ConFoo.CA
DEVELOPER CONFERENCE

Replace complex business logic with Machine Learning models

Michał Żarnecki

CompanyHouse

ConFoo.CA
DEVELOPER CONFERENCE

Technical leader at CompanyHouse AG

- Big data / Data mining / NLP / ML
- Python / PHP
- Real-time processing of trade register data
- Credit reports, owner information
- Enterprise structure

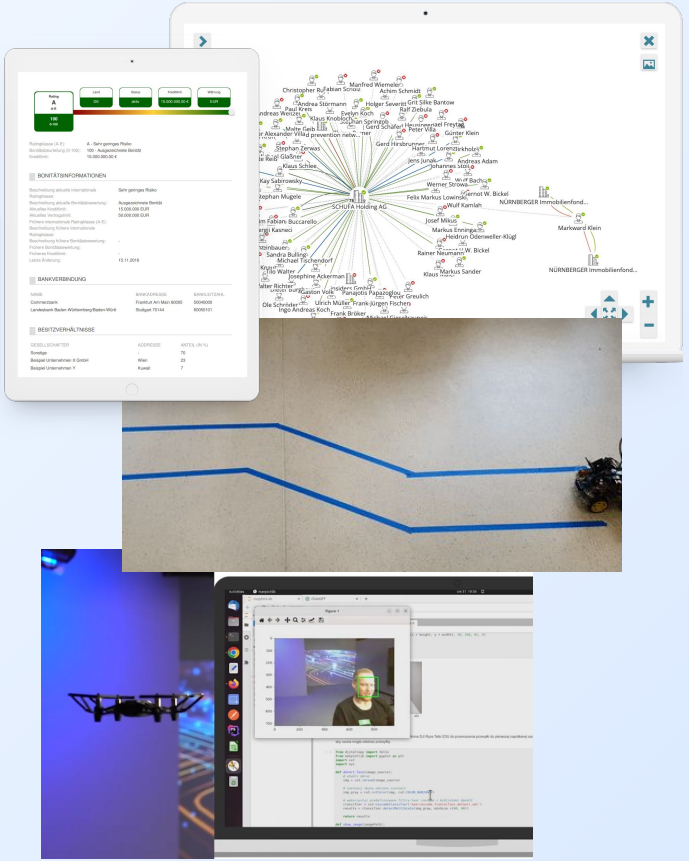
Michał Żarnecki

COLLEGIUM DA VINCI

UCZELNIA LUDZI CIEKAWYCH

Lecturer, Department of Computer Science and Data Analysis

- Modules:
  "Machine learning in Python", "Text data mining", "Gen AI with LLMs"
- E-learning course:
  "Machine Learning - How to use the potential of data, to get better results and make smarter decisions"



GitHub

https://github.com/mzarnecki

Medium

https://medium.com/@brightcode


ConFoo.CA
DEVELOPER CONFERENCE
”
The real voyage of discovery consists
not in seeking new landscapes,
but in having new eyes.
Marcel Proust

Plan

What is Machine learning?

Why we need to replace logic with ML?

Where we can use ML in our projects?

How to pick correct ML algorithm?

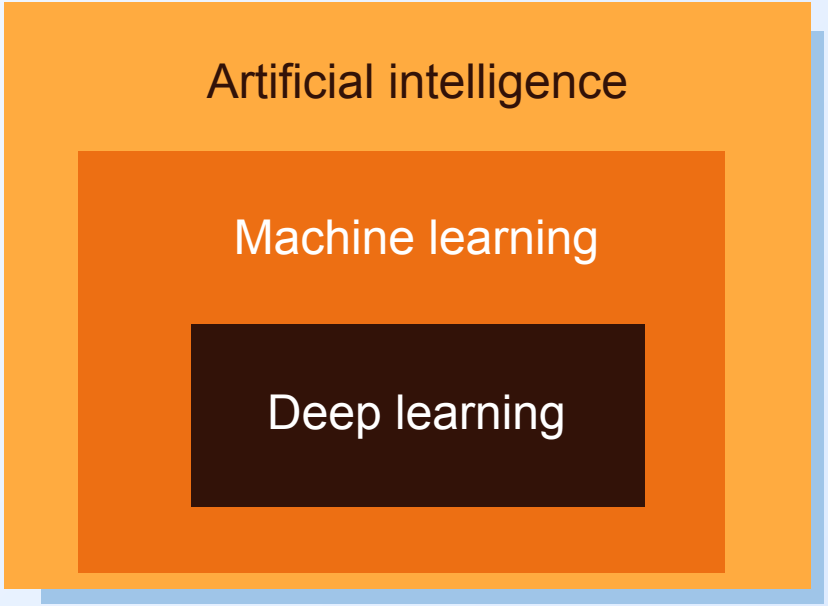
Project 1 - Codebase expert agent

Project 2 - Similar companies finder

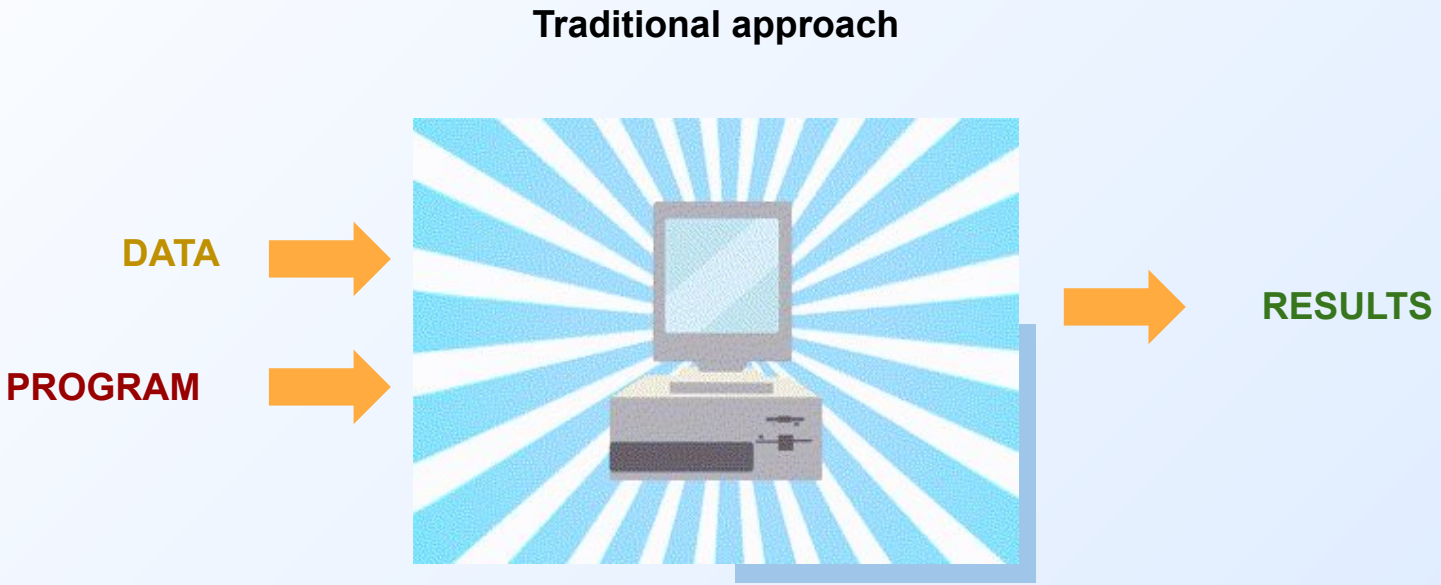
Project 3 - Documents parser

Machine learning

- self-improving programs
- a subset of artificial intelligence
- includes the field of deep learning



ML vs "traditional" programming



ML vs traditional" programming

Using machine learning

DATA

RESULTS
(PARTIAL, EXAMPLES)

PROGRAM

ML vs traditional" programming

Using machine learning

1

DATA

RESULTS
(PARTIAL, EXAMPLES)

PROGRAM

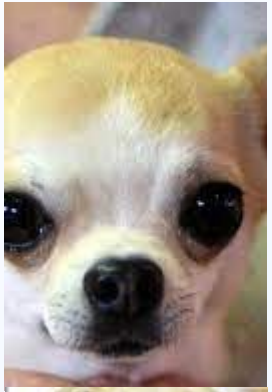
2

DATA

PROGRAM

RESULTS

What is the goal of machine learning?



ML MODEL

(logic)

dog

ML MODEL

(logic)

not a dog

Why we need to replace logic with ML?

ConFoo.CA
DEVELOPER CONFERENCE

- 🍝 Avoid endless if-else spaghetti
- ✏️ Skip hardcoding every single scenario
- 📚 Avoid learning all possible scenarios
- ⏱️ Spam less time debugging
- 💡 Focus on innovation, not maintenance

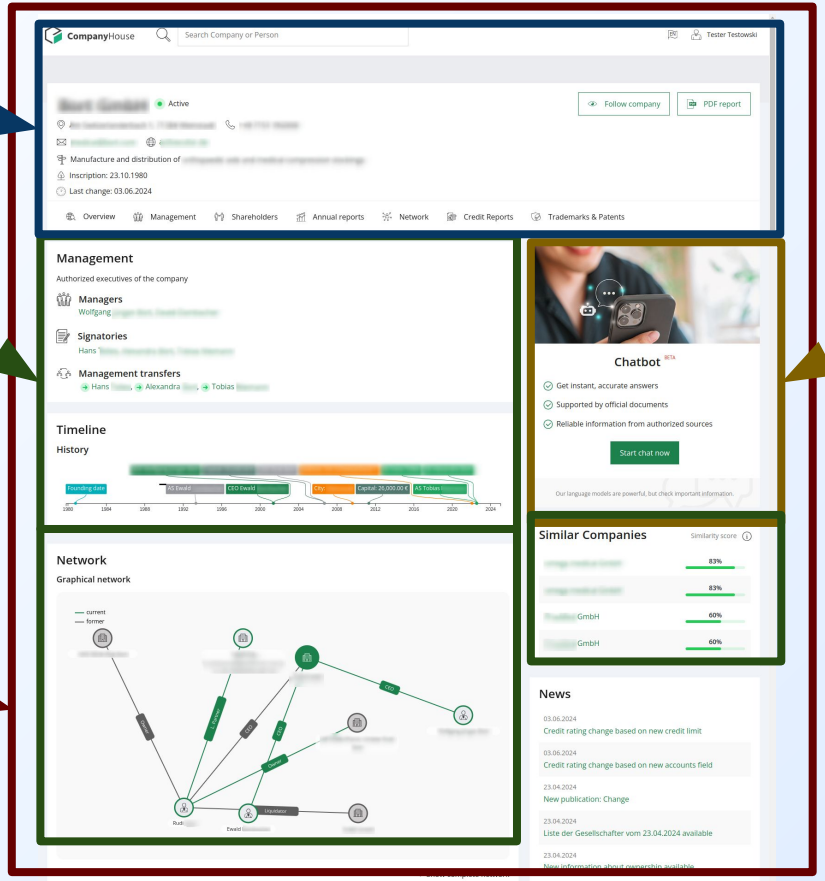
What can be supported with ML?

ConFoo.CA
DEVELOPER CONFERENCE

🌐 Localization

🎯 Business logic
🗂️ Document processing

💻 Code review
🔍 Bug detection/tracking
📊 Requirements analysis



🖥️ Customer service
📂 Knowledge management
📈 Forecasting

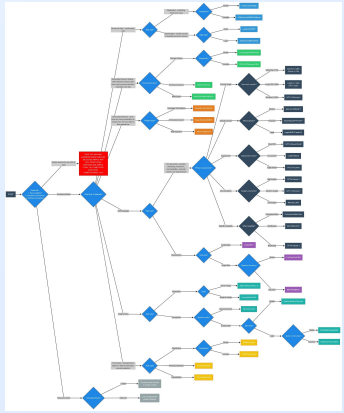
How to pick correct ML algorithm?

ConFoo.CA
DEVELOPER CONFERENCE

Hugging Face Models 1,399,438

1. What type of data or task are you working with?
2. What is the specific goal?
3. How complex is the problem?
4. What is the size of your dataset?
5. What are the performance requirements?

https://medium.com/p/0648ab1e482f

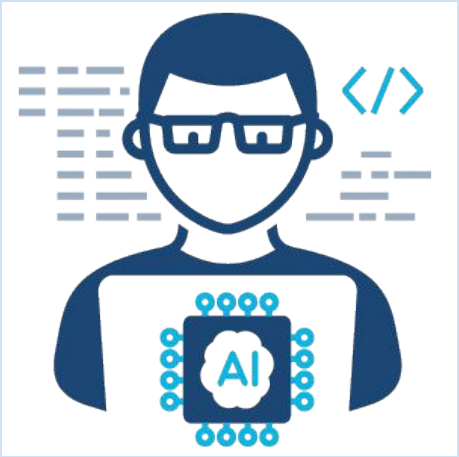


Codebase expert agent

ConFoo.CA
DEVELOPER CONFERENCE

1

Project 1
Codebase expert agent



generate code with AI

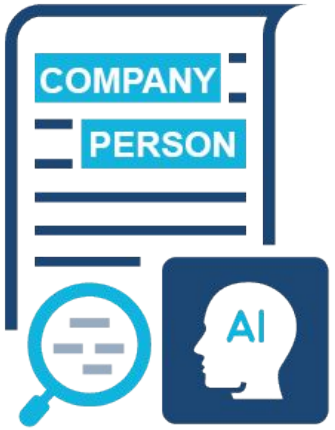
2

Project 2
Similar company finder

use LLMs

3

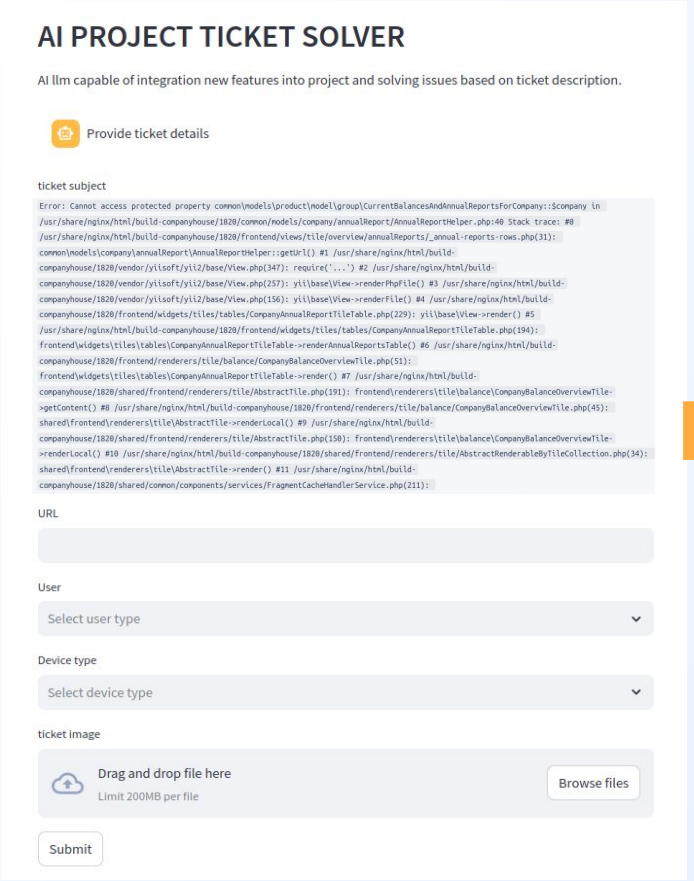
Project 3
Documents parser



train dedicated model

Codebase expert agent

ConFoo.CA
DEVELOPER CONFERENCE



Solution

1. **Identify the Class and Property:**
   - Locate the `CurrentBalancesAndAnnualReportsForCompany` class in the `common/models/product/model/group` directory.
   - Check if the `company` property is defined as protected.
2. **Implement Getter Method:**
   - If the `company` property is protected, create a public getter method within the `CurrentBalancesAndAnnualReportsForCompany` class to access it.
3. **Modify the Code to Use the Getter:**
   - Update the code in `AnnualReportHelper.php` to use the getter method instead of directly accessing the property.

Code Implementation

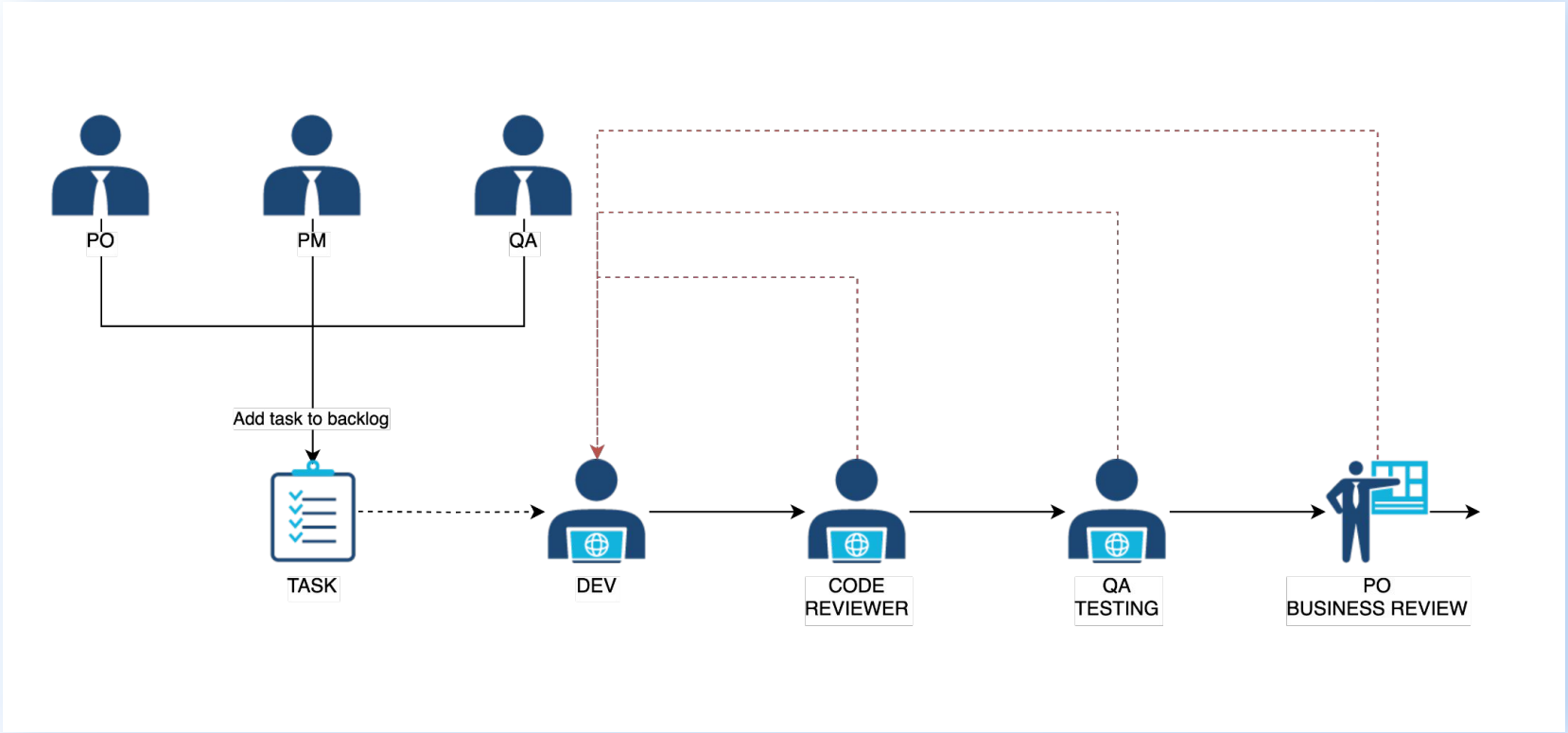
Step 1: Add Getter Method in `CurrentBalancesAndAnnualReportsForCompany` Class

```php
namespace common\models\product\model\group;

class CurrentBalancesAndAnnualReportsForCompany
{
    protected $company;

    // Add a public getter method
    public function getCompany()
    {
        return $this->company;
    }
}
```

