

UNLOCKING DATA: IDENTIFYING NEEDS & COLLABORATIVE APPROACHES

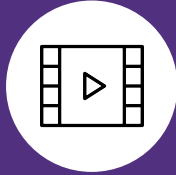
Data Series: Session #2

The Science of Data Questions & Exploring Data Sources

Stefaan G. Verhulst

Wednesday, November 1, 2023

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



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



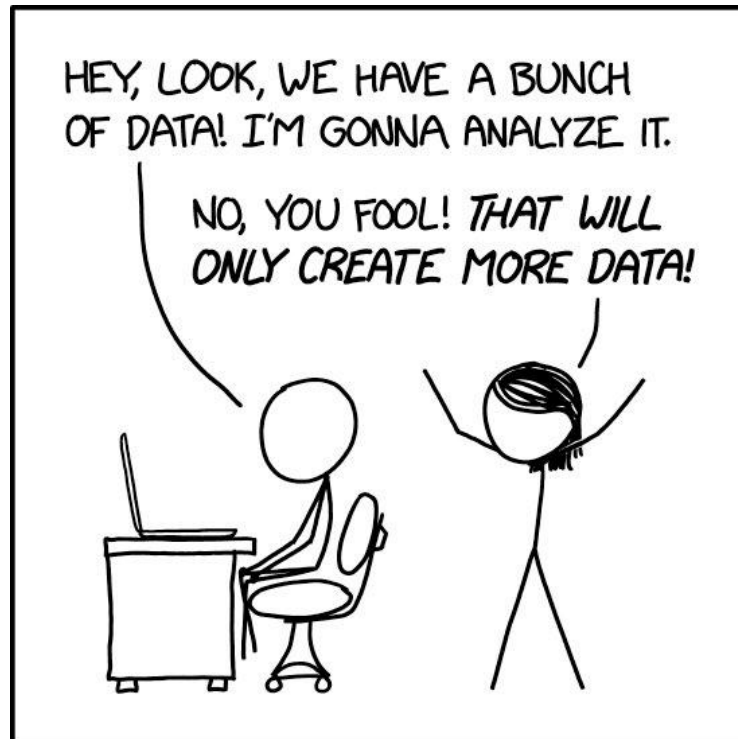


DATA & DATAFICATION: A RECAP

- 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.

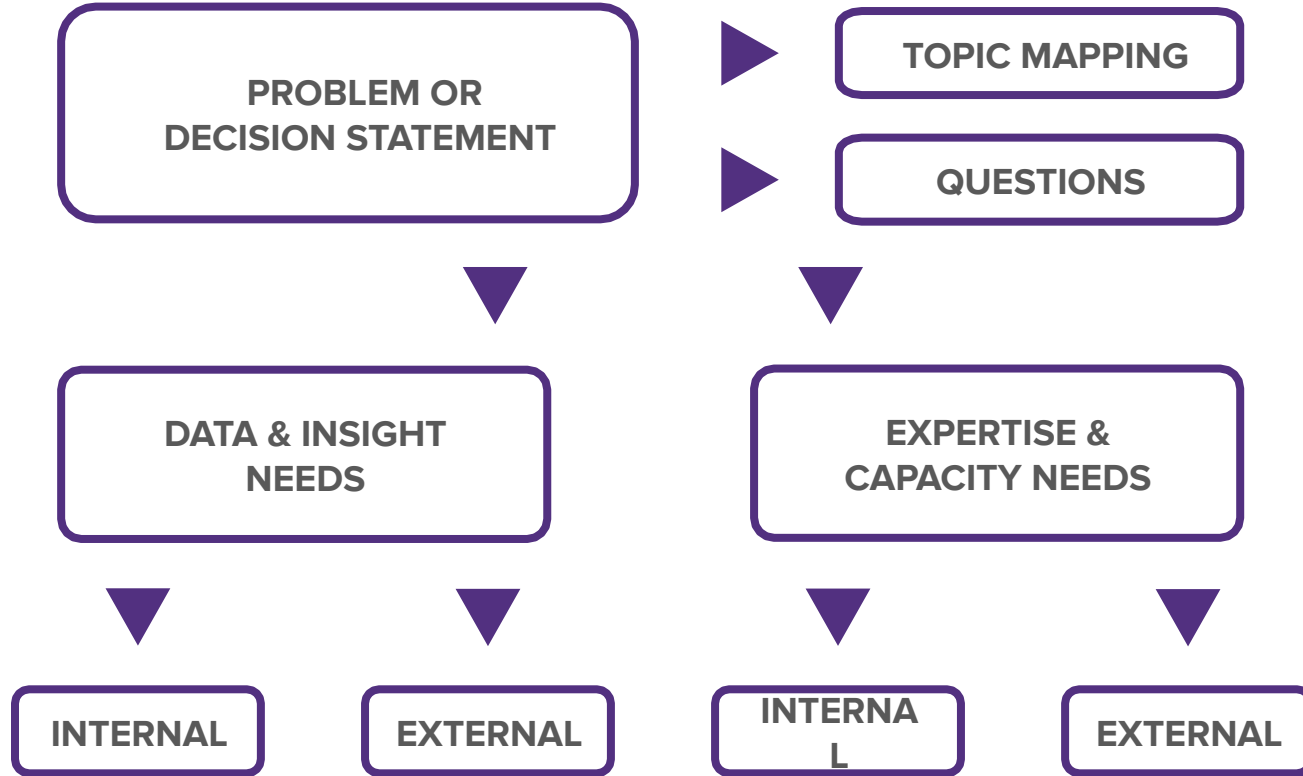


TODAY'S CHALLENGE





GOING FROM PROBLEM TO INSIGHT



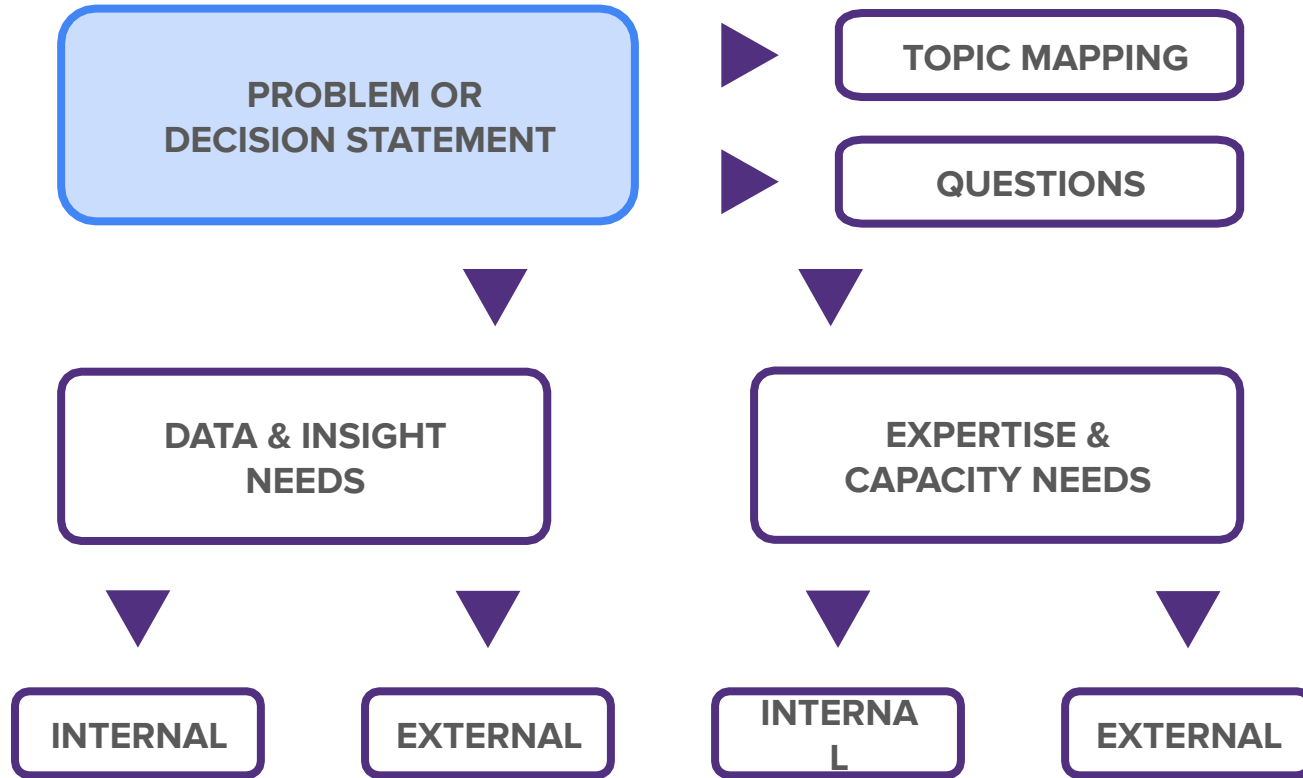


Premise 1:

