

# EMERGING MARKETS: LOCATIONS IN GEO-SPATIAL DOCUMENTS

GYAN RANJAN\*, JUONG-SIK LEE<sup>+</sup>, DEEPTI CHAFEKAR<sup>+</sup> AND UMESH CHANDRA<sup>+</sup>  
 \*Department of Computer Science and Engineering, University of Minnesota, Twin Cities, MN, USA  
<sup>+</sup>Nokia Research Center, Palo Alto, California, USA

## Geo-coding in Emerging Markets

**Problem:** Given a corpus of address texts  $\mathcal{A} = \{A : A \text{ is an address}\}$ , such that  $\forall A \in \mathcal{A}$ , the geo-spatial expanse of  $A$  is known  $(lat_A, lon_A)$ , and a query address  $Q$ , find the address/es that *best* determine the geo-spatial expanse of  $Q$ .

### Challenges in emerging markets:

- Lack of official/standard naming schemes for streets and areas.
- Lack of cartographic data.
- Perceived addresses are different from official postal addresses.

**Question:** How do we design a geo-coding system from scratch for emerging markets?

## Example: A Successful Resolution Here

### Query I:

800, Page Mill Road, Palo Alto, CA 94304.

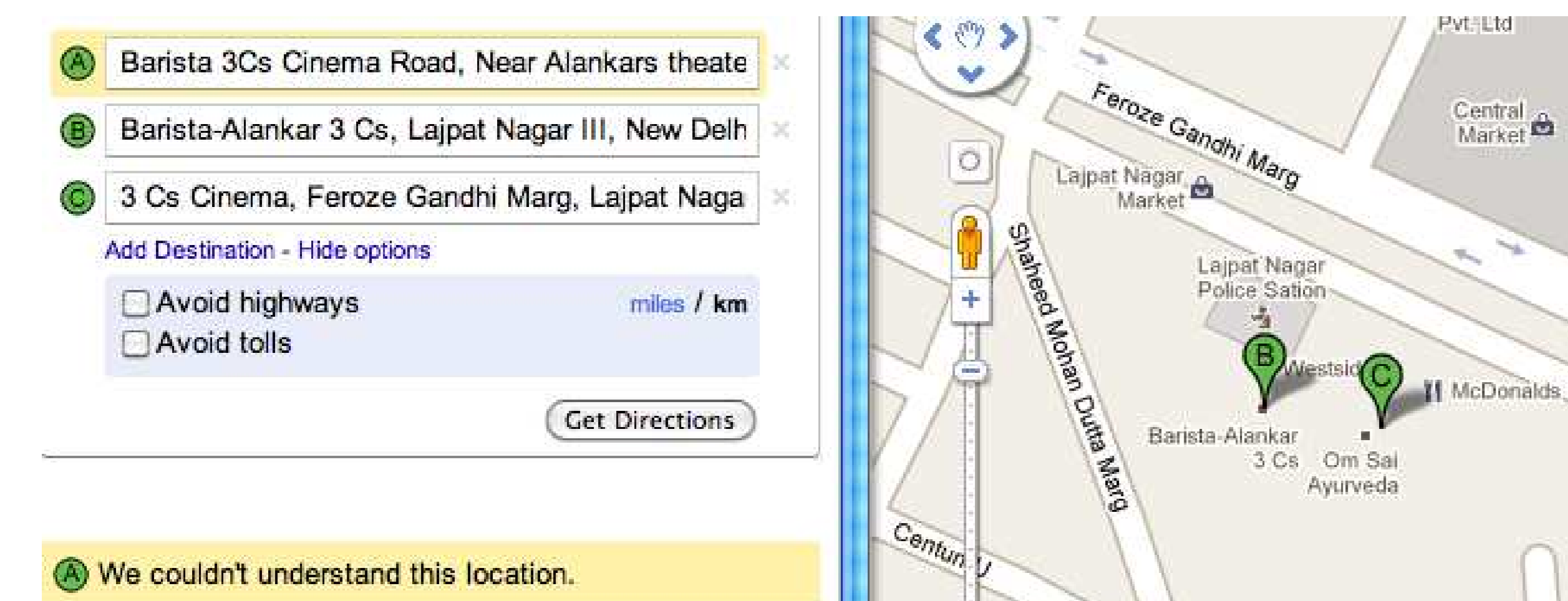


- Block structure in town planning.
- Address resolved to the intersection of 8<sup>th</sup> block on Page Mill Road, in zip-code 94304.
- High granularity resolutions possible.

