

Introduction to Data Science

Logistics for the Lecture

Ziawasch Abedjan & Marius Lindauer



Summer Term 2022;



; based on [DS100]

Large parts of the material (i.e. slides, figures and exercise) were created at [Berkeley] (University of California). We thank them very much for allowing us to re-use their material.

Warning:

- ▶ We changed material...
- ▶ We condensed it...
- ▶ We extended it...
- ▶ We fixed it...

You will be able to ...

1. **identify** steps needed to apply data science
2. **explain** different data processing steps required in data science
3. **choose** a promising combination of approaches to build data science pipelines
4. **evaluate** data science pipelines on different datasets
5. **visualize** data and results in data science
6. **apply** data science to new tasks at hand

- ▶ Motivation [Abedjan & Lindauer]
- ▶ Data Sampling and Probability [Abedjan]
- ▶ Data Preprocessing [Abedjan]
- ▶ Visualization [Lindauer]
- ▶ Intro to Modelling [Abedjan]
- ▶ Simple Linear Regression + Ordinary Least Squares [Lindauer]
- ▶ Feature Engineering [Abedjan]
- ▶ Bias and Variance [Lindauer]
- ▶ Evaluation, Regularization and AutoML [Lindauer]
- ▶ Classification [Lindauer]
- ▶ Inference for Modelling [Abedjan]
- ▶ Conclusion & Ethics [Abedjan]

- ▶ Concepts & details
 - ▶ We provide sufficient details s.t. you can understand and use the techniques
 - ▶ We highly recommend that you dig deeper and read additional material to become a real expert
- ▶ “Classic” lecture
 - ▶ 1 Slot lecture + 1 slot guided exercise
- ▶ Practical exercises and home works
 - ▶ implement it, use it and play with it!



Prof. Dr.
Ziawasch Abedjan



Prof. Dr.
Marius Lindauer



Mahdi Esmailoghli



Tim Ruhkopf

- ▶ Every week new exercise sheet
 - ▶ Exercise focus is aligned with lectures
 - ▶ Do the exercises in the exercise sessions (Thursdays at 11:00am s.t.)
- ▶ Most exercises will be practical, i.e., you have to implement something
 - ▶ Expected work load: 1.5h each week
- ▶ Team work highly recommended, e.g. team size of 3!
- ▶ Home work every 3 weeks, i.e. 4 times overall
 - ▶ You can get 5% as bonus points for the final exam
 - ▶ You need 66% of all homework points to get the bonus points
 - ↪ either you get all the bonus points or none at all

- ▶ Every week new exercise sheet
 - ▶ Exercise focus is aligned with lectures
 - ▶ Do the exercises in the exercise sessions (Thursdays at 11:00am s.t.)
- ▶ Most exercises will be practical, i.e., you have to implement something
 - ▶ Expected work load: 1.5h each week
- ▶ Team work highly recommended, e.g. team size of 3!
- ▶ Home work every 3 weeks, i.e. 4 times overall
 - ▶ You can get 5% as bonus points for the final exam
 - ▶ You need 66% of all homework points to get the bonus points
~> either you get all the bonus points or none at all
- ▶ Don't cheat (incl. plagiarism)
 - ▶ First time cheating: 0 points for exercise / homework
 - ▶ Second time cheating: failing the course

- ▶ Every week new exercise sheet
 - ▶ Exercise focus is aligned with lectures
 - ▶ Do the exercises in the exercise sessions (Thursdays at 11:00am s.t.)
- ▶ Most exercises will be practical, i.e., you have to implement something
 - ▶ Expected work load: 1.5h each week
- ▶ Team work highly recommended, e.g. team size of 3!
- ▶ Home work every 3 weeks, i.e. 4 times overall
 - ▶ You can get **5% as bonus points** for the final exam
 - ▶ You need 66% of all homework points to get the bonus points
 ↪ either you get all the bonus points or none at all
- ▶ Don't cheat (incl. plagiarism)
 - ▶ First time cheating: 0 points for exercise / homework
 - ▶ Second time cheating: failing the course
- ▶ Homework is not mandatory
BUT: quite unlikely that you will pass the course without doing them

- ▶ Every week new exercise sheet
 - ▶ Exercise focus is aligned with lectures
 - ▶ Do the exercises in the exercise sessions (Thursdays at 11:00am s.t.)
- ▶ Most exercises will be practical, i.e., you have to implement something
 - ▶ Expected work load: 1.5h each week
- ▶ Team work highly recommended, e.g. team size of 3!
- ▶ Home work every 3 weeks, i.e. 4 times overall
 - ▶ You can get **5% as bonus points** for the final exam
 - ▶ You need 66% of all homework points to get the bonus points
 ↪ either you get all the bonus points or none at all
- ▶ Don't cheat (incl. plagiarism)
 - ▶ First time cheating: 0 points for exercise / homework
 - ▶ Second time cheating: failing the course
- ▶ Homework is not mandatory
BUT: quite unlikely that you will pass the course without doing them

