MODULE 1

Introduction

CS108 - FOUNDATIONS OF DATA SCIENCE – F22

Venue: CS 140

Lectures: TuTh 8:30am – 9:45am

Labs: 9:05am -9:55am Fri

Instructor: Purity Mugambi

TA: To be Updated

ABOUT YOUR INSTRUCTORS

- Purity Mugambi
  - 5th year PhD candidate
  - Research in development of ML tools to detect (and improve) health equity in patient treatments and outcomes
  - **OH: Th, 10am-11am, LGRC Rm. A205**
- TA – To be updated

DATA SCIENCE

WHAT IS DATA SCIENCE?

DS is the discipline of **drawing useful conclusions from data using computation**

- **Exploration**
  - Identifying patterns in information
  - Uses visualizations
- **Inference**
  - Quantifying whether those patterns are reliable
  - Uses randomization
- **Prediction**
  - Making informed guesses
  - Uses machine learning

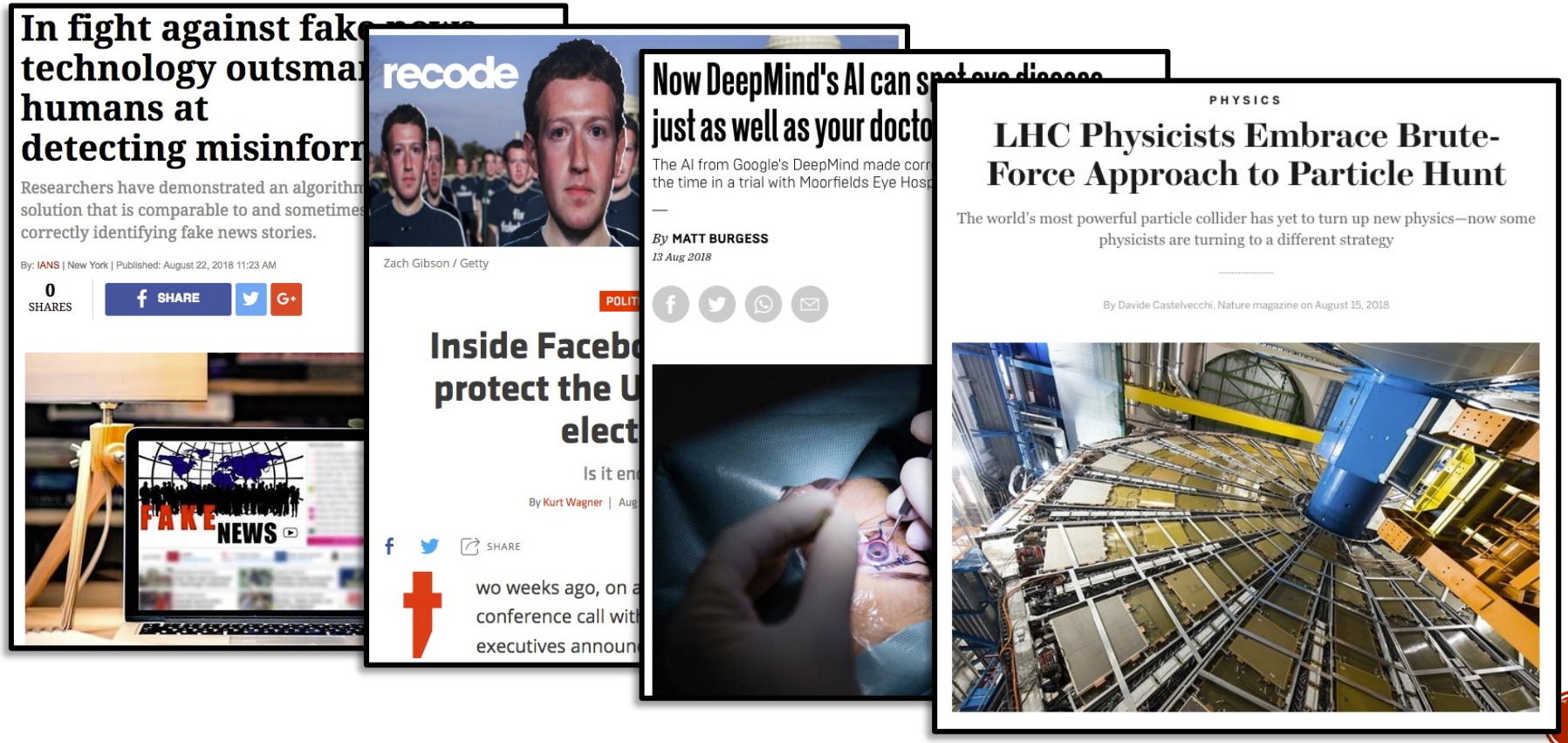
APPLICATIONS

- Data science is driven by applications

- Data analysis is playing an increasingly important role in many fields including:
  - Biology, Chemistry, Economics, Earth Systems, Education, Environmental Science, Finance, Geography, Geology, Kinesiology, Linguistics, Management, Political Science, Public Health, Psychology, Sociology, …

- Every data-driven subject brings new challenges

EXAMPLES


In fight against fake
technology outsma
humans at
detecting misinfor
Researchers have demonstrated an algorithm
solution that is comparable to and sometimes
correctly identifying fake news stories.
By: IANS | New York | Published: August 22, 2018 11:23 AM
0
SHARES
SHARE
FAKE NEWS
recode
Zach Gibson / Getty
Inside Faceb
protect the U
elect
Is it en
By Kurt Wagner | Aug
SHARE
wo weeks ago, on a
conference call wit
executives announ
Now DeepMind's AI can s
just as well as your docto
The AI from Google's DeepMind made corr
