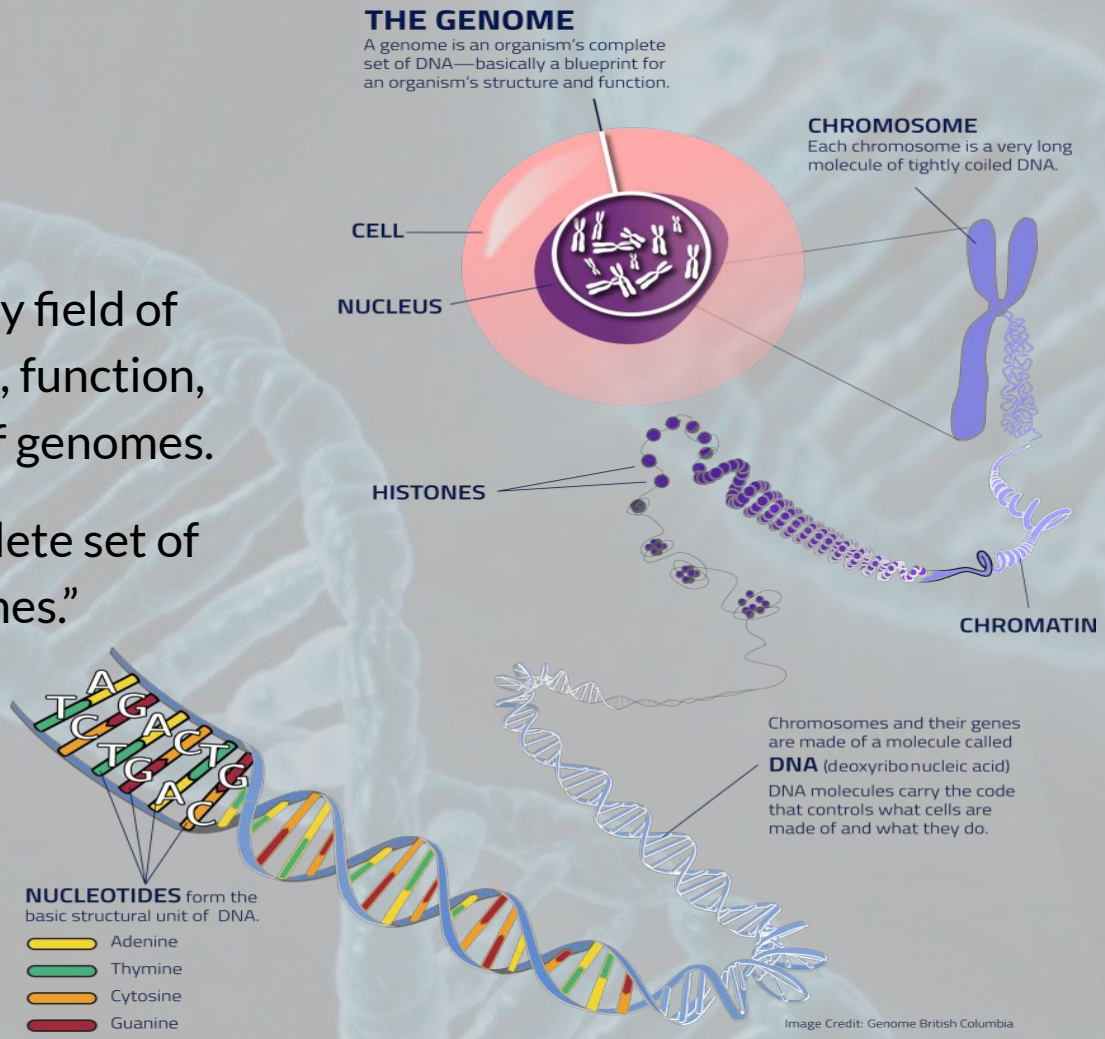
- ❑ Background: Genomics and biological data
- ❑ Project outline
- ❑ Goals and stages
- ❑ Results
- ❑ Project problems
- ❑ Status and future directions

What is Genomics?

From Wikipedia:

"**Genomics** is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes.

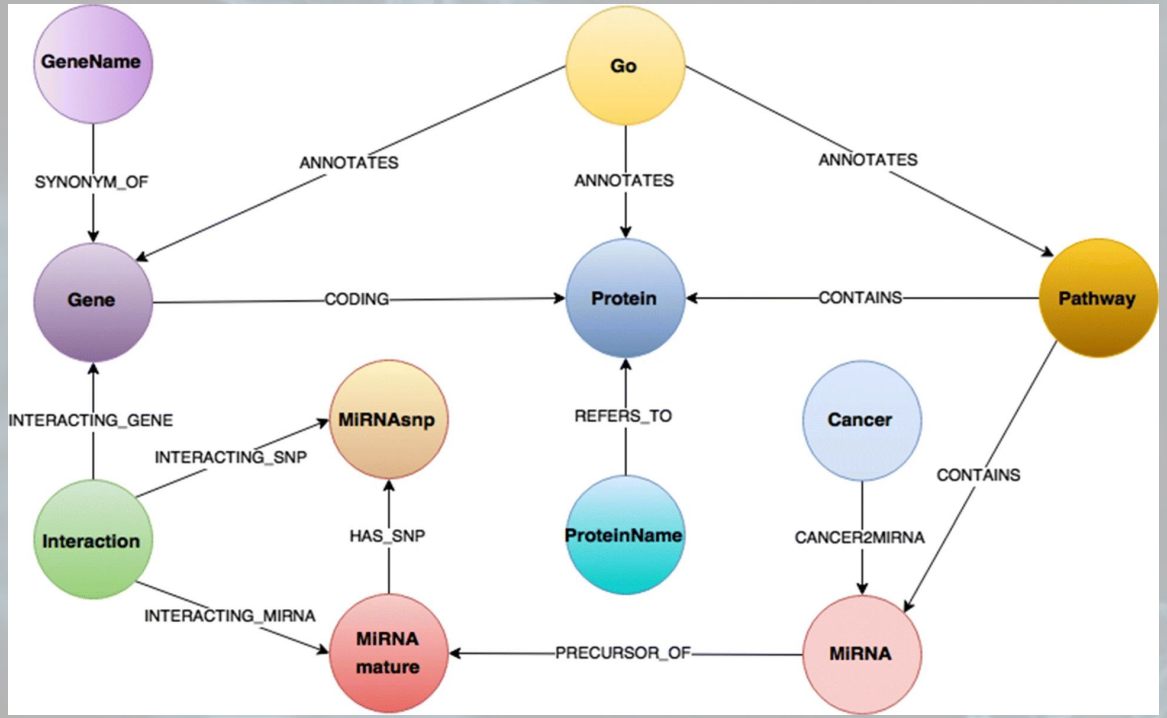
A **genome** is an organism's complete set of DNA, including all of its genes."



