Helmholtz Summer School - From Data to Knowledge

Monday, September 16, 2024 - Friday, September 27, 2024

Overview of courses and lectures

Please find an overview of the
- content,
- prerequisites,
- elements,

- and scheduling
of each course or lecture.

The complete program can be found here.

Also beware of our Opening session and our exciting keynote on Explainable AI and its real-life applications!

Lecture 1 (HMC): "HMC FAIR Friday: The origin of data - The relevance of provenance in the context of FAIR""

Welcome to this Satellite Event!

The lecture series “HMC FAIR Friday” offers lectures on exciting aspects of FAIR data with renowned national and international speakers and is also particularly aimed at interested listeners outside HMC.

Data provenance answers the questions of why and how the data was produced, where, when and by whom. The idea and concept of provenance is about trust, credibility and reproducibility of research. Therefore, collaboration between data users and data producers is required through the provision of provenance metadata. This is important to determine the quality of the data as well as for the trust in the results, their reproducibility or the reusability of data.

Prerequisites:
No prior knowledge or skills are required.

Click on the link below to see the planned date and time for this lecture
HMC FAIR Friday: The origin of data – The relevance of provenance in the context of FAIR

→ **Register here** ←

Lecture 2 (Helmholtz AI): Fairness in Machine Learning

In this talk we survey the role of machine learning methods in questions of social justice and discrimination. First, we take a bird’s eye view on which domains may be particularly affected, how machine learning can sustain or even promote inequalities, and whether there are also opportunities for ML to help reduce or prevent discrimination in practice. Via a deep dive into

automated data-driven decision-making in consequential scenarios, we learn about the interactions of technical aspects with societal questions and introduce a broader perspective of the life-cycle of ML methods. Throughout, we try to give concrete examples of ML models arguably acting “unfair” and try to distill potential mindsets and techniques to avoid such failure modes in the future.

Prerequisites:
No previous knowledge or skills needed

Click on the link below to see the planned date and time for this lecture
Fairness in machine learning

→ **Register here** ←

Lecture 3 (HIDA): Keynote "From Idea to I did: Empowering Women Entrepreneurs"

This keynote builds the central impulse and backbone for the workshop on “empowering women entrepreneurs”, but is also **open to everyone who is interested**! Prof. Dr. Stephanie Birkner will provide valuable insights into the multifaceted landscape of female entrepreneurship, debunking common myths and exploring the critical elements that shape women’s entrepreneurial journeys. By drawing parallels between the cycles of research and startup creation, the discussion will highlight the iterative and dynamic processes that underpin successful entrepreneurial endeavors. Emphasizing the diversity of entrepreneurial thinking and actions, this talk will showcase how varied approaches contribute to a robust and innovative entrepreneurial ecosystem. Discovering and harnessing their own innovative potential, participants will understand that entrepreneurial success is not a one-size-fits-all model but a personalized journey of creativity, resilience, and strategic action. This keynote aims to inspire and equip female entrepreneurs with the insights and tools necessary to navigate and thrive in the entrepreneurial world.

Prerequisites:
No prior knowledge or skills are required.

Click on the link below to see the planned date and time for this lecture
Keynote "From Idea to I did: Empowering Women Entrepreneurs"

→ **Register here** ←

Lecture 4: Introduction to FAIR Data

The digital world forgets nothing, a common saying claims. However, anyone who has ever tried to find data that is only a few years old on the global web and then reuse it with up-to-date software knows that digital information can very well get lost. To ensure that this does not happen to valuable research data, numerous players in science are currently working on making this data as FAIR as possible - findable, accessible, interoperable and reusable.

But what exactly makes my data FAIR? What are the first steps towards this?
The lecture "Introduction to FAIR data" provides an insight into the current work of HMC and answers fundamental questions on the topic of FAIR and Metadata.

Prerequisites:
No prior knowledge or skills are required. This lecture is **open to anyone who is interested**!

Click on the link below to see the planned date and time for this lecture
Lecture: Introduction to FAIR data

→ Register here ←

Course 1 (HMC): Introduction to FAIR and Reusability of Scientific Data

In the **training course "Reusability of Scientific Data"** we focus on the 'R' in the FAIR Principles. This course was developed in collaboration with the Helmholtz Artificial Intelligence team and is specifically tailored to the Research Field Matter.

