

# LOCATION RESOLUTION USING TRANSFORMER



---

**Evelyn Wang, BAM Elevate**  
**June 11, 2024**

Evelyn Wang is a Data Scientist at BAM Elevate and is a presenter at the DataBricks Data +AI Summit. Any graphs, charts, illustrations are for illustrative purposes only and are not intended to be investment advice or investment recommendations.

# Agenda

1 Background & problem statement

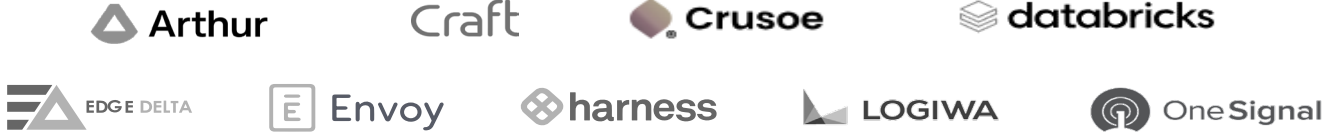


2 Our solution

3 Key takeaways

# Background & problem statement

# BAM Elevate is a late-stage VC firm

BAM Elevate is Balyasny Asset Management's ("BAM") late-stage venture strategy, focused on opportunities in high-growth, tech-enabled sectors

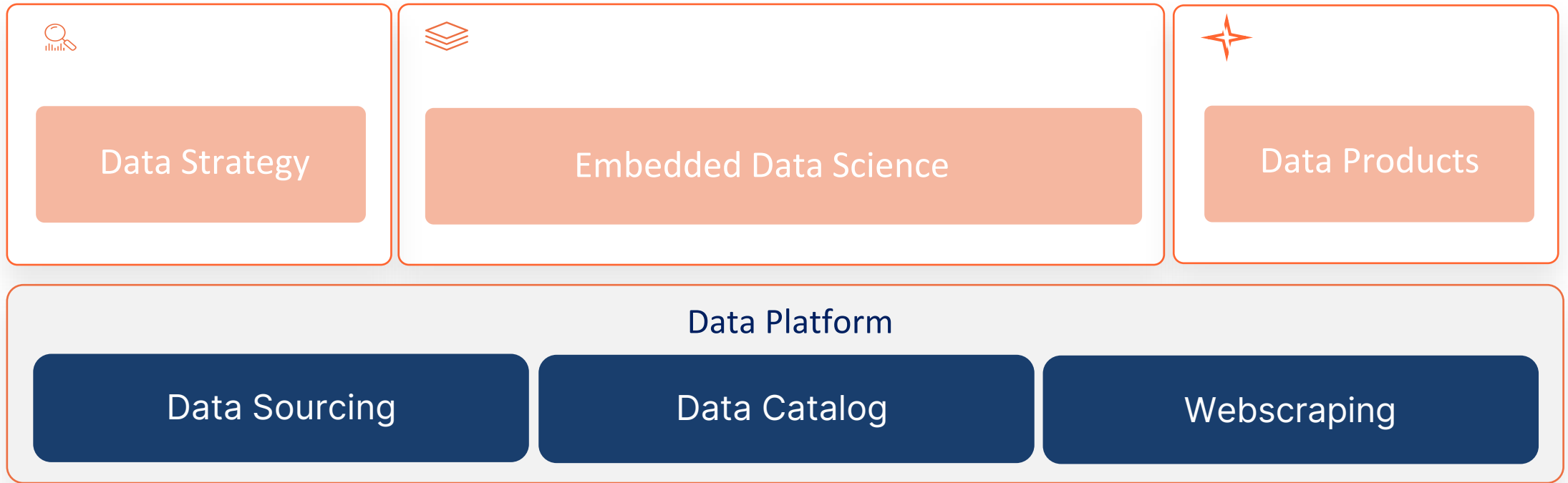
Enterprise	
FinTech	
HealthTech	

The logos shown on this page are the property of the respective companies and their inclusion does not indicate an endorsement of or by Balyasny Asset Management.

©2024 Databricks Inc. — All rights reserved



# We build data products and ML models for ourselves and our portfolio companies



# We needed to process unstructured text data to power an alerting system

## The Problem

One of our portfolio companies provides an enterprise supply chain intelligence platform. They want to automate the extraction of locations from news articles. It's one of the first steps to enable more relevant, timely risk alerts to be sent to their customers so that they can proactively mitigate supply chain risks.

