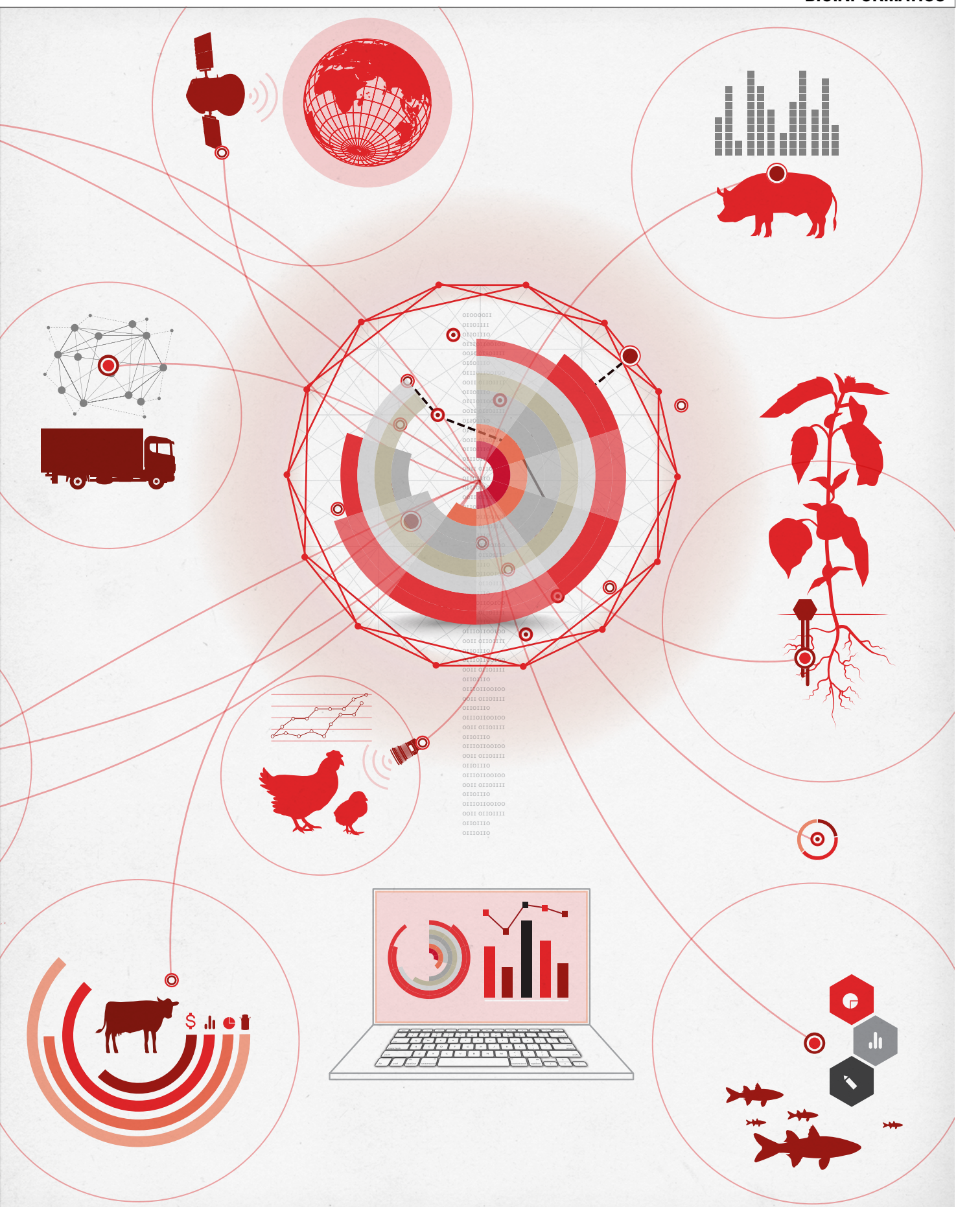
Getting wise to big data

New insights and knowledge are hidden inside the world's fast-growing mountain of digital data. To uncover them, we need smart software and computing power – and an awful lot of data experts. 'In ten years' time, 80 percent of research will be based on the analysis of datasets.'

TEXT RIK NIJLAND **ILLUSTRATIONS** KAY COENEN



At the start of this year, Alphabet became the most valuable company in the world. This was the first time that the ranking was headed by a newcomer that deals in information rather than a manufacturer or oil company, for Alphabet is Google's parent company. The rise of this company – it only entered the top 100 ten years ago – is a clear sign of the value attached to information, IT and datasets.

In numerous fields, the volume of data that is being stored and processed digitally is growing at a dizzying rate. That is happening for example with DNA (see box) but also closer to home.

