Data Science and AI Summer School, 5th edition

https://datascience.ase.ro

July 20-28, 2024

Predeal, Romania

Organized by Bucharest University of Economic Studies, in collaboration with the GATE Institute at Sofia University "St. Kliment Ohridski" and the projects enRichMyData, Graph-Massivizer, UPCAST, INTEND, and InterTwino.

Goal:

The goal of the summer school is to familiarize students with relevant state of the art topics in data science and artificial intelligence (AI). The program will cover fundamentals of data science and AI and focus on the following key topics:

- Data analytics and statistics
- Machine learning and deep learning
- Large Language Models and Conversational AI
- Causal AI
- Time series and graph data
- Data sharing
- Data and AI pipelines (data enrichment pipelines, machine learning pipelines)

The program will consist of a combination of lecture-style talks introducing various data science paradigms and methods, and demo/hands-on sessions. The summer school aims to have a practical orientation, with Python and Jupyter Notebooks being used to exemplify many of the topics covered at the summer school.

At the end of the summer school, the students are expected to have an understanding of key paradigms used in data science and AI and be able to practically apply them in data science and AI projects.

Prerequisites:

Familiarity with computer programming and basic knowledge about Python, interest in working with data, enthusiasm, and willingness to learn new things!

Basic knowledge of linear algebra, probability theory, and knowledge representation would be useful, though not strictly necessary.

Lecturers and instructors:

- **Dan Nicolae** (University of Chicago, USA), Professor, PhD
  https://www.linkedin.com/in/dan-nicolae-221991a
- **Razvan Bunescu** (University of North Carolina at Charlotte, USA), Associate Professor, PhD
  https://www.linkedin.com/in/razvan-bunescu-8097956
- **Anna Fensel** (Wageningen University & Research, the Netherlands), Professor, Dr.
  https://www.linkedin.com/in/anna-fensel-0862501
- **Radu Prodan** (University of Klagenfurt, Austria), Professor, PhD
  https://www.linkedin.com/in/radu-prodan-182812b1
- **Ioan Toma** (Onlim GmbH, Austria), Chief AI Officer, PhD
  https://www.linkedin.com/in/ioantoma
- **Dumitru Roman** (SINTEF / OsloMet – Oslo Metropolitan University, Norway), Senior Research Scientist / Professor, PhD
  https://www.linkedin.com/in/titiroman
- **Jože Rožanec** (Jožef Stefan Institute, Slovenia), Machine Learning Engineer
  https://www.linkedin.com/in/jmrozanec
- **Daniel Thilo Schroeder** (SINTEF / OsloMet – Oslo Metropolitan University, Norway), Research Scientist / Associate Professor, PhD
  https://www.linkedin.com/in/danielschroder555nase
- **Gabriel Terejanu** (University of North Carolina at Charlotte, USA), Associate Professor, PhD
  https://www.linkedin.com/in/gabrielterejanu/
- **Hui Song** (SINTEF, Norway), Senior Research Scientist, PhD
  https://www.linkedin.com/in/hui-song-4132b024
- **Wiktor Sowinski-Mydlarz** (London Metropolitan University, UK and GATE Institute, Bulgaria), Lecturer in Computer Science and Applied Computing, Senior Research Scientist, PhD
  https://www.linkedin.com/in/viktor-sowinski-mydlarz-32a9a5101
- **Roberto Avogadro** (SINTEF AS, Norway), Research Scientist, PhD
  https://www.linkedin.com/in/roberto-avogadro-657744a2
- **Nikolay Nikolov** (SINTEF AS, Norway), Research Scientist
  https://www.linkedin.com/in/nikolay-nikolov-a9672659

