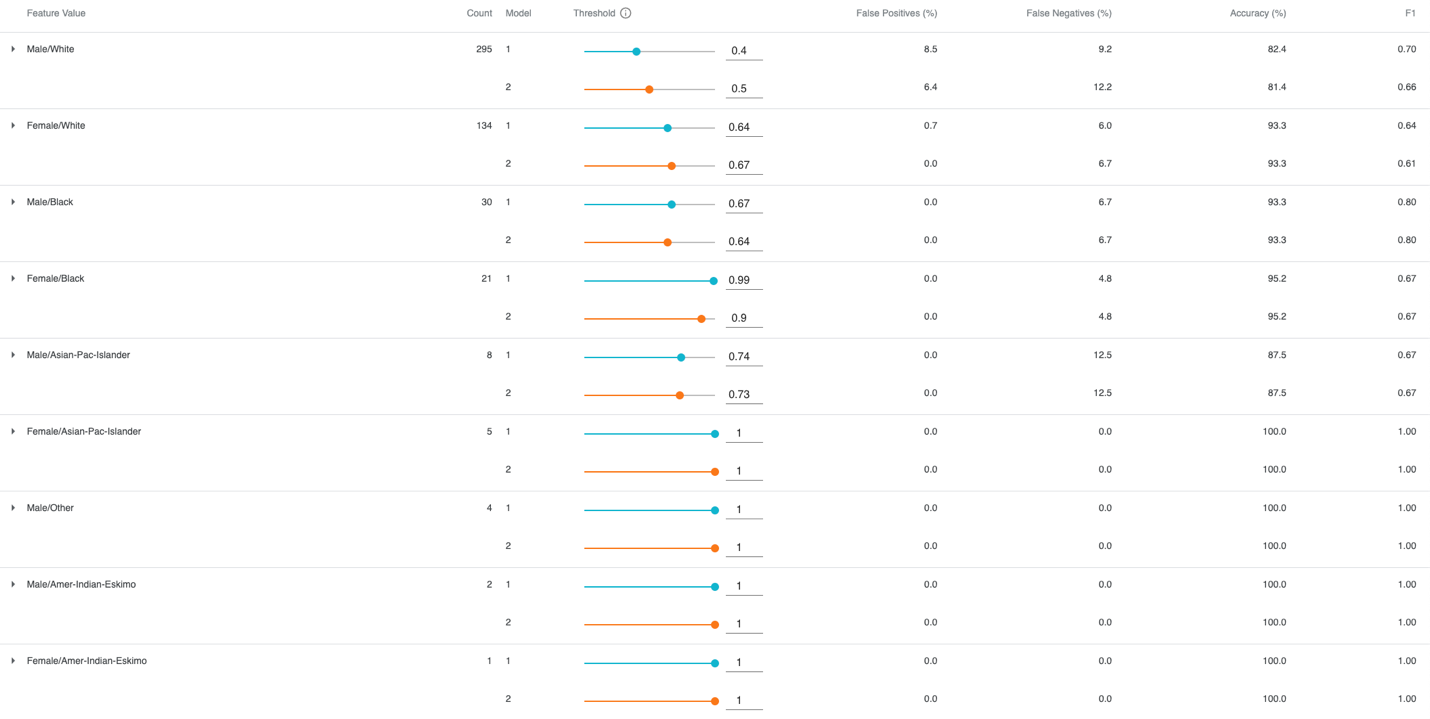
Exercise 6: What If Tool

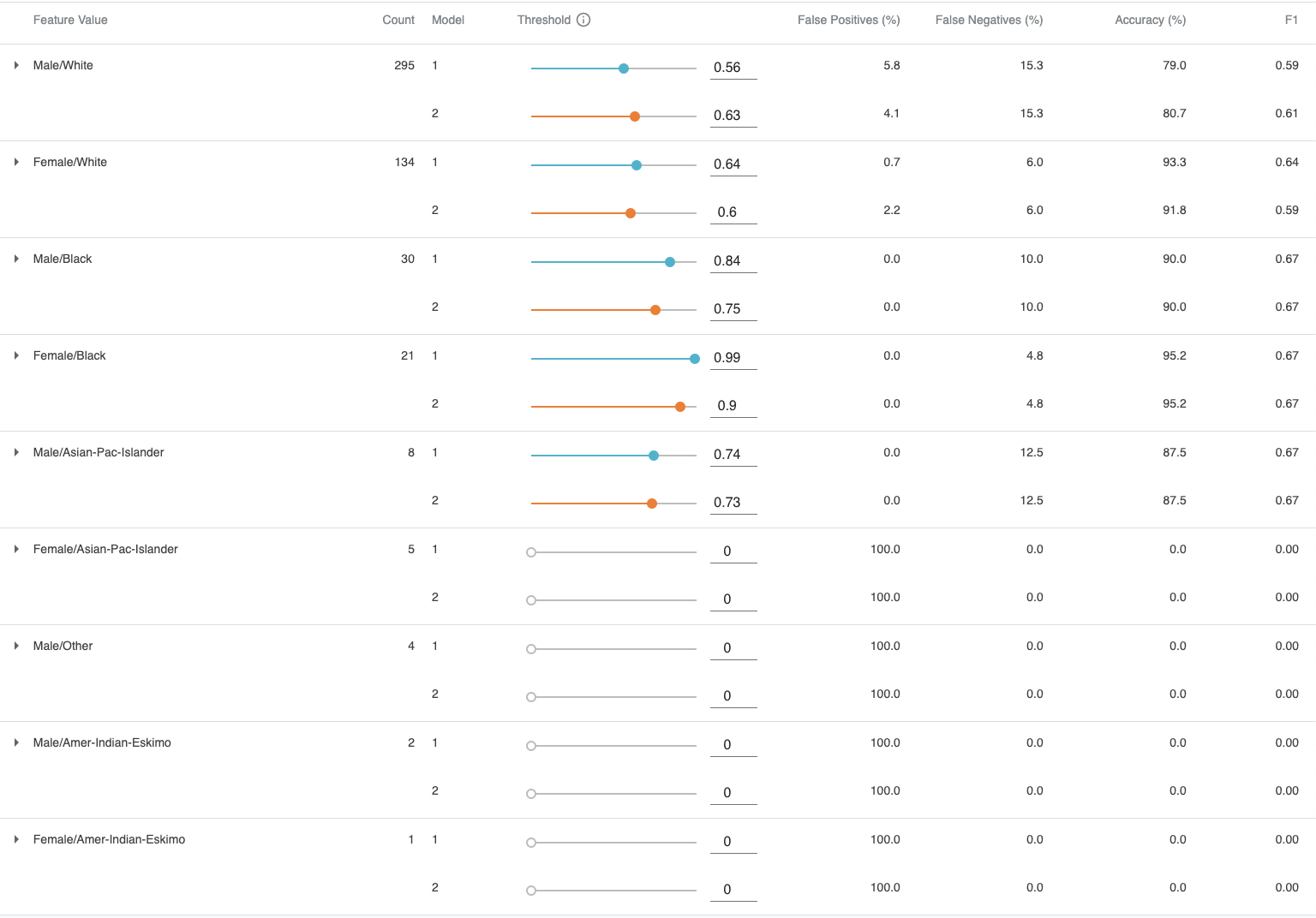
* **Selected web demo:**Compare Income Classification on UCI Census Data
* **Slides:** Sex & Race
* **Which methods seems to be better?**Equal Accuracy

Equal Accuracy seems better because it has the highest accuracy (ranging from 100% to 81.4%), the lowest rate of false positives (ranging from 8.5 to 0) and false negatives (ranging from 12.2 to 0) among other strategies. Details are shown in the below chart.



* **Which methods seems to be worse?**Equal Opportunity

Equal Opportunity seems worse because it has the highest accuracy (ranging from 95.2% to 0%,), and the rates of false positives (ranging from 100 to 0) and false negatives (ranging from 15.3 to 0) also do not show a great advantage over other strategies.

Due to the low counts, all strategies seem to struggle with the extreme minority groups - Asian-Pac-Islander, Amer-Indian-Eskimo and Others. Therefore, evaluating these data characteristics can be less helpful in practice.

Exercise 5: AI Fairness 360

[https://aif360.mybluemix.net/data#Links to an external site.](https://aif360.mybluemix.net/data)

* Select one of the sample data sets**(Step 1)**
* Select 2 of the bias mitigation algorithms**(from Step 3)**and explain any differences in the “Compare original vs. mitigated results” **(Step 4)**when applying the two different methods. Which method seems to be better? Explain.

**Dataset: Adult census income**

**Mitigation: Reweighing algorithm and Adversarial Debiasing algorithm**

* + Differences in the “Compare original vs. mitigated results.”
    - Both algorithms seem to mitigate the bias against the unprivileged groups (Non-white and Female) to acceptable levels for 1-2 previously biased metrics (details below). Using the Reweighing algorithm, the accuracy after mitigation changed from 83% to 82% for Race and 83% to 81% for Sex; While using the Adversarial Debiasing algorithm, the accuracy after mitigation changed from 83% to 77% for Race and 83% to 73% for Sex. Additionally, the difference between “Compare original vs. mitigated results” after applying the Adversarial Debiasing algorithm seems more significant than the Reweighing algorithm.



