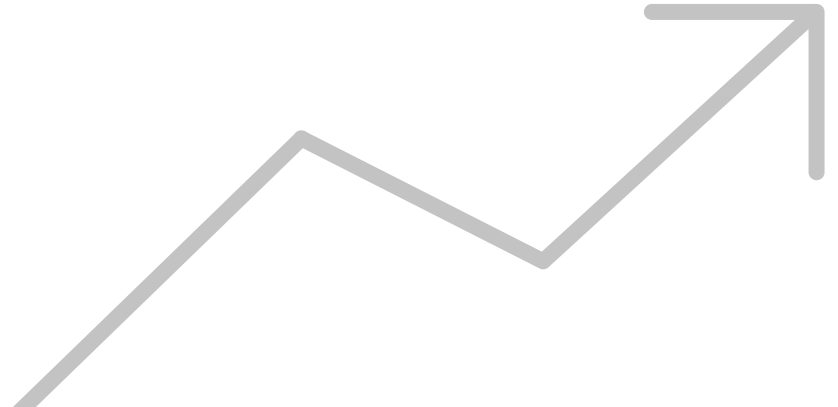
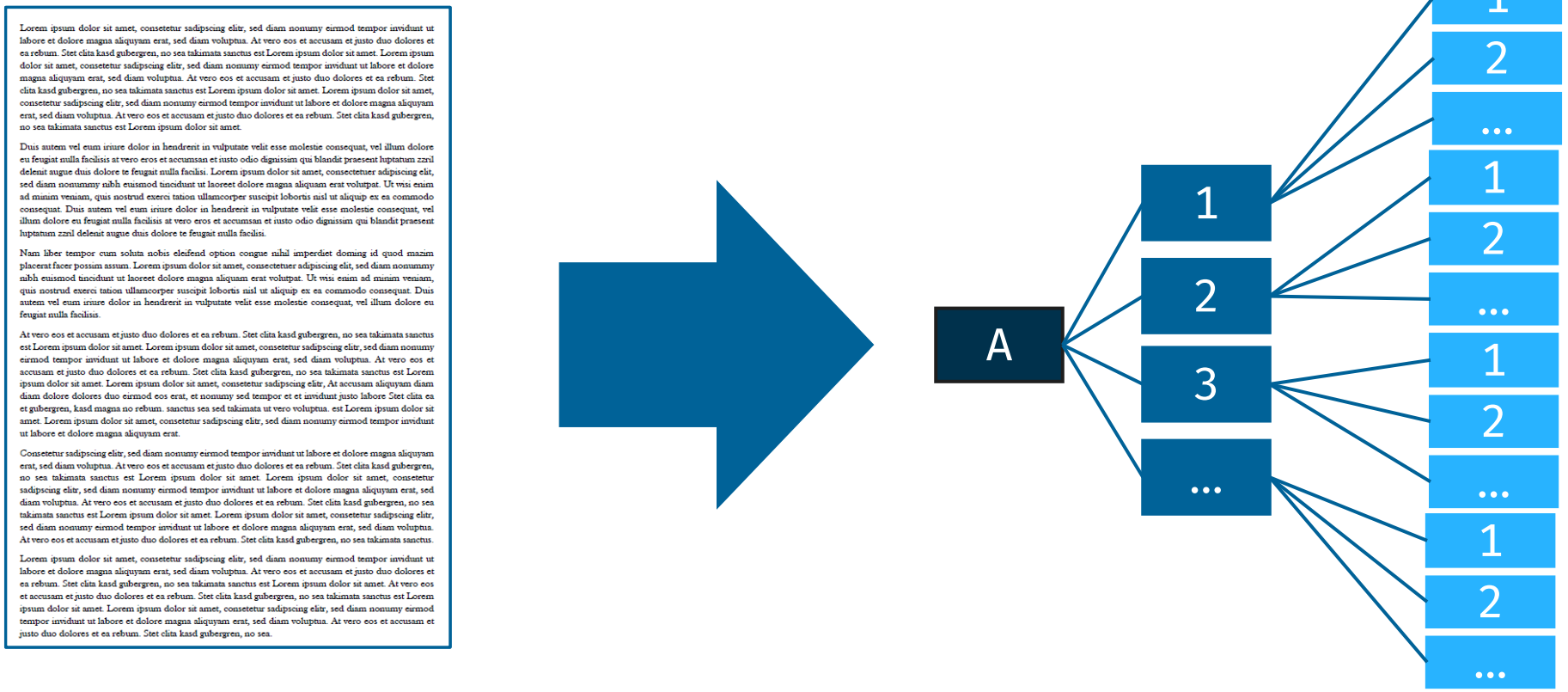
DESTATIS
Statistisches Bundesamt