TENTATIVE SCHEDULE

<table>
<tr><td></td><td rowspan="9">Arrivals (20.07)</td><td>Day 1
(21.07)</td><td>Day 2
(22.07)</td><td>Day 3
(23.07)</td><td>Day 4
(24.07)</td><td>Day 5
(25.07)</td><td>Day 6
(26.07)</td><td>Day 7
(27.07)</td><td rowspan="9">Departures (28.07)</td></tr>
<tr><td>Breakfast (8am-9am)</td><td colspan="7"></td></tr>
<tr><td rowspan="2">Morning session
(9am-12pm)</td><td rowspan="2">Statistics for data science</td><td>Basics of machine learning (ML)</td><td>Prompting techniques and LLM APIs</td><td>Time series: Forecasting, XAI, and databases</td><td>LLMs-based data processing</td><td>LM-based agents for data science</td><td>Data pipeline scheduling on the computing continuum</td></tr>
<tr><td>Linear models for classification</td><td>Data-driven decision making</td><td>Graph data management</td><td>Conversational AI</td><td>Findable, Accessible, Interoperable, Reusable (FAIR) data</td><td>Operationalizing data and ML pipelines</td></tr>
<tr><td>Lunch break and socializing activities
(12am-2.30pm)</td><td colspan="7"></td></tr>
<tr><td rowspan="2">Afternoon session
(2.30pm-5.30pm)</td><td rowspan="2">Statistical learning</td><td>Applied Deep Learning</td><td>Causal AI</td><td>Graph data analytics</td><td rowspan="3">Social event</td><td>Best practices in data sharing</td><td rowspan="2">Deployment, orchestration, monitoring of data and ML pipelines</td></tr>
<tr><td>Basics of Large Language Models (LLMs)</td><td>Time series analysis and forecasting</td><td>Data enrichment</td><td>High performance data processing</td></tr>
<tr><td>Individual group/project work and/or free time
(5.30pm-7pm)</td><td>Intro event</td><td colspan="4"></td><td colspan="2"></td></tr>
<tr><td colspan="5">Dinner (7pm-9pm)</td><td></td><td colspan="2"></td></tr>
</table>

Statistics for data science *(Dan Nicolae)*
- A data science case study
- Foundations of data analysis
- Statistical inference with resampling methods

Statistical learning *(Dan Nicolae)*
- Probability and simulations
- Regression models and inference
- Model Complexity
- Prediction and classification

Basics of machine learning (ML) *(Razvan Bunescu)*
- Feature vector representations
- Occam's razor for ML, intelligence, and science
- Overfitting, underfitting, generalization, and regularization
- ML experiments: training, validation, and testing

Linear models for classification *(Razvan Bunescu)*
- Logistic regression, softmax, and temperature
- ML algorithms in Python: the *sklearn* library
- Linear vs. non-linear classification and deep learning

Applied Deep Learning *(Gabriel Terejanu)*
- From linear models to neural networks
- Why deep neural networks?
- What is an embedding?
- How to make use of pre-trained models?

Basics of Large Language Models (LLMs) *(Razvan Bunescu)*
- Subword tokenization, word embeddings, and neural language models (LMs)
- Encoder, encoder-decoder, and decoder LMs
- Pre-training and fine-tuning

Prompting techniques and LLM APIs *(Razvan Bunescu)*
- Zero-shot and few-shot in-context learning
- Chain-of-thought prompting, retrieval augmented generation, and ReAct
- The chat completion API and LangChain

Data-driven decision making *(Gabriel Terejanu)*
- Introduction to A/B testing for decision making
- Designing effective A/B tests
- Analytical techniques in A/B testing

Causal AI *(Gabriel Terejanu)*

- Importance of causality in AI
- What is a causal model?
- What is an intervention?
- How to estimate causal effects?

