Spatial Approaches to Reducing Error in Geocoded Data

Presented to: Faculty of the Computer Science Department of the University of Southern California
04-01-2010

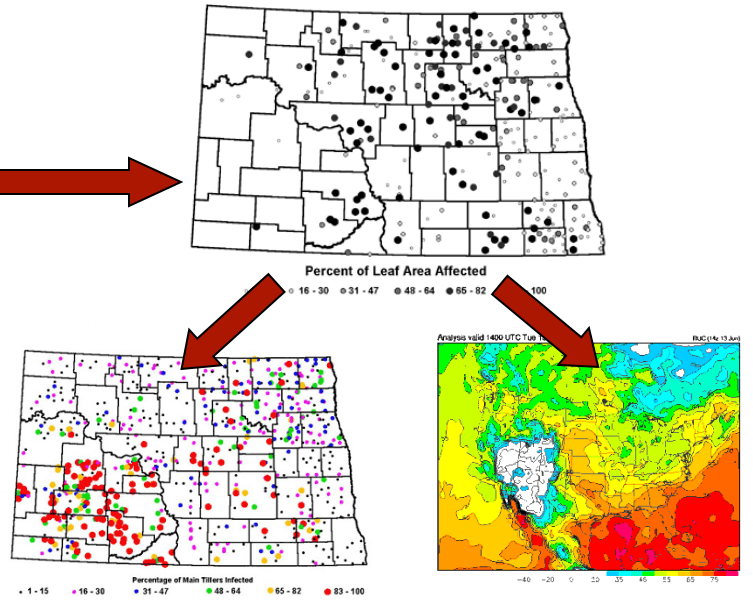
Dan Goldberg
GIS Research Laboratory
Department of Computer Science
University of Southern California
https://webgis.usc.edu

GIS research laboratory
Department of Geography
College of Letters, Arts, and Sciences

(Very) Brief Background

Locational descriptions → Geographic representations

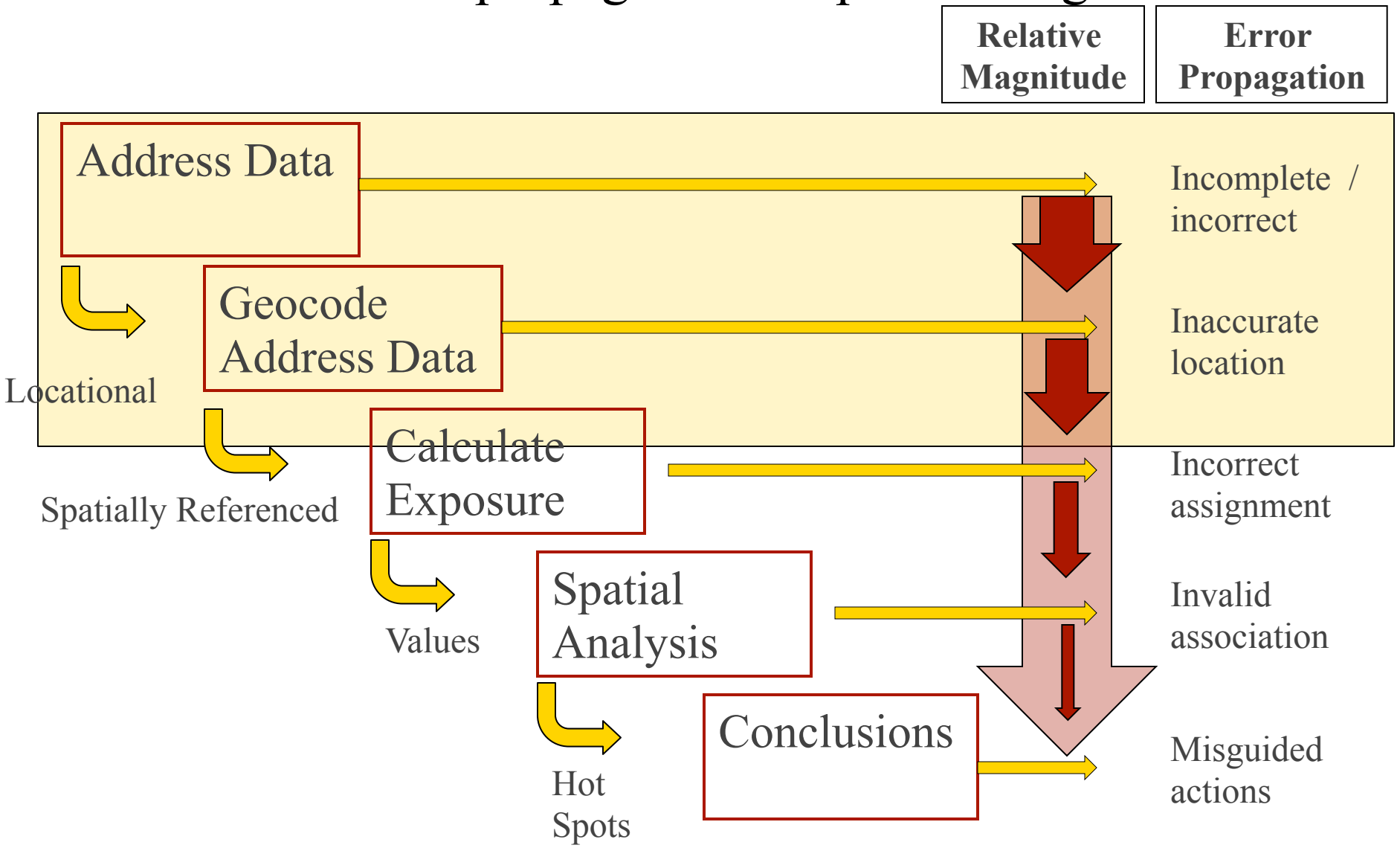
USC GIS Research Laboratory
3620 South Vermont Ave, Los Angeles, CA
Kaprielian Hall, Room 444
Los Angeles, CA 90089-0255



