International Journal of Knowledge, Innovation and Entrepreneurship
Volume 2 No. 2, 2014, pp. 6—14

Models of Knowledge and Big Data

DERYN GRAHAM
University of Greenwich, United Kingdom
Received 27 February 2014; received in revised form 31 July 2014; approved 31 July 2014

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 discusses where Big Data fits in.

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 their association with Big Data. The paper depicts the relationship between these models and discusses where Big Data fits in.

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. 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

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. The main weakness of causal reasoning is scalability; scaling-up from simple (small) to complex (large) problem domains is not easily achieved. The state-space is large for even the simplest of problem domains and can suffer from the Combinational Explosion. The State Space is the space of allowed problem states.

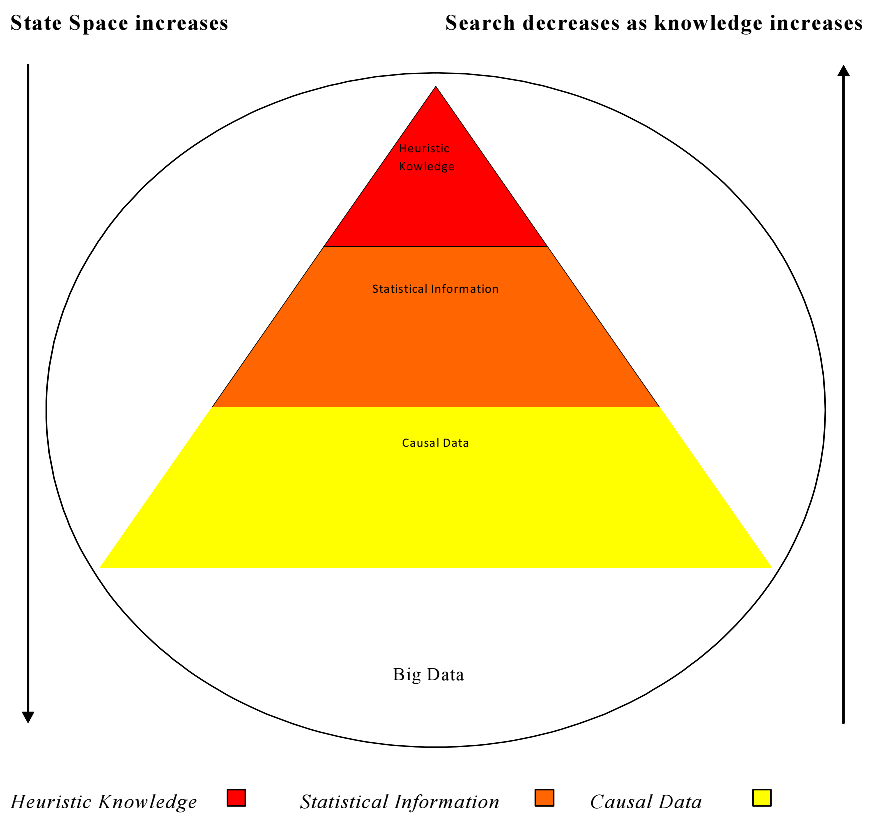

Figure 2: Models of Knowledge within a State Space Pyramid for a Problem Domain

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 Causal Model would consider every possible outcome from every possible combination of moves, i.e. the entire State Space.

The heuristic approach applies "rules of thumb", such as set pieces in Chess, 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. 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 be related 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.

Searching the State Space to identify faults in Figure 2 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 2 above, indistinct.

The data in Figure 3 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 2013 is 80 years.

Figure 3: 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.

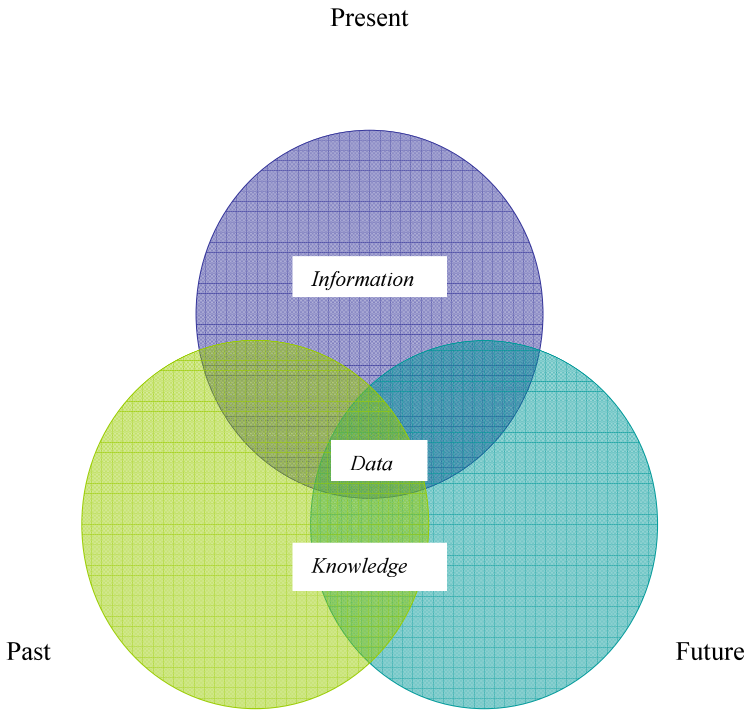

Figure 4: Temporal View of Data, Information, Knowledge (Venn diagram) and Big Data

Figure 4 (Extended from Graham, 2014) adds a temporal dimension. As shown in the Life Insurance Example (Figure 3), 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). It is suggested that knowledge synthesis, on the other hand, 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 information is 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).

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 4 suggests the temporal relationships between data, information and knowledge. Big Data is omnipresent and is therefore not shown in Figure 4. The suggested steps involved are the presentation of external data (facts) and their organisation into information and subsequent analysis to knowledge.

Discussion and Conclusions

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: 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".
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 (e.g. structured, unstructured, etc.).

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 2) 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. 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 4 above 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/statistical analysis and domain application that may convert Big Data into Big Knowledge. It is questionable if Big Information 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 3) 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? This question requires further research.

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.

Correspondence

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

References

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

Graham, D. (2014). 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. (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*, 4^{th} 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); available online: 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.