Machine Learning for Graphs and Sequential Data

Introduction

Lecturer: Prof. Dr. Stephan Günnemann

www.daml.in.tum.de

Summer Term 2021

Data Analytics and
Machine Learning

Roadmap

- Chapter: Introduction

  1. **What will you learn in this lecture?**
  2. Organizational aspects + project tasks

What is this course about?

- In short: A continuation of our intro ML lecture (IN2064) now focusing on advanced learning principles and covering more complex data domains

- Focus on algorithms and general principles, not limited to a single domain

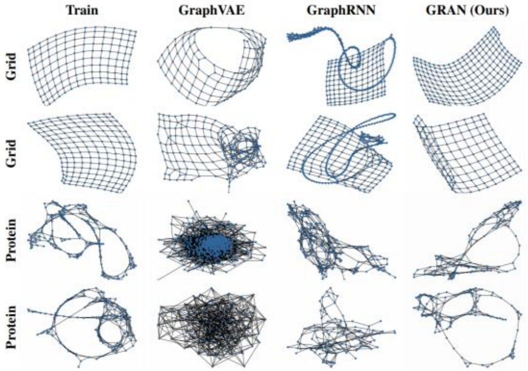
- Project tasks will give you hands-on experience

➢ At the end you should also be able to extend existing techniques and adapt them to applied problems

How do you learn complex distributions?

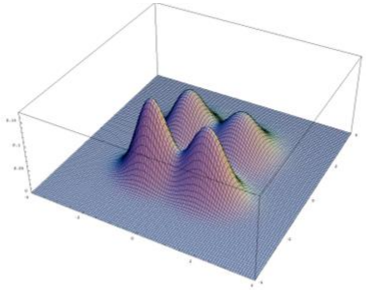
- How can we learn probability distributions over complex real-world data such as images, graphs, and audio signals?

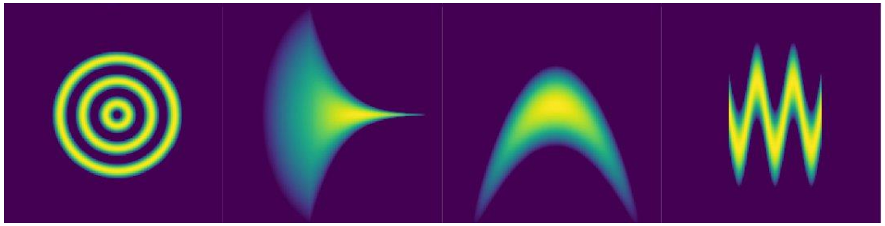
[Khan, 2019]



[Liao+, 2019]

- Distributions in high dimensions defy our intuition

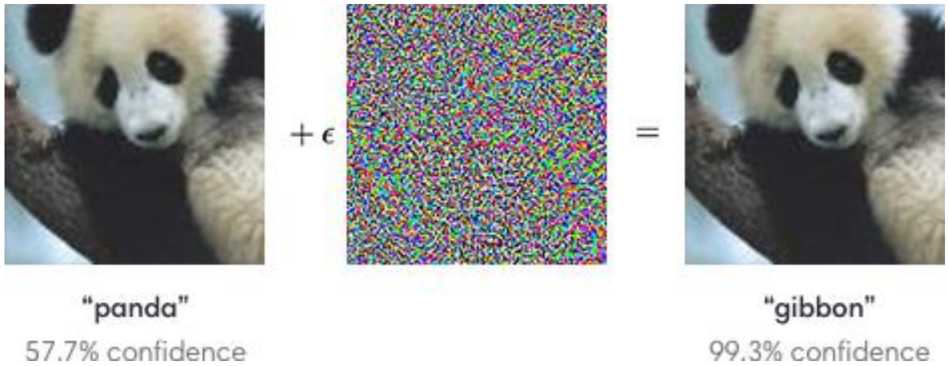




- How do we design flexible and efficient models for these settings?