Spatio-Temporal Analyses

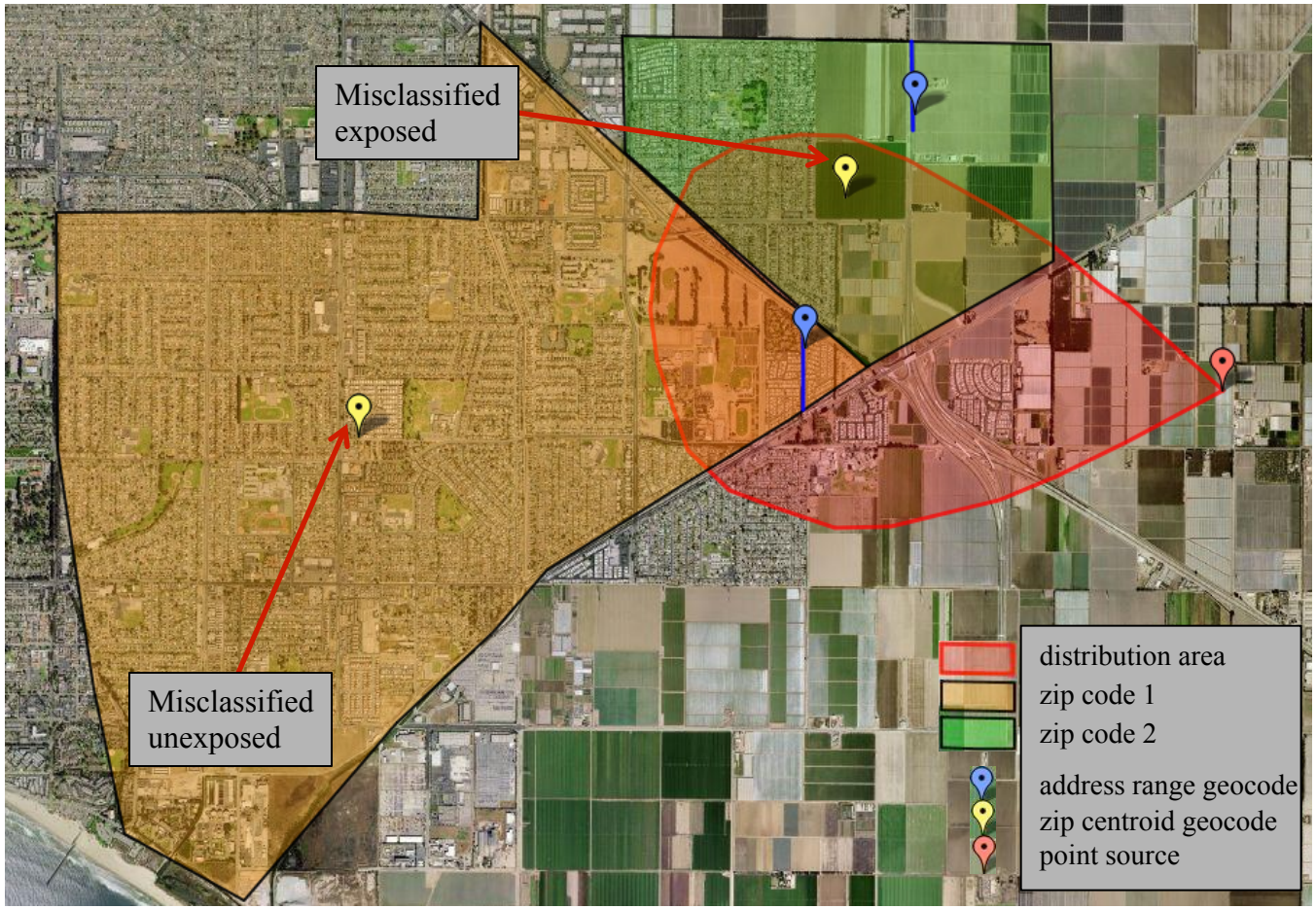
Motivations

- Error introduction/propagation in epidemiological research



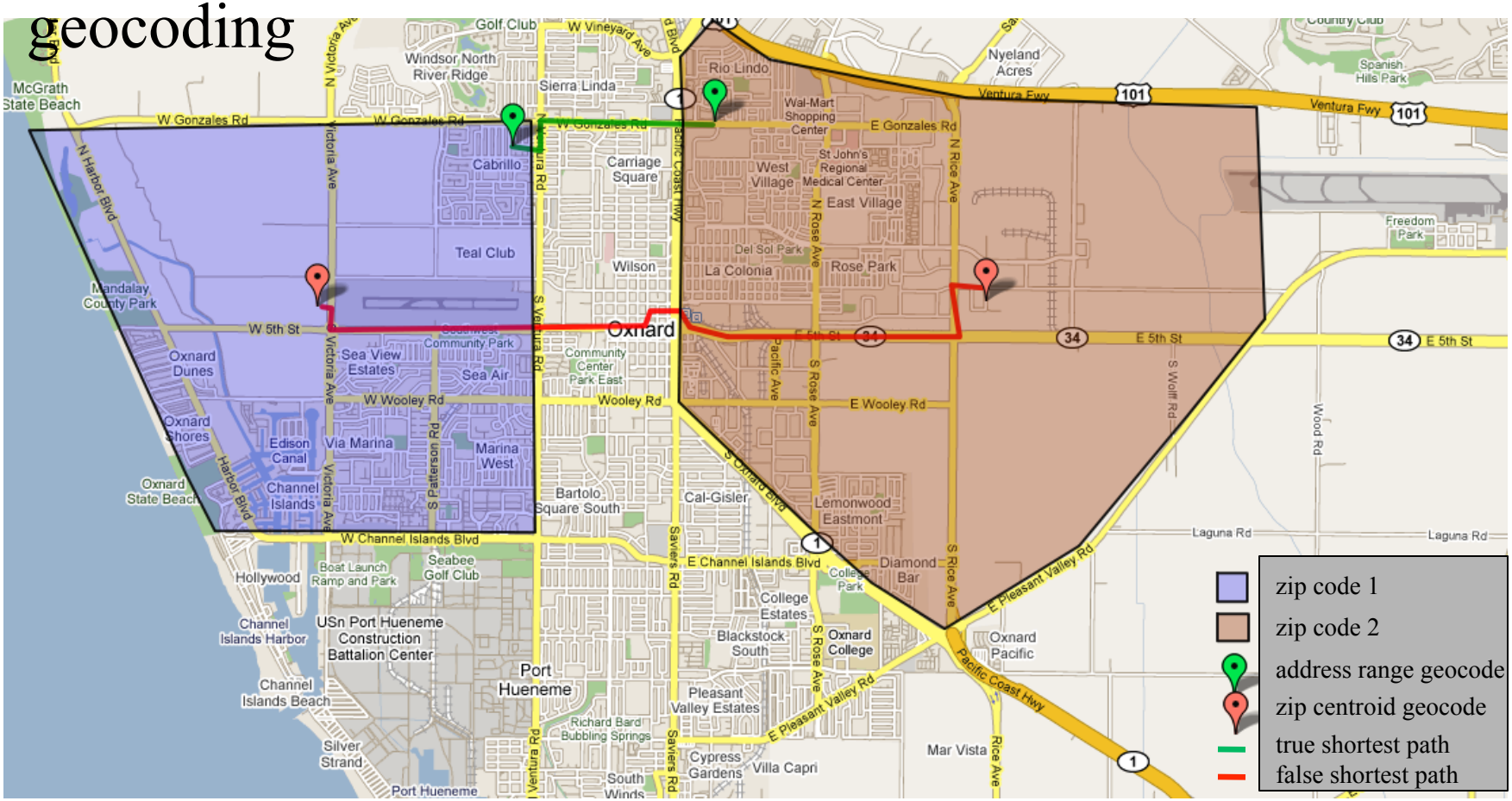
Motivations

- Exposure misclassification from inaccurate geocoding



Motivations

- Accessibility mischaracterization from inaccurate geocoding

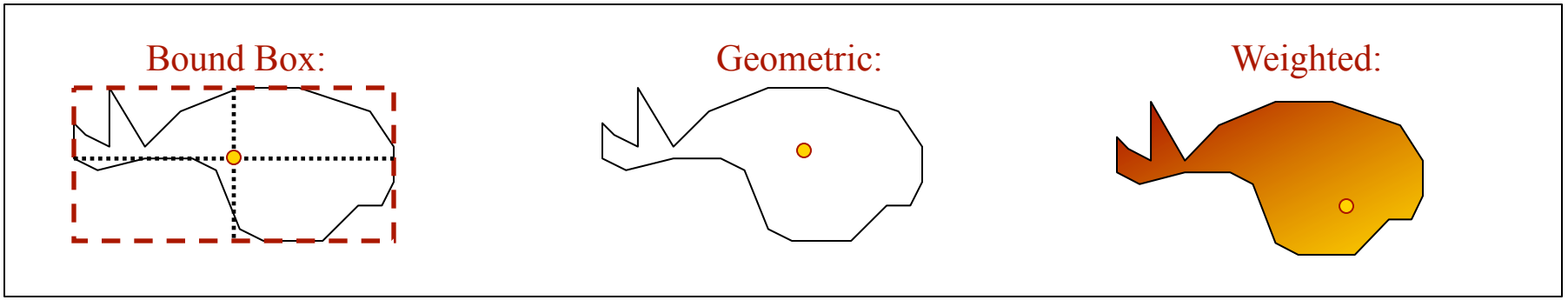


The error from geocoding can be larger than the distance traveled

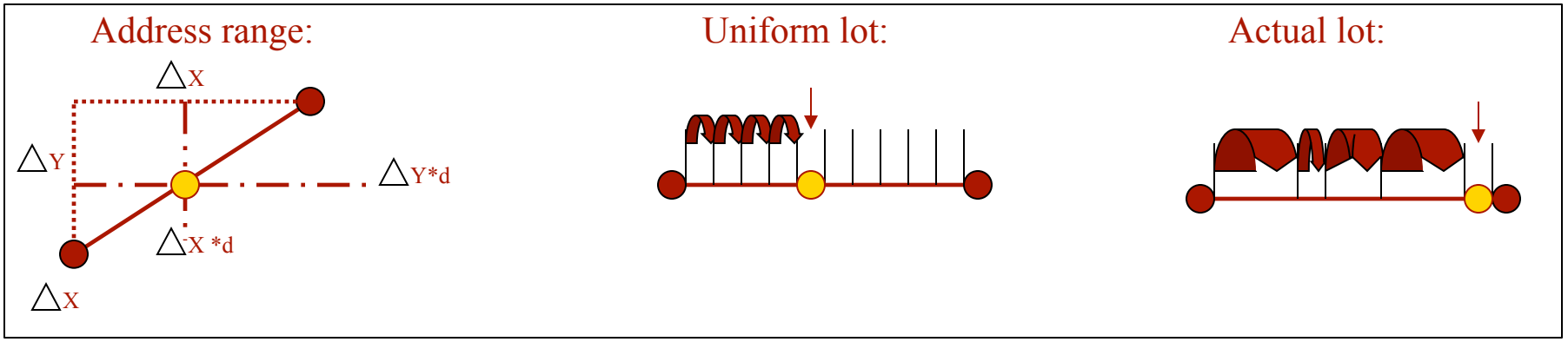
Motivations

- All geocodes with same "quality" do not have the same accuracy or certainty

NAACCR 2: Parcel Centroid



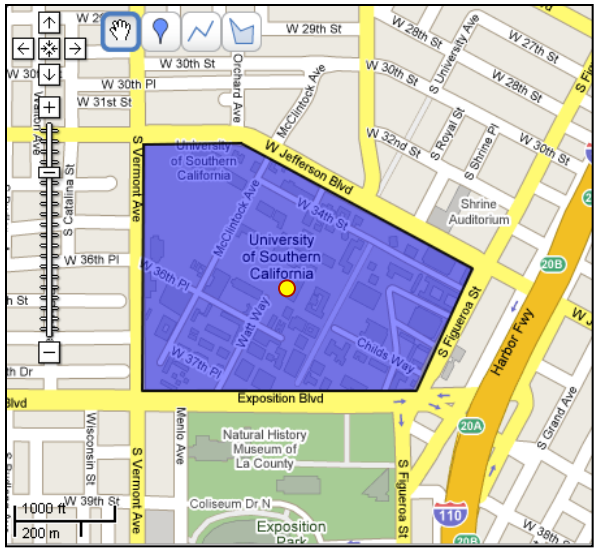
NAACCR 3: Street Address



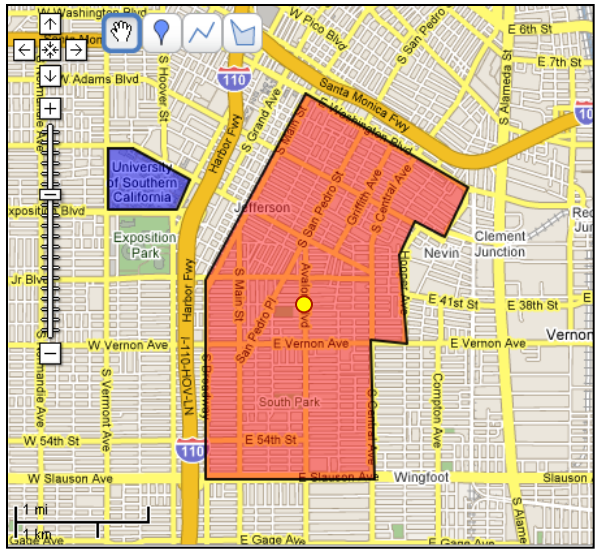
- Qualities of the feature interpolation matters

Motivations

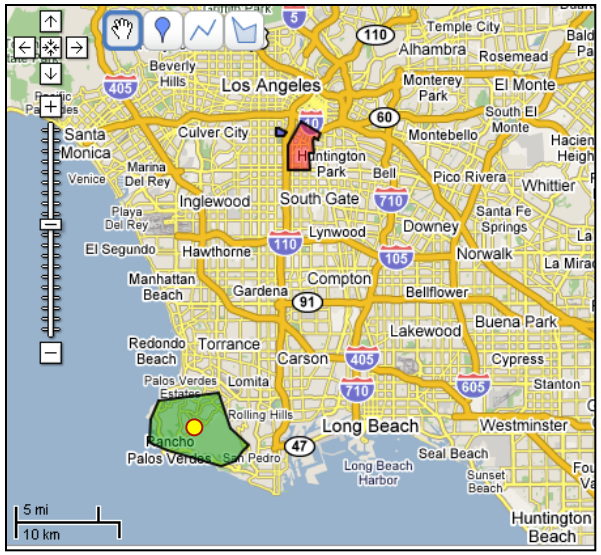
- All geocodes with same "quality" do not have the same accuracy or certainty



90089
~1:10,000 scale



90011
~1:60,000 scale



90275
~1:300,000 scale

- Qualities of the reference features matter

Motivations

– 3620 S. Vermont Ave, Los Angles CA 90089-0255

34.021906,-118.290385

Accuracy = ??

Match rate of geocoder used = ??

Spatial uncertainty of this geocode = ??

Reference data used to produce this geocode = ??

Interpolation assumptions used to produce this geocode = ??

Average spatial uncertainty for other geocodes in the area = ??

Theoretical and Technical Contributions

1. A **theoretical and practical framework** for developing, testing, and evaluating geocoding techniques.

2. A **derivation of the sources and scales** of potential spatial error and uncertainty.

3. A **spatially-varying neighborhood metric** to dynamically score nearby candidate reference features.

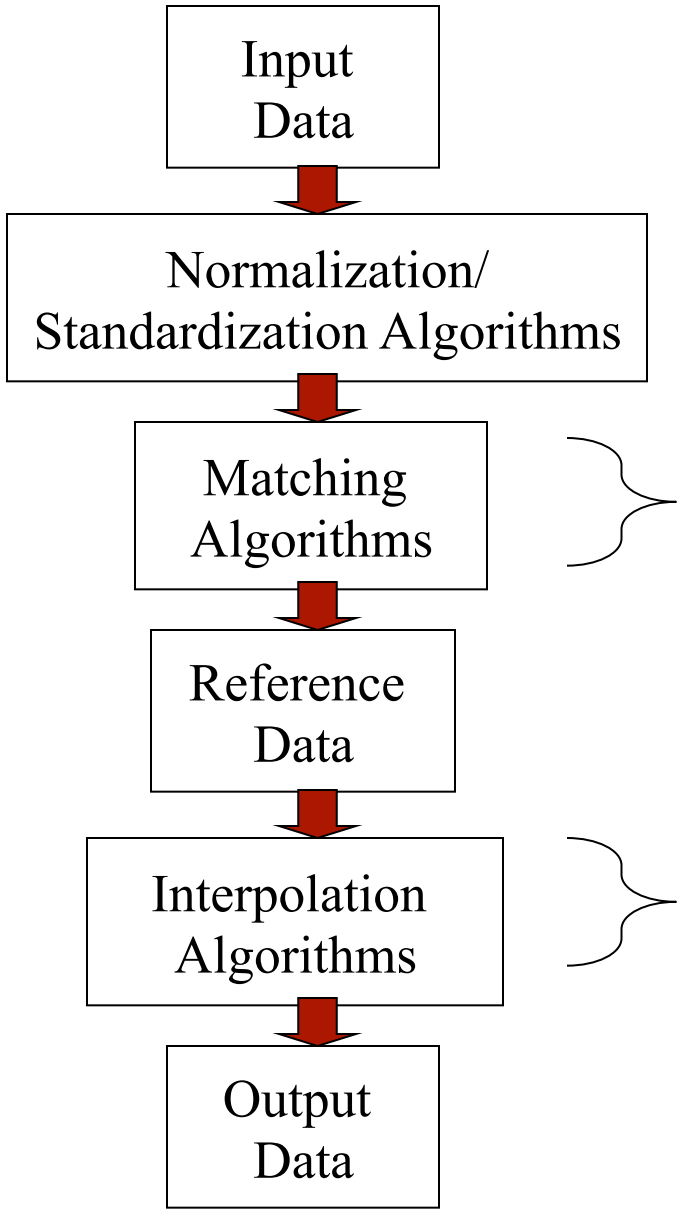
4. A **method to combine multiple layers of reference features** using uncertainty-, gravitationally-, and topologically based-approaches to derive the most likely candidate region.

5. A **rule- and neighborhood-based tie-breaking strategy** that deduces correct candidate selection using relationships between and regions surrounding ambiguous candidate reference features.

A Theoretical Framework for Geocoding Research

How can we model the geocoding process to facilitate an extensible system for describing and reducing spatial uncertainty and error?

