

Introductions and Course Logistics

INST414 - Data Science Techniques

INST414-010*

Data Science
Techniques



**Wait, this isn't
woodshop?**

Lecture:

SKN 0200

Tuesdays

2:00-3:15 PM

Discussion:

HBK 0302J

Friday

9:00-10:15 AM

10:30-11:45 AM



**Wait, this isn't
woodshop?**

Notes about classroom policy

Tuesdays

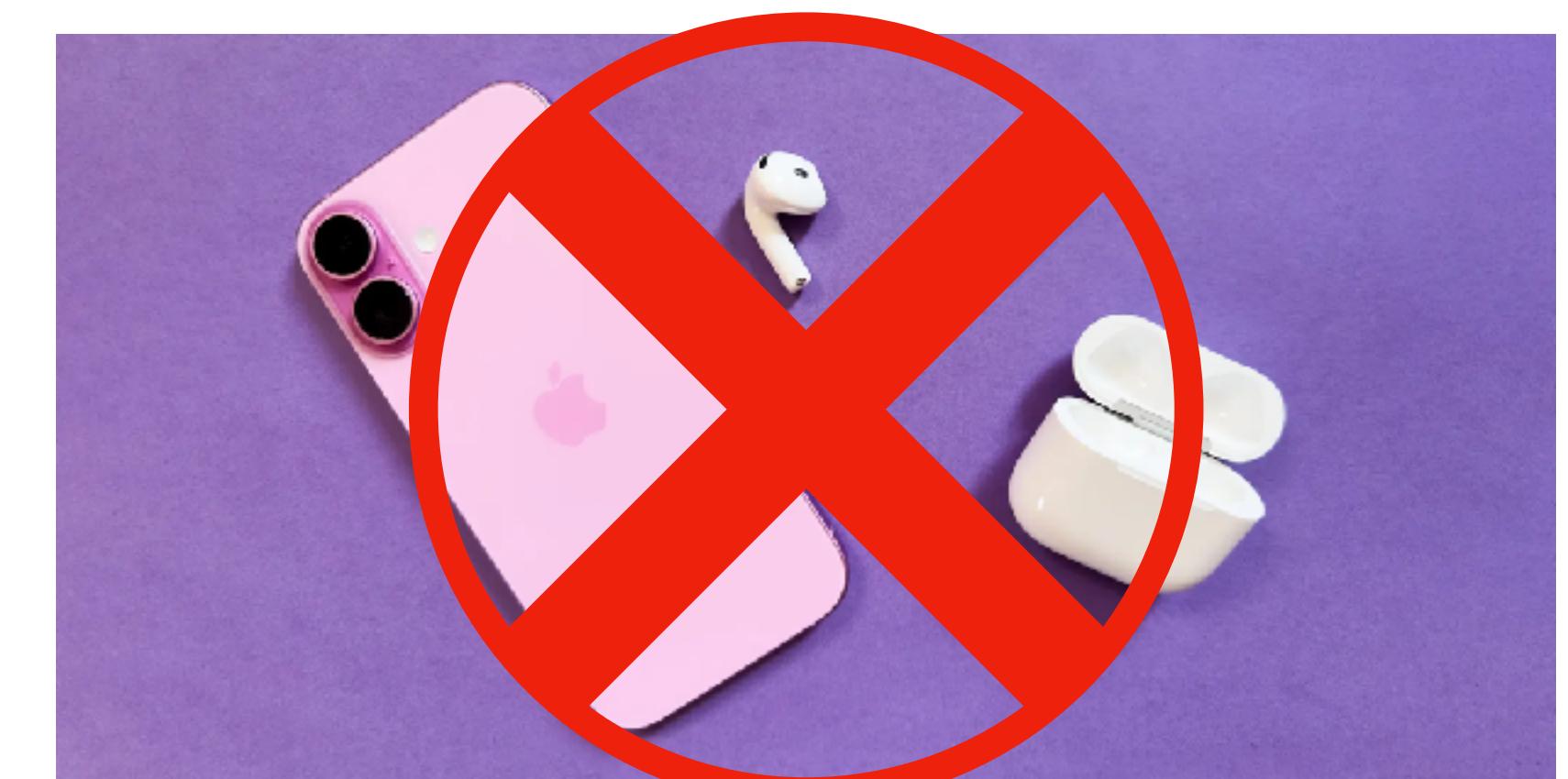


Fridays

Tuesdays



Fridays



Topics for Today

Introducing your instructional team

Course schedule/overview

Assignments over the semester

An intro to data science as a topic



Who am I?

Assistant Professor



COLLEGE OF
INFORMATION
STUDIES

Information and technology for good

Assistant Professor, Informatics

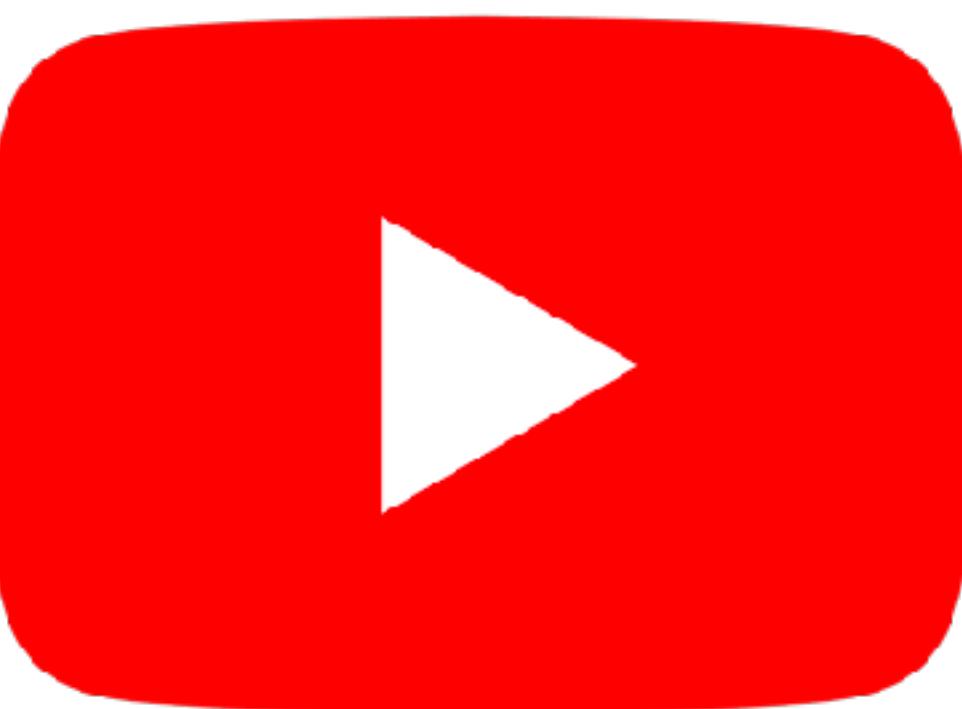
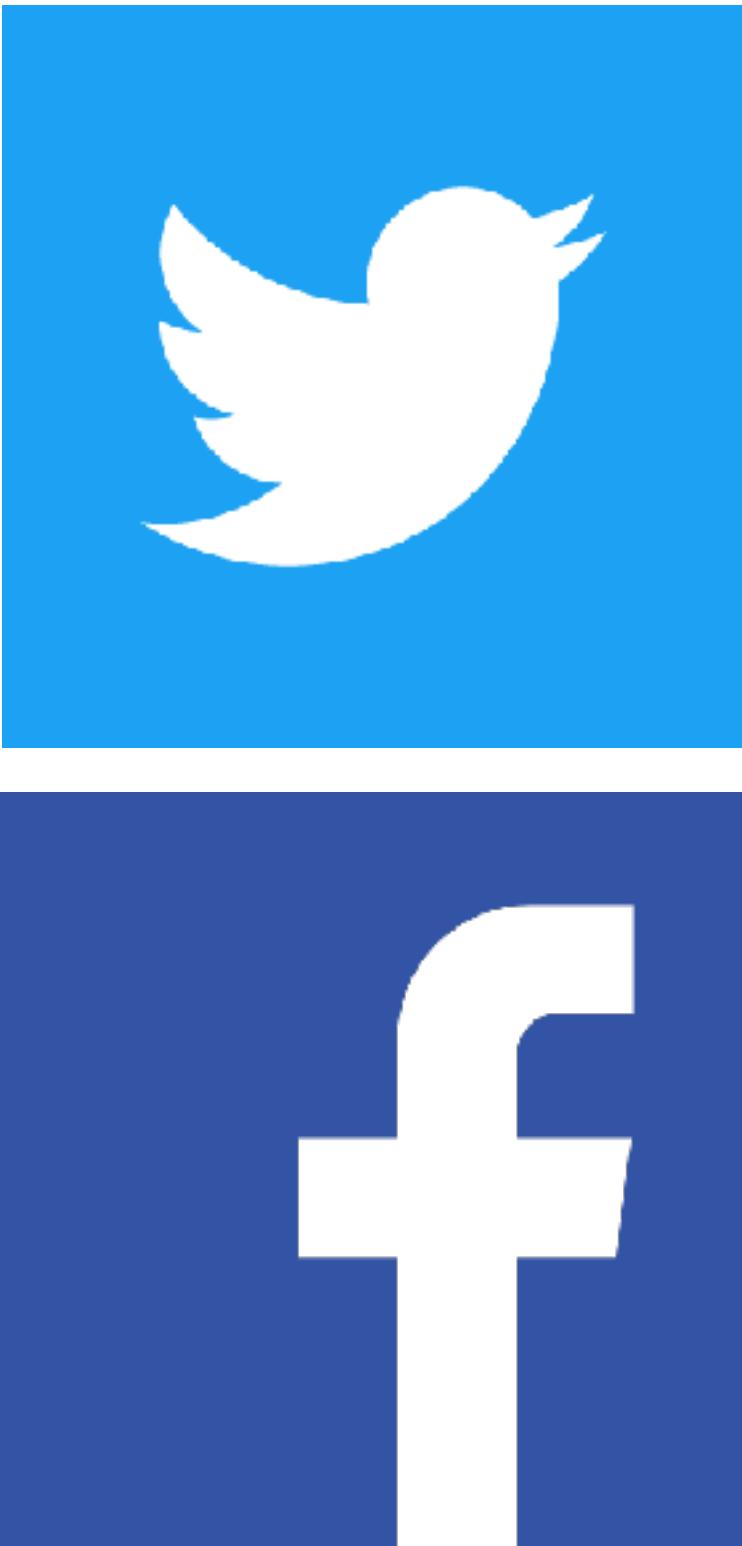
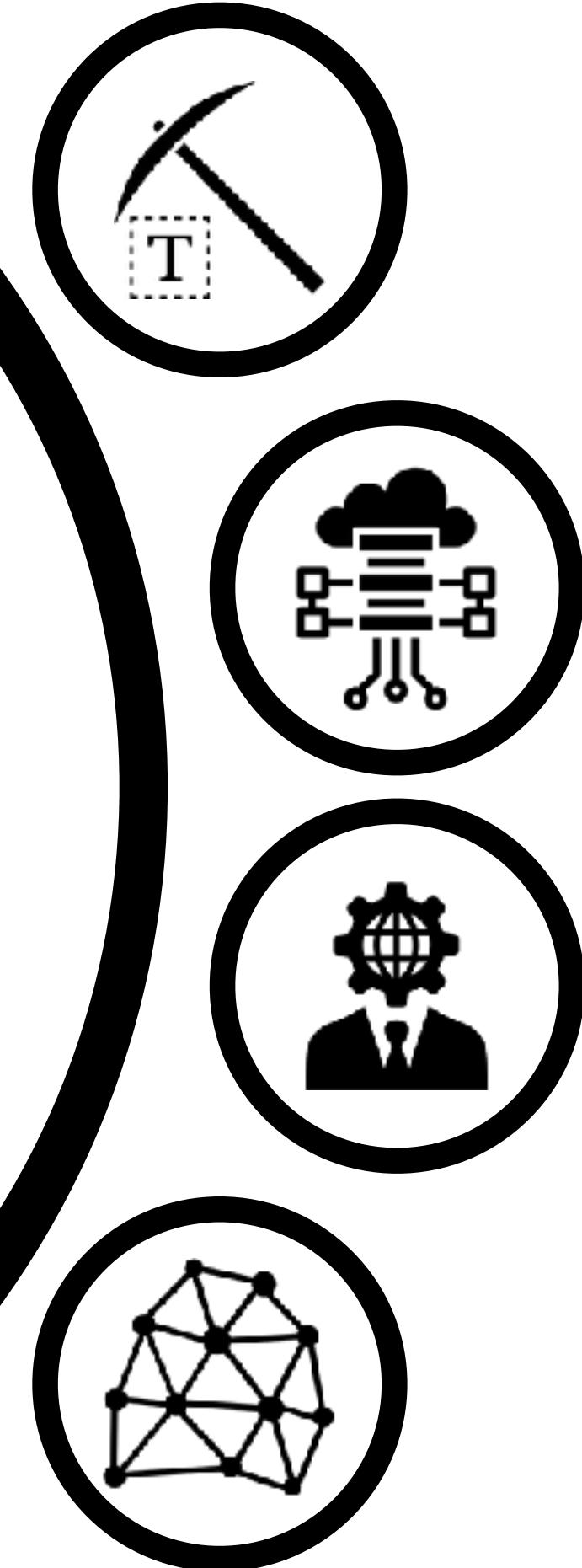
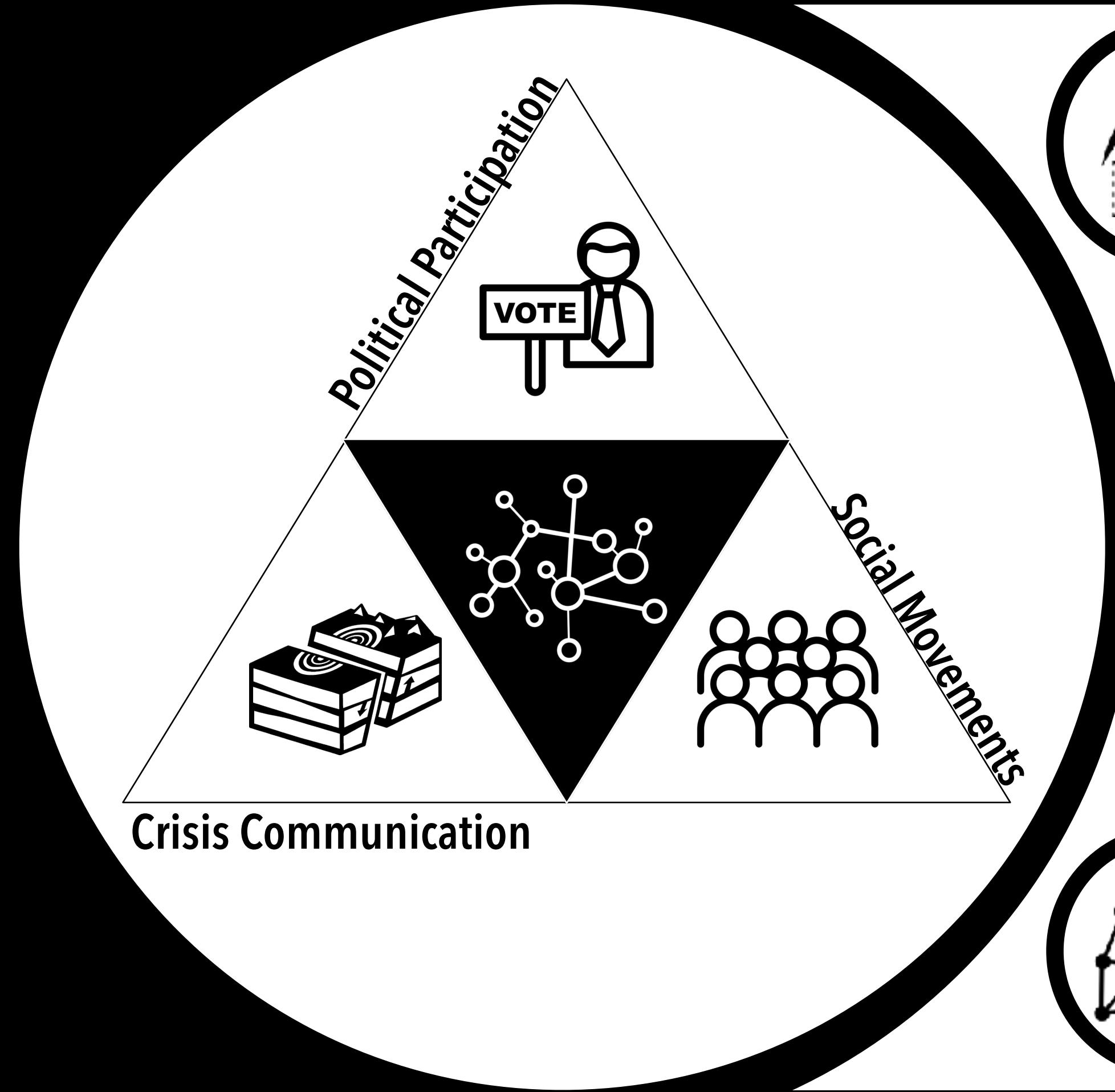


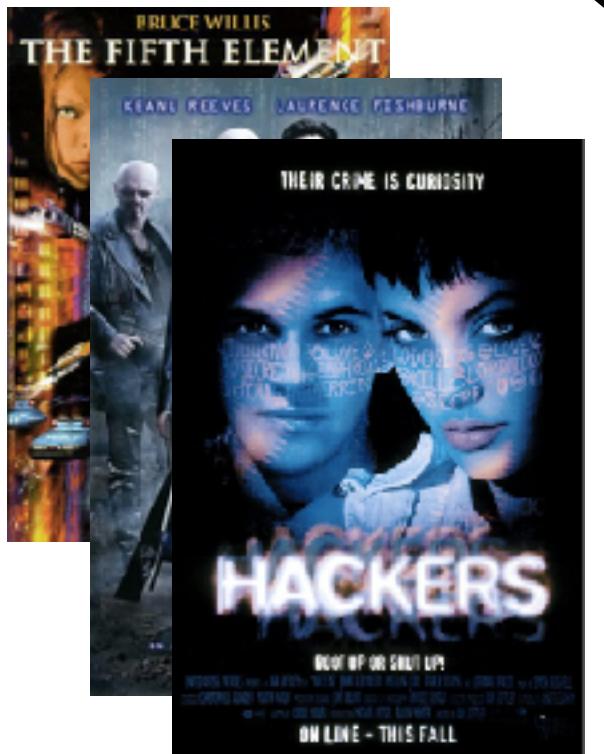
Postdoc, NYU's SMaPP Lab



PhD in CS from UMD







**TEACHER
ASSISTANT**
—
BECAUSE
MIRACLE WORKER
ISN'T AN OFFICIAL
JOB TITLE

Who are your TAs?



