



Certificate of Advanced Studies

Datenanalyse

Daten sind allgegenwärtig in Alltag und Beruf. Aber wie analysieren und bewerten Sie Ihre Datenbestände? Welche Erkenntnisse gewinnen Sie aus ihnen, welche Prognosen können Sie mit ihnen erstellen, wie nutzen Sie Open Data und allgegenwärtige Informationsquellen? Dieses CAS vermittelt Ihnen Wissen, Methoden und Werkzeuge aus Statistik, Informatik und Machine Learning für die Beantwortung dieser Fragen.

Inhaltsverzeichnis

1	Umfeld	3
2	Zielpublikum	3
3	Ausbildungsziele	3
4	Voraussetzungen	3
5	Unterrichtssprache	3
6	Durchführungsort	3
7	Kompetenzprofil	4
8	Kursübersicht	5
9	Kursbeschreibungen	6
	9.1 Tooling und Datenmanagement	6
	9.2 Deskriptive Statistik und mathematische Grundlagen	7
	9.3 Statistisches Testen	7
	9.4 Grafische Datenexploration und Datenvisualisierung	8
	9.5 Regressionsanalyse	9
	9.6 Zeitreihen und Prognosen	9
	9.7 Machine Learning	10
	9.8 Wahltag Kausalanalyse	10
	9.9 Wahltag Shiny	11
	9.10 Wahltag Regression und Zeitreihenanalyse	11
	9.11 Wahltag logistische Regression und Large Language Models (chatGPT)	12
	9.12 Wahltag Datenvisualisierung mit D3.js	12
10	Kompetenznachweis	13
11	Lehrmittel	13
12	Dozierende	14
13	Organisation	14

Stand: 15.10.2024

1 Umfeld

Daten alleine bringen noch keine Erkenntnisse. Entscheidend ist das «Making Sense out of Data»: Wie können Daten beschrieben und analysiert werden, welche Schlussfolgerungen kann man aus ihnen ziehen? Auf dem Markt stehen leicht bedienbare Software-Tools zur Aufbereitung von Daten, zur Analyse und zur Visualisierung zur Verfügung. Das CAS Datenanalyse (DA) vermittelt Ihnen einen praktischen Grundstock an Wissen, wissenschaftlichen Vorgehensweisen und Werkzeugen für die Datenanalyse.

2 Zielpublikum

Das CAS DA richtet sich an Verantwortliche für Datenanalyse-Projekte: Fachpersonen in unterschiedlichsten Branchen und Unternehmen, Informatiker*innen, wissenschaftliche Mitarbeitende, die in Datenanalyse-Projekten arbeiten, Analysen und Studien erarbeiten und selbst Daten auswerten.

3 Ausbildungsziele

- Sie überblicken das Gebiet der Datenanalyse.
- Sie können Daten methodisch aufbereiten, analysieren und visualisieren.
- Sie können die Regressionsanalyse und die Analyse von Zeitreihen als Prognose-Instrumente anwenden.
- Sie lernen das Arbeiten mit der Sprache R (r-project.org), einem der wichtigsten und verbreitetsten Werkzeuge der Datenanalyse.
- Sie bekommen einen Einblick in das maschinelle Lernen (überwachtes und nicht-überwachtes Lernen) und können ausgewählte Methoden anwenden.
- Sie erwerben zu ausgewählten Themen Spezialkenntnisse: Kausalanalyse, Open Data, Operationalisierung von Datenanalyse-Prozessen mit R Shiny.

4 Voraussetzungen

Sie können mit mathematischen Gesetzen umgehen und haben die Bereitschaft zur anwendungsorientierten Arbeit mit statistischer Software.

