Defining Terms: Data, Information and Knowledge

Conference Paper · July 2016

DOI: 10.1109/SAI.2016.7555986

John David Sanders
Bluestone Enterprise Ltd , UK

SAI Computing Conference 2016
July 13-15, 2016 | London, UK

Defining Terms: Data, Information and Knowledge

John Sanders
Bluestone Enterprise Ltd
UK

Abstract—in normal conversation, the meanings of data, information and knowledge can often be used interchangeably. At some level of approximation, these three terms are near enough the same. In the domain of computing and in particular the topic known as artificial intelligence, to accept this generalisation is to lose some resolution in a plausible model of the origin of thought. This paper offers an interpretation of these terms that allows us to preserve a model of adaptive behaviour, which is derived using a basis consisting of progressive evolution acting on control-based systems that are matched to and interacting with, a persistent and consistent environment.

I. INTRODUCTION AND REVIEW OF PAPERS

Frequently definitions of the word "data" make an issue out of its plurality: data, plural of datum[1]. The actual definition tends to use limp phrases such as "facts or pieces of information". If one then asks: what is information? This usually involves words such as knowledge and data. The definitions for knowledge elicit words such as *facts and information* and so we complete an unproductive circle of definitions.

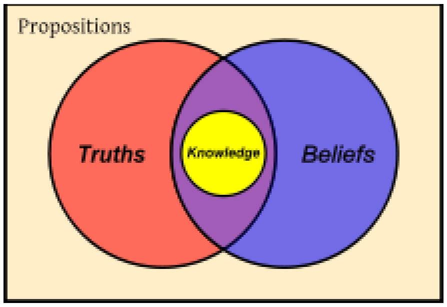


Fig. 1. Knowledge (source Wikipedia)

For the word *knowledge*, we can turn to the philosophers who use the term "epistemology". This involves belief and truth and even logic and proof (Fig. 1). Accepting this diagram at face value, perhaps we can insist that any description of the word knowledge should explain its relationship to belief[2]. Philosophers also identify a theory of knowledge known as Empiricism[1] in which sensory evidence and experimentation are in a process to determine potential knowledge. Empiricism is the basis of scientific reasoning and that often means collecting data, which in turn suggests that data becomes evidence and supports the creation of knowledge.

There are some papers, which attempt to address the nature of data. In general, they define data by way of example, and further, classify it. Intuitively, we all know what data is until we attempt to resolve it.

From paper [2] we can extract the following points:

- *"Data is not the same as information."*
- *"The so-called information overload is in fact a data overload."*
- *"Observation: Unorganized data is of little value."*
- *"Data is independent of a relationship (just numbers or words) until it is linked then it becomes knowledge."*
- *"Data often requires context to make knowledge."*

The phrase also from [2]: "*therefore, the availability of computers did not change the need for the human agent to interpret data into information and infer knowledge through the application of context.*"

In this paper there is a distinction between information and data, and the clue is the use of conceptualisation to model (human agent does the interpreting). To use data seems to require a model, which expects values from the data to be available in order to feed some sort of process. Often the result is successions of symbolic forms defining themselves by accessing previous symbolic forms and inevitably expressing results as symbols rather than addressing reality (a philosophical observation which gives rise to the question such as to what is real (Ontology [3])).

Whatever is *real*, results in either direct responses or models in which responses may need to be determined and then emitted. The use of prediction to establish plausible responses (planning) still cannot be resolved in isolation (i.e. without checking against reality). With the addition of consistency in predictions (inputs mapped to their anticipated responses) as determined in feedback from sensors), we promote our model from a conceptual state (believed) to a higher level of confidence. It can become regarded as fact or true, and hence, a basis for future models. The responses derived from this basis are knowledge [4, p. figure 1].

Knowledge is not the symbolic representation derived in a model from information received, but the emitted response, which, ambiguously, can also be spoken or written words. This is often the source of confusion between the two terms: Knowledge and information[3].

[1] In this paper "data" will be used as a collective noun and the phrase "the data is" will be used.

[2] Descriptions of information being "filtered" or processed to produce knowledge tend to miss this important relationship.

