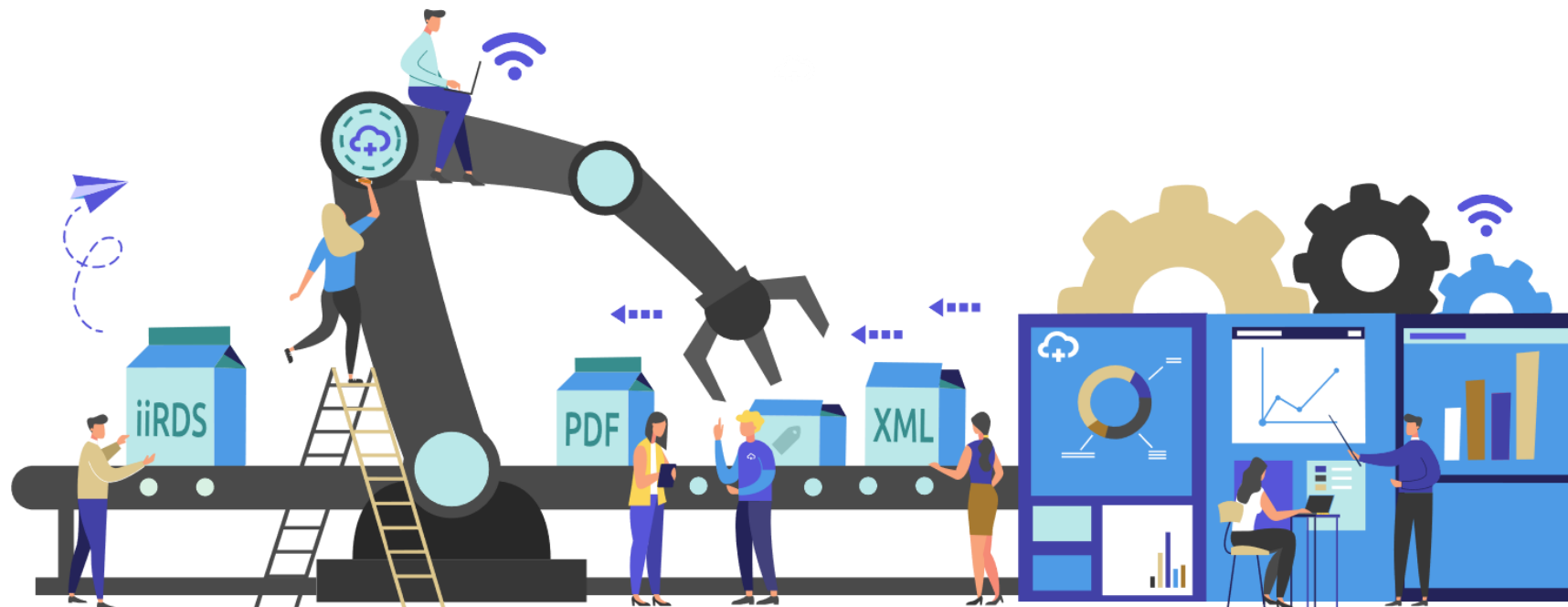


# Neuer Glanz für alte PDFs

Dr. Jan Oevermann, plusmeta GmbH



# Die Frage

**„Wie können automatisiert Metadaten  
für PDF-Dokumente generiert werden?“**

# Methoden

Regelbasiert / Maschinelles Lernen / Linguistische Verfahren

# Verfahren

Regelbasierte Vergabe



Maschinelles Lernen



Deep Learning



Linguistische Verfahren

# Regelbasierte Zuweisung

- Grundlage meist vordefinierten Wertelisten
  - Standardisierbar für informationsbezogene Metadaten
  - Immer individuell für produktbezogene Metadaten
  - Mehrsprachige Bezeichnungen
  - Synonyme und Indikatoren
- Sprachstatistische Abgleiche von n-Grammen
  - Effiziente Spracherkennung
  - Ähnlichkeiten zwischen Texten erkennbar

# Regelbasierte Zuweisung

- Vorhersage über ein Scoring-System und Ranking
  - Durchsucht Textquellen nach Vorkommen
  - Fuzzy Matching mit Distanzbestimmung
  - Scoring-Faktor abhängig von Quelle und Fundort
  - Vorkommenshäufigkeit
- Kein Training mit Beispieldaten notwendig

Automatisches Speichern ☒ PI-Fan\_Eigenschaften - mit Varian... - Zuletzt geändert: Fr um 10:07 Jan Oevermann

Datei Start Einfügen Seitenlayout Formeln Daten Überprüfen Ansicht Add-Ins Hilfe Team

D8

	A	B	C
1	Produktname (DE)	Produktname (EN)	Identifizier (optional)
2	T3-B		<a href="https://metadata.plusmeta.de/products/pi-fan/t3-b">https://metadata.plusmeta.de/products/pi-fan/t3-b</a>
3	T3-H1		<a href="https://metadata.plusmeta.de/products/pi-fan/t3-h1">https://metadata.plusmeta.de/products/pi-fan/t3-h1</a>
4	T5-B		<a href="https://metadata.plusmeta.de/products/pi-fan/t5-b">https://metadata.plusmeta.de/products/pi-fan/t5-b</a>
5	TP-B		<a href="https://metadata.plusmeta.de/products/pi-fan/tp-b">https://metadata.plusmeta.de/products/pi-fan/tp-b</a>
6	Rotor		<a href="https://metadata.plusmeta.de/supplier/rotor-gmbh/rotor">https://metadata.plusmeta.de/supplier/rotor-gmbh/rotor</a>
7	Standplatte	Base plate	<a href="https://metadata.plusmeta.de/supplier/halterungen-gmbh/standplatte">https://metadata.plusmeta.de/supplier/halterungen-gmbh/standplatte</a>
8	Teleskopstange	Telescopic rod	<a href="https://metadata.plusmeta.de/supplier/halterungen-gmbh/teleskopstange">https://metadata.plusmeta.de/supplier/halterungen-gmbh/teleskopstange</a>
9	Heizung	Heating	<a href="https://metadata.plusmeta.de/supplier/kaelteundwaermetechnik-ag/heizung">https://metadata.plusmeta.de/supplier/kaelteundwaermetechnik-ag/heizung</a>
10	Beleuchtung	Lighting	<a href="https://metadata.plusmeta.de/supplier/lightning-ltd/beleuchtung">https://metadata.plusmeta.de/supplier/lightning-ltd/beleuchtung</a>
11	Antrieb	Drive	<a href="https://metadata.plusmeta.de/supplier/drivesandcontrols-ltd/antrieb">https://metadata.plusmeta.de/supplier/drivesandcontrols-ltd/antrieb</a>
12	Anzeige- und Bedienelemente	Control	<a href="https://metadata.plusmeta.de/supplier/drivesandcontrols-ltd/anzeigeundbedier">https://metadata.plusmeta.de/supplier/drivesandcontrols-ltd/anzeigeundbedier</a>

