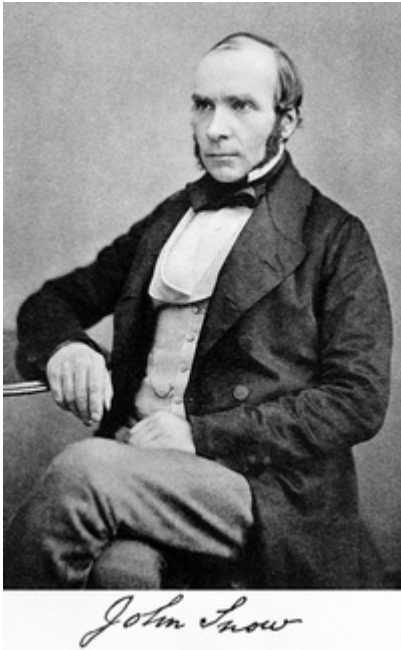


# 1. Dr. John Snow



Dr. John Snow (1813-1858) was a famous British physician and is widely recognized as a legendary figure in the history of public health and a leading pioneer in the development of anesthesia. Some even say one of the greatest physicians of all time.

As a leading advocate of both anesthesia and hygienic practices in medicine, he not only experimented with ether and chloroform but also designed a mask and method how to administer it. He personally administered chloroform to Queen Victoria during the births of her eighth and ninth children, in 1853 and 1857, which assured a growing public acceptance of the use of anesthetics during childbirth.

But, as we will show later, not all his life was just a success. John Snow is now also recognized as one of the founders of modern epidemiology (*some also consider him as the founder of data visualization, spatial analysis, data science in general, and many other related fields*) for his scientific and pretty modern data approach in identifying the source of a cholera outbreak in Soho, London in 1854, but it wasn't always like this. In fact, for a long time, he was simply ignored by the scientific community and currently is very often mythified.

In this notebook, we're not only going to rediscover his "data story", but reanalyze the data that he collected in 1854 and recreate his famous map (also called The Ghost Map).

```
In [1]: # Loading in the pandas module
import pandas as pd

# Reading in the data
deaths = pd.read_csv('datasets/deaths.csv')

# Print out the shape of the dataset
deaths.shape

# Printing out the first 5 rows
deaths.head(5)
```

Out[1]:

	Death	X coordinate	Y coordinate
0	1	51.513418	-0.137930
1	1	51.513418	-0.137930
2	1	51.513418	-0.137930
3	1	51.513361	-0.137883
4	1	51.513361	-0.137883

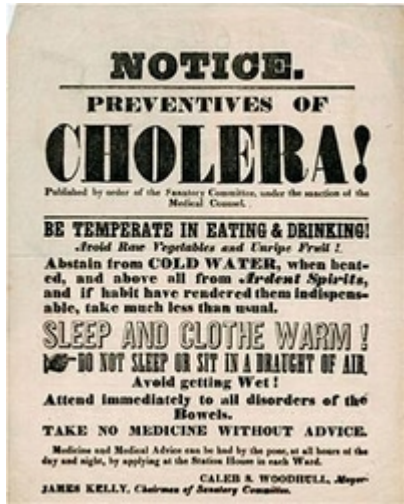
```
In [0]: %%nose

import pandas as pd

def test_deaths_correctly_loaded():
    correct_deaths = pd.read_csv('datasets/deaths.csv')
    assert correct_deaths.equals(deaths), 'The variable deaths should contain the data in datasets/deaths.csv'

def test_deaths_exists():
    assert "deaths" in globals(), \
        'The variable deaths should be defined.'
```

## 2. Cholera attacks!



Prior to John Snow's discovery cholera was a regular visitor to London's overcrowded and unsanitary streets. During the time of the third cholera outbreak, it was one of the most studied subjects (between years 1839-1856 over 700 studies and essays were published in London alone) and nearly all of the authors believed the outbreaks were due to miasma or "bad air".

