Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **1**, Number **4**, October **2016**

doi:10.3934/bdia.2016016

pp. **377–389**

DISENTANGLING DATA, INFORMATION AND KNOWLEDGE

Subrata Dasgupta

Computer Science Trust Fund Endowed Eminent Scholar
School of Computing & Informatics, University of Louisiana at Lafayette
Lafayette, Louisiana 70504, USA

(Communicated by Aijun An)

Abstract. Information, data and knowledge constitute the fundamental 'stuff' of computing and one might assume that in the seven decades since the advent of the modern computer theorists and practitioners of computing can differentiate between the concepts they denote. And, of course, computer scientists do not have exclusive claims over these terms or concepts: sociologists, cultural scholars, economists, historians, natural scientists, philosophers, and the managerial class have them as part of their vocabularies. The surprising fact is that these terms and the concepts they denote are far from distinct. They form a tangled web. In this essay I address the question: *what is the relationship between data, information and knowledge?* I attempt to disentangle – and clarify – how these terms are in fact interpreted by practitioners in such diverse disciplines as information science, historical research, empirical sciences, cognitive science, data mining and computer programming and to identify what appears to be a common thread.

1. **T.S. Eliot's Question, paraphrased.** Information, data, and knowledge constitute the fundamental 'stuff' of computing and one might assume that theorists and practitioners of computing can differentiate between the concepts they denote. As we will see, this would be a mistaken assumption.

Of course, computer scientists and other practitioners of computing do not have an exclusive claim over these terms or concepts. 'Information' and 'data' are also part of the vocabulary of sociologists and cultural scholars, economists, historians, natural scientists and the managerial class. As for 'knowledge', its nature, how it comes into being, and how we acquire it, have engaged thinkers since antiquity. Its investigation has a name all of its own: 'epistemology'. Data, information and knowledge are, thus, ubiquitous across disciplines and practices. Yet what we find is these terms are far from distinct; the corresponding concepts form a tangled web.

Are these three keywords to be understood differently? If so, how? Even a poet was drawn into this matter. In his poem *The Rock* (1934), the American poet T.S. Eliot (1888–1965) famously has the Chorus ask:

- Where is the wisdom we have lost in knowledge?
- Where is the knowledge we have lost in information?[1]

2010 *Mathematics Subject Classification.* 68P01.

Key words and phrases. Information, data, knowledge, science, computer programming, cognitive science, philosophy, symbol structures.

[1] T.S. Eliot, 1954. *Selected Poems.* London: Faber & Faber.

Being a poet, perhaps 'data' did not concern Eliot, but 'wisdom' did. At any rate he was clearly implying a *hierarchy of quality*: wisdom is superior to knowledge, and knowledge is superior to information.

Three quarters of a century later variations on this question still intrigue us, and never more so than in the context of computing. Paraphrasing Eliot's question, we may ask, in the context of computer science: *What is the relationship between the concepts of data, information and knowledge?* In honor of T.S. Eliot's prescience (his wisdom?) let me call this 'Eliot's Question, Paraphrased'.

2. **What information scientists think.** To add to the confusion surrounding the terms 'data', 'information' and 'knowledge', there is a domain of computing known broadly as 'information systems'. The issues of interest in this domain is the design, implementation, organization, behavior, and management of information in the context of such user communities as those devoted to health care, business and finance, management, libraries, and security. The generic term *information science(s)* is usually used to refer to the theory and practice of information systems, and though linked with computer science (as in the oft-used phrase 'computer and information sciences'), a careful distinction is normally made between the two. (In the United States, the word informatics is increasingly used to refer to the study and practice of information systems by itself or as in 'health informatics' and 'bioinformatics'. However, in Europe, the cognate terms *informatique* and *informatik* are used to mean computer science.)

Thus, analogous to computer scientists, there are *information scientists* whose business is not so much information processing or processors but information systems. And not surprisingly, information scientists have taken a keen interest in Eliot's Question, some even including 'wisdom' in the mix.[2]

Thus, Russell Ackoff, an American systems and management scientist, envisoned a pyramidal structure with data at the base, information sitting atop it, knowledge resting on information, and wisdom at the peak.[3] The pyramidal structure implies not only a hierarchy but also abstraction. That is, information is (in some sense) abstracted from data, knowledge is (in some sense) abstracted from information, and wisdom is (in some sense) abstracted from knowledge.

For Ackoff, data is the outcome of observation and, thus, represent "objects, events, and their environments".[4] As for information, Ackoff imagined someone asking certain questions of data, which is "processed" to afford answers, and this latter is information. Thus, in Ackoff's scheme, information is the outcome of *data processing* (in some fashion).

