MODULE 1: DATA FOUNDATIONS

CT ACADEMY | DATA ACTION LAB

1. DATA AWARENESS

DATA FOUNDATIONS

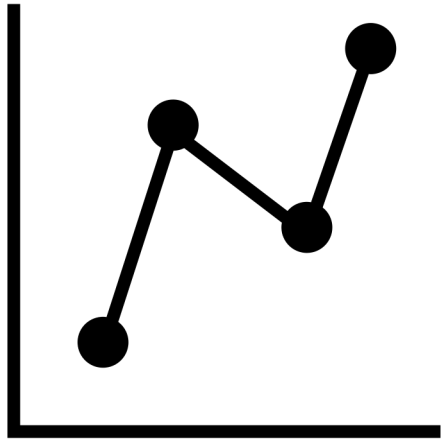
WHAT IS (ARE) DATA?

Should we say, “is data” vs “are data”?

The word data is technically a plural, so “are” is appropriate; the singular is the word **datum** (for a “data point”).

In common usage the word data is used interchangeably (datum has become *passé*): “are” is technically correct but “is” is used all the time!

Being pedantic, if “data” is used as a **mass noun**, then “is” IS appropriate!



PURCH_DOC	AWARD_DATE	SP	ORIGINAL_V	AMENDMENTS	TOTAL_PO_A	NAME
24XXXXXXXX	2015-11-23	TN	24814.30	0.00	24814.30	Canada
24XXXXXXXX	2015-11-23	TN	11327.58	4674.08	16001.66	Canada
24XXXXXXXX	2015-11-23	TN	4860.00	0.00	4860.00	Canada
24XXXXXXXX	2016-05-06	TN	52000.00	0.00	52000.00	Canada
24XXXXXXXX	2014-09-02	TN	23748.68	0.00	23748.68	Canada
24XXXXXXXX	2014-07-24	TN	15943.55	0.00	15943.55	Canada
24XXXXXXXX	2014-07-24	TN	20336.79	0.00	20336.79	Canada
24XXXXXXXX	2014-10-07	TN	29286.40	0.00	29286.40	Canada
24XXXXXXXX	2016-07-28	TN	13800.00	0.00	13800.00	Canada

WHAT IS (ARE) DATA?

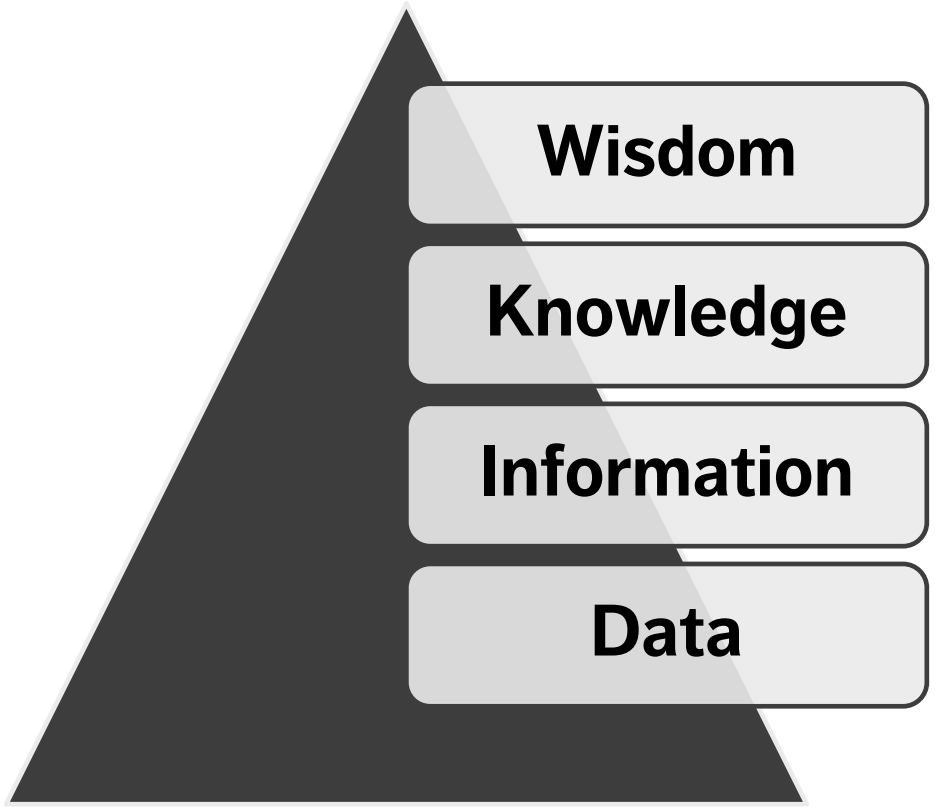
Data can be thought of as **raw "numbers".**

It is often defined as **"a collection of facts from which conclusions may be drawn"**.

Data comes in many different forms and underpins all analyses.

(We will revisit these notions)

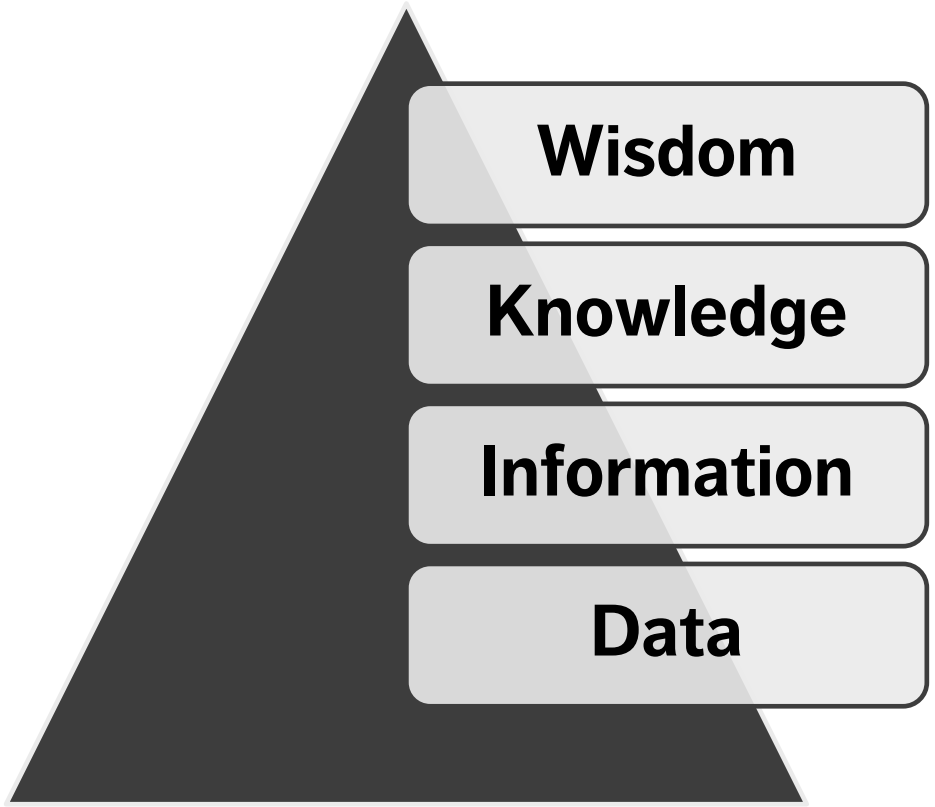
DATA IS A FOUNDATION



DIKW Pyramid

- represents **structure** or functional **relationship** between elements
- we **acquire** data
- **organizing** data gives us information
- using information in **context** yields knowledge
- the correct (and/or incorrect) **application** of information over time makes us wise!

DATA IS A FOUNDATION



DIKW Pyramid Example

- **Data** – individual bank transactions
- **Information** – organizing the transactions into monthly groups (what are my monthly spendings?)
- **Knowledge** – comparing the "spend per month" to a budget
- **Wisdom** – understanding that if I am under budget (over time) I can save up for a vacation!

HOW WE USE DATA

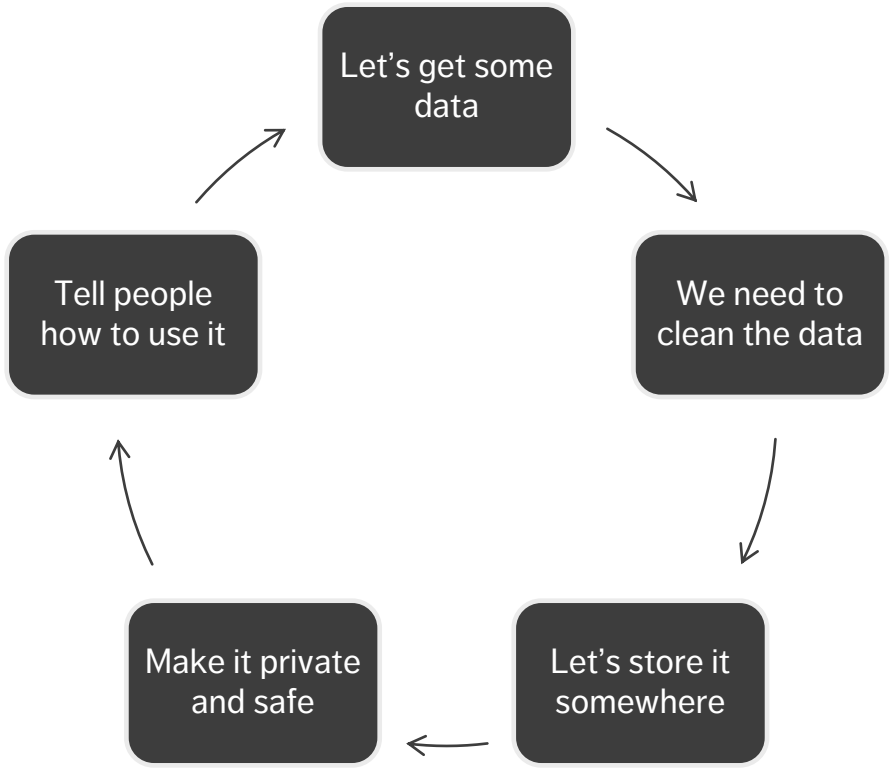
We **analyse** data to tell **stories** that help us make **decisions.**

This requires **data integrity**.

Most departments have **data stewards** who manage the **data lifecycle** (how to acquire, manage, and use data).

Data stewards keep the data in **data assets** and make them available for people to use as they require.

Analysts must **contextualize** the data to help them make effective decisions.



HOW WE USE DATA

What do you do once you have data?

