Spreadsheets and Databases

Will Leeson

Spreadsheets

Documents for organizing data in rows and columns. This alignment can make it easy to give the computer processing instructions - e.g. sum this column. Data is in a cell. Cells can be identified by column and row.

a/1	b	c	d	e
2				
3		c3		
4				
5				

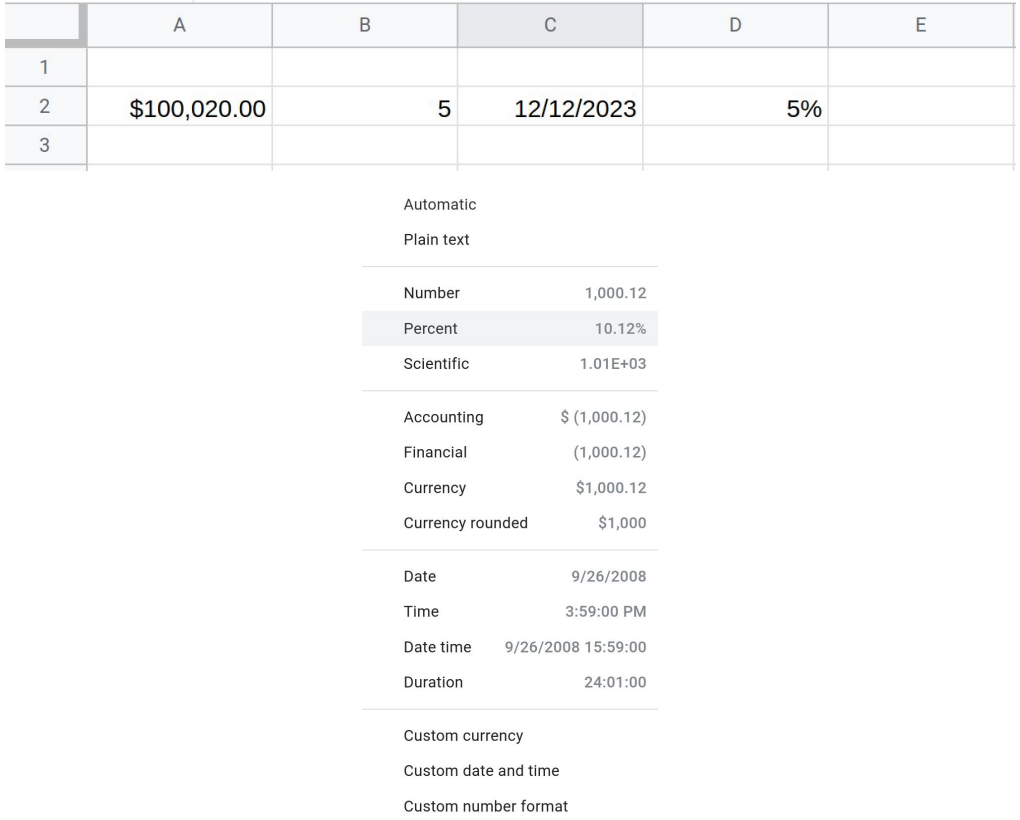
Slide courtesy of Derrick Stone

Spreadsheet software

- Libre Office Calc (Free)
- Microsoft Excel (Purchase/Subscription)
- Google Sheets (Free with Google Account)
- Mac Numbers (Free with Mac purchase)

Cells

- Building block of a spreadsheet
- Can hold different types of data
  - Raw numbers
  - Dates
  - Percents
  - Text
  - Currency
- Can hold “formulas”



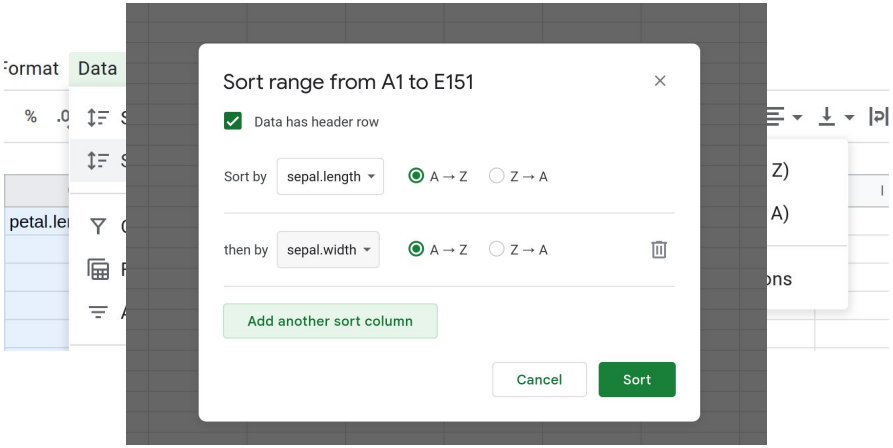
Rows and Columns

- Technically, they are interchangeable
- However, we live in a society
- Columns should be headings
  - Categories
- Rows
  - Values for these categories

