



# Registroit

L'ANAGRAFE DEI DOMINI .IT

REGISTRO .IT È GESTITO DA





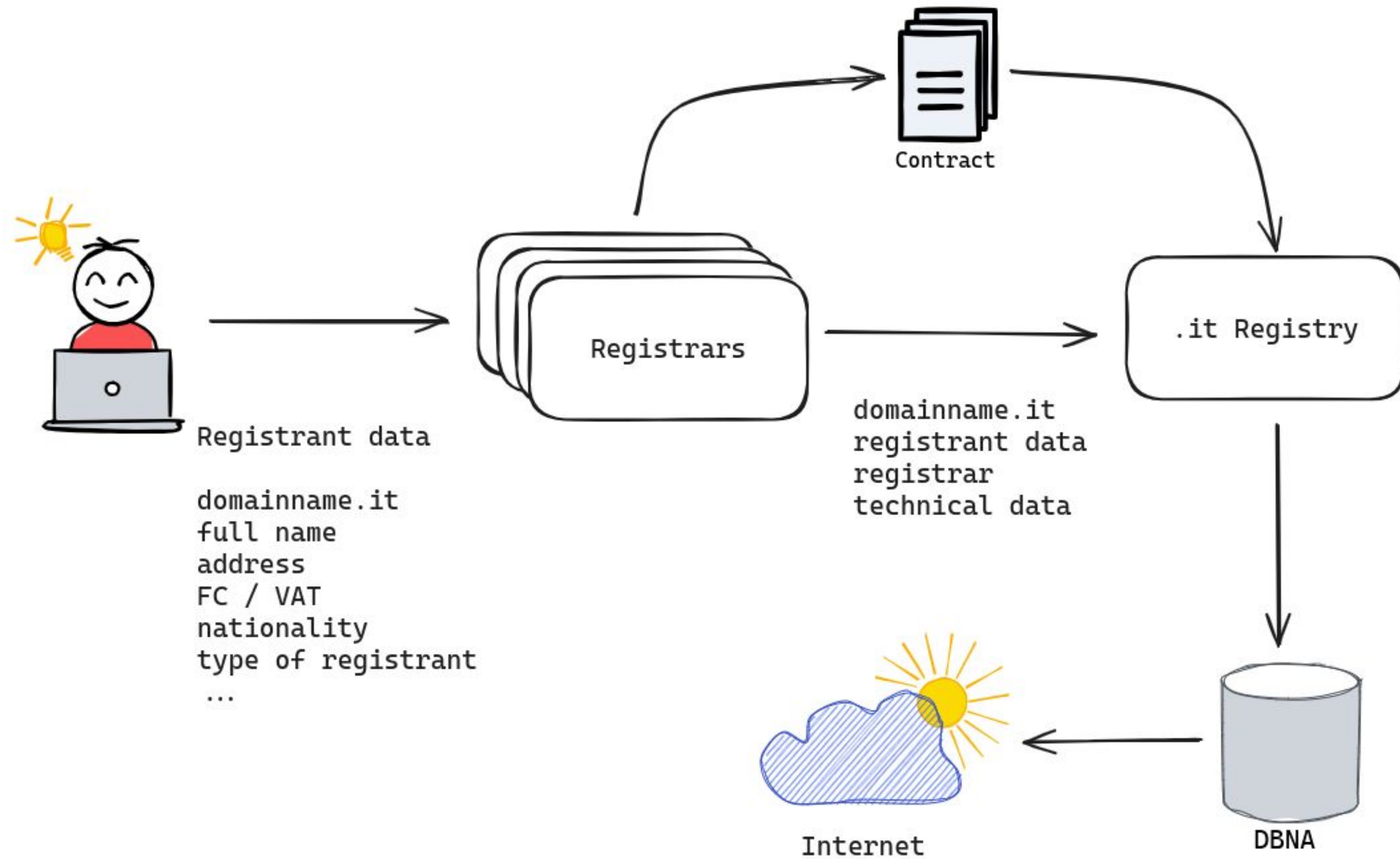
# READS: A REgistrants Anomalies Detection System

