Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction to Data Analysis

DATA COLLECTION & DATA MANAGEMENT

Patrick Boily

Data Action Lab | uOttawa | Idlewyld Analytics

pboily@uottawa.ca

[with files from Jen Schellinck | Sysabee]

OBJECTIVE

We seek data that can:

- provide **legitimate insight** into our system of interest;
- provide **correct**, **accurate** answers to relevant questions;
- **support** the drawing of **valid** conclusions, with the ability to **qualify/quantify** these conclusions in terms of scope and precision.

This cannot be done without **study design:** what data should we collect, and how should we collect it.

MOTIVATIONS FOR DATA COLLECTION

Three functions, historically:

- record keeping (people/societal management)
- science – new general knowledge
- intelligence – business, military? police? social? domestic? personal?

Each of these three functions have traditionally used different **sources** of information.

- they have collected **different types of data**
- they also have **different data cultures** and **terminologies**

DATA IS REAL

Data is a representation, but data is still **physical**.

It has physical properties, it requires physical space & energy to work with it.

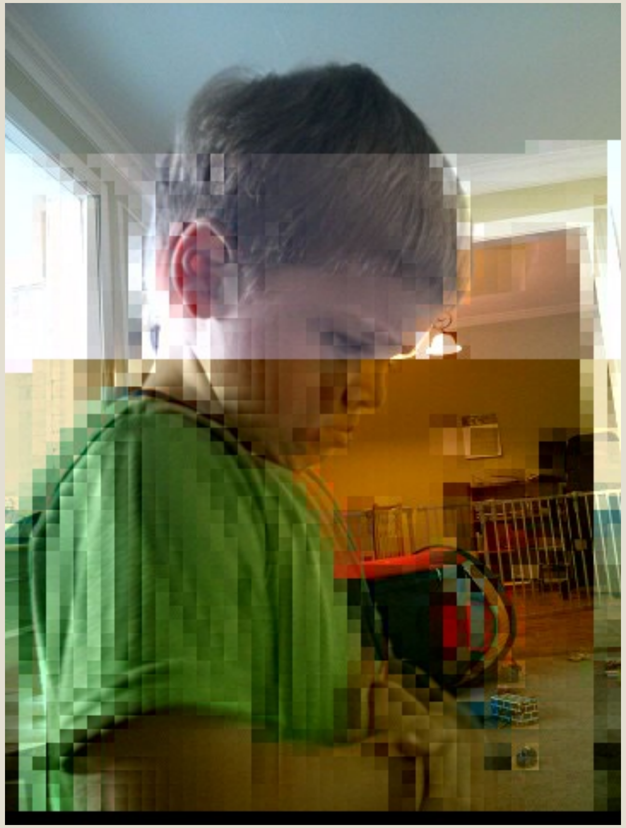
DATA DECAYS

Data ages over time – it has a **shelf life**.

We use the phrase “rotten data” or “decaying data”

- **literally** – the data storage medium might decay
- **metaphorically** – when the data no longer accurately **represents** the relevant objects and relationships or even when those objects no longer exist in the same way

Data must be kept ‘fresh’ and ‘current’, not ‘stale’ (context and model dependent!)