	A	B	C	D	E
1	sepal.length	sepal.width	petal.length	petal.width	variety
2	5	2	3.5	1	Versicolor
3	6	2.2	4	1	Versicolor
4	6.2	2.2	4.5	1.5	Versicolor
5	6	2.2	5	1.5	Virginica
6	4.5	2.3	1.3	0.3	Setosa
7	5.5	2.3	4	1.3	Versicolor
8	6.3	2.3	4.4	1.3	Versicolor
9	5	2.3	3.3	1	Versicolor
10	4.9	2.4	3.3	1	Versicolor
11	5.5	2.4	3.8	1.1	Versicolor
12	5.5	2.4	3.7	1	Versicolor
13	5.6	2.5	3.9	1.1	Versicolor
14	6.3	2.5	4.9	1.5	Versicolor
15	5.5	2.5	4	1.3	Versicolor
16	5.1	2.5	3	1.1	Versicolor
17	4.9	2.5	4.5	1.7	Virginica
18	6.7	2.5	5.8	1.8	Virginica
19	5.7	2.5	5	2	Virginica
20	6.3	2.5	5	1.9	Virginica

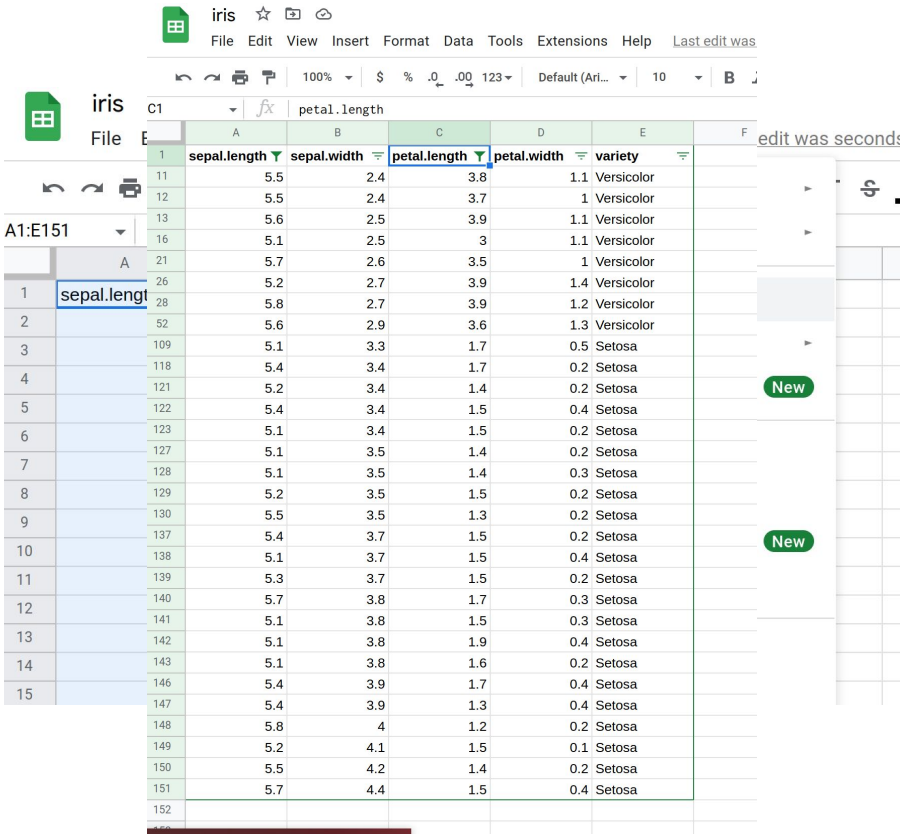
Sorting Data

- Sorting Data can be beneficial
  - See trends
  - Get quick info
- Sort by heading or column
- Can sort by multiple
  - Sort by Column A first
  - If there is a tie, sort by Column C,
  - Etc.