It was John Snow's pioneering work with anesthesia and gases that made him doubt the miasma model of the disease. Originally he formulated and published his theory that cholera is spread by water or food in an essay On the Mode of Communication of Cholera (before the outbreak in 1849). The essay received negative reviews in the Lancet and the London Medical Gazette.

We know now that he was right, but Dr. Snow's dilemma was how to prove it? His first step to getting there was checking the data. Our dataset has 489 rows of data in 3 columns but to work with dataset more easily we will first make few changes.

```
In [2]: # Summarizing the content of deaths
deaths.info()

# Define the new names of your columns
newcols = {
    'Death': 'death_count',
    'X coordinate': 'x_latitude',
    'Y coordinate': 'y_longitude'
}

# Rename your columns
deaths.rename(columns=newcols, inplace=True)

# Describe the dataset
deaths.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 489 entries, 0 to 488
Data columns (total 3 columns):
Death          489 non-null int64
X coordinate    489 non-null float64
Y coordinate    489 non-null float64
dtypes: float64(2), int64(1)
memory usage: 11.5 KB
```

Out[2]:

	death_count	x_latitude	y_longitude
count	489.0	489.000000	489.000000
mean	1.0	51.513398	-0.136403
std	0.0	0.000705	0.001503
min	1.0	51.511856	-0.140074
25%	1.0	51.512964	-0.137562
50%	1.0	51.513359	-0.136226
75%	1.0	51.513875	-0.135344
max	1.0	51.515834	-0.132933

```
In [0]: %%nose

def test_example():
    assert newcols == {
        'Death': 'death_count',
        'X coordinate': 'x_latitude',
        'Y coordinate': 'y_longitude'
    }, \
        'The columns should have the new names (defined by the list/new dictionary).'
```

```
def test_deaths_exists():
    assert "deaths" in globals(), \
        'The variable deaths should be defined/data loaded as DataFrame.'
```

```
def test_deaths_rename():
    correct_deaths_rename = deaths.rename(columns=newcols, inplace=True), \
        'The DataFrame deaths should contain new names of the columns x_latitude and y_longitude.'
```

### 3. You know nothing, John Snow!



It was somehow unthinkable that one man could debunk the miasma theory and prove that all the others got it wrong, so his work was mostly ignored. His medical colleagues simply said: "You know nothing, John Snow!"

As already mentioned John Snow's first attempt to debunk the "miasma" theory ended with negative reviews. However, a reviewer made a helpful suggestion in terms of what evidence would be compelling: the crucial natural experiment would be to find people living side by side with lifestyles similar in all respects except for the water source. The cholera outbreak in Soho, London in 1854 gave Snow the opportunity not only to save lives this time but also to further test and improve his theory. But what about the final proof that he was right?

We now know how John Snow did it, so let's get the data right first.

```
In [3]: # Create `locations` by subsetting only x_latitude and y_longitude from the dataset
locations = deaths[['x_latitude', 'y_longitude']]

# Create `deaths_list` by transforming the DataFrame to list of lists
deaths_list = locations.values.tolist()

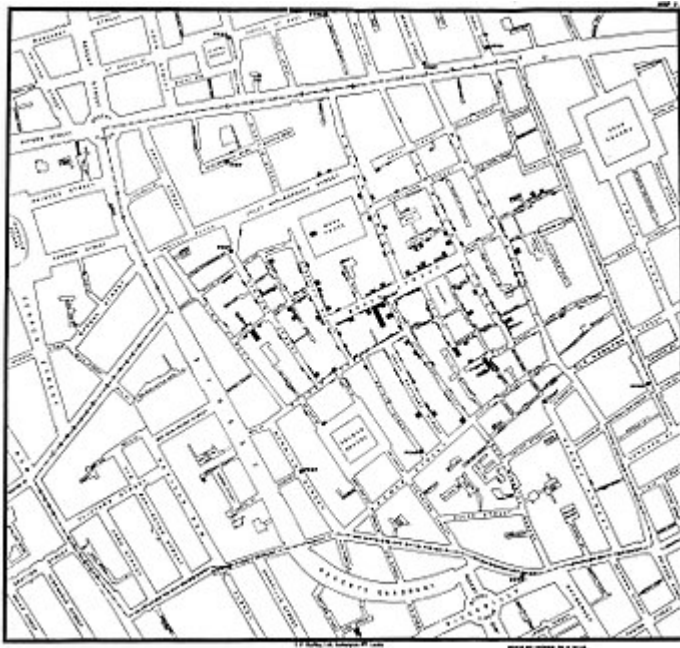
# Check the length of the list
len_deaths_list = len(deaths_list)
print(f"Number of (x_latitude, y_longitude) pairs: {len_deaths_list}")

Number of (x_latitude, y_longitude) pairs: 489
```

```
In [0]: %%nose

def test_example():
    assert deaths_list == locations.values.tolist(), \
        'The columns should have 489 combinations of x,y coordinates.'
    assert len(deaths_list) == 489, \
        'The length of deaths_list should be 489.'
```

