Research Data Management

Data Security, Ethics and Legal Compliance
- Data Protection
- Privacy
- Ethical Use
- Sensitive Data
- Regulatory Compliance

Apply anonymisation, obtain informed consent, manage access securely and comply with institutional and legal standards.

Interoperability and Data Pipelines
- Application Programming Interfaces (APIs)
- PyBIS
- Automation
- Workflow Integration

Demonstrate interoperability by recording data in OpenBIS, processing it via Git-based workflows on Renku, publishing to Zenodo, and archiving securely, in accordance with FAIR principles.

Data Quality and Life Cycle
- **Data Management Plan (DMP)**
- **Data Collection**
- **Storage**
- **Processing**
- **Sharing**
- **Archiving**
- **Reuse**
- **Long-term Preservation**
- **FAIR principles**

Write a DMP to cover all aspects of the data lifecycle for your project application.

Metadata and Standards
- Metadata Schemas
- Naming Conventions
- Ontologies
- Controlled Vocabularies

Using domain standards and consistent naming conventions makes data accessible and valuable beyond the scope of a single project.

Data Governance and FAIR Priciples
- Findable, Accessible, Interoperable, Reusable (FAIR)
- Roles
- Access Management
- Licensing
- Compliance
- Open Access (OA)
- Data Provenance
- Citation Standards
- Persistent Identifiers

Defines roles such as data owners, stewards, custodians, and users to ensure accountability, data quality, and FAIR-compliant management.

RDM Software and Infrastructuvre
- Electronic Lab Notebooks (ELN)
- OpenBIS
- Renku
- Git
- Zenodo
- DMP-Tool
- Long Term Storage

Define clear data models to structure experimental and measurement data, enabling consistent storage, querying and integration with infrastructure tools like openBIS.

Data Stewardship and Training
- Guidance
- Support
- Training
- Best Practices

Participate in trainings to learn about RDM practices and openBIS usage, as well as other RDM tools.

Data Science

Signal processing
- Filtering
- Fourier Transform
- Wavelets
- Feature Extraction

Analyse the chemical composition of mixtures using FTIR: filter to remove noise and peaks are extracted as features.

Removing eye movement signals from EEG data, to reveal the underlying brain activity.

Statistical Analysis
- Descriptive statistics
- Hypothesis testing
- Analysis of Variance (ANOVA)
- Correlation analysis
- Nonparametric methods

Testing if drug treatment significantly increases cell survival.

Comparing tensile strength between two alloys using a t-test or ANOVA.

Exploratory Data Analysis
- Data visualisation
- Summary statistics
- Outlier detection
- Distribution analysis
- Missing value analysis

Analyse the voltage drop distribution per charging cycle.

Compare the thermal conductivity across different material compositions.

Time Series Analysis
- Trend analysis
- Forecasting
- Stationarity analysis
- Autocorrelation

Forecast material degradation over time to prevent failure.

Feature Engineering
- Feature Transformation
- Dimensionality Reduction
- Feature Selection
- Encoding Variables

Compute ratios, such as protein A / protein B concentrations, to reveal regulatory balances or biological relationships.

Generative AI
- **Diffusion Models**
- **Large Language Models (LLMs)**
- **Text-to-Image**
- **Text-to-Speech**

Generate new drug candidates with high binding affinity for target proteins.

Generate new suggestions for new alloys, polymers, or composite materials with the desired properties.

Computer Vision
- **Object Detection**
- **Image Segmentation**
- **Image processing**

Detect tumors or other abnormalities from medical images.

Machine Learning
- **Reinforcement/Un-/Supervised Learning**
- **Machine Learning Operations (MLOps)**
- **Model Explainability**

Predict a material's band gap before synthesis.

Model how proteins and drugs interact to guide experimental design.

Data Engineering
- **Data Pipeline**
- **Data Quality**
- **Schemas**
- **Ontologies**
- **Semantic Data**
- **Data Warehouse**
- **Data Lake**

Make data accessible for analysis and AI models.

Artificial Intelligence

AI Tools and Ecosystem
- OpenAI
- DeepMind
- Generative Pre-trained Transformer (GPT)
- AI Agents
- HuggingFace
- LangChain
- Vector Databases

Deploy an agent that crawls databases, trains a model and emails you the results.

Set up a GPT to read your lab notebooks and write an abstract for you.

