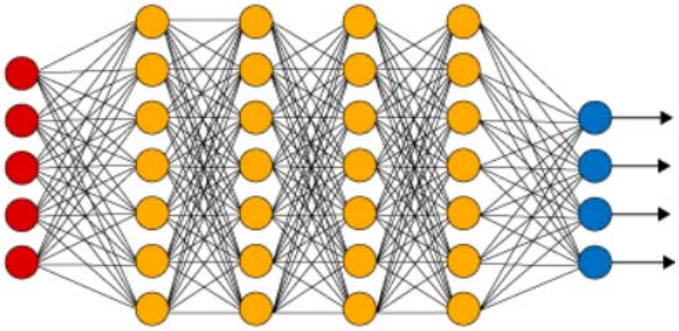
ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

Deep Learning

Prof. Dr. E. Rahm
und Mitarbeiter

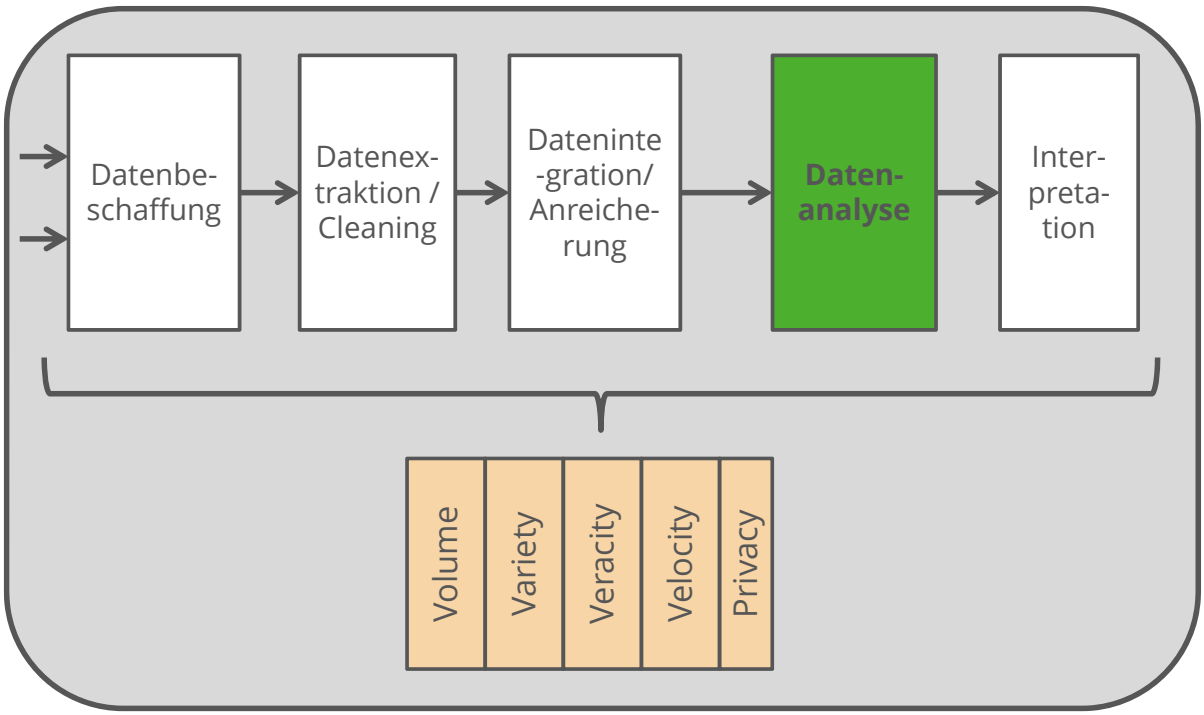
Seminar, WS 2017/18



ScaDS DRESDEN LEIPZIG

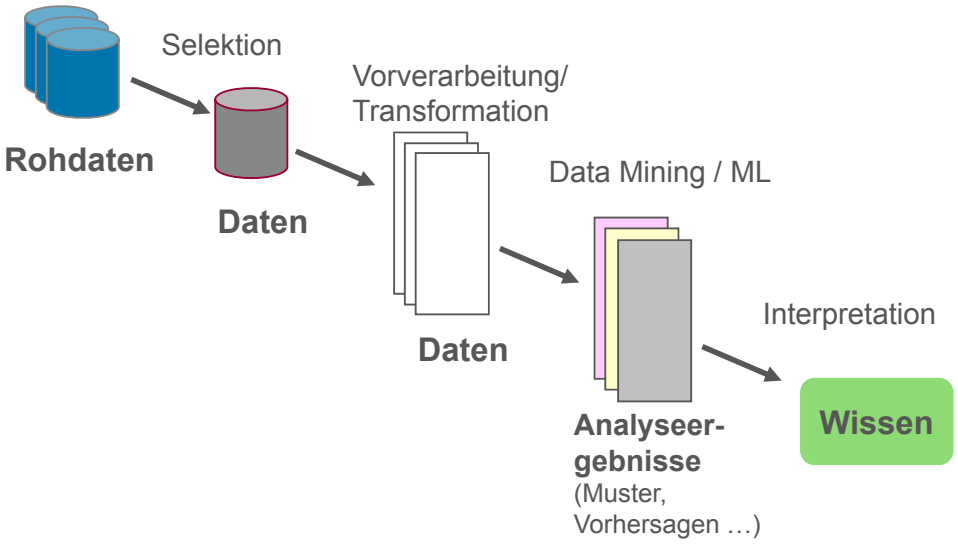
Big Data Analyse-Pipeline

UNIVERSITÄT LEIPZIG



MACHINE LEARNING / KNOWLEDGE DISCOVERY

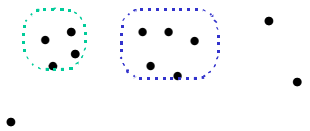
- (semi-)automatische Extraktion von Wissen aus Daten
- Kombination von Verfahren zu Datenbanken, Statistik (Data Mining) und KI (maschinelles Lernen)



TECHNIKEN DES DATA MINING/ML

Clusteranalyse

- Objekte (Kunden, Produkte, ...) werden aufgrund von Ähnlichkeiten in Klassen eingeteilt (Segmentierung)



Assoziationsregeln

- Warenkorbanalyse (z.B. Kunde kauft A und B => Kunde kauft C)
- Nutzung für Kaufvorhersagen / recommendations, Produkt-Bundling, ...

Klassifikation

- Zuordnung von Objekten zu Gruppen/Klassen mit gemeinsamen Eigenschaften bzw. Vorhersage von Attributwerten
- Verwendung von Stichproben (Trainingsdaten)
- Ansätze: Entscheidungsbaum-Verfahren, neuronale Netze, statistische Auswertungen

weitere Ansätze:

- genetische Algorithmen (multivariate Optimierungsprobleme, z.B. Identifikation der besten Bankkunden)
- Regressionsanalyse zur Vorhersage numerischer Attribute . . .

KLASSIFIKATION

Klassifikationsproblem

- gegeben Stichprobe (Trainingsmenge) *O* von Objekten des Formats (a_1, . . ., a_d) mit *Attributen* A_i, 1 ≤ i ≤ d, und Klassenzugehörigkeit c_i, $c_i \in C = \{c_1, . . ., c_k\}$
- gesucht: Klassenzugehörigkeit für Objekte aus D \ O, d.h. *Klassifikator* K : D → C
- weiteres Ziel: Generierung (Lernen) des expliziten Klassifikationswissens (Klassifikationsmodell, z.B. Klassifikationsregeln oder Entscheidungsbaum)

