CSSS/STAT/SOC 321
Case-Based Social Statistics I

Course Introduction

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle

Outline for today

What is Statistics?

Our first case: John Snow's celebrated cholera map

Course details

Some examples to ponder

A typical encounter with “statistics” takes places when we read a news item on a survey:

1. In 1995 USA Weekend magazine asked readers to return a survey with a variety of questions about sex and violence on television. Of the 65,142 readers who responded, 98% were “very or somewhat concerned about violence on TV”. Based on this survey, can we conclude that about 98% of U.S. citizens are concerned about violence on TV? Why or why not?

Some examples to ponder

A typical encounter with “statistics” takes places when we read a news item on a survey:

1. In 1995 USA Weekend magazine asked readers to return a survey with a variety of questions about sex and violence on television. Of the 65,142 readers who responded, 98% were “very or somewhat concerned about violence on TV”. Based on this survey, can we conclude that about 98% of U.S. citizens are concerned about violence on TV? Why or why not?

2. In the November 2004 presidential election, many media outlets reported that exit polls of bellwether districts showed George Bush winning 44% of the Hispanic vote. But telephone polls days before the election showed Bush winning only 32% of the Hispanic vote. Why the discrepancy? Which number would you trust?

What is Statistics?

Popular and traditional meaning:
“Statistics are numbers measured for some purpose”

What is Statistics?

Popular and traditional meaning:
"Statistics are numbers measured for some purpose"

Really, numbers measured for some purpose are *not* statistics, but *data*

Actual scientific meaning:

Statistics is the collection of procedures and principles
for gaining and processing information
in order to make decisions
when faced with uncertainty

What is Statistics?

Popular and traditional meaning:
"Statistics are numbers measured for some purpose"

Really, numbers measured for some purpose are *not* statistics, but *data*

Actual scientific meaning:

Statistics is the collection of procedures and principles
for gaining and processing information
in order to make decisions
when faced with uncertainty

Hence statistics is concerned with:

1. The process of data collection

What is Statistics?

Popular and traditional meaning:
"Statistics are numbers measured for some purpose"

Really, numbers measured for some purpose are *not* statistics, but *data*

Actual scientific meaning:

Statistics is the collection of procedures and principles
for gaining and processing information
in order to make decisions
when faced with uncertainty

Hence statistics is concerned with:

1. The process of data collection
2. Summarizing the information within data

What is Statistics?

Popular and traditional meaning:
"Statistics are numbers measured for some purpose"

Really, numbers measured for some purpose are *not* statistics, but *data*

Actual scientific meaning:

Statistics is the collection of procedures and principles
for gaining and processing information
in order to make decisions
when faced with uncertainty

Hence statistics is concerned with:

1. The process of data collection
2. Summarizing the information within data
3. Proper interpretation of data to answer a scientific research question

Why Statistics?

Why should you learn statistics?

Statistics is the science of uncertainty –
and almost everything we learn is uncertain

Statististics helps us navigate/summarize/infer from oceans of data –
and the computer age has produced vast datasets like never before

Statistics is ubiquitous in scientific fields –
and in most grad programs (business, policy, medicine)

Helps us understand the big picture (general laws)
and how each individual varies from that big picture

Key elements of a statistical study

1. The individuals or objects studied and how they were selected
2. The variables measured about each object
3. The setting or context in which the measurement were made
4. Unmeasured variables on which the subjects vary
5. The magnitude of any claimed effects of differences in measured variables
6. The uncertainty of these claimed effects

Statistics as part of the Scientific Method

1. Observe the world / Read past studies
2. Form a research question based on 1.
3. Build a theory, preferably causal, to answer the question
4. Choose an area to test theory
5. Operationalize theory: Measure variables, generate hypotheses
6. Analyze the data obtained in 5.
7. Report results: is the hypothesis confirmed, or rejected?
8. Replicate & repeat. . .

Statistics as part of the Scientific Method

1. Observe the world / Read past studies
2. Form a research question based on 1.
3. Build a theory, preferably causal, to answer the question
4. Choose an area to test theory [SELECTION]
5. Operationalize theory: Measure variables, generate hypotheses [MEASUREMENT]
6. Analyze the data obtained in 5. [ANALYSIS]
7. Report results: is the hypothesis confirmed, or rejected? [INTERPRETATION]
8. Replicate & repeat...

Populations vs. Sample

Population: Complete set of units of interest in a study

e.g., all American voters;
all students at UW;
all friendships of the students in this class

vs.

Populations vs. Sample

Population: Complete set of units of interest in a study

e.g., all American voters;
all students at UW;
all friendships of the students in this class

vs.

Sample: The subset of the population actually studied, which may be random or non-random; representative or non-representative.

e.g., 1000 voters dialed at random;
500 UW students choosen by random ID number;
the first friendship mentioned by each student in this class.

When the sample includes the entire population, we call it a census

Observation vs. experiment

Observation: A study of the relationship among several variables based on their natural variation

e.g., a longitudinal survey tracks 1000 children over several years, noting how much violent TV each watches and whether they committed violent crimes

vs.

Observation vs. experiment

Observation: A study of the relationship among several variables based on their natural variation

e.g., a longitudinal survey tracks 1000 children over several years, noting how much violent TV each watches and whether they committed violent crimes

vs.

Experiment: A study of the relationship between two or more variables, one of which is controlled by the experimenters

e.g., scientists randomly assign 500 children to a group which watches violent TV, and 500 to a group which does not, then records their rates of criminal activity.

Internal validity: Whether a study is conducted well enough to make valid inferences about the relationship of variables in the population from the sample

Consider the TV violence & crime example: a study with high internal validity is one that correctly estimates the effect of TV violence on the criminal activity of those studied

Internal validity: Whether a study is conducted well enough to make valid inferences about the relationship of variables in the population from the sample

Consider the TV violence & crime example: a study with high internal validity is one that correctly estimates the effect of TV violence on the criminal activity of those studied

Key threats to internal validity:

1. Measurement error: e.g., mistaking children with missing criminal records for non-criminals

Internal validity: Whether a study is conducted well enough to make valid inferences about the relationship of variables in the population from the sample

Consider the TV violence & crime example: a study with high internal validity is one that correctly estimates the effect of TV violence on the criminal activity of those studied

Key threats to internal validity:

1. Measurement error: e.g., mistaking children with missing criminal records for non-criminals
2. Selection bias: e.g., if parents of generally well-behaved children are more likely to be enrolled in the study

Internal validity: Whether a study is conducted well enough to make valid inferences about the relationship of variables in the population from the sample

Consider the TV violence & crime example: a study with high internal validity is one that correctly estimates the effect of TV violence on the criminal activity of those studied

Key threats to internal validity:

1. Measurement error: e.g., mistaking children with missing criminal records for non-criminals
2. Selection bias: e.g., if parents of generally well-behaved children are more likely to be enrolled in the study
3. Confounders: e.g., omitted variables like parental income which also affect crime

Well-run experiments tend to have high internal validity (randomization)

But even well-run observational studies are vulnerable to the above threats

External validity: Whether a study's findings apply to other similar situations in the real world (not just the lab)

External validity: Whether a study's findings apply to other similar situations in the real world (not just the lab)

Possibilities for failure here are endless, especially for experiments:

1. Artificiality of treatment: assigned TV may have less effect than self-selected TV

External validity: Whether a study's findings apply to other similar situations in the real world (not just the lab)

Possibilities for failure here are endless, especially for experiments:

1. Artificiality of treatment: assigned TV may have less effect than self-selected TV
2. Selection bias: what if the participants were recruited by TV ads?

External validity: Whether a study's findings apply to other similar situations in the real world (not just the lab)

Possibilities for failure here are endless, especially for experiments:

1. Artificiality of treatment: assigned TV may have less effect than self-selected TV
2. Selection bias: what if the participants were recruited by TV ads?
3. Duration of treatment: what if it takes 1000s of hours of TV violence?
4. Many more. . .

Well-designed observational studies can have high external validity

But even well-run experiments are vulnerable to the above threats

John Snow Saves London

Cholera outbreaks were common in 19th century London; 10,000s of deaths

Contemporary theories:

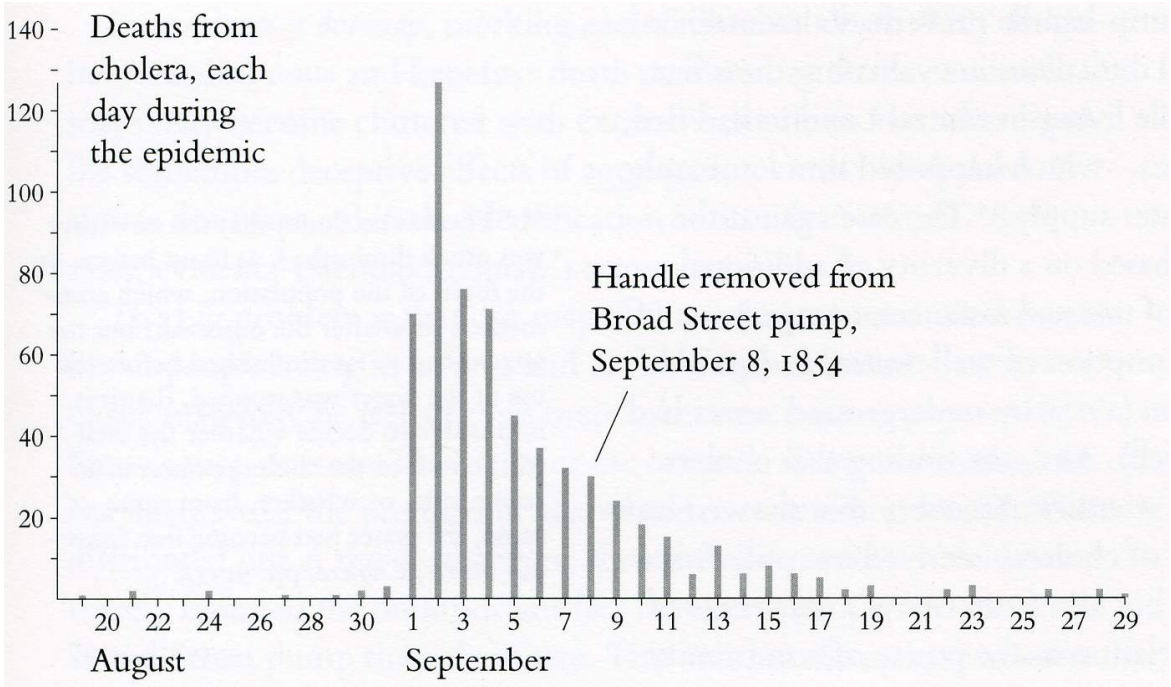
1. Cholera caused by “miasma” in the air coming from swamps
2. Or a “poison” slowly losing strength as it passes from victim to victim?
3. London doctor John Snow thought contaminated water the cause

Outbreak in 1854: 500 deaths in 10 days in Soho

Snow has Broad Street pump handle removed

Did he stop the epidemic? And prove disease can be spread by germs?