2

Daniele Sartiano

21 March 2024

# Domain Name Registration Process



# Why verify the accuracy of registrant data?

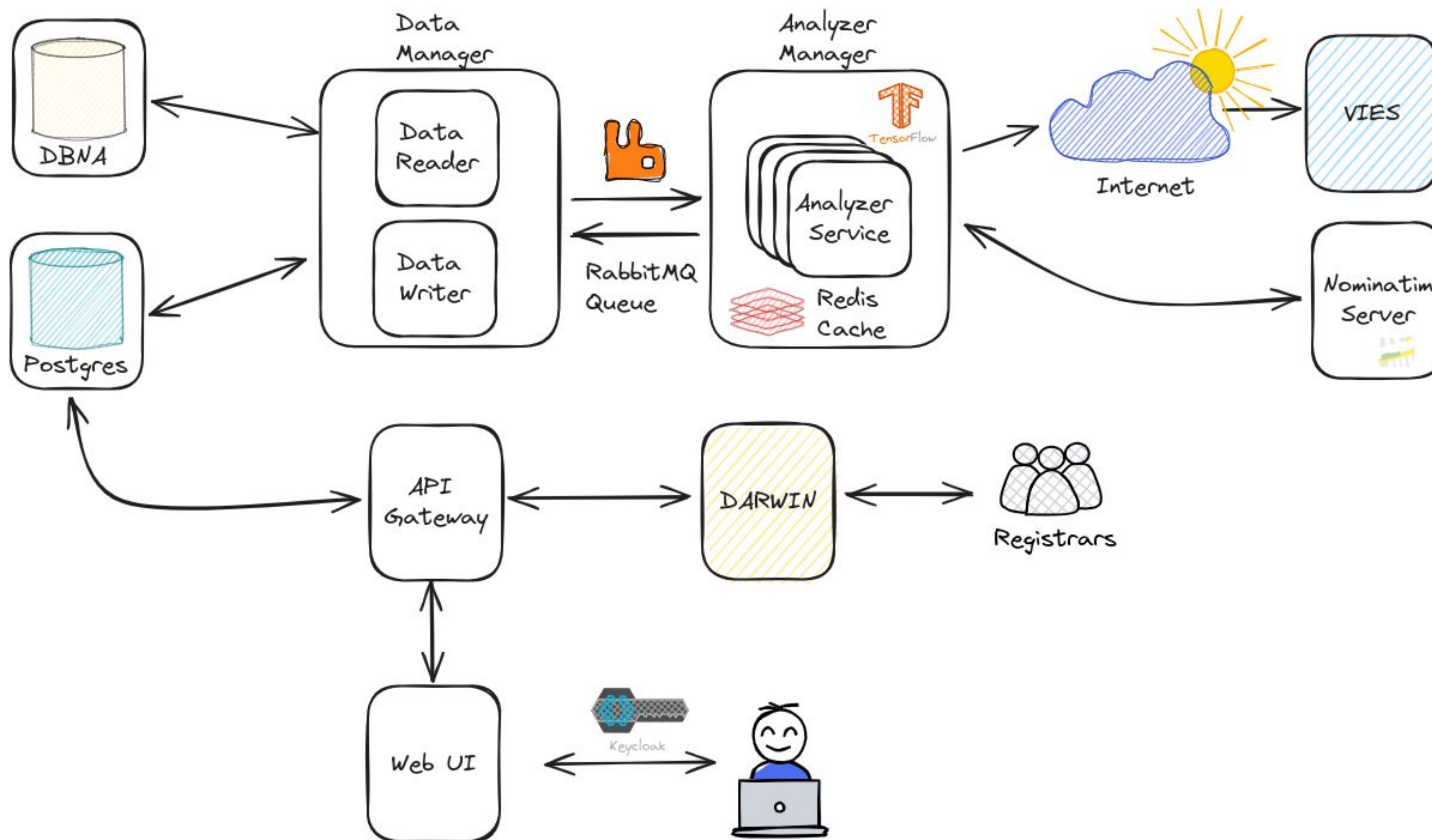
- Ensures the **reliability** and **quality** of the data contained in the .it Registry database (NIS2 Directive)
- Prevents and mitigates the impact of any **security incidents** involving domain names
- Prevents the “**crime of impersonation**” as stated in Article 494 of the Italian Criminal Code
- Ensures **compliance** with **GDPR**: personal data must be “**accurate** and, if necessary, **updated**; all reasonable measures must be taken to promptly delete or rectify inaccurate data with respect to the purposes for which they are processed (accuracy).”
- Some examples of incorrect data provided by registrants include:
  - false names (whether legal entities or individuals)
  - incorrect and inconsistent fiscal code with the corresponding name
  - non-existent addresses
  - invalid VAT numbers

