Khoury College

FOUNDATIONS OF DATA SCIENCE

DS 3000

FALL **2019**

CAGLAR **YILDIRIM**

A little about me

Assistant Teaching Professor
Khoury College of Computer Sciences
Northeastern University

PhD in Human Computer Interaction
Virtual Reality Applications Center
Iowa State University

Khoury College

A little about me

RESEARCH

Human Computer Interaction

Mixed Reality

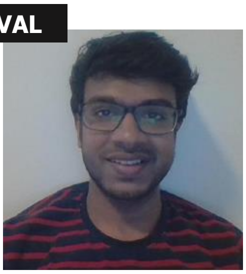
Brain-Computer Interfaces

Video Games

Applied Machine Learning

And Your TAs…

DHAVAL

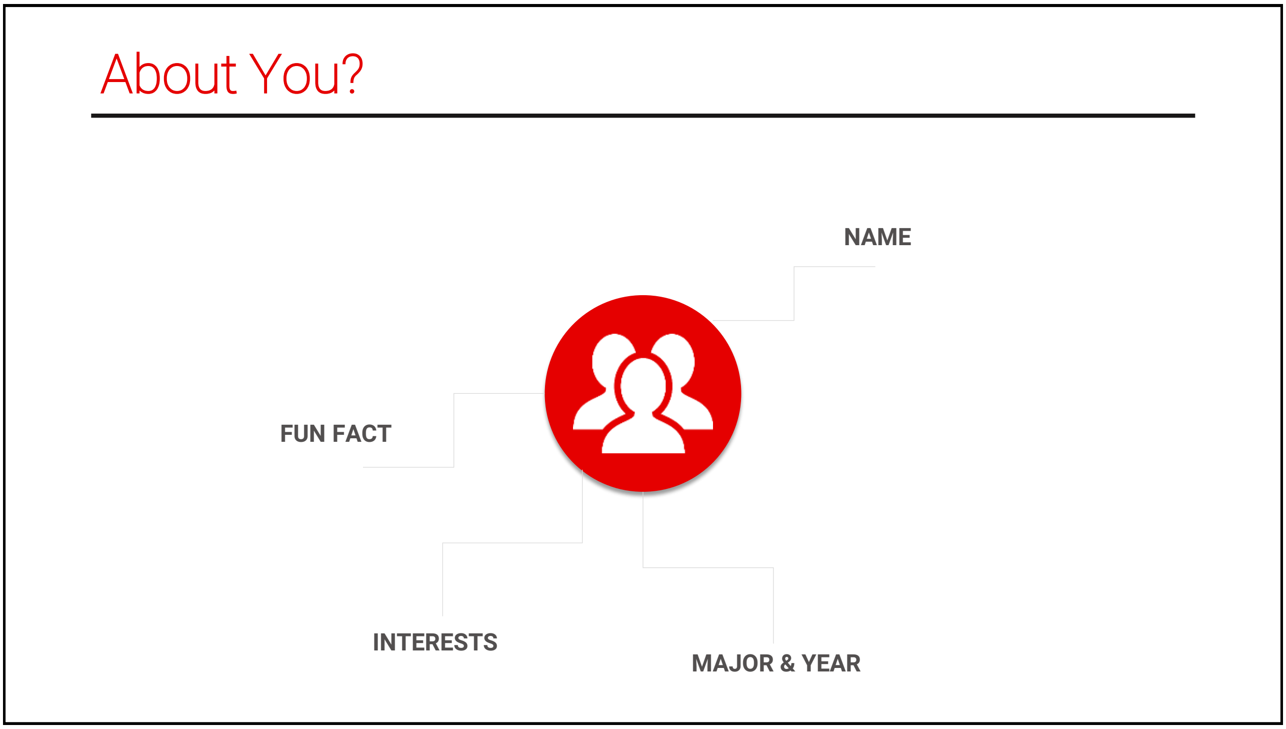


MAHITHA