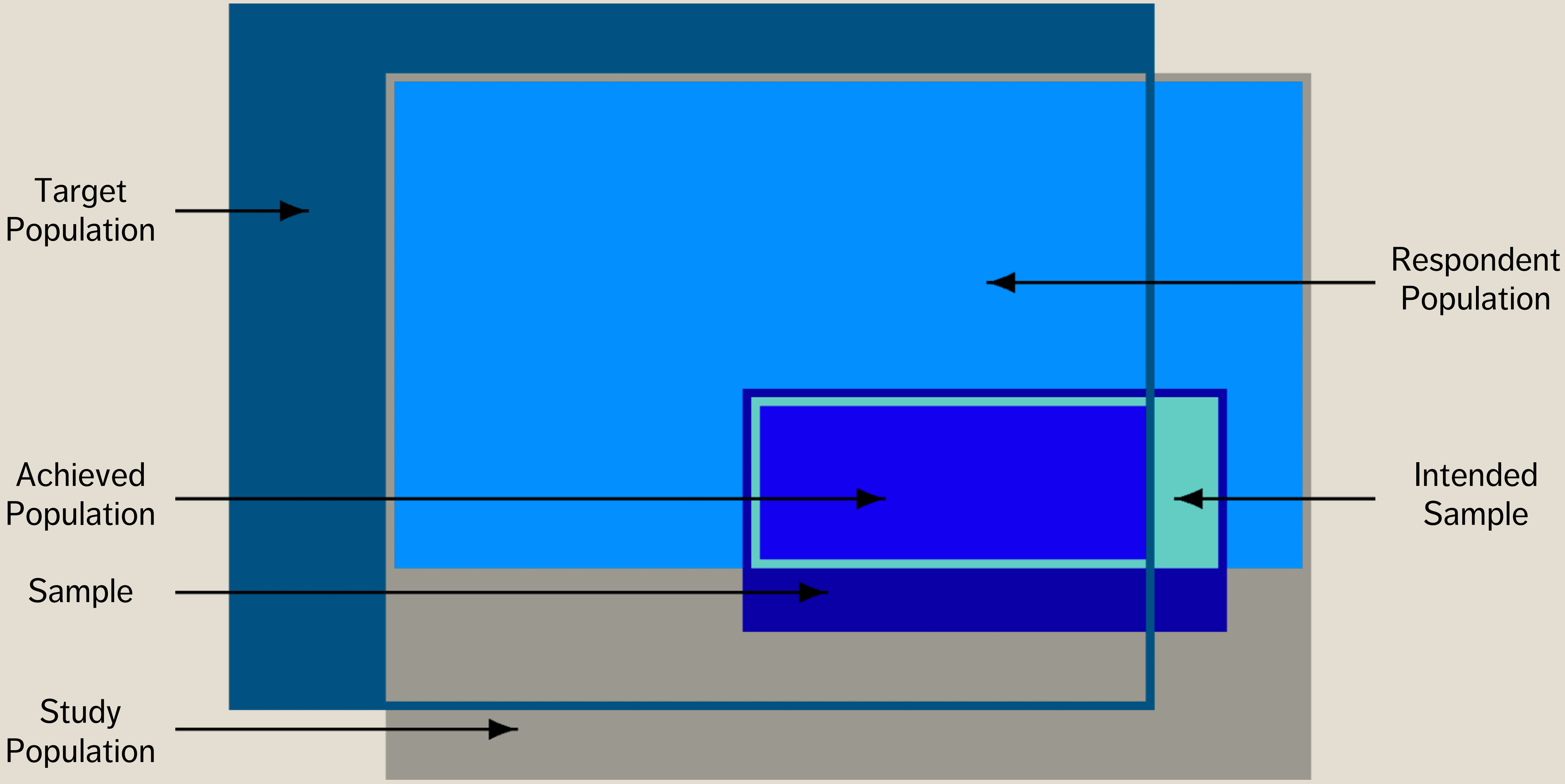


NPS AND PATTERN FISHING

Two separate issues can be combined to cause **problems** with data analysis:

- drawing conclusions (inferences) from a sample about a population that are not warranted by the sample collection method (symptomatic of NPS);
- looking for any available patterns in the data and then coming up with post hoc explanations for these patterns.

Alone or in combination, these lead to poor (and **potentially harmful**) conclusions.



Target Population

Respondent Population

Achieved Population

Intended Sample

Sample

Study Population

STUDY/SURVEY STEPS

Surveys follow the same general steps:

1. statement of objective
2. selection of survey frame
3. sampling design
4. questionnaire design
5. data collection
6. data capture and coding
7. data processing and imputation
8. estimation
9. data analysis
10. dissemination
11. documentation

The process is not always linear, but there is a definite movement from **objective** to **dissemination**.

