Zum State of the Art automatischer Inhaltsanalyse

Michael Scharkow, M.A.

Universität Hohenheim
Institut für Kommunikationswissenschaft (540G)
michael.scharkow@uni-hohenheim.de

Typologie der Verfahren

• deskriptive/explorative Verfahren	• überwachte, gelenkte Verfahren
• keine Hypothesen	• hypothesengeleitet
• keine Kontrolle über das Verfahren	• Kontrolle über die Codierung
• vor allem zur ersten Exploration der Textdaten	• manuelle (Vor-)Arbeit nötig
• vollautomatisch und effizient	• sinnvolle Kategorien

Ungelenkte Verfahren - Mal schauen, was da ist

Textstatistische Verfahren

- Beschreibung des Textes durch Wortfrequenzen, Histogramme

- Wort-, Satz und Textstatistiken sind nützliche Indikatoren für

  - journalistische Praktiken, Genre und Textsorten

  - Autorenschaft/Stilometrie

  - Komplexität der Texte und Lesbarkeit

Beispiele

- Zuordnung von Autoren bei den Federalist Papers
    - spezifisches Vokabular hilft bei der Identifikation
- Messung der Verständlichkeit von (politischen) Texten
    - Wortlänge und Satzlänge als Indikator von Verständlichkeit
    - Umfang des benutzten Wortschatzes als Indikator für Komplexität
- Visualisierung von Textdaten, z.B. für qualitative Interviews oder offene Antworten aus Fragebögen

Co-Occurrence und Latent Semantic Indexing

- Wörter tauchen gern **gemeinsam** in bestimmten Kontexten (Sätzen, Absätzen, Dokumenten) auf

- über alle Dokumente kann daher eine **Ähnlichkeitsmatrix** von Wörtern bzw. Wortgruppen erstellt werden (wie oft tauchen Begriffe zusammen auf)

- **Clusteranalyse** und **Multidimensionale Skalierung** arbeiten mit Ähnlichkeitsmatrizen

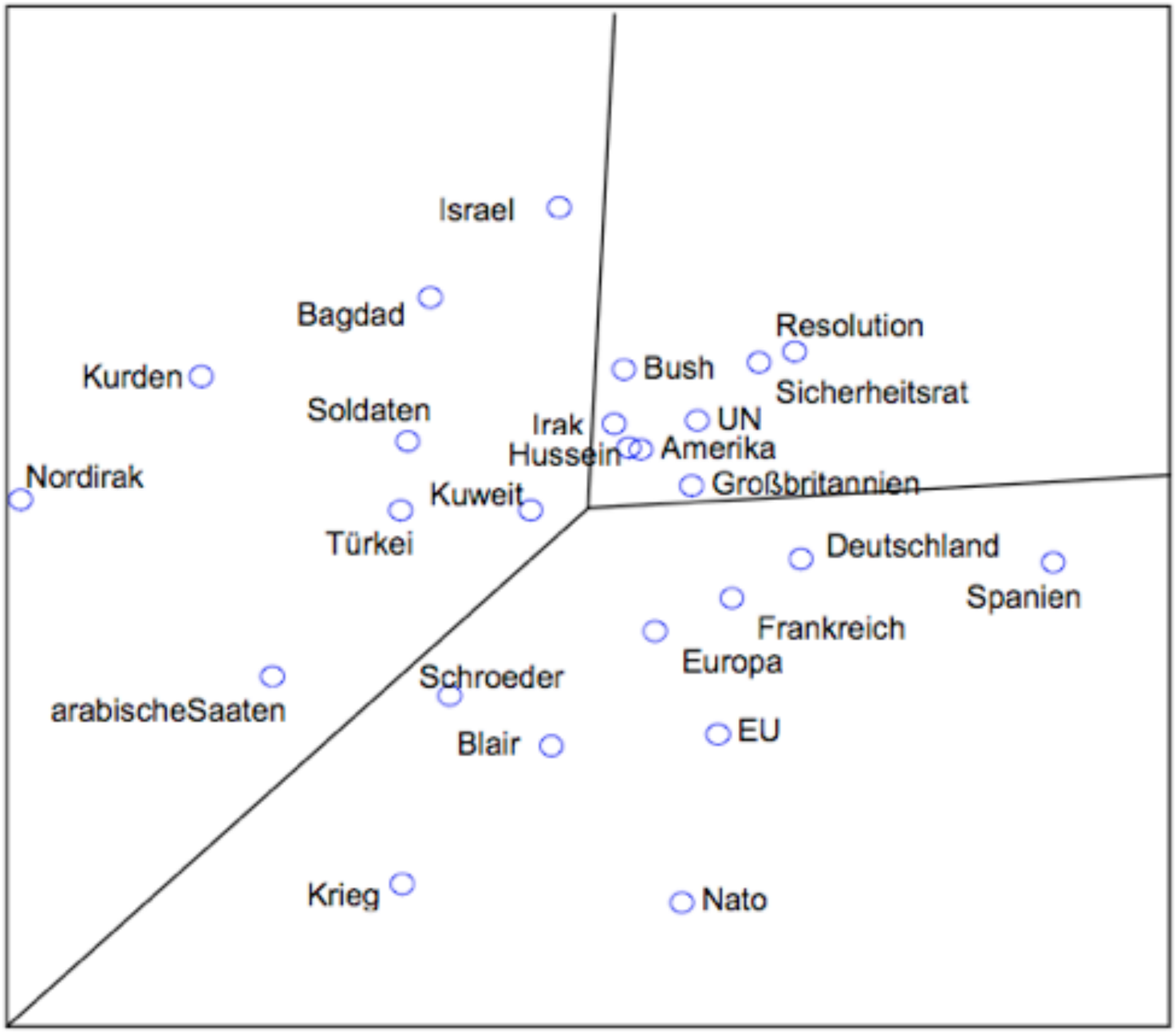
- LSI verdichtet zusammengehörige Wörter zu **Faktoren**, sog. Semantischen Indizes, mit denen weitergearbeitet wird (ähnlich wie bei einer klassischen **Hauptkomponentenanalyse**)