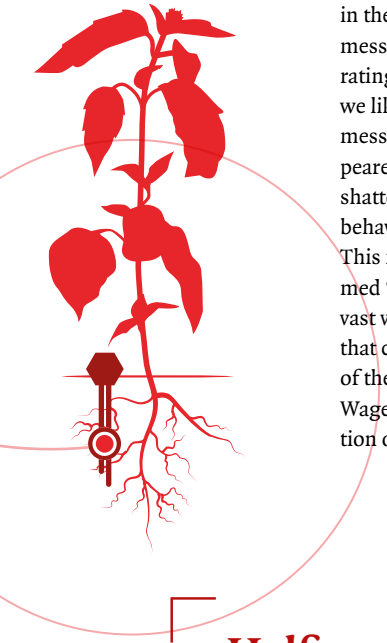
Modern cars send information to the manufacturer day in, day out about RPMs and journey distances, for example. This interesting information can help fine-tune maintenance packages but is also useful input for studies of differences in driving behaviour, a subject that appeals to both researchers and insurers.

STIR-FRYING BROCCOLI

A wealth of knowledge about our preferences is stored in the databases of Google, major retail chains, text messaging, Twitter and car manufacturers, about accelerating fast, what we search for on the Internet or what we like to eat, for instance. An analysis of Dutch Twitter messages showed that the word 'broccoli' often appeared in combination with 'stir-frying'. Not earth-shattering news but an eye-opener on consumer behaviour for the vegetable sector.

This rapidly expanding mountain of digital data is termed 'big data'. The 'big' in big data refers not just to the vast volume but also to Big Brother, the omniscient state that controls everything we do in George Orwell's vision of the future.

Wageningen UR is now taking steps to facilitate exploration of large datasets, for the information being generated in Wageningen's domain is also expected to grow fast. Thanks to modern IT, sensors are already being placed in greenhouses or on tractors to monitor crop growth. All the milking robots, taken together, know almost everything there is to know about hundreds of thousands of cows. Hidden in that mass of data are new insights and knowledge, especially if you can link different datasets, such as data on milk production or feed consumption with genetic information.



Half a million data experts will be needed over the next ten years

TRAILBLAZERS

According to Karin Andeweg, it is crucial for Wageningen to develop this new knowledge further. She and Sander Janssen are the driving force and official trailblazers for big data, which Wageningen UR has declared a priority area. 'Big data seems like a hype but in a few years it'll be commonplace. In ten years' time, we expect 80 percent of research to be based on the analysis and combination of datasets for the generation of new knowledge.'

There are already signs of that future. Wageningen UR is involved in an experiment in Amsterdam in which hundreds of thousands of mobile phones are used to determine where crowds are growing dangerously big on busy days. This experiment also tracks whether the measures taken have any effect. There is no need to send observers out onto the streets.

In the near future, it will no longer be necessary either for researchers to do fieldwork to obtain measurements of crop growth, for instance in wheat fields. In fact, scientists can already use sensors in the field or data from drones or satellites covering the entire wheat field, or several fields at a time, rather than a couple of experimental plots as in the past.

Dealing with such mountains of data is a piece of cake for computers, in principle. If you link the information to data on fertilizer quantities, soil characteristics, spraying and precipitation, you can find out in no time how a particular wheat variety performs under varying conditions. Such research currently takes scientists years as they study various sub-aspects. Farmers will eventually benefit too when all that knowledge is incorporated in useful advice.

A CHANGE OF TACK

The EU thinks that the scientific world will only actually be able to exploit datasets in this way if it changes tack. In mid-April, the European Commission announced that it would be investing billions of euros in data manage-

PHOTO ANP

BIG NUMBERS

Astronomers and physicists have been playing with vast quantities of data for a long time. The results of 600 million collisions per second are recorded during experiments in CERN's particle accelerator. But biologists have been catching up since the turn of the century. The first human genome was deciphered in 2003; this was followed by the thousandth in 2011 and the millionth is expected next year. DNA sequencing capacity triples each year. A comparable data explosion can be seen in proteins and the body's metabolic products. This is turning biology into a data science, says Dick de Ridder, professor of Bioinformatics at Wageningen University. Billions of data items on genomes, genes, proteins and other molecules are combined in huge files and systematically analysed.

'The expectation is that we will be able to read one million billion DNA bases this year,' says De Ridder. 'Bioinformatics specialists are skilled in using those terabytes of data to formulate new biological hypotheses. Whereas biologists currently often outsource data analysis to the bioinformatics guys, I am expecting more and more researchers to sit at the computer working on predictions and then outsource the experimental validation to scientists in the lab.'

