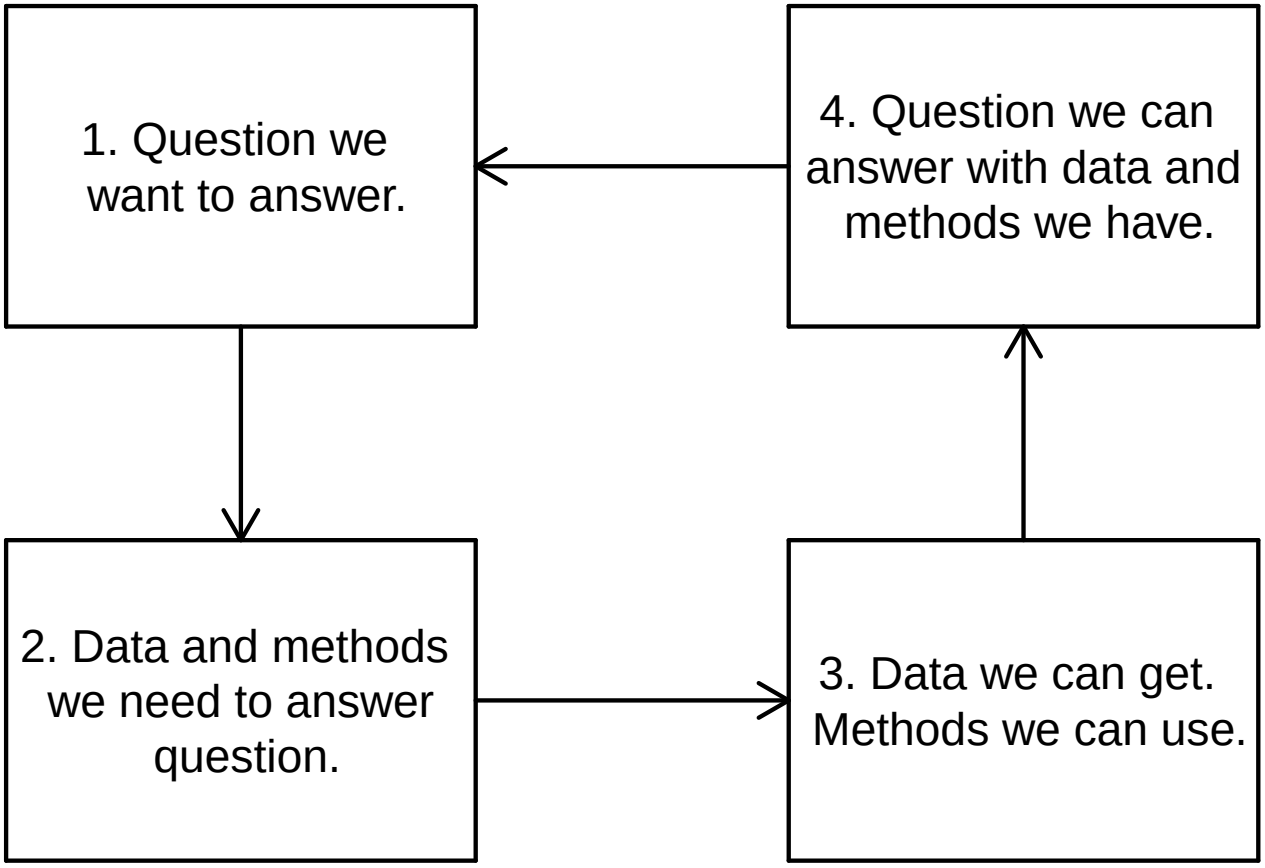
MODULE 16: Spatial Statistics in Epidemiology and Public Health
Lecture 2: Spatial Questions and Answers

Jon Wakefield and **Lance Waller**

How can maps help us with spatial statistics?

- Spatial questions require:
  - Spatial data
  - Spatial methods
  - Spatial answers
- Maps frame questions, data, methods, answers in a spatial setting

The whirling vortex



Maps hold clues...

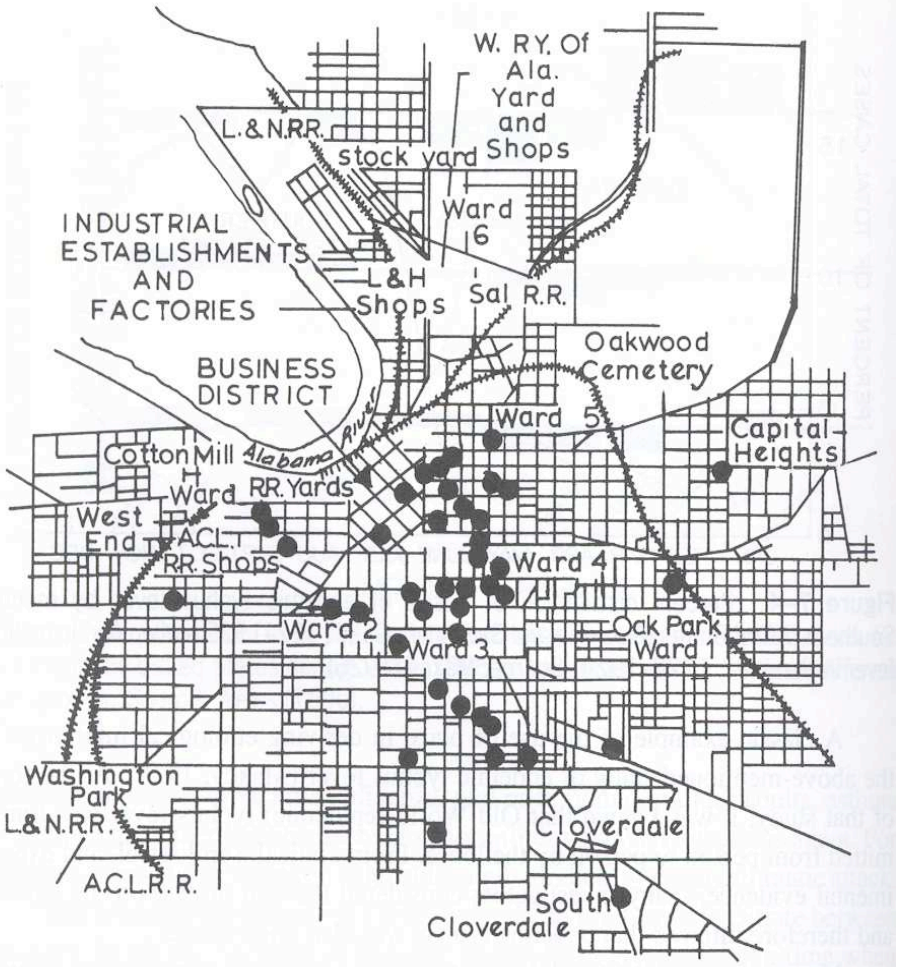


Figure 7–5. Distribution of cases of endemic typhus fever by residence, Montgomery, Alabama, 1922–1925. *Source:* Maxcy (1926).

...but may not reveal them immediately.