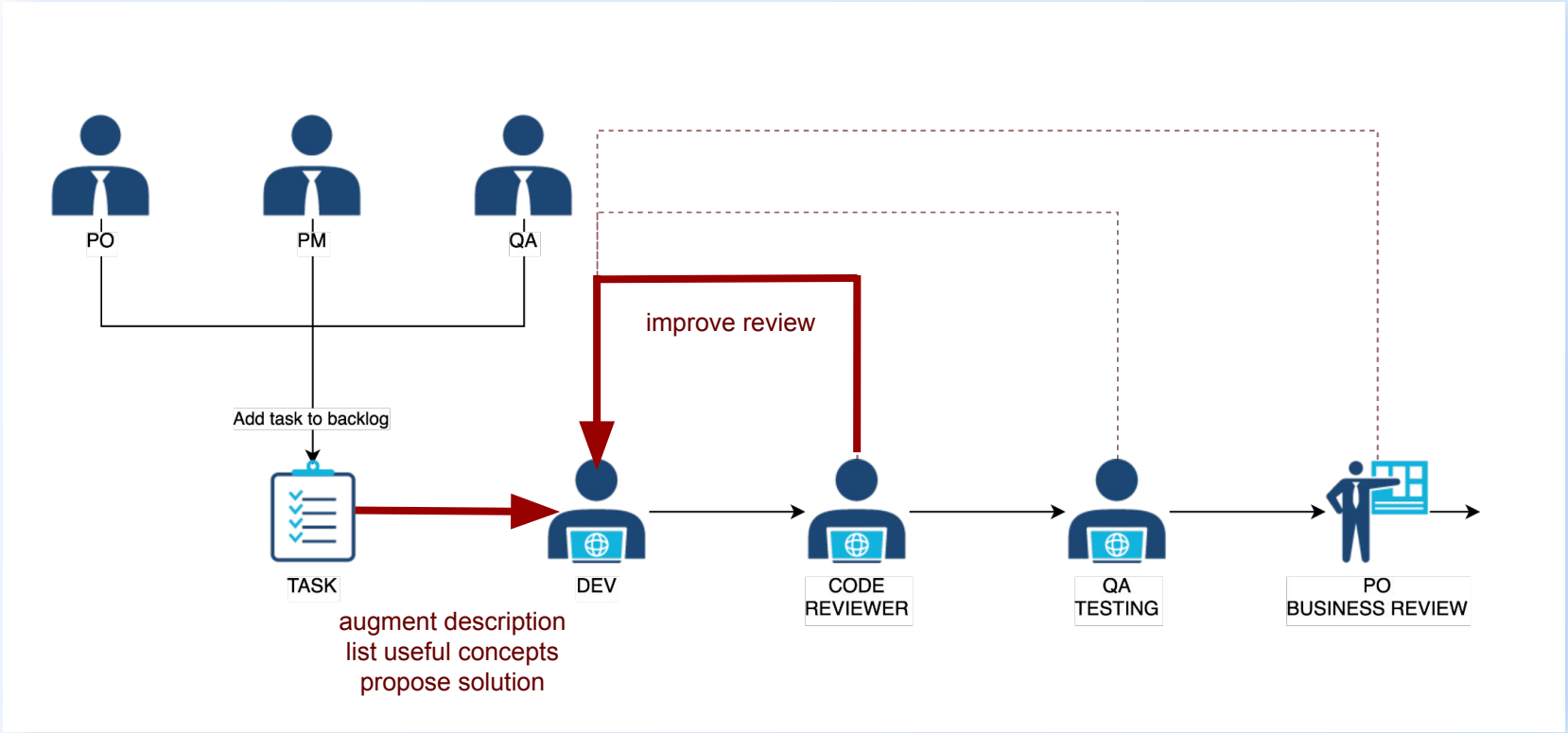
Step 2: Update `AnnualReportHelper.php` to Use the Getter

```php
namespace common\models\company\annualReport;

use common\models\product\model\group\CurrentBalancesAndAnnualReportsForCompany;

```

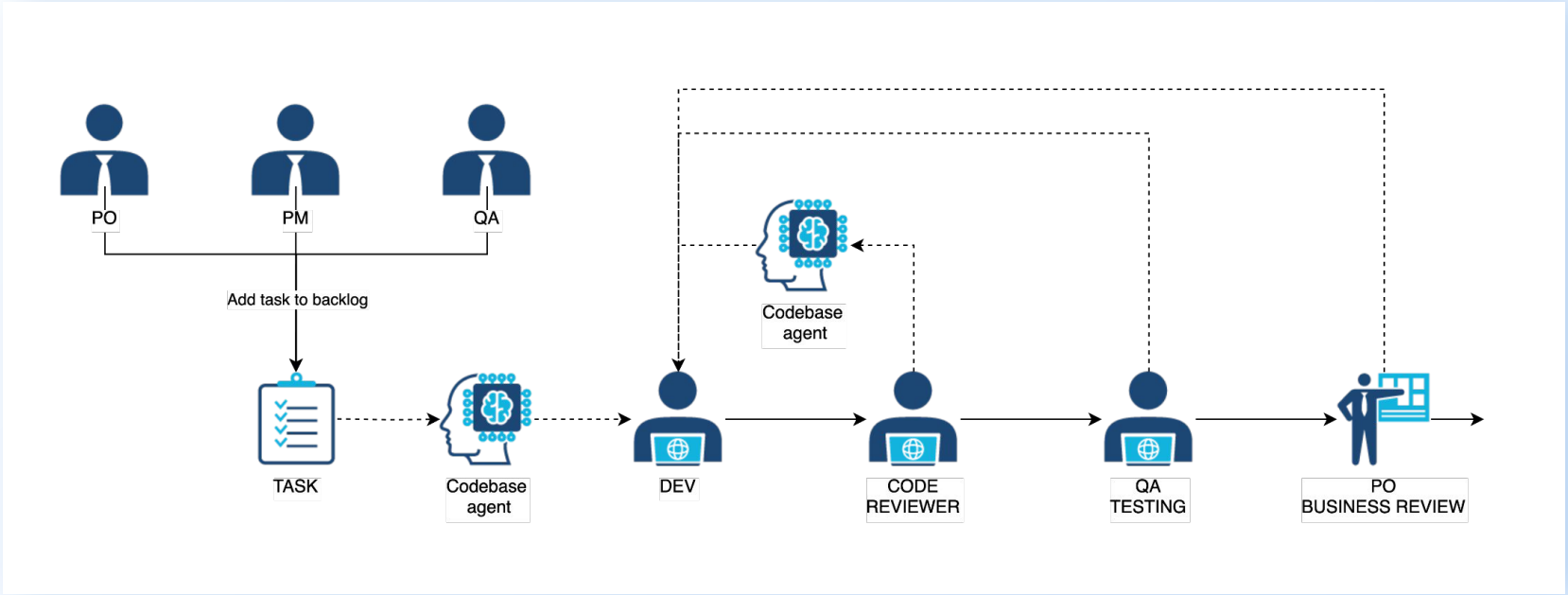
Initial approach - workflow



Initial approach - workflow



Initial approach - workflow



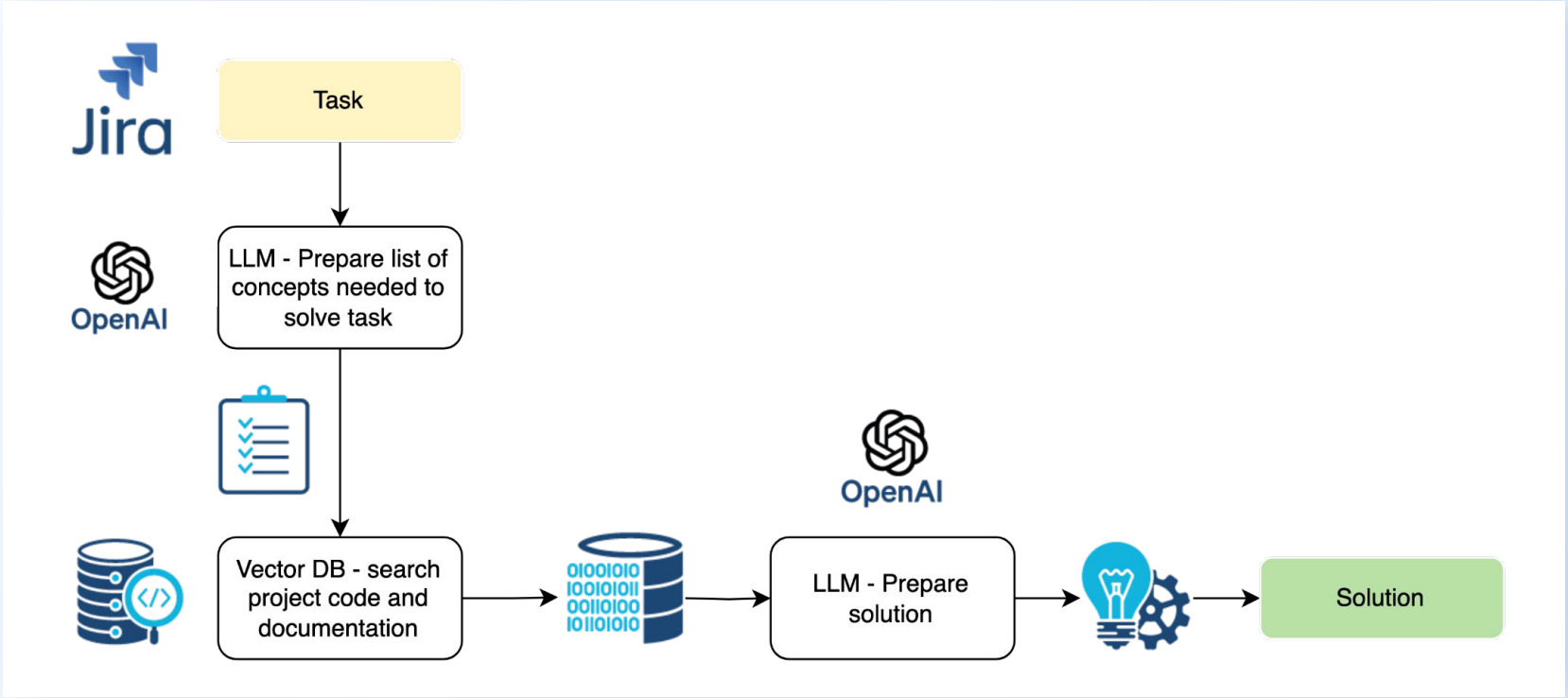
LangChain

LangGraph

Streamlit

Postgres + PG Vector

PoC - architecture



Concepts prompt

ConFoo.CA
DEVELOPER CONFERENCE

```
You are a chatbot tasked with solving software project issues.
You can be also supplied with code solution proposal to review.
The project is {self.project_description} and it's written in
{self.programming_language} language using {self.framework} framework.

Prepare message that will be used for semantic search in database for project code and
project documentation.
This message should contain some code if possible to match also files with code in
vector db.
Prepare message based on issue description below. Say which files should be checked.
Prepare list of information and concepts that are relevant to answering for the
problem described below. Take also into consideration directory structure of the
project


{ticket}\n\n{proj_dir_structure}
```

Example task

Person shown 2x on timeline

+ Add Apps

General PO Zoho

Description • Unsaved changes

The person is shown 2x on the timeline:

Todo

- Check why this person is shown twice and resolve the issue.

Concepts

generated output:

Relevant Information and Concepts:

1. Project Code Analysis:
   - **CompanyTimelineDataProvider Class**: This class is responsible for managing the timeline data for companies. It includes methods like `getDataChartForCompany`, `addCompanyManagersData`, and `addCompanySignatoriesData`, which are crucial for understanding how data is added to the timeline.
   - **PersonFunctionCache and PersonFunctionCacheRepository:** These are used to manage and retrieve cached data about person functions, which might be causing duplicate entries.
2. Project Documentation:
   ...

Semantic Search Query:

- Issue Description: Investigate the duplication of a person on the company timeline. The issue is observed in the `CompanyTimelineDataProvider` class, specifically within methods that handle manager and signatory data.
- Potential Causes: Examine the `PersonFunctionCache` and related repositories for duplicate entries. Check the logic in `addCompanyManagersData` and `addCompanySignatoriesData` for potential flaws in data aggregation or filtering.
- Relevant Classes and Methods:
  - `CompanyTimelineDataProvider::getDataChartForCompany`
  - `CompanyTimelineDataProvider::addCompanyManagersData`
  - `CompanyTimelineDataProvider::addCompanySignatoriesData`
  - `PersonFunctionCacheRepository`
  - `ManagerTile` and `ManagerFilter`

Solution prompt

ConFoo.CA
DEVELOPER CONFERENCE

```
You are a Senior Software Engineer with expertise in code analysis.
You have a strong ability to troubleshoot and  resolve issues based on the
information provided.
If you are uncertain about the answer, simply state that you do not know.

The project is a {self.project_description}.
It is built using {self.programming_language} and the {self.framework} framework.

When analyzing the provided code context, carefully evaluate:

   The structure and connections between different code components
   Key implementation details and coding patterns used
   The file paths and locations of each code snippet
   The type of code element (e.g., class, method, function, etc.)
   The name and purpose of each code segment
...
{ticket}
```

Solution

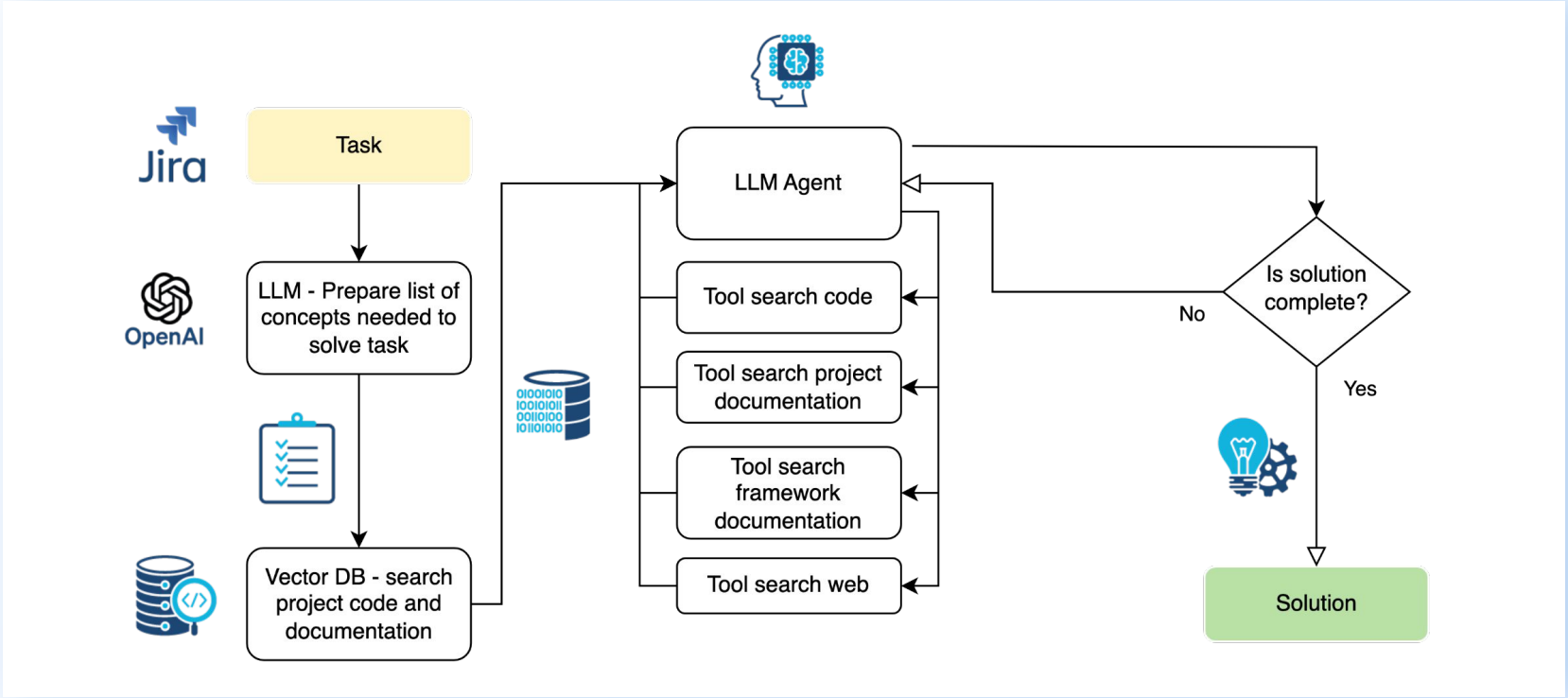
generated output:

Proposed Solution:

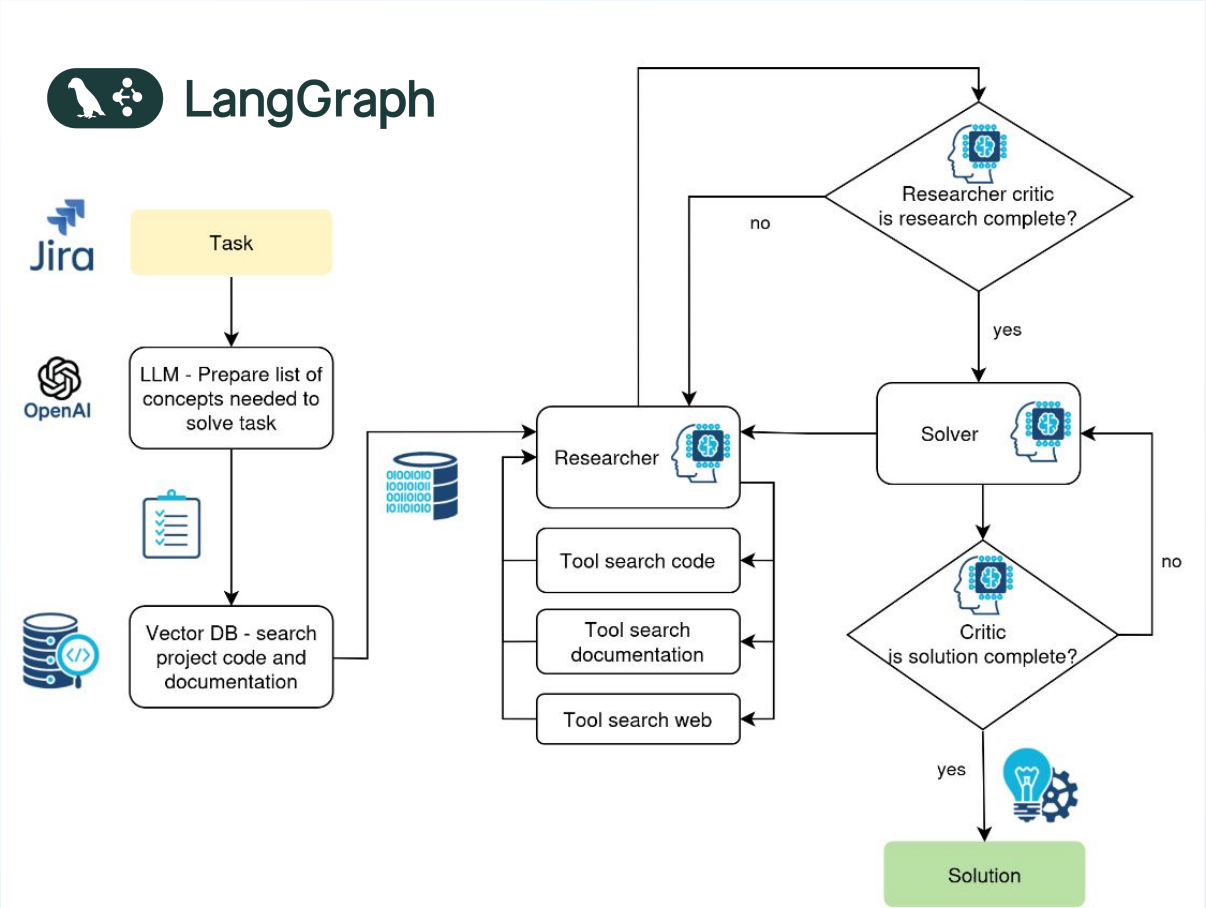
1. Identify Duplicate Entries:
    - In both addCompanyManagersData and addCompanySignatoriesData methods, we need to ensure that duplicate entries are not added to the timeline. This can be achieved by maintaining a set of unique identifiers for each person and checking against this set before adding a new entry.
2. Modify the Code:
    - Update the logic in addCompanyManagersData and addCompanySignatoriesData to check for duplicates before adding data to the timeline.

```
private function addCompanyManagersData(
    CompanyRecord $company,
    CompanyTimelineModel $companyTimelineModel
): void {
    if ($persons = $this->getCompanyManagers($company)) {
        $count = 0;
        $toTimeline = [];
        $uniqueIdentifiers = []; // To track unique entries

        /** @var PersonFunctionCache $person */
        foreach ($persons as $person) {
```

Agent architecture



Final architecture ?



