Data Literacy

and Philippine Public Procurement Data

Ben Hur Pintor

Day 3| 20-22 June 2022

BN
HR

liberty. data. geospatial.

@bnhr.xyz @bnhrdotxyz bnhrdotxyz https://bnhr.xyz

Outline

I. About the event

  A. Code of conduct

  B. About me

  C. About you

  D. Objectives and expectations

  E. Schedule

II. Session 1

  A. Gentle introduction to open data

  B. Open contracting and the OCDS

III. Session 2

  A. What is data literacy?

  B. How do we build individual and organizational data literacy?

IV. Session 3

  A. The Data Pipeline

  B. Data-driven projects using the Data Pipeline

V. Session 4

  A. How to make better data presentations

  B. Free and open source tools for working with data

VI. Session 5

  A. Data ethics, data fallacies, and other considerations

  B. How to start sharing and opening data

VII. Session 6

  A. Data expedition and brainstorming a data-driven procurement-themed project

About me

Ben Hur Pintor

geospatial generalist. open stuff advocate. maptivist/datactivist.

Data Training Lead

Open Knowledge Foundation

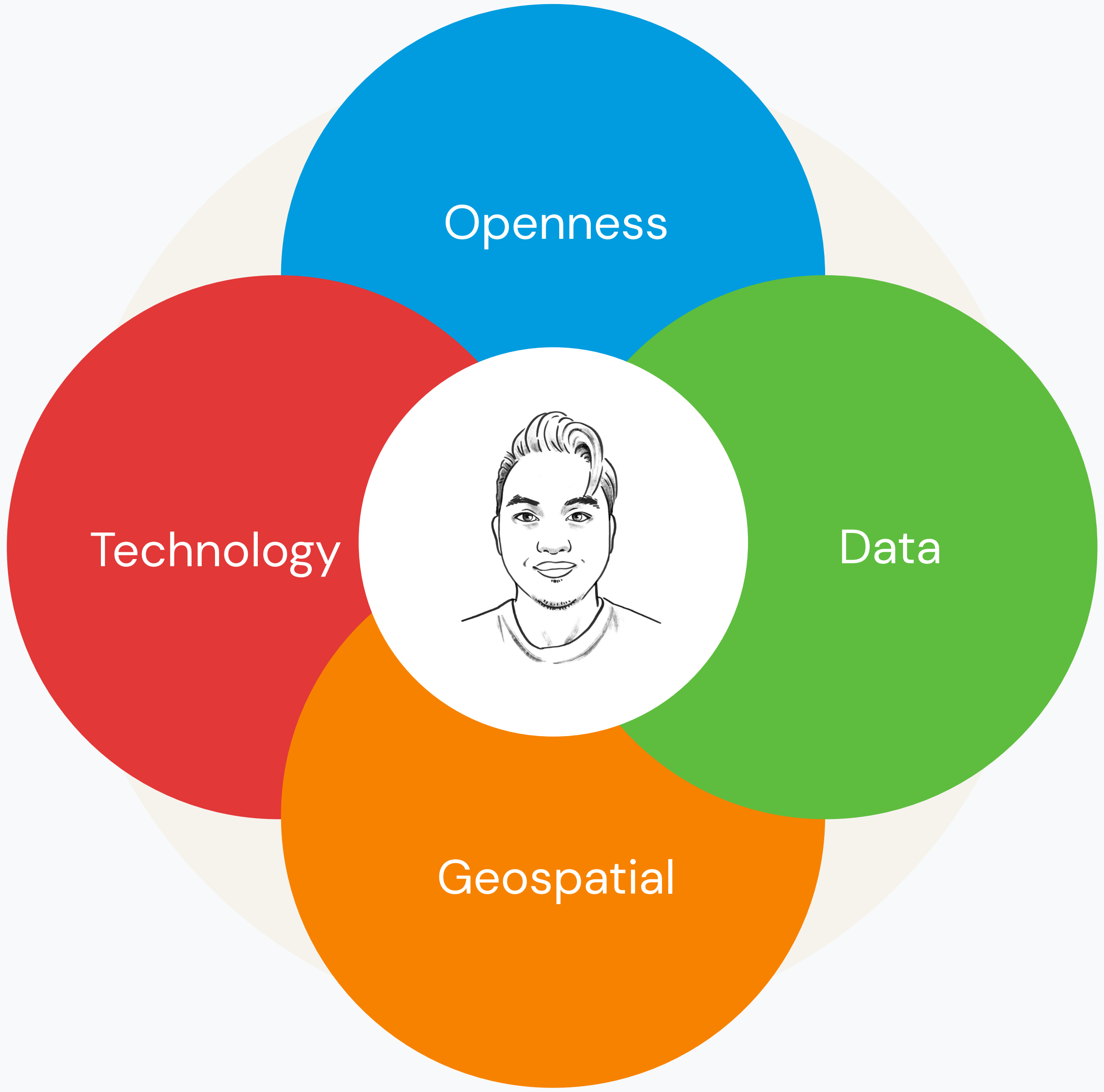
Proprietor

BN HR

openness.
data.
geospatial.

Chief Technology Officer

SmartCT
Citizens x Technology



BN HR

liberty. data. geospatial.

@bnhr.xyz @bnhrdotxyz bnhrdotxyz https://bnhr.xyz

Open Knowledge Foundation

We teach

If you or your organisation wants to learn about data literacy or even develop your skills to expert level, our team is here to help you on your journey.

We build

As open experts, we can create tools and provide services that help people and organisations put their data literacy learnings to work.

We organise

Through campaigning and community building, we're making an open future.

About SmartCT

SmartCT is the first tech non-profit in the Philippines and a pioneer in the smart cities field.

We aim to create a **movement that transforms the way we think, do, and plan smart cities and communities** especially in developing countries such as the Philippines through a co-developed, citizen-centric approach that puts openness and citizens at the heart of the development.

Making Smart Cities Open.

BN
HR

openness.
data.
geospatial.

