Knowledge Discovery in Datenbanken II (IN5043)

Titel	Knowledge Discovery in Databases II	
Typ	Vorlesung mit Übung	
Credits	6 ECTS	
Lehrform/SWS	3V + 2Ü	
Sprache	Deutsch	
Modulniveau	Master	
Arbeitsaufwand	Präsenzstunden	75 Stunden
	Eigenstudium	105 Stunden
	Gesamtaufwand	180 Stunden
Angestrebte Lernergebnisse	Die Studierenden kennen Problemstellungen der Analyse von realen Datenbeständen wie Volumen, Volatilität und Komplexität sowie Ansätze im Umgang mit hochdimensionalen, komplexstrukturierten und verlinkten Daten und im Umgang mit volatilen Datenbeständen sowie verschiedene Szenarien der Datenanalyse in verteilten und parallelen Umgebungen. Sie sind in der Lage, Data Mining Algorithmen für komplexe und verlinkte Objekte zu entwickeln und anzuwenden, parallele und verteilte Algorithmen zur Datenanalyse zu implementieren sowie Data Mining Algorithmen in volatilen Systemen zu entwerfen und zu implementieren. Weiterhin sind sie in der Lage, Knowledge Discovery Prozesse in großen, volatilen und/oder komplexen Datenbeständen zu entwerfen und mit Hilfe der gängigen Softwaretools zu implementieren, die Eignung der vorgestellten Verfahren für gegebene Datenbestände und Anwendungsszenarien zu beurteilen und gut geeignete Verfahren auszuwählen.	
Intended Learning Outcomes	Students are familiar with problems and challenges of the analysis of real data repositories such as volume, velocity and complexity, with approaches to handle high dimensional, complex structured and linked data as well as approaches to handle volatile data. They are familiar with various settings and solution strategies in parallel and distributed Environments. The participants are able to develop and to apply data mining algorithms for complex and linked data, to implement parallel and distributed data mining algorithms, to develop and implement data mining algorithms in volatile systems. Further, they are able to design and to develop knowledge discovery processes in large, volatile and/or	

	complex data facilitating the established tools as well as to evaluate the suitability of the introduced methods for given data sets and applications and to select the well-suited methods.
Inhalt	BigData Analytics und Data Science: • Begriffseinführung(Background) • Herausforderungen (z.B. Volume, Velocity, Variety, Veracity) • Verhältnis zu anderen Gebieten. Data Mining in großen Datenmengen: • allgemeine Lösungsansätze (Sampling, Micro-Clustering, Parallelisierung) • Sampling und Micro-Clustering Ansätze(z.B. Cluster Features, BIRCH, Data Bubbles) • Paralleles Data Mining und Verteiltes Data Mining (Grundprinzip, Workflow, Ansätze zum Parallelisieren von KDD Prozessen) • grundlegende verteilte und parallele Data Mining Algorithmen und ihre Umsetzung • Privacy Preserving Data Mining (Gefahrenpotentiale, einfache Angriffe, grundlegende Maßnahmen: Data Swapping, Data Perturbation, Diskretisierung) • komplexe Attacken auf die Privatsphäre und Gegenmaßnahmen [optional] • Data Mining Algorithmen unter Berücksichtigung der Privatsphäre [optional] Data Mining in Volatilen Datenbeständen: • Stream Data Mining (Grundproblematik, Datenalterung, Konzeptdrift, Online Data Mining und Stream Mining) • grundliegende Algorithmen des Stream Clustering • grundliegende Algorithmen zur Stream Classification • weiterführende Algorithmen zur Aggregation von Datenströmen [optional] • Stream Mining Algorithmen für weitere Data Mining Aufgaben (z.B. Frequent Pattern Mining) [optional] Hochdimensionale Daten: • Feature-Selektion (Redundanz und Relevanz von Merkmalen, Suchraum, Problemkomplexität) • Bewertung von Attributen und Unterräumen (supervised Methoden, unsupervised Methoden) • Suchalgorithmen zur Feature Selektion (Forward Selection, Backward Elimination, Branch and

