Session 2: Ecotoxicological Databases

Theoretical Ecotoxicology

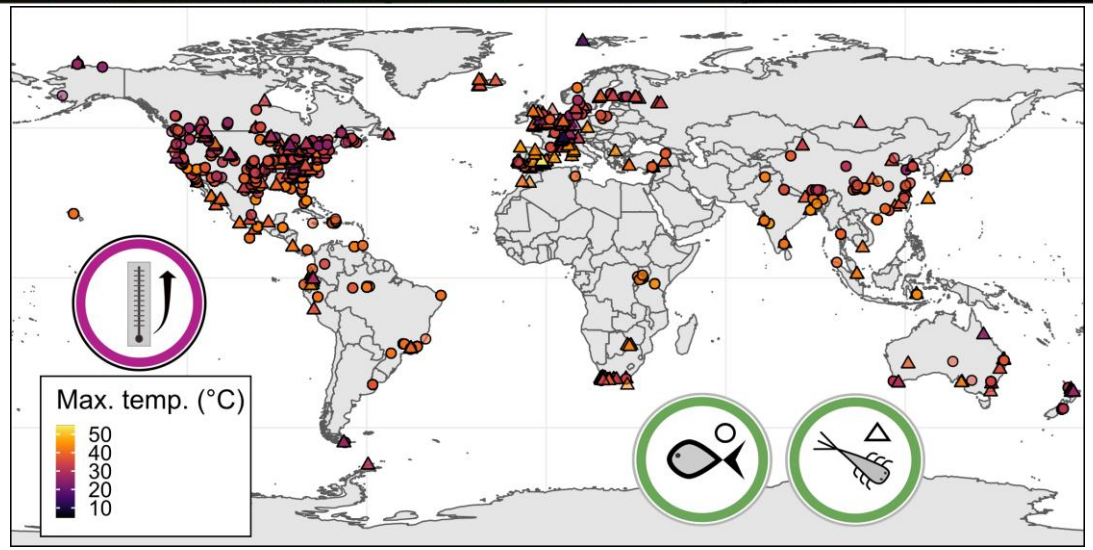
Helena S. Bayat

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

About me

- Helena S. Bayat
- BSc in Environmental Toxicology at UC Davis
  - » Mass spectrometry of aged smoke compounds
- MSc in Environmental Science at University of Copenhagen
  - » Modelling mixture toxicity under stressful food regime
- Currently doing PhD with RESIST project
  - » Multiple stressors in freshwater
  - » Built a database for thermal tolerance



Bayat et al. 2024

Learning objectives

- Understand what a database is and relevance to ecotoxicology
- Introduce a selection of ecotoxicological databases
- Know where to look for ecotoxicological information

What are databases?

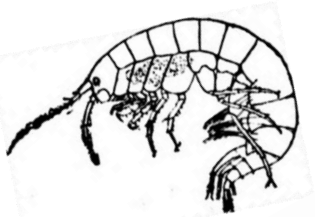
What are databases?

Why do we need them?

Consider:

- You have 5 sampling sites
  - » At each you have taken biological samples
    - › Invertebrates
    - › Fish
  - » You also have chemical data for each
    - › Chemical concentration
  - » How would you collect and store all this different data?

Consider:



Site id	lat	long	date
S1	49.2	8.0	10-10-2022
S2	49.3	8.2	10-10-2022
S3	49.5	8.1	11-10-2022
S4	50.0	7.9	12-10-2022
S5	48.4	8.5	12-10-2022

Chem sample id	Site id	Chemical id	Concentration
C1	S1	7440-50-8	5
C2	S1	2921-88-2	2
C3	S1	138261-41-3	1
C4	S2	7440-50-8	4