**CNBC ENERGY**

### Harvey shuts down largest US oil refinery, bears down on Louisiana plants

PUBLISHED WED, AUG 30 2017-9:56 AM EDT | UPDATED WED, AUG 30 2017-2:43 PM EDT

Tom DiChristopher @TDICHRISTOPHER

SHARE f X in e

**KEY POINTS**

- Tropical Storm Harvey has shuttered about 20 percent of U.S. refining capacity.
- Motiva shut down its Port Arthur, Texas, refinery, the largest in the United States, and three other area refineries were also closed.
- Refiners in Corpus Christi, Texas, are aiming to restart, but facilities in the Houston area remain largely closed.

cnbc TV Power I UP NEXT | c|

BF

**AP**

### Massive cargo ship becomes wedged, blocks Egypt's Suez Canal



# Translating the business problem into a data science solution

## Example Input

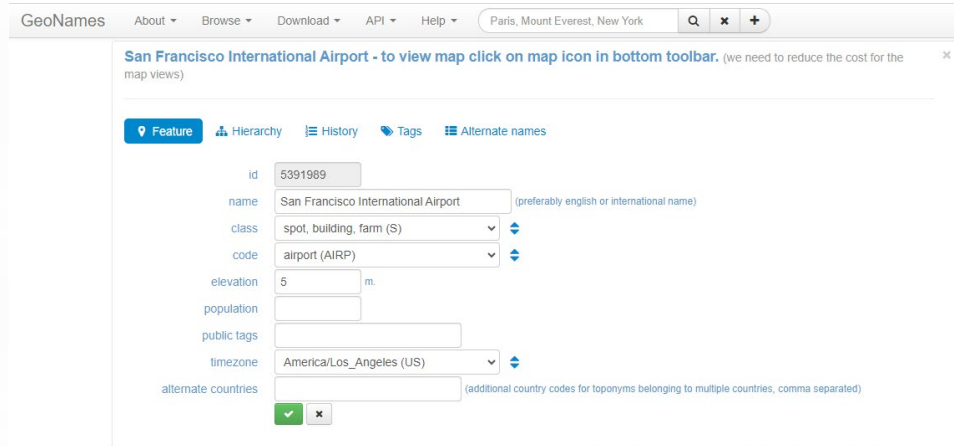
Orders from Federal Aviation Administration towers diverted one corporate jet from **San Francisco International Airport** to **Oakland**, while preventing other planes from touching down at **San Jose Mineta Airport**, ” according to the Chronicle.

## Example Output

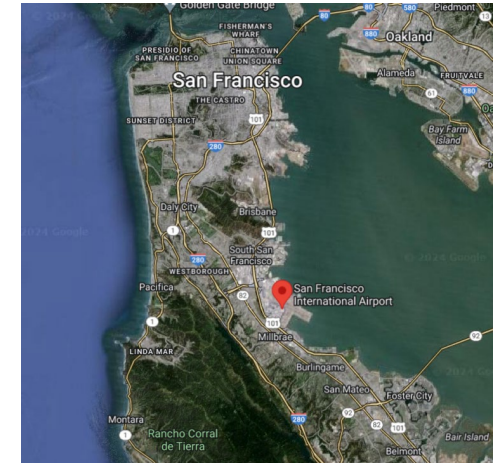
Name: **San Francisco International Airport**  
Start: 84  
End: 119  
Geonames ID: 5391989

Name: **Oakland**  
Start: 123  
End: 130  
Geonames ID: 5378538

Name: **San Jose Mineta Airport**  
Start: 188  
End: 211  
Geonames ID: 5392171



Geonames Record for San Francisco International Airport



Satellite Image of San Francisco International Airport

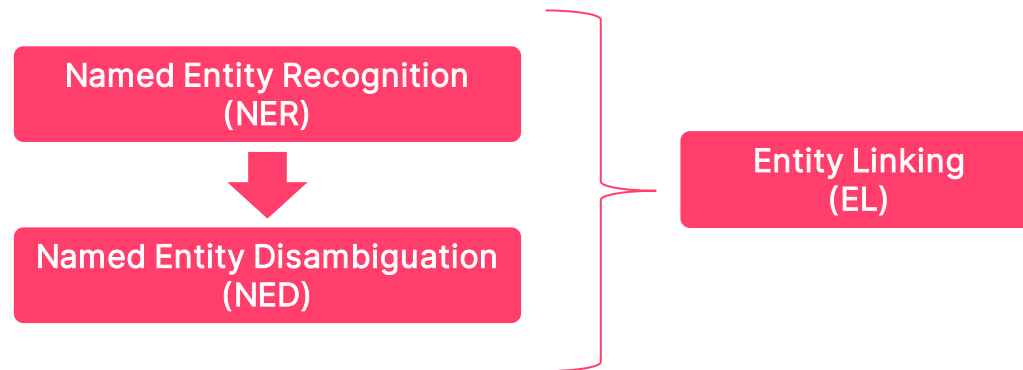
## High-level requirements:

