Big Data Critical Thinking Skills for Analysts—Learning to Ask the Right Questions

Earl H. McKinney Jr.  
Bowling Green State University  
emckinn@bgsu.edu

Bethany D. Niese  
Bowling Green State University  nbiese@bgsu.edu

Abstract

Big data analysis is in high demand; however, little work has identified important critical thinking skills for big data analysts. Here we suggest these critical thinking skills include understanding the limits of measurement and representation, the challenges of the uniqueness of information, and of underdeterminism, and the difficulty of selecting an appropriate software tool. In addition to explaining these topics and their impacts on analysis, we provide example questions for each and suggested overall learning outcomes as well as need for further research.

Key Words

Big data, critical thinking

Introduction

Big data is having a significant impact on business. It is the result of two trends: the cost of storage has plummeted and, revolutionary ways to analyze and interpret data have been developed. As a result, mind numbing amounts of data are flooding the world. Every 48 hours we create as much data as we did from the dawn of civilization to 2003 (Siegler, 2010). For example, every day Facebook produces 10 Terabytes (TB) of data, Twitter contributes another 7TB, and Google processes 24,000 TB of data; Wal-Mart collects more than 2.5 Petabytes of customer transaction data every hour and there are over 4.6 mobile phone subscriptions generating terabytes of data daily (Bollier, 2010; Demirkan & Delen, 2013).

While these numbers are growing at rates that are difficult to keep up with, the main sources of data have remained the same over that past decade. One source stems from scientific advances such as the mapping the human genome and global climate; another from communications—every tweet, post, click, like, email, comment, search and digital media files; and a third source from the healthcare industry with blood tests, MRIs and hospital visits. However, a fourth source of data may make the first three seem almost insignificant. Estimates suggest that the number of devices that generate or store data—e.g. mobile devices, Radio Frequency Identification (RFID) chips or sensors—will soon total more than several trillion. Hyped as the “Internet of Things”, these devices will instrument the world we live in and will sustain the growth of Big Data for many years to come.

New technologies have always changed how work gets done and how we train students for this world. In addition to skills necessary to leverage big data, higher order cognitive skills are also needed to turn data into information. Unfortunately, little work has been done to identify and classify these skills into framework suitable for educational purposes.

Here our objective is to identify and classify a preliminary set of five critical thinking skills. These higher order skills are technology independent and reflect a deeper awareness of the assumptions on which big data analytics rests. We expect this list to grow and become more refined in coming years, the goal of this work is to nominate an initial set with criteria for choosing more.
We have used most of these five topics in senior level/masters level courses on big data in recent years. Employers had persuaded us to include them as they had found many new hires were naïve about important assumptions on which analysis rests. As a result, we have added these topics to our Business intelligence courses which previously consisted of existing technology and statistical tools. We made this change despite incoming student expectations that hands-on training and software use are most needed.

We use exercises relating to these new topics throughout the second half of the course. By waiting until the second half these lessons can be applied to the analysis and projects the students accomplished in the first half of the course. We also attempt to discuss a different topic each week. The overall goal of helping students build an awareness of the many assumptions that underlie Big Data analysis is achieved via repetition—each week a different set of assumptions are uncovered.

**Background**

Many articles address the technical, managerial and social challenges of big data. While little work has identified specific critical thinking skills, the current literature does help us more clearly understand the current challenges with big data from which we define the five critical thinking skills needed by our college graduates.

Most of the challenges of big data address technical and statistical issues. These challenges include designing queries and data structures, how data is prepared for analysis, scaling limitations of data, how to use statistical and other analysis tools are used and how to create and share visualizations (Kaisler et al. 2013).

For example, in a recent Science article, Lazer et al. (2014) highlighted the dangers of ignoring construct validity, reliability and dependencies among data in their analysis of why Google Flu Trend forecasts have consistently been inaccurate. They also warn about data creation as Google changed the way its search algorithm works which impacts how end user searches are classified and recorded.

Other recent research has begun to identify managerial challenges. Kaisler et al. (2013) have suggested big data analysis often suffers from a lack of standards in data collection and storage, and a desire for quantity of data rather than quality. Other issues include ownership of the data (Kaisler, Mondy & Choen 2012), and compliance and security (Gantz & Reinsel (2011). Ayres (2007) also highlights that context is vital to analysis, as hybrid data sets often mislead. Finally, Ritchey (2005) highlights the big data management challenge of structuring socially messy problems that come with incomplete, contradictory and changing requirements.

