

# Modulinhalte

## „Data Science and Big Data “

Studiengangsleitung: Prof. Dr. Claus Weihs

### ALLGEMEINES

#### Übungsaufgaben

- zu den Modulelementen 1-2 gehören Übungsaufgaben, die die Teilnehmenden eigenständig zwischen den Präsenzterminen bearbeiten
- ein Übungszettel umfasst mindestens 5 Aufgaben
- die Inhalte der Modulelemente werden angewandt
- die Reflexions- und die Transferfähigkeit der Teilnehmenden stehen im Vordergrund
- die Ergebnisse der Übungsaufgaben werden an den Übungstagen intensiv diskutiert

#### Fallstudienarbeit

- Bestandteil von Modulelement 3
- Bericht und Quellcode zur Big Data Fallstudie
- Bearbeitung eines vorgegebenen Big Data Fallszenarios
- Übertragung von Studieninhalten auf ein Praxisbeispiel. (Reflexions- und die Transferfähigkeit stehen im Vordergrund)
- Anwendung von Präsentationstechniken
- Ergebnisse werden an Extratermin intensiv diskutiert

## **Abschlussarbeit**

- Bestandteil vom Modulelement 4
- bis max. 30 Seiten
- eigenständige Analyse und Darstellung
- Übertragung von Studieninhalten auf ein Praxisbeispiel, wenn möglich auf der Basis von „eigenen “ Daten
- Präsentation und Diskussion der Arbeit im Abschlusskolloquium

# MODULINHALTE

Modul- element	Modulbaustein	Inhalte
1 Datenmanagement	<b>Einführung</b>	<ul style="list-style-type: none"><li>• Übersicht über das Studium</li><li>• Organisation des Studiums</li><li>• Kurzeinführung in die Studieninhalte</li><li>• Kennenlernen der Teilnehmenden</li></ul>
	<b>Informationssysteme</b>	<b>Datenbanksysteme</b> <ul style="list-style-type: none"><li>• Begriff „Datenbanksystem“, Nutzen, Abstraktionen</li><li>• Relationales Datenmodell (Schlüssel/Fremdschlüssel, Joins)</li><li>• SQL</li></ul> <b>Data Warehousing:</b> <ul style="list-style-type: none"><li>• Data Warehouse-Architektur, Data Warehouse-Entwicklungsprozess</li><li>• Bus Matrix</li></ul>

- Modellierung mittels Star Schema, Fakten, Dimensionen
- Anfragen an Star Schemata mit SQL und mit SQL OLAP-Erweiterungen

#### **NoSQL-Systeme:**

- MapReduce
- Anfragesprachen für MapReduce-Systeme

#### **Bearbeitung großer Datenmengen mit R**

#### **Kurzeinführung in die Software R**

- Daten- und Kontrollstrukturen
- Ein- und Ausgabe größerer Datenmengen
- Datenbankverbindungen
- Analyse großer Daten mit dem Paket data.table

#### **Effiziente Programmierung**

- Funktionale Programmierung: Map-Reduce
- Explizite Parallelisierung mit Paketen parallel und snow
- Parallelisierung auf HPC-Clustern mit dem Paket BatchJobs

## 2 Datenwissenschaft Theorie

### Datenanalyse - Regression

#### Grundlagen


- Lineare Modelle
- Verallgemeinerte Lineare Modelle
- Schätzverfahren
- Residualanalyse
- Diagnostische Plots
- Variablenselektion
- Interpretation

#### Big Data-Analysen

- Penalisierte Regressionsverfahren (Ridge Regression, LASSO)
- Bayes-Verfahren
- Unterraumeinbettungen
- Sampling

#### Evaluation

- Vorhersage
- Qualitätsmaße (AIC, BIC, DIC, Bayes Faktoren, Fehlerraten)

- 
- 
- Tests
  - Resampling

**Datenanalyse -  
Klassifikation**

**Grundlagen und Verfahren**

- Datenunabhängige Verfahren
- Bayes-Verfahren
- Diskriminanzanalyse
- Logistische Regression
- Entscheidungsbäume
- SVM
- Ensemble Verfahren

**Evaluation**

- Resampling
- Interpretation
- Vorhersage
- Konfusionsmaße
- Tuning
- Variablenselektion
- Dimensionsreduktion
- Modellselektion

## **Big Data**

- Viele Variablen: Probleme klassischer Verfahren
- Viele Beobachtungen: Sampling Verfahren



## Datenvisualisierung

### Grundlagen

- Darstellungsformen und deren Bewertung
- Grundlagen zu Visualisierung in R

### Visualisierung metrisch skaliertter Merkmale

- Darstellung univariater, bivariater und multivariater Daten
- Vergleich mehrerer Stichproben

### Visualisierung kategorieller Merkmale

- Kategorielle Merkmale als interessierende Variable
- Kategorielle Merkmale als bedingende Variable

### Visualisierung räumlicher Strukturen

- Punktdaten
- Gitterdaten

### Visualisierung von Zusammenhängen

- Bäume
- Netzwerke

### **Visualisierung bei großen Datensätzen**

- Große Anzahl von Beobachtungen
- Große Anzahl von Variablen
- Interaktive Möglichkeiten zur Visualisierung

### **Statistische Versuchsplanung**

- **Prinzipien der experimentellen Versuchsplanung und deren Analyse**
- **Verfahren und Modelle für die Planung von Experimenten**
- **Fallzahlplanung**
- **allgemeine Guidelines zur Planung von Experimenten**

### **Fallstudie Big Data Analyse**

#### **Fallstudie Big Data –Analyse:**

#### **Analyse eines Datensatzes mit CRISP-DM**

- Überblick über das CRISP-DM Prozessmodell: Warum ein standardisiertes Vorgehen Fehler vermeidet und Kreativität fördert

#### **Große Datensätze**

- Ab wann ist ein Datensatz wirklich big? Was unterscheidet einen Big Data-Fall von einem gewöhnlichen Datensatz?
- Strategien und Technologien zur Analyse riesiger Datenmengen

### **Fallstudie**

- Einführung in die Fallstudie
- Vorstellung der Daten
- Aufbereitung der Daten
- Analyse-Algorithmus für Big Data Analytics

### **Abschlussarbeit**

### **Selbststudium**

Identifikation eines relevanten großen „eigenen Datensatzes“ für die Abschlussarbeit.

### **Prüfungsformen und -leistungen:**

- Abschlussarbeit, maximal 30 Seiten
- Präsentation und Diskussion der Arbeit.