The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (2014)

by Rob Kitchin

Chapter 1: ’Conceptualizing Data’

About

Rob Kitchin is a professor and ERC Advanced Investigator in the National Institute of Regional and Spatial Analysis at Maynooth University, for which he was director between 2002 and 2013. He is one of Ireland's leading social scientists and was the 2013 recipient of the Royal Irish Academy's Gold Medal for the Social Sciences and the Association of American Geographers 'Meridian Book Award' for the outstanding book in the discipline in 2011. After studying at Lancaster University (BSc Geography), Leicester University (MSc GIS) and the University of Wales Swansea (PhD Geography), he took up a post in Queen's University Belfast in 1996, moving to Maynooth University in 1998.

He has published widely across the social sciences, including 23 books and over 140 articles and book chapters. He is editor of the international journal, *Dialogues in Human Geography*, and has been an editor of *Progress in Human Geography* and *Social and Cultural Geography*. He was the editor-in-chief of the 12 volume, *International Encyclopedia of Human Geography*, and edits two book series, *Irish Society* and *Key Concepts in Geography*. He has successfully written or been a principal investigator on forty grants, totalling c.€34m, including funding from PRTLI 2, 4, 5, IRC, ERC, SFI, ESRC, NSF, Interreg and RIA. He is currently a principal investigator on the Programmable City project, the Digital Repository of Ireland, the All-Island Research Observatory and the Dublin Dashboard. He has delivered over 130 invited talks at conferences and universities and his research has been cited over 600 times in local, national and international media.

Conceptualizing Data - Why

“Given their utility and value, and the amount of effort and resources devoted to producing and analysing them, it is remarkable how little conceptual attention has been paid to data in and of themselves” (28)

“when data are the focus of enquiry it is usually to consider, in a largely technical sense, how they should be generated and analysed, or how they can be leveraged into insights and value, rather than to consider the nature of data from a more conceptual and philosophical perspective.” (28)

Conceptualizing Data - Overview

Goal: “To supply an initial conceptual platform [about] . . . the forms, nature and philosophical bases of data”

Position: “Data do not exist independently of the ideas, instruments, practices, context and knowledges used to generate, process and analyze them (Bawker 2005; Gitelman and Jackson 2013)”

Definition in use thus — Capta: “Over time, data came to be understood as being pre-analytical and pre-factual, different in nature to facts, evidence, information and knowledge, but a key element in the constitution of these elements . . .When a fact is proven false, it ceases to be a fact. False data is data nonetheless” (31)

Capta: Data as Rhetorical

Capta:
Data as Rhetorical

part 1

In rhetorical terms, data are that which exists prior to argument or interpretation that converts them to facts, evidence and information (Rosenberg 2013). From this perspective, data hold certain precepts: they are abstract, discrete, aggregative (they can be added together) (Rosenberg 2013), and are meaningful independent of format, medium, language, producer and context (i.e., data hold their meaning whether stored as analogue or digital, viewed on paper or screen or expressed in any language, and 'adhere to certain non-varying patterns, such as the number of tree rings always being equal to the age of the tree') (Floridi 2010). Floridi (2008) contends that the support-independence of data is reliant on three types of neutrality: taxonomic (data are relational entities defined with respect to other specific data); typological (data can take a number of different non-mutually exclusive forms, e.g., primary, secondary, metadata, operational, derived); and genetic (data can have a semantics independent of their comprehension; e.g., the Rosetta Stone hieroglyphics constitute data regardless of the fact that when they were discovered nobody could interpret them).

