Data Mining 101

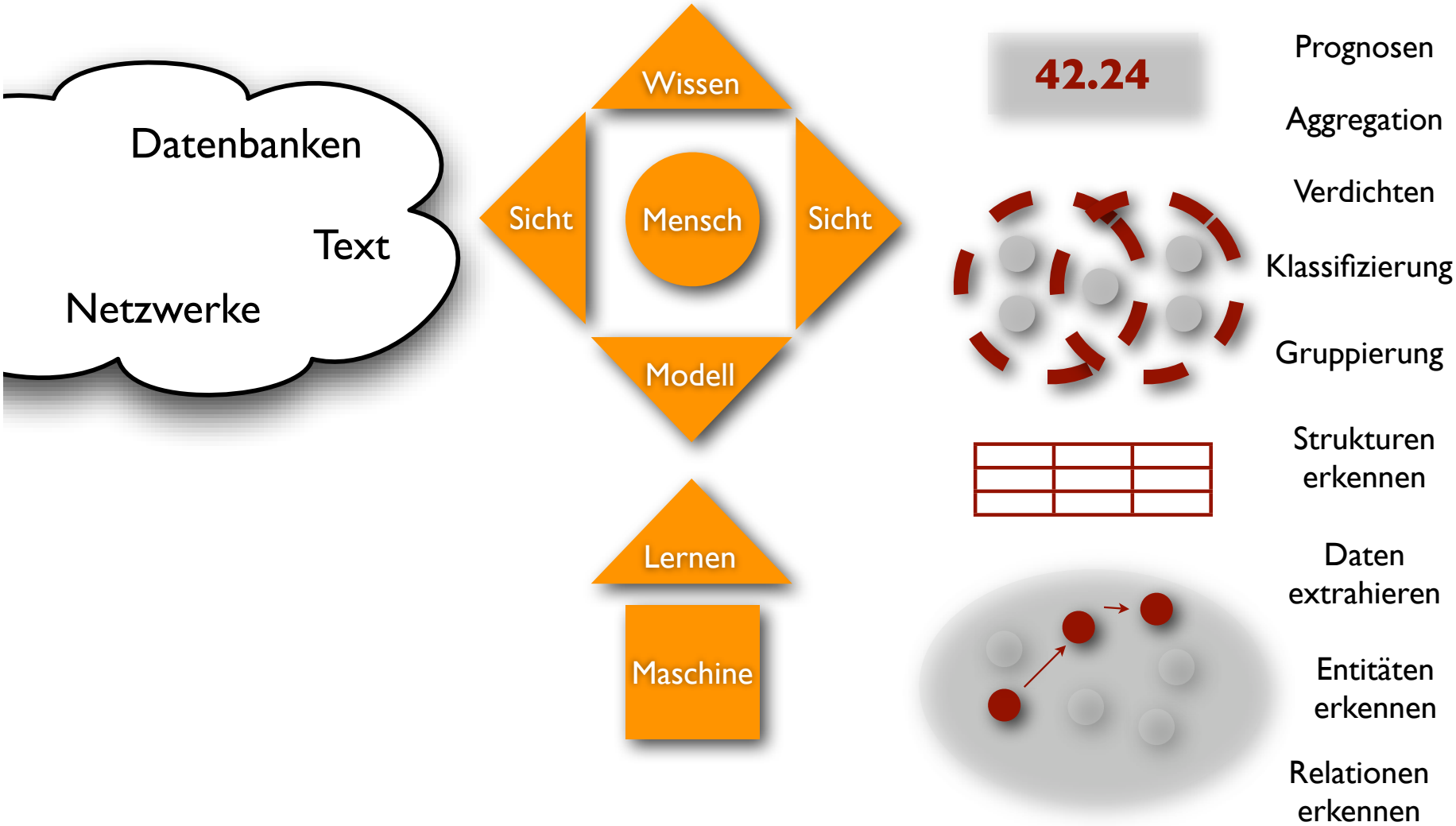
Einführung in die statistische Datenauswertung

