

Semantic Text Matching of Company Policies and Regulatory Documents using Text Similarity Measures

Christoph Erl, 21. August 2017, Munich
Advisor: Jörg Landthaler

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

Motivation

- Problem Statement
- Industry Partner Alyne GmbH

Research Questions

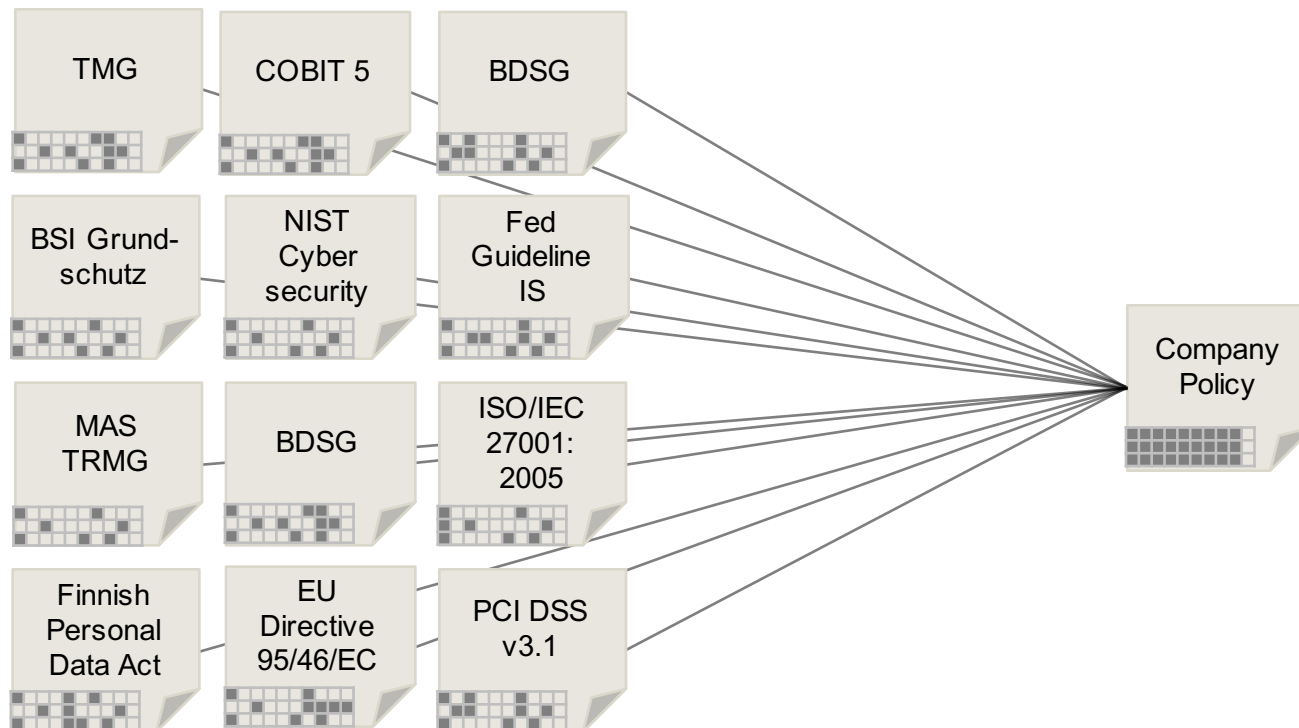
Research Approach

- Basic Concept
- Pre-Processing Technologies
- Editorial Help
- Iterative Approach

Timeline

Problem Statement

- Companies need to comply to international standards and regulations - both growing in number and complexity.
- These **regulatory documents** overlap in requirements and are written on different levels of detail.
- Large enterprises typically maintain individual **company policies**.
- Audits or updates require a **mapping** between regulatory documents and company policies.



