The German Production Pipeline

Frankfurt - Data Model, Re-Structuring & Linking

Dennis Gram, Pantelis Karapanagiotis, Marius Liebald

Groningen Workshop

May 4, 2023

The Data Model

Dennis Gram

Starting Point

The construction of a database is challenging

- Sources can be ambiguous

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential
  - Merging is complicated

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential
  - Merging is complicated
- Creating datasets manually

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential
  - Merging is complicated
- Creating datasets manually
  - Books, journals, loose-leaf editions

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential
  - Merging is complicated
- Creating datasets manually
  - Books, journals, loose-leaf editions
  - Requires immense resources $\rightarrow$ not feasible

Starting Point

The construction of a database is challenging

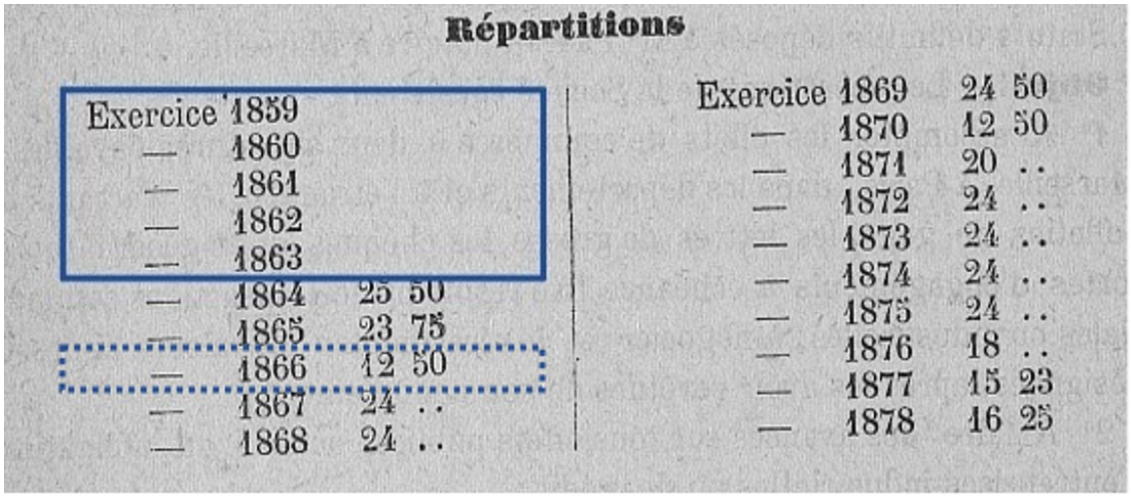
- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential
  - Merging is complicated
- Creating datasets manually
  - Books, journals, loose-leaf editions
  - Requires immense resources $\rightarrow$ not feasible
- Most existing projects do not adhere to common standards

Starting Point

The construction of a database is challenging

- Sources can be ambiguous
  - Incomplete or conflictiong information
- No common identifiers
  - Linking and merging is difficult
  - Metadata standards must be set up and agreed on
- Studies often rely on mixed data
  - Combined datasets are essential
  - Merging is complicated
- Creating datasets manually
  - Books, journals, loose-leaf editions
  - Requires immense resources $\rightarrow$ not feasible
- Most existing projects do not adhere to common standards
  - Individual projects, silo knowledge, no metadata standards

Example: Incomplete or Conflicting information



Répartitions

Exercice	1859	
—	1860	
—	1861	
—	1862	
—	1863	
—	1864	25 50
—	1865	23 75
—	1866	12 50
—	1867	24 ..
—	1868	24 ..
Exercice	1869	24 50
—	1870	12 50
—	1871	20 ..
—	1872	24 ..
—	1873	24 ..
—	1874	24 ..
—	1875	24 ..
—	1876	18 ..
—	1877	15 23
—	1878	16 25

Répartitions annuelles aux actions.

Exerc.	Répartitions. fr.	0/0
1859.....	2.50 ou	4 »
1860.....	11 » —	8.80
1861.....	11 » —	8.80
1862.....	10.85 —	8.68
1863.....	19 » —	15.20
1864.....	25.50 —	20.40
1865.....	23.75 —	19 »
1866.....	23 » ou	18.40
1867.....	24 » —	19.20
1868.....	24 » —	19.20
1869.....	24.50 —	19.60
1870.....	12.50 —	10 »
1871.....	20 » —	16 »
1872.....	24 » —	19.20

Data Model Requirements

A data model

- that covers the specific pecularities of

Data Model Requirements

A data model

- that covers the specific pecularities of
  - historical financial data,

Data Model Requirements

A data model

- that covers the specific pecularities of
  - historical financial data,
  - economic data and

Data Model Requirements

A data model

- that covers the specific pecularities of
  - historical financial data,
  - economic data and
- is flexible enough to reach out for

Data Model Requirements

A data model

- that covers the specific pecularities of
  - historical financial data,
  - economic data and
- is flexible enough to reach out for
  - data of different types (quantitative as well as qualitative)

Data Model Requirements

A data model

- that covers the specific pecularities of
  - historical financial data,
  - economic data and
- is flexible enough to reach out for
  - data of different types (quantitative as well as qualitative)
  - from different historical sources,

Data Model Requirements

A data model

- that covers the specific pecularities of
  - historical financial data,
  - economic data and
- is flexible enough to reach out for
  - data of different types (quantitative as well as qualitative)
  - from different historical sources,
- hence, achieving extensibility.

The German Data Model

- Extensible data model with

The German Data Model

- Extensible data model with
  - a process to digitize and structure large historical datasets.

The German Data Model

- Extensible data model with
  - a process to digitize and structure large historical datasets.
- Proof of concept

The German Data Model

- Extensible data model with
  - a process to digitize and structure large historical datasets.
- Proof of concept
  - Relational implementation of German firm and stock market data for 1920 to 1932, currently working on adding 1896 to 1919.

