

CMSC 473/673
Natural Language Processing
Fall 2021

Frank Ferraro – ferraro@umbc.edu

Instructor: Frank Ferraro

ITE 358/Remote

ferraro@umbc.edu

Monday: 2:15-3

Thursday: 1:00-1:30

by appointment

Natural language processing

Semantics

Vision & language processing

Learning with low-to-no supervision

TA: Ryan Barron

ITE 334/remote

ryanb4@umbc.edu

Monday, Wednesday 4:30-5:30

by appointment

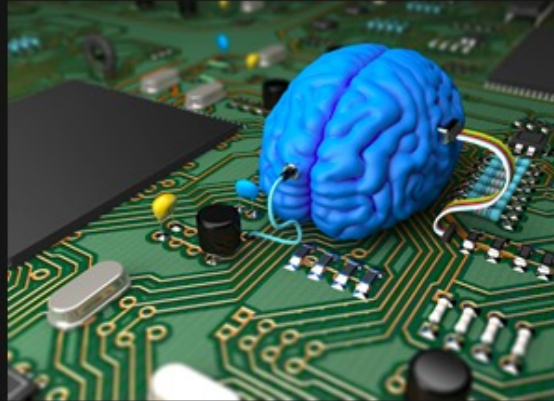
Q: What is NLP
(natural language
processing?)

~~Deep Learning~~

Natural Language Processing



What society thinks I do



What my friends think I do



What other computer scientists think I do



What mathematicians think I do



What I think I do

pytorch

```
from theano import *
```

What I actually do



Artificial Intelligence

News about Artificial Intelligence, including commentary from the New York Times.

Latest

Search

Aug. 26, 2018

Artificial Intelligence Is Now a Pentagon Priority. Will Silicon Valley Help?

The Defense Department, believing that A.I. research should be a national priority, has called on the White House to “inspire a whole of country effort.”

By CADE METZ

August 2018

Aug. 17, 2018

Alexa vs. Siri vs. Google: Which Can Carry on a Conversation Best?

Digital assistants from Amazon, Apple and Google can only have meager back-and-forth exchanges with us. Listen to how that tells us something about where they’re going in the future.

By KEITH COLLINS and CADE METZ



Aug. 16, 2018

Google Employees Protest Secret Work on Censored Search Engine for China

About 1,400 of the internet company’s employees have signed a letter demanding transparency, saying censored search results raise “urgent moral and ethical issues.”

By KATE CONGER and DAISUKE WAKABAYASHI



You’ve probably read about it!

Potential Applications

ASR (automatic speech recognition)

Machine translation

Natural language generation

Document labeling/classification

Document summarization

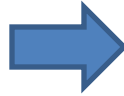
Corpus exploration

Relation/information extraction

Entity identification

Potential Applications

Q: What's an example?



ASR (automatic speech recognition)

Machine translation

Natural language generation

Document labeling/classification

Document summarization

Corpus exploration

Relation/information extraction

Entity identification



Automatic speech
recognition

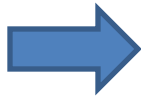
Potential Applications

ASR (automatic speech recognition)

Machine translation

Natural language generation

Q: What's an
example?