Instructional TA: Uday Krishnarajapuram



Grader: Amaan Mohammed



UCA: TBD

TA Office Hours

- Uday will have hours of office hours in-person and online
- Hours will be posted on ELMS/Canvas



WHO ARE YOU PEOPLE?



IT'S TOO MUCH!

#SCHITTSREEK

Pop
TV



Does the professor know you?

Due Dec 7 at 11:59pm | 100 pts



Does the TAs know you?

Due Dec 7 at 11:59pm | 100 pts





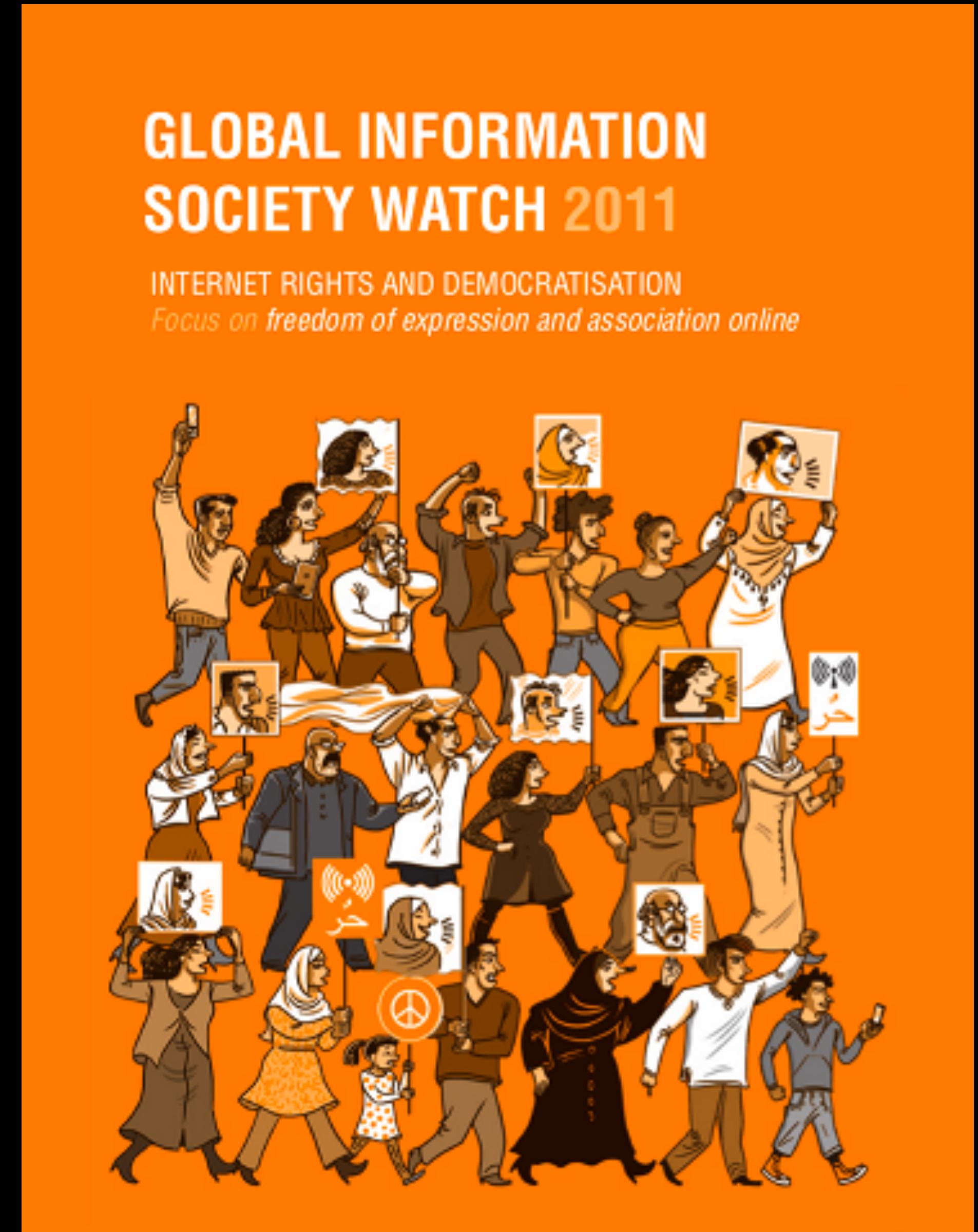
WHERE IS AY-AY-RON RIGHT NOW?

Course Material



Prerequisites

- INST201/301 or BSOS233
- What are the roles of information in society?
- How might tech be problematic, and how might we harness it for positive change?
 - Guide your insights for articles and semester project



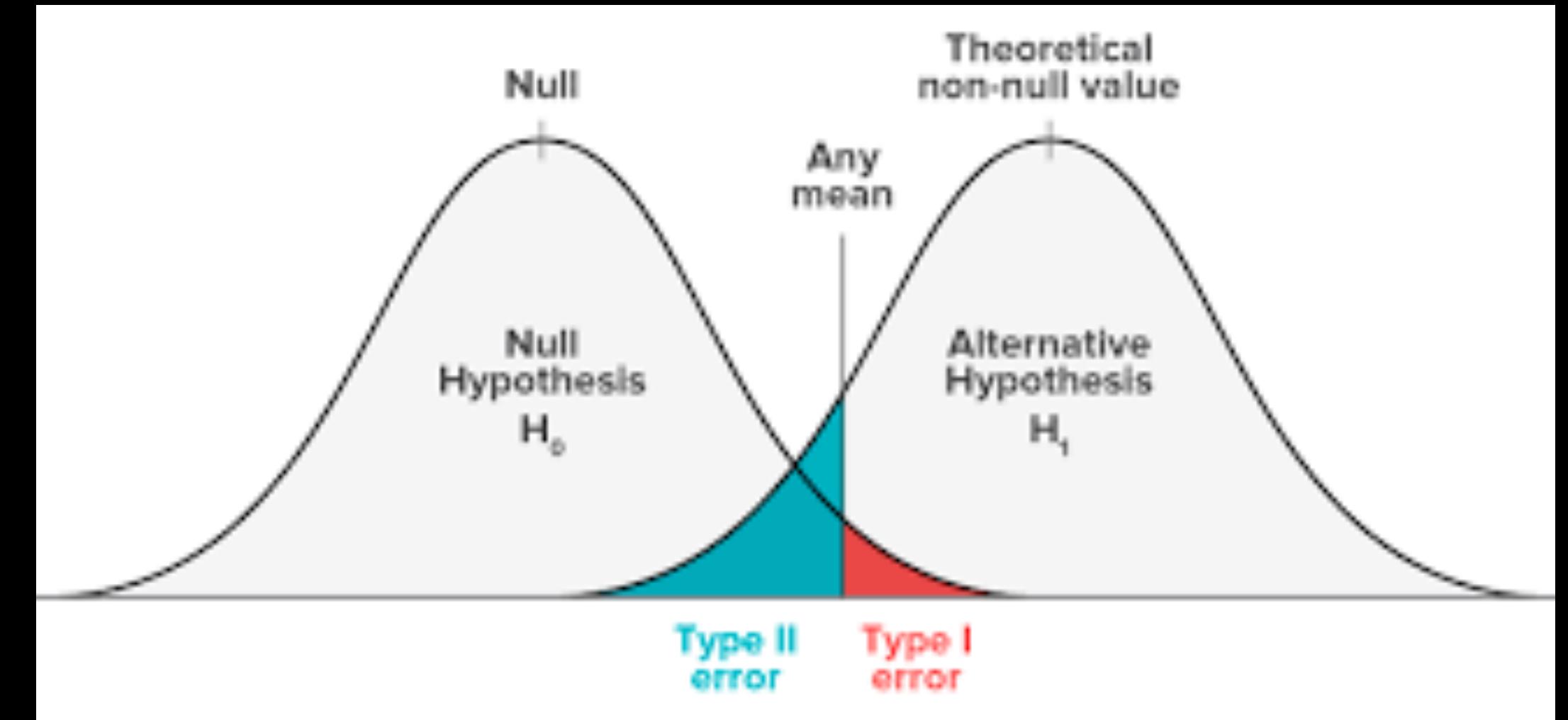
Prerequisites

- AASP101, ANTH210, ANTH260, ECON200, ECON201, GEOG202, GVPT170, PSYC100, or SOCY100



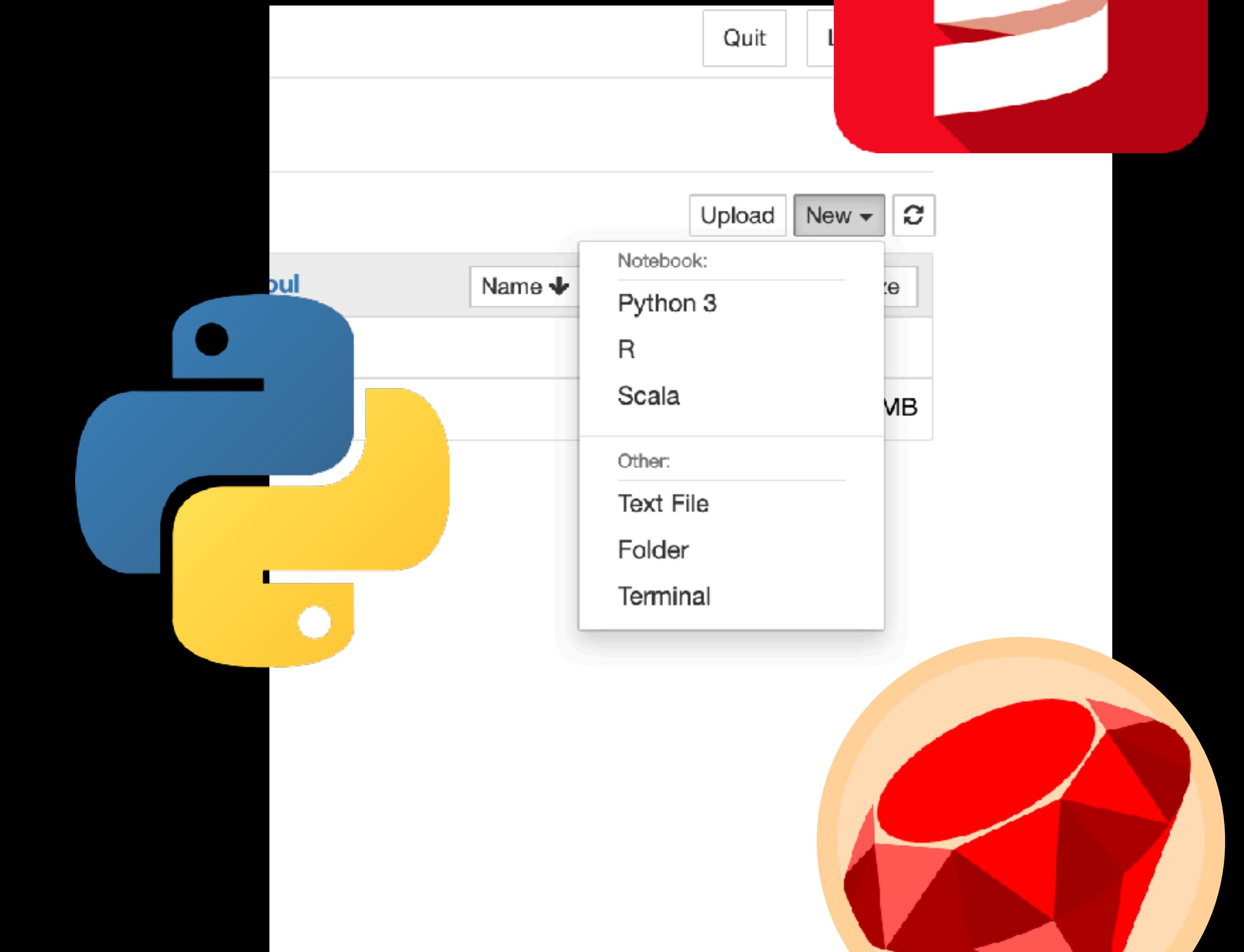
Prerequisites

- STAT100/MATH115 - Stats and precal
- Hypothesis testing
- Regression
- Some linear algebra (vectors and matrices)



Prerequisites

- INST126 - programming
- Proficiency in programming essential to this class
- Python is preferred
 - R, Java, Ruby, Scala, or Perl may be good substitutes



INST414 > Modules

Spring 2024

Home Collapse All View Progress Publish All + Module :

Course Logistics

Class Preparation

- Pre-Class Skill Check - Most Frequent Tokens in Text (Feb 2 | 100 pts)
- Participation - Introductions (100 pts)

Resources

- Syllabus
- Previous Semester's YouTube Playlist
- Course GitHub Repo

Office Hours

- Nidhi's Virtual Office Hours: Tuesdays, 12-1:30 pm
- Buntain Hours: Mondays, 4:00-5:00 PM, and by appt

24/7 Canvas Chat Support
....or call 1-833-566-3347 (staff/faculty)
1-877-399-4090 (students)

Course Status

Unpublish Published

Files

- Inbox
- Syllabus
- Outcomes
- Rubrics
- Quizzes
- Modules
- BigBlueButton
- Collaborations
- Chat
- Panopto Recordings
- New Analytics
- Clickers
- Course Reserves
- Adobe Creative Cloud

Import Existing Content

Import from Commons

Choose Home Page

View Course Stream

Course Setup Checklist

New Announcement

New Analytics

View Course Notifications

To Do

You should see this skill-check in Canvas

Pre-Class Skill Check - Most Frequent Tokens in Text

umd.instructure.com/courses/1361527/assignments/6665554?module_item_id=12579842

Published Edit

24/7 Canvas Chat Support
....or call 1-833-566-3347 (staff/faculty)
1-877-399-4090 (students)

Motivation

This course requires experience in programming, with a pre-requisite of at least one Python-based programming course. To ensure you have the experience necessary to succeed in this course, you should be able to complete the assignment below with minimal new learning (beyond familiarizing yourself with the NLTK library).

If you have trouble completing this skill-check, you will find this class overly difficult. This class assumes a certain level of expertise as we discuss data-science techniques. That is, this course is heavily focused on concepts in data science and how we apply them via programming; the class is not designed to teach you how to program. If you have difficulty here, you are encouraged to reach out to the professor for guidance as soon as possible.

Overview

Write a Python script to read a text file and output the top 10 most common words in the file and their counts. You should use the [Natural Language ToolKit \(NLTK\) library](#) to handle splitting words out of the text for you. NLTK provides a casual tokenizer that works well for "casual" text, such as social media posts (e.g., tweets). Documentation for this tokenizer function is available here: <https://www.nltk.org/api/nltk.tokenize.casual.html>

Requirements:

Your program should satisfy the following requirements.

- Use the TweetTokenizer tokenizer in NLTK (<https://www.nltk.org/api/nltk.tokenize.casual.html>)
- NLTK has multiple tokenizers (e.g., `word_tokenize`), but you should use the TweetTokenizer one, so you have consistent output.
- The input file (i.e., the file to be read by the program) should be named `input.txt`.
- For example, if you wrote a Python program called CommonTokenCounter.py, you would run it on a given text file by calling:
CommonTokenCounter.py input.txt

Related Items

SpeedGrader™

Download Submissions

0 out of 2 Submissions Graded

How to use UMD Canvas ▾

Textbooks

Adopt Textbook

Spring 2024

Home

Announcements

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Outcomes

Rubrics

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

New Analytics

Clickers

Course Reserves

Adobe Creative Cloud

Quiz Extensions

Settings

Account

Dashboard

Courses

Calendar

Inbox

Portfolium

History

Commons

CourseExp

Help

EMT

Settings

You should see this skill-check in Canvas

Sign In Compliance | Discussions: INST414-0102 | Pre-Class Exercise - Most F | University of Maryland, Coll | ScholarOne Manuscripts | Medium

umd.instructure.com/courses/1353667/assignments/6482452?module_item_id=12218829

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello CHI24 SMDRM Dri...

Fall 2023

Testing

I have provided two text files and the appropriate output for you to use to check your work:

- File [test01_cc_sharealike.txt](#) is a brief overview of the Creative Commons Sharealike license
- File [test02_the_last_question.txt](#) is a copy of Isaac Asimov's "The Last Question" short story.

