

# Replace complex business logic with Machine Learning models

Michał Żarnecki



CompanyHouse

## 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



UCZELNIA LUDZI CIĘKAWYCH

## 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"*



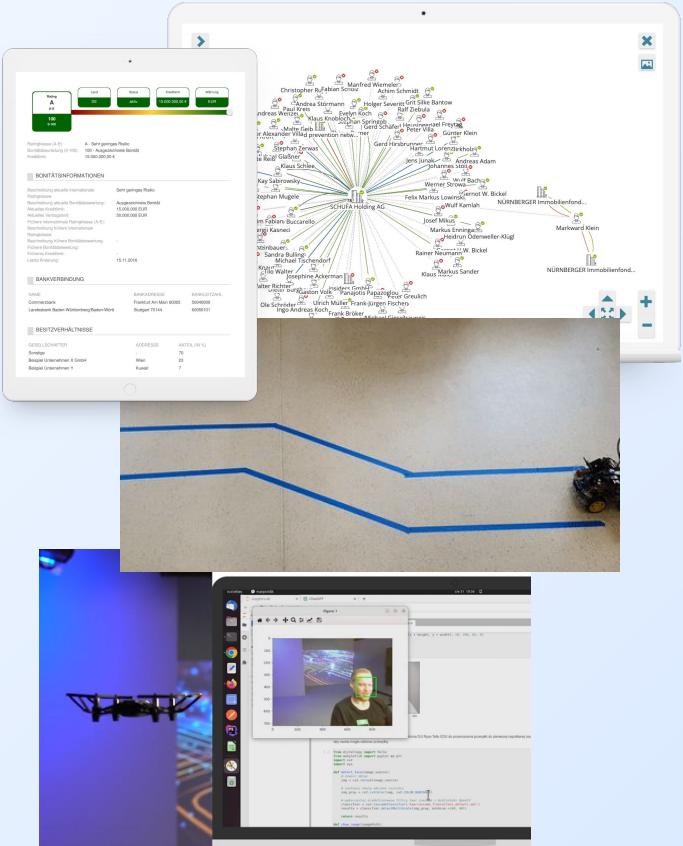
GitHub

2

<https://github.com/mzarnecki>

Medium

<https://medium.com/@brightcode>





“  
*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



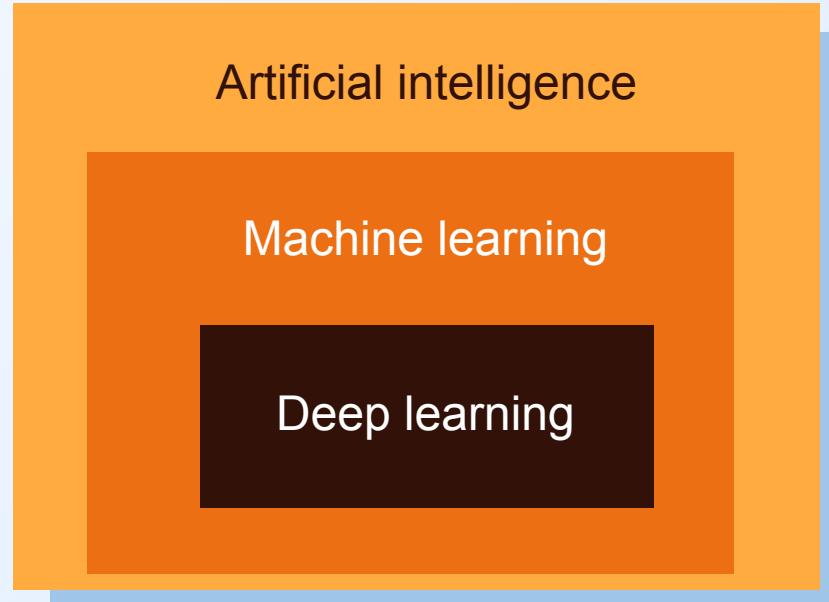
Project 2 - Similar companies finder



Project 3 - Documents parser

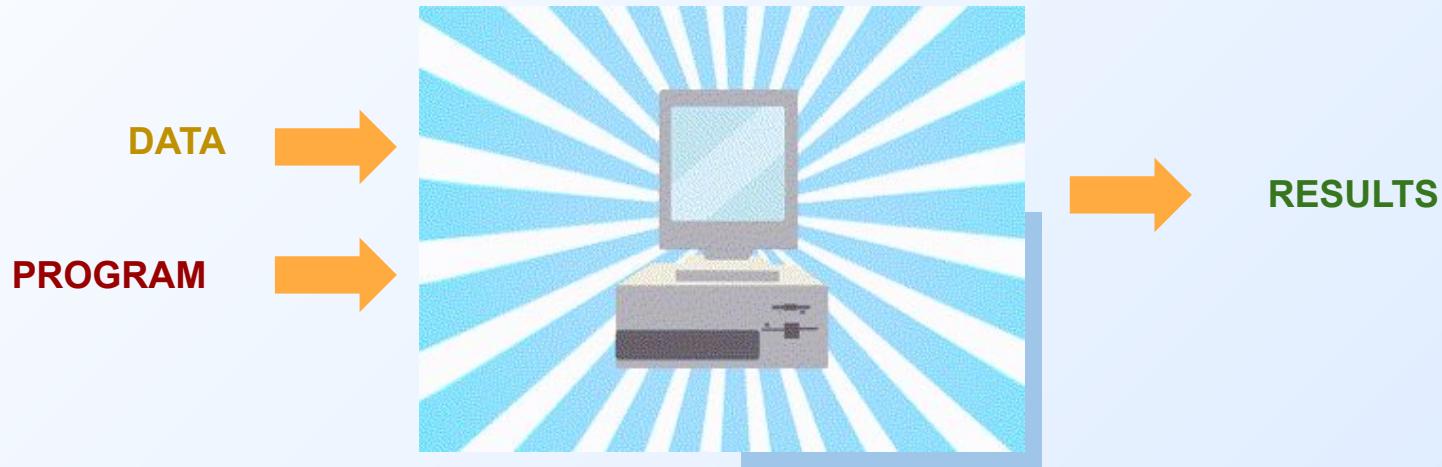
# Machine learning

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



# ML vs "traditional" programming

Traditional approach



# ML vs traditional" programming

Using machine learning



DATA

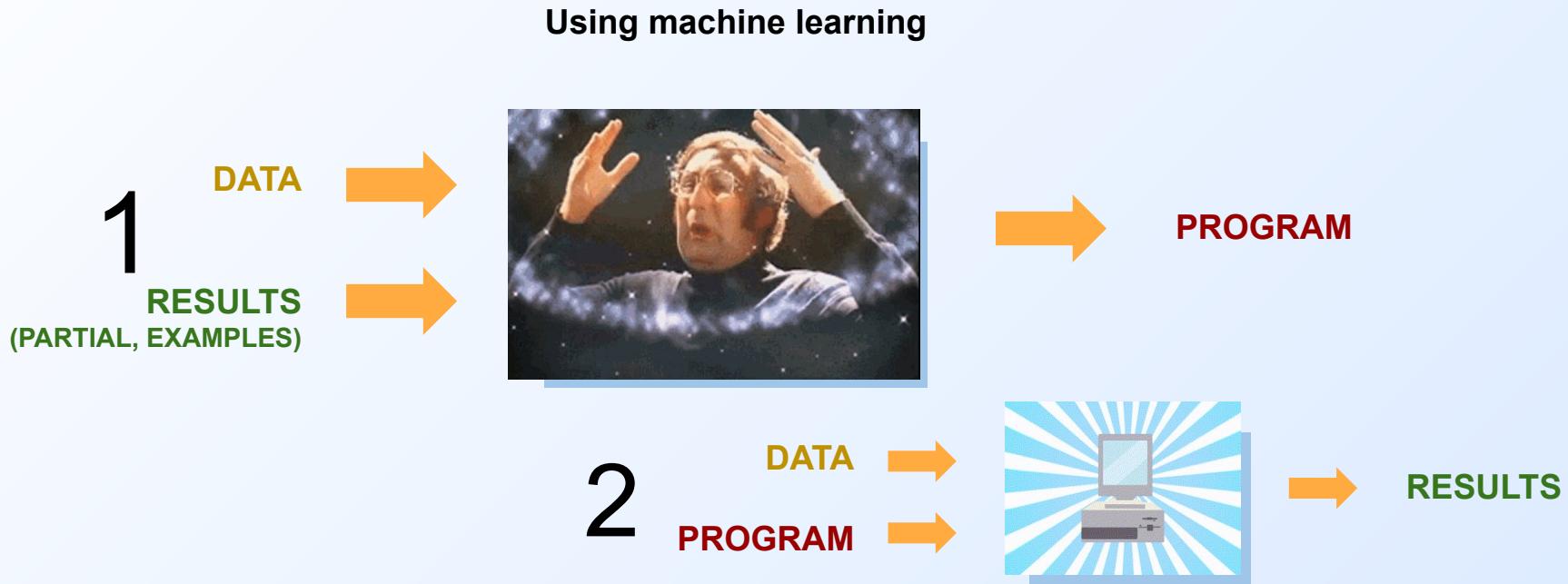


RESULTS  
(PARTIAL, EXAMPLES)

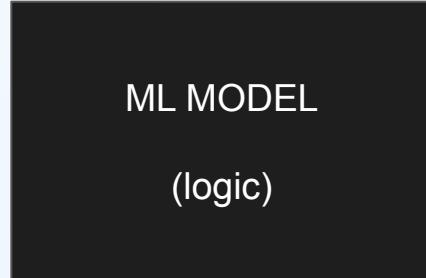


PROGRAM

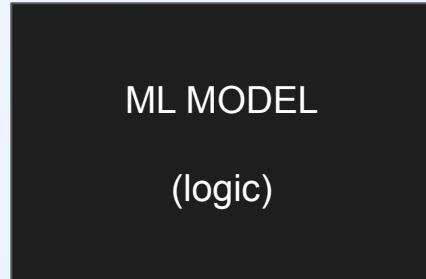
# ML vs traditional" programming



# What is the goal of machine learning?



**dog**



**not a dog**

# Why we need to replace logic with ML?

-  Avoid endless if-else spaghetti
-  Skip hardcoding every single scenario
-  Avoid learning all possible scenarios
-  Spend less time debugging
-  Focus on innovation, not maintenance



# What can be supported with ML?

Localization

Business logic

Document processing

Code review

Bug detection/tracking

Requirements analysis

The screenshot shows a detailed company profile page. At the top, there's a search bar and a user account section. Below that, the company's name and status ('Active') are displayed, along with its incorporation date (23.10.1980) and last change (03.06.2024). The page is divided into several sections: 'Management' (listing authorized executives like Managers Wolfgang and Signatories Hans), 'Timeline' (a horizontal timeline from 1980 to 2024 with colored segments for different periods), 'Network' (a graphical network diagram showing connections between current and former employees), and 'Similar Companies' (a list of companies with similarity scores: 83%, 83%, 60%, 60%). On the right side, there's a 'Chatbot' section with a person holding a phone and a 'Start chat now' button, and a 'News' section listing recent updates.

Customer service

Knowledge management

Forecasting

# How to pick correct ML algorithm?

ConFoo.ca  
DEVELOPER CONFERENCE



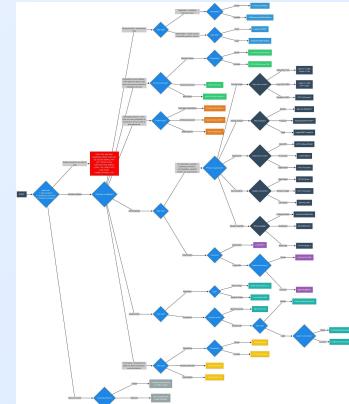
**Hugging Face**

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>

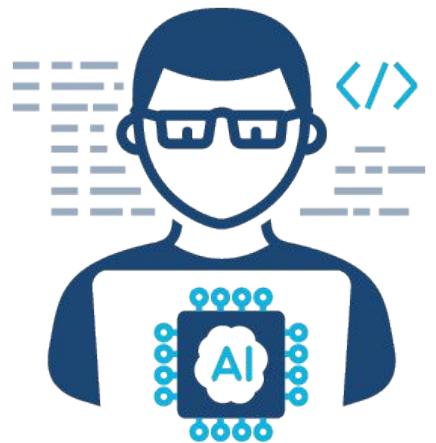


# Codebase expert agent

1

Project 1

**Codebase expert agent**



*generate code with AI*

2

Project 2

**Similar company finder**

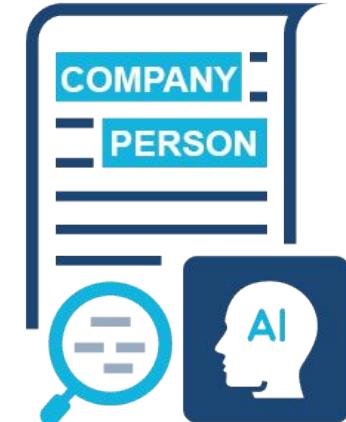


*use LLMs*

3

Project 3

**Documents parser**



*train dedicated model*

# Codebase expert agent

## AI PROJECT TICKET SOLVER

AI is capable of integrating new features into project and solving issues based on ticket description.

