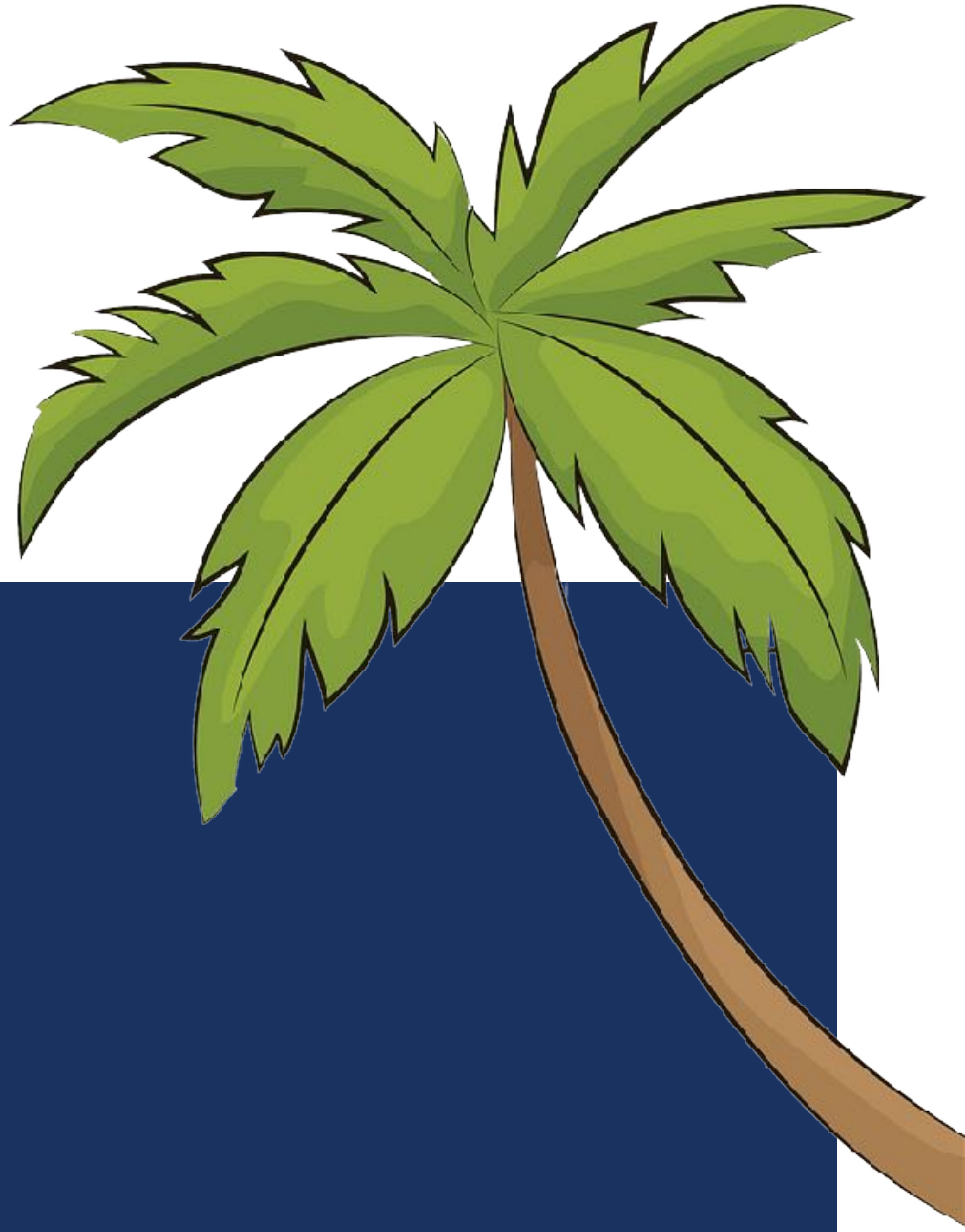
LET'S TALK ABOUT PALM LEAVES FROM MINIMAL DATA TO TEXT UNDERSTANDING

MAGNUS BENDER[1], MARCEL GEHRKE[1], TANYA BRAUN[2]

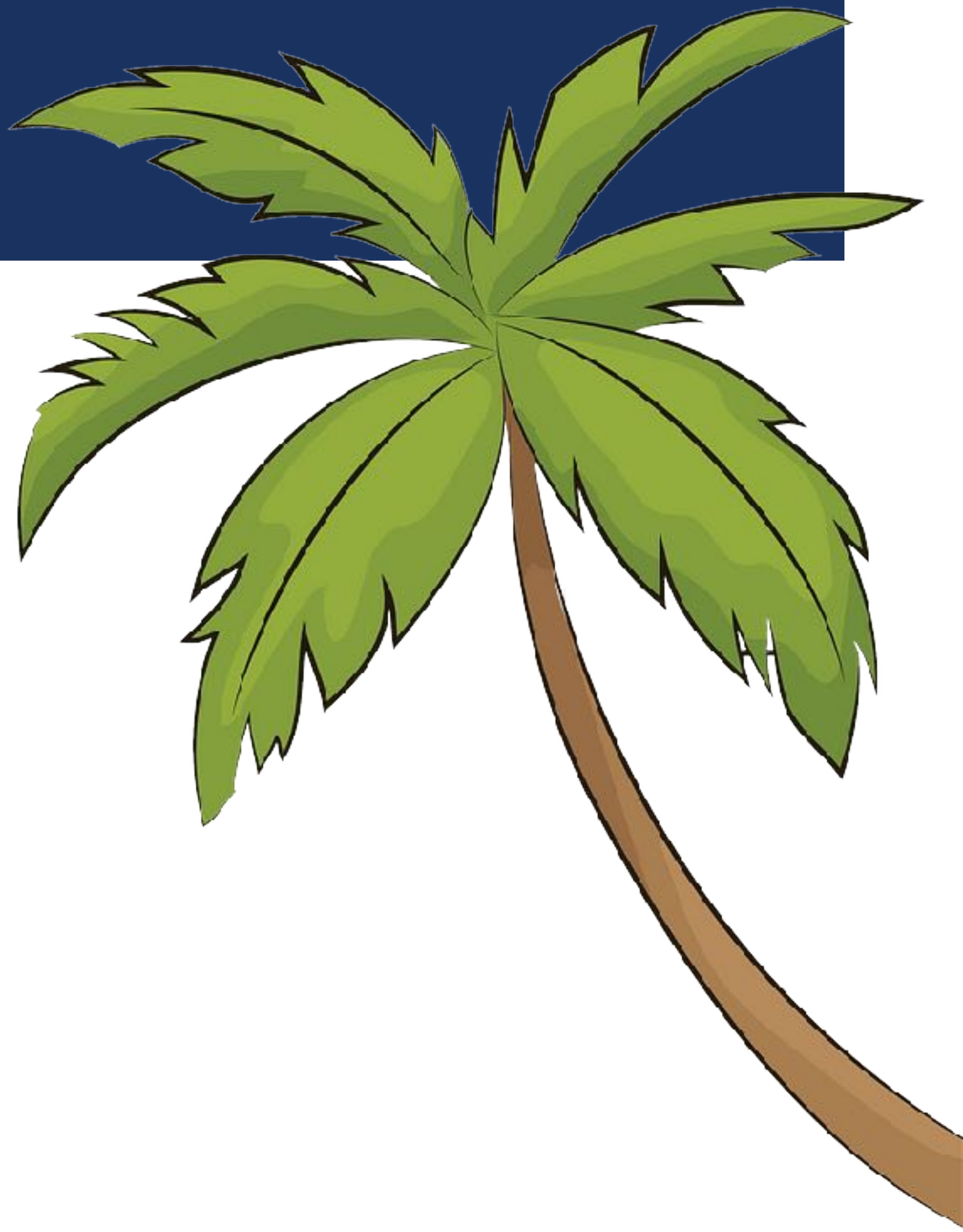


UNIVERSITÄT ZU LÜBECK

[1]Institute of Information Systems, University of Lübeck
[2]Computer Science Department, University of Münster

WWU MÜNSTER

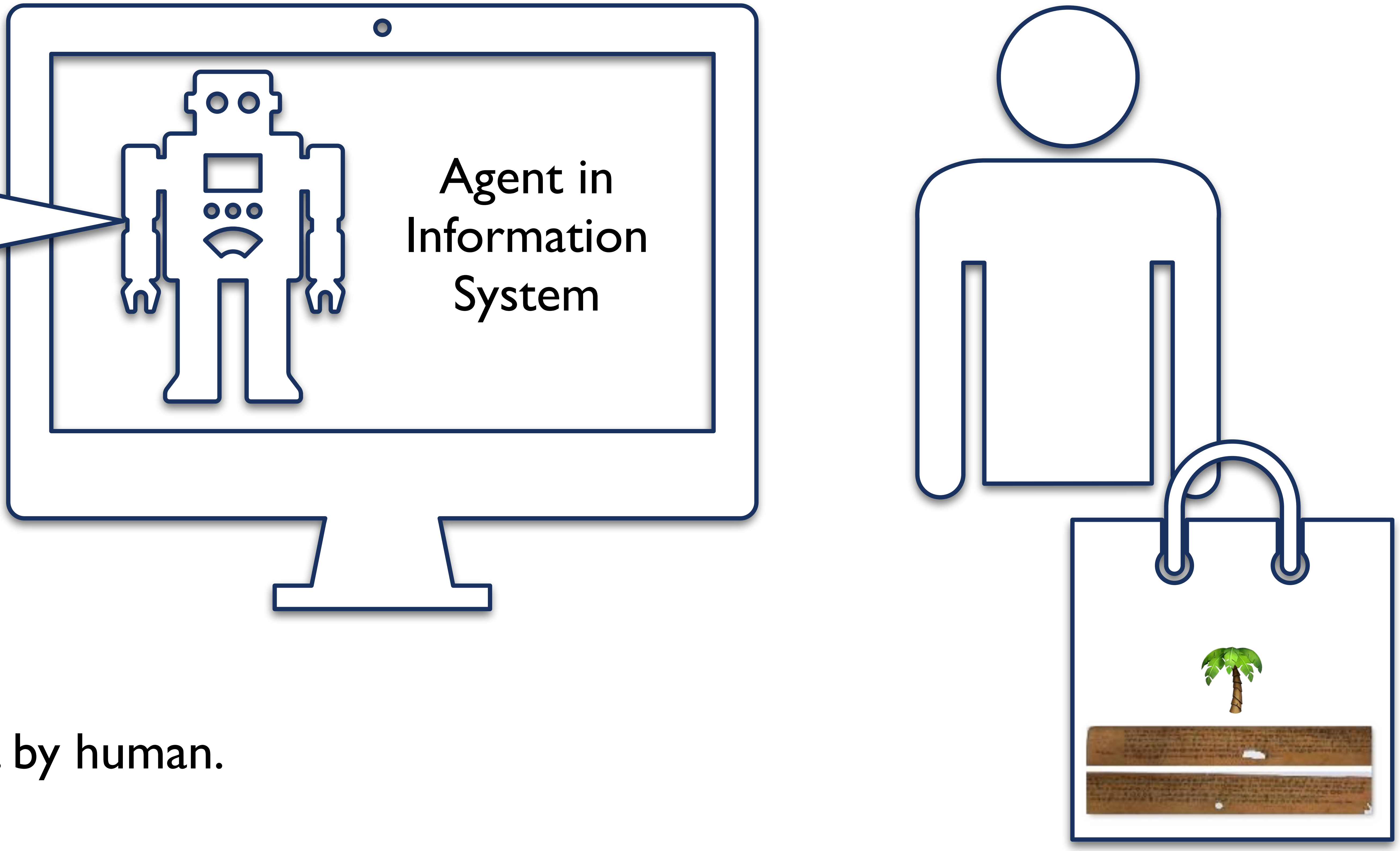
AGENDA



1. Introduction to Semantic Systems [Tanya]
2. Supervised Learning [Marcel]
3. Unsupervised and Relational Learning [Magnus]
   - Unsupervised Estimation of SCDs
   - Continuous Improvement by Feedback
   - Labelling of SCDs
   - Inter- and Intra-SCD Relations
4. Summary [Tanya]

SCENARIO

- Tasks, e.g.,
  - Information Retrieval
  - Corpus Enrichment
- Techniques
  - SCDs



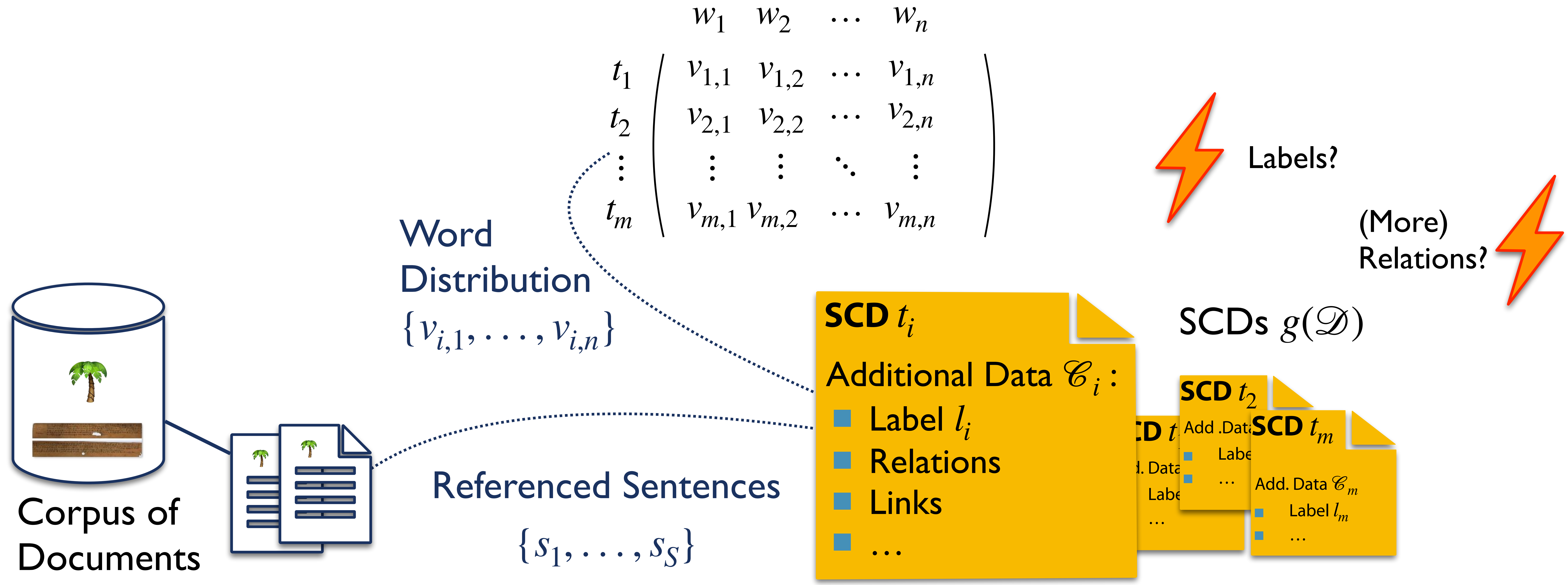
Any corpus brought e.g. by human.

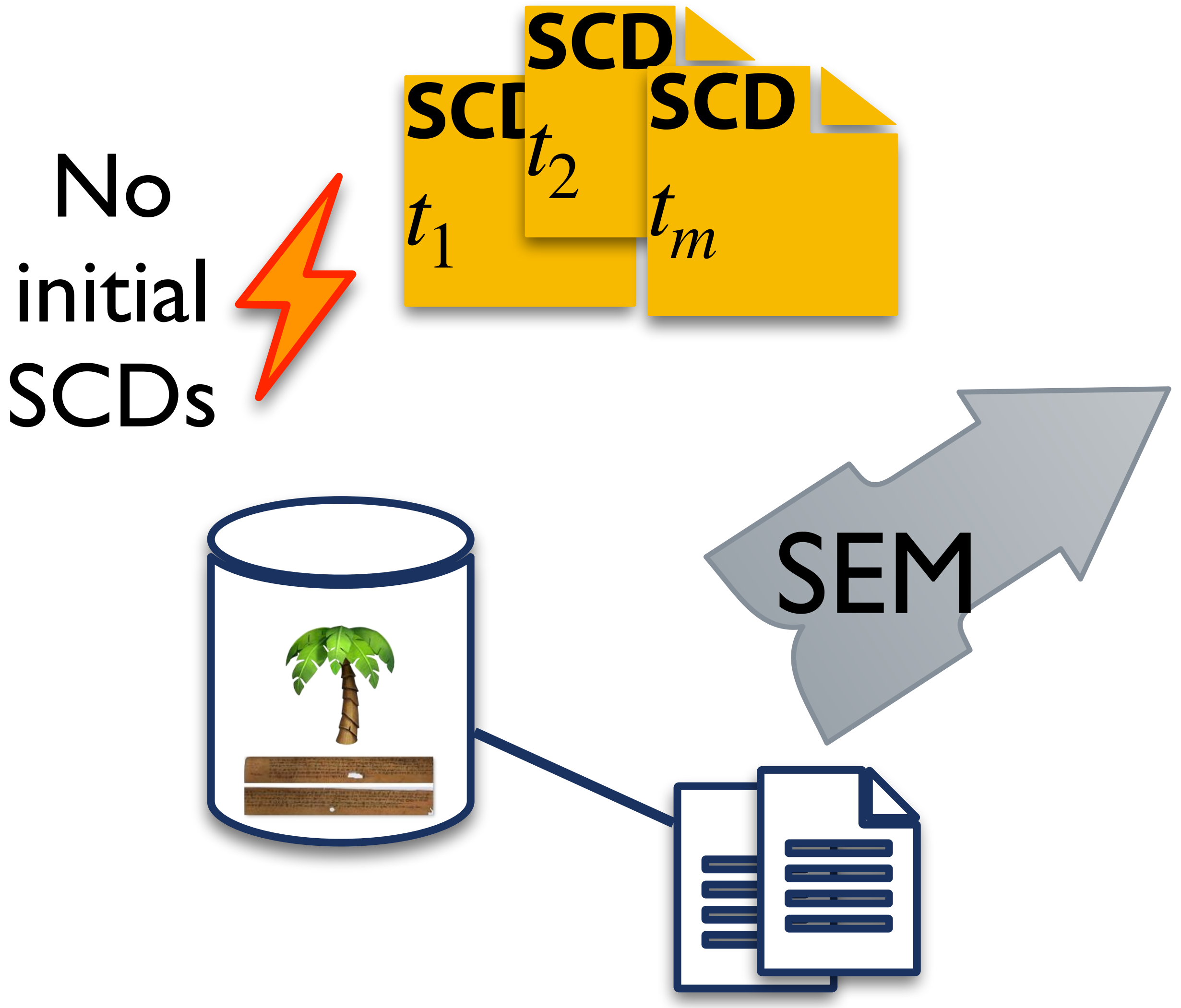
Picture by Eva Wilden, in: Tamil Satellite Stanzas: Genres and Distribution

OVERVIEW

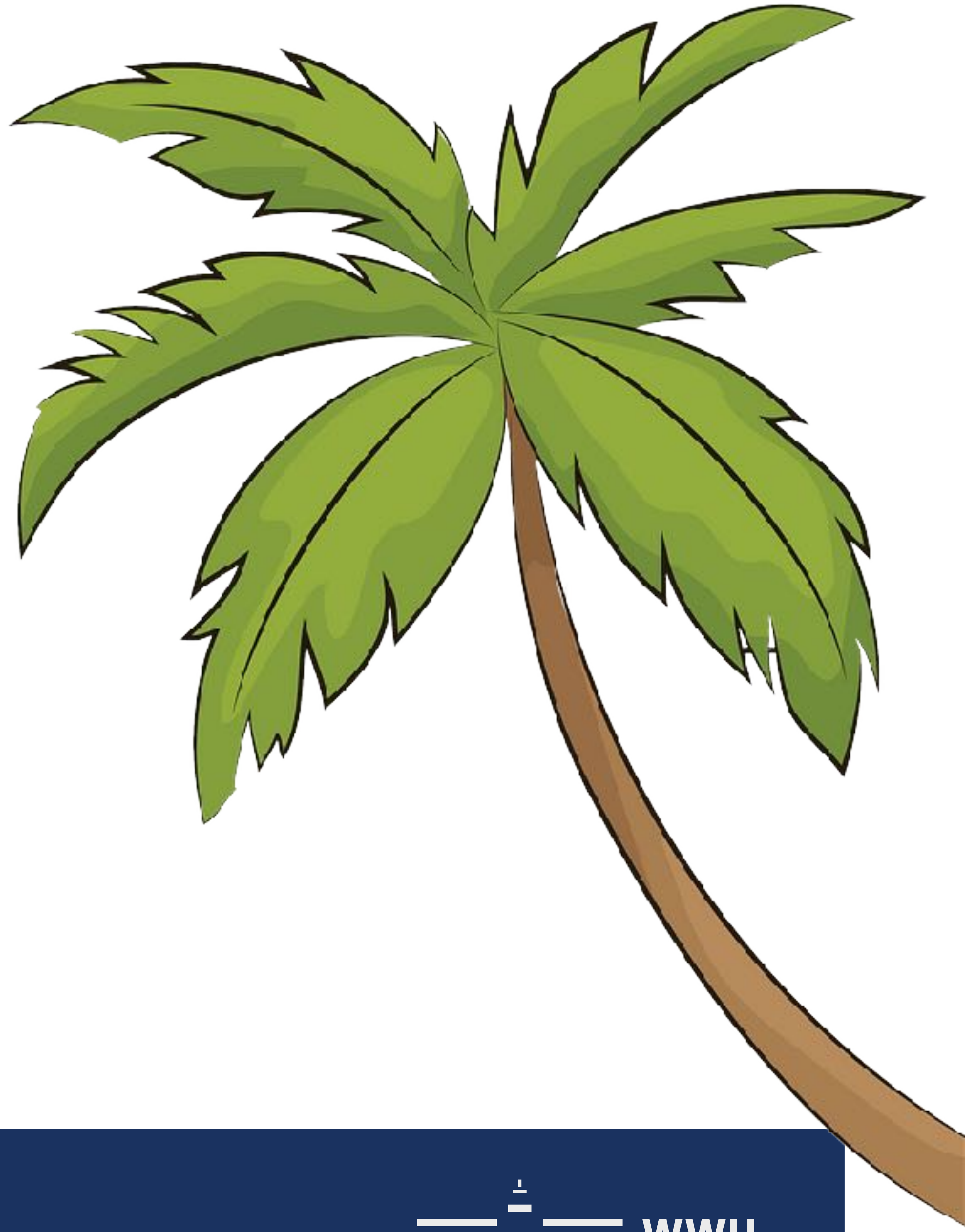
Picture by Eva Wilden, in: Tamil Satellite Stanzas: Genres and Distribution

SCD IN DETAIL





$$\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\ \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix} \end{array}$$



UNSUPERVISED ESTIMATION OF SCDS

USEM – UNSUPERVISED ESTIMATION OF SCD MATRICES

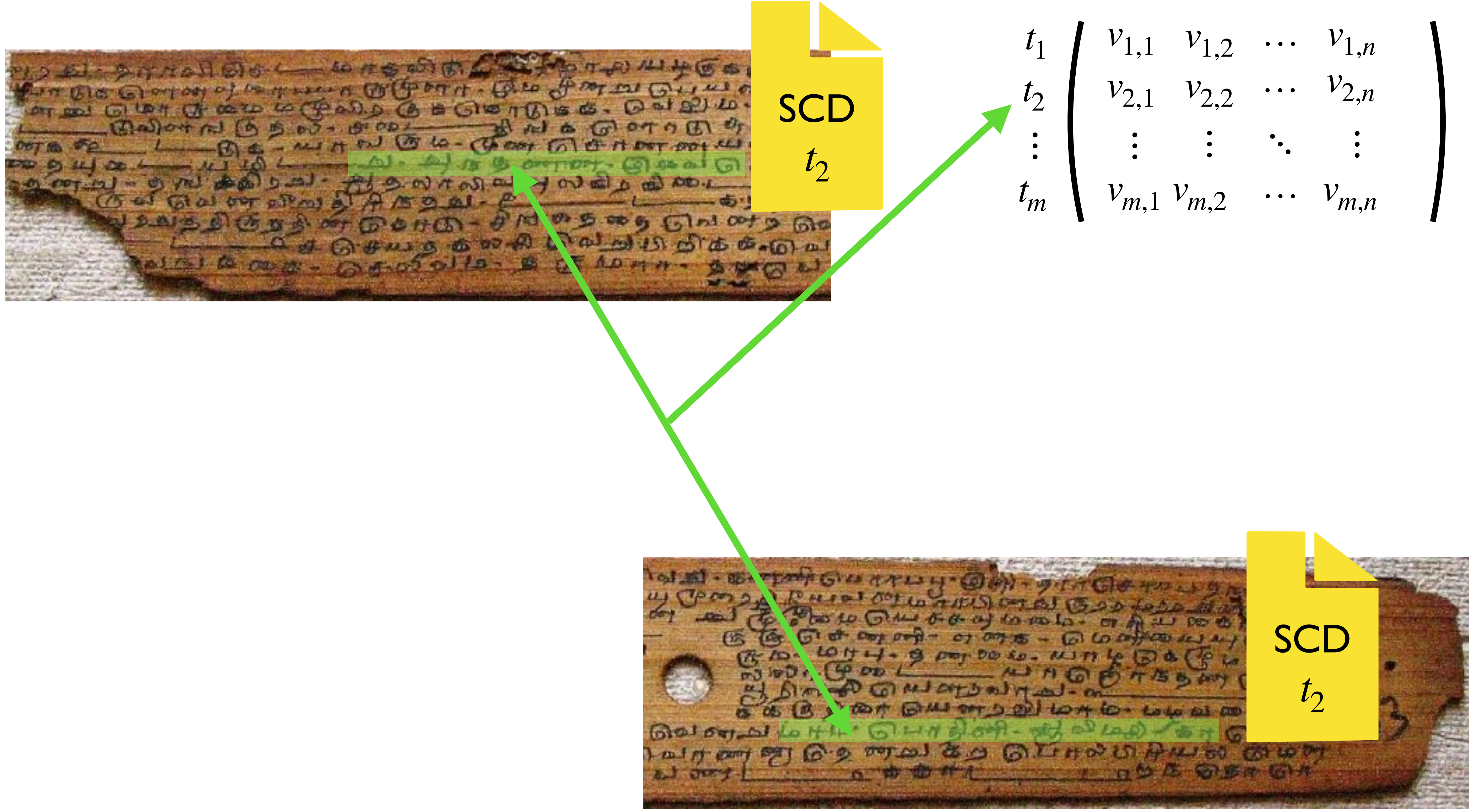
Magnus Bender, Tanya Braun, Ralf Möller, Marcel Gehrke **Unsupervised Estimation of Subjective Content Descriptions** in *17th IEEE International Conference on Semantic Computing (ICSC 2023)*