1. Identify chunks in the text that reference location entities
2. Link each entity to their entry in Geonames (an open source DB containing over 25 million geographical names)

# Entity linking is a foundational NLP task

Entity linking is the task of automatically linking mentions of entities in text to their corresponding entries in a knowledge base

It can be broken down into two subtasks





# NER identifies entities in free text

**Finance:** tag companies in financial texts

**UBS** hopes the flexibility will boost its attractiveness as an employer in the banking sector. It has not yet set a date for employees' return to the office. Only **UBS** workers in roles that require them to be in the office, such as those in supervisory positions, or in trading and branch roles, will have less flexibility, the bank said. However, an internal analysis of the 72,000 **UBS** employees globally showed that around two-thirds are in roles that would allow them to combine working remotely and in the office.

**Healthcare:** tag drug names in texts

A pilot trial assessing the efficacy of **Dabigatran** in patients with atrial fibrillation and a study to prevent venous thromboembolism have both been conducted.

# NED grounds entity mentions to a knowledge base

Finance: link the company to its financial identifiers

UBS

## UBS Group AG (UBSG-CH)

UBS Group AG is a holding company, which engages in the provision of financial management solutions. It operates through the following segments: Global Wealth Management; Personal and Corporate Banking; Asset Management; Investment Bank, and Corporate Center. The Global Wealth Management segment advises and offers financial services to wealthy private clients except those served by Wealth Management Americas which include banking and lending, wealth planning, and investment management. The Personal and Corporate segment offers financial products and services to private, corporate, and institutional clients in Switzerland. The Asset Management segment consists of investment management products and services; platform solutions and advisory support to institutions; wholesale intermediaries, and wealth management clients. The Investment Bank segment comprises investment advice, financial solutions, and capital markets access among corporate, institutional, and wealth management clients. The Corporate Center segment is involved in the services, group asset and liability management and non-core and legacy portfolio. The company was founded on June 29, 1998 and is headquartered in Zurich, Switzerland.

### Sector & Industry

[Finance \(4800\)](#)  
[Major Banks \(4805\)](#)

### Fiscal Year End

December

### Exchange / ISIN

SIX Swiss / CH0244767585

### SEDOL

BRJL176

### Investor Relations Contact

[Martin A. Osinga](#)

### LEI

549300SZJ9VS8SGXAN81

Healthcare: link the drug to its record in DrugBank

Dabigatran

Generic Name	Dabigatran	DrugBank Accession Number	DB14726
Background	Dabigatran is the active form of the orally bioavailable prodrug <a href="#">dabigatran etexilate</a> .		
Type	Small Molecule	Groups	Approved, Investigational

# Entity linking is useful across domains

## Knowledge Graph

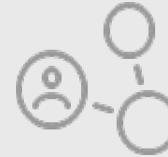


### Enrich a knowledge graph

#### Biomedical example:

Use EL to resolve diseases mentioned in clinical notes and then extract the relationships between these diseases to construct KG facts

## Question-Answering/RAG



### Create more precise queries to improve answer retrieval

#### Retail example:

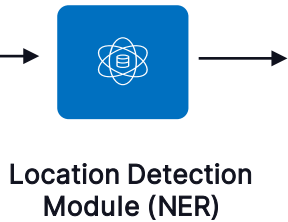
Resolve products mentioned in customer queries so that relevant info can be found in the knowledge base to provide helpful and accurate answers

# Our solution

# Overview

## Input: News article

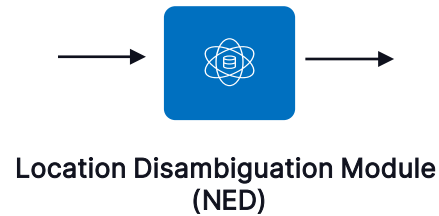
Orders from Federal Aviation Administration towers diverted one corporate jet from San Francisco International Airport to Oakland, while preventing other planes from touching down at San Jose Mineta Airport," according to the Chronicle.