# Textklassifikation

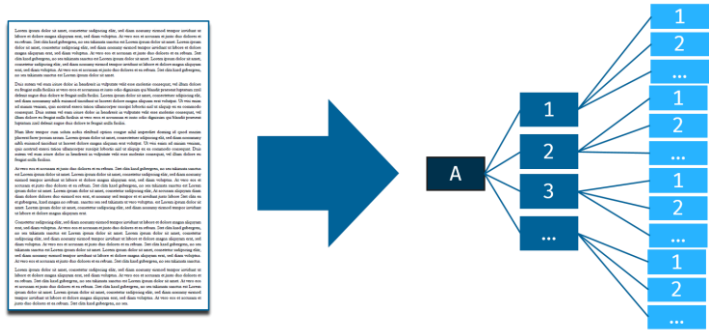
## Maschinelles Lernen zur Klassifikation von freien Texteingaben in der amtlichen Statistik



# Vom Text zur Klasse



# Vom Text zur Klasse

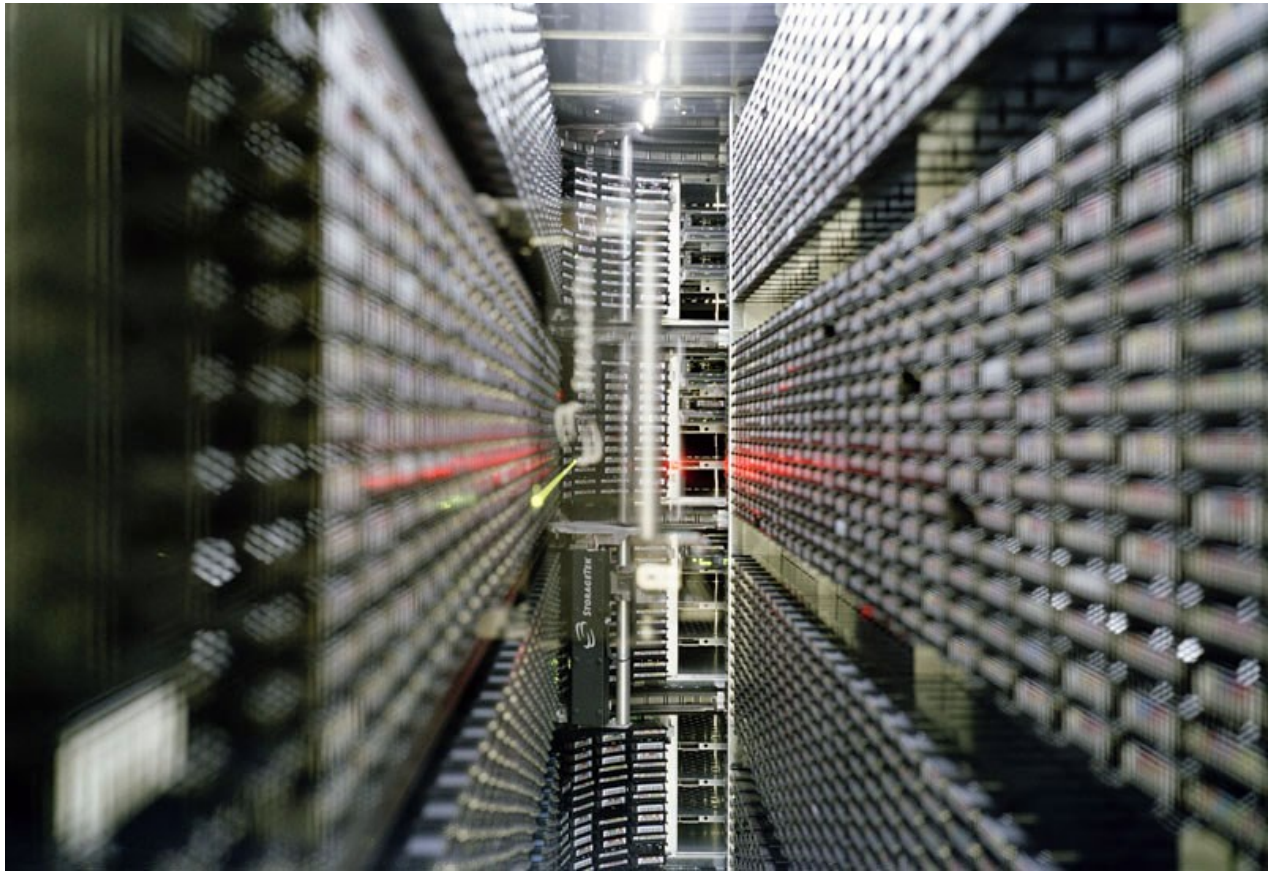


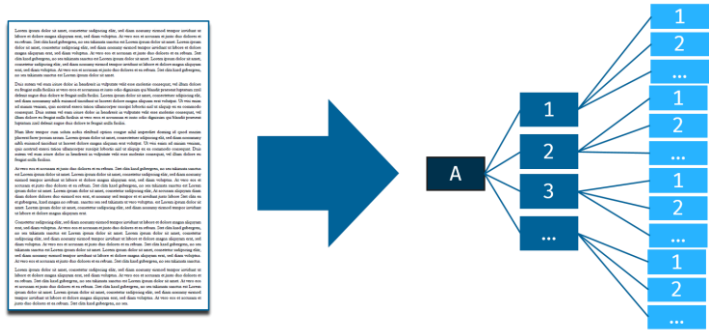

zeitintensiv

personalintensiv

Quellenhinweise: Statistisches Bundesamt https://www.destatis.de/DE/Ueber-uns/Geschichte/_inhalt.html
Fotorechte: © Statistisches Bundesamt

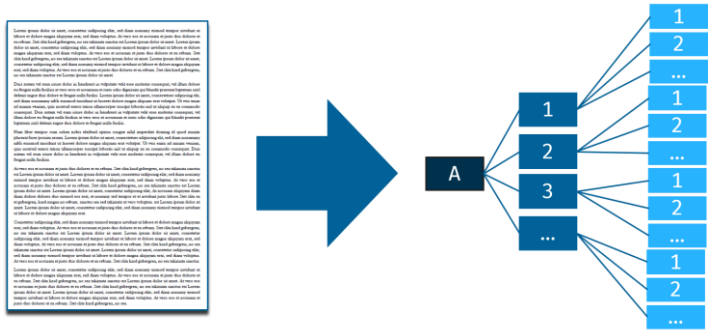
# Vom Text zur Klasse






Verfügbares
Textvolumen steigt

Fotorechte: Wim Klerkx Photography https://www.wimklerkx.nl




# Projekte und Projektideen in der amtlichen Statistik

# Geburtsstaatsklassifikation in den Wanderungsstatistiken

**Trainingsdaten**
rund 6.000.000 Wanderungsfälle jährlich,
teils manuell plausibilisiert

Merkmale Herzugsland, Nationalität,
Geburtsdatum, Religion und *Geburtsort* liegen vor.

**Klassifikation**
Geburtsstaat eines Wanderungsfalles

# Exkurs Freitexte: Beispiel Geburtsort


# COICOP-Klassifikation mit Scannerdaten

## Trainingsdaten

Textbeschreibungen von Produkten + GTIN
regelmäßige Lieferungen großer Datenmengen

## Klassifikation

- Zielgröße: COICOP 10-Steller (> 200 Klassen)
- Mehr als 85% Accuracy bei großen Unterschieden in den Klassen
- Weitere Handelsketten aufnehmen