Big data challenges also include societal and regulatory issues. Questions include how to tradeoff privacy and security, how to promote speech on online communities, and how to address cultural biases in analysis (Boyd & Crawford, 2012) and the challenges of privilege and digital divides (Chan, 2015).

Before specifying our critical thinking skills, we should explain how we use the term critical thinking. We adopt the definition of Scriven & Paul (1987):

> Critical thinking is the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action. It is based on universal intellectual values that transcend subject matter divisions: clarity, accuracy, precision, consistency, relevance, sound evidence, good reasons, depth, breadth, and fairness.

While little research has examined the intersection of Big data and critical thinking skills, others have identified the importance of critical thinking in traditional data analysis, particularly in the analysis of accounting data. Kern (2000) calls for accounting exercises to enhance critical thinking skills while Stanley and Marsden (2012) report on the critical thinking benefits of problem based learning in accounting. This paper makes the argument that critical thinking skills are just as important, if not more so, in big data analytics.
Criteria for critical thinking skills

To provide a context for the five critical thinking skills, we briefly describe the criteria we used to select, improve and retain them.

Each critical thinking skill should:

1. be applicable to most big data applications in any business context. Many other critical thinking skills also reflect critical thinking, but may only apply to particular contexts such as health, auditing, privacy or financial concerns.

2. be technology agnostic. Again many other critical thinking skills are necessary to understand how particular software and database technologies manipulate data. However, here we focus only critical thinking skills that can be applied to any analysis.

3. address higher level Bloom taxonomy (1956) educational objectives of evaluate and create. Rather than the lower taxonomic objective such as know and understand, these critical thinking skills require students to evaluate limitations and create unique insight.

4. require no prior knowledge or practice. Traditional students tend to have limited applicable, professional experience so the skills need to be targeted toward this audience. Critical thinking skills not considered are those that require a technical background or more training in communication or visualization limitations.

Critical thinking skills

Each of the five critical thinking skills is specified below. We explain each, and provide examples of questions for each in the following section.

Five Critical Thinking Skills for Big Data

1. Understand the assumptions necessary for measurement and be able to apply these to analytics.

2. Understand the assumptions necessary for representationalism and be able to apply these to analytics.

3. Be able to explain how insight from data is unique to each individual and the implications of this view.

4. Be able to explain underdeterminism and its implications for analytics.

5. Be able to create a solution based on business process needs and available analytical software.

1. Understand the assumptions necessary for measurement and be able to apply these to analytics.

All big data starts with measurement and the assumptions underlying all measurement are typically underappreciated (Stonebreaker & Hong, 2012). Measurement is the process of assigning numbers or other symbols to things. Measurement assigns numbers to represent magnitudes of pre-existing quantities in proportion to the magnitude. Measurements travel well—they are easy to create, store and transmit widely and they enjoy wide acceptance. While text, dimensions, and other non-quantitative measures can be used in big data, most data, including geographic and mobile measures are quantitative.

Several distinctions about measurement are important to understand as with each distinction, different assumptions are made. Not all of these distinctions can be presented in a limited article. One primary distinction is that measures are either direct (or fundamental) or indirect (or derived). A direct measurement means an attribute of an object can be inspected immediately, for example you can actually pick it up and measure the sides, such as a box, or count items, such as an inventory. An indirect measure is something like the volume of a rock, where you need a procedure or technique to measure it, like
dropping a rock in a container with water, and see how much the water rises. Surveys, test results, subjective ratings are examples of indirect procedures in business. The procedure for creating an indirect measure introduces assumptions that are often overlooked such as how surveys were constructed or ratings combined or dealing with non responses.

A second distinction classifies measures as real and abstract. A real measure represents a physical attribute of an object such as size or weight. An abstract measure is not related to a physical property, but an abstraction such as depreciation, correlation or rank. As with indirect measure, abstract measures introduce assumptions that are often taken for granted, such as why a method to calculate depreciation was chosen, or how correlation was tested.

**Questions to ask:**

1. If a measure is indirect are assumptions about the procedure overlooked?
2. If a measure is abstract are assumptions and choices made to create the measure evaluated?
3. Why did someone measure this, why did they not measure this object another way, what was their goal?
4. Measures are often treated as if they are “true” and objective and exclusive so ask what judgement was used to select a scale or starting and ending points, what were other choices?
5. Measures of people introduce self-awareness bias, ask if subjects knew they were being recorded or measured.
6. Are the arithmetic limits of the scale considered (don’t average rankings or ordinal data, 80 is not twice as hot as 40 degrees)?
7. Measures are not the thing being measured, they objectify it. Pornography egregiously objectifies unique individuals. Are we confusing the measurement of the attribute for the thing itself?