Theoretical Framework



3620 South Vermont Avenue

Transform input to match reference data format

3620 S VERMONT AVE

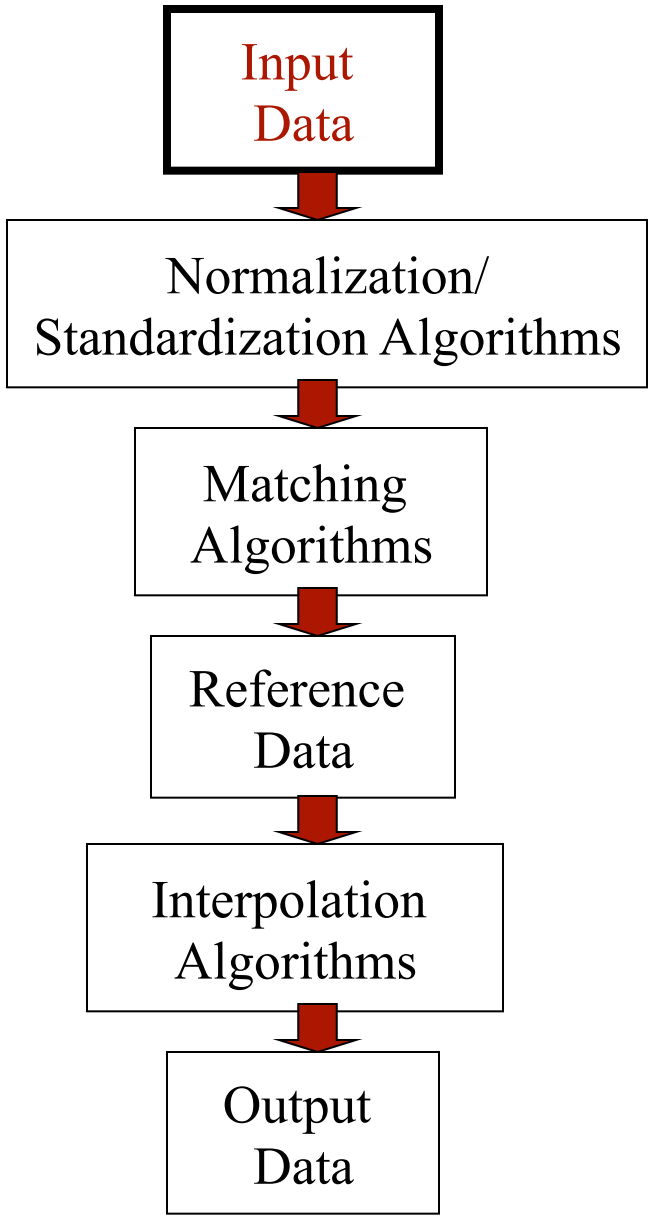
Find a matching geographic feature in reference data

SELECT FromX, FromY, ToX, ToY
FROM SOURCE WHERE
(Start >=3620 AND End <= 3620) AND
(Pre = S) AND
(Name = VERMONT) AND
(Suffix = AVE)

Use matched geographic feature to derive output

Output Point = (20% * △X, 20% * △Y)

Component: Input Data



Error Contribution

Many different types, forms, and formats:

Street Addresses: 3620 South Vermont Ave
Postal Codes: Los Angeles, CA 90089-0255
Named Places: USC Kaprielian Hall
Intersections: Vermont & 36th Place
Relative Descriptions: b/w Bakersfield & Shafter

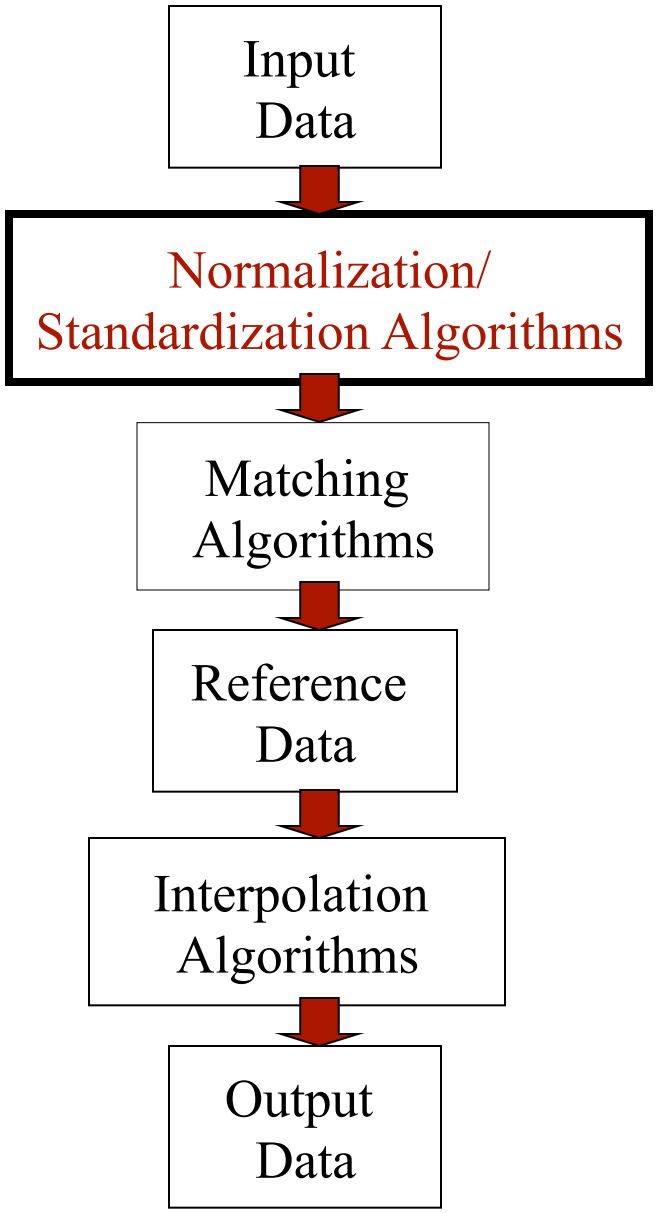
Different levels of information/certainty:

Street Addresses: Somewhere on street
Postal Codes: Somewhere on postal route
Named Places: Absolute location
Intersections: Somewhere near intersection
Relative Descriptions: Somewhere near locations

Incompleteness:
3260 S Vermont ___
3620 _ Vermont Ave
____ _ Vermont Ave

Inaccuracy:
3620 S Verment Ave
362_ S Vermont ___
3260 _ Vermont St

Component: Input Data Cleaning



Error Contribution

- Parsing – Separating components of the address

  Token-Based: relies on formatting

-Normalization – Identifying components of the address

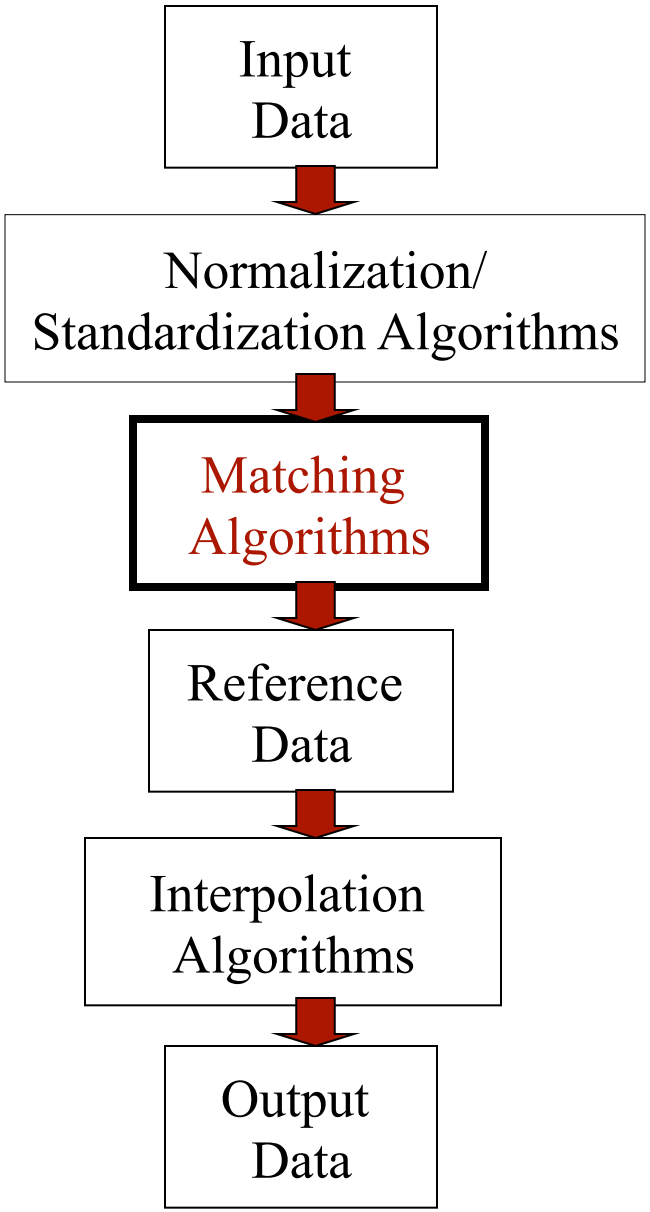
  Substitution-Based: relies on the token ordering
  Context-Based: relies on position and schema knowledge
  Probability-Based: relies on likelihood of occurrence

- Standardization – Formatting components of the address

  Schema mapping: must exist for all reference sources

3620	South	Vermont	Ave	Los Angeles	,	90089
Street Address				City		Zip
90089		St Los Angeles	St	Los Angeles	,	90089
Street Address				City		Zip
23	E	South	St	South Los Angeles	,	90089
Street Address				City		Zip

Component: Matching Algorithms



Error Contribution

-Multiple Match Types – Feature selected from reference set

Exact: A single perfect match
Non-exact: A single non-perfect match
Exact ambiguous: Multiple perfect matches
Non-exact ambiguous: Multiple non-perfect matches
None: No matches

-Multiple Matching Methods – Ways of selecting features

Deterministic: Rule-based, iterative
Probabilistic: Likelihood-based, attribute weighting

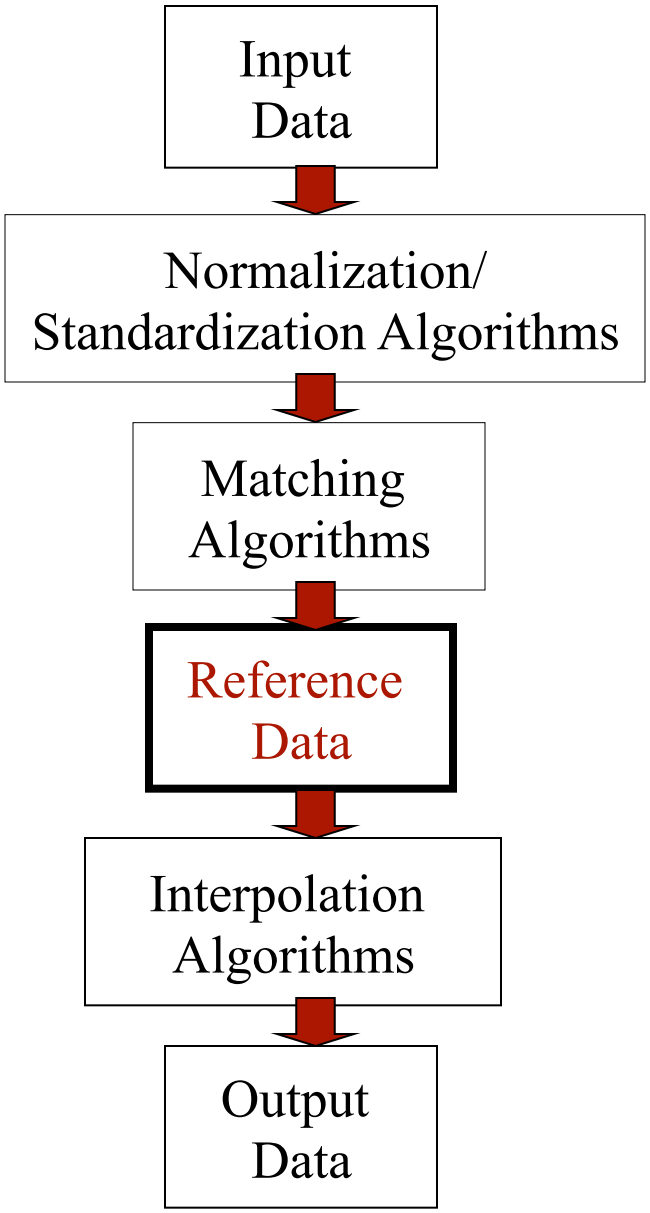
-Multiple Fuzzifying Techniques – Alter input data

Word Stemming: Porter Stemmer
Phonetic Algorithms: Soundex
Attribute Relaxation: Remove attributes and retry match