PRABAL

PRASANNA

XI



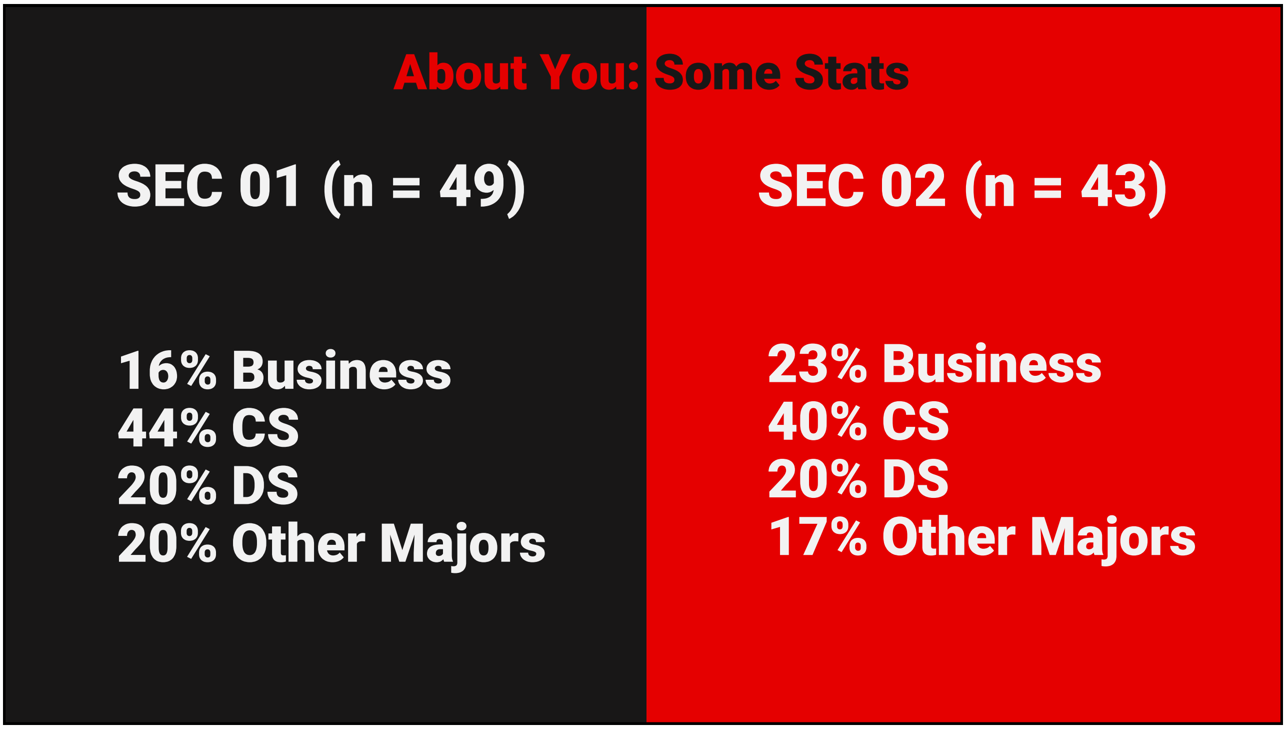
About You?

NAME

FUN FACT

INTERESTS

MAJOR & YEAR



About You: Some Stats

SEC 01 (n = 49)

16% Business
44% CS
20% DS
20% Other Majors

SEC 02 (n = 43)

23% Business
40% CS
20% DS
17% Other Majors

About You: Majors

Physics

Cognitive Psychology

Health Science

Political Science

Mechanical Engineering

Economics

Mathematics

Pharmacy

Information Science

International Affairs

About You: Some Stats

SEC 01 (n = 49)

Freshmen: 1
Sophomore: 14
Junior: 17
Senior: 17

SEC 02 (n = 43)

