*International Journal of Developments in Big Data and Analytics*
*Volume 1 No. 1, 2014, pp. 6—17*

# Big Data and Established Models of Knowledge

DERYN GRAHAM
University of Greenwich, United Kingdom

ABSTRACT This paper focuses on knowledge and describes the relationship between heuristic, causal and statistical models of knowledge and their association with Big Data. These models can be differentiated by the mode of generation; namely the approach used to acquire the knowledge (knowledge acquisition). Causal reasoning, or reasoning from first principles, often uses simulation to obtain the entire set of causes and effects for a complex structure leading to a hierarchy of descriptions. Knowledge-based reasoning tries to emulate the knowledge and experience that an expert applies in diagnostics (the heuristics) through knowledge elicitation techniques such as interviews. Straddling causal and heuristic models of knowledge is the statistical view. This paper depicts the relationships between these models and the State Space, and discusses where Big Data fits in and issues yet to be addressed.

*Keywords:* Knowledge, Causal Reasoning, Heuristics, Statistics, Big Data.

## Introduction

Heuristic, Causal and Statistical models of knowledge and Big Data can be differentiated by the mode of generation; namely the approach used to acquire the knowledge (knowledge acquisition). Causal reasoning, or reasoning from first principles, often uses simulation to obtain the entire set of causes and effects for a complex structure leading to a hierarchy of descriptions. Knowledge-based reasoning tries to emulate the knowledge and experience that an expert applies in diagnostics (the heuristics) through knowledge elicitation techniques such as interviews. Straddling causal and heuristic models of knowledge is the statistical view, where statistical data is usually collected (acquired) from multiple sources such as databases and questionnaires, with further statistics generated by the application of mathematical formulae to produce purely numeric (quantitative) values.

This paper focuses on knowledge and describes the relationship between heuristic, causal and statistical models of knowledge and the State Space, and their association with Big Data. The paper depicts the relationship between these models and discusses where Big Data fits in and suggests issues for Big Data yet to be addressed.

## Models of knowledge

Heuristic, Causal, Statistical and Big Data models can be differentiated by their origin or mode of generation, their quantitative or qualitative characteristics, "format", whether or not domain specific, and their main affinity with data, information or knowledge (Graham, 2014).

Knowledge acquisition for causal reasoning, or reasoning from first principles, often uses simulation to obtain the entire set of causes and effects for a complex structure leading to a hierarchy of descriptions. An example of the use of causal reasoning is Automatic Test Equipment (ATE) for computer hardware fault diagnosis (Graham, 1990). Knowledge is therefore described as a hierarchy of descriptions (behaviours) linking cause (faults) and effect (symptoms). Causal reasoning models are domain specific and numeric data hierarchies.

Knowledge-based reasoning tries to emulate the knowledge and experience that an expert applies in diagnostics (the heuristics) through knowledge elicitation techniques such as interviews, acquiring both qualitative and quantitative values. Knowledge is often expressed in the form of rules. Backwards or forwards chaining through these rules should lead to one or more solution candidates.

Expert or knowledge-based systems separate the domain expertise and knowledge (knowledge-base) from the mechanism (a forward or backward chaining inference engine). "Knowledge-based systems provided clear and logical explanations of their reasoning, use a control structure appropriate to the specific problem domain, and identify criteria to reliably evaluate its performance" (Luger, 2002: 20-21).

These systems require the acquisition of knowledge and expertise, and are more akin to a human expert in a specific domain. They are rule based, applying propositional logic or predicate calculus to reach conclusions based on evidence (attributes of human experts). They enable multiple conclusions with associated degrees of statistical confidence (confidence factors), as well as "How" and "Why" queries. Expert Systems have difficulty in capturing "deep knowledge" and are not truly intelligent, but such systems attempt to encapsulate knowledge and expertise.