Is Deep Learning fragile?

- Neural networks can be tricked by small perturbations



[Goodfellow14]

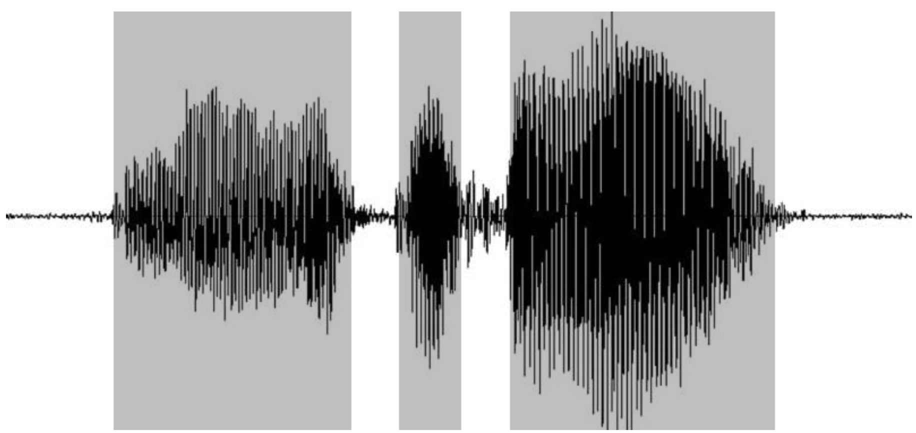
- How can we find such a perturbation?
- How do you measure the susceptibility of a model to these attacks?
- Can we ensure the robustness of our models against such adversarial examples, e.g. through specialized training procedures?

How do you learn from sequences?

- How can we understand written language?

Duchess Kate PERSON did wear an Alexander McQueen BRAND dress ITEM to the wedding OCCASION in the fall of 2017 SEASON . [Akbik18]

- How do you detect patterns in sound waves?

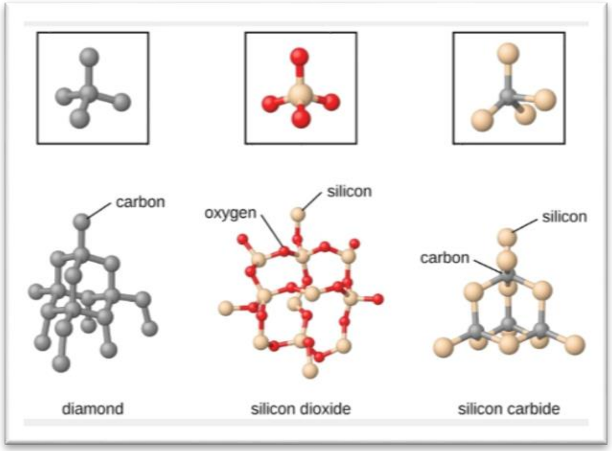


Welcome *to* *MLGS*

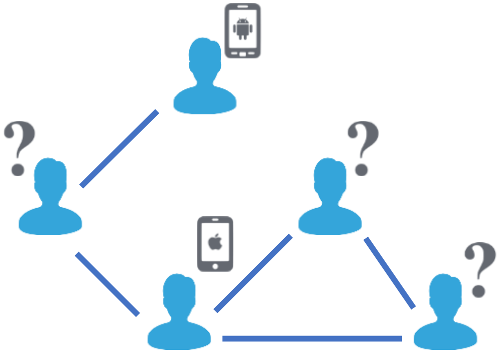
- How can we predict the future?

How to predict…?

- …the properties of molecules?
  - instances are not vectors but graphs!



