190F Foundations of Data Science

Spring 2020

Lecture 2

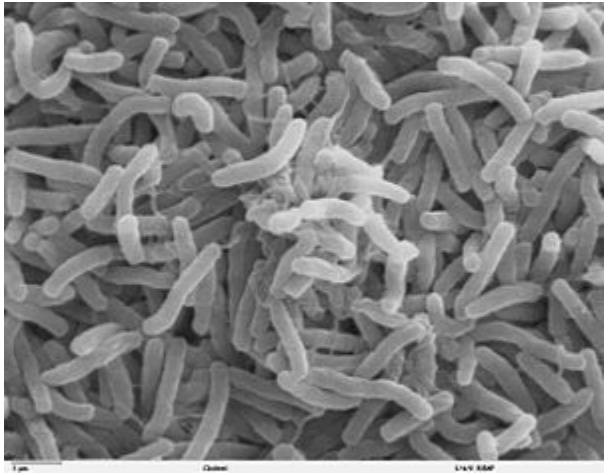
Cause and Effect
Some Python basics

Announcements

Proving Cause and Effect

Broad Street Cholera Outbreak

- The Broad Street cholera outbreak was a severe outbreak of cholera that occurred in 1854 near Broad Street in the Soho district of London, England.
- This outbreak killed 616 people.