Biological data is poorly suited for storage in relational databases.

- Biological data sets have interconnected, overlapping data.
- Data from genomics can be very heterogeneous and come from different levels of biology.
- A genome contains many genes, and genes can occur in multiple genomes (many-to-many relationship).
  a. In a DBMS this results in many joins.
- A relational database makes it difficult to study the evolutionary history of a gene.
- Genetic variants are continually being annotated. The relational database schema has to be revised for the new annotation sources.

Protein Network Graph Models

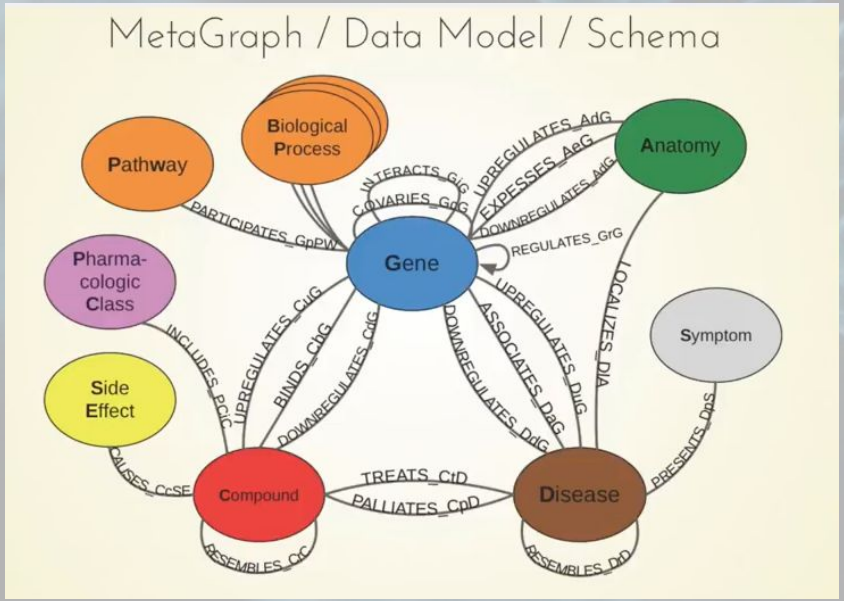


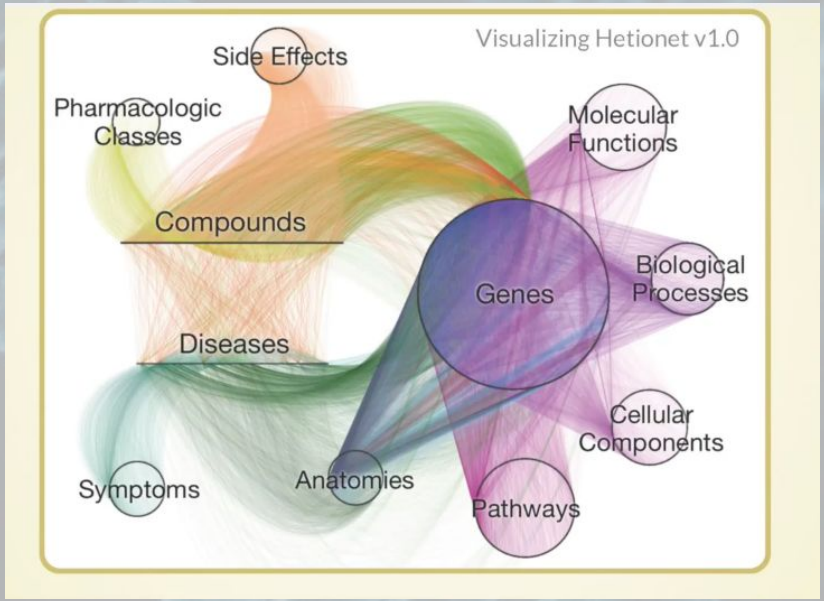
(https://neo4j.com/blog/data-management-systems-biologymedicine/)

Hetionet

- Hetionet: a biology graph database system designed for drug repurposing, finding new uses for existing drugs.

