Science of Data - or what we think about when we worry about data quality

Frauke Kreuter

(Statistics and Data Science in Social Sciences and the Humanities,
Department of Statistics, LMU Munich)

12.01.2022, 16.00 (c.t.)

Online via Zoom
(Meeting-ID: 913-2473-4411; Password: StatsCol21)

Social data do not simply buzz around in space/the cloud, but they are generated, interpreted, and put to use by living persons. Machine learning applications often require human annotation of data. Data are irrelevant without context, and they should be fit for purpose. When data are generated, social interactions determine the data we get and its quality. Data may be incomplete due to historical discrimination, social inequalities in participation in digital technologies, sampling errors, and lack of consent to data linkage, amongst others. What gets collected often has measurement errors. Results can thus become biased, sometimes in surprising and unexpected ways.

Despite these challenges that we wish to identify, understand and resolve wherever possible, data are not only a mainstay of the (social) sciences, but are also becoming an integral part of our (not only digital) lives. It is hard to imagine a more digital, more efficient, more enlightened world without data at its core. Data scientists are challenged to build data pipelines and infrastructure that can handle tomorrow's data requirements. Yet, concerns about data privacy need to be incorporated. If decisions are based on faulty data, they may discriminate against vulnerable groups, or perpetuate wrong decisions from the past into the future. A holistic view on the complete process, from data generation to deployment, is necessary.

In our talk, we introduce a series of projects that recently started, and for which we are eager to find new collaborators and people interested in joining our team. Given the current Omicron concerns and the fact that our group is spread across different locations, we will hold the seminar entirely online.