 Provide ticket details

ticket subject

```
Error: Cannot access protected property common\models\product\model\group\CurrentBalancesAndAnnualReportsForCompany::$company in
/usr/share/nginx/html/build/companyhouse/1828/common\models\company\annualReport\helper.php:40 Stack trace: #0
/usr/share/nginx/html/build/companyhouse/1828/frontend/view/_overview/annualReports/_annual-reports-rows.php(31):
common\models\company\annualReport\AnnualReportHelper::getU('') #1 /usr/share/nginx/html/build-
companyhouse/1828/vendor/yiisoft/yii2/base/view.php(347): require('...')
#2 /usr/share/nginx/html/build-
companyhouse/1828/vendor/yiisoft/yii2/base\View->renderFile() #3 /usr/share/nginx/html/build-
companyhouse/1828/vendor/yiisoft/yii2/base\View.php(156): yii\base\View->renderFile() #4 /usr/share/nginx/html/build-
companyhouse/1828/frontend/widgets/tiles\tables\CompanyAnnualReportTitleTable.php(29): yii\base\View->render() #5
/usr/share/nginx/html/build-companyhouse/1828/frontend\widgets/tiles\tables\CompanyAnnualReportTitleTable.php(194):
frontend\widgets\tiles\tables\CompanyAnnualReportTitleTable->renderAnnualReportsTable() #6 /usr/share/nginx/html/build-
companyhouse/1828/frontend/renderers\tile\balance\CompanyBalanceOverviewViewFile.php(51):
frontend\widgets\tiles\tables\CompanyAnnualReportTitleTable->render() #7 /usr/share/nginx/html/build-
companyhouse/1828/shared\frontend\renders\tile\AbstractViewFile.php(191): Frontend\renders\tile\balance\CompanyBalanceOverviewViewFile-
>getContent() #8 /usr/share/nginx/html/build-companyhouse/1828/frontend\renders\tile\balance\CompanyBalanceOverviewViewFile.php(45):
shared\frontend\renders\tile\AbstractViewFile->renderLocal() #9 /usr/share/nginx/html/build-
companyhouse/1828/shared\frontend\renders\tile\AbstractViewFile.php(190): Frontend\renders\tile\balance\CompanyBalanceOverviewViewFile-
>renderLocal() #10 /usr/share/nginx/html/build-companyhouse/1828/shared\frontend\renders\tile\AbstractRenderableByFileCollection.php(34):
shared\frontend\renders\tile\AbstractViewFile->render() #11 /usr/share/nginx/html/build-
companyhouse/1828/shared\components\services\fragmentCacheHandlerService.php(211):
```

URL

User

Select user type

Device type

Select device type

ticket image

 Drag and drop file here  
Limit 200MB per file

Browse files

Submit

### 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

```
1 namespace common\models\product\model\group;
2
3 class CurrentBalancesAndAnnualReportsForCompany
4 {
5     protected $company;
6
7     // Add a public getter method
8     public function getCompany()
9     {
10         return $this->company;
11     }
12 }
```

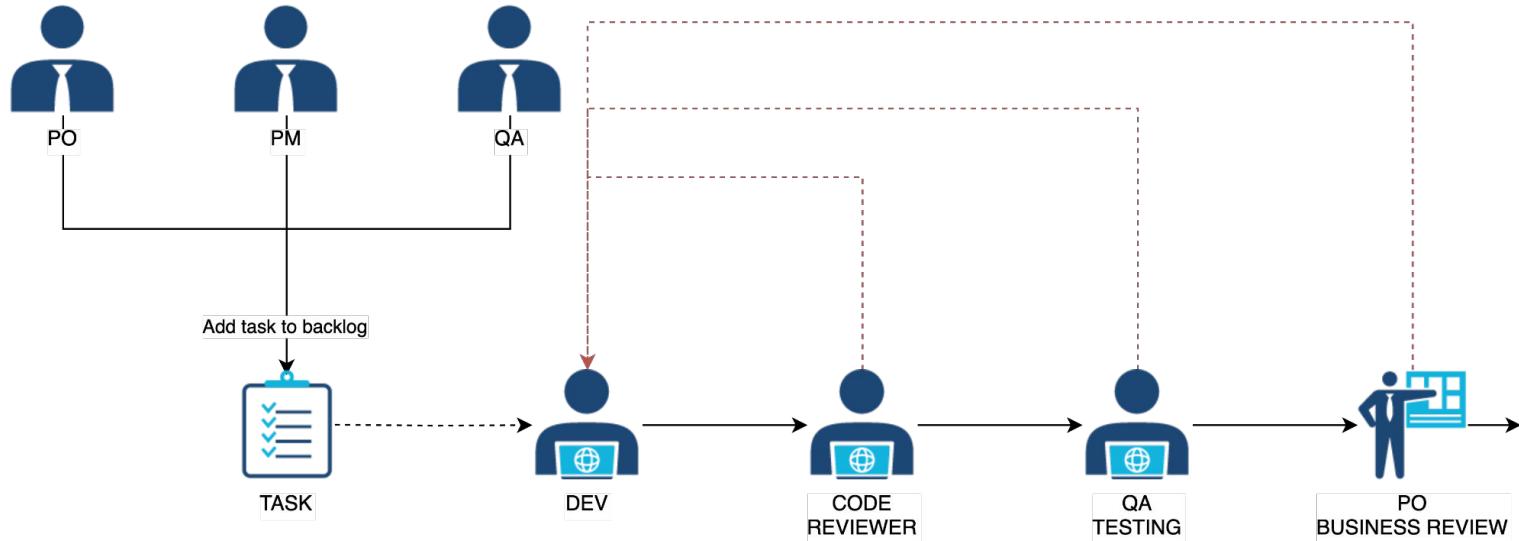


#### Step 2: Update `AnnualReportHelper.php` to Use the Getter

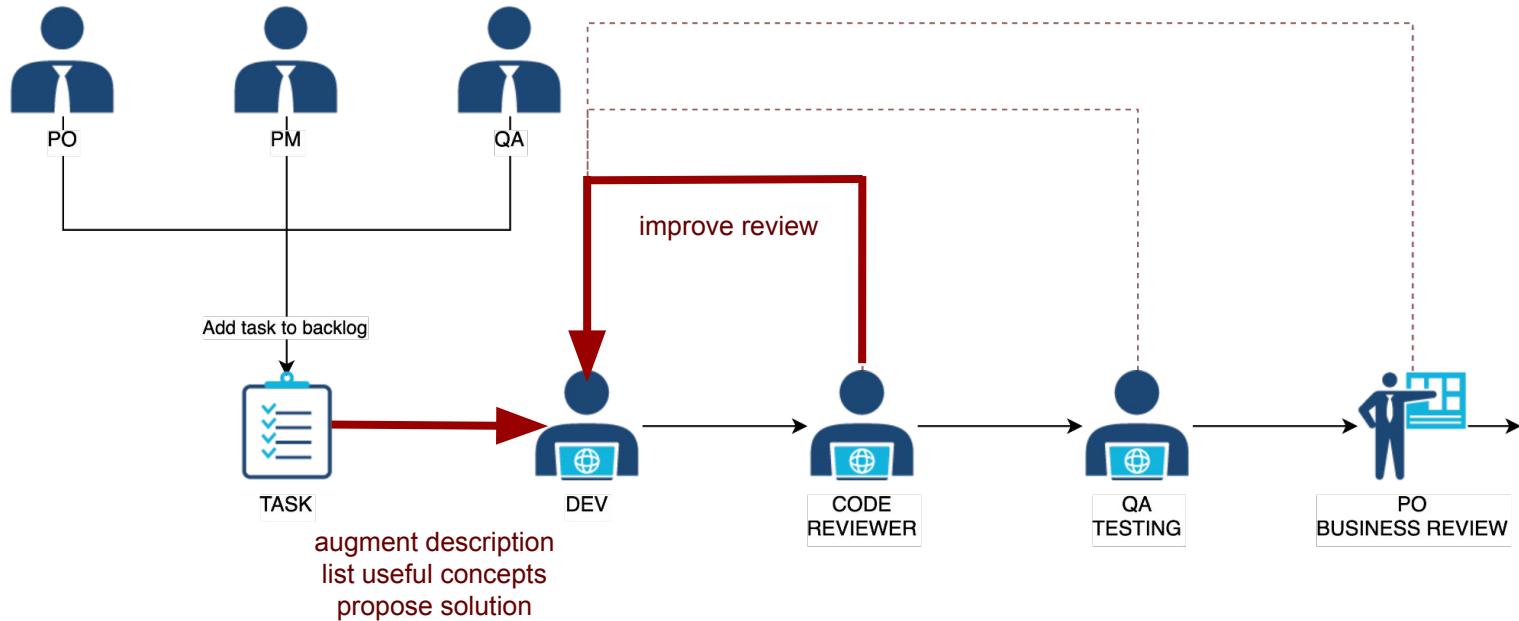
```
1 namespace common\models\company\annualReport;
2
3 use common\models\product\model\group\CurrentBalancesAndAnnualReportsForCompany;
4
```



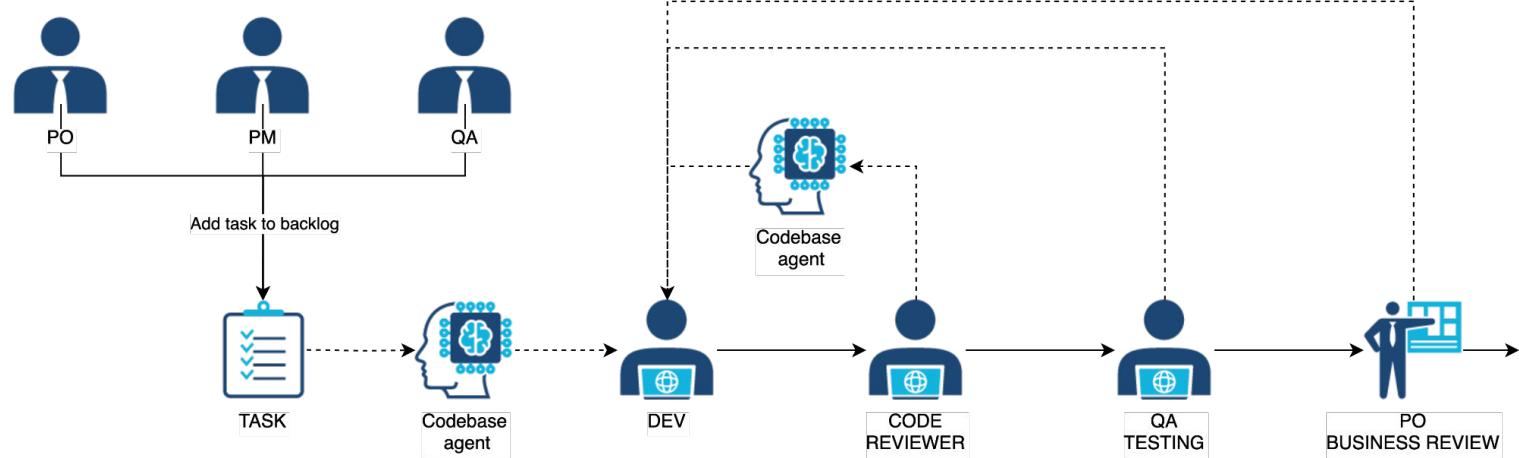
# Initial approach - workflow



# Initial approach - workflow

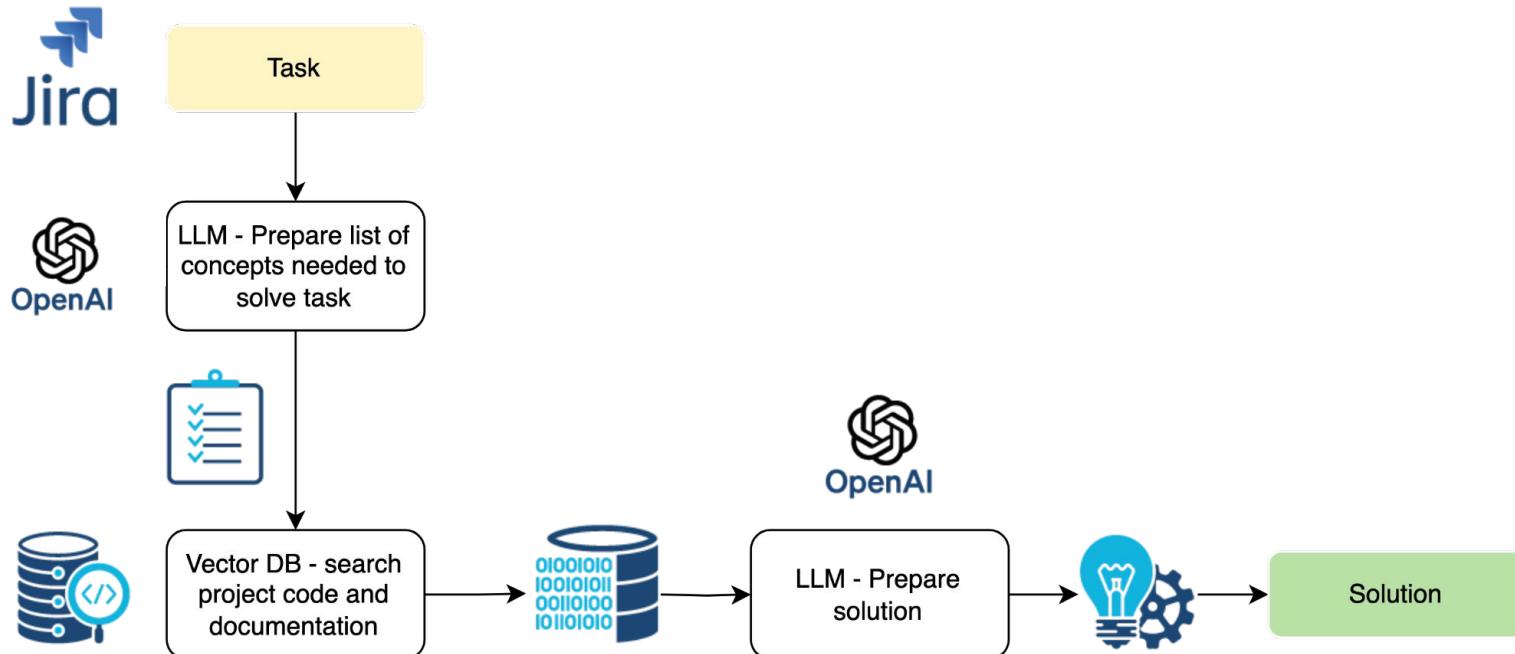


# Initial approach - workflow



Postgres + PG Vector

# PoC - architecture



# Example task

Person shown 2x on timeline

+ Add   @ Apps

General   PO   Zoho

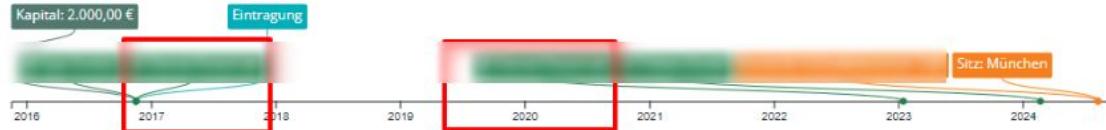
Description • Unsaved changes

The person is shown 2x on the timeline:

## Zeitleiste

### Historie

(1) Person shown 2x as CEO in timeline



### Todo

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

# Concepts prompt

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}`

# 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

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

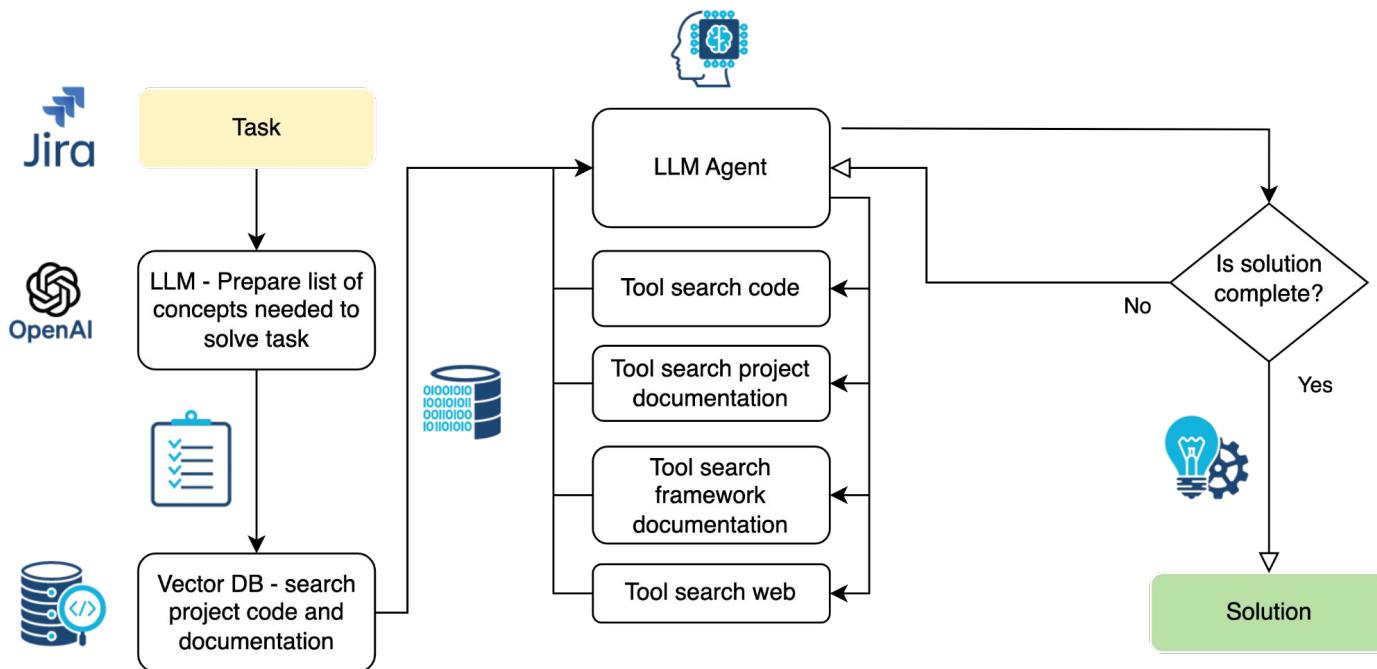
generated output:

## Proposed Solution:

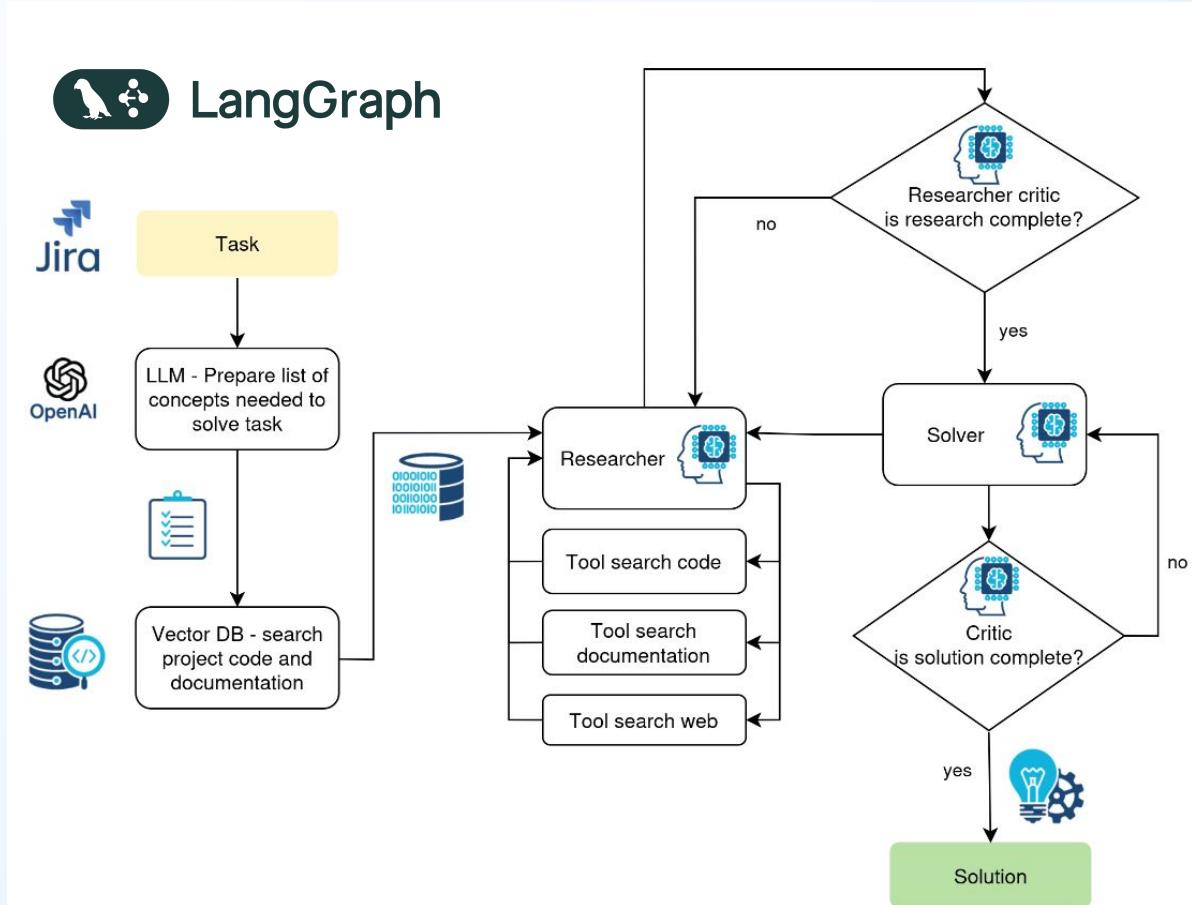
1. Identify Duplicate Entries:
  - o In both `addCompanyManagersData` and `addCompanySignatoriesData` methods, we need to ensure that duplicate entries are not added to the timeline. This can be achieved by maintaining a set of unique identifiers for each person and checking against this set before adding a new entry.
2. Modify the Code:
  - o 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](https://github.com/mzarnecki/ai-codebase-expert)