Test 1: Expected Output for [test01_cc_sharealike.txt](#)

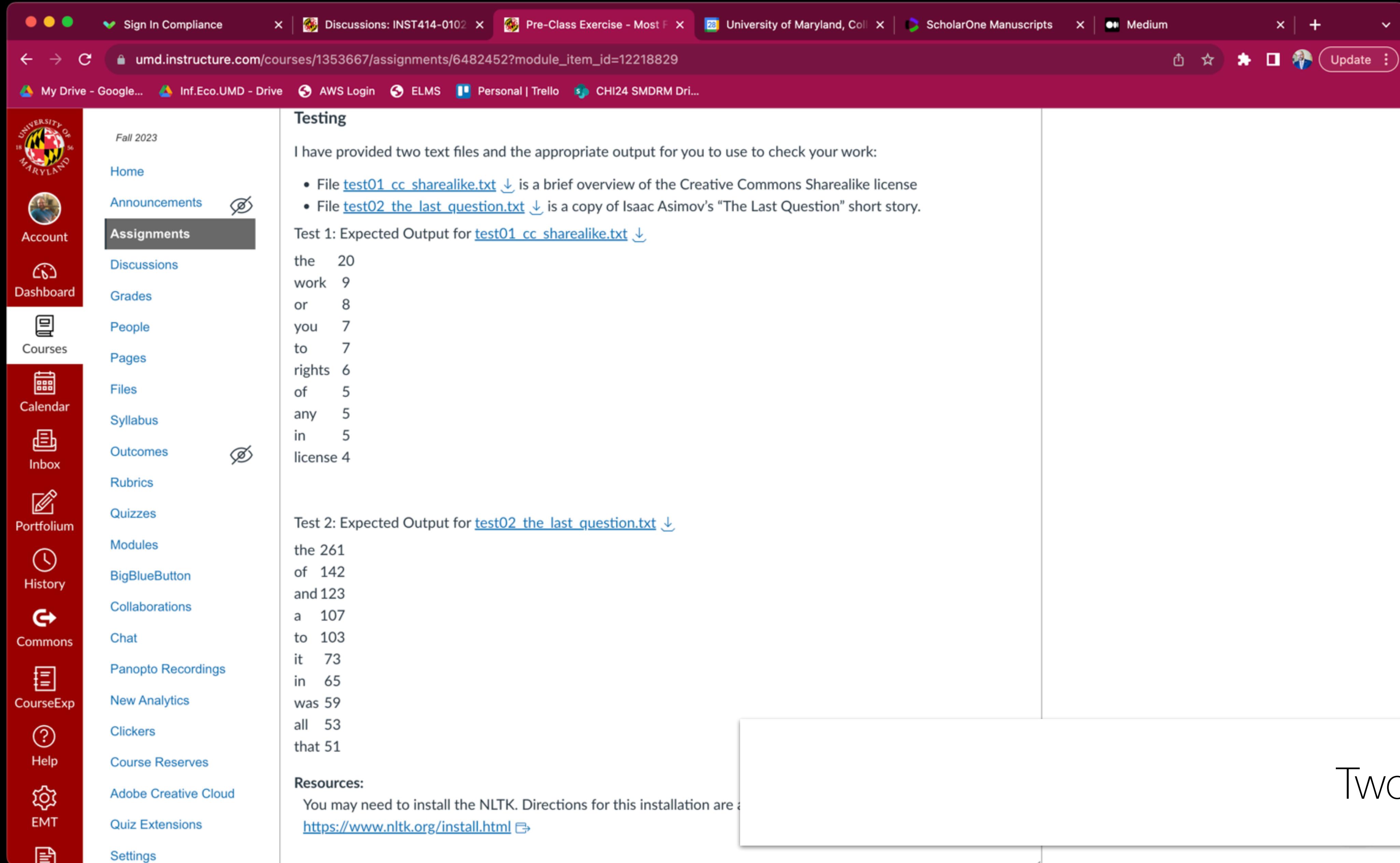
```
the 20
work 9
or 8
you 7
to 7
rights 6
of 5
any 5
in 5
license 4
```

Test 2: Expected Output for [test02_the_last_question.txt](#)

```
the 261
of 142
and 123
a 107
to 103
it 73
in 65
was 59
all 53
that 51
```

Resources:

You may need to install the NLTK. Directions for this installation are at <https://www.nltk.org/install.html>



Two test files

Pre-Class Skill Check - Most Frequent Tokens in Text

umd.instructure.com/courses/1361527/assignments/6665554?module_item_id=12579842

Published Edit

24/7 Canvas Chat Support
....or call 1-833-566-3347 (staff/faculty)
1-877-399-4090 (students)

Motivation

This course requires experience in programming, with a pre-requisite of at least one Python-based programming course. To ensure you have the experience necessary to succeed in this course, you should be able to complete the assignment below with minimal new learning (beyond familiarizing yourself with the NLTK library).

If you have trouble completing this skill-check, you will find this class overly difficult. This class assumes a certain level of expertise as we discuss data-science techniques. That is, this course is heavily focused on concepts in data science and how we apply them via programming; the class is not designed to teach you how to program. If you have difficulty here, you are encouraged to reach out to the professor for guidance as soon as possible.

Overview

Write a Python script to read a text file and output the top 10 most common words in the file and their counts. You should use the [Natural Language ToolKit \(NLTK\) library](#) to handle splitting words out of the text for you. NLTK provides a casual tokenizer that works well for "casual" text, such as social media posts (e.g., tweets). Documentation for this tokenizer function is available here: <https://www.nltk.org/api/nltk.tokenize.casual.html>

Requirements:

Your program should satisfy the following requirements.

- Use the TweetTokenizer tokenizer in NLTK (<https://www.nltk.org/api/nltk.tokenize.casual.html>)
- NLTK has multiple tokenizers (e.g., `word_tokenize`), but you should use the TweetTokenizer one, so you have consistent output.
- The input file (i.e., the file to be read by the program) should be named `input.txt`.

For example, if you wrote a Python program called `CommonTokenizer`, you would run it on a given text file by calling:

```
python CommonTokenizer.py input.txt
```

Related Items

SpeedGrader™

Download Submissions

0 out of 2 Submissions Graded

How to use UMD Canvas ▾

Textbooks

Adopt Textbook

Spring 2024

Home Announcements Assignments Discussions Grades People Pages Files Syllabus Outcomes Rubrics Quizzes Modules BigBlueButton Collaborations Chat Panopto Recordings New Analytics Clickers Course Reserves Adobe Creative Cloud Quiz Extensions Settings

Account Dashboard Courses Calendar Inbox Portfolium History Commons CourseExp Help EMT Settings

Due Sept 5

Questions about the skill-check?

INST414-0103: Data Science X +

umd.instructure.com/courses/1361527

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello UMIACS Object St...

UNIVERSITY OF MARYLAND

INST414 > Modules

Spring 2024

Home

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

Clickers

Collapse All

Course Logistics

Class Preparation

Pre-Class Skill Check - Most Frequent Tokens in Text

Feb 2 | 100 pts

Participation - Introductions

100 pts

Resources

Syllabus

Previous Semester's YouTube Playlist

24/7 Canvas Chat Support

....or call 1-833-566-3347 (staff/faculty)
1-877-399-4090 (students)

View Course Stream

View Course Calendar

View Course Notifications

To Do

Quiz, Week 2

INST414-0103: Data Science Techniques- Spring 2024 cbuntain
7 points | Jan 31 at 11:59pm

In-class Exercise, Week 2

INST414-0103: Data Science Techniques- Spring 2024 cbuntain
100 points | Jan 31 at 11:59pm

Pre-Class Skill Check -

<https://elms.umd.edu/>

github.com/cbuntain/umd.inst414

Product Solutions Open Source Pricing

Search Sign in Sign up

cbuntain / umd.inst414 Public

Notifications Fork 6 Star 4

Code Issues Pull requests Actions Projects Security Insights

main 1 branch 0 tags Go to file Code

Cody Buntain Added TA Example ... 71dedf3 on Apr 25, 2022 15 commits

Module	Description	Time
Module01	More Examples	last year
Module02	Fixed a bug in the IMDB data and added export functionality	10 months ago
Module03	Movie Similarity Based on User Ratings	10 months ago
Module04	Simple k-Means Example	9 months ago
Module05	Added Module 5 example	9 months ago
Module06	Added TA Example	9 months ago
data	Examples for Genre Classification and Revenue Regression	9 months ago
.gitignore	Initial commit	last year
LICENSE	Initial commit	last year
README.md	Initial commit of examples	last year

README.md

Example Code for INST414 at UMD's iSchool

Course Material for INST414, Data Science Techniques

Instructor: Cody Buntain

About

Course Material for the iSchool at UMD's INST414, Data Science Techniques

Readme Apache-2.0 license 4 stars 3 watching 6 forks

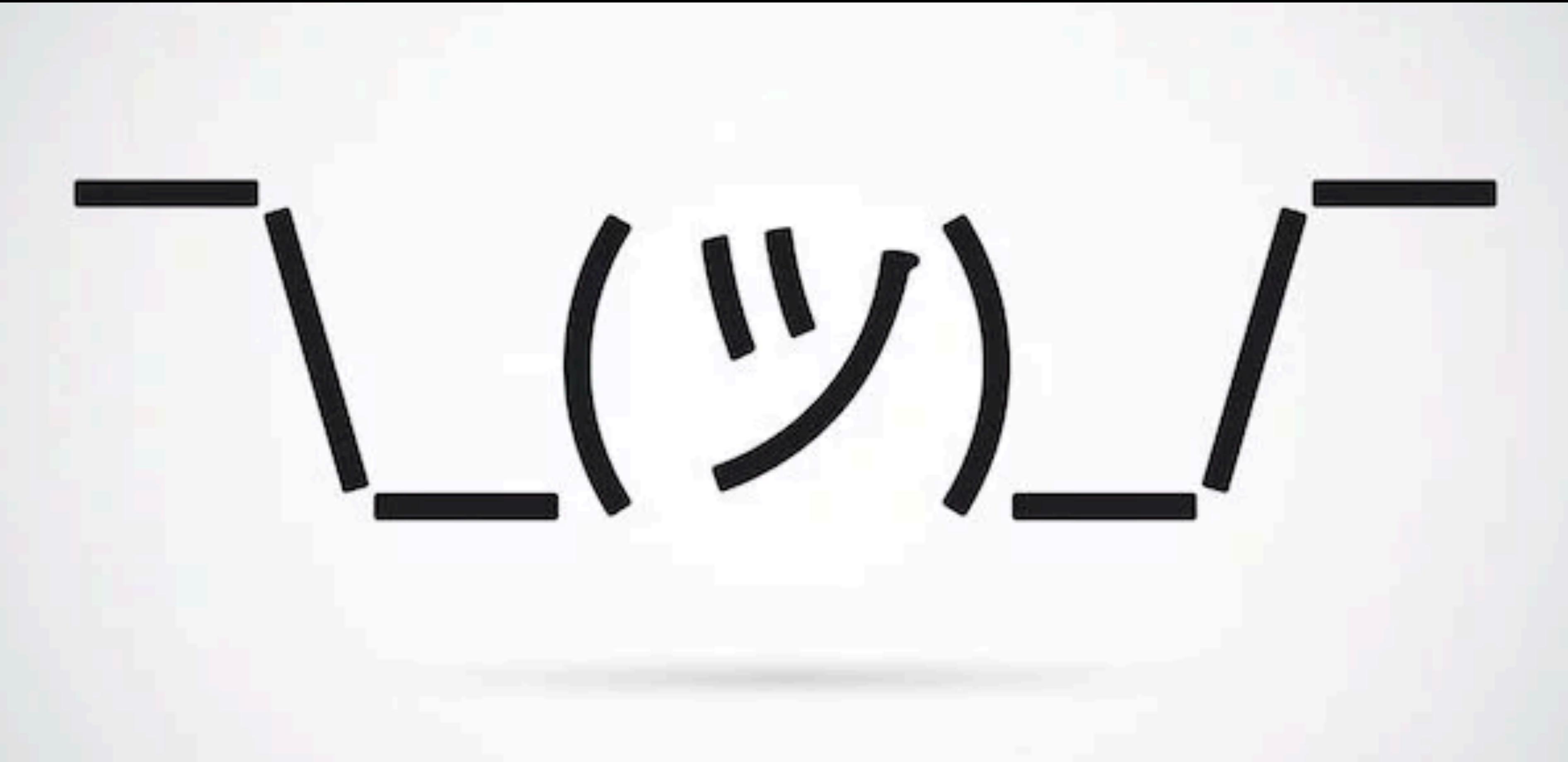
Releases No releases published

Packages No packages published

Languages Jupyter Notebook 100.0%

GitHub Repo for the class

What questions do you have?



What is INST414?

Seven Course Modules

Module 1 - Data Science and Motivations

Module 2 - Web Data as Graphs

Module 3 - Similarity, Dimensionality
Reduction, and Data Cleaning

Module 4 - Clustering

Module 5 - Probability and Bayes

Module 6 - Supervised Machine Learning

Module 7 - Evaluation

Six Core Learning Objectives



Six Core Learning Objectives

1. Collect and clean large-scale datasets

The image shows a composite of two screenshots. On the left, there's a dark background featuring several large, semi-transparent social media icons: a red YouTube play button, a blue Twitter bird, a blue Facebook 'f', and a pink Instagram camera icon. Overlaid on this is a white speech bubble containing the text "Google BigQuery" next to a blue hexagonal icon with a magnifying glass and a bar chart. On the right, there's a screenshot of a GitHub repository page for "public-apis/public-apis". The URL in the address bar is "github.com/public-apis/public-apis#sports--fitness". The repository has 3.1k stars and 108k forks. The main interface shows a list of commits from "yannbertrand" and other contributors, along with sections for "About", "Releases", "Packages", and "Contributors". At the bottom of the GitHub page, it says "A collective list of free APIs for use in software and web development."

Six Core Learning Objectives

2. Articulate the math behind supervised and unsupervised techniques

Table 1: Distance Metrics for kNN

Distance	Equation
Euclidean	$d = \sqrt{\sum_{j=1}^n (x_{sj} - x_{tj})^2}$
City Block	$d = \sum_{j=1}^n x_{sj} - x_{tj} $
Chebyshev	$d = \max_j \{ x_{sj} - x_{tj} \}$
Cosine	$d = 1 - \frac{\sum_{j=1}^n x_{sj} x_{tj}}{\sqrt{\sum_{j=1}^n x_{sj} x_{sj}} \sqrt{\sum_{j=1}^n x_{tj} x_{tj}}}$
Correlation	$d = 1 - \frac{(x_s - \tilde{x}_s)(x_t - \tilde{x}_t)'}{\sqrt{(x_s - \tilde{x}_s)(x_s - \tilde{x}_s)'} \sqrt{(x_t - \tilde{x}_t)(x_t - \tilde{x}_t)'}}$

$$\hat{\alpha} = \min_{\alpha} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = \min_{\alpha} \sum_{i=1}^n \varepsilon_i^2$$