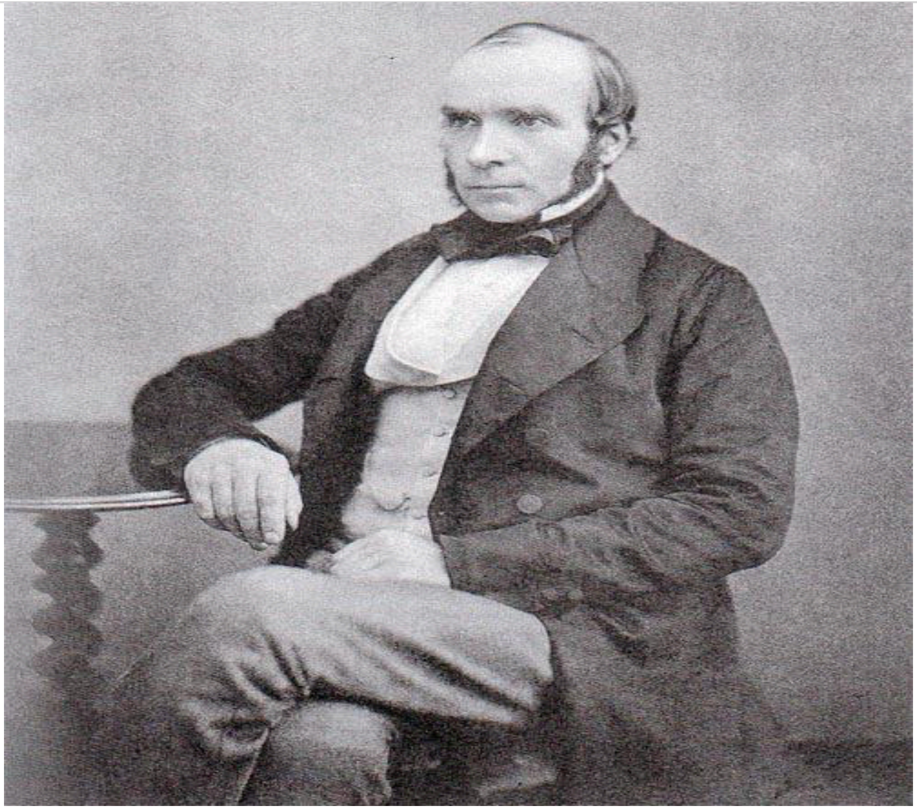


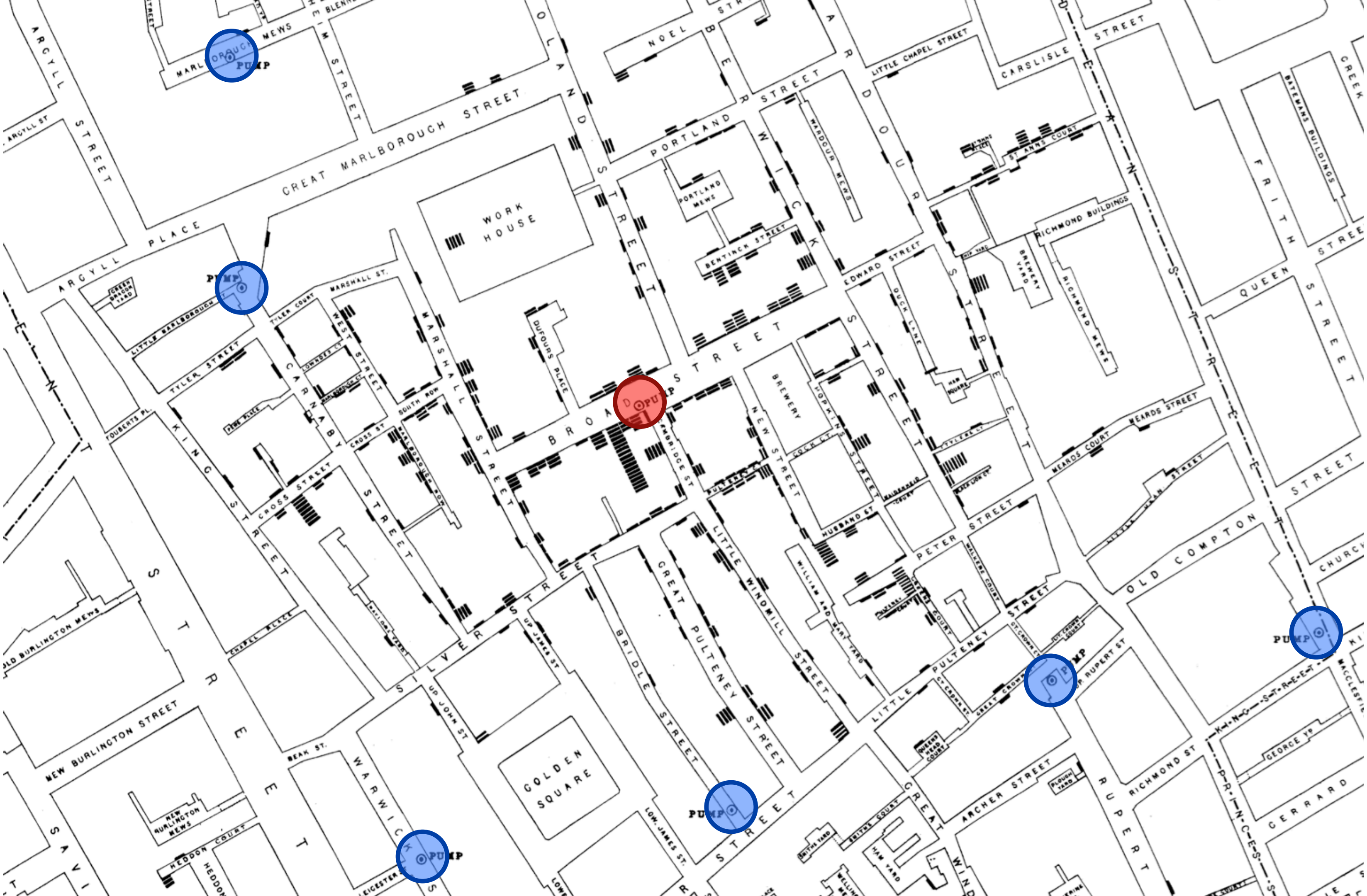
Two Theories of Cholera

- **Miasma theory:** cholera was caused by **particles in the air**, or "miasmata", which arose from decomposing matter or other dirty organic sources.

- **Germ theory:** the principal cause of cholera was a germ cell that had not yet been identified, but was **transmitted through food or drink**.