[2] See, e.g., M. Zeleny, 1987. "Management Support System: Towards Integrated Knowledge Management", Human Systems Management, 7,1, pp 59–70; R.L. Ackoff, 1989. "From Data to Wisdom", Journal of Applied Systems Analysis, 16, pp 3–9; L. Floridi, 2004. "Information", pp 40–62 in L. Floridi (ed.), Philosophy of Computing and Information. Oxford: Blackwell; J. Rowley, 2007. "The Wisdom Hierarchy: Representation of the DIKW Hierarchy", Journal of Information Science, 33 (2), pp 163–180; M. Frické, 2007. "The Knowledge Pyramid: A Critique of the DIKW Hierarchy", Journal of Information Science, 35, PP 131–142; C. Zins, 2007. "Conceptual Approaches for Defining Data, Information and Knowledge", Journal of the American Society for Information Science and Technology, 58, 4, pp 487–493; L. Floridi, 2010. Information: A Very Short Introduction. Oxford: Oxford University Press; T.D. Robertson, 2013. "The Data/Information/Knowledge/Wisdom Hierarchy Goes to Seminary", Advances in the Study of Information and Religion, 3: http://digitalcommons.kent.edu/asir/vol3/iss1/7. Retrieved April 4, 2014.

[3] Ackoff, op cit.

[4] Ackoff, op cit: p. 3.

As we have noted, the terms 'data', 'information' and 'knowledge', being in common use, mean all things to all people, even to 'experts'. Chaim Zins, an Israeli information scientist, surveyed and documented the views of forty seven "leading" scholars from sixteen different countries, and elicited some 130 definitions covering these terms. A broad but not consensual view according to the survey was that data constitutes the "raw material" for information, and information likewise for knowledge.[5] We see here, again, the presence of hierarchy and abstraction. Another widely shared view was that knowledge is *subjective*: it is not only the synthesized outcome of a "knowing" person's mind, it *resides* only in that person's mind.[6] This suggests (though Zins does not explicate this point) that a concomitantly held view is that information is objective — that it exists 'outside the mind'.

Unlike his surveyed subject, Zins himself makes an interesting distinction between the subjective and objective realms: he allows for each of data, information and knowledge to exist *both* subjectively and objectively, leading to six possibilities (Zins prefers the term 'universal' to objective but here we will use the latter term.)

[a] *Data*: In the subjective realm, they are "sensory stimuli which we pursue through our senses". In the objective realm, they are "sets of signs that represent empirical stimuli or perceptions".[7]

[b] *Information*: In the subjective realm it is "*the meaning of* ... [the] sense stimuli [the data]" and constitutes "empirical knowledge". In the objective realm, it is "a set of *signs* which *represents* empirical knowledge".[8]

[c] *Knowledge*: Subjectively, it is a thought in one's mind which the thinker *justifiably believes* to be true. Objectively, knowledge is "a set of *signs* that *represents* one's subjective knowledge."[9]

Notice that in Zins' view there is no such thing as non-empirical information. A purely formal definition such as is common in mathematics (e.g., 'a straight line is the shortest distance between two points', or 'zero is a number') would not constitute information, either subjective or objective. This leaves non-empirical statements in limbo in Zins' scheme. Moreover, his distinction between subjective and objective is the distinction between entities perceived, understood or thought in an individual's mind and their (symbolic) *representation* outside the mind. If I have understood Zins correctly, there is no objectivity outside representation by signs (symbols). Thus, the trees, grass, fence, and squirrels in my back yard I am observing through the window of my study as I write these words, are not objective; only their symbolic representations are (by words; by images?).

For Martin Krické, data is whatever is recordable such that an interpretation of the recording yields a true statement.[10] So, if the interpretation is correct there is no possibility for 'false data'. Moreover, he holds that data can be more than observables. One can imagine a situation wherein a mathematical proposition (such as, 'there are four prime numbers between 10 and 20) which is mathematically factual but not observable might serve as data.

Krické pays careful attention to the diverse meanings of 'information', ranging from everyday use (often conflating it with 'knowledge') to technical definitions in

[5] Zins, op cit: p. 479.

[6] Ibid.

[7] Zins, op cit: p. 487. Italics added.

[8] Ibid. Italics added.

[9] Ibid. Italics added.

[10] Krické. Op cit.

such fields as communication engineering. Ultimately, his distinction between information (as useful in information science) and knowledge is the distinction between what he terms *weak knowledge* and *strong knowledge.* By strong knowledge Krické means personal (individual) and socially shared *beliefs that are justified as true.* Weak knowledge is almost like strong knowledge but not quite: it constitutes of beliefs that are true but not justified. For Krické, *information is weak knowledge.*

Krické's condition for information-hood might be exemplified as follows: consider the statement 'in a right-angled triangle with sides a, b forming the right angle and c the hypotenuse, $c^2 = a^2 + b^2$.' This is, of course, Pythagoras' theorem, and it happens to be a true statement in plane geometry. If a person believes this statement (having read it or been told it by someone), then *for that person*, this is weak knowledge, hence information. On the other hand if a person (say a high school mathematics student) can also justify (prove) the statement this would constitute (strong) knowledge for that person. Thus one person's knowledge is another‘s information.

