# AI for DB (AI4DB)

**Homework 3 (AI4DB Problems, Techniques, and Challenges)**

**Instructor:** Xiang Lian

In this homework, there are 5 questions, covering the learned indexes and learning-based approaches for cardinality/cost estimations. The maximum mark for this homework is **120 points**, which will be later scaled (to **12 points**).

1. **[AI4DB Problems, Techniques, and Challenges, 10 points]** List pros and cons of AI4DB techniques and their challenges.

2. **[AI4DB Problems, Techniques, and Challenges, 20 points]** List two (2) AI for DB problems and their applications.

3. **[Learned Indexes, 10 points]** What triggers the structural modifications of the learned index? Please provide at least two (2) causes. Read the following reference and find out the answer.

***Reference***

Zhaoyan Sun, Xuanhe Zhou, and Guoliang Li. Learned Index: A Comprehensive Experimental Evaluation, PVLDB, 2023. URL: <https://www.vldb.org/pvldb/vol16/p1992-li.pdf>

4. **[Learned Indexes, 30 points]**

4(a). Please describe the projection-based learned indexes using the Z-order curve. **[10 points]**

4(b). For Z-order curve, with 16 × 16 pixels (or cells), please use the bit-shuffling method to find the mappings between cell locations and Z-values below. Please illustrate how to calculate the Z-value for the cell with the 2D location (6, 15) ***[note: cell locations start from 0 on each dimension]***. **[10 points]**

4(c). Given a Z-value 101 (decimal number), please identify the 2D location of the cell corresponding to this Z-value. **[10 points]**

5. **[Learning-based Approaches for the Cardinality Estimation, 30 points]** Please read Section 6 (Model Training) of the following paper, and describe the processing of the model training, including, but not limited to, the format of training data and the loss function used.

***Reference***

Yaoshu Wang, Chuan Xiao, Jianbin Qin, Xin Cao, Yifang Sun, Wei Wang, and Makoto Onizuka. 2020. Monotonic cardinality estimation of similarity selection: A deep learning approach. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. 1197–1212.

**Bonus Question** **[20 points]** Challenge yourself to answer this bonus question for extra credits.

**[Learned Bloom Filter]** Read Section 5.1 of the following paper and answer the question: How to set threshold *t* in the learned bloom filter?

***Reference***

Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, Neoklis Polyzotis. The Case for Learned Index Structures. SIGMOD Conference, 2018.



## Submitting Your Assignment

*All work must be your own. Copying other people’s work or from the Internet is a form of plagiarism and will be prosecuted as such.*

You may submit a Microsoft Word (.docx) document as an attachment using the **Canvas** Assignment tool. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

* Select this assignment in **Canvas** Assignment tool where you submit your work.
* You can submit multiple times, and only the last submission attempt will be considered for grading.
* Submissions sent by email will NOT be accepted.