UNLOCKING DATA: IDENTIFYING NEEDS & COLLABORATIVE APPROACHES

Data Series: Session #1

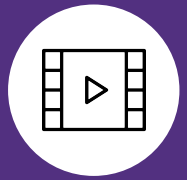
An Introduction to Data & its Re-Use

Stefaan G. Verhulst

Tuesday, October 24, 2023

THEGOVLAB innovate(us)

ZOOM HOUSEKEEPING



The session will be recorded, and the recording will be shared as part of the InnovateUS workshops.

Mute your microphone when not talking and turn on your video (if feasible).

Share any comments or questions in the chat.

ABOUT THE GOVLAB

GOVLAB

The Governance Lab (The GovLab) is an action-oriented research center that seeks to improve people's lives by changing how we govern using new technologies.

Learn more at: govlab.org.

The **Open Data Policy Lab** is a resource hub supporting decision-makers as they work toward accelerating the responsible reuse and sharing of open data for the benefit of society and the equitable spread of economic opportunity.

Learn more at: opendatapolicylab.org.

DATA CAN HELP TO INNOVATE HOW WE SOLVE PUBLIC PROBLEMS

Observation 1: *Datafication has transformed the data landscape.*

DATAFICATION & DIGITALIZATION

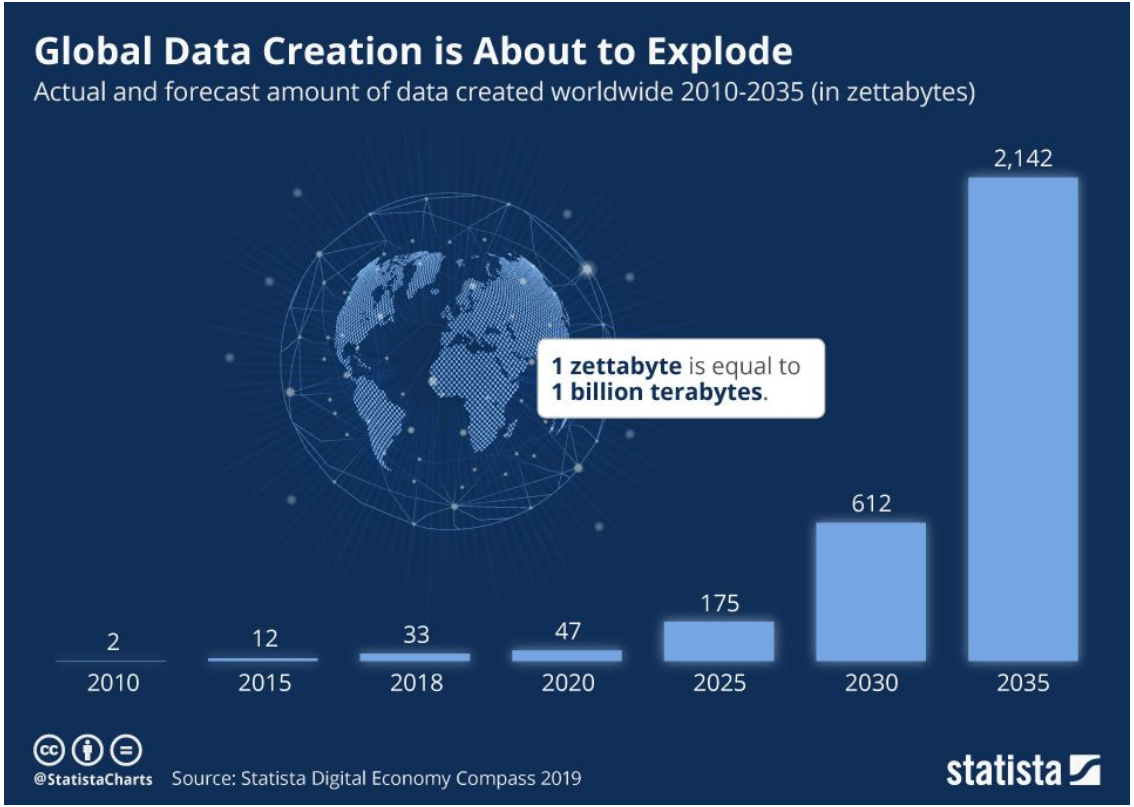
Changes in the way data is **collected**

Changes in the way data is **stored**

Changes in **computation & analytic capacities**

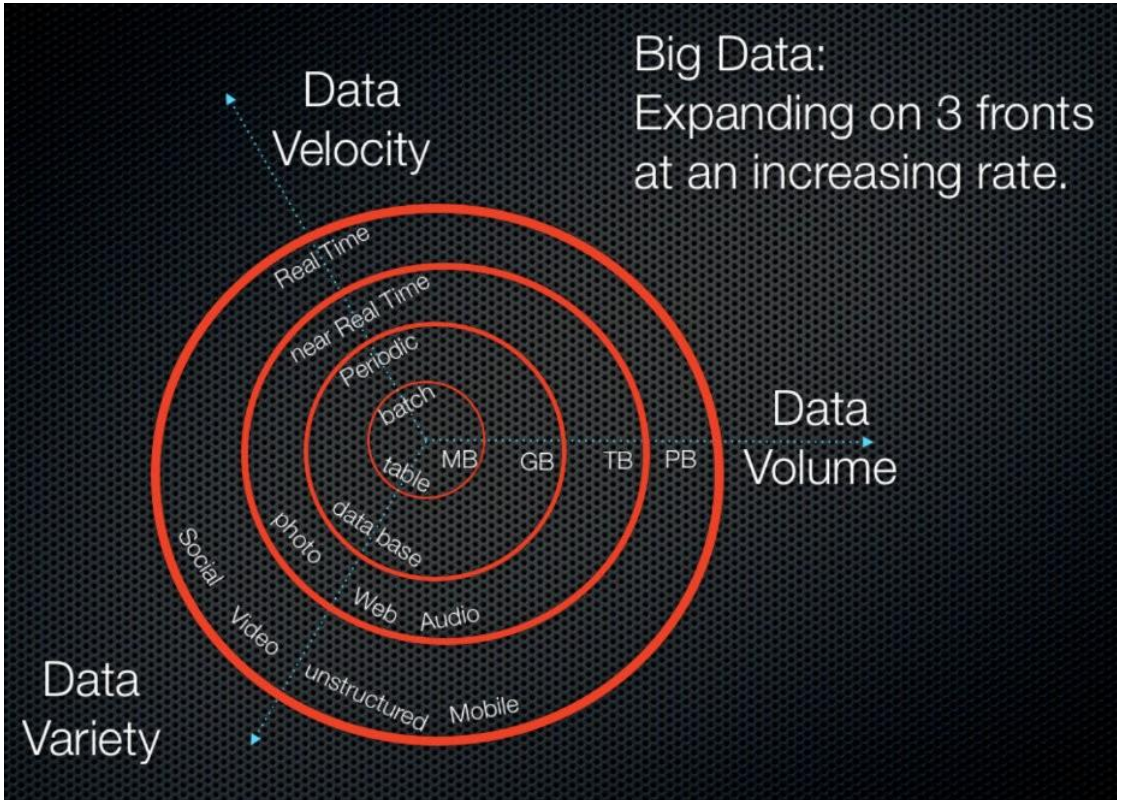
Changes in the **use of & reliance on data**

DATAFICATION & DIGITALIZATION



SOURCE: Martin Armstrong. 2019. "Global Data Creation is About to Explode." *Statista.*

“BIG DATA”



SOURCE: Divya Soubra. 2012. “The 3Vs That Define Big Data.” *Data Science Central*.

