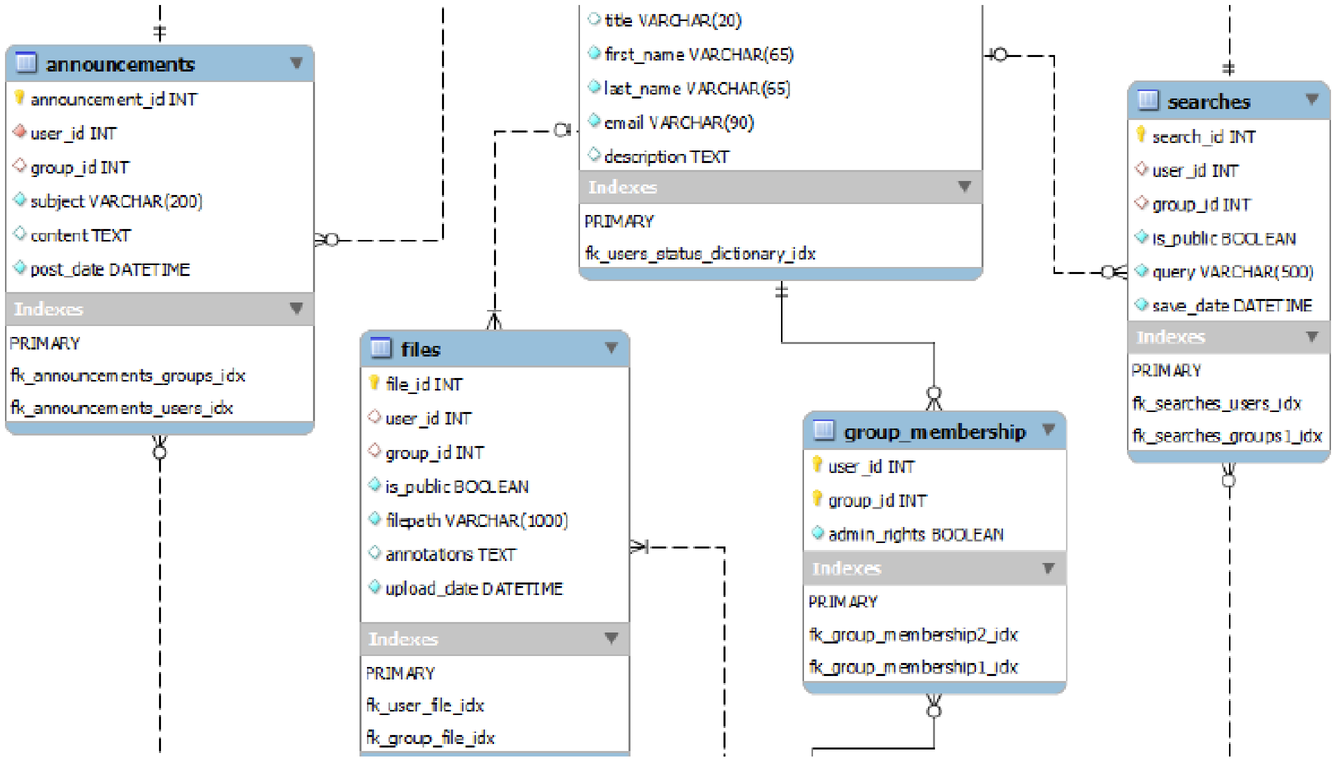
# Introduction to Database Design and Normalisation



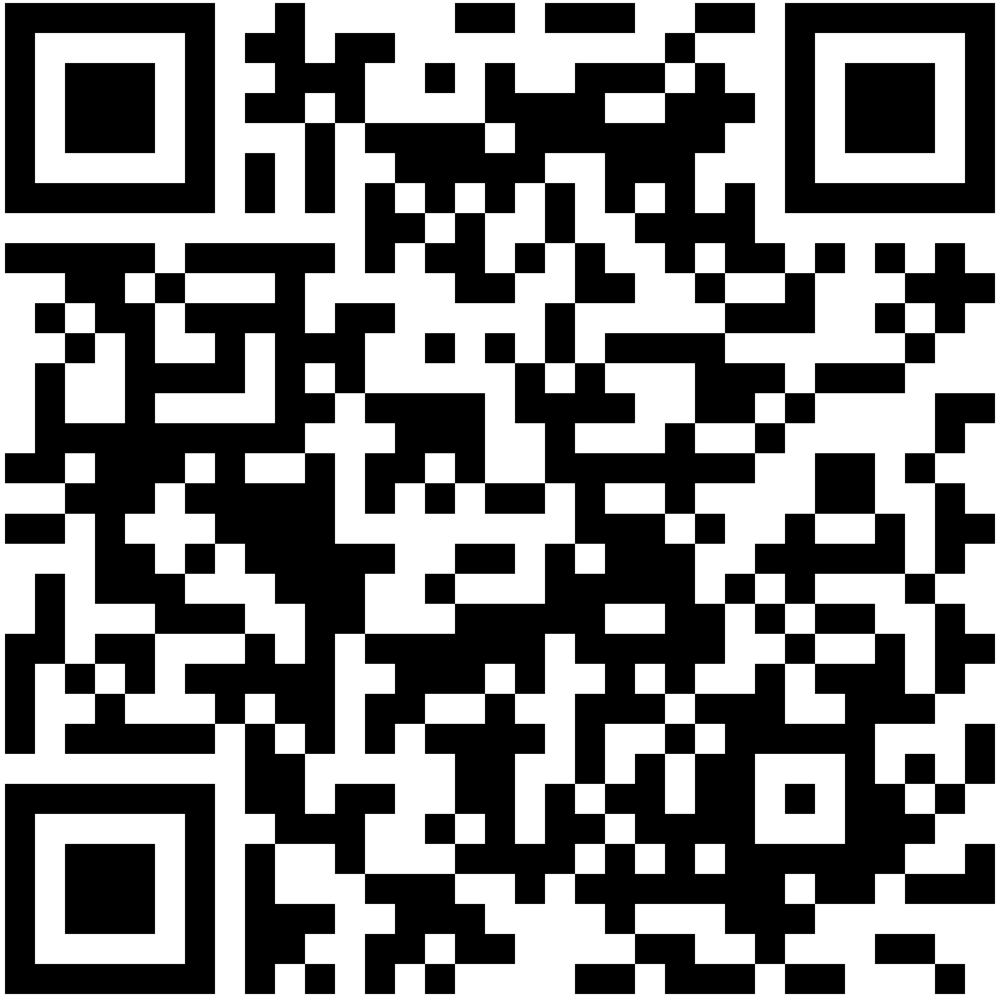
Dr Tomasz Kurowski

t.j.kurowski@cranfield.ac.uk

11 February 2025

www.cranfield.ac.uk

# Module engagement QR code

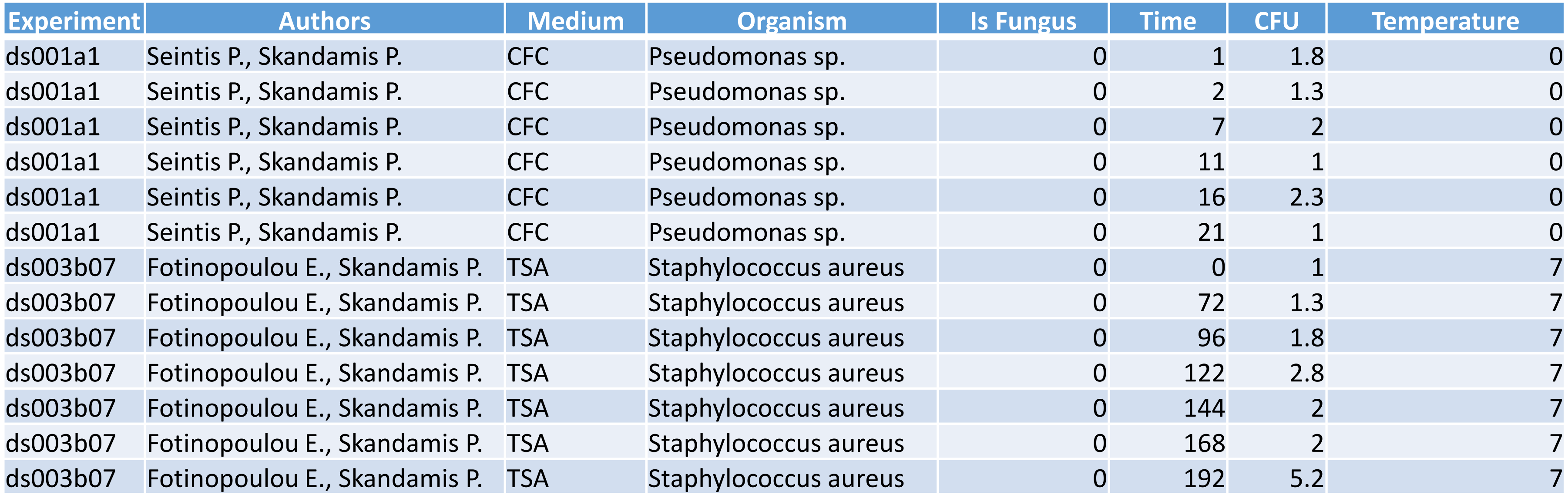


If you are unable to scan this code, please contact SAS Admin –

**seeaadmin@cranfield.ac.uk**

# Problems of data storage

Tabular data storage - spreadsheets, comma-separated value (CSV) files...



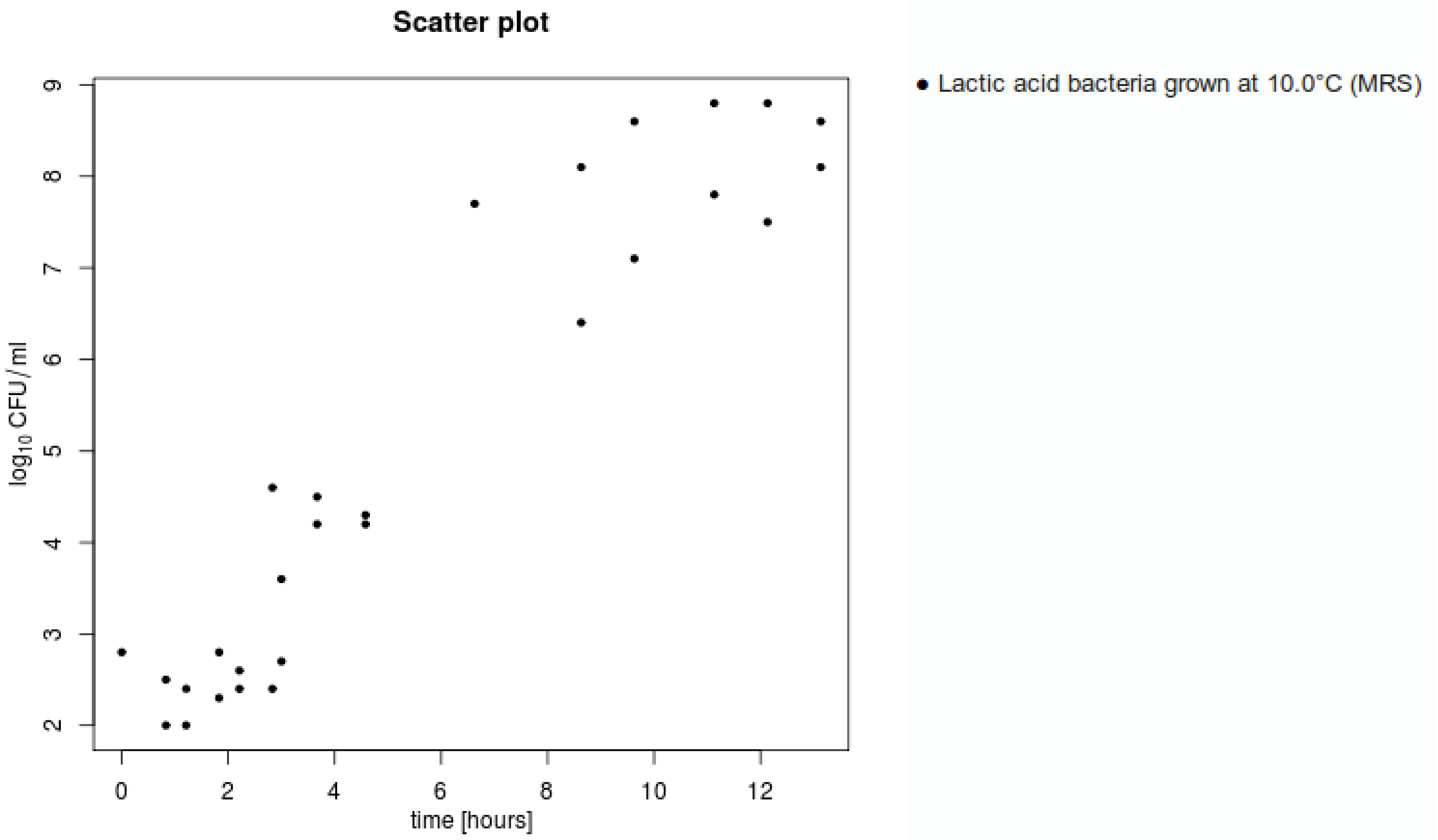
| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 1 | 1.8 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 2 | 1.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 7 | 2 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 11 | 1 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 16 | 2.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 21 | 1 | 0 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 0 | 1 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 72 | 1.3 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 96 | 1.8 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 122 | 2.8 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 144 | 2 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 168 | 2 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 192 | 5.2 | 7 |

