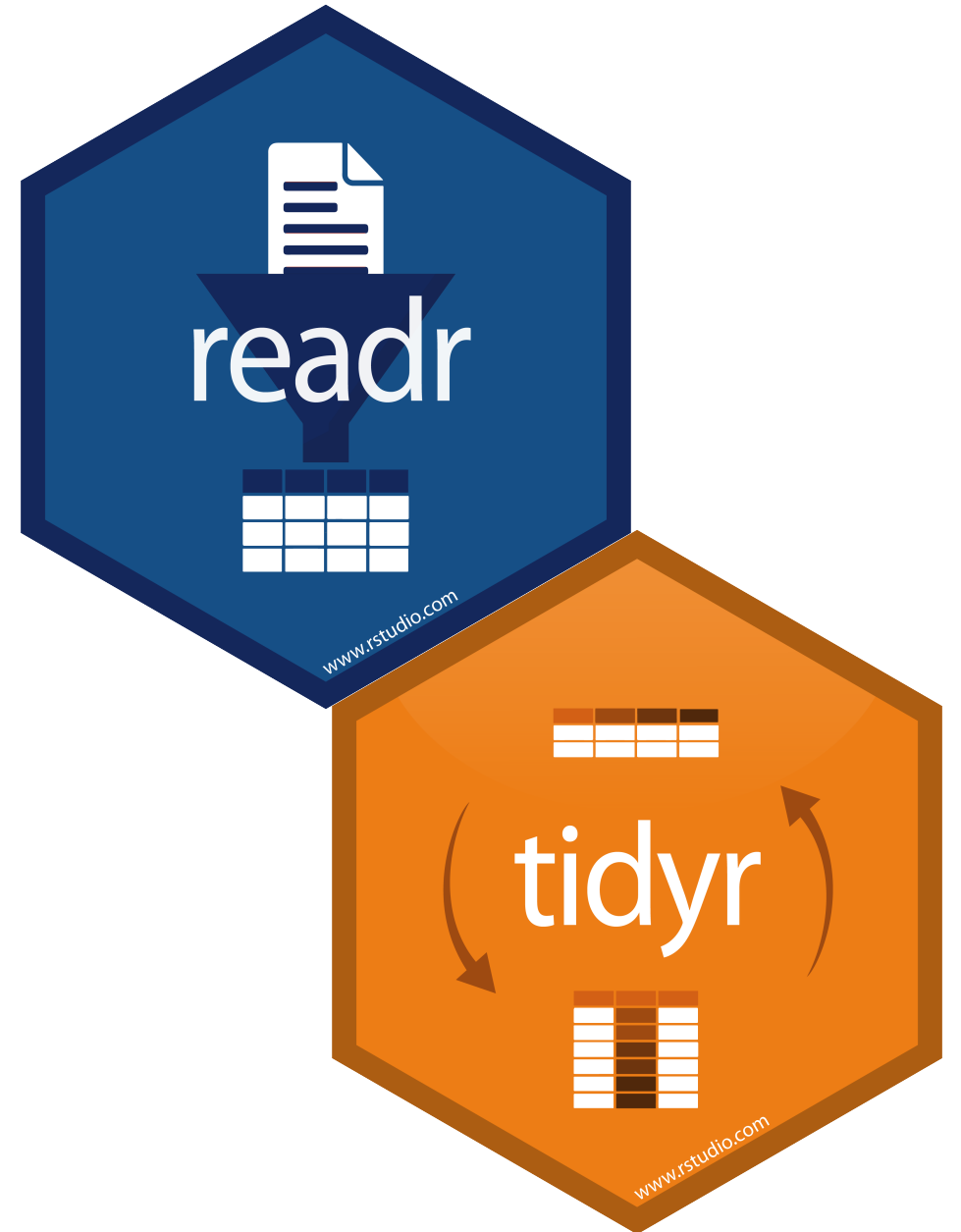


Reading Data with `readr` and Tidying Data with `tidyr`

2020-07-14



Reading Data

Data File Formats

Data is stored in plain text files with a **delimiter** specifying the boundaries between data entries. The most common delimiters are tabs or commas.

tab separated values (TSV)

```
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
5.1 3.5 1.4 0.2 setosa
4.9 3 1.4 0.2 setosa
4.7 3.2 1.3 0.2 setosa
4.6 3.1 1.5 0.2 setosa
5 3.6 1.4 0.2 setosa
```

comma separated values (CSV)

```
Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Species
5.1,3.5,1.4,0.2,setosa
4.9,3,1.4,0.2,setosa
4.7,3.2,1.3,0.2,setosa
4.6,3.1,1.5,0.2,setosa
5,3.6,1.4,0.2,setosa
```

