

# The Problem: Text => Geo Locations in Text

This and That and the Other street in Porters Lake Nova Scotia

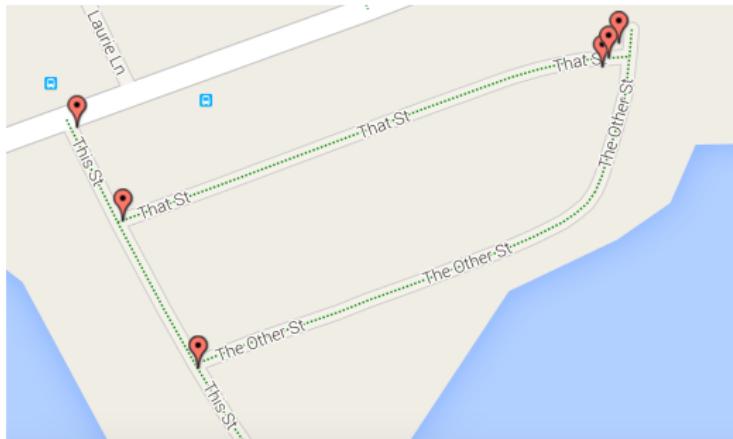
Q. How many locations are in this text?

# The Problem: Text => Geo Locations in Text

This and That and the Other street in Porters Lake Nova Scotia

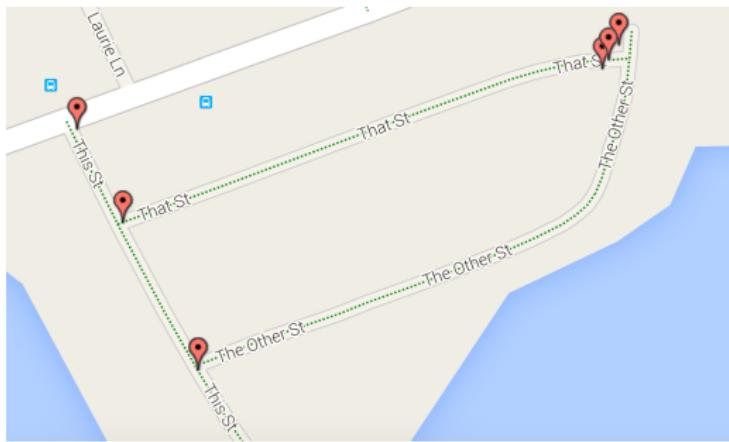
Q. How many locations are in this text?

A. 6.



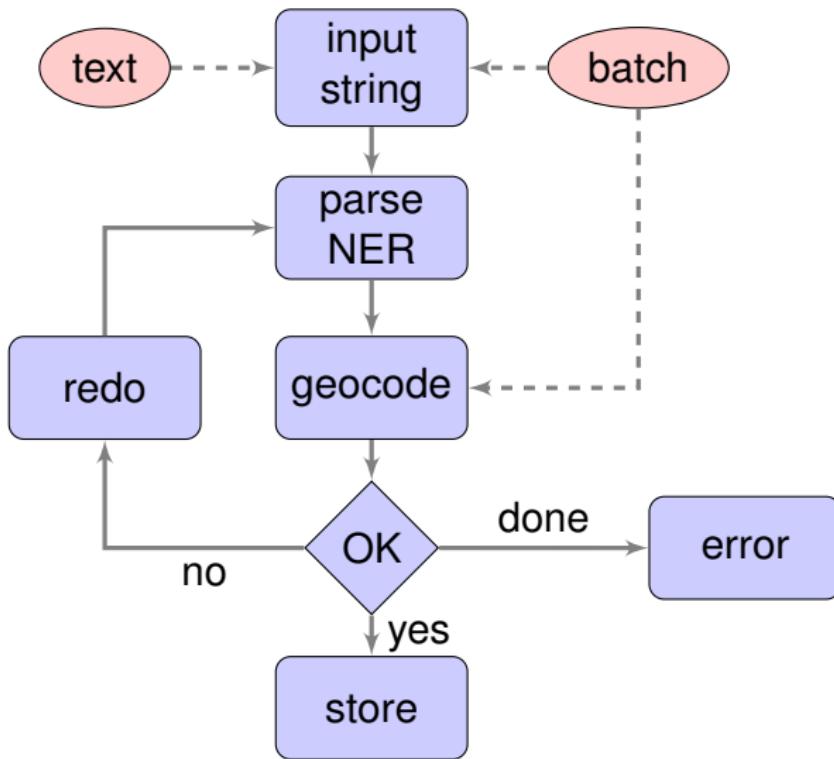
# The Problem: Text => Geo Locations in Text