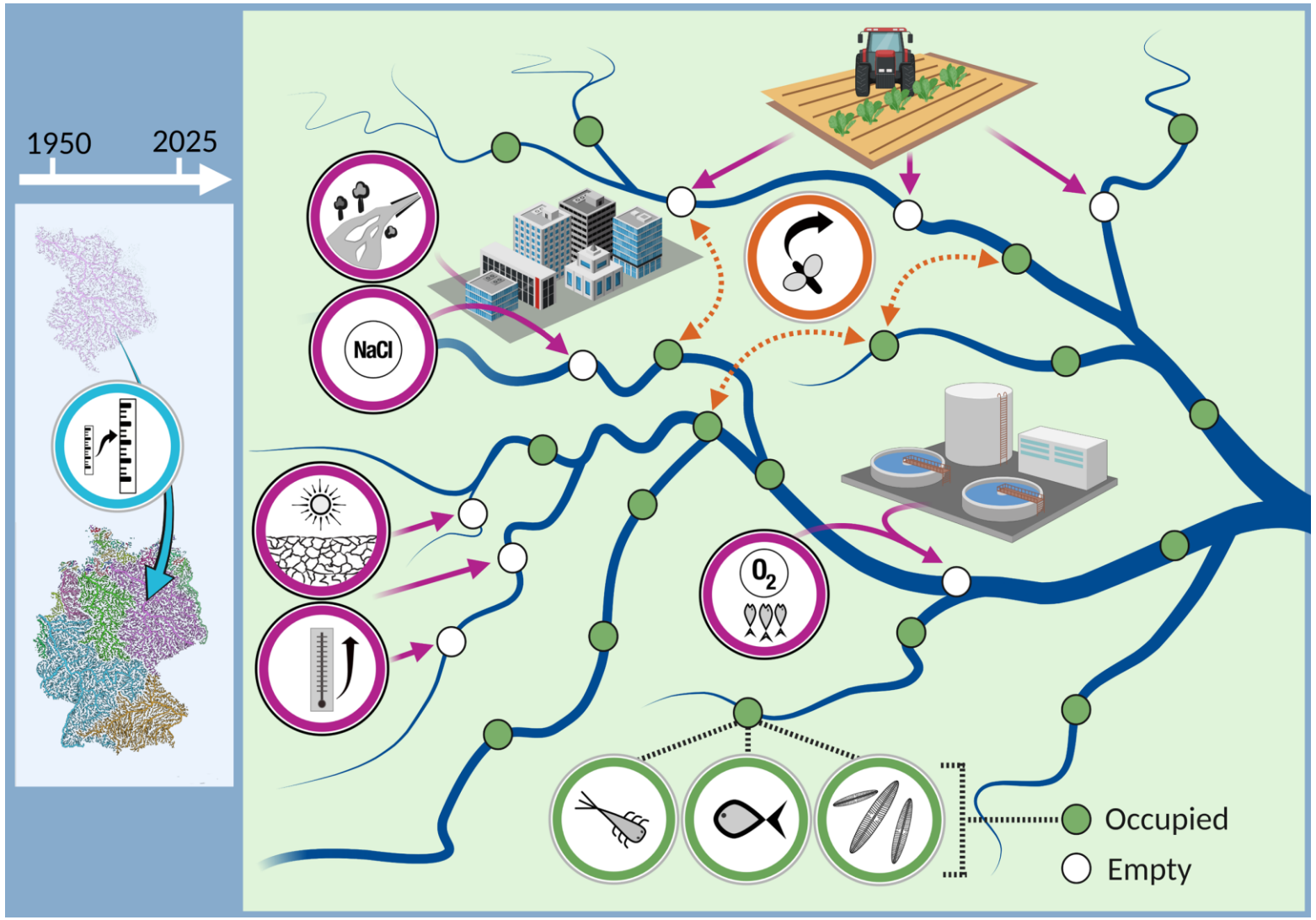
Chemical id	Chemical name	Usage date
7440-50-8	copper	8-2022
2921-88-2	chlorpyrifos	10-2022
138261-41-3	imidacloprid	7-2022

Fish sample id	Site id	Cottus	Salmo
F1	S1	5	2
F2	S2	2	3
F3	S3	3	4
F4	S4	4	2
F5	S5	1	0

Invert sample id	Site id	Gammarus	Hydropsyche	Rhyacophila
I1	S1	10	2	5
I2	S1	8	5	1
I3	S1	7	1	0
I4	S2	12	0	1
I5	S2	4	1	2

Tables in wide format

Consider:



Relational databases

- Relational database management systems (RDMS)
- Developed by Edgar F. Codd for IBM in 1970

