Swiss Re

Catching the big corporates – record linkage algorithm for company names

Aurélien Severino, Antoine Logean, Tobias Bröckl, Cristina Turcu, Péter Szirmai, Danilo Biella, Nikita Kuksin

Context: *What is in my contract?*

- Reinsurance ?

→ "insurance for insurers":

- Reinsurance contract ?

→ insures/covers an insurance *portfolio*:

Swiss Re

↓

typical reinsurance contract →

Insurance Company XYZ

Portfolio of XYZ
Google
Apple
Facebook
Amazon
Microsoft
… *(& many more!)*

The tough job of identifying companies

Status quo *"What companies are in my contract?"*

1. We need a 'reference' **company repository**
2. We need to **map it** to your portfolio

Problem *"Real-life data is messy & complex"*

- Google ⟷ Alphabet (? ¿)
- BMW ⟷ Bayerische Motorwerke AG (? ¿)

Our Solution ***"CorpFinder"***

R package for record linkage of company names

- Using similarity-based string matching; taking into account corporate ownership tree; explicitly accounting for legal entity suffixes; using ad-hoc deduplication approaches…

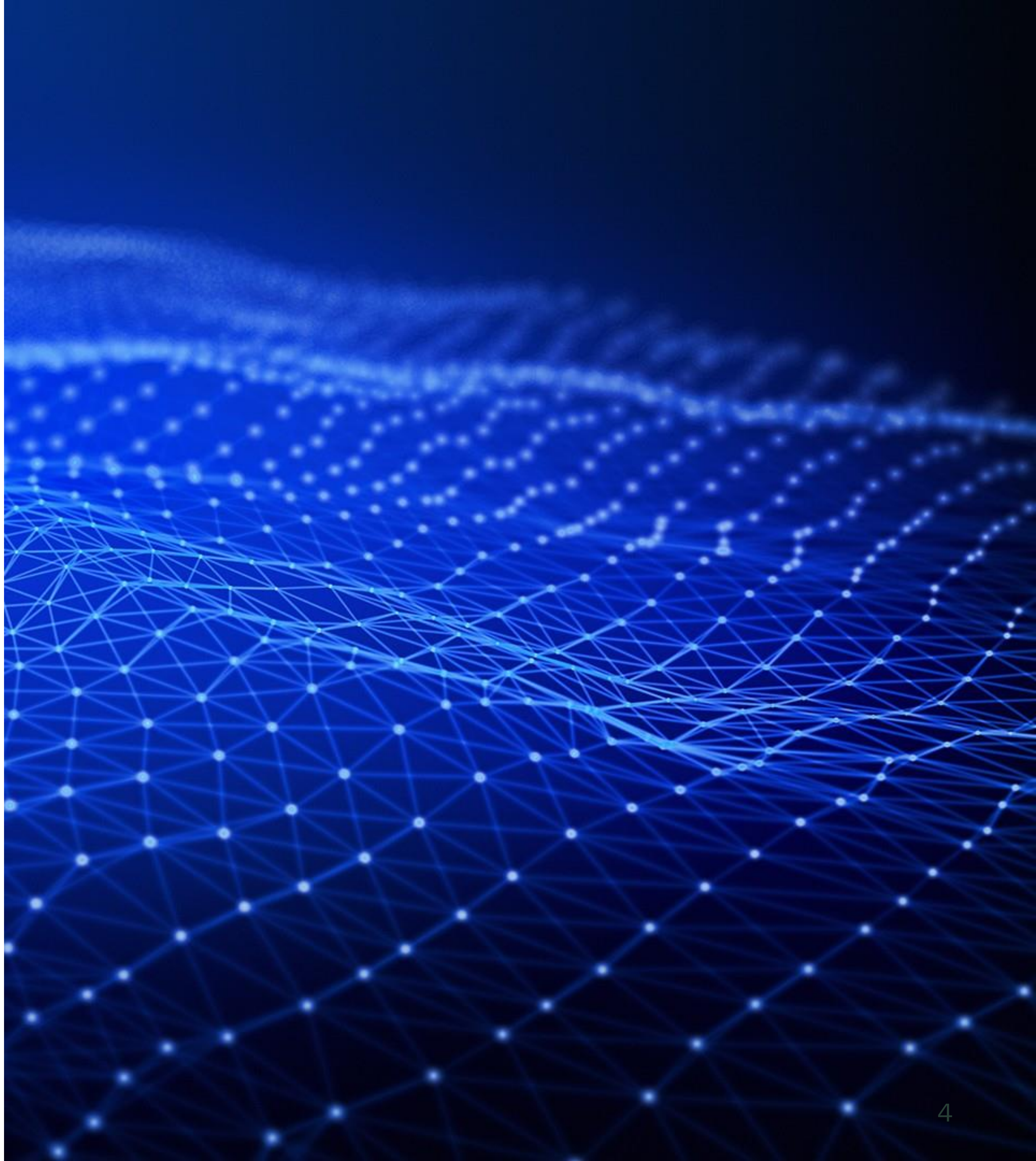
Why to search for large corporates?

