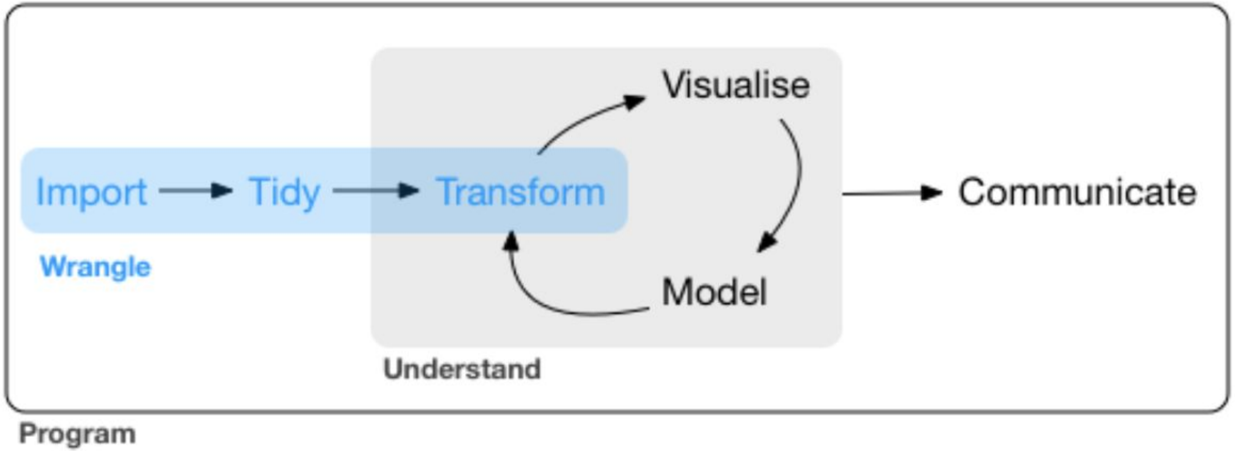
3.1 Data wrangling

Applied Data Analysis (ADA)

Oxford DH Summer School - 2019

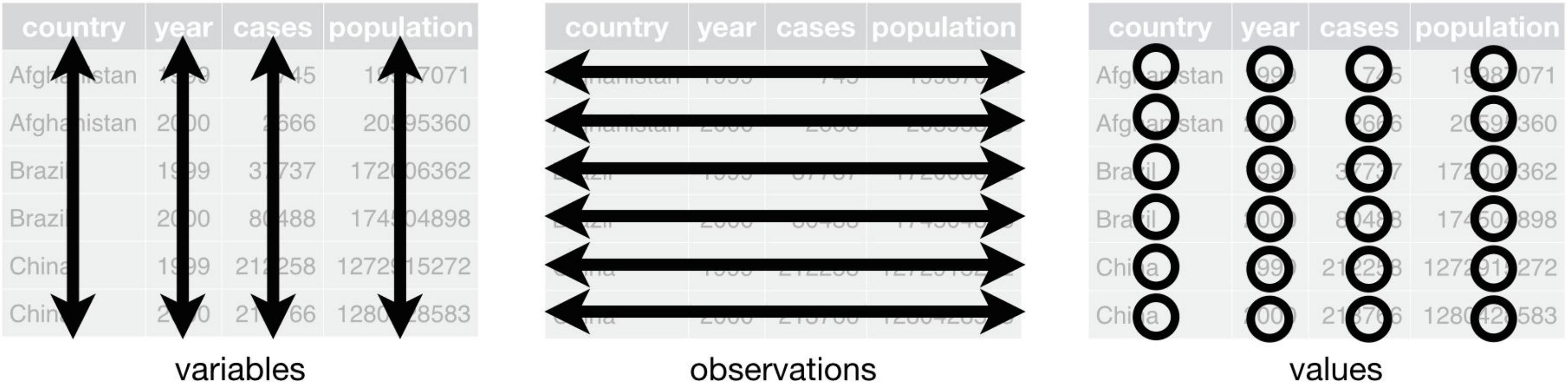


https://r4ds.had.co.nz/tidy-data.html

http://users.dimi.uniud.it/~massimo.franceschet/ds/syllabus/learn/database/algebra.html

There are few interrelated rules which make a dataset tidy:

1. The **dataset** is organized into a collection of **tables** (or relations, or data frames).
2. Every **table** contains data for a single **observation type** (or class).
3. Each **variable** (or attribute) must have its own **column**.
4. Each **observation** (or instance, or tuple) must have its own **row**.
5. Each **value** must have its own **cell**.