*Start with the problem,
not the data.*



GOING FROM PROBLEM TO INSIGHT





DEFINE THE PROBLEM WELL

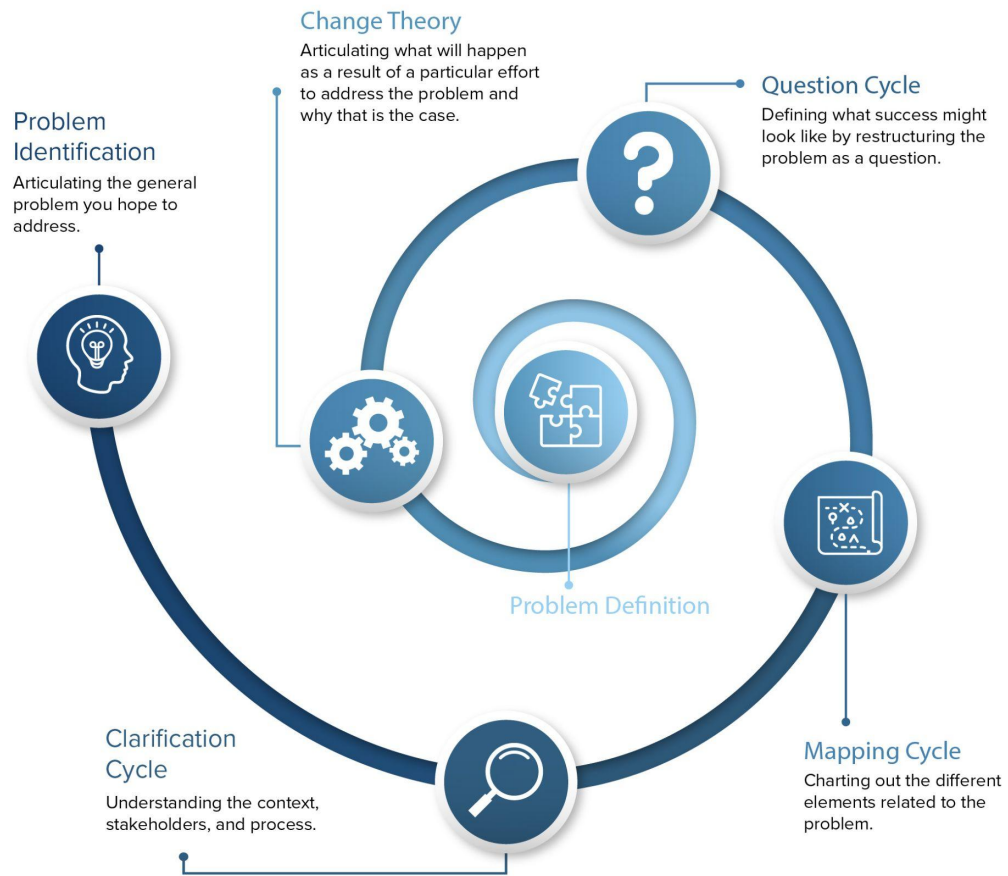
DO NOT:

- Develop solutions in search of a problem...
- Focus on symptoms...not the problem

DO:

- Deconstruct: “Why is this problem taking place?”
- Reframe: “How might others look at it”





THE PROBLEM DEFINITION TOOL



THE PROBLEM DEFINITION TOOL



PROBLEM IDENTIFICATION

Articulating the general problem they hope to address;



CLARIFICATION CYCLE

Understanding the context, stakeholders, and process;



MAPPING CYCLE

Charting out the different elements related to the problem;



QUESTION CYCLE

Defining what success might look like by restructuring the problem as a question;



CHANGE THEORY

Articulating what will happen as a result of a particular effort to address the problem and why that is the case.



IMPLICATIONS

Each problem operates within a unique and dynamic environment.

- The importance of lived experience (thick data, complementing big data)
- Be adaptive as the context evolves
- Be aware of untested assumptions



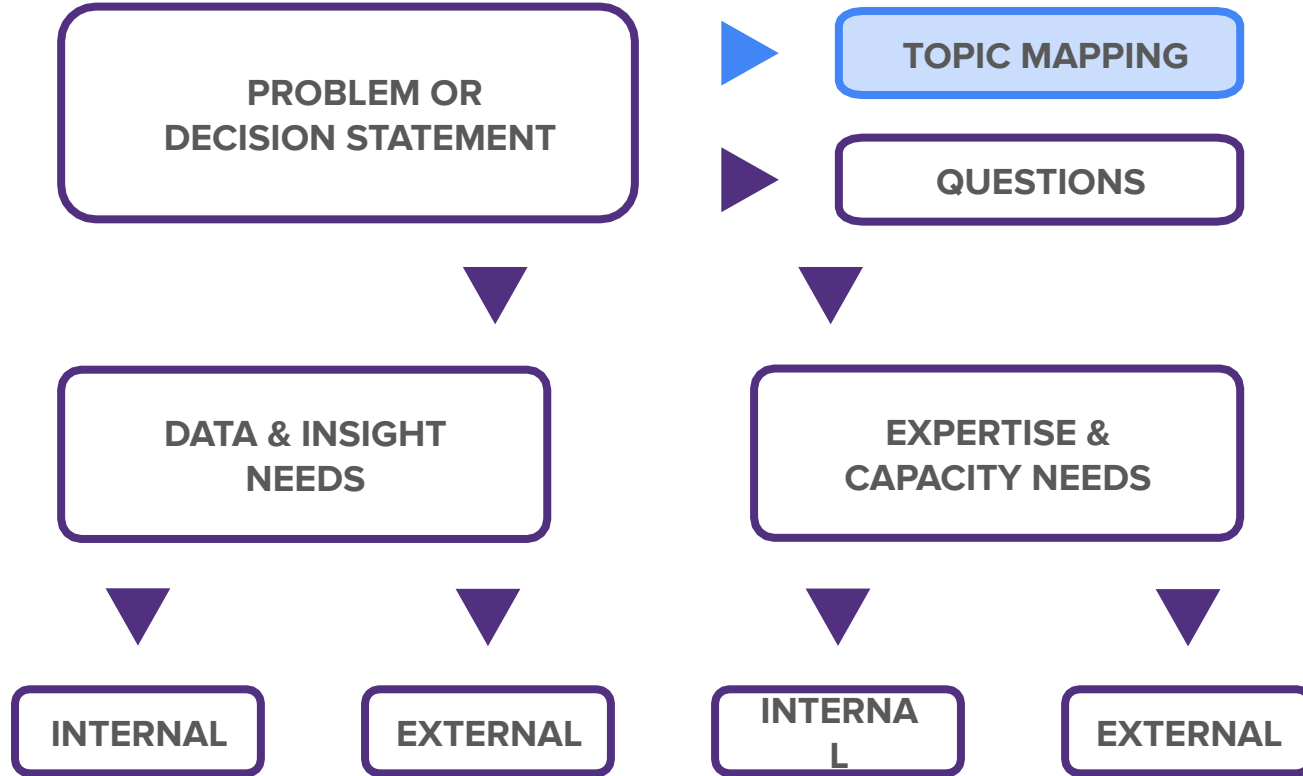


Premise 2:

*A participatory process
can enable a new kind
of “question science”.*



GOING FROM PROBLEM TO INSIGHT





THE 100 QUESTIONS INITIATIVE





ELEMENTS OF THE 100 QUESTIONS INITIATIVE

Developing a Topic Mapping



Identifying & Engaging with Bilinguals

BILINGUAL COMMUNITY

"Bilinguals" are practitioners who possess both relevant domain knowledge and data science expertise. They are experts who understand the importance of data, are aware of the ways data can inform decision-making, and can provide new actionable insights to educate the community on where and how to leverage data responsibly.

Using a Taxonomy of Data-Actionable Questions



Clustering Questions

WORKING DOCUMENT: TO BE FINALIZED LIST OF CLUSTERED QUESTIONS

Food Systems Sustainability Domain: Master Document of Sourced Questions - Grouped by Sub-topic

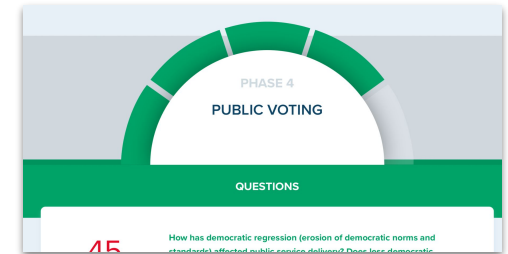
Prioritizing Big, Data-Actionable Questions

3+ [FOOD PRODUCTION AND PROCESSING]
Please select your top 10 questions across the 5 subgroups.

Choose as many as you like

<input type="checkbox"/>	A [Production Tools] How are the existing datasets helping agriculture players? Where are the gaps and limitations in harmonizing indicators for sustainability in food systems?
<input type="checkbox"/>	B [Production Tools] What is the main reason why modern biotechnology is not widely adopted?
<input type="checkbox"/>	C [Consumption Patterns] What are the health risks and benefits of the currently prevailing high-calorie diets and what are the health, socio-

Public Voting





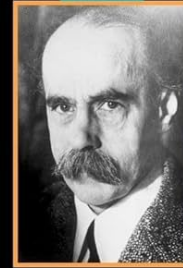
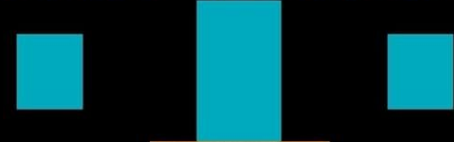
TOPIC MAPPING

Builds a “gestalt” of a problem field

- **Gestalt:** “Seeing things as a whole”



D. Brett King
Michael Wertheimer



Max Wertheimer
& Gestalt Theory



TOPIC MAPPING



DEVELOP “ACTOR MAPS”