For me, as a lay person in the matter of climate science, the statement 'global climate is warming in part because of the emission from carbon fuels' is weak knowledge, that is, information: I believe it is true (because of the experts stating it so) though I personally cannot justify it. For climatologists, however, this statement would constitute (strong) knowledge.

3. **What philosophers believe.** As it happens, epistemologists, though not interested particularly in information (let alone data), have concerned themselves with the distinction between knowledge and belief.[11] One philosopher, Norman Malcolm, has explored the distinction between weak and strong knowledge, but not in Krické's sense.

For Malcolm, to know in the strong sense is "when a person's statement 'I know that p is true' implies that the person making the statement would look upon nothing whatever as evidence that p is false."[12] The issue of justification does not arise here. If I say "The world was created in six days," this would be strong knowing on my part if I discount all geological evidence to the contrary. On the other hand, one knows something in the weak sense when one does not exclude the possibility that it could be proved wrong. Thus, my knowing that 'Pluto is a planet' would constitute weak knowing if I'm prepared to accept evidence that Pluto is not a planet. (This suggests that *all* empirical knowledge – for example, natural and artificial science knowledge – is almost certainly weak in Malcolm's sense; and even parts of mathematical (that is, non-empiricial) knowledge may be weak since it may well be that the reasoning underlying such knowledge may be found to be flawed.[13]) By this view, knowing in the strong sense is simply dogmatic belief.

We turn to another philosopher, Luciano Floridi, a philosopher of information, for a more formal treatment of the concept of information and its relation to data.[14] For

[11] The literature on epistemology is vast and goes back centuries. For a glimpse of the issues involved, see, e.g., B. Russell, 1948. *Human Knowledge: Its Scope and Limits.* New York: Simon & Schuster; A.J. Ayer, 1956. *The Problem of Knowledge.* Harmondsworth: Penguin Books; A.P. Griffith (ed.), 1967. *Knowledge and Belief.* London: Oxford University Press; M. Polyani, 1962. *Personal Knowledge.* Chicago: University of Chicago Press; K.R. Popper, 1972. *Objective Knowledge.* Oxford: Clarendon Press.

[12] N. Malcolm [1952] 1967. "Knowledge and Belief," pp 69–81 in Griffith, op cit: p. 72.

[13] See, e.g., J. Avigad & J. Harrison, 2014. "Formally Verified Mathematics," Communications of the ACM, 57, 4, pp 66–75.

[14] Floridi 2004, op cit: Floridi 2010, op cit.

Floridi, as for Ackoff and Zins, data precedes information in the sense that without data there can be no information. Data exists, according to Floridi, only when there is an *absence of uniformity* between two states of a system. More formally, a *datum* exists whenever there are two *uninterpreted* variables, x, y such that $x \neq y$.[15]

Thus data intrinsically has no meaning; it merely signifies the presence of a difference. In the case of a traffic light, for example, the presence of a red signal is a datum because it could have been otherwise: yellow or green. But a (faulty) traffic light that shows none of these colors is also a datum because of its difference from red, yellow and green – just as silence can be data as in the Sherlock Holmes story in which the dog does not bark in the night because this condition differs from the possibility of the dog that does bark in the night.

The only condition that is genuinely *dataless* is by "the elimination of all possible differences."[16] Consider, for example, a neighborhood in a city, or even an entire city, wherein traffic lights do not work at all, and have not worked in years. This would be a dataless condition according to Floridi, in that there is total absence of difference. The traffic light is *always* black.

Given this definition of data, Floridi defines information as follows: Δ is an instance of information if and only if : (a) Δ consists of one or more data elements; (b) The data elements are *well formed*; (c) The well-formed data elements are *meaningful*.

Condition (a) stipulates that information is composed of data; condition (b) specifies that the data elements follow some rules of syntax; condition (c) stipulates that the data elements together must possess semantics. Consider, e.g., linguist Noam Chomsky's famous example "colorless green ideas sleep furiously," which though syntactically correct (is well-formed) is meaningless.[17] As per Floridi's conditions, this statement is not information, whereas the sentence "the man hit the dog" is.

As for *knowledge*, Floridi has little to say about it: he notes that it bears a 'family resemblance' to information in that they are both meaningful entities; however, they differ in that while information elements are singular, isolated entities, knowledge constitutes a relationship between information items so that one can make sense of one piece of information by its relation to others.[18]

The view thus far is clearly quite unsatisfactory: no evidence of consensus or even a broad agreement about the fundamental nature of the concepts data, information and knowledge. Eliot's Question, Paraphrased remains unanswered. The situation is rather like that of biologists in the 19th century who agreed that the 'stuff' of biology was something called 'life' but they could not agree what constituted life and what demarcated life from nonlife.[19] Eventually, the biologists realized that one cannot identify life by a single criterion; rather, living entities manifest a number of characteristics, and it is these, collectively, that 'define' life.

4. **From data to knowledge: An empirical scientist at work.** Computing practitioners, information specialists, and philosophers are, of course, not the only

[15] Floridi 2004, op cit: p. 43.

