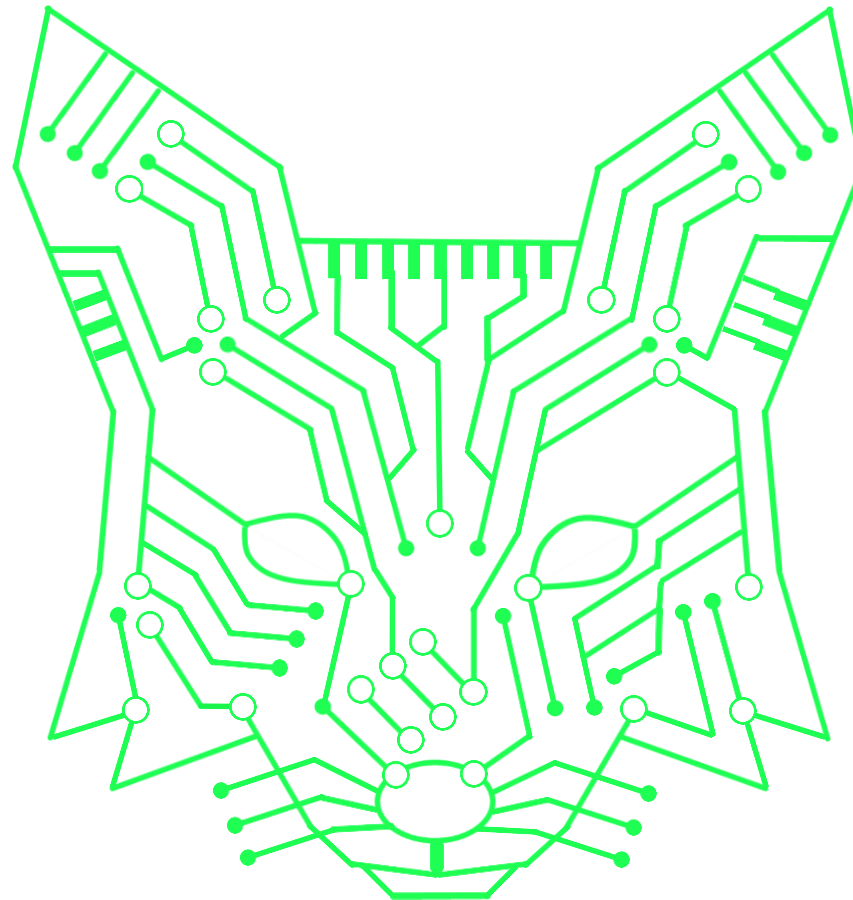
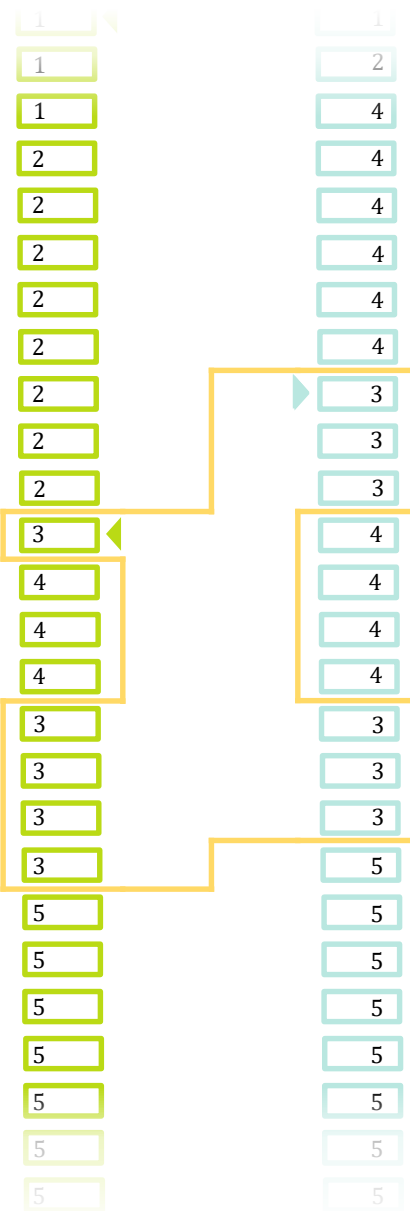


SearchEngine: a Holistic Approach to Matching

Thorsten Doherr | Leibniz Centre for European Economic Research, Germany | <https://github.com/ThorstenDoherr/searchengine>



Blocking



Pairwise Comparisons

21 million IAB records
24 million creditreform records
≈ 500*10¹² comparisons

Blocking

Segmentation of the solution space along arbitrary exclusion restrictions

For example:

- Address blocking (postal code, city)
- Spatial blocking after geocoding
- Overlapping Canopies, i.e. based on first 4 characters of any word in the firm name
- ...
- Combination of methods
- Overlapping blocks to avoid false negatives

Comparisons

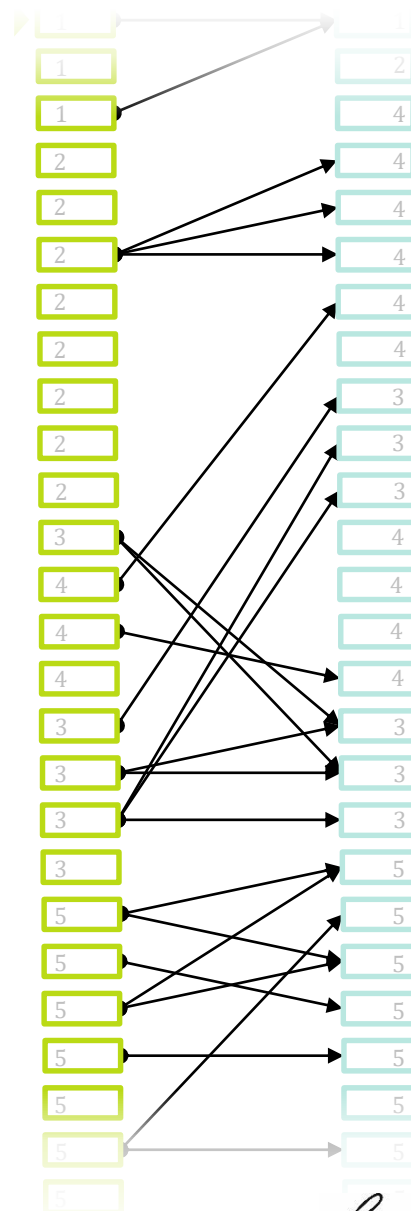
Scoring based on...

- string comparison functions
- word or token frequencies
- spatial distances
- ...

Multiple scores or vectors require statistical methods like ML to handle curse of dimensionality

vs.

Searching



Index Based Search

Complexity increment close to log₂

Search Heuristic

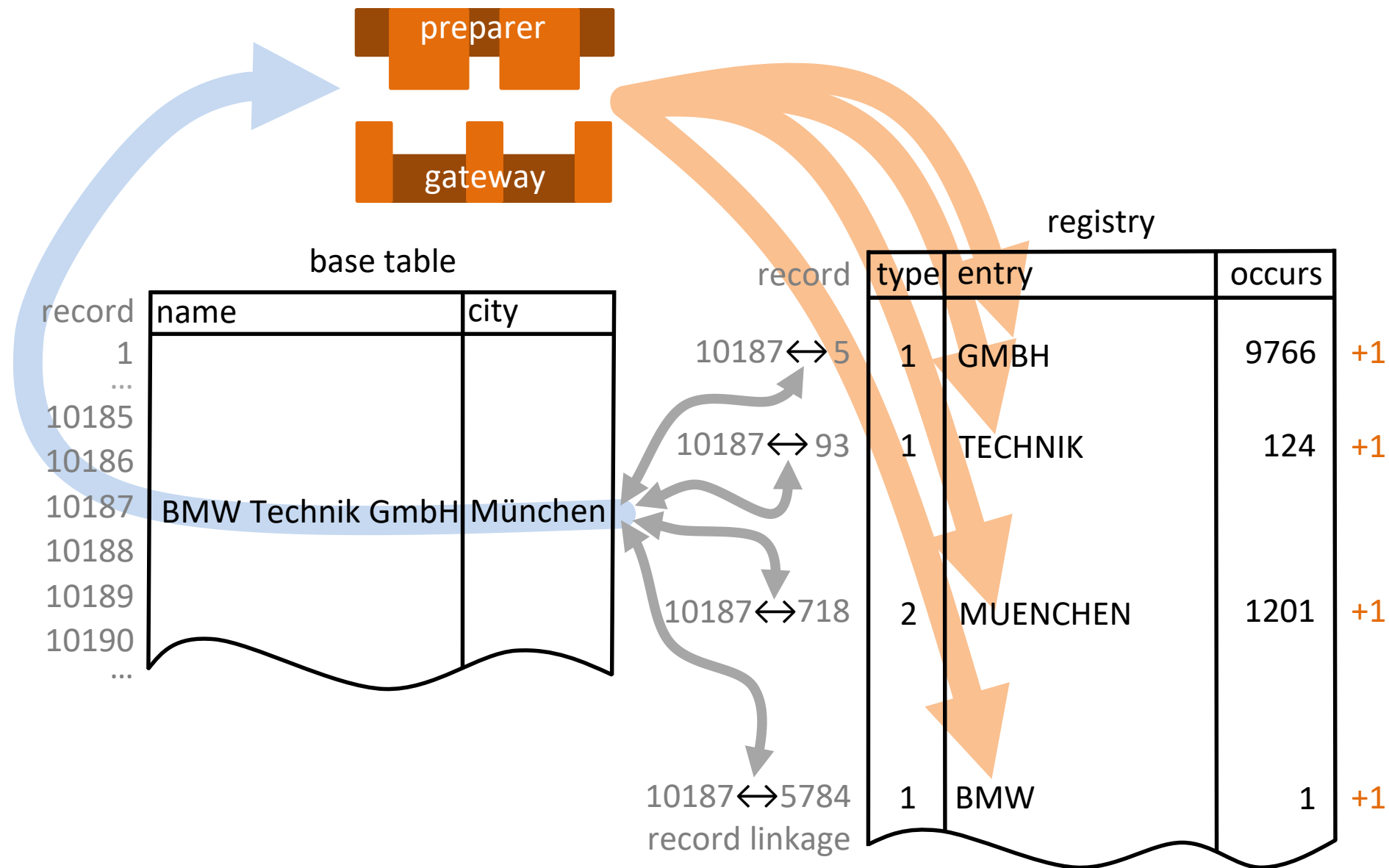
Word frequency based heuristic identifies candidates

- Base table defines heuristic meta data and provides candidates
- Search table provides the search terms
- Individual blocking is not bound to exclusion restrictions
- Search strategy requires several runs to minimize false negatives