INCREASE IN UNSTRUCTURED 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

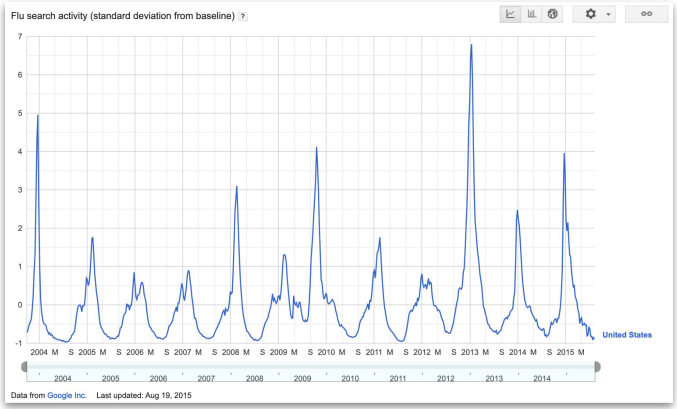
No Fixed Fields
Various formats, sizes and structures

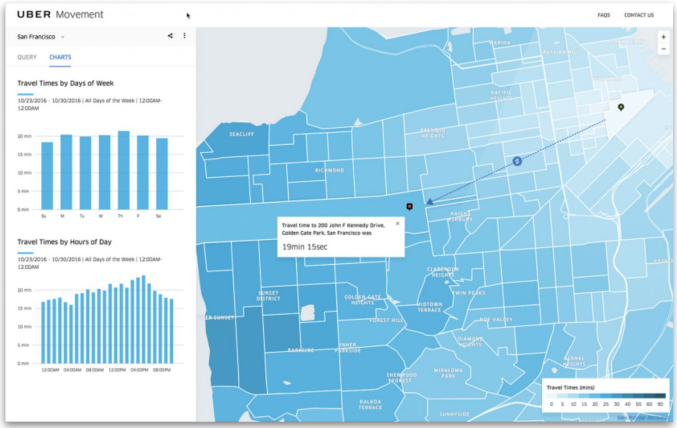
Texts, Audio, Pictures, Social Media

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.

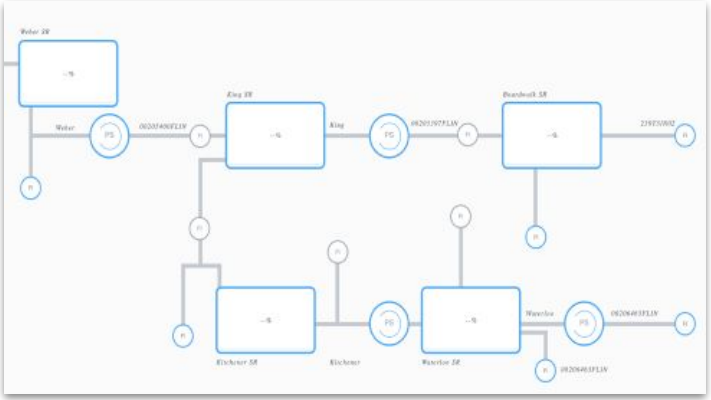
BIG DATA: VARIETY

- Data resulting from **consumption, commercial and financial transactions**
  - The Bureau of Economic Analysis used daily card data (including credit, debit, and gift cards) to measure the reduction in revenue of local businesses around the time of the pandemic.
- Data resulting from **communicating and engaging in social interactions**–for pleasure, study and/or work
  - The Center for Disease Control & Google co-created Google Flu Trends using search queries and relevant terms to estimate influenza levels in over 25 countries.
- Data resulting from having people and products **moving around**
  - Uber shares aggregated and anonymized ride data with city planners via its Uber Movement platform.





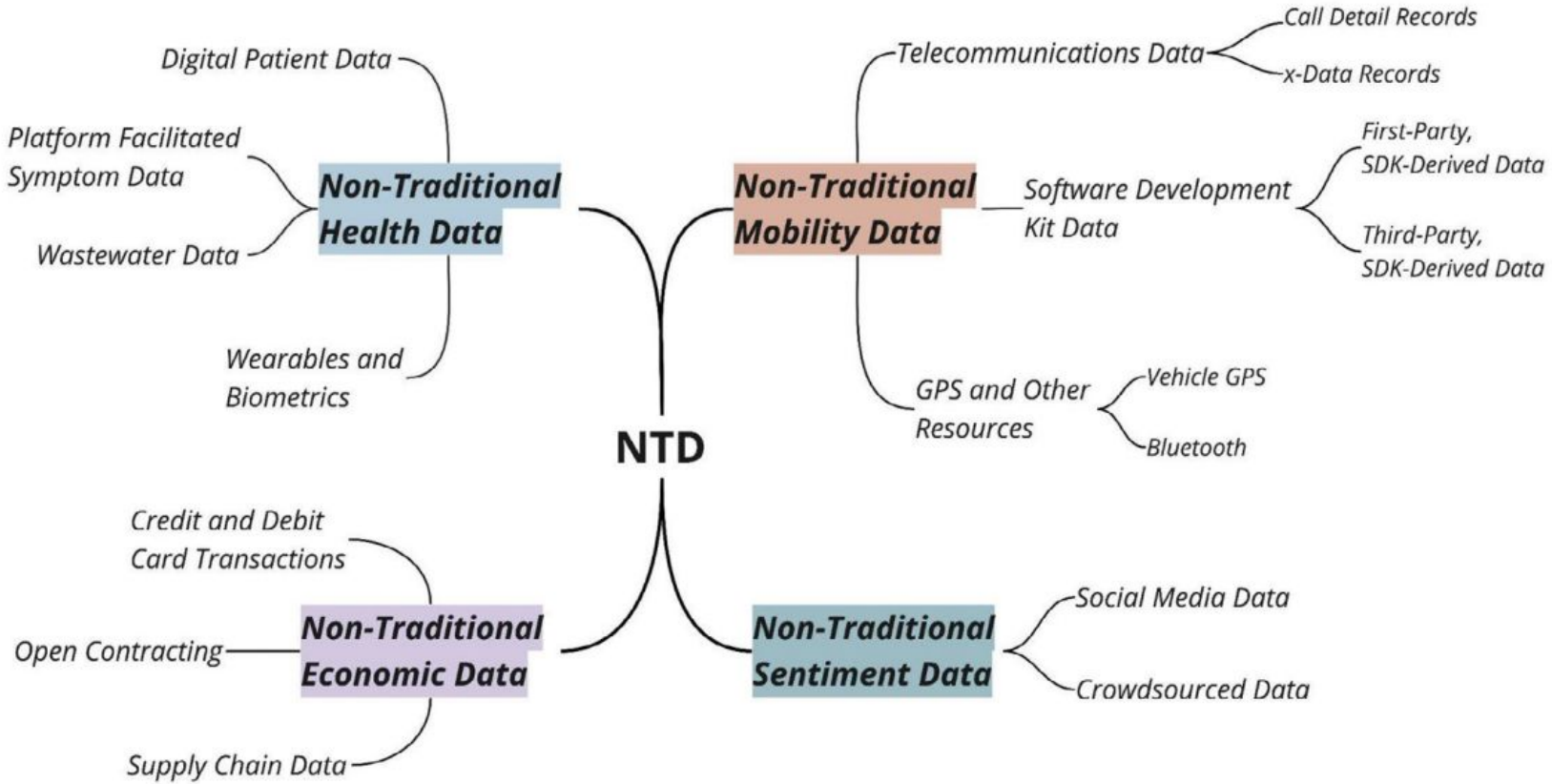
BIG DATA: VARIETY



Innovyze
An AUTODESK company

- Data resulting from **media and entertainment consumption**
  - Researchers at Michigan State University analyzed YouTube comments to understand perceptions of antimarijuana PSA on YouTube.
- Data emerging from **producing products and goods**
  - Cisco's Internet of Everything provides consumers and policymakers with the means to trace a food product back along its entire chain of production.
- Data emerging from **managing infrastructures and natural assets**.
  - Innovyze's Emagin technology collects data from waterways and applies AI to increase reliability of drinking water treatment systems, and optimize water networks and machinery.

