THEGOVLAB

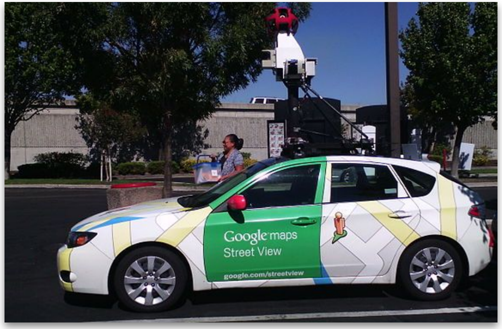
Revolutionizing Transportation Planning Responsibly

How to advance Data Collaboration in a systematic, sustainable and responsible way

Stefaan Verhulst

September 2024

TODAY'S PUBLIC PROBLEMS REQUIRES INNOVATION IN HOW WE MAKE DECISIONS

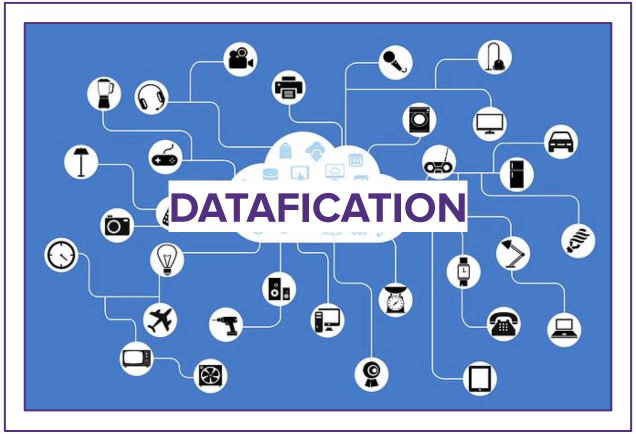


DATAFICATION

Volume

Veracity

DIGITAL TWIN



ARTIFICIAL INTELLIGENCE

Velocity

Variety

VARIETY OF DATA

STRUCTURED

Fixed Fields
Relational Database
Spreadsheets

Sales data, Birthdates,
Zip codes

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

SEMI-STRUCTURED

Tagged/metadata
XML or HTML tagged
text

Email, RSS feeds

```
<University>
 <Student ID="1">
   <Name>John</Name>
   <Age>18</Age>
   <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
   <Name>David</Name>
   <Age>31</Age>
   <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

UNSTRUCTURED

No Fixed Fields
Various formats, sizes
and structures

Texts, Audio, Pictures,
Social Media

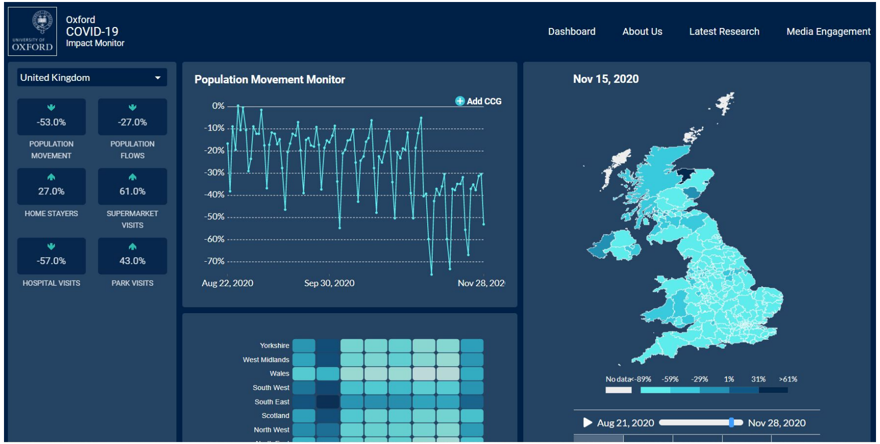
The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

‘NON-TRADITIONAL’ DATA

Non-traditional data (NTD) refers to data that is:

- Digitally captured, mediated or observed;
- Using new instrumentation mechanisms;
- Often privately held, and;
- Can be re-used for purposes unrelated to its initial collection.

Example: Oxford Covid-19 Impact Monitor



The Oxford COVID-19 Impact Monitor used mobile phone data to understand, predict and control the course of Covid-19.

Definition of non-traditional data derived from https://www.gdi.manchester.ac.uk/research/publications/di/dd-wp92

Example from Oxford University and the Internet Archive: https://web.archive.org/web/20201109160135/https://oxford-covid-19.com/

NON-TRADITIONAL DATA

- Data resulting from **consumption, commercial and financial transactions**;
