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.

bfh.ch/ti/cas-da

Inhaltsverzeichnis

Stand: 22.09.2022

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.

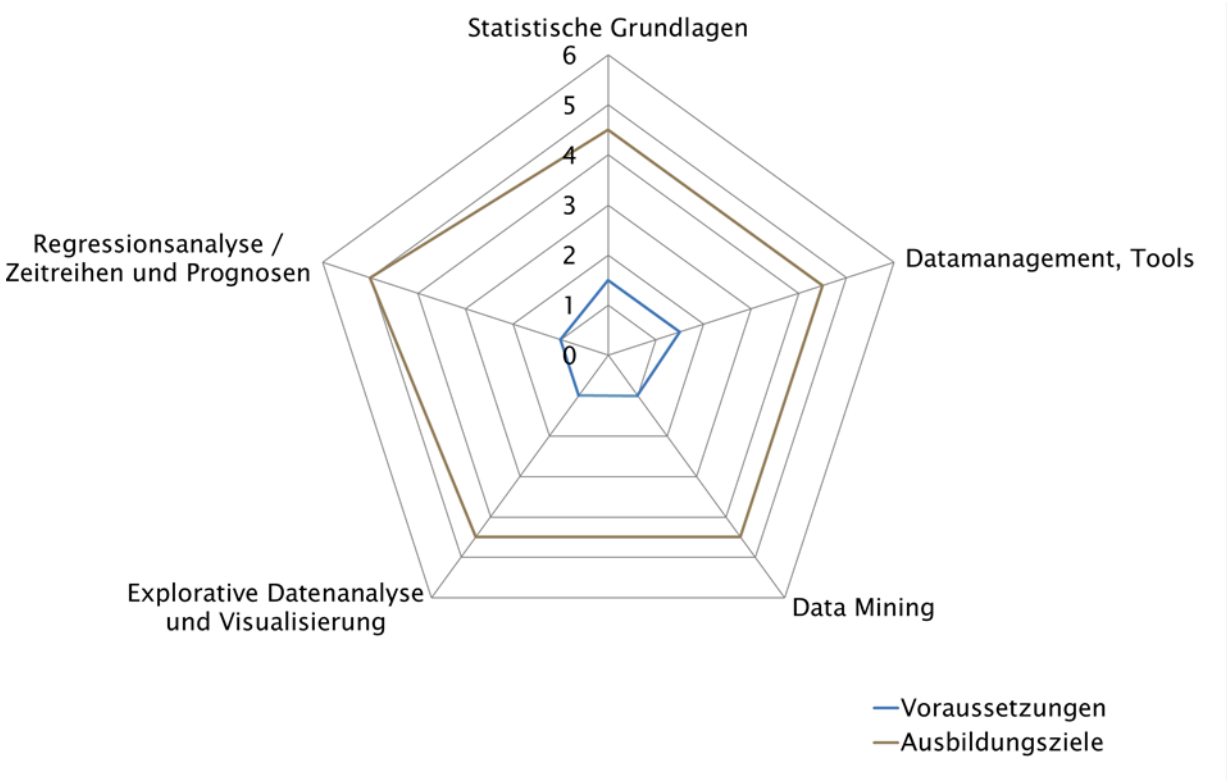
3 Ausbildungsziele

- 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 Kompetenzprofil



Kompetenzstufen

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

6 Kursübersicht

Kurs / Lehreinheit	Lektionen	Dozierende
Tooling und Datenmanagement	24	Rudolf Farys Lukas Hobi
Deskriptive Statistik und mathematische Grundlagen	12	Michel Krebs
Statistisches Testen	28	Michel Krebs
Grafische Datenexploration und Datenvisualisierung	20	Oliver Hümbelin Christian Schneider
Regressionsanalyse	16	Raul Gimeno
Zeitreihenanalyse und Prognosen	16	Raul Gimeno
Machine Learning	16	Werner Dähler
Optionale Wahltage – Kausalanalyse – Open Data – Shiny – Zeitreihen und Regression Erweiterungen – Datenvisualisierungmit 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.

