Scalable ML

The course discusses advances in reasoning and learning over large and heterogeneous data.

Administrative

- **Time**: 4-5:20pm Tuesday/Thursday
    - **Place**: Zoom Meeting ID 927 8648 5093
- **Instructor**: [Arash Termehchy (mailto:termehca@eecs.oregonstate.edu)](mailto:termehca@eecs.oregonstate.edu)
- **Office hour**: Thursday 3-4pm Zoom Meeting ID 342-275-2195
- **Discussion forum** : Slack

Prerequisites

Students should have good programming skills and mathematical maturity. CS 540, CS 440 or equivalent courses are required. Please contact the instructor if you are not sure whether you have the required background.

Communications with Instructor

Students are *strongly encouraged* to ask questions from the instructor and participate in the discussions in order to understand the course materials. They should use the class Slack channel to post their questions. If students do not want to publicly share their questions, they may direct message the instructor on Slack.

Evaluation

Paper Review

All students should read all posted papers before the class and direct message a short, around 300-words, summary of the paper to the instructor on Slack before 12:00 pm on the day the paper is presented in the class. The students may and are encouraged to read the paper together. But, each student should submit a separate review. Late submissions are not accepted. Each student may skip two papers without any penalty. The presenter of a paper does not need to submit any review for the paper. Each review should answer the following questions about the paper:

- What is the problem discussed in the paper?
- Why is it important?
- What are the main ideas of the proposed solution for the problem?

Paper Presentation

The purpose of presenting papers is for students to learn the concepts and methods discussed in the papers deeply and improve their skills in understanding and presenting scientific papers. After the first five lectures, each student will present one paper of his/her choice in the class. Students should prepare their presentation carefully and make sure that the audience understand the main ideas of the paper. Students should finish their presentations with a ***critical assessment*** of the paper: ***its strong points, weak points, and possible extensions.***

Students should select their papers from the schedule of the course within ***the first two weeks of the class.*** The students will be graded based on the ***their quality and clarity of presentations, their knowledge of the material, and their abilities to initiate and lead discussions*** about the paper in the class and Slack. Presenters should share their power point slides by 5:00 pm a week before the presentation date to the instructor on Slack. Each presentation will take about 45 minutes, which leaves about 30 minutes for questions and discussions about the paper and its topic.

Discussion

Each presentation will follow with a discussion about the paper, its strong points, weak points (mostly), and possible extensions. All students should participate in the discussions and ask at least one question each week in the class.

Research Survey and Proposal

Students should pick a specific topic relevant to the papers and subjects discussed in the course and write a survey paper on the topic. The survey paper should describe and analyze the important recent works in the area. It should also proposed and describe a couple of interesting open problems related to its topic. The survey may be done in groups of 2 students. Students should form their groups and find the subject for their survey within the first three weeks of the class. The group will present their survey in the last week of the class. The surveys are graded based on

- The completeness of the survey and its coverage of the latest ideas on the topic.
- The analysis of the strengths and weaknesses of the current approaches.
- The interesting open problem(s) with some reasonable and concrete ideas for the solutions.

Grading Scheme

All grade disputes must be received via Slack within **one week** of receiving the grade. We use the following grading scheme.

- Paper review - 20%
- Paper presentation -25%
- Participation and discussion -10%

- Survey - 45%

Grade **Average**

A	**90 or greater**
A-	**87 - 89**
B+	**83 - 86**
B	**80 - 82**
B-	**77 - 79**
C+	**73 - 76**
C	**70 - 72**
C-	**67 - 69**
D+	**63 - 66**
D	**60 - 62**
D-	**57 - 59**
F	**less than 57**

List of Recommended Papers

Learning & Reasoning Over Graphs & Relational Data

Arbitrary-Order Proximity Preserved Network Embedding (https://www.kdd.org/kdd2018/accepted-papers/view/arbitrary-order-proximity-preserved-network-embedding) . KDD, 2018

Hyperbolic geometry for low-dimensional knowledge graph embeddings (https://grlearning.github.io/papers/101.pdf) , ACL, 2020

Hyperbolic Attention Networks, (https://arxiv.org/pdf/1805.09786.pdf) 2018

Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks (https://dl.acm.org/doi/10.1145/3318464.3389742) , SIGMOD 2020

Semi-Supervised Classification with Graph Convolutional Networks (https://arxiv.org/abs/1609.02907) , ICLR, 2017

metapath2vec: Scalable Representation Learning for Heterogeneous Networks (https://www.kdd.org/kdd2017/papers/view/metapath2vec-scalable-representation-

(https://www.kdd.org/kdd2017/papers/view/metapath2vec-scalable-representation-
learning-for-heterogeneous-networks), KDD 2017

[HetGNN: Heterogeneous Graph Neural Network](http://www.shichuan.org/hin/time/2019.KDD%202019%20Heterogeneous%20Graph%20N) (http://www.shichuan.org/hin/time/2019.KDD%202019%20Heterogeneous%20Graph%20N, KDD, 2019

[Differentiable Inductive Logic Programming for Structured Examples](https://arxiv.org/abs/2103.01719) (https://arxiv.org/abs/2103.01719), AAAI 2021

[SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver](https://arxiv.org/abs/1905.12149) (https://arxiv.org/abs/1905.12149), ICML 2019

Survey: [A Survey on Network Embedding, IEEE TKDE](https://ieeexplore.ieee.org/document/8392745) (https://ieeexplore.ieee.org/document/8392745), 2019

Survey: [Deep Learning on Graphs: A Survey](https://arxiv.org/pdf/1812.04202.pdf) (https://arxiv.org/pdf/1812.04202.pdf), 2019

