

## Teil I: Kurzbericht

# EDV-TEK: Entstehung, Diffusion und Verwertung von Technologien zur Eindämmung des Klimawandels

Förderkennzeichen: 16DKWN059

## 1 Ursprüngliche Aufgabenstellung und Stand der Forschung

Das Projekt EDV-TEK untersuchte Technologien zur Eindämmung des Klimawandels (TEKs) entlang von vier Dimensionen: Identifikation, Entstehung, Diffusion und Kommerzialisierung. Ziel war es, mithilfe von Verfahren des Deep Learnings – insbesondere neuronaler Netze, großer Sprachmodelle und Graph-neuronaler Netze – neue Erkenntnisse über den Lebenszyklus von TEKs zu gewinnen.

Zum Zeitpunkt des Projektbeginns wurden TEKs überwiegend anhand statischer Patentklassifikationssysteme (CPC Y02, IPC Green Inventory, OECD ENV-TECH) identifiziert. Die systematische Anwendung moderner Verfahren der natürlichen Sprachverarbeitung (NLP) auf Patentvolltexte zur Klassifikation, zur Messung des Neuheitsgrades sowie zur Analyse von Diffusionsmustern und Kommerzialisierungspotenzialen stellte eine Forschungslücke dar. Insbesondere fehlte es an Verfahren, die textuelle und strukturelle Informationen (Zitationsnetzwerke, Autorenetzwerke) gemeinsam nutzen.

## 2 Ablauf des Vorhabens

Das Vorhaben gliederte sich in fünf Arbeitspakete. Zunächst wurde eine umfassende PATSTAT-PostgreSQL-Datenbank mit Volltextdaten des EPO und USPTO aufgebaut (AP 0, nicht Teil der ursprünglichen Vorhabenbeschreibung). Darauf aufbauend wurden in AP 1 verschiedene Klassifikationsverfahren – klassische neuronale Netze, vortrainierte Sprachmodelle und Graph-neuronale Netze – zur Identifikation von TEKs entwickelt und verglichen. In AP 2 wurde der wissenschaftliche Ursprung von TEKs über Patent-Paper-Verknüpfungen (OpenAlex, Reliance on Science) untersucht und der Neuheitsgrad mittels semantischer Distanzmaße bestimmt. AP 3 analysierte die Diffusion von TEKs innerhalb und zwischen CPC-Y02-Technologieklassen sowie auf internationaler Ebene (US ↔ EP). In AP 4 wurde die Kommerzialisierung über ein Matching von Crunchbase- und PATSTAT-Daten untersucht. Die Arbeitspakete 2, 3 und 4 differenzierten dabei jeweils nach dem Neuheitsgrad der Patente. Der Projektfortschritt wich zeitlich um etwa drei bis fünf Monate von der ursprünglichen Planung ab, was durch eine kostenneutrale Laufzeitverlängerung bis zum 30.11.2025 kompensiert wurde.

Gefördert durch:



### **3 Wesentliche Ergebnisse (und ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen)**

Im Rahmen des Projektes wurde eine PATSTAT-PostgreSQL-Datenbank, welche ca. 1,87 Millionen TEK-Patente und zugehörigen Volltextdaten inkludiert, aufgebaut. Der Vergleich der Klassifikationsverfahren in AP 1 zeigte, dass vortrainierte Sprachmodelle (DistilBERT mit 76,75 % Accuracy, DistilRoBERTa mit 76,69 %) die höchste Klassifikationsgenauigkeit erreichten, gefolgt von klassischen neuronalen Netzen (Feed-Forward Network: 74,24 %) und Graph-neuronalen Netzen (FullGNN: 71,44 %). In AP 2 wurde der wissenschaftliche Ursprung von TEKs über Patent-Paper-Verknüpfungen analysiert – das SimpleBaselineModel erreichte 99,93 % AUC bei der Link Prediction – und der Neuheitsgrad von ca. 318.000 Patenten mittels semantischer Distanzmaße nach Shibayama et al. bestimmt. Die Diffusionsanalyse in AP 3 identifizierte über 3,5 Millionen Diffusionsverbindungen und zeigte, dass High-Novelty-Patente durchschnittlich 2,9 Jahre schneller diffundieren als Low-Novelty-Patente. In AP 4 wurden über ein mehrstufiges Matching-Verfahren ca. 1,16 Millionen deduplizierte Firmen-Patent-Verknüpfungen hergestellt und mit 467.532 Funding-Datensätzen für 17.608 Unternehmen verknüpft. High-Novelty-Patente zeigten dabei eine höhere Median-Finanzierung (11 Mio. USD) als Medium- (8,7 Mio.) und Low-Novelty-Patente (8,5 Mio.).

Die im Projekt aufgebaute Dateninfrastruktur, die entwickelten Methoden und die gewonnene Expertise fließen nachhaltig in die Lehre (Masterkurs „Deep Learning for Social Analytics“), in Abschlussarbeiten sowie in die Forschung am Institut für Unternehmertum ein. Darüber hinaus sollen sowohl sämtliche Programmierskripte (über GitHub) als auch die Projektergebnisse und ein umfassender Leitfaden zur Software (über die Projektwebseite) der wissenschaftlichen Community zur Nachnutzung dienen. Insbesondere die Skripte zum Aufbau der Patstat PostgreSQL-Datenbank, die trainierten Modelle zur Patentklassifikation und das Matching-Verfahren zwischen Patenten und Unternehmungen sind vielversprechende Hilfestellungen für die wissenschaftliche Nachnutzung.