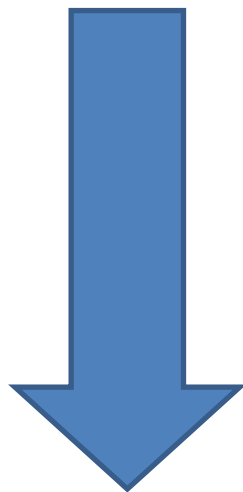
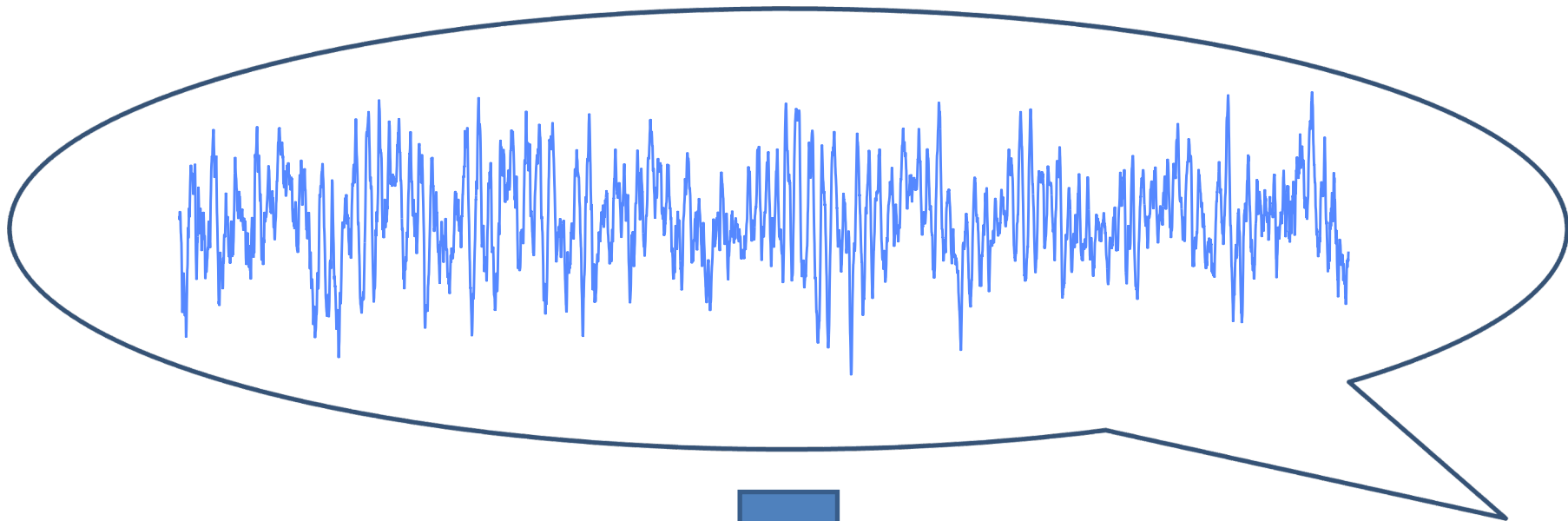
Document labeling/classification

Document summarization

Corpus exploration

Relation/information extraction

Entity identification



Document
classification





Google Translate

Document
classification



Machine
translation

Potential Applications

ASR (automatic speech recognition)

Machine translation

Natural language generation

Document labeling/classification

Q: What's an
example?

