What Can DB Researchers Do?

Input Source Selection

- Unverified ratings on a review website easily manipulated.
- Even after verified purchase, subject to selection bias:
  - Younger, more tech savvy people more likely to post
  - Very unhappy or very happy people more likely to post
- What about racial or gender or ethnic bias in the real world that is accurately captured?

Input Data Preparation

Must correct for known input source biases.

Systematic bias in missing or erroneous input can cause cleaned data sets to become biased. E.g. average of missing data could be very different from average of all data.

Poor choice of outlier detection can systematically remove data points associated with small minorities. (E.g. Facebook real names policy).

Data Representation

- Representation choices have to be made, and implicit assumptions creep in.
  - Permitting only two values for gender excludes other possibilities
  - Assuming a non-null value for an address field excludes the homeless.
- We cannot leave everything open: some system structuring is required.
- Can we develop systematic methods that minimize the likelihood of un undesired limitation?

Representative Sampling by Dataset Construction

Today, we often have access to many different data sources containing relevant information,   Each may have its own biases and its own domains of coverage. They may also very in cost, quality, and other criteria.

Can we construct a good sample data set by merging information from multiple data sources?

Ist there a tradeoff of cost and quality?

Data Profiling

Having onbtained a data set, before we use it for analysis, we must profile it.

How should one conduct a focused profiling that can illuminate possible bias?

For starters, we could report on representation of protected groups. But there are many intersectional and correlational questions of interest, too.

Algorithm Design

Problems can very quickly become computationally challenging, because many concerns are fundamentally combinatorial in representation.

Database insprired algorithms can make a big difference

In particular, top-k queries have been studied extensively, and may be valuable in addressing problems of ranking.

DB Integration

Currently, all the data processing required for responsible scoring occurs at the application level, on top of a database system.

There are likely to be significant performance benefits to integrating this functionality into the data management system.

Such integration will also provide a mechanism to build in data quality guarantees as a service to downstream learning applications.

Result Presentation

Cherry picking is a major problem in data analytics.

It is easy to end up with unstable results, even unintentionally.

Techniques to ensure robustness of results are a good topic of research, as are techniques that can minimize the likelood of misleading.

Conclusions

Database community has much to contribute to responsible Data Science.

In this tutorial we saw, in particular, challenges to address in ranking and scoring.

Many of these challenges carry over to other analytics and AI problems.