[16] Floridi 2010, op cit: p. 23.

[17] N. Chomsky, 1957. *Syntactic Structures*. The Hague: Mouton: p. 15.

[18] Floridi 2010, op cit: p. 51.

[19] E. Mayr, 1982. *The Growth of Biological Thought*. Belknap Press of Harvard University Press: pp 51-53.

communities occupied with the issues of data, information and knowledge. Academic scholars of different stripes are in the business of gathering such 'stuff', manipulating it, making inferences from it, and so on. It might be useful to see what some of them have to say on this matter.

"You see but you do not observe." So Sherlock Holmes admonished his friend and amanuensis Dr. Watson in the first of the Holmes short stories, *A Scandal in Bohemia* (1891). What Holmes was saying was that there is something particular to *observing*: seeing is necessary but not sufficient for observing.

The empirical scientist, in particular, thrives on observation, either 'in the wild' or in the laboratory, through naked eyes or via instruments. What does the scientist observe and what does he *do* with his observations?

The physiologist and creativity researcher Robert Root-Bernstein in his remarkable book *Discovering* (1989) written in the form of imaginary conversations between several young fictional thinkers, scientific and otherwise – has one of his characters say the following:

First thing students see in X-ray diffraction photographs is the large black (or white) spot at the center, yet it is totally meaningless, since it's where the X-ray beam passes through the crystal without being diffracted. Everything they see is data, of course, but the trick is to figure out what part of it is *meaningful* data.[20]

So Root-Bernstein seems to believe that data is 'out there' to be seen, and the scientist's observation is preceded by this data. But amidst that data is whatever is *meaningful* to the observer. The observer *interprets and thereby confers* semantic content to the data. In Floridi's terms it becomes information.

Let us follow one young cognitive scientist's thinking about the relationship between seeing, observing, and interpreting, and how she connected these activities to data, information and knowledge in actually *practicing* science.

Our investigator, Carley Faughan, was concerned with executing her doctoral research in primate cognition, and so she was using such terms as 'data' and 'observation' *unselfconsciously.* It is this unselfconscious use of the terms in her dissertation that is of interest here.[21]

This researcher observes, by way of experiments in a laboratory setting, her chimpanzee subjects attempting to use a stick with a hook at one end to reach through bars and access a piece of food that was beyond the animal's arm reach. So what the investigator is doing is observing the behavior of her chimpanzee subjects. She tells us in her dissertation that the *aim* of these experimentally-controlled observations is to *gather data in order to resolve certain questions concerning* the nature of chimpanzee intelligence.

She clearly views that the outcome of her observations is *data*: "I will present data ...";[22] "I collected data ...".[23] However, she also tells us that she "was able to recover each occurrence of the behavior, as well as *information* about the time of its occurrence."[24] It is not clear, though, whether she uses the word 'information' as a synonym for data.

[20] R. Root-Bernstein, 1989. *Discovering.* Cambridge, MA: Harvard University Press: p. 92.

[21] C.E. Faughan, 2014. "Social and Physical Cognition in Chimpanzees (Pan tryglodytes): Preliminary Investigation of Domain-General versus Domain-Specific Intelligence." Ph.D Dissertation, University of Louisiana at Lafayette.

[22] Faughan, op cit: p. 30.

[23] Faughan, op cit: p. 37.

[24] Ibid. Italics added.

The data she gathers based on the observations and presented in various tables, charts and graphs were time measures of certain social activities and behaviors of the chimpanzees (in play or grooming or in contact with other chimpanzees). In addition, there were scores on a particular standard 'social responsiveness test'.

What is noteworthy is that this data gathering was a *selective* process. The data was collected in the context of an experiment the investigator had designed in response to her *research goal* – to elicit an answer to a particular question concerning primate cognition. Data did not precede observation, *pace* Root-Bernstein; data *is* observation. Faughan must have seen much that was going on in her chimpanzee laboratory, but what she observed *was* data. The date she gathered was, thus, *intrinsically meaningful*. They afforded certain kinds of *information* about chimpanzee behavior. Here, I use the word 'information' to mean 'something told' about something, an (apparantly true) assertion about something in the world. Thus, the measurement of a particular chimpanzee's time spent in grooming (this was gathered or observed data) *becomes* a piece of information about that animal's grooming behavior. In this sense *data and information are identical*.

But then our researcher goes into an activity she calls 'data analysis':[25] *interpreting* the data she has collected. But, *pace* Floridi's model, her data is not meaningless: she *gathered* it as relevant to her research goal. So, in interpreting her (already meaningful) data, she is seeking *more general* meaning in the data. If her data is already information, her 'data analysis' as interpretation is intended to construct 'higher-level' or more general information about her subjects' behaviors. Moreover, interpretation is always contextual. How one interprets is determined by the circumstances that drive the interpretation: goals and what the interpreter already *knows* about her subject matter. In the case of this researcher this meant that her data analysis was as much driven by the specific goal of her research as was data gathering. The results of her interpretation included *categorization* of the data in specific ways, *finding relationships* between various behavioral and social data, and *computing* certain features of her chimpanzee collective.