By the end of the course, participants will have a comprehensive understanding of the importance and necessity of making data reusable, a key aspect of the FAIR principles. While the course is open to everyone, it is particularly aimed at newly hired postdocs and PhD students in the Research Field Matter who are at the beginning of their research. For these researchers, it is crucial to learn how to apply data reusability methods in the early stages of their projects. They will also gain hands-on experience in selected steps necessary for making research data reusable by others.

Participants will understand the significance of the FAIR principles, particularly the reusability of data within the Research Field Matter and the AI field. Additionally, the course will feature a lightning talk on data reusability in the Research Field Matter.

Prerequisites:
No prior knowledge or skills are required.

Click on the links below to see the planned dates and times for the elements of this course
Lecture: Introduction to FAIR data
Training Course: Reusability of Scientific Data

→ Register here ←

Course 2 (HMC): Introduction to Scientific Metadata

This course is an ***entry-level introduction*** to the fundamentals of scientific metadata for PhD students, early-career researchers, and postdocs. In this course we will look at the intricate relationship between (digital) research data, metadata, and knowledge; discuss why metadata is critical in today's research; and explain some of the technologies and concepts related to structured machine-readable metadata.

Have you ever struggled to make sense of scientific data provided by a collaborator? Or, even worse, to understand your own data five months after publication... Do you have difficulties in meeting the data description requirements of your funding agency? Do you want your data to have lasting value; but don't know how to ensure that?
Precise and structured descriptions of research data are key for scientific exchange and progress - and for ensuring recognition of your effort in data collection. The solution: make your data findable, accessible, interoperable, and reusable - by describing them with metadata.

This course is targeted especially at scientific staff and researchers in the Helmholtz Research Field Information but is open to anyone who would like to better understand research data annotation with metadata.

You will learn:
- about the differences between, and the importance of data & metadata
- how to annotate your research data with structured metadata
- how to find and evaluate a suitable metadata framework and data repository
- how to use basic Markdown / JSON / XML
- which tools are already available to level up your metadata annotation game
- why structured metadata is important and how it can increase your scientific visibility

Prerequisites:
No prior knowledge or skills are required.

Click on the links below to see the planned dates and times for the elements of this course
Fundamentals of Scientific Metadata: Why context matters - PART 1
Fundamentals of Scientific Metadata: Why context matters - PART 2

→ **Register here** ←

Course 3 (HIDA / Helmholtz AI): From Idea to I Did: Empowering Women Entrepreneurs

This workshop is designed to inspire and equip early-career female scientists with the knowledge and tools necessary to explore entrepreneurship as a viable career path. Recognizing the unique challenges faced by women in a predominantly male-dominated field, the workshop aims to foster awareness, resilience, and confidence among participants. Our speakers being experts in the field, this workshop complements findings and knowledge from research with first hand practical experience that brings the academic perspective to life.

After completing this workshop you will
1. Beware of entrepreneurship as an alternative career path

2. Understand gender-specific challenges
3. be equipped with essential tools and resources to succeed
4. Recognize personal strength and build confidence

Prerequisites:
No prior knowledge or skills are required

Click on the links below to see the planned dates and times for the elements of this course
Keynote "From Idea to I did"
Workshop "From Idea to I did": Empowering Women Entrepreneurs

→ Register here ←
PLEASE NOTE THAT THE KEYNOTE IS OPEN TO ALL SUMMER SCHOOL PARTICIPANTS! REGISTER IF YOU ARE INTERESTED!

Course 4 (Helmholtz AI): Introduction to Machine Learning with scikit-learn

Welcome to the Scikit-Learn Bootcamp, a four-day workshop designed to provide data scientists with a comprehensive introduction to machine learning (ML) using the popular Scikit-Learn library in Python. This hands-on workshop will cover the entire ML workflow, from data preprocessing and feature engineering to model training and evaluation. We will cover the fundamentals of ML, including regression and classification algorithms, model evaluation metrics, and hyperparameter tuning. Participants will learn how to use Scikit-Learn python library to implement popular ML algorithms. Throughout the workshop, participants will engage in hands-on exercises and case studies to reinforce their learning and build practical ML skills. By the end of the Scikit-Learn Bootcamp, participants will have a solid foundation in ML concepts and techniques, as well as the confidence and skills to apply Scikit-Learn in real-world data science projects.