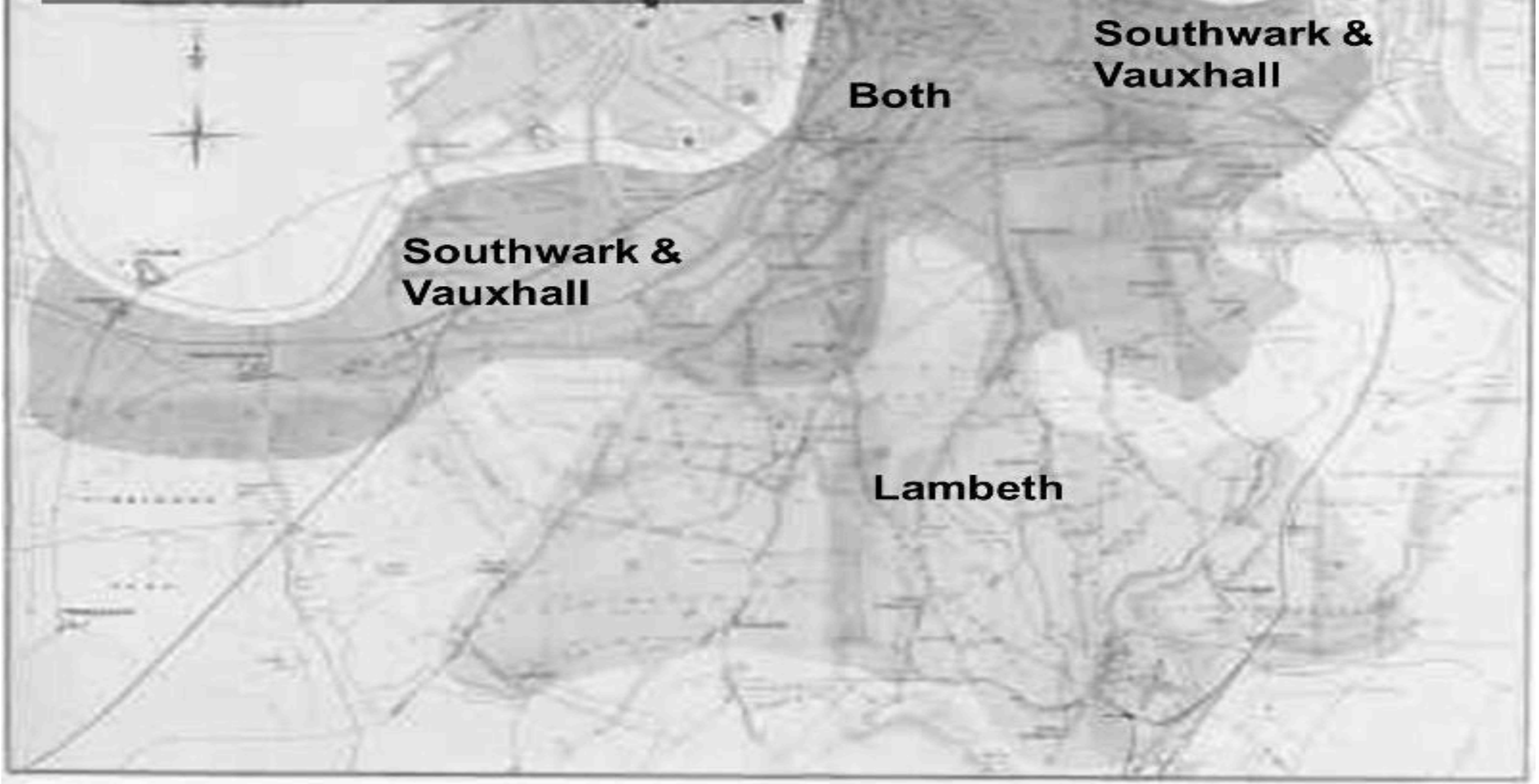
John Snow, 1813-1858




MARLBOROUGH MEWS
PUMP
BLENHEIM STREET
NOEL STREET
BERWICK STREET
ARGYLL STREET
LITTLE ARGYLL ST
GREAT MARLBOROUGH STREET
POLAND STREET
PORTLAND STREET
LITTLE CHAPEL STREET
CARLISLE STREET
DEAN STREET
FRITH STREET
BATEMANS BUILDINGS
GREEK STREET
PORTLAND MEWS
WARDOUR MEWS
WARDOUR STREET
BENTINCK STREET
ST ANNS COURT
RICHMOND BUILDINGS
RICHMOND MEWS
BREWERY YARD
QUEEN STREET
ARGYLL PLACE
WORK HOUSE
GREEN DRAGON YARD
LITTLE MARLBOROUGH ST
TYLER COURT
MARSHALL ST
MARSHALL STREET
TYLER STREET
WEST STREET
DUFOURS PLACE
BROAD STREET
EDWARD STREET
DUCK LANE
NEW SQUARE
MEARDS STREET
MEARDS COURT
MARLBOROUGH ROW
CARNABY STREET
FOUBERTS PL
KING STREET
CROSS STREET
SOUTH ROW
CAMBRIDGE ST
NEW STREET
BREWERY
HOPKINS STREET
HUSBAND ST
PETER STREET
OLD COMPTON STREET
CHURCH STREET
REGENT STREET
SILVER STREET
CHAPEL PLACE
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
WILLIAM AND MARY YARD
BRIDLE STREET
GOLDEN SQUARE
WARWICK STREET
BEAK ST
UP JAMES ST
UP JOHN ST
LITTLE PULTENEY STREET
GREAT CROWN CT
RUPERT STREET
ARCHER STREET
RICHMOND ST
KING STREET
PRINCES STREET
GERRARD
GEORGE YD
MACCLESFIELD
QUEENS HEAD COURT
PLOUGH YARD
SMITHS COURT
HAM YARD
NEW BURLINGTON STREET
OLD BURLINGTON MEWS
NEW BURLINGTON MEWS
HEDDON COURT
HEDDON
SAVILE
LEICESTER ST
LOW JAMES ST

LEXINGTON STREET
JOHN SNOW
JOHN SNOW
JOHN SNOW
SALOON
SALOON

London Water Supply



Comparison

- **Treatment group:** Do receive the treatment
- **Control group:** Do not receive the treatment

Snow’s “Grand Experiment”

“… there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded …”

- The two groups were *similar except for the treatment*.

Snow’s table

Supply Area	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37

Key to establishing causality

If the treatment and control groups are *similar apart from the treatment*, then differences between the outcomes in the two groups can be ascribed to the treatment.

Confounding

- If the treatment and control groups have systematic differences other than the treatment, then it might be difficult to identify causality.
- Such differences are often present in **observational studies**.
- When they lead researchers astray, they are called confounding factors.

Randomization and Confounding

- If you assign individuals to treatment and control **at random**, then the two groups are likely to be similar apart from the treatment.

- You can account – mathematically – for variability in the assignment.

- **Randomized Controlled Experiments are the gold standard for establishing cause and effect.**

RCEs vs Observational Studies

- **Question:** If randomized controlled experiments can establish causality while observational studies are subject to confounding, why are so many studies observational?