- Data resulting from **communicating and engaging in social interactions** - for pleasure, study and/or work;
- Data resulting from having people and products **moving around**;
- Data resulting from **media and entertainment consumption**;
- Data emerging from **producing products and goods**;
- Data emerging from **managing infrastructures and natural assets**.

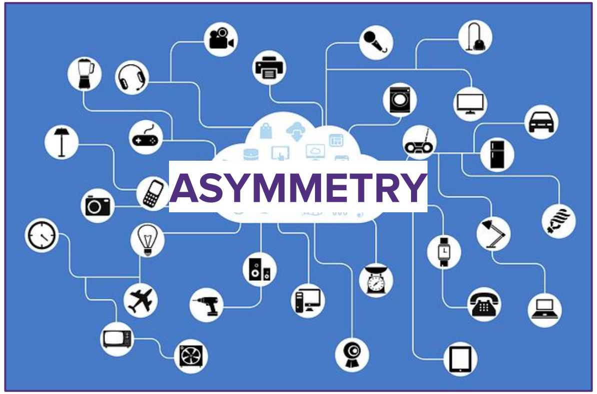
THE NEED FOR COLLABORATION

Volume

MIS-USE

Veracity

DEMAND



SUPPLY

Velocity

RE-USE

Variety

DATA COLLABORATIVES: MATCHING DEMAND & SUPPLY

DATA COLLABORATIVES: OPERATIONAL MODELS

- Public Interfaces
- Data Pooling
- Prizes and Challenges
- Trusted Intermediary
- Intelligence Generation
- Research and Analysis Partnerships

PUBLIC INTERFACES

A single data holder provides access to certain types of pre-processed and/or data-driven tools for public use.

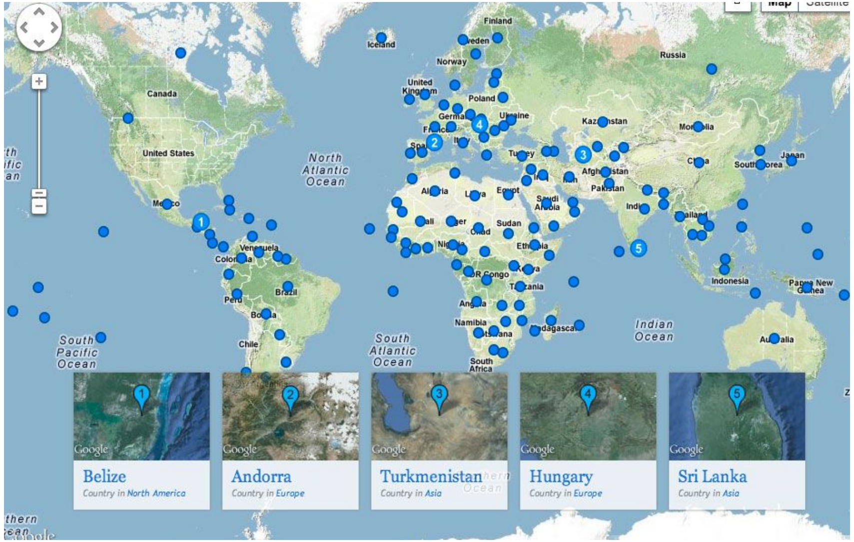
Data Holders: One data holder

Data Users: Many data users

Main Types:

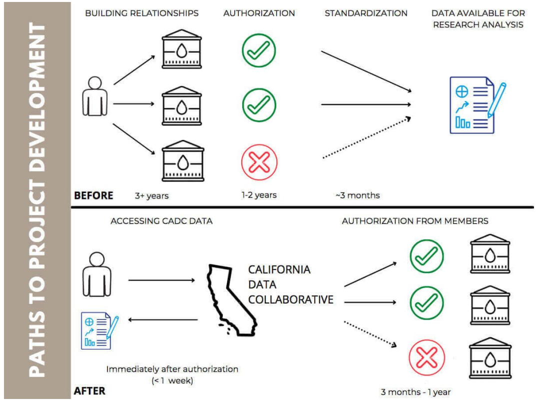
- Application Programming Interfaces (APIs)
- Data Platforms

Google Earth Outreach



API: Google Earth Outreach

CALIFORNIA DATA COLLABORATIVE



Private Data Pool: California Data Collaborative

DATA POOLING

Data holders pool datasets as a collection designed to be accessible by multiple parties.

Data Holders: Some data holders
Data Users: Some data users

Main Types:

- Public Data Pools
- Private Data Pools

PRIZES & CHALLENGES

Data holders make data available to participants who compete to solve problems or pioneer innovative uses of data for the public interest.

Data Holders: One or more
Data Users: Many data users

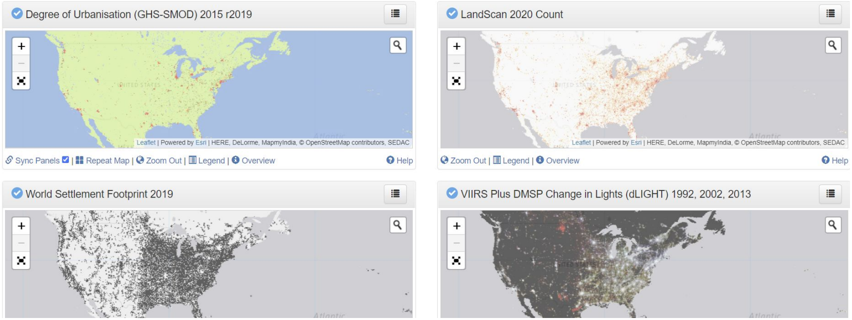
Main Types:

- Open Innovation Challenges
- Selective Innovation Challenges

Türk Telekom
Değerli Hissettirir

Selective Innovation Challenges:
