Human data science

DL Oberski

Utrecht University and University Medical Center Utrecht

Abstract

Most data science is about people, and opinions on the value of human data differ. The author offers a synthesis of overly optimistic and overly pessimistic views of human data science: it should become a science, with errors systematically studied and their effects mitigated – a goal that can only be achieved by bringing together expertise from a range of disciplines.

"Data are people!"

On the second of July, 1832, a 29-year-old data scientist named André-Michel Guerry presented a short manuscript at the prestigious *Académie Française*. His talk, published as *Essai sur la statistique morale de la France* (1833), changed the way we view data, and it would also change the speaker's own life, earning him an award and a coveted place in the academy. Up until that moment, Guerry had been an unassuming lawyer from a provincial town, and had published works such as *Ancient folkloric chants of Poitou*, which he presumably wrote as a hobby to distract him from his initially dreary-seeming day job of compiling the French state's crime data. During this time as a data manager, though, something marvelous appears to have occurred to Guerry: data on people were everywhere.

Readers in the era of personal computers, mobile phones, satellites, wearables, electronic health records, Google, business intelligence software, the internet, and the modern surveillance state will be as surprised by this insight as they are by tap water. But Guerry (along with his contempory Adolphe Quetelet) showed the world that we need only reach out and analyze data that are already out there to learn more about the world and change our understanding of, for example, crime (Friendly, 2007). In the time of Guerry and Quetelet, routinely produced and available data were overwhelmingly about people, because they were produced for public administration by the state (hence "state-istics"). Our modern era is no different, in that the data leveraged in data science are predominantly collected in the course of public or private administration, and therefore about people.

In short, data science, since time immemorial, is mostly about humans: to mangle a quote from 1973 cult movie *Soylent Green*, "data is people!".

If data is people, how good are people? This appears to be a subject of some disagreement among philosophers, the comprehensive study of which is left as an exercise for the reader. For understanding of the key issues, here the author will rely on TV. In the show *The Good Place* (NBC, 2016–2020), two opposing theses are proposed, which align closely with current data science practice. These are: (1) everything is fine; (2) people are terrible. The purpose of this opinion piece is to combine these apparently contradictory theses into a new suggestion: that neither blind optimism nor blind despair are warranted, and data science must concentrate on developing the methodology to deal with the nature of humans and their data.

Idea #1: Everything is fine

Data science has long been recognized as carrying great potential to improve people's lives and gain insight. Some examples of applications currently at various stages of development and deployment include automatic segmentation of radiology images, processing of vast amounts of -omics data in bioinformatics, early warning on sepsis in neonatal intensive care units, prediction of housing market dynamics, or modeling of the spread of both infectious disease *and* anti-vaccination opinions through social networks.

A "can-do" attitude. A striking trait of many data scientists is a "can-do" attitude to data. For example, in making the beautiful graph reproduced in Figure 1, Guerry (1864) realized that England and France employed different definitions of crimes, that there were ample missing data, and that the English data were less precise—note, for example, the differences in binning across the two columns (for a more detailed description of this and other visualizations, see Friendly (2007)). But, thankfully, none of those issues appears to have plunged him into an impassable mental labyrinth of anxieties. Data science's "can-do" attitude may have arisen from its partial roots in engineering,as engineers are used to making do; if MacGyver had spent his time philosophizing about the adhesive properties of duct-tape, he never would have escaped that shark tank. This positive spirit plays an important role in data science's current popularity: it is exciting to think that we can understand more and make better decisions, purely using "found data" from all types of human activity. When Guerry presented his findings to the *Académie*, many of its illustrious members must have been pleasantly surprised that all the human data lying around could produce such interesting results.

Data science improves continually. The inner circle in Figure 2 shows an approach to (human) data science in an idealized cycle of investigation and ”product” building (where a ”product” can also be an insight). I will briefly illustrate. Ideally, one would start with the *problem formulation*. For example, to reduce the costs and improve the benefits of treatment for a new infectious disease, by designing a system that will screen potential patients more accurately than is done currently. People that do not need treatment will receive it less often, reducing costs; and patients that do need treatment can be recognized earlier, improving survival. We then need *data*. The ideal, perfectly measured and perfectly randomized, data are not available. But we make do with a large amount of data from an app in which people register their symptoms, as well as from the routine care processes of general practitioners (GPs) and hospitals, in which we find survival data and background variables. The next task is to assess how survival might be *predicted* from symptoms and background variables, so that we can screen people. ("Machine learning" is a commonly used term for this step, although some readers might prefer some other term, such as "prediction modeling".) Once that is done, and evaluated to everyone's satisfaction, it is time to *implement* the screening, perhaps by changing the warning system in the app, or by developing a software plug-in to the GP's electronic health record. And after the system has been implemented, it will influence actual *decision-making*, for example by guiding the decision to refer a patient to further care or advising home confinement. We can then keep observing and improving the system.

What could possibly go wrong?

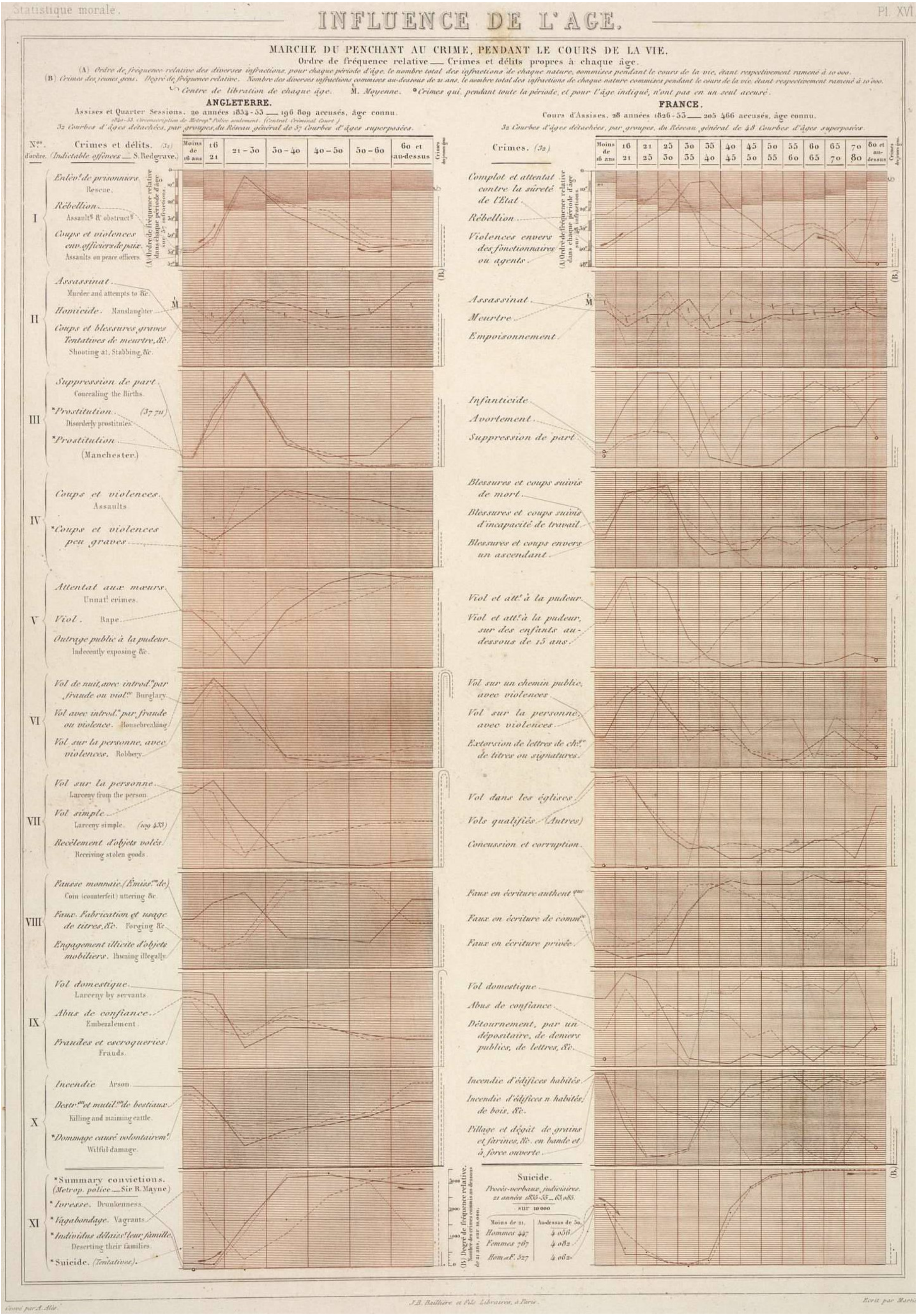

Figure 1: Guerry's visualization of relative crime statistics (vertical axes) across age groups (horizontal axes) from England (left-hand column) and France (right-hand column). Crimes with similar life-course patterns are grouped together. Courtesy of Universitat de Barcelona, Biblioteca Patrimonial Digital. License: Creative Commons Public Domain Mark 1.0.

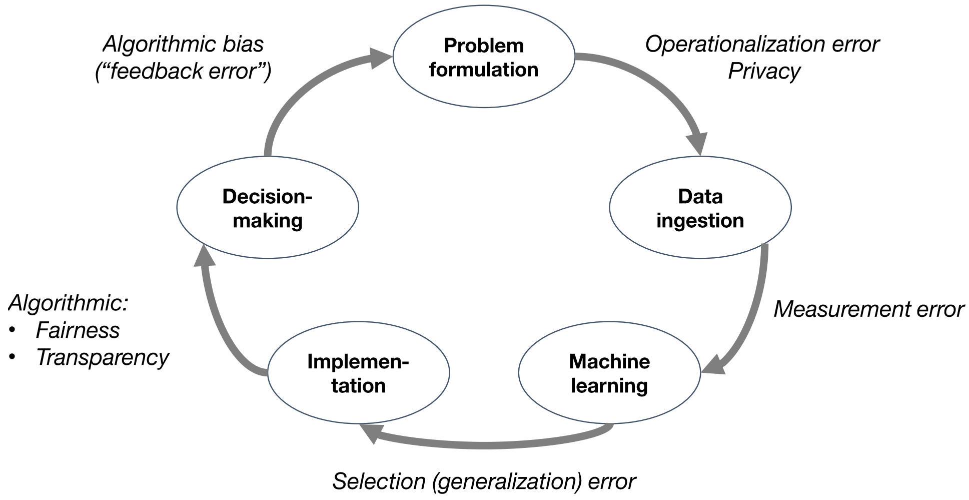

Figure 2: At each step of the typical human data science cycle, errors occur.

Idea #2: People are terrible

Among the various objects that fall within the purview of statistics, one of the most important and the most difficult to deal with, consists in the enumeration and classification of human actions...

Lacroix on A-M. Guerry's Essai sur la statistique morale de la France, 1833

”Pobody's nerfect.”

Eleanor
The Good Place

When dealing with human data, the can-do "everything is fine" is one idea. Its anti-thesis, embodied in TV's *The Good Place* by the evil demon Shawn, is that humans are the worst. Specifically: they are hard to predict, they differ, they have rights, and they are like ourselves.

Humans are just hard to predict. We do not currently understand the health and societies of humans–or other complex organisms for that matter–the way we understand, say, bridgebuilding. Two centuries of quantitative social science have shown clear patterns in averages such as those shown in Figure 1. But, even using more data and more sophisticated models, it can be surprisingly difficult to improve upon existing, simplistic models for individual human health and behavior. For example, Christodoulou et al. (2019)'s systematic review of machine learning approaches to clinical prediction concluded "no performance benefit" of modern machine learning approaches over a simple benchmark; following the illustration, we may doubt whether our elaborate cycle will improve on a common-sense rule that simply says people who look sick should see a doctor. In sociology, Salganik et al. (2020) reported on the Fragile Families Challenge, a mass collaboration effort to measure the predictability of life outcomes such as layoff, GPA, or material hardship, that "despite using a rich dataset and applying machine-learning methods optimized for prediction, the best predictions were not very accurate and were only slightly better than those from a simple benchmark model".

Humans differ. "How are you feeling?" An innocuous question used in polite conversation, in social surveys, and at the doctor's office. But, as doctors know, no two people answer this question in the same way. They might give the same *answer*, but that does not mean they feel the same. Or they might give *different* answers but feel similar (a technical doctor-term for this is "crybaby"). This becomes a measurement error problem when self-reported symptoms are used to predict survival, which leads to problems when symptoms are used as ground-truth in an ML model. And before you start thinking self-reports are worse than so-called "objective data" from administrative records, wearables, apps, and the like, those data have huge errors as well (see, Oberski et al., 2017, for references). Selection error is a problem too. For example, looking only at Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies gives experimental results in behavioral science that are probably atypical for humans in general. In our illustration, not everyone uses an app. That is fine as long as future app users are similar to current app users; but, presumably, by improving the app more people will want to use it. So we are changing the population. Because populations can shift over time, selection errors may lead to generalization problems.

Humans have rights.

> "'Fair' is the stupidest word humans ever invented, except for 'staycation'."
>
> *Shawn (demon)*
> *The Good Place*

Common decency, and – for the demons among my readers – many laws, require that data science, when applied to humans, makes decisions that are fair and that do not have harmful side-effects, for example through confidentiality breaches. In many cases, some degree of explainability or transparency of the decisions is also necessary. This means that prediction performance or the rigor of the scientific approach is not the only parameter to be optimized: instead, the entire cycle must be enacted under constraints, which can be considerable, and which can have dire consequences when violated (see Obermeyer et al., 2019, for an example).

Humans "R" us. Finally, dear reader, you and I are not the only smart ones. Our data subjects and product users are smart like us. When we advise them at the end of Figure 2's "inner" data science cycle, and try to go back to the beginning, they will think about our advice and either (1) ignore it or (2) not ignore it. Both can be bad. When they ignore us, the potential value we created with our data science cycle is wasted. But things can get even worse when people actually listen to us: people might trust the results too much, for example, making decisions that do not properly account for the associated uncertainties, and leading to potentially large losses. In our illustration, patients kept at home by a human doctor might be followed up more than patients being kept at home by an all-knowing computer, for instance. Another problem is that of algorithmic feedback. To illustrate, suppose that, in our data, people with a fever survive equally well compared with other symptomatic patients, so that this variable is not used to predict survival. But perhaps this was because the feverish go to the hospital early and therefore get better treatment, canceling out their otherwise bleaker prospects. The absence of an observed association has erroneously convinced our model that people with fever have nothing to worry about. After implementing the model, this (1) worsens their survival and (2) decreases model accuracy (I leave it up to the reader's conscience to decide which is the more pressing reason for caution).

In short, the second thesis is: sure, it seems attractive to enter the inner circle in Figure 2. But at each step of the process, in the outer circle of Figure 2, demons stand ready with the pitchforks of measurement and selection error, fairness, transparency, and privacy risks, and algorithmic bias, to torment us. All who enter here, despair!

A better idea: Let's science our problems

> In football, trying to run out the clock and hoping for the best never works. It's called 'prevent defense.' You don't take any chances and just try and hang on to your lead. But prevent defense just *prevents* you from winning. It's always better to try something
>
> *Jason Mendoza*
> *The Good Place*

In summary, an, admittedly caricatured, picture of the current situation is this: we either assume everything is fine (idea #1), or we dismiss human data out of hand (idea #2).

Because neither is particularly constructive, we should do neither. Instead, we should concentrate on the **science of human data science**. Good human data science:

1. Acknowledges that errors – in data, modeling decisions, and implementation – are inevitable;
2. Studies the degree to which errors actually affect the output;
3. Removes these effects where necessary, either by preventing the errors or, when this is not possible or cost-effective, by correcting for their effects.

In keeping with data science's engineering roots, good human data science is also good engineering. A good engineer does not just plow ahead, regardless of the terrain. Instead, she designs a solution that uses the available materials, while controlling the risks. These risks have been recognized from the very beginning for human data science; even the person who wrote Guerry's introduction called human data the "most difficult to handle". And he was probably being polite.

Here are some potential things one could do to mitigate risks by approaching human data science as a science:

- Work closely with domain experts to formulate a useful problem and operationalize it into a data science project appropriately;

- Work closely with domain experts to understand the evidence that including certain variables (features) will yield a net benefit;
- Don't assume measurement error will "cancel out". Use statistical models to estimate and correct for the effect of measurement errors;
- Use causal models and research designs to estimate and correct for the effects of selection error, including missing data;
- Investigate any threats to fairness of decision-making, and use appropriate techniques to ensure models do not learn biases;
- Use appropriate techniques to explain the model results to users and data subjects;
- Work with cognitive scientists to evaluate whether the above actually worked;
- Work with social scientists and domain experts to investigate how the data product affects daily practice.

How do we work towards achieving these lofty goals?

First, **latent variable models and causal modeling can provide a convenient framework** to think about some of these issues. For example, latent variable models are a convenient framework to estimate and correct for data issues (Oberski et al., 2017), causal modeling is a useful tool in dealing with selection error (Mohan and Pearl, 2014), and both causal modeling and latent variables are useful tools when investigating model output's fairness (see Boeschoten et al., 2020, and references therein).

Second, human data scientists need to routinely take a much wider view of "model evaluation"; a successful data science project must not only predict well in out-of-sample data, but also: generalize to other contexts in which it might be applied, generalize to the concept it was intended to study, reflect uncertainty accurately, improve on what was already there, have benefits that outweigh its costs, and actually achieve its stated goals in social reality. To verify all of these, the **comprehensive scientific study of data science projects should become routine**.

Third, we should collaborate much more. I do not pretend to be the first person to point out these issues are important; many of the points above are already entire research fields, covering swathes of statistics, computer science, ethical and legal scholarship, domain sciences such as medicine and biology, and social science. But, in my opinion, we are not collaborating enough, and we are not bringing in a large enough diversity of perspectives. In going through the data science cycle, it is all too easy to focus on what interests us, to the detriment of the project as a whole. No one person, including your author, is immune to this problem! For example, some research in data science ethics focuses exclusively on ethical, legal, and societal aspects, and pass over existing mathematical tooling to weigh the issues in a practical engineering context. Conversely, some other research focuses entirely on the mathematical tooling without considering whether it will be of practical use in a social context: have I done justice to people's intelligence and autonomy; will people actually react the way I assume they will? Of course no one person can do everything. That is exactly why we should work harder to **unite *all* the fields that work on these problems**.

I am aware both that my to-do list for the science of human data science is long, and that it should be longer. But, if we really want to put human data to good use, it is the work we need to do. The demons are legion, and nobody ever said it was going to be easy to avoid them. But, as André-Michel Guerry, Jason Mendoza, and many others have shown, it will be well worth it. So: let's science the heck out of human data science!

Acknowledgements

Thanks to Laura Boeschoten, Dong Nguyen, Ayoub Bagheri, Michael Friendly, and Peter van der Heijden for their useful comments.

The author was supported by the Dutch Research Council (NWO) [VIDI grant VI.Vidi.195.152].

Biography

Daniel Oberski holds a joint appointment as associate professor of data science methodology at Utrecht University, department of Methodology & Statistics, and at the University Medical Center Utrecht (UMCU), department of Biostatistics. His work focuses on latent variable modeling and data science applications in the social, behavioral, and biomedical sciences. He is lead data scientist of UMCU's "digital health" program, which works to implement data science in clinical care at the hospital. He also leads the social data science team at the national research infrastructure for the social sciences in the Netherlands, ODISSEI.

References

Boeschoten, L., van Kesteren, E.-J., Bagheri, A., and Oberski, D. L. (2020). Fair inference on error-prone outcomes. *arXiv:2003.07621 [cs, stat]*. arXiv: 2003.07621 version: 1.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22.

Friendly, M. (2007). A.-M. Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis. *Statistical Science*, 22(3):368–399.

Guerry, A.-M. (1833). Essai sur la statistique morale de la France précédé d'un rapport à l'Académie des sciences, par Mm. Lacroix. Crochard, Paris. OCLC: 1096788219.

Mohan, K. and Pearl, J. (2014). Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems*, pages 1520–1528.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Oberski, D. L., Kirchner, A., Eckman, S., and Kreuter, F. (2017). Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models. *Journal of the American Statistical Association*, 112(520):1477–1489.

Salganik, M. J. et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403.