ENVIRONMENTAL SCANNING



PUBLIC, INCLUSIVE &
INTERACTIVE ENGAGEMENT



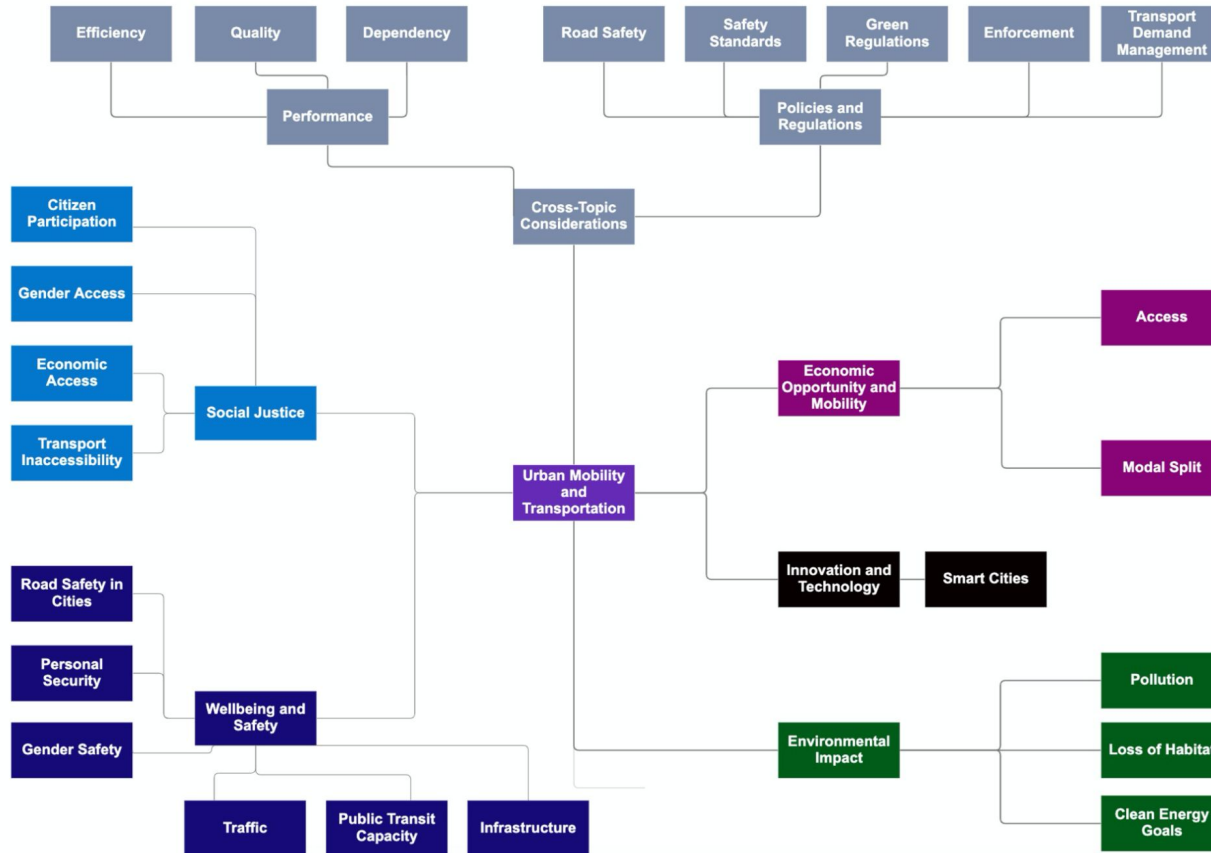
ASSESSING THE STATE OF THE
FIELD & KPIS

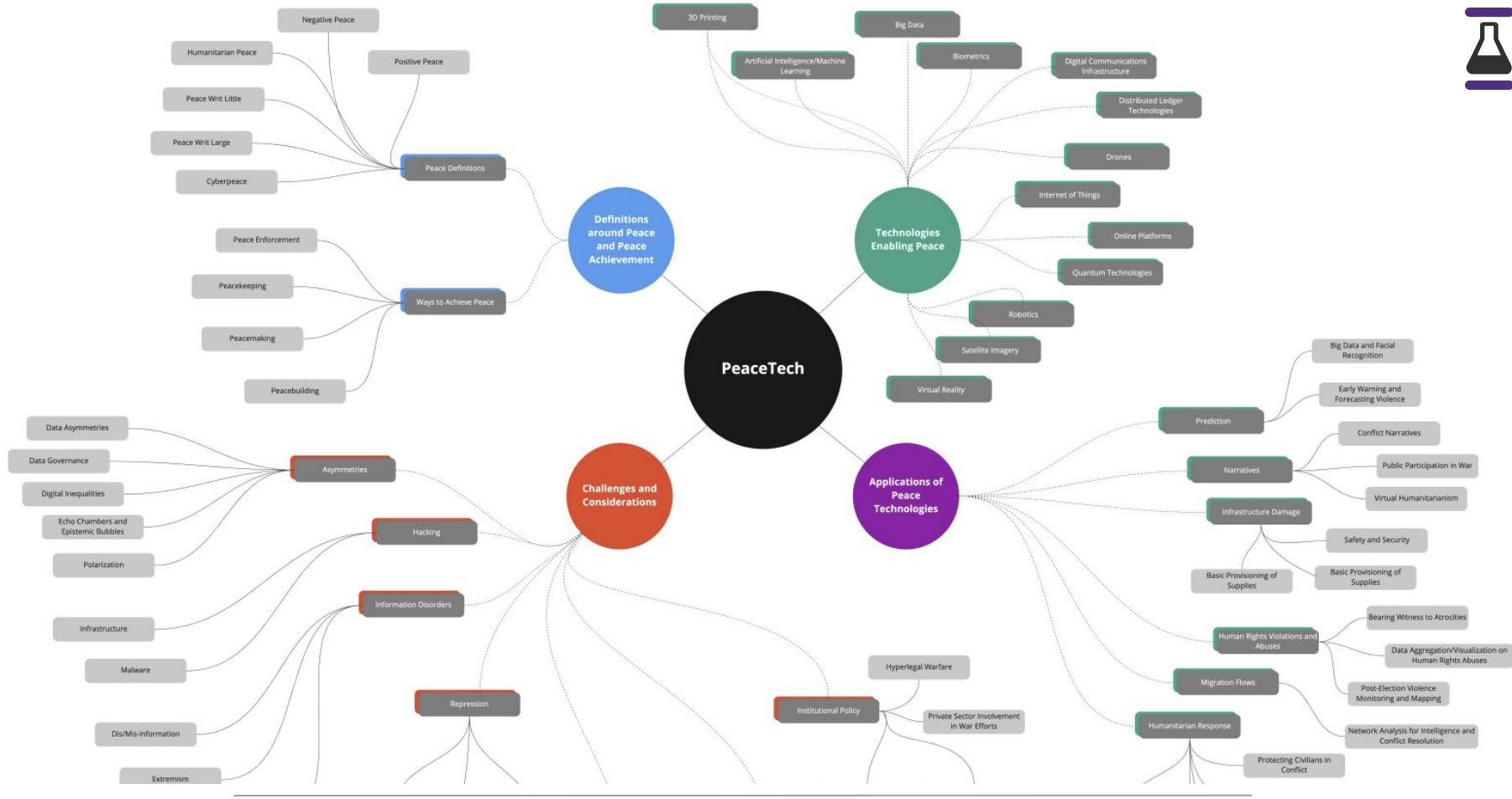


MORE RESPONSIBLE DATA & AI
USE



TOPIC MAPPING



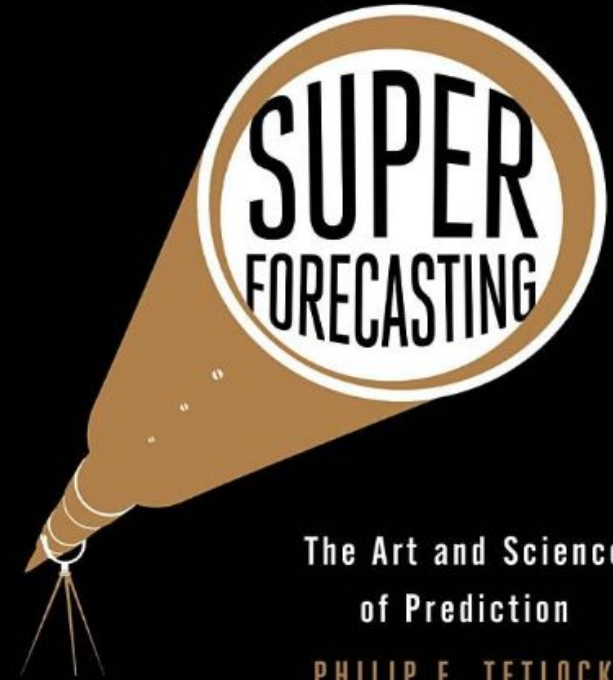


Selection from The GovLab's PeaceTech Topic Map

BILINGUALS: SUPER QUESTIONERS

“**Bilinguals**” are practitioners who possess both relevant domain knowledge and data science expertise. They are experts who understand the importance of data, are aware of the ways data can inform decision-making, and can provide new actionable insights to educate the community on where and how to leverage data responsibly.