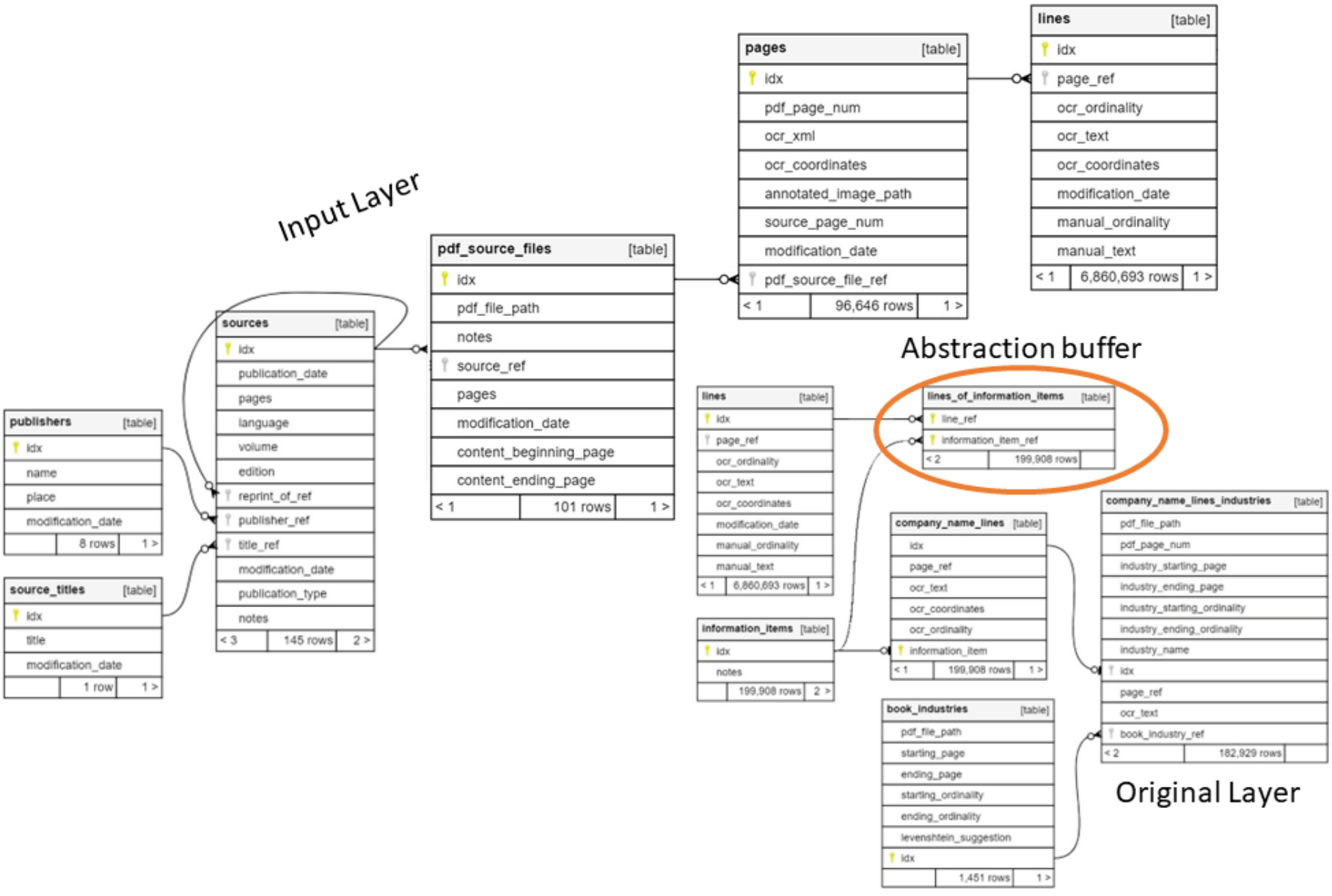
The German Data Model

- Extensible data model with
  - a process to digitize and structure large historical datasets.
- Proof of concept
  - Relational implementation of German firm and stock market data for 1920 to 1932, currently working on adding 1896 to 1919.
- Our implementation allows researchers to validate the digitized data against the original sources.

The German Data Model

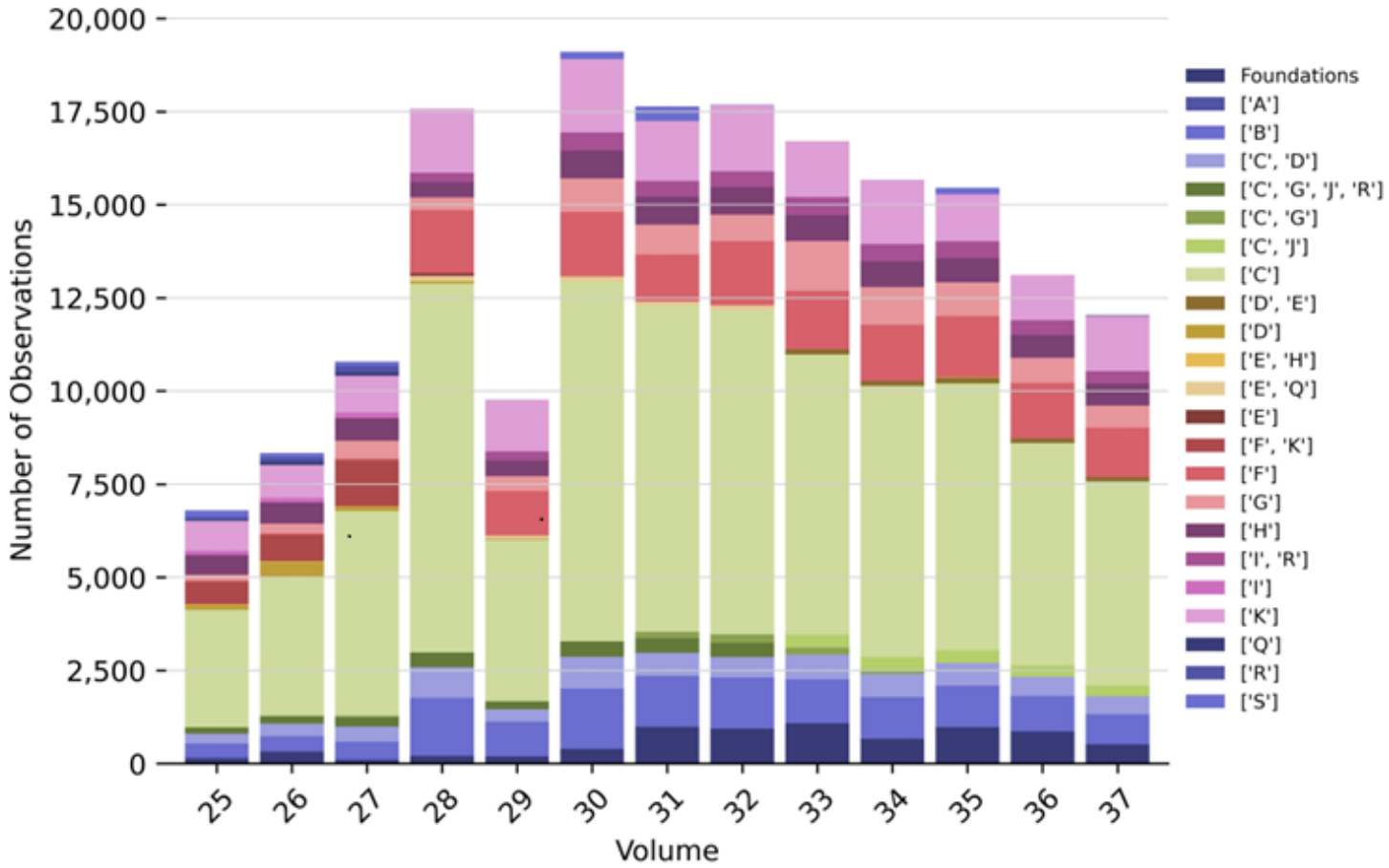
- Extensible data model with
  - a process to digitize and structure large historical datasets.
- Proof of concept
  - Relational implementation of German firm and stock market data for 1920 to 1932, currently working on adding 1896 to 1919.
- Our implementation allows researchers to validate the digitized data against the original sources.
- A detailed description of the data model can be found in Gram *et al.* (2022).

Data Model


Input Layer
Abstraction buffer
Original Layer
publishers [table]
idx
name
place
modification_date
8 rows 1 >
source_titles [table]
idx
title
modification_date
1 row 1 >
sources [table]
idx
publication_date
pages
language
volume
edition
reprint_of_ref
publisher_ref
title_ref
modification_date
publication_type
notes
< 3 145 rows 2 >
pdf_source_files [table]
idx
pdf_file_path
notes
source_ref
pages
modification_date
content_beginning_page
content_ending_page
< 1 101 rows 1 >
pages [table]
idx
pdf_page_num
ocr_xml
ocr_coordinates
annotated_image_path
source_page_num
modification_date
pdf_source_file_ref
< 1 96,646 rows 1 >
lines [table]
idx
page_ref
ocr_ordinality
ocr_text
ocr_coordinates
modification_date
manual_ordinality
manual_text
< 1 6,860,693 rows 1 >
lines [table]
idx
page_ref
ocr_ordinality
ocr_text
ocr_coordinates
modification_date
manual_ordinality
manual_text
< 1 6,860,693 rows 1 >
lines_of_information_items [table]
line_ref
information_item_ref
< 2 199,908 rows
information_items [table]
idx
notes
199,908 rows 2 >
company_name_lines [table]
idx
page_ref
ocr_text
ocr_coordinates
ocr_ordinality
information_item
< 1 199,908 rows 1 >
book_industries [table]
pdf_file_path
starting_page
ending_page
starting_ordinality
ending_ordinality
levenshtein_suggestion
idx
1,451 rows 1 >
company_name_lines_industries [table]
pdf_file_path
pdf_page_num
industry_starting_page
industry_ending_page
industry_starting_ordinality
industry_ending_ordinality
industry_name
idx
page_ref
ocr_text
book_industry_ref
< 2 182,929 rows

