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 cllect, and how should we collect it.

DATA SOURCES

DATA COLLECTION AND DATA MANAGEMENT

FUNDAMENTAL QUESTIONS

Why do we collect data? What can we **do** with data?

Where does data come from?

What does ‘a **collection**’ of data look like? How could it be described?

Do we need to distinguish between data, information, knowledge?

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 CULTURES AND TERMS

Business Intelligence:

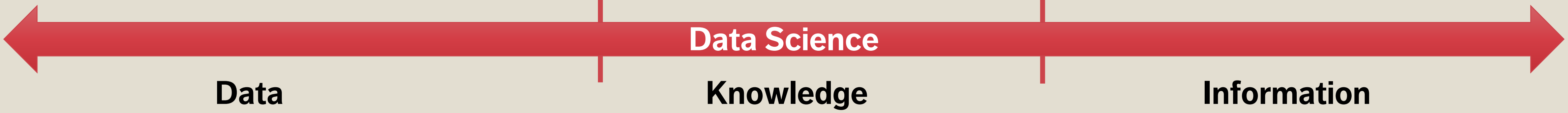
- data warehouse + data mart
- data 'dimension' (= data set)
- hierarchical data (slices)
- data element
- dimension table + fact table

Science/Statistics:

- experimental data
- trials
- participants
- variables
- correlation

Record Management:

- information architecture
- file plan
- information resource
- field
- form and subject



Data Science

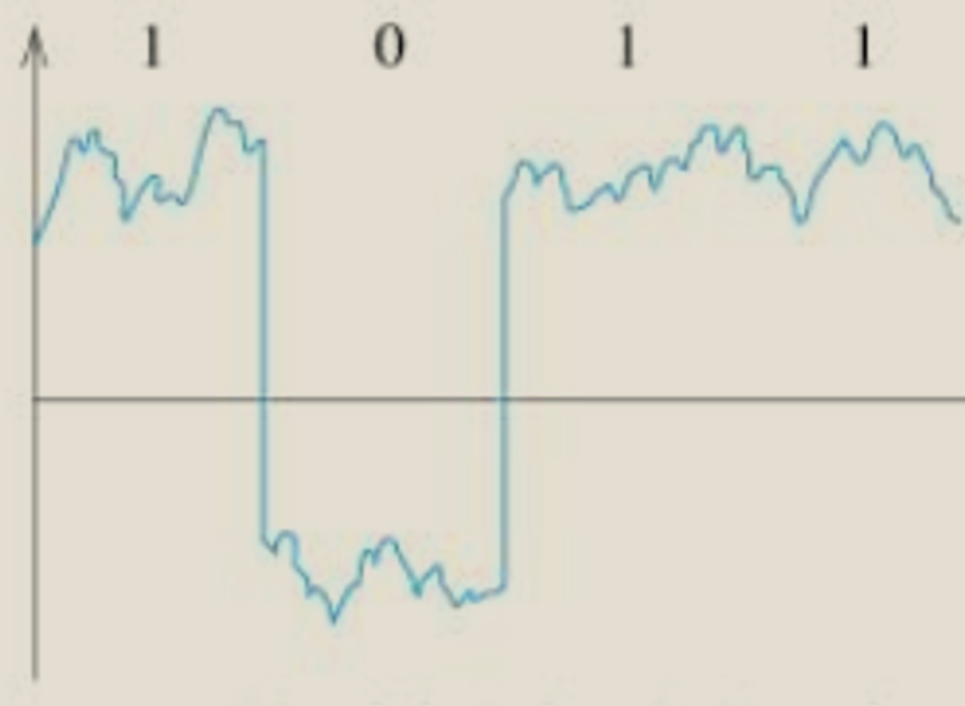
Data **Knowledge** **Information**

COMPUTERS AND DATA

Computer/information science has its own theoretical, **fundamental** viewpoint about data, and information.

Data becomes digital: computers operate over data in a fundamental sense – 1's, 0's representing numbers, letters, etc.

Pragmatically, data is now stored on computers, and is accessible through world-wide computer networks.



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.