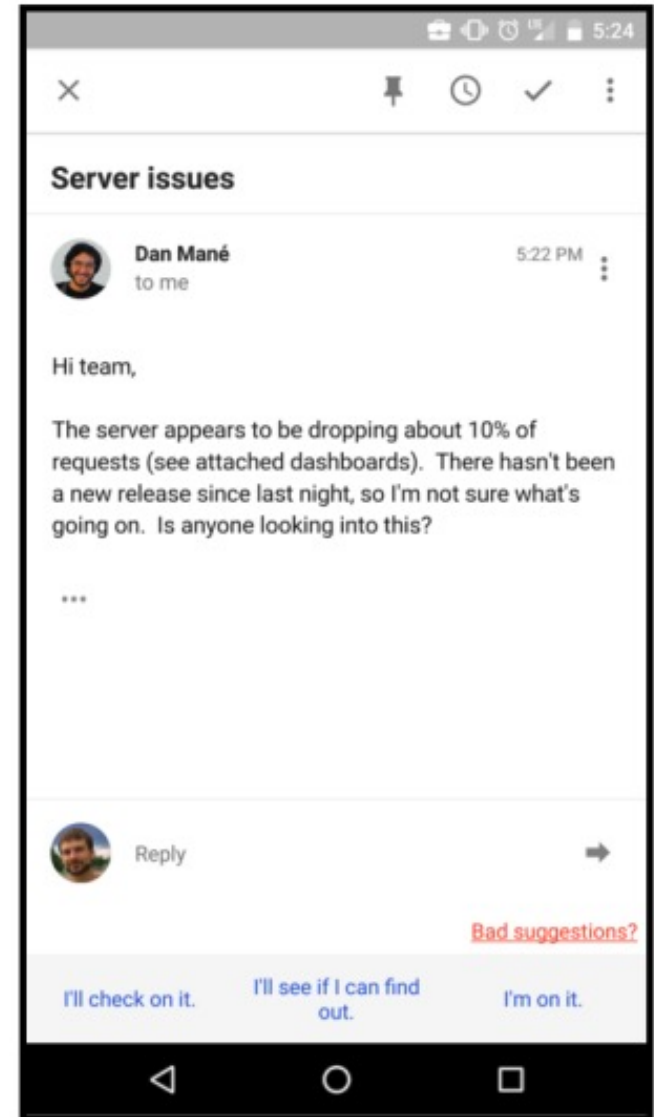
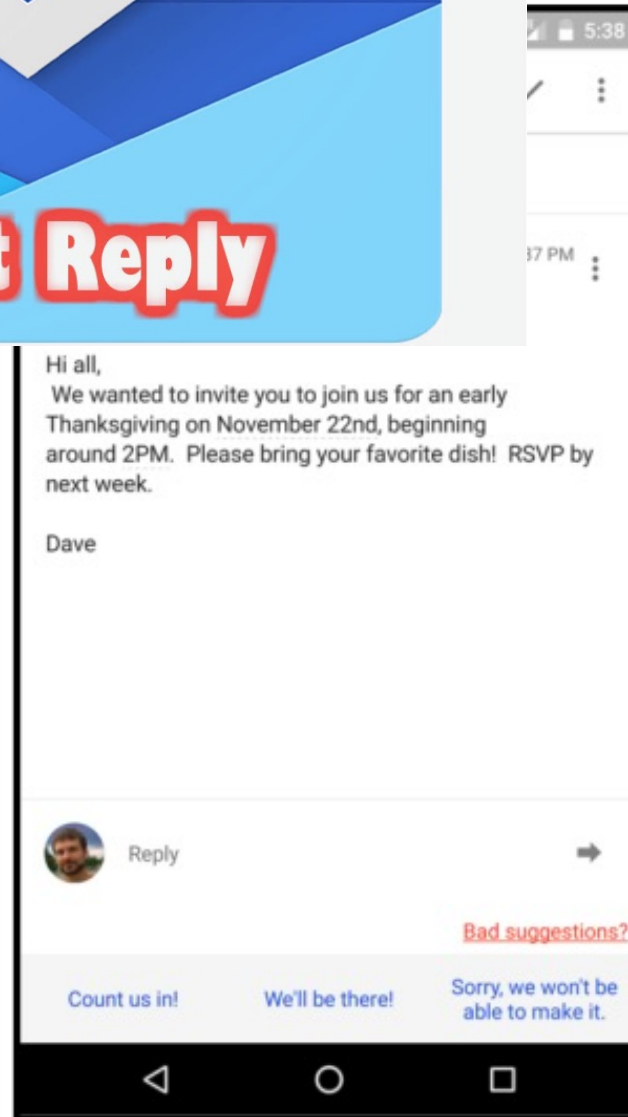


Document summarization

Corpus exploration

Relation/information extraction

Entity identification



Natural language
generation

A massive climate change study is canceled ... because of climate change



By **Doug Criss**, CNN

Updated 10:37 AM ET, Tue June 20, 2017

Story highlights

Arctic sea ice has traveled farther south than normal along Newfoundland's northeast coast

An icebreaker has been repeatedly diverted to take part in rescue operations

(CNN) — A \$17 million study of climate change in the Canadian Arctic has been nixed for now -- because of climate change.