Redundant data, no integrity protection, no associated query system

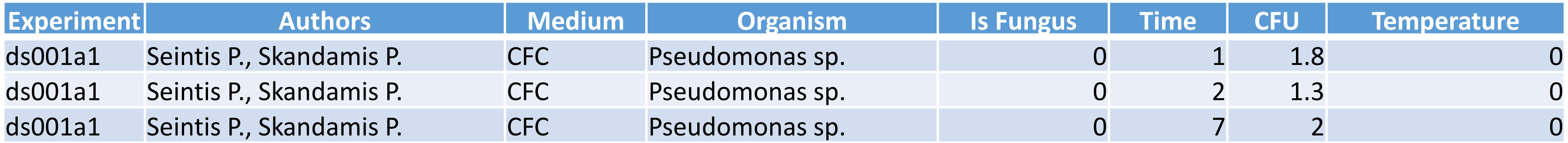
**Spreadsheet available on Canvas!**

# Example data

Time series of bacterial growth in specific conditions:



# Problems of data storage



| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 1 | 1.8 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 2 | 1.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 7 | 2 | 0 |

**Experiment** - alphanumeric experiment identifier

**Authors** - list of authors

**Medium** - name of medium used for microorganism growth

**Organism** - name of studied organism

**Is Fungus** - 1 if organism is a fungus, 0 if it is not

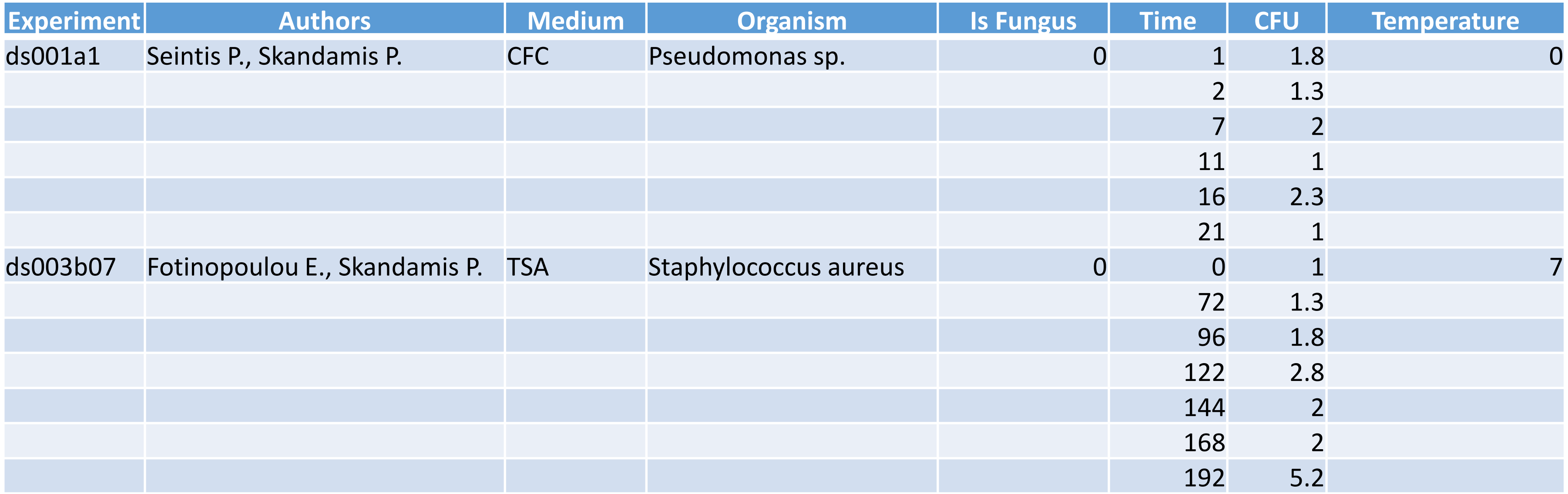
**Time** - time point of data collection [hours]

**CFU** - $\log_{10}$ of Colony Forming Unit concentration

**Temperature** - temperature during experiment [°C]

# Problems of data storage

You could make it prettier...



| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 1 | 1.8 | 0 |
| | | | | | 2 | 1.3 | |
| | | | | | 7 | 2 | |
| | | | | | 11 | 1 | |
| | | | | | 16 | 2.3 | |
| | | | | | 21 | 1 | |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 0 | 1 | 7 |
| | | | | | 72 | 1.3 | |
| | | | | | 96 | 1.8 | |
| | | | | | 122 | 2.8 | |
| | | | | | 144 | 2 | |
| | | | | | 168 | 2 | |
| | | | | | 192 | 5.2 | |

But does this make it better?

# Databases

What is a database for?

- Storage of data
- Organising data
- Providing a system of accessing data and interacting with it

↓

**CRUD** – **C**reating, **R**eading, **U**pdating, **D**eleting

# Databases

What is a database for?

- Storage of data
- Organising data
- Providing a system of accessing data and interacting with it

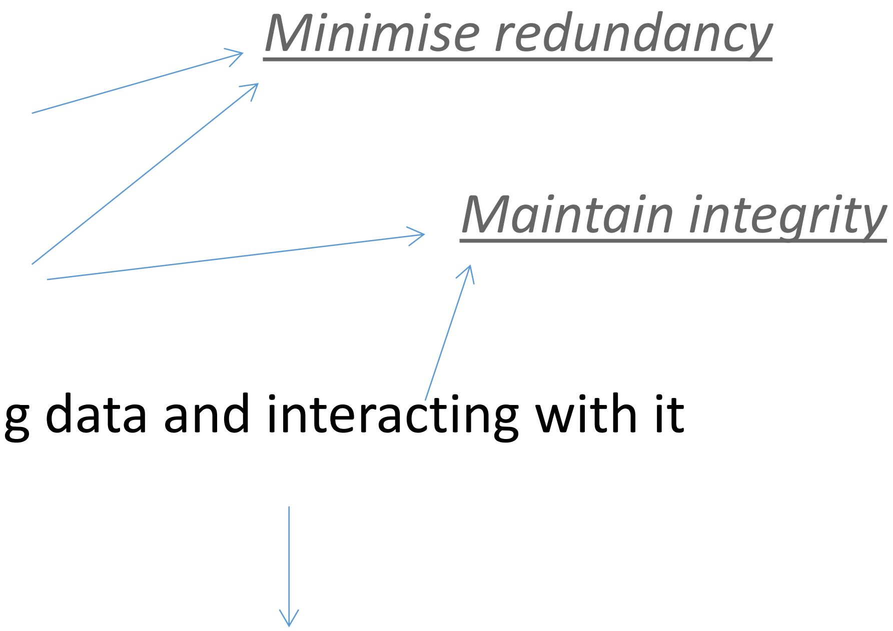
*Minimise redundancy*

**CRUD** – **C**reating, **R**eading, **U**pdating, **D**eleting

# Databases

What is a database for?

- Storage of data
- Organising data
- Providing a system of accessing data and interacting with it



*Minimise redundancy*

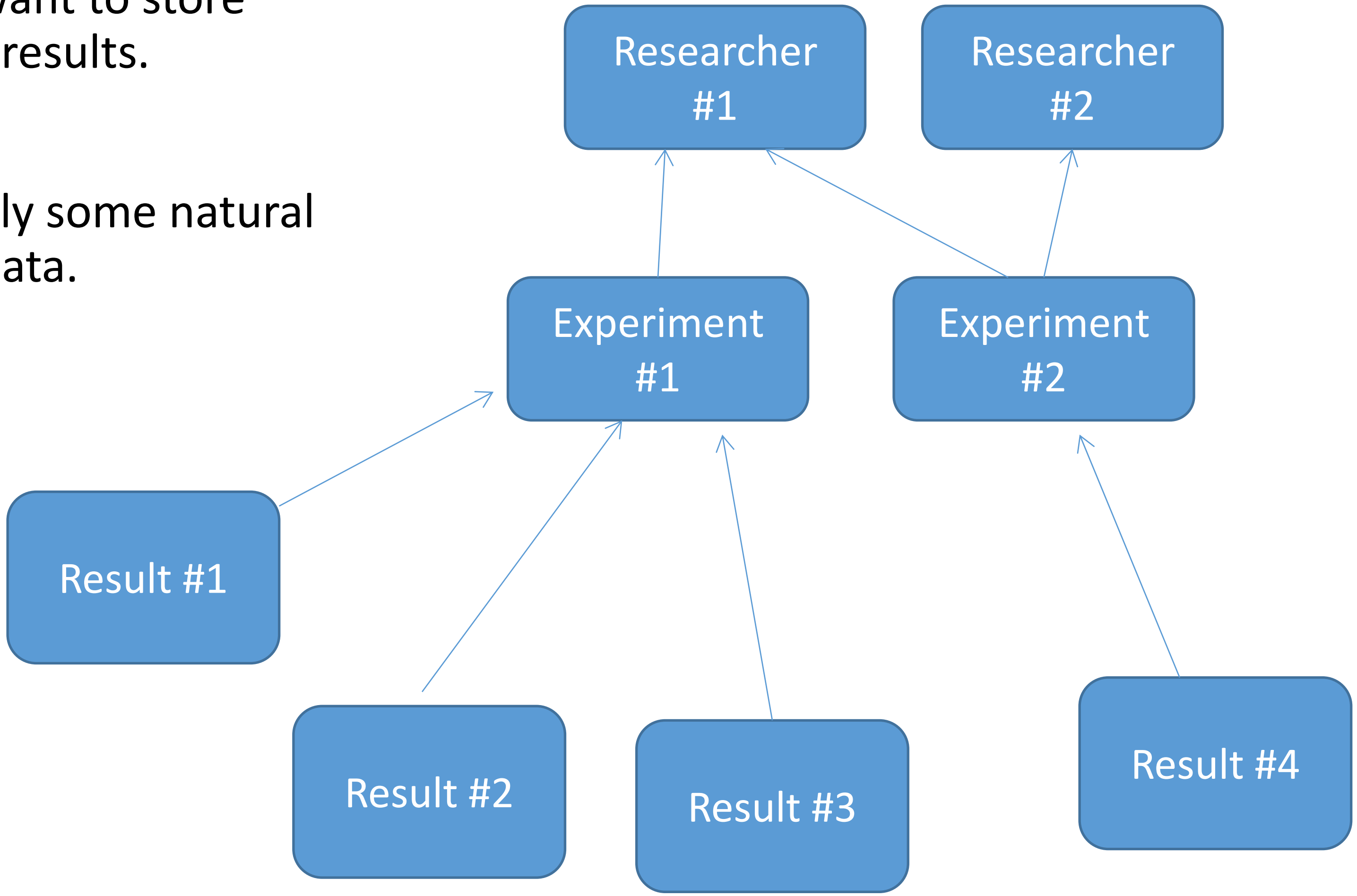
*Maintain integrity*

**CRUD** – **C**reating, **R**eading, **U**pdating, **D**eleting

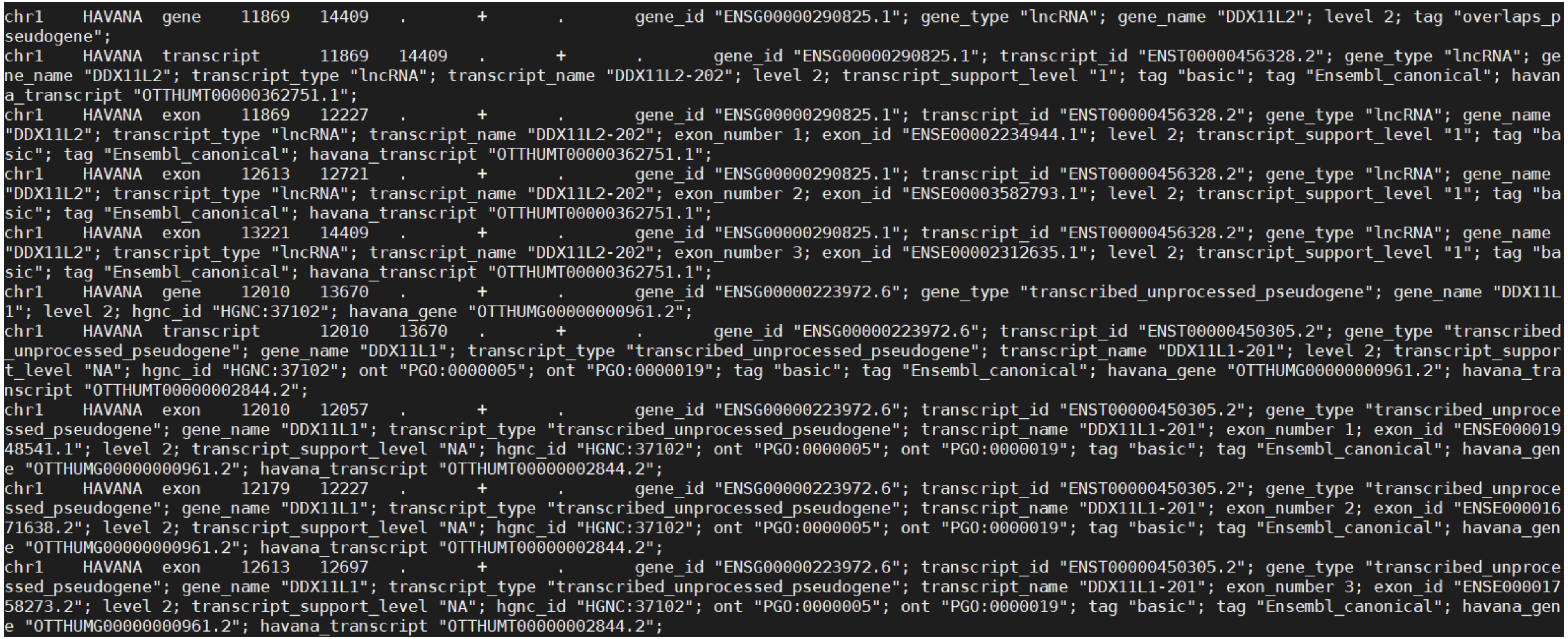
# Relationships between data

Let’s say we want to store experimental results.