NON-PROBABILISTIC SAMPLING

Nonprobabilistic sampling (NPS) methods (designs) select sampling units from the target population using subjective, non-random approaches.

- NPS are quick, relatively inexpensive and convenient (no frame required).
- NPS methods are ideal for exploratory analysis and survey development.

Unfortunately, NPS are often used instead of probabilistic designs (not good)

- the associated selection bias makes NPS methods inferentially unsound;
- automated data collection often fall squarely in the NPS camp – we can analyze data collected with a NPS approach, but not generalize the results to the target population.

PROBABILISTIC SAMPLING

Probabilistic sample designs are usually more **difficult** and **expensive** to set-up (due to the need for a quality frame), and take longer to complete.

They provide **reliable estimates** for the attribute of interest and the **sampling error**, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

SAMPLING DESIGNS

Different **sampling designs** have distinct advantages and disadvantages.

They can be used to compute estimates

- for various population attributes: mean, total, proportion, ratio, difference, etc.
- for the corresponding 95% confidence intervals.

We might also want to compute sample sizes for a given **error bound** (an upper limit on the radius of the desired 95% CI), and how to determine the **sample allocation** (how many units to be sampled in various sub-population groups).

PROBABILISTIC SAMPLING DESIGNS

Simple random sampling (SRS)

Stratified random sampling (STS)

Systematic sampling (SYS)

Cluster sampling (CLS)

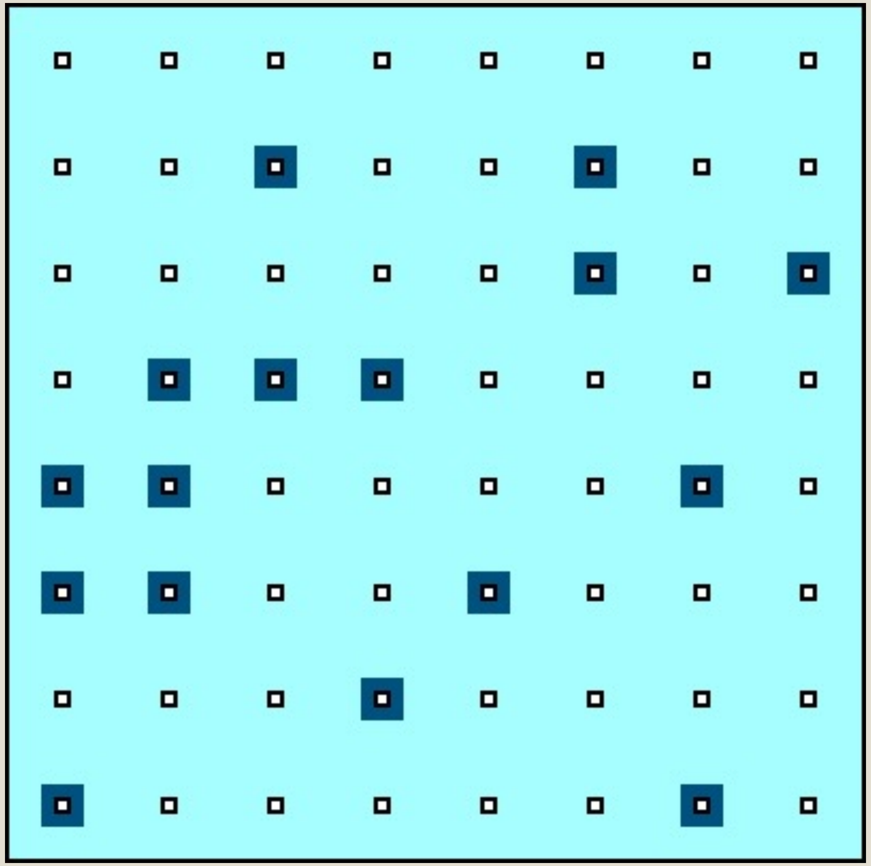
Probability proportional-to-size sampling (PPS)