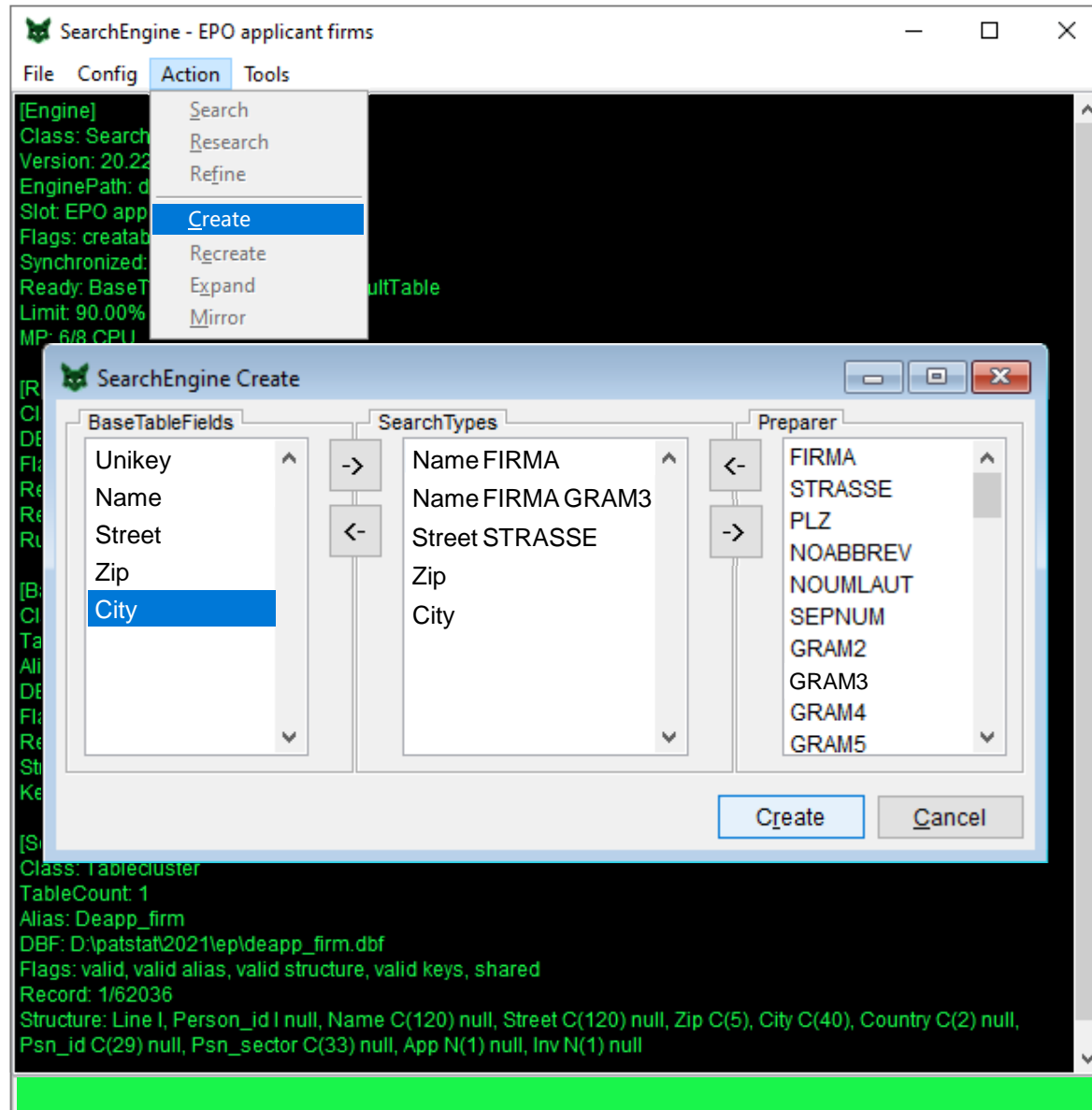
Holistic Approach

- Similarity between search record and candidates is already established
- Only filtering of false positives is required
- Meta data of the search heuristic is repurposed to create variation
- Training sample composition and ML integration is trivial

Components of the SearchEngine



Preparer Gateway



- Preparers are normalization/tokenization directives. In combination with search fields, they define search types. By default, all fields will be normalized: upper case, only alpha-numeric characters, word tokens
- FIRMA, STRASSE, PLZ are custom preparers defined in the "searchengine_firma_strasse.xml" file. They handle the idiosyncrasies of German firm names and addresses (legal forms, concatenated street types) and are not required for other languages. Languages allowing the (arbitrary) concatenation of words pose a challenge for word based algorithms.
- GRAMn preparers implement a computational linguistics method by creating overlapping tokens of size n, i.e. THORSTEN → THO HOR ORS RST STE TEN. Other linguistic preparers implement SOUNDEX, METAPHONE and COLONE but only GRAM can handle erroneously concatenated/split words. These preparers destroy information for the sake of increased robustness against misspellings, therefore they will also be referred as destructive preparers → increased risk of false positives

Relative Identification Potential

| 1 name | 2 street | 3 zip | 4 city |
|--------------------------------|--------------------|-------|----------|
| BMW FORSCHUNG UND TECHNIK GMBH | HANAUER STRASSE 46 | 80992 | MUENCHEN |

| type | entry | occurs | $IP = \frac{1}{occurs}$ $share = \frac{IP}{\sum IP}$ $rIP = share * weight$ | | | | | |
|------|-----------|---------|---|-----------|---|----------|---|---------|
| ... | ... | ... | | | | | | |
| 1 | BMW | 552 | → | 0.0018116 | → | 76.863% | → | 53.804% |
| 1 | FORSCHUNG | 1980 | → | 0.0005051 | → | 21.428% | → | 15.000% |
| 1 | TECHNIK | 25552 | → | 0.0000391 | → | 1.660% | → | 1.162% |
| 1 | UND | 1190073 | → | 0.0000008 | → | 0.036% | → | 0.025% |
| 1 | GMBH | 3353864 | → | 0.0000003 | → | 0.013% | → | 0.009% |
| ... | ... | ... | Σ | 0.0023569 | Σ | 100.00% | Σ | 70.000% |
| 2 | HANAUER | 6851 | → | 0.0001460 | → | 90.511% | → | 9.051% |
| 2 | 46 | 65931 | → | 0.0000152 | → | 9.494% | → | 0.941% |
| 2 | STRASSE | 7410645 | → | 0.0000001 | → | 0.084% | → | 0.008% |
| ... | ... | ... | Σ | 0.0001613 | Σ | 100.000% | Σ | 10% |
| 3 | 80992 | 3905 | → | 0.0002561 | → | 100.000% | → | 10% |
| ... | ... | ... | Σ | 0.0002561 | Σ | 100.000% | Σ | 10% |
| 4 | MUENCHEN | 316874 | → | 0.0000032 | → | 100.000% | → | 10% |
| ... | ... | ... | Σ | 0.0000032 | Σ | 100.000% | Σ | 10% |

Relative Identity

| 1 name | | | | | 2 street | | | 3 zip | 4 city | $\sum r_{IP}$ |
|--------------------------------|-----------|-----|---------|------|----------------------|---------|-----|-------|----------|---------------|
| BMW | FORSCHUNG | UND | TECHNIK | GMBH | HANAUER | STRASSE | 46 | 80992 | MUENCHEN | |
| 53.8 | 15.0 | 0.0 | 1.2 | 0.0 | 9.1 | 0.0 | 0.9 | 10 | 10 | 100.00% |
| BMW FORSCHUNG UND TECHNIK GMBH | | | | | HANAUER STRASSE 46 | | | 80992 | MUENCHEN | 100.00% |
| BMW FORSCHUNG U TECHNIK GMBH | | | | | HANAUER STRASSE 46 | | | 80992 | MUENCHEN | 99.98% |
| BMW TECHNIK UND SERVICE GMBH | | | | | HANAUER STRASSE 48 | | | 80992 | MUENCHEN | 84.06% |
| BMW STIFTUNG HERBERT QUANDT | | | | | HANAUER STRASSE 46 | | | 80992 | MUENCHEN | 83.80% |
| BMW MAENNERCHOR MUENCHEN EV | | | | | DACHAUER STRASSE 371 | | | 80992 | MUENCHEN | 73.81% |

Threshold: 70%

Not Commutative ($a \rightarrow b \neq b \rightarrow a$)

| 1 name | | | | 2 street | | | 3 zip | 4 city | $\sum r_{IP}$ |
|-----------------------------|-------------|----------|-----|----------------------|---------|-----|-------|----------|---------------|
| BMW | MAENNERCHOR | MUENCHEN | EV | DACHAUER | STRASSE | 371 | 80992 | MUENCHEN | |
| 18.0 | 51.2 | 0.8 | 0.0 | 1.2 | 0.0 | 8.8 | 10 | 10 | 100.00% |
| BMW MAENNERCHOR MUENCHEN EV | | | | DACHAUER STRASSE 371 | | | 80992 | MUENCHEN | 100.00% |
| MANNERCHOR RIESENFELD EV | | | | ABBACH STRASSE 27 A | | | 80992 | MUENCHEN | 71.23% |

Threshold: 70%

Candidate Retrieval

| | | | | | | | | | |
|---------|-------------|---------|------------|-----------|-----------|---------|---------|---------|-----------|
| 1 BMW | 1 FORSCHUNG | 3 80992 | 4 MUENCHEN | 2 HANAUER | 1 TECHNIK | 2 46 | 1 UND | 1 GMBH | 2 STRASSE |
| 53.8 | 15.0 | 10.0 | 10.0 | 9.1 | 1.2 | 0.9 | 0.0 | 0.0 | 0.0 |
| 2150179 | 2172419 | 2179375 | 2188161 | 2182779 | 2186398 | 2187556 | 2188225 | 2188227 | 2188228 |

sort order

registry

| type | entry | occurs |
|------|-----------|--------|
| ... | ... | ... |
| 1 | BMW | 552 |
| 3 | LENIN | 552 |
| ... | ... | ... |
| 1 | FORSCHUNG | 1980 |
| ... | ... | ... |
| 3 | 80992 | 3950 |
| ... | ... | ... |

N_{registry}

regindex

| record | index |
|---------|----------|
| ... | ... |
| 2150179 | 4636758 |
| 2156731 | 4637310 |
| ... | ... |
| 2172419 | 7854367 |
| ... | 7856347 |
| ... | ... |
| 2179375 | 11241655 |
| ... | 11245605 |
| ... | ... |

$N_{\text{registry}} + 1$

base

| record | target |
|---------|---------|
| ... | ... |
| 4636758 | 10187 |
| ... | 15633 |
| ... | 27996 |
| ... | ... |
| ... | 7267831 |
| 4637310 | ... |
| ... | ... |
| ... | ... |
| ... | ... |

$\sum \text{occurs}$

| BUFFER | AGGREGATION | Σ | +21.2 |
|--------|-----------------------|----------|-------|
| | BMW, FORSCHUNG, 80992 | 78.8 | 100.0 |
| | BMW, FORSCHUNG | 68.8 | 90.0 |
| | BMW, 80992 | 63.8 | 85.0 |
| | BMW | 53.8 | 75.0 |
| | FORSCHUNG, 80992 | 25.0 | 46.2 |
| | FORSCHUNG | 15.0 | 36.2 |
| | 80992 | 10.0 | 31.2 |

Efficiency Puzzle

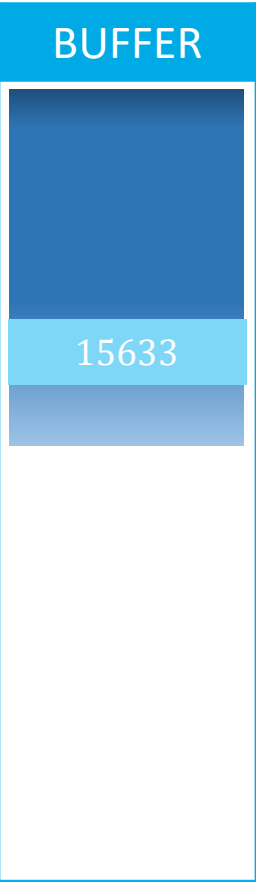
The retrieval also stops early when the interim total of the collated RIDs exceeds 100—threshold, because at least one of those words is necessary to push the Identity over the threshold. In this example retrieval actually stops already after “BMW”.

fixed size

Identity Completion

| | | | | | | | | | |
|--------|-------------|---------|-----------|-----------|---------|------------|---------|---------|-----------|
| 1 BMW | 1 FORSCHUNG | 3 80992 | 2 HANAUER | 1 TECHNIK | 2 46 | 4 MUENCHEN | 1 UND | 1 GMBH | 2 STRASSE |
| 53.8 | 15.0 | 10.0 | 9.1 | 1.2 | 0.9 | 10.0 | 0.0 | 0.0 | 0.0 |
| 250179 | 2172419 | 2179375 | 2182779 | 2186398 | 2187556 | 2188161 | 2188225 | 2188227 | 2188228 |

sort order



| baseindex | |
|----------------|--------|
| record | index |
| | ... |
| 15633 | 124935 |
| 15634 | 124944 |
| | ... |
| $N_{base} + 1$ | |

| reg | | | |
|---------------|---------|----------|------|
| record | target | | |
| | ... | | |
| 124935 | 2150179 | BMW | 53.8 |
| | 2179375 | 80992 | 10.0 |
| | 2182779 | HANAUER | 9.1 |
| | 2186398 | TECHNIK | 1.2 |
| | 2187556 | 46 | 0.9 |
| | 2188021 | SERVICE | |
| | 2188161 | MUENCHEN | 10.0 |
| | 2188225 | UND | 0.0 |
| | 2188227 | GMBH | 0.0 |
| | 2188228 | STRASSE | 0.0 |
| 124944 | ... | | 83.8 |
| $\sum occurs$ | | | |

Depth
The default value for the BUFFER is 262,144. It can be increased to accommodate more homogenous databases or for specific search purposes, i.e. selection by legal form. The higher the Depth of the BUFFER the higher the amount of bycatch to be rejected at the first stage.