## Teil II: Eingehende Darstellung

# EDV-TEK: Entstehung, Diffusion und Verwertung von Technologien zur Eindämmung des Klimawandels

Förderkennzeichen: 16DKWN059  
 Laufzeit: 01.09.2022 bis 30.11.2025

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter den Förderkennzeichen 16DKWN059 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei dem/der Autor/in bzw. den Autor/inn/en.

Finanziert durch die Europäische Union – NextGenerationEU. Die geäußerten Ansichten und Meinungen sind ausschließlich die des Autors/der Autoren und spiegeln nicht unbedingt die Ansichten der Europäischen Union oder der Europäischen Kommission wieder. Weder die Europäische Union noch die Europäische Kommission können für sie verantwortlich gemacht werden.

## Inhaltsverzeichnis

<b>1</b>	<b>Aufgabenstellung</b> .....	<b>2</b>
<b>2</b>	<b>Wissenschaftlicher und technischer Stand, an den angeknüpft wurde</b> .....	<b>2</b>
<b>3</b>	<b>Planung und Ablauf des Vorhabens</b> .....	<b>3</b>
<b>4</b>	<b>Inhaltliche Ergebnisse</b> .....	<b>4</b>
<b>5</b>	<b>Wichtigste Positionen des zahlenmäßigen Nachweises</b> .....	<b>8</b>
<b>6</b>	<b>Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten</b> .....	<b>8</b>
<b>7</b>	<b>Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen im Sinne des Verwertungsplans</b> .....	<b>9</b>
<b>8</b>	<b>Fortschritt auf dem Gebiet des Vorhabens während der Durchführung des Vorhabens bei anderen Stellen</b> .....	<b>10</b>
<b>9</b>	<b>Erfolgte oder geplante Veröffentlichungen der Projektergebnisse</b> .....	<b>10</b>
<b>10</b>	<b>Literaturverzeichnis</b> .....	<b>10</b>



## 1 Aufgabenstellung

Das Projekt EDV-TEK hatte zum Ziel, Technologien zur Eindämmung des Klimawandels (TEKs) umfassend entlang ihres Lebenszyklus zu untersuchen. Die zentrale Forschungsfrage war, wie moderne Verfahren des Deep Learnings und der natürlichen Sprachverarbeitung genutzt werden können, um (1) TEKs in Patentdaten zu identifizieren, (2) ihren wissenschaftlichen Ursprung und Neuheitsgrad zu bestimmen, (3) ihre Diffusion innerhalb und zwischen Technologiebereichen zu messen und (4) Muster ihrer Kommerzialisierung zu analysieren.

Die Aufgabenstellung gliederte sich in folgende Arbeitspakete:

AP 0 – Aufbau der Dateninfrastruktur: Erstellung einer PATSTAT-PostgreSQL-Datenbank mit integrierten Volltextdaten des EPO und USPTO als Grundlage für alle nachfolgenden Analysen.

AP 1 – Identifikation von TEKs: Entwicklung und Vergleich verschiedener Klassifikationsverfahren (klassische neuronale Netze, vortrainierte Sprachmodelle, Graph-neuronale Netze) zur Identifikation von TEKs auf Basis von Patentvolltexten und Netzwerkstrukturen.

AP 2 – Entstehung von TEKs: Untersuchung des wissenschaftlichen Ursprungs von TEKs über Patent-Paper-Verknüpfungen sowie Messung des Neuheitsgrades mittels semantischer Distanzmaße.

AP 3 – Diffusion von TEKs: Analyse der Verbreitung von TEKs innerhalb von CPC-Y02-Technologieklassen, zwischen diesen Klassen und auf internationaler Ebene (US ↔ EP), differenziert nach dem Neuheitsgrad der Patente.

AP 4 – Kommerzialisierung von TEKs: Untersuchung der kommerziellen Verwertung von TEKs durch Matching von Patentdaten mit Unternehmensdaten (Crunchbase) und Analyse von Finanzierungsmustern.

Im Kontext der Förderrichtlinie zur „Stärkung der Datenkompetenzen des wissenschaftlichen Nachwuchses“ (DKWN) zielte das Projekt darauf ab, die Verbindung fachlicher Expertise im Bereich der Technologie- und Innovationsforschung mit datenwissenschaftlichen Methoden des Deep Learnings exemplarisch umzusetzen und so zur Datenkompetenzsteigerung in der Fachdisziplin beizutragen.

## 2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

**Patentklassifikation von TEK-Technologien.** Die Identifikation von TEKs stützte sich zum Projektbeginn primär auf drei etablierte Klassifikationssysteme: die CPC-Y02-Klassifikation des Europäischen Patentamts, die IPC Green Inventory der WIPO sowie die OECD-Envtech-Klassifikation. Diese Systeme basieren auf manuell definierten Patentklassen und decken den Bereich der Klimaschutztechnologien umfassend ab. Die automatisierte Klassifikation auf Basis von Patentvolltexten mittels Deep Learning befand sich noch in einem frühen Forschungsstadium.

**NLP und Deep Learning für Patentdaten.** Im Bereich der Textklassifikation hatten sich vortrainierte Sprachmodelle wie BERT und seine Varianten (DistilBERT, RoBERTa) als leistungsstarke Verfahren etabliert. Deren spezifische Anwendung auf Patentvolltexte zur Klassifikation von TEKs war jedoch kaum untersucht. Ebenso waren Graph-neuronale Netze (GNNs) wie GraphSAGE und GAT für Knotenklassifikation und Link Prediction in heterogenen Graphen verfügbar, aber ihre Anwendung auf Patent-Zitationsnetzwerke in Kombination mit textuellen Features stellte eine methodische Lücke dar.