- Established in 2019
- Part enterprise, part advocacy
- Provides training, support, and consulting services on open data, open source, data literacy, and free and open source software for geospatial applications (FOSS4G)

QGIS Certifying Organization

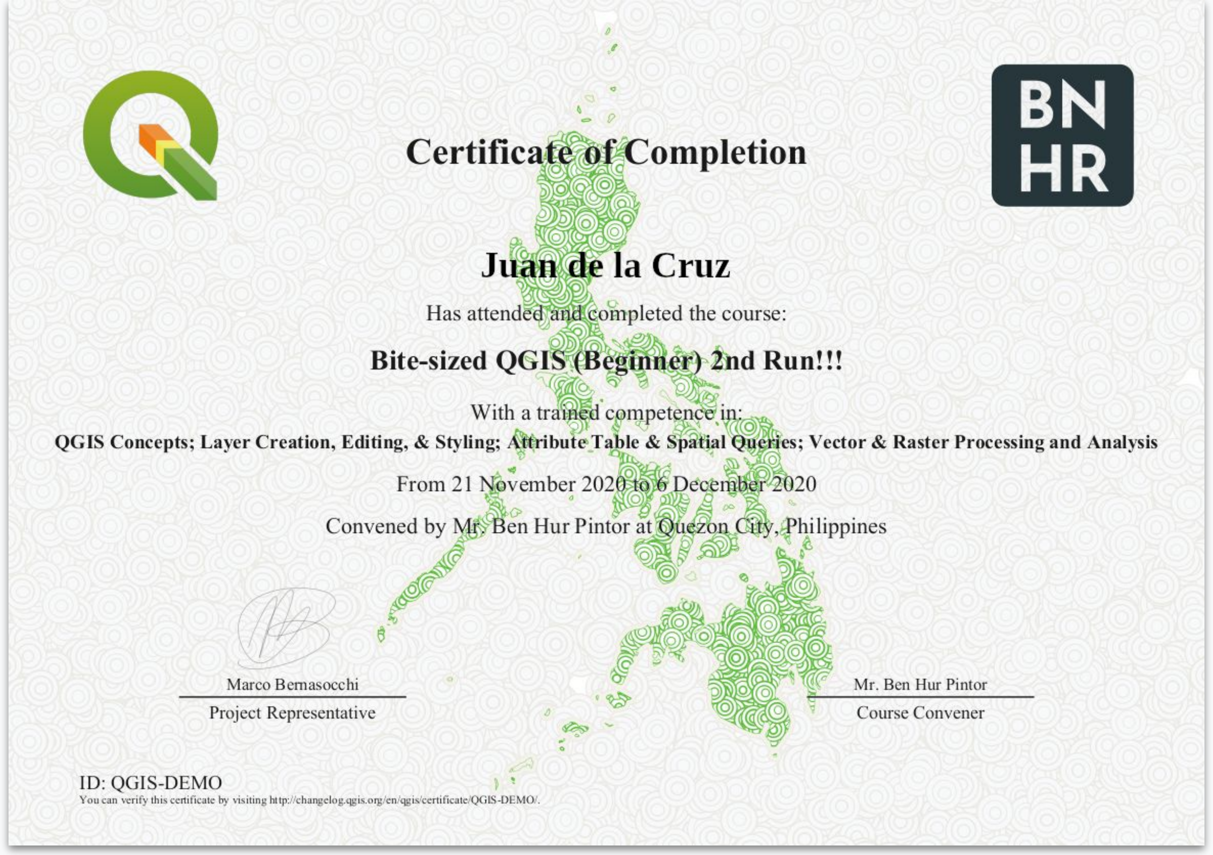
- Courses are vetted by the QGIS Project Steering Committee
- Can issue official QGIS Certificate

QGIS Sustaining Member

- Financially supports the QGIS Project
- Currently the only (and probably first) sustaining member from the Philippines

Some people I've worked with:

Philippine Statistics Authority · Solid · Responsive · World-class
SAS
THIBI
ADB
Hivos people unlimited
webcast TECHNOLOGIES, INC.
AERO 360 SOLUTIONS BEYOND DRONES



BN
HR

Certificate of Completion

Juan de la Cruz

Has attended and completed the course:

Bite-sized QGIS (Beginner) 2nd Run!!!

With a trained competence in:
QGIS Concepts; Layer Creation, Editing, & Styling; Attribute Table & Spatial Queries; Vector & Raster Processing and Analysis

From 21 November 2020 to 6 December 2020

Convened by Mr. Ben Hur Pintor at Quezon City, Philippines

Marco Bernasocchi
Project Representative

Mr. Ben Hur Pintor
Course Convener

ID: QGIS-DEMO
You can verify this certificate by visiting http://changelog.qgis.org/en/qgis/certificate/QGIS-DEMO/.

**“Statistics never lie,
but lovers often do...”**

Antonio vs. Reyes, 484 SCRA 353 (2006)

With all due respect to Justice Tinga but **statistics and data do lie.**

Data fallacies

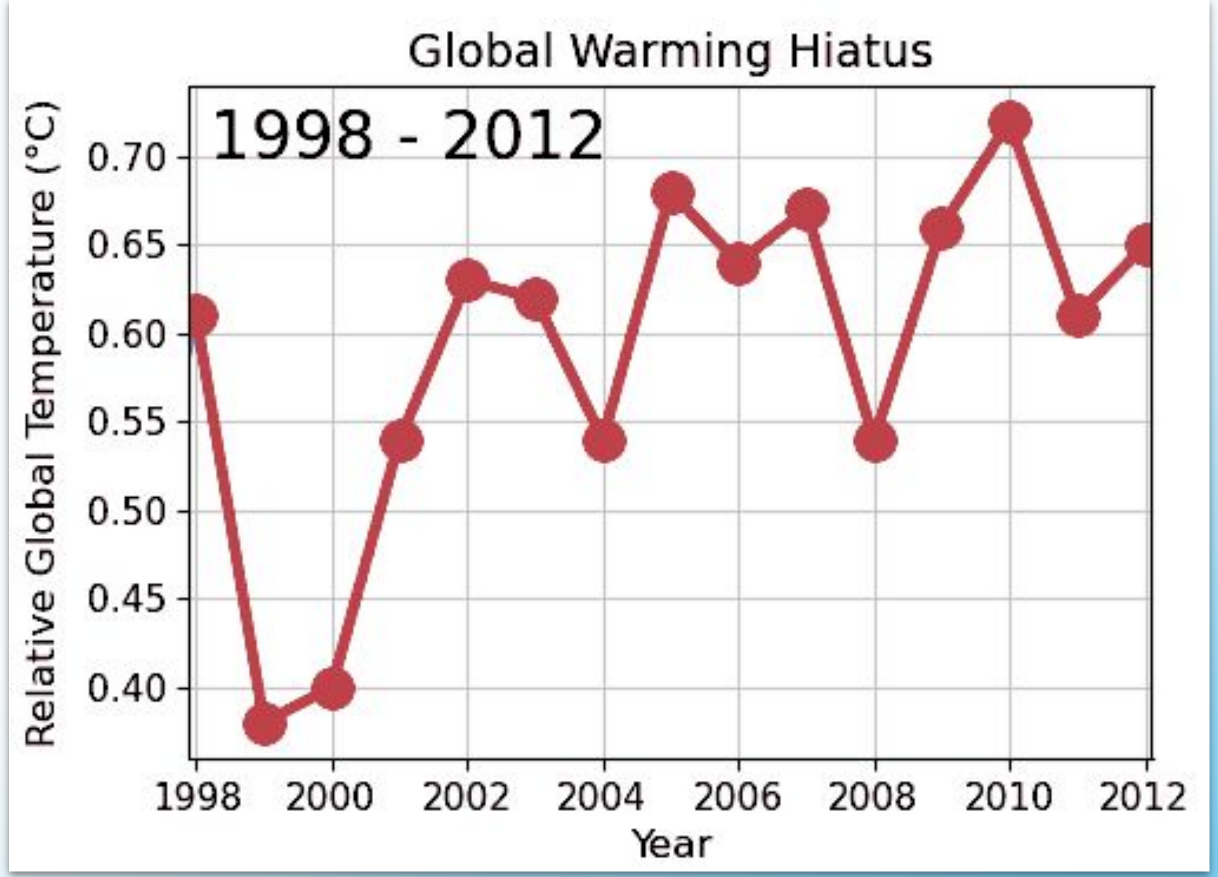
Cherry picking

What it is

- Also known as suppressing evidence or the fallacy of incomplete evidence.
- Selecting, using, and presenting only the subset of the data that agree or fit with your claims and beliefs.
- This becomes really dangerous when paired with people's confirmation bias.