← Oblivious to surplus words of candidates

Weak Search Terms: Berlin GmbH, Berlin

| identity | name | street | zip | city |
|----------|--|------------------------------|-------|--------|
| 80.00 | CPB Cassetten-Produktion GmbH Berlin | Ballinstr. 16-18 | 12359 | Berlin |
| 80.00 | SCHERDEL Berlin GmbH & Co. KG | Kanalstr. 66-74 | 12357 | Berlin |
| 80.00 | Münchhoff GmbH Berlin | Nollendorfplatz 6 | 10777 | Berlin |
| 80.00 | Elektron Berlin GmbH Fertigung elektrotechnischer Spezialartikel | Saatwinkler Damm 60 | 13627 | Berlin |
| 80.00 | KMB Kabel-Maschinen-Berlin GmbH | Innungsstr. 56 | 13509 | Berlin |
| 80.00 | Alois Dallmayr Kaffee Berlin GmbH u. Co. KG | Haberstr. 9-13 | 12057 | Berlin |
| 80.00 | Mampe GmbH Berlin | Grenzallee 22-34 | 12057 | Berlin |
| 80.00 | KBA - Berlin GmbH | Mertensstr. 127-131 | 13587 | Berlin |
| 80.00 | Beiersdorf Manufacturing Berlin GmbH | Franklinstr. 1 | 10587 | Berlin |
| 80.00 | Centro-Boden Berlin KG Wohn- und Gewerbebauten im Centrum GmbH & Co. | Joachimstaler Str. 14 | 10719 | Berlin |
| 80.00 | Amcor Specialty Cartons Berlin GmbH | Haberstr. 5 | 12057 | Berlin |
| 80.00 | Helmholtz-Zentrum Berlin für Materialien und Energie GmbH | Hahn-Meitner-Platz 1 | 14109 | Berlin |
| 80.00 | LAROSÉ Hygiene-Service GmbH Berlin | Grünauer Str. 116-120 | 12557 | Berlin |
| 80.00 | Hifi Elements Berlin GmbH | Hubertusstr. 7 | 12163 | Berlin |
| 80.00 | FB Fernsehdienst in Berlin GmbH | Bismarckstr. 71 | 12157 | Berlin |
| 80.00 | Klosterfrau Berlin GmbH | Motzener Str. 41 | 12277 | Berlin |
| 80.00 | DG Leasing Berlin GmbH | Turmstr. 77 | 10551 | Berlin |
| 80.00 | DG Leasing Berlin GmbH & Co. Miet + Leasing KG | Turmstr. 77 | 10551 | Berlin |
| 80.00 | Henning Berlin GmbH & Co. | Potsdamer Str. 8 | 10785 | Berlin |
| 80.00 | "Pia Rucci" Sportswear Bekleidungsvertriebs GmbH, Berlin | Lützowufer 12-13 | 10785 | Berlin |
| 80.00 | Ideal Automotive Berlin GmbH | Zerpenschleuser Ring 22 | 13439 | Berlin |
| 80.00 | Saarberg Handel Berlin GmbH | Quedlinburger Str. 11 | 10589 | Berlin |
| 80.00 | Messe Berlin GmbH | Messedamm 22 | 14055 | Berlin |
| 80.00 | BAO BERLIN International GmbH | Fasanenstr. 85 | 10623 | Berlin |
| 80.00 | Metallbauzaun-Montagen Roden GmbH Berlin | Saatwinkler Damm 25- 26 | 13627 | Berlin |
| 80.00 | tip AUTO-Berlin GmbH | Schulzendorfer Str. 23-24Hof | 13347 | Berlin |
| 80.00 | CVB Albert Carl GmbH Berlin | Oberlandstr. 22-25 | 12099 | Berlin |
| 80.00 | Breuer Service Berlin GmbH Industrie- und Kraftfahrzeugreinigung | Alte Jakobstr. 135 | 10969 | Berlin |
| 80.00 | Tyler Berlin GmbH | Flankenschanze 28 | 13585 | Berlin |
| 80.00 | FRÜH - HERBST Anlagentechnik Berlin GmbH | Lankwitzer Str. 23-25 | 12107 | Berlin |
| 80.00 | PFENNIGs Feinkostfabrik Berlin Albert Pfennig + Sohn GmbH & Co. | Ringbahnstr. 22-30 | 12099 | Berlin |
| 80.00 | Wachschutz Berlin Werner Loesch GmbH & Co. | Odenwaldstr. 26 | 12161 | Berlin |
| 80.00 | Kaiser Kabel GmbH Kabel- und Freileitungswerk Berlin | Gradestr. 100 | 12347 | Berlin |
| 80.00 | KKB Küchenkomplettbau GmbH Berlin | Reuchlinstr. 10-11 | 10553 | Berlin |
| 80.00 | Tetra Pak Berlin GmbH & Co TPB KG | Hennigsdorfer Str. 159 | 13503 | Berlin |
| 80.00 | Häfele Berlin GmbH & Co. KG | Schichauweg 50 | 12307 | Berlin |
| 80.00 | Van Houten Industrie Berlin GmbH | Grenzallee 4- 6 | 12057 | Berlin |
| 80.00 | Möbelkiste GmbH u. Co. Handelsgesellschaft Berlin | Bundesallee 36 | 10717 | Berlin |
| 80.00 | Kamps Berlin Geschäftsführungs GmbH | Bergiusstr. 26-28 | 12057 | Berlin |
| 80.00 | Novetta Berlin KG NB-Nahrungsmittelgesellschaft mbH & Co. | Sonnenallee 221 | 12059 | Berlin |
| 80.00 | Tyler Refrigeration Berlin GmbH & Co. | Urbanstr. 116 | 10967 | Berlin |
| 80.00 | Greve-Chemotechnik Berlin GmbH | Lindenufer 39 | 13597 | Berlin |
| 80.00 | Hilfswerk-Siedlung GmbH Evangelisches Wohnungsunternehmen in Berlin | Irchblick 13 | 14129 | Berlin |
| 80.00 | Werbedienst Berlin GmbH | Ringbahnstr. 16-20 | 12099 | Berlin |
| 80.00 | Berlin Los Angeles Platz Value Added I, GmbH & Co. KG | Los-Angeles-Platz 1 | 10789 | Berlin |
| 80.00 | Kraft Jacobs Suchard Berlin GmbH & Co. KG | Nobelstr. 1- 21 | 12057 | Berlin |
| 80.00 | GmbH & Co. Berlin KG | Rosendammallee 9 | 13407 | Berlin |

25000

Weak Search Terms...

- have few words with high frequencies
- are often affected by missing search fields
- contain redundancy, i.e. city name in firm name

This leads to...

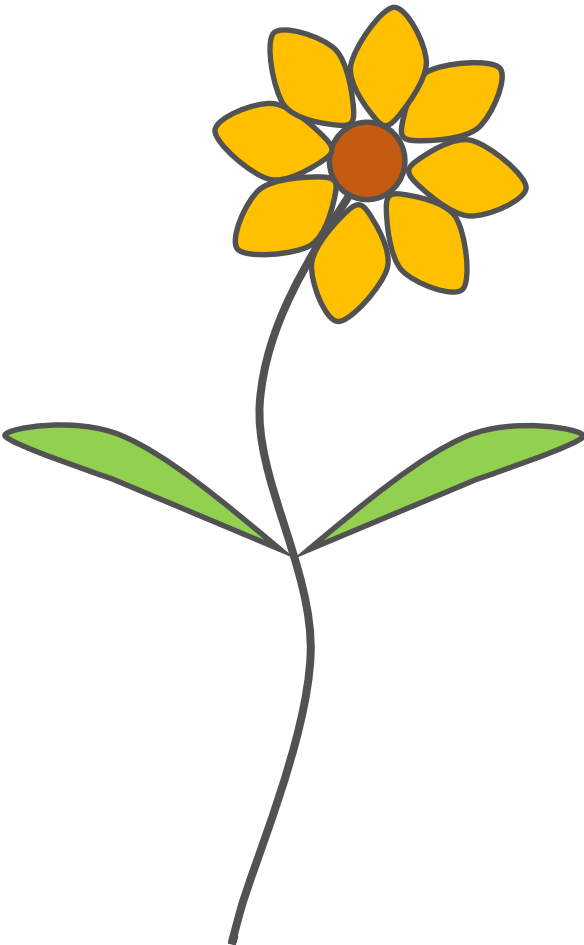
- bloated candidate lists of unrelated false positives
- ambiguous candidates without variance in the identity
- waste of time trying to find true positives

Potential solution

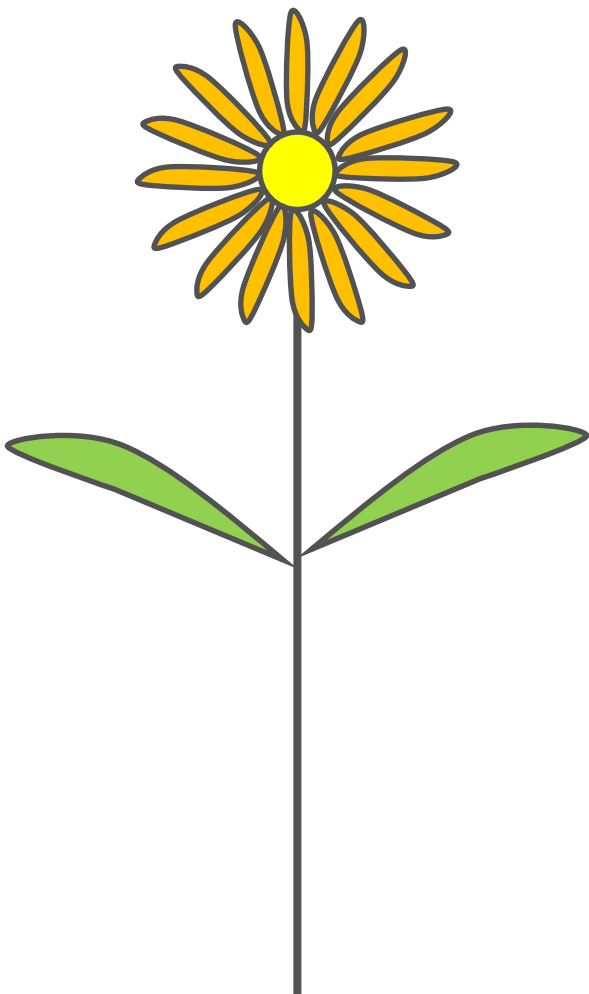
1. Sort candidates by identity in descending order
2. All candidates down to a cutoff point are deemed plausible, i.e. top 10 candidates
3. The identity of the candidate at the cutoff point becomes the new threshold for this search term to prevent arbitrary dismissal of candidates

! No variance in the identity → no cutoff

Searched Flower



Found Flower



$$J(S, F) = \frac{|\{S \cap F\}|}{|\{S \cup F\}|}$$

$$\frac{4}{7 + 3 \text{ feedback}}$$

The numerator 4 is derived from the set of features shared by both flowers: $|\{ \text{petal}, \text{leaf}, \text{petal_yellow}, \text{center_empty} \}|$. The denominator 7 is derived from the set of features in either flower: $|\{ \text{petal}, \text{leaf}, \text{center_empty}, \text{petal_yellow}, \text{center_brown}, \text{petal_empty}, \text{center_yellow} \}|$. The additional 3 in the denominator is labeled 'feedback'.

