Data Mining

MIE524

Eldan Cohen

Slides modified from Leskovec, Rajaraman, Ullman (http://www.mmds.org)

Data contains value and knowledge

Data Mining

- **But to extract the knowledge data needs to be**
  - **Stored**
  - **Managed**
  - **And ANALYZED ← this class**

Data Mining ≈ Predictive Analytics ≈ Data Science ≈ Machine Learning ≈ Data-Centric AI

What is Data Mining?

- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

This class: MIE524

Data Mining

This class: MIE524

Algorithms for Scalable Data Mining and Machine Learning

This class: MIE524

Algorithms for Scalable
Data Mining
and Machine Learning

What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is labeled
  - Data is infinite/never-ending (?)

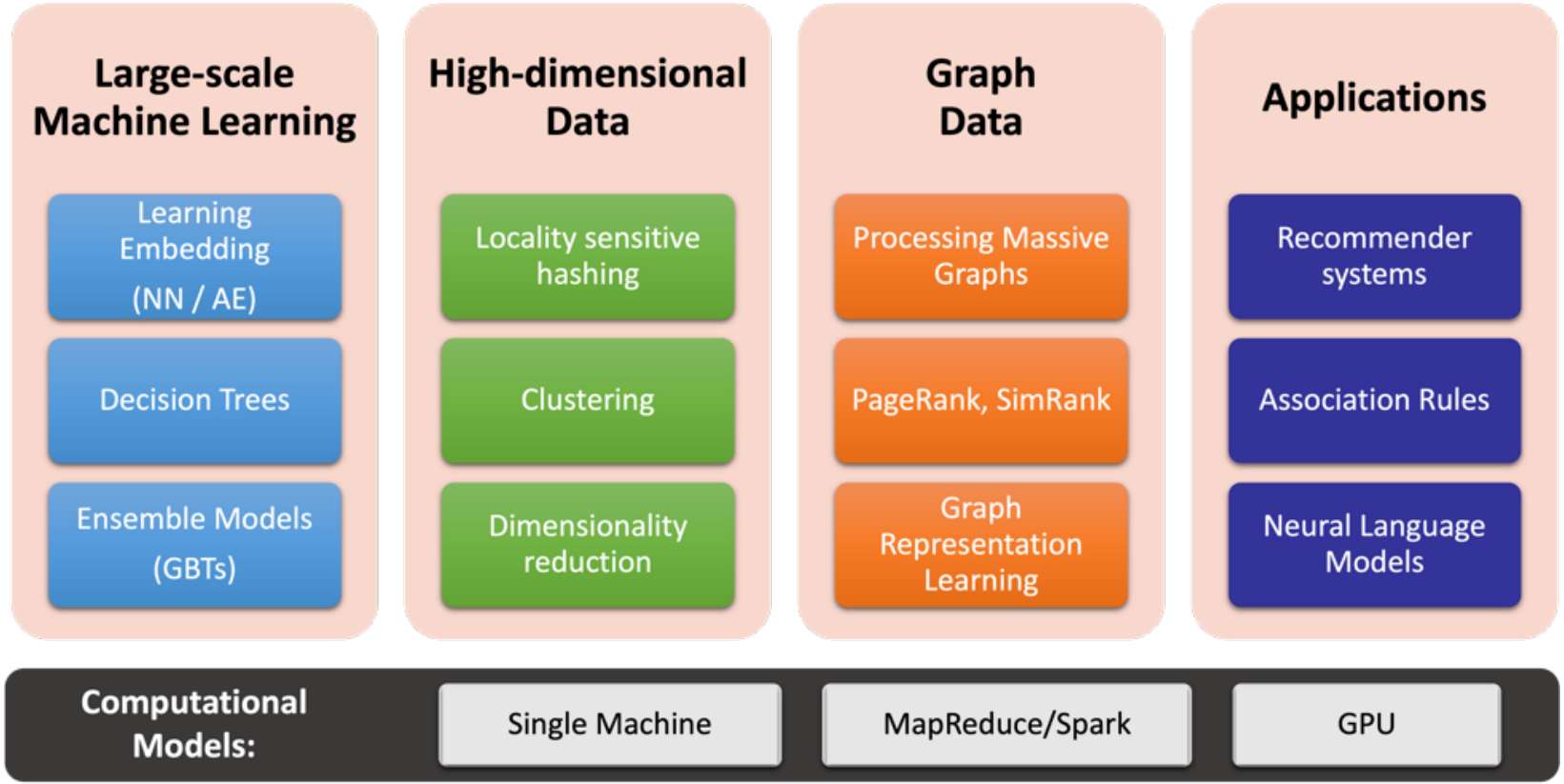
- **We will learn to use different models of computation:**
  - Single machine in-memory
  - MapReduce/Spark
  - GPU, mini-batch

What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Association Rules
  - Tentative: Language Models