Copyrighted Material
NEW YORK TIMES BESTSELLER



The Art and Science
of Prediction

PHILIP E. TETLOCK
DAN GARDNER

“The most important book on decision making since Daniel Kahneman’s
Thinking, Fast and Slow.” —JASON ZWEIG, *The Wall Street Journal*

Copyrighted Material



SIX LESSONS FROM THE 100 QUESTIONS INITIATIVE

1. The **value of data depends on the questions** we seek to answer
2. Data intelligence depends tapping into the **collective intelligence** when formulating questions
3. Knowing what questions to answers requires **knowing what the terrain looks like**
4. There is a need to develop a **taxonomy of questions**
5. **Not all questions are equal**
6. A **question centric approach** enables data responsibility



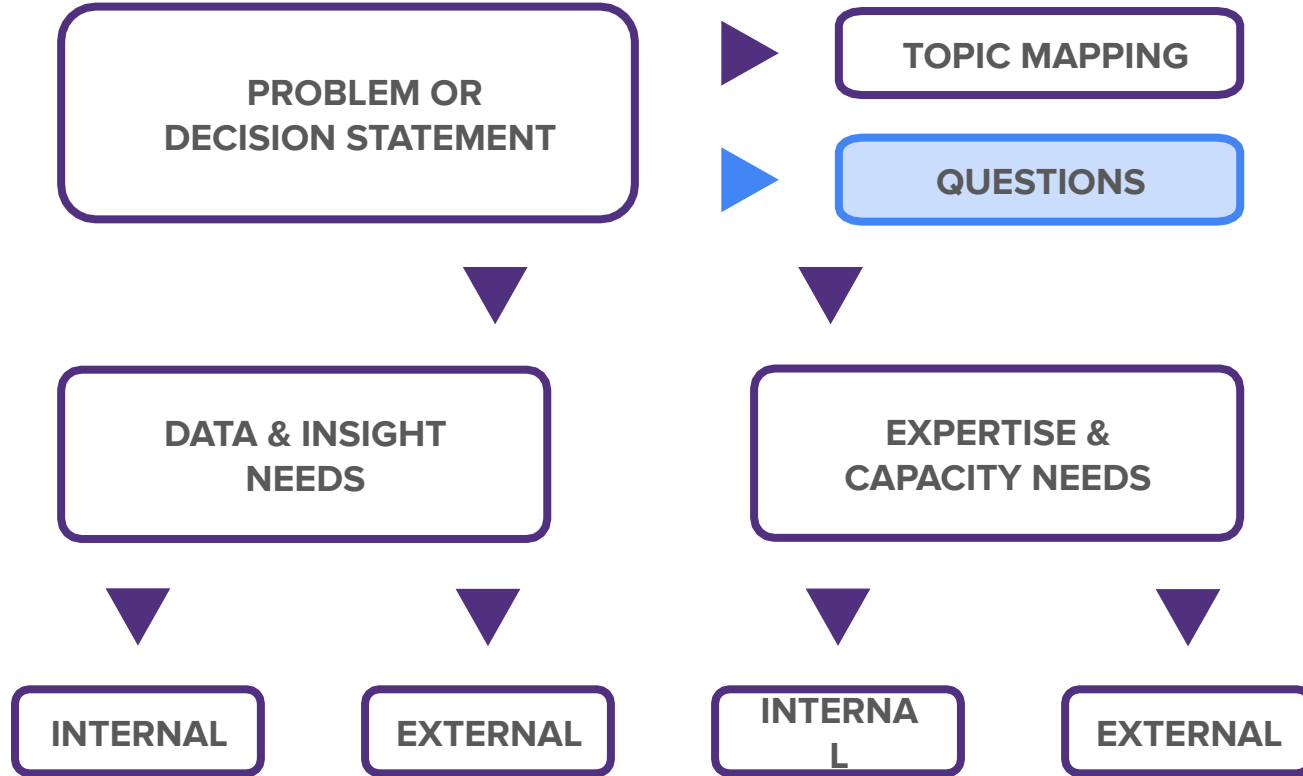


Premise 3:

*Different questions
facilitate different types
of insights.*



GOING FROM PROBLEM TO INSIGHT





USING A TAXONOMY OF DATA-ACTIONABLE QUESTIONS

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



DESCRIPTIVE QUESTIONS

What has happened?

What is the current situation?

Who? When? Where?

- Provides a baseline (“Ground Truth”)
- Quantitative and qualitative
- Situates description in context



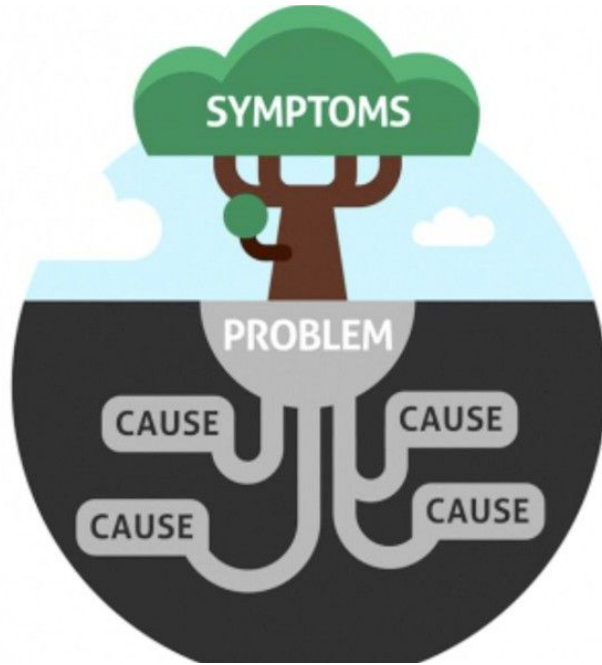


DESCRIPTIVE QUESTIONS: CHALLENGES

- Selective framing and naming
- Perception bias
- Lack of qualitative knowledge & insights



DIAGNOSTIC QUESTIONS



SOURCE: Georgia Lobban. 2019. [“Address the Root Cause, Not the Symptoms”](#) *LinkedIn*.

Why did this happen?

Why are we in the current situation?

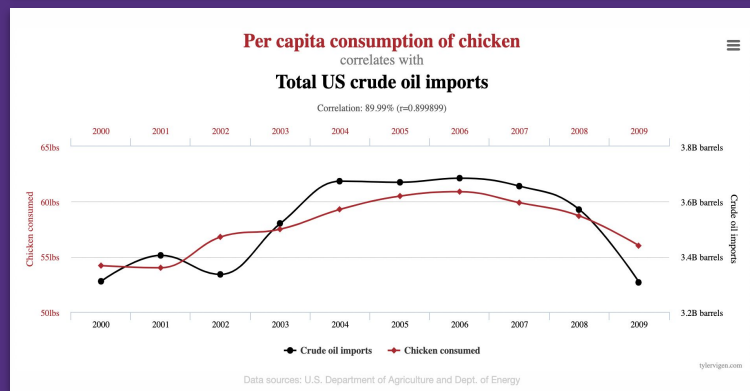
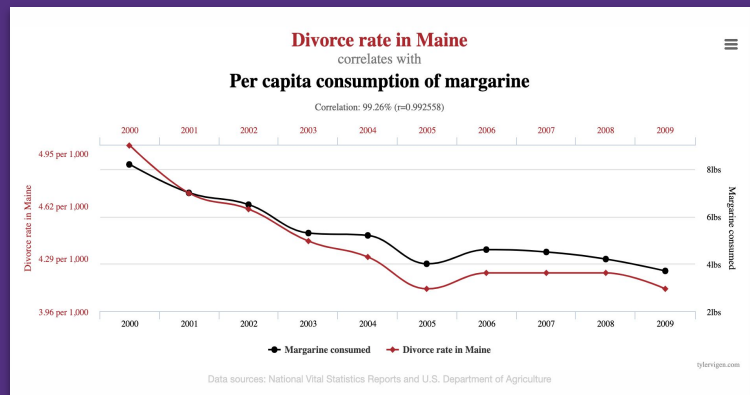
What is the relationship between different phenomenon?

- Examining Root Causes vs Symptoms
- Requires data aggregation

CORRELATION VS. CAUSATION

Correlation does not imply causation.

- Correlation: the size and direction of a relationship between 2 or more variables.
- Causation: one event is the result of the other event (cause and effect).
- It's important to identify the relationship between variables to understand how to achieve a desired outcome.



SOURCE: Tyler Vigen. [Spurious Correlations.](#)



PREDICTIVE QUESTIONS

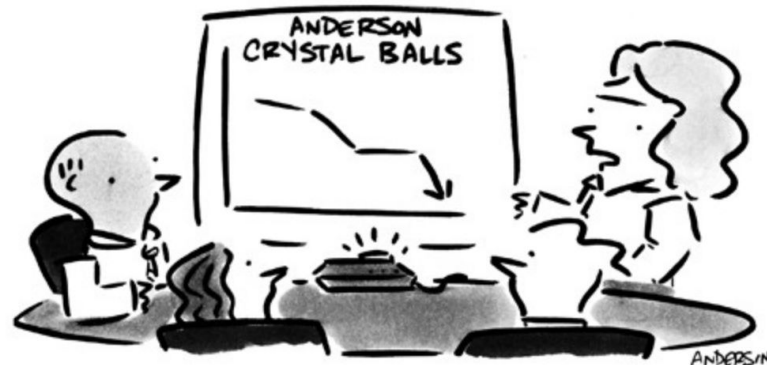
What will happen next?

How will the current situation evolve (based upon past experiences)?

Inferences: If X then Y

© MAZK ANDERSON

WWW.ANDERTOONS.COM

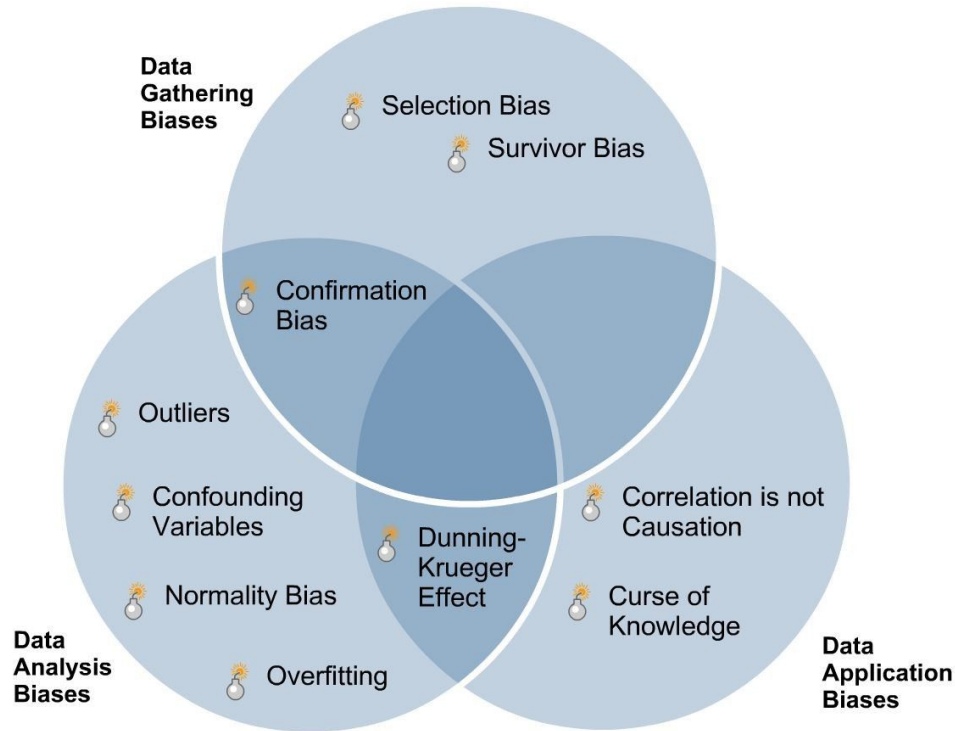


"Seriously?! No one saw this coming?!"

