MT5125 MSc in Business Administration

2020/2021

Strat. Consultancy Project I & Data Analytics

DCU BUSINESS SCHOOL

Dr. Damien Dupré
damien.dupre@dcu.ie
Assistant Professor - Business Research Methods

dotLAB
The Irish Institute of Digital Business

1. Data Storage and Access

DCU BUSINESS SCHOOL

dotLAB
The Irish Institute of Digital Business

Big Data Everywhere!

BIG DATA

Data that is TOO LARGE & TOO COMPLEX for conventional data tools to capture, store and analyze.

The 3V's of Big Data

VOLUME

VARIETY

VELOCITY

Shares traded on US Stock Markets each day:

7 Billion

Data generated in one flight from NY to London:

10 Terabytes

Number of tweets per day on Twitter:

400 Million

Number of 'Likes' each day on Facebook:

3 Billion

90%

OF THE WORLD'S DATA WAS GENERATED IN THE **LAST TWO YEARS**

What is Big Data

- Too large or too complex to be handled by conventional tools
- Microsoft Excel’s Limits (current version)
  - Total number of rows: 1,048,576 rows
  - Total number columns: 16,384 columns

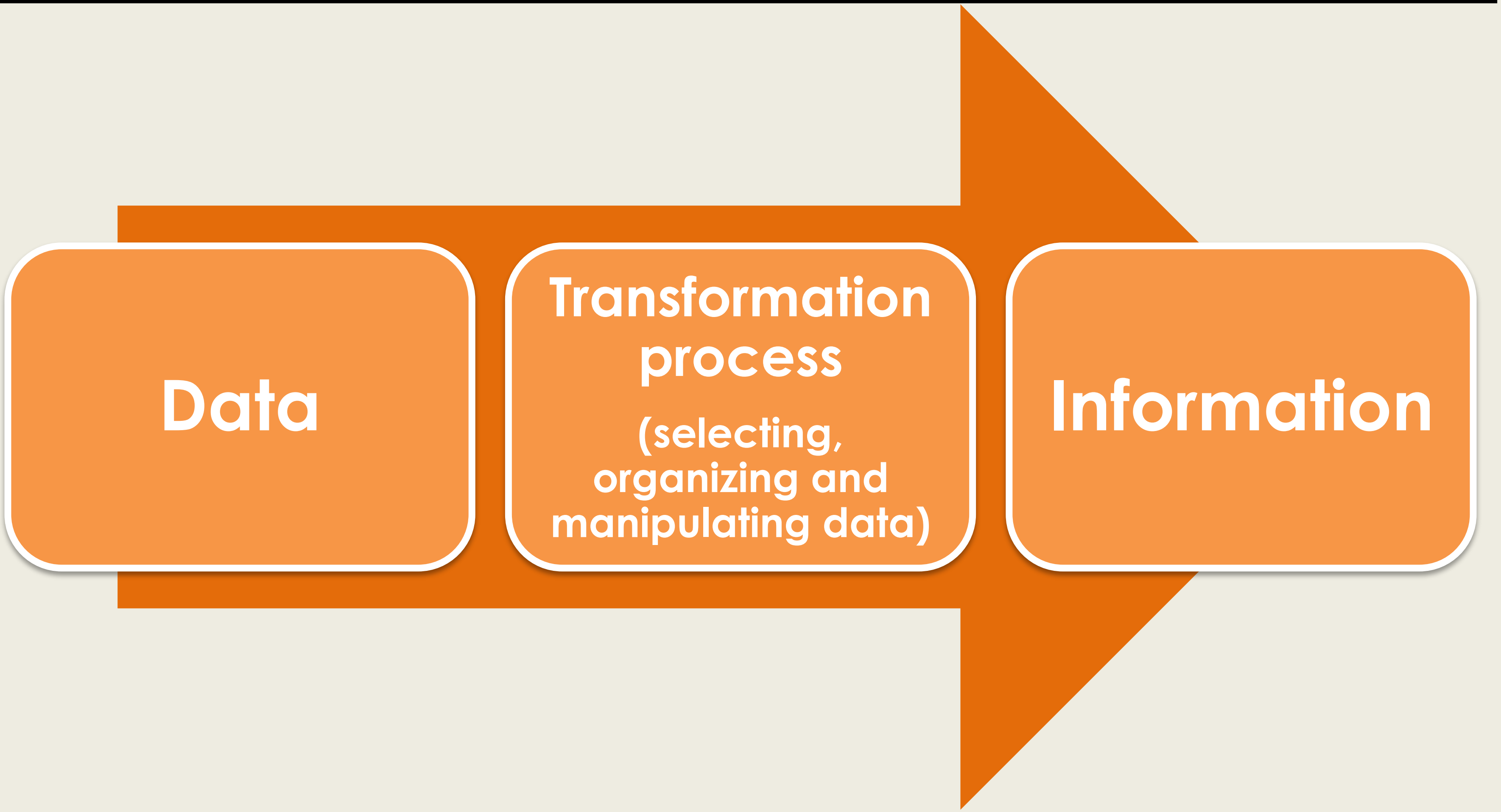
Data vs. Information (1)

- Without data an organization could not successfully complete most business activities
- **Data** consists of **raw facts**
- **Information** is one of an organisation's most valuable resources
- Often confused with the term **data**
- To transform **data** into useful **information**

Data vs. Information (2)

- Example: Sales Manager
  - Knowing number of sales for each representative
    - (fact – data)
  - Knowing total monthly sales
    - (transformed – information)

Data vs. Information (3)



Data

Transformation process

(selecting, organizing and manipulating data)

Information

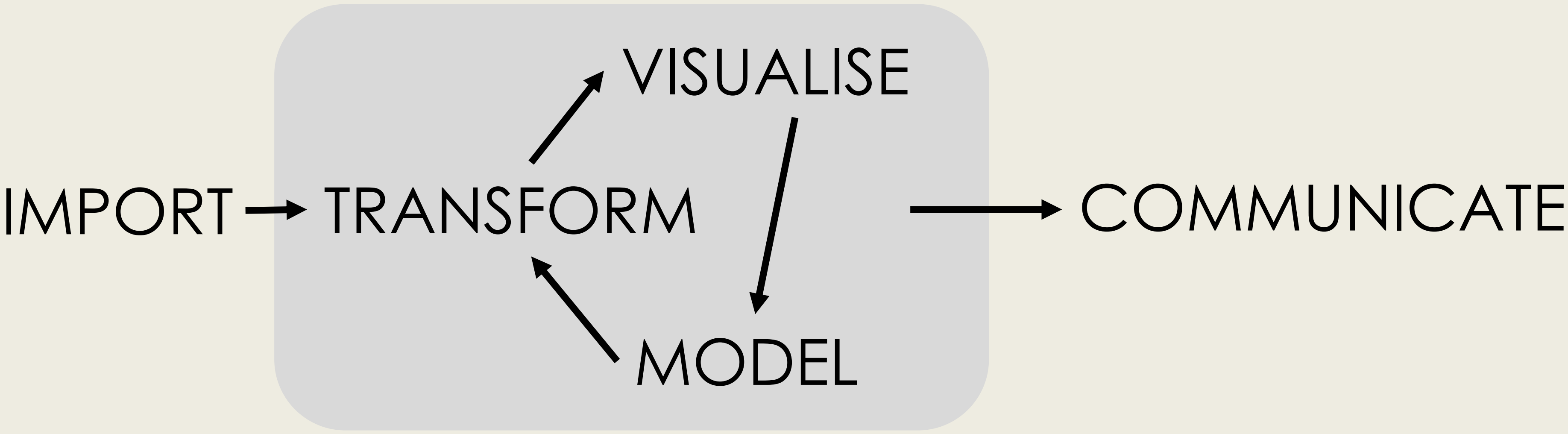
Value of Information

- Goals
  - Helps decision makers achieve organisational goals
- Performance
  - Valuable information helps people and organisations perform
- Accuracy
  - Inaccurate/Incomplete information leads to Poor Decisions and can result in High Cost for the organisation

Data Analytics

- The science of using data to build models that lead to better decisions that in turn add value to individuals, companies and institutions
- The analysis of data, typically large sets of data, by the use of mathematics, statistics, and computer software

Data Analytics Tasks



IMPORT

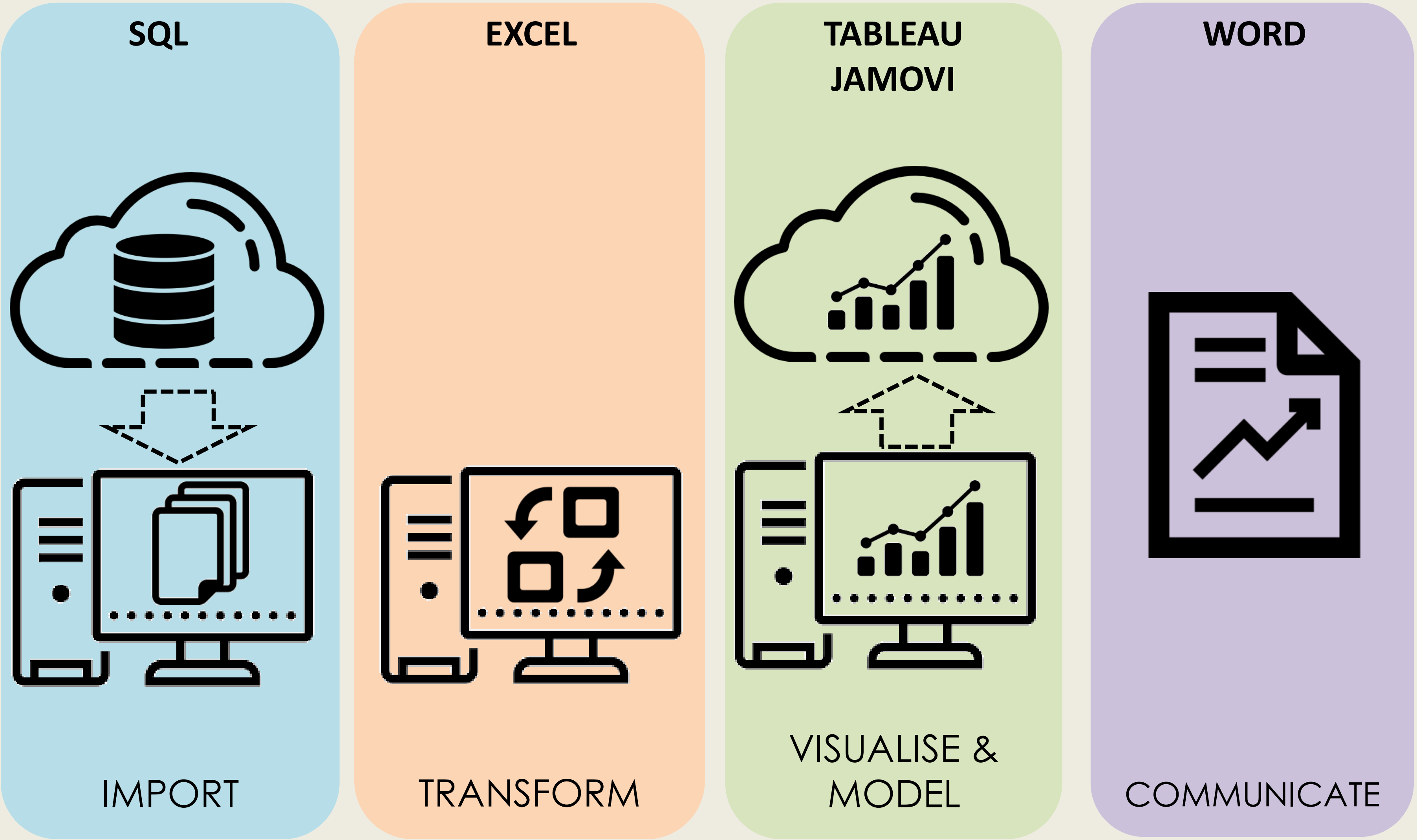
TRANSFORM

VISUALISE

MODEL

COMMUNICATE

Data Analytics Tools



SQL

IMPORT

EXCEL

TRANSFORM

TABLEAU
JAMOVI

VISUALISE &
MODEL

WORD

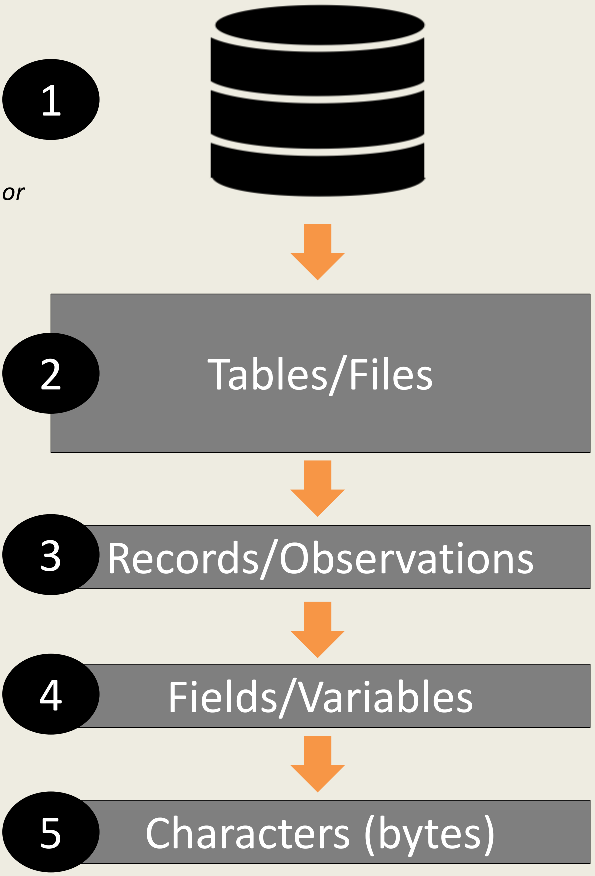
COMMUNICATE

DATA STORAGE

The Hierarchy of Data

Levels

Database



1

A database contains one or more Tables/Files

2 Tables/Files

A Table/File contains a number of records/observations

3 Records/Observations

A record/observation contains a number of fields/variable

4 Fields/Variables

5 Characters (bytes)

Example

Personnel file
Department file
Payroll file

(Project database)

098 - 40 - 1370 Fiske, Steven 01-05-1985
549 - 77 - 1001 Buckley, Bill 02-17-1979
005 - 10 - 6321 Johns, Francine 10-07-1997

(Personnel file)

098 - 40 - 1370 Fiske, Steven 01-05-1985

(Record containing SSN, last and first name, hire date)

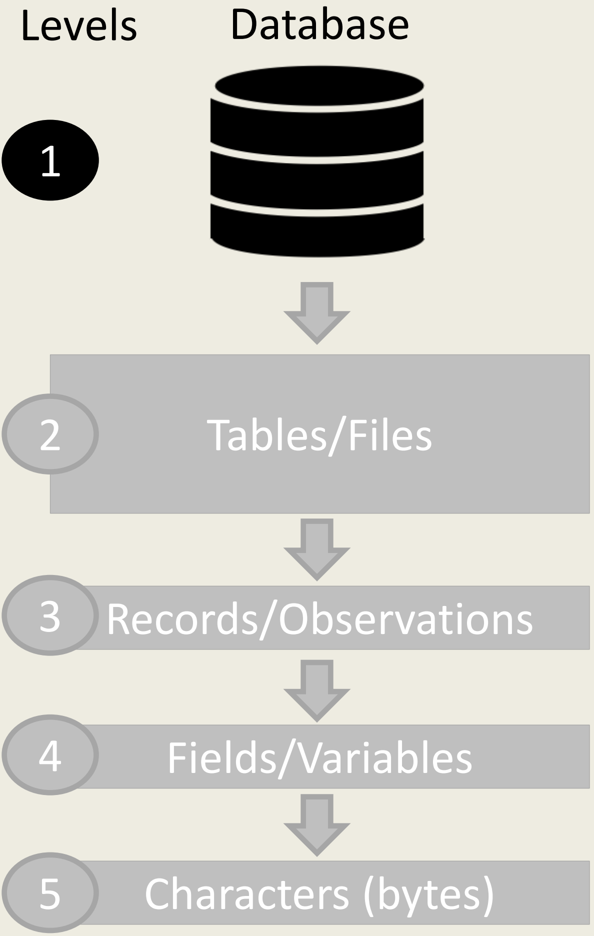
Fiske

(Last name field)

1000110

(Letter F in ASCII)

Database Access



Levels Database

1

2 Tables/Files

3 Records/Observations

4 Fields/Variables

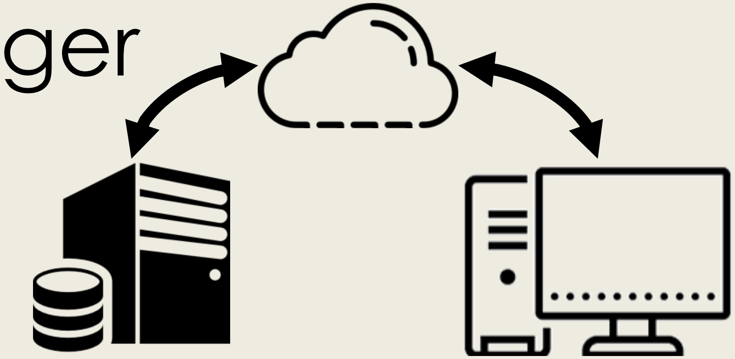
5 Characters (bytes)

- When building a database, organizations must consider:
  - ***Content:*** What data should be collected and at what cost?
  - ***Access:*** What data should be provided to which users and when?
  - ***Logical structure:*** How should data be arranged so that it makes sense to a given user?

Access to Databases

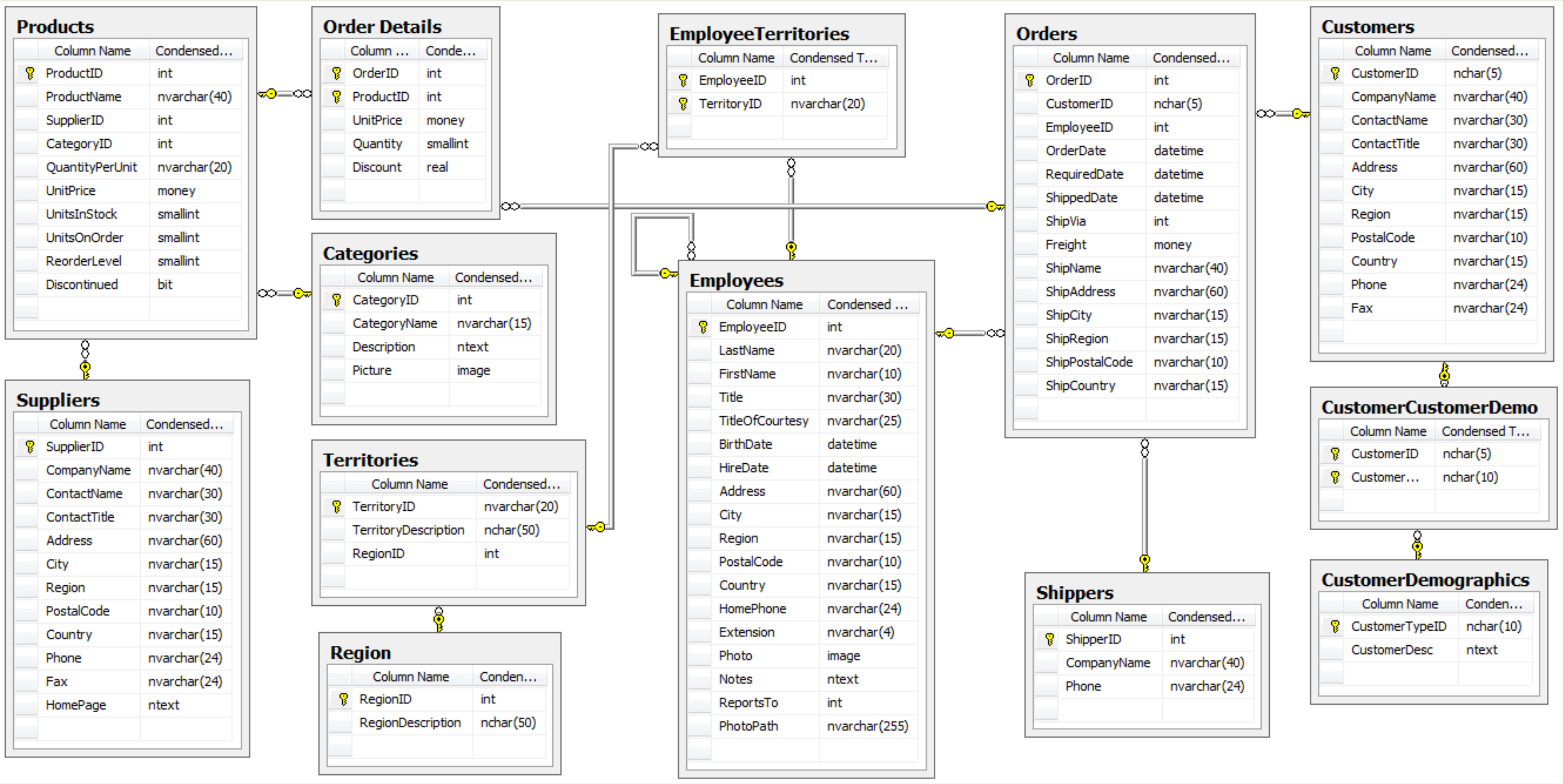
Allows dozens or hundreds of people to access the same database system at the same time:

- Hosted on a remote “cloud” server (usually)
- Database client manager
  - Host address (IP or URL)
  - Guest Access
    - TCP Port (e.g. 5432)
    - Login/password
- Gateway for another tool



Relational Database

Northwind PostgreSQL Database



Products

	Column Name	Condensed...
🔑	ProductID	int
	ProductName	nvarchar(40)
	SupplierID	int
	CategoryID	int
	QuantityPerUnit	nvarchar(20)
	UnitPrice	money
	UnitsInStock	smallint
	UnitsOnOrder	smallint
	ReorderLevel	smallint
	Discontinued	bit

Order Details

	Column ...	Conde...
🔑	OrderID	int
🔑	ProductID	int
	UnitPrice	money
	Quantity	smallint
	Discount	real

EmployeeTerritories

	Column Name	Condensed T...
🔑	EmployeeID	int
🔑	TerritoryID	nvarchar(20)

Orders

	Column Name	Condensed...
🔑	OrderID	int
	CustomerID	nchar(5)
	EmployeeID	int
	OrderDate	datetime
	RequiredDate	datetime
	ShippedDate	datetime
	ShipVia	int
	Freight	money
	ShipName	nvarchar(40)
	ShipAddress	nvarchar(60)
	ShipCity	nvarchar(15)
	ShipRegion	nvarchar(15)
	ShipPostalCode	nvarchar(10)
	ShipCountry	nvarchar(15)

Customers

	Column Name	Condensed...
🔑	CustomerID	nchar(5)
	CompanyName	nvarchar(40)
	ContactName	nvarchar(30)
	ContactTitle	nvarchar(30)
	Address	nvarchar(60)
	City	nvarchar(15)
	Region	nvarchar(15)
	PostalCode	nvarchar(10)
	Country	nvarchar(15)
	Phone	nvarchar(24)
	Fax	nvarchar(24)

Categories

	Column Name	Condensed...
🔑	CategoryID	int
	CategoryName	nvarchar(15)
	Description	ntext
	Picture	image

Suppliers

	Column Name	Condensed...
🔑	SupplierID	int
	CompanyName	nvarchar(40)
	ContactName	nvarchar(30)
	ContactTitle	nvarchar(30)
	Address	nvarchar(60)
	City	nvarchar(15)
	Region	nvarchar(15)
	PostalCode	nvarchar(10)
	Country	nvarchar(15)
	Phone	nvarchar(24)
	Fax	nvarchar(24)
	HomePage	ntext

Territories

	Column Name	Condensed...
🔑	TerritoryID	nvarchar(20)
	TerritoryDescription	nchar(50)
	RegionID	int

Region

	Column Name	Conden...
🔑	RegionID	int
	RegionDescription	nchar(50)

Employees

	Column Name	Condensed ...
🔑	EmployeeID	int
	LastName	nvarchar(20)
	FirstName	nvarchar(10)
	Title	nvarchar(30)
	TitleOfCourtesy	nvarchar(25)
	BirthDate	datetime
	HireDate	datetime
	Address	nvarchar(60)
	City	nvarchar(15)
	Region	nvarchar(15)
	PostalCode	nvarchar(10)
	Country	nvarchar(15)
	HomePhone	nvarchar(24)
	Extension	nvarchar(4)
	Photo	image
	Notes	ntext
	ReportsTo	int
	PhotoPath	nvarchar(255)

CustomerCustomerDemo

	Column Name	Condensed T...
🔑	CustomerID	nchar(5)
🔑	Customer...	nchar(10)

Shippers

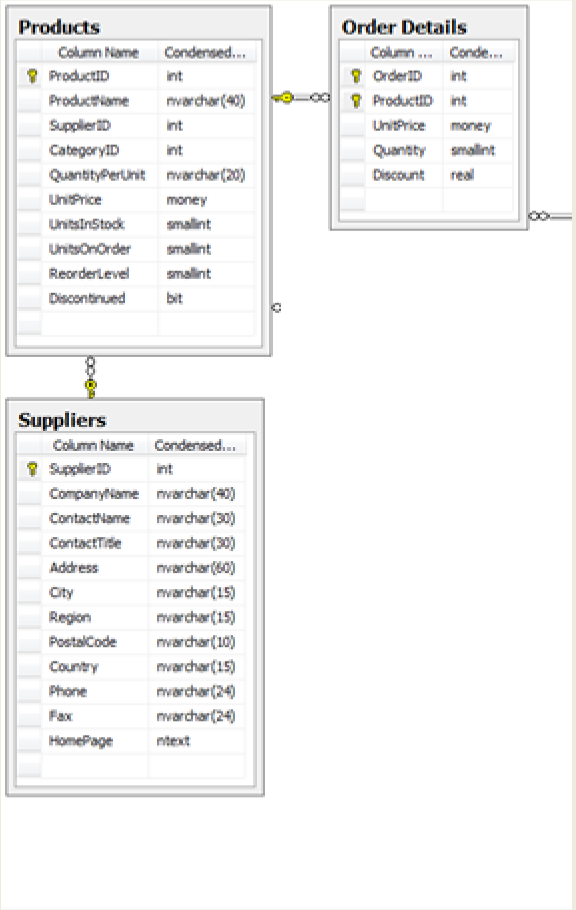
	Column Name	Condensed...
🔑	ShipperID	int
	CompanyName	nvarchar(40)
	Phone	nvarchar(24)

CustomerDemographics

	Column Name	Conden...
🔑	CustomerTypeID	nchar(10)
	CustomerDesc	ntext

The Relational Principle

- Every product gets ONE record in the Products table
- Every supplier gets ONE record in the Suppliers table
- Rows in different tables can be related to another using a shared key
- There can be multiple product records for a given supplier



Products

	Column Name	Condensed...
🔑	ProductID	int
	ProductName	nvarchar(40)
	SupplierID	int
	CategoryID	int
	QuantityPerUnit	nvarchar(20)
	UnitPrice	money
	UnitsInStock	smallint
	UnitsOnOrder	smallint
	ReorderLevel	smallint
	Discontinued	bit

Order Details

	Column ...	Conde...
🔑	OrderID	int
🔑	ProductID	int
	UnitPrice	money
	Quantity	smallint
	Discount	real

Suppliers

	Column Name	Condensed...
🔑	SupplierID	int
	CompanyName	nvarchar(40)
	ContactName	nvarchar(30)
	ContactTitle	nvarchar(30)
	Address	nvarchar(60)
	City	nvarchar(15)
	Region	nvarchar(15)
	PostalCode	nvarchar(10)
	Country	nvarchar(15)
	Phone	nvarchar(24)
	Fax	nvarchar(24)
	HomePage	ntext

The Relational Principle

- Related records can be found using a shared key
  - Shared key = identifier that is:
    - Unique to each table
    - Can be referenced by another table
  - E.g.,

    Products.ProductID = Order Details.ProductID

Database Schema

- The schema describes all tables/files and all fields/variables
  - Describes relationship between tables
  - Crucial in enabling retrieval of desired data
- Very important
  - Must understand schema for accurate querying
  - Wrong understanding = wrong results

Database Queries

- Database can be made of millions/billions data spread on hundreds of tables/files
- A Query is a set of instructions to retrieve, sort and format returning data
  - E.g., "find me all customers in my database"
  - Query = extracting information out of the database and process them into something (e.g., MS Excel)

Query with SQL

- Structured Query Language
- Way to obtain ONE file with the information you need ONLY
- This is the main SQL statement you need to understand for querying:

```
SELECT *
FROM table_name;
```

Translation: “Show me the data from all the fields from the table ‘table_name’”

Basic Syntax of SQL SELECT

```
SELECT field_name_1, field_name_2
FROM table_name;
```

- Show me the data from the fields ‘field_name_1’ and ‘field_name_2’ from the table ‘table_name’

- Example:

```
SELECT ProductID, ProductName
FROM Products;
```

Basic Syntax of SQL SELECT

```
SELECT field_name_1, field_name_2
FROM table_name
WHERE field_name_1 = ‘X’;
```

- Show me the data from the fields ‘field_name_1’ and ‘field_name_2’ from the table ‘table_name’ corresponding to ‘X’ in the field ‘field_name_1’

- Example:

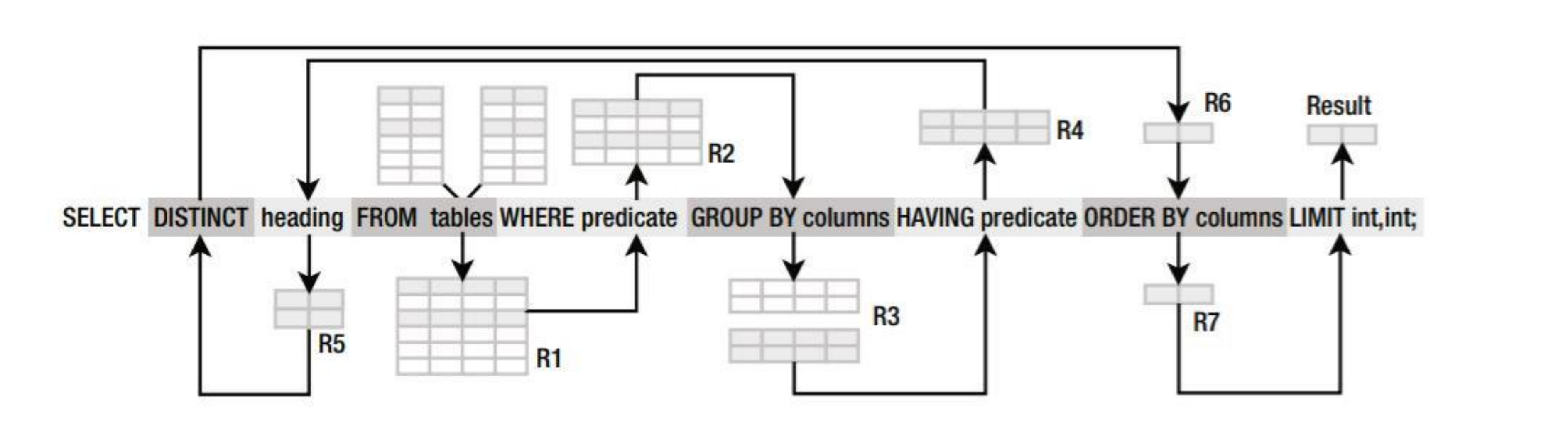
```
SELECT ProductID, ProductName
FROM Products
WHERE ProductName = ‘macbook’;
```

More Possibilities with SELECT

```
SELECT [DISTINCT|COUNT] field_names
FROM table_name_1
[WHERE conditions]
[GROUP BY field_name]
[ORDER BY field_name]
[LEFT|RIGHT|INNER JOIN table_name_2]
[ON table_name_1.field_name = table_name_2.field_name];
```

- More: https://beginner-sql-tutorial.com/sql-commands.htm

More Possibilities with SELECT



```
SELECT DISTINCT heading FROM tables WHERE predicate GROUP BY columns HAVING predicate ORDER BY columns LIMIT int,int;
```

Read/Write Order ⇨

Processing Order →

Words of Caution

- Easy to build queries that
  - Retrieve nonsense
  - Never complete, end up completely bogging down the database
- Understanding Schema is a way to prevent that

EXERCISE: PRACTICE SQL

Practice SQL

1. On the loop page of the module, download the document called “northwind_onlinedemo.txt” on your desktop.
2. Open your web browser and go to https://sqliteonline.com/ (free emulation of SQL servers).
3. Click right (Win)/double (Mac) on demo, use DROP

Practice SQL

4. Copy-Paste the text of the file “northwind_onlinedemo.txt” (3556 lines) and press Run on the top menu bar (you should see the 15 table appears on the left box).
5. Select all the lines in the box (CTRL + A or Cmd + A) and delete them.

All the game of this tutorial will be to create new tables that can be downloaded for our analyses.

Practice SQL

- Run the following commands:

```
SELECT *
FROM customers
```

```
SELECT ProductName, UnitsInStock * UnitPrice AS profit_max
FROM products
```

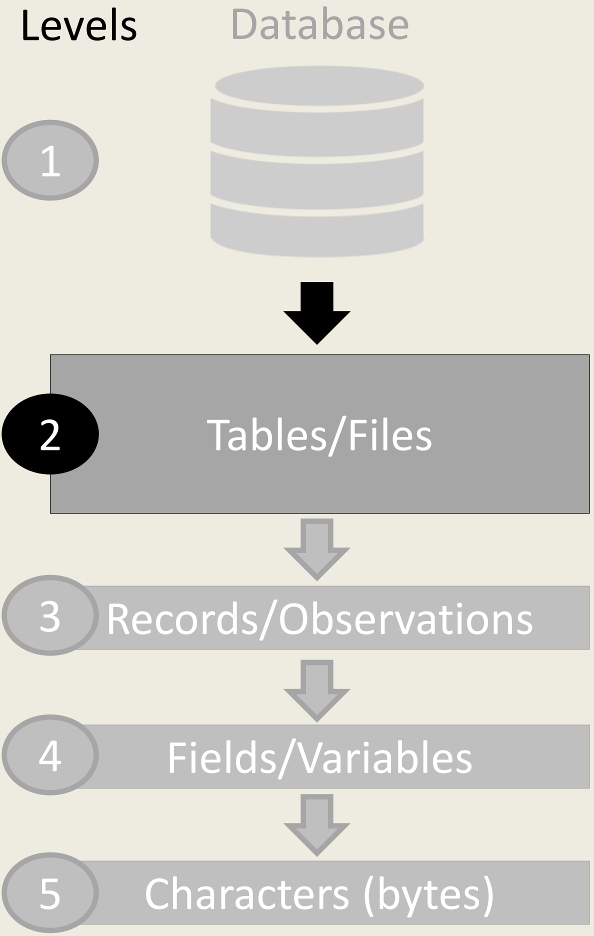
```
SELECT *
FROM customers
WHERE Country = "Mexico"
```

```
SELECT  COUNT (ContactName), Country
FROM customers
GROUP BY Country
```

```
SELECT *
FROM  orders
INNER JOIN customers ON orders.CustomerID =  customers.CustomerID
```

MANAGE TABLES/FILES

Tables/Files Access



Levels

Database

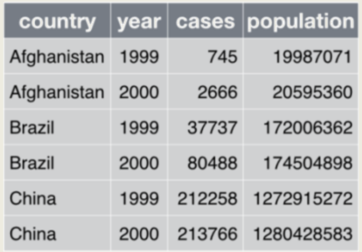
1

2 Tables/Files

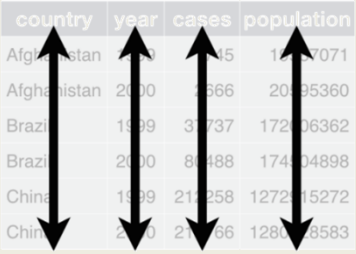
3 Records/Observations

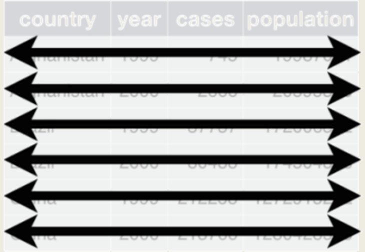
4 Fields/Variables

5 Characters (bytes)



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	1280428583

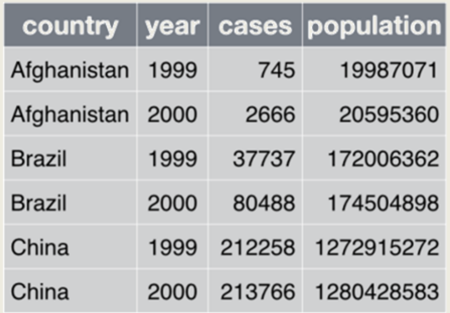




1. Each variable has its own column
2. Each observation is placed in its own row
3. Each value is placed in its own cell

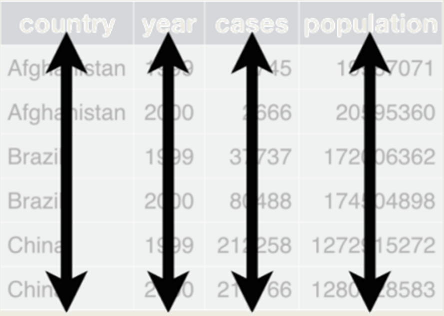
File/Table Structure (1)

Table/File

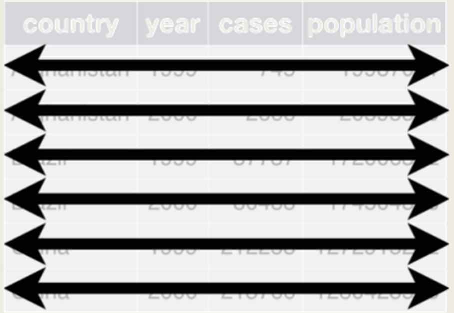


country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Field/Variable



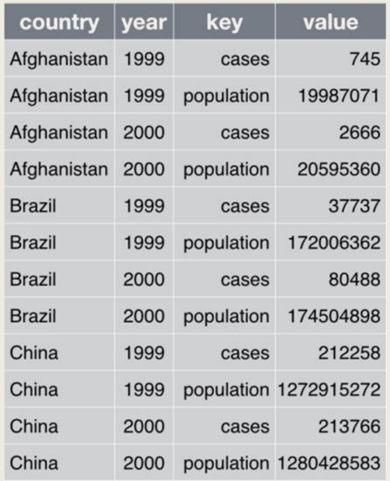
Record/Observation



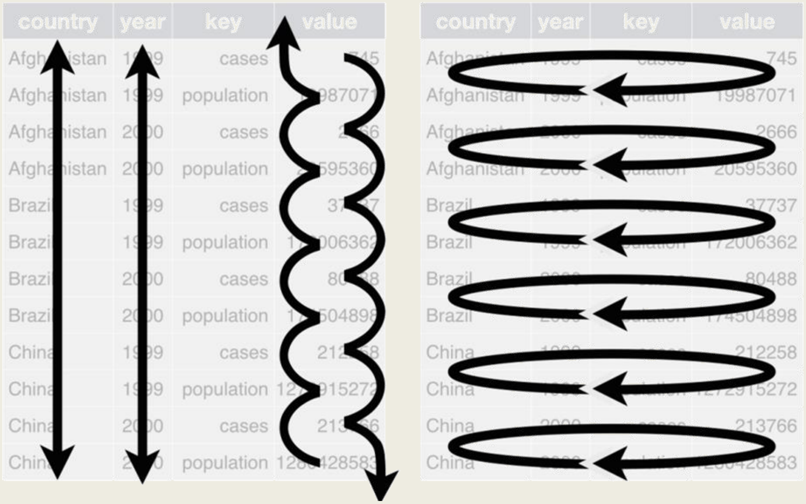
1. Each variable has its own column
2. Each observation is placed in its own row
3. Each value is placed in its own cell

File/Table Structure (2)

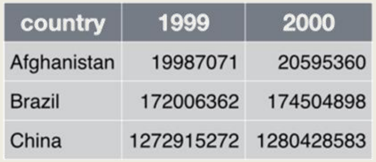
- Long format



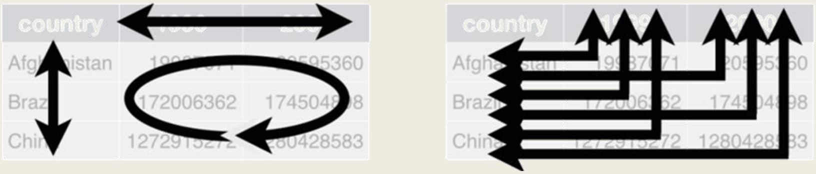
country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583



- Wide format



country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583



Fields/Variables

- “Field” when the observation is recorded
- “Variable” when all the observations are agglomerated
- When processing Files, we are using variables

What is a Variable?

- Series of observations/records with different values
  - if the same value is used this is not a variable
- Have different types:
  - Character
  - Numeric
  - Date
  - …

Naming Conventions

- for files and variables
  - no white space “ “!
- Choose either
  - Camel Case
    - E.g. *someVar*, *someClass*, *somePackage.xyz*
  - Pascal Case
    - E.g. *SomeVar*, *SomeClass*, *SomePackage.xyz*
  - Snake Case
    - E.g. *some_var*, *some_class*, *some_package.xyz*

Files Types/Formats

File Extension	Comment
.doc/.docx/.pdf/.jpg/.mp3/.avi/...	Data can be access but needs file processing or OCR
.xls/.xlsx	MS Excel format (not open) which contains interface, metadata and figures. To avoid
.ods/.ots	Open Office spreadsheet format, contains interface, metadata and figures but is open
.csv	**Comma Separated Value is the most common data format. Open and light**
.txt	Similar to CSV, can be tab separated
.json	JavaScript Object Notation (https://www.json.org/), semi-structured data file
.sav	SPSS format (not open) which contains interface, metadata and figures. To avoid

Convert Data to Information

1. Extract relevant data from the database with a query
2. Check that the structure of the obtained file is compatible with your analysis
3. Process these data with mathematic/statistic calculations

Convert Data to Information

- Receiving data extracted from a database is an optimal way to preform analyses
- However, it is usual to access data that are gathered and analysed in an Excel file
  - Local only
  - No update possible
  - Messy and unstructured

Convert Data to Information

- Next Lecture we will see how to clean and prepare data for our analyses

Homework for Next Lecture

- Next time: Data Cleaning and Transformations with Excel
  - Mandatory
    - https://www.udemy.com/course/excel_quickstart/
  - Optional but suggested
    - https://www.udemy.com/course/ten-excel-features-every-analyst-should-know/
  - Just have a look for your interest
    - https://www.udemy.com/course/excel-dashboards-in-an-hour/

DATABASE
USERS
ID (INT)
NAME (TEXT)
SURNAME (TEXT)
NICKNAME (TEXT)
EMAIL (TEXT)
PASSWORD (TEXT)
ORDERS
ID (INT)
ORDER_TIME
USER_ID
PRODUCT_ID
PRODUCTS
ID (INT)
NAME (TEXT)
DESCRIPTION
PRICE (DOUBLE)
PICTURE (TEXT)
ENABLED (BOOL)
CODE (INT)
IN-STOCK (INT)
CATEGORY_ID
CATEGORY
ID
NAME
DESCRIPTION
PICTURE
WEB SERVER
FIREWALL

QUESTIONS?