Abstract

What is data? That question is the fundamental investigation of this dissertation. I have developed a methodology from social-scientific processes to explore how different people understand the concept of data, rather than to rely on my own philosophical intuitions or thought experiments about the "nature" of data. The evidence I have gathered as to different individuals' constructions of data can be used to inform further inquiry of data and the design of information systems.

My research demonstrates that people have different constructions of data. The methodology of the , created for this dissertation, has proven able to probe those understandings. The , loosely based on a and combined with ideas from , provides a way of discovering practical definitions of hard-to-operationalize terms like *data*. The process of repeatedly categorizing various items as data allows the methodology to explore how participants actually use the term, rather than relying on theoretical dictionary-based definitions.

Analysis of the interviews found three different constructions of data: data as communications, a container for meaning; data as subjective observations, sense-impressions filtered by knowledge; and data as objective facts, measurements revealing the relationships of reality[1].

[1] For a longer summary of this research, look at Appendix D. The peer-reviewed paper on page ?? was presented at the IEEE 5th International Conference on Computer Sciences and Convergence Information Technology in Seoul, Korea during the process of writing the thesis.

2 Introduction

In Information Systems and Technology studies (IST), I have noticed that practitioners use and understand the term “data” differently than the people they are helping. The purpose of this research is to explore the different conceptions of data that may exist beyond the domain of IST and demonstrate a methodology that allows practitioners to access the conceptions of data present in their workplace.

Exploring a conception of data is fundamentally a philosophical problem. A person’s conception of data stems from the affordances they attach to it, their belief in its underlying qualities, and their differentiation between data and non-data. However, this philosophical problem cannot be solved through intuition alone: a methodology is necessary to extract a person’s conception of data.

These individual conceptions can then be formalised as “philosophies of data.” By ‘philosophies’ we mean answers to the questions like, ‘What is data?’, ‘What is data for?, ‘How do I know the data is reliable?’, and ‘What are the properties of data?’ While individuals may not “have philosophies,” understanding that individuals engage philosophically with their conceptions of data allows the creation of a tool to probe those philosophical conceptions of data in a workplace. By probing conceptions, the IST practitioner effectively uncovers de facto philosophies of data in individuals.

This research, however, does not propose to uncover fundamental philosophies of data, only some common conceptions of data that may exist in workplaces. These different conceptions of data can produce frustration, error, and miscommunication if people with different conceptions interact unknowingly. Conceptions of data include context, reliability, constraints as to its nature (can it be a description, must it be a number), the means of collection, and the means of manipulation.

I have created a methodology called the Social Data Flow Network (SDFN). This interview technique has elicited people’s conceptions of data (their de facto philosophical approaches towards knowing that something is or is not data), demonstrating three different conceptions within a particular industrial research workplace. A survey developed from the SDFN technique hints that there may be different conceptions of data present in the intelligence analysis community and the IST practitioner community.

It is my hope that IST practitioners can use the SDFN I have developed to make better interfaces and databases: through the understanding of a client’s expectations of data, the system can provide natural interaction methods that conform to the client’s expectations of what data is and is not. The SDFN might also be used within an organization to reduce miscommunication and error: the explicit definition of one particular conception of data for a workplace.

2.1 Methodological Summary

The primary result of this thesis is the methodology of the Social Data Flow Network. The SDFN uses repeated categorization to explore how individuals group informational or communicative flows into categories. By eliciting categorizes that focus on data, information, and knowledge, the participants use the categorization to operationalize their epistemological understanding of data: they indicate what is and is not data and how it becomes information and knowledge. This elicitation helps both the interviewer and the participant to discover their own situational conceptualization of data.

The repeated categorization allows participants to generate and resolve cognitive dissonance situated around the differences between their theoretical definitions of data and their practical uses and categorizes of data. In interviews, participants demonstrated a refined understanding of their own conceptions of data at the end of the interview, catalyzed through their participation in the SDFN.

The SDFN involves the articulation of roles as entities, descriptions of content flows between those entities and the categorization of those flows as data, information, knowledge, or other. Participants iterate over a task domain defined at the start of the interview, discussing all the entities and flows between those entities involved in the task. The interview concludes with an opportunity for the participant to self reflect on their "philosophy" of data, discussing what they categorize data as and how it becomes information and/or knowledge.

A scenario based survey, inspired by the SDFN was also trialled with less satisfactory results. While the survey did demonstrate that intelligence officers, IST professionals, and other industrial research employees did have different conceptions of data, it did not do so with any statistical rigor nor with the depth of discussion that the interviews provided.