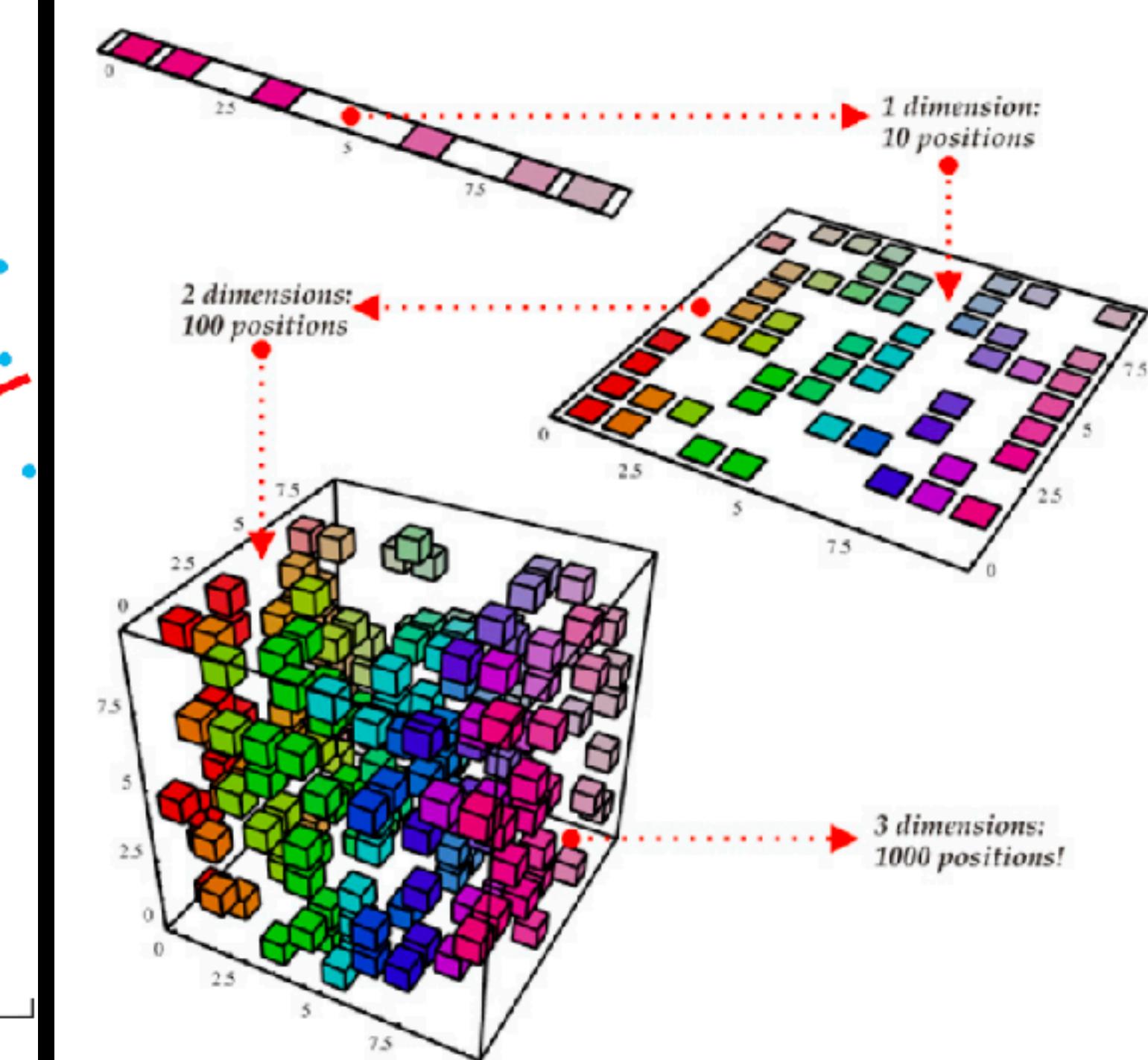
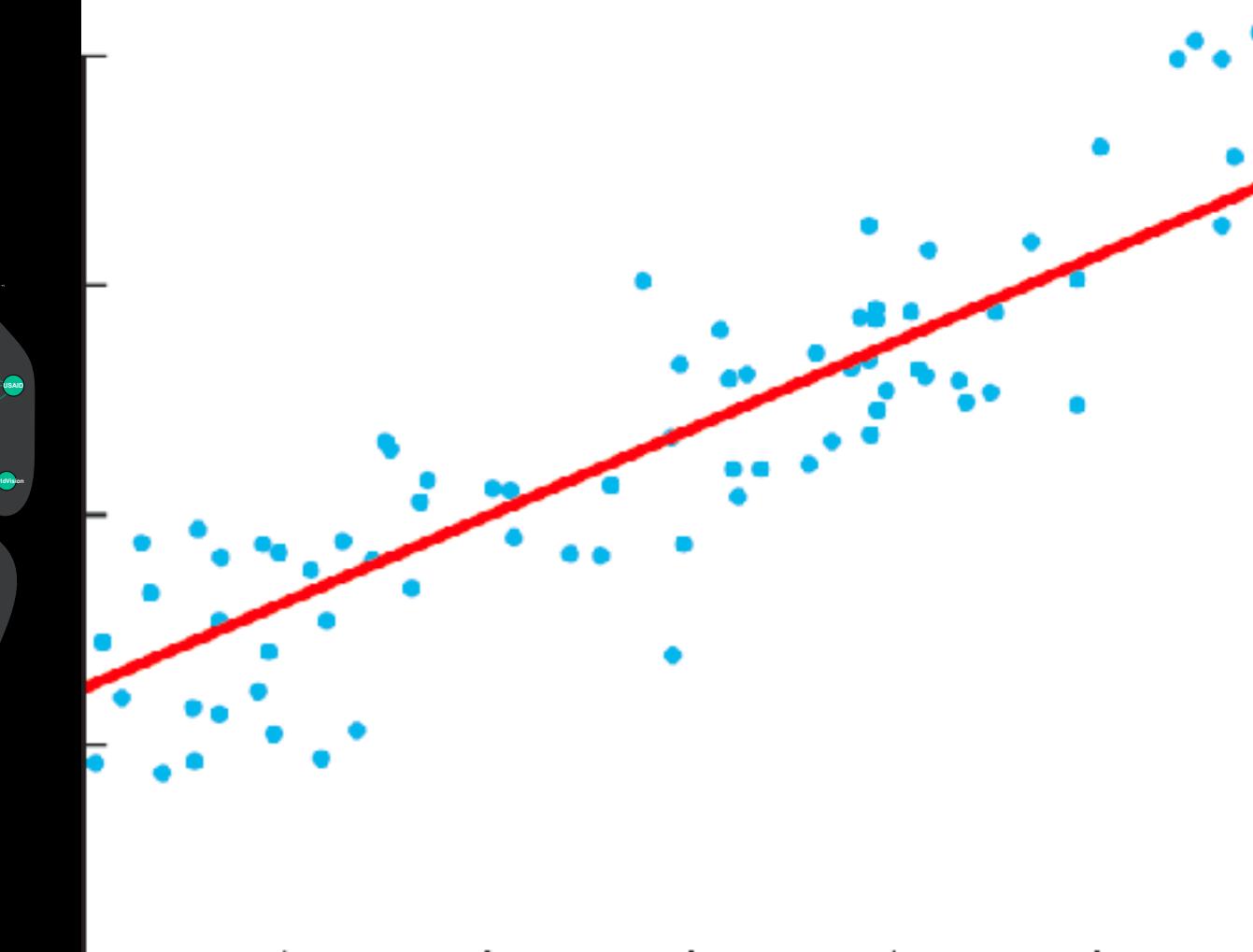
Six Core Learning Objectives

3. Execute supervised and unsupervised machine learning techniques



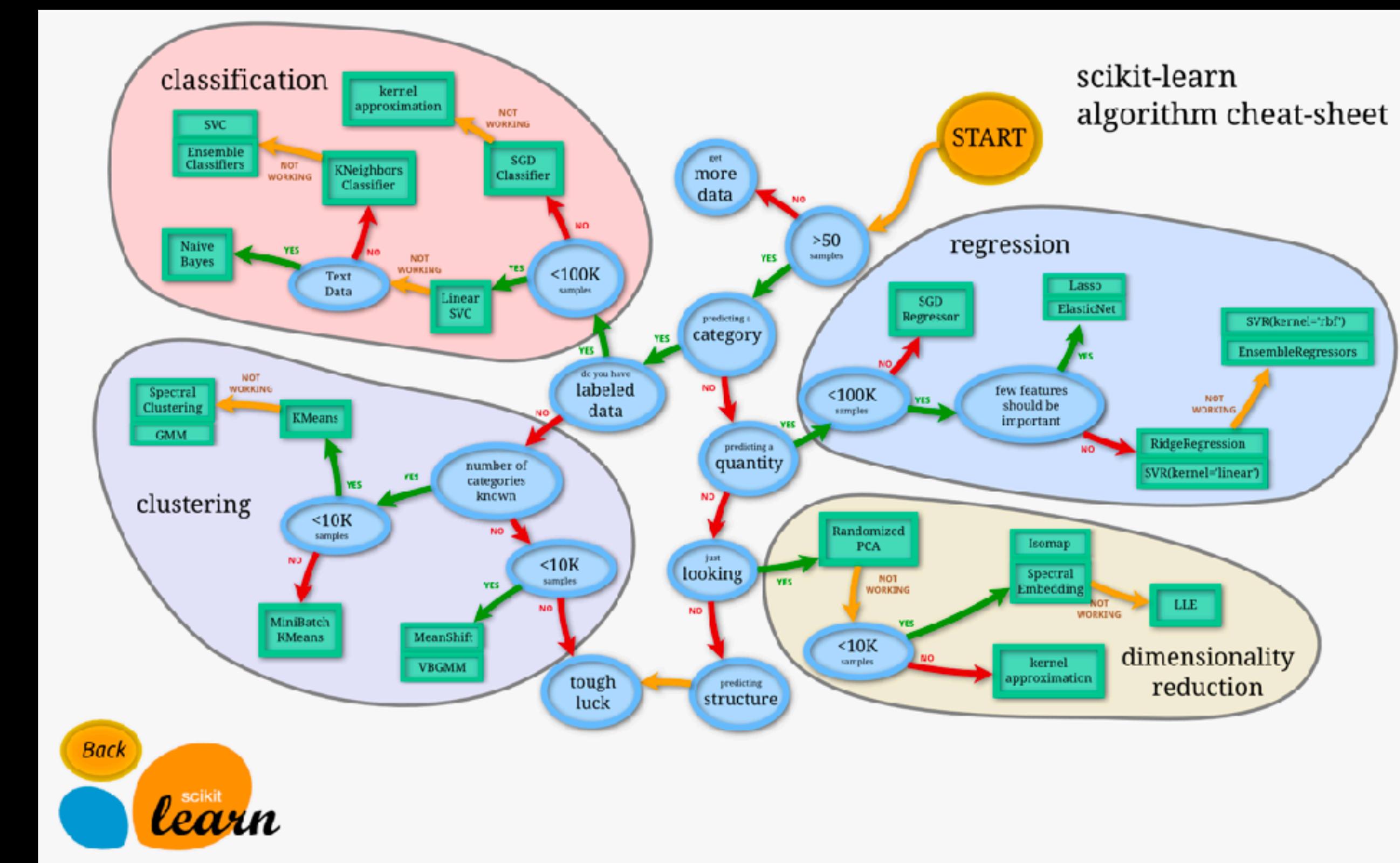
Building a Regression Model

The line summarizes the relationship between x and y.



Six Core Learning Objectives

4. Select and evaluate various types of machine learning techniques



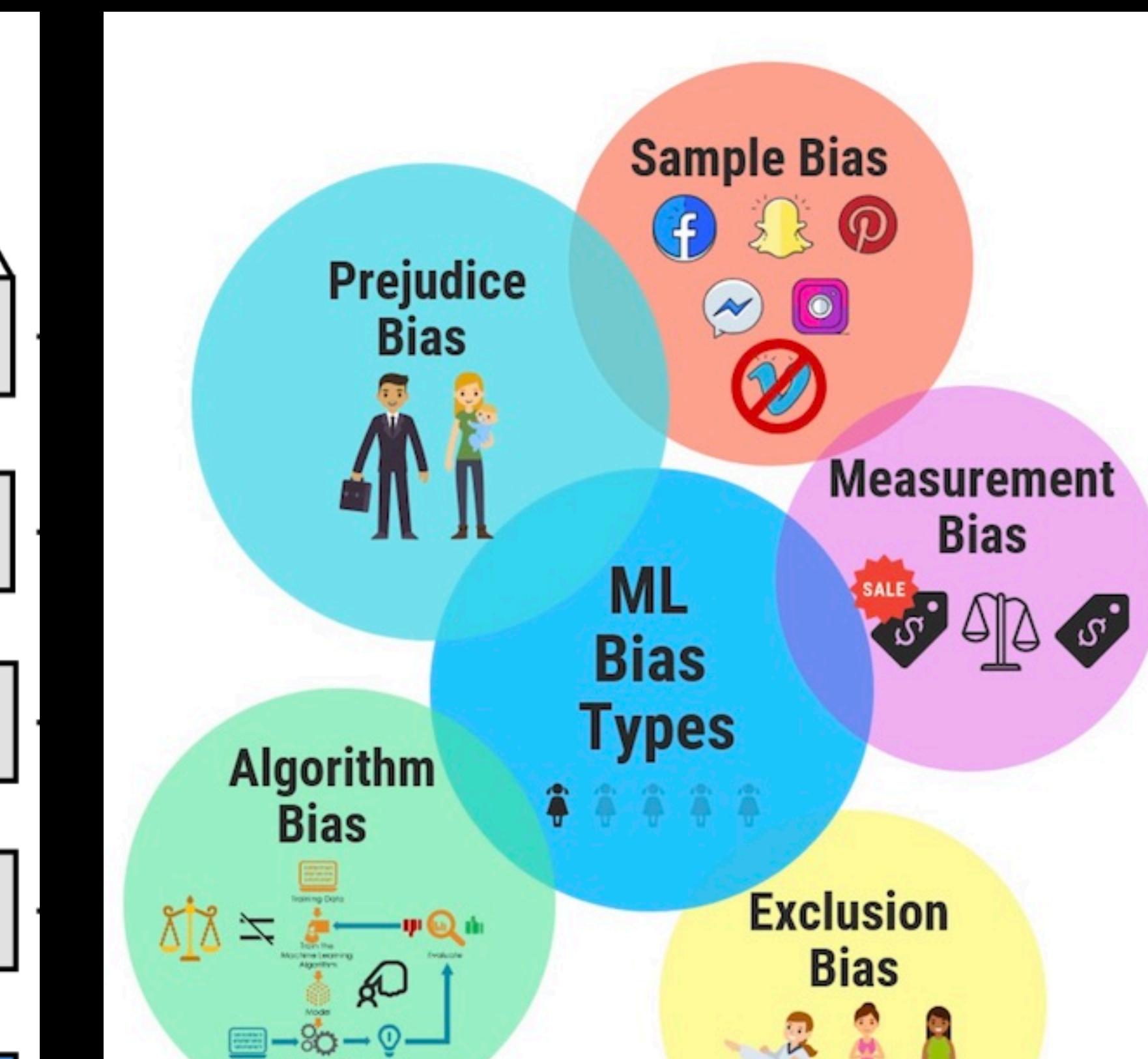
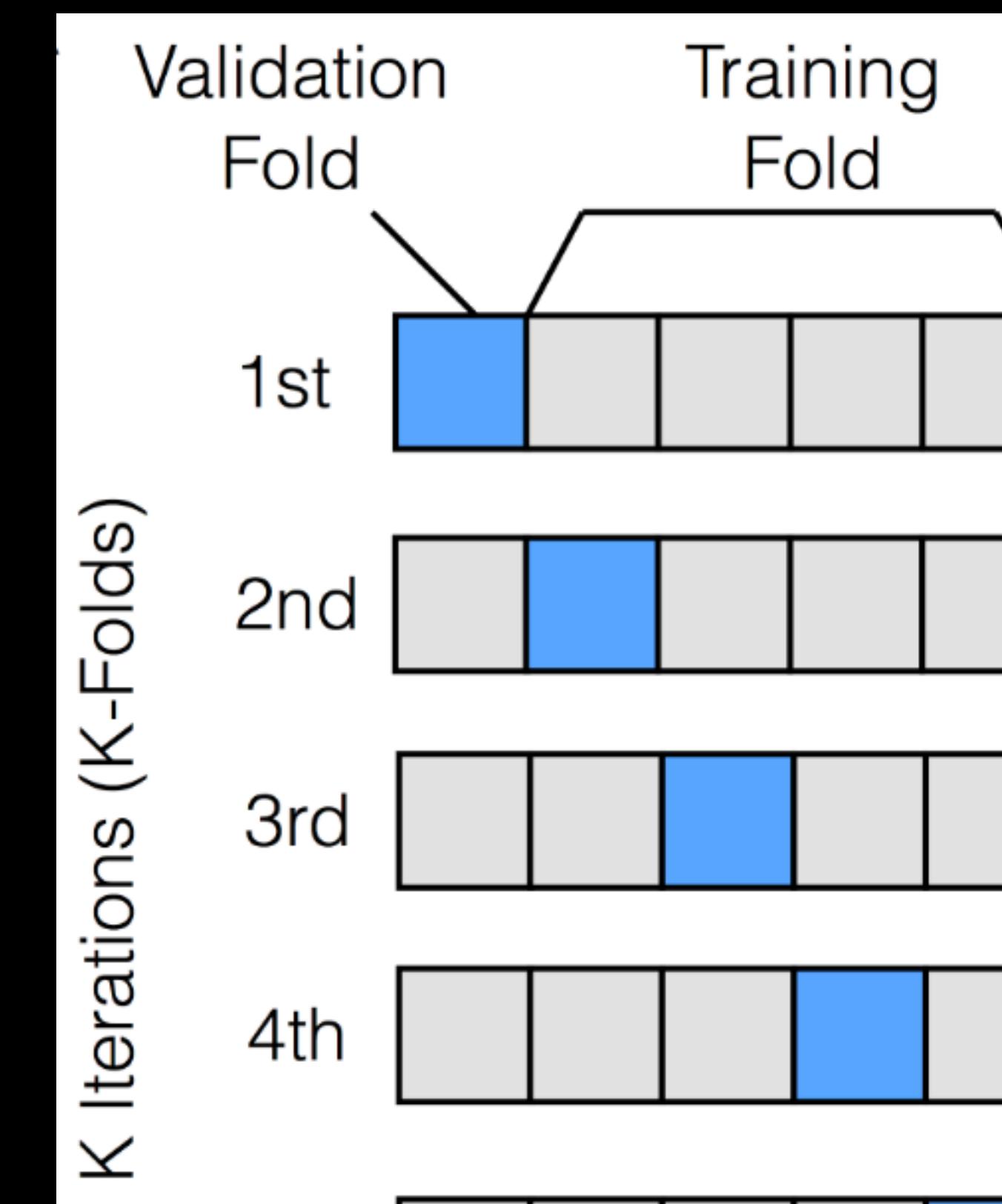
Six Core Learning Objectives

5. Explain the results coming out of the models

000 Medium

Six Core Learning Objectives

6. Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach



How to be successful in this class

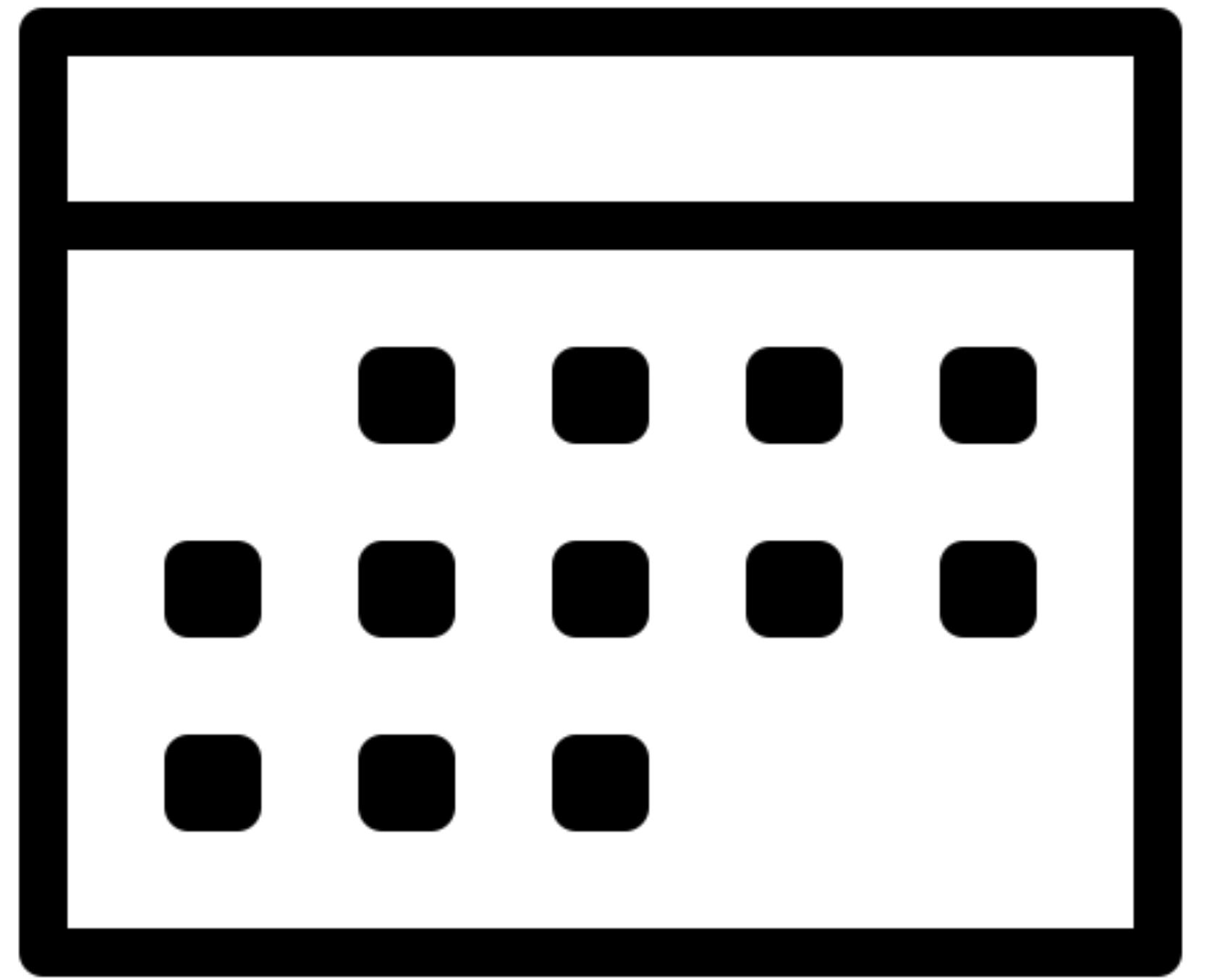
Get comfortable with programming (more than Excel)

Do the work, but don't stress about being perfect

Be clear on why you made choices, and you'll be fine

Many opportunities for extra credit

What questions do you have?