Türk Telekom’s Data 4 Refugees Challenge

POPGRID
DATA COLLABORATIVE



Data Brokerage: The POPGRID Data Collaborative

TRUSTED INTERMEDIARIES

Third-party actor mediates collaboration between (private sector) data providers and data users from the public sector, civil society, or academia.

Data Holders: One or more
Data Users: One or more

Main Types:

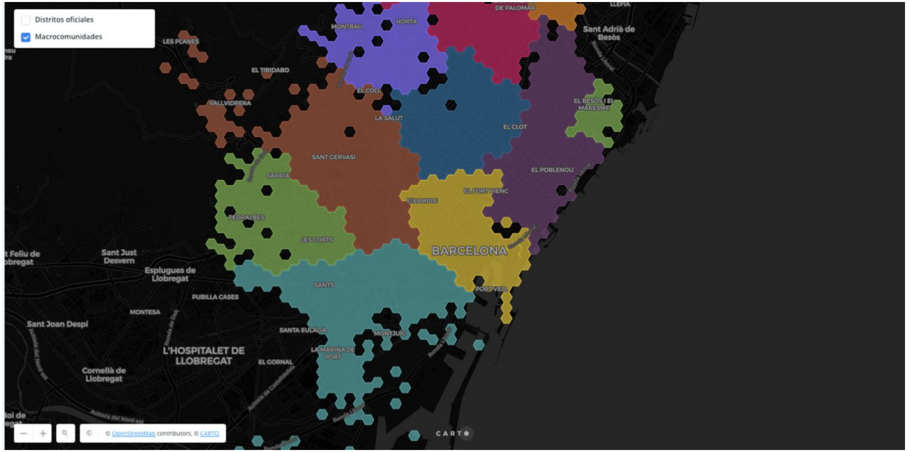
- Data Brokerage
- Third-Party Analytics

INTELLIGENCE GENERATION

No data is shared with external parties, instead the results of analysis within the data holder's organization are shared with external actors.

Data Holders: One data holder

Data Users: No external data users

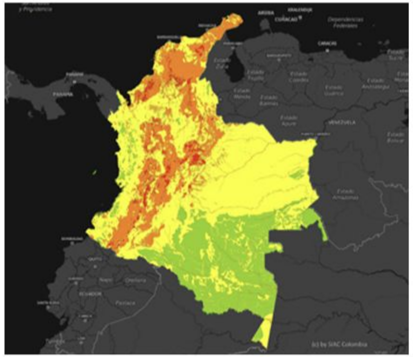


Intelligence Generation: BBVA Urban Discovery

Building communities resilient to climatic extremes

Figure 1

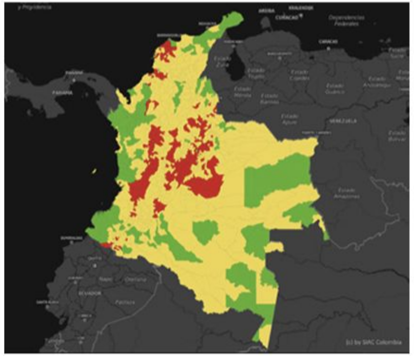
Potential Impact of Climate Change in Colombia 2011-2040.



Source: Sistema de Información Ambiental de Colombia.

Figure 2

Adaptation capacity for municipalities of Colombia to Climate Change



Source: Sistema de Información Ambiental de Colombia.

Data Transfer: UN FAO's Building Communities Resilient to Climatic Extremes

RESEARCH & ANALYSIS PARTNERSHIPS

A pairing between (private sector) data providers and data analysts or data users from the public sector, civil society or academia.

Data Holders: One or more
Data Users: One or more

Main Types:

- Data Transfer
- Data Fellowship

CHALLENGES FACING DATA COLLABORATIVES

Lack of Awareness & Data Literacy

There often exists a lack of awareness and appreciation regarding the potential of data sharing.

Limited Capacity

Organizations can lack technical knowledge, financial resources, or simply the awareness needed to participate in a collaborative.

Absence of Trust