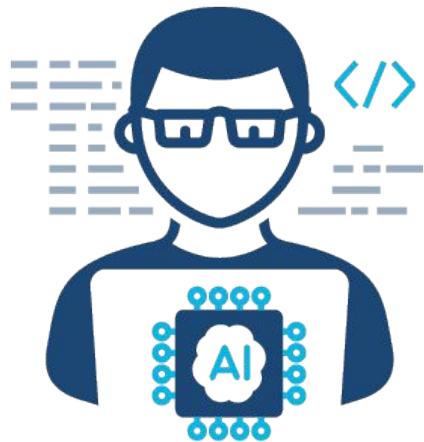


# Similar company finder

1

Project 1

**Codebase expert agent**



*generate code with AI*

2

Project 2

**Similar company finder**

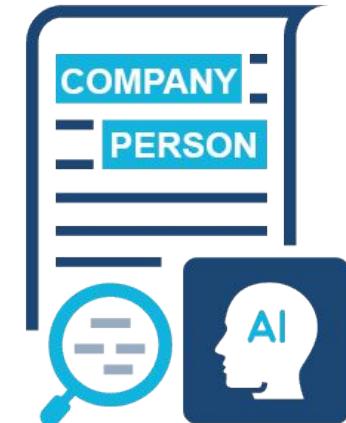


*use LLMs*

3

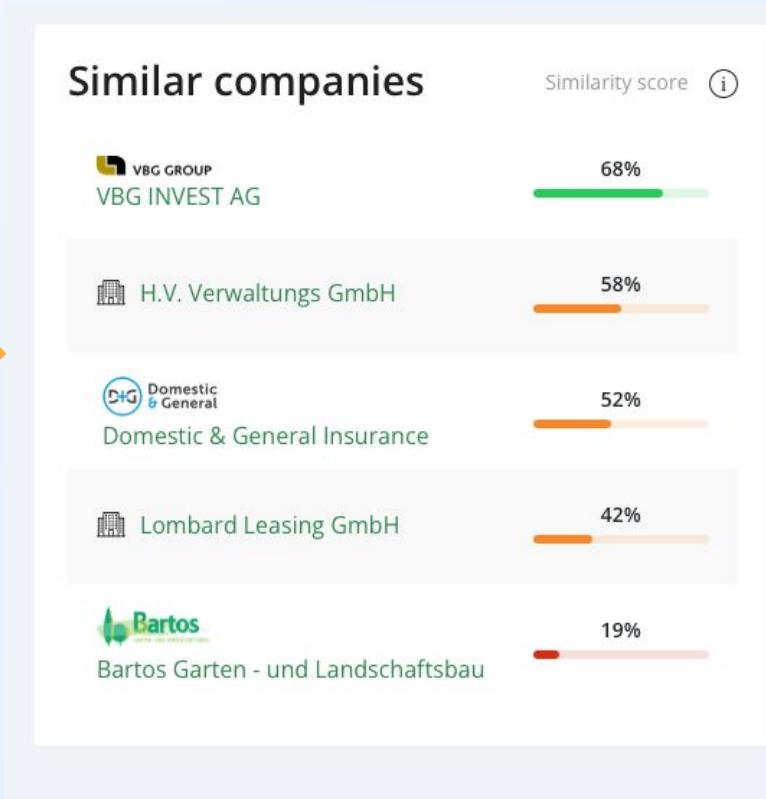
Project 3

**Documents parser**

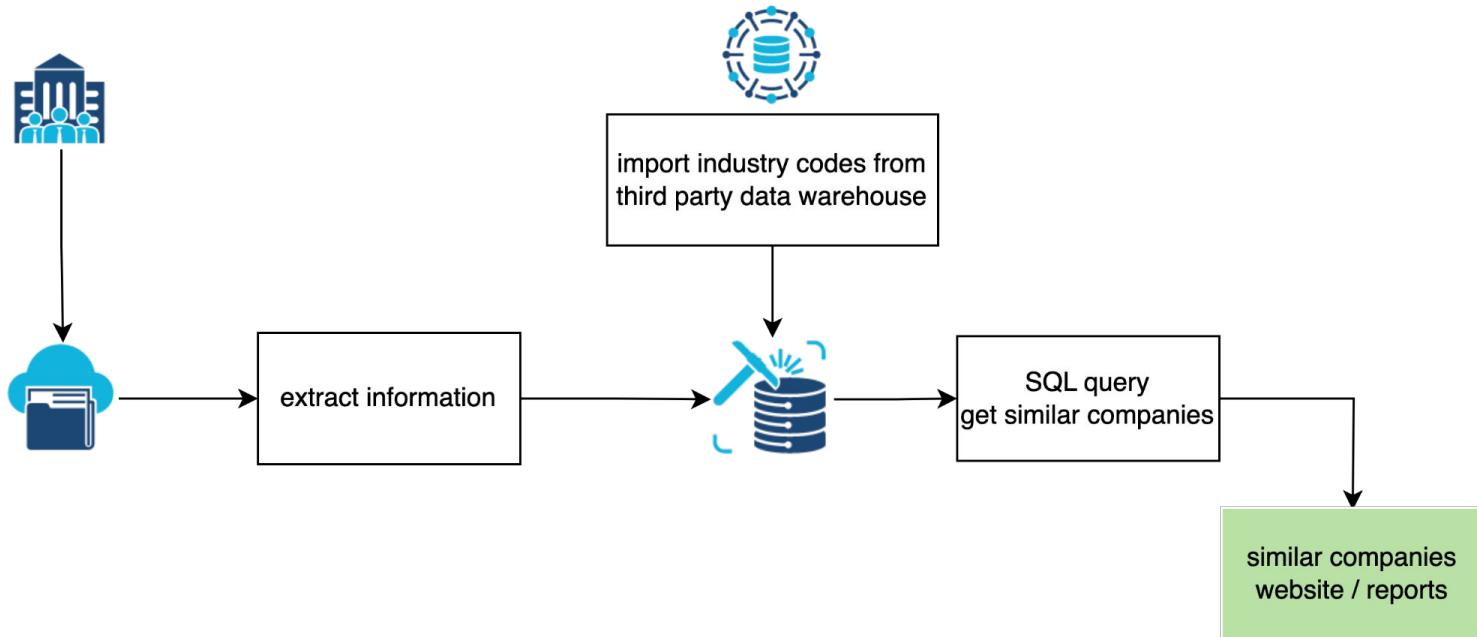


*train dedicated model*

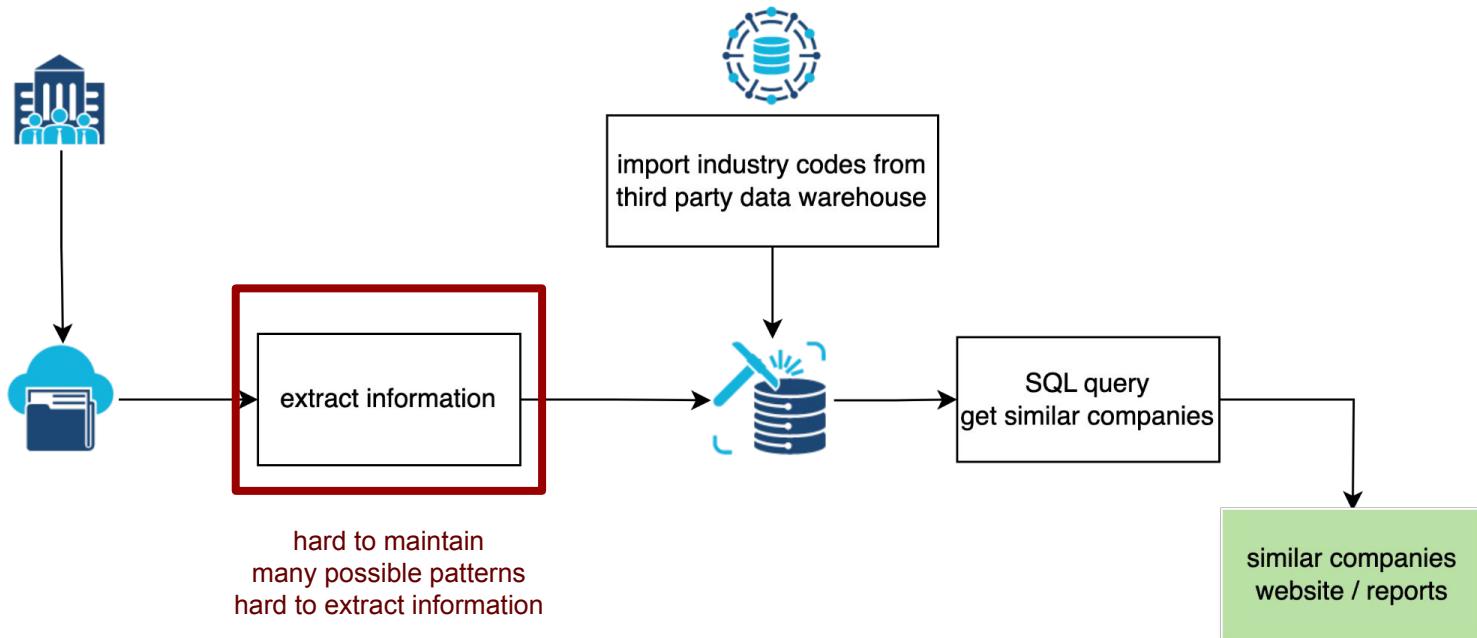
# Similar company finder



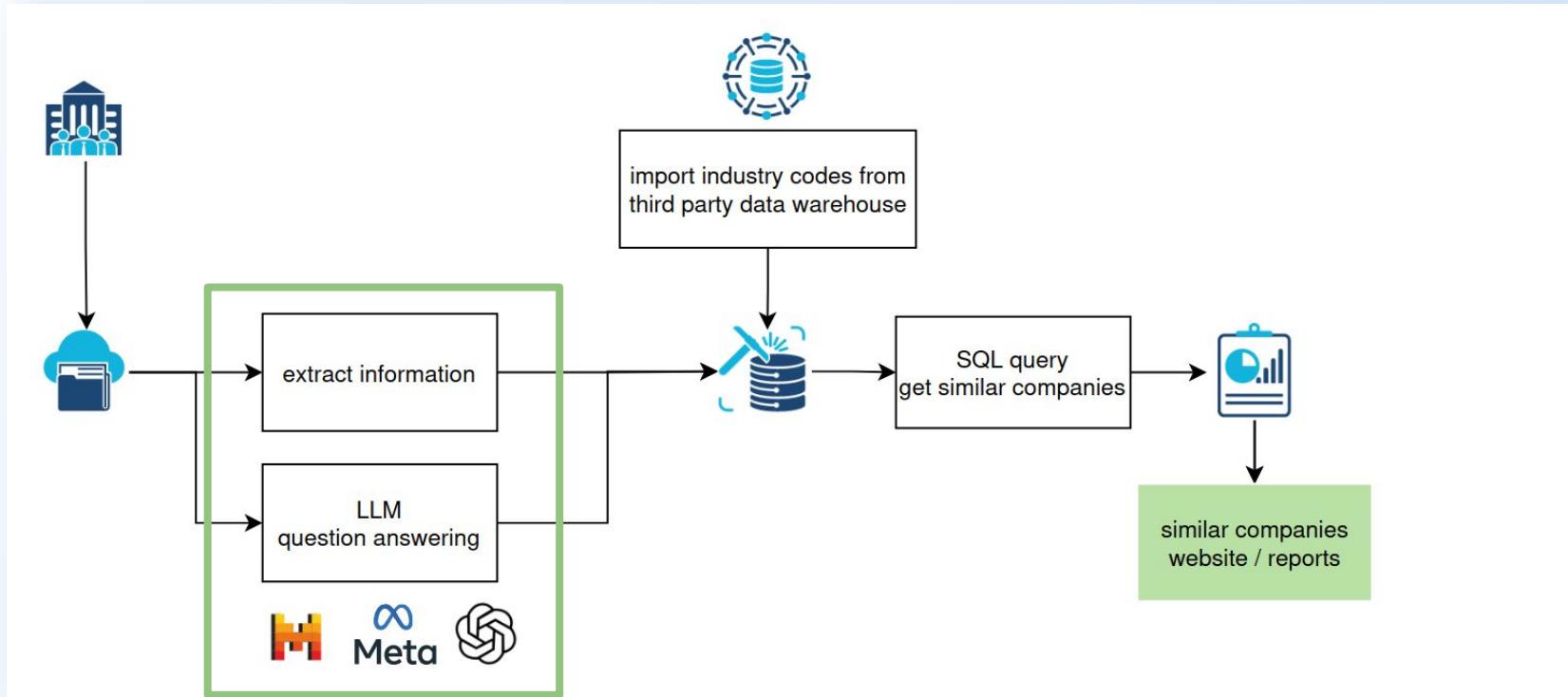
# PoC - architecture



# PoC - architecture



# architecture



# Extract information



```
<li><a href="index.html">Home</a></li>
<li><a href="home-events.html">Home Events</a></li>
<li class="has-children"> <a href="#" class="current">Header Options</a>
  <ul>
    <li><a href="tall-button-header.html">Tall Button Header</a></li>
    <li><a href="image-logo.html">Image Logo</a></li>
    <li class="active"><a href="tall-logo.html">Tall Logo Images</a></li>
  </ul>
</li>
<li class="has-children"> <a href="#">Carousels</a>
  <ul>
    <li><a href="variable-width-slider.html">Variable Image Width</a>
    <li><a href="testimonial-slider.html">Testimonial Sliders</a>
    <li><a href="featured-work-slider.html">Featured Work Slides</a>
    <li><a href="equal-column-slider.html">Equal Column Slider</a>
    <li><a href="video-slider.html">Video Slider</a></li>
    <li><a href="mini-bootstrap-carousel.html">Mini Sliders</a></li>
  </ul>
</li>
```

HTML (purified)



prompt



answer 1

answer 2

answer 3

# Extract information



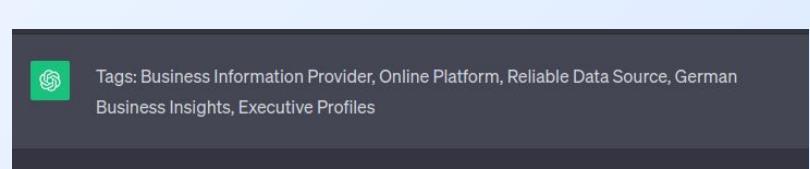
The screenshot shows the homepage of CompanyHouse. At the top, there's a navigation bar with the CompanyHouse logo, a search bar, and a 'PREMIUM MITGLIEDSCHAFT' button. Below the header, a large image of a woman in a blue suit and a man in a grey suit shaking hands. To the left of the image, the text 'Keine Risiken eingehen.' is displayed in bold. Below this, it says 'Unbegrenzt Informationen zu Firmen und Managern abrufen.' On the right side of the image, there are four icons: '8.5 M' (with a person icon), 'STRUKTUREN UND NETZWERK ERHEBEN' (with a network icon), 'BONITÄTSPRÜFUNG UND DOKUMENTE' (with a document icon), and 'IN ECHTZEIT AKTUALISIERT' (with a clock icon). A purple button at the bottom left says 'MEHR ERFAHREN'. On the far left, under the heading 'Was ist CompanyHouse?', there's a green-bordered box containing text about the platform's mission.

Mit den Daten von mehr als 10 Millionen Unternehmen und Führungskräften ist **CompanyHouse** eine vertrauenswürdige Online-Plattform, die Unternehmen, Investoren, Forschern und Fachleuten wertvolle Einblicke in Unternehmensprofile, Finanzdaten und Informationen über Geschäftsführer bietet. **CompanyHouse** bietet eine umfangreiche Sammlung aktueller Informationen und ist eine unverzichtbare Quelle für alle, die umfassende und zuverlässige Informationen über Unternehmen in Deutschland suchen.

What industry does the company operate in?

What products and services does the company offer?

Which tags best describe this company?



The summary box contains the CompanyHouse logo and a list of tags: Business Information Provider, Online Platform, Reliable Data Source, German Business Insights, Executive Profiles.

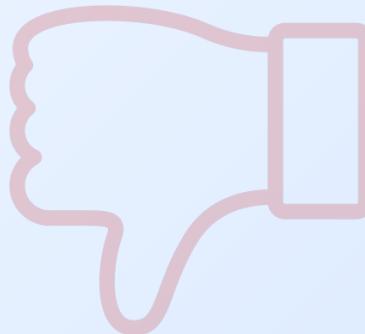
# Extract information

**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.**



The best 5 tags to describe this company based on the provided content are:

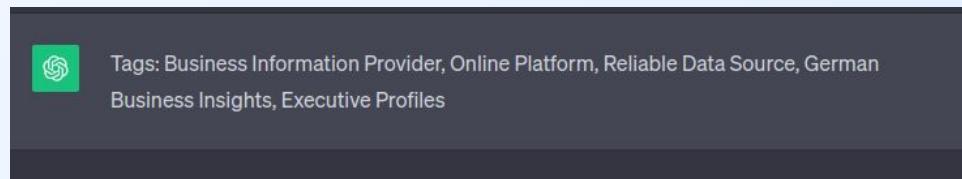
1. Business Information Provider
2. Online Platform
3. Reliable Data Source
4. German Business Insights
5. Executive Profiles

# Extract information

*Provide tags for company based on website content.*

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

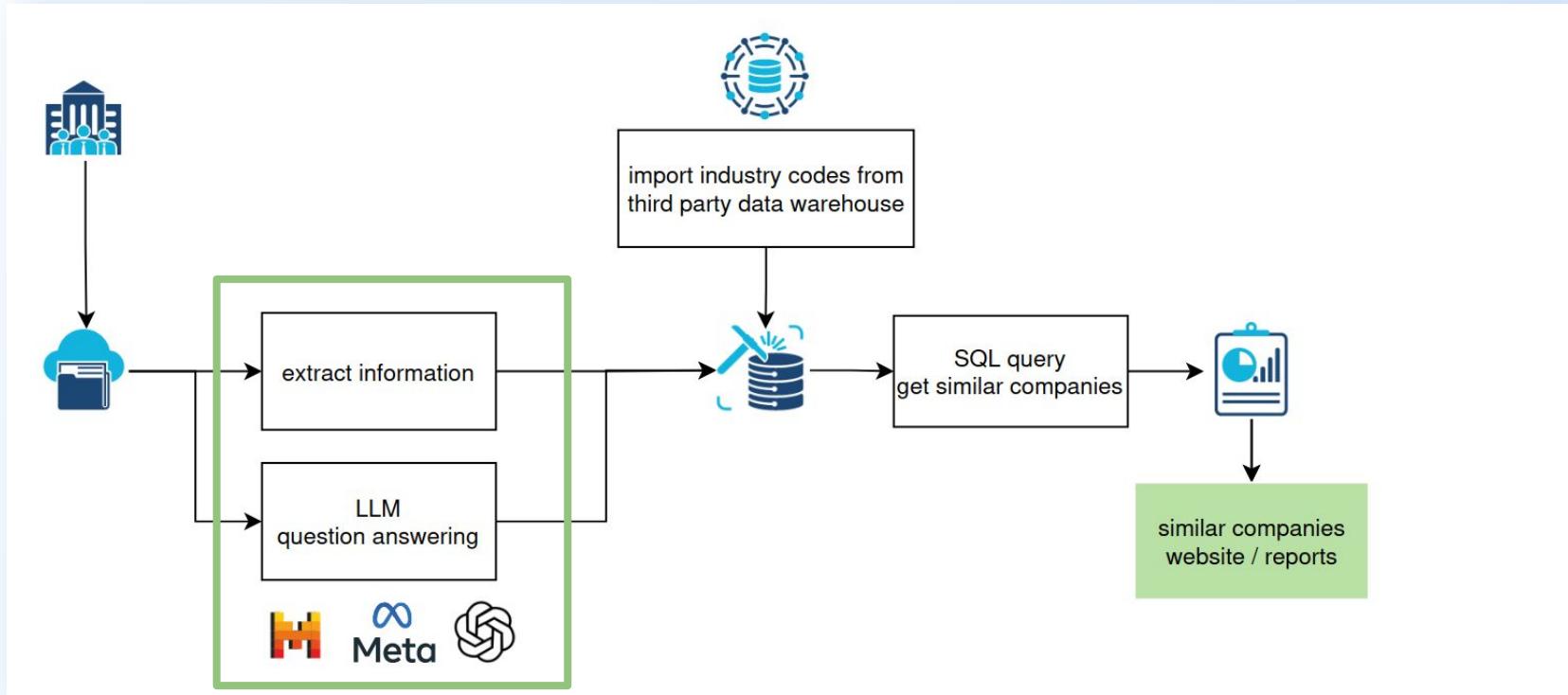
**What are the best tags to describe this company?  
Provide no more than 5 in format Tags: tag1, tag2.**