-Multiple Scoring Methods – compute a candidate score

Relative attribute weighting
Match-Unmatch weighting

Component: Reference Data



Error Contribution

-Multiple Data Types

Point-based: ZCTA and Place Centroids
Linear-Based: Street Centerlines
Areal Unit-Based: Parcels, ZCTA and Place Boundaries

-Wide spectrum of accuracies/completeness

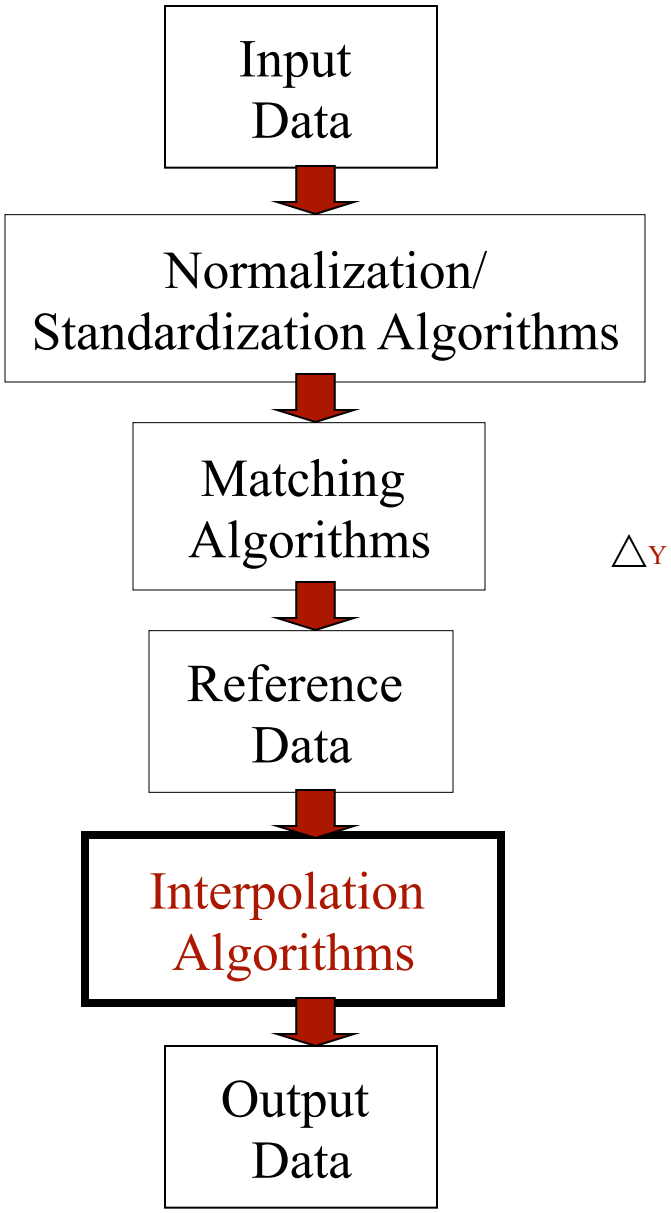
Commercial vs. Public
- Attribute accuracy – spatial and non-spatial
- Attribute completeness – spatial and non-spatial
- Feature complexity – simple vs. polylines

Local Scale vs. National Scale
- Census Place Boundaries vs. Local Neighborhoods

-Wide spectrum of cost/availability

Free vs. Costly: TIGER/Lines vs. TeleAtlas
Available vs. Not: Address points – CA. vs. N. Carolina

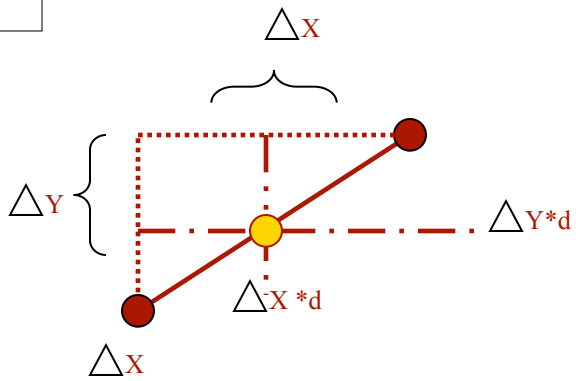
Component: Interpolation Algorithms

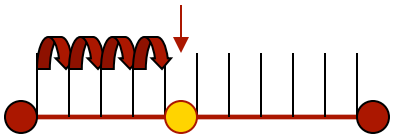


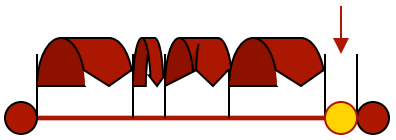
Error Contribution

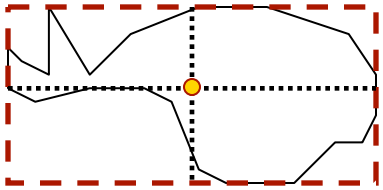
- Many methods of interpolation

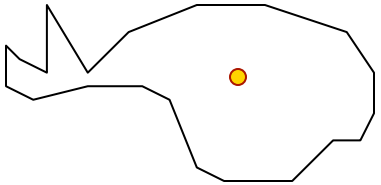
Depend on reference feature type
Depend on info available (assumptions)

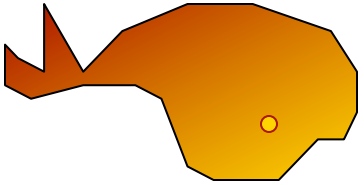




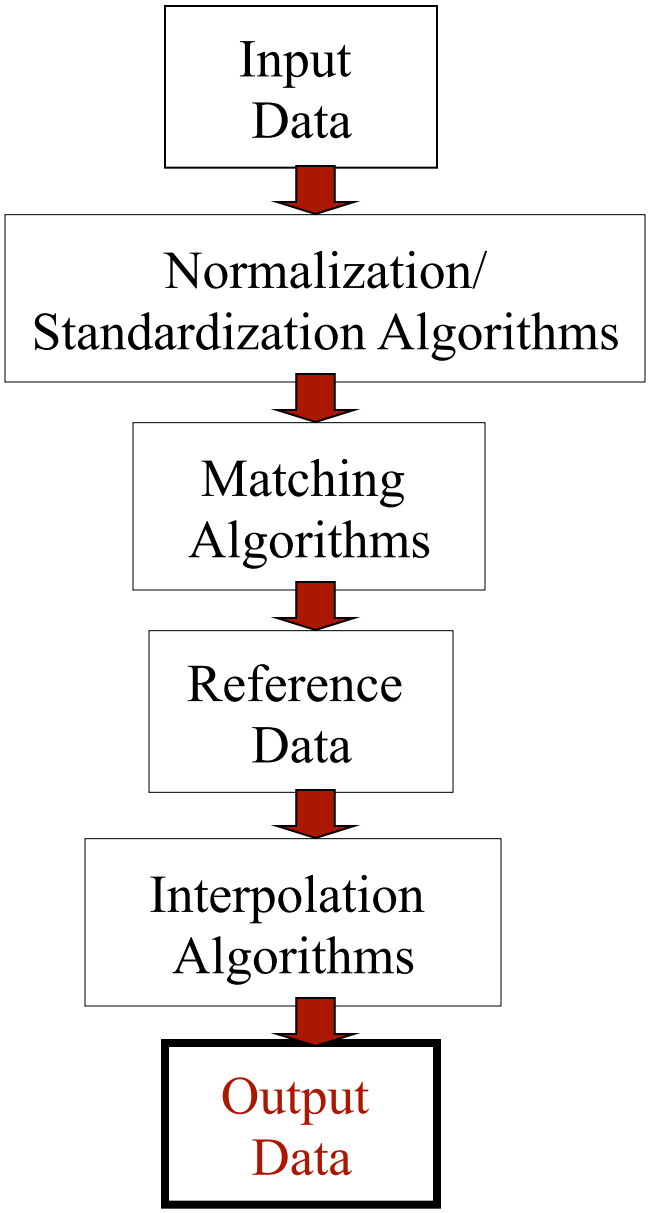








Component: Interpolation Algorithms



Error Contribution

- Lack of Process Transparency

  -Nothing reported about the decisions made or alternatives

- Output Data Type: Only Geographic Coordinates

  -Lose data required for determining true accuracy

- Output Accuracy: Feature Match Type + Probability

  -Nothing that indicates direction

  -Nothing that indicates distance

  -Nothing that indicates certainty area or surface

A Spatially-Varying Block-Distance Candidate Scoring Approach

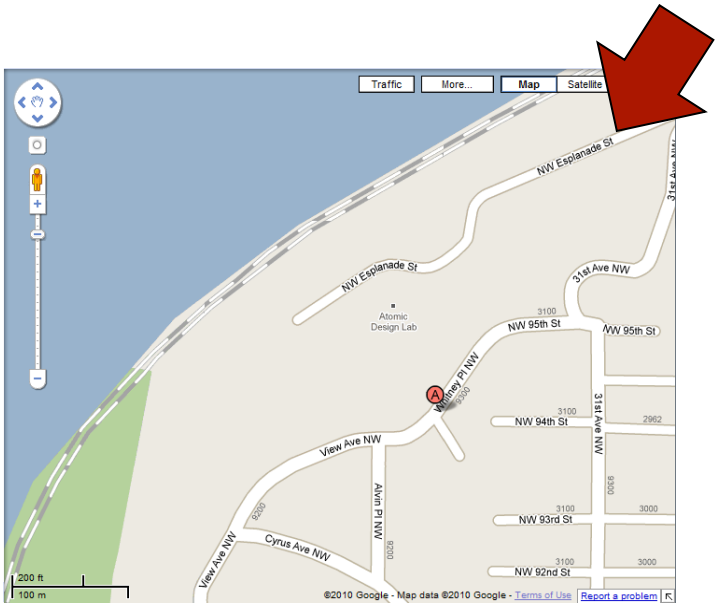
Can nearby candidate reference features be used to overcome inaccuracies and incompleteness in reference data sources?

Spatially-Varying Block-Distance Feature Scoring - Motivation

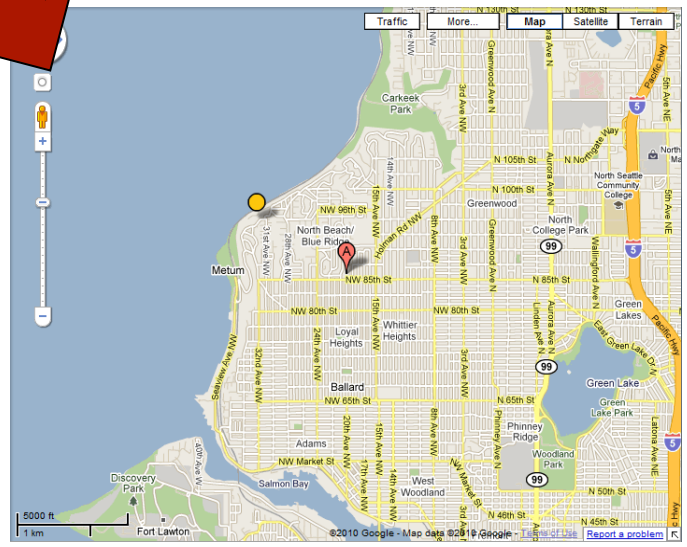
Problems:

1) Address ranges in reference data files are often inaccurate
2) Leads to false negative non-matches
3) Results in reversion to lower level geographic matches

