*Acta Alimentaria, Vol. 49 (1), pp. 1–3 (2020)*
*DOI: 10.1556/066.2020.49.1.1*

**Editorial**

# BIG DATA AND FOOD SCIENCE

Food Science is certainly not immune to the Big Data buzz. Nor should it be. Apart from information, food is the most obvious candidate in the physical world that links people – a true representative of globalisation. Everybody needs it every day; it is grown, harvested, processed, mixed and transported long distances while passing through many hands, cities and countries, before arriving at our dining tables. At all these stages, a vast amount of interactions and processes take place affecting food choice, safety and waste, just to name a few. In the United States, according to the National Sustainable Agriculture Information Service, various produce items travel more than 2000 km on average before reaching the consumer. A case in point was the finding[1], that today the total amount of food crossing international borders is more than the world's total food production. How is this possible? Simply because it has become typical that food items go through many countries before consumption, or even return to their origin after a long journey. Data about the global food transport networks are freely available (see for example the UN-operated ComTrade information centre[2]) but the big question is what to do with and how to make sense of such an abundance of data?

The demand to analyse the data deluge created a new discipline on its own, referred to as data science. The term "Big data" emerged in the first five years of this century, with the completion of the Human Genome project and the explosion in the world's digital storage capacity[3]. Commonly three V-s identify its main attributes: *volume, velocity* and *variety*[4]. Recently a fourth V joined the others: *veracity*, referring to the uncertainty of the data. In the newly emerging age of fake news it has become crucial that the data should be verifiable or at least their uncertainty be quantifiable.

For a quick exercise, let's see an example how big data can be used to get an idea on global questions. Accessing Google Scholar on 18th May 2019, the word "food" returned 6.5 million search results, 1% of which also contained the term "big data". The word "business" returned with ca. 4 million scientific pages, 7% of which contained "big data". The result is almost identical for the search word "politics" instead of business. Even swapping "food" for "health" does not improve this ratio significantly. Intuitively many academics would agree that currently life sciences are lagging behind social sciences in using IT and computational methods, but the above exercise provides an objective quantification to support this view.

Of course, there are some success stories in food data science too. These are: tracking food items, identifying consumer choice and demand, optimizing food distribution and waste.

[1] Ercsey-Ravasz, M., Lakner, Z., Toroczkai, Z. & Baranyi, J. (2012): Complexity of the international agro-food trade network and its impact on food safety. *PLoS ONE, 7,* 10.1371
[2] https://www.comtrade.com/ (last accessed: 4 December 2019)
[3] Hilbert, M. & López, P. (2011): The world's technological capacity to store, communicate, and compute information. *Science, 332,* 60–65.
[4] Wamba, S.F., Akter, S., Edwards, A., Chopin, G. & Gnanzou, D. (2015): How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *Int. J. Prod. Econ., 165,* 234–246.

On the 12th Dubai International Food Safety Conference, the organizing Dubai Municipality proudly showed their impressive "Food Watch System" in action[5].

Consumers can learn all they need (ingredients and their history) about food items to be purchased from retailers who participate in the scheme. The system runs on blockchain technology and is easy to use via a mobile app reading the QR code of the item. Another example for advanced technology is the system applied by the retailer Kroger in the USA, where infrared body-heat sensors and bespoke software track how customers are moving through the store, thus predicting how many cashiers to deploy.

However, not only retail but also food safety services can benefit from big data technology. "Yelp" is a crowd-sourced review website allowing users to submit reviews of local businesses, including restaurants. The reviews can be searched by a text-mining program for keywords such as "sick" or "vomit", from which epidemiologists can find out a possible food poisoning much earlier than the current medical reporting system makes it possible (let alone that ca. 90% of milder food poisoning cases remain unreported). This is an example where consumers' communication is exploited for public health and not simply for marketing purposes.

A much (ab)used field, where big data will be (and already is) vital, is nutrition and healthy diet. Dubious labels and ads on "super foods" can be predatory tricks on unassuming consumers. A new initiative that seems to be destined to become a new omics-discipline could help with this. It is dubbed as "Foodomics" and it will help to make sure that science-based information will be available to accept or reject health benefit claims. The idea behind foodomics is the recognition that, among the body's external stimuli, it is food that is the easiest to control for positive effects on our health.