Motivation

Industry Partner Alyne GmbH



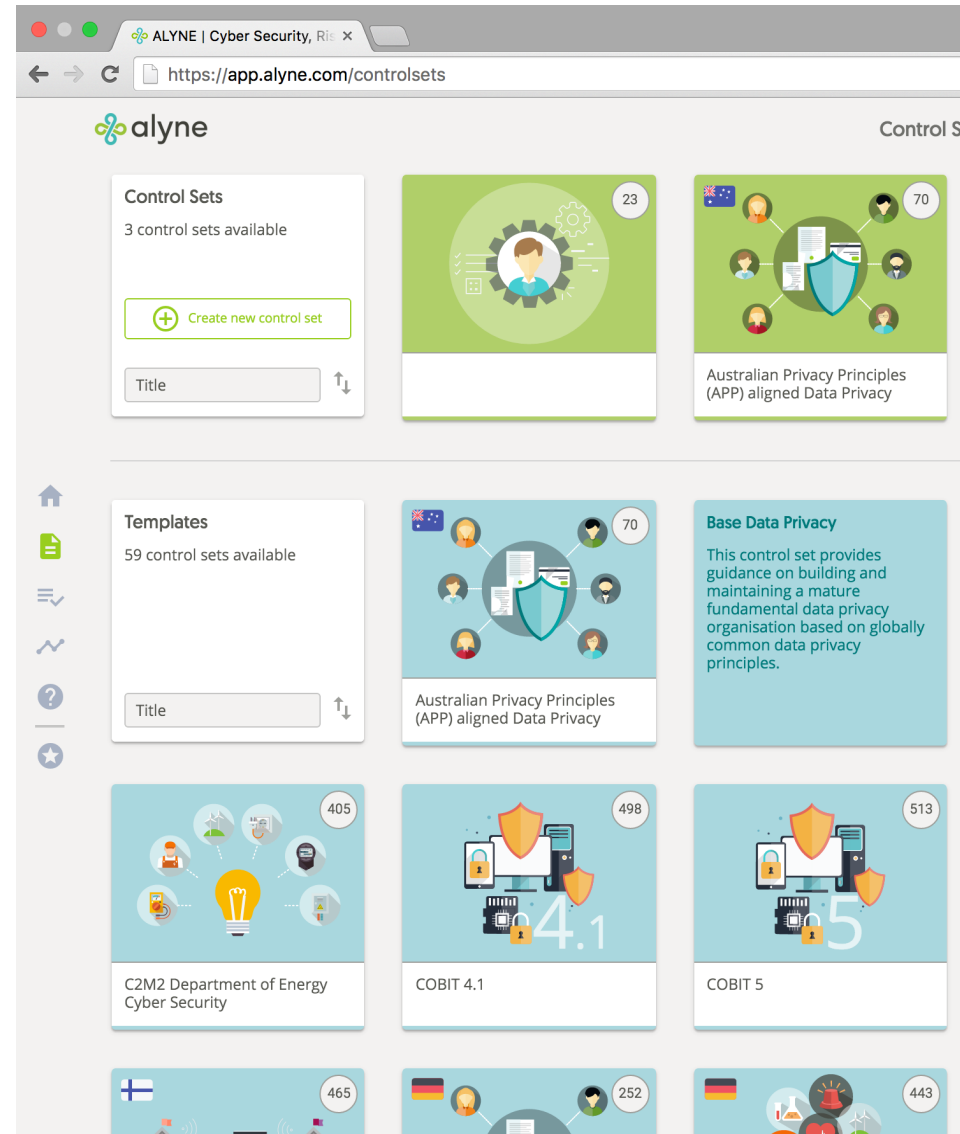
Startup: 4 Founders, Launch in 2015

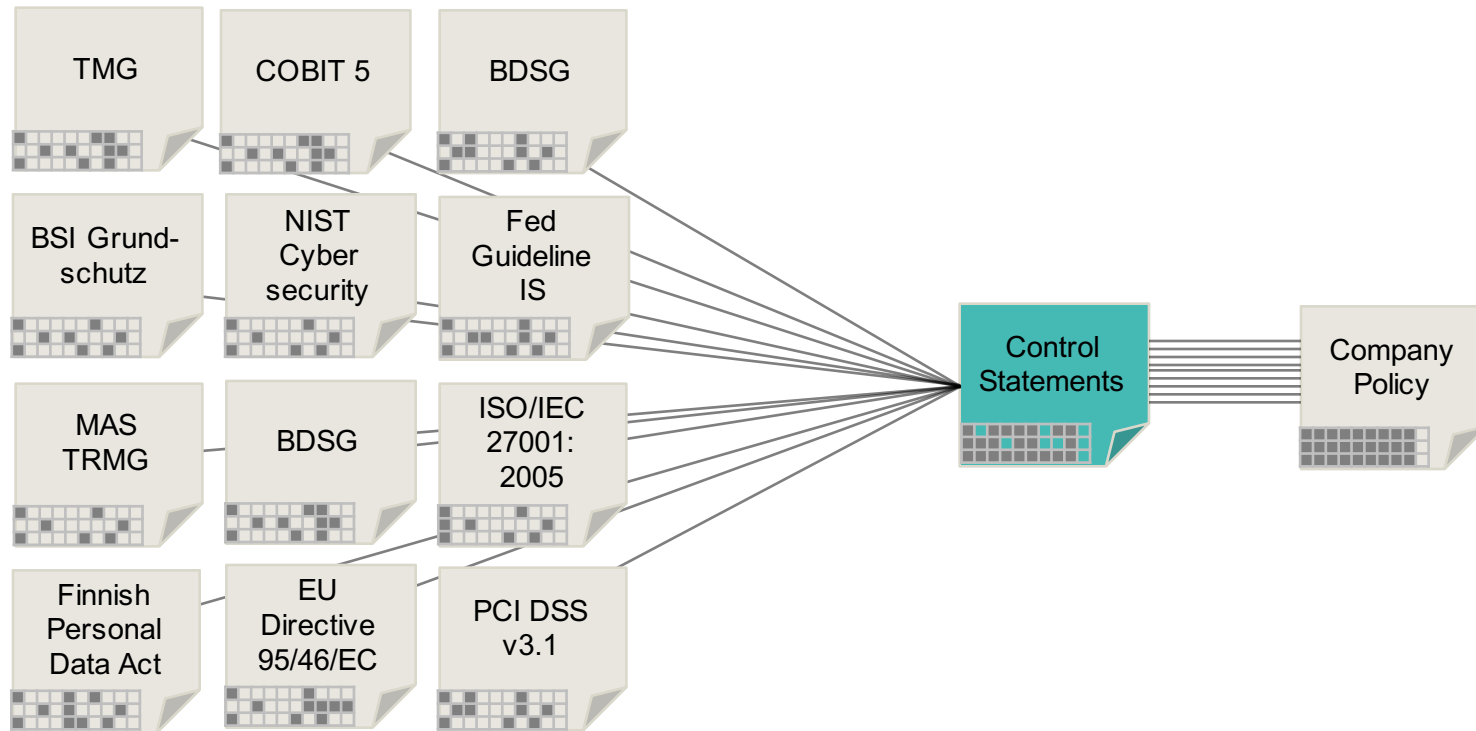
Solution: Software-as-a-Service (SAAS)

Type: B2B

Business Idea:

- Alyne delivers a software for Risk Management and Compliance
- Generalized **control statements** and risks serve as a glue between **company policies** and **regulatory documents**





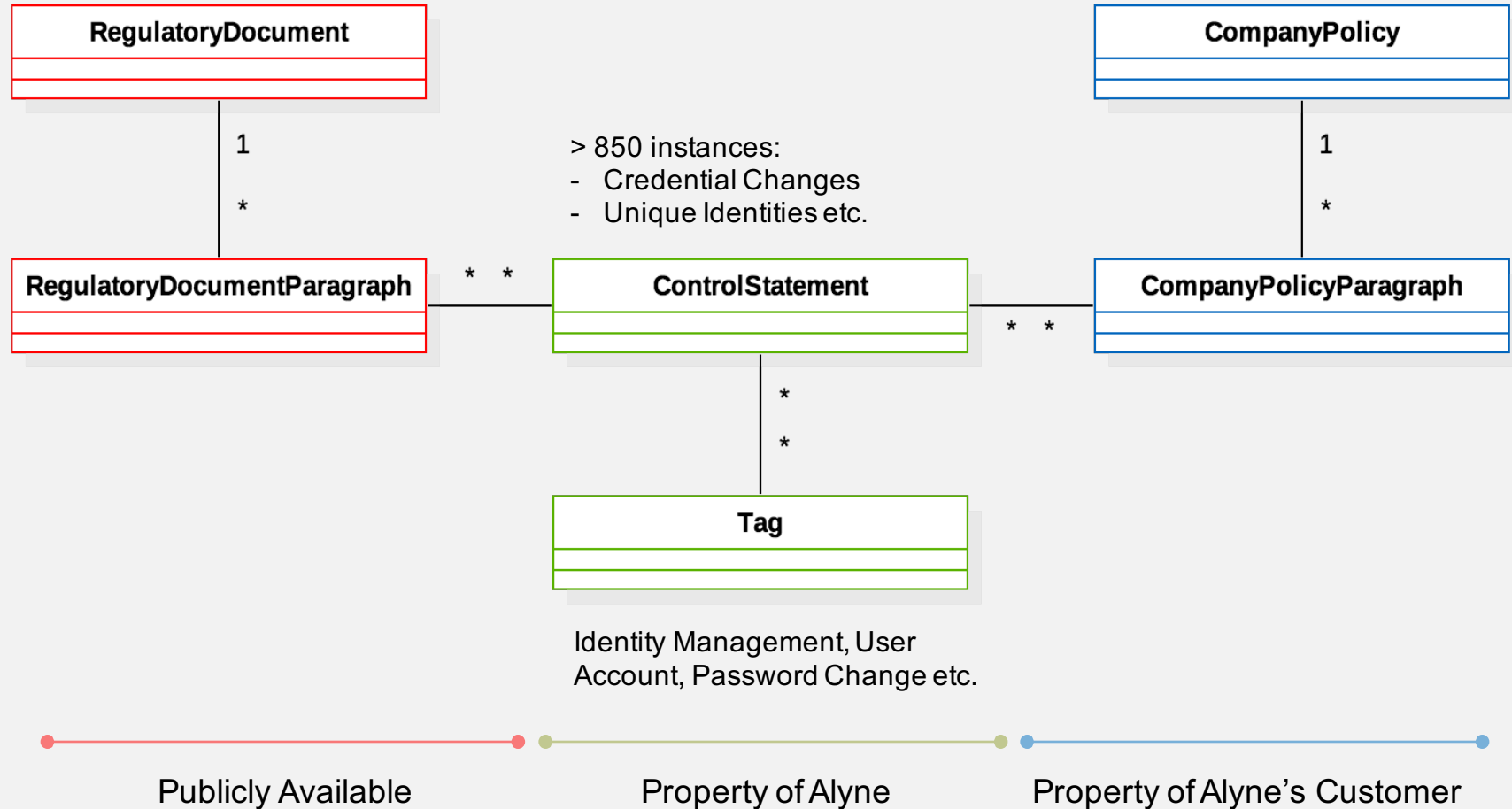
Motivation

Industry Partner Alyne GmbH – Data Model



29 Instances:

- EU-DSGVO
- BDSG etc.



29 Instances:

- EU DIRECTIVE
-

RegulatoryDocument

1

*

RegulatoryDocumentParagraph

4.5.2016

EN

Official Journal of the European Union

L 119/1

I

(Legislative acts)

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee ⁽¹⁾,

Having regard to the opinion of the Committee of the Regions ⁽²⁾,

Acting in accordance with the ordinary legislative procedure ⁽³⁾,

Whereas:

(1) The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

(2) The principles of, and rules on the protection of natural persons with regard to the processing of their personal data should, whatever their nationality or residence, respect their fundamental rights and freedoms, in particular their right to the protection of personal data. This Regulation is intended to contribute to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market, and to the well-being of natural persons.

