MODULE 1: DATA FOUNDATIONS

CT ACADEMY | DATA ACTION LAB

4. DATA COLLECTION

DATA FOUNDATIONS

EXERCISE

In groups, answer the following questions:

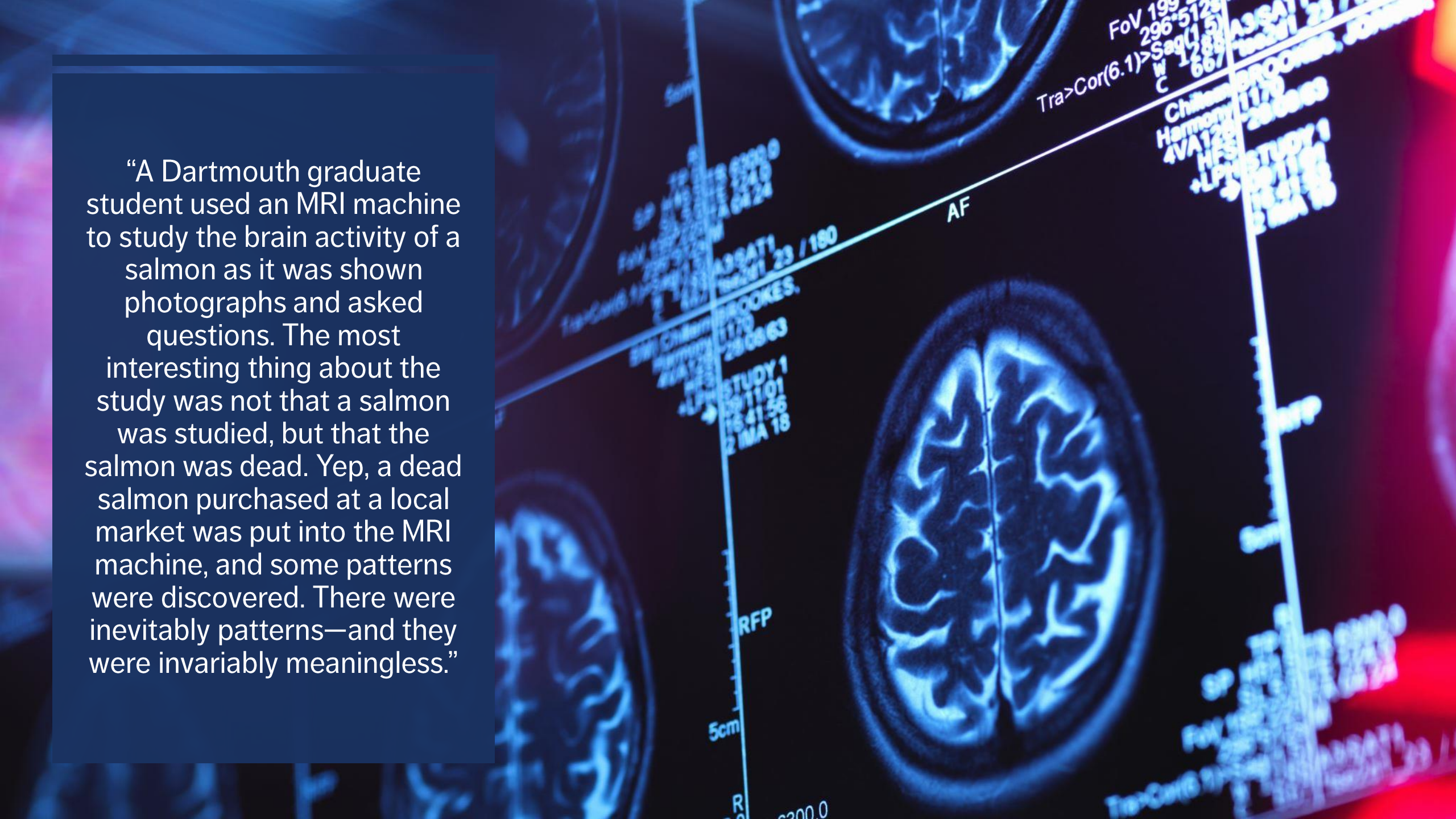
1. does your group create or generate data? if so, what data?
2. do you use data from external sources? if so, which ones?
3. how many sources of data (e.g., databases) does your group use, roughly speaking?
4. do you publish analysis of data internally to your group? externally? both?