the time in a trial with Moorfields Eye Hosp
By MATT BURGESS
13 Aug 2018
PHYSICS
LHC Physicists Embrace Brute-Force Approach to Particle Hunt
The world's most powerful particle collider has yet to turn up new physics—now some physicists are turning to a different strategy
By Davide Castelvecchi, Nature magazine on August 15, 2018

COURSE STRUCTURE

WHAT DOES THE COURSE COVER?

- An introduction to programming in Python with a focus on manipulating, visualizing, and analyzing data.
- An introduction to statistics that is grounded in computer simulations.
- An introduction to predictive modeling and machine learning.

COURSE COMPONENTS AND GRADING

- Lectures (often interactive)
- Weekly labs including attendance (pass/fail)
- Weekly graded homework assignments
- Midterm exam & final exam

Homework	35%
Labs	20%
Midterm Exam	20%
Final Exam	25%

COURSE TECHNOLOGY

- **Moodle:** Online gradebook
- **Github.io:** Course website (lecture slides, demos, assignments, labs, syllabus, etc.)
- **DataHub:** Web-based Python compute environment for completing labs and homework assignments.
- **Gradescope**: Assignments, labs, and exams submission system
- **Links to all resource can be found on Moodle.**

COURSE POLICIES

- Late Homework
- Re-grades
- Academic Honesty
- https://umass-data-science.github.io/190fwebsite/policies/

COLLABORATION POLICY

Asking questions is highly encouraged

- You can discuss homework and lab questions with each other
- Do not take notes or pictures out of discussions
- The work you turn in must be your own

The Limits of collaboration

- Don't share solution material of any type with each other
- Copying solutions from any source will be dealt with under UMass' Academic Honesty procedures: https://www.umass.edu/honesty/