## Location Detection Output:

- A list of locations that were detected and their positions in the input

```
[{'end': 119,
  'entity_group': 'LOC',
  'score': 0.922234,
  'start': 84,
  'word': ' San Francisco International Airport'},
 {'end': 130,
  'entity_group': 'LOC',
  'score': 0.97684264,
  'start': 123,
  'word': ' Oakland'},
 {'end': 211,
  'entity_group': 'LOC',
  'score': 0.93859875,
  'start': 188,
  'word': ' San Jose Mineta Airport'}]
```

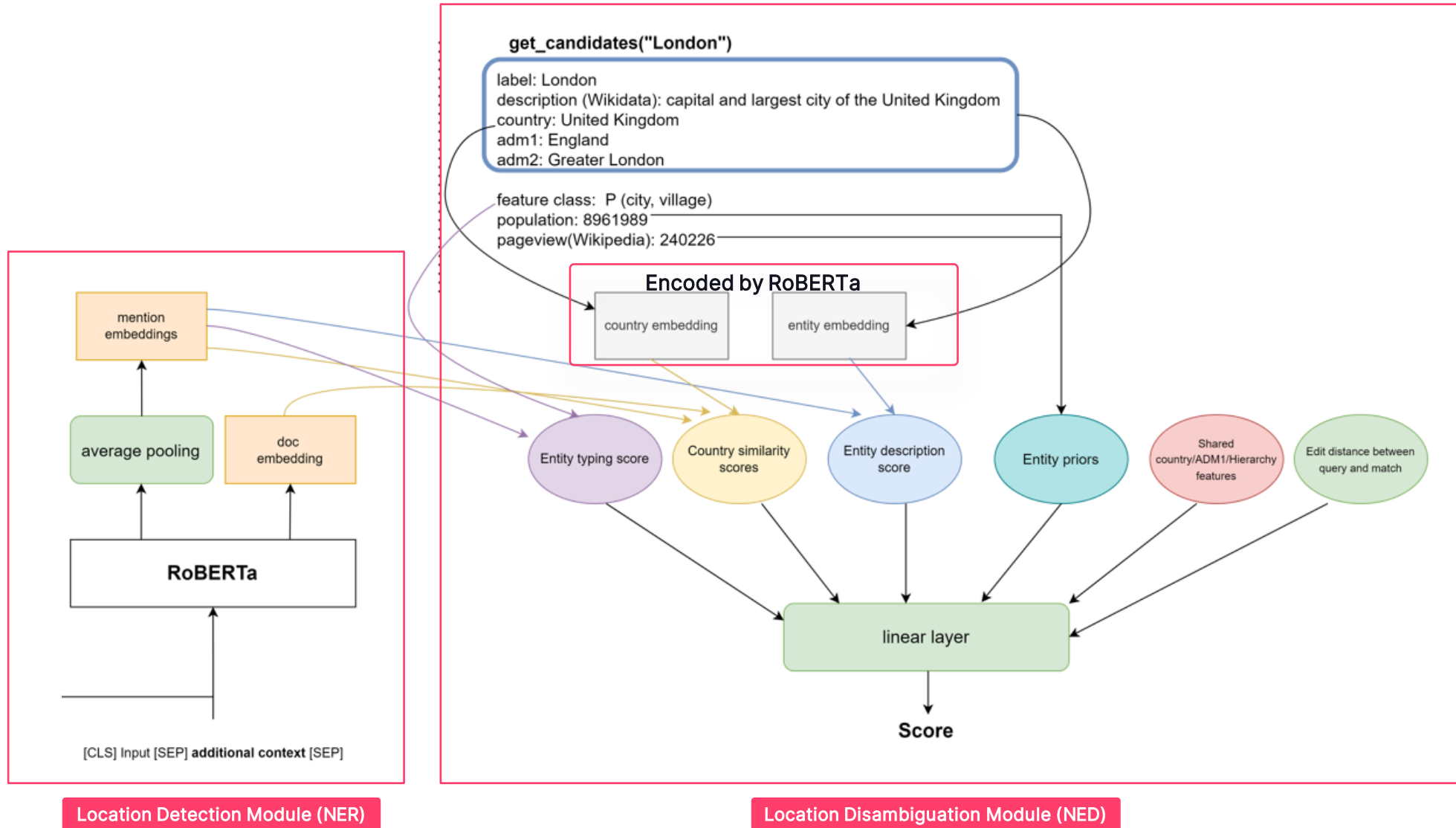


## Final Output:

- List of detected locations (from previous step) enriched with their Geonames information (i.e. id, coordinates)

```
[{'feature_code': 'AIRP',
  'feature_class': 'S',
  'country_code3': 'USA',
  'lat': 37.61882,
  'lon': -122.3758,
  'name': 'San Francisco International Airport',
  'admin1_code': 'CA',
  'admin1_name': 'California',
  'admin2_code': '081',
  'admin2_name': 'San Mateo County',
  'geonameid': '5391989',
  'score': 1.0,
  'search_name': 'San Francisco International Airport',
  'start_char': 84,
  'end_char': 119},
 {'feature_code': 'PPLA2',
  'feature_class': 'P',
  'country_code3': 'USA',
  'lat': 37.80437,
  'lon': -122.2708,
  'name': 'Oakland',
  'admin1_code': 'CA',
  'admin1_name': 'California',
  'admin2_code': '001',
  'admin2_name': 'Alameda County',
  'geonameid': '5378538',
  'score': 0.9999984502792358,
  'search_name': 'Oakland',
  'start_char': 123,
  'end_char': 130},
 {'feature_code': 'PPLA2',
  'feature_class': 'P',
  'country_code3': 'USA',
  'lat': 37.33939,
  'lon': -121.89496,
  'name': 'San Jose',
  'admin1_code': 'CA',
  'admin1_name': 'California',
  'admin2_code': '085',
  'admin2_name': 'Santa Clara County',
  'geonameid': '5392171',
  'score': 0.9993213415145874,
  'search_name': 'San Jose Mineta Airport',
  'start_char': 188,
  'end_char': 211}]
```

