[The original, German version of this article appeared on May 3, 2021, as part of the AI column “Aus dem Maschinenraum der KI” in the financial section of the German national daily newspaper Die Welt, p. 10.]

[Translated with www.DeepL.com/Translator (free version) - AI technology made in Europe, see https://en.wikipedia.org/wiki/DeepL_Translator, and subsequently polished and modified by the author.]

Messages from the AI engine room

AI and data to fight diseases

Is data scarcity resulting in a standstill or is it an overexploitation of our health data?

By Kristian Kersting and Andreas Maier

In simple terms, one of the most amazing capabilities of Artificial Intelligence (AI) is to detect and evaluate patterns in the real world as well as deviations from and matches to them — in particular when we cannot see them by naked eye or are lost in observation overload. This feature of AI techniques enables more accurate predictive models within climate research and deeper insights into the behavior of bees as well as traffic flows in cities, so that routes can be optimized and CO2 pollution reduced in turn.

But AI is not just about increased efficiency: it saves lives! Medical applications of AI are manifold: AI systems evaluate X-rays, detect suspicious irregularities in ECGs and brain waveforms, they plan and support operations, and they are learning to diagnose diseases, from eye diseases to gum disease or any dental decay. Actually, Jens Baas, CEO of one of Germany’s major health insurances, Techniker Krankenkasse, is convinced that medical findings without the support of a digital system would be considered bad medical practice in just a few years.

The capabilities of AI grow faster, the more data is available to the system as “food”, i.e., for training: from images of benign and malignant skin changes, over meteorological or economic columns of numbers to the local statistics of Corona infections or the game strategies of successful soccer teams. Thanks to the digital revolution, more and more data is generated and collected across the world. According to current estimates, more than 22 zettabytes per year. If you were to load this amount into 1 GB USB sticks and line them up, you would get the length of about 3,000 million soccer fields!

This flood of data is unprecedented in human history. However, just collecting data is not (yet) enough. Modern speech recognition systems, for example, are trained with up to one million hours of speech. But to be efficient, these recordings must be post-processed by humans: Does someone say "dog" or does she say "and"? Sometimes a cough, sometimes a verbal interruption makes the annotation difficult. Every hour of speech data requires about ten hours of annotation. Based on the current minimum wage in Germany of 9.50 Euros, a speech recognition system would require an initial investment of at least 95 million Euros — and then not a single line of AI has even been programmed!

In medicine, things get even more complicated: every cancer patient produces countless data in the course of his or her illness: Laboratory findings, radiological images, pathological incisions or physicians’ letter. But until now, this data has been scattered, varies greatly in quality, is legally protected for good reasons, and is often not machine-readable.

Fortunately, the German Federal Ministry of Education and Research (BMBF) medical informatics initiative is gathering momentum, bringing together data of different types across sites. This will advance AI research and, in turn, medicine.

However, data protection experts warn against this combination. Sure, individual persons cannot be identified on the basis of age and gender. But if the place of residence is known, the group of people can be clearly narrowed down, and DNA, a fingerprint or the voice ultimately make an assignment crystal clear.

For this reason, there are still rather few public data sets in medicine and healthcare for, say, medical imaging — but the fewer the data, the greater the inaccuracies and uncertainties. As with speech recognition, millions of health data are needed for AI systems to show their merits in diagnostics and therapy, as long as we do not manage to make AI less “data-hungry”. But we have to be careful: while voice data for AI system is already predominantly in the grip of companies, this may not be desirable for health data.

Denmark therefore pools its medical data under state control and makes it available to research and companies; other initiatives such as the registered association Medical Data Donors ask patients to donate data to research. This way, protected by high ethical and legal standards, important data can be collected and shared worldwide. However, some AI approaches can actually be trained in an "encrypted" way so that the data never leaves the hospital. This so-called federated learning — participating computers train with local data sets only and do not share the data with each other globally — protects medical data from being accessed by global players. We have to act fast. Large corporations are already fuse medical data into massive data lakes.

Our health is invaluable and must be protected, but at the same time the use of our health data must be enabled, and the not always helpful, sometimes even obstructive requirements of data protection must be carefully monitored. But this balancing is worth the effort: AI gives us all a greater chance of a healthy and happy life.

Kristian Kersting is Professor of AI and Machine Learning at TU Darmstadt, co-director of the Hessian Center for AI (hessian.ai), and winner of the "German AI Award 2019". His AI column "Aus dem Maschinenraum der KI" appears regularly in the German national daily newspaper “Die Welt.” **Andreas Maier** is Professor of Pattern Recognition at the University of Erlangen-Nuremberg — the AI hub for medicine of Bavaria's High-Tech Agenda. He is the Chairman of Medical Data Donors e.V., member of the Steering Committee of the European Time Machine project, and winner of an ERC Synergy project.