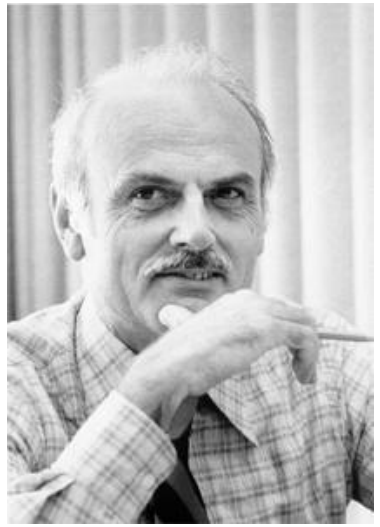
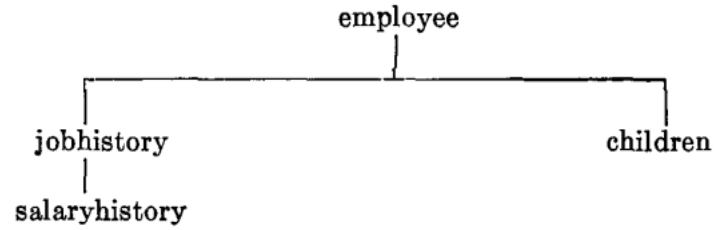


Image from IBM

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence be-



employee (*man#*, name, birthdate, jobhistory, children)
jobhistory (*jobdate*, title, salaryhistory)
salaryhistory (*salarydate*, salary)
children (*childname*, birthyear)

Fig. 3(a). Unnormalized set

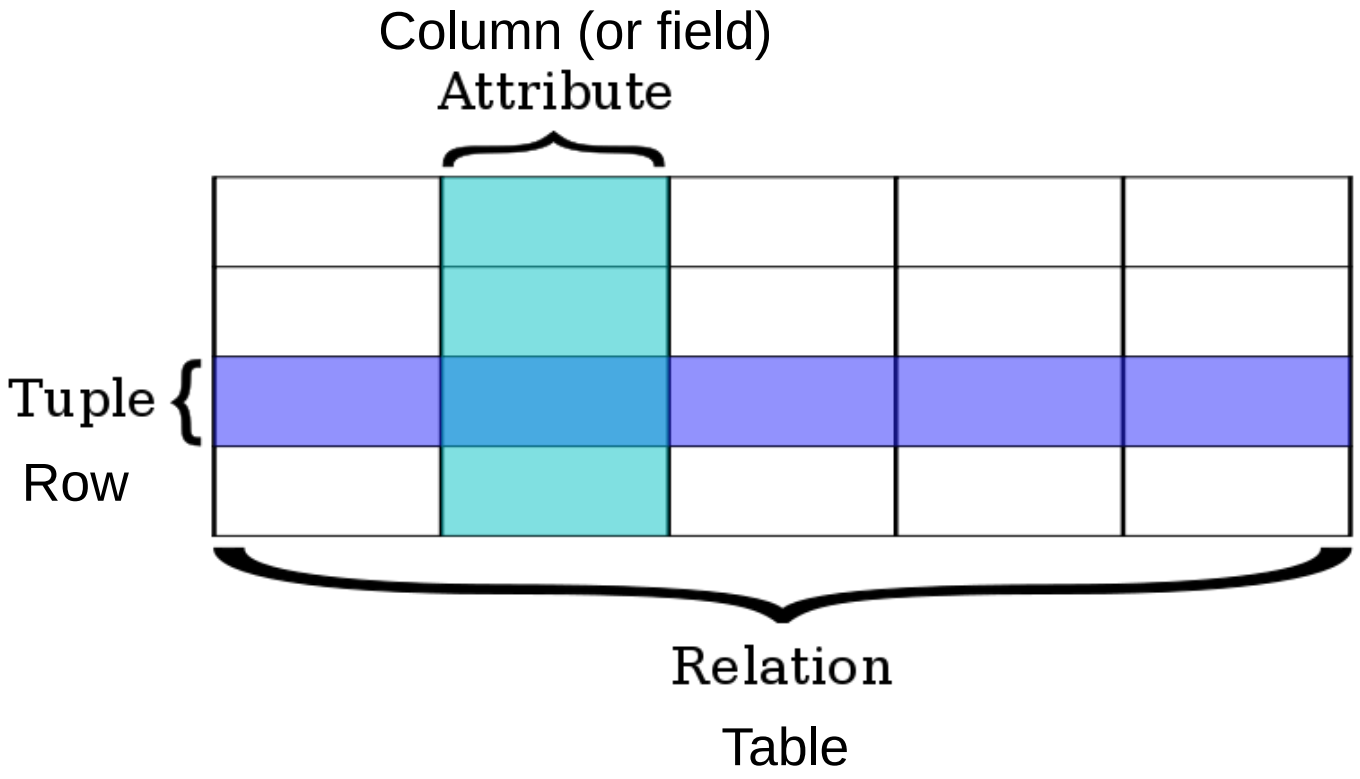
employee′ (*man#*, name, birthdate)
jobhistory′ (*man#*, *jobdate*, title)
salaryhistory′ (*man#*, *jobdate*, *salarydate*, salary)
children′ (*man#*, *childname*, birthyear)

