Integrating Online and Geospatial Information Sources

Craig Knoblock

Cyrus Shahabi
Jose Luis Ambite
Maria Muslea

Snehal Thakkar
Jason Chen
Mehdi Sharifzadeh

University of Southern California

Introduction

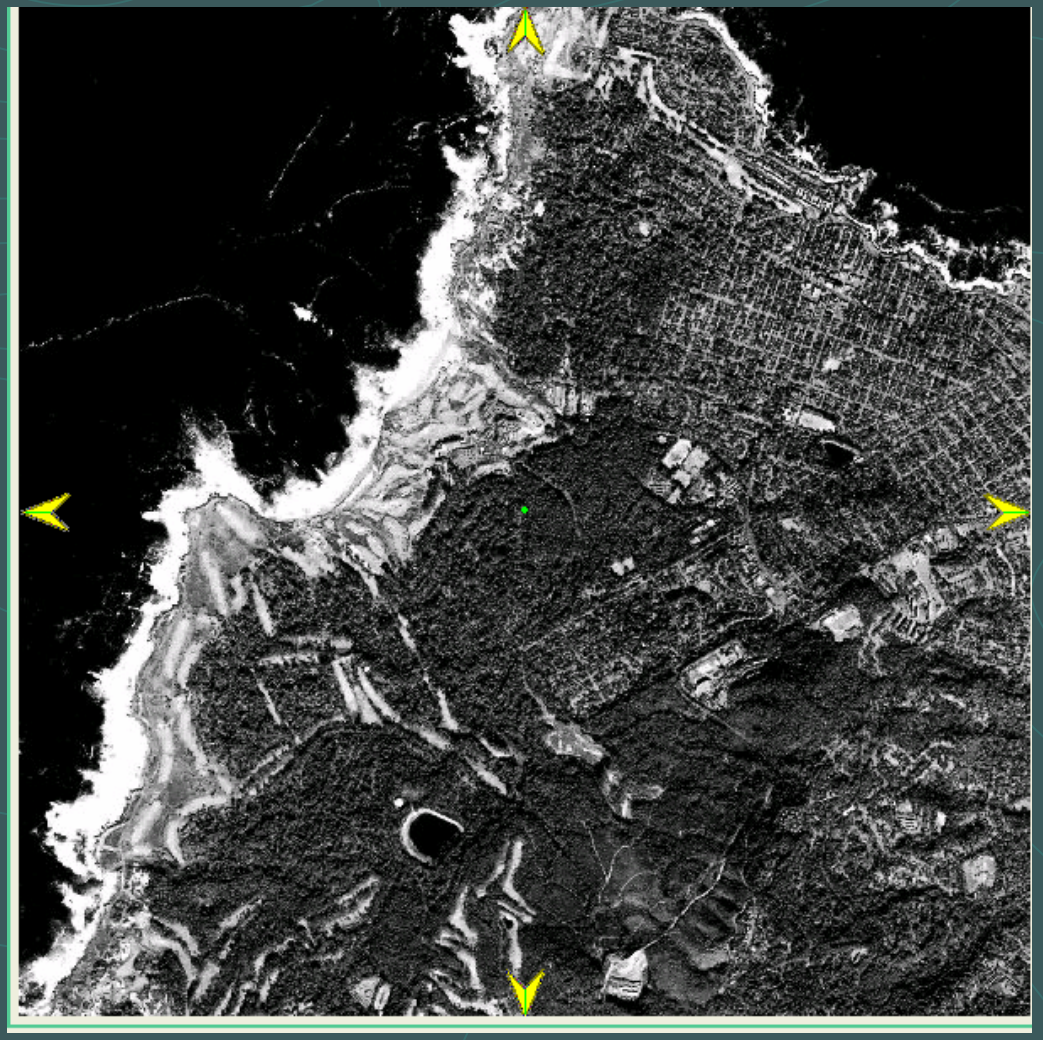
- Geospatial data sources have become widely available
- Huge amount of data available online that can be related to these geospatial sources
- Challenge is to support the dynamic integration of these two types of sources

Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

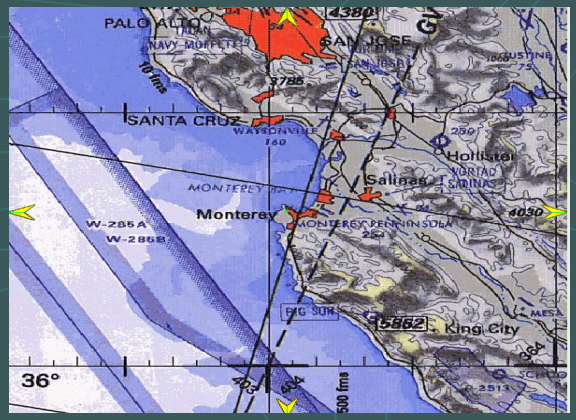
Geospatial Data Sources

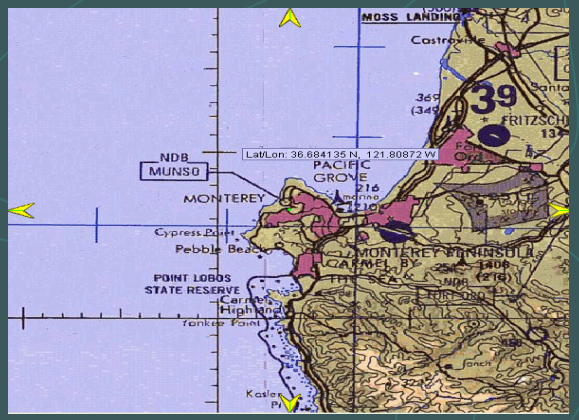
- Imagery

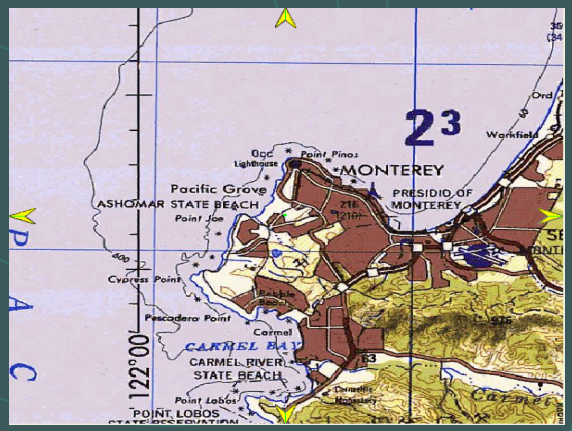


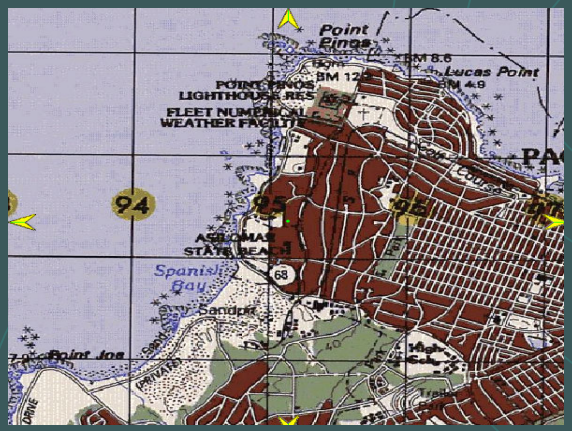
Geospatial Data Sources

- Imagery
- Maps



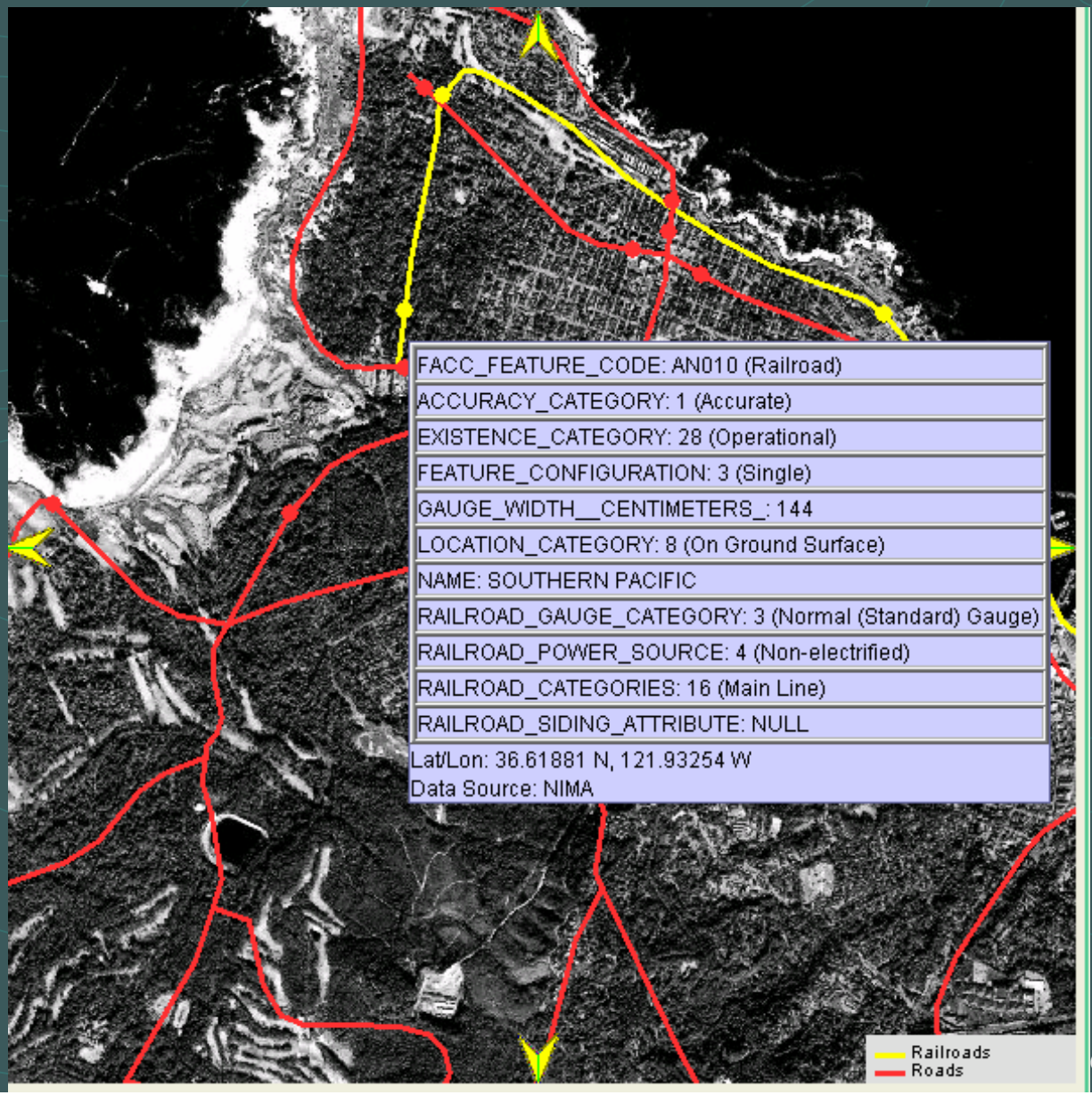






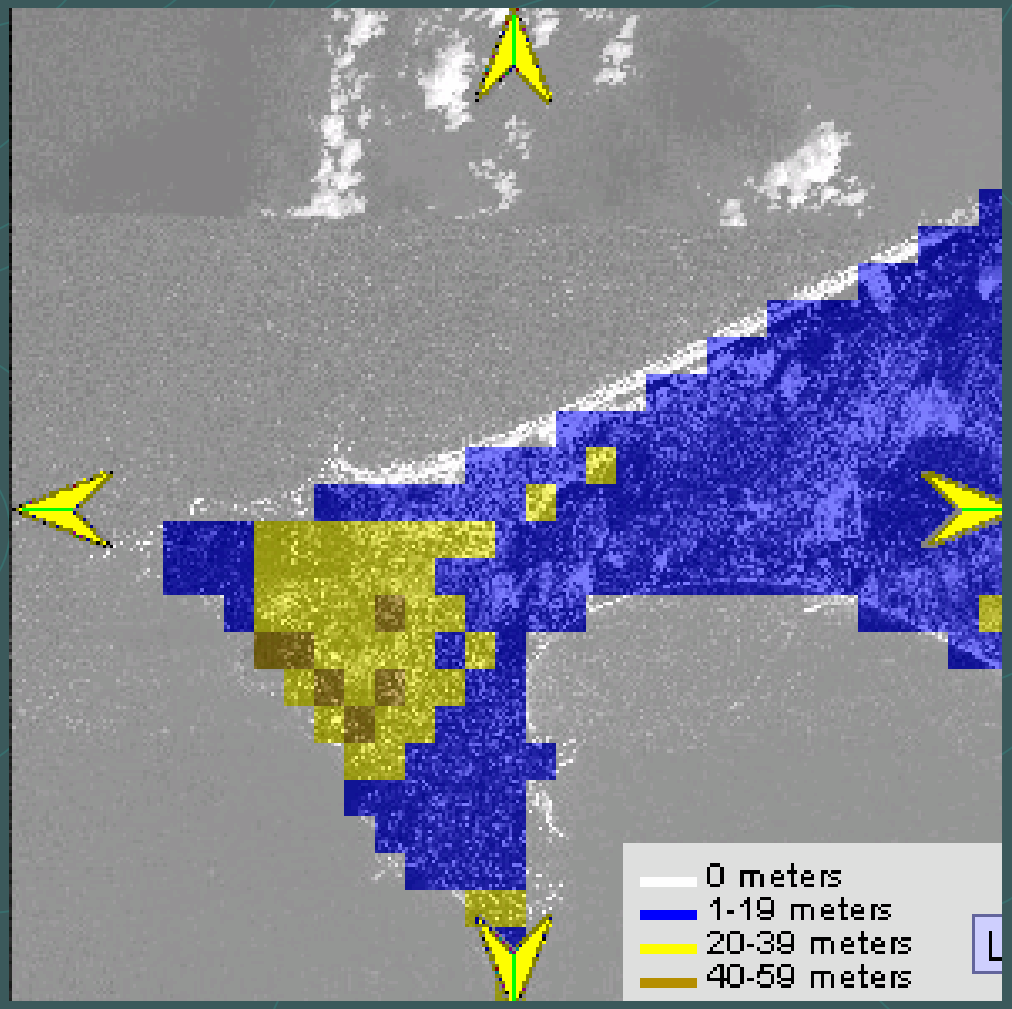
Geospatial Data Sources

- Imagery
- Maps
- Vectors



Geospatial Data Sources

- Imagery
- Maps
- Vectors
- Elevations



Geospatial Data Sources

- Imagery
- Maps
- Vectors
- Elevations
- Points

Name	YAKHCHALSAZIYEZ
Type	MFG
Feature Class	Spot Feature
Short Desc	factory
Long Desc	one or more buildings where goods are manufactured, processed or fabricated
Lat/Lon	35.7061 N, 51.2744 E
Data Source	NIMA

TerraWorld System

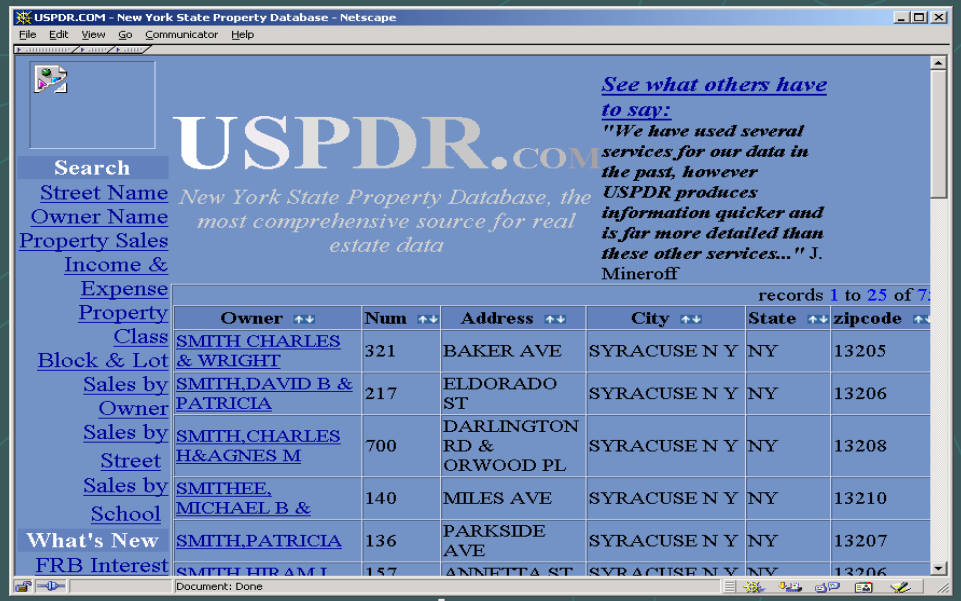
- Data from the National Imagery and Mapping Agency (NIMA)
  - Includes imagery, map, vector, elevation, and point data
  - Covers most of the world (including the oceans!)
- Hardware
  - 8 High-end Dell Servers
    - Separate servers for imagery & maps, vectors, databases, and web servers
  - Storage Attached Network (SAN)
    - 3 terabytes of storage
    - Provides high-speed data access to all servers

Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

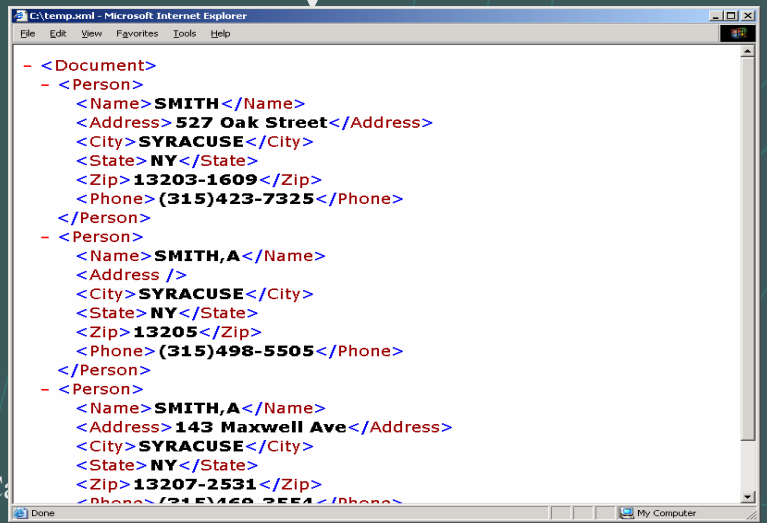
Semi-structured Data Sources

- Property tax sites



Semi-structured Data Sources

- Property tax sites
- Telephone books



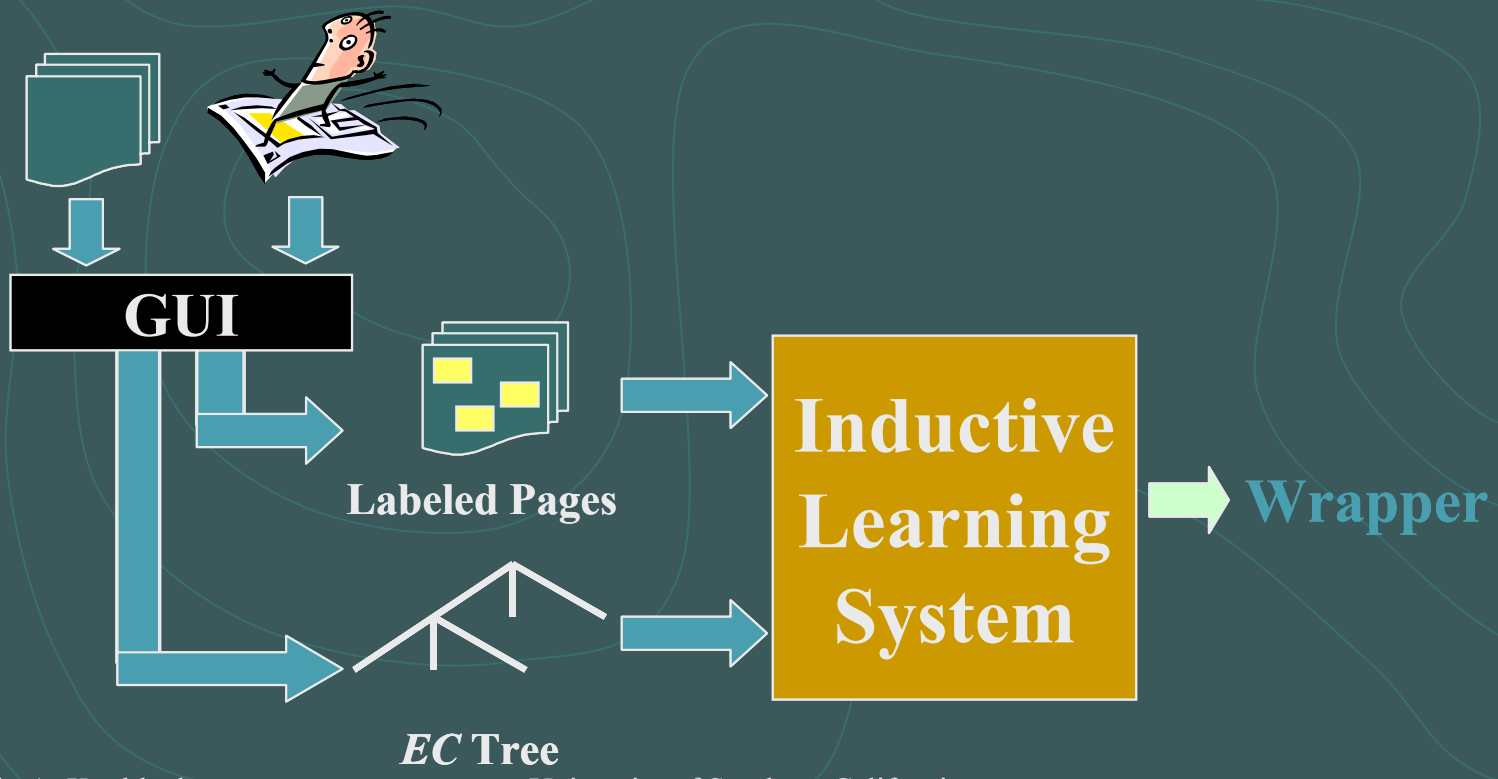
Semi-structured Data Sources

- Property tax sites
- Online telephone books
- Railroad schedules
- …


```
  </ROW>
 </TRAIN>
 <TRAIN>
  <ROW>
   <CITY>Tehran</CITY>
   <TIME>14:00</TIME>
  </ROW>
  …
 </TRAIN>
</IRANIAN_RAILWAYS>
```

Machine Learning of Wrappers

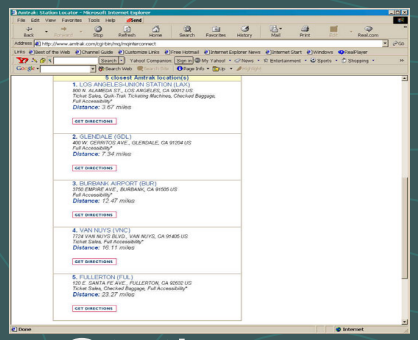
- Developed machine learning techniques for rapidly extracting data from semi-structured sources (wrapper)
- Started a spin-off company from ISI (Fetch Technologies) that has commercial product based on this work



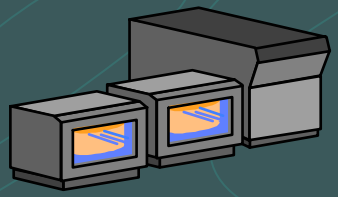
Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

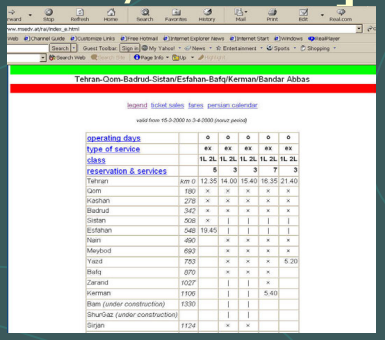
Combining Online Schedules with Vectors and Points [Shahabi et al., 2001]



Stations



Railroads



Schedules

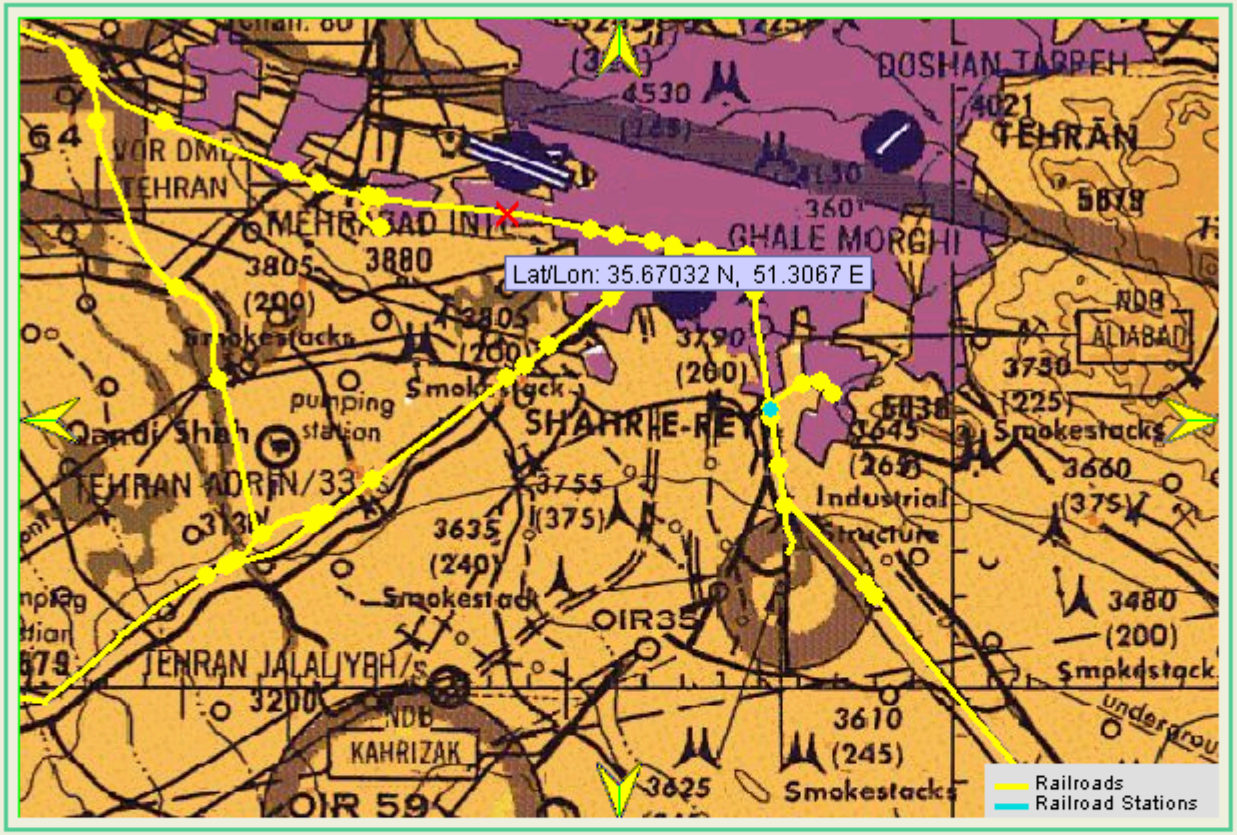
- How do we efficiently determine which trains will pass a given point or region
  - Railroad vectors specify all possible paths of the trains
  - Stations show the locations of the stops
  - Schedules provide the detailed timetable and stops

Integrating Schedules with Vector Data

- Approach:
  - Create a wrapper for the online schedule and download it to a database
  - Match the names of the stations in the online schedule with the names of the stations in the gazetteer
    - Exploits work we have done on record linkage across sources
  - Align the points in the gazetteer with the vector data of the railroads
  - Find the shortest paths between the stations
  - Compute the trains that will pass a given region within some time interval
    - Determines how much real paths can deviate from the shortest distance between two points to compute this efficiently

Integrating Schedules with Vectors

Integrating Schedules with Vectors



Map

ID	Train Name	Departure City	Departure Time	Arrival City	Arrival Time	Estimated Time
0	Mehrabad-Tehran	MEHRABAD	03:45	TEHRAN	04:59	04:48
1	Tehran-Mehrabad	TEHRAN	02:12	MEHRABAD	03:24	02:24
90	Khoramshahr/Bandar	RUDESHUR	15:30	TEHRAN	17:35	17:20
95	Khoramshahr/Bandar	RUDESHUR	11:05	TEHRAN	13:12	12:57
100	Tehran-Qom-Arak-Ahv	TEHRAN	18:00	RUDESHUR	20:05	18:45
105	Tehran-Qom-Arak-Ahv	TEHRAN	14:05	RUDESHUR	16:15	14:52

Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

Aligning Vectors with Imagery (Chen et al., 2003)