How to avoid it

- As a creator, always be faithful to your data and results especially when they do not fully agree with your claims.
- As a consumer, ask for the complete dataset or ask yourself: "What am I not being told?"



By selecting or cherry-picking data, the trend of global warming appears to mistakenly stop*, as in the period from 1998 to 2012, which is actually a random contrary fluctuation.*

image source: Physikinger, CC0, via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Global_warming_hiatus.gif

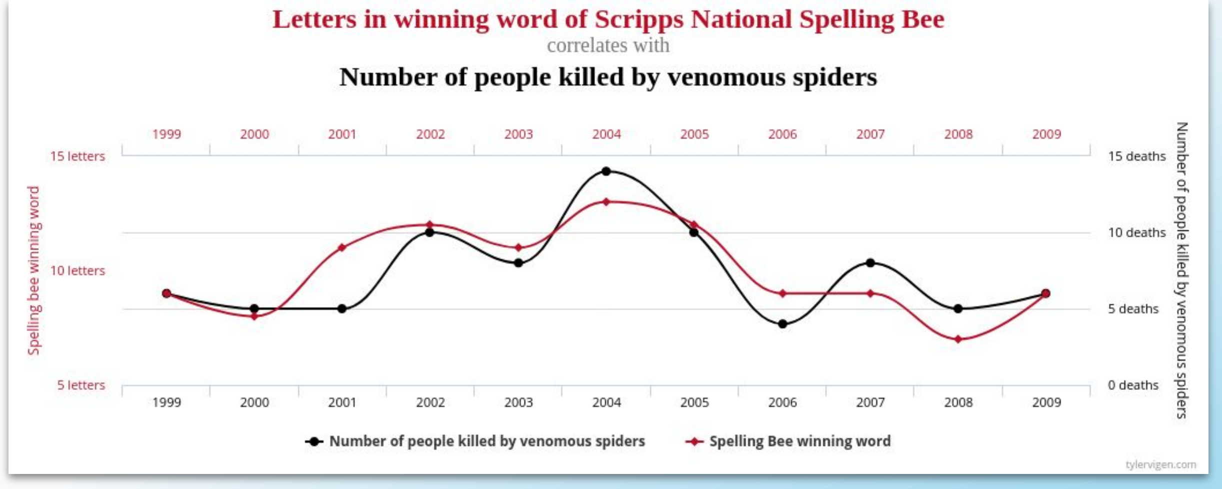
Data dredging

What it is

- The misuse of data analysis to find patterns in data that can be presented as statistically significant.
- Seeking correlation where there is none. Performing countless statistical tests on data and reporting the ones that show correlation.
- If you combine enough time with a large enough dataset, you are bound to find things that appear to be correlated.

How to avoid it

- Always be upfront with what you are testing—e.g. using a hypothesis in the analyze step of the data pipeline.
- Accept that sometimes things that seem to be correlated aren't.



An example of data produced by ***data dredging through a bot*** *operated by Tyler Vigen,* ***apparently showing a close link between the best word in a spelling bee competition and the number of people in the US killed by venomous spiders****.*

It's obviously a coincidence: ***with so many possible comparisons of data of things happening in the world, it is easy to find some unrelated data*** *that shows similar trends.*

image source: Tyler Vigen, CC BY 4.0, via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Spurious_correlations_-_spelling_bee_spiders.svg

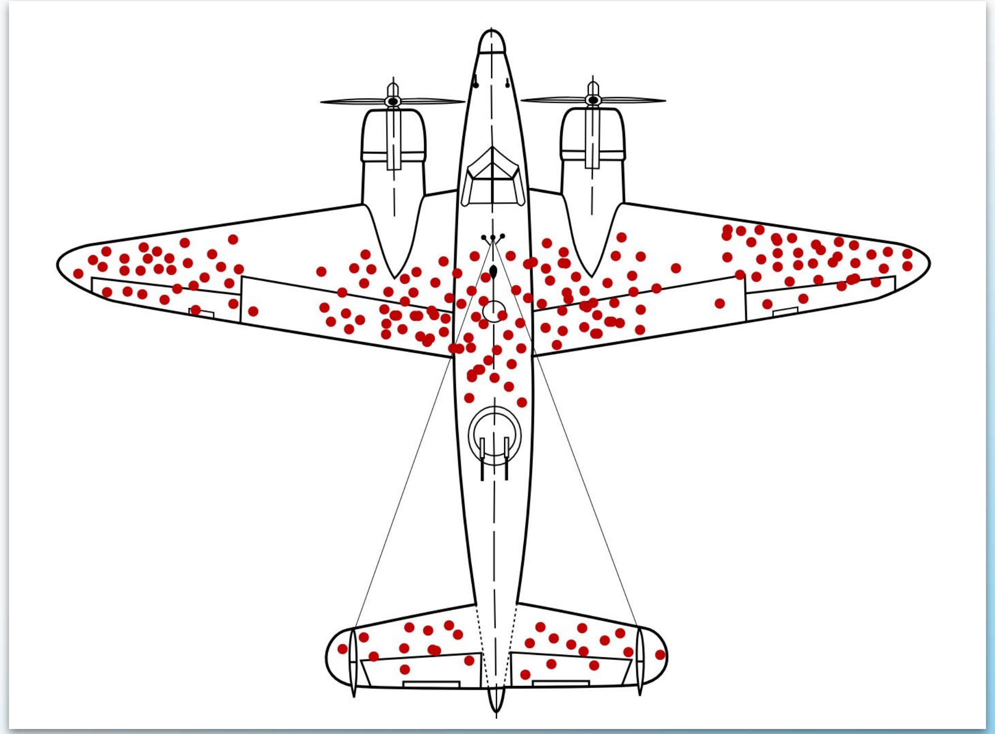
Survivorship bias

What it is

- Logical error of drawing conclusions from an incomplete dataset composed of data that has survived a selection process and overlooking those that did not, typically because of their lack of visibility.
- Example:
  - Bullet patterns of WW2 aircrafts returning from the war
  - College dropouts being billionaires—for every dropout who became a billionaire, how many thousands more did not?