$$I_J(S, F, f) = I(S, F)(1 - f) + J(S, F)f$$
$$f \in [0,1]$$

10% Feedback Effect on Berlin GmbH, Berlin

| identity | name | street | zip | city |
|----------|--|---------------------------|-------|--------|
| 80.00 | Berlin GmbH | Oudenarder Str. 31 | 13347 | Berlin |
| 80.00 | Berlin | | | Berlin |
| 80.00 | C/O Berlin GmbH & Co. KG | Auguststr. 5a | 10117 | Berlin |
| 80.00 | C/O Berlin GmbH & Co. KG | Oranienburger Str. 35/ 36 | 10117 | Berlin |
| 80.00 | KG Berlin | Kottmeierstr. 1- 3 | 12459 | Berlin |
| 80.00 | Gesellschaft Berlin GmbH | Holzhauser Str. 52-56 | 13509 | Berlin |
| 80.00 | Immobilien Berlin GmbH | Barstr. 29 | 10713 | Berlin |
| 80.00 | Berlin Bau GmbH | Breite Str. | 10178 | Berlin |
| 80.00 | Berlin Bau - GmbH Berlin | | | Berlin |
| 80.00 | Berlin GbR | | | Berlin |
| 78.51 | AG Berlin | Schlüterstr. 36 | 10629 | Berlin |
| 78.37 | Thomas Berlin | An den Bänken 3 | 12589 | Berlin |
| 78.15 | DR Berlin UG | Saarbrücker Str. 36 | 10405 | Berlin |
| 78.09 | Baugesellschaft Berlin GmbH | Pfalzbürger Str. 15 | 10719 | Berlin |
| 77.97 | SE Berlin Beteiligungs GmbH | Pichelswerderstr. 3-5 | 13597 | Berlin |
| 77.89 | Bau und Immobilien Berlin GmbH | Lindenstr. 4 | 12621 | Berlin |
| 77.70 | Walter Berlin GmbH | Bürgerstr. 25- 27 | 12347 | Berlin |
| 77.70 | Walter Berlin GmbH | Brauhofstr. 4a | 10587 | Berlin |
| 77.65 | Immobilien Service Berlin | Lampesteig 10 | 13409 | Berlin |
| 77.51 | Martin Immobilien Berlin GmbH | Hönower Str. 1 | 10318 | Berlin |
| 77.41 | Jürgen Berlin Handelsgesellschaft mbH | Sagritzer Weg 12 | 13435 | Berlin |
| 77.29 | In-Berlin e.V. | Kaiser-Friedrich-Str. 88 | 10585 | Berlin |
| 77.09 | AM Berlin GmbH | Prenzlauer Allee 53 | 10405 | Berlin |
| 77.00 | Joachim Berlin | Randowstr. 12 | 13057 | Berlin |
| 76.99 | KG Beteiligungs-mbH Berlin | | | Berlin |
| 76.85 | Management Holding GmbH & Co. KG Berlin | | | Berlin |
| 76.82 | Beteiligungs-Management-Gesellschaft Berlin GmbH | | | Berlin |
| 76.76 | Marketing Service Berlin GmbH | Taubenstr. 19-23 | 10117 | Berlin |
| 76.68 | Walter Immobilien Berlin | Hochkönigweg 44 | 12349 | Berlin |
| 76.57 | Trockenbau Berlin | Küstriner Str. 7- 8 | 13055 | Berlin |
| 76.56 | Grundstücksgesellschaft mbH Berlin | Kaiser-Friedrich-Str. 41 | 10627 | Berlin |
| 76.54 | Ulrich Bau Berlin GmbH | Mühlenstr. 8a | 14167 | Berlin |
| 76.51 | Michael Berlin Ingenieurbüro | Rochowstr. 1b | 10245 | Berlin |
| 76.49 | Kurt Berlin GmbH | Baumschulenstr. 72 | 12437 | Berlin |
| 76.34 | Immobilien Haus Berlin | Kleiststr. 3-6 | 10787 | Berlin |
| 76.14 | Thomas-Haus Berlin e.V. | Peter-Lenne-Str. 4 | 14195 | Berlin |
| 76.07 | Metallbau Berlin GmbH | Müggelseedamm 128 | 12587 | Berlin |
| 76.04 | BERLIN VERLAG GmbH & Co.KG | Greifswalder Str. 207 | 10405 | Berlin |
| 76.04 | BERLIN VERLAG GmbH & Co.KG | Greifswalder Str. 207 | 10405 | Berlin |
| 76.01 | Claudia Berlin | Paul-Dessau-Str. 9 | 12679 | Berlin |
| 75.95 | Bäckerei Berlin | Oranienstr. 67 | 10969 | Berlin |
| 75.93 | Schneider Bau Berlin | Bismarckstr. 39 | 10627 | Berlin |
| 75.91 | Auto-Service Berlin GmbH | Katzlerstr. 15 | 10829 | Berlin |
| 75.86 | Büro-Service-GmbH Berlin | Malschweg 12 | 13593 | Berlin |
| 75.85 | Immobilien Vertrieb Berlin e.V. | Kurfürstendamm 208 | 10719 | Berlin |
| 75.85 | Metallbau Service Berlin GmbH | Heinersdorfer Str. 4 | 13129 | Berlin |
| | Management-Gesellschaft Berlin mbH | | | Berlin |

Containment of weak search terms

Cutoff

Defines the upper limit for a reasonable number of candidates, i.e. 10.

Activation

If the candidate list meets this threshold, Feedback will be applied. Usually, it should equal the Cutoff.

Feedback

Defines the magnitude of the feedback effect as discount on the Identity.

When Activation > 0 and Cutoff > 0:

→ Temporary Feedback effect to create variation for Cutoff

- Usually, 10% feedback suffices
- Effect will be undone before Threshold validation

Otherwise:

→ Permanent Feedback effect affecting Identity vs. Threshold

The interaction of these three settings prevents unnecessary, time consuming feedback calculations. Containment will keep candidates with the least amount of relevant noise.

3-Gram Search for “Blaupause”: BLA LAU AUP UPA AUS USE

| identity | name |
|----------|--|
| 100.00 | Blaupause Bootsbau GmbH |
| 100.00 | Blaupause KfK Verwaltungs GmbH |
| 100.00 | Projekt Blaupause e.V. |
| 100.00 | Alexander Schilder 'Blaupause' |
| 100.00 | blaupause e.V. |
| 100.00 | BLAUPAUSE e.V. - mobile und flexible Hilfen für Menschen mit einer Alkoholerkrankung |
| 100.00 | Blaupause - Initiative für mentale Gesundheit im Gesundheitswesen e.V. |
| 100.00 | Martin Alexander Rieger Blaupausen Medien Multimediaagentur |
| 100.00 | Blaupause UG Die Agentur für mehr |
| 100.00 | Blaupauser Zeichenbüro e.U. |
| 100.00 | Blaupauser Projektentwicklung GmbH |
| 100.00 | Blaupause Interior Design e.U. |
| 100.00 | Anja Thomä und Lena Dreesmann GbR "Blaupause" |
| 100.00 | Anja Thomä Blaupause papeterie |
| 100.00 | Barnimer Alternative e.V. Jugendclub Blaupause |
| 100.00 | Jens Naumann Jan Welsch Blaupause GbR |
| 100.00 | Agentur Blaupause 36 UG |
| 91.10 | Blaupark Living Ulm K20 Projekt GmbH |
| 91.10 | Ute Nißle-Klammt Blaupark -Apotheke |
| 91.10 | Blaupapier Immobilienverwertungsges.m .b.H. |
| 91.10 | Blaupapier GmbH |
| 91.10 | Blaupark Living Ulm K20 Projekt GmbH |
| 91.03 | Heike Hartung Plaupause |
| 91.03 | Plaupause - Bistro & Spätshop UG |
| 89.68 | Simon Huber Filmproduktion graupause |
| 86.68 | Seniorenzentrum Lopaupark GmbH |
| 86.68 | paupau GmbH |
| 86.68 | paupau Deli UG |
| 86.68 | Paulsen Baupartner GmbH |
| 86.68 | Karl Paul Kaupa Bettengeschäft |
| 86.68 | fraupaul e.U. |
| 83.80 | Bausen & Markwart Ausbaupartner GmbH |
| 83.80 | Baupartner BAUSTOFFSERVICE Schönhausen GmbH |
| 83.80 | Krause Baupartner |

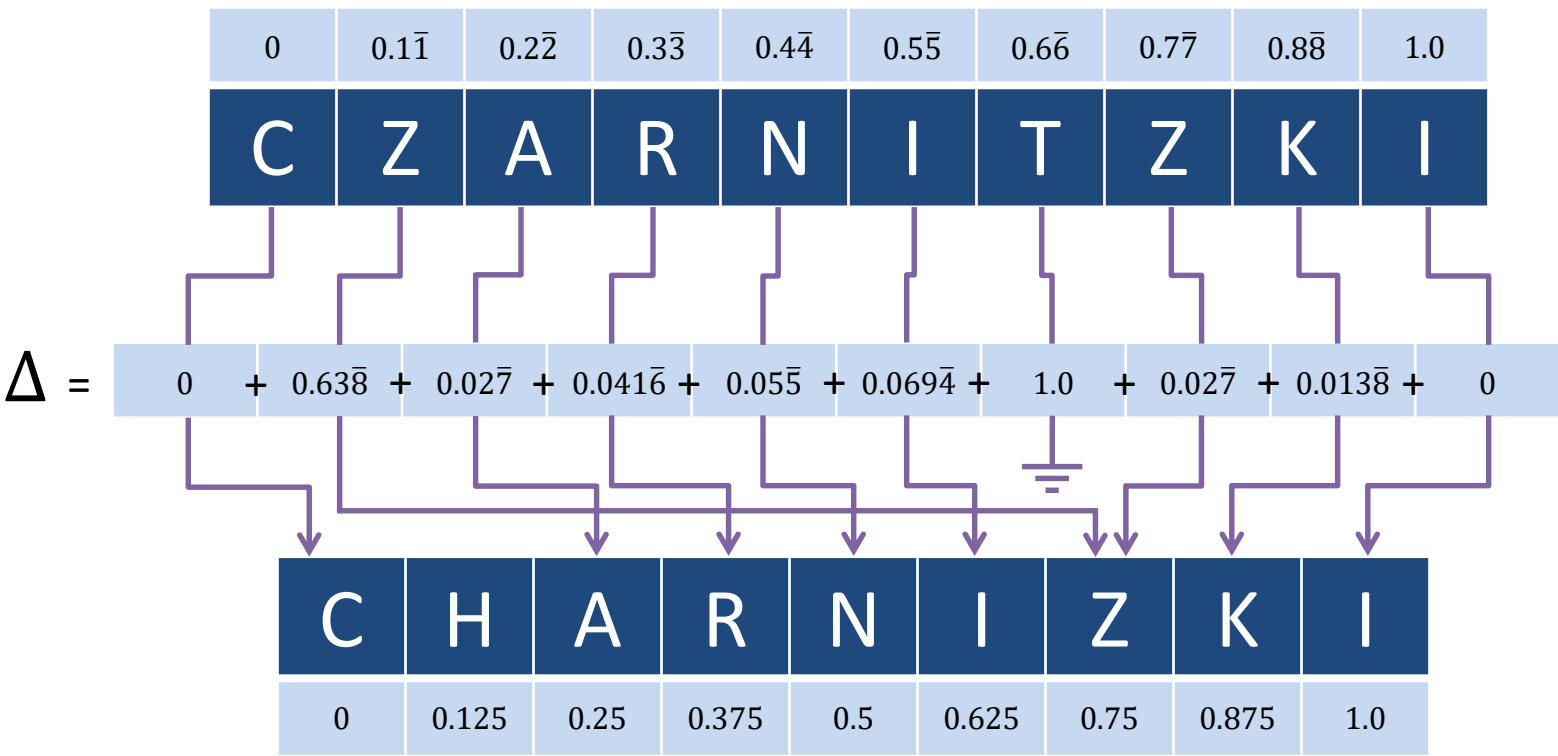
Words responsible for retrieval:

blaupause
blaupausen
blaupauser
blaupark
blaupapier
plaupause
graupause
lopaupark
paupau
paulsen baupartner
paul kaupa
fraupaul
bausen ausbaupartner
baupartner baustoffservice schönhausen

Linguistic preparers destroy information for the sake of recalling misspelled terms at the expense of precision. The genuinely misspelled entries are drowned in a deluge of false positives. Usually, linguistic methods are applied in interactive environments, where results are eye-balled by humans discriminating between real misspellings and mechanical clutter. We have to replicate this visual screening via a flexible string distance function, which is independent of word positioning.