- 1 THIS ST AND THAT ST, PORTERS LAKE, NS
- 2 THIS ST, PORTERS LAKE, NS
- 3 THAT ST, PORTERS LAKE, NS
- 4 THE OTHER ST AND THAT ST, PORTERS LAKE, NS
- 5 THE OTHER ST AND THIS ST, PORTERS LAKE, NS
- 6 THE OTHER ST, PORTERS LAKE, NS



Reqs: Identify addresses, intersections, city names, province/state.

# The Solution: Geocoding



# Geocoding Spans Many Fields

- 1 linguistics (matching/translating across different languages)
  - 2 data processing (normalization, standardization and input)
  - 2.1 data structures (R-trees, KD-trees,...)
  - 3 natural language processing (parsing, named entity recognition)
  - 4 computational geometry (point in polygon)
  - 5 pattern recognition (fuzzy match)
  - 6 geography (dealing with projections)
  - 7 Ai (learning, hidden markov models)
  - ... and a few others (tokenization, data cleanup, UI..)
- AND** Testing, testing and more testing

# And there are Many Geocoders.

And Many more are being built, Plus a few I've tested:

- ① Google Geocoder (Coverage: 99%, Accurate 93%) (Canada)
- ② HERE.com (Coverage 98%, Accurate 92%) (Canada)
- ③ Nominatim (Coverage 80%, Accurate 57%) (Canada)
- ④ Geocoder.ca (Coverage 99%, Accurate 94%) (Canada)
- ⑤ Geocode.xyz (Coverage 80%, Accurate 58%) (Spain)
- ⑥ Mapzen.com (Coverage 86%, Accurate 80%) (Spain)

Download test data and results here: <https://github.com/eruci/openaddresses/tree/master/test>

## Why create a new Geocoder?

No Geocoder does it all. Google Geocoder (presumably the most complete in the bunch) does not

- ① Geocode parcel data (avail as opendata in Canada and USA)
- ② Extract location data from text
- ③ Do address parsing and standardization (incl postal codes)
- ④ Return all addresses that match a partial address
- ⑤ Provide 100% coverage (open problem)
- ⑥ Provide 100% accuracy (open problem too)

## Accuracy and Coverage Differ Because

- a Geocoding is an imprecise process and various Geocoders fail in various ways.
- b Ambiguities, incomplete data, incorrect data, software bugs, are the main causes