Übersicht Organisationen **Produkte** Organisation <> Produkt Produl ...

Quelle: plusmeta GmbH / Inhalt: PI-Fan Project

### Eigenschaft

Klasse  
**Dokumentart**

Datentyp  
**Listenwert**

Bezeichnung (de)  
**Montageanleitung**

Bezeichnung (en)  
**Assembly instructions**

Identifikator/URI  
**iirds:AssemblyInstructions**

Indikatoren

Installation MHB Assembly Instruction Erection Procedure

Montage-Anweisung Aufbauanleitung Arbeitsanleitung

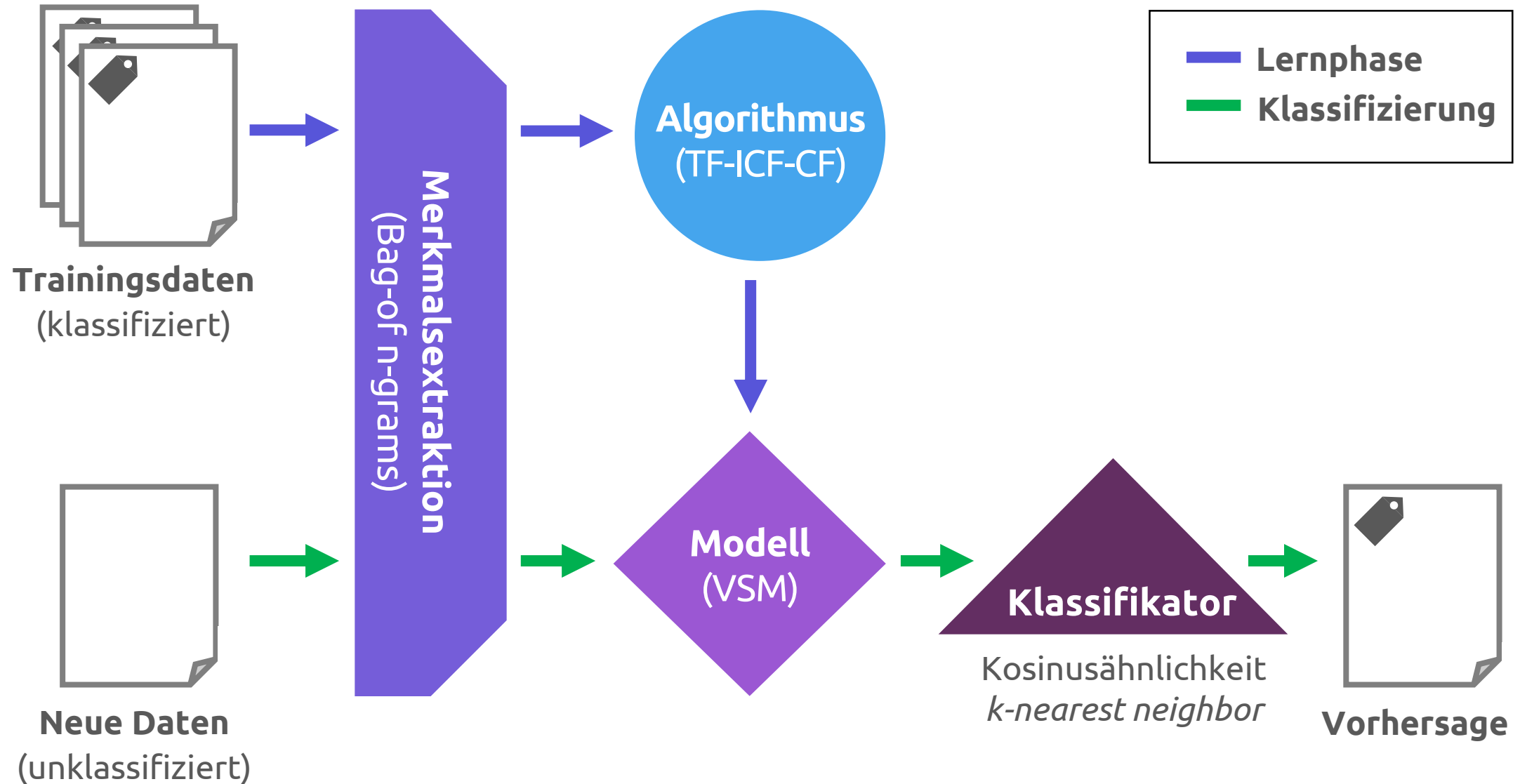
Montageanweisung Einbauanleitung Montagehandbuch