Challenges and lessons learned

Project 1 - Codebase expert agent

1. LLMs are bad in multitasking unless split to dedicated agents
2. Well defined list of concepts is crucial as a separate step
3. Match all needed code files is challenging
4. Consider privacy issues with LLM API usage
5. **Best results having synergy between AI tool and human expert**

Github sample project
beta

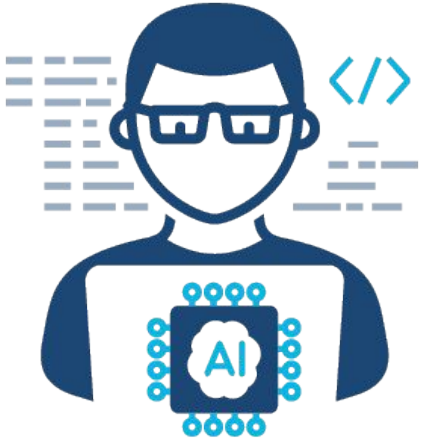
https://github.com/
mzarnecki/ai-codebase-expert

Similar company finder

ConFoo.CA
DEVELOPER CONFERENCE

1

Project 1
Codebase expert agent



generate code with AI

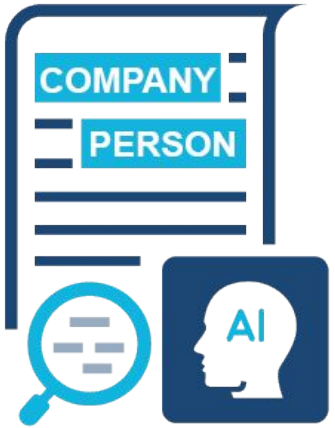
2

Project 2
Similar company finder

use LLMs

3

Project 3
Documents parser



train dedicated model

Similar company finder

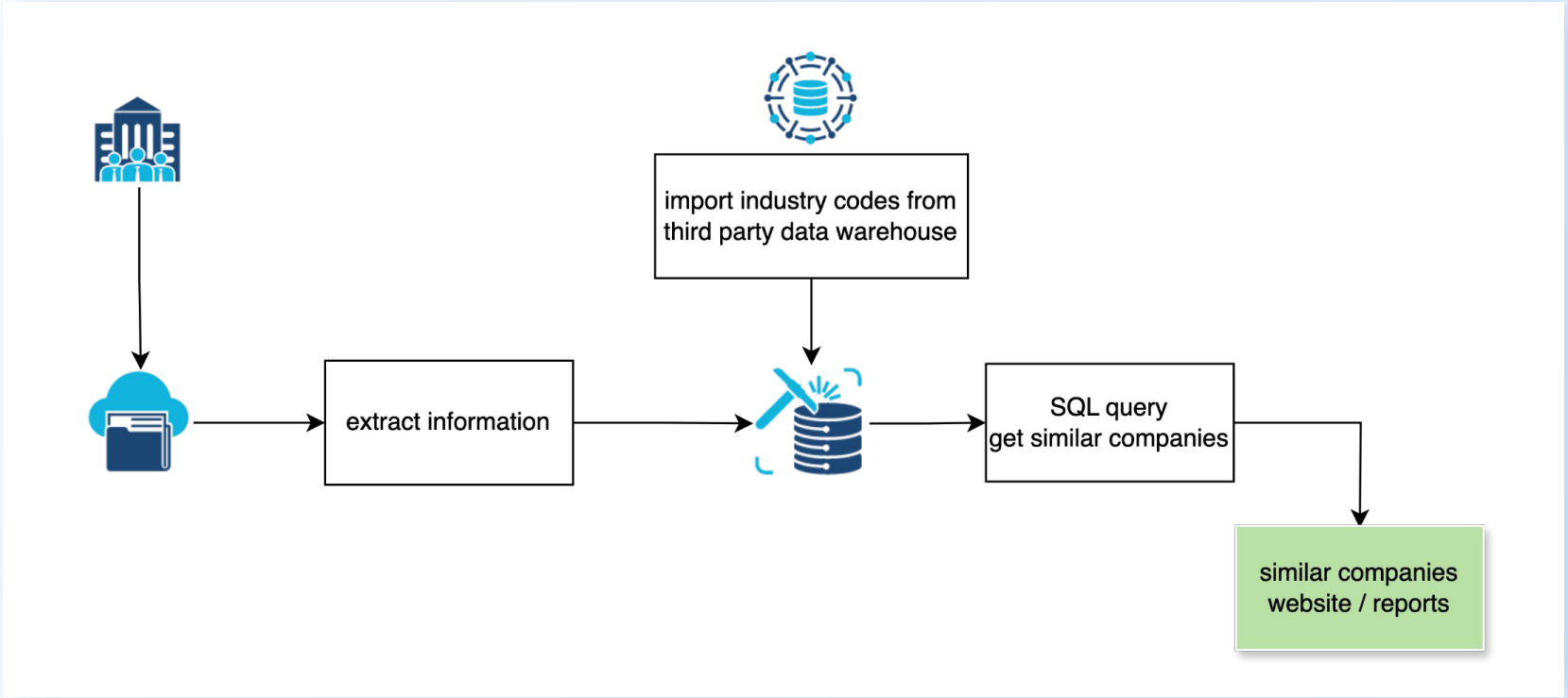
ConFoo.CA
DEVELOPER CONFERENCE

Similar companies

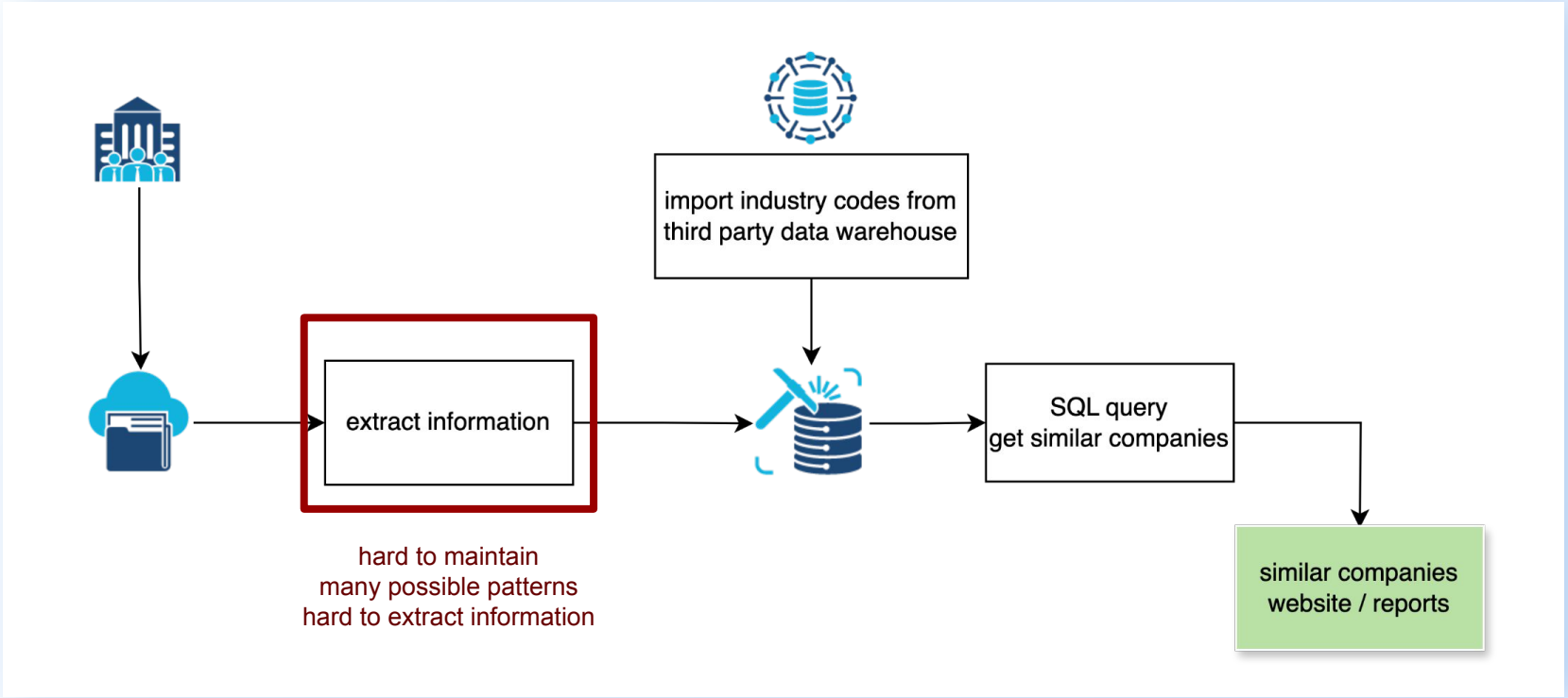
Similarity score

VBG GROUP VBG INVEST AG	68%
H.V. Verwaltungs GmbH	58%
Domestic & General Domestic & General Insurance	52%
Lombard Leasing GmbH	42%
Bartos Bartos Garten - und Landschaftsbau	19%

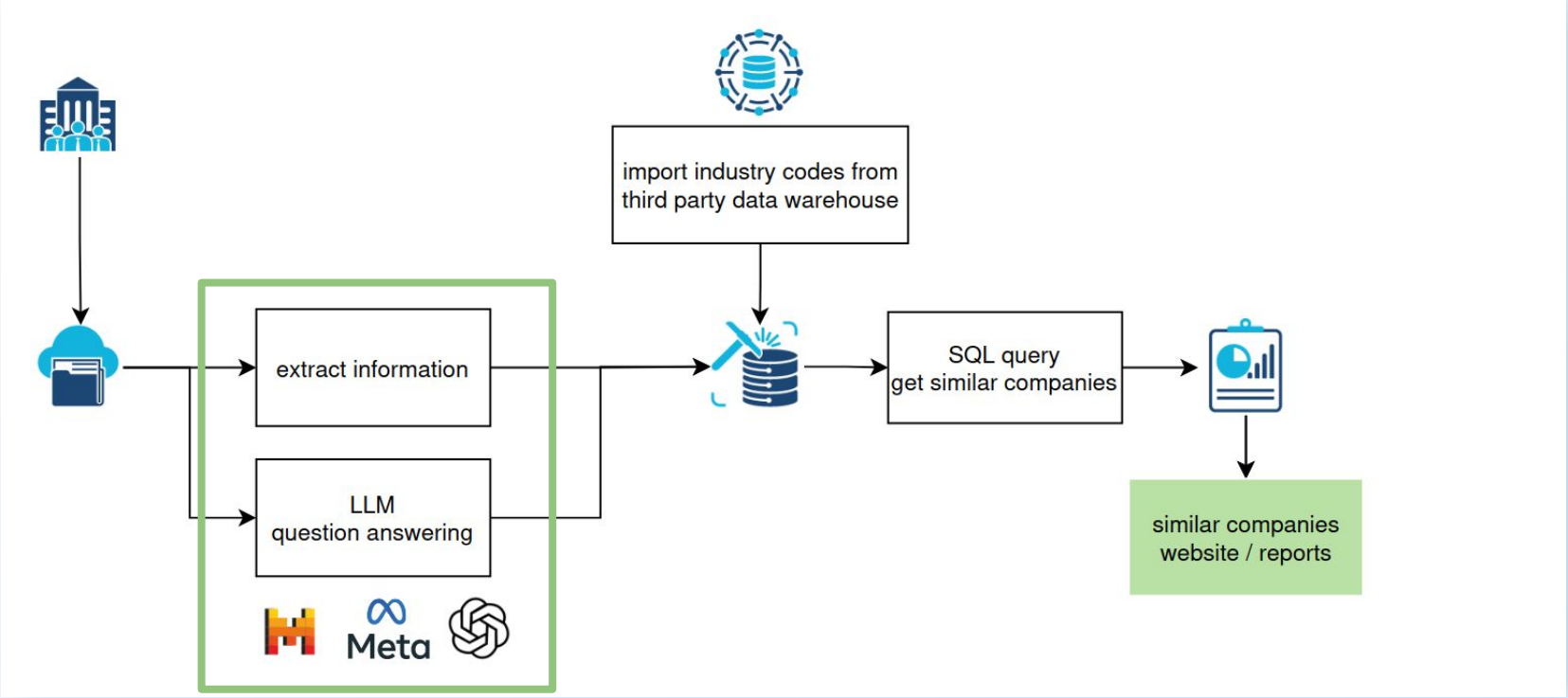
PoC - architecture



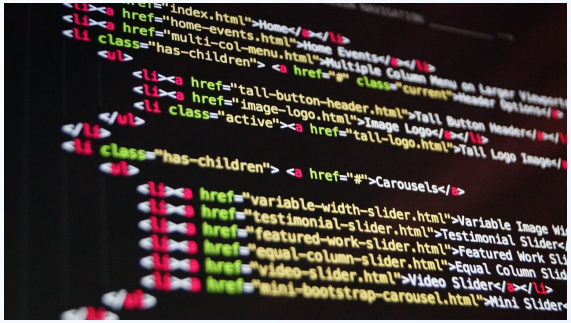
PoC - architecture



architecture



Extract information



HTML (purified)

prompt

LLM
API

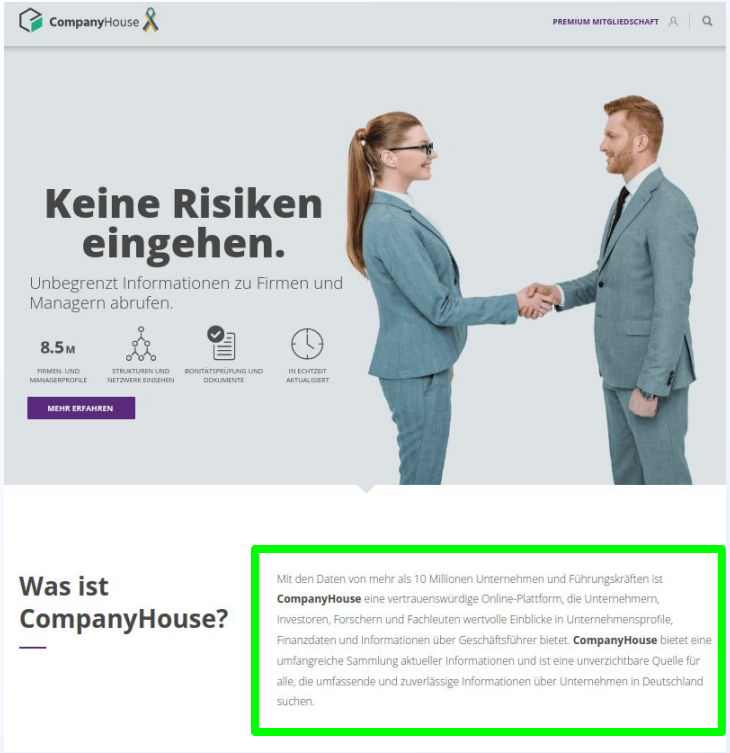
answer 1

answer 2

answer 3

32

Extract information



What industry does the company operate in?

What products and services does the company offer?

Which tags best describe this company?

Tags: Business Information Provider, Online Platform, Reliable Data Source, German Business Insights, Executive Profiles

33

Extract information

ConFoo.CA
DEVELOPER CONFERENCE

Provide tags for company based on website content.

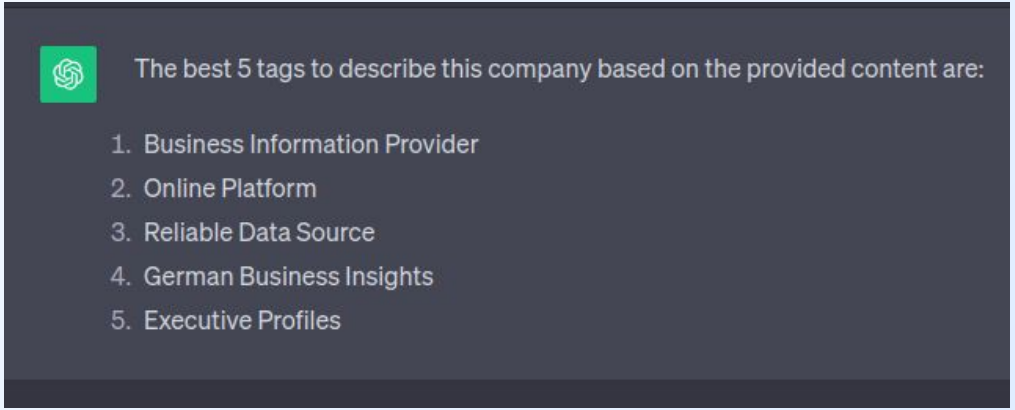
Based on the provided website content, here are some potential tags for the company:

1. Company Name: CompanyHouse
2. Industry: Business Information
3. Services: Business data, Financial information, Executive profiles
4. Geographic Focus: Germany
5. Data Size: More than 10 million companies and executives
6. User Audience: Entrepreneurs, Investors, Researchers, Professionals
7. Online Platform: Yes
8. Valuable Insights: Yes

Extract information

Provide tags for company based on website content.

What are the best tags to describe this company? Provide no more than 5.