## Example: And a Not-So-Successful One There

### Query II:

Barista, 3 Cs Cinema Road, Alankars Theater, Lajpat Nagar III, New Delhi.

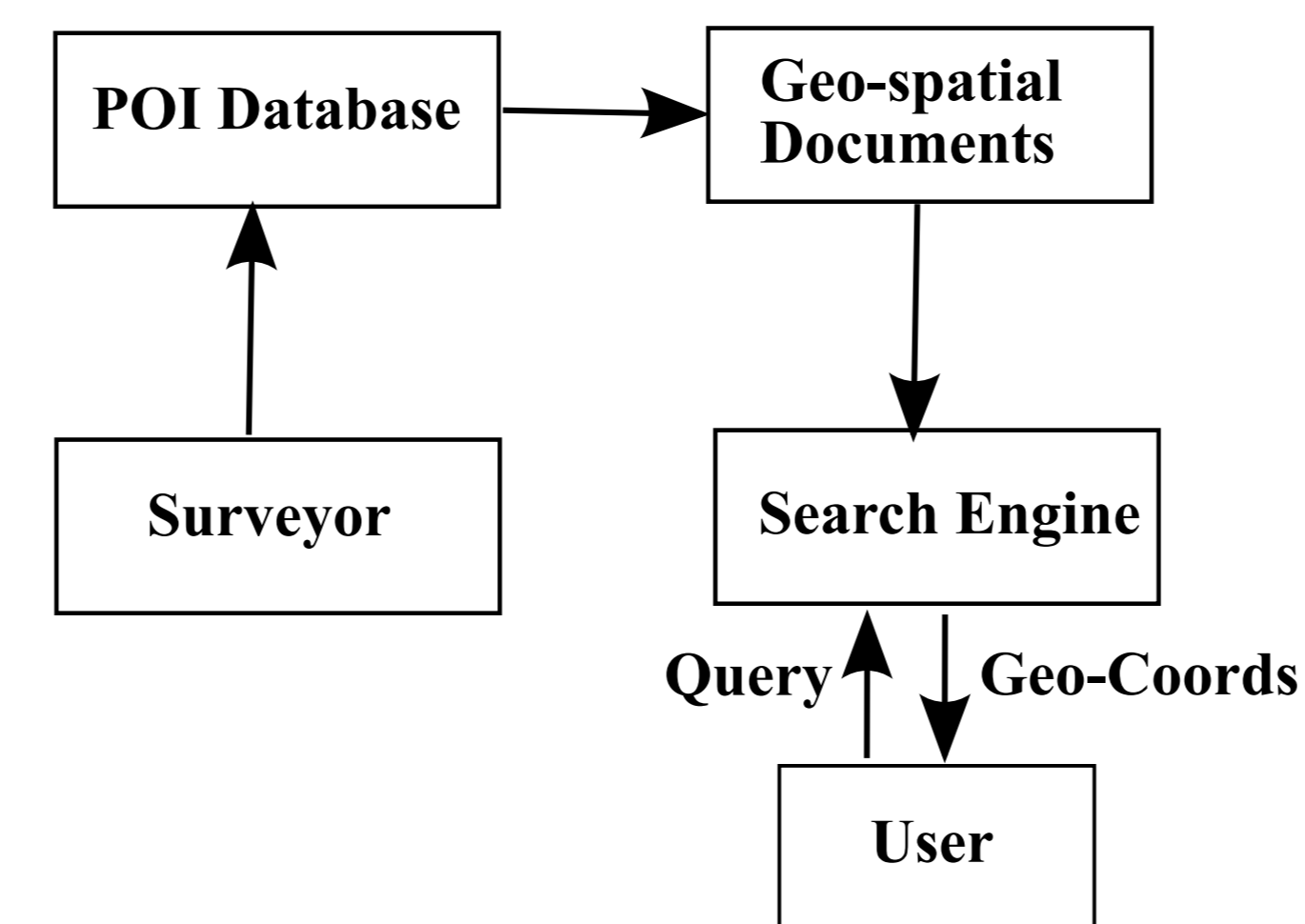


- **POSTAL:** Feroze Gandhi Marg, **PERCEIVED:** 3 Cs Cinema Road.
- The queried address is a combination of two addresses in corpus.
- The result is a **FALSE** negative.