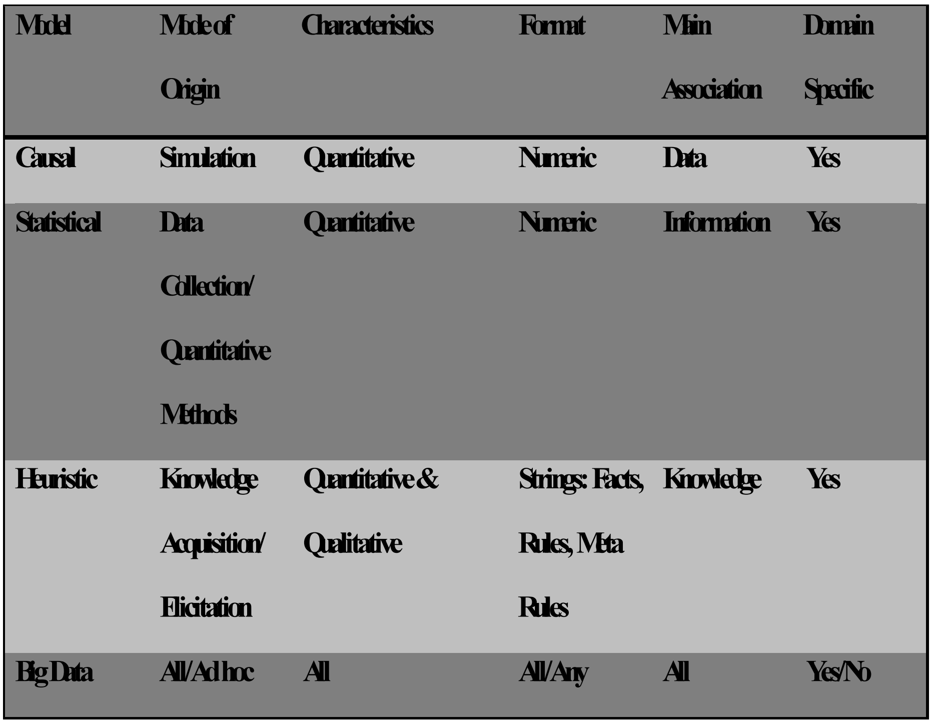
Straddling causal and heuristic models of knowledge is the statistical view where data can originate from multiple sources and there is no single knowledge acquisition approach. In addition, statistical information is the result of the application of mathematical formulae. Most statistics are domain specific and take the form of statistical data or information (when analysed). Statistics may aid the identification of knowledge, by statistical weighting (such as confidence factors) or search. The model is purely numeric and quantitative, and statistical data is usually collected (acquired) from multiple sources such as databases and questionnaires, with further statistics generated by the application of mathematical formulae.

Causal, heuristic and statistical models are likely to be domain specific because of the Combinatorial Explosion (described later).

## Characteristics of models of Data, information and knowledge

Graham (2013) depicted the "transformations" from data to information and then from information to knowledge, discriminating between data, information and knowledge through the dimension of time for the purpose of learning (competence achievement). Humans do appear to take in raw data with a specific goal, to organise the data so that it has meaning, and to analyse this information (compare and contrast, etc elements of Bloom's (1956) taxonomy) to a more structured form, namely knowledge. This knowledge or expertise is the basis of knowledge-based systems and heuristic knowledge models.

Causal, statistical and heuristic models have been differentiated by their main affinities to data, information and knowledge, respectively, in Figure 1 below.



| Model | Mode of Origin | Characteristics | Format | Main Association | Domain Specific |
|---|---|---|---|---|---|
| Causal | Simulation | Quantitative | Numeric | Data | Yes |
| Statistical | Data Collection/ Quantitative Methods | Quantitative | Numeric | Information | Yes |
| Heuristic | Knowledge Acquisition/ Elicitation | Quantitative & Qualitative | Strings: Facts, Rules, Meta Rules | Knowledge | Yes |
| Big Data | All/Ad hoc | All | All/Any | All | Yes/No |

*Figure 1: Characteristics of Causal, Statistical, Heuristic and Big Data Models of Data, Information and Knowledge*

### Pros and cons of models of domain knowledge

The State Space is the space of allowed problem states. State Space may take the form of a tree, or (when it is possible to return to a previously visited state), a graph. In all but trivial cases, it is not possible to explore State Space fully (i.e. until every path reaches a goal state or a dead end). If the branching factor (the number of successors to a given state) is b and the tree is explored to a depth N, there will $b^N$ nodes at the Nth level. The classical example is a Chess Board. The State Space is large for even the simplest of problem domains and can suffer from the Combinational Explosion.