How to avoid it

- Always try to look at the full picture and ask yourself if you are overlooking anything or if something is missing in your data.
- Ask yourself: "Did your data undergo any selection or trimming process prior to your analysis?"



This hypothetical pattern of damage of returning aircraft shows locations where they can sustain damage and still return home.

***If the aircraft was reinforced in the most commonly hit areas**, this would be a result of **survivorship bias because crucial data from fatally damaged planes was being ignored; those hit in other places presumably did not survive**.*

image source: Martin Grandjean (vector), McGeddon (picture), Cameron Moll (concept), CC BY-SA 4.0 , via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Survivorship-bias.svg

Sampling bias

What it is

- Occurs when a sample is selected in a way such that some members of the intended population have a lower or higher chance of being included in the sample.
- Results in conclusions drawn from a dataset that is not representative of the population you are trying to understand.
- Example:
  - Using an online poll to determine whether students are in favor of online classes.

How to avoid it

- Always try to use a representative sample of your population in your analysis.
- Choose an appropriate and robust sampling method.

DO YOU PREFER CATS OR DOGS?

Asking people at a dog show whether they like cats or dogs.

image source: Sampling Bias, via Geckoboard

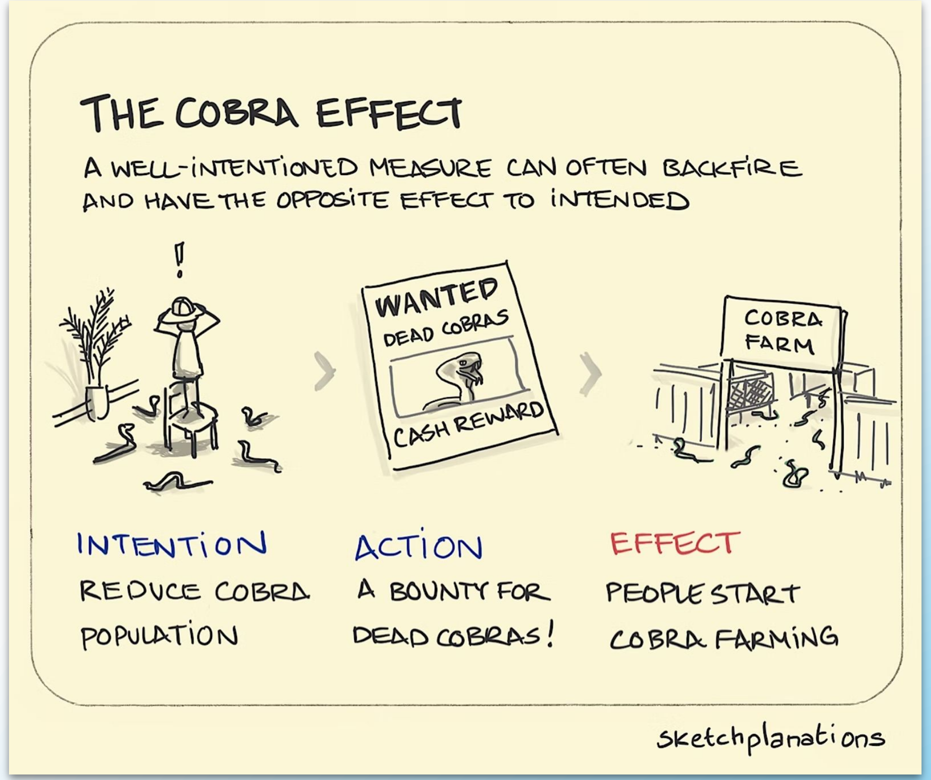
Cobra effect

What it is

- When an attempted solution to a problem somehow makes it worse as an unintended result of using incorrect stimulation or wrong incentives.

How to avoid it

- Be careful what you are incentivizing because incentives generally increase the likelihood of what you are incentivizing.


THE COBRA EFFECT
A WELL-INTENTIONED MEASURE CAN OFTEN BACKFIRE AND HAVE THE OPPOSITE EFFECT TO INTENDED
!
WANTED
DEAD COBRAS
CASH REWARD
COBRA FARM
INTENTION
REDUCE COBRA POPULATION
ACTION
A BOUNTY FOR DEAD COBRAS!
EFFECT
PEOPLE START COBRA FARMING
sketchplanations

The story goes something like this: back in colonial India the top Brit in charge decided there were too many cobras around Delhi. ***To reduce the population they put in place a cash reward, or bounty, for anyone who brought in a dead cobra****. The intention was clear. Legend has it that* ***people did bring in the cobras reliably because some enterprising souls had started breeding cobras for the very purpose of getting the bounty****.*

image source: The cobra effect, via Sketchplanations

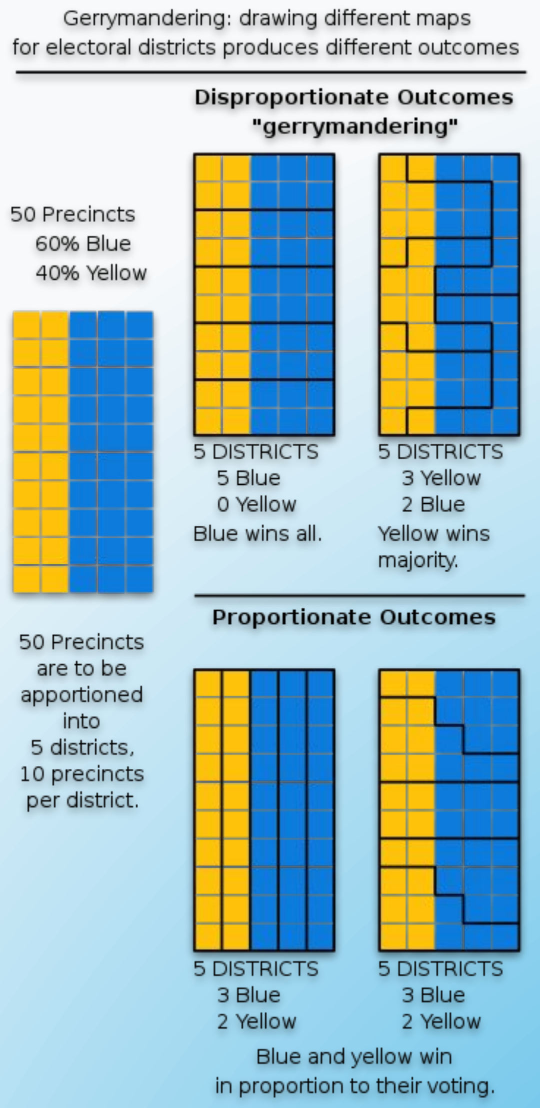
Gerrymandering/MAUP

What it is

- In both gerrymandering and the Modifiable Areal Unit Problem (MAUP), the outcome of an event (e.g. election, analysis) can vary depending on how you divide the area of interest.

