Contagious Diseases in the United States:
Trends and Patterns in the Past 100 Years

INFSCI 2415 Information Visualization

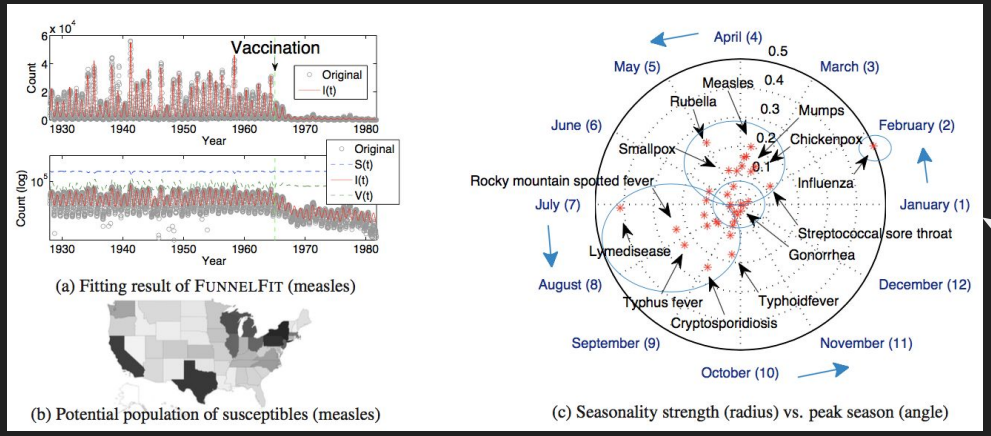
Group 5
Chen Song, Jiexiao He, Jingran Xie

Motivation

- Contagious diseases are always important issues in human history:
  - Black death killed 30-60% of Europe's population during 14th century;
  - Smallpox was a main factor in choosing the heir to the throne during Qing Dynasty in China.

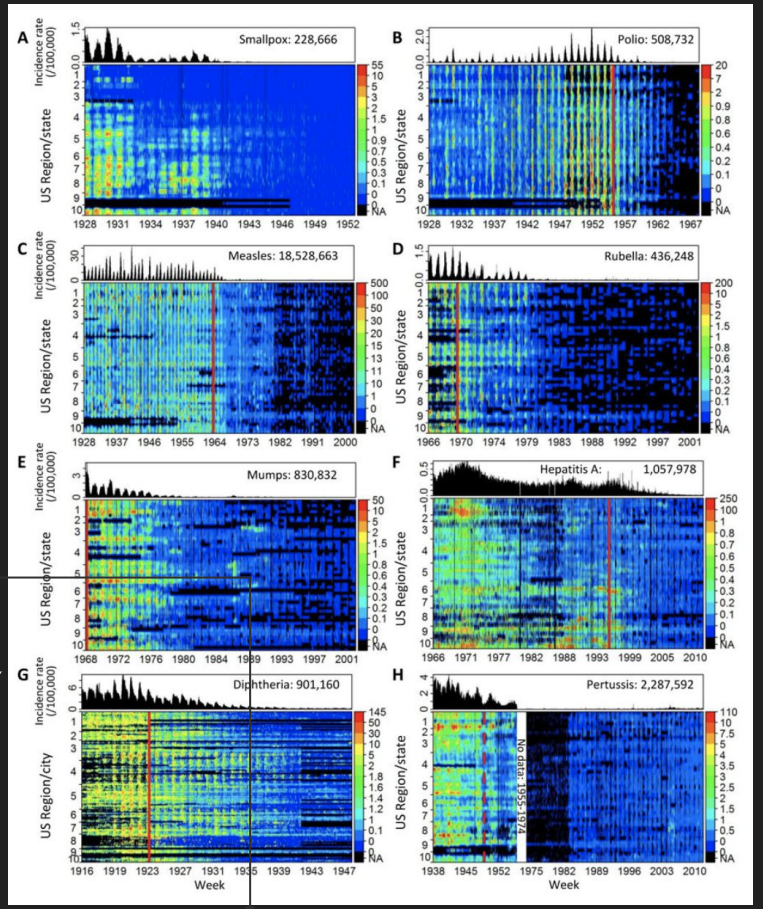
- It is important to understand the trends and patterns in contagious diseases spreads and preventions.