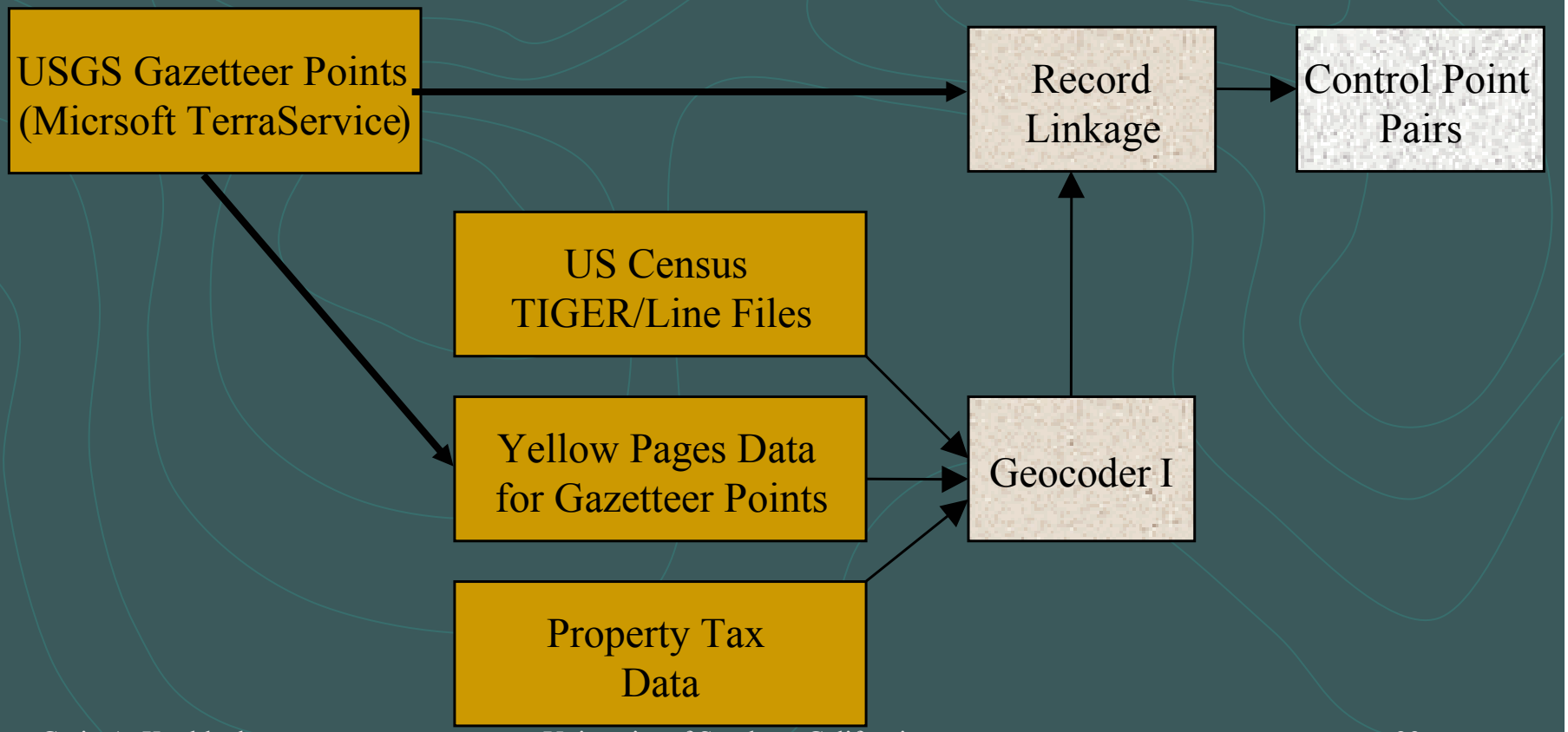
- Integration Challenges
  - Different geographic projections
  - Global transformations do not exist
  - Previously this was performed by:
    - Manually identifying control points
    - Applying conflation techniques

Conflation

- Conflation: Compiling two geo-spatial datasets by establishing the correspondence between the matched entities and transforming other objects accordingly.
- Requires identifying matched entities, named control points, on the image and the vectors
  - Each pair of corresponding control points from the two datasets indicates corresponding positions on each datasets
  - Existing algorithms only deal with vector to vector spatial data integration or accomplish imagery to vector data integration manually
  - We explored two techniques
    - Control points generated from online sources
    - Control points produced from localized image processing

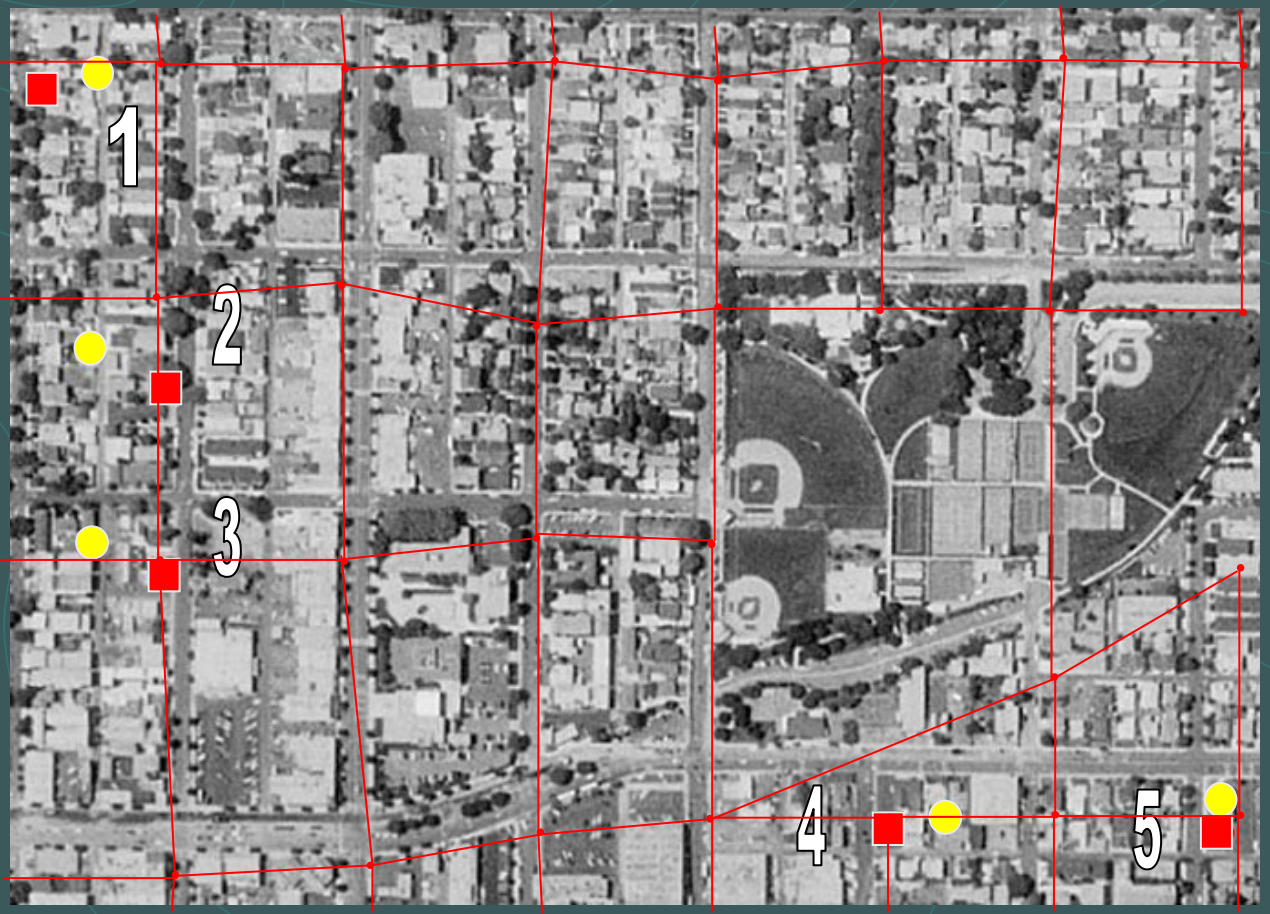
Finding Control Points Using Online Sources

- Online sources can be used to locate points on vector data



Finding Control Points Using Online Sources

Control Point Pairs



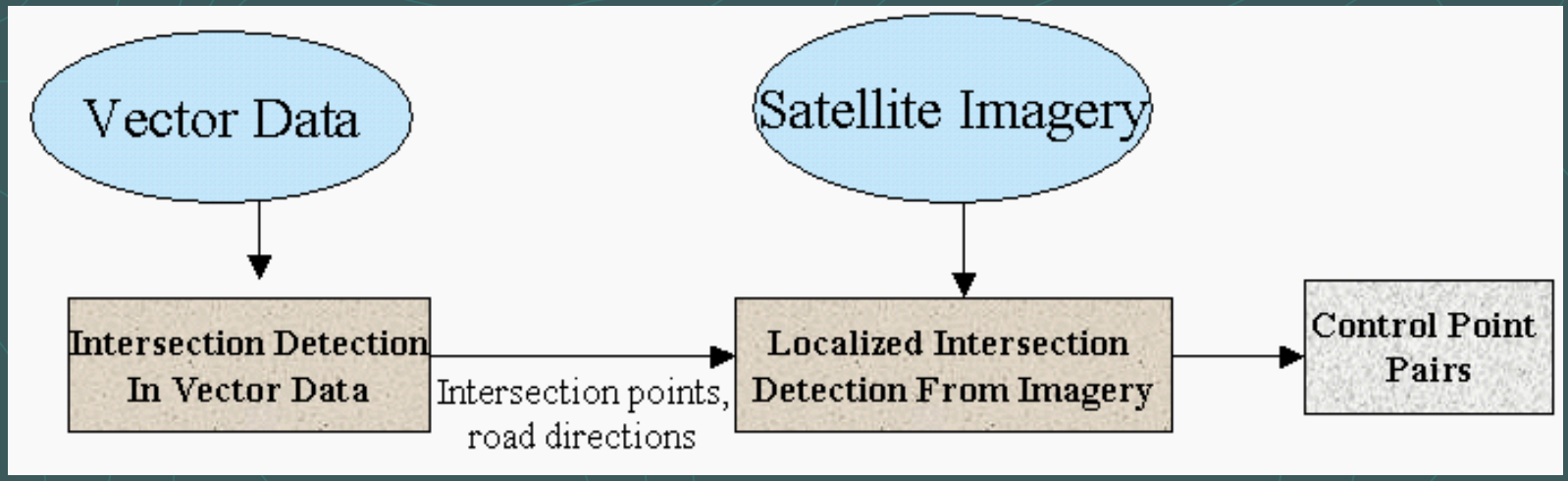
Features Previously Identified on Imagery
(Yellow points)

Feature Name	Latitude	Longitude
Church of Christ	33.91971	-118.40790
El Segundo Christian Church	33.91811	-118.41790
El Segundo Public Library	33.92391	-118.41690
El Segundo Foursquare Church	33.92154	-118.41750
First Baptist Church	33.92531	-118.40990