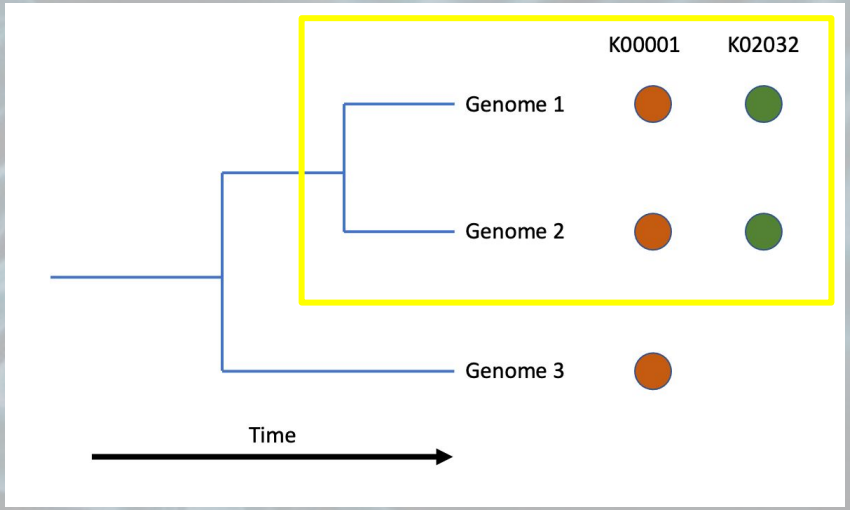
- It's much cheaper and safer to find a new use for drugs that we already know are safe for humans rather than designing a new compound from scratch.