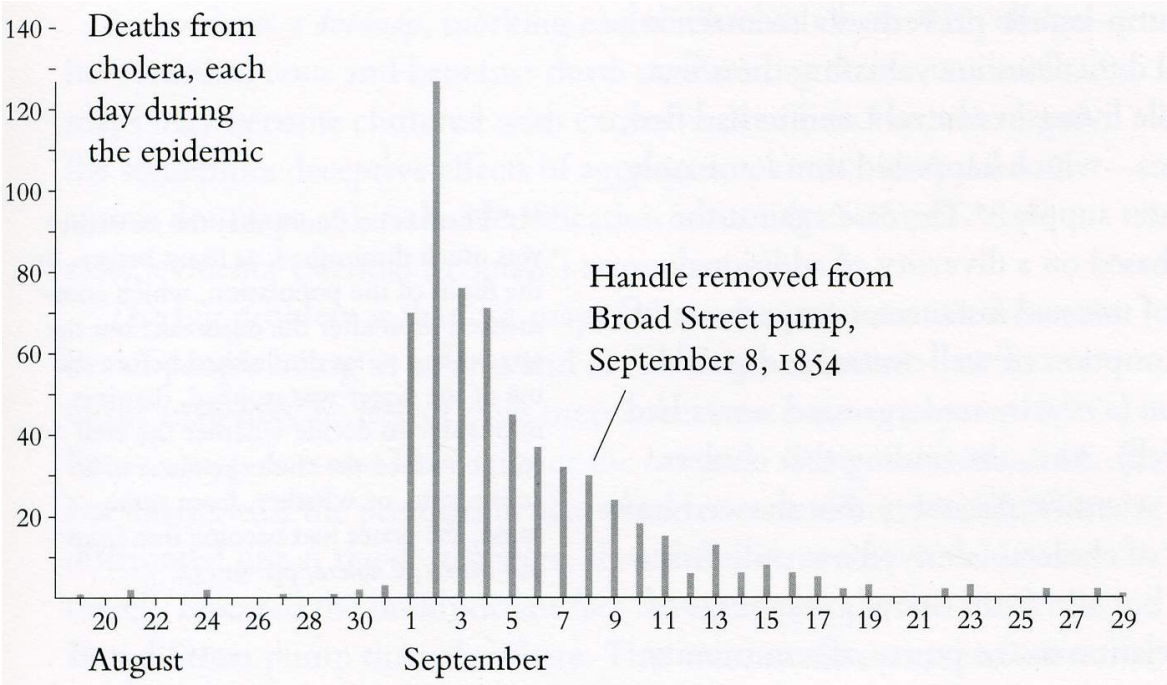
How might the newspaper "analyze" John Snows's intervention?



(plot from Tufte, *Visual Explanations*)

- Overwhelming tendency to view time series data this way
  Doesn't help us make inferences about the data
- The data aren't being compared to any other variables:
  time series plots don't help us devise a model of the data

How might the newspaper "analyze" John Snows's intervention?



(plot from Tufte, *Visual Explanations*)

- Can we do better? Specify a research question?
- Translate it into variables? Formulate some hypotheses?
- Think about internal and external validity?

Snow's spatial analysis

In 1954, London water was provided by competing private firms

Residents would walk to the nearest street pump for water

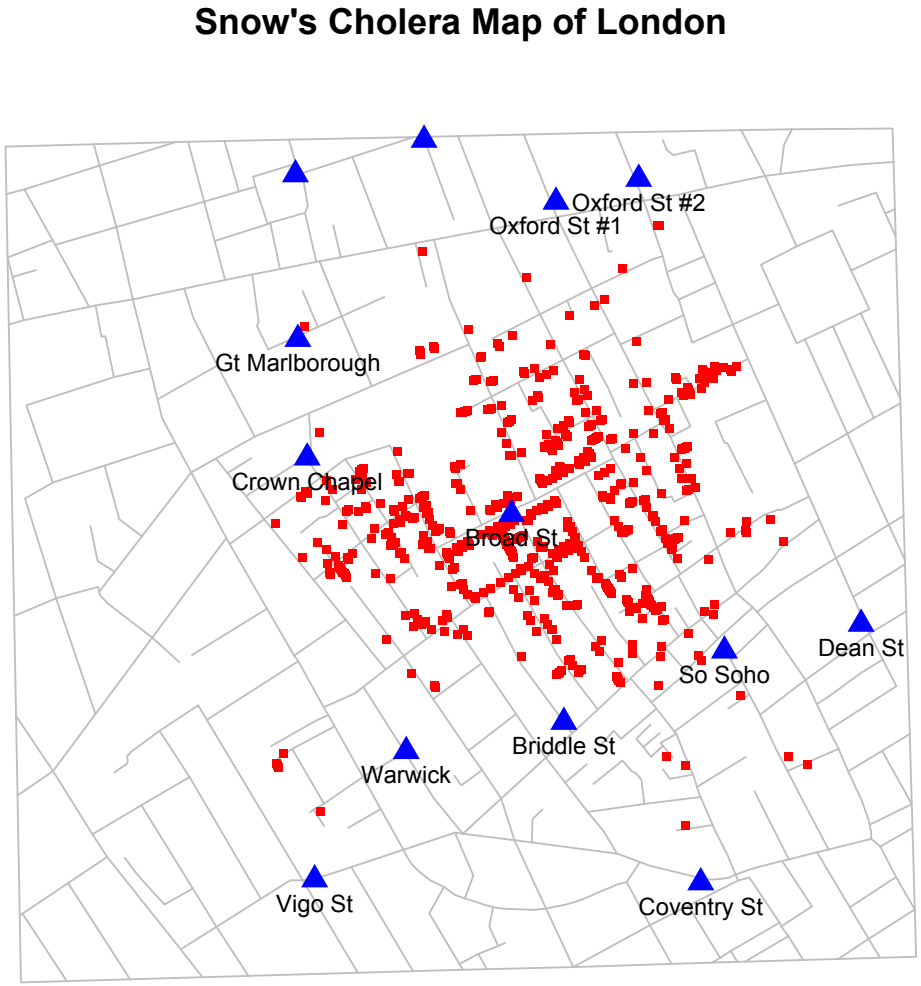
Snow recorded the location of each death in real time

Placed these spatial data on a map along with the *water pumps*

Was one pump, from a particular company, contaminated with cholera?

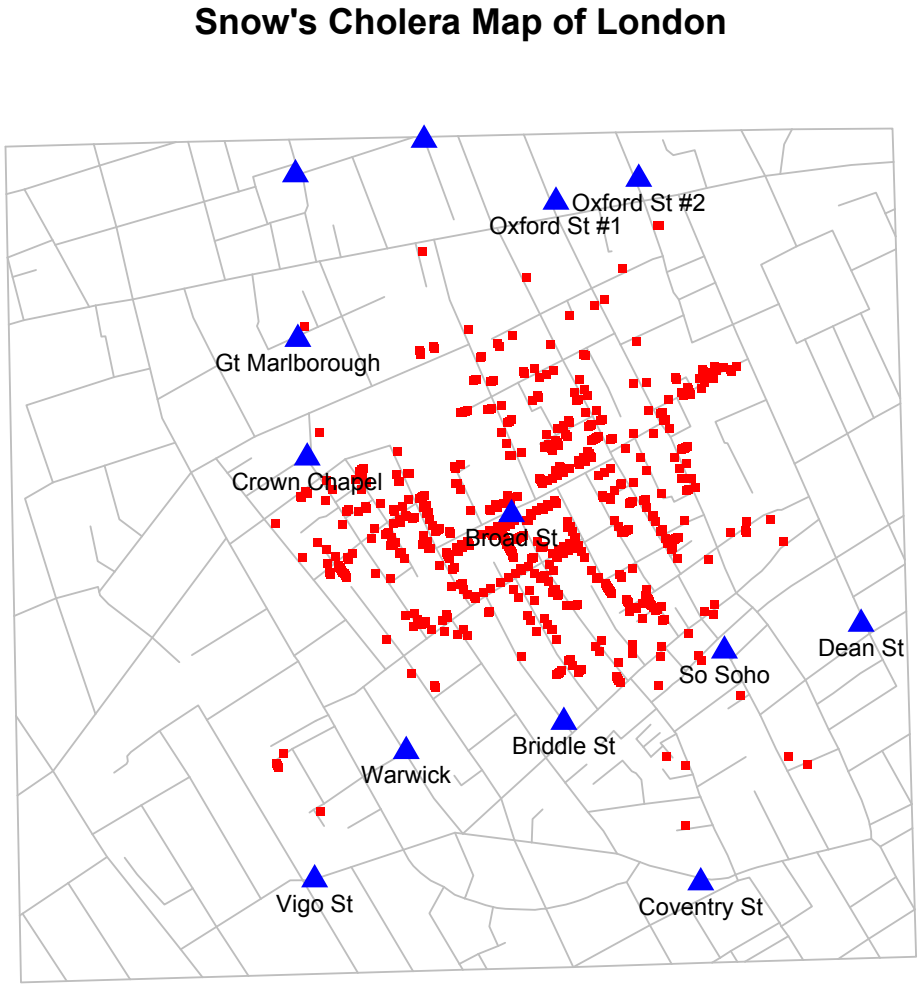
Snow's spatial analysis: Tufte redrawing

Snow's spatial analysis: Slide friendly version



What kind of sample did Snow collect?