Faughan goes on to discuss her fundamental discoveries. These include general (statistical) assertions, including her own answer to her primary overarching questions, for example, whether there was any evidence of domain-general intelligence amongst her subject population.

As an experimental scientist, Faughan clearly took data as meaningful observation. This meaningfulness was determined by her research goals. *Her observations (that is, data) were goal-driven.* This coheres well with philosopher of science Karl Popper's dictum that *all observation is theory-driven*: One cannot observe, Popper claimed, unless one has a 'theory' or a set of expectations to guide what one observes.[26] The data she gathered and presented in tables *said* something about her chimpanzee subjects' behavior.

Of the various theoretical propositions on data and information advanced by the information scientists and discussed in the preceding sections, *none cohered with the assumptions of data as meaningful, goal-driven and selective observations* we find in this working scientist's actual practice. There are 'family resemblances' with one or the other propositions in one respect or another, but neither Ackoff nor Zins nor Krické nor Floridi offered a satisfactory theoretical basis for our observation of this

[25] Faughan, op cit: p. 41.

[26] K.R. Popper, 1965. *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Harper & Row: p. 46.

particular scientist's practice.

The very idea of data as the outcome of goal-driven observations is what Sherlock Holmes probably meant when he admonished his friend of seeing but not observing. Artificial intelligence (AI) researchers Jeffrey Shrager and Pat Langley, who have investigated computational approaches to scientific discovery also coupled observations with context. For them, data does not result *from* observation; rather (as we saw in the case of Carley Faughan), observation *is* data, and such observation-data represent recordings of the "environmental setting" in which the observer (observing directly[27] or through "sensors or measuring instruments"[28]) "must select some aspects to record and some to ignore".[29]

According to Floridi, meaningful data is information. The tables of observed data Faughan constructed are, thus, *also* information, since they are assertions about chimpanzee behavior. We might label them 'low-level information'. Thus data and information become synonymous.

In contrast, the outcome of Faughan's data analysis - presented as more tables and bar charts - denotes her chimps' *patterns of behavior* grounded in the data. Each of these patterns *taken in isolation* is an assertion or declaration about some general feature of her subjects' behavior. They are generalizations she believes are true (in some sense). A lay reader of her dissertation can assimilate these assertions as tokens of *information.* But the lay reader is (normally) not in a position to draw broad generalizations, or to relate and connect these tokens of information to other (known) aspects of chimpanzee behavior. Faughan, on the other hand, could *integrate* these individual pieces of information and *connect* them with other aspects of chimpanzee behavior, and with the observed data gathered and published by other primate scientists; and *infer* and *predict* some conclusions.

In other words, Faughan could integrate these tokens of information into an *existing* structure of theories, data, facts, hypotheses, laws, models, and so on. *This* is what constitutes *knowledge.* (This coheres with Floridi's concept of knowledge). What is merely information to the lay reader becomes part of the knowledge possessed by the primate cognition community.

5. **Data, knowledge and facts in the historian's world.** Historians, rather more than scientists, like to ruminate upon their craft. They reflect on the nature of history (its ontology), the techniques and methods of its investigation (its methodology) and the nature of historical knowledge (its epistemology). In this endeavor, they are joined by philosophers of history who bring their own critical tradition to the table.[30] Amongst their many concerns are the role and nature of *data, knowledge* and *facts.*

[27] J. Shrager & P. Langley, 1990. "Computational Approaches to Scientific Discovery," pp 1–26 in J. Shrager & P. Langley (ed.), *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, CA: Morgan Kaufmann: p.6.

[28] Shrager & Langley, op cit: p. 3.

[29] Shrager & Langley, op cit: 6.

[30] The literature on the nature of the historian's enterprise certainly reach back to the 19th century if not earlier. Here are just a few references to both classic and modern writings on the subject. H. Butterfield [1931] 1973. *The Whig Interpretation of History.* Harmondsworth: Penguin Books; R.G. Collingwood, 1946. *The Idea of History.* Oxford: Clarendon Press; M. Bloch, 1953. *The Historian's Craft.* New York: Vintage Books; E.H. Carr [1961] 1964. *What is History?* Harmondsworth: Penguin Books; G.R. Elton [1967] 1984. *The Practice of History.* London: Fontana; F. Braudel, 1980. *On History.* Chicago: University of Chicago Press; J. Appleby, L. Hunt & M. Jacob, 1994. *Telling the Truth About History.* New York: W.W. Norton; R.J. Evans, 2000. *In Defence of History.* London: Granta Books; J. L. Gaddis, 2002. *The Landscape of*

Here, I will only present in summary my interpretation of some of the salient views of a few influential historians.

In the historian's world, there are the following entities:

(a) *Events* that actually occurred in the past.

(b) A *record* of these events in the archives – meaning broadly, the document archives and archeological traces.

(c) The historian's *knowledge* of the past – as, say, a network of facts, theories, interpretations, causal relationships, beliefs, etc. about the past.

