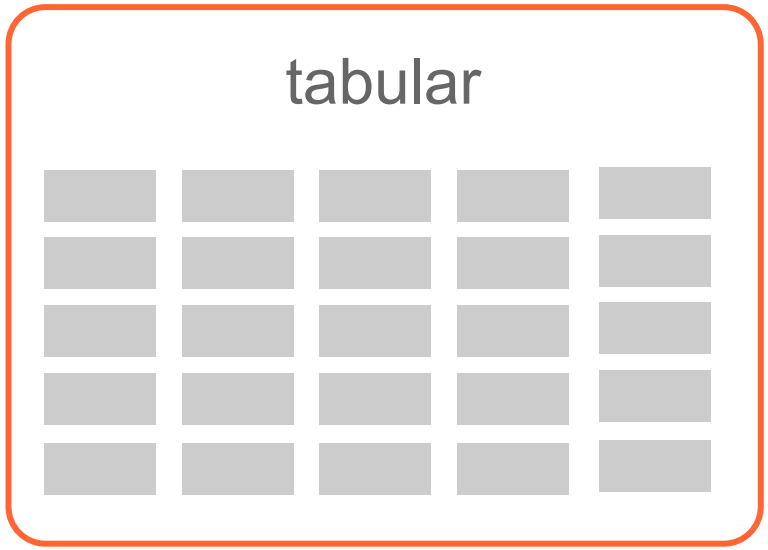
Data Tables

Stat 133 by Gaston Sanchez

Tabular vs Non-tabular



First part of the course

By the way

Data is ready to be analyzed

Textbook / Classroom monumental assumption

VS

Data requires some processing

Rarely taught and/or discussed

Tables

A Data Table

Name	Gender	Homeland	Birthyr	Mass	Height	Jedi
Luke	male	Tatooine	-99999	77kg	1.72m	yes
Leia	female	Alderaan	19BBY	49kg	1.50m	no
Obi-Wan	male	Stewjon	57BBY	77kg	1.82m	yes
Han	MALE	Corellia	29BBY	80kg	1.80m	no
Anakin	male	Tatooine	41.9BBY	84kg	1.88m	NA
Amidala	female	Naboo	46BBY	45kg	1.65m	no

Name	Gender	Homeland	Birthyr	Mass	Height	Jedi
Luke	male	Tatooine	-99999	77kg	1.72m	yes
Leia	female	Alderaan	19BBY	49kg	1.50m	no
Obi-Wan	male	Stewjon	57BBY	77kg	1.82m	yes
Han	MALE	Corellia	29BBY	80kg	1.80m	no
Anakin	male	Tatooine	41.9BBY	84kg	1.88m	NA
Amidala	female	Naboo	46BBY	45kg	1.65m	no

BBY: *Before the Battle of Yavin*

Tabular Data

Most data analysis methods (e.g. statistical, predictive models, machine learning, data mining, AI algorithms) require data in tabular format

Not necessarily the way data is collected, stored, organized, formatted

And even when data is in a table format ...

Name	Gender	Homeland	Birthyr	Mass	Height	Jedi
Luke	male	Tatooine	-99999	77kg	1.72m	yes
Leia	female	Alderaan	19BBY	49kg	1.50m	no
Obi-Wan	male	Stewjon	57BBY	77kg	1.82m	yes
Han	MALE	Corellia	29BBY	80kg	1.80m	no
Anakin	male	Tatooine	41.9BBY	84kg	1.88m	NA
Amidala	female	Naboo	46BBY	45kg	1.65m	no

there's still a lot of work to do
to have data ready to be analyzed

Types of Tables

Leia

Luke

Han

Heterogeneous data

Name	Sex	Height	Weight
Luke	male	1.72	77
Leia	female	1.50	49
Han	male	1.80	80

Binary (e.g. presence-absence)

Which products do you buy?

	Beer	Wine	Juice	Coffee	Tea
Luke	1	1	1	0	0
Leia	0	1	1	0	1
Han	1	0	1	1	0

(1 = yes, 0 = no)

Modalities

How often do you drink the following products?

	Beer	Wine	Juice	Coffee	Tea
Luke	3	2	3	1	1
Leia	1	2	3	2	3
Han	3	2	2	3	1

(1=never, 2=sometimes, 3=always)

Preference

How much do you like the following juice flavors?