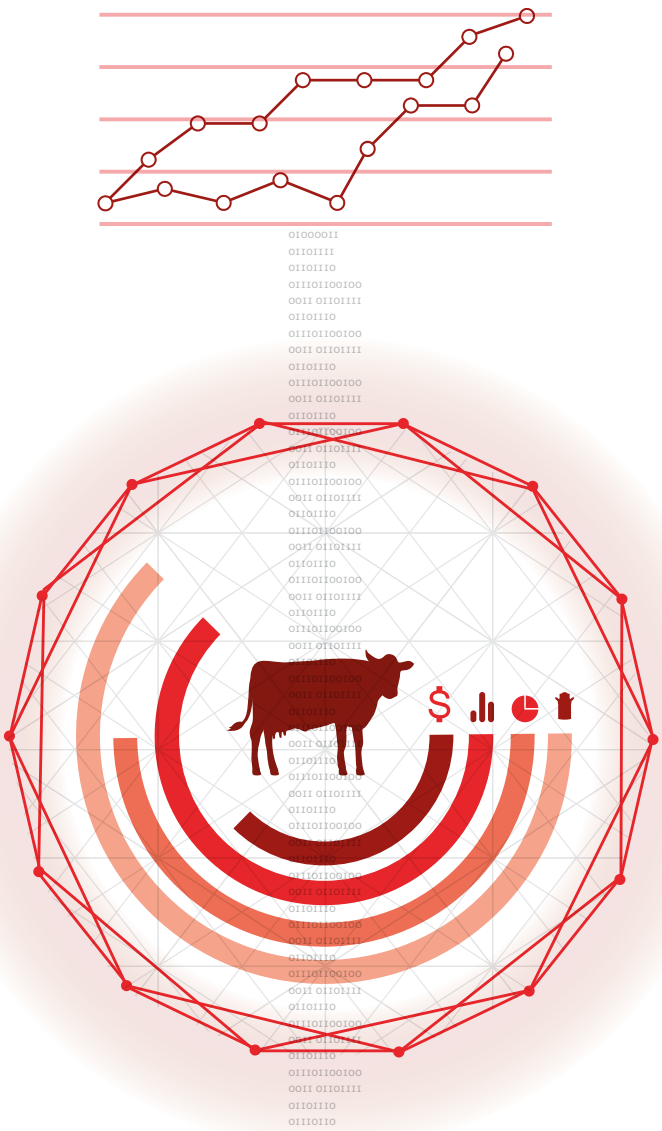
'If you want to keep up, you need to invest'

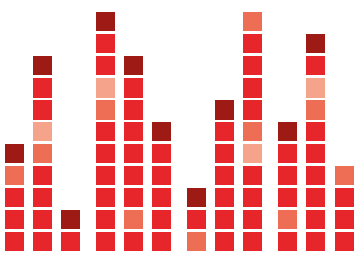
ment. Research results are often not stored properly: scientific journals do not have the space or it may not be possible to access or link the files properly due to the use of different formats. As a result, clever software and plenty of computing power are needed to extract the relevant information. That applies in Wageningen too, says Andeweg. 'We need to invest in hardware, in employees with IT skills and in data scientists. And we need to train people so that they can work in line with this new approach to research.' During Wageningen University's Foundation Day celebrations in early March, Rector Magnificus Arthur Mol called data science 'a new frontier'. Wageningen UR is considering the possibility of setting up a data sciences centre to strengthen that new area of expertise.

At the event, guest speaker Laxmi Parida from IBM Research talked about the options her company offers. A key role is played by Watson, IBM's self-learning supercomputer named after a former company CEO. It is able to process information in no time, search literature from the entire world (40 million documents in 15 seconds), detect associations and appraise the value of information thanks to advanced artificial intelligence. 'To develop computer programs that can reason and learn to solve complex problems, you need the combined expertise of computer scientists and specialists in the discipline in question,' says Parida. 'Those two groups need to constantly work together on programs such as Watson to "teach" it how to solve such problems.' She thinks that data science will become a new scientific discipline that will bring together researchers and companies from different fields. 'It is all about context, context, context.'

HIGH-POTENTIAL TODDLER

'Watson is a toddler that still has to be taught everything, but it also has huge potential,' says Richard Visser, professor of Plant Breeding at Wageningen UR. >





Last year, he contacted IBM Research to suggest using the supercomputer for potato breeding. Visser expects Watson Potato, as the system is known informally, to become an important tool that will make the breeding process more efficient and help get better results. But first Watson has to learn to focus. 'If it sees the word "tuber", it mustn't end up with dahlias,' explains Visser.