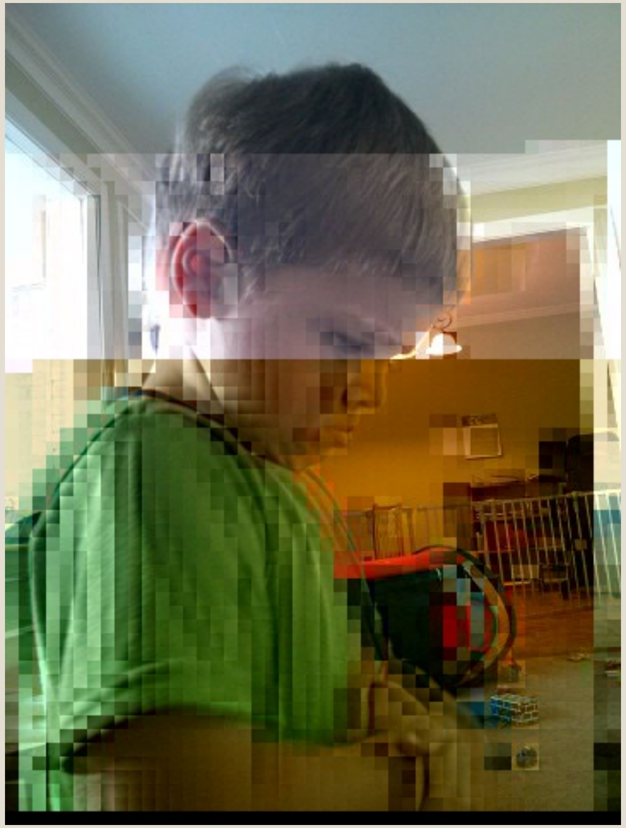
DATA DECAYS

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!)



SAMPLING THEORY AND STUDY DESIGN

DATA COLLECTION AND DATA MANAGEMENT

“The latest survey shows that 3 out of 4 people make up 75% of the population”

D. Letterman

[G. Smith, The Exaggerated Promise of So-Called Unbiased Data Mining]

"A Dartmouth graduate student used an MRI machine to study the brain activity of a salmon as it was shown photographs and asked questions. The most interesting thing about the study was not that a salmon was studied, but that the **salmon was dead**. Yep, a dead salmon purchased at a local market was put into the MRI machine, and some patterns were discovered. There were inevitably patterns—and they were invariably meaningless."

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.

STUDIES, SURVEYS, AND SAMPLING MODELS

A **survey** is any activity that collects information about characteristics of interest:

- in an **organized** and **methodical** manner;
- from some or all **units** of a population;
- using **well-defined** concepts, methods, and procedures, and
- compiles such information into a **meaningful** summary form.

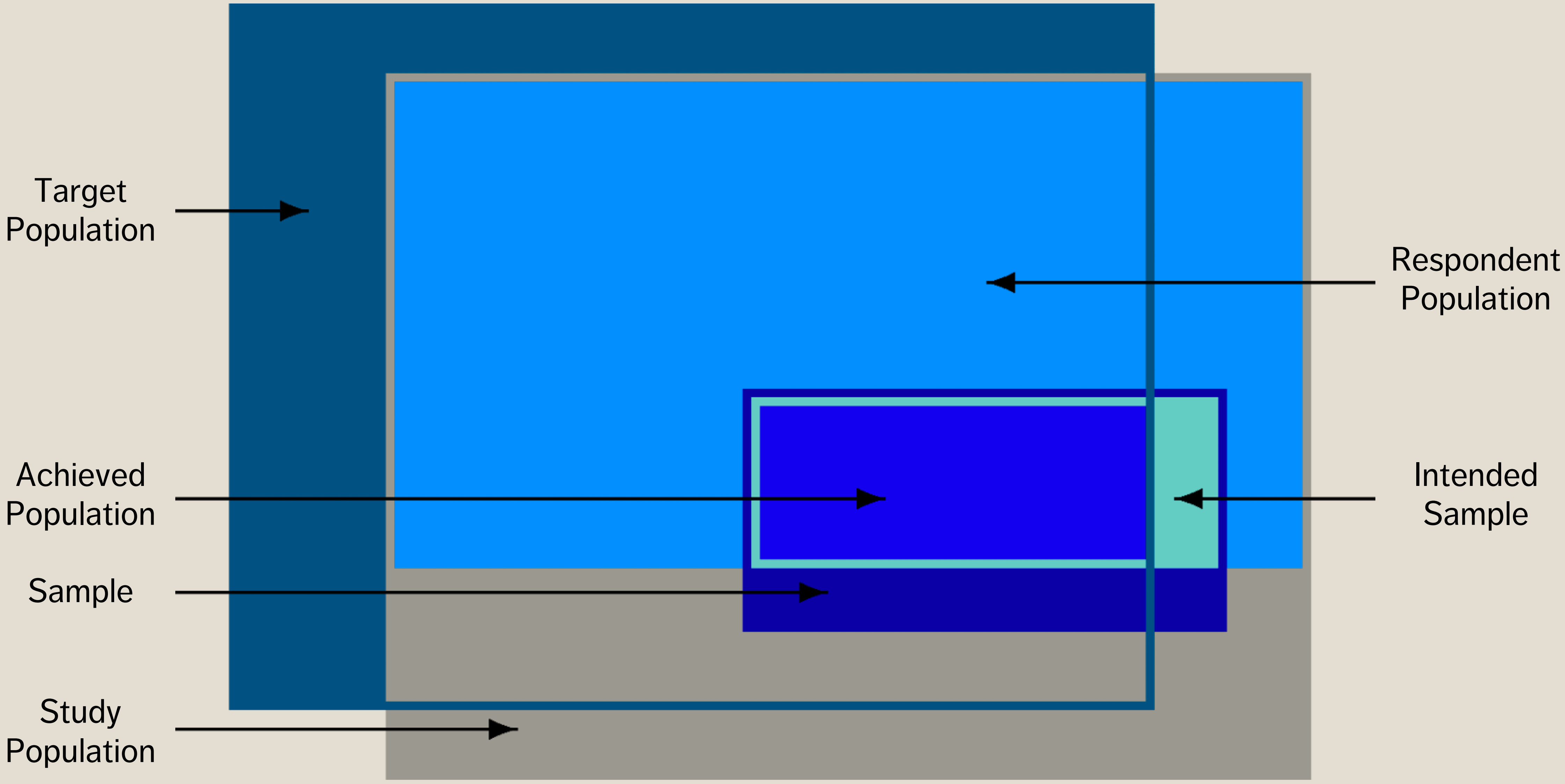
A **census** is a survey where information is collected from all units of a population, whereas a **sample survey** uses only a fraction of the units.