THE GOAL OF GOOD STUDY/SAMPLING DESIGN

We need data that can:

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

This starts with **study design** – what data to collect and how it should be collected



"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."

PATTERN FISHING / NON-PROBABILISTIC SAMPLING

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 AND SURVEYS

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.

SAMPLING MODELS

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.

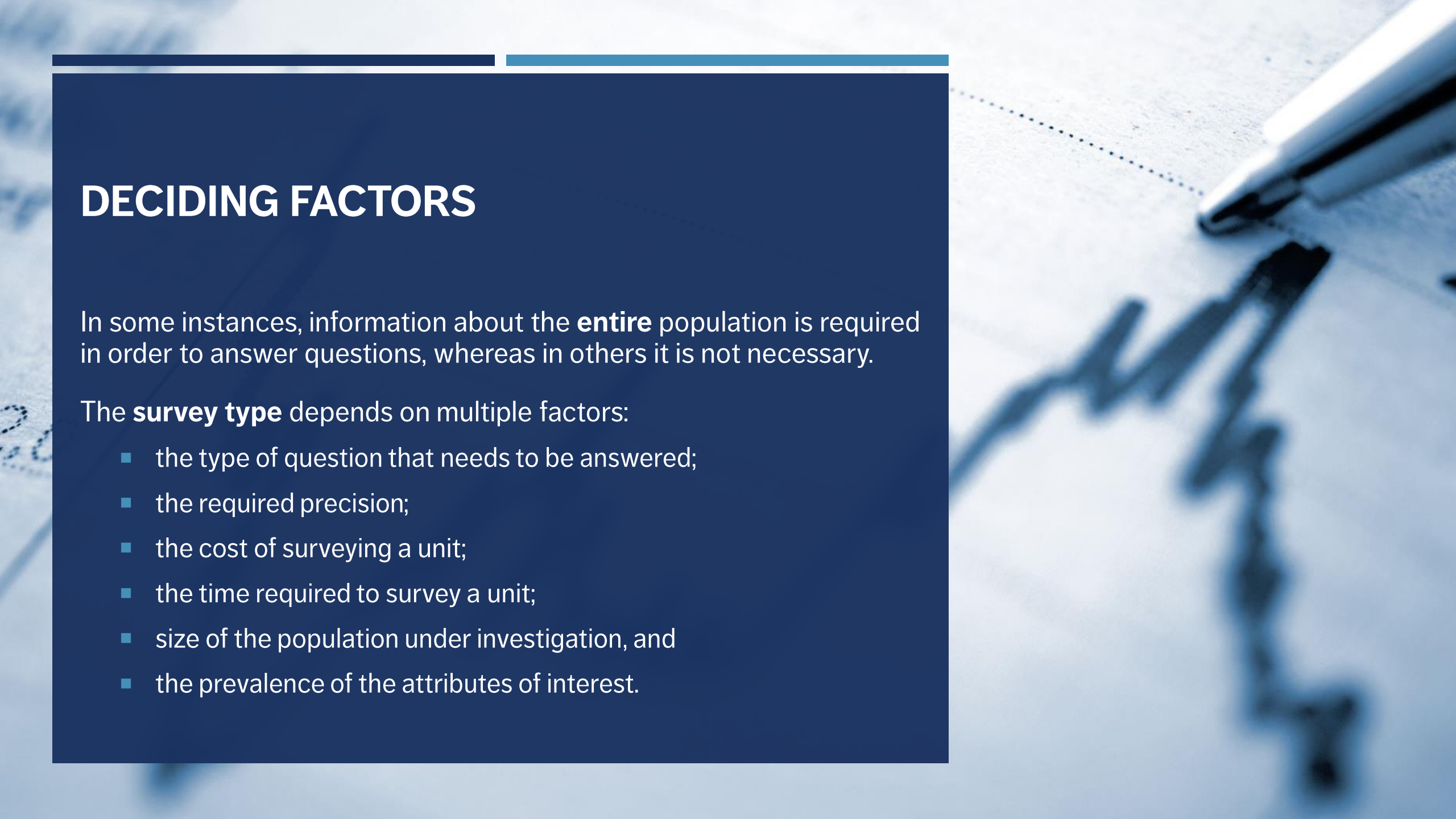
When survey sampling is done properly, we may be able to use various **statistical methods** to make **inferences** about the **target population** by sampling a (comparatively) small number of units in the **study population**.