Filtering Data

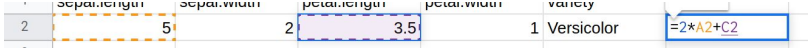
- Isolates specific sections of the data
- Filters are boolean formulas
  - X = (sepal.length > 5)
  - Y = (petal.length < 4)
  - X ∧ Y

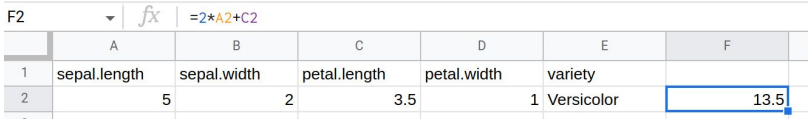


	sepal.length	sepal.width	petal.length	petal.width	variety
11	5.5	2.4	3.8	1.1	Versicolor
12	5.5	2.4	3.7	1	Versicolor
13	5.6	2.5	3.9	1.1	Versicolor
16	5.1	2.5	3	1.1	Versicolor
21	5.7	2.6	3.5	1	Versicolor
26	5.2	2.7	3.9	1.4	Versicolor
28	5.8	2.7	3.9	1.2	Versicolor
52	5.6	2.9	3.6	1.3	Versicolor
109	5.1	3.3	1.7	0.5	Setosa
118	5.4	3.4	1.7	0.2	Setosa
121	5.2	3.4	1.4	0.2	Setosa
122	5.4	3.4	1.5	0.4	Setosa
123	5.1	3.4	1.5	0.2	Setosa
127	5.1	3.5	1.4	0.2	Setosa
128	5.1	3.5	1.4	0.3	Setosa
129	5.2	3.5	1.5	0.2	Setosa
130	5.5	3.5	1.3	0.2	Setosa
137	5.4	3.7	1.5	0.2	Setosa
138	5.1	3.7	1.5	0.4	Setosa
139	5.3	3.7	1.5	0.2	Setosa
140	5.7	3.8	1.7	0.3	Setosa
141	5.1	3.8	1.5	0.3	Setosa
142	5.1	3.8	1.9	0.4	Setosa
143	5.1	3.8	1.6	0.2	Setosa
146	5.4	3.9	1.7	0.4	Setosa
147	5.4	3.9	1.3	0.4	Setosa
148	5.8	4	1.2	0.2	Setosa
149	5.2	4.1	1.5	0.1	Setosa
150	5.5	4.2	1.4	0.2	Setosa
151	5.7	4.4	1.5	0.4	Setosa

Formula Basics

- A formula defines how a cell will be calculated
- Begins with “=”
  - This tells software “I’m a formula”
- Made up of some combination of:
  - Hardcoded values (42, -2, 3.1415, etc.)
  - Cell references (A1, B12, etc.)
  - Functions (+, -, Average(), etc.)





Keyboard shortcuts

- Ctrl + (Key) may do something helpful
  - It may be Cmd + (Key) if you have a Mac
  - Ctrl + (Key) means click Ctrl and then (Key), don’t click the + key
  - Unless it's Ctrl + +
- Ctrl + C = Copies the selected thing
- Ctrl + V = Pastes the selected thing
- Ctrl + X = Cuts the selected thing
- Ctrl + Z = Undoes last action
  - Can be pressed multiple times
- Ctrl + Shift + Z = Redoes last undid thing
- Ctrl + D = Copies cell down selected rows

Repeating Formulas

F2 =2*A2+C2

	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	13.5
3	6	2.2	4	1	Versicolor	
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	

F2 is selected

Repeating Formulas

F2 | fx | =2*A2+C2

	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	13.5
3	6	2.2	4	1	Versicolor	
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	

Ctrl + C copies it

Dashed border shows its selected by clipboard

Repeating Formulas

F3 | fx | =2*A3+C3

	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	13.5
3	6	2.2	4	1	Versicolor	16
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	

Ctrl + V pastes it

Repeating Formulas

F3:F5 | fx | =2*A3+C3

	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	13.5
3	6	2.2	4	1	Versicolor	16
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	

Select F3, hold shift, hit down twice (or click F5)

