

In most data courses, we assume that data is available in CSVs, which then can be manipulated in R/Python using data frames

CSV

id	id2	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6
1	1	0.21	0.89	0.04	0.75	0.85	0.66
2	1	0.5	0.56	0.52	0.21	0.89	0.45
3	1	0.43	0.56	0.48	0.36	0.33	0.59
4	2	0.52	0.43	0.3	0.1	0.43	0.22
5	2	0.23	0.37	0.15	0.87	0.6	0.67
6	2	0.96	1	0.67	0.93	0.26	0.6
7	2	0.32	0.21	0.82	0.15	0.41	0.15
8	2	0.9	0.76	0.24	0.65	0.61	0.81
9	2	0.37	0.83	0.28	0.7	0.43	0.21
10	2	0.65	0.87	0.03	0.88	0.01	0.07
11	3	0.26	0.44	0.47	0.53	0.01	0.83
12	3	0.5	0.47	0.77	0.48	0.58	0.52

Data Frame

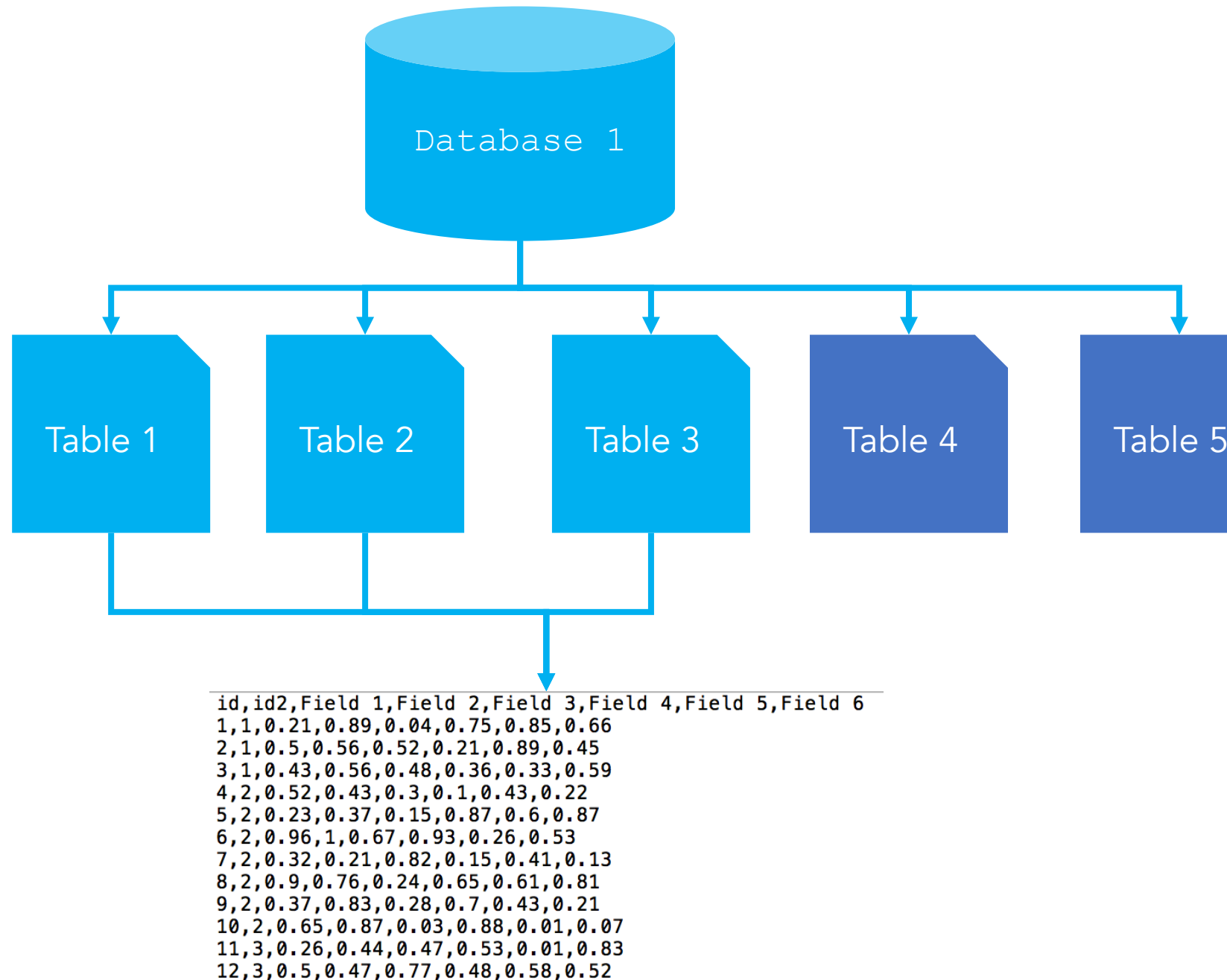
id	Field 1	Field 2	Field 3	Field 4
1	0.48864085	0.00366587	0.58474211	0.51938667
2	0.33876445	0.09769656	0.84166132	0.1030575
3	0.0278745	0.36673626	0.11164558	0.04700535
4	0.1763168	0.82417691	0.19927198	0.4810012
5	0.56424078	0.79502561	0.23455883	0.34132441
6	0.42115835	0.45674212	0.62794548	0.59208026
7	0.6158106	0.89786742	0.86752992	0.39943343
8	0.23344419	0.66442167	0.22883527	0.87996448
9	0.42737816	0.82922444	0.9400483	0.59218112
10	0.85595686	0.41012588	0.64374624	0.09399705
11	0.49237375	0.02784349	0.88235821	0.72331339
12	0.61511126	0.10467654	0.89272309	0.24495752