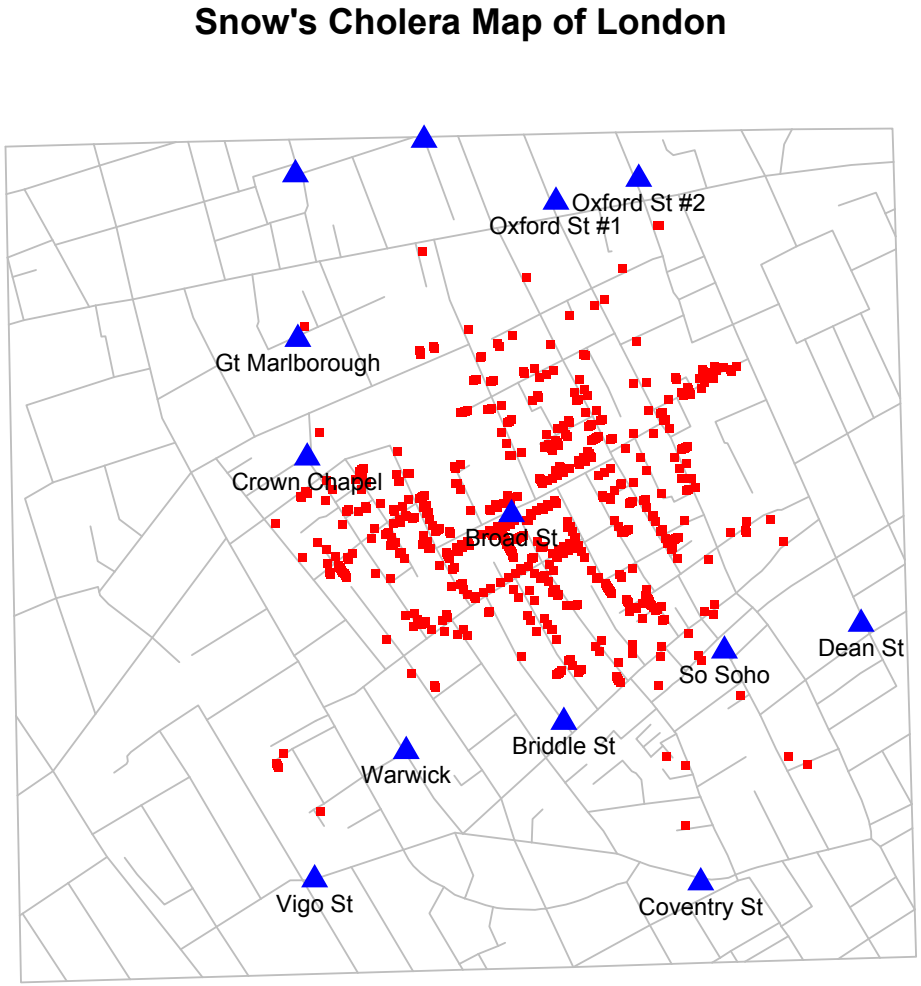
Snow's spatial analysis: Slide friendly version



What kind of sample did Snow collect?

A census of cholera victims – but what about non-victims?

Snow's spatial analysis: Slide friendly version

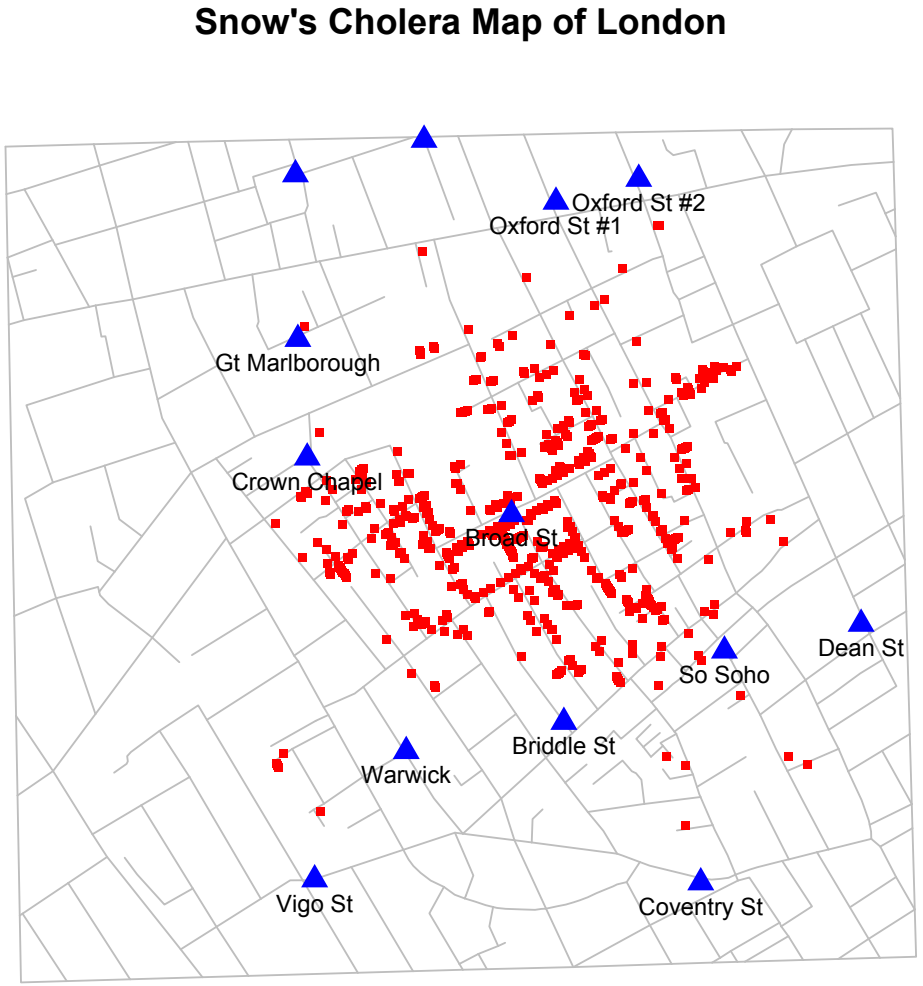


What kind of sample did Snow collect?

A census of cholera victims – but what about non-victims?

Is this an observational study or experiment?

Snow's spatial analysis: Slide friendly version



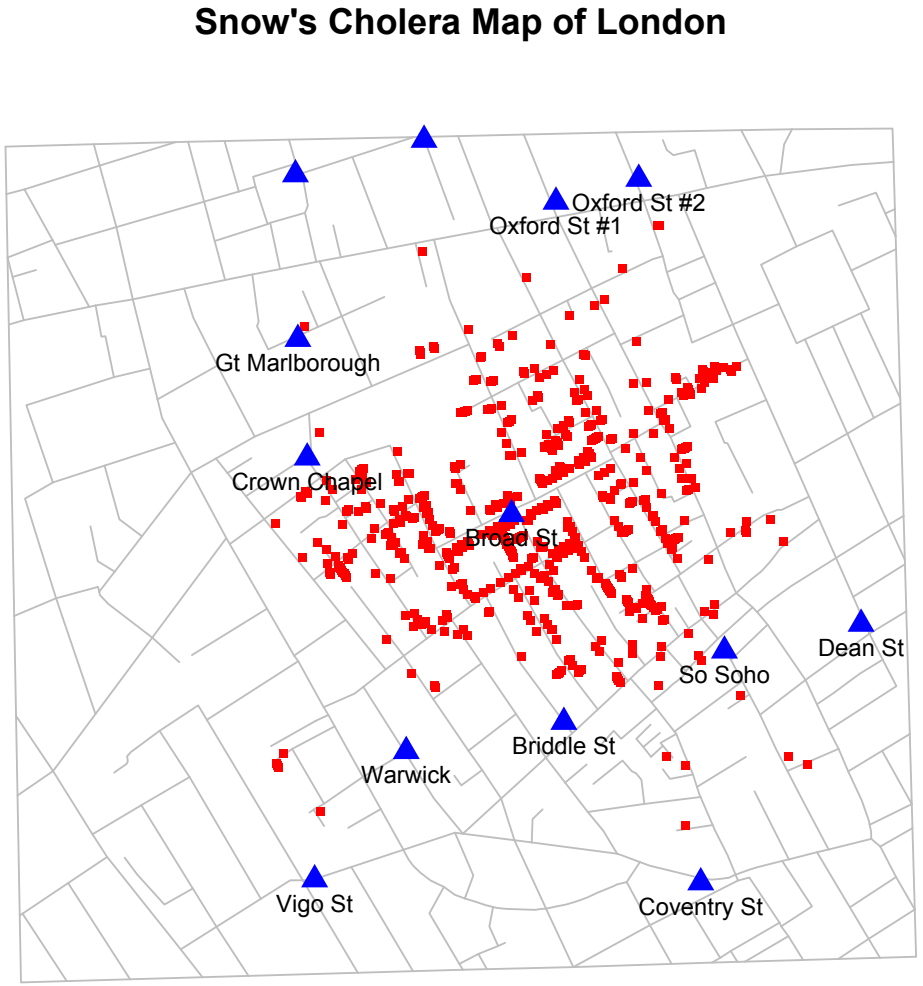
What kind of sample did Snow collect?

