Module title	**Handling and Visualising Data**
Code	D1
Degree Programme	Master of Science in Life Sciences
Workload	3 ECTS (90 student working hours) - Asynchronous and synchronous distance learning, decentralized teaching: 32 h - Self-study: 58 h (20 h self-study before module starts)
Module Coordinator	**Name**: Dr. Manuel Gil **Phone**: +41 (0)58 934 57 44 **Email**: manuel.gil@zhaw.ch **Address**: ZHAW Life Sciences und Facility Management, Schloss 1, 8820 Wädenswil
Lecturers	• Dr. Manuel Gil, ZHAW • Dr. Simone Ulzega, ZHAW
Entry requirements	1. **Basic statistics** experience at the bachelor level is necessary, including: descriptive statistics, basics of probability theory, probability distributions, basic hypothesis testing, and correlation measures. Prior to the course, students will be provided with a detailed list of topics and corresponding references to learning materials. 2. Students require some **experience with the software R**. Prior to the course (one month in advance) preparatory e-learning material will be provided as part of the self-study. Students are expected to work through the material before the course starts and will be evaluated with an entry test. 3. The following open source **software has to be installed** on the students' notebooks: • RStudio • Apache Open Office Base Details (download and installation instructions) will be provided on Moodle prior to the course.
Learning outcomes and competences	After completing the module, students will be able to: • apply programming structures in R (variables, if-statement, loops, functions) • organise data, control data quality, • work with relational databases with graphical user interfaces (GUI), • understand the application of semantic web concepts (triple stores, ontologies) for biological data integration, • reformat, prepare and process data for further analysis, • import data (into statistics software), • handle missing data (imputation), • describe data, check skewness, outliers or unequal variance and quantify these phenomena, • use robust measures of location and scatter to protect from outliers, • understand the grammar of graphics (and apply it with ggplot2), • produce quick exploratory plots as well as publication quality plots of the data, • use different types of plots, adapted to the data type (independent or correlated data such as time series or spatial data, univariate and multivariate data),

	• weigh advantages and disadvantages of different plot types (e.g. what is hidden/glossed over in a particular plot, what is the minimal/maximal sensible sample size for a particular plot, what plot is suited to illustrate which type of relation, etc.), • produce "meaningful" plots, suited to visualize the answer to the research question (e.g. integrating regression lines into scatter plot) or to display the extracted information, • apply principles of good graph design.
Module contents	<u>Introduction to R *(self-study with e-learning)*</u> • Basic R (import/export of data, command line, basic plotting, basic commands) • Programming structures (variables, if-statement, loops, functions) <u>Introduction to the topic "Handling and visualising data" *(lecture)*</u> <u>Organising data *(lectures and exercises)*</u> • Flat files and redundant data • Relational databases (concepts and querying with a GUI) • Semantic Web technology (Triple, RDF, Ontologies) • Tidy data in R • Classifying and treating missing data <u>Exploring and describing Data *(lectures and exercises)*</u> • Measures of location and scatter • Skewness, outliers, unequal variance <u>Visualising data *(lectures and exercises)*</u> • Grammar of graphics • Plots in R with *ggplot2* • Design characteristics of good plots <u>Project work *(self-study)*</u> • Apply and reinforce the material
Teaching / learning methods	The central teaching will consist of lectures, exercises and a group project. During the local coaching the students will continue/complete the work on the exercises and projects from the central teaching. Thus, the local coaching will supplement the central teaching and allow the students to interact personally with a coach to ask questions and obtain closer supervision. Local coaching can be timed flexibly, subject to taking place between the central teaching slots. The self-study will consist of e-learning units (in particular to prepare for the entry requirements), online tutorials, additional reading, and a project work. For the *Introduction to R* e-learning unit, beginners will require 15-25 hours to work carefully through the tutorial. About 10h are reserved for the completion of the *project work,* and 10h for exam preparation.

Assessment of learning outcome	- In-class entry exam during the first week of the module (open book, 25%) - The final assessment of learning outcome contributes 75% and will either be a written exam, or a project work. This will be decided one month before module starts. Both cases are individual and open book.
Format	7-weeks
Timing of the module	For ZHAW and FHNW: Autumn semester, CW 38-44 For BFH and HES-SO: Spring semester, CW 8-14
Venue	online / decentralized teaching at respective school
Bibliography	<u>Pre-course work</u> Peter Kauf, R online course, provided on Moodle <u>Course material</u> Wickham, Hadley, 2014. "Tidy data." *Journal of Statistical Software*59.10: 1-23. Wickham, Hadley, 2010. "A layered grammar of graphics." *Journal of Computational and Graphical Statistics* 19.1: 3-28. Wickham, Hadley, 2016. *ggplot2: elegant graphics for data analysis*. Springer. Tufte, Edward, and P. Graves-Morris, 2014. "The visual display of quantitative information.; 1983."
Language	English
Links to other modules	This module is the basis for module D2 “Design and Analysis of Experiments” and module D3 “Modelling and Exploration of Multivariate Data”.
Comments	Material treated during local teaching is relevant for the exam.
Last Update	03.08.2021