Frank Kaufer

Potsdam, 07.05.2008

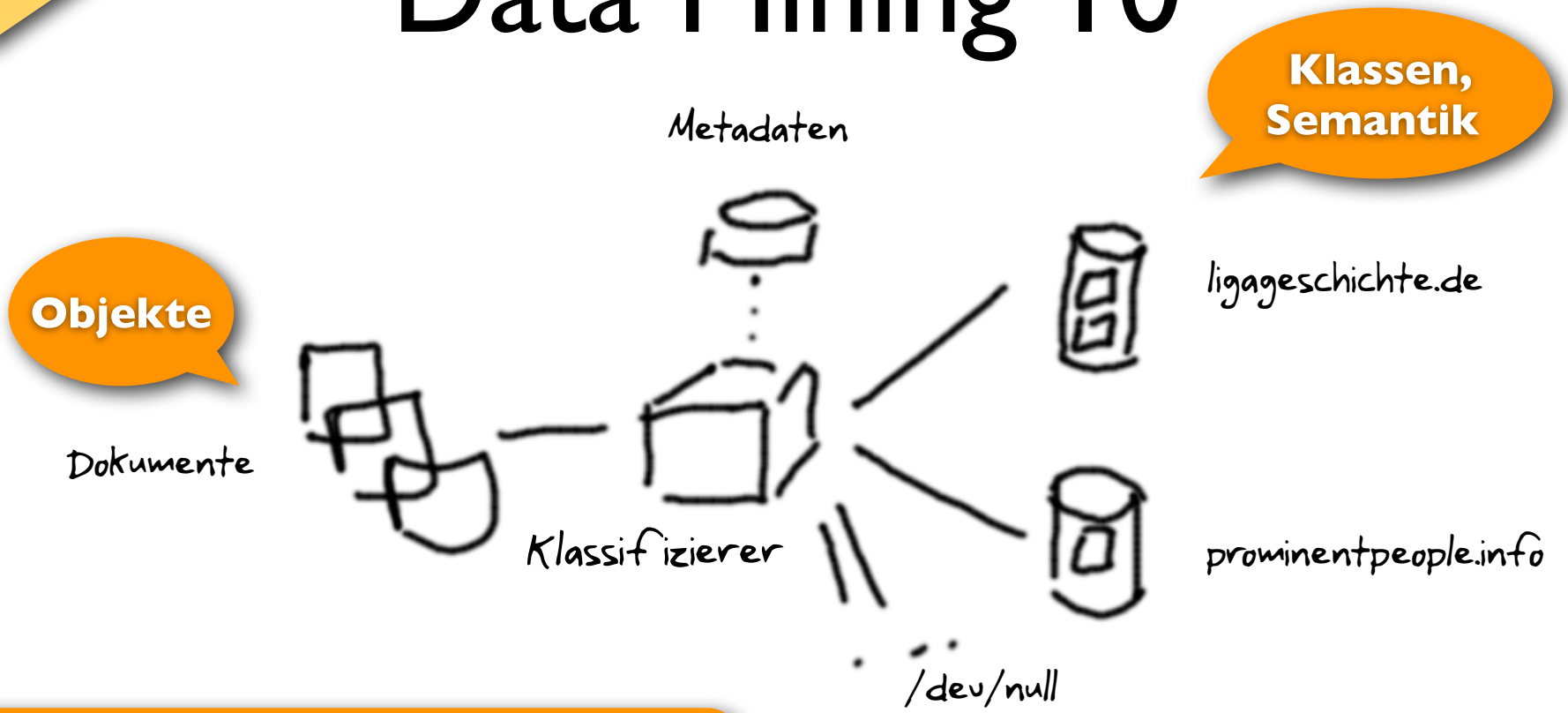
www.ligageschichte.de

Data Mining I



Dokument-
klassifizierung

Data Mining 10



if ... then ... else ...

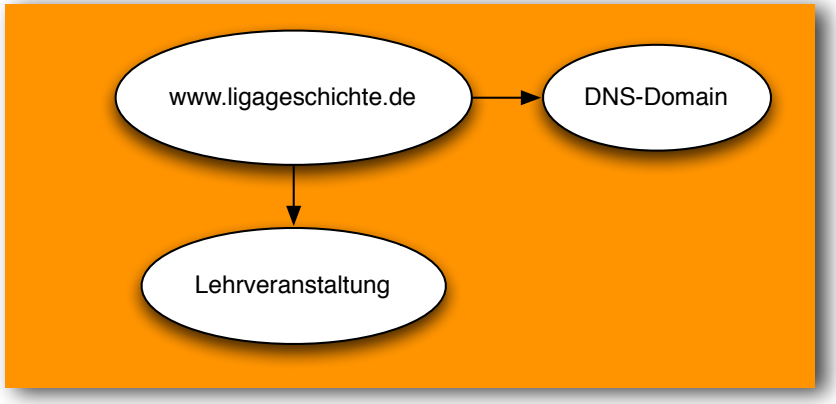
switch ... case ... default ...

$\ldots \wedge \ldots \vee \ldots \rightarrow \ldots$

Formale Semantik

Definieren vs.
Lernen

Data Mining 101



Semantik*en*

... es gibt mehrere Wahrheiten!

Semantik

Paris.getSemantics()

Paris

mit allen Sinnen

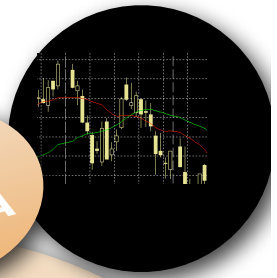
Paris

mit allen Sinnen

CDG

LFPG

SZE.PA
EAD.PA



CAC40

Paris

mit allen Sinnen

Paris

Französische Regierung

27. Juli 2005

SPIEGEL ONLINE

URANANREICHERUNG

Paris verliert die Geduld mit Iran

Paris stimmt zu

© DIE ZEIT, 15.09.1972 Nr. 37

Die Hoffnungen, daß bei der Tagung de
Außenminister der erweiterten EWG ei
Gipfelkonferenz fallen werde, haben si
französische Außenminister Schumann

DER TAGESSPIEGEL

LIBANON-MISSION

Kurz vor der Katastrophe

9.11.2006 14:45 Uhr

Paris/Tyrus - Die Flüge israelischer Kampfflugzeuge über dem Libanon haben schwere Spannungen zwischen Frankreich und Israel ausgelöst. Wie die Regierung in Paris bekannt gab, standen französische Soldaten der UN-Friedenstruppe im Libanon (Unifil) Ende Oktober kurz davor, Raketen auf israelische Kampfjets abzufeuern. Die Soldaten seien "nur zwei Sekunden" davor gewesen, "auf die Flugzeuge zu feuern, die direkt unsere Truppen bedroht haben", sagte Verteidigungsministerin Michèle Alliot-Marie. Israels Armee erklärte, sie wisse nichts von einem solchen Zwischenfall. Paris bestellte Israels Botschafter ein und verlangte ein Ende der Libanon-Überflüge. Unterdessen wurde ein neuer Überflug gemeldet.

Paris

Pariser (Einwohner)

Gott ganz nahe ...

NINA GRUNENBERG | © DIE ZEIT, 31.07.1987 Nr. 32

Den schönsten Trostpreis erhielt John Vinocur. Seit Februar dieses Jahres ist er Chefredakteur der *international Herald Tribüne,* an der die *New York Times* zu einem Drittel Eigentümerrechte hat. Die Rückkehr nach Paris, dem Sitz der Zeitung, machte Vinocurs Glück vollkommen. Jetzt kann er wieder versuchen, „wenigstens für ein paar Minuten am Tag wie Gott in Frankreich zu leben": „Man kann es schaffen", versichert er ernsthaft wie einer, der im Training ist: Zum Beispiel morgens um halb acht Uhr, wenn Paris noch schläft und er allein auf der Straße ist, in ein Cafe geht, seine Zeitung liest und in ein Croissant beißt - „dann bin ich Gott ganz nahe".

Paris

Französische Wirtschaft

Paris kauft am meisten

Ähnlich wie der gesamte Außenhandel entwickelte sich, der Warenverkehr mit den anderen EG-Ländern. Mit 52 Prozent der gesamten Importe und 47 Prozent Anteil am deutschen Export hatten die acht Staaten einen überragenden Anteil am deutschen Außenhandel. Frankreich steigerte seine Käufe im vergangenen Jahr von 19,4 auf 23,2 Milliarden Mark und führt mit Abstand die

Paris

Pariser Börse

finanzen.net

20.09.2005 14:11

Börsen in Europa: Märkte ignorieren sämtliche Störfeuer

Ein weiterer Gewinner des Ölpreises ist der französische Mineralölkonzern TOTAL, der an der Heimatbörse ebenfalls hinzugewinnen kann. Die France Telecom begibt über 37 Millionen neue Aktien zur Finanzierung der Übernahme von Amena, den drittgrößten Mobilfunkbetreiber in Spanien. Die Aktien werden im Rahmen einer Privatplatzierung ausgegeben. Dem Kurs schadet die Nachricht nicht, In Paris gewinnt das Papier rund 1 Prozent. Leichte Gewinne verbucht auch der Versicherungskonzern Assurances Générales de France, der heute bekannt gab, dass man seinen Gewinn im ersten Halbjahr 2005 ausweiten konnte.