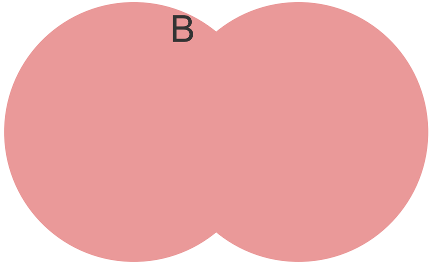
Sets

A = {a1, a2, b3}, B = {b1, b2, a1}

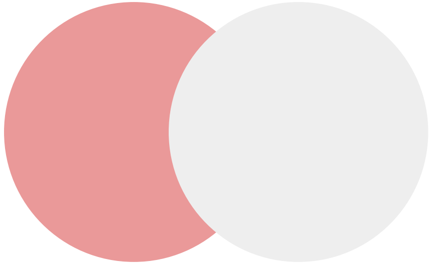
Sets for us are tables. Crucially, we need to define keys for every observation of the set/table.

Operations on sets

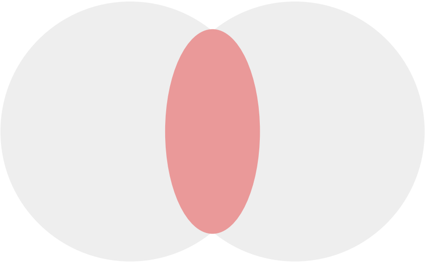
A ∪ B



A / B



A ∩ B



Tables

A

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The force awakens	2015	J. J. Abrams

Projection

A [Title, Director] ⟶

Title	Director
Blade runner	Ridley Scott
Star wars - The empire strikes back	Irvin Kershner

Tables

A

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The force awakens	2015	J. J. Abrams

Selection

A [Year > 1981] ⟶

Title	Year	Director
Blade runner	1982	Ridley Scott

Tables

A

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The force awakens	2015	J. J. Abrams

Union

$A \cup B \rightarrow$

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner
Star wars - The force awakens	2015	J. J. Abrams

Tables

A

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The force awakens	2015	J. J. Abrams

Difference

A / B →

Title	Year	Director
Star wars - The empire strikes back	1980	Irvin Kershner

Tables

A

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The force awakens	2015	J. J. Abrams

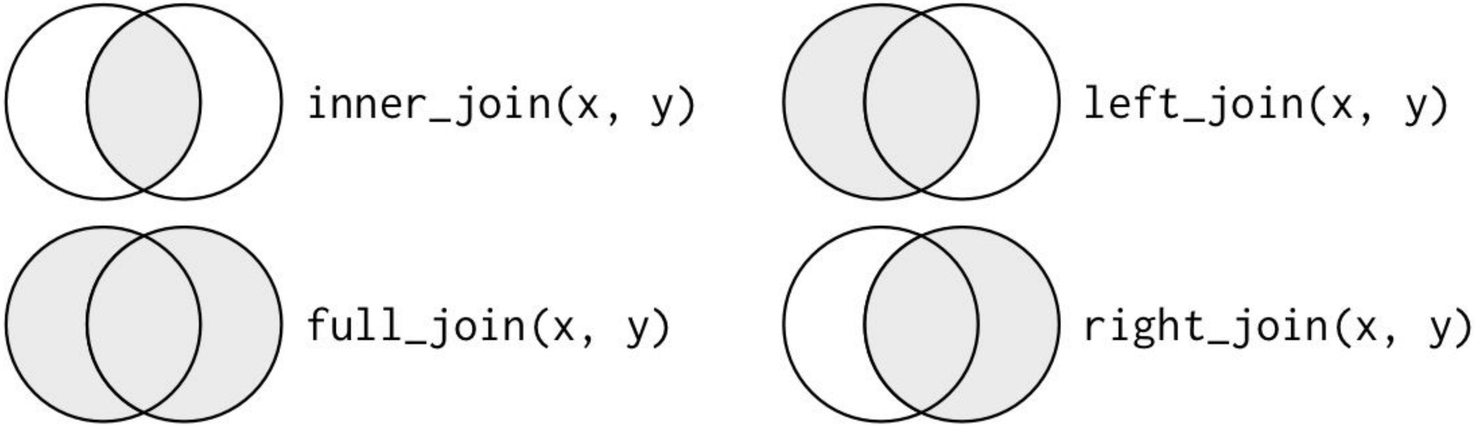
Intersection

$A \cap B \longrightarrow$

Title	Year	Director
Blade runner	1982	Ridley Scott

Joins

We consider tables with different variables for the same observational unit (class) and (possibly) the same observations.