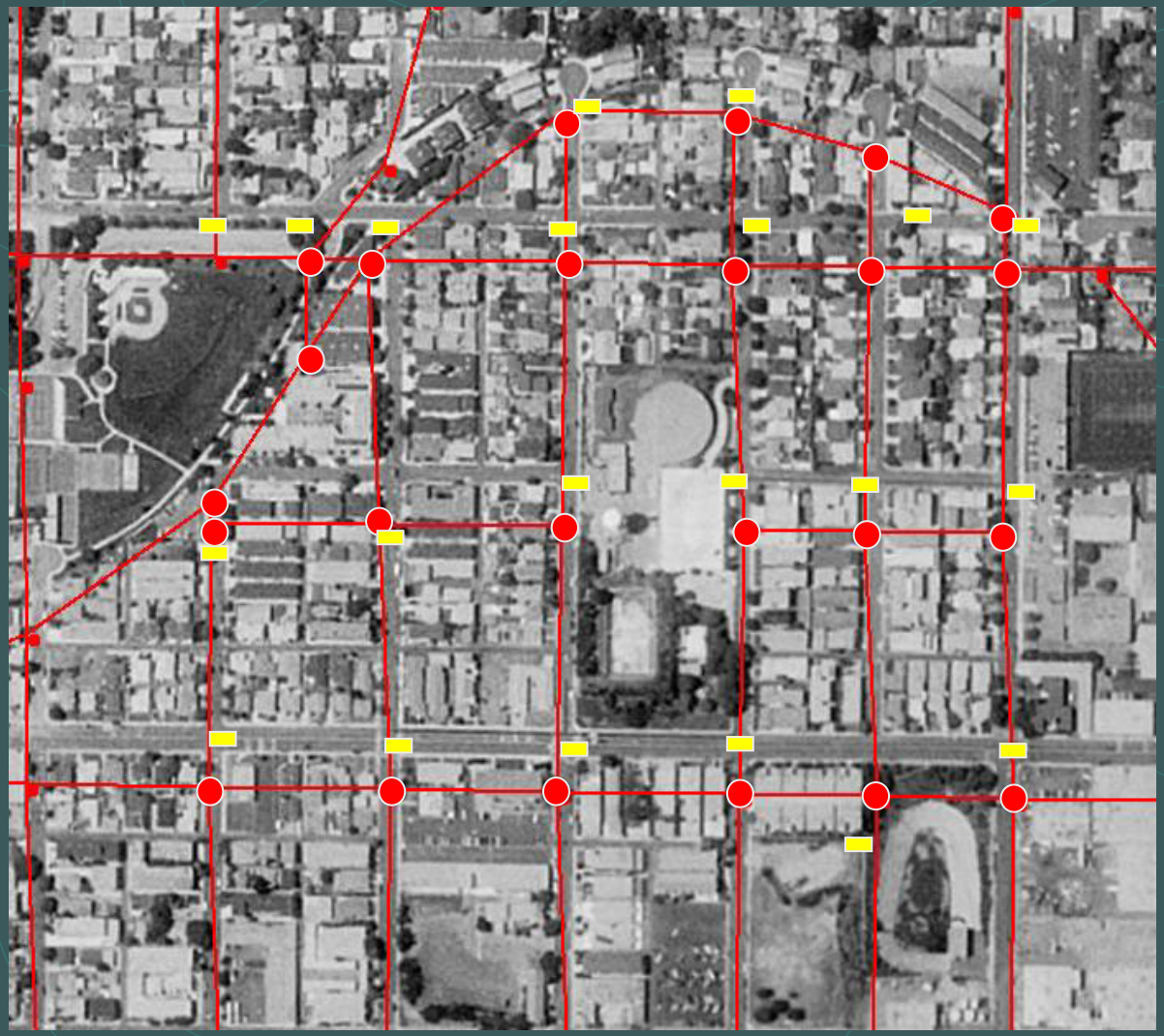
Points on vector data
(Red points)

Feature Name	Address
Church of Christ El Segundo Hilltop Community	717 East Grand Ave
El Segundo Christian Church	223 West Franklin Ave
El Segundo Public Library	111 W Mariposa Ave
Foursquare Church Of El Segundo	429 Richmond Street
First Baptist Church of El Segundo	591 East Palm Avenue

Finding Control Points Using Localized Image Processing



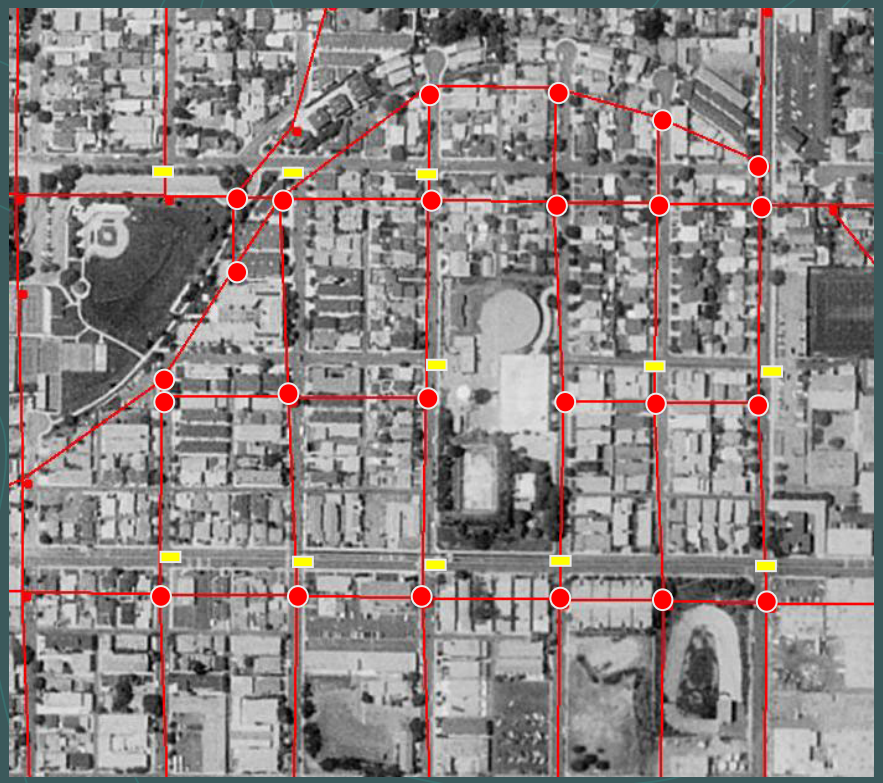
Resulting Control Point Pairs



Intersection Points Located on Vector Data (Red points)

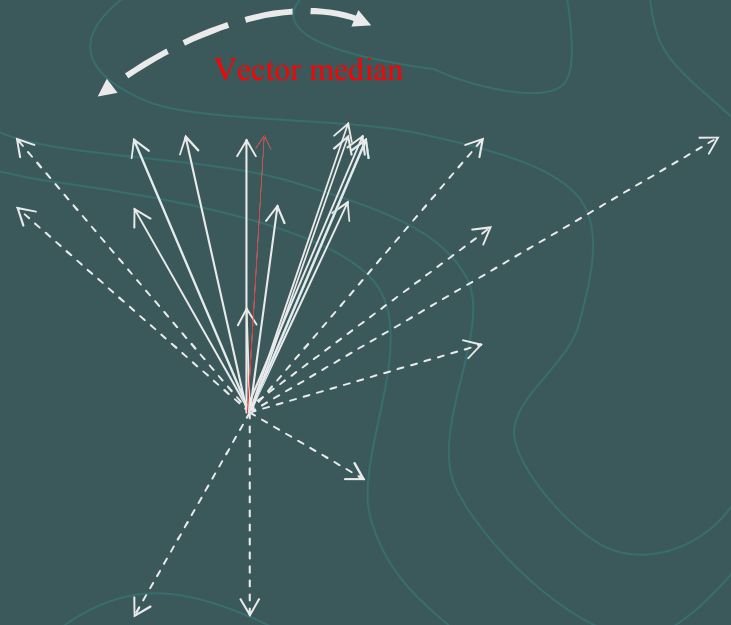
Intersection Points Detected on Imagery (Yellow points)

Filtering Control Points
Vector Median Filter



After Filtering

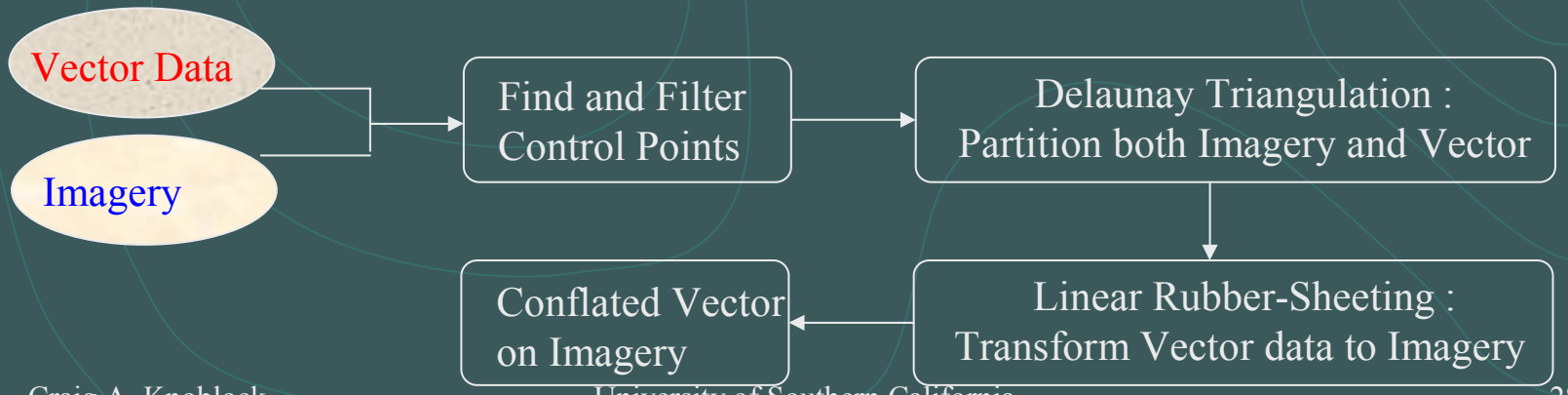
Keep half control-point vectors



Control-point vectors

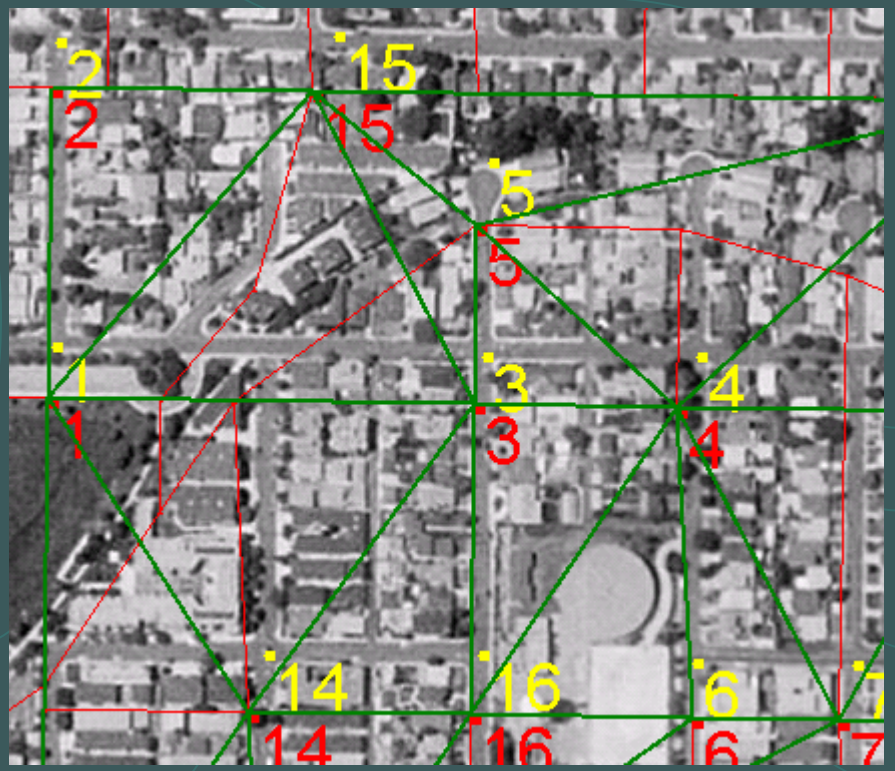
Conflating Imagery and Vector Data

- Conflate imagery and vector data by computing the transformations between the control point pairs and transforming other objects accordingly
- Two steps
  - Delaunay Triangulation
    - Partition the space into multiple triangles
  - Linear Rubber-Sheeting
    - Stretching of vector data within each triangle as if it was made of rubber



