Statistische Modellierung in der Computerlinguistik

Sebastian Pado

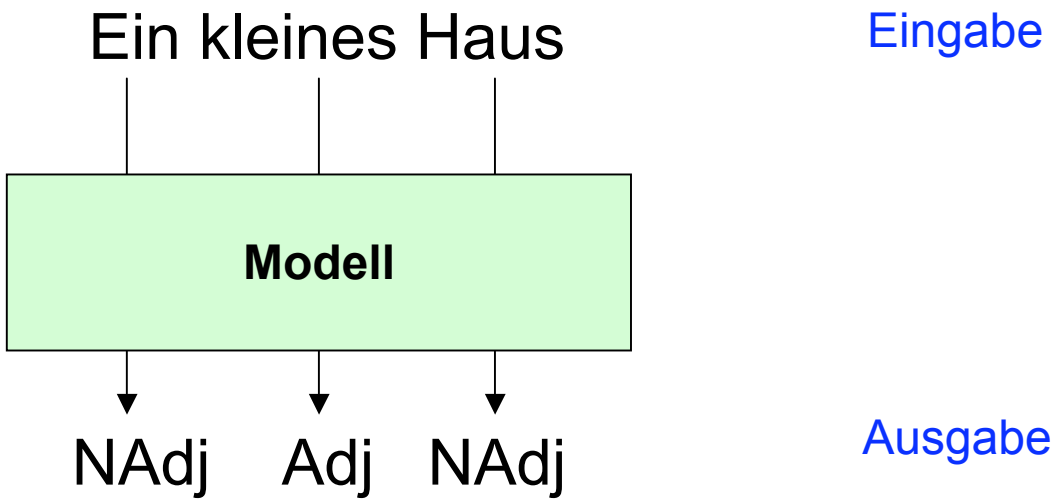
17.01.2006

Computerlinguistik 2005 (Rückblick)

- Ziel:
  - Modelle zur automatischen linguistischen Analyse von Text
- Herausforderungen:
  - Mehrsprachigkeit
  - Große Textmengen, gemischte Qualität
- Wünsche an ein Analysemodell:
  - Robust
  - Mit geringem Aufwand zu konstruieren

Beispielaufgabe: Adjektiverkennung

- Handelt es sich bei einem Wort (in einem fortlaufenden Text) um ein Adjektiv?

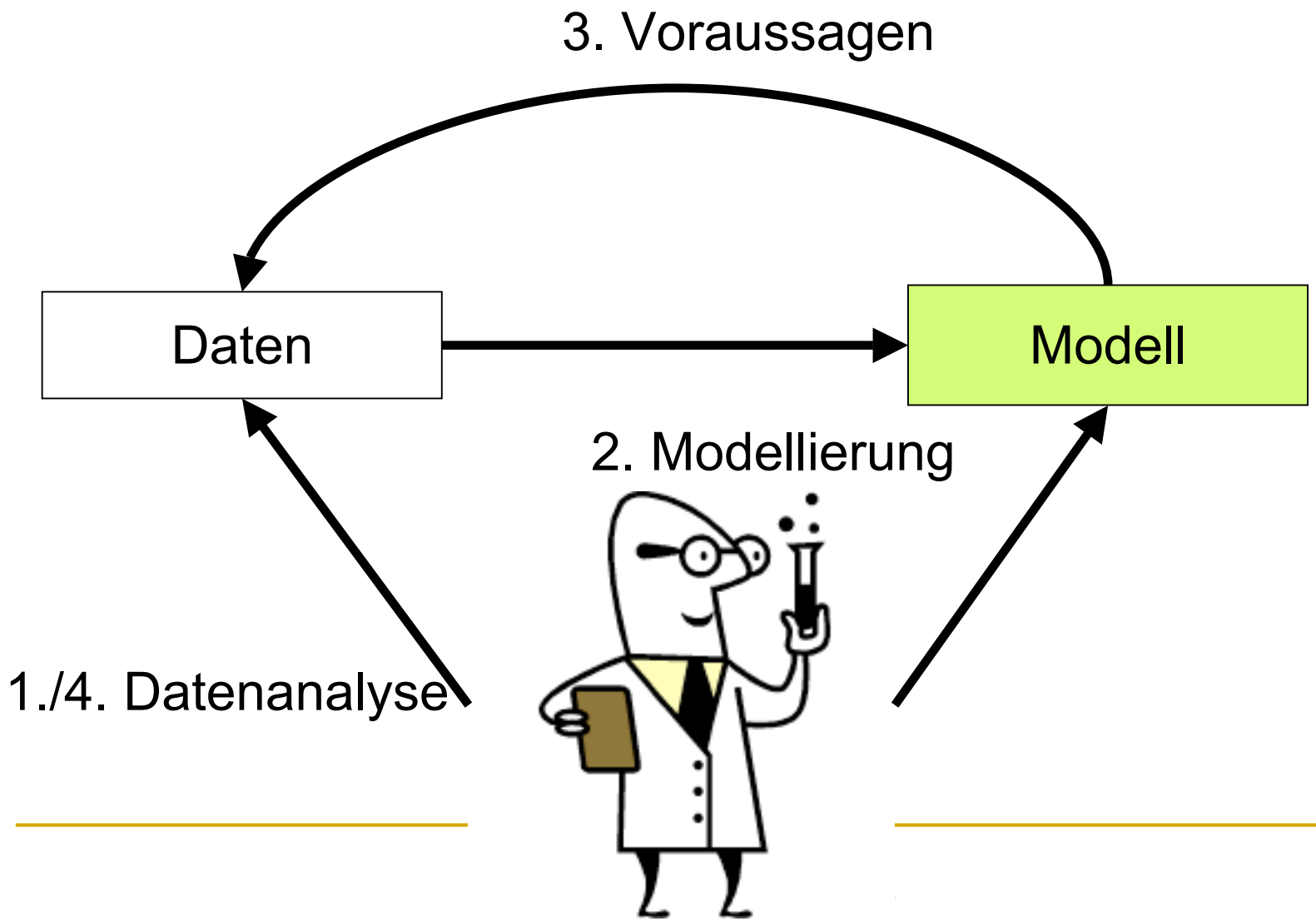


Modellierung der Adjektiverkennung

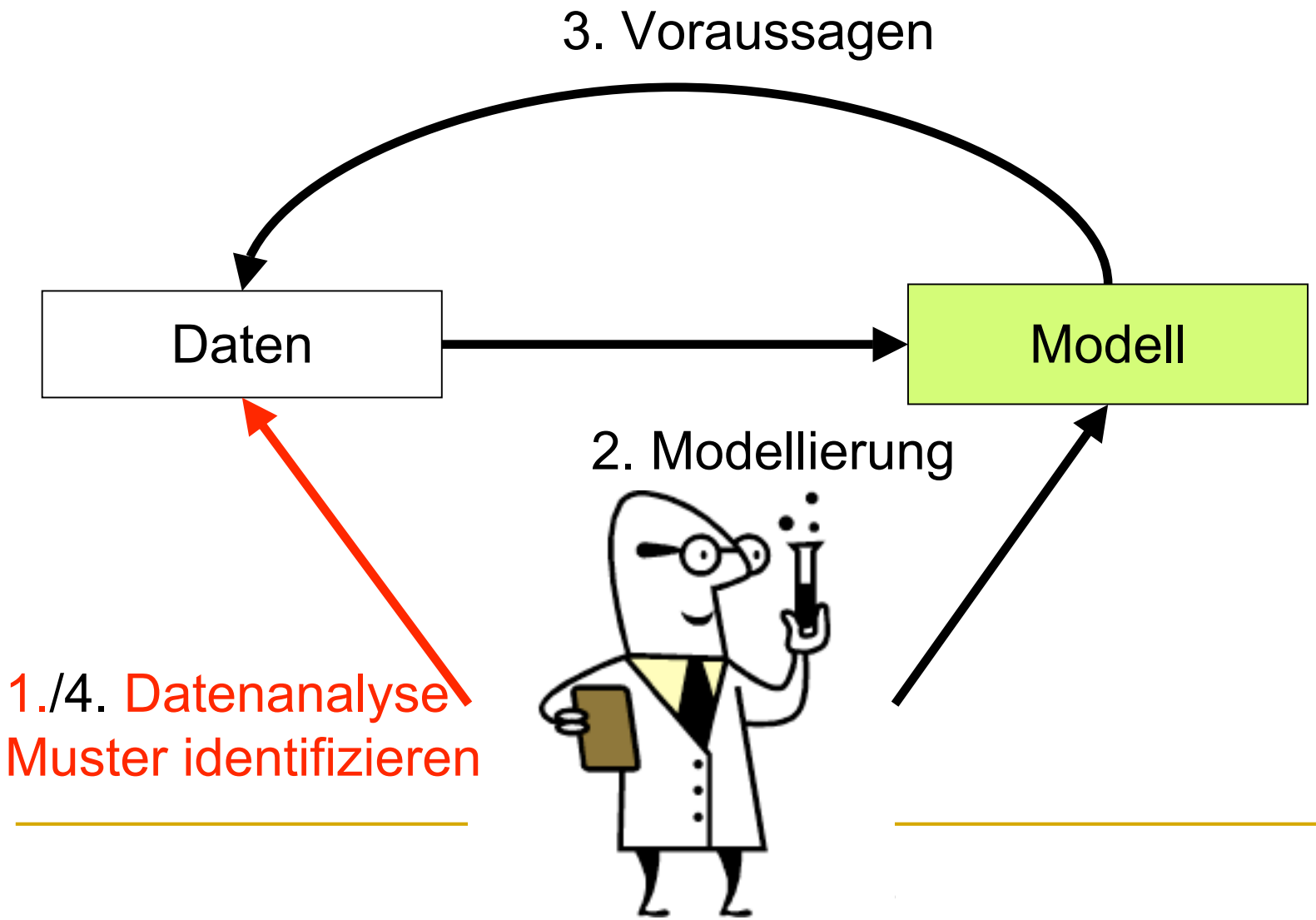
- **Lexikon**basiertes Modell
  - Prüfe, ob Wort in Adjektivliste ist
    - Problem: Adjektive sind offene Wortklasse