	Orange	Apple	Grapefruit	Carrot
Luke	3	5	2	1
Leia	1	4	5	2
Han	2	1	3	5

(1=don't like at all, 5=like it very much)

Cross-table

Character	Episode IV	Episode V	Episode VI
Luke	254	128	112
Leia	57	114	56
Han	153	182	124
Vader	41	56	43
Emperor	0	5	39

Number of dialogues per character

Proximities & Dissimilarities

	Luke	Leia	Han	Vader	Emperor
Luke	1	0.8	0.7	0.3	0.2
Leia	0.8	1	0.8	0.2	0.2
Han	0.7	0.8	1	0.1	0.1
Vader	0.3	0.2	0.1	1	0.9
Emperor	0.2	0.2	0.1	0.9	1

1 = perfect similarity, 0 = totally different

Plain Text Formats (tabular data)

Leia Skywalker
Female
1.50m tall

Luke Skywalker
Male
1.72m tall

Han Solo
Male
1.80m tall

How to store tables in a file?

name	gender	height
Leia Skywalker	female	1.50
Luke Skywalker	male	1.72
Han Solo	male	1.80

Analyst /Scientist

Tables in Spreadsheets

	A	B	C
1	name	gender	height
2	Leia Skywalker	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Storing a data table

Many people enter and store their data in spreadsheets

e.g. MS Excel, Google Sheets, Apple Numbers

Using spreadsheet provides a nice graphical display of a table's content

Using spreadsheet software brings (a deceptive) comfort

Formatted content in spreadsheets

	A	B	C
1	**name**	**gender**	**height**
2	Leia Skywalker	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Formatted content in spreadsheets

	A	B	C
1	**name**	**gender**	**height**
2	Leia Skywalker	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Formatted content in spreadsheets

	A	B	C
1	**name**	**gender**	**height**
2	Leia Skywalker	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Formatted content in spreadsheets

	A	B	C
1	**name**	**gender**	**height**
2	Leia Skywalker	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

Avoid Excel

	A	B	C
1	name	gender	height
2	Leia Skywalker	female	1.50
3	Luke Skywalker	male	1.72
4	Han Solo	male	1.80

What happens when you try to open a spreadsheet file in a text editor?

Spreadsheet inconveniences

Excel files (.xls) are NOT text files

They are **enriched** files with added format elements

Cannot be opened with a text editor

You depend on ***proprietary*** software

Still wanna save tables as `.xlsx` (or `.xls`)?

Every time you save a
data file in **.xlsx** format ...

