What Big Data does not know and the consequences for protecting biodiversity

Darko D. Cotoras, PhD
California Academy of Sciences

The recent accumulation of large databases, development of statistical methods and availability of computational resources has opened the “new” field of Data Science or “Big Data”. Using these novel techniques, we can explore huge amounts of data to find patterns and correlations that were hidden from us before. This has proven to be the case in a wide variety of fields, from market studies to microbiology. There is a lot of enthusiasm about all the possibilities that this new field offers, but what Big Data does not know?

The answer is rather trivial. Data Science is useless without data. It is therefore important to wonder: which questions do we not have data to answer? Even more importantly, which questions do we lack sufficient data to answer, yet still try to answer them anyway? This situation is present in many fields, without being the exception studies in biodiversity and conservation. Historically, there has been a vast tradition of exploration and documentation of the natural world, which gave us a current total of 1.9 million described species. However, theoretical estimations could go up to more than 11 million. This discrepancy between described and theoretical estimations is a small example of how little we know about the incredible diversity in natural world. Our state of the knowledge looks even more precarious if we consider the availability of data about natural history, physiology or genomics. There still a lot of foundational work to do.

Species that we have named correspond to a non-random sample of the biodiversity. So far, we have information on organisms which were the easiest to collect and study. In other words, as expected, we started with the "low hanging fruits". The small and hard to find organisms are still “black matter” for biologists. They have an effect on the ecosystem, but we do not know what they are.

This lack of knowledge could have deep negative consequences for us and the rest of life on Earth. Given our current model of economic development, we have affected profoundly the planetary ecosystem, by altering natural cycles, changing weather patterns and causing extinctions. Not having an understanding of the consequences of these disruptive actions is like playing Russian roulette. Without a proper knowledge of the natural world it will not be clear once we will reach the point of no return ("tipping point") and the bullet of environmental collapse will be fired.

Here, is where Big Data comes into play. The information about species distributions, ecological data and weather patterns have proven to have strong predictive power and be informative to explain nature. The issue is that these data sets are far from being complete. Many aspects need to be better documented in order to have more precise and solid understanding of the already well-known general trends.

Documenting biodiversity is critical, as data everyday it is literally been erased by extinctions. It is like rescuing books from a library on fire, but worst. On the case of biodiversity, the books are at the same time the building blocks of the library itself. Therefore, today more than ever, there is an essential need to go into the field, look for new species, and learn more about the ones already discovered.

When the biodiversity research is focused on data sets without the proper curation of an expert on the group, the results should be more than questionable. Moreover, the experience of rearing, collecting and describing species produces a body of metadata that it is always present in the mind of the expert naturalist, but not necessarily coded into the data sheets. After data analysis, proper expectations can only be well evaluated by someone who knows the organisms and has access this wealth of metadata. These insights many times will provide more meaningful expectations than artificially created null distributions.

Reducing species exploration efforts might be a mistake that we will not be able to make up for in the future. In the current biodiversity crisis, the exploration of threatened ecosystems is an urgent duty. But, the excitement for Data Science has shifted attention towards research programs with a lot of statistics and little biology.

As funding is limited, a new field dealing with computers and artificial intelligence, might attract more resources than the naturalistic practice, which for centuries has been already done. The same way as in real species competition, this uneven fight for economic resources is driving professional naturalists to extinction.

All the work involved on intimately getting to know a group of organisms or ecosystem takes time. Time to go to the field, time to visit museums, time to do experiments, time to breed organisms. But in an academic system thirsty for quick publications, the time investment required to create this knowledge will hardly be able to respond to the established productivity goals. For example, to produce a single taxonomic revision, it could take several years of work. While, large data synthesis analysis, which feed on those same revisions, could be done in comparatively less time. Both types of work are essential and complementary, but in the academic market of papers, one is definitely more attractive than the other. While a taxonomic revision is highly specialized and slow to produce, the data synthesis work addresses more general questions and is faster to generate. Therefore, dedication to taxonomy or other naturalistic approaches, could turn into a professional disadvantage, which translates into part of the biodiversity discovery work been discouraged to be done. In parallel to the species extinction, those who can recognize them also go extinct, reducing our capabilities to properly react to climate change.

Big Data is an exceptionally powerful tool, which relies on a well develop body of evidence. Today, as we enter what has been called the Sixth Mass Extinction, the time is little and the data, perhaps not so big…

Acknowledgements: I would like to thank Peter Jaeger and Paloma Medina for suggestions on this essay.

Figure legends:

Figure 1: Despite a long history of biological research, a major part of the biodiversity remains unknown. A large part of Big Data is still on the making. Parque Nacional Queulat, Chile. Photo credit: Darko D. Cotoras

Figure 2: Harvestmen are among the many groups that need more studies. On the picture, a male from the genus *Sadocus* (Gonyleptidae) endemic to the forests of southern Chile. Photo credit: Darko D. Cotoras

Figure 3: Scientists such as Alexander von Humboldt have based their discoveries on fieldwork and natural history. In front of the Museo Nacional de Historia Natural in Santiago, a bust of the explorer recognizes his important role on shaping our current view of nature. Photo credit: Darko D. Cotoras