Indeed, let's see what we, humans, are given and what we can choose regarding our metabolic health. Apart from singular interventions, we cannot change the genome, which determines for example the enzymes produced for the digestive mechanism. The Human Genome Project gave some hope that soon we would be able to read phenotypic traits from our three billion nucleotide base pairs. Still, it was only recently[6] when a complex trait, the susceptibility to obesity at birth, was predicted from a genome-wide polygenic score, integrating "big data" from 2.1 million common genetic variants.

We can, however, change the food, water, drug, etc that we ingest, as well as our physical and social environment - though the latter is much more difficult. All these constitute our "exposome", the subject of another "omics" field. Food is the most plausible part of the exposome, the easiest to change, while it can significantly affect both the gene expressions of the genome and another "-ome": the microbiome. The focus of foodomics is the complex triangular set of interactions between food, the microbiome and the host physiology controlled by the genome.

Medical genetic databases have been providing opportunities for a long time to use big data technologies to study the genome and recently more and more structured databases are available for the microbiome, too[7]. Databases on the bacterial genome, or their kinetic

[5] https://www.dm.gov.ae/en/Business/FoodSafetyDepartment/Pages/Food-Watch-System.aspx (last accessed: 4 December 2019)

[6] Khera, A.V., Chaffin, M., Wade, K.H., ..., Timpson, N.J., Kaplan, L.M. & Kathiresan, S. (2019): Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell, 177,* 587–596.

[7] Marvin, H.J.P., Janssen, E.M., Bouzembrak, Y., Hendriksen, P.J.M. & Staats, M. (2017): Big data in food safety: An overview. *Crit. Rev. Food Sci. Nutr., 57,* 2286–2295.

responses to the environment, such as ComBase[8] have been around for long but not yet for predicting an optimal diet that would prevent certain illnesses and extend life expectancy. A recent article[9] by Albert-László Barabási, probably the most highly cited data scientist, aimed at integrating big data and network methodology in nutrition science. Such events herald a new research field whose findings should help us to judge whether health benefit claims for certain foods are justified.

A key for food science to catch up with the utilisation of big data (as with all new emerging technological fields) is education. The time has long gone when agriculture and food universities could neglect computational methods. It was shown[10] by systematic text mining, that one third of the million-strong biomedical abstracts of recent decades used some sort of statistical P-value. However, after choosing a smaller sample and inspecting their full text, it turns out that the majority of those do not see the link of these P-values with confidence intervals or other quantifications of uncertainty. Based on the personal experience of the author of this article and agreeing with the above authors, the bottle-neck in life sciences is not necessarily the need for more data, more sophisticated statistical techniques or more advanced software packages; but rather the correct interpretation of traditional data analysis concepts, among which the P-values, confidence intervals, etc. are just singled out examples.

To quantify how certain you are in your findings is a basic task in science. Such quantitative investigations will be even more important when analysing and interpreting big data. To generate the required mentality, we would need much more interdisciplinary research, especially for PhD students. The examples when a student has co-supervisors from both "wet" (laboratory or field-oriented) and "dry" (data-oriented) scientific fields are still scarce. It is commonly acknowledged that interdisciplinary research is in fact a bridge-building between two fields, therefore it is time consuming and patience demanding. Under the current publication pressure, there is no time for this during the 3–4 years of a PhD studentship. The recent decline in support for academic institutions (poignantly exemplified by the fact that the university which has been arguably the most successful in interdisciplinary research in Hungary has been forced out of the country) makes the situation even worse. The risk is real that food (and generally life-) scientists will pick up the concept of big data as a superficial buzzword rather than a serious aid for research.

J. Baranyi*
Institute of Nutrition
University of Debrecen

[8] www.combase.cc (last accessed 4 December 2019)
[9] Barabási, A., Menichetti, G. & Loscalzo, J. (2019): The unmapped chemical complexity of our diet. *Nat. Food, 1,* doi: 10.1038/s43016-019-0005-1
[10] Chavalarias, D., Wallach, J.D., Li, A.H.T. & Ioannidis, J.P.A. (2016): Evolution of reporting P values in the biomedical literature 1990–2015. *JAMA, 315,* 1141–1148.
* Jozsef.Baranyi@gmail.com