## 4. The Ghost Map



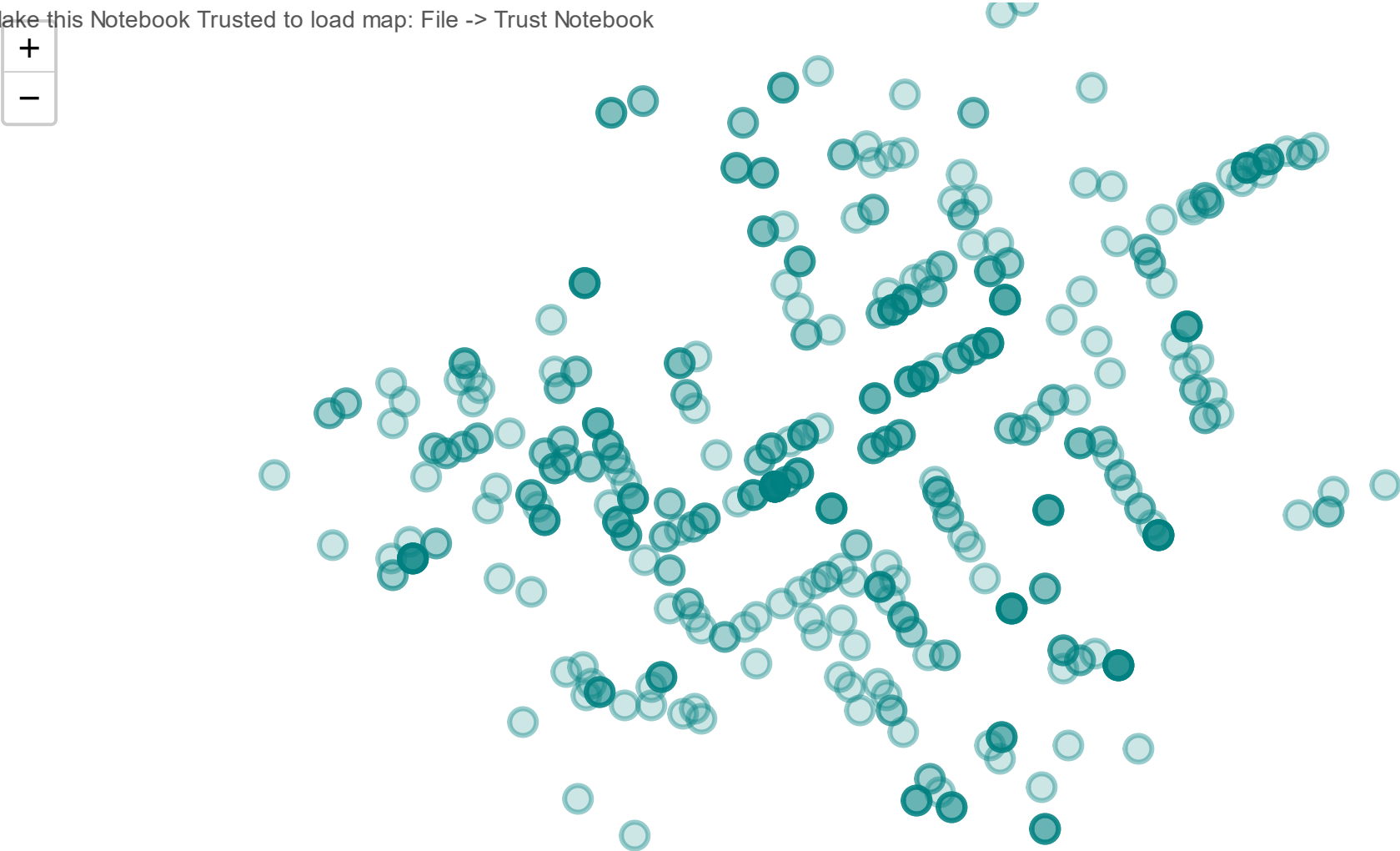
His original map, unfortunately, is not available (it might never even existed). We can see the famous one that he drew about a year later in 1855, though, and it is displayed in this cell. Because the map depicts and visualizes the deaths sometimes it is called also **The Ghost Map**.