part 2

A taxonomy/map of ‘data’ as a term

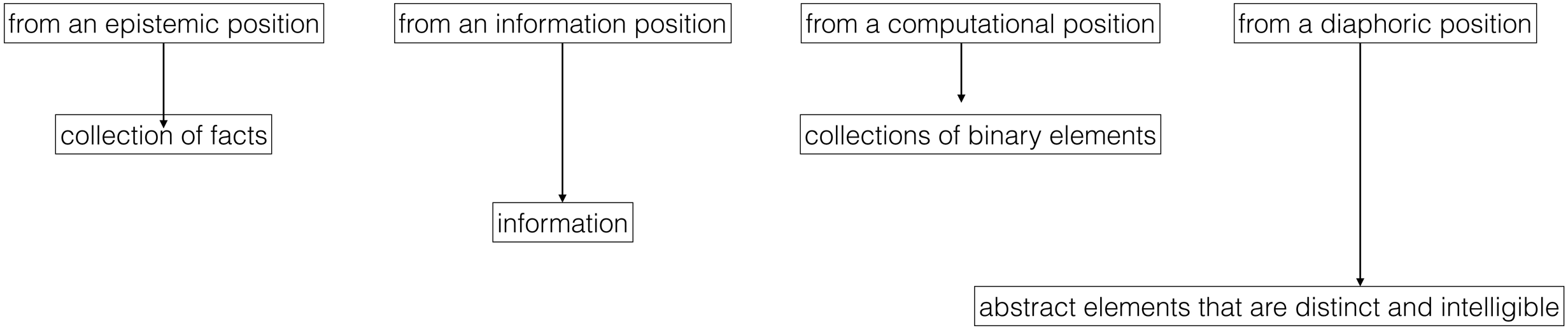
from an epistemic position

from an information position

from a computational position

from a diaphoric position

A taxonomy/map of ‘data’ as a term



A taxonomy/map of ‘data’ as a term

from an epistemic position

collection of facts

provide basis for reasoning & constitute evidence

from an information position

information

constitute representative info that can be stored/processed/analysed, but are not necessarily facts

from a computational position

collections of binary elements

constitute inputs & outputs of computation, have to be processed (ex: a DVD)

from a diaphoric position

abstract elements; distinct and intelligible

are meaningful because they capture and denote variability

A taxonomy of kinds of data

structured/semi-structured/unstructured

structure

captured, exhaust, transient, derived

source

primary, secondary, tertiary

producer

quantitative/qualitative

form

indexical, attribute, metadata

type

A taxonomy of kinds of data

quantitative/qualitative

form

Quantitative: numeric records, relate to physical properties of phenomena (length, height, distance, etc), or rank/class/etc

Qualitative: non-numeric - texts, pictures, art, video, sounds, and music.

Is art data?

In contrast, *qualitative data* are non-numeric, such as texts, pictures, art, video, sounds, and music. While qualitative data can be converted into quantitative data, the translation involves significant reduction and abstraction and much of the richness of the original data is lost by such a process. Consequently, qualitative data analysis is generally practised on the original materials, seeking to tease out and build up meaning and understanding rather than subjecting the data to rote, computational techniques. However, significant progress is being made with respect to processing and analysing qualitative data computationally through techniques such as machine learning and data mining (see Chapter 6).

A taxonomy of kinds of data

captured, exhaust, transient, derived

source

how data are generated

There are two primary ways in which data can be generated. The first is that data can be *captured* directly through some form of measurement such as observation, surveys, lab and field experiments, record keeping (e.g., filling out forms or writing a diary), cameras, scanners and sensors. In these cases, data are usually the deliberate product of measurement; that is, the intention was to generate useful data. In contrast, *exhaust data* are inherently produced by a device or system, but are a by-product of the main function rather than the primary output (Manyika et al. 2011). For example, an electronic checkout till is designed to total the goods being purchased and to process payment, but it also produces data that can be used to monitor stock, worker performance and customer purchasing. Many software-enabled systems produce such exhaust data, much of which have become valuable sources of information. In other cases, exhaust data are *transient* in nature; that is, they are never examined or processed and are simply discarded, either because they are too voluminous or unstructured in nature, or costly to process and store, or there is a lack of techniques to derive value from them, or they are of little strategic or tactical use (Zikopoulos et al. 2012; Franks 2012). For example, Manyika et al. (2011: 3) report that 'health care providers... discard 90 percent of the data that they generate (e.g., almost all real-time video feeds created during surgery)'.