**Implication:** Playing SAFE may result in FALSE NEGATIVES.

## A Framework for Emerging Markets

### The System



- A corpus of geo-annotated address texts for Points-Of-Interests in a city.
- An index and search mechanism.
- An interactive client.

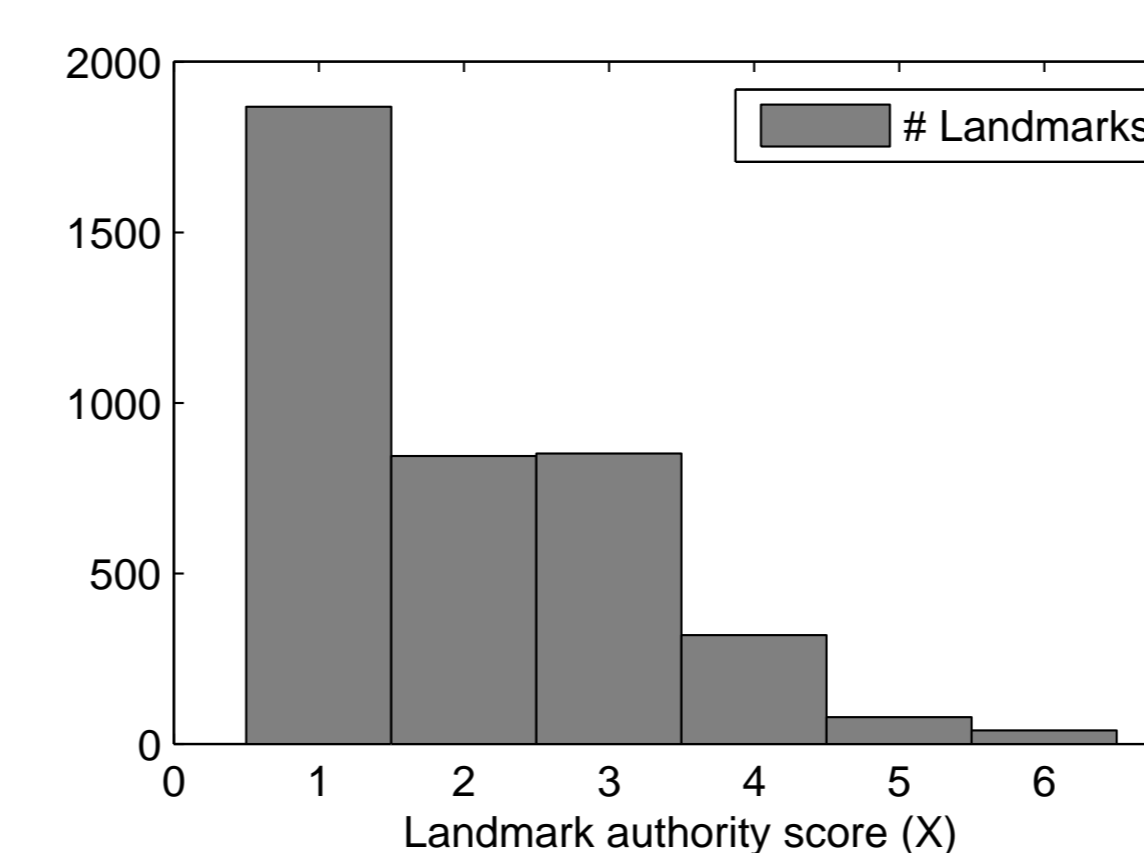
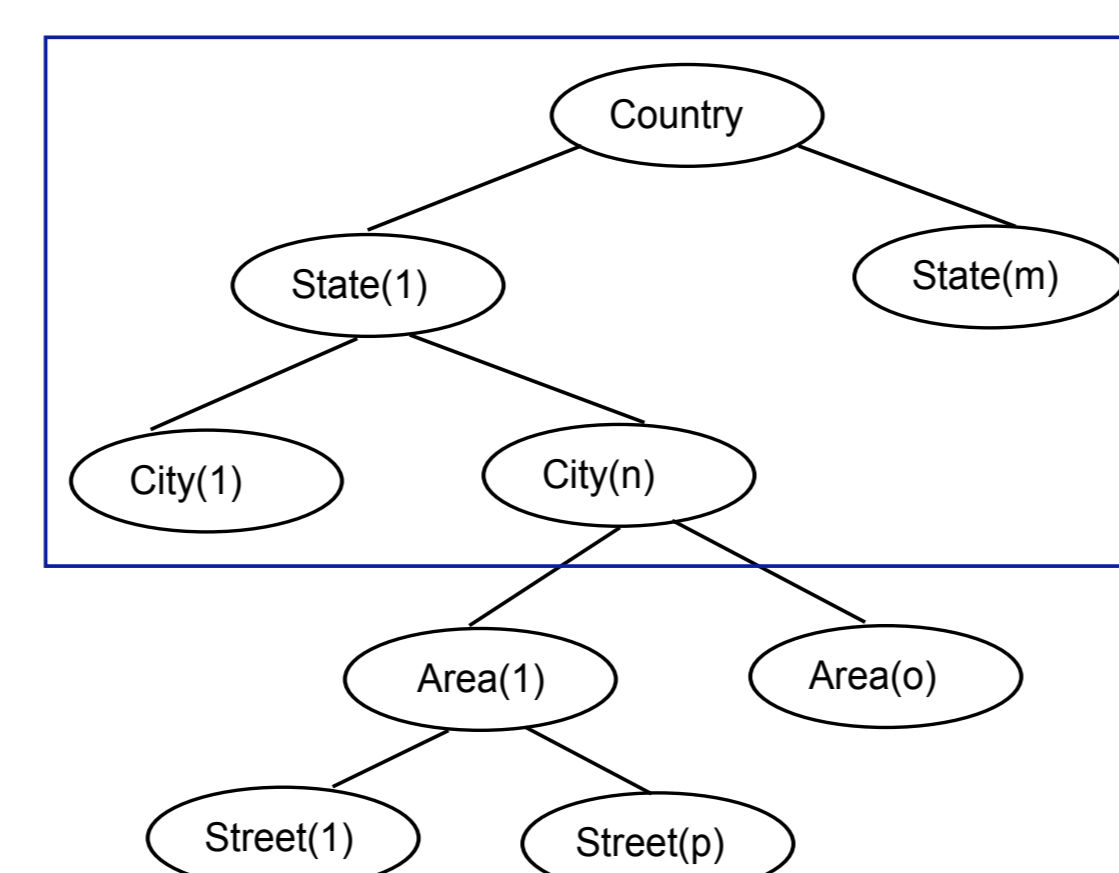
### The Survey