COICOP: Klassifikation der Verwendungszwecke des Individualverbrauchs (Classification of Individual Consumption by Purpose)
Foto: commons.wikimedia.org

# Klassifikation der Wirtschaftszweige in den Gewerbeanzeigen

Quelle: Deutsche Fotothek

## Trainingsdaten

Freitextangaben von Gewerbeanmeldungen

## Klassifikation

- Zielgröße WZ-4-Steller (ca. 600 Klassen)
- Knapp 60% Accuracy bei großen Unterschieden in den Klassen
- Weitere Hilfsmerkmale könnten die Ergebnisse verbessern

Foto: Bundesarchiv

# Freitexte in den freiwilligen Haushaltserhebungen

## Trainingsdaten

Freitextangaben zu Ausgaben und Aktivitäten

## Klassifikation

- Systematik der Ein- und Ausgaben (ca. 250 Klassen)
- ca. 89% Accuracy bei großen Unterschieden in den Klassen
- Besonderheiten beim Meldeverhalten könnten ins Modell integriert werden

Foto: commons.wikimedia.org

# Klassifikation der Handelsregister- und Insolvenzbekanntmachungen

*Projektidee*

**Trainingsdaten**
frei verfügbare, unstrukturierte und heterogene Bekanntmachungen im Internet

**Ziel**

- Identifikation bestimmter thematischer Sachverhalte für Fachstatistiken
- Automatische Zuordnung zu Einheiten im Unternehmensregister

Foto: snappygoat.com

DESTATIS
Statistisches Bundesamt

# Kontakt

Statistisches Bundesamt
Gustav-Stresemann-Ring 11
65189 Wiesbaden

www.destatis.de

www.destatis.de/kontakt

Ansprechpartner
Joerg Feuerhake
joerg.feuerhake@destatis.de
Telefon +49 611 75-4116