God kills a kitten

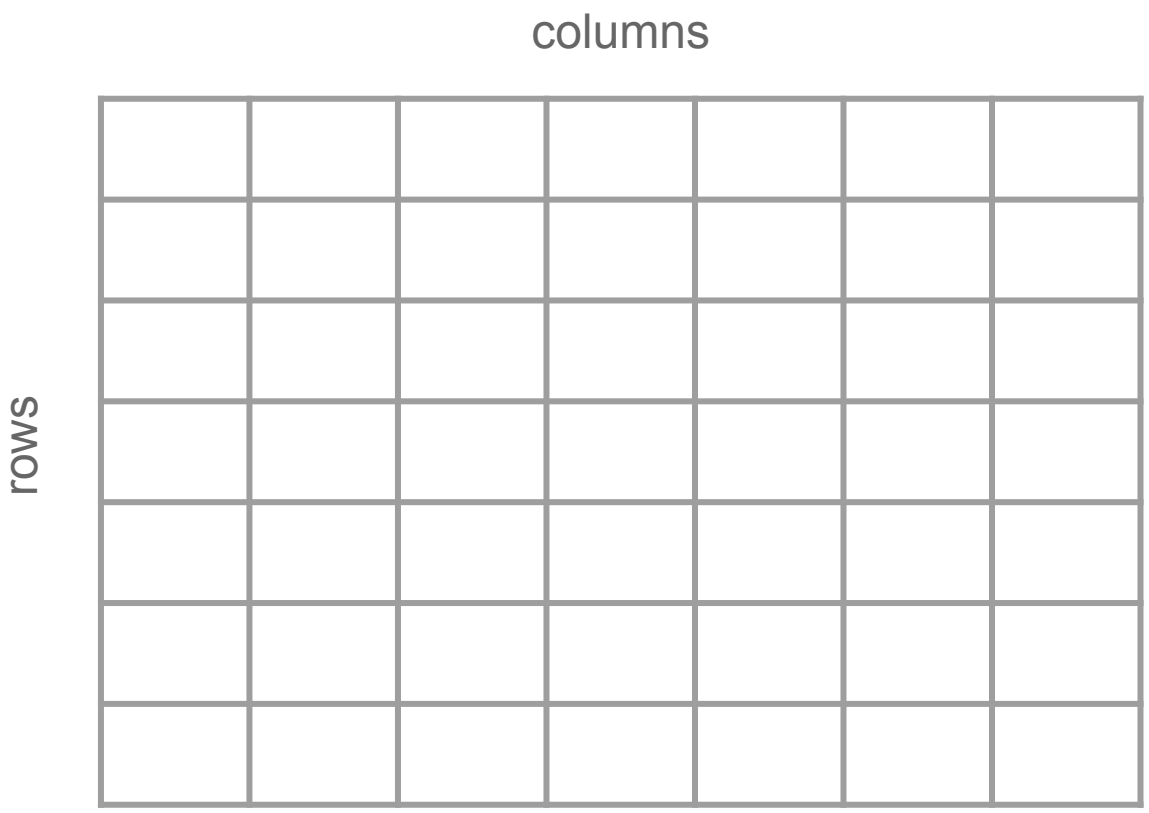
Good Practice

Whenever you work with some source of data stored in a native spreadsheet format (e.g. .xls, .xlsx, .numbers), always generate a text version (e.g. **`.csv, .txt, .dat`**)

Tabular (rectangular) Data

How to store data cells?

What type of format?



How do we “encode” cells in a file?

Character Delimited Text

A common way to store data in tabular form is via text files

To store the data we need a way to separate data values

Each line represents a “row”

The idea of columns is conveyed with delimiters

Plain Text Formats

There are 2 main subtypes of plain text format, depending on how the separated values are identified in a row

1) Delimited formats

2) Fixed-width formats

Common Delimited Formats

Delimiter	Description
`" "`	White space
`","`	Comma
`"\t"`	Tab
`";"`	Semicolon

Consider the following data table

Name	Gender	Homeland	Born	Jedi
Anakin	male	Tatooine	41.9BBY	yes
Amidala	female	Naboo	46BBY	no
Luke	male	Tatooine	19BBY	yes
Leia	female	Alderaan	19BBY	no
Obi-Wan	male	Stewjon	57BBY	yes
Han	male	Corellia	29BBY	no
Palpatine	male	Naboo	82BBY	no
R2-D2	unknown	Naboo	33BBY	no

Tab delimited format

Each value separated by a tab `"\t"`
Typical file extensions: txt

```
Name      Gender   Homeworld  Born     Jedi
Anakin    male     Tatooine   41.9BBY  yes
Amidala   female   Naboo      46BBY    no
Luke      male     Tatooine   19BBY    yes
Leia      female   Alderaan   19BBY    no
Obi-Wan   male     Stewjon    57BBY    yes
Han       male     Corellia   29BBY    no
Palpatine male     Naboo      82BBY    no
R2-D2     unknown  Naboo      33BBY    no
```

Space delimited format

Each value separated by a space " "
Typical file extensions: txt

```
Name Gender Homeworld Born Jedi
Anakin male Tatooine 41.9BBY yes
Amidala female Naboo 46BBY no
Luke male Tatooine 19BBY yes
Leia female Alderaan 19BBY no
Obi-Wan male Stewjon 57BBY yes
Han male Corellia 29BBY no
Palpatine male Naboo 82BBY no
R2-D2 unknown Naboo 33BBY no
```

Comma delimited format

Each value separated by a comma ","
Typical file extensions: csv

```
Name,Gender,Homeworld,Born,Jedi
Anakin,male,Tatooine,41.9BBY,yes
Amidala,female,Naboo,46BBY,no
Luke,male,Tatooine,19BBY,yes
Leia,female,Alderaan,19BBY,no
Obi-Wan,male,Stewjon,57BBY,yes
Han,male,Corellia,29BBY,no
Palpatine,male,Naboo,82BBY,no
R2-D2,unknown,Naboo,33BBY,no
```

Semicolon (european) delimited format

Each value separated by a semicolon "**;**"
Typical file extensions: csv

```
Name;Gender;Homeworld;Born;Jedi
Anakin;male;Tatooine;41.9BBY;yes
Amidala;female;Naboo;46BBY;no
Luke;male;Tatooine;19BBY;yes
Leia;female;Alderaan;19BBY;no
Obi-Wan;male;Stewjon;57BBY;yes
Han;male;Corellia;29BBY;no
Palpatine;male;Naboo;82BBY;no
R2-D2;unknown;Naboo;33BBY;no
```