Paris

Das französische NOK

http://sport.orf.at/050706-20098/20099bigstory_txt.html

Paris knapp gescheitert

Franzosen können die Niederlage nicht fassen.

Londons Sieg als Gastgeber der Olympischen Spiele 2012 hat in Frankreich abgrundtiefe Enttäuschung ausgelöst.

http://sport.orf.at/050706-20098/

London sticht Paris aus

Paris

Paris Saint Germain Football-Club

Paris gewinnt Ligapokal

Veröffentlicht: 29.03.2008 - 22:16 Uhr Quelle: transfermarkt.de

Der abstiegsbedrohte Traditionsverein Paris St. Germain gewann am Abend den französischen Ligapokal gegen den ebenfalls abstiegsbedrohten RC Lens. Held des Abends war Bernard Mendy (Bild), der einen Elfmeter in der 92. Minute zum 2:1-Sieg verwandelte

Niederlage für Lyon im Spitzenspiel - Auch Paris verliert

Tabellenführer Olympique Lyon hat auf dem Weg zum erhofften erneuten Meisterti
einstecken müssen. Der Titelverteidiger unterlag am Sonntagabend im Spitzenspiel bei
der Serienmeister mit 67 Punkten sechs Spieltage vor Saisonende noch sechs Zähler V
mit 52 Punkten auf Rang vier. Paris St. Germain musste eine 0:1-Niederlage beir
Ligapokalsieger rutschte mit 35 Zählern als 18. auf einen Abstiegsrang.

22:52 06-04-2008

AMERIKA WOCHE

Paris unter Druck

Semantik ungewiss

REUTERS

Italien-Trainer Donadoni: In Paris unter Druck

SPIEGEL ONLINE

04. September 2006, 15:26 Uhr

REVANCHE GEGEN FRANKREICH

Zitternder Weltmeister

Nach Todesschuss auf Fan: Paris unter Druck

Paris (dpa) - 25.11.2006, 13:52 Uhr

FUSSBALL24.DE

Nach dem tödlichen Schuss eines sc
Polizisten auf einen Hooligan gerät d
französische Fußballverein Paris Sain
Germain (PSG) wegen seiner rassisti
Fans zunehmend unter Druck.

FINANCIAL TIMES
DEUTSCHLAND

18. 06. 2006

Suez setzt Paris unter Druck

Der Versorger Suez hat der Regierung ein Ultimatum für die geplante Fusion im Volumen von 72 Mrd. Euro mit Gaz de France (GdF) gestellt. Paris müsse die dafür nötige Gesetzesänderung bis zum Sommer auf den Weg bringen.

115,70 -	13,49	• Tiscali		19,34	-0,90

PARIS *unter Druck*

CAC-40:
5761,97 (-1,54%)
Vortag: 5852,35

07.02. (16:30 Uhr)
Veränderung seit
30.12.99: -3,30%
29.12.00: -2,77%

51,00 -	35,70	• Accor	47,20	0,20

SPIEGEL

DER SPIEGEL 20/1969 vom 12.05.1969, Seite 29

"ES WÄRE FALSCH, PARIS UNTER DRUCK ZU SETZEN"

SPIEGEL-Interview mit Bundesaußenminister Willy Brandt Ober die Bonner Europa-Politik nach de Gaulle

AOL

Donnerstag, 01. Mai 2008

Spiele

"The Prison Life": Paris unter Druck

Im Knast steht Paris am längeren Hebel - lassen Sie die prominente Knastjoule am Fließband schuften:Nummernschilder stanzen in "The Prison Life

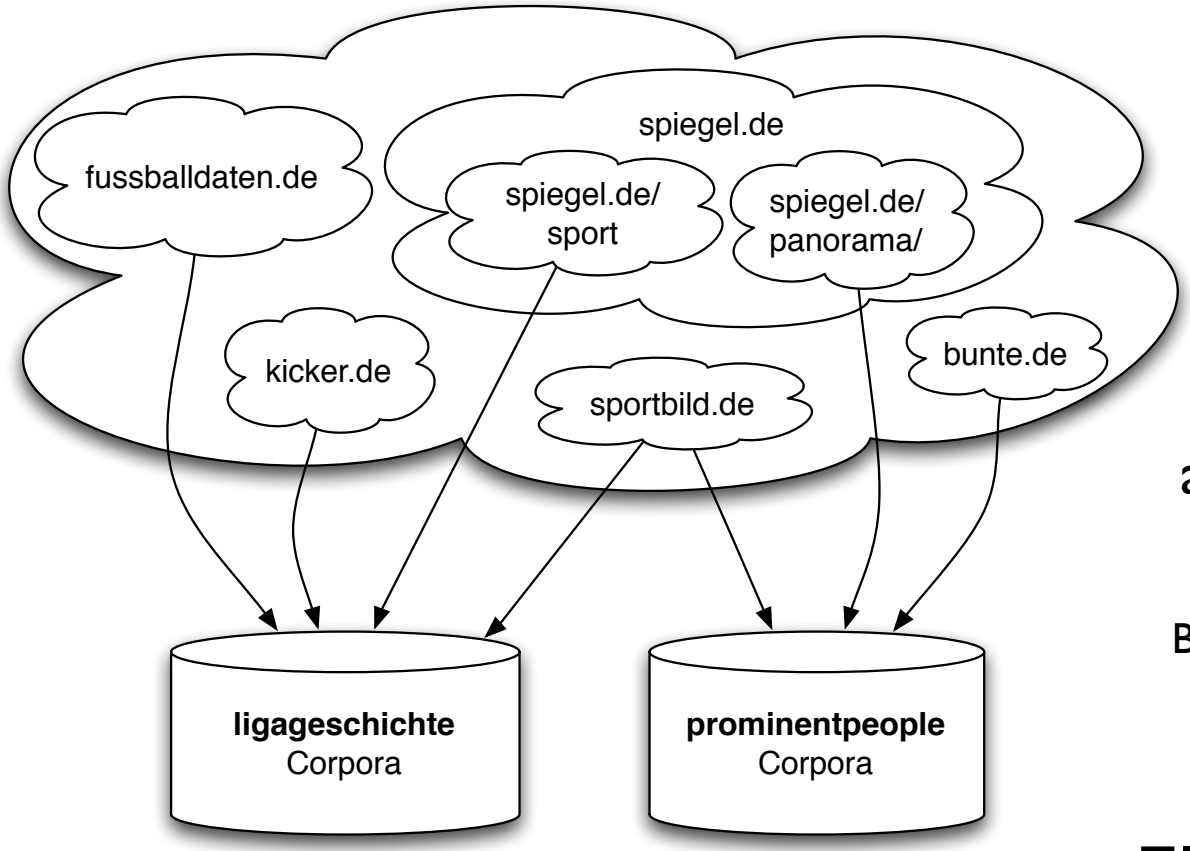
Kontext

Kontext

- Was macht einen Kontext aus?
- Was/Wer grenzt ihn gegenüber anderen ab?
  - Herkunft: Autor, Produzent
  - Übertragung: Darstellung, Zeit, Motivation
  - Ziel: Leser, Konsument

WER stellt WANN und WARUM WEM WELCHE Information WIE dar?