2. **Understand the assumptions necessary for representationalism and be able to apply these to analytics.**

Once a measurement is recorded it can be used as a sign of something. Signs and the objects they represent are the subject of semiotic theory (Peirce, 1907). A sign and its object create meaning; they are a model of something to someone (Floridi, 2005; Zoglauer, 1996). In representation theory, a sign represents something else (Beynon-Davies, 2009; Stamper, 1985; Vigo, 2011). For example, when dark clouds (sign) portend a storm (object) to a hiker (observer), the dark clouds are information about the storm for the hiker. Similarly, when an end user (observer) examines an icon (sign) of a dollar bill (object); the icon is information about the dollar bill for the end user. Common IS signs that represent objects include topologies, E/R diagrams, database records, and hashtags in social media.

However, in use, a representation becomes a more vague, less rigorous idea. While the theory requires that a sign, object, and observer be specified; often only a sign is specified. A representation becomes the ring of the doorbell, the dark clouds, the accounting number, or the E/R diagram. When individuals do this, they frequently assume the object that they believe the sign represents is known by all.

Data are signs and by themselves data do nothing. Only when individuals link the data, the sign, to objects is meaning created. Every decision made using data requires this step, but often the assumptions about which sign goes with which object are ignored.

**Questions to ask:**

1. What object is the sign (the data) applied to?
2. What other signs might be better for that object?

3. Is it more appropriate to apply this sign (a $7.00 sale) to other objects (to habit, to great sales reps)?

4. Why do we think this sign and object is appropriate, do we attribute this sign to that object out of tradition or culture perhaps?

3. **Be able to explain how insight from data is unique to each individual and the implications of this view.**

Measures lead to signs which can be linked to objects, but individuals, both analysts and decision makers, create their own meaning by choosing different combinations of signs and objects (Checkland & Holwell, 1998; Boland, 1987; MacKay, 1969; Dretske, 1981). Different choices for signs and objects create different meanings. Boland (1987) defines meaning as a change in a person from an encounter with data that changes knowledge beliefs, values or behavior.

The meaning an individual conceives likely differs from the meaning other individuals conceive. At times these differences can be minor and intersubjective agreement is reasonably assumed, but at other times these differences can be significant, surprising, useful, and informative.

For example, for one database designer dirty data (sign) means more money is needed to improve a poor cleansing process (object), to another it means more fines should be assessed to motivate lazy employees (object).

Different people will create different meaning. A good way to understand this reflected in the saying, “The map is not the territory.” A territory (object) can be displayed in a variety of maps (sign), and different individual map makers will choose to create maps with different meanings.

**Questions to ask:**

1. How is my interpretation of the data unique to me, are others interpreting the data differently?

2. How are others creating meaning, what patterns am I missing?

3. Will different people use that same sign for another object?

4. How can we reach consensus on the meaning of a sign and object?

4. **Be able to explain underdeterminism and its implications for analytics.**

Underdeterminism, more precisely, underdeterminism of scientific theory, is a concept in philosophy of science that states that the evidence available to us at a given time may be insufficient to determine what beliefs we should hold in response to it (Babich, 1993). In other words, for any data set, a theory that explains patterns in the data underdetermines the data, there will always be other, at least equally plausible, explanations for the observed patterns. This epistemological challenge in science has profound implications for the confidence that any theory will provide the truth about the observed pattern. This challenge is not limited to scientific theories, but also extends to any conclusion about any data set.

Therefore any big data analytic observation about patterns in a data set must be carefully stated. We can only conclude the presence of a pattern, and cannot rule out that there are other explanations for the pattern.

One way epistemologists explain this is, once again, with the metaphor of map making. Many “true” maps can be created for a territory as many display distinct or different attributes about the territory such as resources, roads or populations. All can be true. Further, underdeterminism means that maps can contradict each other and still be “true”.

---

Critical Thinking Skills for Big Data Analysts
Common expressions such as “the analysis proves” or “the numbers speak for themselves” while frequently used in analysis indicate far greater confidence in analysis than is warranted when underdeterminism is recognized.

**Questions to ask:**

1. Do I realize that there are other explanations as plausible as the one presented?
2. What is our stopping rule, when should we quit looking for other explanations?
3. For patterns explained to me, do I simply accept them as true, or do I realize they are only one of many possible explanations?
4. Are we overstating conclusions, such as...the data says, this proves that...?