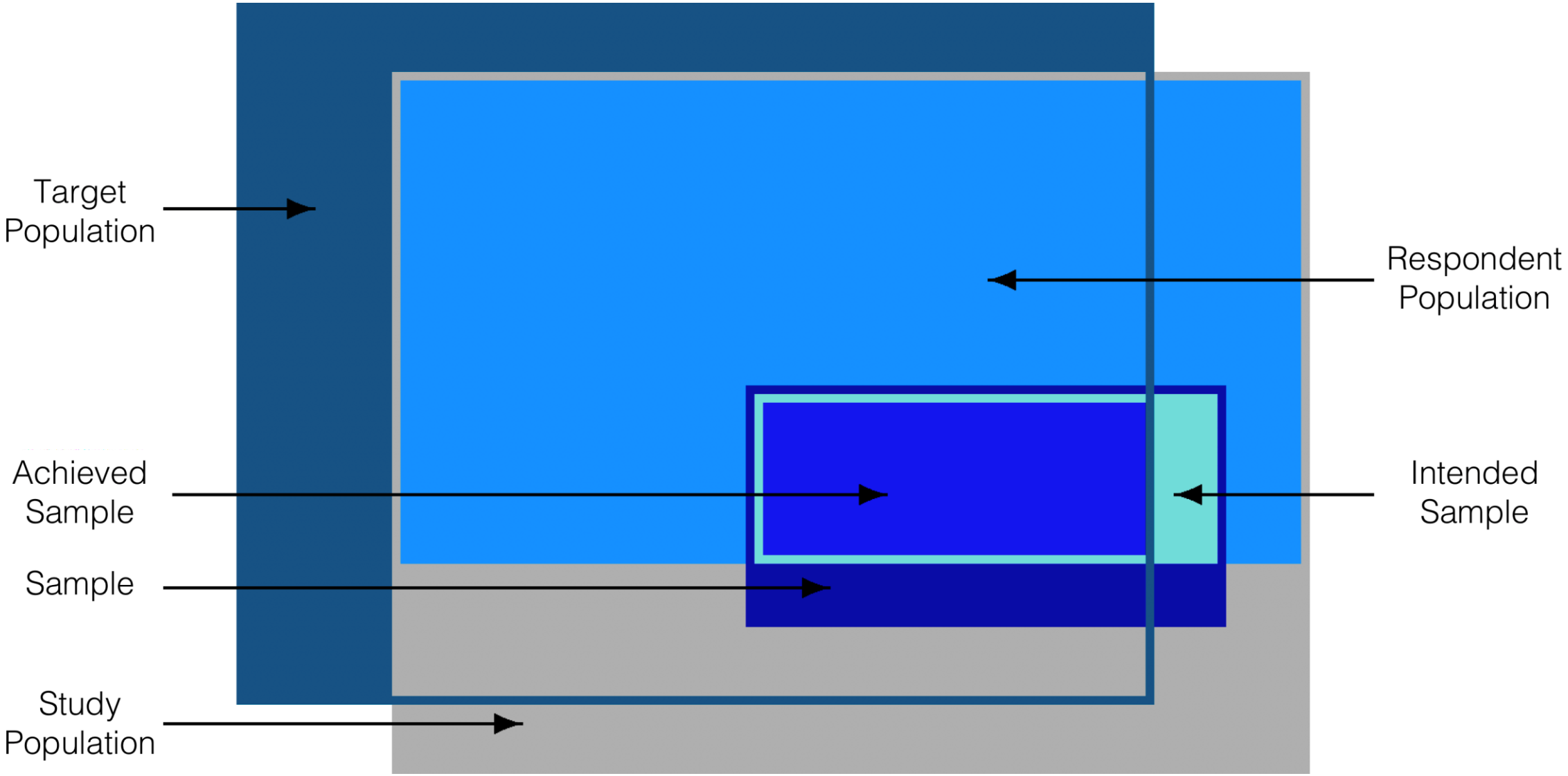
DECIDING FACTORS

In some instances, information about the **entire** population is required in order to answer questions, whereas in others 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 Sample
Intended Sample
Sample
Study Population