Abgrenzung zum Clustering

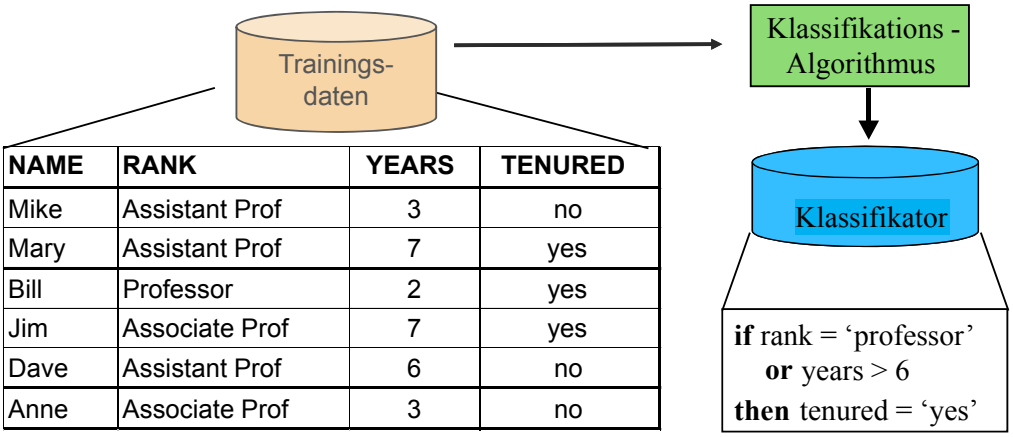
- Klassifikation: Klassen vorab bekannt, Nutzung von Trainingsdaten
- Clustering: Klassen werden erst gesucht, keine Trainingsdaten (unsupervised)

Klassifikationsansätze

- Entscheidungsbaum-Klassifikatoren
- Neuronale Netze
- Bayes-Klassifikatoren (Auswertung bedingter Wahrscheinlichkeiten)
- Support Vector Machine (SVM), lineare Regression ...