How to avoid it

- Always consider the scale and how you group your data when doing your analysis. Try to see if your results also vary when you vary your scale.



Different ways to apportion electoral districts leads to different election results

image source: M.boli, CC BY-SA 4.0, via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:DifferingApportionment.svg

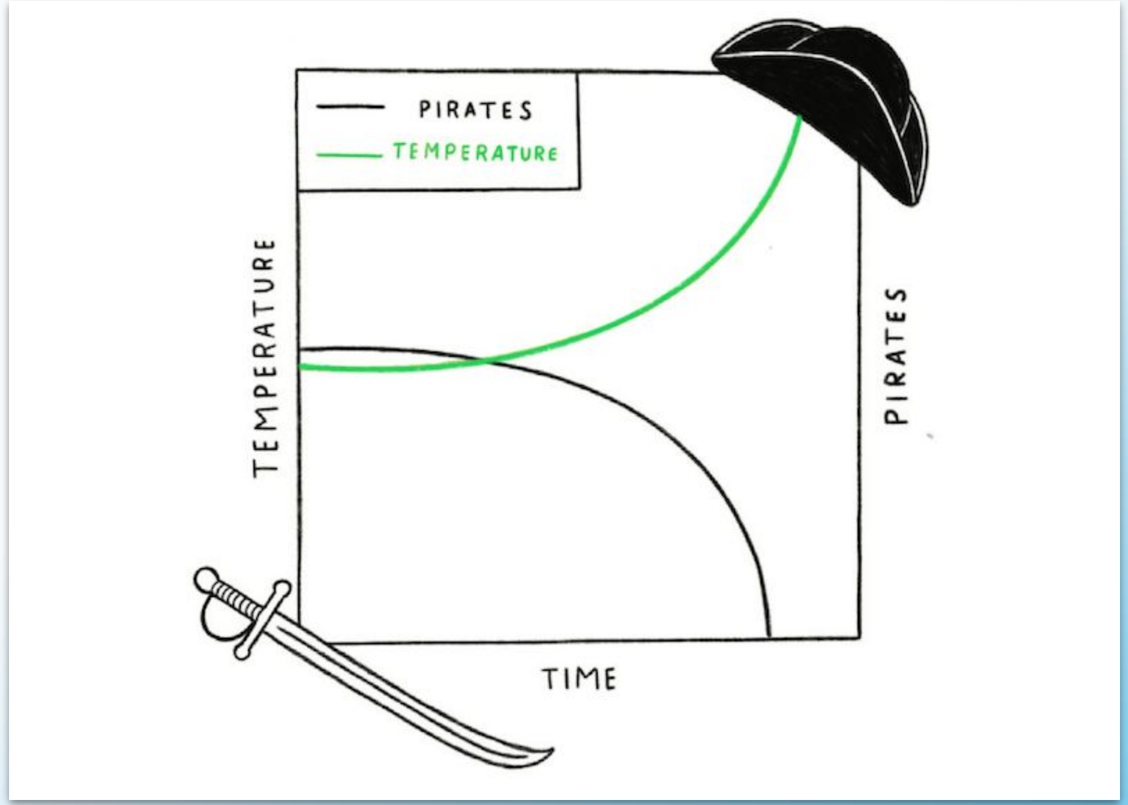
False causality

What it is

- The belief that because two events occur together or immediately after one another then one must have caused the other.
- Correlation does not imply causation.

How to avoid it

- Never assume causation based on correlation alone.



Global temperatures have steadily risen over the past 150 years and the number of pirates has declined at a comparable rate. No one would reasonably claim that the reduction in pirates caused global warming or that more pirates would reverse it.

image source: Sampling Bias, via Geckoboard

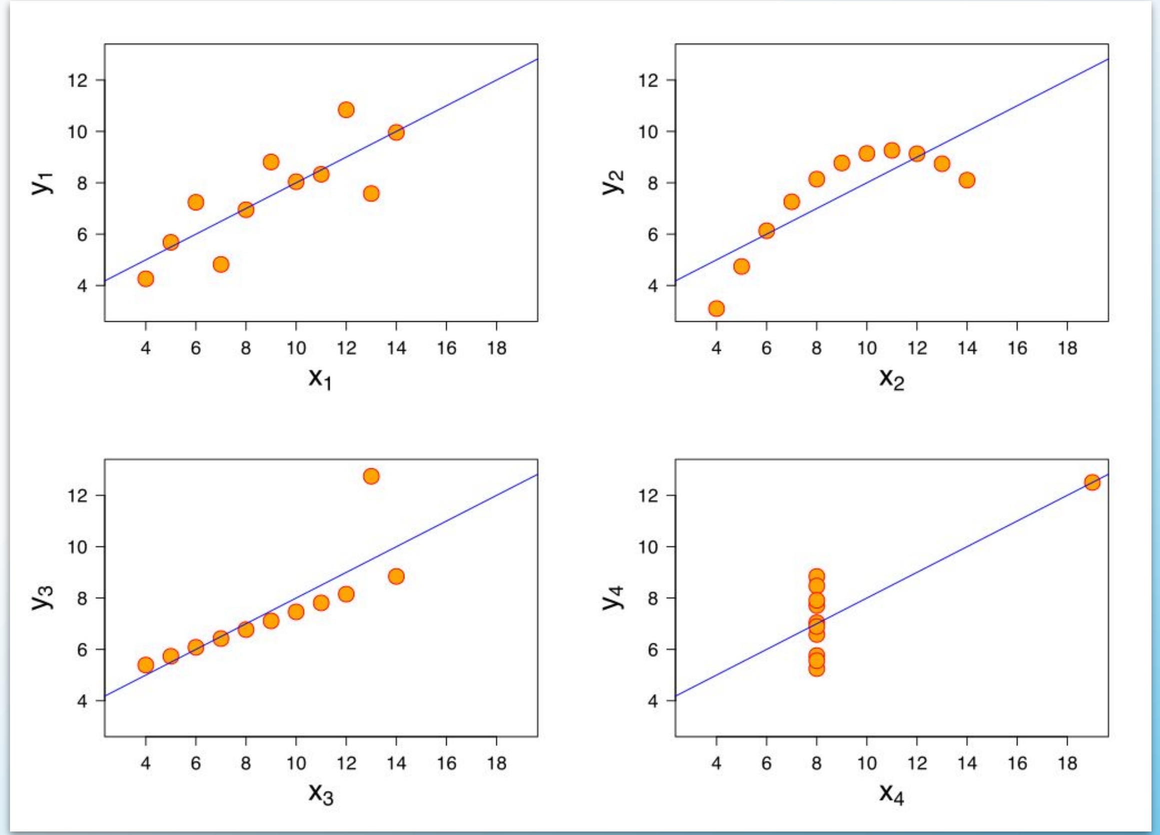
Danger of summary metrics

What it is

- Reliance on summary metrics blur out differences in the dataset. Some datasets may have the same summary metrics (e.g. mean, variance, correlation) but be totally different from each other.
- Example:
  - Anscombe’s quartet

How to avoid it

- As a provider, show or open the data used for the study instead of just the summary statistics.
- As a consumer, always look or ask for the data behind the summary statistics used.



