1) Extend the ideas for learned index structures to multiple dimensions, Learned Rtrees, Learned High dimensional index structures, metric trees. Some references:

<https://arxiv.org/abs/1412.6622>

<http://www-db.deis.unibo.it/research/papers/SEBD97.pdf>

<https://pdfs.semanticscholar.org/5444/cbc603c1b1d4e16bd8cadc9e55c6ce690cdd.pdf>

2) Selectivity Estimation: Extend ideas of selectivity estimation using deep learning to estimate selectivities of string collections. Review of past bibliography on the topic and then focus on the development of novel techniques using deep learning primitives. Some references

<https://www.cs.ubc.ca/~rng/psdepository/edbt2009.pdf>

# **Estimating the selectivity of approximate string queries, Mazeika, et. al.**

<https://www.cs.ubc.ca/~rng/psdepository/edbt2009.pdf>

http://www.cs.columbia.edu/~gravano/Papers/2004/icde04.pdf

3) Study sensitivity of cleaning and entity matching algorithms to training data. The basic idea is to see how sensitive a deep entity resolution algorithm is to training data. Some references to consult:

<https://arxiv.org/abs/1511.04599>

<https://arxiv.org/abs/1610.08401>

You can download the code from the paper “Deep Learning for Entity Matching a design space exploration” along with the datasets. The aim would be to study how changes in the test data, affect the accuracy of the algorithms reported.

4) Data Explorations: Utilize ideas from neural cubes for intelligent roll ups and drill downs or other types of data exploration. You will probably have to read ahead the data exploration papers and see if ideas from neural cubes can be applied to other types of data cube analysis. If you can obtain the code from the paper implementation, understanding the package and exploring if data cube style exploration can be conducted, it would be a good first step.

5) Extend ideas from the papers in class for video query processing to other types of interesting queries on video streams. You will have to read ahead the papers in the course syllabus. For example, what is the equivalent of a group by query? What are the tradeoffs to compute them and can we come up with efficient algorithms?

6) Extend ideas from the papers we read in class on video query processing to other query types as well as to multiple cameras (fixed angle). Multiple cameras would be the equivalent of a join operation. Assume the re-identification is given as a specific function that correctly identifies objects across cameras. What types of queries are of interest in this setting? Come up with query examples and see what are the associated algorithms to compute them.

7) Repeat the experiments in the Learned Indexes paper for the case of Bloom Filters and aim to quantify how bad the proposal will get if the learned distribution and the query distribution changes. Essentially expand on the ideas of the paper in class by Mintzemacher, and aim to quantify the critique that paper makes.

<https://arxiv.org/pdf/1802.00884.pdf>

https://arxiv.org/pdf/1901.00902.pdf