Replicated sampling (RES)

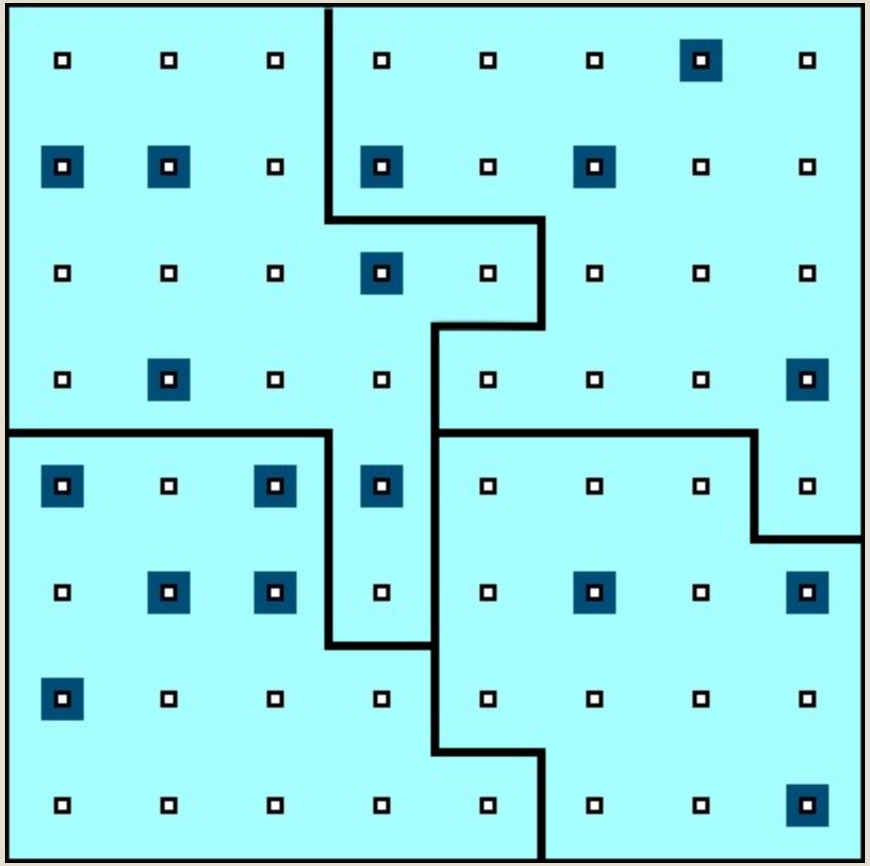
Multi-stage sampling (MSS)

Multi-phase sampling (MPS)

SAMPLING DESIGNS

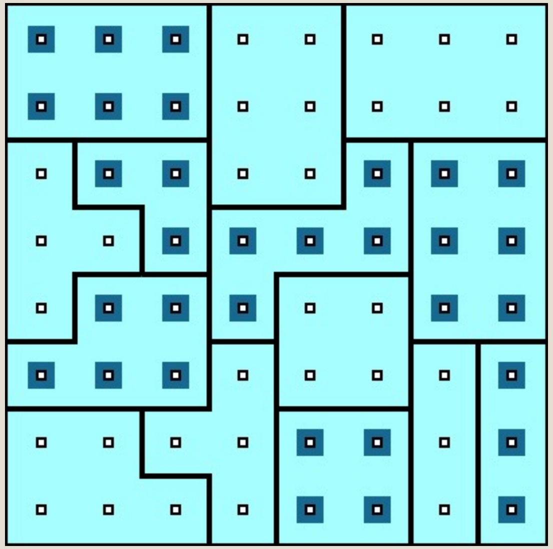


Simple Random
Sampling (SRS)