Firm Industry Distribution per Volume

NACE Rev. 2 Section	Description
A	Agriculture, Forestry and Fishing
B	Mining and Quarrying
C	Manufacturing
D	Electricity, Gas, Steam and Air Conditioning
E	Water Supply, Sewerage, Waste Management and Remediation Activities
F	Construction
G	Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
I	Accomodation and Food Service Activities
H	Transportation and Storage
J	Information and Communication
K	Financial and Insurance Activities
L	Real Estate Activities
M	Professional, Scientific and Technical Activities
N	Administrative and Support Service Activities
O	Public Administration and Defence; Compulsory Social Security
P	Education
Q	Human Health and Social Work Activities
R	Arts, Entertainment and Recreation
S	Other Service Activities
T	Activities of Households as Employers; Undifferentiated Goods- and Services-Producing Activities of Households for Own Use
U	Activities of Extraterritorial Organisations and Bodies



Validate against the original data sources

Lübecker Maschinenbau-Ges. in Lübeck, Direktion

- Digitized data (taken from input layer)
  - {Coup.-V.:,4,J.,(K.), Direktion:,Carl,"Mette","Emil,Wischow.}
- Derived data for information item Direktion (original layer)
  - Carl Mette, Emil Wischow

Lübecker Maschinenbau-Ges. in Lübeck, Prokuristen

- Digitized data (taken from input layer)
  - {Prokuristen:,H.,"Kaeferstein,",Chr.,"Behrens,",W.,"Priegnitz,",E.,"Tscha
- Derived data for information item Prokurist (original layer)
  - H. Kaeferstein, Chr. Behrens, W. Priegnitz, E. Tscharncke, Paul Flitner

Re-Structuring

Marius Liebald

Data Extraction Logic

- Line type recognition
  1. Line types definition
  2. Random forest classifiers using layout features & string similarities
  3. ANN classifier using
     - random forest predictions (probabilities)
     - contextual features
- Line collapsing
  0. Non-content line removal
  1. Assign lines to items
  2. Assign items to companies
  3. Assign companies to industries
- Variable parsing
  - Variable specific
  - Useful functions stored in folder
    - string_decomposer
    - date_detection
- Linking (i.e., Matchting)

Line types defintion

1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie."

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion

1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie.“

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion

1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie."

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion

1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie."

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion

1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie."

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion

1454 Leder-Fabriken.

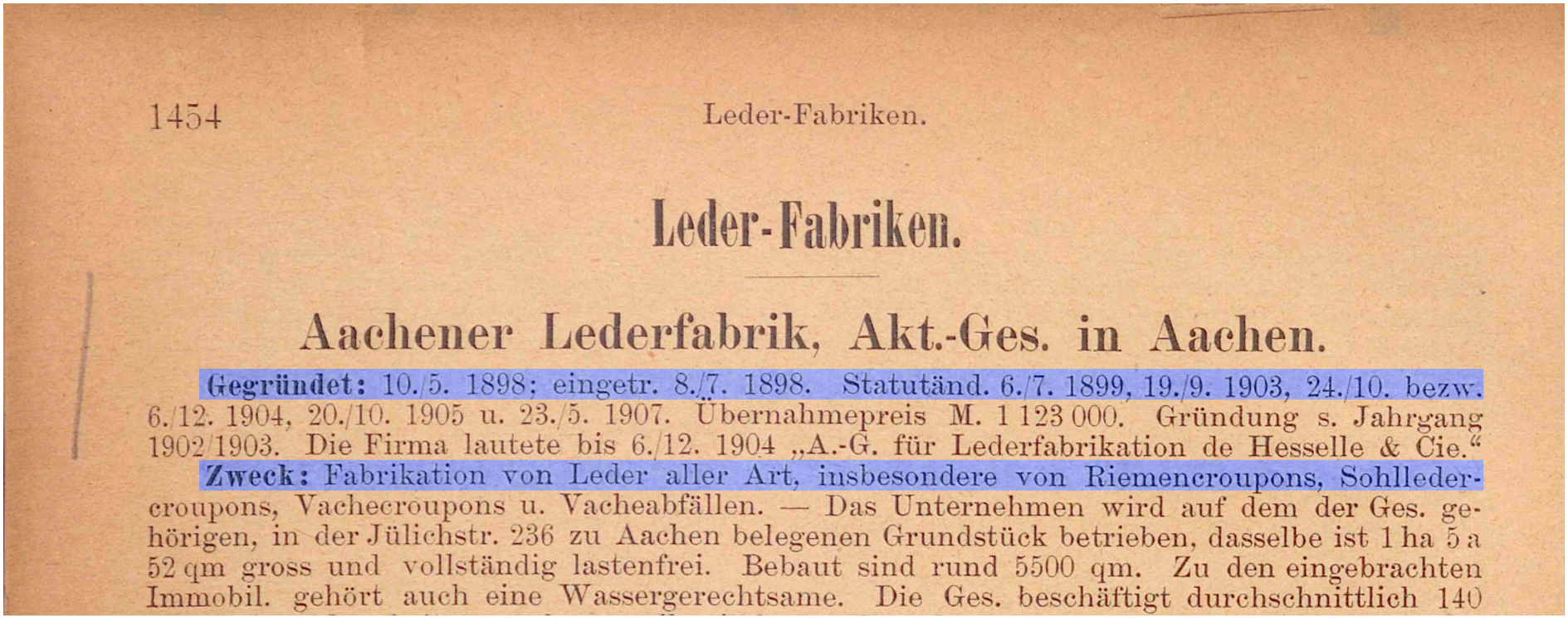
Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie."

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion



1454 Leder-Fabriken.

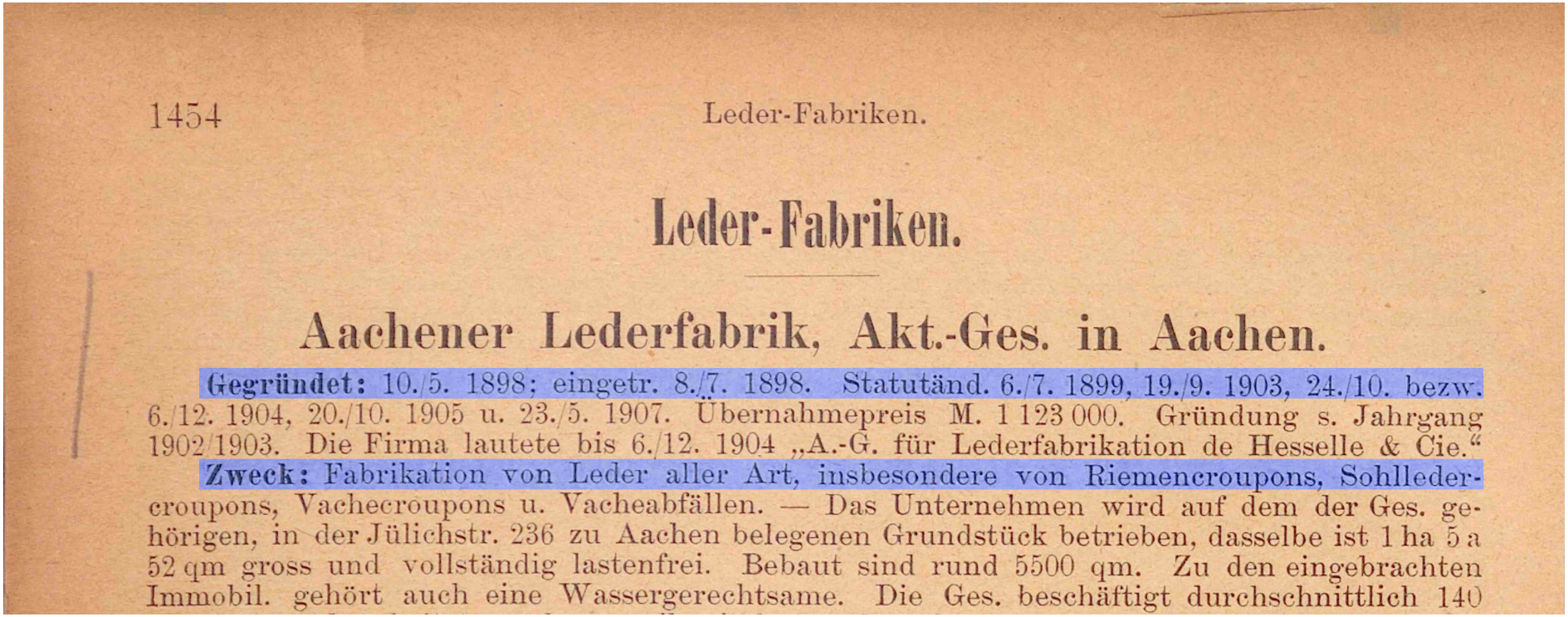
Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie.“

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlledercroupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

Line types defintion



1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie.“

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlleder-croupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140

→ In total, 19 different line types

Layout variables

- (line) textbox width
- (line) textbox center
- (line) textbox height
- (line) textbox height-width ratio
- **vertical distance to (line) textbox above/below**
- **avg. (word) textbox height**
- share of alphabetical characters
- share of non-alphanumerical characters
- **string distance to any industry**
- string distance to any item
- ...