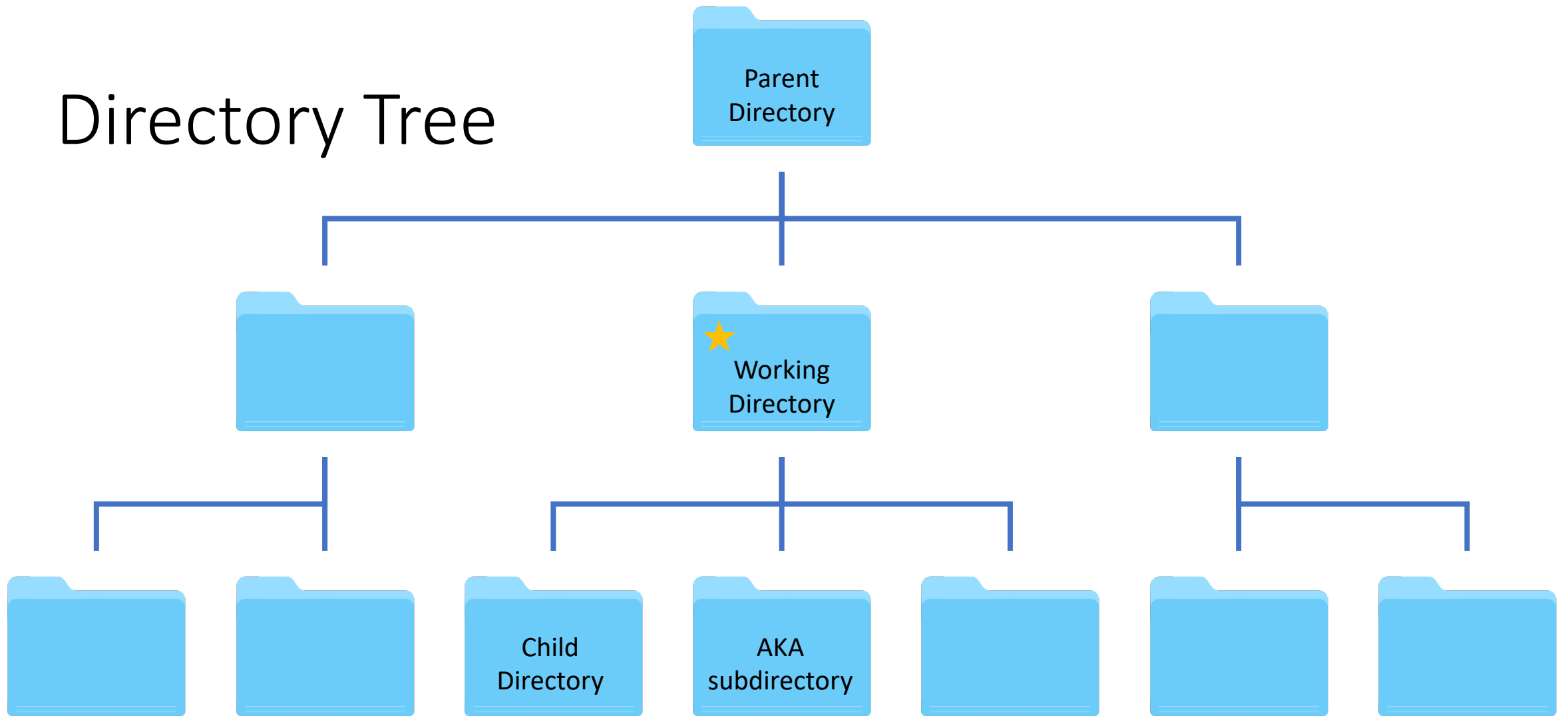
spaces

```
|Sepal.Length Sepal.Width Petal.Length Petal.Width Species|
5.1 3.5 1.4 0.2 setosa
4.9 3 1.4 0.2 setosa
4.7 3.2 1.3 0.2 setosa
4.6 3.1 1.5 0.2 setosa
5 3.6 1.4 0.2 setosa
```

Or any other character (BUT NEVER DO THIS)

```
Sepal.Length/Sepal.Width/Petal.Length/Petal.Width/Species
5.1/3.5/1.4/0.2/setosa
4.9/3/1.4/0.2/setosa
4.7/3.2/1.3/0.2/setosa
4.6/3.1/1.5/0.2/setosa
5/3.6/1.4/0.2/setosa
```