- **Muster**basierte Modelle
  - Option 1: Schreibe explizite Regeln der Form
    `Wenn <Muster>, dann Adj/NAdj`
    - **Symbolisches** Modell
  - Option 2: Lerne Abhängigkeiten zwischen Mustern und "Adjektivheit" aus Korpus
    - **Statistisches** Modell

Empirische Modellierung



Empirische Modellierung



Daten

- Meistens zwischen 70 und 90% der Daten
- Grundlage der Modellierung (Schritt 1)
- Quelle fuer Freuquenzen
- (Schritt 2)

- Überprüfung des Modells an unabhängigen Daten (Schritt 4)

Datenanlyse: Informative Muster für die Adjektiverkennung

Intuitive Frage: **Woran erkenne ich ein Adjektiv**?

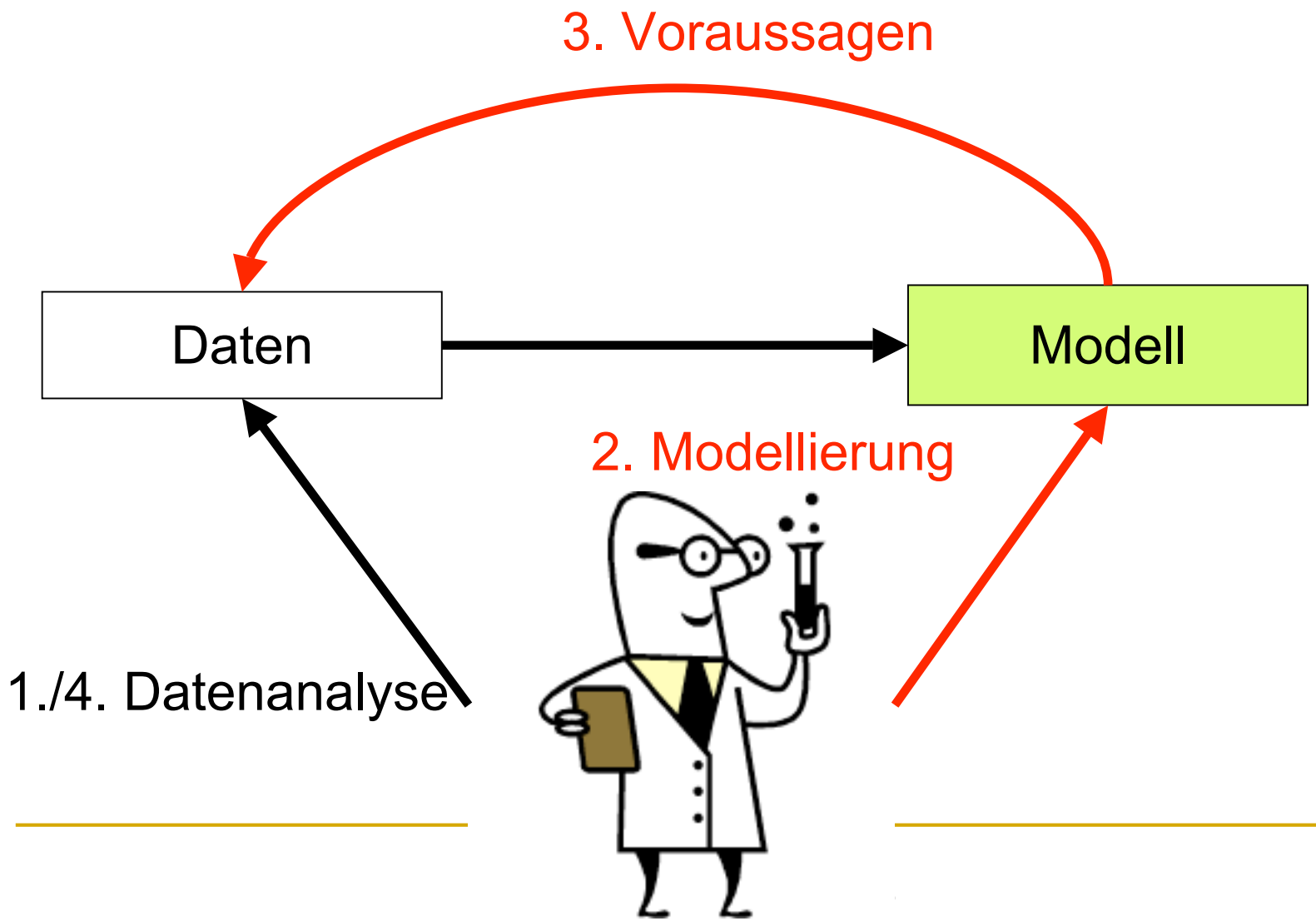
Ich moechte Ihnen fuer Ihren Bericht ueber
den siebenten Bericht ueber
staatliche Beihilfen in
der europaeischen Union danken.

Datenanlyse: Informative Muster für die Adjektiverkennung

Intuitive Frage: **Woran erkenne ich ein Adjektiv**?

- Nächstes Wort ist großgeschrieben (kapitalisiert)
  - `Der **siebente** Bericht`
- Vorheriges Wort ist Artikel
  - `Der **siebente** Bericht`
- Wort selbst ist nicht kapitalisiert
  - `Der **siebente** Bericht`
- Wort selbst ist kein Artikel