(3) Directive 95/46/EC of the European Parliament and of the Council ⁽⁴⁾ seeks to harmonise the protection of fundamental rights and freedoms of natural persons in respect of processing activities and to ensure the free flow of personal data between Member States.

⁽¹⁾ OJ C 229, 31.7.2012, p. 90.

⁽²⁾ OJ C 391, 18.12.2012, p. 127.

⁽³⁾ Position of the European Parliament of 12 March 2014 (not yet published in the Official Journal) and position of the Council at first reading of 8 April 2016 (not yet published in the Official Journal). Position of the European Parliament of 14 April 2016.

⁽⁴⁾ Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ L 281, 23.11.1995, p. 31).

Paragraphs

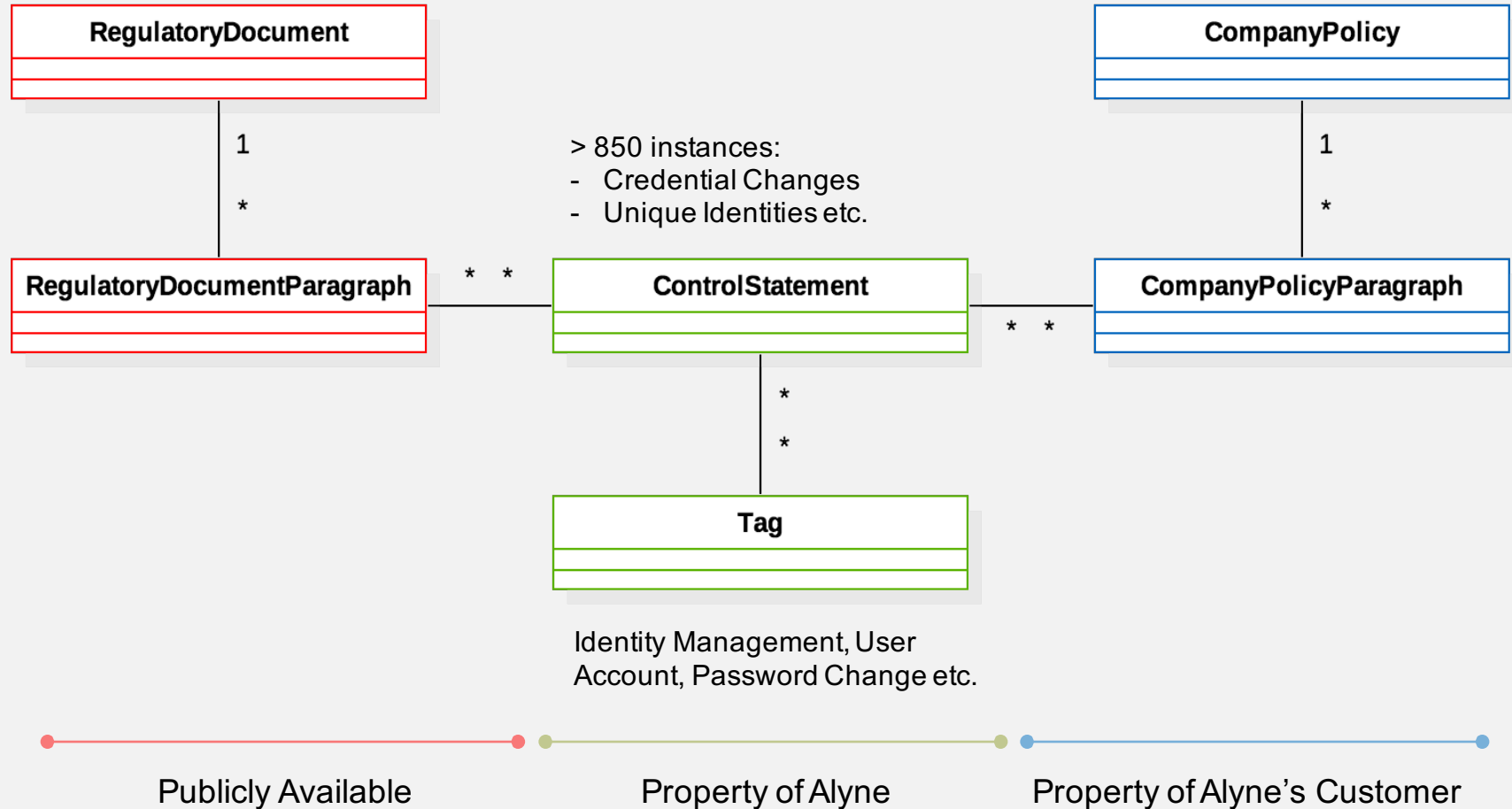
Motivation

Industry Partner Alyne GmbH – Data Model



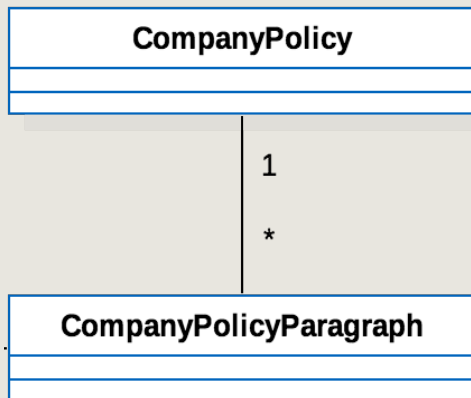
29 Instances:

- EU-DSGVO
- BDSG etc.



29 Instances:

- FU-DSO-VG
-



GENERICORP

2. Company Policies

Paragraphs

2.1. Documenting Incidents

CO-008: All technology related incidents are appropriately identified, recorded and classified.

- 2.1.1. All Incident Records must be raised and managed within the approved Incident Management toolset.
- 2.1.2. Incident Records must be raised for any issue identified within any of the in-scope IT environments (i.e. PROD, DR, TEST and DEV).
- 2.1.3. All relevant information relating to the nature and impact of the incident must be logged within the Incident Record at the point of Registration.
- 2.1.4. All Incidents must be managed in accordance with the published Incident Management Process.
- 2.1.5. All Incidents must be assigned a priority rating which is calculated based on the impact and urgency of the incident (as defined automatically by the toolset or by the end user)
- 2.1.6. Any Incidents that are identified as being triggered by Change must be linked to the relevant Change Record

2.2. Incident Resolution

CO-009: All technology related incidents are resolved promptly or escalated as necessary.

- 2.2.1. All actions taken to investigate, diagnose and resolve an Incident, together with any additional relevant information received, must be documented within the Incident Record to ensure that there is a full audit trail for future reference.
- 2.2.2. The root cause of the incident must be documented within the Incident Record, where this has been identified during the Incident lifecycle
- 2.2.3. Any Incident that is determined to be Severity 3: High or above Major Incident, must be escalated to the Major Incident Management team promptly – either directly in the toolset or via direct contact to the Major Incident Management Team. Any Incident that is determined to be Severity 3: Medium or below must be managed accordingly to the Major Incident Management sub-process
- 2.2.4. Where an incident can't be resolved at first line, it must be functionally escalated to the appropriate technical support team.
- 2.2.5. Where a change activity is implemented in order to fix an incident, the Change Record must be related to the Incident.