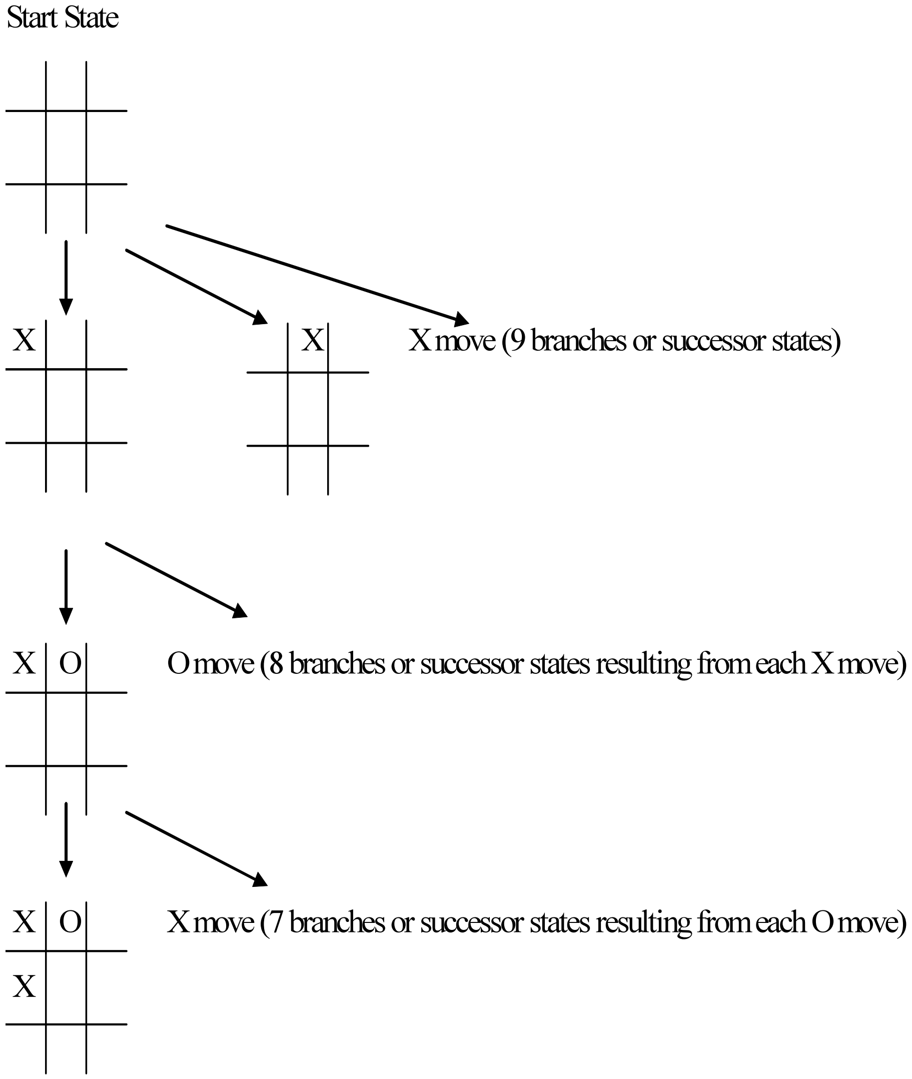

Start State
X
X
X move (9 branches or successor states)
X
O
O move (8 branches or successor states resulting from each X move)
X
O
X
X move (7 branches or successor states resulting from each O move)

*Figure 2: Partially developed tree for the game of Tic-Tac-Toe*

Consider for example, the simple game of Tic-Tac-Toe. The initial state (the start or root node) is a blank 3x3 matrix. There are two players (X and O) who take it in turns to enter an X or O respectively. For a player to win, the objective is to achieve a horizontal, vertical or diagonal line (goal state). Figure 2 shows a partially developed tree for the game. From the start node, the first player (shown to be X) has nine possible positions (9 branches) from which to select. The second player (O in this instance) then has eight possible positions (branches) from which to choose. Player X subsequently has 7 branches, and so on, and so forth until either X wins, O wins or stalemate is reached. However, not all positions are equal. Figure 3 shows that the centre of the matrix, co-ordinates (2, 2) can lead to 4 possible complete lines (wins or goal states): 1 horizontal, 1 vertical and 2 diagonal, whilst corner coordinates (1, 1); (1, 3); (3, 1); (3, 3), can lead to 3 possible complete lines (1 vertical, 1 horizontal, 1 diagonal) or goal states. These facts may be known experientially (heuristic rules or cases), statistically (probability of an X or an O win), or through simulation (causal reasoning).