(d) *Questions* posed by the historian about the events that occurred in the past.

Historians seem to vary somewhat in their understanding of how these elements 'fit' together in historical investigations. For the British historian and philosopher R. G. Collingwood, historical data corresponded to elements of (b) but selected by the historian according to (c) and (d).[31] For another British historian E.H. Carr, the only events of the past of interest to historians are 'historical facts', those elements selected from the archives (b) in particular context (c and d), and these historical facts constitute his data.[32] For the German-British historian G. R. Elton, historical data (evidence) are elements of (b) that combine with (c) and (d) to elicit the facts of the past (a) that will, in turn, modify the historian's knowledge.[33] For the contemporary British historian Richard Evans, events of the past (a) are established as information about the past by virtue of the data (b), knowledge (c) and questions posed (d).[34] Such information, in turn, is also data that enter into the production of new information or knowledge. In the view of contemporary American historians Joyce Appleby, Lynn Hurst and Margaret Jacob, the archival material (b) becomes data relative to the context. (c and d)[35]

The common ground amongst these views is twofold:

[1] The historian brings to her task a context that is a blend of her personal historical *knowledge*, and the *questions* she asks.

[2] The historian's *data* which enter into her reasoning are some subset of the archived material selected by way of the abovementioned context. (This can also be compared to the scientist's data elicited by observation directed by the scientist's goals; see section 4 above.)

As for *information*, this is not really a preferred concept in the historian's world. However, there are two ways in which the historian's entities seem to cohere most closely with Luciano Floridi's definition of information: (a) as the records of past events in the archives; and (b) the actual facts (events) of history.

6. **What cognitive scientists say about information and knowledge.** We have already encountered a young cognitive scientist in the practice of her craft (section 4). But more generally, cognitive scientists seem to hover uncertainly between information and knowledge when it comes to theorizing about their discipline.

In his comprehensive history of cognitive science, *The Mind's New Science* (1985), psychologist Howard Gardner defines cognitive science as an empirical discipline that strives to "explain human knowledge" – its nature, structure, development

History. Oxford: Oxford University Press; D. Cannadine (ed.), 2002. *What is History Now?* Basingstoke: Palgrave Macmillan.

[31] Collingwood, op cit: p. 243.

[32] Carr, op cit: pp 9-12.

[33] Elton, op cit: pp 76–81.

[34] Evans, op cit: p. 76.

[35] Appleby, Hunt & Jacob, op cit: pp 26, 73, 262.

and how it is employed.[36] On the other hand, two of the begetters of the 'cognitive revolution' – which forged the discipline of cognitive science out of such traditionally disparate fields as psychology, computer science, linguistics, anthropology, and philosophy – Allen Newell and Herbert Simon (along with collaborator Cliff Shaw) – described, in a seminal paper published in 1958, an "information processing model" of human problem solving.[37] Some psychologists have espoused the 'information processing paradigm' as central to cognitive psychology.[38] For example, John R. Anderson stipulated that :

"Cognitive psychology is dominated by the *information processing approach* which analyzes cognitive processes into a sequence of ordered stages. Each stage reflects an important step in the processing of cognitive information."[39]

Likewise Newell and Simon stated that "information processing is the leading contemporary point of view in cognitive psychology".[40] Yet not only Gardner but cognitive scientists in general broadly construe cognition as a *knowledge-based syste*m that effects thinking, understanding, perceiving, visualizing and, more generally, making meaning of our experiences. A significant subject matter in cognitive science is knowledge *representation* – that is, how the mind/brain complex organizes knowledge.[41]

So how do cognitive scientists distinguish, if at all, between information and knowledge?

For Paul Rosenbloom, a cognitive and computer scientist, information is "any content expressed in some medium ... that *resolve uncertainty*".[42] Rosenbloom's 'information' is sweeping in its embrace, including:

"... numeric values and measurements ... GPS coordinates ... strings of characters fragments of texts ... audio and/or video files ... knowledge about the world ... modeling how things work ... programs ..."[43]

This eclectic view does not insist that information has to be (perhaps approximately) true, for information can be "about our world ... about imaginary or virtual worlds, or about no world at all."[44] It also encompasses "knowledge about the world such as the generalization that *all men are equal.*"[45] So in Rosenbloom's view knowledge is a *kind* of information, in contrast, say, to Floridi's view that knowledge *comprises of* information.

A textbook on cognitive psychology published in 1979 mixed 'information' and 'knowledge' in such a manner that the reader is left the impression that the authors use the term synonymously:

[36] H. Gardner, 1987. *The Mind's New Science.* New York: Basic Books: p. 6.

[37] A. Newell, C.J. Shaw & H.A. Simon, 1958. "Elements of a Theory of Human Problem Solving," *Psychological Reviews*, 65, pp 151–166. See also the magisterial A. Newell & H.A. Simon, 1972. *Human Problem Solving.* Englewood-Cliffs, NJ: Prentice-Hall.