There is usually some natural hierarchy to data.

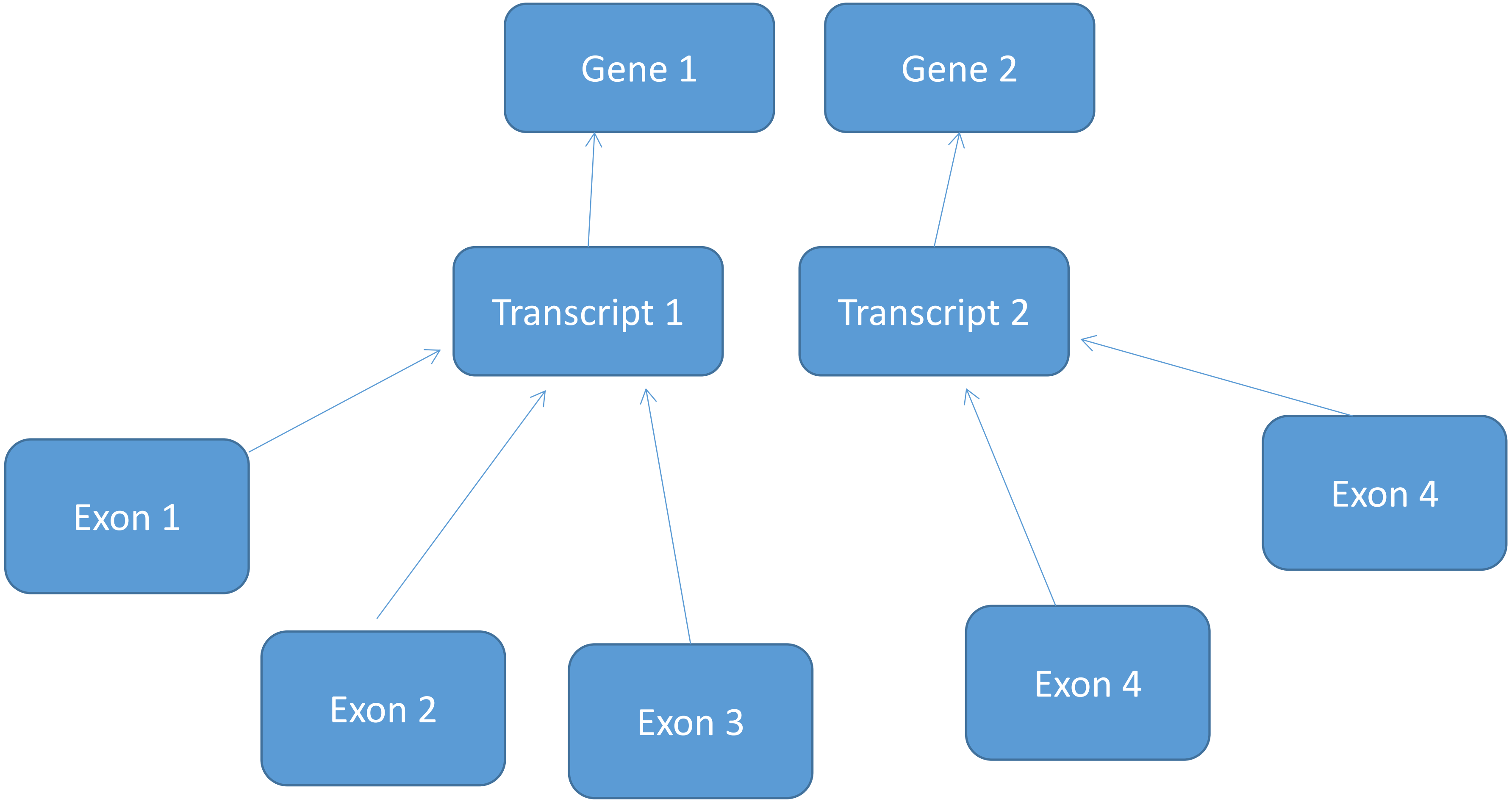


# In your Java assignment

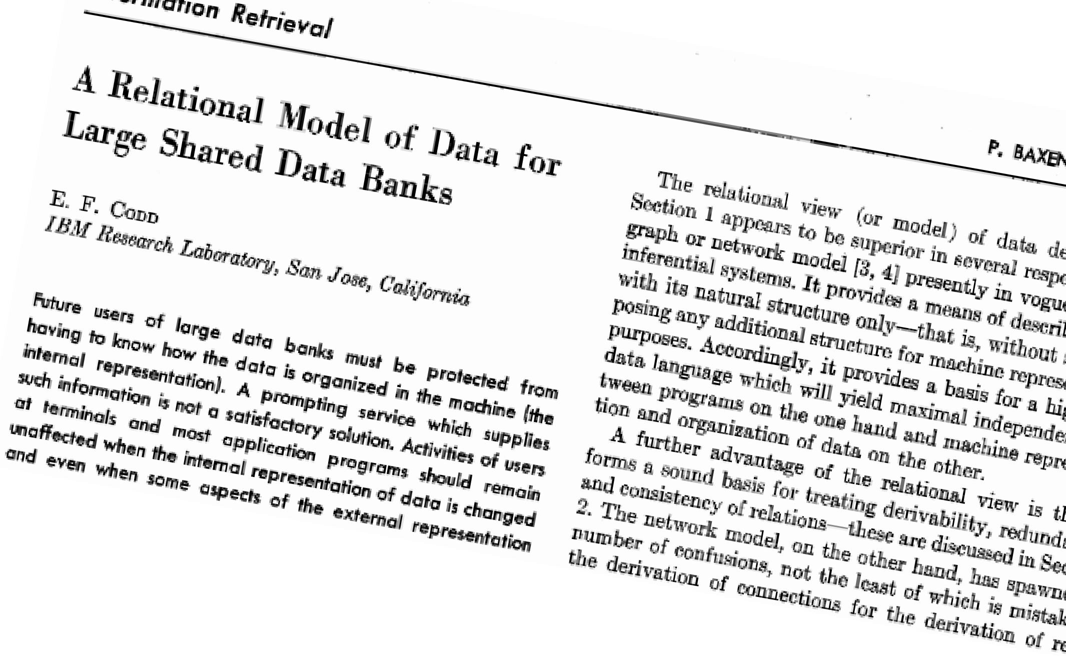


```
chr1    HAVANA  gene    11869   14409   .       +       .       gene_id "ENSG00000290825.1"; gene_type "lncRNA"; gene_name "DDX11L2"; level 2; tag "overlaps_p
seudogene";
chr1    HAVANA  transcript      11869   14409   .       +       .       gene_id "ENSG00000290825.1"; transcript_id "ENST00000456328.2"; gene_type "lncRNA"; ge
ne_name "DDX11L2"; transcript_type "lncRNA"; transcript_name "DDX11L2-202"; level 2; transcript_support_level "1"; tag "basic"; tag "Ensembl_canonical"; havan
a_transcript "OTTHUMT00000362751.1";
chr1    HAVANA  exon    11869   12227   .       +       .       gene_id "ENSG00000290825.1"; transcript_id "ENST00000456328.2"; gene_type "lncRNA"; gene_name
"DDX11L2"; transcript_type "lncRNA"; transcript_name "DDX11L2-202"; exon_number 1; exon_id "ENSE00002234944.1"; level 2; transcript_support_level "1"; tag "ba
sic"; tag "Ensembl_canonical"; havana_transcript "OTTHUMT00000362751.1";
chr1    HAVANA  exon    12613   12721   .       +       .       gene_id "ENSG00000290825.1"; transcript_id "ENST00000456328.2"; gene_type "lncRNA"; gene_name
"DDX11L2"; transcript_type "lncRNA"; transcript_name "DDX11L2-202"; exon_number 2; exon_id "ENSE00003582793.1"; level 2; transcript_support_level "1"; tag "ba
sic"; tag "Ensembl_canonical"; havana_transcript "OTTHUMT00000362751.1";
chr1    HAVANA  exon    13221   14409   .       +       .       gene_id "ENSG00000290825.1"; transcript_id "ENST00000456328.2"; gene_type "lncRNA"; gene_name
"DDX11L2"; transcript_type "lncRNA"; transcript_name "DDX11L2-202"; exon_number 3; exon_id "ENSE00002312635.1"; level 2; transcript_support_level "1"; tag "ba
sic"; tag "Ensembl_canonical"; havana_transcript "OTTHUMT00000362751.1";
chr1    HAVANA  gene    12010   13670   .       +       .       gene_id "ENSG00000223972.6"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L
1"; level 2; hgnc_id "HGNC:37102"; havana_gene "OTTHUMG00000000961.2";
chr1    HAVANA  transcript      12010   13670   .       +       .       gene_id "ENSG00000223972.6"; transcript_id "ENST00000450305.2"; gene_type "transcribed
_unprocessed_pseudogene"; gene_name "DDX11L1"; transcript_type "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-201"; level 2; transcript_suppor
t_level "NA"; hgnc_id "HGNC:37102"; ont "PGO:0000005"; ont "PGO:0000019"; tag "basic"; tag "Ensembl_canonical"; havana_gene "OTTHUMG00000000961.2"; havana_tra
nscript "OTTHUMT00000002844.2";
chr1    HAVANA  exon    12010   12057   .       +       .       gene_id "ENSG00000223972.6"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unproce
ssed_pseudogene"; gene_name "DDX11L1"; transcript_type "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-201"; exon_number 1; exon_id "ENSE000019
48541.1"; level 2; transcript_support_level "NA"; hgnc_id "HGNC:37102"; ont "PGO:0000005"; ont "PGO:0000019"; tag "basic"; tag "Ensembl_canonical"; havana_gen
e "OTTHUMG00000000961.2"; havana_transcript "OTTHUMT00000002844.2";
chr1    HAVANA  exon    12179   12227   .       +       .       gene_id "ENSG00000223972.6"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unproce
ssed_pseudogene"; gene_name "DDX11L1"; transcript_type "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-201"; exon_number 2; exon_id "ENSE000016
71638.2"; level 2; transcript_support_level "NA"; hgnc_id "HGNC:37102"; ont "PGO:0000005"; ont "PGO:0000019"; tag "basic"; tag "Ensembl_canonical"; havana_gen
e "OTTHUMG00000000961.2"; havana_transcript "OTTHUMT00000002844.2";
chr1    HAVANA  exon    12613   12697   .       +       .       gene_id "ENSG00000223972.6"; transcript_id "ENST00000450305.2"; gene_type "transcribed_unproce
ssed_pseudogene"; gene_name "DDX11L1"; transcript_type "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-201"; exon_number 3; exon_id "ENSE000017
58273.2"; level 2; transcript_support_level "NA"; hgnc_id "HGNC:37102"; ont "PGO:0000005"; ont "PGO:0000019"; tag "basic"; tag "Ensembl_canonical"; havana_gen
e "OTTHUMG00000000961.2"; havana_transcript "OTTHUMT00000002844.2";
```

# In your Java assignment



# Relational model



ormation Retrieval

P. BAXEN

## A Relational Model of Data for Large Shared Data Banks

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

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation

The relational view (or model) of data de
Section 1 appears to be superior in several respe
graph or network model [3, 4] presently in vogue
inferential systems. It provides a means of describ
with its natural structure only—that is, without
posing any additional structure for machine repres
purposes. Accordingly, it provides a basis for a hi
data language which will yield maximal independe
tween programs on the one hand and machine repre
tion and organization of data on the other.

A further advantage of the relational view is t
forms a sound basis for treating derivability, redunda
and consistency of relations—these are discussed in Sec
2. The network model, on the other hand, has spawn
number of confusions, not the least of which is mistak
the derivation of connections for the derivation of re

General, theoretical model for organising databases.