- identify what decisions you, or your team (or your boss) needs to make
- identify the data that you need to make that decision
- get the data, check it to make sure it's ok
- you can then do one of more of the following: **analyze** the data, **visualize** the data, **summarize** the data
- turn the data into **information** and **knowledge**
- make your decisions or provide your outputs to the stakeholders who need them

DATA RELATED TERMINOLOGY

Here are a few common data **buzzwords**:

Term	Description	Example(s)
metadata	data about data	a document that describes what column titles mean in a spreadsheet
data asset	a system or program that stores data	Excel spreadsheets, SAP, PeopleSoft, Access Databases, Web Tables
reference data	data that is common across data assets	list of Provinces, list of countries, list of branches in a department
master data	data that we use to run our business	employee names, transaction amounts
data inventory	a list of data assets	tools like Microsoft Purview maintain lists of data assets
data catalog	descriptions of data	precisely defines words the business uses, e.g., “FTE”, “Headcount”
data model	how data interrelates	“linking” together financial and HR data through a PRI

INTRODUCING ROLES & RESPONSIBILITIES

Who uses data?

- if you are using data, you are a **data consumer**
- if you a responsible for the integrity of the data you are a **data steward**
- **data trustees** are accountable for the data
- If you input data into a system, or acquire it from somewhere and add it to a data asset you are known as a **data contributor**
- if you help to manage the systems in which the data resides, you are a **data custodian**

EXAMPLE : USE OF DATA – HOW MUCH MONEY DO WE HAVE?

We need to see how much money is available in the department ("**free balance**"). We run a report from the relevant data asset, getting the data on what we:

- have spent up to this time (the "**actuals**");
- have committed to spend (the "**commitments**"), and
- think we should have spent (the "**budget**").

Data stewards check if there are any problems with the data and fix them any such problems.

Finance asks business to validate the amounts to see if they are accurate and that nothing is missing.

Finance then applies the following formula and stores the result for its financial reporting obligations:

$$\text{Free Balance} = \text{Budget} - (\text{Actuals} + \text{Commitments})$$

EXAMPLE: USE OF DATA – MAKING DEPARTMENTS SAFER

The occupational health and safety group at a department wants to make our environment safer.

Every time a health and safety incident occurs, the **details are recorded** (type of incident, when/where/how it happened, etc.).

This data is **tracked** and **analyzed**.

If trends are seen in the data (for example a lot of slips and trips happen at a particular location) then the team **decides to intervene**, and steps are taken to **mitigate the issue** (e.g., coating the floor with a non-slip surface, etc.).

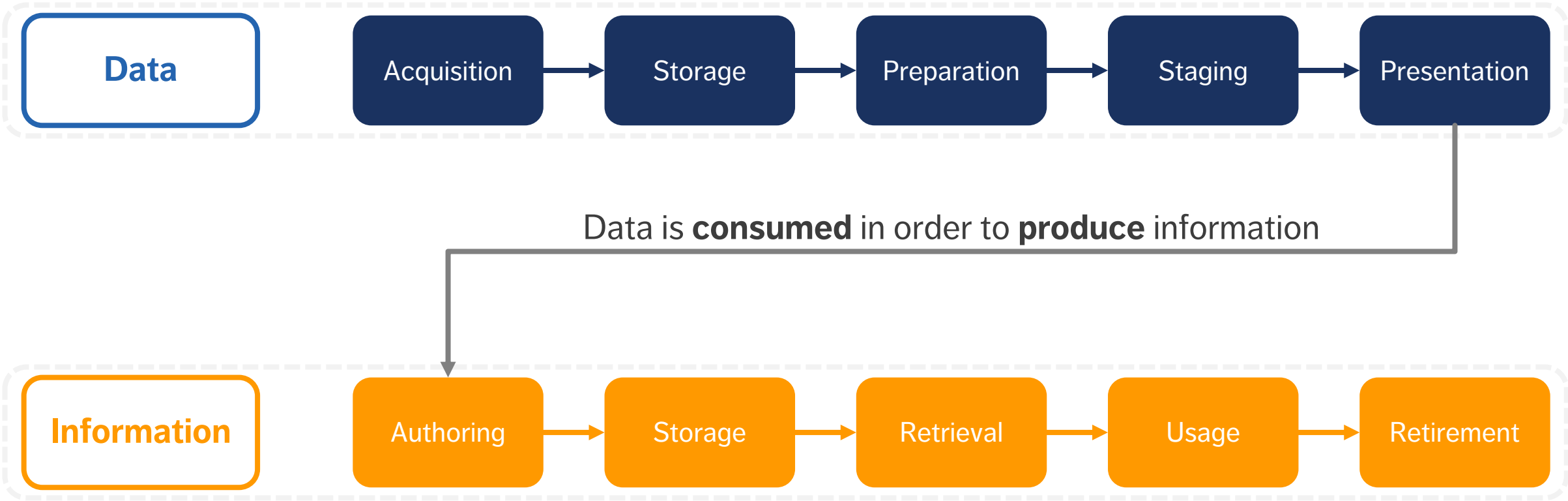
More data is collected after mitigation so that the group can measure "**improvements**".

DATA AND INFORMATION LIFE CYCLES

Data and information "**life cycles**" describe all the steps that happen between data collection and data not being needed anymore.

As data **gets turned into** information, we need to really understand both of these life cycles (they are different).

DATA AND INFORMATION LIFE CYCLES



DATA AS A STRATEGIC ASSET

In a previous version of the Privy Council Data Strategy, "**data as an asset**" was a foundational pillar, defined as:

> "The government has the data it needs, which are fit for use, discoverable, and available. Includes, for example:
>
> - planning and stewardship;
> - use;
> - quality;
> - storage, and
> - sharing and access."

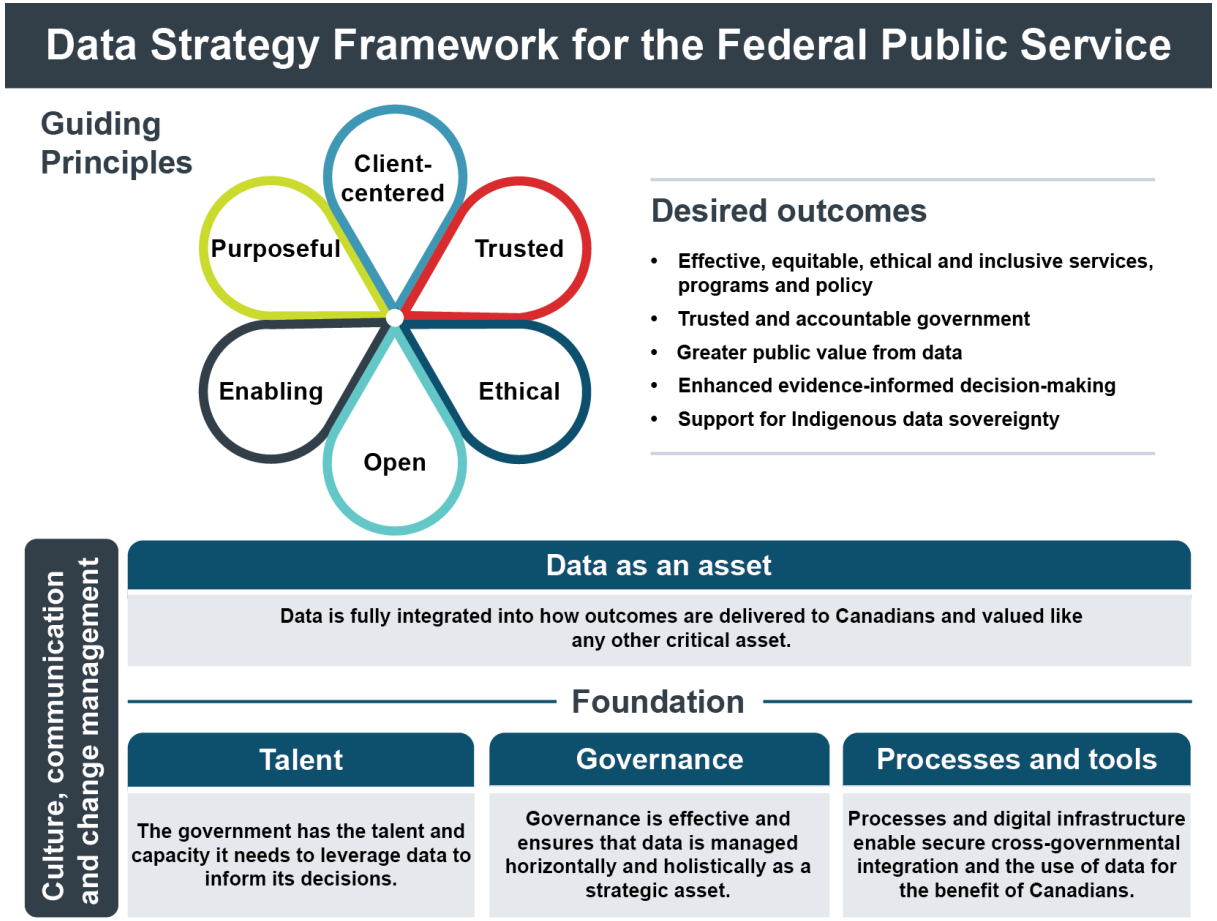
GOVERNMENT OF CANADA DATA PICTURE

The GoC has released a refresh on their original **2018 data strategy**.

All departments are required to implement its policies, directives, and procedures.

This is typically the responsibility of a "**Chief Data Steward**".

In this new model "Data as a Strategic Asset" is now an **output** of the implementation of the guiding principles.



DATA LITERACY

To support the data strategy, the GoC requires that GoC employees be **data literate**. Supporting data literacy is the **GoC Data Competency Framework** that this set of courses is built around.

> "Having a data literate workforce is at the core of modernization efforts. This Data Competency Framework is meant to support conversations and aims to advance data literacy by creating a shared understanding and language about data competencies for all federal public servants."

A department's level of data literacy is usually identified through **surveys**; gaps are addressed through **training**, **education**, **mentorship**, and other learning methods.

DATA LITERACY

The Data Competency Framework consists of four sections that are divided into three proficiency levels:

Sections:

1. Data Concepts and Culture
2. Data Governance, Collection, and Stewardship
3. Analytics and Evaluation
4. Data Systems and Architecture

Proficiency Levels:

1. Foundational: defining the core level of understanding and awareness
2. Intermediate: putting theory into practice
3. Advanced: applications and enabling others

DATA ROLES & RESPONSIBILITIES

Data roles and responsibilities are “required” by TBS as part of GoC Data Strategy.

They align business operations with **data governance** activities, helping managers and supervisors to define and assign **accountability** & **responsibility** to employees.

Explicit R&R help GoC employees understand how they fit in with their department’s data activities.

Note that one person can be assigned **multiple R&R** at the same time (it is possible to be both a data contributor and a data consumer simultaneously!)