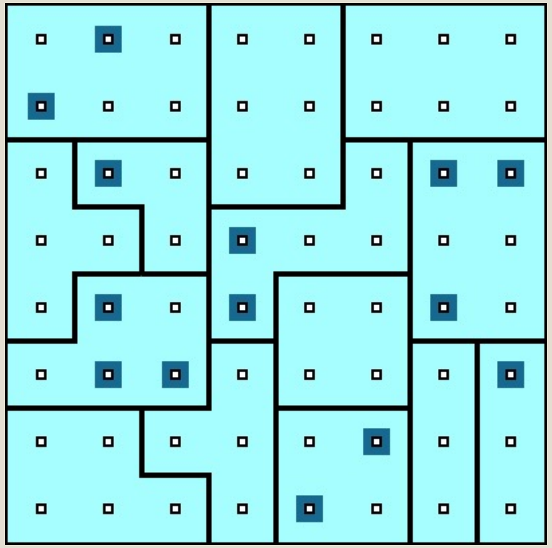


Stratified random
sampling (STS)

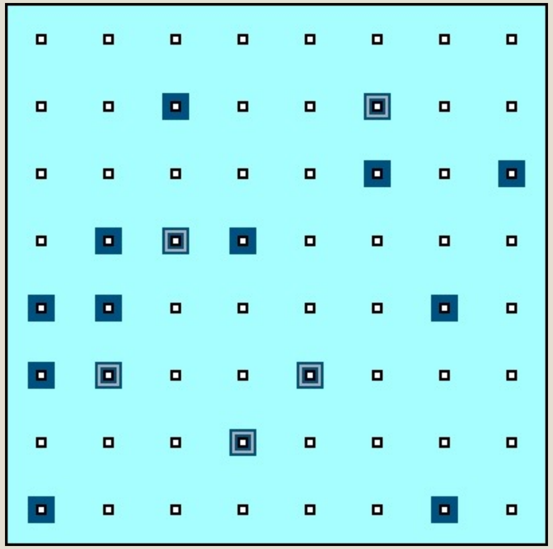
OTHER SAMPLING DESIGNS



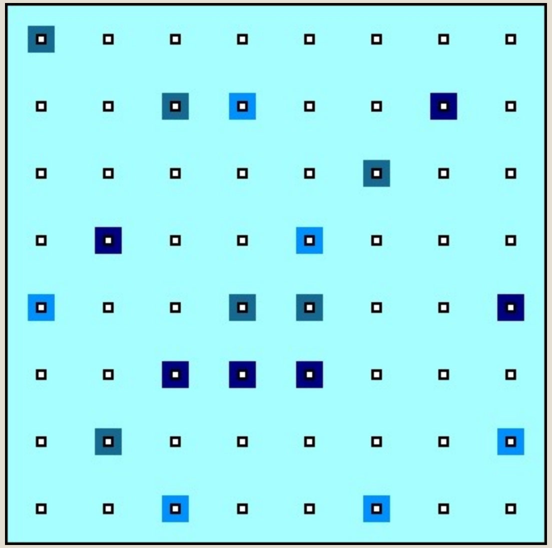
Cluster Sampling (ClS)



Multi-Stage Sampling (MSS)



Multi-Phase Sampling (MPS)



Replicated Sampling (ReS)

AUTOMATED DATA COLLECTION CHECKLIST

With regards to social scientific data:

- sparse financial resources
- little time or desire to collect data by hand
- want to work with up to date, high-quality data sources
- document process from data collection to publication for reproducibility

Issues with manual collection:

- non-reproducible process
- prone to errors and cumbersome
- subject to heightened risks of “death by boredom”

Advantages:

- reliability
- reproducibility
- time-efficient
- higher quality datasets

WEB SCRAPING DATA QUALITY

First-hand information: for example, a tweet, or a news article.