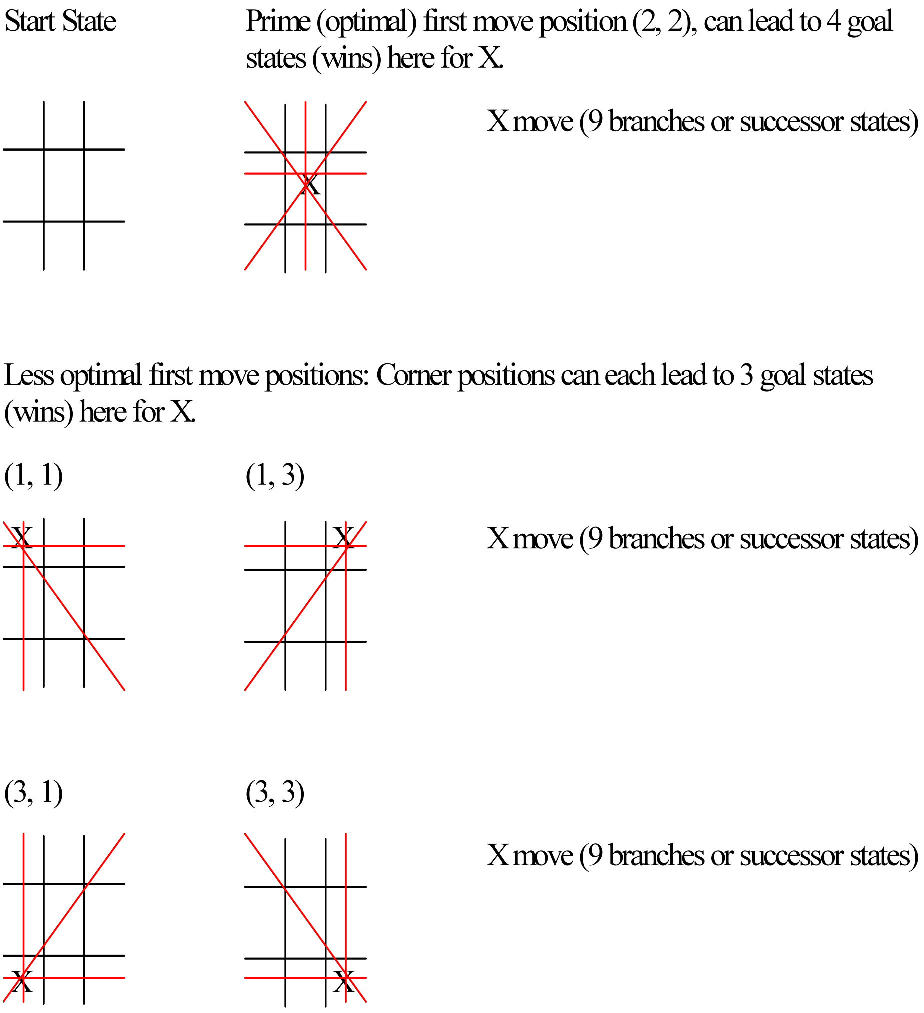

Start State
Prime (optimal) first move position (2, 2), can lead to 4 goal states (wins) here for X.
X
X move (9 branches or successor states)
Less optimal first move positions: Corner positions can each lead to 3 goal states (wins) here for X.
(1, 1)
X
(1, 3)
X
X move (9 branches or successor states)
(3, 1)
X
(3, 3)
X
X move (9 branches or successor states)

*Figure 3: Start state and optimal X first move positions for the game of Tic-Tac-Toe*

Causal, knowledge-based reasoning and statistical models have their advantages and disadvantages. The main advantage of causal reasoning is that it is definitive; causes and effects (states and their pathways) can be clearly defined (known). The main weakness of causal reasoning is scalability; scaling-up from simple (small) to complex (large) problem domains is not easily achieved. The Causal Model would consider every possible outcome from every possible combination of moves, i.e. the entire State Space.