When survey sampling is done properly, we may be able to use various **statistical methods** to make **inferences** about the **target population** by sampling a (comparatively) small number of units in the **study population**.

DECIDING FACTORS

Sometimes, information about the **entire** population is required in order to answer questions; at other times it is not necessary. The **survey type** depends on multiple factors:

- the type of question that needs to be answered;
- the required precision;
- the cost of surveying a unit;
- the time required to survey a unit;
- size of the population under investigation, and
- the prevalence of the attributes of interest.


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**.

SURVEY FRAMES

The **frame** provides the means of **identifying** and **contacting** the units of the study population. It is generally costly to create and to maintain.

The ideal frame contains ID, contact, classification, maintenance, and linkage data. It must minimize the risk of **under/over-coverage**, as well as the number of duplications and misclassifications.

A statistical sampling approach is contraindicated unless the frame is

- **relevant** (it corresponds, and permits accessibility to, the target population),
- **accurate** (the information it contains is valid),
- **timely** (it is up-to-date), and **competitively priced**.

SURVEY ERROR

$$\text{Total Error} = \underbrace{\text{Sampling Error}}_{\text{survey, not census}} + \underbrace{\text{Measurement Error}}_{\text{observations not measured accurately}} + \underbrace{\text{Non-Response Error}}_{\text{non-respondents having systematic observation differences}} + \underbrace{\text{Coverage Error}}_{\text{frame decay and/or corruption}}$$

Statistical sampling can help provide estimates, but importantly, it can also provide some control over the **total error** (TE) of the estimates.

Ideally, TE= 0. In practice, there are two main contributions to TE: **sampling errors** (due to the choice of sampling scheme), and **non-sampling errors** (everything else).

NON-SAMPLING ERROR

Non-sampling error can be controlled, to some extent:

- **coverage error** can be minimized by selecting high quality, up-to-date frames;
- **non-response error** can be minimized by careful choice of the data collection mode and questionnaire design, and by using “call-backs” and “follow-ups”;
- **measurement error** can be minimized by careful questionnaire design, pre-testing of the measurement apparatus, and cross-validation of answers.

In practice, these suggestions are not that useful in modern times.

This explains, in part, the over-use of **web scraping** and **non-probabilistic sampling**.

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.

NPS METHODS

There are contexts where NPS methods might fit a client's or an organization's need, but they must be informed of the drawbacks, and presented with some probabilistic alternatives.

- **Haphazard:** person on the street, depends on availability of units, interviewer bias
- **Volunteer:** self-selection bias
- **Judgement:** biased by inaccurate preconceptions about the target population
- **Quota:** exit polling, ignores non-response bias
- **Modified:** starts probabilistic, switches to quota as a reaction to high non-response rates
- **Snowball:** "pyramid" scheme

How could NPS methods work against collecting inclusive, representative data?

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)

SIMPLE RANDOM SAMPLING (SRS)