Quell-/Zielkontext Domänen



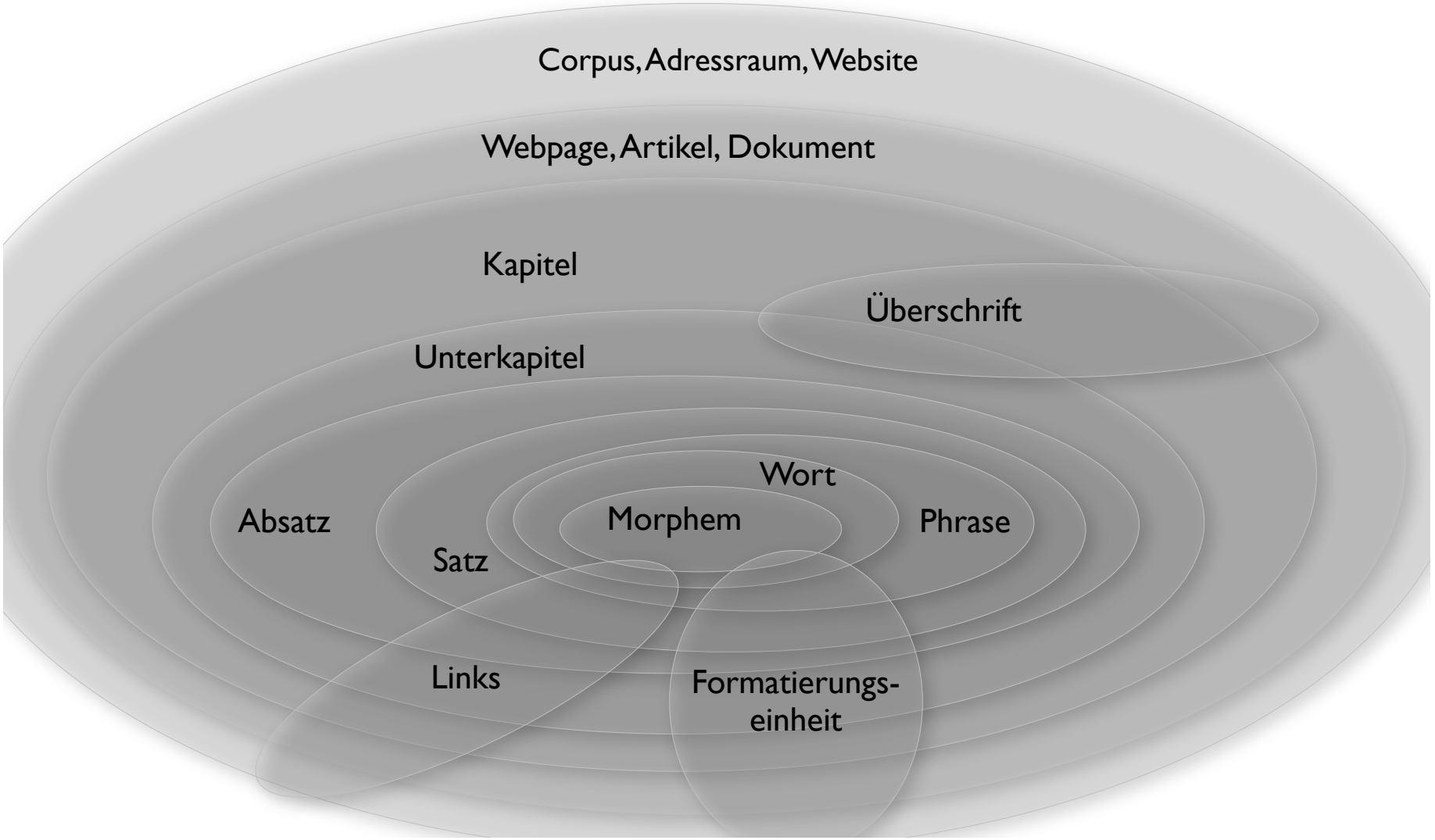
Beckenbauer, Matthäus

"nach Diskobesuch aus Kader gestrichen"

Nennung von Fußballern in BUNTE im Vergleich zu anderen Sportlern -> gesellschaftliche Bedeutung von Fußball

Floskeln

Struktureller Kontext



Kontext Dokument

http://www.kicker.de/news/fussball/bundesliga/startseite/artikel/378042

Überblick: Fernduell von Werder und S04

Ärgert Koller seinen Ex-Klub?

Vor dem Schaulaufen des designierten Meisters FC Bayern [...] wird sich bei den abstiegsbedrohten Rostockern zeigen.

Nürnberg ... Dortmund ... nach dem ... Bielefeld ... Arminia am Samstag gegen Bochum ... Dortmund ...

Geht dem HSV die Luft aus?

Hamburger SV ... Rostock ... HSV-Treffer ... Hansa ... nach dem ... Cottbus ...

Fernduell um Platz zwei

Hannover ... "Königsblauen" ... in der Torschützenliste bis auf Platz zwei vorgearbeitet.

Platz zwei in der Tabelle hat Bremen inne, das die auswärtsschwachen Cottbuser (1/6/7) zu Gast hat

Befreiungsschlag für Bielefeld?