Other characters as delimiters

Each value separated by a bar "|"

You can make up your own extension: *bar*

```
Name|Gender|Homeworld|Born|Jedi
Anakin|male|Tatooine|41.9BBY|yes
Amidala|female|Naboo|46BBY|no
Luke|male|Tatooine|19BBY|yes
Leia|female|Alderaan|19BBY|no
Obi-Wan|male|Stewjon|57BBY|yes
Han|male|Corellia|29BBY|no
Palpatine|male|Naboo|82BBY|no
R2-D2|unknown|Naboo|33BBY|no
```

CSV files

The Comma-Separated Value (CSV) format is a special case of a plain text format.

Although not a formal standard, CSV files are very common and are a quite reliable plain text delimited format that at least solves the problem of where the fields are in each row of the file.

CSV files are a common way to transfer data from a spreadsheet to other software

What if a comma is part of a field?

Commas in fields

Name	Gender	Homeland	Born
Skywalker, Anakin	male	Tatooine	41.9BBY
Amidala, Padme	female	Naboo	46BBY
Skywalker, Luke	male	Tatooine	19BBY
Organa, Leia	female	Alderaan	19BBY
Kenobi, Obi-Wan	male	Stewjon	57BBY
Solo, Han	male	Corellia	29BBY

Commas within a comma delimited format

Commas that are part of a field must be surrounded by quotes ","

```
Name,Gender,Homeworld,Born,Jedi
"Skywalker, Anakin",male,Tatooine,41.9BBY
"Amidala, Padme",female,Naboo,46BBY
"Skywalker, Luke",male,Tatooine,19BBY
"Organa, Leia",female,Alderaan,19BBY
"Kenobi, Obi-Wan",male,Stewjon,57BBY
"Solo, Han",male,Corellia,29BBY
```

There’s also fixed-width

Fixed-width format

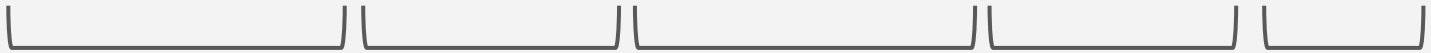
Each value separated with a pre-specified width (in characters)

```
Name      Gender   Homeworld Born     Jedi
Anakin    male     Tatooine  41.9BBY  yes
Amidala   female   Naboo     46BBY    no
Luke      male     Tatooine  19BBY    yes
Leia      female   Alderaan  19BBY    no
Obi-Wan   male     Stewjon   57BBY    yes
Han       male     Corellia  29BBY    no
Palpatine male     Naboo     82BBY    no
R2-D2     unknown  Naboo     33BBY    no
```

Fixed-width format

Each value separated with a pre-specified width (in characters)

```
Name       Gender   Homeworld  Born     Jedi
Anakin     male     Tatooine   41.9BBY  yes
Amidala    female   Naboo      46BBY    no
Luke       male     Tatooine   19BBY    yes
Leia       female   Alderaan   19BBY    no
Obi-Wan    male     Stewjon    57BBY    yes
Han        male     Corellia   29BBY    no
Palpatine  male     Naboo      82BBY    no
R2-D2      unknown  Naboo      33BBY    no
```



10 8 10 8 4

Pros and Cons of data tables

Advantages

Simplicity

Common formats (csv, tsv, txt, dat, *etc*)