[38] R. Lachman, J.L. Lachman & E.C. Butterfield, 1979. Cognitive Psychology and Information Processing. Hillsdale, NJ: Lawrence Erlbaum Associates; J.R. Anderson, 1980. Cognitive Psychology and Its Implications. San Francisco: W.H. Freeman; J. L. Bermúdez, 2014. Cognitive Science (2nd ed.) Cambridge: Cambridge University Press: pp 19-23; esp. Part III, pp 138–269.

[39] Anderson, op cit: p. 3.

[40] A. Newell & H.A. Simon [1975] 1987. "Computer Science as Empirical Inquiry: Symbols and Search" pp 287–313 in Anon 1987. *ACM Turing Award Lectures. The First Twenty Years 1966-1985.* New York: ACM Press / Reading, MA: Addison-Wesley: p. 299.

[41] Anderson, op cit: pp 61 et seq; pp 223 et seq.

[42] P.S. Rosenbloom, 2013. On Computing. Cambridge, MA: MIT Press: p. 9. Italics added.

[43] Ibid.

[44] Ibid.

[45] Ibid.

"Cognitive psychologists within the information processing paradigm ... have defined the area of study as the way man collects, stores, modifies and interprets environmental information or information stored internally. They are interested in knowing how he adds information to his permanent knowledge of the world, how he accesses it again, and how he uses his knowledge in every facet of human activity."[46]

7. **'Mining' data and 'discovering' knowledge.** As if the data/information/knowledge complex is not befogged enough, the past three decades have witnessed the emergence of a specialty that straddles computer science, information science and statistics. It goes by the names *data mining* or *knowledge discovery in data bases* (depending on the background of the specialists involved), though there are those who distinguish between the two.

A recent textbook defines data mining as a process that discovers "interesting patterns and knowledge" from large volumes of data.[47] For these authors, J. Han, M. Kamber and J. Pei, knowledge is characterized as "interesting patterns" where, by 'interesting' they mean a regularity that is "easily understood by human beings", that is also "valid on new or test data with some degree of certainty", is "potentially useful", and "novel".[48] Thus the knowledge 'discovered' must not only be new, useful and comprehensible, it must also possess a degree of generality and predictive authenticity.

The authors sharply distinguish between data mining and *information retrieval* wherein input queries applied to a data base retrieve "useful information" from the data base. Such information, they assert, "do not reflect sophisticated patterns or regularities *buried* in the data base[49] – hence data mining. It is not clear, though, where the boundary lies between 'useful information' and 'sophisticated patterns'. Nor is it clear whether they distinguish between information and data.

An earlier publication, authored by U.M. Fayyad, G. Piatetcki-Shapiro and P. Smyth, discriminated between data mining and knowledge discovery. The latter, the authors stipulated, discovers "useful knowledge from data", while the former uses algorithms to extract "patterns from data", without the further knowledge discovery steps that that related the extracted pattern to "prior knowledge" or that interpret the pattern.[50]

So, here, knowledge is "useful information" extracted from the data.[51] For Fayyad et al data is "a set of facts".[52] And a pattern is an expression in a language describing facts that are a subset of the facts; but that pattern must be 'simpler' (in some sense) than the listing of the facts in the subset. By 'simpler', Fayyad *et al* are presumably meaning an abstraction.

A pattern, then, is not intrinsically knowledge. Fayyad *et al* identify an attribute they call "interestingness" which differs from the 'interestingness' of Han *et al.* The latter was correlated to human understandability, while for Fayyad *et*

[46] Lachman, Lachman & Butterfield, op cit: pp 6-7.

[47] J. Han, M. Kamber & J. Pei, 2012. *Data Mining: Concepts and Techniques.* Amsterdam: Elsevier / Waltham, MA: Morgan Kaufman: p. xxiii.

[48] Han, Kamber & Pei, op cit: p. 21.

[49] Han, Kamber & pei, op cit: p. 154. Italics added.

[50] U.M. Fayyad, G. Piatetsky-Shapiro & P. Smyth, 1996. "From Data Mining to Knowledge Discovery: An Overview," pp 1–34 in U. M. Fayyad, G. Piatetski-Shapiro, P. Smyth & R. Uthuruswamy (ed.), 1996. *Advances in Knowledge Discovery and Data Mining.* Menlo Park, CA: AAAI Press / Cambridge, MA: MIT Press: p. 3.

[51] Ibid.

[52] Fayyad, Piatetski-Shapiro & Smyth, op cit: p. 6.

al interestingness is a measure of "pattern value". A pattern becomes knowledge if, given a threshold value (specified by humans) the interestingness is greater than the threshold.[53]

So it seems that according to both Han *et al* and Fayyad *et al* patterns are information but only 'useful' or 'interesting' information is knowledge. There is more data than patterns (which are abstractions of data) and more patterns than knowledge (which abstracts from information). However, what makes patterns (i.e., information) 'interesting' and thus knowledge is a matter of personal judgement, hence arbitrary. The hierarchy of data from which patterns (information) is abstracted, and knowledge which is extracted from the information coheres with Ackoff's data-information-knowledge pyramidal hierarchy.