The CSV might be a custom view from a database.

Views are the results of a pre-defined query.

What if you need to build your own view?



Option #1

Download the whole database and export individual tables.

Not recommended as databases can be huge



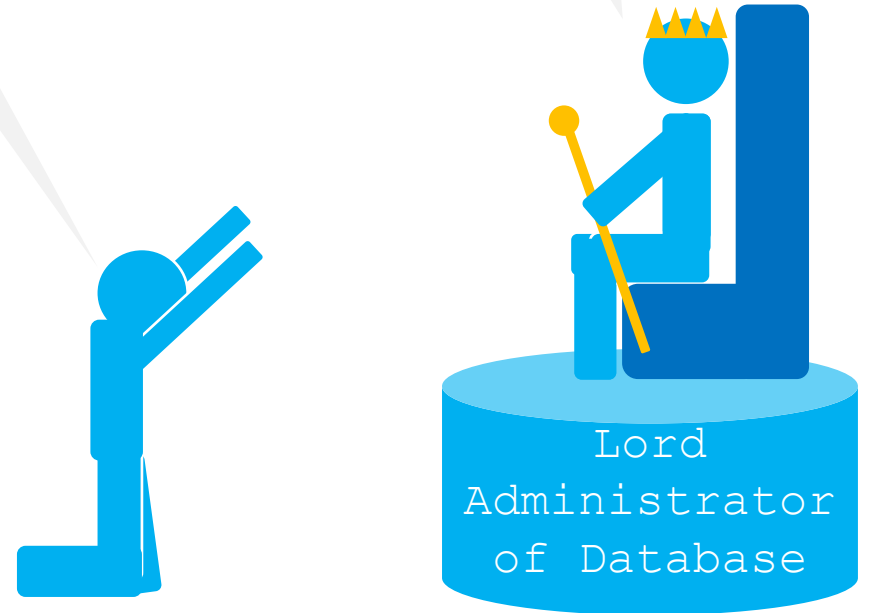
Option #2

Beg the database administrator (the gate keeper) to build specific views and extract tables for you.

Too many custom modifications means your work will be largely dependent on one person to give you data.

Oh, exalted one, I missed a variable in my previous request...

Silence, database mortal!

