



From the ML Model to Practice

Case Study on NLP-based Decision-Making on the Eligibility of Security Prospectuses



Maximilian König
AI Solution Architect



Bernd Rusitschka
AI Expert in DG
Markets



Janek Blankenburg
AI Application Engineer



Philipp Rothhaar
Expert in DG Markets

In collaboration with further colleagues from DG Markets and Prof. Christian Hänig and Serhii Hamotskyi from Anhalt University of Applied Sciences

Agenda



Status quo ante Deciding the Eligibility of Securities' Prospectuses

Training a model Proof of Concept with a fine-tuned model

Integration ... of the model into the business process

Learnings ... from the process



Status Quo Ante

Deciding the Eligibility of Securities' Prospectuses

Status Quo Ante

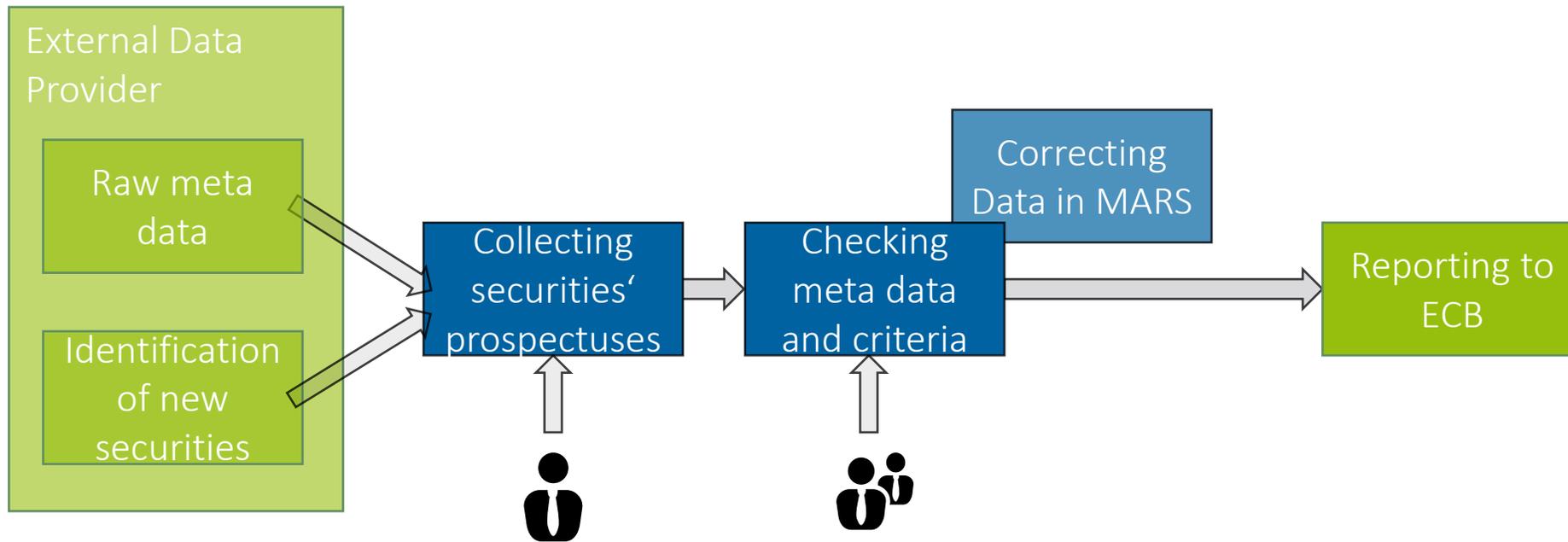
Deciding the Eligibility of Securities Prospectuses



- NCBs report daily new **eligible marketable assets** to ECB, which collects them into **EADB (Eligible Assets Database)**
- Checking a security / asset for eligibility is based on harmonized criteria (Guideline (EU) 2015/510)
- The reporting contains the **eligible assets** as well as related **meta data**
- Several eligibility criteria are established based on a security's prospectus
 - So far this is achieved by manually checking / reading the prospectuses in a four-eyes principle
- At Deutsche Bundesbank (BBk) the (BBk-made) application MARS is used for collecting the securities' data and reporting them to ECB

Status Quo Ante

Process Flow



Manual assessment is time-consuming and repetitive

Training a model

Proof of concept with a fine-tuned model

Research Project Automatic Annotation

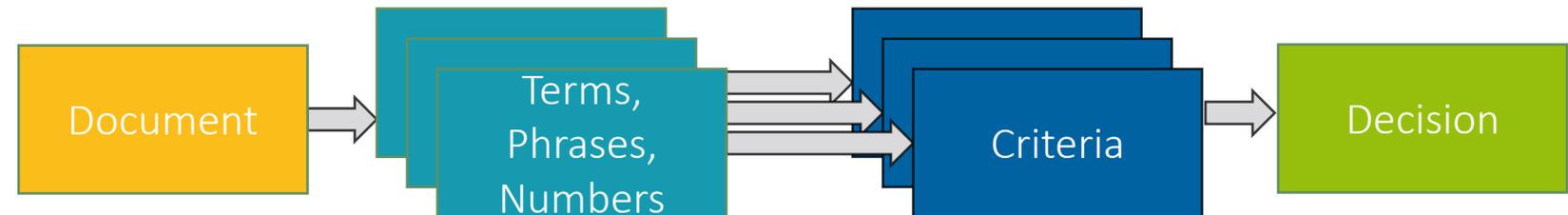
Proof of Concept using NLP

§ 5 (Status)