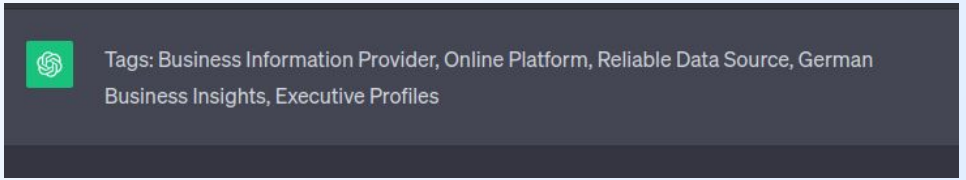


Extract information

ConFoo.CA
DEVELOPER CONFERENCE

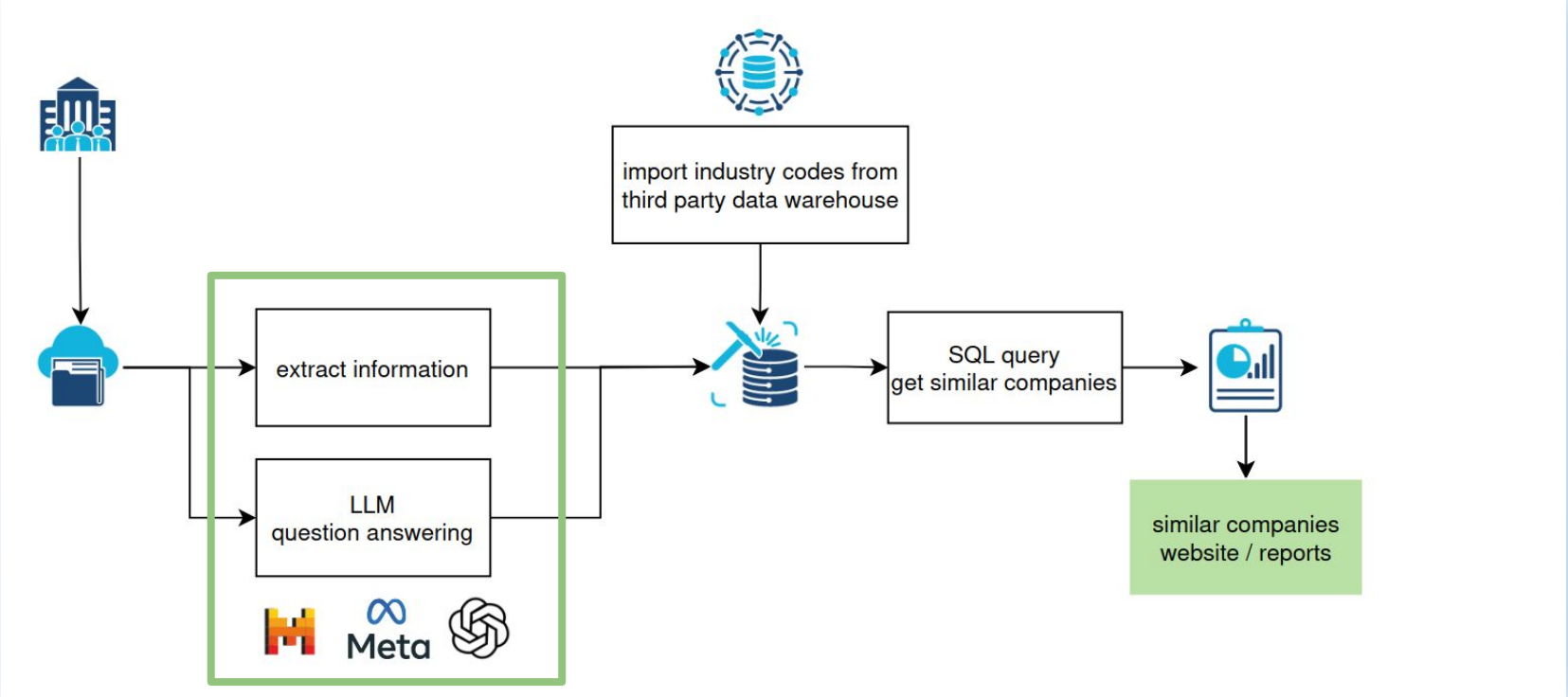
Provide tags for company based on website content.

What are the best tags to describe this company?
Provide no more than 5.



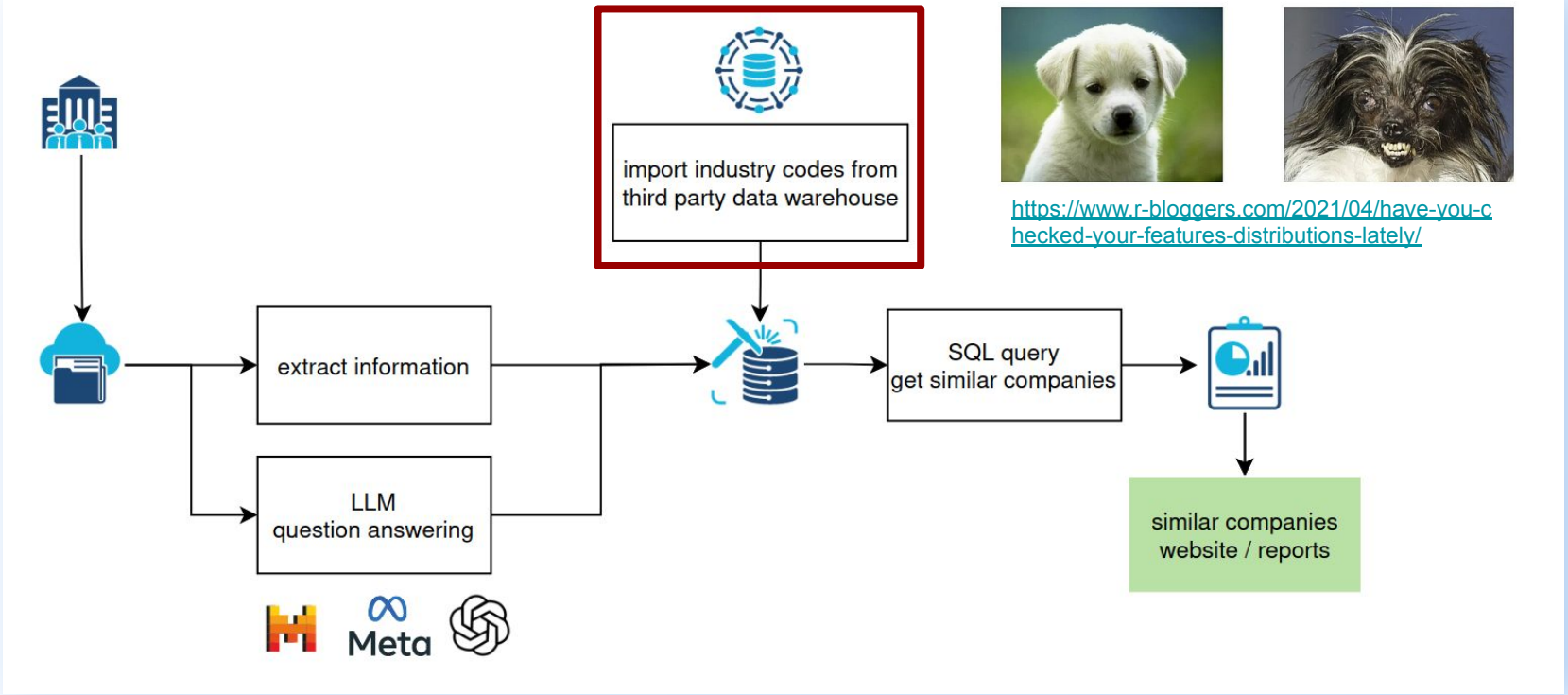
What are the best tags to describe this company?
Provide no more than 5 in format Tags: tag1, tag2.

architecture



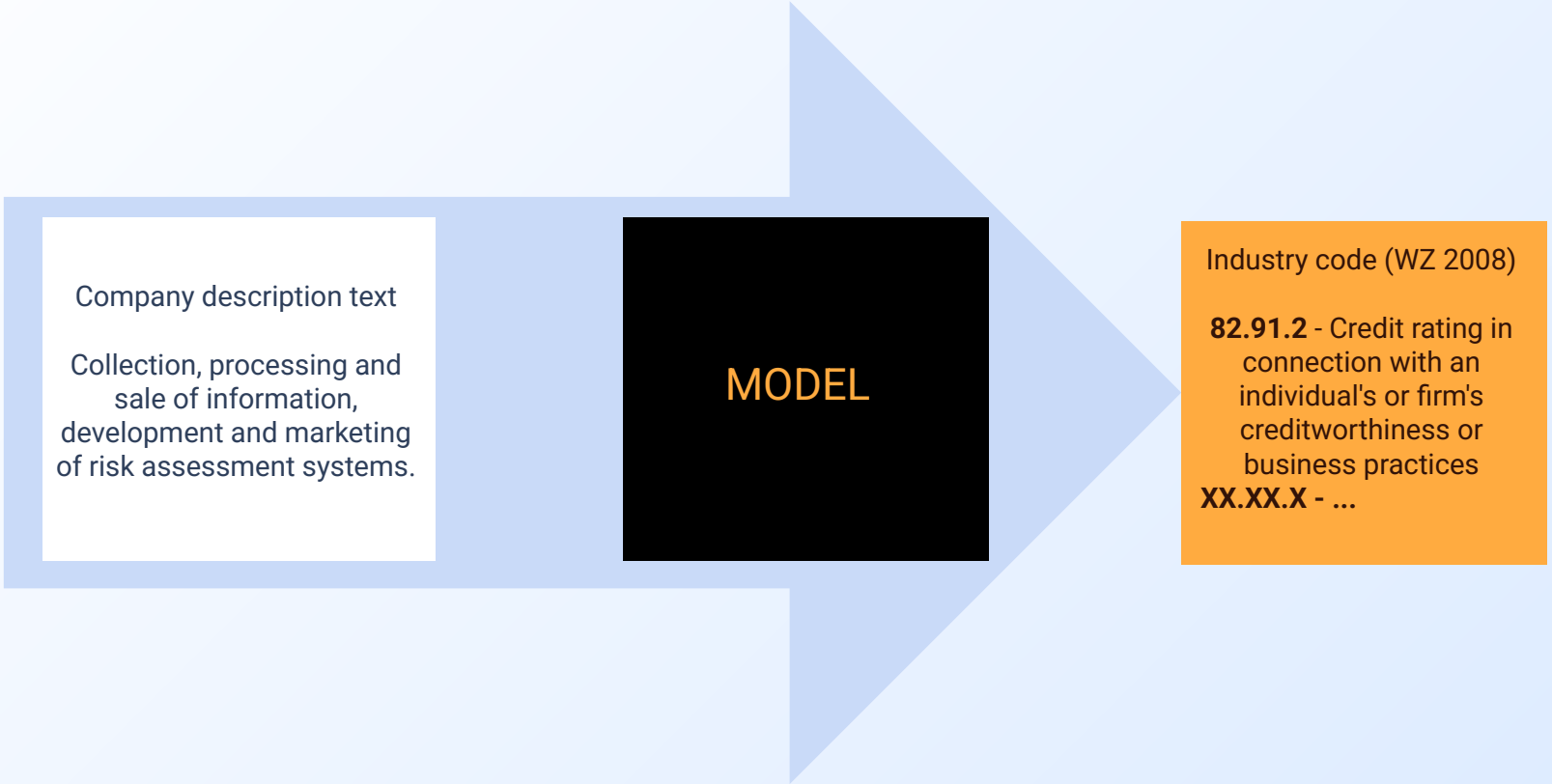
architecture

data quality issues
no control of data processing



https://www.r-bloggers.com/2021/04/have-you-checked-your-features-distributions-lately/

The goal


Company description text
Collection, processing and sale of information, development and marketing of risk assessment systems.
MODEL
Industry code (WZ 2008)
82.91.2 - Credit rating in connection with an individual's or firm's creditworthiness or business practices
XX.XX.X - ...

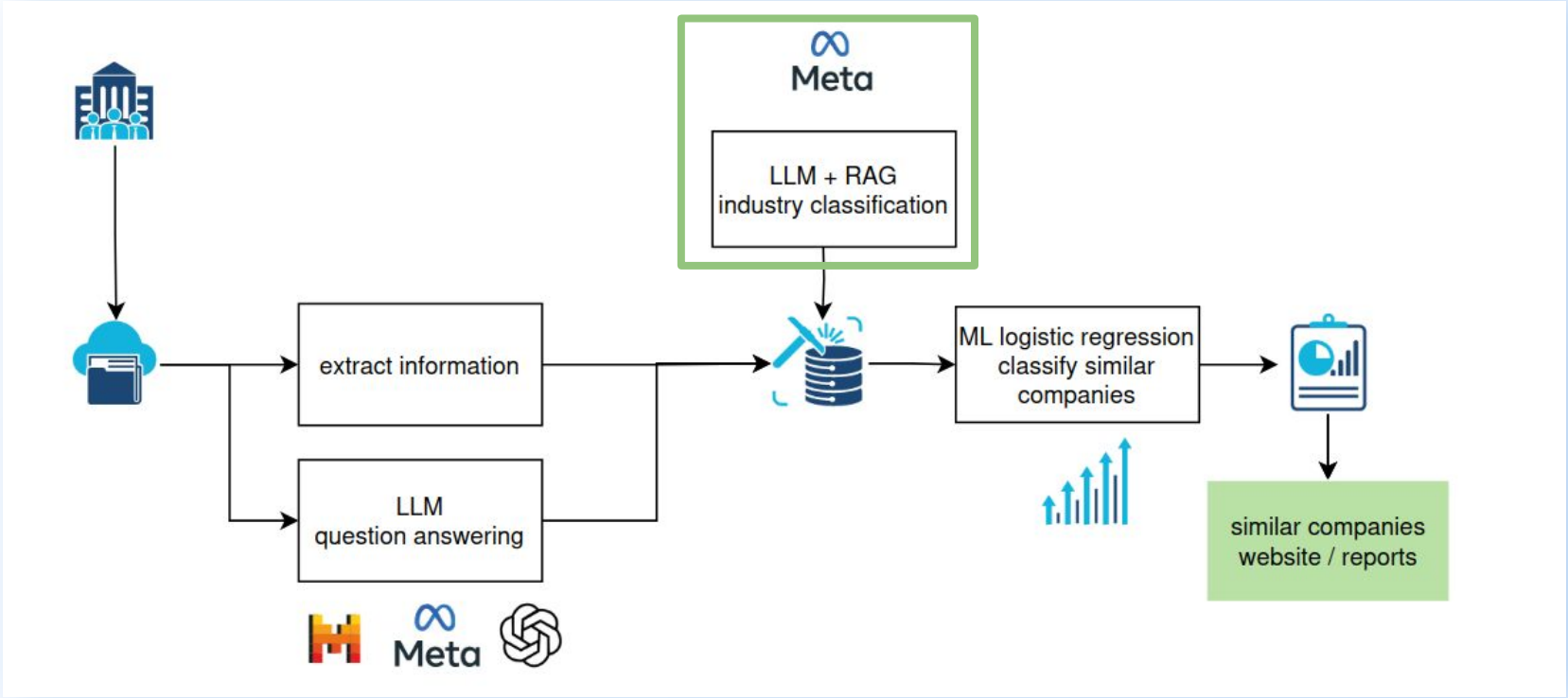
```
system_prompt = """You are AI assistant with access to descriptions of industry codes and company
information.
   Based on the provided information, your role is to match industry codes to the data provided.
   Do not infer or guess information that is not explicitly stated in the provided information.

   PROVIDE ONLY THE MAIN_CODE AND OTHER_CODES KEYS WITH THEIR RESPECTIVE VALUES;
   NEVER INCLUDE ANY EXPLANATION OR CONTEXT IN THE OUTPUT

   the example output should be:
    {
       'main_code': '11.11.1'
       'other_codes': [22.22.2, 33.33.3]
    }"""

user_prompt = f"""
   Below you can find description of the data for the company:
   {query}\n\n
   Here are the descriptions of German industry codes (Klassifikation der Wirtschaftszweige, Ausgabe
2008).
   Base your answers only on the codes provided below; do not use any other codes apart from those
below:\n
"""

for idx, doc in enumerate(selected_documents):
   user_prompt += f"Example {idx+1}: {doc}\n"
```

```
# call GPT-4 API for classification results
response = openai.OpenAI().chat.completions.create(
    model="gpt-4",
    messages=[
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": user_prompt}
    ]
)
```

```
# call GPT-4 API for classification results
response = openai.OpenAI().chat.completions.create(
    model="gpt-4",
    messages=[
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": user_prompt}
    ]
)
```

Classification Result:
'main_code': '01.11.0'
'other_codes':['01.61.0']

Final approach - architecture

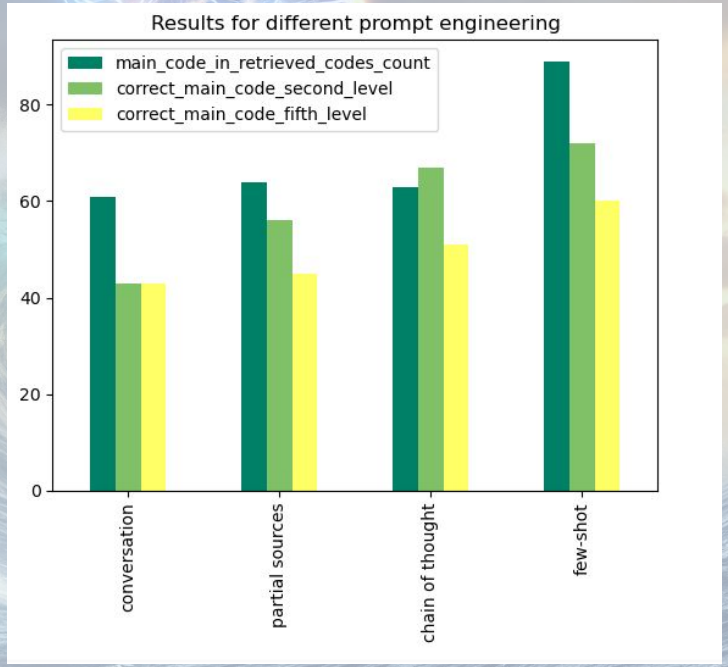


Challenges and lessons learned

Project 3 - Similar companies finder

GitHub
code samples

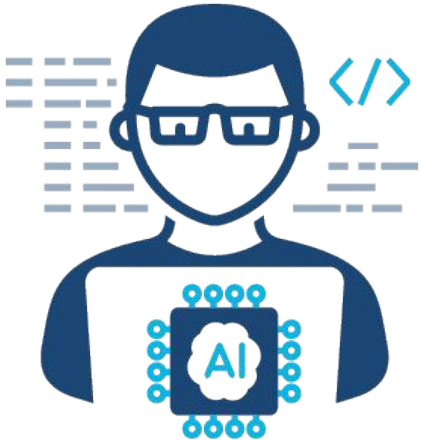
1. Poor results of data extraction can be improved with LLMs
2. Data quality is crucial
3. Try different prompt techniques
4. Try different LLM and embedding models