A team of scientists from the University of Manitoba and four other schools were in the middle of the first leg of a four-year study of how climate change is affecting the areas around the Hudson Bay, the [university said in statement](#). The study, named BaySys, started last month, and the scientists were traveling on the Canadian Research Icebreaker CCGS Amundsen.

But because of warmer temperatures in the Arctic, hazardous sea ice is traveling farther south than usual. The Amundsen, which is part of the Canadian Coast Guard fleet, has been diverted several times because its ice-breaking capabilities have been needed to help out in rescue efforts along Newfoundland's northeast coast. All of the delays and concerns about safety forced



Top stories



Photo of dog carrying red bag goes viral



North Korea launches missile amidst US-South Korea tensions

Document
summarization

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Learn different ways that success and progress can be measured in NLP

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Learn different ways that success and progress can be measured in NLP

Relate to statistics, machine learning, and linguistics

Implement NLP programs

Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Learn different ways that success and progress can be measured in NLP

Relate to statistics, machine learning, and linguistics

Implement NLP programs

Read and analyze research papers

Practice your (written) communication skills

Course Topics

- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maxent/log-linear models, and distributed/vector-valued representations)

Course Topics

- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maxent/log-linear models, and distributed/vector-valued representations)
- sequences of latent variables (e.g., hidden Markov models, some basic machine translation alignment)

Course Topics

- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maxent/log-linear models, and distributed/vector-valued representations)
- sequences of latent variables (e.g., hidden Markov models, some basic machine translation alignment)
- trees and graphs, as applied to syntax and semantics

Course Topics

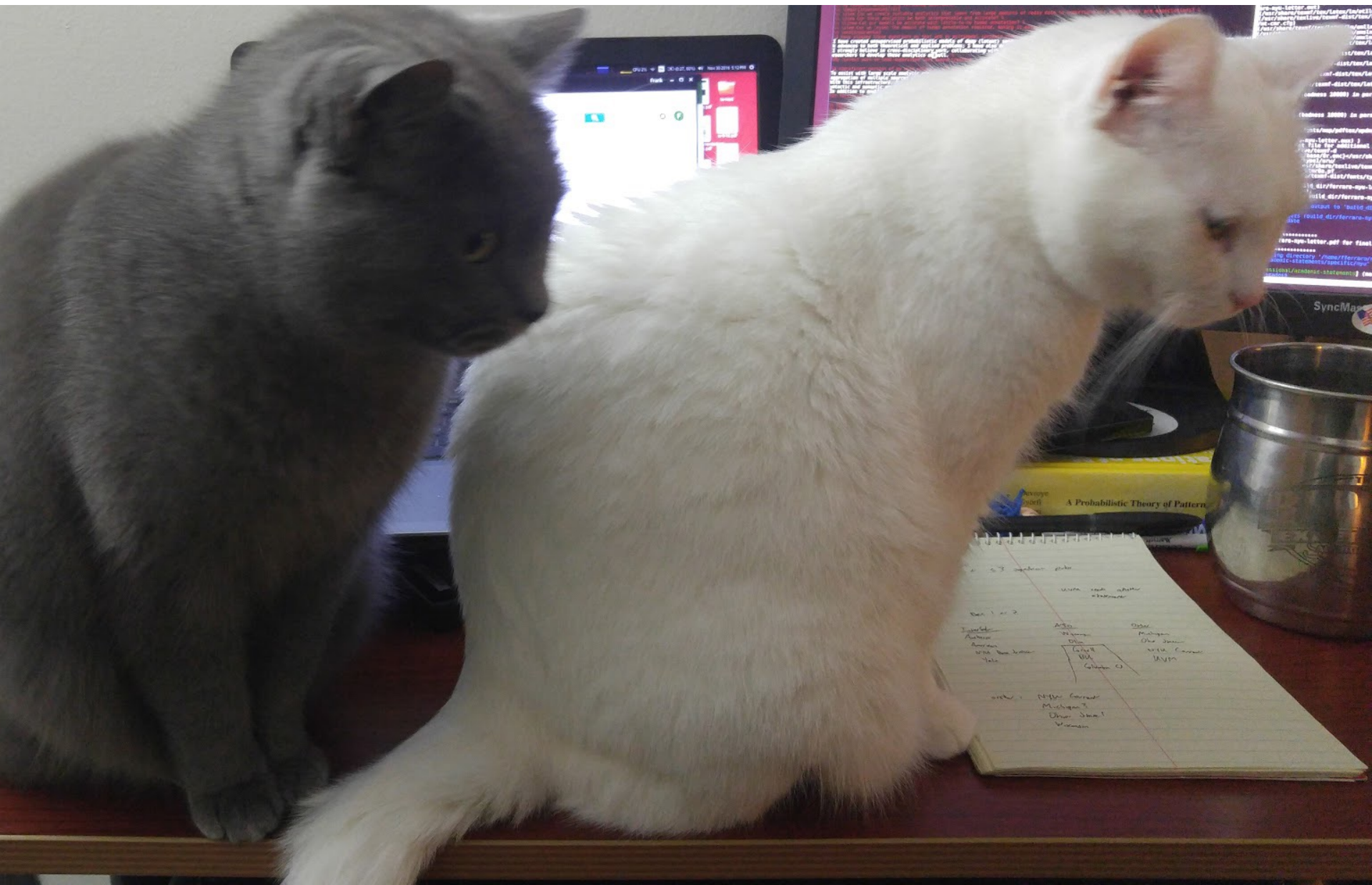
- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maxent/log-linear models, and distributed/vector-valued representations)
- sequences of latent variables (e.g., hidden Markov models, some basic machine translation alignment)
- trees and graphs, as applied to syntax and semantics
- some discourse-related applications (coreference resolution, textual entailment), and