Conflating Imagery and Vector Data: Delaunay Triangulation

- Sub-divide the vector data into multiple triangles using the control points as vertices, then construct the corresponding triangles on the imagery



Red lines : Original Road Network

Point : Control Point Pairs

Green lines: Delaunay Triangulation

Conflating Imagery and Vector Data: Linear Rubber-Sheeting

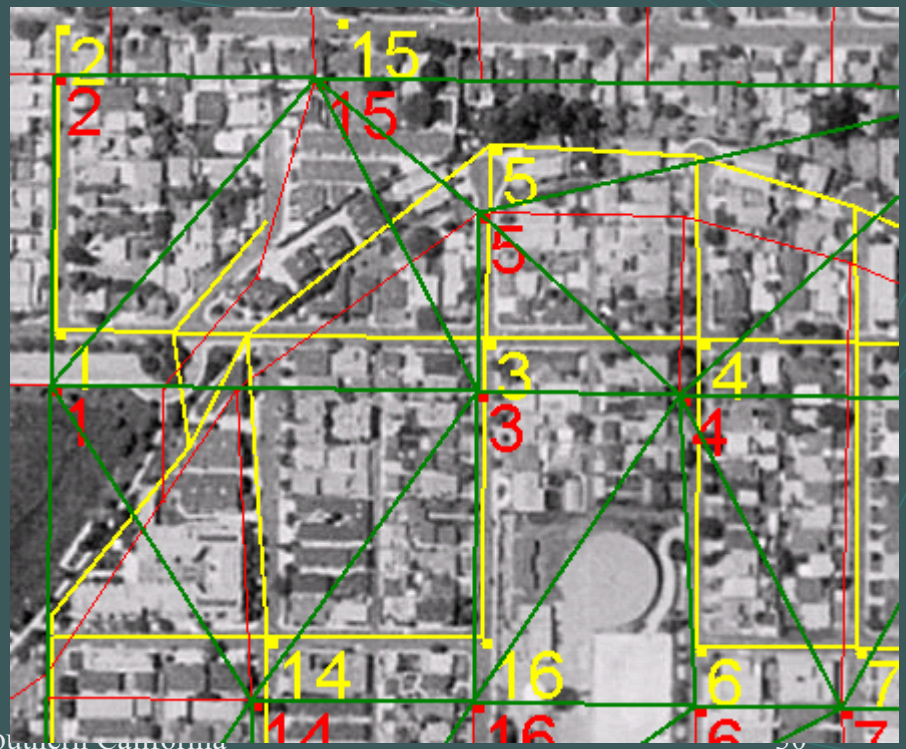
- Imagine stretching a vector map as if it was made of rubber
- Deform algorithmically, forcing registration of control points over the vector data with their corresponding points on the imagery

Red lines : Original Road Network

Yellow lines : Conflated Road Network

Point : Control Point Pairs

Green lines: Delaunay Triangulation



Results

El Segundo Dataset	Mean Displace.	Std Dev	Mean + Std Deviation
Original TIGER/Lines	26.19	5	(21.19, 31.19)
Using Online Sources	15.92	8.38	(7.54, 24.3)
Using Local Image Pro	8.61	6	(2.61, 14.61)

Conflation Results of Using Localized Image Processing

Before Conflation

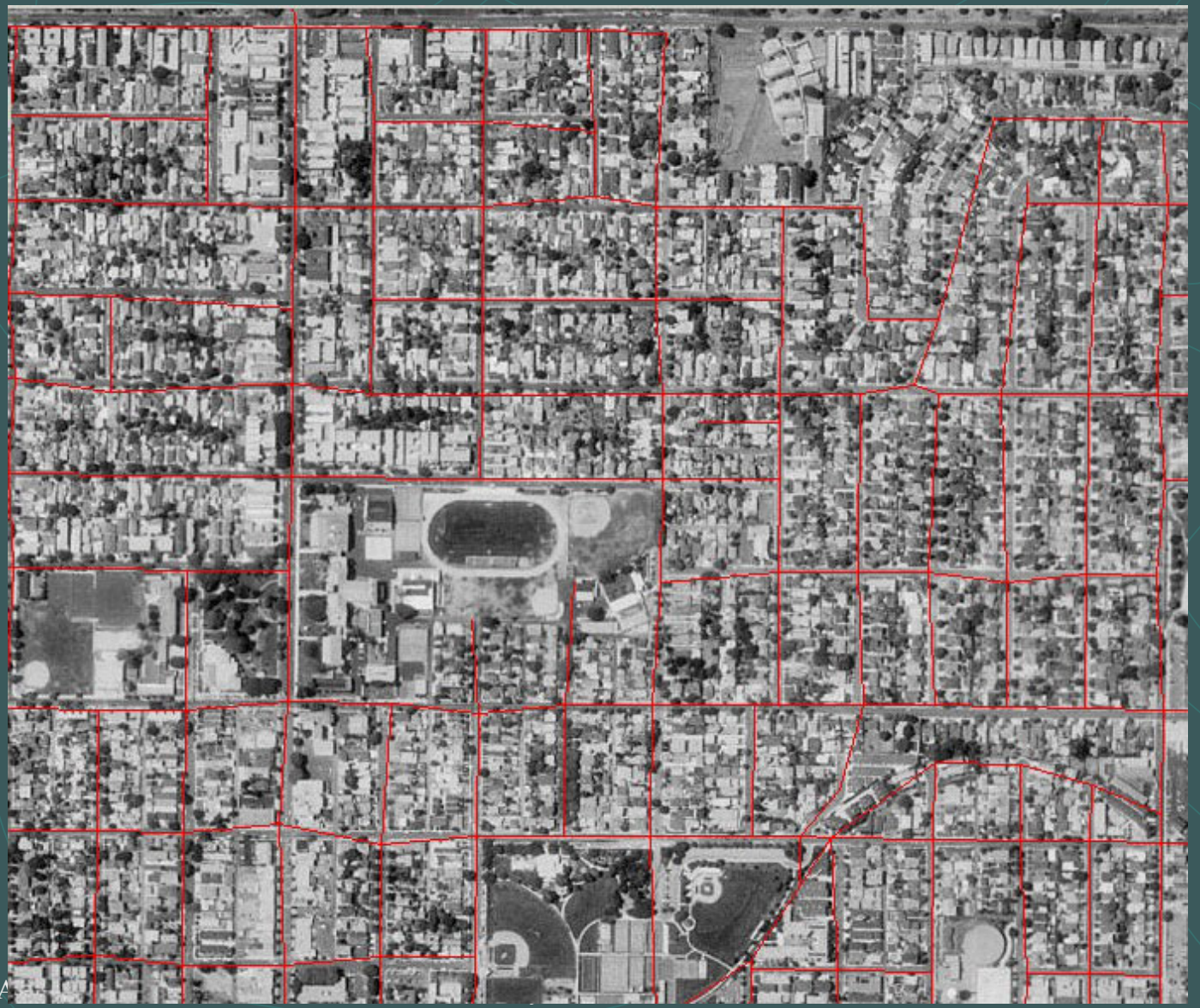
After Conflation

Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

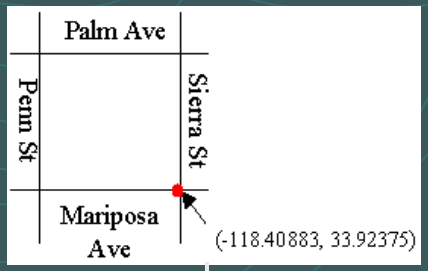
Identifying Structures in Imagery

Locate the Roads in the Image



Exploiting Online Sources to Accurately Identify Structures in Imagery

Satellite Image
Terraserver

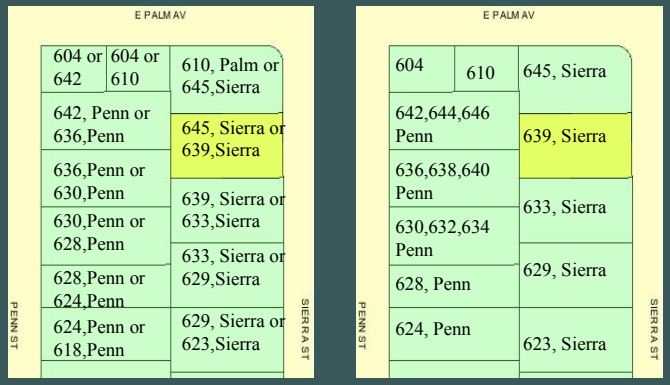


Street Vector Data
Corrected Tiger Line Files

Street Address	City, State Zipcode
642 Penn St	El Segundo, CA 90245
640 Penn St	El Segundo, CA 90245
636 Penn St	El Segundo, CA 90245
604 Palm Ave	El Segundo, CA 90245
610 Palm Ave	El Segundo, CA 90245
645 Sierra St	El Segundo, CA 90245
639 Sierra St	El Segundo, CA 90245