Fig. 3(b). Normalized set

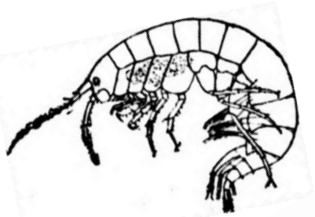
Relational databases

- Data is accessed using Structured Query Language (SQL)
  - » Standardized in 1987, ISO 9075:1987
- Tables connected by keys

```
database.schema.table.column
```



Consider:



Site id	lat	long	date
S1	49.2	8.0	10-10-2022
S2	49.3	8.2	10-10-2022
S3	49.5	8.1	11-10-2022
S4	50.0	7.9	12-10-2022
S5	48.4	8.5	12-10-2022

Chem sample id	Site id	Chemical id	Concentration
C1	S1	7440-50-8	5
C2	S1	2921-88-2	2
C3	S1	138261-41-3	1
C4	S2	7440-50-8	4

Chemical id	Chemical name	Usage date
7440-50-8	copper	8-2022
2921-88-2	chlorpyrifos	10-2022
138261-41-3	imidacloprid	7-2022

Fish sample id	Site id	Cottus	Salmo
F1	S1	5	2
F2	S2	2	3
F3	S3	3	4
F4	S4	4	2
F5	S5	1	0

Invert sample id	Site id	Gammarus	Hydropsyche	Rhyacophila
I1	S1	10	2	5
I2	S1	8	5	1
I3	S1	7	1	0
I4	S2	12	0	1
I5	S2	4	1	2

Tables in wide format

SQL software

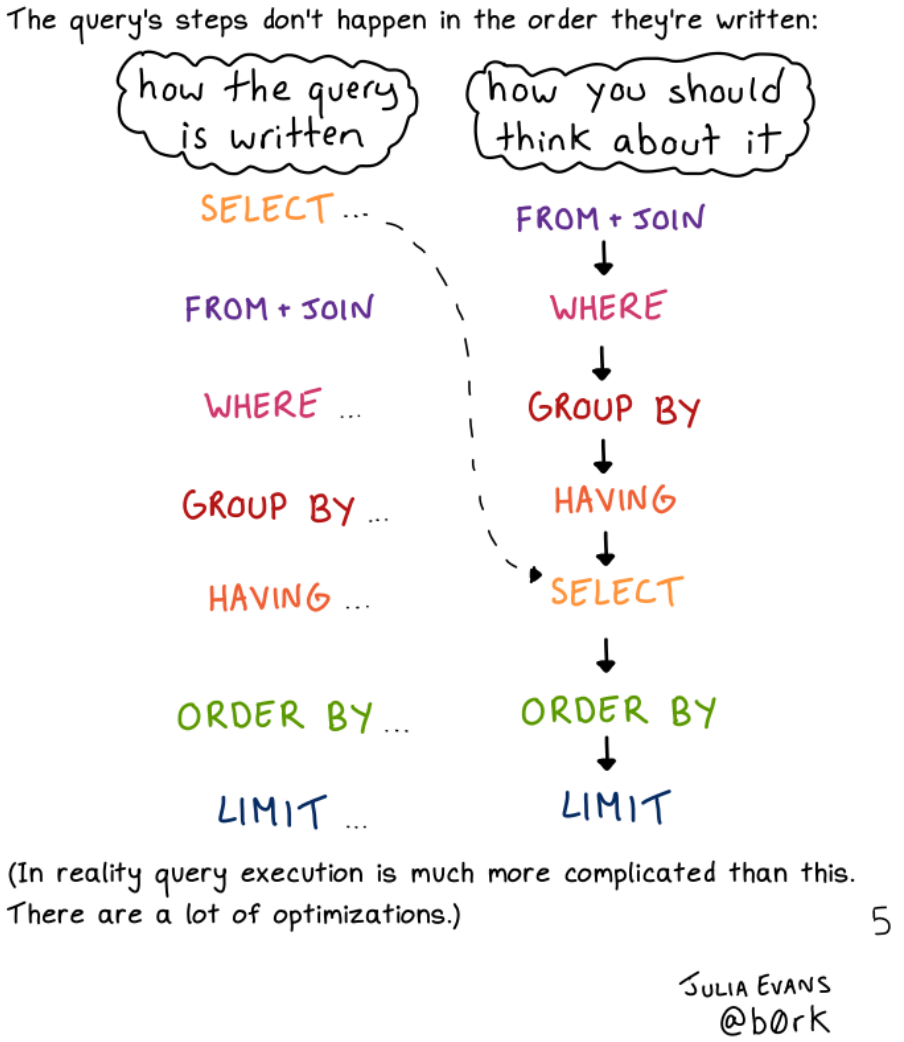
Free and open source (FOSS)