Die Schuldverschreibungen begründen nicht besicherte und nicht nachrangige Verbindlichkeiten der Emittentin. Bei Emission handelt es sich bei den Schuldverschreibungen um bevorrechtigte Schuldtitel (Senior Preferred Schuldverschreibungen), die nicht den durch § 46f Absatz 5 in Verbindung mit Absatz 6 KWG gesetzlich bestimmten niedrigeren Rang haben.

x status_nicht_nachrangig_bevorrechtigt

- Task at hand: Identifying in PDF-Documents a given number of terms, phrases, numbers etc. that form the basis for the decision

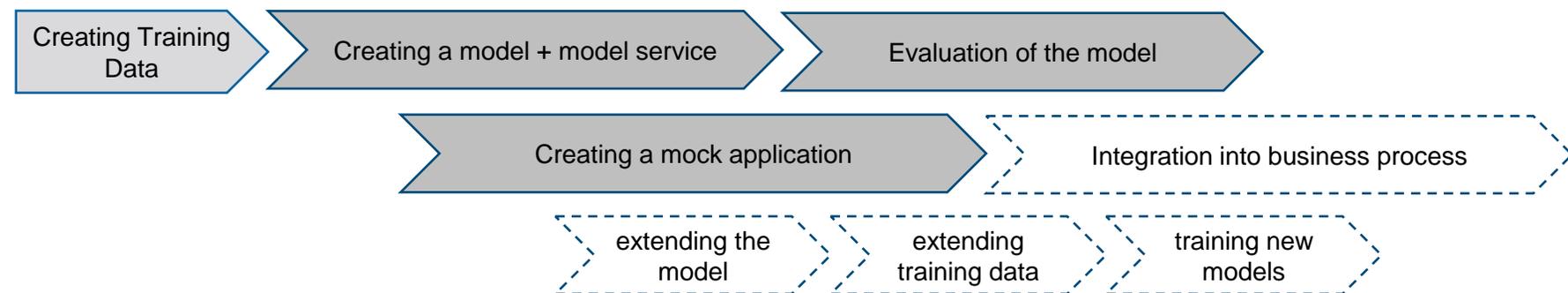


- In ML terms: Multiclass/multilabel Classification Task (≈ 20 categories)

Research Project Automatic Annotation

Starting Point

- At the start of the project (early 2022, pre „GPT breakthrough“):
 - No German-language domain specific (i.e. financial) language model available
 - Hence 2-Step modelling process:
 - (1) Fine-tuning a language model for German financial documents
 - (2) Training a multilabel classifier on top of the language model
 - No public dataset available -> creating training data is the first step



Creating Training Data

Data collection

In case of Notes listed on the Official List of the Luxembourg Stock Exchange and traded on the Regulated Market "Bourse de Luxembourg" or publicly offered in the Grand Duchy of Luxembourg, the Final Terms will be published in electronic form on the website of the Luxembourg Stock Exchange (www.lse.lu). Furthermore, the aforementioned Final Terms will be published in electronic form on the website of DZ BANK AG (www.dzbank.de), in case of Notes listed on any other stock exchange or traded on any other regulated market or publicly offered in one or more member states of the European Economic Area (excluding the Grand Duchy of Luxembourg), the Final Terms will be published in electronic form on the website of DZ BANK AG (www.dzbank.de).

15 February 2022
15. Februar 2022

Final Terms
Endgültige Bedingungen

EUR 10,000,000 Multi-currency Fixed Rate Preferred Senior Notes of 2022(2027) (the "Notes")
EUR 10,000,000 mehrfach-kündbare festverzinsliche befristete Schuldverschreibungen (die "Schuldverschreibungen") von 2022(2027) (die "Schuldverschreibungen")

Issued pursuant to the begehren aufgrund des

DZ BANK AG
Sonderliche Bedingungen
Direct Issuance Programme
dated 4. June 2021
datiert 4. Juni 2021

of
der

DZ BANK AG
Deutsche Zentral-Gesamtsbank AG, Frankfurt am Main
LEI: 529600RCAH1G0L0007
(having its registered office at Platz der Republik, 60325 Frankfurt am Main, Federal Republic of Germany)
(mit eingetragenem Sitz in Platz der Republik, 60325 Frankfurt am Main, Bundesrepublik Deutschland)

Issue Price: 100 per cent during the subscription period
from 18 February 2022 to 18 February 2022 (each date including).
The selling price of the Notes is free to trade after the expiry of the subscription period.

Ausgabepreis: 100 % während der Zeichnungsfrist
vom 18. Februar 2022 bis 18. Februar 2022 (jeweils einschließlich)
Nach Ablauf der Zeichnungsfrist ist der Verkauf der Schuldverschreibungen freibestehend.

