When Data Lies
The Boundaries of Data Cleansing

In order to gain insights from data analysis, the underlying data set has to offer an appropriate level of data quality [3]. The cost-intensive data quality assessment and data cleansing might be in vain when the data set itself is not trustworthy and contains misleading or wrong information because the insights based on such data are not valuable.

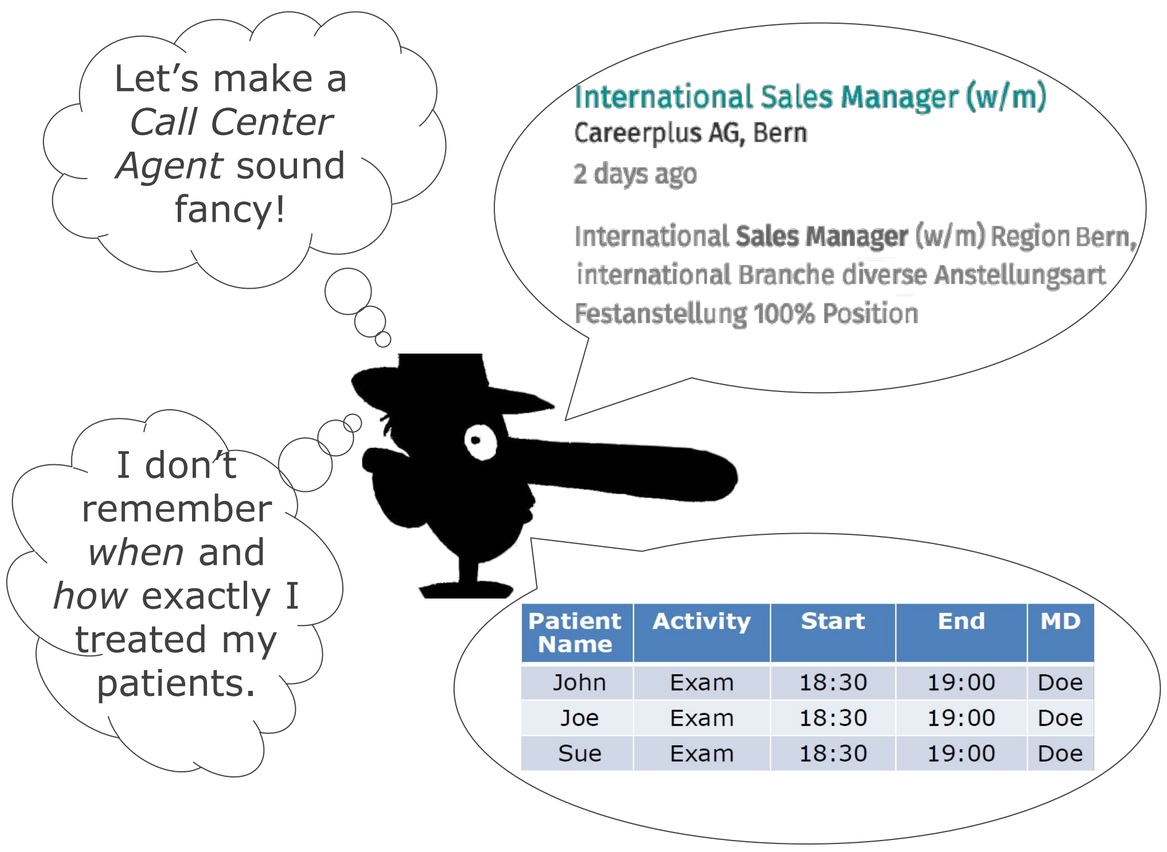
Data Collection Issues

There are various possibilities to collect data, such as surveys, crawling web content, manual logging, tracking, sensors etc. Any of such ways can provide false data, e.g. due to attacks. Problems especially arise when data is collected in a non-automated fashion or when data is extracted from human generated content. In such cases, the data may be entered wrong unintentionally or even on purpose. The latter may happen when authors try to hide the real meaning of a text, such as a job ad to attract more applicants [2].

Automated Logging

To achieve a more objective data collection, a switch from manual logging to automated logging is possible, e.g. using a real-time event system [1]. Unfortunately, the generated data then is not in the desired format. Thus, strategies like *interaction mining* need to be pursued.

Using Domain Knowledge such as Duration, Participants and Precedence to Identify Activities from Interaction Data.

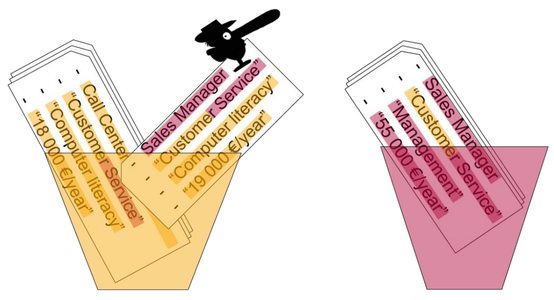


Job Title Classification

is a type of text classification used in the field of job recommender systems for clustering jobs to compare them to seekers' needs. LinkedIn currently uses *phrase-based classification* which is omitting long description fields. Short fields are known to contain most information on the class. Another approach is the *semantic enrichment* where class information is augmented with contextually relevant terms from the job description [4].

The first approach fails to recognize the actual job type due to the misleading title in the call center agent job ad from the above example.

Clustering, using title and description fields, reveals that the, as Sales Manager disguised, job actually is a Call Center Agent job:



Call Center Agent Job Cluster

Sales Manager Job Cluster

Conclusion

Any collection technique has its flaws, be it only a tempered sensor. In certain cases data cleansing might not help when a great amount of the data is based on misleading or wrong information. Therefore, it is necessary to discover ways to mitigate this issue.

For some cases specific strategies can be found to cope with the corrupt data or to change the data collection technique.

As shown in [1], a switch to automated logging may be possible but additional problems have to be solved by strategies such as *interaction mining* to recover the originally desired data format. For the problem demonstrated in [2], a different solution can be found, namely *job title classification*.

For any other domain the problem remains and individual strategies have to be developed.

Klara Seitz
Master Student

Hasso Plattner Institute, Potsdam, Germany
E-Mail: klara.seitz@student.hpi.de

Poster for the *Data Engineering Lecture Series*

References:

[1] Matthias Weidlich, "Queue Mining", Ringvorlesung, 2017
[2] Fabian Abel, "Data Engineering for Job Recommendations", Ringvorlesung, 2017
[3] Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. 1996
[4] Y. Zhu, F. Javed, O. Ozturk: "Semantic similarity strategies for job title classification". arXiv 2016

HPI Hasso Plattner Institut