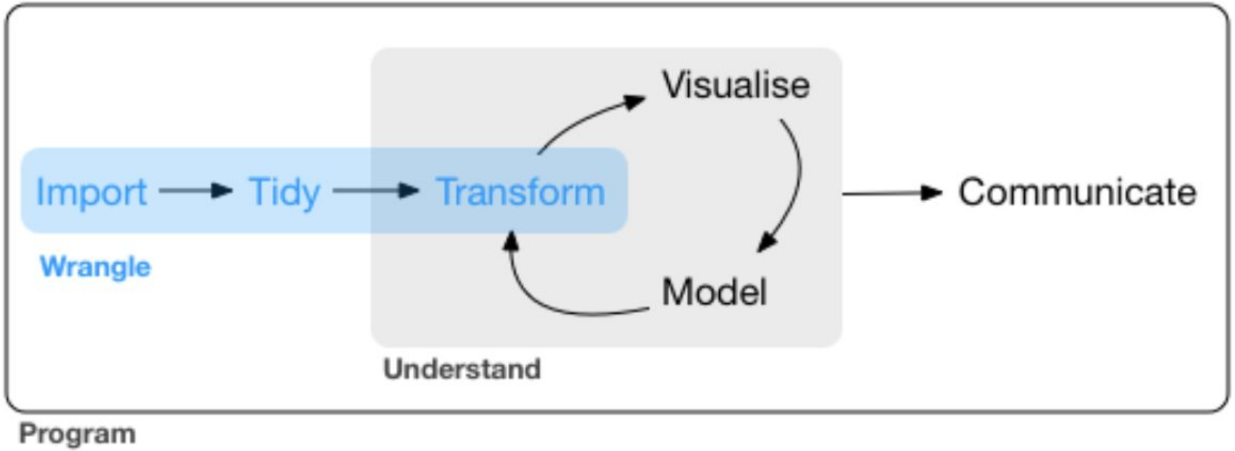
2.2 Tidy data

Applied Data Analysis (ADA)

Oxford DH Summer School - 2022



https://r4ds.had.co.nz/tidy-data.html

A motivating example

Consider a spreadsheet like this one:

Manuscript	Users	Conservation note	Last consulted	User 1	User 2	User 3
BL.201	John, James, Janeth	Approved for consultation	March 2010	John	James	Janeth
BL.301	Mary	Only under supervision	10-12-2009	Mary		
BL.401	Susan, Mark	OK	10 May	Mark		

Let’s say the sheet contains years of consultation activity at your library, and you want to analyse it. Can you think about any issue you might face?

Vocabulary of tidy data

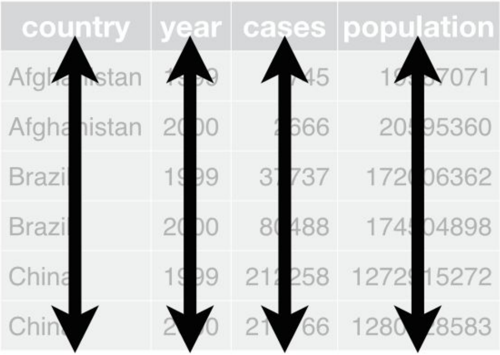
Data structure: the way data is organised. E.g., in 2D tables made of rows and columns.

Data semantics: data is a collection of **values**. Every value belongs to a **variable** and an **observation**. In our example, an observation is a consulted manuscript, a variable is the date it was last consulted.

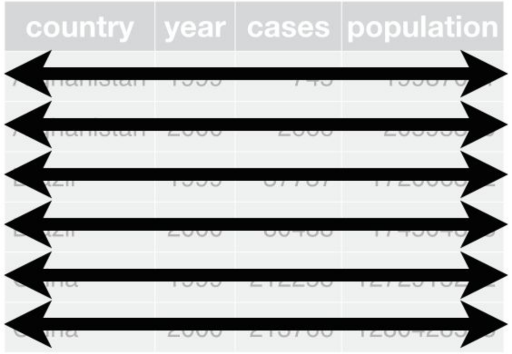
Tidy data is a standard way to map the meaning of a dataset to its structure.