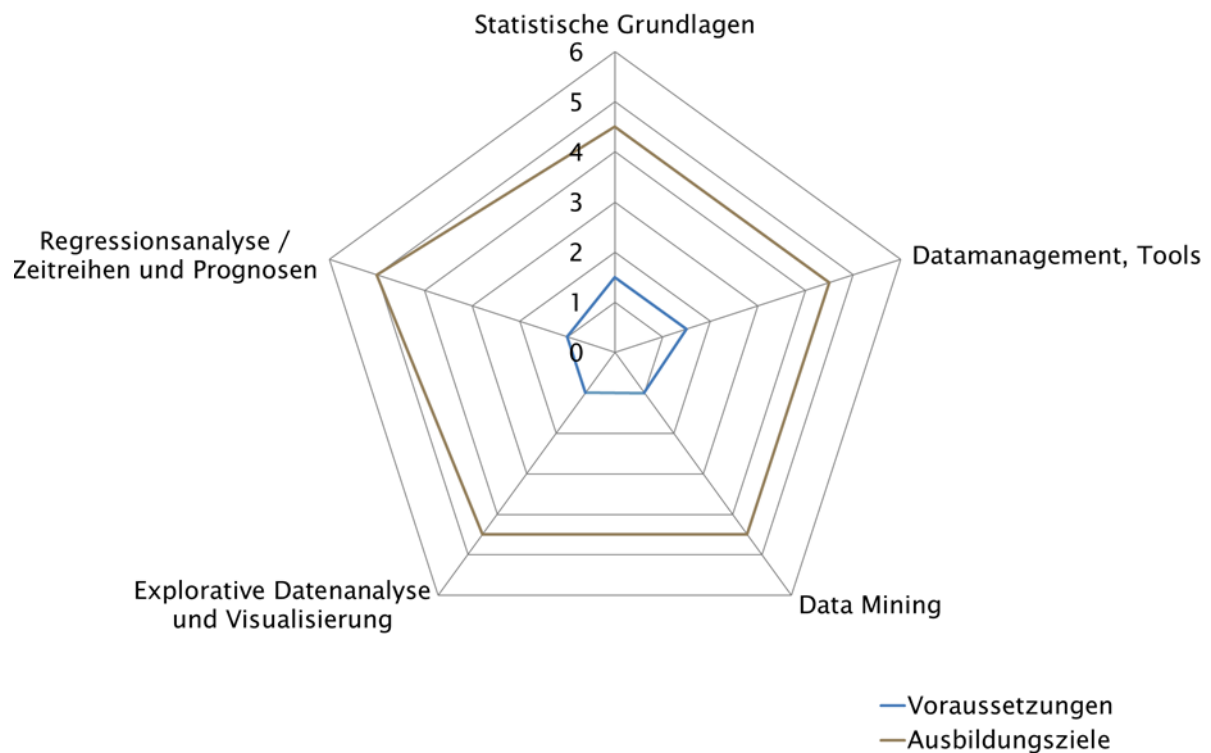
5 Unterrichtssprache

Die Unterrichtssprache ist Deutsch, die Unterlagen sind teilweise in Englisch.

6 Durchführungsort

Berner Fachhochschule, Weiterbildung, Aarbergstrasse 46 (Switzerland Innovation Park Biel/Bienne), 2503 Biel,
Telefon +41 31 848 31 11, E-Mail weiterbildung.ti@bfh.ch.

7 Kompetenzprofil



Kompetenzstufen

1. Kenntnisse/Wissen
2. Verstehen
3. Anwenden
4. Analyse
5. Synthese
6. Beurteilung

8 Kursübersicht

Kurs / Lehreinheit	Lektionen	Dozierende
Tooling und Datenmanagement	24	Rudolf Farys Lukas Schlittler Olivier Lehmann
Deskriptive Statistik und mathematische Grundlagen	12	Michel Krebs
Statistisches Testen	28	Michel Krebs
Grafische Datenexploration und Datenvisualisierung	20	Oliver Hümbelin Christian Schneider Lukas Schlittler
Regressionsanalyse	16	Raul Gimeno
Zeitreihenanalyse und Prognosen	16	Raul Gimeno
Machine Learning	16	Violeta Vogel
Optionale Wahltage – Kausalanalyse – Shiny – Zeitreihen und Regression: Erweiterungen – Datenvisualisierung mit D3.js	je 8	Diverse
Total obligatorische Kurse	132-172	

Das CAS umfasst insgesamt 12 ECTS-Credits. Für die einzelnen Kurse ist entsprechend Zeit für Selbststudium, Prüfungsvorbereitung etc. einzurechnen.