The field of data sharing is characterized by a pervasive absence of trust. There may be value in a data sharing framework to address.

Transaction Costs

Preparing data, de-risking data, preparing legal agreements, and establishing governance structures all take resources.

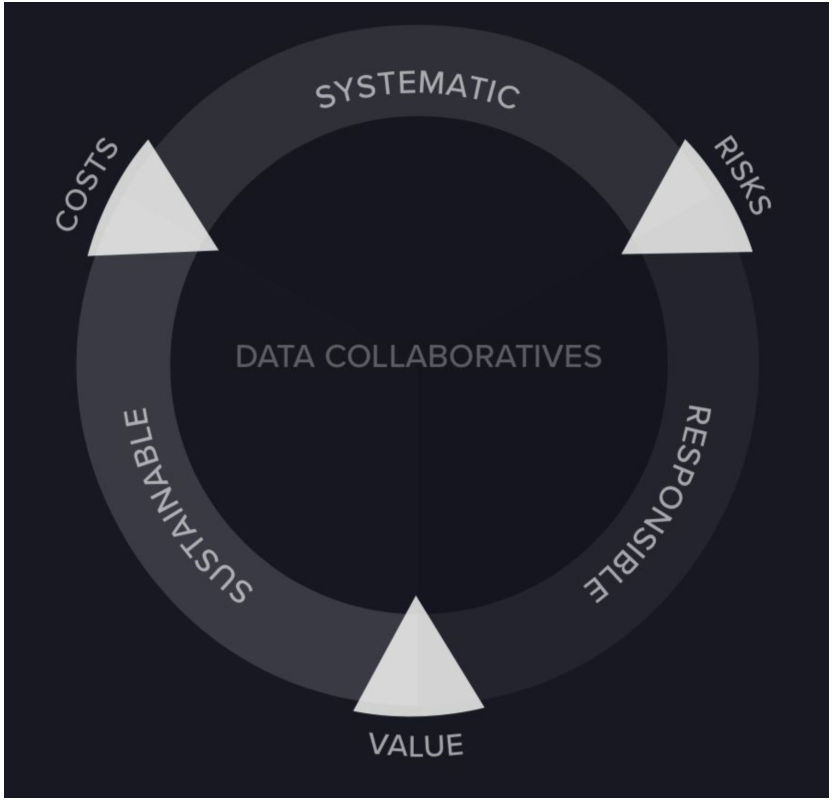
Uncertainty and Unclear Incentives

Organizations, particularly in the private sector, can have concerns that data reuse won't advance goals but will instead lead to risks (e.g. data leaks, reputational loss)

Limited Community of Practice

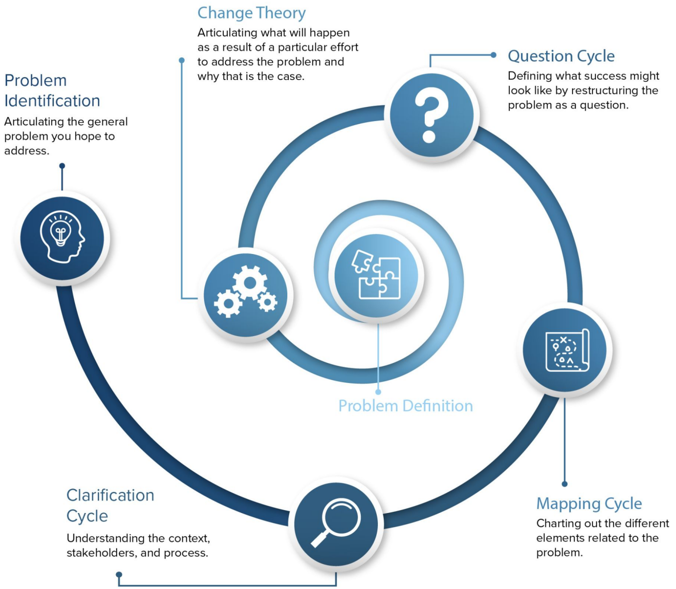
Successful initiatives need a community of practice that can provide an established knowledge base (e.g. case studies, lessons learned).

HOW TO BE MORE SYSTEMATIC SUSTAINABLE, AND RESPONSIBLE



Pathways Toward Systematic, Sustainable and Responsible Data Collaboration

1. STRENGTHEN DEMAND



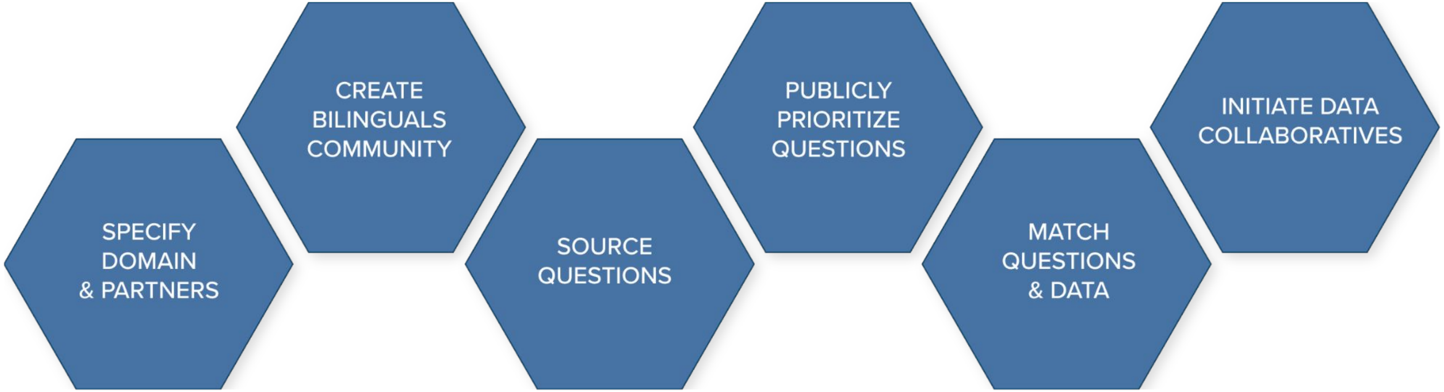
the100questions.org

bit.ly/ODPLProbTool

DEVELOP A NEW SCIENCE OF QUESTIONS

https://the100questions.org/

DEVELOP A NEW SCIENCE OF QUESTIONS

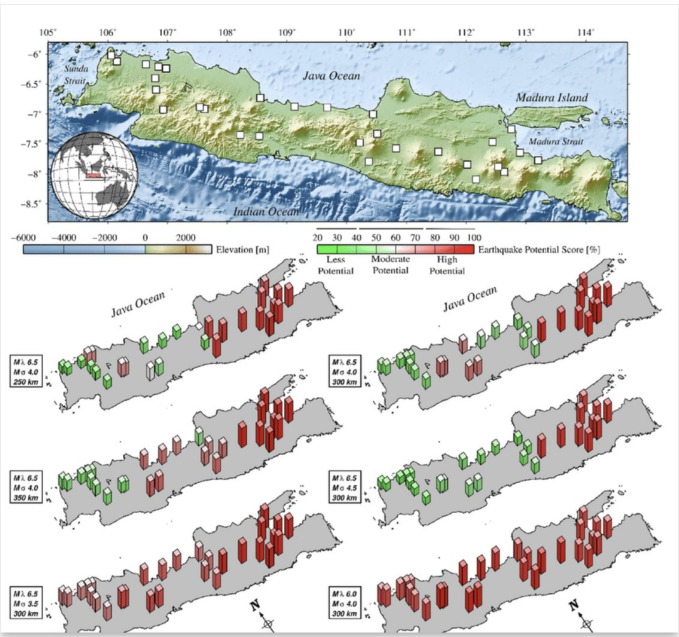


https://the100questions.org/

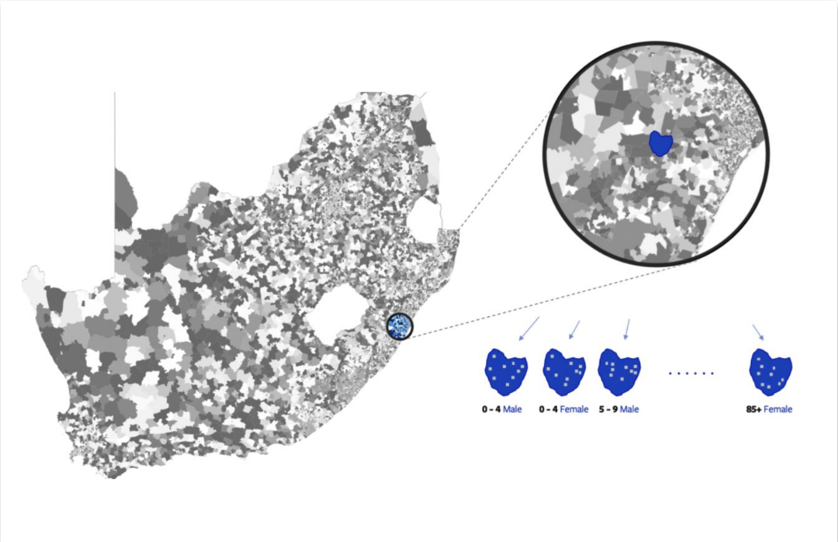
TAXONOMY OF QUESTIONS

BACKWARD LOOKING	SITUATION ANALYSIS **DESCRIPTIVE** *WHAT HAPPENED?*	CAUSE AND EFFECT **DIAGNOSTIC** *WHY DID IT HAPPEN?*
FORWARD LOOKING	FORECASTING **PREDICTIVE** *WHAT WILL HAPPEN?*	EXPERIMENTATION (WHAT IF?) **PRESCRIPTIVE** *WHAT SHOULD HAPPEN?*

DESCRIPTIVE

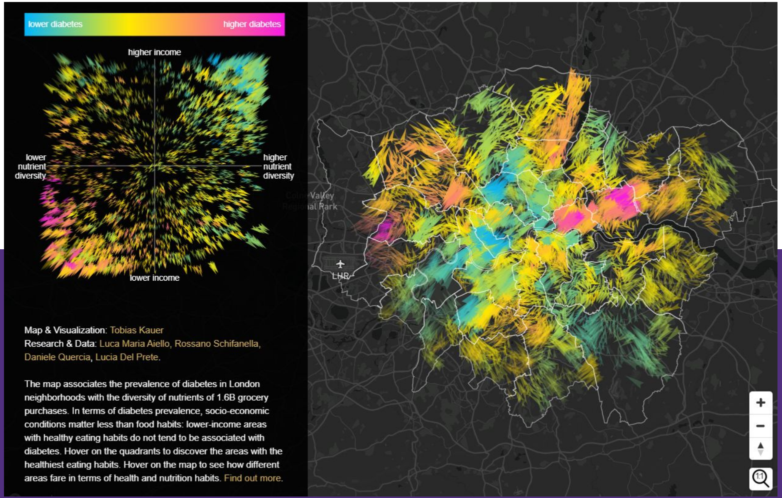


Nowcasting

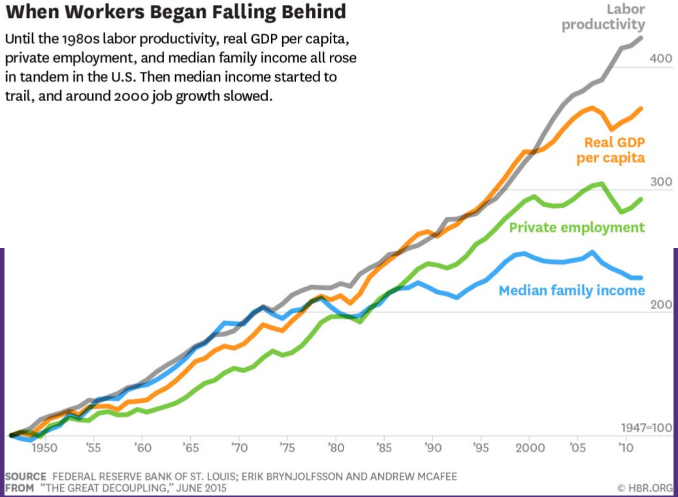


Population Density

DISRUPTIVE DESCRIPTION

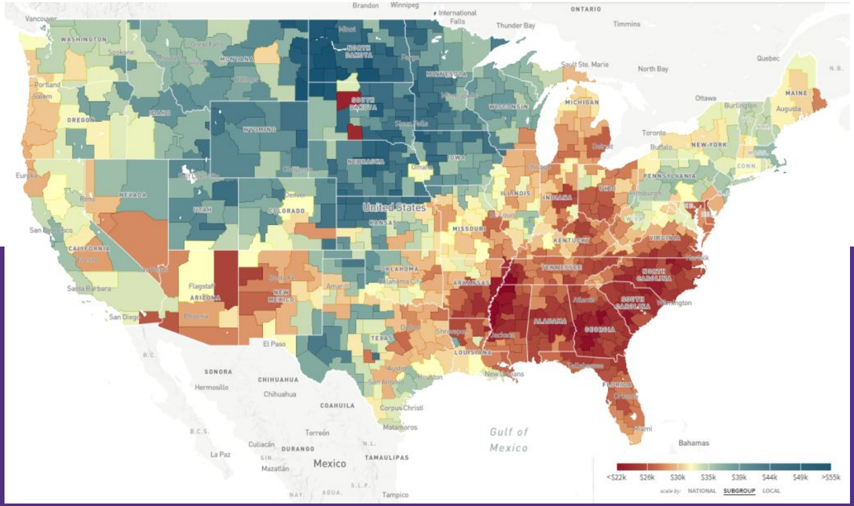


Poor But Healthy Visualization
https://goodcitylife.org/food/

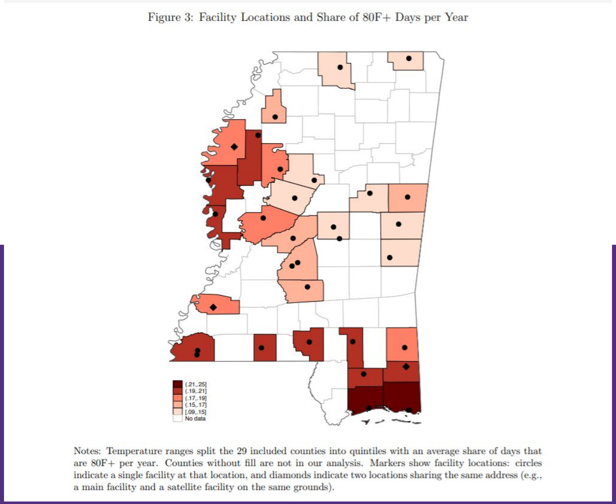


The Great Decoupling, HBR
https://bit.ly/3ibN5Jn

DIAGNOSTIC

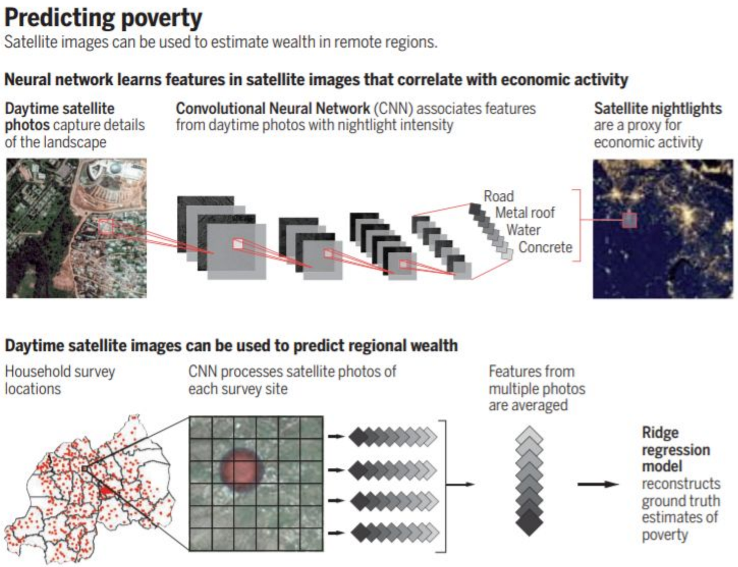


Neighborhoods Matter, Opportunity Insights