Introduced in the 1970s by E.F. Codd.

Currently the most popular approach, relational databases are a $100 billion industry.

Data arranged in a collection of tables (n-ary relations subsets of the Cartesian product of n domains).

Separates the concepts of “data” and “schema” (structure of data)

# Relational model

Data is stored in multiple named **tables** (**relations**).

| Experiments | | | |
|---|---|---|---|
| Experiment_ID | Medium | Organism | Temperature |
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# Relational model

Each table has a list of named **columns** (**attributes**).

Each column has a **type** (**domain**), e.g. INT for integer or CHAR for character strings.

| Experiments | | | |
|---|---|---|---|
| Experiment_ID | Medium | Organism | Temperature |
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# Relational model

Each table **row** (**record, tuple**) represents an entry and has to be unique.

| Experiments | | | |
|---|---|---|---|
| **Experiment_ID** | **Medium** | **Organism** | **Temperature** |
| **ds001a1** | **CFC** | **Pseudomonas sp.** | **0** |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# Primary Keys

In order to be unique, each table needs to have at least one unique column or combination of columns. One of them is the **Primary Key** (**PK**), which serves as an identifier of that row.

There may be multiple **Candidate Keys**, but only one **Primary Key**.

| Experiments | | | |
|---|---|---|---|
| **Experiment_ID** | **Medium** | **Organism** | **Temperature** |
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# Primary Keys

In order to be unique, each table needs to have at least one unique column or combination of columns. One of them is the **Primary Key** (**PK**), which serves as an identifier of that row.

There may be multiple **Candidate Keys**, but only one **Primary Key**.

| Experiments | | | |
|---|---|---|---|
| **Experiment_ID** | **Medium** | **Organism** | **Temperature** |
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# But what about this?!

What are the candidate keys? What should be the Primary Key?

| Employees | |
|---|---|
| **First_Name** | **Last_Name** |
| Tomasz | Kurowski |
| Fady | Mohareb |
| James | Smith |
| … | … |

# Compound Key

A combination of attributes which are not themselves unique can be a Candidate/Primary Key.

The key is (First_Name, Last_Name)

| Employees | |
|---|---|
| **First_Name** | **Last_Name** |
| Tomasz | Kurowski |
| Fady | Mohareb |
| James | Smith |
| ... | ... |

But is that enough?

# Surrogate Key

We could simply add a unique identifier ourselves – typically an integer. This is called a **Surrogate Key**.

Non-Surrogate Keys are sometimes called **Natural Keys**.

| Employees | | |
|---|---|---|
| Employee_id | **First_Name** | **Last_Name** |
| 1 | Tomasz | Kurowski |
| 2 | Fady | Mohareb |
| 3 | James | Smith |
| ... | ... | ... |

Some people add surrogate keys to all tables, avoiding natural keys.

Others claim using natural keys is superior.

It depends...

# Foreign Keys

Tables may also contain Foreign Keys. These refer to a Primary Key of a different table. This creates a relationship between two tables!

Child table

**Measurements**

| Time | CFU | Experiment_ID |
|---|---|---|
| 1 | 1.8 | ds001a1 |
| 2 | 1.3 | ds001a1 |
| 7 | 2 | ds001a1 |
| 11 | 1 | ds001a1 |
| 16 | 2.3 | ds001a1 |
| 1 | 2.3 | ds001b1 |
| 7 | 1.6 | ds001b1 |
| 11 | 2.4 | ds001b1 |
| 16 | 2.5 | ds001b1 |
| 21 | 3.5 | ds001b1 |
| 28 | 5.1 | ds001b1 |
| 32 | 4.3 | ds001b1 |

Parent table

**Experiments**

| Experiment_ID | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

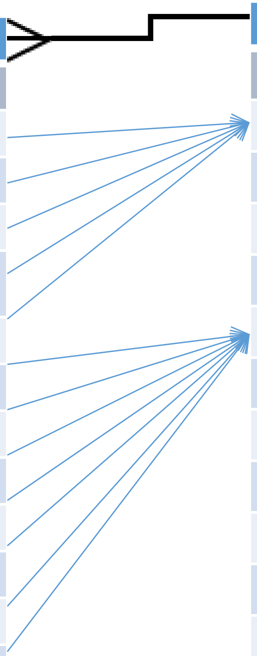
# Cardinality

**Experiments** has a one-to-many relationship with **Measurements**

Usually symbolised by a „crow's foot"

Child table

| Measurements | | |
|---|---|---|
| **Time** | **CFU** | **Experiment_ID** |
| 1 | 1.8 | ds001a1 |
| 2 | 1.3 | ds001a1 |
| 7 | 2 | ds001a1 |
| 11 | 1 | ds001a1 |
| 16 | 2.3 | ds001a1 |
| 1 | 2.3 | ds001b1 |
| 7 | 1.6 | ds001b1 |
| 11 | 2.4 | ds001b1 |
| 16 | 2.5 | ds001b1 |
| 21 | 3.5 | ds001b1 |
| 28 | 5.1 | ds001b1 |
| 32 | 4.3 | ds001b1 |



Parent table

| Experiments | | | |
|---|---|---|---|
| **Experiment_ID** | **Medium** | **Organism** | **Temperature** |
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# Cardinality

**One-to-many** relationships are the most common type in relational databases.

**One-to-one** relationships are also common, but usually such tables can be merged and remain seperate for convenience or performance.

**Many-to-many** relationships are **NOT ALLOWED**!

# Database normalisation

How to make a „good” database?

**Normal forms** - list of conditions which a correct relational database should fulfil

You can treat it like a check-list:

1. Put all data in table
2. Check if it fulfils first set of conditions (is in normal form 1)
   a. If no, modify or split the table to make it fit and check again
   b. If yes, move on to step 3
3. Check if it fulfils second set of conditions (is in normal form 2)
   a. If no, modify or split the table to make it fit and check again
   b. If yes, move on to step 4
4. Check if It fulfils third set of conditions (is in normal form 3)…

When you are more experienced you can design the whole database first, and THEN do the check-list to see if you have missed anything.

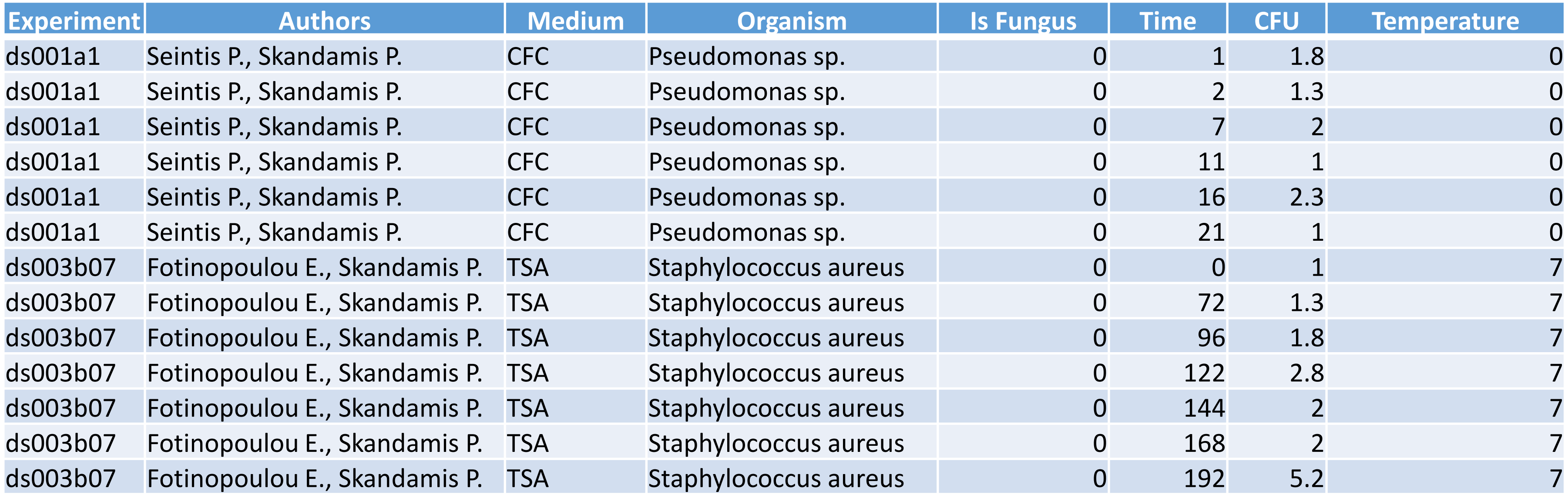
# Database normalisation

1. First Normal Form (1NF)
2. Second Normal Form (2NF)
3. Third Normal Form (3NF)
4. Boyce-Codd Normalisation
5. Fourth Normal Form (4NF)
6. Domain-key Normal Form (5NF)

**We will stop at 3NF!**

# Data normalisation

The first step is putting everything in a single table – we already have that!



| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 1 | 1.8 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 2 | 1.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 7 | 2 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 11 | 1 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 16 | 2.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 21 | 1 | 0 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 0 | 1 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 72 | 1.3 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 96 | 1.8 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 122 | 2.8 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 144 | 2 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 168 | 2 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 192 | 5.2 | 7 |

Time to normalise this!

**Download the spreadsheet from Canvas**

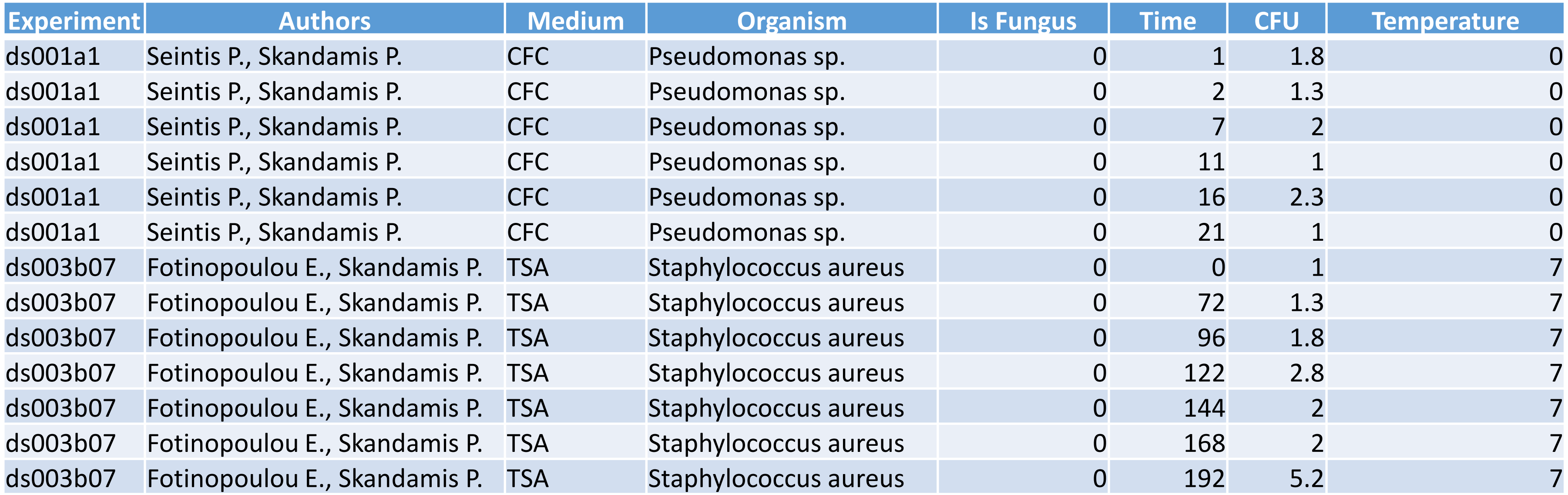
# First normal form (1NF)

Criteria:

1. Each table must have a primary key (rows should not repeat)
2. Values in the table should be atomic
3. There should be no repeating groups

# Criterion 1 – Primary Key

What are our **Candidate Keys**?



| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 1 | 1.8 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 2 | 1.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 7 | 2 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 11 | 1 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 16 | 2.3 | 0 |
| ds001a1 | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | 21 | 1 | 0 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 0 | 1 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 72 | 1.3 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 96 | 1.8 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 122 | 2.8 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 144 | 2 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 168 | 2 | 7 |
| ds003b07 | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | 192 | 5.2 | 7 |