Magnus Bender, Tanya Braun, Ralf Möller, Marcel Gehrke **Unsupervised Estimation of Subjective Content Descriptions in an Information System** in *International Journal of Semantic Computing, 2023*

UNSUPERVISED ESTIMATION OF SCD MATRICES

- Estimate SCDs in an unsupervised manner
- Focus on identifying similar sentences
- Estimate an SCD matrix
- Select the *best* from multiple matrices

$$
\begin{array}{c@{\;}c}
 & \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\
\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_m \end{array} &
\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix}
\end{array}
$$

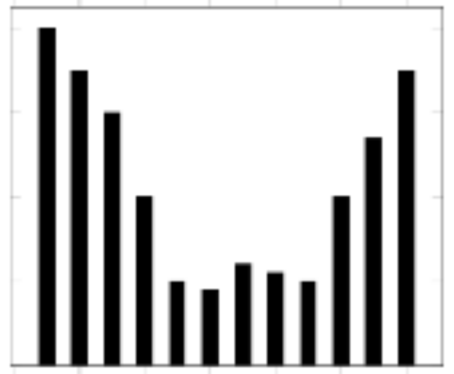


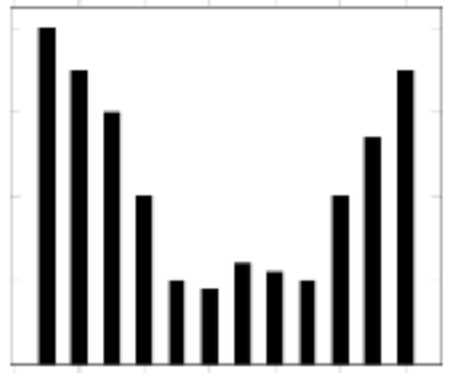
Picture by Eva Wilden, in: Tamil Satellite Stanzas: Genres and Distribution

IDEA: USEM

1. Initially, one SCD for each sentence

$$\begin{array}{cc} & \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\ \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_m \end{array} & \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix} \end{array}$$





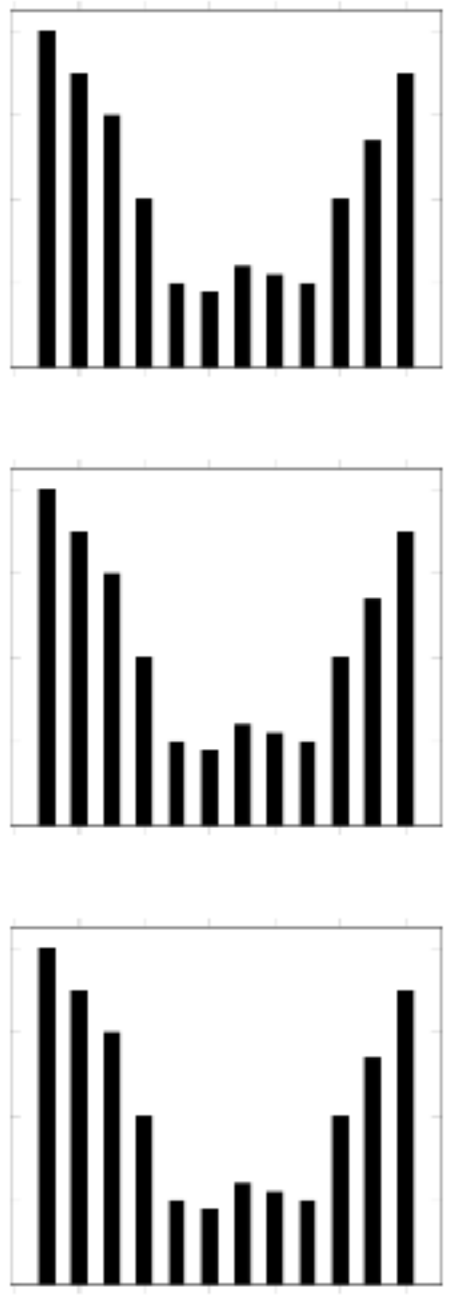
$$d = (\overbrace{w_1, w_2, w_3, w_4, w_5, w_6}, \overbrace{w_7, w_8, w_9, w_{10}, w_{11}}, w_{12}, \ldots, w_N)$$

IDEA: USEM

1. Initially, one SCD for each sentence

2. Identify similar distributions (sentences)
   - Greedy
   - K-Means
   - DBSCAN

$$\begin{array}{c|cccc} & w_1 & w_2 & \cdots & w_n \\ \hline \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_m \end{array} & \multicolumn{4}{c}{\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix}} \end{array}$$



Similar?

Similar?

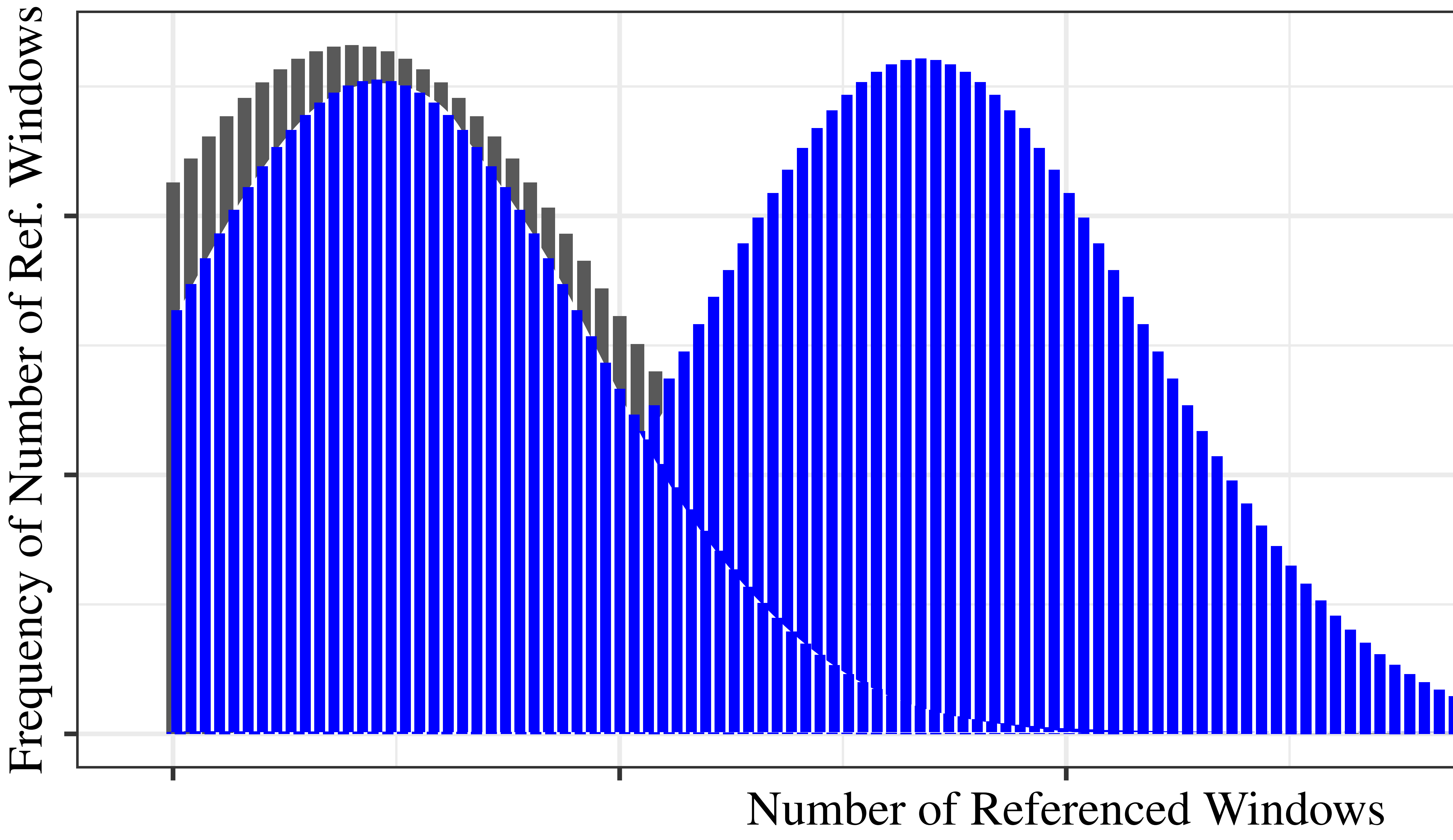
Only K rows, multiple sentences represented in each SCD

$$\begin{array}{c|cccc} & w_1 & w_2 & \cdots & w_n \\ \hline \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_K \end{array} & \multicolumn{4}{c}{\begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{K,1} & v_{K,2} & \cdots & v_{K,n} \end{pmatrix}} \end{array}$$

3. Merge the similar sentences (incrementally)

SCD MATRIX MODEL SELECTION

- **Problem:** Three Methods → multiple matrices
- **Goal:** Identify best hyperparameters for USEM
  - Method (one of DBSCAN, K-Means, Greedy) and
  - Hyperparameters for method
- **Idea:** Run USEM multiple times and choose best resulting matrix
- **Quality Score:** Similarity to optimal histogram depicting the different numbers of windows referenced in an SCD matrix →



Gray assumed optimal, left blue best matrix, right blue weaker matrix.

ALGORITHMS

Algorithm UnSupervised Estimator for SCD Matrices $\delta(\mathcal{D})$

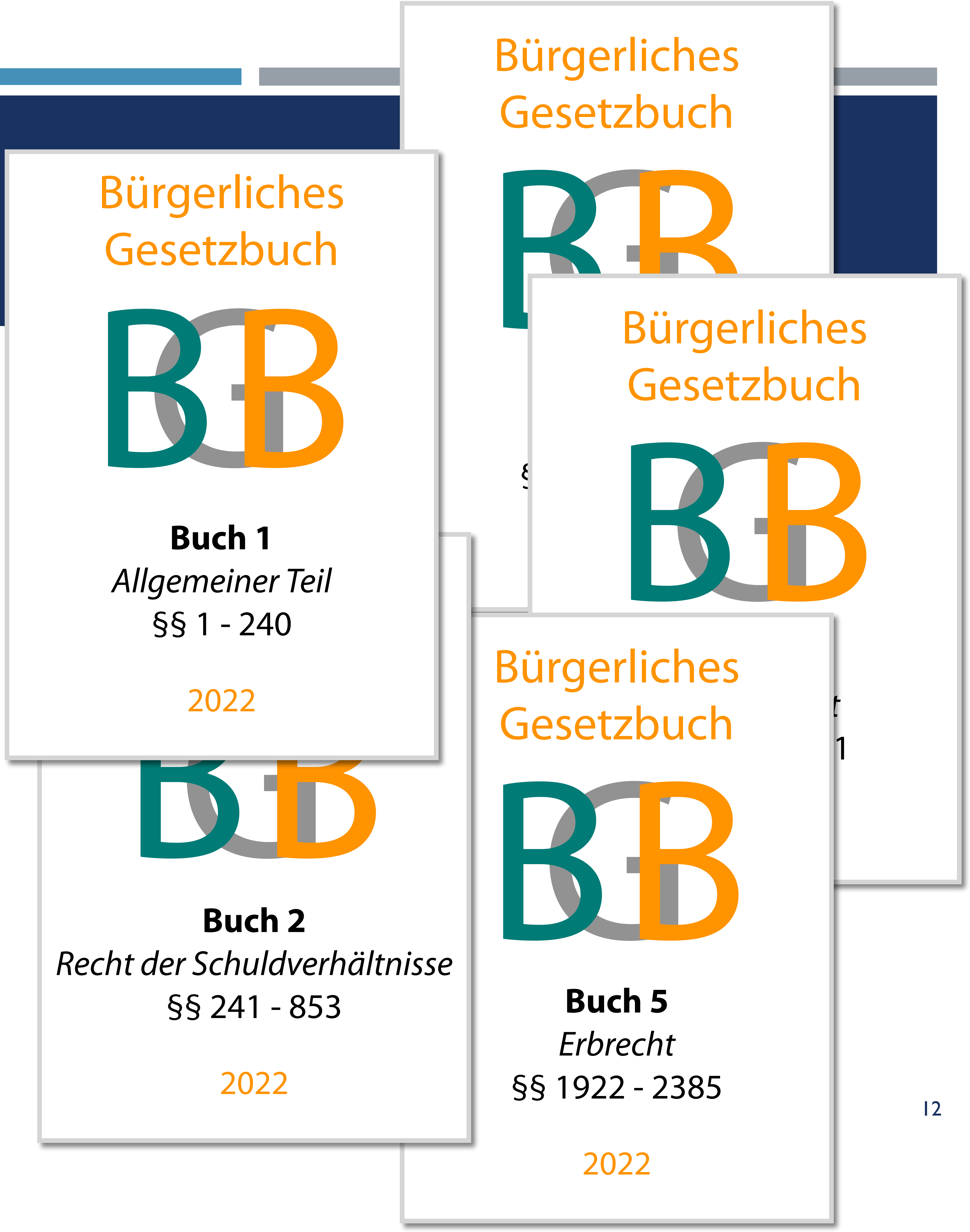
function USEM$(\mathcal{D}, m, [\theta,] [K,] [\varepsilon, ms])$
  Input: Corpus $\mathcal{D}$; Method with hyperparameters, i.e.,
    m = `Greedy` and threshold θ,
    m = `K-Means` and number of SCDs K, or
    m = `DBSCAN`, distance ε, and threshold ms
  Output: SCD-word distribution matrix $\delta(\mathcal{D})$
  Initialize an $(\sum_{d \in \mathcal{D}} M^d) \times L$ matrix $\delta(\mathcal{D})$ with zeros
  $l \leftarrow 0$
  ▷ Build initial SCD matrix
  for each document $d \in \mathcal{D}$ **do**
    for each sentence $s^d \in d$ **do**
      for each word $w_i \in s^d$ **do**
        $\delta(\mathcal{D})[l][w_i] \mathrel{+}= I(w_i, s^d)$
      $l \leftarrow l + 1$
  ▷ Use method m to merge rows
  if m = `Greedy` **then**
    repeat ▷ Detect similar rows and merge
      $(r_i, r_j) \leftarrow$ MOSTSIMILARROWS$(\delta(\mathcal{D}))$
      $\delta(\mathcal{D})[r_i] \leftarrow \delta(\mathcal{D})[r_i] + \delta(\mathcal{D})[r_j]$ ▷ Sum rows
      $\delta(\mathcal{D})[r_j] \leftarrow Nil$ ▷ Delete row
    until SIMILARITY$(r_i, r_j) < \theta$
  else ▷ Create clusters of similar rows
    if m = `K-Means` **then**
      $clusters \leftarrow$ KMEANS$(\delta(\mathcal{D}), K)$
    else
      $clusters \leftarrow$ DBSCAN$(\delta(\mathcal{D}), \varepsilon, ms)$
    for each cluster $c \in clusters$ **do**
      ▷ Create sum of all cluster's rows in first row
      $r_i \leftarrow$ FIRSTROW(c)
      $\delta(\mathcal{D})[r_i] \leftarrow \sum_{r_j \in c} \delta(\mathcal{D})[r_j]$
      for each row $r_j \in c$ **do**
        if $r_i \neq r_j$ **then** ▷ Delete all non-first rows
          $\delta(\mathcal{D})[r_j] \leftarrow Nil$
  return $\delta(\mathcal{D})$

Algorithm SCD Matrix Model Selection

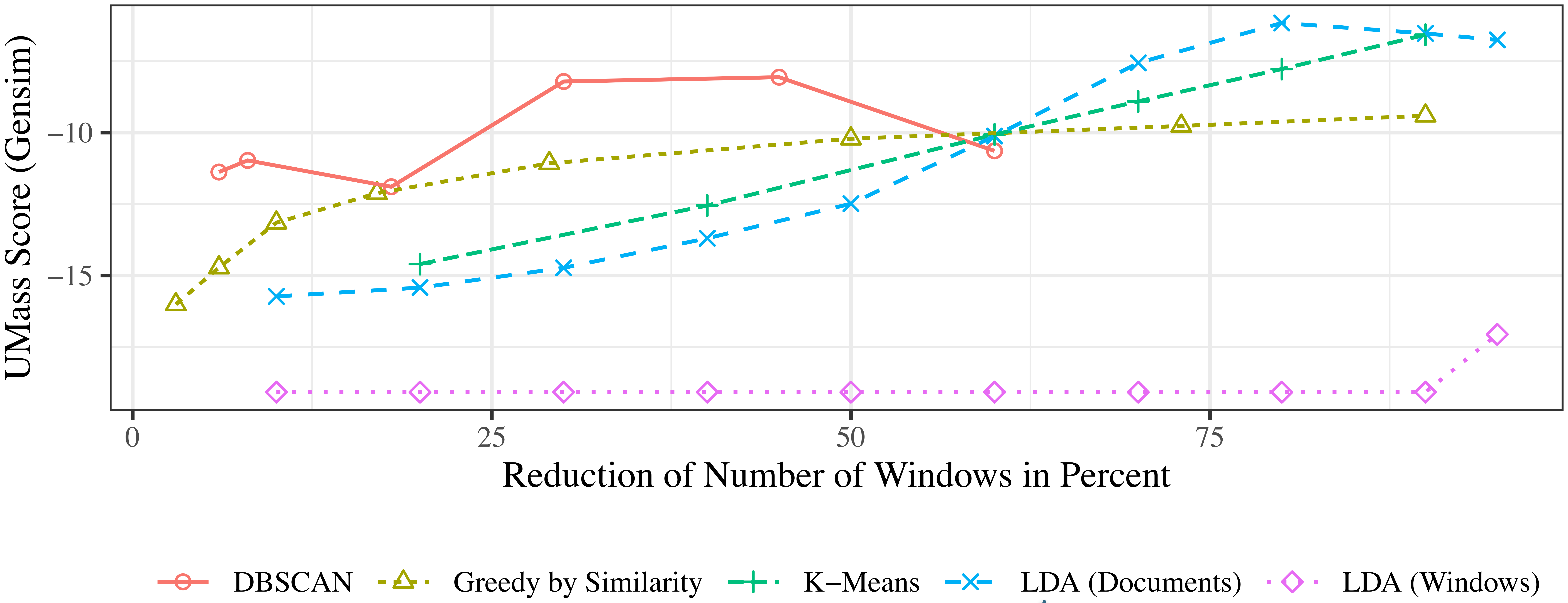
function ESTIMATEBESTMATRIX$(\mathcal{D})$
  Input: Corpus $\mathcal{D}$
  Output: *Best* SCD-word distribution matrix $\delta(\mathcal{D})$
  $sim_{best} \leftarrow 0$
  $\delta_{best} \leftarrow Nil$ ▷ Iterate all methods
  for each method $m \in$ {`Greedy`, `K-Means`, `DBSCAN`} **do**
    ▷ Take a set of hyperparameters depending on method
    if m = `Greedy` **then**
      $H \leftarrow (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$ ▷ Values of θ
    if m = `K-Means` **then**
      $M' \leftarrow \sum_{d \in \mathcal{D}} M^d$ ▷ Number of sentences in $\mathcal{D}$ to calculate Ks
      $H \leftarrow (\lfloor M' \cdot 0.8 \rfloor, \lfloor M' \cdot 0.6 \rfloor, \lfloor M' \cdot 0.4 \rfloor, \lfloor M' \cdot 0.3 \rfloor, \lfloor M' \cdot 0.2 \rfloor, \lfloor M' \cdot 0.1 \rfloor)$
    else
      $H \leftarrow ((0.3, 1), (0.5, 10), (0.5, 5), (0.5, 2), (0.7, 10))$ ▷ Tuples of ε, ms
    for each hyperparameter $h \in H$ **do**
      $\delta(\mathcal{D}) \leftarrow$ USEM$(\mathcal{D}, m, h)$ ▷ Run USEM
      ▷ Calculate score using Hellinger distance to normal distribution
      $sim \leftarrow 1 - \mathrm{HD}(\mathrm{SCALE}([0, 100], \delta(\mathcal{D})), \mathcal{N}([0, 100], \mu = 10, \sigma^2 = 15))$
      if $sim > sim_{best}$ **then**
        $sim_{best} \leftarrow sim$
        $\delta_{best} \leftarrow \delta(\mathcal{D})$
  return δ_{best}

EVALUATION & EXAMPLE: DATASET

- Bürgerliches Gesetzbuch (BGB)
  - German civil code (German language)
- Why BGB?
  - Easily to process
  - Uniform style of writing
- Identify and present similar paragraphs
  - Compare USEM to LDA topic model
- Only example for a corpus



USEM VS. LDA



Well-know competitor

Provides SCD matrix

USAGE EXAMPLE

„An association whose purpose is not directed towards a commercial business operation acquires legal capacity through entry in the register of associations at the competent local court..“

„An association whose purpose is to engage in commercial business shall acquire legal capacity, in the absence of special federal law, through state conferral. The grant is due to the state in whose territory the association has its registered office.“

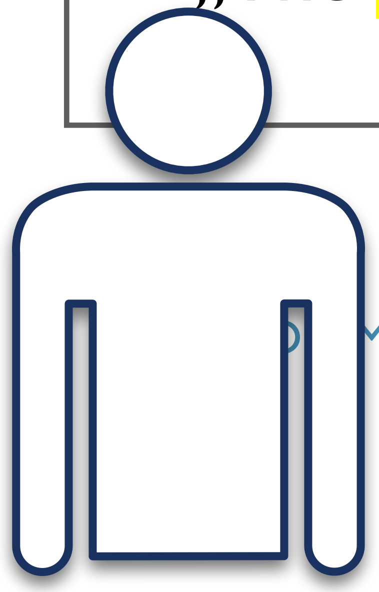
„The seat of an association, unless otherwise provided, is the place where the administration is conducted.“

„The seat of a foundation, unless otherwise provided, is the place where the administration is conducted.“

Find similar sentences in SCD matrix

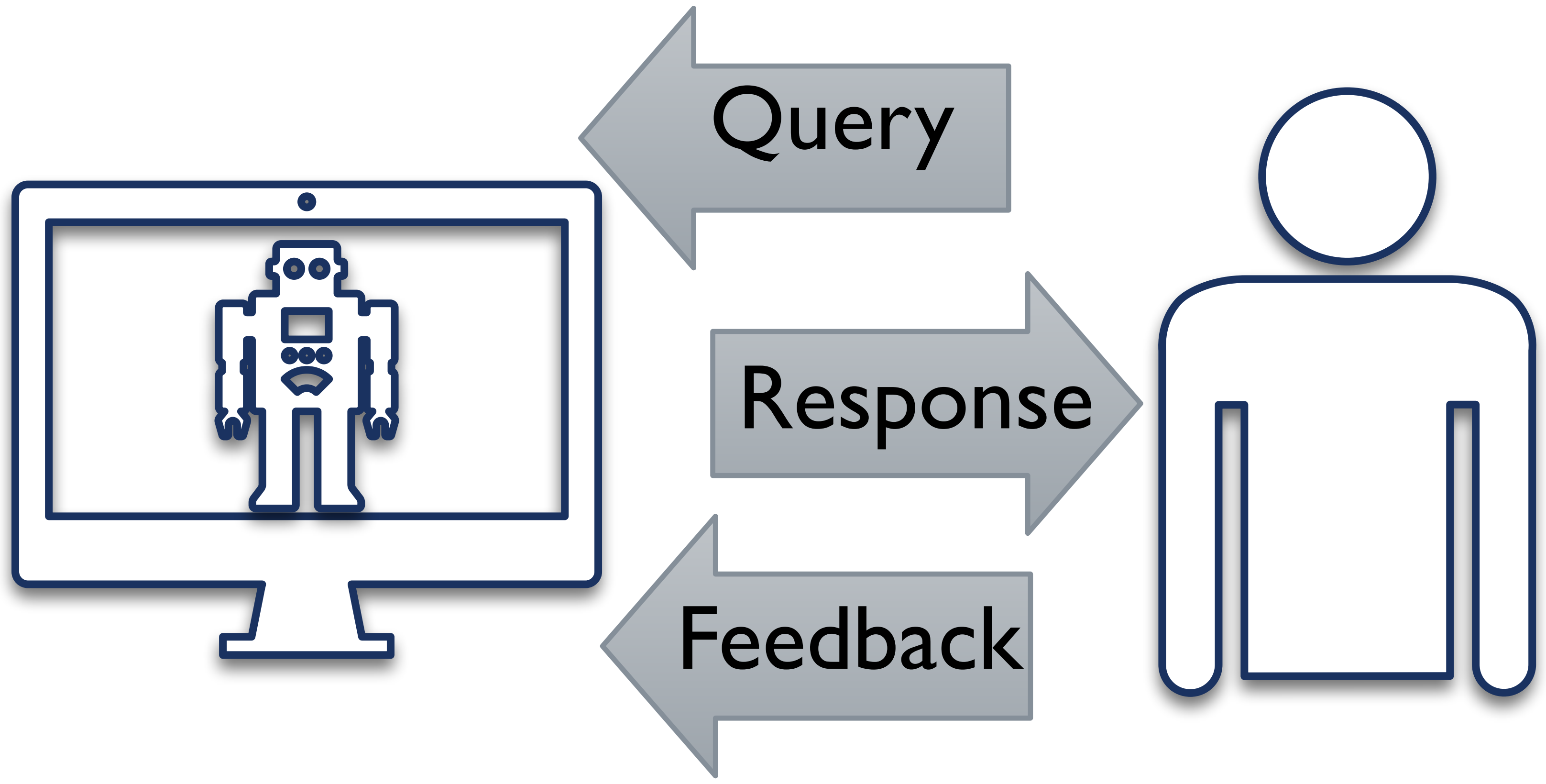
$$
\begin{array}{c c}
 & \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\
\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_m \end{array} & \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix}
\end{array}
$$

Find sentences referenced by same SCD



USEM

§§ 21, 22, 24, 83 BGB (https://www.gesetze-im-internet.de/bgb/)



How to incorporate Feedback?