# Model Architecture Overview



# Location detection module

# Off-the-shelf NER taggers are trained on old and unspecialized data

While off-the-shelf Transformers-based models have exhibited great performance, there are limitations:

- **Temporal drift:** model training typically uses standard datasets (i.e. CoNLL-2003) that are decades old
- **Not optimized for locations:** struggle with identifying all locations in a text (low recall) and underperforms on fine-grained location types (i.e. airports, train stations)



# We annotated a custom training dataset and augmented it with public datasets

- We manually annotated 1914 locations in 251 supply chain news articles
- In addition to the custom dataset, we augmented it with five other publicly available datasets resulting 2x more locations in the training corpus vs. training data used by off-the-shelf models

1 Pacific Gas and Electric executives announced they will send crews to west Cottonwood<sup>LOC</sup> and Happy Valley<sup>LOC</sup> to install undergrounding power lines throughout 2023. That could mean traffic controls and some waits and detours for motorists, limited street parking at times for residents. But placing the lines underground will reduce the risk the utility company's equipment could spark wildfires, and will increase the power grid's reliability, according to a company announcement issued Tuesday. Undergrounding can cut "risk of ignitions along undergrounded circuits by 99%, reduce annual spending on temporary repairs and other recurring costs such as vegetation management; and curb the need for wildfire safety-related power outages. ". In Shasta County<sup>LOC</sup>, PG&E plans to install 77 miles of underground power lines this year, PG&E spokesman Paul Moreno said. The company's work schedule through 2023 covers the following areas:

Annotated Example of Our Data

Dataset	Descriptions
CoNLL++	Reuters news articles published in 2020
CoNLL-2003	Reuters news articles published in 1997
GeoWebNews	200 news articles collected in 2018
LGL	588 human-annotated news articles published by 78 local newspapers
TR-News	Created from news articles of different sources

Public datasets we used

	Articles	Sentences	Tokens	Unique Tokens	LOC
Training Set	899	28877	565690	41797	14256
Development Set	194	6675	135800	17679	3548
Test Set	195	6917	131000	17917	3457

Our Dataset

English data	LOC
Training set	7140
Development set	1837
Test set	1668

CoNLL-2003



# Fine-tuning a small LLM can yield robust performance

- Using LLMs like GPT-4 and Llama2 for location NER are expensive, have higher latency, and are not well-suited
- We fine-tuned a RoBERTA model (open-source, 125M parameters) on our training corpus

## Overall performance

	Precision	Recall	F1
Our model	0.78	0.86	<b>0.82</b>
bert-base-NER	0.63	0.66	0.64

Performance on the test set (N=195 articles)

Precision: what proportion of the locations identified were correct?

Recall: what proportion of actual locations were identified?

F1: weighted average of precision and recall

Description	Unique Count	Our model Recall	bert-base-NER Recall
Country, State	744	<b>0.75</b>	0.62
City, Town	574	<b>0.89</b>	0.69
Region, Continent	75	<b>0.76</b>	0.68
Building, Landmark, Road	34	<b>0.83</b>	0.68
Mountain, Lake	28	<b>0.79</b>	0.54

Breakdown by location types



# Location disambiguation module

# State of the art EL models only link to general knowledge bases

- State of the art EL systems (i.e. ReFINEd, BLINK, GENRE) are transformers-based while existing geoparsers are mostly rule-based or ML-based
- However, these systems only link to Wikipedia, which has less coverage of locations than Geonames