We now know how John Snow did it and have the data too, so let's recreate his map using modern techniques.

```
In [6]: # Plot the data on map (map location is provided) using folium and for loop for plotting all the points
import folium

map = folium.Map(location=[51.5132119,-0.13666], tiles='Stamen Toner', zoom_start=17)
for point in range(0, len(deaths_list)):
    folium.CircleMarker(deaths_list[point], radius=8, color='teal', fill=True, fill_color='teal', opacity = 0.4).add_to(map)
map
```

Out[6]: Make this Notebook Trusted to load map: File -> Trust Notebook



In [0]: %%nose

```
def test_example():
    assert deaths_list == locations.values.tolist(), \
        'The columns should have 489 combinations of x,y coordinates.'
    assert len(deaths_list) == 489, \
        'The length of deaths_list should be 489.'
```

## 5. It's the pump!



After marking the deaths on the map, what John Snow saw was not a random pattern (we saw this on our recreation of The Ghost Map too). The majority of the deaths were concentrated at the corner of Broad Street (now Broadwick Street) and Cambridge Street (now Lexington Street). A cluster of deaths around the junction of these streets was the epicenter of the outbreak, but what was there? Yes, a water pump.

John Snow at the time already had a developed theory that cholera spreads through water, so to test this he marked on the map also the locations of the water pumps nearby. And here it was, the whole picture.

By combining the location of deaths related to cholera with locations of the water pumps, Snow was able to show that the majority were clustered around one particular public water pump in Broad Street, Soho. Finally, he had the proof that he needed.