EMERGENCE OF NOWCASTING



BIASED INFERENCES



SOURCE: Martin J. Eppler. 2021. [“Detecting Data Distortions: The Three Types of Biases every Manager and Data Scientist should know.”](#) *LinkedIn*.



PRESCRIPTIVE QUESTIONS

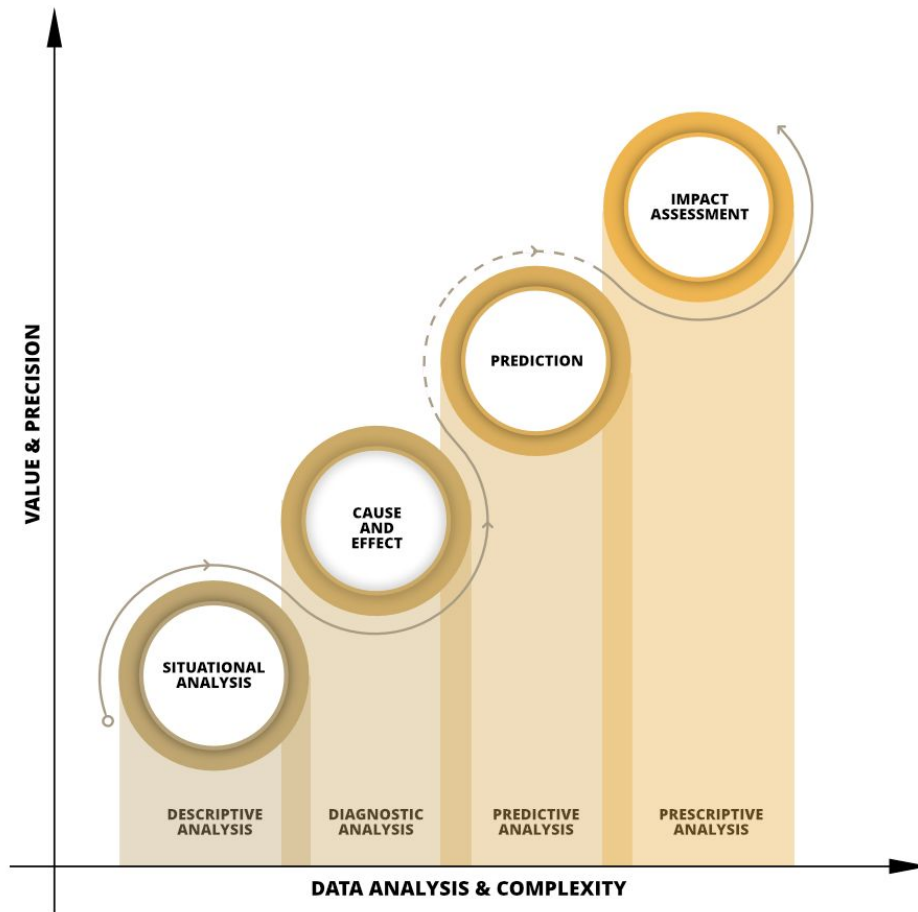


What's the impact of certain interventions?

What if?

- Informs Change Theories
- Informs automated decision-making and reflection

HIERARCHY OF QUESTIONS





CLUSTERING

QUESTIONS THAT MATTER

- **Practical and/or Scientific Impact** — potential to transform our understanding of the problem, fuel innovation or improve how we address societal challenges;
- **Novelty** — questions that have not been answered or analyzed before with answers that will confirm, refute, or extend previous findings;
- **Feasibility and Actionability** — likelihood that the question could be answered with existing data and methods;
- **Quality** — overall alignment with the above criteria.

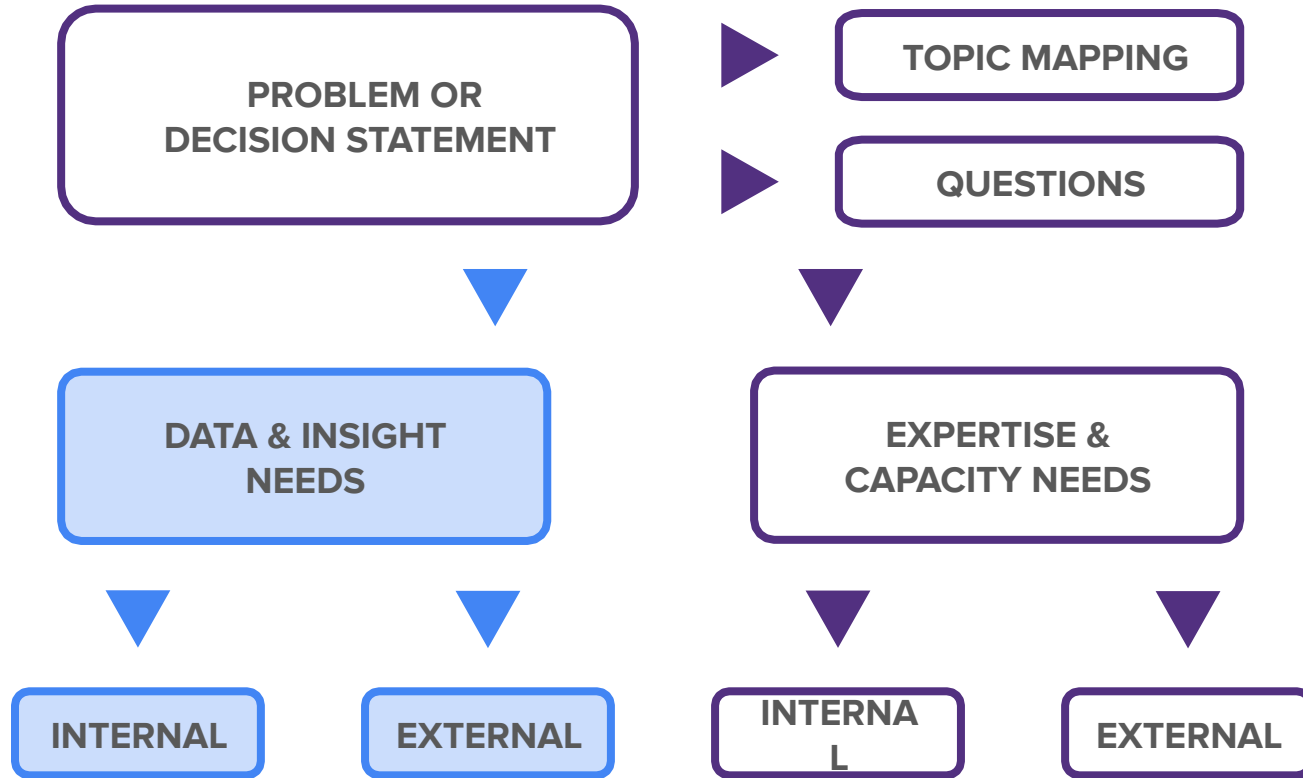


Premise 4:

Focus on the minimum viable data.