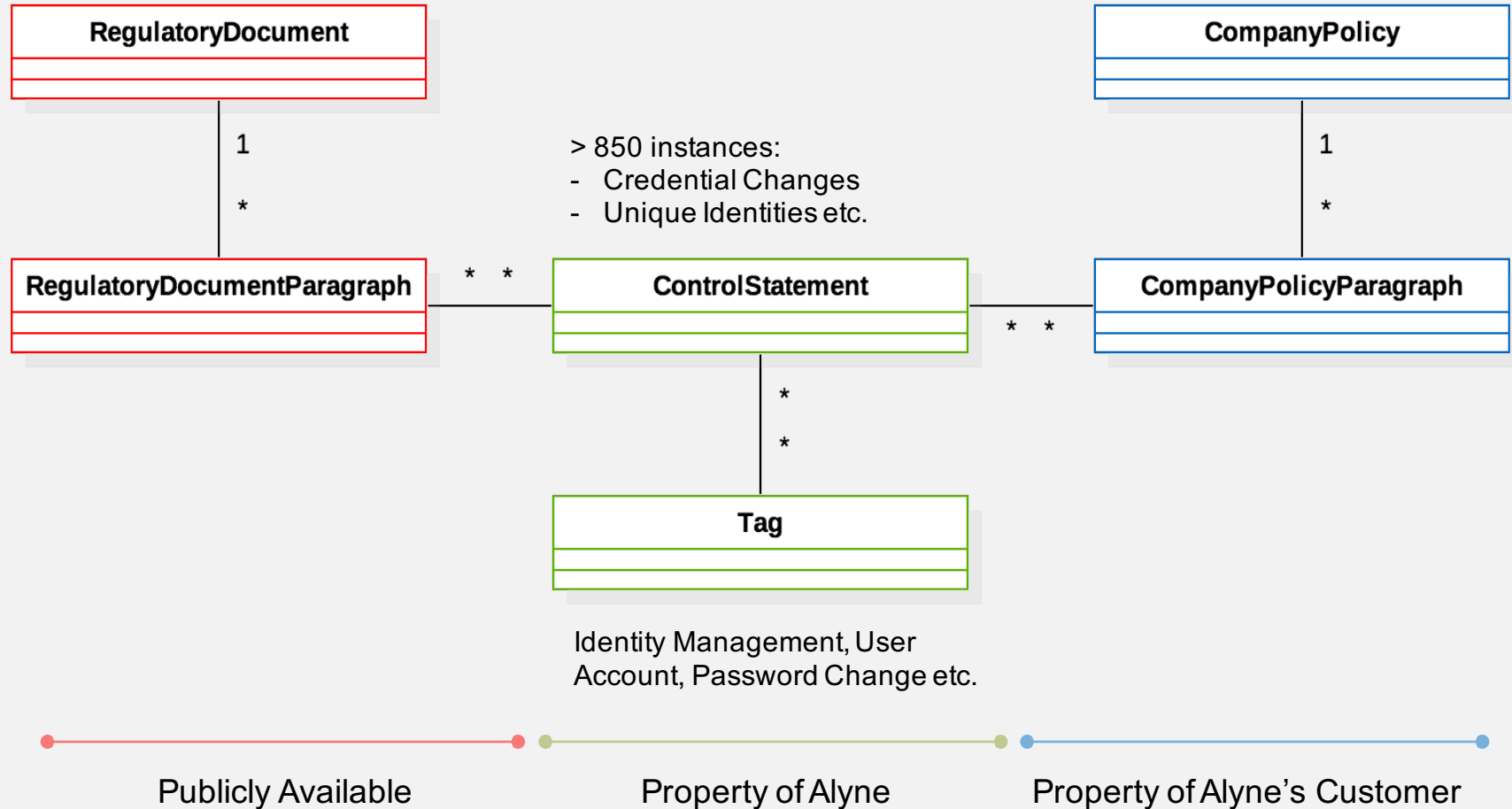
Motivation

Industry Partner Alyne GmbH – Data Model



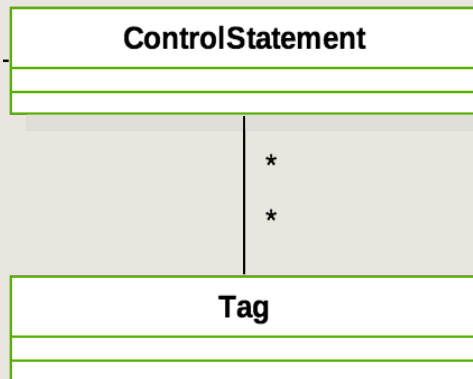
29 Instances:

- EU-DSGVO
- BDSG etc.



29 Instances:

- FU 100000
-



All archived information shall be retrievable within a timeframe without impacting dependent business processes.

Archiving

Availability

Data Lifecycle

Data

Retrieval

Data Deletion

Each release must include release notes defining the content of changes to the application, the impact on functionality for application users and potential fall back procedures.

Release Management

Software

Release Notes

Business Owner

All cryptographic keys shall be randomly generated, independently from other cryptographic keys.