Empirische Modellierung

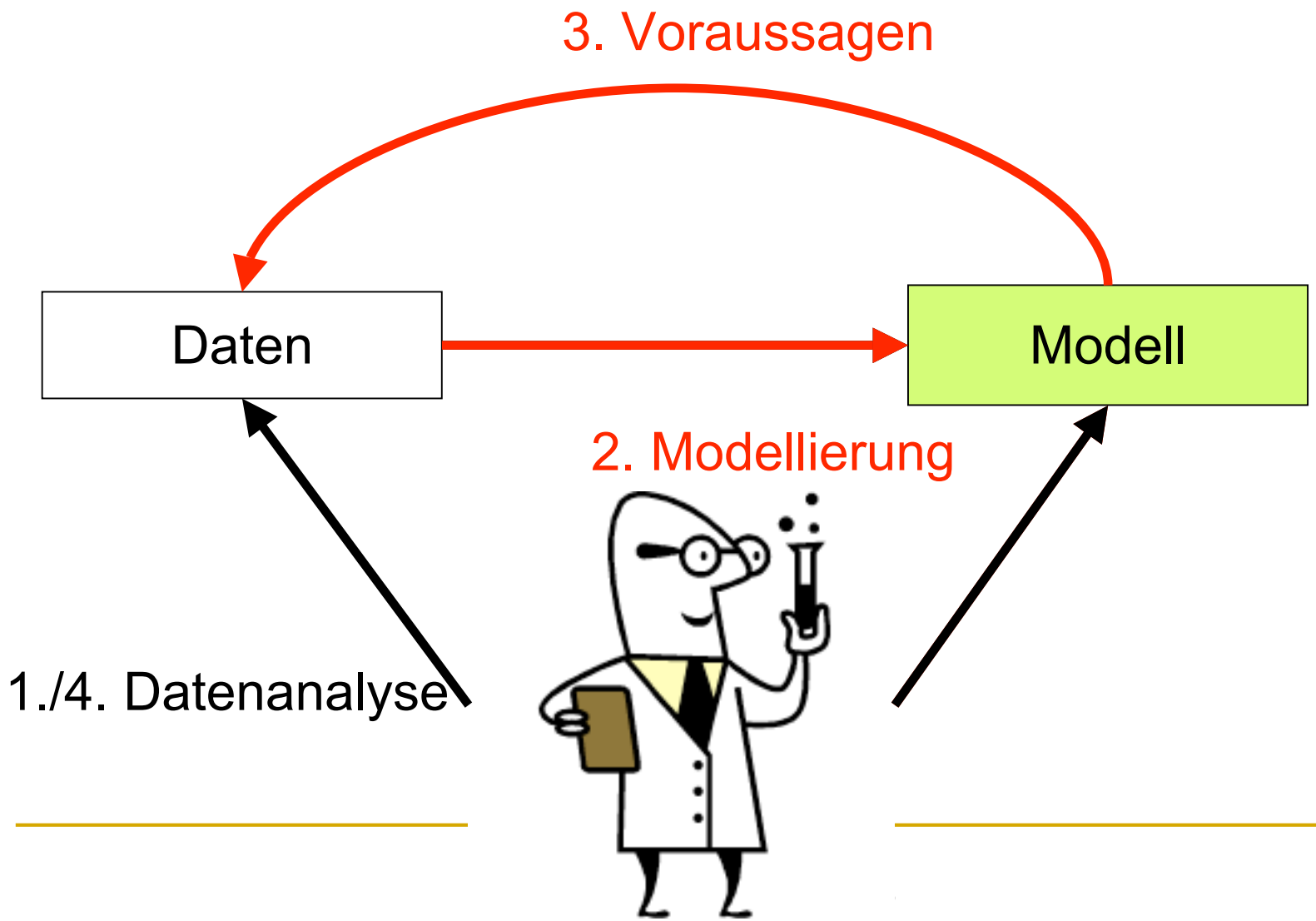


Symbolische Regeln

- Welche expliziten Regeln können wir aufstellen?
  - Nächstes Wort großgeschrieben ⇒ Adj
    - Der Mann (Korrektheitsproblem)
  - Nächstes Wort kapitalisiert und Wort kein Artikel ⇒Adj
    - Ich sehe Peter (Korrektheitsproblem)
  - Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel ⇒ Adj
    - Große Bedenken (Vollständigkeitsproblem)

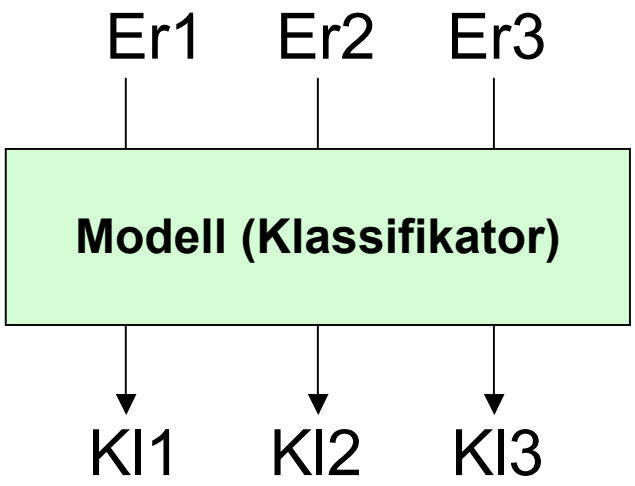
Es ist schwer, explizite Regeln zu schreiben,
die sowohl korrekt als auch vollständig sind

Empirische Modellierung



Adjektiverkennung als Klassifikation

- **Alternative Idee:** Zusammenhang zwischen Mustern und “Adjektivheit” automatisch (maschinell) lernen!
- **Klassifikationsaufgabe**: jedem Eingabe-Ereignis wird eine Klasse (aus bekannter Menge) zugeordnet



Ereignisse

(hier: Worte)

Klassen

(hier: Adj/NAdj)

- Gegenbeispiel: Lemmatisierung

Statistische Klassifikation

- Schritt 1: Modell lernen ("**Training**")
  - Modell ist eine Tabelle mit **Häufigkeiten** von **Ereignissen** und dazugehörigen **Klassen**
    - Häufigkeitsverteilung/ Wahrscheinlichkeitsverteilung

Ereignis	Kl. 1	Kl. 2
Er1	10	20
Er2	2	0
...	...	...

- Schritt 2: Modell anwenden ("**Klassifikation**")
  - Gegeben neues **Ereignis**
    - Schlage in Wahrscheinlichkeitsverteilung nach, welches die häufigste / wahrscheinlichste **Klasse** ist
      - Ereignis 1: Kl. 2
      - Ereignis 2?

NB. Es gibt auch andere Varianten statistischer Klassifikation!

Schritt 1: Training...

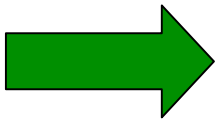
- Wie kommen die “informativen Muster” ins Spiel?
  - Definition der *Ereignisse*
  - *“Featurisierung”*

Was ist ein Ereignis? (I)

- Im einfachstes Fall: ein Wort

Trainingskorpus

Ereignis	Ein kleines Haus
Klasse	Nadj Adj Nadj



Modell

Ereignis	Nadj	Adj
ein	1	0
kleines	0	1
Haus	1	0

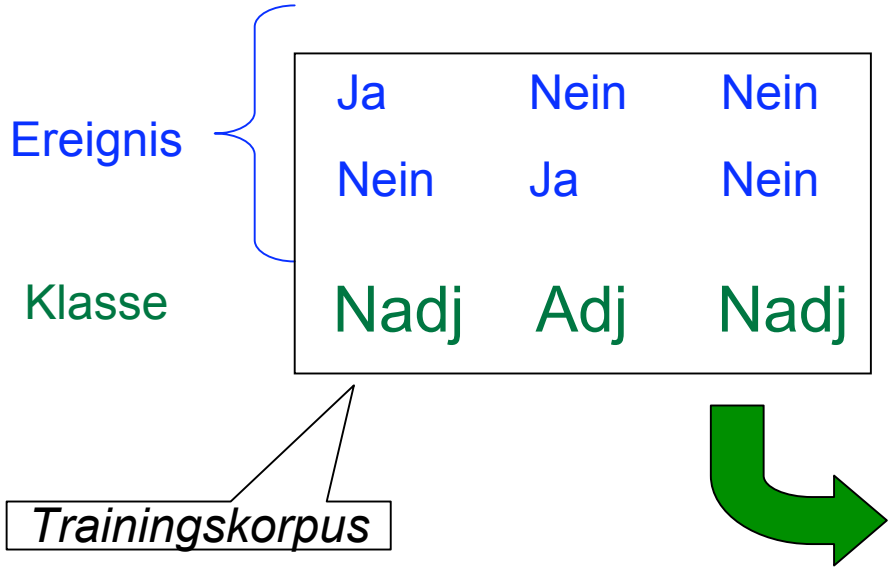
- Keine gute Idee!
  - Klassifikation eines neuen Satzes:
    - Wortliste!

Ereignis	Ein großes Haus
Klasse	Nadj ?? Nadj

Was ist ein Ereignis? (II)

- Besser: Ereignis ist Kombination von **informativen Mustern** (“**Features**”)