Census Master Address File

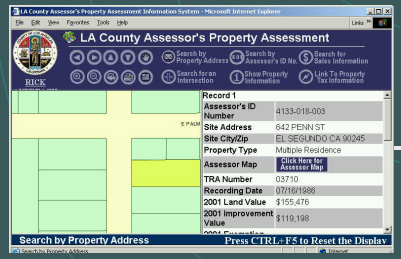
Constraint Satisfaction



Initial Hypothesis Result After Constraint Satisfaction

Address	Latitude	Longitude
642 Penn St	33.923413	-118.409809
640 Penn St	33.923412	-118.409809
636 Penn St	33.923412	-118.409809
604 Palm Ave	33.923414	-118.409809
610 Palm Ave	33.923414	-118.409810
645 Sierra St	33.923413	-118.409810
639 Sierra St	33.923412	-118.409810

Geocoded Houses

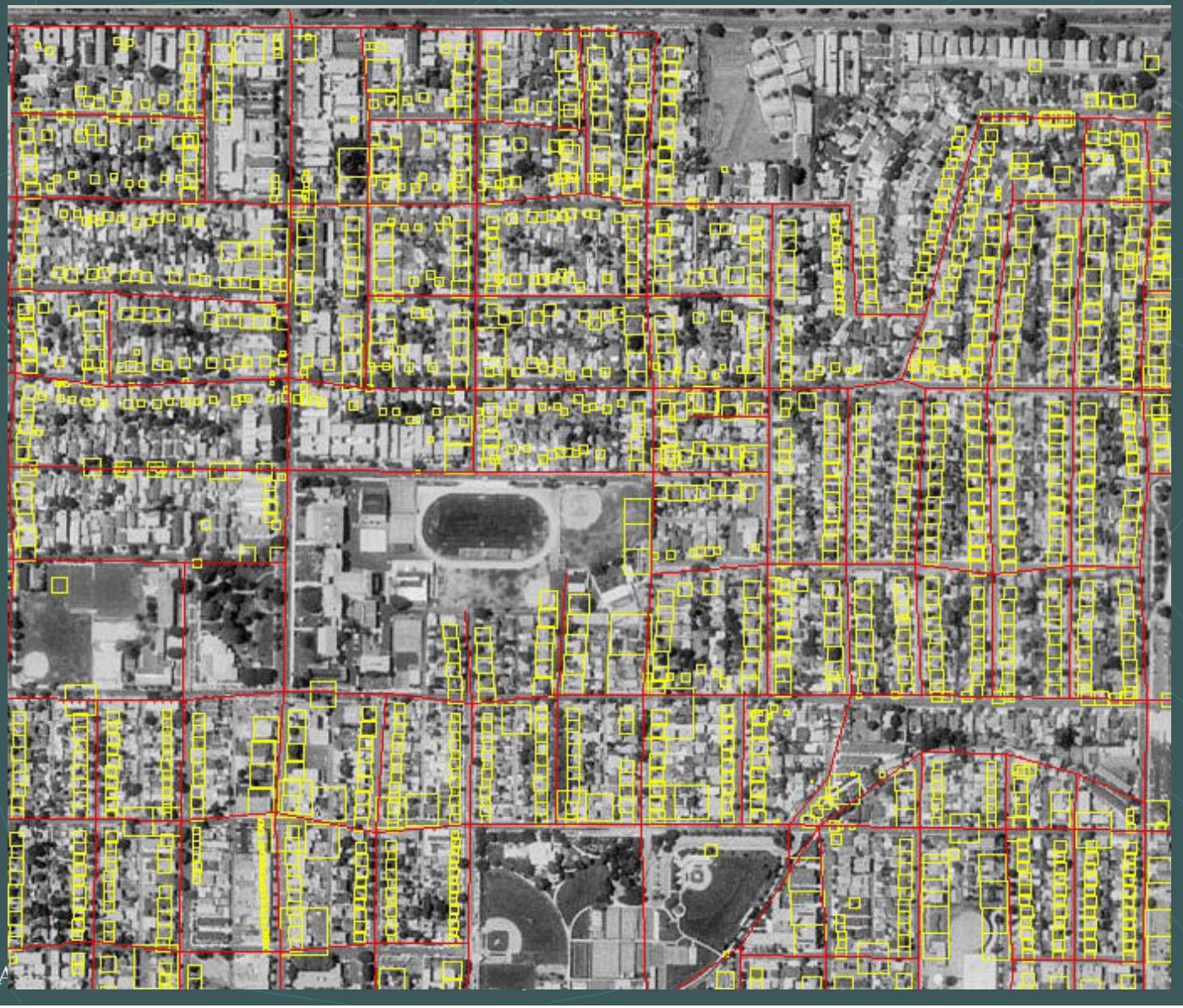


Los Angeles County Assessor's Site
Property Tax Records

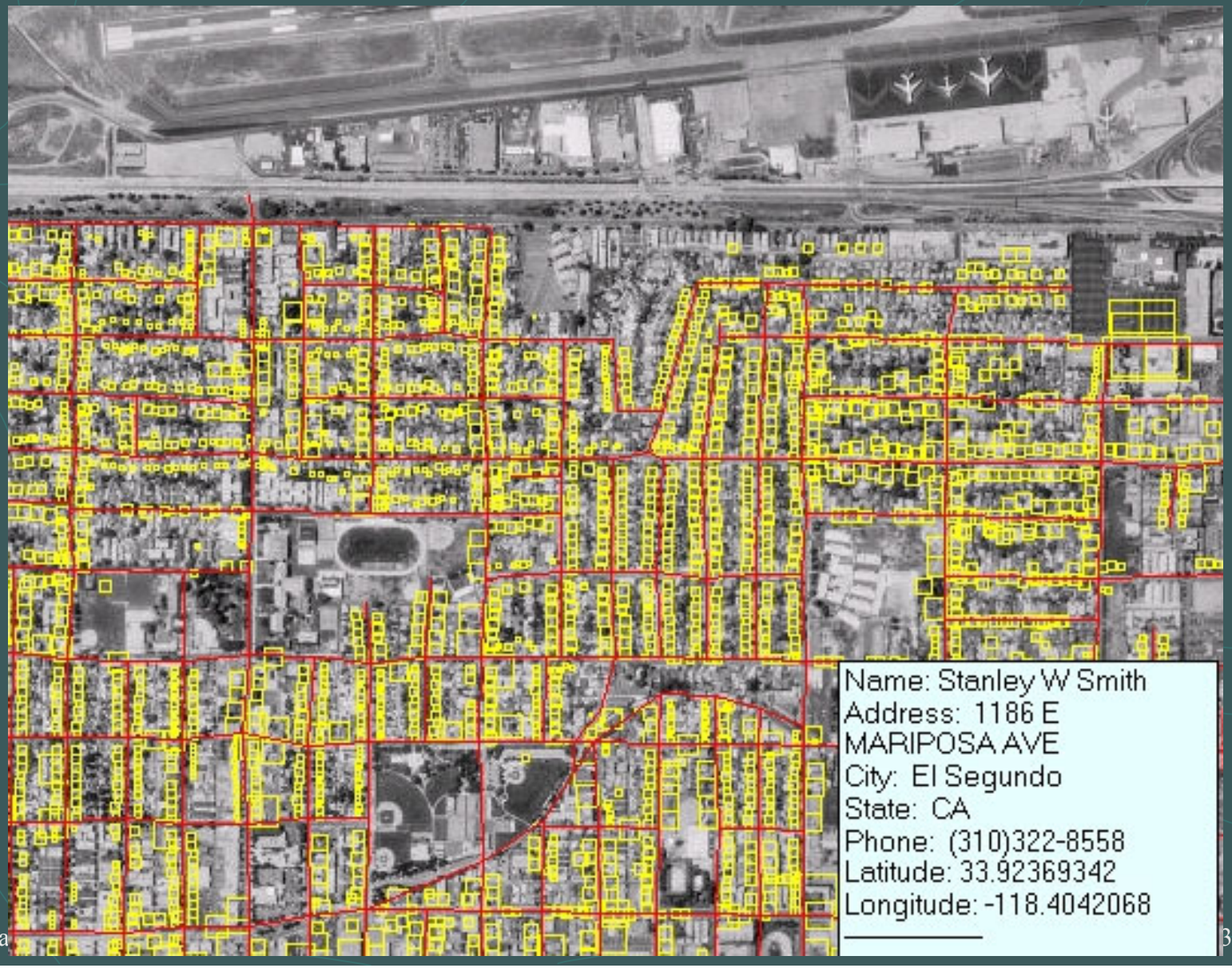
Address	# units	Area (sq ft)	Lot size
642 Penn St	3	1793	135.72 * 53.33
604 Palm Ave	1	884	69 * 42
610 Palm Ave	1	756	66 * 42
645 Sierra St	1	1337	120 * 62
639 Sierra St	1	1408	121*53.5