A census of cholera victims – but what about non-victims?

Is this an observational study or experiment?

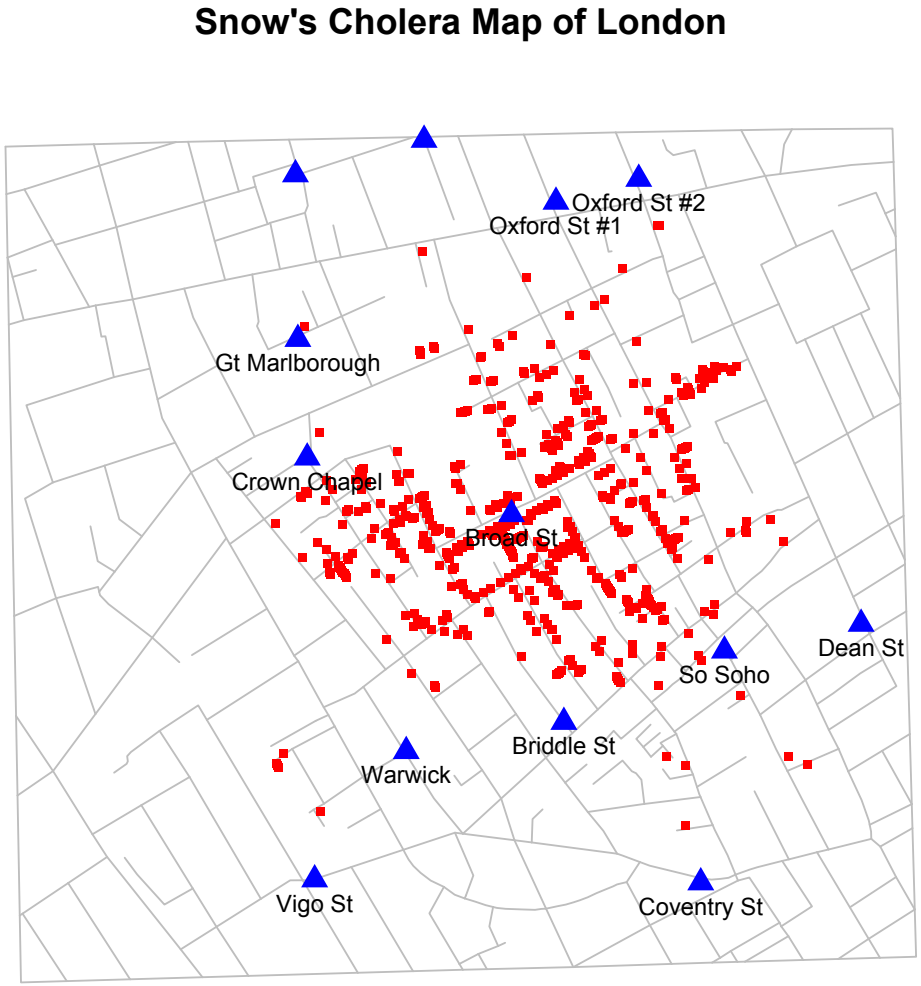
Combines features of both: a natural experiment

Snow's spatial analysis: Slide friendly version



How do we assess the relationship between deaths (red dots) and pumps (blue triangles)?

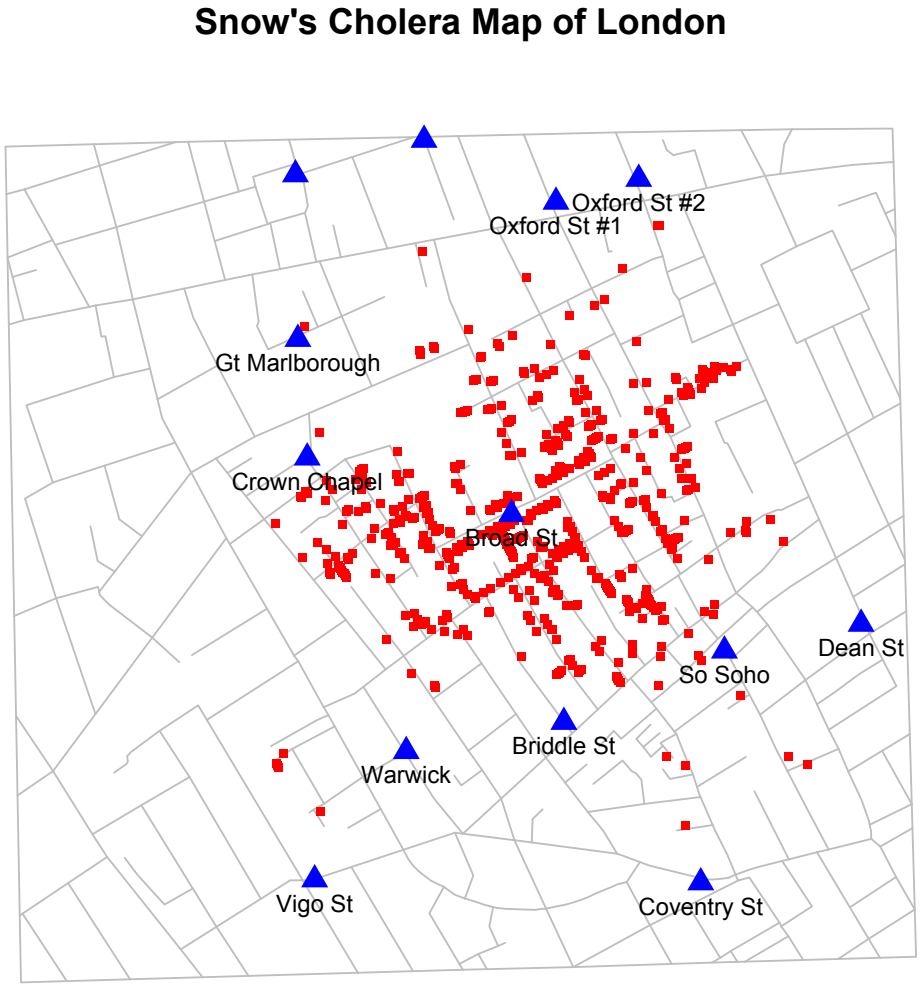
Snow's spatial analysis: Slide friendly version



How do we assess the relationship between deaths (red dots) and pumps (blue triangles)?

Are we convinced that a relationship exists?

Snow's spatial analysis: Slide friendly version

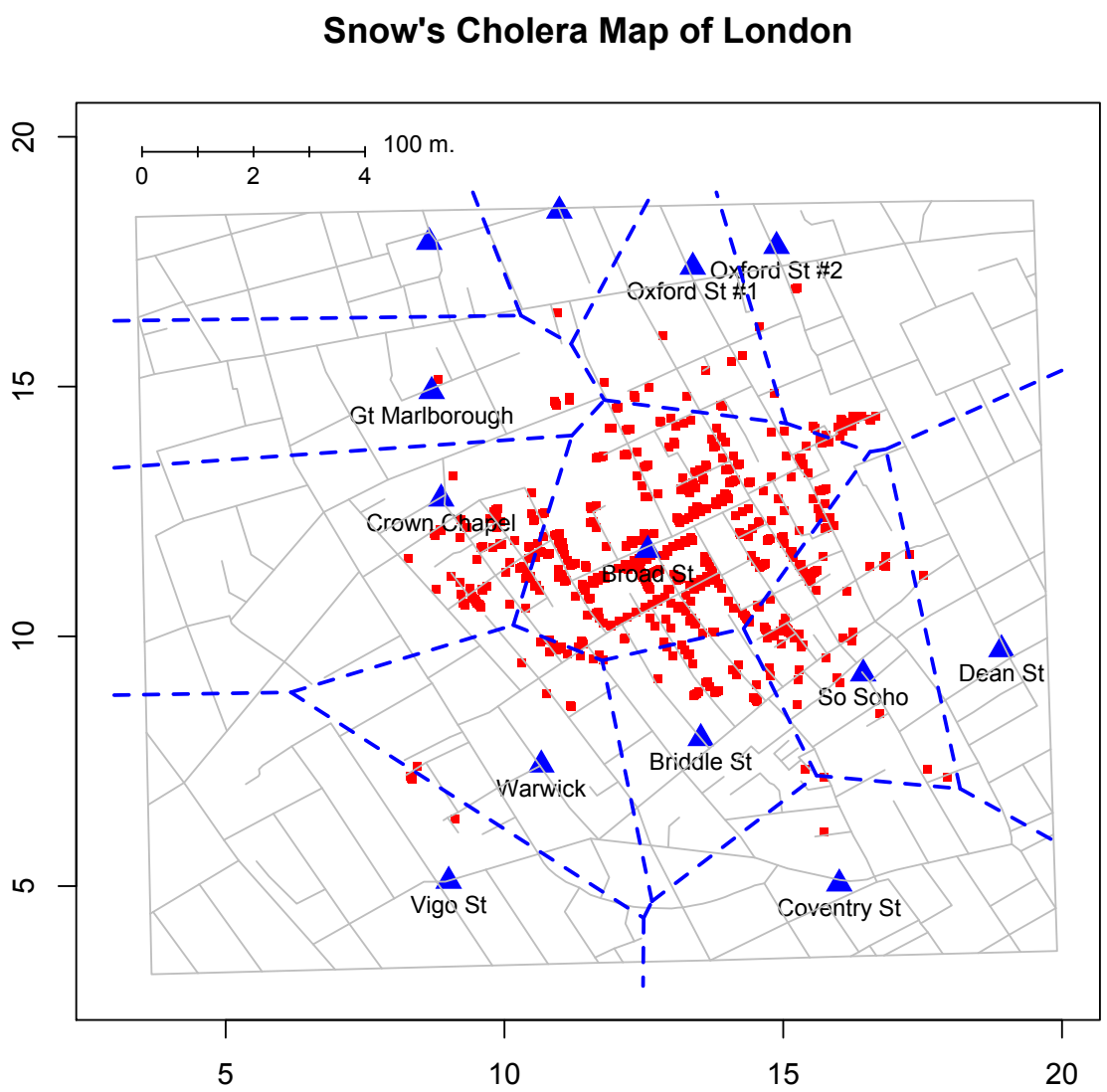


How do we assess the relationship between deaths (red dots) and pumps (blue triangles)?

