An Example

Consider the following dataset showing four female tennis players. Notice that the last column is trying to put multiple pieces of information in one value.

	A	B	C
1	**player**	**nationality**	**majors_won**
2	Ashleigh Barty	AUS	Wimbledon|French Open
3	Emma Raducanu	GBR	US Open
4	Barbora Krejčíková	CZE	French Open
5	Naomi Osaka	JPN	Australian Open|US Open

unit of observation: **Player**

First Normal Form (1NF)

Here is an alternative that only has one piece of information in each cell. This table is said to be in the **First Normal Form (1NF)**.

1NF requires that the data be in a tabular format with only one individual piece of information in each cell. Here is one way to do that using the same unit of observation (note that a long format might be better for many kinds of analysis).

	A	B	C	D	E	F
1	**player**	**nationality**	**won_wimbledon**	**won_us_open**	**won_french_open**	**won_australian_open**
2	Ashleigh Barty	AUS	1	0	1	0
3	Emma Raducanu	GBR	0	1	0	0
4	Barbora Krejčíková	CZE	0	0	1	0
5	Naomi Osaka	JPN	1	0	0	1

unit of observation: **Player**

Another Example

Here is another example of a similar dataset. This one gives information about each Grand Slam itself. It is in **1NF** but notice that it duplicates information.

	A	B	C	D	E
1	**tournament**	**year**	**tournament_country**	**winner**	**winner_nationality**
2	Australian Open	2020	AUS	Sofia Kenin	USA
3	French Open	2020	FRA	Iga Świątek	POL
4	Wimbledon	2020	GBR		
5	US Open	2020	USA	Naomi Osaka	JPN
6	Australian Open	2021	AUS	Naomi Osaka	JPN
7	French Open	2021	FRA	Barbora Krejčíková	CZE
8	Wimbledon	2021	GBR	Ashleigh Barty	AUS
9	US Open	2021	USA	Emma Raducanu	GBR

unit of observation: **tournament x year**

Second Normal Form (2NF)

We can fix the some problems of duplication by creating two different tables. Namely, we want all of the information in the table to relate to the unit of observation. All of the columns should be part of a *candidate key* or defined by the entire candidate key. The data is now follows the rules for the **second normal form (2NF)**.

	A	B	C	D
1	**tournament**	**year**	**winner**	**winner_nationality**
2	Australian Open	2020	Sofia Kenin	USA
3	French Open	2020	Iga Świątek	POL
4	Wimbledon	2020		
5	US Open	2020	Naomi Osaka	JPN
6	Australian Open	2021	Naomi Osaka	JPN
7	French Open	2021	Barbora Krejčíková	CZE
8	Wimbledon	2021	Ashleigh Barty	AUS
9	US Open	2021	Emma Raducanu	GBR

unit of observation: **tournament x year**

	A	B
1	**tournament**	**tournament_country**
2	Australian Open	AUS
3	French Open	FRA
4	Wimbledon	GBR
5	US Open	USA

unit of observation: **tournament**

Third Normal Form (3NF)

There is still a kind of duplication in the previous dataset because the winner nationality is a function of the winner name. Putting the tables into the **third normal form**, or 3NF, requires fixing this dependency as well.

	A	B	C
1	**tournament**	**year**	**winner**
2	Australian Open	2020	Sofia Kenin
3	French Open	2020	Iga Świątek
4	Wimbledon	2020	
5	US Open	2020	Naomi Osaka
6	Australian Open	2021	Naomi Osaka
7	French Open	2021	Barbora Krejčíková
8	Wimbledon	2021	Ashleigh Barty
9	US Open	2021	Emma Raducanu

unit of observation: **tournament x year**

	A	B
1	**tournament**	**tournament_country**
2	Australian Open	AUS
3	French Open	FRA
4	Wimbledon	GBR
5	US Open	USA

unit of observation: **tournament**

	A	B
1	**player**	**nationality**
2	Sofia Kenin	USA
3	Iga Świątek	POL
4	Naomi Osaka	JPN
5	Barbora Krejčíková	CZE
6	Ashleigh Barty	AUS
7	Emma Raducanu	GBR

unit of observation: **player**