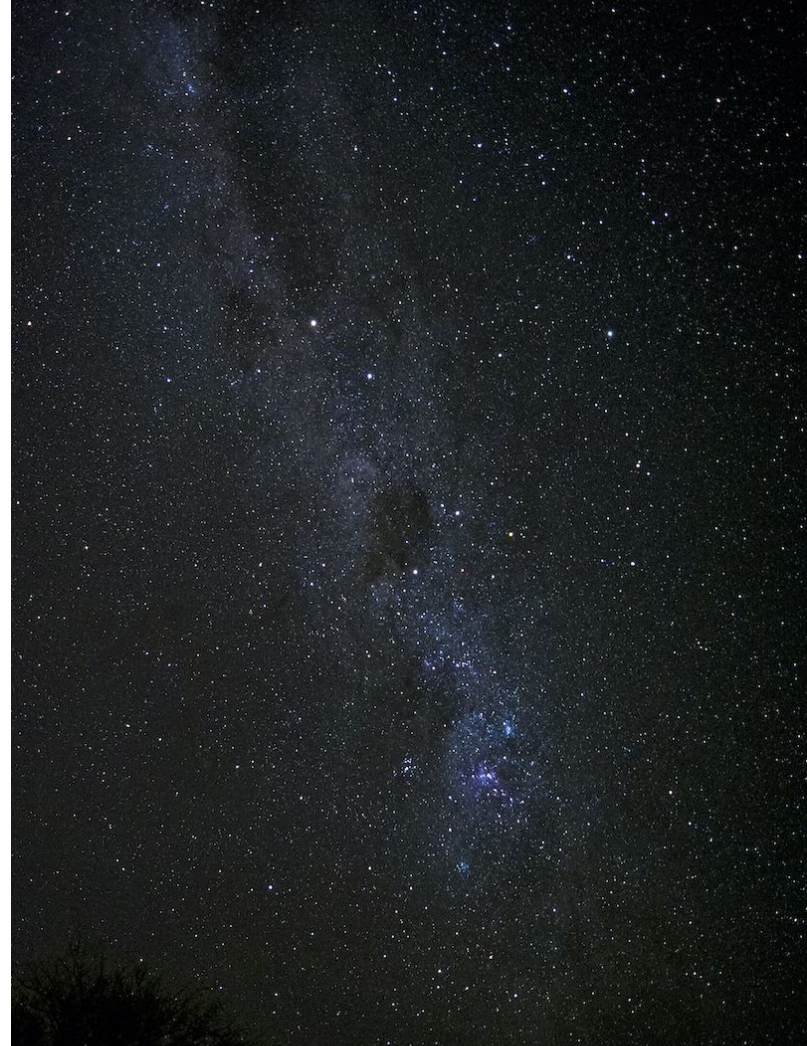
GOING FROM PROBLEM TO INSIGHT





MINIMUM VIABLE DATAPPOINT

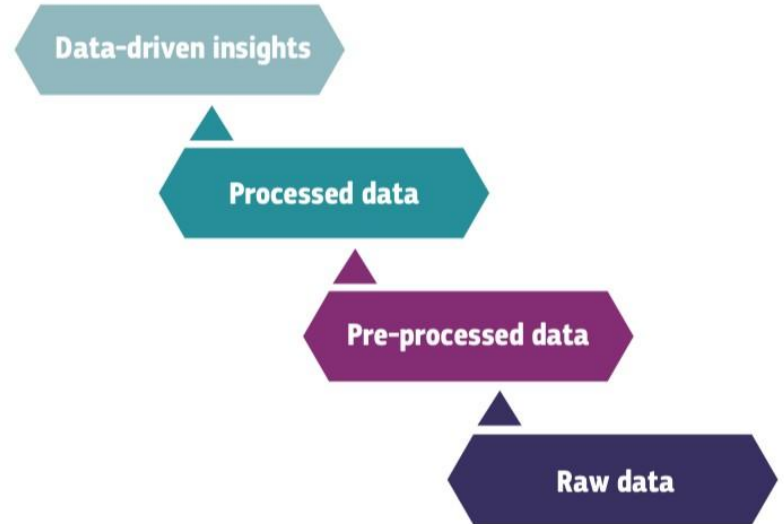
- Your **minimum viable datapoint** is the most minimal amount of data that you need to make progress in answering your question.
- By focusing on a minimum viable datapoint, we ensure that we are using data in a **targeted and responsible manner** and **avoiding inefficiency** by only using relevant data that is proportional to the question you need to answer.





MINIMUM VIABLE DATAPOINT

- What is the type of question being asked?
- Level of disaggregation? Sample size?
- Other data elements vis-à-vis the problem definition:
 - Timescale?
 - Geolocation?
 - Level of abstraction / detail?
 - Comparability / Interoperability?





INTERNAL DATA & INSIGHTS

- For internal data and insights, the use of **data systems or asset mapping tools** can be useful here. These tools can help **keep track of your internal resources** and **help define strategies** for their use going forward.
- The **RD4C Data Ecosystem Mapping Tool** is one example of a systems and asset mapping tool for data stewardship.



RESPONSIBLE DATA FOR CHILDREN

RD4C DATA ECOSYSTEM MAPPING TOOL

VERSION 1 - 2020

Across contexts and regions, the children's data ecosystem is complex and constantly shifting. New data systems, stakeholders, opportunities, and risks arise regularly. Actors who seek to ensure responsible data for children in their context often need to make decisions without a detailed understanding of the current situation.

The RD4C Data Ecosystem Mapping Tool intends to help these actors to identify the systems generating data about children and the key components of those systems. After using this tool, users will be positioned to understand the breadth of data they generate and hold about children; assess data systems' redundancies or gaps; identify opportunities for responsible data use; and achieve other insights.

Through this work, actors can enable the development of a data systems inventory and determine whether the data systems align with the RD4C Principles: Purpose-Driven, People-Centric, Participatory, Protective of Children's Rights, Proportional, Professionally Accountable, and Prevention of Harms Across the Data Lifecycle. After mapping the current state of the children's data ecosystem, users can 1) conduct more in-depth analyses of the policies, guidelines, and risks present in their context through the RD4C Opportunity and Risk Diagnostic Tool; and 2) clarify the decision-making structures affecting these data systems through the RD4C Decision Provenance Mapping Tool.

The RD4C Data Ecosystem Mapping Tool encourages users to capture three types of information about the data systems they use and manage:

- ▶ **WHY** the system exists. This information is important for determining if the system and associated activities are **Purpose-Driven**, **People-Centric**, and if it involves **Proportional** data collection and retention practices.
- ▶ **WHO** is involved in its management and use. Awareness of these stakeholders can help determine whether current data practices are **Professionally Accountable** and **Participatory**.
- ▶ **WHAT** data is held on the system. The types of data held on the system, as well as its inputs and outputs, should inform appropriate actions to be **Protective of Children's Rights** and ensure **Prevention of Harms Across the Data Lifecycle** from the data's initial collection through its eventual use.



DATA TYPOLOGY

DATA COLLABORATIVES | EXCHANGING DATA

EXCHANGING DATA: TYPES OF DATA

Registration records, data included in government transactions, and crowdsourced data. For example, patient health systems records shared by 10 biopharmaceutical companies in the Accelerating Medicines Partnership.

Personal information actively and intentionally shared by an individual, entity or group for a specific reason.

Information free from personally identifiable elements that is actively shared by an individual, entity or group for a specific reason.

Citizen science data, computer system logs, and data on the domain name system. For example, crop data shared through computer systems logs in Intel's Big Data for Precision Farming Initiative.

Internet usage data, commercial transactions like credit card data, and records of energy usage. For example, anonymized energy usage data shared by Dutch energy company, Enexis.

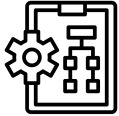
Information with potentially personally identifiable data that is passively collected by an entity prior to any use.

Information with no personally identifiable elements that is passively collected by an entity prior to any use.

Satellite and aerial imagery. For example, geolocational data on the movement of fishing vessels shared by Global Fishing Watch.