NON-TRADITIONAL DATA



IMPLICATIONS

Harnessing Datafication for Good

Utilize new data sources, which can unveil actionable insights & optimize decision-making processes in real-time.

Data Innovation Labs

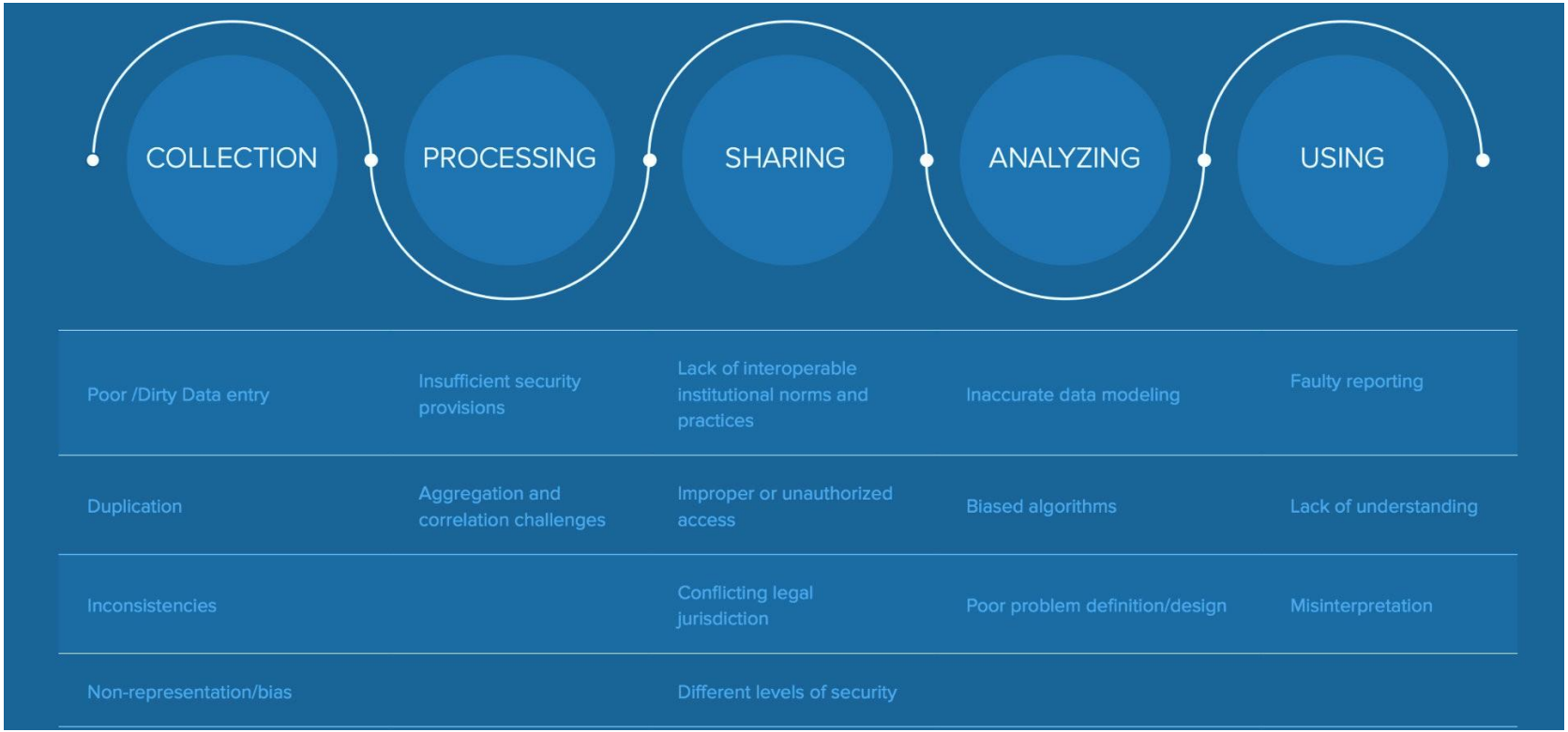
Establish data innovation labs within government departments to explore new data analysis techniques, tools, & collaborative models. These labs can serve as a breeding ground for innovative data-driven solutions to state-level challenges.

Public-Private Tech Innovation Partnerships

Form partnerships with tech companies, startups, & academia to drive technological & data analytics innovation. These collaborations can bring in fresh perspectives & expertise, accelerating the development of innovative data solutions.

Observation 2:
Data is not a thing, it is a process.

DATA AS A PROCESS: THE DATA LIFECYCLE



IMPLICATIONS

Requires Collaboration Across Stages

- Effective collaboration among departments and agencies is needed at and across the stages.

End-to-End Resource Allocation

- Efficient allocation of resources to ensure each stage is adequately supported.

Data Quality and Integrity

- Ensuring data quality and integrity throughout the lifecycle for accurate insights.

Risk Management and Compliance

- Identification and mitigation of risks at each stage to prevent data breaches, loss, or misuse.
- Adherence to legal, ethical, and procedural guidelines at every stage.

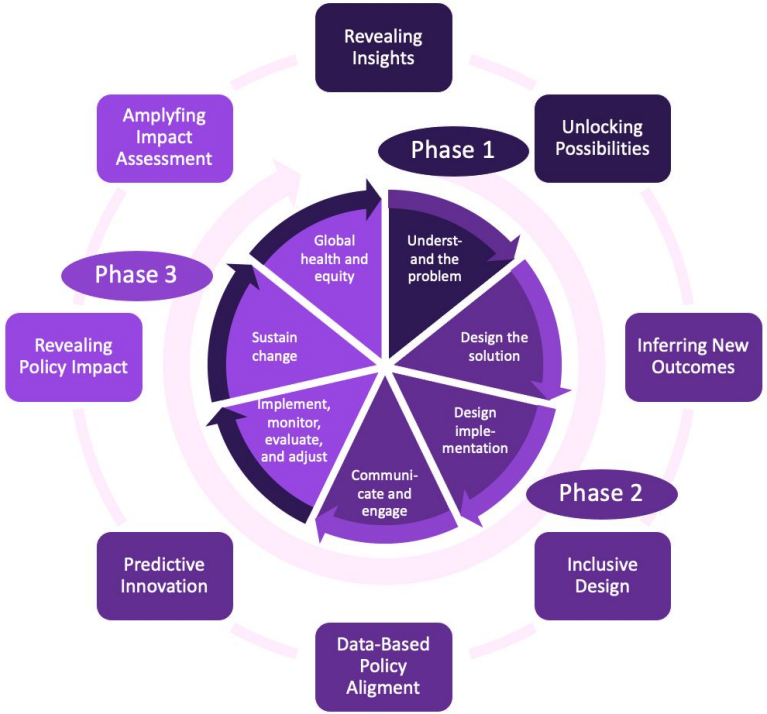
Functions and Training

- Dominant focus on analysis/data science ignores other important roles such as data stewardship

Observation 3:
Data in itself has no value.

CHALLENGE:
FROM DATA TO DECISION INTELLIGENCE

CHALLENGE:
STAGES OF THE DECISION CYCLE



IMPLICATION

1: (First Mile Considerations) Understand Decision-Makers' Needs

Clarity on Questions and Objectives:

- Align data collection and analysis with decision-makers' questions, intentions and goals.

Anticipate Needs at Different Stages:

- Align insights with different needs according to the stage of decision making

2: (Last Mile Considerations) Enhancing Decision-Making Process

Actionable Intelligence:

- Provide decision-ready insights and possible recommendations.