Time series analysis and forecasting *(Jože Rožanec)*

- Introduction to time series
- Analysis tools and real-world examples
- Time series forecasting

Time series: Forecasting, XAI, and databases *(Jože Rožanec)*

- Using network models to represent and forecast time series
- Introduction to explainability methods
- Introduction to time series databases

Graph data management *(Dumitru Roman)*

- Intro to graph data structure
- Knowledge Graphs
- Graph data management (graph databases with Noe4j, graph data model, graph construction and querying)

Graph data analytics *(Daniel Schroeder, Dumitru Roman)*

- Complex Network and Graph Analysis in IGraph
- Intro to Graph Neural Networks in PyG
- Graph data visualization

Data enrichment *(Roberto Avogadro, Dumitru Roman)*

- Data linking
- Tabular data enrichment
- Human-in-the-Loop (HITL) for data enrichment

LLMs-based data processing *(Ioan Toma)*

- Introduction to LLM-based Data Processing
- Knowledge Extraction using LLMs
- Document Classification, Summarization and Comparison using LLMs

Conversational AI *(Ioan Toma)*

- Conversational AI setup and designing a chatbot interface
- Semantic Knowledge Graphs and their role in Conversational AI
- Building a chatbot using Onlim Conversational AI framework

LLM-based agents for data science *(Hui Song)*

- Use of ChatGPT Data Analyst to process data files, generate data processing code
- Development of LLM-based agents for multi-phase data processing tasks
- Multi-agents for complex and collaborative data processing

Findable, Accessible, Interoperable, Reusable (FAIR) data *(Anna Fensel)*

- Introduction to FAIR data. Examples from agri-food and health domains
- How to make data FAIR? Open data, closed data and everything in between
- Research data infrastructures

Best practices in data sharing *(Anna Fensel)*

- Legal compliance (GDPR, AI Act, Data Act)
- Consent, contracts and licenses, empowered with knowledge graphs
- Incentivising data sharing

High performance data processing ***(****Radu Prodan)*

- Parallel computing architectures
- Multiprocessing
- Parallel algorithms
- Parallel computing for AI and data science

Data pipeline scheduling on the computing continuum *(Nikolay Nikolov)*

- Introduction to pipeline scheduling in the context of big data and distributed applications
- The importance and challenges of pipeline scheduling
- Solutions and practical approaches to pipeline scheduling

Operationalizing data and ML pipelines *(Wiktor Sowinski-Mydlarz)*

- Contemporary Data Processing
- GATE Institute Data Platform
- Alternatives and Decisions
- The Lifecycle

Deployment, orchestration, monitoring of data and ML pipelines *(Wiktor Sowinski-Mydlarz)*

- Data Spaces: Decentralized Supply and Consumption of Data Services
- Private Cloud for Big Data Processing
- Platform support
- Resources: Free products and software bibles

Software (preliminary)**:** Software tools/services to be used during the sessions include:

- Anaconda (https://www.anaconda.com): Installation instructions for various platforms can be found at: https://docs.anaconda.com/anaconda/install
    - A number of relevant tools and libraries that we will use can be configured from Anaconda: Python 3, NumPy, SciPy, Matplotlib, Jupyter Notebook, Ipython, Pandas, and Scikit-learn.
- Other Python packages: statsmodels, transformers, lingam
- Onlim Platform (https://app.onlim.com/): Conversational and Knowledge Graph Platform. Accounts can be created https://auth.onlim.com/auth/realms/onlim/login-actions/registration?client_id=onlim&tab_id=gmTCMEh3-6U
- Neo4j (https://neo4j.com): Installation and documentation can be found at https://neo4j.com/developer/get-started.We will use the online sandbox service provided at https://neo4j.com/sandbox, so no installation on local machines is needed for experimenting with Neo4j. Alternatively you can download and install Neo4j Desktop, which provides a convenient way for developers to work with local Neo4j databases (this can be downloaded from https://neo4j.com/download-center/#desktop). We will also use Neo4j Graph Data Science (https://neo4j.com/product/graph-data-science) which comes with Neo4j.
- Docker (https://www.docker.com): An open-source containerization platform that will be used for ML pipelines. Installation instructions can be found at https://docs.docker.com/engine/install.
- SIM-PIPE (https://github.com/DataCloud-project/SIM-PIPE): An open-source tool for dry running of Big Data Pipelines using sample data. The tool allows evaluating pipeline performance and resource requirements at scale. An open version of the tool is available on https://simpipe.sct.sintef.no.