Documents Parser

1

Project 1
Codebase expert agent



generate code with AI

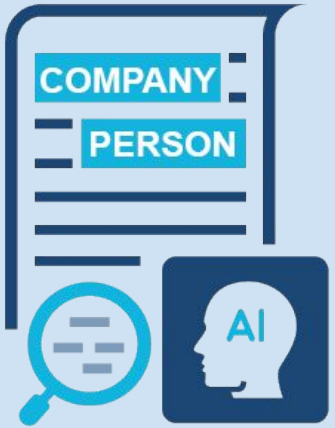
2

Project 2
Similar company finder

use LLMs

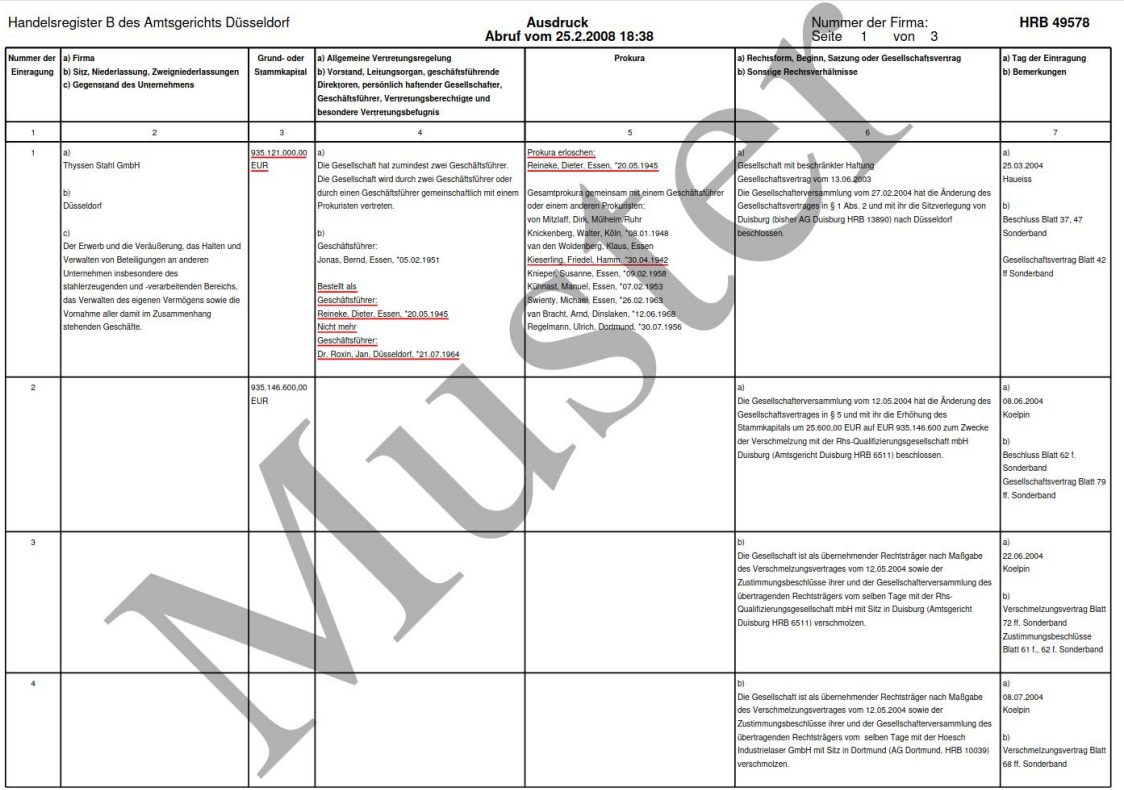
3

Project 3
Documents parser



train dedicated model

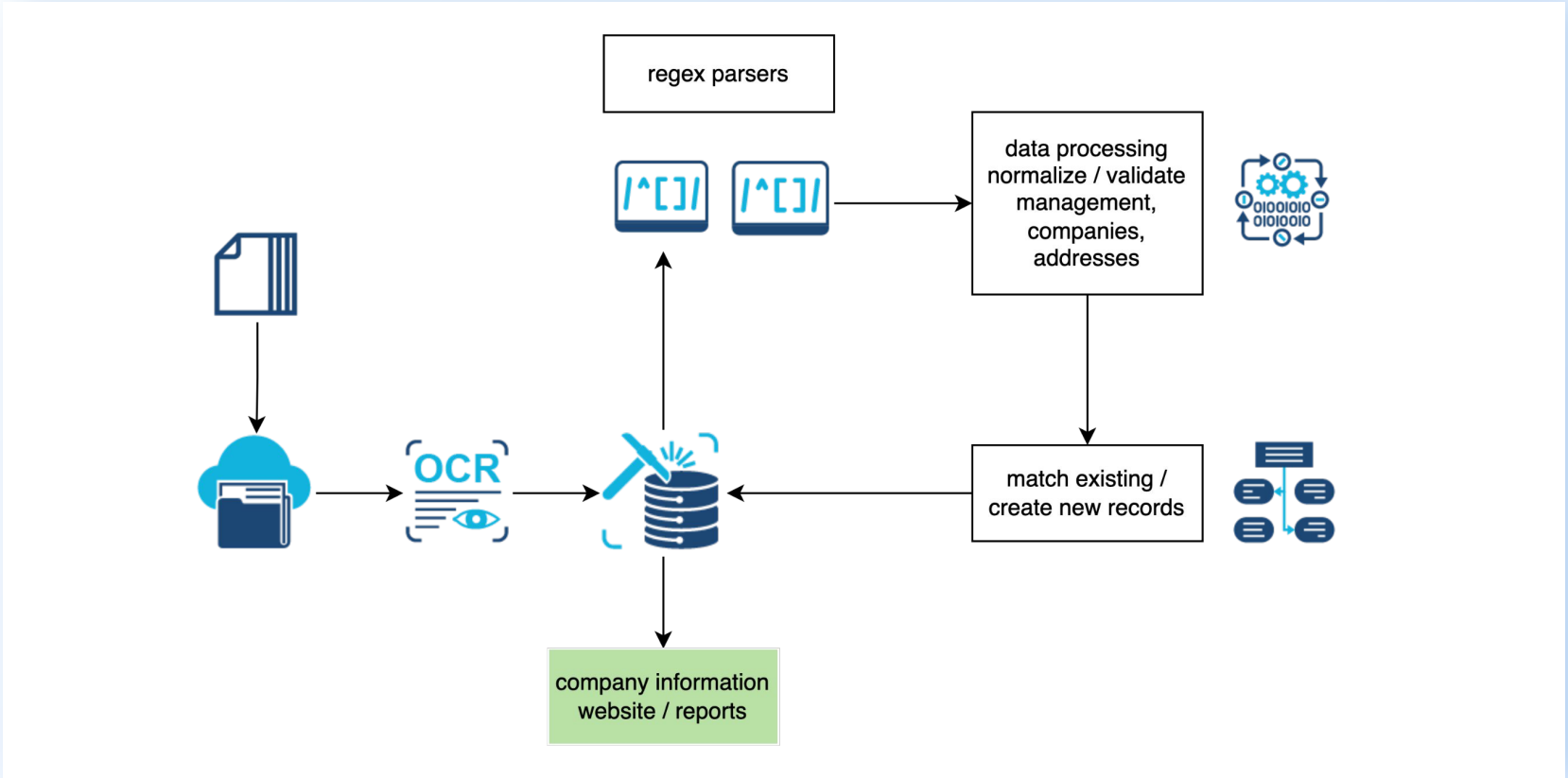
Extract information from documents



Handelsregister B des Amtsgerichts Düsseldorf — **Ausdruck Abruf vom 25.2.2008 18:38** — Nummer der Firma: **HRB 49578** — Seite 1 von 3

Nummer der Eintragung	a) Firma b) Sitz, Niederlassung, Zweigniederlassungen c) Gegenstand des Unternehmens	Grund- oder Stammkapital	a) Allgemeine Vertretungsregelung b) Vorstand, Leitungsorgan, geschäftsführende Direktoren, persönlich haftender Gesellschafter, Geschäftsführer, Vertretungsberechtigte und besondere Vertretungsbefugnis	Prokura	a) Rechtsform, Beginn, Satzung oder Gesellschaftsvertrag b) Sonstige Rechtsverhältnisse	a) Tag der Eintragung b) Bemerkungen
1	2	3	4	5	6	7
1	a) Thyssen Stahl GmbH b) Düsseldorf c) Der Erwerb und die Veräußerung, das Halten und Verwalten von Beteiligungen an anderen Unternehmen insbesondere des stahlerzeugenden und -verarbeitenden Bereichs, das Verwalten des eigenen Vermögens sowie die Vornahme aller damit im Zusammenhang stehenden Geschäfte.	935.121.000,00 EUR	a) Die Gesellschaft hat zumindest zwei Geschäftsführer. Die Gesellschaft wird durch zwei Geschäftsführer oder durch einen Geschäftsführer gemeinschaftlich mit einem Prokuristen vertreten. b) Geschäftsführer: Jonas, Bernd, Essen, *05.02.1951 Bestellt als Geschäftsführer: Reineke, Dieter, Essen, *20.05.1945 Nicht mehr Geschäftsführer: Dr. Roxin, Jan, Düsseldorf, *21.07.1964	Prokura erloschen: Reineke, Dieter, Essen, *20.05.1945 Gesamtprokura gemeinsam mit einem Geschäftsführer oder einem anderen Prokuristen: von Mitzlaff, Dirk, Mülheim/Ruhr Knickenberg, Walter, Köln, *08.01.1948 van den Woldenberg, Klaus, Essen Kieserling, Friedel, Hamm, *30.04.1942 Knieper, Susanne, Essen, *09.02.1958 Kühnast, Manuel, Essen, *07.02.1953 Swienty, Michael, Essen, *26.02.1963 van Bracht, Arnd, Dinslaken, *12.06.1968 Regelmann, Ulrich, Dortmund, *30.07.1956	a) Gesellschaft mit beschränkter Haftung Gesellschaftsvertrag vom 13.06.2003 Die Gesellschafterversammlung vom 27.02.2004 hat die Änderung des Gesellschaftsvertrages in § 1 Abs. 2 und mit ihr die Sitzverlegung von Duisburg (bisher AG Duisburg HRB 13890) nach Düsseldorf beschlossen.	a) 25.03.2004 Haueiss b) Beschluss Blatt 37, 47 Sonderband Gesellschaftsvertrag Blatt 42 ff Sonderband
2		935.146.600,00 EUR			a) Die Gesellschafterversammlung vom 12.05.2004 hat die Änderung des Gesellschaftsvertrages in § 5 und mit ihr die Erhöhung des Stammkapitals um 25.600,00 EUR auf EUR 935.146.600 zum Zwecke der Verschmelzung mit der Rhs-Qualifizierungsgesellschaft mbH Duisburg (Amtsgericht Duisburg HRB 6511) beschlossen.	a) 08.06.2004 Koelpin b) Beschluss Blatt 62 f. Sonderband Gesellschaftsvertrag Blatt 79 ff. Sonderband
3					b) Die Gesellschaft ist als übernehmender Rechtsträger nach Maßgabe des Verschmelzungsvertrages vom 12.05.2004 sowie der Zustimmungsbeschlüsse ihrer und der Gesellschafterversammlung des übertragenden Rechtsträgers vom selben Tage mit der Rhs-Qualifizierungsgesellschaft mbH mit Sitz in Duisburg (Amtsgericht Duisburg HRB 6511) verschmolzen.	a) 22.06.2004 Koelpin b) Verschmelzungsvertrag Blatt 72 ff. Sonderband Zustimmungsbeschlüsse Blatt 61 f., 62 f. Sonderband
4					b) Die Gesellschaft ist als übernehmender Rechtsträger nach Maßgabe des Verschmelzungsvertrages vom 12.05.2004 sowie der Zustimmungsbeschlüsse ihrer und der Gesellschafterversammlung des übertragenden Rechtsträgers vom selben Tage mit der Hoesch Industrielaser GmbH mit Sitz in Dortmund (AG Dortmund, HRB 10039) verschmolzen.	a) 08.07.2004 Koelpin b) Verschmelzungsvertrag Blatt 68 ff. Sonderband

Use regex parsers - architecture



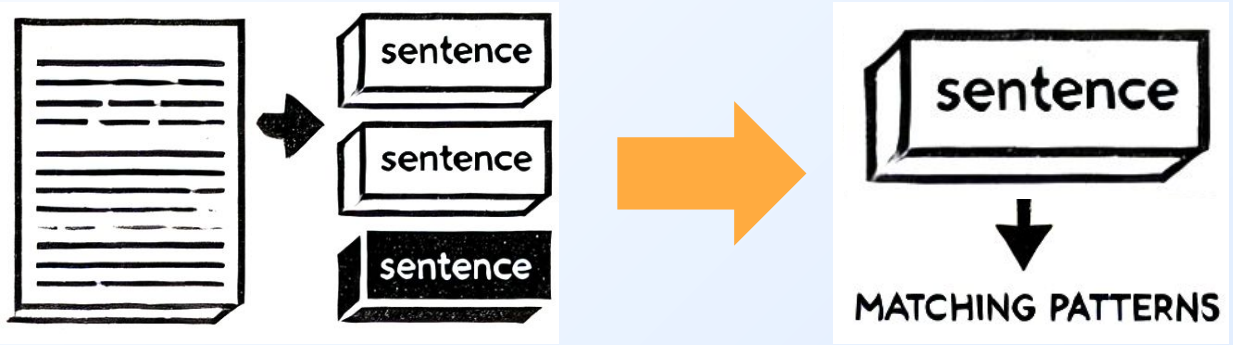
First approach - regexp

Assumptions:

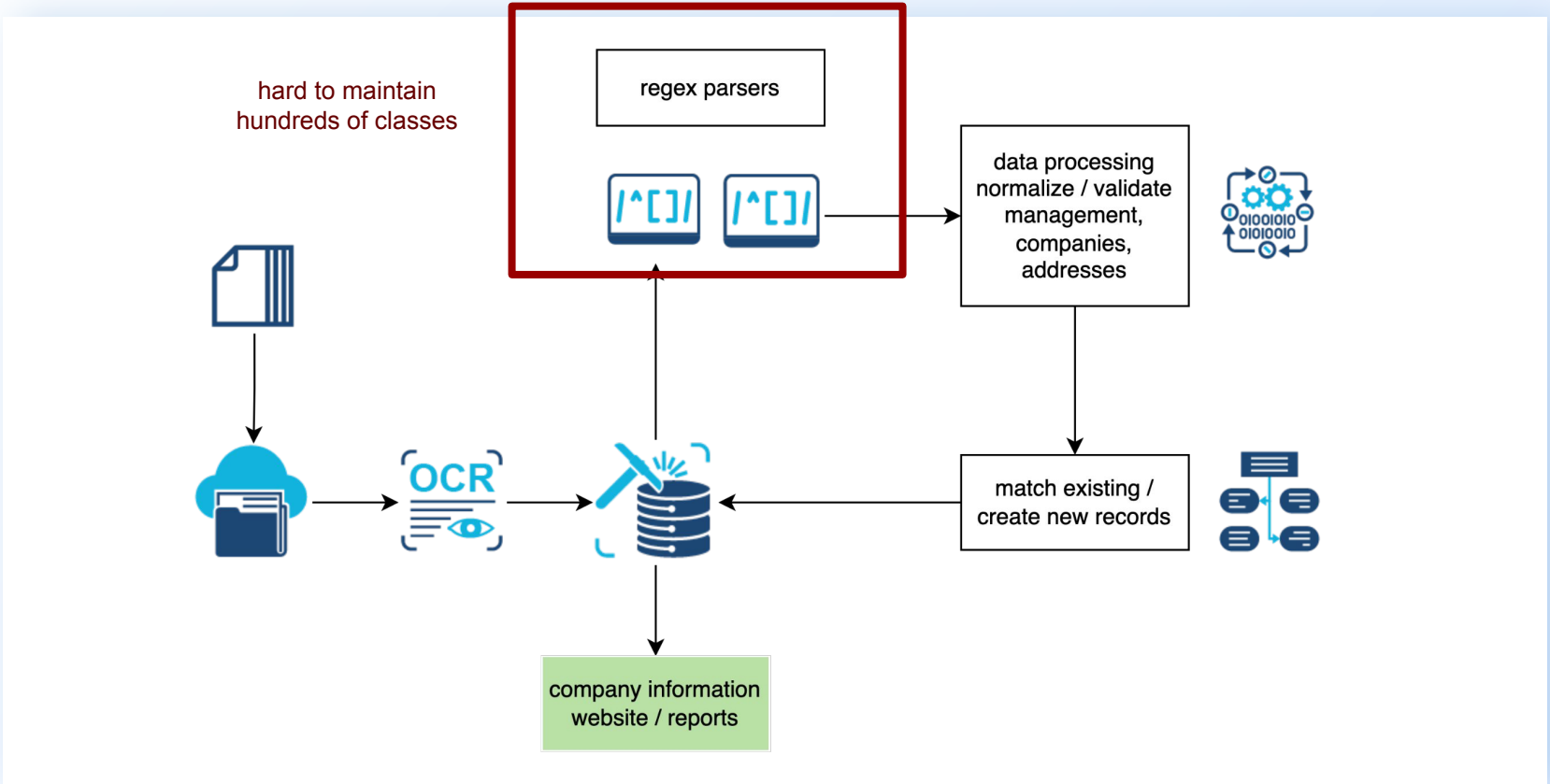
- Conditions are based on the order of words in the sequence
- Multiple possible variants: **<ScientificTitle> LastName, FirstName, City, <BirthDate>, <SigningAuthority>**

*Management Board: Prof. Dr. Doe, John, Saint-Ambroise, *01.02.1976*

Managing Director: Michael, Json, NY, 19.04.1987, Managing Director, with sole power of representation



Use regex parsers - architecture

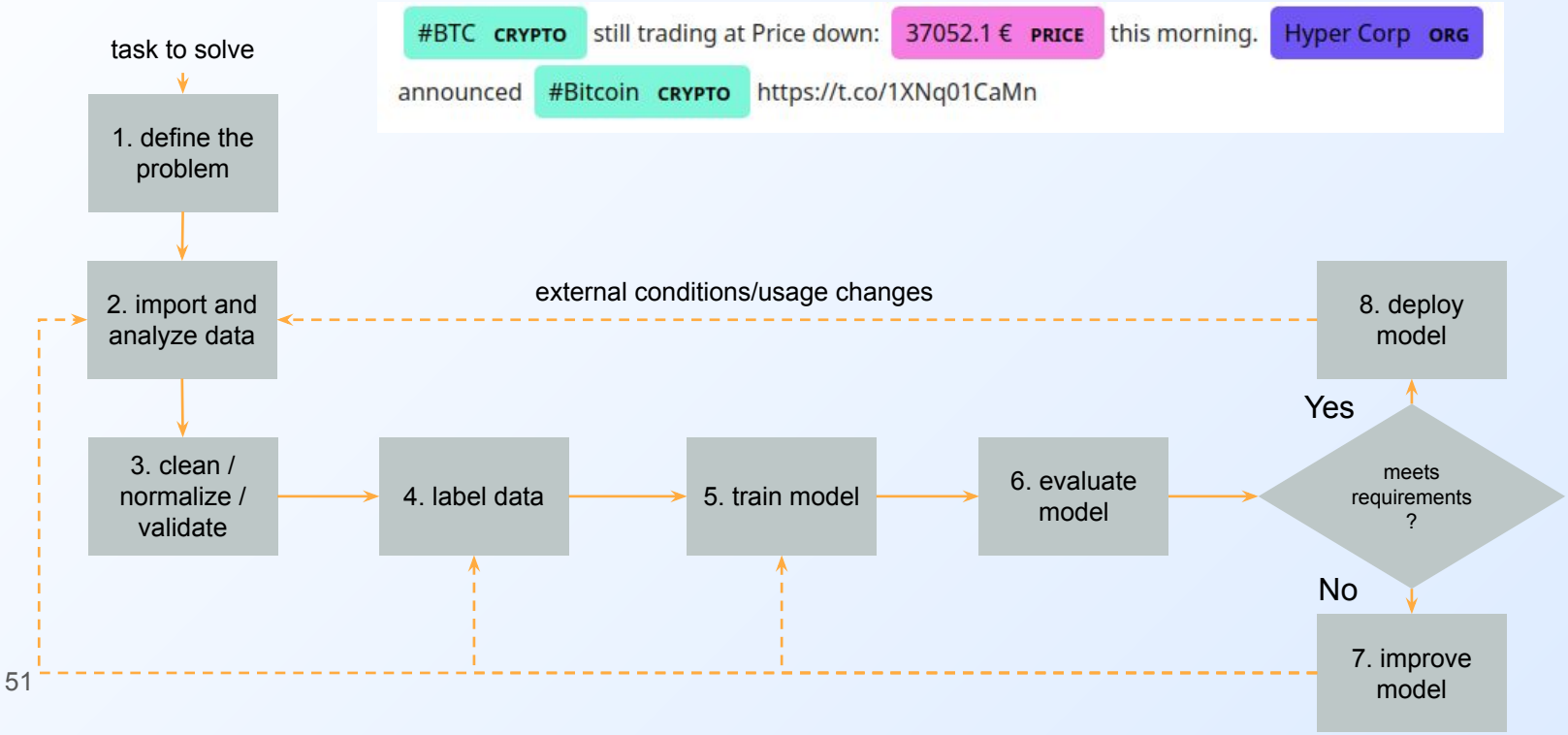


Challenges

- Assumptions may not cover 100% of cases
- Different structure, different order of phrases
- Foreign names, multi-word names, titles
- Complicated logic to cover all cases
- Good accuracy for structured texts
- Low accuracy when there are many variants