	Bound) • Feature-Reduktion und Lernen von Abstandsmaßen (Begriffserklärung und Zusammenhang) • Lineare Feature-Reduktion (Hauptkomponentenanalyse, Singulär Wert Zerlegung) • Clustering in hochdimensionalen Datenmengen (Ansätze, Top-Down, Bottom up, Locality Assumption) • Clustering Algorithmen für hochdimensionale Daten (z.B. Clique, Subclu, 4C, Proclus, CASH, Co-Clustering) • fortgeschrittene Verfahren(z.B.: Fischer Faces, RCA, LMNN) [optional] • Manifold Lerner [optional] Zusammengesetze Datenobjekte: • Grundbegriffe des Ensemble Learning Möglichkeiten zur Generierung von Diversität, Ergebniskombination • Ensemble-Techniken (z.B. Bagging, Boosting, ECOC) • Multiview Data Mining (Zusammengesetzte Datenräume, Multiview-Distanzen, Multiview-Algorithmen, Kombination von Kernelfunktionen) • Multi-Instanz Data Mining (Begriffsklärung und Abgrenzung) • Multi-Instanz Distanzmaße und Kernel (z.B. Hausdorff Distanz) • Multi-Instanz Data Mining Algorithmen (Multi-Instanz Lernen, konzeptbasiertes Lernen). Link Mining und Graph Mining: • Einführung und Graphmining Tasks (z.B. Link Prediction, Dense Subgraph Discovery, Zentralitätsmaße, SubgraphMining) • Abstandsmaße zwischen Graphen (Graph-Isomorphie, Graphkernel, Distanzmaße) • Abstandsmaße in Netzwerken (z.B. Random Walk with Repeat, kürzester Pfad) • Zentralität in Netzwerken (z.B. PageRank, Betweenness Centrality) • Link-Prediction (z.B. Matrix-Faktorisierung) • Finden häufiger Teilgraphen (Subgraph-Isomorphie, Normalformen, Algorithmen z.B. GSPAN).
Contents	Big Data Analytics and Data Science:

	• Introduction to the topic and background • Challenges (volume, velocity, variety, veracity) • Relationship to other research areas Data Mining in Large Data Repositories: • General approaches(sampling, micro-clustering, parallel computing) • Sampling and micro-clustering techniques(e.g. cluster features, BIRCH, data bubbles) • Parallel and distributed data mining (general principles, workflows, approaches to parallel knowledge discovery) • Basic parallel and distributed data mining algorithms and their implementation • Privacy Preserving Data Mining (risks, simple attacks, basic methods: data swapping, data perturbation, discretization) • Complex attacks to privacy and counter measures [optional] • Privacy preserving data mining algorithms. Data Mining on Volatile Data: • Stream data mining(basic problem setting, aging, concept drift, online and streams data mining) • Algorithms for stream clustering • Algorithms for stream classification • Advanced techniques for data aggregation in data streams [optional] • Stream mining algorithms for further data mining tasks (e.g. frequent pattern mining in streams). High Dimensional Data: • Feature selection (redundance and relevance of features, search space, problem complexity) • Feature and subspace evaluation(supervised and unsupervised criteria) • Search algorithms for feature selection(forward selection, backward elimination, branch and bound) • Feature reduction and metric learning(definitions and connection with related approaches) • Linear feature reduction (principle component analysis, singular values decomposition) • Clustering in high dimensional data spaces (top-down approach, bottom-up approach, locality assumption) • Algorithms for clustering high-dimensional data (e.g. Clique, SubClu, 4C, Proclus, CASH, Co-Clustering)