Second-hand data: data that has been copied from an offline source or scraped from elsewhere.

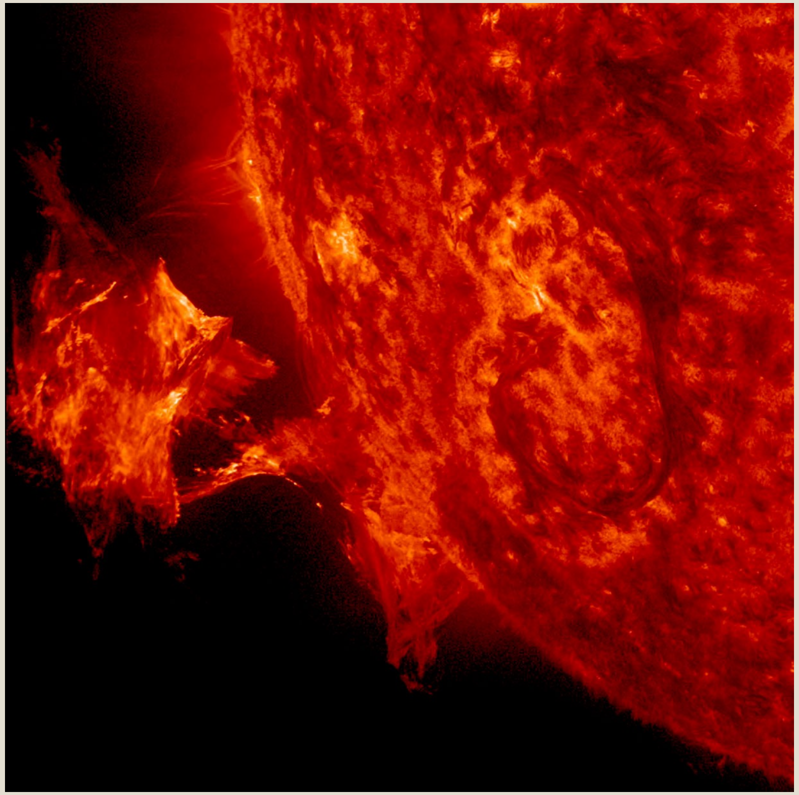
- Sometimes one can’t remember or retrace the source of such data.
- Does it still make sense to use it? It depends.

Any use of secondary data requires **cross-checking** and **validation**.

STRUCTURED/UNSTRUCTURED DATA

A major motivator for new developments in database types and other data storing strategies is the increasing availability of **unstructured** data and '**blob**' data:

- **structured data**: labeled, organized, discrete structure is constrained and pre-defined
- **unstructured data**: not organized, no specific pre-defined structure data model (text)
- **blob data**: **B**inary **L**arge **Ob**ject (BLOb) – images, audio, multi-media