Term-Dokument-Matrix (Co-Occurrence)

	Peter	sehen	Student	Seminar	im	Telefon	Tag	am
d1	1	1	1	1	1	0	0	0
d2	0	0	0	0	0	1	1	2
d3	1	1	0	0	0	0	2	1

Beispiel Co-Occurrence-Analyse (Landmann/Züll)

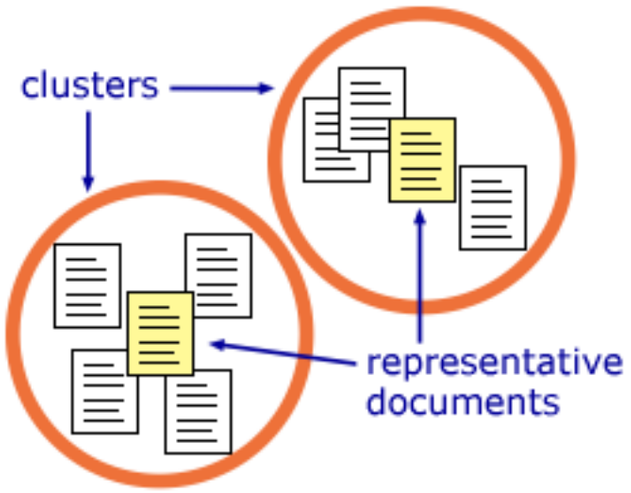
Abbildung 7: Multidimensionale Skalierung der Texte der Frankfurter Allgemeinen Zeitung zu Kriegsbeginn



Automatische Dokumentklassifikation

- Term-Dokument-Matrix lädt zum Clustern von **Dokumenten** ein, die ein ähnliches Wortprofil haben