Course Schedule

Seven Course Modules

Module 1 - Data Science and Motivations

Module 2 - Web Data as Graphs

Module 3 - Similarity, Dimensionality
Reduction, and Data Cleaning

Module 4 - Clustering

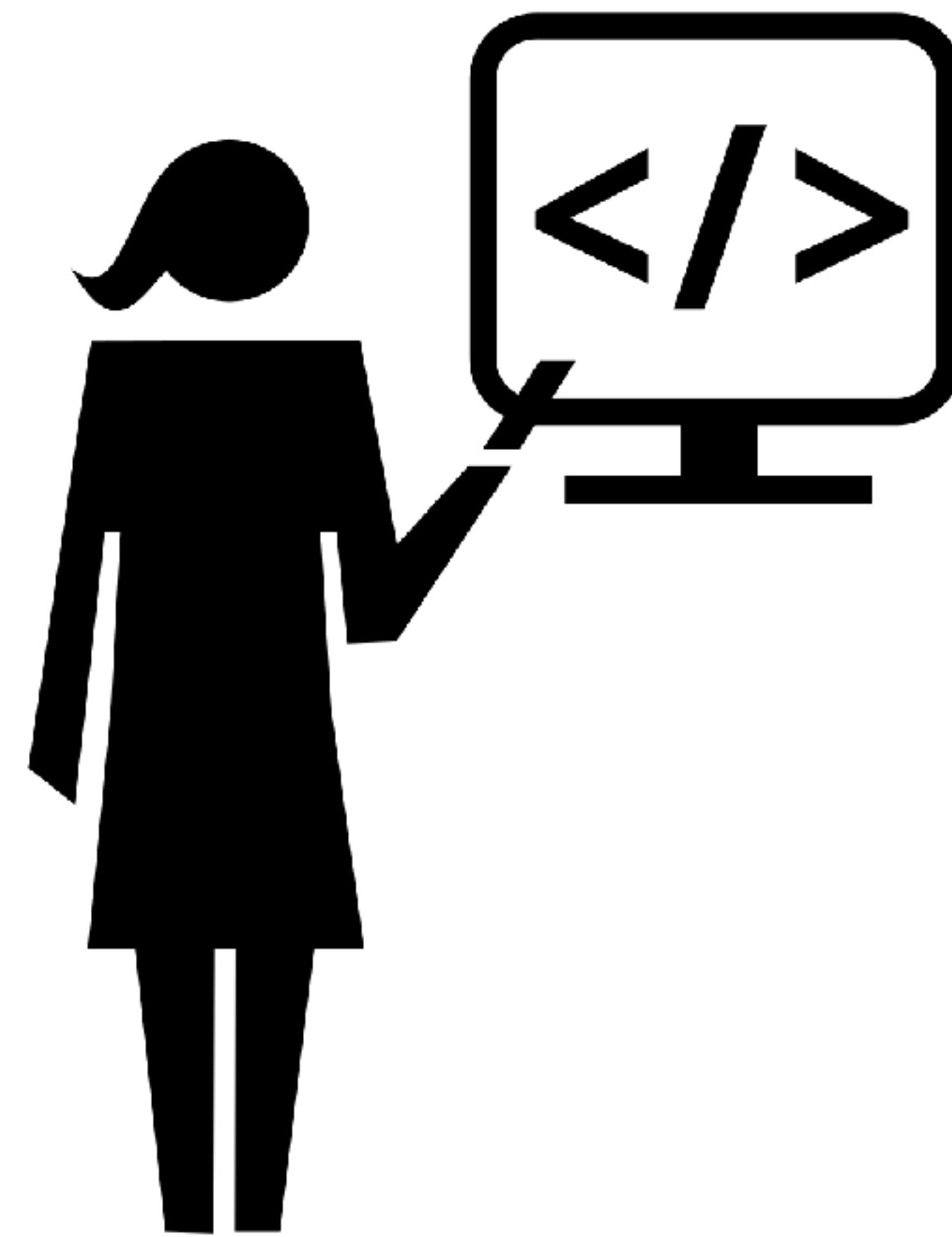
Module 5 - Probability and Bayes

Module 6 - Supervised Machine Learning

Module 7 - Evaluation

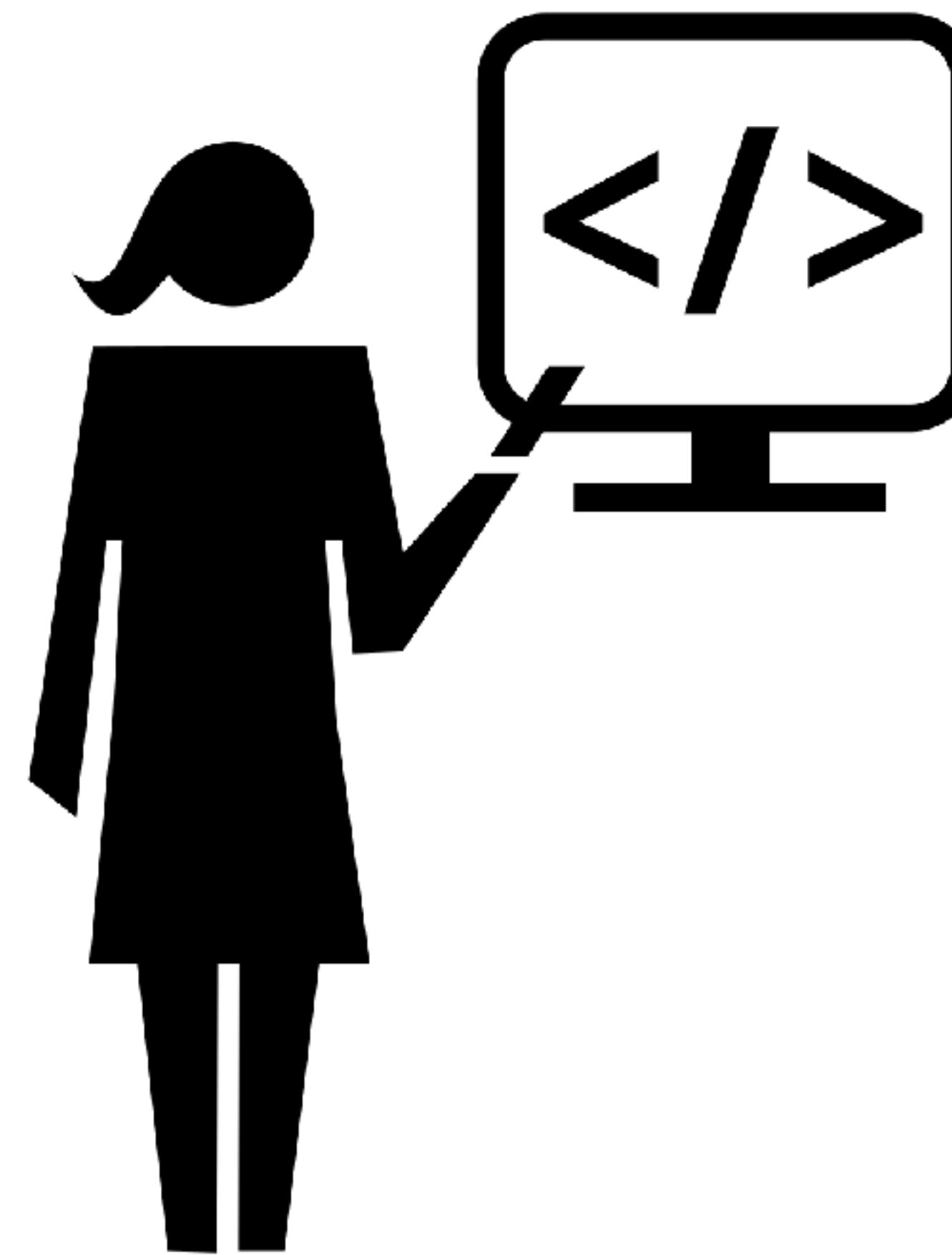
An Average Week in INST414

Tuesdays



An Average Week in INST414

Tuesdays



Fridays



INST414 Grading Distribution and Policies

NETFLIX

Four Parts to Your Grade

Module Assignments - 35%

In-Class Labs/Quizzes - 25%

Semester Project - 30%

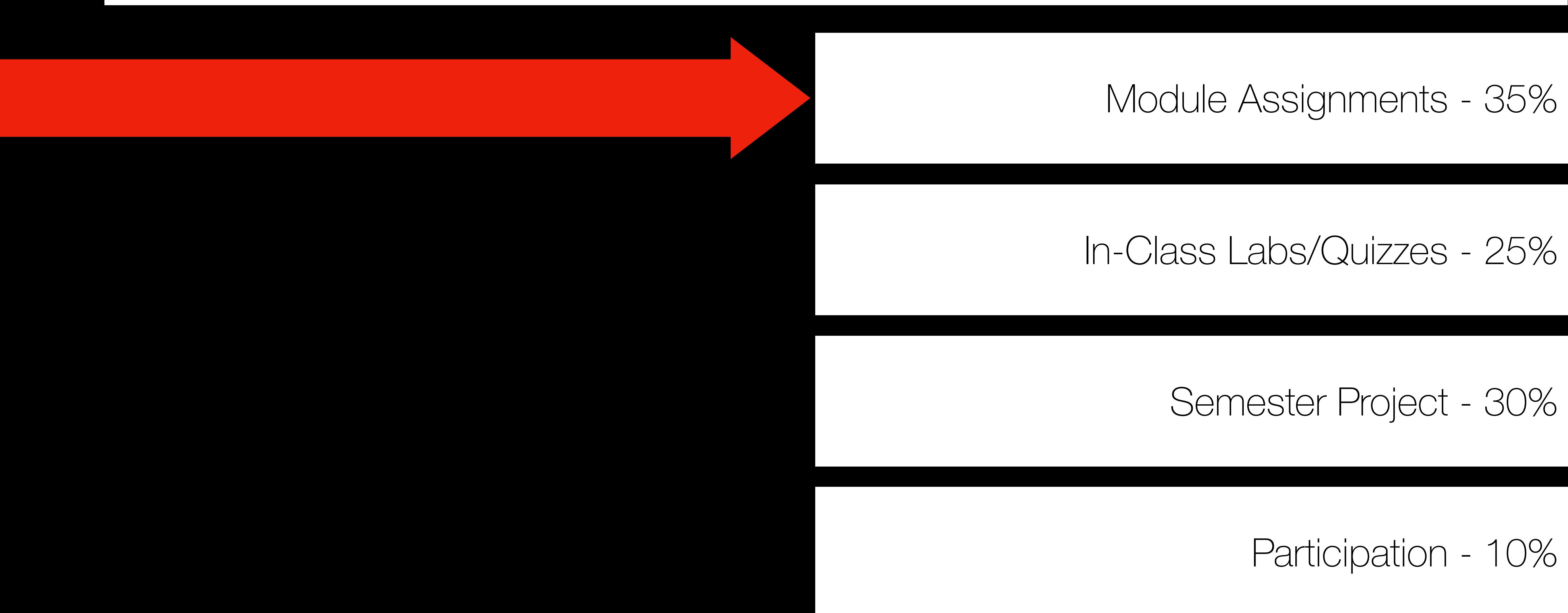
Participation - 10%

Late Policy – Amnesty Days

- Two days where you can submit any late Medium post for full credit
- October 24 – Amnesty Day 1
 - Anything due before this day
- December 2 – Amnesty Day 2
 - Anything due after Amnesty Day 1



Four Parts to Your Grade





Module Assignments (Homework)

Assignments: INST414-0103 X +

umd.instructure.com/courses/1361527/assignments

Relaunch to update :

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello UMIACS Object St...

Spring 2024

Home

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

Clickers

Course Reserves

Adobe Creative Cloud

Module Assignments 35% of Total

Pre-Class Skill Check - Most Frequent Tokens in Text
Available until Feb 9 at 11:59pm | Due Feb 2 at 11:59pm | -/100 pts

Foundation - Medium Signup
Available until Feb 9 at 11:59pm | Due Feb 9 at 11:59pm | -/100 pts

Foundation - GitHub Signup
Available until Feb 9 at 11:59pm | Due Feb 9 at 11:59pm | -/100 pts

Module 1 Assignment
Available until May 1 at 11:59pm | Due Feb 9 at 11:59pm | -/100 pts

Module 2 Assignment
Available until May 1 at 11:59pm | Due Feb 23 at 11:59pm | -/100 pts

Module 3 Assignment
Available until May 1 at 11:59pm | Due Mar 8 at 11:59pm | -/100 pts

Module 4 Assignment
Available until May 1 at 11:59pm | Due Mar 29 at 11:59pm | -/100 pts

Module 6 Assignment
Available until May 1 at 11:59pm | Due Apr 26 at 11:59pm | -/100 pts

Extra Credit - Sharing Resources

Semester Project 30% of Total

Reset Student Leave Student View

You are currently logged into Student View

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

Module 1 Assignment

umd.instructure.com/courses/1361527/assignments/6665528

Relaunch to update

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello UMIACS Object St...

Spring 2024

Module 1 Assignment

Due: Fri Feb 9, 2024 11:59pm

100 Points Possible

Attempt 1 In Progress
NEXT UP: Submit Assignment

Add Comment

Account

Dashboard

Courses

Calendar

Inbox

Portfolium

History

CourseExp

Help

EMT

Previous

Submit Assignment

Next

Unlimited Attempts Allowed

Available until May 1, 2024 11:59pm

Details

Write a 1,000-word Medium post describing an exploratory analysis of a question for which 1) you have some expertise, and 2) you can describe who needs the answer to this question, and 3) you can explain how this answer enables a particular decision. This assignment will serve as a baseline for your work in this course, and at the end of the semester, you can review it so evaluate whether you would do things differently.

Your post should include the following:

- Describe a question you think can be answered from data, what specific stakeholder is asking this question, and what decision(s) the answer to this question will inform.
- Describe the data that could answer this question, what fields it contains, and why it is relevant to your question.
- Explain how you collected some subset of this data (e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from data archive).
- Perform an exploratory data analysis that helps answer the question you posed.
- Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.
- Include figures or tables summarizing your findings.
- Discuss the limitations of your analysis. What's missing? How might it be biased?
- Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.

When you have written your post, publish it via Medium, **add your post it to the class publication via Medium**, and submit the URL to it via ELMS.

Note: The publication name for the class is "[Data Science Techniques \[INST414\]](#)"

Tag your story as "inst414spr24a01" - Articles with this tag will automatically be added to the appropriate assignment tab after I accept on the class

6d You are currently logged into Student View

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

Reset Student Leave Student View

Module 1 Assignment

umd.instructure.com/courses/1361527/assignments/6665528

Relaunch to update

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello UMIACS Object St...

Module 1 Assignment

Spring 2024 Due: Fri Feb 9, 2024 11:59pm

100 Points Possible

Attempt 1 In Progress NEXT UP: Submit Assignment

Add Comment

Account

Dashboard

Courses

Calendar

Inbox

Portfolium

History

CourseExp

Help

EMT

Previous Next

Unlimited Attempts Allowed
Available until May 1, 2024 11:59pm

Details

Write a 1,000-word Medium post describing an exploratory analysis of a question for which 1) you have some expertise, and 2) you can describe who needs the answer to this question, and 3) you can explain how this answer enables a particular decision. This assignment will serve as a baseline for your work in this course, and at the end of the semester, you can review it so evaluate whether you would do things differently.