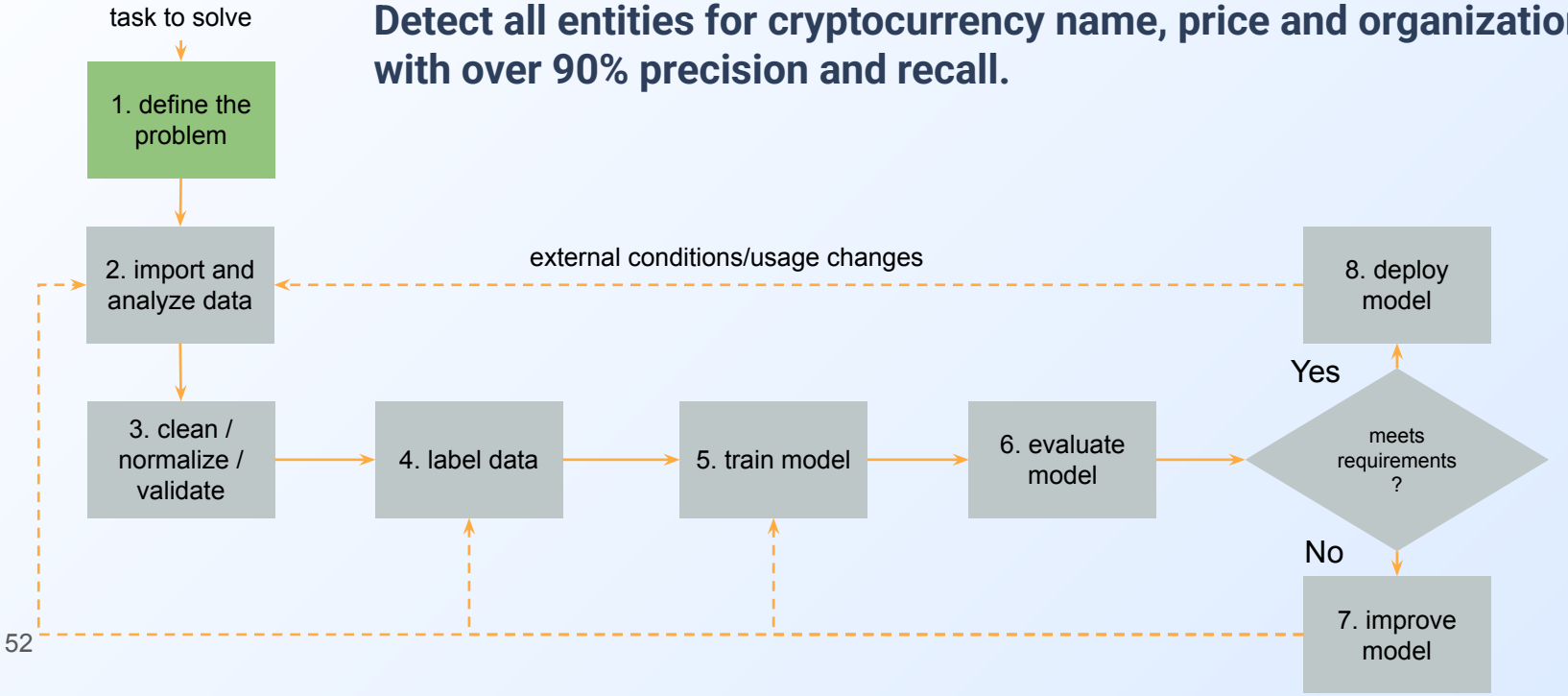

```
(?<main_header>(?:
Änderung\sder\sFirma:\s+)?(?:HRB\s\d+(?:\s|-)\p{Lu}\p{Ll}?\s--\s(?:infolge\sSitzverleg
ung\s)?jetzt:\s+)?(?:\d{0,2}[\d\p{Ll}]\s)?\b(?<trade_register_number>((?<number_prefix
_ar>\d+)\s)?(?<type>HRA|HRB|GnR|AR|PR|VR|HR\sA|HR\sB)\s?(?<number_first>\d+)(?:\s?(?<s
lash>\/))?(?:\s?(?<number_second>\d{2,4}))?(?:\s?(?<type_end>\p{L}{1,3}))?)\b(?:-\p{Lu
}\.?)?(?:\s\((?:früher|bisher|vorher)\s(HRA|HRB|GnR|AR|PR|VR|HR\sA|HR\sB)\s\d+,?\sAG\s
\p{Lu}[\p{L}
]+\p{L}\))?(?:(?:\s--\sAG)?\s(?<court>Bad\sHomburg\sv\.\sd\.\sH\.|Bad\sHomburg\sv\.d\.
Höhe|Bad\sHomburg\sv\.\sd\.\sHöhe|Bad\sHomburg\sv\.d\.\sHöhe|Bremerhaven|Frankfurt\s\(
Oder\)|Frankfurt\(Oder\)|Freiburg\sim\sBreisgau|Freiburg\sim\sBr\.|Freiburg\si\.Br\.Be
rlin|Berlin-Charlottenburg|Charlottenburg\s\(Berlin\)|Charlottenburg|Charlottenburg\sB
erlin|Frankfurt\sOder|Frankfurt\sa\.M\.|Frankfurt\/Main|Frankfurt\sa\.\sMain|Homburg\s
\(Saar\)|Kempten|Kempten\sAllgäu|Kempten\/Allgäu|Kempten\(Allgäu\)|Ludwigshafen|Ludwig
shafen\sam\sRhein|Ludwigshafen\sa\.Rhein|Ludwigshafen\sa\.\sRhein|Ludwigshafen\/Rhein|
Oldenburg|Offenbach|St\.\sIngbert|St\.\sWendel|Weiden|Weiden\si\.d\.\sOPf\.|Weiden\si\
.d\.OPf\.|Weiden\si\.d\.OPf|Weiden\si\.d\.\sOPf|Dorsten|Karlsruhe|Detmold|Hameln|Konst
anz|Straubing|Degendorf))?(?:(?:\s-\s)?|(?:\s?-{2,3}\
\(?(?<date>(?:(?<year>\d{4})(?<date_parts_separator_1>[\-\/\.\s\\]{1,2})(?<month>\d{1,
2})\g{date_parts_separator_1}(?<day>\d{1,2})|(?<day2>\d{1,2})(?<date_parts_separator_2
>[\-\/\.\s\\]{1,2})(?:(?<month2>\d{1,2})((\g{date_parts_separator_2})|(?<=\p{Ll}))|(?<
month3>(?:Januar|Februar|März|April|Mai|Juni|Juli|August|September|Oktober|November|De
zember))(?:\s)?)(?<year2>\d{2,4}))\)?)|):(?:\s?\d{2}\.\d{2}\.(?:19:|20)\d{2})?\s?(?:(
?:(?:[Bb]isherige\s)?(?:Firma|Genossenschaft|Gen\.|Name)?(?:\s?\((?:[Bb]isher|[Vv]orhe
r|[Ff]rüher)[^)]+\))?(?:(?:\sund|,)\sSitz)?:?\.?[\s;](?:(?:\/\sName\s)?vormals:\s)?(?:
der\sFirma\s)?)|(?:(?:(?:Berichtigung\sund\s)?Ergänzung\s|(?:Betr\.\sFa\.:\s)?Berichti
gung(?:\sder\sVeröffentlichung)?\s|Berichtigung\/Ergänzung\s)?(?:der|d\.)\s(?:Eintragu
ng|Eintr\.)\s(?:vom|v\.)\s(?:.*?\d{4}[.:]?)\s)|(?:Berichtigung\szur\sEintragung\s(?:vo
m|v\.)\s(?:.*?\d{4}[.:]?)\s)|(?:Die\sEintragung\svom(?:\s\d{1,2}\.)?(?:\s\d{1,2}\.)?.*
(?:\sFirma(?:\slautet\srichtig:)?\s|\swie\sfolgt\sberichtigt:\s))|(?:Die\sEintragung\s
erfolgte\sam(?:\s\d{1,2}\.)?(?:\s\d{1,2}\.)?\s\d{4}:\s)|(?:Eintragung\svon\sAmts\swege
n:\s)|(?:(durch\s)?
Übernahme|Umschreibung)\saus\sdem\s.+?\sFirma\sund\sSitz:\s|)?(?:(?:früher|vorher|\(bi
sher\)):\s(?:(?<old_company>[^,]+)(?:,|.\sDie\sFirma|.\s\()(?:.+\)?,?)?\s?jetzt\)?:\s)
?)?(?<company>.+?))(?:\s?(?:\(Zweigniederlassung.+\))?(?:,?\sSitz:?|(?:\smit\sSitz)?\s
in|,)\s(?<address>(?:(?<=^|\s|\()(?:(?:(?<zip>\d{5}))|(?:(?<zip_former>(?:O|W)-\d{4}))
|(?:(?<zip_former_not_prefixed>\d{4})))(?=\s|$|\.|\)))\s)?(?<=^|\s)(?:(?<district>(?:\p
{Lu}(?:\p{Ll}){2,18}(?:(?:\s|-)\(?\p{L}{2,15}\)?(?:(?:(?:\s|-)\p{L}{3,14}(?:\s\p{Lu}(?
:\p{Ll}){3,9}))?)|(?:\s\(Hameln\-Pyrmont\)))?)?))\/\s?(?:OT|Orts\.|Ortsteil)\s)?(?<city
>(?!OT\s)(?:\p{Lu})(?:(?:(?:(?:\p{Ll}+(?(?<=St)\.)?(?:-\p{Lu}\p{Ll}+)?(?:(?:(?:(?:\s(?
:(?!und)\p{Ll}{1,6}(?:\s|\.\s?)){1,2})|(?:\s|\s?-\s?|\s?\/\s?(?!\p{Lu}\p{Lu})|(?<=\.))
)(?:(?<city_g>(?!D(?:as|ie|er)\s)(?!OT\s)(?!Vorsitzender)\p{Lu}{1,3}\p{Ll}*(?(?<!\p{L}
{7})(?<=\p{Ll}\.)?)|\(\g<city_g>(?:(?:\s|-)\g<city_g>){0,3}\))){0,5})?)))(?:(?<!\.)\p
{Ll}|(?:(?<=\)|\.))|(?<=\s\p{Lu}{2})|(?<=\s\p{Lu}{3})))(?:,?\s(?:OT|Orts\.|Ortsteil)\s
(?<district_2>\g<district>))?(?=\s|,|\.|$|\)|;|\(|(?:(?<!\.)\p{Lu})|\/)(?:,\s(?<branch
>Zweigniederlassung\ der\ (?<company2>.+)\ mit\ (?:dem\ )?Sitz\ in\
(?<city2>(?!OT\s)(?:\p{Lu})(?:(?:(?:(?:\p{Ll}+(?(?<=St)\.)?(?:-\p{Lu}\p{Ll}+)?(?:(?:(?
:(?:\s(?:(?!und)\p{Ll}{1,6}(?:\s|\.\s?)){1,2})|(?:\s|\s?-\s?|\s?\/\s?(?!\p{Lu}\p{Lu})|
(?<=\.)))(?:(?<city2_g>(?!D(?:as|ie|er)\s)(?!OT\s)(?!Vorsitzender)\p{Lu}{1,3}\p{Ll}*(?
```

Solution - ML model for Named Entity Recognition



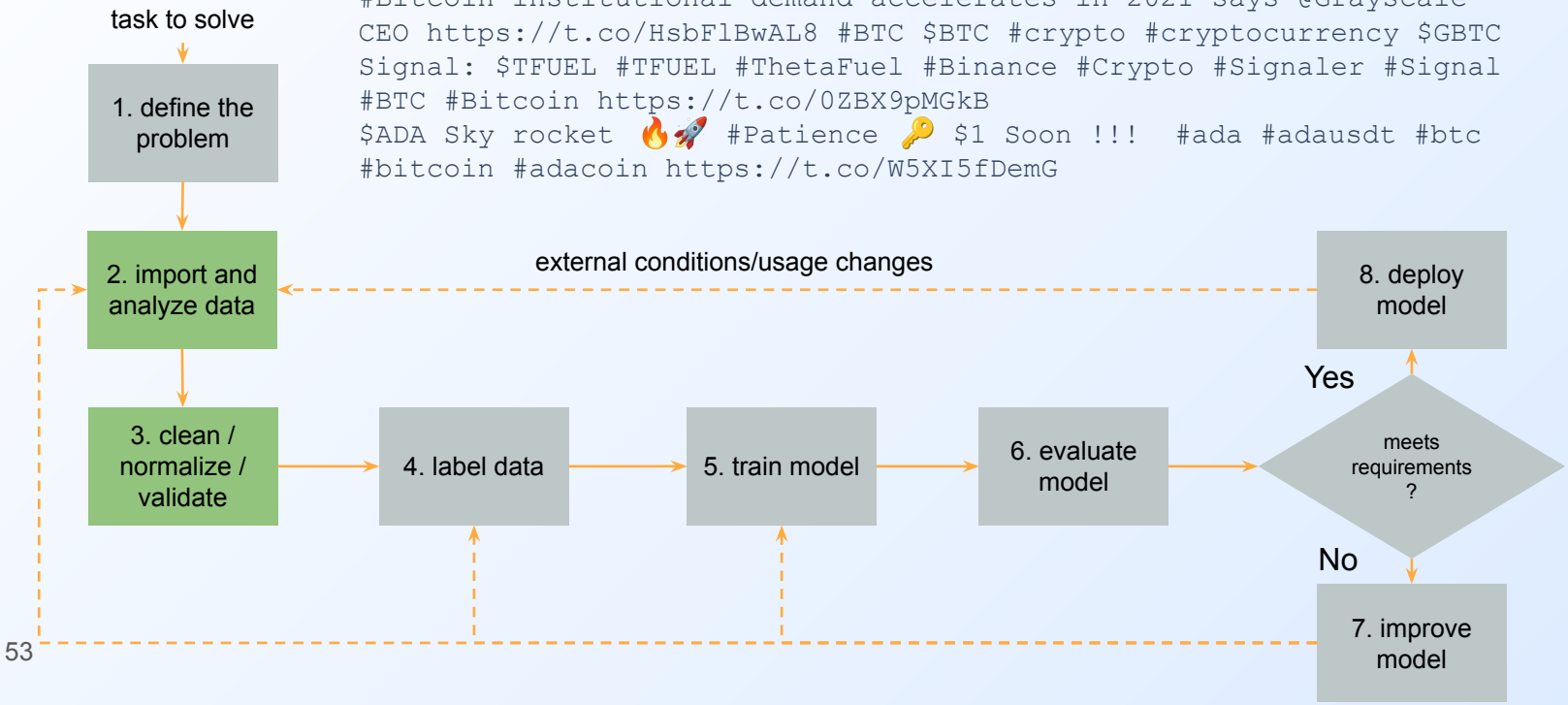
Solution - ML model for Named Entity Recognition

Detect all entities for cryptocurrency name, price and organization with over 90% precision and recall.



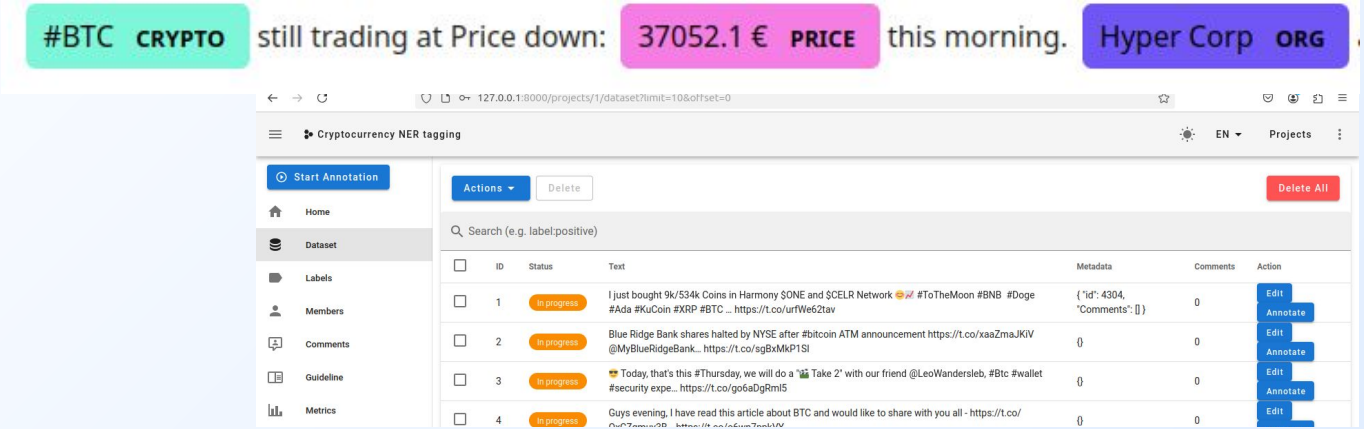
Import, analyze and clean data

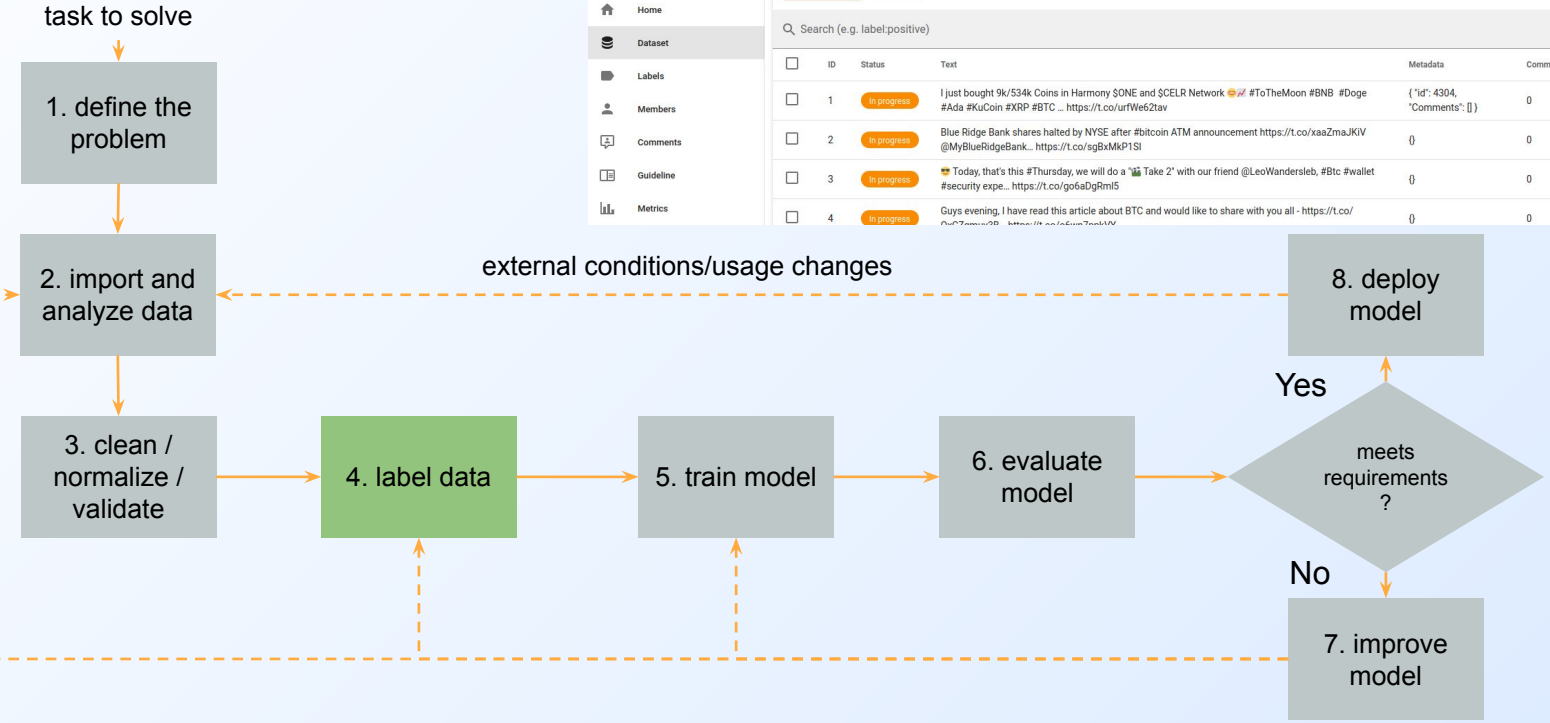

```
https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

$BTC A big chance in a million! Price: \4898631.0 (2021/02/11 08:45)
#Bitcoin #FX #BTC #crypto
#Bitcoin institutional demand accelerates in 2021 says @Grayscale
CEO https://t.co/HsbFlBwAL8 #BTC $BTC #crypto #cryptocurrency $GBTC
Signal: $TFUEL #TFUEL #ThetaFuel #Binance #Crypto #Signaler #Signal
#BTC #Bitcoin https://t.co/0ZBX9pMGkB
$ADA Sky rocket 🔥🚀 #Patience 🔑 $1 Soon !!!  #ada #adausdt #btc
#bitcoin #adacoin https://t.co/W5XI5fDemG
```



Label data

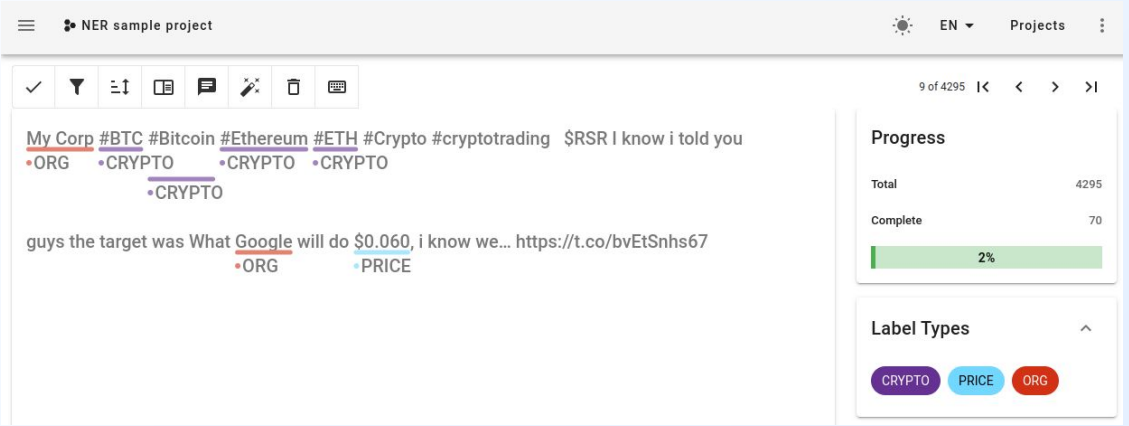
doccano





Label data

doccano



task to solve

1. define the problem

2. import and analyze data

3. clean / normalize / validate

4. label data

5. train model

6. evaluate model

meets requirements ?