Cryptography

Key Generation

Key Management

Encryption

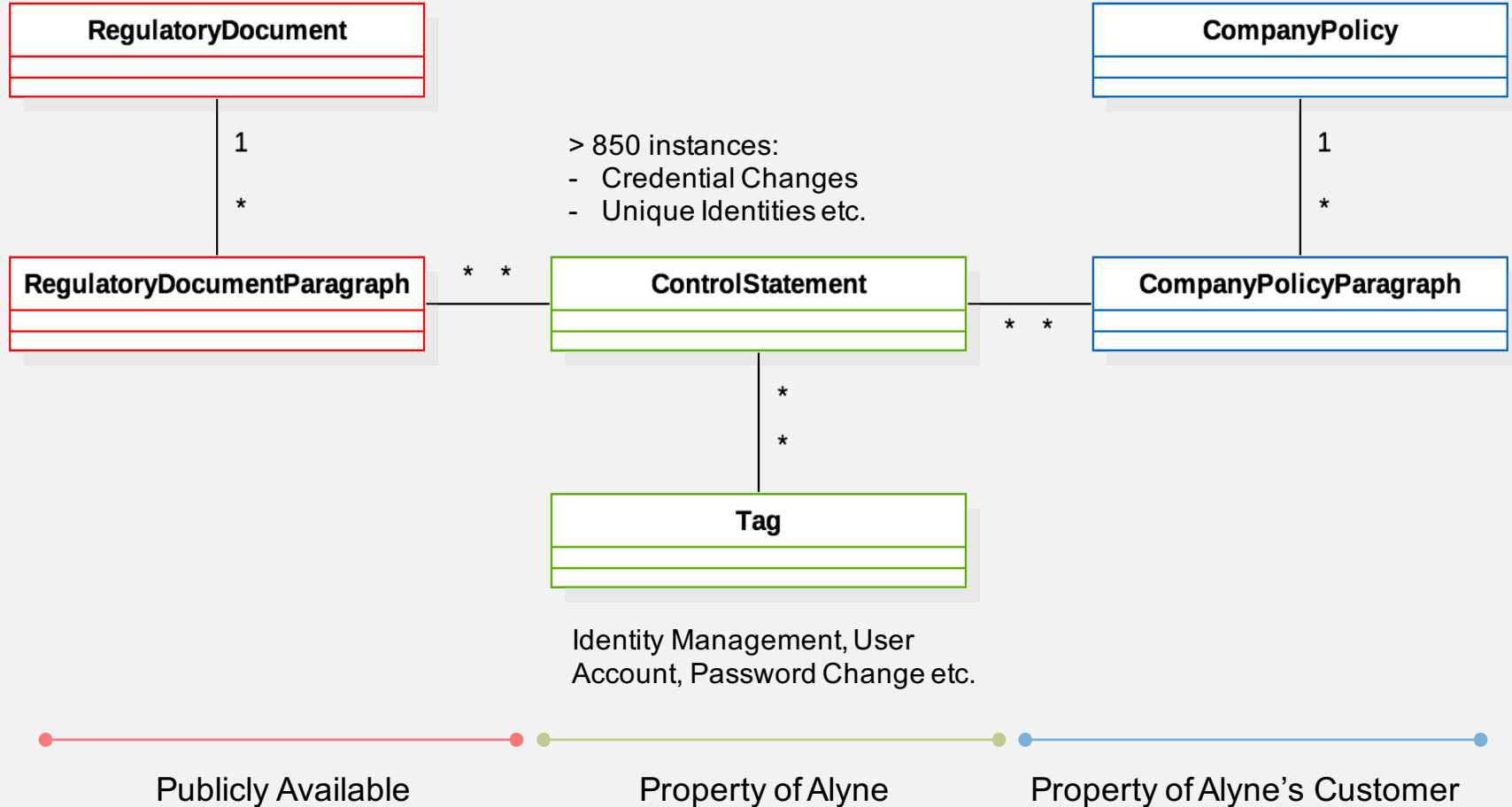
Motivation

Industry Partner Alyne GmbH – Data Model

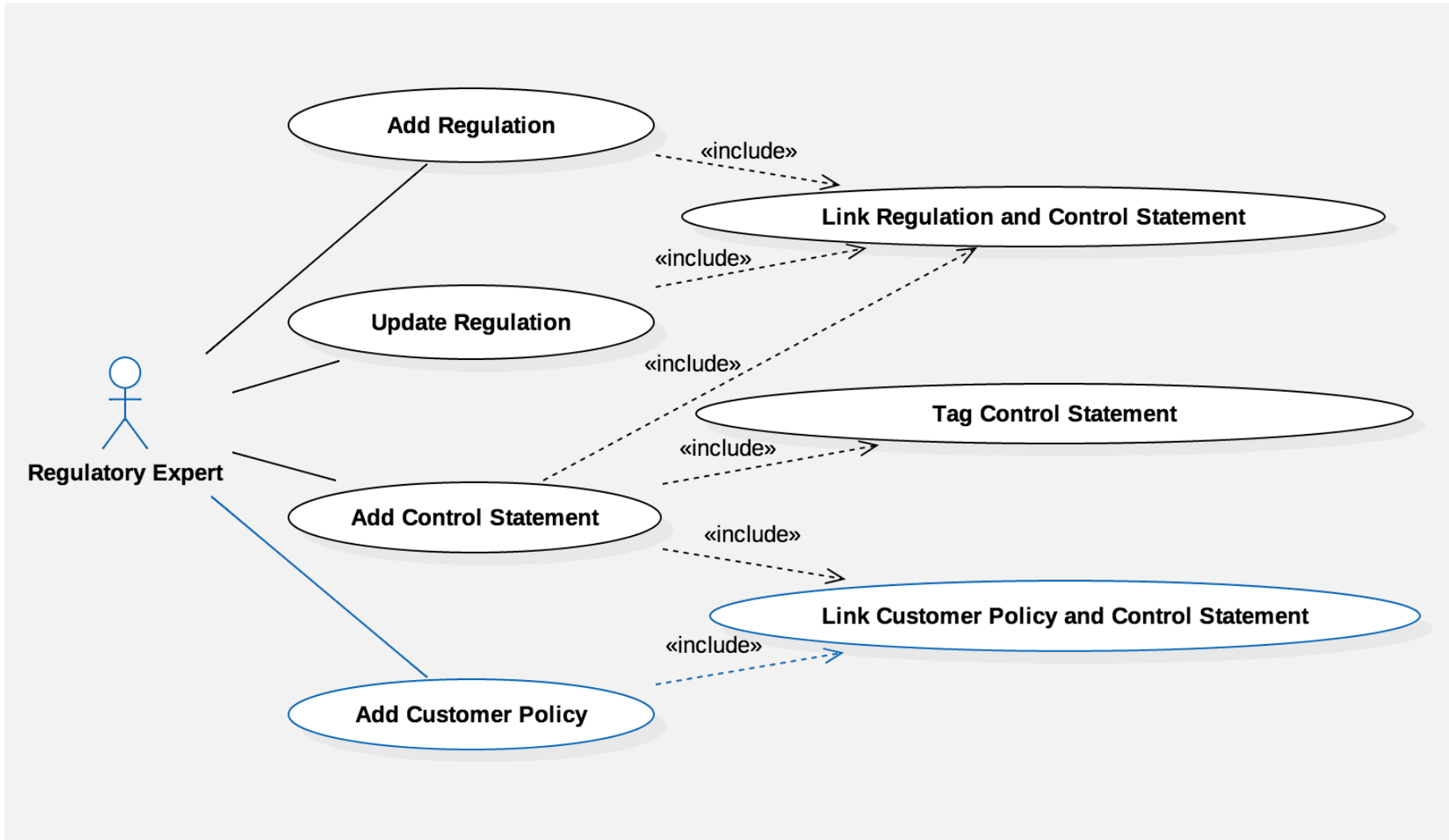


29 Instances:

- EU-DSGVO
- BDSG etc.



Linking Customer Policies and Control Statements takes the largest volume.



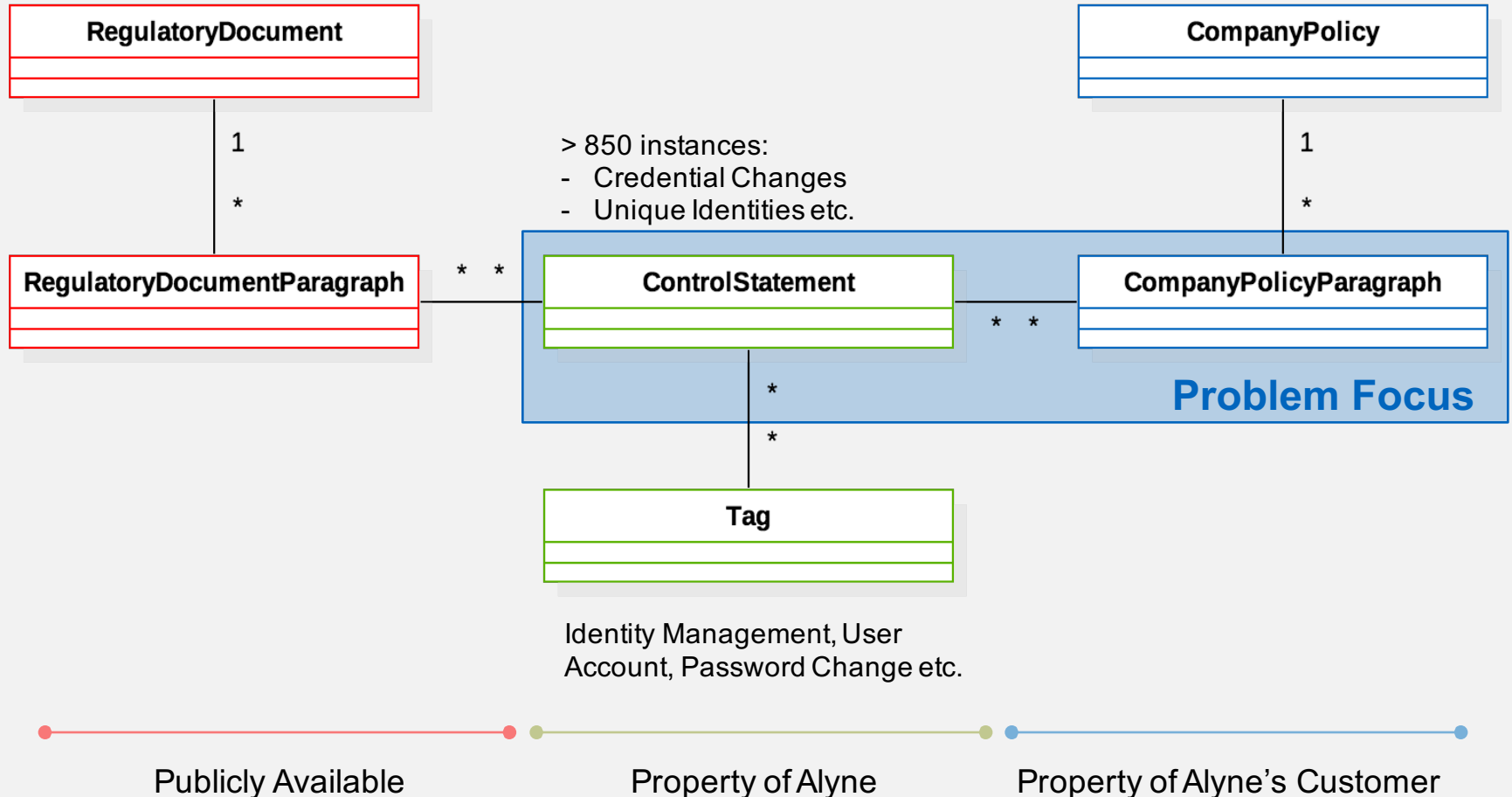
Motivation

Industry Partner Alyne GmbH – Data Model



29 Instances:

- EU-DSGVO
- BDSG etc.



