What would you do with 212.765.957 DVDs?

Gregor Breitschopf, Gerrit Merz und Adrian Pucher

Build 30 towers equivalent to the height of the Burj Khalifa

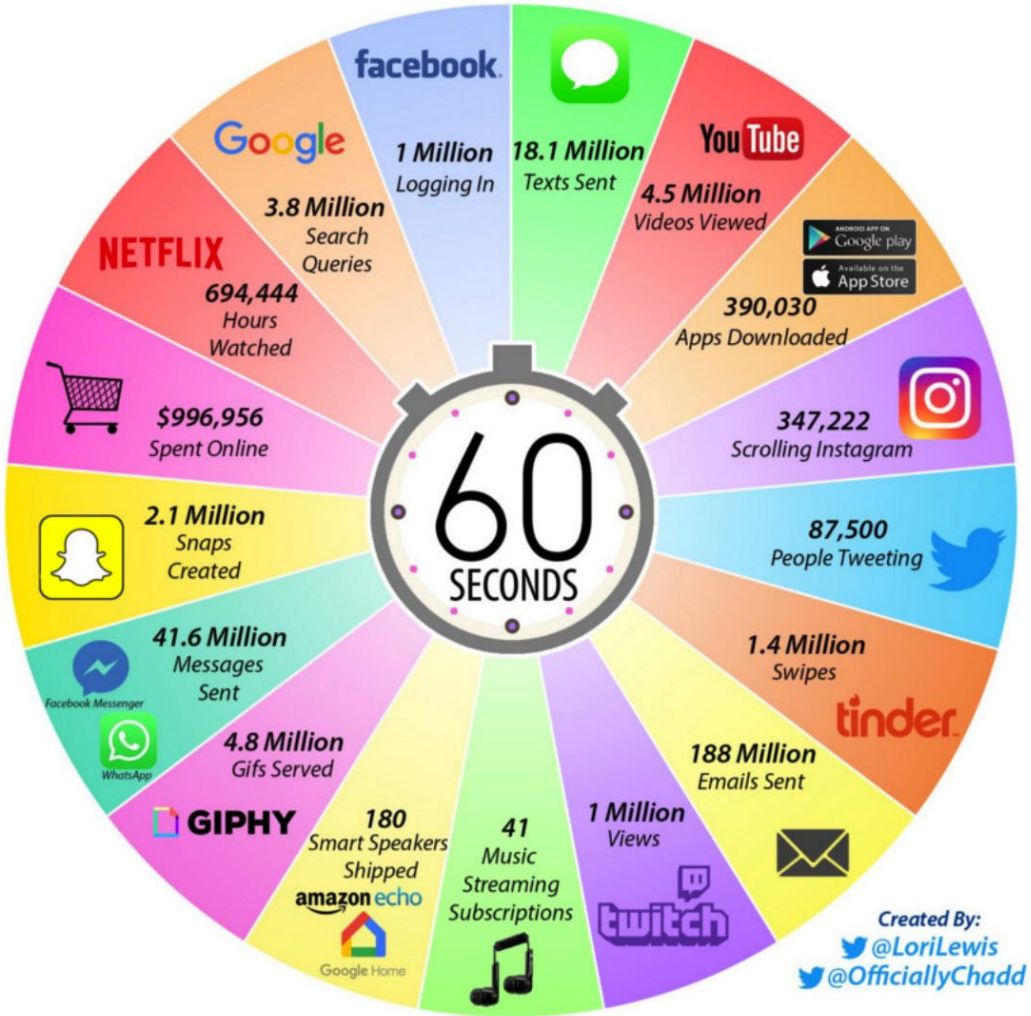
Cover the size of around 300 soccer fields

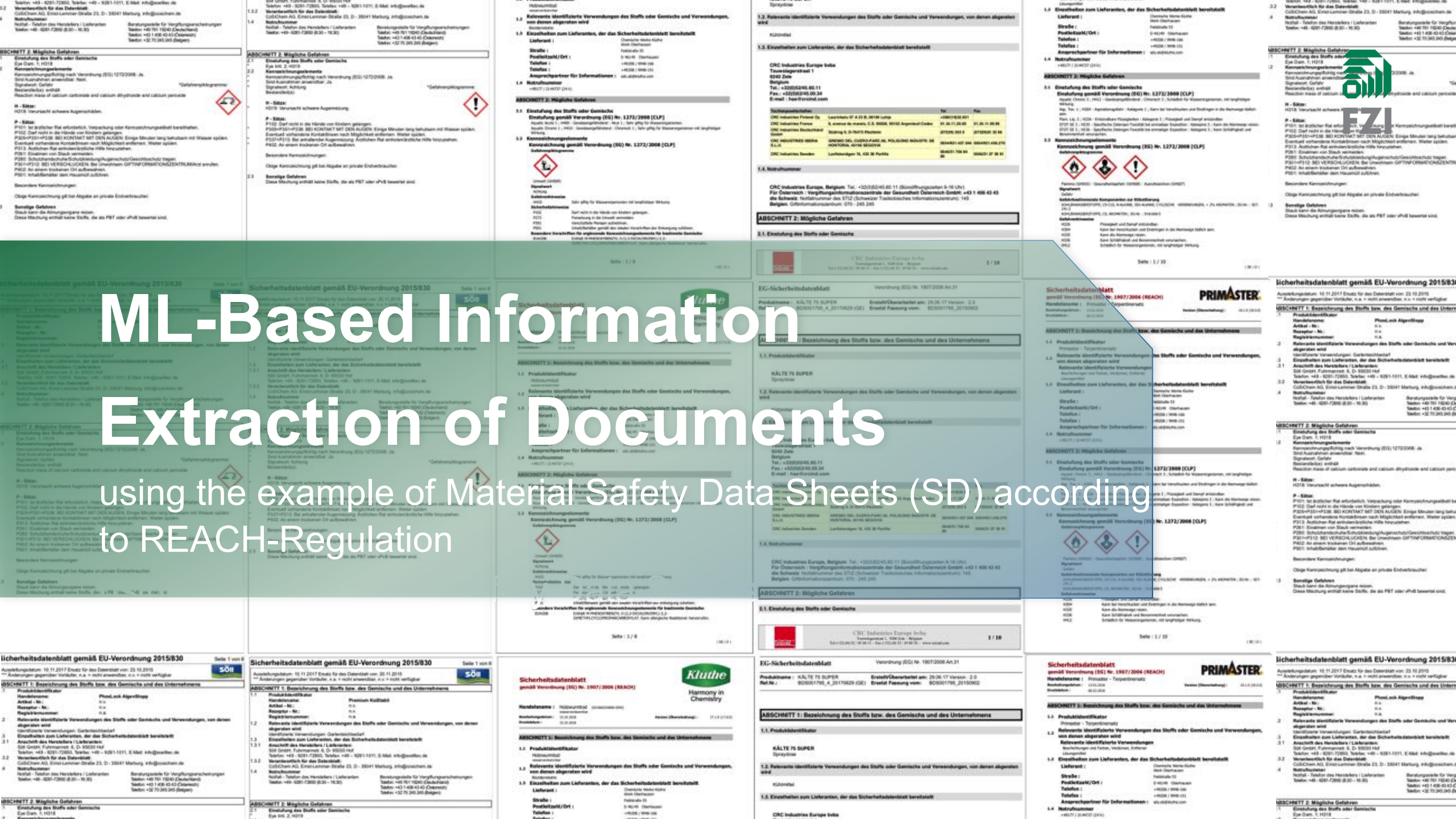
Right answer in the context of this seminar:

By 2025...

we could store the amount of data that is produced each day...

2019 This Is What Happens In An Internet Minute

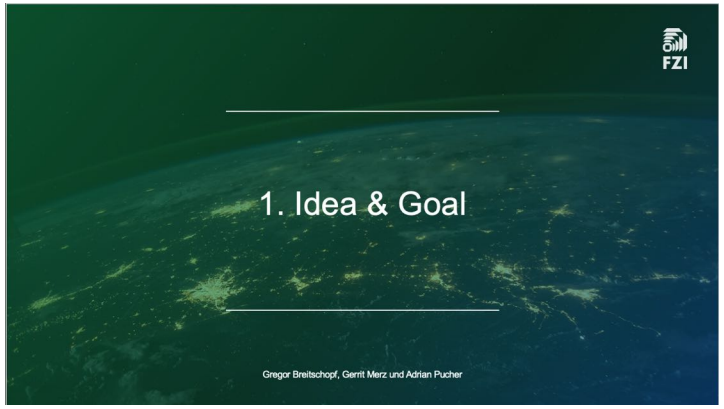




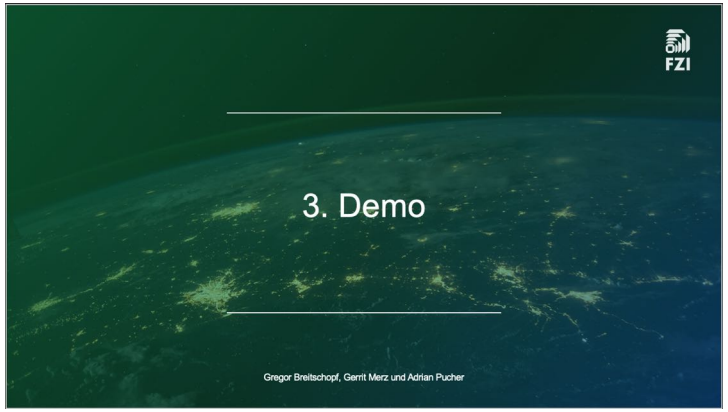
ML-Based Information Extraction of Documents