Your post should include the following:

- Describe a question you think can be answered from data, what specific stakeholder is asking this question, and what decision(s) the answer to this question will inform.
- Describe the data that could answer this question, what fields it contains, and why it is relevant to your question.
- Explain how you collected some subset of this data (e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from data archive).
- Perform an exploratory data analysis that helps answer the question you posed.
- Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.
- Include figures or tables summarizing your findings.
- Discuss the limitations of your analysis. What's missing? How might it be biased?
- Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.

When you have written your post, publish it via Medium, **add your post it to the class publication via Medium**, and submit the URL to it via ELMS.

Note: The publication name for the class is "[Data Science Techniques \[INST414\]](#)"

Tag your story as "inst414spr24a01" - Articles with this tag will automatically be added to the appropriate assignment tab after I accept on the class

You are currently logged into Student View

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

Reset Student Leave Student View

Module 1 Assignment

umd.instructure.com/courses/1361527/assignments/6665528

Relaunch to update

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello UMIACS Object St...

Spring 2024

Module 1 Assignment

Due: Fri Feb 9, 2024 11:59pm

100 Points Possible

Attempt 1 In Progress
NEXT UP: Submit Assignment

Add Comment

Account

Dashboard

Courses

Calendar

Inbox

Portfolium

History

CourseExp

Help

EMT

Previous

Submit Assignment

Next

Unlimited Attempts Allowed

Available until May 1, 2024 11:59pm

Details

Write a 1,000-word Medium post describing an exploratory analysis of a question for which 1) you have some expertise, and 2) you can describe who needs the answer to this question, and 3) you can explain how this answer enables a particular decision. This assignment will serve as a baseline for your work in this course, and at the end of the semester, you can review it so evaluate whether you would do things differently.

Your post should include the following:

- Describe a question you think can be answered from data, what specific stakeholder is asking this question, and what decision(s) the answer to this question will inform.
- Describe the data that could answer this question, what fields it contains, and why it is relevant to your question.
- Explain how you collected some subset of this data (e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from data archive).
- Perform an exploratory data analysis that helps answer the question you posed.
- Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.
- Include figures or tables summarizing your findings.
- Discuss the limitations of your analysis. What's missing? How might it be biased?
- Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.

When you have written your post, publish it via Medium, **add your post it to the class publication via Medium**, and submit the URL to it via ELMS.

Note: The publication name for the class is "[Data Science Techniques \[INST414\]](#)"

Tag your story as "inst414spr24a01" - Articles with this tag will automatically be added to the appropriate assignment tab after I accept on the class

6d You are currently logged into Student View

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

Reset Student Leave Student View

INST414 – Medium Post Grading Rubric, Assignment 1

Name _____ Topic _____

#	Item Name	Max	Earned
1	Identify a non-obvious insight you want to extract from your data, describing what decision(s) this insight might inform.	10	
2	Describe the data that could answer this question, where it lives, and why it's relevant.	10	
3	Explain how you collected the data, e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from a data archive.	10	
4	Perform some exploratory data analysis that helps inform the decision you posed in your introduction.	10	
5	Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.	10	
6	Include figures or tables summarizing your findings.	10	
7	Conclude with a discussion of the limitations of your analysis. What's missing? How might it be biased?	10	
8	Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.	10	
9	Excellence	10	
10	Length	10	
	TOTAL	100	

Comments:

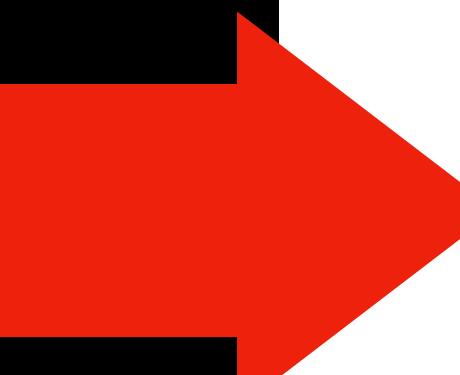
All assignments have an associated rubric

INST414 – Medium Post Grading Rubric, Assignment 1

Name _____ Topic _____

#	Item Name	Max	Earned
1	Identify a non-obvious insight you want to extract from your data, describing what decision(s) this insight might inform.	10	
2	Describe the data that could answer this question, where it lives, and why it's relevant.	10	
3	Explain how you collected the data, e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from a data archive.	10	
4	Perform some exploratory data analysis that helps inform the decision you posed in your introduction.	10	
5	Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.	10	
6	Include figures or tables summarizing your findings.	10	
7	Conclude with a discussion of the limitations of your analysis. What's missing? How might it be biased?	10	
8	Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.	10	
9	Excellence	10	
10	Length	10	
	TOTAL	100	

Comments:



GitHub Update

github.com

Jordan's Style https://inf.ethz.ch/... tome - Synology... Twitter Google Cl... TMRC Priority - Google... GO GDrive INCAS - Chat Other Bookmarks

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

cbuntain

Top Repositories New

Find a repository...

phomemes/phomemes.github.io
crisisfacts/crisisfacts.github.io
incas-upscale/response_rates
aidancrowl-INCAS/ta2-model-server
trecis/trecis.github.io
incas-upscale/upscale
crisisfacts/utilities
cbuntain/umd.inst414
cbuntain/TwitterStreamFilter
cbuntain/multimodal.consistency
SMAPPNYU/ss1-infeco
SMAPPNYU/youtube-data-api
cbuntain/CrossPlatformAnalytics
gmuric/INCAS
cbuntain/BurstyTwitterStreams
cbuntain/njit.is688
Iceberg-
ng/Moral_Foundation_FrameAxis
Iceberg-
ng/bbannotator
Iceberg-
ng/geolocator_domain
cbuntain/ta2-model-server

Join GitHub Global Campus!

Prepare for a career in tech by joining GitHub Global Campus. Global Campus will help you get the practical industry knowledge you need by giving you access to industry tools, events, learning resources and a growing student community.

Follow your Expert
Breaking into tech: internship edition with Helen Huang
Level up your code with TwilioQuest
Learning by teaching for your community - Cassidy Williams
Science & Technology
Talk Shows & Podcasts
Special Events

Popular offers you have not claimed:
Curated Experiences with popular offers:
Virtual event kit

January 22, 2021 Level up your code with TwilioQuest Arieli Kanter
February 1, 2021 GitHub Campus Experts applications are open Juan Pablo Flores Cortés
Artificial Intelligence 6 assignments
Lists and Loops Due by May 1, 2021, 12:00 PST
Week Five: Functions Due by Aug 16, 2021, 14:00 PST

Claim more offers

Connect your local Expert Visit their profile
View projects at our gallery Visit the Student Gallery
Learn more about an event Click on an event
Watch a Campus TV episode Visit GitHub Education on Twitch
Claimed a Student Pack offer See popular offers

Join Global Campus

Following For you Beta

edsu starred umd-mith/webvtt-player · 1h

umd-mith/webvtt-player

A React audio player & transcription viewer.

JavaScript Star 28 Updated Jan 4

Start coding instantly with GitHub Codespaces

Spin up fully configured dev environments on powerful VMs that start in seconds. Get up to 60 hours a month of free time.

Get started

Trending Repositories

brycedennan / imaginAlry

AI imagined images. Pythonic generation of stable diffusion images.
4.5k Python

trekhleb / javascript-algorithms

Algorithms and data structures implemented in JavaScript with explanations and links to further readings
160k JavaScript

LazyVim / LazyVim

Neovim config for the lazy

We'll review GitHub later

Why do I give these kinds of homework assignments?

Alternatively, why are you evaluated by your ability to communicate your analysis thoughtfully and concisely?

Medium

[HOME](#) [POPULAR](#) [MOMENTUM](#) [CORONAVIRUS](#) [ONEZERO](#) [ELEMENTAL](#) [GEN](#) [ZORA](#) [FORGE](#) [HUMAN PARTS](#) [MARKER](#)



How Medium Works With Writers

We're investing heavily in editorial and the Partner Program. Here's why.

Siobhan O'Connor in 3 min read
Oct 3, 2018 • 4 min read



Spark performance tuning from the trenches

[Yann Moisan in Teads Engineer](#)
May 29, 2018 · 9 min read

You Will Destroy Yourself Financially If You Save

Tim Denning in Mind Ca
May 13 · 6 min read ★

Tips and tricks for Medium writers

Make the best out of Medium as a writ...

Medium in 3 min read
Oct 2, 2017 · 5 min read

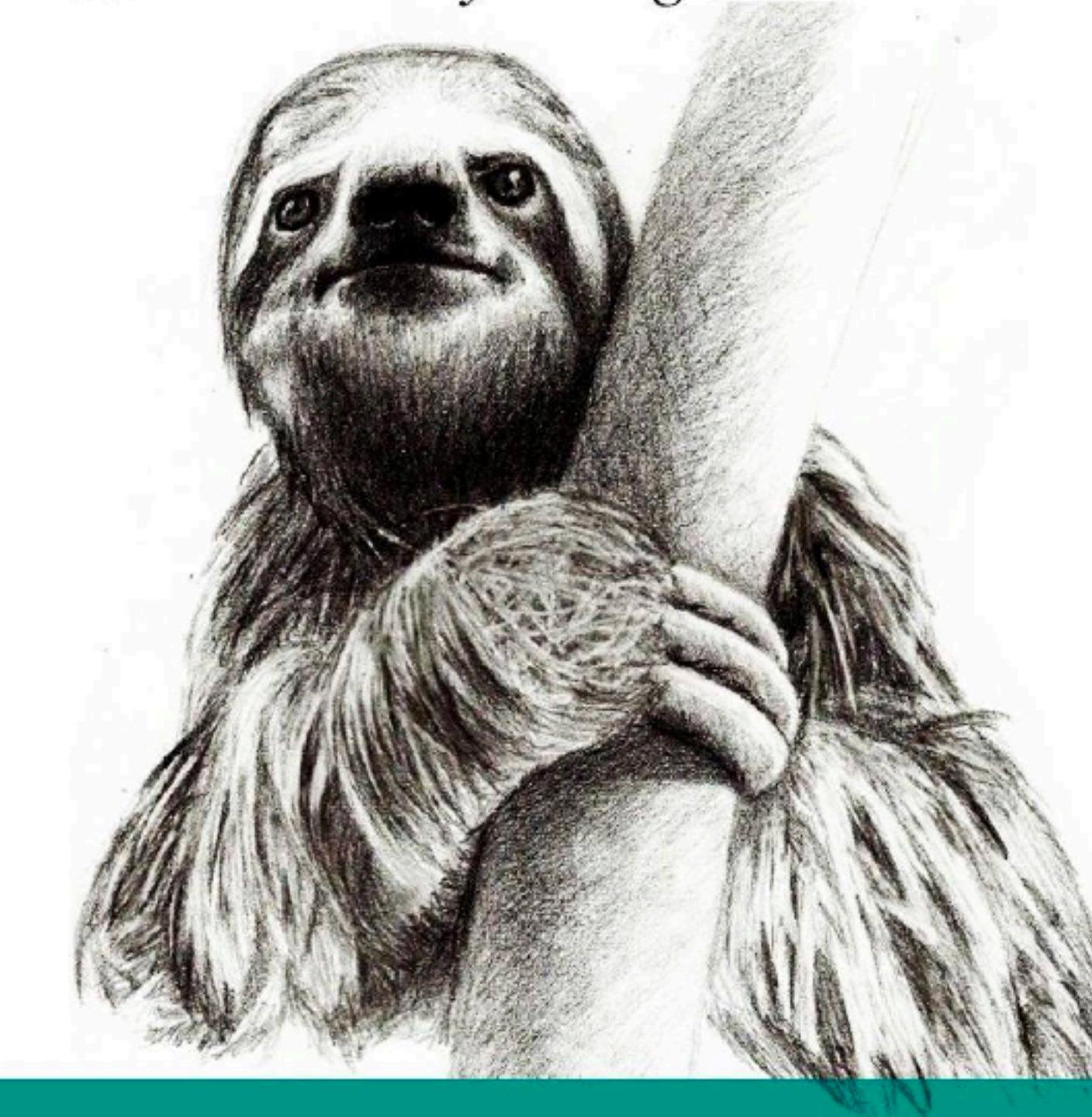
Experience suggests the public-facing nature of Medium leads to higher quality work

MOMENTUM

Popular on Medium

If a student is unable or does not want to use the Medium platform,
email me before Sept 5

Cutting corners to meet arbitrary management deadlines



Essential

Copying and Pasting from Stack Overflow

O'REILLY®

The Practical Developer
@ThePracticalDev

ChatGPT: Optimizing Language X +

openai.com/blog/chatgpt/ Update :

Introducing ChatGPT research release Try Learn more >

OpenAI API RESEARCH BLOG ABOUT

Syllabus for INST414-0102: Da X +

umd.instructure.com/courses/1353667/assignments/syllabus Update :

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello CHI24 SMDRM Dri...



Fall 2023

Home

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

Clickers

Course Reserves

Adobe Creative Cloud

Dashboard

Courses

Calendar

Inbox

Portfolium

History

CourseExp

Help

Use of AI and ChatGPT

In this course, you are welcome to use artificial intelligence (AI)-powered programs such as ChatGPT or Bing AI to help you create outlines or first drafts of your work. These tools can be a great starting place, and I encourage you to spend time editing your final draft into something you are proud of. It is your responsibility to verify that any information you get in these drafts is accurate and represents your own point of view well. If you have any questions about this policy and what constitutes acceptable or unacceptable use of AI-based tools, please do not hesitate to ask me -- I would be happy to talk with you about this!

Participation Policy

While the class is offered online and synchronously, interaction within and among the students is still an essential part of the learning experience. Such points include responses to discussion questions the instructors posts on ELMS, asking questions of fellow students' presentations, and consistent engagement with the instructors. These engagements will count for the class-participation portion of the final grade.

Late Policy

This class does not institute a late policy for module-level assignments. Module assignments will be made available throughout the semester with suggested deadlines, and students are encouraged to submit them by that deadline. No late penalty will be assessed against that deadline, however.

The motivation for this "lax" policy comes from pedagogical evidence suggesting late penalties are not helpful in getting students to submit work. Rather, module assignments will shared on Medium, and students are encouraged to assess the rate of submissions from their fellow students.

Sign In Compliance X INST414: Data Science Technic +

← → C medium.com/inst414-data-science-tech

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello CHI24 SMDRM Dri...

Search icon, Share icon, Bell icon, Profile icon, Magnifying glass icon

INST414: Data Science Techniques

Explore applications of data science techniques to unstructured, real-world datasets