One random forest classifier per line type (and document)

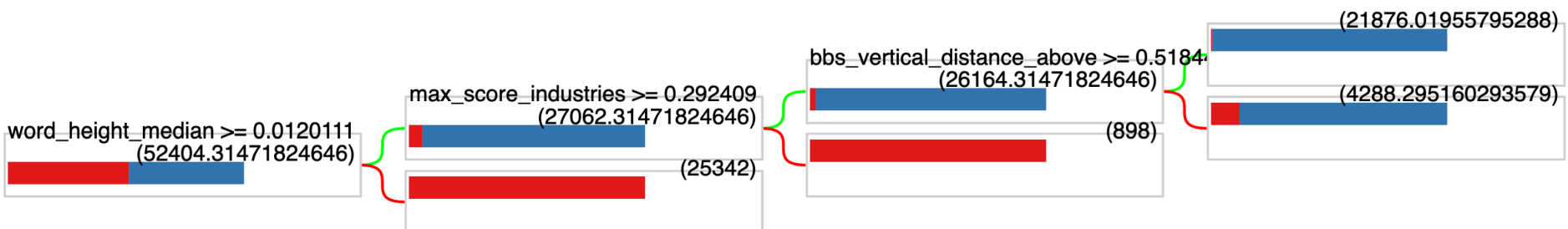
1454 Leder-Fabriken.

Leder-Fabriken.

Aachener Lederfabrik, Akt.-Ges. in Aachen.

Gegründet: 10./5. 1898; eingetr. 8./7. 1898. Statutänd. 6./7. 1899, 19./9. 1903, 24./10. bezw. 6./12. 1904, 20./10. 1905 u. 23./5. 1907. Übernahmepreis M. 1 123 000. Gründung s. Jahrgang 1902/1903. Die Firma lautete bis 6./12. 1904 „A.-G. für Lederfabrikation de Hesselle & Cie.“

Zweck: Fabrikation von Leder aller Art, insbesondere von Riemencroupons, Sohlleder-croupons, Vachecroupons u. Vacheabfällen. — Das Unternehmen wird auf dem der Ges. gehörigen, in der Jülichstr. 236 zu Aachen belegenen Grundstück betrieben, dasselbe ist 1 ha 5 a 52 qm gross und vollständig lastenfrei. Bebaut sind rund 5500 qm. Zu den eingebrachten Immobil. gehört auch eine Wassergerechtsame. Die Ges. beschäftigt durchschnittlich 140



ANN classifier

Random forest prediction output

- For each line: one probability score for each line type
- Probability scores range between 0 and 1

ANN classifier input (for line i)

- random forest prediction $_{i+n}$ $\forall n \in [-5, 5]$
- for instance
  - if $n = 0 \rightarrow$ prediction of line i
  - if $n = 1 \rightarrow$ prediction of line below line i
  - if $n = -1 \rightarrow$ prediction of line above line i

Direktion: Rud. Türk. **Prokuristen:** Fabrik-Direktor A. Rommeney; Kassierer Max Diessler, Fritz Best.

Aufsichtsrat: (3—7) Vors. Rentier Arthur Pekrun, Weisser Hirsch; Stellv. Bank-Dir. Max Gentner, Rechtsanw. Dr. Elb, Dresden; Komm.-Rat Fr. Lindemann, Halberstadt; Dir. Curt Fochtmann, Meissen.

Zahlstellen: Eigene Kassen; Dresden u. Wernigerode: Mitteldeutsche Privatbank; Wernigerode: Heinr. Schmidt; Halberstadt: Mooshake & Lindemann. *

Gust. Schaeuffelen'sche Papierfabrik in Heilbronn a. N.

Bilanz am 30. Juni 1909: Aktiva: Immobil. 1 161 111, Masch. 664 711, Vorräte 594 737, Bar 2596, Wechsel 27 361, Effekten 6500, Debit. 366 282. — Passiva: A.-K. 857 142, R.-F. 40 017, Einlagen 459 481, Hypoth.-Anleihen 1 100 000, Kredit. 339 913, Gewinn 26 745. Sa. M. 2 823 301.

Gewinn- u. Verlust-Konto: Debet: Betriebsausgaben 1 429 120, Gewinn 26 745. Sa. M. 1 455 865. — Kredit: Betriebseinnahmen M. 1 455 865.

Direktion: Carl Schaeuffelen, G. Hub. **Aufsichtsrat:** Vors. A. Schmidt.

Cellulose-Fabrik Hof in Hof-Moschendorf.

Gegründet: 29./6. 1892 durch Übernahme der Fabrik Wiede & Co. in Moschendorf für M. 450 000. Letzte Statutänd. v. 28./4. 1900.

Zweck: Herstellung von Cellulose und damit in Zus.hang stehender Fabrikate.

Kapital: M. 450 000 in 450 Aktien à M. 1000.

1. Assign lines to items
2. Assign items to companies
3. Assign companies to industries

Useful functions

```
class Common_operations():
        def __init__(self):
                pass

        def string_decomposer(self, string:str, comp_regex:dict):
                """
                [...]
                """

        def date_detection(self, string:str, extract=True, replace=True):
                """
                [...]
                """
```

string_decomposer()

```
example_pl = """
        Gewinn- u. Verlust-Konto: Debet: Item A 111,
        Item B 222. - Kredit: Item C 333, Item D 444. Sa. M. 777.
        """

result = Common_operations().string_decomposer(
        string= example_pl,
        comp_regex = {'debet':['Debet:'], 'kredit' : ['Kredit', 'Credit:'],
                'sum':['Sa\.?\s?M\.?']}
        )

print(result)
        {
        'debet': 'Item A 111, Item B 222. -',
        'kredit': ': Item C 333, Item D 444.',
        'sum': '777.',
        'begin': 'Gewinn- u. Verlust-Konto:'
        }
```

date_detection()

```
example_str = """Mein Geburtstag ist am 3. Aug. 1990.
Der Meiner Nichte am 8./7. 2020"""

result = Common_operations().date_detection(
        string = example_str,
        extract = True,
        replace = False
        )

print(result)
(
'Mein Geburtstag ist am 3. Aug. 1990. Der Meiner Nichte am 8./7. 2020',
[datetime.date(1990, 8, 3), datetime.date(2020, 7, 8)]
)
```

Linking

Pantelis Karapanagiotis

Matching Data is Hard

- Adam et al., 2021 . Data extraction and matching The EurHisFirm experience. Methodological Advances in the Extraction and Analysis of Historical Data.
- Cule et al., 2020 . Data Connecting Case Study (EurHisFirm M6.2).
- Poukens, 2018 . Report on the Inventory of Data and Sources (EurHisFirm D4.2).
- Karapanagiotis, 2019 . Technical document on national data models (EurHisFirm D5.1).

Can we Develop Reusable Tools?

User requirements:

1. meaningful match suggestions

Can we Develop Reusable Tools?

User requirements:

1. meaningful match suggestions
2. in reasonable execution time

Can we Develop Reusable Tools?