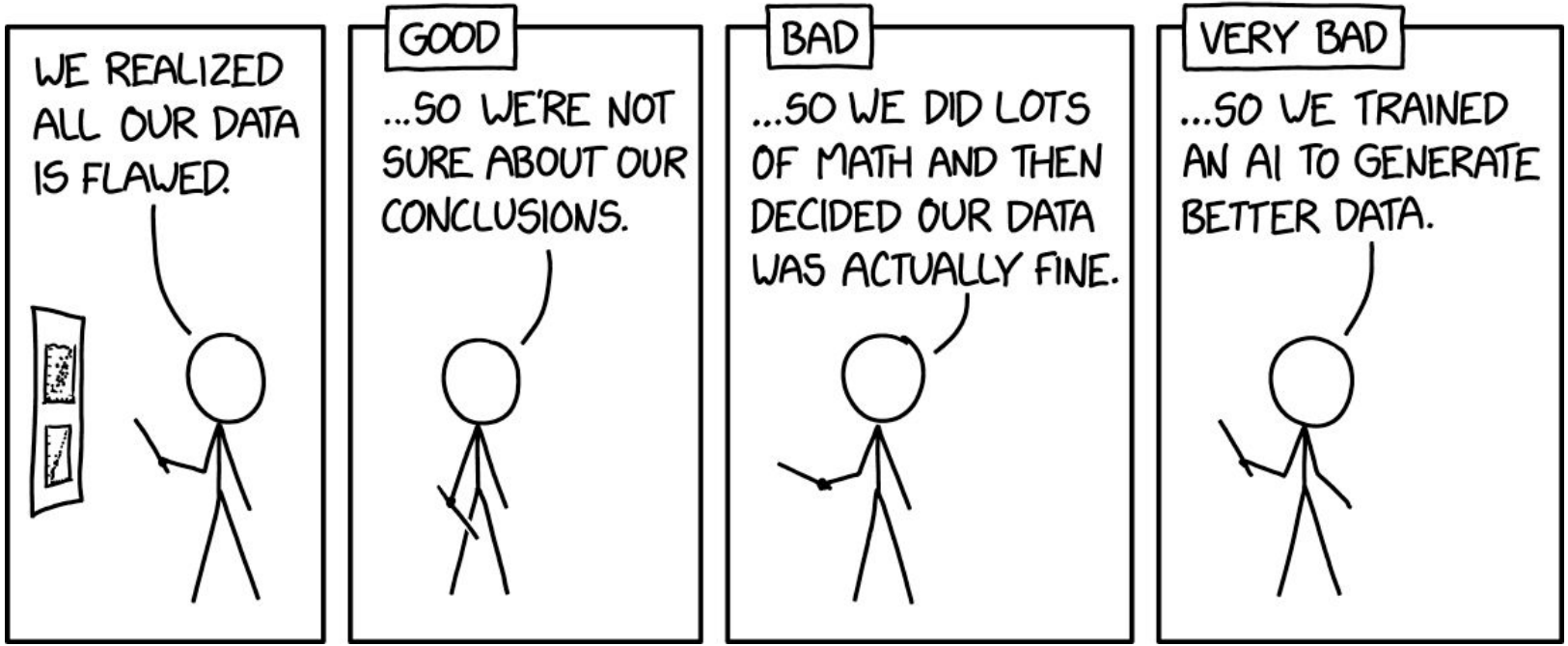
Innovate in how decisions are made

- Consider new approaches like Decision Accelerator Labs

Observation 4

Data is never “raw” or “born”—it is created.

DATA CREATION & CONTEXT



SOURCE: xkcd.com. "Flawed Data." https://xkcd.com/2494/.

DATA CREATION & CONTEXT

Data Construction:

- Created through human-designed processes
- Context gives data its meaning.

Bias Introduction:

- Arises from data collection methods, question framing, and respondent demographics.
- Influenced by the choice of metrics and measurement techniques.

Selective Collection:

- Choices on what data to collect/omit are driven by practical, political, or theoretical factors.
- Represents a selective view of reality.

DATA PROCESSING & INTERPRETATION

Instrumentation & Measurement:

- Instruments/methodologies embody assumptions and limitations.
- Shape the data collected and its accuracy.

Data Cleaning and Transformation:

- Processes to handle errors, missing values, or outliers can modify original data.
- May introduce additional biases or assumptions.

Representation and Visualization:

- Choice of representation can highlight or obscure data insights.
- Influences human interpretation and decision-making.

Human Interpretation:

- Data is ultimately interpreted by humans with their own biases and objectives.
- Emphasizes the importance of ethical considerations in data science.

BIAS IN DATA & ALGORITHMS

RESEARCH

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

EMORY UNIVERSITY

Emory News Center

AI systems can detect patient race, creating new opportunities to perpetuate health disparities

STAT+

A STAT INVESTIGATION

Could AI tools for breast cancer worsen disparities? Patchy public data in FDA filings fuel concern

By Casey Ross Feb. 11, 2021

The Verge / Tech / Reviews / Science / Entertainment / More +

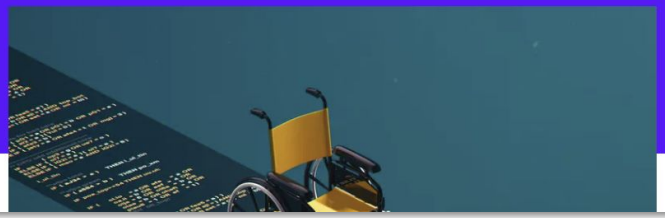
SCIENCE

What happens when an algorithm cuts your health care

By Colin Lecher

Illustrations by William Joel; Photography by Amelia Holowaty Krales

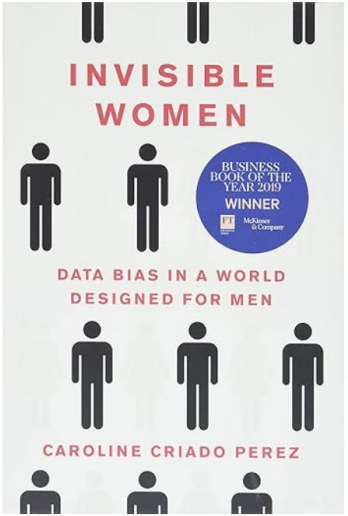
Mar 21, 2018, 9:00 AM EDT | 0 Comments / 0 New

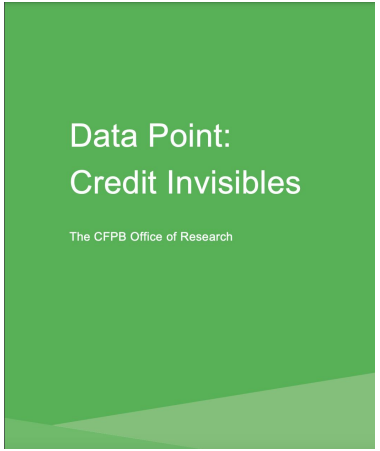


DATA INVISIBLES

Data Invisibles: Individuals who are excluded from data sets, re-inforcing existing social exclusions or discrimination.

Many invisibles located in developing countries, are non-Western citizens, or come from minority or disadvantaged groups that might already be excluded.





IMPLICATIONS

Acknowledge and Account for Bias

- Biases throughout the data process pose a risk to data-driven decision making. Identifying.
- Mitigating bias helps improve the outcomes of data-informed interventions.

The Role of Different Data Actors

- Each stage of the data process calls for a different set of skills and expertise, related not only to data analytics but also to strategy, communication and more.

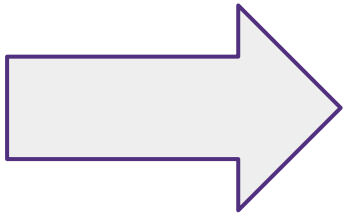
Collaboration and Diversification

- Collaborating with diverse individuals and organizations as suppliers of data and expertise can help address bias.

