CS 2731
Introduction to Natural Language Processing

Session 1: Course introduction and NLP basics

Michael Miller Yoder

August 26, 2024

School of Computing and Information

Overview: Course introduction and NLP basics

- Introductions
- What is NLP?
- Course logistics

About Michael Miller Yoder

- You can call me "Michael"
- Teaching faculty, Pitt School of Computing and Information
- PhD, Language Technologies Institute at CMU (2021)
- **Research interests:**
  - NLP
  - computational social science
  - ethics and bias in AI

Michael's office hours

- By appointment in person at Sennott Square 6309 or on Zoom
- Sign up for a slot here
  - Link also posted on course website
- Drop in to ask questions about the course or anything else

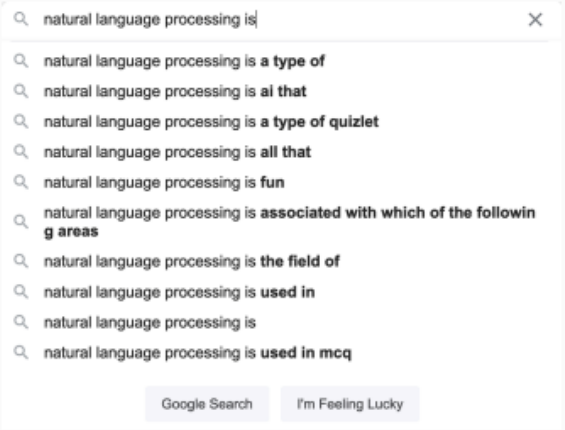
Introductions

1. What is your name?
2. What is your program/year/research interests?
3. What is a language other than English that you speak, or some your ancestors spoke?
4. [Optional] Is there anything that makes you interested in NLP or excited to take this class?

What is natural language processing (NLP)?

NLP is Everywhere

Did you ever wonder how web search engines work...



...or how Google can anticipate what you're searching for?

That's NLP!

Slide credit: David Mortensen

NLP is Everywhere

Did you ever wonder how ChatGPT generates language?

That’s NLP!

NLP is Everywhere

Did you ever wonder how digital assistants work?

That's NLP!

Slide credit: David Mortensen

NLP is Everywhere

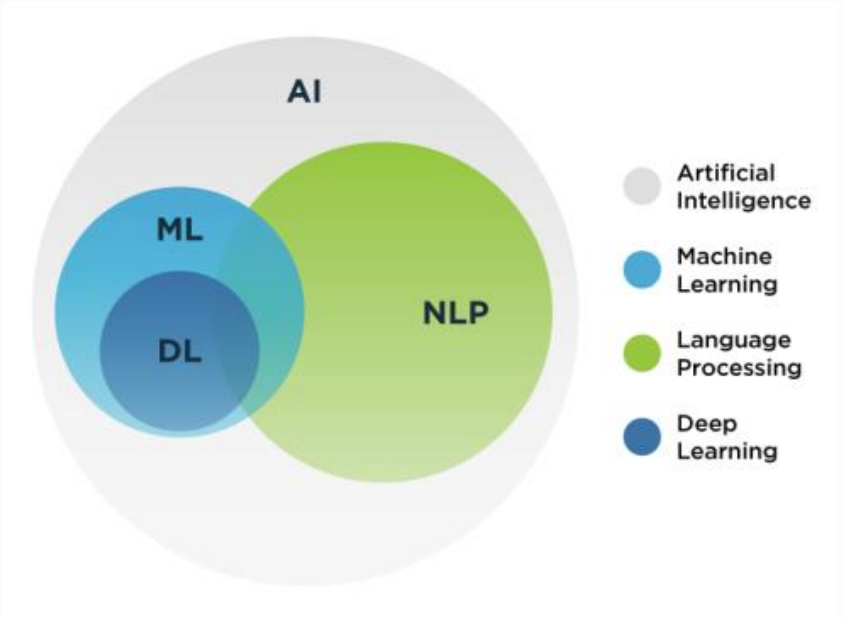
Did you ever wonder how the government is spying on your every word?

That's also NLP!