SURVEY ERROR

Total Error = Sampling Error + Measurement Error + Non-Response Error + Coverage Error

- Sampling Error: survey, not census
- Measurement Error: observations not measured accurately
- Non-Response Error: non-respondents having systematic observation differences
- Coverage Error: 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 **nonsampling errors** (everything else).

PROBABILISTIC SAMPLING

Probabilistic sample designs are usually more **difficult** and **expensive** to set-up (due to the need for a quality survey 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% CI.

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).

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.

OPEN SOURCE SOFTWARE

Another trend:

- growth and increasing popularity and power of **open source software** (source code can be inspected, modified, and enhanced by anyone).

Community aspect → ever-changing and improving

R and **Python** are open source software that can be used for data analysis in the social sciences and other domains.

They incorporate **interfaces** to other programming languages and software **solutions**.

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.

DATA SOURCES (TRADE-OFFS)

Automated vs. Traditional

Accuracy vs. Completeness

Coverage vs. Validity

Speed vs. Cost

etc.

DATA COLLECTION PROCESS (5 STEPS)

1. Know exactly what kind of information you need

- Specific: GDP of all OECD countries for last 10 years; sales of top 10 shoe brands in 2017
- Vague: people's opinion on shoe brand X

2. Find out if there are any web data sources that could provide direct or indirect information on your problem

- Easier for specific facts: shoe store's webpage will provide information about shoes that are currently in demand i.e. sandals, boots, etc.
- Tweets may contain opinion trends on *anything*
- Commercial platforms can provide information on product satisfaction

DATA COLLECTION PROCESS (5 STEPS)

3. Develop a theory of the data generation process when looking into 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 (5 STEPS)

4. Balance advantages and disadvantages of potential data sources

- Validate the quality of data used
- Are there other independent sources that provide similar information to crosscheck against
- 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

EXERCISE

You must estimate the yearly salary of all GoC financial professionals.

What data sources do you need? What obstacles might stand in your way?

FRIENDLY COOPERATION WITH API

Application program interface (API) are sets of routines, protocols, and tools for building software applications.

Many APIs restrict the user to a certain amount of API calls per day (or some other limits).

These limits should be obeyed.

CONTACT DATA PROVIDERS

Any data accessed by HTTP forms is stored in some sort of database.

Ask proprietors of the data first if they will grant access to the database or files.

The larger the amount of data you want, **the better it is for both parties to communicate before starting to harvest data**.

For small amounts of data, that's less important.

EXERCISE

Identify 3 ways in which your organization collects data.

For each way, identify potential issues with how the data is collected and how you manage those issues (if any issues exist).

SUPPLEMENTAL MATERIAL

4. DATA COLLECTION

STUDY/SURVEY STEPS

Studies or 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 ideal frame contains identification data, contact data, classification data, maintenance data, and linkage data, and must minimize the risk of **undercoverage** or **overcoverage**, as well as the number of duplications and misclassifications (although some issues that arise can be fixed at the data processing stage).

A statistical sampling approach is contraindicated unless the selected frame is

- **relevant** (that is, 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**.

MODES OF DATA COLLECTION

Paper-based vs. computer-assisted

- **self-administered questionnaires** are used when the survey requires detailed information to allow the units to consult personal records; associated with high non-response rate.
- **interviewer-assisted questionnaires** use well-trained interviewers to increase the response rate and overall quality of the data; face-to-face vs. telephone.
- **computer-assisted interviews** combine data collection and data capture, which saves time.
- unobtrusive direct observation
- diaries to be filled (paper or electronic)
- omnibus surveys, email, Internet, and social media

NONSAMPLING ERROR

Nonsampling error can be controlled, to some extent:

- **coverage error** can be minimized by selecting high quality, up-to-date survey 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 (landline-based survey frames are becoming irrelevant due to demographics, response rates for surveys that are not mandated by law are low, etc.).