https://opportunityinsights.org/neighborhoods/



The Causal Effect of Heat on Violence: Social Implications of Unmitigated Heat Among the Incarcerated
https://www.nber.org/papers/w28987

PREDICTIVE



Fighting poverty with data, *Science*
https://bit.ly/36FfORG

Special Collection: Lessons learned from telco data informing COVID-19 responses: toward an early warning system for future pandemics?

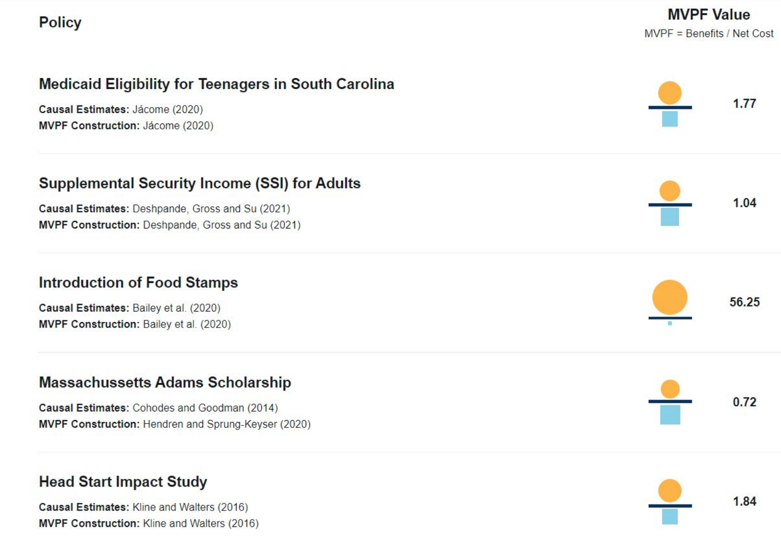
Data & Policy Blog Follow
Jun 28 · 7 min read

Photo by Jeremy Zero on Unsplash

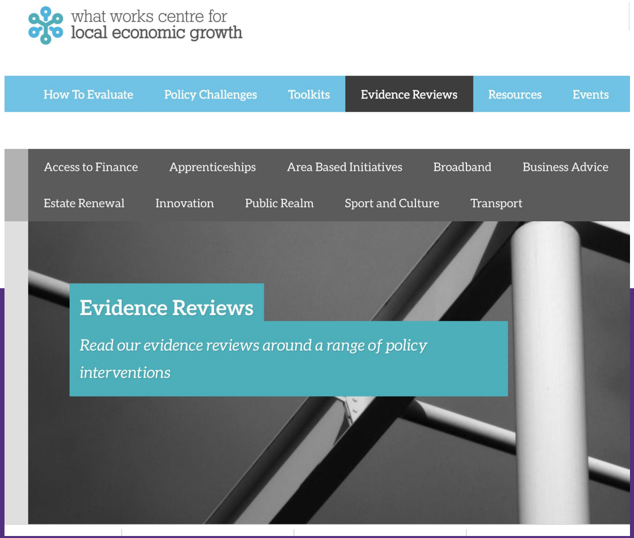
In this blog, Guest Editors Richard Benjamins (Telefónica), Jeanine Vos (GSMA) and Stefaan Verhulst, Data & Policy Editor-in-Chief, introduce the first set of peer-reviewed, open access articles in a Data & Policy special collection dedicated to Telco Big Data Analytics for COVID-19.

Predicting future pandemics with data
https://www.cambridge.org/core/journals/data-and-policy

PRESCRIPTIVE



The Policy Impacts Library
https://bit.ly/36IFUD2



What Works Center
https://whatworksgrowth.org/

2.STEWARD THE SUPPLY

https://datastewards.net/

INVEST IN DATA STEWARDS

NURTURE DATA COLLABORATIVE TO SUSTAINABILITY

- Strategize for scaling and sustaining data collaboratives
- Share insights to build the societal and business case for data collaboration

INTERNAL COORDINATION AND STAFF ENGAGEMENT

- Gain approval from and coordinate the actors within the company
- Map and match staff with skills to positions within the collaboration

DISSEMINATION AND COMMUNICATIONS OF FINDINGS

- Raise awareness of findings
- Communicate with actors on issues such as regulatory compliance and contractual obligations

PARTNERSHIP AND COMMUNITY ENGAGEMENT

- Vet and engage with possible partners
- Inform beneficiaries of the insights generated

DATA AUDIT, ETHICS, AND ASSESSMENT OF VALUE AND RISK