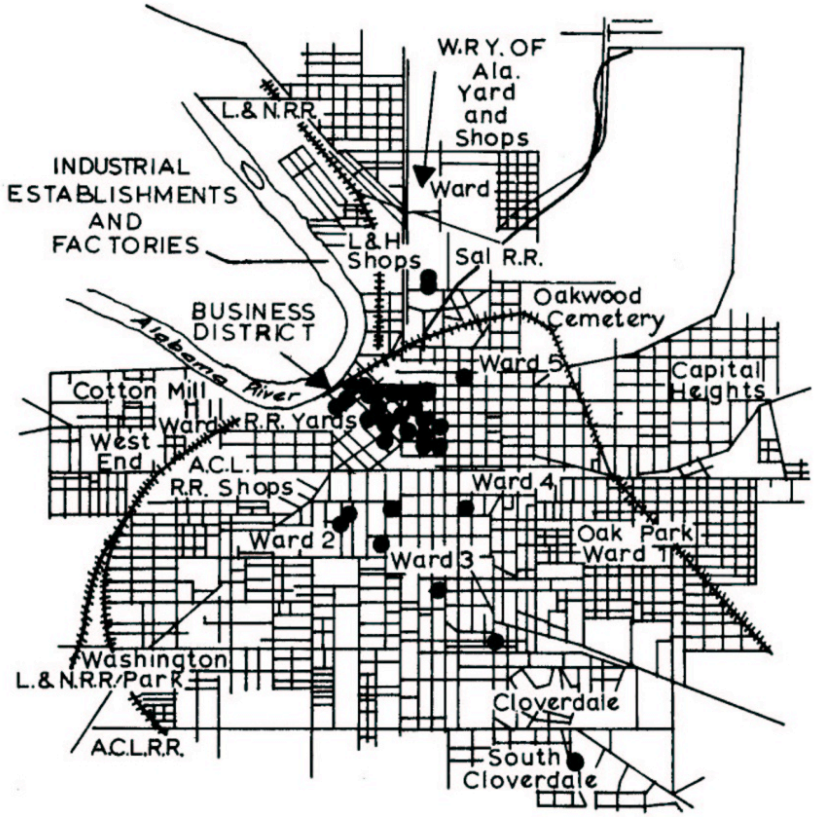
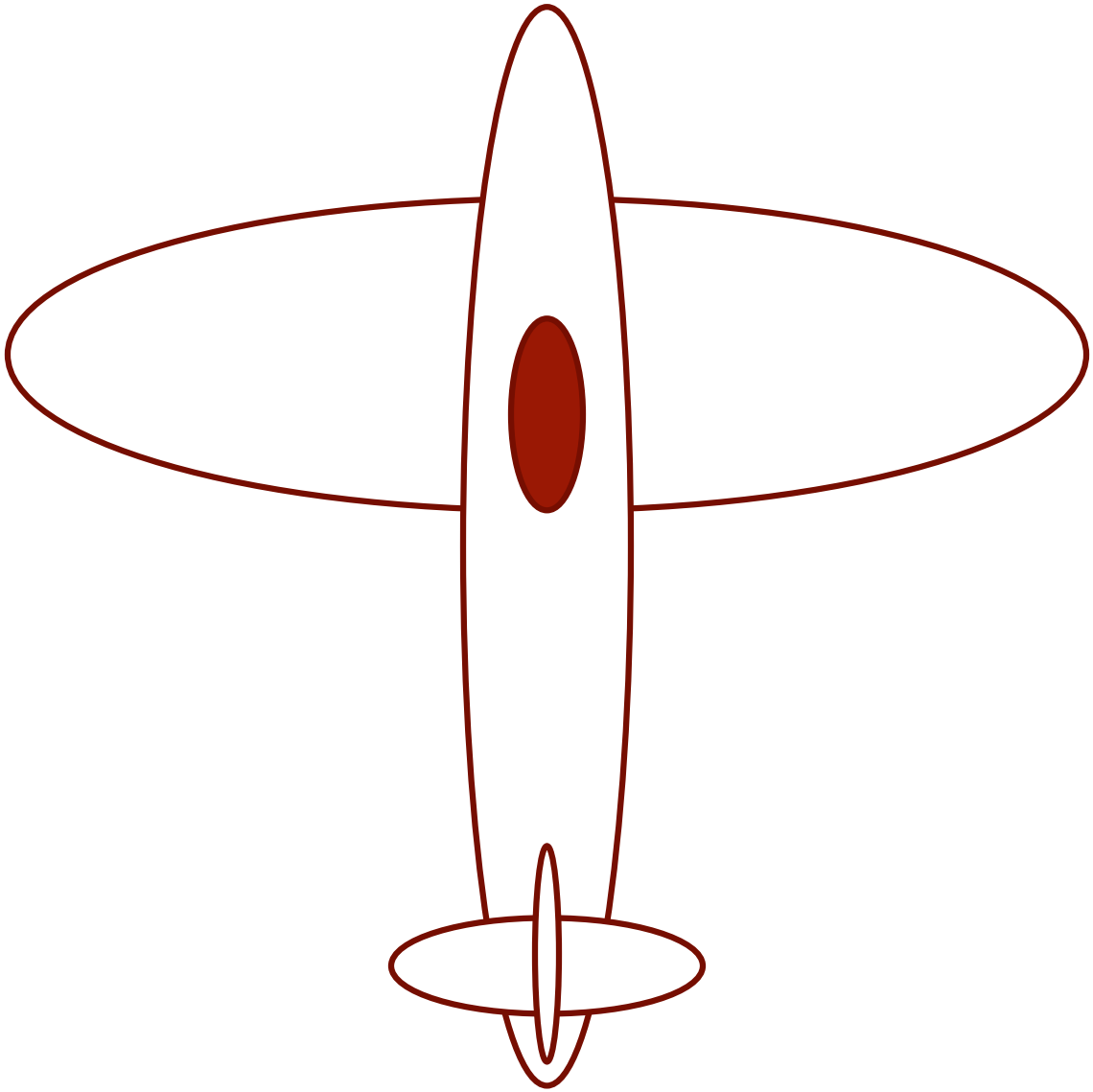


Figure 7–6. Distribution of cases of endemic typhus fever by place of employment or, if unemployed, by place of residence, Montgomery, Alabama, 1922–1925. *Source:* Maxcy (1926).

- Lillenfeld and Stolley (1994, *Foundations of Epidemiology, 3rd Ed.*. Oxford pp. 136-140).