Proprietary

SQL queries

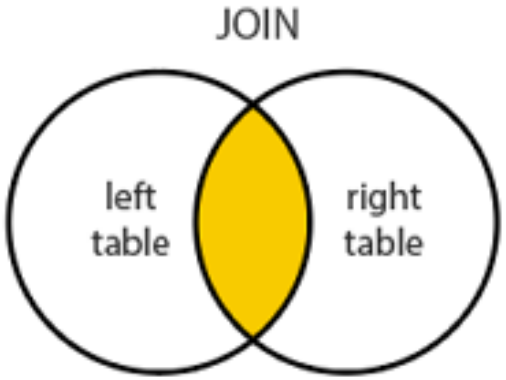
- Query: instructions for retrieving subsets of data

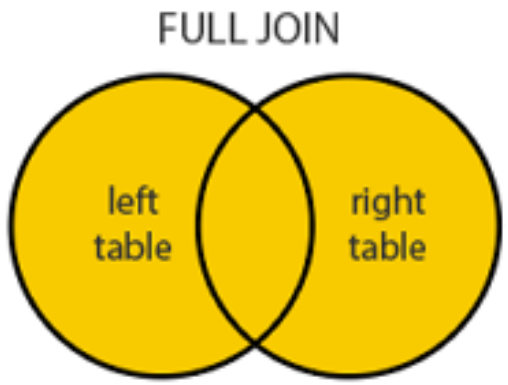
```
SELECT
  city_name,
  country,
  population
FROM cities
WHERE country = 'Austria'
```