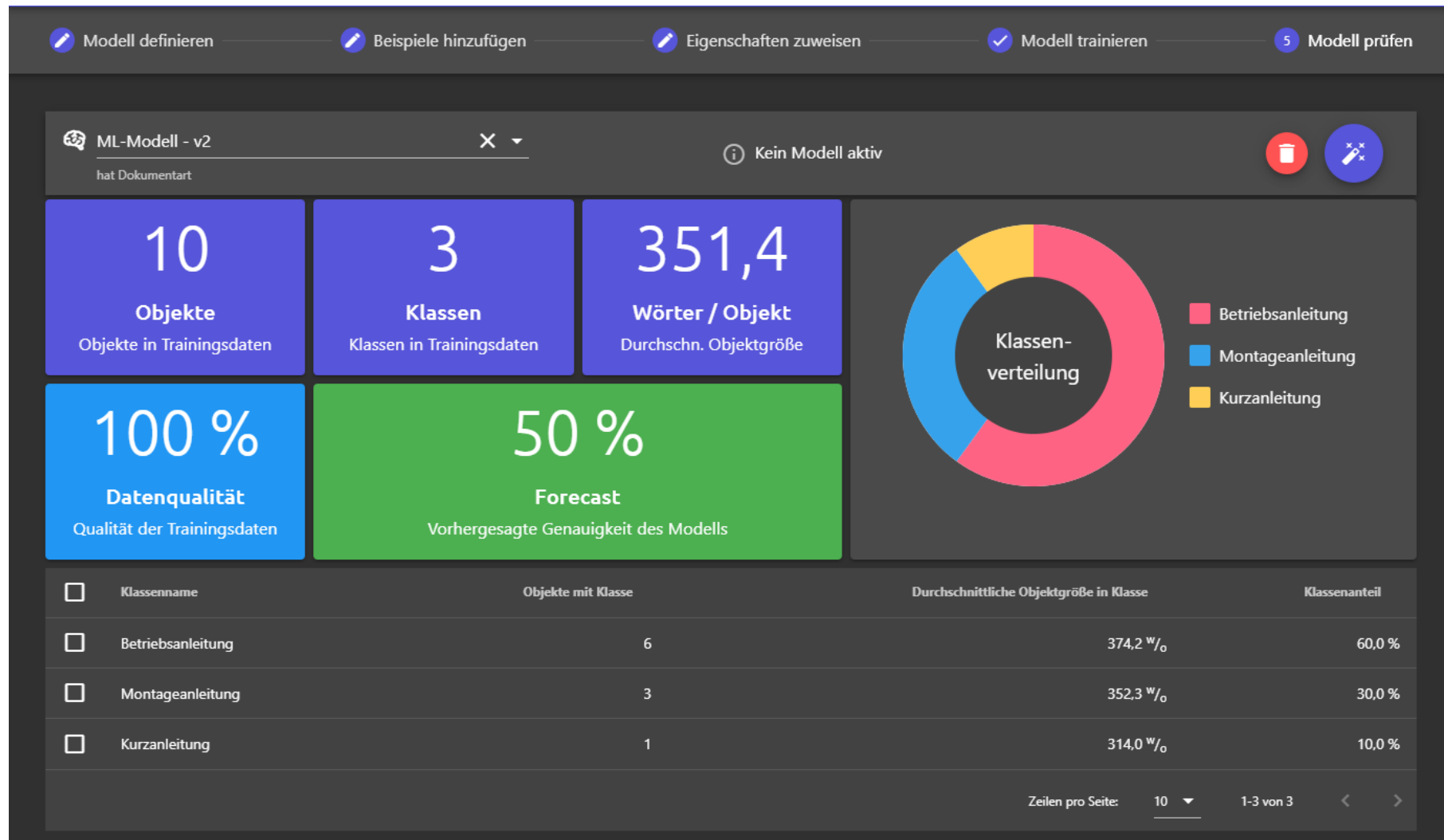
Quelle: plusmeta-Plattform

# Machine / Deep Learning

- Wissen wird auf Basis von Erfahrung generiert
- Training des Systems über gelabelte Beispiele (Goldstandard)
- Erkennen und Gewichten von charakteristischen Textmustern
- Vorhersage über Abgleich mit hinterlegtem Modell







Quelle: plusmeta-Plattform

# Linguistische Verfahren

- Basierend auf grammatischem Modell der Sprache
- Kann Textfunktionen erkennen (PoS-Tagging)
- Häufige Anwendung: Named Entity Recognition
- Kann eigene Werte und Listen erzeugen ohne Vorwissen

Um PDF-Dateien in moderne Nutzungsszenarien einzubinden, werden Metadaten und Struktur benötigt. Mit Hilfe regelbasierter Ansätze oder Methoden der Künstlichen Intelligenz können Informationen automatisiert extrahiert und ausgewertet werden. Im Vortrag werden verschiedene Möglichkeiten beleuchtet, deren Vorteile abgewogen und Fallstricke der Metadatengenerierung anhand von Beispielen aufgezeigt.

Art: Schlagwort

#### ✓ Schlagwort

Metadaten	2	■ ■
Struktur	1	■
Methode	1	■
Vorteil	1	■
PDF-Datei	1	■
Vortrag	1	■
Intelligenz	1	■
Fallstrick	1	■