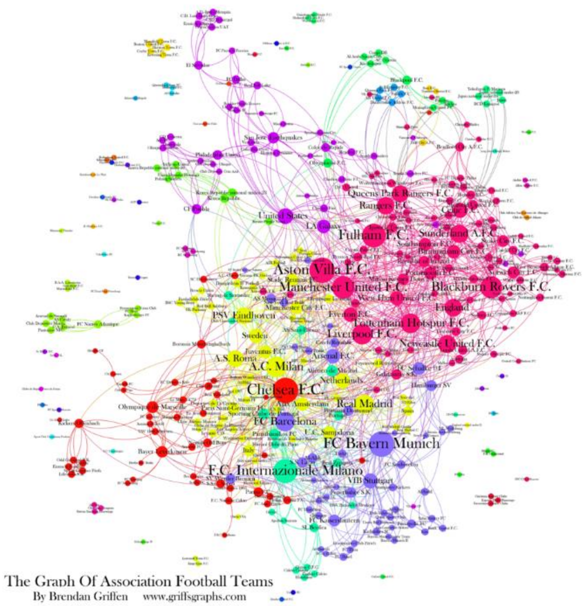
- …the preference of a user in a social network?
  - instances are connected by a graph



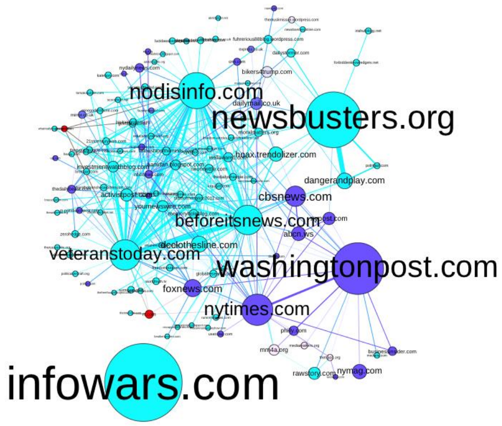
How to use/incorporate the graph structure for classification?

How to exploit graph/network data…?

- … to find communities of people?



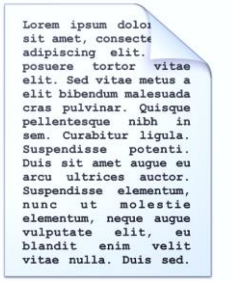
- … to detect relevant websites?



← or the most favorite football clubs ☺

The Challenge of Non-I.I.D. data

- Classical Machine Learning approaches assume data instances to be i.i.d.
  - Independent, identically distributed
- However, most real world datasets show dependencies
  - Temporal data/sequences
    - Text: sequence of words (discrete values),
      Sensor measurements: time series (continuous values)
    - Current value depends on previous values
  - Graphs/networks
    - Already seen before: molecules, social networks, knowledge graphs, connected devices (IoT)
    - Arbitrarily complex dependency structure possible



- How to design machine learning techniques that consider these dependencies (and thus, hopefully, lead to better results)?

Contents of this Course

- Deep Generative Models
  - Normalizing flows
  - Variational inference
  - VAE, implicit models, GAN
- Robustness
  - Adversarial examples
  - Attacks & defenses, certificates
- Temporal & Sequential Data
  - Autoregressive models, HMM
  - Deep learning on sequences: RNN, LSTM, embeddings
  - Temporal point processes
- Graphs
  - Laws & generative models
  - PageRank
  - Unsupervised representation learning: spectral clustering, embeddings
  - Supervised learning: label propagation, GNNs & their limitations

Roadmap

- Chapter: Introduction
  1. What will you learn in this lecture?
  2. **Organizational aspects + project tasks**

Course Organization

- Lecturer
  - Prof. Dr. Stephan Günnemann
- Teaching assistants
  - Marin Biloš, Bertrand Charpentier, Simon Geisler, Marten Lienen, Jan Schuchardt, Oleksandr Shchur

- 5 ECTS
- Language: English

- Weekly ungraded exercise sheets
- Graded project tasks every 2-3 weeks

- Final exam + repeat exam
- Project tasks can grant a bonus of up to 0.5

Schedule

- Due to COVID-19
  - Lecture and tutorial videos released each Monday via Moodle
  - External materials such as videos and papers
  - Q&A sessions via Zoom: Thursdays 4:00pm – 6:00pm