Course Topics

- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maxent/log-linear models, and distributed/vector-valued representations)
- sequences of latent variables (e.g., hidden Markov models, some basic machine translation alignment)
- trees and graphs, as applied to syntax and semantics
- some discourse-related applications (coreference resolution, textual entailment), and
- special and current topics (e.g., fairness and ethics in NLP).

Course Topics

- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maxent/log-linear models, and distributed/vector-valued representations)
- sequences of latent variables (e.g., hidden Markov models, some basic machine translation alignment)
- trees and graphs, as applied to syntax and semantics
- some discourse-related applications (coreference resolution, textual entailment), and
- special and current topics (e.g., fairness and ethics in NLP).
- modern, neural approaches to NLP, such as recurrent neural networks and transformers (e.g., BERT or GPT-2).



Administrivia

Web Presence

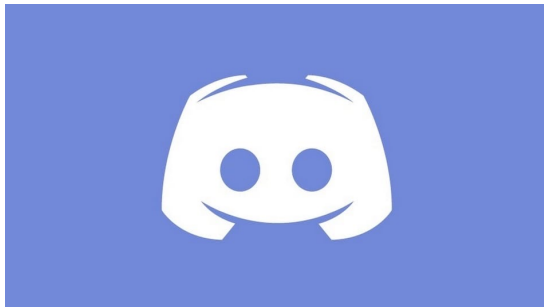
WWW

Schedule, slides, assignments,
readings, materials, syllabus here

<https://www.csee.umbc.edu/courses/undergraduate/473/f21>



<https://discord.gg/REUT7xyW> (or check email, or email me)



Course announcements,
Q&A, discussion board here

Expectations about Discord Server

For *you* to make communication and **asking** and **answering** questions *easy*

- Chat can be nice...
- ... but it can also be a distraction
- No “always on” view

Please Read the Syllabus (On the Website)

<https://www.csee.umbc.edu/courses/undergraduate/473/f21/content/materials/syllabus.pdf>