GETTING HELP

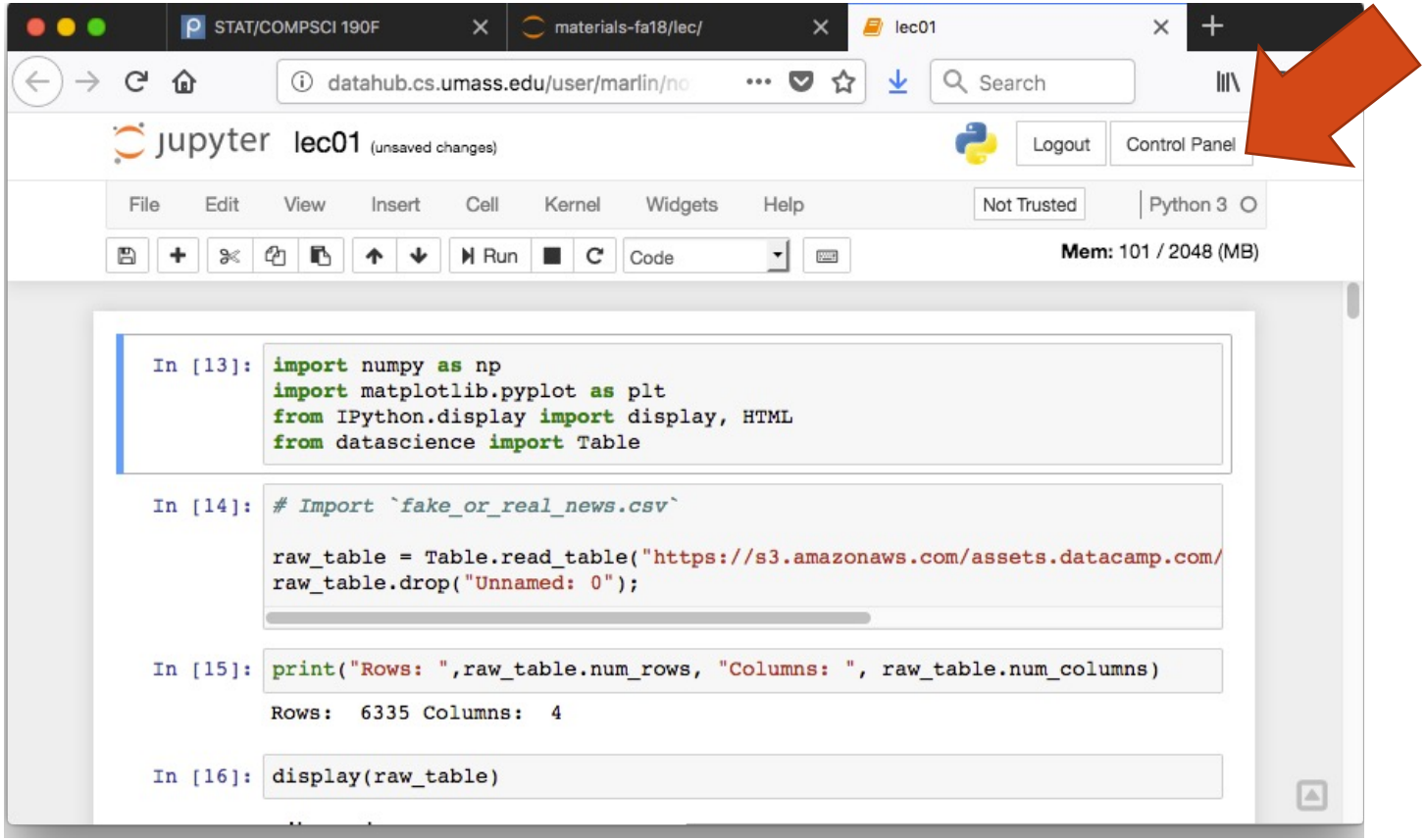
- The course staff are here to help you be successful in the course!
- When you need help come to office hours.
- The lab sessions are also a good time to ask questions and get help.