Yes

8. deploy model

external conditions/usage changes

No

7. improve model

```
[
 {
   "id":4304,
   "text":"I just bought 9k Coins in Harmony $ONE and $CELR Network 😊📈 #BTC",
   "Comments":[],
   "label":[[60,64,"CRYPTO_NAME"]]
 },
 {
   "id":2,
   "text":"Blue Ridge Bank shares halted by NYSE after #bitcoin ATM announcement",
   "Comments":[],
   "label":[[0,15,"ORGANIZATION"],[33,37,"ORGANIZATION"],[44,52,"CRYPTO_NAME"]]
 },
...
]
```

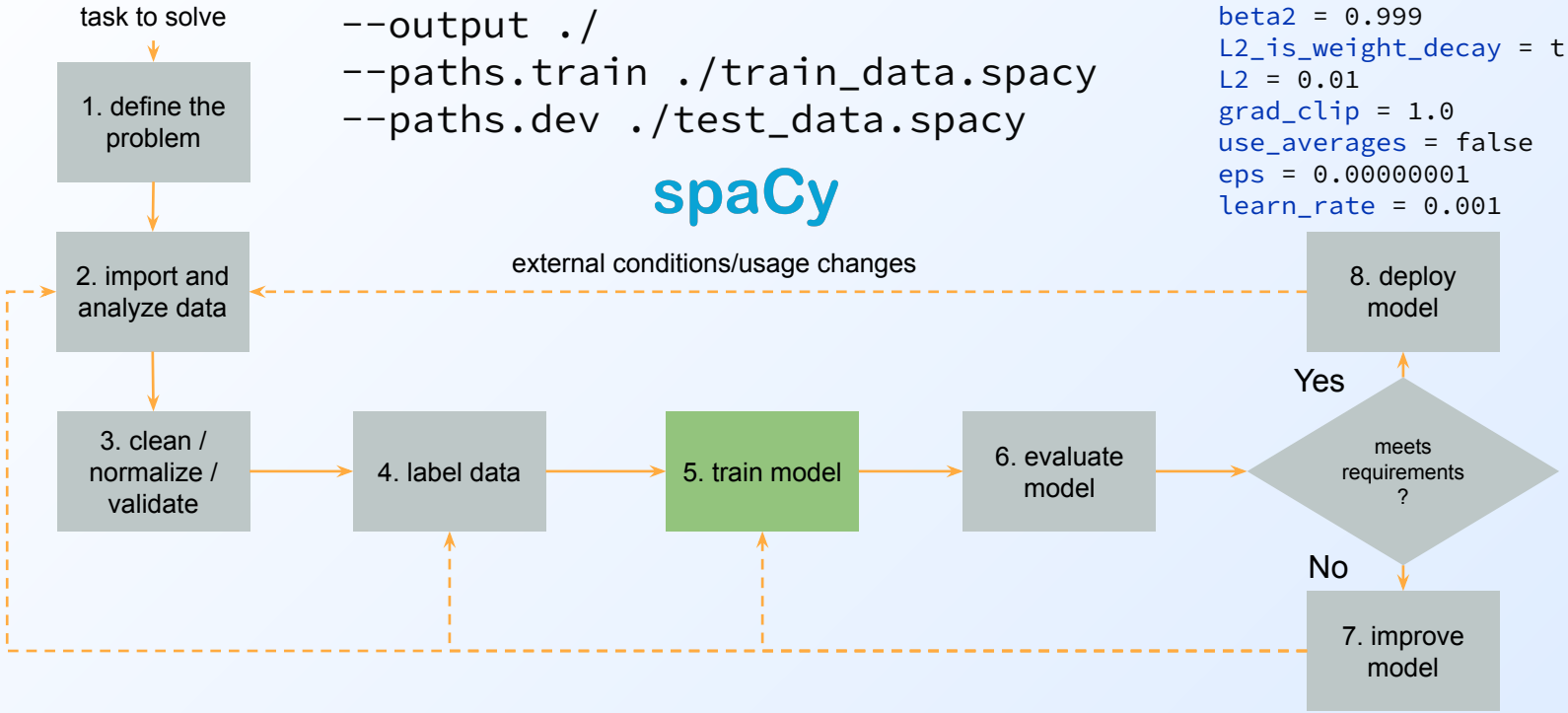
```
from spacy.tokens import DocBin
from tqdm import tqdm

doc_bin = DocBin()

for training_row  in tqdm(training_data['annotations']):
    text = training_row['text']
    labels = training_row['entities']
    doc = nlp.make_doc(text)
    ...
    doc_bin.add(doc)
```

Train model


```
python -m spacy train config.cfg
--output ./
--paths.train ./train_data.spacy
--paths.dev ./test_data.spacy
```

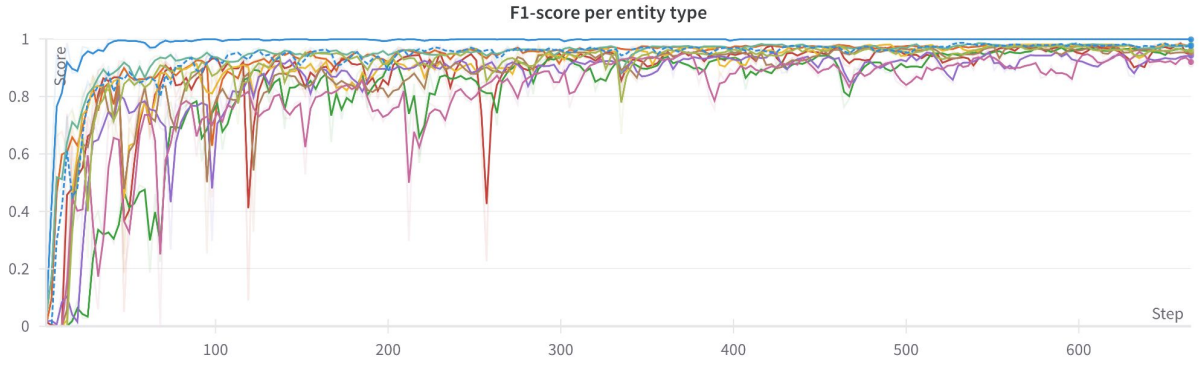
spaCy

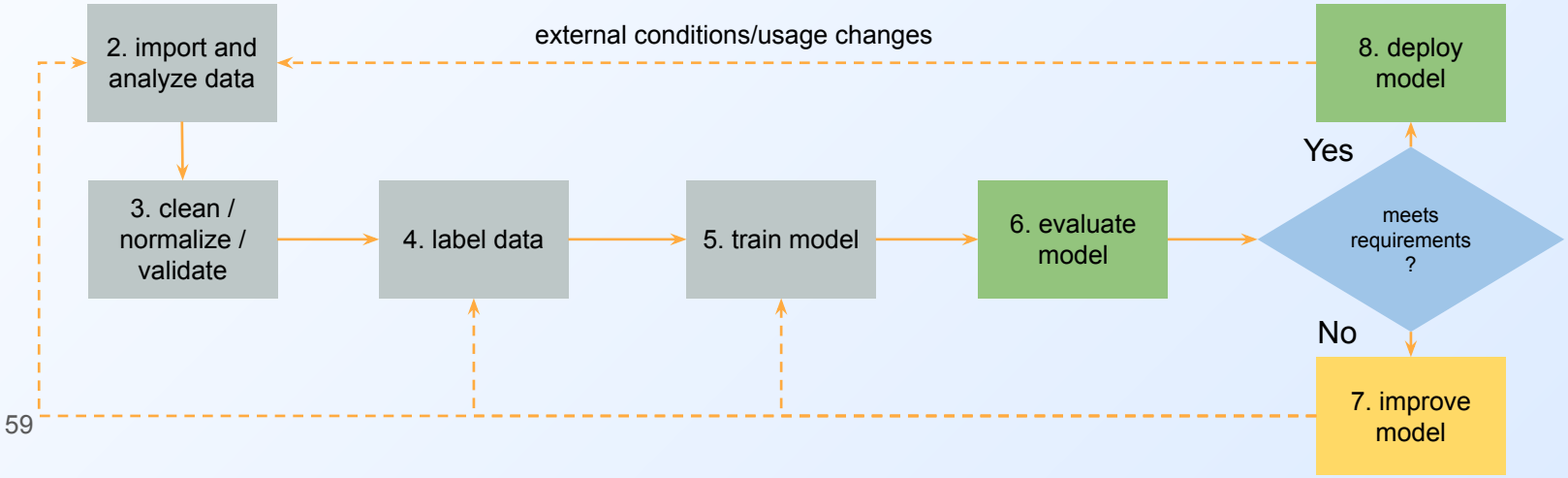
```
[training.optimizer]
@optimizers = "Adam.v1"
beta1 = 0.9
beta2 = 0.999
L2_is_weight_decay = true
L2 = 0.01
grad_clip = 1.0
use_averages = false
eps = 0.00000001
learn_rate = 0.001
```



Evaluation

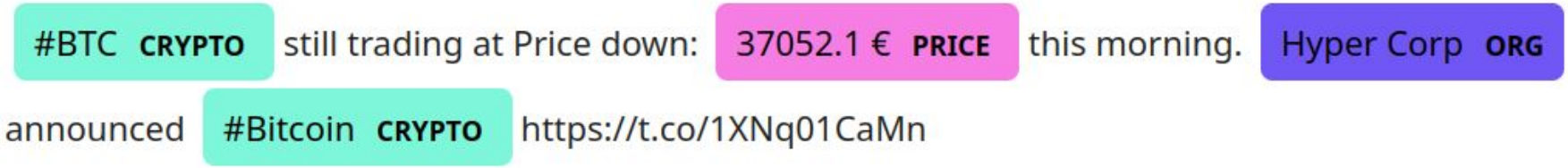
```
665: 0.99917 (0.99882) ents_per_type.Birth Date.f
665: 0.9796 (0.98102) ents_per_type.Function.f
665: 0.97913 (0.98148) ents_per_type.Court Town.f
665: 0.97886 (0.97919) ents_per_type.City of Birth.f
665: 0.97625 (0.97735) ents_per_type.Person.f
665: 0.96159 (0.96314) ents_per_type.Signing Authority
665: 0.95969 (0.95868) ents_per_type.Company.f
665: 0.95383 (0.95413) ents_per_type.Registration Num
665: 0.94575 (0.94231) ents_per_type.Company City.f
665: 0.94197 (0.94268) ents_per_type.Foreign Country.f
665: 0.92037 (0.91071) ents_per_type.Scientific Title.f
Press CMD+C or CTRL+C to copy this data
```



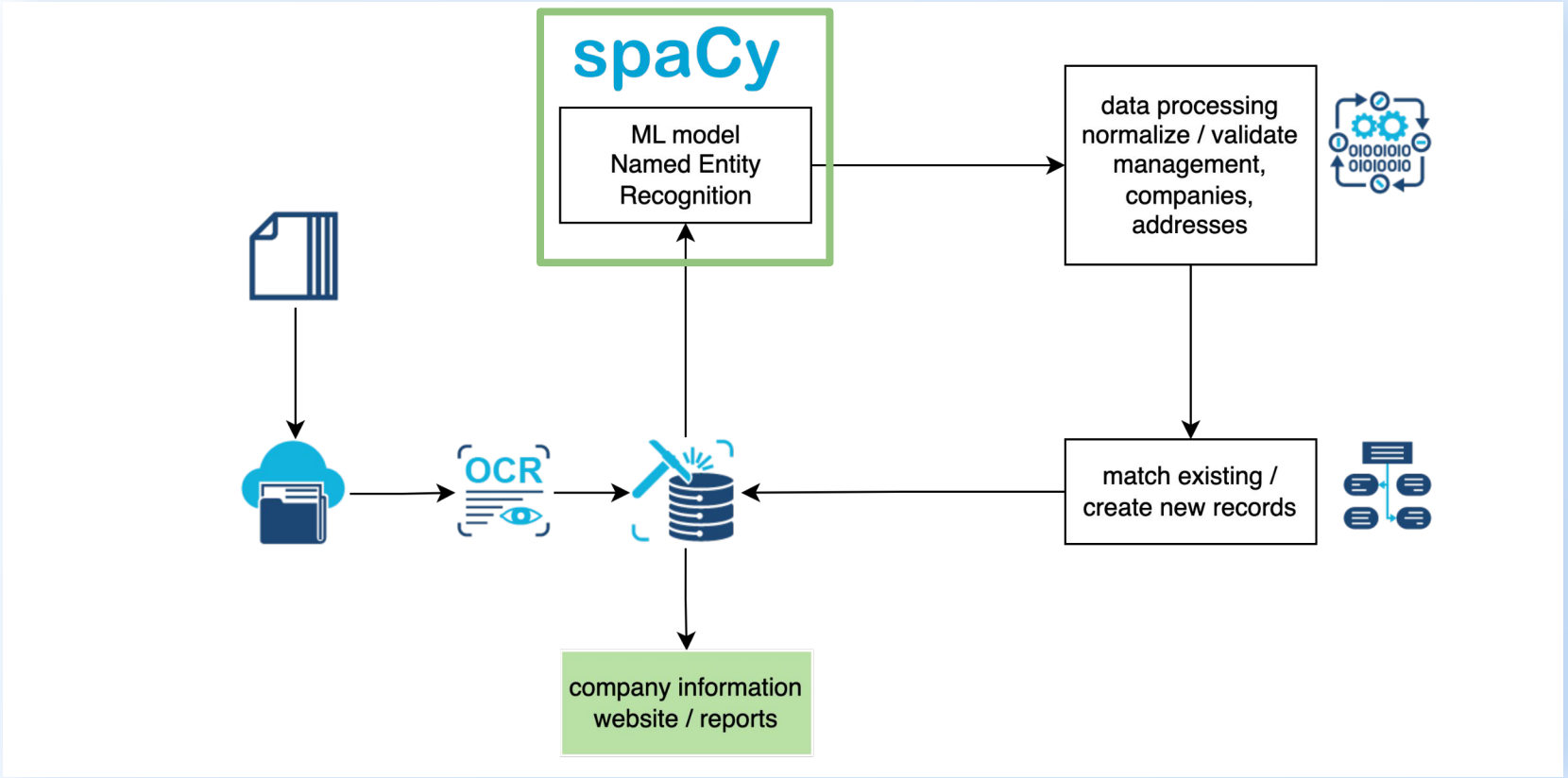


```
1  nlp_ner = spacy.load("model-best")
2
3  doc = nlp_ner("#BTC still trading at Price down: 37052.1 € this morning. Hyper Corp
    announced #Bitcoin https://t.co/1XNq01CaMn")
4
5  colors = {"PRICE": "#F67DE3", "CRYPTO": "#7DF6D9", "ORG":"#7156F6"}
6  options = {"colors": colors}
7
8  spacy.displacy.render(doc, style="ent", options= options, jupyter=True)
```

[46]



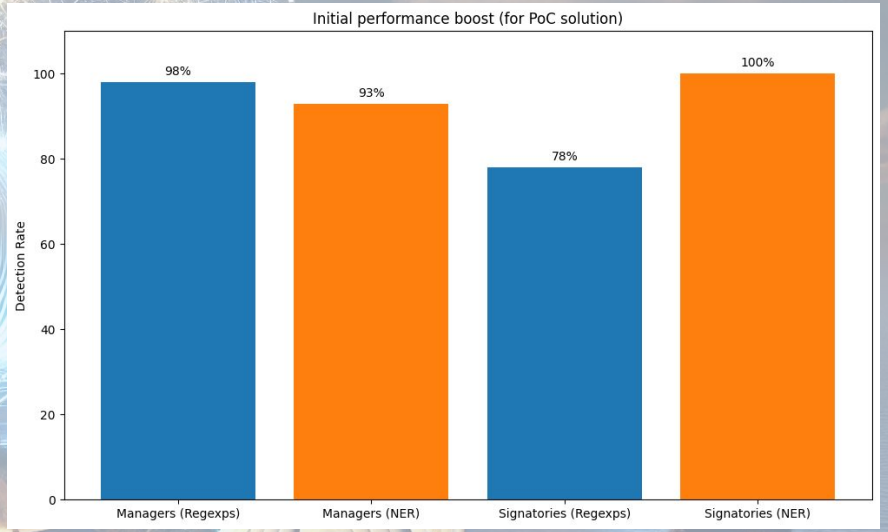
spaCy NER - architecture



Challenges and lessons learned

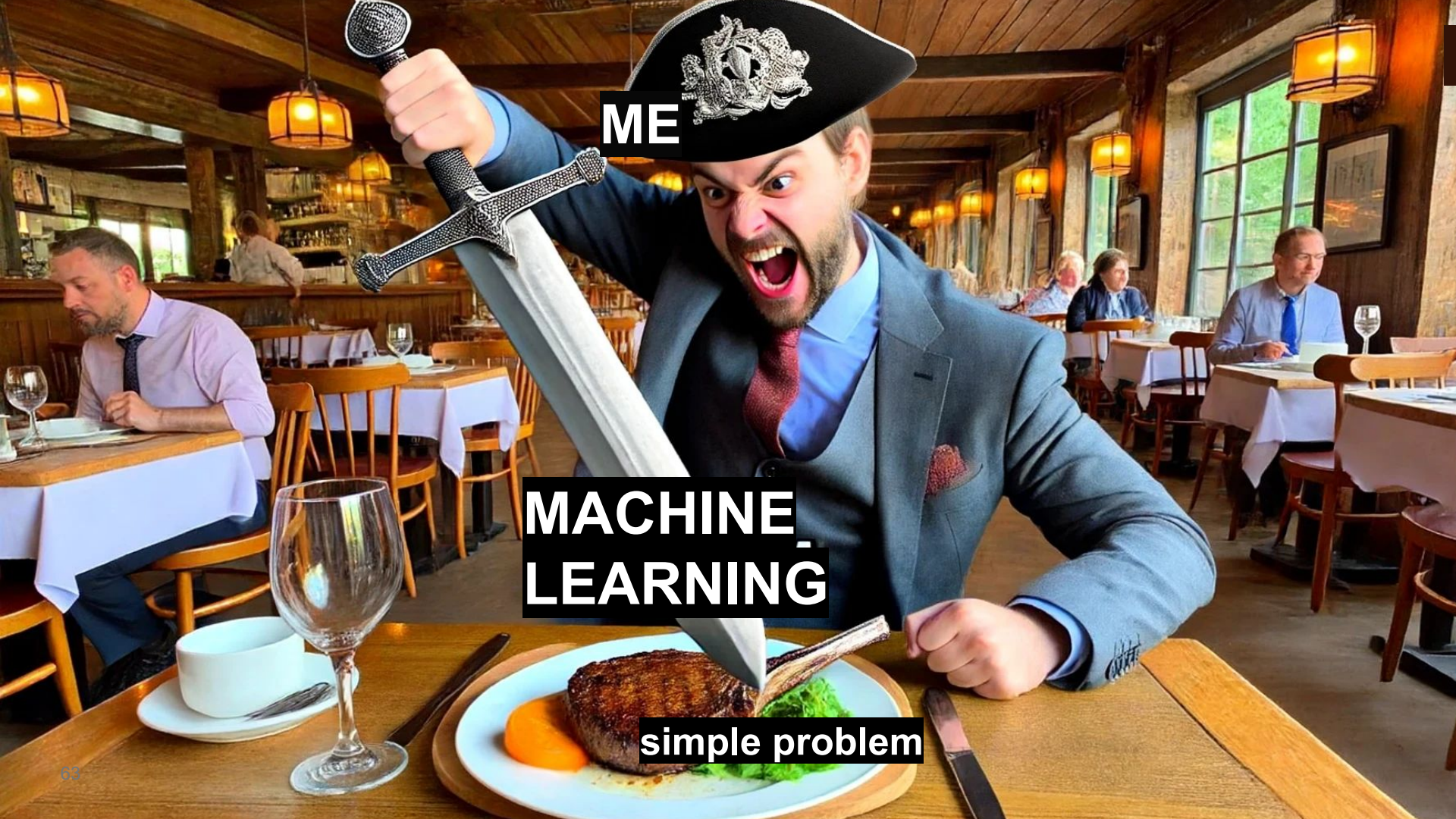
Project 2 - Document parser

1. Even 20% more recognized cases
2. No control of single conditions / need to retrain model
3. Less time spent on maintenance
   - tagging: 700 documents per task ~ 40h of work
   - model preparation and train: ~40h of work