Patterns and conclusions

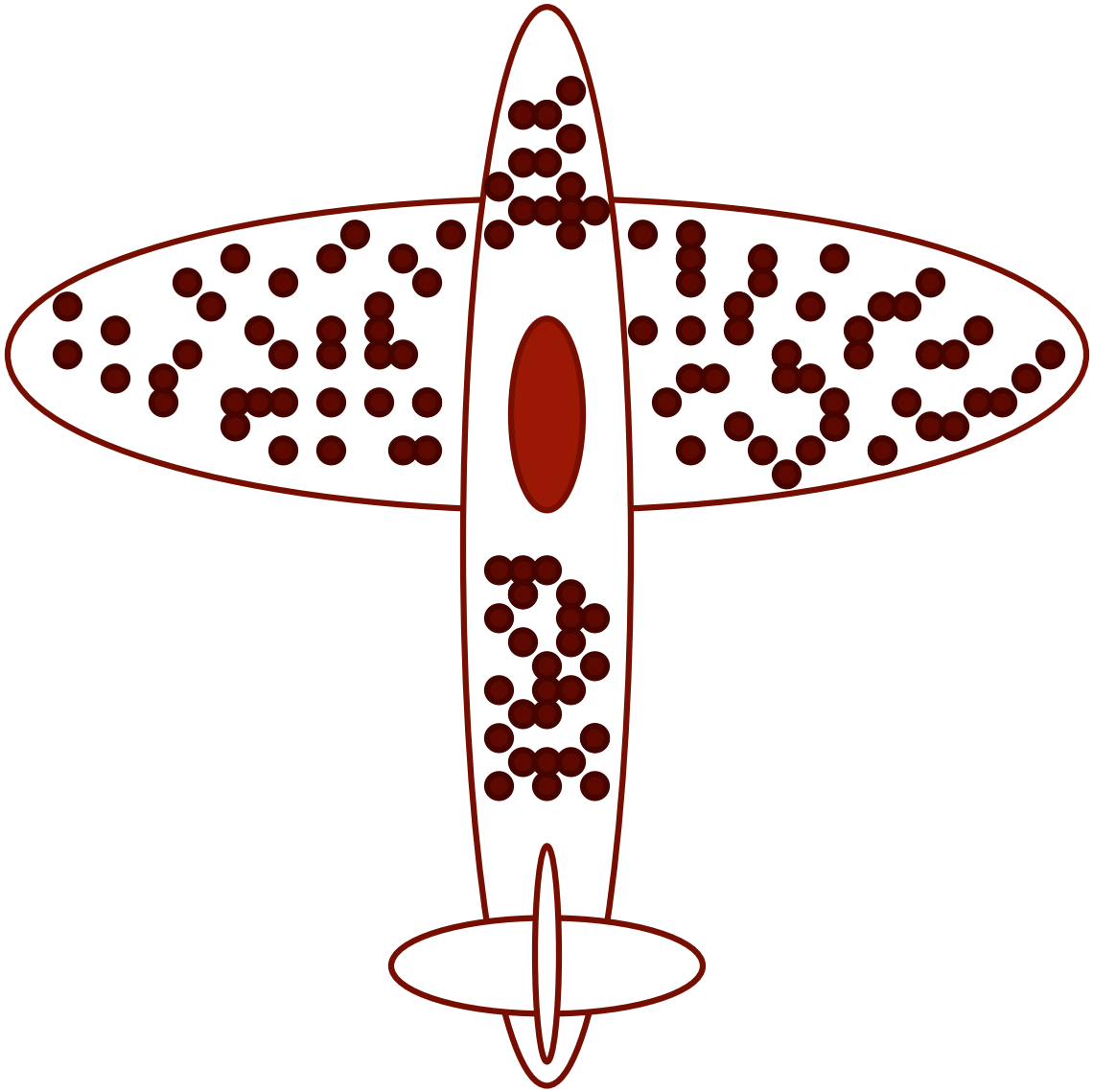


Patterns and concluions

- Abraham Wald, WWII, Pacific airbase.
- Where to put (limited amount of) extra armor?
- Outline of plane placed on hanger wall.
- Add a dot for observed damage on returning planes.

- Wainer (1997, Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot, Springer).

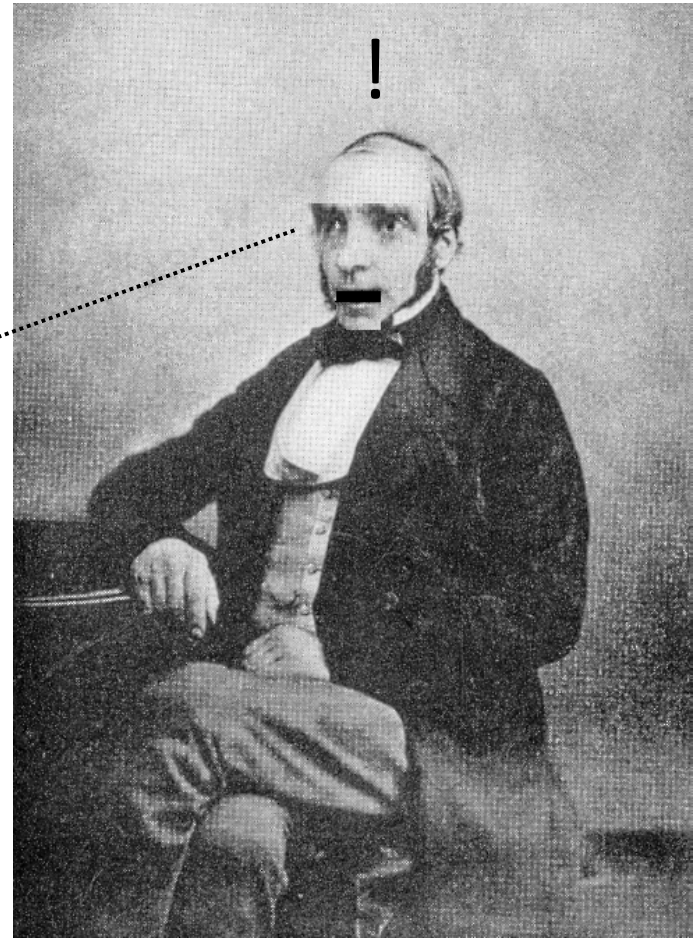
Patterns and conclusions



Maps and Health

John Snow

Snow, J. (1949) *Snow on Cholera.* Oxford University Press: London.



Short version of the story

- "In 1854, Londoners were dropping like flies from cholera until Dr. Snow figured out that the bacteria were carried by water. The water pump he turned off, thereby saving countless lives, was near the site of this pub."
- John Snow Pub entry in *Access London* tour guide, Harper-Collins, 2005.

Truth a little more complicated

- Brody et al. (2000) Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *Lancet*
- Koch (2005) *Cartographies of Disease: Maps, Mapping, and Medicine*. ESRI Press
- Johnson (2006) *The Ghost Map*. Riverhead Books

Edmund Cooper's map

- Map for Metropolitan Commission of Sewers, September 1854
- (Snow's map in December)
- In response to public concern that sewer works had disturbed an ancient pit containing bodies from the plague of 1665.
- Theory: Cholera clustered near gully holes.
- Map revealed this was not the case.