Python

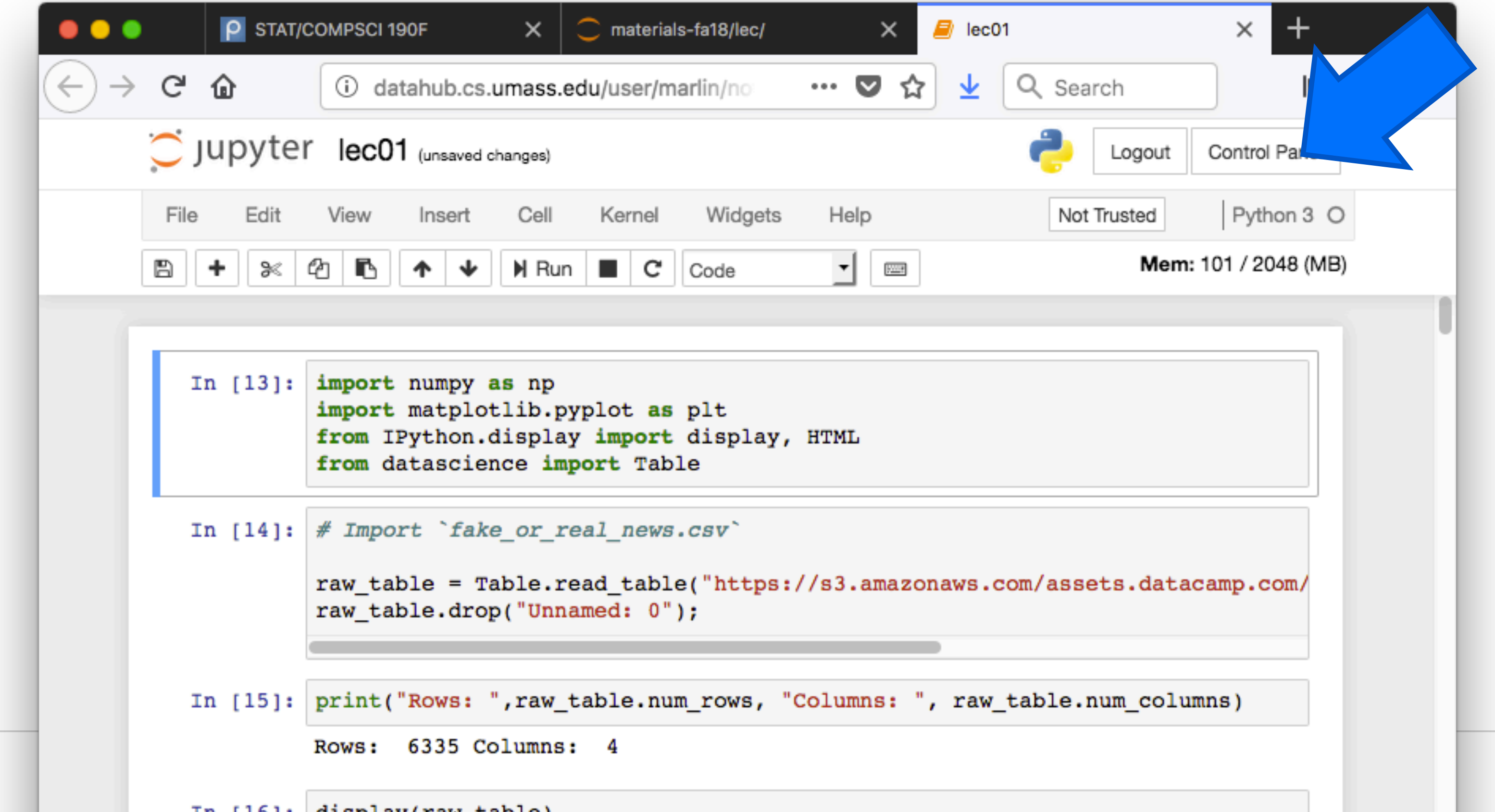
Datahub

- If you have an @umass.edu email address, you should be able to access the course's data hub. (If not, please email me at chosman@math.umass.edu)

- Datahub uses UMass Google Apps authentication. Use your @umass.edu email address and Spire password to log in. It takes a minute to start up.

- When you're done working with the Datahub, make sure to shut your datahub server down and then log out.