Github sample

https://github.com/mzarnecki/train-ner-model-with-spacy


ME
MACHINE LEARNING
simple problem

ConFoo.CA
DEVELOPER CONFERENCE

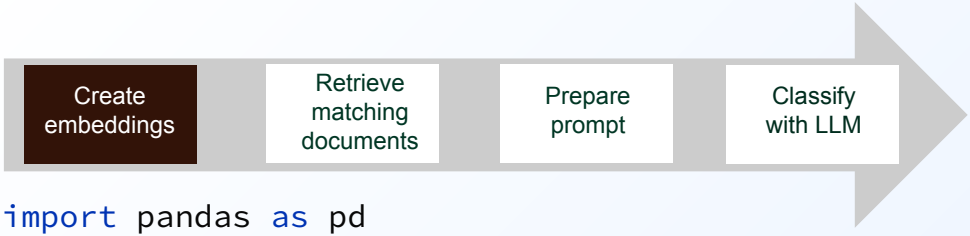
leave your feedback

Thank you for watching

Michał Żarnecki

Contact: michal@zarnecki.pl
github: https://github.com/mzarnecki
LinkedIn: https://www.linkedin.com/in/micha%C5%82-%C5%BCarnecki-47219355/

Model	Context Window Size	Pricing Details
GPT-4	8,192 or 32,768 tokens	OpenAI offers GPT-4 with two context window options: 8,192 tokens and 32,768 tokens. Pricing varies based on the context window size and usage. For detailed pricing, refer to OpenAI's official documentation.
Claude 3	Up to 1 million tokens	Anthropic's Claude 3 model features a context window of up to 1 million tokens, accommodating extensive inputs. Pricing details can be found on Anthropic's official website.
Llama 3.1 (405B model)	128,000 tokens	Meta's Llama 3.1 405B model offers a context window of 128,000 tokens. As an open-source model, it is available for use without licensing fees.
DeepSeek-V3	128,000 tokens	DeepSeek's V3 model provides a context window of 128,000 tokens. The model is open-source, allowing for free use and modification.
Google Gemini 1.5	Up to 2 million tokens	Google's Gemini 1.5 model supports a context window of up to 2 million tokens, enabling the processing of extensive data. Pricing information is available through Google's official channels.
Mistral 7B	Not specified	Mistral AI's Mistral 7B model is open-source and free to use. Specific context window size details are not provided in the available sources.



LangChain

PostgreSQL
pg_vector

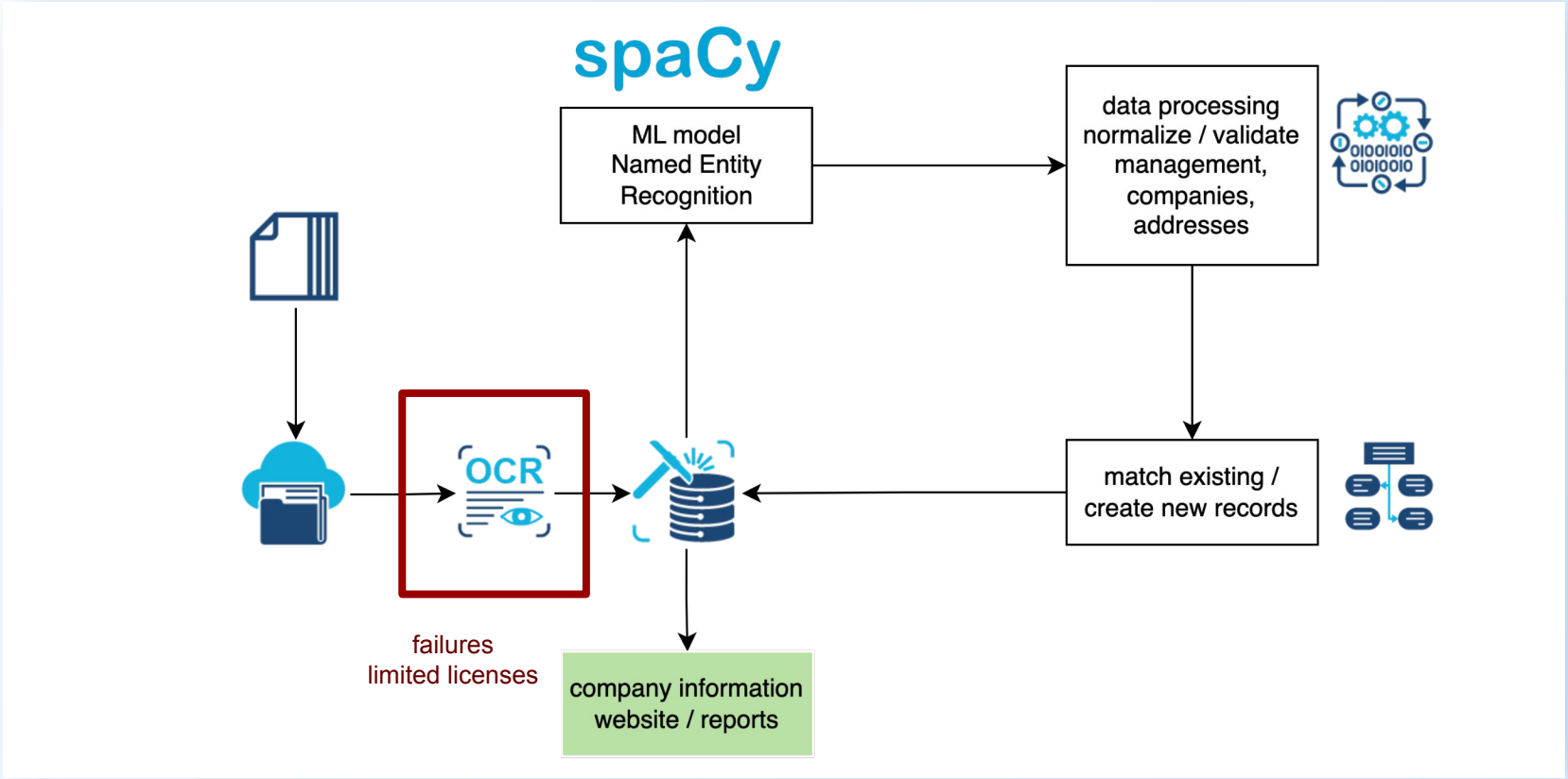
```
import pandas as pd

# sample documents
documentsDf = pd.read_csv('data/industry_guidelines.csv', quotechar='"', header=None)
documents = documentsDf.iloc[0].to_list()

------------------------------
# sample documents
documents = [
    """
Industry code: 01

covers the two activities of production of crop products and production of animal products.
It also includes organic farming and the cultivation of genetically modified crops...
does not include:
- processing of agricultural products other than preparation
...
```

Create embeddings | **Retrieve matching documents** | Prepare prompt | Classify with LLM

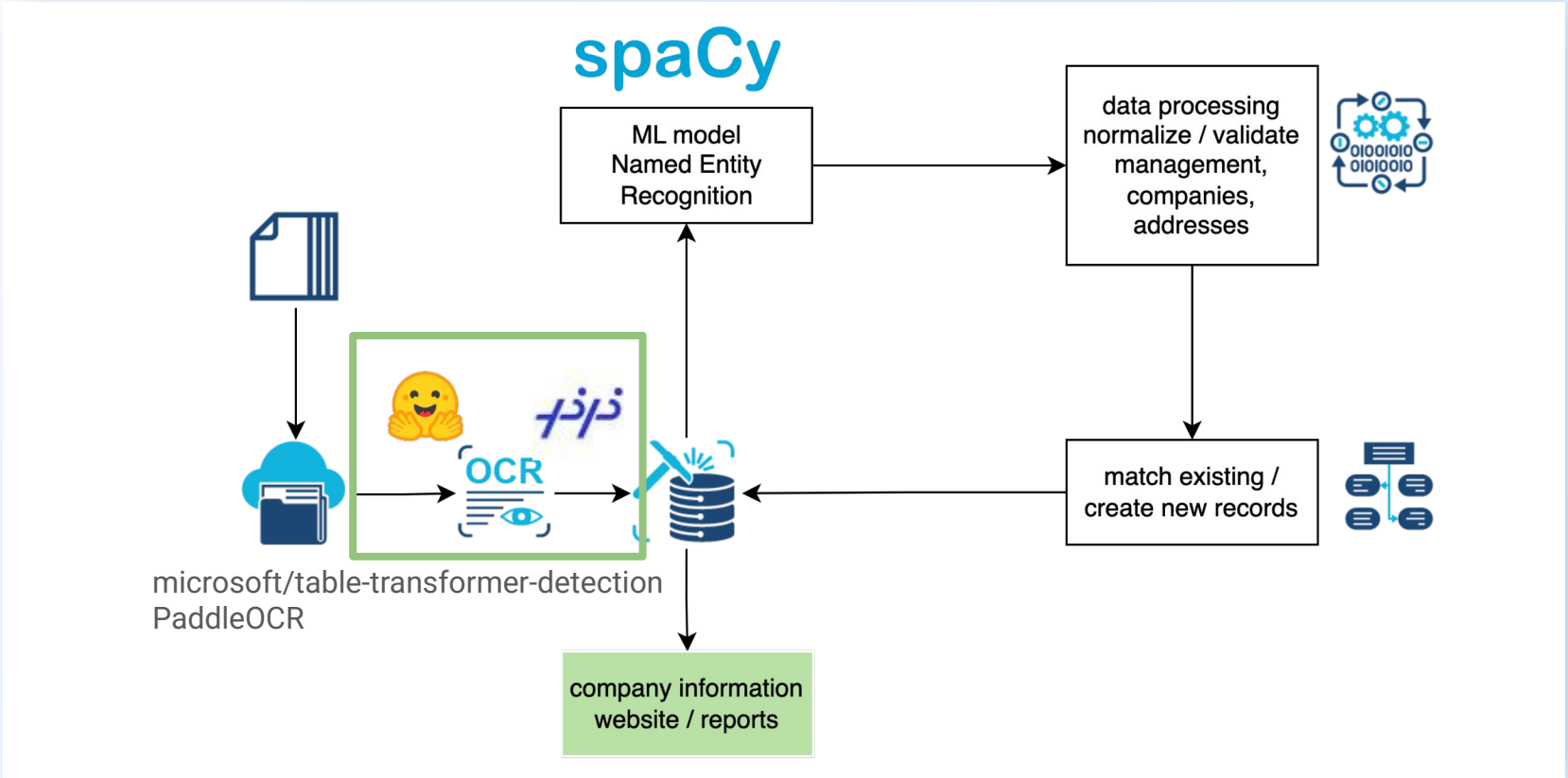
```
# company description to classify that will be used as query text
query = """
Company is producing crop products.
"""
# retrieve top_k similar documents using cosine similarity score
retriever = faiss.as_retriever(search_type='similarity', search_kwargs={"k": 3})
selected_documents = retriever.invoke(query)
```

```
[
        0.12233  , -0.40965  ,  0.26148  , -0.51868  , -0.60459  , -0.84374  , -0.11416  ,  0.89577  , -0.23875  , -0.73708  ,
        -1.0004  , -0.098071 ,  0.18746  ,  0.14397  ,  0.47586  , 0.40886  ,  0.87898  ,  0.18226  , -0.32419  ,  0.54662  ,
        -0.46989  ,  0.90234  , -0.34768  ,  0.10074  ,  0.35826  , 0.31507  , -0.65955  , -0.042604 ,  0.010843 ,  0.050955 ,
        0.31938  , -0.10816  ,  0.22426  , -0.22073  ,  0.15903  , 0.16886  , -0.035559 ,  0.35242  ,  0.73245  ,  0.29862  ,
        -0.26791  , -0.09938  , -1.1552  ,  0.14245  ,  0.28243  , 0.50401  ,  0.27086  , -0.11168  , -0.2424  ,  0.29003  ,
        -1.3194  , -0.19863  ,  0.27979  ,  0.12743  , -0.5687  , -1.1821  ,  0.15884  ,  0.0027925,  0.10209  , -0.10743  ,
        0.084864 ,  0.5932  ,  0.96694  ,  0.79069  ,  0.59454  , -0.061516 ,  0.25304  , -0.0038544,  0.17405  ,  0.08916  ,
        -0.82945  , -0.56059  , -0.29723  ,  0.39116  ,  0.14941  , -0.22372  , -0.3033  ,  0.12273  , -0.58946  ,  0.53444  ,
        -0.39859  ,  0.37937  , -0.2814  , -0.19535  ,  0.35982  , 0.24495  , -0.15736  , -0.45703  ,  0.71712  ,  0.75183  ,
        -0.43934  , -0.49642  ,  0.2353  , -0.039732 , -0.47302  , -0.15553  , -0.11614  ,  0.59744  ,  0.15876  , ...
]
```

spaCy NER - architecture



spaCy NER - architecture

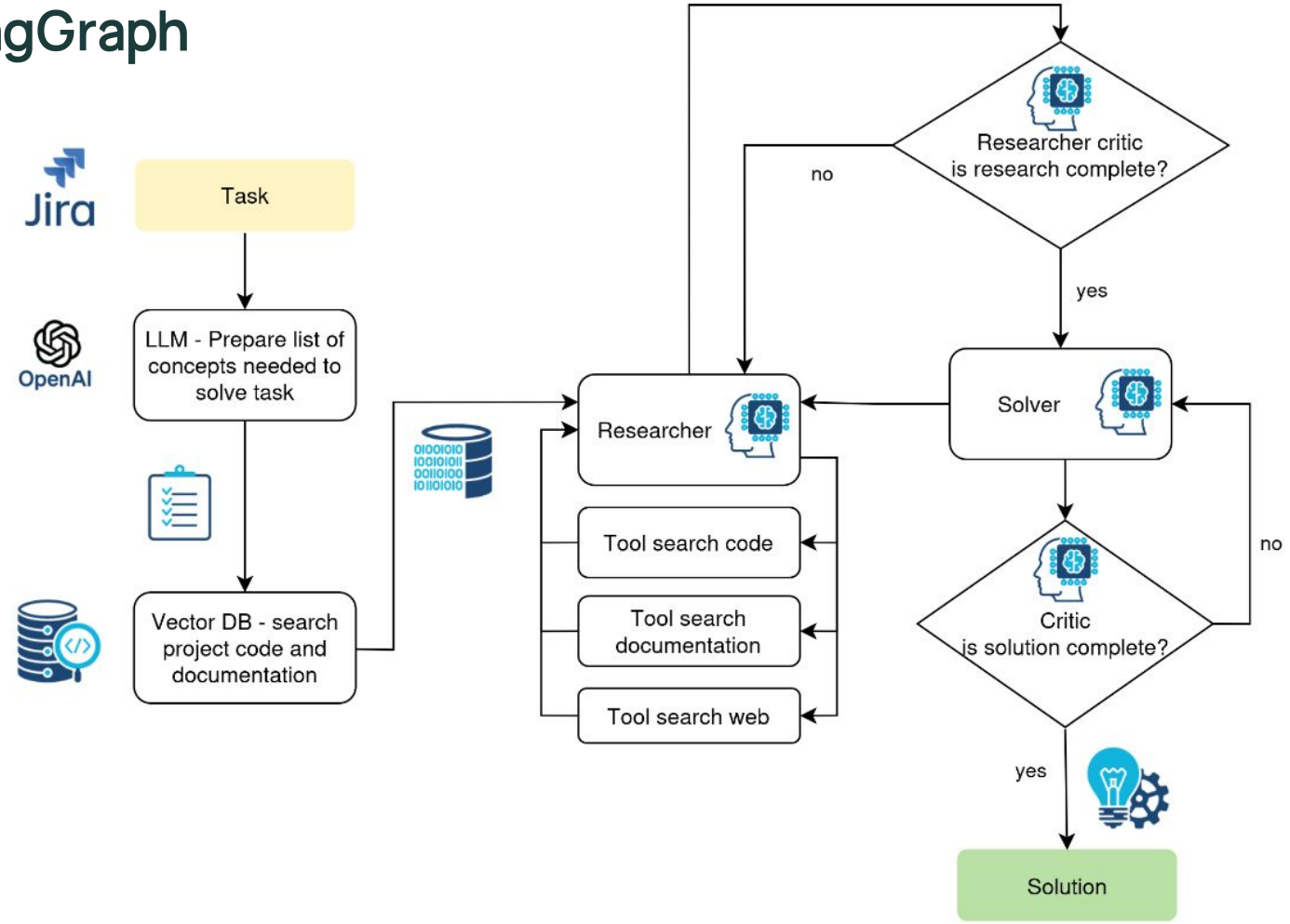


Solution prompt

ConFoo.CA
DEVELOPER CONFERENCE

```
When addressing questions:

    Cite specific sections of the code and their corresponding locations
    Clarify the rationale behind the implementation choices
    Offer suggestions for improvements if applicable
    Keep the overall architecture and goals of the codebase in mind
    Always base your answers on the provided code context
    Adopt a methodical, step-by-step approach to problem-solving
    Use retriever tools to search codebase

  Using these guidelines, create a solution for the ticket described below.
  You can also be provided with a proposed solution for the code to
evaluate.
  Write working code solution, add explanation and additional instructions
  related to using this code if needed. \n\n
  {ticket}
```

LangGraph


Jira
Task
OpenAI
LLM - Prepare list of concepts needed to solve task
Vector DB - search project code and documentation
Researcher
Tool search code
Tool search documentation
Tool search web
Researcher critic is research complete?
no
yes
Solver
Critic is solution complete?
no
yes
Solution