9800 View Ave, Seattle WA 98117



Address range doesn't exist

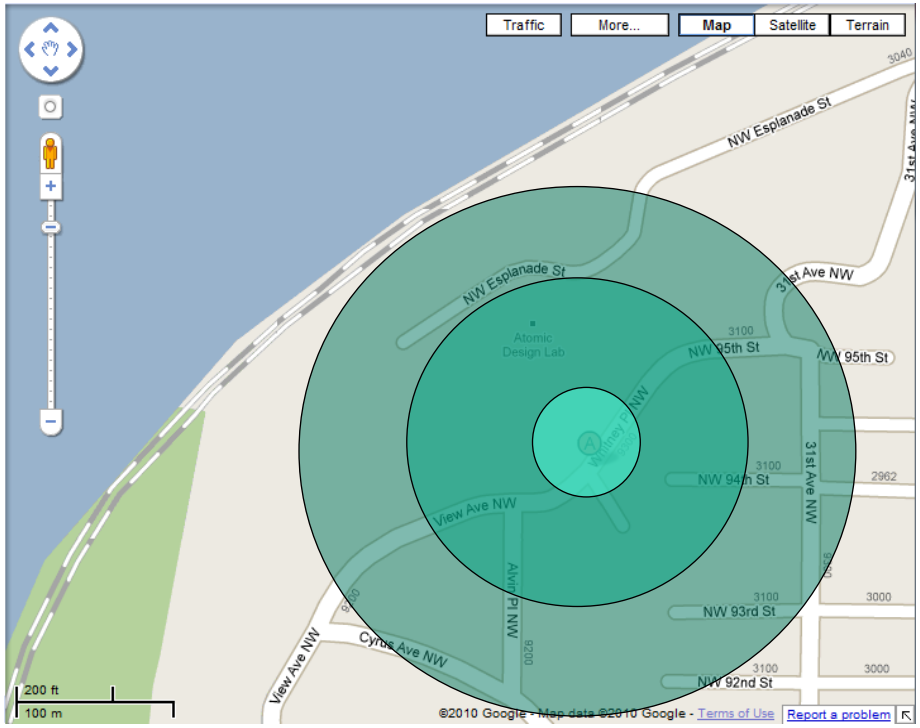


Reverts to ZIP 98117

Spatially-Varying Block-Distance Feature Scoring - Intuition

A better approach:

1) Proportionally weight the closest reference features by their distance away in number of blocks
2) Choose the reference feature with the highest score within the search radius threshold (max number of blocks away)



Intuitions:

1) If we exclude the address number from the matching algorithm, we will have a large candidate set of all streets in the region with the correct name and regional attributes (ZIP, city) differing only by their address ranges

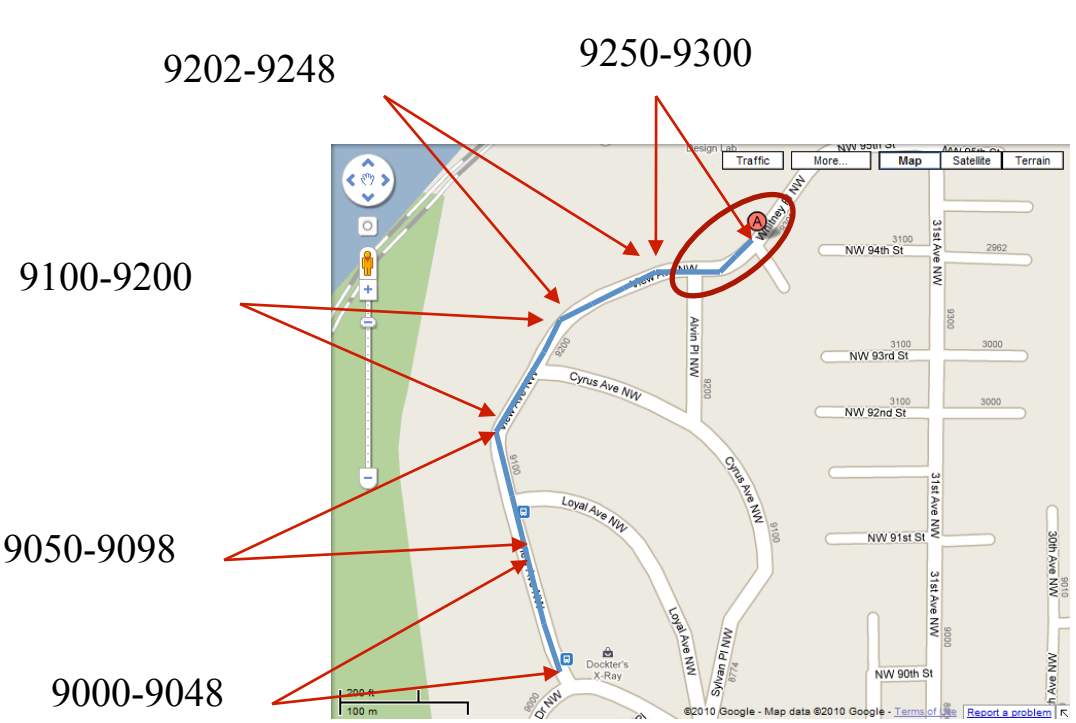
2) We can score them based on how many blocks they are from the input address

   9300-9400 Block of View Ave is ~ 4 blocks away from 9800 View Ave

Spatially-Varying Block-Distance Feature Scoring - Implementation

Implementation:

1) User provides max search radius: B_M
2) Use a loose query to obtain all streets with correct name in correct City/Zip
3) Determine average block size for the region: $|b|$
4) Calculate the distance from the input address to the closest end of the available reference features: d
5) Calculate the block distance between reference feature and input address: B_d
6) Calculate the proportional weight $\dot{w}_i$ based on the block distance and maximum search radius B_M



B_M=10

Input: 9800 View Ave

Item	Range	Size	d	B_d	$\dot{w}_i$
0	9000-9048	48	752	13.8	1.38
1	9050-9098	48	702	12.9	1.29
2	9100-9200	100	600	11.2	1.12
3	9202-9248	48	552	10.4	1.04
4	9250-9300	50	500	9.5	0.95

$|b| = 58.8$

$$d = Min\left(Abs(a_r - N_s), Abs(a_r - N_e)\right)$$

$$B_d = 1 + \frac{d}{|b|}$$

$$\dot{w}_i = \frac{B_d}{B_M}$$

Spatially-Varying Block-Distance Feature Scoring - Evaluation

(1) Determine if nearby match scoring can overcome address range attribute problems of reference data and improve match rates in address range geocoders

(2) Ensure that the output of such an approach is consistent with higher accuracy geocoders that do not suffer from such reference source errors

Sample Data:

Medicare National Provider Identification Number (NPI)

22,984 records in LA County after removing duplicates

Test Geocoders:

(1) USC Geocoder with StreetMap North America

(2) ESRI ArcGIS Address Locator with StreetMap North America
- Comparable because it uses the same reference data set
- Shows how prevalent the problem is

(3) Google, Microsoft Bing, and Yahoo! With buildings, parcel centroids, streets
- When one or more geocoders with access to reference data agree with output, it indicates a strong likelihood of being correct

Spatially-Varying Block-Distance Feature Scoring - Results

(1) High level of agreement between USC (243) and ESRI (241) geocoders for records that fail to match
- Shows that the USC geocoder performs on-par with existing state-of-the-art
- Shows that nearby matching is needed because these 243 records would have reverted to ZIP

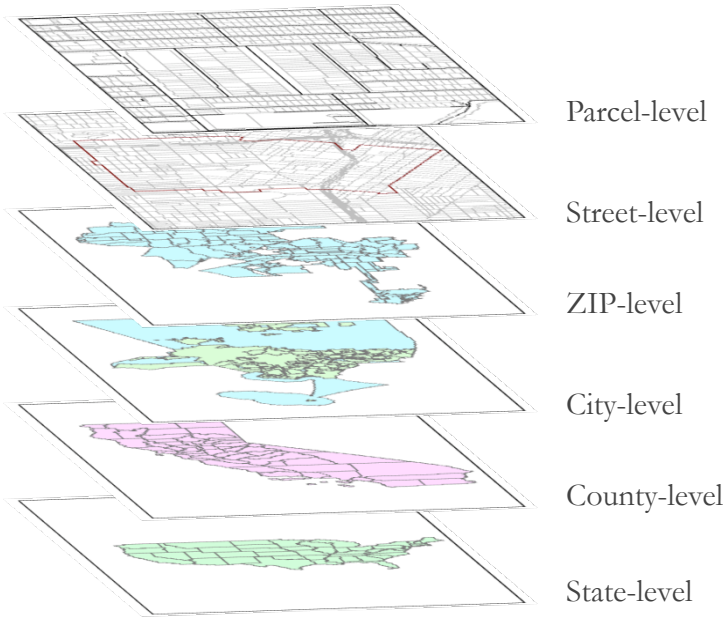
(2) Average distance between the nearby output and closest online geocoder output is 135 m
- Shows that the nearby placement of the USC geocoder puts the output close to the correct location

A Best-Match Candidate Selection Approach

When multiple candidate geocodes are available from several reference layers, what is the best strategy to pick the most accurate one?

Best-Match Candidate Selection Criteria - Motivation

Many different reference layers available



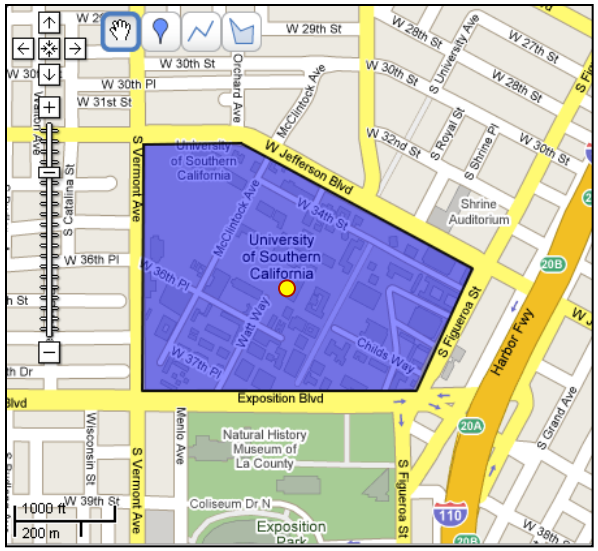
NAACCR GIS Coordinate Quality Codes

Code	Description
1	GPS
2	Parcel centroid
3	Complete street address
4	Street intersection
5	Mid-point on street segment
6	USPS ZIP5+4 centroid
7	USPS ZIP5+2 centroid
8	Assigned manually
9	USPS ZIP5 centroid
10	USPS ZIP5 centroid of PO Box or RR
11	City centroid
12	County centroid

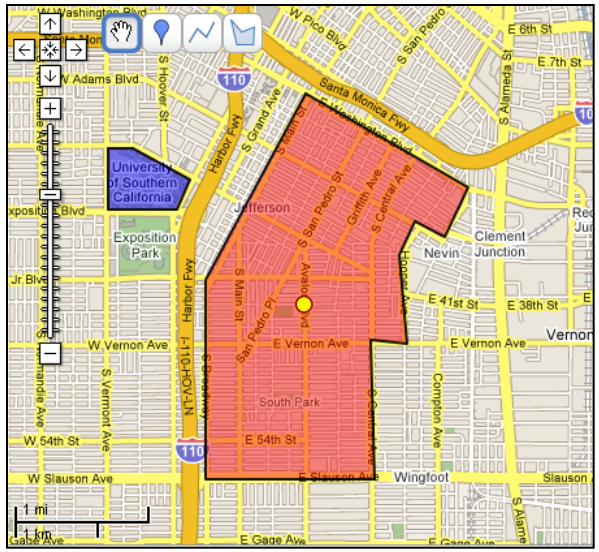
Hierarchy-based best match criterion

Best-Match Candidate Selection Criteria - Motivation

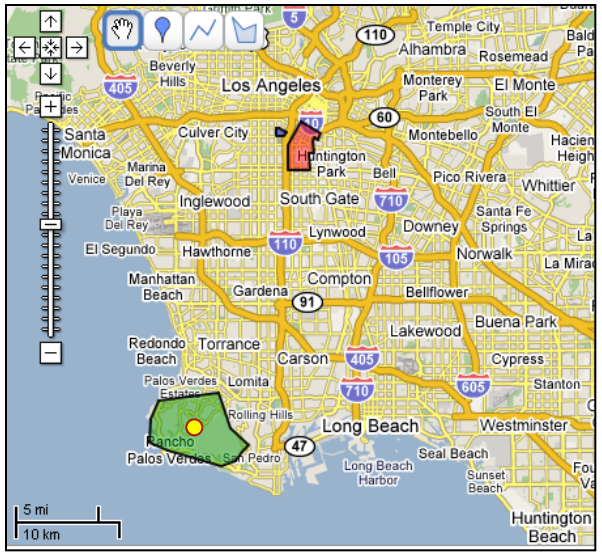
- All reference features of the same class do not have the same accuracy or certainty