HIGH LEVEL REVIEW OF DIFFERENT R&R

Data Trustees

- ensure strategic management of assigned data assets as well as compliance with departmental and enterprise data (**related strategies**, **regulations**, **policies**, **directives**, and **standards**)
- executives with business accountability and intermediate level of technical knowledge (typically at **director level**)

Data Stewards

- advise on, enact, and help enforce **data policies** and **standards**
- **operations-focused**
- they have a mix of business and technical background (**branch representatives**)

HIGH LEVEL REVIEW OF DIFFERENT R&R

Data Custodians

- ensure the safe **custody** and **integrity** of hosted data, and safeguard the enterprise data repository
- normally **operations-focused**, with a technical background (**IT-focused**)

Data Contributors

- ensure that the data they provide to the Department (including third-party data) aligns with all technical and business **policies**, **procedures**, and **standards**
- **operations-focused**, with a "**business**" background and some technical expertise for the systems they typically use

HIGH LEVEL REVIEW OF DIFFERENT R&R

Data Consumers

- ensure that usage of data **supports** departmental and government objectives and mandates
- anyone within the organization can play that role, typically with a "**business**" background

WHERE DO YOU FALL?

Where do you fit? You may have been assigned a specific role (e.g., data steward), but regardless of role assignment it is highly probable that you are a **data consumer**. If you do any of the following you can count yourself as part of that role:

- exporting data from any system;
- create a spreadsheet that people use to make decisions;
- get data from outside of the department and use it in internal or external reports, etc.

You may also be a **data contributor** if you do things like:

- do research and gather data that adds to corporate knowledge;
- enter data into a system (e.g., call center employee entering case data);
- entering overtime into a salary system, etc.

Some roles are **assigned**, some are **inherited** (based on what you do).

EXAMPLES: R&R

Example: Call Center Agent

You take your first call of the day, the information from the call is entered in the call center system (**data contributor**).

A list of calls appears on your screen. You prioritize the callers on the list and select your next call (**data consumer**).

You realized that you entered a wrong piece of data that you can't overwrite, you call your data steward for them to fix it (**data consumer**).

You have a responsibility to review your team call performance for the day and provide feedback to your supervisor. You download a system extract and do the calculations in Excel before forwarding the results (**data contributor**).

EXAMPLES: R&R

Example: Program Director

You are providing a new service as part of the program you run. You review and approve a new database to track the program data (**data trustee**).

Once the database is up and running, you start to review and make decisions on the reports obtained from it (**data consumer**).

Your data steward identifies a major issue that is escalated to you for approval (**data trustee**).

You request that the system be integrated into an existing system (**data trustee**).

You use the reports from the new integrated system to help you administer your program (**data consumer**).

EXAMPLES: R&R

Example: IT Support Technician

You get a call from a data steward wanting you to update a business rule in a database (**data custodian**).

You get the information you need and apply the rule (**data custodian**).

You then export data from the database into a report so you can check that the change was correctly implemented (**data consumer**, **data steward**).

You then update the system with the information and close the ticket (**data contributor**).

2. DATA ETHICS

DATA FOUNDATIONS

MOTIVATION AND POLICY DRIVERS

Unethical and irresponsible handling of data assets and A.I. can have a broad range of consequences:

- Decision-making and policies resulting in harms (e.g., stigma, financial loss, etc.) to individuals and communities
- Violations of personal privacy
- A.I. models that are difficult to understand and can behave in unintended manners
- Loss of public trust, hindering the ability to meaningfully engage with Canadians

Policy drivers for ethical handling of data:

- Federal Data Strategy Roadmap / 2023–2026 Data Strategy for the Federal Public Service
- Departmental Data Strategy – Ethical use of data as an asset
- Canada's Digital Ambition 2022

Policy drivers for ethical handling of Indigenous data:

- UNDRIP / UNDA / UNDA Action Plan
- Departmental Reconciliation Strategy
- 2023–2026 Data Strategy for the Federal Public Service

Policy drivers for responsible use of A.I.:

- Federal Responsible A.I. Guiding Principles
- TBS Directive on Automated Decision-Making
- Government of Canada Digital Standards: Playbook

PRINCIPLES AND STANDARDS

- Privacy Act
- Statistics Act
- Policy on Government Security
- Policy on Privacy Protection
- Privacy Impact Assessments
- Levels of Security
- Gender-Based Analysis+
- Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans
- Model Policy on Scientific Integrity
- Directive on Automated Decision-Making
- Disaggregated Data

COMMONLY USED TERMS

- Data ethics
- Governance
- Consent
- Bias and discrimination
- Inclusiveness
- Fairness
- Accountability

WHAT ARE ETHICS?

Broadly speaking, ethics refers to the study and definition of right and wrong conducts.

We all have a personal ethical system, don't we?

- be honest
- be fair
- be objective
- be responsible
- be compassionate
- etc.

WHAT ARE ETHICS?

Influential *Western* ethical theories:

- Kant's **golden rule** (do unto others as you would have them do unto you),
- **consequentialism** (the ends justify the means)
- **utilitarianism** (act in order to maximize positive effect)

Influential *Eastern* ethical theories:

- **Confucianism** (virtue from people and motives, not from outcomes)
- **Taoism** (case-by-case appropriateness of action determines morality)
- **Buddhism** (harmony and self-restraint to avoid causing harm)

WHAT ARE ETHICS?

Ubuntu ethical tradition:

- **tension** between individual and universal rights
- **global** context of life
- **solidarity**

Maori *tikanga*:

- connection with **spiritual** realm
- **respect** for all things
- **self-determination** and **reciprocity**

WHAT ARE DATA ETHICS?

Data ethics is a branch of ethics that evaluates data practices, including the **collection**, **generation**, **analysis**, and **dissemination** of data, that have the potential to adversely impact people and society.

Mission 3 (Enabling Data-Driven Services) of the 2023-26 data strategy refresh from TBS, states that GoC entities will ensure...

> "... the responsible, ethical and transparent sharing and use of data are key to enabling the delivery of better services to people in Canada."

THE NEED FOR ETHICS

When large scale data collection first became possible, there was to some extent a "**Wild West**" mentality to data collection and use. Whatever wasn't proscribed from a technological perspective was allowed (if not mandatory).

Now, however, professional codes of conduct are being devised, for example, for data scientists, which outline responsible ways to practice data science – i.e., ways that are **legitimate** rather than fraudulent, as well as **ethical**, rather than unethical.

THE NEED FOR ETHICS

Although this puts some **extra** responsibility onto data scientists, it also provides them with protection from people who hire them to carry out data science in questionable ways – **they can refuse on the grounds that it is against their professional code of conduct**.

Does your organization have a code of ethics for its data scientists or other data professionals? For its employees?

GUIDING PRINCIPLES

The Cambridge Dictionary defines a "**Guiding Principle**" as:

"an idea that influences you very much when making a decision or considering a matter."

For example:

"Equality of opportunity has been the government's guiding principle in its hiring policies."

EXAMPLE OF GUIDING PRINCIPLES

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, **except** where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Isaac Asimov's *3 Laws of Robotics*

BEST PRACTICES

“Do No Harm”: data collected from an individual **should not be used to harm** the individual. This may be difficult to apply in practice.

Informed Consent: covers a wide variety of ethical questions, but mainly:

- individuals must **agree to the collection and use** of their data
- individuals must have a **real understanding of what they are consenting to**, and of **possible consequences** for them and others.

BEST PRACTICES

Respect "Privacy": dearly-held principle. Excessively hard to maintain in the age of constant trawling of the Internet for personal data.

Keep Data Public: another aspect of data privacy – some (all? most? any?) data should be kept **public**.

Opt-In/Opt-Out: informed consent requires the ability to **not consent** (to opt out).

- tacit vs. stated consent

BEST PRACTICES

Anonymize Data: removal of identifying fields from the dataset prior to analysis.

“Let the Data Speak”:

- no cherry picking
- importance of validation
- correlation and causation
- repeatability

"And yes, **transparency is also the trick to protecting privacy**, if we empower citizens to notice when neighbors infringe upon it. Isn't that how you enforce your own privacy in restaurants, where people leave each other alone, because those who stare or listen risk getting caught?"

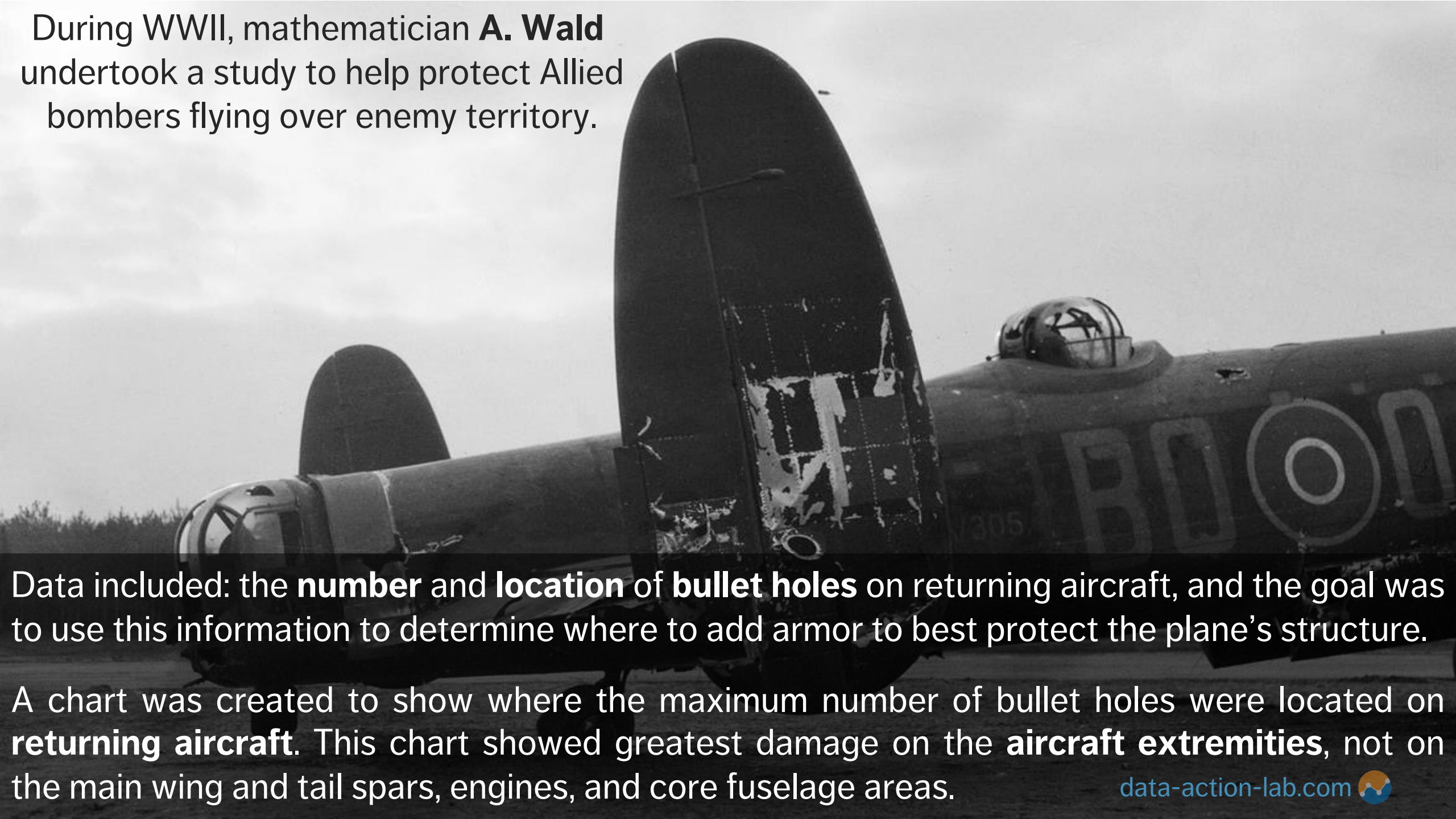
David Brin, *The Transparent Society*

BIAS

A **cognitive bias** is a systematic error in thinking that occurs when people get information in the world around them, and their processing and interpreting of this information affects the decisions and judgments that they make.