# READS: REgistrants Anomalies Detection System

- A **semi-automatic** system has been designed and implemented to analyze registrants data using **NLP techniques, heuristics, and external services**
- For each registrant stored in the DBNA, an **anomaly score** is calculated to identify the severity of any eventual detected anomaly
- Specifically, the system analyzes the following information:
  - name and organization
  - address
  - fiscal code/VAT number
  - nationality
  - registrant type



# Architecture



# Registrants aggregation

- In an ideal scenario, a registrant would have multiple domains under a single contact
- However, in reality, it is not rare to have the **same registrant replicated multiple times**, resulting in **many contacts** in the Registry database that correspond to a **single real registrant**
- Furthermore, unfortunately, it often happens that the same registrant is duplicated with **some variations** in data

# Registrants aggregation<sup>(Cont)</sup>

CNR | via Giuseppe Moruzzi 1, 56124, Pisa, Italy

C.N.R. | via G. Moruzzi 1, 56124, Pisa, Italy

CNR | Moruzzi 1, 56100, Pisa, Italy

C.N.R | via Moruzzi 1, 56124, Pisa, Italy

CNR | via Giuseppe Moruzzi 1, 56124, Pisa, Italy





# Registrants aggregation<sup>(Cont)</sup>

Registrants are aggregated if the following fields match:

- **RegCode** (VAT number/Fiscal Code)
- Normalized **Organization** (lowercase, removed the punctuation and tokens sorted "My Beautiful Company s.p.a." --> "spa my beautiful company" --> "beautifulcompanymyspa")
- **Entity Type**
- **Country**
- **Nationality**
- **City**

The total number of aggregate registrants is **1.9 million**, compared to **2.9 million** registrants (entries in the database)

# Anomalies Detection

The system calculates an **anomaly score** for each aggregation and it is based on the checks performed on the relevant fields:

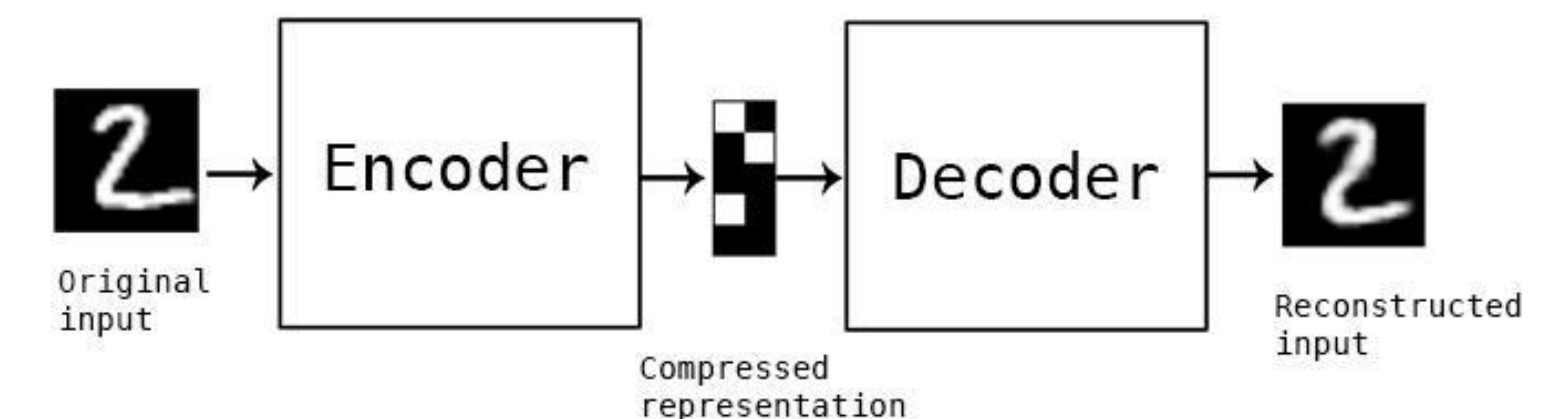
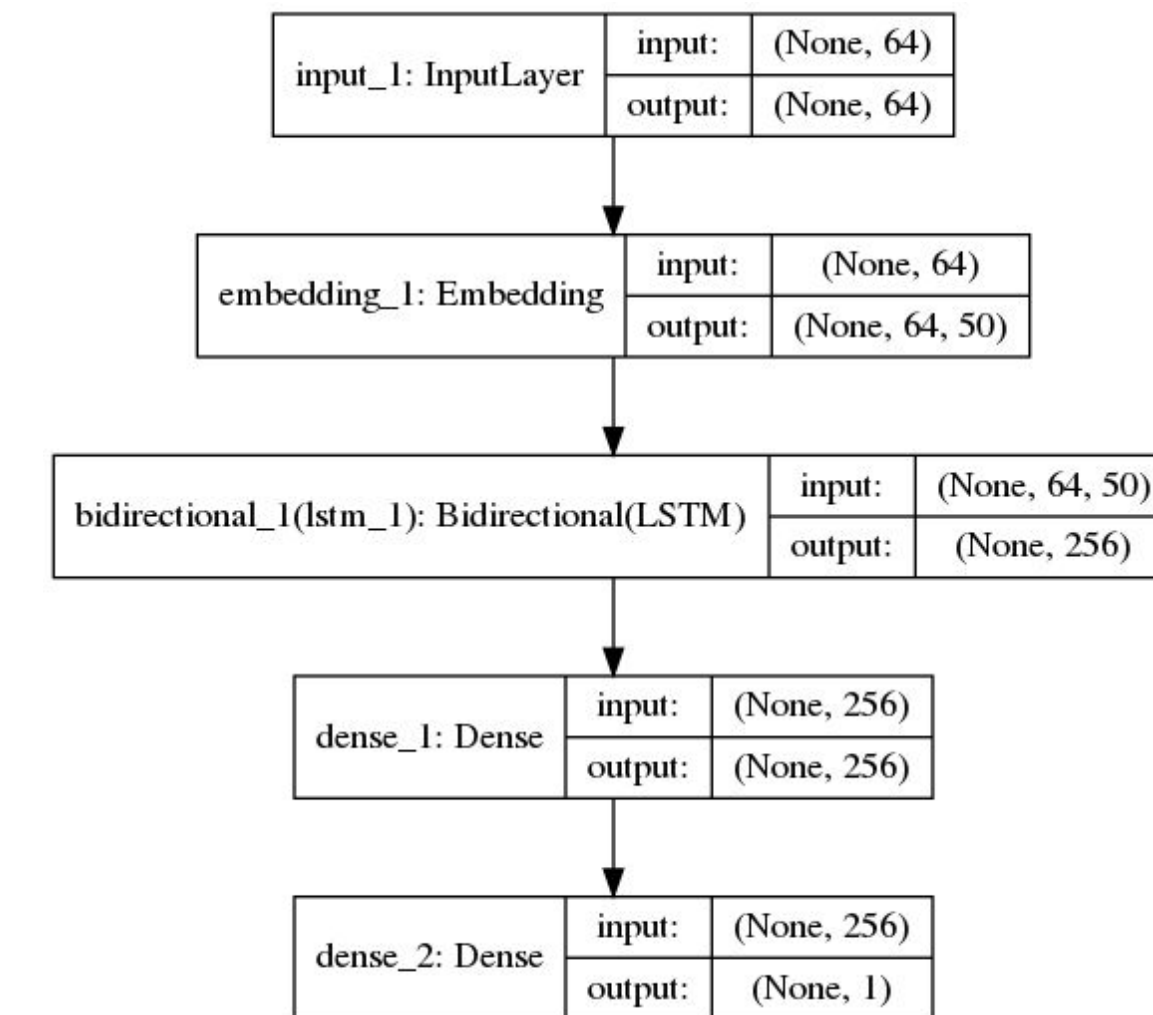
- **Syntax validity of name/address** through **Deep Learning** based Anomaly Detectors
- The **probability of the name/address belonging to a certain language** is calculated (language identification)
- Check if an address is recognized by **Open Street Map** service
- Check the match between **the fiscal code** and **the registrant's name**
- Check the match between company name, VAT number, and address using the **VIIES** (Vat Information Exchange System) service provided by the European Commission
- **Inconsistencies** between multiple fields: for example for a legal entity, the nationality must match the country of the registrant

# Deep Learning based Anomaly Detectors



To identify invalid personal names, company names and addresses, two **Anomaly Detectors based on Deep Learning** have been trained:

- **Supervised Anomaly Detector**
  - **Training Data:** It is trained using a **silver dataset** combining authentic and synthetic data
  - **Architecture:** Embedding and Bidirectional LSTM layers
  
- **Unsupervised Anomaly Detector**
  - **Training Data:** It is trained using **authentic** data
  - **Architecture:** It is based on **Autoencoder**  
Anomalies are flagged when the model fails to correctly reproduce a string, suggesting the presence of a fake name or address



# Darwin: a Registrars web app to fix the Registrants anomalies

- An application dedicated to managing the anomaly correction requested by **READS**
- **Registrars** can correct anomalies and submit any required documentation
- **Registry** can validate the corrections made by the Registrars
- Each **request** is related to a **registrant aggregation** calculated by READS
  - potentially impacting on one or more registrant contacts in the DBNA
- **DARWIN** interacts with domains and contact objects in the DBNA, applying the necessary modifications



Three overlapping green circles of varying shades are located in the top-left corner of the slide.