ASSIGNMENT 1 ASSIGNMENT 2 ASSIGNMENT 3 ASSIGNMENT 4 ASSIGNMENT 6 FINAL PROJECT REPORTS PRIOR SEMESTERS

Latest

Prince Okpoziakpo
May 17 • 16 min read

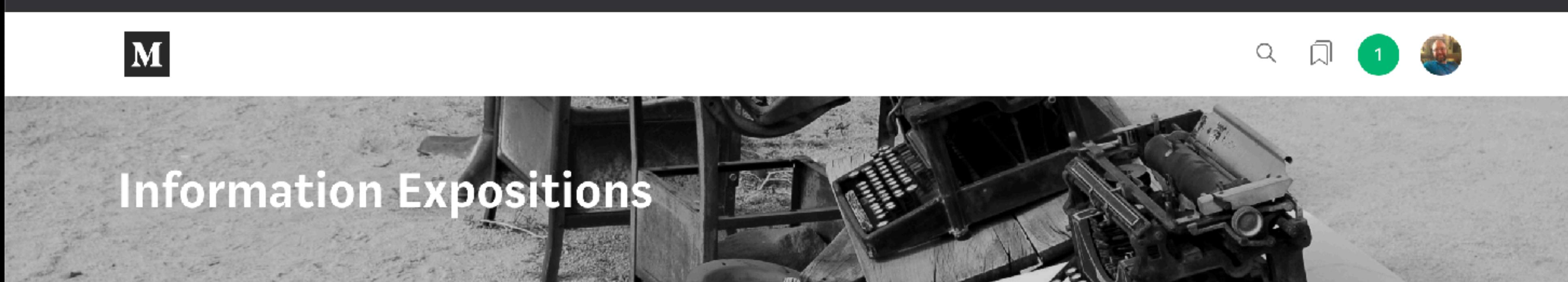
INST414: Data Science Techniques

This publication includes student-authored

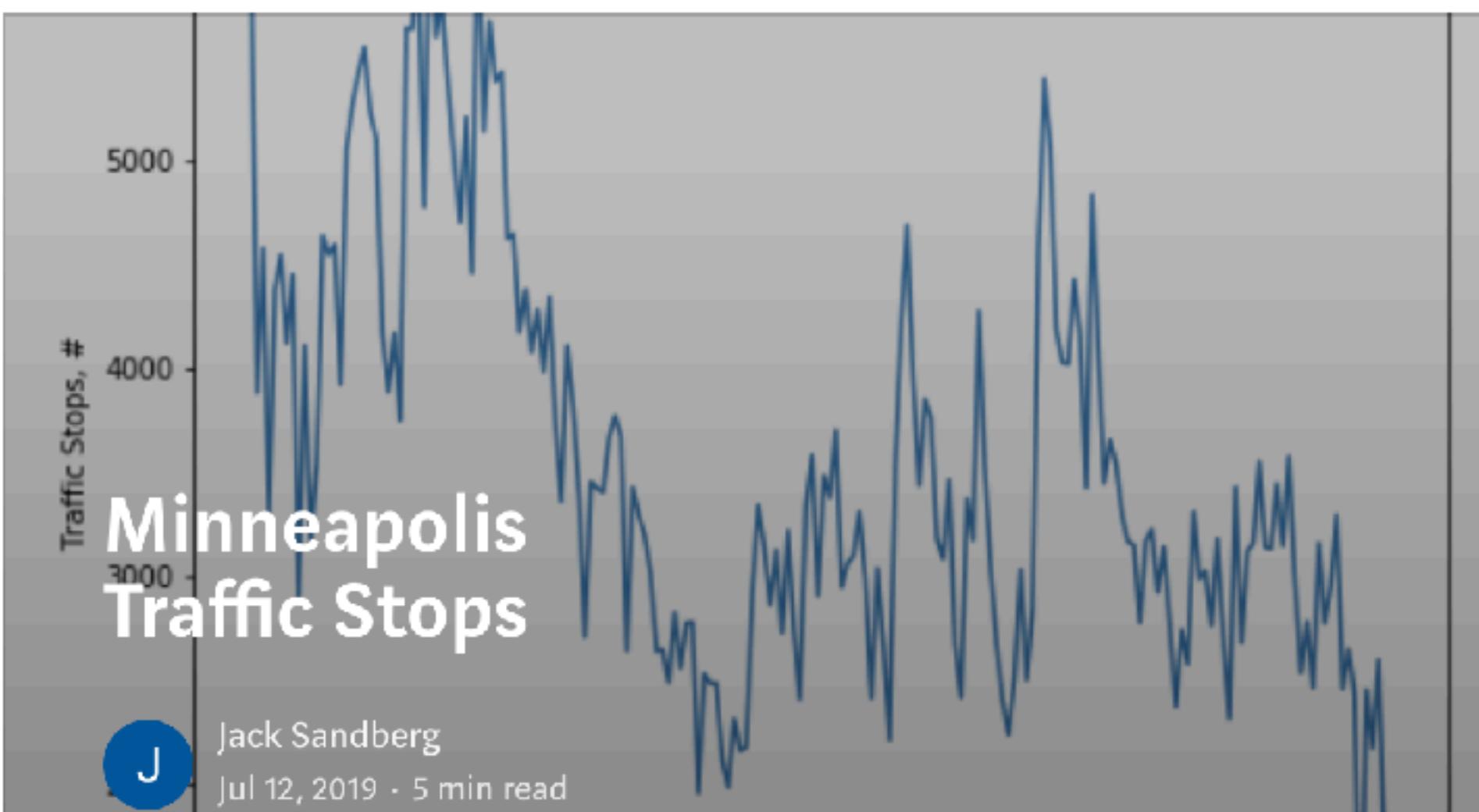
Finding Relationships and Similarities Within Books: The Foundation for A Book Recommendation Tool

Read more...

Can see examples here:
<https://medium.com/inst414-data-science-tech>



ASSIGNMENT 1 ASSIGNMENT 2 ASSIGNMENT 3 ASSIGNMENT 4 ASSIGNMENT 5 ASSIGNMENT 6 ASSIGNMENT



Or here:

<https://medium.com/information-expositions>

```
In [4]:  
In [71]:  
Out[71]:  
  
In [5]:  
Out[5]:
```

Tn [71]

Out[71]

1

20

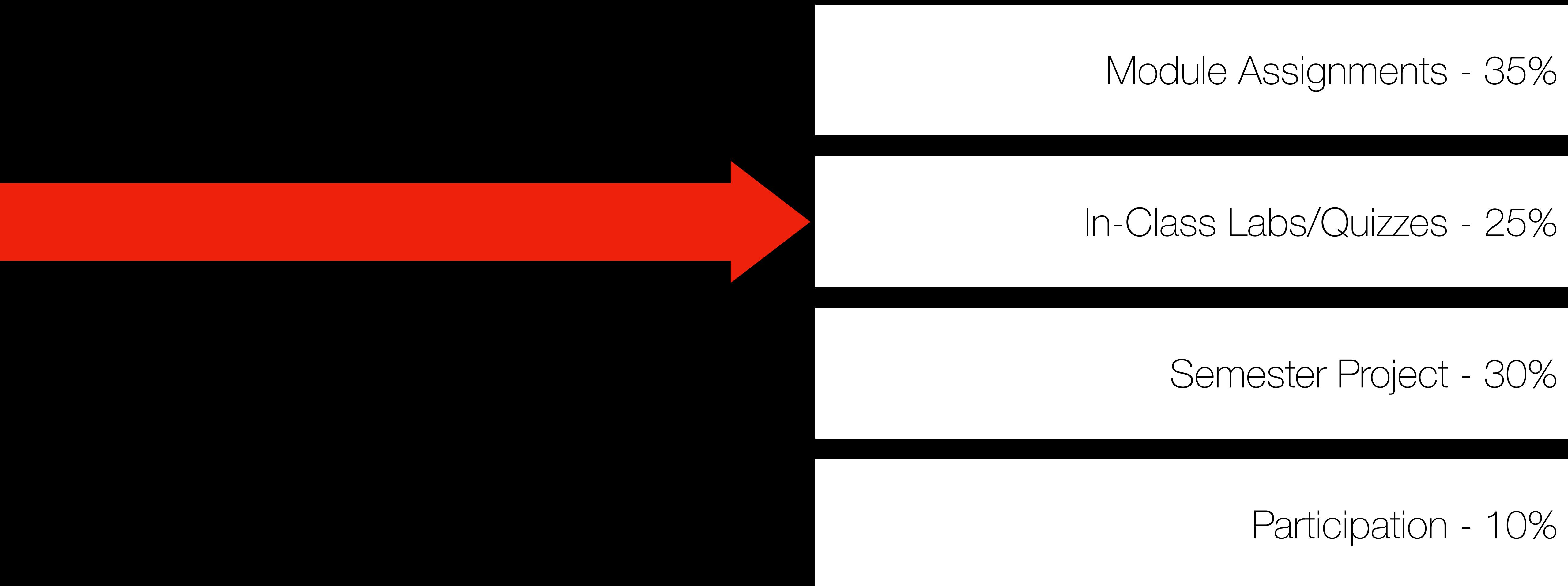
1

0	5318711	HN178009	02/15/2007 01:00:00 PM	0240XX W BLOOMINGDALE AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
---	---------	----------	------------------------------	---------------------------------	------	--------------------	---------------	--------

How To Become The

We'll go over Medium later

Four Parts to Your Grade



Exercises...

The screenshot shows a GitHub repository page for 'countain / umd.inst414' with a public notebook named '03-Graphs.ipynb'. The notebook title is 'Creating Networks from JSON Data'. It describes reading data from 'imdb_recent_movies.json' and constructing a graph of actors. The code includes imports for matplotlib, json, random, numpy, pandas, and networkx, and starts to define a graph 'g' by opening the JSON file and adding nodes based on actor IDs and names.

```
In [1]: %matplotlib inline

In [2]: import json
import random

import numpy as np
import pandas as pd
import networkx as nx

In [ ]:

In [12]: g = nx.Graph()

with open("../data/imdb_recent_movies.json", "r") as in_file:
    for line in in_file:

        this_movie = json.loads(line)

        for actor_id,actor_name in zip(this_movie['actor_ids'],this_movie['actor_names']):
            g.add_node(actor_id, name=actor_name)
```

B Graph Example

Download imdb_recent_movies.json file from Canvas

graph of actors based on costar roles

graph and randomly walk to tenth node

times, counting frequency of endpoints

Print top 10 most frequent endpoints

Quizzes...

Assignments: INST414-0102: +

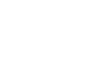
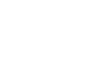
umd.instructure.com/courses/1353667/assignments

My Drive - Google... Inf.Eco.UMD - Drive AWS Login ELMS Personal | Trello CHI24 SMDRM Dri...

Fall 2023

Home Announcements  Assignments  Discussions Grades People Pages Files Syllabus Outcomes  Rubrics Quizzes Modules BigBlueButton Collaborations Chat Panopto Recordings New Analytics Clickers Course Reserves Adobe Creative Cloud Quiz Extensions Settings

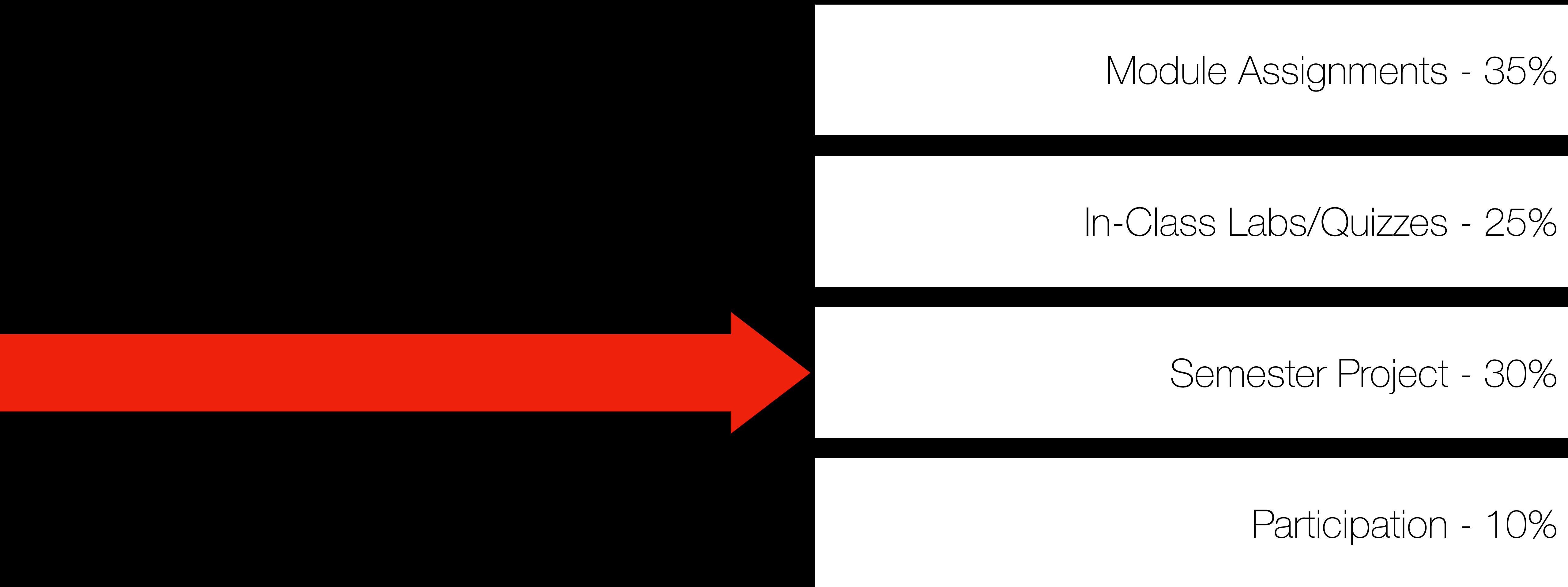
Exercises and Quizzes 25% of Total + :

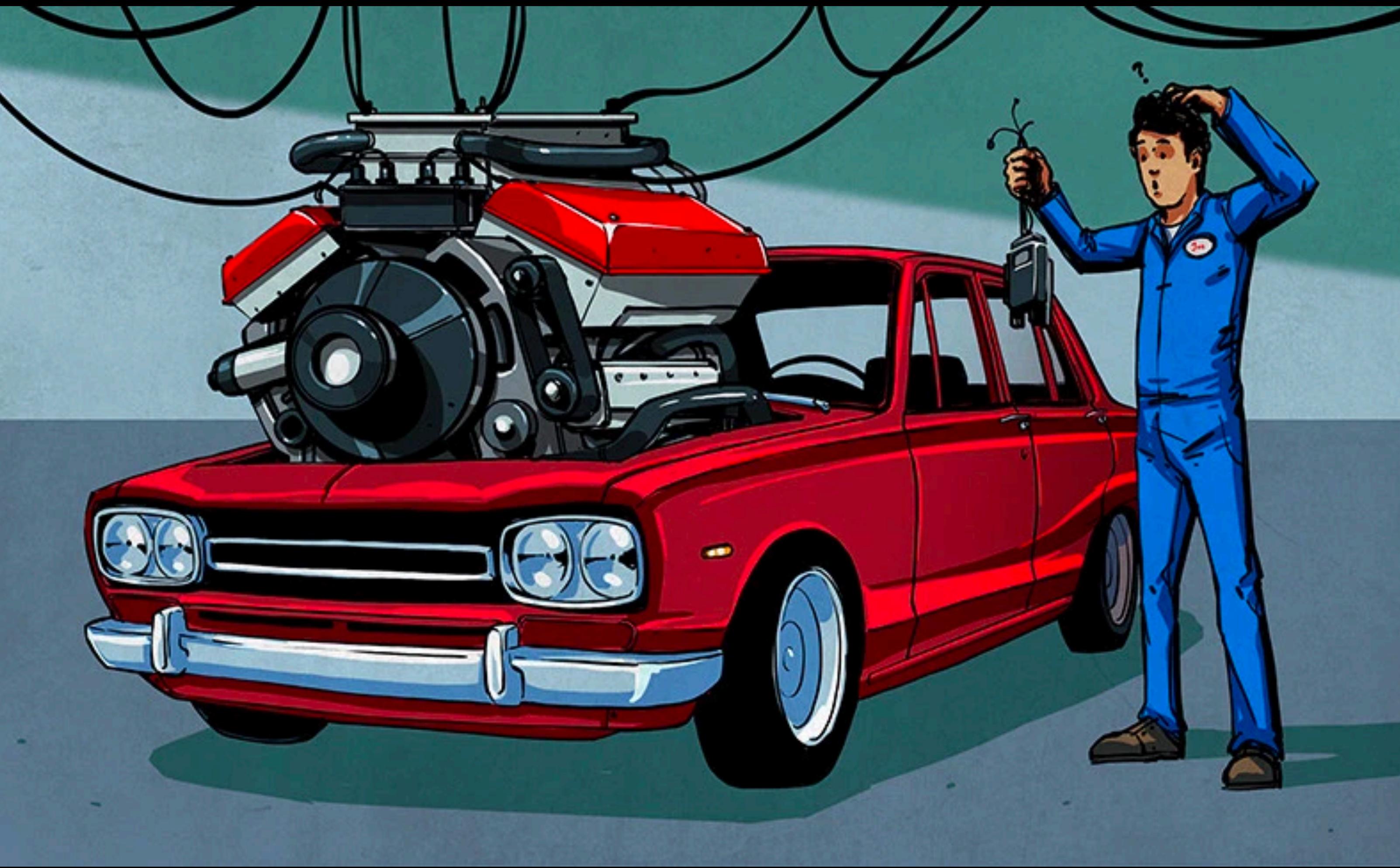
- Quiz, Week 14 Available until Nov 29 at 11:59pm | Due Nov 29 at 11:59pm | 6 pts  
- Quiz, Week 13 Available until Nov 20 at 11:59pm | Due Nov 20 at 11:59pm | 6 pts  
- Quiz, Week 12 Available until Nov 15 at 11:59pm | Due Nov 15 at 11:59pm | 8 pts  
- Quiz, Week 9 Available until Oct 25 at 11:59pm | Due Oct 25 at 11:59pm | 6 pts  
- Quiz, Week 8 Module 04 - Clustering and Unsupervised Learning Module | Available until Oct 18 at 11:59pm | Due Oct 18 at 11:59pm | 6 pts  
- Quiz, Week 7 Available until Oct 11 at 11:59pm | Due Oct 11 at 11:59pm | 6 pts  
- Quiz, Week 6 Available until Oct 4 at 11:59pm | Due Oct 4 at 11:59pm | 7 pts  
- Quiz, Week 5 Module 03 - Similarity, Dimensionality Reduction, and Cleaning Module | Available until Sep 27 at 11:59pm | Due Sep 27 at 11:59pm | 7 pts  
- Quiz, Week 4 Available until Sep 20 at 11:59pm | Due Sep 20 at 11:59pm | 7 pts  
- Quiz, Week 3 Available until Sep 13 at 11:59pm | Due Sep 13 at 11:59pm | 8 pts  
- Quiz, Week 2 Module 01 - Data Science and Motivations Module | Available until Sep 6 at 11:59pm | Due Sep 6 at 11:59pm | 10 pts  

In-class Exercise, Week 2

In-person during Discussion

Four Parts to Your Grade





Semester Project

Final Report on Medium

In-Class Presentation on your Expansion

Select a Module Assignment to Expand

Module 1
Assignment

Module 2
Assignment

Module 3
Assignment

Module 4
Assignment

Module 6
Assignment

The semester project is meant to give students an opportunity to go deeper on a particular subject and dataset that are of interest to the student

Engagement with and getting to know the data is one of the most important parts of data science

Need to do a good job your Module assignments

Four Parts to Your Grade



Module Assignments - 35%

In-Class Labs/Quizzes - 25%

Semester Project - 30%

Participation - 10%



DUDE, I DON'T EVEN KNOW YOU

BULLET TRAIN



I don't even know who you are.

Do I know who you are?