Motivation

- Problem Statement
- Industry Partner Alyne GmbH

Research Questions

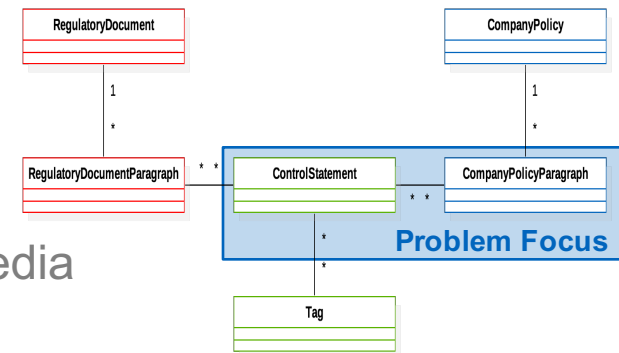
Research Approach

- Basic Concept
- Pre-Processing Technologies
- Editorial Help
- Iterative Approach

Timeline

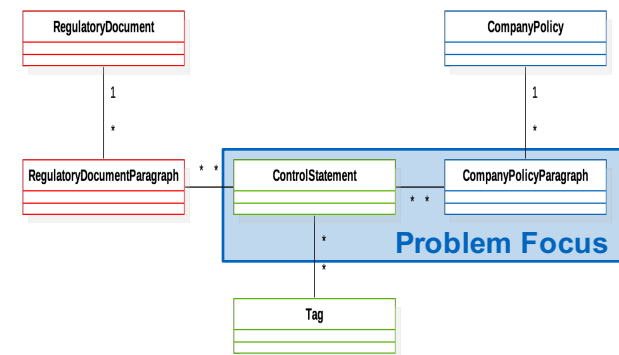
Research Questions (1)

1. How good is a text similarity approach suited to solve the semantic text matching problem?
 - Slide 19
2. What text similarity technology performs best: TF, TF-IDF or Word Embeddings?
 - Slide 19
3. What pre-process technologies are suited for this approach?
 - Slide 21
4. How much editorial help does improve this approach?
 - Slide 22
5. What are good corpora to train word embeddings for this approach?
 - Company Policies, Regulatory Documents, Wikipedia
 - Which parameters? window size, vector size etc.



Optional

6. Can chapter or paragraph context help to improve the results?
7. Can an intersection approach help to improve the precision (identify only relevant problems)?



Motivation

- Problem Statement
- Industry Partner Alyne GmbH

Research Questions

Research Approach

- Basic Concept
- Pre-Processing Technologies
- Editorial Help
- Iterative Approach

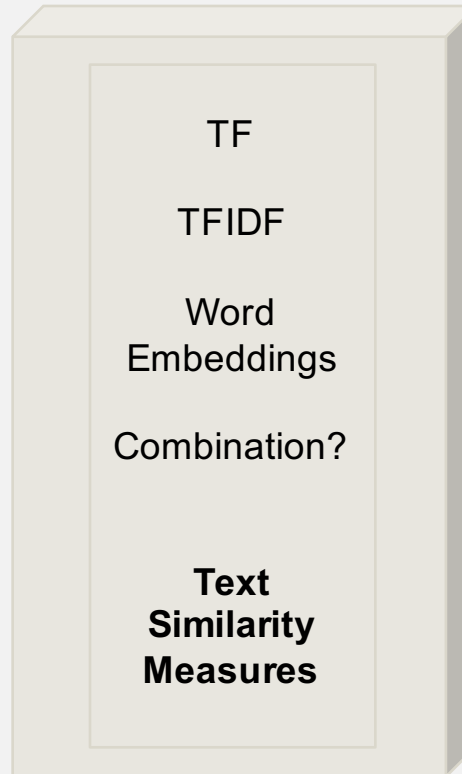
Timeline

A **recommender system** supports experts to **align control statements and company policies** using **text similarity measures**. (Semantic Text Matching Problem)

Important records shall be in accordance with statutory, regulatory, contractual, and business requirements.

Verify that all data expected for processing are processed [...] in a timely manner, and all output is delivered in accordance with business requirements.

Define and implement procedures for backup and restoration of systems [...] in line with business requirements and continuity plan.



All archived information shall be retrievable within a timeframe without impacting dependent business processes.

Recovery from backup media shall be tested at least every year.

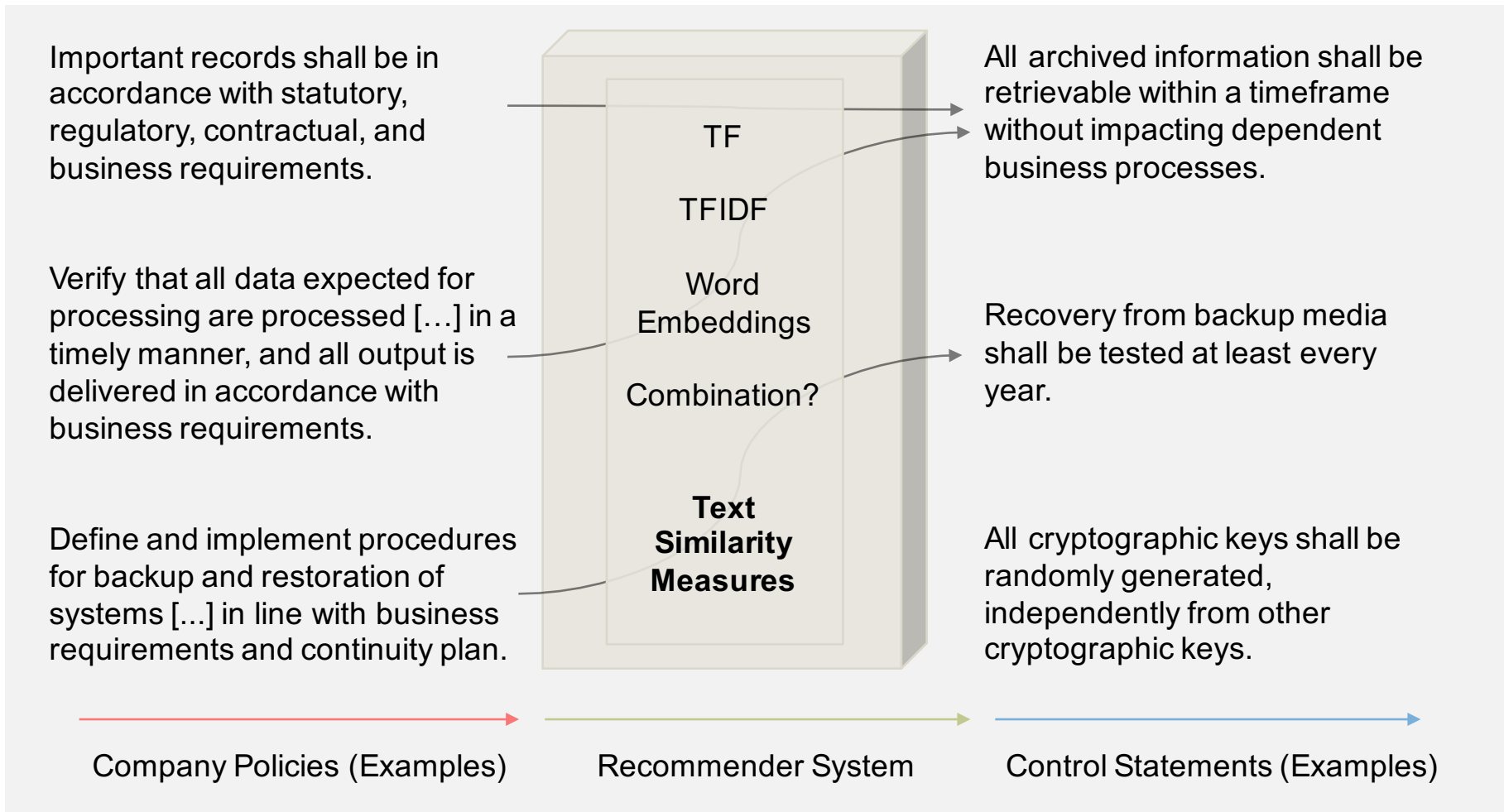
All cryptographic keys shall be randomly generated, independently from other cryptographic keys.

