Tools and Data Sources

Michael F Goodchild
University of California
Santa Barbara

Outline

- GIS 101
- Spatial disconnects
- Tracking data
- Data sources

Geographic information system

- System to acquire, store, transform, analyze, display, share, archive geographic information
- Geographic information
  - information about the specific characteristics of places on or near the Earth's surface
  - <***x***,***z***> where ***x*** is a location in space-time and ***z*** is some set of general properties
  - often aggregated to statements about regions <***R***,***z***> where ***R*** is standardized

Motivations

- Map compilation and editing
- Measurement from maps
- Economies of scale
  - build the foundation to handle a particular data type, and additional functions can be added quickly and cheaply
  - doing anything conceivable with geographic information

GIS data types: discrete objects

- Points
  - instances of a disease
  - residential location
- Lines
  - roads, rivers, tracks of individuals
- Areas
  - reporting zones (counties, tracts)
  - areas of risk (buffer zone around nuclear plant)
- Associated attributes
- Relationships to other objects

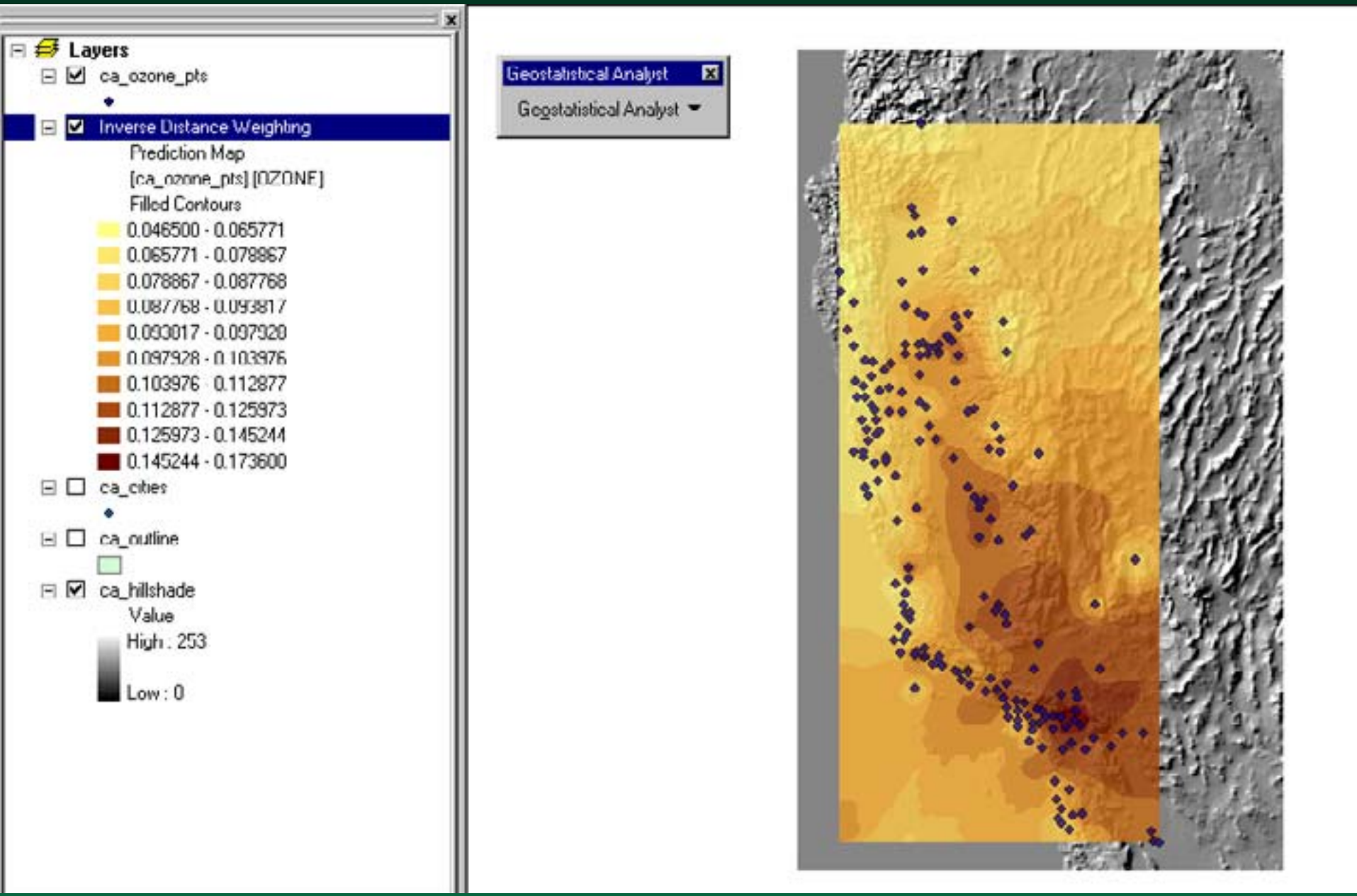
Location as attribute

- The data table

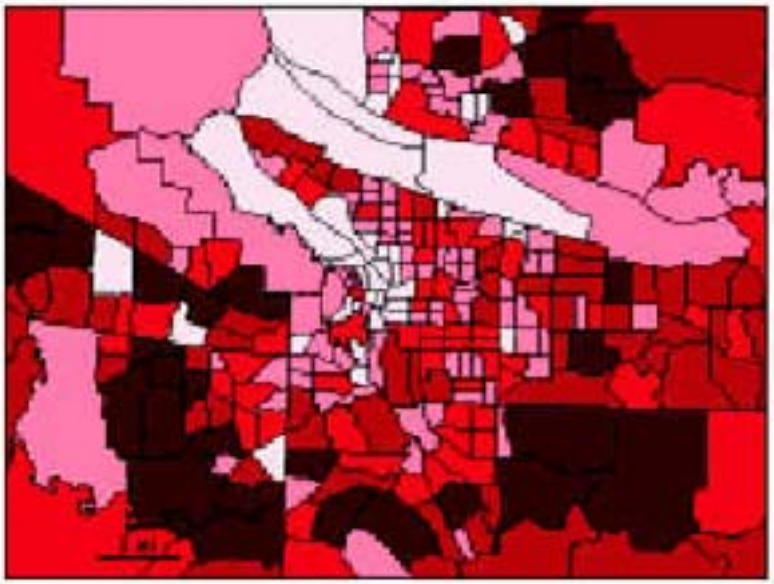
Tract	Pop	Location	Shape
1	3786	*x,y*	
2	2966	*x,y*	
3	5001	*x,y*	
4	4983	*x,y*	
5	4130	*x,y*	
6	3229	*x,y*	
7	4086	*x,y*	
8	3979	*x,y*	

GIS data types: continuous fields

- Variable as function of location
  - $z = z(x,y)$
  - elevation as the common metaphor
  - but *z* may be a class (land use type)
  - exactly one value at any location
- Environmental risk factors
  - individuals (objects) moving through continuously varying exposure (fields)


Layers
ca_ozone_pts
Inverse Distance Weighting
Prediction Map
[ca_ozone_pts] [OZONE]
Filled Contours
0.046500 - 0.065771
0.065771 - 0.078867
0.078867 - 0.087768
0.087768 - 0.093817
0.093817 - 0.097928
0.097928 - 0.103976
0.103976 - 0.112877
0.112877 - 0.125973
0.125973 - 0.145244
0.145244 - 0.173600
ca_cities
ca_outline
ca_hillshade
Value
High : 253
Low : 0
Geostatistical Analyst
Geostatistical Analyst

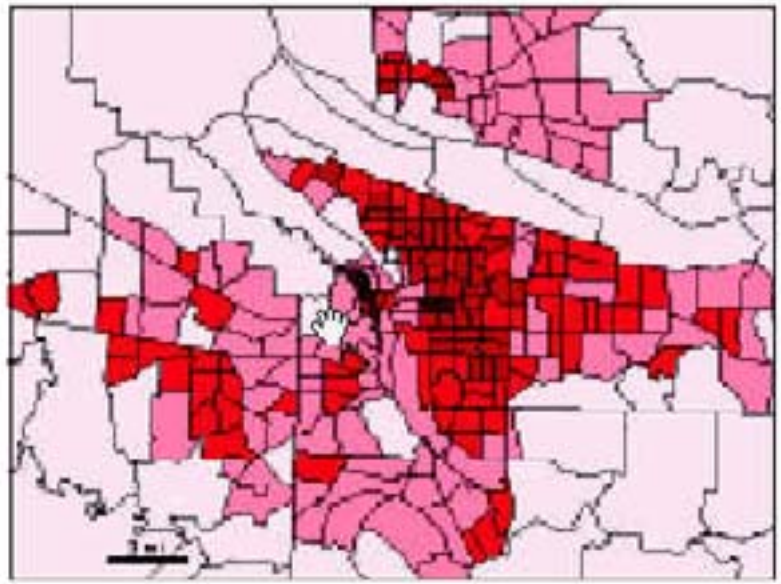
If you want to know approximately how many people each census tract has, map total population.



132 - 2224
2225 - 3900
3901 - 5507
5508 - 7732
7733 - 12611

Census tracts by total population.

If you want to know where most of the people are concentrated. map population density.