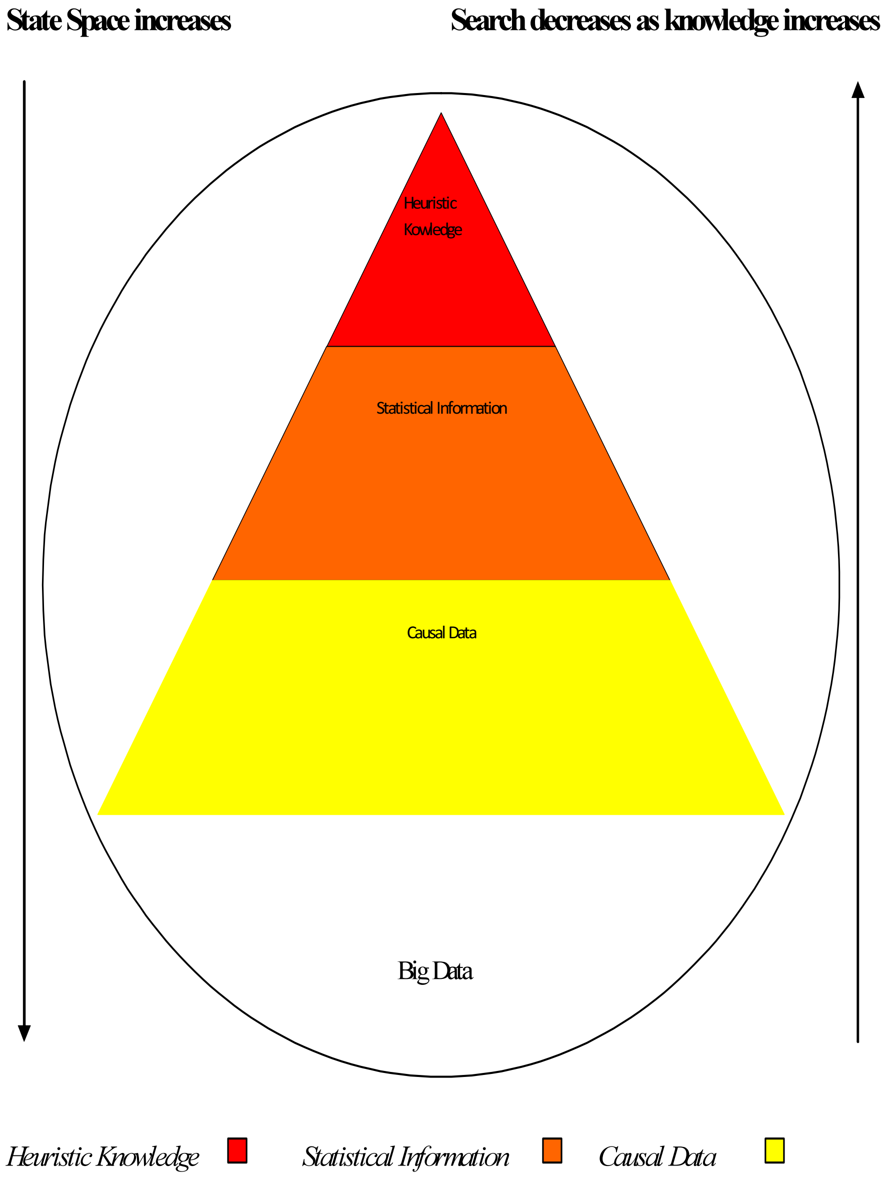


*Figure 4: Models of Knowledge within a State Space Pyramid for a Problem Domain*

The heuristic approach applies "rules of thumb", such as set pieces (cases) in Chess or Tic-Tac-Toe, using knowledge to guide the search (of the State Space). Knowledge-based reasoning has the opposite issues to causal reasoning; its heuristic approach effectively contracts the State Space, but the heuristics may not be as well defined.

The statistical outlook covers both causal and heuristic models. The heuristics are also likely to map against probabilities (of decision and goal outcomes) which would be experientially realised by human experts, i.e. guide search as simply demonstrated for Tic -Tac-Toe, that not all start positions are equal. The main advantage of the statistical model is its simplicity; purely numeric and quantitative, it is usually combined with other models to provide information (to guide search and contract the State Space), for example in knowledge bases where statistical probabilities are employed to provide confidence factors (the measurement of confidence or belief in a given solution).

Causal reasoning is strongly associated with quantitative data whilst knowledge-based reasoning has a greater affinity with qualitative (heuristic) "data". This is reflected by the fact that causal reasoning applications are often automated (such as ATE) analysing numeric data. Knowledge-based reasoning involves knowledge acquisition and some elicitation of rules from human experts using qualitative methods such as interviews.