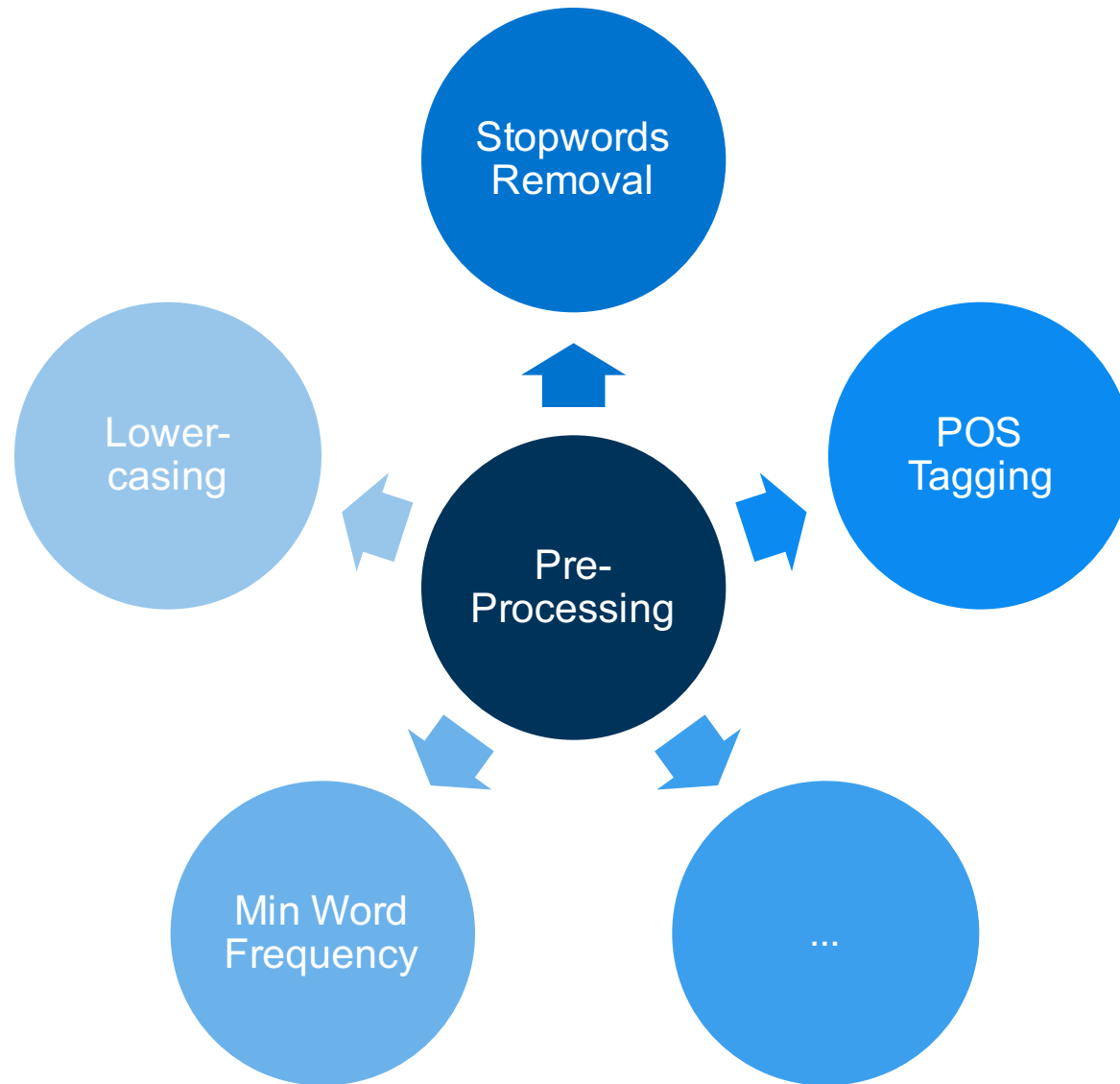
Company Policies (Examples)

Recommender System

Control Statements (Examples)

A **recommender system** supports experts to **align control statements and company policies** using **text similarity measures**. (Semantic Text Matching Problem)





- Control Statement +

Title

Subtitle

Tags

or

...

Beispiel:

All cryptographic keys shall be randomly generated, independently from other cryptographic keys.

Cryptography

Key Generation

Key Management

Encryption

Transform

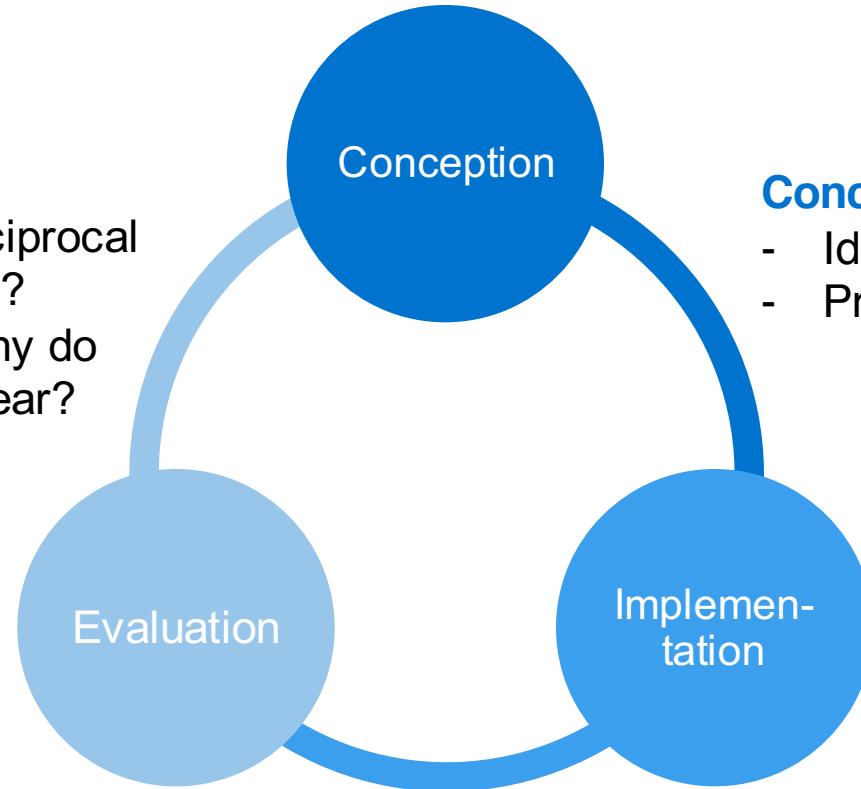


All cryptographic keys shall be randomly generated, independently from other cryptographic keys.

**Cryptography Key Generation
Key Management Encryption.**

Evaluation

- Measure: Mean Reciprocal Rank (MRR), others?
- Error Inspection: Why do False Positives appear?