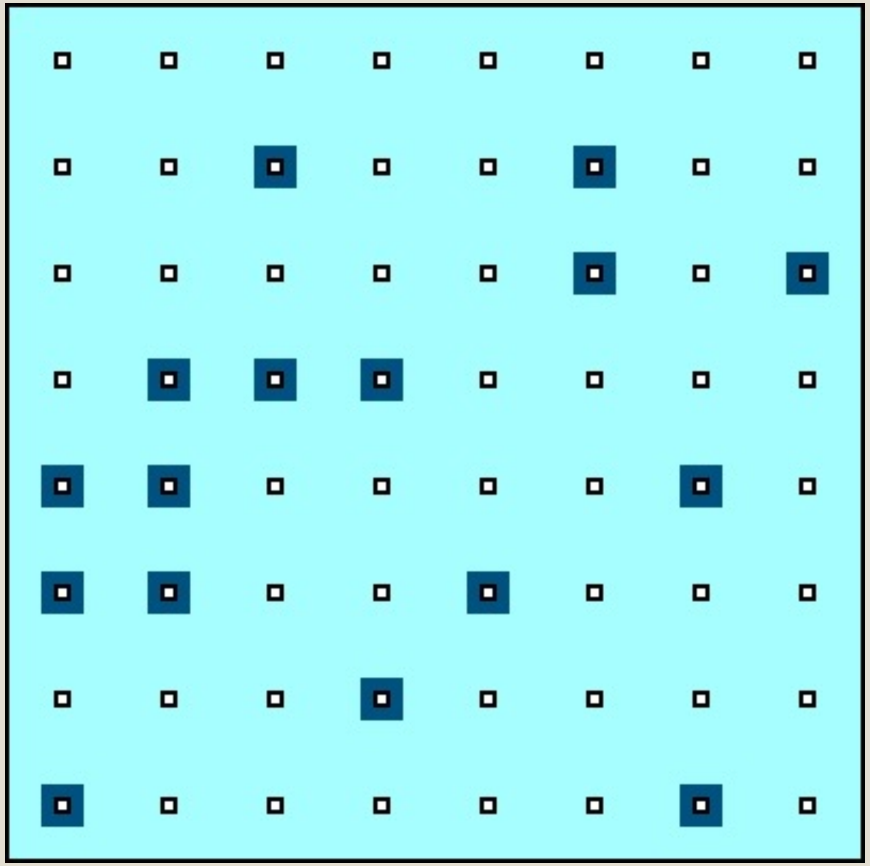
n units are selected randomly from the frame.

Advantages:

- easiest sampling design to implement
- sampling errors are well-known and easy to estimate
- does not require auxiliary information

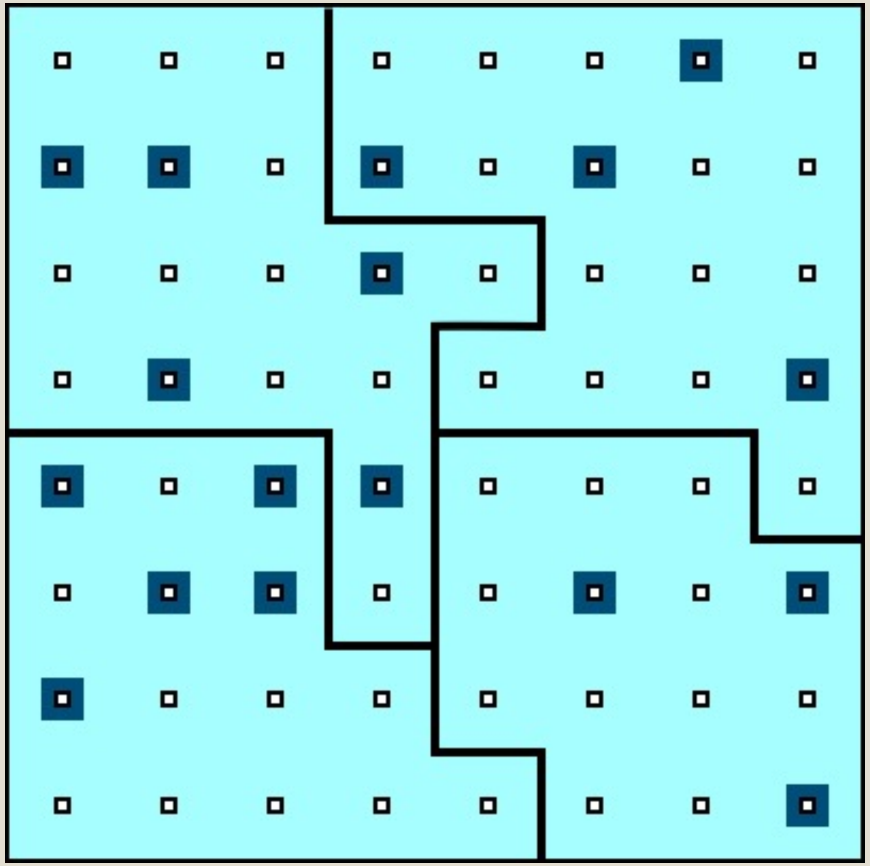
Disadvantages:

- makes no use of auxiliary information
- no guarantee that the sample is representative
- costly if sample is widely spread out, geographically



STRATIFIED RANDOM SAMPLING (StS)

$n = n_1 + \cdots + n_k$ units are selected randomly from k frame **strata**.