The combines two concepts for a novel purpose. It is a graph[2] that combines the

[2] A graph, strictly speaking, is any diagram that contains edges and nodes. A node is the component of a graph that is a point. The point can be labeled or unlabeled. The node is the element of the graph that is a representation of a thing. Sometimes the thing being represented is a computer or a person, or a place, but in any event the node represents a noun. Edges on the other hand are the relationships or connections between nodes. An edge represents a "flow" of action or stuff between nodes. Edges traditionally have served as network links, roads, phone lines, and simple representations of adjacency. A graph is a non-topological method of representing the relationships between entities through edges and nodes.

Edges can be directed: they show a flow or relationship from one node to another. The direction on the edge indicates the direction of relationship. For example, consider Alice and Bob. To represent Alice sending a letter to Bob, we would make both of them nodes and draw a directed edge from Alice to Bob indicating the one way flow of the letter. By adding the concept of directionality to edges, a causal element is introduced to the representation-specifically, that the originating node idea of the social network with that of the data flow diagram. In social network analysis, it is possible to represent interactions between people, a social network, through graphs. Each node on a graph represents a person and each edge represents some sort of connection between people, as a function of the interactions of interest to the researcher[[error 2]].

The Data Flow Diagram[[error 2]] contributes its diagrams to the . A originally was designed for structured programming. The document produced by the would combine the delineation of a universe of discourse via the context diagram with the highly precise definition of flows into and out of that diagram. A ()[[error 2]] is the term used for defining the topic under consideration. Everything within the is relevant and must be modeled. Everything outside the is irrelevant. Interestingly, as the was repurposed for business modeling, the remained the same: it is still asking, "What bit of reality do we care about right now?"

The would then be refined through a process of "zooming in" on that context diagram to expose the transformations required to produce the outputs from the inputs. Each additional level would seek to conserve inputs and outputs, and thereby produce a diagram that could be mapped to the functions and variables necessary for a structured program.

The contributes great ideas to the . It contributes the idea that data *is* something that can be modeled. The conception of data embodied by the is that the modeler can translate reality into data-as-bits and that data could be described through text. All actions in the data flow diagram are considered either flows or transformation. Data flows from sources through transformations, and out into sinks. The sources and sinks are entities outside the scope of the diagram. By decomposing these transformations into ever simpler and more detailed sets of sub-transformations, modelers could design an entire software system intended to process and transform data. The modeler acts as translator: taking the described reality by the client and forcing it into a computerized mold. Repurposing the methodology of the by subtracting the modeler's translation suggests that it might be possible to use my method to probe and document a client's subjective reality.

The also contributes an iterative structure for the definition of reality. The iterative techniques explore the in order of increasing specificity from the vague context diagram describing the universe of discourse to highly detailed sub-sub-sub (etc.)

causes a relationship to the recipient node. This addition of causality then precipitates the idea of connectiveness.

A node may or may not be reachable by other nodes. A graph or subgraph where every node can be reached from every other node is called a strongly connected graph. A graph where that's not true is weakly connected. When we apply the idea of strongly connected graphs to social networks, we can identify small groups by identifying strongly connected subgraphs within a larger, weakly connected graph.

transformations required deep in the diagram. By starting with broad generalizations, the insured that the client was thinking about the whole task and did not immediately become fixated on any one aspect. With the iterating across each declared "transformation" and decomposing it, the details of each transformation were both evoked and then situated in the scaffolding of the broader context. The requirement to conserve inputs and outputs eliminated any question of missing aspects of the diagram or other design-based blind alleys. The idea of iterative exploration and definition is extremely valuable to the .

The Social Network Graph provides the concept of a social network[3] to the . The Social Network Graph also contributes a novel idea about the *scope* of edges. Edges in the were simple *flows* of data, representing the movement of trivial signs. In the social network graph, edges can be individual communications, orders, relationships, and objects. The huge diversity of edge types suggested by a social network graph, when combined with the , ruins the for its original purpose: the modeling of software systems. However, they also suggest different possible models that can be applied to the format.

In communicative analysis, social network graphs are used for linguistic analysis[4]. It is possible to explore the control structures of a group by noting, with an edge, who is talking to whom. By exploring the frequency and directionality of those notes, analysts gain insights into the power and influence roles of social networks. As such, the "thought leaders" of the small group can be identified.