Related Work



Explored 5 properties of 56 contagious diseases: (P1) disease seasonality; (P2) disease reduction effect; (P3) local/state-level sensitivity; (P4) external shock events; (P5) detect incongruous values. (Matsubara, Yasuko, et al. 2014)

Explored the effects of vaccination: “declines in the incidence of contagious diseases in the United States over the past century. However, some contagious diseases are now on the rise despite the availability of vaccines. ” (Van Panhuis, Willem G., et al. 2013)



Data: Project Tycho Level 1 Data

- Project Tycho contains three datasets:
  - Level 1 data contains different types of counts of 8 diseases in 50 states and 122 cities from 1916 to 2010 which have been standardized in a common format.
  - Level 2 data contains informational counts of 50 diseases in 50 states and 1284 cities from 1888 to 2014 which have been reported in a common format.
  - Level 3 data contains different types of counts of 58 diseases and 81 disease subcategories in 3026 cities which have not been standardized.
- We use level 1 data to design and test our visualizations:
  - includes counts at the state level for smallpox, polio, measles, mumps, rubella, hepatitis A, and whooping cough, and at the city level for diphtheria.
  - 7 fields: epi_week, state, loc, loc_type, disease, cases, incidence_per_100000.

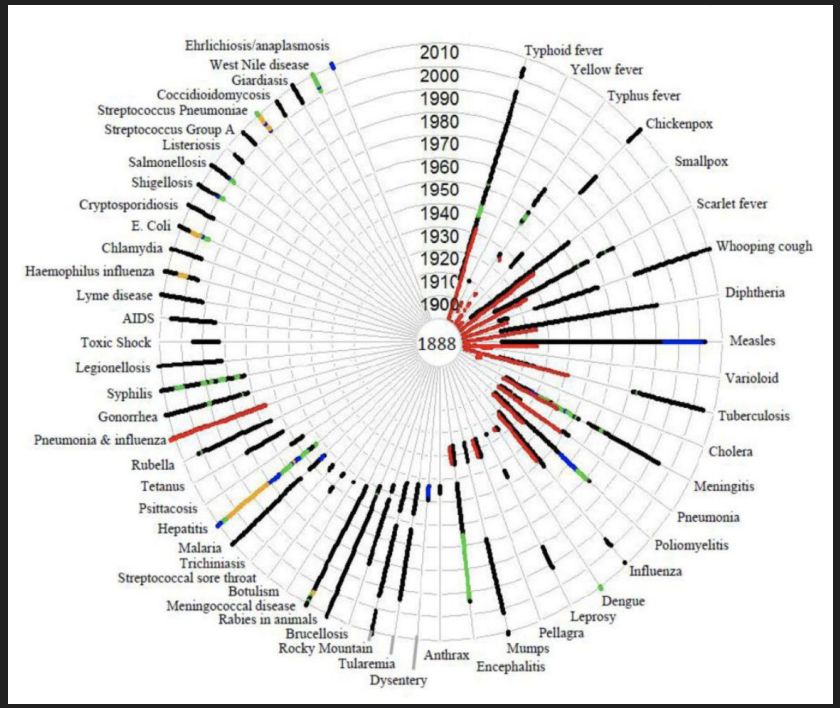
Project Goal

Reveal Trends and Patterns in Contagious Diseases' spread and prevention:

- **Trends**: diseases declined over time, mainly due to vaccination, but could be also significantly affected by general economic developments, i.e. availability of public health program or clear water.
- **Seasonal Patterns**: some diseases have peaks during warmest or coldest weather, while some other diseases may not be affected by weather.
- **Spatial Patterns**: some places are more vulnerable to a particular disease, maybe because of the local natural environment.

Challenges

- **Data availability**: 1) the spatial and temporal range covers every disease is different to each other, and 2) some diseases are already disappeared in the recent years, which give us a lot of zeros in the dataset.
- **Scalability**: the data size is relatively large (100+ years * 52 week/year * 8 diseases * 50+ states&cities = 2,000,000+ observations).



Van Panhuis, Willem G., et al. 2013

Data Preparation

- Separate and aggregate the data based on the visualization needs.
  - Handle missing data
  - Aggregate the cities to states for the disease has city level information
  - Aggregate 52 weeks to a year
  - Aggregate 50+ states & regions to the whole country
- Add GDP per capita as the economic issue that will influence the diseases.
- Rebuild data into proper format

Visual Design Overview:

Trend

Best encodings for time -series relationship: lines to emphasize the overall shape of the data.