Slide credit: David Mortensen

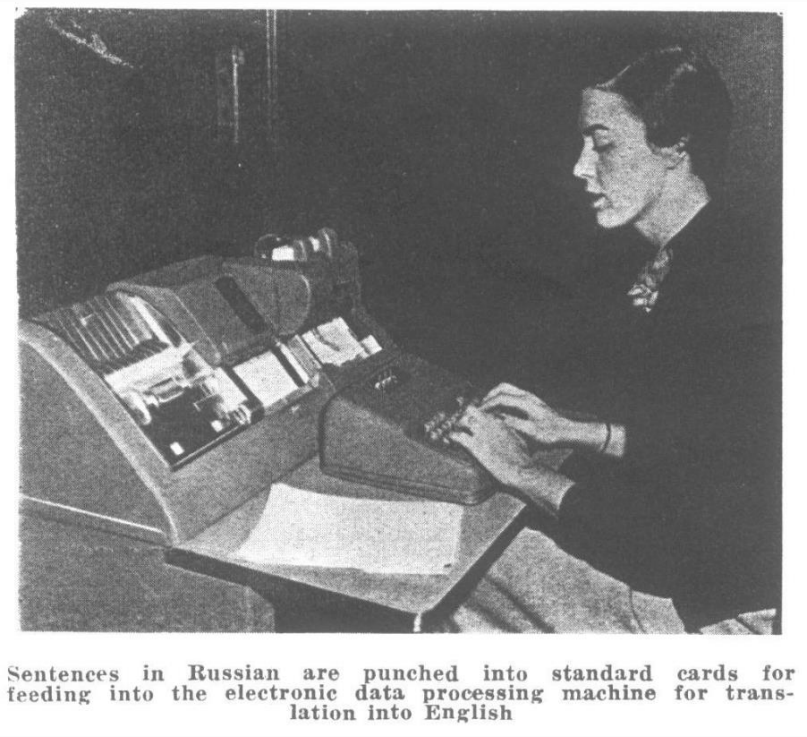
NLP is the Computational Analysis and Synthesis of Speech and Language

- NLP is one of the most important AI fields today
- It is about processing language with computers
- Engineering focus—solving practical problems



Slide credit: David Mortensen

A brief history of NLP



Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

The Georgetown-IBM Experiment.
Credit: John Hutchins

- 1950s: **foundations**
  - Turing Test ("Computing Machinery and Intelligence" paper)
  - Georgetown-IBM Experiment translating Russian to English

A brief history of NLP

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

The Georgetown-IBM Experiment.
Credit: John Hutchins

- 1950s: **foundations**
  - Turing Test ("Computing Machinery and Intelligence" paper)
  - Georgetown-IBM Experiment translating Russian to English
- 1960s-1980s: **symbolic reasoning**
  - ELIZA, rule-based parsing, hand-built conceptual ontologies

A brief history of NLP

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

The Georgetown-IBM Experiment.
Credit: John Hutchins

- 1950s: **foundations**
  - Turing Test ("Computing Machinery and Intelligence" paper)
  - Georgetown-IBM Experiment translating Russian to English
- 1960s-1980s: **symbolic reasoning**
  - ELIZA, rule-based parsing, hand-built conceptual ontologies
- 1990s-2010s: **statistical NLP**
  - Learn patterns from large corpora (feature-based machine learning)

A brief history of NLP

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

The Georgetown-IBM Experiment.
Credit: John Hutchins

- 1950s: **foundations**
  - Turing Test ("Computing Machinery and Intelligence" paper)
  - Georgetown-IBM Experiment translating Russian to English
- 1960s-1980s: **symbolic reasoning**
  - ELIZA, rule-based parsing, hand-built conceptual ontologies
- 1990s-2010s: **statistical NLP**
  - Learn patterns from large corpora (feature-based machine learning)
- 2000s-today: **neural NLP**
  - SOTA on many tasks from "deep" layers of neural networks

NLP and Computational Linguistics

- These terms are often used interchangeably
- If you want to make a distinction:
  - Computational linguistics is the scientific study of language using computers
  - Natural language processing is the development of computational tools to process human language (engineering-focused)
- "Natural language" = human languages (not programming languages)

The other NLP 😂

Neuro-linguistic programming (pseudoscience)

Course objectives and overview

Learning objectives

At the end of this course, a student will be able to structure an NLP system to achieve a desired outcome from language data.

