Naturally digital: the big data of nature

Norfarhan Mohd Assaad[1] & Rahmad Akbar[2*]

[1]Center for Frontier Sciences, Universiti Kebangsaan Malaysia, Bandar Baru Bangi, Malaysia.

[2]Department of Immunology, University of Oslo, Oslo, Norway

[*]Correspondence

Abstract. Nature is, perhaps, the largest source of data known to man. Advancements in DNA sequencing technologies have allowed scientists, in particular those from the field of life science, to accumulate ultra large datasets and ask ever increasingly complex questions. Here, we briefly discuss the great promise, opportunities, and challenges presented by the explosion of data in a world where nature is becoming increasingly digitized.

One most often associates nature with, plants, animals, or geographic topologies (rivers, lakes, oceans and etc.). This "traditional" association, however, is changing very rapidly in particular for modern day life scientists. Thanks to the Next Generation Sequencing technology (NGS) and since the completion of the Human Genome Project in 2003, biological data, in the form of DNA/protein sequences, continues to grow at an unprecedented pace. To appreciate the sheer scale of the data coming out of the life science field, let us use the internet as a measuring stick. It was estimated that up to the year 2016 the internet accumulates a total of 7.7 zettabyte of data. If you are not familiar with the unit zettabyte, in numbers one zetta is a factor of 10^{21}, in words a zettabyte is one thousand billion billion bytes. Now, in 2016 alone, the number of DNA (bases) was pushing 13 quadrillion, a factor of 10^{15} or in words one thousand trillion bases. Bearing in mind that the estimate for the DNA bases is only a one year estimate, whereas, the estimate for the internet data bytes is a cumulative estimate since the 90s, one can easily recognize the grandness of the data coming out of biology and life science in general. Nature is big data.

With big data comes big responsibility...and opportunity. Well, the original quotes actually goes like this: "with great power comes great responsibility" but today, data is almost

equivalent to power and power provides opportunities. Massive corporations such as Google and Facebook center their products and services around data retrieval and, more importantly, data analytics. We are going to share (what now is) a typical online experience. We needed new parts for our bicycles (yes we admit that as scientists we tend to have nerdy personalities but we do go out there to bike and socialize with other humans–smiley). Like anyone would, we went to our web browser to look for these parts and bought it online. A couple of days later, when browsing unrelated things, one of the advertisements that the browser displayed to us was an ad for a bike shop. As we spend more and more time online, we notice that the browser (or more precisely the corporation behind this browser, Google) "understand" our behaviour better and better. What is amazing about this is the fact that this "understanding" is achieved largely via big data analytics. Typically these corporations employ a class of techniques known as Deep Learning. A deep learning model is trained on massive amount of user data allowing it to make tailored and accurate predictions for each user. The phrase artificial intelligence (AI) is often used to describe this type of systems to the general public.

For us life scientists, understanding what advertisements to serve to a particular user may not be of interest. But, questions such as what treatment to give to a patient given his/her previous medical history or what DNA changes (mutation) associated with the survival of a virus (pathogen) can be framed in a similar if not identical fashion to the advertisement problem. That is, we can **deep learn nature** just as we **deep learn user data from the internet**. For example, using deep feed-forward neural networks (a type of deep learning algorithm) Akbar and colleagues identified cavities on the surface of proteins that may potentially influence the way these proteins deliver their functions within cells. Mohd Assaad and colleagues demonstrated the potency of high-throughput sequencing (NGS, big biology

data) in identifying genetic changes (polymorphisms) associated with the fitness and survival of plant pathogens.

Importantly for Malaysia and the neighboring Southeast Asian nations, big data thrives in the agriculture sector. For instance, a collaboration between technology giants such as Fujitsu, a financial corporation ORIX, and agricultural corporations Masuda Seed and Smart Agriculture Iwata in Japan gave rise to a smart farming program that leverages on technology to yield superior produce. Specifically, the smart farming program uses connected-sensors to manage and manipulate the farming conditions. On broader terms, these connected-sensors are a part of a larger family of connected devices known conceptually as the Internet of Things (IoT). In practise, these connected devices collect huge amount of data which is then used to train deep learning algorithms (or any other model) to optimize the farming conditions. Similar collaboration between Fujitsu and Aeon Agri Create was implemented in Vietnam as well. Japan is obviously a high-income country but the country never took its eyes off agriculture. Surprisingly for Malaysia, where agriculture is still one of the core economic activities, there has been very little buzz over smart farming. REDtone, a Malaysian telecommunication and digital infrastructure provider, listed IoT based farming on its website but the details remain sparse. Further, we have yet to see serious initiatives aimed specifically to digitize the nation's agriculture sector from the government or industry alike.

Health care is another major sector undergoing revolution in the age of big data. Companies such as 23andMe which attracted big money (and data hungry) investor such as Google, illustrates this quite nicely. Here is a little snippet we took from 23andMe website:

"You are made of cells. And the cells in your body have 23 pairs of chromosomes. Your chromosomes are made of DNA, which can tell you a lot about you. Explore your 23 pairs today. Find out what your 23 pairs of chromosomes can tell you."

We must admit the allure of knowing everything about ourselves just from our DNA sequences is quite powerful if not sirenical. When we did our Bachelor of Science degrees at University Kebangsaan Malaysia (UKM) in 2007, we experimented with but a tiny fraction of DNA and we were extremely excited about it. Today you can get your entire genome sequenced and analyzed albeit by (data hungry) corporations. The data from personal genome sequencing opens a plethora of opportunities. Chief among them is personalized and precision medicine. In personalized medicine, therapeutic and preventive cares leverage the patient's genetic information and combine it with big data analytic techniques such as deep learning algorithms to come up with a therapy most suitable to the patient. Not too dissimilar to the "what advertisement to serve for this user" problem described earlier.