Best encodings for correlation: points and a trend line in the form of a scatter plot.

Seasonal Patterns

What we want to seek: seasonality -- using cyclic arrangement to find seasonal patterns.

Spatial Patterns

Using map with animation to show the spatial patterns changing over time.

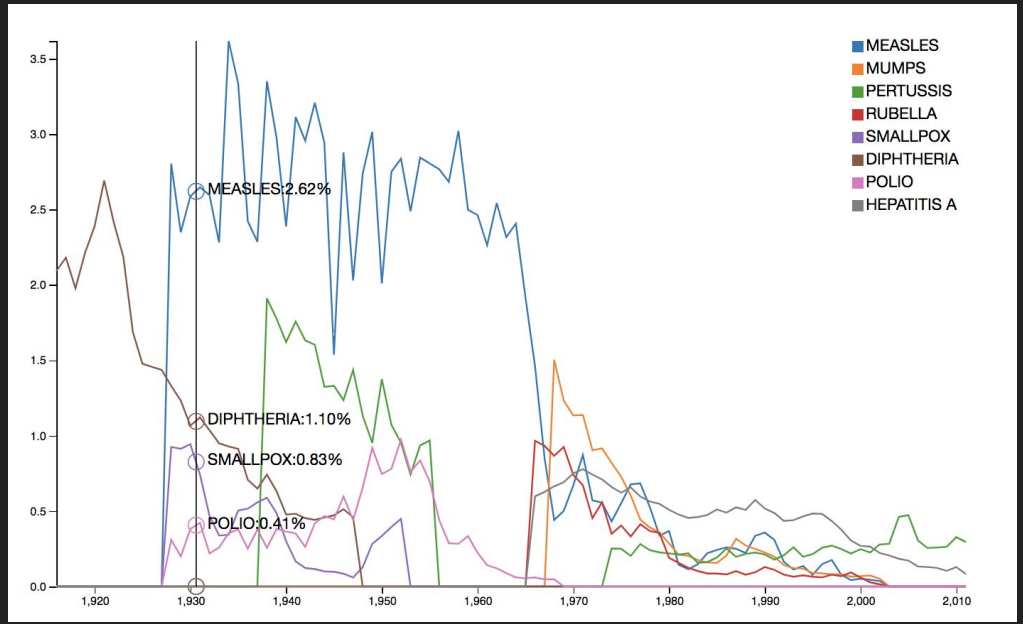
Still need to compare the trend in each state: small multiple line charts.

Trends in the Past 100 years:

- Disease declined due to vaccine usage: not only affected by the invention of vaccine, but also the availability of the vaccine to public;
- Disease declined due to economic development: financial ability to provide public health program, clear water, etc;
- Disease re-emerge because of some individuals refuse to get vaccinated after the risk declined

Trends in the Past 100 years:

An overview of diseases decline and re-emerge



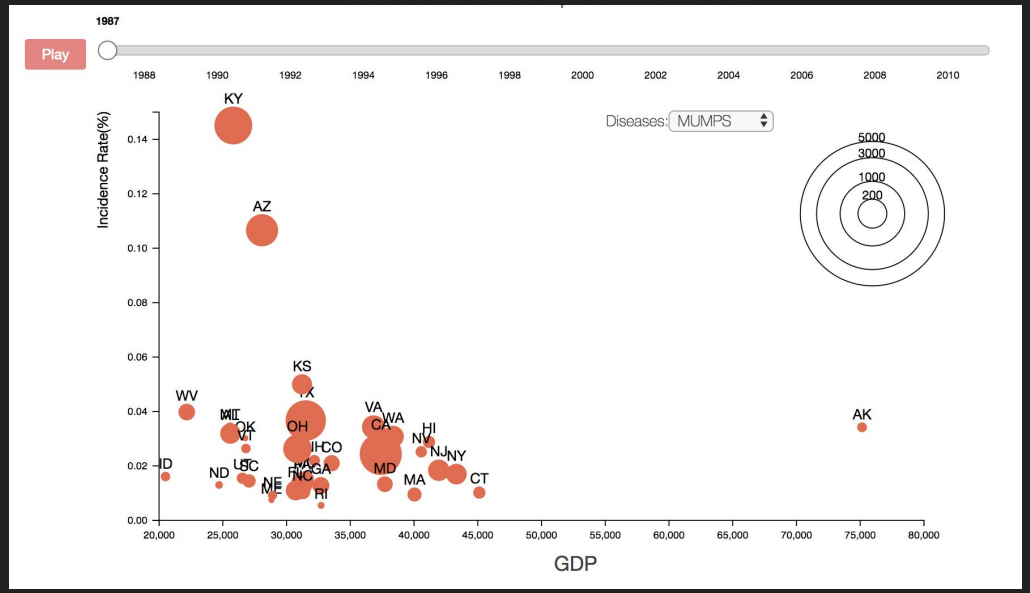
- Layout: multi line chart, each line represents a disease
- Visual encoding:
  - X-axis: time (year)
  - Y-axis: incidence rate (%)
  - Color: category of diseases
- Interaction:
  - Mouse over: detail number of each point