Candidate Keys are unique columns or combinations of columns

# Criterion 1 – Primary Key

The combination of Experiment and Time is a good candidate key

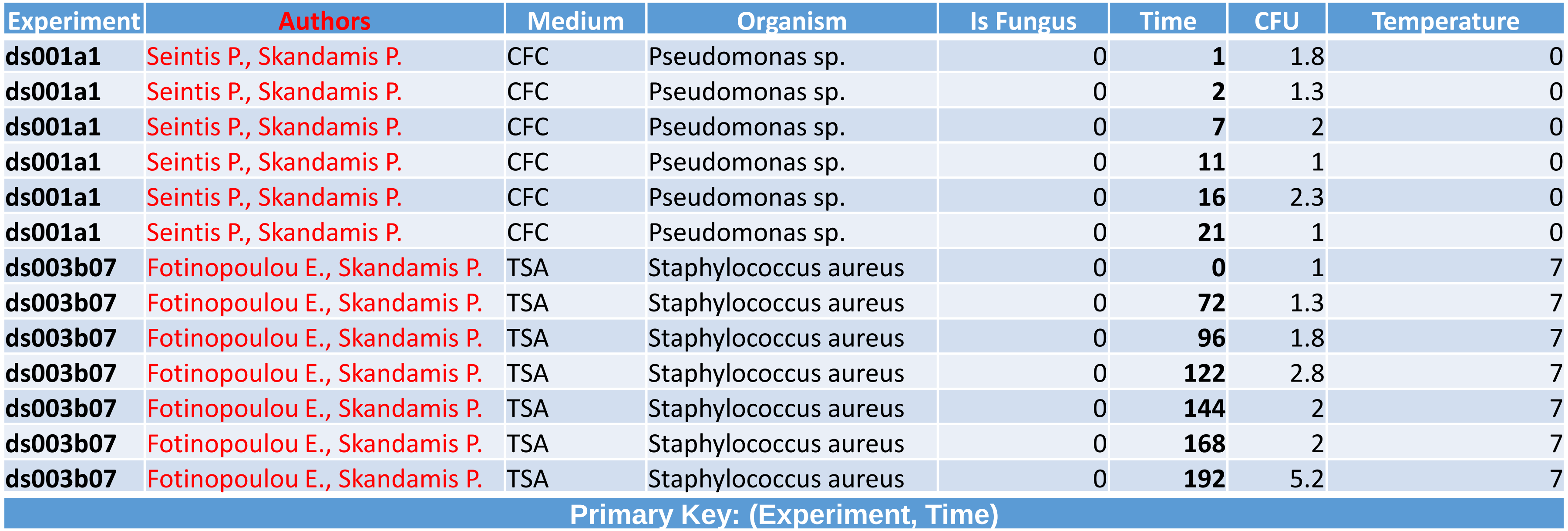
| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **16** | 2.3 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **21** | 1 | 0 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **0** | 1 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **72** | 1.3 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **96** | 1.8 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **122** | 2.8 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **144** | 2 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **168** | 2 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **192** | 5.2 | 7 |

**Primary Key: (Experiment, Time)**

We could also simply add a surrogate key. How would you name it?

# Criterion 2 – Atomic?

A table cell should only include a single value of a given type.



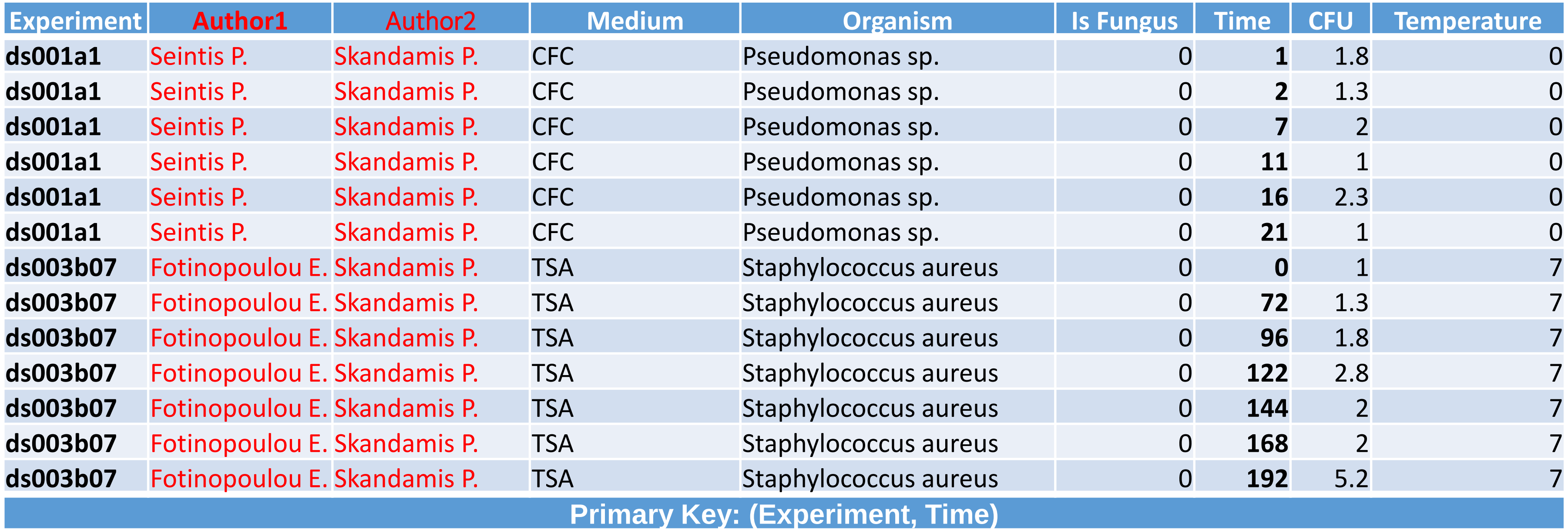
| Experiment | Authors | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **16** | 2.3 | 0 |
| **ds001a1** | Seintis P., Skandamis P. | CFC | Pseudomonas sp. | 0 | **21** | 1 | 0 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **0** | 1 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **72** | 1.3 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **96** | 1.8 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **122** | 2.8 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **144** | 2 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **168** | 2 | 7 |
| **ds003b07** | Fotinopoulou E., Skandamis P. | TSA | Staphylococcus aureus | 0 | **192** | 5.2 | 7 |

**Primary Key: (Experiment, Time)**

How do we split this?

# Criterion 2 – Atomic?

## Add an extra column?



| Experiment | Author1 | Author2 | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|---|
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **16** | 2.3 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **21** | 1 | 0 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **0** | 1 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **72** | 1.3 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **96** | 1.8 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **122** | 2.8 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **144** | 2 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **168** | 2 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **192** | 5.2 | 7 |

**Primary Key: (Experiment, Time)**

# Criterion 2 – Atomic?

We don’t know if there are always going to be two authors.

| Experiment | Author1 | Author2 | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|---|
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **16** | 2.3 | 0 |
| **ds001a1** | Seintis P. | Skandamis P. | CFC | Pseudomonas sp. | 0 | **21** | 1 | 0 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **0** | 1 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **72** | 1.3 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **96** | 1.8 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **122** | 2.8 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **144** | 2 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **168** | 2 | 7 |
| **ds003b07** | Fotinopoulou E. | Skandamis P. | TSA | Staphylococcus aureus | 0 | **192** | 5.2 | 7 |

**Primary Key: (Experiment, Time)**

...and these are the „repeating groups" from criterion 3

# Criterion 2 – Atomic?

Duplicate rows for each author?

| Experiment | Author | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| **ds001a1** | Seintis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Skandamis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Seintis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | Skandamis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **...** | ... | ... | ... | ... | **...** | ... | ... |

**Primary Key: (Experiment, Time)**

# Criterion 2 – Atomic?

This could technically work as 1NF.

| Experiment | Author | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| **ds001a1** | Seintis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Skandamis P. | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | Seintis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | Skandamis P. | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **...** | ... | ... | ... | ... | **...** | ... | ... |

**Primary Key: (Experiment, Time)**

But it introduces even more redundancy.

# Criterion 2 – Atomic?

Probably best solution: split the table!

(for now, let's keep the entire primary key)

**Authorships**

| Name | Measurement ID | |
|---|---|---|
| **Seintis P.** | **1** | |
| **Skandamis P.** | **1** | |
| **Seintis P.** | **2** | |
| **Skandamis P.** | | |
| **...** | | |

**Primary Key: (Name, Experiment, Time)**
**Foreign Key: (Experiment, Time)**

**Measurements**

| Experiment | Measurement ID | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|---|
| **ds001a1** | **1** | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | **2** | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | **3** | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | **4** | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **...** | | ... | ... | ... | **...** | ... | ... |

**Primary Key: (Measument ID)**

So, are we in 1NF?

# Second normal form (2NF)

Criteria:

1. Must be in 1NF
2. Non-prime attributes must depend on entire Primary Key, not only part of it

(non-prime attribute – attribute which is not part of any candidate key in the referenced table)

# Criterion 2

What do Medium, Organism, Is_Fungus, CFU, and Temperature depend on?

**Authors**

| Name | Experiment | Time |
|---|---|---|
| **Seintis P.** | **ds001a1** | **1** |
| **Skandamis P.** | **ds001a1** | **1** |
| **Seintis P.** | **ds001a1** | **2** |
| **Skandamis P.** | **ds001a1** | **2** |
| **...** | **...** | **...** |

**Primary Key: (Name, Experiment, Time)**
**Foreign Key: (Experiment, Time)**

**Experiment Data**

| Experiment | Medium | Organism | Is Fungus | Time | CFU | Temperature |
|---|---|---|---|---|---|---|
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **...** | ... | ... | ... | **...** | ... | ... |

**Primary Key: (Experiment, Time)**

# Criterion 2

Medium, Organism, Is_Fungus, and Temperature do not change over Time. They only depend on the Experiment.

CFU depends both on the Experiment and Time, the entire Primary Key

| Authors | | |
|---|---|---|
| Name | Experiment | Time |
| **Seintis P.** | **ds001a1** | **1** |
| **Skandamis P.** | **ds001a1** | **1** |
| **Seintis P.** | **ds001a1** | **2** |
| **Skandamis P.** | **ds001a1** | **2** |
| **...** | **...** | **...** |
| Primary Key: (Name, Experiment, Time)<br>Foreign Key: (Experiment, Time) | | |

| Experiment Data | | | | | | |
|---|---|---|---|---|---|---|
| Experiment | Medium | Organism | Is Fungus | Time | CFU | Temperature |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **1** | 1.8 | 0 |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **2** | 1.3 | 0 |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **7** | 2 | 0 |
| **ds001a1** | CFC | Pseudomonas sp. | 0 | **11** | 1 | 0 |
| **...** | ... | ... | ... | **...** | ... | ... |
| Primary Key: (Experiment, Time) | | | | | | |

# Criterion 2

## We split the Experiment_Data table!

**Authors**

| Name | Experiment | Time |
|---|---|---|
| Seintis P. | ds001a1 | 1 |
| Skandamis P. | ds001a1 | 1 |
| Seintis P. | ds001a1 | 2 |
| Skandamis P. | ds001a1 | 2 |
| ... | ... | ... |

**Primary Key: (Name, Experiment, Time)**
**Foreign Key: (Experiment, Time)**

**Experiments**

| Experiment | Medium | Organism | Is Fungus | Temperature |
|---|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 0 | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 0 | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 0 | 15 |
| ... | ... | ... | ... | ... |

**Primary Key: (Experiment)**

**Datapoints**

| Experiment | Time | CFU |
|---|---|---|
| ds001a1 | 1 | 1.8 |
| ds001a1 | 2 | 1.3 |
| ds001a1 | 7 | 2 |
| ds001a1 | 11 | 1 |
| ... | ... | ... |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

What about the Authors table?

The authors of a given experiment do not depend on time.

# Criterion 2

## That’s better.

**Authors**

| Name | Experiment |
|---|---|
| **Seintis P.** | **ds001a1** |
| **Skandamis P.** | **ds001a1** |
| **Seintis P.** | **ds001a2** |
| **Skandamis P.** | **ds001a2** |
| **...** | **...** |

**Primary Key: (Name, Experiment)**
**Foreign Key: (Experiment)**

**Experiments**

| Experiment | Medium | Organism | Is Fungus | Temperature |
|---|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 0 | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 0 | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 0 | 15 |
| ... | ... | ... | ... | ... |

**Primary Key: (Experiment)**

**Datapoints**

| Experiment | Time | CFU |
|---|---|---|
| **ds001a1** | **1** | 1.8 |
| **ds001a1** | **2** | 1.3 |
| **ds001a1** | **7** | 2 |
| **ds001a1** | **11** | 1 |
| **...** | **...** | ... |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

Are we in 2NF?

# Third normal form (3NF)

Criteria:

1. Must be in 2NF
2. Non-prime attributes must not depend on other non-prime attributes

# Criterion 2

Does any non-prime attribute depends on another?

**Authors**

| Name | Experiment |
|---|---|
| **Seintis P.** | **ds001a1** |
| **Skandamis P.** | **ds001a1** |
| **Seintis P.** | **ds001a2** |
| **Skandamis P.** | **ds001a2** |
| **...** | **...** |

**Primary Key: (Name, Experiment)**
**Foreign Key: (Experiment)**

**Experiments**

| Experiment | Medium | Organism | Is Fungus | Temperature |
|---|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 0 | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 0 | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 0 | 15 |
| ... | ... | ... | ... | ... |

**Primary Key: (Experiment)**

**Datapoints**

| Experiment | Time | CFU |
|---|---:|---:|
| **ds001a1** | **1** | 1.8 |
| **ds001a1** | **2** | 1.3 |
| **ds001a1** | **7** | 2 |
| **ds001a1** | **11** | 1 |
| **...** | **...** | ... |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

# Criterion 2

## Is_Fungus depends on Organism!

**Authors**

| Name | Experiment |
|---|---|
| Seintis P. | ds001a1 |
| Skandamis P. | ds001a1 |
| Seintis P. | ds001a2 |
| Skandamis P. | ds001a2 |
| ... | ... |

**Primary Key: (Name, Experiment)**
**Foreign Key: (Experiment)**

**Experiments**

| Experiment | Medium | Organism | Is Fungus | Temperature |
|---|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 0 | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 0 | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 0 | 15 |
| ... | ... | ... | ... | ... |

**Primary Key: (Experiment)**

What shall we do?