Stopping Your Data Hub



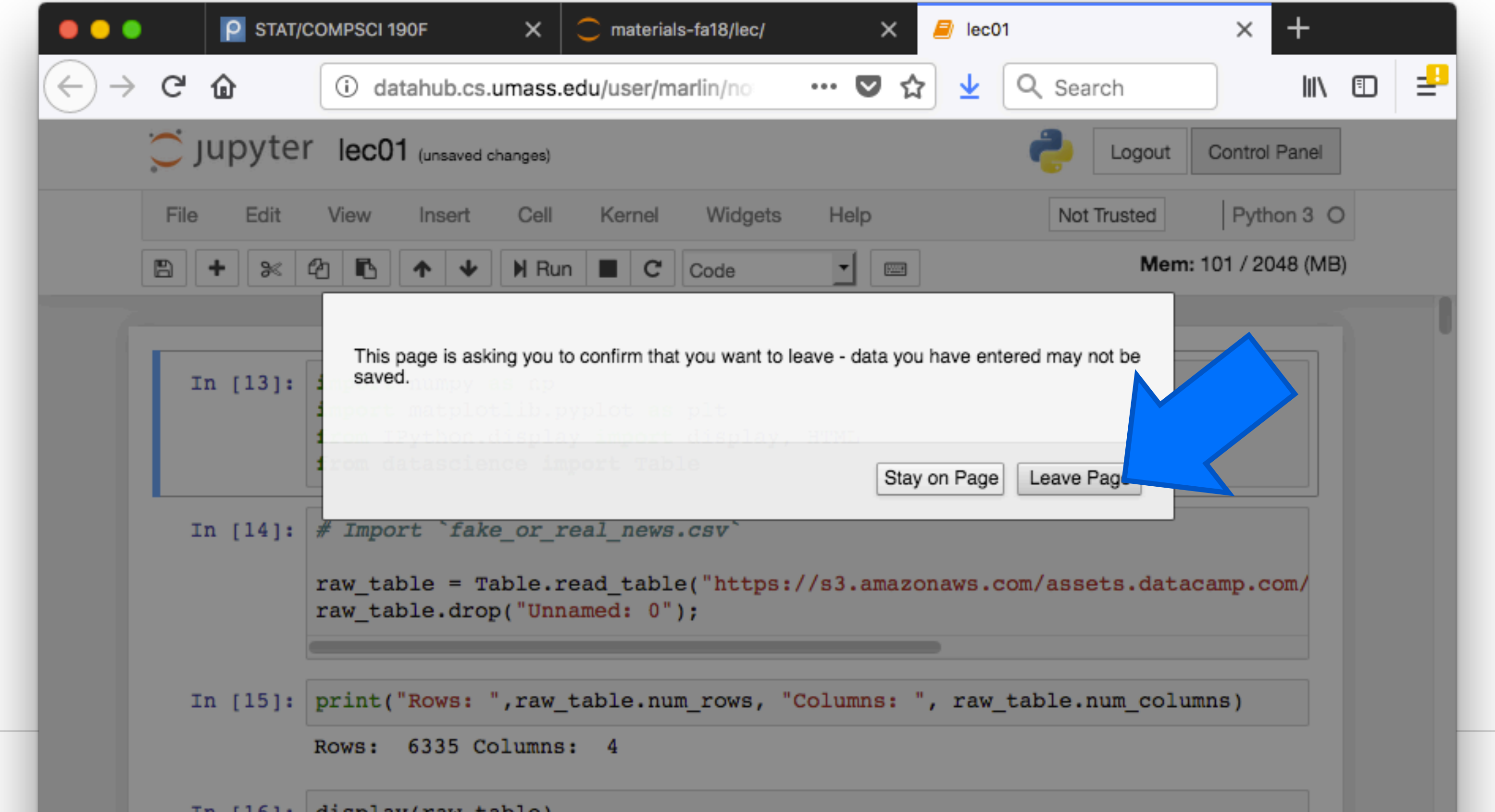
```
In [13]: import numpy as np
         import matplotlib.pyplot as plt
         from IPython.display import display, HTML
         from datascience import Table
```

```
In [14]: # Import `fake_or_real_news.csv`

         raw_table = Table.read_table("https://s3.amazonaws.com/assets.datacamp.com/
         raw_table.drop("Unnamed: 0");
```

```
In [15]: print("Rows: ",raw_table.num_rows, "Columns: ", raw_table.num_columns)

         Rows:  6335 Columns:  4
```

```
In [16]: display(raw_table)
```

