Databases

What is a database?

- A systematic (structured) collection of data records for the purpose of retrieving records upon a query.
- Sequential storage:
  - Index cards, magnetic tape
  - Exhaustive search, but fast processing of a batch when you reach it
- Random access:
  - Retrieve a record using an index
  - Index has to be built, updated; slower to add new records.

Tables

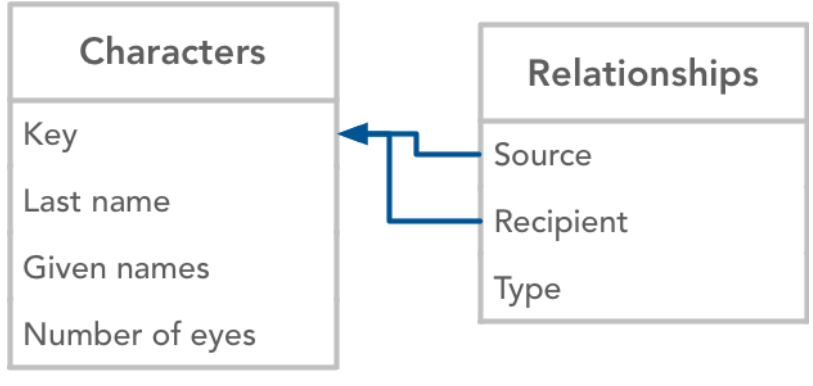
- A set of related data, where each record is represented by a *row* and each variable (field) appears in a *column*.
- For example:

Key	Last name	Given names	Number of eyes
1	Fry	Philip J.	2
2	Leela	Turanga	1
3	Rodriguez	Bender Bending	2

- Can you add another row to this table?

Database schema

- A formal model of how data are organized.
- Basically, how fields in different tables are related to each other.
- The defining characteristic of relational databases.



Relational databases

- A relational database is implemented with a *structured query language* (SQL).
- SQL is used to create, update, and query a relational database.
- First developed by IBM in 1980's, based on work by Ted Codd.
- Later other companies developed commercial implementations of SQL, such as Oracle.
- MySQL is an open-source implementation.
- Many languages now support SQL in different ways.

Non-relational databases

- Although relational databases are very common, some agencies dealing with massive amounts of data (*e.g.*, Google, Amazon, Netflix) have adopted a different approach.
- A non-relational database or *NoSQL* has no fixed database schema - great when database has to adapt to new demands.
- NoSQL allows the addition of new fields in real-time.
- Supports distributed computing (multiple servers/sites).

Querying a database

- SELECT is an SQL command that returns one or more fields from records from one or more tables

```
select LAST_NAME from CHARACTERS where NUMBER_OF_EYES is 2
```

```
Fry
Rodriguez
```

Is a spreadsheet application the same as a DMS?

- DMS = database management system.
- A spreadsheet app like MS Excel can store data in a structured format.
- A relational database can be built in Excel using a "master" spreadsheet.
- Performance of Excel does not scale with size of database.
- Database management systems like SQL enable multiple users to query and modify the database -- Excel is a single-user application.

Front-ends

- Most web pages are *dynamic content* - the information that gets displayed in your browser is the result of a database query.
- In contrast, the web resources for this class are *static* web pages that were rendered from text files in a format called *Markdown*.
- Code responsible for database transactions take place in the *back end*.
- The *front end* is what you see: a web interface with forms and other data from which the back-end will compose a query.

Genbank

- https://www.ncbi.nlm.nih.gov/genbank/
- Perhaps the largest public repository of genetic sequences in the world.
- Maintained by the US National Center for Biotechnology Information (NCBI)
- Have you used Genbank for your research? Have you deposited sequences in Genbank?

Origins of Genbank

- The Atlas of Protein Sequence and Structure (Dayhoff and Ecks)
- Dr. Margaret Oakley Dayhoff was the first professional curator of a sequence database and arguably the founder of bioinformatics (*more later!*).
- Both Dayhoff and the Los Alamos National Laboratory (headed by Walter Goad) submitted proposals to NIH to form what is now Genbank.
- Dayhoff managed the *Atlas* as a proprietary resource, whereas Goad intended the database to be a free public resource.

Origins of Genbank

> *When science funding agencies and the scientific community finally recognized the potential of sequence databases for the production of knowledge, [Dayhoff] lost the contract to build such a database to a physicist with no prior experience in sequence collecting [...] she died of heart failure eight months after the contract was awarded.*

Bruno J. Strasser, *The Experimentalist's Museum*.

Demonstrate a Genbank query

NCBI Short Read Archive

- Genbank mostly comprises sequences generated by Sanger method
- Next-generation sequencing (NGS) platforms now produce millions of sequences in a single run.
- The SRA stores NGS data - requires special toolkit to download records.
- Data are annotated by:
  - Project metadata (SRP)
  - Experiment (SRX)
  - Run number (SRR)
  - Sample number (SRS)

Demonstrate fasterq-dump

In-class assignment

Find an infectious disease database

1. Record URL
2. Briefly describe the database's focus/purpose.
3. Provide the number of records, if available.
4. Compose a simple query and describe it here.
5. Provide the first result of the simple query in (4).

Further readings:

- Early History of SQL
- Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965
- The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine
- Understanding SRA Search Results