Die Wahltage stehen zur freien Verfügung. Sie werden durchgeführt, wenn mindestens 8 Teilnehmende eingeschrieben sind.

9 Kursbeschreibungen

9.1 Tooling und Datenmanagement

Lernziele	<p>Einführung in das Statistiksoftwarepaket R, welches sich zunehmend zu einer Standardsprache der Datenanalyse entwickelt. Folgende Themen sollen behandelt werden: Grundlegende Funktionsweise von R, Datenmanagement, einfache Auswertungen sowie die Einbindung von R in den persönlichen Workflow (Umgang mit unterschiedlichen Datenquellen/-formaten und Einbindung von Resultaten in die Textverarbeitung (Word/Latex/HTML).</p> <p>Die Teilnehmenden werden befähigt, R für eigene Anwendungen einzusetzen, und kennen die wichtigsten «Anlaufstellen» (Literatur und Onlinehilfen) um das bestehende Wissen weiterzuentwickeln und auf neue Anwendungen auszuweiten.</p>
Themen und Inhalte	<ul style="list-style-type: none">– Benutzung von R-Studio– Klassen und Datentypen– Einlesen und Aufbereitung von Daten: Datensätze laden, verbinden, umformen, aggregieren und exportieren.– Funktionen der Datenmanipulation– Reguläre Ausdrücke– Webscraping– Export von Resultaten
Lehrmittel	<ul style="list-style-type: none">– Skript/Readings auf eLearning Plattform– https://www.datacamp.com/courses/free-introduction-to-r– Onlinehilfen wie Stack Overflow, u.v.a.m.

9.2 Deskriptive Statistik und mathematische Grundlagen

Lernziele	Die Teilnehmenden <ul style="list-style-type: none">– erlernen die grundlegenden Konzepte der deskriptiven Statistik. Nach Absolvierung des Moduls sind sie in der Lage, Daten aufzubereiten und zu präsentieren.– kennen Matrizen und sind in der Lage, elementare Matrizenoperationen korrekt durchzuführen.
Themen und Inhalte	<ul style="list-style-type: none">– Statistische Kennzahlen– Verteilungen– Lage- und Streuungsmasse– Quantile– Bivariate Datenanalyse– Matrizen und Matrizenoperationen
Lehrmittel	<ul style="list-style-type: none">– Folien/Skript/Readings auf E-Learning-Plattform– Literaturempfehlung Nr.1

9.3 Statistisches Testen

Lernziele	Die Teilnehmenden <ul style="list-style-type: none">– erlernen die Grundlagen der Wahrscheinlichkeitsrechnung und der schliessenden Statistik.– kennen insbesondere die statistischen Konzepte der Schätzung, des Hypothesentests sowie des Vertrauensintervalls und können diese in der Praxis anwenden.
Themen und Inhalte	<ul style="list-style-type: none">– Wahrscheinlichkeitsrechnung– Zufallsvariable– Summen von Zufallsvariablen– Vertrauensintervalle und Hypothesentests– Lineare Einfachregression– Schätzen– Bestimmtheitsmass– Prognose
Lehrmittel	<ul style="list-style-type: none">– Folien/Skript/Readings auf E-Learning-Plattform– Literaturempfehlung Nr. 1