Data Extracted from On-line Site

Identifying Structures in Imagery



Labeling Structures in Imagery



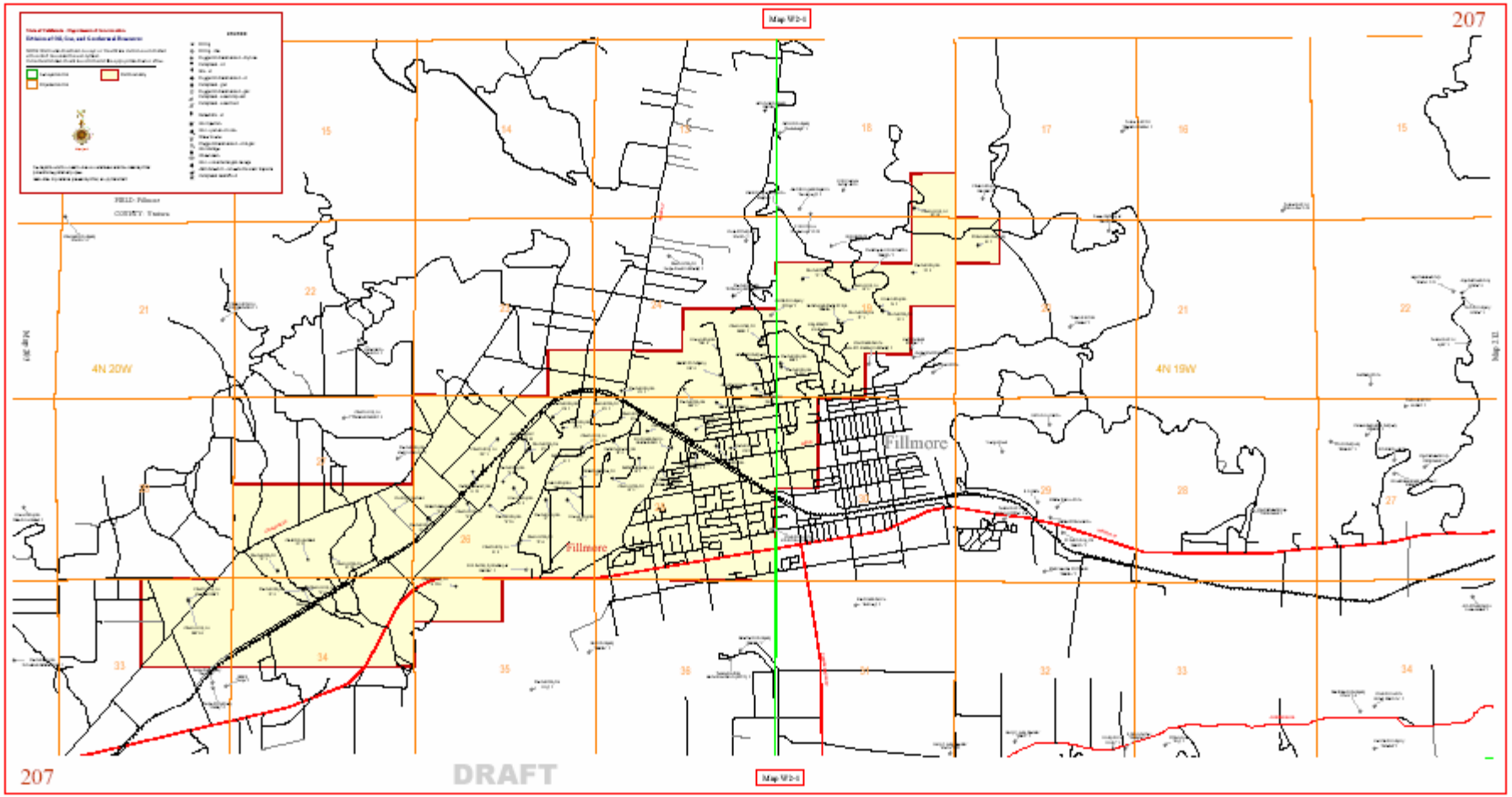
Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

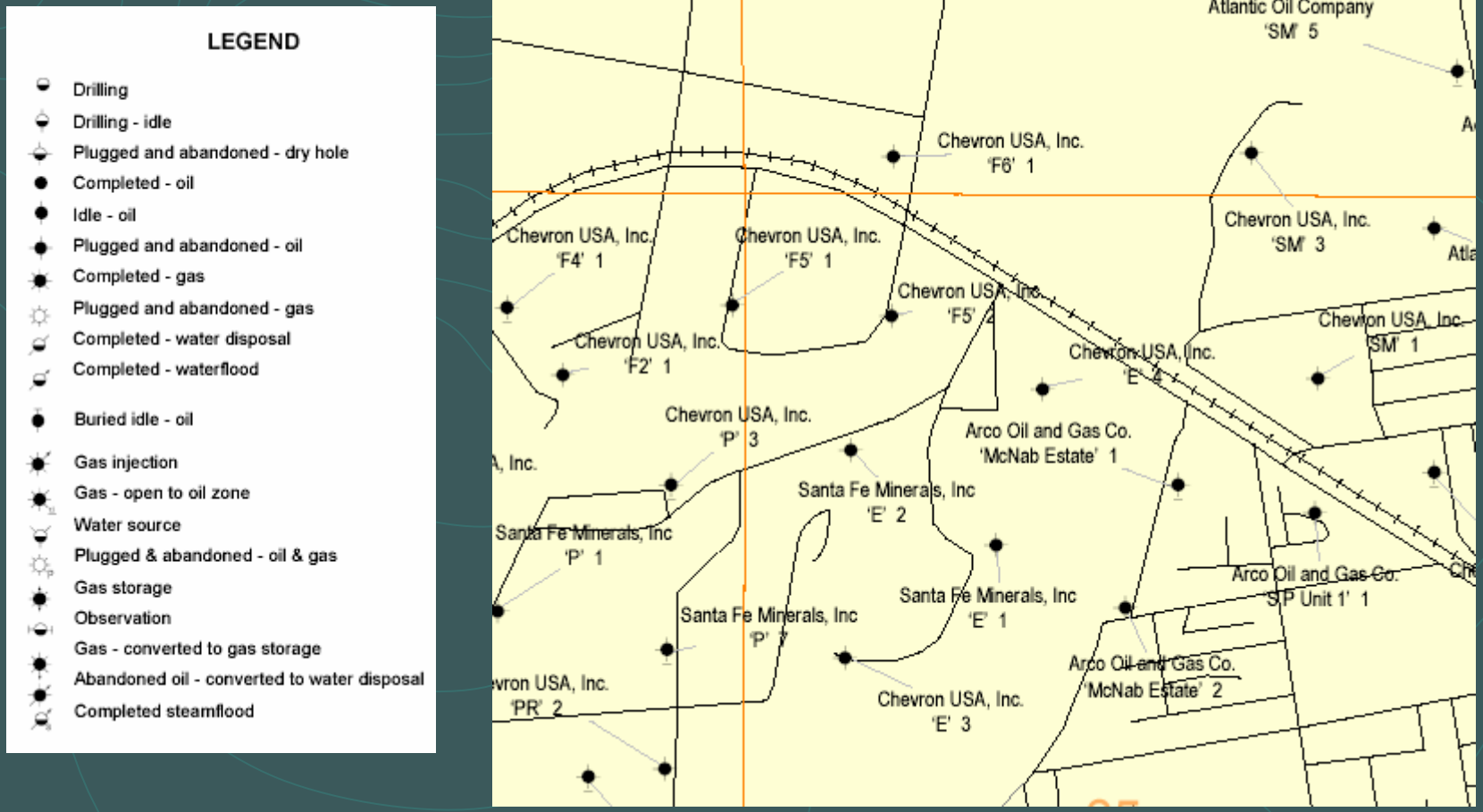
Integrating Vectors and Points with Online Oil Field Maps

- Goal: Determine which houses are built over abandoned oil wells
  - Integrate the online oil maps with street vector data
  - Challenge:
    - Not given lat/long coordinates of maps
    - Given a database of some of the oil wells on the maps
- Source : California Dept. of Conservation, Division of Oil, Gas and Geothermal Resources
  - http://www.consrv.ca.gov/DOG/maps/index_map.htm
  - Maps: in PDF format.
  - Wells information : vector(point) dataset contains, for example, status/operator/lat/long

Sample Oil Map



Sample Oil Map (Zoom In)



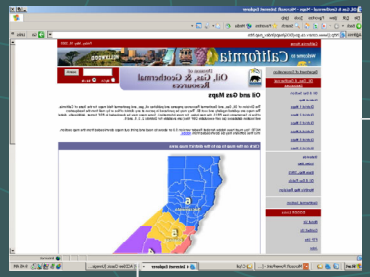
Vector Data (Online Wells Info)

apinumber	STATUS...	OPERATOR	LEASE	WELLNO_	SEC	map	RNG	TWN	LATITUDE	LONGITUDE
11102789	004	Wasibi Oil Co	Well No.	B	34	207	19W	05N	34.491765000000001	-118.856876
11106431	014	Arizona Oil Company	Well No.	1	13	207	20W	04N	34.431190999999998	-118.92027899999
11106042	006	Texaco E & P Inc	Standrad-Arundell	1	17	207	19W	04N	34.429599000000003	-118.885361
11106013	006	Jahn's Oil Company	Goodenough	1	18	207	19W	04N	34.42944	-118.918792
11106004	006	C. W. Colgrove	Bursin	46-18	18	207	19W	04N	34.424782999999998	-118.912588

apinumber	STATUS...	OPERATOR	LEASE	WELLNO_	SEC	map	RNG	TWN	LATITUDE	LONGITUDE
11106014	006	Ken-Cal Oil Company	Calumet	1	34	207	19W	04N	34.379790999999997	-118.86073
11106017	006	Harry C. Long, O...	Basolo	1-31	31	207	19W	04N	34.379472999999997	-118.902154
11106432	014	Arizona Oil Company	Well No.	2	13	207	20W	04N	34.312612000000001	-119.326383000000
11100048	015	Atlantic Oil Company	SM	4	24	207	20W	04N	34.310805000000002	-119.331548
11100053	015	Chevron USA, Inc.	P	3	26	207	20W	04N	0.0	0.0

- Issue : Some wells are detected on the maps while not found on the vector data, and vice versa.

Integration Approach (Work in Progress)



PDF

PDF to Image : Ghostscript (GSView)

Image

Extracting well points

Extracted Points

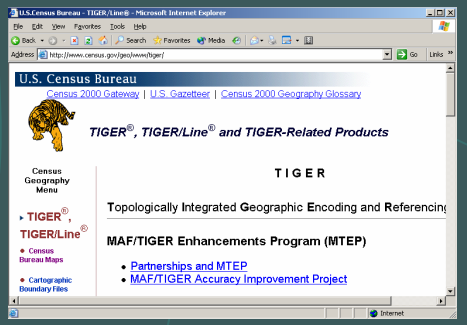
Online Wells-Info (*.dbf)

Georeferenced Map

Well points matching

Integration

DB Points



Corrected Vector Data

Vector datasets (point datasets)

Vector datasets (line datasets/ TIGERLines)

Outline

- Geospatial Data Sources
- Semi-structured Data Sources
- Integrating Semi-structured and Geospatial Sources
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Discussion and Future Work

Discussion

- Described four example applications
  - Combining online schedules with vectors and points
  - Using online sources and image processing to align vectors and imagery
  - Exploiting property records to identify structures in imagery
  - Integrating vectors and points with online oil field maps
- Goal is not to develop the specific applications, but to develop the techniques for automatically integrating these diverse types of sources

Future Work

- Build a general framework for integrating online and geospatial data sources
- Our previous integration work focused on integrating structured data (e.g., SIMS & Ariadne projects at USC)
- Extend this to support geospatial data types (imagery, maps, vectors, elevations, points)
- Develop integration techniques over these types
  - Conflation → integration imagery and vectors
  - Moving object queries → queries across time and space
  - Constraint satisfaction → integrating different types of data
- Investigate approaches to rapidly and automatically integrating these sources