Moreover, by graphing flows of communication, it is possible to identify small groups within larger groups, as these small groups will communicate strongly between each other and vaguely to nodes outside. In other circles, this behavior is known as siloing[[error 2]]. One design intent of the is to confer the ability to identify siloing. By rendering flows between members of an organization, it should be possible to identify strongly connected sub-graphs, which suggest communicative silos within that organization.

The social network graph contribution alters the diagramming rules of the . Social network entities can be any actor that participates in a communication. The is a diagram exploring flows of data between actors, instead of flows between transformations. By creating a web of affiliation[[error 2]] between these entities, it should

[3] A social network graph is a mapping of a person's relationships with other people into non-topological graph format. Each relationship is a directed edge; each person, a node. The social network graph is used in many different fields: communications, social media, and sociology are some of them. In many ways, the idea of the social network graph is strongly related to the ideas of actor-network theory[[error 2]].

[4] figure ?? provides a trivial example of linguistic analysis as applied to a set of twitter replies during a conference. The different line weights are used to denote quantity of communications along a radially distributes set of nodes. Other approaches can be far more complex, looking at patterns beyond simple frequency[[error 2]].

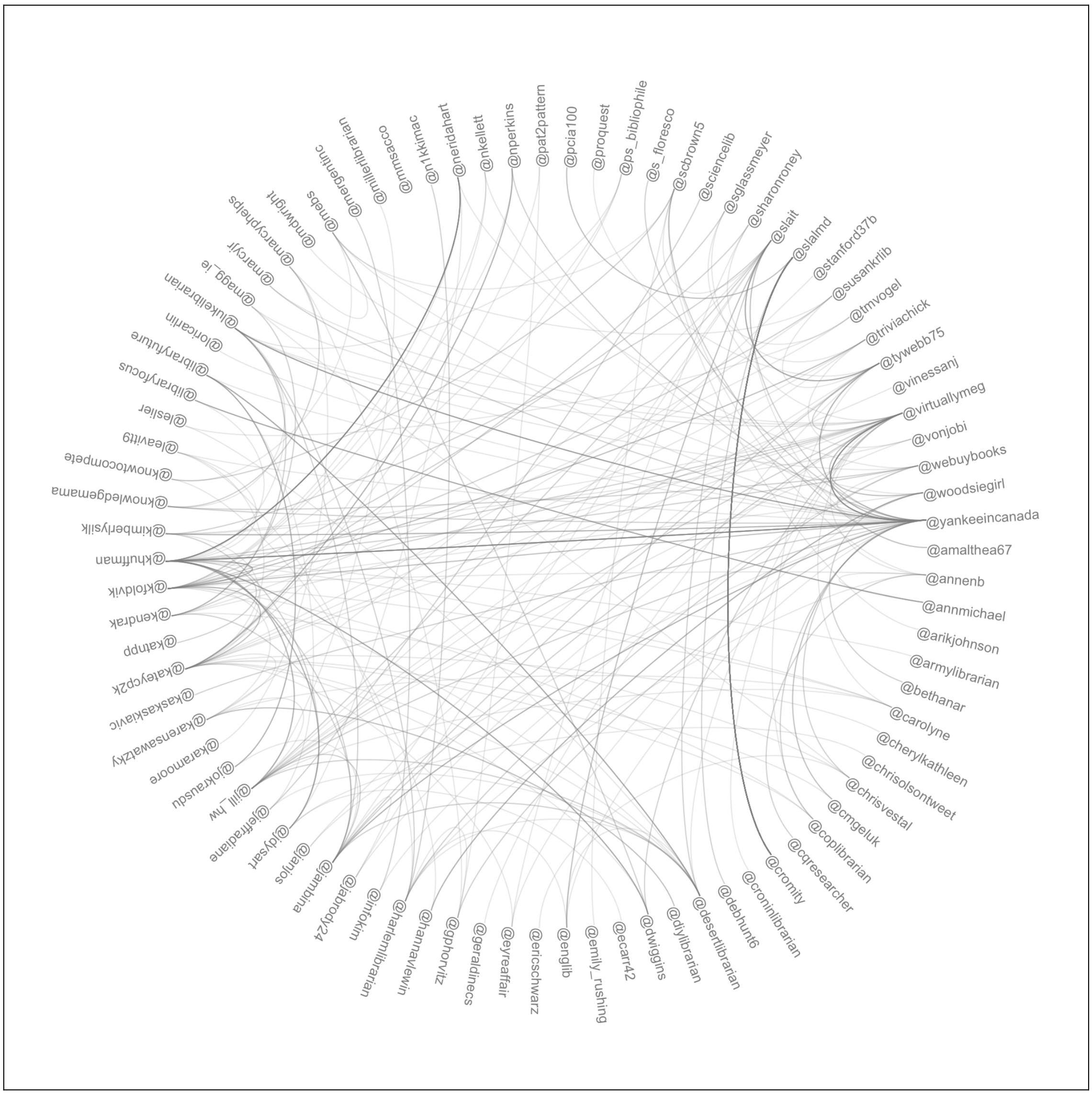

Figure 2.1 Social network graph of #sla2009 tweet replies to June 19, 2009 “The thicker the line, the more times you sent an @reply to that person. The more lines you have, the more @replies to different people you sent. If you don’t appear on the graph, but know that you sent out @replies, it’s because the person you sent your @reply to never sent out an @reply and so that person won’t appear on the graph and unfortunately, you can’t either! Interestingly, a few people only sent replies to themselves, so they do appear on the graph as a line that goes back to themselves.” -Image used with permission, created by: Daniel P. Lee, MLIS.

be possible to describe the communicative realities that an individual perceives. It should therefore be possible to explore how they understand the nature of data by exploring how they describe its movement from entity to entity in the .

Despite the terminology of actors, and the use of a social network, my research does not yet incorporate actor-network theory[[error 2]]. While Latour's work offers many useful ideas for understanding the world, it still imposes a framework from which biases may be imparted. Therefore, while I do not use actor-network theory here, it may be useful in later research exploring the implications of held philosophies on Latour's work.

The does not try to be explanatory, comprehensive, or objective. The point of the is to reveal part of how the participant understands a concept, not to build upon that understanding nor transform it into a model for a computer system. Consequently, no design provisions in the methodology allow two or more peoples' categories to be reconciled. More work will be necessary before the can be used directly as a design methodology.

2.2 Analysis and Results