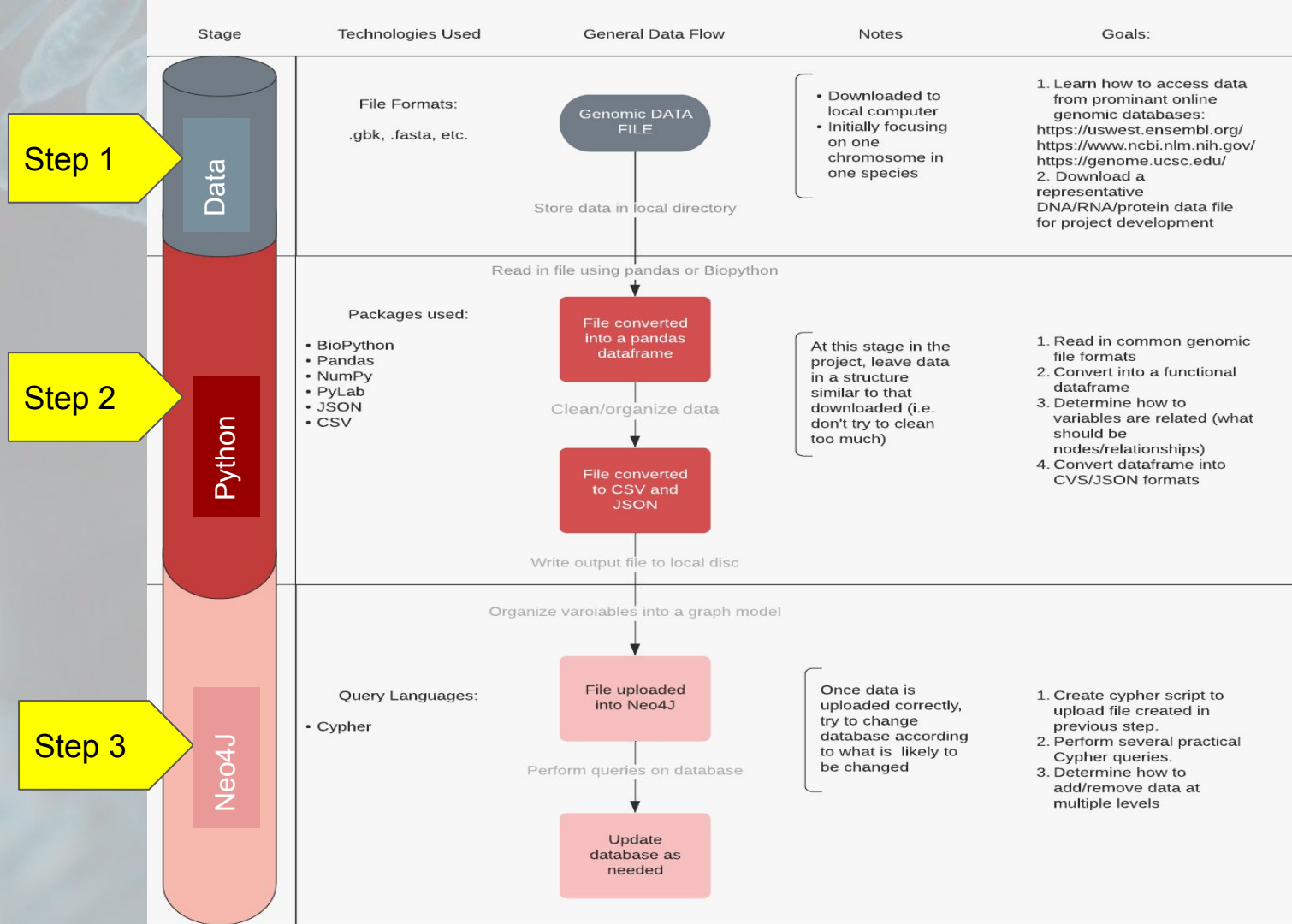


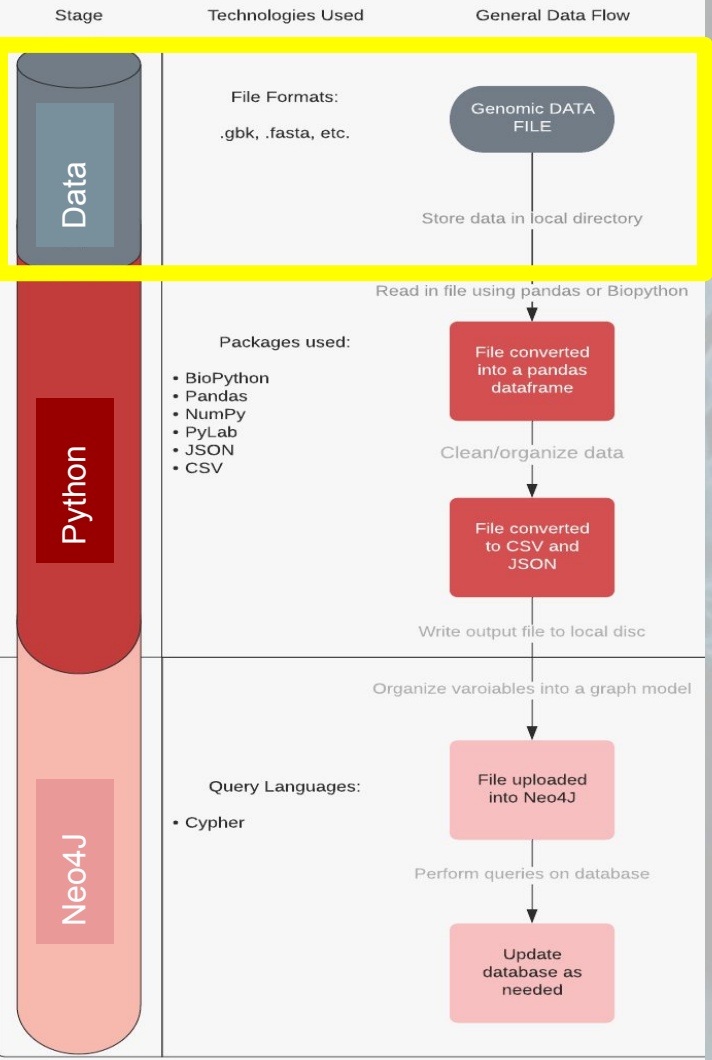
Why is this interesting?

- Genes with similar DNA sequences tend to fulfill similar functions. These genes can be grouped into clusters. These clusters can be queried for in graph databases.
- The clusters allow scientists to discern what an organism can do biochemically.
- The evolutionary history of a cluster of genes can be determined by where it occurs phylogenetically.
- Example: Genomes 1 and 2 are the most closely related since they share more genes than they do with genome 3.



Project Outline

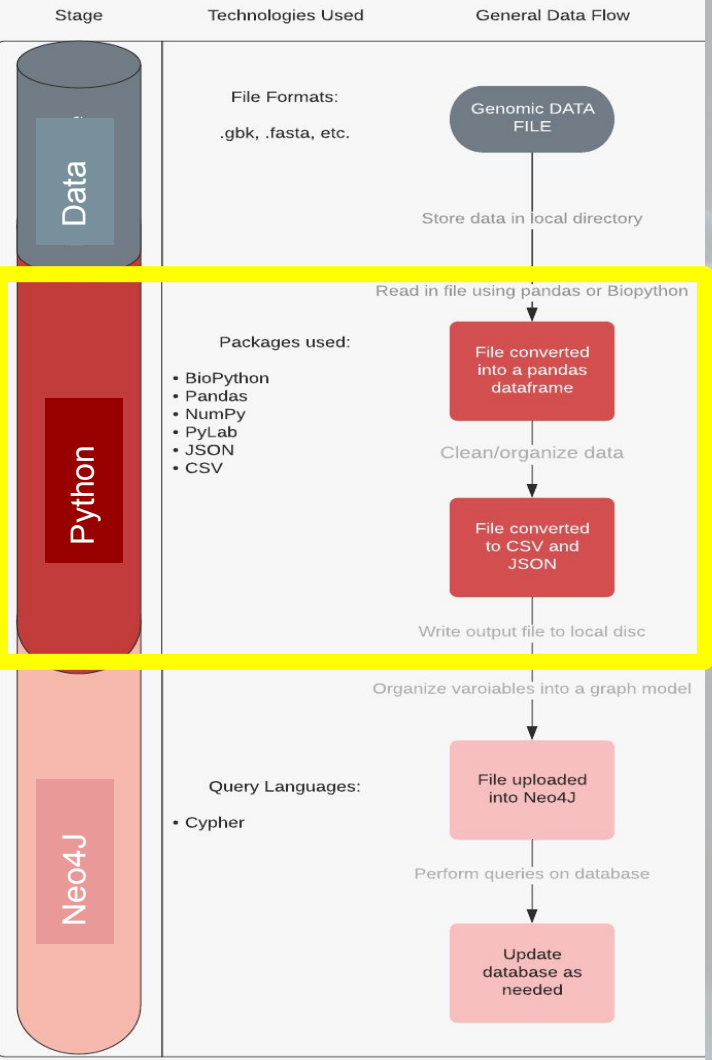




STAGE 1 - Data Collection

Goals for Step 1:

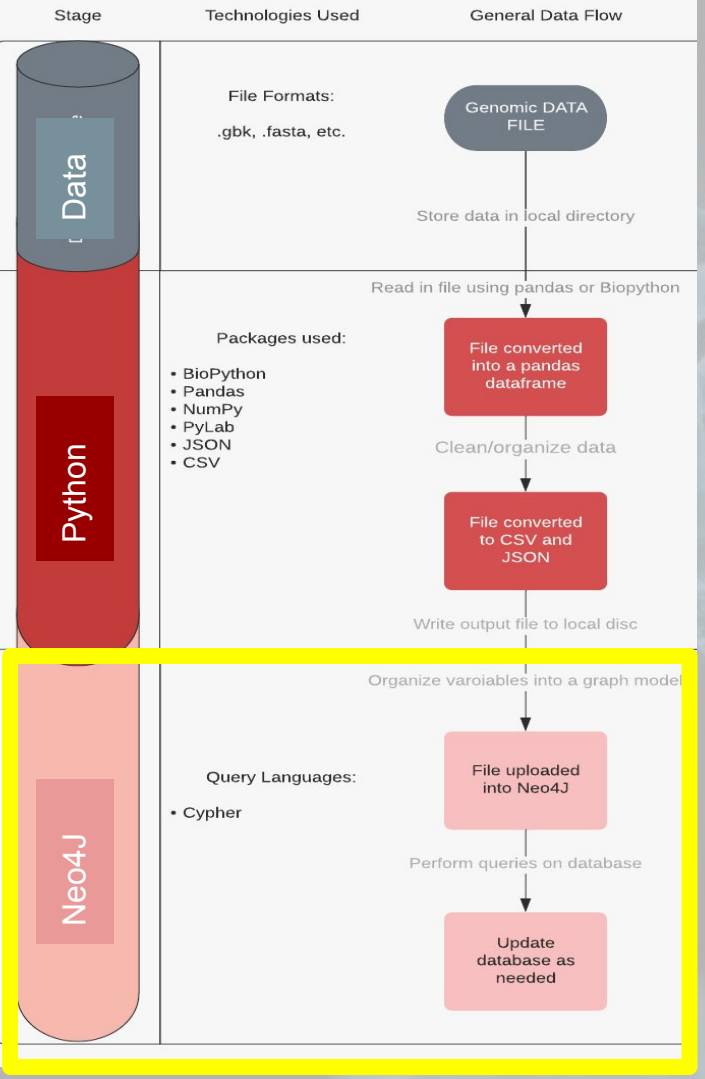
1. Research what data is included in each of the common file formats (fasta, gbk, etc.).
2. Visit genomic database sites and pick one to use.
3. Download several file types to the local machine and explore/clean in Python.
4. Determine what parts of the genomic data would logically fit into a graph database.



STEP 2 - Data preparation and model creation

Goals for Step 2:

1. Observe the relationships between variables and create a model of what a gene expression network would look like in a graph database.
2. Create a graph schema of the previous goal (1) for Neo4J.
3. Clean data file from Step 1 in such a way it can be easily uploaded into Neo4J.
4. Save resulting file as a CSV file.



STEP 3 - Upload into Neo4J

Goals for Step 3:

1. Learn how to upload the data file.
2. Learn Cypher queries to create database.
3. Determine how well it represents the system.
4. Investigate how changes are made to the database.

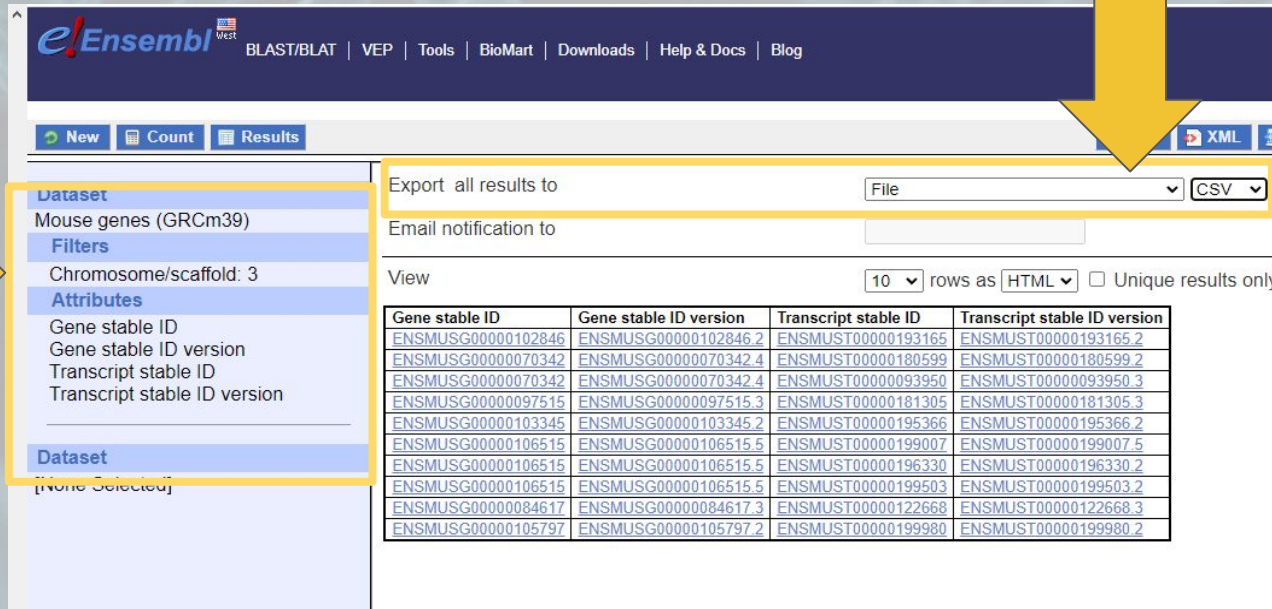
Results: Ensembl - BioMart

- Because of its easy-to-use data-mart, the genomics database at https://uswest.ensembl.org/index.html was chosen as the main data access site.