ON USING DATAHUB

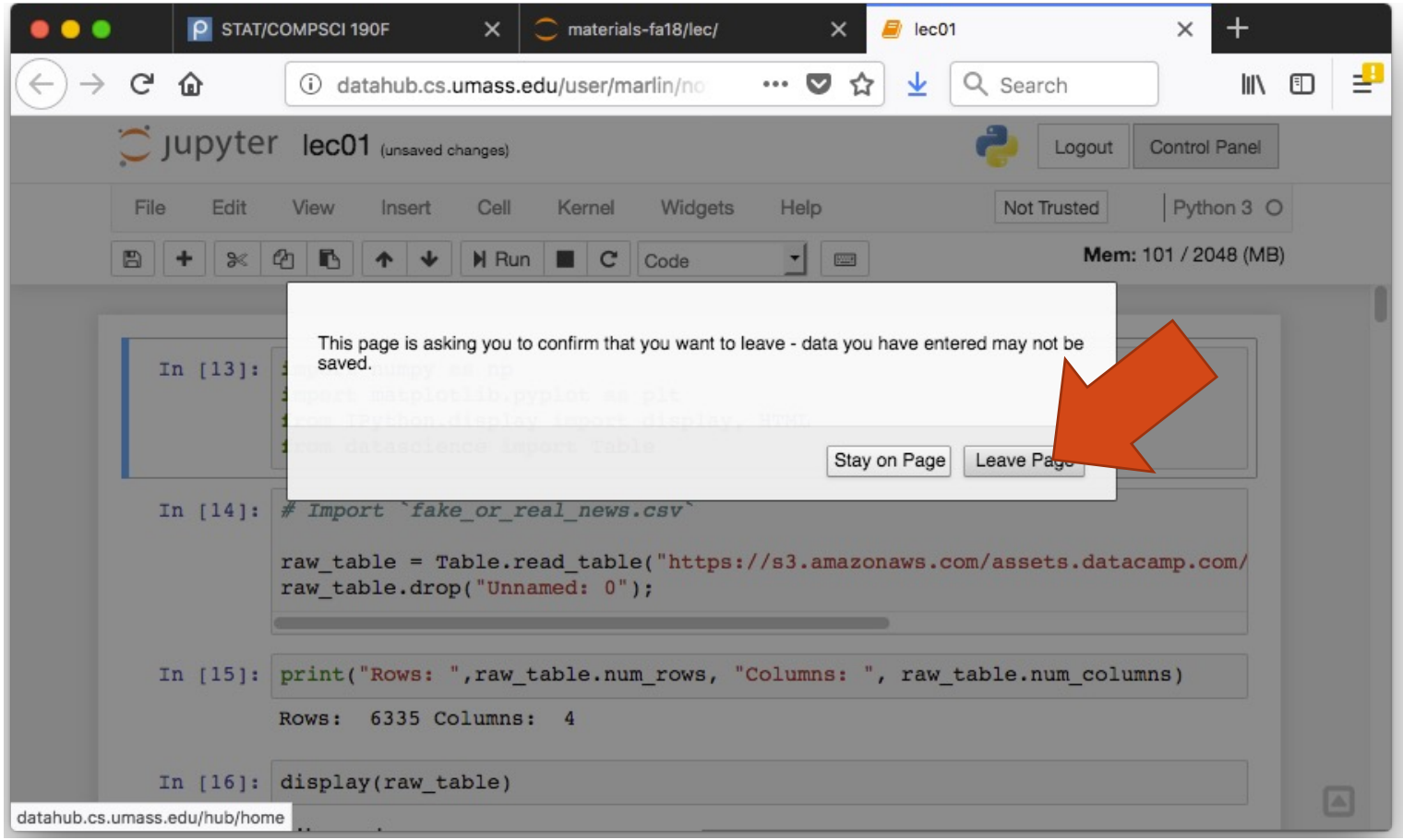
DATAHUB

- If you have an @umass.edu email address, you should now be able to access the course's data hub.
- 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.