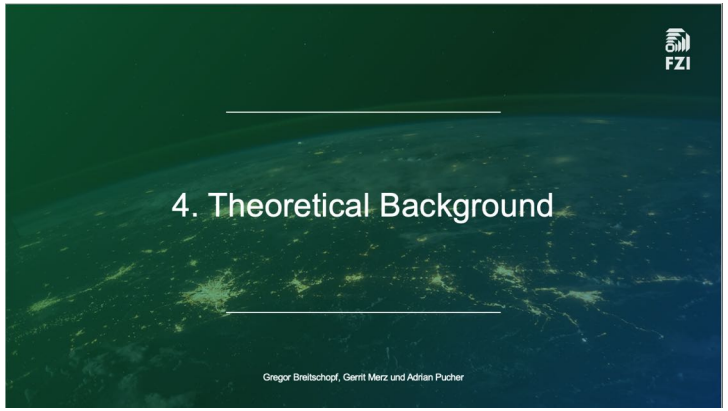
using the example of Material Safety Data Sheets (SD) according to REACH-Regulation

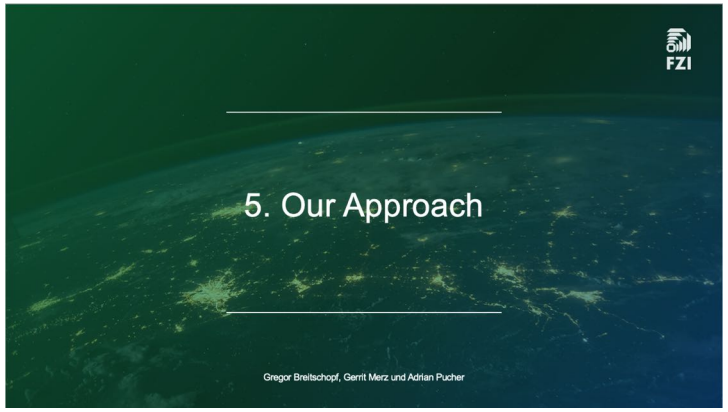
Agenda


1. Idea & Goal
Gregor Breitschopf, Gerrit Merz und Adrian Pucher


2. Finding the Right Use Case
Gregor Breitschopf, Gerrit Merz und Adrian Pucher


3. Demo
Gregor Breitschopf, Gerrit Merz und Adrian Pucher

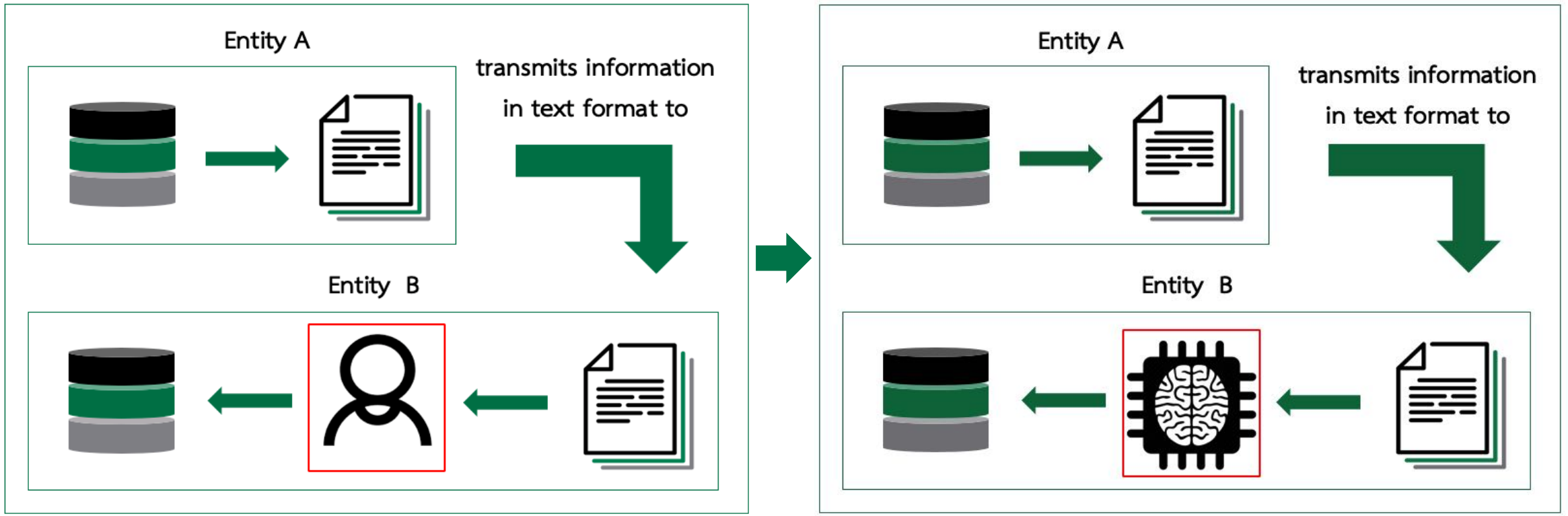

4. Theoretical Background
Gregor Breitschopf, Gerrit Merz und Adrian Pucher


5. Our Approach
Gregor Breitschopf, Gerrit Merz und Adrian Pucher


6. Summary
Gregor Breitschopf, Gerrit Merz und Adrian Pucher

1. Idea & Goal

Idea

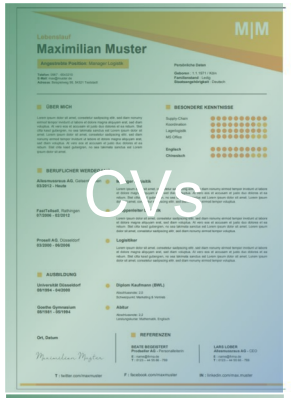

Entity A
transmits information
in text format to
Entity B
Entity A
transmits information
in text format to
Entity B

Goal

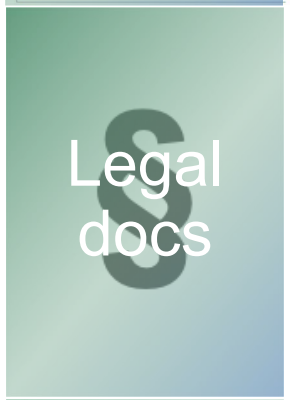
Problem	Solution	Goal
Manual Extraction	ML-Based Approach	Replicate and extend ML-Based Approach to our Use Case
E.g. time-consuming, prone to error	Extracting Contract Elements, Chalkidis et al. (2017)	See chapter 5 „Our Approach“

2. Finding the Right Use Case

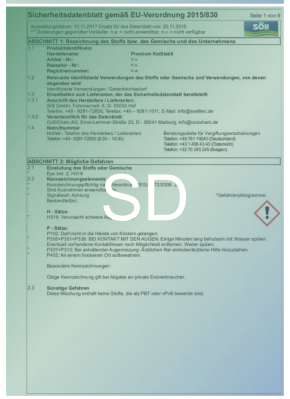
Potential Data Basis



- Decent structured documents
- × Large available data basis
- × No data protection concerns



- × Decent structured documents
- Large available data basis
- No data protection concerns



- Decent structured documents
- Large available data basis
- No data protection concerns

What are Safety Datasheets (SD)?

- SDs…
  - are documents that transmit safety-related information on substances and mixtures
  - are needed if Entity A sells a chemical/substance to Entity B
  - intend to provide the necessary user with the necessary data and handling recommendations when dealing with such substances
- Legal basis for preparation of SDs: REACH Regulation

Source: Leitfaden Sicherheitsdatenblätter, April 2008

How is a Safety Data Sheet structured?