[3] We can capture knowledge by writing it down and hence it provides a source of information! The point here is that the source of information

Knowledge should imply persistence and viability (as a test). Knowledge is ultimately the product of the process[4] that determined its status (e.g. processes: evolution, thought).

In [8] the confusion between data and information is highlighted and in [8] the authors attempt to find definition(s) or at least a basis of agreement. Starting with the Shannon Paper [7] and looking at subsequent authors from this period [9] results in a bias towards information being equivalent to communications between systems aware of a common basis (e.g. language). But in [8] the notion of data as potential-information and the concept of acting on the information (referred to in the paper as *impact*) brings it back to the mind and the need to "comprehend.

A. *Information and data*

In the early days of computing after the Second World War, communications and the physical nature of signals, encoding and decoding were prominent.

In [6] Dr Weaver has written some philosophical notes on an interpretation of Shannon's paper [7]. In this book, he refers to information as the measure of what could be transmitted. At the same time, he clearly stated that any associated meaning belongs to a colloquial use of the word information. He points out that representation relies on knowledge residing with a receiver or a transmitter (not the intermediate hardware but the final receiver (and transmitter) – e.g. a person). Thus a single pulse = 0 may represent the pre-arrange block of text known as the "King James Version of the Bible", while pulse = 1 represents "*yes*". With this example, the separation of information from meaning is quite clear. The book expands on the analysis of channel capacities and the effects of noise on channels and describes more efficient encodings; even then, data does not appear as part of explanation. This description of the definition of information is all to do with consciously using changes in a given signal to relocate something that is known in one place and needs to be known elsewhere. Besides transmit and receive, we need to encode and decode and further we need an interpreter. The simplest decoder uses a memory address. Store a set of responses in some memory devices, send the devices to others, when you want something known – transmit an address. The receiving *other* then applies the address. The stored set of: addresses and associated output, implies data, the meaning of it, is *potentially* knowledge but it still needs to be emittable. (The emitted response can be seen as an information flow to a tightly coupled system in which we can safely anticipate the outcome from our emission i.e. we *know* what will happen).

B. *Pyramid of Data, Information, knowledge... and wisdom*

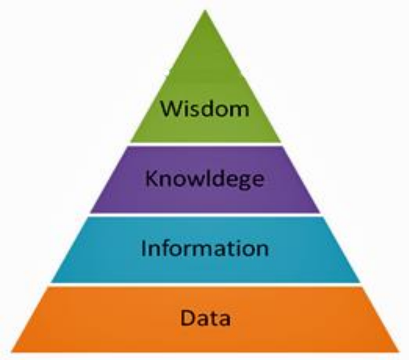


Fig. 2. DIKW diagram

This diagram (the DIKW model [10] [11] [12]) in its many forms has been around since before 1982.

Diagrams of this form suggest that there is a kind of hierarchy (increasing in some property towards the apex) perhaps progressive refinement or some sort of filtering? It is representative of a flow of progressive organisation where knowledge is definitely more organised than information or data. Perhaps wisdom belongs elsewhere - systems that can use knowledge, data and information cannot be wise unless they are self-referential (i.e. think).

Perhaps there is a layer or demarcation missing - a step or junction between information/data and knowledge. This step would require a model that can evaluate information/data that has been organised to provide a basis (or *belief-set)*. Knowledge implies a migration from belief to certainty.

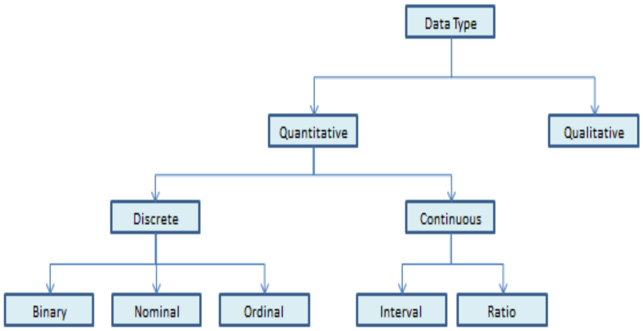
II. On Data

Data representation and usage now exceeds the simple idea of data just being a collection of measurements. This section will touch on the nature of data.