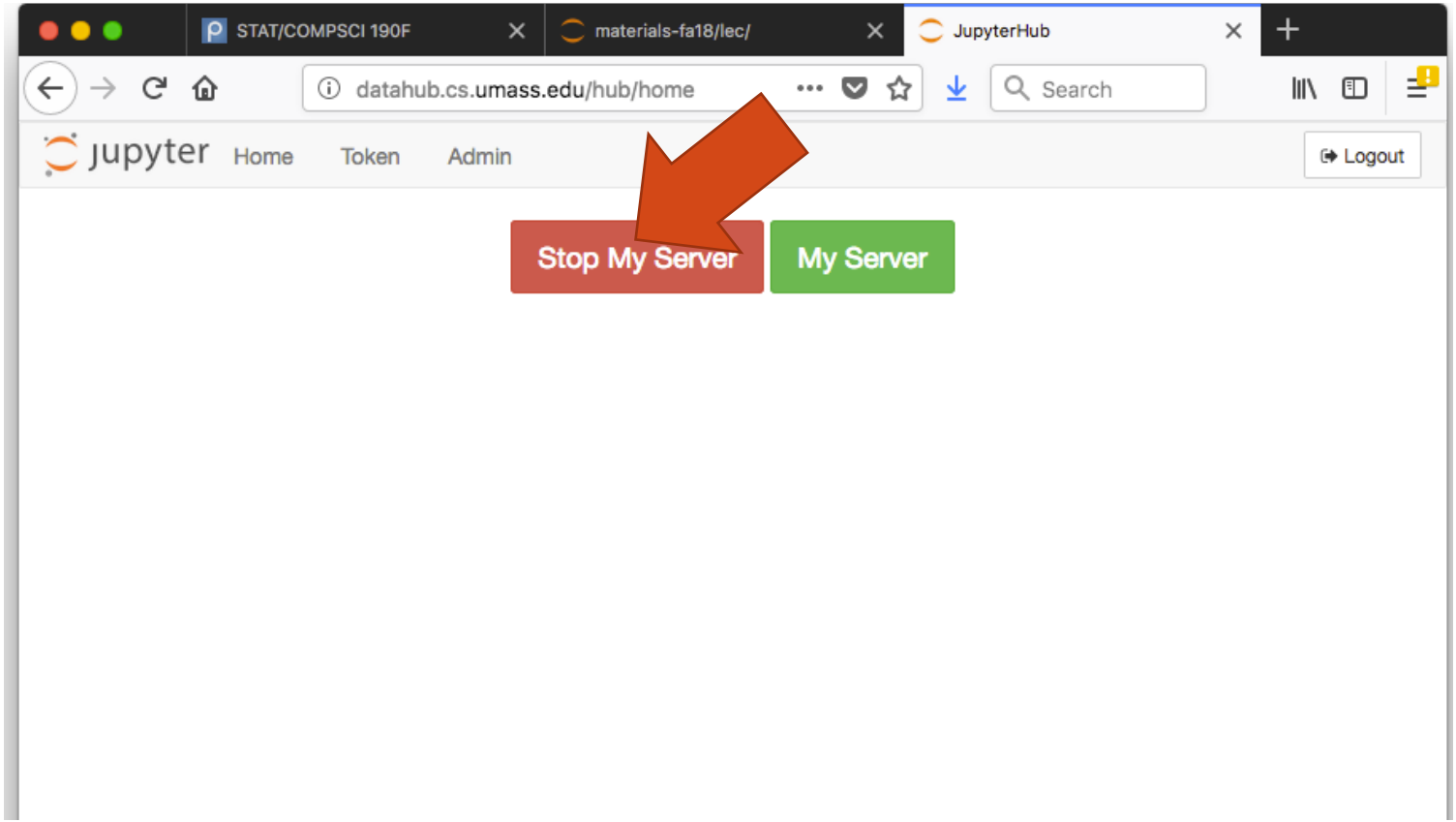
STOPPING YOUR DATA HUB



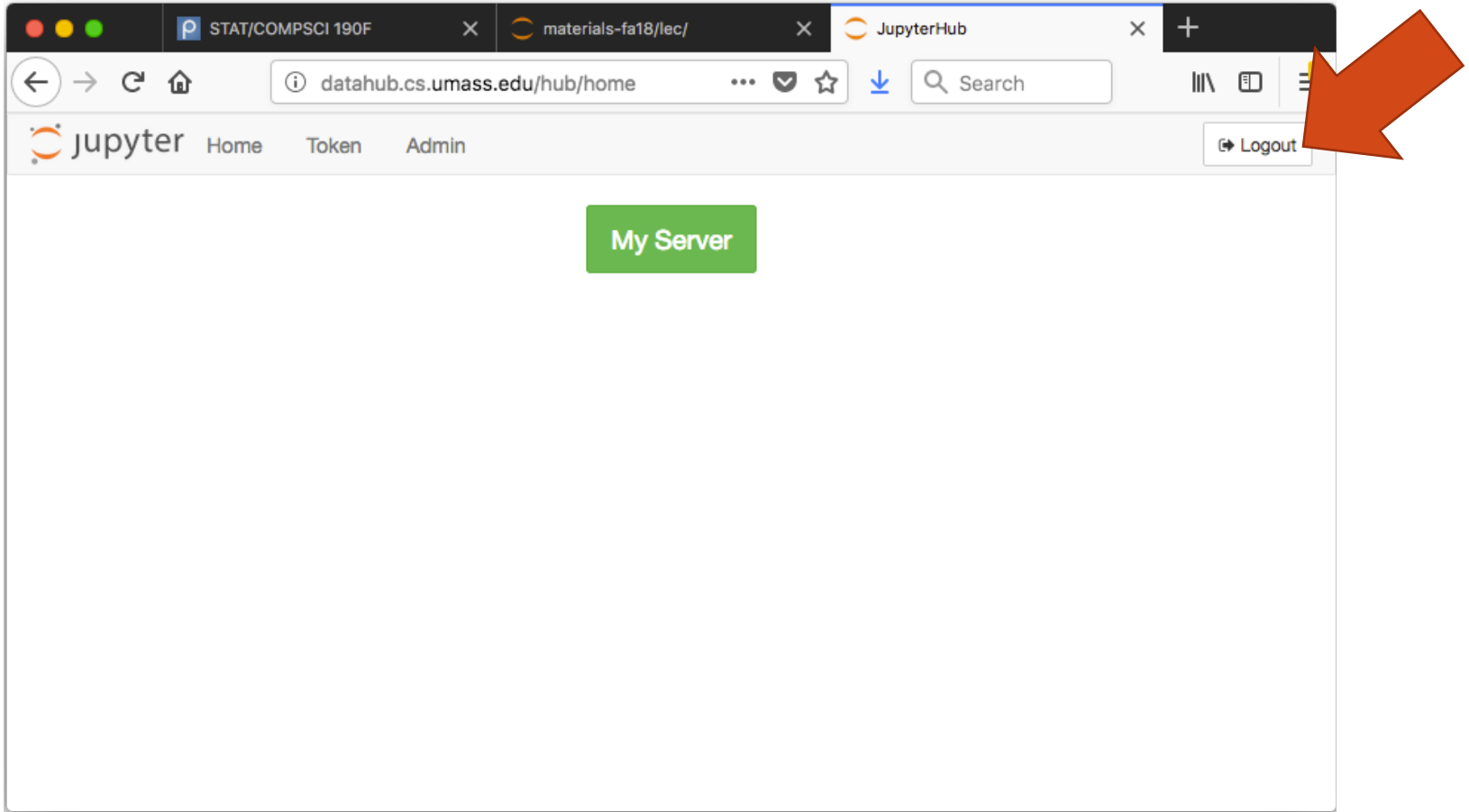
STOPPING YOUR DATA HUB



STOPPING YOUR DATA HUB



STOPPING YOUR DATA HUB

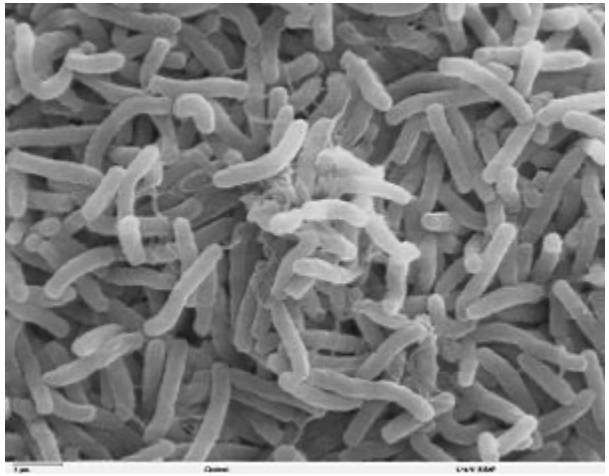


CAUSE AND EFFECT

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.