Discussion:

- Why multiple line chart
  - Line chart is a good model to show the general trend of the disease
  - Advantage of multiple line chart: state that has the highest value of rate can be shown easily. Available to see the disease re-emerge
  - Disadvantage of multiple line chart: difficult to show the accuracy rate when select a specific line.
- Special challenge: Each disease not present in a same year or rate range.
  - In order to show the year the diseases occurred, each disease's year and rate is different and adjust to the data size.
- Data scalability issue:
  - Pre-aggregate the data to get the cases of the whole year instead of the week.
  - Disadvantage: Some states have low disease rate. Many lines aggregate in a small area and it is hard to select.

Trends in the Past 100 years:

The relationship between diseases and economy

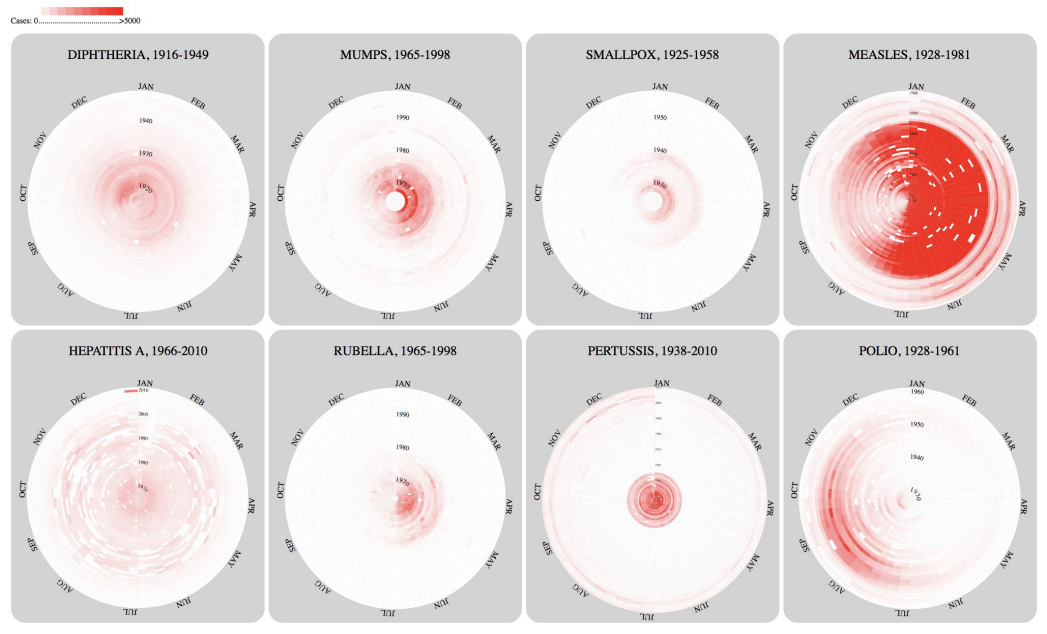


- Layout: scatterplot, each scatter represent the value of a disease in a state
- Visual encoding:
  - X-axis: GDP per capita($)
  - Y-axis: incidence rate(%)
  - Size: cases
  - Text label: state name
  - Animation: time
- Interaction:
  - Slide: time
  - Select box: category of diseases