1580

Least Relative Character Position Delta (LRCPD)



$$lrcpd(word1, word2) = 1 - \frac{\Delta(word1, word2)}{len(word1)} = 1 - \frac{1.875}{10} = 0.8125$$

Candidate Refinement (Visual Screening)

search = any search field affected by a destructive preparer

found = corresponding entry of the Candidate

run basic preparer over search and found

delta = 0

go through all words of search → word1

max = 0

go through all words of found → word2

max = max(max, lrcpd(word1, word2))

if max == 1 exit loop

delta += max

delta = delta/wordcount(search)

remove all blanks from search and found

return max(delta, lrcpd(search, found))

- Refinement ignores word-positioning, is commutative and returns a percentage → easy integration
- Refinement is integrated into the Identity by replacing the components of search types affected by destructive Preparer

3-Gram Search for “Blaupause” after LRCPD-Refinement

| identity | name |
|----------|--|
| 100.00 | Blaupause Bootsbau GmbH |
| 100.00 | Blaupause KfK GmbH & Co. KG |
| 100.00 | Projekt Blaupause e.V. |
| 100.00 | Alexander Schilder 'Blaupause' |
| 100.00 | blaupause e.V. |
| 100.00 | BLAUPAUSE e.V. - mobile und flexible Hilfen für Menschen mit einer Alkoholerkrankung |
| 100.00 | Blaupause - Initiative für mentale Gesundheit im Gesundheitswesen e.V. |
| 100.00 | Blaupause UG Die Agentur für mehr |
| 100.00 | Blaupause Interior Design e.U. |
| 100.00 | Anja Thomä und Lena Dreesmann GbR "Blaupause" |
| 100.00 | Anja Thomä Blaupause papeterie |
| 100.00 | Barnimer Alternative e.V. Jugendclub Blaupause |
| 100.00 | Jens Naumann Jan Welsch Blaupause GbR |
| 100.00 | Agentur Blaupause 36 UG |
| 88.89 | Heike Hartung P laupause |
| 88.89 | P laupause - Bistro & Spätshop UG |
| 85.00 | Martin Alexander Rieger Blaupause n Medien Multimediaagentur |
| 85.00 | Blaupause r Zeichenbüro e.U. |
| 85.00 | Blaupause r Projektentwicklung GmbH |

Refinement

and integration are subsequent processes following retrieval. They are optional. A second Threshold can be specified to filter candidates thereafter. Even the direction of the Refinement can be altered:

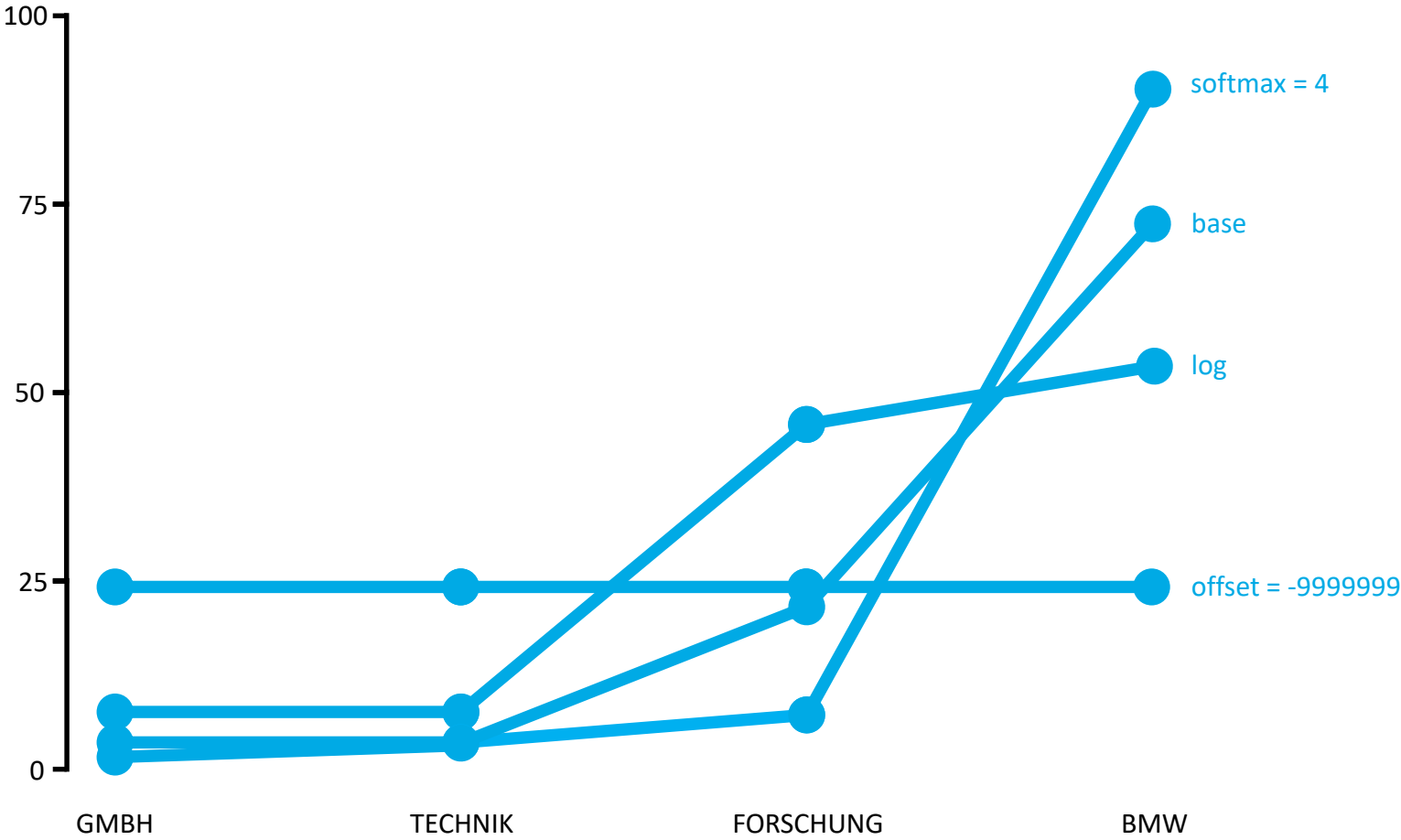
- **Compare Searched with Found** is the default direction and mimics the general SearchEngine behavior
- **Dynamic compare** compares in both directions and uses the lowest result → suitable for person names
- **Compare Found with Searched** reverses the default direction → more noise in the base table (rarely used)
- **No automatic refine/research on destructive preparer** skips the whole Refinement part (not advised but required for educational purposes)

Advice

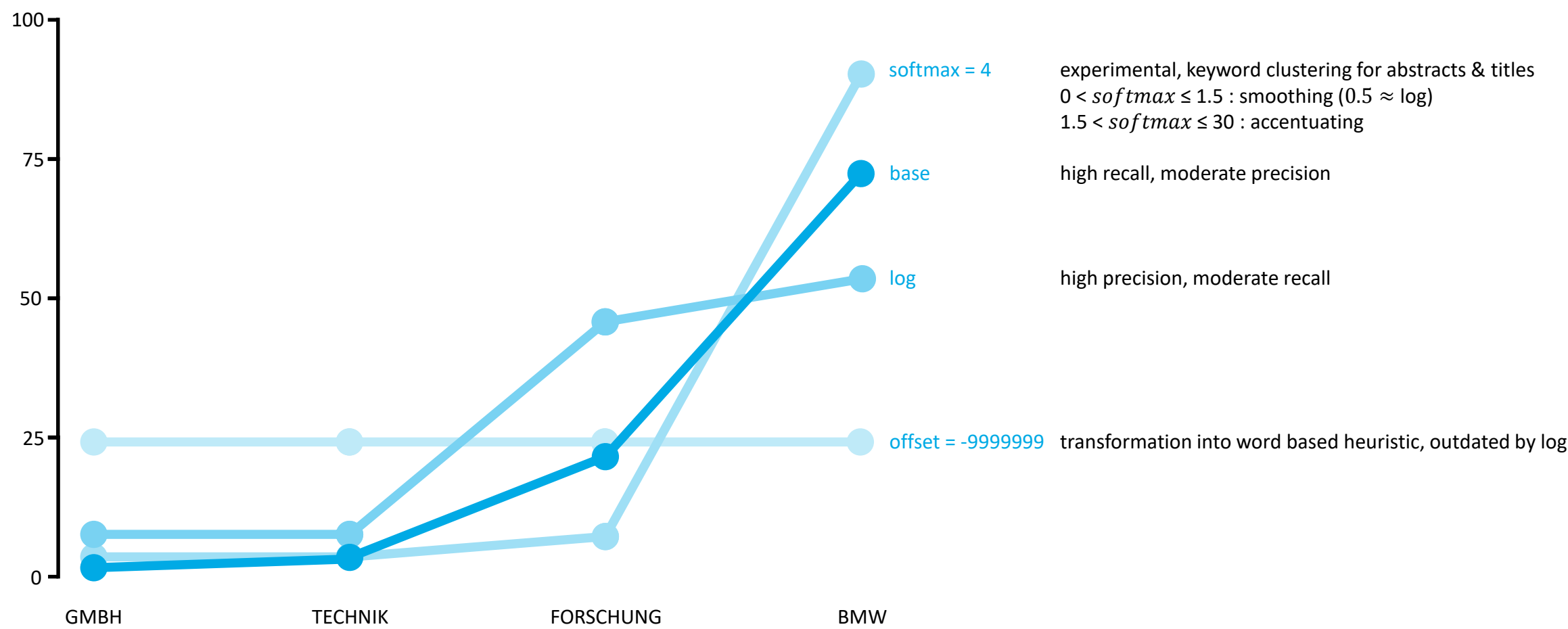
Refined components of the Identity lack the intrinsic frequency based heuristic. Every word has the same “Identification Potential”. This can be considered a tacit loss in information conveyed by the Identity. Plus, retrieval using destructive Preparer and the consecutive Refinement are slow.

Therefore search types using destructive preparers should be used sensibly and only after search steps based on conventional preparers to fill the gaps caused by misspellings.

Smoothing & Accentuating of the rIP



Smoothing & Accentuating of the rIP

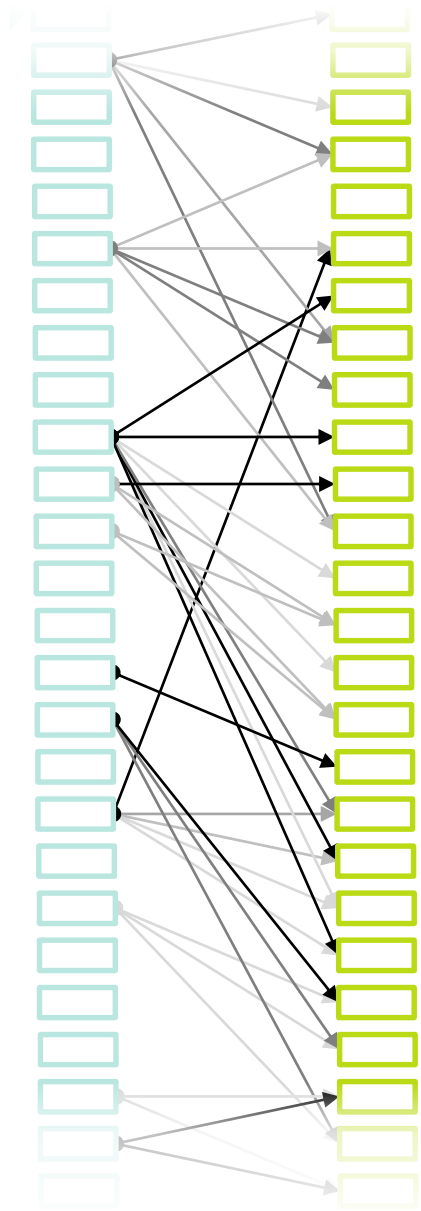


3-Gram Search for “Blaupause” with Log-Smoothing vs. Refinement

| identity | name | identity | name |
|----------|--|----------|--|
| 100 | Blaupause Bootsbau GmbH | 100.00 | Blaupause Bootsbau GmbH |
| 100 | Blaupause KfK GmbH & Co. KG | 100.00 | Blaupause KfK GmbH & Co. KG |
| 100 | Projekt Blaupause e.V. | 100.00 | Projekt Blaupause e.V. |
| 100 | Alexander Schilder 'Blaupause' | 100.00 | Alexander Schilder 'Blaupause' |
| 100 | blaupause e.V. | 100.00 | blaupause e.V. |
| 100 | BLAUPAUSE e.V. - mobile und flexible Hilfen für Menschen mit einer Alkoholerkrankung | 100.00 | BLAUPAUSE e.V. - mobile und flexible Hilfen für Menschen mit einer Alkoholerkrankung |
| 100 | Blaupause - Initiative für mentale Gesundheit im Gesundheitswesen e.V. | 100.00 | Blaupause - Initiative für mentale Gesundheit im Gesundheitswesen e.V. |
| 100 | Martin Alexander Rieger Blaupausen Medien Multimediaagentur | 100.00 | Blaupause UG Die Agentur für mehr |
| 100 | Blaupause UG Die Agentur für mehr | 100.00 | Blaupause Interior Design e.U. |
| 100 | Blaupause Zeichenbüro e.U. | 100.00 | Anja Thomä und Lena Dreesmann GbR "Blaupause" |
| 100 | Blaupause Projektentwicklung GmbH | 100.00 | Anja Thomä Blaupause papeterie |
| 100 | Blaupause Interior Design e.U. | 100.00 | Barnimer Alternative e.V. Jugendclub Blaupause |
| 100 | Anja Thomä und Lena Dreesmann GbR "Blaupause" | 100.00 | Jens Naumann Jan Welsch Blaupause GbR |
| 100 | Anja Thomä Blaupause papeterie | 100.00 | Agentur Blaupause 36 UG |
| 100 | Barnimer Alternative e.V. Jugendclub Blaupause | 88.89 | Heike Hartung Plaupause |
| 100 | Jens Naumann Jan Welsch Blaupause GbR | 88.89 | Plaupause - Bistro & Spätshop UG |
| 100 | Agentur Blaupause 36 UG | 85.00 | Martin Alexander Rieger Blaupausen Medien Multimediaagentur |
| 83.54 | Heike Hartung Plaupause | 85.00 | Blaupause Zeichenbüro e.U. |
| 83.54 | Plaupause - Bistro & Spätshop UG | 85.00 | Blaupause Projektentwicklung GmbH |

Why should “UPA” be much more relevant then “BLA”? Log-smoothing suppresses the undeserved dominance of specific n-grams. Still, refinement should always be engaged for n-grams because the SearchEngine notoriously disregards positioning. N-grams with and without smoothing complement each other, hence alternate both in dedicated search runs for best effect.

