## 645 Reproducibility Mini-Project

The mini-project is a collaborative assignment, which is due at the end of the semester. Students will be working in teams of 3 (no exceptions please). You will be working on reproducing partial results from a research paper, selected only from the list we provide below.

**General project guidelines**

* **Step 1**: Self-organize in teams of three. We suggest using Piazza to connect with other students if you need help finding teammates. It is your responsibility to self-organize into teams in a timely fashion and you shoul.
* **Step 2**: (**Due Nov 2**) With your team, please select your preferred mini-project and a second choice, from those listed below. Notify the instructor and TA by email (cc’ing all group members) of your preferred choices. You will likely receive your first choice, but we may impose a balancing constraint that not more than ½ the groups works on any one project.
  + A representative of the group should send a single email, cc’ing all group members, noting your first and second project choices and any extensions you hope to explore in the context of your project choices (this latter part is not a binding commitment).
* **Step 3**: Complete the reproducibility tasks for your chosen project. The mini-project has a single deliverable at the end of the semester. It is your responsibility to pace your work appropriately.
  + To reproduce the research successfully, you need to carefully read and understand the original paper (and possibly supporting research). Your implementation must be your own! You are not allowed to use code related to this research that you find online or that you may obtain from others.
* **Step 4**: Report on your findings. You will submit a final report detailing what you did (your implementation, datasets, evaluation methodology, findings, and discrepancies with the results in the original paper). Some of the results may not be perfectly reproducible for a variety of reasons (e.g., changes in the datasets, differences in parameter tuning or reference systems, etc); that's okay but should be accompanied by an explanation.
  + You may also extend the experiments you do in any meaningful way to improve or enhance your project, or help explain your findings, in which case you should explain the steps you took and your additional results.
* **Step 5**: Submit your group report by Dec 20, 10pm. No late submissions will be accepted. Also submit individually a brief paragraph explaining your personal contribution to the work of the project.

**Project 1: Similarity Search in High Dimensions**

[*Weber et al. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In VLDB, 1998.*](https://pdfs.semanticscholar.org/63ea/eb0c48175065ffd096aad10aed712c6d7bbb.pdf)

Implement the VA-File and one other hierarchical method (either the X-Tree or R\*-Tree) and compare the performance on both synthetic and real data. You will reproduce Fig 14 (using synthetic data) and Fig 15 (using a real dataset). For the real dataset, find an appropriate high-dimensional dataset to evaluate on.

Extensions:

1. Study the impact of the data distribution on the performance of the above techniques by generating synthetic high dimensional data from a more interesting parameterized distribution.
2. Invent a “learned index structure” for high dimensional data and compare its performance with the above techniques.

**Project 2: “Picasso” Plan Diagrams**

[*Reddy and Haritsa. Analyzing Plan Diagrams of Database Query Optimizers. VLDB 2005.*](http://www.vldb.org/archives/website/2005/program/paper/fri/p1228-reddy.pdf)

Using some relational database system (e.g. Postgres), perform experiments that allow you to reproduce a plan diagram like Figure 2(a) which focuses on TPCH Query 8 and a cost diagram like 1(b). This will involve loading the TPCH benchmark database, running many parameterized versions of Query 8, parsing/analyzing the EXPLAIN output, and then producing plots. You should assess the results you generate to see if the interesting features observed in the paper occur, for example, complex patterns, non-monotonic costs, duplicates, islands, footprints, etc.

Extensions:

1. Run with an alternative database system (especially a commercial system if you used an open source system above)
2. Evaluate one or more additional queries on the same database.

**Project 3: SampleJoin**

[*Zhao et al. Random Sampling over Joins Revisited. SIGMOD 2018.*](https://www.cs.utah.edu/~lifeifei/papers/samplejoin.pdf)

Implement the framework for random sampling from joins so that you can instantiate the various algorithms considered in the paper (e.g. Extended Olken, Exact Weight, etc). Reproduce Figure 3 (on TPCH data) and perform the statistical test described in Sec 6.4 to validate that the samples are uniform (using query QX).

Extensions:

1. Also reproduce Figure 4 (on social graph data).
2. Reproduce the scalability evaluation presented in Fig 5.