Stat 190F
Spring 2020

Foundations of Data Science

Lecture 1

Introduction
Cause and effect

Welcome to 190F: Foundations of Data Science

Carrie Hosman

chosman@math.umass.edu

Instructor

LGRT 1614

Tom Cook

tjcook@math.umass.edu

TA

Data Science

What is Data Science?

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 may 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 news, technology outsmarts humans at detecting misinformation

Researchers have demonstrated an algorithm-based automated solution that is comparable to and sometimes better than huma correctly identifying fake news stories.

By: IANS | New York | Published: August 22, 2018 11:23 AM

0 SHARES

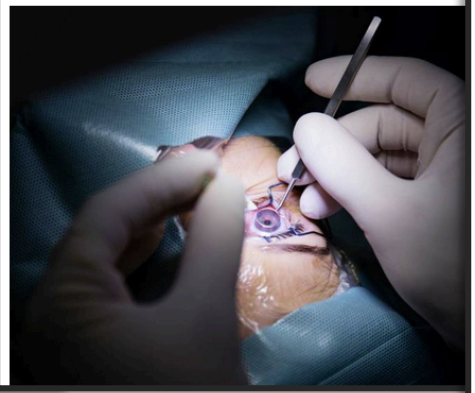
SHARE

Now DeepMind's AI can spot eye disease just as well as your doctor

The AI from Google's DeepMind made correct diagnoses 94.5 the time in a trial with Moorfields Eye Hospital

—

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 Tue/Thu
- Weekly labs on Mondays (1:25pm in LGRC A210, 9:05am in Morrill 3 Room 212)
- Weekly homework assignments
- Evening Midterm exam & final exam

Homework	35%
Labs	15%
Midterm Exam	25%
Final Exam	25%

Course Technology

- **Moodle:** Online gradebook
- **Piazza:** Online discussion forums, course Q&A, announcements, and instructor DM.
- **Github.io:** Course website (lecture slides, demos, assignments, labs, etc.)
- **DataHub:** Web-based Python compute environment for completing labs and homework assignments.

- **Links to all resource can be found on moodle and the course website.**

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 or post on the Piazza discussion forums.
- The lab sessions are also a good time to ask questions and get help.

Let’s Dive In!

Association vs Causation

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

tha

Eating chocolate regularly reduces your chance of heart failure according to new study

Busi

Huddersfield Examiner • yesterday

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?