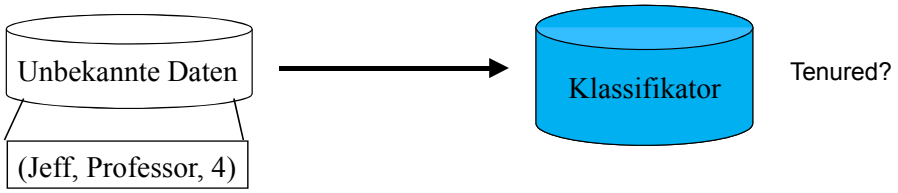
KLASSIFIKATIONSPROZESS

1. Konstruktion des Klassifikationsmodells

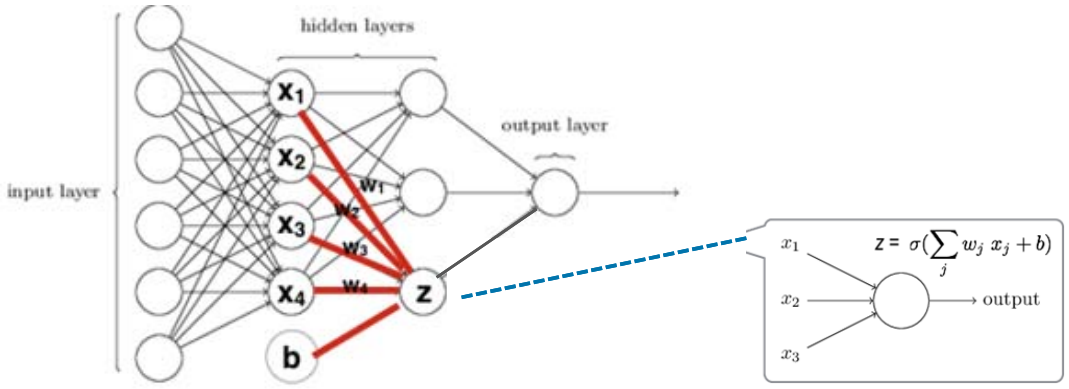


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

2. Anwendung des Modells zur Vorhersage (Prediction)



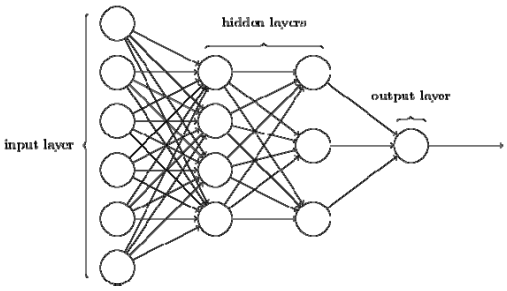
NEURONALE NETZE



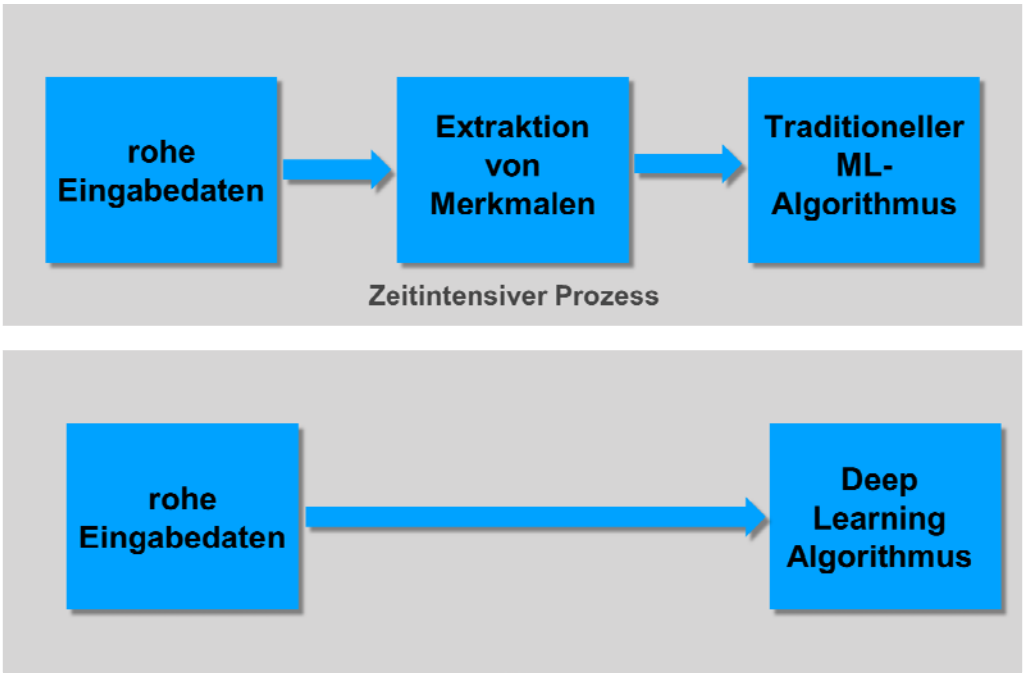
- Neuronales Netz (NN) besteht aus mehreren Schichten
  - Eingabe-/Ausgabeschicht
  - Mind. einer verdeckten (hidden) Schicht
- jede Schicht besteht aus mehreren Neuronen, welche mit anderen Neuronen verbunden sind
- Verbindungen / Kanten verwenden Zahlen, z.B. Gewichte ($w_i \in \mathbb{R}$)
- Deep Learning: mehrere hidden layers