Observation 5
Data re-use provides the real opportunity.

KEY CHARACTERISTICS OF DIGITAL DATA

- Digital data is **intangible and easy (and often cost-free) to copy and replicate**.
- It is a **non-rivalrous good**.
- The value of any particular item of digital data often **expands when made accessible and combined with other data**.

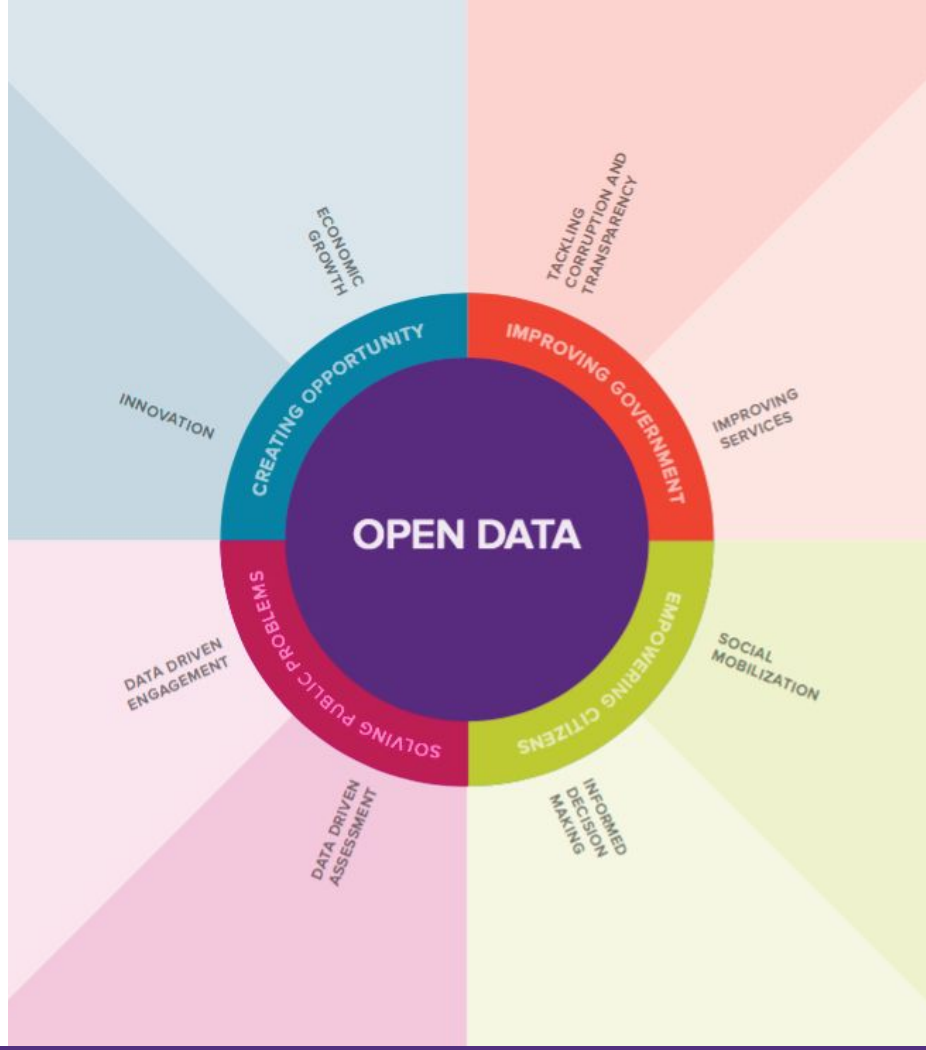


THE POTENTIAL OF DATA RE-USE

THE CASE FOR DATA RE-USE

Data re-use can:

1. Improve decision making by fostering greater accountability, transparency, efficiency and sustainability.
2. Empower stakeholders to transform their ecosystems, demand change and enable new forms of social mobilization.
3. Create new economic opportunities bringing greater prosperity.
4. Facilitate scientific discovery and innovation for more effective outcomes.



THE POTENTIAL OF DATA RE-USE: ADVANCING PUBLIC SERVICE PROVISION

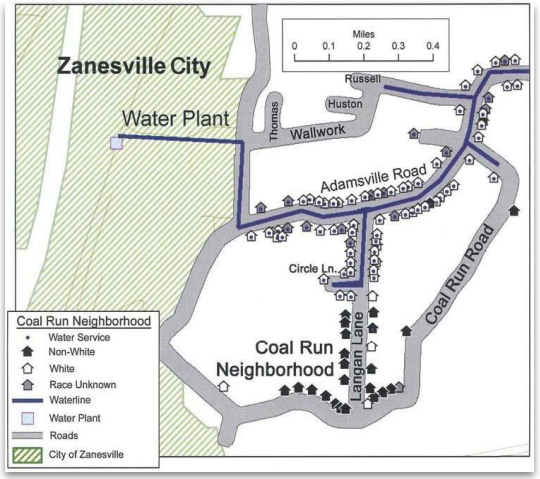
- Open data is improving government, primarily by enhancing public services and resource allocation, and tackling corruption and increasing transparency.
- **Example:** Canada's T3010 Charity Information Return Data
    - Charitable data from the Canada Revenue Agency's information return has been available to the public since 1975. In 2013, all data sets since 2000 were transcribed and made available online via the government's data portal under a commercial open data license.
    - The resulting data set has been used to explore the state of the nonprofit sector, improve advocacy by creating a common understanding between regulators and charities, and create intelligence products for donors, fundraisers and grant-makers.

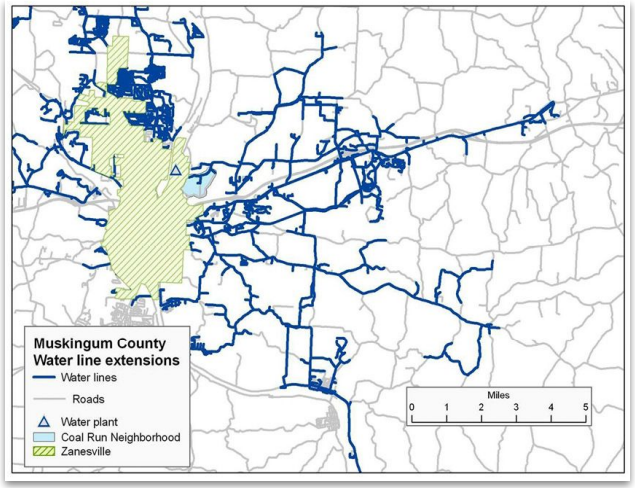
Canada

canada revenue agency

THE POTENTIAL OF DATA RE-USE: EMPOWERING INDIVIDUALS

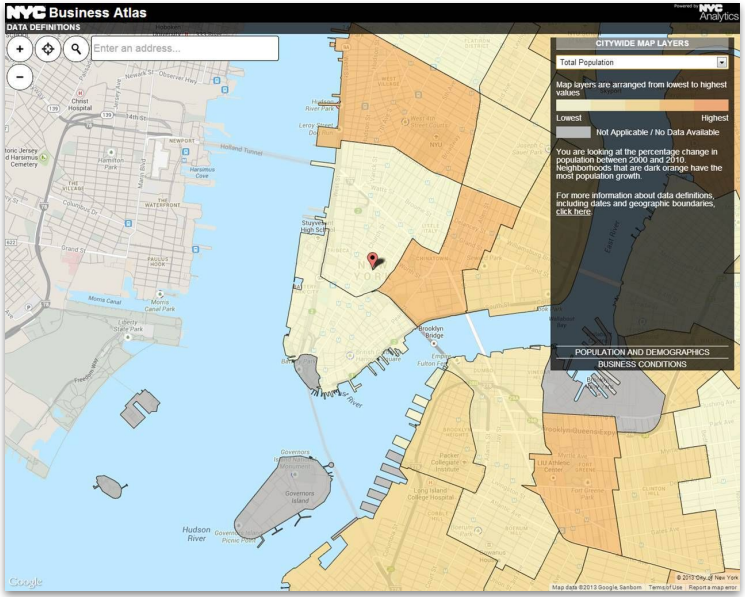
- Open data is empowering citizens to take control of their lives and demand change by enabling more informed decision-making and new forms of social mobilization, both in turn facilitated by new ways of communicating and accessing information.
- **Example:** Kennedy vs. the City of Zanesville
  - One of the key pieces of evidence used during the case was a map derived from open data from the water company displaying houses connected to the water line and data showing town demographics.
  - The insights from the map showed significant correlation between the houses occupied by the white residents of Zanesville and the houses hooked up to the city water line, and the case went in favor of the African-American plaintiffs, awarding them a $10.9 million settlement.