90089
~1:10,000 scale



90011
~1:60,000 scale



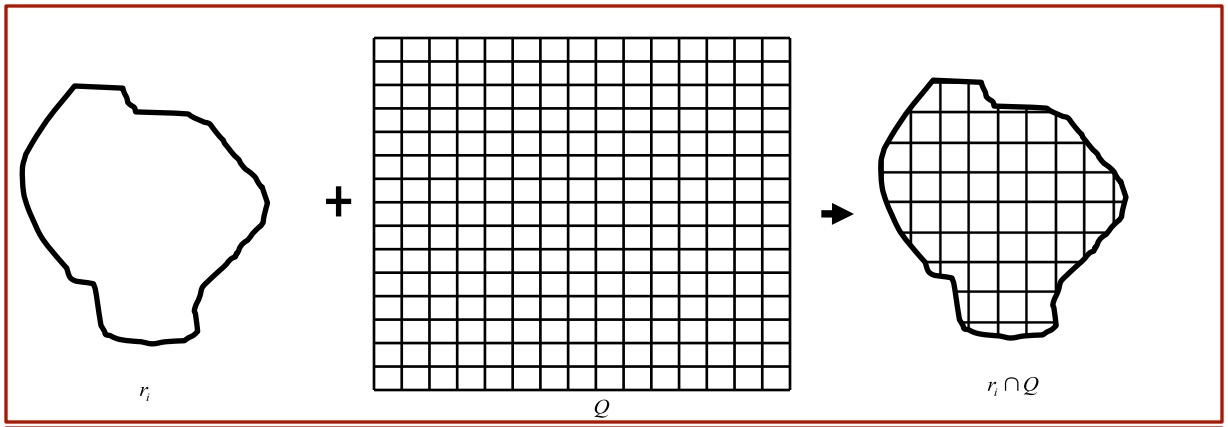
90275
~1:300,000 scale

- A hierarchy-based approach is not a quantitative method to choose the most optimal output

Best-Match Candidate Selection Criteria – Uncertainty-Method

(1) Uncertainty-based criterion

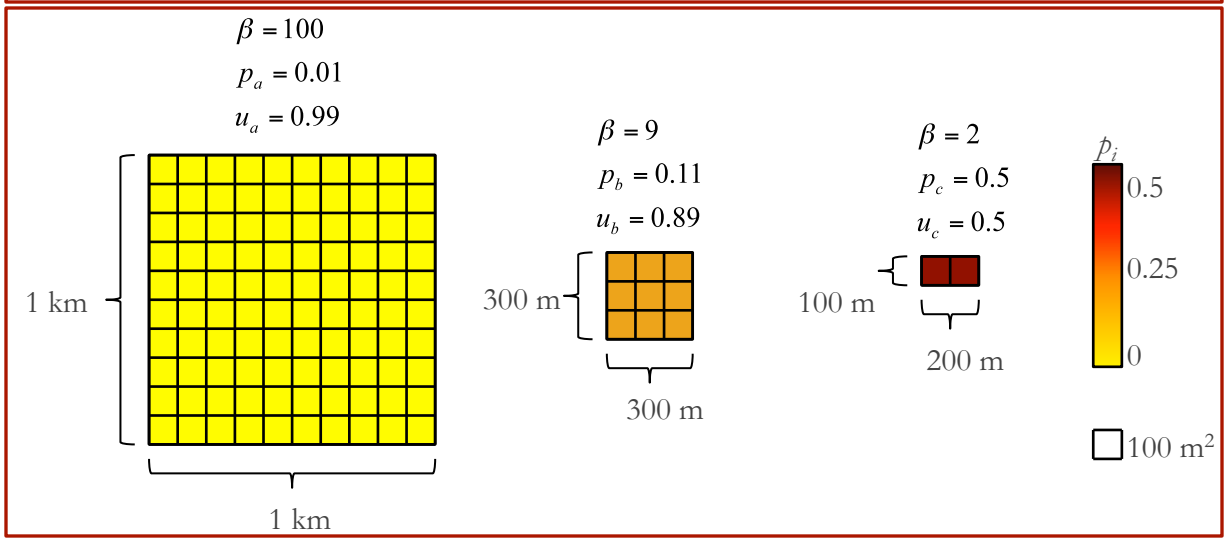
- Use the area of the feature as a proxy for uncertainty, pick the candidate with minimum



$$\beta = \left(\frac{Area(q)}{Area(r_i)}\right)^2$$

$$p_i = p(g_i) = \frac{1}{\beta}$$

$$u_i = 1 - p_i$$



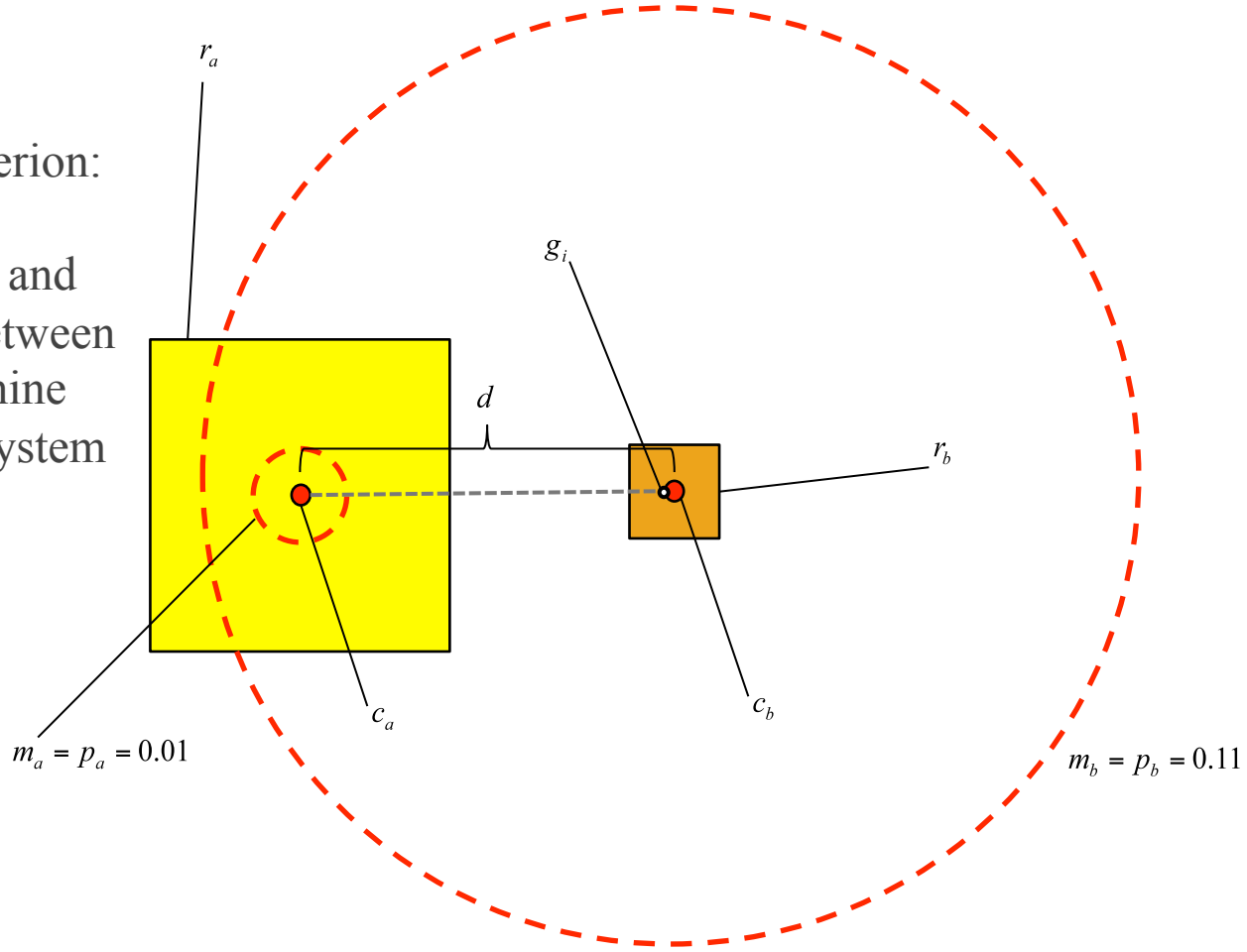
Best-Match Candidate Selection Criteria – Gravitational-Method

(2) Gravitationally-based criterion:

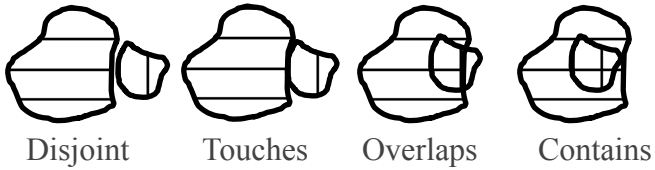
Use the area of the feature and the spatial relationships between all other features to determine the center of mass of the system

$$M_y = \sum_{i=1}^{n} m_i x_i, M_x = \sum_{i=1}^{n} m_i y_i$$

$$\bar{x} = \frac{M_y}{m}, \bar{y} = \frac{M_x}{m}$$

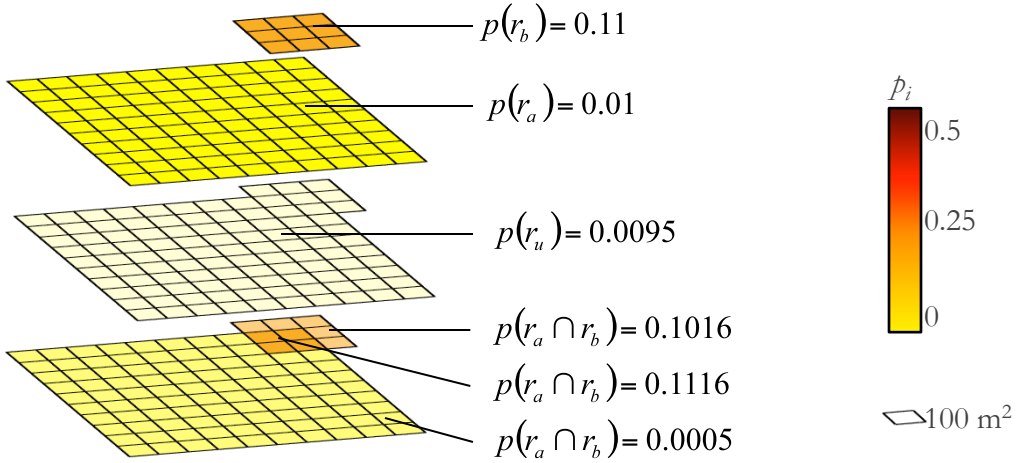


Best-Match Candidate Selection Criteria– Topological-Method



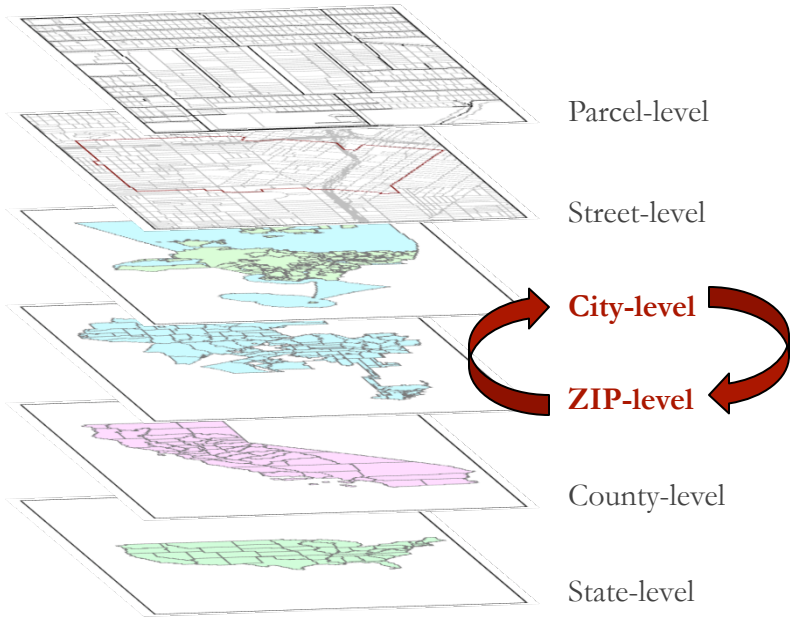
(2) Topologically-based criterion:

Use the area of the feature and the topological relationships between all other features to distribute the uncertainty across the whole system



Best-Match Candidate Selection Criteria – Hierarchy-Reversed

Many different reference layers available



NAACCR GIS Coordinate Quality Codes

Code	Description
1	GPS
2	Parcel centroid
3	Complete street address
4	Street intersection
5	Mid-point on street segment
6	**City centroid**
7	USPS ZIP5+4 centroid
8	USPS ZIP5+2 centroid
9	Assigned manually
10	USPS ZIP5 centroid
11	USPS ZIP5 centroid of PO Box or RR
12	County centroid

Hierarchy-based best match criterion

Best-Match Candidate Selection Criteria – Evaluation

(1) Does the spatial accuracy of geocodes improve if we simply reverse the order of layers in a hierarchy-based approach?
(2) Does accuracy improve when utilizing an uncertainty-based based approach instead of any type of hierarchy-based approach?
(3) What level of spatial improvement is possible when using either the gravitationally- or topologically-based approach over the uncertainty-based approach?