- **We will learn various "tools":**
  - Linear algebra (SVD, Rec. Sys.)
  - Optimization (stochastic gradient descent)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH)

MIE524: Course Topics (Tentative)



MIE1520: Learning with Graphs and Sequences

- Graduate research course
- Winter 2025
- Deep neural architectures for sequences and graph data
  - Transformers, Language Models
  - Graph Neural Networks (GNNs)
  - Advanced training techniques (e.g., using RL) and inference algorithms (e.g., beam search and variants).
- Research focus

About The Course

MIE524: Where to Look

- **Course logistics**
  - Syllabus (on Quercus)
  - Quercus (Announcements, Pages, …)
- **Lectures**
  - Slides posted on Quercus
  - Reading posted on Quercus
- **Assignment and Lab Links**
  - Quercus Modules
- **All questions**
  - Office Hours (lecture materials), Piazza (assignment questions)
- **Quizzes**
  - During lab time, after tutorials

How NOT to succeed in the course?

- **Lectures & Labs**
  Skip *lectures* and/or *labs*
- **Assignments**
  Miss assignment submission *deadlines (see assignment instructions)*
  Don’t do assignments yourself
- **Quizzes**
  Don’t attend quizzes *(missed quiz = 0)*
- **Office Hours and Piazza**
  Be inactive on *Piazza* discussion forum
- **Quercus**
  Ignore *Quercus* announcements and update notifications

Labs and Assignments

- Alternate Labs
  - Week 0 (this week):
    Optional Python tutorial
  - Week 1 (next week):
    Post Assignment #1
    Tutorial on new assignment
    No code review (first week)
  - Week 2:
    Supervised lab time (by TA) to work on the assignment
    In some cases: additional tutorial material
  - Week 3:
    Post Assignment #2
    Tutorial on new assignment
    Quiz on previous Assignment #1 material
  - ….

Labs and Assignments

- Assignments and Labs:
  - Python + Jupyter notebooks
  - Environment: ECF + Google Colab
    - Verify access before lab
  - Why Python:
    - Rich libraries make it tool of choice for data analysis in industry
    - Portable (not OS dependent) and shareable (e.g., Jupyter)
  - Why Jupyter notebooks:
    - Easy to structure code and re-run parts of code
    - Interactive data science environment (code, visualization, analysis (text)). Can be exported as PDF/HTML report
    - Very popular, used in industry, extended to other languages.

Labs and Assignments

- Assignments
  - Can involve coding and free-text questions, implementation of algorithms, application of algorithms, data analysis.
  - All submitted code and answers should be your own work.
  - Co-pilot and other generative AI solutions are not allowed.
  - Submission via GitHub (instruction on first lab).

- Late submission
  - Projects submitted up to 48 hours late will be given a 30% late penalty (of assignment maximum mark). Projects submitted 48 hours late or more will be given a mark of zero.

Labs and Assignments

- Quizzes
  - Either a **Quercus Quiz** or a one-to-one quiz with TA/instructor
  - Can include different questions, e.g., coding questions, free text questions, multiple choice, etc.
  - Material related to the assignment (both theoretical background and coding skills)
  - To succeed:
    - Complete the assignment
    - Understand your code
    - Understand your answers
    - Understand the (theoretical and practical) material related to the assignment

Labs and Assignments

- Assignment plan (**tentative**):
  - Assignment 1: Spark, Association rules
  - Assignment 2: Gradient boosting, NN autoencoders
  - Assignment 3: Locality-sensitive hashing, Dimensionality reduction
  - Assignment 4: Graph processing
  - Assignment 5: Neural language models

This course requires you to write substantial code in Python

Grading through a combination of manual assignment grading + post-assignment quizzes

The course does not teach Python.
Optional python tutorial during lab today

Grading

- Assignments:
  - Deliverables: 20% of your final grade
  - Post-assignment in-class quizzes: 15% of your final grade
- Midterm: 25% of your final grade
- Final Exam: 40% of your final grade
  - **The Final Exam is mandatory and will result in course grade of incomplete (INC) assigned on the transcript if not attempted.**

Quercus Page

- Syllabus
- Assignment List
- Readings
- Lecture slides
- Lab materials
- Piazza (soon)
- ...

Textbook

Mining of Massive Datasets 3rd Edition. Leskovec, Rajaraman, and Ullman. Cambridge University Press. 2020.

- Available online: http://www.mmds.org/

- Readings from other textbooks may be provided

Communication

- **Questions on course content**
  - Office hours (tentatively): Wednesday 2-3pm in BA8106
  - Additional hours will be scheduled before midterm/final
- **Questions on homework assignments**
  - During labs
  - On Piazza
- **Advice on career, grad school, AI/ML projects, etc.**
  - Happy to meet with MIE524 students to discuss the above or anything else
  - Email to set up a meeting

To-Do: Complete GitHub Information Form

Due date: September 12

Link on **Assignment List** on Quercus