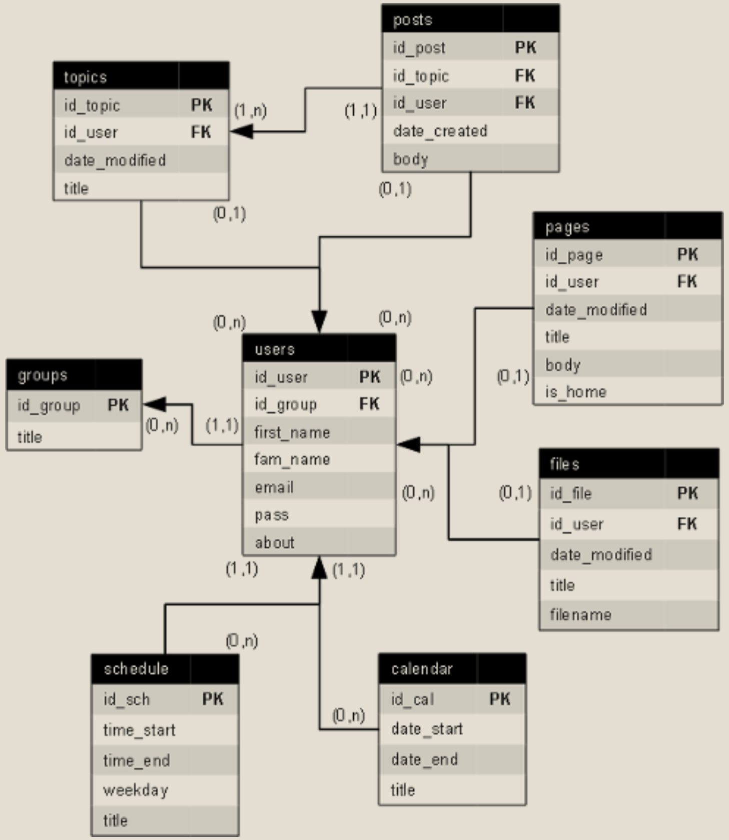
RELATIONAL DATABASES

Data is stored in a series of **tables**.

Broadly speaking, each table represents an object and some properties related to this object.

Special columns in the tables **connect** object instances across tables (allowing for merges).

The traditional approach to data storage.