NONPROBABILISTIC 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 survey frame required).
- NPS methods are ideal for exploratory analysis and survey development.

Unfortunately, NPS are often used instead of probabilistic designs (problematic)

- the associated selection bias makes NPS methods unsound when it comes to inferences (they cannot be used to provide reliable estimates of the sampling error, the only component of TE under the analyst's direct control);
- automated data collection often fall squarely in the NPS camp – we can still analyze data collected with a NPS approach, but may not generalize the results to the target population.

NPS METHODS

Haphazard

- man on the street, depends on availability of units and interviewer bias

Volunteer

- self-selection bias

Judgement

- biased by inaccurate preconceptions about the target population

Quota

- exit polling, ignores non-response bias

NPS METHODS

Modified

- starts probabilistic, switches to quota as a reaction to high non-response rates

Snowball

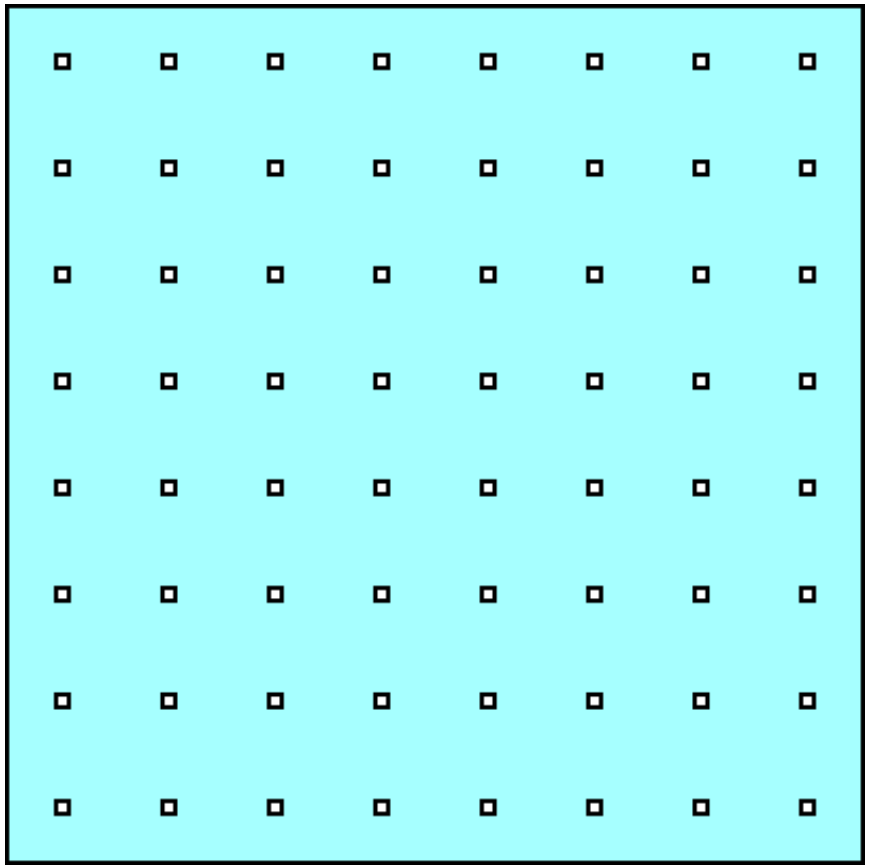
- “pyramid” scheme

There are contexts where NPS methods might fit a client’s or an organization’s need (and that remains their decision to make, ultimately), but they must be informed of the drawbacks, and presented with some probabilistic alternatives.

SAMPLING UNIVERSE

Goal: estimate the true population attributes μ, σ^2, τ, p *via* the sample population attributes $\bar{y}$, s^2, $\hat{\tau}$, $\hat{p}$, n, and the size N of the target population.

We look for **confidence intervals** (typically 95%).