Issue Date: 17 February 2022
Valutierungstag: 17. Februar 2022
Series No.: A1702
Serien-Nr.: A1702

Number of prospectus: 413
Issuing period: 2021 - 2022
Eligible documents: 369
Ineligible documents: 44
Training set: 272
Test set: 141 + 141

Data annotation

PART I TERMS AND CONDITIONS
TEIL I ANLEHBERINGANGEN

This PART I of these Final Terms is to be read in conjunction with the A1 Terms and Conditions of Fixed Rate Preferred Senior Notes (the "Terms and Conditions") set forth in the Prospectus. Capitalised Terms not otherwise defined in this PART I of these Final Terms shall have the same meaning as specified in the Terms and Conditions.

Dieser TEIL I dieser Endgültigen Bedingungen ist in Verbindung mit dem A1-Anleihebedingungen für festverzinsliche befristete mehrfach-kündbare Schuldverschreibungen (die "Anleihebedingungen") zu lesen, die im Prospekt enthalten sind. Begriffe, die in diesem TEIL I dieser Endgültigen Bedingungen nicht definiert sind, haben die gleiche Bedeutung, wie sie in den Anleihebedingungen festgelegt sind.

Abhängig von dem TEIL I dieser Endgültigen Bedingungen sind sub-paragraphen in die paragraphen und sub-paragraphen der Terms and Conditions.

Abhängig von dem TEIL I dieser Endgültigen Bedingungen sind Paragraphen und Absätze zusammenzufassen auf die Paragraphen und Absätze der Anleihebedingungen.

The provisions in this PART I of these Final Terms and the Terms and Conditions, taken together, shall constitute the terms and conditions applicable to the Terms of Notes (the "Conditions").
Die Angaben in diesem TEIL I dieser Endgültigen Bedingungen zusammengelesen mit den Bestimmungen der Anleihebedingungen bilden die für die Festsätze von Schuldverschreibungen anwendbaren Bedingungen (die "Bedingungen").

Language of Conditions
Sprachliche Bedingungen

These Final Terms are drawn up in German and English (German text controlling and binding).
Diese Endgültigen Bedingungen sind in Deutsch und Englisch (deutscher Text maßgeblich und bindend).

§1 CURRENCY / DENOMINATION / FORM / DEFINITIONS
§1 WÄHRUNG / STÜCKELUNG / FORM / DEFINITIONEN

Sub-paragraph (A)
Absatz (A)

Principal Senior Notes
Hauptverschreibungen

Currency and Denomination
Währung und Stückelung

Duration
Dauer

Aggregated Principal Amount
Gesamtnennbetrag

Specified Denomination/Principal Amount
Festgelegte Stückelung/Nennbetrag

Sub-paragraph (B)
Absatz (B)

Partnership (Global Issuer)
Gesellschaft (Gesamtschuldner)

Sub-paragraph (C)
Absatz (C)

Clearing System
Clearing-System

Clearing System
Clearing-System

Clearstream Banking AG

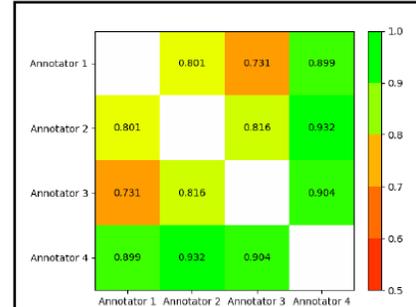
Annotation tool: Konfuzio
Annotation types: ~40

Disregarding of pages without annotations for training and validation purposes.

Annotation statistics

Target type	Train	Test
coupon fixed	431	375
coupon variable index	56	84
coupon variable margin	38	42
coupon variable operator	37	43
coupon variable tenor	45	75
currency	514	577
early redemption amount	64	52
early redemption	177	108
isin	421	417
principal amount	784	800
redemption at maturity amount	26	42
redemption at maturity	370	347
special termination	96	109
special termination amount	61	63
status non preferred	56	47
status senior non preferred	488	333
type of instrument	431	422

Inter-annotator agreement



Average inter-annotator agreement scores

Test set was used to measure IAA. Therefore, every prospectus in the test set was annotated by a second analyst. 4 analysts served as annotators in total.