**Reliance on Science.** Die Arbeit von Marx und Fuegi (2022) zur Verknüpfung von Patenten und wissenschaftlichen Publikationen stellte einen umfangreichen Datensatz von Patent-Paper-Zitationen und Patent-Paper-Paaren bereit, der als Grundlage für die Analyse des wissenschaftlichen Ursprungs von TEKs diente.

**Messung von Neuartigkeit.** Das Verfahren von Shibayama et al. (2021) zur Messung von Novelty in der Wissenschaft mittels Wortvektoren bot einen methodischen Ansatzpunkt. Die Übertragung auf Patente und speziell auf TEKs war jedoch noch nicht erfolgt.



**Technologiediffusion.** Die Messung der Diffusion von Technologien über Patentzitationen war ein etabliertes Verfahren. Die ergänzende Messung semantischer Diffusion über Wort-Embeddings und Ähnlichkeitsmaße stellte einen neueren Ansatz dar, der im Kontext von TEKs noch nicht systematisch angewendet worden war.

**Kommerzialisierung und Startup-Finanzierung.** Die Verknüpfung von Patentdaten mit Unternehmensdatenbanken wie Crunchbase erforderte robuste Name-Matching-Verfahren. Bestehende Ansätze nutzten mehrstufige Matching-Pipelines mit Harmonisierung, Blocking und verschiedenen String-Ähnlichkeitsmaßen, waren jedoch nicht spezifisch für den Kontext von TEK-Patenten und Startup-Finanzierung angewendet worden.

### 3 Planung und Ablauf des Vorhabens

#### Ursprüngliche Planung

Das Vorhaben war ursprünglich für eine Laufzeit von 36 Monaten (01.09.2022 bis 31.08.2025) geplant. Die Arbeitspakete waren sequenziell angelegt, wobei AP 0 (Datenbankaufbau) die Grundlage für alle folgenden APs bildete. Die APs 1 bis 4 bauten aufeinander auf, da insbesondere die Ergebnisse der Identifikation (AP 1) und der Neuheitsmessung (AP 2) in die nachfolgenden Analysen zur Diffusion (AP 3) und Kommerzialisierung (AP 4) einfließen.

#### Tatsächlicher Ablauf

Der Aufbau der PATSTAT-Datenbank (AP 0) erfolgte in den ersten Monaten des Projektes und umfasste den Import von elf PATSTAT-Datensätzen (Stand: Spring 2023) in eine PostgreSQL-Datenbank mit vollständiger Tabellenstruktur, Primär- und Fremdschlüsseln sowie die Integration der EPO- und USPTO-Volltextdaten.

Die Arbeiten an AP 1 (Identifikation) wurden Ende 2022 und Anfang 2023 durchgeführt. Die Identifikation der TEK-Patente erfolgte über drei Klassifikationssysteme (CPC Y02, IPC Green Inventory, OECD Envtech), und es wurden insgesamt drei Klassen von Klassifikationsmodellen entwickelt und verglichen. Die Arbeitspakete 2.1 (Patent-Paper-Verknüpfungen) und 2.2 (Neuheitsmessung) wurden im Laufe des Jahres 2023 begonnen, im Jahr 2024 vertieft und abgeschlossen.

Die Diffusionsanalysen in AP 3 wurden im zweiten Halbjahr 2024 und ersten Halbjahr 2025 durchgeführt. Hierbei erwies sich die Berechnung der semantischen Diffusion als rechenintensiver als ursprünglich geplant, da die FAISS-basierte Ähnlichkeitssuche über große Patentmengen auf einem HPC-Cluster durchgeführt werden musste.

Das Arbeitspaket 4 (Kommerzialisierung) wurde ab Anfang 2025 bearbeitet. Das mehrstufige Matching-Verfahren zwischen Crunchbase- und PATSTAT-Daten erforderte eine umfangreiche Entwicklungsarbeit, insbesondere bei der Namensharmonisierung und dem Umgang mit großen Datenmengen.

#### Abweichungen von der ursprünglichen Planung

Der Projektfortschritt wich zeitlich um etwa drei bis fünf Monate von der Vorhabenskizze ab. Dies war insbesondere auf die Überarbeitung des Ansatzes in AP 2 (Entstehungsanalyse, Patent-Paper Paare), den höher als erwarteten Rechenaufwand bei der Diffusionsanalyse (AP 3) sowie die Komplexität des Crunchbase-PATSTAT-Matchings (AP 4) zurückzuführen. Die Abweichung wurde durch eine kostenneutrale Laufzeitverlängerung bis zum 30.11.2025 kompensiert. Die methodischen Zielsetzungen und die inhaltliche Ausrichtung des Vorhabens blieben unverändert.



## 4 Inhaltliche Ergebnisse

### AP 0 – Aufbau der PATSTAT-Postgres-Datenbank

Als Grundlage für alle nachfolgenden Analysen wurde eine umfassende PATSTAT-PostgreSQL-Datenbank aufgebaut. Der Import umfasste die PATSTAT Global Spring 2023-Daten aus elf Datenpartitionen mit insgesamt ca. 30 Tabellen (u.a. tls201\_appln, tls206\_person, tls207\_pers\_appln, tls209\_appln\_ipc, tls211\_pat\_publn, tls212\_citation, tls224\_appln\_cpc). Es wurden vollständige Primär- und Fremdschlüsselbeziehungen erstellt sowie Indizes auf den relevanten Abfragespalten angelegt. Ergänzend wurden die EPO-Volltextdaten (Claims, Descriptions, Abstracts) und die USPTO-Volltextdaten (Brief Summary, Claims, Description) über PatentsView integriert. Die Datenbank dient als zentrale Infrastruktur für die Patentforschung am Institut und wird über das Projektende hinaus genutzt.