- **k-Means** (partitionierend) Clusteranalyse als häufigstes Verfahren



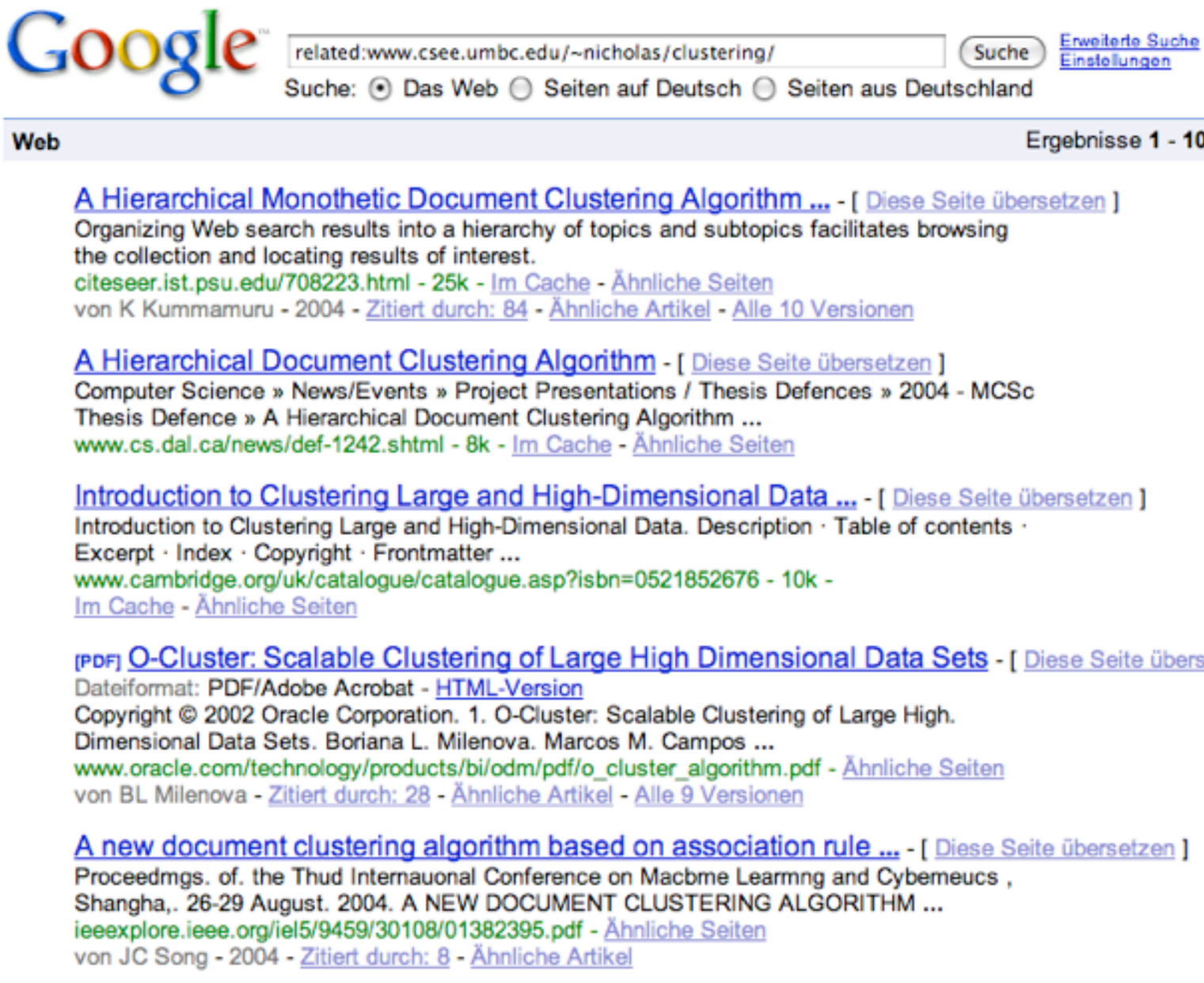
- Hierarchische Clusteranalyse (agglomerativ) skaliert sehr schlecht mit vielen Fällen und Variablen

- **Clusteranzahl** und **Interpretation** unklar, oft wenig sinnvolle Kategorien

Term-Dokument-Matrix (Document-Clustering)

	Peter	sehen	Student	Seminar	im	Telefon	Tag	am
d1	1	1	1	1	1	0	0	0
d2	0	0	0	0	0	1	1	2
d3	1	1	0	0	0	0	2	1

Beispiel Clustering von Web-Dokumenten (Google)



Gelenkte Verfahren - es wird ernst!