Sample Data:

3,329 GPS locations of Best Western Hotels (2,093) and Target Stores (1,649)

Test Geocoders:

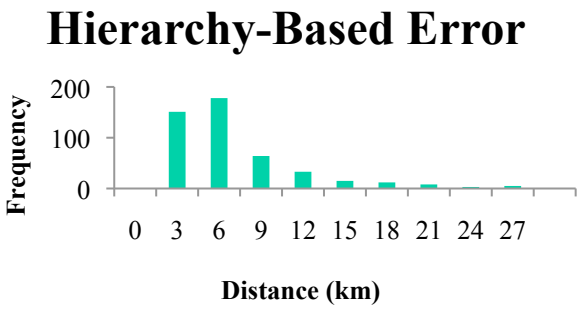
USC Geocoder with:

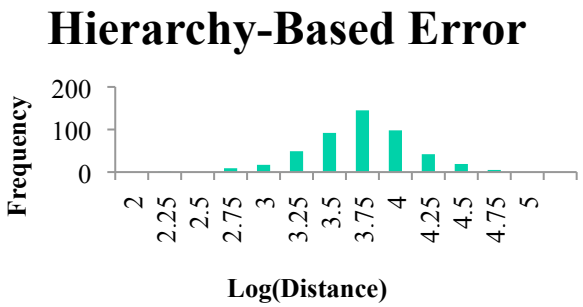
(a) Census TIGER/Line, ZCTA, and Place Files
- Represents a geocoder with a high proportion of street-level matches

(b) Census ZCTA, and Place Files
- Represents a geocoder with an extremely low proportion of street-level matches

Best-Match Candidate Selection Criteria – Results

Sample Dataset	Reference layers	n % of total	Hierarchy mean (km)	Hierarchy Reversed mean (km)	Uncertainty mean (km)	Gravitational mean (km)	Topological mean (km)
Best Westerns	(a)	255 12.2%	7.445	3.737	2.888	2.864	2.842
Target Stores	(a)	222 13.5%	4.422	5.274	3.685	3.367	3.319
Combined	(a)	477 12.8%	6.038	4.452	3.259	3.098	3.064
Combined	(b)	3329 89%	4.927	4.418	2.845	2.672	2.626





(1) Reversing a feature hierarchy improves results in some areas (rural) but not in others (urban)

(2) Using any of the alternative methods improves spatial accuracy over a hierarchy-based approach
- Students t-test shows that these spatial improvements are statistically significant (α=.05, $p < .001$)

(3) Gravitationally- and topologically-based improvements are only statistically significantly when the reference features topologically overlap

Intelligent Tie-Breaking Approaches

When two or more candidates are each equi-probable, what techniques can be used to reason about which is more likely to be correct?

Intelligent Tie-Breaking Approaches – Motivation

Problem:

1) Address data often results in ties because of input/reference data errors/incompleteness

Id	Source	Address	Ambiguous matches	Ambiguity reason	Type of ambiguity
1	(a)	701 N Main Street Colfax Wa 99111-2120	631-699 block of N Main Street 703-707 block of N Main Street	Between known address ranges	Reference data incompleteness
2	(b)	626 W Route 66 Glendora Ca 91740	616-698 block of Route 66 (W) 624-630 block of Route 66 (E)	E/W missing from reference features	Reference data incompleteness
3	(b)	8354 Natalie Lane West Hills Ca 91304	8338-8398 block of Natalie Lane 8336-8498 block of Natalie Lane	Overlapping address ranges	Reference data incorrectness
4	(b)	439 S 97th Street Los Angeles Ca 90003	439 E 97th Street 439 W 97th Street	Incorrect pre-directional	Input data incorrectness
5	(b)	222 Market Street Inglewood Ca 90301	222 N Market Street 222 S Market Street	Missing pre-directional	Input data incompleteness

- Most systems revert up the hierarchy (ZIP, City, State)
- Some systems require user intervention to correct the tie (manual process)
- Others choose one of the ambiguous matches at random (automatic flip-a-coin)

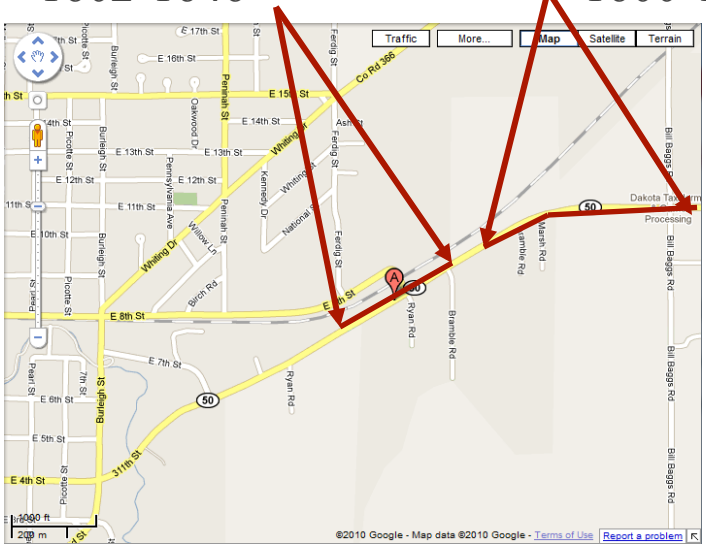
Intelligent Tie-Breaking Approaches – Motivation

Address Range Ambiguities:

One reference feature contains another

1602-1646

1600-1800



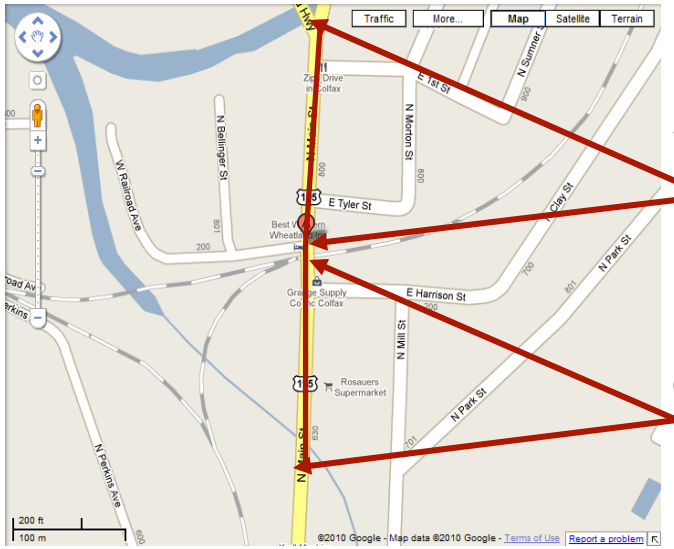
1607 E Highway 50 Yankton, SD

Should choose the smaller more specific one because it is more likely to be an updated/corrected version

701 N Main St, Colfax WA 99111

703-707

631-699



Input address missing from reference

Should choose the one that is on the correct block range (700 block)

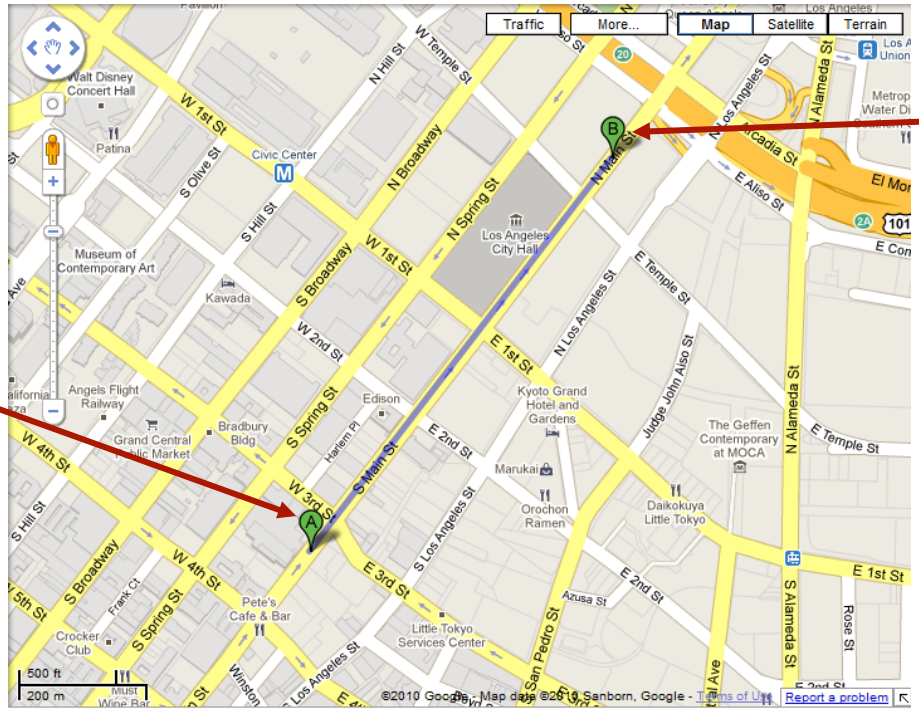
Intelligent Tie-Breaking Approaches – Motivation

Directional Ambiguities:

300 E Main St, Los Angeles, CA 90013 – Directional is incorrect

300 N Main St

300 S Main St

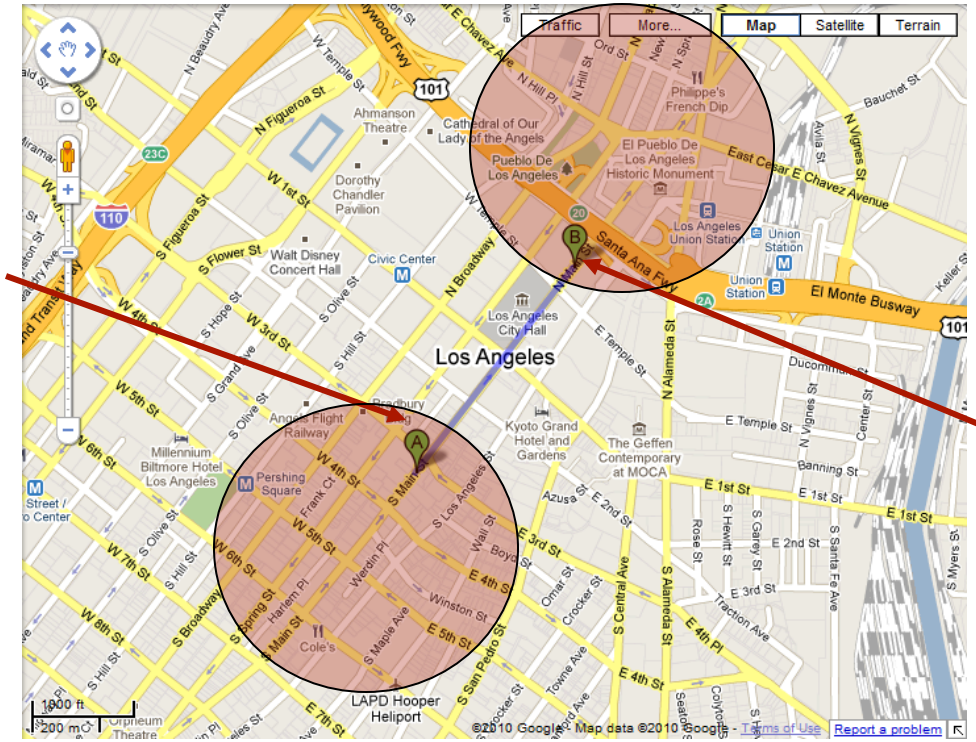


Intelligent Tie-Breaking Approaches – Geo-Intelligence

1) Use information drawn from the other street segments in the region around each candidate to determine if one is more likely than the other based on the attributes present