CMSC 473/673 — Introduction to Natural Language Processing
Fall 2021 (3 credits)
Version 0.9 — August 30th, 2021

Contents

1 Course Information	1
1.1 Masking Requirements: tl;dr	1
1.2 Meetings and Instructors	2
1.3 Texts, Readings, and Discussion	2
1.4 Topics	3
1.5 Goals	3
2 Coursework and Evaluation	3
3 Dates and Deadlines	4
3.1 Important Dates	4
3.2 Extensions and Late Policy	4
4 Academic Honesty	4
5 Students with Accommodations	5
6 Inclusion	6
7 Covid-19 Specifics	7
7.1 Attendance/In-Person Expectations	7
7.2 Masking	7
8 Version Changes	7

1 Course Information

Please see [this Google doc](#) for UMBC Policies and Resources during COVID-19. This Google doc contains important information regarding equity and inclusion.

1.1 Masking Requirements: tl;dr

Masks fully covering your mouth and nose must be worn at all times in the classrooms or (non-remote) office hours. As a result, please do not eat in the classroom or office hours. If you need to take a drink of water, take it, but then replace your mask when done taking that drink. Reminders will be given; please adjust your mask if needed. Repeated non-compliance or disruptions will

Do wear a mask that



- Covers your nose and mouth and secure it under your chin.
- Fits snugly against the sides of your face.

At all times in class or
office hours!

How NOT to wear a mask



Around your neck



On your forehead



Under your nose



Only on your nose



On your chin



Dangling from one ear



On your arm

Academic Integrity

Super important: I take it **very** seriously

You are responsible for your (& your group's) own work: if in doubt, ask!

Penalties could include 0 on the assignment, course failure, suspension, or expulsion (not exhaustive)

(Max. Lower Bounds for) Final Grades

473

$\geq$	Letter
90	A
80	B
70	C
65	D
0	F

673

$\geq$	Letter
90	A-
80	B-
70	C-
65	D
0	F

Grading

Component	473	673
Assignments	45%	30%
Midterm	10%	10%
Graduate Paper	---	30%
Course Project	45%	30%

Computation of Component Grades

See the syllabus!

Each component (e.g.,
“Assignment” component) is:
 $\max(\text{micro-average}, \text{macro-average})$

Computation of Component Grades

See the syllabus!

Each component (e.g., “Assignment” component) is:

$\max(\text{micro-average}, \text{macro-average})$

65/90

95/100

95/110

100/110

↑
Assignment grades
(not representative)

Computation of Component Grades

See the syllabus!

Each component (e.g., “Assignment” component) is:

$\max(\text{micro-average}, \text{macro-average})$

65/90

95/100

95/110

100/110

Assignment grades
(not representative)

$$\text{microaverage} = \frac{65 + 95 + 95 + 100}{90 + 100 + 110 + 110} \approx 86.59\%$$

$$\text{macroaverage} = \frac{1}{4} \left(\frac{65}{90} + \frac{95}{100} + \frac{95}{110} + \frac{100}{110} \right) \approx 86.12\%$$

We'll learn what these
are in the semester

Computation of Component Grades

See the syllabus!

Each component (e.g., “Assignment” component) is:

$\max(\text{micro-average}, \text{macro-average})$

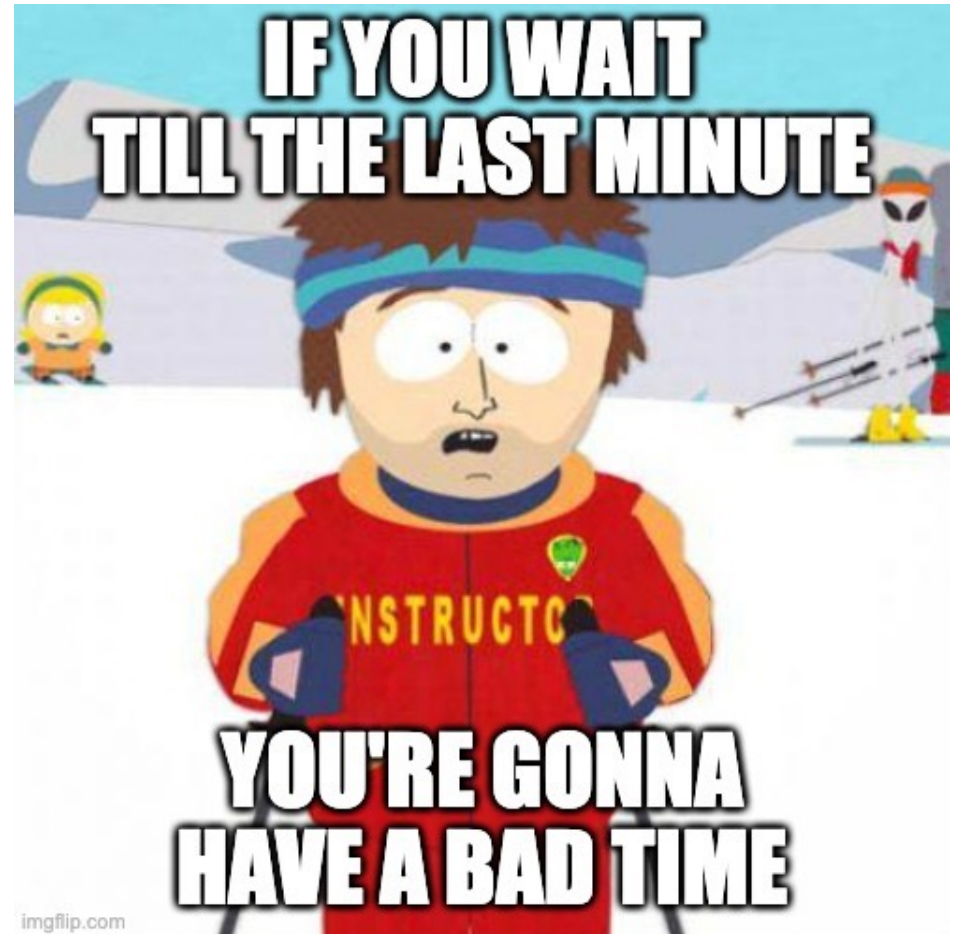
65/90
95/100
95/110
100/110

$$\text{microaverage} = \frac{65 + 95 + 95 + 100}{90 + 100 + 110 + 110} \approx 86.59\%$$