9.4 Grafische Datenexploration und Datenvisualisierung

Lernziele	<p>Die Teilnehmenden</p> <ul style="list-style-type: none">– können Nutzen und Grenzen von explorativer Datenanalyse und Datenvisualisierungen im Prozess der Datenanalyse einschätzen.– sind mit den zentralen Techniken der Datenexploration vertraut und können diese mit R umsetzen.– sind fähig, basierend auf den Gestaltungs-Prinzipien der Datenvisualisierung, anschauliche Grafiken zu erstellen.– lernen die Möglichkeiten von interaktiven Datenvisualisierungen kennen und können eigene, einfache Applikationen programmieren.
Themen und Inhalte	<ul style="list-style-type: none">– Bedeutung und Funktion von explorativer Datenanalyse und Datenvisualisierungen im «Epicycles of Analysis»– Techniken der Datenexploration mit R und ggplot2<ul style="list-style-type: none">– «Grammar of Graphics»– Klassische univariate Techniken: Balken- und Kuchendiagramm, Histogramm, Boxplots– Bi- und multivariate Techniken: Streudiagramme, Techniken zum Vergleich von Verteilungen und zur Visualisierung von Entwicklungen über die Zeit, Mosaikplot, Correlogram und generalisierte Scatterplot-Matrizen, small multiples mit Trellis-Plots, Radar-Charts– Erkennen räumlicher Muster, Choroplethenkarte und Punkteverteilungskarten– Visualisierungen als Mittel der Kommunikation<ul style="list-style-type: none">– Gestalt-Prinzipien der Datenvisualisierung– Interaktive Grafiken als Webapplikationen <p>Einführung in R Shiny</p>
Lehrmittel	<ul style="list-style-type: none">– Folien/R-Skripte und über E-Learning bereitgestellte Texte– Literaturempfehlungen Nr. 3,4

9.5 Regressionsanalyse

Lernziele	Die Teilnehmenden <ul style="list-style-type: none">– lernen die Regressionsanalyse als ein vielseitiges und klassisches Instrument kennen, mit dem Beziehungen zwischen abhängigen und unabhängigen Variablen hergestellt und Prognosen erstellt werden können.– verstehen die Grundannahmen der linearen Regressionsmodelle, um valide und verlässliche Ergebnisse zu gewährleisten.– können Methoden und Kriterien zur Überprüfung eines Modells anwenden, mögliche Einschränkungen und Modellfehler identifizieren und die Qualität der Prognosen einschätzen.
Themen und Inhalte	<ul style="list-style-type: none">– Lineare Regression– Multiple lineare Regression– Test auf lineare Restriktionen– T-Test und F-Test für die Regressionskoeffizienten– Prognosen– Informationskriterien
Lehrmittel	<ul style="list-style-type: none">– Folien/Buch/Readings auf E-Learning-Plattform– R-Skripte

9.6 Zeitreihen und Prognosen

Lernziele	Die Teilnehmenden <ul style="list-style-type: none">– kennen die Eigenschaften und Charakteristika von Moving-Average und autoregressiven Prozessen.– können eine Zeitreihe anhand verschiedener Methoden glätten.– können zwischen trend-stationären und differenz-stationären Prozessen unterscheiden.
Themen und Inhalte	<ul style="list-style-type: none">– Glättungsverfahren:<ul style="list-style-type: none">– gleitende Durchschnitte– exponentielle Glättung– Holt-Verfahren– Hot-Winters-Verfahren– Saisonbereinigung: additives und multiplikatives Modell– Regressionsverfahren
Lehrmittel	<ul style="list-style-type: none">– Folien/Readings auf E-Learning-Plattform– R-Skripte

9.7 Machine Learning

Lernziele	<p>Die Teilnehmenden</p> <ul style="list-style-type: none"> – kennen die Anforderungen von Machine-Learning-Methoden an die Daten – können praxisnahe Daten selbständig in der für Machine Learning geforderten Form bereitstellen – kennen die Wirkungsweise ausgewählter Machine-Learning-Methoden – können alternative (auch selbst recherchierte) Methoden selbstständig in den Machine Learning Workflow einbinden, optimieren und testen
Themen und Inhalte	<ul style="list-style-type: none"> – Explorative Datenanalyse (Wiederholung aus Deskriptive Statistik und Datenvisualisierung) – Datenaufbereitung (Feature Engineering) als Konsequenz aus den Erkenntnissen der Explorativen Datenanalyse und mit Sicht auf Machine Learning – Unüberwachtes Lernen, Methodenübersicht <ul style="list-style-type: none"> – Clustering – Dekomposition – Überwachtes Lernen, Methodenübersicht <ul style="list-style-type: none"> – Klassifikation – Regression – Validierung <ul style="list-style-type: none"> – Validierungstechniken <p>Performancemetriken</p>
Lehrmittel	<ul style="list-style-type: none"> – Kuhn und Johnson, Applied Predictive Modeling, Springer 2013 – weitere werden ggf. bei Kursbeginn bekannt gegeben