Now F3 through F5 is selected

Repeating Formulas

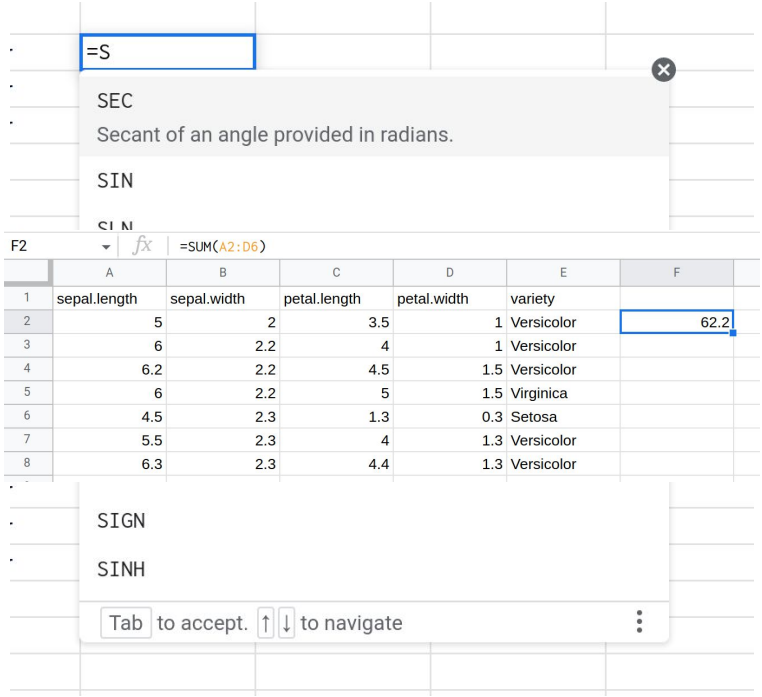
F3:F5 | fx | =2*A3+C3

	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	13.5
3	6	2.2	4	1	Versicolor	16
4	6.2	2.2	4.5	1.5	Versicolor	16.9
5	6	2.2	5	1.5	Virginica	17

Ctrl + D copies formula from F3 to selected cells

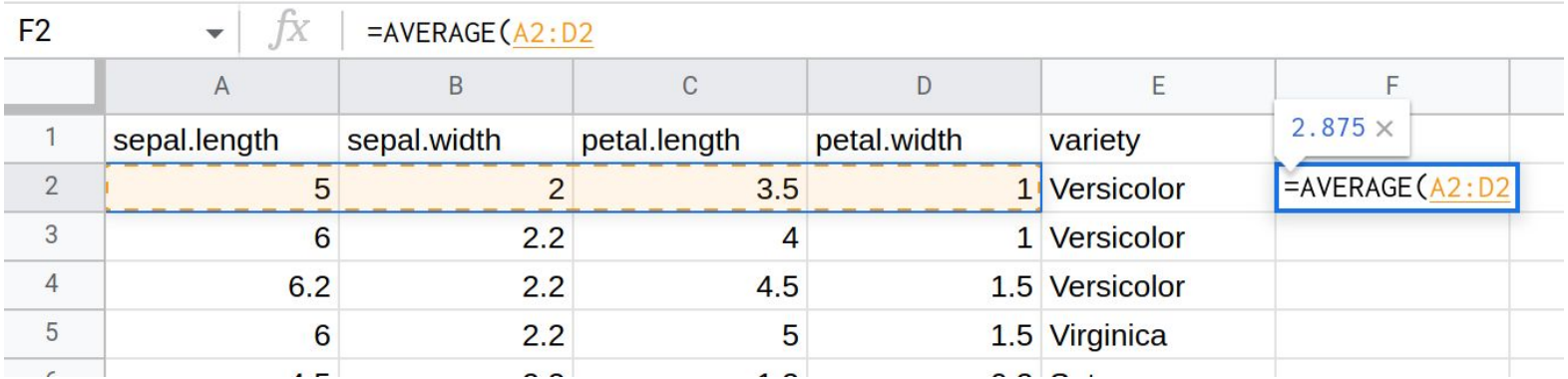
Formulas over Ranges

- Some formulas operate over ranges
  - Sum, Average, Min, Max, etc.
- Ranges can be
  - Values across a Row (A1:A10)
  - Values across a Column (A1:F1)
  - Both (A1:F10)
  - A set of cells (A1, B3, F7)