A. *Types of data (as opposed to data-types).*

In computing the different representations of standard binary items such as string, integer, and float are referred to as *types*. We may also attempt the classification by applicable properties of data.

1) *Different types of data (source http://www.sigmamagic.c) om/forum/archives/176*



is not the same as knowledge (it is essential to include the resulting appropriate response).

[4] The originating process is evolution and it is likely that thought arose from the organisational effect that you get when outcomes can be anticipated.

Qualitative data: non-numeric: e.g. colour, members of a set. Quantitative data: numeric or ordered: e.g. Size, number of some common attribute (e.g. for animals it could be legs). The box "Binary" is a little ambiguous, but could be a reference to 2–state types such as ("true/false") or ("on /off"), ("yes/no"). Ordinal implies that a set of items is arranged in rank order (with respect to some common property). Nominal – elements of a set (no ordering). An Enumerated set (frequently used in computer languages) is a set in which the list of items are numbered sequentially (e.g. from left to right, or top to bottom) but they need not be ordinal.

2) Types classified by use

Source http://www.paragonian.org/glossary.shtml

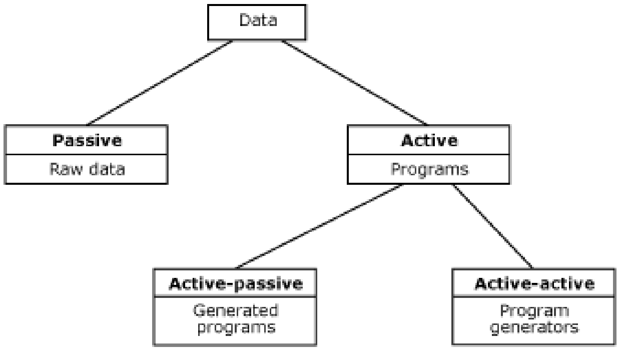


Fig. 3. Data usage

Because data can be interpreted by hardware as a set of instructions, it is classified above to be active (when used as instructions) and passive when input into a loaded software process. The significance of classification by usage is that data must be processed and therefore some sort of stepwise procedure has been planned for it. All data implies a degree of pre-planning (i.e. collection of data for a model or processing as a set of instructions for hardware).

III. Sensors, information and knowledge

Books (and pictures, screens, sound etc) can be regarded as data (rather than knowledge or information). Information flows when the sensors (eyes and ears) interact with a data-source (e.g. words) which result in internalised responses that can trigger mental images or a mind-generated context. This flow of symbolic forms (states) must mirror real input and their attendant responses to produce a stream of potential information and subsequent potential responses to it. (Potential in that it is available symbolically and, although associated with real responses they are not yet emitted.) The symbolic image which is invested with the environment's consistency and modelled accordingly can then yield images of interacting with the real world and the consequential images used to construct sequences of likely interactions with the model environment. It therefore carries with it the test of viability albeit masked by complex sequences. Through this tenuous process and somewhat amazingly, a book can be used to impart knowledge!

1) Signal processing

Some engineering solutions process sensor information in order to improve the ratio of signal energy to background noise energy and at the same time can extract from signals directional measurements. They can also classify the signal to find characteristics that will help in source-identification. The results, after processing the signals, inform the sensors' creators as to the disposition, type, and nature of the signal sources. Thus, identification and location at a given moment can be determined. This is planned e.g. signal processing model, but **not** by the system undertaking the processing.

For biological systems, the sensors tend to be simpler but in greater abundance and higher densities. Their disposition helps directionality and they are often combined (including different types) to give a more complete spatial sample. They provide, without any planning, a "snapshot" at an instance from several sensor types. This process addresses an associative set of likely states (i.e. the context - context generation is an example of sensor-data fusion [13], [14]. Defence-system builders have known the importance of sensor-fusion for decades. It usually involves the production of a composite view of the theatre of operations collated from many sensors. This forms a current picture (a basis/context) for planned operations.