# We created a new Transformer-based EL system for Geonames

## Overview

### 1. Candidate generation

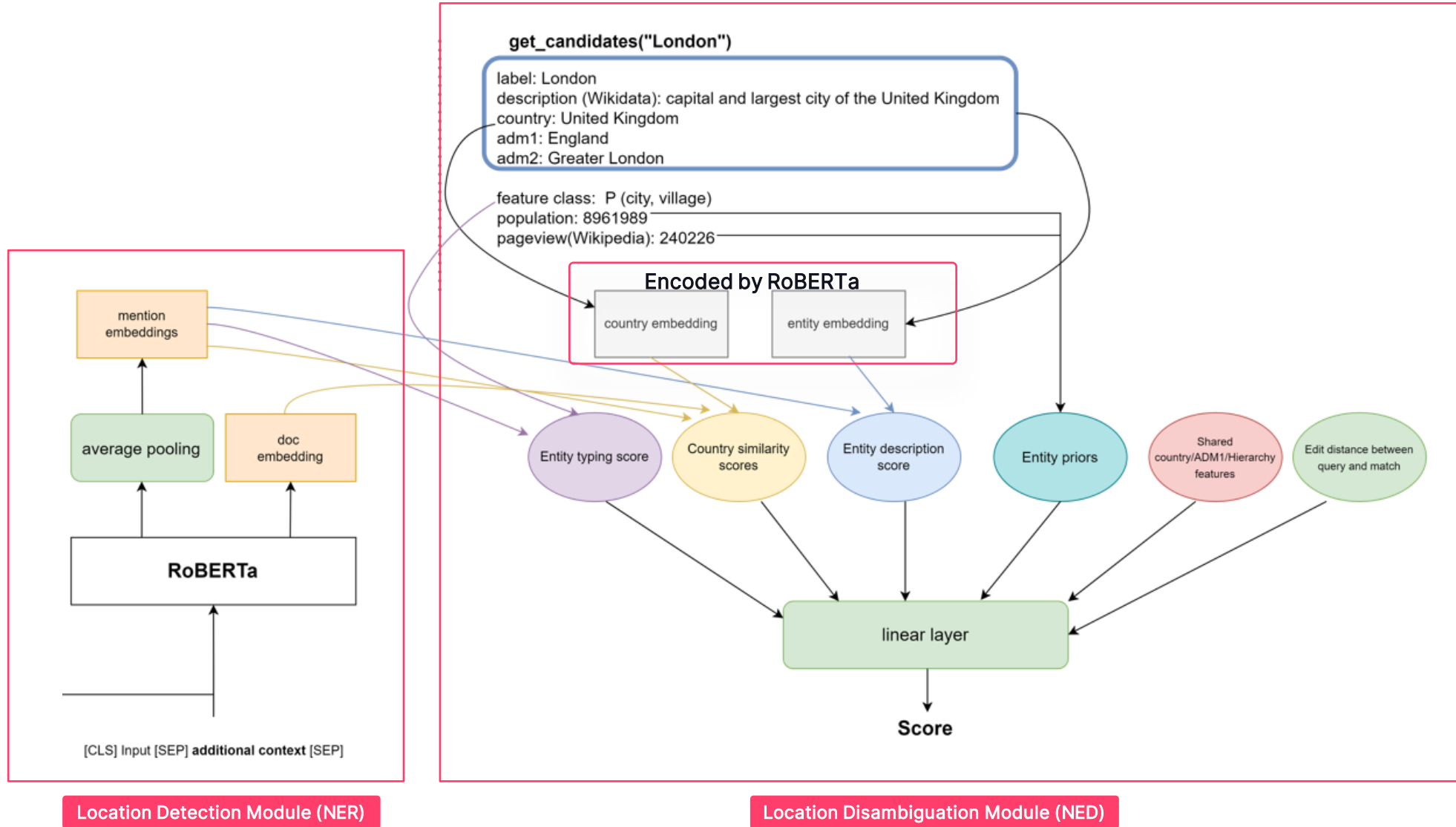
Select top-K candidate locations for identified location (from the previous step) using Elasticsearch

### 2. Evaluate the candidate

Transformer-based classifier evaluates the match between the input location and candidate by producing a score that is based on the location type, description, candidate priors (i.e population size) and other similarity features