THE POTENTIAL OF DATA RE-USE: SUPPORTING BUSINESS & SMEs

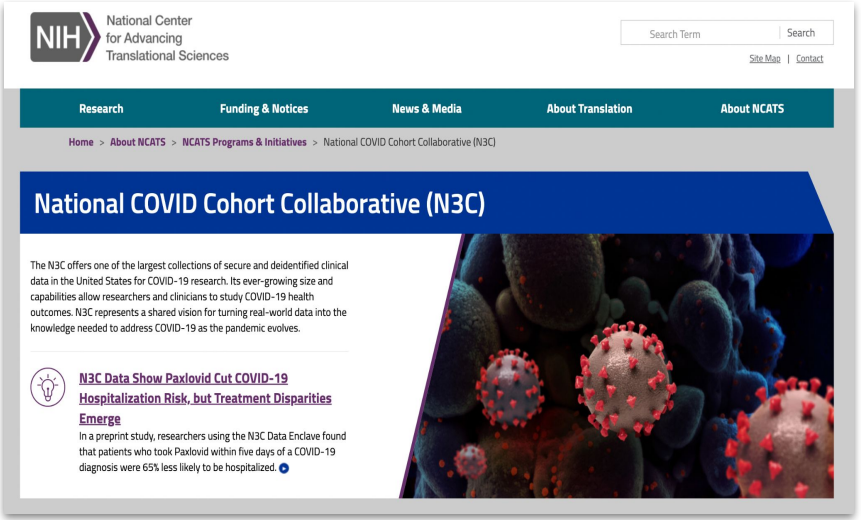
- Open data is used primarily to serve the Business-to-Business (B2B) markets, followed by the Business-to-Consumer (B2C) markets. A number of the companies studied serve two or three market segments simultaneously.
- Open data is usually a free resource, but SMEs are monetizing their open-data-driven services to build viable businesses.
- **Example:** MODA's New York City Business Atlas



NYC Analytics

THE POTENTIAL OF DATA RE-USE: SOLVING PUBLIC PROBLEMS

- Open data is playing an increasingly important role in solving big public problems, primarily by allowing citizens and policymakers access to new forms of data-driven assessment of the problems at hand. It also enables data-driven engagement producing more targeted interventions and enhanced collaboration.
- **Example:** National COVID Cohort Collaborative Data Enclave
    - The effort collects and harmonizes electronic health records into a common data model and provides a way for select partners to easily access this clinical data to inform their research on COVID-19 interventions and long-term care.



IMPLICATIONS

Publishing with Purpose

- Open data and data re-use can create substantial value provided it is done with intent. Opening up datasets in isolation, with little thought on how it can be used, is unlikely to yield substantial value.

Embrace the FAIR principles

- Findable
- Accessible
- Interoperable
- Re-usable

Social License

- Social license refers to the public acceptance of business practices or operating procedures used by a specific organization or industry. Seeking social license can proactively address concerns over misuse, privacy violations, and surveillance.

Observation 6
Data likely resides elsewhere.

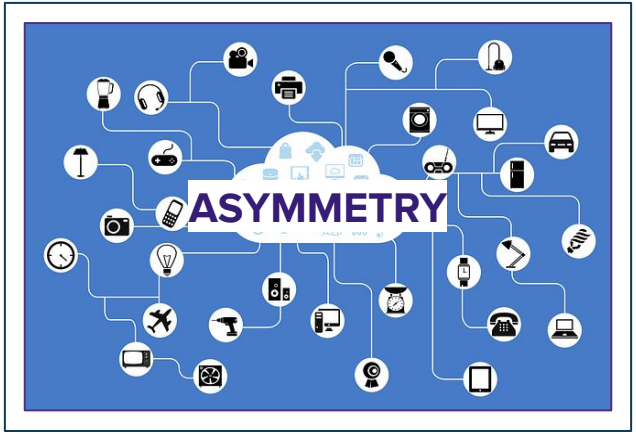
THE EMERGENCE OF DATA ASYMMETRIES

Volume

MIS-USE

Veracity

DEMAND



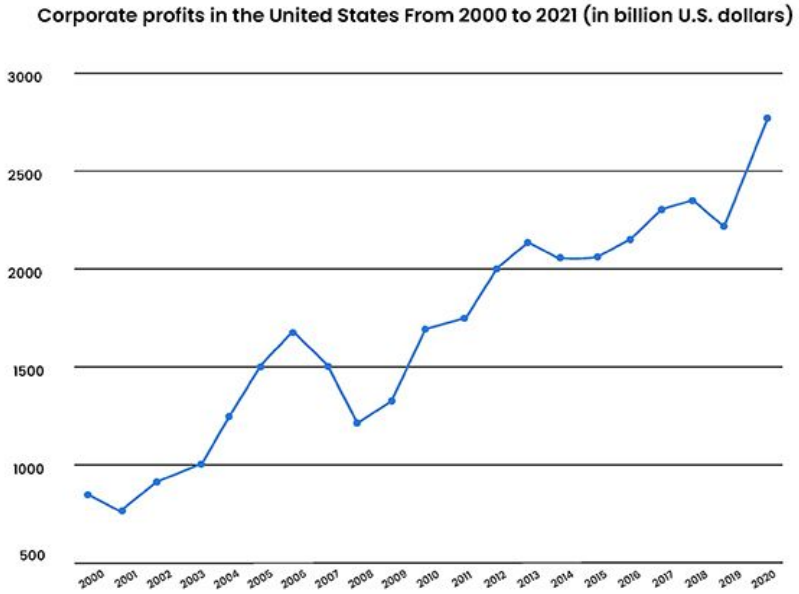
SUPPLY

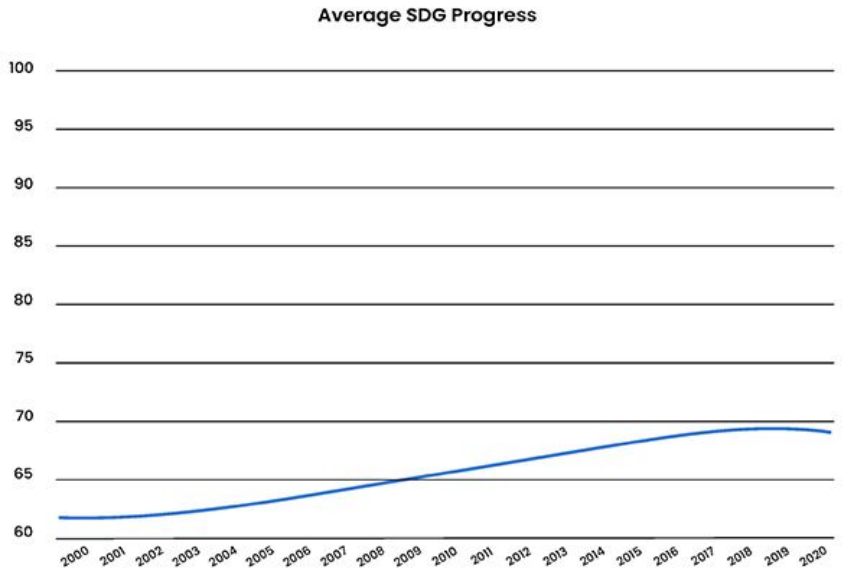
Velocity

RE-USE

Variety

DATA FOR SOCIAL IMPACT CASE STUDY





SOURCE: Jason Saul and Kriss Deiglmeier. 2023. "Unlocking the Power of Data Refineries for Social Impact." *Stanford Social Impact Review.*

DATA & POWER ASYMMETRIES

Sir Francis Bacon

“Knowledge itself is power.”