Tables

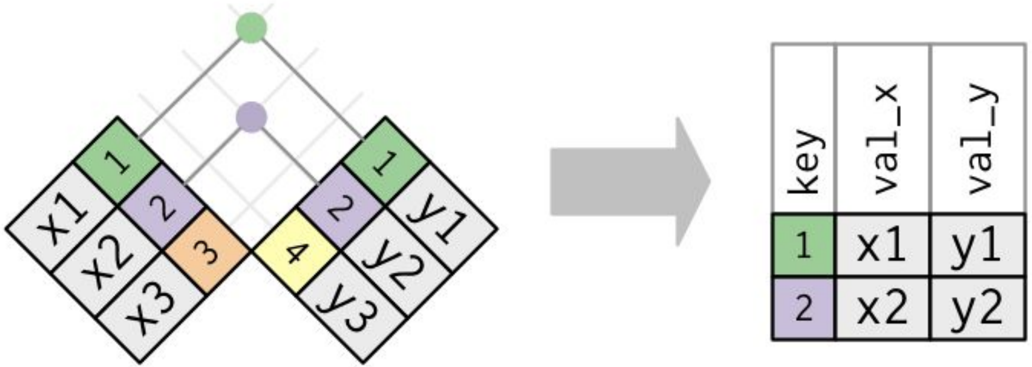
A

Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Budget	Box office
Blade runner	28M	33.8M
Star wars - The force awakens	306M	2,068M

Inner join



key	val_x	val_y
1	x1	y1
2	x2	y2

Tables

A

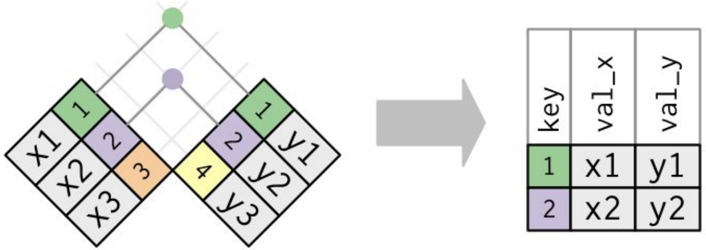
Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Budget	Box office
Blade runner	28M	33.8M
Star wars - The force awakens	306M	2,068M

Inner join

Title	Year	Director	Budget	Box office
Blade runner	1982	Ridley Scott	28M	33.8M



key	val_x	val_y
1	x1	y1
2	x2	y2

Tables

A

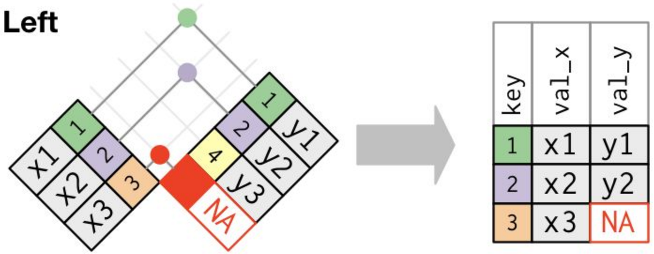
Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Budget	Box office
Blade runner	28M	33.8M
Star wars - The force awakens	306M	2,068M

Left join

Title	Year	Director	Budget	Box office
Blade runner	1982	Ridley Scott	28M	33.8M
Star wars - The empire strikes back	1980	Irvin Kershner	NA	NA



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Tables

A

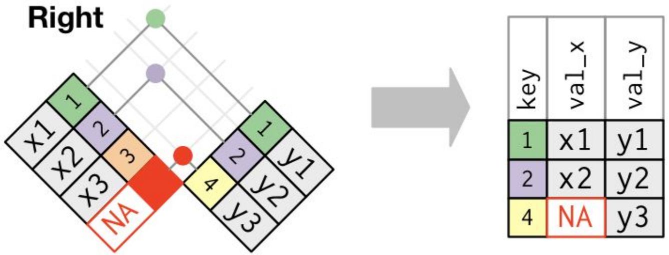
Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Budget	Box office
Blade runner	28M	33.8M
Star wars - The force awakens	306M	2,068M