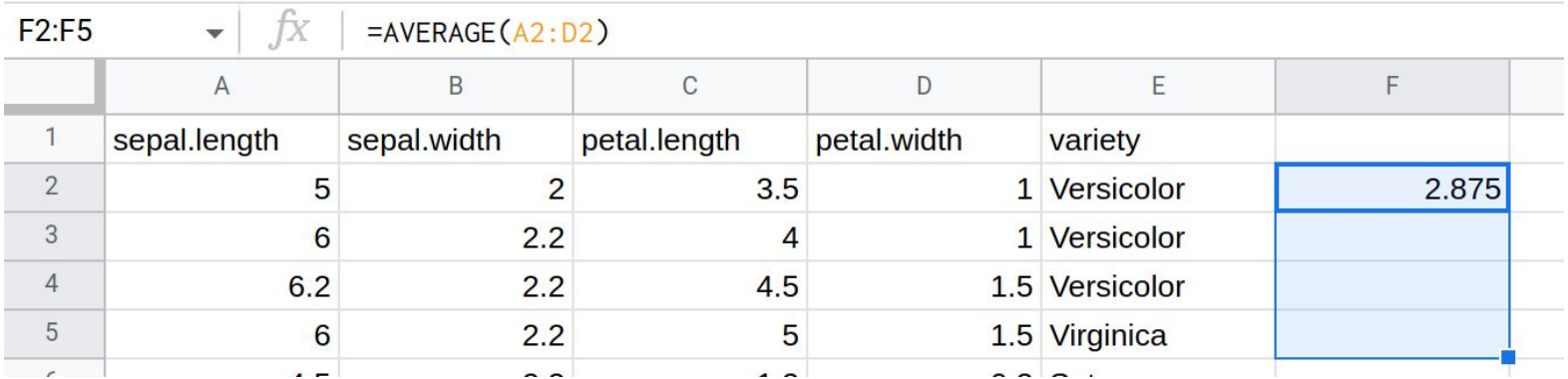
	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	62.2
3	6	2.2	4	1	Versicolor	
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	
6	4.5	2.3	1.3	0.3	Setosa	
7	5.5	2.3	4	1.3	Versicolor	
8	6.3	2.3	4.4	1.3	Versicolor	

Repeating Formulas Over Ranges



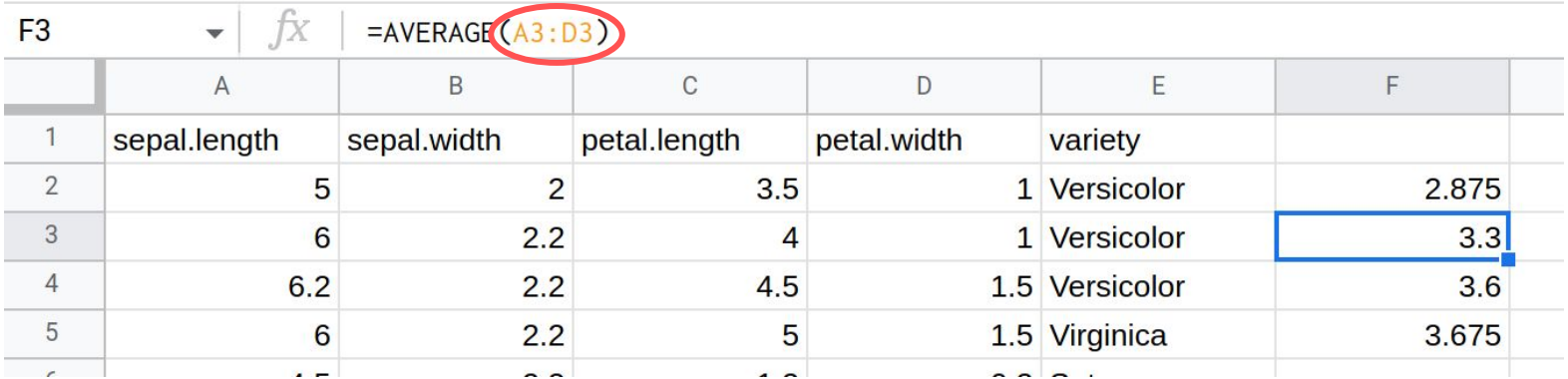
	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	
3	6	2.2	4	1	Versicolor	
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	

Repeating Formulas Over Ranges



	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	2.875
3	6	2.2	4	1	Versicolor	
4	6.2	2.2	4.5	1.5	Versicolor	
5	6	2.2	5	1.5	Virginica	

Repeating Formulas Over Ranges



	A	B	C	D	E	F
1	sepal.length	sepal.width	petal.length	petal.width	variety	
2	5	2	3.5	1	Versicolor	2.875
3	6	2.2	4	1	Versicolor	3.3
4	6.2	2.2	4.5	1.5	Versicolor	3.6
5	6	2.2	5	1.5	Virginica	3.675

Repeating Formulas Over Ranges

F2 `=AVERAGEIF(E$2:E$151, E2,A$2:A$151)`

	A	B	C	D	E	F	G	H
1	sepal.length	sepal.width	petal.length	petal.width	variety			
2	5	2	3.5	1	Versicolor	=AVERAGEIF(E$2:E$151, E2,A$2:A$151)		
3	6	2.2	4	1	Versicolor			
4	6.2	2.2	4.5	1.5	Versicolor			
5	6	2.2	5	1.5	Virginica			
6	4.5	2.3	1.3	0.3	Setosa			
7	5.5	2.3	4	1.3	Versicolor			
8	6.3	2.3	4.4	1.3	Versicolor			