Looking at fault diagnosis, the complete causal model for a system or device would possess all possible faults (causes) for all possible symptoms (effects), i.e. the entire state-space for a given hardware device domain. Both the heuristic and statistical models can be mapped onto the causal model. It is suggested that the relationship between the heuristic and statistical models may be a close one, with both the heuristic and statistical models homing in on the most common faults, as might be experienced by human experts and is therefore experientially based. In the statistical model, this would relate to the frequencies of faults in terms of probabilities, whereas in the heuristic model, this might equate to experience. The heuristic model can therefore be skewed by extraneous cases when the experience gained is not a true indication of the actual fault frequency (merely an aberration or "blip").

Searching the State Space to identify faults in Figure 4 advocates a heuristic approach first because of its reduced State Space, before considering the use of the statistical, and, if all else fails, causal reasoning (or reasoning from first principles) being employed to identify faults and solutions. The divisions between models are likely to be fuzzy and, unlike the depiction in Figure 4, indistinct.

The data in Figure 5 could be data held in a database, i.e. a conventional source acquired by conventional knowledge acquisition means, and is domain specific. The quantitative data would tally with statistical data. The data could be converted into statistical information through the application of statistical formulae, possibly via an Information System. The accrued data in a data warehouse could be converted in to knowledge through techniques such as data-mining, pattern recognition and machine learning. Knowledge-based systems are often front-ends to data warehouses and databases.

*Life Insurance Example*

Data: Mr. (male) John Smith died in London, England on the 1st February 2003, aged 74 years.

123456SMITHJOHNMLONDONENGLAND0102200374

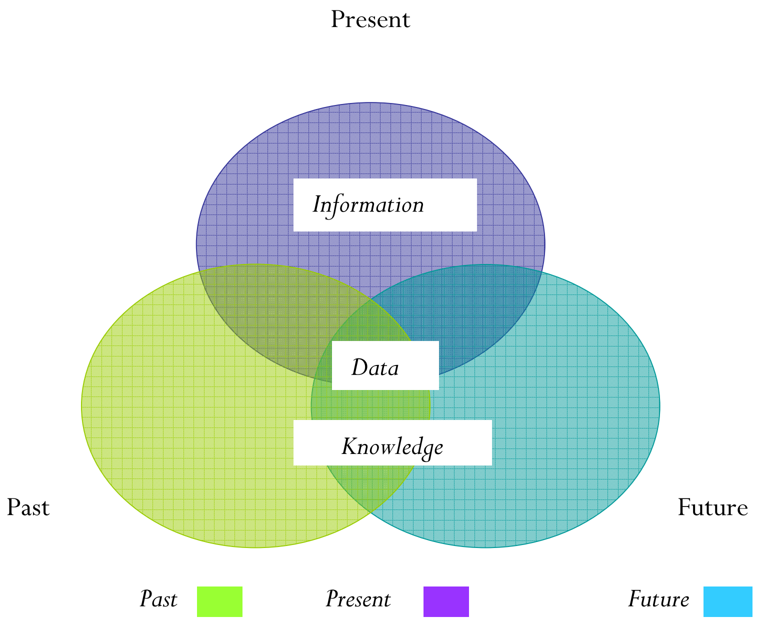
Mr. (male) Peter Brown died in Stafford, England on the 23rd September 2003, aged 69 years.
789101BROWNPETERMSTAFFORDENGLAND2309200369
Etc.........

Information: The average life expectancy of men in England in 2003 was 73 years.
Knowledge: The predicted life expectancy of men in England in 2014 is 80 years.

*Figure 5: Data, information and knowledge: Life Insurance Example (Extended from Graham 2013, p.176).*

The actual alphanumeric data strings are given below the more readable description of the data beginning with six digit identifiers. Age is given as an attribute, but could be calculated if the Date Of Birth (DOB) is known. The causal model would encompass all the data (states) for all criteria; there is no contraction or reduction of the state-space. Figure 6 (Extended from Graham, 2014a) adds a temporal dimension. As shown in the Life Insurance Example (Figure 5), data is absolute and with a value independent of time. This is not true of information; information must be timely if it is to be informative and of value, and usually deals with the now (present). So in the example above, the average life expectancy of men in England being 73 years was only information in 2003, and constituted historical data in 2014. It is suggested that knowledge synthesis, can take place at any point in time post the processing of information, relying on past, historical information (recent or otherwise) to enable future predictions. For example, the employment of data mining: historical (past) data and current information are mined to make (future) predictions and hypotheses. Although knowledge is employed in the present, the creation of new knowledge is perhaps associated more with the past (events) and the future (predictions).

