SEMINAR
KLASSIFIKATION & CLUSTERING
EINFÜHRUNG

Stefan Langer
CIS – Universität München
Wintersemester 2024/25
stefan.langer@cis.uni-muenchen.de

Anmeldung & Kommunikation

- Bereits erfolgt über LSF

- Nachrichten an:
  - stefan.langer@lmu.de
    - An diese Adresse bitte alle seminarbezogene Kommunikation

Scheinvoraussetzungen

- Regelmäßige Teilnahme und Abgabe der Übungen
- Anwendung eines ausgewählten Algorithmus auf verschiedenen Datensätzen – Präsentation und Dokumentation der Ergebnisse
- Abgabe der Präsentation (als PDF)

Links

- Seminar-Homepage:
  - https://www.cis.lmu.de/~stef/seminare/klassifikation_2024/

Passwort Datensammlung: ‘classify!’

Referatsthemen

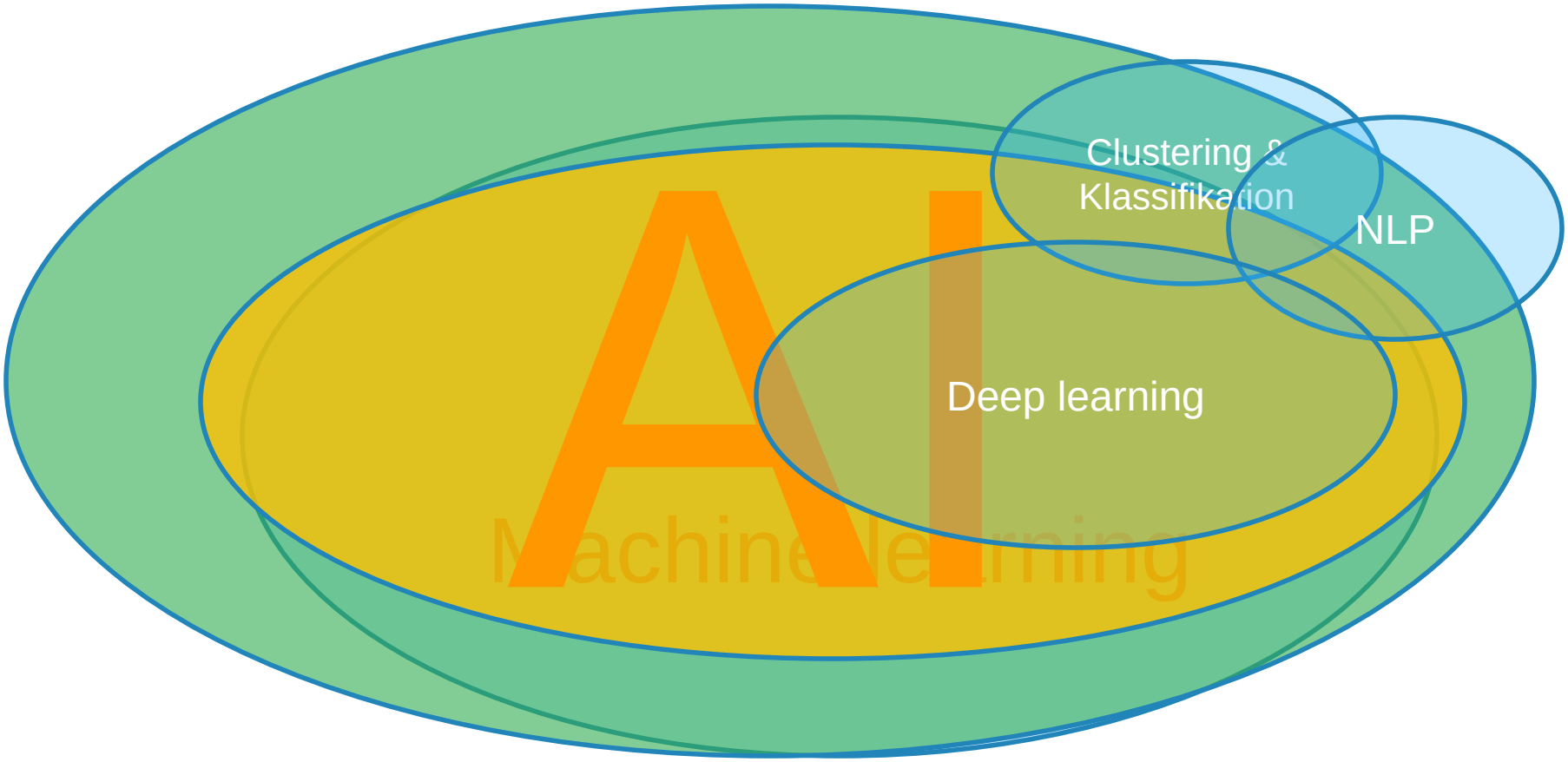
- Vorstellung eines von Ihnen gewählten Klassifikations- oder Clusteringalgorithmus oder Teilaspekt eines Algorithmus
  - K-nearest-neighbour
  - Entscheidungsbäume
  - Support vector machines
  - Latent semantic indexing
  - Neuronale Netze (versch. Varianten)
  - Klassifikation mit Large language Models