In biological systems, a context is a receding memory of previous inputs and responses, which modify the set of potential responses to the current input-set. Assuming that the brain arose as a control-orientated system, the picture is in terms of responding rather than a traditional database of events and objects. This is why I believe that it is important that any definition of knowledge should include the emitted response. In a control-based system, the responses are addressed, so that they can more easily meet the time constraints for that action. Control-base systems can be described as being feedback orientated – in this case the direct link from input to response and then subsequently modifying feedback can track the emissions that constitute a viable control-path. At this stage, the system is entirely automatic. Simply emitting plausible responses to input-events in a timely manner and these being suitable (i.e. not prejudicing the system). Over time this constitutes a test of viability. In this situation, no planning is available to anticipate suitable responses to events. (The mechanism of thought is needed for anticipation and planning.) For biological systems sensor-processing is analogous to many individual independent processes, which can interact associatively. This also promotes tight-coupling with the environment and its associated time-constraints. There are information-flows and states (which arise automatically (thought not involved)), but there is no actual (static) data. States may be regarded as a precursor to data in the development of memory for the control part of the system.

The flow of input information, its association with context and the responses are determined by the connectivity of the neural paths, and further by building adaptive connections, the flow can modify (refine) the responses in a heuristic (and convergent) manner [15]**.** As the response becomes appropriate, the adaptive component must stabilize to become a regular connection. Hence, the connections between input and response become persistent. The information-flow transitions from delivering a plausible response to delivering a viable response (its viability-status with respect to an environment) (i.e. it has become knowledge). No words or language was needed to complete this transition. Nor at this stage do we invoke thought. This stage requires a locally consistent environment, a reasonably matched[5] biological system with ever-increasingly organised responses at two levels: the internal level (pathway adaptation) and genetically, the system layout (for that organism). The pre-planning does not yet exist, but the heuristic refinement process (evolution) takes care of propagating the long-term adaptation. The intermediate states are not part of the real initial input but through pathway adaptation they refine (and classify) the connections and hence the outcomes. A key point for this early data-like form is that it is essentially representative of some driven, processing-mechanism.

Remembering the review-section I.A, this is similar to the ideas of from [8] in that data is not significant but the relationship between information, transmission and interpretation (acceptable response) is emphasized. In addition (because of thought and language) there is the warning about colloquial meanings for information and knowledge.

For review-section I.B: data and information have exchanged positions in the DIKW pyramid and knowledge has acquired the label *viability* along with a relationship to output (i.e. a complete feedback cycle). Data and its transient purpose (intermediate steps) reveal that the ideas in II.A.2) are significant.

B. Signals, information and data

The transformation from a continuous (temporal) domain to a spatial (discrete) representation of that domain in which a planned process can be implemented, marks the transition between information and data. All data is planned and representative of an information flow. However, Information flows can also be planned (communications) and are continuous (it is, perhaps, this apparent ambiguity that gives rise to the confusion between information and data). A planned information flow (communication) converts to a time-based flow by superimposing changes on it. This then is processed to retrieve the original data. The word process is synonymous with concept of interacting with the continuous flow to yield states (spatial representations that have a lifespan). Since data can be spatially representative of an information flow, is always predictable (consistent to a degree) and invariably discrete. Data can be stored (spatial property) information (signals) cannot be stored but can be represented as a data[6] sequence and hence be recoverable, but always as an approximation[7].

1) Separation of Terms

The term information belongs to the continuous flow of energy that we observe using sensors. It is essentially time and space located. It is also continuous, which for processing requires a transformation[8] by sampling in real-time. Each sample at a given moment is a measurement-value of some distinct property in time and space; it comes from the environment and so is consistent with respect to events occurring in that environment. The events may generate or moderate the energy flows to generate some sort of signal-set. The received signal was not pre-planned and it may or may not require a response. At its simplest, the input may elicit a response directly (if required) and it will appear that the event will (autonomically) address the response. If that response is *always* appropriate (for that input) (i.e. maintains system viability with respect to the environment), then we can call it knowledge.