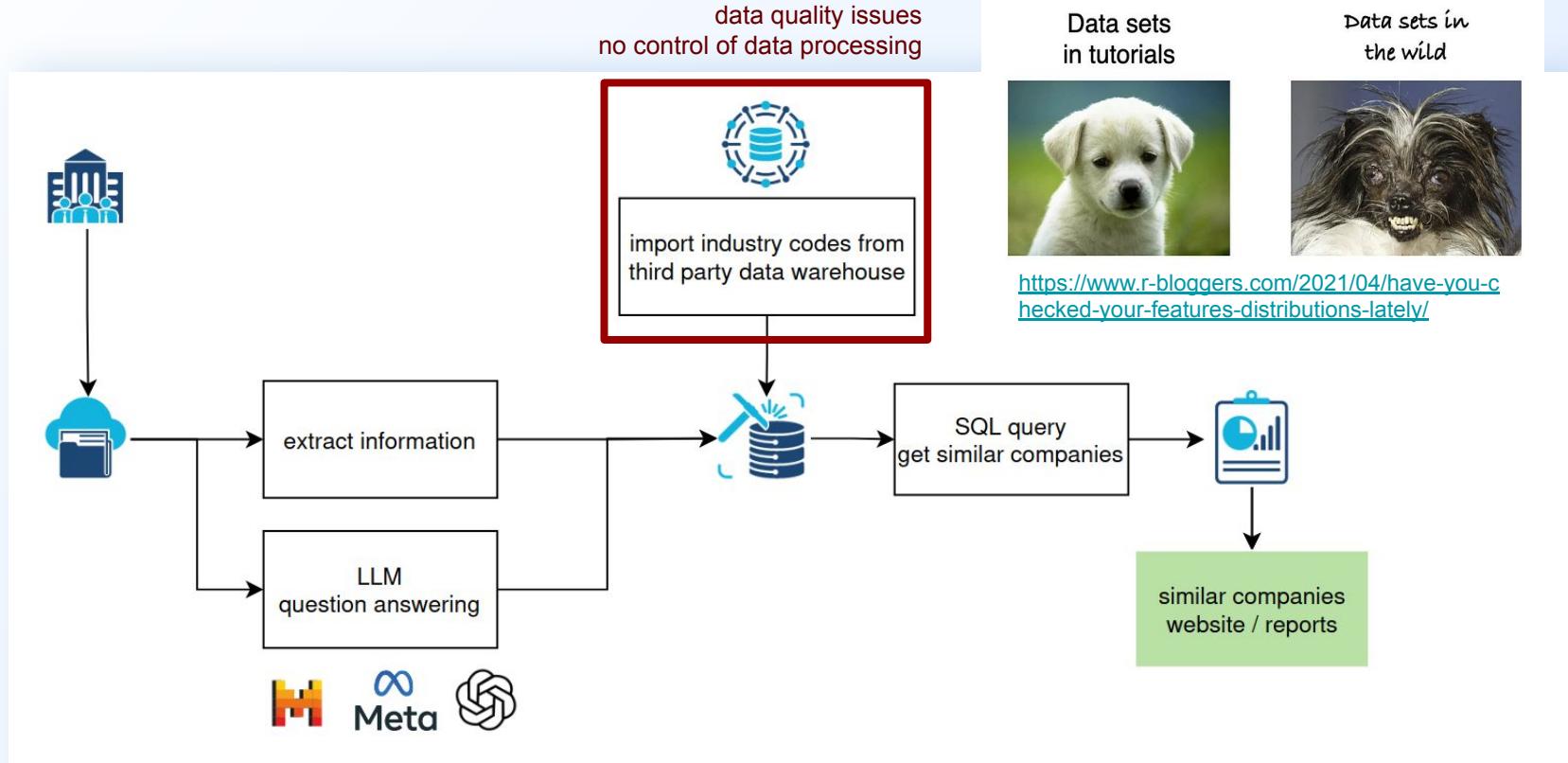
Tags: Business Information Provider, Online Platform, Reliable Data Source, German Business Insights, Executive Profiles



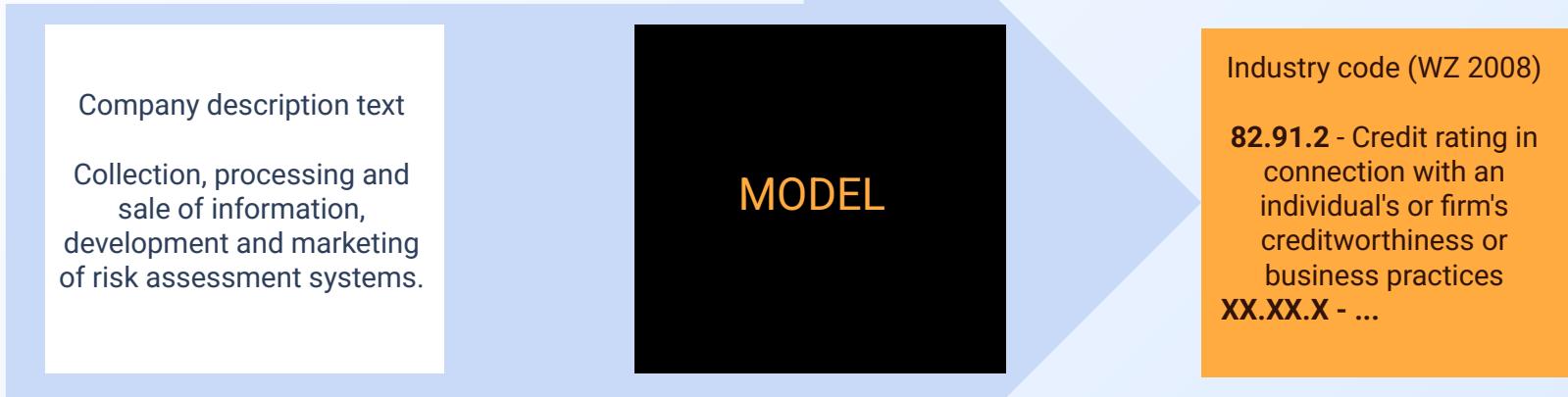
# architecture



# architecture



# The goal



```
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}
    ]
)
```

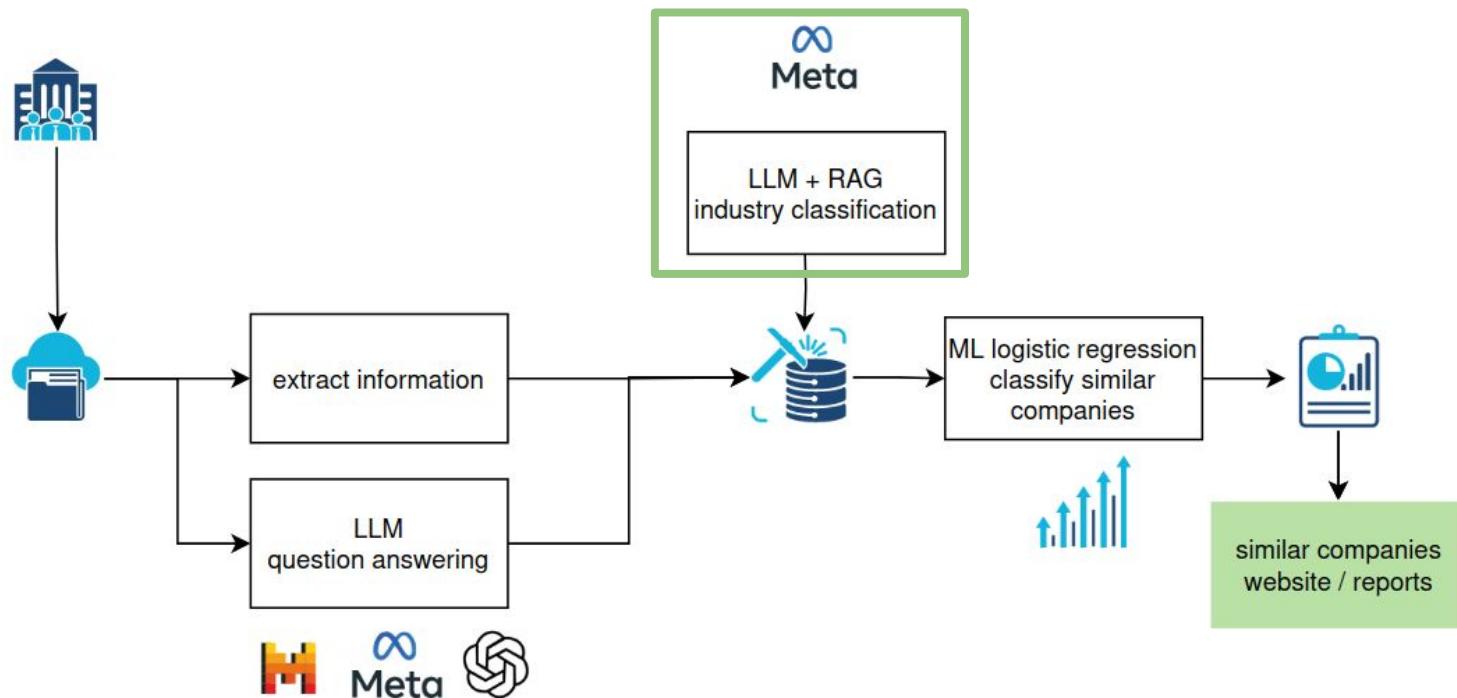


```
# call GPT-4 API for classification results
response = openai.OpenAI().chat.completions.create(
    model="gpt-4",
    messages=[
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": user_prompt}
    ]
)
```

Classification Result:  
'main\_code': '01.11.0'  
'other\_codes':['01.61.0']