SAMPLING UNIVERSE

Target population:

- N units and measurements $\mathcal{U} = \{u_1, \dots, u_N\}$

True population attributes:

- mean μ, variance σ^2, total τ, proportion p

Sample population:

- n units and measurements $\mathcal{Y} = \{y_1, \dots, y_n\} \subseteq \mathcal{U}$

Sample population attributes:

- sample mean $\bar{y}$, sample variance s^2, sample total $\hat{\tau}$, sample proportion $\hat{p}$

PROBABILISTIC SAMPLING DESIGNS

Simple random sampling (SRS)

Stratified random sampling (StS)

Systematic sampling (SyS)

Cluster sampling (ClS)

Replicated sampling (ReS)

Multi-stage sampling (MSS)

Multi-phase sampling (MPS)

Probability proportional-to-size sampling (PPS)

SIMPLE RANDOM SAMPLING (SRS)

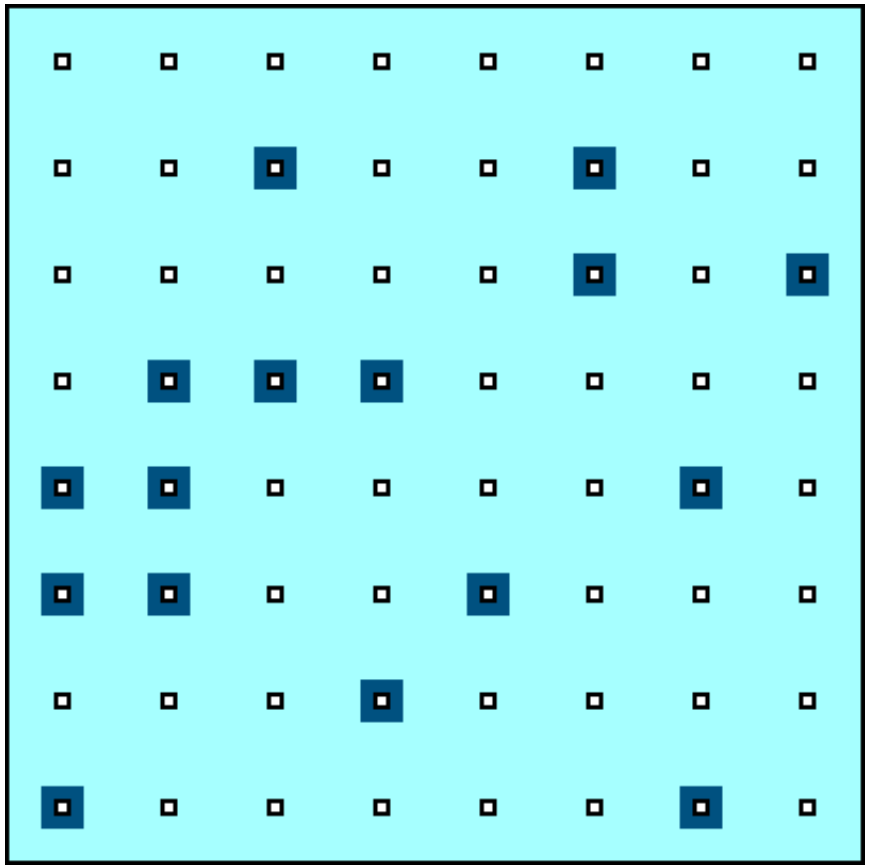
In SRS, we select n units randomly from the frame.

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)