Unfocused Base Table



Firms are not in the Focus of the Data Collector

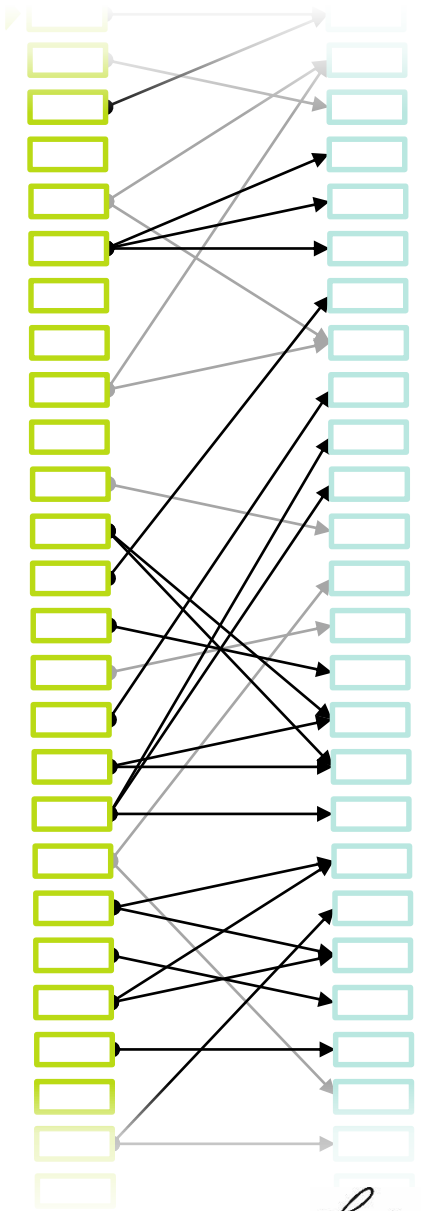
- Firms do not have a dedicated key
- Patent authorities and similar
- Scraped web data
- Focused on subordinate entities to firms, i.e. establishments, departments

Compound Search Strategy

- Low threshold as all potential candidates need to be captured: the good, the bad and the ugly
- Merging: candidate list of subsequent search runs have to be merged with existing lists
- Prone to false positives
- Candidates of a search record have variable identities

vs.

Focused Base Table



Data Collector Curates Firm Data

- Firms have a dedicated key, i.e. bvdid, tax number
- Duplicates are avoided
- Orbis, Compustat, creditreform, authorities

Incremental Search Strategy

- Multiple incremental search runs
- Darwinian: only the best candidates are picked
- Completing: search records that already have candidates are omitted from subsequent search runs
- Search parameters are gradually relaxed between runs
- Candidates of a search record have uniform identities

Decision Space

Search Table providing Search Terms

Base Table providing Candidates/Heuristic

unfocused or focus on subordinate entities vs. curated, focused on the search topic
concise representation of the search topic vs. data littered with additional noise and clutter
small or specialized dataset vs. sufficiently large unbiased sample/population

General Workflow

- Preparing the Data (search and base table)
 - Harmonize idiosyncratic disparities between the datasets, like consistently using different kinds of abbreviations. “Univ” vs. “University” vs. “Uni”, “limited” vs. “ltd”, “Strasse” vs. “Str.” and so on
Optionally: development of custom Preparers (there are already some in the “preparer” directory)
 - Removal of duplicate entries in the data especially in unfocused datasets, i.e. patent applicants, but also in panel data. This removes unwanted distortions of the heuristic and redundancies. Keep a linkage to the original data!
 - The SearchEngine imports tab-delimited text files with column headers (not too fancy). Avoid unnecessary fields! Avoid control characters in the data!
- Create the SearchEngine
 - Choose the base table by balancing the following priorities: being more focused, having more clutter, being significantly larger
 - Define search types by adding preparers to search fields. Usually, every search field has a conventional search type and only the most relevant field(s) will get an additional destructive search type using a n-gram preparer.
- Search Strategy for Firm Matching
 - Define a weight distribution according the relevancy of the search fields, ignoring additional destructive search types, e.g. 70% for relevant fields (firm name) and 30% for auxiliary fields (address). Define a containment by choosing a justifiable size for a large candidate list.
 - **Incremental**: complete search terms without candidates from earlier runs, gradually reduce retrieval restrictions, Darwinian (keep only the best)
 - Start with a high threshold requiring auxiliary fields using only conventional search types. You may even prelude with a run based on log smoothed conventional search types for the relevant fields followed by a non-smoothed version.
 - Switch to destructive search types by switching the corresponding weights and perform two search runs, one with log smoothing and one without smoothing of the destructive search types.
 - Lower the threshold forfeiting the necessity of auxiliary fields and repeat the search steps.
 - **Compound**: merge search runs based on a reasonably low threshold
 - Start with a conventional search run by setting the weights of the destructive search types to zero
 - Switch to destructive search types by switching the corresponding weights and perform two search runs, one with log smoothing and one without smoothing of the destructive search types.
- Concentrate on appropriate retrieval and less on identity refinement because the identity will be replaced by the meta vector carrying much more similarity indicators for the SearchEngine Machine Learning approach SEMML.

Search Strategy Components

- Threshold
 - Interaction with the search type weights constitutes the search strategy
 - A threshold above the firm name weight enforces partial address similarity
 - If the threshold is below the firm name weight, similarity in the address increases the leeway
- Redistribution of search type weights
 - Increasing/decreasing the impact of search types on the relative Identity
 - The more weight on address search types, the higher the leeway for the firm name
 - A weight of zero deactivates a search type for strategies with dedicated misspelling steps:
 - Start with runs not capturing misspellings → associated search types (n-grams) has a weight of zero
 - Continue with misspelling runs → switch weight to linguistic search types, set conventional to zero
- Containment
 - Keep weak search terms at bay by assessing a sensible cutoff: maximum expected number of plausible candidates
- Capturing misspellings
 - Activating linguistic search types (n-grams) by assigning a weight > 0 and deactivating the corresponding conventional search type
 - Usually only relevant search fields like the firm name are equipped with linguistic preparer
- Smoothing of the rIP distribution
 - A smoothed distribution requires more words to match → dominant words lose dominance
 - Higher precision at the expense of recall
 - Some search contexts fare better with smoothing, i.e. street addresses or person names
 - N-grams covering misspellings may become dominant preventing matches → two search runs: one with and one without smoothing
- Strategies over multiple search steps
 - Incremental: exclude search records with candidates from subsequent search runs, Darwinian (keeping only the best candidates) → [Focused Base Table](#)
 - Compound: candidates of subsequent runs are merged (union of candidate sets) → [Unfocused Base Table](#)

Search Strategy: Establishment Panel vs. Company Panel (MUP)

- Data
 - Mannheimer Unternehmens Panel (MUP 2000-2021, German Company Panel)
 - Removing duplicates → firm name & address aggregates: 24 million
 - Betriebsdatenpanel (Establishment Panel 2000-2020)
 - Harmonizing street addresses due to systematic changes over years
 - Removing duplicates → establishment & address aggregates: 21 million
- SearchEngine
 - Base table: MUP
 - Search types: name FIRMA, name FIRMA GRAM3, street STRASSE, zip, city
 - Darwinian search mode: pick only the candidates with the highest Identity
 - Incremental: skip search records with candidates in subsequent search runs

- Strategy

| run | description | threshold | establishments | candidates |
|-----|---|-----------|----------------|------------|
| 1 | name 70, name 3-gram 0, street 10, zip 10, city 10 | 85 | 9,393,950 | 14,531,161 |
| 2 | name 0, name 3-gram 70 log, street 10, zip 10, city 10 | 85 | 294,805 | 356,327 |
| 3 | name 0, name 3-gram 70, street 10, zip 10, city 10 | 85 | 32,075 | 39,234 |
| 4 | name 70 -9999999, name 3-gram 0, street 10, zip 10, city 10 | 85 | 841,062 | 1,296,231 |
| 5 | name 70, name 3-gram 0, street 10, zip 10, city 10 | 65 | 6,964,399 | 43,400,697 |
| 6 | name 0, name 3-gram 70 log, street 10, zip 10, city 10 | 65 | 571,931 | 801,739 |
| 7 | name 0, name 3-gram 70, street 10, zip 10, city 10 | 65 | 141,064 | 202,351 |
| | | | 18,239,286 | 60,627,740 |