**Datapoints**

| Experiment | Time | CFU |
|---|---|---|
| ds001a1 | 1 | 1.8 |
| ds001a1 | 2 | 1.3 |
| ds001a1 | 7 | 2 |
| ds001a1 | 11 | 1 |
| ... | ... | ... |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

# Criterion 2

## Looks like we are in 3NF!

**Authors**

| Name | Experiment |
|---|---|
| **Seintis P.** | **ds001a1** |
| **Skandamis P.** | **ds001a1** |
| **Seintis P.** | **ds001a1** |
| **Skandamis P.** | **ds001a1** |
| **...** | **...** |

**Primary Key: (Name, Experiment)**
**Foreign Key: (Experiment)**

**Experiments**

| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ... | ... | ... | ... |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

**Organisms**

| Organism | Is Fungus |
|---|---|
| Pseudomonas sp. | 0 |
| Lactic acid bacteria | 0 |
| Enterobacteriaceae | 0 |
| Yeasts-moulds | 1 |
| ... | ... |

**Primary Key: (Organism)**

**Datapoints**

| Experiment | Time | CFU |
|---|---|---|
| **ds001a1** | **1** | 1.8 |
| **ds001a1** | **2** | 1.3 |
| **ds001a1** | **7** | 2 |
| **ds001a1** | **11** | 1 |
| **...** | **...** | ... |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

# Normalisation summed up

Data depends on:

**1NF**: *The key,*

**2NF**: *the whole key,*

**3NF:** *and nothing but the key.*

So, are we done yet?

# A look at our relationships

What type of relationship do Experiments and Datapoints have?

**Experiments**

| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| … | … | … | … |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

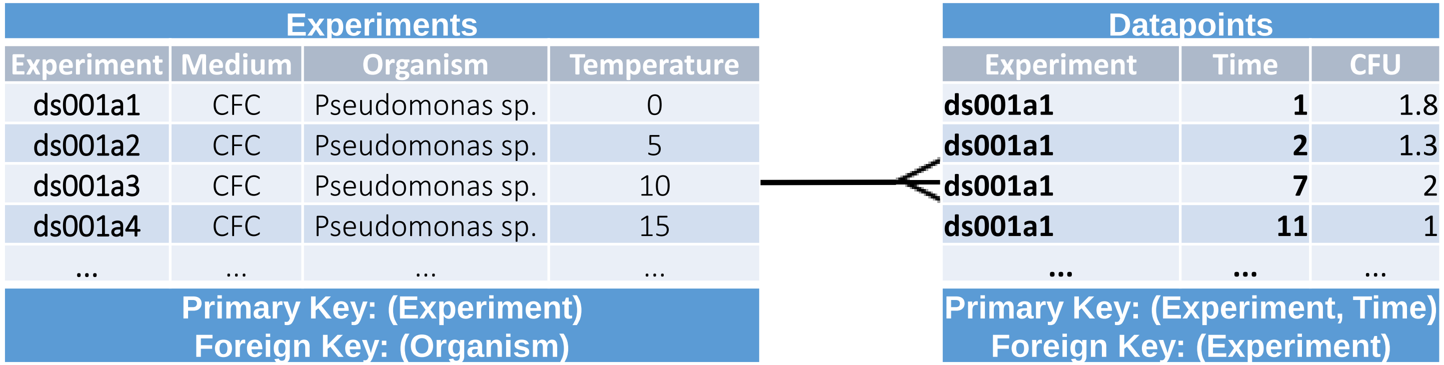
**Datapoints**

| Experiment | Time | CFU |
|---|---|---|
| ds001a1 | 1 | 1.8 |
| ds001a1 | 2 | 1.3 |
| ds001a1 | 7 | 2 |
| ds001a1 | 11 | 1 |
| … | … | … |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

# A look at our relationships

Answer: One-to-Many



Experiments

| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| … | … | … | … |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

Datapoints

| Experiment | Time | CFU |
|---|---|---|
| **ds001a1** | **1** | 1.8 |
| **ds001a1** | **2** | 1.3 |
| **ds001a1** | **7** | 2 |
| **ds001a1** | **11** | 1 |
| **…** | **…** | … |

**Primary Key: (Experiment, Time)**
**Foreign Key: (Experiment)**

# A look at our relationships

What is the relationship between Organisms and Experiments?

**Organisms**

| Organism | Is Fungus |
|---|---|
| Pseudomonas sp. | 0 |
| Lactic acid bacteria | 0 |
| Enterobacteriaceae | 0 |
| Yeasts-moulds | 1 |
| ... | ... |

**Primary Key: (Organism)**

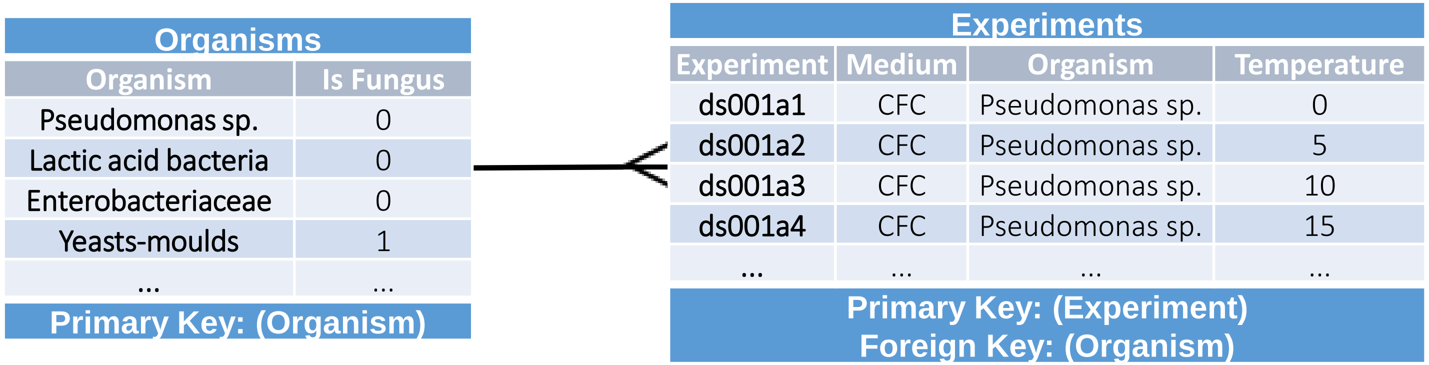
**Experiments**

| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ... | ... | ... | ... |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

# A look at our relationships

Answer: One-to-Many



**Organisms**

| Organism | Is Fungus |
|---|---|
| Pseudomonas sp. | 0 |
| Lactic acid bacteria | 0 |
| Enterobacteriaceae | 0 |
| Yeasts-moulds | 1 |
| ... | ... |

**Primary Key: (Organism)**

**Experiments**

| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ... | ... | ... | ... |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

# A look at our relationships

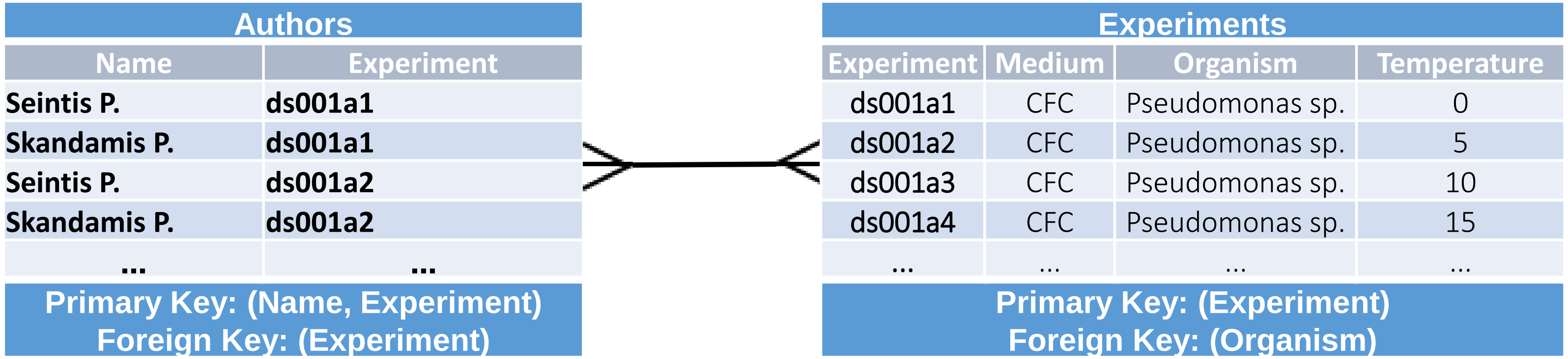
What about Authors and Experiments?

| Authors | |
|---|---|
| Name | Experiment |
| **Seintis P.** | **ds001a1** |
| **Skandamis P.** | **ds001a1** |
| **Seintis P.** | **ds001a2** |
| **Skandamis P.** | **ds001a2** |
| **...** | **...** |
| **Primary Key: (Name, Experiment)**<br>**Foreign Key: (Experiment)** | |

| Experiments | | | |
|---|---|---|---|
| Experiment | Medium | Organism | Temperature |
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ... | ... | ... | ... |
| **Primary Key: (Experiment)**<br>**Foreign Key: (Organism)** | | | |

# A look at our relationships

Many-to-Many!



**Authors**

| Name | Experiment |
|---|---|
| **Seintis P.** | **ds001a1** |
| **Skandamis P.** | **ds001a1** |
| **Seintis P.** | **ds001a2** |
| **Skandamis P.** | **ds001a2** |
| **...** | **...** |

**Primary Key: (Name, Experiment)**
**Foreign Key: (Experiment)**

**Experiments**

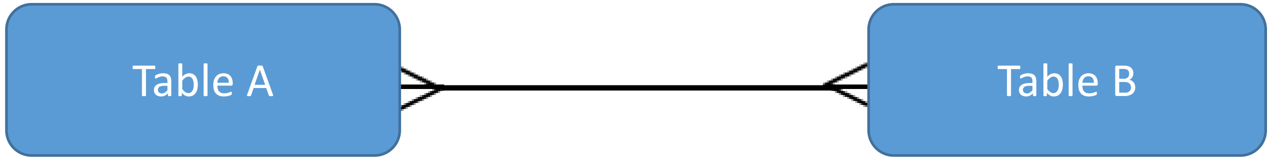
| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ... | ... | ... | ... |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

How do we fix this? This is related to higher normal forms, but…

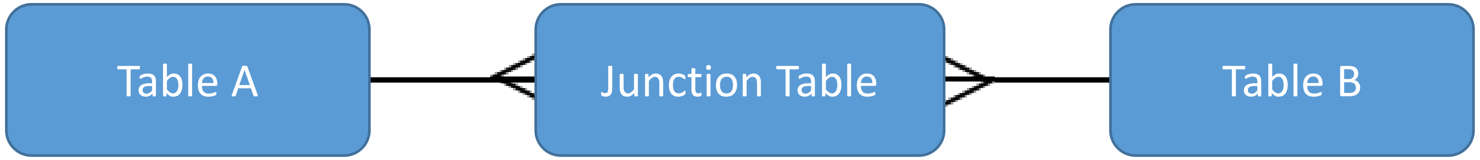
# Junction tables

Many-to-Many relationships can be represented by introducing a **Junction Table**.