Thus a possible definition of knowledge might be: *knowledge is the persistent, appropriate response to a given input*. The process of acquiring knowledge involves adaptation – this, at the autonomic level (no thought involved), can be achieved by modifying connections [15] for changing the transfer rates along neural paths. This change of weight ultimately should converge to a stable path-transfer rate to perform consistent addressing. The initial rate is a trial rate - (analogous to a plausible (i.e. believed) likely rate) and then the final rate (after some time) is the conversion to experience (knowledge). Adaptation is the key ingredient in his case: if we assume that intelligence is the property of a system that defines its ability to adapt then even path-weight changes implies systems with intelligence. We rarely think of intelligence as the consequence of automatic "built-in path-weight changes". Thus we might infer that the application of intelligence on sources of information yields applicable responses – this provides a route in an evolving system to add the further adaptation using reflexive thought to modify responses. The expectation or "belief" is the current state of the system and as that current-state is modified and stabilized with respect to a given environment to a fixed addressable response (knowledge). Information is the flow of sensor-input that stimulates the system. At this point there is no apparent data. However, locally, states arise, which route the flow of information through relatively fixed paths to address responses (analogous to combinatorial logic for electric circuits). If this feeds back internally (symbolic responses to current state), it may be viewed as a process (or extending the electrical analogy: sequential logic). The conversion to a sequence from a spatial representation (addressed states) marks the information to data transition. This uses **states** as spatial transformations, which always have a distinct lifespan and are consequently a primitive form of data. The process here is predominantly decoding and addressing (associative addressing as found in neural systems).

IV. Synthesis

Simple signals often have the results of interactions superimposed on them and, in essence, are information carriers rather than data items. If we have a simple signal and then superimpose changes in some manner then we can arrange to decode those further changes on receipt of this "modulated"

[5] Matched = time to respond is of the order of the time in which significant changes occur.

[6] Compare this to the wave-form and then measurement process in Quatum Mechanics. Data = observed state, information the waveform

[7] Even if the only thing missing is noise – which is unwanted anyway. Sometimes data looks more perfect than information!

[8] Replaces a continuous function of time and space with discrete sequence of measurements to approximate the signal at a given time and location

signal. Because we know what was done we can processes the information to recover the overlain interactions. Encoding changes onto a communications-signal is commonplace and allows us to move the information from place to place. Given that the encoding was planned, it can be decoded to recover the data and regenerate the sequence of changes. It may well be used to generate further sequences and even new signals (information flows). The signal (information) served as a means to disseminate and perhaps even invoke responses to the message.

1) Signal –data transition

The transition from signals to a static representation is the transition from an information flow to a data representation. The representation was planned for, in order to store/process it. Information belongs to the temporal or continuous phenomena while data is a discrete, spatial, time-independent representation which is often quantized or approximated (at the very least it has accuracy limits).

2) Colloquial Information

It is perfectly acceptable to say we get information (and even knowledge) from a book. But the end product is *data* stored and knowledge (that can be addressed). Sight (and sound) is an information flow. The translation to words (or even images) is a symbol-to-state transformation with an opportunity to relate them to internal models. The models modify paths and address outcomes (which return symbolically rather than be emitted). This changes paths and hence our responses in the future, so reading (and listening) to a book processes an information-flow to store changes in a mind. We are processing data (words) generating a flow (information) and synthesizing data changes (pathway connections). We have learned to say “get information” ... which is part of the process and even “acquire knowledge” which is the end result. We don’t tend to say “store potential state “ as a result of reading – this being, perhaps, the more accurate description.

In our animal-past the need to remain viable means that, the more attuned our responses are to the input that we receive, then the more *successful* we are. At first we have simple tuning, in which paths become more entrenched the more use they get (see Hebb rule in [16]). This becomes model-based, in which internal feedback allows us to explore the unfolding context in our minds and hence perceive likely outcomes. Our beliefs (anticipation) about the outcome become entrenched as they prove to be successful. Then we no longer need to model - simply address the usual (but now refined) response – this is also knowledge. This mechanism is a function of a property that reflects our adaptability - its units should measure viability acquired from the *older process*[9] which defines viability by not dying, and procreating (propagating our genes). Not something, one would normally associate with intelligence.