Spring 2025

[Home](#)[Assignments](#)[Discussions](#)[Grades](#)[People](#)[Pages](#)[Files](#)[Syllabus](#)[Quizzes](#)[Modules](#)[BigBlueButton](#)[Collaborations](#)[Chat](#)[Panopto Recordings](#)[Clickers](#)[Course Reserves](#)[Adobe Creative Cloud](#)[Top Hat](#)[Lucid \(Whiteboard\)](#)

Available until Mar 7 at 11:59pm | Due Mar 7 at 11:59pm | -/100 pts

In-class Exercise, Week 8

Available until Mar 21 at 11:59pm | Due Mar 21 at 11:59pm | -/100 pts

In-class Exercise, Week 9

Available until Mar 28 at 11:59pm | Due Mar 28 at 11:59pm | -/100 pts

In-class Exercise, Week 10

Available until Apr 4 at 11:59pm | Due Apr 4 at 11:59pm | -/100 pts

In-class Exercise, Week 11

Available until Apr 11 at 11:59pm | Due Apr 11 at 11:59pm | -/100 pts

In-class Exercise, Week 12

Available until Apr 18 at 11:59pm | Due Apr 18 at 11:59pm | -/100 pts

In-class Exercise, Week 13

Available until Apr 25 at 11:59pm | Due Apr 25 at 11:59pm | -/100 pts

▼ Participation

10% of Total

Participation - Introductions

Due Feb 10 at 11:59pm | -/100 pts

Does the professor know you?

Due May 11 at 11:59pm | -/100 pts

Does the TAs know you?

Due May 11 at 11:59pm | -/100 pts



Fall 2025



Account



Dashboard



Courses



Calendar



Inbox



Portfolio



History



Commons



CourseExp



Help



EMT



Home

Announcements



Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Outcomes



Rubrics

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

Clickers

Course Reserves

Adobe Creative Cloud

Quiz Extensions

Course Analytics

Ally Course Accessibility

Medium Peer Review

PublishedAssign ToEdit⋮

24/7 Canvas Chat Support

...or call 1-833-566-3347 (staff/faculty)

1-877-399-4090 (students)

Related Items

SpeedGrader

Canvas Guides ▾

Textbooks

Adopt Textbook

Peer reviews to give you experience reading other people's work

Write a peer review of one of your fellow student's Medium posts from this semester. This peer review serves to expose you to how others frame and communicate data science for their stakeholders and give you practice in thinking critically and considering how a reader may read your own content. You will reflect on what makes an impactful question/actionable insight, effective analysis, and high-quality technical writing. Your constructive feedback will help your classmates improve their communication skills and create a better project portfolio for recruiters.

Unlike the Medium posts, your review will not be public, but we will share its contents and authorship with the post's author.

You should submit a Word document containing this review, where each question below is answered in a separate section. Include the following sections:

- Selected Post. Include a link to and the title of the post you will be reviewing.
- Stakeholders and Questions: Explain whether you think the stakeholder and their question are clearly identified. Does the article clearly outline the decisions available to the stakeholder and how an answer will help them make this decision?
- Data and Collection: Explain what source of the data used, its form, biases? Does this description include the work? Is sufficiently relevant to the article's posed question?
- Quality of Analysis: Describe the extent to which you think the methods and techniques are explained clearly.

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

TRY CHATGPT ↗

November 30, 2022
13 minute read

Introducing ChatGPT research release [Try ↗](#) [Learn more >](#)

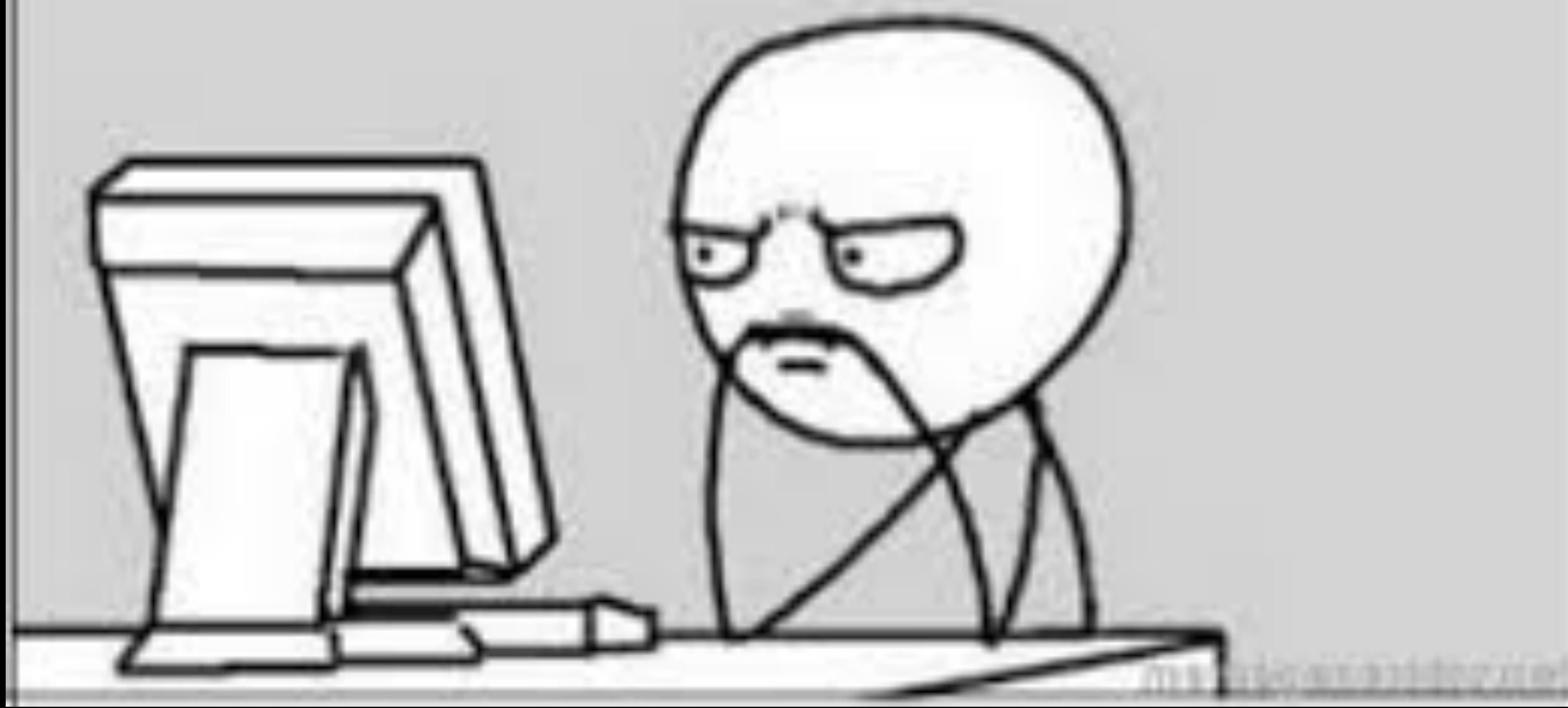
API RESEARCH BLOG ABOUT

TRY CHATGPT ↗

November 30, 2022
13 minute read

As AI tools get better, we need to get better reading and assessing their output

**ANXIOUSLY WAITING FOR YOUR NEXT
ASSIGNMENT!**



For Next Time

Pre-Class Skill Check - Most Frequent Tokens in Text

umd.instructure.com/courses/1361527/assignments/6665554?module_item_id=12579842

Published Edit

24/7 Canvas Chat Support
....or call 1-833-566-3347 (staff/faculty)
1-877-399-4090 (students)

Motivation

This course requires experience in programming, with a pre-requisite of at least one Python-based programming course. To ensure you have the experience necessary to succeed in this course, you should be able to complete the assignment below with minimal new learning (beyond familiarizing yourself with the NLTK library).

If you have trouble completing this skill-check, you will find this class overly difficult. This class assumes a certain level of expertise as we discuss data-science techniques. That is, this course is heavily focused on concepts in data science and how we apply them via programming; the class is not designed to teach you how to program. If you have difficulty here, you are encouraged to reach out to the professor for guidance as soon as possible.

Overview

Write a Python script to read a text file and output the top 10 most common words in the file and their counts. You should use the [Natural Language ToolKit \(NLTK\) library](#) to handle splitting words out of the text for you. NLTK provides a casual tokenizer that works well for "casual" text, such as social media posts (e.g., tweets). Documentation for this tokenizer function is available here: <https://www.nltk.org/api/nltk.tokenize.casual.html>

Requirements:

Your program should satisfy the following requirements.

- Use the TweetTokenizer tokenizer in NLTK (<https://www.nltk.org/api/nltk.tokenize.casual.html>)
- NLTK has multiple tokenizers (e.g., `word_tokenize`), but you should use the TweetTokenizer one, so you have consistent output.
- The input file (i.e., the file to be read by the program) should be named `input.txt`.
- For example, if you wrote a Python program called CommonTokenCounter.py, you would run it on a given text file by calling:
CommonTokenCounter.py input.txt

Related Items

SpeedGrader™

Download Submissions

0 out of 2 Submissions Graded

How to use UMD Canvas ▾

Textbooks

Adopt Textbook

Spring 2024

Home

Announcements

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Outcomes

Rubrics

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

New Analytics

Clickers

Course Reserves

Adobe Creative Cloud

Quiz Extensions

Settings

Account

Dashboard

Courses

Calendar

Inbox

Portfolium

History

Commons

CourseExp

Help

EMT

Settings

You should see this skill-check in Canvas



Foundation - GitHub Signup Foundation - Medium Signup

umd.instructure.com/courses/1340623/assignments/6190135

INST414 > Assignments

Spring 2023

Home

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

Clickers

Course Reserves

Adobe Creative Cloud

Course Policies

Foundation - Medium Signup

Due: Wed Feb 1, 2023 11:59pm

Attempt 1 **IN PROGRESS**
Next Up: Submit Assignment

100 Possible Points

Add Comment

Unlimited Attempts Allowed

Details

Sign up for a Medium account and send me your Medium username (the piece with the @-symbol, as highlighted below).

Medium - Where good ideas find their audience

Good afternoon

Gabriel Faucher in Towards Data Science

Python: Detecting Twitter Bots with Graphs and Machine Learning

Aug 7, 2020 · 9 min read

Nechu BM in Towards Data Science

How to build an encoder decoder translation model using LSTM...

Yong Cui in Towards Data Science

How to Calculate the Number of Parameters in Keras Models

Sep 29, 2020 · 5 min read

Libor Vanek in Towards Data Science

LATEST FROM FOLI

Cody Buntain @cbuntain

Write a story

Stories

Stats

Design your profile

Settings

Programming

Reading list

Publications

Submit Assignment

Next >

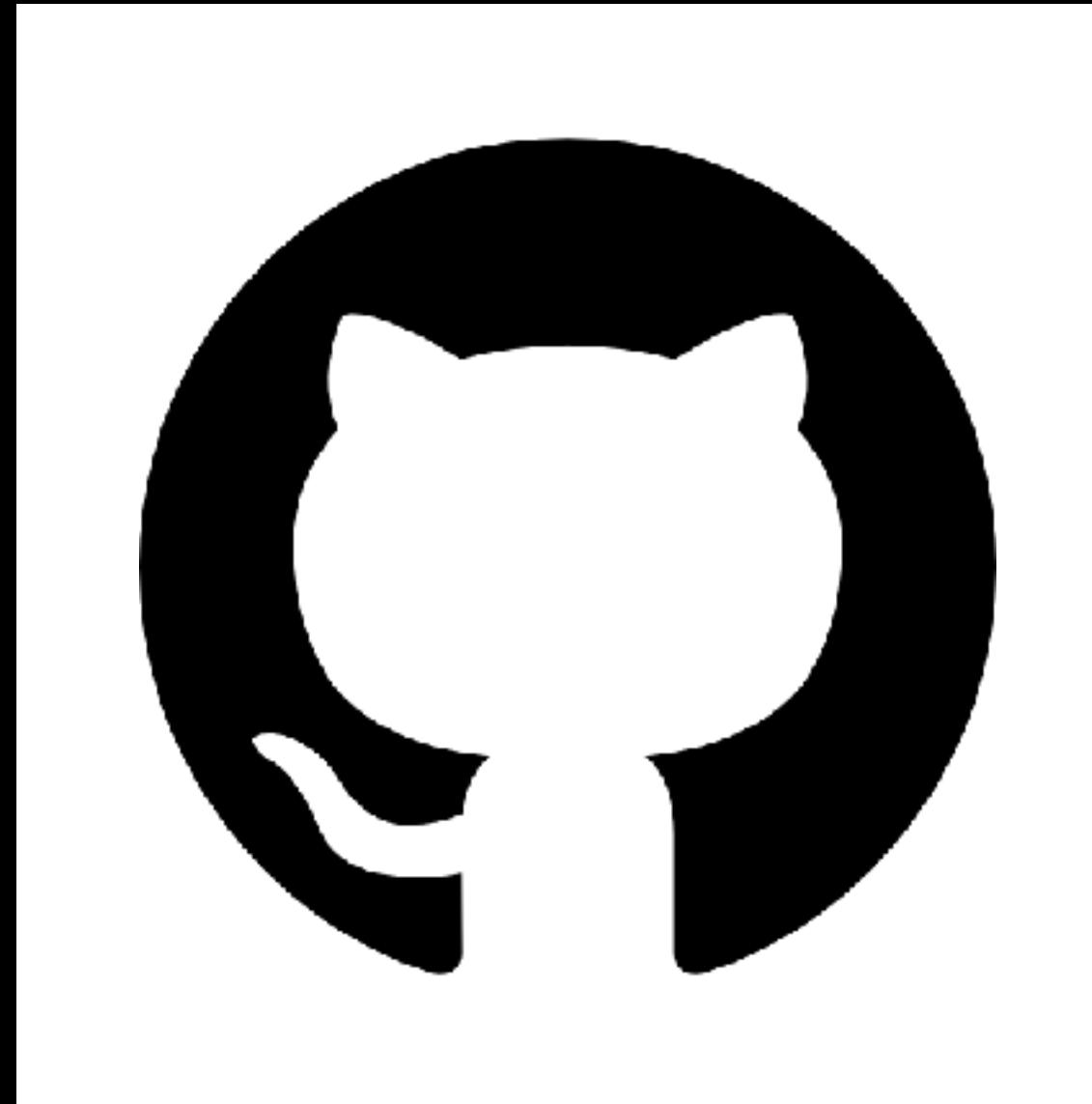
< Previous

You are currently logged into Student View

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

Reset Student

Leave Student View



Foundation - GitHub Signup

umd.instructure.com/courses/1340623/assignments/6301321

INST414 > Assignments

Spring 2023

Foundation - GitHub Signup

Due: Wed Feb 1, 2023 11:59pm

Attempt 1 IN PROGRESS

Next Up: Submit Assignment

100 Possible Points

Add Comment

Unlimited Attempts Allowed

Details

As your Medium posts require to share a link to your GitHub repository, you also need to have a GitHub account. Sign up for GitHub (it's free), and submit your account name here.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

Submit Assignment

Course Policies

Previous Next

You are currently logged into Student View

Reset Student Leave Student View

Reset the test student will clear all history for this student, allowing you to view the course as a brand new student.



Topic: Participation - Introductory

umd.instructure.com/courses/1340623/discussion_topics/4854786

INST414 > Discussions > Participation - Introductions

Spring 2023

Home

Assignments

Discussions

Grades

People

Pages

Files

Syllabus

Quizzes

Modules

BigBlueButton

Collaborations

Chat

Panopto Recordings

Clickers

Course Reserves

Adobe Creative Cloud

Account

Dashboard

Courses

Calendar

Inbox

Portfolium

History

CourseExp

Help

EMT

Course Policies

This is a graded discussion: 100 points possible due Jan 27

Participation - Introductions

Use this space to introduce yourself, answering at least the following:

What do you prefer to be called (include pronunciation if possible)?

What are your preferred pronouns (e.g., he/him, she/her, they/them)?

Search entries or author

Unread

Subscribe

Reply

Avani Badugu (she/her) 12:10pm

My name is Avani Badugu and I am a senior. I use she/her pronouns.

Reply

Noah Blackwell 2:55pm

Resetting the test student will clear all history for this student, allowing you to view the course as a brand new student.

Reset Student

Leave Student View



Account



Dashboard



Courses



Calendar



Inbox



Portfolium



History



Commons



CourseExp



Help



EMT



Course Policies



Spring 2025

In-class Exercise, Week 1

Published

Assign To

Edit

⋮

24/7 Canvas Chat Support

...or call 1-833-566-3347 (staff/faculty)

1-877-399-4090 (students)

Related Items

SpeedGrader

[How to use UMD Canvas ▾](#)

Textbooks

Adopt Textbook

Instructions

You will be told to focus on one of the following exercises. Read the scenario and consider the prompt. Then, describe an answer to the provided prompt. We will review answers in class. At the end of the exercise, upload a document describing your solution to the exercise.

For your answer to the prompt, provide the following:

- Identify one single main stakeholder you are assisting.
- What domain expertise do you think is necessary to answer this prompt?
- What data do you think would be ideal data to address this prompt?
- How might you collect this data?

Exercise 1. Airline Networks and Pandemic Prevention

Scenario: Consider the case of the Global Health Organization (GHO), a (fictional) organization charged with preventing the spread of disease. In this fictitious context, a disease is spreading internationally, and a major vector of this spread is travel through international airports (1,200 or so exist in the world today). The GHO has a treatment that can effectively cure an infection of the disease, but this treatment provides **no lasting immunity**—that is, if you are exposed to the disease again post-treatment, you can still get it and pass it on. GHO resources are limited such that they can only deploy this treatment at 20 airports at a given time.

Prompt: Describe a data-driven strategy for the GHO to identify 20 airports at which they should deploy their treatment.

Exercise 2. Recommending Videos

Scenario: The developers of the new short-video app, MeTok, have watched the current video. To ensure user satisfaction, the next video should somehow be enjoyable to the user. As the app developer, you have access to the user's video-watching history and other on-app meta-data, like friends and contacts.

Friday, an in-class exercise on comms



What questions do you have?

Prof. Cody Buntain | @codybuntain | cbuntain@umd.edu