We will now do the same and add the locations of the pumps to our recreation of The Ghost Map.

```
In [7]: # Import the data
pumps = pd.read_csv('datasets/pumps.csv')

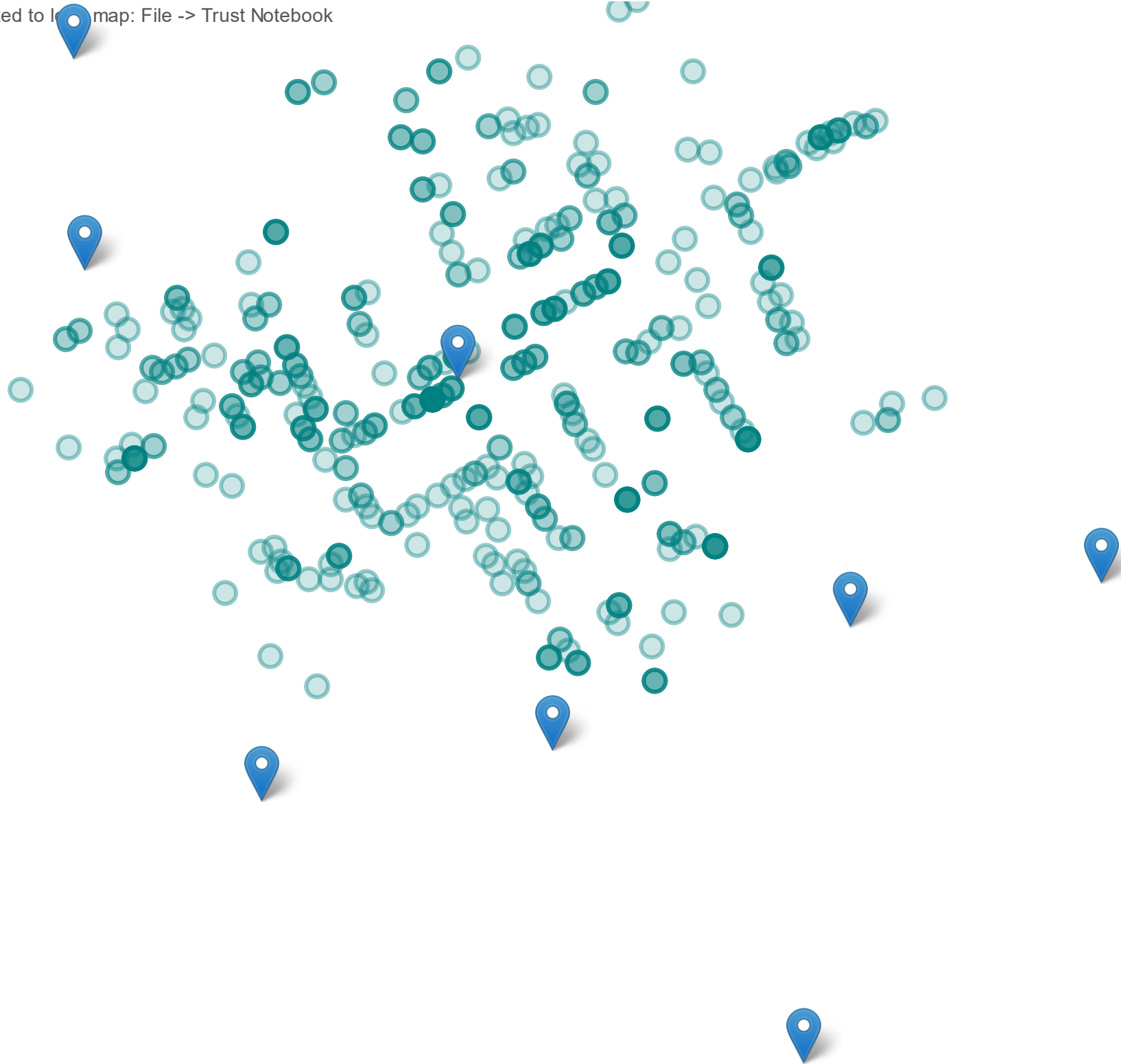
# Subset the DataFrame and select just ['X coordinate', 'Y coordinate'] columns
locations_pumps = pumps[['X coordinate', 'Y coordinate']]

# Transform the DataFrame to list of lists in form of ['X coordinate', 'Y coordinate'] pairs
pumps_list = locations_pumps.values.tolist()

# Create a for loop and plot the data using folium (use previous map + add another layer)
map1 = map
for point in range(0, len(pumps_list)):
    folium.Marker(pumps_list[point], popup=pumps['Pump Name'][point]).add_to(map1)
map1
```



Out[7]: Make this Notebook Trusted to load map: File -> Trust Notebook



Leaflet (<http://leafletjs.com>)

In [0]: %%nose

```
def test_example():
    assert deaths_list == locations.values.tolist(), \
        'The columns should have 489 combinations of x,y coordinates.'
    assert len(deaths_list) == 489, \
        'The length of deaths_list should be 489.'
    assert pumps_list == locations_pumps.values.tolist(), \
        'The columns should have 8 combinations of x,y coordinates.'
    assert len(pumps_list) == 8, \
        'The length of deaths_list should be 8.'
```

## 6. You know nothing, John Snow! (again)



So, John Snow finally had his proof that there was a connection between deaths as a consequence of the cholera outbreak and the public water pump that was probably contaminated. But he didn't just stop there and investigated further.