Quelle: <http://intrafind.org/demo/taggingservice/>

# Konfidenz & Provenienz

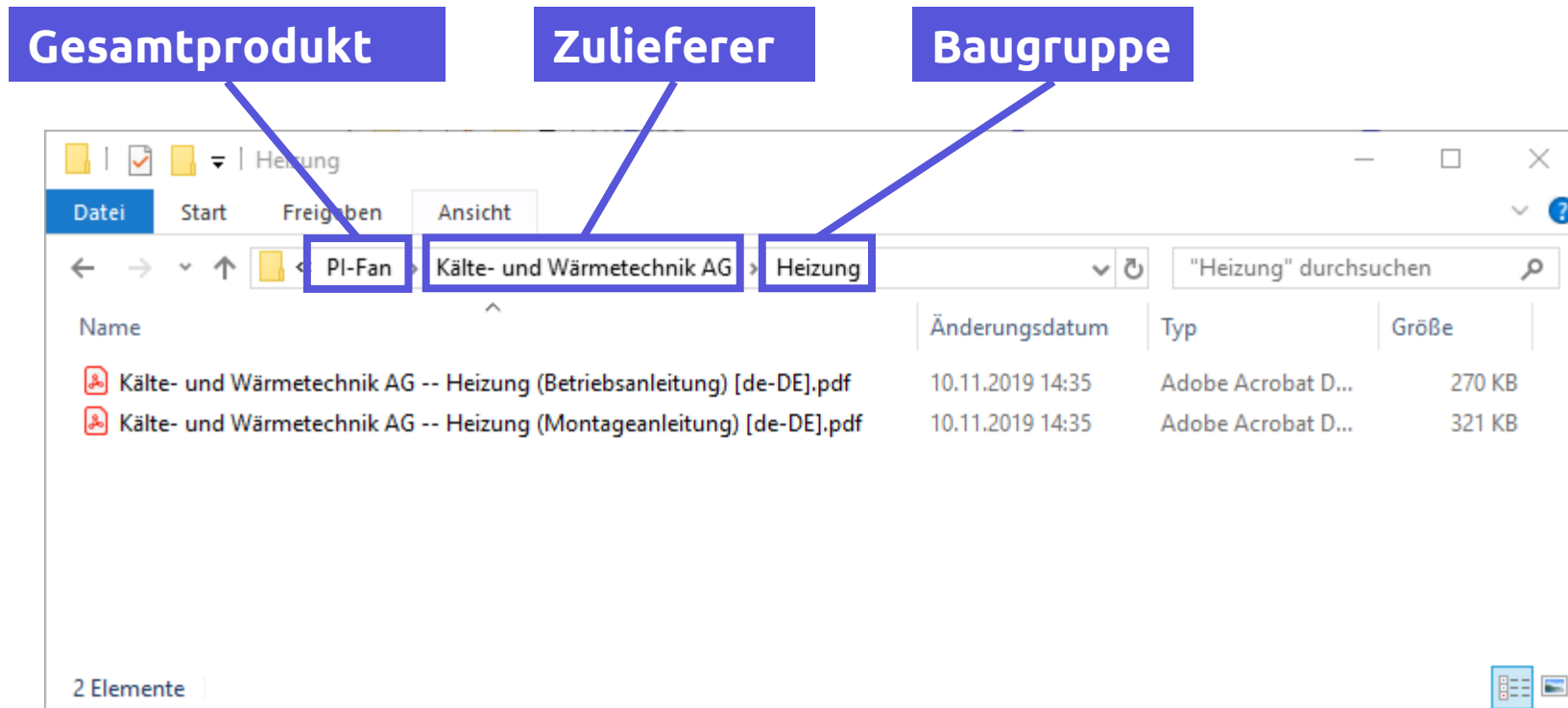
- Übergreifendes Konfidenzmaß
  - Aussage darüber, wie sicher die Vorhersage ist
  - Dient zur Qualitätssicherung und Kontrolle durch Menschen  
(*Human-in-the-Loop*)
- Herkunft von Metadaten kennzeichnen
  - Durch Mensch oder Maschine vergeben
  - Welche Methode kam unter welchen Bedingungen zum Einsatz

# Metadaten

Ort / Name / Datei / Text / Beziehungen

# Ablageort

- Metadaten in Ordnerstruktur kodiert
- Geteilter Ordner / SharePoint / WebDAV / DMS
- Unterschiedliche Abteilungen nutzen verschiedene Ablagen
- Mehrfachzuordnungen problematisch



Quelle: plusmeta GmbH



# Dateinamen

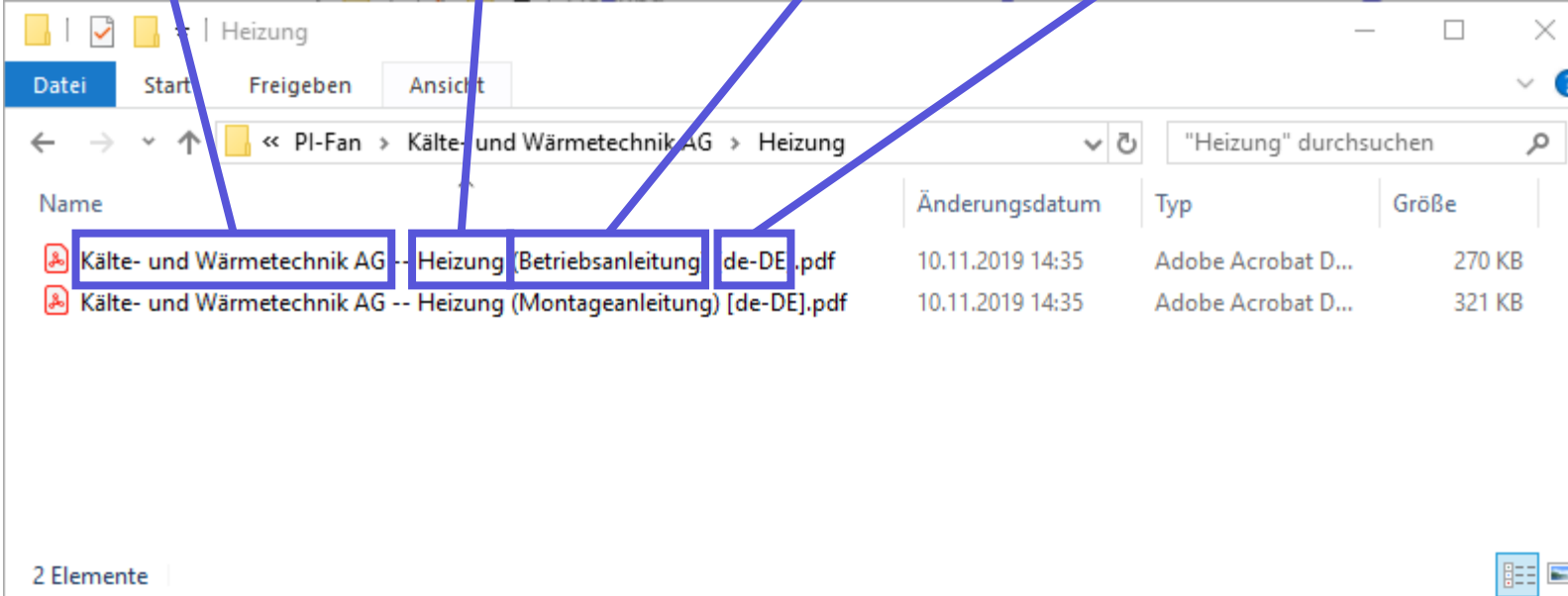
- Oft regelmäßig aufgebaut / interne Logik
- Immer lesbarer/zugänglicher Text
- Fast immer Sprache / Versionsinformationen enthalten
- Teilweise verfälscht durch Zeichenrestriktionen