Training Data

| searched | found | identity | equal | name | street | zip | city | score | cnt | run |
|----------|----------|----------|-------|--|---|-------|--------------------|-------|-----|-----|
| 5935 | | | 1 | Isotopen Technologien München AG | Rathausplatz 5 | 83435 | Bad Reichenhall | 2.65 | 5 | |
| 5935 | 684067 | 70.00 | | ITM Isotopen Technologien München AG | Theaterstr. 23 | 80333 | München | 2.65 | 5 | 5 |
| 5935 | 19425145 | 70.00 | | ITM Isotopen Technologien München AG | Walter-Meissner-Str. 2 | 85748 | Garching | 2.65 | 5 | 5 |
| 5935 | 19425146 | 70.00 | | ITM Isotopen Technologien München AG | Lichtenbergstr. 1 | 85748 | Garching | 2.65 | 5 | 5 |
| 5935 | 21419919 | 70.00 | | ITM Isotopen Technologien München AG | Schleißheimer Str. 91a | 85748 | Garching | 2.65 | 5 | 5 |
| 5935 | 21419920 | 70.00 | | ITM Isotopen Technologien München AG | Walther-von-Dyck-Str. 4 | 85748 | Garching | 2.65 | 5 | 5 |
| 3885 | | | 9 | Universität Stuttgart | Keplerstrasse 7 | 70174 | Stuttgart | 0.04 | 4 | |
| 3885 | 16183176 | 100.00 | 1 | Universität Stuttgart | Keplerstr. 7 | 70174 | Stuttgart | 0.04 | 4 | 1 |
| 3885 | 17625066 | 100.00 | | Akademische Motorsportgruppe an der Universität Stuttgart | Keplerstr. 7 | 70174 | Stuttgart | 0.04 | 4 | 1 |
| 3885 | 17706003 | 100.00 | | Vereinigung von Freunden der Universität Stuttgart | Keplerstr. 7 | 70174 | Stuttgart | 0.04 | 4 | 1 |
| 3885 | 17706007 | 100.00 | | Förderkreis Betriebswirtschaft an der Universität Stuttgart e.V. | Keplerstr. 7 | 70174 | Stuttgart | 0.04 | 4 | 1 |
| 40529 | | | 9 | Klaus Sindel Rusi-Kosmetik-Pinsel-Brushes GmbH | Ansbacher Strasse 53 | 91572 | Bechhofen | 0.05 | 4 | |
| 40529 | 19723005 | 85.69 | | Hauck Pinsel GmbH | Ansbacher Str. 47a | 91572 | Bechhofen | 0.05 | 4 | 9 |
| 40529 | 19700672 | 83.76 | | Johann Führ & Söhne Pinselfabrik GmbH | Ansbacher Str. 27-29 | 91572 | Bechhofen | 0.05 | 4 | 9 |
| 40529 | 19707191 | 82.98 | | Elco-Pinsel GmbH | Ansbacher Str. 86 | 91572 | Bechhofen | 0.05 | 4 | 9 |
| 40529 | 19773153 | 81.77 | | Ernst Bock & Sohn Pinselfabrik GmbH | Ansbacher Str. 68 | 91572 | Bechhofen | 0.05 | 4 | 9 |
| 25276 | | | 1 | RHODIA AG | Engesserstrasse 8 Postfach 1320 | 7800 | Freiburg | 2.41 | 4 | |
| 25276 | 4371760 | 80.32 | | RHONE-POULENC RHODIA AG | Engesserstr. 8 | 79108 | Freiburg | 2.41 | 4 | 1 |
| 25276 | 5408910 | 80.32 | | RP Rhodia AG | Engesserstrasse 8 | | Freiburg | 2.41 | 4 | 1 |
| 25276 | 5408911 | 80.32 | | RP Rhodia AG | Engesserstr. 8 | 79108 | Freiburg | 2.41 | 4 | 1 |
| 25276 | 15531543 | 80.32 | | Rhodia Acetow AG | Engesserstr. 8 | 79108 | Freiburg | 2.41 | 4 | 1 |
| 52085 | | | 1 | Karlsruher Institut für Technologie | Körperschaft des öffentlichen Rechts,Kaiserstrasse 12 | 76131 | Karlsruhe | 0.15 | 3 | |
| 52085 | 16156922 | 90.02 | 9 | Akademische Fliegergruppe am Karlsruher Institut für Technologie | Kaiserstr. 12 | 76131 | Karlsruhe | 0.15 | 3 | 1 |
| 52085 | 16156923 | 90.02 | 9 | Akademische Fliegergruppe am Karlsruher Institut für Technologie e.V. | Kaiserstr. 12 | 76131 | Karlsruhe | 0.15 | 3 | 1 |
| 52085 | 16240255 | 90.02 | | KIT Karlsruher Institut für Technologie | Kaiserstr. 12 | 76131 | Karlsruhe | 0.15 | 3 | 1 |
| 20010 | | | 1 | Dynamic Microsystems Semiconductor Equipment GmbH | Im Wiesengrund 17 | 78315 | Radolfzell | 0.09 | 2 | |
| 20010 | 16318610 | 82.32 | | DMS DYNAMIC MICRO SYSTEMS SEMICONDUCTOR EQUIPMENT GMBH | Im Wiesengrund 17 | 78315 | Radolfzell | 0.09 | 2 | 2 |
| 20010 | 16318611 | 82.32 | | DMS Dynamic Micro Systems Semiconductor Equipment GmbH | Im Wiesengrund 17 | 78315 | Radolfzell | 0.09 | 2 | 2 |
| 19904 | | | 1 | KARL BROTMANN CONSULTING GmbH | von Scheffel-Strasse 34 | 92224 | Amberg | 0.10 | 2 | |
| 19904 | 19376777 | 98.80 | | Karl Brotzmann Consulting GmbH | Von-Scheffel-Str. 34 | 92224 | Amberg | 0.10 | 2 | 3 |
| 19904 | 19776576 | 98.80 | | K a r l B r o t z m a n n C o n s u l t i n g GmbH | Von-Scheffel-Str. 34 | 92224 | Amberg | 0.10 | 2 | 3 |
| 22278 | | | 1 | Eerec Technology GmbH Development & Design | Borntalstrasse 9 | 36460 | Merkers/Thür. | 0.11 | 2 | |
| 22278 | 7273363 | 81.58 | | EuRec Technology GmbH Development & Design | Borntalstr. 9 | 36460 | Merkers-Kieselbach | 0.11 | 2 | 2 |
| 22278 | 7273364 | 81.58 | | EuRec Technology GmbH Development & Design | Borntalstr. 9 | 36460 | Merkers | 0.11 | 2 | 2 |
| 33792 | | | 1 | A L M Ü PRAZISIONSWERKZEUG GmbH | Ohmder Strasse 12 | 73119 | Zell | 1.74 | 1 | |
| 33792 | 15622884 | 98.70 | | ALMÜ Präzisions-Werkzeug GmbH | Ohmder Str. 12 | 73119 | Zell | 1.74 | 1 | 8 |
| 3092 | | | 1 | Gesellschaft zur Förderung angewandter Optik, Optoelektronik, Quantenelektronik und Spektroskopie e.V. | Rudower Chaussee 29 (IGZ) | 12489 | Berlin | 2.04 | 1 | |
| 3092 | 483360 | 82.25 | | OPTOSENS Optische Spektroskopie und Sensortechnik GmbH | Rudower Chaussee 29(IGZ) | 12489 | Berlin | 2.04 | 1 | 7 |

Meta Vector Components

Heuristic

Absolute Identification Potential: $aIP(w) = 1 - \ln(occ(w, st_w)) / \ln(maxocc(st_w))$

Report only the n largest aIP in descending order for...

...matching words in search term and candidate

...words exclusive to the candidate

...words exclusive to the search term (requires auxiliary registry of the search table)

for all search types, i.e.: name = 5, name GRAM3 = 15, street = 3, zip = 1, city = 2

Visual

Asymmetric string distances based on maximizing word-by-word comparisons between search term and candidate and vice versa for all search fields (independent of the positioning of words).

Overlap

Similarities between search field components, i.e. city name repeats in firm name

Descriptive

Candidate block statistics

- Number of candidates for the same search record
- Number of distinct identities among those candidates
- Percentile rank position within candidates
- Standard deviations of the string distances (calculated externally)

- Relatively slim parameter set per observation (around 110 variables)
 - Low risk of over-specification: having too many variables for too little data
- No semantics
 - “second hand metal wares” and “scrapyard” are not identified as tantamount
 - Lenient labeling required (use post-processing with core data to settle ambiguities)

SEML (SearchEngine Machine Learning): Establishments vs. Company Panel

- SEML components
 - `brain.ado` – Neural Network module for Stata
 - `seml_train.do`
 - Iterates hidden neuron setups: [25], [50], [100], [25x25], [50x50], [100x100]
 - Reports the best performing NN pertaining out-of-sample prediction
 - `seml_think.do` – implements the winning NN setup on the whole meta data
 - BYOD – bring your own device, i.e. Tensor Flow, Keras, Random Forest, ...
- Training sample
 - 2,000 establishments paired with 5,523 candidates
 - 569 pairings retained for out-of-sample prediction

- Confusion matrix

| Probit | True | False |
|-----------|--------|--------|
| Positive | 180 | 10 |
| Negative | 367 | 12 |
| Recall | 93.75% | 97.35% |
| Precision | 94.74% | 96.83% |
| Accuracy | 96.13% | |

| NN[25x25] | True | False |
|-----------|--------|--------|
| Positive | 184 | 3 |
| Negative | 374 | 8 |
| Recall | 95.83% | 99.20% |
| Precision | 98.40% | 97.91% |
| Accuracy | 98.07% | |

- Quality assessment
 - 85% of the establishments matched to a company
 - Final quality assessment is difficult due to structural disparities, i.e. self-employment vs. owner operated firm

Search Strategy: German EPO Firm Applicants vs. Company Panel (MUP)

- Preparation
 - Mannheimer Unternehmens Panel (MUP 2000-2021, German Company Panel)
 - Removing duplicates → firm & address aggregates: 24 million
 - German EPO Applicants (Patstat 2021)
 - Removing person owned patents with less than 10 patents
 - Removing duplicates → firm & address aggregates: 62,036
- SearchEngine
 - Base table: MUP
 - Search types: name FIRMA, name FIRMA GRAM3, street STRASSE, zip, city
 - Darwinian search mode: pick only the candidates with the highest Identity
 - Incremental: skip search records with candidates in subsequent search runs
- Strategy

| run | description | threshold | applicants | candidates |
|-----|---|-----------|------------|------------|
| 1 | name 70, name 3-gram 0, street 10, zip 10, city 10 | 79 | 56,603 | 108,775 |
| 2 | name 0, name 3-gram 70, street 10, zip 10, city 10 | 79 | 682 | 977 |
| 3 | name 0, name 3-gram 70 log, street 10, zip 10, city 10 | 79 | 275 | 385 |
| 4 | name 70 -9999999, name 3-gram 0, street 10, zip 10, city 10 | 79 | 167 | 358 |
| 5 | name 70, name 3-gram 0, street 10, zip 10, city 10 | 65 | 3,078 | 8,251 |
| 6 | name 70 log, name 3-gram 0, street 10, zip 10, city 10 | 65 | 69 | 134 |
| 7 | name 0, name 3-gram 40, street 20, zip 20, city 20 | 80 | 17 | 20 |
| 8 | name 0, name 3-gram 40 log, street 20, zip 20, city 20 | 80 | 34 | 53 |
| 9 | name 0, name 3-gram 40, street 20, zip 20, city 20 | zealous | 35 | 61 |
| | | | 60,960 | 119,014 |

SEML: German EPO Applicants vs. Company Panel

- Training sample
 - 3,264 applicants paired with 6,535 candidates
 - 677 pairings retained for out-of-sample prediction
 - + completely scrutinized search runs 4 & 6 to 9 due to marginal representation

- Confusion matrix

| Probit | True | False |
|-----------|--------|--------|
| Positive | 575 | 13 |
| Negative | 70 | 19 |
| Recall | 96.80% | 84.34% |
| Precision | 97.79% | 78.65% |
| Accuracy | 95.27% | |

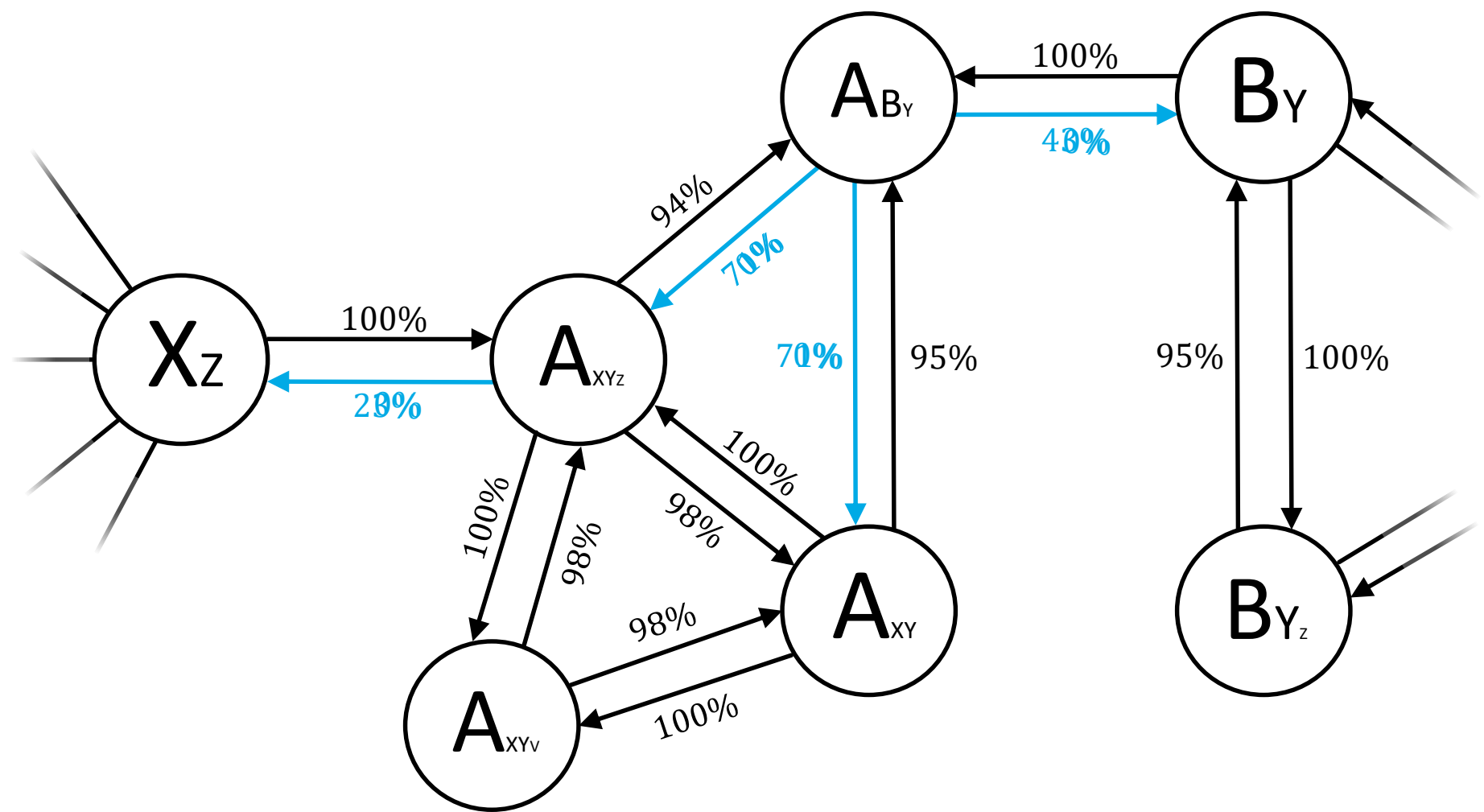
| NN[25] | True | False |
|-----------|--------|--------|
| Positive | 590 | 5 |
| Negative | 78 | 4 |
| Recall | 99.33% | 93.98% |
| Precision | 99.16% | 95.12% |
| Accuracy | 98.67% | |