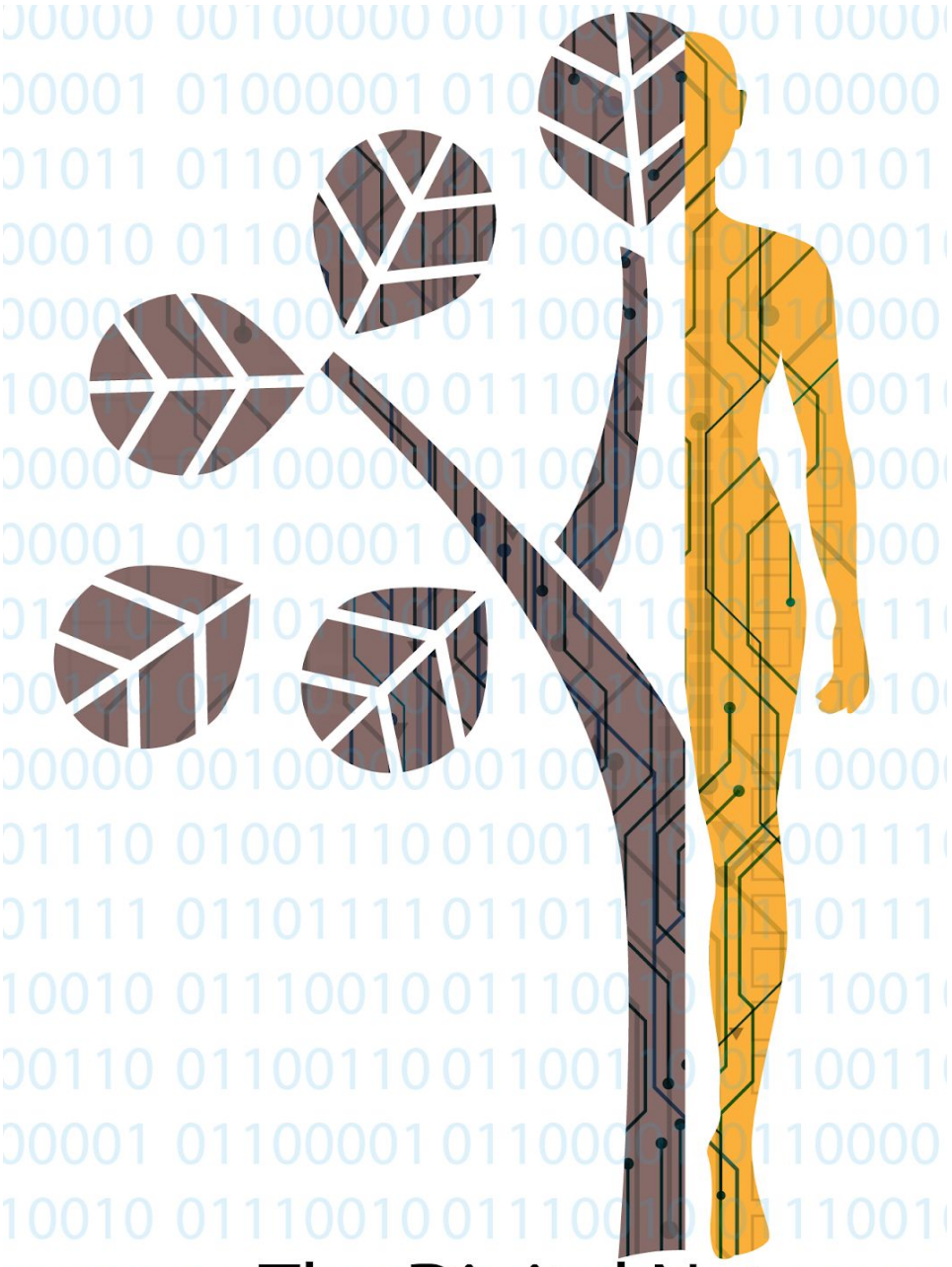
When there is light there must be shadow. Stories of digital revolution in life science do not always end beautifully. The meteoric rise and fall of Theranos, a Silicon Valley based company infamously known for its lab-on-chip blood testing technology, remind us that the digital dream can deprive even the best of us from our senses. There is also a matter of privacy in data management and collection. Much like the heated discussion surrounding how to best handle privacy and user data on the internet, which manifested as the General Data Protection Regulation (GDPR) in the European Union (EU) and has been in effect since May 2018, similar or even stronger arguments must be weighted when the genetic data of the planet is at stake.

Although we only briefly discuss two sectors: agriculture and health care. It is enough to take us to the obvious question: are we, as a nation, ready for the digital nature? From our perspective as scientists and academicians, ready or not, the digital nature is already here. Either we march forward with it or get left behind. Unfortunately, the academic sector (universities and research institutes) appears to still shy away from taking full advantage of big data analytics. From personal communications and observations, many of our life scientists remain anchored to legacy niche research areas and our students receive very little training on the increasingly important computational competence necessary to thrive in the era of digital nature. Do you think we are ready?

Author information

Dr. Norfarhan Mohd Assaad, Center for Frontier Sciences, Universiti Kebangsaan Malaysia, Bandar Baru Bangi, Malaysia. E-mail: n_farhan@ukm.edu.my. Phone (office): 03-89215401

Dr. Rahmad Akbar, Department of Immunology, University of Oslo, Oslo, Norway. E-mail: rahmad.akbar@medisin.uio.no. Phone (office): 22 85 50 50 extension 02770



The Digital Nature

Imense data coming out of the life science field in particular health care and agriculture demands digitalization. Algorithms, computational modeling, and big data analytics are but a neccessity to modern day life scientists.