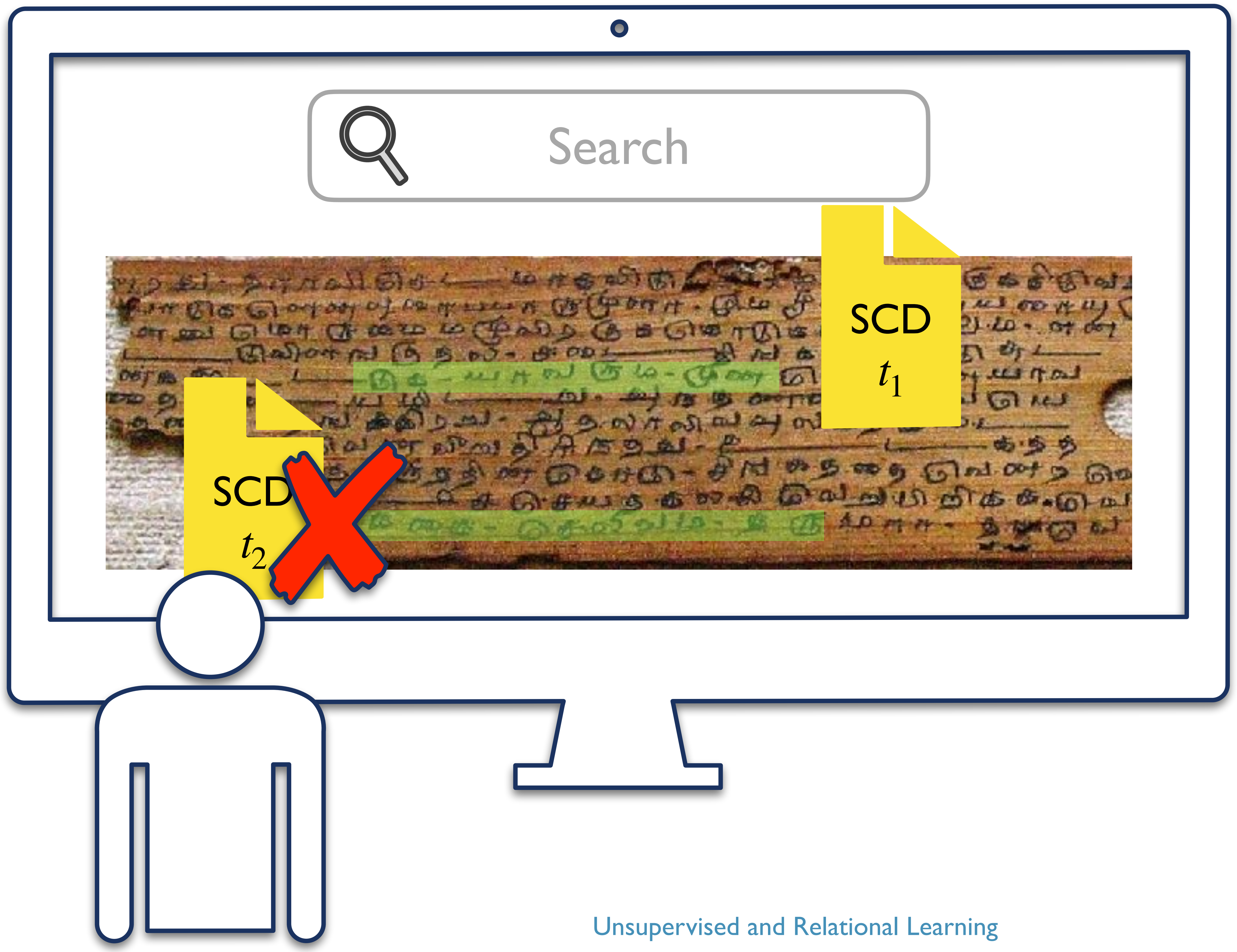
CONTINUOUS IMPROVEMENT BY FEEDBACK

FRESH – FEEDBACK-RELIANT ENHANCEMENT OF SUBJECTIVE CONTENT DESCRIPTIONS BY HUMANS

Magnus Bender, Kira Schwandt, Ralf Möller, Marcel Gehrke **FrESH – Feedback-reliant Enhancement of Subjective Content Descriptions by Humans** @ *Humanities-Centred AI (CHAI) Workshop at KI2023*

FRESH – FEEDBACK-RELIANT ENHANCEMENT OF SUBJECTIVE CONTENT DESCRIPTIONS BY HUMANS

- Information retrieval service uses SCDs
  - Corpus of documents associated with SCD
  - SCD matrix for corpus
- **Problem:** Faulty SCDs, faulty content like *fake-news,* or privacy-protected content
  - Delete from corpus ✓
  - Retrain SCD matrix from scratch?
  - Update SCD matrix ✓



Picture by Eva Wilden, in: Tamil Satellite Stanzas: Genres and Distribution

UPDATE SCD MATRIX:
DELETE SINGLE SENTENCE

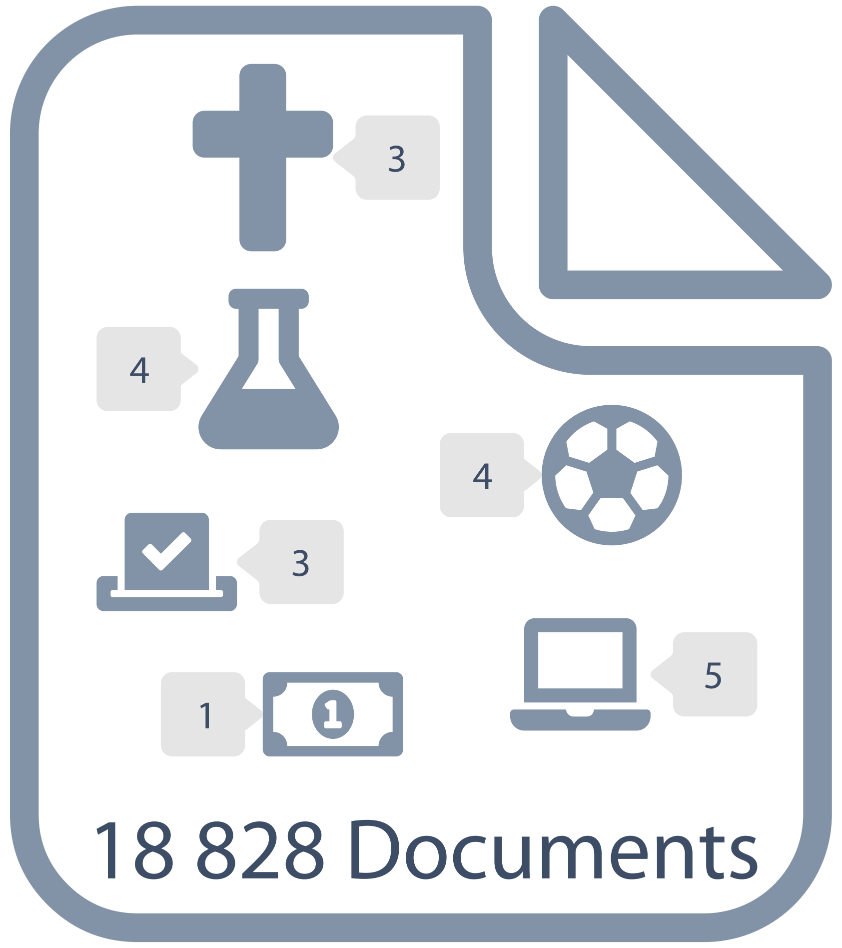
- Update distribution (matrix row) of SCD
- Revsere SEM for sentences p and SCD
- Cases
  - **C1:** Sentence and SCD known
  - **C2:** SCDs not known
    → MPS²CD
  - **C3:** Distribution instead of frequencies in matrix
    → Assume factor
  - C2+C3 may be combined

Algorithm Feedback-reliant Enhancement of SCDs

function FRESH(SCD Matrix $\delta(\mathcal{D})$, Set of faulty Sentences p)
    for each $(s, t) \in p$ **do**
        if $t = nil$ **then**
            $t = \text{MPS}^2\text{CD}(\delta(\mathcal{D}), s)$
        if DISTRIBUTIONMATRIX(δ) **then**
            $m = \min_{j=1,\ldots,n;\ \delta(\mathcal{D})[t][j]>0} \delta(\mathcal{D})[t][j]$
        else
            $m = 1$
        for each word $w_i \in s$ **do**
            $\delta'(\mathcal{D})[t][w_i] -= I(w_i, s) \cdot m$
    return $\delta(\mathcal{D})$

EVALUATION

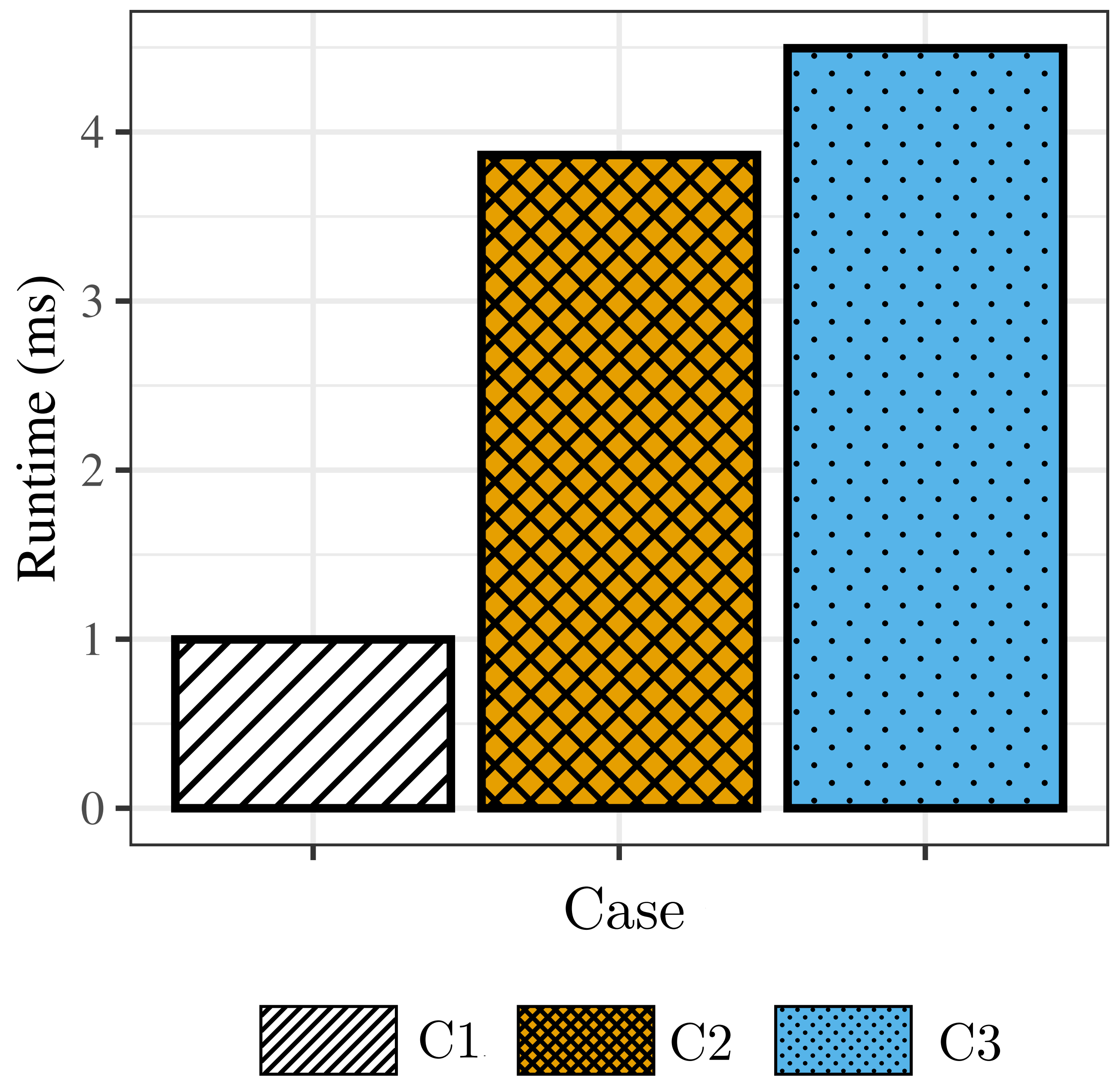
- Corpora
  - Assumed faulty $\mathscr{D}_s$
  - Assumed correct $\mathscr{D}_k$
  - Full corpus $\mathscr{D}_f = \mathscr{D}_s \cup \mathscr{D}_k$
- Workflow
  1. SCD matrices $\delta(\mathscr{D}_f)$ and $\delta(\mathscr{D}_k)$
  2. Run update $\delta' = \mathrm{FrESH}\big(\delta(\mathscr{D}_f, \mathscr{D}_s)\big)$
  3. Evaluate distance between $\delta(\mathscr{D}_k)$ and δ'

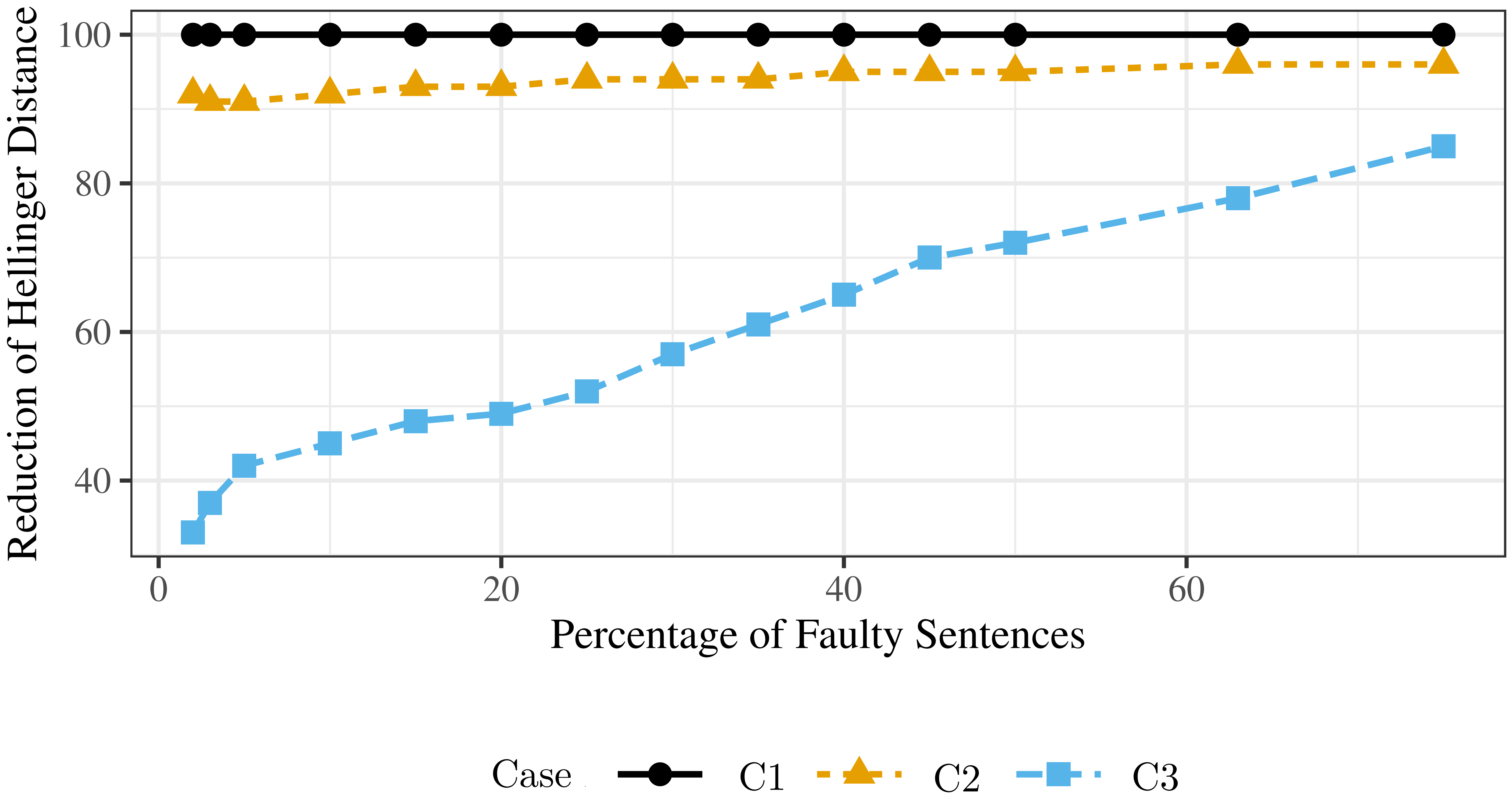


- Dataset
  - 20 newsgroups
  - SCD using SEM and Open IE

$$HD_t(\delta', \delta(\mathscr{D}_k)) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}\left(\sqrt{\delta'[t][i]} - \sqrt{\delta(\mathscr{D}_k)[t][i]}\right)^2}$$

RESULTS:
RUNTIME & DELETION ACCURACY





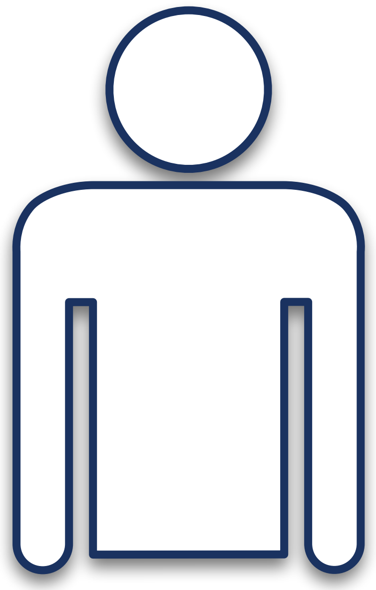
Reduction: Initially distance $\delta(\mathscr{D}_f)$ and $\delta(\mathscr{D}_k)$, finally $\delta(\mathscr{D}_k)$ and δ' → reduction 100 % deletion exact; Percentage: size of $\mathscr{D}_s$

USAGE EXAMPLE

„The seat of an association, unless otherwise provided, is the place where the administration is conducted.“

„The seat of a foundation, unless otherwise provided, is the place where the administration is conducted.“

„A minor child shares the parents' residence.“



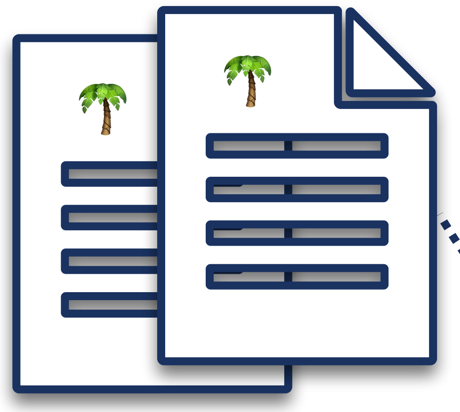
„The seat of an association, unless otherwise provided, is the place where the administration is conducted.“

„The seat of a foundation, unless otherwise provided, is the place where the administration is conducted.“

$$
\begin{array}{c}
 \\
t_1 \\
t_2 \\
\vdots \\
t_m
\end{array}
\begin{array}{c}
\begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\
\begin{pmatrix}
v'_{1,1} & v'_{1,2} & \cdots & v'_{1,n} \\
v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
v_{m,1} & v_{m,2} & \cdots & v_{m,n}
\end{pmatrix}
\end{array}
$$

FrESH

§§ 11, 24, 83 BGB (https://www.gesetze-im-internet.de/bgb/)



Referenced Sentences

$\{s_1, \ldots, s_S\}$

SCD t_i

Additional Data $\mathscr{C}_i$:

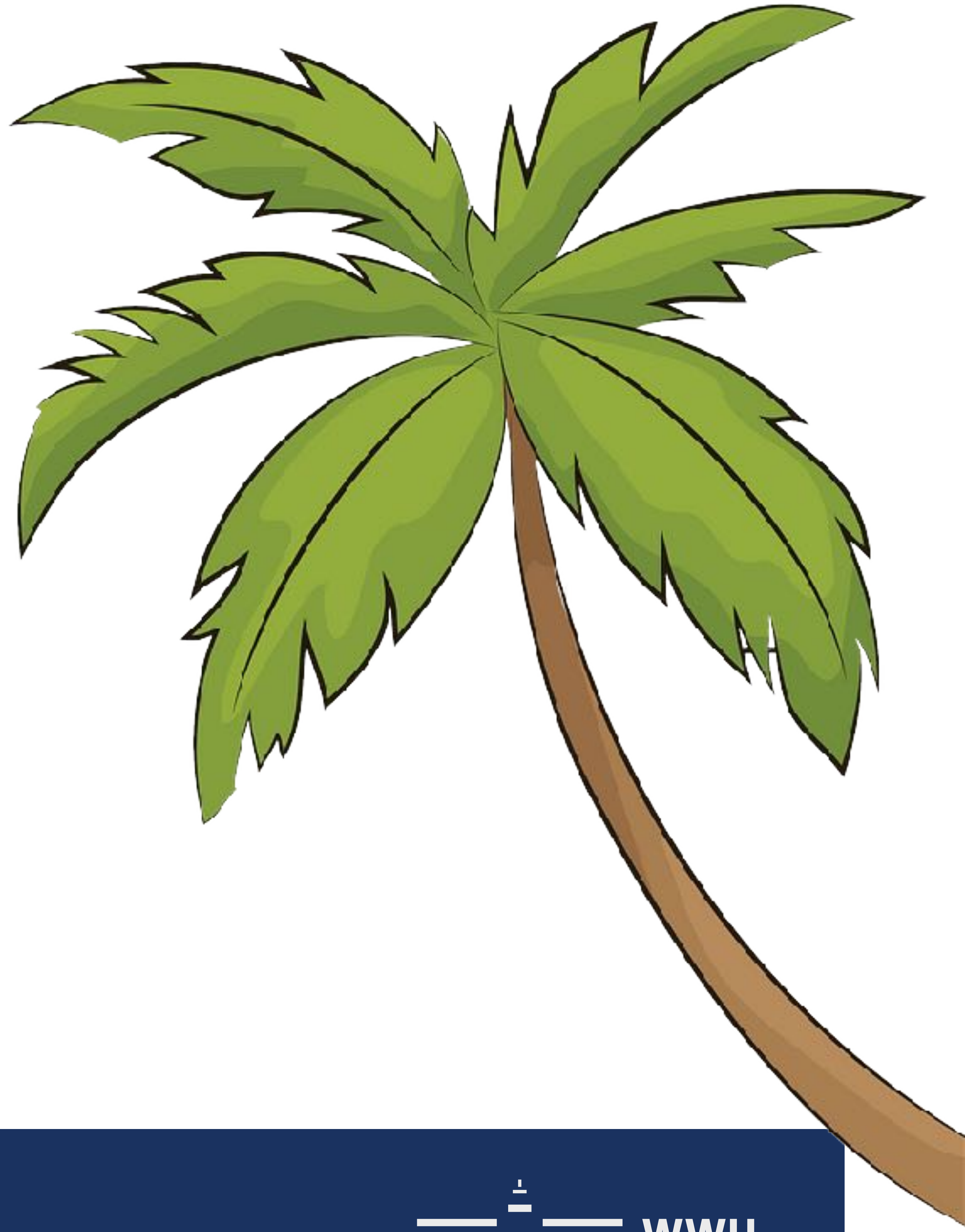
- Label l_i
- Relations
- Links
- …

Labels?