# Final approach - architecture



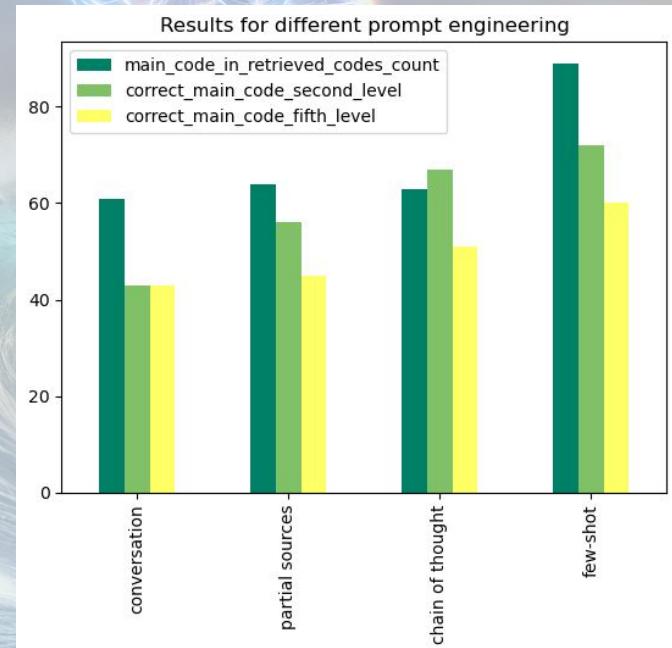
# Challenges and lessons learned

## Project 3 - Similar companies finder

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



GitHub  
code samples

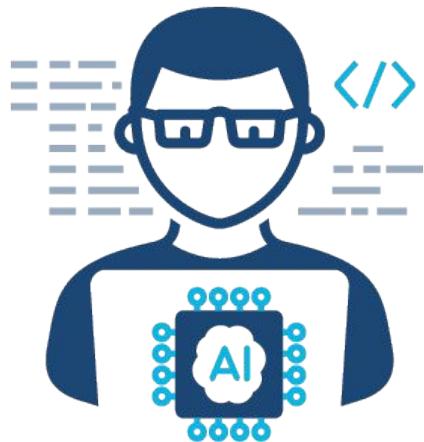


# Documents Parser

1

Project 1

**Codebase expert agent**



*generate code with AI*

2

Project 2

**Similar company finder**

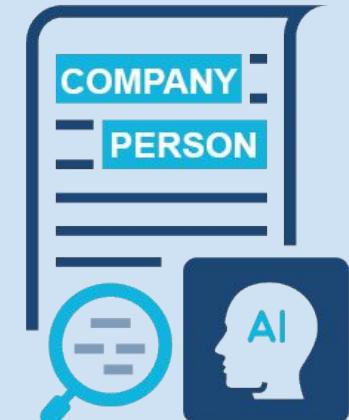


*use LLMs*

3

Project 3

**Documents parser**

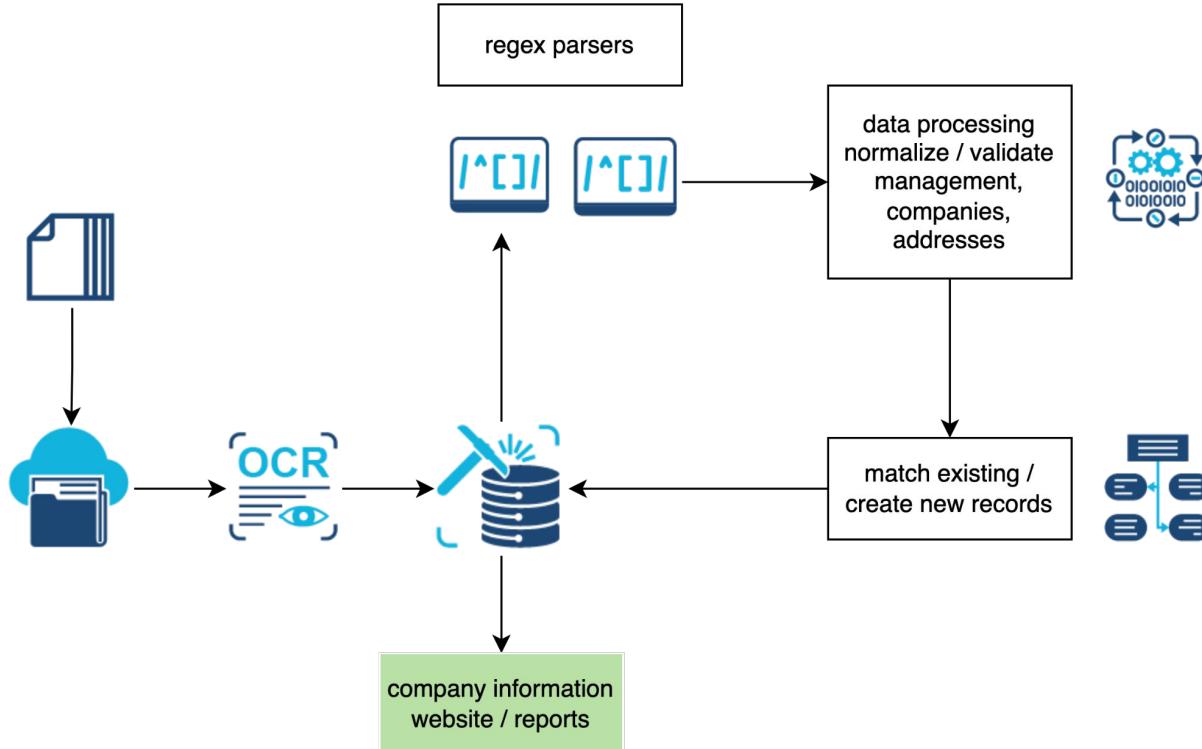


*train dedicated model*

# Extract information from documents

| Handelsregister B des Amtsgerichts Düsseldorf |  |                          | Ausdruck<br>Abruf vom 25.2.2008 18:38  |   | Nummer der Firma:<br>Seite 1 von 3  |   | HRB 49578 |
|---|--|--------------------------|--|---|---|---|-----------|
| Nummer der Eintragung                         | a) Firma<br>b) Sitz, Niederlassung, Zweigniederlassungen<br>c) Gegenstand des Unternehmens   | Grund- oder Stammkapital | a) Allgemeine Vereinungsregelung<br>b) Vorstand, Leitungsräte, geschäftsführende Direktoren, persönlich haftender Gesellschafter, Geschäftsführer, Verrechnungsberechtigte und besondere Vereinigungsbefugnis  | Prokura   | a) Rechtsform, Beginn, Satzung oder Gesellschaftsvertrag<br>b) Sonstige Rechtsverhältnisse  | a) Tag der Eintragung<br>b) Bemerkungen |           |
| 1   | 2  | 3                        | 4  | 5   | 6   | 7                                       |           |
| 1   | a) Thyssen Stahl GmbH<br>b) Düsseldorf<br>c) Der Erwerb und die Veräußerung, das Halten und Verwalten von Beteiligungen an anderen Unternehmen insbesondere des stahlerzeugenden und -verarbeitenden Bereichs, das Verwahren des eigenen Vermögens sowie die Vornahme aller damit im Zusammenhang stehenden Geschäfte.<br><u>Dr. Roxin, Jan, Düsseldorf, *21.07.1964</u> | 935.121.000,00<br>EUR    | a) Die Gesellschaft hat zumindest zwei Geschäftsführer. Die Gesellschaft wird durch zwei Geschäftsführer oder durch einen Geschäftsführer gemeinschaftlich mit einem Prokuristen vertreten.<br><br>b) Geschäftsführer:<br>Jonas, Bernd, Essen, *05.02.1951<br><br><u>Bestellt als Geschäftsführer:</u><br><u>Reinike, Dieter, Essen, *20.05.1945</u><br><u>Nicht mehr Geschäftsführer:</u><br><u>Dr. Roxin, Jan, Düsseldorf, *21.07.1964</u> | <u>Prokura erloschen:</u><br><u>Reinike, Dieter, Essen, *20.05.1945</u><br><br>Gesamtprokura gemeinsam mit einem Geschäftsführer oder einer anderen Prokuristen:<br>von Mitzaff, Dirk, Mülheim/Ruhr<br>Krickenberg, Walter, Köln, *08.01.1948<br>von den Woldeberg, Klaus, Essen<br><u>Kiesering, Friedel Hamm, *30.04.1942</u><br><u>Kriespel, Susanne, Essen, *09.02.1958</u><br><u>Kühnast, Maguel, Essen, *07.02.1953</u><br><u>Swinty, Michael, Essen, *26.02.1963</u><br><u>van Bracht, Arno, Dinslaken, *12.06.1968</u><br><u>Regelmann, Ulrich, Dortmund, *30.07.1956</u> | a) Gesellschaft mit beschränkter Haftung<br>Gesellschaftsvertrag vom 13.06.2003<br>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.<br><br>a) 25.03.2004<br>Hauseiss<br>b) Beschluss Blatt 37, 47<br>Sonderband<br>Gesellschaftsvertrag Blatt 42 ff. Sonderband  |   |           |
| 2   |  | 935.146.600,00<br>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.<br><br>a) 08.06.2004<br>Koepin<br>b) Beschluss Blatt 62 ff.<br>Sonderband<br>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 übertragenen Rechtsträgers vom selben Tage mit der Rhs-Qualifizierungsgesellschaft mbH mit Sitz in Duisburg (Amtsgericht Duisburg HRB 6511) verschmolzen.<br><br>a) 22.06.2004<br>Koepin<br>b) Verschmelzungsvertrag Blatt 72 ff. Sonderband<br>Zustimmungsbeschlüsse Blatt 61 ff., 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 übertragenen Rechtsträgers vom selben Tage mit der Hoesch Industrielaer GmbH mit Sitz in Dortmund (AG Dortmund, HRB 10639) verschmolzen.<br><br>a) 08.07.2004<br>Koepin<br>b) Verschmelzungsvertrag Blatt 68 ff. Sonderband  |   |           |

# Use regex parsers - architecture



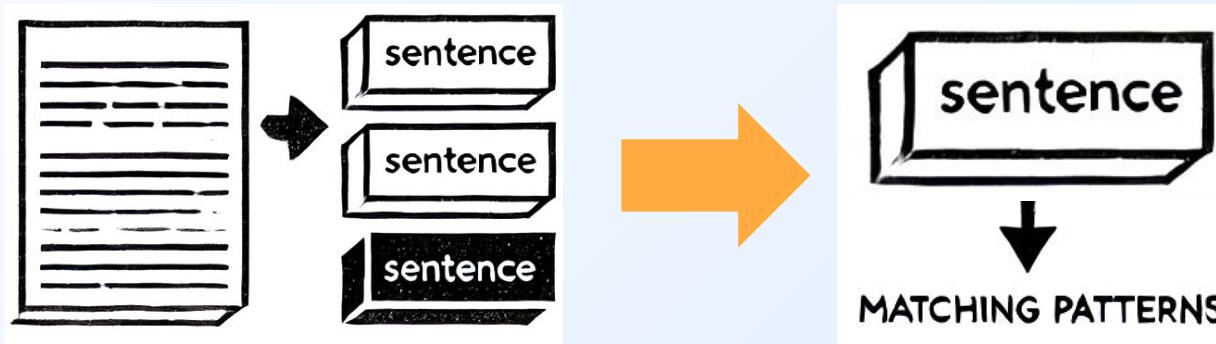
# First approach - regexp

## Assumptions:

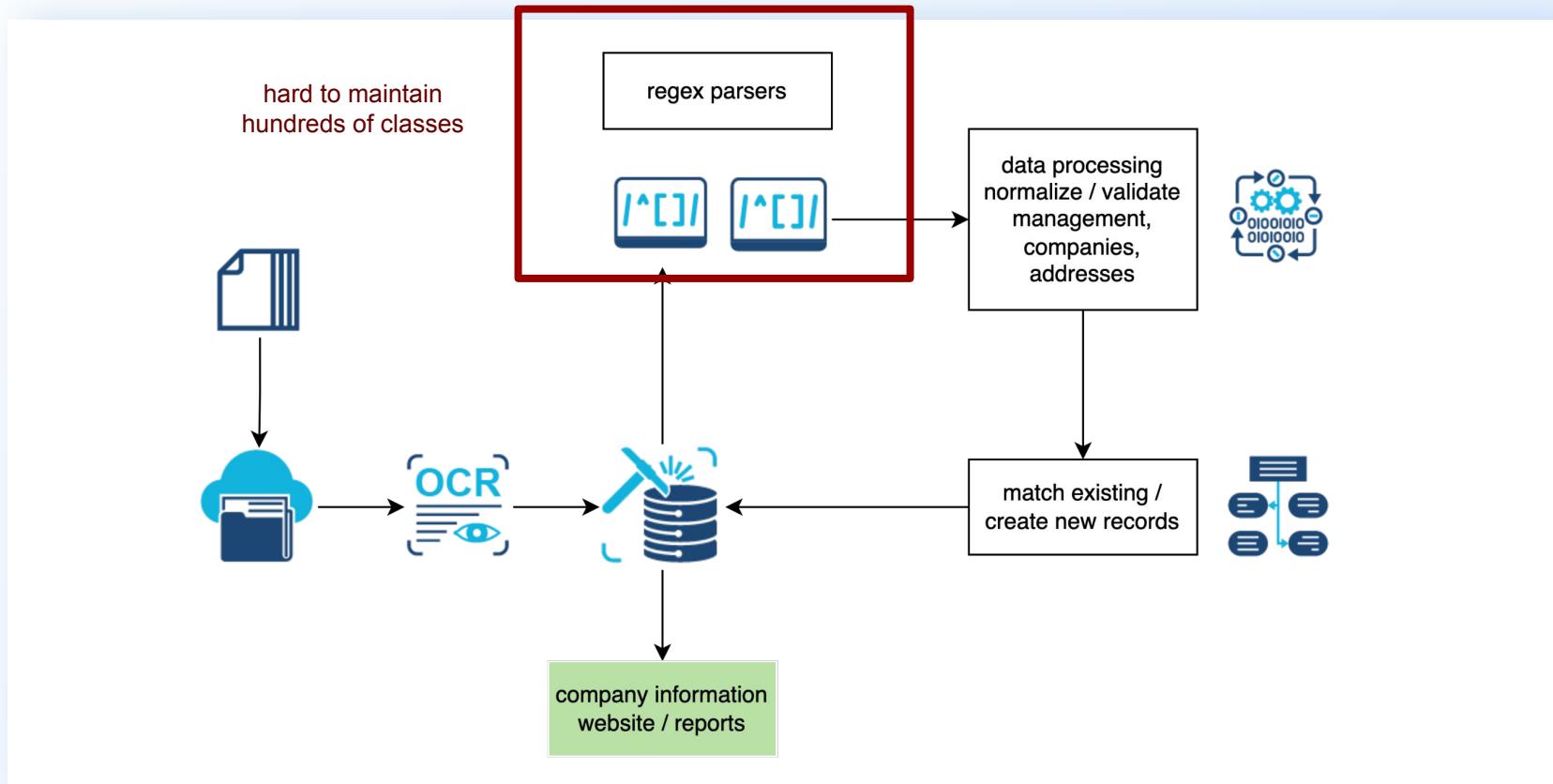
- Conditions are based on the order of words in the sequence
- Multiple possible variants: <ScientificTitle> LastName, FirstName, City, <BirthDate>, <SigningAuthority>

*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

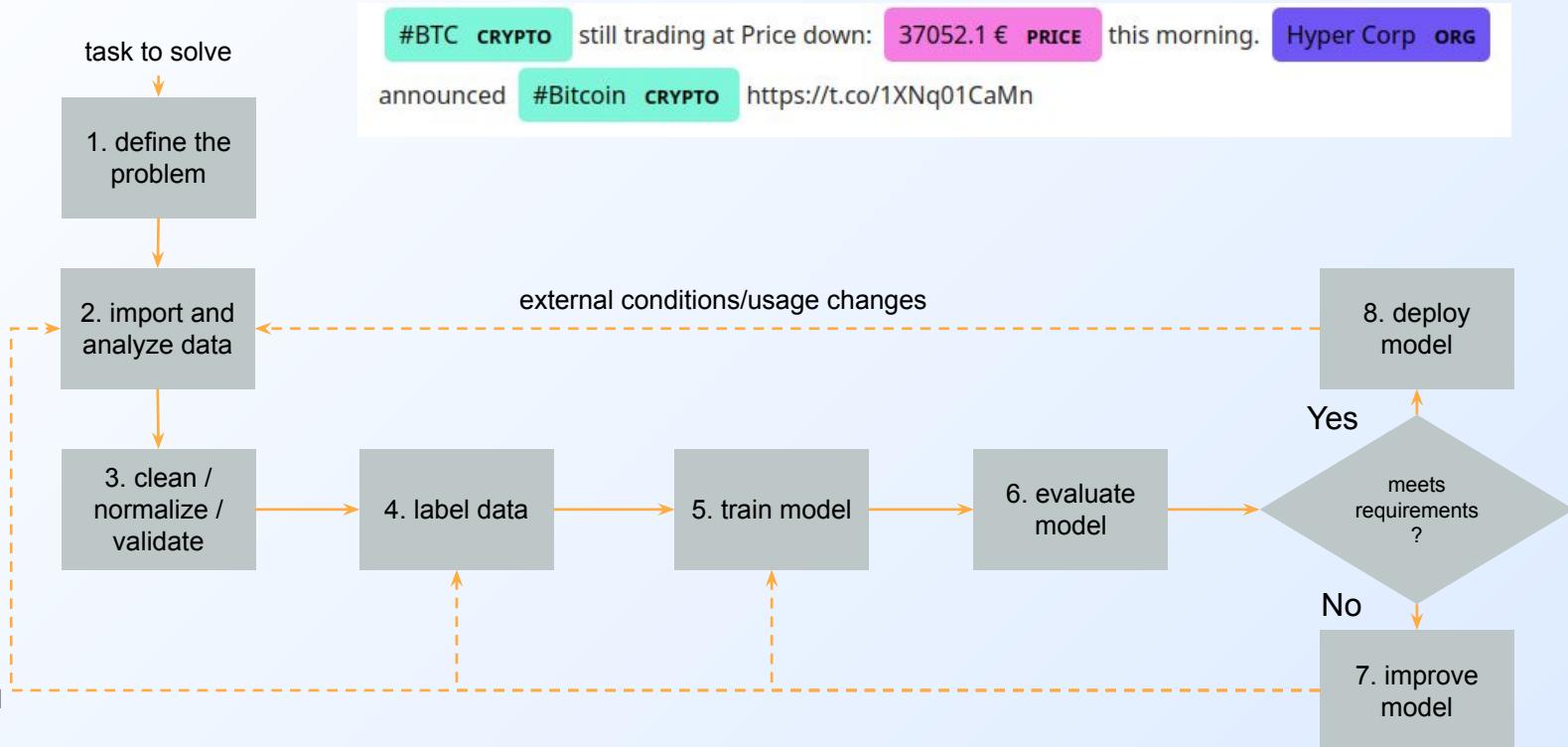


- Assumptions may not cover 100% of cases
  - Different structure, different order of phrases
  - Foreign names, multi-word names, titles
  - Complicated logic to cover all cases
  - Good accuracy for structured texts
  - Low accuracy when there are many variants

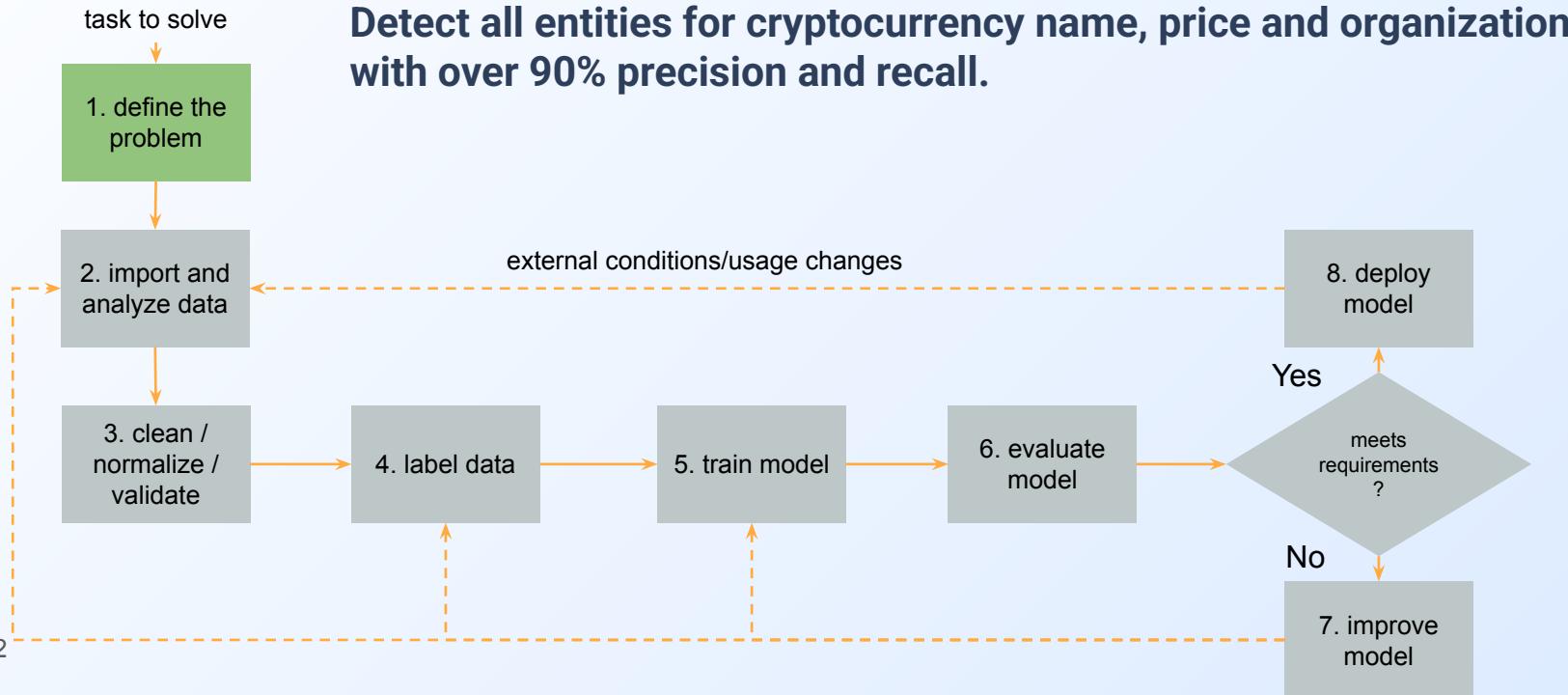


- ✓ Implementations of RegexParser 1,002 results
  - ✓ Unclassified 1,002 results
    - ✓ companyhouse 1,002 results