He was looking for anomalies now (we would now say "outliers in data") and found two in fact where there were no deaths. First was brewery right on the Broad Street, so he went there and learned that they drank mostly beer (in other words not the water from the local pump, which confirms his theory that the pump is the source of the outbreak). The second building without any deaths was workhouse near Poland street where he learned that their source of water was not the pump on the Broad Street (confirmation again). The locations of both buildings are visualized also on the map on the left.

He was now sure, and although officials did not trust him nor his theory they removed the handle to the pump next day, 8th of September 1854. John Snow later collected and published in his famous book also all the data about deaths in chronological order, before and after the peak of the outbreak and we will now analyze and compare the effect when the handle was removed.

```
In [8]: # Import the data the right way
dates = pd.read_csv('datasets/dates.csv', parse_dates=['date'])

# Set the Date when the handle was removed (8th of September 1854)
handle_removed = pd.to_datetime('1854-09-08')

# Create new column `day_name` in `dates` DataFrame with names of the day
dates['day_name'] = dates['date'].dt.day_name()

# Create new column `handle` in `dates` DataFrame based on a Date the handle was removed
dates['handle'] = dates['date'] > handle_removed

# Check the dataset and datatypes
dates.info()

# Create a comparison of how many cholera deaths and attacks there were before and after the handle was removed
comparison = dates.groupby(['handle']).sum()
comparison
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43 entries, 0 to 42
Data columns (total 6 columns):
order      43 non-null int64
date       43 non-null datetime64[ns]
attacks    43 non-null int64
deaths     43 non-null int64
day_name   43 non-null object
handle     43 non-null bool
dtypes: bool(1), datetime64[ns](1), int64(3), object(1)
memory usage: 1.8+ KB
```

Out[8]:

	order	attacks	deaths
handle			
False	231	528	500
True	715	43	116

```
In [0]: %%nose

def test_dates_exists():
    assert "dates" in globals(), \
        "The variable dates should be defined."

def test_dates_correctly_loaded():
    correct_dates = pd.read_csv('datasets/dates.csv', parse_dates=['date'])
    handle_removed = pd.to_datetime('1854/9/8')
    correct_dates['day_name'] = correct_dates.date.dt.weekday_name
    correct_dates['handle'] = correct_dates.date > handle_removed
    assert correct_dates.equals(dates), "The variable `dates` should contain the data in dates.csv"

def test_date_correctly_converted():
    assert dates.date.dtype == pd.to_datetime(pd.Series("1854/9/8")).dtype, \
        "The column date should be converted using the pd.to_datetime() function"
```

## 7. The picture worth a thousand words



Removing the handle from the pump prevented any more of the infected water from being collected. The spring below the pump was later found to have been contaminated with sewage. This act was later recognized as an early example of epidemiology, public health medicine and the application of science (the germ theory of disease) in a real-life crisis.

A replica of the pump, together with an explanatory and memorial plaque and without a handle was erected in 1992 near the location of the original close to the back wall of what today is the John Snow pub. The site is subtly marked with a pink granite kerbstone in front of a small wall plaque.

We can learn a lot from John Snow's data. We can take a look at absolute counts, but this observation could lead us to a wrong conclusion so let's take a different look on the data using Bokeh.

Thanks to John Snow we have the data in chronological order (i.e. as time series data), so the best way to see the whole picture is to visualize it and look at it the way he saw it while writing *On the Mode of Communication of Cholera* (1855).