Register Places	Register Places	Register Places
Name (eg: Nokia Office) abc	Street Info	Area Name (eg: Marathalli Flyover) abc
Building Number	Landmark (eg: near New Horizon College)	State (eg: Karnataka)
Building Name	Additional Landmark abc	Pincode (eg: 560050)
Next Back	Next Back	Next Back

- Dedicated surveyors visit and annotate addresses for Points-Of-Interests in **SIX** Indian cities.
- Manual classification of sub-texts in addresses such as street names and landmarks.
- **POSTAL** and **PERCEIVED** addresses collected for each Point-Of-Interest.
- Each surveyed address has a pair of latitude-longitude co-ordinate associated with it.

**Comment:** With increasing adoption of GPS enabled devices, such surveys can be crowd-sourced in the not-so-distant future.

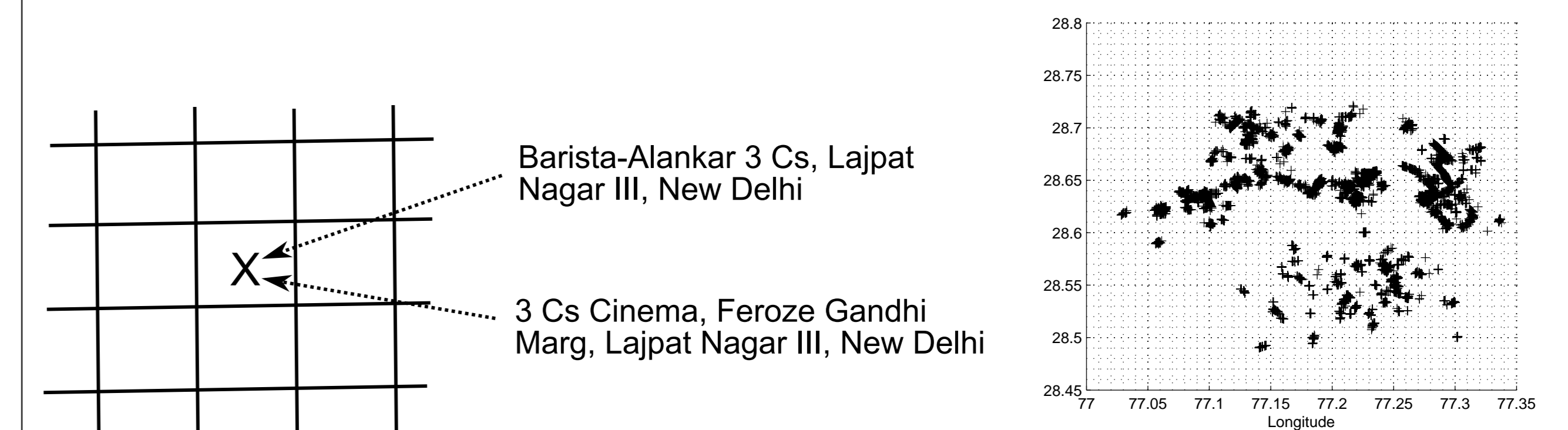
## Landmarks in Addresses



- Loose geospatial hierarchy in addresses.
- Standardization to the granularity of a city.
- Significant presence of landmarks in surveyed addresses.

**Reduced problem:** Geo-coding of address queries to within a city.

## From Addresses to Geo-spatial Documents

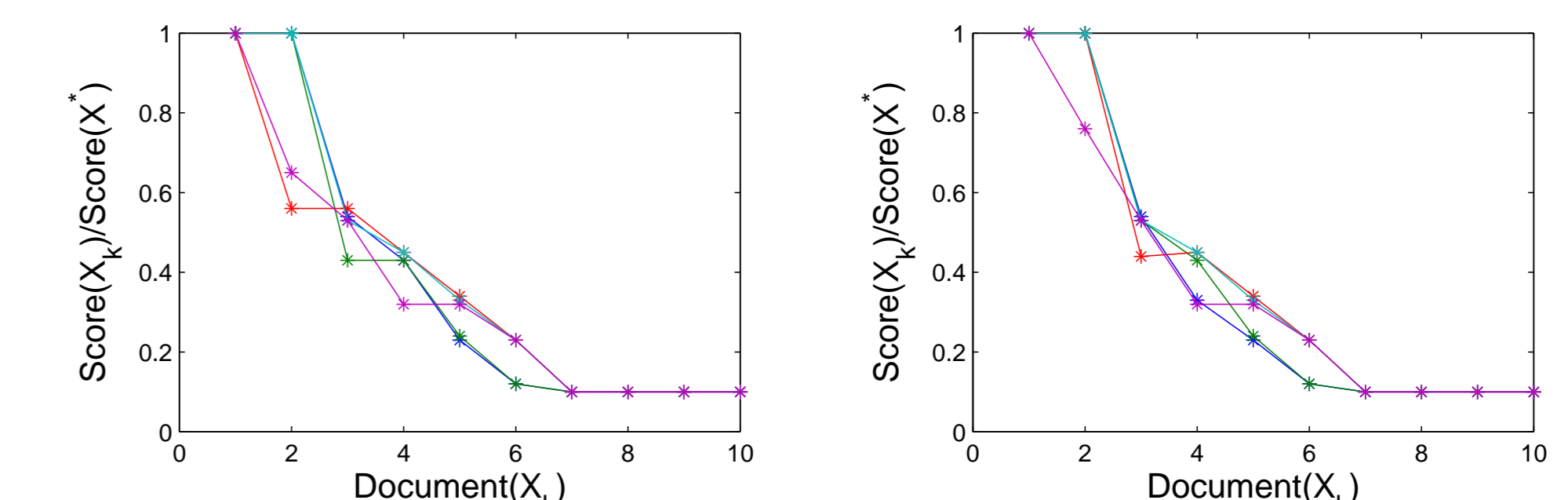


- A grid-like representation of the city.
- Each address maps to a grid-cell i.e. a geo-spatial document.
- A geo-spatial document is a combination of all addresses mapping to it.