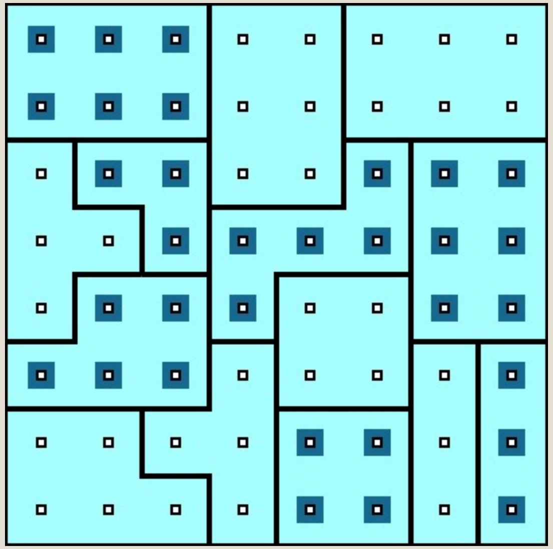
Advantages:

- may produce smaller error bounds than SRS
- may be less costly if elements are conveniently strat.
- may provide estimates for sub-populations
- increases odds of inclusive, representative data

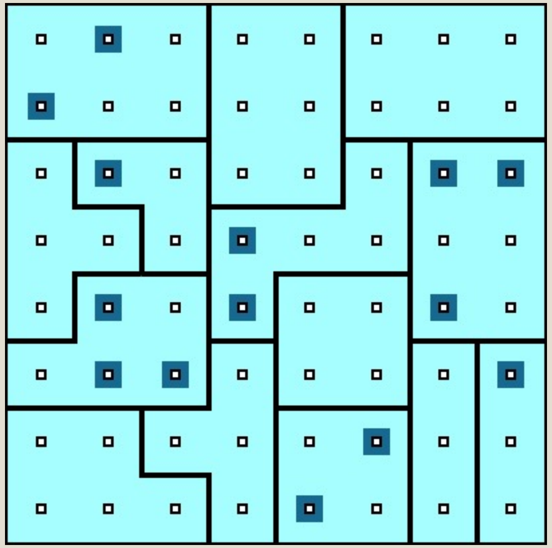
Disadvantages:

- no major disadvantage
- if there are no natural ways to stratify the frame into homogeneous groupings, StS≈SRS

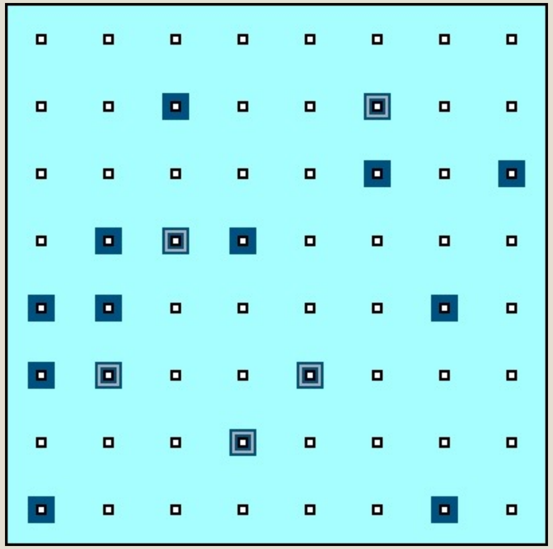
OTHER SAMPLING DESIGNS



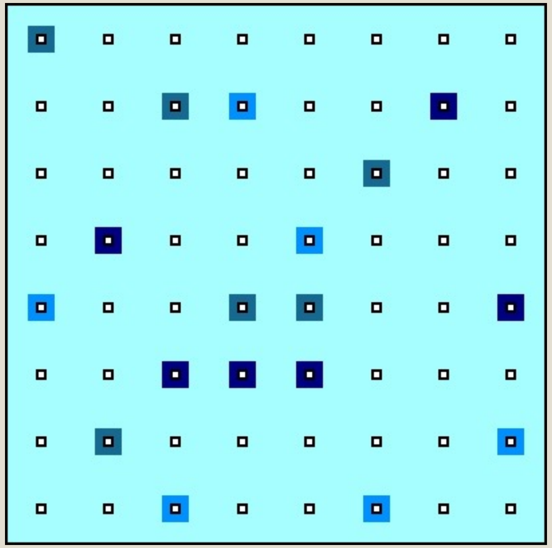
Cluster Sampling (ClS)



Multi-Stage Sampling (MSS)



Multi-Phase Sampling (MPS)



Replicated Sampling (ReS)

WEB SCRAPING & AUTOMATED DATA COLLECTION

DATA COLLECTION AND DATA PROCESSING

“The streets of the Web are paved with data that can’t wait to be collected.”

Munzart, Rubba, Meissner, Nyhuis, Automated Data Collection with R

WORLD WIDE WEB

The way we **share**, **collect**, and **publish** data has changed over the past few years due to the ubiquity of the *World Wide Web* (WWW).

Private businesses, **government**, and **individual users** are posting and sharing all kinds of data and information.

At every moment, new channels generate vast amounts of data on human behaviour.