- Assess the value and risk of using data
- Consider the ethical implications and validate ways to measure impact

https://bit.ly/3Ok5jHq

DATA AUDIT (MVDpoints)

WHAT	• What type of data can be used? (modality) • Where can the data originate from? (source)
HOW	• How does the data need to prepared / transformed? • How does the data need to be stored / managed?
WHEN	• How old can the data be? (independently of relevance) • How long is it ok to use it?
WHAT CONTEXT	• What application can the data and its underlying applications be used for?

DATA QUALITY

- Accuracy
- Precision
- Relevance
- Transparency
- Timeliness
- Consistency

Datasheets for Datasets 15

Movie Review Polarity — **Thumbs Up? Sentiment Classification using Machine Learning Techniques**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

[1] All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

> these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

Is there a label or target associated with each instance? If so, please provide a description.

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

GEN AI DATA: Garbage in/Garbage out

Training Data (Inputs)

Any data used to develop an AI (pre-training and post-tuning)

- Raw data (often agentless)
- "Human" data

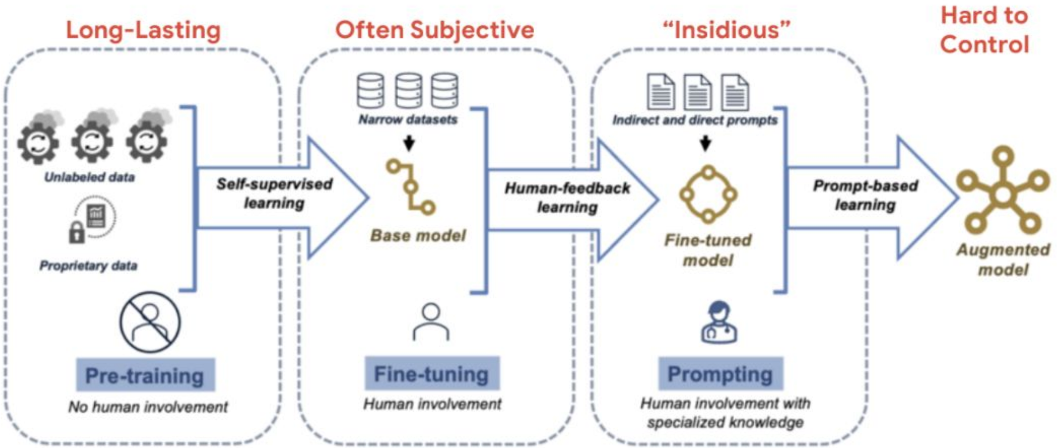
Prompts (Interactions)

Data

Generated Data (Outputs)

Output of a trained AI

- By extension, synthetically generated data



Making Data “AI ready”

	“Classical” Data Science / ML	Gen AI / LLMs
Model type	Supervised	Semi-Supervised ● “Deep” architectures more likely to hallucinate
Data type	Structured	Unstructured
Data quantity	M (depends on model)	XL (depends on architecture)
Modality	Defined, specific	Any, mixed
Provenance	Often enterprise-owned	Anywhere, can be open source
Use case	Targeted use cases	Generative AI, AGI = potentially any