George A. Akerlof

“The Market for Lemons”

AI & SOCIETY
https://doi.org/10.1007/s00146-022-01410-5

CURMUDGEON CORNER

The ethical imperative to identify and address data and intelligence asymmetries

Stefaan G. Verhulst[1]

Received: 16 August 2021 / Accepted: 9 February 2022

Francis Bacon is generally known as the man who developed the scientific method[1] in the sixteenth century, but he is also credited with coining the aphorism that “knowledge is power” (Meditationes Sacrae, 1597). As economies and society in general have grown more complex, the truth of this aphorism has only increased. In 1970, George Akerlof published “The Market for Lemons,”[2] his seminal paper on information asymmetries and how access to information can define markets and establish winners and losers. Although Akerlof focused on the used car market, his insights applied across sectors and industries. In 2001, Akerlof, Michael Spence and Joseph Stiglitz were awarded[3] the Nobel Prize in Economics “for their analyses of markets with asymmetric information.”

The insight that knowledge, resulting from having access

issue. Just what are data asymmetries? How do they emerge, and what form do they take? And how do data asymmetries accelerate information and other asymmetries? What forces and power structures perpetuate or deepen these asymmetries, and vice versa? I argue that it is a mistake to treat this problem as homogenous. In what follows, I suggest the beginning of a taxonomy of asymmetries. Although closely related, each one emerges from a different set of contingencies, and each is likely to require different policy remedies. The focus of this short essay is to start outlining these different types of asymmetries. Further research could deepen and expand the proposed taxonomy as well help define solutions that are contextually appropriate and fit for purpose.

TYPES OF DATA ASYMMETRIES

BUSINESS-TO-CONSUMER (B2C)

Increasingly common with the datafication of consumption patterns.

BUSINESS-TO-GOVERNMENT (B2G)

Hampers the ability of government to develop data-driven policies or target services.

BUSINESS-TO-BUSINESS (B2B)

Spurred by the emergence of large data monopolies that dominate their sectors and the broader economy.

GOVERNMENT-TO-SOCIETY (G2S)

Results when data collected by the government is hoarded without transparency or making it accessible to society writ large.

DATA COLLABORATIVES: MATCHING DEMAND & SUPPLY

"The term data collaborative refers to a new form of collaboration, beyond the public–private partnership model, in which participants from different sectors—in particular companies—exchange their data to create public value."

— *Stefaan Verhulst, "Data Collaboratives: Exchanging Data to Improve People's Lives"*

IMPLICATIONS

Data Stewards

- Invest in the human infrastructure to make data collaboration more systematic, sustainable and responsible

Explore # Data Collaborative Models

- Different Data collaborative Models are fit for purpose in how they bridge gaps between businesses, government, and society,

Reducing Transaction Costs

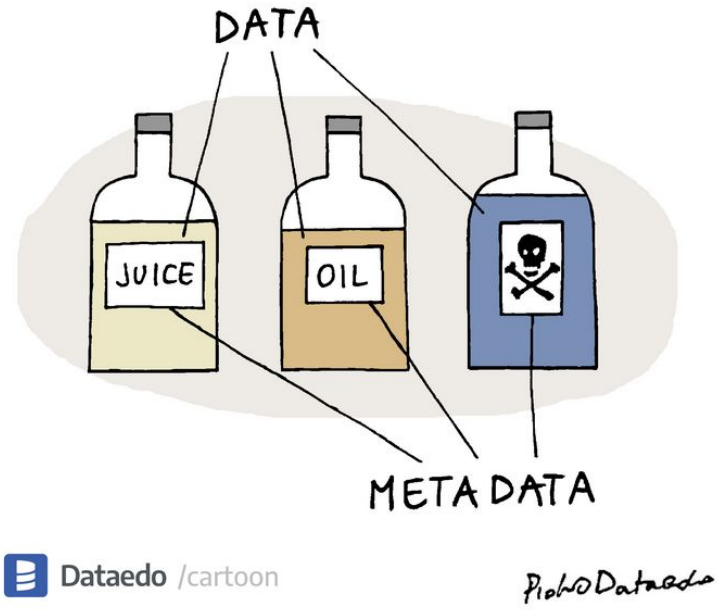
- Coordinating with partners can be expensive and time consuming. By developing agreements and frameworks for collaboration, organizations can reduce the costs associated with data re-use.

Observation 7

Metadata is the new data.

THE GROWING IMPORTANCE OF METADATA

- Metadata can be considered “data about the data.”
- It answers the question of the who, what, when, where, why and how of a dataset.
- This includes critical information about a dataset including the format, quality, provenance and origins of the data.
- Metadata helps improve the discoverability and interoperability of data.



SOURCE: Piotr Kononow. 2022. “Data vs Metadata #4.” *Dataedo*.

DATA TAGGING

Draft 1.0 - Data Tagging Criteria and Exercise, The GovLab September 4, 2020

Data Life Cycle Questions	Release Risk Factors	High/Low-Risk Tagging	Open vs Closed? (Spectrum of Conditionality)
How was the data acquired or collected?	Was Consent Obtained (for Reuse)? Data Lineage? Obtained Ethical Clearance through Review Board?		
What data rights are associated with the data?	Licensing Regime? Ownership Expectations? Chain of Trust?		
Are there laws and regulations that need to be complied with?	Regulatory compliance? Cross-jurisdictional considerations?		
How is the data currently stored and processed?	Security? Ease of Access? Auditability?		