$$\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\ \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix} \end{array}$$

Word Distribution

$\{v_{i,1}, \ldots, v_{i,n}\}$



LABELLING OF SCDS

LESS – LEAN COMPUTING FOR SELECTIVE SUMMARIES

UNIVERSITÄT ZU LÜBECK

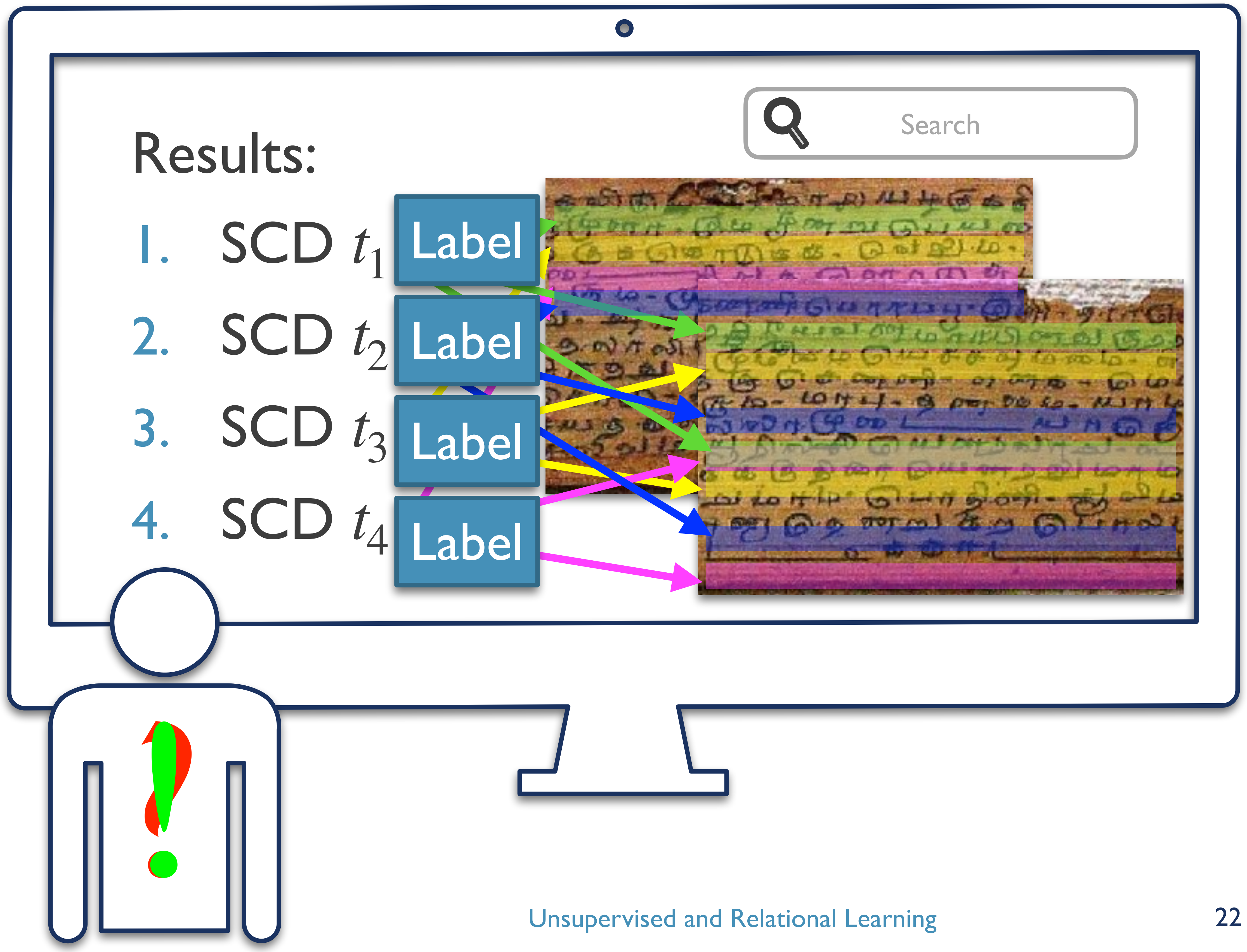
WWU MÜNSTER

Magnus Bender, Tanya Braun, Ralf Möller, Marcel Gehrke **LESS is More: LEan Computing for Selective Summaries** @ *KI 2023: Advances in Artificial Intelligence. Lecture Notes in Computer Science, Springer.*

LABELS AS DESCRIPTIONS FOR SCDS

- User browses corpus with SCDs
- SCDs represent concepts mentioned in corpus
- SCDs contain references to sentence
- **Problem:** System needs to describe SCDs to user

➡ **Solution:** Label for SCDs



Picture by Eva Wilden, in: Tamil Satellite Stanzas: Genres and Distribution

INFORMATION SOURCES

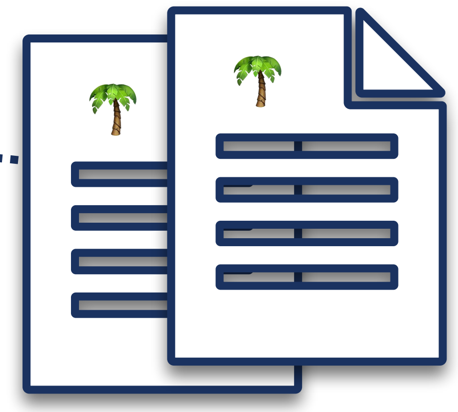
- Available per SCD
  - ✓ References sentences $\{s_1, \ldots, s_S\}$
  - ✓ Word distribution $\{v_{i,1}, \ldots, v_{i,n}\}$
  - ✗ Label
  - ✗ Other data like *relations*
- Formalised problem

$$l_i = \underset{l_j \in \text{ all possible labels}}{\operatorname{argmax}} \; Utility\Big(l_j, t_i = \big((v_{i,1}, \ldots, v_{i,n}), \{s_1, \ldots, s_S\}\big)\Big)$$

SCD t_i

Additional Data $\mathscr{C}_i$:

- Label l_i
- Relations
- Links
- …



Referenced Sentences

$\{s_1, \ldots, s_S\}$

Word Distribution

$\{v_{i,1}, \ldots, v_{i,n}\}$

$$\begin{array}{c c} & \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\ \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_m \end{array} & \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{pmatrix} \end{array}$$

LABEL CANDIDATES & UTILITY OF LABELS

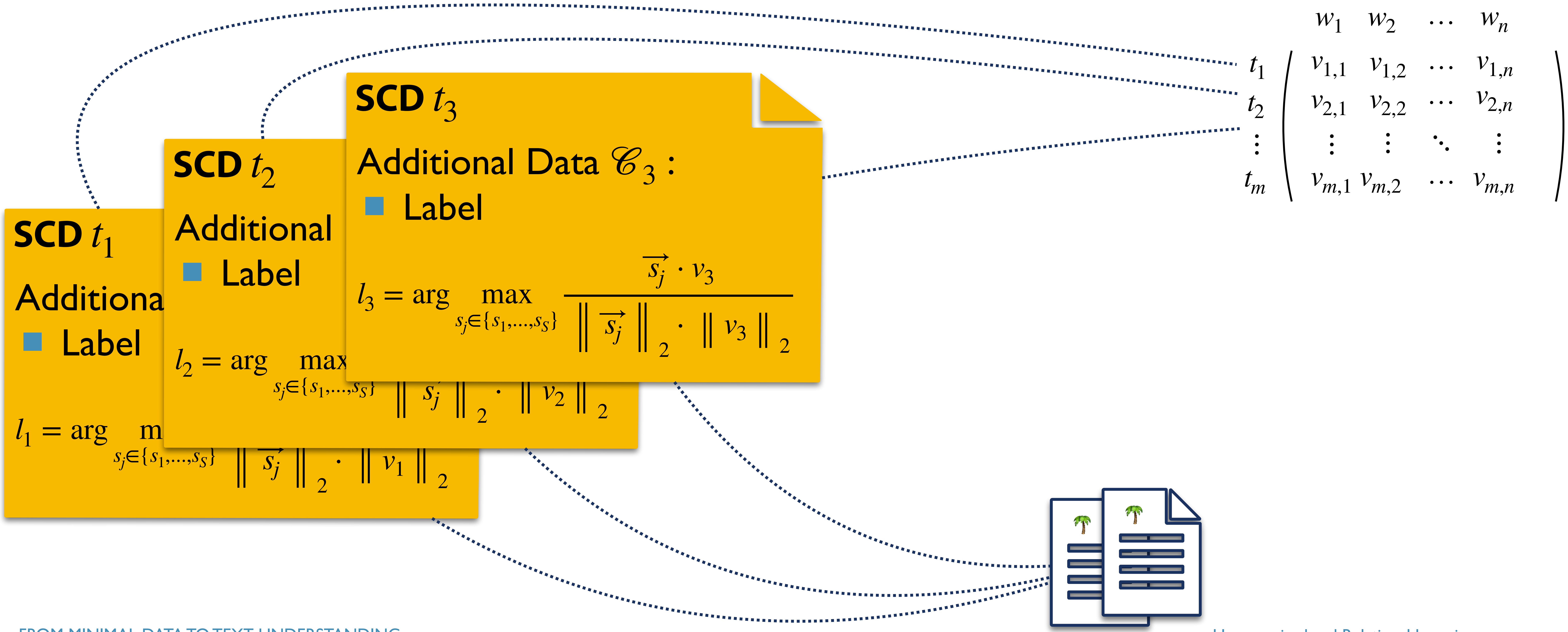
- What should a label look like?
  - Sequence of words like a short description
  - Summary of SCDs
- **Candidates:**
  - Referenced sentences $\{s_1, \ldots, s_S\}$
  - Reformulate problem

$$l_i = \arg \max_{s_j \in \{s_1, \ldots, s_S\}} Utility(s_j, (v_{i,1}, \ldots, v_{i,n}))$$

- What is a *good* label?
  - Similar to references sentences of SCDs
  - Word distributions generates sentences
- **Utility:** Cosine similarity
  - Use word distribution $\{v_{i,1}, \ldots, v_{i,n}\}$
  - Reformulate problem

$$l_i = \arg \max_{s_j \in \{s_1, \ldots, s_S\}} \frac{\vec{s_j} \cdot v_i}{\| \vec{s_j} \|_2 \cdot \| v_i \|_2}$$

APPROACH – LESS



ALGORITHM

Algorithm LEan computing for Selective Summaries

function LESS($\mathcal{D}$)

    Input: Corpus $\mathcal{D}$

    Output: SCD matrix $\delta(\mathcal{D})$; SCD set $g(\mathcal{D})$ containing labels l_i for SCDs t_i

    $\delta(\mathcal{D}) \leftarrow \text{USEM}(\mathcal{D})$ ▷ Run USEM

    $g(\mathcal{D}) \leftarrow \{\}$ ▷ Initialize empty SCD set $g(\mathcal{D})$

    for each row of matrix $i = 1, ..., K$ **do**

        $v_i \leftarrow g(\mathcal{D})[i]$ ▷ Extract SCD-word distribution

        $\{s_1, ..., s_S\} \leftarrow \textsc{ReferencedSentences}(i)$ ▷ Get referenced sentences

        $l_i \leftarrow \arg\max_{s_j \in \{s_1, ..., s_S\}} \dfrac{\vec{s_j} \cdot v_i}{\|\vec{s_j}\|_2 \cdot \|v_i\|_2}$ ▷ Compute label

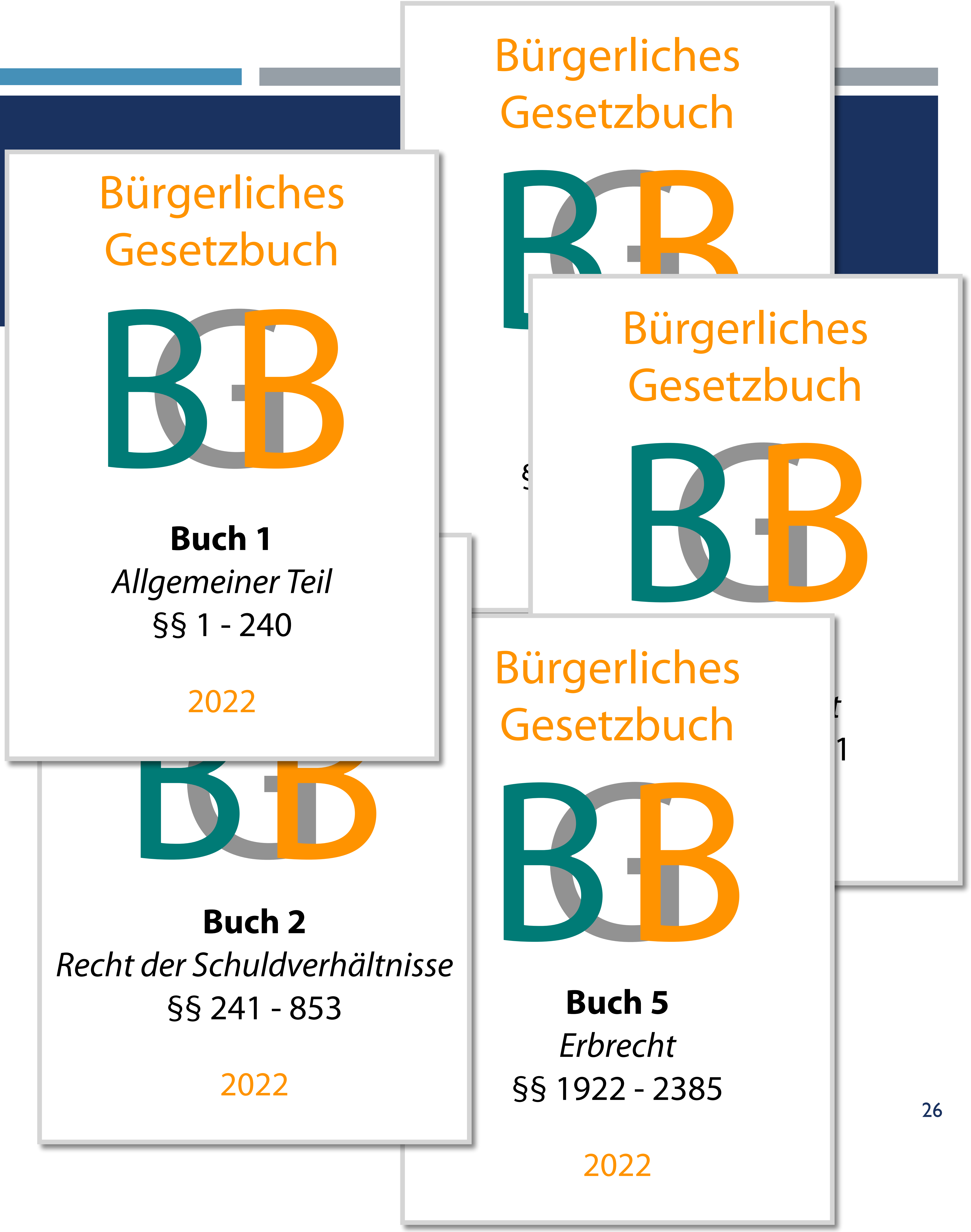
        $t_i \leftarrow (\{l_i\}, \{s_1, ..., s_S\})$ ▷ Compose associated SCD with computed label

        $g(\mathcal{D}) \cup \{t_i\}$ ▷ Add to SCD set

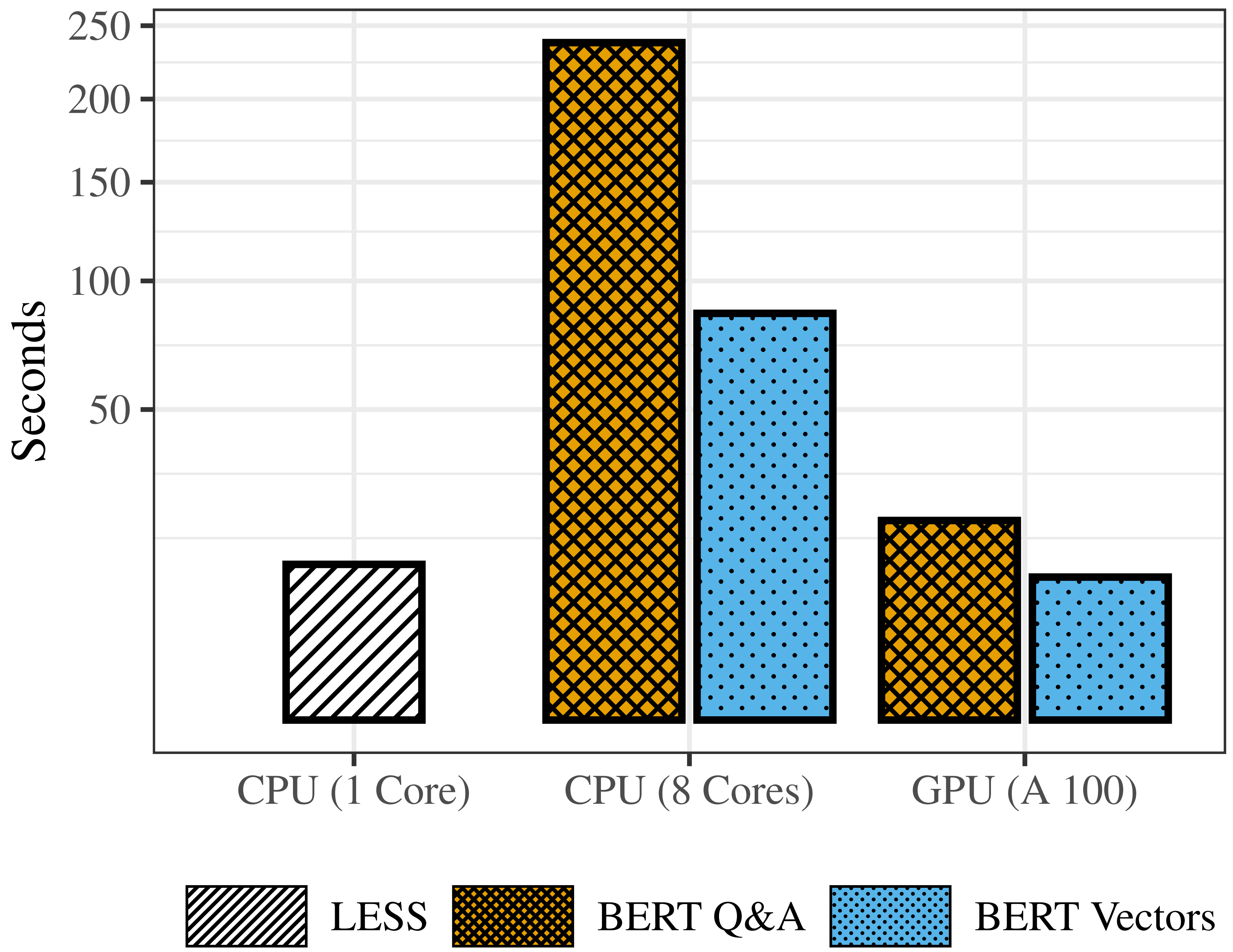
    return $\delta(\mathcal{D})$, $g(\mathcal{D})$

EVALUATION & DATASET: AGAIN BGB

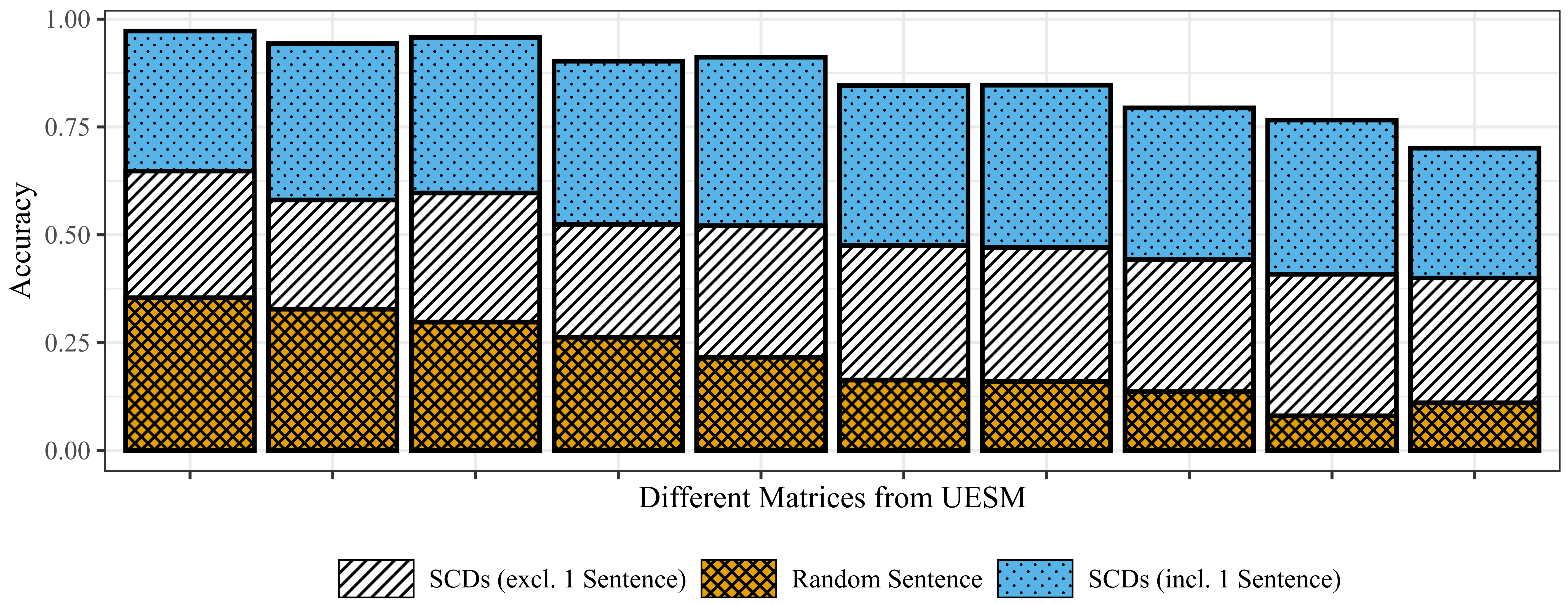
- Bürgerliches Gesetzbuch (BGB)
  - German civil code (German language)
- First run USEM
- Second add labels with LESS
  - Compare to BERT-based approach



RUNTIME AND ACCURACY



Two techniques using BERT; 1 or 8 Intel CPU cores and single NVIDIA A100 40GB GPU



Random sentence: Theoretical accuracy a random approach would result in.

USAGE EXAMPLE

Association Commercial Business Operation

„An association whose purpose is not directed towards a commercial business operation acquires legal capacity through entry in the register of associations at the competent local court.“

„An association whose purpose is to engage in commercial business shall acquire legal capacity, in the absence of special federal law, through state conferral. The grant is due to the state in whose territory the association has its registered office.“

Seat Foundation Place Administration

„The seat of an association, unless otherwise provided, is the place where the administration is conducted.“

„The seat of a foundation, unless otherwise provided, is the place where the administration is conducted.“

$$
\begin{array}{cc}
 & \begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\
\begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_m \end{array} &
\begin{pmatrix}
v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\
v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
v_{m,1} & v_{m,2} & \cdots & v_{m,n}
\end{pmatrix}
\end{array}
$$

USEM
+ LESS

§§ 21, 22, 24, 83 BGB (https://www.gesetze-im-internet.de/bgb/)

SCD t_i

Additional Data $\mathscr{C}_i$:

- Label l_i
- Relations
- Links
- …

Word Distribution $\{v_{i,1}, \dots, v_{i,n}\}$

$$
\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array}
\begin{array}{c}
\begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \\
\begin{pmatrix}
v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\
v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
v_{m,1} & v_{m,2} & \cdots & v_{m,n}
\end{pmatrix}
\end{array}
$$

Relations?