Stopping Your Data Hub

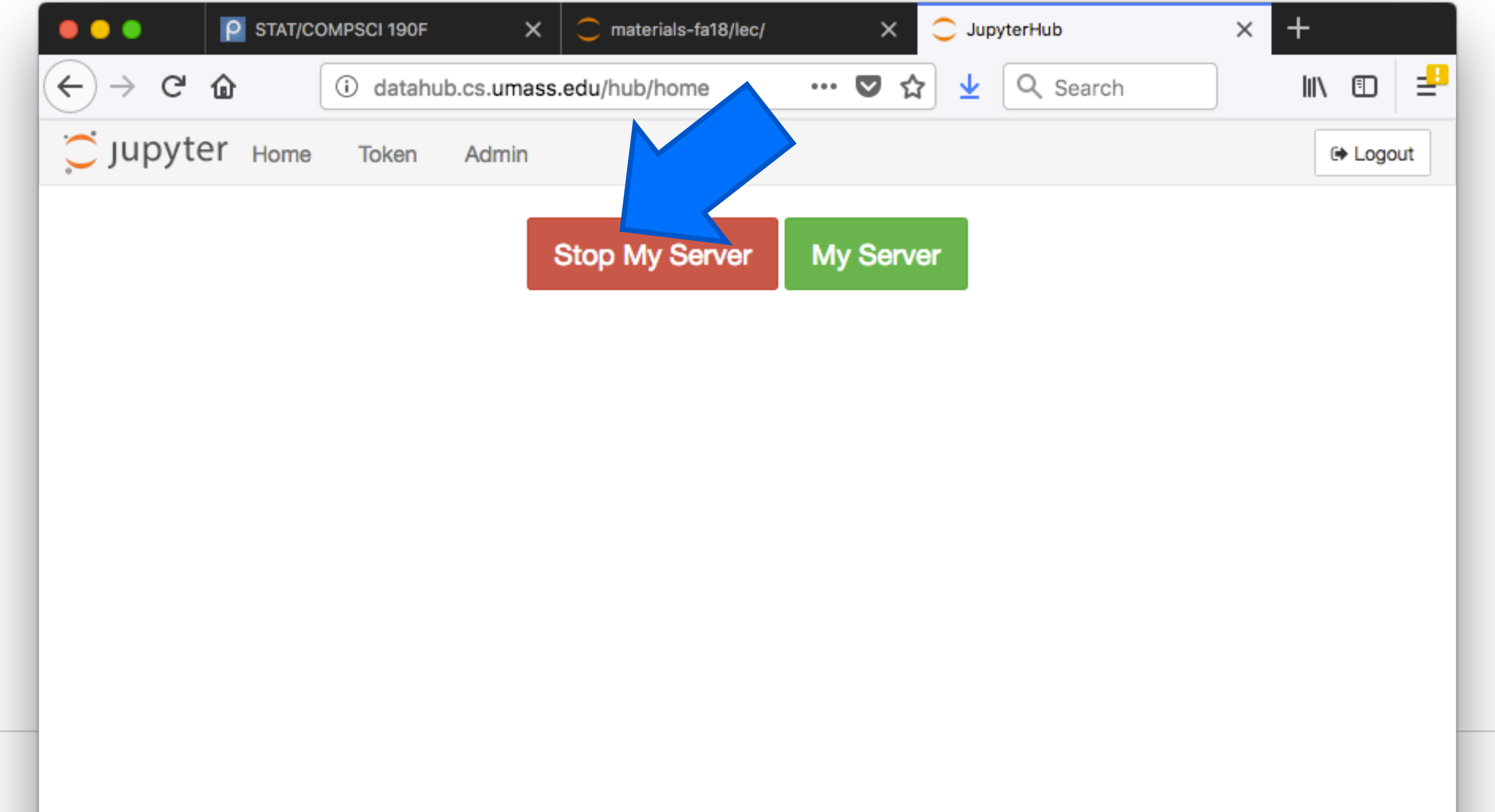


This page is asking you to confirm that you want to leave - data you have entered may not be saved.

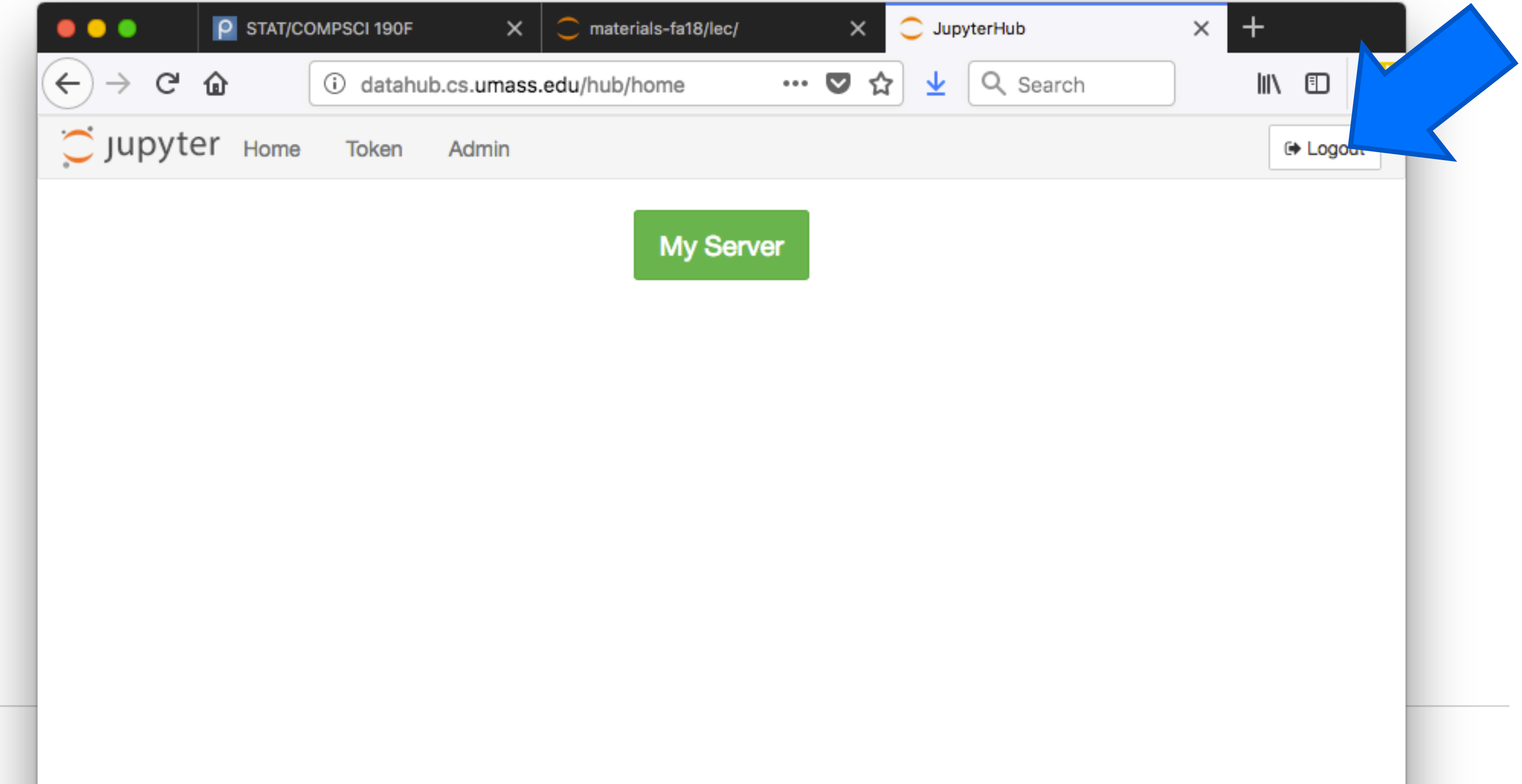
Stay on Page | Leave Page

```
In [13]: import numpy as np
         import matplotlib.pyplot as plt
         from IPython.display import display, HTML
         from datascience import Table

In [14]: # Import `fake_or_real_news.csv`

         raw_table = Table.read_table("https://s3.amazonaws.com/assets.datacamp.com/
         raw_table.drop("Unnamed: 0");

In [15]: print("Rows: ",raw_table.num_rows, "Columns: ", raw_table.num_columns)

         Rows:  6335 Columns:  4

In [16]: display(raw_table)
```

