Data sec:data

First, we determine features of people that might influence their probability of winning a Nobel Prize. They are elaborated in Section ssec:features. Next, we research whether and where this information can be found. The main data source are knowledge bases, also known as Linked Open Data in the Semantic Web. We can use SPARQL queries on an endpoint to easily retrieve structured data wc3SPARQL. The used knowledge bases are discussed in Section ssec:knowledgebases. Needed information that is not (yet) linked has to be scraped from websites. These are discussed in Section ssec:additional. We use Perl scripts and regular expressions to extract the data from plain HTML text. All the retrieved data and all the used scripts can be found at the webpage for this paper: kaw.wardschodts.ws.

Feature selection ssec:features In order to determine whether someone will win a Nobel Prize, a set of reasonable features that the model can base its decisions on is needed. A summary of the chosen features is given here. itemize

Year and country of birth. Currently the average age of a laureate is 59 age. Evidently, this plays a role in the decision if somebody makes a chance for a Nobel Prize. For example, young scientists are far less likely to win an award because they have yet to prove themselves. Next to this, we are also interested to see if the country of a person has an influence on this.

Alma mater. We suppose that there's a link between the university where one graduated and his chances of winning a Nobel Prize. For example, admission to prestige universities is only granted to the best candidates. In order to extract a usable feature from an alma mater, we use its award score, which is based on how many Nobel Prize and Field medal winners it produced. In case someone has studied at multiple universities that have a score higher than 0, the average is used. Unranked universities automatically have a score of 0 (which is an inherent property of how it is calculated), if this is the case for someone with multiple universities, then these universities are omitted from the average.

Work productivity. A measure of work productivity can also be useful to include. This is however not trivial to determine, as not many properties are available for the majority of scientists and politicians. In order to solve this, we attempted defining a proxy measure for work productivity for the training data and testing individuals. For the scientists and economists from the training dataset, their number of publications might be an appropriate measure for productivity. For politicians, the number of speeches they make are perhaps a good indication of how productive they are. Because this is a sensitive subject, Appendix sec:productivity provides a thorough discussion of why we chose these features and how the research question is slightly weakened by our assumption. Furthermore, in Section ssec:quality on quality assessment, we show why we deem both to be a reasonable proxy by comparing their distributions.

Popularity. A popularity measure is easier to define. Facebook automatically generates pages for individuals that exist on Wikipedia. Therefore all people that we consider have at least such automatically generated page. The number of likes these pages have can then serve as a measure for popularity.