Freshmen: 1
Sophomore: 7
Junior: 16
Senior: 18

About the Course

DS SKILLS

DS TOOLS

Machine Learning

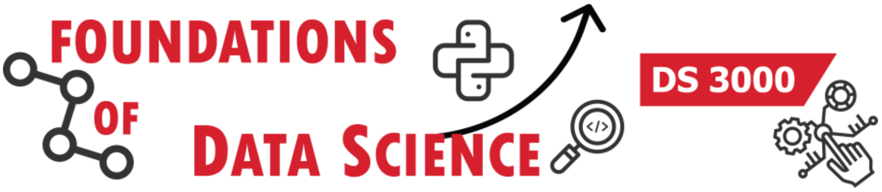
Data Processing

Data Scraping

Data Loading

Statistical Inference

Predictive Analytics



Hypothesis Testing

Data Visualization

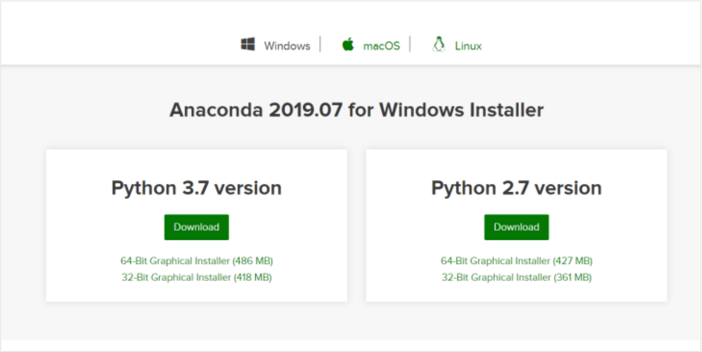
Data Modeling

Data Wrangling

Experiments

Descriptive Analytics

Jupyter Notebook & Anaconda



Anaconda 2019.07 for Windows Installer

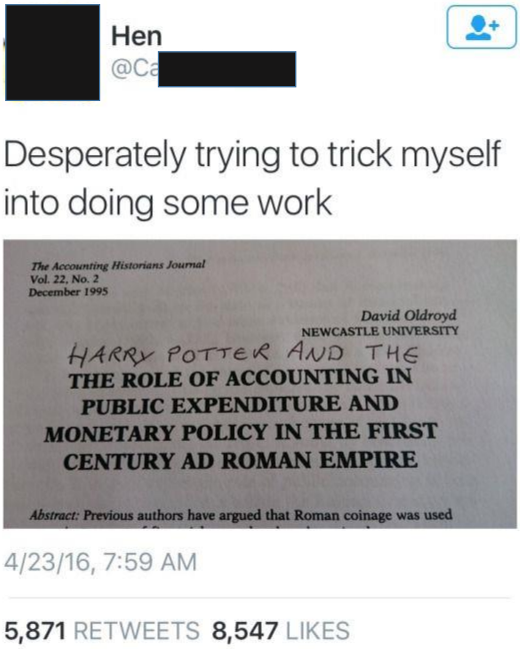
Python 3.7 version

64-Bit Graphical Installer (486 MB)
32-Bit Graphical Installer (418 MB)

Python 2.7 version

64-Bit Graphical Installer (427 MB)
32-Bit Graphical Installer (361 MB)

Download:
https://www.anaconda.com/distribution/



[redacted] Hen
@Ca[redacted]

Desperately trying to trick myself into doing some work

The Accounting Historians Journal
Vol. 22, No. 2
December 1995

David Oldroyd
NEWCASTLE UNIVERSITY

HARRY POTTER AND THE
THE ROLE OF ACCOUNTING IN PUBLIC EXPENDITURE AND MONETARY POLICY IN THE FIRST CENTURY AD ROMAN EMPIRE

Abstract: Previous authors have argued that Roman coinage was used

4/23/16, 7:59 AM

5,871 RETWEETS 8,547 LIKES

Received: 10 August 2018 | Revised: 16 October 2018 | Accepted: 23 October 2018
DOI: 10.1111/gcb.14506

WILEY Global Change

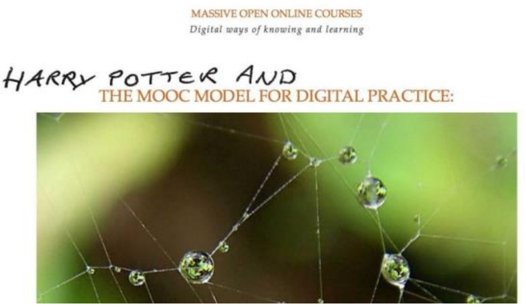
PRIMARY RESEARCH ARTICLE

HARRY POTTER AND
The influence of climatic legacies on the distribution of dryland biocrust communities

David J. Eldridge[1] | Manuel Delgado-Baquerizo[2,3]

[1]Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences,University of New South Wales, Sydney,New South Wales,Australia
[2]Departamento de Biología y Geología, Física y Química Inorgánica, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, Móstoles, Spain

Abstract
Predicting the distribution of biocrust species, mosses, lichens and liverwor
ated with surface soils is difficult, but climatic legacies (changes in climate
last 20 k years) can improve our prediction of the distribution of biocrus
To provide empirical support for this hypothesis, we used a combination o



MASSIVE OPEN ONLINE COURSES
Digital ways of knowing and learning

HARRY POTTER AND
THE MOOC MODEL FOR DIGITAL PRACTICE:

Harry Potter and Foundations of Data Science

What will you do in this course?

Syllabus

Prerequisites

Schedule

Blackboard Page

Final Project

Sorry, can't do this!