Repeating Formulas Over Ranges

F2 | fx | =AVERAGEIF(E$2:E$151, E2,A$2:A$151)

	A	B	C	D	E	F	G	H
1	sepal.length	sepal.width	petal.length	petal.width	variety			
2	5	2	3.5	1	Versicolor	5.936		
3	6	2.2	4	1	Versicolor			
4	6.2	2.2	4.5	1.5	Versicolor			
5	6	2.2	5	1.5	Virginica			
6	4.5	2.3	1.3	0.3	Setosa			
7	5.5	2.3	4	1.3	Versicolor			
8	6.3	2.3	4.4	1.3	Versicolor			

Repeating Formulas Over Ranges

F5 | fx | =AVERAGEIF(E$2:E$151, E5, A$2:A$151)

	A	B	C	D	E	F	G	H
1	sepal.length	sepal.width	petal.length	petal.width	variety			
2	5	2	3.5	1	Versicolor	5.936		
3	6	2.2	4	1	Versicolor			
4	6.2	2.2	4.5	1.5	Versicolor			
5	6	2.2	5	1.5	Virginica	6.588		
6	4.5	2.3	1.3	0.3	Setosa			
7	5.5	2.3	4	1.3	Versicolor			
8	6.3	2.3	4.4	1.3	Versicolor			

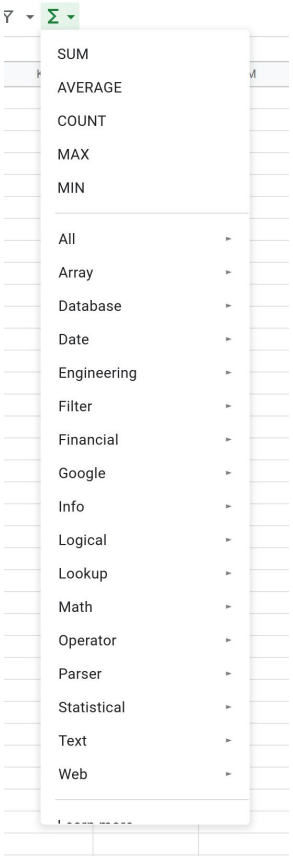
Repeating Formulas Over Ranges

F6 | fx | =AVERAGEIF(E$2:E$151, E6, A$2:A$151)

	A	B	C	D	E	F	G	H
1	sepal.length	sepal.width	petal.length	petal.width	variety			
2	5	2	3.5	1	Versicolor	5.936		
3	6	2.2	4	1	Versicolor			
4	6.2	2.2	4.5	1.5	Versicolor			
5	6	2.2	5	1.5	Virginica	6.588		
6	4.5	2.3	1.3	0.3	Setosa	5.006		
7	5.5	2.3	4	1.3	Versicolor			
8	6.3	2.3	4.4	1.3	Versicolor			

Some Helpful functions

- Logical Operators
  - And, Or, ExOR
  - Can be helpful with other functions
- If Operators
  - COUNT**IF**(s), SUM**IF**, MAX**IF**
  - Using logical operators, can selectively apply
- The list goes on
  - Use Google to find a function that matches your needs
  - “Excel find cell matching value”



Databases

Database

Definition - Database

An organized collection of data stored and accessed electronically

Disclaimer:

We are talking about “relational databases”

Spreadsheet vs Database

Conceptually no difference.

Slide courtesy of Derrick Stone

Spreadsheets vs Databases

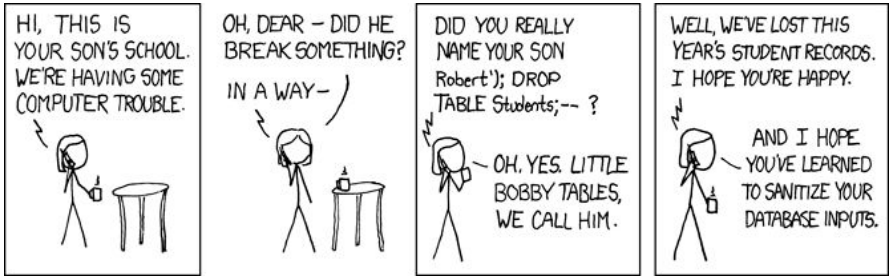
- Single user
- Single file
- Library of common functions
- Row and column limits low (relatively)

- Multiple users, as well as possible use by external applications
- Multiple files… can include files as an element!
- Common functions, plus a library of programmable functions, plus a built-in processing oriented programming language
- Much larger data within a particular cell, many more possible rows of data, binary files
- Tools provided for more complex analysis
- Space for large amounts of data is used more efficiently

Slide courtesy of Derrick Stone

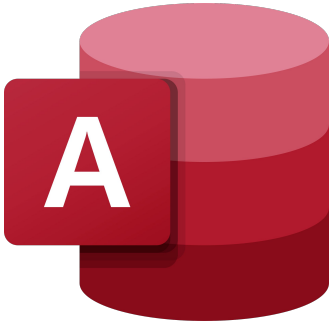
Tables

- Databases consist of “Tables”
- Roughly equivalent to a spreadsheet
- Tables consist of
  - Columns/Fields - The header
    - Name
    - Date of birth
    - Location of Birth
  - Row - Individual entry in database



Database Management System (DBMS)

- Provides a way to
  - Store
  - Access
  - Update
- Support multiple users at once
- Support authorization levels
- Support distributed storage
- Enforce a “schema”
  - The way in which tables are connected



Database setup

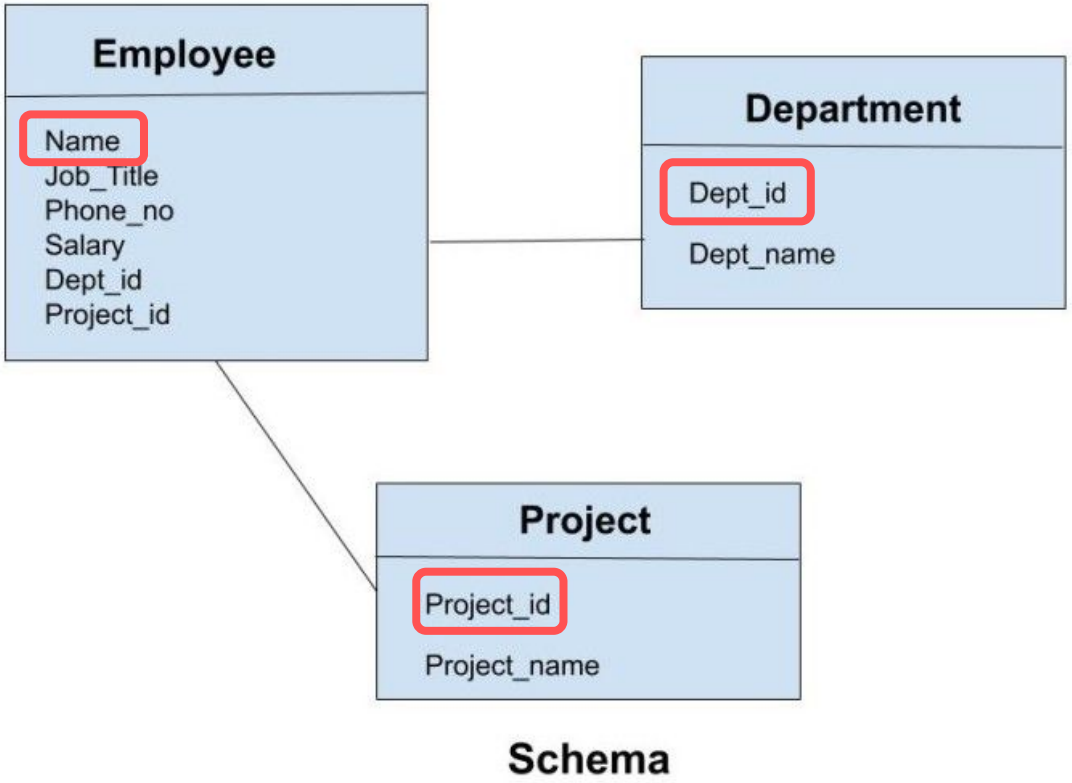
- Requires an understanding of what the database is
  - What information is being stored
  - What kind of data is stored
    - Name (Text)
    - Salary (Number)
    - Date (Formatted Number)
  - Can add constraints to fields
    - Is is optional?
    - Should we sort it based on this field?
  - Each table should have a “primary key”

Primary Key

- A field used to identify specific data points
  - Typically not directly tied to data
- Allows connecting different tables
- Primary Key can differ (and probably should) between tables

ID	First Name	Last Name	Date of Birth
1	Michael	Jordan	2/17/1963
2	Michael	Jordan	2/9/1987

Schema Example



How do we manipulate the database?