Bios

Dan Nicolae (University of Chicago, USA), Professor, PhD

Dan Nicolae is the Elaine M. and Samuel D. Kersten, Jr. Distinguished Service Professor of Statistics at University of Chicago where he has served as chair of the Department of Statistics from 2016-2022, section chief for the Section of Genetic Medicine from 2015-2016, and is currently founding co-Director of the Data Science Institute. Originally from Craiova, Dan Nicolae graduated from "Facultatea de Matematica" of University of Bucharest in 1995, and has obtained his PhD in Statistics from University of Chicago in 1999. His research seeks to understand the role of genetic, genomic and environmental factors, and their interactions, in the development of common/complex diseases. A statistical geneticist and a mathematical statistician, he specializes in developing methodological advances for large data problems in biology and medicine. Particular interests include functional genomics, microbiome, integration of omics data, networks, and systems biology. The statistical and computational methods developed by his group are based on foundations in high-dimensional statistical inference, machine learning and data science.

Razvan Bunescu (University of North Carolina at Charlotte, USA), Associate Professor, PhD

Razvan Bunescu is an Associate Professor in the Department of Computer Science at the University of North Carolina at Charlotte. Previously, he was a Professor in the School of Electrical Engineering and Computer Science at Ohio University. He received the PhD degree in computer science from the University of Texas at Austin in 2007, with a dissertation on machine learning methods for information extraction. Prof. Bunescu has co-authored over 120 papers in the general area of machine learning, with applications in natural language processing, music information retrieval, biomedical informatics, performance efficient computer architecture, computational creativity, and more recently computer science education. His work has been funded by grants from the National Science Foundation, the National Institutes of Health, the US Air Force, and Microsoft.

Anna Fensel (Wageningen University & Research, the Netherlands), Professor, Dr.

Prof. Dr. Anna Fensel (former name: Anna V. Zhdanova) is a Full Professor in Artificial Intelligence and Data Science at Wageningen University & Research, Wageningen, the Netherlands. Previously, she was an Associate Professor and Senior Assistant Professor at Semantic Technology Institute (STI) Innsbruck, Department of Computer Science, University of Innsbruck, Austria, and prior to this worked as a Senior Researcher at FTW – Telecommunications Research Center Vienna, Austria, and a Research Fellow at the University of Surrey, UK, and as a project employee at DERI

Innsbruck, University of Innsbruck, Austria. Anna has earned both her habilitation (2018) and her doctoral degree (2006) in Computer Science at the University of Innsbruck. Prior to that, she has received a diploma in Mathematics and Computer Science equivalent to the Master degree in 2003 from Novosibirsk State University, Russia. Anna has been extensively involved in European and national projects related to semantic technologies, linked data and knowledge graphs, e.g. as a consortium coordinator (EURegio project REEeT, FFG projects TourPack, SESAME-S and SESAME, ÖAD project RESIDE), a partner project manager (HE SoilWise, H2020 OntoCommons, H2020 smashHit, Eurostars WordLiftNG, Interreg KI-Net, BMWi-FFG CampaNeo, H2020 ENTROPY, FFG DALICC, FFG OpenFridge, FP7 CSA BYTE, FP7 CA BIG, FP7 CA PRELIDA, FP7 STREP m:Ciudad, FP7 NoE PlanetData, AAL PeerAssist, FP6 IST IP SPICE, FP5 IST project Esperonto) as well as a technical contributor in other numerous projects. She has been a co-organizer or a Program Committee member of more than 100 scientific events including being in chair roles at top events such as RuleML+RR, SEMANTiCS and ESWC, an editor and reviewer for numerous journals and a project proposals evaluator for funding agencies (EU H2020 and FP6, Eureka-Eurostars, national funding). She is a (co-)author of ca. 150 refereed publications. Webpage: https://sites.google.com/site/annafensel.