SANDBOXES

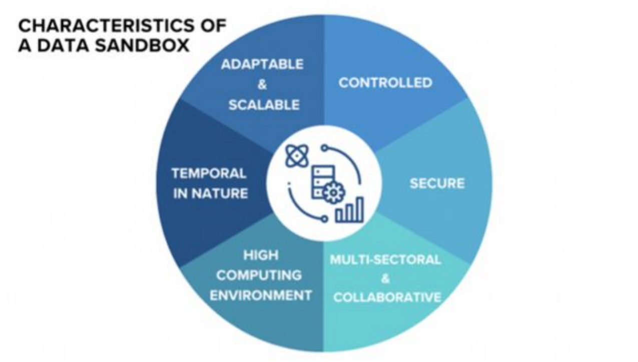
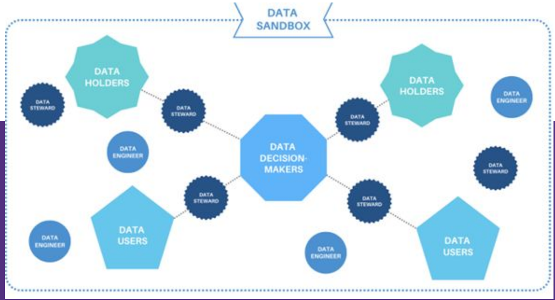


Figure 1. The six characteristics of a data sandbox. Image by The GovLab.



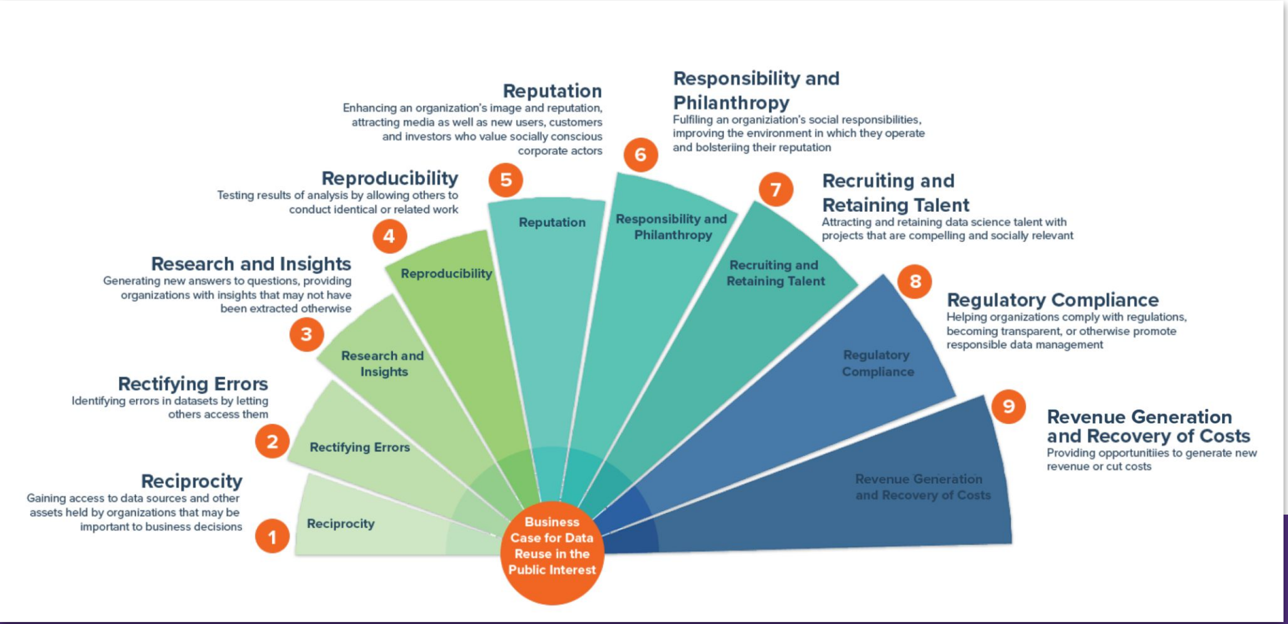
bit.ly/3QyBjL8

3. MATCHING SUPPLY AND DEMAND

BLT

- *Business*
- *Legal and Governance*
- Technical

BUSINESS CASE FOR PROVIDING ACCESS



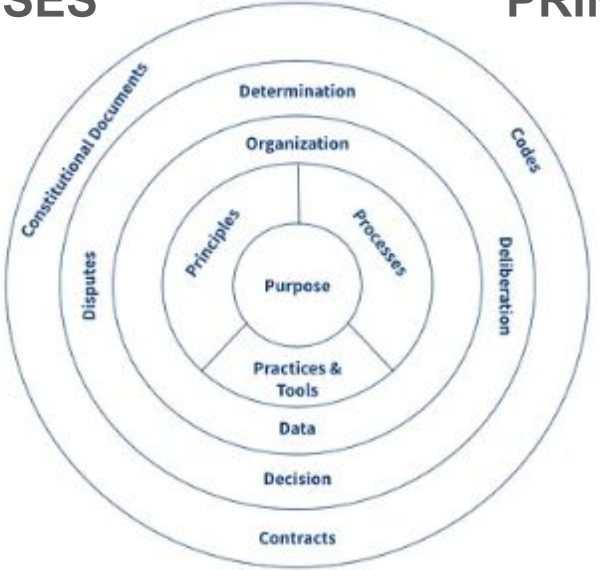
bit.ly/3QyBjL8

GOVERNANCE

PROCESSES

PRINCIPLES

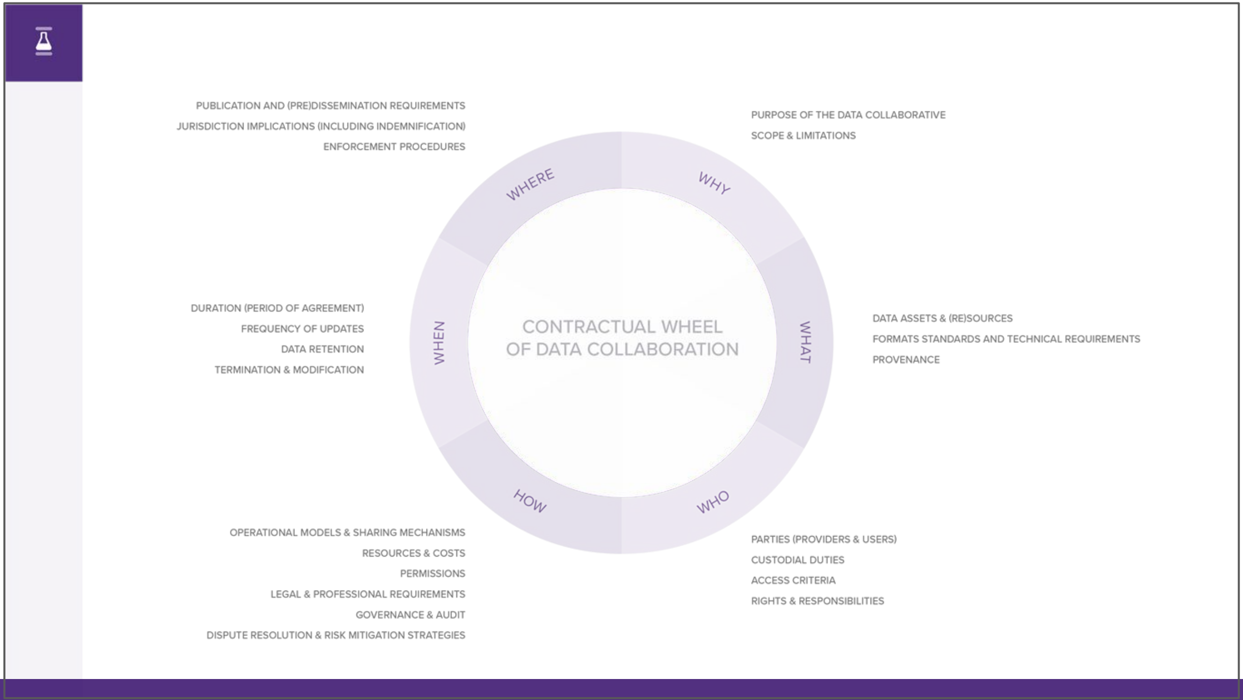
LEGITIMACY/
TRUST



EFFECTIVENESS/
AGILITY

PRACTICES

GOVERNANCE: DATA SHARING AGREEMENTS



https://contractsfordatacollaboration.org/

4. STRENGTHEN THE SOCIAL LICENSE

https://thedataassembly.org/

The Data Stewardship Canvas

Designed by:

Date:

Version:

The Data Stewardship Canvas is a step by step process that maps a data steward's journey when building a data collaborative to support data re-use—whether the data steward is requesting or providing access to data. The steps of the canvas seek to create a systematic and responsible approach to effectively re-using data for positive social and economic outcomes.

1. Defining the Demand for Data

- What is the problem you seek to solve?
- Do you need to scope out the domain using a topic map?
- Are there certain issues you ought to prioritize based on their need, externalities and feasibility?
- What is the guiding question leading this project?

In the Toolkit:

Problem Definition Tool

R-Search Methodology

Open Data Demand Assessment and Segmentation Methodology

2. Defining the Supply of Data

- What are the expertise and capacity needs for this project?
- What is the minimal amount of data needed to make progress towards answering the question?
- What are the different data sources available for this project?

In the Toolkit:

RD4C Data Ecosystem Mapping Tool