Zulieferer

Baugruppe

Dokumentart

Sprache



Name	Änderungsdatum	Typ	Größe
Kälte- und Wärmetechnik AG -- Heizung Betriebsanleitung de-DE .pdf	10.11.2019 14:35	Adobe Acrobat D...	270 KB
Kälte- und Wärmetechnik AG -- Heizung (Montageanleitung) [de-DE].pdf	10.11.2019 14:35	Adobe Acrobat D...	321 KB

2 Elemente

Quelle: plusmeta GmbH

# Dateinamen: Beispiele

- 63407004\_DE.pdf
- DCBS ChamberSystemManual\_2\_3a\_en.pdf
- PSGDFP01\_DE\_04.pdf
- SB30-50-1AV-40-BE-de-10.pdf
- Operating\_instructions\_microScan3\_Core\_I\_O\_de\_IM0063754.PDF
- 1992-08-01\_Verfahrensdatenblatt\_Hydraulikpumpe.pdf

Tag-Matching

Trennungsmarker

Eigene Synonyme

Fuzzy Matching

# Dateimetadaten

- Qualität bestimmt durch PDF-Generator
- Mögliche Quelle mit hoher Gewichtung
- Typische Inhalte: Autor, Titel, Software, Erstellungsdatum
- Viele (auch eigene) Metadaten möglich, jedoch selten benutzt

Dokumenteigenschaften

Beschreibung Sicherheit Schriften Benutzerdefiniert Erweitert

Beschreibung

Datei: Kälte- und Wärmetechnik AG -- Heizung (Betriebsanleitung) [de-DE].pdf

Titel: Die Heizung

Verfasser: Kälte- und Wärmetechnik AG

Thema: Betriebsanleitung

Stichwörter: T3-H1 | X3-H1

Erstellt am: 24.09.2019 00:42:28

Geändert am: 24.09.2019 00:42:28

Anwendung: Microsoft® Word für Office 365

Erweitert

PDF erstellt mit: Microsoft® Word für Office 365

PDF-Version: 1.7 (Acrobat 8.x)

Speicherort: C:\Users\jan.oevermann\Downloads\PI-Fan\Kälte- und Wärmetechnik AG\Heizung\

Dateigröße: 269,47 KB (275.941 Byte)

Seitenformat: 210 x 297 mm      Seitenanzahl: 3

PDF mit Tags: Ja      Schnelle Webanzeige: Nein

OK      Abbrechen

Titel

Autor

Dokumentart /  
Gültigkeit

Verwaltungs-  
metadaten

Quelle: plusmeta GmbH

#### Beschreibung

Datei: EDBS ChamberSystemManual\_2\_3a\_en.pdf

Titel: STD\_EDBS\_V2\_2.book

Verfasser: Eric

Name Quelldatei

#### Beschreibung

Datei: PSGDFP01\_DE\_04.pdf

Titel: Anleitung

Verfasser: "Schmidt, Uwe"

Zusatzinformationen

#### Beschreibung

Datei: SB30-50-1AV-40-BE-de-10.pdf

Titel: Betriebsanleitung - SUNNY BOY 3.0 / 3.6 / 4.0 / 5.0

Verfasser: SMA Solar Technology AG

Thema: Betriebsanleitung

Metadaten  
wie erwartet

#### Beschreibung

Datei: Operating\_instructions\_microScan3\_Core\_I\_O\_de\_IM0063754.PDF

Titel: microScan3 Core I/O, 8016345

Verfasser: SICK AG | Subject to change without notice

Thema: Betriebsanleitung

Stichwörter: 8016345/11JQ/1371238224232/6.13/2018-10-25/de

# Extrahierbarer Text

- Bei gescannten oder handschriftlichen Dokumenten ist meist kein eingebetteter Text vorhanden
- Textextraktion nicht trivial bei PDF-Dokumenten:
  - Mehrspaltiger Text (z.B. eine Sprache pro Spalte)
  - Tabellarischer Inhalt oder Diagramme mit Text
- Fokus auf Zuweisung Metadaten, nicht Rekonstruktion des Texts
- Textmenge muss definiert sein – mehr ist nicht gleich besser

1992	Tag	Name	Benennung	Zeichn. Nr. 1/2		
Bearb.	29.04.	Jenft	or 2,5 m <sup>3</sup>	10 093/92 h		
Gepr.	31.07.					
Stück	Benennung	DIN	Werkstoff	Teil	Maße	Atteste
1	Klöpperboden	28011 17440	1.4571	1	3/8 DVI 1600 x 13 min	HP7/2 + 8/1 W2 3.1B
1	Klöpperboden	28011 17440	1.4571	2	3/8 DVI 1600 x 10 min	HP7/2 + 8/1 W2 3.1B
1	Klöpperboden	28011 17440	1.4571	3	VI 1700 x 7 min	HP7/2 + 8/1 W2 3.1B
1	Mantelblech	17440	1.4571	4	4983 x 1340 x 14	W2 3.1B
1	Mantelblech	17440	1.4571	5	5319 x 1245 x 7	W2 3.1B
1	Mannloch DN 500			6	nach Zeichn. 10 093/92-1	Schmiedeteil W2/9 3.1B
1	Flansch HG 300	28437 17440	1.4571	7	PN10	
				8		
1	Flansch C150-168,3	2633 17440	1.4571	9	PN16	W2/9 3.1B
1	Rohr	17457	1.4571	10	φ168,3 x 5 x ~160	W2 3.1B
± 1	Flansch C80-88,9	2633 17440	1.4571	11	PN16	W2/9 3.1B
± 1	Rohr	17457	1.4571	12	φ88,9 x 4 x ~140	W2 3.1B
8	Flansch C50-60,3	2633 17440	1.4571	13	PN16	W2/9 3.1B
1	Rohr (8-teilig)	17458	1.4571	14	φ60,3 x 3,6 x ~1280	W2 3.1B
1	Konus	17440	1.4571	15	φ1700/1600 x 80 (7)	W2 3.1B
1	Konus	17440	1.4571	16	φ270/160 x 50 (6)	W2 3.1B

## Abweichende Grundlinie

## Oberfläche

Schweißnähte sauber verputzt, nicht verschliffen.