Bielefeld ... Arminia ... Nürnberg ... Ostwestfalen ... Bochum ... Nürnberg

Kontext Satz

- "Paris unter Druck"

- "Flug nach Paris"
- "One night in Paris"

Kontext Wort/Entität

Paris Saint Germain Football-Club

Paris

FC Paris St. Germain

Paris Saint Germain

Paris St. Germain

PSG

Paris Saint-Germain

Paris Saint-Germain FC

Paris SG

Kontext Wort/Entität

Paris Saint-Germain

1. Fußballverein Paris **Saint-Germain** Football Club (PSG)
2. Vorstadt **Saint-Germain**-en-Laye
3. Stadtteil **Saint-Germain**-de-Prés

Paris en detail

- **Paris Saint Germain (PSG)**: Kader 2008, 2007, ..., Fußballabteilung, 1. Mannschaft Herren, Frauenmannschaft, Jugend, Unternehmen/Gesellschaft/Verein, Stadion, Verwaltungssitz

- **Stadt Paris**: geographisch, geometrisch (Ort, Ausdehnung), politisch (Ile de France, Stadtteil), historisch/zeitbezogen, Menge von Menschen, kulturell, städtebaulich
  "*Wieviele Einwohner hat Paris?*"
  - Wann? Wer zählt als Einwohner?

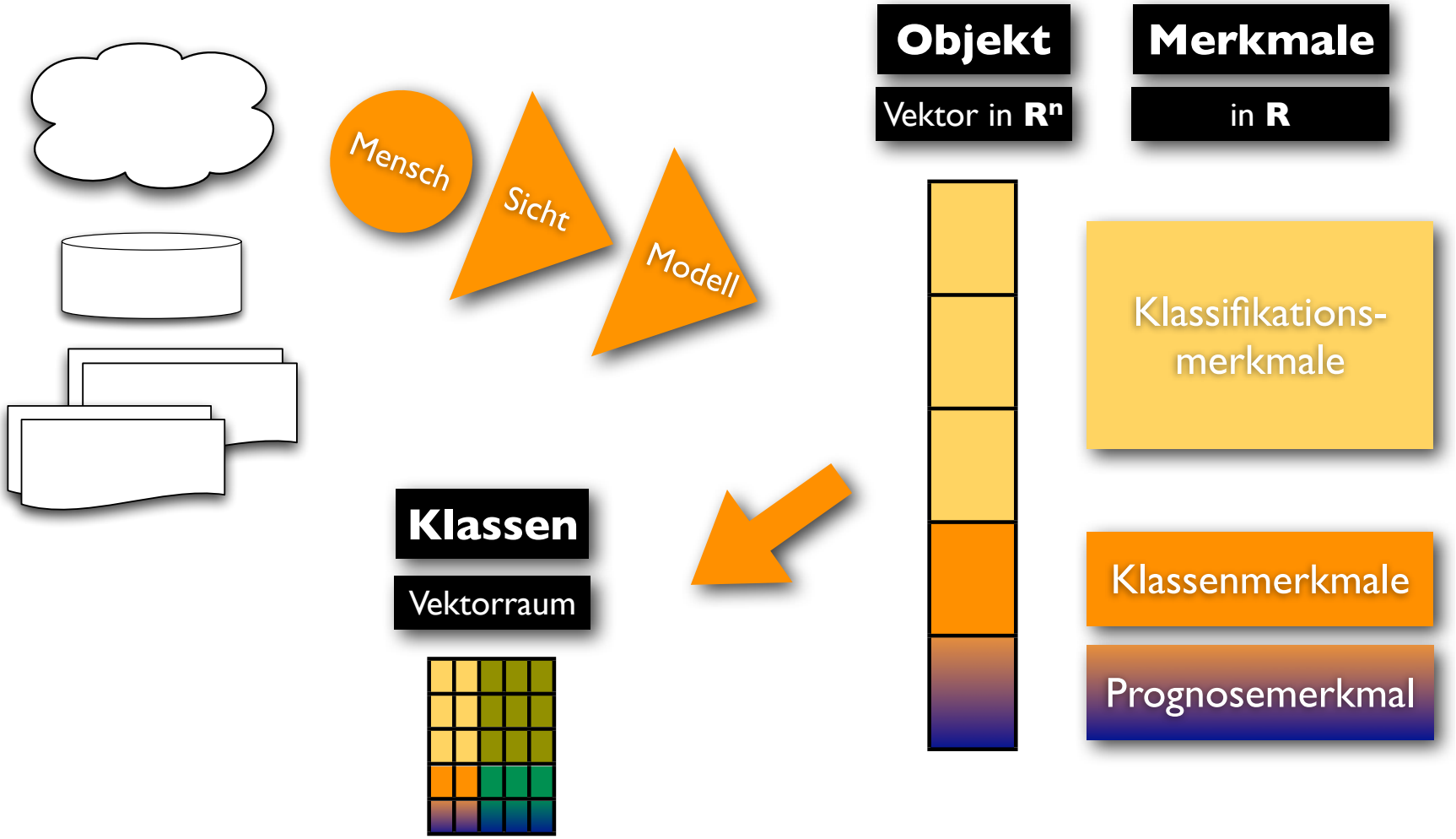
- **Paris Saint Germain**: (1) Fußballverein Paris Saint-Germain Football Club (PSG), (2) Stadtteil Saint-Germain-de-Prés, Vorstadt Saint-Germain-en-Laye

Kontext durch Metadaten

- Ordner, Tags, Partitionen
- Graphen, Listen, Relationen
- Domänendaten, Ontologien
- Formale Logik
- Statistik, Wahrscheinlichkeiten
- **Lernen vs. Spezifizieren** (Berechnung vs. Eingabe)
- **Implizit vs. Explizit** (Duck-Typing vs. Benennung)

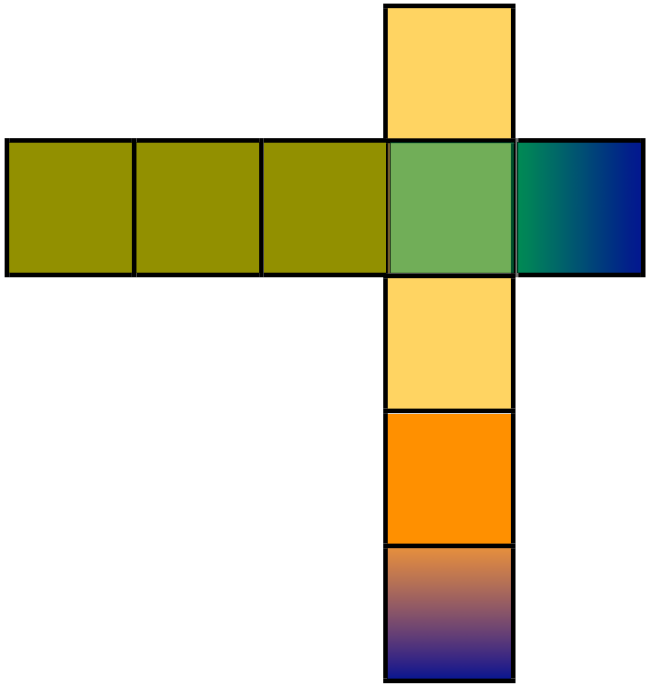
Modelle

Objekte und Klassen



Aggregation

- Kontexte
- Gewichte
- Merkmale
- Objekte
- Klassifizierer



Merkmale

Feature

- Alle bestimmbaren Daten eines Kontextes
- Daten und Metadaten
  Wörter, Phrasen, Formatierung,
  Wörterbucheintrag, Vernetzung
- Merkmal vorhanden/nicht vorhanden
- Gewicht: Häufigkeit, Bedeutung