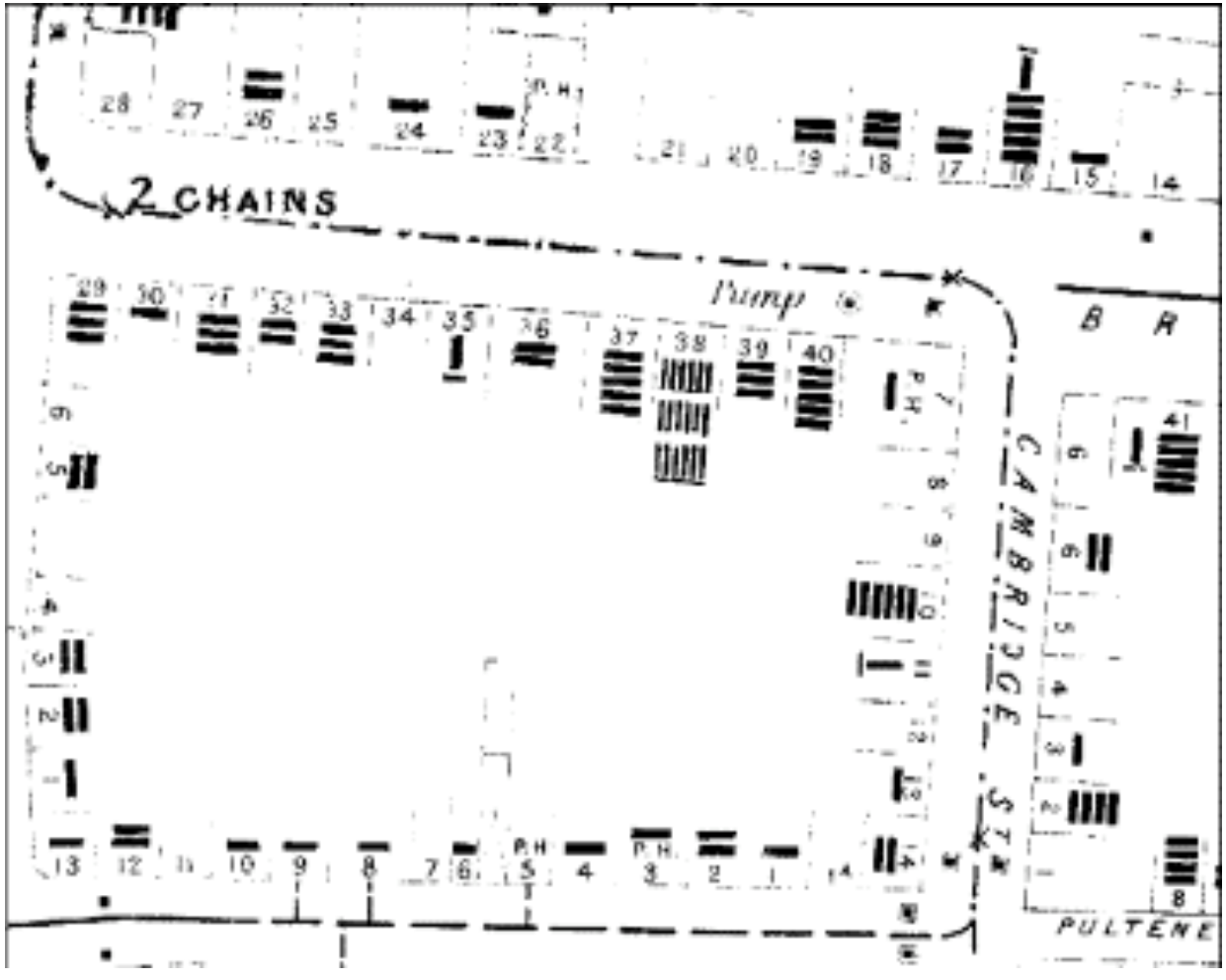
Cooper's map


28
27
26
25
24
25
P.H.
22
21
20
19
18
17
16
15
14
Temple
BROA
CAMBRIDGE ST.
TENT ST.
29
30
31
32
33
34
35
36
37
38
39
40
P.H.
7
8
9
10
11
12
13
14
6
5
4
3
2
1
13
12
11
10
9
8
7
6
P.H.
5
4
P.H.
3
2
1
PULTE
21

Board of Health map

- Includes additional cases.
- Pretty obvious, isn't it?
- Remove the handle!!

Board of Health map



Well...

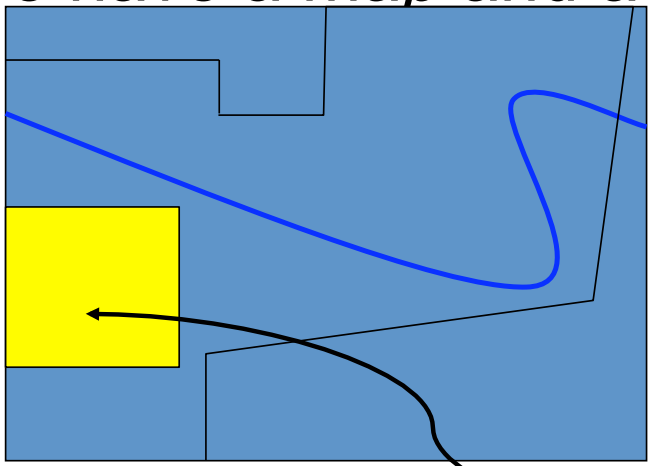
- "This certainly looks more like the effect of an atmospheric cause than any other; if it were owing to the water, why should not the choler have prevailed equally everywhere where the water was drunk? (Parkes, 1855).
- Same data, similar map, different conclusions.
- What do we learn from this?
  - A map is not enough, we need to understand the spatial nature of the data, apply spatial methods, get spatial answers.

What can we do with a map?

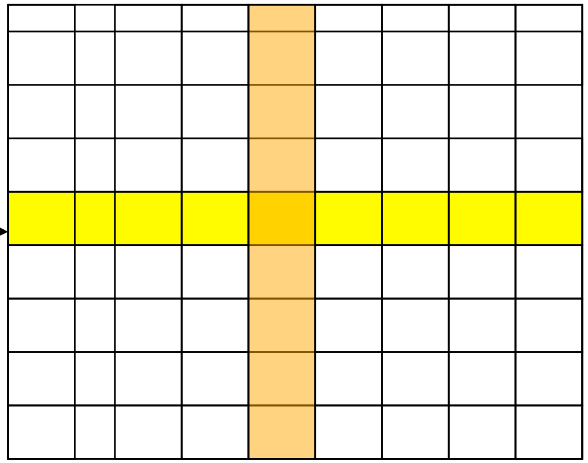
- Merriam-Webster online: Map = "a *representation* usually on a flat surface of the whole or part of an area."
- Note "representation" means "not an exact duplicate"!
- *Thematic* maps include *locations* and *attributes* associated with the locations.
- Think of a *map* of locations linked to a *table* of attribute values.

Maps and tables

We have a *map* and a *table.*



A column is associated with a particular attribute.



a row in the table corresponds to the collection of attribute values for a particular geographic feature in the map.

Geographic Information Systems (GIS)

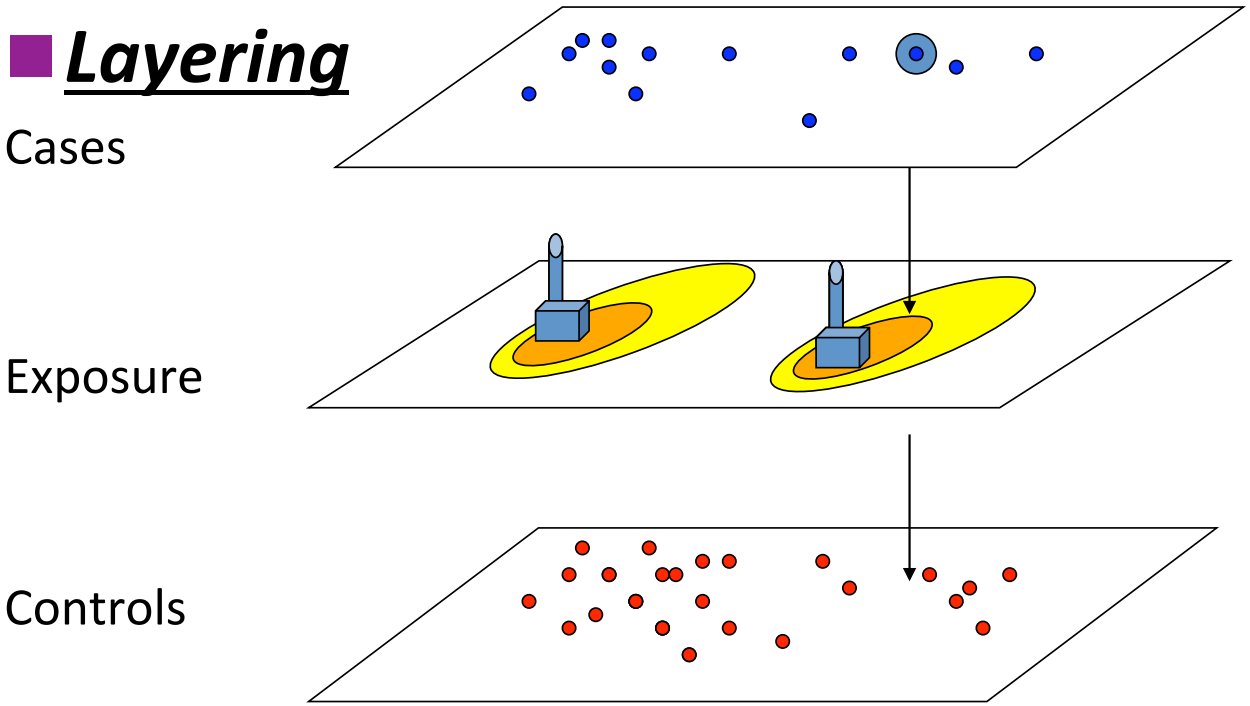
- A *geographic information system* is "a technology designed to capture, store, manipulate, analyze, and visualize georeferenced data" (Goodchild, Parks, and Steyaert 1993).
- GIS is a database system containing locations for every value and allowing operations (search, sorting, etc.) based on locations as well as attributes.
- Allows maps of attribute values.

What does a GIS do?

- Think of data sets as "layers".
- For example:
  - One layer of case locations (points).
  - One layer of road locations (lines).
  - One layer of population levels (areas).
  - One layer of vegetation type (satellite image (raster)).