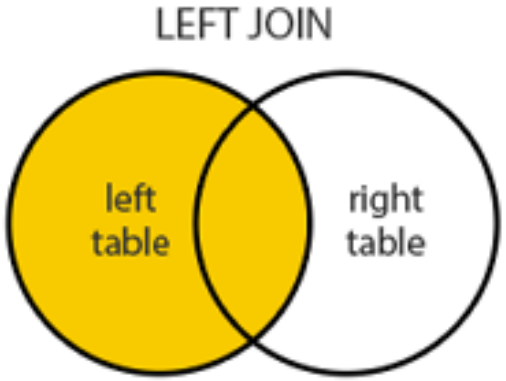


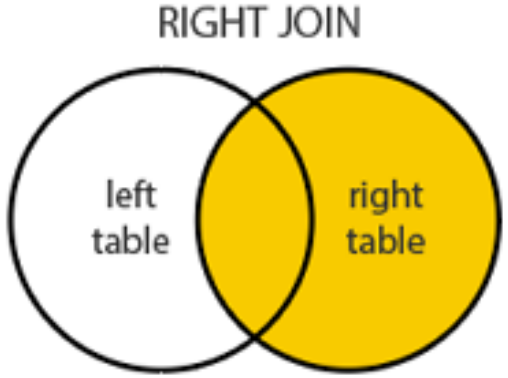
SQL queries

- INNER JOIN
- LEFT JOIN
- RIGHT JOIN
- FULL JOIN









SQL vs R

SQL

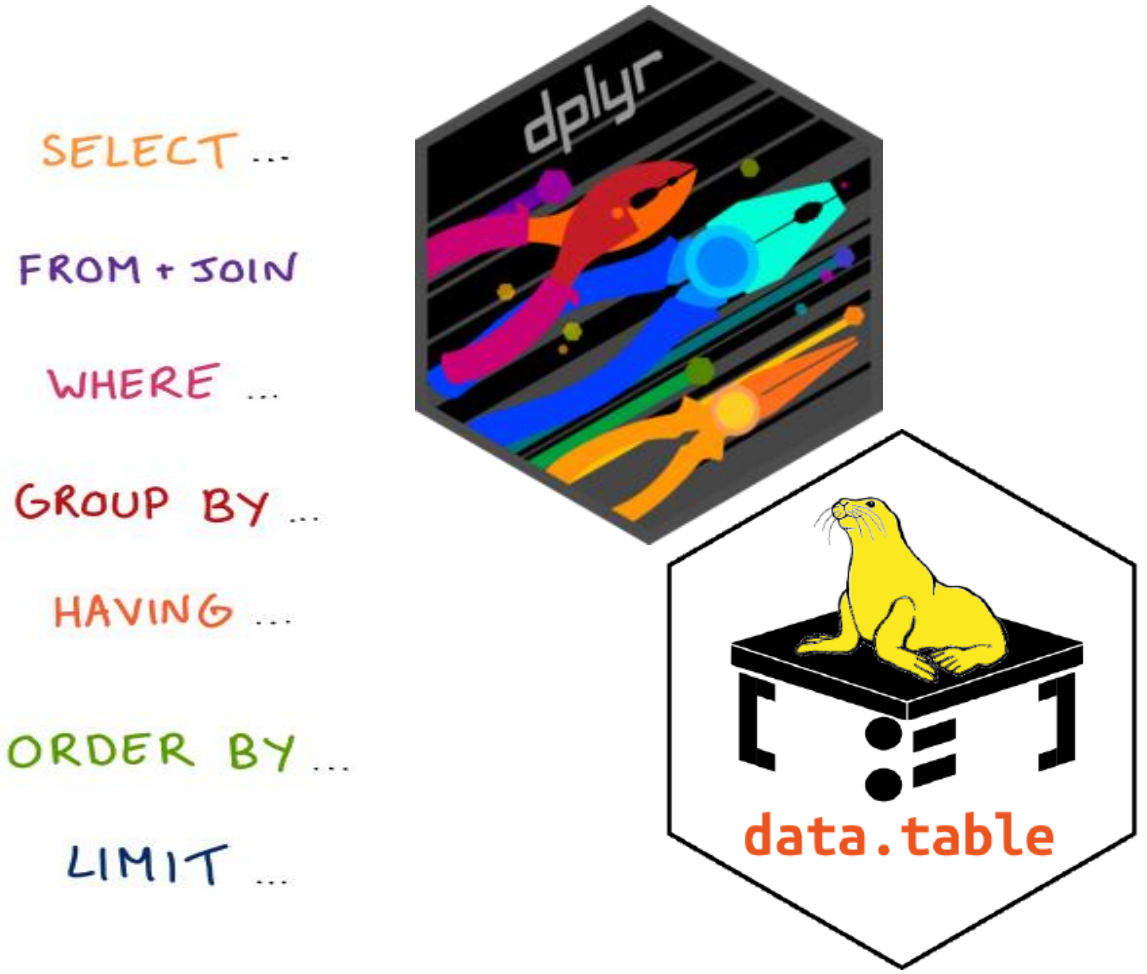
- Used with databases
- Data on disk
- Unlimited rows, 250-1600 columns
- Can be slow

R

- Programming language
- Data in memory
- Limited but fast
- Can use packages to connect to databases and send SQL queries

SQL logic in R

- Joins
- Selecting, filtering, grouping, etc
- Part of the tidyverse and data.table