Erdblechteile allseitig matt gebeizt und passiviert

Stahlfeste gesandstrahlt SA 2 1/2

Grundansicht mit 2-komp.-Zinkstaubfarbe, Schichtdicke 80 µm

Korrosionsschutzsystem 612 (Höls WN 68 - 001)

## Notizen

## Kennzeichnung

Der Apparat ist an gut sichtbarer Stelle mit weißer Farbe in 5 cm Schriftgröße

wie folgt gekennzeichnet.

Anlagen - Nord →

## Werkstoffe / Werkstoffgruppe 6/1

Teile	Werkstoffe	Werkstoffblatt	Prüfgrundlage	Nachweis
		DIN oder VdTÜV	AD-Merkblatt	DIN 50049
1-3, 11, 12, 16-22, 31	14571	17440	W 2	3.1 B
5, 15-15	14571	17440	-	-
46-48	14301	17440	-	-
4, 9	RS137-2	17100	W 1	3.1 B
23, 27	RS137-2	17100	W 9	Stempel
36-45	RS137-2	17100	-	-

## Tabellen

## Durchgestrichener Text



# Dokumenttext

- Besonderer Fokus Titelseite und deren Formatierung
- Wichtigste Informationen für Dokumentklassifikation befinden sich auf den ersten Seiten (ca. erste 1000 Wörter ausreichend)
- Inhaltsverzeichnisse, Leerseiten, Schmutztitel, allg. Vorworte
- Geringere Gewichtung als andere Fundstellen
- Unterschied Dokument vs. (Daten-)Blatt / Plan

Produkt(-reihe)

Der PI-Fan

Dokumentart

Sicherheitsanleitung

Logo als Bild



Quelle: plusmeta GmbH / Inhalt: PI-Fan Project

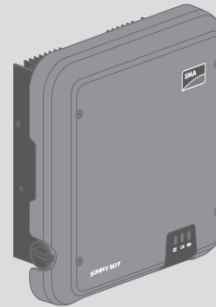
Vakuumsauggreifer ESG

**FESTO**



Quelle: Festo AG & Co. KG

Betriebsanleitung  
SUNNY BOY 3.0 / 3.6 / 4.0 / 5.0



DEUTSCH

SB30-50-1AV-40-BE-de-10 | Version 1.0

Quelle: SMA Solar Technology AG

microScan3 Core I/O

Sicherheits-Laserscanner

BETRIEBSANLEITUNG

**SICK**  
Sensor Intelligence.



Quelle: SICK AG

Betriebsanleitung



de

Modell:

**LR 1200 / LR 1250**

Seriennummer:

**135xxx**



000 - TT.MM.JJJJ

**LIEBHERR**

Quelle: Liebherr-International S.A.

# Beziehungswissen

- Typische Beziehungen zur Ableitung erweiterter Metadaten
  - Produkt -> Hersteller
  - Produkt -> Baugruppe
  - Baugruppe -> Zulieferprodukt
- Kann als zusätzliche Wissensquelle herangezogen werden

# Herausforderungen

- Sprache wird z.T. als Grundlage für OCR oder ML benötigt
- Reihenfolge kann relevant sein (siehe Produkt/Hersteller)
- Auswahl des Klassifikators (regelbasiert vs. ML)
- Kenntlichmachung gegenüber dem Benutzer
- Unterschiedliche Ansprüche je Metadatum

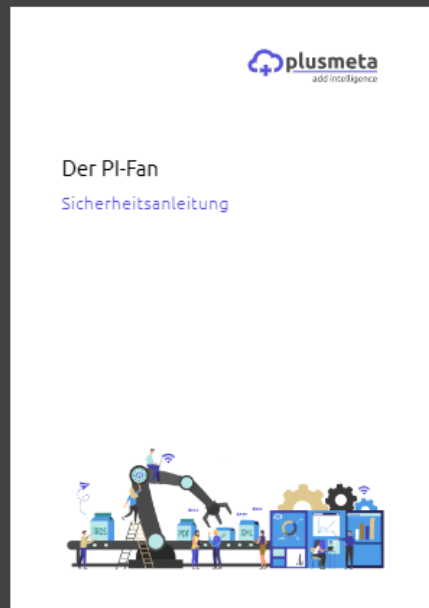
Alle Objekte

Allgemeine Metadaten

T Titel  Sprache

1 DOKUMENT  
PI-Fan\_Sicherheitsanleitung\_T3-B\_T5-B\_TP-B\_de-DE.pdf

✓ METADATEN ABNEHMEN



1 / 3



Titel*	T	PI-Fan AG - Sicherheitsanleitung - T5-B / TP-B / T3-B - Deutsch (Deutschland)	▼	
Sprache*		Deutsch (Deutschland) 	▼	
Produkt		T5-B  TP-B  T3-B 	▼	
Organisation		PI-Fan AG 	▼	
Dokumentart		Sicherheitsanleitung 	▼	

Quelle: plusmeta-Plattform

# Struktur

Inhaltsverzeichnis / Klassifikation

# Ebenen





# Inhaltsverzeichnis

- Zur kapitelbasierten Strukturierung geeignet.  
Keine zwangsläufige Entsprechung mit ursprünglicher Modularisierung
- Sind Herstellerübergreifend stark unterschiedlich
- Ebene zur Strukturierung muss bestimmt werden
- Kleinste Auflösung: Seite bzw. Beginn Kapitels