The human brain is powerful but subject to imperfections. Cognitive biases are often a result of the brain's attempt to **simplify information processing**. They often work as rules of thumb that help us make sense of the world and reach decisions with relative speed.

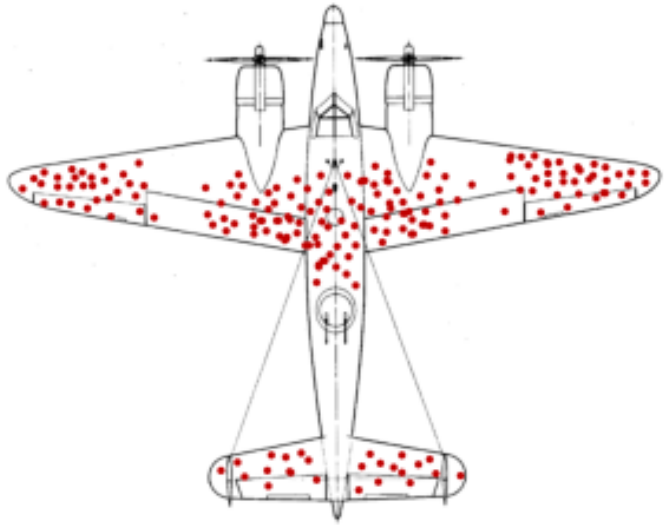
During WWII, mathematician **A. Wald** undertook a study to help protect Allied bombers flying over enemy territory.



Data included: the **number** and **location** of **bullet holes** on returning aircraft, and the goal was to use this information to determine where to add armor to best protect the plane's structure.

A chart was created to show where the maximum number of bullet holes were located on **returning aircraft**. This chart showed greatest damage on the **aircraft extremities**, not on the main wing and tail spars, engines, and core fuselage areas.

BIAS



As such, the Air Ministry wanted to add armor to the **extremities**. Wald suggested they were **dead wrong**.

To avoid **“survivorship bias”,** armor should be added to the areas with the **fewest holes**: if no returning planes had holes in their wing spars and engines, then even a few holes in those locations were **deadly**.

Take-Away: the data that is missing may be as important to story than the data that is there. Storytelling is not always an obvious endeavour.

[20 Cognitive Biases That Screw Up Your Decisions, *Business Insider*]

1. Anchoring bias.

People are **over-reliant** on the first piece of information they hear. In a salary negotiation, whoever makes the first offer establishes a range of reasonable possibilities in each person's mind.

2. Availability heuristic.

People **overestimate the importance** of information that is available to them. A person might argue that smoking is not unhealthy because they know someone who lived to 100 and smoked three packs a day.

3. Bandwagon effect.

The probability of one person adopting a belief increases based on the number of people who hold that belief. This is a powerful form of **groupthink** and is reason why meetings are often unproductive.

4. Blind-spot bias.

Failing to recognize your own cognitive biases is a bias in itself. People notice cognitive and motivational biases much more in others than in themselves.

5. Choice-supportive bias.

When you choose something, you tend to feel positive about it, even if that **choice has flaws**. Like how you think your dog is awesome — even if it bites people every once in a while.

6. Clustering illusion.

This is the tendency to **see patterns in random events**. It is key to various gambling fallacies, like the idea that red is more or less likely to turn up on a roulette table after a string of reds.

7. Confirmation bias.

We tend to listen only to information that confirms our **preconceptions** — one of the many reasons it's so hard to have an intelligent conversation about climate change.

8. Conservatism bias.

Where people favor prior evidence over new evidence or information that has emerged. People were **slow to accept** that the Earth was round because they maintained their earlier understanding that the planet was flat.

9. Information bias.

The tendency to **seek information when it does not affect action**. More information is not always better. With less information, people can often make more accurate predictions.

10. Ostrich effect.

The decision to **ignore dangerous or negative information** by "burying" one's head in the sand, like an ostrich. Research suggests that investors check the value of their holdings significantly less often during bad markets.

[20 Cognitive Biases That Screw Up Your Decisions, *Business Insider*]

11. Outcome bias.

Judging a decision based on the **outcome** — rather than how exactly the decision was made in the moment. Just because you won a lot in Vegas doesn't mean gambling your money was a smart decision.

12. Overconfidence.

Some of us are **too confident about our abilities**, and this causes us to take greater risks in our daily lives. Experts are more prone to this bias than laypeople, since they are more convinced that they are right.

13. Placebo effect.

When **simply believing** that something will have a certain effect on you causes it to have that effect. In medicine, people given fake pills often experience the same physiological effects as people given the real thing.

14. Pro-innovation bias.

When a proponent of an innovation tends to **overvalue its usefulness** and undervalue its limitations. Sound familiar, Silicon Valley?

15. Recency.

The tendency to weigh the **latest information** more heavily than older data. Investors often think the market will always look the way it looks today and make unwise decisions.

16. Salience.

Our tendency to focus on the **most easily recognizable features** of a person or concept. When you think about dying, you might worry about being mauled by a lion, as opposed to what is statistically more likely, like dying in a car accident.

17. Selective perception.

Allowing our expectations to **influence how we perceive** the world. An experiment involving a football game between students from two universities showed that one team saw the opposing team commit more infractions.

18. Stereotyping.

Expecting a group or person to have certain qualities without having real information about the person. It allows us to quickly identify strangers as friends or enemies, but people tend to **overuse and abuse** it.

19. Survivorship bias.

An error that comes from focusing only on surviving examples, causing us to **misjudge a situation**. For instance, we might think that being an entrepreneur is easy because we haven't heard of all those who failed.

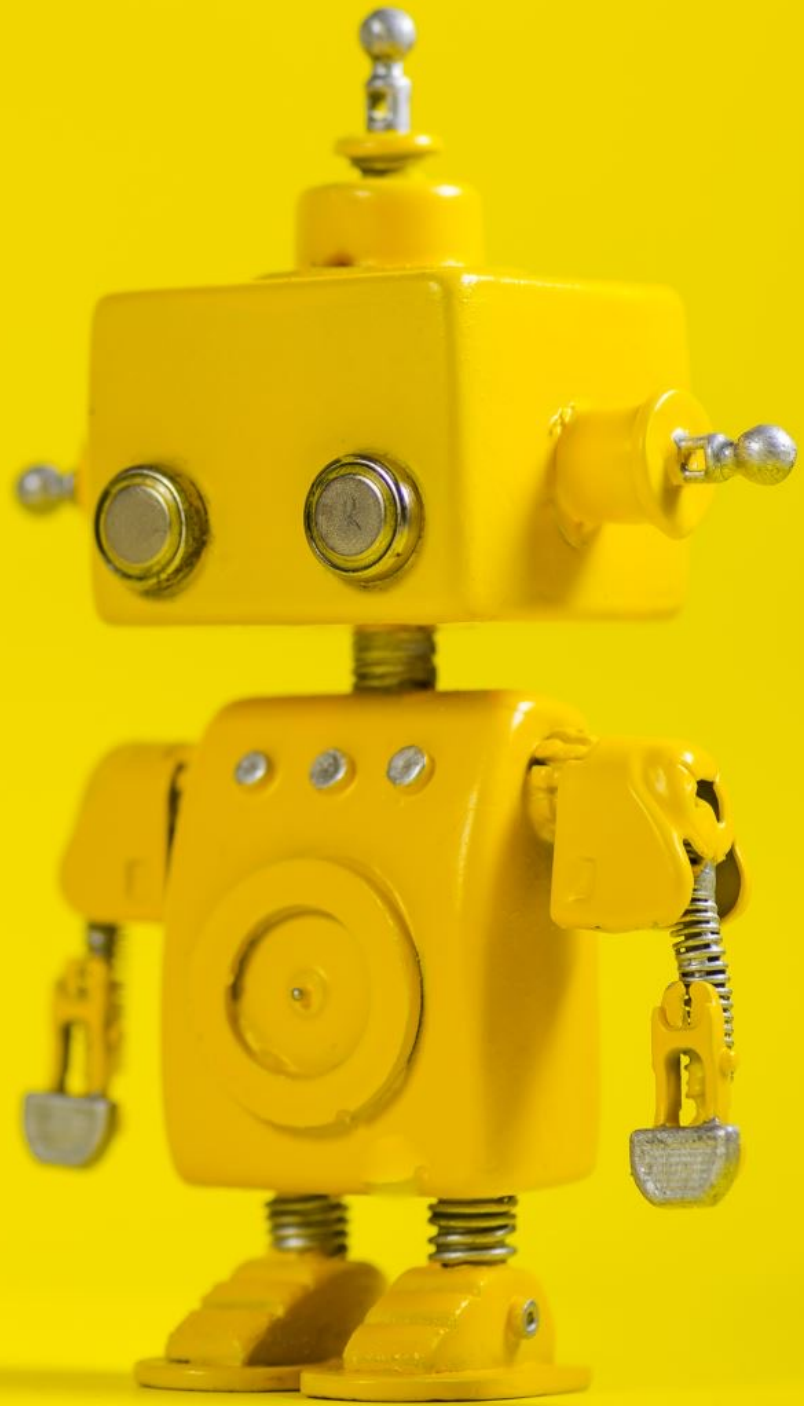
20. Zero-risk bias.

Sociologists have found that **we love certainty** — even if it's counterproductive. Eliminating risk entirely means there is no chance of harm being caused.

ETHICS IN A.I.

The explosion in the use of "**Artificial Intelligence**" has required the creation of GoC guidance on the responsible use of A.I.

The GoC policy also includes the *Directive of Automated Decision-Making* and the *Guide on the use of generative A.I.*



CASE STUDY: HIRING

Your company is always looking for the most talented people, especially for technical positions.

Corporate policy **supports** diversity and inclusion.

The hiring process is time-consuming, and you are concerned about **personal biases** of panel members influencing the decisions.