User requirements:

1. meaningful match suggestions
2. in reasonable execution time
3. using information from multiple entity characteristics

Can we Develop Reusable Tools?

User requirements:

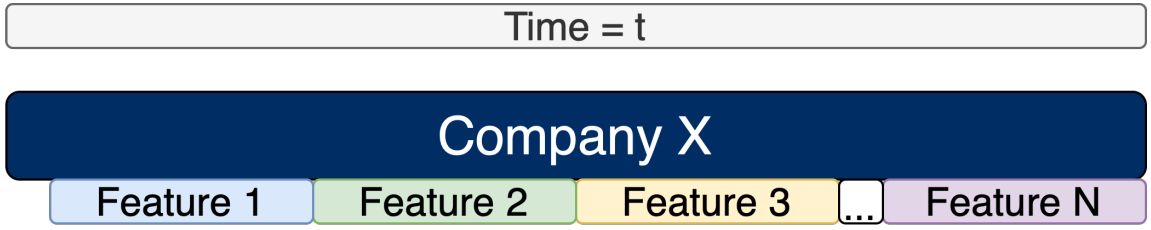
1. meaningful match suggestions
2. in reasonable execution time
3. using information from multiple entity characteristics
4. applicable in different matching contexts

Can we Develop Reusable Tools?

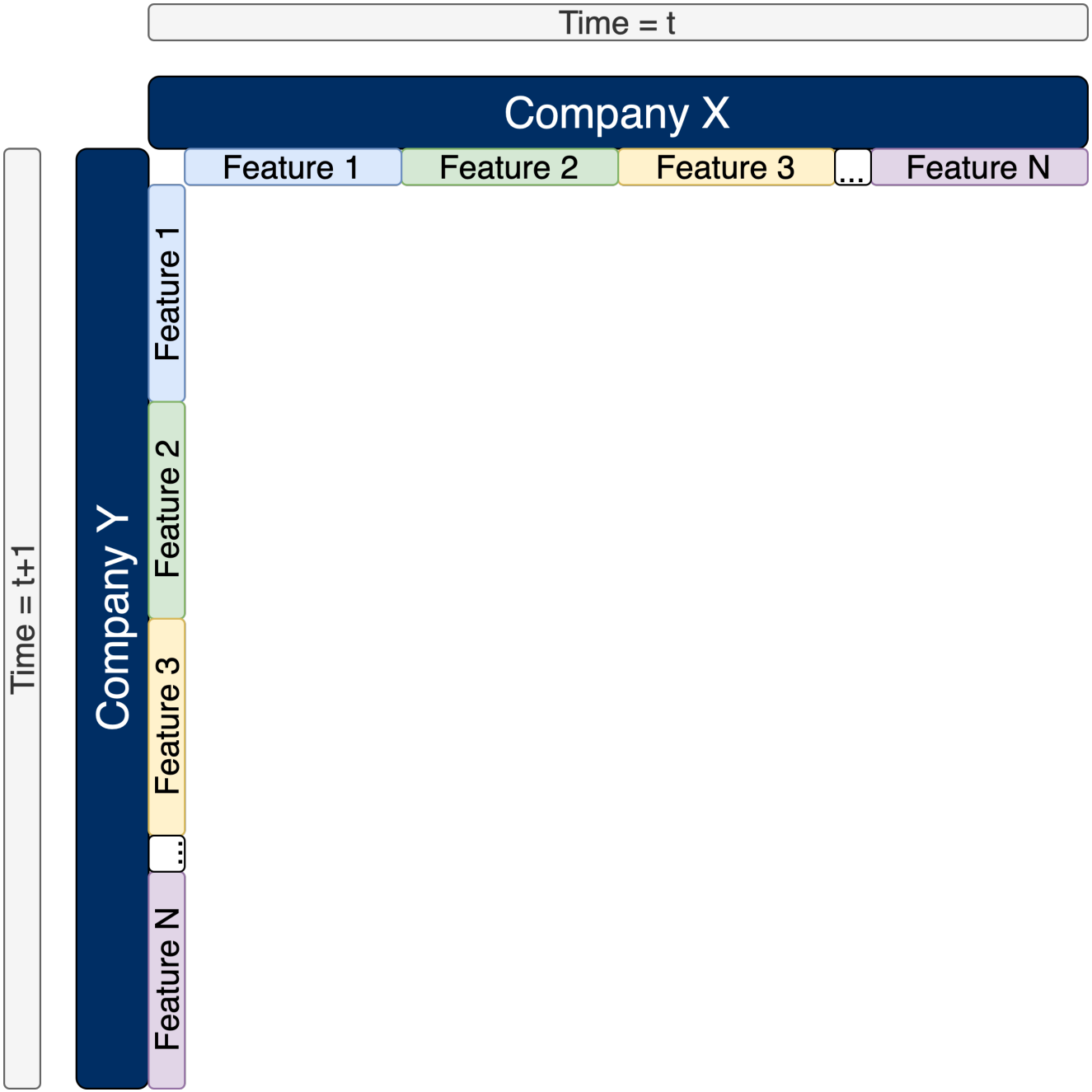
User requirements:

1. meaningful match suggestions
2. in reasonable execution time
3. using information from multiple entity characteristics
4. applicable in different matching contexts
   - in particular for heterogeneous country data