Are we convinced that a relationship exists?

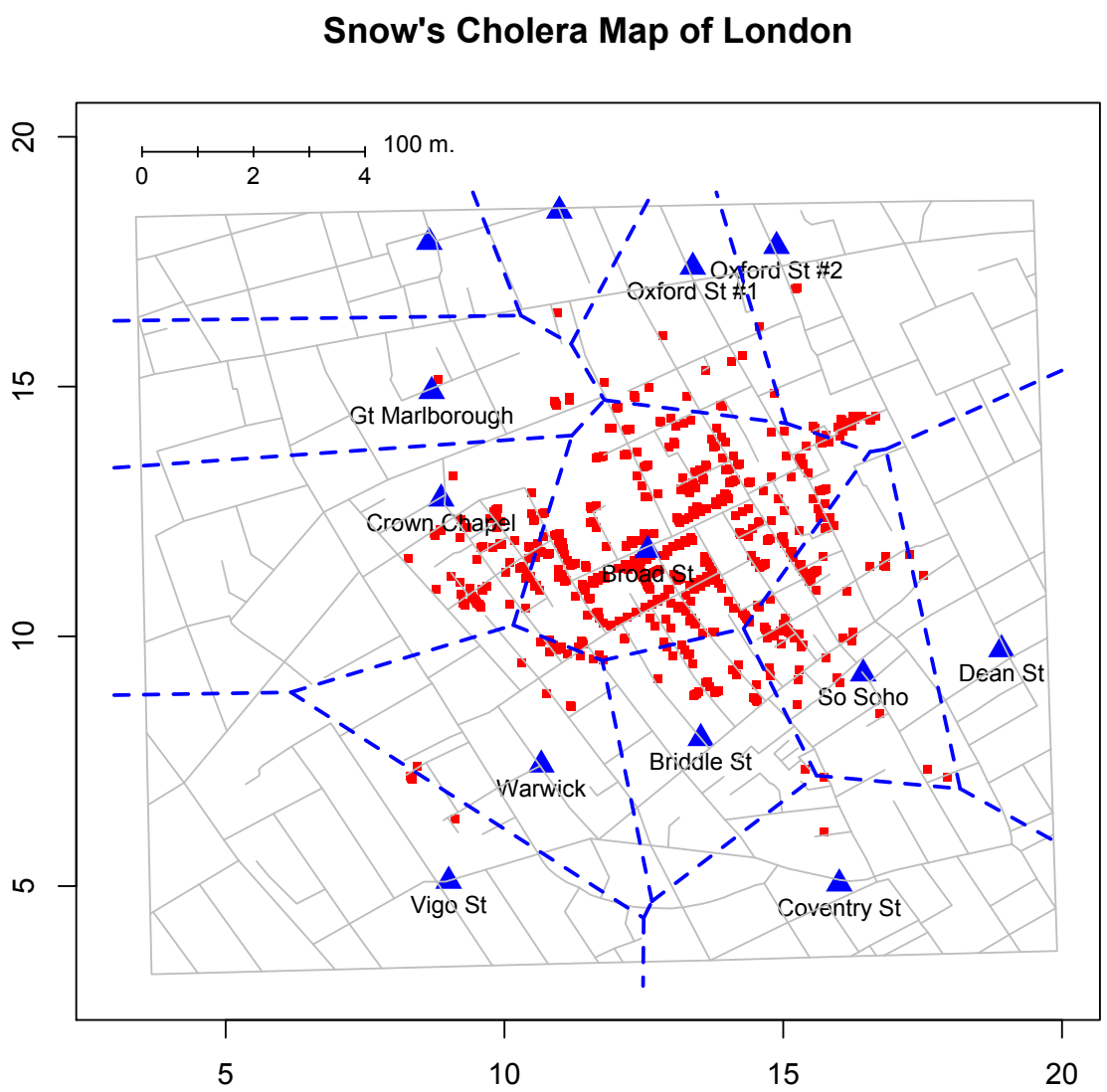
What additional variables should we measure?

Snow's spatial analysis: A simple visual model (Tobler 1994)



Fact: For any spot x on the map, there is a closest pump A

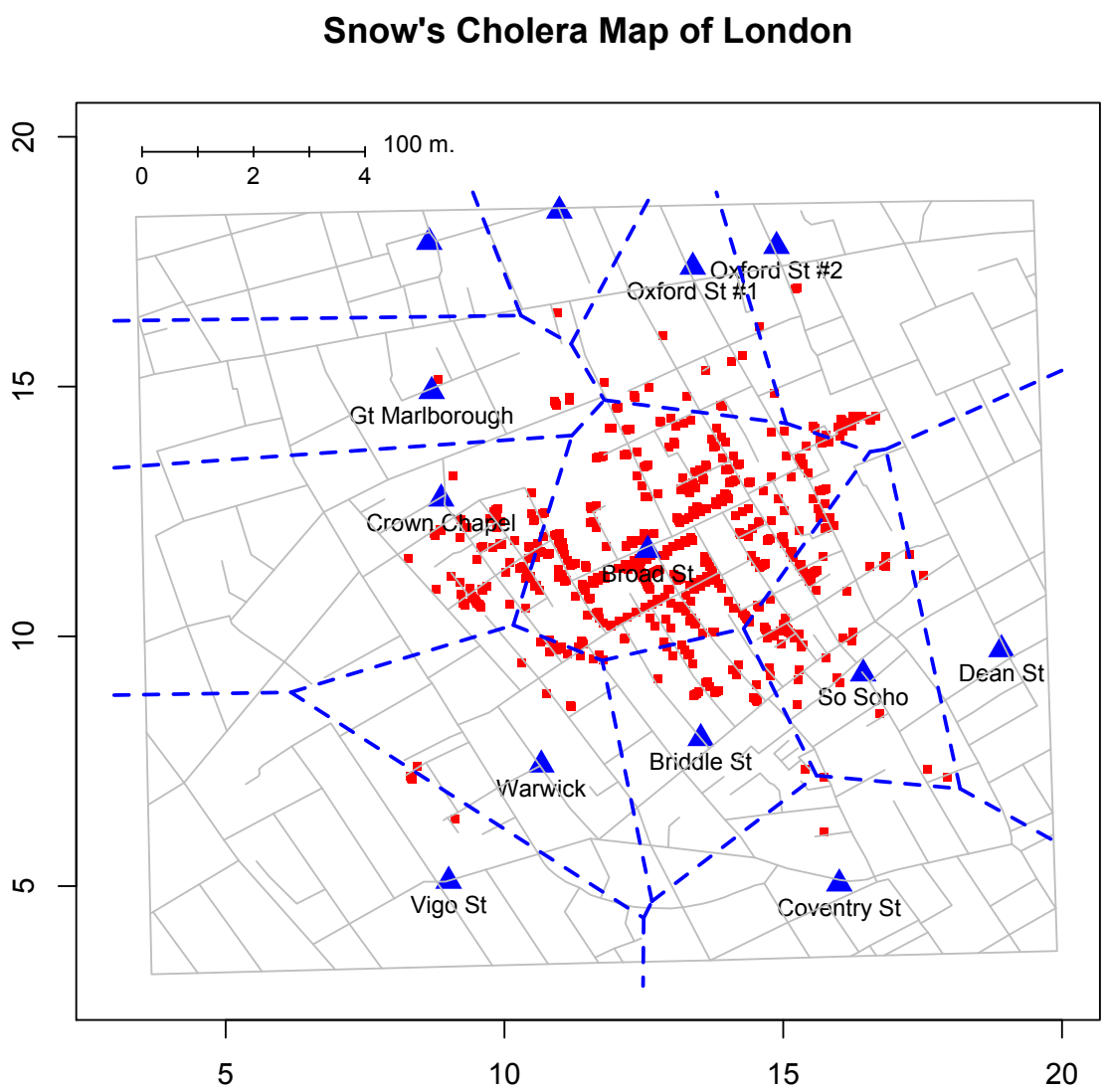
Snow's spatial analysis: A simple visual model (Tobler 1994)



Fact: For any spot x on the map, there is a closest pump A

Definition: The set of all points x closest to pump A is the *Voronoi cell* of pump A

Snow's spatial analysis: A simple visual model (Tobler 1994)