Basic GIS operation 1: Layering

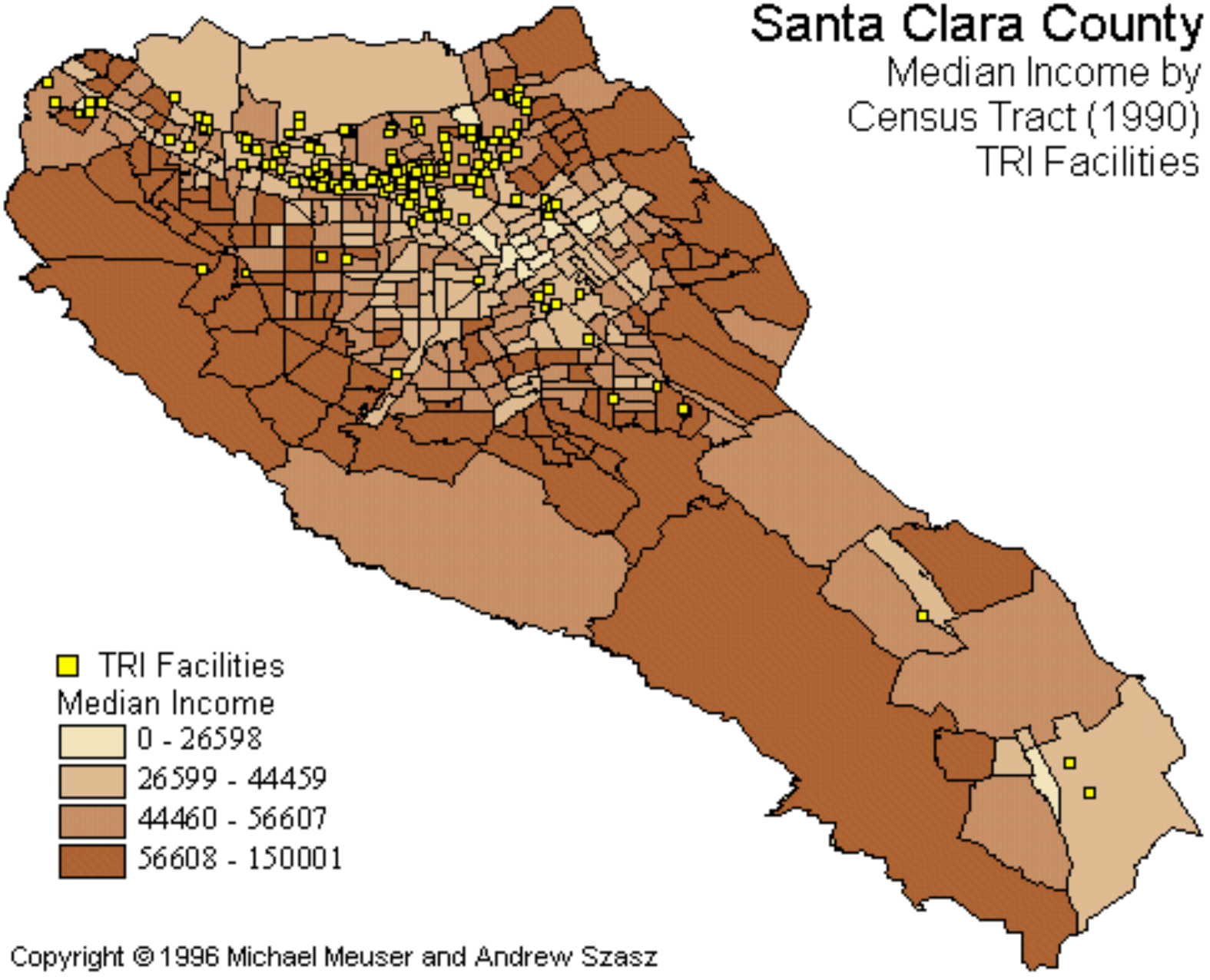
- ***Layering***



What questions can we answer with layering?

- Do certain features in layer A occur in the same (or similar locations) as features in layer B.
- Examples
  - Spatial case-control study.
  - Bars and DUI arrests.
  - Library locations and school performance.
  - Environmental justice.

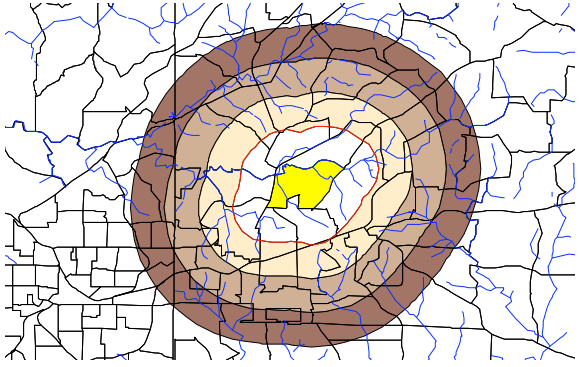
Layering example: Environmental justice



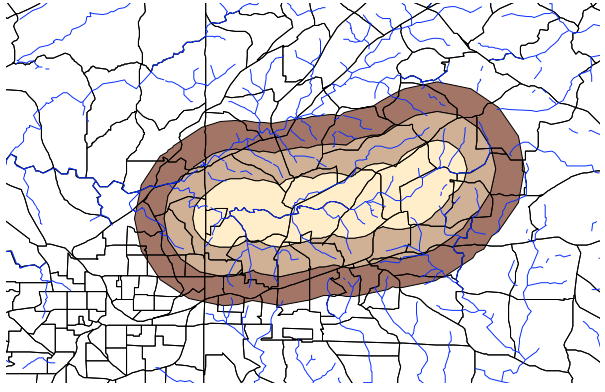
Basic GIS operation 2: Buffering

- ***Buffering***
  - Find areas within a user-specified distance of:
    - points
    - lines
    - areas

Buffers around an area



Buffers around a line feature



What questions can you answer with layering and buffering?

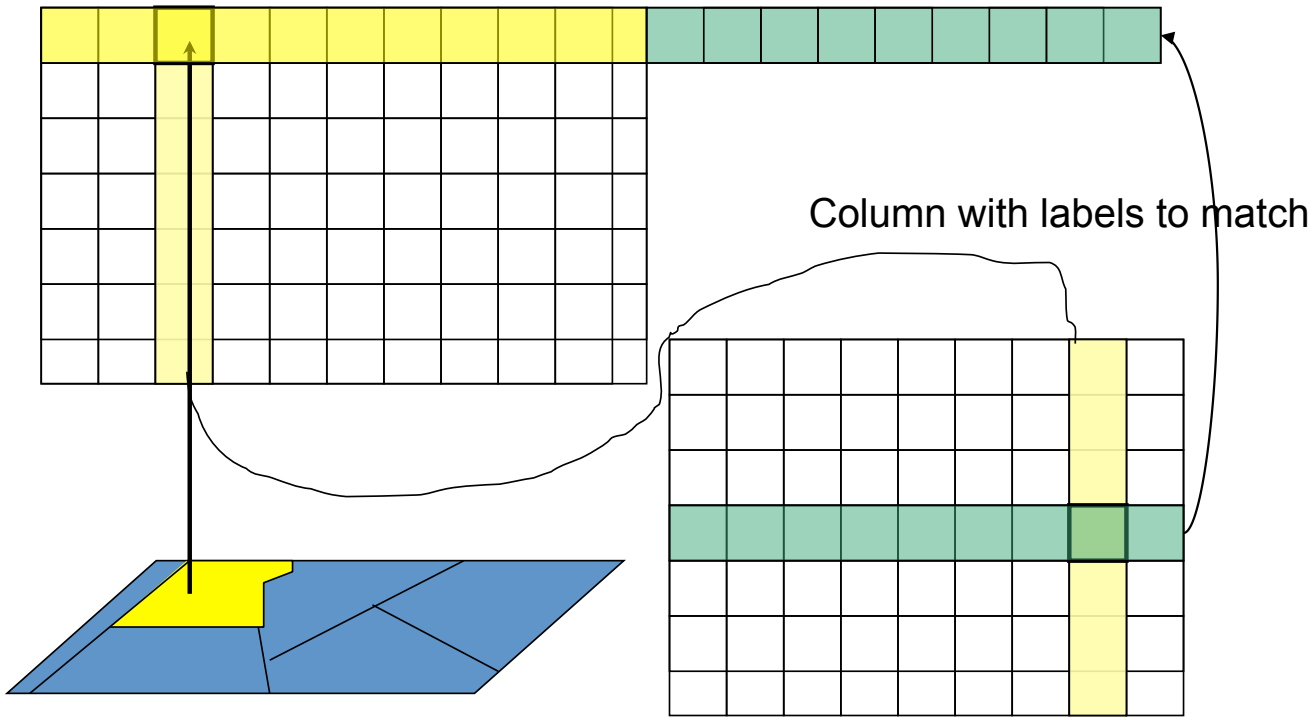
- Layer 1: Pollution sources
- Layer 2: Residents experiencing health effects (cases)
- Layer 3: Residents without health effects (controls)
- Question 1: What fraction of cases are within a given distance of a pollution source?
- Question 2: What fraction of controls are within a given distance of a pollution source?
- Question 3: Are these the same?
- This is the quintessential GIS environmental health study.