5. **Be able to create a solution based on business process needs and available analytical software**

Vendors continue to provide an increasing number of technical solutions. According to the 2016 Gartner Magic Quadrant for Business Intelligence, there are 24 competing software manufacturers, each with many levels of solutions (Gartner, 2016). With such an overwhelming number of solutions, today’s analysts need to be sufficiently aware of the assumptions necessary to use each solution, their challenges, strengths, needs and audiences so the appropriate toolset can be chosen. And with such a large number of new solutions, analysts need to be able to anticipate the challenges of using a particular solution without a complete understanding of all the attributes of the software.

In addition, analysts need to understand the business process from which the data emerges. Again, analysts examine data from a wide variety of processes and their knowledge of the process will always be incomplete. The challenge for the analyst is to know when they know enough about a process to appropriately select a software solution.

**Questions to ask:**

1. What are the primary areas in my company that can most benefit from conversion from data into information?
2. Who are the primary audiences and how do they interact with the data? For example, if the audience is C-level executives, they’d likely need visualization functionality that can drill down into underlying data.
3. How well are the business processes linked to these primary areas been understood?
4. What is our current level of data quality? Do we have a centralized database or information silos? If the data quality is high, most of the work to match software to needs will be on the front-end. However, if the data is unstable or in multiple locations (which increases the chances of data collisions) then a stable back-end solution will first need to be determined.

**Learning Outcomes**

While these five critical thinking skills help students recognize the challenges of analysis and how to ask insightful questions, they also help achieve more specific learning outcomes. Here we describe the four learning outcomes we currently use. For simplicity they are written here as they would appear in a syllabus or class exercise.

1. Understand the Difficulties of Effective Analysis. Students often underappreciate the assumptions made in conducting an analysis.
2. Understanding the Limits of Analysis. Students typically overstate the certainty of analytical conclusions. Common expressions such as “the analysis proves” or “the numbers speak for themselves” are understandable, but indicate far greater confidence in analysis than is warranted.

3. Flourish in Uncertainty. Most students assume the intellectual posture that problems have a finite or given amount of information. Students believe that if they would get that information they will be done. Many academic exercises reinforce this—read a case study, use the “information” there and make a decision, follow the syllabus in order to receive a good grade. In contrast, we believe the chief characteristic of the authentic business context is that the uncertainty is infinite; the potential data is overwhelming and very little information is ever created from it. The more uncertainty in a context, the greater the need for analysis and the greater variety of conclusions. Our educational outcome is for you to appreciate the value of uncertainty in actual business contexts and limits it places on confidence in the results of one analysis.

4. Understand the Value of Curiosity. Analytical insight is a creative effort. This creative effort is fueled by curiosity. When you quit being curious about the data, you quit being helpful. Stay curious, you were born that way for a reason.

Caveats

While our goal is to help students understand some of the assumptions and limits that undergird big data analysis, we are not at all questioning the value of need to conduct effective analysis. We simply are echoing calls from the workplace that our students not only be able to conduct analysis, but understand the limits of the analysis.

We also recognize that this list of five critical thinking skills is only a start. More critical thinking skills need to be examined and shared with students. The limitations of big data analysis needs some type of model or taxonomy so that students can conduct more structured evaluation of analysis and new skills can be identified. Other key critical thinking skills may involve privacy, how to manage iterative analysis, how to state a clear problem or question and gain buy-in from stakeholders, how to articulate or explain the assumptions or uncertainty of the data to the data consumers. Finally, we recognize one key limitation to our list of critical thinking skills. Analysts often do not have access to how measurements were collected for the data they are analyzing.

Big data analysis may appear to remove ambiguity but only shifts it. Some questions are answered but more are created. Understanding the ambiguity of business and accounting data is unlike engineering and math. Even with the biggest data there will be uncertainty and with the uncertainty the need for judgment, choice, trust, and caution.

Conclusion

In the current age of big data, analysts need not only the skills to analyze data, but higher order critical thinking skills about analysis. Here we initiated the conversation about these skills by explaining five—understanding the limits of measurement and representation, the challenges of the uniqueness of information, of underdeterminism, and of selecting an appropriate software tool. Our goal is to help budding analysts ask insightful questions and apply them to business goals, opportunities and issues. The critical thinking skills are presented here as well as example questions from each and the overall learning outcomes (goals) and need for further research.

REFERENCES


Zoglauer, T. “Can information be naturalized?” In Information: New Questions to a Multidisciplinary Concept, K. Kornwachs and K. Jacoby (Eds.), Berlin: Akademie-Verlag, 187-207.