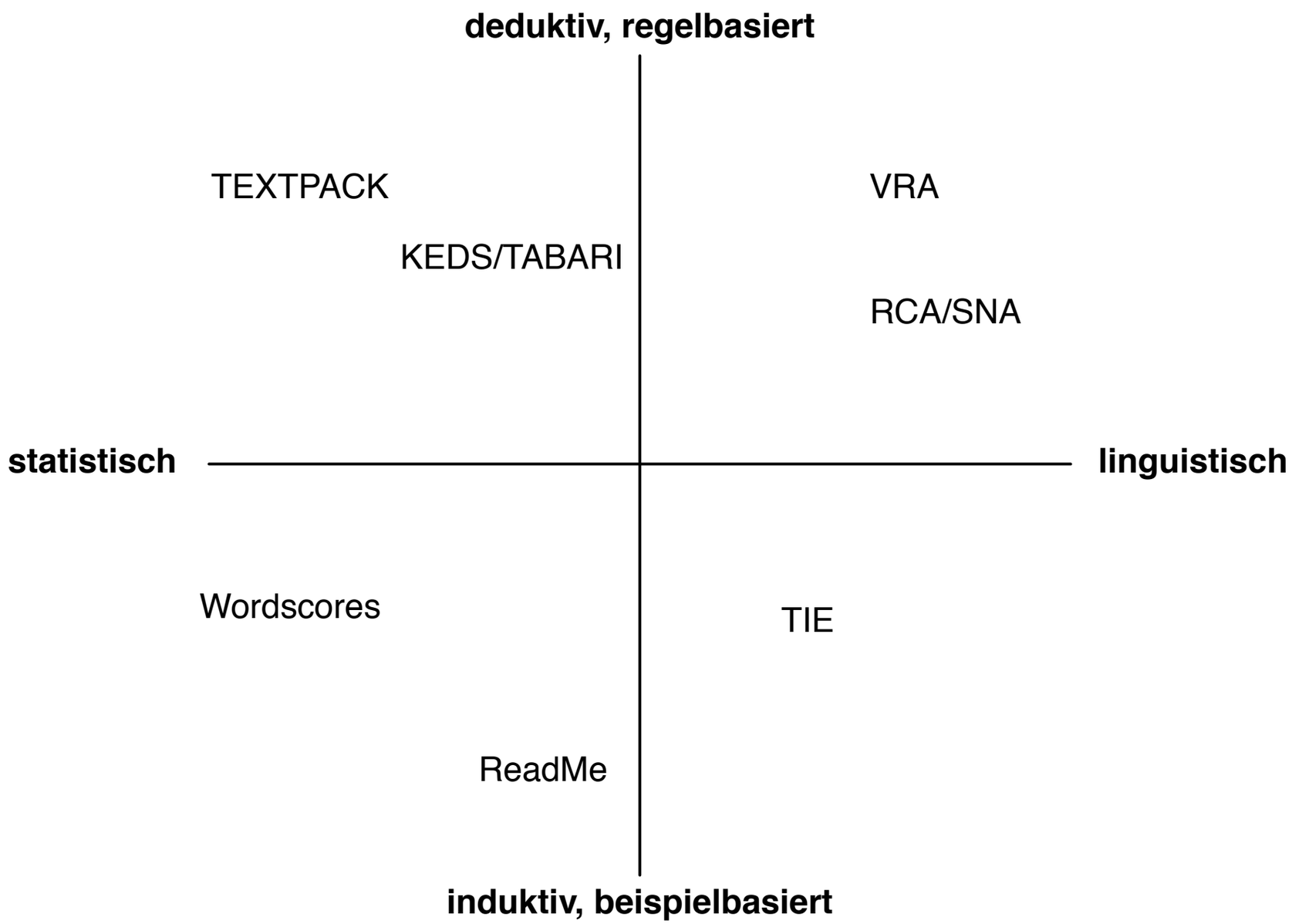
Gelenkte Verfahren - Induktion und Deduktion

- Standardfall für die Kommunikationsforschung
- **Wir haben Hypothesen, die wir überprüfen wollen!**
- Computersoftware muss gezielt gesteuert werden, aber wie?
  - durch explizite Begriffe - `^(Finanz|Banken|Kredit)krise*$`
  - durch explizite Regeln - `html.title.content.string.split[1]`
  - durch annotierte Beispiele - `doc1 = SPORT; doc2 = WIRTSCH`