Prerequisites: In order to participate in the course successfully, we expect the participants to have a basic knowledge of Python programming (defining variables, writing functions, importing modules).
Some prior experience with the NumPy, pandas and Matplotlib libraries is recommended but not required.

Click on the links below to see the planned dates and times for the elements of this course
Introduction to Machine Learning with scikit-learn (Part 1)
Introduction to Machine Learning with scikit-learn (Part 2)
Introduction to Machine Learning with scikit-learn (Part 3)
Introduction to Machine Learning with scikit-learn (Part 4)

→ Register here ←

Course 5 (Helmholtz AI): Introduction to Statistical Learning

The class covers foundations and recent advances of machine learning techniques, including:
- Basic concepts: Linear regression, nearest neighbour, parametric vs. non-parametric methods, Bayesian classifiers, the curse of dimensionality, model accuracy, bias-variance trade-off
- Linear classifiers: linear regression for classification (discriminative model), linear discriminant analysis (generative model) - Nonlinear classifiers with Ensemble learning: Decision trees, random forests, boosting
- Unsupervised learning: Gaussian mixture models, k-means Our course aims to provide participants with not only a theoretical foundation, but also the practical skills needed to use and develop effective machine learning solutions to a wide variety of problems. We illustrate the use of the models in the tutorial throughout the course with methods implemented in Python.

Prerequisites:
The participants are expected to know linear algebra and multivariate calculus, basic concepts from linear functional analysis and basic concepts in probability theory. We ask all participants to be able to run Python (>3.5) within a Jupyter notebook on their computer or use Google Colab notebook.

Click on the link below to see the planned date and time for this course
Introduction to Statistical Learning

→ Register here ←

Course 6 (Helmholtz AI): Introduction to Uncertainty Quantification (UQ) in ML

In this half-day workshop, we will provide an overview of uncertainty quantification (UQ) techniques and their importance in developing robust and reliable machine learning (ML) models. The workshop will begin with an introduction to the basic concepts of UQ, including aleatoric and epistemic uncertainty, and their impact on ML model performance. We will then delve into popular UQ methods, such as Bayesian methods, Monte Carlo dropout, and deep ensembles, and their implementation in various ML models. The workshop will include hands-on tutorials and case studies to help participants apply UQ techniques to real-world ML problems. The goal of this workshop is to equip data scientists with the foundational knowledge and practical skills necessary to incorporate UQ in their ML models and to make informed decisions when deploying ML models in high-stakes applications. By the end of the workshop, participants will have a better understanding of the importance of UQ in ML and the tools and techniques available to quantify and reduce uncertainty in ML models.

Prerequisites: participants should be familiar with torch and neural network basics.

Click on the link below to see the planned date and time for this course
Introduction to Uncertainty Quantification (UQ) in ML

→ Register here ←

Course 7 (Helmholtz AI): Explainable Artificial Intelligence

During this course participants will get an introduction to the topic of Explainable AI (XAI). The goal of the course is to help participants understand how XAI methods can help uncover biases in the data or provide interesting insights. After a general introduction to XAI, the course goes deeper into state-of-the-art model agnostic interpretation techniques as well as a practical session covering these techniques. Finally, we will focus on two model specific post-hoc interpretation methods, with hands-on training covering interpretation of random forests and neural networks with imaging data to learn about strengths and weaknesses of these standard methods used in the field.

Prerequisites:
Basic understanding of ML models (Random Forest and CNNs). Basic knowledge of Python.

Click on the link below to see the planned date and time for this course
Introduction to eXplainable Artificial Intelligence

→ **Register here** ←

Course 8 (Helmholtz AI): A practical guide to dimensionality reduction

Dimensionality reduction is a common data preprocessing step preceding the application of supervised and unsupervised learning methods in AI modeling. After motivating the use of dimensionality reduction and highlighting its role in data exploration, this course gives an introduction to three types of dimensionality reduction approaches: feature transformation, feature aggregation, and feature selection. Course participants will have the opportunity to discover and compare the main methods for each approach in a hands-on experience, using jupyter notebooks on a real-world high-dimensional gene expression dataset.

Prerequisites: Basic knowledge of Python and Machine Learning.