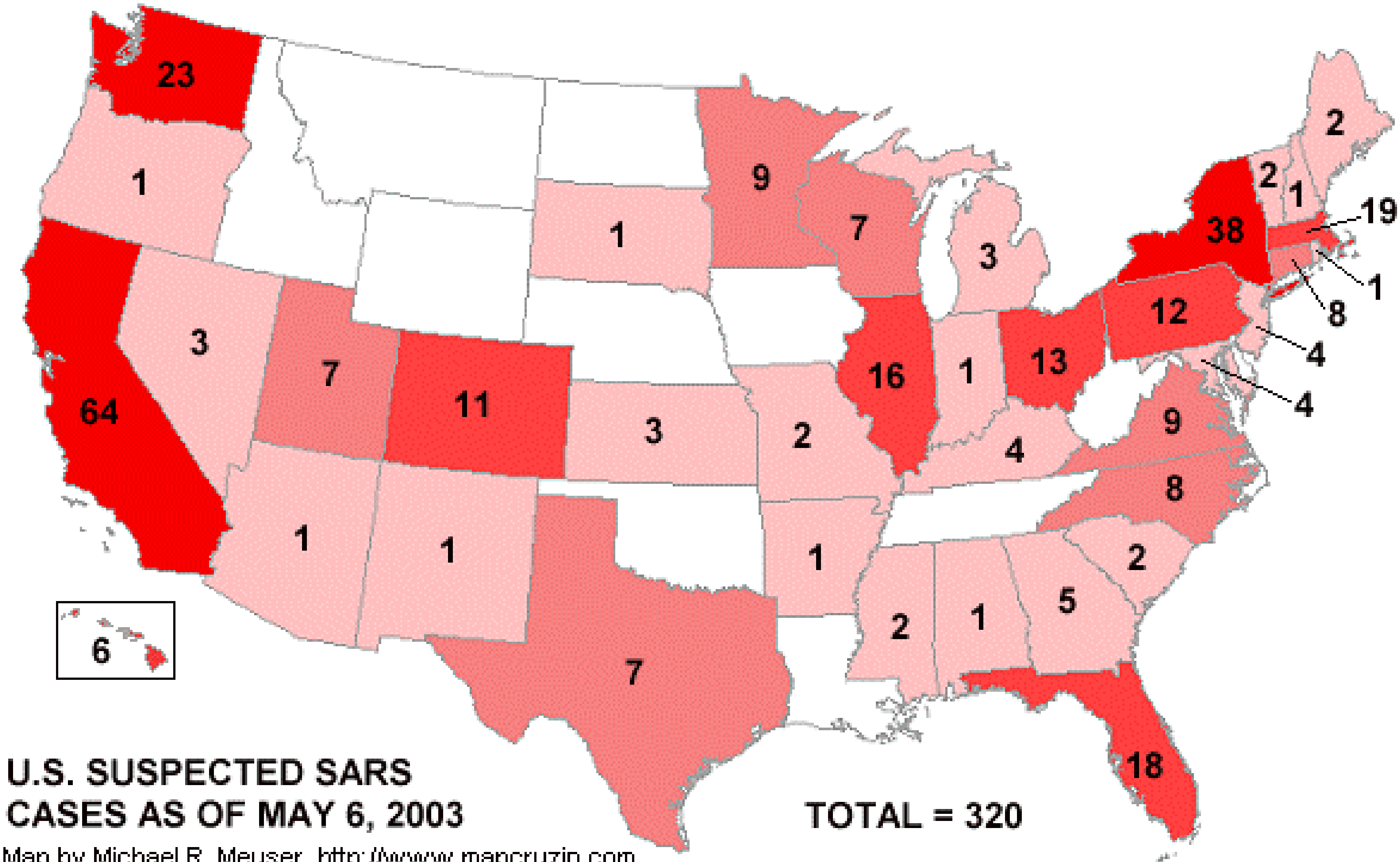


8 - 1882
1883 - 4409
4410 - 6906
6907 - 11111
11112 - 21334

Census tracts by people per square mile.