- Practice material and exercises uploaded to Moodle
  - Ungraded exercise sheets weekly
  - Project tasks every 2-3 weeks

Timetable

Week		Topic	Project
1	Apr 19	Normalizing Flows	
2	Apr 26	Variational Inference	Normalizing Flows
3	May 03	VAE + GAN	
4	May 10	Robustness 1	
5	May 17	Robustness 2	Robustness
6	May 24	AR + Markov Chains	
7	May 31	Hidden Markov Models	
8	Jun 07	RNN + Embeddings	Word2vec
9	Jun 14	Temporal Point Processes	
10	Jun 21	Graph Laws	
11	Jun 28	PageRank	PageRank
12	Jul 05	Node Embeddings	
13	Jul 12	Label Propagation, GNN	GNN
14	Jul 19	Limitations of GNNs	

Prerequisites

- The course is designed for Master students of Computer Science (and specializations such as Data Engineering and Analytics, Games Engineering, etc.)

- Prerequisites:
  - Knowledge about the standard Machine Learning concepts (i.e. content of our lecture IN2064)
    - We assume the basic concepts are clear; no repetition!
    - We strongly recommend that you attend IN2064 first before taking this class

  - Knowledge about:
    - Algorithms and Data structures
    - Programming
    - Mathematics: Linear Algebra, Statistics, Optimization

Course Material + Announcements

- Video lectures and project submission on Moodle
- All course materials (slides, exercises, project descriptions) will be uploaded to Moodle
- Use Piazza to ask questions! (please avoid sending e-mails)

https://piazza.com/tum.de/summer2021/in2323

Access Code: **mlgs2021**

- Please read the guidelines for using Piazza

Exercises and Project Tasks

- Exercise sheets
  - Exam preparation
  - Solutions in the tutorials
  - Due to the high number of registrations, we are unable to provide corrections to your solutions

- Project tasks
  - Get hands-on experience with advanced machine learning methods
  - Improve your final grade! (details later)

Project Format

- Project Format
  - data files
  - jupyter notebook with task description and boilerplate code such as data loading, plotting
- Python + (data) science ecosystem
  - Pytorch
  - Numpy / Scipy
- Groups of up to 3 members (free formation)
- Deliverables
  - One filled-in jupyter notebook per group
    - Well documented, runnable & self-contained
    - Any results and plots included (i.e. submit with output)

Project topics

- Five project tasks on the following topics
  - Normalizing flows
  - Adversarial examples
  - Word2vec
  - PageRank, spectral methods
  - Graph neural networks

- The specific tasks and all details will be described in the corresponding exercise sheets

Graded online exercise & final grade

- Graded online exercise: 90 minutes
  - Date will be announced via TUMonline
  - Conducted via TUMexam
  - Open book

```
def final_grade(exam_grade, project_grade):
    if exam_grade > 4.0:
        return exam_grade
    else:
        bonus = max(0.0, 4.0 - project_grade) / 6
        return max(1.0, exam_grade - bonus)
```

→ The project is voluntary and can only improve the final grade. The project bonus applies only if you passed, and you cannot improve beyond 1.0

Our Group's Focus

Reliable Machine Learning for **Non-Independent** Data

Robustness/Uncertainty

- Data corruptions, adversaries
- Certificates

Non-independent data

- Temporal/sequence data
- Graph data

- Interested? We offer:
  - Bachelor/Master theses, Guided Research projects, HiWi positions
- More details on specific topics closer to the end of the semester

References

Figures taken from

- Goodfellow et al. 2014, https://arxiv.org/abs/1412.6572
- Akbik et al. 2018, https://research.zalando.com/welcome/mission/research-projects/flair-nlp/
- Khan 2019, https://heartbeat.fritz.ai/stylegans-use-machine-learning-to-generate-and-customize-realistic-images-c943388dc672
- Liao et al. 2019, https://arxiv.org/abs/1910.00760