Directory Tree



File Path: `working_directory/child_directory`

Tidying Data

Tidy Data

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280425583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280425583

observations

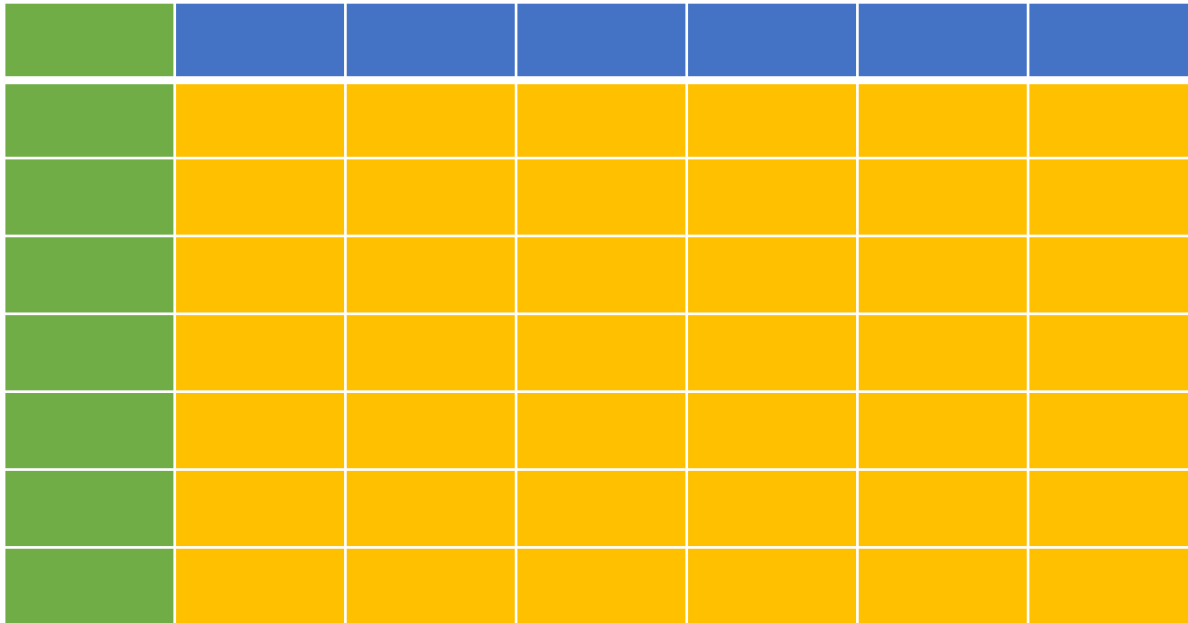
country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280425583

values

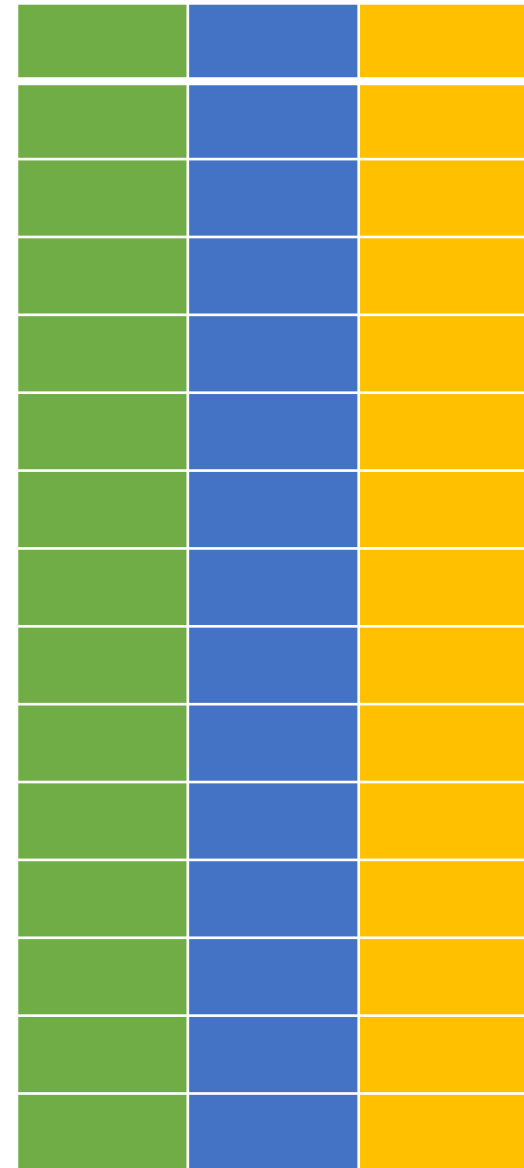
1. Each variable is in a column.
2. Each observation is a row.
3. Each value is a cell.

Wide vs Skinny Data

Wide



Skinny



DEMO