Poliboy — pflegt wie neu

Seite: 1/6

Sicherheitsdatenblatt
gemäß 1907/2006/EG, Artikel 31

Druckdatum: 31.03.2015 Versionsnummer 8 überarbeitet am: 31.03.2015

ABSCHNITT 1: Bezeichnung des Stoffs bzw. des Gemischs und des Unternehmens

· **1.1 Produktidentifikator**
· **Handelsname: BIO Möbel-Reiniger**
· **Artikelnummer:** 03
· **1.2 Relevante identifizierte Verwendungen des Stoffs oder Gemischs und Verwendungen, von denen abgeraten wird**
Keine weiteren relevanten Informationen verfügbar.
· **Verwendung des Stoffes / des Gemisches** Reinigung von Holzoberflächen
· **1.3 Einzelheiten zum Lieferanten, der das Sicherheitsdatenblatt bereitstellt**
· **Hersteller/Lieferant:**
POLIBOY
Brandt & Walther GmbH
Tornéestr. 5
28865 Lilienthal
info@poliboy.de
+49(0)4298-4662-0
· **Auskunftgebender Bereich:** POLIBOY Abteilung Labor Tel.: +49(0)4298-4662-61
· **1.4 Notrufnummer:**
Deutschland:
Giftinformationszentrum-Nord: Tel.: +49(0)551-19240

ABSCHNITT 2: Mögliche Gefahren

· **2.1 Einstufung des Stoffs oder Gemischs**
· **Einstufung gemäß Verordnung (EG) Nr. 1272/2008**
Das Produkt ist gemäß CLP-Verordnung nicht eingestuft.
· **Einstufung gemäß Richtlinie 67/548/EWG oder Richtlinie 1999/45/EG** Entfällt.
· **Besondere Gefahrenhinweise für Mensch und Umwelt:**
Das Produkt ist nicht kennzeichnungspflichtig auf Grund des Berechnungsverfahrens der "Allgemeinen Einstufungsrichtlinie für Zubereitungen der EG" in der letztgültigen Fassung.
· **Klassifizierungssystem:**
Die Klassifizierung entspricht den aktuellen EG-Listen, ist jedoch ergänzt durch Angaben aus der Fachliteratur und durch Firmenangaben.
· **2.2 Kennzeichnungselemente**
· **Kennzeichnung gemäß Verordnung (EG) Nr. 1272/2008** entfällt
· **Gefahrenpiktogramme** entfällt
· **Signalwort** entfällt
· **Gefahrenhinweise** entfällt
· **2.3 Sonstige Gefahren**
· **Ergebnisse der PBT- und vPvB-Beurteilung**
· **PBT:** Nicht anwendbar.
· **vPvB:** Nicht anwendbar.

ABSCHNITT 3: Zusammensetzung/Angaben zu Bestandteilen

· **3.2 Chemische Charakterisierung: Gemische**
· **Beschreibung:** Gemisch aus nachfolgend angeführten Stoffen mit ungefährlichen Beimengungen.
· **Gefährliche Inhaltsstoffe:**

CAS: 68515-73-1 NLP: 500-220-1 Reg.nr.: 01-2119488530-36	Alkylpolyglycosid C8-10 Xi R41 Eye Dam. 1, H318	1-≤2,5%

(Fortsetzung auf Seite 2)
DE

Material Safety Data Sheet **Chem AG**

according to Regulation (EG) No. 1907/2006

Chem AG: Karlsruhe | Product name: SwipeX56

Version number: 2.0 Date: 26.07.2019 Page: 1 out of 6

Chapter 1: Name of the Substance

Subchapter 1.1: Product Name

SwipeX56

Subchapter 1.2: Usage

Cleaning

Chapter 2: Possible Hazards

Signal word: Danger

Chapter 3: Composition

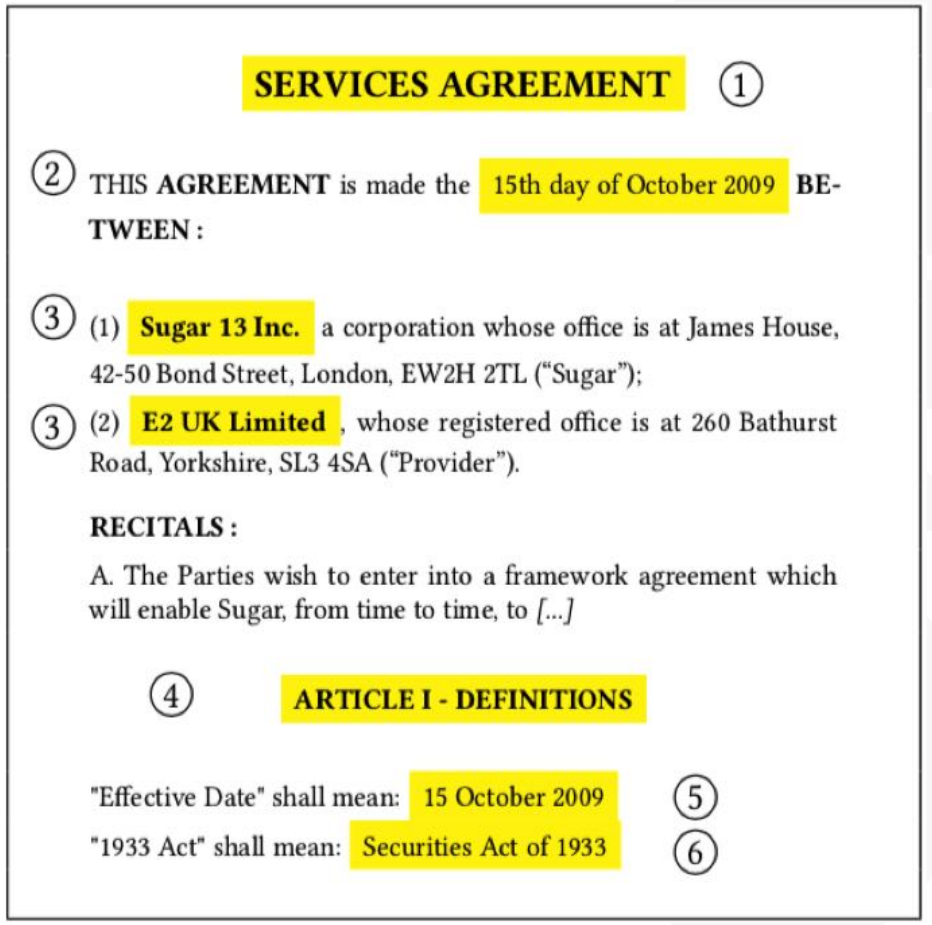
Name	CAS-No.	%-range
Calciumchlorid	1-22691-02-7	20-30%

3. Demo

4. Theoretical Background

Extracting Contract Elements, Chalkidis et al. (2017)

- Automated contract elements extraction with ML models
- Data set: 3.500 English gold contracts



SERVICES AGREEMENT (1)

(2) THIS **AGREEMENT** is made the 15th day of October 2009 **BETWEEN :**

(3) (1) **Sugar 13 Inc.** a corporation whose office is at James House, 42-50 Bond Street, London, EW2H 2TL ("Sugar");

(3) (2) **E2 UK Limited**, whose registered office is at 260 Bathurst Road, Yorkshire, SL3 4SA ("Provider").

RECITALS :

A. The Parties wish to enter into a framework agreement which will enable Sugar, from time to time, to *[...]*

(4) **ARTICLE I - DEFINITIONS**

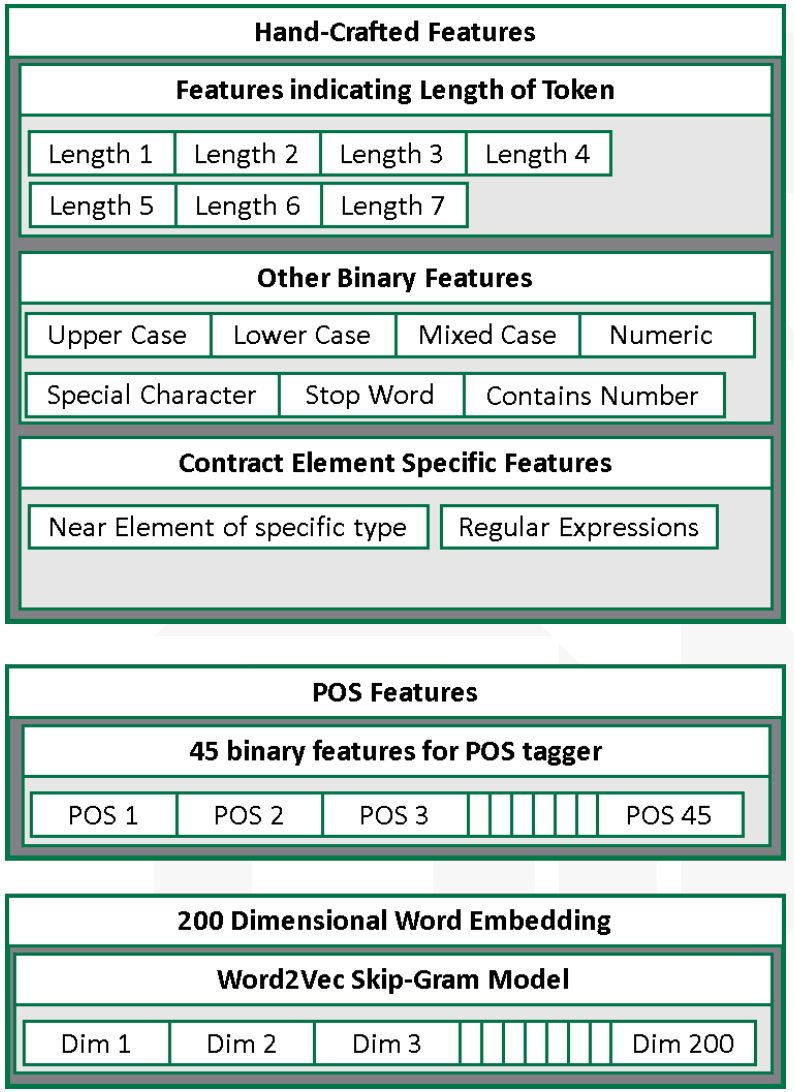
"Effective Date" shall mean: 15 October 2009 (5)

"1933 Act" shall mean: Securities Act of 1933 (6)

Contract Element Type
Title
Parties
Start
Effective
Termination
Period
Value
Gov. Law
Jurisdiction
Legisl. Refs.
Headings