Radu Prodan (University of Klagenfurt, Austria), Professor, PhD

Radu Prodan is a professor in distributed systems at the Institute of Software Technology, University of Klagenfurt. He received his PhD in 2004 from the Vienna University of Technology and was Associate Professor until 2018 at the University of Innsbruck, Austria. His research interests include performance, optimization, and resource management tools for distributed and parallel systems. He participated in numerous national and European projects. Presently he coordinates the Horizon 2020 project ARTICONF that researches a decentralized platform and ecosystem for next-generation social media applications. He authored over 200 publications and received two IEEE best paper awards.

Ioan Toma (Onlim GmbH, Austria), Chief AI Officer, PhD

Ioan Toma is the Chief AI Officer and co-founder of ONLIM GmbH, an Austrian startup focusing on Chatbots and Intelligent Assistants, being responsible for the research activities of the company. Ioan's current research areas include Conversational AI and Knowledge Graphs. Ioan received a Ph.D. in Computer Science from the University of Innsbruck, Austria and a Master's degree in Computer Science from the Technical University of Cluj-Napoca, Romania. Ioan has been involved in numerous research projects at national and European levels. He authored over 85 articles as book chapters, conference papers, workshops papers and journal articles.

Dumitru Roman (SINTEF / OsloMet – Oslo Metropolitan University, Norway), Senior Research Scientist / Professor, PhD

Dumitru Roman works as a Senior Research Scientist at SINTEF AS (Norway) and Professor at OsloMet – Oslo Metropolitan University (Norway). He has wide experience with initiating, leading, and carrying out data-driven and research-intensive projects, participating in dozens of large international projects during the past two decades in which he has collaborated with large numbers of private companies, public sector organizations, universities, and research institutes. He is currently active in the data management, data science and AI fields, focusing on innovation projects enabling data-driven business products and services.

Jože Rožanec (Jožef Stefan Institute, Slovenia), Researcher / Machine Learning Engineer

Jože is a Researcher at the Artificial Intelligence Laboratory (Jožef Stefan Institute), and a machine learning engineer. He collaborates with the American Slovenian Education Foundation, where he leads multiple activities for Fellows and Alumni. Over more than ten years, he worked in software engineering and machine learning-related roles for several companies (e.g., Mercado Libre, Navent, Globant). His research interests include machine learning methods for recommendations, fraud detection, demand forecasting, active learning, and explainable artificial intelligence (XAI).

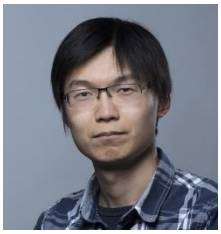
Daniel Thilo Schroeder (SINTEF / OsloMet – Oslo Metropolitan University, Norway), Research Scientist / Associate Professor, PhD

Dr. Daniel Thilo Schroeder, residing in Oslo, is a Research Scientist and Associate Professor specializing in big data analytics, complex networks and digital communication. He is a member of the Smart Data group at SINTEF, contributing to the development of sustainable platforms for processing extreme data and building high-quality, FAIR-compliant datasets that fortify the effectiveness of AI applications. In his position as Associate Professor at Oslo Metropolitan University, Dr. Schroeder explores the influence of digital communication on conflict development and mediation in sub-Saharan Africa. During his postdoc at the Simula Research Laboratory, he engaged in pivotal work to expand the application of deep learning to unstructured data through the development of computational frameworks. Earning his PhD from the Technical University of Berlin, Dr. Schroeder has been involved in multiple projects centered around understanding and mitigating the rapid spread of online misinformation.