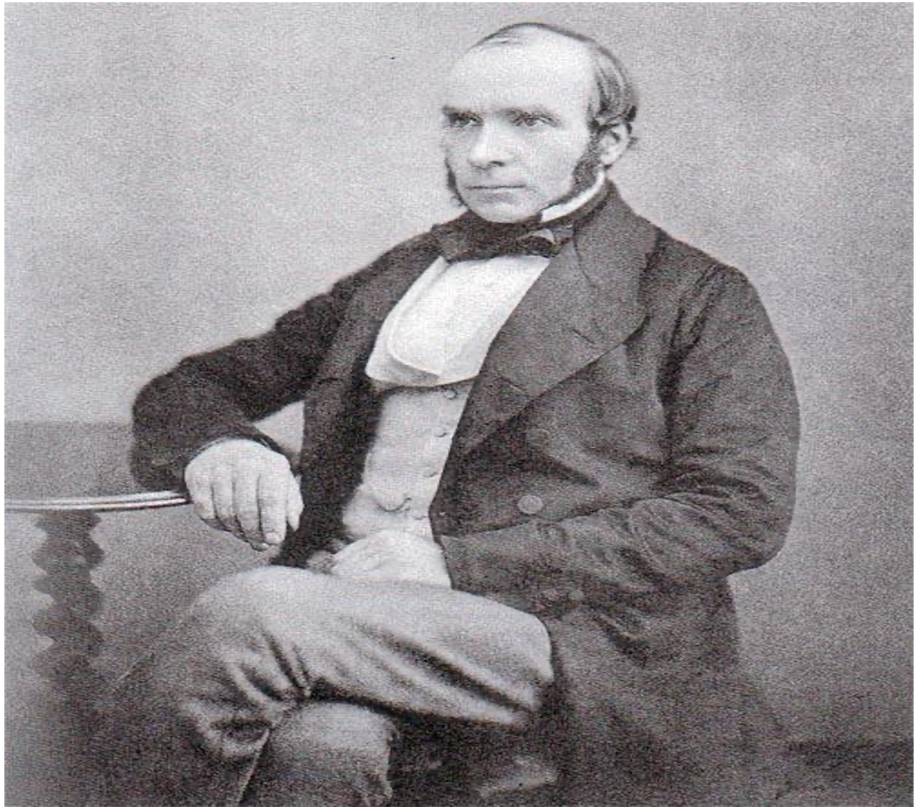


TWO THEORIES OF CHOLERA

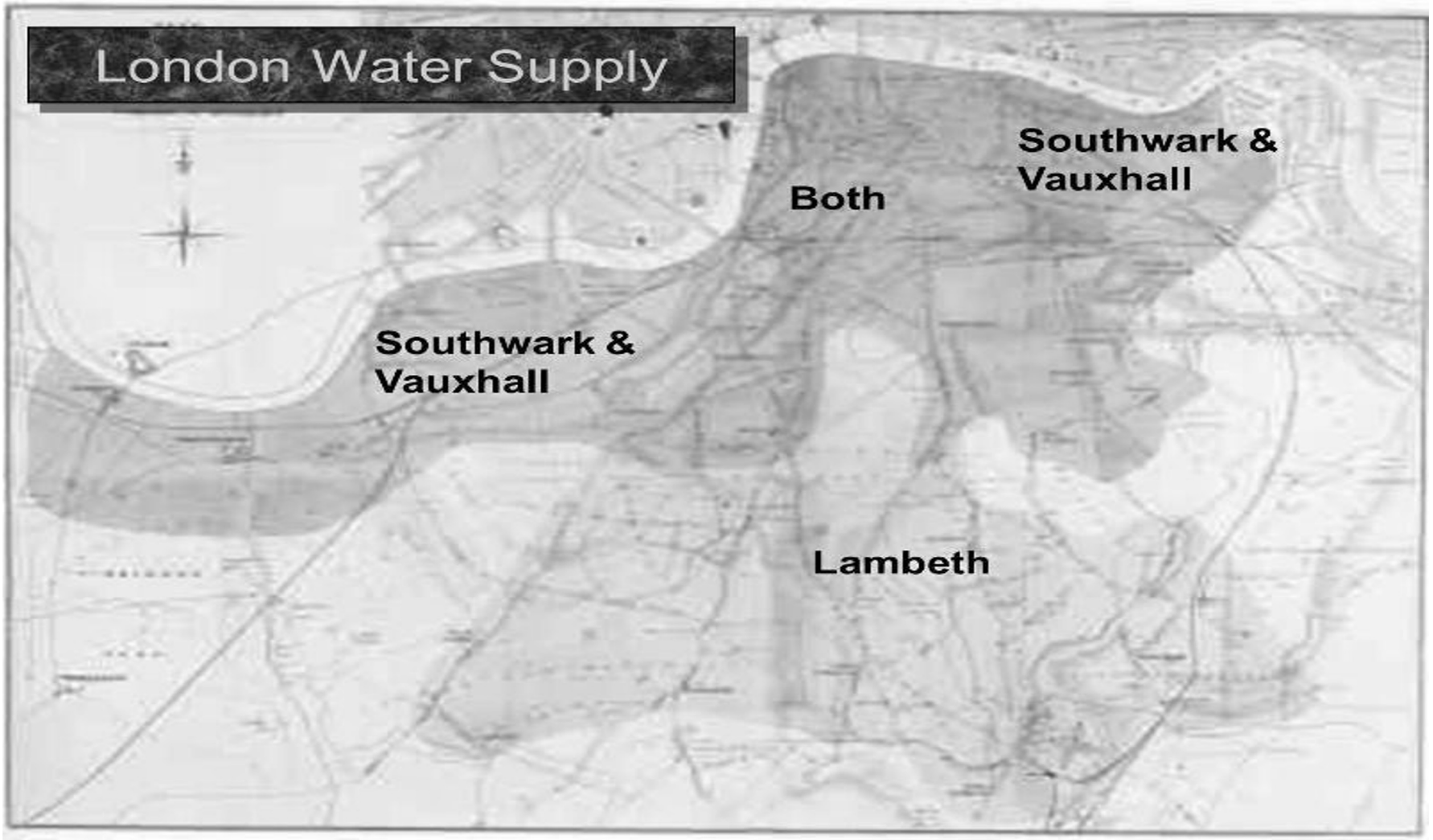
- **Miasma theory:** cholera was caused by **particles in the air**, or "miasmata", which arose from decomposing matter or other dirty organic sources.