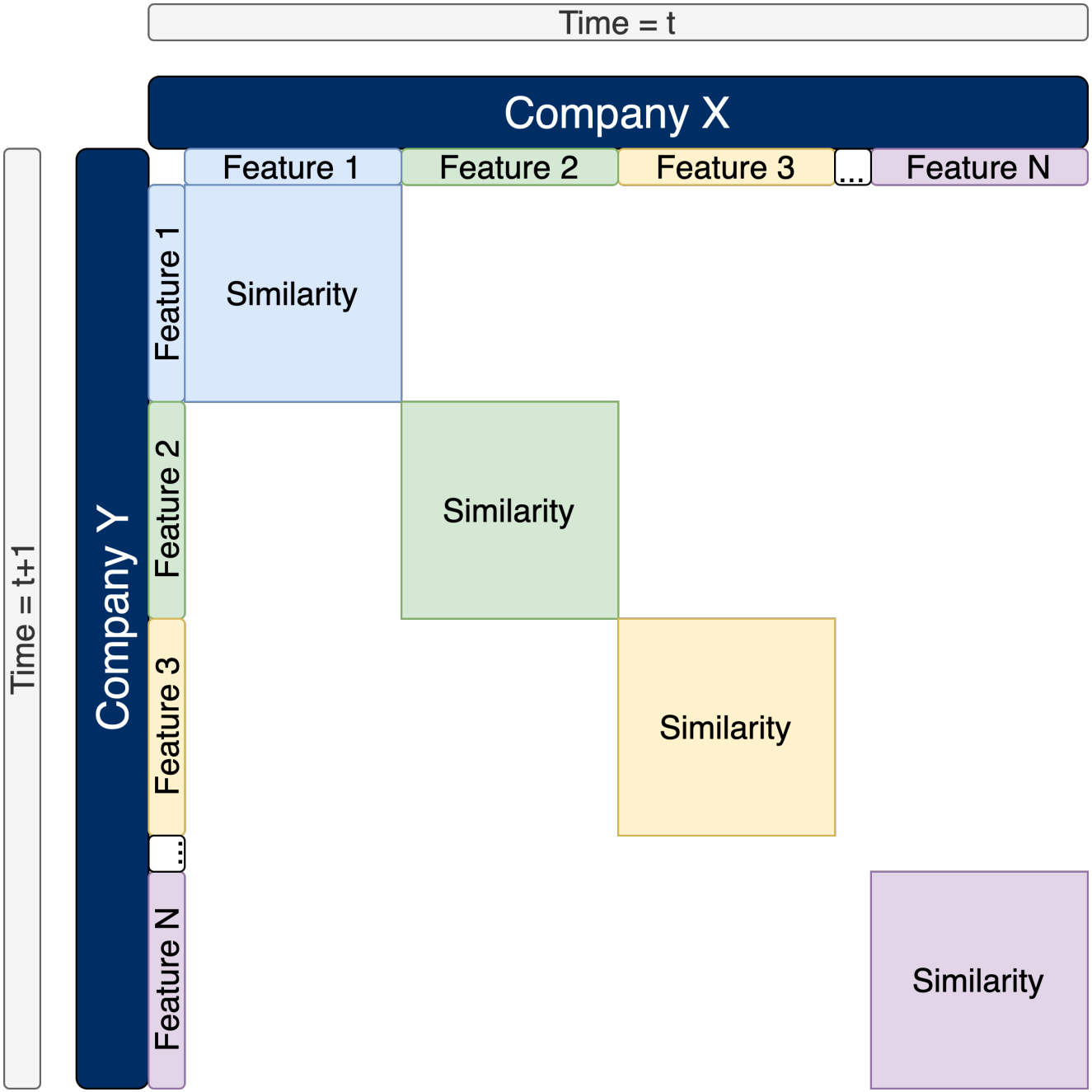
A Matching Problem Example



A Matching Problem Example



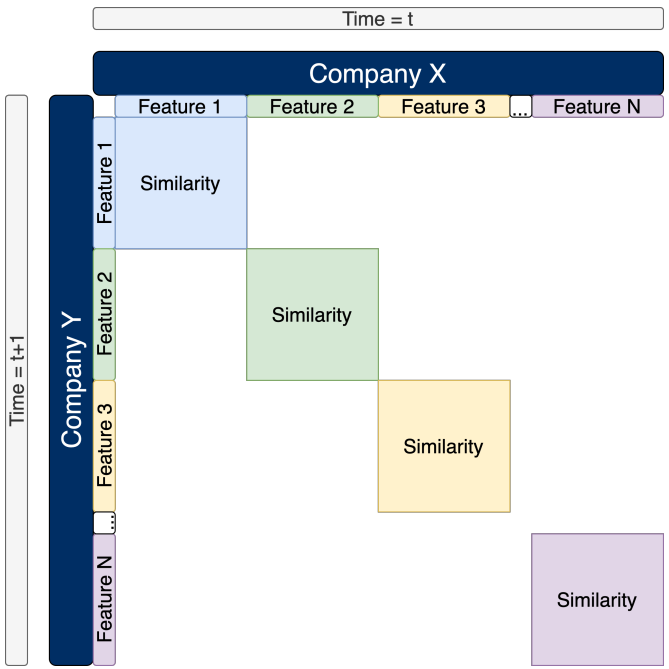
A Matching Problem Example



A Matching Problem Example

Using multiclass classification

- Use the firms of Right as output classes.
- Train a model to classify each f_l from *Left* to one of the output classes.
- But:
  - Say, *Right* is the master database.
  - It gets updated and now contains additional firms.
  - The model will not give classification probabilities for the new firms.



Matching as a Binary Classification Problem

- Instead, we can approach the problem from another perspective.
- Make pairs of records for each firm f_l in the *Left* and f_r in the *Right* data.
- Classify the pairs (f_l, f_r) as a match $label = 1$ or no match $label = 0$.
- But:
  - This requires cross-joining the *Left* and *Right* data.
  - The memory requirements to store the transformed data can quickly render the solution infeasible.
  - Even for small data sources, say $N_L = 5 \cdot 10^3$ and $N_R = 10^4$, the matching pairs $N_{LM} = 5 \cdot 10^7$ require 100s of Gb.

Matching with Blocking

- One solution is to use blocking (Doll et al., 2021).
- Exclude some potential pairs based on pre-defined criteria before training.
- E.g., match *Left* with firms from *Right* that have the save foundation year.
- This effectively reduces the memory requirements problem.
- But:
  -
  - The blocking criteria require having already expertise with the data,
  - are not re-usable in different contexts, and
  - might even be different for heterogeneous country data.
- Can we do better?

Matching with a Similarity Encoder

- Encoders are used in natural language processing models (see e.g. Vaswani et al., 2017).
- They reduce text data to vectors used to train models.
- These models use a huge amount of data.
- The encoding is calculated on the fly.
- How can we use this idea?

Matching with a Similarity Encoder

Using a similarity encoder

1. Pick a pair (f_l, f_r) of *Left* and *Right* firms.

Matching with a Similarity Encoder

Using a similarity encoder

1. Pick a pair (f_l, f_r) of *Left* and *Right* firms.
2. Instruct how the features of f_l and f_r are associated.

Matching with a Similarity Encoder

Using a similarity encoder

1. Pick a pair (f_l, f_r) of *Left* and *Right* firms.
2. Instruct how the features of f_l and f_r are associated.

```
similarity_map = {
    "company_name": {
        "discrete": discrete_si
        "partial": partial_ratio
    },
    "address": {
        "partial": partial_ratio
    },
    "purpose": {
        "sort": token_sort_ratio
    },
    "foundation": {
        "discrete": discrete_si
        "partial": partial_ratio
    },
}
```

Matching with a Similarity Encoder

Using a similarity encoder

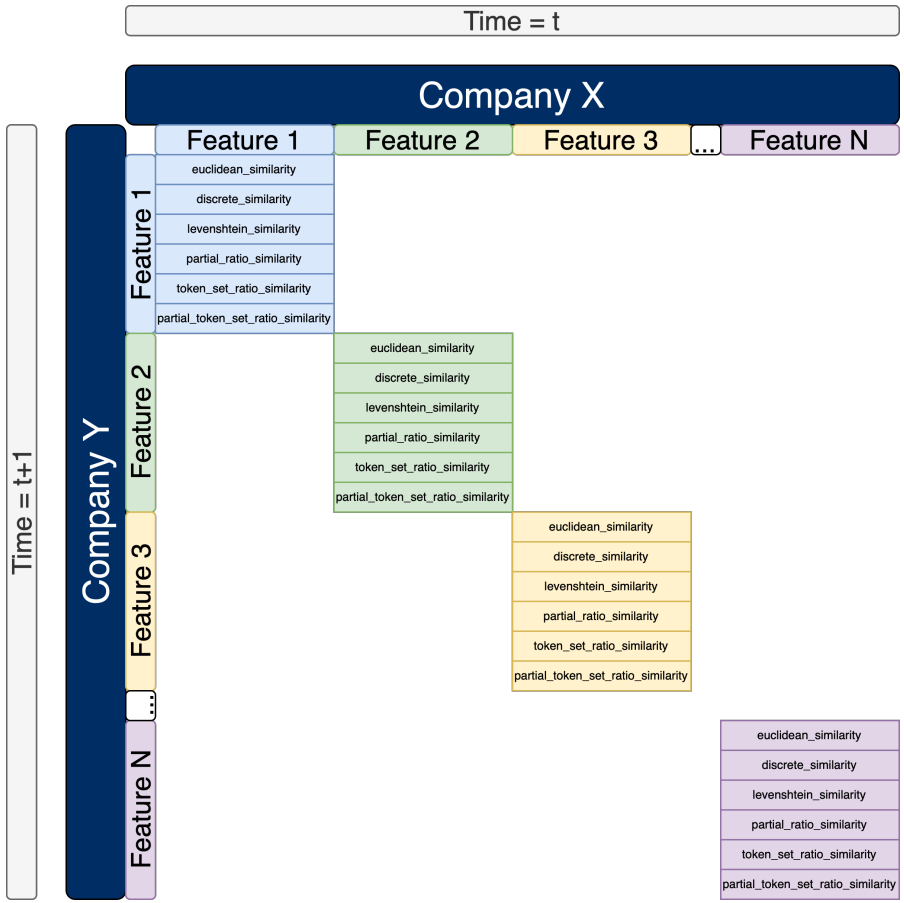
1. Pick a pair (f_l, f_r) of *Left* and *Right* firms.
2. Instruct how the features of f_l and f_r are associated.

```
similarity_map = {
    ...
    "address<->address1": {
        "partial": partial_rati
    },
    "address<->address2": {
        "partial": partial_rati
    },
    ...
}
```