Ein kleines Haus



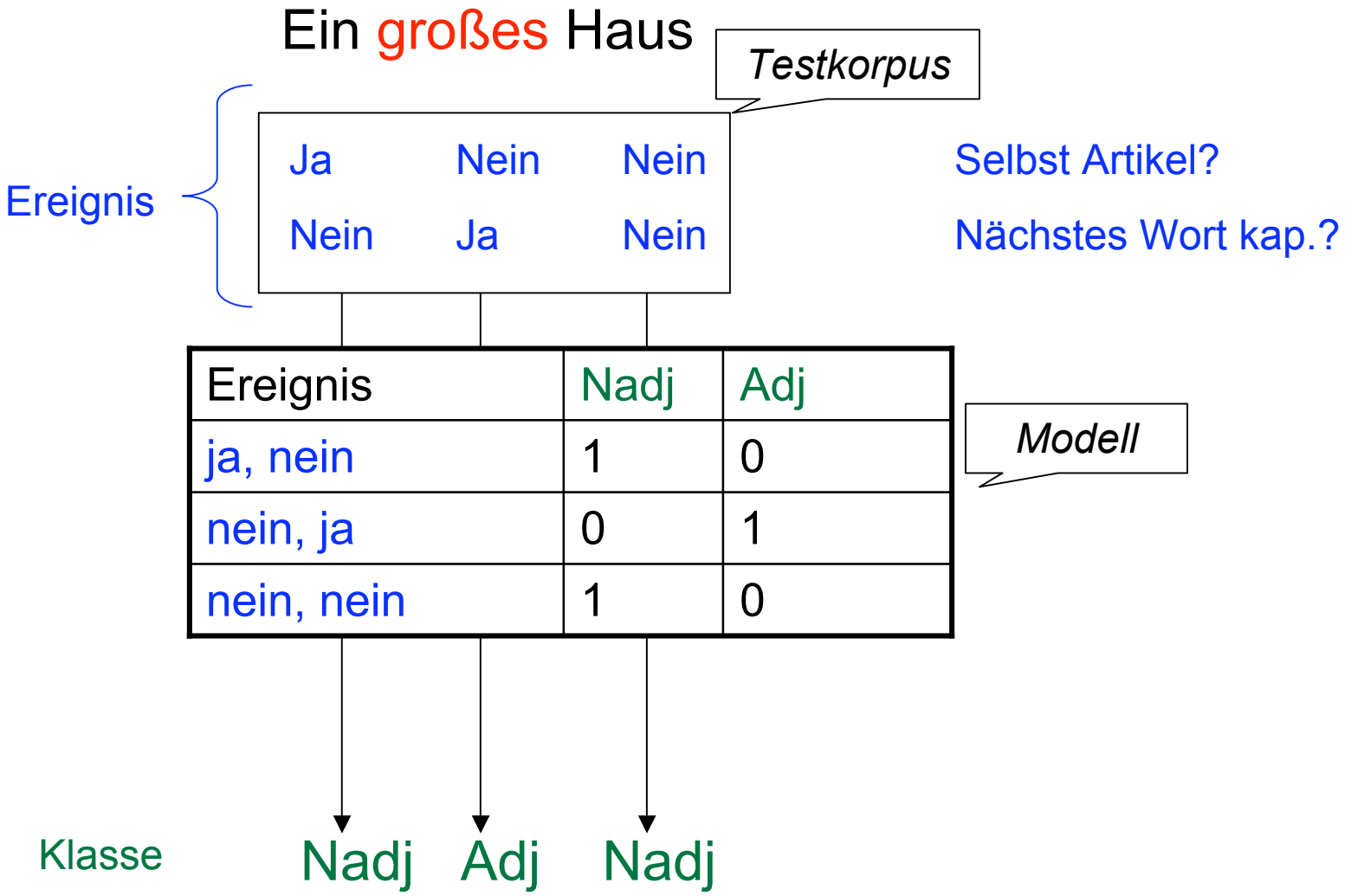
Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

Modell

Schritt 2: Klassifikation

- Dasselbe Spiel:
  - Repräsentation von Ereignissen durch informative Muster (Featurisierung)

Klassifikation eines neuen Satzes



Was "bedeutet" das Modell?

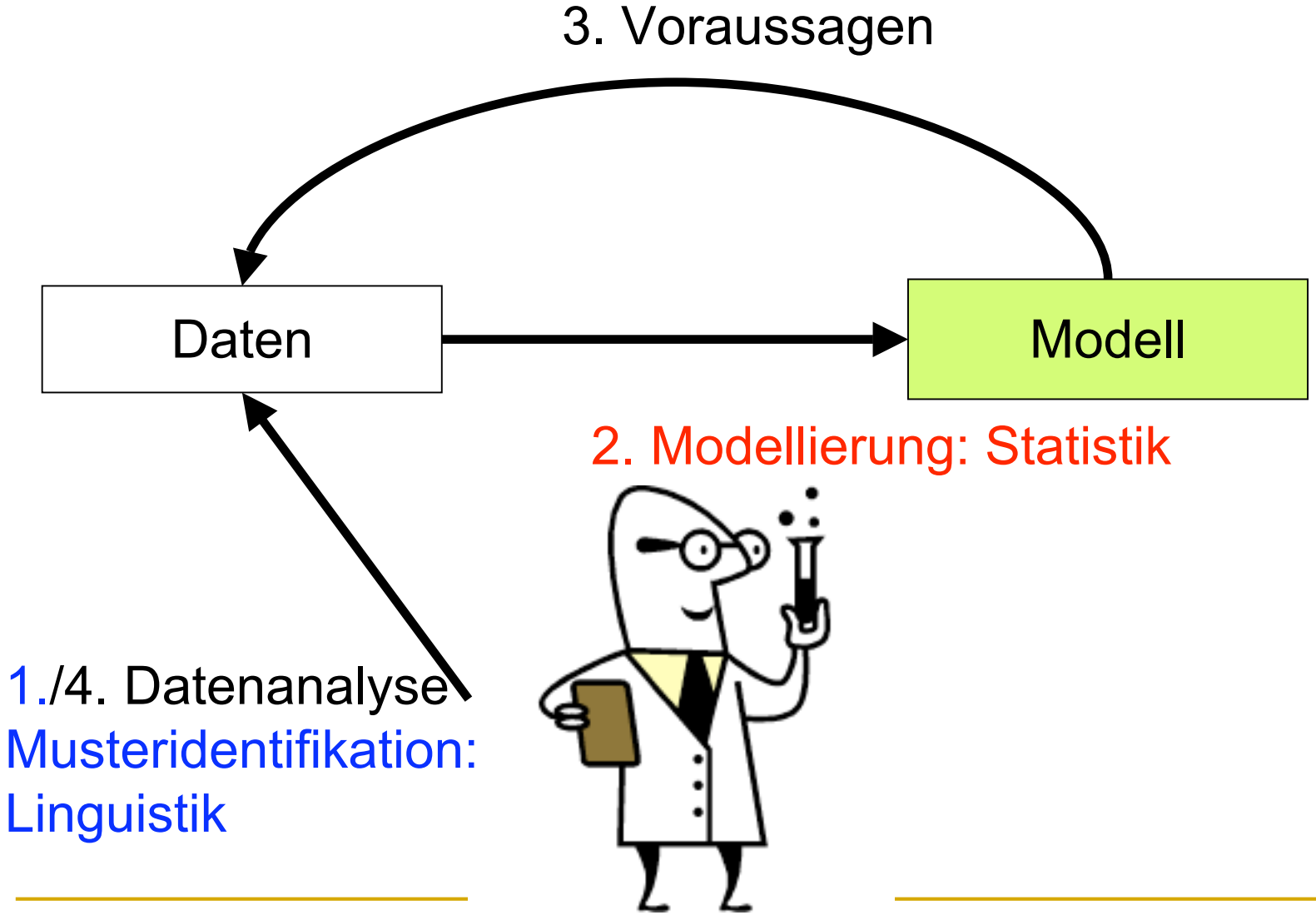
- Features:
  - Selbst Artikel?
  - Nächstes Wort kap.?

Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

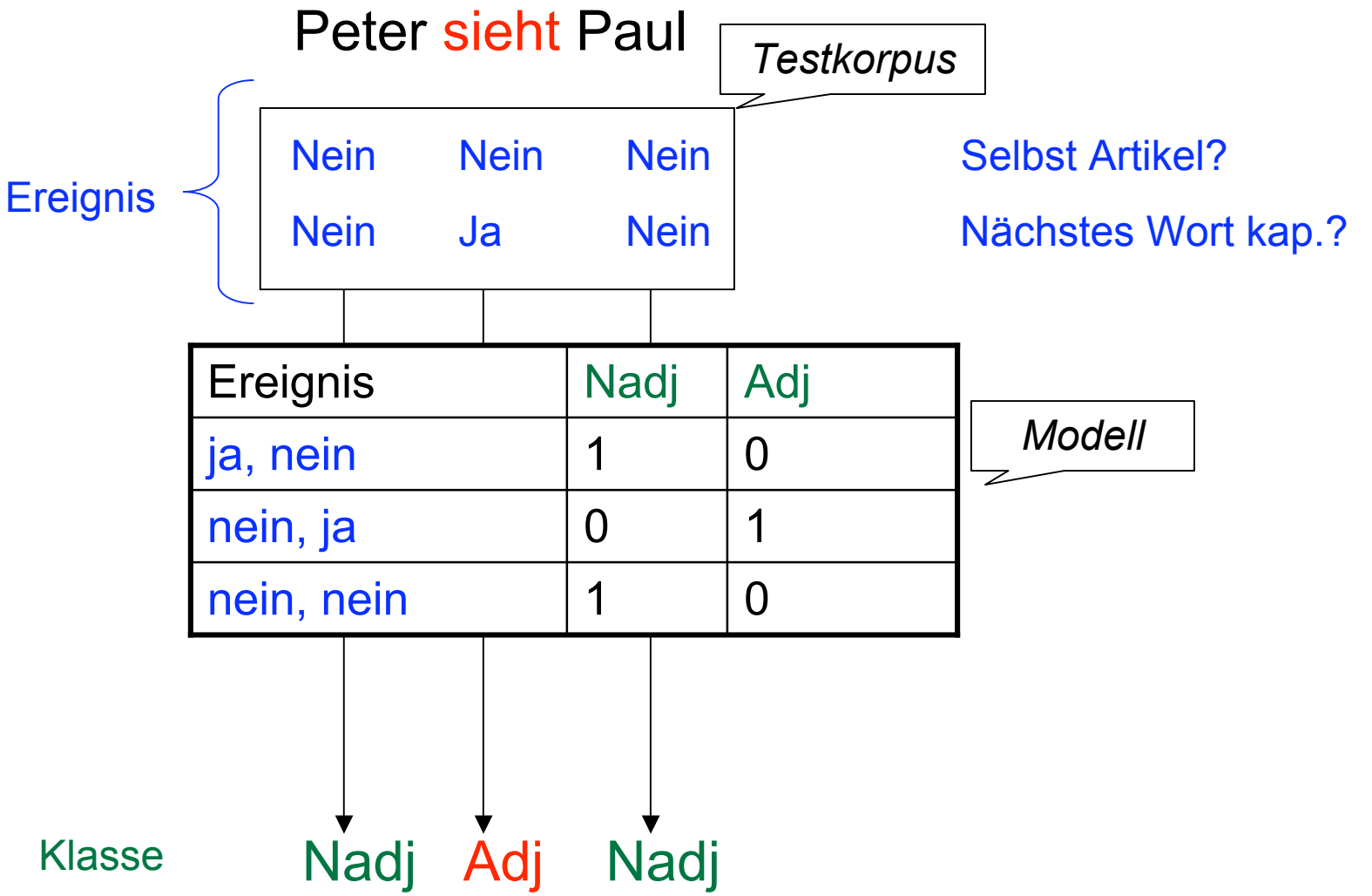
- Zeile 1: Artikel, nächstes Wort nicht kapitalisiert: Kein Adjektiv
- Zeile 2: Kein Artikel, nächstes Wort kapitalisiert: Adjektiv
- Zeile 3: Kein Artikel, nächstes Wort nicht kap.: Kein Adjektiv

- Also doch Regeln: nur eben automatisch gelernte...

Arbeitsteilung



Fehler


Peter sieht Paul
Testkorpus
Ereignis
Nein Nein Nein
Nein Ja Nein
Selbst Artikel?
Nächstes Wort kap.?
Ereignis | Nadj | Adj
ja, nein | 1 | 0
nein, ja | 0 | 1
nein, nein | 1 | 0
Modell
Klasse
Nadj Adj Nadj

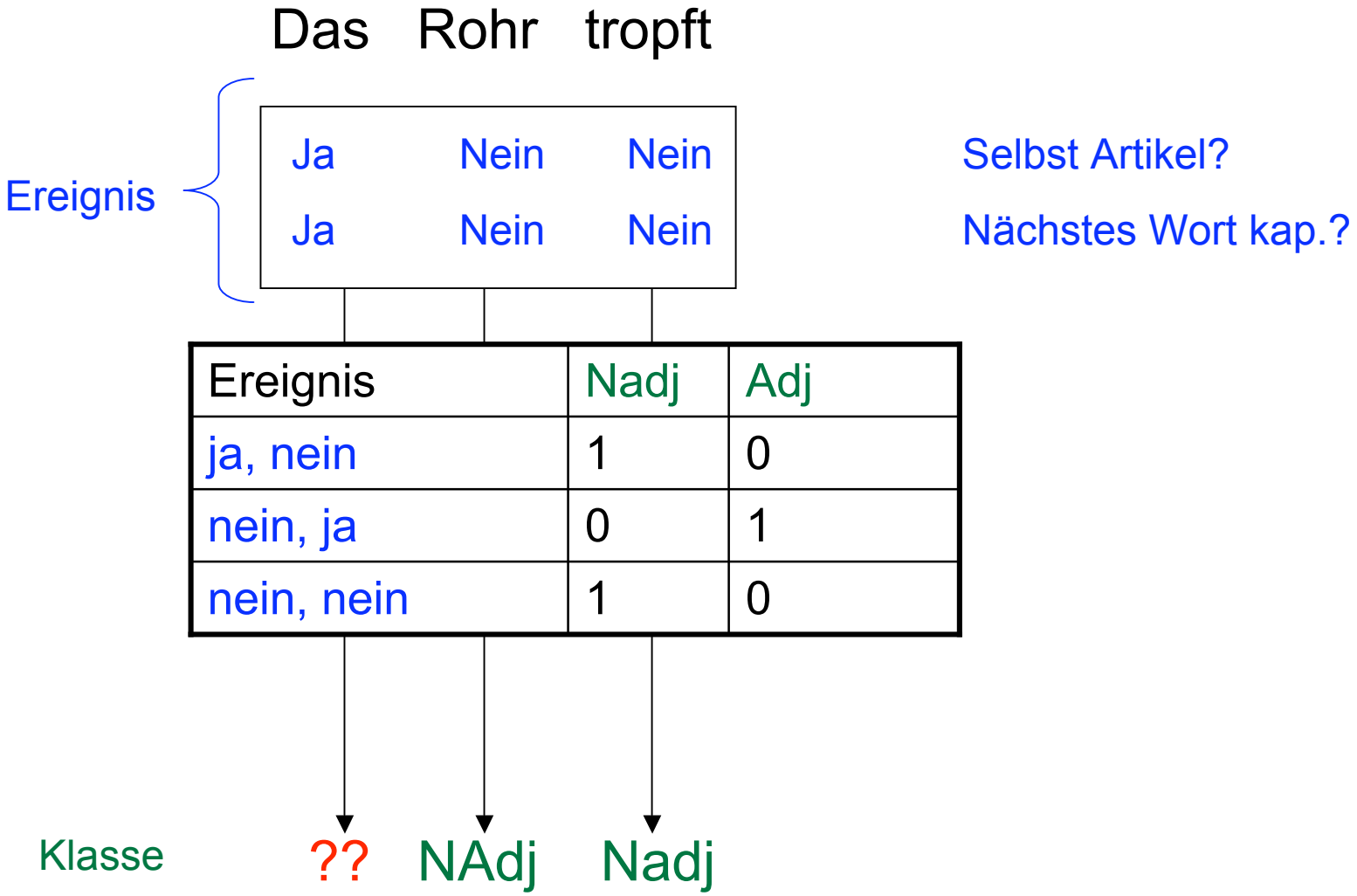
Intelligentere Modelle = mehr Features

- Welche weiteren Muster könnte man ausnutzen, um Adjektive zu identifizieren?
  - Wortendungen (morphologische Information)
    - Attributive Adjektive können nicht auf -t enden wie “fehlt”
  - Gradpartikel stehen fast immer vor Adjektiven
    - “sehr”, “besonders”, ...
  - Kombination von existierenden Features
    - Voriges Wort Artikel UND selbst nicht kapitalisiert
  - ...
- Wieso verwendet man nicht einfach alle Features, die einem einfallen?