- Implementierung und Anwendung auf die gestellten Datensets

Ziele des Kurses

- Grundlagen der Klassifikation und des Clusterings
- Algorithmen und Algorithmentypen
  - Neuronale Architekturen
- Vorverarbeitung von Daten
- Featureauswahl für Textklassifikation, inkl. linguistische Grundlagen
- Vorhandene Tools und Bibliotheken (v.a. open source)
- Anwendungen

Klassifikation / Clustering @ AI



Clustering vs. Klassifikation

Klassifikation	Clustering
Vordefinierte Klassen	Cluster werden algorithmisch erzeugt
Trainingskorpus erforderlich	Kein Trainingskorpus erforderlich
Ähnlichkeitmaß zwischen Objekten	Ähnlichkeitsmaß zwischen Objekten
Hierarchisch oder flach	Hierarchisch oder flach
Einfache Algorithmen verfügbar	Relativ komplexe und aufwändige Algorithmen
Evaluation nach Recall/Precision u.ä. Maßen; Testkorpus erforderlich	Evaluation schwierig; Testkorpus erforderlich; Erstellungskriterien unklar

Machine Learning

- Machine learning is about the increasing a computer‘s ability to solve a task by learning data
- Clustering and classification learning algorithms are instances of machine learning algorithms
- Example 1: A set of document vectors is presented together with the classes the document fall into → the computer learns the classification rules from the data
- Example 2: A set of documents is presented to the computer. The program analyses the data, tries to identify patterns and clusters the data automatically.

Classification algorithms w/o machine learning

- Are there any classification algorithms that do not use machine learning?
- → Hand crafted rules in decision tree-like algorithms (but decision trees can also be learned automatically)

Klassifkationsaufgaben: Einteilung der Tiere

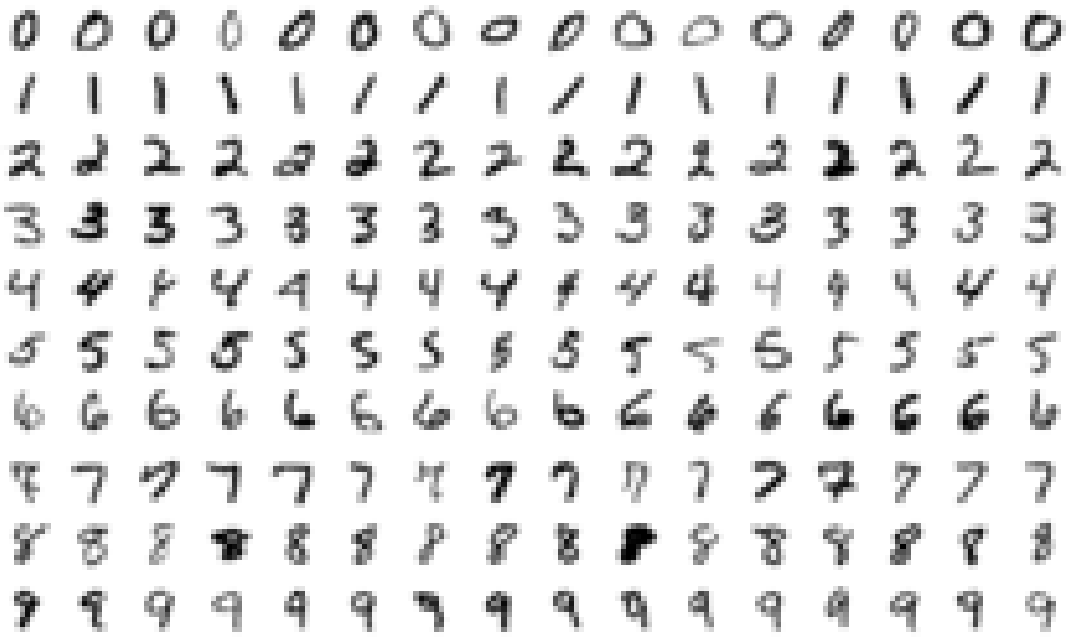
(angeblich chinesische Enzyklopädie, nach Borges (1942): "The Analytical Language of John Wilkins")