Extracting Contract Elements, Chalkidis et al. (2017)

- Automated contract elements extraction with ML models
- Data set: 3.500 English gold contracts
- Feature groups
  - Hand-Crafted Features
  - POS Features
  - Word Embedding:
    pre-trained with Word2Vec: 750.000 contracts
  - Sliding window for context:
    best window size = 13



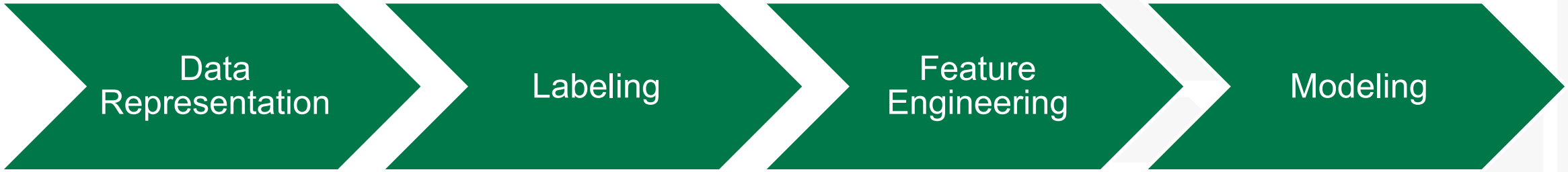
Extracting Contract Elements, Chalkidis et al. (2017)

- Automated contract elements extraction with ML models
- Data set: 3.500 English gold contracts
- Feature groups
  - Hand-Crafted Features
  - POS Features
  - Word Embedding:
    pre-trained with Word2Vec: 750.000 contracts
  - Sliding window for context:
    best window size = 13
- Models: Logistic Regression and SVM

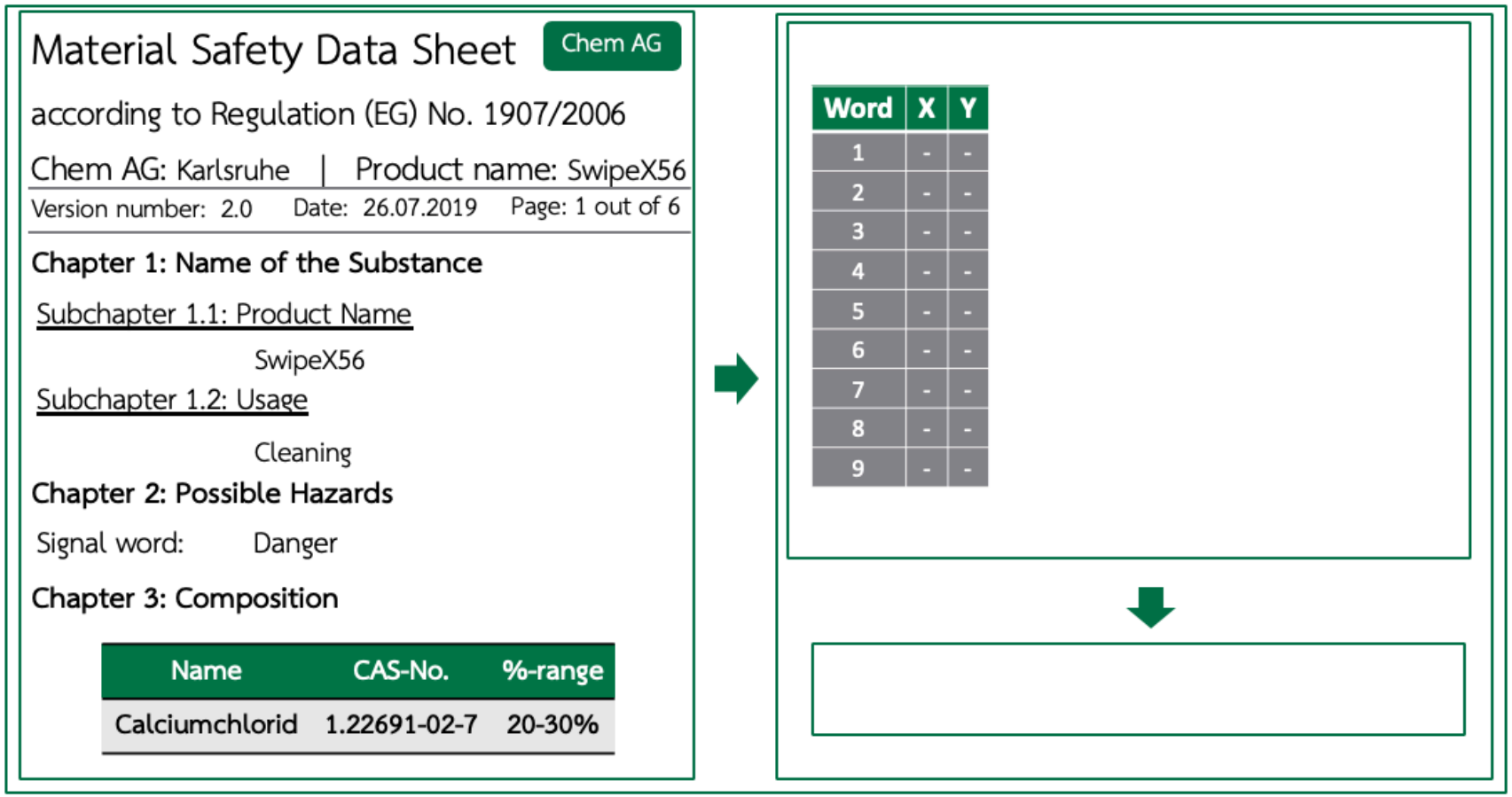
Contract Element Type	SW-LR-ALL			SW-SVM-ALL		
	P	R	F1	P	R	F1
Title	0.91	0.91	0.91	0.91	0.91	0.91
Parties	0.92	0.85	0.89	0.92	0.87	0.89
Start	0.79	0.96	0.87	0.78	0.96	0.86
Effective	0.71	0.63	0.67	0.67	0.79	0.72
Termination	0.68	0.86	0.76	0.54	0.95	0.69
Period	0.61	0.74	0.67	0.55	0.83	0.66
Value	0.70	0.56	0.62	0.68	0.61	0.64
Gov. Law	0.92	0.96	0.94	0.91	0.97	0.94
Jurisdiction	0.86	0.77	0.81	0.82	0.82	0.82
Legisl. Refs.	0.84	0.83	0.83	0.83	0.88	0.86
Headings	0.71	0.92	0.80	0.71	0.92	0.80
Macro-average	0.79	0.82	0.80	0.76	0.86	0.80

5. Our Approach

Overview


Data Representation
Labeling
Feature Engineering
Modeling

First step: Data representation (Principle)



First step: Data representation (Implementation)

PDF Miner: Data representation

```
Set Data: PDF Safety Datasheets
Initialize PDFMiner

For safetysheet in Data:
    For page in safetysheet:
        process page content to get information
        translate information into objects
            For obj in objects:
                For textbox in objects:
                    For textline in textbox:
                        For character in textline:
                            combine characters to words
                            Get x-coordinate
                            Get y-coordinate
                            Get font size
                            Get font type

                            Store information
```

First step: Data representation (Data set)