9.8 Wahltag Kausalanalyse

Allgemein	<p>Datenanalyt*innen sind in der Regel weniger an reinen Korrelationen interessiert, sondern an Fragen zu Ursache und Wirkung. Kausale Zusammenhänge sind jedoch mit Ausnahme experimentell erhobener Daten nur schwer nachzuweisen. Experimente sind jedoch in vielen Fällen nicht durchführbar, z.B. weil sie zu teuer oder unethisch sind. Auch lassen sich experimentelle Ergebnisse oft nicht auf die reale Welt eins zu eins übertragen. Basierend auf dem Potential Outcomes Framework wurden in den letzten zwei Jahrzehnten verschiedene Methoden zur Herleitung von Ursache/Wirkungs-Beziehungen für nicht experimentelle Daten entwickelt bzw. wiederentdeckt und erfreuen sich zunehmender Beliebtheit in der Datenanalyse.</p>
Lernziele	<p>Nach Abschluss des Moduls sind die Teilnehmenden sensibilisiert für typische Probleme bei der Datenanalyse und haben ein Grundverständnis, wann und wie kausale Methoden zum Einsatz kommen können. Praktische Erfahrungen mit diesen Methoden werden anhand von Hands-on-Sessions mit R geübt.</p>
Themen und Inhalte	<ul style="list-style-type: none"> – Potenzial Outcomes Framework – gerichtete azyklische Graphen – Difference-in-Differenz-Schätzer – Regression Discontinuity Design <p>Überblick zu weiteren Verfahren wie z.B. Matching und Instrumentalvariablen</p>

9.9 Wahltag Shiny

Themen und Inhalte	Die Kommunikation der Resultate ist ein wichtiger Teil des datenwissenschaftlichen Prozesses. Dashboards sind eine beliebte Möglichkeit, Daten in einer zusammenhängenden, visuellen Darstellung zu präsentieren. In diesem Kurs lernen Sie, wie Sie Ihre Ergebnisse mit Hilfe von R-Paketen zu einem ausgefeilten Dashboard zusammenstellen können. Das Spektrum der Möglichkeiten reicht vom Hinzufügen einiger Zeilen in R-Markdown zu Ihrem bestehenden Code bis zu reichhaltigen, interaktiven Shiny-Projekten (→ shiny.rstudio.com).
--------------------	---

9.10 Wahltag Regression und Zeitreihenanalyse

Themen und Inhalte	Halbtag für Regressionsanalyse Viele reale Daten enthalten kategoriale Variablen (Geschlecht, Region, Bildungsniveau), die nicht durch metrische Variablen (Grösse, Gewicht) dargestellt werden können. Dummy-Variablen ermöglichen es, diese kategorialen Daten in die Regressionsmodelle aufzunehmen, sodass die Analyse umfassender und realistischer wird. Bei Dummy-Variablen handelt es sich um binäre Variablen, die nur die Werte 0 und 1 annehmen können. Dichotome Variablen lassen sich durch eine einfache Transformation leicht in eine Dummy-Variable überführen: Liegt eine festgelegte Ausprägung vor, nimmt die Variable den Wert 1 an, liegt sie dagegen nicht vor, so nimmt die Variable den Wert 0 an. Dummy-Variablen sind auch nützlich, um Interaktionen zwischen kategorialen und metrischen Variablen zu modellieren. Dadurch lassen sich komplexere Zusammenhänge darstellen, was das Verständnis für multivariate Beziehungen verbessert.
Themen und Inhalte	Halbtag für Zeitreihenanalyse Häufig ist es von Interesse, nicht nur die Vergangenheit mit der Schätzung eines Modells beschreiben und erklären zu können, sondern auch Prognosen für die Zukunft abgeben zu können. Verschiedene Kennzahlen zur Beurteilung der Vorhersagen werden eingeführt, welche die Prognosequalität über die gesamte Stichprobe analysieren. Das Konzept der Scheinregression wird illustriert, bei der ein statistisch signifikanter Zusammenhang zwischen zwei Variablen festgestellt wird, der sachlogisch aber nicht zu begründen ist. Wichtige stochastische Prozesse werden erläutert, wie z.B. Random Walk und autoregressiver Prozess, welche für das Verständnis von ARIMA-Modellen notwendig sind.

