Ethics Module

Introduction to Machine Learning

Adya Danaditya & Dani Grodsky

10–315 Introduction to Machine learning: Overview

- High-level summary of existing learning objectives: understand ML principles; select and apply appropriate supervised learning algorithms (Decision Trees, Neural Nets, SVM, Bayes Nets)
- Semester-long course, ~100 students per semester
- Two 80-minute lectures & one 80-minute recitation each week
- Previously used active learning techniques: Think-Pair-Share, in-class polling, breakout room discussion

Constraints:

- Desire to have specific cases and questions to help guide any in-class discussion - to reduce any need to spitball
  Strategy - *Paper/article based ethics discussion*
- Ethical concepts can be hard to test / assess but there should be some graded component
  Strategy - *Questions in assignment and participation grading in discussion boards*
- It is difficult to find full lecture time before the last week of class
  Strategy - *Small chunks scattered throughout the semester*

Our Ethics Plan

Visual cue that will pop up throughout the class as we talk about ethics

The module we envision consists of:

1. **Short inclusion in the intro lecture** that:
   - ○ Touch on logistics & teases upcoming activities regarding ethics
   - ○ Highlights the general topic and instances of <u>dataset bias</u>

2. **Learning activities** (homework and/or class) on ethical tie-ins relevant to chosen class topics scattered throughout the semester

   Short review in lecture

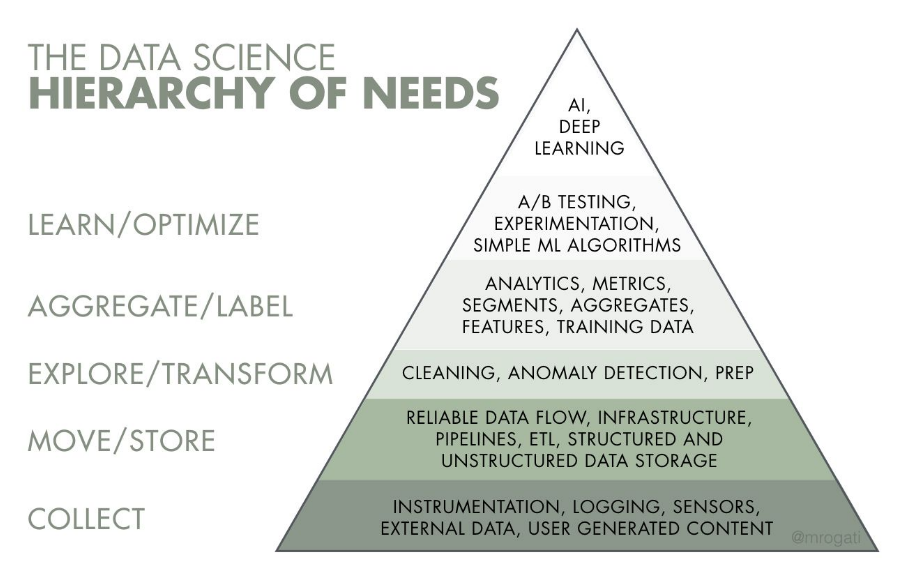
   Prep through assigned work

   Reflect via discussion board

3. **Team-based activity** spanning the last two classes

Opening Lecture

Ethics Introduction and Dataset Bias


THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

Learning objective: Assess how characteristics of the dataset and its collection can affect analysis outcomes

Short segment on data collection and data set bias:

1. **Introduced through the data science 'hierarchy of needs':**
   - Proper data collection is the foundation of everything

2. **Explored through two relevant case studies**:
   - Amazon's hiring algorithm, which was trained on previous resumes (of mostly men) and flagged woman-related parameters as negative
   - Skin cancer detection model that was trained on mostly images of lighter-skinned patients and does not generalize well to darker-skinned people

Learning activities and the topics involved

Core Concept

Deep Dive

Real-World Application

What is explainability, why is explainability desirable most of the time, and how the opinion inside the AI community differs on this?

Via lecture in class

Do a deeper reading on the topic - investigate nuances, current state of affairs and how it played out in different industries and domains

Through reading material

Comment on the consequences of having explainability or lack thereof in real-world applications (criminal justice, health, finance and science)

Via discussion board

Learning objective: Define the concept of explainability as well as its consequences in a real-world application

Learning activities and the topics involved

Core Concept

Deep Dive

Reflections

How recommender systems influence short and long-term user happiness and societal well being

Via lecture in class

Explore topic deeper: Personal preference is often malleable; Use of multi-stakeholder instead of user-centered approach to address externalities, etc.

Via podcast listening

Answer podcast related prompts:

"Before I ever use Twitter, my political views were some set of beliefs. And then after I use Twitter, my political views were a different set of beliefs. **I changed as a person from that interaction.**" Have you had an experience where it seemed that a recommender system noticeably influenced your beliefs, decisions or actions?

