DSAA5002

Data Mining and Knowledge Discovery in Data Science

Li, Jia

DSA Thrust, Information Hub
HKUST Guangzhou

Fall Term

Seq 1, 2025

This is the DSAA5002 L1, E4 102.

There is another 5002 L2, Lecture Hall C, Week 1 by Prof. Jing Tang, and from Week 2 by Prof. Jeffrey Xu Yu.

Instructor

LI,JIA W2-605 jialee@ust.hk
Office hour: Wes, 3:30PM-4:30PM

TA

Zhang, Qifan qzhang297@connect.hkust-gz.edu.cn
Peng, Miao peng885@connect.hkust-gz.edu.cn
Huang, Feiyu fhuang743@connect.hkust-gze.edu.cn
Linghu, Han hlinghu866@connect.hkust-gz.edu.cn

Time and Venue

Mon 3:00PM – 5:50PM E4 102

Course Page

https://sites.google.com/view/lijia/courses/dsaa5002

Prerequisites

Data Structure and Algorithms

- Decision Tree
- Hierarchical Clustering

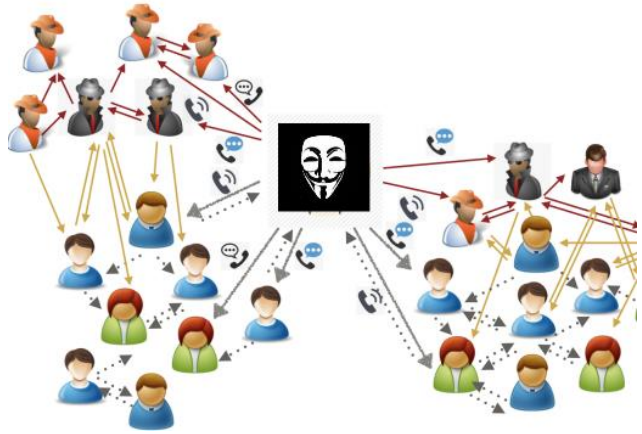
Linear Algebra

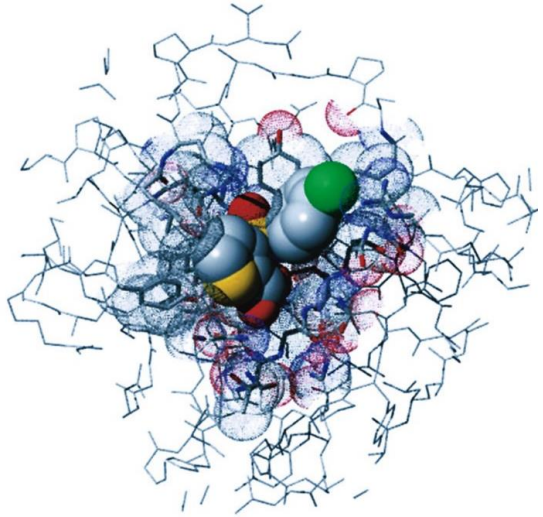
- Spectral Clustering

Probability Theory

- GMM HMM
- Expectation Maximization

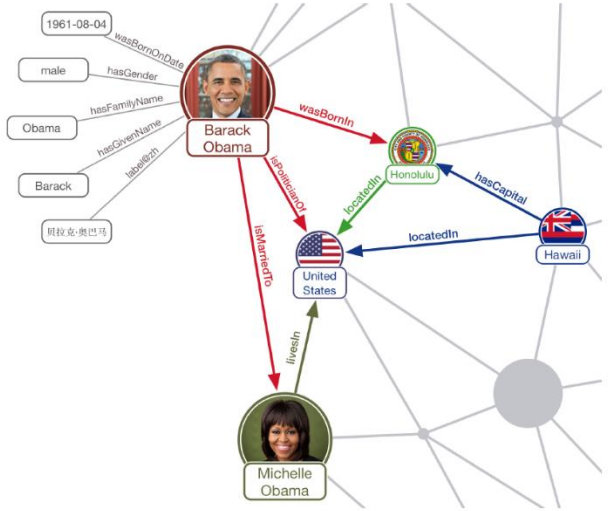
Data Mining Applications

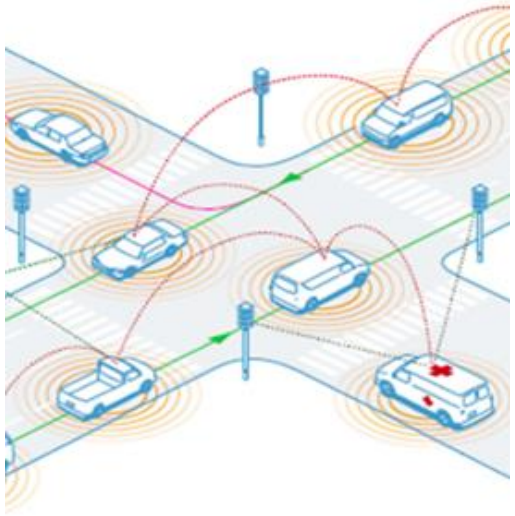




PaLM 2

ChatGPT





In Risk Management

Does the model fit into real scenario?

Is the model robust/efficient enough?

Sense-making

Assessment Scheme