- Using Ensembl, I collected data on the first 100 genes on chromosome 3 in the mouse, chicken, and human.

File download options

Data Filters



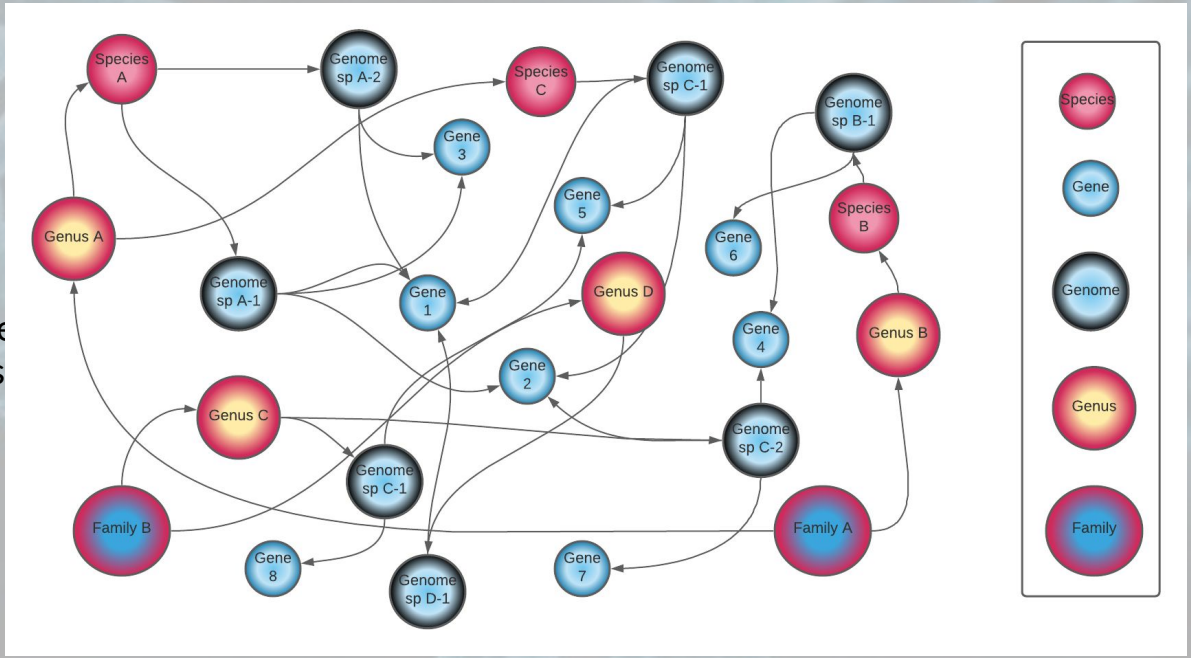
Gene stable ID	Gene stable ID version	Transcript stable ID	Transcript stable ID version
ENSMUSG00000102846	ENSMUSG00000102846.2	ENSMUST00000193165	ENSMUST00000193165.2
ENSMUSG00000070342	ENSMUSG00000070342.4	ENSMUST00000180599	ENSMUST00000180599.2
ENSMUSG00000070342	ENSMUSG00000070342.4	ENSMUST00000093950	ENSMUST00000093950.3
ENSMUSG00000097515	ENSMUSG00000097515.3	ENSMUST00000181305	ENSMUST00000181305.3
ENSMUSG00000103345	ENSMUSG00000103345.2	ENSMUST00000195366	ENSMUST00000195366.2
ENSMUSG00000106515	ENSMUSG00000106515.5	ENSMUST00000199007	ENSMUST00000199007.5
ENSMUSG00000106515	ENSMUSG00000106515.5	ENSMUST00000196330	ENSMUST00000196330.2
ENSMUSG00000106515	ENSMUSG00000106515.5	ENSMUST00000199503	ENSMUST00000199503.2
ENSMUSG00000084617	ENSMUSG00000084617.3	ENSMUST00000122668	ENSMUST00000122668.3
ENSMUSG00000105797	ENSMUSG00000105797.2	ENSMUST00000199980	ENSMUST00000199980.2