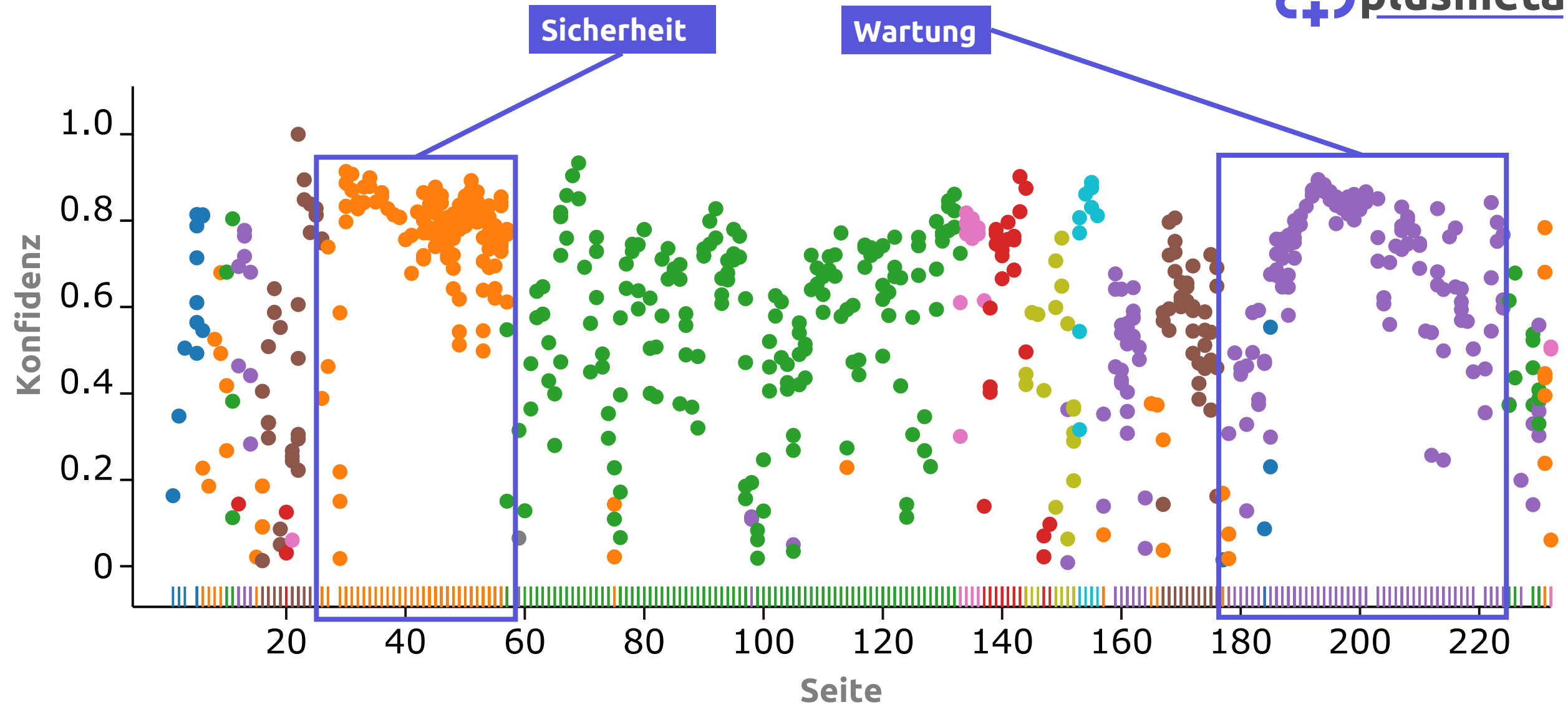
<b>1</b>	<b>Produktbeschreibung</b>	<b>49</b>
1.1	Typenschild	49
1.2	Konformitätserklärung	51
1.3	Bestimmungsgemäße Verwendung	52
1.4	Sonderbetriebsarten	54
1.5	Nicht bestimmungsgemäße Verwendung	55
1.6	Maschine	57
1.6.1	Maschine Übersicht	57
1.6.2	Technische Daten	58
1.6.3	Sicherheitseinrichtungen	63
1.6.4	Überwachungseinrichtungen	64
1.7	Grundgerät	66
1.7.1	Grundgerät Übersicht	66

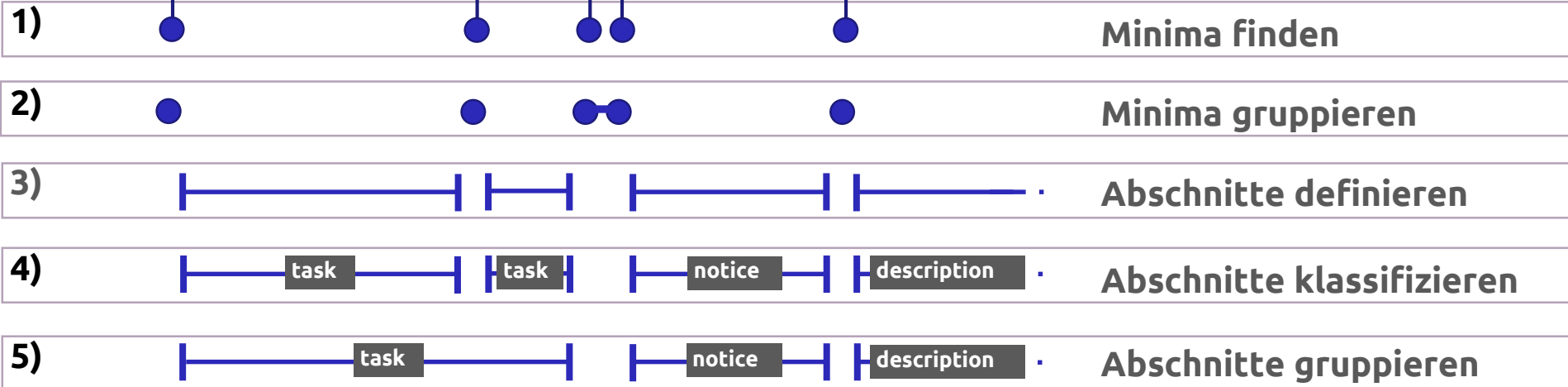
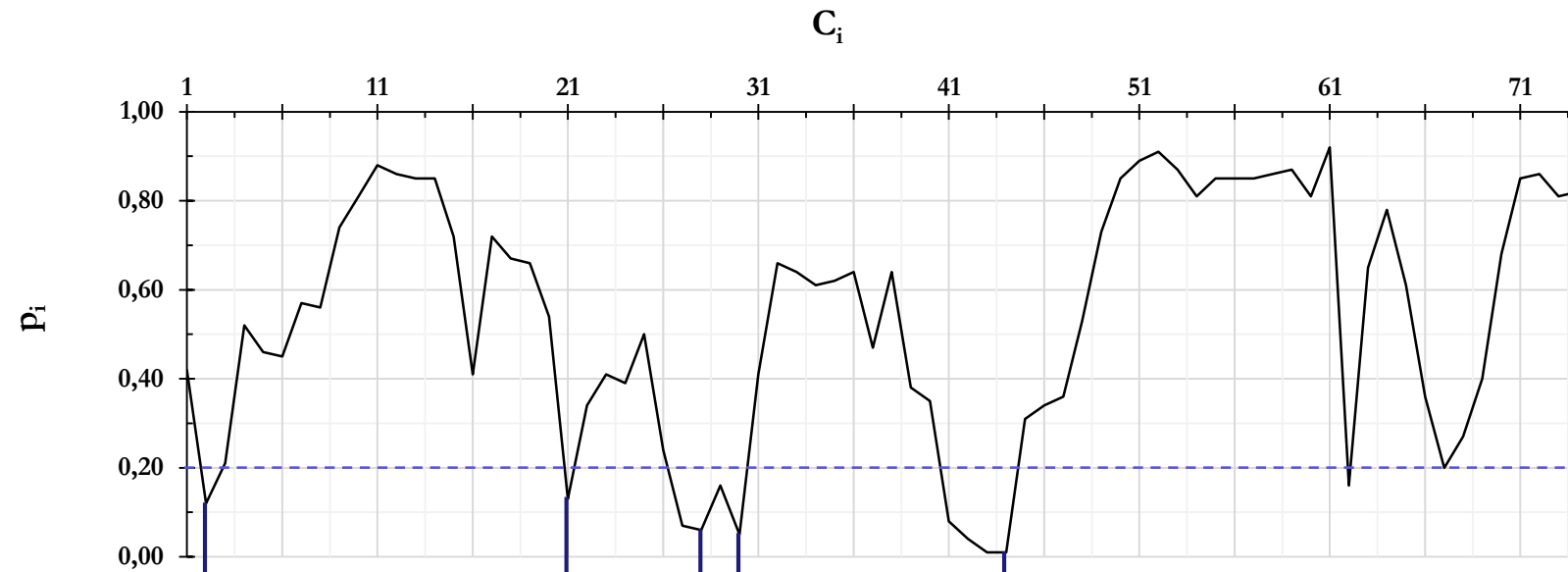
**Segmentgrenzen**