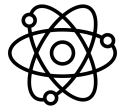
EXTERNAL DATA & INSIGHTS: THE DATA RE-USE ECOSYSTEM



ADMINISTRATIVE DATA



**OPEN
GOVERNMENT DATA**



**OPEN
SCIENCE**



**CIVIL
SOCIETY**



**CROWDSOURCED
DATA**



**PRIVATE SECTOR
PUBLIC INTERFACES**

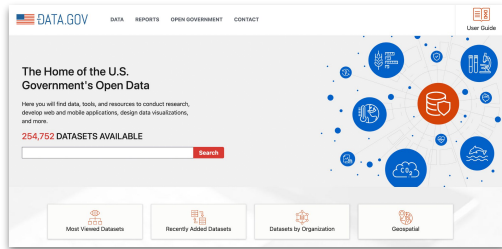


**B2G
ECOSYSTEM**

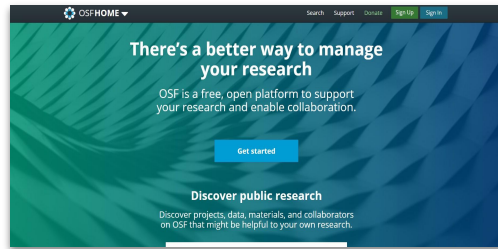


EXTERNAL DATA & INSIGHTS: THE DATA RE-USE ECOSYSTEM

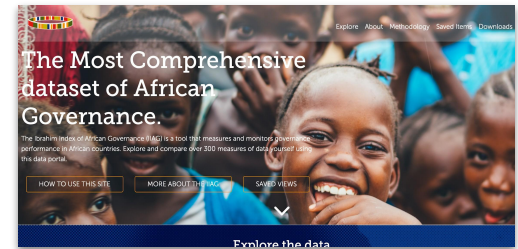
Open Government Data



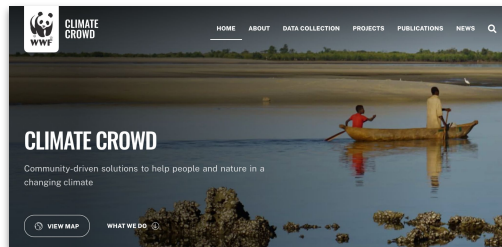
Open Science Research Portals



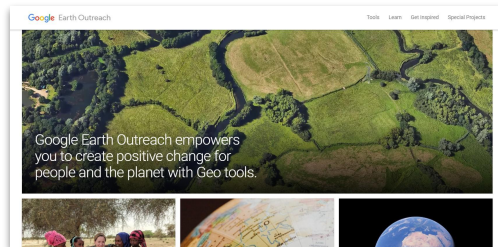
Civil Society Data Portals & Products



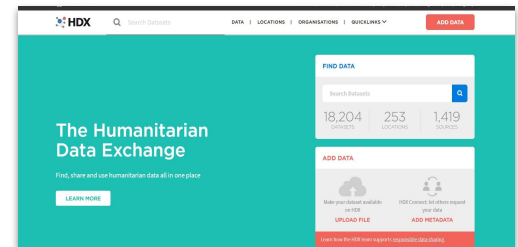
Crowdsourced Data



Private Sector Public Interfaces



Repositories of Repositories



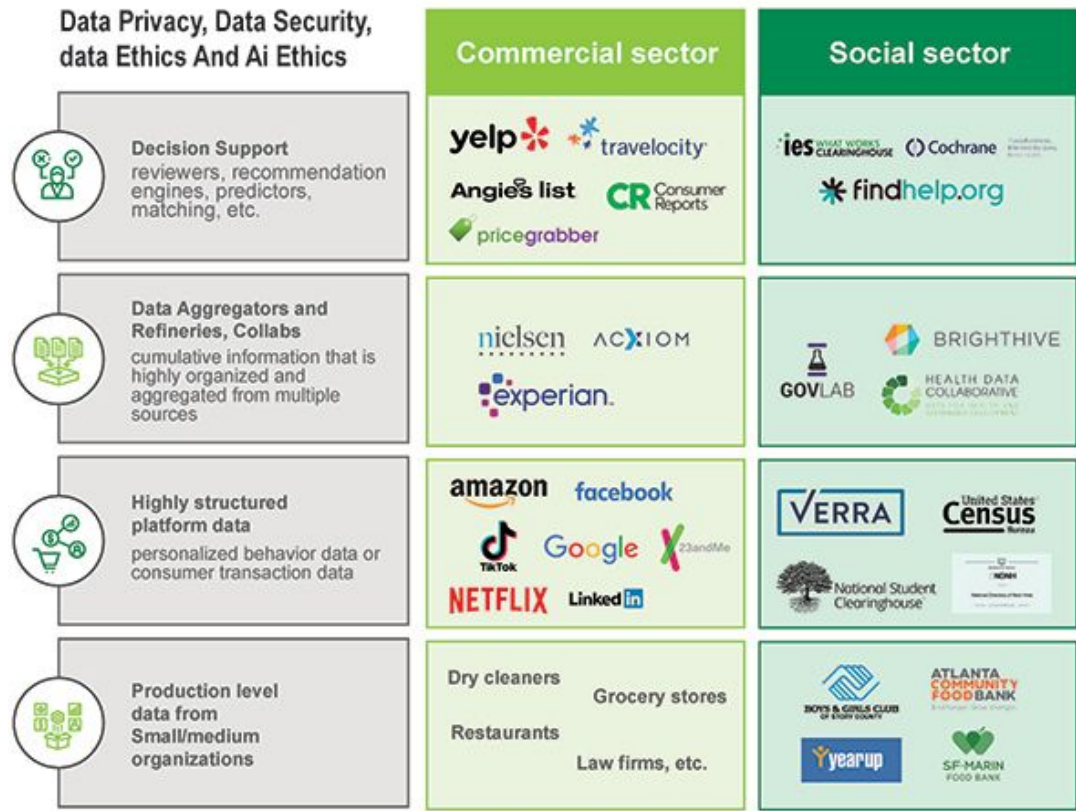


MAPPING THE B2G DATA ECOSYSTEM

An initial snapshot of the complex private data ecosystem, indicating the variety of data that is currently being generated and/or captured by mainly private actors. The snapshot includes the following key information:

Type of Data Generated	Business Actors involved	Includes PII ?	Structured/ Unst. (Semi St.) ?	Location Logged?	Time Stamp?	Example Public Interest Use Cases	Possible use case and Government Re-users
Data resulting from consumption, commercial and financial transactions							
E-Receipts (often emailed) ; Purchase habits, Shopping history and preference information	Retailers and Merchants (with a wide variety of size and sophistication). E-Commerce Sites Utility companies (Registered) Data Brokers	Purchase data; becomes PII when connected with other ID information	S	Y	Y	Deborah Estrin has worked with " Small Data " including receipts to nudge people into buying health food options. Data Does Good , meanwhile, allowed individuals to share their online shopping history anonymously, which is bundled with other user data and sold to raise money for charities.	Statistical Agencies
Loyalty/Reward Card information	Supermarkets; Pharmacies; Airlines and Car rental; Banking institutions	Y	S	Y	Y	The United Kingdom's Consumer Data Research Center collects loyalty and rewards program data from retail and service businesses. This information is available to researchers in criminology, health, transportation services, and other fields.	
Account	E-Commerce	Y	US	N	N		

MAPPING SOCIAL SECTOR DATA ECOSYSTEM



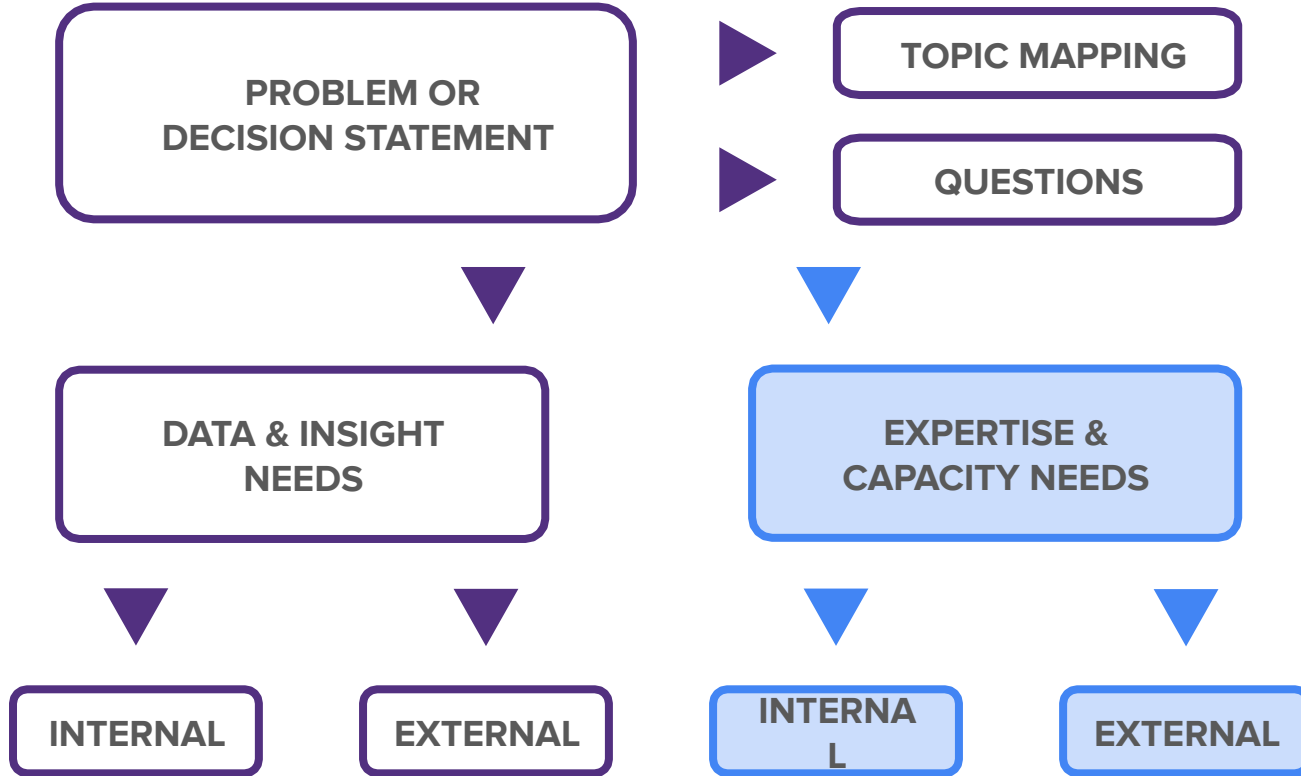


Premise 5:

*Without the right
expertise, the data's
impact will be limited.*



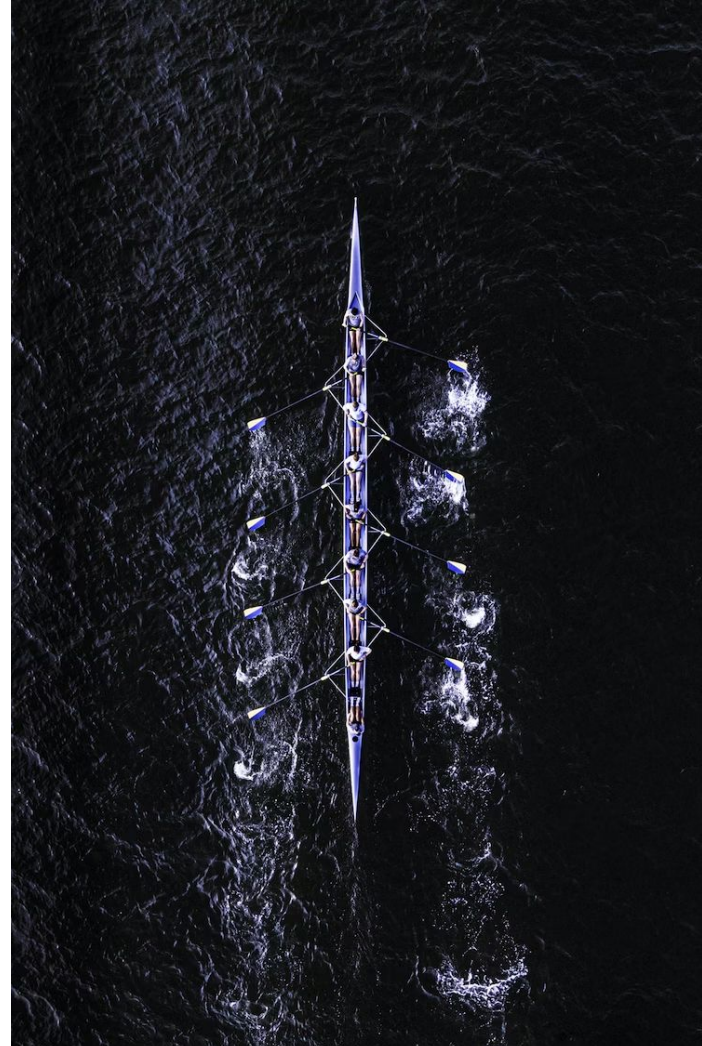
GOING FROM PROBLEM TO INSIGHT





EXPERTISE & CAPACITY NEEDS

- **Expertise** refers to the level of **technical know-how** and **domain expertise** needed to analyze the data to generate insights towards the problem statement.
- This includes **lived experience** (thick data).
- This can be found **internally** and **externally**.
- One of the main roles of a data steward is to **identify, document and manage these people assets**.