# Solution - ML model for Named Entity Recognition



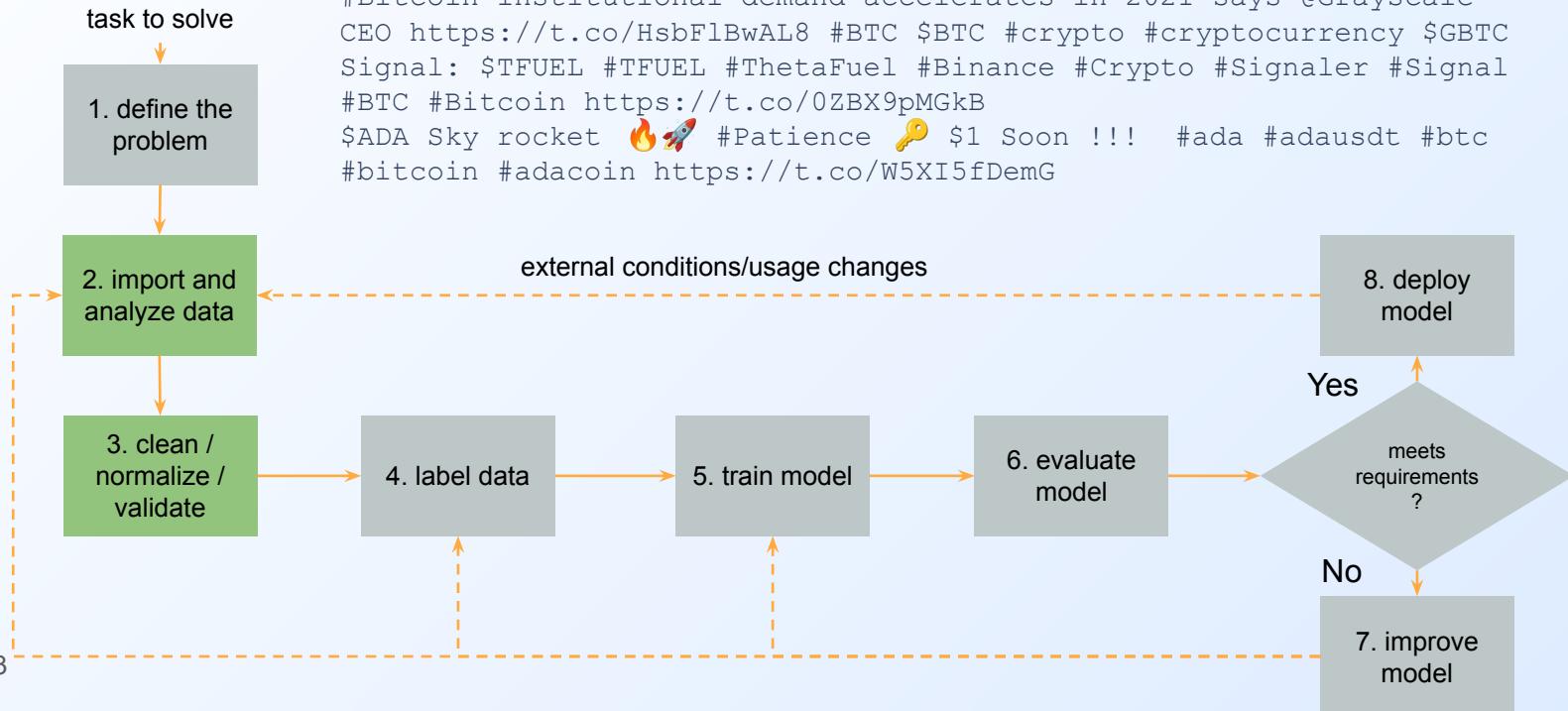
# Solution - ML model for Named Entity Recognition



# 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



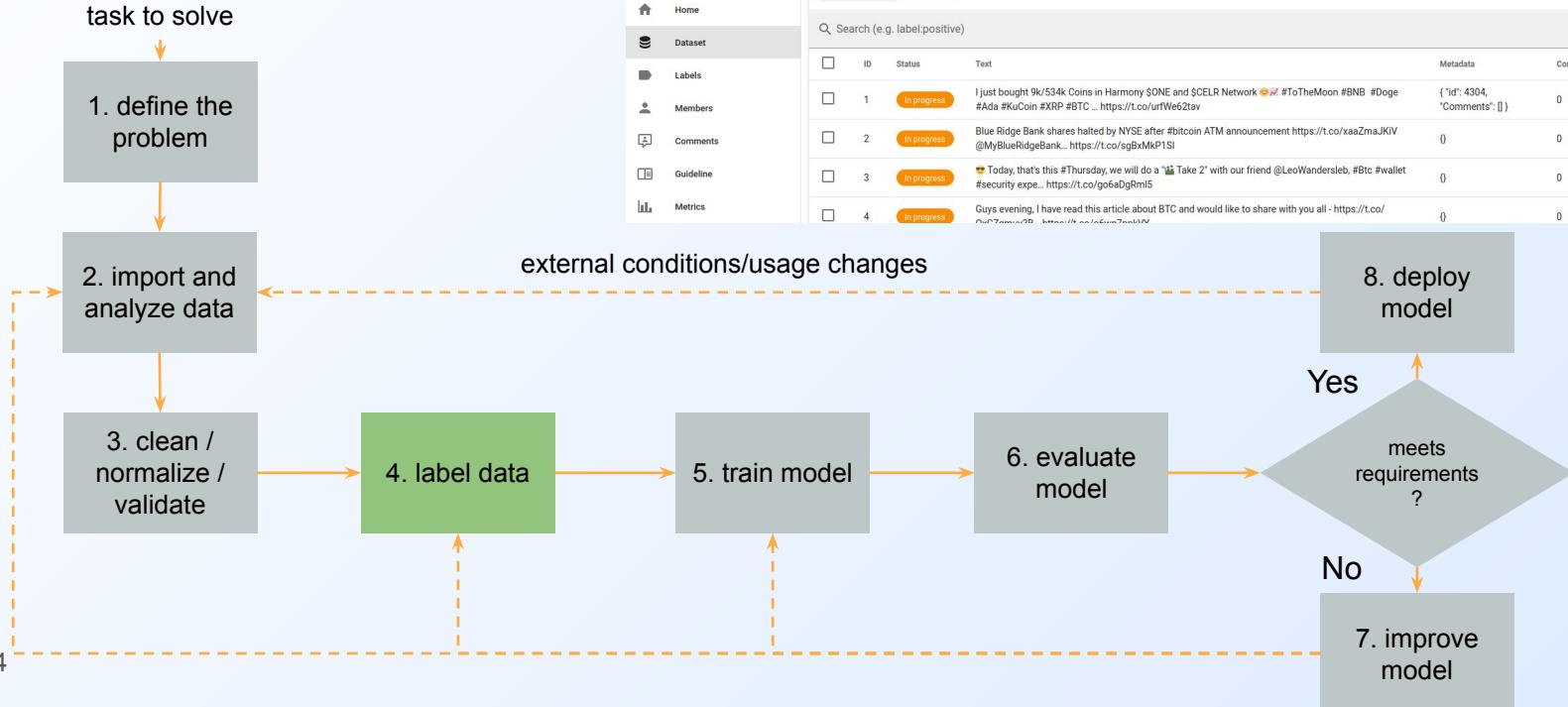
#BTC CRYPTO still trading at Price down: 37052.1 € PRICE this morning. Hyper Corp ORG

127.0.0.1:8000/projects/1/dataset?limit=10&offset=0

Cryptocurrency NER tagging

| ID | Status      | Text  | Metadata                       | Comments | Action  |
|----|-------------|---|--------------------------------|----------|---|
| 1  | in progress | I just bought 9k/534k Coins in Harmony \$ONE and \$CELR Network 🚀 #ToTheMoon #BNB #Doge #Ada #KuCoin #XRP #BTC ... https://t.co/urfw62tw  | { "id": 4304, "Comments": [] } | 0        | <button>Edit</button> <button>Annotate</button> |
| 2  | in progress | Blue Ridge Bank shares halted by NYSE after #bitcoin ATM announcement https://t.co/xaaZmaJKIV @MyBlueRidgeBank... https://t.co/sgBxMkP1SI |                                | 0        | <button>Edit</button> <button>Annotate</button> |
| 3  | in progress | Today, that's this #Thursday, we will do a "Take 2" with our friend @LeoWandersleb, #Btc #wallet #security expo. https://t.co/g6g4DgRml5  |                                | 0        | <button>Edit</button> <button>Annotate</button> |
| 4  | in progress | Gus evening, I have read this article about BTC and would like to share with you all - https://t.co/Our77zqDB ... https://t.co/zaeTnV     |                                | 0        | <button>Edit</button> <button>Annotate</button> |

Actions Delete Delete All



# Label data



task to solve

1. define the problem

NER sample project

My Corp #BTC #Bitcoin #Ethereum #ETH #Crypto #cryptotrading \$RSR I know i told you  
 •ORG •CRYPTO •CRYPTO •CRYPTO  
 •CRYPTO

guys the target was What Google will do \$0.060, i know we... https://t.co/bvEtSnhs67  
 •ORG •PRICE

Progress

Total 4295  
 Complete 70  
 2%

Label Types

CRYPTO PRICE ORG

2. import and analyze data

3. clean /  
normalize /  
validate

external conditions/usage changes

4. label data

5. train model

6. evaluate model

8. deploy model

Yes

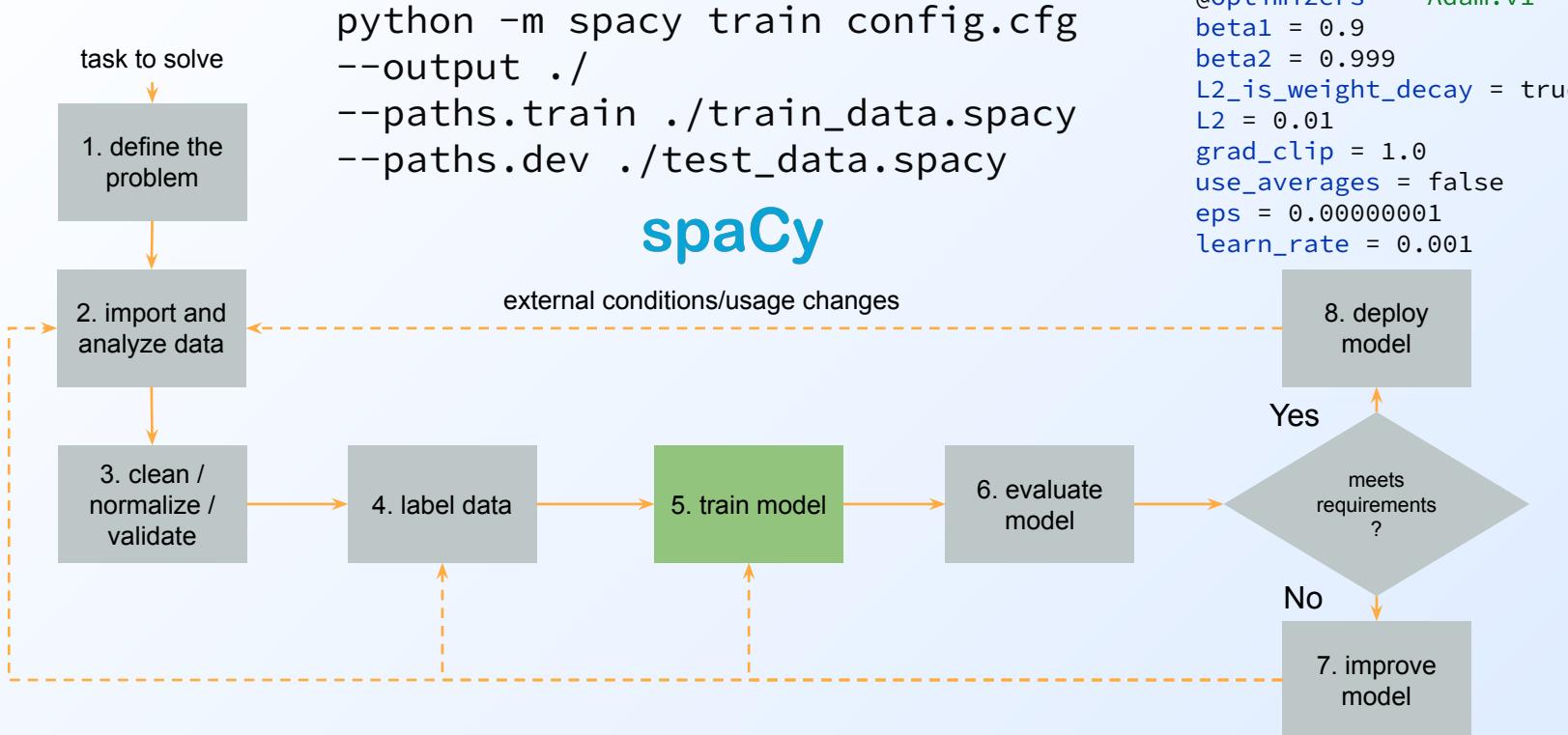
meets  
requirements  
?