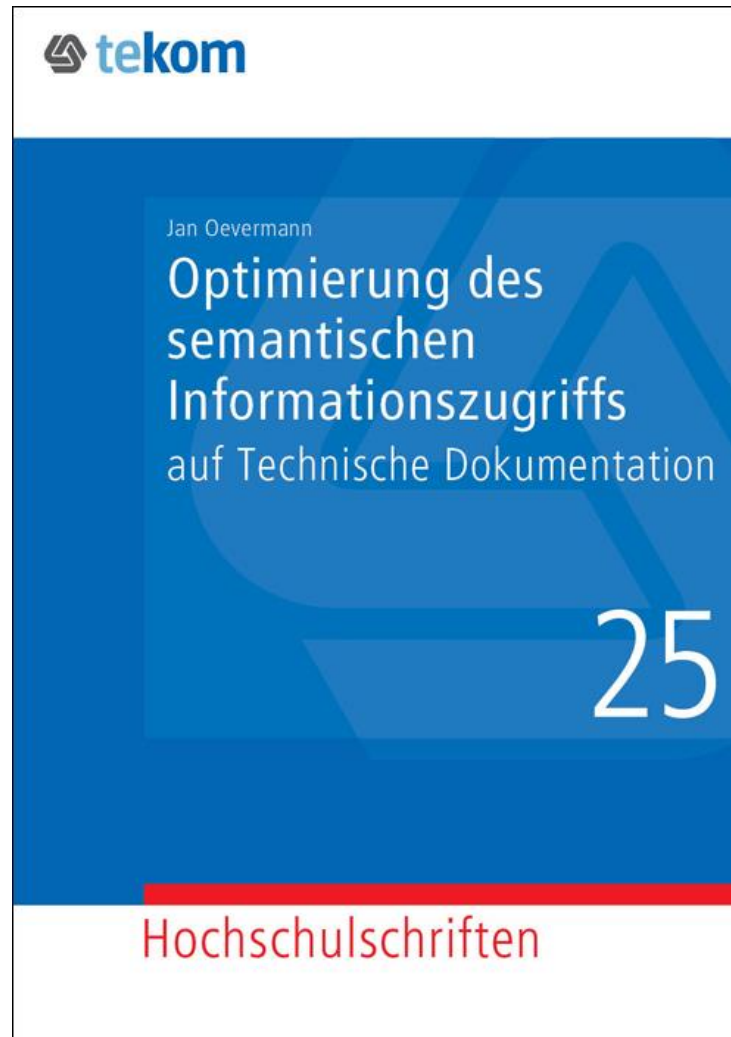
**Ebenenauswahl**

# Klassifikationsbasiert

- Basierend auf zuvor trainiertem Modell einer Klassifikation
- Extrahierter Text wird zerteilt und klassifiziert
- Kann auf semantischer Ebene eine Segmentierung vornehmen
- Strukturierte Trainingsdaten zwingend notwendig
- Auflösung auf Seitenbereiche bzw. Seitenteile







Quelle: <https://www.tekom.de/die-tekom/publikationen/fachbuecher>



**Dr. Jan Oevermann**

jan@plusmeta.de

plusmeta GmbH



Das Bewertungstool steht Ihnen auch noch nach der Tagung zur Verfügung!

**tekom** **tcworld**  
Jahrestagung 2019 conference 2019  
STUTTGART, 12. – 14. NOVEMBER STUTTGART, NOVEMBER 12 – 14

Ihre Meinung ist uns wichtig! Sagen Sie uns bitte, wie Ihnen der Vortrag gefallen hat. Wir freuen uns auf Ihr Feedback unter:

**<http://ai12.honestly.de>**

oder scannen Sie den QR-Code