Merkmale: Kriterien

- Häufiges Auftreten
- Nicht ganz so häufiges Auftreten
- Teilen Objektmenge möglichst gut auf
- Redundanz vermeiden, evt. Aggregation

Transformation

- Text in Vektormodell (Information Retrieval)
- Aggregation von Daten
- Nominale Merkmale
- Abstrakte Merkmale (gut)
- Semantische Äquidistanz der Merkmalskalen

Textaufbereitung

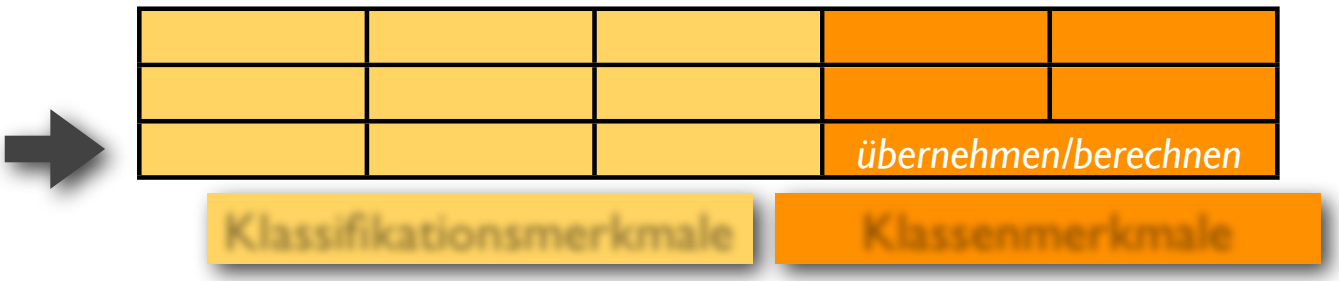
- **Tokenisierung**: Interpunktion, Formatierung, Komposita: Zerlegung (de), Bestimmung (en)
- **Stopwörter** ("und", "der", "Drucken")
- **Intrinsische Normalisierung**: Groß-/ Kleinschreibung, Lemmatisierung, Stemming, Abkürzungen, Phonetische Normalisierung
- **Extrinsische Normalisierung:** TF/IDF, Abbildung auf Synonymmenge Klasse in Ontologie, Übersetzung
- **Analyse/Annotierung**: POS, Phrasenerkennung, grammatikalische Analyse

Anwendungen

Klassifikation (Themen)

Topic Detection

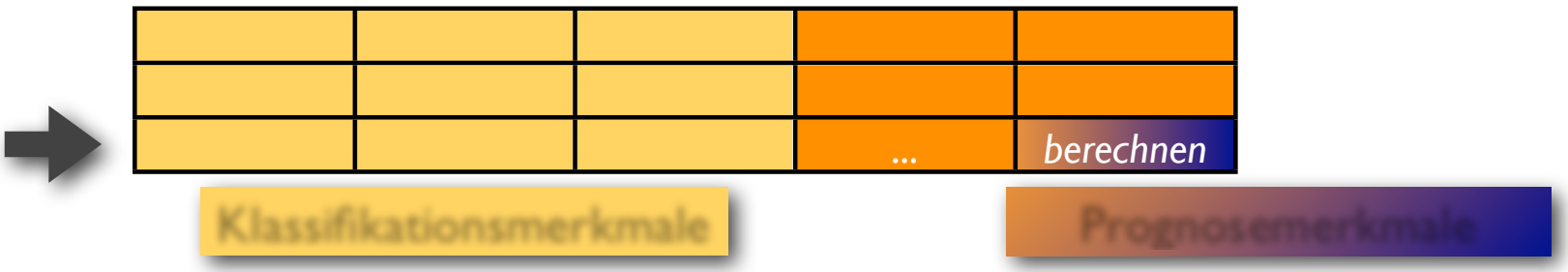
1. Wörter/Phrasen in Kontext
2. Zuordnung von Wörtern zu Themen
    1. Clustering: implizite Themen/Merkmale
    2. Klassifizieren: explizite Themen/Merkmale
3. Themenranking je Kontext
4. Trenderkennung über zwei Kontextdimensionen



Prognose

nach/mit Klassifikation

- **Mannschaftsprognose**:
  nach Anz. Nationalspieler, letzten k Saisonabschlüsse, Budget, ... gruppieren
  Saisonabschluss abschätzen
- **Spielerprognose** (wer kann mit wem): gemeinsame regionale Herkunft, Altersunterschied, gem. Siege/ Niederlage/Erfolge,
  numerischer Positionsunterschied
  gemeinsame Erfolge abschätzen



Informationsextraktion

- Entitäten
- Zusammenhänge
  - Regeln
  - Beziehungen
- Techniken: Matching gegen Metadaten, linguistische Analyse, domänenspezifische Heuristiken und Regeln, Templates, Regex, Lernen

Maschinelles Lernen

Über- vs. Unüberwacht

- **Unüberwacht** (*unsupervised*): Klassenkriterien unbestimmt, ergeben sich durch Ähnlichkeit von Instanzen, i.A. muss Anzahl/Größe der Klassen festgelegt werden
- **Überwacht** (*supervised*): Klassenkriterien sind vorbestimmt oder werden durch Eingaben gelernt

Unüberwachtes Lernen

- **k-Means** Clustering
- Neuronale Netzwerke
- Assoziationsregeln

Überwachtes Lernen

- **Naive Bayes**, Fisher
- Entscheidungsbäume
- Neuronale Netzwerke
- Support-Vektor-Maschinen

Training

- Initiale Trainingsmenge
- Training während Benutzung
- Maschinelle Lerner
  - starten unsicher
  - konvergieren aber und werden sicherer

Bayessche Klassifikation

Klassifikation & Prognose

- Spam/kein Spam
- Arbeit/Privat
- Klatsch/Sport/Politik/Wirtschaft
- Bericht über Mannschaft X, Y, Z
- Gutes Spiel/Schlechtes Spiel