*Figure 6: Temporal View of Data, Information, Knowledge (Venn diagram) and Big Data*



Causal models are likely to be temporally independent data hierarchies. Statistical models generate information and are of the "now" (present). Knowledge-based models fit more with the future predictions based upon past (historic) events. Figure 6 suggests the temporal relationships between data, information and knowledge. Big Data is omnipresent and is therefore not shown in Figure 6. The suggested steps involved are the presentation of external data (facts) and their organisation into information and subsequent analysis to knowledge.

## Discussion and Conclusions

There are multiple definitions of Big Data, basically Big Data refers to very large datasets. Big data is commonly agreed to have some combination of five characteristics: Volume, Variety, Velocity, Value and Veracity. Volume is where the amount of data to be stored and analysed is sufficiently large to require special consideration. Variety refers to the data being of multiple types and from multiple sources, such as structured data held in tables or objects for which metadata is well defined, for example, semi-structured data in documents, where the metadata is contained internally (XML documents), or unstructured data such as photographs, video, or any other form of binary data. Velocity refers to the data being produced at high rates and operating on "stale" data is not valuable. Value is where the data has perceived or quantifiable benefit to the enterprise or organisation using it. Finally, Veracity is where the correctness of the data can be assessed. Big data can exploit cloud data, adding public cloud data to private cloud data (Gordon 2013).

McKinsey Global Institute (Neaga and Hao, 2013) suggested models for Big Data characteristics based on the source, with the main key characteristics being those of volume, variety, value and veracity.

Attributes for each modelled characteristic (Neaga and Hao, 2013, p. 36):

> *"Volume: Data at Rest – Terabytes to exabytes of existing data to process.*
>
> *Velocity: Data in Motion – Streaming data, milliseconds to seconds to respond.*
>
> *Variety: Data in Many Forms – Structured, unstructured, text, multimedia".*
>
> *Here, an additional characteristic is Veracity:*
>
> *"Veracity: Data in Doubt – Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximation".*

These characteristics have an implicit temporal element (data at rest, for example) and associated definitions of data, information and knowledge, and relationships with heuristic, causal and statistical models.

So where does Big Data fit? The term "Big Data" is all encompassing as it fits anywhere and everywhere within the domain specific State Space pyramid (Figure 4) and, more importantly, outside. The distinguishing feature of Big Data is its method of collection, often more ad hoc than by design. Much of the knowledge embodied within causal reasoning, heuristic reasoning and statistical models is methodically sought and structured. Big Data is often a bi-product of other things; data stored in public and private clouds or gleaned through social media interactions. Big Data originates from multiple sources; as sensor data, from social media, as well as conventional databases etc,

etc. Big Data that is outside the domain specific State Space pyramid is not data specific to a given domain nor, as data, is it temporally specific as indicated by Figure 6 and supported by McKinsey's model, it exists in the past, the present and the future. It is the filtering and processing through machine learning or statistical analysis and domain application that may convert Big Data into Big Knowledge. It is questionable if Big Information truly exists because of domain specifics combined with temporal relevance.

Big Data includes specific domain information and knowledge "reformed" as data. For example, knowledge and information associated with life insurance (Figure 5) could be "reformed" as Big Data looking at how many people both are born and die in England.

Big Data is everywhere and "everywhen" because everything (data, information and knowledge) begins with data and data is temporally independent. Curran (Sumner 2013) argued that "data centres will be the engine rooms driving the 'Fourth Industrial Revolution', which will see the internet of things and big data transform the way modern businesses operate and societies function" (p. 16).

There is a temptation to use Big Data simply because it is there. A significant proportion of Big Data is likely to be spurious to any specific application or domain. One domain source of Big Data has apparently been utilised successfully for another unrelated domain; the use of an earthquake aftershocks mathematical prediction model applied to crime prediction in Los Angeles (MIT, 2013)—could this be the identification of a natural generic pattern for seemingly disparate phenomena or a unique feature of earthquake models? This question requires further research.