8. **A view from computer programming.** The time has come (as a character from *Alice in Wonderland* may well have said) to talk of computer scientists and computing practitioners. Keeping in mind that computer science is also regarded as the science of information processing, how do *computer scientists and practitioners actually* relate data to information, and information to knowledge. Consider, specifically, a view of this matter from the perspective of computer programming.

"There seems to be confusion between the words *information* and *data* ... When a scientist conducts an experiment in which some quantity is being measured we have four things present each of which is called "information": (a) the true value of the quantity; (b) the approximation to this true value that is actually obtained by the measuring device; (c) a representation of the measured value of (b); and (d) the concept learned by the scientists by a study of the measurement. The word "data" is *most appropriately* applied to (c), and the word "information" when used in a technical sense should be further qualified by stating *what kind of information* is meant."[54]

Thus wrote computer scientist Donald Knuth in 1966. Knuth, then, held that while information referred to the actual or approximate values of entities being observed in an experimental setting, and/or the concepts associated with such observation, data was the *representation* of such information. This, is a quintessentially computer scientist's perception, for it was the advent of computing that ushered representation into scientific consciousness.

Knuth was as much an artificer as a theorist: though not an industrial practitioner, he was as much a prolific creator of computational artifacts in the realm of large scale (commercial grade) programming, algorithms, and programming languages as he was an acute computer scientist. And so we must pay close attention to his identification of data with the representation of information. He was not the only computer scientist to do this. A major division of the computer science community which I will call (for want of a better term) 'programming theorists' came to concern themselves with data; specifically:

(a) The identification of the proper *types* of *data objects* to represent information about the world
(b) The rules for *manipulating and processing* data objects of certain types to compute new data objects of similar or different types.

[53] Fayyad, Piatetski-Shapiro & Smyth, op cit: P. 8.

[54] D.E. Knuth [1966] 1996. "Algorithms, Programs, and Computer Science," pp 1–3 in D.E. Knuth 1996. *Selected Papers in Computer Science.* Stanford, CA: Center for the Study of Language and Information: p. 1. Italics added.

From the perspective of a major class of computing practitioners (algorithm designers, programmers, software engineers, and programming language designers and implementers) and computer scientists (programming theorists, language theorists, software theorists) what *really* matters is not information *per se*, nor knowledge per se, but the nature, structure and types of data *objects* that can represent information or knowledge about the world in a form appropriate for both human understanding *and* automatic processing by computers; and how to operate on, and process, these data objects. From their perspective, computer science is not so much the science of information processing but rather the science of the automatic processing of *data objects that represent information.* In fact, the words 'information' and 'knowledge' occur rather sparsely in the literature on programming.[55]

9. **Data, information and knowledge are all symbol structures.** The problem, as we have seen, is the entanglement of the terms 'information', 'data' and 'knowledge' and the concepts they signify. All three terms are ubiquitous in computational discourse. The stuff of computing, in practice, seems not only information but also data and even knowledge. Taking liberty with T.S. Eliot's famous question, it seemed necessary to ask: What is the relationship between data, information and knowledge in the computational context?

To add to the confusion, we also saw that these terms are used in a range of other disciplines including information science (a 'cousin' of computer science), experimental sciences, philosophy, historical research, and cognitive psychology. We witnessed a diversity of interpretations across these enterprises. A tangled and elusive web of interpretations, in fact.

Ultimately, despite the manifold interpretations, information, data, and knowledge all *represent* things in the world; they are all *about* things in the world. They are, so to speak, different incarnations – *avatars* – of the same things. In the computational realm, data, information and knowledge are all *symbol structures*. Symbols are entities (signs, marks, physical states, etc.) that *represent* or *denote* other entities. Computational artifacts are, therefore, *symbol processing systems*. In some context we call these symbol structures information, in others, data, in still others we call them knowledge. For the present, as an answer to the Eliot Question, Paraphrased, we may want to follow Allen Newell and Herbert Simon who envisioned computer science as *the science of symbol processing.*[56]

E-mail address: subratadasgupta1044@gmail.com

[55] In an informal and quite random survey of some seventy five books in my library on programming methodology, programming languages, operating systems, and software engineering, the word 'information' collectively appeared no more than half-a-dozen entries in the indexes! 'Data' or some term that included 'data'; appeared multiple times in practically all the indexes.

[56] A. Newell & H.A. Simon [1975] 1987. *Computer Science as Empirical Inquiry: Symbols and Search*, pp 287–313 in Anon. 1987. *ACM Turing Award Lectures. The First Twenty Years 1966-1985.* New York: ACM Press / Reading, MA: Addison-Wesley; A. Newell 1980. "Physical Symbol Systems:, Cognitive Science," 4, pp 135–183; P.S. Rosenbloom, 2013. *On Computing.* Cambridge, MA: MIT Press: pp 9–10.