Learning objectives

When coming across a natural language problem, students will be able to:

- Recognize the class of tasks that a specific natural language task belongs to
- Explain the basics of language structure from linguistics (morphology, syntax, semantics, discourse) that are relevant to NLP
- Preprocess text into a machine-readable format
- Extract needed features from text for a variety of tasks
- Identify a suitable model to tackle the task
- Evaluate algorithms for that task
- Identify potential ethical pitfalls in an NLP system and how to potentially address them
- Communicate motivation, key components, and implications of an approach to NLP tasks in writing

Core tasks and applications of NLP

APPLICATIONS

machine translation

speech recognition & synthesis

chatbots

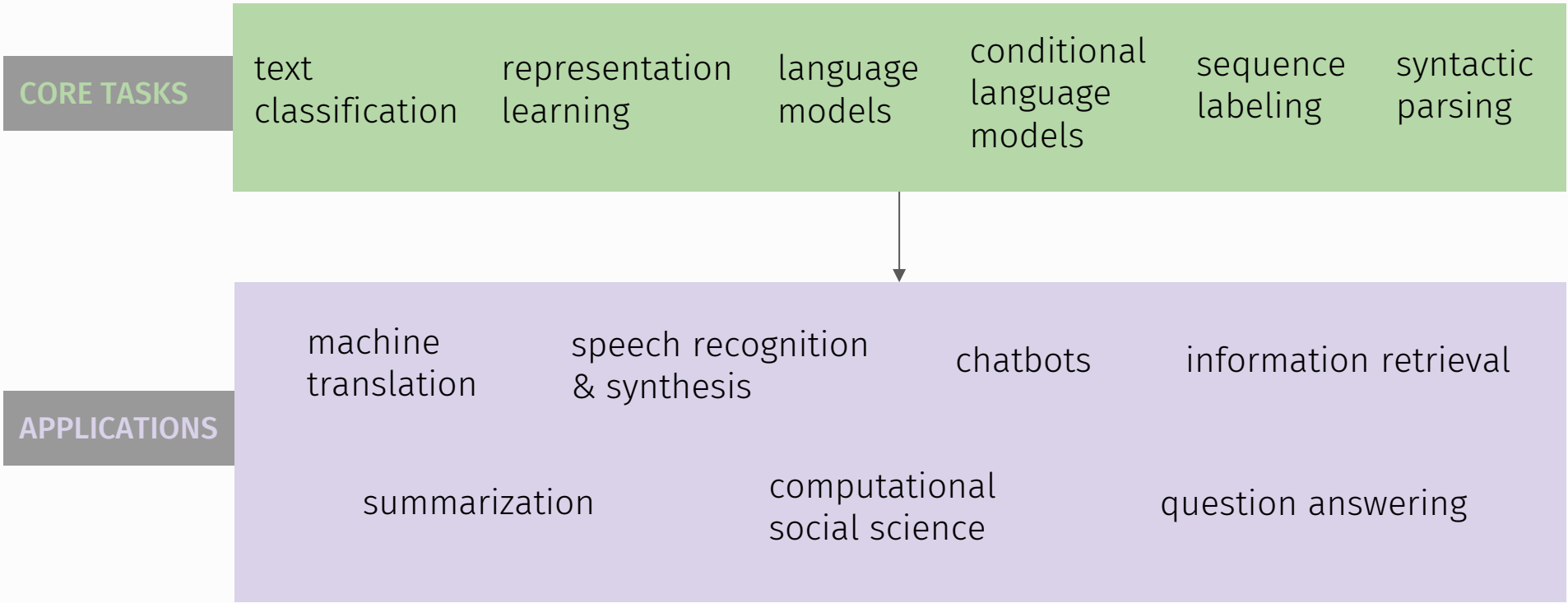
information retrieval

summarization

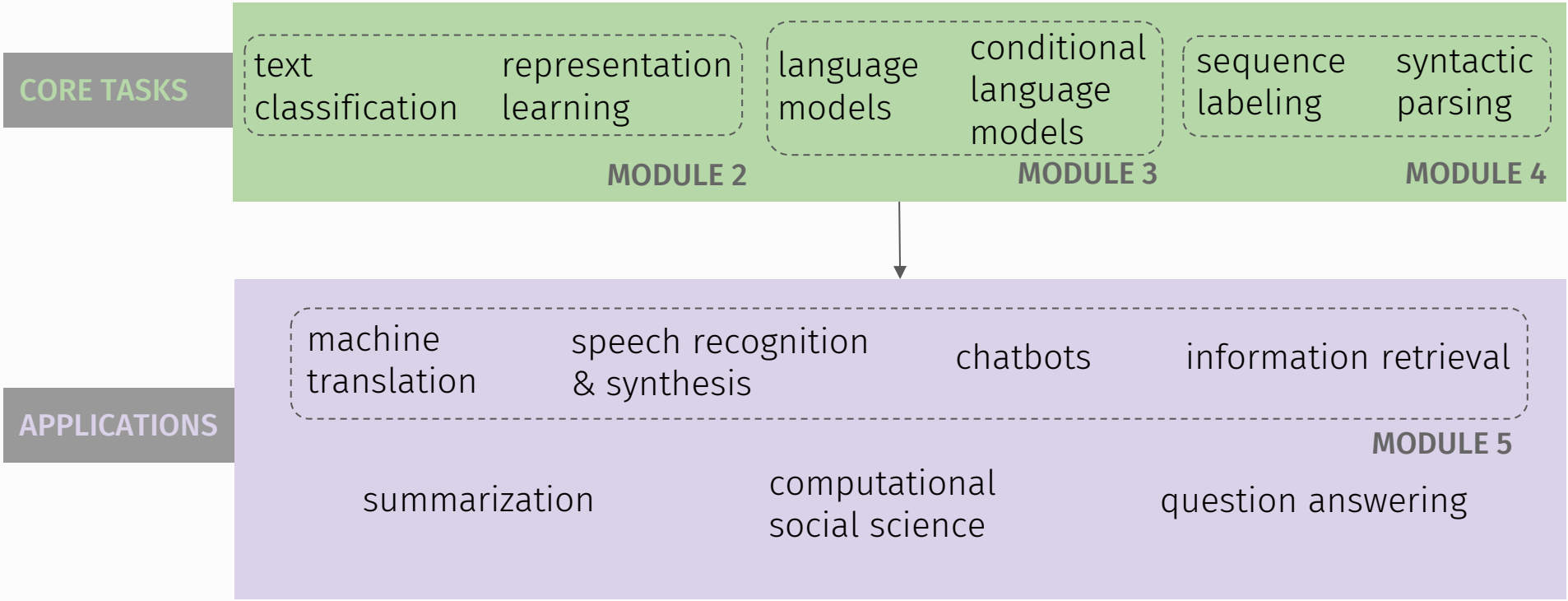
computational social science

question answering

Core tasks and applications of NLP



Core tasks and applications of NLP



Approaches covered in this course

For most NLP tasks, we will cover:

- Classic approaches: symbolic, statistical, feature-based approaches
- Contemporary approaches: neural network-based

Resources

Textbook (free)

- Dan Jurafsky and James H. Martin, *Speech and Language Processing*, 3rd edition draft, 2024-02-03.
- Available completely free online: https://web.stanford.edu/~jurafsky/slp3/
- Why do the readings?
  - Learn better: get the information from readings and lectures
  - Spend class time more efficiently: come with questions
  - Reading quizzes due 11:59pm day of lecture

Lectures

- Cover the most important parts of the course content
- Students are expected to attend each lecture