	• Advanced methods for supervised metric learning(e.g. Fisher faces, RCA, LMNN) [optional] • Manifold learning [optional] Compound Data Objects: • Basic concepts of Ensemble learning (methods for generating diversity, combination functions) • Ensemble techniques (e.g. Bagging, Boosting, ECOC) • Multiview Data Mining (Composed feature spaces, Multiview distance measures, Multiview algorithms, kernel combination) • Multi-Instance Data Mining (Definition and connection to multiview data) • Multi-Instance distance measures(e.g. Hausdorff distance) • Multi-Instance data mining algorithms (multi-instance learning, concept-based learning). Link Mining and Graph Mining: • Introduction to graph mining tasks (e.g. link prediction, dense subgraph discovery, centrality measures, subgraph mining) • Distance measures between graphs (graph isomorphism, graph kernels, topological descriptors) • Distance measures in graphs (e.g. random walk with repeat, shortest path) • Centrality in networks(e.g. pagerank, Betweeness centrality) • Link-Prediction (matrix factorization, classification) • frequent subgraph mining (subgraph isomorphism, normal forms, algorithms e.g. GSPAN)
Prüfung	Prüfungsleistung (benotet): -Klausur: 90 min Wiederholungsklausur zu Ende des Semesters. Details werden zu Beginn des Moduls bekannt gegeben. In der Klausur weisen die Studierenden nach, inwieweit sie die vorgestellten Modelle, Methoden und Algorithmen verstanden haben, komprimiert wiedergeben, anwenden sowie auf verwandte Problemstellungen übertragen können. In der Klausur werden 7 bis 10 Aufgaben gestellt, die eine eigenständige Beschreibung der Prozessdefinition,

	Entwicklung und Anwendung von Data Mining Algorithmen (z.B. subspace clustering), Entwurf paralleler und verteilter Algorithmen (z.B. naive Bayes Modelle in Spark), Anwendung von Knowledge Discovery Prozessen für volatile Systeme (z.B. Micro Clustering, Hoeffding Bäume) und die Bewertung der Verfahren für gegebene Datenbestände und Anwendungsprozesse (z.B. Zentralität in sozialen Netzwerken) erfordern.
Examination	Examination requirements (graded): - written exam: 90 min A makeup exam will be offered at the end of the semester, details will be announced at the beginning of the course. Within the written exam, students demonstrate that they understand the presented models, methods, and algorithms, that they can reproduce and apply them as well as that they can transfer and extend models and methods to similar problems. The written exam consists of 7 to 10 assignments which require an independent description of process definitions, the application and development of data mining methods (e.g. subspace clustering), the development of parallel and distributed data mining methods (e.g. naive Bayes models via Spark), the application of data mining techniques for volatile systems (e.g. micro clustering, Hoeffding trees) and the evaluation of methods for given data sets and application scenarios (e.g. centrality in social networks).
Literatur/Literature	Han J., Kamber M., Pei J.: Data Mining: Concepts and Techniques, 3. Auflage, Morgan Kaufmann, 2011 Tan P.-N., Steinbach M., Kumar V.: Introduction to Data Mining, Addison-Wesley, 2006 Mitchell T. M.: Machine Learning, McGraw-Hill, 1997.
Medienformen	Beamer-Präsentation, Tafelpräsentation, Handout
Media	slides show, blackboard presentation, handouts
Lehr- und Lernmethode	Vorlesung, Übung, Aufgaben zum Selbststudium. Das Modul besteht aus einer Vorlesung und Übungen in kleinen Gruppen (als Tutorübungen). In den Hausaufgaben, die freiwillig abzugeben sind, analysieren die Studierenden die in der Vorlesung vorgestellten Modelle, Methoden und Algorithmen, wenden diese auf konkrete Daten an und erweitern diese für ähnliche Problemstellungen. In den Hausaufgaben werden selbständig anspruchsvolle Übungsaufgaben bearbeitet, die ähnlich zu den

	Klausuraufgaben sind (siehe oben) und daher zur Vorbereitung darauf dienen. In der Übung werden mögliche Lösungsstrategien der Aufgaben zum Selbststudium diskutiert.
Teaching and Learning Methods	Lecture, tutorial, assignments for individual study. Within the assignments (the submission is optional) students analyze the models, methods, and algorithms presented in the corresponding lectures, apply them to real data, and extend these to similar problems. The assignments consist of demanding problems similar to the assignments in the written exam (for details see above) and serve as a preparation for the exam. Within the tutorials possible approaches for solutions of the assignments will be discussed.
Turnus	Wintersemester
Modulverantwortlicher	PD Dr. Matthias Schubert
Dozenten	PD Dr. Matthias Schubert