Matching with a Similarity Encoder

Using a similarity encoder

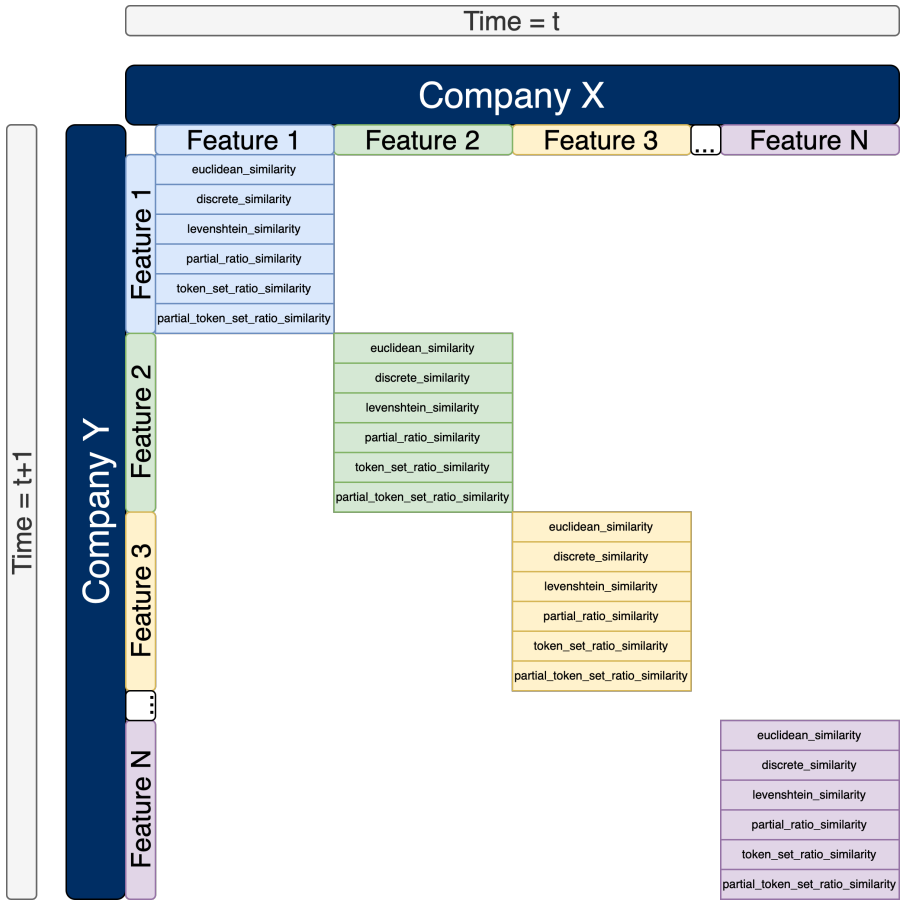
1. Pick a pair (f_l, f_r) of *Left* and *Right* firms.
2. Instruct how the features of f_l and f_r are associated.
3. Calculate various similarities for each feature on the fly.



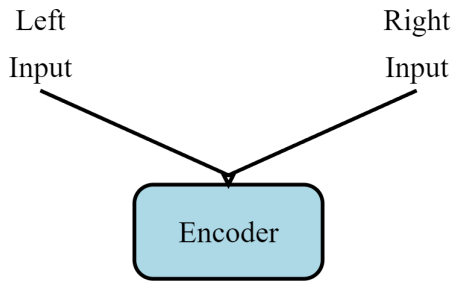
Matching with a Similarity Encoder

Using a similarity encoder

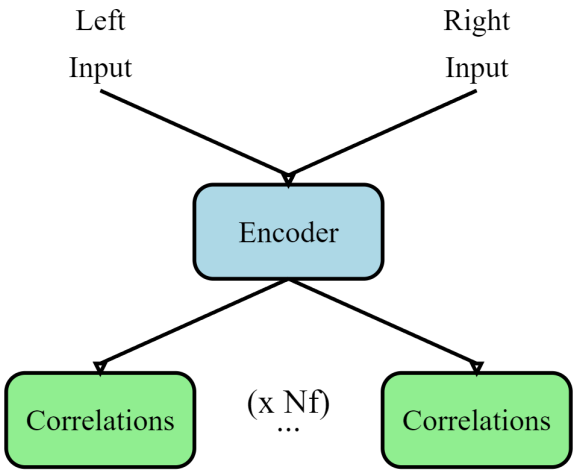
1. Pick a pair (f_l, f_r) of *Left* and *Right* firms.
2. Instruct how the features of f_l and f_r are associated.
3. Calculate various similarities for each feature on the fly.
4. Train the binary matching model using these similarities.