Conception

- Identify Data Source
- Prepare Sandbox

Implementation

- TF, TF-IDF, word embeddings
- Editorial Help
- Java based toolkit DeepLearning4J

Motivation

- Problem Statement
- Industry Partner Alyne GmbH

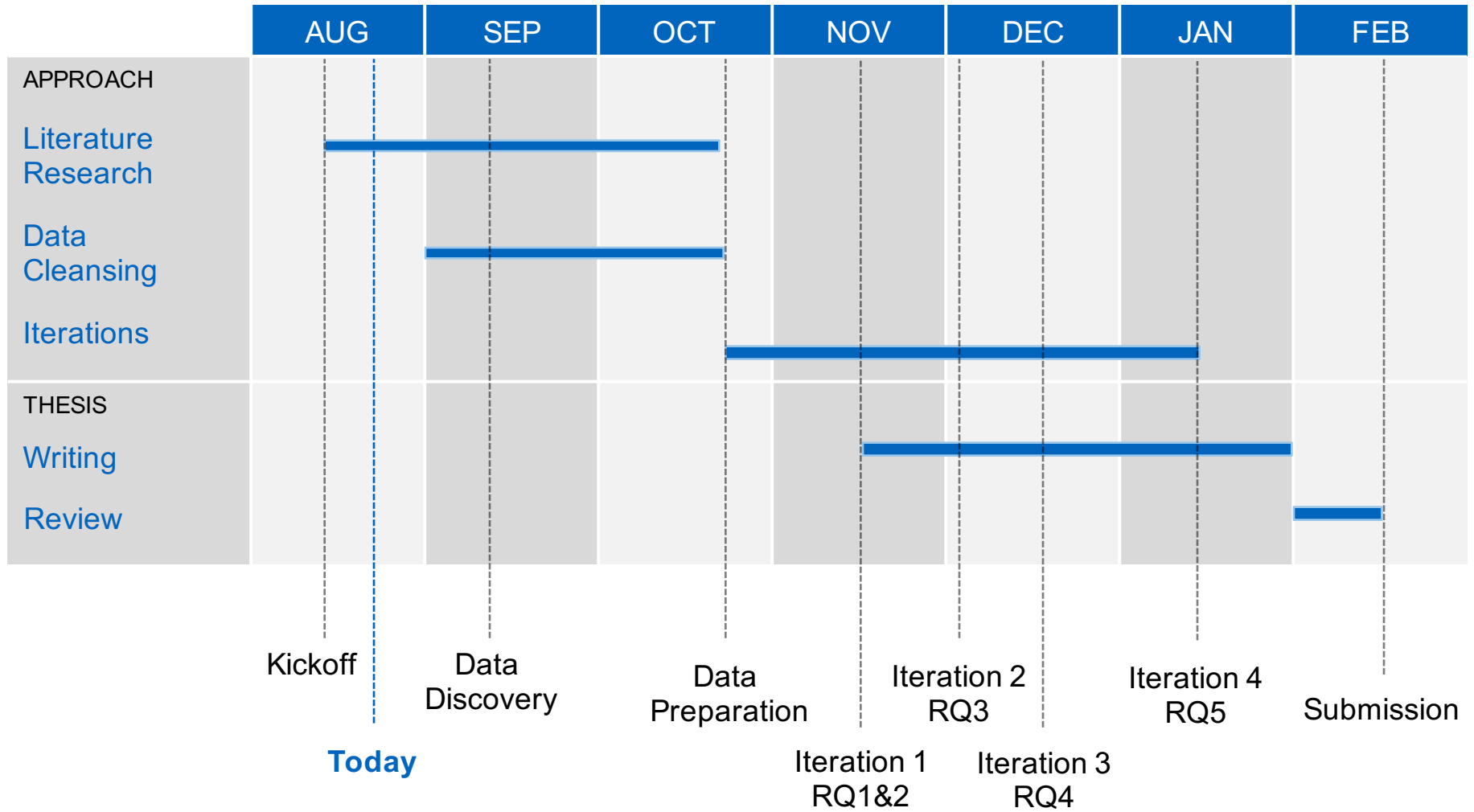
Research Approach

- Basic Concept
- Pre-Processing Technologies
- Editorial Help
- Iterative Approach

Research Questions

Timeline

Timeline





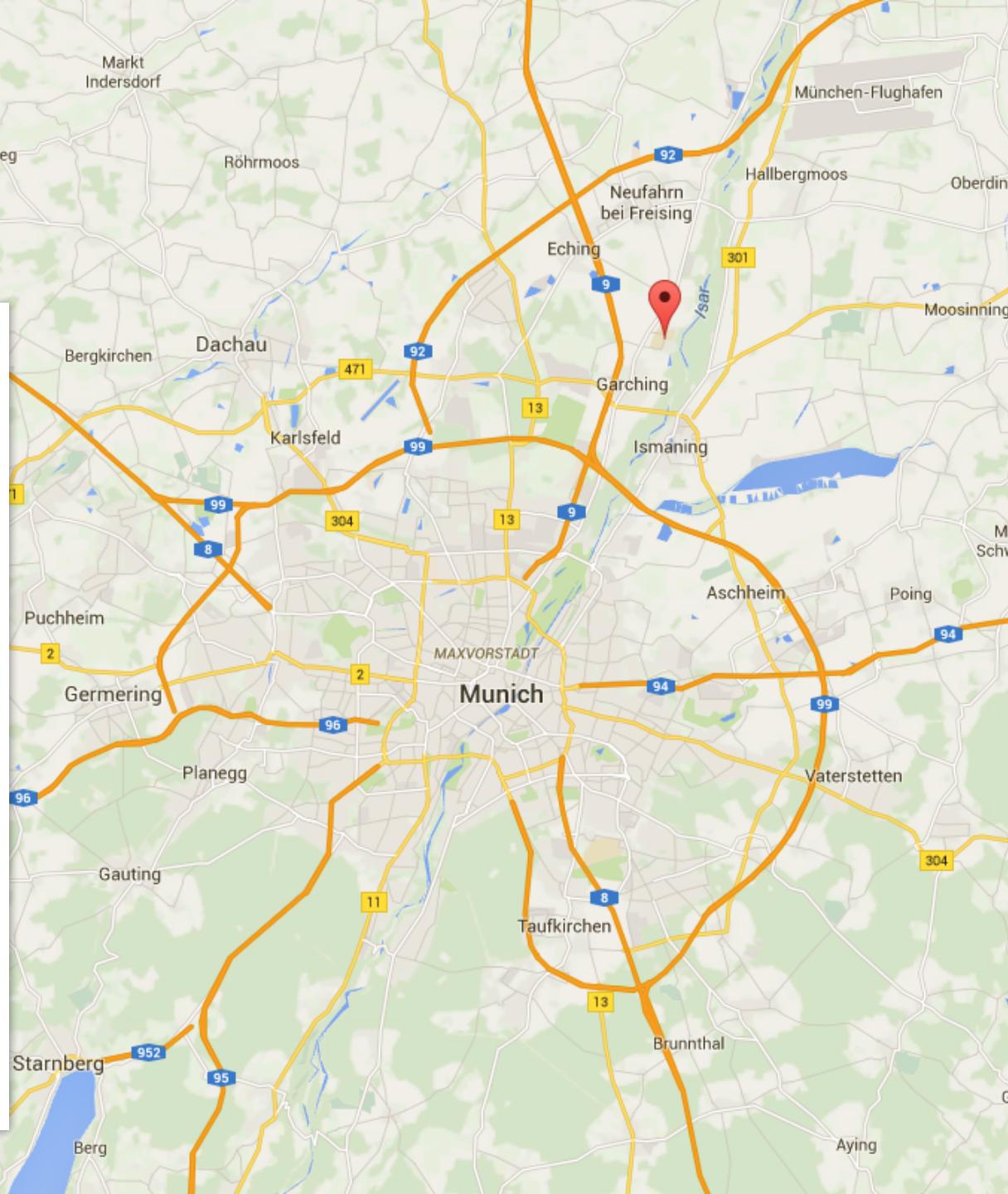
B. Sc.

Christoph Erl

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

christoph.erl@tum.de
www.matthes.in.tum.de



Backup Slides

Research Approach

Iterative Approach

Evaluation

- Measure: Mean Reciprocal Rank (MRR), others?
- Error Inspection: Why do False Positives appear?

Discovery

- Business Domain RegTech
- Data Source: MongoDB Collection Exports (JSON)

Data Preparation

Data Preparation

- Prepare Sandbox:
Load & Survey:
(Control Statements
+ Editorial Help
+ Pre-Processing)
- Clean & Normalize

Model Building

- (Parallelisation/Optimisation, not topic in thesis)

Model Planning

- TF, TF-IDF, word embeddings
- Java based toolkit DeepLearning4J