With the help of an outstanding A.I. team, you **automate** this process.

The A.I.-assisted processes finds talented people, who fit into the organizational culture, and who like their jobs (low turnover).

CASE STUDY: AMAZON HIRING A.I.

But... more likely to get hired if your name was **Jared** and you played **lacrosse**.

A.I. was behaving in a **biased manner**, not recommending women be hired.

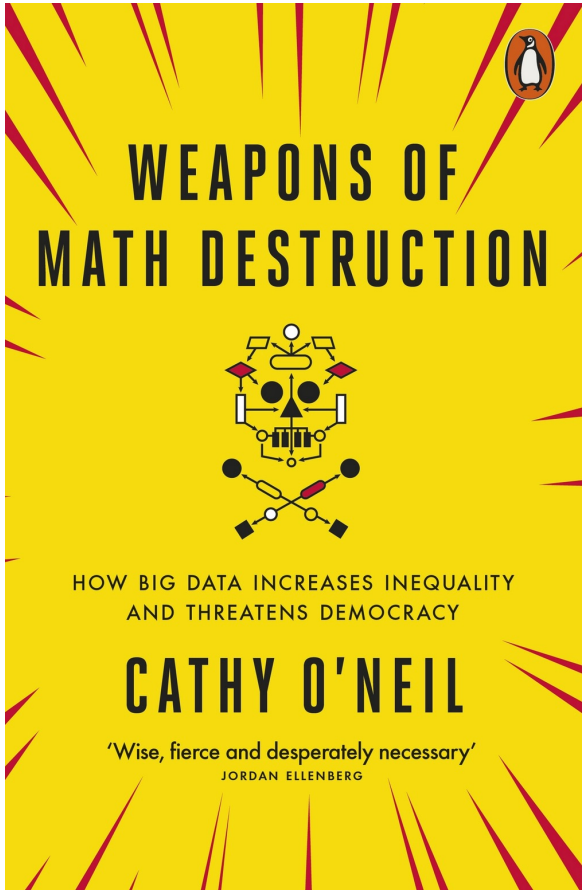
Amazon was not confident they could **remove the bias** or identify biased behaviours in the future, so they project was **scrapped**.

THE THREAT

In her book about data power, Dr. Cathy O'Neil presents several cautionary examples and tales.

"A computer program could speed through thousands of résumés [...] and sort them into neat lists [...]. This not only saved time but also was marketed as fair and objective. After all, it didn't involve prejudiced humans digging through reams of paper, just machines processing cold numbers. [...]

The math-powered applications driving the data economy were based on choices made by fallible human beings. Some of these choices were no doubt made with the best intentions. Nevertheless, many of these models and algorithms encoded human prejudice, misunderstanding and bias into the software systems that increasingly managed our lives."



A.I. ETHICS GUIDING PRINCIPLES

Uses:

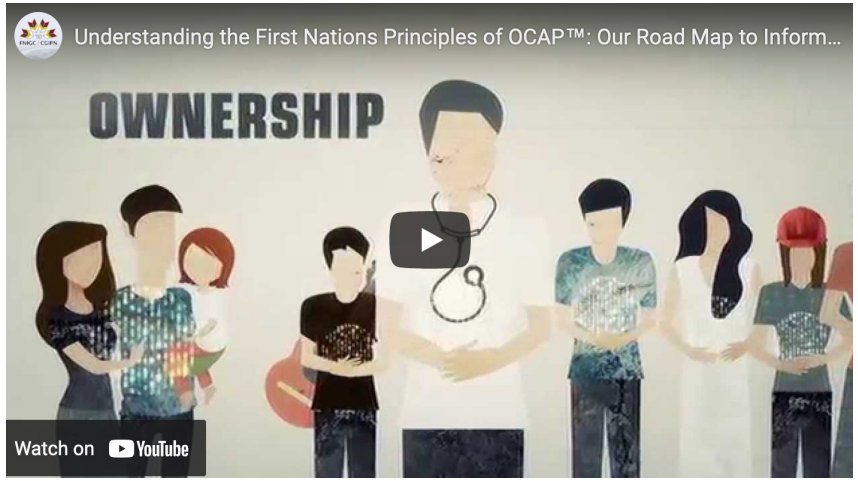
- privacy and security
- transparency
- accountability
- methodology and data quality
- model fairness
- model explainability
- indigenous data sovereignty

DATA ETHICS

Data ethics questions:

- Who, if anyone, owns data?
- Are there limits to how data can be used?
- Are there value-biases built into certain analytics?
- Are there categories that should not be used in analyzing personal data?
- Should some data be publicly available to all researchers?

Are there lessons to be learned from the First Nations Principles of OCAP®? (**ownership, control, access, possession**)



[https://fnigc.ca/ocap-training]

FIRST NATIONS DATA

First Nations Principles of **OCAP®**:

- **Ownership:** cultural knowledge, data, and information is owned by First Nations communities
- **Control:** First Nations communities have the right to control all aspects of research and information management that impact them
- **Access:** First Nations communities must have access to information and data about themselves no matter where it is held
- **Possession:** First Nations communities must have physical control of relevant data

[OCAP® is a registered trademark of the *First Nations Information Governance Centre* (FNIGC), fnigc.ca/OCAP]

[https://datascience.virginia.edu/center-data-ethics-and-justice]

DATA ETHICS

Some examples of data science ethics questions (University of Virginia's *Centre Data Ethics and Justice*):

- **who**, if anyone, owns data?
- are there **limits** to how data can be used?
- are there **value-biases** built into certain analytics?
- are there categories that should **never** be used in analyzing personal data?
- should some data be **publicly available** to **all** researchers?

LEGAL CONSIDERATIONS USING DATA

Profiling:

- are you using personal data to draw inferences that are unfair, unethical or discriminatory?

Surveillance:

- are people being placed in a perpetual line-up?

Liability:

- are you liable for what an A.I. does?

EMERGING LEGAL TRENDS

Canada

GoC: Algorithmic Impact Assessment prior to the production of any Automated Decision System

Privacy Commissioner (Personal Information Protection and Electronic Documents Act):

- Defines automated decision systems any tech that assists or replaces the judgment of humans.
- Need to give people an explanation of the prediction/recommendation, and how their personal info was used.

Europe

General Data Protection Regulation (GDPR)

Article 22: not subject to a decision based solely on automated processing (with exceptions)

Article 15: if subject to such a decision, have right to meaningful information about the logic involved.

DATA ETHICS GUIDING PRINCIPLES

1. Public Benefit
2. Privacy and Security
3. Transparency
4. Accountability
5. Methodology and Data Quality
6. Indigenous Data Sovereignty

CODES OF CONDUCT

A **code of conduct** is a set of rules outlining the norms, rules, and responsibilities or proper practices of an individual party or an organization (in medicine, we have the *Hippocratic Oath*).

Many professional organizations are starting to integrate data ethics into their **professional** designation's codes of conduct.

The Government of Canada has a general "Values and Ethics Code for the Public Sector" in which the use of data is **implied**.

The 2023-2026 Data Strategy explicitly identifies ethical use of data as a guiding principle.

There are other subject-specific policies such as the Tri-Council Policy on Ethical Conduct for Research Involving Humans, depending on areas of expertise.

PROTECTING AND SHARING CONFIDENTIAL DATA

Privacy is protected by laws and other measures including the Statistics Act, the Privacy Act, the Directive on Security Management and by GoC Levels of Security.

In short, the data in documents/information with a higher classification rating than "unclassified" can **only be shared with personnel with the relevant level of screening** and on a "**need to know**" basis, with documents being held at a site with the appropriate organization screening.

Type	Information and assets	Organization screening	Personnel screening
Classified	Top Secret	Facility security clearance (Top Secret)	Top Secret
Classified	Secret	Facility security clearance (Secret)	Secret
Classified	Confidential	Facility security clearance (Confidential)	Secret
Protected	Protected C	Designated organization screening	Enhanced reliability status
Protected	Protected B	Designated organization screening	Enhanced reliability status
Protected	Protected A	Designated organization screening	Reliability status

DECISION-MAKING

Ethical research groups have identified different approaches to ethical decision making. The simplest being the **Blanchard-Peale framework** which is summarized as:

1. Is it legal?
2. Is it fair?
3. How does it make me feel?

Other approaches: **Markkula Centre framework** (utilitarianism, rights approach, fairness, common good approach, virtue approach), **issue-contingent model** (recognize issue, make judgement, establish moral intent, engage in behaviour).

The key concept is that decision-making for the organization must first be analyzed – however decisions are made, guidance is provided to help decision makers if issues must be addressed.

ETHICS AND THE DATA LIFECYCLE

If we remember from the Data Awareness module that there are a number of steps in the data lifecycle. We need to consider ethics at each stage

Do we **acquire** data in an ethical and unbiased manner? It is **stored** safely? When we prepare it do we introduce biases? Is it **staged** safely and when we **present,** are we representing all the actors in a fair and ethical manner?

3. DATA GOVERNANCE

DATA FOUNDATIONS

WHAT IS DATA GOVERNANCE?

Data governance is a concept that enables an organization to ensure that **high data quality** exists throughout the **complete life cycle** of the data.

Focusing on data governance allows the organization to have data that:

- is **available** when needed;
- is **usable** when accessed;
- is **consistent** when analyzed;
- has **integrity** and is of **high quality**, and
- is **secure** and **trustworthy**.

WHAT IS DATA GOVERNANCE?

Data governance encompasses:

- **people**;
- **processes**, and
- **information technology**.

It is required to create a **consistent** and **proper handling** of an organization's data, across the enterprise.

It provides the **foundation**, **strategy**, and **structure** to ensure that data is managed as an **asset** and transformed into **meaningful information**.