WORLD WIDE WEB

There was a time in the recent past where both scarcity and inaccessibility of data was a problem for researchers and decision-makers. That is **emphatically** not the case anymore.

Data abundance carries its own set of problems:

- tangled masses of data
- traditional data collection methods and classical (small) data analysis techniques may not be sufficient anymore

WEB DATA SCRAPING EXAMPLE: NEW PHONE

Let’s say you want to know what people think of a new phone. Standard approach: market research (e.g. telephone survey, reward system, etc.).

Pitfalls:

- *unrepresentative sample*: the selected sample might not represent the intended population
- *systematic non-response*: people who don’t like phone surveys might be less (or more) likely to dislike the new phone
- *coverage error*: people without a landline can’t be reached, say
- *measurement error*: are the survey questions providing suitable info for the problem at hand?

WEB DATA SCRAPING EXAMPLE: NEW PHONE

These solutions can be **costly**, **time-consuming**, **ineffective**.

Proxies are indicators that are strongly related to the information of interest, without measuring it directly.

If **popularity** is defined as large groups of people preferring one product over a competitor, then sales statistics on a commercial website may provide a proxy for popularity.

Rankings on Amazon could provide a more **comprehensive** view of the phone market than a traditional survey.

WEB DATA SCRAPING EXAMPLE: NEW PHONE

Representativeness of the **listed products**

- are all phones listed?
- if not, is it because that website doesn’t sell them?
- is there some other reason?

Representativeness of the **customers**

- are there specific groups buying/not-buying online products?
- are there specific groups buying from specific sites?
- are there specific groups leaving/not-leaving reviews?

Truthfulness of customers and **reliability** of reviews.

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

AUTOMATED DATA COLLECTION CHECKLIST

Is **web scraping** really necessary?

Criteria:

- do you plan to repeat the task from time to time e.g. to update your database?
- do you want others to be able to replicate your data collection process?
- do you deal with online sources of data frequently?
- is the task non-trivial in terms of scope and complexity?
- if the task can be done manually, do you lack the resources to let others do the work?
- are you willing to automate the process by means of programming?

If most answers are “Yes”, then automated collection may be the right choice.

GAC example: collecting data at mission from social media to see trends in local opinion.

DATA COLLECTION PROCESS

1. **Know exactly what kind of information you need**
   - Specific: sales of top 10 shoe brands in 2017
   - Vague: people's opinion on shoe brand X

2. **Find web data sources that could provide direct/indirect information**
   - Easier for specific facts: shoe store's webpage provides information about shoes that are currently in demand, such as sandals, boots, etc.
   - Tweets may contain opinion trends on anything
   - Commercial platforms can provide information on product satisfaction

DATA COLLECTION PROCESS

3. **Develop a theory data generation processes for potential sources**
   - When was the data generated?
   - When was it uploaded to the Web?
   - Who uploaded the data?
   - Are there any potential areas that are not covered? consistent? accurate?
   - How often is the data updated?

DATA COLLECTION PROCESS

4. **Balance advantages and disadvantages of potential data sources**
   - Validate the quality of data used
   - Are there independent sources that provide similar information?
   - Can you identify original source of secondary data?

5. **Make a decision**
   - Choose data source that seems most suitable
   - Document reasons for this decision
   - Collect data from several sources to validate data sources

DATA QUALITY

Questions:

- what type of data is most suited to answer the questions?
- is the quality of the data sufficiently high to answer the questions?
- is the information systematically flawed?
- is the data being used because “it’s the best data we have”?

Data quality depends on the **application**.

- a sample of tweets collected on a random day could be used to analyze the use of a hashtags or the gender-specific use of words
- not as useful if collected during Game 7 of the Stanley Cup Finals (**collection bias**)

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**.

WEB SCRAPING LEGALITY

What is a spider?

- Programs that graze or crawl the web for information rapidly
- Jumps from one page to another, grabbing the entire page content

Web scraping requires taking specific information from specific websites (which is the stated goal): how is that **different** from a spider?

“Scraping inherently involves **copying**, and therefore one of the most obvious claims against scrapers is copyright infringement.”

WEB SCRAPING LEGALITY

Crawling another company's information to process and resell it is a common complaint.

Ethical Guidelines:

- work as transparently as possible
- document data sources at all time
- give credit to those who originally collected and published the data
- if the data is collected by another agency, get permission to reproduce it
- don't do anything illegal

DATA AND KNOWLEDGE MODELING

DATA COLLECTION AND DATA MANAGEMENT

CONTEXTUAL METADATA

Something is lost when we move from conceptual models to either a data or a knowledge model.

One way of keeping the context is to provide rich **metadata** – data **about** the data.