Right join

Title	Year	Director	Budget	Box office
Blade runner	1982	Ridley Scott	28M	33.8M
Star wars - The force awakens	NA	NA	306M	2,068M



key	val_x	val_y
1	x1	y1
2	x2	y2
4	NA	y3

Tables

A

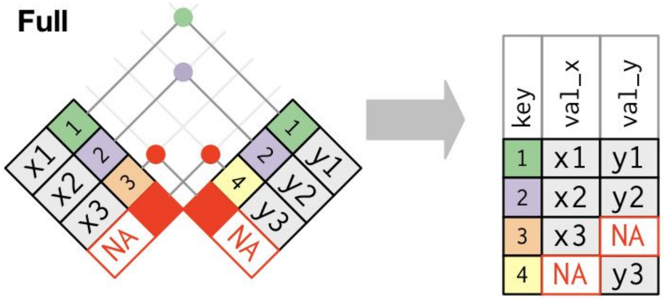
Title	Year	Director
Blade runner	1982	Ridley Scott
Star wars - The empire strikes back	1980	Irvin Kershner

B

Title	Budget	Box office
Blade runner	28M	33.8M
Star wars - The force awakens	306M	2,068M

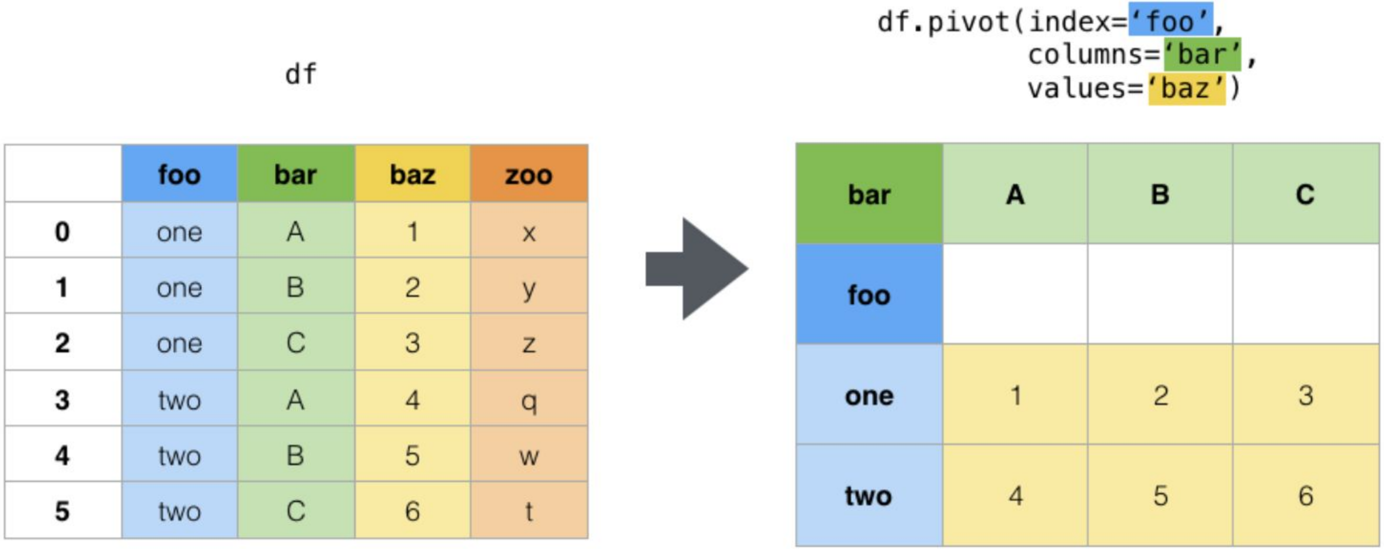
Full join

Title	Year	Director	Budget	Box office
Blade runner	1982	Ridley Scott	28M	33.8M
Star wars - The empire strikes back	1980	Irvin Kershner	NA	NA
Star wars - The force awakens	NA	NA	306M	2,068M



key	val1_x	val1_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

Pivoting



df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

```
df.pivot(index=‘foo’,
         columns=‘bar’,
         values=‘baz’)
```

bar	**A**	**B**	**C**
foo			
one	1	2	3
two	4	5	6

SQL and relational databases

The databases we have explored so far are called *relational databases*. Examples include MySQL and Postgresql.

Relational databases mostly implement a unified query language called *SQL* (Structured Query Language): https://en.wikipedia.org/wiki/SQL.

NoSQL

Document databases

XML, json, .. Examples: BaseX, MongoDB.

Graph databases

Linked data, graphs. Examples: Allegro, Neo4j.

See: https://en.wikipedia.org/wiki/NoSQL.