### AP 1 – Identifikation von TEKs

#### Extraktion der TEK-Patente

Die Identifikation der TEK-Patente erfolgte über drei komplementäre Klassifikationssysteme: die CPC-Y02-Klassifikation (Klassen Y02A bis Y02W), die IPC Green Inventory der WIPO (6.847 spezifische Umwelttechnologie-Codes) und die OECD-Envtech-Klassifikation. Die Extraktion beschränkte sich auf erteilte EP- und US-Patente. Die Vereinigung dieser drei Quellen ergab einen Datensatz von ca. 1,87 Millionen eindeutigen TEK-Patenten mit zugehörigen Titel-, Abstract- und Claims-Daten in englischer Sprache.

Für das Training der Klassifikationsmodelle wurden zudem ca. 3 Millionen Non-TEK-Patente extrahiert, die über Zitationsbeziehungen mit den TEK-Patenten verbunden sind. Diese strategische Auswahl der Negativbeispiele stellt sicher, dass die Klassifikationsaufgabe anspruchsvoll und realitätsnah ist, da die Kontrollgruppe thematisch in der Nähe der TEK-Patente liegt.

#### Vergleich der Klassifikationsverfahren

Es wurden insgesamt neun Modelle aus drei Architekturklassen entwickelt, trainiert und auf einem einheitlichen Datensatz (je 100.000 TEK- und Non-TEK-Patente, 80/10/10-Split) verglichen.

**Klassische neuronale Netze.** Es wurden vier Architekturen implementiert: ein Feed-Forward Network (FNN) mit EmbeddingBag-Aggregation, ein Convolutional Neural Network (CNN) mit Multi-Filter-Architektur (Filtergrößen 3, 4, 5), ein Long Short-Term Memory (LSTM) und ein Recurrent Neural Network (RNN).

**Vortrainierte Sprachmodelle (LLM-Classfier).** Zwei Modelle wurden für die binäre Klassifikation feinabgestimmt: DistilBERT (distilbert-base-uncased) und DistilRoBERTa (distilroberta-base), jeweils mit einer maximalen Sequenzlänge von 512 Tokens und über 5 Epochen.

**Graph-neuronale Netze (GNN-Classfier).** Für die GNN-basierte Klassifikation wurde ein heterogener Graph aufgebaut mit 200.000 Patentknoten (768-dimensionale DistilBERT-Embeddings), 535.000 Autorenknoten, 532.510 Patent-zu-Patent-Zitationen und 756.150 Autor-zu-Patent-Verbindungen. Es wurden vier GNN-Architekturen verglichen: ein SimplifiedGNN (nur Patent-Zitationskanten), ein FullGNN (alle Kantentypen), ein EnhancedGNN (mit GAT-Attention und Residualverbindungen) und ein TransformerGNN (Multi-Head-Attention mit Multi-Scale-Feature-Aggregation).

#### Ergebnisse der Klassifikation

Die folgende Tabelle zeigt die Ergebnisse aller neun Modelle auf dem Testset:

Rang	Modell	Test Accuracy	F1-Score	Precision	Recall
1	DistilBERT	76,75%	76,72%	76,91%	76,75%
2	DistilRoBERTa	76,69%	76,68%	76,72%	76,69%
3	Feed-Forward NN	74,24%	74,10%	74,71%	74,24%
4	FullGNN	71,44%	71,45%	71,49%	71,44%



5	CNN	70,62%	70,21%	71,77%	70,62%
6	EnhancedGNN	69,69%	69,69%	69,76%	69,69%
7	SimplifiedGNN	68,23%	68,23%	68,24%	68,23%
8	LSTM	65,16%	64,84%	65,69%	65,16%
9	RNN	54,98%	54,97%	54,98%	54,98%

Die vortrainierten Sprachmodelle erreichten die beste Gesamtperformance mit ausgeglichener Precision-Recall-Balance. Bemerkenswert ist, dass das einfache Feed-Forward Network (74,24 %) die komplexeren CNN-, LSTM- und sogar die GNN-Architekturen übertraf. Dies zeigt, dass semantische Bag-of-Words-Repräsentationen für Patent-Abstracts bereits sehr effektiv sind, da technische Schlüsselwörter wichtiger sind als sequenzielle Struktur.

Die GNN-Modelle (68–71 %) blieben hinter den textbasierten Modellen zurück. Die Analyse ergab, dass Patent-Abstracts bereits die wesentliche technologische Information enthalten und die zusätzlichen strukturellen Signale aus dem Zitations- und Autorennetzwerk oft redundant zu den Textinformationen sind. Das FullGNN konvergierte bemerkenswert früh nach nur einer Trainingsepoche, wobei weitere Epochen zu Overfitting führten.