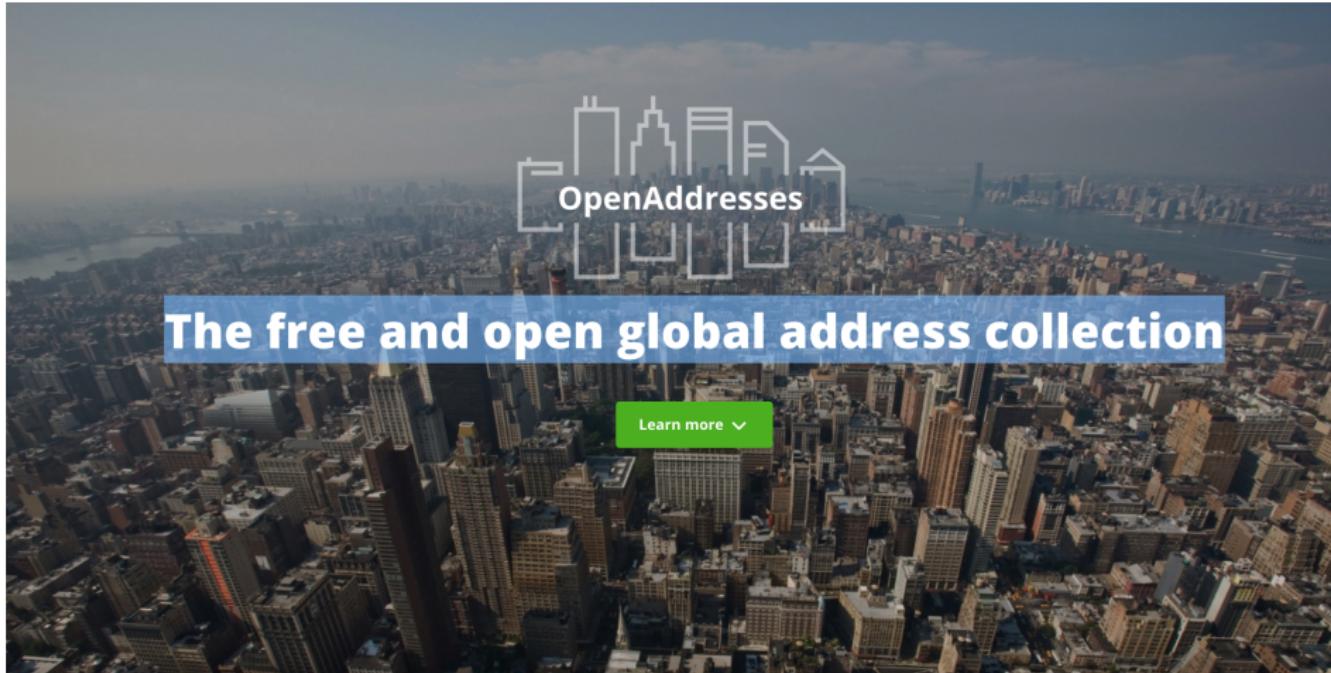
# The key ingredients of the solution

- 1 DATA
- 2 A good parser

# Parsing

- 1 Fuzzy vs Exact (correct spelling errors)
- 2 Partial vs Complete (fill in missing location entities)
- Quick demo: <http://geocoder.ca/textscan>

# Data



over 218,548,165 addresses (was half that only 6 months ago)

# Even Google Maps (presumably the best) Fails!

- ✖ Even in well mapped big cities.

# Even when you Google wrong location you get:



wrong location |



wrong location

wrong location **on google maps**

wrong location **on google**

wrong location **on facebook**

wrong location **tinder**

wrong location **on iphone**

wrong location **on phone**

wrong location **android**

wrong location **on traffic ticket**

wrong location **on parking ticket**



: Wrong Location Google Maps!.

# Carrer de Colomines, 2 Entresuelo 200AA, Barcelona, Catalunya 08003, Spain

≡ Carrer de Colomines, 2 Entresuelo 2<sup>a</sup>, Barcelona, C. X

🔍

📍 Search nearby: hotels · restaurants

Ank

Bruc 19, Entresuelo 2<sup>a</sup>  
08010 Barcelona  
Spain

Directions Save +34 717 12 75 79

Send to device

Ad Book a room

Check-in Thu, Aug 20 Check-out Fri, Aug 21

Booking.com \$104 / night Book

BudgetPlaces.com \$104 / night Book

View 1 more booking option at \$103

Street View...