3) State versus Data

Electronic circuits use data storage devices; these are known as memory (e.g. dynamic and static random access/read only memory) or simple latch devices (flip-flops). These are addressable - i.e. set a binary value on the address-lines of the device (clock/enable signal to transition) and an 8/16/32/64 bit binary values will emerge on the output lines. Such circuits do not directly exist in the brain, but with feedback, a similar effect can be created. Normally it is a case of hold onto the state of a path sufficiently long enough to trigger an outcome should further input suggest it. This transient, short-term memory is more normally thought of as “state. The brain stores events in the short-term using this kind of memory device. These states are defined in terms of pathways and their links from input to responses, along with naturally occurring propagation delays. States are a primitive form of data. They can also be modified by changing the properties of neural pathways [16]

V. Conclusion

Information: This is a (usually time-dependent) flow (e.g. of energy), which may be modified by the effect of interactions and changes that occur along each channel/route taken though the local space. Information can be regarded in terms of signals.

Data: A spatial representation organized from an information flow. Generally, the organization comes from a mind and the representation is constructed symbolically. Any purpose is defined by processing or decoding. Data only exists in the presence of a modelling agent that is able to manipulate static components to form a representation (typical static components might be chemical (DNA), pathway thresholds, magnetic polarity, static charges, electrical voltages (e.g. binary states). For the more primitive form of data we have states (electrical, mechanical, neural, chemical short-lived intermediate instances). The “purposeful” mechanism for data in biological systems comes initially from the organising (heuristic) process of evolution. This creates data using chemical (amino acids) sequences.

Information and communications: An information flow can be constructed by superimposing data forms (represented in terms of properties of the given flow). The encoded signal can then be processed later to yield a data sequence and further re-processed for some underlying purpose. Information such as this is used for communications between transmitters and receivers. Again, this mechanism requires a plan manager [4] and so communications is a data-orientated information flow (as in paper [8] where the encoding and transmission of information lead to knowledge after interpretation). The encoding process must also be modelled.

Knowledge: appropriate, persistent response to a given input. The appropriateness is determined by external interactions, which generally maintain viability for the responder. For a model-based system: agreed/accepted outcomes from given inputs to widespread, long-term, consistent set of beliefs. The model’s basis is strictly a belief-set, but approximates to truth/knowledge if use/acceptance is, for example, successful and widespread.

A. The AI aspect

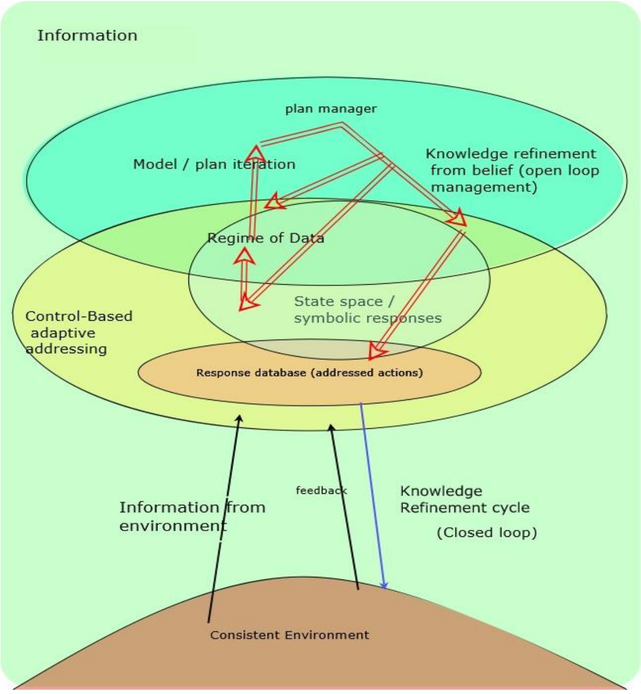
There is nothing new in associating biological systems with basic control theory [17], but in many years of the computer

[9] Reference to evolution (as an heuristic process)

age this was overlooked in favour of linguistic[10] models. The key aspects of control are perhaps feedback, and convergence (closed-loop, with feedback between responder and environment).