## AP 2 – Entstehung von TEKs

### AP 2.1 – Patent-Paper-Verknüpfungen und GNN Link Prediction

Zur Untersuchung des wissenschaftlichen Ursprungs von TEKs wurden die Patente mit wissenschaftlichen Publikationen verknüpft. Als Datenquellen dienten der Reliance-on-Science-Datensatz (Marx und Fuegi, 2022) mit Patent-Paper-Zitationen und Patent-Paper-Paaren sowie die OpenAlex-Datenbank, aus der Titel, Abstracts, Autorenschaften und Zitationsbeziehungen der verknüpften Publikationen extrahiert wurden. Der Fokus lag auf US-Patenten, da die Patent-Paper-Pair-Daten nur für US-Patente verfügbar sind.

Nach Filterung umfasste der Datensatz 1.642.111 Patente, 1.137.787 wissenschaftliche Publikationen, ca. 5,6 Millionen Patent-Paper-Zitationen und 97.928 validierte Patent-Paper-Paare. Das extreme Klassenungleichgewicht (19 Millionen negative zu einem positiven Beispiel) erforderte die Generierung von Hard Negative Samples.

Für die GNN-basierte Link Prediction wurde ein heterogener Graph aufgebaut mit Patent- und Paper-Knoten (Features: paraphrase-MiniLM-L6-v2-Embeddings, 384-dimensional) und vier Kantenarten: Patent-zitiert-Patent, Paper-zitiert-Paper, Patent-zitiert-Paper und Patent-Paper-Paar (Zielgröße). Die Ergebnisse auf dem Testset zeigen:

Modell	AUC	Average Precision	F1-Score	Precision	Recall
SimpleBaselineModel	99,93%	99,70%	98,87%	97,86%	99,90%
PatentPaperLinkPredictor	99,75%	98,96%	96,07%	92,65%	99,75%

Das SimpleBaselineModel – das ohne Message Passing auskommt und nur auf transformierten Embeddings basiert – übertraf den komplexeren PatentPaperLinkPredictor mit GNN-Architektur in allen Metriken. Dieses Ergebnis zeigt, dass die semantische Textähnlichkeit zwischen Patent- und Paper-Abstracts das stärkste Signal für den wissenschaftlichen Ursprung liefert und zusätzliche Graph-Strukturen keinen proportionalen Mehrwert bieten.

### AP 2.2 – Messung des Neuheitsgrades

Die Messung des Neuheitsgrades erfolgte nach dem Verfahren von Shibayama et al. (2021). Für jedes Patent mit mindestens zwei zitierten wissenschaftlichen Publikationen wurden die Embeddings der zitierten Paper ermittelt und paarweise Cosinus-Distanzen berechnet. Der Neuheitsgrad wurde über verschiedene Perzentile der Distanzverteilung operationalisiert (q100, q99, q95, q90, q80, q50), wobei hohe Distanzen auf die Kombination thematisch weit entfernter wissenschaftlicher Quellen und damit auf einen höheren Neuheitsgrad hinweisen.



Die Berechnung erfolgte parallelisiert über 32 CPU-Kerne und ergab Novelty-Scores für 318.104 Patente. Die folgende Tabelle zeigt die Verteilung der Novelty-Scores:

Metrik	Mittelwert	Median	Standardabweichung	90. Perzentil	95. Perzentil
Novelty_q100	0,783	0,814	0,247	1,083	1,122
Novelty_q99	0,759	0,801	0,226	1,020	1,050
Novelty_q95	0,727	0,768	0,206	0,962	1,002
Novelty_q90	0,703	0,740	0,194	0,927	0,973
Novelty_q80	0,667	0,696	0,180	0,877	0,935
Novelty_q50	0,582	0,591	0,159	0,770	0,855

Die Patente wurden auf Basis des novelty\_q100-Scores in drei Neuheitskategorien eingeteilt: High Novelty (31.820 Patente,  $\geq$  90. Perzentil), Medium Novelty (127.232 Patente, 50.–90. Perzentil) und Low Novelty (159.052 Patente,  $<$  50. Perzentil). Diese Einteilung floss als Unterscheidungskriterium in die nachfolgenden Diffusions- und Kommerzialisierungsanalysen ein.

### AP 3 – Diffusion von TEKs

#### Methodischer Ansatz

Die Diffusion von TEKs wurde auf drei Ebenen analysiert: innerhalb von CPC-Y02-Technologieklassen (Within-Class), zwischen Technologieklassen (Between-Class) und international (US  $\leftrightarrow$  EP). Für jede Ebene wurden zwei Diffusionsmechanismen gemessen: die zitationsbasierte Diffusion über direkte Zitationsbeziehungen sowie die semantische Diffusion über die Cosinus-Ähnlichkeit von Wort-Embeddings (Schwellenwert: 0,92–0,98).

Als Embedding-Modell wurde ClimateBERT (distilroberta-base-climate-f) verwendet, da es speziell auf klimarelevante Texte vortrainiert ist. Die Ähnlichkeitssuche erfolgte über FAISS (Facebook AI Similarity Search). Berücksichtigt wurden 770.168 Y02-klassifizierte Patente im Zeitraum 1980–2023. Die Embeddings wurden auf einem GPU-Cluster berechnet und die eigentliche Diffusionsanalyse parallelisiert auf einem HPC-Cluster durchgeführt.

#### Ergebnisse der Diffusionsanalyse

Insgesamt wurden über 3,5 Millionen Diffusionsverbindungen identifiziert.