The Periodic Table of Open Data: A User's Guide

3. Making a Value Proposition

- What is this project's value to society?
- What is the return on investment of this project?
- Do the benefits of this project outweigh the costs?

In the Toolkit:

A User's Guide to the 9Rs Framework.

Cost-Benefit Analysis: Data Collaboration

4. Assessing the Risk

- What are the risks of this project across the Data Lifecycle?
- What are the potential externalities (including environmental externalities) of this project?

In this Toolkit:

RD4C Opportunity and Risk Diagnostic

RD4C 22 Questions Audit Tool

5. Matching Demand & Supply: Operational Models

- How is the data going to flow between the project partners?
- What does a fit-for-purpose collaborative model look like?

In the Toolkit:

Data Collaboratives Canvas

RD4C Decision Provenance Mapping Tool

6. Matching Demand & Supply: Governance

- What are the 4 Ps of data governance for this project?
- Who is going to govern this project and how?

In the Toolkit:

Data Responsibility Journey

Contractual Wheel of Data Collaboration

7. Matching Demand & Supply: Tech Infrastructure

- What data standards will improve the interoperability of the data?
- How can the data be handled to balance privacy with efficiency?
- Who can access and re-use the data?

In the Toolkit:

Data Tagging Criteria and Exercise

8. Using Data Responsibly

- What are the ethical implications of this project?
- Do you need to establish a social license for this project?
- How can you assess and mitigate the environmental impact of your project?

In the Toolkit:

Data Responsibility Journey

9. Measuring Impact

- How will you capture the impact and success of this project?
- How will you know when to end this project?

In the Toolkit:

Building a Logic Model to Assess Impact

DATA COLLABORATIVES: MATCHING DEMAND & SUPPLY

THEGOVLAB