Stopping Your Data Hub



Stopping Your Data Hub

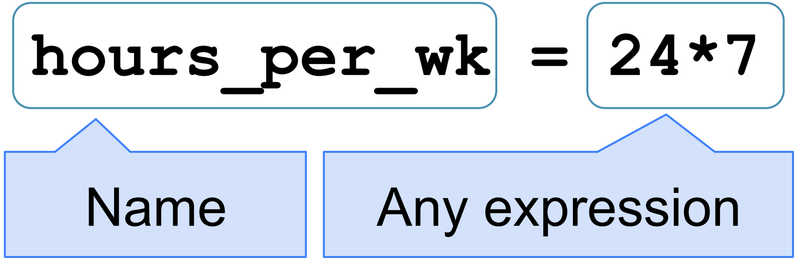


Programming Languages

- Python is popular both for data science & general software development
- Mastering the language fundamentals is critical
- Learn through practice, not by reading or listening
- Follow along: Click on today’s “Notebook” on the course website

Names

Assignment Statements

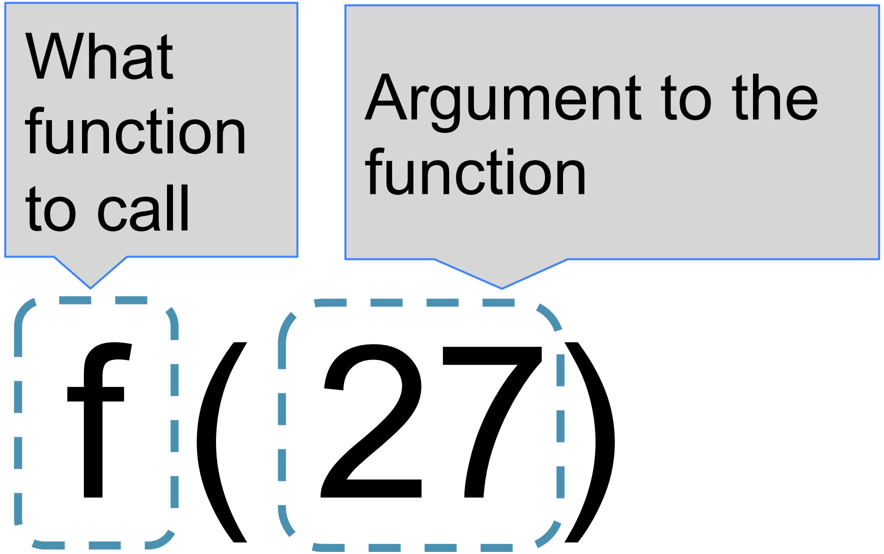


```
hours_per_wk = 24*7
```

- Statements don't have a value; they perform an action
- An assignment statement changes the meaning of the name to the left of the = symbol
- The name is bound to a value (not an equation)