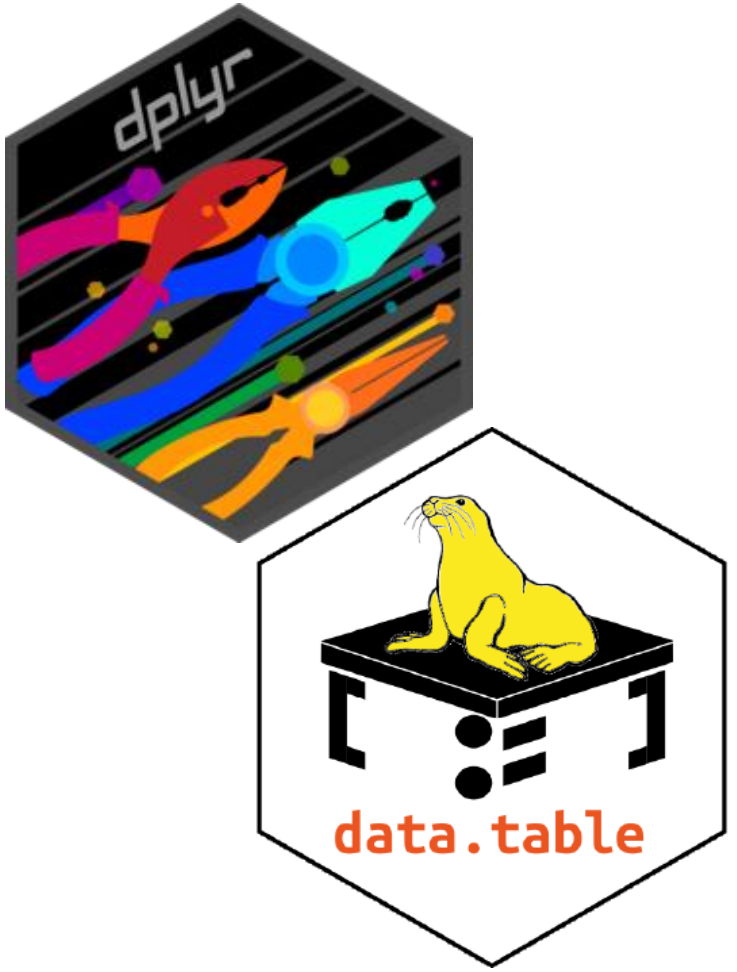
SQL logic in R

- Joins
- Selecting, filtering, grouping, etc
- Part of the tidyverse and data.table

For more information:

R for Data Science
https://r4ds.hadley.nz/

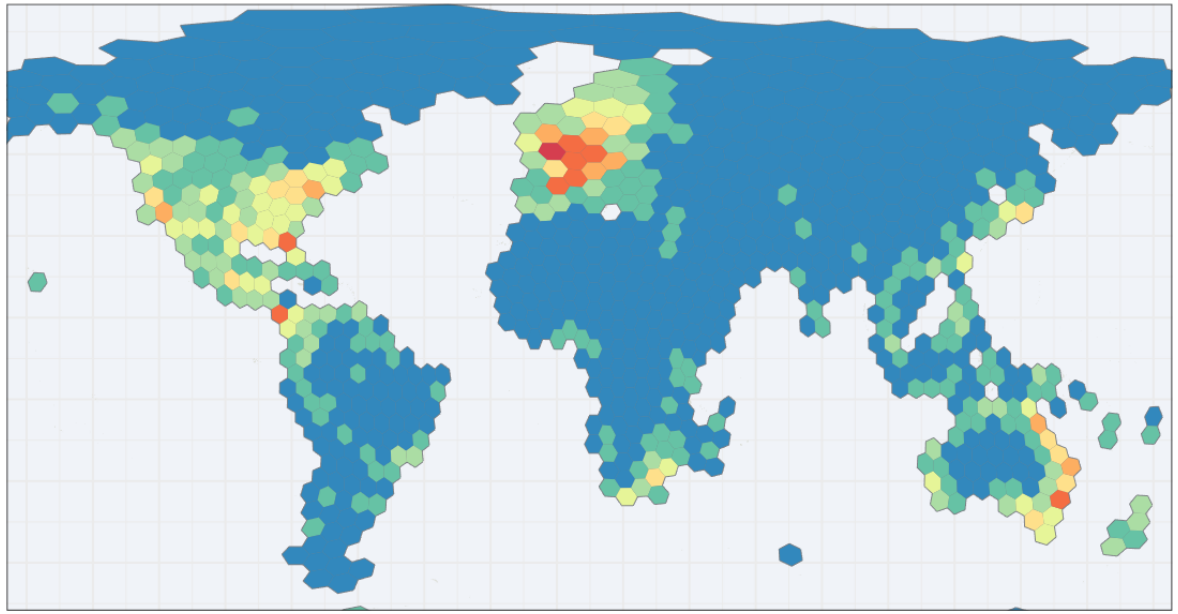
data.table
https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html



Large databases for ecological data

GBIF
Global Biodiversity Information Facility

WoRMS
World Register of Marine Species



freshwaterecology.info

EPA ECOTOX Database

- Large!
- Ecotoxicological test results from published studies and grey literature
- Automatically updated
- Hosted by the United States Environmental Protection Agency (USEPA)

Some reminders for working in R

- Working directory
- Make a folder → keep code organized
- Make comments
- Troubleshooting
  - » Check if working directory is set to correct folder
  - » Check spelling
  - » Extra or wrongly placed dots, dashes, etc cause errors in the code
  - » Use Stack Overflow or AI tools when stuck

Time for an exercise

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken