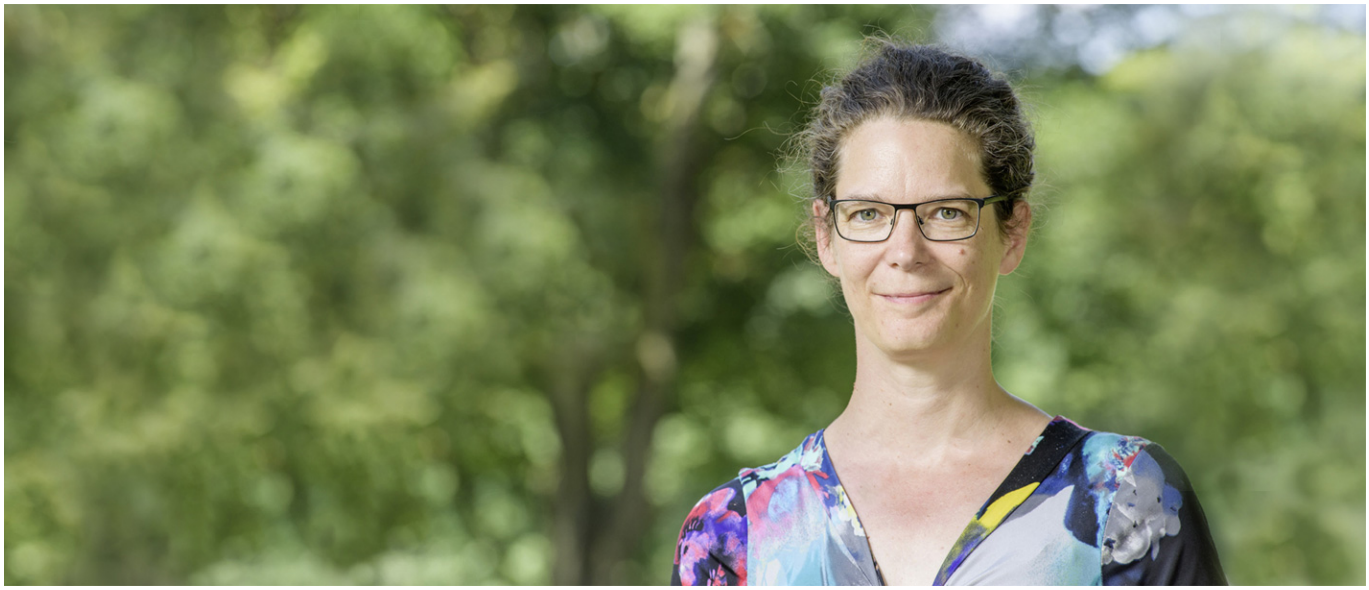
Datum: 10. November 2017



Big Data – the beginning of a golden age for empirical labour market research? An interview with IAB researcher Frauke Kreuter

Martin Schludi

Big Data! This is the new magic word in labour market research. But what is this really about? What is the use of it? And what does it mean for data protection and the training and further training of our researchers? Prof. Frauke Kreuter, Head of the Statistical Methods Centre (Kompetenzzentrum Empirische Methoden – KEM) in the IAB, answers these questions in an interview for IAB-Forum.

Recently, a lot has been written about "Big Data", in particular in connection with social networks. What exactly does this term mean?

This term is, indeed, used a lot, but people mean different things by it. One thing is for sure: Big data is not so much about the actual size, because large data volumes have existed in the field of physics, astronomy, geosciences, etc. for a long time. What is typically understood under big data are large amounts of unstructured data originating as a by-

product of other activities, i.e., not collected for this purpose, or for which no own measurement series was established. For example, texts and images are created in digital form in social networks which are now available as data and can be evaluated by us.

> "What is typically understood under big data are large amounts of unstructured data originating as a by-product of other activities."

The Federal Employment Agency (Bundesagentur für Arbeit – BA) has been working with huge data sets for many years. What are the differences between big data and these administrative data records?

Administrative data records also arise as a by-product of other activities, in this case from social security notifications. This is why, for example in the US, administrative data are often also discussed under the keyword big data. When people differentiate between big data and administrative data, it is often because big data are connected to at least three characteristics, some of which are not quite true for administrative data: Volume, i.e., very large volumes of data; this holds true for the BA's administrative data as well. Velocity, i.e., a constant stream of new data; this is not always the case with administrative data. Notifications are typically not in real-time as is the case, e.g., with Twitter. Here, the data set expands permanently with each new tweet. Variety, i.e., a variety of data types; this is also not necessarily the case; typically, data have already been coded, there is much less free text, and rarely pictures or audio files – however, there are differences between various administrative sources. A fourth characteristic, veracity, i.e., the truthfulness of data, is something that is important for all types of data. Especially recently, there has been lots of talk about 'fake news' and 'bots'. This means: With Twitter, as with all the other data, you have to look very closely at what you are actually measuring.

Big data are often very unstructured data. What does this mean for the quality of these data?

Whether data are structured or unstructured does not necessarily say anything about the quality. However, you cannot discuss data quality independently of the respective question. In simple terms, you can differentiate between two different types of problems: Missing data for certain groups of people, events, or points in time, hence the question "How suitable are the data for general statements?" and the question "Do the data actually measure what you are interested in?".

Can you give us an example?

Let's assume I am interested in how the salaries of men and women differ in the Federal Republic of Germany, then I must make sure that I have enough data for the entire country. Since salary differences could vary regionally or by industry. However, if I am interested in whether a new headache pill has the same effect on men and women, it may be less important to have data from all parts of the country and of employees in all industries – at least if it can be assumed that this does not affect the effects of the pill. Measuring what you actually want to measure is a major challenge especially with unstructured data. There are some great developments in text analysis, but, depending on the context, it can still be difficult to have an algorithm decide automatically what a "like" on Facebook means or whether a sentence was serious or ironic. Some social scientists currently launch into Google queries. Here, too, it is not always clear why someone enters a specific search term, and what this search means in aggregated terms. For a while, Google very successfully correlated queries with flu epidemics. This was turned into an index to predict flu epidemics. After a few years, however, unfortunately, this was no longer successful since some terms only correlated highly for a while but had nothing to do with the flu epidemics in terms of content. Transparency of algorithms and data-generating mechanisms can help here.

> "You can create exciting analyses by linking different data sources with each other."

Can big data also be linked with conventional data?

Yes. Frequently, you can create exciting analyses by linking different data sources with each other. Facebook, for example, has an incredible volume of user data. Nevertheless, Facebook frequently conducts additional surveys among its users, because they found out that from behaviour alone they often cannot derive what users like or don't like about the platform with sufficient precision. In this case, you can establish a direct connection of big data and conventional data at the user level. In most cases, e.g., when evaluating Google search terms, this is quite complicated. The search terms are usually available at the aggregate level, i.e., at the level of states or cities. In general, many big data sources are geocoded and can be linked via geographical coordinates.

Can you explain this with an example?

My current favourite project uses photos visitors take in African national parks, and the

associated geographic coordinates to determine the size of the zebra population. Ideas of this kind also exist in migration research. For example, there are attempts to observe the movements of language groups on the basis of geopositions of Twitter posts.

How can big data generate an added value for labour market research? Do you agree with the opinion that a golden age is dawning for labour market research due to big data?

This implies that labour market research has been in a bad state – which I think is not true. Golden age in the sense that there is a gold-rush mood regarding what could be done with big data and where treasures can be collected – including the expected glitches. For sure! (laughs)

We certainly have not yet exploited all the data sources in labour market research. In Germany, administrative data, as you mentioned yourself, have been used in research for quite a while. In the U.S. such data have not been as easily accessible. In fact this fall a report from the Commission for Evidence Based Policy Making was published urging the government to make legislative changes that allow for more access. If this happens this would surely boost certain labour market research, we already see this in pilot activities around the NYU Administrative Data Research Facility. Some do call the work happening there a revolution.

And what about social media data?

It is fascinating to watch what is happening in that arena. One example, American colleagues at the University of Michigan were very successful with the evaluation of Twitter data and the reproduction of an indicator for unemployment insurance statistics. In Germany, private twitter posting is less popular and some of these things would likely not work as well. But the evaluation of network platforms like Xing or LinkedIn or the use of the BA's Jobbörse has only just begun. The Junior Researchers at IAB constantly develop new ideas for the use of new data sources – stay tuned.

> "The Junior Researchers at IAB constantly develop new ideas for the use of new data sources – stay tuned."

Does the IAB already conduct research using big data?

Yes. I have mentioned the BA's Jobbörse, a project by Christian Hutter has been working on

this for some time (more information can be found here). Other projects are currently beginning. Perhaps you have heard about the Marienthal study? The study observed an entire city, and measured and wrote down by hand how the activities of the people in the city changed after a large factory closed down. Today, you can measure in completely different ways, with our mobile phones many of us already have a measuring device in their pocket which you can, of course, also use for research purposes. Together with the group of Mark Trappmann, we are currently developing an app which allows a modern form of the Marienthal study. Of course, this will only be possible if we find enough volunteers to participate, and if there are not too large selectivities. We will examine this as well.

Does the use of big data also threaten data protection in labour market research?

Data protection is very important to the IAB. This will not change in the context of big data. What is interesting for us all is how the new EU General Data Protection Regulation, coming into force in May 2018, will be implemented. I think it is important that everyone providing data has a basic understanding about what happens with their data, why it is important to provide them, and what risks exist. There is often diffuse scepticism because not enough is known about which measures are already taken to protect data. However, the landscape generally changes if more and more data are available which can be linked to each other. On the one hand, we can gain valuable insights thanks to big data. On the other hand, we must make sure that this happens in good faith and in the interest of the general public, and that data protection provisions are complied with. Some research data centres in Germany are already working on this subject – here in the IAB as well. Stefan Bender, the former Head of the BA's Research Data Centre at the IAB, is co-editor of an interesting book on this topic: Privacy, Big Data and the Public Good. I highly recommend this book to anyone interested in more details.

What challenges arise for the training and further training of scientific staff due to big data?

Education and awareness-raising for data protection issues and maintaining informational self-determination in general – this is very important to me – as well as awareness for data quality which we have already discussed. In the past, further trainings mainly covered learning evaluation techniques. This is still important and many things are happening in this field but what is new is the need to know which data are suitable for which questions and understanding the data-generating processes. There is also much to be learnt in terms of preparation of data, i.e., how to transform pictures or texts into figures. This is why we are currently working on establishing several training opportunities. Around the earlier mentioned ARDF there is a whole training series aimed at government employees working

with administrative data. Key is here to learn the techniques while answering research questions.

> “We are building an international continuing education study programme for survey and data science .”

For a wider audience and a broader range of topics we are building an international continuing education study programme. The project is funded by the Federal Ministry of Education and Research within the framework of the federal-and-state competition “Advancement through Education: Open Universities” (Aufstieg durch Bildung: offene Hochschulen). The IAB is a partner, the University of Mannheim which organises this course of study in Survey and Data Science in cooperation with the University of Maryland (survey-data-science.net) is the main responsible body.

The questions were asked by Martin Schludi.