# SDSC3002 Intro to Data Mining

Yu Yang
yuyang@city.edu.hk

# Outline

What is Data Mining?

SDSC3002

# Data, Information & Knowledge

- Data
    - Phenomenon observed by people, due to some underlying mechanism
- Information
    - Organized data that has meaning and value
- Knowledge
    - The concept of understanding information based on identified patterns that provide insights to applications

# Data Mining

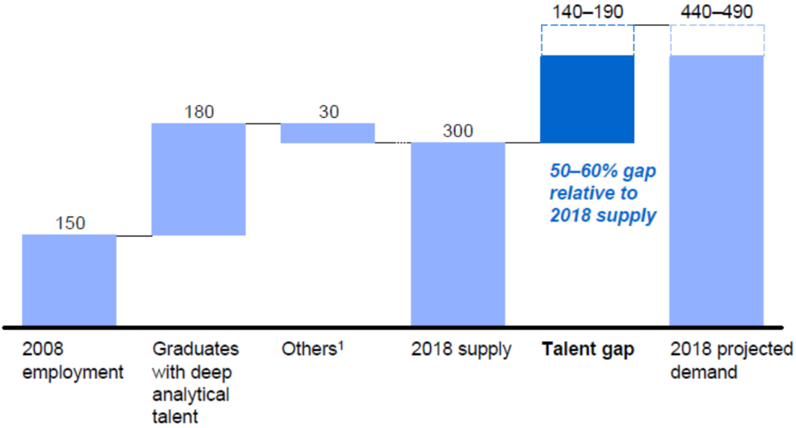
- **K**nowledge **D**iscovery from **D**ata
  - Synonym of data mining
  - ACM KDD is the best data mining conference
- Identifying interesting patterns/knowledge from (big) data
- Example: Recommender Systems
  - Data: users' browsing/purchasing/like data
  - Knowledge: a mapping function that maps a (user, item) pair to a clicking/purchasing probability
  - Insight: displaying related and personalized items to users to enhance user stickiness and increase revenue

# Demand for Data Mining

**Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018**

**Supply and demand of deep analytical talent by 2018**

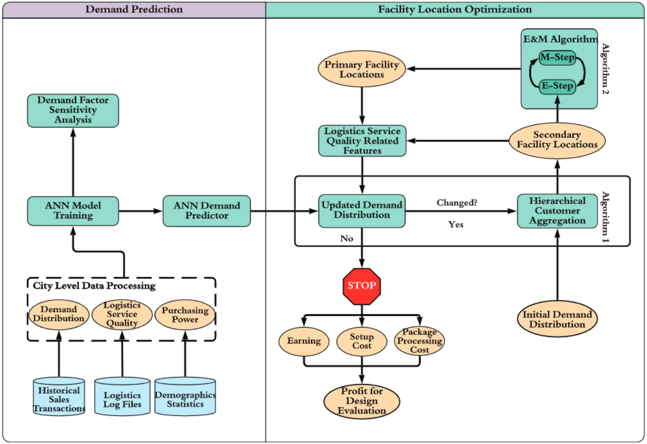
Thousand people



# Data Mining Tasks

- Descriptive task
  - Find human-interpretable patterns that describe the data
  - Examples: itemset mining, community detection
- Predictive task
  - Use some variables to predict unknown or future values of other variables
  - Examples: spam email detection, recommender systems
- Complex task
  - Combine descriptive/predictive methods with decision making methods
  - Examples: item bundling, warehouse inventory design

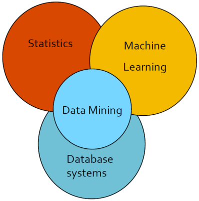
# Example: Data-Driven Distribution Network Design


Demand Prediction
Facility Location Optimization
Demand Factor Sensitivity Analysis
ANN Model Training
ANN Demand Predictor
City Level Data Processing
Demand Distribution
Logistics Service Quality
Purchasing Power
Historical Sales Transactions
Logistics Log Files
Demographics Statistics
E&M Algorithm
M-Step
E-Step
Algorithm 2
Primary Facility Locations
Logistics Service Quality Related Features
Secondary Facility Locations
Updated Demand Distribution
Changed?
Yes
No
Hierarchical Customer Aggregation
Algorithm 1
STOP
Initial Demand Distribution
Earning
Setup Cost
Package Processing Cost
Profit for Design Evaluation

Liu, Junming, et al. "Iterative Prediction-and-Optimization for E-Logistics Distribution Network Design." INFORMS Journal on Computing (2021).

# Data Mining: Cultures

- Sub-communities in Data Mining
  - Database: big data, “simple” queries
    - DM as complex queries, scalability/efficiency
  - Machine Learning: small data, complex models
    - DM as inference of models, prediction accuracy
  - Theory: (randomized) algorithms



# Related Courses

- SDSC2102 Statistical Methods and Data Analysis
  - 3002 covers non-statistical methods and more “practical” methods
- SDSC3006 Fundamentals of Machine Learning I
  - Machine learning is a very important tool in Data Mining, large overlap
  - “practical” machine learning
  - More business applications
- SDSC3001 Big Data: The Arts and Science of Scaling
  - 3002 is more application-driven

# Outline

What is Data Mining?

SDSC3002

# Contents

- Data Mining Methodologies
  - Similarity/Distance
  - Clustering
  - Graph Mining
  - Data Privacy
- Data Mining Applications
  - Market Basket Analysis
  - Recommender Systems
  - Online Advertising
  - Social Network

# Course Logistics

- Course Website:
  `https://canvas.cityu.edu.hk/courses/46765`
- Office hour: 14:50-15:50 every Wednesday (starting from week 2)
- Textbook: Mining of Massive Datasets, *by Jure Leskovec, Anand Rajaraman, Jeff Ullman*
- Grading Scheme
  - 4 Assignments: 10% each
  - Group Project (max. 4 students per group): 20%
  - Midterm: 10%
  - Final exam: 30%

# Acknowledgement

- Some of the contents originate from Jure Leskovec's slides for CS246Stanford