Four different datasets look identical when examined using simple summary statistics, but vary considerably when graphed.

image source: Anscombe's quartet, via Wikimedia user Schutz, via Wikimedia Commons, https://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Anscombe%27s_quartet_3.svg/1024px-Anscombe%27s_quartet_3.svg.png

Other data fallacies

- **Simpson's paradox** - A phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.
- **Gambler's fallacy** - The erroneous belief that if an event occurs more frequently than normal in the past then it is less likely to occur in the future, when it has already been established that the probability of such events do not depend on what happened in the past.
- **Hawthorne effect** - Also known as the Observer Effect. This is the phenomenon where the actions and behaviors of the subjects of a study change because they are aware that they are being observed/monitored.
- **McNamara fallacy** - Being too focused on what can easily observed and assuming that does that cannot are irrelevant. This leads to decisions based solely on quantitative observations (i.e., metrics, hard data, statistics) while all qualitative factors are ignored.
- **Publication bias** - Refers to the fallacy that the outcome of a research or experiment influences whether it is published instead of the robustness of the methodology. This results in an imbalance in published papers in favor of positive results when, in reality, more researches using the same methodology but showing negative or inconclusive results may exist.

Data bias

Data bias

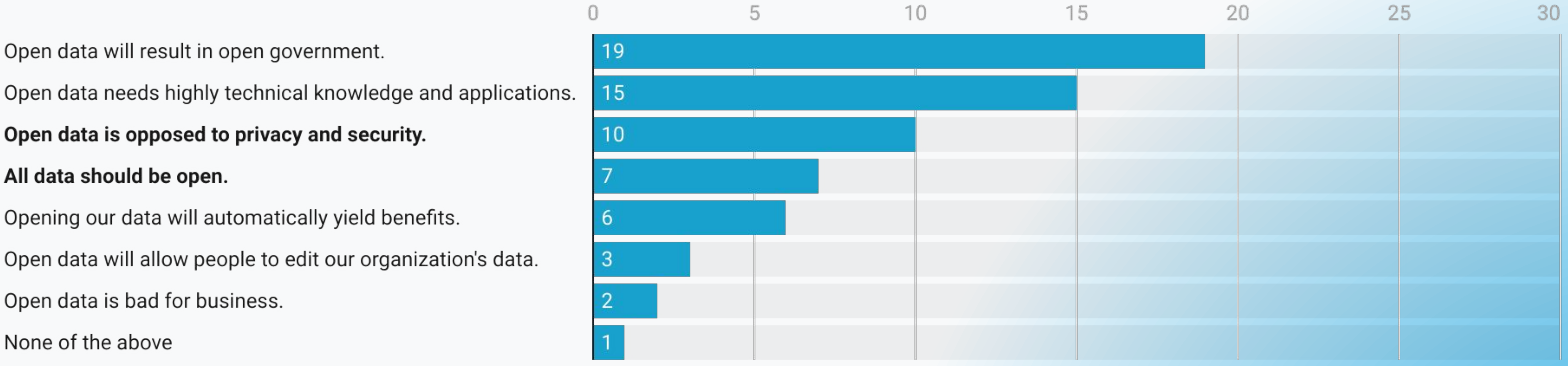
Data bias is an issue in data ethics. This refers to the fact that **human biases are reflected, propagated, and are amplified with data**.

Some examples where bias can arise include:

- Survey questions are constructed with a particular intent/framing
- Selective collection of data from a particular group
- Underlying bias in the data sources

Data privacy and security

Which of these things about open data have you heard about or believe?