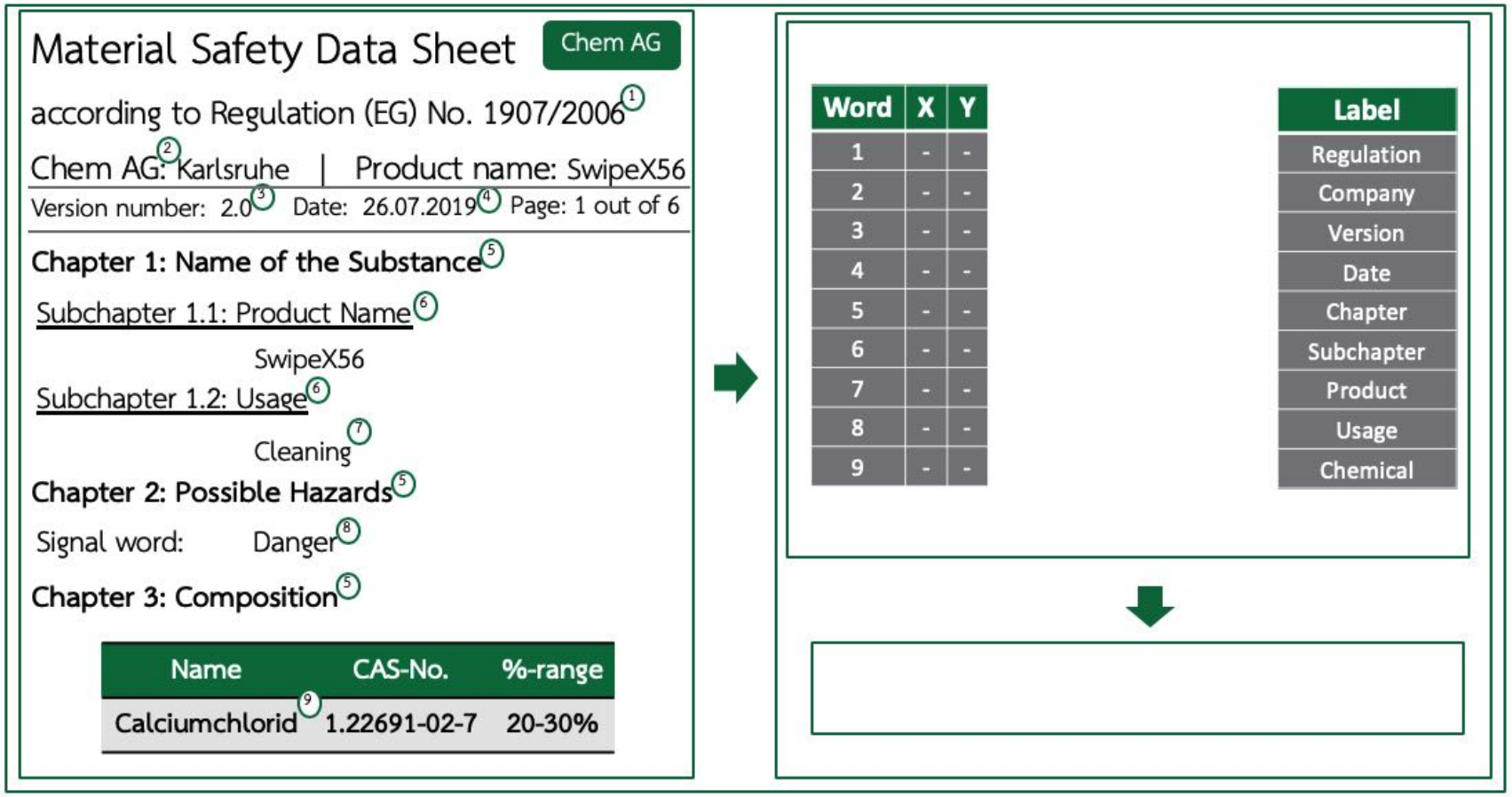
Before Ycord_average_sort

Doc	Page	Ycord_first	Xcord_first	Font_name	word
001_sd.pdf	1	768.29	479.5	CAAAAA+ArialMT	Seite
001_sd.pdf	1	768.29	504.8	CAAAAA+ArialMT	1
001_sd.pdf	1	768.29	513.09	CAAAAA+ArialMT	von
001_sd.pdf	1	768.29	531.8	CAAAAA+ArialMT	6
001_sd.pdf	1	767.446	58.3	BAAAAA+Arial-BoldMT	Sicherheitsdatenblatt
001_sd.pdf	1	767.446	204.74	BAAAAA+Arial-BoldMT	gemäß
001_sd.pdf	1	767.446	253.684	BAAAAA+Arial-BoldMT	EU-Verordnung
001_sd.pdf	1	767.446	360.28	BAAAAA+Arial-BoldMT	2015/830

After Ycord_average_sort

Doc	Page	Ycord_average	Xcord_first	Font_name	word
001_sd.pdf	1	767.868	58.3	BAAAAA+Arial-BoldMT	Sicherheitsdatenblatt
001_sd.pdf	1	767.868	204.74	BAAAAA+Arial-BoldMT	gemäß
001_sd.pdf	1	767.868	253.684	BAAAAA+Arial-BoldMT	EU-Verordnung
001_sd.pdf	1	767.868	360.28	BAAAAA+Arial-BoldMT	2015/830
001_sd.pdf	1	767.868	479.5	CAAAAA+ArialMT	Seite
001_sd.pdf	1	767.868	504.8	CAAAAA+ArialMT	1
001_sd.pdf	1	767.868	513.09	CAAAAA+ArialMT	von
001_sd.pdf	1	767.868	531.8	CAAAAA+ArialMT	6

Second step: Labeling (Idea)



Second step: Labeling (Implementation) – Part 1

Labels with multiple tokens: Company, Chapters, Subchapters, Usecase

```
Data: Tokens
Define Start Words                                  (first words of labels)
Define Stop Words                                   (last words of labels)

For token in Data:
    If token is in Start Words:
        Label next token while Stop Condition is False:
        Stop Condition is True if:
            token is in Stop Words
            token has different format
            token is in new line
```

Second step: Labeling (Implementation) – Part 2

One-Token-Labels (Regex): Reach Regulation

```
Data: Tokens
Define Regular Expressions

For token in Data:
    If token is in Regular Expressions:
        Label token
```

One-Token-Labels (catch word): Version Nr., Signal, Dates

```
Data: Tokens
Define Catch Words

For token in Data:
    If token(-1) is in Catch Words:
        Label token
```

Second step: Labeling (Implementation) – Part 3

CAS number, name and percentage

Data: Tokens
Define Regular Expressions for **CAS numbers**

For token in **Data:**
    If token is in Regular Expressions:
        Label token as CAS number and **save** in CAS List

Use CAS List to **get** List of chemicals
Create Dictionary to map CAS number to chemical names

For token in **Data:**
    If token is in CAS List:
        Inspect context tokens around token:
            If context token is in CAS number – chemical name Dictionary:
                Label token as chemical names
            If context token is giving information about % of chemical:
                Label token as chemical percentage

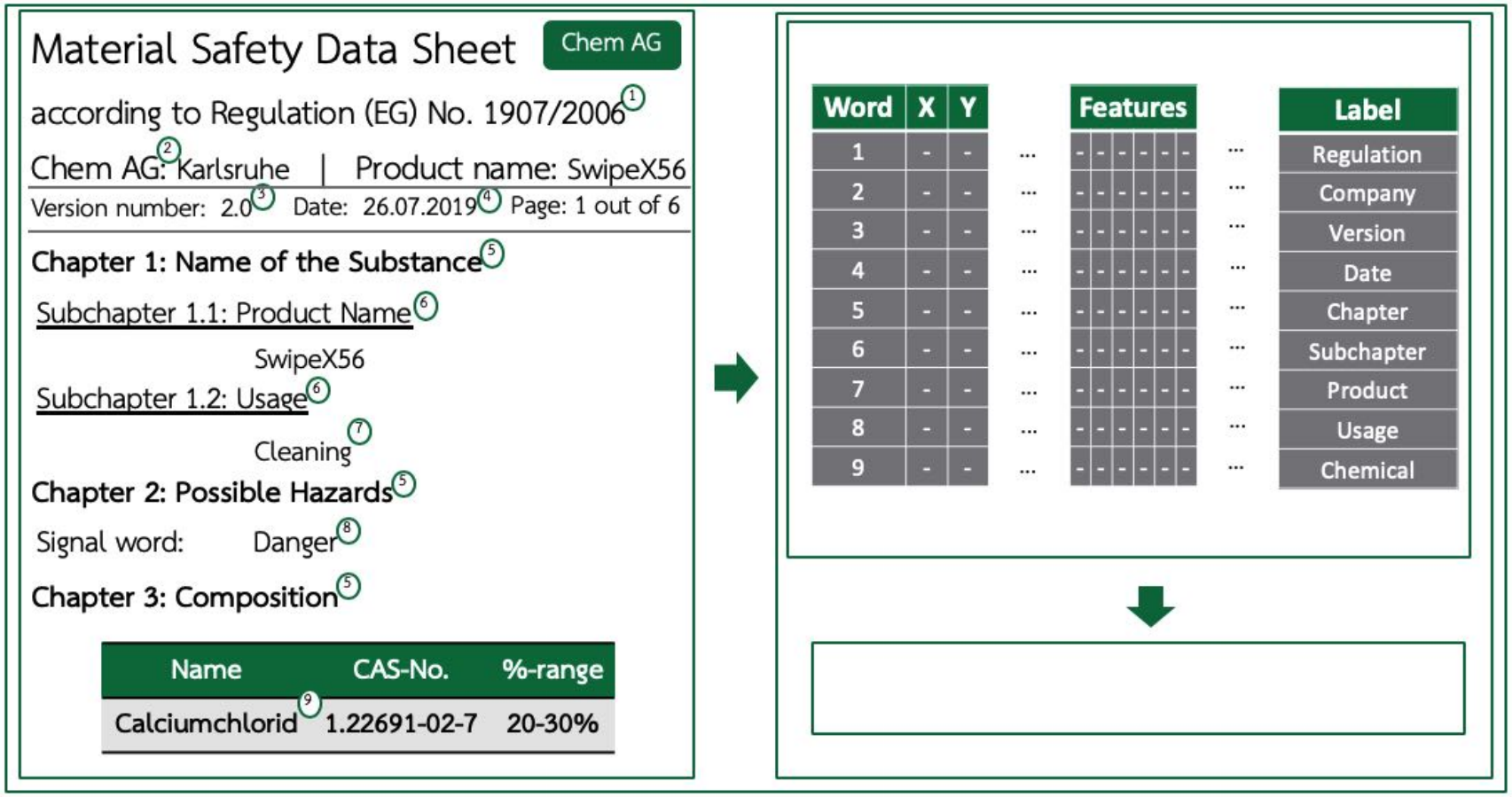
Second step: Labeling (Data set) – Part 1