Klassisch: Regeln

- Manuell -> aufwändig
- Schwer zu bestimmen
- Adaptivität

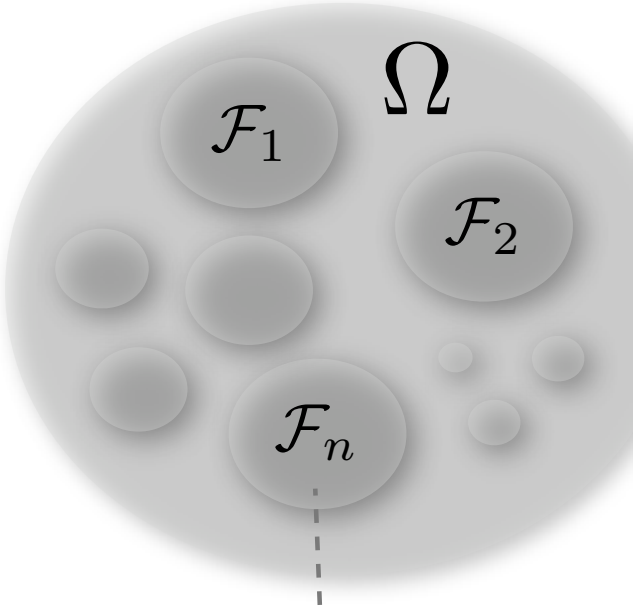
- **Besser: Regeln implizit und individuell lernen**

Wahrscheinliche Merkmale

Merkmals raum

X

$P(X \in \mathcal{F}_i)$



Objekt

Wahrscheinlichkeit,
dass Objekt ein Merkmal besitzt,
einer Merkmalsklasse angehört

Merkmalsklasse

Neue Email

Wahrscheinlichkeit, dass Email beruflich ist

Beruf

Privat

P.

$$P(F_i) := P(X \in \mathcal{F}_i)$$

$$P(F_1, ..., F_k) := P(X \in \mathcal{F}_1 \cap \ldots \cap \mathcal{F}_k)$$

Wahrscheinlichkeiten

$$P(F_i) := P(X \in \mathcal{F}_i)$$

Woher nehmen,
wenn nicht stehlen?

$$P(F_i) := P(X \in \mathcal{F}_i)$$

$$P(F_i) := \frac{|\mathcal{F}_i|}{|\Omega|}$$

Statistische Datenanalyse

Schätzen

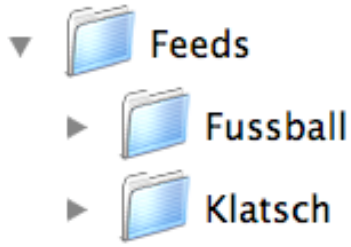
Zählen!

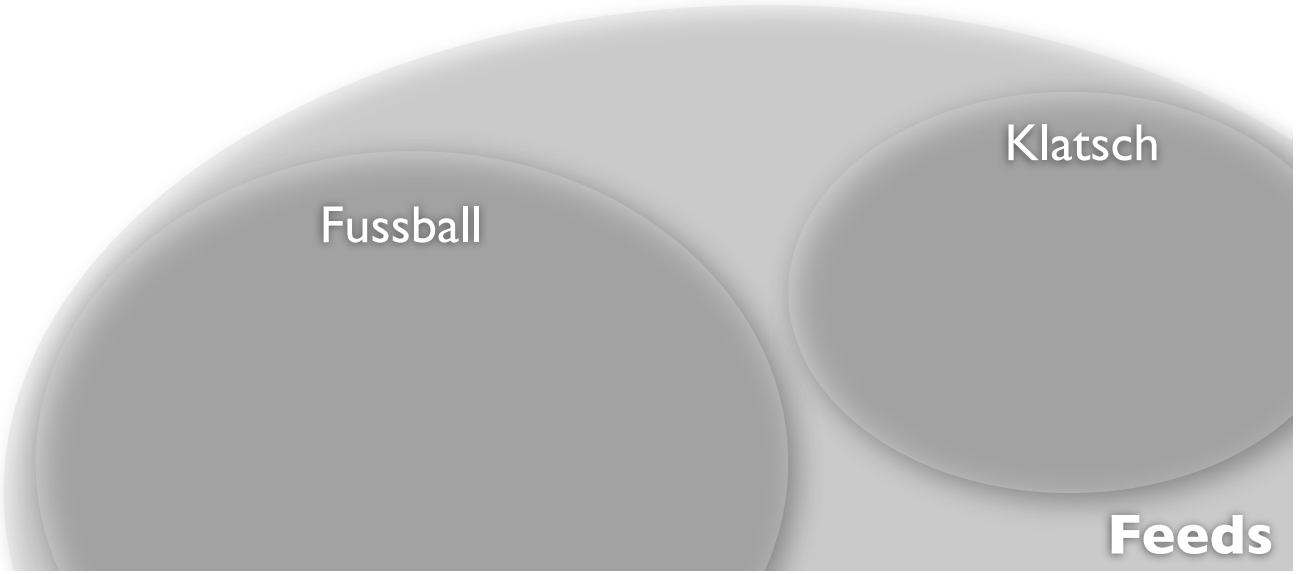
Klassifikation?

$$P(KlatschFeedItem) := \frac{|Klatsch|}{|Feeds|}$$

$$P(FussballFeedItem) := \frac{|Fussball|}{|Feeds|}$$

Beschreibende Statistik





Klassifikation!