[https://dama.org/content/body-knowledge]

DATA GOVERNANCE

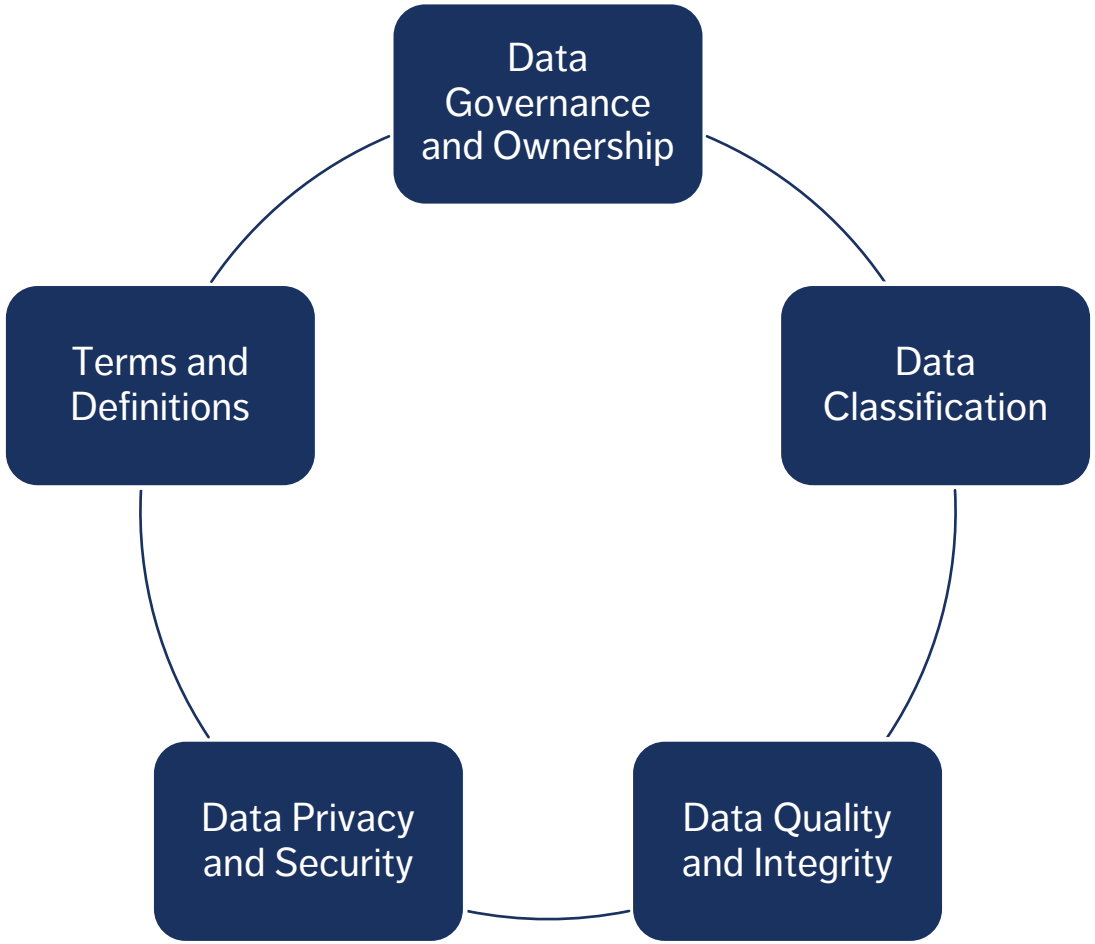
Data Management Association (DAMA)

- DMBOK 2 (*Data Management Body of Knowledge*)
- well-detailed & thorough
- sections “reasonably” aligned with GoC approach
- backed by a professional organization
- but … not government focused

DATA GOVERNANCE

Placeholder – go through each of the sections in a little more detail. It's going to be especially important for things like metadata, reference data etc.

DATA GOVERNANCE



DATA GOVERNANCE IN THE GOC

Central point of reference for GoC (**Digital Government** website):

- Strategic plans, policies, standards and guidelines related to government digital services

Report to the **Clerk of the Privy Council:**

- A Data Strategy Roadmap for the Federal Public Service

Treasury Board Secretariat (selection):

- Policy on Service and Digital and Digital Operations Strategic Plan: 2018-2022
- Government of Canada Strategic Plan for Information Management and Information Technology 2017 to 2021
- Government of Canada Cloud Adoption Strategy: 2018 update

Industry Canada:

- Canada's Digital Charter in Action: A Plan by Canadians, for Canadians

ACCOUNTABILITY AND RESPONSIBILITY

	Data Trustee	Data Steward	Data Custodian	Data Contributor	Data Consumer
Definition	A person who has governance and compliance responsibility for a set of data assets	A person who has business accountability for a set of data assets	A person who has technical accountability for a set of data assets	A person who creates or collects data that is relevant to the organization	A person who uses data to enable business outcomes
Accountability summary	▪ Compliance ▪ Risk ▪ Oversight ▪ Approval ▪ Champion ▪ Issue resolution	▪ Accuracy ▪ Consistency ▪ Business requirements ▪ Metadata definition and management ▪ Data Quality ▪ Data Curation ▪ Fitness for Purpose ▪ Governance ▪ Inventory ▪ RDM ▪ Role Management	▪ Security ▪ Access Management ▪ Availability ▪ Capacity ▪ Continuity ▪ Safeguarding ▪ Implementation ▪ Technical standards ▪ Configuration ▪ Control ▪ Modeling ▪ Versioning ▪ Change Management	▪ Data Acquisition and entry ▪ Data Quality ▪ Metadata Preparation ▪ Ethical and secure gathering of data ▪ Identification of issues ▪ Identification of new data sources	▪ Ethical use ▪ Report on Data Quality ▪ Report on fitness for purpose ▪ Identification of business and data rules ▪ Identification and reporting of data control ▪ Use in line with governance

GOALS OF DATA GOVERNANCE

1. Create self-service data culture
2. Establish internal rules for data use
3. Implement compliance requirements
4. Improve internal and external comms
5. Increase value of data
6. Reduce costs
7. Continually manage risks
8. Ensure continued existence

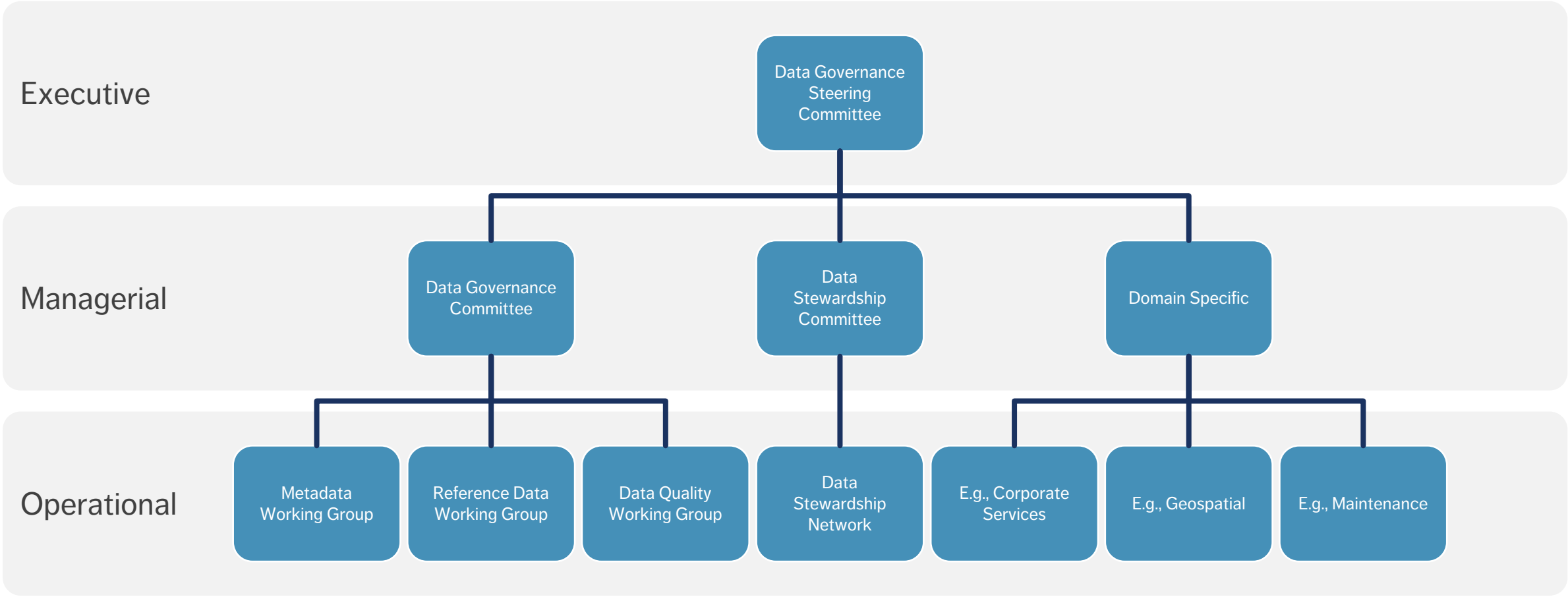
COMMON ISSUES WITH DATA GOVERNANCE

Pioneered, planned, and projected by the most technical resources.

Verbiage is all over the place.

‘Donut meetings’...

POSSIBLE STRUCTURE



4. DATA COLLECTION

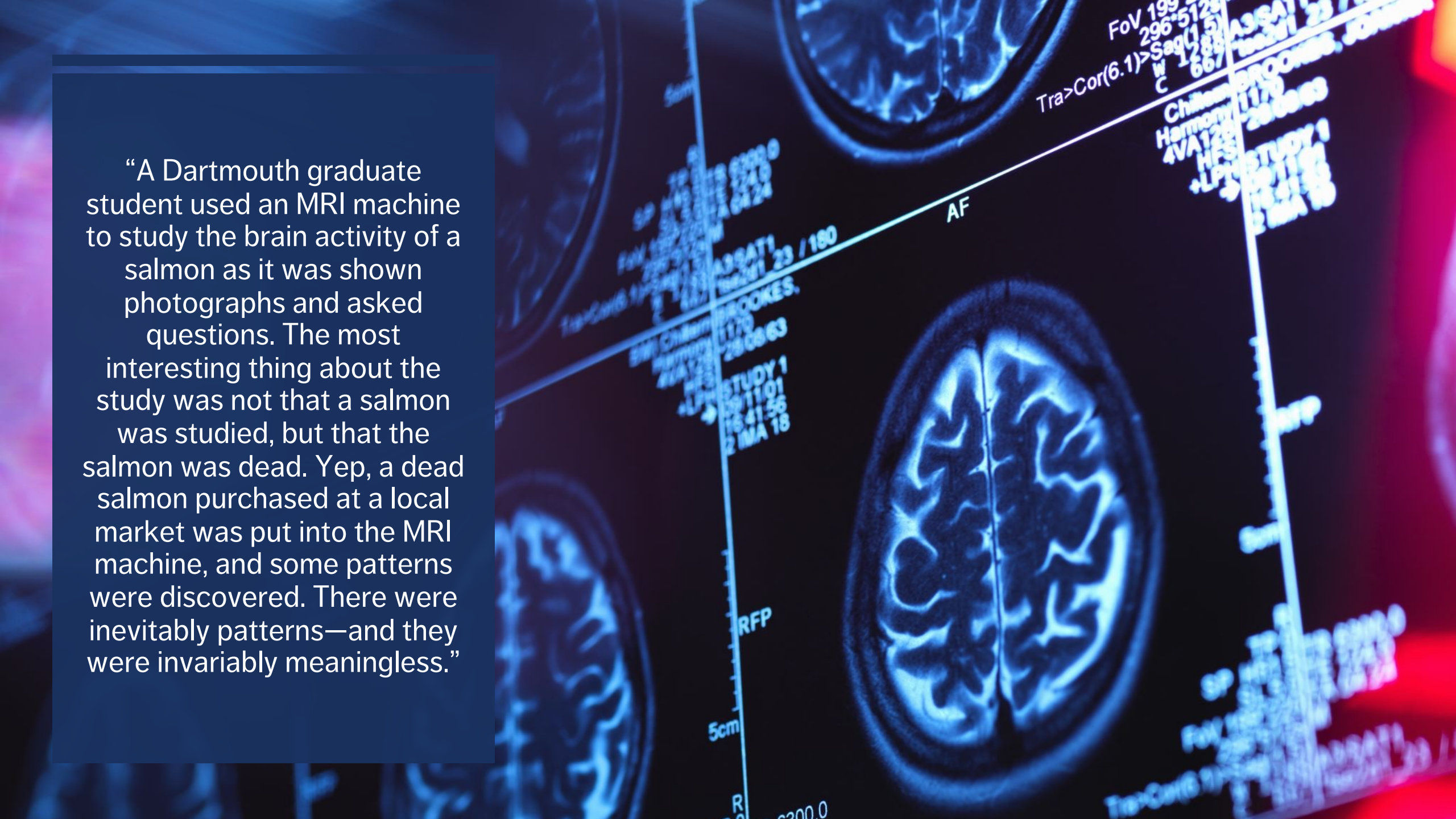
DATA FOUNDATIONS

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.