DIGGING DEEPER

Once it has been through that learning process, Visser thinks Watson Potato will be a valuable instrument for scanning literature, for example on the position of a specific gene. 'Watson can search much faster than we can and look at far more literature. We are capable of finding out that this gene is in the upper part of chromosome 3, for example, but that's still an area with 1000 genes. That is more than we can manage,' says Visser. 'Watson can dig deeper, comparing those 1000 genes with the DNA from other plants, or even fungi, mussels or birds, to see what is known about the function of those genes there. Then you can probably limit the list of candidates to five or ten genes, and as the computer gathers more and more data you may hit the jackpot straight off.'

The computer learns from its mistakes. If experts tell it that the answers are going in the right direction, it continues along that path but if the feedback is negative, it abandons that route. 'So it genuinely learns just like a human,' says Visser, 'and builds up expertise. We will be able to improve the selection of our breeding material enormously by linking information about external characteristics, DNA composition, yields and growing conditions.'

They may be able to get even more benefit out of Watson in the longer term, for example by putting it to work on complex problems. 'Which genes are responsible for the fact that some potatoes remain firm when boiled while others turn to mush? We don't really have a clue at the moment.'

MAKE AGREEMENTS

Despite the inviting prospects offered by big data, there are still hurdles looming ahead. Who should be allowed to make agreements about the data on a wheat field? The owner of the drone, the organization that is able to interpret the data, or the farmer? Do the data collected day in, day out by the milking robot belong to the manufacturer or the livestock farmer? That is still virtually uncharted territory. Who owns the results if a company and a university combine their datasets?

According to Ben Schaap, it is best for innovation if everyone has access to data files. Schaap has been sent on secondment by Wageningen UR to the global lobby organization Global Open Data for Agriculture and Nutrition (GODAN). Its sponsors include the US, the UK, the Netherlands and the FAO, while its 250 partners include both major companies, such as IBM and Syngenta, and local African NGOs.

'Openness ensures a level playing field for everybody,' says Schaap. 'If access to data is restricted, parties with more money and power have control. A multinational can buy information and a one-person company can never compete with that. If the information from weather stations or satellites is made public, anyone can have a go with it — not just a multinational but also a whizz kid. Open data means that anyone can develop applications and farmers are not dependent on a single organization that also supplies them with seed, fertilizers or crop protection products, for instance.'

That is why Schaap feels it is really important for publically funded data to be made available to all with no strings attached. 'That includes the datasets generated by the universities and institutes. The Dutch government and Dutch Organization for Scientific Research (NWO) are now making open science a condition for grants. Researchers applying for money from the Gates Foundation or the European research programme Horizon 2020 are also required to publish their data.'

'If access to data is restricted, parties with more money will have control'

COMMERCIAL INTERESTS

Schaap can understand that companies will probably not be keen to let others use their data. But he still expects them to make exceptions if the data do not represent a competitive advantage for them. 'Syngenta released a dataset on a pesticide for mosquitoes; this was not a priority area for the company and malaria researchers were really pleased to have it. You should see this as their contribution to making the world a better place, along the same lines as a reduction in CO_2.'

Yet he sees open data as more than simply charity. 'There are also companies that are prepared to create an open data landscape,' says Schaap. 'They feel that exchanging data should be made easier so that more players can work with one another's data in developing useful applications, for example in precision agriculture. The idea is that innovating is not something you can do on your own anymore. Open science enables collaboration with smart inventors. Syngenta says to startups: "Please make use of our research data. If you come up with an interesting application, we may want to take over your company".'

Plant breeding expert Visser is also in favour of open access, but only under certain conditions. 'You don't want any old whizz kid from Russia getting his hands on the data. That's what American researchers who are funded by the National Science Foundation find so frustrating at the moment. If they sequence a genome, then it has to be published on the web the very next day. They don't have time to look at it properly themselves first. Others sit back and wait for this, say "thank you very much" and publish some nice results.'

Certain rules should be imposed even in open science, thinks Visser. 'Perhaps the user should say first what they are intending to do with the data. If that's close to our own areas of interest, it's only logical if we then make arrangements about collaborating or if we say we need to be given time first to publish a couple of papers.'

HARDWARE NEEDED

In March, a number of scientific institutions including Wageningen UR published the FAIR Guiding Principles in Nature in an effort to facilitate cooperation. Data should be *findable*, *accessible*, *interoperable* and *reusable*.

A good start, says Visser. 'But you also need hardware. It doesn't matter where you store the data, whether that's with IBM, in the cloud or on SARA, we will still have to buy our own computer infrastructure for Wageningen and have people who know how to operate it. Otherwise you'll be dependent on others for everything. If you want to keep up, you need to invest.' ■

www.wageningenur.nl/en/bigdata