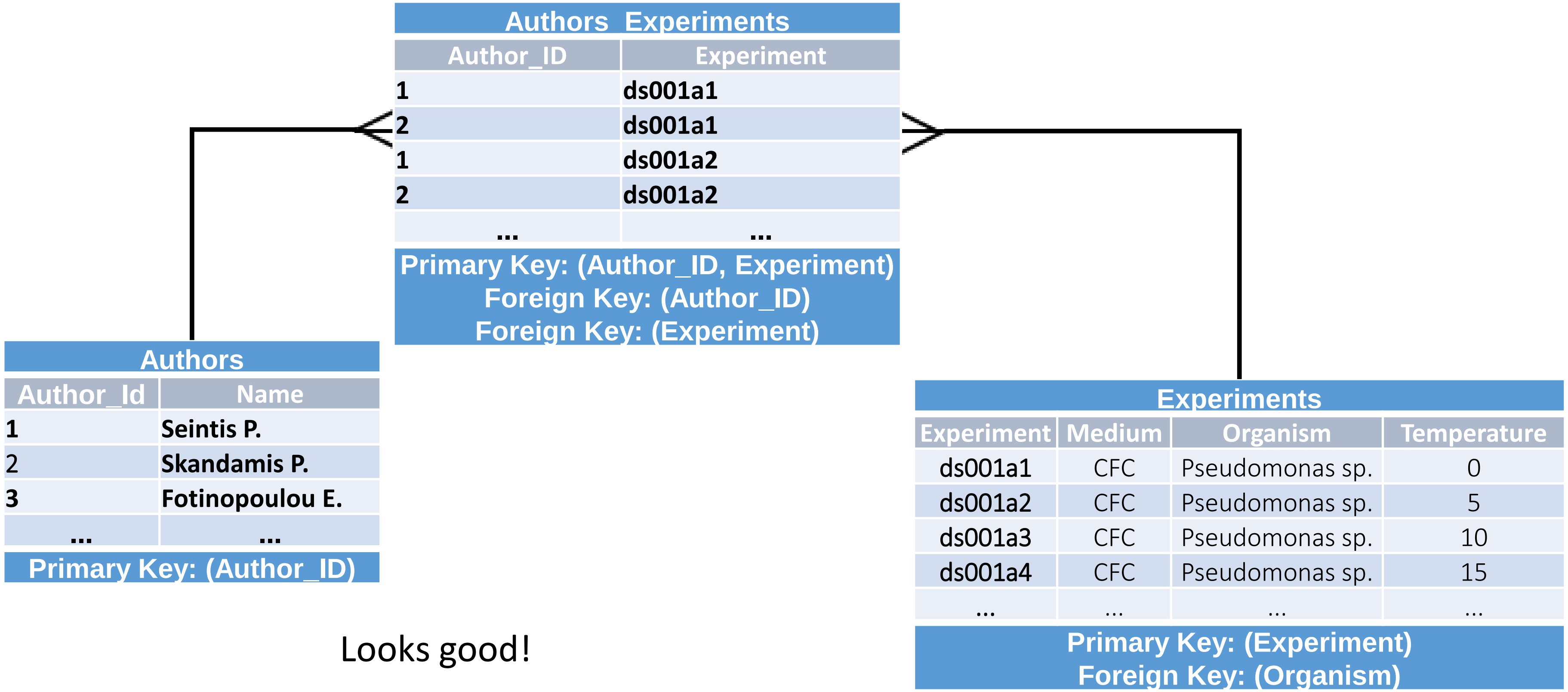


One Many-to-Many relationship is replaced by two One-to-Many relationships.

The Junction table has one Foreign Key referencing Table A, and a second Foreign Key referencing Table B.



# A look at our relationships



**Authors_Experiments**

| Author_ID | Experiment |
|---|---|
| 1 | ds001a1 |
| 2 | ds001a1 |
| 1 | ds001a2 |
| 2 | ds001a2 |
| ... | ... |

**Primary Key: (Author_ID, Experiment)**
**Foreign Key: (Author_ID)**
**Foreign Key: (Experiment)**

**Authors**

| Author_Id | Name |
|---|---|
| 1 | Seintis P. |
| 2 | Skandamis P. |
| 3 | Fotinopoulou E. |
| ... | ... |

**Primary Key: (Author_ID)**

**Experiments**

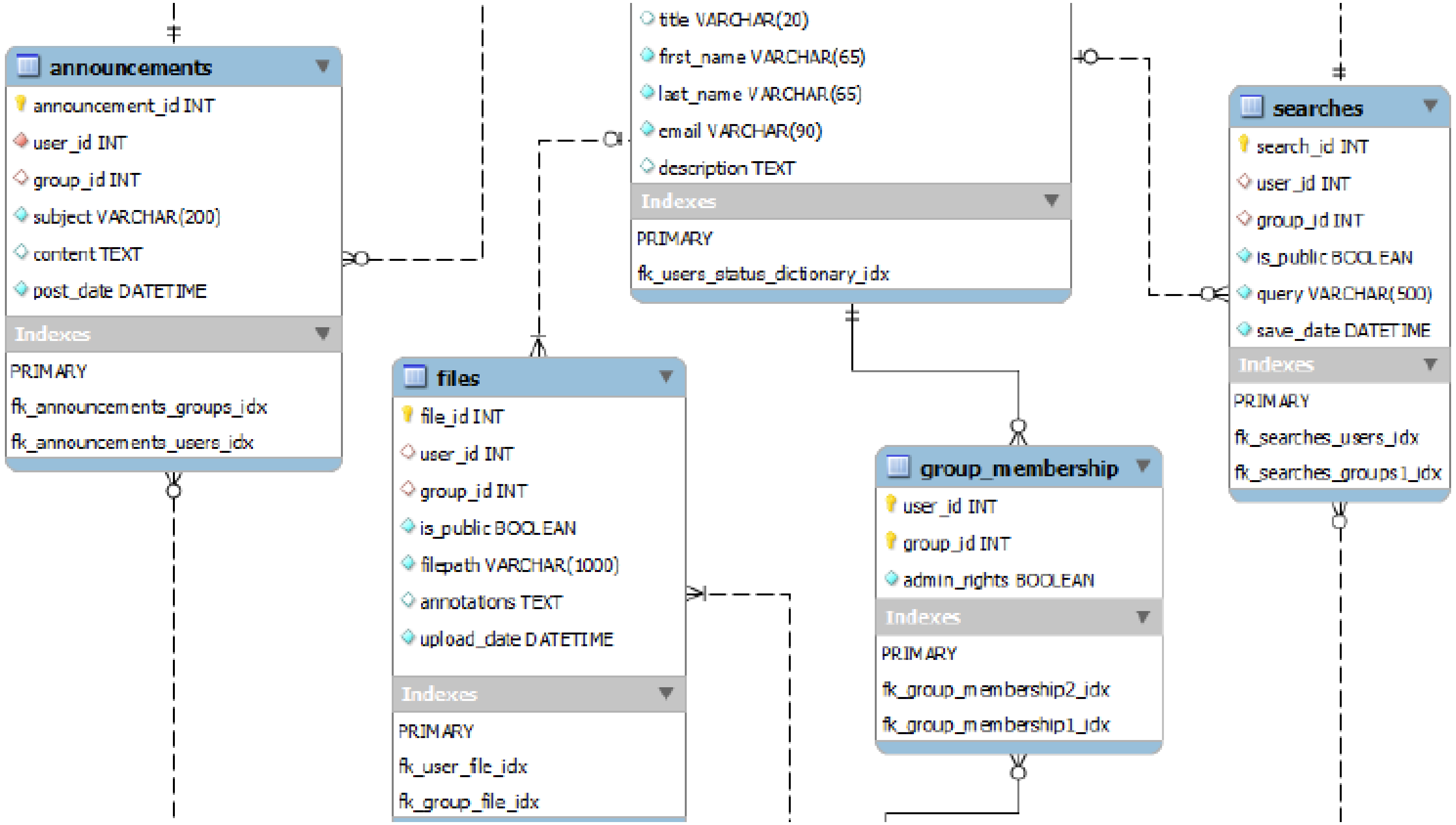
| Experiment | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ... | ... | ... | ... |

**Primary Key: (Experiment)**
**Foreign Key: (Organism)**

Looks good!

"Author_Experiments" is a conventional name, but it may be better to use something meaningful, like "Author_Participation"

# EER diagrams



title VARCHAR(20)
first_name VARCHAR(65)
last_name VARCHAR(65)
email VARCHAR(90)
description TEXT
Indexes
PRIMARY
fk_users_status_dictionary_idx

**announcements**
announcement_id INT
user_id INT
group_id INT
subject VARCHAR(200)
content TEXT
post_date DATETIME
Indexes
PRIMARY
fk_announcements_groups_idx
fk_announcements_users_idx

**files**
file_id INT
user_id INT
group_id INT
is_public BOOLEAN
filepath VARCHAR(1000)
annotations TEXT
upload_date DATETIME
Indexes
PRIMARY
fk_user_file_idx
fk_group_file_idx

**group_membership**
user_id INT
group_id INT
admin_rights BOOLEAN
Indexes
PRIMARY
fk_group_membership2_idx
fk_group_membership1_idx

**searches**
search_id INT
user_id INT
group_id INT
is_public BOOLEAN
query VARCHAR(500)
save_date DATETIME
Indexes
PRIMARY
fk_searches_users_idx
fk_searches_groups1_idx

# ACID – Database transactions

A transaction is a single operation on a relational database. Transactions may include **multiple** changes.

**A**tomicity – either succeed completely or fail completely

**C**onsistency – any written data must respect all database constraints

**I**solation – concurrent transactions are equivalent to sequential transactions

**D**urability – committed transactions remain committed

# SQL

**SQL - S**tructured **Q**uery **L**anguage

By far the most popular language for using relational databases.

Relatively few keywords and simple syntax.

Standardised by ISO and ANSI.

**Declarative** – you describe **what** you want it to do, not **how**. Relatively few keywords and simple syntax.

https://dev.mysql.com/doc/refman/5.7/en/sql-statements.html

# SELECT

```
SELECT column1, column2
FROM table_name
WHERE condition;

SELECT Organism FROM Experiments
WHERE Experiment_ID=’ds001a2’;

SELECT Time, CFU FROM Measurements
WHERE Experiment_ID=’ds001a2’;

SELECT * FROM Experiments;
```

# INSERT

```
INSERT INTO table_name (column1, column2)
VALUES (value1, value2)
WHERE condition;
```

```
INSERT INTO Measurements(Time, CFU)
VALUES (15, 32.5) WHERE
Experiment_ID=’ds001b2’;
```

# UPDATE

```
UPDATE table_name
SET column1=value1, columne2=value2
WHERE condition;

UPDATE Experiments
SET Organism=’E.coli’, Medium=’CFC’
WHERE Experiment_ID=’ds001a2’;
```

# DELETE

```
DELETE FROM table_name
WHERE condition;
```

```
DELETE FROM Measurements
WHERE Experiment_ID=’ds001a2’;
```

# CREATE

```
CREATE TABLE table_name (
 column1 type,
 column2 type,
 …
);

CREATE TABLE IF NOT EXISTS datapoints (
 experiment_id VARCHAR(10),
 time DOUBLE NOT NULL,
 cfu DOUBLE NOT NULL,
 PRIMARY KEY (experiment_id, time),
 FOREIGN KEY (experiment_id)
      REFERENCES experiments
(experiment_id)
);
```