DEEP LEARNING

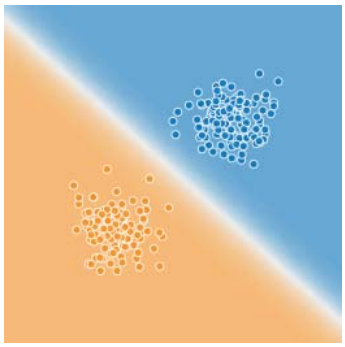
- Nutzung tiefer neuronaler Netze zum Lernen einer Datenrepräsentation auf großen Mengen an Trainingsdaten (Feature Engineering)
- Nutzung des gelernten Wissens für Klassifikation, Vorhersagen …
- zahlreiche Anwendungsfälle
  - Erkennung von Bildern
  - Erkennung von Handschriften
  - Spracherkennung
  - Verarbeitung von Texten …
- verschiedene Varianten von Netzen
  - Convolutional deep neural networks
  - Recurrent neural networks , u.a. LSTM (Long short-term memory)
  - Autoencoder networks (Erzeugung verbesserter Repräsentationen)



AUTOMATISIERTE ERMITTLUNG RELEVANTER FEATURES

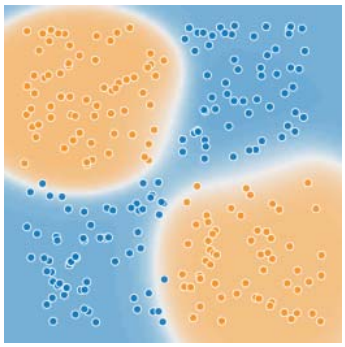


KOMPLEXE MODELLBILDUNG

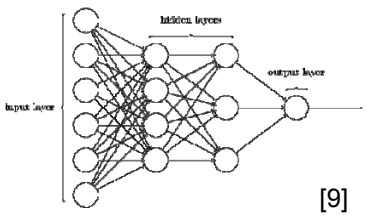


Lineares Entscheidungsmodell

$$h_\Theta(x) = \Theta_0 + \Theta_1 x$$



Kombination von mehreren nicht linearen Funktionen mit neuronalem Netz

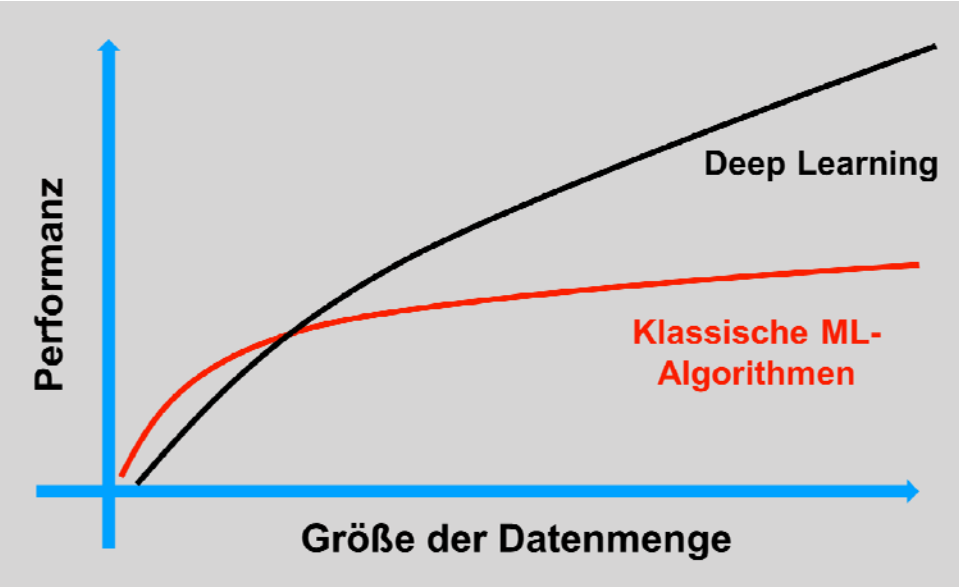


[9]