Via discussion board

Learning objective: Explain ethical considerations related to the creation and use of recommender systems

Culminating Learning Activity – First Step

Healthcare

Privacy

Financial Health

Sustainability

Education

Civic Engagement

Breakout rooms: Have student groups pick one of six topics (left) and develop a pitch for a ML application for good in that topic area

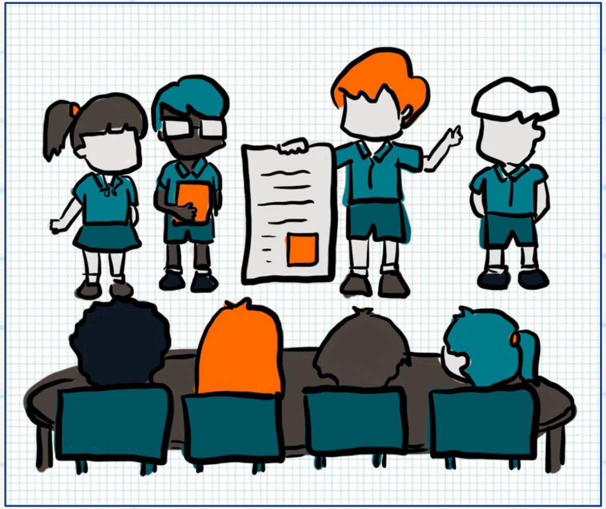
Pitch: 6-8 sentences, including brief plan for data collection and model implementation

Share: Post pitch in discussion board for other groups to see and partner group to review

Learning objective: Generate applications of Machine Learning for social good and summarize ethical considerations of your own and ideas and that of others

ML for Good

Culminating Learning Activity – Second Step



Groups will be partnered with another team that selected the same topic to do a **peer review** based on ethical considerations of the idea

Prompts are provided to help kick start group's review:

- Problems that may arise from the data collection / selection
- Risks of overgeneralization
- Impacts to end-users and society:
  - Is there a potential for the model's impact to change over time? If so, how might monitoring play a role?
  - What should an end-user do if an issue arises? Who is responsible?
  - Could it be misused intentionally or unintentionally?

A random selection of teams will be chosen to speak for up to 3 minutes about highlights of their peer review discussion, the rest of the groups will post takeaways on discussion board

Post Class

Survey and Grading

Grading components:

- Would take 3% of the whole class grade, as part of the 5% participation grade
- The 2 discussion activities (Explainability, Recommender Systems) will amount to 1% and the ML for Good activity would take the rest - and they are mostly graded by participation

Post class-survey

Survey Questions:

1. How relevant do you think the ethics material (those signified by the Ethics and ML logo) is to the class topic being discussed?
2. Do you think the ethics material infused in this class is valuable to you as a CS graduate?
3. What ethics activity do you like the most?
   a. The introduction lecture and the concept of dataset bias
   b. The discussion around explainability
   c. The discussion around recommender systems
   d. The final class activity
4. What can be improved if we want to continue this part of the class moving forward, and do you have any other comments on the whole activity?

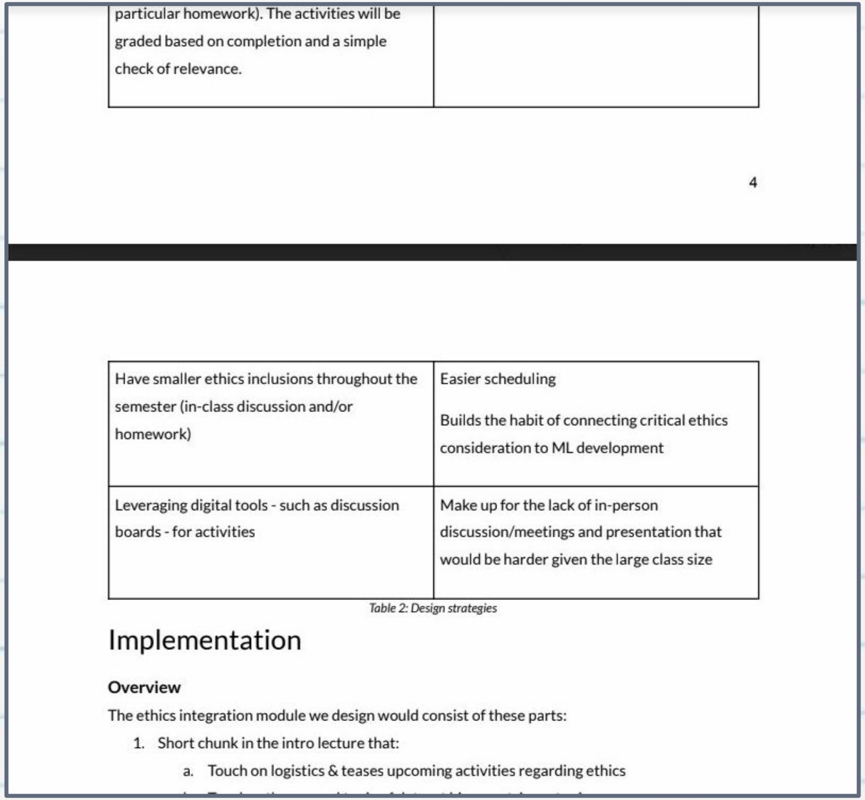
- See engagement rate and quality of student comprehension from survey
- Feedback should be used to design further iterations of this module in the future

Implementation

Instructor's Pack

Teaching Materials:

- Slides + notes
- All related papers
- Templates for groups
- Homework question prompts
- Post-class survey detailed wording



particular homework). The activities will be graded based on completion and a simple check of relevance.	

4

Have smaller ethics inclusions throughout the semester (in-class discussion and/or homework)	Easier scheduling Builds the habit of connecting critical ethics consideration to ML development
Leveraging digital tools - such as discussion boards - for activities	Make up for the lack of in-person discussion/meetings and presentation that would be harder given the large class size

Table 2: Design strategies

Implementation

Overview

The ethics integration module we design would consist of these parts:

1. Short chunk in the intro lecture that:
   a. Touch on logistics & teases upcoming activities regarding ethics

Thanks!