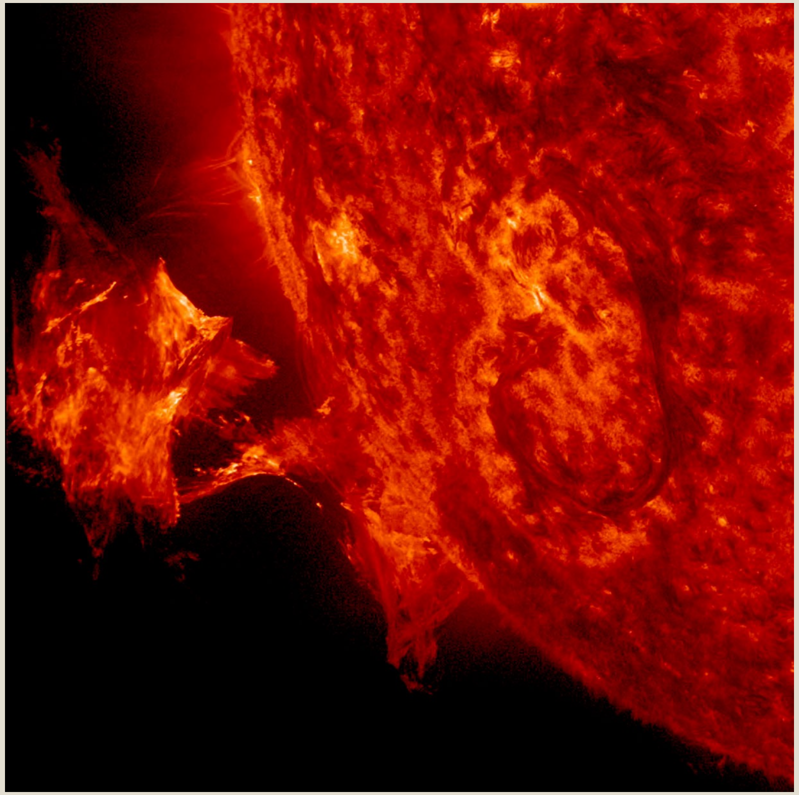
Metadata is crucial when it comes to carrying out strategies for working across datasets.

Ontologies can also play a role here.

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



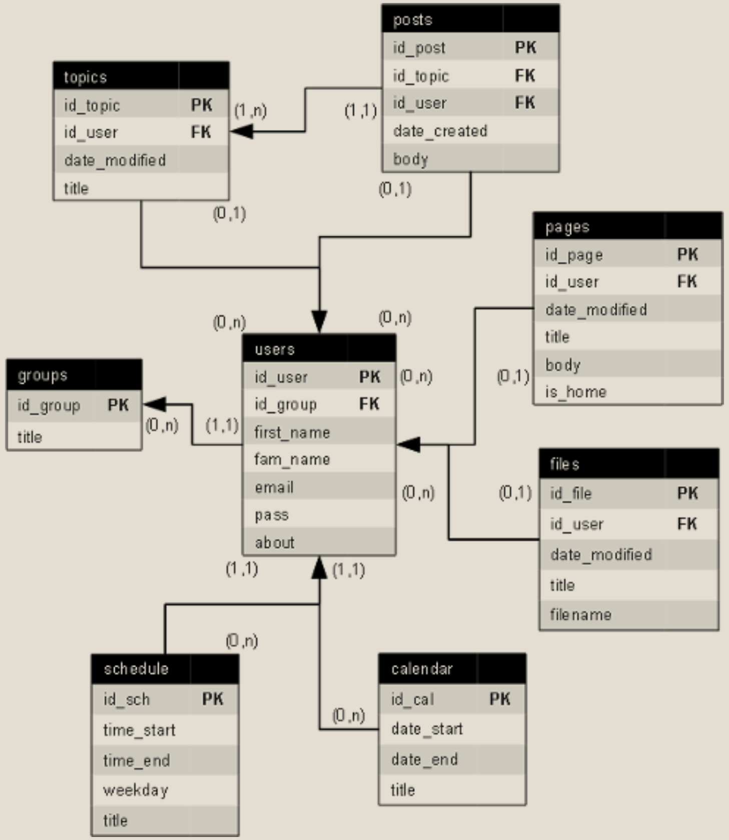
RELATIONAL DATABASES

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.



STORES AND DATABASES

Relational Database:

- widely supported, well understood, works well for many types of systems and use cases, difficult to change once implemented, doesn't deal with relationships well

Key-Value Stores:

- can take any sort of data, no need to know much about its structure in advance, missing values don’t take up space, can get messy, difficult to find specific data

Graph Databases:

- fast and intuitive for heavily relation-based data, might be the only option in this case as traditional databases may slow to a crawl, probably overkill in other cases, not yet widely supported

