以下两项数据来自网址：[Searching on Sorted Data - Harvard Dataverse](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JGVF9A&version=4.0)

注：Mit在kraska中提供的数据集均为二进制文件，这样是为了减少数据读取占用的时间以及内存占用。这些数据集都用了特殊的符号对于样本X与带预测值y进行分割

1. Book\_200M\_uint32：是mit开源提供的一项关于图书的，已经经过排序处理的数据集。内部是用32位模式下的unsigned int 图书信息。这部分数据集在利用rust环境下的cargo命令进行实验时，采用了线性回归层linear与三阶样条函数cubic相结合的方式进行优化测试。
2. Wiki\_ts\_200M\_uint64：是mit开源提供的来自wiki百科的已经排序的数据集，内部是64 位模式下的unsigned int。此数据集在测试时，用采用jeff论文中各种底层的模型的组合优化，其中包括线性回归，逻辑回归，cdf方法，样条函数。

经过实验发现，jeff提出的这一方法在多次训练后，组合形成的这一套底层结构大小是基本确定的，约为384M，他们最终的准确度会随着数据集的不同有着小幅度变化，约为0.0001。

1. 其余数据集：200M, 350M, 7M, 1.9G这四个：利用自己编写的create函数生成，其中加入了适当的噪音，制定了他们的数据分布。

200M：随机分布

350M：指数分布

1.9G：正态分布

注：自动模型选择部分的实验，内存占用主要在数据集处理部分（因为kraska直接用的二进制，所以他们没有这个消耗）。经测试得出，8G内存下测试1.9G数据集，内存占用达到600M。数据处理时间达到了30分钟以上，需要优化。