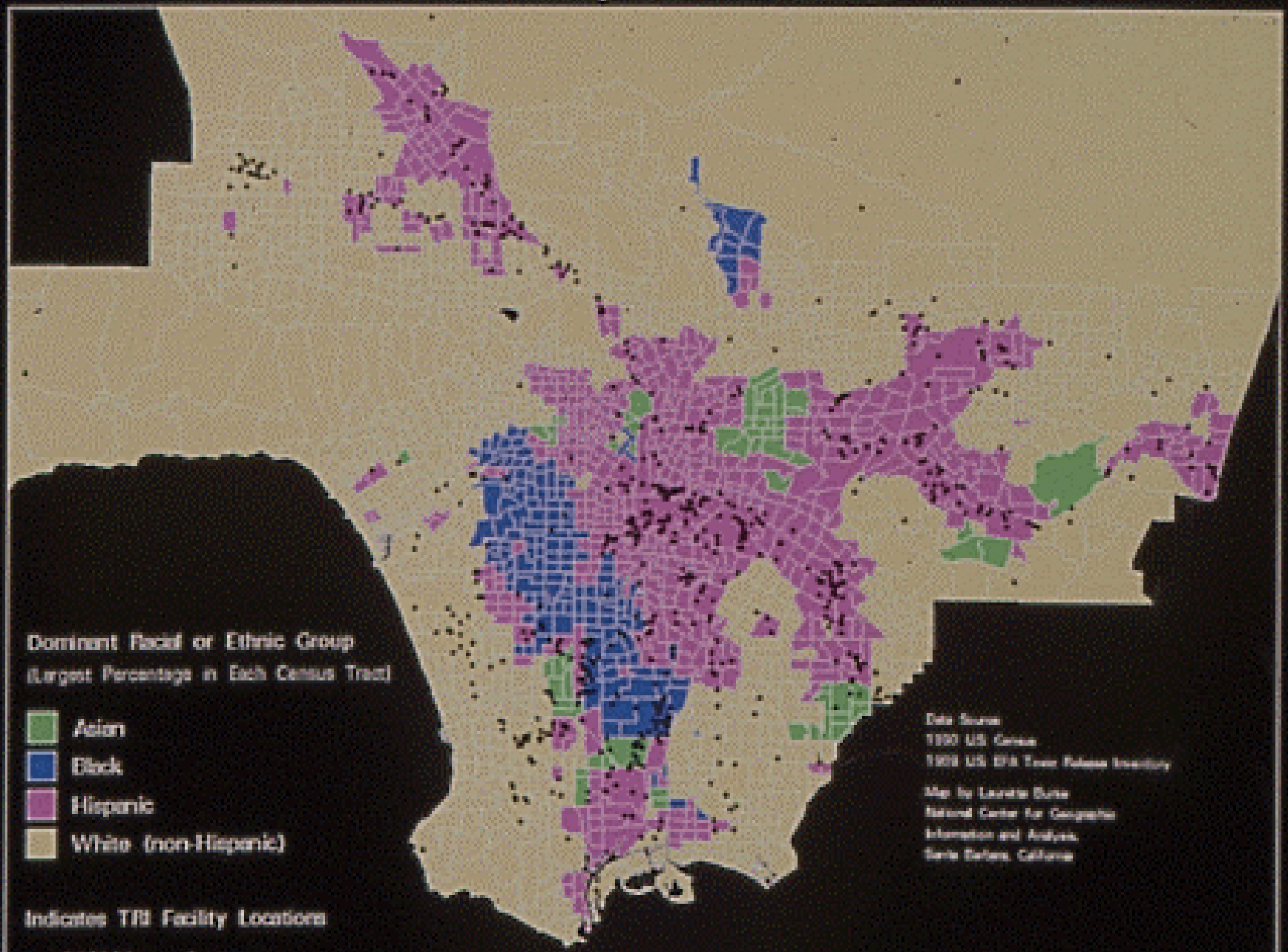
Spatial disconnect

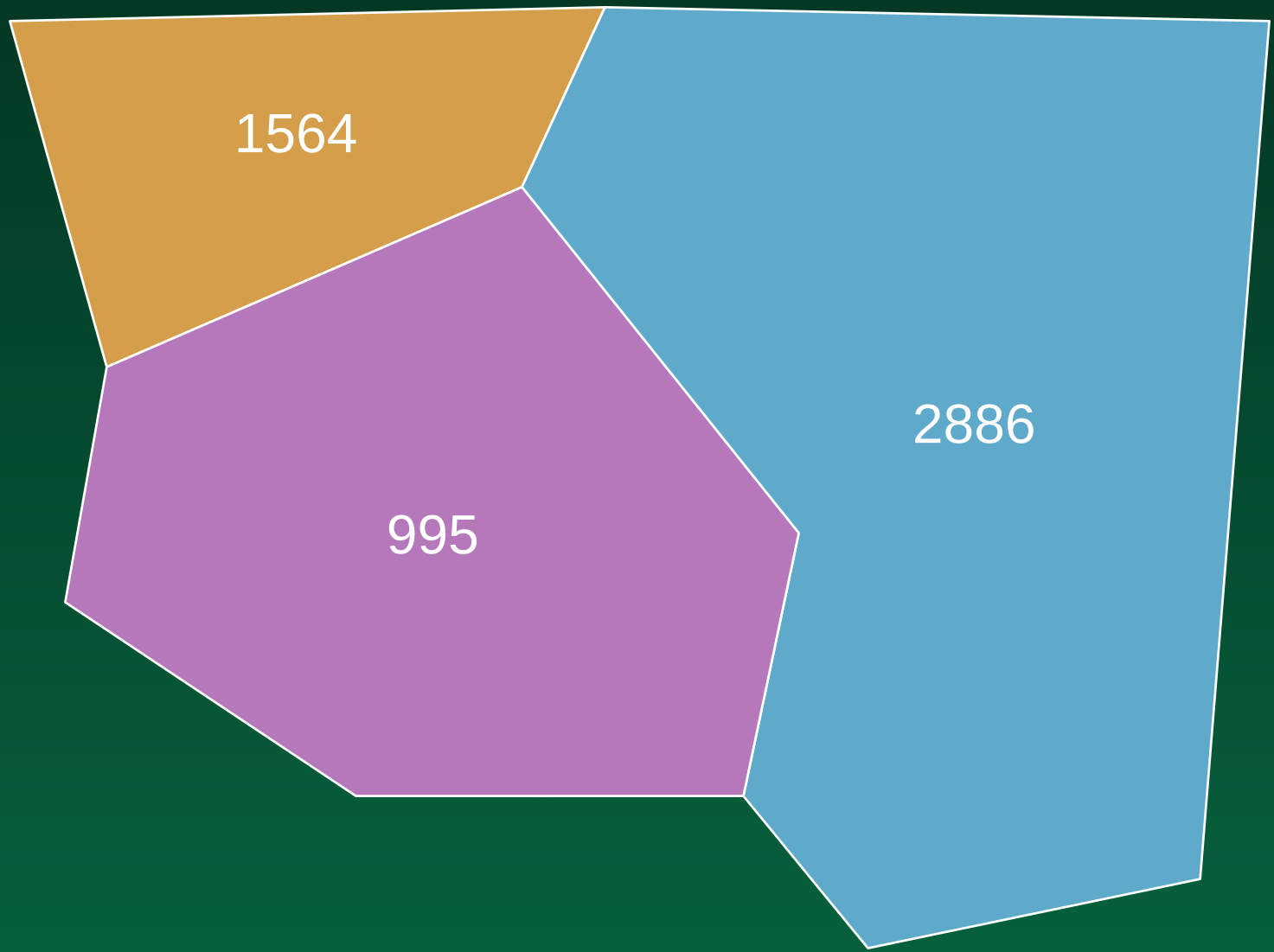
- Many types of data are available only in aggregated form
  - to protect confidentiality
  - for economy
- Basis of aggregation is standardized
  - FIPS
- What if the units of aggregation don't match the units of analysis?



23
1
64
3
7
11
9
1
3
7
16
3
2
1
13
4
2
38
2
1
19
1
8
12
4
4
9
8
2
5
1
2
1
1
1
7
18
6
U.S. SUSPECTED SARS
CASES AS OF MAY 6, 2003
TOTAL = 320
Map by Michael R. Meuser, http://www.mapcruzin.com

Race, Ethnicity and TRI Facilities



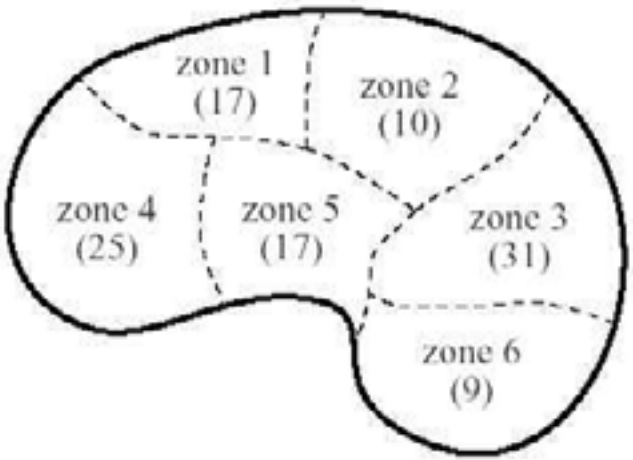

1564
2886
995
1990

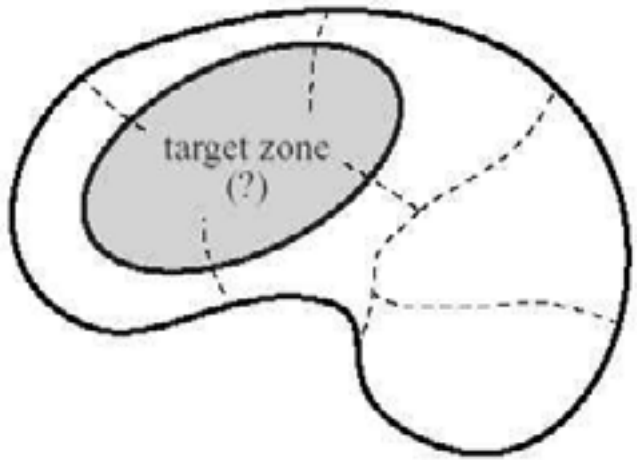
Scale issues

- Ecological fallacy
  - cannot infer individual behavior from data on aggregates
  - model-based solutions
- Modifiable Areal Unit Problem
- Areal interpolation
  - given attributes for source zones, estimate attributes for target zones

The Modifiable Areal Unit Problem

- Openshaw and Taylor
  - 99 counties of Iowa
  - % Republican voters, % over 65
- 48 regions: -.548 to +.886
- 12 regions: -.936 to +.996
- Solutions:
  - manipulate to determine range
  - strengthen theoretical framework