- **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



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


London Water Supply
Southwark &
Vauxhall
Both
Southwark &
Vauxhall
Lambeth

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
Rest of London	256,423	1,422	59

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*.

COMPARISON

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

ASSOCIATION VS CAUSATION

CHOCOLATE AND HEART DISEASE: STUDY

Chocolate, Chocolate, It's Good For Your Heart, Study Finds

June 19, 2015 · 5:03 AM ET
Heard on Morning Edition

ALLISON AUBREY

- **Population** (individuals, study subjects, participants, units): 20K *European adults followed for 12 years.*
- **Treatment:** *chocolate consumption*
- **Outcome:** *heart disease*

https://www.npr.org/sections/thesalt/2015/06/19/415527652/chocolate-chocolate-its-good-for-your-heart-study-finds

CHOCOLATE AND HEART DISEASE: ASSOCIATION

Question 1: Is there any association (any relationship) between chocolate consumption and heart disease?

- **Data:** "Among those in the top tier of chocolate consumption, 12 percent developed or died of cardiovascular disease during the study, compared to 17.4 percent of those who didn't eat chocolate."
- **Answer:** Yes, this points to an **association**

CHOCOLATE AND HEART DISEASE: CAUSATION

Question 2: Does chocolate consumption lead to a reduction in heart disease?

- **This questions asks about causality**
- This question is often harder to answer.
- "[The study] doesn't prove a cause-and-effect relationship between chocolate and reduced risk of heart disease and stroke." - *JoAnn Manson, chief of Preventive Medicine at Brigham and Women's Hospital, Boston*

CHOCOLATE AND HEART DISEASE: ALTERNATIVES

Question 3: Is the fact that people ate more chocolate the only possible cause for the observed effect of decreased heart disease risk?

- For example, suppose the people who ate more chocolate tended to live in European countries with better health care?

- What if wealthier people eat more chocolate and can also afford better health care?

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.

EXAMPLES

No level of alcohol consumption is healthy, scientists say

Fo

Dairy and meat 'beneficial for heart health and longevity'

Medic

Eating cheese may be associated with a lower risk of death — and scientists are zeroing in on why it might be a better choice than milk

Business Insider • today

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?