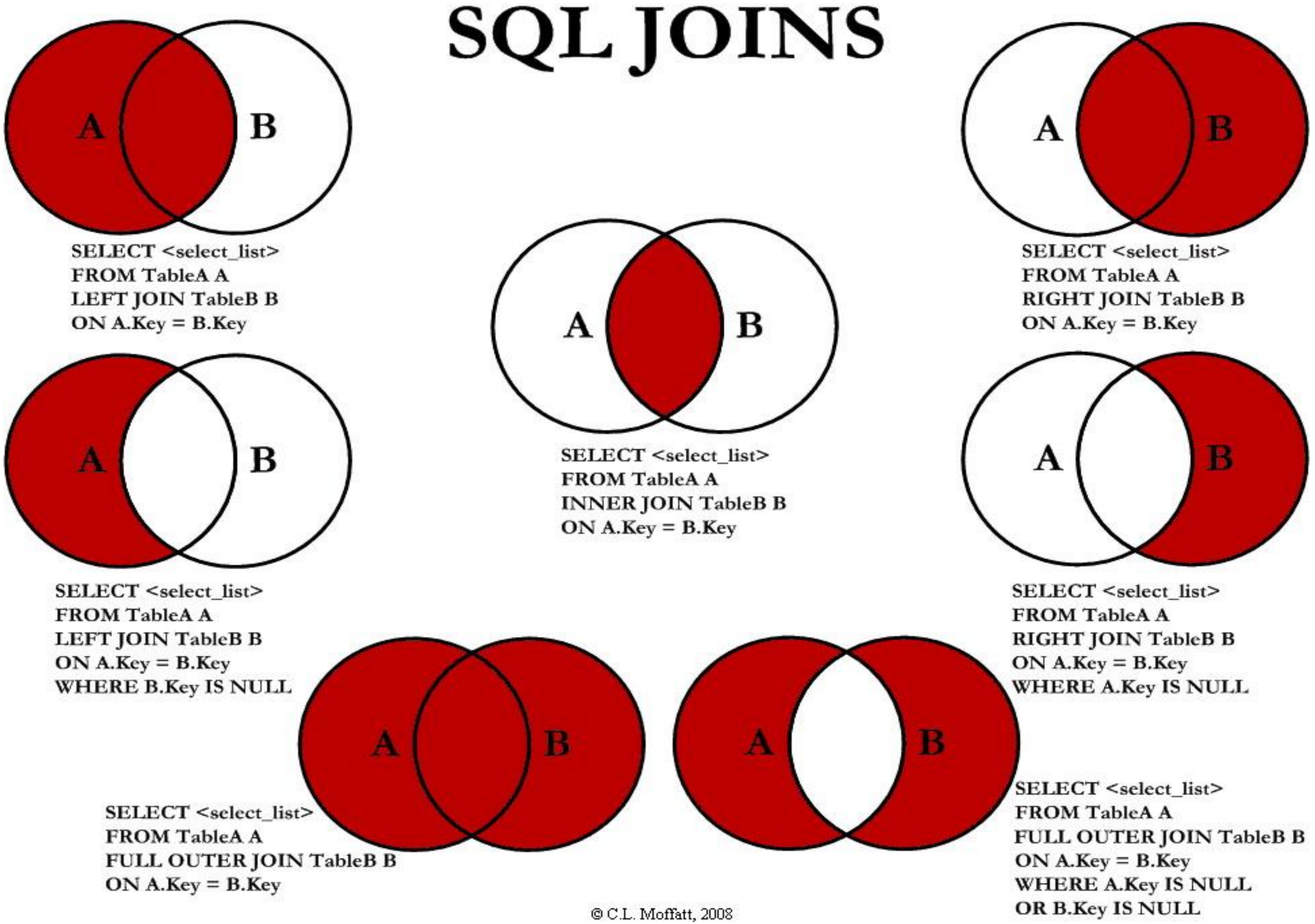
# Joins

**INNER JOIN**: Returns all rows when there is at least one match in BOTH tables

**LEFT JOIN:** Return all rows from the left table, and the matched rows from the right table

**RIGHT JOIN**: Return all rows from the right table, and the matched rows from the left table

**FULL JOIN**: Return all rows when there is a match in ONE of the tables



# SQL JOINS

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```

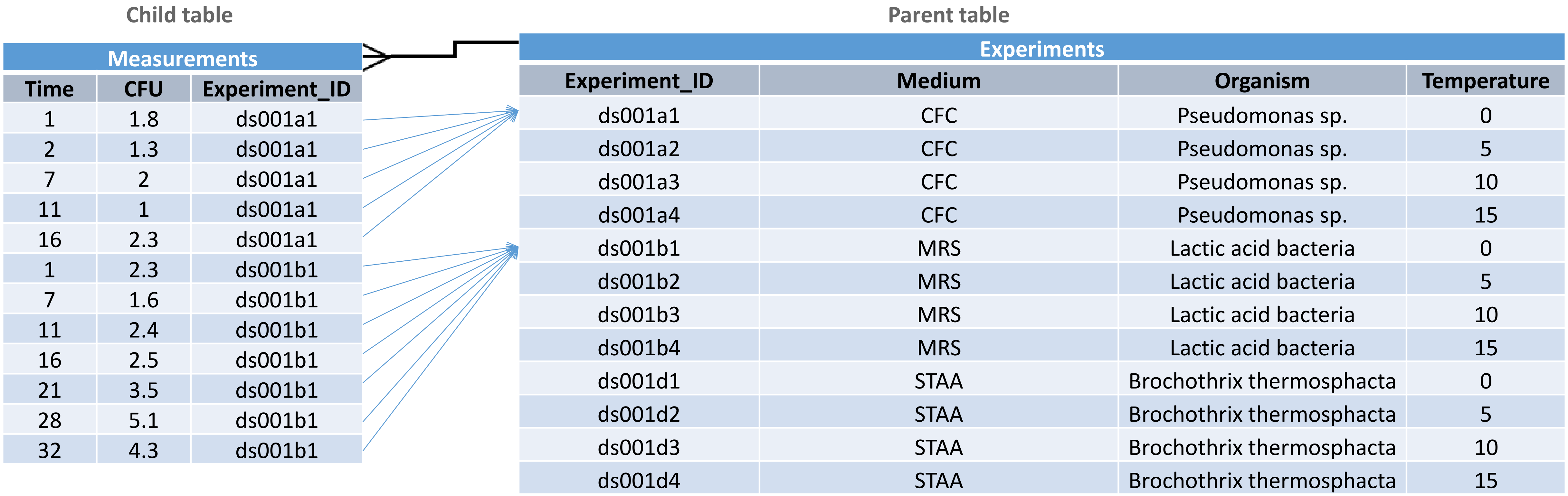
```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

© C.L. Moffatt, 2008

www.codeproject.com/Articles/33052/Visual-Representation-of-SQL-Joins

# Measurements and Experiments



Child table

**Measurements**

| Time | CFU | Experiment_ID |
|---|---|---|
| 1 | 1.8 | ds001a1 |
| 2 | 1.3 | ds001a1 |
| 7 | 2 | ds001a1 |
| 11 | 1 | ds001a1 |
| 16 | 2.3 | ds001a1 |
| 1 | 2.3 | ds001b1 |
| 7 | 1.6 | ds001b1 |
| 11 | 2.4 | ds001b1 |
| 16 | 2.5 | ds001b1 |
| 21 | 3.5 | ds001b1 |
| 28 | 5.1 | ds001b1 |
| 32 | 4.3 | ds001b1 |

Parent table

**Experiments**

| Experiment_ID | Medium | Organism | Temperature |
|---|---|---|---|
| ds001a1 | CFC | Pseudomonas sp. | 0 |
| ds001a2 | CFC | Pseudomonas sp. | 5 |
| ds001a3 | CFC | Pseudomonas sp. | 10 |
| ds001a4 | CFC | Pseudomonas sp. | 15 |
| ds001b1 | MRS | Lactic acid bacteria | 0 |
| ds001b2 | MRS | Lactic acid bacteria | 5 |
| ds001b3 | MRS | Lactic acid bacteria | 10 |
| ds001b4 | MRS | Lactic acid bacteria | 15 |
| ds001d1 | STAA | Brochothrix thermosphacta | 0 |
| ds001d2 | STAA | Brochothrix thermosphacta | 5 |
| ds001d3 | STAA | Brochothrix thermosphacta | 10 |
| ds001d4 | STAA | Brochothrix thermosphacta | 15 |

# Measurements and Experiments

```
SELECT Organism, Medium, Time, CFU

FROM Experiments JOIN Measurements ON

Experiments.Experiment_ID=Measurements.Experiment_ID WHERE Temperature=0;
```

| Result | | | |
|---|---|---|---|
| **Organism** | **Medium** | **Time** | **CFU** |
| Pseudomonas sp. | CFC | 1 | 1.8 |
| Pseudomonas sp. | CFC | 2 | 1.3 |
| Pseudomonas sp. | CFC | 7 | 2 |
| Pseudomonas sp. | CFC | 11 | 1 |
| Pseudomonas sp. | CFC | 16 | 2.3 |
| Lactic acid bacteria | MRS | 1 | 2.3 |
| Lactic acid bacteria | MRS | 7 | 1.6 |
| Lactic acid bacteria | MRS | 11 | 2.4 |
| Lactic acid bacteria | MRS | 16 | 2.5 |
| Lactic acid bacteria | MRS | 21 | 3.5 |
| Lactic acid bacteria | MRS | 28 | 5.1 |
| Lactic acid bacteria | MRS | 32 | 4.3 |

# RDBMS

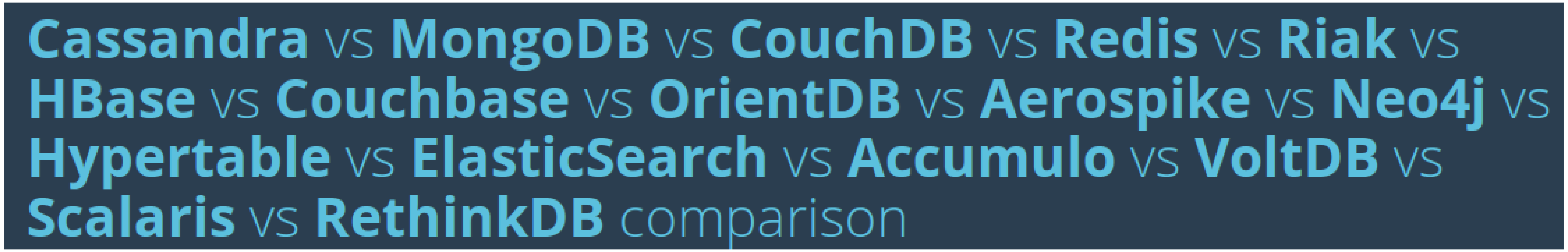
Relational database management systems

- SQLite
- MySQL
- Oracle
- Microsoft SQL Server
- MariaDB
- PostgreSQL

# NoSQL

Non-SQL…

…or Not Only SQL



**Cassandra** vs **MongoDB** vs **CouchDB** vs **Redis** vs **Riak** vs **HBase** vs **Couchbase** vs **OrientDB** vs **Aerospike** vs **Neo4j** vs **Hypertable** vs **ElasticSearch** vs **Accumulo** vs **VoltDB** vs **Scalaris** vs **RethinkDB** comparison

https://kkovacs.eu/cassandra-vs-mongodb-vs-couchdb-vs-redis

# SQLite

Free, open source.

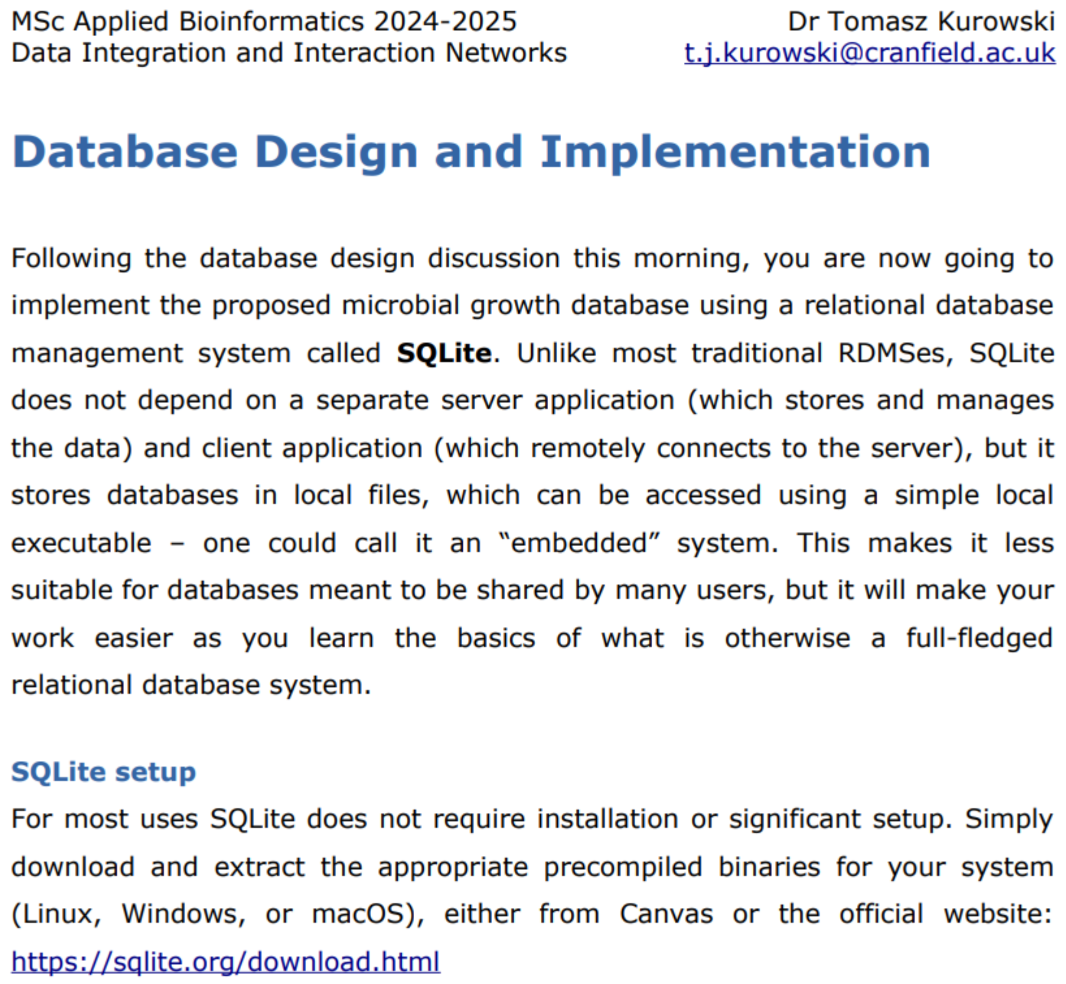
No client-server setup - "embedded" approach.

Databases stored in local files.

Good for small projects and learning SQL.

One of the most popular pieces of software in the world – you already have it!

# That’s all!



MSc Applied Bioinformatics 2024-2025
Data Integration and Interaction Networks

Dr Tomasz Kurowski
t.j.kurowski@cranfield.ac.uk

## Database Design and Implementation

Following the database design discussion this morning, you are now going to implement the proposed microbial growth database using a relational database management system called **SQLite**. Unlike most traditional RDMSes, SQLite does not depend on a separate server application (which stores and manages the data) and client application (which remotely connects to the server), but it stores databases in local files, which can be accessed using a simple local executable – one could call it an “embedded” system. This makes it less suitable for databases meant to be shared by many users, but it will make your work easier as you learn the basics of what is otherwise a full-fledged relational database system.

### SQLite setup

For most uses SQLite does not require installation or significant setup. Simply download and extract the appropriate precompiled binaries for your system (Linux, Windows, or macOS), either from Canvas or the official website: https://sqlite.org/download.html