- dem Kaiser gehörige,
- einbalsamierte,
- gezähmte,
- Milchschweine,
- Sirenen,
- Fabeltiere,
- streunende Hunde,
- in diese Einteilung aufgenommene,
- die sich wie toll gebärden,
- unzählbare,
- mit feinstem Kamelhaarpinsel gezeichnete,
- und so weiter,
- die den Wasserkrug zerbrochen haben,
- die von weitem wie Fliegen aussehen.

Klassifikation – Daten

- Erfahrungen aus der Praxis zur Voranalyse der Daten:
  - Analyse, Definition und Abgrenzung der Klassen ist Teil der Klassifikationsaufgabe
    - Machen die definierten Klassen Sinn? Gibt es Überschneidungen?
    - Sind die Daten korrekt / sinnvoll gelabelt?
    - Welcher Typ von Klassenstruktur Binär? Multiclass / Multilabel?
    - Reichen die gelabelten Daten aus?

Klassifikationsaufgaben



Klassifikationsaufgaben

ویلیام شکسپیر در ۲۶ آوریل سال ۱۵۶۴ در انگلستان در شهر استراتفورد متولد شد. شهرت شکسپیر به عنوان شاعر، نویسنده، بازیگر و نمایشنامه نویس منحصربه‌فرد است و برخی او را بزرگ ترین نمایشنامه نویس تاریخ می‌دانند، اما بسیاری از حقایق زندگی او مبهم است.

FA

Medan Brand var eit drama Ibsen sleit lenge med, kom Peer Gynt omntrent av seg sjølv. Dramaene står i et refleksjonstilhøve til kvarandre

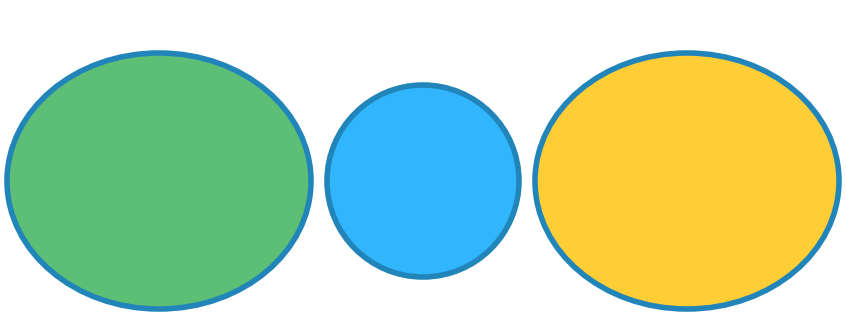
aw än wård schööld än stiped. Sunt 2004 sääkert et friisk-gesäts et rucht aw e friiske spräke önj e öfentlikhäid än stalt fååst, dåt följk ham tu e friiske följkefloose bekåne mätj

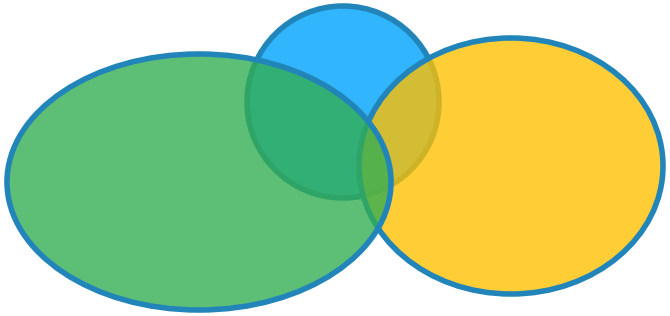
Dekred eus ar 26 a viz Du 1790 war al le ret: e penn-kentañ 1791 e voe graet al le ouzh ar Roue, ar vro hag al lezennoù gant Le Coz, Ollitraut, Serandour ha Le Gac