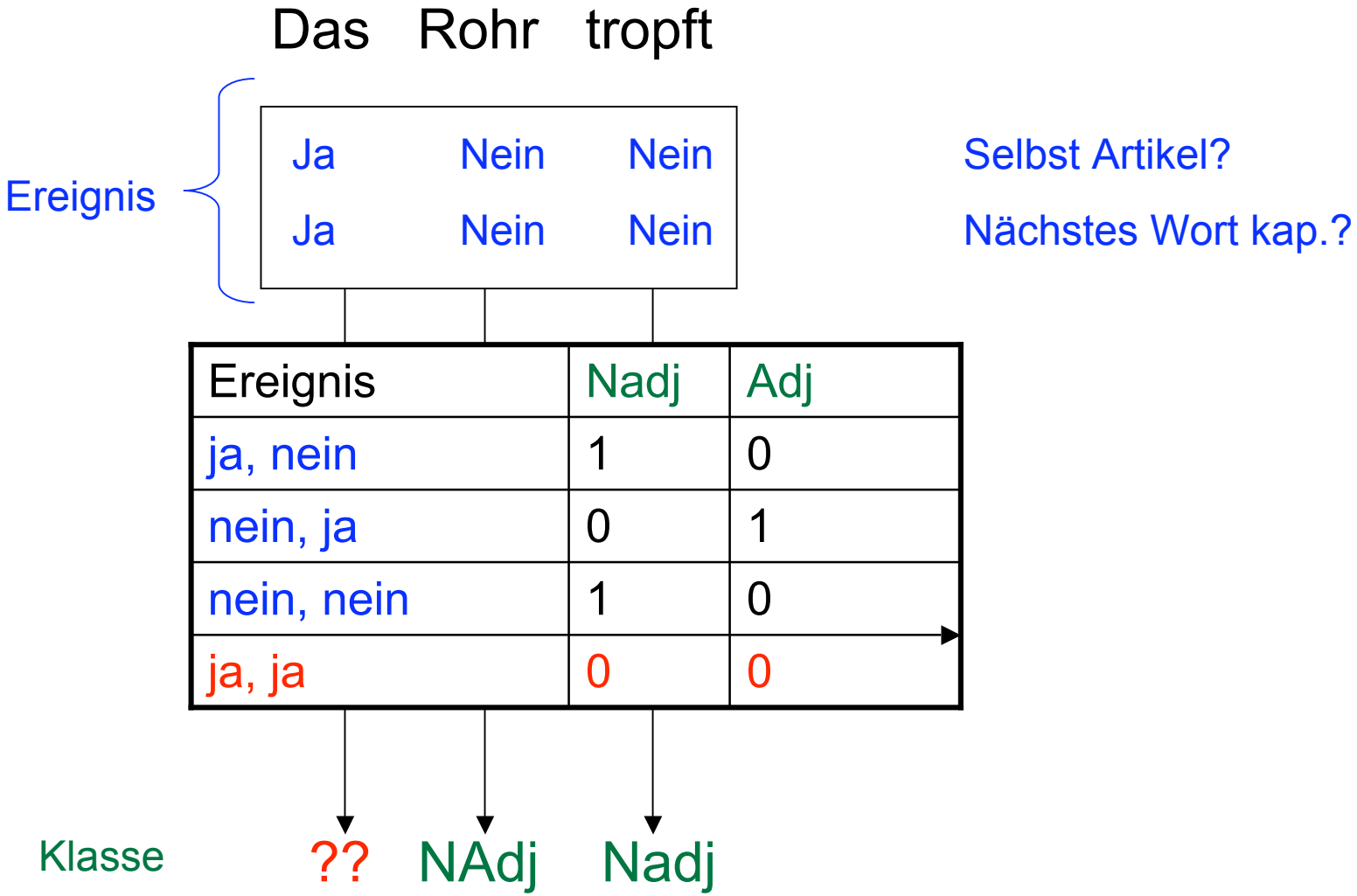
Größe des Ereignisraumes

- Wieviele Zeilen hat die Modell (Tabelle)?
  - Anzahl möglicher verschiedener Ereignisse
    - Produkt der Anzahl möglicher Werte aller Features
      - Beispiel: Selbst Artikel? x Nächstes Wort kapitalisiert = 2 x 2 = 4
      - Lexikalische Features: Wort alleine > 10.000
- Frequenzen in Trainingskorpus werden auf Zeilen (Ereignisse) verteilt
  - Wenn Trainingskorpus < mögliche Ereignisse, können in den Testdaten ungesehene Ereignisse auftreten: Modell kann keine Vorhersage machen
- Das “Sparse Data”-Problem

Sparse Data: Beispiel



Sparse Data: Beispiel



Das Sparse-Data-Dilemma

- Je mehr Features, desto besser die Datenlage für die Entscheidung
- Je mehr Features, auf desto mehr Ereignisse verteilen sich die Trainingsdaten

- Richtlinien für die Identifikation von Features:
  - Wenige gute Features sind besser als viele mittelmäßige
  - Bevorzuge Features mit wenigen Werten

Weitere Strategien

- Strategie 1: Bessere statistische Modelle
  - Entscheidung nicht direkt aus Frequenzen abschätzen
- Strategie 2: Lerne einfachere Modelle (weniger Features) und komplexere Modelle (mehr Features) und kombiniere

Exkurs: Interpretation von Präpositionalphrasen (PPs)

Friedrich sieht den Mann **mit dem Fernrohr**

- Aufgabe: Binäre Klassifikation
  - Klasse 1: „VP“. Das Sehen passiert mit dem Fernrohr
    - Friedrich [$_{VP}$ sieht [$_{NP}$ den Mann] [$_{PP}$ mit dem Fernrohr]]
  - Klasse 2: „NP“. Der Mann hat das Fernrohr:
    - Friedrich [$_{VP}$ sieht [$_{NP}$ den Mann [$_{PP}$ mit dem Fernrohr]]]
- Welche Muster helfen bei dieser Entscheidung?

Features für PP-Anbindung (I)

Friedrich sieht den Astronomen mit dem Fernrohr (NP)

Friedrich sieht den Stern mit dem Fernrohr (VP)

- Feature 1: Kopf der NP (n_1)

Features für PP-Anbindung (II)

Friedrich sieht den Astronomen mit der Gitarre (NP)

Friedrich stört den Astronomen mit der Gitarre (VP)

- Feature 2: Kopf der VP (v)

Features für PP-Anbindung (III)

Friedrich stört den Mann mit der Sonnenbrille (NP)

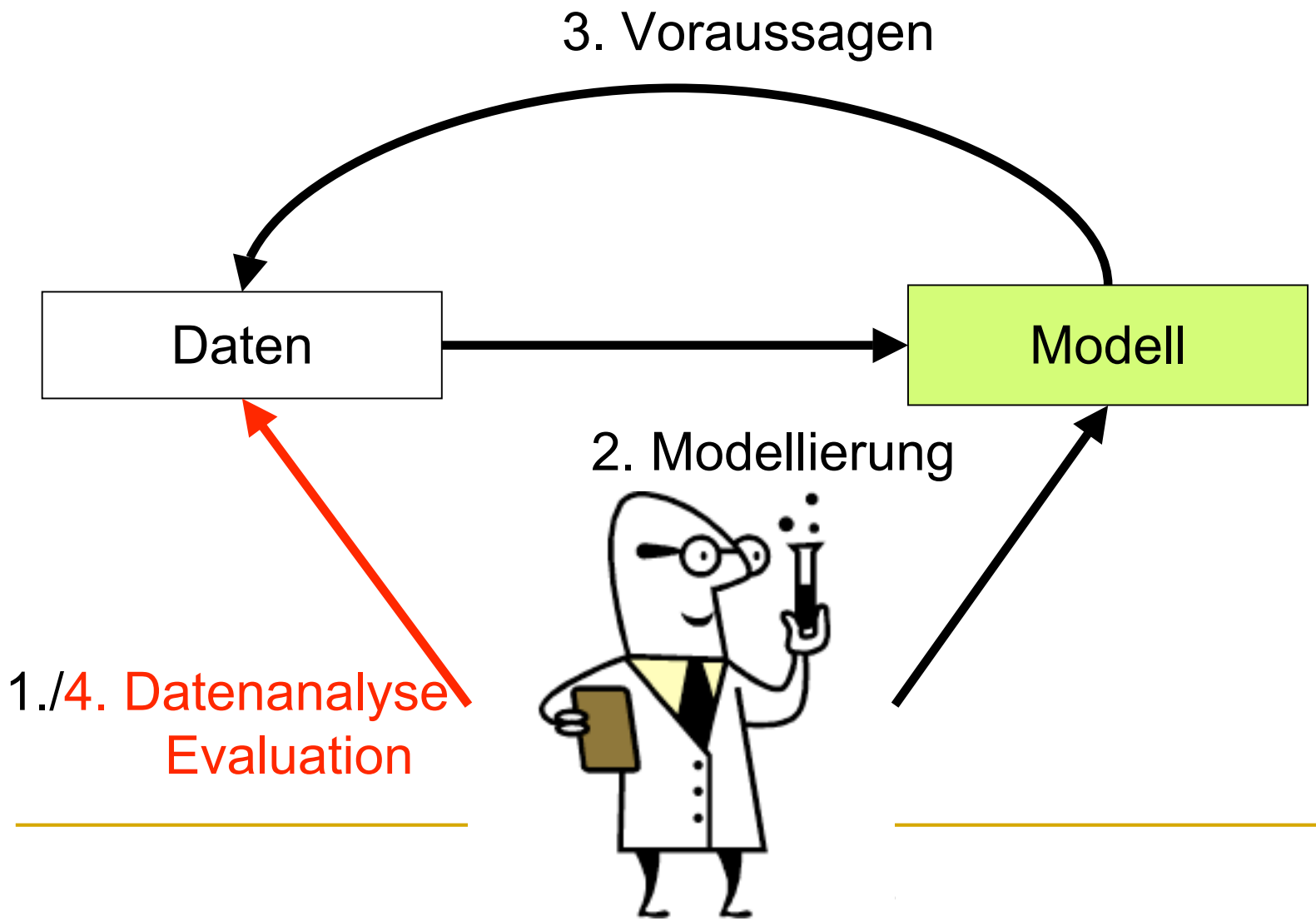
Friedrich stört den Mann mit der Gitarre (VP)