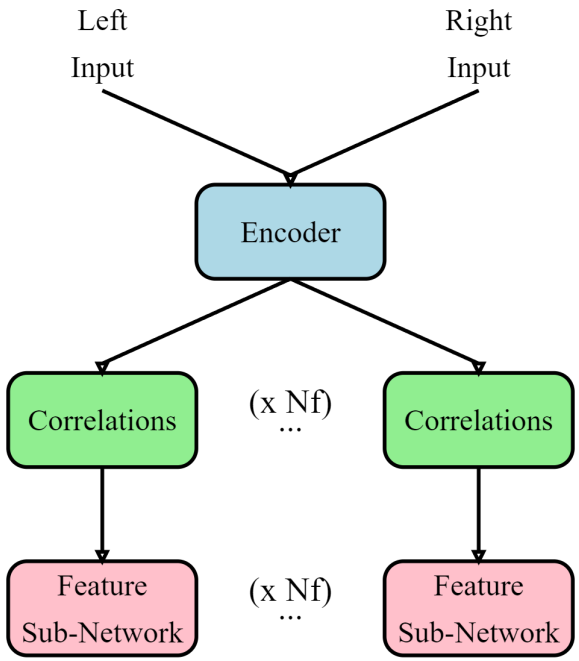
Network Architecture



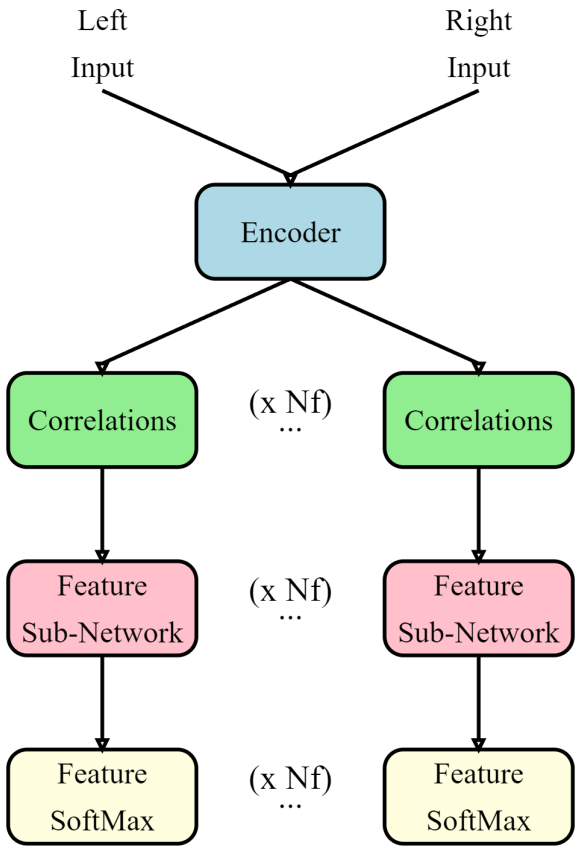
Network Architecture



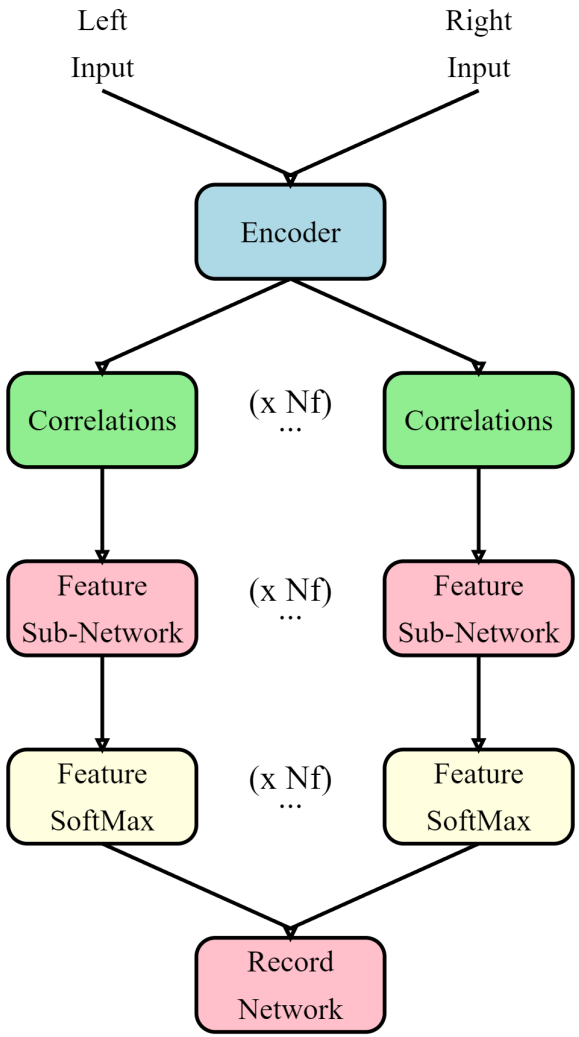
Network Architecture



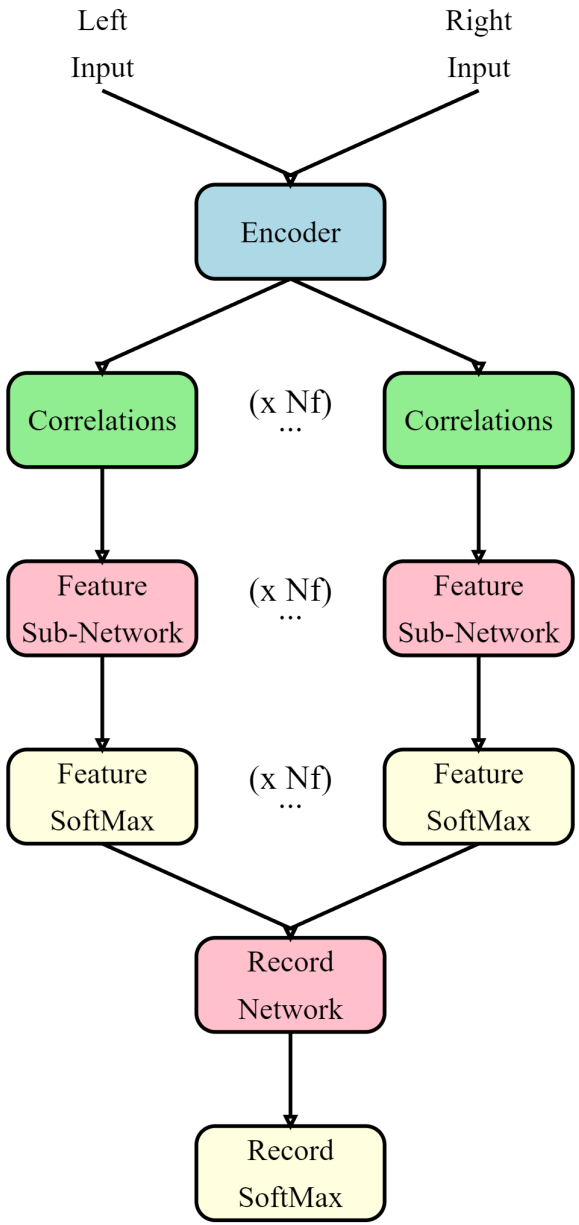
Network Architecture



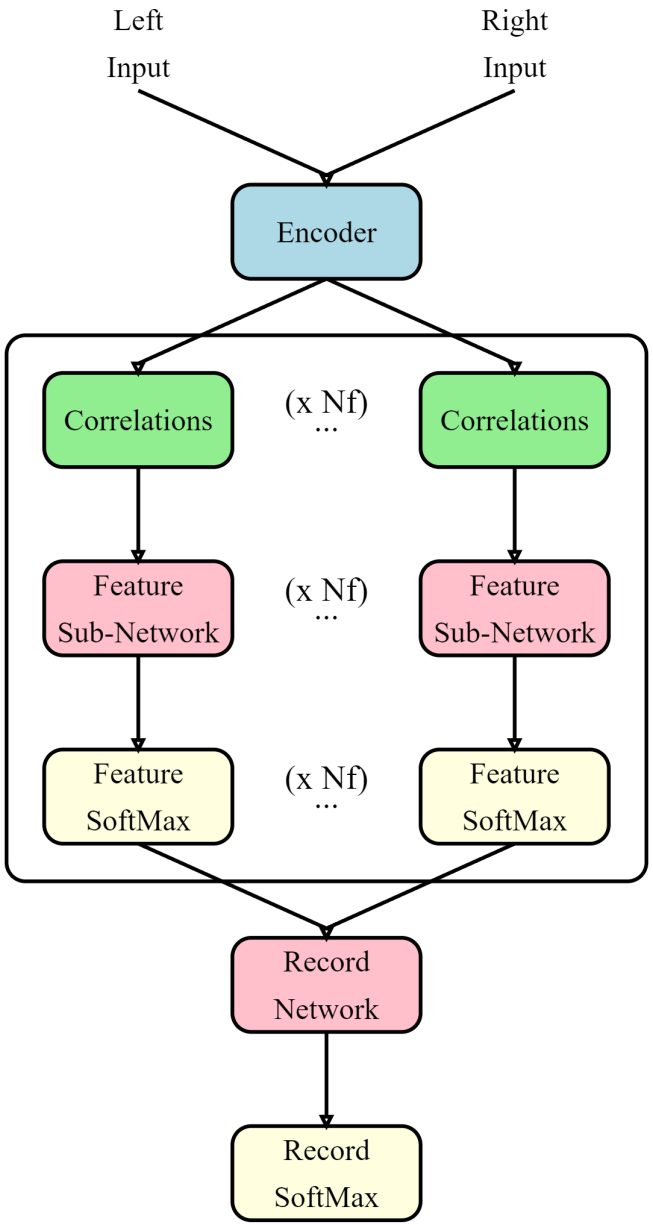
Network Architecture



Network Architecture



Network Architecture



Usage (Pilot)

```
model = match.MatchingModel(similarity_map)
```

Usage (Pilot)

```
model = match.MatchingModel(similarity_map)

model.compile(
    loss="binary_crossentropy",
    optimizer=tensorflow.keras.optimizers.Adam(learnin
    metrics=evaluation_metrics
)
```

Usage (Pilot)

```
model = match.MatchingModel(similarity_map)

model.compile(
    loss="binary_crossentropy",
    optimizer=tensorflow.keras.optimizers.Adam(learnin
    metrics=evaluation_metrics
)

train_left, train_right, train_matches = load_train_da
model.fit(train_left, train_right, train_matches, epoc
```

Usage (Pilot)

```
model = match.MatchingModel(similarity_map)

model.compile(
    loss="binary_crossentropy",
    optimizer=tensorflow.keras.optimizers.Adam(learnin
    metrics=evaluation_metrics
)

train_left, train_right, train_matches = load_train_da
model.fit(train_left, train_right, train_matches, epoc

train_evaluation = model.evaluate(train_left, train_rig
```

Usage (Pilot)

```
model = match.MatchingModel(similarity_map)

model.compile(
    loss="binary_crossentropy",
    optimizer=tensorflow.keras.optimizers.Adam(learnin
    metrics=evaluation_metrics
)

train_left, train_right, train_matches = load_train_da
model.fit(train_left, train_right, train_matches, epoc

train_evaluation = model.evaluate(train_left, train_rig

predictions = model.predict(train_left, train_right)
suggestions = model.suggest(train_left, train_right, 3
```

Bibliography

Gram, D., Karapanagiotis, P., Liebald, M. and Walz, U. (2022). Design and implementation of a historical german firm-level financial database. *J. Data and Information Quality*, **14** (3).