- **Attendance will be taken via Top Hat at a number of random class sessions**
- Slides will be provided in advance of each lecture for note-taking
- There are no current plans for recording lectures

Infrastructure

- Website
  - https://michaelmilleryoder.github.io/cs2731
  - https://tinyurl.com/nlppitt
  - Up-to-date syllabus and schedule
  - Lecture slides
  - Homework assignment and project instructions
- Canvas
  - Submit assignments
    (homeworks and project milestones)
  - Post on discussion forums
  - Receive course announcements
  - Check your grade

Programming languages and software

- **Python will be the expected programming language used in assignments**
- Python-based data science packages (numpy, pandas, jupyter, scikit-learn, pytorch) will be used and encouraged in both assignments and the project
- If you have zero familiarity with Python (no shame):
  - Check out the **Tutorials on Python and data science** section of the course website under **Learning resources**
- Let us know if you want to use other languages for assignments (it's probably fine)
- You can use whatever you want for the project

Assessments

Assessment overview

Assessment	Points	Percentage of grade
Homeworks (4 total)	224	44.8%
Project	203	40.6%
Reading quizzes	33	6.6%
Discussion posts	15	3.0%
Participation	25	5.0%

No exams

Homework assignments

- 4 total
- 11.2% of total course grade each
- Most will have a written component (working through an algorithm, e.g.) and a coding component
- Due ~15 to 18 days after they are released
- Descriptions will be on the course website
- Submitted through Canvas

Project

NLP is inherently hands-on. The course project will demonstrate an ability to build a system that **makes a contribution** to NLP research or practice.