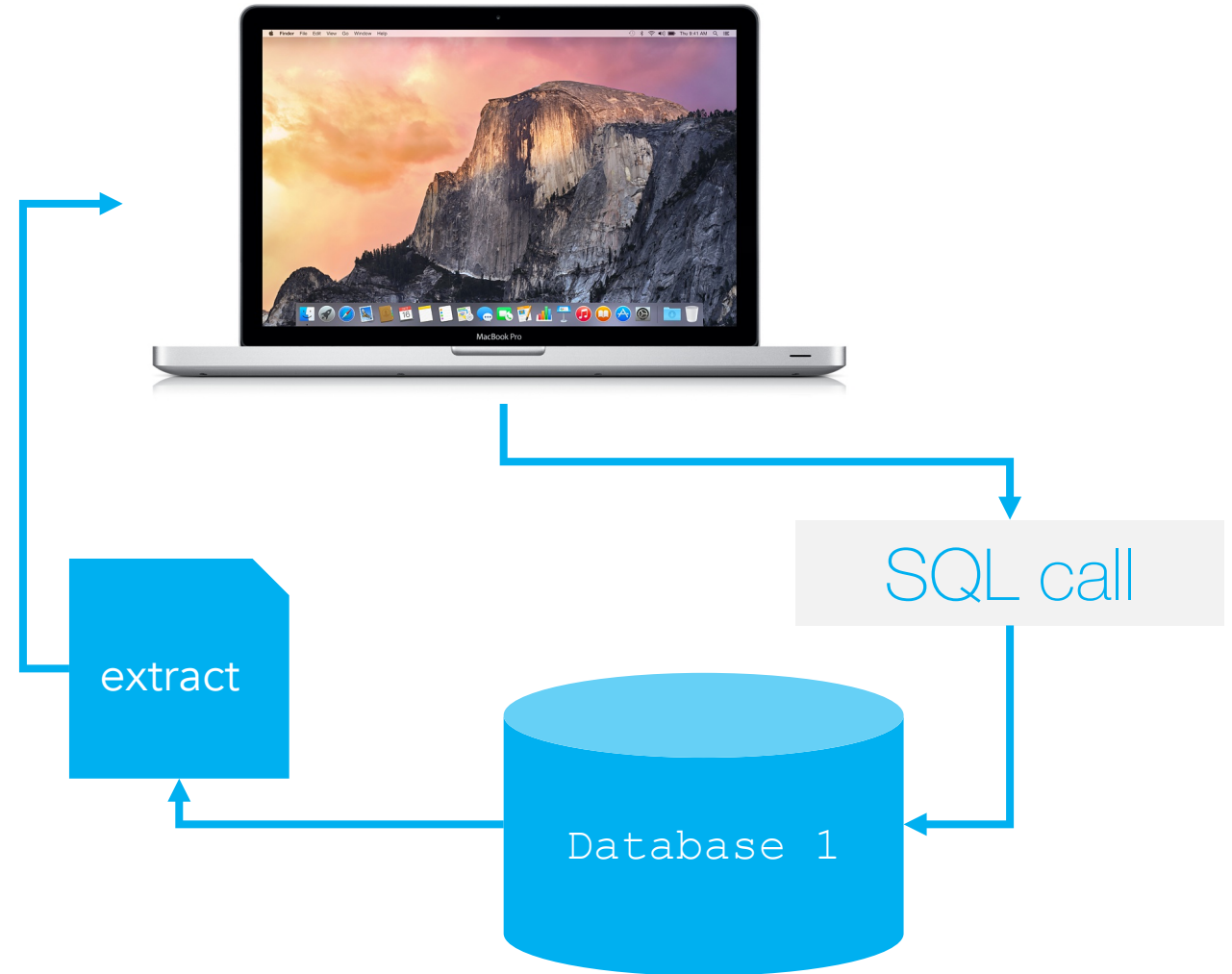


Option #3

Learn to query
databases using

**Structured Query
Language** or **SQL**

Structured Query
Language (See-Qwell
or S. Q. L.) is used to
interact with and
manage relational
databases.



What's a Relational Database

Relational databases organizes information into one or more tables such that each table is an entity type, a columns contain attributes that describe the entity and may be used to link tables

ssn	firstname	lastname
123-45-6789	john	Sobrenome
912-34-5678	Jorge	Apellido
891-23-4567	Gianni	Cognome

state	desc
1	New York
2	Virginia
3	Washington

id	ssn	state	income
1	123-45-6789	1	\$100,000
2	912-34-5678	1	\$30,000
3	891-23-4567	3	\$75,000

SQL CRUD Operations

CRUD Operations are four basic operations that SQL can help undertake in a database


- Create – **INSERT** statement
- Read – **SELECT** statement
- Update – **UPDATE** statement
- Delete – **DELETE** statement

SQL CRUD Operations

CRUD Operations are four basic operations that SQL can help undertake in a database

- Create – **INSERT** statement
- Read – **SELECT** statement
- Update – **UPDATE** statement
- Delete – **DELETE** statement

Data scientists and analysts largely work in this step but may occasionally work with all four



SQL A few examples

```
SELECT *  
FROM tbl
```

R equivalent
> tbl

“Retrieve all fields (*) and records FROM table named tbl”

SQL A few examples

```
SELECT income  
FROM tbl
```

```
R equivalent  
> tbl[, c("income")]
```

"Retrieve 'income' field and records FROM table named tbl"

SQL A few examples

```
SELECT income, state  
FROM tbl
```

```
R equivalent  
> tbl[, c("income", "state")]
```

"Retrieve 'income' and 'state' fields and records
FROM table named tbl"

SQL A few examples

```
SELECT income, state  
FROM tbl  
LIMIT 10
```

R equivalent
> tbl[1:10, c("income", "state")]

"Retrieve first 10 records of 'income' and 'state' fields
FROM table named tbl"

SQL A few examples

```
SELECT income, state  
FROM tbl  
WHERE region = 'Midwest'  
LIMIT 10
```

R equivalent

```
> a <- tbl[tbl$region == "Midwest",  
c("income", "state")  
> b <- a[1:10,]
```

"Retrieve first 10 records of 'income' and 'state' fields FROM table named tbl if region is in the 'Midwest' "

SQL

How to use

- Usually, SQL statements would be executed via command line or an Interactive Development Environment (IDE) to interact with databases
- In R, SQL can only be used via the library *sqldf*, which operates on data frames.
 - Good way to get accustomed

<Code Time/>

US' and UK's Export and Financial Sanction Lists