There are few interrelated rules which make a dataset tidy:

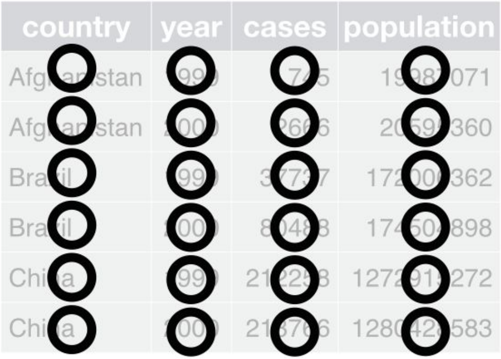
1. The **dataset** is organized into a collection of **tables** (or relations, or data frames).
2. Every **table** contains data for a single **observation type** (or entity, or class).
3. Each **variable** (or attribute) must have its own **column**.
4. Each **observation** (or tuple, or instance) must have its own **row**.
5. Each **value** must have its own **cell**.



variables



observations



values

Tidy motivating example

Manuscript ID	Conservation note	Last consulted *(calculated!)*
BL.201	Approved	03-2010
BL.301	Supervised	12-2009
BL.401	Approved	05-2009

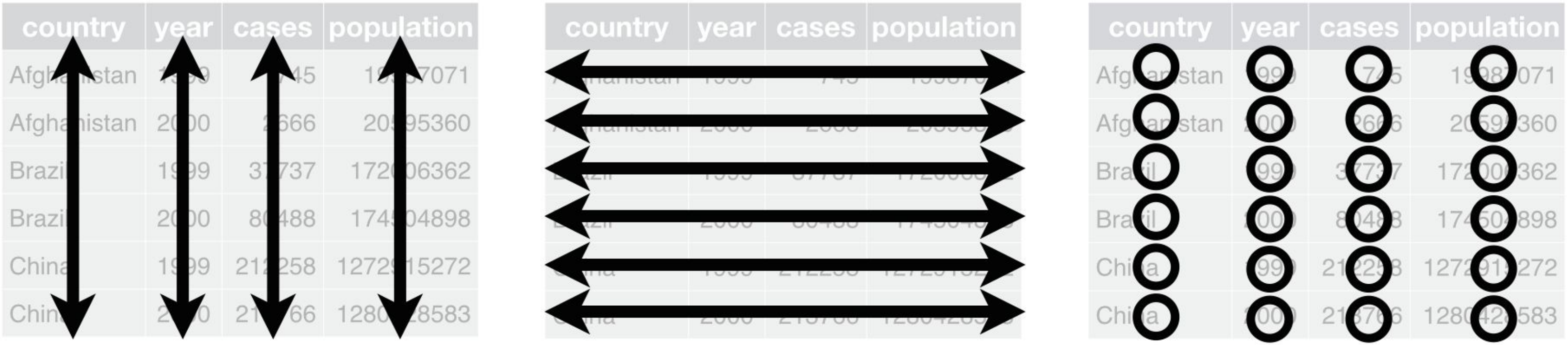
User ID	Name
0	John
1	Mary
2	Susan

User ID	Manuscript ID	End of consultation
0	BL.201	01-2008
1	BL.301	12-2009
2	BL.401	05-2009

Note: not all observations are reported.

Motivations:

1. Same approach for all data.
2. Minimises redundancy.
3. Maximises intrinsic uniformity (1 column/variable has 1 data type, etc.) and ease of manipulations.



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

variables

observations

values

Five most common problems with messy datasets:

1. Column headers are values, not variable names *(User1, User2, …).*
2. Multiple variables are stored in one column *(Users).*
3. Variables are stored in both rows and columns.
4. Multiple types of observational units are stored in the same table *(Manuscripts and users).*
5. A single observational unit is stored in multiple tables. Note: this does not necessarily applies for observation types. *Right?*

Manuscript	Users	Conservation note	Last consulted	User 1	User 2	User 3
BL.201	John, James, Janeth	Approved for consultation	March 2010	John	James	Janeth
BL.301	Mary	Only under supervision	10-12-2009	Mary		
BL.401	Susan, Mark	OK	10 May	Mark		

The entity-relationship model

A conceptual model of the data, it defines a **conceptual data schema**. It does not describe actual data. In object-oriented programming, we reason about classes and not their instances.

An E-R model contains the following components:

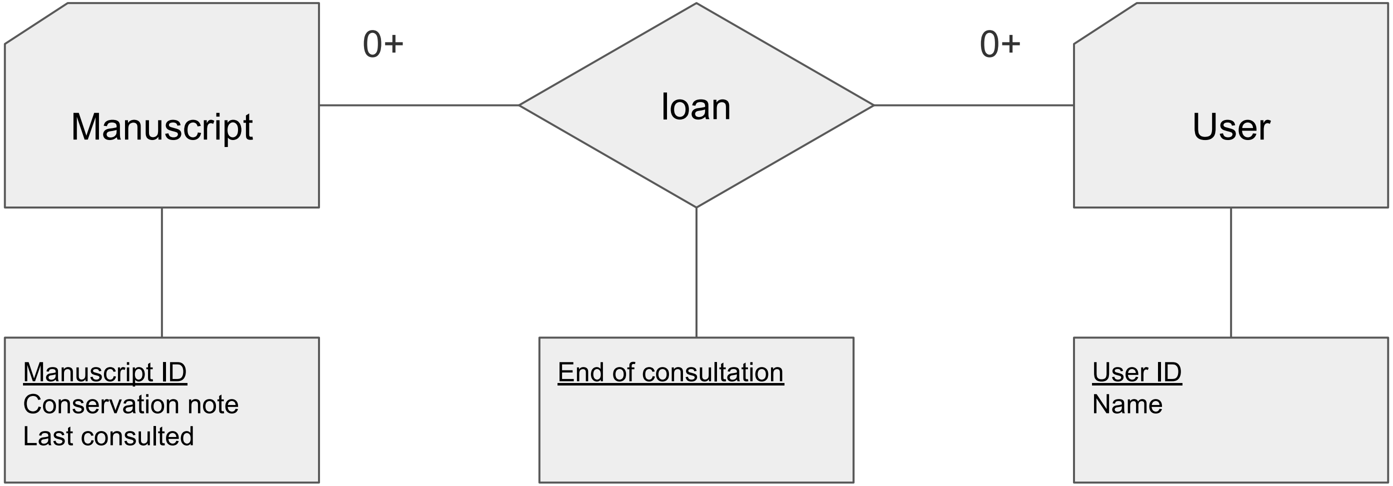
- **Entities** (observational types): a complex concept we want to model. E.g. books and persons.
- **Relationships**: a logical tie between entities. An instance of a relationship is given between two instances of entities. E.g. a person (entity) can be the author (relationship) of a book (entity).
- **Attributes** (variables): entities and relationships can possess atomic attributes. E.g. a book can have a publication year.
- **Keys**: every instance of an entity must be uniquely identifiable via a key, one or more of its attributes which, in combination, are unique for the given instance.
- **Cardinalities** of relationships: one to one, one to many, many to many.

http://users.dimi.uniud.it/~massimo.franceschet/ds/syllabus/learn/database/ER.html

A motivating example

Manuscript	Users	Conservation note	Last consulted	User 1	User 2	User 3
BL.201	John, James, Janeth	Approved for consultation	March 2010	John	James	Janeth
BL.301	Mary	Only under supervision	10-12-2009	Mary		
BL.401	Susan, Mark	OK	10 May	Mark		

A motivating example



The relational model

It defines a **logical data schema**, independent of the physical model (i.e. how the data is stored and how the queries are actually implemented). The relational model is composed of: 1) **data structures** and 2) **integrity constraints** defined over them (which we won’t consider here).

The main concept is that of **relation**, whose representation is a **table**:

1. A relation/table consists of columns and rows. Each column is an attribute (variable), each row is a tuple (observation). Each row/column intersection contains a single (atomic) value.
2. Each attribute is associated with a **domain**: a set of values it can take.
3. Every relation must have a **primary key**: a combination of attribute values that uniquely identify every observation.
4. Attributes, excluding those part of a key, can have null values.

http://users.dimi.uniud.it/~massimo.franceschet/ds/syllabus/learn/database/RM.html

A motivating example

Manuscript

Manuscript ID
Conservation note
Last consulted

User

User ID
Name

Loan

Manuscript ID
User ID
End of consultation

Loan

Foreign keys:
Manuscript ID -> Manuscript(Manuscript ID)
User ID -> User(User ID)

Tidy motivating example

Manuscript ID	**Conservation note**	**Last consulted *(calculated!)***
BL.201	Approved	03-2010
BL.301	Supervised	12-2009
BL.401	Approved	05-2009

User ID	**Name**
0	John
1	Mary
2	Susan

Note: an integrity constraint for this dataset is that two users cannot load a manuscript before the end of its last consultation has passed.

Note: not all observations are reported.

User ID	**Manuscript ID**	**End of consultation**
0	BL.201	01-2008
1	BL.301	12-2009
2	BL.401	05-2009

A second look at our vocabulary

Several traditions are focusing on roughly the same concept: representing complex data.
Tidy data is a framework for statisticians. The E-R and relational model come from the database community; classes and instances from object oriented programming.

We use the tidy vocabulary from now on. A glossary goes as follows:

1. **Observational type**: entity, class.
2. **Table**: relation.
3. **Observation**: tuple, instance.
4. **Variable**: attribute.

Key concepts we also use in the tidy setting:

1. **Domain** of a variable: the values it can take.
2. **Key**: one or more variables whose values identify observations within a table.
3. **Cardinality** of relationships between tables.