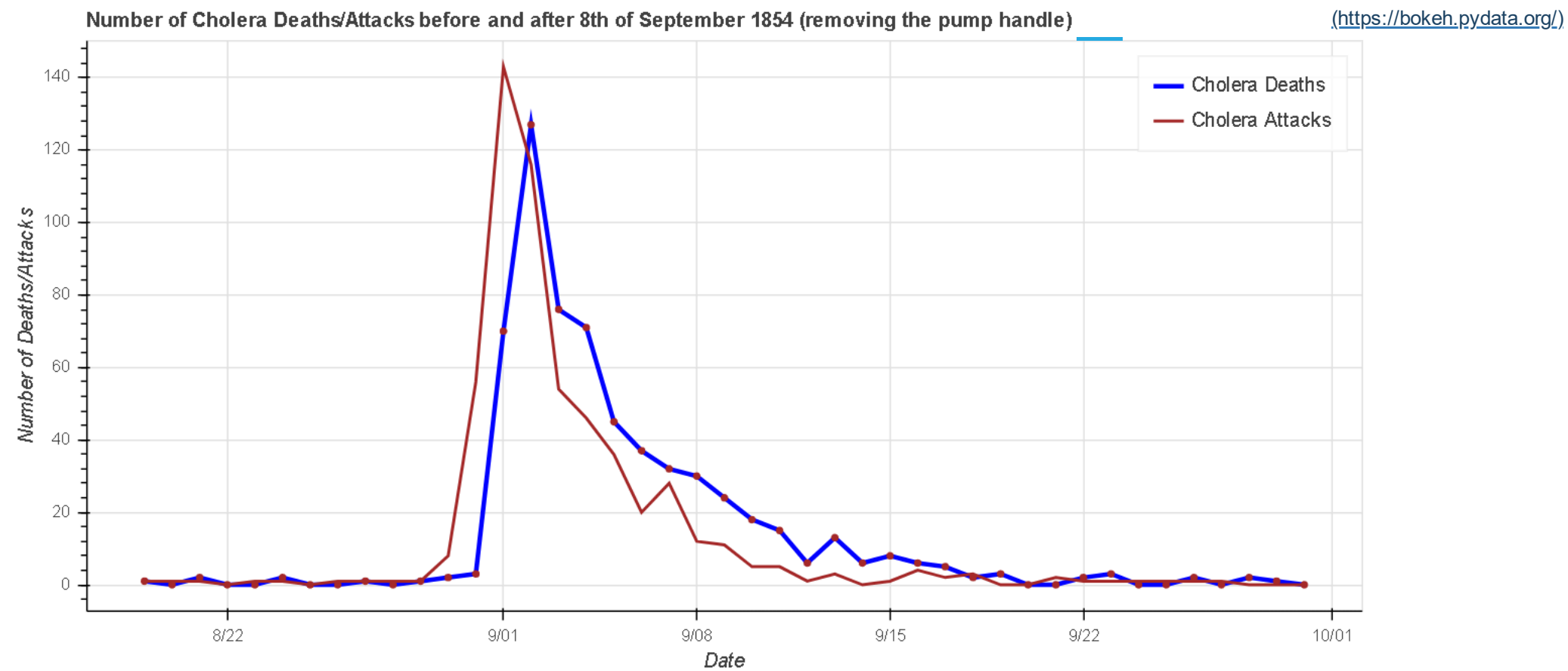
```
In [9]: import bokeh
from bokeh.plotting import output_notebook, figure, show
output_notebook(bokeh.resources.INLINE)

# Set up figure
p = figure(plot_width=900, plot_height=450, x_axis_type='datetime', tools='lasso_select, box_zoom, save, reset, wheel_zoom',
          toolbar_location='above', x_axis_label='Date', y_axis_label='Number of Deaths/Attacks',
          title='Number of Cholera Deaths/Attacks before and after 8th of September 1854 (removing the pump handle)')

# Plot on figure
p.line(dates['date'], dates['deaths'], color='blue', alpha=1, line_width=3, legend='Cholera Deaths')
p.circle(dates['date'], dates['deaths'], color='brown', nonselection_fill_alpha=0.2, nonselection_fill_color='grey')
p.line(dates['date'], dates['attacks'], color='brown', alpha=1, line_width=2, legend='Cholera Attacks')

show(p)
```