Drama radio enwog gan Dylan Thomas a gyhoeddwyd yn 1954 yw Under Milk Wood. Mae'r ddrama yn disgrifio digwyddiadau mewn un diwrnod yn unig, yn y pentref dychmygol Llareggub, er cred llawer fod nifer o'r cymeriadau yn seiliedig ar bobl go iawn ag oedd yn byw yn Nhalacharn.

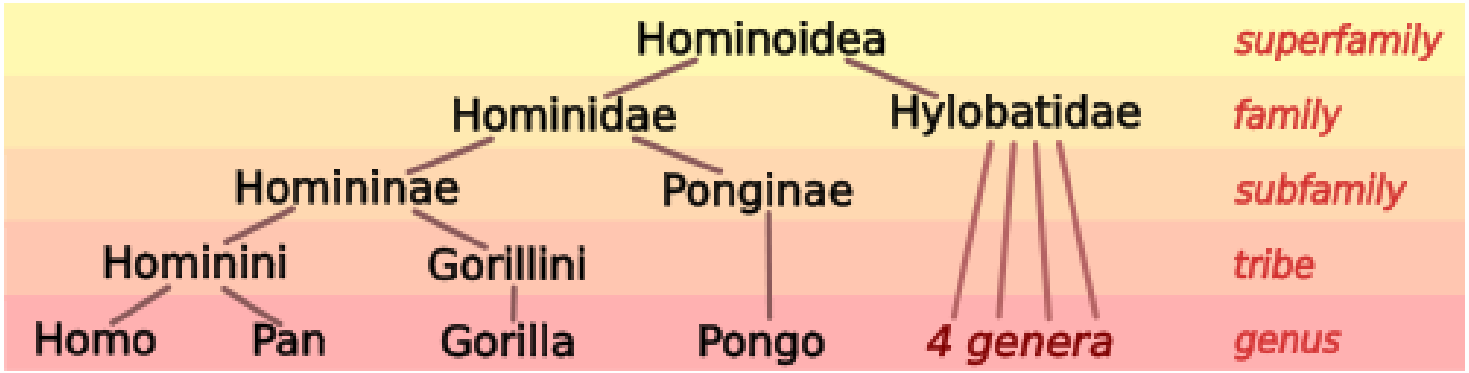
Einstufige / flache Klassifikation

- 2 – N Klassen (binär ←→ Multiclass)
- Sonderfall binäre Klassifikation: nur zwei Klassen (auch interpretierbar als eine Klasse mit Entscheidung über Zugehörigkeit versus Nichtzugehörigkeit (= Zugehörigkeit zur Komplementklasse)
- Klassen können disjunkt sein, oder sich überlappen (Multilabel)