Click on the links below to see the planned dates and times for the elements of this course
A practical guide to dimensionality reduction (Part 1)
A practical guide to dimensionality reduction (Part 2)

→ **Register here** ←

Course 9 (Helmholtz AI): Large Language Models

In this course, we will explain what large language models (LLMs) are and how they generally operate. We will also introduce guidelines for writing effective prompts for these models. We present 'Blabladoor,' a tool we developed for interacting with open-source language models. We will

demonstrate how to use the interface and its integration with Visual Studio4\ Code. Additionally, we will provide practical examples to illustrate the applications of LLMs.

Prerequisites:
No prior knowledge or skills are required.

Click on the link below to see the planned date and time for this course
Effective Use of Large Language Models and Prompt Engineering

→ **Register here** ←

Course 10 (Helmholtz Imaging): Regularization in Image Reconstruction: From Model to Data Driven Methods

In this course, we are going to provide the participants with knowledge about the typical mathematical tasks and caveats of image reconstruction problems. This covers advanced forward models and uncertainty, regularizing the reconstruction in order to prevent artifacts caused by noisy data and model errors, and eventually computational tasks. The participants will get the chance to test different image reconstruction and regularization schemes in the hands-on tutorial session.

Prerequisites:
Basic Knowledge in Coding, Basics of Image Reconstruction

Click on the links below to see the planned dates and times for the elements of this course
Lecture: Regularization in Image Reconstruction: From Model to Data Driven Methods
Tutorial: Regularization in Image Reconstruction: From Model to Data Driven Methods (Part 1)
Tutorial: Regularization in Image Reconstruction: From Model to Data Driven Methods (Part 2)

→ **Register here** ←

Course 11 (Helmholtz Imaging): Introduction to Image Registration

Image registration is essential for aligning multiple images—taken at different times, perspectives, or different sensors—so they match up perfectly. This process is crucial for effectively integrating, comparing and analyzing scientific data.
We'll explore key algorithms and methods, including intensity-based and feature-based approaches, and demonstrate these techniques on scientific images. The workshop will also touch on advanced topics like multimodal image registration and machine learning based approaches.

Prerequisites:
none.

Click on the link below to see the planned date and time for this course
Seminar: Intro to Image Registration

→ **Register here** ←

Course 12 (Helmholtz Imaging): 3D visualization

This seminar will focus on visualizing volumetric datasets, particularly those that are voxel-based. We'll look at various tools and strategies for displaying these datasets in 3D. You'll learn how to convert 3D segmentations into meshes and import them into Blender. We'll also go over some basic techniques in Blender to help you create beautiful volumetric data renderings. Additionally, I'll demonstrate how to upload your data for use with browser-based 3D visualization tools, making it easier for everyone to interactively explore these datasets.

Prerequisites:
none.

Click on the link below to see the planned date and time for this course
Seminar: Visualization of volumetric datasets

→ **Register here** ←

Course 13 (HIFIS): First steps with Python

This course covers the basic language and programming concepts. This fundamental knowledge is to be used as a starting point for self-guided learning during and beyond the course time.

All workshop days cover alternating sequences of theoretical input and hands-on exercises, during which the instructors are available for quick feedback and advice.

Prerequisites:
No prior knowledge or skills are required.

Click on the links below to see the planned dates and times for the elements of this course
First steps with Python (Part 1)
First steps with Python (Part 2)
First steps with Python (Part 3)

→ **Register here** ←

Course 14 (Helmholtz AI): Time-dependent Generative Models

This course is an introduction to current state-of-the-art generative models. In particular, we will focus on a family of time-dependent generative models (e.g., diffusion models), which have risen in popularity and shown unprecedented generative capabilities in recent years. These are based on a process that turns randomly sampled noise into data over time.In detail, you will learn how to construct a generative model based on a differential equation that interpolates between a noise and data distribution, resulting in both diffusion and flow matching models. Afterward, you will learn about two popular applications of these models: current state-of-the-art Text-to-Image generative models and protein generative models.This course is targeted at any researcher interested in learning about the methodology behind current state-of-the-art generative models. The goal of this course is to obtain an understanding of these models and how to train and effectively use them yourself.

Helpful prerequisites:
- Basic knowledge of Machine Learning & Deep Learning (e.g. backpropagation and neural networks)
- Basic knowledge of first-year university-level Mathematics (e.g. differential equations and matrix calculus)

Click on the link below to see the planned date and time for this course
Time-dependent Generative Models

→ **Register here** ←

Course 15 (Helmholtz AI): Fantastic Vision Language Models and how to compress them

This course provides an introductory exploration into Vision Language Models (VLMs), which leverage large language models as the backbone, and delves into techniques for their compression. We'll explore various open-source VLMs of differing sizes that have demonstrated remarkable capabilities in visual understanding and reasoning tasks.