Furthermore, issues exist with Big Data, these are suggested to be:

1. Noise—Data quantity, data quality and relevance (problems associated with the five characteristics). Determining what data is relevant and of value, separating this data from the "noise".
2. Over filtering or analysis—Big data can be overly filtered, analysed and refined so that it will inevitably match the set hypothesis. An example would be the over training of neural networks.
3. Quantum or random element – Spontaneous, unforeseen original data.

An example of the random element is the appearance of "Loom Bands" in the toy market. Previous, past data (big or otherwise) did not predict the appearance and popularity of this toy, there was no trend to analyse. The next new toy trend is normally based on fashion, films and market creation (existing data). Loom Bands could not be predicted because the data was not available, it did not exist. This relates to incomplete knowledge, but in reality, this is an example of a quantum factor – the spontaneous creation of original data. In a temporal context, predictive knowledge is usually based on past data or present information. Knowledge is gained by the analysis of past data to predict future trends, but prior to the appearance of Loom Bands, the trend data simply did not exist.

To summarise, Big Data analytics is equivalent to the question of "Life, the Universe and Everything" (Adams, 2009); you may have the answer(s) in the data, but then you need to work out what the question is, and that may prove more difficult! This paper has looked at models of knowledge (causal, heuristic and statistical) which have been evaluated in terms of their origins and existence within the State Space, and the acquisition and synthesis of data to information and knowledge in a temporal context. This has led to the identification of Big Data, its derivation and position within the State Space and within the context of time, and the suggested issues of: Noise, over filtering or analysis and the Quantum or random element.

*Correspondence*

Deryn Graham, PhD
Faculty of Business
University of Greenwich,
30 Park Row, London, SE10 9LS
United Kingdom
Email: D.Graham@gre.ac.uk
Tel.: +44(0)20 8331 9358

**Acknowledgements**
Loom Bands and The Hitchhikers Guide to the Galaxy are registered trade names.

## References

Adams, D. (2009). *The Hitchhiker's Guide to the Galaxy*, Macmillan Children's Books, Unabridged Edition.

Bloom, B. (1956). The taxonomy of educational objectives: the classification of educational goals, handbook one: the cognitive domain, New York: Mc Kay.

Gordon, K. (2013). What is Big Data? In: *ITNOW September 2013*, pp. 12-13.

Graham, D. (2014a). The temporal dimensions and implications of E-Learning. To appear in: E. Barbera & M. Clara (Eds.), The Temporal Dimensions of E-Learning, Special Issue of *ELEA E-Learning.*

Graham, D. (2014). The Relationship between Heuristic, Causal and Statistical Models of Knowledge and Big Data. *Research Papers on Knowledge, Innovation and Enterprise, Volume II 2014, KIE Conference Book Series*, J. Ogunleye (Editor), 2014 International Conference on Knowledge, Innovation & Enterprise, pp. 8-17.

Graham, D. (2013) Chronology of Competence Achievement. *International Journal of Knowledge, Innovation and Entrepreneurship (IJKIE), Vol. 1, Nos. 1-2, September 2013,* pp. 171-184.

Graham, D. (1990). *Knowledge Elicitation: A Case Study in Computer Fault Diagnosis and Repair*. Unpublished PhD Thesis, Brunel University.

Luger, G. F. (2002). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 4th edn, Harlow, Essex, England, Addison Wesley, Person Education Limited.

McKinsey Global Institute (2011). Big Data: The next frontier for innovation, competition and productivity. In: Neaga, I., Hao, Y. (2013). Towards Big Data Mining and Discovery, *Short Research Papers on Knowledge, Innovation and Enterprise, Part 2 - Innovation, KIE Conference Book Series*, J. Ogunleye, D. Heger and U. H. Richter (Editors), 2013 International Conference on Knowledge, Innovation & Enterprise, pp. 35-43.

MIT Technology Review (2013). www.technologyreview.com/news/428354/la-cops-embrace-crime-predicting-algorithm/ Accessed 06/02/14.

Neaga, I., Hao, Y. (2013). Towards Big Data Mining and Discovery, *Short Research Papers on Knowledge, Innovation and Enterprise, Part 2 - Innovation, KIE Conference Book Series*, J. Ogunleye, D. Heger and U. H. Richter (Editors), 2013 International Conference on Knowledge, Innovation & Enterprise, pp. 35-43.

Sumner, S. (2013). Data Centre Summit: insight and foresight. In: *Computing*, 3 October, pp. 14-16.