Results - Graph modeling of genomics data

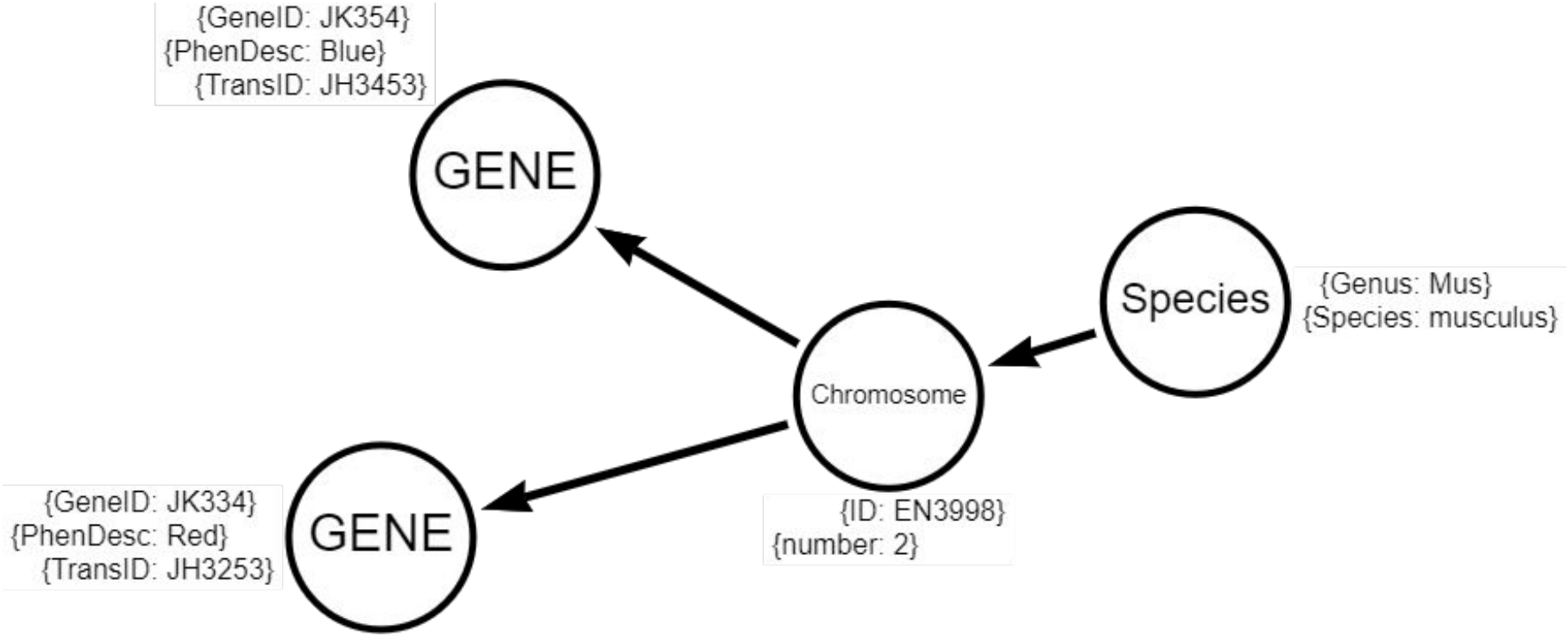
- A graph model of the data was created, and then I made a "practice" database in Neo4J (without data).

- As a result, I was able to determine how to structure the relationships and what data to use.

- A schema was created using the Arrows application.



Neo4J/Cypher graph - Arrows web application



Results: Completed Neo4J graph

- The data was placed into the Neo4J's import directory.
- Cypher scripts were used to upload data and create relationships.

Node creation

```
CREATE (:Order {name: 'Rodentia'})
```

Loading in data

```
LOAD CSV WITH HEADERS FROM
'file:///smmouse' AS line
CREATE (:Gene {name: line.transcript_id})
```

Adding relationships

```
MATCH (a:Chromosome),
(b:Gene),(s:Species)
WHERE a.number = '11' AND s.species =
'musculus'
CREATE (a)-[r:CONTAINS]->(b)
CREATE (s)-[t:CONTAINS]->(a)
RETURN type(r),(t)
```

```
MATCH (a:Chromosome), (b:Gene)
WHERE a.number = '11'
CREATE (a)-[r:CONTAINS]->(b)
RETURN type(r)
```

Final Database

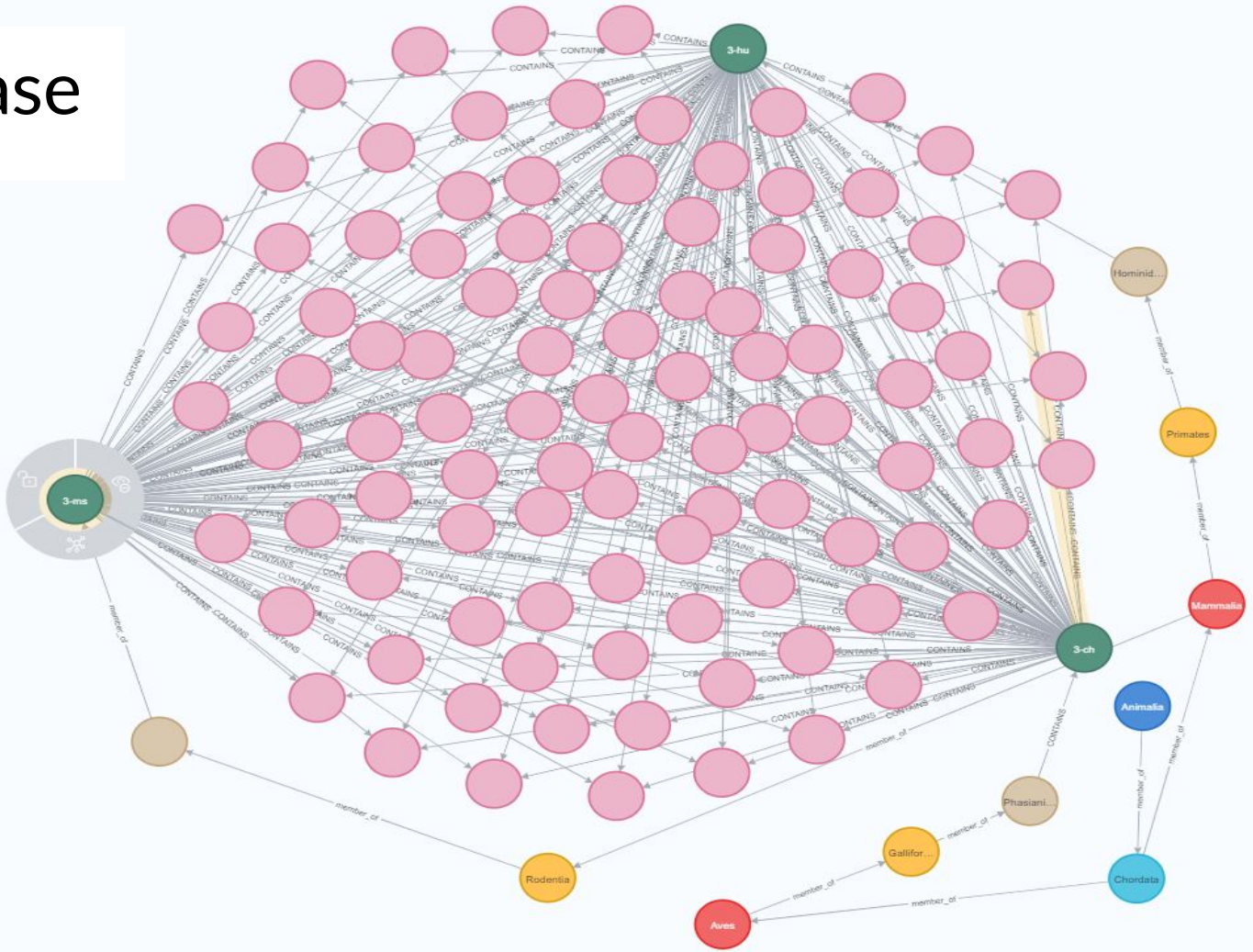
KEY:

Chromosome/ Species:

Gene

Classification nodes:

-Family
-Order
-Class
-Phylum
-Kingdom



Problems with project

1. Uploading local files to Neo4J.

- Neo4J will only read-in files in its import directory.
- Each Neo4J server has its own import folder. If you create a second or third Neo4J server on your local machine, you will need to make sure you have your data in the correct import folder for each.
- This isn't clearly mentioned in the documentation.

2. Genomic data storage variety.

- Many different file types.
- Each file type contains different information, much of it overlapping.
- Most countries have their own website and database for submitted genetic data.
- This also means that data standards and lexicon may not be globally consistent.

Project status

- The major project goals have been met.
- To get data from Ensembl, I initially planned to use an API. However, the API proved to be too difficult to use. Instead, I downloaded the CSV file directly.

Future directions

- Directly link genomic data with Python using Ensembl's API (i.e., no need to download data manually).
- Using Neo4j inside Python.
- Using Neo4J, create a more complex network model (i.e., add more data to the nodes and use the built-in algorithms).
- Integrating multiple sources of genomic data into Neo4J using Python.

Conclusions

- The amount of variability in genomic data formats must be reduced.
- Neo4J appears to be an appropriate database to handle complex genomic data.
- The design of a model and schema for Neo4J was less time consuming than setting up a relational database schema, and was more intuitive and reflective of biological relationships.
- Graph databases are more approachable to non-programmers due to their intuitive nature, however there still appears to be a reluctance to embrace them.
- Data practices in genomics should be standardized. It is very likely that the genomic database landscape will turn into a swamp if a set of best practices is not adopted soon.

Questions?

How much data does genomics produce?

Statistics from genome.gov:

- Genomics research will likely generate between **2 and 40 exabytes** of data within the next decade.

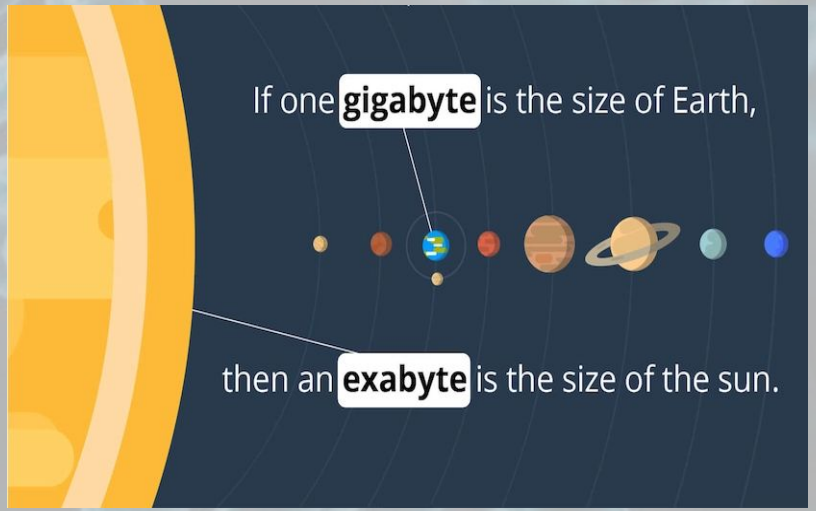
- Roughly **2 to 40 billion gigabytes** of genomics data are generated each year.

- The data of a single human genome sequence takes up **200 gigabytes**.

1 gigabyte = 1,000,000,000 bytes

1 exabyte = 1,000,000,000,000,000,000 bytes

Five exabytes could store all of the words ever spoken by human beings.



Why is this interesting?

There are many uses for graph databases in biology.

- Graph databases have been used to help detect cancer causing genes by researching which genes are expressed along with the cancer (gene coexpression).

- Graph databases allow for other types of biological network analysis, such as population genetics.

- To the right is the gene network for the tiger shark. Circle size represents the number of individuals with a particular gene, and the color represents location.