Word	Label	Regulation	Company	Version No.	Chapter	Subchapter	Usage Pro	Usage Con	Signal Word	Print Date	Revision Date	Validation Date	Old Version Date	Chemical
2015/830	Regulation	1	0	0	0	0	0	0	0	0	0	0	0	0
10.11.2017	Print date	0	0	0	0	0	0	0	0	1	0	0	0	0
2.0	Version	0	0	1	0	0	0	0	0	0	0	0	0	0
ABSCHNITT	Chapter	0	0	0	1	0	0	0	0	0	0	0	0	0
1	Chapter	0	0	0	1	0	0	0	0	0	0	0	0	0
Bezeichnung	Chapter	0	0	0	1	0	0	0	0	0	0	0	0	0
des	Chapter	0	0	0	1	0	0	0	0	0	0	0	0	0
Stoffs	Chapter	0	0	0	1	0	0	0	0	0	0	0	0	0
1.1	Subchapter	0	0	0	0	1	0	0	0	0	0	0	0	0
Produktidentifikator	Subchapter	0	0	0	0	1	0	0	0	0	0	0	0	0
Chem	Company	0	1	0	0	0	0	0	0	0	0	0	0	0
AG	Company	0	1	0	0	0	0	0	0	0	0	0	0	0
Keine	Useage Con	0	0	0	0	0	0	1	0	0	0	0	0	0
Poliermittel	Useage Pro	0	0	0	0	0	1	0	0	0	0	0	0	0
Gefahr	Signal Word	0	0	0	0	0	0	0	1	0	0	0	0	0
Sauerstoff	Chemical	0	0	0	0	0	0	0	0	0	0	0	0	311
7782-44-7	Chemical	0	0	0	0	0	0	0	0	0	0	0	0	321
10%	Chemical	0	0	0	0	0	0	0	0	0	0	0	0	331

Second step: Labeling (Data set) – Part 2

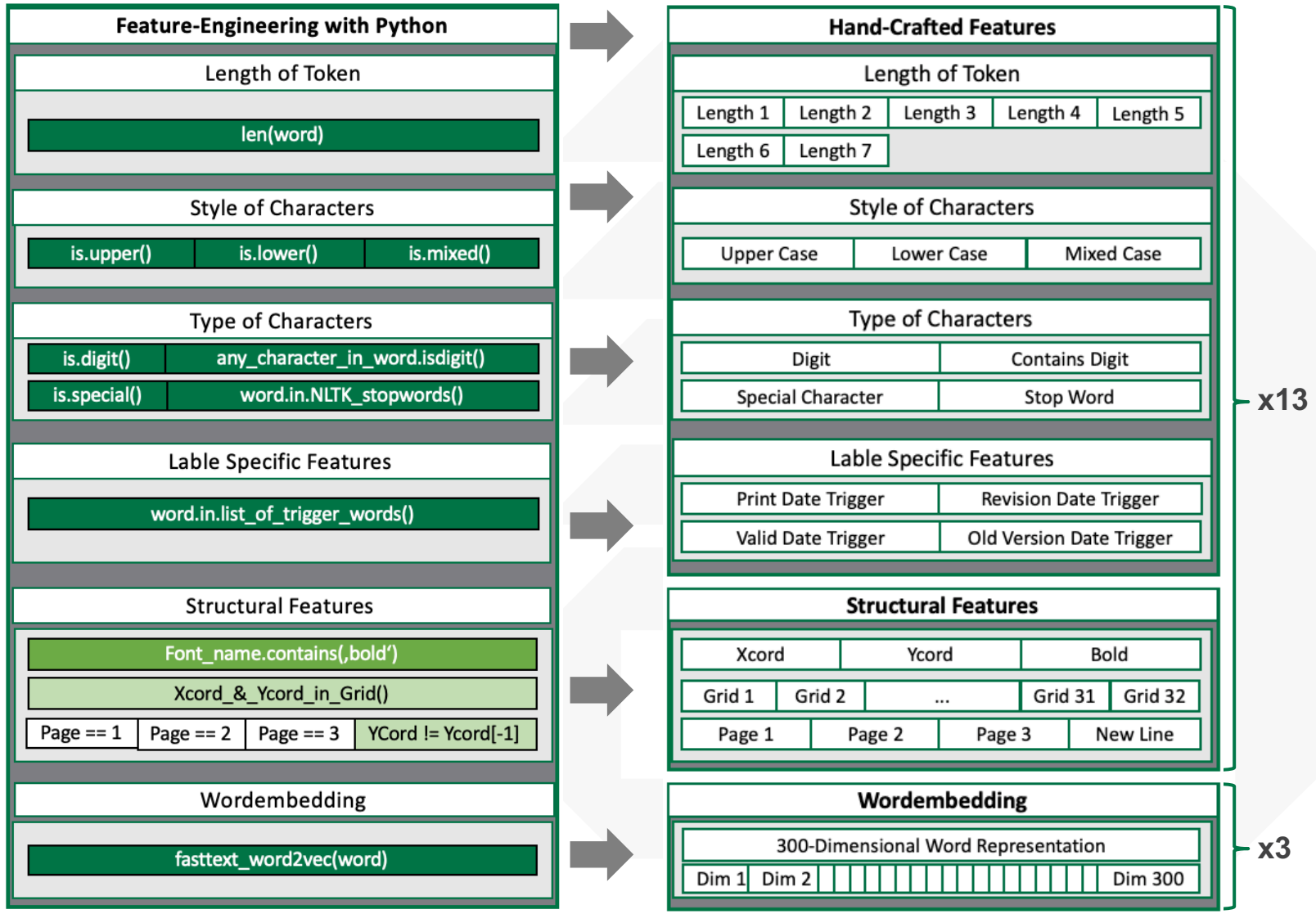
Word	Label	Label Group	Date	Chemical
2015/830	Regulation	1	0	0
10.11.2017	Print date	0	1	0
2.0	Version	2	0	0
ABSCHNITT	Chapter	3	0	0
1	Chapter	3	0	0
Bezeichnung	Chapter	3	0	0
des	Chapter	3	0	0
Stoffs	Chapter	3	0	0
1.1	Subchapter	4	0	0
Produktidentifikator	Subchapter	4	0	0
Chem	Company	5	0	0
AG	Company	5	0	0
Keine	Usage Con	6	0	0
Poliermittel	Usage Pro	7	0	0
Gefahr	Signal Word	8	0	0
Sauerstoff	Chemical	0	0	311
7782-44-7	Chemical	0	0	321
10%	Chemical	0	0	331

Third step: Features (Idea)

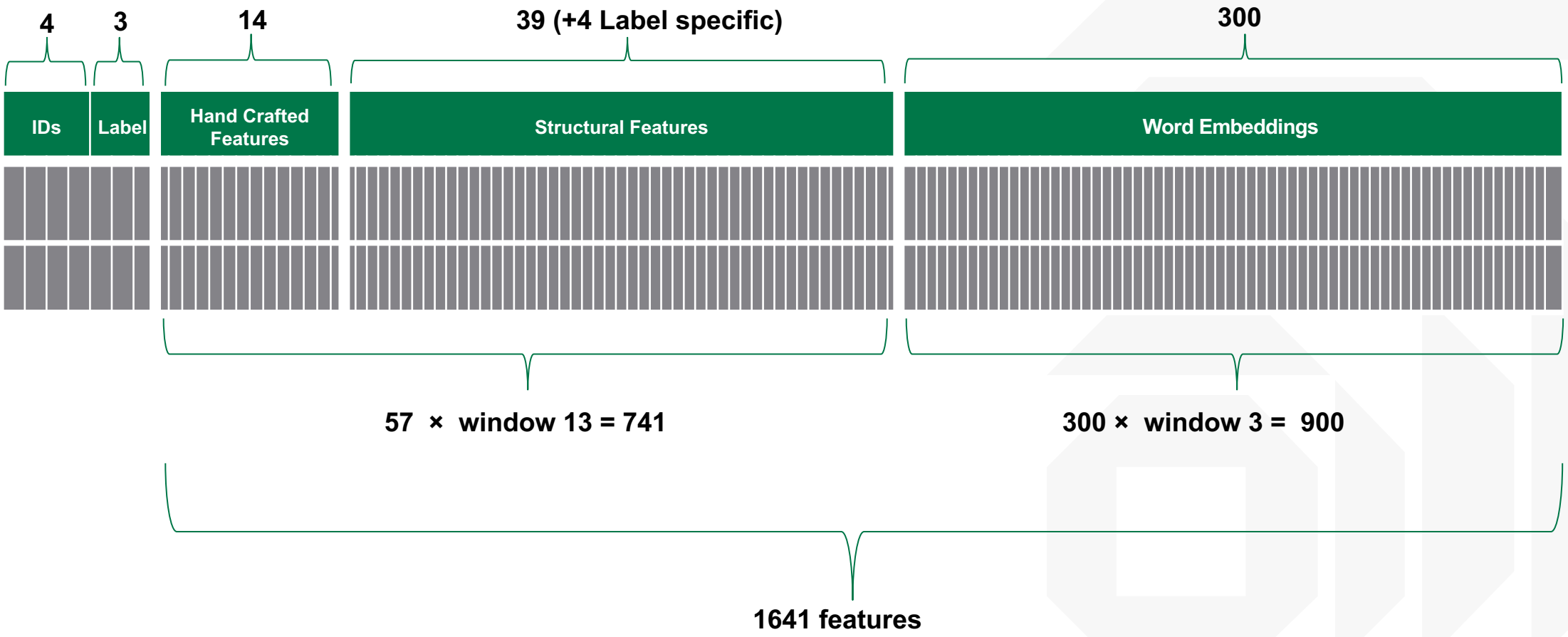


Third step: Features (Implementation)

