## Technology and data - server costs note

During the meeting, Talkwalker said that they are spending approximately 85.000 euro on renting servers each month which is surprisingly low considering they rent over 2.000 servers each month. Hence, a more moderate estimate of the monthly rental costs per server is euro. In comparison, a small/medium server with eight cores costs 165,4 euros per month on Google cloud, and a large server with 32 cores cost 660,5 euros per month also on Google cloud. That means if Talkwalker’s servers are small/medium sized, then they would have to pay approximately four times more on Google cloud. If they use large servers, then they would have to pay around fifteen times more on Google cloud. Of course, the prices mentioned are pocket listing prices and one would probably be able to get considerable rebates. But that doesn’t change the fact that Talkwalker’s monthly server costs seem to be remarkably low.

## Technology and data - GDPR note

It should be noted that with the new data privacy act new requirements for storage of data are put in place. Considering that Talkwalker is hosting their servers at Hetzner, which primarily provides services for small customers, then the implementation of GDPR might require Talkwalker to change their server provider to one of the big three: Azure, AWS or Google cloud. This will drastically increase monthly server costs.

# Machine Learning Infrastructure and Data Science

## General conclusion

The Talkwalker data-science/machine-learning team seems to be highly skilled and have succeeded in implementing quite challenging ML models. Furthermore, it should be noted that the ML models put into production seem to be highly optimized thereby allowing them to scale data sets of considerable sizes.

The ML team does, however, seem to be biased towards more academically challenging problems instead of tackling more fundamental issues. Hence, one gets the impression that there might be a series of "easy wins" where considerable user value could be added without much effort.

In summary, we assess that Talkwalker has a highly skilled ML team that has implemented ML in a series of fascinating use cases.

## In-depth explanation

From an academic perspective, Talkwalker seems to have an excellent data-science/machine-learning team.

They have a good framework, where they prototype models in python and R and then use direct C++ bindings to enable fast and decentralized models in production. Here it’s noteworthy that they have extended the Caffe library to get a good production speed for computationally heavy ML models. In general, the ML models they put in production seems to be highly optimized.

It should also be noted that the level of implementation difficulty, for the models implemented by the Talkwalker team, is quite high. This testifies to the team’s academic prowess.

The team does, however, seem to be slightly biased towards using more ML-based approaches. This is exemplified by their story clustering approach, where they classify the stories using gradient boosting[[1]](#footnote-1). A more traditional clustering approach would be used for instance t-SNE to cluster the stories through dimensionality reduction visually. As noted at the meeting the large volume of data that Talkwalker handles renders typical clustering approaches which are practically difficult to implement due to the algorithmic complexity of clustering algorithms. One could, however, just cluster small subsets of the data, for instance, clustering at the level of some of the common queries.

While only clustering small subqueries might not be the most academically correct way of doing things, we anticipate they could provide the user with useful insights (and let’s admit it, everyone likes fireworks).

This brings us to another critical point. The Talkwalker team seems to prioritize academically tricky and interesting problems higher than more fundamental issues. This gives the impression that there could be a lot of "easy wins", where ML could add a lot of value to the product. These "easy wins" have, however, not been implemented yet due to a continued focus on use cases, which from an academic point of view, are more interesting. Given that the ML team seems to be highly skilled no technical boundaries are preventing them from implementing these "easier and less exciting” use cases.

Another interesting aspect of the ML teams work is that they have assembled a relatively big proprietary dataset of images of different brand logos[[2]](#footnote-2). The data set seems to be somewhat high quality as a result of the team’s ability to bite the bullet and invest a great deal of work in assembling the initial dataset, themselves. Along with that, they have refrained from using Amazon Mechanical Turk and instead are using higher cost options to ensure good quality.

Due to the lack of high-quality public datasets, for instance, logos, having such a proprietary dataset is without a doubt an essential component in getting an ML framework to work correctly. Going forward having such large proprietary data sets could prove very valuable. It should, however, be noted that firms are offering, for instance, logo recognition as a service (see <https://www.logograb.com/>). How Talkwalker’s ML models perform at recognizing logos compared to, for example, Logograb is challenging to say. However, even though Logograb might compare favorably to Talkwalker, it could from a production point of view be very valuable that Talkwalker can do the logo detection and classification in-house.

Furthermore, the ML team have expressed a desire to explore the possibility of providing their users with more "explanatory predictions" for example, being able to say we predict this will happen because of these factors. The ML team would obtain these “explanatory predictions” by leveraging additive regression trees. Their argument for doing so was that the prediction problems are inherently non-linear and complex therefore requiring highly flexible functional approximators such as additive regression trees. Although this might be true, we believe that it would be possible to quickly provide the users with a lot of value using simple linear-in-parameters regression models. If the model assumptions, for more standard statistical models, are not met by the data, these models can be used as a quick and dirty baseline, which could potentially give the user the insights he or she needs.

1. Their story clustering approach is to use the Okapi BM25 method to extract features and then use some of the most informative features – a few hundred features – in a gradient boosting classifier. [↑](#footnote-ref-1)
2. The position of the logos in the images has also been annotated with bounding box coordinates. [↑](#footnote-ref-2)