Within-Class-Diffusion. Die folgende Tabelle zeigt die durchschnittlichen Diffusionszeiten innerhalb der acht Y02-Klassen:

Y02-Klasse	Beschreibung	Durchschnittliche Diffusionszeit
Y02T	Transport	5,7 Jahre
Y02D	IT-Effizienz	5,8 Jahre
Y02C	Kohlenstoffmanagement	5,9 Jahre
Y02E	Energieerzeugung	6,2 Jahre
Y02W	Abfallmanagement	6,5 Jahre
Y02P	Industrieprozesse	6,7 Jahre
Y02B	Gebäude	6,8 Jahre
Y02A	Adaption Klimawandel	8,1 Jahre

Y02T (Transport) und Y02D (IT-Effizienz) zeigten die schnellste Within-Class-Diffusion, während Y02A (Adaption an den Klimawandel) mit 8,1 Jahren die langsamste aufwies.

Between-Class-Diffusion. Die stärksten Diffusionspfade zwischen verschiedenen Y02-Klassen bestanden zwischen Y02E (Energieerzeugung) und Y02T (Transport) mit durchschnittlichen Diffusionszeiten von 10,2 bis 10,8 Jahren. Y02W (Abfallmanagement) und Y02C (Kohlenstoffmanagement) zeigten mit 11,3 bis 11,5 Jahren die langsamsten Between-Class-Diffusionszeiten.

Internationale Diffusion. Die bidirektionale Analyse zwischen US- und EP-Patenten ergab insgesamt über 2,2 Millionen internationale Diffusionsverbindungen. Nur ca. 0,21 % der US-Patente zitierten EP-



Patente, während ca. 4,41 % der EP-Patente US-Patente zitierten. Die Richtung US → EP war mit ca. 1,23 Millionen Verbindungen leicht stärker ausgeprägt als EP → US (ca. 0,98 Millionen). Die schnellste internationale Diffusion zeigte sich im Bereich Y02D (IT-Effizienz) mit 7,2 Jahren. Ca. 95 % der internationalen Diffusionsverbindungen basierten auf semantischer Ähnlichkeit, was die Bedeutung impliziter Wissensflüsse gegenüber expliziten Zitationen unterstreicht.

Differenzierung nach Neuheitsgrad. Alle drei Diffusionstypen wurden nach den Neuheitskategorien aus AP 2.2 differenziert:

Neuheitskategorie	Within-Class	Between-Class	International	Durchschnitt gesamt
High Novelty	5,2 Jahre	8,9 Jahre	9,1 Jahre	7,7 Jahre
Medium Novelty	6,8 Jahre	10,8 Jahre	10,9 Jahre	9,5 Jahre
Low Novelty	7,9 Jahre	12,1 Jahre	11,8 Jahre	10,6 Jahre

High-Novelty-Patente diffundierten durchschnittlich 2,9 Jahre schneller als Low-Novelty-Patente. Der Effekt war bei der Between-Class-Diffusion am stärksten ausgeprägt (3,2 Jahre Unterschied). Dieses Ergebnis widerspricht der Erwartung, dass radikalere Innovationen aufgrund höherer Adoptionsbarrieren langsamer diffundieren, und deutet darauf hin, dass neuartige Technologiekombinationen im TEK-Bereich schneller aufgegriffen werden.

#### AP 4 – Kommerzialisierung von TEKs

##### Matching-Verfahren

Zur Verknüpfung der Patentdaten mit Unternehmensdaten wurde eine mehrstufige Matching-Pipeline zwischen der Crunchbase-Datenbank und PATSTAT implementiert (Tarasconi et al. 2017). Die Namensharmonisierung umfasste die Unicode-Normalisierung, die Entfernung von Rechtsformbezeichnungen in über 20 Sprachen sowie akademischer Titel und Interpunktion. Das Company Matching erfolgte in vier Stufen:

- (1) Perfect Match nach Harmonisierung mit Prefix-basiertem Blocking,
- (2) Alphanumeric Match nach Entfernung aller Nicht-Alphanumerischen Zeichen,
- (3) Jaro-Winkler-Fuzzy-Matching mit Token-basierter Gewichtung (Schwellenwert: 85 %) und
- (4) Levenshtein-Distanz-Matching mit Validierung über Erfinder-Übereinstimmung.

PATSTAT-Erfinder wurden anhand von fünf Kriterien disambiguiert (gemeinsame Patentfamilie, IPC4-Codes, Land, Zeitraum, Co-Erfinder). Crunchbase-Personen wurden über Bigramm-Ähnlichkeit mit den disambiguierten Erfindern verknüpft.

##### Ergebnisse

Das Matching ergab ca. 1,16 Millionen deduplizierte Firmen-Patent-Verknüpfungen und 400.778 Personen-Patent-Verknüpfungen. Nach Verknüpfung mit Crunchbase-Funding-Rounds umfasste der finale Datensatz 467.532 Patent-Finanzierungs-Verbindungen für 17.608 eindeutige Unternehmen mit durchschnittlich 2,88 Funding-Runden pro Unternehmen.

Die Analyse nach Y02-Klassen zeigte, dass Y02E (Energieerzeugung) mit 1.838 Unternehmen und 12.501 Patenten die am stärksten vertretene Klasse ist, gefolgt von Y02A (Anpassung) mit 1.571 Unternehmen und Y02P (Industrieprozesse) mit 1.236 Unternehmen. Y02C (Kohlenstoffmanagement) war mit 73 Unternehmen und 181 Patenten die am schwächsten vertretene Klasse.