ScaDS DRESDEN LEIPZIG

DEEP LEARNING UND BIG DATA

UNIVERSITÄT LEIPZIG



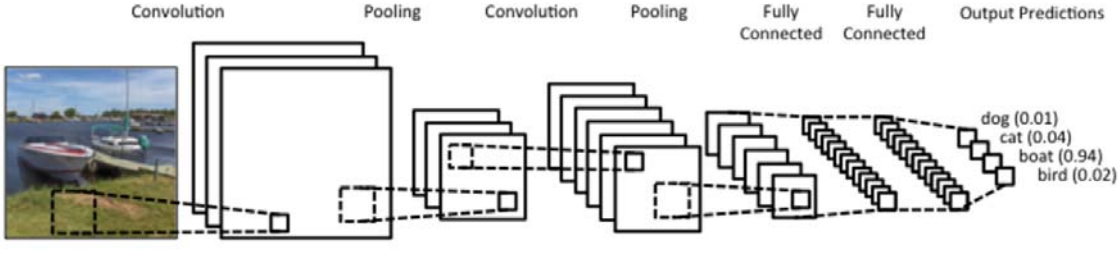
ScaDS DRESDEN LEIPZIG

BILDKLASSIFIZIERUNG

UNIVERSITÄT LEIPZIG

Nutzung z.B. von *Convolutional Neural Networks*

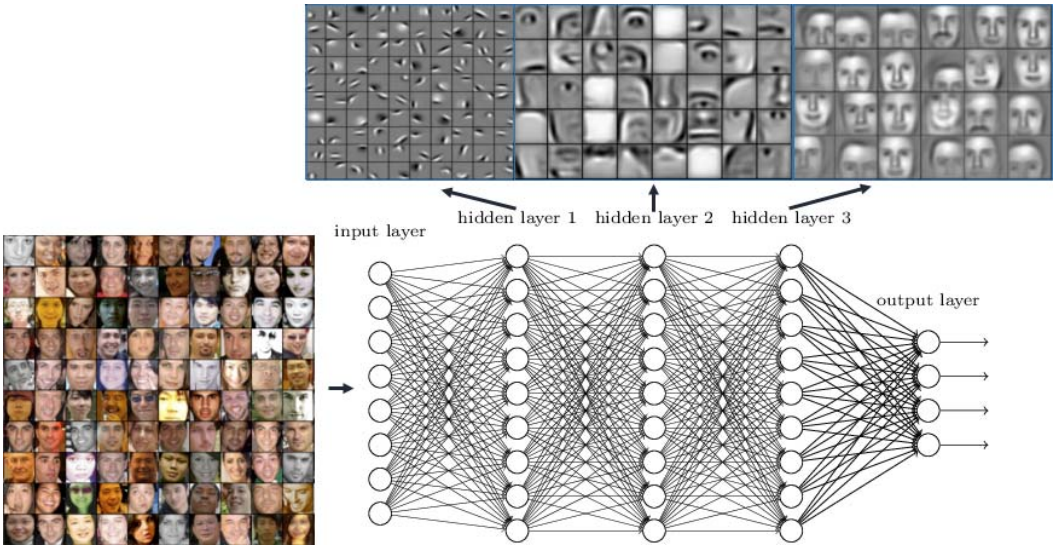
- lokale Filter fassen Pixelaktivität zusammen (convolutional layer)
- nur ausgewählte Informationen daraus werden weitergereicht und somit überflüssige Information verworfen (pool layer)
- dieser Vorgang kann wiederholt Anwendung finden



www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp

GESICHTSERKENNUNG

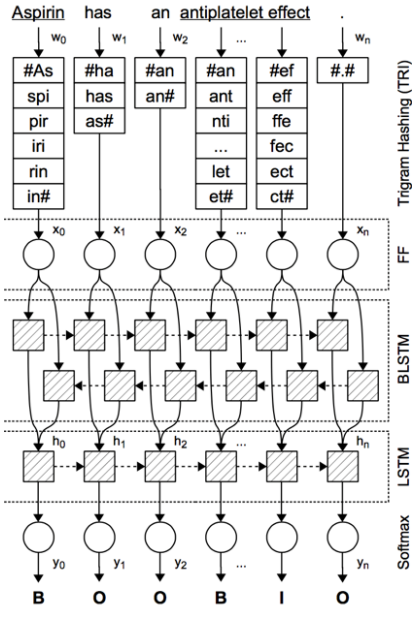
- Neuronale Netze lernen Merkmale der Eingabedaten in Form von aufeinander aufbauenden Konzepten.
- hierarchische Repräsentation der Daten (Farbwerte der Pixel): Kanten -> Teile des Gesichts -> gesamtes Gesicht