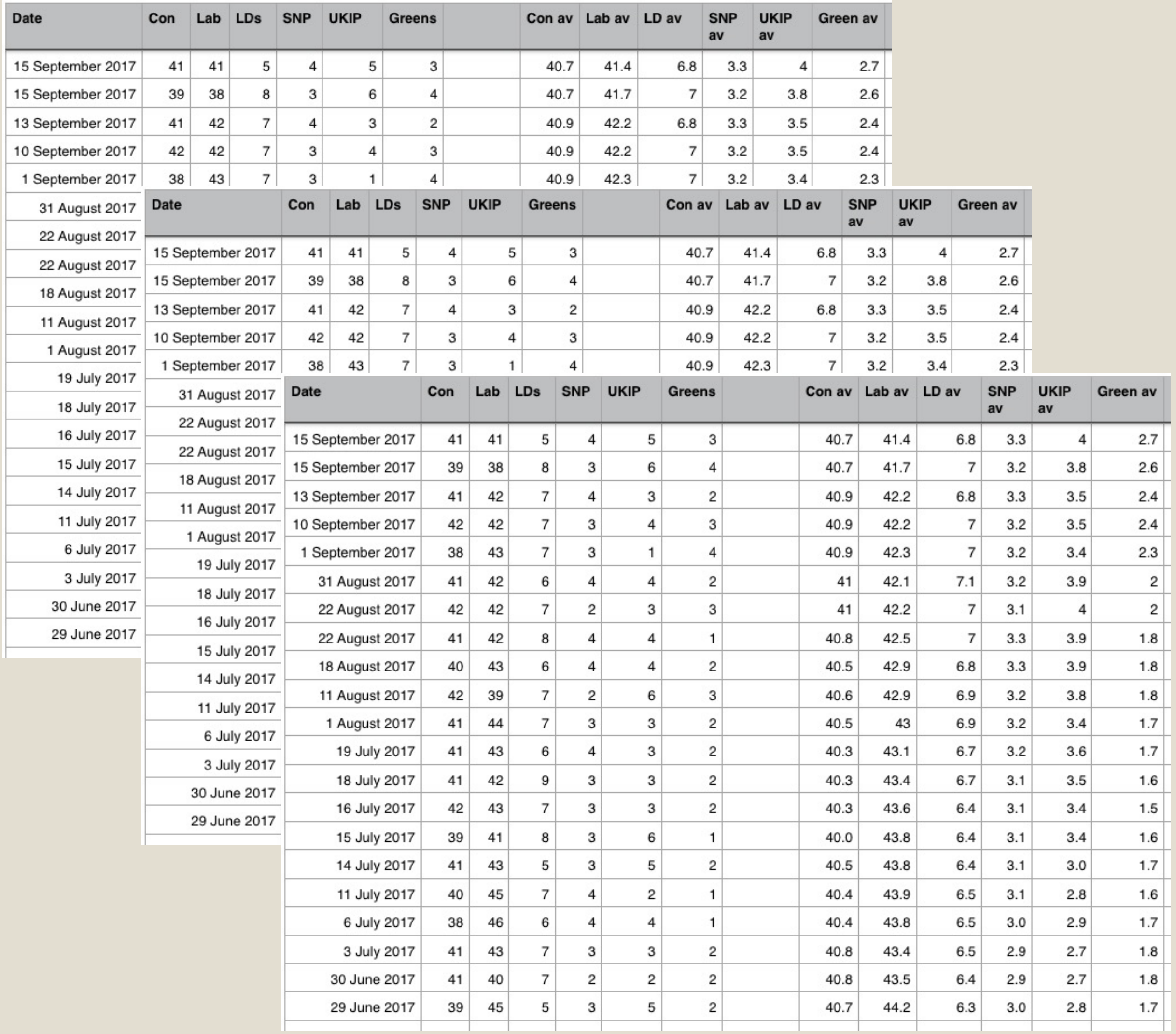
FLAT FILES AND SPREADSHEETS

What about keeping data in a single giant table (spreadsheet)?

Or multiple spreadsheets?

How bad can it be?

Wayne Eckerson coined the term ‘spreadmart’ to describe a situation with many (*ad hoc*) spreadsheets as a data strategy.



Date	Con	Lab	LDs	SNP	UKIP	Greens		Con av	Lab av	LD av	SNP av	UKIP av	Green av
15 September 2017	41	41	5	4	5	3		40.7	41.4	6.8	3.3	4	2.7
15 September 2017	39	38	8	3	6	4		40.7	41.7	7	3.2	3.8	2.6
13 September 2017	41	42	7	4	3	2		40.9	42.2	6.8	3.3	3.5	2.4
10 September 2017	42	42	7	3	4	3		40.9	42.2	7	3.2	3.5	2.4
1 September 2017	38	43	7	3	1	4		40.9	42.3	7	3.2	3.4	2.3
31 August 2017	41	42	6	4	4	2		41	42.1	7.1	3.2	3.9	2
22 August 2017	42	42	7	2	3	3		41	42.2	7	3.1	4	2
22 August 2017	41	42	8	4	4	1		40.8	42.5	7	3.3	3.9	1.8
18 August 2017	40	43	6	4	4	2		40.5	42.9	6.8	3.3	3.9	1.8
11 August 2017	42	39	7	2	6	3		40.6	42.9	6.9	3.2	3.8	1.8
1 August 2017	41	44	7	3	3	2		40.5	43	6.9	3.2	3.4	1.7
19 July 2017	41	43	6	4	3	2		40.3	43.1	6.7	3.2	3.6	1.7
18 July 2017	41	42	9	3	3	2		40.3	43.4	6.7	3.1	3.5	1.6
16 July 2017	42	43	7	3	3	2		40.3	43.6	6.4	3.1	3.4	1.5
15 July 2017	39	41	8	3	6	1		40.0	43.8	6.4	3.1	3.4	1.6
14 July 2017	41	43	5	3	5	2		40.5	43.8	6.4	3.1	3.0	1.7
11 July 2017	40	45	7	4	2	1		40.4	43.9	6.5	3.1	2.8	1.6
6 July 2017	38	46	6	4	4	1		40.4	43.8	6.5	3.0	2.9	1.7
3 July 2017	41	43	7	3	3	2		40.8	43.4	6.5	2.9	2.7	1.8
30 June 2017	41	40	7	2	2	2		40.8	43.5	6.4	2.9	2.7	1.8
29 June 2017	39	45	5	3	5	2		40.7	44.2	6.3	3.0	2.8	1.7