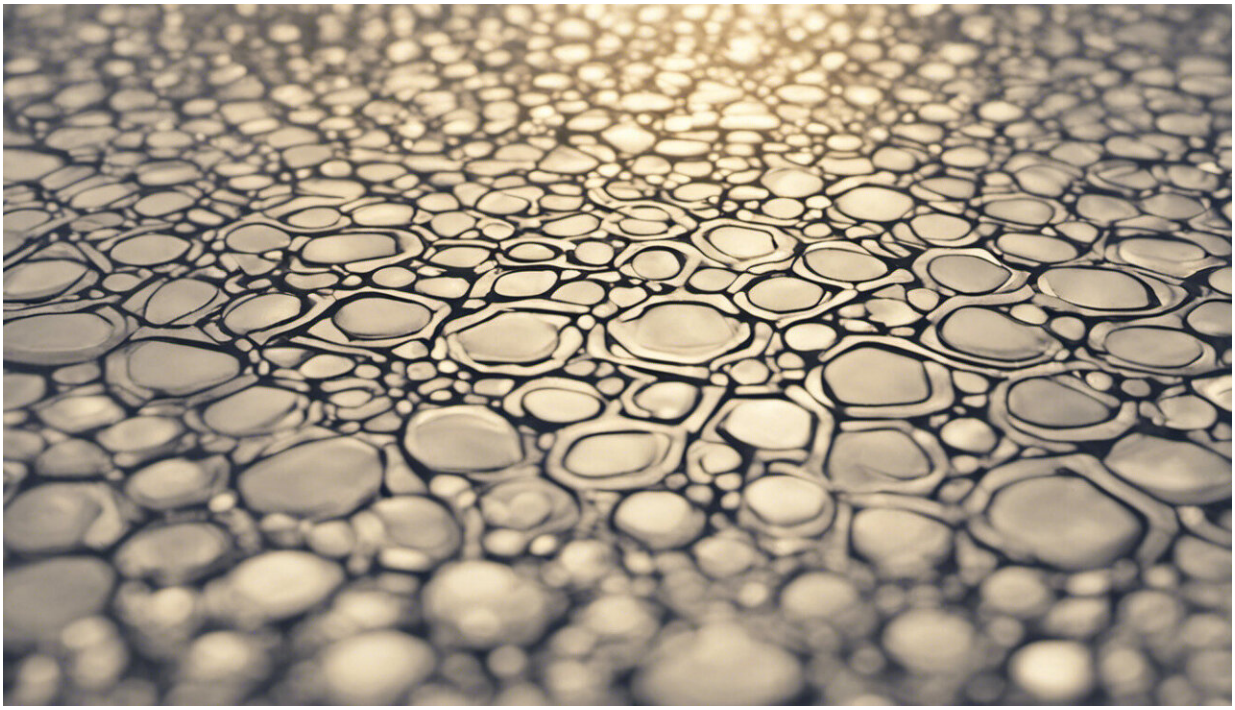
Size doesn't matter in Big Data, it's what you ask of it that counts

March 17 2016, by Malte Ebach, Unsw Australia



Credit: AI-generated image (disclaimer)

Big Data is changing the way we do science today. Traditionally, data were collected manually by scientists making measurements, using microscopes or surveys. These data could be analysed by hand or using simple statistical software on a PC.

Big Data has changed all that. These days, tremendous volumes of information are being generated and collected through new technologies, be they large telescope arrays, DNA sequencers or Facebook.

The data is vast, but the kinds of data and the formats they take are also new. Consider the hourly clicks on Facebook, or the daily searches on Google. As a result, Big Data offers scientists the ability to perform powerful analyses and make new discoveries.

The problem is that Big Data hasn't yet changed the way many researchers ask scientific questions. In biology in particular, where tools like genome sequencing are generating tremendous amounts of data, biologists might not be asking the right kinds of questions that Big Data can answer.

Questions

Asking questions is what scientists do. Biologists ask questions about the living world, such as "how many species are there?" or "what are the evolutionary relationships between rats, bats and primates?".

The way we ask questions says a lot about the type of information we use. For example, systematists like myself study the diversity and relationship between the many species of creatures throughout evolutionary history.

We have tended to use physical characteristics, like teeth and bones, to classify mammals into taxonomic groups. These shared characteristics allow us to recognise new species and identify existing ones.

Enter Big Data, and cheap DNA sequencing technology. Now systematists have access to new forms of information, such as whole genomes, which have drastically changed the way we do systematics. But

it hasn't changed the way many systematists frame their questions.

Biologists are expecting big things from Big Data, but they are finding out that it initially delivers only so much. Rather than find out what these limitations are and how they can shape our questions, many biologists have responded by gathering more and more data. Put simply: scientists have been lured by size.

Size matters

Quantity is often seen as a benchmark of success. The more you have, the better your study will be.

This thinking stems from the idealistic view of complete datasets with unbiased sampling. Statisticians call this "n = all", which represents a data set that contains *all* the information.

If all the data was available, then scientists wouldn't have the problem of missing or corrupted data. A real world example would be a complete genome sequence.

Having all the data would tell us everything, right? Not exactly.

From 2004 to 2006, J. Craig Venter led an expedition to sample genomes in sea water from the North Atlantic. He concluded he had found 1,800 species.

Not so fast. He did, in fact, find thousands of unique genomes, but to determine whether they are new species will require Venter and his team to compare and diagnose each organism, as well as name them.

So, in answer to the question: "how many species are there in this bucket of water?", Big Data gave the answer of 1.045 billion base pairs. But

1.045 billion base pairs could mean any number of species.

Size doesn't matter, it is what we ask of our data that counts.

Wrong questions

Asking impossible questions has been the bane of Big Data across many fields of research. For example, Google Flu Trends, an initiative launched by Google to predict flu epidemics weeks before the Centers for Disease Control and Prevention (CDC), made the mistake of asking a traditionally framed question: "when will the next flu epidemic hit North America?".

The data analysed were non-traditional, namely the number and frequency of Google search terms. When compared to CDC data, it was discovered that Google Flu Trends missed the 2009 epidemic and over-predicted flu trends by more than double between 2012 and 2013.

In 2013, Google Flu Trends was abandoned as being unable to answer the questions we were asking of it. Some statisticians blamed sampling bias, others blamed the lack of transparency regarding the Google search terms. Another reason could simply be that the question asked was inappropriate given the non-traditional data collected.

Big Data is being misunderstood, and this is limiting our ability to find meaningful answers to our questions. Big Data is not a replacement for traditional methods and questions. Rather, it is a supplement.

Biologists also need to adjust the questions aimed at Big Data. Unlike traditional data, Big Data cannot give a precise answer to a traditionally framed question.

Instead Big Data sends the scientist onto a path to bigger and bigger

discoveries. Big and traditional data can be used together can enable biologists to better navigate their way down the path of discovery.

If Venter actually took the next step and examined those sea creatures, we could make a historic discovery. If Google Flu Trends asked "what do the frequency and number of Google search terms tell us?", then we may make an even bigger discovery.

As we incorporate Big Data into the existing scientific line of enquiry, we also need to accommodate appropriate questions. Until then, biologists are stuck with impossible answers to the wrong questions.

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

Source: The Conversation

Citation: Size doesn't matter in Big Data, it's what you ask of it that counts (2016, March 17) retrieved 17 July 2024 from https://phys.org/news/2016-03-size-doesnt-big.html