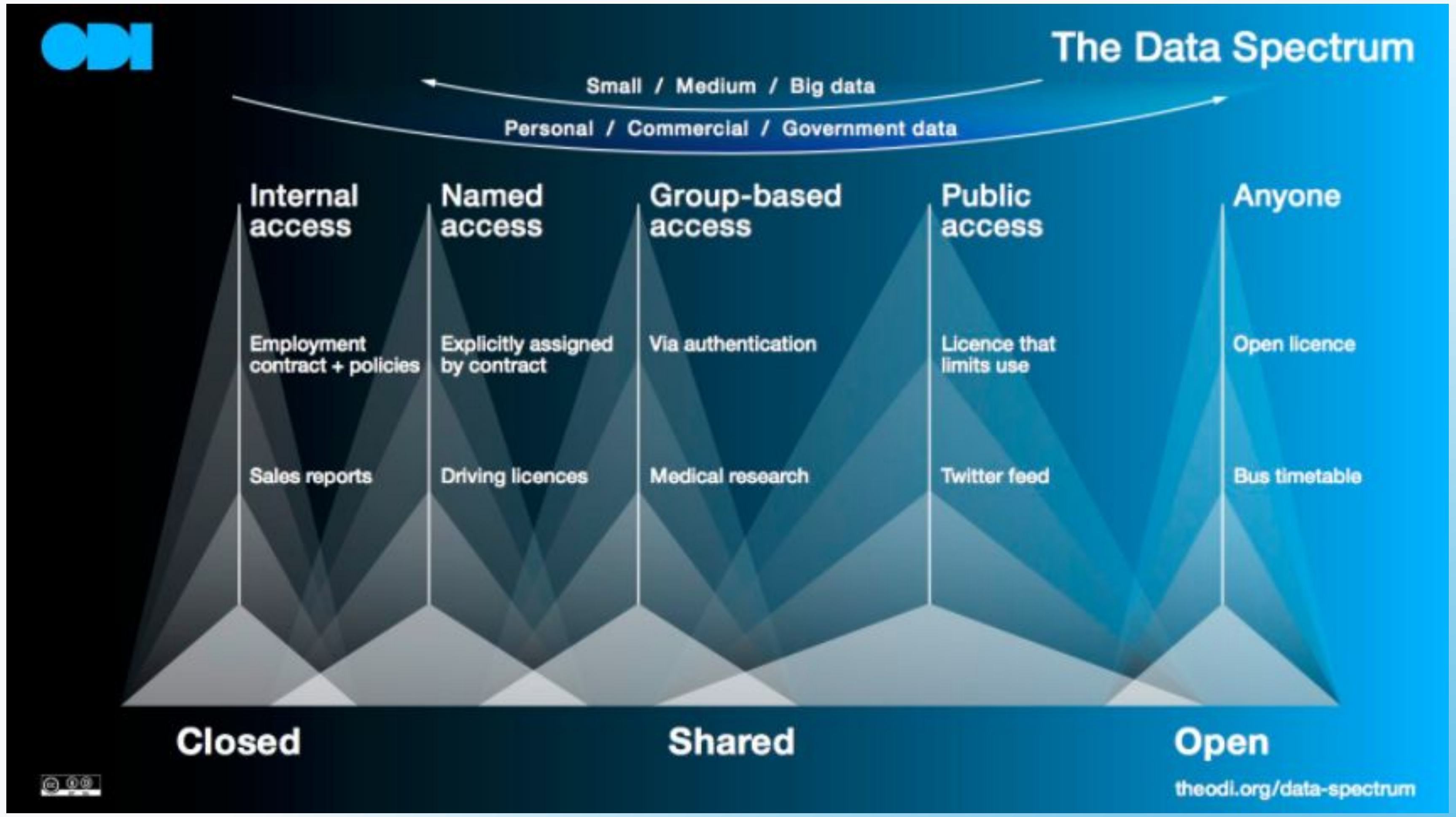


From a total of 30 respondents selecting multiple options

Chart: BNHR • Created with Datawrapper

The data spectrum

Open data doesn’t mean everything is open but respects the rights and privacy of data actors. The openness of data isn't binary, it's a spectrum that ranges from closed to shared to public.



The Data Spectrum by the ODI.

In the Philippines...

- Data Privacy Act of 2012 (Republic Act 10173)
- **Personal information**—any information whether recorded in a material form or not, from which the **identity of an individual is apparent** or can be **reasonably and directly ascertained** by the entity holding the information, or when **put together with other information** would **directly and certainly identify an individual**
    - full name + photograph
    - full name + address

In the Philippines... (1)

- **Privileged information**—any and all forms of data which under the Rules of Court and other pertinent laws constitute privileged communication
- **Sensitive personal information**—personal information:
  - About an individual's race, ethnic origin, marital status, age, color, and religious, philosophical or political affiliations;
  - About an individual's health, education, genetic or sexual life of a person, or to any proceeding for any offense committed or alleged to have been committed by such person, the disposal of such proceedings, or the sentence of any court in such proceedings;
  - Issued by government agencies peculiar to an individual which includes, but not limited to, social security numbers, previous or current health records, licenses or its denials, suspension or revocation, and tax returns; and
  - Specifically established by an executive order or an act of Congress to be kept classified.

In the Philippines... (2)

- **Transparency**—the user must know what data is being processed, how and why it is being processed, and consent to the processing.
- **Legitimate purpose**—the purpose of the processing must be legitimate and necessary.
    - processing data for employment, insurance, government ID
- **Proportionality**—the amount and kinds of data being processed must be proportionate to the purpose of the processing; the amount of time that the data is stored must also be proportionate to the purpose.
    - collect the least amount of personal information needed to accomplish the objectives

Improving data privacy and security

- Individual or enterprise (organizational)?
    a. Personal Digital Privacy and Security: Paano maging private, secure, at anonymous sa internet
- Varies depending on the person or organization (Threat modelling)

Threat modelling

Five questions:

1. **What** do I want to protect?
2. **Who** do I want to protect it from?
3. **How likely** is it that I will need to protect it?
4. **How bad** are the consequences if I fail?
5. **How much** trouble am I willing to go through to try to prevent potential consequences?

Some considerations

- **Digital privacy and security are shared responsibilities**—everyone plays a role in your and your organization's digital security and privacy.
- **You cannot protect yourself against everything all the time**—what's more important is for you to understand the threats you face and how you can counter them.
- **There is no such thing as fully protected just better protected**—you can be compromised even w/o your fault:
    - Mass data breaches
    - Data brokering
    - People close to you being compromised
    - Social engineering

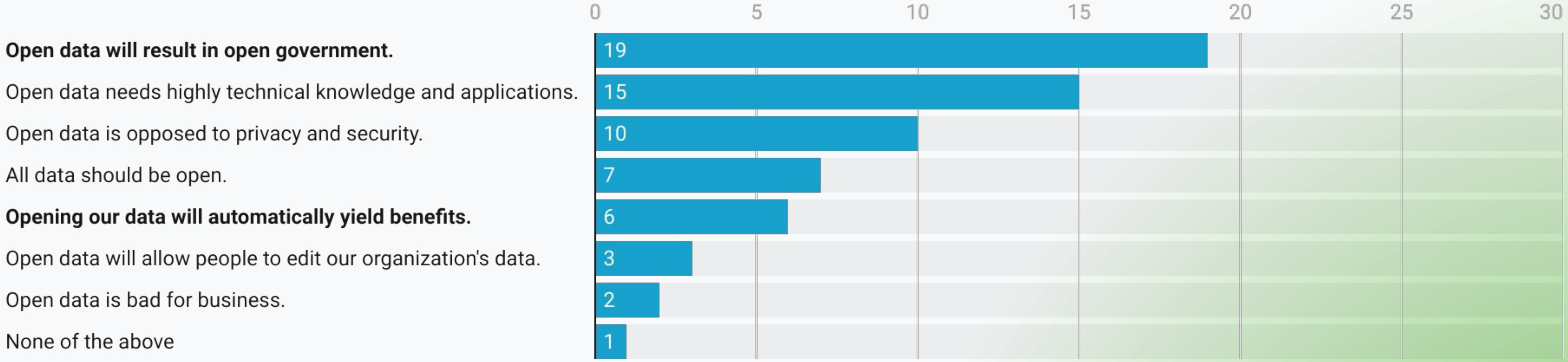
Beware of openwashing

Openwashing

- To **spin a product or company as open, although it is not**. Derived from 'greenwashing.' ([Michelle Thorn])
- (n.) **Having an appearance** of open-source and open-licensing **for marketing purposes,** while continuing proprietary practices. ([Audrey Watters])
- Just because an organization or agency says they are open or smart does not mean that they are.