7 Kursbeschreibungen

7.1 Tooling und Datenmanagement

Lernziele	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). 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.
Themen und Inhalte	– 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	– Skript/Readings auf eLearning Plattform – https://www.datacamp.com/courses/free-introduction-to-r – Onlinehilfen wie Stackoverflow, uvvm.

7.2 Deskriptive Statistik und mathematische Grundlagen

Lernziele	Die Teilnehmenden – 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	– Statistische Kennzahlen – Verteilungen – Lage- und Streuungsmasse – Quantile – Bivariate Datenanalyse – Matrizen und Matrizenoperationen
Lehrmittel	– Folien/Skript/Readings auf eLearning Plattform – Literaturempfehlung Nr.1

7.3 Statistisches Testen

Lernziele	Die Teilnehmenden – 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	– Wahrscheinlichkeitsrechnung – Zufallsvariable – Summen von Zufallsvariablen – Vertrauensintervalle und Hypothesentests – Lineare Einfachregression – Schätzen – Bestimmtheitsmass – Prognose
Lehrmittel	– Folien/Skript/Readings auf eLearning Plattform – Literaturempfehlung Nr. 1

7.4 Grafische Datenexploration und Datenvisualisierung

Lernziele	Die Teilnehmenden – 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 Graphiken zu erstellen. – lernen die Möglichkeiten von interaktiven Datenvisualisierungen kennen und können eigene, einfache Applikationen programmieren.

Themen und Inhalte	– Bedeutung und Funktion von explorativer Datenanalyse und Datenvisualisierungen im «Epicycles of Analysis»
	– Techniken der Datenexploration mit R und ggplot2
	  – «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
	  – Gestalt-Prinzipien der Datenvisualisierung
	  – Interaktive Graphiken als Webapplikationen
	  – Einführung in R Shiny

Lehrmittel	– Folien/R-Skripte und über e-learning bereitgestellte Texte – Literaturempfehlungen Nr. 3,4

7.5 Regressionsanalyse

Lernziele	Die Teilnehmenden – lernen die Regressionsanalyse als vielseitiges und klassisches Instrument kennen, mit dem Beziehungen zwischen abhängigen und unabhängigen Grössen hergestellt und Prognosen erstellt werden können. – können Methoden und Kriterien zur Überprüfung eines Modells, möglicher Einschränkungen, möglicher Modellfehler und zur Einschätzung der Prognosequalität anwenden.
Themen und Inhalte	– Lineare Regression – Multiple lineare Regression – Test auf lineare Restriktionen – T-test und F-Test für – Prognosen – Informationskriterien
Lehrmittel	– Folien/Buch/Readings auf eLearning Plattform – R-Skripte

7.6 Zeitreihen und Prognosen

Lernziele	Die Teilnehmenden – 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	– Glättungsverfahren:
	  – Gleitende Durchschnitte
	  – Exponentielle Glättung
	  – Holt-Verfahren
	  – Hot-Winters Verfahren
	– Saisonbereinigung: Additives und multiplikatives Modell
	– Regressionsverfahren

Lehrmittel	– Folien/Readings auf eLearning Plattform – R-Skripte

7.7 Machine Learning

Lernziele	– Die Teilnehmenden 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 selber recherchierte) Methoden selbständig in den Machine Learning Workflow einbinden, optimieren und testen –

Themen und Inhalte	– Explorative Datenanalyse (Wiederholung aus Deskriptive Statistik und Datenvisualisierung)
	– Datenaufbereitung (Feature Engineering) als Konsequenz aus den Erkenntnissen der Explorativen Datenanalyse und mit Sich auf Machine Learning
	– Unüberwachtes Lernen, Methodenübersicht
	  – Clustering
	  – Dekomposition
	– Überwachtes Lernen, Methodenübersicht
	  – Klassifikation
	  – Regression
	– Validierung
	  – Validierungstechniken
	  – Performancemetriken
	–

Lehrmittel	– Kuhn und Johnson, Applied Predictive Modeling, Springer 2019 – weitere werden ggf. bei Kursbeginn bekannt gegeben

7.8 Wahltag Open Data

Lernziele	Die Teilnehmenden setzen sich mit den Möglichkeiten global verfügbarer Daten auseinander.