- ▶ Lecture session every Monday (14:00am s.t.) and exercise session every Thursday (11:00 am s.t.)
- ▶ Use the forum in StudIP for all kind of questions
- ▶ Don't send us emails
 - ~> Only in case of emergencies

- ▶ Basics in **Statistics** (mandatory)
 - ▶ We will cover many concepts, but you need a basic understanding of the underlying math.
- ▶ Programming in **Python** (mandatory)
 - ▶ all exercises will require that you implement something in Python
 - ▶ We will show you basics at the beginning in the exercises.
However if you never used Python before, it could get quite hard for you.
- ▶ **English** (mandatory)
 - ▶ You can ask us any question also in German. However, we will reply in English. So, you need to understand us ;-)

Final Project Exam – Tentative Plan!

- ▶ Written Exam
- ▶ Show us that you understood the concepts
- ▶ Be a master of all the algorithms we showed you
- ▶ Be able to read and check code
- ▶ ...

- ▶ There are also awesome online courses (MOOCs), teaching you many concepts
 - ▶ Data science is such a big field
 - ▶ Don't expect that you will learn everything in a single lecture / course
 - ↪ there might be parts you can only learn in our lecture (e.g., AutoML)
 - ▶ Applied Data Science with Python Specialization by the University of Michigan
 - ▶ Machine Learning by Andrew NG
- ▶ Kaggle for online competitions, datasets and code
 - ▶ If you can become a grand master at Kaggle, you have very good chances for job offers
 - ▶ Participate in competitions, look for help in forums, polish your skills!
 - ▶ Approaching (almost) any machine learning problem by Abhishek Thakur (first 4x Grant Master at Kaggle)

"Introduction to Data Science" is a basic lecture.
It is the first time we teach it at the Leibniz University Hannover.

"Introduction to Data Science" is a basic lecture.
It is the first time we teach it at the Leibniz University Hannover.

Opportunities:

- ▶ Get all the basics you need to do your own first data science projects
- ▶ Perfect foundation for other AI courses at the LUH (most of them in the masters; see next slide)

Risks:

- ▶ You will find some typos and issues in the slides;
please tell us if you find something
- ▶ We will not cover deep neural networks

→ Give us some feedback and we will improve the course!

AI Courses @ LUH

Summer

Winter

Basics
Advanced Core Lectures
Application-driven
Seminars/Labs

Lab: ML for Games (Rosenhahn)	Seminar & Lab: AI (Nejdl)	Seminar & Lab: AutoML (Lindauer)	Seminar: Human-Computer Interaction (Rohs)	...
Seminar: Adv. Topics in Database Systems (Abedjan)	Lab: Hardware Architectures for AI (Blume; starting soon)	Seminar: Computer Vision (Rosenhahn)	Pro-Seminar: Auto. Bildinterpretation (Rosenhahn)	Seminar: Reinforcement Learning (Lindauer)
Data- and learning- based control (Müller, Lilge, Lopez)	Bildgebende Systeme für die Medizintechnik (Blume, Ostermann, Rosenhahn)	Mobile Interaktion (Rohs)	Interaktive System (Rohs)	AI for Automotive Industry (Nolting)
Machine Learning for Graphs (Khosla and Funke)	Multimedia Retrieval (Ewerth)	Interpretable Machine Learning (Lindauer + Anand)	Social Responsible in AI (Lindauer; starting 2022)	Knowledge Engineering and Sem. Web (Auer)
Multi-Agent Interactions and Games (Kudenko)	Information Retrieval (Nejdl)	Reinforcement Learning (Lindauer)	AutoML (Lindauer)	Web Science (Nejdl)
Natural Language Processing (Ganguly)	Hardware Architectures for AI (Blume; starting soon)	Data Integration (Abedjan)	Deep Learning (Anand)	Computer Vision (Rosenhahn)
Artificial Intelligence I+II (Nejdl)	Databases I+II (Abedjan)	Machine Learning (Rosenhahn)	Data Science (Abedjan + Lindauer; starting 2022)	

Introduce yourself!

- ▶ What drives you for being here?
- ▶ What interests you in data science?
- ▶ Do you already have hands-on experience in data science?
- ▶ Are you looking for team members for exercises and home works?

Questions?