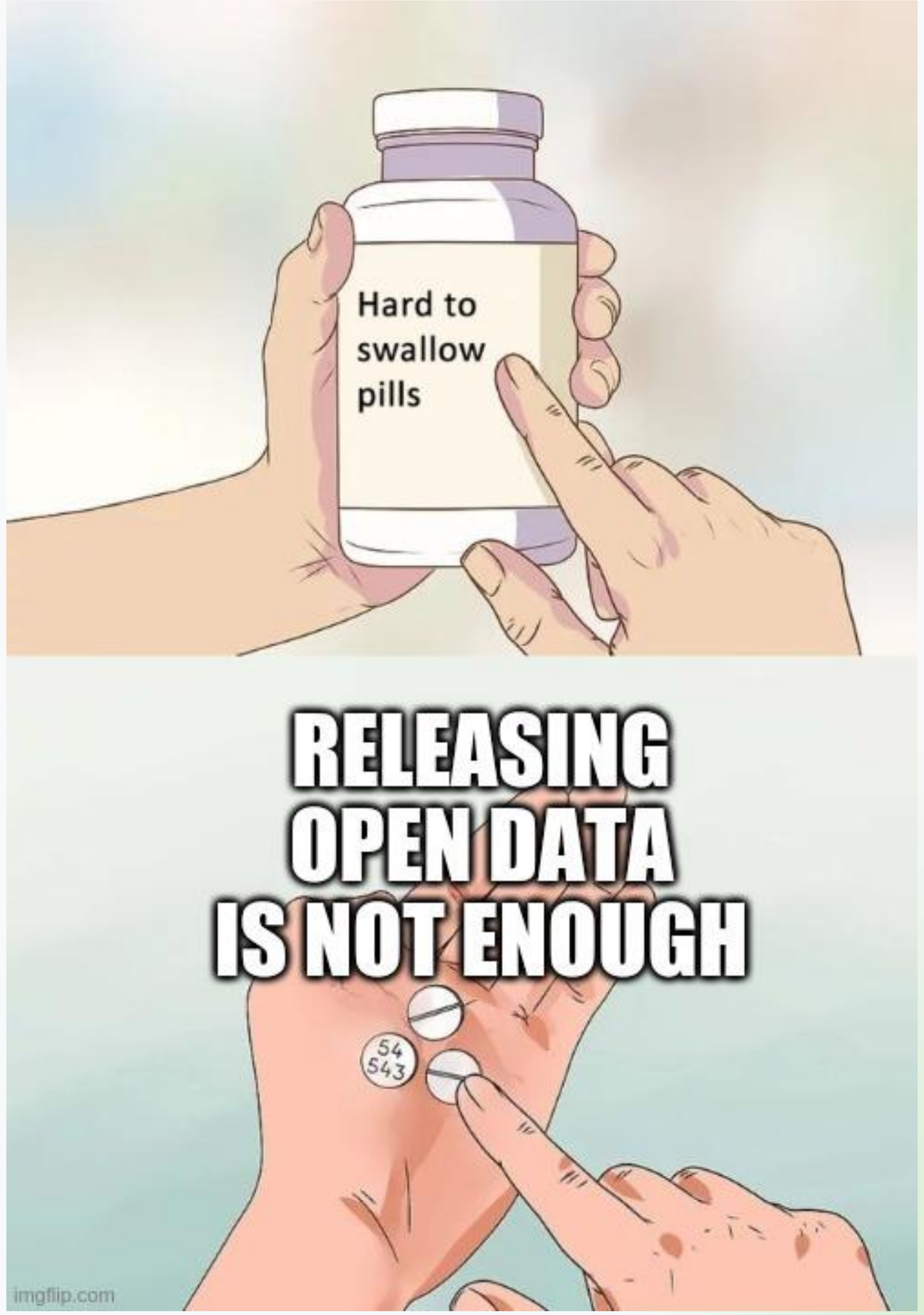
How to: share and open data

Which of these things about open data have you heard about or believe?



From a total of 30 respondents selecting multiple options

Chart: BNHR • Created with Datawrapper


Hard to swallow pills
RELEASING OPEN DATA IS NOT ENOUGH
54 543
imgflip.com

It does not end with data release

- The end goal of any open data project **isn’t simply to release data**.
- We do not set out to create **data dumps**.
  - Data portals that hold data but aren’t updated or whose data aren’t being used.
- An open data project should **spark conversation, facilitate communication** between the data provider and the data user, and **disturb the status quo**.
- **Open data without users is no better than closed data**.

Be open by design

Deliberate and conscious openness

- Open data isn't an afterthought but as a core principle.
- Give permission in advance
- Don't make it hard to work with your data or tech. Make it available and discoverable.
- Use open standards and practice interoperability.
- Prefer open source instead of proprietary and closed solutions.
- Use an open license.
- Co-develop and co-create from the start. Don't make it hard for people to collaborate with you.
- See also How to publish open data: a list of advice and tools

Be open by default

- Assumption is open first then find restrictions (not the other way around)
- Aside from ethical and legal considerations (e.g. constitutionally and statutorily protected data), public data and data collected by public institutions should be open.

Co-creation

People before systems (1)

- Design for people not machines
- Engage early and engage often
  - Identify your key stakeholders and assumptions early.
  - Identify the needs and use-cases of your stakeholders.
  - Give your users a say in the design and development process
  - Validate your assumptions with actual users
  - Iterate

People before systems (1)

- Utilize user and people-centered design tools
  - Persona exercises
  - User Stories
  - Journey Mapping
  - UI Sketching
- Ensure data is used after release
  - Provide incentive for users to use the data.
  - Focus on telling compelling stories rather than just the data.
  - Forge collaborations among the data users and data providers.
  - Build grassroots communities around open data.

Co-creation isn’t…

- Putting different people in one room and expecting them to generate a solution themselves.
- Gathering and consolidating ideas from different people.
- Finding the “middle-ground” solution between stakeholders who have varying, and oftentimes competing, interests.

Co-creation is about...

- Confronting the inherent power relations and inequalities between your stakeholders. Those who have more must be willing to give-up more and those with the most to lose should have a say in things.
- Acknowledging historical injustices and their present-day impacts.
- Helping everyone develop a shared point of view no matter how difficult it is.

Five-star open data

5 ☆ Open Data

- Suggested by Sir Tim Berners-Lee (inventor of the WWW) as a deployment scheme for open data.
- Stars are awarded according to the state of open data
- Provides a good framework for assessing where you are in terms of opening your data and what steps to take in order to move forward.
- https://5stardata.info/en/

5 ☆ Open Data

- ★ Data is available *(in whatever format)* under an open license
- ★★ Data is available as structured, machine-readable data *(e.g. Excel instead of PDFs)*
- ★★★ Data is available in a non-proprietary open format *(e.g. CSV instead of Excel)*
- ★★★★ Uniform Resource Identifier (URI) is used to identify things so others can point to your data
- ★★★★★ Data is linked to other data in order to provide context

Brainstorming:

procurement-themed data-driven project

Brainstorming a data-driven project (1)

Objectives

1. For you to create a plan/proposal for a procurement-related data-driven project using the data pipeline as a template.

General instructions

- You will be divided into groups/breakout rooms
- You will be tasked with brainstorming a procurement-related data-driven project using the data pipeline as a template (120 mins)
    - During this time, I will join each of the breakout rooms in case you need help or have questions and clarifications.
- Each group will be asked to give presentation about your group's data-driven project (5-7 minutes per group)

Brainstorming a data-driven project (2)

Specific instructions

- Make a copy of the Data-driven project brainstorming document.
- Share with me a link to your group's brainstorming document.
- Discuss within your group what project you want to propose and add the necessary information in the brainstorming document.
- Create a presentation about your project that includes all the important information—you may refer to this template (you can replace the theme).
- Present your project.

Thank you!

BN
HR

openness.
data.
geospatial.

- @bnhr.xyz
- @bnhrdotxyz
- @bnhrdotxyz
- bnhr.xyz