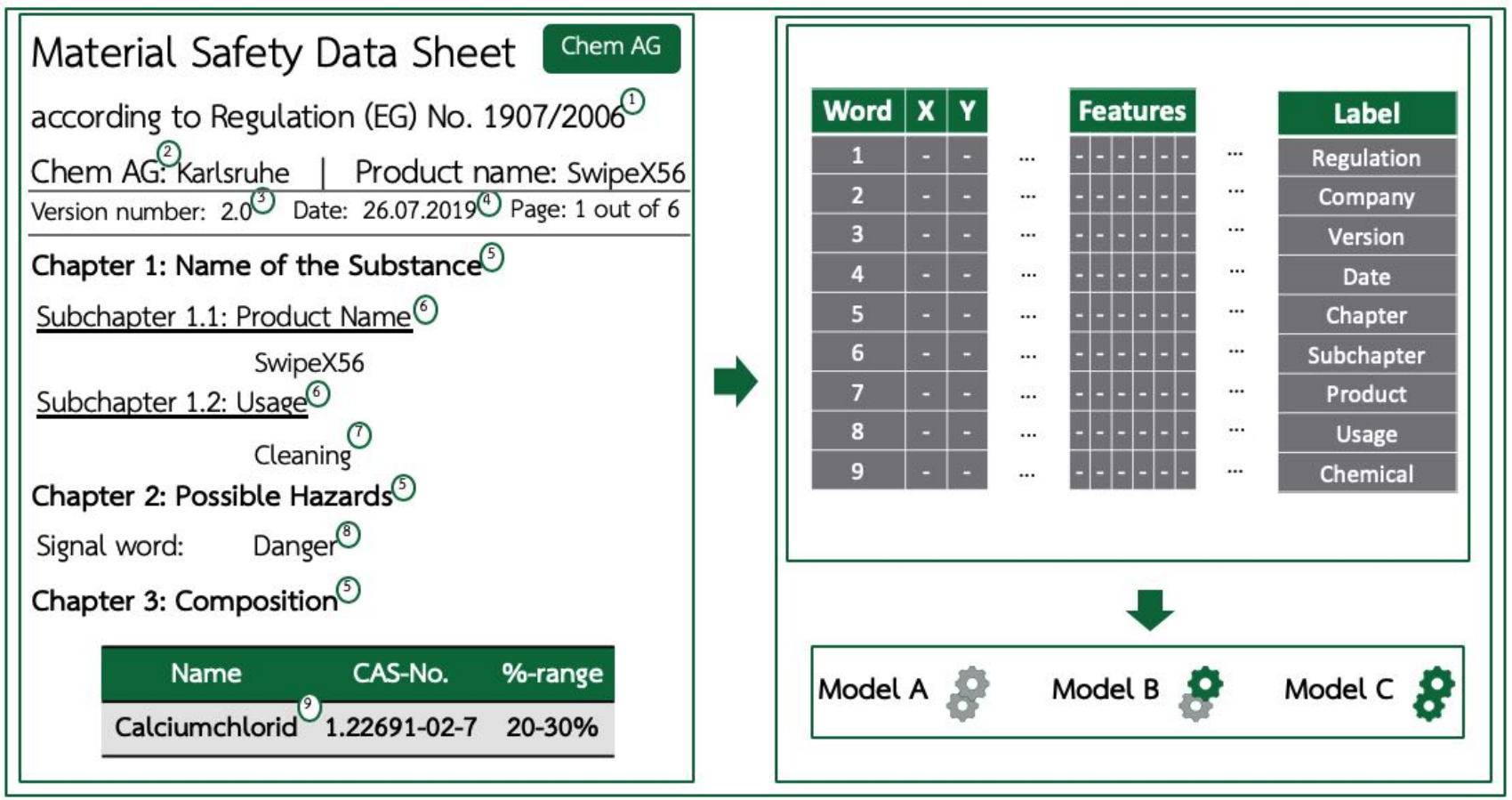
P.	Xcord _first	Ycord _average	Font_name	word
1	58.3	767.868	Arial-BoldMT	Datenblatt
1	204.74	767.868	Arial-BoldMT	gemäß
1	253.684	767.868	Arial-BoldMT	EU-Verordnung
1	360.28	767.868	Arial-BoldMT	2015/830
1	479.5	767.868	ArialMT	Seite
1	504.8	767.868	ArialMT	1
1	513.09	767.868	ArialMT	von
1	531.8	767.868	ArialMT	6



Third step: Features (Data set)



Fourth step: Models (Idea)



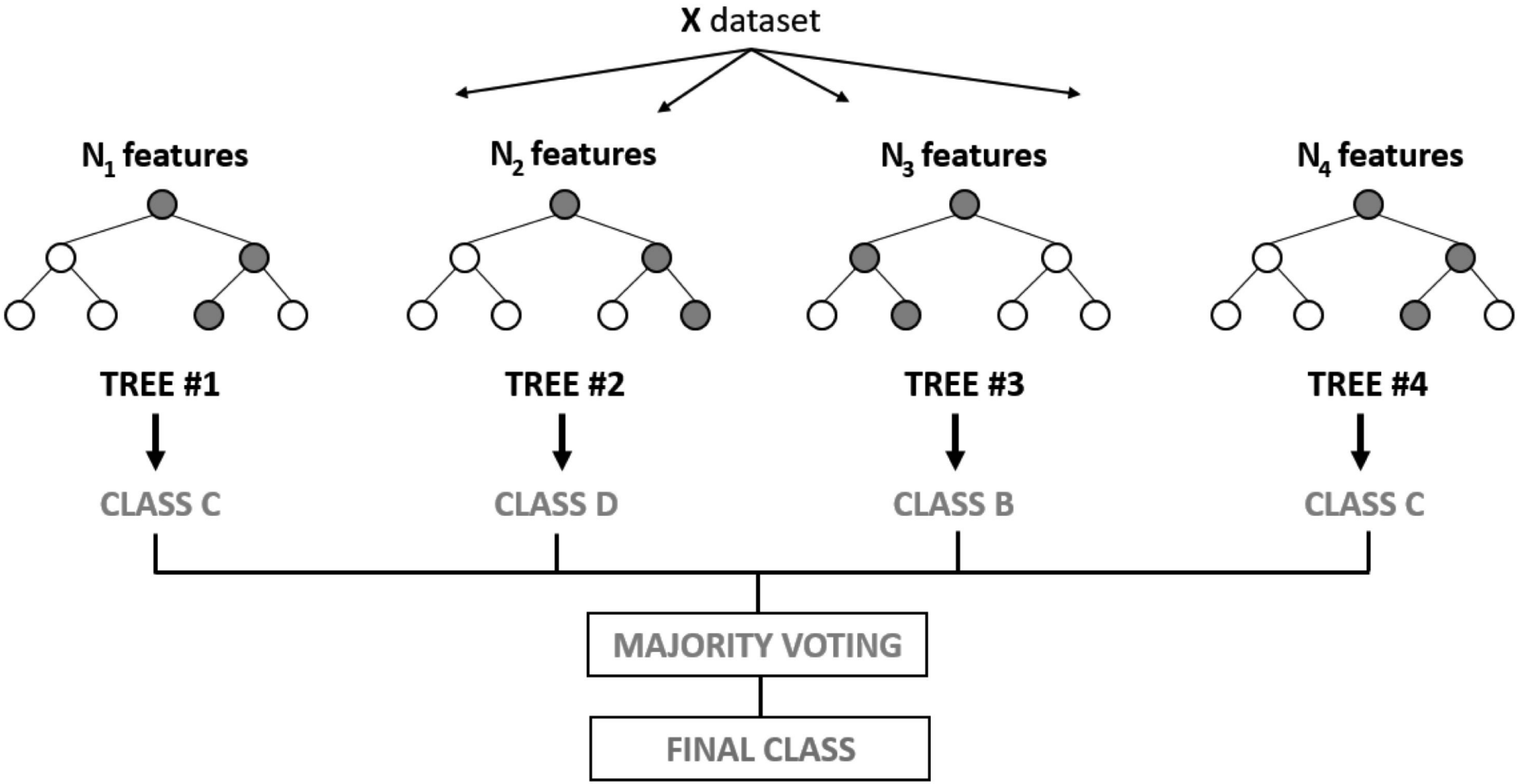
Fourth step: Models (Implementation)

- Classification problem
- Single model for each combination,

$$\sum \textbf{feature\_group} \times \textbf{labelsets} \times \textbf{windowfactor} = (7 \times 2 \times 2) + (6 \times 1 \times 2) + (1 \times 3 \times 2) = 46$$

- Written script for automated feature/label selection, training and testing
- Tested algorithms from Chalkidis et al. (2017):
  - Support Vector Machine
  - Logistic Regression
  - But: Time expensive training on our data set
- Best runtime performance with Random Forest Classifier on our hardware

Fourth step: Models (Implementation)



Data Basis

Sample size	Number of SDs
Size of initial sample	741
Duplicates	-21
Importing problems	-17
Exporting problems	-1
Formatting problems	-175
Final data set	**527**
Training	**395**
Test	**132**