### Training Details

1. Pre-training on Wikipedia-derived data
2. Fine-tuning on training corpus (custom + public datasets)



# Our EL system is faster and more accurate

Performance on location disambiguation only

- Exact Match: 0.76
- Accuracy@161km: 0.86

End to end performance

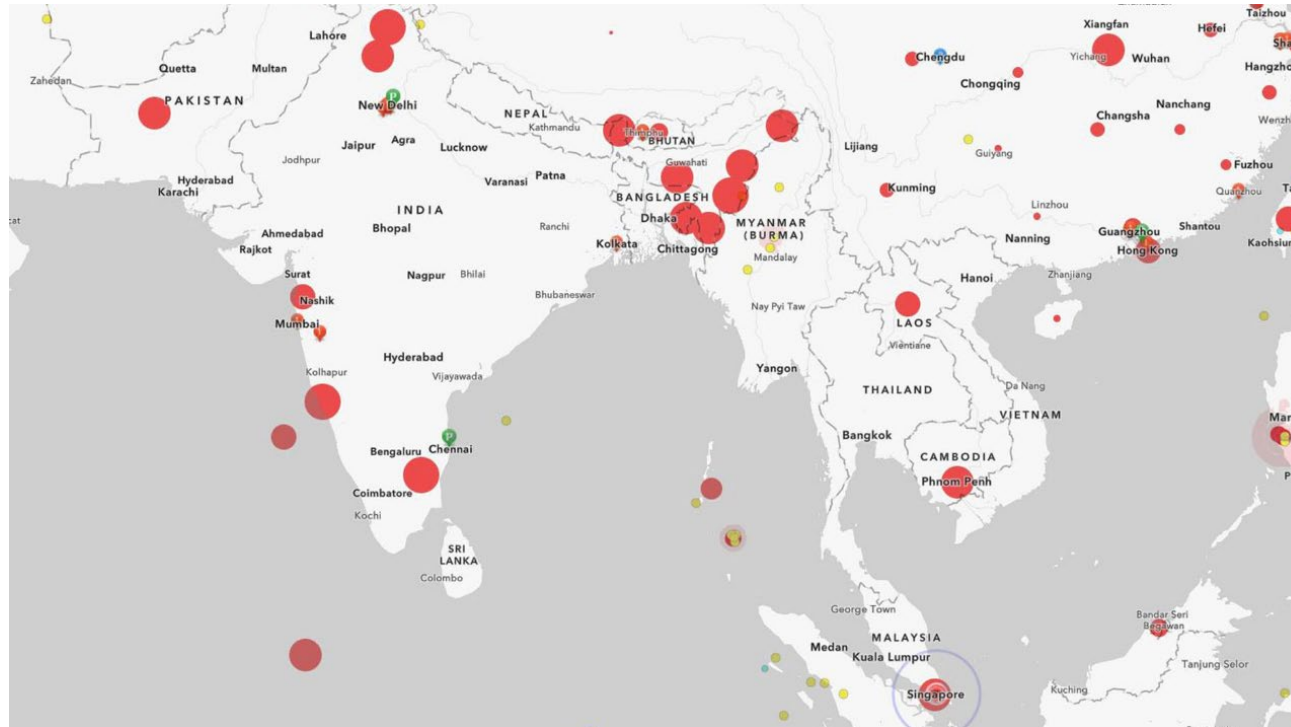
Methods	Precision	Recall	F1	Inference speed (per news article)
Our system	0.64	0.67	<b>0.65</b>	<b>1s</b>
ReFINED	0.70	0.48	0.57	3.5s
Naïve Baseline (first result returned by ES)	0.18	0.19	0.18	0.5s

Description	Our system Recall	ReFINED Recall
Country, State	<b>0.71</b>	0.57
City, Town	<b>0.80</b>	0.45
Region, Continent	0.33	0.47
Building, Landmark, Road	<b>0.59</b>	0.10
Mountain, Lake	<b>0.57</b>	0.43



# Powering AI-driven alerts and beyond

- Our portfolio company is integrating this system into their platform to enhance signal and event tracking by locations and deliver proactive risk insights via AI-driven alerts
- They are also looking to scale our approach and learnings to other use cases (i.e. company resolution)



# Key Takeaways



# Takeaway #1: Don't re-invent the wheel

- **Consider off-the-shelf solutions first**
  - There are several off the shelf entity linking models with robust performance for generic use cases (i.e. linking to Wikipedia)
- **For more niche, domain-specific EL tasks, it might make sense to try finetuning a custom model**
  - Finetuning smaller LLMs can yield competitive performance
  - Compared to prompting, it's easier to tailor, evaluate and more cost-effective