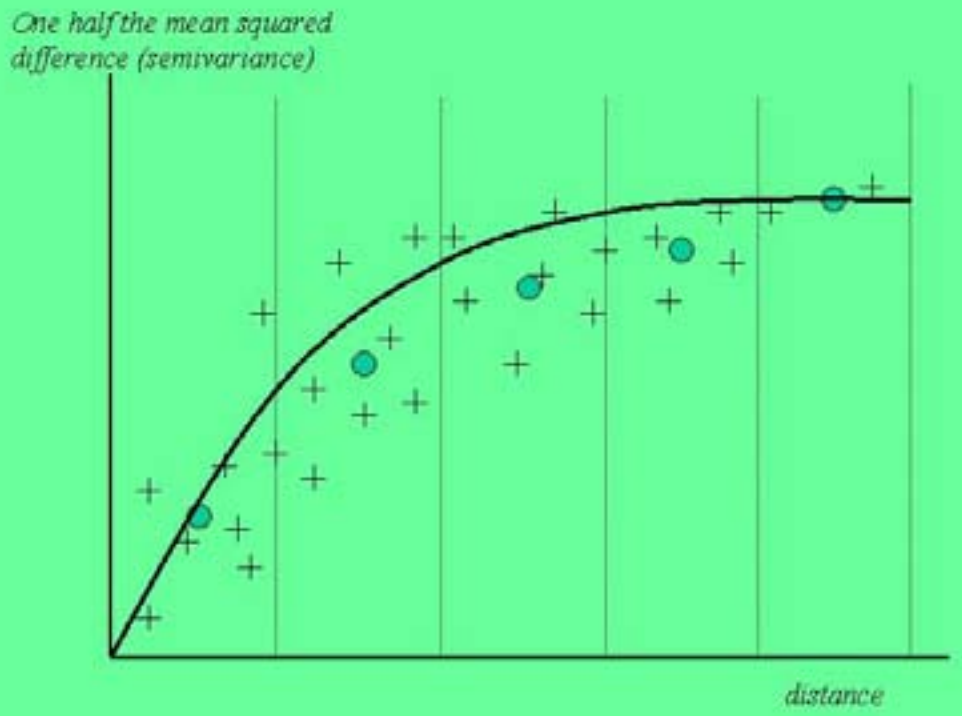


Source: Sadahiro

http://www.csis.u-tokyo.ac.jp/dp/9.pdf

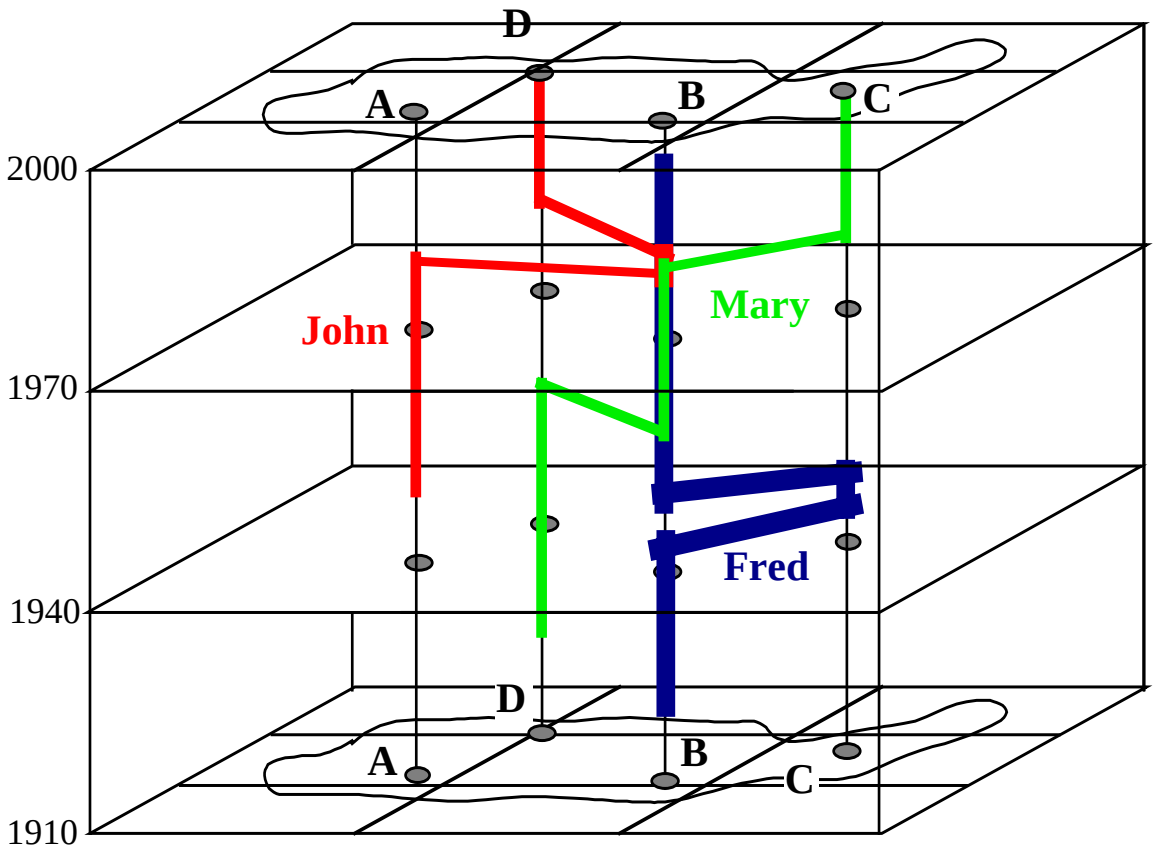
Alternative bases

- Assumptions about $p(x,y)$, the underlying field of population density
  1. constant within source zones
  2. constant within target zones
  3. constant within control zones
  4. maximally smooth (Tobler)
  5. known variogram (Kyriakidis)

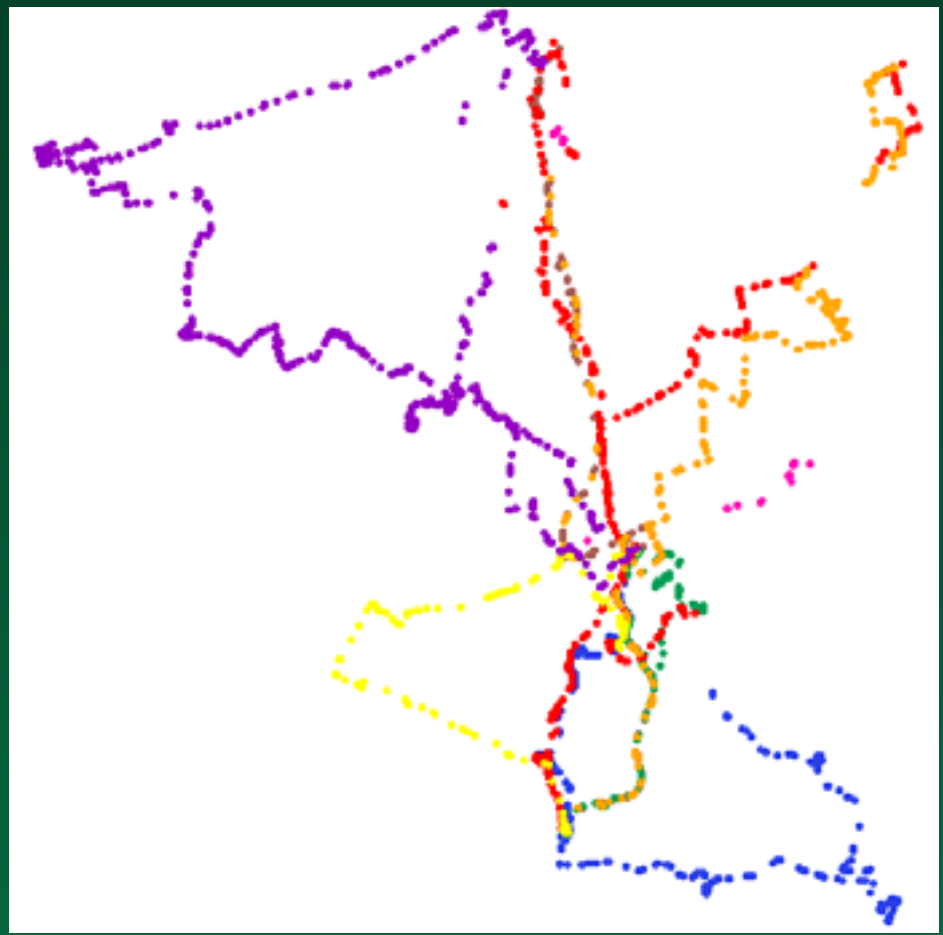

One half the mean squared
difference (semivariance)
distance

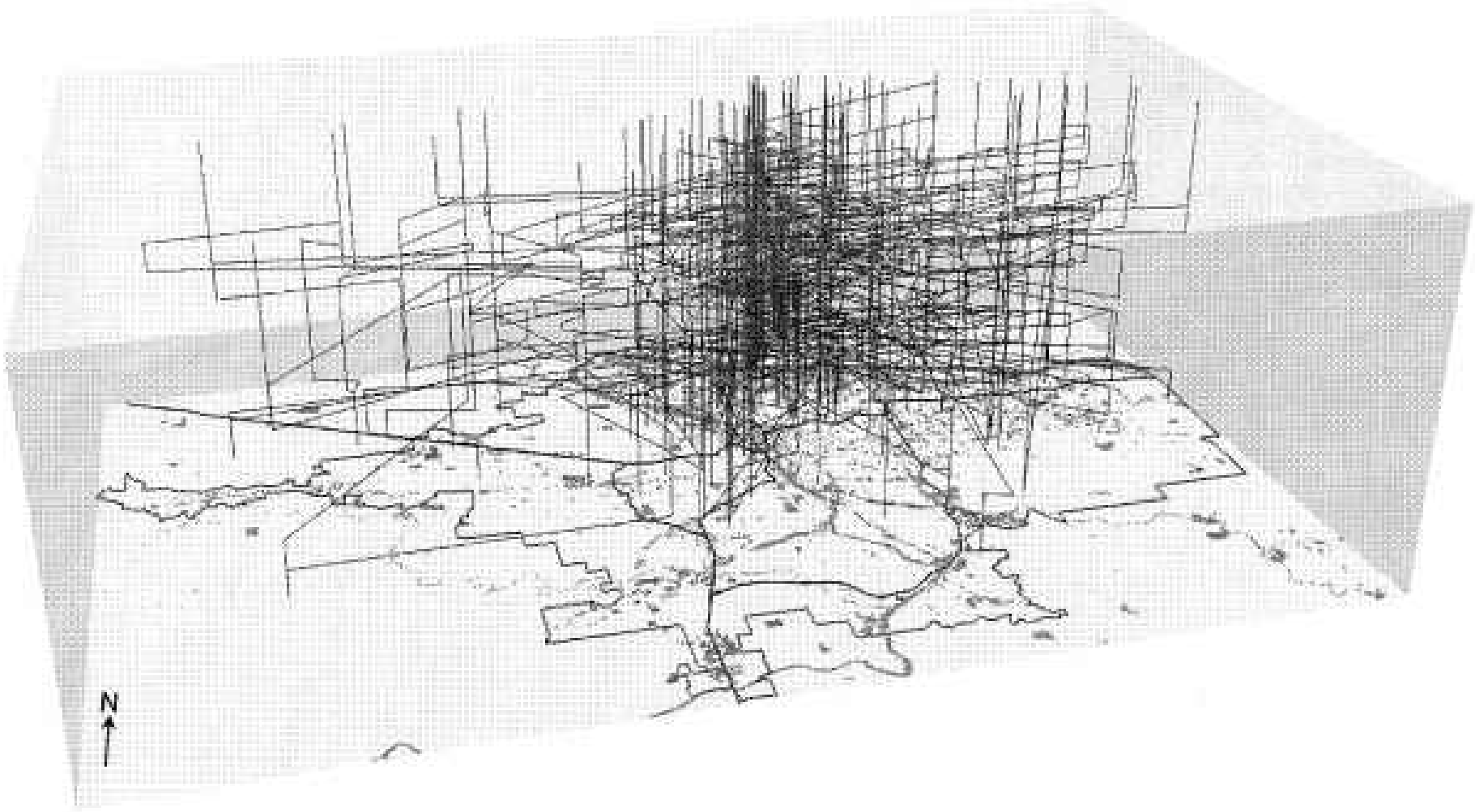
Tracking data

- Data on the locations (and activities) of organisms or people through time
  - diurnal (work/home)
  - annual (travel)
  - lifetime (migration)


D
A
B
C
2000
John
Mary
1970
Fred
1940
D
A
B
C
1910

A week in Jonathan Raper's life




N

Probes

- 20 million US vehicles now equipped with GPS
  - and many 3G cellphones
- Inference from tracks
  - potholes
  - emergencies
- Modeling and simulation of exposure

Simulations

- 1.8 vehicles per driveway
- Driver behavior influenced by:
  - lane width
  - slope
  - view distances
  - traffic control mechanisms
  - information feedback
  - driver aggressiveness
- 770 homes
  - clearing times > 30 minutes