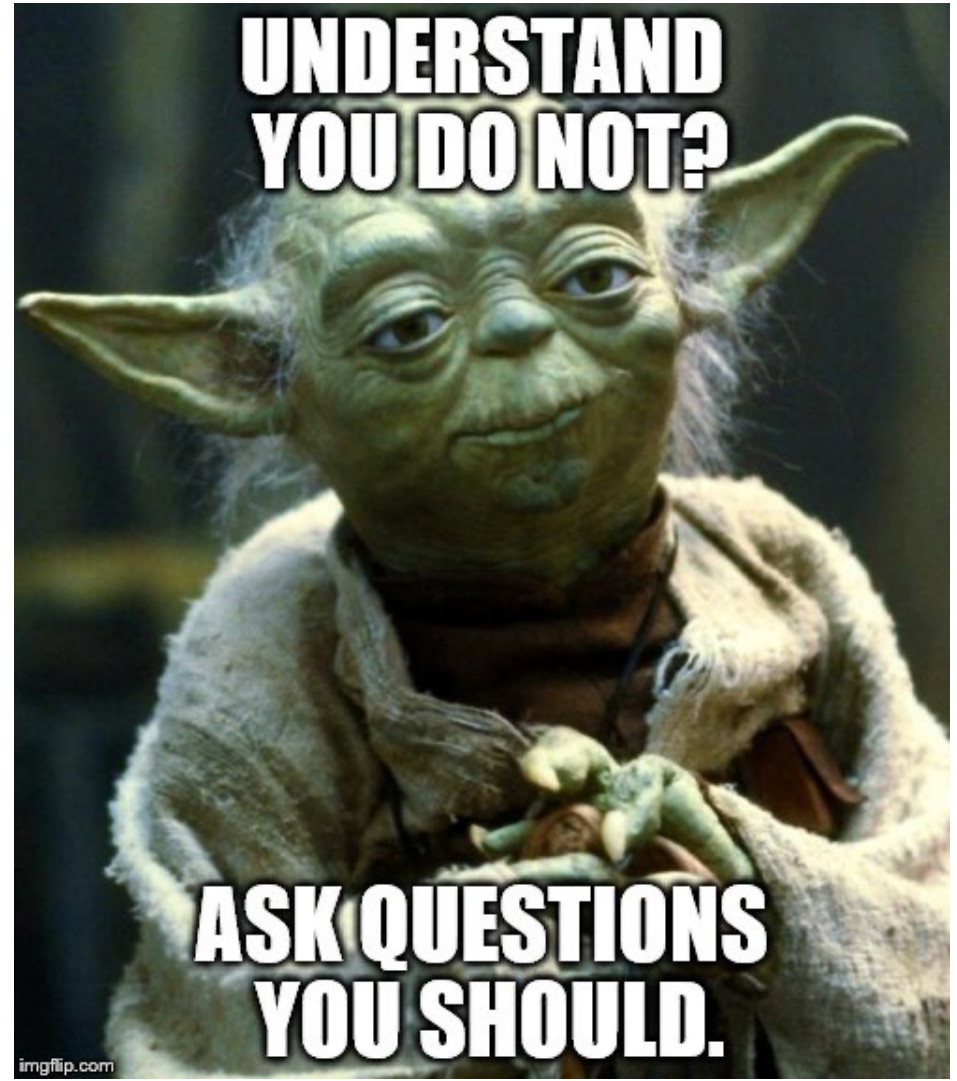
$$\text{macroaverage} = \frac{1}{4} \left(\frac{65}{90} + \frac{95}{100} + \frac{95}{110} + \frac{100}{110} \right) \approx 86.12\%$$



Tip #1 for
Assignments



Tip #2 for Assignments



Running the Assignments

A "standard" x86-64 Linux machine, like `colab` or `gl`

A passable amount of memory (2GB-4GB)

Modern but not necessarily cutting edge software

Don't assume a GPU (if you want to write CUDA yourself,
talk to me)

If in doubt, ask first

Running the Project

An x86-64 Linux machine

Memory and hardware constraints lifted
(somewhat)

If in doubt, ask first

Programming Languages for Assignments

Use the tools you feel comfortable with, though some will require
Python libraries

Python+numpy, C, C++, Java, Matlab, ...: OK

Libraries: Generally OK, as long as you don't use their
implementation of what you need to implement

Math accelerators (blas, numpy, etc.): OK

If in doubt, ask first

Programming Languages for the Project

Use the tools you feel comfortable with

Python+numpy, C, C++, Java, Matlab, ...: OK (straight Python may not cut it)

Libraries: **Use what you want**

Math accelerators (blas, numpy, etc.): OK

Late Policy

Everyone has a budget of 10 *late days*

Late Policy

Everyone has a budget of 10 *late days*

If you have them left: assignments turned in after the deadline will be graded and recorded, no questions asked

Late Policy

Everyone has a budget of 10 *late days*

If you have them left: assignments turned in after the deadline will be graded and recorded, no questions asked

If you don't have any left: they get a 0, but still turn assignments in

Late Policy

Everyone has a budget of 10 *late days*

Use them as needed throughout the course

They're meant for personal reasons and emergencies

Do not procrastinate

Late Policy

Everyone has a budget of 10 *late days*

Contact me privately if an extended absence will occur

You must know how
many you've used

Stretch
break!

