Today

- Meeting everybody in class
- Course topics
- Course logistics
Instructor

Li Xiong
- Web: [http://www.mathcs.emory.edu/~lxiong](http://www.mathcs.emory.edu/~lxiong)
- Email: [lxiong@emory.edu](mailto:lxiong@emory.edu)
- Office Hours: M 2-3pm W 11:15-12:15pm or by appt
- Office: MSC E412
About Me

http://www.mathcs.emory.edu/~lxiong

• Undergraduate teaching
  - CS170 Intro to CS I
  - CS171 Intro to CS II
  - CS377 Database systems
  - CS378 Data mining

• Graduate teaching
  - CS550 Database systems
  - CS570 Data mining
  - CS573 Data privacy and security
  - CS730R/CS584 Topics in data management – big data analytics

• Research
  - Data privacy and security
  - Spatiotemporal data management
  - health informatics

• Industry experience (software engineer)
  - SRA International
  - IBM internet security systems
TA
Meet everyone in class

- Group introduction (2-3 people)
- Introducing your group
  - Name
  - Goals for taking the class
  - Something interesting to share with the class
Today

- Meeting everybody in class
- Course topics
- Course logistics
Evolution of Sciences

- Before 1600, **empirical science**
  - Knowledge must be based on observable phenomena
  - Natural science vs. social sciences
- 1600-1950s, **theoretical science**
  - Motivate experiments and generalize our understanding (e.g. theoretical physics)
- 1950s-now, **computational science**
  - Traditionally meant simulation (e.g. computational physics)
  - Evolving to include information management
- 1960-now, **data science**
  - Flood of data from new scientific instruments and simulations
  - Ability to economically store and manage petabytes of data online
  - Accessibility of the data through the Internet and computing Grid
  - Scientific information management poses Computer Science challenges: acquisition, organization, query, analysis and visualization of the data

Evolution of Data and Information Science

- 1960s:
  - Data collection, database creation, network DBMS

- 1970s:
  - Relational data model, relational DBMS implementation

- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases

- 2000s
  - Stream data, web technology (XML, data integration), social network

- 2010s
  - Data science, deep learning, social media
"THAT’S your Ark for the Big Data flood? Noah, you will need a lot more storage space!"
The 5 V’s of Big Data

1. **Volume**: The size of the data.
2. **Velocity**: The speed at which the data is generated.
3. **Variety**: The different types of data.
4. **Veracity**: The trustworthiness of the data in terms of accuracy.
5. **Value**: Just having Big Data is of no use unless we can turn it into value.
Transforming the world with data

- Precision medicine
- Enriched daily lives and social systems
Value of Data

- Precision medicine
Value of Data

- GPS traces, call records
- Syndromic surveillance, social relationships
Value of Data

- Shopping history
- Recommendations
What the class is about
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining really means knowledge mining
  - We are drowning in data, but starving for knowledge!
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, information harvesting, business intelligence, etc.
Knowledge Discovery (KDD) Process

Data Cleaning

Data Integration

Databases

Data Warehouse

Task-relevant Data

Selection and transformation

Data Mining

Pattern Evaluation

Knowledge
Data Mining: Confluence of Multiple Disciplines

Data Mining

- Machine Learning
- Statistics
- Visualization
- Other Disciplines
- Database Technology
- Artificial Intelligence
Data Mining Functionalities

- Predictive: predict the value of a particular attribute based on the values of other attributes
  - Classification
  - Regression

- Descriptive: derive patterns that summarize the underlying relationships in data
  - Pattern mining and association analysis
  - Cluster analysis
Evolution of Data Analytics
Class Topics

- Classical data mining and machine learning algorithms
  - Frequent pattern mining
  - Classification
  - Clustering
- Data mining applications and challenges
  - Data streams and spatiotemporal data (data velocity)
  - Truth discovery (data veracity)
  - Recommender systems
Frequent pattern mining and association analysis

- **Frequent pattern**: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
  - Frequent sequential pattern
  - Frequent structured pattern

- **Applications**
  - Basket data analysis — Beer and diapers
  - Web log (click stream) analysis
  - DNA sequence analysis

- **Challenge**: efficient algorithms to handle exponential size of the search space

- **Topics**
  - Algorithms: Apriori, Frequent pattern growth, Vertical format
  - Closed and maximal patterns
  - Association rules mining
Cluster and outlier analysis

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Unsupervised learning (vs. supervised learning)
  - Maximizing intra-class similarity & minimizing interclass similarity

- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
  - E.g. Extreme large purchase
Clustering Analysis

Topics
- Partitioning based clustering: k-means
- Hierarchical clustering: classical, BIRCH
- Density based clustering: DBSCAN
- Model-based clustering: EM
- Cluster evaluation
- Outlier analysis
Classification and prediction

- Classification: construct models (functions) that describe and distinguish classes for future prediction
- Prediction/regression: predict unknown or missing numerical values
- Derived models can be represented as rules, mathematical formulas, etc.

Topics
- Classification: Decision tree, Bayesian classification, Neural networks, Support vector machines, kNN
- Regression: linear and non-linear regression
- Ensemble methods
Spatiotemporal data mining

- Trajectory mining
- Time series
- Applications
  - Mobility study
  - Traffic prediction
  - Location recommendation
Recommender Systems

Too much information!!
Big Data and Privacy

“I’m Big Data, and this is my friend No Privacy.”
Top-10 Data Mining Algorithms (ICDM 2006)

- #1: C4.5 Decision Tree - Classification
- #2: K-Means - Clustering
- #3: SVM – Classification
- #4: Apriori - Frequent Itemsets
- #5: EM – Clustering
- #6: PageRank – Link mining
- #7: AdaBoost – Boosting
- #7: kNN – Classification
- #7: Naive Bayes – Classification
- #10: CART – Classification
Top 10 Machine Learning Algorithms (Kdnuggets 2017)

1. Linear regression
2. Logistic regression
3. k-means
4. SVMs
5. Random Forest
6. Matrix Factorization/SVD
7. Gradient Boosted Decision Trees
8. Naïve Bayes
9. Artificial Neural Networks
10. Bayesian Networks
Today

- Meeting everybody in class
- Course topics
- Course logistics
Textbooks

- Data mining: concepts and techniques. J. Han, M. Kamber, Jian Pei. 3rd edition

- Mining of massive datasets. J. Leskovec, A. Rajaraman, J. Ullman
  - Online version: [http://www.mmids.org](http://www.mmids.org)

- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, 2013

- P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Wiley, 2005
Data (Mining) Conferences

- Data mining
  - SIGKDD, ICDM, SDM, CIKM, PAKDD ...
- Data management
  - SIGMOD, VLDB, ICDE, EDBT, CIKM ...
- Machine learning
  - ICML, NIPS, AAAI, ...
Workload

- ~4 programming/written assignments
  - Implementation of classical algorithms and competition!
- 1 open-ended course project (team of up to 3 students) with project presentation
  - Application and evaluation of existing algorithms to interesting data, data challenges
  - Design of new algorithms to solve new problems
- 1 midterm
- 1 final exam
Grading

- Assignments: 40
- Final project: 15
- Midterm: 20
- Final: 25
Late Policy

- Late assignment will be accepted within 3 days of the due date and penalized 10% per day
- 2 late assignment allowances, each can be used to turn in a single late assignment within 3 days of the due date without penalty.
Some expectations

- Participate in class, think critically, ask questions
- Start on assignments and projects early
- Enjoy the class!
Today

- Meeting everybody in class
- Course topics
- Course logistics

- Next lecture: data preprocessing