These questions of interest are posted to the reader to keep in mind in the results section. My personal analysis, presented after the "raw data," uses these questions of interest as framing devices for my reflections on the individual interviews.

2.3 Questions of Interest and the Methodology of Analysis

My "hypotheses" are described as questions of interest to reflect the rapid iterative nature of abductive explorations. They provide research directions that act as broad guides to the formation of a universe of discourse for future research rather than predictive statements about reality.

The intent of the questions is to frame analyses and guide it towards useful and interesting areas. We need to consider how the evidence relates to these questions of interest.

Each interview, after transcription, was subjected to recursive analysis for my personal reflections on the interviews. I summarized six to ten lines of each interview in a one-line summary. Then between three and six summaries were summarized, filtering for statements about the user's conception of data. Although self-transcription transmits personal bias, two significant factors prevent a traditional double-blind study. An untested methodology is no place for the mass utilization of volunteer interviewers. The limited scope allowed me to retain control of the interview process and to provide for the best possible interviews for each participant while retaining the basics of the . Because I conducted each interview, the bias would have already been introduced; providing for pious-sounding human coding would have lent false reliability to something inherently subjective.

My personal reflections are very simple. I have tried to extract each participant's intuitions about data from the recursive analysis.

2.3.1 Question of interest 1: Do People have different philosophies of data?

If this research produces nothing else, it must investigate whether people have different conceptions of data. This idea was the central intuition that prompted this research, and its testing will demonstrate whether or not there is anything to my intuition.

As the organizing factor of my analysis, this question of interest will focus my activities. It will justify further research on the philosophy of data from my experimental results, or else its demonstrable failure will justify not doing so.

The question "Do people have different philosophies of data?" defines an overly large universe of discourse, one impossible to study at a useful level of granularity in one research project. The very breadth of the question precludes the determination of any useful and specific facts about the world besides simple exploration of the assertion that people have different understandings of data. The intent of this research and of this question is to generate interest in the research of the philosophy of data and to demonstrate that there are questions to research.

I want to see if, beyond my intuitive insight, people actually have different conceptions of data or if my perception of different philosophies is an artifact of the requirements-gathering process of designing a database. It is therefore not sufficient to state that people have different understandings of data depending on whether they are dealing with it in a technical or scientific context. We must look for evidence.

This question of interest, in its reach, is not ambitious. It suggests no predictions about peoples' conceptions of data, how they act with different realities of data, or any other fact about the world. Instead, it simply directs us to see if there is anything of interest for further explorations.