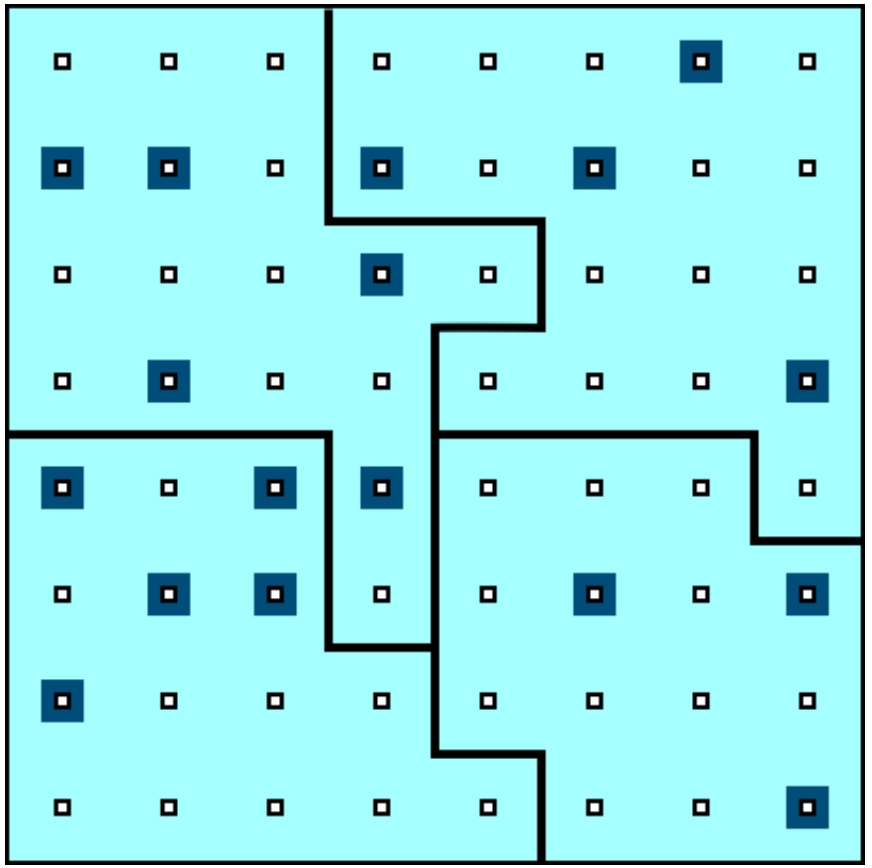
In StS, $n = n_1 + \cdots + n_k$ units are randomly drawn from k **strata**.

Advantages:

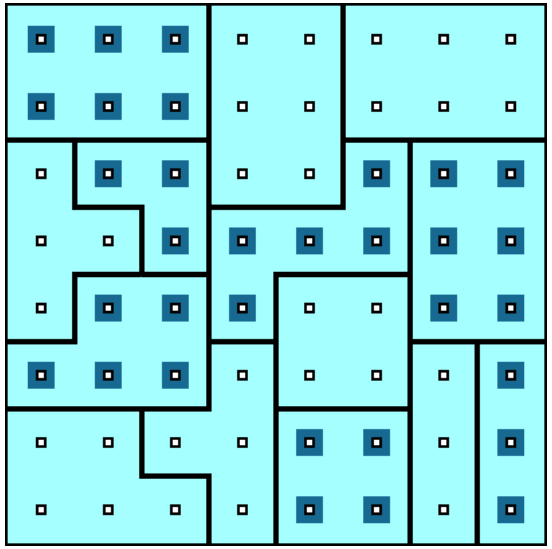
- may produce smaller error bound on estimation than SRS
- may be less expensive if elements are conveniently strat.
- may provide estimates for sub-populations

Disadvantages:

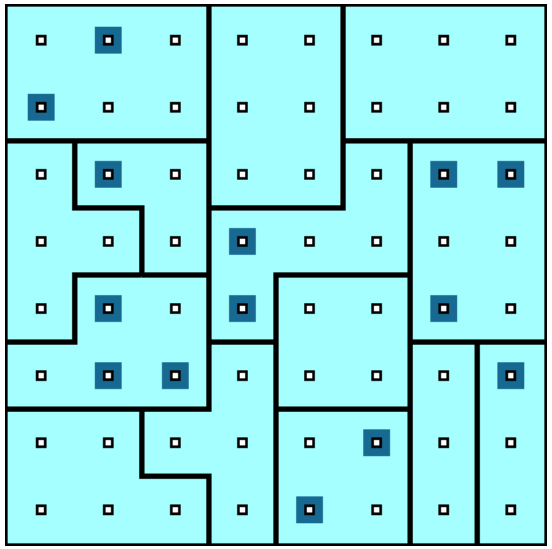
- no major disadvantage
- if there are no natural ways to stratify the frame into homogeneous groupings, StS is roughly equivalent to SRS