Nach Filterung auf Patente mit verfügbaren Novelty-Scores (122.456 TEK-Finanzierungs-Verbindungen, 33,4 % des Gesamtdatensatzes) ergab die Differenzierung nach Neuheitskategorien statistisch signifikante Unterschiede in den Finanzierungsbeträgen (Kruskal-Wallis-Test,  $p < 0,001$ ). High-Novelty-Patente wiesen eine höhere Median-Finanzierung (11.000.000 USD) und eine höhere Funding-Success-Rate (87,5 %) auf als Medium-Novelty (8.713.627 USD, 85,9 %) und Low-Novelty-Patente (8.500.000 USD, 85,0 %). Gleichzeitig zeigten die durchschnittlichen Finanzierungsbeträge pro Runde ein differenziertes Bild: High-Novelty lag bei 78,2 Mio. USD (1.081 Unternehmen), während Medium- und Low-Novelty deutlich höhere Durchschnitte aufwiesen (400,4 bzw. 525,5 Mio. USD). Dies deutet darauf hin, dass High-Novelty-Patente häufigere, aber kleinere Finanzierungsrunden anziehen.



## Beitrag zur Datenkompetenzsteigerung (DKWN)

Das Projekt hat in mehrfacher Hinsicht zur Datenkompetenzsteigerung in der Fachdisziplin der Technologie- und Innovationsforschung beigetragen. Die Zusammenarbeit zwischen fachlicher Expertise (Technologie- und Innovationsforschung, Unternehmertum) und datenwissenschaftlichen Methoden (Deep Learning, NLP, Graph-neuronale Netze) hat neue Forschungserkenntnisse ermöglicht, die mit traditionellen bibliometrischen Methoden nicht erzielbar gewesen wären. Insbesondere die Kombination von textuellen Features (Embeddings) mit strukturellen Features (Zitationsnetzwerke) in den GNN-Modellen und die Anwendung domänenspezifischer Sprachmodelle (ClimateBERT) auf Patentdaten stellen methodische Beiträge dar. Die im Projekt gewonnene Datenkompetenz fließt unmittelbar in den Masterkurs „Deep Learning for Social Analytics“ am Institut für Unternehmertum ein, wodurch die Datenkompetenzsteigerung über das Projekt hinaus nachhaltig in die Lehre integriert wird.

## 5 Wichtigste Positionen des zahlenmäßigen Nachweises

Die Ausgaben verteilten sich im Wesentlichen auf folgende Positionen:

Personalausgaben (Pos. 0812/0817): 311.813,61 (Jürgen Thiesen, Jonas Wilinski, Joschka Schwarz und Oliver Mork)  
 Hilfskräfte (Pos. 0822): -  
 Dienstreisen (Pos. 0846): 1.657,81 (DRUID Konferenz, BMBF/BMFTR Statusveranstaltung)  
 Sonstige Sachausgaben (Pos. 0843): 7.673,50 (Lizenz Crunchbase und PATSTAT Datensatz)  
 Geräte (Pos. 0850): -

Die Ausgabenplanung weicht von der Vorhabenbeschreibung ab, da eine Mittelumwidmung von Sach- in Personalmittel stattgefunden hat.

## 6 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Der Aufbau einer eigenen PATSTAT-PostgreSQL-Datenbank (AP 0) war notwendig, da die Volltextanalysen in den nachfolgenden Arbeitspaketen den direkten und performanten Zugriff auf Millionen von Patentdokumenten erforderten. Die Integration der EPO- und USPTO-Volltextdaten in eine einheitliche Datenbankstruktur mit Indizierung ermöglichte Abfragen, die über die Online-Schnittstellen der Patentämter nicht in der benötigten Geschwindigkeit und dem benötigten Umfang möglich gewesen wären.

Die Entwicklung und der Vergleich dreier Klassen von Klassifikationsmodellen (AP 1) war notwendig, um die Eignung verschiedener Verfahren für die TEK-Identifikation empirisch zu bewerten. Der systematische Vergleich klassischer neuronaler Netze, vortrainierter Sprachmodelle und Graph-neuronaler Netze lieferte Erkenntnisse, die bei der Verwendung nur einer Modellklasse nicht gewonnen worden wären. Insbesondere das Ergebnis, dass einfache Feed-Forward-Netze die komplexeren GNN-Architekturen übertreffen, hat methodische Relevanz für die Forschungsgemeinschaft.

Die Nutzung von HPC-Ressourcen für die Diffusionsanalyse (AP 3) war angemessen, da die FAISS-basierte Ähnlichkeitssuche über Hunderttausende von Patenten und die parallelisierte Berechnung von Diffusionsmetriken den Rechenkapazitäten einzelner Arbeitsplatzrechner nicht entsprechen konnte. Die Verwendung von ClimateBERT als domänenspezifischem Embedding-Modell war durch die klimarelevante Fragestellung begründet.

Das mehrstufige Matching-Verfahren in AP 4 war notwendig, da einfache Name-Matching-Verfahren aufgrund der unterschiedlichen Schreibweisen in PATSTAT und Crunchbase zu hohen Fehlerquoten geführt hätten. Die Implementierung eines vierstufigen Verfahrens mit Validierung über Erfinder-



Übereinstimmungen gewährleistete eine hohe Matching-Qualität bei gleichzeitig handhabbarem Rechenaufwand.

## 7 Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen im Sinne des Verwertungsplans

### Wissenschaftliche Verwertung

Publikationen in wissenschaftlichen Zeitschriften und Journals, welche auf den Projektergebnissen basieren, werden für die Zukunft angestrebt. Neben der Veröffentlichung in wissenschaftlichen Medien sollen die Projektwebseite, sowie das GitHub Code Repository die wissenschaftliche Verwertung unterstützen.