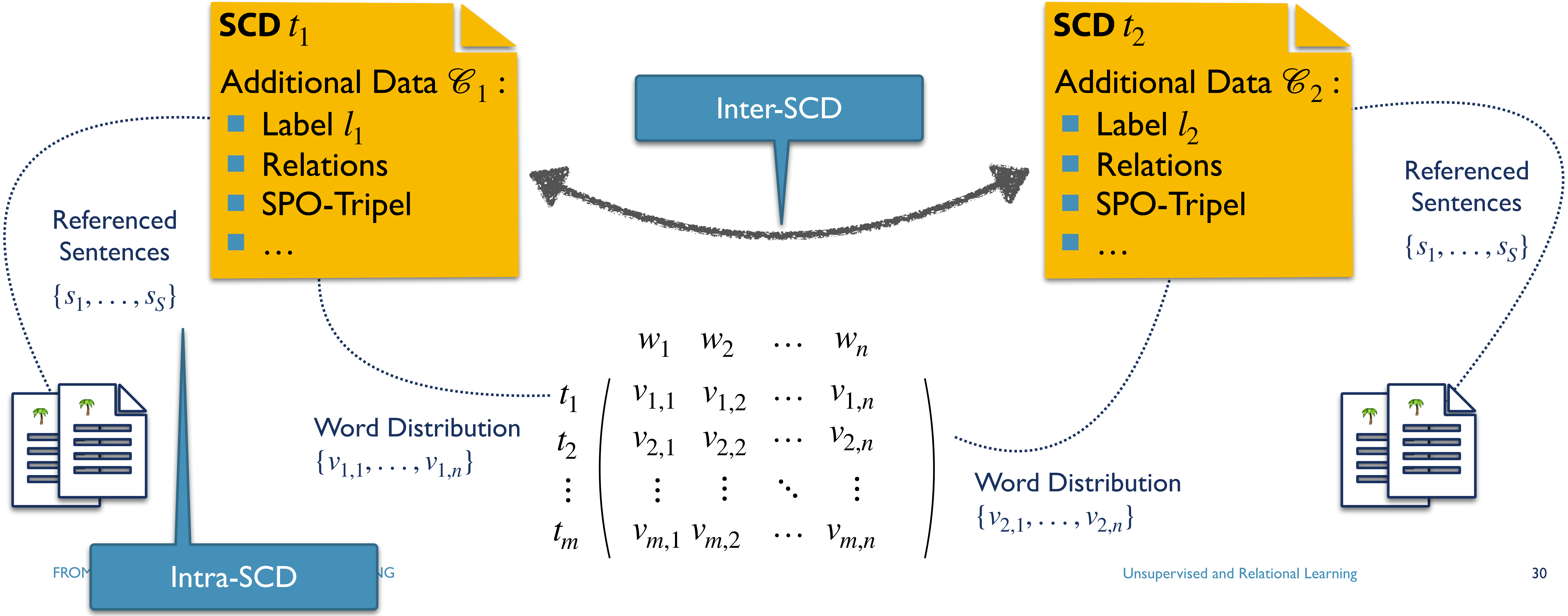
INTER- AND INTRA-SCD RELATIONS

ENRICHING A CORPUS WITH DOCUMENTS USING THE INTER-SCD RELATION COMPLEMENT

Magnus Bender, Felix Kuhr, Tanya Braun **To Extend or not to Extend? Complementary Documents** in *16th IEEE International Conference on Semantic Computing (ICSC 2022)*

Magnus Bender, Felix Kuhr, Tanya Braun **To Extend or not to Extend? Enriching a Corpus with Complementary and Related Documents** in *International Journal of Semantic Computing, 2022*

RELATIONS AMONG SCDS



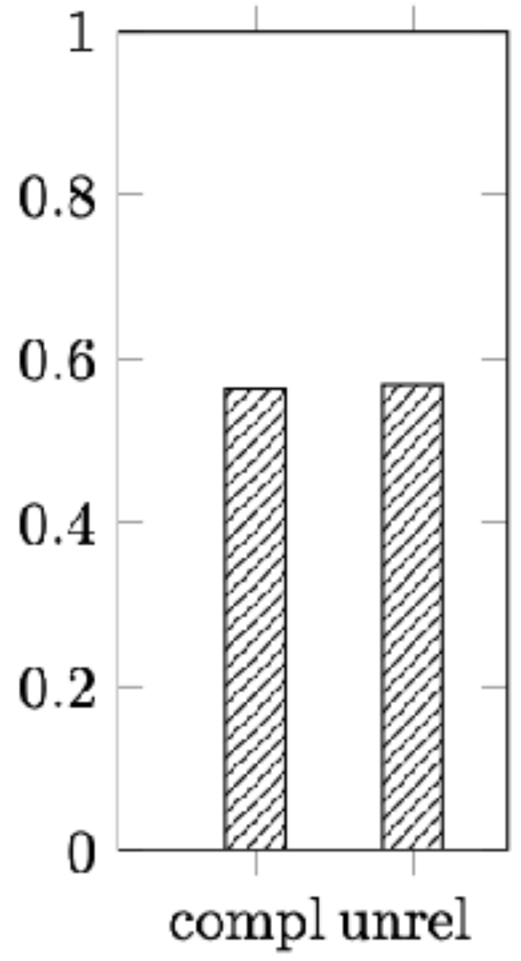
EXAMPLE INTER-SCD RELATION: COMPLEMENT

- **Goal:** Identify documents that are complementary to a corpus/ a document in a corpus
  - Binary classification problem:

    $Complement = true$ or $Complement = false$
- Solution approach to corpus enrichment uses cosine similarity at its core
  - Sequence of similarity values between vector representations of SCDs and the words in the new document
- Also applies to many document retrieval approaches: return documents similar in some regard
  - Topic distribution similar, entities match (equality), etc.

- Corpus $\mathscr{D}_r$ on sporting events
  - Olympics 2020, UEFA Euro 2020

- Corpus $\mathscr{D}_c$ with *complementary* documents
  - Covid-19 spread in cities



Similarity values of complements and unrelated documents for corpus enrichment.

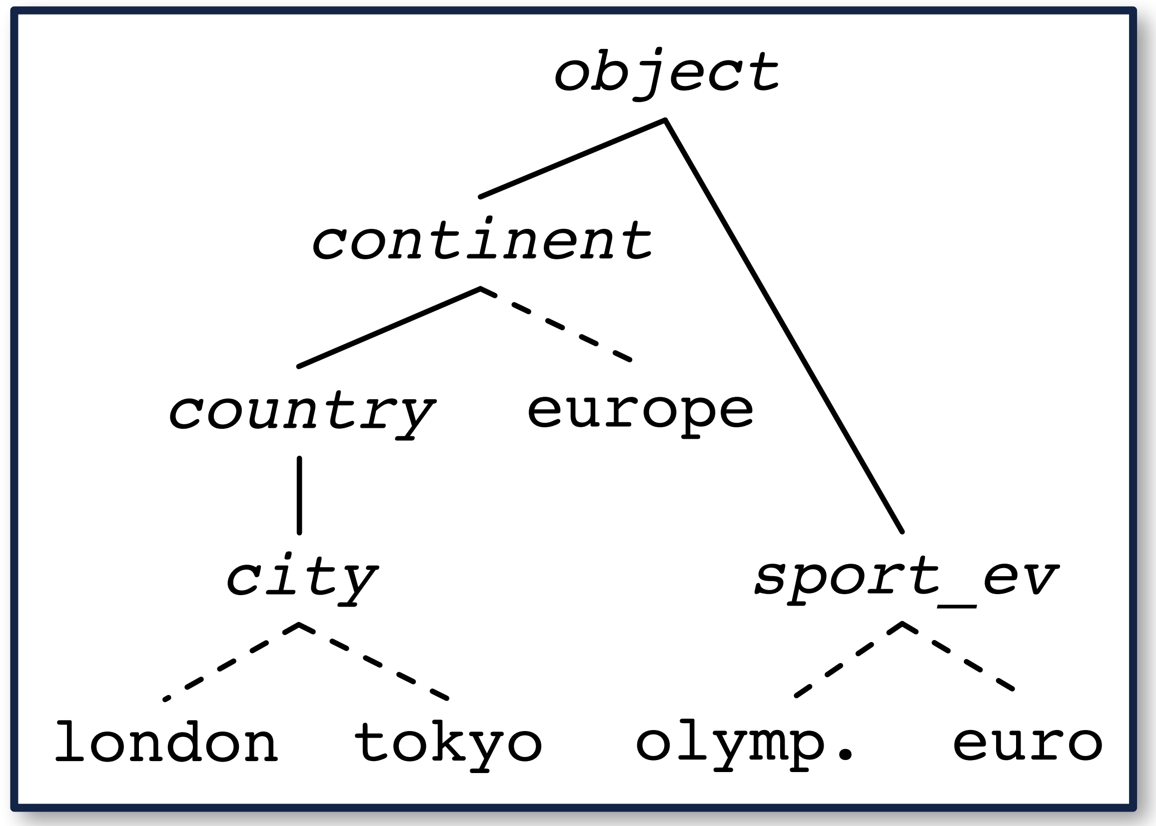
Problem: How do we *formally* define complementarity accounting for semantics?

Problem: Similarity-based approaches might only provide more of the same.

HOW TO GRASP COMPLEMENTARITY ON A FORMAL LEVEL?

- Use SCDs specifically in the form of SPO-Triples
  - SPO-Triple: <subject, predicate, object>
  - Extract for any document, e.g., with OpenIE tools
- Together with a taxonomy
  - Hierarchy of concepts
  - Dictionary of synonyms
- Allow for grasping complementarity on a semantic level by
  - Looking at shared concepts in the SPO triples
  - While also accounting for hierarchy and synonyms

- Words of *complement* very different compared to corpus
  - Different (topic / SCD) distributions
  - Likely to be classified as unrelated



- t_1: <Olympics '21, in, Tokyo>
- t_2: <UEFA euro '20, in, Europe>

- t_3: <Covid-19, in, Tokyo>
- t_4: <Covid-19, in, London>

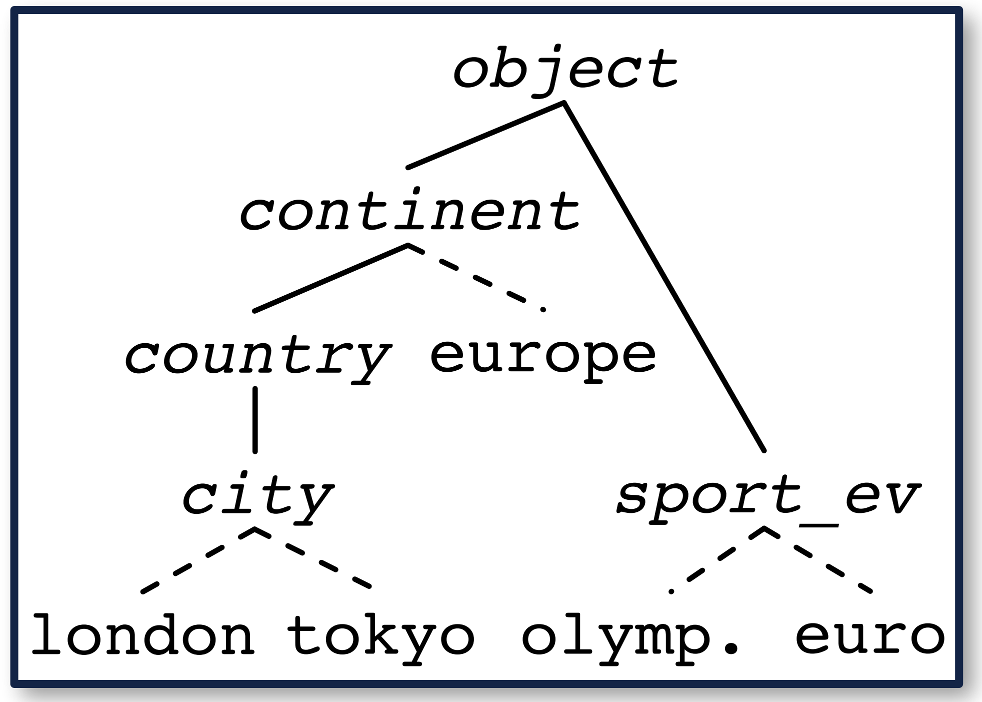
A FORMAL DEFINITION: COMPLEMENTARY SCDS

- Let $x^\uparrow$ refer to the concept or meaning of x
- Seven types of complementarity between SCDs t_i, t_j

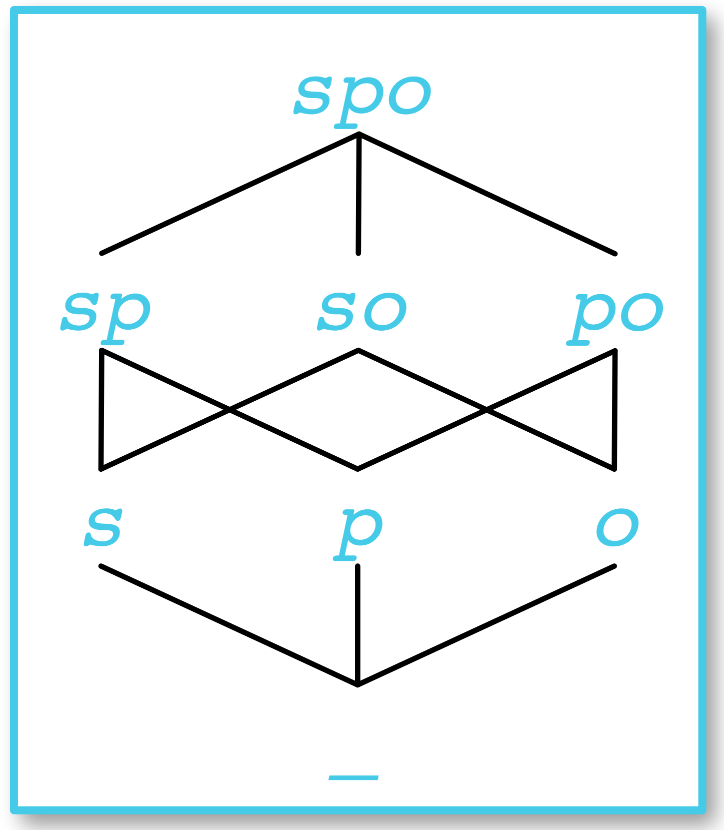
1. s $\quad t_i = \langle s^\uparrow, p_i, o_i \rangle, t_j = \langle s^\uparrow, p_j, o_j \rangle$
2. p $\quad t_i = \langle s_i, p^\uparrow, o_i \rangle, t_j = \langle s_j, p^\uparrow, o_j \rangle$
3. o $\quad t_i = \langle s_i, p_i, o^\uparrow \rangle, t_j = \langle s_j, p_j, o^\uparrow \rangle$
4. sp $\quad t_i = \langle s^\uparrow, p^\uparrow, o_i \rangle, t_j = \langle s^\uparrow, p^\uparrow, o_j \rangle$
5. so $\quad t_i = \langle s^\uparrow, p_i, o^\uparrow \rangle, t_j = \langle s^\uparrow, p_j, o^\uparrow \rangle$
6. po $\quad t_i = \langle s_i, p^\uparrow, o^\uparrow \rangle, t_j = \langle s_j, p^\uparrow, o^\uparrow \rangle$
7. spo $\quad t_i = \langle s^\uparrow, p^\uparrow, o^\uparrow \rangle, t_j = \langle s^\uparrow, p^\uparrow, o^\uparrow \rangle$

- Types gets more strict $\rightarrow$ Order in lattice

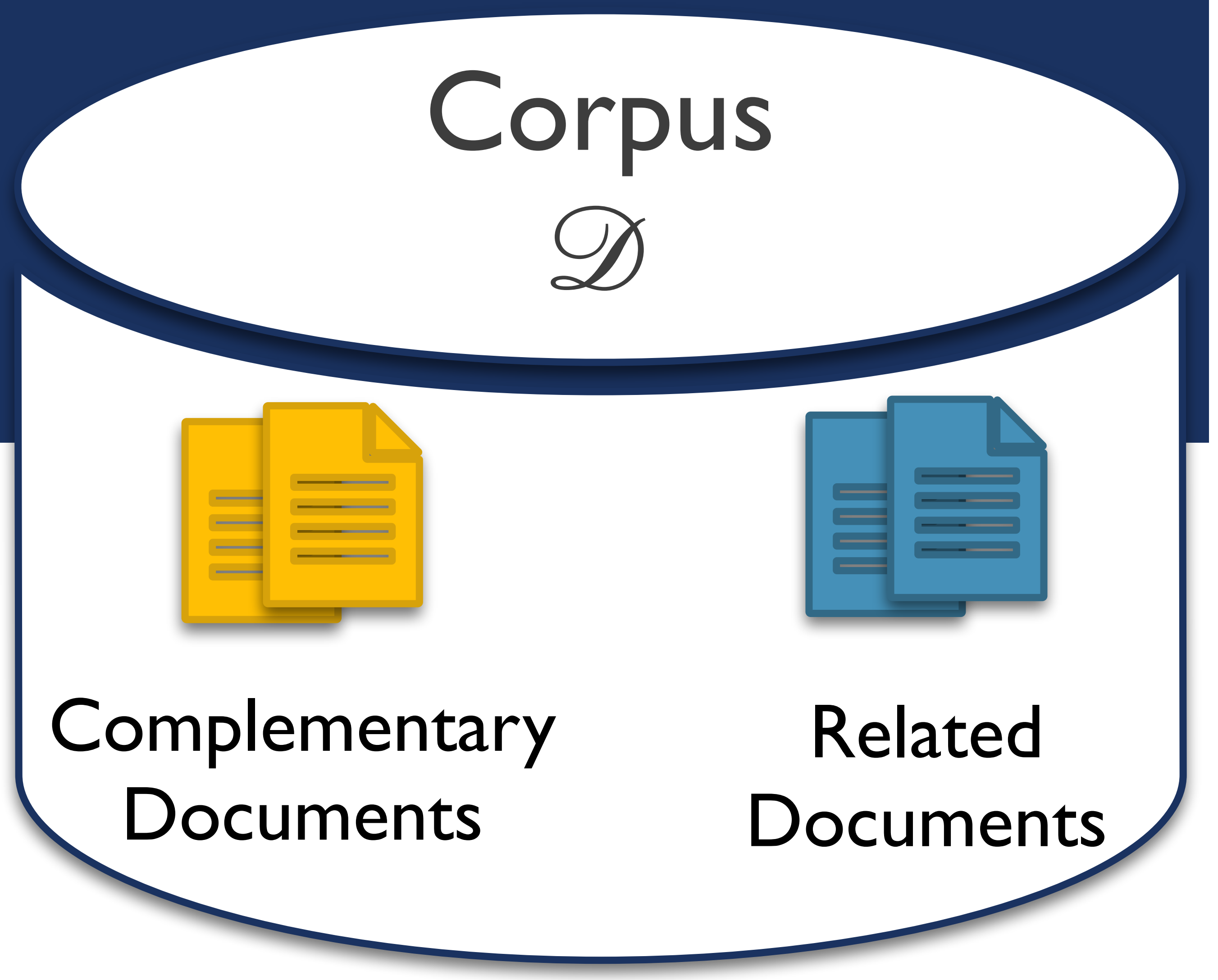
- t_1: `<Olympics '21, in, Tokyo>`
- t_2: `<UEFA euro '20, in, Europe>`
- t_3: `<Covid-19, in, Tokyo>`
- t_4: `<Covid-19, in, London>`



- t_1, t_3 o-complementary
  - s_1, s_3 share *object*; $p_1 = p_3$; $o_1 = o_3$
  - And p, po complementary
  - Same holds for t_1, t_4
- spo-complementary
  - All three items share same concept or are identical
- s-complementary
  - s shares same concept, other two different



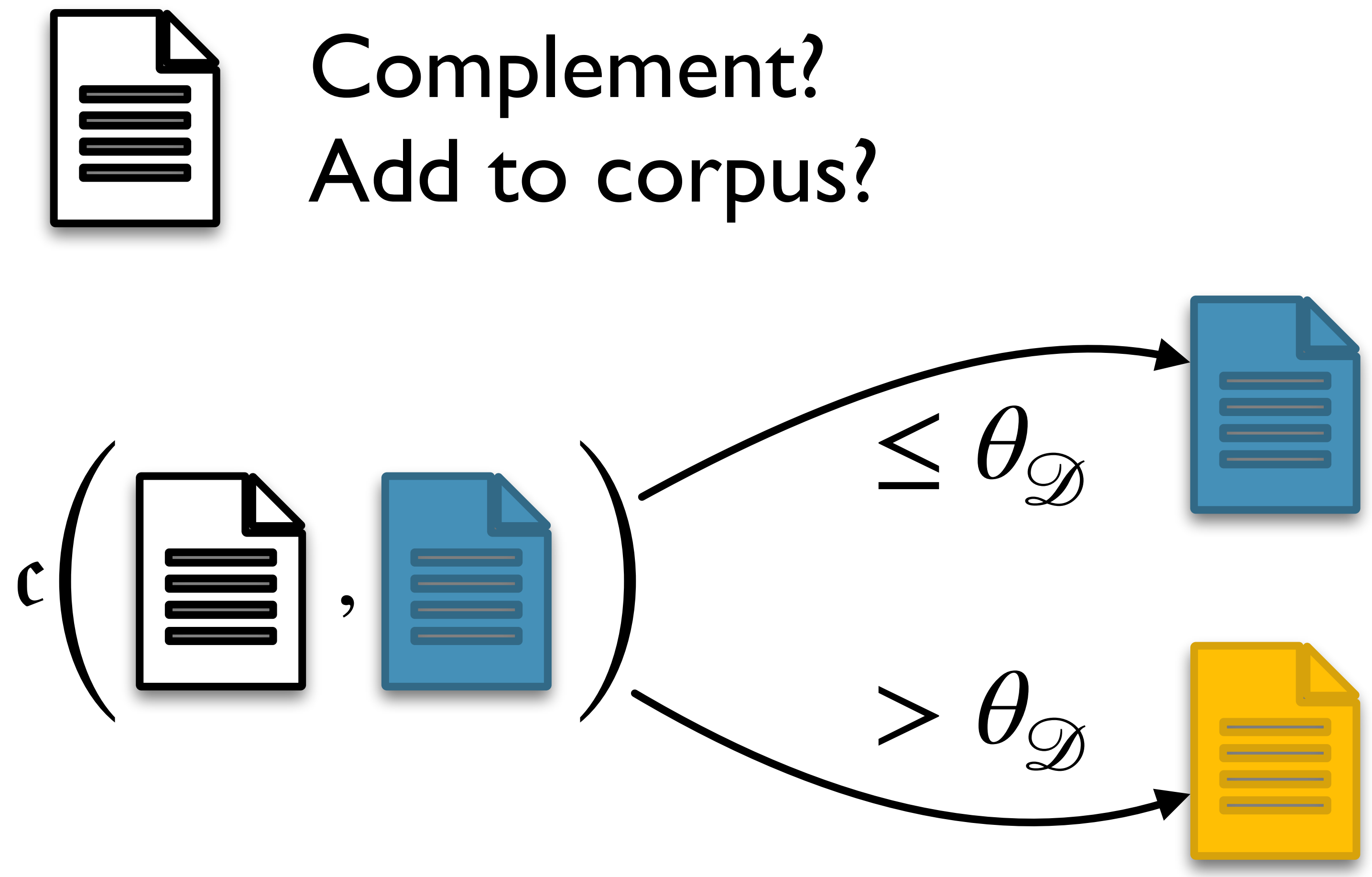
CORPUS ENRICHMENT: COMPLEMENTARY DOCUMENTS



- Let $\mathfrak{C}_x(t_i, t_j), x \in \mathscr{X} = \{s, p, o, sp, so, op, spo\}$ be an indicator function
  - Returns 1 if $t_i,\ t_j$ x-complementary; otherwise 0
  - $\mathfrak{C}_x$ is symmetric, i.e., $\mathfrak{C}_x(t_i, t_j) = \mathfrak{C}_x(t_j, t_i)$
- Complementarity value between documents d', d:

$$\mathfrak{c}(d', d) = \sum_{t_i \in g(d')} \sum_{t_j \in g(d)} \sum_{x \in \mathscr{X}} w_x \mathfrak{C}_x(t_i, t_j)$$