Ank

Mango Outlet

Grupotel Gran Via 678

Carrer de les Corts Catalanes

Carrer del Bosc

Carrer d'AUS

Carrer de la Llúria

Carrer dels Sants

Carrer de Caspe

Ank

: my Airbnb in Barcelona: Wrong!.

## GeoCode.xyz

API

BARCELONA, ES » 2 COLOMINES, BARCELONA, ES » [41.3855238000,2.1789661000 Directions](#)

2 COLOMINES Calle, BARCELONA, ES ( BARCELONA,ES polygon ) [Directions](#) [Reverse Geocode](#)

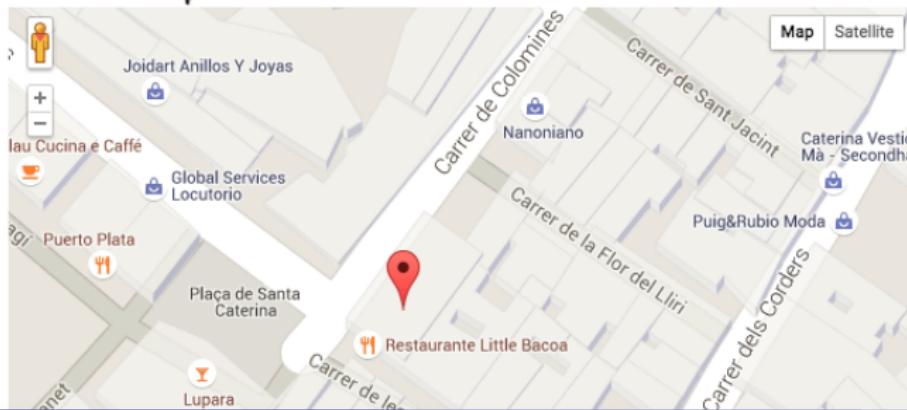
Confidence Score: 0.0

Is the location shown in the Map incorrect?

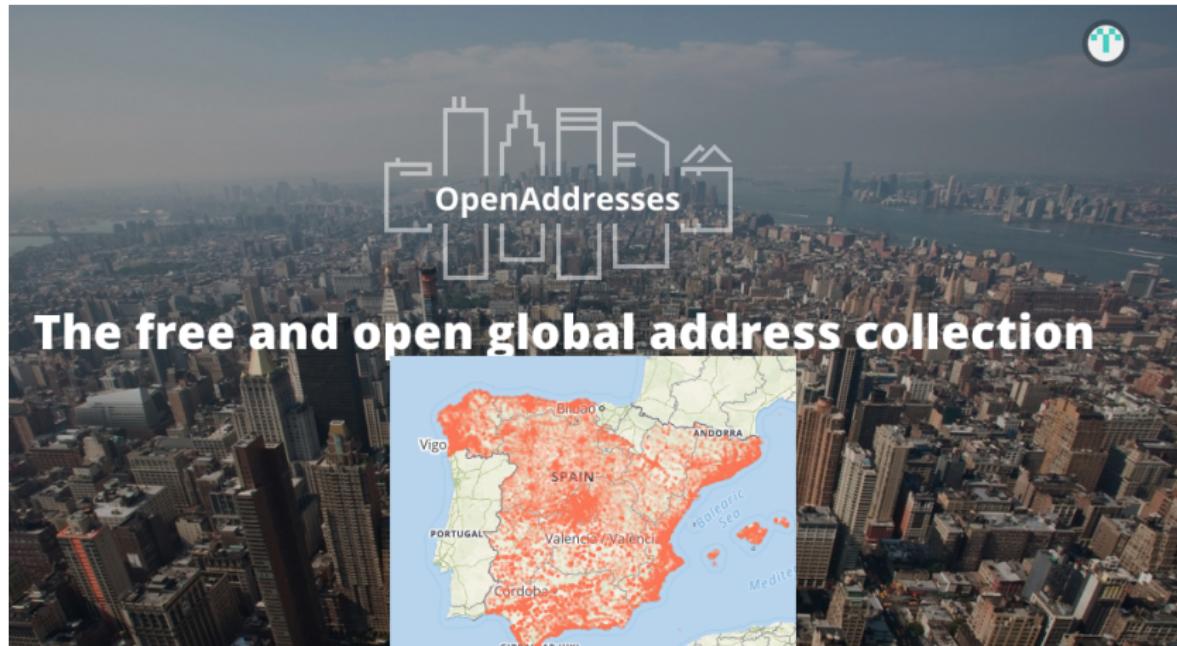
Then click here to send your corrections