A taxonomy of kinds of data

structured/semi-structured/unstructured

structure

captured, exhaust, transient, derived

source

primary, secondary, tertiary

producer

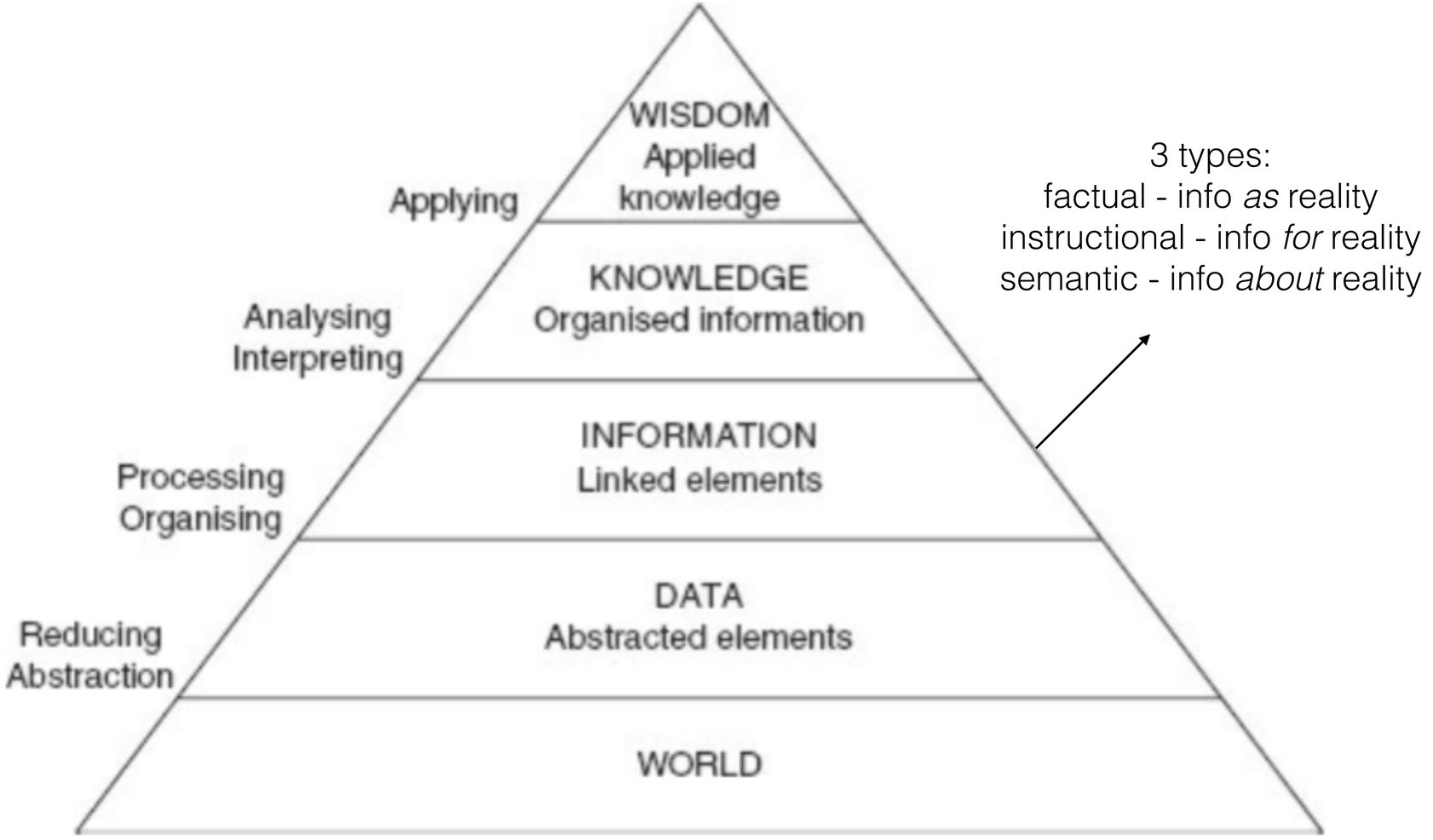
quantitative/qualitative

form

indexical, attribute, metadata

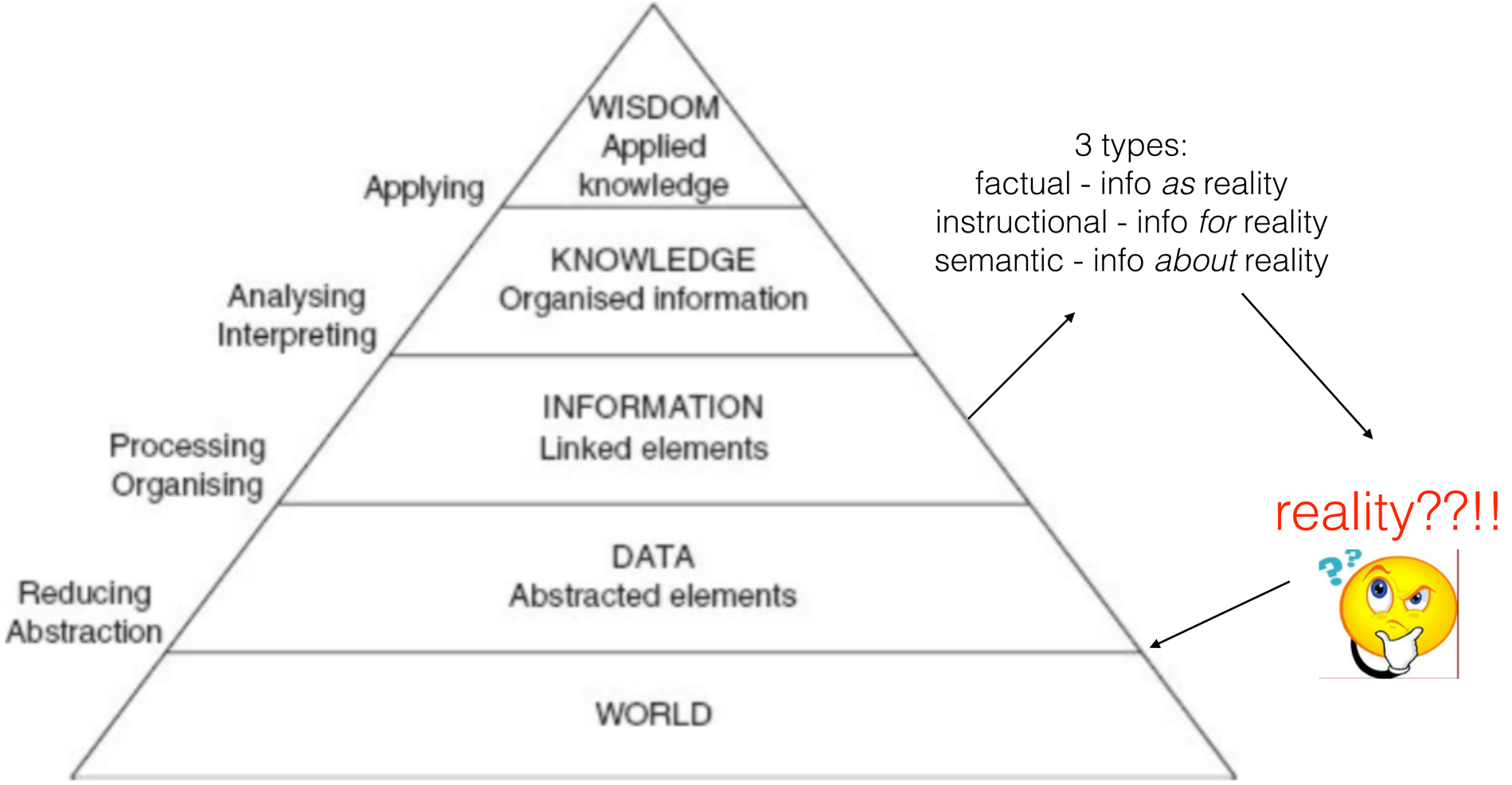
type

“What unites these various kinds of data is that they form the base or bedrock of a knowledge pyramid: data precedes knowledge, which precedes understanding and wisdom (Adler 1986; Weinberger 2011). (42)



3 types:
factual - info *as* reality
instructional - info *for* reality
semantic - info *about* reality

“What unites these various kinds of data is that they form the base or bedrock of a knowledge pyramid: data precedes knowledge, which precedes understanding and wisdom (Adler 1986; Weinberger 2011). (42)



NOT so much the concept …
but the ***value*** 

Regardless of how it is conceived, Floridi (2010) notes that given that information adds meaning to data, it gains currency as a commodity. It is, however, a particular kind of commodity, possessing three main properties (which data also share):

Non-rivalrous: more than one entity can possess the same information (unlike material goods)

Non-excludable: it is easily shared and it takes effort to seek to limit such sharing (such as enforcing intellectual property rights agreements or inserting pay walls)

Zero marginal cost: once information is available, the cost of reproduction is often negligible.

While holding the properties of being non-rivalrous and non-excludable, because information is valuable many entities seek to limit and control its circulation, thus increasing its value. Much of this value is added through the processes enacted in the information life cycle (Floridi 2010):

ethically

politically and economically

Framing data

temporally and spatially

technically

philosophically

Framing data

ethically

the debate on the ethics of what data are generated and the means of production, how data are shared, traded, protected, and employed. (how individuals are ‘sorted’)

politically and economically

how are data normatively conceived (within the context of nation-states), how regulated, and funding/investing, the unfolding of capitalism, traded as commodity

temporally and spatially

data and data assemblages have histories and geographies. how data are processed/analysed mutates over time and place

technically

the extend that methods of capture/measurement generate certain, clean, accurate data

philosophically

the notion that a ‘sensor produces an objective, realist worldview of things as they are’ vs. ‘data already cooked’ view.

Thinking critically about databases and data infrastructures

As with conceptualising data more generally, it is then important to think critically about the nature of databases and data infrastructures, their sociotechnical production, and how they reflect rationalities about the world at the same time as they reproduce and reinforce such rationalities. Such critical reflection has been largely absent with respect to big data, open data, and the scaling of small data, with the focus to date being more technical and instrumental in nature.