**Modified problem:** Geo-coding of address queries is a problem in the search space of geo-spatial documents.

## Experiments

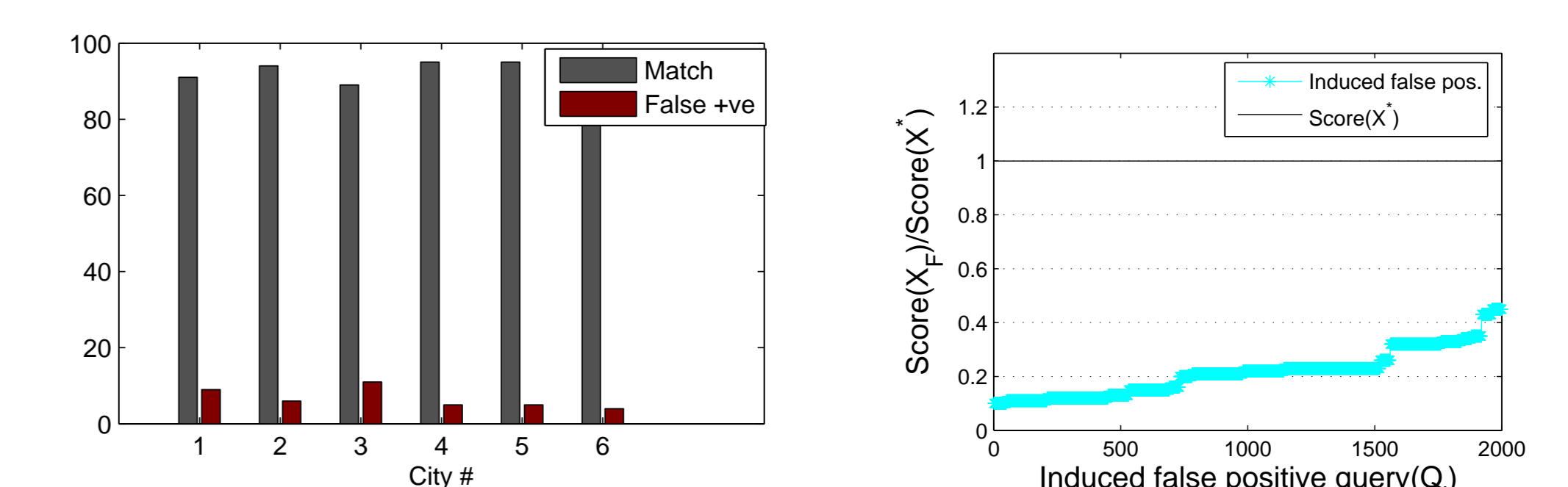
### False Negatives



- **In-corporus** and **Partial in-corporus** addresses as queries.
- Perfect resolution to mapping geo-spatial document, **No False Negatives**.

### Induced False Positives

- Barista, 3 Cs Cinema Road, Near Alankar theatre, Lajpat Nagar II New Delhi. [**Corpus**]
- Dominos Pizza, Feroze Gandhi Marg, near 3 Cs cinema, Lajpat Nagar II, New Delhi. [**Corpus**]
- Mc Donalds, 3 Cs Cinema Road, Lajpat Nagar II, New Delhi. [**Corpus**]
- Feroze Shopping center, Alankar Marg, Near Mc Donalds, Lajpat Nagar II, New Delhi. [**Query**]



**Conclusion:** Not plagued by FALSE NEGATIVES, low risk of FALSE POSITIVES.

## Summary

- A simple first principles approach to designing a geo-coding engine for emerging markets.
- Geo-coding of addresses modified to a problem in the document search space.
- Minimizes **FALSE NEGATIVES** and incurs few **FALSE POSITIVES**.