Midterm Examination (50%) Early Nov, open book with only printed material (No Internet Access), samples will be provided by exercises (1, 2, 3).

Individual Project (50%) Late Dec

- Presentation (25%) oral presentation with slides, 5 min each; (about 50% students to present online, the remaining ones need to upload videos.)
- Report (25%)

Project Requirement

A research topic related to course material, ACM format with strict 6 pages limitation, see the following for reference

https://kdd.org/kdd2021/calls/view/call-for-research-track-papers

1.The report should at least consist of introduction, related work, methodology and experiment. Theoretical deviation is not a necessity but encouraged.

2. Use concise and clear language.

3. Clearly declare your difference with previous works.

4. If there is any theoretical deviation, check your assumption and make sure it is non-fragile.

Report will be graded based on:

- Writing (25%)
- Novelty (25%)
- Experiment (25%)
- Others (25%)

Outline

Data
Types of data
Normalization/cleaning
Similarity

Classification Models
Decision tree
Generalization theorem
SVM
Ensembles
Kernel methods

Cluster Models
K-means
Hierarchical clustering
Spectral clustering

Association Analysis
Apriori

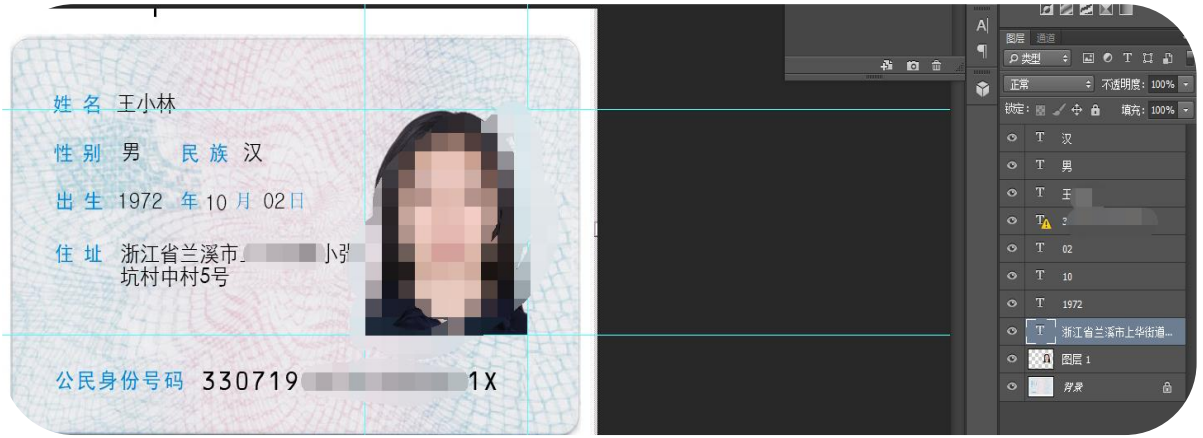
Graph Analytics
PageRank
HITS and SimRank