Discussion:

- Why scatterplot
  - Line chart is a good model to show the general trend of the disease
  - Advantage of multiple line chart: state that has the highest value of rate can be shown easily. Available to see the disease re-emerge
  - Disadvantage of multiple line chart: difficult to show the accuracy rate when select a specific line.
- Special challenge: Each disease not present in a same year or rate range.
  - In order to show the year the diseases occurred, each disease's year and rate is different and adjust to the data size.
- Data scalability issue:
  - Pre-aggregate the data to get the cases of the whole year instead of the week.
  - Disadvantage: Some states have low disease rate. Many lines aggregate in a small area and it is hard to select.

Seasonal Patterns:



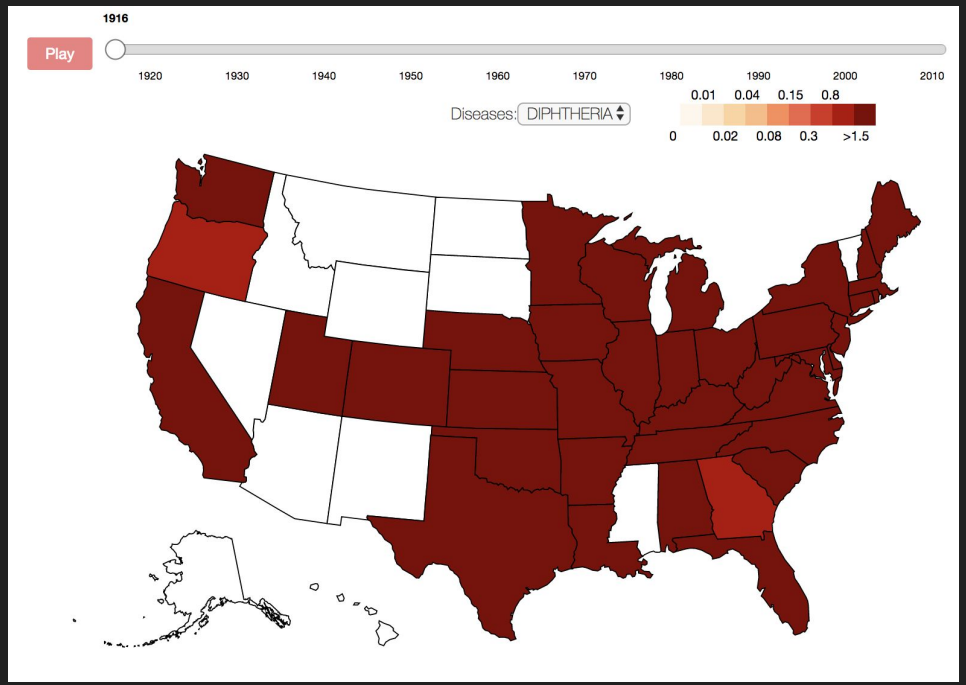
- Layout: multiple circular heat maps
- Visual encoding:
  - Angle: same week of each year
  - Radius: year
  - Color: cases across the States
- Interaction
  - Click to switch between small multiple and detail individual graphs
  - Mouse over in detail graphs: detail information of each grid

Discussion:

- Why circular heat map
  - Heat map is one of the best way to find clusters
  - Advantage of circular heat map: easily to find repeated patterns even the pattern appears at the beginning or ending of each time interval
  - Disadvantage of circular heat map: uneven size of each grid
- Why small multiples: compare across diseases
- Special challenge: ranges of cases are very different across the diseases
  - In Measles the max is over 50,000, while in other diseases the max is about 4000-7000
  - Use [0, 5000] as domain in color encoding
- Data availability issue: each circular heat map covers different time interval
- Data scalability issue:
  - Pre-aggregate the data to get the cases in the whole country instead of in each state
  - Disadvantage: each state could have different seasonal patterns due to the local natural environment, but this visualization can not show the difference among states

Spatial Patterns:

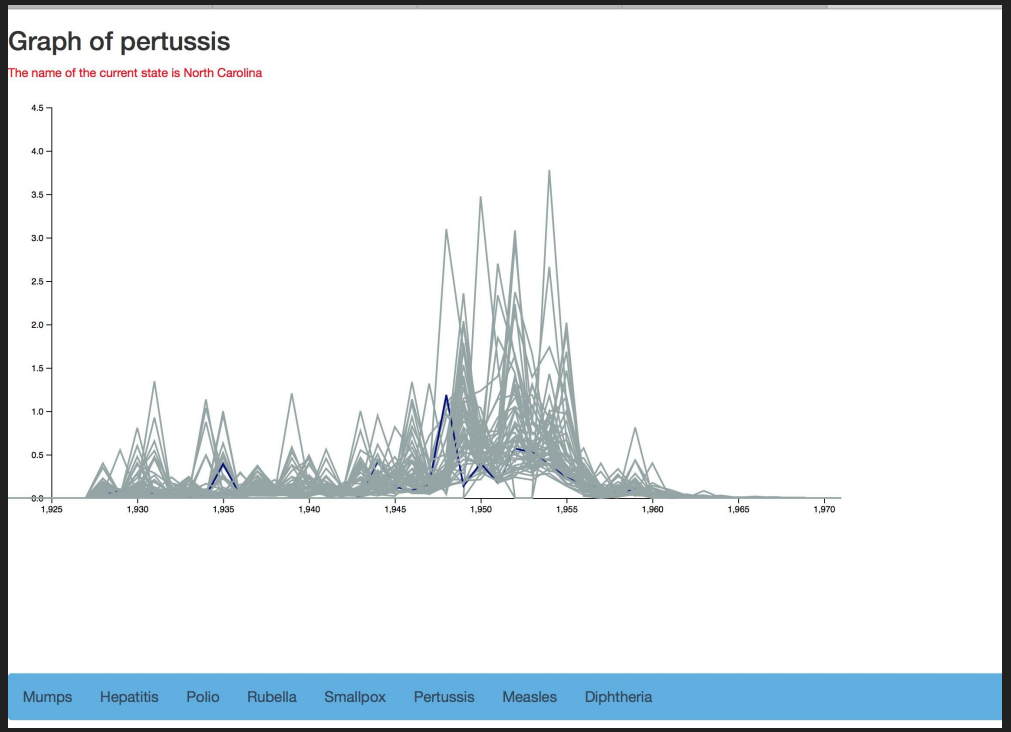
An overview through map



- Layout: map
- Visual encoding:
  - X & Y: geo-spatial location
  - Color: incidence
  - Animation: time
- Interaction:
  - Mouse over: detail information of each state in a year
  - Mouse over: on legend highlight states of that value

Spatial Patterns:

Compare the trends across 50 states



- Layout: multiple line chart with menu bar, each diseases have the bar chart line
- Visual encoding:
  - X-axis: time
  - Y-axis: rate of the disease
- Interaction:
  - Mouseover: information of the current state, the color of the clicked state will change.

Discussion:

- Why multiple line chart
  - Line chart is a good model to show the general trend of the disease within 50 states.
  - Advantage of multiple line chart: state that has the highest value of rate can be shown easily. Available to see the disease re-emerge
  - Disadvantage of multiple line chart: difficult to show the accuracy rate when select a specific line.
- Special challenge: Each disease not present in a same year or rate range.
  - In order to show the year the diseases occurred, each disease's year and rate is different and adjust to the data size.
- Data scalability issue:
  - Pre-aggregate the data to get the cases of the whole year instead of the week.
  - Disadvantage: Some states have low disease rate. Many lines aggregate in a small area and it is hard to select.

Future Work:

- Try to link the visualizations in a same topic:
  - Link the scatterplot and the line chart so that the time slide and control both of them
  - Link the map and the small multiple line chart so that when select and highlight a state, the small line chart of the state could also be highlighted
- Add narratives about brief history of the diseases:
  - Explosions in the history
  - Vaccine innovation and usage
  - Important historical events, i.e. World War II
- Try the same thing using Level 2 data

Reference:

1. Project Tycho: https://www.tycho.pitt.edu/.
2. GDP per capita: Bureau of Economic Analysis (https://bea.gov).
3. Matsubara, Yasuko, et al. "FUNNEL: automatic mining of spatially coevolving epidemics." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.
4. Van Panhuis, Willem G., et al. "Contagious diseases in the United States from 1888 to the present." The New England journal of medicine 369.22 (2013): 2152.

THE END. THANK YOU!