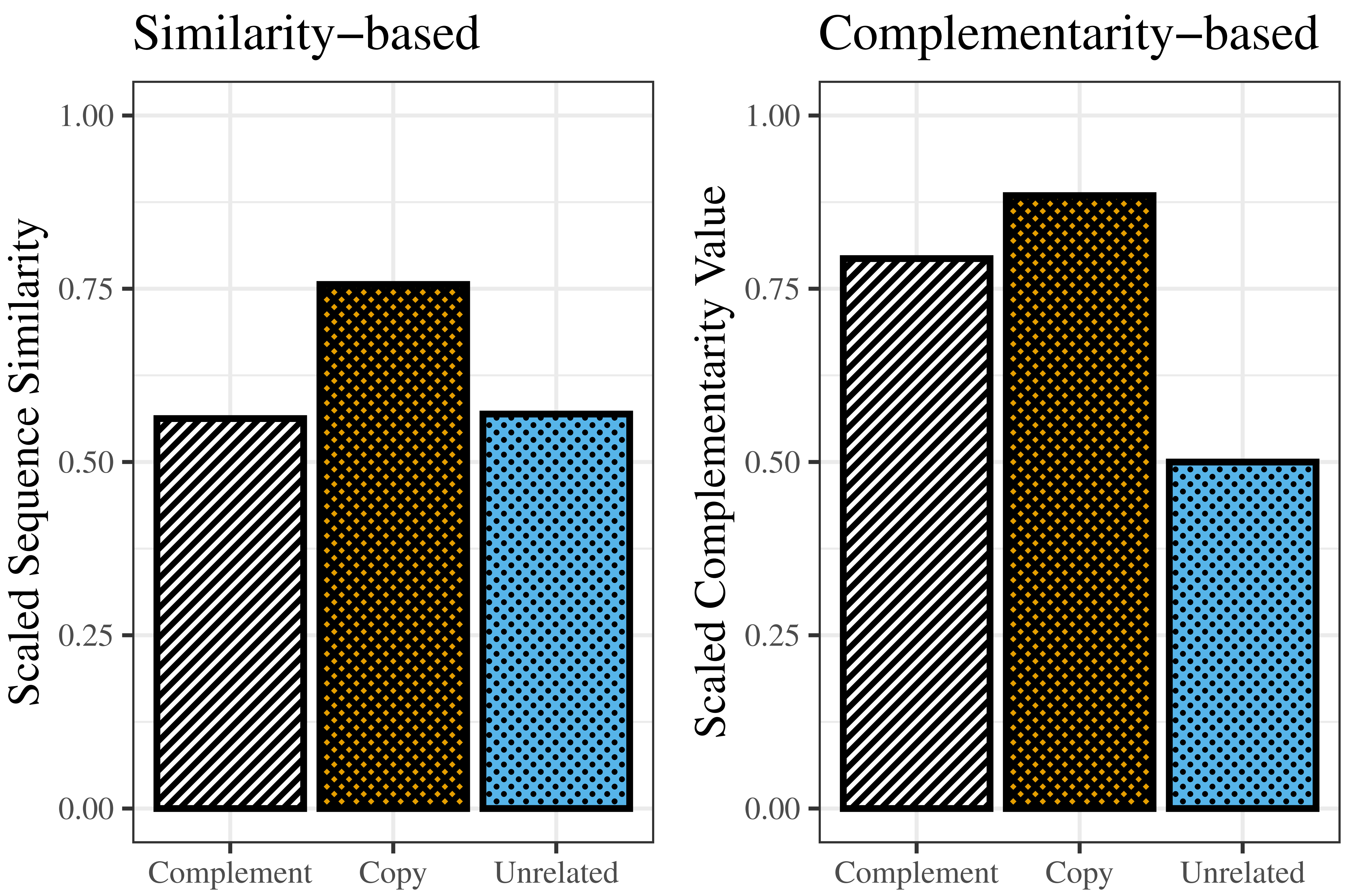
  - Sum over all pairs of SCDs $t_i \in g(d'), t_j \in g(d)$, indicating if $t_i,\ t_j$ are x-complementary
  - $\mathfrak{c}$ is symmetric, i.e., $\mathfrak{c}(d',\ d) = \mathfrak{c}(d,\ d')$
  - Assign weights w_x, $\sum_{w \in \mathscr{X}} w_x = 1$ to complementarity types x to encode which complementarity interested in



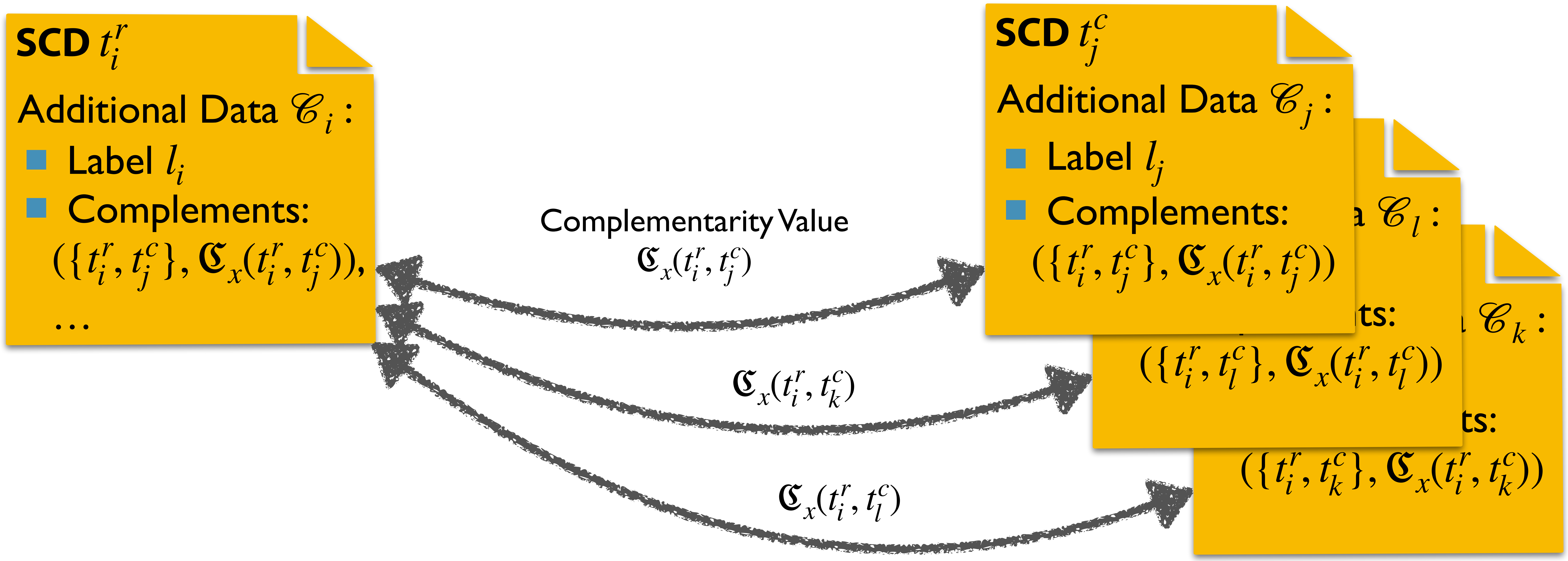
COMPLEMENT DETECTION

- Similarity-based technique does not distinguish between *complement* and *unrelated*
- Complementarity-based technique uses $\mathfrak{c}_x(d', d)$

➡ Resulting values differ for *complement* and *unrelated*



BACK TO RELATIONS: COMPLEMENTARITY BETWEEN SCDS



RELATIONS IN SCD MATRIX: COMBINED SCD MATRIX

- Combine SCDs of two different corpora in one SCD matrix
  - Corpus $\mathscr{D}_r$ related documents
  - Corpus $\mathscr{D}_c$ with complementary documents
- Model the relations among the SCDs in the matrix
  - Filter matrix to keep only SCDs from $\mathscr{D}_c$ which are complementary
- Adapted of MPS²CD yields negative similarity value for complementary SCDs

$$
\delta_c(\mathscr{D}_r, \mathscr{D}_c) = \begin{array}{c} \\ t_1^r \\ \vdots \\ t_{m_r}^r \\ t_1^c \\ \vdots \\ t_{m_c}^c \end{array}
\begin{array}{c}
\begin{array}{ccccc} w_1 & w_2 & w_3 & \cdots & w_n \end{array} \\
\begin{pmatrix}
v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,V} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
v_{m_r,1} & v_{m_r,2} & v_{m_r,3} & \cdots & v_{m_r,n} \\
v_{m_r+1,1} & v_{m_r+1,2} & v_{m_r+1,3} & \cdots & v_{m_r+1,n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
v_{m_r+m_c,1} & v_{m_r+m_c,2} & v_{m_r+m_c,3} & \cdots & v_{m_r+m_c,n}
\end{pmatrix}
\end{array}
$$

$\mathfrak{C}_x(t_i^r, t_j^c)$

$\mathfrak{C}_x(t_k^r, t_l^c)$

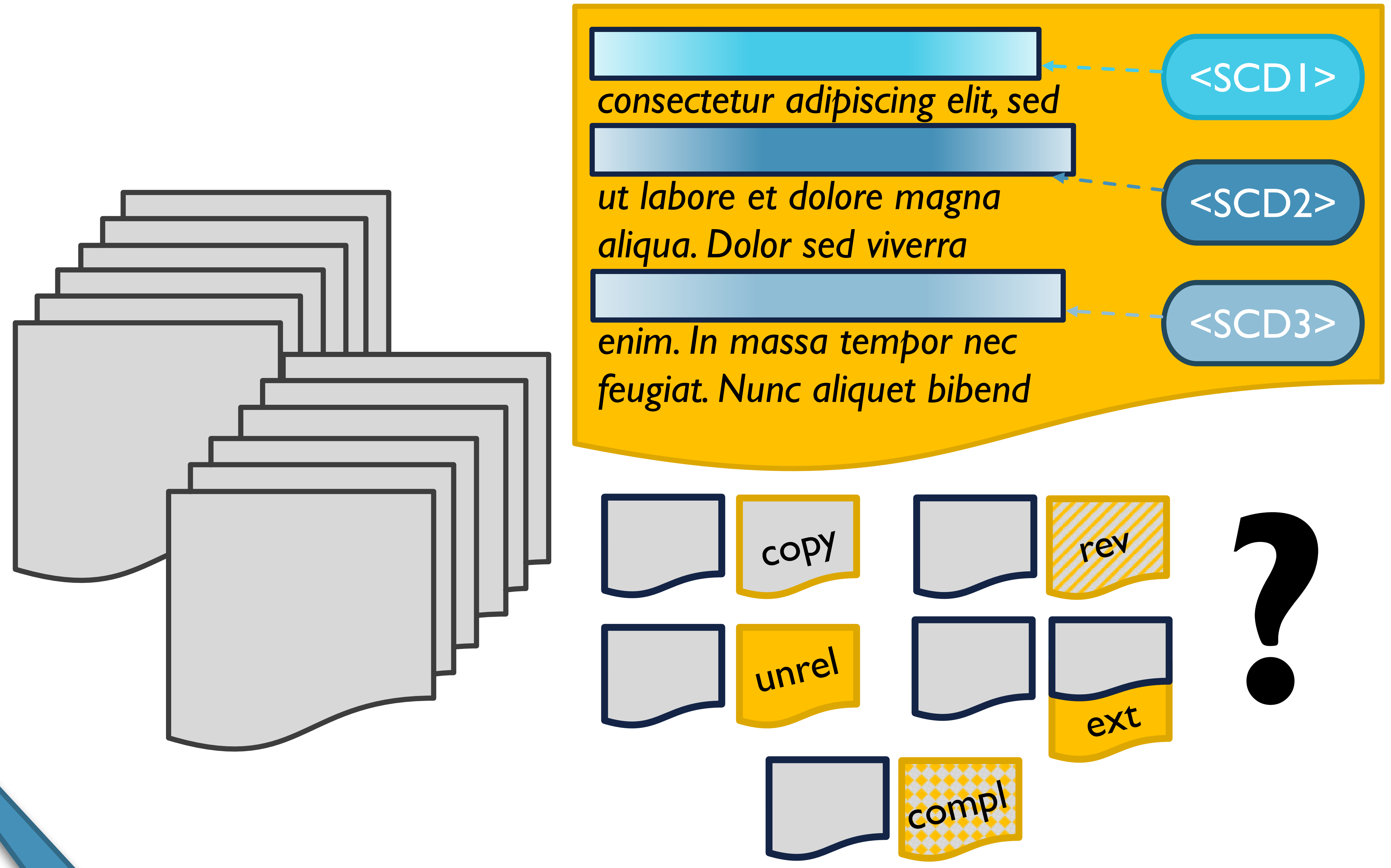
May be generalised to any type of corpora and relations among them.

CORPUS ENRICHMENT INCL. COMPLEMENTS

- Same problem as earlier

  Classify a new document before adding to corpus.
- Now five types $\mathscr{Y} = \{copy, ext, rev, unrel, compl\}$
- Find most probable type $\arg\max_{y \in \mathscr{Y}} P(Type = y \mid d', \mathscr{D})$

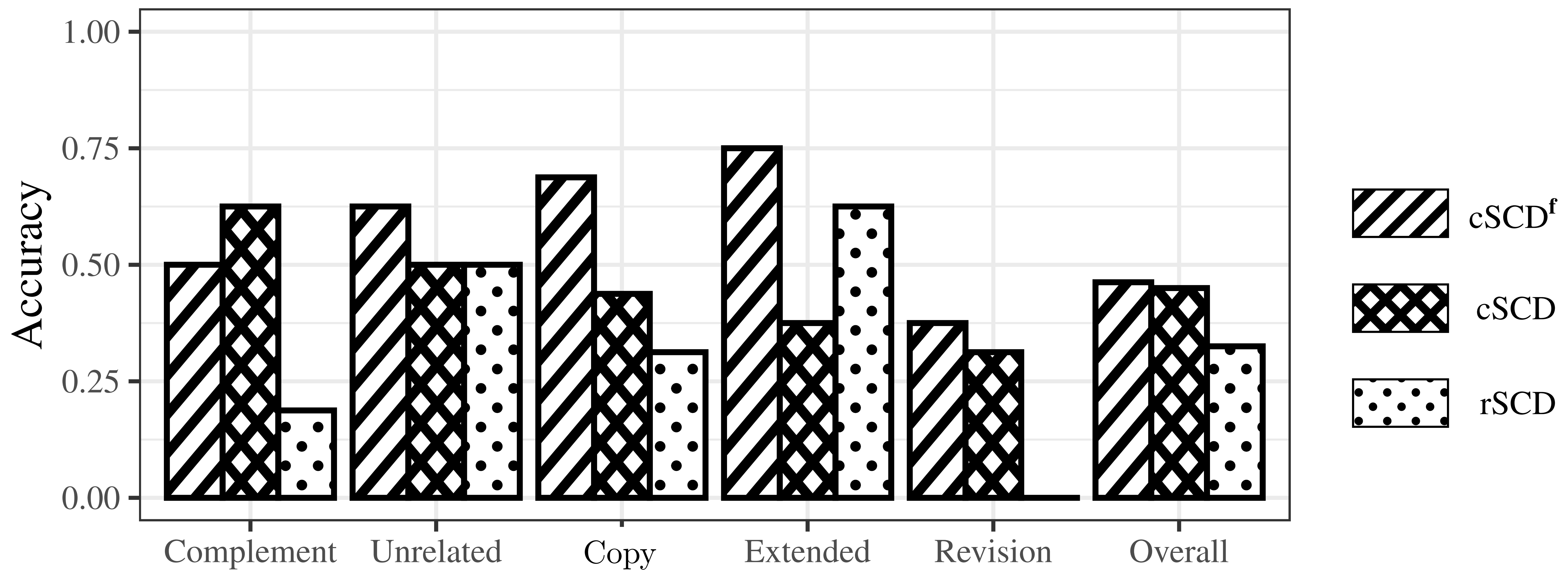
1. Build combined SCD matrix (needs corpus of related and complementary documents, use $\mathfrak{c}(d', d)$)
2. Filter matrix by removing complementary documents with no relation to related document
3. Train an HMM on MPS²CD similarity values for classification
4. Run MPS²CD on new documents and use HMM



$\mathfrak{c}(d', d)$ only in 1. needed 🙂
Querying taxonomy quite costly.

RESULTS:
COMPLEMENT DETECTION BY COMBINED SCD MATRIX

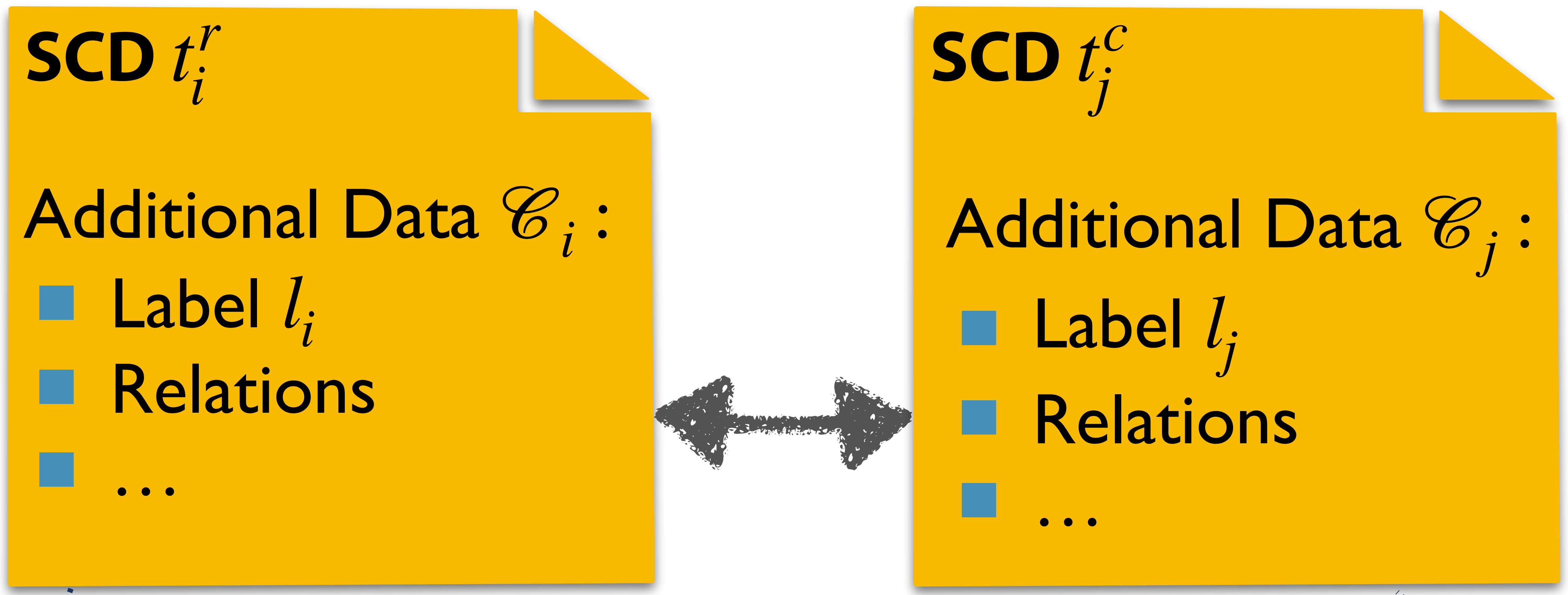
- Document classification accuracy using
  - Combined SCD matrix
    → cSCD
  - Filtered combined SCD matrix
    → cSCD^f
  - Related (normal) SCD matrix
    → rSCD



GENERALISE RELATIONS AND COMBINED MATRIX

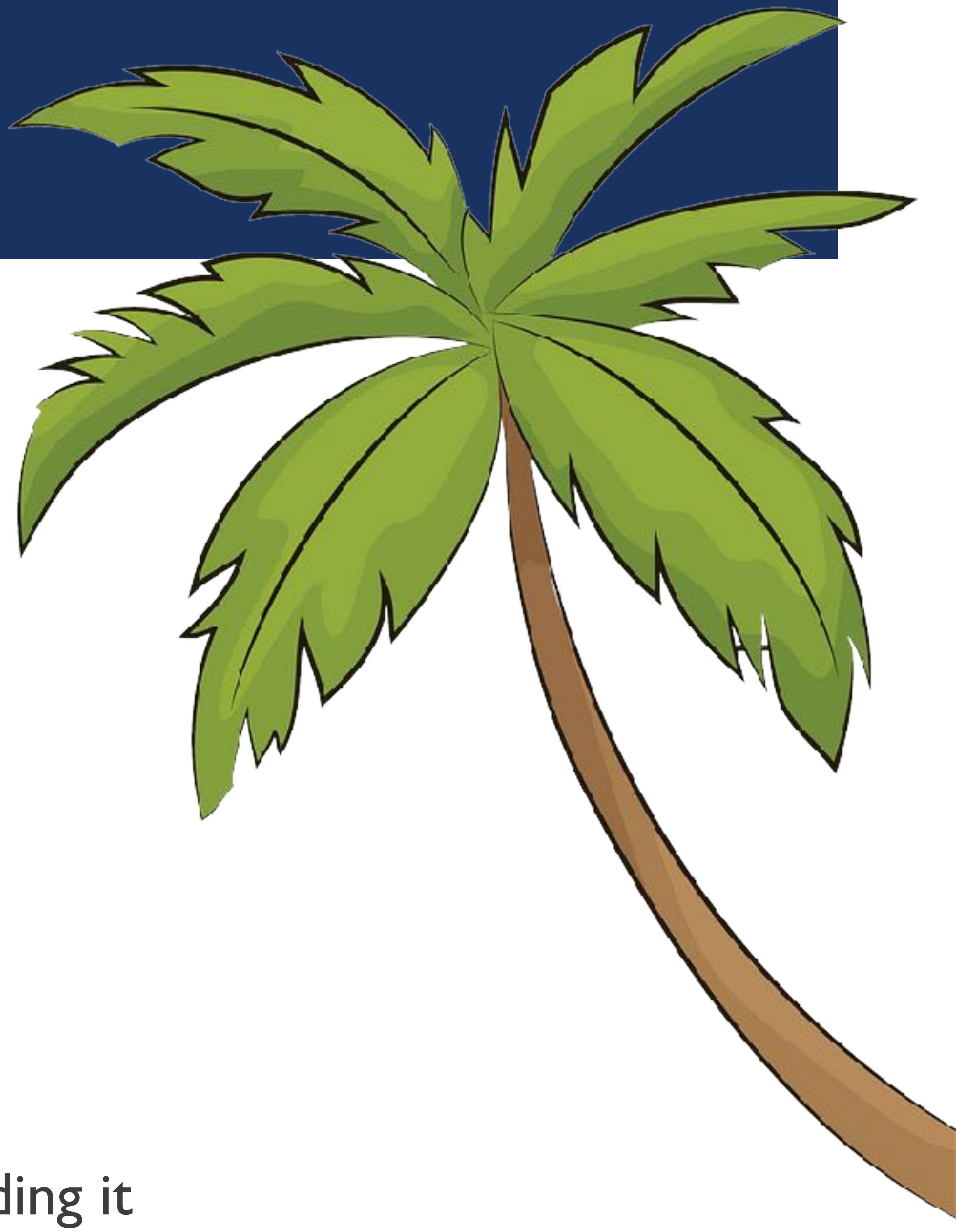
- Inter-SCD relations
  - Stored as links in additional data
  - Represented by combined SCD Matrix
    - Adapted MPS²CD yields adjusted similarity value
    - Apply techniques originally for related corpora
  - ➡ Example type complement used for corpus enrichment and document classification

- Intra-SCD relations
  - Referenced Sentences
  - Word-Distribution in matrix



$$\delta_c(\mathscr{D}_r, \mathscr{D}_c) = \begin{array}{c} \\ t_1^r \\ \vdots \\ t_{m_r}^r \\ t_1^c \\ \vdots \\ t_{m_c}^c \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & \cdots & w_n \end{array} \\ \begin{pmatrix} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,V} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{m_r,1} & v_{m_r,2} & v_{m_r,3} & \cdots & v_{m_r,n} \\ v_{m_r+1,1} & v_{m_r+1,2} & v_{m_r+1,3} & \cdots & v_{m_r+1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{m_r+m_c,1} & v_{m_r+m_c,2} & v_{m_r+m_c,3} & \cdots & v_{m_r+m_c,n} \end{pmatrix} \end{array}$$