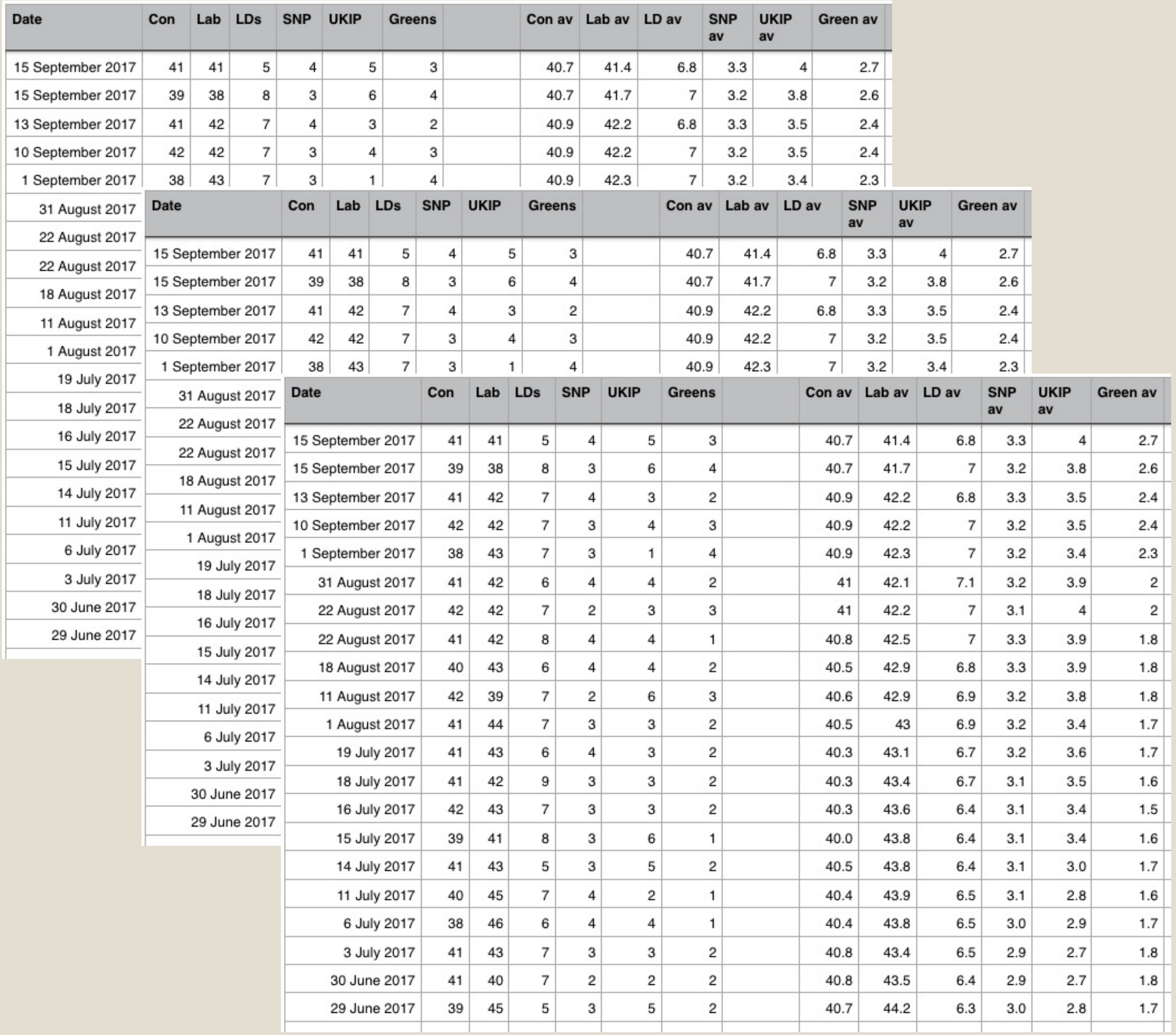
FLAT FILES AND SPREADSHEETS

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

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													
22 August 2017													
22 August 2017													
18 August 2017													
11 August 2017													
1 August 2017													
19 July 2017													
18 July 2017													
16 July 2017													
15 July 2017													
14 July 2017													
11 July 2017													
6 July 2017													
3 July 2017													
30 June 2017													
29 June 2017													

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													
22 August 2017													
22 August 2017													
18 August 2017													
11 August 2017													
1 August 2017													
19 July 2017													
18 July 2017													
16 July 2017													
15 July 2017													
14 July 2017													
11 July 2017													
6 July 2017													
3 July 2017													
30 June 2017													
29 June 2017													

FLAT FILES AND SPREADSHEETS

Pros:

- very efficient if collecting data only once, about one particular type of object
- some types of analysis require all the data in one place
- easy to read into analysis software and do operations over the entire dataset

Cons:

- very hard to manage data integrity if continually collecting data
- not ideal for system data involving multiples types of objects and relationships
- can be very difficult to carry out data querying operations

DATABASE MANAGEMENT

Once data has been collected, it must also be **managed**.

Fundamentally, this means that the database must be maintained, so that the data is

- accurate,
- precise,
- consistent
- complete

Don’t let your data lake turn into a data swamp!