Gelenkte Verfahren - Ansatz und Gegenstand

- **Bag-of-Words-Verfahren** betrachten Texte als eine (ungeordnete) Ansammlung von Wörtern bzw. Zeichen
    - rein lexikalische Ebene, Analyseeinheit ist der Text
    - klassische Inhaltsanalyse (Codierung von Themen, Nachrichtenfaktoren)
- **Syntaktisch-semantische Ansätze**
    - Texte bestehen aus Aussagen bzw. Propositionen
    - Computer soll die Bestandteile der Aussage erkennen und analysieren

Typologie gelenkter Verfahren



Diktionärsbasierte Verfahren

- **Urgestein** der automatischen Inhaltsanalyse (General Inquirer seit den 60ern)

- Forscher definieren **Wortlisten** (Diktionäre), nach denen dann Dokumente klassifiziert werden
  - „Wenn das Dokument das Wort „Fußball“ enthält, wird es als Sportmeldung codiert.“

- **Freitextrecherche** über Suchmaschinen funktioniert ähnlich, allerdings ist die Grundgesamtheit oft unbekannt, also Validität der Ergebnisse unklar.

- vollständige Reliabilitität und gute Effizienz, Validität bei komplexen Konstrukten oft sehr problematisch und Diktionärsentwicklung aufwändig

Aufgabe - Entwickelt ein Diktionär für ...

- Medienresonanzanalyse bei Volkswagen (leicht!)
    - Polo, Golf, Vfl Wolfsburg, Betriebsrat + Prostituierte, ...
- Nachrichtenfaktor Schaden (auch nicht schwer, aber lang)
    - sterben, Unfall, verunglücken, Tote, Verletzte, Absturz, Amok, ...
- Wirtschaftsnachrichten - ???
- Personalisierung der Wahlkampfberichterstattung - ???
- Postmaterialismus in Wahlprogrammen - ???

Regelbasierte linguistische Verfahren

- Umwandlung von Aussagen in analysierbare **Graphenstruktur**

- Information wird aus den Dokumenten extrahiert, d.h. offene Frage „Wer tut was mit wem?“ soll beantwortet werden

- **Syntaktische Parser** und andere Verfahren funktionieren nur bei domainspezifischen Texten, Beispiel: Extraktion von int. Ereignissen aus Tickermeldungen

- sowohl automatisches Parsing als auch die Definition von Codierregeln sind **extrem aufwändig**, meistens werden Diktionäre zur Hilfe geholt

- Verfahren eignen sich super für Online-Inhaltsanalyse, weil HTML, XML, Mails stark strukturiert sind

Induktive überwachte Textklassifikation

- da die Definition von Diktionären bzw. Regeln sehr aufwändig und wenig dem manuellen **beispielgeleiteten Codierprozess** entspricht: lernende Klassifikatoren