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"]]  
 },  
 ...  
 ]
```

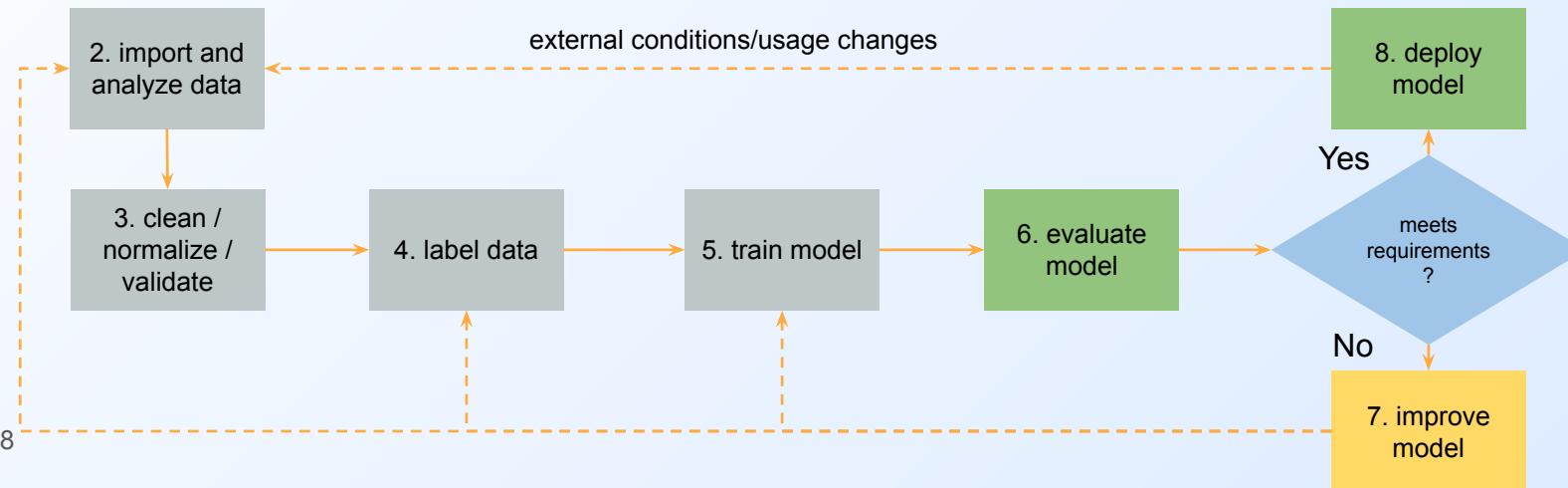
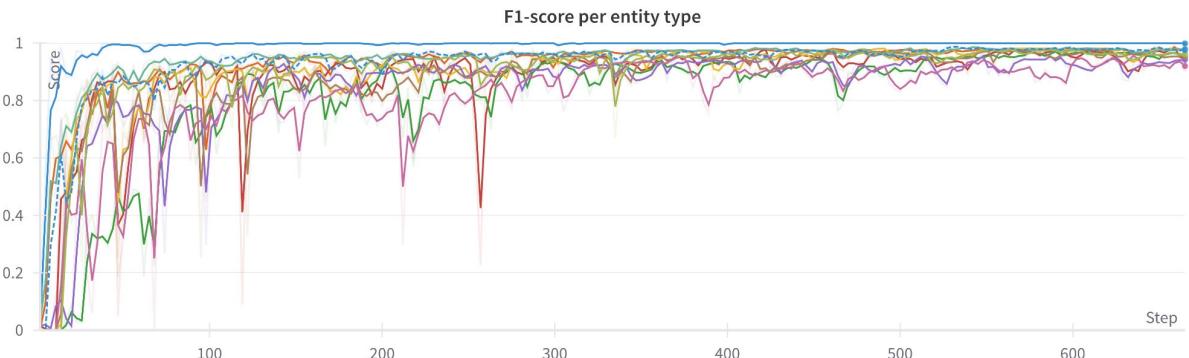
# Train model



# 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.f
- 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

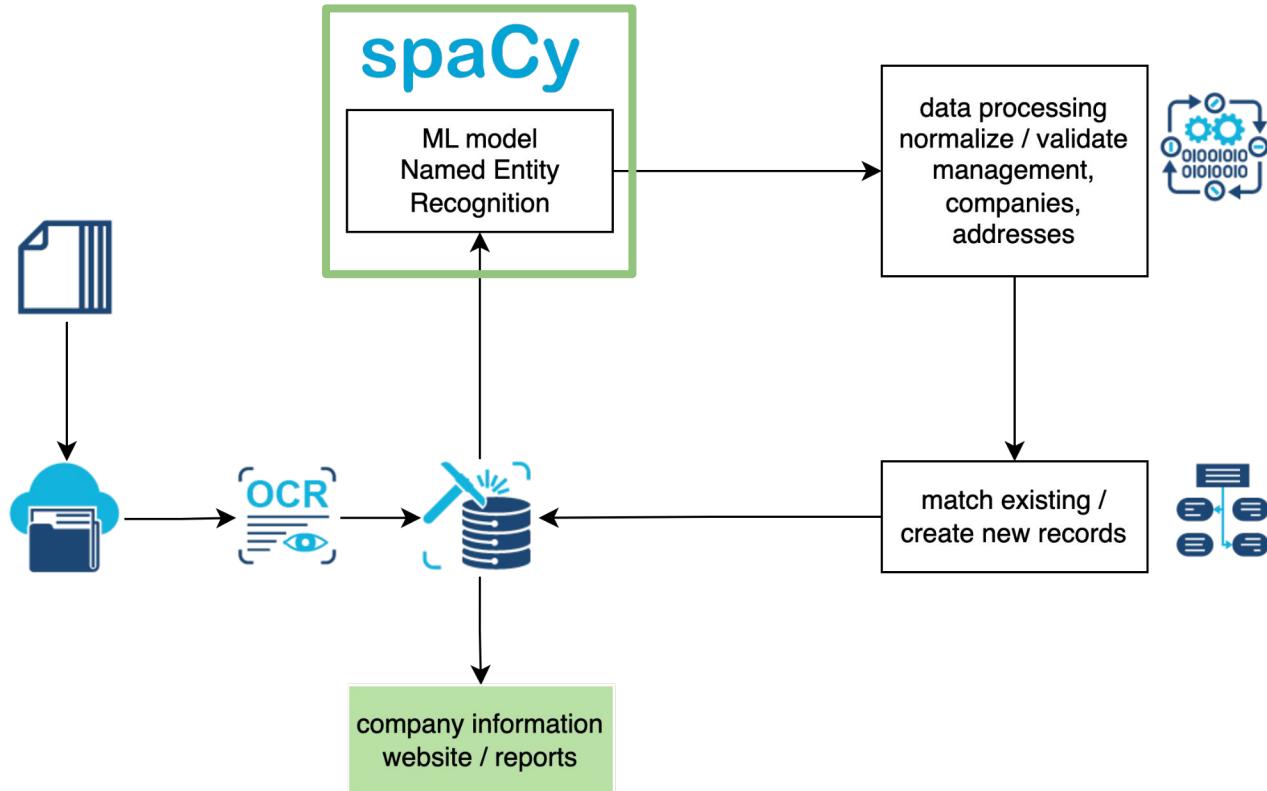
Press CMD+C or CTRL+C to copy this data



```
1 nlp_ner = spacy.load("model-best")
2
3 doc = nlp_ner("#BTC still trading at Price down: 37052.1 € this morning. Hyper Corp
4 announced #Bitcoin https://t.co/1XNq01CaMn")
5
6 colors = {"PRICE": "#F67DE3", "CRYPTO": "#7DF6D9", "ORG": "#7156F6"}
7 options = {"colors": colors}
8
9 spacy.displacy.render(doc, style="ent", options= options, jupyter=True)
[46]
```

#BTC CRYPTO still trading at Price down: 37052.1 € PRICE this morning. Hyper Corp ORG  
announced #Bitcoin CRYPTO <https://t.co/1XNq01CaMn>

# spaCy NER - architecture



# Challenges and lessons learned

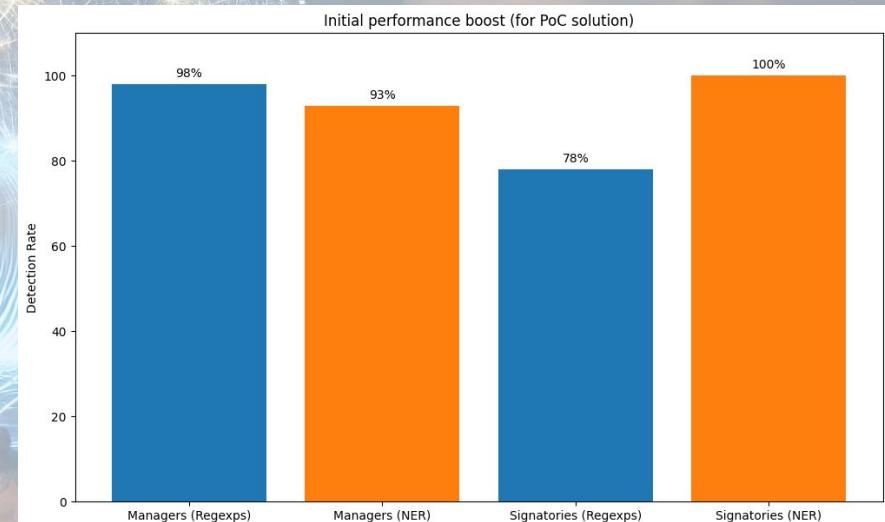
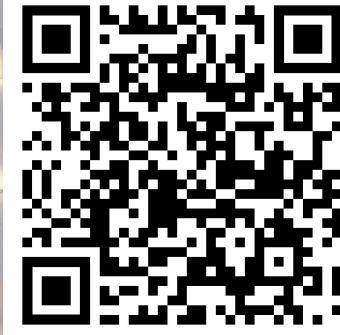
## Project 2 - Document parser

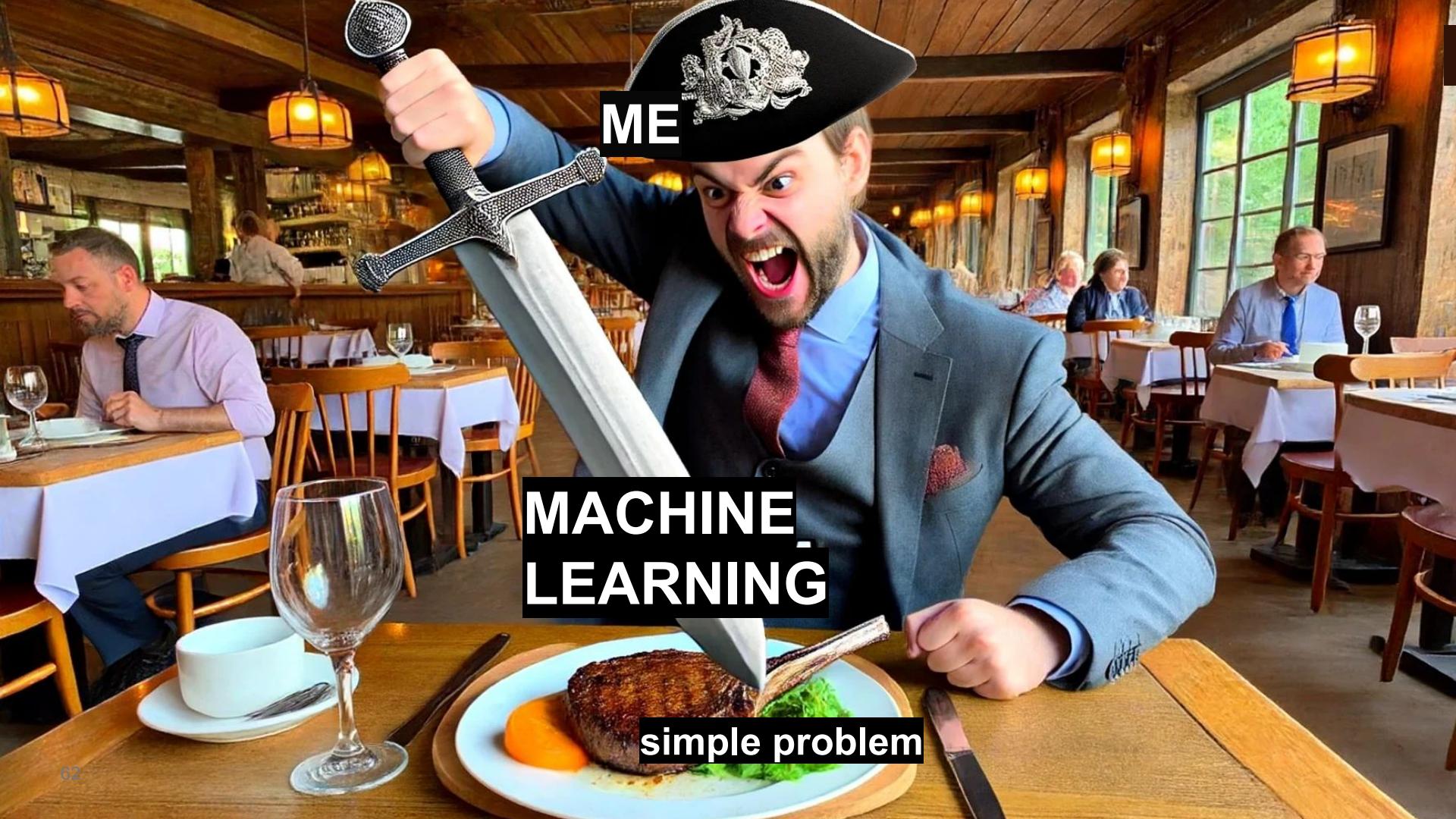
1. Even 20% more recognized cases
2. No control of single conditions / need to retrain model
3. Less time spent on maintenance
  - tagging: 700 documents per task ~ 40h of work
  - model preparation and train: ~40h of work



**Github sample**

[https://github.com/mzarnecki/  
train-ner-model-with-spacy](https://github.com/mzarnecki/train-ner-model-with-spacy)





ME

# MACHINE LEARNING

simple problem

leave your feedback

# Thank you for watching

Michał Żarnecki

Contact: michal@zarnecki.pl

github: <https://github.com/mzarnecki>

LinkedIn: <https://www.linkedin.com/in/michal-zarnecki-47219355/>