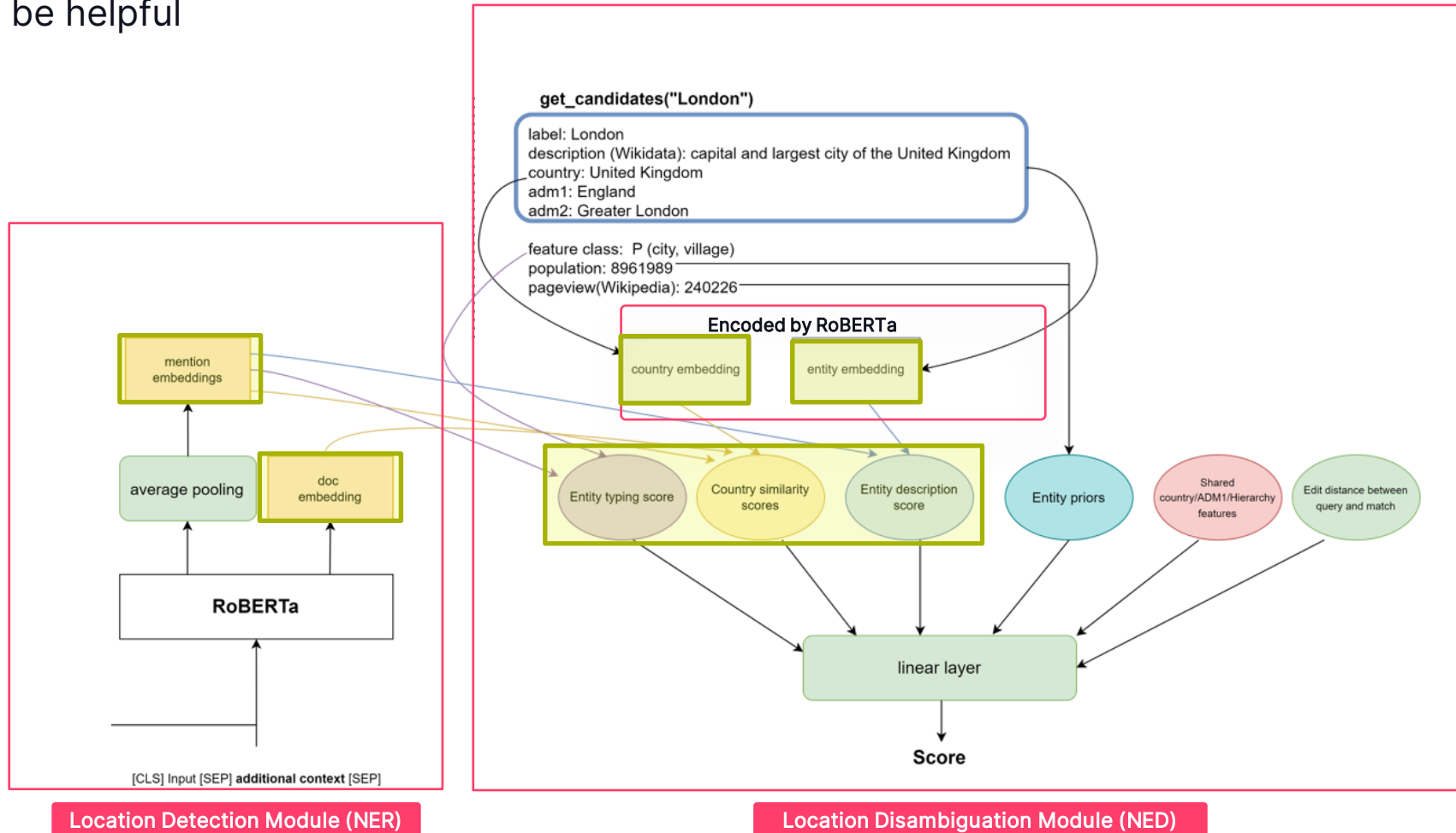
Methods	Precision	Recall	F1	Inference speed (per news article)
Our system	0.64	0.67	<b>0.65</b>	<b>1s</b>
ReFINED	0.70	0.48	0.57	3.5s
Naïve Baseline (first result returned by ES)	0.18	0.19	0.18	0.5s

- **Be creative with your datasets**
  - Whenever possible, do try to collect high quality annotations for your specific task – it's worth the lift!
  - We bootstrapped our custom dataset with publicly available datasets and Wikipedia-derived data



# Takeaway #2: Context matters

Especially for longer documents, any component you can build into your model architecture to leverage the context can be helpful



We computed scores between contextual embeddings of the detected location and the candidate



# Takeaway #3: Thoughtful deployment

- Consider hosting your model behind a REST endpoint
- Work with business stakeholders to set up guardrails for model errors



# Appendix



# Pretrain Data Details

- Idea: train on a large, generic corpus before adapting the model to Craft dataset
- How the data is derived:
  - From Kensho Derived Wikimedia Dataset (a link annotated corpus of English Wikipedia pages)
    - Select if the linked Wikipedia pages from the text has a Geonames ID
  - Stats:
    - # of pages: 462032
    - # of unique Geoname id: 99557
    - # of countries: 249

feature_class	Total ID	Unique ID	description
P	543009	72559	city, village
S	114619	17031	spot, building, farm
T	33046	6670	mountain, hill, rock
A	407787	2963	country, state, region
H	7497	182	stream, lake
L	4105	142	parks, area
R	29	6	road, railroad
V	11	3	forest, heath
NA	4626	1	NaN