Gabriel Terejanu (University of North Carolina at Charlotte, USA), Associate Professor, PhD

Gabriel Terejanu is an Associate Professor of Computer Science at the University of North Carolina at Charlotte, with prior academic roles at the University of South Carolina and a fellowship at the University of Texas at Austin. His industry background includes positions as a software engineer and quantitative researcher in IT and financial services. Dr. Terejanu's research focuses on causal modeling, uncertainty quantification, machine learning, and integrating physics-based models with data-driven approaches. His interdisciplinary endeavors have propelled advancements across sectors ranging from agriculture to aerospace and from materials discovery to understanding polarization on social media. He has received funding for his work from the Army Research Office, Toyota Research and Development, Lowe's Innovation Fund, the National Institute of Food and Agriculture, and the National Science Foundation. Webpage: https://www.UncertaintyQuantification.org.

Hui Song (SINTEF, Norway), Senior Research Scientist, PhD



Hui Song is a senior research scientist with SINTEF Digital, Norway. His research interest includes software and AI engineering, with a particular focus on the DevOps of software and AI applications on the cloud-edge computing continuum. He has a long experience in EU- and Norway-funded research and innovation projects, and is currently coordinating the Horizon Europe project INTEND.

Wiktor Sowinski-Mydlarz (London Metropolitan University, UK and GATE Institute, Bulgaria), Lecturer in Computer Science and Applied Computing, Senior Research Scientist, PhD

Dr. Viktor Sowinski-Mydlarz is a senior researcher at GATE and post-doctoral researcher at the Cyber Security Research Centre of London Metropolitan University. His PhD is the area of hybrid frameworks for data processing, which combine logical and machine learning methods for data analysis. He is highly experienced in software integration, particularly in containerization, orchestration and monitoring of data analytics engines on the cloud. Dr. Sowinski-Mydlarz has been working with Prof. Vassilev over the last 5 years on a number of projects, funded by Lloyds Banking Group and Innovate UK. He is currently working on the visualization of air quality in Sofia, Bulgaria on Cesium Ion and integration with GATE Data Platform. He also lectures in Cloud Computing and Internet of Things and teaches Cyber Security Fundamentals and Ethical Hacking.

Roberto Avogadro (SINTEF AS, Norway), Research Scientist, PhD

Roberto Avogadro is a Researcher at SINTEF Digital in the Smart Data group. He holds a PhD in computer science from the University of Milano-Bicocca. His research within the Smart Data Group is centered on the data linking problem, exploring how AI-based solutions can be applied to this area. Roberto has recently joined SINTEF, bringing his academic expertise into the practical realm of smart data challenges.

Nikolay Nikolov (SINTEF AS), Research Scientist

Nikolay Nikolov is a Research Scientist at SINTEF Digital Smart Data group. His current research is focused around novel methods to support the lifecycle of Big Data pipelines processing, enabling their definition, model-based analysis and optimization, simulation, and deployment on top of decentralized heterogeneous infrastructures. Nikolay has been doing applied research related to data management, data integration, data enrichment, big data, and the semantic web since 2014 as part of SINTEF Digital. During the recent years, he has been involved in research, implementation and technical coordination in the context of several national and international research projects in the area of data-driven innovation. Nikolay holds a joint Erasmus Mundus M.Sc. degree in Service Engineering from Stuttgart University, University of Crete and Tilburg University. Nikolay's main interest and focus is on approaches for supporting the lifecycle of Big Data pipelines on the Computing Continuum.

Organizing Team

1. Dan Nicolae (program co-chair) - LinkedIn
2. Razvan Bunescu (program co-chair) - LinkedIn
3. Dumitru Roman (program co-chair) - LinkedIn
4. Vasile Alecsandru Strat (general co-chair) - LinkedIn
5. Sylvia Ilieva (general co-chair) - LinkedIn
6. Ahmet Soylu (scientific advisor) - LinkedIn
7. Raluca Dana Caplescu (organization) - LinkedIn
8. Iva Krasteva (organization) - LinkedIn
9. Irena Pavlova (organization) - LinkedIn
10. Cosmin Proscanu (organization) - LinkedIn
11. Anca Bogdan (promotion) - LinkedIn
12. Dessislava Petrova Antonova (promotion) - LinkedIn
13. Orlin Kouzov (promotion) - LinkedIn