- Self-selected topic, type of research contribution, and idea
  - Can fit with your research interests outside of this class
  - Come up with your own idea or choose one of the example project ideas
- We will solicit ideas for the projects through a form, to be advertised to all students anonymously
- On another form, you will rank the project ideas students have based on your interests, as well as provide any preferences on who you will work with
- Groups of 2-4 will be assigned by the instructor and TA based on project idea interests, skills, and group preferences from students
  - There will be group member evaluation
- Types of contributions: new dataset and/or annotations, new approach/application, new evaluation, new survey

Project components

Component	Points	Percentage of course grade
Interest survey response	5	1%
Project area and type of contribution	10	2%
Proposal and literature review	35	7%
Peer review	2	0.4%
Proposal presentation	*None*	*None*
Basic working system report	30	6%
Final presentation	*None*	*None*
Final report	121	24.2%

Reading quizzes

- On Canvas
- Quick checks for comprehension
- Designed to motivate you to do the reading and come to class
- Simple, auto-graded (generally multiple choice or short answer)
- Indicate most confusing topic (not graded)
    - Helps the instructor know what needs review
- Only 6.6% of your course grade total
- The lowest 2 quiz scores will be dropped
- **Due by 11:59pm on days with class sessions**
- Can't redo after they're due

First will be due next Wednesday, Sep 4

Discussion posts

- There will be 3 required discussion posts from readings,
- Often on the social impact of NLP (bias, transparency, etc)
- Respond to a prompt
- Post on a discussion forum on Canvas
- Add your own ideas, respond to others
- Minimum 100 words with a substantive idea or response

Participation grade

- Class interactions (activities, discussions) are better with more people in class
- Incentives to come to class and engage
- 5% participation grade
  - 3%: attendance on a random subset of class sessions, taken via Top Hat
  - 2%: engagement
    - Have you ever asked a question in class, afterward or over email?
    - Do you participate in in-class activities?
    - If yes to either, you will be fine

Policies

Late work

- Students are granted 5 total late days across all homework assignments without penalty.
- After those five late days, you will be penalized **20% for each day that your submission is late** except in extreme unforeseen circumstances.
- Group project work will be penalized 20% for each day late. No late work will be accepted for the final project report.

Academic integrity

- Students in this course will be expected to comply with the University of Pittsburgh's Policy on Academic Integrity. Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity
- Discussing tools, concepts, and formalisms is acceptable collaboration
- Sharing code is prohibited

Generative AI policy

- You are allowed to use generative AI (ChatGPT, DALL-E, GitHub Copilot, etc) in some circumstances
  - Exposes you to the current capabilities and limitations of such systems
- Allowed use:
  - **Use as an aid, not for a finished product.** Generating ideas, study guides, bibliographies (watch for hallucinations, though) is ok. Drafting entire homework assignments or project reports, even if you revise the draft, is not ok.
  - **Cite its use.** Citing the generative AI's tool contribution to your work is required. See the APA guidelines on how to cite ChatGPT.
  - **You are responsible for the work you turn in.** LLMs and other generative AI systems can and do generate biased, socially problematic language and assert unfounded claims.
- When in doubt, ask instructor if specific uses are ok. There will be no retaliation for asking.

Disability rights

Many people have disabilities. **We view disabilities as deficits not in disabled people but in the institutions and societies that are structured to disadvantage disabled people.**

If you have a disability (visible or invisible), please let us know as soon as possible (you don't need to tell us the nature of the disability). You are encouraged to work with Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term. DRS will work with you to determine reasonable accommodations for this course. This might include lecture materials that are usable by people with visual disabilities, sign language interpretation, captioning, flexible due dates, etc.

Adapted from policies by David Mortensen and Lori Levin

Maintaining scholarly discourse

In this course we will be discussing some complex issues. It is essential that we **approach this endeavor with our minds open** to evidence that may conflict with our presuppositions. Moreover, **it is vital that we treat each other's opinions and comments with courtesy even when they diverge and conflict with our own**. We must avoid personal attacks and the use of ad hominem arguments to invalidate each other's positions. Instead, we must develop a culture of civil argumentation, wherein all positions have the right to be defended and argued against in intellectually reasoned ways. It is this standard that everyone must accept in order to stay in this class; a standard that applies to all inquiry in the university, but whose observance is especially important in a course whose subject matter is so emotionally charged.

Questions?