DETERMINING RELEVANT DISCIPLINES



LEGAL



**DATA SCIENCE &
ANALYTICS**



FINANCE



DATA ENGINEERING



**MARKETING &
COMMUNICATIONS**



MANAGEMENT



OPERATIONS & HR




DOMAIN EXPERTISE



ASSESSING DATA SKILLS

data.org

Search data.org 

Data Maturity Assessment

[Overview](#)

[About](#)

[Your Results](#)

[Request a Demo](#)

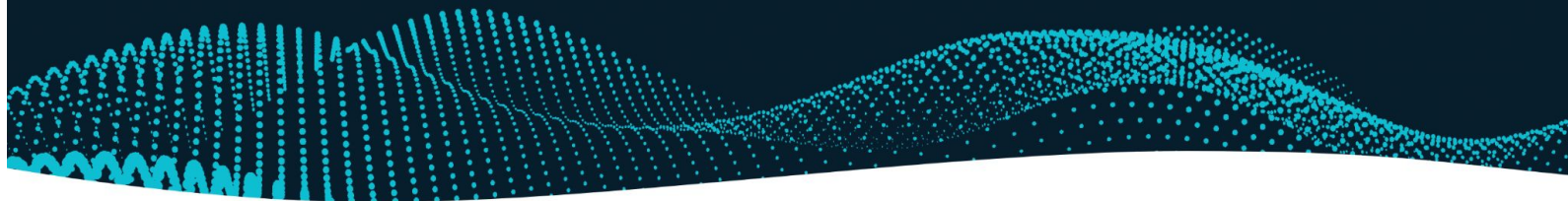
[LET'S BEGIN !\[\]\(6059a5aa8b4ca7bb793408023d6c6e42_img.jpg\)](#)

Start Your Journey Today.

The data maturity assessment offers social impact organizations (SIOs) a snapshot view of their data maturity journey today, and relevant tools and resources to move forward. Use the results to identify ways to strengthen your practice, and share them to communicate areas of opportunity.

[Take the assessment](#)

[Request a demo](#)



SMARTER STATE

SMARTER HEALTH

BOOSTING ANALYTICAL
CAPACITY AT NHS

Beth Simone Noveck
Stefaan Verhulst
Andrew Young
María Hermosilla
Anirudh Dinesh
Juliet McMurren

PUBLISHED FEBRUARY 2017

 **GOVLAB**

With Support from NHS England



DATA ANALYTICAL SKILLS

Example capacities:

- Structured analysis of patterns
- Unstructured data analytical capability
- Predictive capability
- Decision support capability
- Traceability
- Data governance
- Implementation, communication and visualisation



TAXONOMY OF DATA SKILLS

	Examples of technical skills	Examples of translational skills	Examples of Roles
Basic data skills	<ul style="list-style-type: none">» Basic use of statistical software (e.g., Excel)» Data collection and entry» Basic data analysis and manipulation» Basic data visualization	<ul style="list-style-type: none">» Communication (verbal/written)» Ability to interpret and understand key takeaways from data» Collaboration and teamwork» Critical thinking, problem-solving» Data-driven decision-making	<ul style="list-style-type: none">» All non-data specialist professionals who use data in decision-making» Data entry roles» Data collection officer
Intermediate data skills	<ul style="list-style-type: none">» Ability to perform complex functions (e.g., create and maintain complex spreadsheets on Excel; using Stata)» Some programming languages (R, SQL, Python)» Data quality management (DQM)» More complex data visualization» Some knowledge of machine learning» Statistical knowledge (e.g., sampling techniques, ability to design research tools)	<ul style="list-style-type: none">» Ability to understand the nuances and takeaways of data analysis» Ability to communicate results in the organization and to the public» Data presentation skills» Critical and structured thinking» Continuous learning» Ability to work in teams	<ul style="list-style-type: none">» Research analyst» Data analyst» Data associate» M&E associate/officer» Database administrator» Data coordinator» Cloud operations associate
Advanced data skills	<ul style="list-style-type: none">» AI and machine learning» Deep learning» Advanced data analytics and modeling» Predictive analytics» Advanced data visualization» Advanced cloud computing and engineering skills» Deep theoretical knowledge of statistics» Ability to plan and/or manage the entire data lifecycle	<ul style="list-style-type: none">» Data strategy (developing and leading the implementation of strategies)» Ability to work with and lead data teams» Data stewardship» Data ethics» Storytelling skills» Critical and structured thinking	<ul style="list-style-type: none">» Data scientist» Machine learning engineer» AI engineer» Cloud computing engineer» Data engineer» Data architect» Head of research/analytics



EXTERNAL EXPERTISE & CAPACITY: EXPERT NETWORKING

VIVO creates an **integrated record** of the scholarly work of your organization

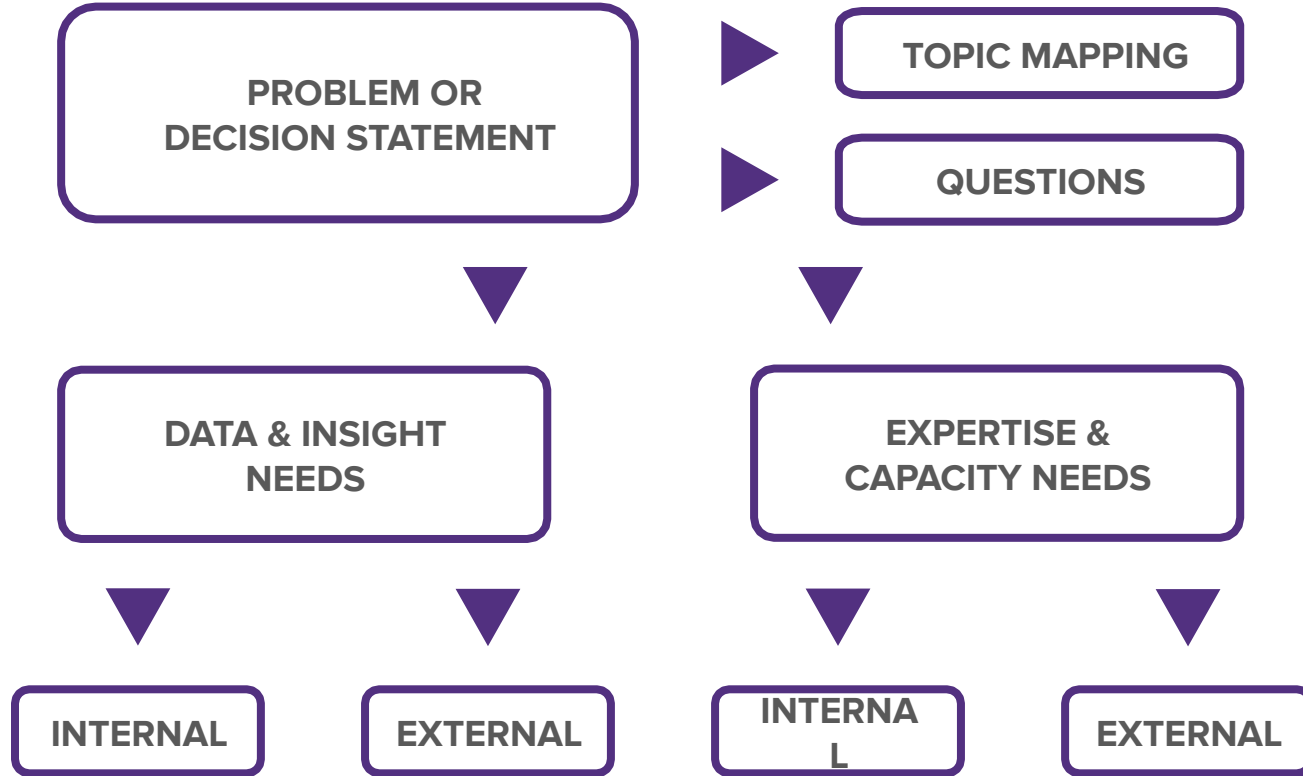
VIVO is member-supported, open source software and an ontology for representing scholarship. VIVO supports recording, editing, searching, browsing, and visualizing scholarly activity.

VIVO is used in more than 60 institutions in more than 20 countries world wide, representing hundreds of thousands of scholars and millions of scholarly works.





GOING FROM PROBLEM TO INSIGHT





SUMMARY

1

Start with the problem, not the data.

2

A participatory process can enable a new kind of “question science”.

3

Different questions facilitate different types of insights.

4

Focus on the minimum viable data.

5

Without the right expertise, the data’s impact will be limited.



UNLOCKING DATA: IDENTIFYING NEEDS & COLLABORATIVE APPROACHES

Session #3: Data Collaboration & Governance

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





STAY IN TOUCH & RECEIVE UPDATES



DATA STEWARDS

The Data Stewards Network (DSN) connects responsible data leaders from the private and public sectors seeking new ways to create public value through cross-sector data collaboration. Watch this space for regular insights and outputs from the Network.



Data Stewards Network

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



@TheGovLab



datastewards@thegovlab.org



@The Governance Lab



www.thegovlab.org