Call Expressions

Anatomy of a Call Expression



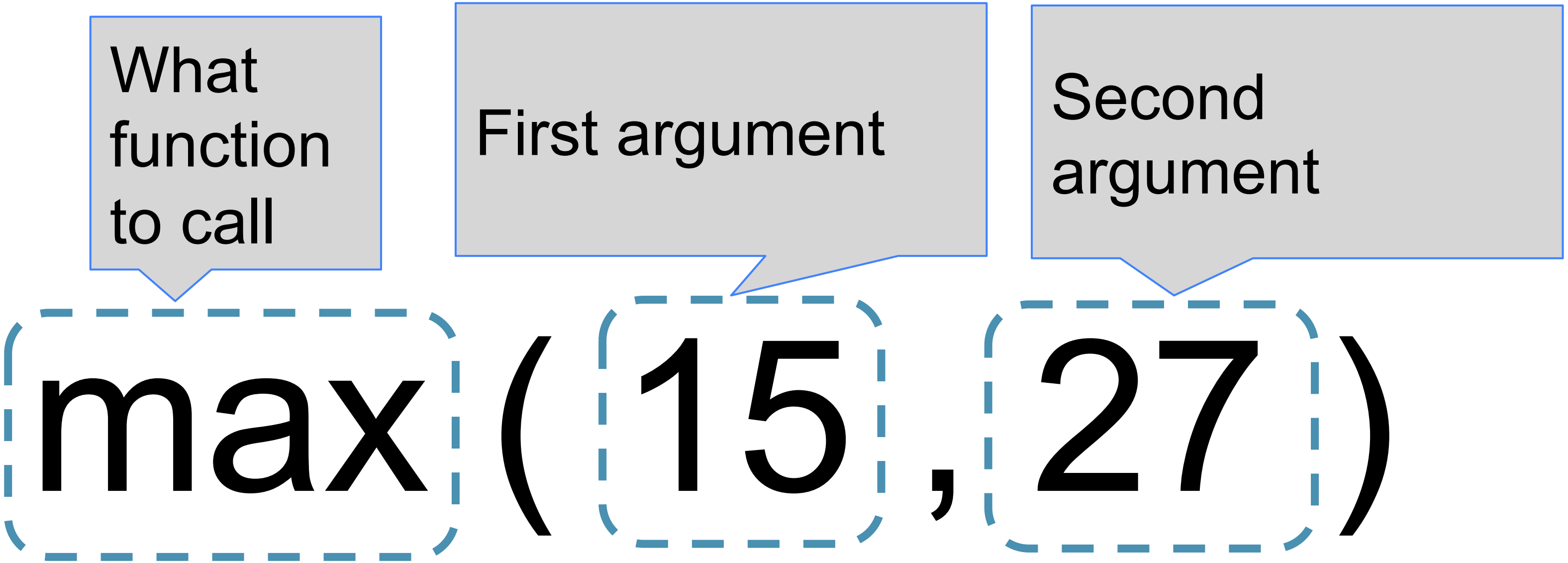
What function to call

Argument to the function

f(27)

"Call f on 27."

Anatomy of a Call Expression



What function to call

First argument

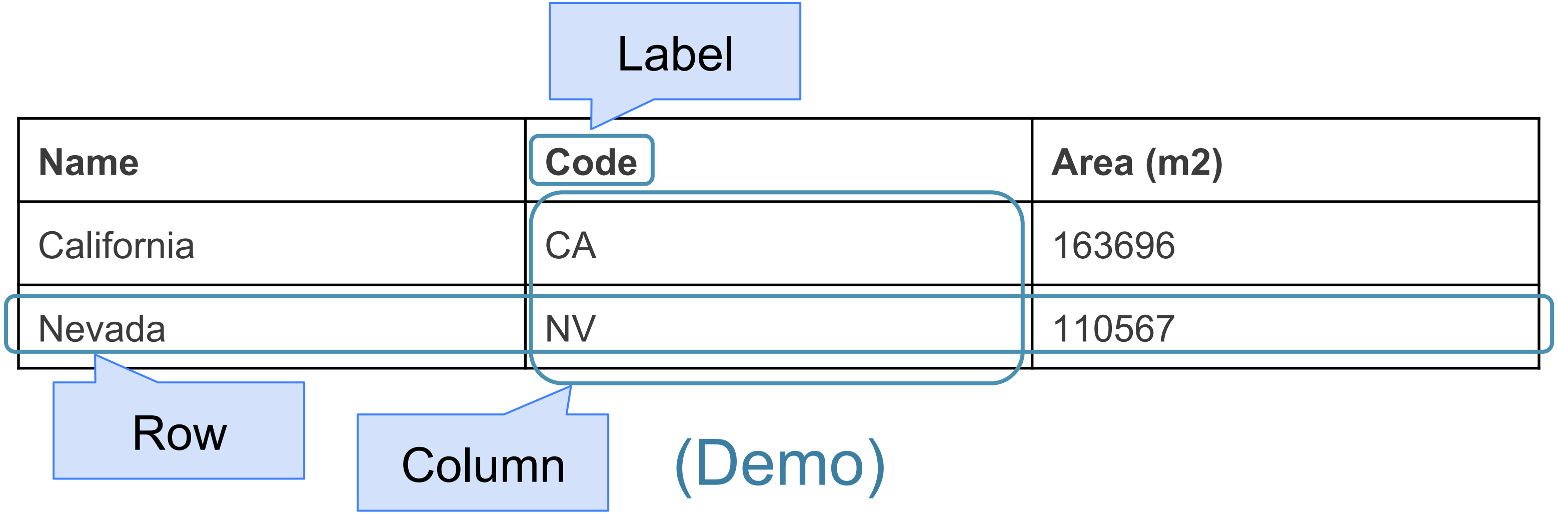
Second argument

```
max(15,27)
```

Tables

Table Structure

- We organize our data in tables
- A Table is a sequence of labeled columns
- Data within a column should be of the same "type"



Name	Code	Area (m2)
California	CA	163696
Nevada	NV	110567

(Demo)

Table Operations

- `t.select(label)` - constructs a new table with just the specified columns
- `t.sort(label)` - constructs a new table, with rows sorted by the specified column

(Demo)

The where method

- `t.where(label, condition)` - constructs a new table with just the rows that match the condition

(Demo)