Themen und Inhalte	– Die Relevanz von Daten in einer zunehmend digitalisierten Welt
	  – für ein globales Wissenssystem
	  – für Ökonomie und Wirtschaft
	– Open Data in der Schweiz
	– Öffentlich zugängliche Daten
	  – Open Data nutzen (Gruppenarbeit)
	
	  – Was sind die wichtigsten Schritte bei der Beschaffung von Open Data zur Dialog und Wiederverwendung?
	  – Welche Schlüsselfragen muss man bei der Veröffentlichung von Open Data wissen (Lizenzen, Formate, Metadaten, Feedback, Community).
	  – Wie funktionieren Datenkataloge, API-Schnittstelle, Datenportale und Open-Government-Projekte?

7.9 Wahltag Kausalanalyse

Allgemein	Datenanalysten 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.
Lernziele	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.
Themen und Inhalt	– Potenzial Outcomes Framework – Gerichtete azyklischen Graphen – Difference-in-Differenz Schätzer – Regression Discontinuity Design – Überblick zu weiteren Verfahren wie z.B. Matching und Instrumentalvariablen

Lehrmittel	– Folien/Skript/Readings auf eLearning Plattform
	– Bücher zur Orientierung und Vertiefung:
	  – Angrist/Pischke, 2008: Mostly Harmless Econometrics
	  – Morgan/Winship, 2011: Counterfactuals and Causal Inference

7.10 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 von Hinzufügen einiger Zeilen in RMarkdown zu Ihrem bestehenden Code bis zu reichhalteigen, interaktiven Shiny-Projekten (→ shiny.rstudio.com).

7.11 Wahltag Regression und Zeitreihenanalyse

Themen und Inhalte	**Wahltag für Regressionsanalyse** Die Dummy-Variablen und die Logarithmierung der Variablen werden in die Regressionsanalyse einführen. Bei der multiplen Regressionsanalyse wurde bis jetzt ein metrisches Skalenniveau für die benutzten Variablen vorausgesetzt. Sollen nun nominalskalierte Variablen in eine solche Analyse einfliessen, können sogenannte Dummy-Variablen gebildet werden. 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. Durch bestimmte Transformationen der Variablen der Regressionsgleichung können viele gekrümmte, nichtlineare Beziehungen dargestellt werden und immer noch das lineare Regressionsmodell verwendet werden.
Themen und Inhalte	**Wahltag 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 des ARIMA-Modells notwendig sind. Das ARIMA-Modell ermöglicht die Analyse und Prognose von Zeitreihen. Es handelt sich um eine leistungsstarke Modellklasse, die den autoregressiven und den moving average Teil des ARMA-Modells um die Differenzierung und Integration zur Trendbeseitigung und Herstellung der Stationarität erweitert.

7.12 Wahltag Datenvisualisierung mit D3.js

Themen und Inhalte	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. 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 Beispielen bekommen Sie ein Intuition für die Möglichkeiten und Grenzen von D3.

8 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'	0 – 100 %
Machine Learning	2	Übungen + Schriftlich 45' / Open Book, Laptop	0 – 100 %
Wahltage	0	Nicht Teil des Kompetenznachweises	
Gesamtgewicht / Erfolgsquote	10		0 – 100 %

Jeder Studierende kann 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.

9 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.	Introduction to Modern Time Series Analysis	Uwe Hassler, Gebhard Kirchgässner, Jürgen Wolters	Springer	2013	978-3-642-44029-8
3.	R Graphics Cookbook, 2nd edition (r-graphics.org)	Winston Chang	O'Reilly UK Ltd.	2018	9781491978597
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.

10 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
Werner Dähler	IBM	werner.daehler@bfh.ch
Lukas Hobi	BFH	lukas.hobi@bfh.ch
Oleg Lavrovsky	datalets.ch	oleg@datalets.ch
Christian Schneider	christianschneider.ch	info@chrischne.io

11 Organisation

CAS-Leitung:
Prof. Dr. Arno Schmidhauser, Departement Technik und Informatik
Tel: +41 31 84 83 275
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 84 83 211
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
Email: weiterbildung.ti@bfh.ch

bfh.ch/ti/weiterbildung
bfh.ch/ti/cas-da