Fact: For any spot x on the map, there is a closest pump A

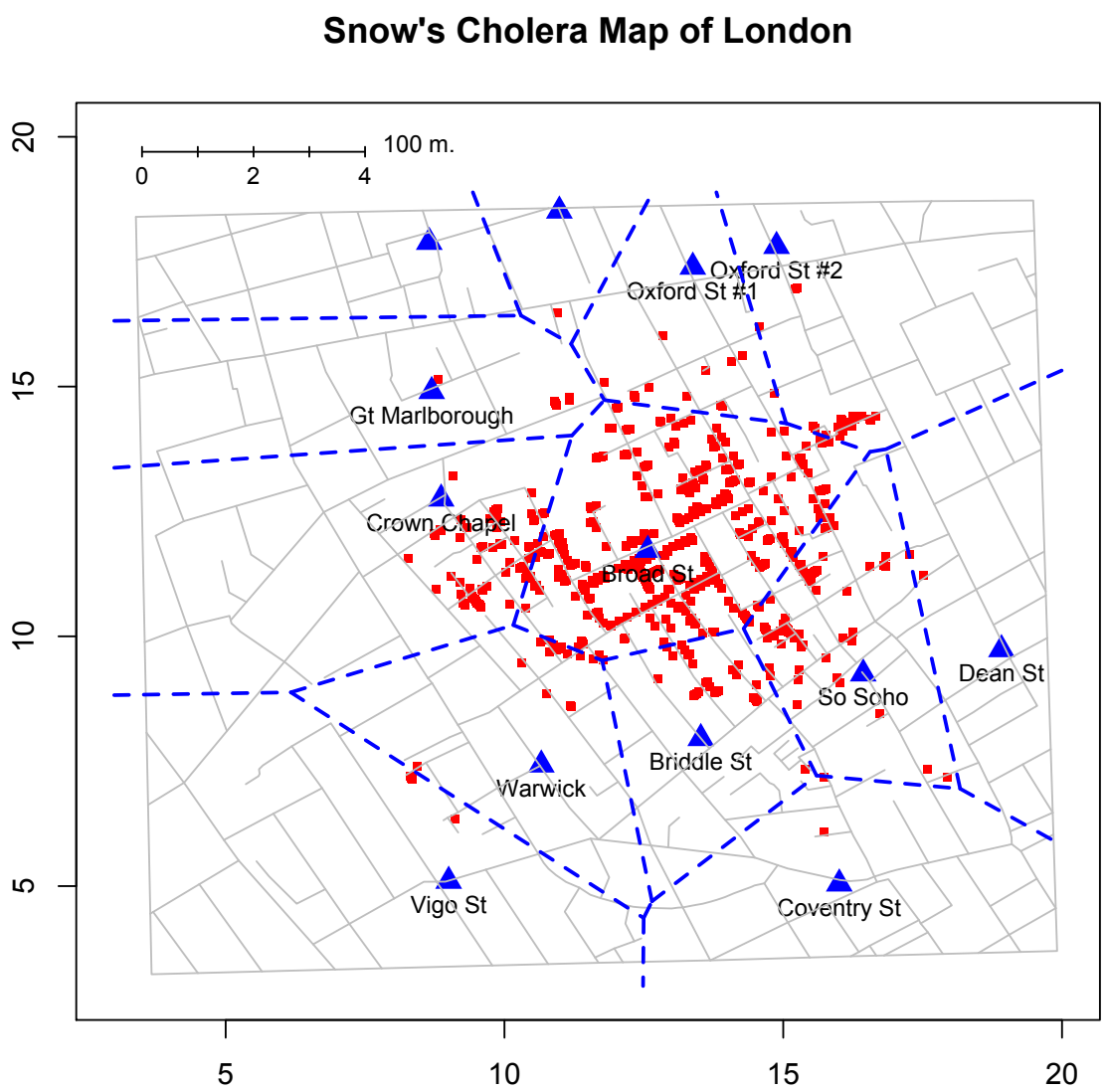
Definition: The set of all points x closest to pump A is the *Voronoi cell* of pump A

Modeling Assumptions:

Some (not all) pumps are contaminated

People use the closest pump

Snow's spatial analysis: A simple visual model (Tobler 1994)



Fact: For any spot x on the map, there is a closest pump A

Definition: The set of all points x closest to pump A is the *Voronoi cell* of pump A

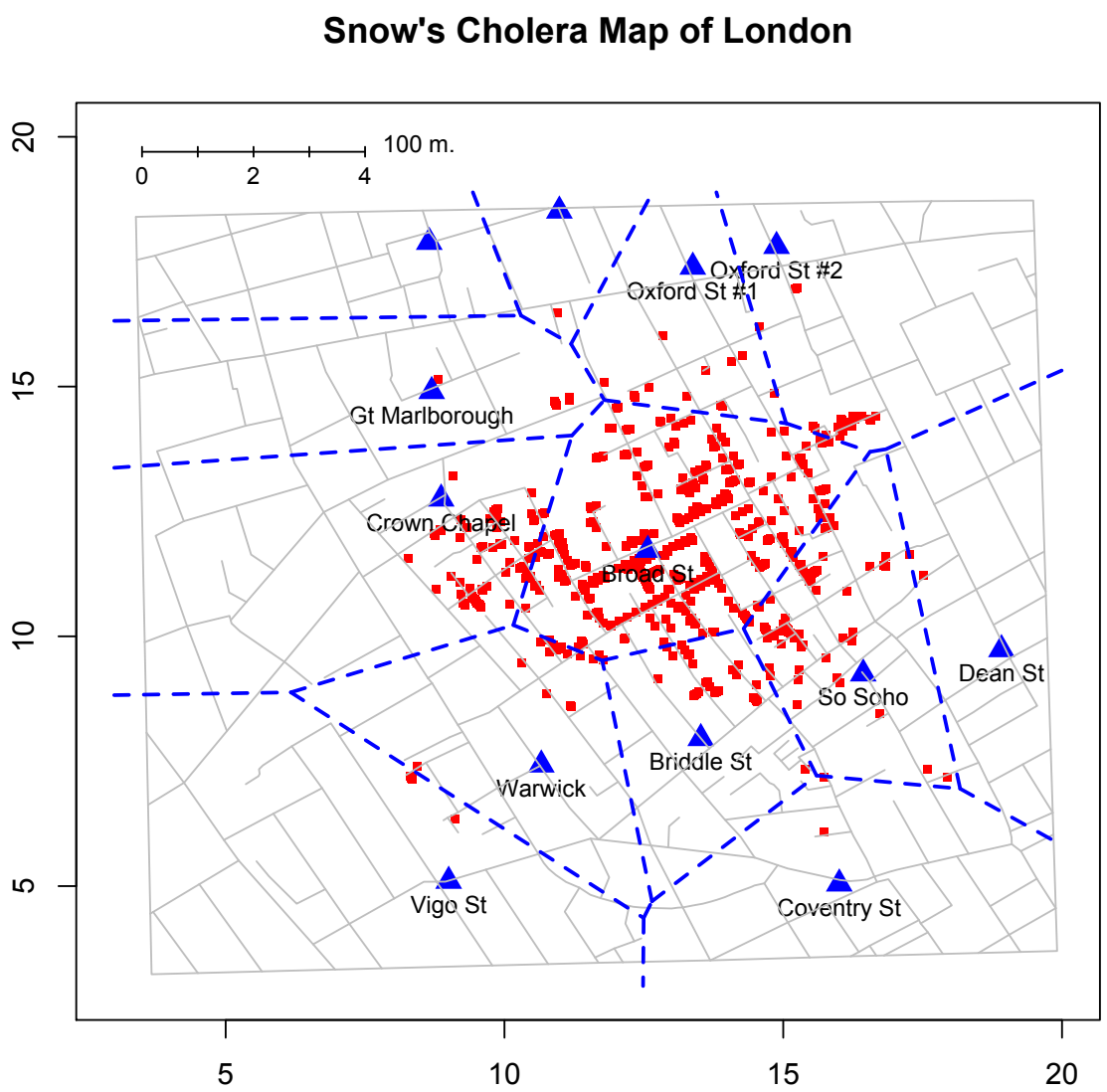
Modeling Assumptions:

Some (not all) pumps are contaminated

People use the closest pump

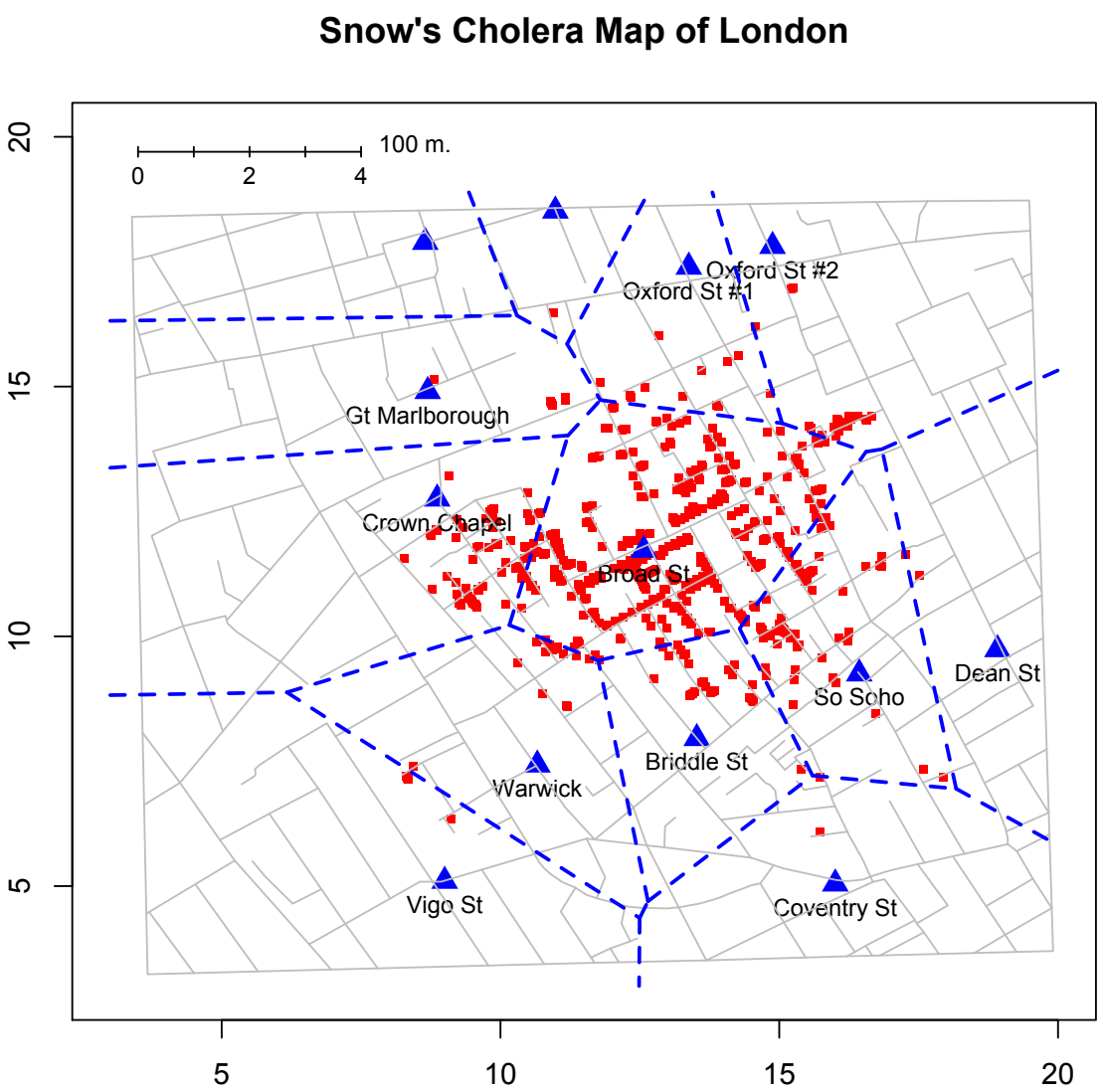
Model prediction: Pattern of deaths should match Voronoi cell boundaries

Snow's spatial analysis: A simple visual model (Tobler 1994)



Problems?

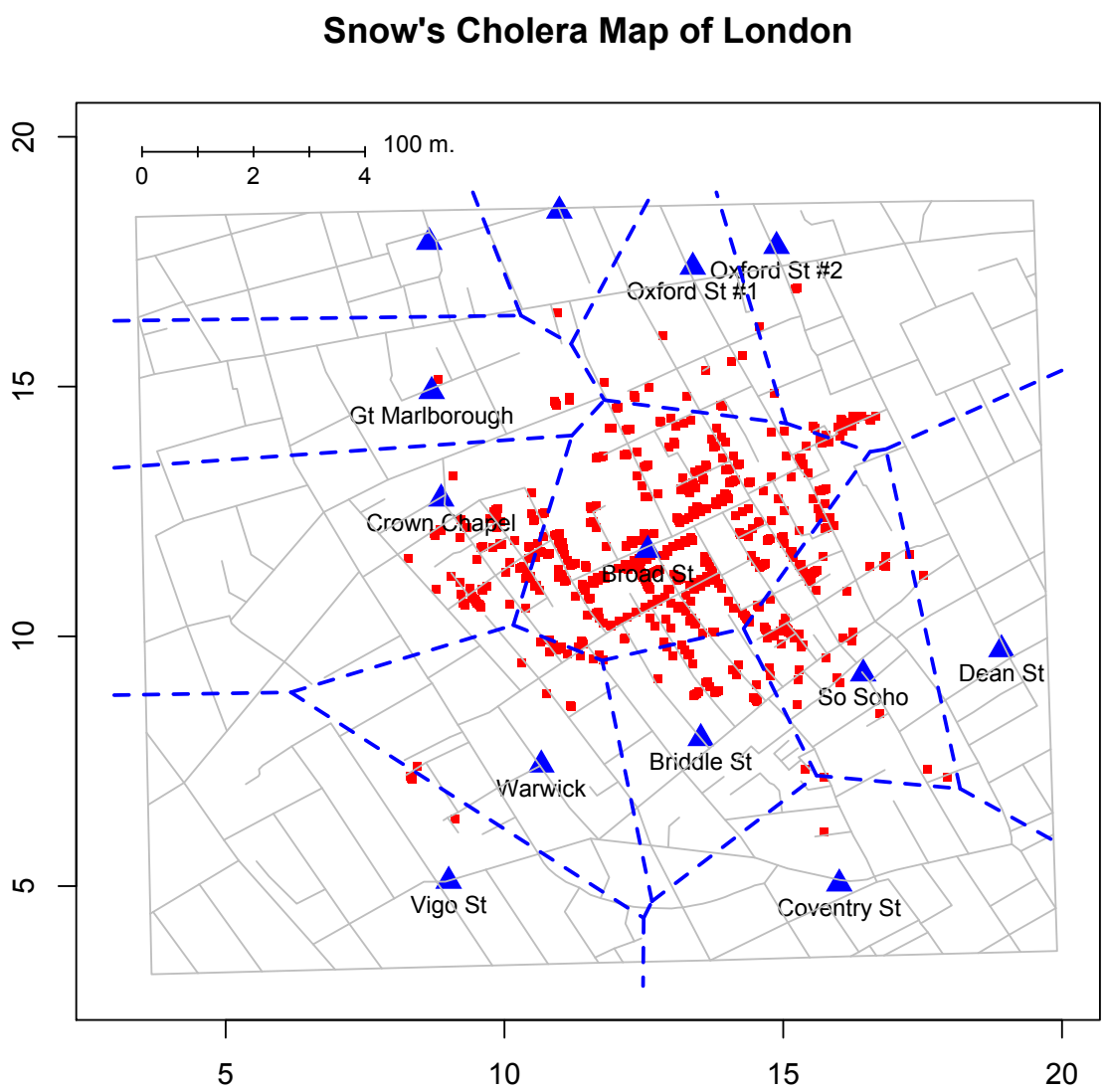
Snow's spatial analysis: A simple visual model (Tobler 1994)



Problems?

Distance in a city isn't really Euclidian – the built environment lengthens some paths.

Snow's spatial analysis: A simple visual model (Tobler 1994)

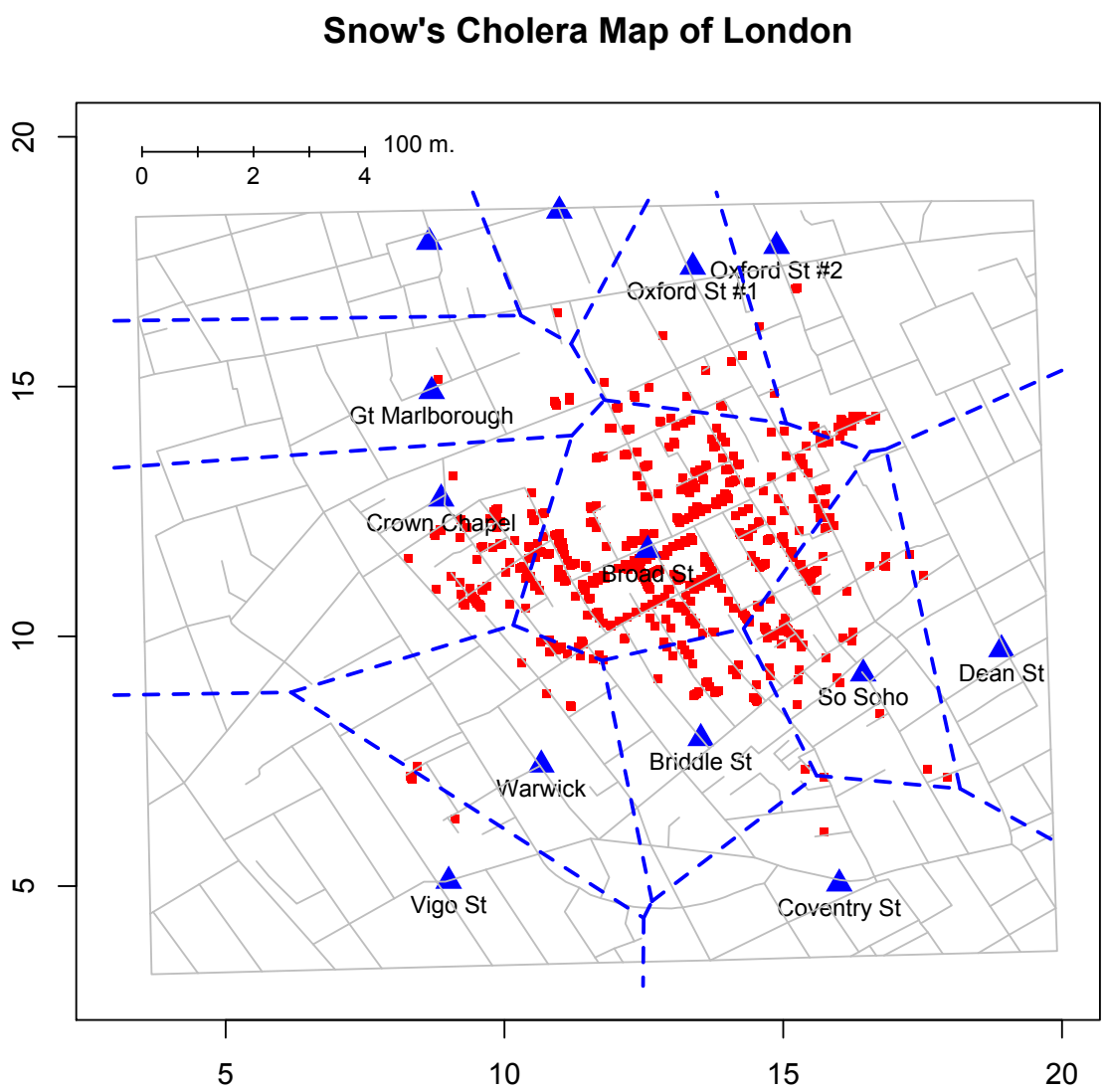


Problems?