Data preprocessing

1st step: Extraction of JSON-formatted raw data containing the annotations from the annotation tool
2nd step: Conversion and transformation of extracted data into dataset for token classification (BIO encoding)

Endgültige Bedingungen
Final Terms

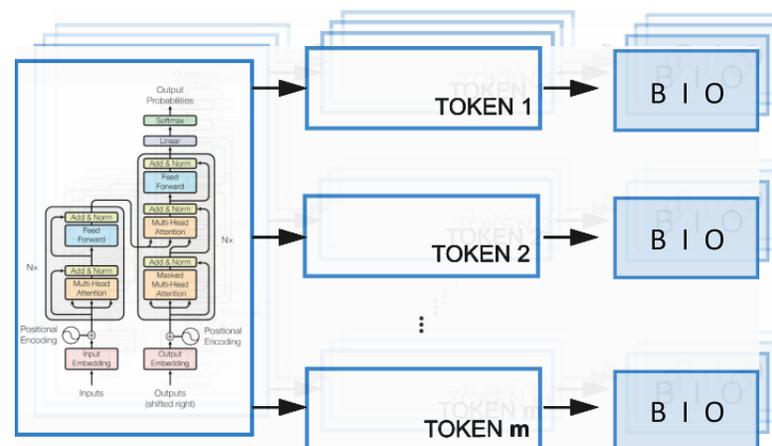
EUR 10,000,000 einfach-kündbare 0,35% Schuldverschreibungen fällig am 30. Mai 2031 (die "Notes")

Principal amount	B	I	O	O	O	O	O	O	O	O	O	O	O
Type of instrument	O	O	O	O	O	O	B	O	O	O	O	O	O
Currency	B	O	O	O	O	O	O	O	O	O	O	O	O

- Implementation of dataset classes using Hugging Face Datasets framework
- Challenge: overlapping text sequences belonging to different annotation types

1. Conversion PDF -> Text (including OCR)
2. Text processing and clean-up (e.g. extraction of German parts of bilingual docs, analysis of textboxes, ...)
3. *Embedding (Text to vectors) using fine tuned language model*
4. *Labelling with multilabel classifier*
5. Decision based on deterministic rules (derived from EU Guideline)