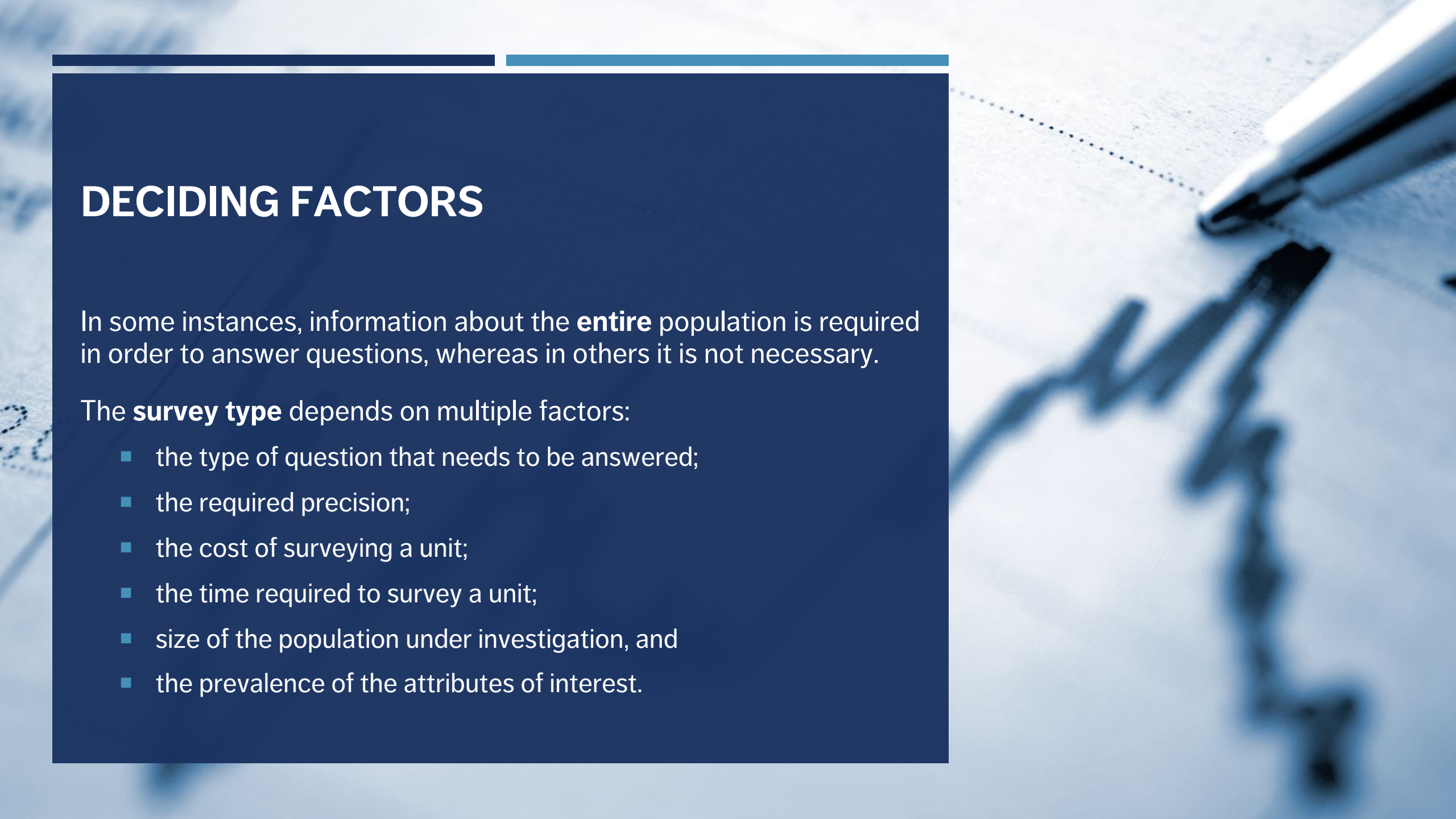
When survey sampling is done properly, we may be able to use various **statistical methods** to make **inferences** about the **target population** by sampling a (comparatively) small number of units in the **study population**.

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.

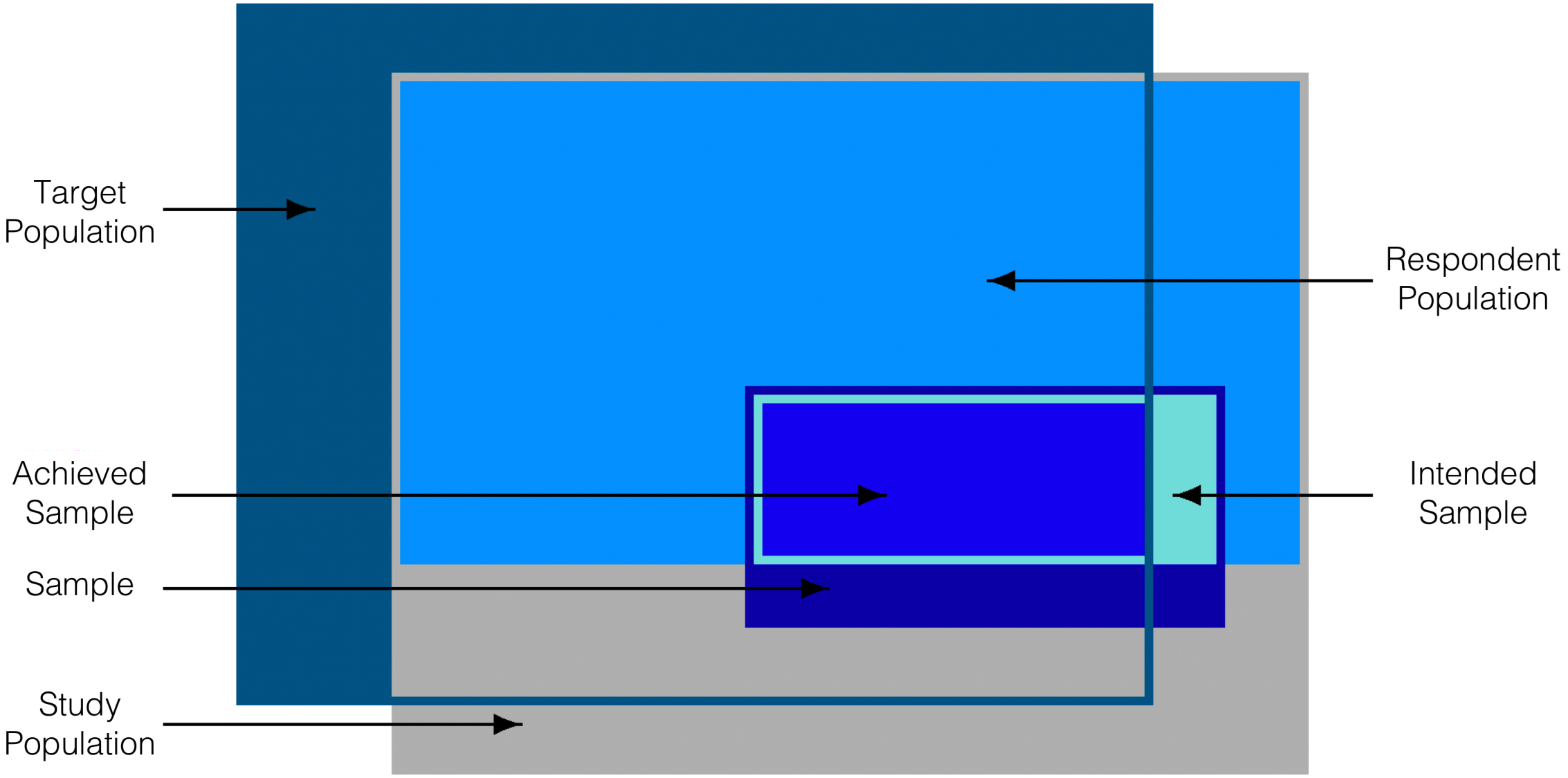


STUDY/SURVEY STEPS

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.


Target Population
Respondent Population
Achieved Sample
Intended Sample
Sample
Study Population

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

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).

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.

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).

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 (CIS)

Replicated sampling (ReS)

Multi-stage sampling (MSS)

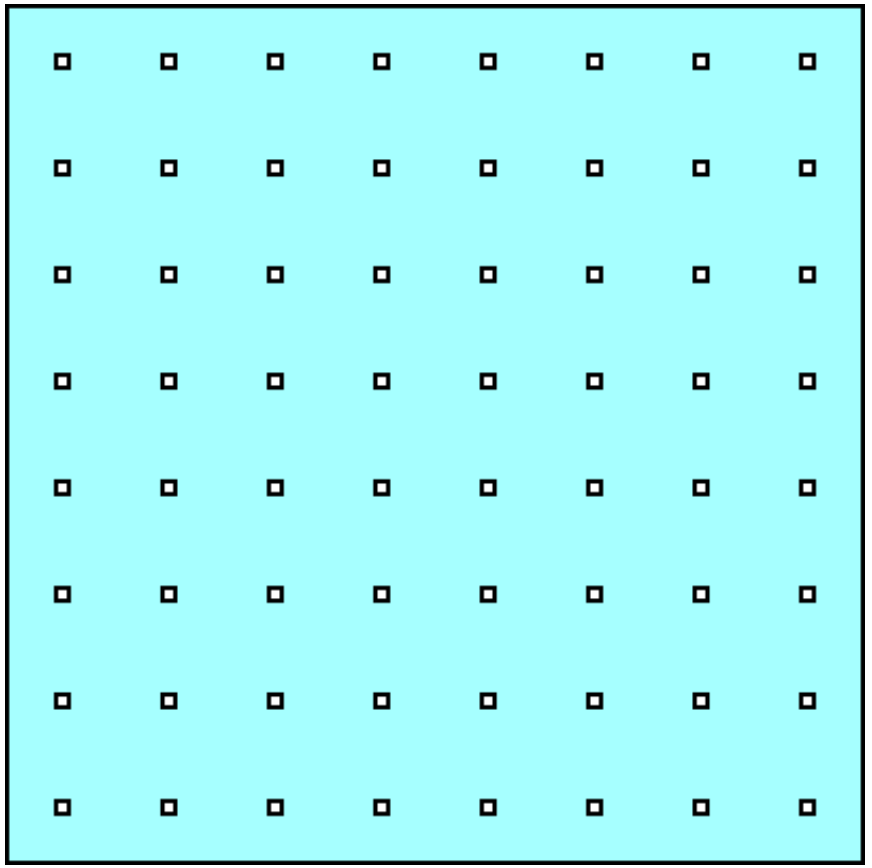
Multi-phase sampling (MPS)

Probability proportional-to-size sampling (PPS)

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%).