Sharing Sensitive Data with Confidence: The Datatags System

Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai

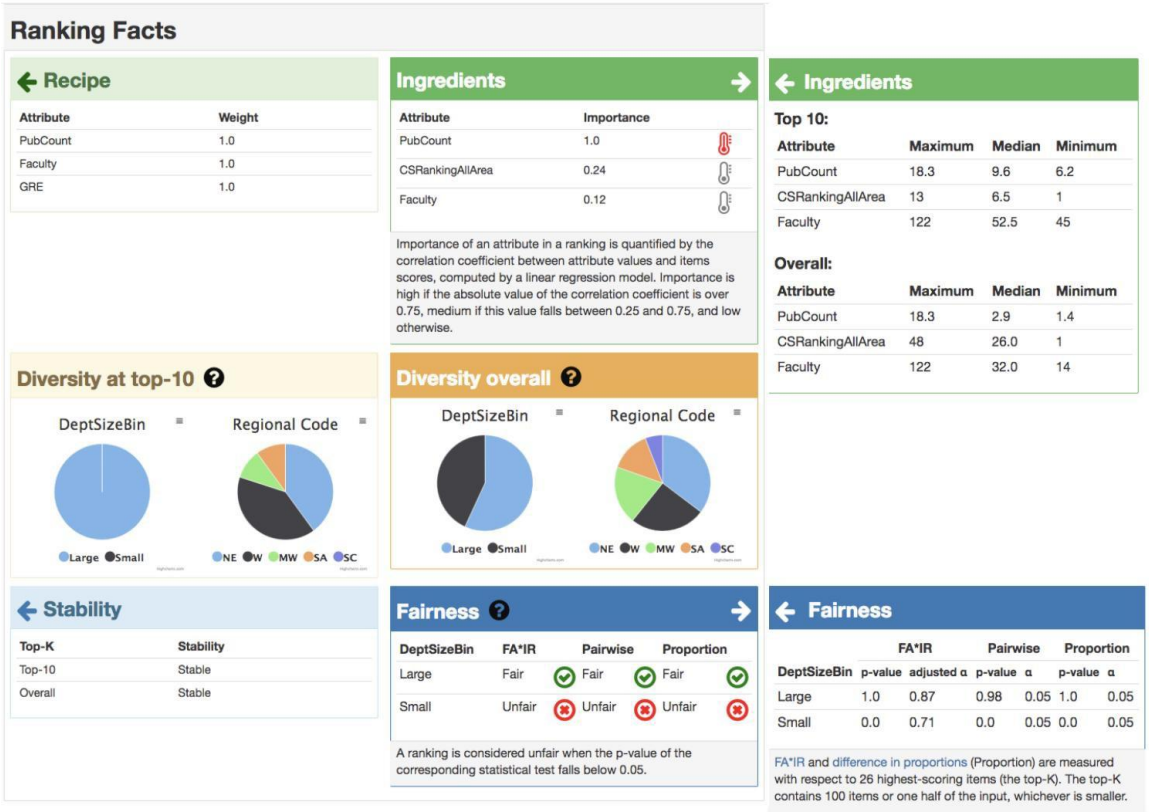
Highlights

- We introduce datatags as a means of specifying security and access requirements for sensitive data.
- The datatags approach reduces the complexity of thousands of data-sharing regulations to a small number of tags.
- We show implementation details for medical and educational data and for research and corporate repositories.

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Definitions for each of six ordered Blue to Crimson sample datatags.

NUTRITIONAL LABELS



Ranking Facts

Recipe

Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Ingredients

Attribute	Importance
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Importance of an attribute in a ranking is quantified by the correlation coefficient between attribute values and items scores, computed by a linear regression model. Importance is high if the absolute value of the correlation coefficient is over 0.75, medium if this value falls between 0.25 and 0.75, and low otherwise.

Ingredients

Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
CSRankingAllArea	13	6.5	1
Faculty	122	52.5	45

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
CSRankingAllArea	48	26.0	1
Faculty	122	32.0	14

Diversity at top-10

Diversity overall

Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

A ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05.

Fairness

	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	α
Large	1.0	0.87	0.98	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

FA*IR and difference in proportions (Proportion) are measured with respect to 26 highest-scoring items (the top-K). The top-K contains 100 items or one half of the input, whichever is smaller.

SUMMARY

1. Datafication has transformed the data landscape.
2. Data is not a thing, it is a process.
3. Data in itself has no value.
4. Data is never “raw” or “born” – it is created.
5. Data re-use provides the real opportunity.
6. Data likely resides elsewhere.
7. Metadata is the new data.

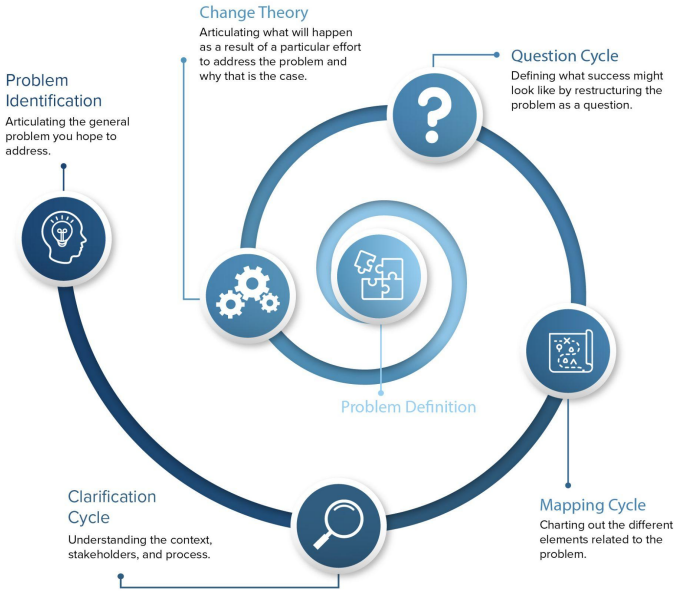
UNLOCKING DATA: IDENTIFYING NEEDS & COLLABORATIVE APPROACHES

Session #2: The Science of Data Questions & Exploring Data Sources

- The Science of Questions
- Minimum Viable Data
- Exploring Data Sources

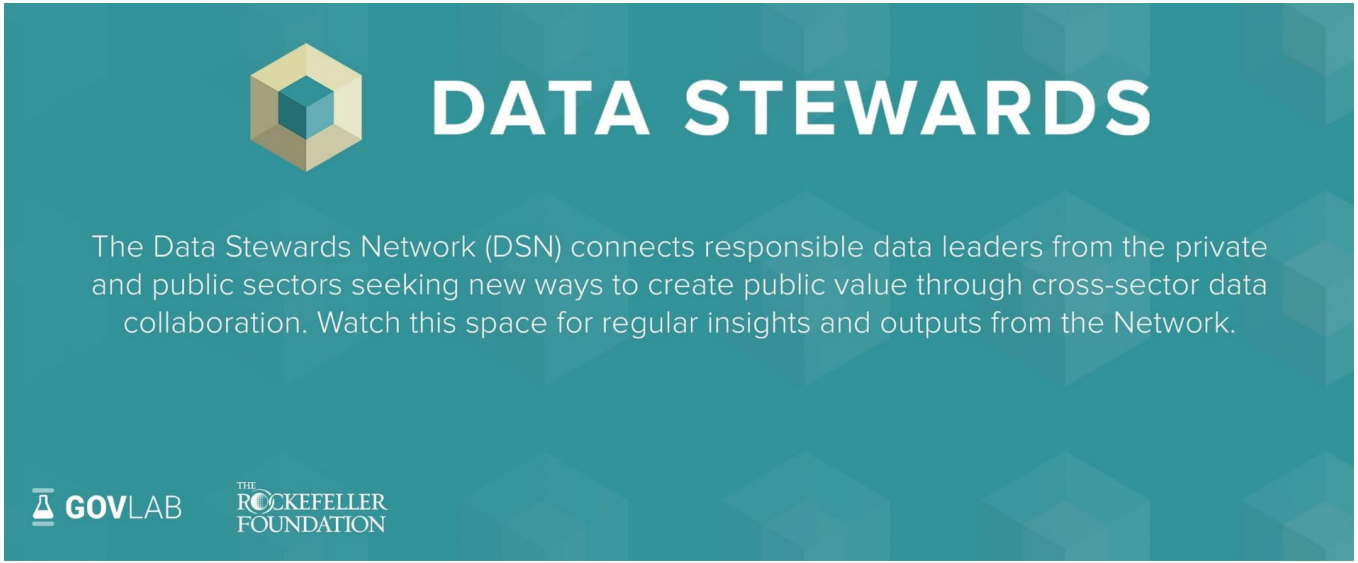
Session #3: Data Collaboration & Governance

- Data Collaboratives
- Governance and Data Sharing Agreements
- Technical Infrastructure for Data Collaboration



THE PROBLEM DEFINITION TOOL

STAY IN TOUCH & RECEIVE UPDATES



Data Stewards Network

https://medium.com/data-stewards-network

@TheGovLab

datastewards@thegovlab.org

@The Governance Lab

www.thegovlab.org