Final Project

GOAL

To become an independent data scientist and work on a full DS project from conceptualization to reporting

DELIVERABLES

- Topic Proposals
- Dataset
- Data Analysis Plan
- Jupyter Notebook (aka DS Report)
- Poster Presentation

Foundations of Data Science

Data Science

What is **DATA**?

What is **SCIENCE**?

What is **DATA SCIENCE**?

Data

Derived from the Latin word “datum”

Datum means *given*

Plural version of datum

Datum is and data are

Anything you can parameterize

Giving a metric to things we observe and measure

Science

Science Council:

The pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence.

Science

Science Council:

The **pursuit** and **application** of knowledge and understanding of the natural and social world following a systematic methodology based on evidence.

The Scientific Method

Hypothesize

Operationalize

Measure

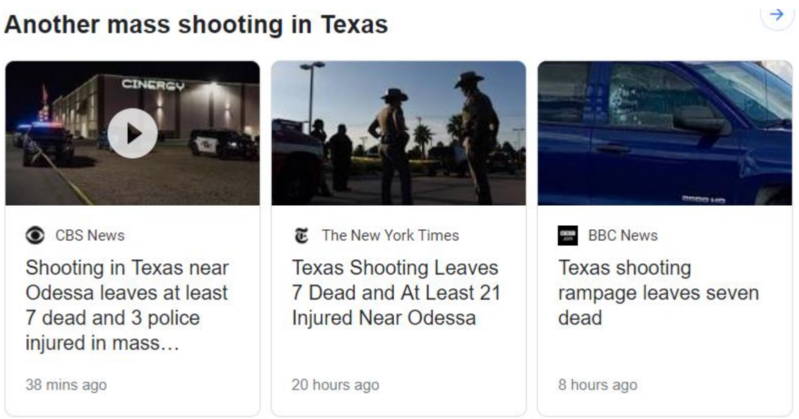
Evaluate

Report/revise/replicate

Hi, I'm **HOMER**!

The Scientific Method: A Sad Example

Are mass shootings on the rise in the US?



Another mass shooting in Texas

CBS News

Shooting in Texas near Odessa leaves at least 7 dead and 3 police injured in mass...

38 mins ago

The New York Times

Texas Shooting Leaves 7 Dead and At Least 21 Injured Near Odessa

20 hours ago

BBC News

Texas shooting rampage leaves seven dead

8 hours ago

The Scientific Method: A Sad Example

Hypothesize

Mass shootings are more common nowadays than before.

Operationalize

Variables: **# of mass shootings** over **years**

Measure

Obtain secondary, historical data and retrieve the above variables

Evaluate

Analyze data to examine the change in # of mass shootings over years

Report/revise/replicate

Interpret and report your findings

Data Science

This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.

DS Initiative, University of Michigan

Data Science

The interdisciplinary study and practice of computationally extracting meaningful insights from data

Three components:

Exploration ➔ identifying patterns in data (messing around)

Prediction ➔ making informed guesses

Inference ➔ quantifying our degree of certainty

Data Scientist

“Data Scientist” means a professional who uses scientific methods to liberate and create meaning from raw data.

Data Science Association's “Professional Code of Conduct”

Data Scientist (n.):

A person who is better at statistics than any computer scientist and better at computer science than any statistician.

Data Scientist (n.):

A data wizard who extracts meaningful insights from data using computation and statistics.

Five Main Activities in DS

1. Data Exploration and Preparation

2. Data Representation and Transformation

3. Computing with Data

4. Data Visualization and Presentation

5. Data Modeling

David Donoho (2015)

Why Python for DS?

One of the most popular interpreted programming languages

Easy to read, understand, and write Python code

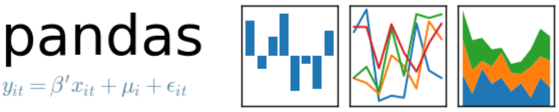
Improved support for data analysis libraries

e.g., NumPy, pandas, and scikit-learn

A common must-have skill for many DS positions

Essential Python Libraries for DS

Numerical Python



Data Manipulation

Data Visualization

Statistical tests

Machine Learning

Jupyter Notebook

A browser-based interactive GUI to the IPython shell

Now standard programming environment for DS projects

Reminders

Read Donoho (2015)

Sections 1, 2, and 8 only

Syllabus and Scheduled Quiz

Due Monday, September 9

Prior Learning Assessment

Due Friday, September 6 (preferred), or Saturday, September 7