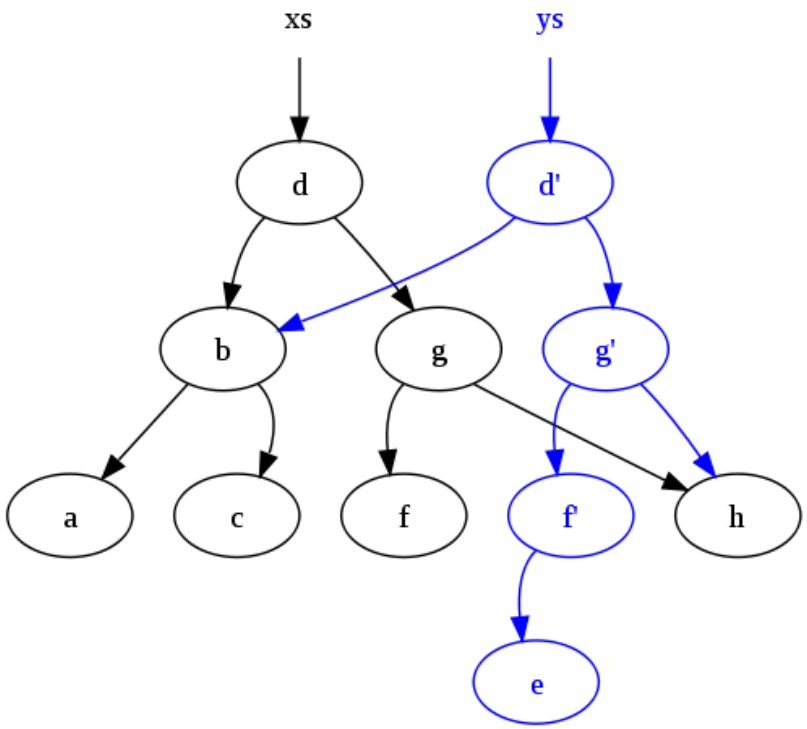


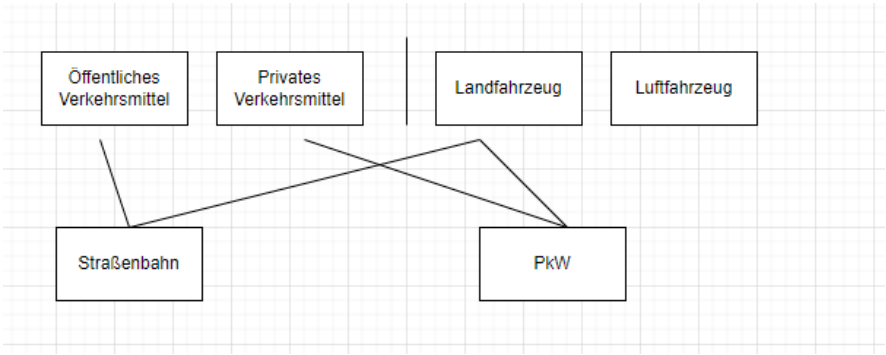


Monohierarchie (Baum)



Polyhierarchie (gerichteter, azyklischer Graph)





Merkmale

- Objekte, die klassifiziert werden sollen werden durch Merkmale ('features') definiert
- Merkmale in der Textklassifikation sind: Buchstaben, Wörter, Phrasen, Textlänge, Satzlänge im Text u.a.; evt. Metamerkmale wie Entstehungsdatum, Autorin …
- Eine große Rolle bei Klassifikation & Clustering spielt die Merkmalsauswahl

Komponenten eines Klassifikationssystems

- Zur **Klassifikation** selbst braucht man:
  - Einen Klassifikator (trainierter Klassifikationsalgorithmus)
  - Daten, die klassifiziert werde sollen
  - Eine Featureauswahl für die Daten
- Zur **Erstellung** eines Klassifikators benötigt man:
  - Einen Trainingsalgorithmus
  - Ein Trainingsset
- Zur **Evaluierung** eines Klassifikators benötigt man:
  - Ein Testset
  - Trainingsset und Testset müssen disjunkt sein
  - Der Trainingsalgorithmus muss auf den Klassifikationsalgorithmus zugeschnitten sein

Klassifikation in Anwendungen

- Spamfilter (binäre Klassifikation)
- Sprachenidentifizierung (Suchmaschinen, Textverarbeitung)
- Sentimentanalyse
- News-Suche, vertikale Suchmaschinen
- Nachrichtenklassifikation
- Autorenerkennung
- Trend-Vorhersage

Klassifikation in der Linguistik und Computerlinguistik

- Syntax: Wortarten – Automatisierung: Tagging
- Semantik: Wortbedeutungen – Automatisierung z.B. automatische Wortbedeutungsdisambiguierung
- Sprachenerkennung
- Textklassifikation (text categorization)
  - Inhaltlich
  - Sentiment-Analyse
  - Autorenerkennung
  - ...

Textklassifikation – Text-/Dokumententypen

- Unstrukturierte Texte (nur-Text-Dokumente)
- Semistrukturierte Texte (z.B. HTML, PDF)
- Strukturierte Texte (Datenbanken; XML)

Clustering – Typen von Algorithmen

- Konnektivitätsbasiert (z.B. Single-Link/Complete-Link im hierarchischen Clustering)
- Centroid-basiert – z.B. K-Means
- Dichtebasiert – z.B. DBSCAN und Optics