Thinking critically about databases and data infrastructures

Data infrastructures || The Archive

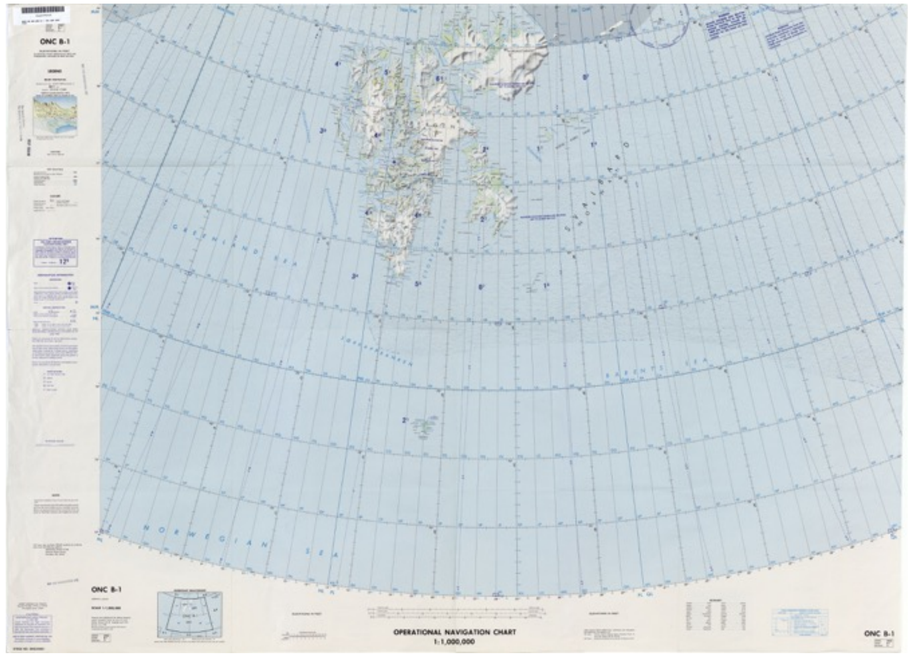
Derrida Deconstruction

> Databases create silences, adding to the inherent gaps in the data, as much as they reveal relationships between data and enable some questions to be answered; they constrain and facilitate through their ontology, producing various presences and absences of relations (Ruppert 2012; Vis 2013). Drawing on Derrida, Bowker (2005: 12) thus argues that databases and archives are jussive: they 'tell us what we can and cannot say' by defining what is remembered and what is ignored and forgotten.

Thinking critically about databases and data infrastructures

Data infrastructures || The Immutable Mobile

Latour



In other words, databases and data infrastructures do not simply support research, they fundamentally change the practices and organisation of research – the questions asked, how they are asked, how they are answered, how the answers are deployed, who is conducting the research and how they operate as researchers (see Chapter 8). For example, in her study of the evolution of the Canadian Census and the Atlas of Canada, Lauriault (2012) details how each has developed recursively and iteratively based on models of the world which construct ways to imagine and produce Canada. She argues that the data archives and the data themselves constitute an institutional 'extrasomatic memory system that allows for the telling of stories about the nature of Canada... [through] maps, graphs, models and statistics which rely on sensors, data, interoperability and web mapping standards, portals, metadata and models, science, and open architectures' (p. 27). In turn, these stories modulate the underlying models and thus the data infrastructure mutates, inflecting the means through which the stories are created.

NOT so much the concept …
but the ***value*** 

Is art data?

TROPICAL MINING STATION

A spatial product that extracts surplus value from the process of cryptocurrency mining to produce space.

October 2-4
11:00-17:00

Future Firm
49 S. Morgan St. #1
Chicago, IL 60608

Tropical Mining Station is a Partner Event to the Chicago Architecture Biennial.

http://www.investopedia.com/terms/b/blockchain.asp

https://docs.google.com/document/d/1ndBAcD72sHJnH0GpWY0pe8_KhkSZHFNTqBAHGBmY9N8/edit#

Art project: asking about mining cryptocurrency as a way to produce space

Tech/investment project: Using blockchain to authenticate art

http://techcrunch.com/2015/09/27/using-the-blockchain-to-the-fight-crime-and-save-lives/?utm_source=I%20am%20Dinesh%20Agarwal&utm_medium=I%20Love%20Your%20Content%20and%20shared%20it%20with%20my%20audience&utm_campaign=Follow%20Dinesh%20on%20Twitter%20dinwal