2D clip

3D clip

Analysis of tracks

- Visualization
- Conflation with risk fields
- Models
  - random tracks
  - track density

Modeling risk fields

- Interpolation from point samples
  - CA ozone
- Plume models
  - atmospheric
  - subsurface
- Spatial interaction models
  - negative exponential decline from source

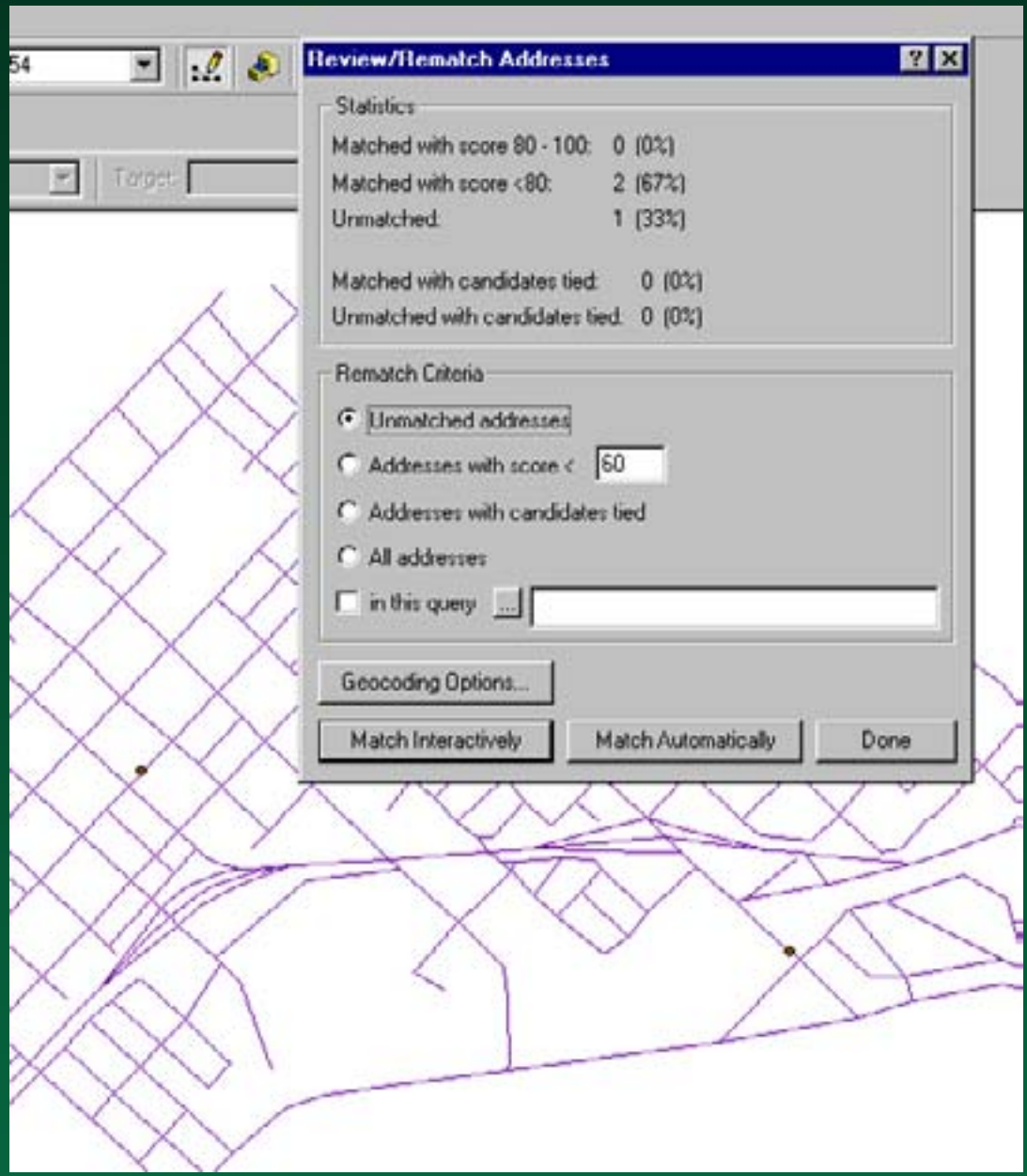
Data sources

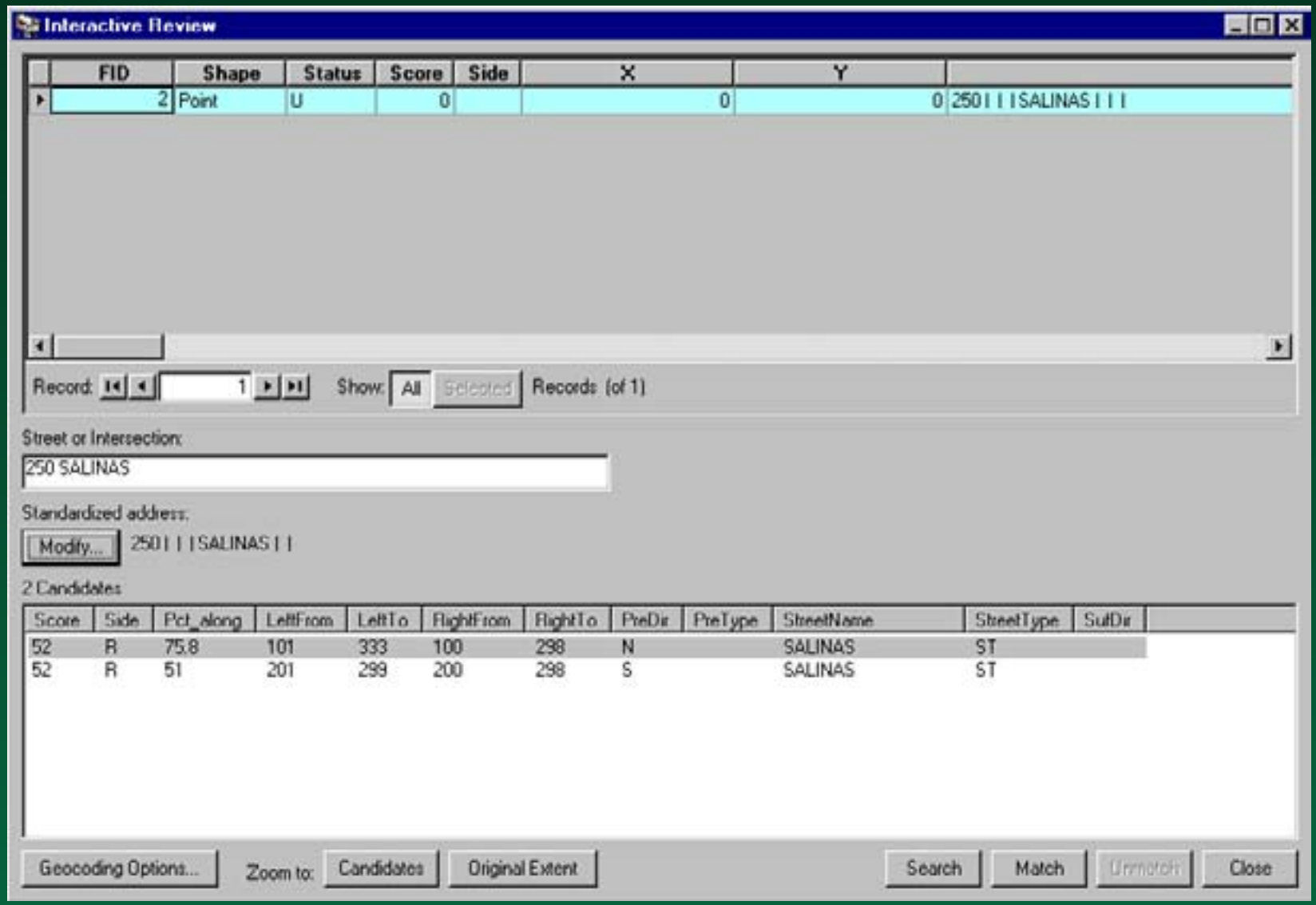
- Framework data
  - data used for georeferencing
  - to which other data can be added
  - quality control by public agencies
- The framework layers
  - street centerlines
  - topography
  - geodetic control
  - boundaries
  - imagery
  - hydrography
  - cultural features