3 + 4

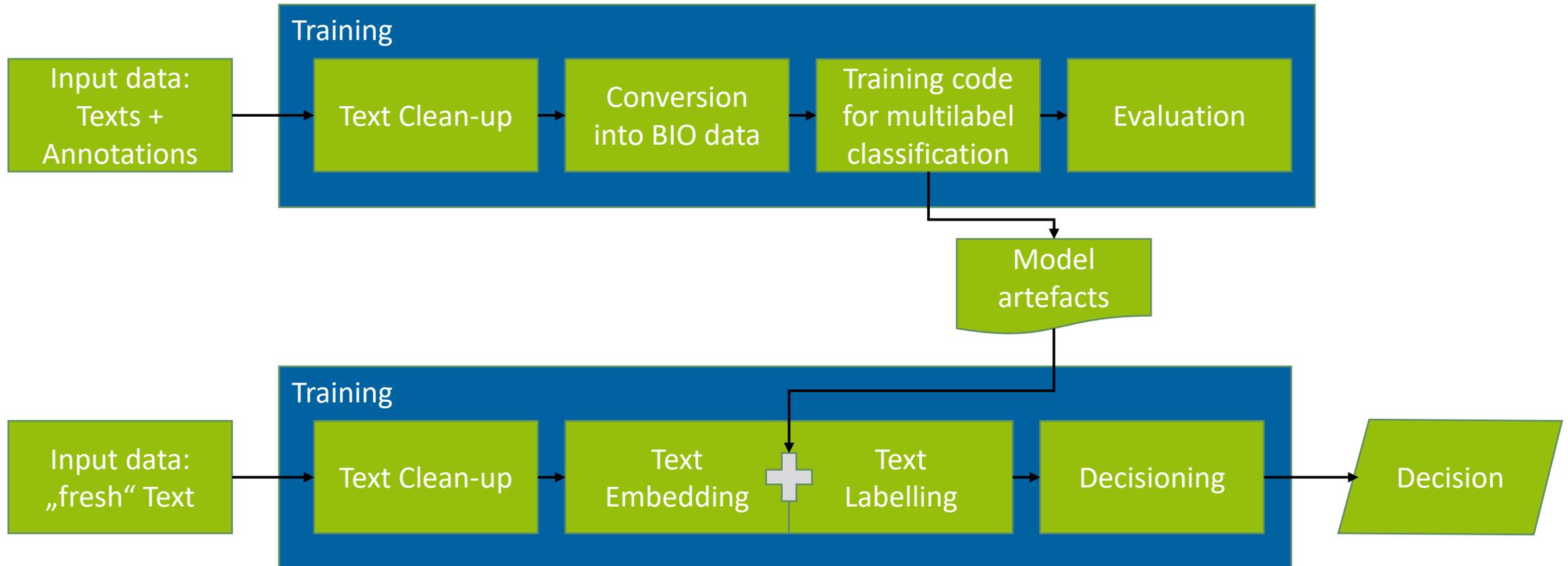


Integration

of the model into the business process

Operating the Model

Model Training and Decisioning



Integration into Business Process

Process Needs

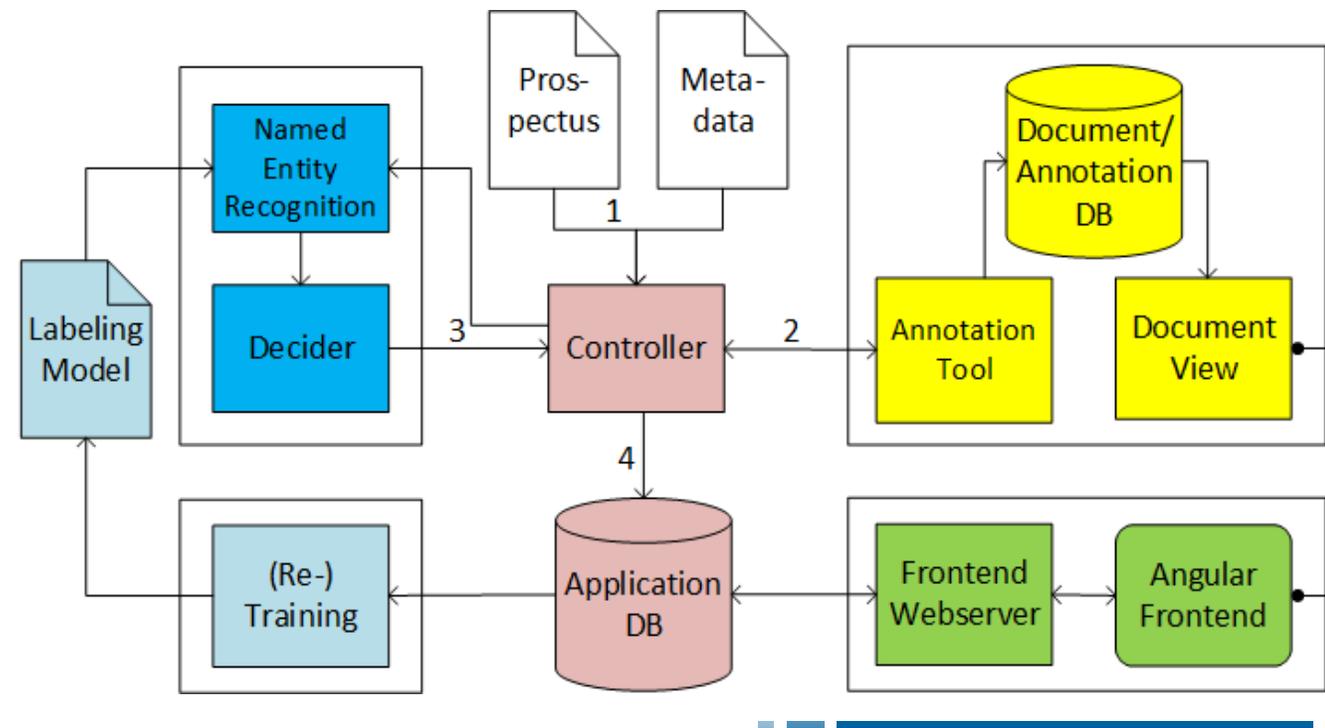


- Given a document the experts needs:
 - a. the decision of the model,
 - b. the criteria causing that decision and
 - c. (optimally) the relevant passages in the document (or relevant meta data) to
- check the validity of the ML decision.
- If the model makes a mistake, the expert needs to **overwrite that decision** and
- (optimally) collect the data for future model improvements (retraining)
- If retraining is undertaken, we need both valid as well as invalid model decisions.

Integration into Business Process

Overview of Application Architecture

- Containerized application with communication via REST
- Integration into the actual business application (MARS) open as of yet



Implications of Using ML in the Process



- Using an ML model can reduce processing time by replacing manual reading with reviewing found passages



- An ML model will always have a chance for error, but the **accuracy can reach the Inter Annotator Agreement (IAA)** at the least