Basic GIS operation 3: Joining

- The spatial "join":
- Have:
  - Attribute table linked to map
  - 2nd table of data over same features
  - Common identifier in both
- Want:
  - Add (join) attributes in 2nd table to first table
  - How: Link tables based on common attributes
  - Need: One-to-one correspondence

Basic GIS operation 3: Spatial Join

Visually...



Joining: More detail

- Imagine two tables
  - Table A linked to a map
  - Table B in Excel (not linked to map)
- Can I add the data from Table B to Table A?
- Yes, if I have a field in both tables to tell me where data from a row in Table B should go in Table A (which row in Table A)?
- Relational databases don't actually merge the tables into a new one, they just hook the two tables together.
- But you can save it as a new table with both sets of data.

Joins have a *direction*

- Add *source table* to *destination table*.
- Direction matters:
  - Suppose have a shapefile (with map) of states and a table (no map) of demographics.
  - Souce (demographics) to destination (states) gives shapefile with mappable demographics for each state.
  - Source (states) to destination (demographics) gives table (no map).

Cardinality

- Cardinality = numbers.
- For joins we can have:
  - One-to-one: (one demographic record for each state).
  - One-to-many: (many source records to one destination record, demographics for each year for each state).
  - Many-to-one: (many destination records match single source record, single state to many cities).
  - Many-to-many: (many destinations to many sources, students and classes).

Rule of Joining

- There must be one and only one record in the source table for each record in the destination table.
- One-to-one? OK.
- Many-to-one? OK (join counties to states).
- One-to-many? No joining (but you can *relate*).
  - You can asociate records between tables but you cannot *join* the tables into one.
- Many-to-many? No joining or relating.

Many to many

- Chaos!
- Students in classes.
- No joining, no relating.
- A one-to-many relate could be done for each class to get a student list, OR
- A (different) one-to-many relate could give a class list for each student, BUT
- No single, master relate gives both.

What questions can you answer with joins?

- How long have cases resided in their current residence?
- Layer 1: Map of case residences.
- Data table 1: Tax records for all residences, including length of ownership.
- Joined data: Location of case residences and how long families have owned the residence.
- Main point: Can add data to layers to create new data!

A detailed example

- Guthe et al. (1992, *Environmental Research*
  - Lead exposure in children in Newark/East Orange/Irvington, NJ.
  - Used existing data to predict populations of children at high risk of lead exposure.

Components

- Existing data
  - GIS links disparate data sources to address issue none was specifically designed to address (the analysis of "found" data).
- Predict populations.
  - Analysis does not predict individual exposures but predicts groups of children with risk of high exposure.
- Risk
  - Study doesn't find which children *had* high exposure, it finds groups of children *likely* to have experienced high exposure.
- Data and analysis choices changes the question.

What questions can we answer and what data do we have?

- Where is the lead (likely to be)?
  - Near waste sites.
  - Paint in older houses.
  - Pipes in older houses.
- Where are the children?
  - Census (decennial).
  - Birth records (mobile population).
  - School enrollment.
- What data help?

Guthe et al's data

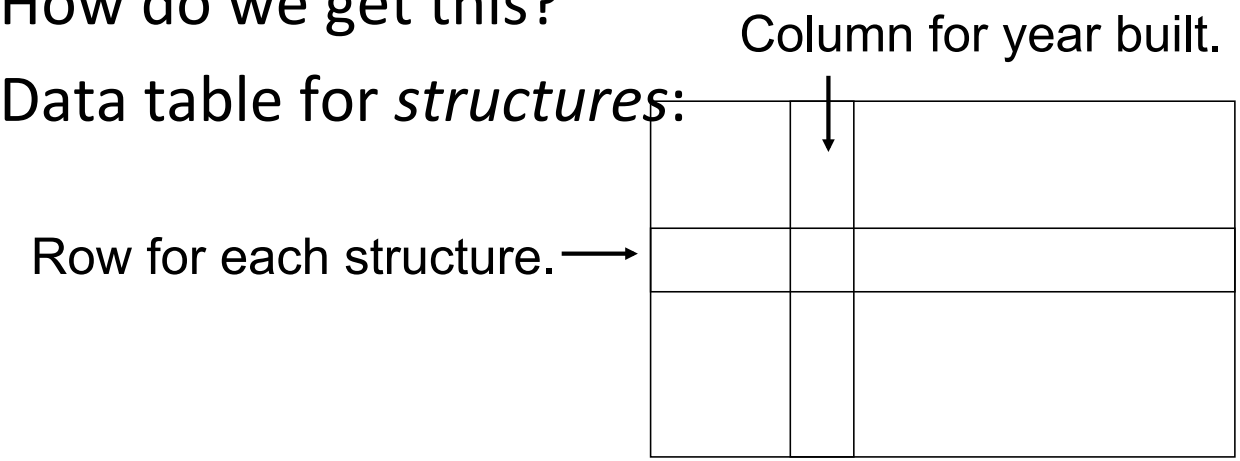
- Geographic data
  - Census tract boundaries
  - Locations of lead sources from industrial and hazardous waste sites
    - NJ Dept of Environmental Protection and Energy
  - Vehicle traffic miles/road
    - NJ DOT
- Blood lead screening records from county health department.
- Spatial query of census tracts with:
  - 620 or more structures built before 1940 AND
  - 290 or more children aged < 5 years.
- Guthe et al. report good but imperfect correlation between CTs with predicted high blood lead and those with high screening values.

How do we do this?

- *Query* the table.
- Example: Recall lead level case study and suppose we want to identify census tracts with ¿620 structures built before 1940 and containing $>$ 290 children aged $<$ 5 yrs.

Step 1

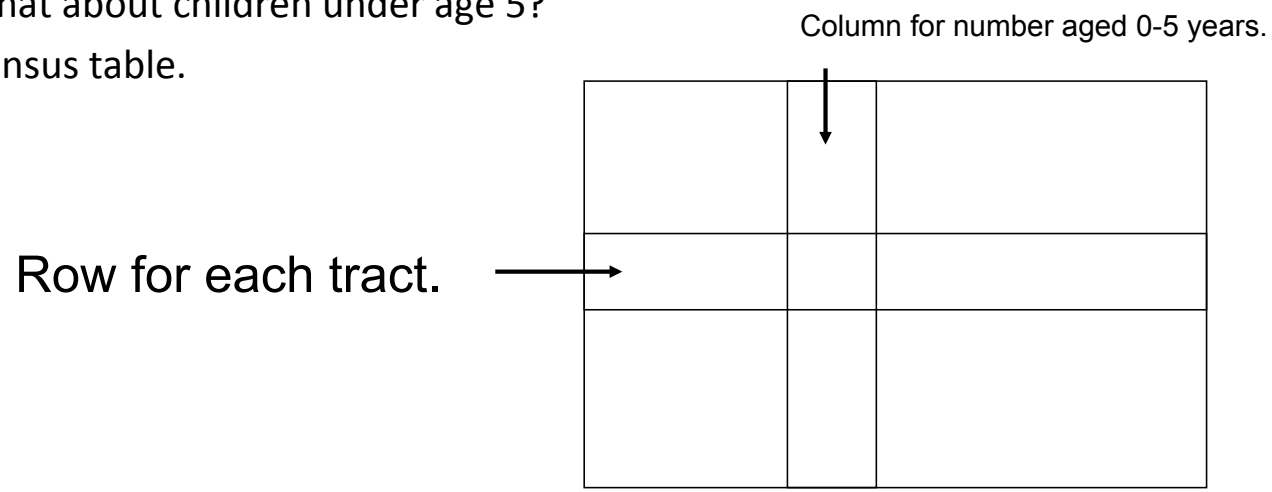
- How do we get this?
- Data table for *structures*:



- *Select* structures built before 1940.
- Display on the map.