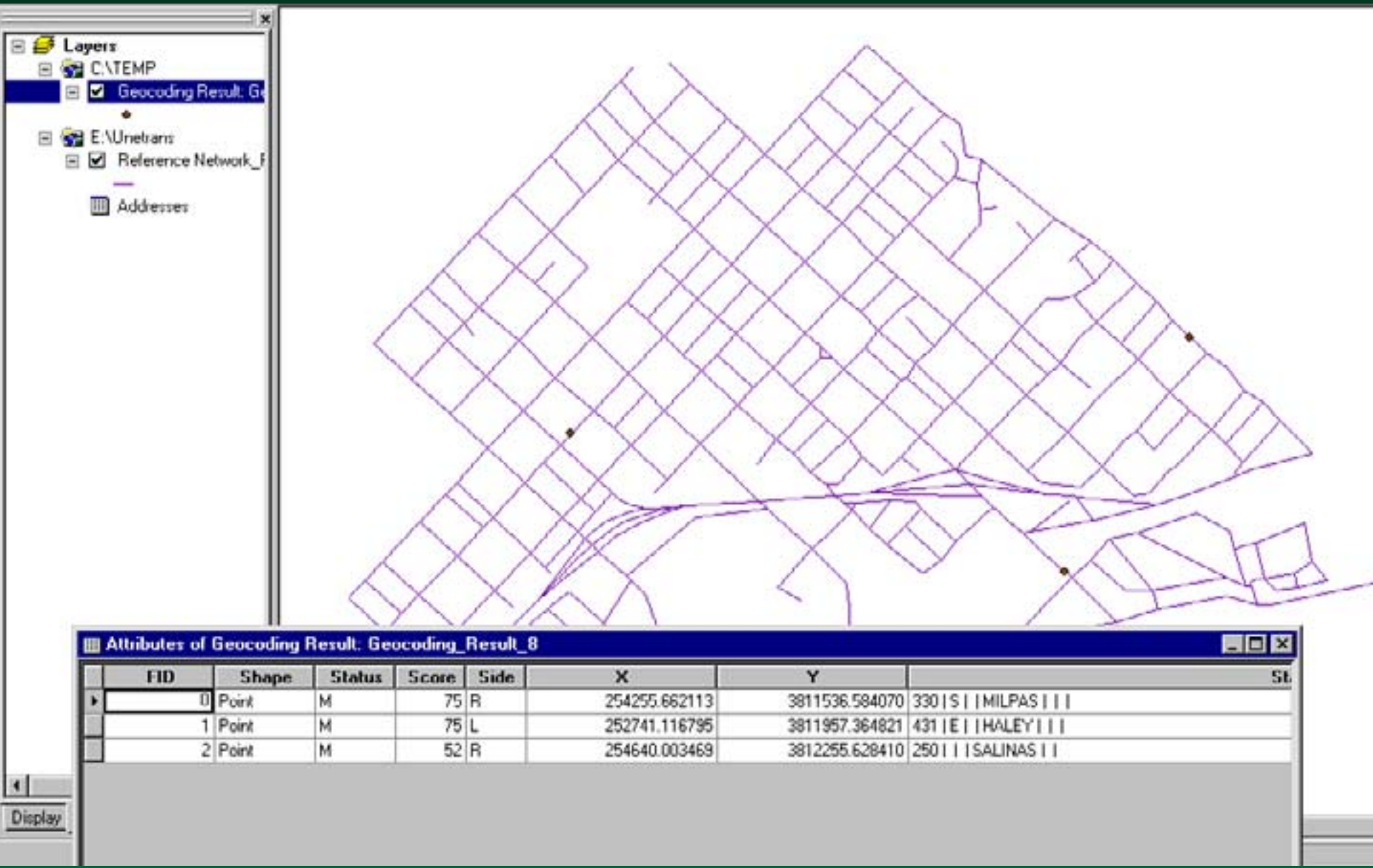
Geocoding

- Conversion of street addresses to coordinates
  - requires a street centerline database with address ranges
  - automatic for ~80% of addresses
  - the remaining 20% pathological


Attributes of Addresses
OID
Street_nam
0
330 S MILPAS
1
431 E HALEY
2
250 SALINAS
Record:
1
Show:
All
Selected
Records (0 out of 3 Selected.)
Options


Review/Rematch Addresses
Statistics
Matched with score 80 - 100: 0 (0%)
Matched with score <80: 2 (67%)
Unmatched: 1 (33%)
Matched with candidates tied: 0 (0%)
Unmatched with candidates tied: 0 (0%)
Rematch Criteria
Unmatched addresses
Addresses with score < 60
Addresses with candidates tied
All addresses
in this query
Geocoding Options...
Match Interactively
Match Automatically
Done


Interactive Review
FID
Shape
Status
Score
Side
X
Y
2
Point
U
0
0
0
250 | | | SALINAS | | |
Record:
1
Show:
All
Selected
Records (of 1)
Street or Intersection:
250 SALINAS
Standardized address:
Modify...
250 | | | SALINAS | |
2 Candidates
Score
Side
Pct_along
LeftFrom
LeftTo
RightFrom
RightTo
PreDir
PreType
StreetName
StreetType
SufDir
52
R
75.8
101
333
100
298
N
SALINAS
ST
52
R
51
201
299
200
298
S
SALINAS
ST
Geocoding Options...
Zoom to:
Candidates
Original Extent
Search
Match
Unmatch
Close


Layers
C:\TEMP
Geocoding Result: G
E:\Unetrans
Reference Network_F
Addresses
Display
Attributes of Geocoding Result: Geocoding_Result_8
FID | Shape | Status | Score | Side | X | Y | St
0 | Point | M | 75 | R | 254255.662113 | 3811536.584070 | 330 | S | | MILPAS | | |
1 | Point | M | 75 | L | 252741.116795 | 3811957.364821 | 431 | E | | HALEY | | |
2 | Point | M | 52 | R | 254640.003469 | 3812255.628410 | 250 | | | SALINAS | |


display.june.stp - Display
File View Map GPS Comm Window Help
A B C D E F G H I
Ready
Time: 1.00
241697.0
3814757.1
Zone 11
no gps
no gps
CLARENCE WARD MEMC
0 - 0
0 - 0
Start
fwy-ce.bmp - Paint
C:\Vital
display.june.stp - Dis...
13:19

Online data sources

- www.geographynetwork.com
- FGDC National Geospatial Data Clearinghouse
  - www.fgdc.gov
- USGS EROS Data Center
- State and local data warehouses
  - data or maps

Evolving trends in GIS software

- The georelational model
  - related tables
- Object-oriented modeling
  - objects as instances of general classes
  - classes as specializations of more general classes (inheritance)
  - methods associated with classes (encapsulation)
  - associations between objects

Specialized GIS data models

- The basic elements built into the GIS
  - points, lines, areas
- How these elements are specialized in application domains (e.g. health risk perception)
  - track as a class of line
  - disease instance as a class of point

Unified Modeling Language

- Visual representation of a data model
  - conventional symbols
  - implemented in Visio
- Creation of database layout
  - using CASE tools
  - building tables
  - populate tables with data

ESRI ArcGIS

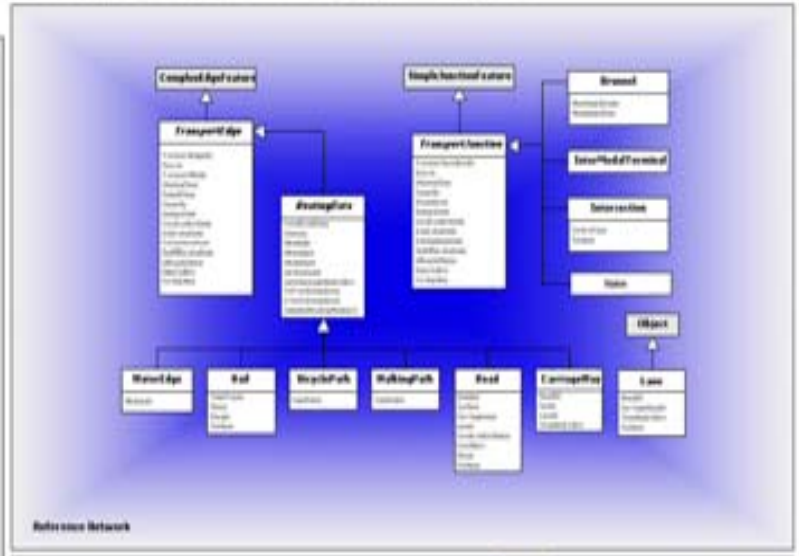
- ArcInfo version 8
- Specialized data models
  - water utilities (ArcFM)
  - hydrology
  - transportation (Unetrans)
  - health data model http://www.ncgia.ucsb.edu/projects/health/

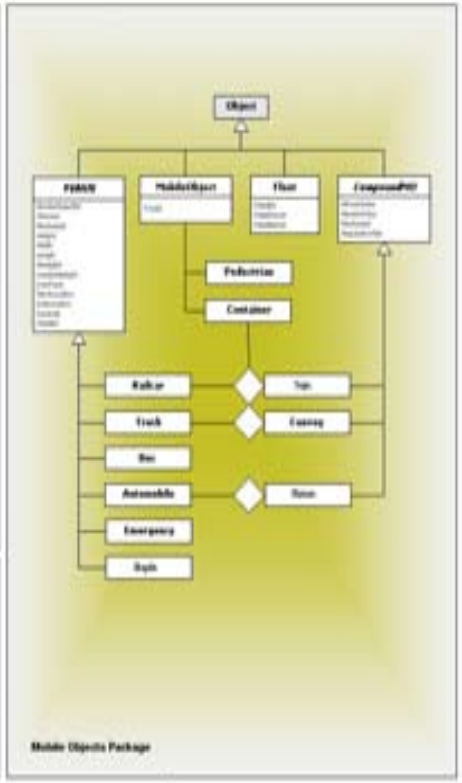
Objective

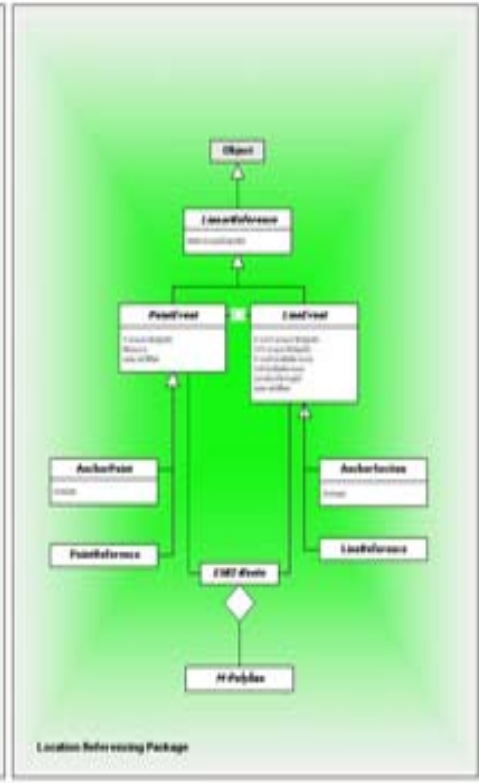
- Helping users by providing a database framework that includes familiar elements
  - contains the core items
  - is easy to extend and specialize
  - add new attributes
  - add specialized classes