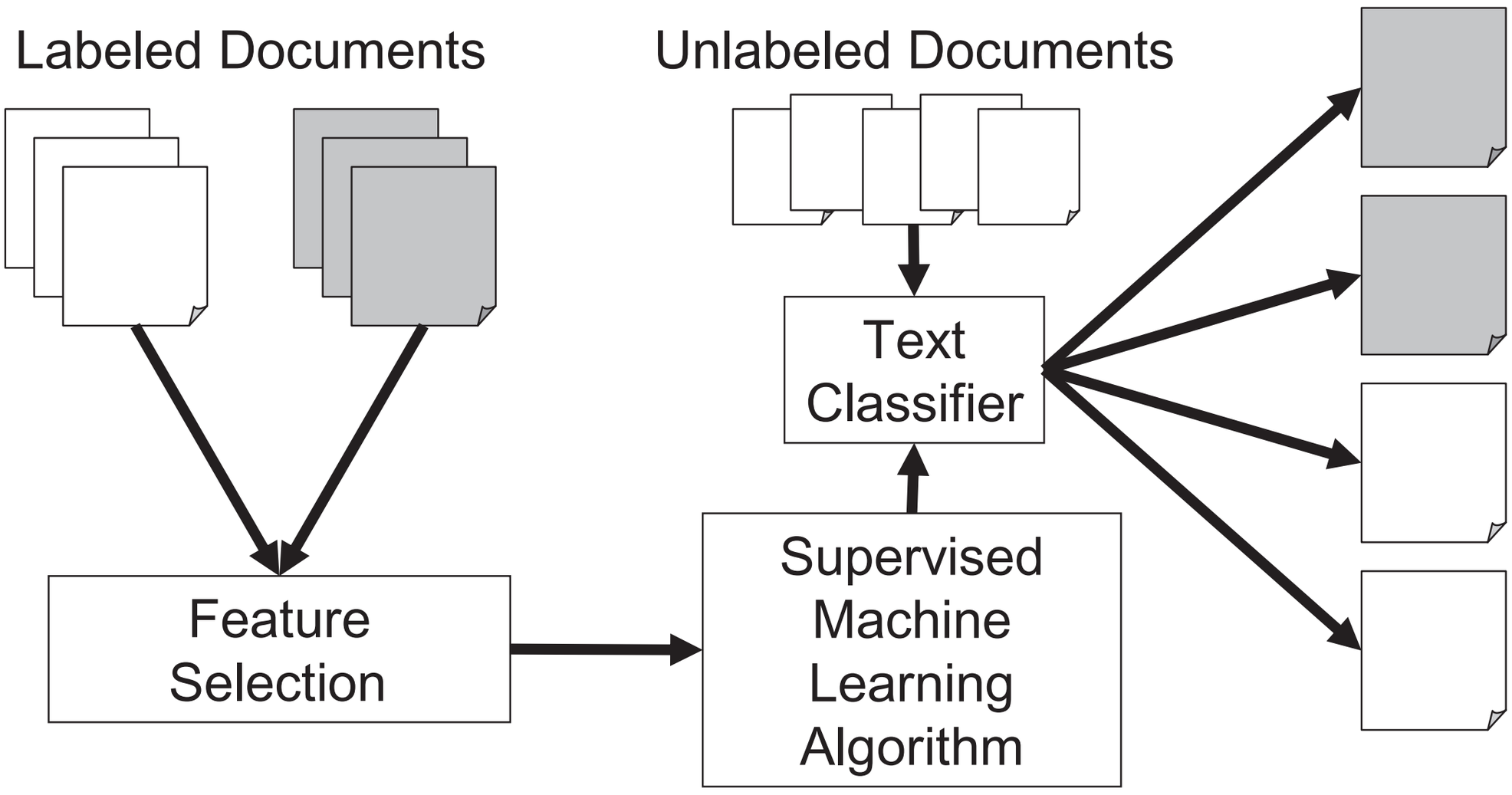
- Klassifikator erhält **Trainingsdaten**, die Rohtext und Kategorie enthalten, extrahiert daraus Feature-Gewichte (vgl. Regressionskoeffizienten), mit denen dann neue Dokumente klassifiziert werden

- man erhält ein **probabilistisches Diktionär**, das man nicht selbst definiert

- Beispiele: Wordscores zur Links-Rechts-Codierung von Wahlprogrammen

- viel Trainingsmaterial und gute Algorithmen verbessern die Validität

Funktionsweise induktiver Klassifikation



Induktive Informationsextraktion

- siehe linguistische, regelbasierte Codierung, aber **ohne explizite** Regeln
- der Computer soll **selbst lernen**, was ein Akteur, eine Handlung, ein Termin, ein Ort ist
- Annotierung der Aussagen und Inferenz auf **andere Texte** mit **anderen Strukturen** und **anderen Inhalten** (!)
- funktioniert bislang nur **sehr begrenzt**
- wäre aber das mächtigste Verfahren neben der vollautomatischen Textinterpretation

Fazit

- alle Verfahren haben in bestimmten Situationen noch Sinn
- Diktionäre eignen sich bei einfachen Konstrukten, Eigennamen, Marken etc.
- Lesbarkeit und Stilometrie sind bislang relativ selten in kommwiss. Forschung (außerhalb von Hohenheim!)
- Induktive Verfahren sind toll, weil sie anschlussfähig an bestehende manuelle Codierpraxis sind und trotzdem super mit vielen, vielen Dokumenten skalieren
- Linguistische Verfahren sind noch nicht weit genug für Standardfragestellungen.

Danke für die Einladung und die Aufmerksamkeit!

michael.scharkow@uni-hohenheim.de