20th LCDS meeting: Big Data, Policy & Infrastructure (12 June 2017)

Abstracts

Beyond effective heuristics: Predictable, robust and performant AI - Holger H. Hoos (LIACS)

The success of data science relies on two key ingredients: Informative data and effective procedures for extracting insights from data. The procedures used in this context often solve challenging computational problems, and they are constructed, to an increasing extent, automatically, using sophisticated machine learning and optimisation methods. These methods heavily rely on heuristics, and although they achieve remarkable performance, their behaviour can be brittle - i.e., vary substantially and unexpectedly with input data.

In this talk, I will outline research aimed at achieving increased predictability and robustness of high-performance AI algorithms, drawing on several lines of work on performance prediction, algorithm selection and automatic algorithm design and covering examples from machine learning and several prominent, NP-hard AI problems.

Complexity in company turnover as measured from big data - F.P. Pijpers (CBS)

There are many factors that determine the turnover of companies registered in the Netherlands. Competition among companies as well as mutual exchange of goods, services, and assets play a role, as does the interaction with consumers or international competition. One of the outcomes of this complex interaction process is the distribution function for company turnover.

Statistics Netherlands / CBS has quite detailed information about (all) Dutch companies, including turnover, from a variety of administrative registers through chambers of commerce (KvK), tax returns, and VAT registrations. Added to this there is information collected through surveys and currently there are efforts to enhance this further by collecting data with web-scraping.

The combination of all of these data satisfies some of the criteria (volume, variety, velocity) of Big data, but the analysis also requires some of the tools and framework of the modelling of complex systems. This talk will present some early results as well as the importance for official national statistics of such research.

A data-driven supply-side method to accurately estimate cross-border online consumption of goods - Quinten Meertens (CBS)

Over the past two decades, the internet economy has grown immensely. As internet access among consumers increased as well, retail trade through online channels accelerated. A relatively recent shift in retail trade is online sales across national borders. This poses significant questions for policymakers on the effects on job markets, international trade and GDP estimates. Accurate estimates are crucial in answering such policy questions.

Earlier research on the drivers and impediments of online consumption, especially across borders, has shown that a website's language plays a crucial role, and not so much the company's geographical location. The consequence is that any consumer survey approach to measure cross-border online consumption is strongly biased. We propose a novel data-driven supply-side approach that is far more accurate. In particular, we show the vastness of the downward bias of a consumer survey approach: close to a factor of 5 in 2015 for The Netherlands.