2.3.2 Question of interest 2: Can my methodology probe people's philosophies of data?

My methodology has a simple job: to assess what people mean when they use the term "data." This question of interest is designed as a sanity check. I am investigating a new idea with an untested methodology. It is vital to consider that the success or failure of Question of interest 1 is directly modulated by the success or failure of Question of interest 2. Therefore, the methodology itself deserves distinct analysis.

The methodology should be of use to more people than just those investigating peoples' conceptions of data. If the methodology is useful and judged to add value to Question of interest 1, analysis of the methodology should indicate whether other people could use it to investigate matters of interest to them.

Question of interest 2 is asking: do these results make sense? Sense-making is a matter of internal and external consistency. This question should force me to explore whether the correlates with interview results and whether the types of results make sense relative to the survey.

Beyond consistency, I must also ask: Is it possible to get these results from this methodology? In this case, I need to make sure that I am not reading imaginary meaning in the tea leaves of the results. Because this kind of external self-reflection is difficult, the question must be simplified to: Do the results surprise me? If they do not have elements of surprise, then the probability that I am projecting meaning into them must be strongly considered.

All of these are very self-critical questions, as they must be to explore the impact of an untested methodology. I am trying to consider whether my methodology can present a persuasive story, and if it can, does it?

2.4 Interview Analysis

My interview analysis discovered three different conceptions of data. It would be hard to deny that interviews I and II have data as communication, III and IV have subjective observations (with IX hinting at them) and the rest considering data as objective fact. With these broad differences evident, I feel question of interest 1 has been satisfied.

The observation constructions differ strikingly from the numeric constructions, possibly differing on a fundamental perception of reality. As one interview is trying to render the relationships between matter in the world as numbers (objectivist), another is suggesting that everything emits data and we must filter it. The conflict is records versus measurements versus signs. Does data measure objective reality, record subjective reality, or merely transmit signs? Numbers are seen as a result of precision most of the time, whereas observations are building their way towards knowledge.

2.4.1 Data as communications

Data, in the communicative sense, merely requires signs and things to communicate with those signs. The data can be rendered as bits or marks on paper, but it is seen as a factor of semiotic import rather than as something to be discovered or filtered. This construction is substantively different from the other two inasmuch as it does not uphold data to be an aspect of reality. Instead, data is produced as a function of human intent. Because this understanding does not concern itself with interactions of the real, there is a far greater difference between this and the other two than between the subjective-objective philosophies. However, the passivity of this construction allows it to accept facts produced from either source as something to be encoded, stored, and transmitted. Significant research needs to be done to explore how this construction of data relates to the other two.

2.4.2 Data as subjective observations

Data, in the subjective observations, requires contextualization and filtering. Everything emits data as sense impressions[5] that can be captured by us. Thus, to perform sense-making activities, we must filter and contextualize the interesting data so that it can become information.

Subjective data lends itself more to cyclic hierarchies, where data begets the information and knowledge used to collect more data, reflecting an interestingly constructivist view of knowledge. There is quite a lot here available to future research, and I do not feel sufficiently confident in my sample size to make any assertions as to relationships between data and the various philosophies of knowledge or science, though the subjective nature of observations may tend slightly more towards Latour or Feyerabend.

Of more interest is that this inherently subjective data is constructed from the mind's impressions of the surroundings, rather than revealed through measurement of the surroundings. The understanding of the embodiment of data is a significant difference between the two understandings of data.

[5] Like the ancient Aristotelian idea of species (particles of sensation). Light was the medium that visual species traveled within.
While this ancient philosophy of image is not hugely useful to us, the same intuitions that led to it could have some parallels with data as subjective observations. This research area could make an interesting bridge between intuitive and experimental philosophies.

2.4.3 Data as measured facts

Objective data comes with its own context “baked in.” It is, in many ways, rare: it requires positive effort to generate, and higher quality data requires a commensurate increase in effort. Data requires analysis to uncover the extant patterns of reality, and with enough data, knowledge about the singular real can be generated.

Objective data requires that data be a fact, usually a numerical, reproducible representation of reality that conveys an understanding of measurement quality and units. Objective data is not filtered, because it is collected with prior intent and all elements of the “data set” may produce interesting patterns.

Both humans and sensors can reveal objective data, which is embodied in the things being measured. There seems to be no significant link with any of the major philosophies of science. Although my investigations did not explore confirmation, falsifiability, or paradigms, there seems to be a common understanding that data-as-fact accurately represents the universe within the constraints of measurement. This may be because the participants believed data to be a building block upon which their hypotheses or understanding of the universe could be built.