Expectation Maximization
GMM/HMM
Topic models

Dimension Reduction
PCA

Anomaly Detection

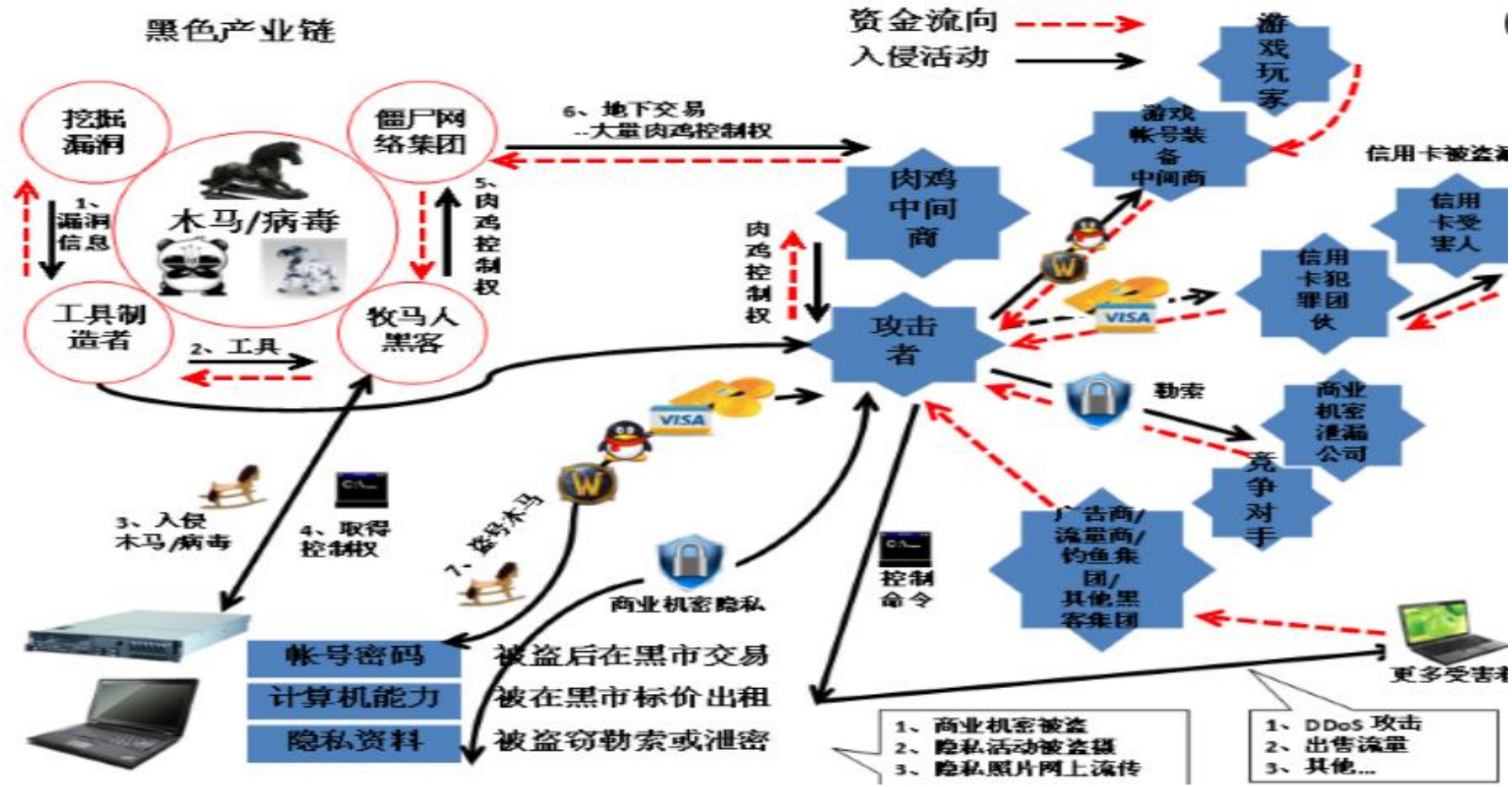
Data Mining is NOT Just a Course



Data Mining is NOT Just a Course

Data Mining is NOT Just a Course

Data Mining is NOT Just a Course



Data Mining is NOT Just a Course

Data Mining is NOT Just a Course

Data Mining is NOT Just a Course

Q&A