- Quality assessment
 - 59,630 of 62,036 applicants assigned covering 98.81% of all German firm patents

Entity Resolution: Search Table = Base Table

| searched | found | identity | equal | name | zip | city | street | country |
|----------|--------|----------|-------|--|-------|----------------|--|---------|
| 203346 | | | | SCHERING AKTIENGESELLSCHAFT PATENTE | 13342 | BERLIN | MUELLERSTRASSE 178 POSTFACH 65 03 11 | |
| 203346 | 203346 | 100.00 | | SCHERING AKTIENGESELLSCHAFT PATENTE | 13342 | BERLIN | MUELLERSTRASSE 178 POSTFACH 65 03 11 | DE |
| 71132 | | | | HOECHST SCHERING AGR EVO GMBH | 13342 | BERLIN | GERICHTSTRASSE 27 | |
| 71132 | 71132 | 100.00 | | HOECHST SCHERING AGR EVO GMBH | 13342 | BERLIN | GERICHTSTRASSE 27 | DE |
| 71132 | 66486 | 95.00 | | HOECHST SCHERING AGR EVO GMBH | 13347 | BERLIN | GERICHTSTRASSE 27 | DE |
| 71132 | 73565 | 90.00 | | HOECHST SCHERING AGR EVO GMBH | 13509 | BERLIN | MIRAUSTRASSE 54 | DE |
| 323602 | | | | SCHERING AG | 13342 | BERLIN | MUELLERSTRASSE 170 178 | |
| 323602 | 323602 | 100.00 | | SCHERING AG | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 323602 | 389129 | 99.57 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 323602 | 199658 | 99.29 | | SCHERING AKTIENGESELLSCHAFT | 13342 | BERLIN | PATENTE MUELLERSTRASSE 178 POSTFACH 65 03 11 | DE |
| 323602 | 203346 | 99.29 | | SCHERING AKTIENGESELLSCHAFT PATENTE | 13342 | BERLIN | MUELLERSTRASSE 178 POSTFACH 65 03 11 | DE |
| 323602 | 402998 | 95.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 323602 | 180193 | 94.73 | | SCHERING AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 323602 | 303857 | 94.73 | | SCHERING AG | 13353 | BERLIN WEDDING | MUELLERSTRASSE 178 | DE |
| 323602 | 397563 | 94.73 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 323602 | 264 | 94.57 | | SCHERING AKTIENGESELLSCHAFT | 13342 | BERLIN | | DE |
| 323602 | 71132 | 94.57 | | HOECHST SCHERING AGR EVO GMBH | 13342 | BERLIN | GERICHTSTRASSE 27 | DE |
| 323602 | 171208 | 94.29 | | SCHERING AKTIENGESELLSCHAFT | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 323602 | 435123 | 94.29 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | PATENTS LICENSING MUELLERSTRASSE 178 | DE |
| 402998 | | | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | |
| 402998 | 402998 | 100.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 402998 | 397563 | 99.73 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 402998 | 435123 | 99.40 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | PATENTS LICENSING MUELLERSTRASSE 178 | DE |
| 402998 | 389129 | 94.68 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 518249 | | | | BAYER PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | |
| 518249 | 518249 | 100.00 | | BAYER PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 397563 | 100.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 402998 | 100.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 518249 | 435123 | 98.76 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | PATENTS LICENSING MUELLERSTRASSE 178 | DE |
| 518249 | 441578 | 98.76 | | BAYER PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 543908 | 94.52 | | BAYER PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 389129 | 93.76 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |

Intransitive Similarity Network: Directed Graph



Search

- Required: high threshold, i.e. 90%
- Connections in both directions enable transitivity
- Connections in one direction cause intransitivity

Mirroring

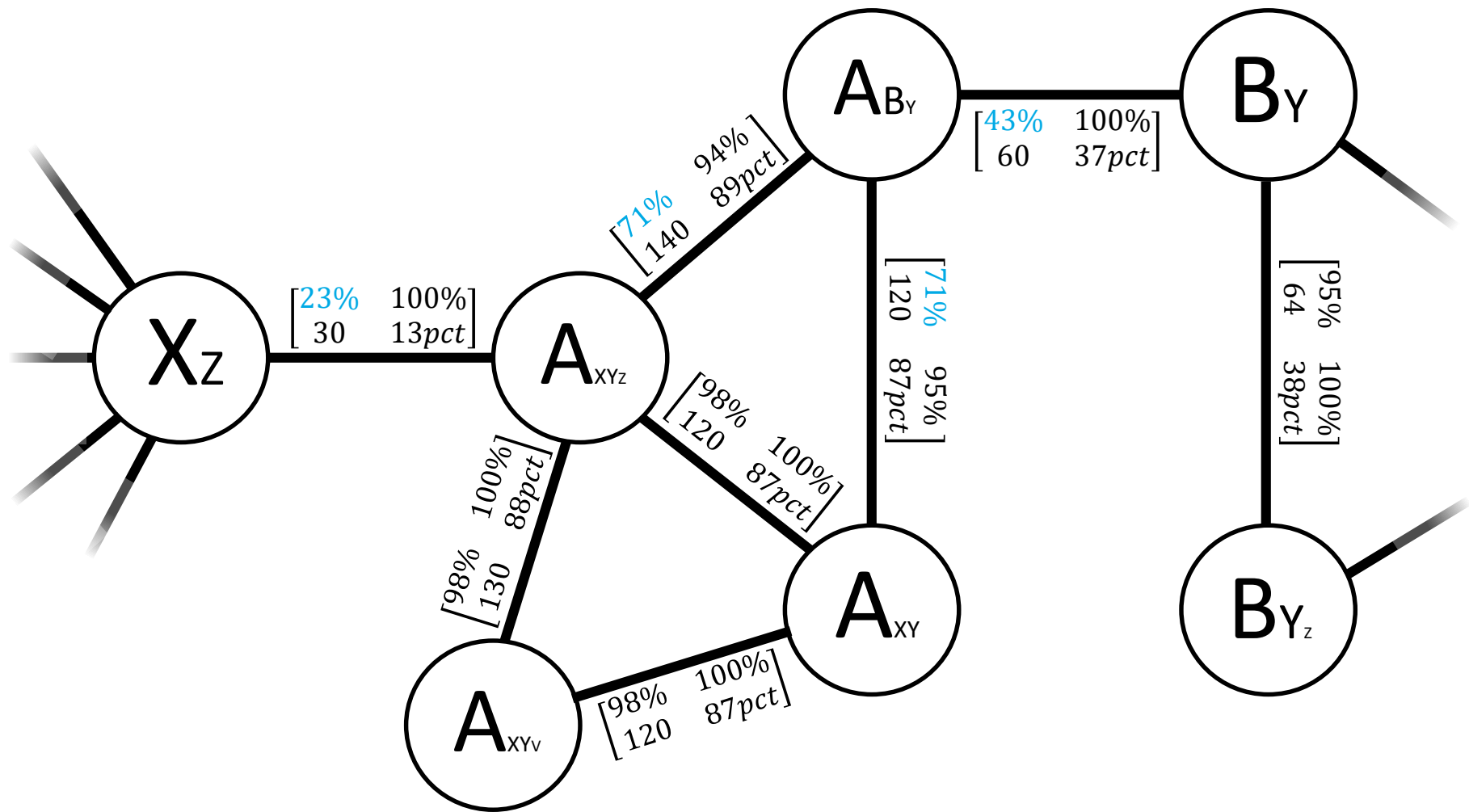
- Creates a **dummy run n** for all reversed connections not yet established
- Mirrored connections always have Identity zero

Research

- Reevaluates Identity for selected runs given current settings
- Apply on **dummy run n**
- Log-smoothing on critical search types is advised as retrieval is of no concern

$$A_{BY} \rightarrow AB_Y$$

Intransitive Similarity Network: Undirected Graph



quality attributes

$$\begin{bmatrix} min & max \\ s & p \end{bmatrix}$$

min, max of the identities
of both directions

s = Score

The score of a word is its associated weight divided by its frequency. The score of a search term is the sum off all word scores.

The minimum score of both involved search terms is used.

p = Score Percentile

The score has an arbitrary value range while the percentile stays within [0,100].

Nested Cascaded Traversal

It is like Spelunking!

- Choose an entrance to an unexplored cave (starting node).
- Shine with your flashlight into all tunnels (connections) branching out from the current room (node). Avoid all tunnels that appear too derelict according to safety guidelines (rules imposed on the quality attributes).
- Mark every passed room and tunnel to not run in circles or getting lost while backtracking from a dead end.
- If you have explored too many rooms, your expedition is abandoned and a new one with stricter guidelines enters the cave entrance.

Separates cascaded rules. Current rule is active unless the threshold of the subsequent rule is breached. Each new rule has to be more restrictive than the previous one to guarantee non-overlapping clusters.

Comma

Separates cascades. Consolidates the clusters into hyper-nodes. The quality attributes of connections between encased nodes of different hyper-nodes are aggregated to the respective maximum forming a hyper-network.

Semicolon

$min \geq 90$ or $p \geq 70$ and $min \geq 70 @ 0$, $min \geq 90 @ 21$; $min \geq 90 @ 4$

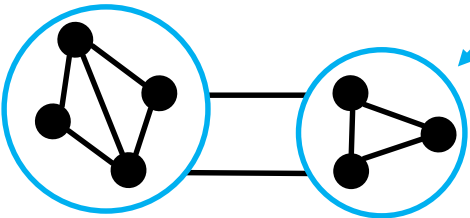
Rule

Imposes restrictions on the validity of connections based on the quality attributes. During traversal, these imply the boundaries of the resulting cluster.

Cluster Threshold

Forces the activation of the associated Rule when the intermediate cluster size attains the number. Resets the traversal to the starting node under the new regime.

Hyper-Network



The threshold refers to hyper-nodes. In this example, three unobserved transitions between hyper-nodes are condoned until the rule kicks in to enforce transitivity.

- First cascade establishes transitive clusters of high coherence
- Optional second cascade for audaciously intransitive transitions exploiting the tendency of thickets being less contained than plausible clusters (most of the time)

Clustering German EPO Applicants by Name

| 2. Cascade | 1. Cascade | name |
|------------|------------|---|
| 17840 | 17840 | HANS KNOELL INSTITUT FUER NATURSTOFF FORSCHUNG |
| 17840 | 17840 | HANS KNOELL INSTITUT FUER NATURSTOFF FORSCHUNG E V |
| 17840 | 17840 | HANS KNOELL INSTITUT LEIBNIZ INSTITUT FUER NATURSTOFF FORSCHUNG |
| 17840 | 25877 | LEIBNIZ INSTITUT FUER NATURSTOFF FORSCHUNG UND INFEKTIONS BIOLOGIE E V HANS KNOELL INSTITUT |
| 17840 | 25877 | LEIBNIZ INSTITUT FUER NATURSTOFF FORSCHUNG UND INFEKTIONS BIOLOGIE E V HANS KNOELL INSTITUT HKI |
| 17840 | 25877 | LEIBNIZ INSTITUT FUER NATURSTOFF FORSCHUNG UND INFEKTIONS BIOLOGIE HANS KNOELL INSTITUT |
| 17840 | 25876 | LEIBNIZ INSTITUT FUER NATURSTOFF FORSCHUNG UND INFEKTIONS BIOLOGIE |

| 2. Cascade | 1. Cascade | name |
|------------|------------|------------------------------------|
| 13887 | 13887 | FISCHER FORTUNA GMBH |
| 13887 | 13887 | FORTUNA MASCHINENBAU HOLDING AG |
| 13887 | 13887 | FORTUNA SPEZIALMASCHINEN GMBH |
| 13887 | 13887 | FORTUNA WERKE MASCHINENFABRIK GMBH |
| 13887 | 15891 | DR SCHICK GMBH |
| 13887 | 15891 | GEORG SCHICK DENTAL GMBH |
| 13887 | 14322 | FORTUNA VERTRIEB DR G SCHICK GMBH |



SearchEngine

<https://github.com/ThorstenDoherr/searchengine>

Brain – Neural Network for Stata

<https://github.com/ThorstenDoherr/brain>
ssc install brain

Thank you for your attention
Time for questions