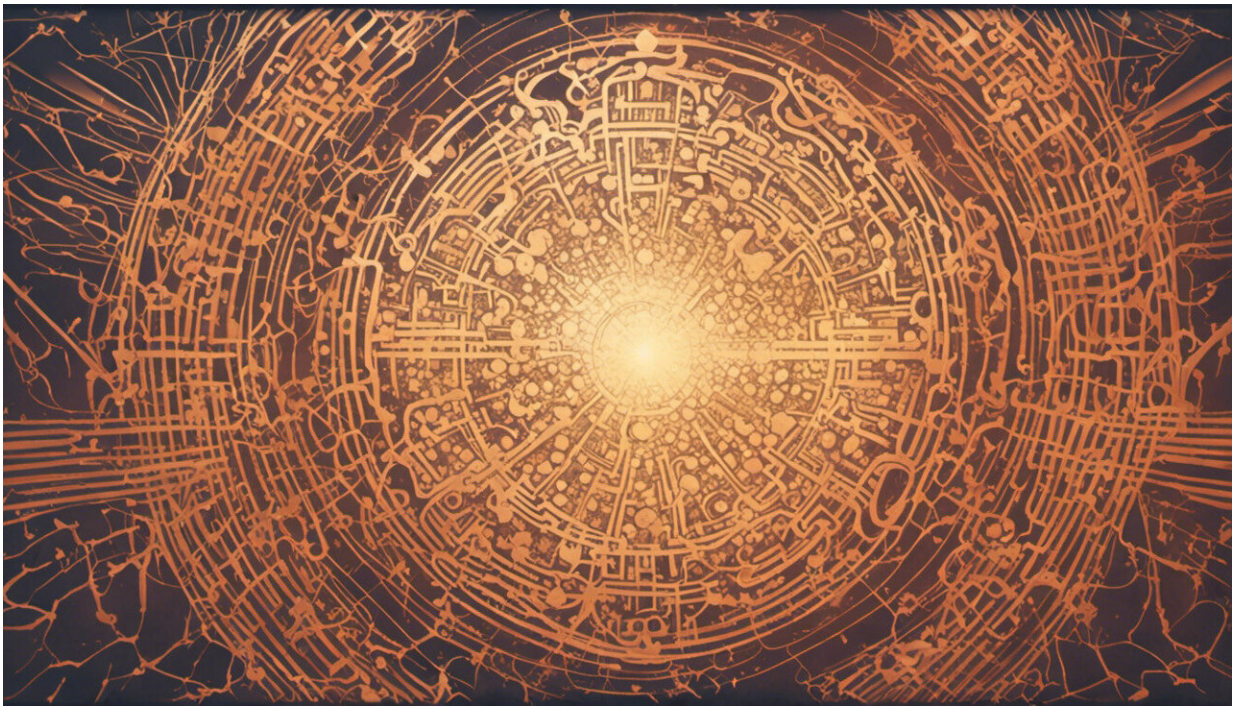
Use it or lose it—the search for enlightenment in dark data

December 22 2016, by Shazia Sadiq



Credit: AI-generated image (disclaimer)

Big data is big news these days. But most organisations just end up hoarding vast reams of data, leaving them with a massive repository of unstructured – or "dark" – data that is of little use to anyone.

Given the potential benefits of big data, it's crucial that we find better

ways to gather, store and analyse data in order to make the most of it.

Stories of big data successes have triggered significant investments in big data initiatives. This has prompted many organisations to gather significant volumes of external and internal data into so-called "data lakes". These are repositories that contain data in any format, whether structured, like databases, or unstructured, like emails or audio and video.

As a result, the growth in the amount of data being generated, collected and stored continues at an exponential rate.

But according to a recent IBM study, more than 80% of all data is inactive, unmanaged, often unstructured, lacking meaningful metadata, and even unknown to the organisation. The proportion of this dark data is expected to reach 93% by 2020.

For example, data generated from vehicle on-board devices can be expected to reach 350MB of data every second. Where does all this data go and who is using it?

Organisations can also generate significant internal data. For example, a recent study found that a company with 1,500 employees had around 2.5 million spreadsheets, each of which were only used by 12 people on average.

What's more, there is evidence of a variety of unstructured data such as document versions, project notes and emails that is left behind from organisational processes and subsequently sits dormant in data servers.

Use it or lose it

Lessons learnt from years of research in information system use have

shown that the assumption that "more is better" when it comes to data is unfounded.

Even in traditional IT projects that follow carefully crafted analysis and design life cycles, the misalignment between perceived and actual value has been a notoriously difficult problem, often leading to poor returns on investment.

In big data projects, the data can often be externally sourced with little or no knowledge of its schemata, quality or expected utility. Thus the risk of making investments that will not deliver is greatly heightened.

The old adage of "use it or lose it" is by no means obsolete, and brings attention back to the purpose of how we use big data. Organisations may retain data for a variety of reasons, including data retention regulations, but perceived future value is typically the main reason.

Although storage is relatively cheap, given the volume of data being assimilated, the maintenance and energy consumption of data centres is not trivial. Furthermore, there are costs and risks related to the security of such unmanaged data.

Thus defining the purpose is pivotal to ensure that big data investments are targeted towards a meaningful problems, and data collection and storage is well justified.

Approaches such as design thinking, which encourages people to use creative solution-focused thinking, are proving to be highly successful in genuine problem formulation for big data.

When appropriately applied, design thinking can equip data scientists to bring together desirability (customer need) and viability (business value) with technological feasibility, and thereby guide them towards

developing meaningful solutions.

Garbage in, garbage out

When the gap between data creation and use becomes larger, it makes it more likely that data quality decreases. This means an organisation will have to employ a lot of effort cleaning old data if it wants to use it today.

According to the US Chief Data Scientist DJ Patil:

"Data is super messy, and data cleanup will always be literally 80% of the work. In other words, data is the problem."

Earlier this year, a group of global thought leaders from the database research community outlined the grand challenges in getting value from big data. The key message was the need to develop the capacity to "understand how the quality of that data affects the quality of the insight we derive from it".

The golden principle of "garbage in, garbage out" is still true in the context of big data. Without scientifically credible knowledge that provides the ability to efficiently evaluate the underlying quality characteristics of the data, there is a significant risk of organisations and governments accumulating large volumes of low value density data, or investing in low return-on-investment data products.

Moreover, the lack of knowledge on the underlying data (distributions, semantics and other nuances) could result in analytical traps, where the data analysis can lead to erroneous, and possibly dangerous, conclusions.

Data exploration is emerging as a promising approach to empower users with exploratory capabilities to investigate the quality of the data and gain awareness of data's shortcomings in terms of their intended use, and

do so before they invest in expensive data cleaning and curation tasks.

The search for enlightenment from the data deluge will consume the energy and investments of the data-driven society in the foreseeable future. Whereas there is immense power in the scale of data, when left unattended will propel organisations into the abyss of dark data.

All this underscores the growing need for well-trained data scientists who have the ability to articulate a well-justified business, scientific or social purpose and align it with the technological efforts for data collection, storage, curation and analysis.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Use it or lose it—the search for enlightenment in dark data (2016, December 22) retrieved 3 May 2024 from https://phys.org/news/2016-12-itthe-enlightenment-dark.html