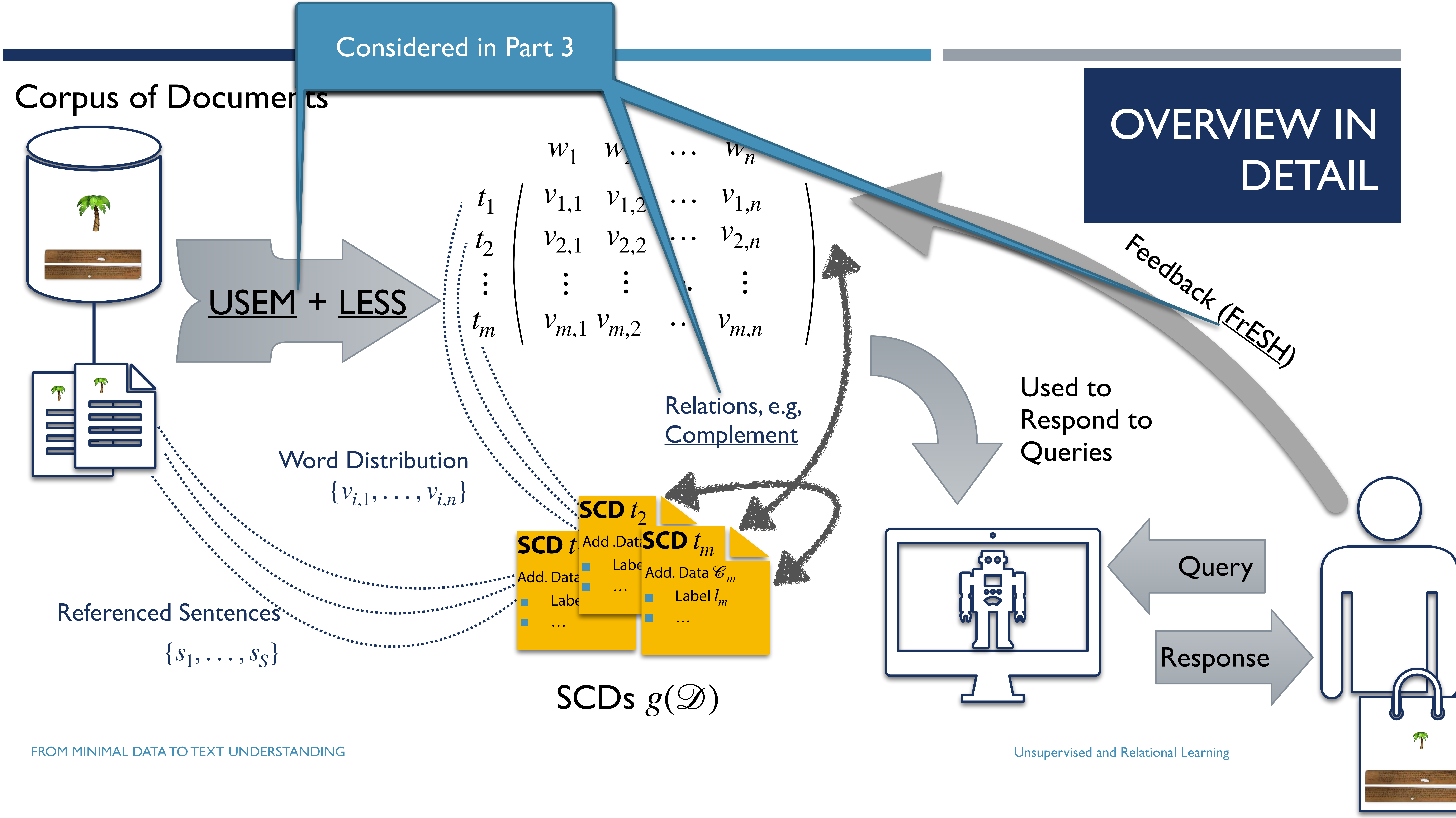
INTERIM SUMMARY

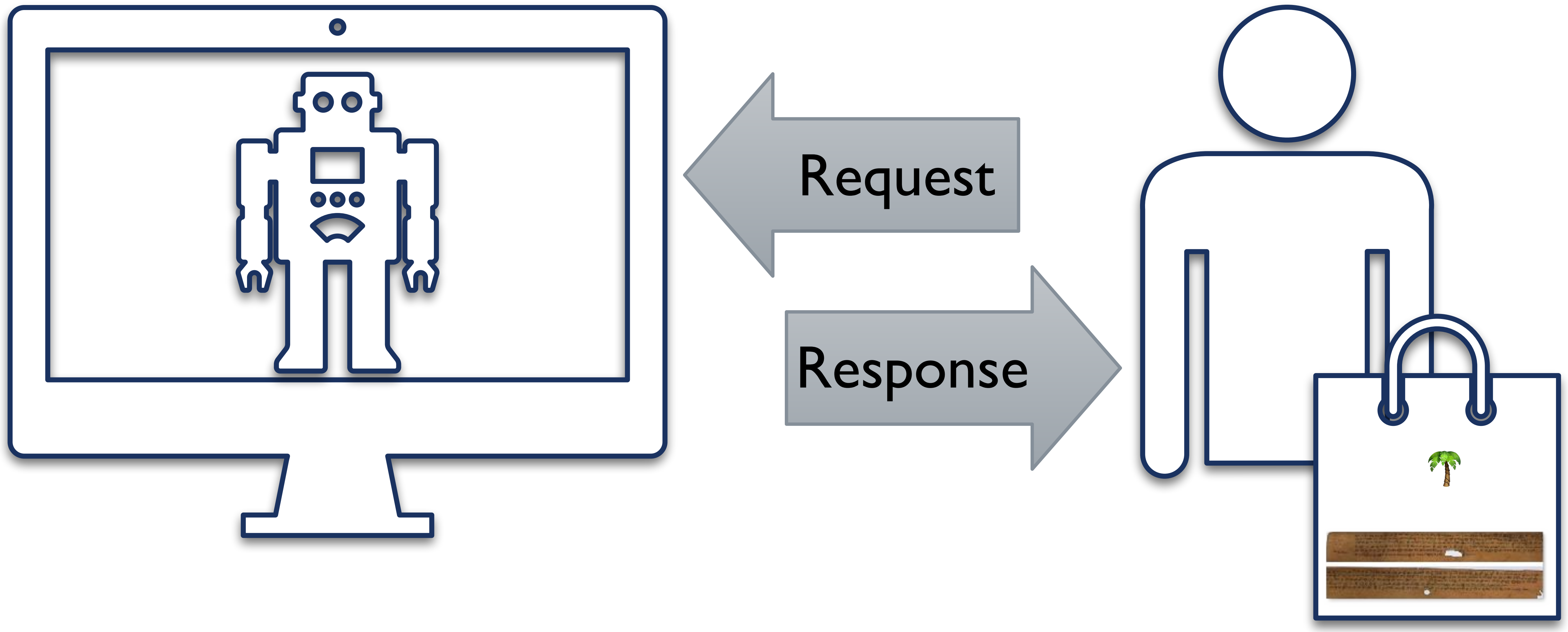


1. Unsupervised Estimation of SCDs
   - SCDs (an SCD matrix) for any corpus
2. Continuous Improvement by Feedback
   - Feedback from users used to update and enhance SCD matrix
3. Labelling of SCDs
   - SCDs get a human friendly label for display and description
4. Intra- and Inter-SCD Relations
   - Intra: Each SCD references sentences, has word distribution, and data incl. label
   - Inter: SCDs have relations, e.g., complement, among each other

➡ Apply SCD on any corpus (e.g., small and without initial SCDs) to help understanding it

OVERVIEW IN DETAIL





INFORMATION SYSTEM

BASED ON SCDS USING USEM, LESS, FRESH, …

Backup

Magnus Bender, Tanya Braun, Ralf Möller, Marcel Gehrke **Unsupervised Estimation of Subjective Content Descriptions in an Information System** in *International Journal of Semantic Computing, 2023*

← Previous ↖ Parent → Next

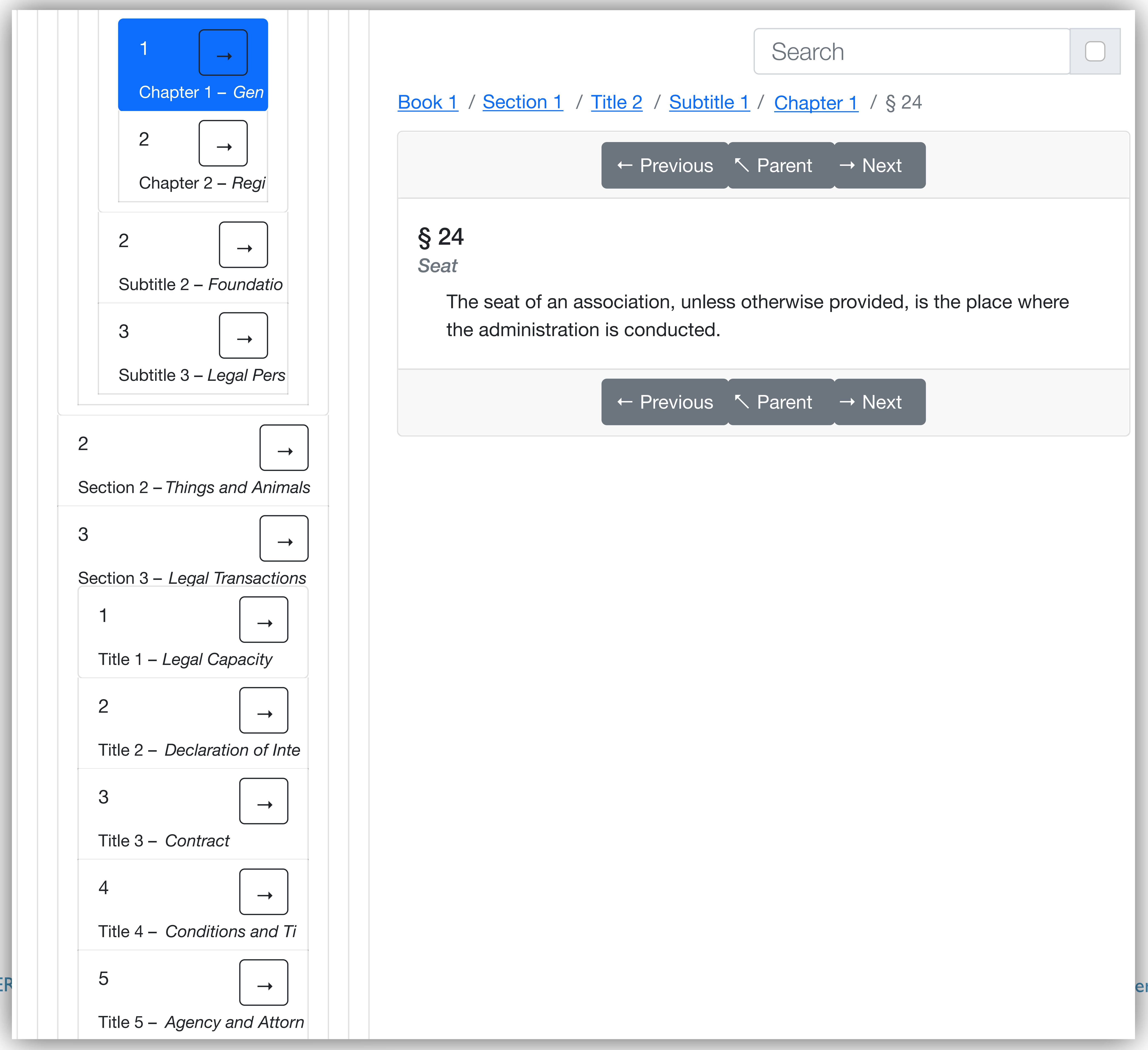
§ 83

Foundation upon Death

[1]If the foundation transaction consists of a disposition upon death, the probate court shall notify this to the competent authority for recognition, unless requested by the heir or the executor. [2]If the foundation transaction does not satisfy the requirements of § **81** (1) **1** sentence 3, the foundation shall be given articles of association by the competent authority before recognition or incomplete articles of association shall be supplemented; the will of the founder shall be taken into account. [3] The seat of an foundation, unless otherwise provided, is the place where the administration is conducted. [4]In case of doubt, the founder's last domicile in Germany shall be deemed to be the registered office.

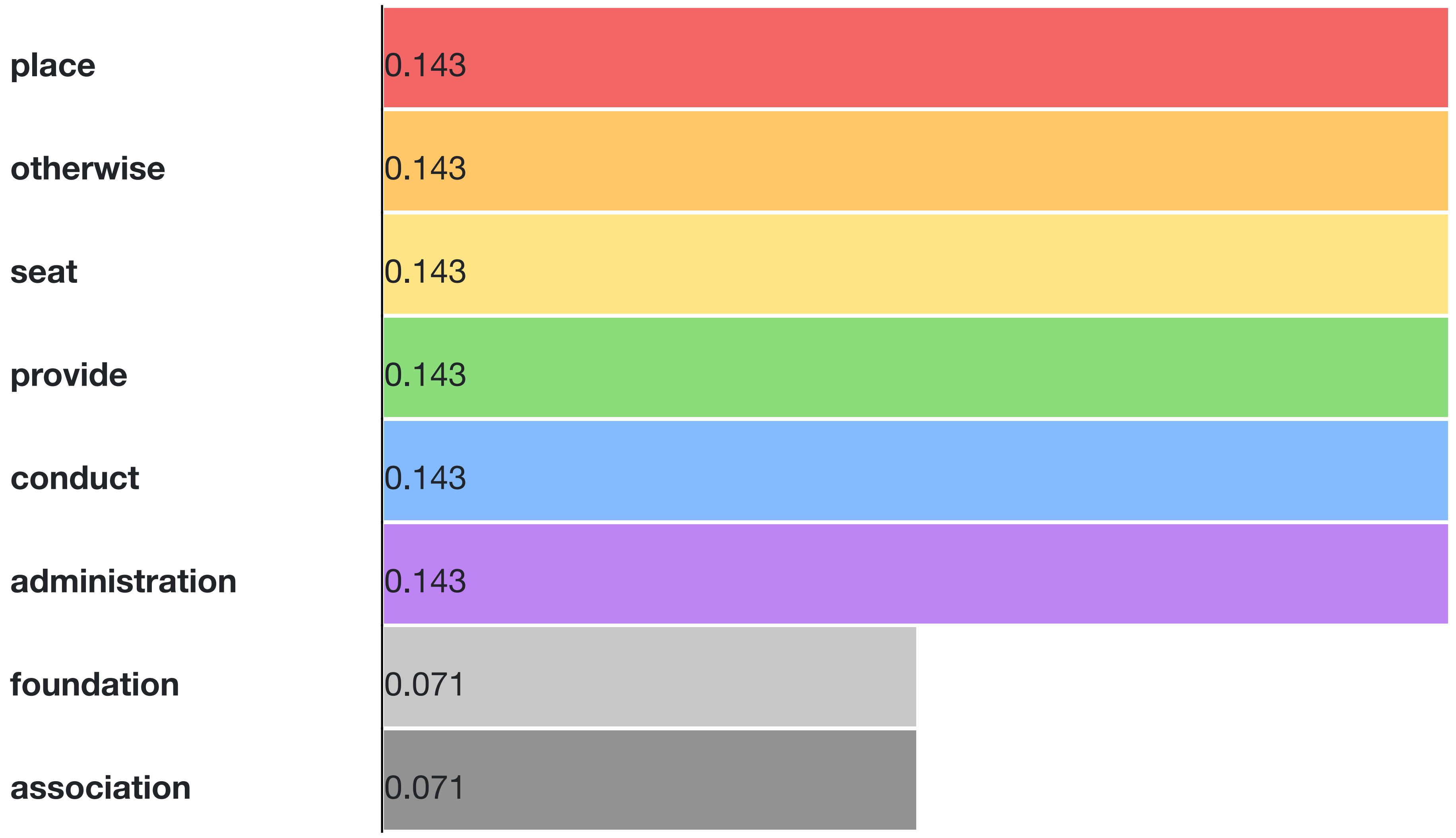
← Previous ↖ Parent → Next

Hover SCD Windows


1
Chapter 1 – Gen
2
Chapter 2 – Regi
2
Subtitle 2 – Foundatio
3
Subtitle 3 – Legal Pers
2
Section 2 – Things and Animals
3
Section 3 – Legal Transactions
1
Title 1 – Legal Capacity
2
Title 2 – Declaration of Inte
3
Title 3 – Contract
4
Title 4 – Conditions and Ti
5
Title 5 – Agency and Attorn
Search
Book 1 / Section 1 / Title 2 / Subtitle 1 / Chapter 1 / § 24
← Previous ↖ Parent → Next
§ 24
Seat
The seat of an association, unless otherwise provided, is the place where the administration is conducted.
← Previous ↖ Parent → Next

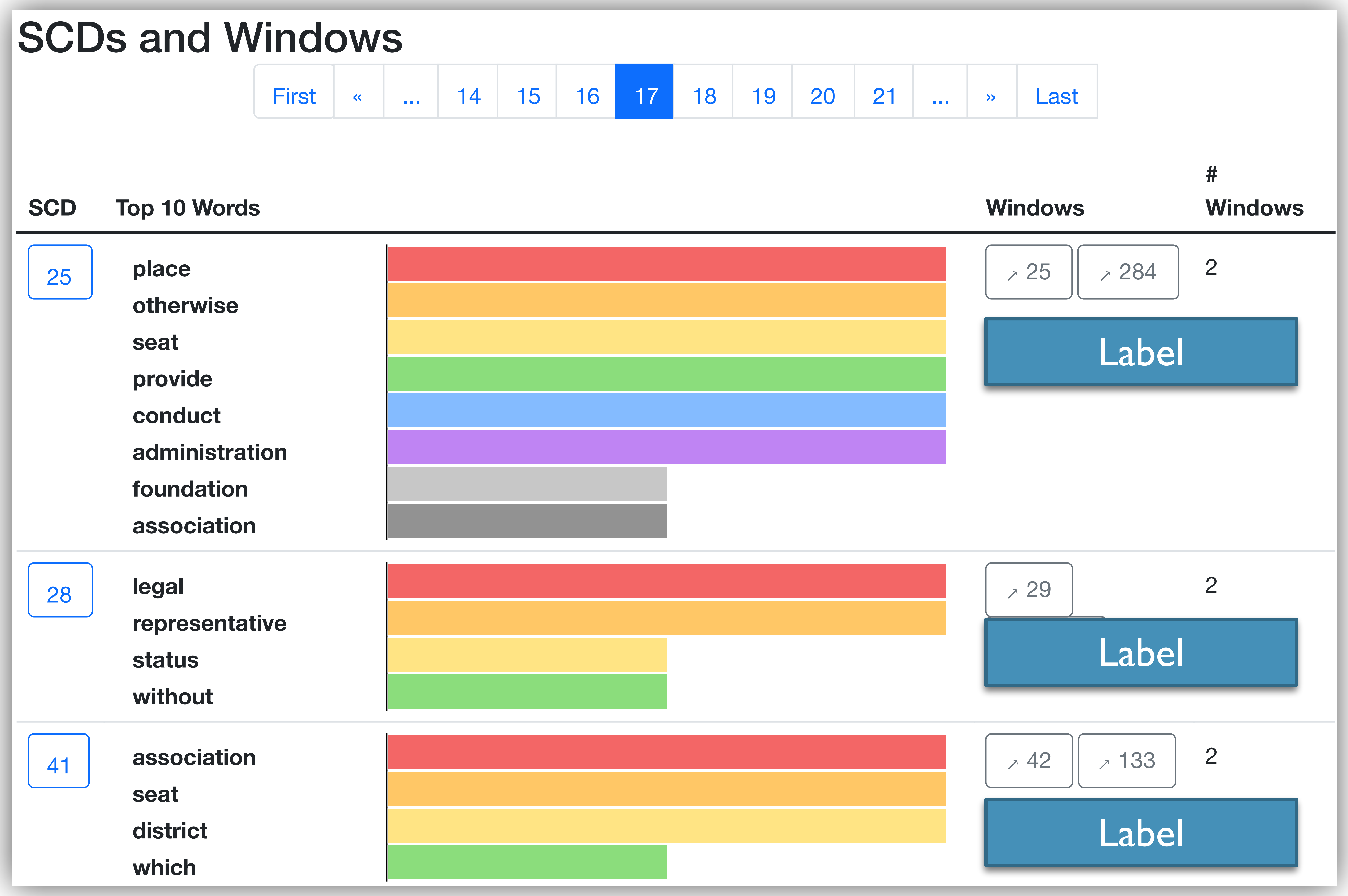
← Entire Matrix | ← Previous | SCD ID 25 | Next →

Words of this SCD's Distribution

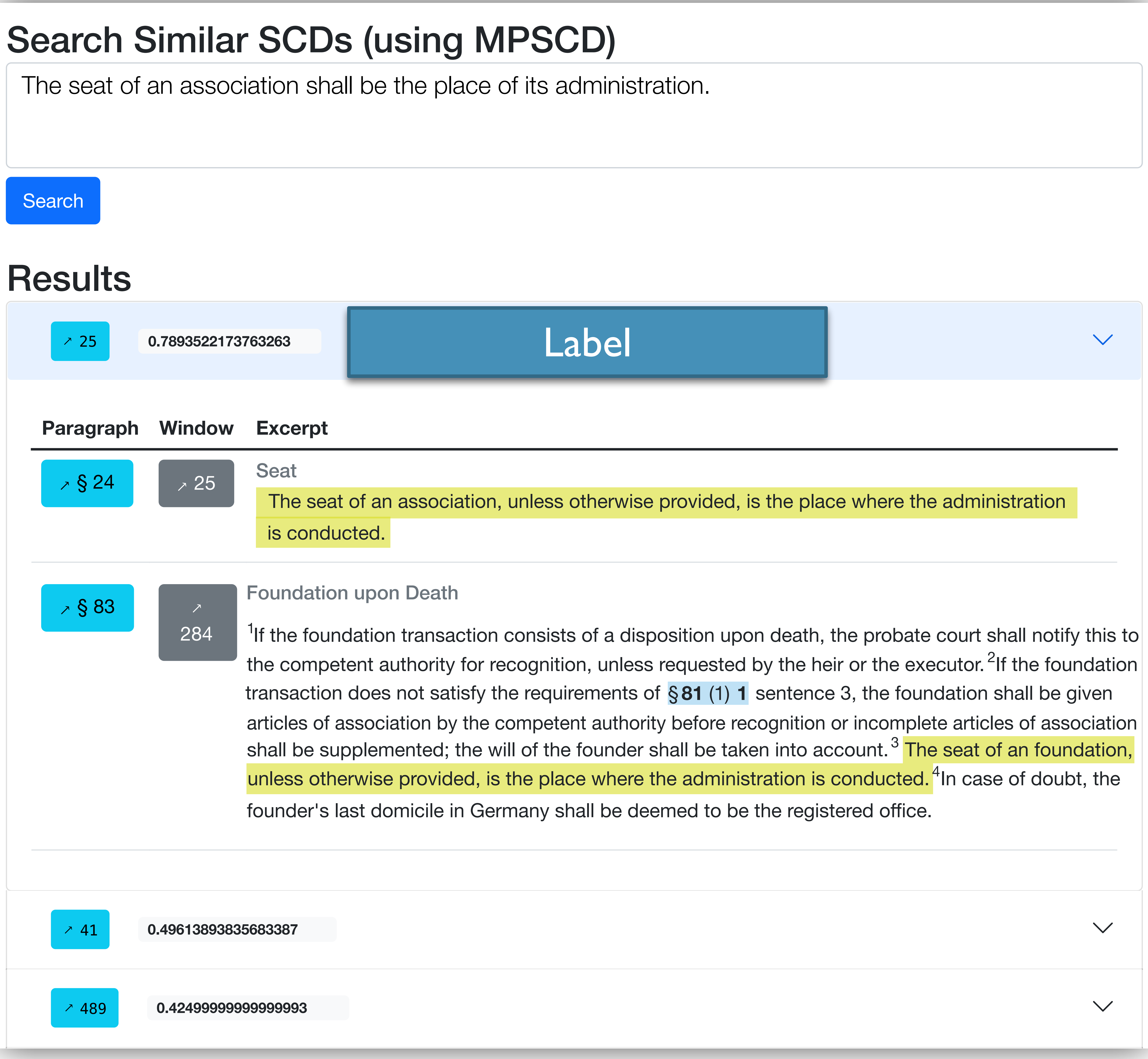


Windows of this SCD 2

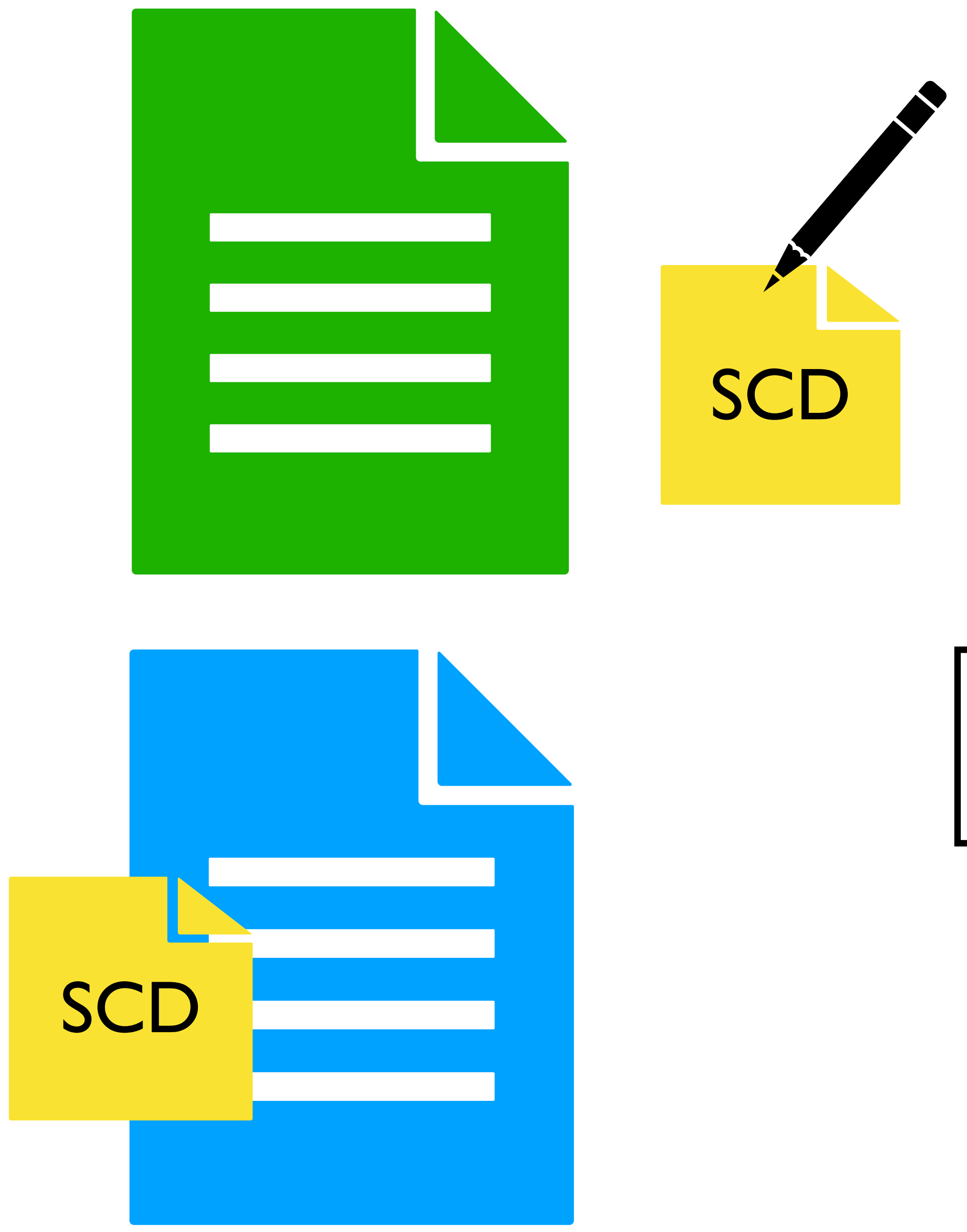
Paragraph	Window	Excerpt
↗ § 24	↗ 25	Seat The seat of an association, unless otherwise provided, is the place where the administration is conducted.
↗ § 83	↗ 284	Foundation upon Death [1]If the foundation transaction consists of a disposition upon death, the probate court shall notify this to the competent authority for recognition, unless requested by the heir or the executor. [2]If the foundation transaction does not satisfy the requirements of § **81** (1) **1** sentence 3, the foundation shall be given articles of association by the competent authority before recognition or incomplete articles of association shall be supplemented; the will of the founder shall be taken into account. [3] The seat of an foundation, unless otherwise provided, is the place where the administration is conducted. [4]In case of doubt, the founder's last domicile in Germany shall be deemed to be the registered office.