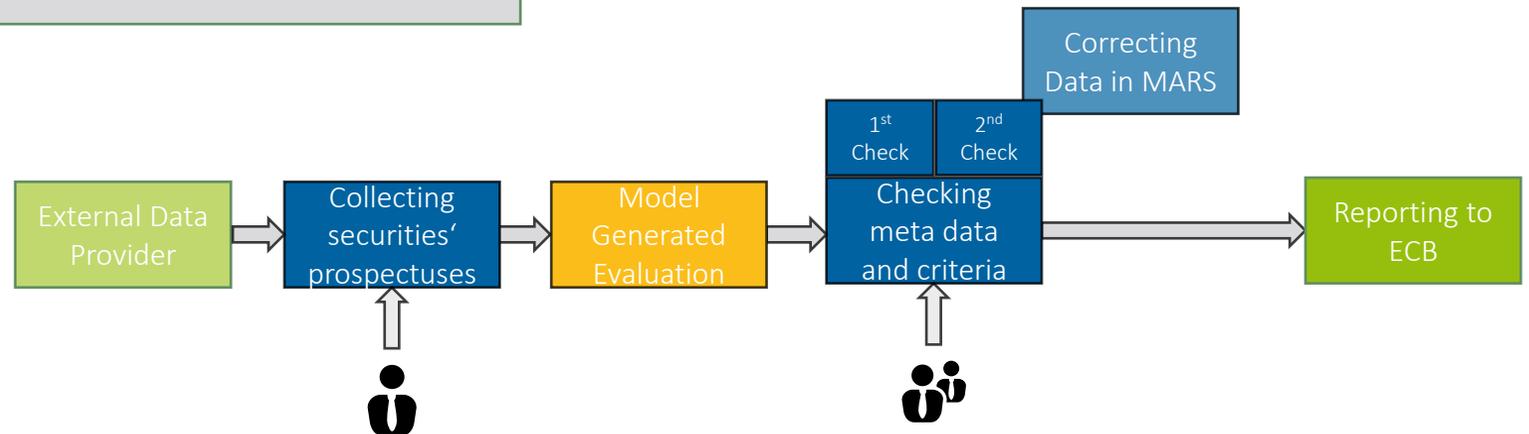
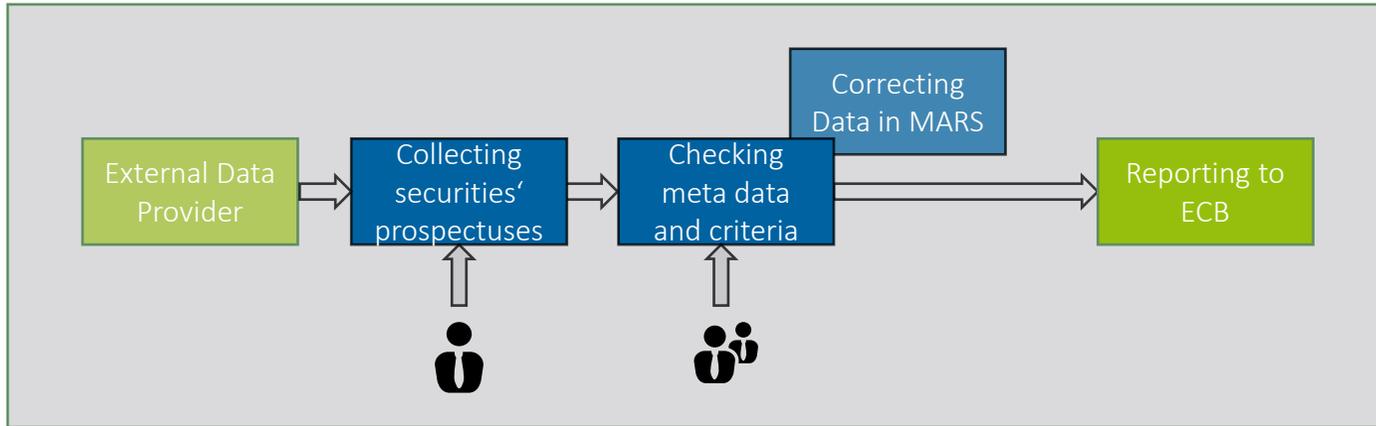
- Current legal environment requires a „**human in the loop**“
 - If model accuracy is (acceptably) high, the four eyes principle (as well as the review by two experts) could be replaced by a simple review
(2 pairs of human eyes \Rightarrow „AI eyes“ + 1 pair of human eyes)



- Using an ML model will require:
 - **continuous monitoring** of model performance
 - **continuous improvement** of model mistakes and training data

Evolving the „4 Eyes Principle“

New Process Flow – Proof of Concept



External Interface

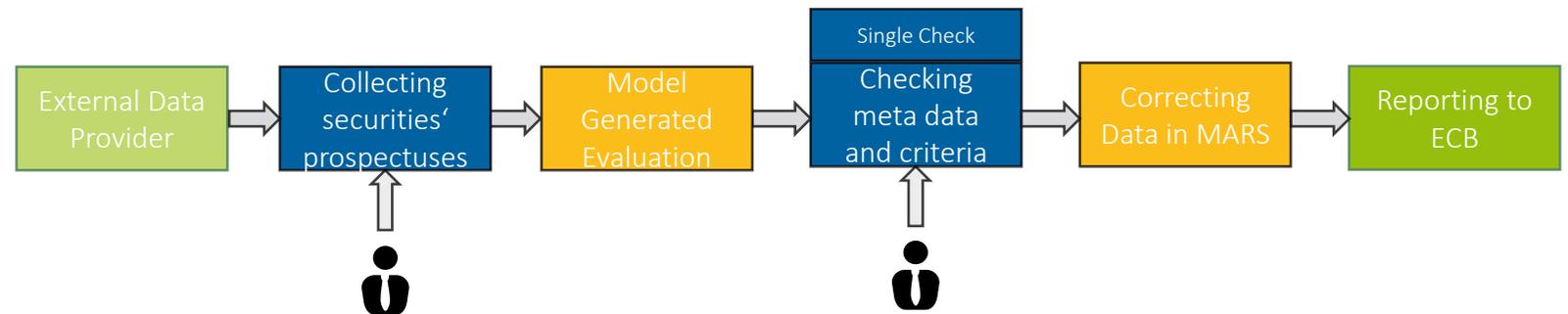
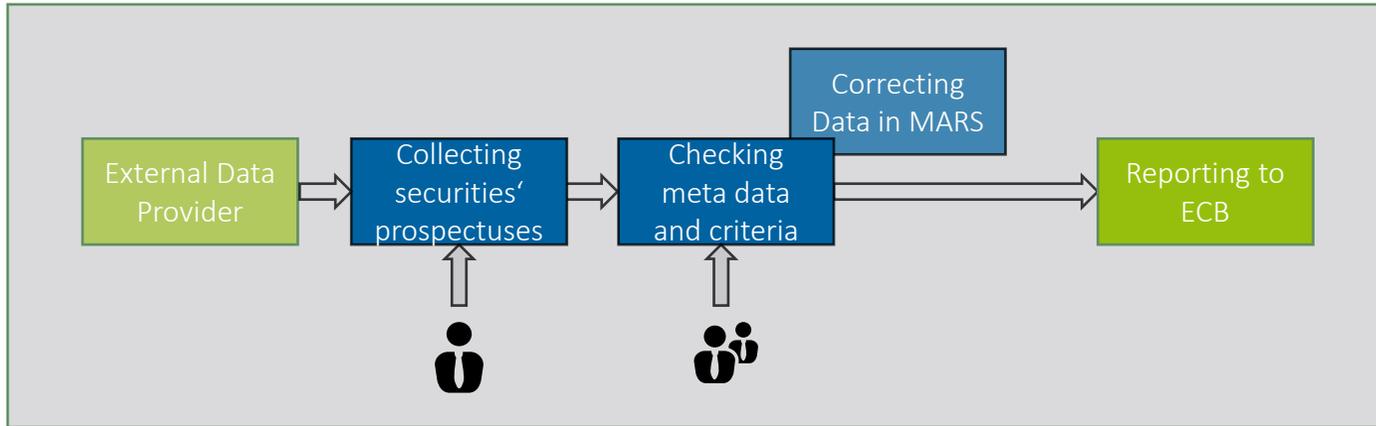
Human Interaction