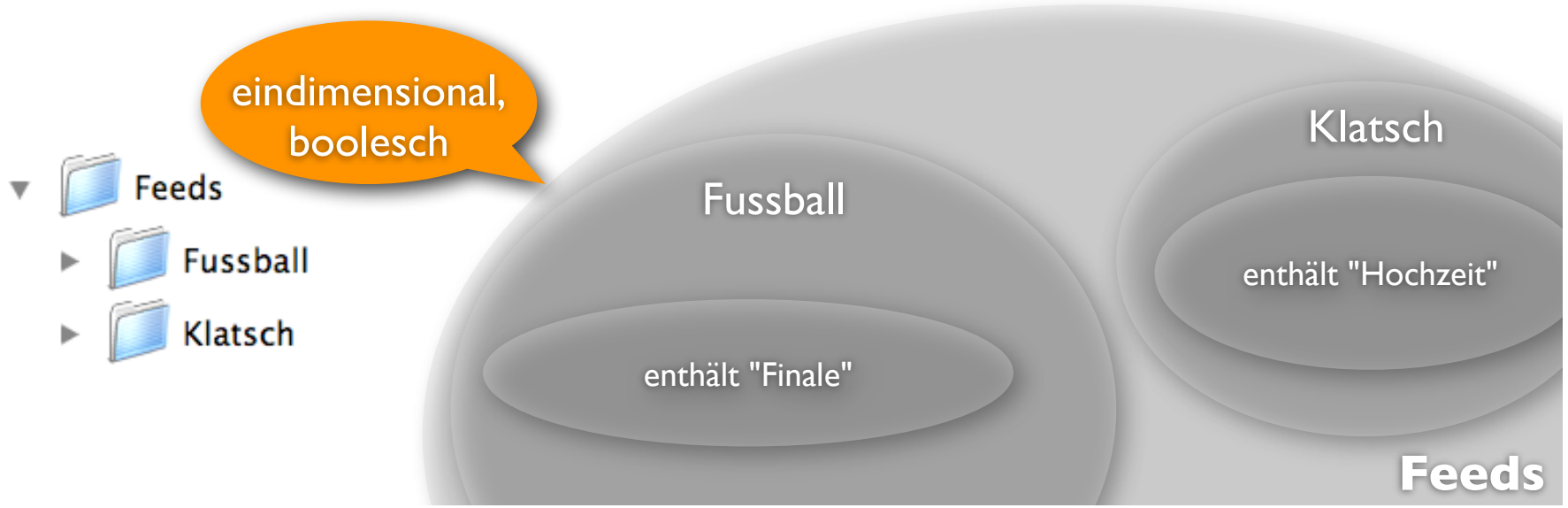
$$P(Klasse|Merkmal) = \frac{P(Klasse \cap Merkmal)}{P(Merkmal)}$$

$$P(Fussball|Finale) = \frac{P(Fussball \cap Finale)}{P(Finale)}$$

$$P(Fussball|Hochzeit) = \frac{P(Fussball \cap Hochzeit)}{P(Hochzeit)}$$

$$P(Klatsch|Finale) = \frac{P(Klatsch \cap Finale)}{P(Finale)}$$

$$P(Klatsch|Hochzeit) = \frac{P(Klatsch \cap Hochzeit)}{P(Hochzeit)}$$



Naive Bayes

$$P(Fussball|Finale, Hochzeit, Beckenbauer) = \frac{P(Fussball, F, H, B)}{P(F, H, B)}$$

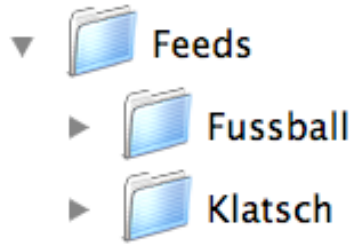
$$= \frac{P(Fussball) \times P(F, H, B|Fussball)}{P(F, H, B)}$$

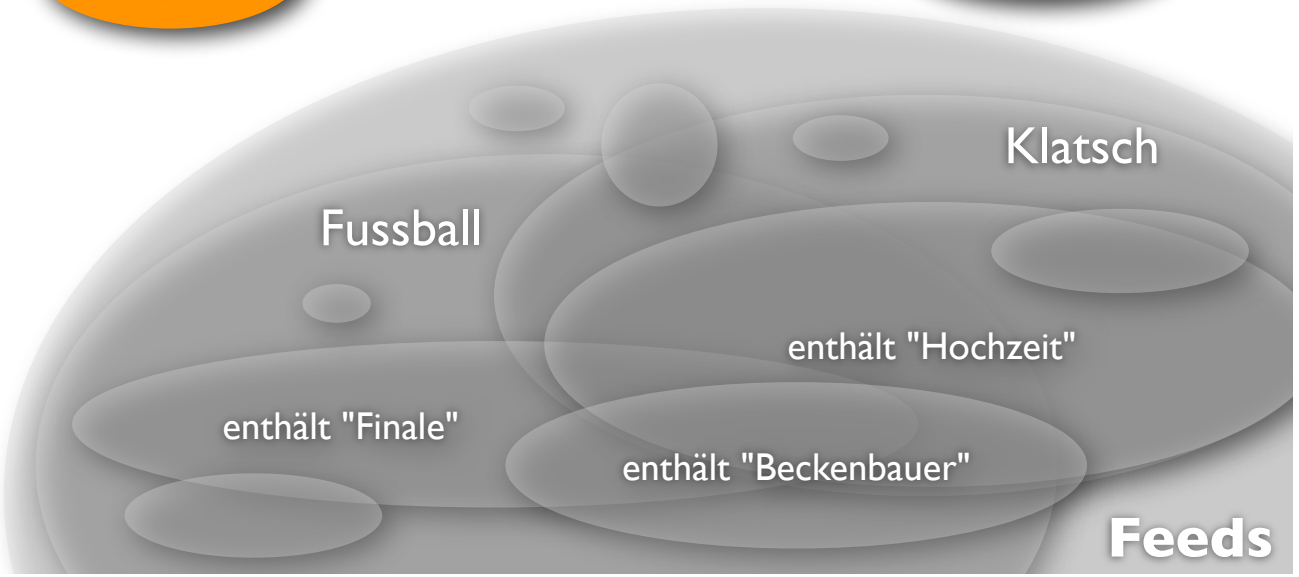
Bayes

$$= P(Fussball) \times P(F|Fussball) \times P(H|Fussball) \times P(B|Fussball) \times \frac{1}{P(F,H,B)}$$

Naive

irrelevant





Klassenauswahl

$$P(Fussball|Finale, Hochzeit, Beckenbauer)$$
$$= P(Fussball) \times P(F|Fussball) \times P(H|Fussball) \times P(B|Fussball)$$

$$P(Klatsch|Finale, Hochzeit, Beckenbauer)$$
$$= P(Klatsch) \times P(F|Klatsch) \times P(H|Klatsch) \times P(B|Klatsch)$$

Wähle Klasse mit größtem Wert
(beachte Mindestwerte, Mindestabstände)

Unabhängige, partionierende Merkmale

Klasse **Bayern Gewinnt**

Luca Toni zuletzt stark
Luca Toni trifft
Luca Toni in Startformation
Luca Toni ist Italiener

Ribery zuletzt stark
Luca Toni trifft
Kahn im Tor
kein internationales Spiel zuvor

Naive Bayes funktioniert in der Praxis auch bei nicht vollständiger Unabhängigkeit sehr gut!

Gastvortrag

Named Entity Recognition

Alexander Kuscher
Christoph Tiele