9.1.1 Wahltag logistische Regression und Large Language Models (chatGPT)

Themen und Inhalte	Halbtag für logistische Regression Die logistische Regression wird in vielen Bereichen angewendet, wie z.B. in der Medizin zur Vorhersage von Krankheiten (krank/nicht krank), in den Sozialwissenschaften zur Analyse von Wahlergebnissen (wählen/nicht wählen), und im Marketing zur Vorhersage von Kundenverhalten (Kauf/kein Kauf). Im Gegensatz zur linearen Regression, bei der die abhängige Variable kontinuierlich ist, wird bei der logistischen Regression die abhängige Variable typischerweise als zweiwertig angenommen, d.h., sie hat nur zwei Kategorien (Ja/Nein, 0/1, Erfolg/Misserfolg). Die logistische Regression modelliert die Wahrscheinlichkeit, dass ein bestimmtes Ereignis eintritt (z.B. Wahrscheinlichkeit von «Ja» statt «Nein»). Sie liefert als Ergebnis Wahrscheinlichkeitswerte zwischen 0 und 1.
Themen und Inhalte	Halbtag für Large Language Modelle Ein neuronales Netz besteht aus mehreren Schichten von «Neuronen», die jeweils eine ähnliche Transformation wie die logistische Regression vornehmen. Jedes Neuron in einer Schicht berechnet eine gewichtete Summe der Eingabewerte, wendet eine Aktivierungsfunktion an (z. B. die Sigmoid-Funktion) und gibt einen Wert weiter. Studierende bekommen eine Einführung in Large Language Models (LLM-Modelle), welche eine spezielle Klasse von neuronalen Netzmodellen darstellen, die darauf trainiert sind, menschliche Sprache zu verstehen und zu generieren. Diese Modelle zeichnen sich durch ihre grosse Anzahl an Parametern und ihre Fähigkeit aus, komplexe Aufgaben im Bereich der natürlichen Sprachverarbeitung (NLP) zu bewältigen. Konkrete Anwendungen werden mittels ChatGPT betrachtet.

9.1.2 Wahltag Datenvisualisierung mit D3.js

Themen und Inhalte	<p>In diesem Kurs lernen Sie D3 kennen, eine etablierte und weit verbreitete JavaScript Library für professionelle Datenvisualisierungen im Web. D3 ist nicht einfach eine Ansammlung von Charts, sondern ein umfassendes Toolkit für die Programmierung massgeschneiderter und eigenständiger, den Daten angepassten Visualisierungen.</p> <p>Sie programmieren mit Hilfe von D3 eine Visualisierung von A-Z. Dabei lernen Sie die einzelnen Schritte von der Datenaufbereitung über das Datenmapping bis zum Erstellen von grafischen und interaktiven Elementen kennen. Durch die eigene Anwendungserfahrung und die zahlreichen Beispiele bekommen Sie eine Intuition für die Möglichkeiten und Grenzen von D3.</p>
--------------------	---

10 Kompetenznachweis

Für die Anrechnung der 12 ECTS-Credits ist das erfolgreiche Bestehen der Qualifikationsnachweise (Prüfungen, Projektarbeiten) erforderlich, gemäss folgender Aufstellung:

Kompetenznachweis	Gewicht	Art der Qualifikation	Erfolgsquote Studierende
Tooling und Datenmanagement	2	Übungen / Hausaufgabe	0 - 100 %
Deskriptive Statistik und statistisches Testen	2	Übungen + schriftlich 60' / Open Book, Laptop	0 - 100 %
Grafische Datenexploration und Datenvisualisierung	2	Übungen + Hausaufgabe	0 - 100 %
Regressionsanalyse, Zeitreihen und Prognosen	2	Schriftlich 60' / Open Book, Laptop	0 - 100 %
Machine Learning	2	Übungen + schriftlich 45' / Open Book, Laptop	0 - 100 %
Wahltag	0	Nicht Teil des Kompetenznachweises	
Gesamtgewicht / Erfolgsquote	10		0 - 100 %

Studierende können in einem Kompetenznachweis eine Erfolgsquote von 0 bis 100% erreichen. Die gewichtete Summe aus den Erfolgsquoten pro Thema und dem Gewicht des Themas ergibt eine Gesamterfolgsquote zwischen 0 und 100%. Der gewichtete Mittelwert der Erfolgsquoten der einzelnen Kompetenznachweise wird in eine Note zwischen 3 und 6 umgerechnet. Die Note 3 (gemittelte Erfolgsquote weniger als 50%) ist ungenügend, Die Noten 4, 4.5, 5, 5.5 und 6 (gemittelte Erfolgsquote zwischen 50% und 100%) sind genügend.

11 Lehrmittel

Für das Einlesen und als Begleitmaterial werden nachfolgend aufgeführte Bücher empfohlen. Die Beschaffung liegt im Ermessen der Studierenden.

Nr.	Titel	Autoren	Verlag	Jahr	ISBN-Nr.
1.	Statistik ohne Angst vor Formeln	Andreas Quatember	Pearson Studium	2020	978-3-86894-410-5
2.	Applied Predictive Modeling	Max Kuhn, Kjell Johnson	Springer	2013	978-1-4614-6848-6
3.	<u>R Graphics Cookbook, 2nd edition (r-graphics.org)</u>	Winston Chang	O'Reilly UK Ltd.	2024	Online / E-Book
4.	Storytelling with Data: A Data Visualization Guide for Business Professionals	Cole Nussbaumer Knaflic	Wiley	2015	978-1119002253

Weitere Empfehlungen und Hinweise bei den einzelnen Lehrveranstaltungen.

12 Dozierende

Vorname Name	Firma	E-Mail
Michel Krebs	BFH	michel.krebs@bfh.ch
Oliver Hümbelin	BFH	oliver.huembelin@bfh.ch
Rudolf Farys	UniBe	rudolf.farys@soz.unibe.ch
Raul Gimeno	BFH	rauldiego.gimeno@bfh.ch
Violeta Vogel	BFH	violeta.vogel@bfh.ch
Lukas Schlittler	BFH	lukas.schlittler@bfh.ch
Christian Schneider	Data Anatomy GmbH	info@chrischne.io
Olivier Lehmann	BFH	olivier.lehmann@bfh.ch

13 Organisation

CAS-Leitung:

Prof. Dr. Arno Schmidhauser, Departement Technik und Informatik

Tel: +41 31 848 32 75

E-Mail: arno.schmidhauser@bfh.ch

Prof. Dr. Oliver Hümbelin, Departement Soziale Arbeit

Tel: +41 31 848 36 97

E-Mail: oliver.huembelin@bfh.ch

CAS-Administration:

Andrea Moser

Tel: +41 31 848 32 11

E-Mail: andrea.moser@bfh.ch

Während der Durchführung des CAS können sich Anpassungen bezüglich Inhalten, Lernzielen, Dozierenden und Kompetenznachweisen ergeben. Es liegt in der Kompetenz der Dozierenden und der Studienleitung, aufgrund der aktuellen Entwicklungen in einem Fachgebiet, der konkreten Vorkenntnisse und Interessenslage der Teilnehmenden, sowie aus didaktischen und organisatorischen Gründen Anpassungen im Ablauf eines CAS vorzunehmen.

Berner Fachhochschule

Technik und Informatik

Weiterbildung

Aarbergstrasse 46 (Switzerland Innovation Park)

2503 Biel

Telefon +41 31 848 31 11

E-Mail: weiterbildung.ti@bfh.ch

bfh.ch/ti/weiterbildung

bfh.ch/ti/cas-da