Automatic Process



Evolving the „4 Eyes Principle“

New Process Flow – 1st Evolution



External Interface

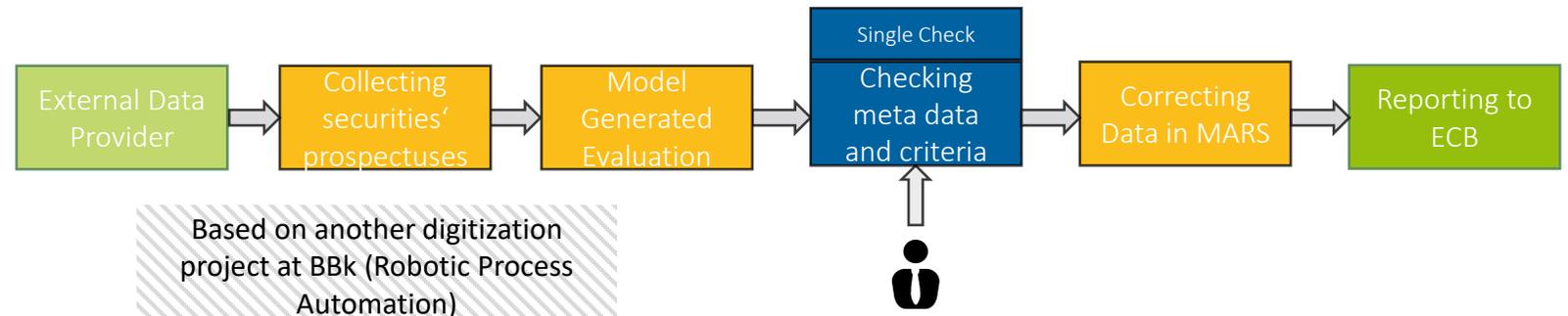
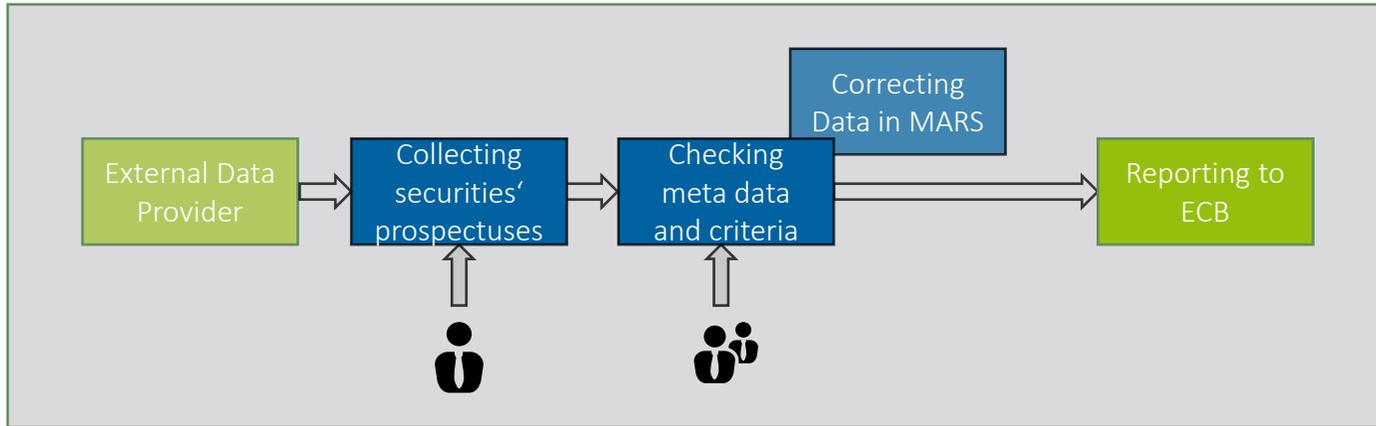
Human Interaction