300 E Main St, Los Angeles, CA 90013 – Directional is incorrect

300 S Main St

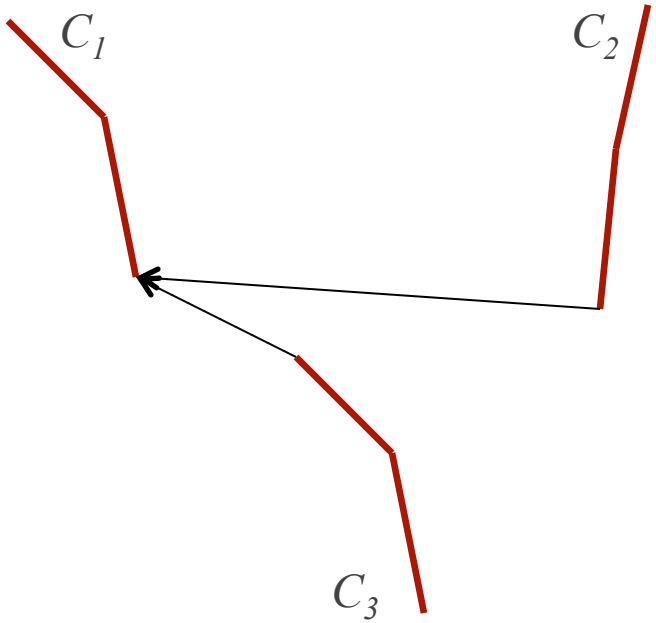


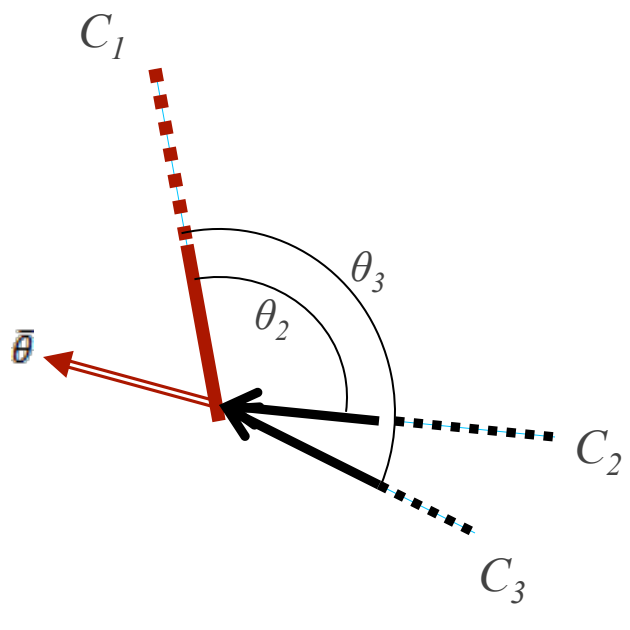
300 N Main St

Intelligent Tie-Breaking Approaches – Geo-Intelligence - Considerations

1) What direction should the regions expand in?
   (a) We want to the regions to be expanding away from each other in opposite directions

   Solution:
   (a) Find the point closest to all other candidates
   (b) Connect the closest point on all other candidates to this point
   (c) Use the average of incoming angles from other candidates



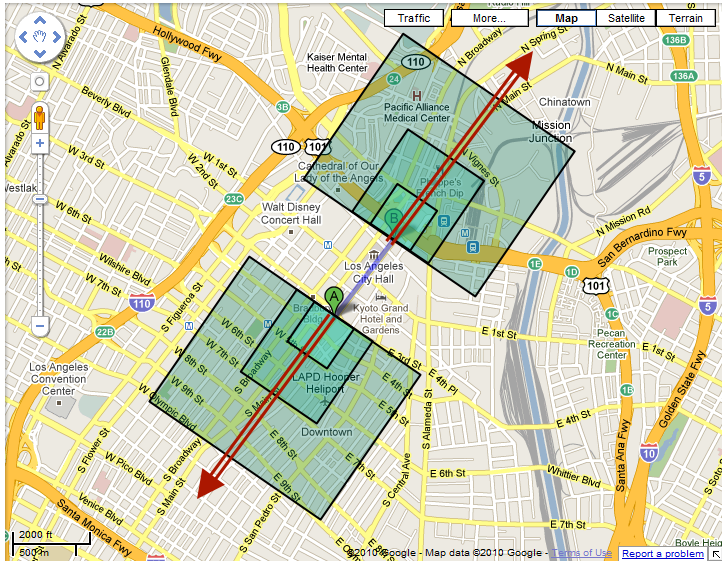


Intelligent Tie-Breaking Approaches – Geo-Intelligence - Considerations

1) How big should the region around a candidate be?
   (a) We want to keep the regions as small as possible to facilitate rapid processing
   (b) We want the regions to be large enough to include sufficient information useful for discriminating between two separate areas

Solution: Iteratively grow the region until no further useful information is being added

- Query all street segments in region defined by bounding box
- Keep a vector of attributes occurrence counts
- Use Shannon's information entropy metric (diversity index) to determine when we have outgrown the immediate region around the candidate (ZIP/City)



$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

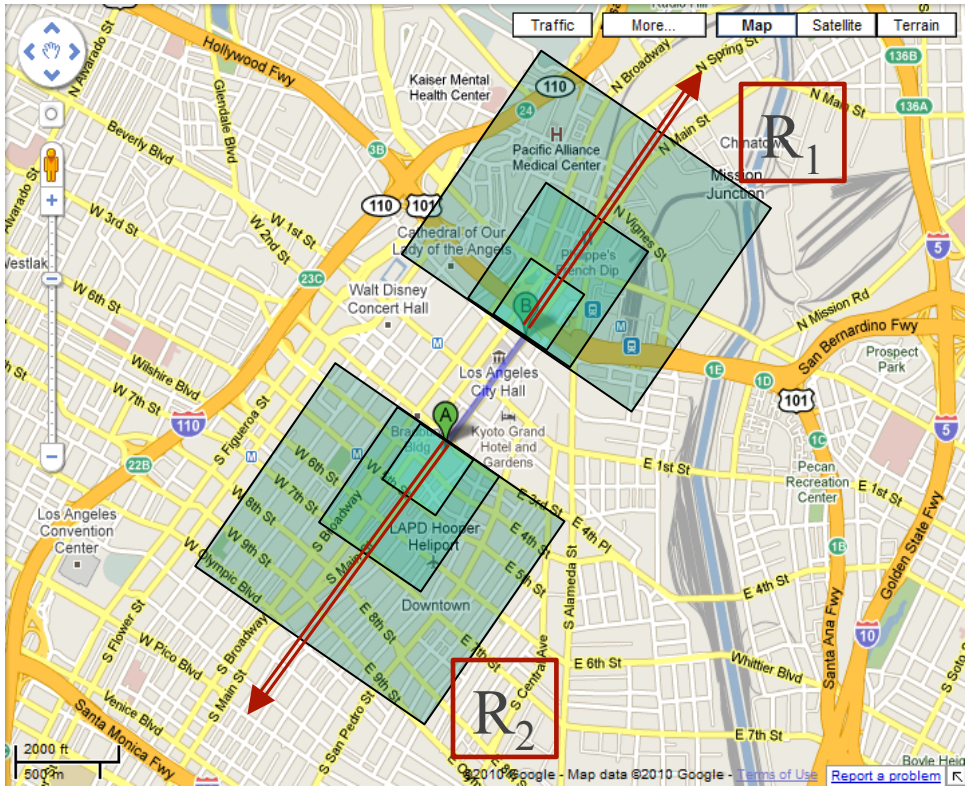
S = each attribute value (e.g., 90013, 90007)
p_i = the number of occurrences at each expansion

Intelligent Tie-Breaking Approaches – Geo-Intelligence- Frequencies

(1) Store the frequencies of the attributes in each region in vector *d*< ... >

(2) Determine a probability that the input address is located in each candidate region by the prevalence of the attribute value in question

300 E Main St, Los Angeles, CA 90013



Attr.	d_{R1}< … > (56 Segments)	d_{R2}< … > (87 Segments)
E	30	15
W	2	22
N	23	0
S	0	44

$$||[\bar{E}]|| = \sum_{i=1}^{n} ||[E_i]||$$ - Total number of segments in a region

$$p(r_i) = \frac{d_i(c_q)}{||[\bar{E}]||}$$ - Probability of picking a street segment with the correct attribute value in a region

$$\bar{p}(r_i) = \frac{p(r_i)}{\sum_{i=1}^{n} p(r_i)}$$ - Probability of picking a street segment with the correct attribute value in a region normalized by all regions

Intelligent Tie-Breaking Approaches – Evaluation

(1) How often do ambiguous reference features occur which prevent successful geocoding?
(2) What level of spatial error improvement results from the various alternative approaches?
(3) What level of spatial uncertainty improvement results from the various alternative approaches?

Sample Data:

Source	Note	Original count	Count after pre-processing
USC WebGIS transactions	Unknown and/or widely-varying quality	12,119,850	6,354,666
Medicare NPI file	Government list of self-reported data	2,903,156	1,086,196
LA County address points	Government list of cleaned data	2,890,639	2,890,639
Best Western hotels	Official company-reported list	2,074	2,074
Target stores	Official company-reported list	1,648	1,648
Totals		17,917,367	10,335,223

Test Geocoders:

(1) Census TIGER/Line, ZCTA, and Place Files

(2) ESRI ArcGIS Address Locator with StreetMap North America
- Shows how prevalent the problem is even with top-notch reference data

Random (Flip-a-coin):

(1) The random approach was run 5 times and the mean spatial error was used for analysis

Intelligent Tie-Breaking Approaches – Results

Dataset	Record count	USC street-level match ESRI street-level match	USC ambiguous ESRI ambiguous
USC WebGIS	6,354,666	4,752,122 (74.8%) 4,510,710 (71%)	45,941 (0.72%) 158,292 (2.49%)
NPI	1,086,196	946971 (87.2%) 922,955 (85%)	7,628 (0.7%) 19,748 (1.82%)
LA County	2,890,639	2,649,239 (91.6%) 2,564,852 (88.7%)	6,345 (0.22%) **15,785 (0.55%)**
Best Western hotels	2,074	1,772 (85.4%) 1,666 (78.6%)	17 (0.82%) 88 (4.24%)
Target stores	1,648	1,374 (83.4%) 1,295 (78.6%)	10 (0.61%) 34 (2.1%)
Total	10,335,223	8,351,478 (80.8%) 8,001,478 (77.4%)	59941 (0.58%) 193947 (1.88%)

(1) USC geocoder results in fewer ties than ESRI Address Locator

(2) Tie occur frequently even in high-quality data

(3) Ties are most prevalent because of address number ambiguities in the reference data (44% of cases)

(4) Directional ambiguities are also quite prevalent (> 30% of cases)

(5) Geo-intelligence chose the correct **82% of the time with address range** rules, **and 98% of the time with bounding box directional** approach

Count	% of total ambiguous records	Attribute	Cause
26,284	43.85	Number	Incomplete/ Incorrect
11,407	19.03	Pre-directional	Incomplete
3,874	6.46	Post-directional	Incomplete
3,695	6.16	Pre-directional	Incorrect
2,179	3.64	Suffix	Incomplete
1,591	2.65	City	Incorrect
1,334	2.23	Name	Incorrect
732	1.22	Zip	Incomplete
713	1.19	Post-directional	Incorrect
230	0.38	Zip	Incorrect
116	0.19	Pre-type	Incorrect
80	0.13	City	Incomplete
22	0.04	Zip	Incorrect
8	0.01	Post-qualifier	Incomplete
6	0.01	Pre-type	Incomplete
1	0.01	Pre-qualifier	Incorrect

Count	% of total	Address Range Relation
10,114	38.95	Contains
8,992	34.63	Next to
4,059	15.63	Overlap
2,379	9.16	Disjoint
420	1.62	Equivalent reversed

Conclusions

- Geocoding systems need to be open boxes
  - Users need to know what happened, why, and what the alternatives were to have confidence in fitness-for-use

- The USC WebGIS Geocoding framework aims to achieve these goals by providing an open source extensible approach to addressing the problem

- Our novel approaches
  (1) Improve match rates using nearby candidates instead of reverting to a lower level of geographic match
  (2) Reduce spatial error and uncertainty by using an uncertainty-driven approach to pick the most-likely output location given the candidates available and their spatial and topological relationships
  (3) Use intelligent tie breaking strategies that deduce the most likely outcome by interrogating the region around the ambiguous matches and investigating the relationships between their attributes

Current Status

- https://webgis.usc.edu
- Production system
  - > 3,000 users
  - 15 million geocodes produced
- .Net implementation on top of SQL server for reference data
- TIGER/Lines, LA County Parcel Data
  - Actively adding more parcel data
- Code is being reviewed, cleaned, and finalized before open sourcing

Thanks!

- Advisors and Committee Members
  - John Wilson, USC Geography
  - Craig Knoblock, USC Computer Science
  - Myles Cockburn, USC Preventive Medicine
  - Ulrich Neumann, USC Computer Science