Drag the marker ( ) to correct this location

41.3855238000, 2.1789661000  
 Geocode this Location on a Map



# Correct, but, Coverage for Spain is limited...



: [Data]

# Geoparsing and Geocoding in unstructured text

Text (from wikipedia entries to microblog posts) => Geocoded Locations.

- 1 extraction
- 2 disambiguation
- 3 geolocation

Demo: <http://geocode.xyz>

# In case internet does not work for the demo

**GeoCode.xyz** API

2 locations within a 1.084 km radius, found in this text:

The most important museums of Amsterdam are located on the Museumplein (Museum Square), located at the southwestern side of the Rijksmuseum.

Reprocess text and Download results On a Map

Match	Location	Confidence Score
1	Amsterdam, NL	0.7
2	Museumplein, Amsterdam, NL	0.3

Click a marker for more information.

: [Demo geocode.xyz]

# In case internet does not work for the demo

GeoCode.xyz API

Found 2 locations in this line:

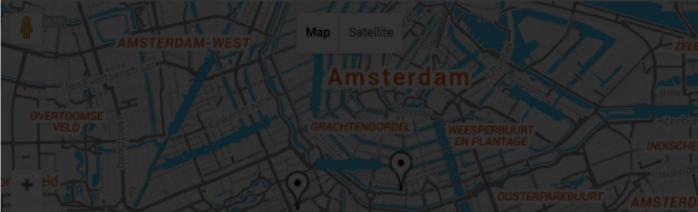
The most important museums of Amsterdam<sup>2</sup> are located on the Museumplein<sup>2</sup> (Museum Square), located at the southwestern side of the Rijksmuseum

2 locations within a 1.084 km radius of the Rijksmuseum.

The most important museum in Amsterdam is the Rijksmuseum.

Reprocess text and Data

1. [Amsterdam, NL](#) (Confidence: 0.7)  
2. [MUSEUMPLEIN, AMSTERDAM, NL](#) (Confidence: 0.3)



A map of Amsterdam showing the central area with the Grachtengordel (canal belt) and various neighborhoods labeled: AMSTERDAM-WEST, OVERTOMSE VELD, GRACHTENGORDEL, WEESPERBUIKT EN PLANTAGE, DOSTERPARKBUURT, INDIISCHE BUURT, and ZEEBURG. A red dot marks the location of the Museumplein.

Match	Location	Confidence Score
1	Amsterdam, NL	0.7
2	Museumplein, Amsterdam, NL	0.3

Click a marker for more information.

: [Demo geocode.xyz]

# Coding a Geocoder that does this is easy (in theory)

But.. making it recognize over 90% of input at over 90% accuracy requires at least these steps

- 1 importing and parsing country specific data from openaddresses.io (suffixes, prefixes, city names, numbering schemes)
- 2 cleaning up errors post import.
- 3 test and pick away at errors, one at a time

In theory, there is no difference between theory and practice. But, in practice, there is.

That is where you come in!

## Source Code / Data

Source code and Data: <http://geocode.xyz>

Just grab the server image on AWS, it is free for a micro instance

G  
e  
o  
c  
o  
d  
e  
r  
.c  
a

If you need help: e: [eruci@geocoder.ca](mailto:eruci@geocoder.ca) twitter: @geolytica  
*PS. One more thing. The core module is 47355 lines of Perl code.*