The notions: viability, closed-loop control (with respect to the environment) and open-loop modelling i.e. planning with adaptive modification of responses supports bio-systems as an evolved capability. The meanings of information, data and knowledge should coincide with this notion. It can also support a definition of intelligence in terms of adaptability with respect to an environment and its attendant process of evolution in which the test of viability (which includes pro-creation to pass on that viability) is the key ingredient.

Fig. 4. The domains of Information and Data



Information flow is everywhere. Knowledge is determined (ultimately) in terms of viability (this often includes *agreement* [11] in the currently new memetic environments (social and economic) [18]). Data is always a symbolic entity such as language or numbers, and so requires a plan-based modelling, which uses symbolic forms (labels, words, and numbers) to represent responses and inputs to and from the environment. The internal process (thought) is naturally recursive and mirrors likely feedback as seen by the autonomic system-component .

VI. Bibliography

[1] B. Russell, The History of |Western Philosophy, George Allen and Unwin, 1961.

[2] J. Pohl, *Transition from data to Information*, 2001.

[3] T. Gruber, From: Encyclopedia of Database Systems. Section: Ontology, L. Liu and M. Tamer Özsu, Eds., Springer-Verlag, 2009.

[4] J. R. Hobbs and A. S. Gordon, “Toward a Large-scale Formal Theory of Commonplace Psychology for Metacognition,” University of Southern California, 2006.

[5] D. Creager and A. Simpson, “Towards a fully generic theory of data,” Oxford University Computing Laboratory, Circa 2004.

[6] C. Shannon and W. Weaver, The mathematical Theory of Communication, University of |Illinois Press, 1949.

[7] C. E. Shannon, “A Mathematical Theory of Communication,” *The Bell System Technical Journal*, vol. 37, pp. 379-423, 1948.

[8] C. T. Meadow and W. Yuan, “Measuring the impact of information: Defining the concepts,” *Information Processing & Management*, vol. 33, no. 6, pp. 697-714, 1997.

[9] L. Brillouin, Science and Information Theory, New York: Dover (2013) Re-publication from Academ ic press, 1962.

[10] “DIKW pyramid,” Wikipedia, 21 09 2015. [Online]. Available: https://en.wikipedia.org/wiki/DIKW_Pyramid. [Accessed 21 September 2015].

[11] J. Rowley, “The wisdom hierarchy: representation of the DIKW hierarchy,” *Journal of Information Science*, pp. 163-180, 2007.

[12] D. Weibberger, “The problem with the Data-Information-Knowledge-Wisdom Hierarchy,” 02 2010. [Online]. Available: https://hbr.org/2010/02/data-is-to-info-as info-is-not. [Accessed 21 09 2015].

[13] L. F. Pau, “Sensor Data Fusion,” *Journal of Intelligent and Robotic Systems*, pp. 103-166, 1988.

[14] E. L. Waltz and D. M. Buede, “Data Fusion and Decision Support for Command and Control,” *IEEE Transactions on Systems, Man and Cybernetic* , Vols. SMC-16, no. 6, 1986.

[15] O. Paulsen and T. J. Sejnowski, “Natuiral Patterns of activity and long-term synaptic plasticity,” *Current Opinion in Neurology*, vol. 10, no. 2, pp. 172-179, 2000.

[16] J. C. Principe, N. R. Euliano and C. W. Lefebvre, “Hebbian Learning and Principal Component Analysis,” in *Neural and Adaptive Systems: Fundamentals Threough Simulation*, John Wiley and Sons, 1999, p. Ch 6.

[17] N. Wiener, Cybernetics:, MIT press, 1961.

[18] R. Dawkins, The Selfish Gene, Oxford: Granada Publishing Ltd, 1976, pp. 203-217.

[19] N. M. Schmitt and R. F. Farwell, Understanding Automation Systems, Dallas: Howard W Sams & Company, 1984.

[10] Word orientated models rather than direct sensor data

[11] This is often results in a weaker product