Step 2

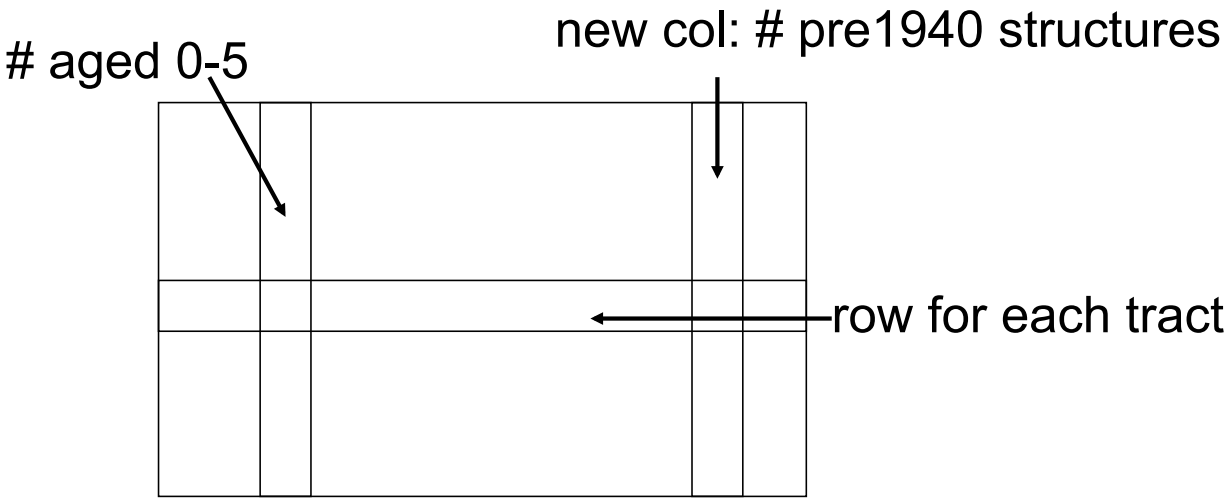
- What about children under age 5?
- Census table.



- Need: number of selected structures (pre 1940) in each census tract.
- How do we get this?

Step 3

- *Layer* the structure map (points) onto the census tract map (areas).
- *Summarize*: sum the number of structure features within each tract.
- Now we have an expanded tract table:



- Now *select* tracts with more than 290 children aged 0-5 AND more than 620 pre-1940 structures.
- Finally, *display* selected tracts on map.
- In summary:
  - *Selecting* in table.
  - *Layer* via spatial location.
  - *Summarize* on map (assign point features to areas).
  - *Summarize* within table (numbers within areas).
  - *Display* areas on map.

Using layering, buffering, and joining in creative ways

- Xiang et al. (2000, *Environmental Research*.
- Question: Relationship between maternal exposure to pesticides and adverse birth outcomes?
- Epidemiologic studies inconsistent.
- Weld County CO (North of Denver): Corn, beans, sugar beets, alfalfa hay.
- What data can we get?

Data (3 layers)

- Remotely sensed (satellite, raster) data on crop type
  - 28.5 $\times$ 28.5 m resolution
  - 1991 and 1993 (1992 cloudy)
  - Rule: If 1991 and 1993 match, assign same crop to 1992. If differ, no crop for 1992.
- Locations of rural residences from directory of Weld County.
  - Extract maternal residence locations from all live births registered with CO Dept of Public Health and Environment.
  - Create attribute table for each maternal address location including: sex of baby, weight, gestational age, maternal age, maternal education, maternal smoking during pregnancy.
- CO Pesticide Use Survey (1992) (TABLE ONLY)
  - Which chemicals applied to which crops, portion of acres treated, applications per season, application rate.
  - No location information, summarize by crop type.

Using our GIS operations

- 300m and 500m buffer around each maternal residence.
- Calculate area within buffer associated with each crop, link crop to pesticide.
  - Pixel within buffer associated with crop type by Layer 1.
  - Crop type associated with pesticide by Layer 3.
  - Assume pesticide applied at average rate and amount to pixel in buffer.
- Challenges:
  - Only 125 residences linked to specific location from address.
  - Difficult to pinpoint specific chemicals so summarize usage by crop type.

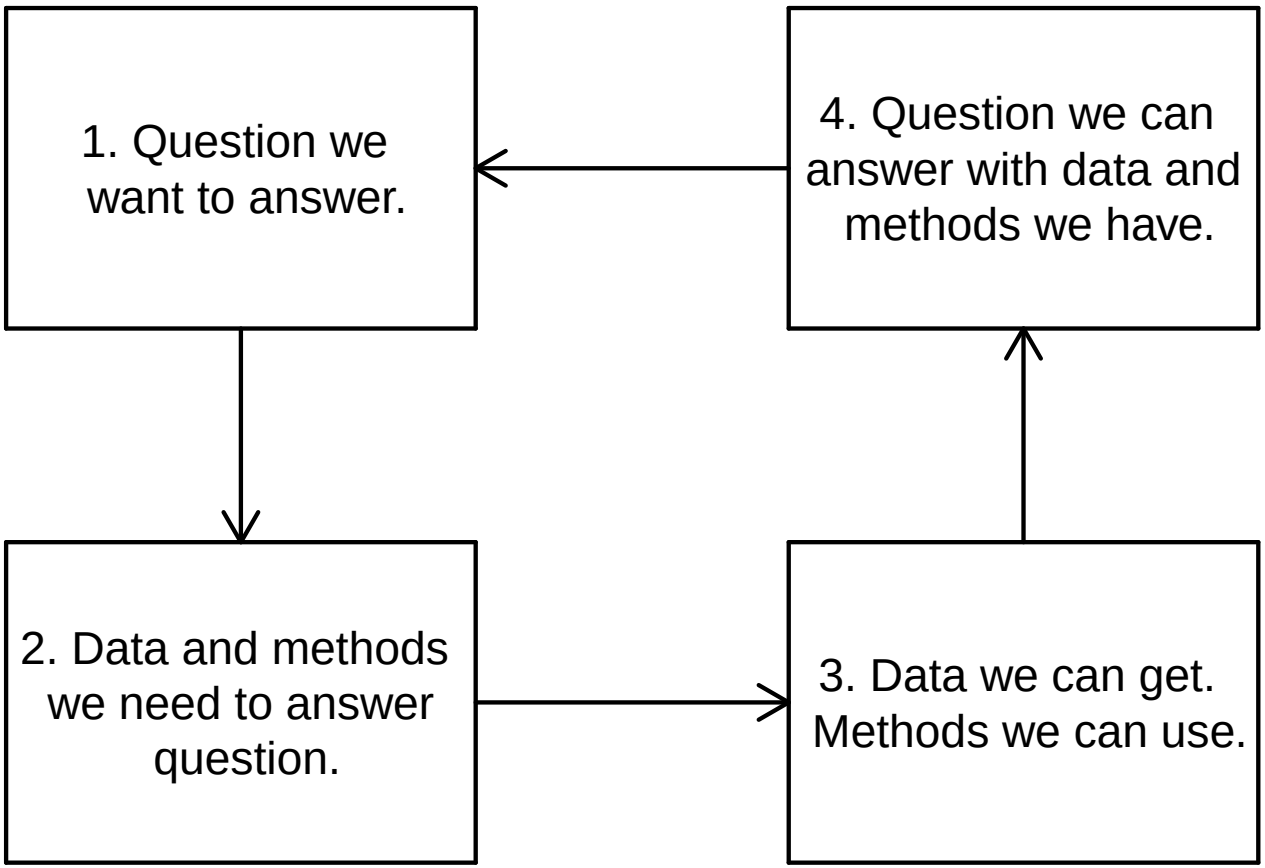
Interpreting results

- Careful interpretation: "While RS/GIS technology may enhance epidemiologic research, it will not replace the traditional epidemiological methods and approaches involving accurate measurements of environmental exposures."
- Moral: GIS allows creation of (sometimes very sensible) exposure surrogates, but does not offer same level of accuracy as exposure measurement.