https://www.slideshare.net/Tricode/deep-learning-stm-6

TEXT/SPRACHVERARBEITUNG

- Lernen der Nachbarschaft von Wörtern (*word embeddings*) in Text, um deren semantische Ähnlichkeit zu ermitteln
- trainierte Datenrepräsentationen nutzen für weitere ML-Aufgaben, zB
  - Named Entity Recognition
  - Machine Translation
  - Spracherkennung
- häufiger Einsatz von *Recurrent Neural Networks (RNN)*
- vortrainierte Vokabulare: word2vec, glove



RNN Pipeline für
Named Entity Recognition

https://arxiv.org/abs/1608.06757

ScaDS DRESDEN LEIPZIG

UNIVERSITÄT LEIPZIG

SEMINAR

ScaDS DRESDEN LEIPZIG

SEMINARZIELE

UNIVERSITÄT LEIPZIG

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- Erarbeitung + Durchführung eines Vortrags unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- schriftliche Ausarbeitung zum Thema
- Hilfe und Feedback durch zugeteilten Betreuer

SEMINAR: ANRECHNUNGSMÖGLICHKEITEN

- Masterstudium, insbesondere für Schwerpunkt „Big Data“
  - Teil der Module Moderne Datenbanktechnologien
  - Seminarmodul
- Bachelorstudium
  - Seminarmodul

SCHEINVERGABE / MODULPRÜFUNG

- selbständiger Vortrag mit Diskussion (ca. 45 Minuten)
  - Abnahme der Folien durch Betreuer
- schriftliche Ausarbeitung (ca. 15 Seiten)
  - Abnahme der Ausarbeitung durch Betreuer
  - Abgabe-Deadline 31.3.2018
- aktive Teilnahme an allen Vortragsterminen
- Modul-Workload: 30h Präsenzzeit,
  120 h Selbststudium

ScaDS DRESDEN LEIPZIG

SEMINAR (3)

UNIVERSITÄT LEIPZIG

- Themenzuordnung
  - Koordinierungstreffen mit Betreuer innerhalb der nächsten zwei Wochen, d.h. bis spätestens 3.11.2017
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 3.11.2017

- Vortragstermine
  - freitags, Ritterstr, ab 5. 1. 2018 (ggf. ab 12.1.)
  - max. 2 Doppelstunden ab 13:30 Uhr

Themen	Betreuer	max. #Themen	Termin	Studenten
Einführung ML / DL **Arten neuronaler Netze: Autoencoder, CNN, RNN**	Sehili Christen / Lin	1 2	12.1.	Stritz Saalmann Hüsselhuth
Systeme TensorFlow Caffe & Chainer DeepLearning4J	Nentwig Peukert Franke	3	12.1. 19.1	Türkyer (Kim) Stelzer
Image Processing ImageNet Classification Identity-Preserving Face Space	Sehili Rostami	2	19.1. 26.1.	Stut (Yuan) Kuber
NLP – Sprachverarbeitung Word Representations Named Entity Recognition Machine Translation	Alkhouri Alkhouri Lin	3	26.1.	Bordewisch Stoudte Gro
Graph classification	Kricke	1	2.2.	Michael
Privacy and Security Anomaly Detection Privacy-preserving DL Malware classification	Grimmer Franke Franke	3	2.2.	Neumann
Life Science Alzheimer diagnosis Prediction of protein functions	Christen Christen	2	2.2.	Schreiblehner