# Future Works

- Dataset collection
- Human in the Loop
- Domain names analysis for detecting malicious activities
- LLM based Anomaly Detector

# LLM for anomaly detection: preliminary tests

Prompt:

Given the registration data for a .it domain name, where the registrant is described as a **natural person** of **Italian nationality**, please evaluate the provided details for potential anomalies. The data includes:

**Fiscal code:** oodds742837482

**Name:** Topolino

**Address:** via Moruzzi 1112, 42513, Pisa, Italia

Are there any anomalies in this data that could indicate errors in data entry or potential fraud, considering typical formats and conventions for such information in Italy? Consider also the consistency between the union of the fields. Respond only with a JSON format with the fields "regcode", "name" and "address" as keys, and the anomalies recognized as values.

# LLM for anomaly detection<sup>(Cont)</sup>

```
{
```

```
  "regcode": "Incorrect length, not adhering to standard 16-character format. The structure does not match typical 'codice fiscale' conventions, indicating potential data entry errors or fraud.",
```

```
  "name": "Name 'Topolino' refers to a fictional character, which is highly unusual for official documents and suggests potential inaccuracies or misuse.",
```

```
  "address": "The postal code '42513' does not match the geographical location of Pisa, indicating an error. Additionally, the street number '1112' is unusually high for Italian addresses, raising questions about its accuracy."
```

```
}
```

# LLM for anomaly detection<sup>(Cont)</sup>

Prompt:

Given the registration data for a .it domain name, where the registrant is described as an **Italian company**, please evaluate the provided details for potential anomalies. **The VAT number corresponds to the company "POSTE ITALIANE SPA" with address "VIALE EUROPA 190 00144 ROMA RM"** . The provided data of the registrant includes:

VAT number: 01114601006  
Organization: Pippo srl  
Address: via Roma 1, 11111, Roma, Italia

Are there any anomalies in this data that could indicate errors in data entry or potential fraud, considering typical formats and conventions for such information in Italy?

Consider also the consistency between the union of the fields.

Respond only with a JSON format with the fields "regcode", "name" and "address" as keys, and the anomalies recognized as values.



# LLM for anomaly detection<sup>(Cont)</sup>

```
{
```

```
  "regcode": "Mismatch between the VAT number '01114601006' and the provided organization name. The VAT number corresponds to 'POSTE ITALIANE SPA', not 'Pippo srl', indicating a potential error or fraudulent intent.",
```

```
  "name": "The organization name 'Pippo srl' does not match the company associated with the provided VAT number, suggesting inaccuracies.",
```

```
  "address": "The postal code '11111' is not a valid postal code for Roma or any other location in Italy, further indicating potential data entry errors or the use of a placeholder."
```

```
}
```

# LLM for anomaly detection<sup>(Cont)</sup>

Prompt:

Given the registration data for a .it domain name, where the registrant is described as a **natural person** of **Italian nationality**, please evaluate the provided details for potential anomalies. The data includes:

**Fiscal code:** RSSMRA80L05F593A

**Name:** Mario Rossi

**Address:** via Roma 1, 11111, Roma, Italia

Are there any anomalies in this data that could indicate errors in data entry or potential fraud, considering typical formats and conventions for such information in Italy? Consider also the consistency between the union of the fields. Respond only with a JSON format with the fields "regcode", "name" and "address" as keys, and the anomalies recognized as values.

# LLM for anomaly detection<sup>(Cont)</sup>

```
{  
  
  "regcode": "No anomaly detected based on format.",  
  
  "name": "No anomaly, common Italian name.",  
  
  "address": "The postal code '11111' is not a valid postal code for any  
location in Italy, indicating a potential error in data entry or fraudulent  
intent."  
  
}
```



# Questions?

Daniele Sartiano

[daniele.sartiano@iit.cnr.it](mailto:daniele.sartiano@iit.cnr.it)

<https://www.iit.cnr.it/daniele.sartiano/>