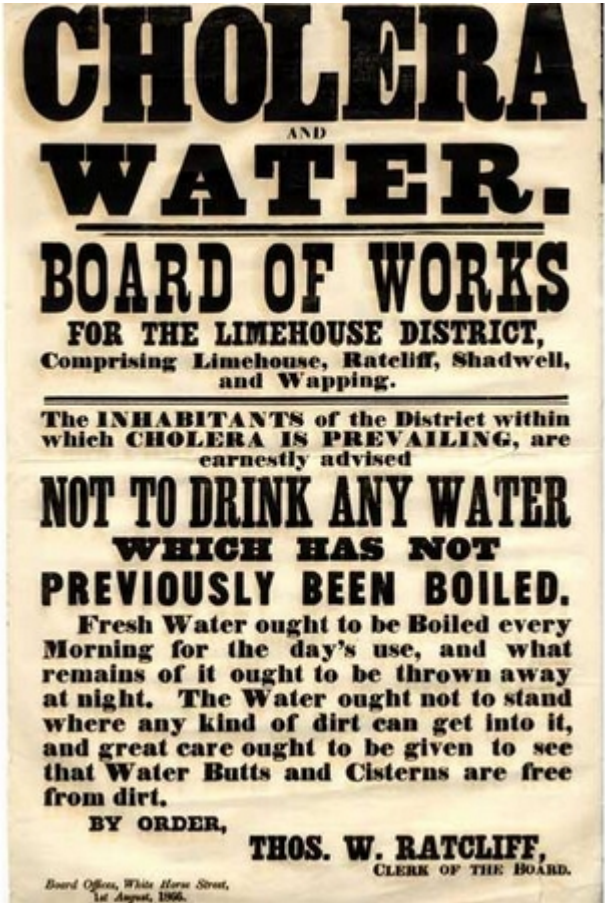
<https://bokeh.pydata.org/> successfully loaded.



```
In [0]: %%nose

def test_dates_exists():
    assert "dates" in globals(), \
        "The variable dates should be defined."
```

## 8. John Snow's myth & Did we learn something?



From the previous interactive visualization, we can clearly see that the peak of the cholera outbreak happened before removing the handle and it was already in decline (downside trajectory) before the 8th of September 1854.

This different view on the data is very important because in case that we compare just absolute numbers this could lead us to wrong conclusion that removing the handle on Broad Street pump for sure stopped the outbreak, which is simply not true (it surely did help but did not stop the outbreak) and John Snow was aware of this (he just did what needed to be done and never aspired to become a hero).

But people love stories about heroes and other myths (definitely more than science or data science). According to John Snow's myth, he was the superhero who in two days defied their equals by hypothesizing that cholera was a waterborne disease. Despite no one listening to him, he bravely continued drawing his map, convinced local authorities to remove the handle of the infected water pump with his findings, and stopped the outbreak. John Snow saved the lives of many Londoners.

If we take a better look behind this story, we can find also the true John Snow, who was fighting the disease with limited tools and wanted to get proof that he was right and "knew something" about cholera. He just did what he could with limited time and always boiled his water before drinking.

```
In [0]: # Based on John Snow's map and the data John Snow collected, what would you say?
john_snow_knows_nothing = False
```

```
In [0]: %%nose

def test_john_snow_knows_nothing():
    assert john_snow_knows_nothing == False, \
        "You know nothing, John Snow! (Would you agree with this statement?)"
```