Distance in a city isn't really Euclidian – the built environment lengthens some paths.

What about outliers? Can our theory be right if some cases lie outside Voronoi cell of Broad St. Pump?

Snow's spatial analysis: A simple visual model (Tobler 1994)



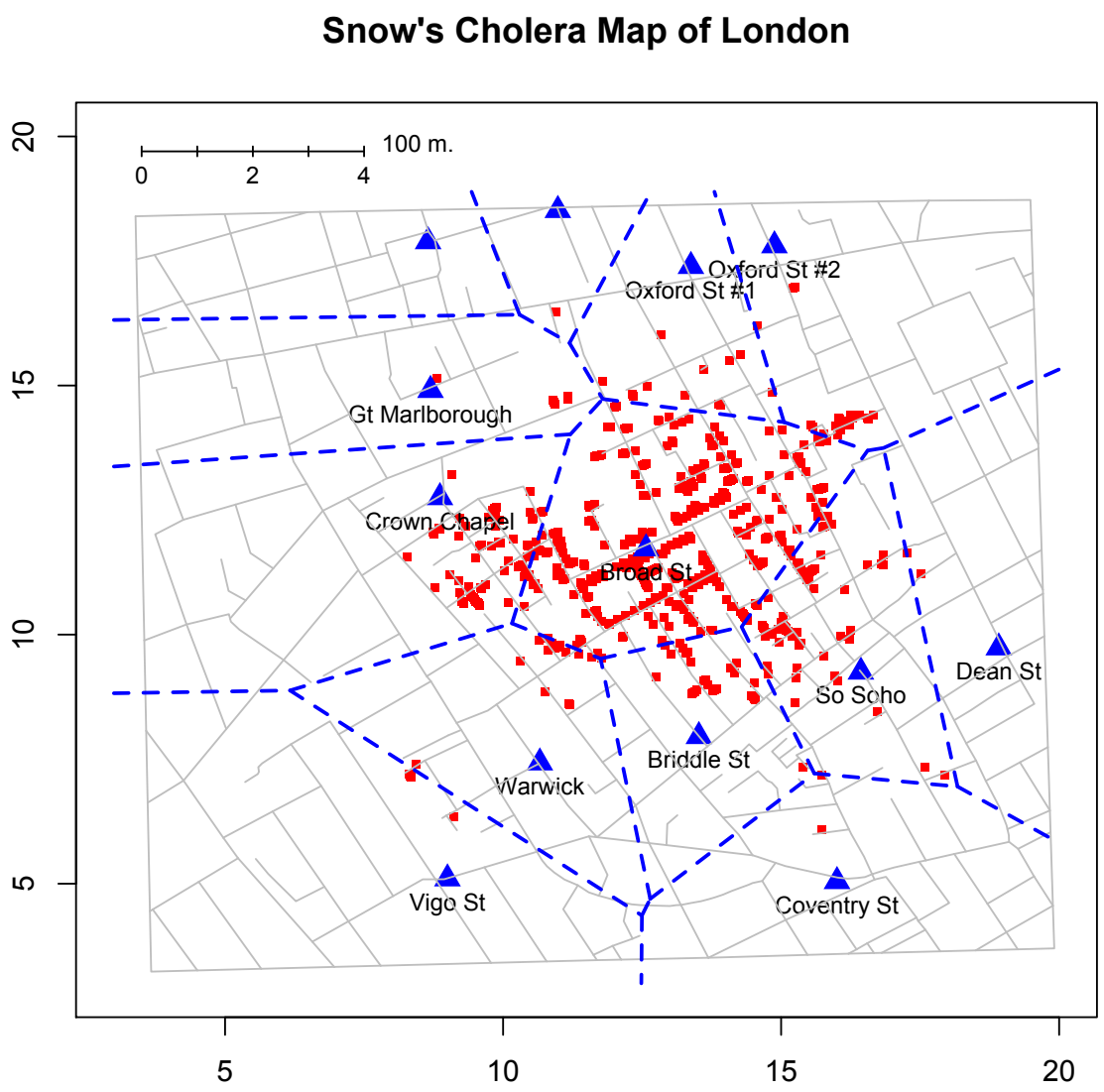
Problems?

Distance in a city isn't really Euclidian – the built environment lengthens some paths.

What about outliers? Can our theory be right if some cases lie outside Voronoi cell of Broad St. Pump?

Outliers could point to missing variables *or* simple randomness

Snow's spatial analysis: A simple visual model (Tobler 1994)



Problems?

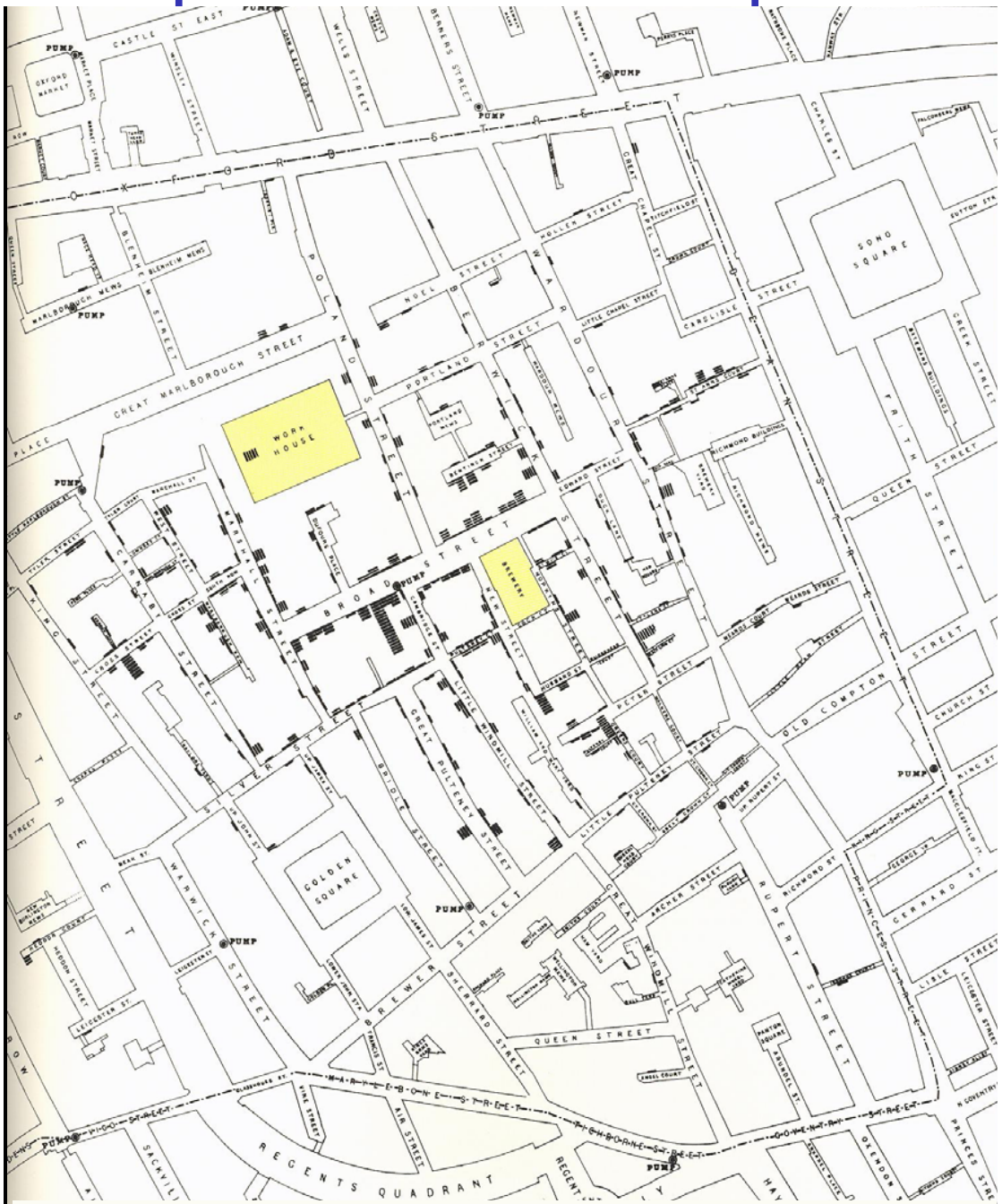
Distance in a city isn't really Euclidian – the built environment lengthens some paths.

What about outliers? Can our theory be right if some cases lie outside Voronoi cell of Broad St. Pump?

Outliers could point to missing variables *or* simple randomness

Is our model deterministic or probabilistic?

What explains outliers in this map?



Three cases:

1. A prison (work house) with its own well.
2. A brewery with its own water source. Saved by the beer.
3. Some distant deaths attributed to preference for Broad St. water.

John Snow stops the Cholera epidemic

Snow used his data and map to convince officials to remove the handle from the Broad Street pump.

Credited with stopping the outbreak and providing first experimental evidence for germs

Some questions to consider later:

1. Did the Broad Street Pump really cause the cholera outbreak?
2. Did removing the handle stop it?
3. Can we measure our uncertainty about our answers to 1 and 2?