Label	Tokens	Label	Tokens
Chapter	23.501	Usage Con	3.680
Subchapter	24.318	Chemical	6.133
Version No.	1.283	Print Date	443
Regulation	1.113	Revision Date	435
Signal word	439	Validation Date	15
Company	3.081	Old Version Date	59
Usage Pro	421	No Label	473.048
Sum	**537.969**		

Results: No Window – Recall

Labels	HC			ST									WE			ST + HC			ST + WE			HC + WE			ST + HC + WE		
	P	R	F1	P			R			F1			P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Chapter	0,90	0,13	0,23	0,85			0,72			0,78			0,66	0,39	0,49	0,92	0,88	0,90	0,96	0,93	0,94	0,73	0,39	0,51	0,96	0,93	0,95
Subchapter	0,00	0,00	0,00	0,88			0,82			0,85			0,78	0,33	0,47	0,94	0,89	0,92	0,95	0,92	0,93	0,77	0,36	0,49	0,96	0,94	0,95
Version No.	0,00	0,00	0,00	0,91			0,85			0,88			0,00	0,00	0,00	0,95	0,89	0,92	0,93	0,88	0,91	0,00	0,00	0,00	0,94	0,90	0,92
Regulation	0,00	0,00	0,00	0,88			0,86			0,87			0,00	0,00	0,00	0,97	0,94	0,96	0,94	0,90	0,92	0,00	0,00	0,00	0,97	0,94	0,95
Signal word	0,00	0,00	0,00	0,76			0,61			0,68			0,89	1,00	0,94	0,91	0,70	0,79	1,00	0,93	0,96	0,89	1,00	0,94	1,00	0,95	0,97
Company	0,00	0,00	0,00	0,92			0,77			0,84			0,81	0,52	0,64	0,91	0,80	0,85	0,94	0,88	0,91	0,82	0,48	0,61	0,92	0,90	0,91
Usage Pro	0,00	0,00	0,00	0,63			0,26			0,36			1,00	0,05	0,09	0,88	0,28	0,43	0,87	0,47	0,61	0,69	0,07	0,12	0,92	0,48	0,63
Usage Con	0,00	0,00	0,00	0,86			0,72			0,78			0,53	0,14	0,22	0,90	0,75	0,82	0,91	0,78	0,84	0,55	0,14	0,23	0,92	0,78	0,85
Chemical	0,01	0,01	0,01	0,49			0,27			0,33			0,04	0,01	0,02	0,46	0,27	0,33	0,46	0,28	0,33	0,04	0,01	0,02	0,44	0,27	0,32
Print Date	0,00	0,00	0,00	0,94	0,00	0,95	0,86	0,00	0,84	0,90	0,00	0,89	0,00	0,00	0,00	0,94	0,94	0,94	0,93	0,91	0,92	0,00	0,00	0,00	0,93	0,93	0,93
Revision Date	0,00	0,00	0,00	0,96	0,00	0,93	0,85	0,00	0,84	0,90	0,00	0,88	0,00	0,00	0,00	0,93	0,91	0,92	0,90	0,87	0,89	0,00	0,00	0,00	0,93	0,91	0,92
Validation Date	0,00	0,00	0,00	0,50	0,00	0,50	1,00	0,00	1,00	0,67	0,00	0,67	0,00	0,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	0,00	0,00	1,00	1,00	1,00
Old Version Date	0,00	0,00	0,00	1,00	0,00	1,00	0,86	0,00	0,86	0,92	0,00	0,92	0,00	0,00	0,00	1,00	0,86	0,92	1,00	0,86	0,92	0,00	0,00	0,00	0,92	0,86	0,89
Macro-average	0,07	0,01	0,02	0,71	0,45	0,71	0,76	0,49	0,76	0,48	0,33	0,48	0,36	0,19	0,22	0,90	0,78	0,82	0,91	0,82	0,85	0,35	0,19	0,22	0,91	0,83	0,86

HC: Hand-Crafted Features ST: Structural Features WE: Word Embeddings

➡ HC not useful; ST very strong; WE shows solid results with vocab. words

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Results: Window – Recall

Labels	HC			ST									WE			ST + HC			ST + WE			HC + WE			ST + HC + WE		
	P	R	F1	P			R			F1			P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Chapter	1,00	1,00	1,00	0.99			0.91			0.95			0.96	0.96	0.96	1,00	0.96	0.98	1,00	0.99	0.99	0.99	1,00	1,00	1,00	0.99	0.99
Subchapter	0,99	0,82	0,97	0.97			0.88			0.92			0.94	0.87	0.90	1,00	0.92	0.96	0.99	0.96	0.98	0.98	0.96	0.97	0.99	0.96	0.98
Version No.	1,00	0,36	0.94	1,00			0.89			0.94			0.97	0.71	0.82	1,00	0.90	0.95	0.97	0.92	0.94	0.99	0.94	0.96	0.98	0.91	0.94
Regulation	0.73	0.86	0.79	1,00			0.89			0.94			0.69	0.61	0.65	1,00	0.91	0.95	1,00	0.94	0.97	0.73	0.89	0.80	1,00	0.94	0.97
Signal word	1,00	0.86	0,90	1,00			0.81			0.89			1,00	0.98	0.99	1,00	0.82	0.90	1,00	0.98	0.99	1,00	0.97	0.99	1,00	0.97	0.99
Company	0.89	0.99	0.88	0.99			0,81			0.89			0.87	0.94	0.91	1,00	0.86	0.92	1,00	0.96	0.98	0.9	0.95	0.93	1,00	0.95	0.98
Usage Pro	1,00	0.76	0.53	1,00			0.11			0,20			0.96	0.32	0.48	1,00	0.21	0.34	1.0	0.59	0.74	1,00	0.49	0.66	1,00	0.63	0.77
Usage Con	0.98	0.94	0.86	0.98			0.73			0.84			0.91	0.57	0.70	0.99	0.79	0.88	0.97	0.80	0.88	0.96	0.79	0.87	0.99	0.79	0.88
Chemical	0.37	0.22	0.26	0.52			0.28			0.35			0.11	0.03	0.04	0.52	0.28	0.34	0.48	0.28	0.34	0.37	0.23	0.27	0.48	0.29	0.34
Print Date	0.92	0.32	0.48	1,00	1,00	1,00	0.86	0.03	0.98	0.93	0.05	0.94	0.77	0.19	0.30	1,00	0.88	0.94	0.97	0.93	0.95	0.85	0.31	0.45	0.99	0.93	0.96
Revision Date	0.95	0.71	0.81	0.99	0,00	1,00	0.86	0,00	0.83	0.92	0,00	0.91	0.66	0.56	0.60	1,00	0.91	0.95	0.99	0.93	0.96	0.97	0.70	0.81	1,00	0.96	0.98
Validation Date	0,00	0,00	0,00	1,00	0,00	1,00	1,00	0,00	1,00	1,00	0,00	1,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	0.0	0,00	1,00	1,00	1,00
Old Version Date	0.79	0.79	0.79	0.92	0,00	0.92	0.86	0,00	0.86	0.89	0,00	0.89	0,00	0,00	0,00	1,00	0.86	0.92	1,00	0.86	0.92	0.69	0.79	0.73	1,00	0.93	0.96
Macro-average	0.82	0.66	0.71	0.95	0.73	0.95	0.76	0.49	0.77	0.82	0.54	0.82	0.68	0.52	0.57	0.96	0.79	0.85	0.95	0.86	0.90	0.80	0.69	0.73	0.96	0.87	0.90

HC: Hand-Crafted Features ST: Structural Features WE: Word Embeddings

➡ Significant improvements for HC and WE with window

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

6. Summary

Discussion

- High Precision and strong Recall within all models containing Structural Features
- Window leads to improvements for all feature sets, significantly for Hand-Crafted Features and Word Embeddings
- Chemicals as triple labels are hard to detect
- Possible reasons for high scores
  - Very similar structure of the SDs
  - Similar data in Training/Test Split
  - Wanted for cases with a limited vocabulary

Discussion

- Limitations on data set and computing power
  - Quite small amount of TP in the data set: unbalanced data set
  - No self-trained German Word Embedding
  - No computationally expensive models like in our case SVM
- Manual labeling is prone to error
- Strongly paper-oriented approach

Outlook

- Testing with
  - Balanced data set
  - Other models (e.g. convolutional or recurrent neural networks)
  - Parameter tuning
- Extracting other kind of information (e.g. contact person, new labels)
- Transferability to other documents

Takeaways

- **Bottom line: start with a labeled data set**
- Assigning X- and Y-coordinates using PDFMiner also suitable for other PDF documents
- Use Pickle format for storing processed dataframe, CSV gets too slow over time
- Use a combination of Jupyter Notebooks and code editor to implement and debug your code

Project KPIs

- Project hosting and version control: GitHub
  - https://github.com/adrianpu/ML-based-information-extraction-safety-datasheets
- Programming language: Python
- Software: Visual Studio Code, Jupyter Notebooks
- In total: 4 scripts with around 1.900 rows of code
  - Datamining ~ 350 rows
  - Labeling ~ 900 rows
  - Feature Engineering ~ 500 rows
  - Modeling ~ 150 rows


THANK
YOU