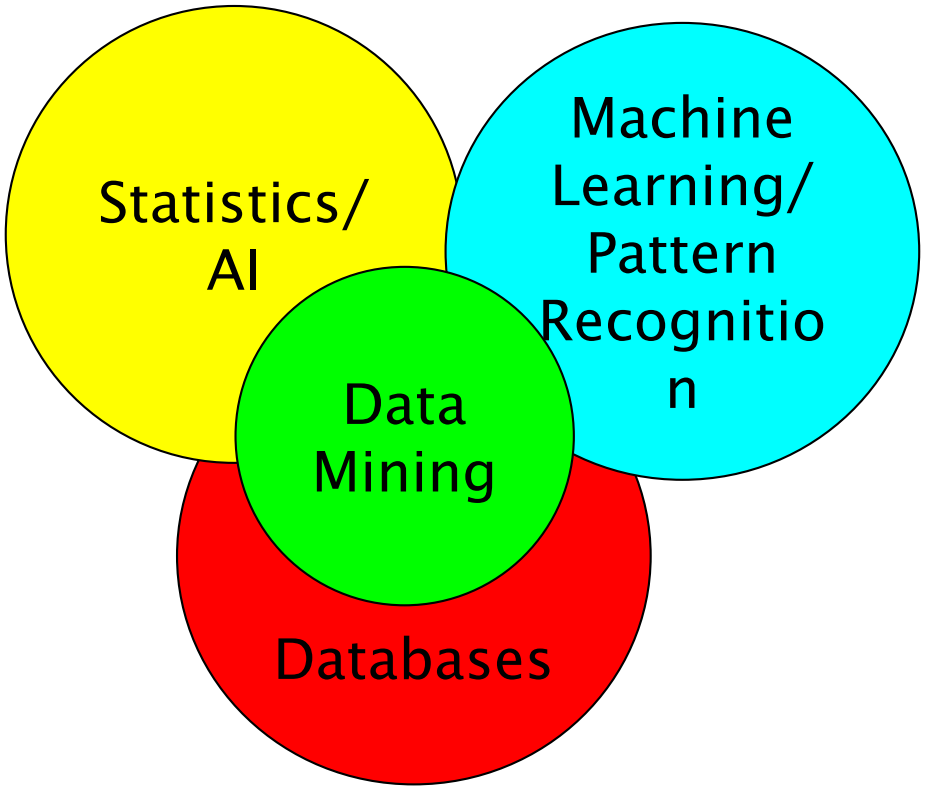
CS426 REVIEW

Spring 2017

MINING MASSIVE DATASET

- Overlaps with machine learning, statistics, artificial intelligence, databases, but more stress on
  - **Scalability** of number of features and instances
  - **Algorithms** and **architectures**
  - Automation for handling **large data**



WHAT WE HAVE COVERED

- OpenMP
- Pthread
- MPI
- MapReduce
- DISC

- Association rule discovery
- Dimension reduction
- Information retrieval
- Clustering
- Classification
- Finding similar items
- Recommendation systems
- Search engines
- Link analysis
- Advertising on the Web
- Mining data streams

MAIN TOPICS

- MapReduce
- Association rules
- Apriori
- PCY
- Frequent itemsets
- Recommender systems
- PageRank
- SVM
- Perceptron
- Naïve Bayes
- kNN
- LSH
- MinHash
- Simularity
- k-means
- BFR
- CURE
- SVD
- Matrix factorization
- Collaborative Filtering
- Stream sampling

HOW IT ALL FITS TOGETHER

- Based on different types of data:
  - Data is **high dimensional**
  - Data is a **graph**
  - Data is **never-ending**
  - Data is **labeled**
- Based on different models of computation:
  - **MapReduce**
  - **Streams**
  - **Batch (offline) vs. Active (online) algorithms**
  - **Single machine in-memory**

HOW IT ALL FITS TOGETHER

- **Based on different applications:**
  - **Recommender systems**
  - **Market basket analysis**
  - **Link analysis**
  - **spam detection**
  - **Duplicate detection** and **similarity search**

- **Based on different "tools":**
  - **Linear algebra:** SVD, Matrix factorization
  - **Optimization:** Stochastic gradient descent
  - **Dynamic programming:** Frequent itemsets
  - **Hashing:** LSH