### Nachhaltige Nutzung der Dateninfrastruktur

Die im Projekt aufgebaute PATSTAT-PostgreSQL-Datenbank wird am Institut für Unternehmertum der TUHH über das Projektende hinaus nachhaltig genutzt. Mehrere Doktorand:innen und Studierende am Institut nutzen die Datenbank für ihre Dissertationen sowie Abschluss- und Projektarbeiten. Die Datenbank stellt damit eine langfristige Forschungsinfrastruktur dar, die weit über den ursprünglichen Projektkontext hinauswirkt.

### Integration in die Lehre

Die im Projekt erworbenen Fähigkeiten, Daten und Aufgabenstellungen sind zu einem großen Maße in den Masterkurs „Deep Learning for Social Analytics“ (<https://tuhhstartupengineers.github.io/master-deep-learning-social-analytics/>) am Institut für Unternehmertum eingeflossen. Der Kurs, der seit 2024 angeboten wird, nutzt Projektdaten und -methoden in den Vorlesungsinhalten und in den von den Studierenden zu bearbeitenden Projekten. Die Daten wurden hierfür explizit aufbereitet (Patent-Paper-Pairs, Paper-Paper-Zitationen, Paper-Metadaten). Die Integration in die Lehre stellt sicher, dass die im Projekt aufgebaute Datenkompetenz nachhaltig an den wissenschaftlichen Nachwuchs weitergegeben wird.

### Abschlussarbeiten und Promotionsvorhaben

Mehrere Masterarbeiten haben unmittelbar auf den Inhalten und Daten des Projektes EDV-TEK aufgebaut.

Darüber hinaus orientiert sich eine Promotionsarbeit unmittelbar an den Inhalten des Projektes.

### Methodische Anschlussforschung

Die im Projekt aufgebaute Expertise in der Anwendung großer Sprachmodelle und Attention-Mechanismen auf domänenspezifische Texte bildet die Grundlage für aktuelle Forschung im Bereich der Keyphrase Extraction auf Basis von Large Language Models und Attention Maps. Diese Anschlussforschung überträgt die methodischen Kompetenzen, die im Kontext der Patentklassifikation gewonnen wurden, auf neue Anwendungsfelder.

### Code und Datensätze

Der Fortschritt der Codedatenbank ist im öffentlichen GitHub-Repository ([https://github.com/juergenct/Projekt\\_EDV-TEK](https://github.com/juergenct/Projekt_EDV-TEK)) abgelegt und steht der wissenschaftlichen Community zur Nachnutzung zur Verfügung. Ergänzend dokumentiert die Projektwebseite ([https://tuhhstartupengineers.github.io/Website\\_EDV-TEK/](https://tuhhstartupengineers.github.io/Website_EDV-TEK/)) die Ergebnisse und Methoden in aufbereiteter Form.



## 8 Fortschritt auf dem Gebiet des Vorhabens während der Durchführung des Vorhabens bei anderen Stellen

Während der Laufzeit des Projektes sind im Bereich der TEK-Identifikation und -Analyse relevante Fortschritte erzielt worden. Insbesondere die rasante Entwicklung großer Sprachmodelle (Large Language Models) hat neue Möglichkeiten für die Textanalyse von Patentdaten eröffnet. Die im Projekt verwendeten Modelle (DistilBERT, DistilRoBERTa) stellen etablierte Architekturen dar, die durch die Fortschritte im Bereich der LLMs bestätigt und erweitert wurden. Die Berücksichtigung weiterer Fortschritte im Bereich großer Sprachmodelle und im Deep Learning ist in die methodische Arbeit des Projektes eingeflossen.

## 9 Erfolgte oder geplante Veröffentlichungen der Projektergebnisse

### Lehrveranstaltungen (Wissenschaftskommunikation)

Masterkurs „Deep Learning for Social Analytics“ am Institut für Unternehmertum der TUHH (seit 2024), in dem Projektdaten, -methoden und -aufgabenstellungen unmittelbar in die Lehre integriert wurden.

URL: <https://tuhhstartupengineers.github.io/master-deep-learning-social-analytics/>

### Code-Repository und Projektwebseite

GitHub-Repository: [https://github.com/juergenct/Projekt\\_EDV-TEK](https://github.com/juergenct/Projekt_EDV-TEK)

Projektwebseite: [https://tuhhstartupengineers.github.io/Website\\_EDV-TEK/](https://tuhhstartupengineers.github.io/Website_EDV-TEK/)

Neben der Veröffentlichung/Nutzung in Lehrveranstaltungen, im Code-Repository und auf der Projektwebseite sind weitere wissenschaftliche Publikationen in Planung.

## 10 Literaturverzeichnis

Marx, Matt, und Aaron Fuegi. „Reliance on Science: Worldwide Front-page Patent Citations to Scientific Articles.“ *Strategic Management Journal* 43, Nr. 8 (2022): 1624–1654.

Marx, Matt. „Reliance on Science.“ Zenodo, 2024. <https://doi.org/10.5281/zenodo.11461587>.

Shibayama, Sotaro, Deyun Yin, und Kuniko Matsumoto. „Measuring Novelty in Science with Word Embedding.“ *PLOS ONE* 16, Nr. 7 (2. Juli 2021). <https://doi.org/10.1371/journal.pone.0254034>.

Tarasconi, Gianluigi and Menon, Carlo. „Matching Crunchbase with patent data“ *OECD Science, Technology and Industry Working Papers* 2017/07. 2017. <https://dx.doi.org/10.1787/15f967fa-en>