SIMPLE RANDOM SAMPLING (SRS)

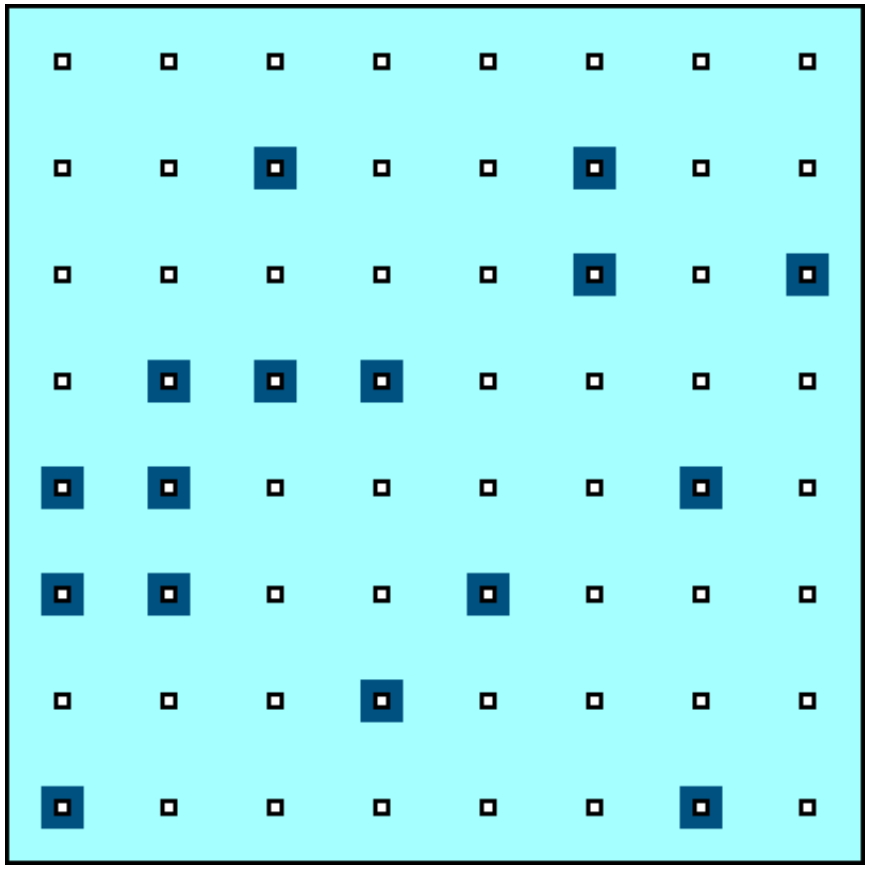
In SRS, we select n units randomly from the frame.

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)

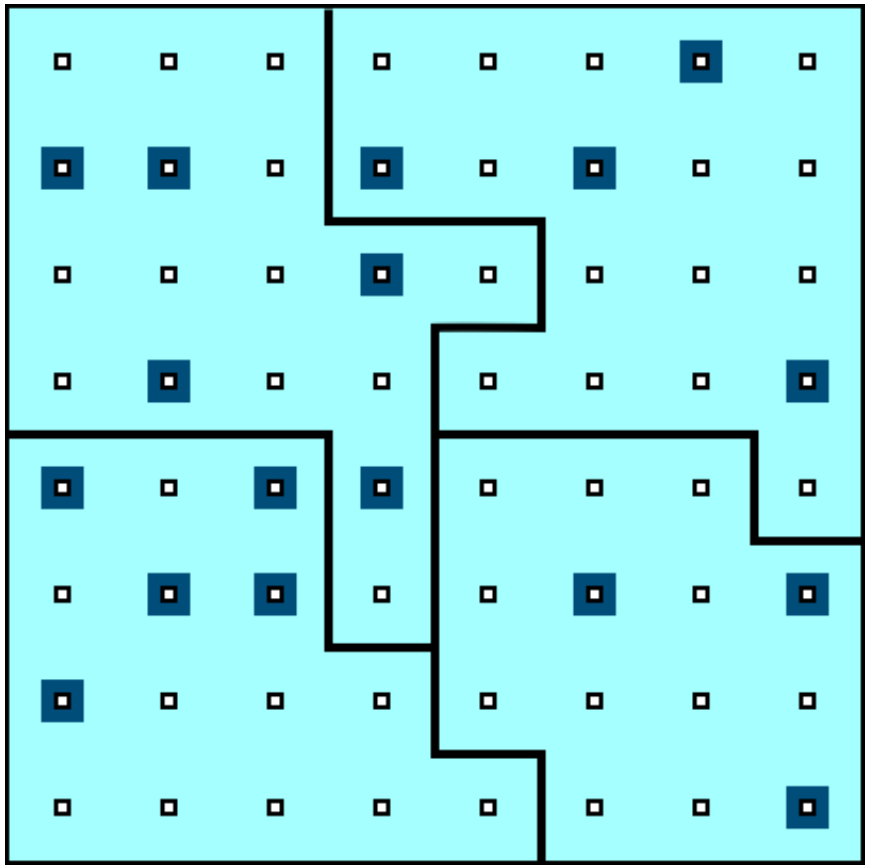
In StS, $n = n_1 + \cdots + n_k$ units are randomly drawn from k **strata**.

Advantages:

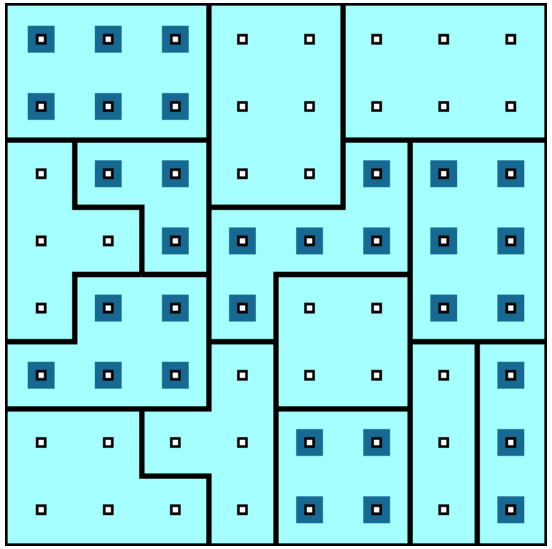
- may produce smaller error bound on estimation than SRS
- may be less expensive if elements are conveniently strat.
- may provide estimates for sub-populations

Disadvantages:

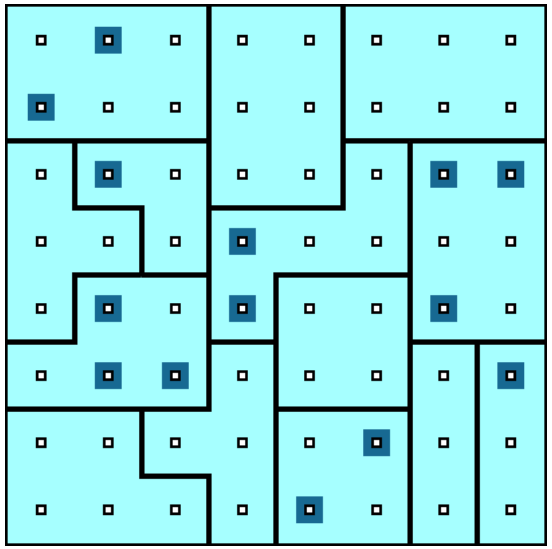
- no major disadvantage
- if there are no natural ways to stratify the frame into homo-geneous groupings, StS is roughly equivalent to SRS



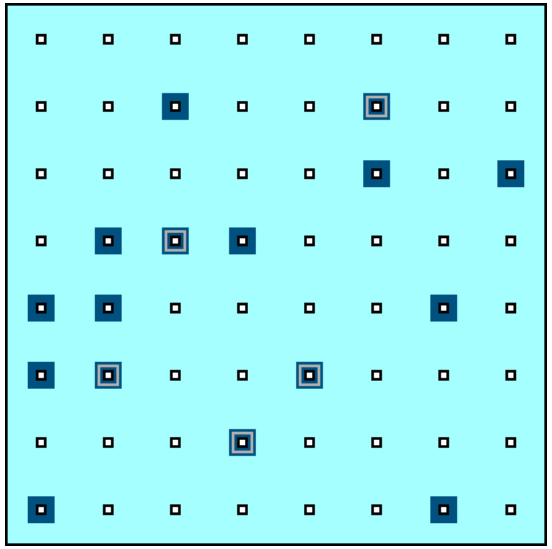
OTHER PROBABILISTIC SAMPLING DESIGNS



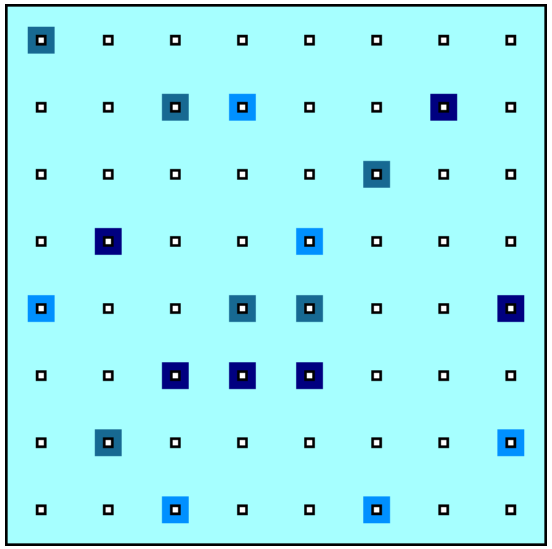
Cluster Sampling (CIS)



Multi-Stage Sampling (MSS)



Multi-Phase Sampling (MPS)



Replicated Sampling (ReS)

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.

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.

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

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?**

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.

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.

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.

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