Future Directions
- Quantum AI
- Synthetic Data
- AI Safety
- Artificial General Intelligence (AGI)
- Human-AI Collaboration and Societal Impact

Suggestions for experiments that humans have not yet imagined are made by AI.

Ethics, Security and Privacy
- Handling of Sensitive Data
- General Data Protection Regulation (GDPR)
- Data Protection Act
- Cybersecurity with AI
- Bias
- Explainability
- Transparency

Ensure collaborators or systems are trusted to access sensitive data, using verified identities.

Who is responsible when AI makes a mistake?

Models can hallucinate plausible sounding compounds or crystal structures that violate basic laws of nature.

Compute

Cloud Computing
- Amazon Web Service (AWS)
- Azure
- Google Cloud
- Scalability
- On-Demand Resources
- Cloud Services
- Distributed Systems
- Performance Optimisation
- Computational Science

Enable declarative, self-service deployment of software allowing to provision, update, and manage applications automatically without manual intervention.

Rent cloud GPUs on-demand to run compute-intensive tasks only when needed.

Use finite element analysis to simulate and study structural engineering problems, predicting stress, strain, and deformation before physical testing.

Platform Engineering
- Kubernetes
- Container Orchestration
- Policy as Code
- GitOps Workflows
- Automated Reconciliation

Leverage GPU nodes on supercomputing infrastructure to train a large neural network on massive datasets.

Computing Infrastructure
- High-Performance Computing (HPC)
- Swiss National Supercomputing Centre (CSCS)
- Supercomputers
- Clusters
- Job Scheduling

Use HPC clusters designed for large-scale data processing to process terabytes of experimental data.

Setting up Jupyter notebooks with the appropriate kernel and extensions for interactive data analysis.

Computing Environment
- Integrated Development Environment (IDE)
- Package Managers
- Dependencies
- Reproducibility

Create an isolated Pyton environment for your machine learning poject with specific library versions.

Storage Solutions
- Data persistence
- File Systems
- Object Storage
- Backup
- Archiving
- Data Lifecycle

Set up automated backup strategies to safely store intermediate results from long-running simulations.

Archive completed project data for long-term preservation. Implement proper archiving solutions with metadata.

Algorithms and Computation
- **Data Structures**
- **Optimisation**
- **Graph Algorithms**
- **Computational Efficiency**
- **Parallel Computing**
- **Distributed Systems**
- **Simulations**

Choose the right algorithm: For example use stochastic gradient descent instead of standard gradient descent to efficiently train large-scale machine learning models.

Parallel processing to reduce runtime and improve efficiency.

Test and verify to ensure reliability and reproducibility in research experiments.

Code Optimisation
- **Profiling**
- **Performance Optimisation**
- **Parallel Processing**
- **Vectorisation**
- **Memory Management**

Profile your code to identify bottlenecks and optimise the slow parts, reducing simulation times from hors to minutes.

Software Engineering

Research Software Engineering (RSE)
- **Intersection of traditional software engineering and scientific research**

Bridging the gap between traditional software development and scientific research needs.

Development
- **Debugging**
- **Testing**
- **Scaling**
- **Version Control (Git)**
- **Dependency Management**
- **Programming Languages**
- **Python**
- **Open Source**

Choose the best Python framework for web development or for data analysis.

Deployment
- **Production Environment**
- **Containerisation**
- **Cloud Deployment**
- **Monitoring**
- **Release management**
- **Continuous Integration/Continuous Delivery or Deployment (CI/CD)**
- **DevOps**

Automate builds, tests and releases to sustain the quality and longevity of research software.

Deploy and maintain research software reproducibly across different environments; minimise downtime and risk with staged releases and monitoring.

Software Architecture
- **Design Patterns**
- **Microservices**
- **Scalability**
- **Modularity**
- **System Design**
- **Architectural Patterns**

Design systems to remain reliable and performant under high load, ensuring continuous availability and the ability to scale.

Software Lifecycle
- **Requirements**
- **Design**
- **Implementation**
- **Testing**
- **Deployment**
- **Maintenance**
- **Expectation Management**
- **Documentation**
- **Sustainability**
- **Open Source**
- **Licensing**

Adapt the lifecycle to research, including rapid prototyping, changing requirements and reproducibility, as well as handling of requirements throughout a project.

Empa