Tutorial: [Modeling Data With Networks + Network Embedding: Problems, Methodologies and Frontiers](https://ivanbrugere.github.io/kdd2018/) (https://ivanbrugere.github.io/kdd2018/), KDD 2018

Tutorial: [Multi-modal Network Representation Learning](https://chuxuzhang.github.io/KDD20_Tutorial.html) (https://chuxuzhang.github.io/KDD20_Tutorial.html), KDD 2020

Position paper: [Relational inductive biases, deep learning, and graph networks](https://arxiv.org/pdf/1806.01261.pdf) (https://arxiv.org/pdf/1806.01261.pdf), 2018

Data Augmentation, Preparation, & Communication

[Data Programming: Creating Large Training Sets, Quickly](https://arxiv.org/abs/1605.07723) (https://arxiv.org/abs/1605.07723), NeurIPS 2016

Snorkel: Rapid Training Data Creation with Weak Supervision (https://arxiv.org/abs/1711.10160) , PVLDB 2018

Learning to Compose Domain-Specific Transformations for Data Augmentation (http://arxiv.org/abs/1709.01643) , NeurIPS 2017

A Group-Theoretic Framework for Data Augmentation (https://arxiv.org/abs/1907.10905) , NeurIPS 2020

Synbols: Probing Learning Algorithms with Synthetic Datasets, (https://proceedings.neurips.cc/paper/2020/hash/0169cf885f882efd795951253db5cdfb-Abstract.html) NeurIPS 2020

Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases (https://proceedings.neurips.cc/paper/2020/hash/22f791da07b0d8a2504c2537c560001c-Abstract.html) , NeurIPS 2020

RandAugment: Practical Automated Data Augmentation with a Reduced Search Space (https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html) , NeurIPS 2020

Unsupervised Data Augmentation for Consistency Training, (https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html) NeurIPS 2021

Multi-agent cooperation and the emergence of (natural) language (https://arxiv.org/pdf/1612.07182.pdf) , ICLR 2017

Deep Learning for Entity Matching: A Design Space Exploration (http://pages.cs.wisc.edu/~anhai/papers1/deepmatcher-sigmod18.pdf) ,SIGMOD 2016

Deep Communicating Agents for Abstractive Summarization (https://arxiv.org/abs/1803.10357)

Human-Data System Interaction

[Cooperative Inverse Reinforcement Learning (http://people.eecs.berkeley.edu/~russell/papers/russell-nips16-cirl.pdf)](http://people.eecs.berkeley.edu/~russell/papers/russell-nips16-cirl.pdf), NeurIPS, 2016

[Inverse Reward Design (http://people.eecs.berkeley.edu/~russell/papers/nips17-ird.pdf)](http://people.eecs.berkeley.edu/~russell/papers/nips17-ird.pdf), NeurIPS 2017

[On the Utility of Learning about Humans for Human-AI Coordination (https://arxiv.org/pdf/1910.05789.pdf)](https://arxiv.org/pdf/1910.05789.pdf), 2020

Tutorial: [The Neuroscience of Reinforcement Learning (http://www.princeton.edu/~yael/ICMLtutorial2009)](http://www.princeton.edu/~yael/ICMLtutorial2009), ICML 2009

Learning Scalable Algorithms and Data Structures

[ALEX: An Updatable Adaptive Learned Index (https://dl.acm.org/doi/10.1145/3318464.3389711)](https://dl.acm.org/doi/10.1145/3318464.3389711), SIGMOD 2020

[DBEst: Revisiting Approximate Query Processing with Machine Learning Models (https://dl.acm.org/doi/10.1145/3299869.3324958)](https://dl.acm.org/doi/10.1145/3299869.3324958), SIGMOD 2019

PGMJoins: Random Join Sampling with Graphical Models, SIGMOD 2021

[Multidimensional Adaptive & Progressive Indexes (https://www.inf.ufpr.br/eduardo/papers/ICDE2021.pdf)](https://www.inf.ufpr.br/eduardo/papers/ICDE2021.pdf), ICDE 2021

Students with Disabilities

Accommodations are collaborative efforts between students, faculty and Disability Access Services (DAS). Students with accommodations approved through DAS are responsible for contacting the faculty member in charge of the course prior to or during the first week of the term to discuss accommodations. Students who believe they are eligible for accommodations but who have not yet obtained approval through DAS should contact DAS immediately at (541) 737-4098."

Students with documented disabilities who may need accommodations, who have any emergency medical information the instructor should be aware of, or who need special arrangements in the event of evacuation, should make an appointment with the instructor as early as possible, and no later than the first week of the term. Class materials will be made available in an accessible format upon request

Tentative Schedule

This is the anticipated schedule and may be updated over the term.

Date	Topic	Reading
March 30	**Overview (https://canvas.oregonstate.edu/courses/1811614/files/85982796/download?download_frd=1)**	Course introduction & l
Background		
April 1	**Learning Over Structured Data:** Graph Analysis	**Authorative Sources i et al. 1998. (https://w** Review Needed] **The PageRank Citatio Page et al. 1999 (htt**
April 6	**Learning Over Structured Data:**	
April 8	**Human-AI Interaction:** Human Learning	**The Data Interaction G (https://dl.acm.org/doi/1**
April	**Data Preparation**	**Learning Over Dirty D (http://web.engr.oregons**
April 15	**Learning Algorithms & Data Structures**	**The Join Game (http://web.engr.oregons**
April 20		

April 22		
April 27		
April 29		
May 4		
May 6	Midterm Survey Presentation	
May 11		
May 13		

May 18		
May 20		
May 25		
May 27		
Survey & Proposal Presentation		
June 1	Survey Presentations	

June 3	Survey Presentations	