Large Corporate Risks (LCRs) present special characteristics for a (re)insurer:

- **Deep pockets**:
  - Reputation leads to legal costs.
  - Reparation costs are large.
- **Risk Accumulation**:
  - Complex subsidiary structure.
  - Accumulation of risk exposure.

Entity (record) linkage

- **Deterministic record linkage**
  - ifelse(*ID1 == ID2, Person1 == Person2, Person1!=Person2)*
- **Probabilistic record linkage**
  - String distance measures
    => Better control/interpretability than ML
    => Efficient when only one dimension available
- **Machine Learning methods**
  - Regression
  - Naïve Bayes method
  - GNN
  - ...



Our modular solution

Component I
Company name normalization

Component II
Matching to reference list

Component III
Disambiguation

Component I: *Name normalization*

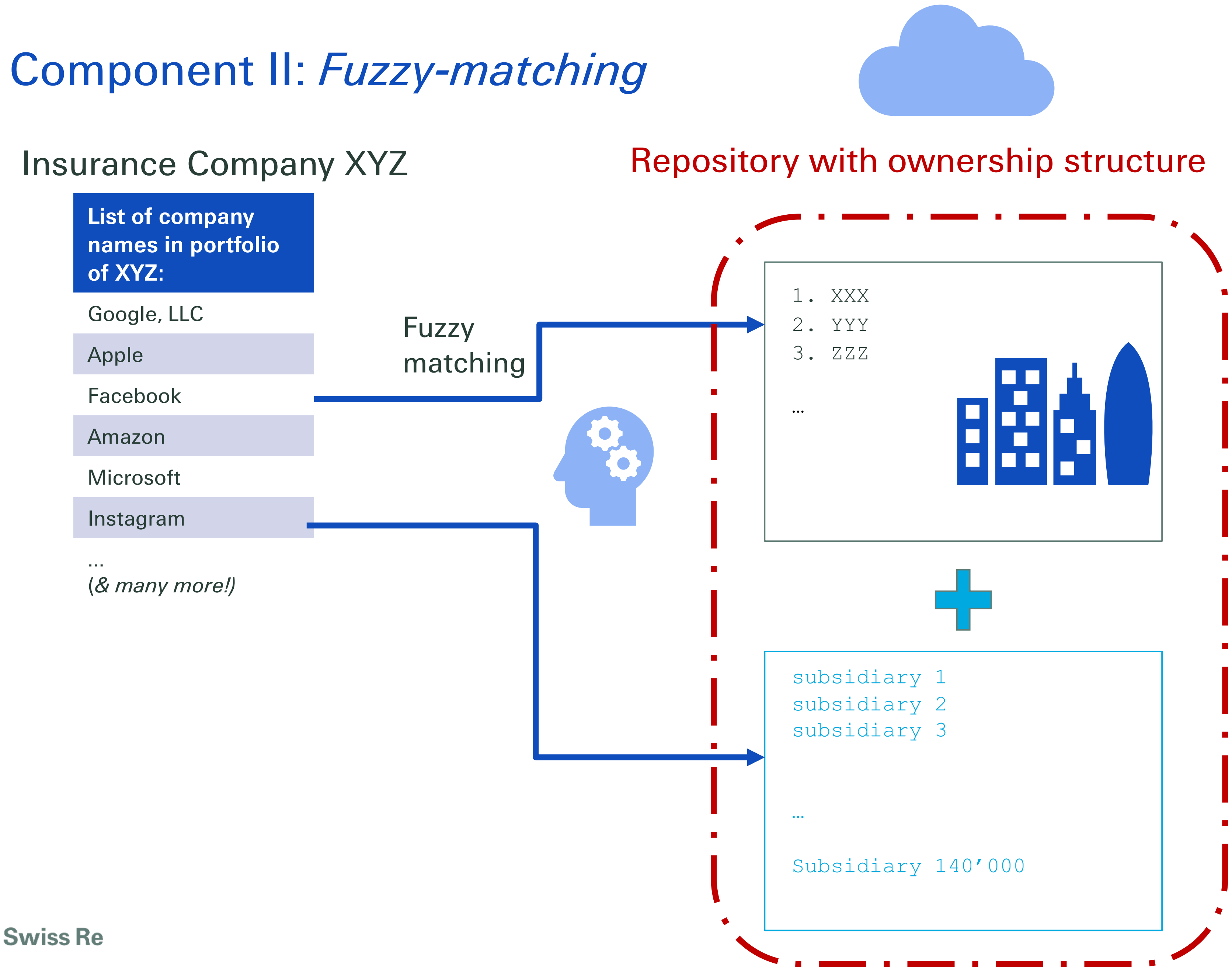
- Specificity of company names: often contain legal suffixes or prefixes, and in no consistent way:
  - *Apple Inc.* vs Apple vs. *Apple Incorporated* vs. *Apple Inc* describe the same entity.
  - Specific to languages, countries and legal structures: *pjsc* [russian], *oyj* [finish], *sa/nv* [belgium], etc...