Automatic Process



Evolving the „4 Eyes Principle“

New Process Flow – 2nd Evolution



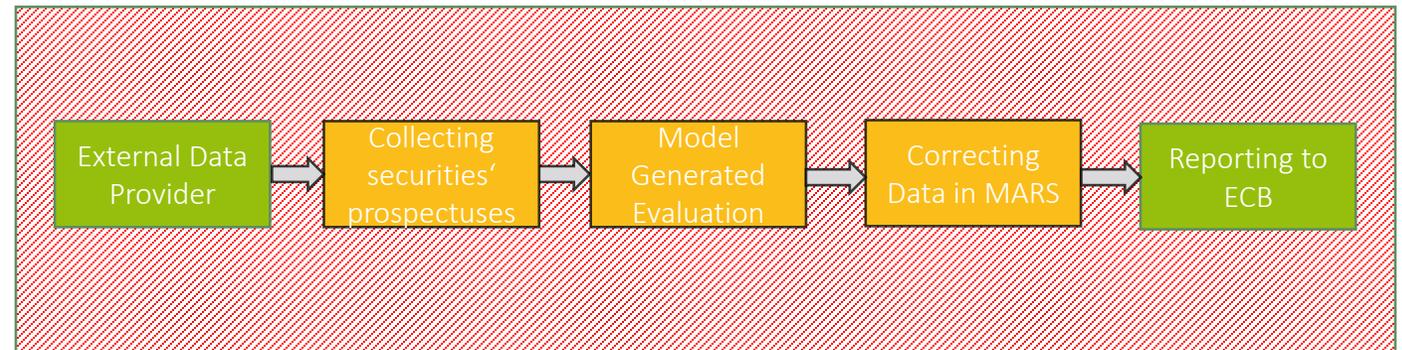
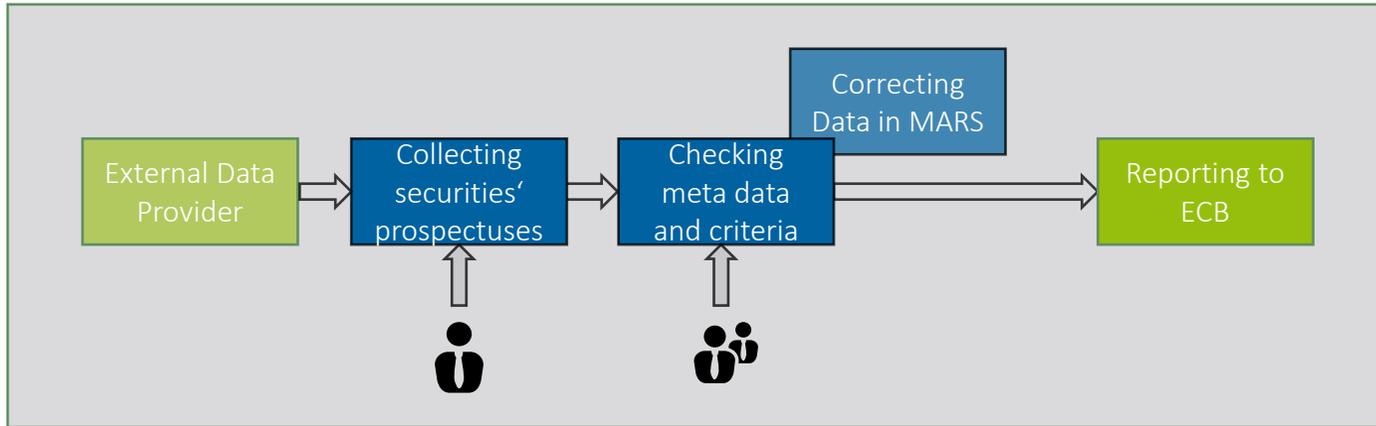
External Interface

Human Interaction

Automatic Process

Evolving the „4 Eyes Principle“

Currently not Possible: Fully Automated Process – No Human in the Loop



External Interface

Human Interaction

Automatic Process



Learnings

from the Process

Learnings from the Project



- Creating training data is highly costly



- Understanding the business process is key
 - if only part of the process is automated, **the benefit may not outweigh the complexity**



- Building the necessary environment is highly complex
 - The codebase of the proof of concept easily reaches **10'000 lines of code**



- **Integration into production is hard**, in particular if it necessitates new components, e.g.
 - Application for creating and storing text annotations
 - ML model monitoring and model archives (MLOps)
 - GPUs for model training

Questions?

SCAI

Service and Community Center for Artificial Intelligence

scai@bundesbank.de