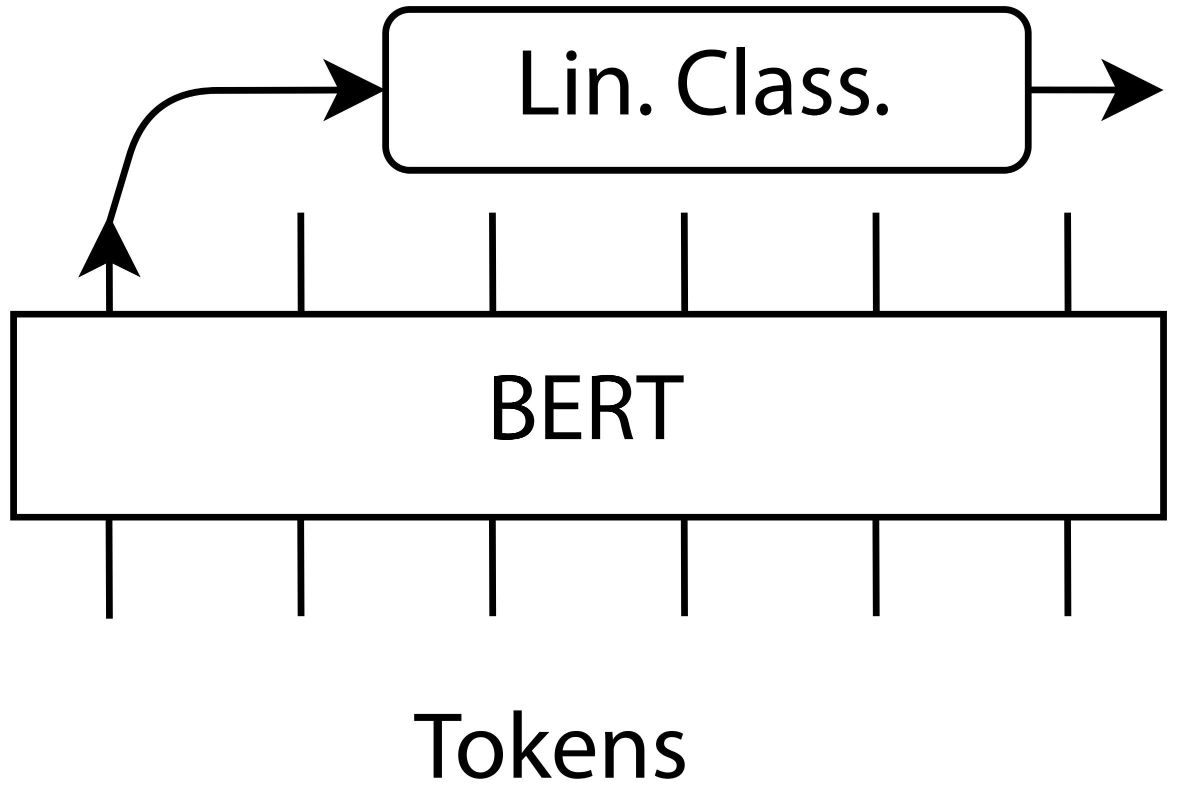


SCDs and Windows
First
«
...
14
15
16
17
18
19
20
21
...
»
Last
SCD
Top 10 Words
Windows
Windows
25
place
otherwise
seat
provide
conduct
administration
foundation
association
25
284
2
Label
28
legal
representative
status
without
29
2
Label
41
association
seat
district
which
42
133
2
Label


Search Similar SCDs (using MPSCD)
The seat of an association shall be the place of its administration.
Search
Results
↗ 25
0.7893522173763263
Label
Paragraph
Window
Excerpt
↗ § 24
↗ 25
Seat
The seat of an association, unless otherwise provided, is the place where the administration is conducted.
↗ § 83
↗ 284
Foundation upon Death
1If the foundation transaction consists of a disposition upon death, the probate court shall notify this to the competent authority for recognition, unless requested by the heir or the executor. 2If the foundation transaction does not satisfy the requirements of §81 (1) 1 sentence 3, the foundation shall be given articles of association by the competent authority before recognition or incomplete articles of association shall be supplemented; the will of the founder shall be taken into account. 3 The seat of an foundation, unless otherwise provided, is the place where the administration is conducted. 4In case of doubt, the founder's last domicile in Germany shall be deemed to be the registered office.
↗ 41
0.49613893835683387
↗ 489
0.42499999999999993

1. Identify SCDs among text – iSCD
2. Most probably suited SCD – MPS²CD



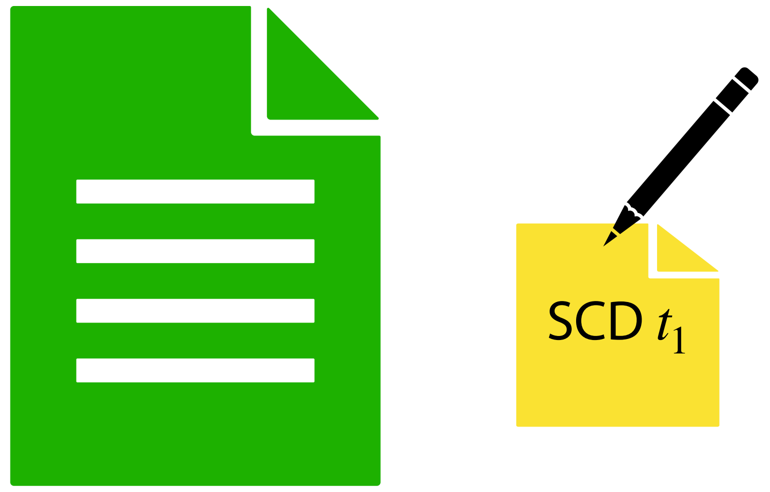


LLMS IN REPLACEMENT FOR SCDS

ESTIMATING CONTEXT-SPECIFIC SCDS USING BERT

Backup

Magnus Bender, Felix Kuhr, Tanya Braun, Ralf Möller **Estimating Context-Specific Subjective Content Descriptions using BERT** in *16th IEEE International Conference on Semantic Computing (ICSC 2022)*



TASK:
APPLY BERT ON SCDS

1 Identify SCDs among text – iSCD

- Given a text document d' where SCDs and content are interleaved
- Asked for set $g(d)$ containing SCDs and the content of text document $d \subseteq d'$

$d' =$ (“We visited the bisons <u>large animals</u> in the zoo <u>a place where non-domestic animals are exhibited</u>.”)

2 Most probably suited SCD – MPS²CD

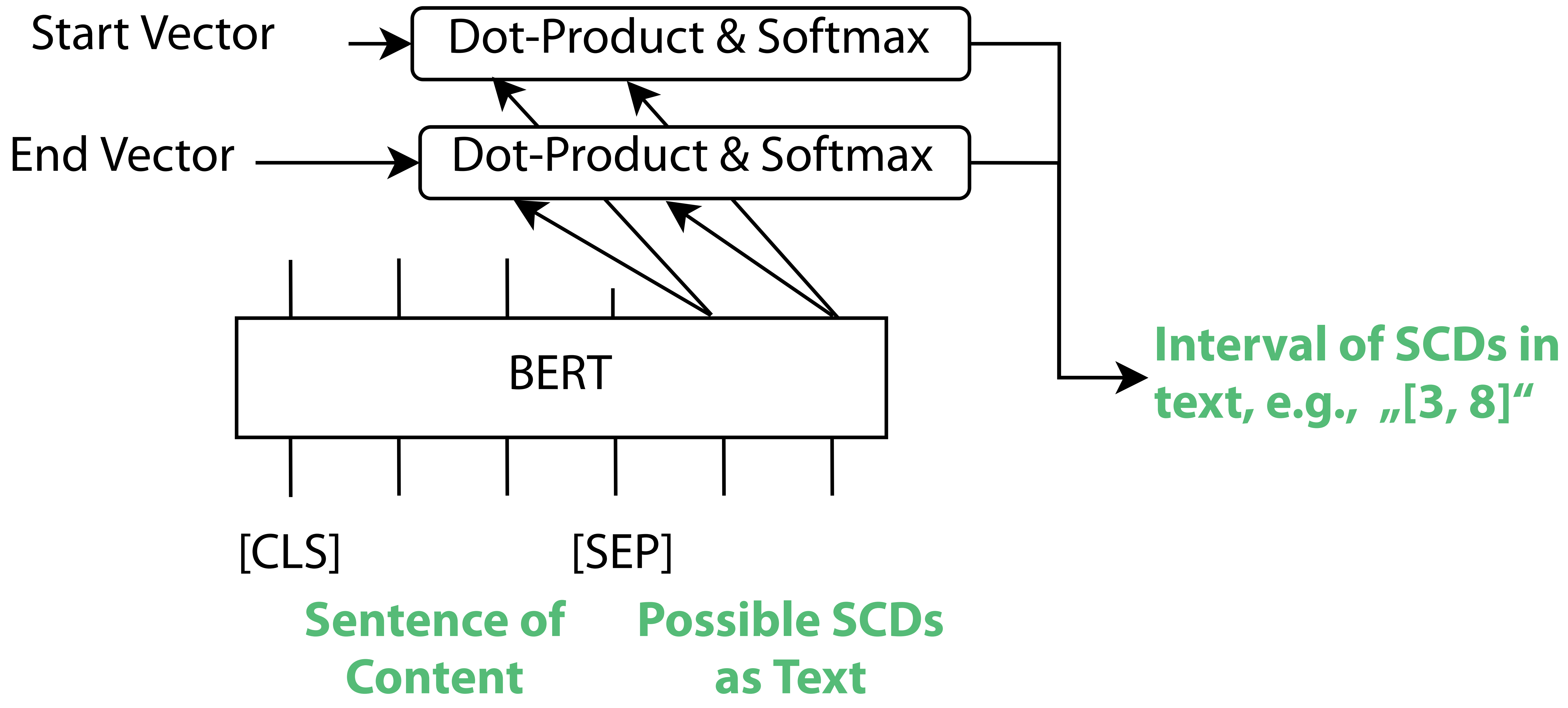
- Given a text document d without associated SCDs
- Asked for set $g(d)$ containing best suited SCDs t for d

$d =$ (“We visited the bisons in the zoo.”)

$d =$ (“We visited the bisons in the zoo.”)
$g(d) =$ { (“large animals”, 4),
(“a place where non-domestic animals are exhibited”, 7) }

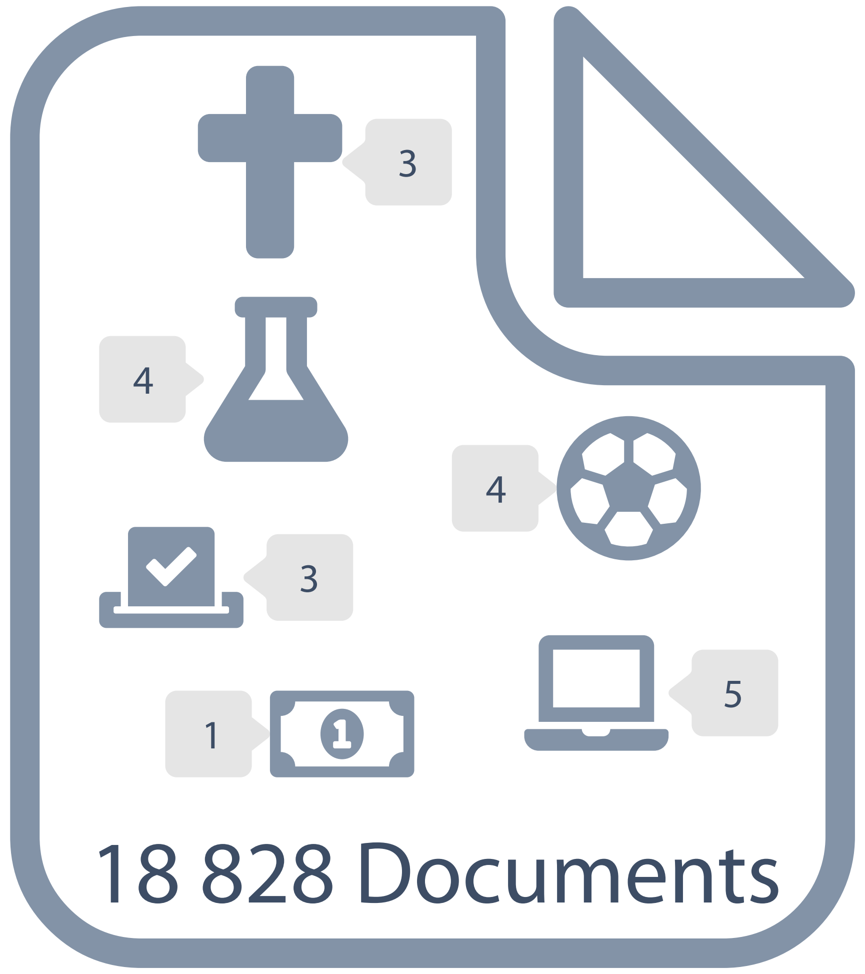
APPROACH: APPLYING BERT ON SCDS

- iSCD
  - BERT Classify
  - BERT Next
- MPS²CD
  - BERT Choose
  - BERT Highlight

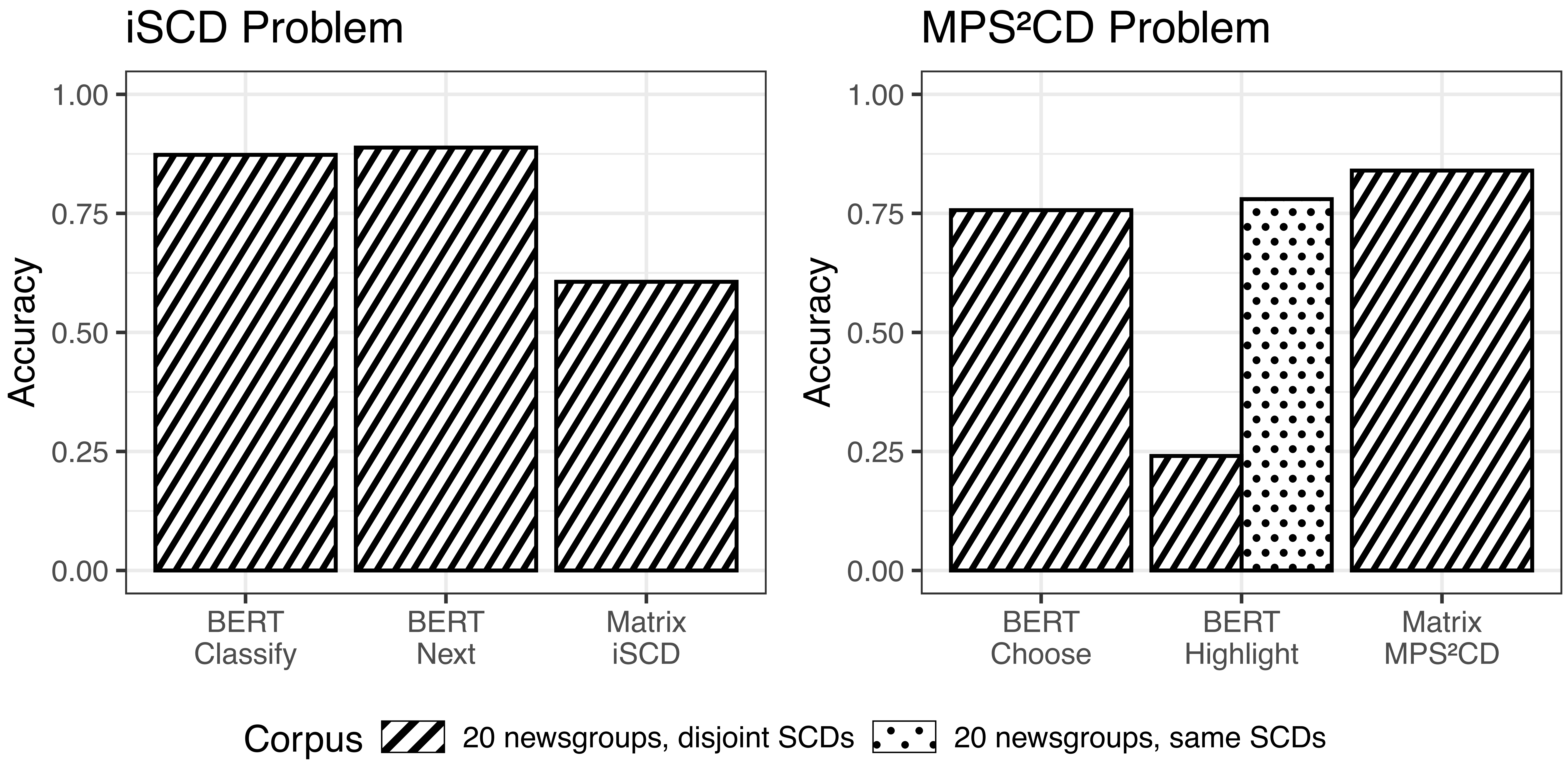


EVALUATION

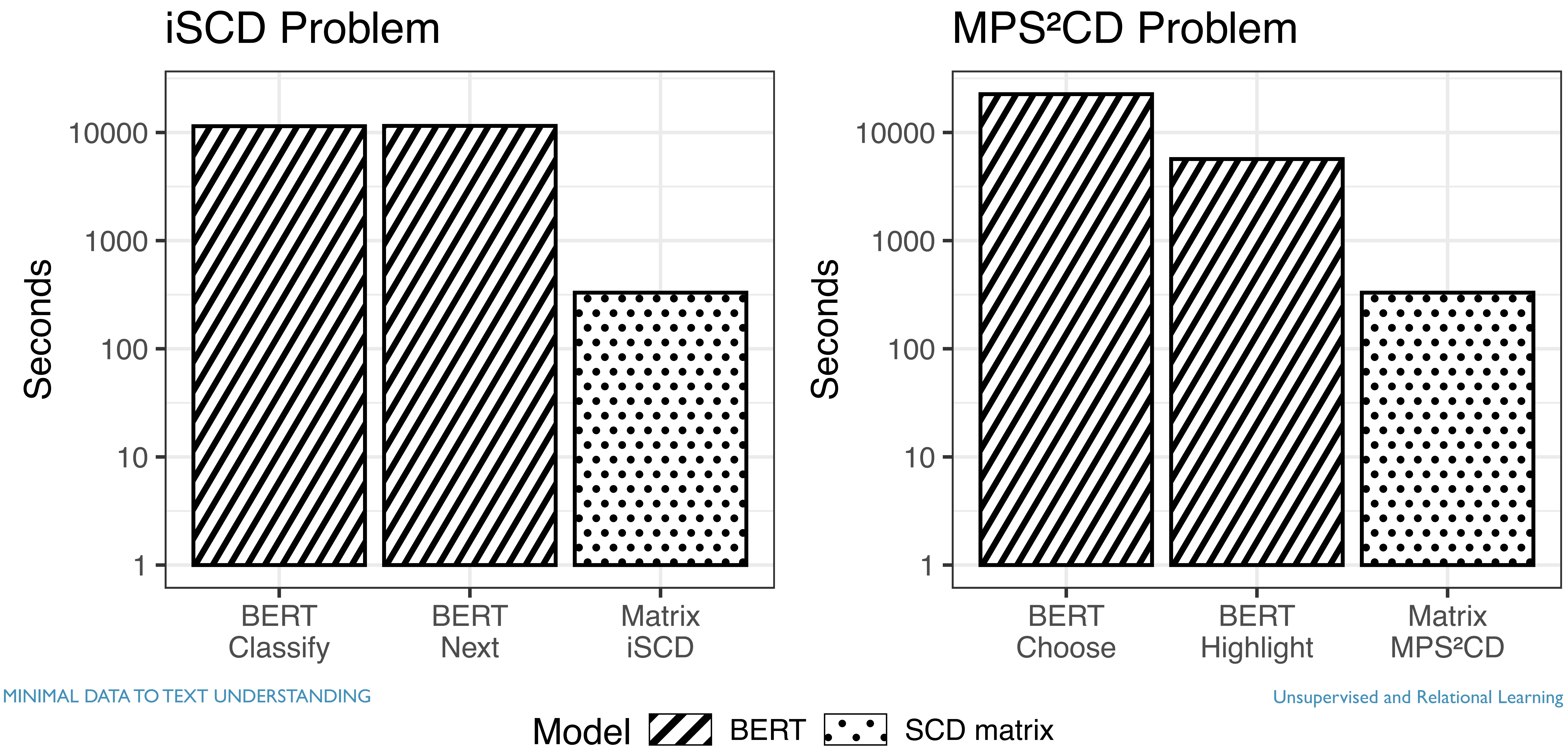
- Corpus
  - 20 newsgroups
  - Definitions from Wiktionary
- Dataset
  - 80 % training and 20 % testing
  - Disjoint and same sets of definitions for SCDs
- Hardware
  - NVIDIA DGX A100 320GB
  - 8 Intel 6248 with 2.50GHz (3.90GHz), 16GB RAM
- Model
  - “Bert-Base-Uncased“



RESULTS: ACCURACY



RESULTS: RUNTIME



CONCLUSION

- BERT and the SCD matrix solve the MPS^{2}CD and iSCD problem well
- BERT needs much more time and computational resources in contrast to the SCD matrix

„We demonstrate that BERT is able to grasp the concept of SCDs, in a way that BERT can be trained to solve SCD-related tasks.“

AGENDA

1. Introduction to Semantic Systems [Tanya]
2. Supervised Learning [Marcel]
3. Unsupervised and Relational Learning [Magnus]
   - Unsupervised Estimation of SCDs
   - Continuous Improvement by Feedback
   - Labelling of SCDs
   - Inter- and Intra-SCD Relations
4. Summary [Tanya]