Can be opened and modifed with a text editor

Can also be opened in spreadsheet software

Easy to undestand for most users

Can be read in data analysis software

Disadvantages

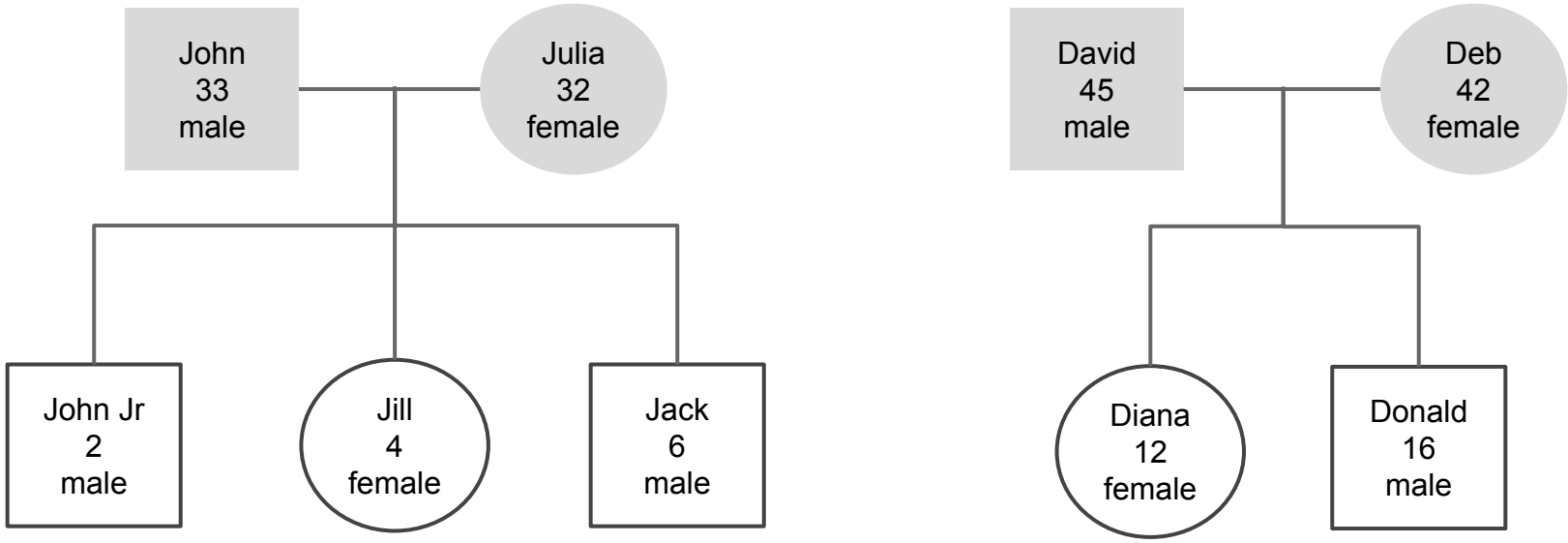
Not good for hierarchical or nested structures

Values may not be well self-described

Difficult to include metadata

Long column names are cumbersome

Hierarchical Data



Father	Mother	Name	Age	Sex
		John	33	male
		Julia	32	female
John	Julia	Jack	6	male
John	Julia	Jill	4	female
John	Julia	John Jr	2	male
		David	45	male
		Debbie	42	female
David	Debbie	Donald	16	male
David	Debbie	Dianne	12	female

HTML Tables

https://en.wikipedia.org/wiki/List_of_world_records_in_swimming

Event	Time		Name	Nationality	Date
50 m freestyle	20.91		César Cielo	Brazil	18 December 2009
100 m freestyle	46.91		César Cielo	Brazil	30 July 2009
200 m freestyle	1:42.00		Paul Biedermann	Germany	28 July 2009
400 m freestyle	3:40.07		Paul Biedermann	Germany	26 July 2009
800 m freestyle	7:32.12		Zhang Lin	China	29 July 2009
1500 m freestyle	14:31.02		Sun Yang	China	4 August 2012
50 m backstroke	24.04		Liam Tancock	Great Britain	2 August 2009
100 m backstroke	51.85	r	Ryan Murphy	United States	13 August 2016
200 m backstroke	1:51.92		Aaron Peirsol	United States	31 July 2009
50 m breaststroke	26.42	sf	Adam Peaty	Great Britain	4 August 2015
100 m breaststroke	57.13		Adam Peaty	Great Britain	7 August 2016
200 m breaststroke	2:07.01		Akihiro Yamaguchi	Japan	15 September 2012
50 m butterfly	22.43	sf	Rafael Muñoz	Spain	5 April 2009
100 m butterfly	49.82		Michael Phelps	United States	1 August 2009
200 m butterfly	1:51.51		Michael Phelps	United States	29 July 2009
200 m individual medley	1:54.00		Ryan Lochte	United States	28 July 2011
400 m individual medley	4:03.84		Michael Phelps	United States	10 August 2008

Organizing Data in Spreadsheets

If you decide to work with spreadsheets software to organize and prepare your data tables, take a look at some excellent recommendations by Karl Broman:

http://kbroman.org/dataorg/