大数据算法课程（2020年春）

丁虎

huding@ustc.edu.cn

中国科学技术大学，计算机科学与技术学院

基本内容（10-12周）

1. 时间复杂度+渐近性分析
2. 常用概率不等式
3. 降维方法：PCA/近似PCA, JL-Transform
4. JL-Transform的应用：矩阵乘法、压缩感知、SVD、线性回归
5. 最近点查询：kd-tree, Quad-tree, R-tree; LSH; product quantization
6. 机器学习基础：VC dimension/ PAC learning, AdaBoost, multiplicative weights update method
7. 聚类及近似算法：k-means, k-means++, hierarchical clustering, DBSCAN
8. 支持向量机：Gilbert algorithm, nu-SVM, soft margin SVM, SMO, Core-SVM, Kernel method
9. 核心集技术，次线性算法，流算法
10. 分布式计算，通信复杂度

考核标准

- **Report+Presentation**
  - 可以两人一组，或者单人一组（标准一样，但最后评分会对单人组有一定照顾）
  - 可以在老师指定的主题里面任选一个，提交中文报告，内容包括对于现有方法的总结回顾，论文实验复现（报告中提供复现的细节及代码），创新和改进。
  - 评分：总结回顾+实验复现+创新改进+现场答辩表现，各占25%
  - 答辩时间：最后3-5周，每组15分钟左右（视具体组数进行调整），包括现场提问
  - 优秀率不超过40%
  - 严禁抄袭（其他组报告或者已经发表的论文和代码），一旦发现，记为不及格

助教

- 秦睿哲 red46@mail.ustc.edu.cn
- 方佳艳 jyfang@mail.ustc.edu.cn

参考教材

- Foundations of Data Science， Avrim Blum, John Hopcroft, and Ravindran Kannan https://www.cs.cornell.edu/jeh/book.pdf
- Mathematic fundation for data analysis, Jeff Phillips http://www.cs.utah.edu/~jeffp/M4D/M4D.html