Query Languages

- Domain-Specific Language (DSL)
- “Query” the database
  - Please database, do X for me
- Defines how to
  - Update database
  - Retrieve information from database
- Common examples
  - SQL (Structured Query Language)
  - XQuery (XML Query)

Retrieving Data

`SELECT <columns>`	5.
`FROM <table>`	1.
`WHERE <predicate on rows>`	2.
`GROUP BY <columns>`	3.
`HAVING <predicate on groups>`	4.
`ORDER BY <columns>`	6.
`OFFSET`	7.
`FETCH FIRST`	8.

Retrieving Data

```
SELECT *
 FROM  Book
 WHERE price > 100.00
 ORDER BY title;
```

Retrieving Data (Subqueries)

```
SELECT isbn,
       title,
       price
 FROM  Book
 WHERE price < (SELECT AVG(price) FROM Book)
 ORDER BY title;
```

Updating the Database

```
CREATE TABLE example(
 column1 INTEGER,
 column2 VARCHAR(50),
 column3 DATE NOT NULL,
 PRIMARY KEY (column1, column2)
);
```

Updating the Database

```
INSERT INTO example
 (column1, column2, column3)
 VALUES
 ('test', 'N', NULL);
```

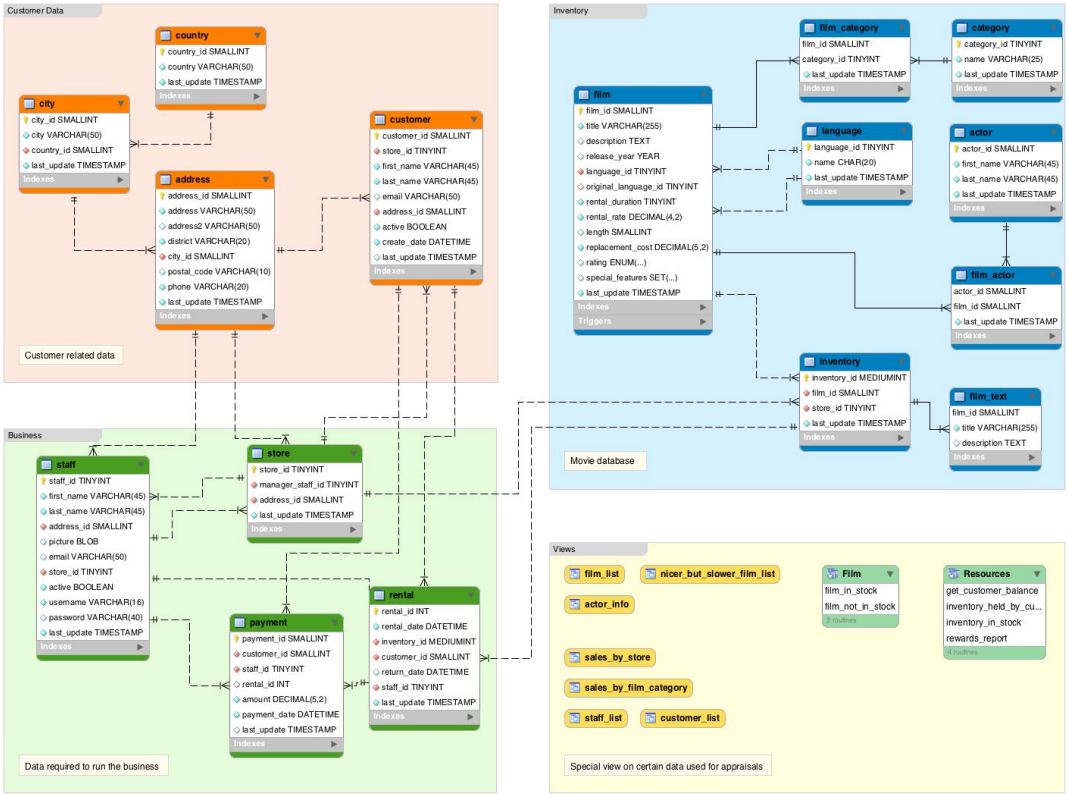
Updating the Database

```
UPDATE example
 SET column1 = 'updated value'
 WHERE column2 = 'N';
```

Updating the Database

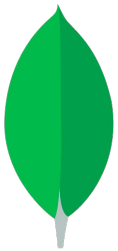
```
DELETE FROM example
 WHERE column2 = 'N';
```

Databases get complicated



NoSQL

- “Non-SQL”
- Several Types
  - Key-Value
  - Document Storage
  - Graph
- Sacrifice structure for flexibility
- Typically scales better
  - Speed
  - Space



mongoDB®

Gremlin

$G = (V, E)$