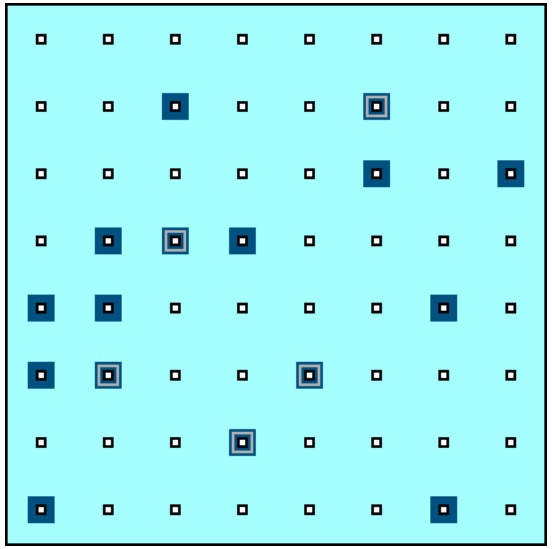
OTHER PROBABILISTIC SAMPLING DESIGNS



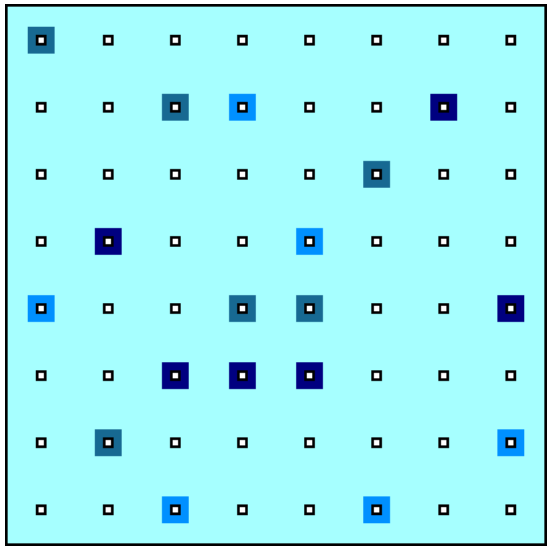
Cluster Sampling (CIS)



Multi-Stage Sampling (MSS)



Multi-Phase Sampling (MPS)



Replicated Sampling (ReS)

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 QUALITY – NEW PHONE

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

Proxies – indicators that are strongly related to the product's popularity, 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 vs. traditional survey.

POTENTIAL ISSUES – 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.

IS WEB SCRAPING LEGAL?

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 you did not collect the information, you probably need permission to reproduce it
- Don’t do anything illegal.

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

IS WEB SCRAPING LEGAL?

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

Scraping is taking specific information from specific websites (which is the goal): how are these **different**?

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

LEGAL CASES – WEB SCRAPING

eBay vs. Bidder's Edge (BE)

- BE used automated programs to crawl information from different auction sites.
- Users could search listings on the BE webpage instead of going to individual auction sites.
- BE accessed eBay's sites ~100 000 times / day (1.53% of # of requests, 1.1% of total data transferred by eBay) in 1999.
- eBay alleged damages of up to $45k- $62K in a 10 month period.
- BE didn't steal information that wasn't public, but excessive traffic was demanding on eBay's servers.
- **Your verdict?**

LESSONS LEARNED

It is not clear which scraping actions are illegal and which are legal.

Re-publishing content for commercial purposes is considered more problematic than downloading pages for research/analysis.

Robots.txt: *Robots Exclusion Protocol* is a file that tells scrapers what information on the site may be harvested.

Be friendly! Not everything that can be scraped needs to be so. Scraping programs should behave “nicely”, provide the data you seek, and be efficient, in this order.

SCRAPING DO'S AND DON'T'S

1. Stay identifiable

2. Reduce traffic

- Accept compressed files
- If scraping the same resources multiple times, check first if it has changed before accessing again
- Retrieve only parts of a file

SCRAPING DO'S AND DON'T'S

3. Do not bother server with multiple requests

- Many requests per second can bring smaller servers down
- Webmasters may block you if your scraper behaves this way
- One or two request per second is fine

4. Write modest scraper (efficient and polite)

- No reason to scrape pages daily or repeat same task over and over; make your scraper as efficient as possible
- Do not over-scrape pages
- Select resources you want to use and leave the rest untouched