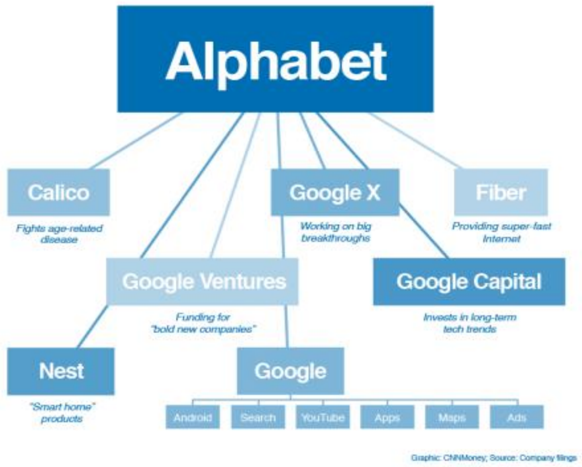
- Hence: legal suffixes are isolated away from company names, based on an ad-hoc 'legal dictionary' (but they are kept and stored for use in *Component III*)

- Various additional normalization steps (stopwords removal, accent/special characters standardization, etc.)

- Considering 'aliases': e.g. BMW vs Bayerisches motorwerk

Component II: *Fuzzy-matching*





In practice

- Take ‘*normalized names*’ as input on both sides and compute pairwise distance
- Pick best score (as given by metric) as ‘*match*’, using a threshold

Component II: *Fuzzy-matching*

Which metric do we pick?

- Levenshtein
- Jaro-Winkler
- Jaccard (or token-based)

How to set a threshold?

- Fine-tune using a test-set to keep sensitivity/recall balanced

How to improve performance?

- Parallelize
- Cache
- Reduce search space

Component III: *Disambiguation*

- **Problem**: How to handle matches with equally good score?
  - E.g. *"Coca-Cola"* vs. *"Coca-Cola US"* and *"Coca-Cola UK"*
  - *"Freedom"* vs. *"Freedom Corp."* and *"Freedom SAS"*
  - ...

- **Our solution**:
  - **'Forbidden' associations:** e.g. *Apple Ltd* cannot match *Apple Inc*
  - **Different countries**: e.g. *Apple SAS* cannot match with *Apple Inc*.
  - **Matches on same tree:** If matches belong to the same ownership tree, the entry is matched to the root of the entries

Details of the package

- Not on GitHub but planning to make it available by 2021

- Access to our package
  - **Shiny application for internal users:** file upload for fuzzy matching
  - **Package deployed on Cran**:
    - Exported functions set up using one list of configuration
    - Fuzzy matching with a pre-defined or any user-defined list

Outlook

- Open-source package on GitHub
- Inclusion of multiple dimensions
- Probabilistic record linkage as a reference for ML techniques
- User-trained ML algorithm