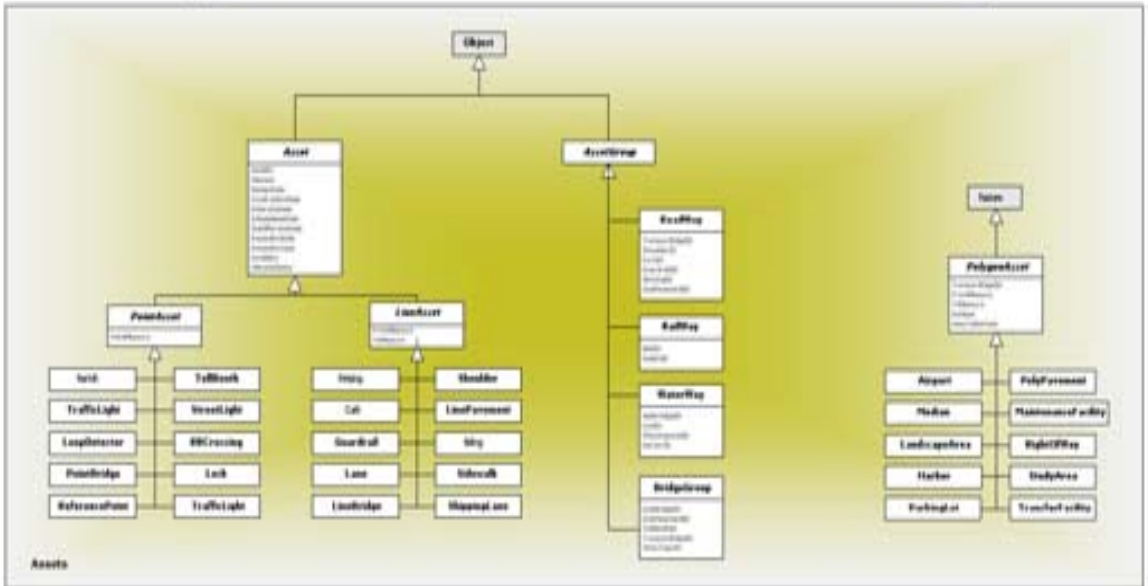
ArcGIS Transportation Data Model

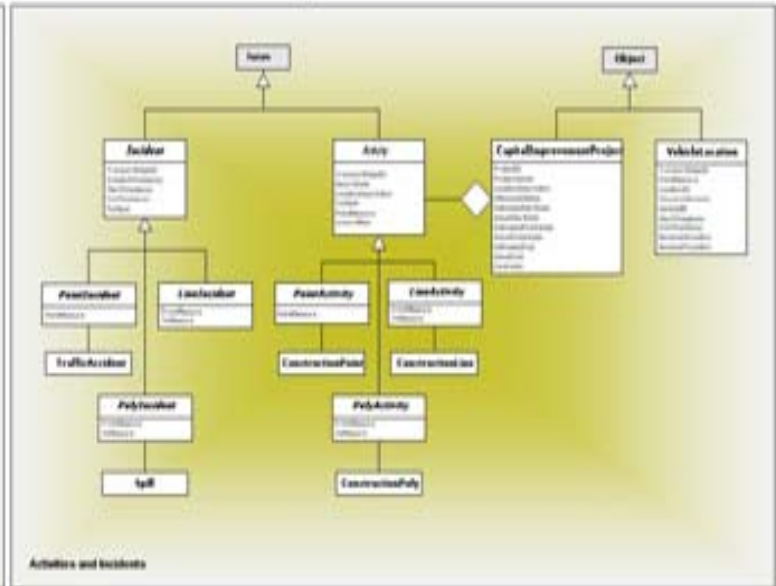
Feb. 21, 2001

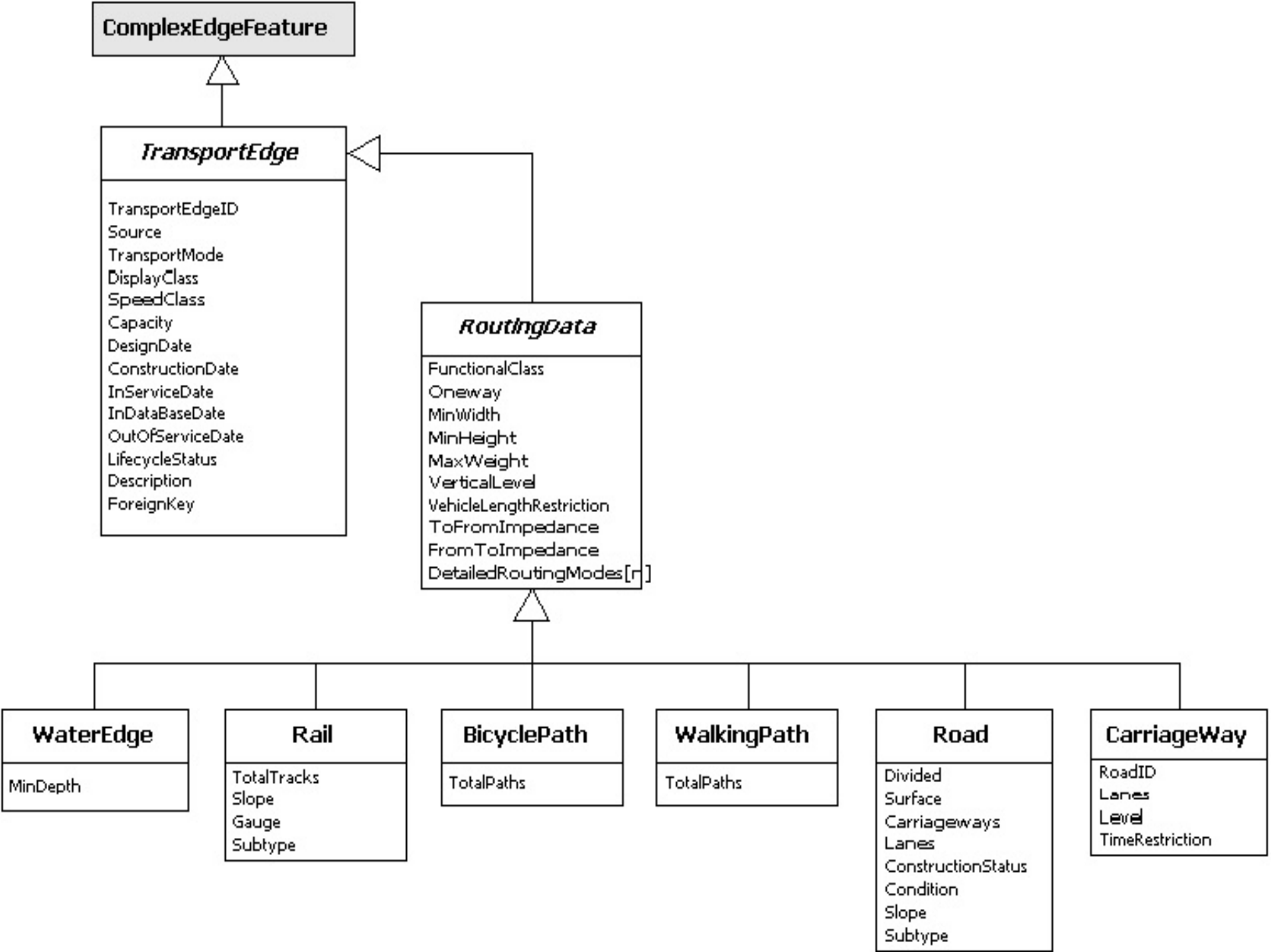


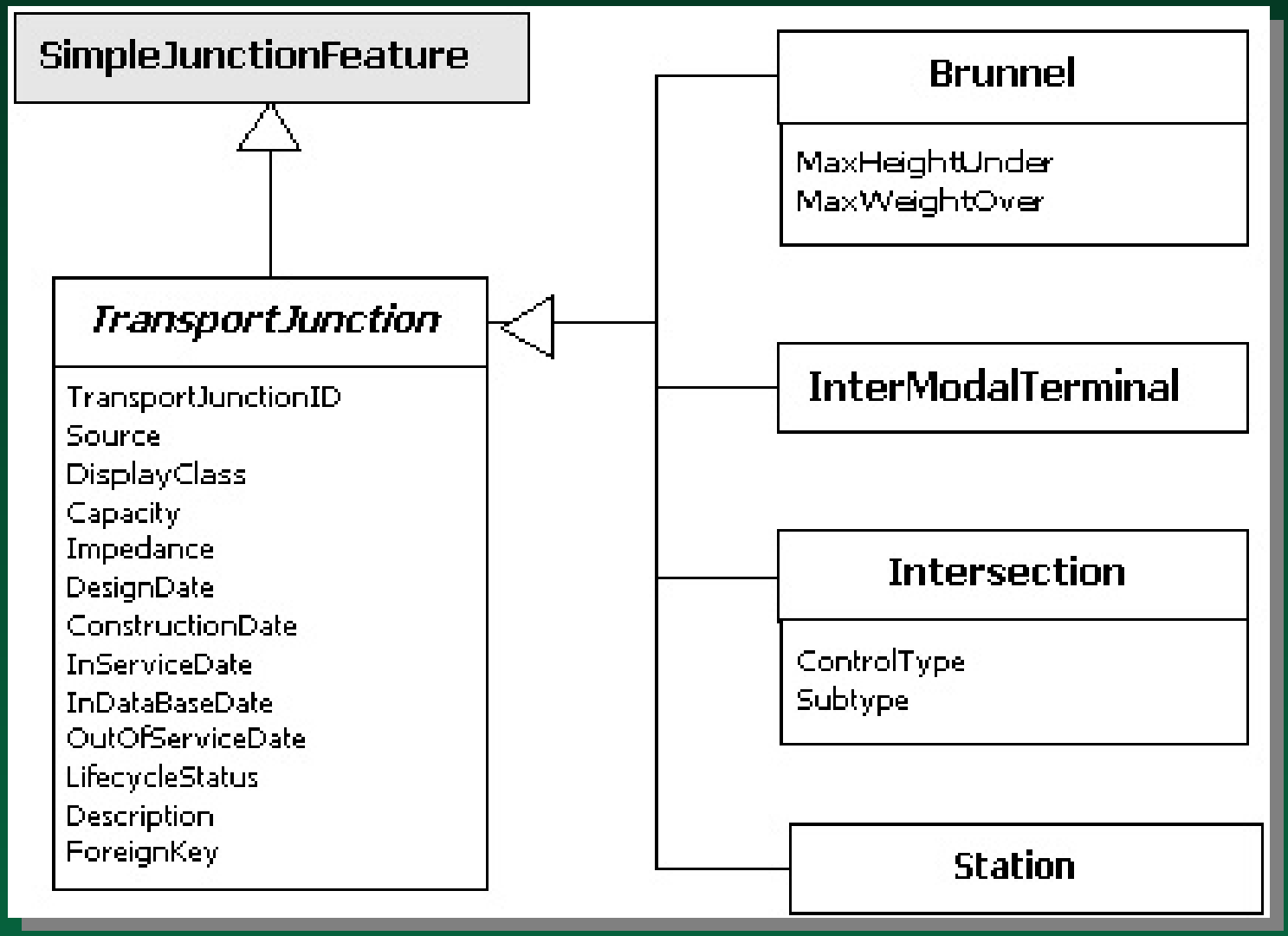






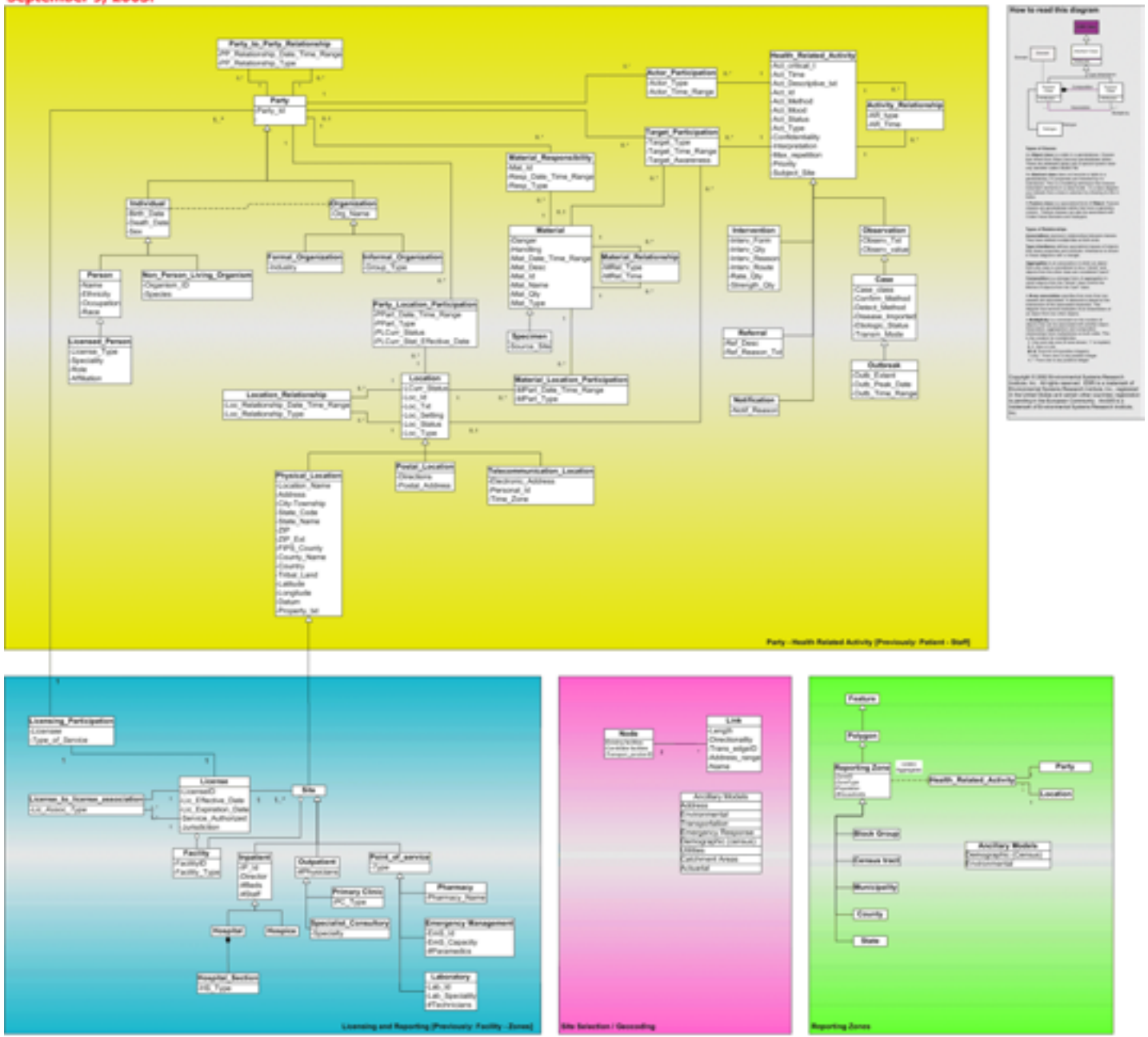



ComplexEdgeFeature
TransportEdge
TransportEdgeID
Source
TransportMode
DisplayClass
SpeedClass
Capacity
DesignDate
ConstructionDate
InServiceDate
InDataBaseDate
OutOfServiceDate
LifecycleStatus
Description
ForeignKey
RoutingData
FunctionalClass
Oneway
MinWidth
MinHeight
MaxWeight
VerticalLevel
VehicleLengthRestriction
ToFromImpedance
FromToImpedance
DetailedRoutingModes[n]
WaterEdge
MinDepth
Rail
TotalTracks
Slope
Gauge
Subtype
BicyclePath
TotalPaths
WalkingPath
TotalPaths
Road
Divided
Surface
Carriageways
Lanes
ConstructionStatus
Condition
Slope
Subtype
CarriageWay
RoadID
Lanes
Level
TimeRestriction


SimpleJunctionFeature
TransportJunction
TransportJunctionID
Source
DisplayClass
Capacity
Impedance
DesignDate
ConstructionDate
InServiceDate
InDataBaseDate
OutOfServiceDate
LifecycleStatus
Description
ForeignKey
Brunnel
MaxHeightUnder
MaxWeightOver
InterModalTerminal
Intersection
ControlType
Subtype
Station

Health Data Model

September 9, 2003.



What are the limitations?

- Dominated by COTS software
  - priorities defined by largest markets
  - research tools often built as extensions
  - limited infrastructure for sharing
    - http://arcscripts.esri.com/
  - ArcGIS 9.0 release
  - COM integration
- Weak representation of time
  - cartographic legacy
  - limited data, theory, models, tools

More limitations

- Lack of lifetime-scale tracking data
  - space-time life histories
- Lack of comprehensive data on risk


your location
909 W Campus Ln, Goleta, CA 93117
Distance from the nearest nuclear waste route: 1.5 miles
Distance from Diablo Canyon, the nearest waste source: 78.7 miles
[more]
Isla Vista
Pacific Ocean
0 1mi
Copyright 2003 by EWG, ESRI, GDT
DISTANCE TO A PROPOSED NUCLEAR WASTE ROUTE
within 1 mile
within 2 miles
within 5 miles
Hospitals
Proposed Route
Highway
Rail
Schools