- Feature 3: Kopf der NP in der PP (n_2)

Entscheidung bei PP-Anbindung

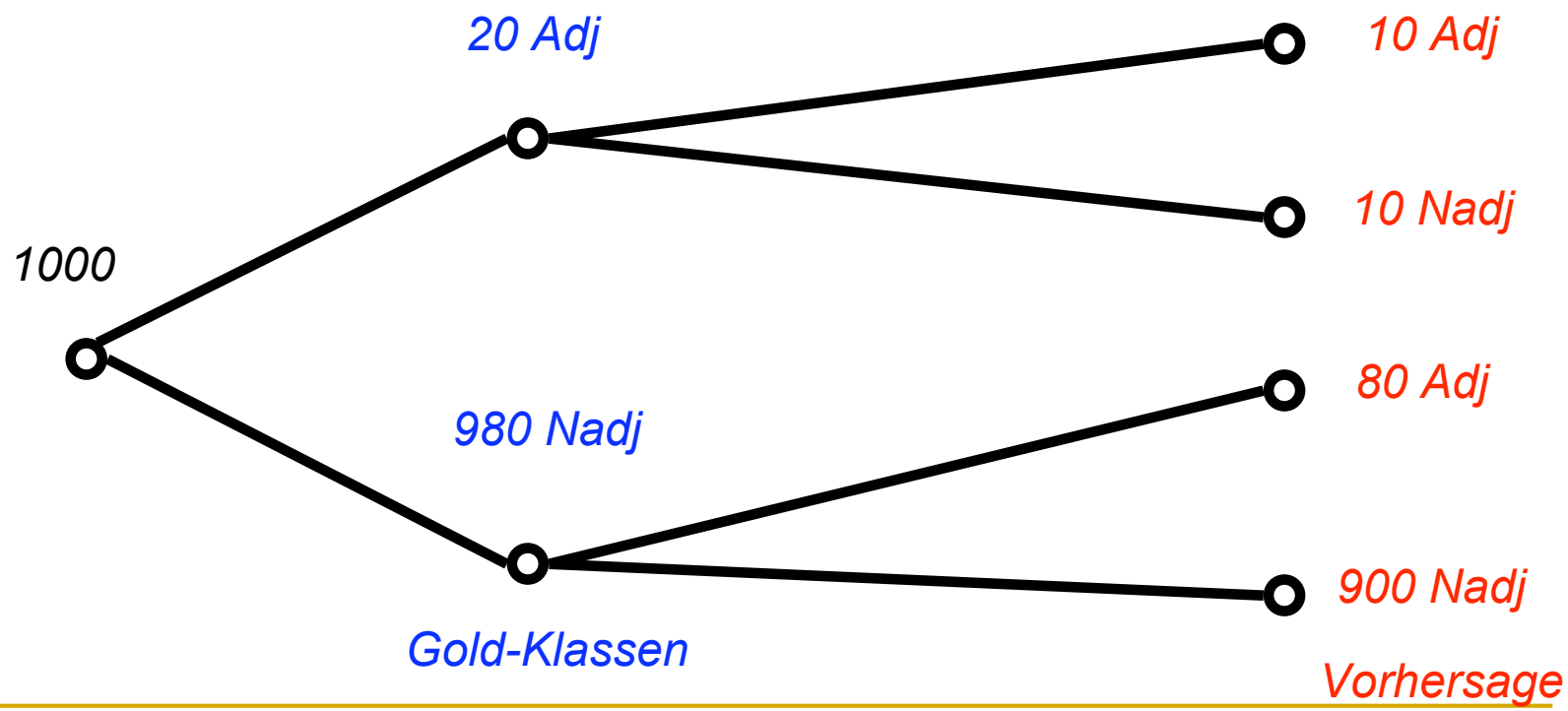
- 95% der Ereignisse (v,n_1,n_2) in den Testdaten kommen nicht in den Trainingsdaten vor
- Lösung: „back-off“
  - Wenn du komplexen Modell vertrauen kannst, nimm es; sonst nimm einfacheres Modell
- Beispiel: Klassifiziere (stört,Mann,Gitarre)
  - Wenn (stört,Mann,Gitarre) gesehen, klassifiziere Tripel
  - Sonst versuche Paare (stört,Mann), (stört, Gitarre), (Mann,Gitarre)
  - Wenn auch nicht gesehen, versuche (stört), (Mann), (Gitarre)

Empirische Modellierung



Beispielevaluation

- 1000 Instanzen
- Aufgabe: Klassifikation nach Adjektiv / Nicht-Adjektiv



Accuracy / Error

- Einfachstmögliche Evaluation
- Accuracy = (# richtige Instanzen) / (# alle Instanzen)

$$= \frac{\text{(\# Instanzen: Vorhersage = Gold-Klasse)}}{\text{(\# alle Instanzen)}}$$

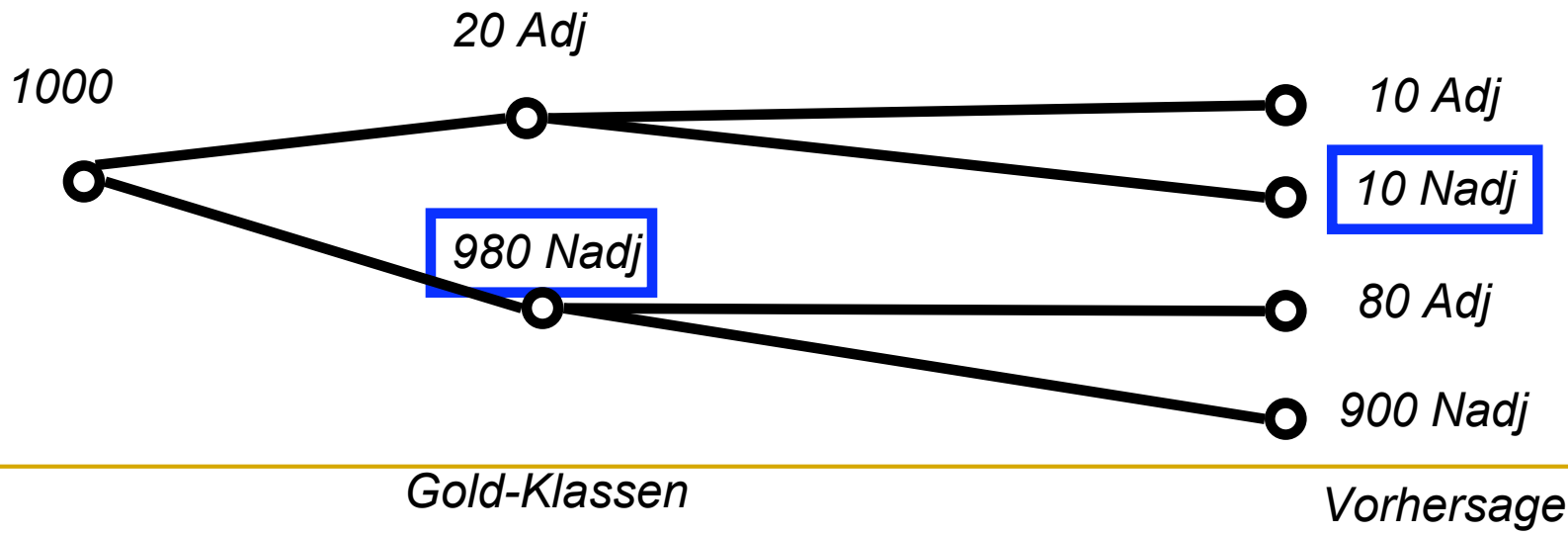
  - Hier: (10 + 900) / 1000 = 910 / 1000 = 91%
- Error = (1 - Accuracy)
  - Hier: 9%

Probleme

- Problem 1: Macht keinen Unterschied zwischen (Gold-)Klassen (Adj vs. Nadj)
  - Wieso unterscheiden?
    - In der Computerlinguistik sind die interessanten Klassen oft klein
    - In Gesamtevaluation geht Qualität der kleinen Klasse unter
  - Klassenspezifische Accuracy / Error
    - 10 / 20 korrekt für Adj: 50%
    - 900 / 980 korrekt für Nadj: 91.8%