Basic GIS operations

- Layering
- Buffering
- Joining
- All GISs do these. They do more, but all of these basic operations are included.

The whirling vortex



GIS analysis

- What can you do with these three operations?
- The key to GIS analysis is to break your problem down into steps consisting of these operations.
- What question(s) do you have?
- What data would you need?
- What data can you get?
- Can you layer, buffer, join data to enable summaries relating to your question (or parts of it)?
- What answers can you provide?

Scenario 1: Toxic train

- At 2 a.m. your GIS hotline rings and you are informed that a train transporting chlorine gas just derailed in near Helena, Montana creating a large cloud of chorine gas. You are asked to coordinate the GIS component of the response.
- What questions?
- What data do you want?
- What data can you get (in what time)?
- What questions can you answer (in what time)?

Scenario 2: Site Selection

- Your city has landed a contract with a large firm to locate their new manufacturing plant somewhere within the city limits. Your GIS team has been asked to evaluate seven proposed sites for a single new manufacturing facility and you wish to ensure that no sociodemographic group is disproportionately impacted by the new facility. How can GIS help you rank the sites?
- What questions?
- What data do you want?
- What data can you get (in what time)?
- What questions can you answer (in what time)?

Scenario 3: Hurricane Help

- A large coastal city contacts your GIS team and asks for two analyses. The first is a plan to aid in the event of a predicted hurricane. The second is a plan for response after a hurricane hits. How do the two tasks differ? Do they require the same data?
- What questions?
- What data do you want?
- What data can you get (in what time)?
- What questions can you answer (in what time)?

Doesn't GIS do statistics?

- GIS analysis can
  - Show patterns,
  - Illustrate areas with high crude rates,
  - Show if high rates are near exposure sources.
- Statistical analysis needed to see
  - Patterns different from random allocation (constant risk)?
  - Are highest rates higher than expected?
  - Are high rates associated with high exposures?

GIS and statistics

- Some tools and toolboxes available, but few and specific.
- Do we really want SAS to do GIS?
- Do we really want ArcGIS to do statistics?
- Both based on objects and operations, but different objects and operations.

Spatial analysis vs. spatial data analysis

- Spatial analysis
  - Combining GIS operations to get summaries, sometimes complicated summaries.
- Spatial data analysis
  - Statistical tests of pattern (Do we have a cluster?)
  - Regression (are values associated?)
  - Prediction (what is the temperature *here*?)
- But little consistency in usage across literatures.

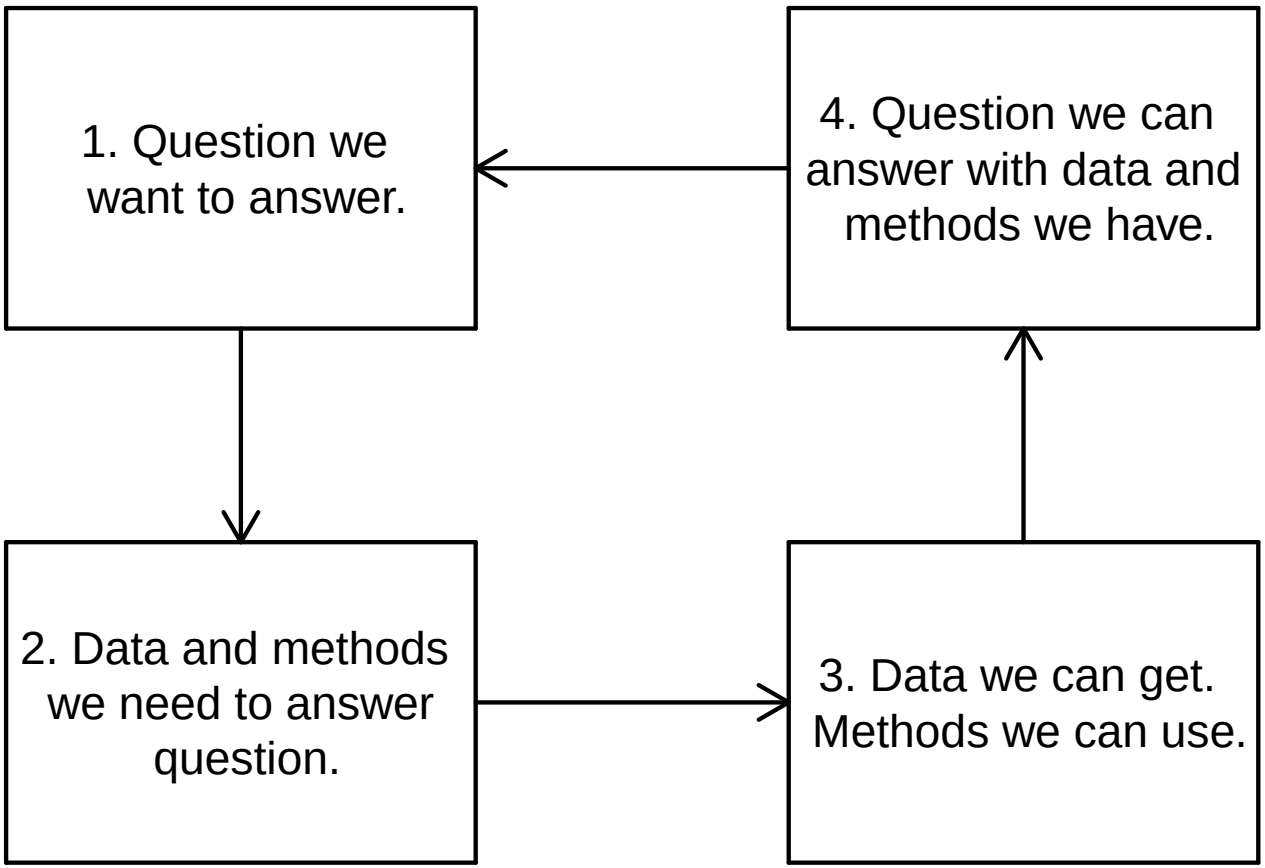
Disciplines and spatial statistics

- Many disciplines have their own rules of thumb with spatial analysis.
- The key questions and methods vary from discipline to discipline.
- Geography: Spatial autocorrelation (Moran's I, LISAs, spatial regressions).
- Ecology: Associations and diffusion (Mantel tests).
- Criminology: Hotspots.
- Epidemiology: Clusters, Poisson/logistic regression.

Role of statistics

- Different methods are fine for different questions, or different data restrictions.
- Statistical thinking thinking places question in probabilistic setting, and builds inference on data-based summaries.
- Compare methods on performance (probability of proper classification, probability of detection, probability of false alarms).

The whirling vortex



Summary

- Maps are cool.
- Maps place data spatially.
- Spatial data enable answers for spatial answers.
- Spatial data also allow spatial statistics.
- So what spatial statistics can we do?