During this course, you will:
1. Understand Vision Language Models: Gain a foundational understanding of what VLMs are and the principles behind training them.
2. Recognize Computing Costs: Learn about the computational demands of VLMs and why compression is essential for deployment on edge devices.
3. Learn Compression Techniques: Explore methods to compress VLMs, including pruning and knowledge distillation.
This session is designed for researchers with an interest in VLMs and their compression. By the end, attendees will have a solid grasp of VLMs and practical approaches to make these models more efficient.

Helpful Prerequisites:
• Basic knowledge of Machine Learning and Deep Learning (e.g., backpropagation, neural networks)

• Basic knowledge of Large Language Models

Click on the links below to see the planned dates and times for the elements of this course
Fantastic Vision Language Models and how to compress them

→ **Register here** ←

Course 16 (Helmholtz AI / HIDA): From Idea to I did: Playing lean

Followed by an introductiory lecture on Lean Methodology for Startups, this course package offers a deep dive into the lean methodology by "Playing Lean", an engaging hands-on "flight simulator" for Lean Startup and innovation. The creators of the game have partnered up with Alexander Osterwalder, inventor of the Business Model Canvas and one of the great minds of Lean Startup. The Playing Lean board game teaches highly valuable lessons on Lean Startup, creates interest in Business Model Canvas, goes deep on the Value Proposition Canvas or running lean with the Lean Canvas.

Prerequisites:
No prior knowledge or skills needed.

Click on the links below to see the planned dates and times for the elements of this course
Lecture: Introduction to Lean Methodology for Startups
Workshop - Playing Lean

→ **Register here** ←

Course 17 (HIFIS): Data processing with Pandas & Data plotting with Matplotlib

This course days will give a hands-on, fundamental introduction to the data processing framework Pandas and the data plotting framework Matplotlib. These frameworks are written in Python and very popular in all data science areas thanks to their wide variety of functionality and usability.

All workshop days cover alternating sequences of theoretical input and hands-on exercises, during which the instructors are available for quick feedback and advice.

Prerequisites
Basic knowledge of the Python language (variables, functions, loops, conditions; attaiable through the course 13 "First steps with Python").

Click on the links below to see the planned dates and times for the elements of this course
Introduction to Pandas

Introduction to Matplotlib
Hands - on Exercises with Pandas

→ **Register here** ←

Course 18 (HIFIS): Introduction to Git and GitLab

The workshop provides a practical introduction into the usage of the version control system Git in combination with the collaboration platform GitLab. The topics are presented using a live coding approach which allows you to directly try out all steps on your own.

The workshop covers the following topics:
- Initial Git Setup
- Basic Git workflow
- Feature branch workflow
- Working with remote repositories in GitLab
- GitLab contribution workflow using issues and merge requests

Prerequisites: This is an introductory workshop. No special knowledge is required.

Prerequisites:
No prior knowledge or skills are required

Click on the links below to see the planned dates and times for the elements of this course
Introduction to Git
Introduction to GitLab

→ **Register here** ←

Course 19 (HIFIS): Continuous Integration with GitLab

What is continuous integration and why do we need it? Figure the following scenario:

A team of scientists is working on a little project that takes astronaut data from Wikidata to analyse the time humans spent in space as well as the age distribution of the astronauts. The project quickly gained attraction and a lot of users as well as contributors joined the project. After some time it became hard for the maintainers to ensure new functionality is properly tested. It also frequently happened that contributors followed a different code style or forgot to add license information.

Verifying those criteria manually is tedious and not promising in the long run. This is why the team aims at automating as much as possible to save their valuable time. Luckily, they found a tool called GitLab CI which they can use to automate those tasks. This course provides an introduction to this

tool.

Prerequisites:
Participants should have basic programming skills, be familiar with the basic operations of Git, and GitLab and ideally already have made some initial experience with a Unix-shell. No further software needs to be installed.
Please make sure that you can log in to codebase.helmholtz.cloud via the Helmholtz AAI before participating in the workshop.

Click on the links below to see the planned dates and times for the elements of this course
Continuous Integration in GitLab (Part 1)
Continuous Integration in GitLab (Part 2)

→ **Register here** ←