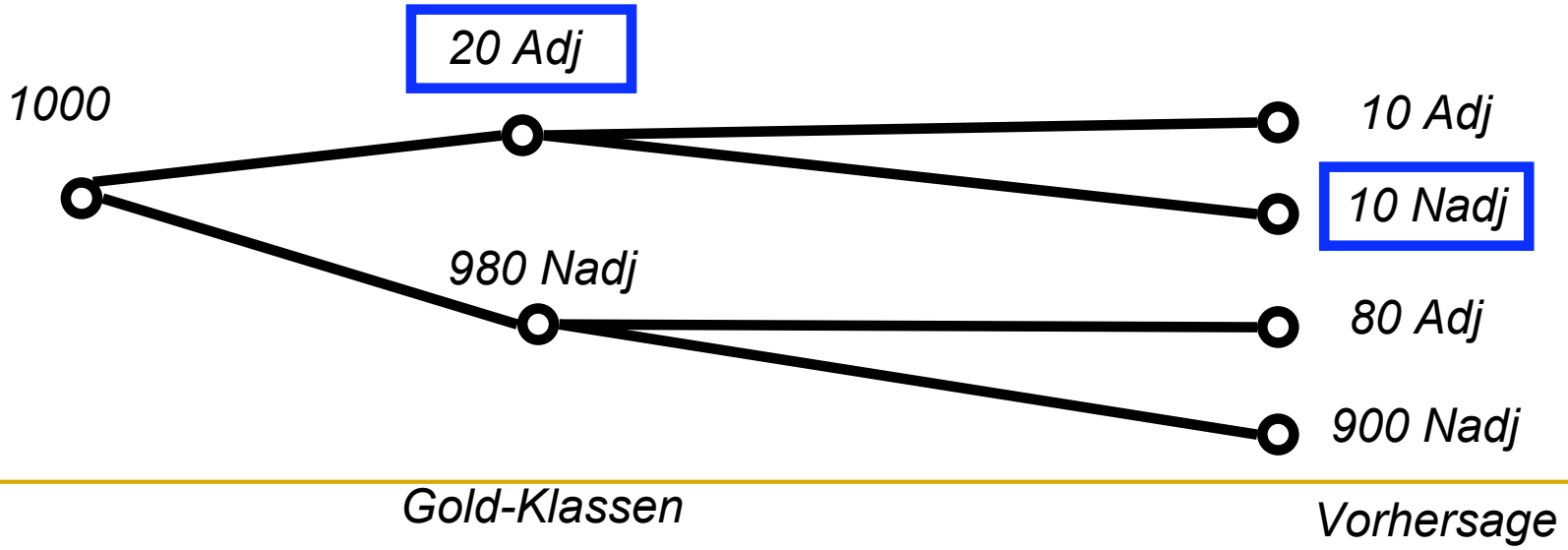
Klassenspezifische Fehler

- Bei *klassenspezifischer* Betrachtung unterscheidet man zwei Fehlerarten:
  - Korrektheitsfehler von X (“false positive”):
    - Eine Instanz ist kein X, wird aber vom Modell als X klassifiziert



Klassenspezifische Fehler

- Bei *klassenspezifischer* Betrachtung unterscheidet man zwei Fehlerarten:
  - Korrektheitsfehler von X (“false positive”):
    - Eine Instanz ist kein X, wird aber vom Modell als X klassifiziert
  - Vollständigkeitsfehler von X (“false negative”):
    - Eine Instanz ist ein X, wird aber vom Modell nicht als X klassifiziert



Klassenspezifische Fehler

- Bei *klassenspezifischer* Betrachtung unterscheidet man verschiedene Fehler:
  - Korrektheitsfehler von X (“false positive”):
    - Eine Instanz ist kein X, wird aber vom Modell als X klassifiziert
  - Vollständigkeitsfehler von X (“false negative”):
    - Eine Instanz ist ein X, wird aber vom Modell nicht als X klassifiziert
- Wieso unterscheiden?
  - Fehler können unterschiedlich wichtig sein
    - Erkennung von Adjektiven
      - Korrektheitsfehler der Klasse Adj sind meist schlimmer als Vollständigkeitsfehler (Weiterverarbeitung)
    - Qualitätskontrolle von Hardware-Komponenten
      - Vollständigkeitsfehler der Klasse “defekt” schlimm

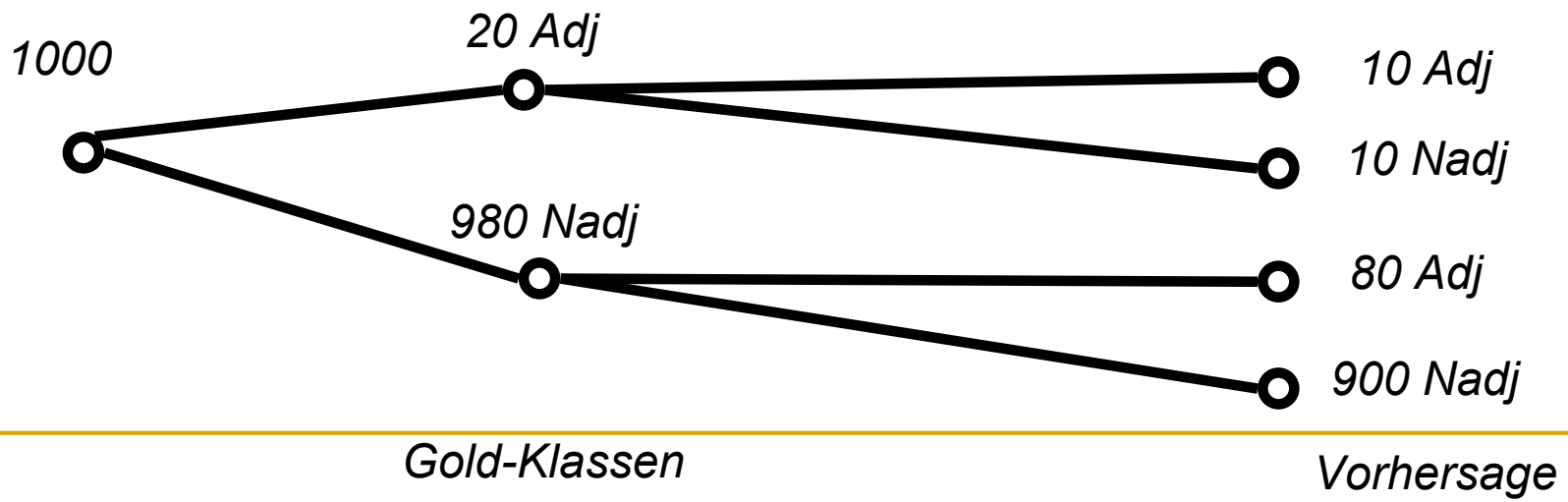
Konfusionsmatrix

	Echtes X	Echtes nicht-X
Als X klassifiziert	gut	schlecht (Korr.fehler)
Als nicht-X klass.	schlecht (Vollst.fehler)	gut

- Methode, um klassenspezifische Frequenzen für verschiedene Fehlerarten aufzuschreiben

Konfusionsmatrix für Klasse Adj

	Echtes X	Echtes nicht-X
Als X klassifiziert	10	80
Als nicht-X klass.	10	900



Recall

	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Welcher Anteil der echten X wurde als X klassifiziert? (Vollständigkeit)
- Werte zwischen 0 und 1 (höher = besser)

Recall für Klasse Adj

	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Hier: 10/(10+10) = 0.5
- Interpretation: Die Hälfte aller echten Adjektive wurde durch das Modell richtig erkannt

Precision

	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Welcher Anteil der als X klassifizierten Instanzen ist wirklich ein X? (Korrektheit)
- Werte zwischen 0 und 1 (höher = besser)

Precision für Klasse Adj

	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Hier: 10 / (10+80) = 11%
- Interpretation: Wenn das Modell behauptet, eine Instanz sei ein Adj, ist das nur 11% der Fälle wahr

Precision und Recall

- Precision und Recall können i.A. nur zusammen betrachtet werden
  - Hohe Precision, hoher Recall: fast perfekte Klassifikation
  - Niedrige Precision, niedriger Recall: sehr schlechte Klassifikation
  - Hohe Precision, niedriger Recall: “vorsichtiges Modell”
    - Findet nicht alle Instanzen von X
    - Klassifiziert fast keine Nicht-Xe als X
  - Niedrige Precision, hoher Recall: “mutiges Modell”
    - Findet fast alle Instanzen von X
    - Klassifiziert auch Nicht-Xe als X

Extremfälle..und die Kombination

- Extremfälle
  - Modell klassifiziert alles als X
    - Recall 100%, Precision sehr niedrig
  - Modell klassifiziert nichts als X
    - Recall 0%, Precision nicht definiert (0/0)

- F-Score: Kombination aus P und R:
  - Ein Maß für „Gesamtgüte“ der Klassifikation
    - Werte zw. 0 und 1 (höher = besser)
    - Bevorzugt „true positives“

$$F = \frac{2PR}{P+R}$$

F-Score für Klasse Adj

	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

- Precision: 10 / (10+80) = 0.11
- Recall: 10 / (10+10) = 0.5
- F-Score: (2 * 0.5 * 0.11) / (0.5 + 0.11) = 0.18