* + Which method seems to be better? Explain.
    - The Adversarial Debiasing algorithm acts better than the Reweighing algorithm. After applying the Reweighing algorithm, the Bias against the Non-white group was reduced to acceptable levels for only 1 of 2 previously biased metrics. In contrast, the Adversarial Debiasing algorithm improved the results to 2 of 2. In addition, Race and Sex have 0 of 5 metrics, still indicating bias for the unprivileged group after applying the Adversarial Debiasing algorithm. Therefore, The Adversarial Debiasing algorithm seems to be the better option for the Adult census income dataset.

Exercise 4: Predictive Algorithms

I used 7625 Cottonwood Drive, Georgetownship, Michigan to report the below:

* Crime Layer
  + Slightly below avg. (<0.62)
* Income level
  + $​63,400 ± $17,244
* Majority Demographic
  + White 89.5%

Case Study 6: Predictive Algorithms

* Is it ethical for Facebook to mine its users’ posts for signals that those users are about to go through a break up? Is it ethical for the company to then help its clients target their ads based on this research?
  + No, it is not ethical for Facebook to mine its users’ posts for signals that those users are about to go through a breakup. When going through a breakup, people are likely to be in a vulnerable or emotional mental state, with a weaker risk-taking ability. Therefore, the conclusions drawn by Facebook could easily pose more significant potential risks to these vulnerable users if used for other unknown purposes.
  + No, it is unethical for the company to help its clients target their ads based on this research. As mentioned before, target users in an emotionally unstable state pose an unknown risk to them, which is greater than that of other ordinary users.
* Is what Facebook is doing so much different than what other companies do? Explain.
  + Personalized advertising has become increasingly popular as an effective marketing tool, while users' data and privacy have been traded off instead. In this case, Facebook's finding is essentially about discovering users' vulnerabilities and corresponding behavioural patterns, then making it possible for other parties to exploit such characteristics of their users. Other companies may not use the "vulnerability" of their users but identify and leverage users’ other factors, such as "personal preferences" and "shopping habits," to promote the ads. These behaviours are fundamentally similar—for instance, Amazon Ads and Youtube promotions.
* Identify one use/application in which this type of prediction might cause significant ethical harms, then briefly describe the specific harms that might be caused and who they might affect.
  + Let's say winery factories use this finding from Facebook to hold a wine-tasting event and further promote their wine products. According to the prediction, several users who are mentally vulnerable will be more likely to accept the invitation and participate in the event. These users will indeed create profits for the winery business. Still, the users are more likely to expose themselves to the risk of excessive alcohol dependence, which is increased by their ongoing vulnerability. Therefore, these users who attend the event could become addicted to alcohol, harming themselves physically and mentally.

Case Study 5: Facial Recognition

* Who are the stakeholders involved? Who should be consulted about such a project's goals and development?
  + The stakeholders involve the guests who would visit the Shelter, staff working there, residents living there, and the company that develops the facial recognition system. All the stakeholders should be consulted about such a project's goals and development to ensure the system's proper use and address the ethical issues the system may cause.
* Do you think there should be governmental regulations imposed on the company? What about an internal ethics board? What about an external ethics board? Why?
  + Yes, governmental regulations should be imposed on the company, and both internal and external ethics boards should be involved.
  + First, the facial recognition system is expected to follow the data use ethics regulated by governments; secondly, as mentioned in the lectures, the system may bring potential ethical issues against protected classes due to the less recognition accuracy of minorities. The governmental regulations and internal and external ethics board should work together to ensure that the company is regulated and operates following established rules and that proper accountability and problem-handling processes are in place to ensure that issues are addressed in a timely and appropriate manner when they arise.
* What are some other ethical issues that any designers/developers of such an application would need to address? Which of these ethical issues would you prioritize, and why?
  + Several ethical issues need to be addressed by the designers/developers as follows.
    - Privacy. The facial recognition system needs to request explicit permissions from the users to collect and use the data.
    - Abuse use of data. The designers should ensure the security of the user data and prevent using the data from being used for other purposes and by other parties.
    - Racial profiling. The designers should be able to mitigate the racial profiling that is potentially caused by the unsupervised learning of the AI agent and increase the facial recognition accuracy for different racial groups.

I would prioritize the users' privacy because the earlier the design involves the users, the better outcomes usually the plan would have – having users' consent is the first step. Additionally, the other two issues can be mitigated to some extent if the privacy issue is appropriately addressed.

* Thinking of protected classes, which groups could this most disadvantage? Why?
  + Overall, minority groups are more or less disadvantaged due to the lack of sample data, and I think Race is most disadvantaged compared to other protected classes (e.g., Sex and Age). The reason is that for facial recognition algorithms, the different facial and skin colour characteristics of different races are intuitive and are directly parsed into data to train the AI agent, so when the facial reorganization system is put into use, Race should be the most affected.
* Thinking of protected classes, do you think the algorithms are able to distinguish between members of different groups based on facial images? If yes, which ones?
  + Yes, the algorithms can distinguish between members of different groups based on facial images with less accuracy than expected. Taking Race as an example, the algorithm should be able to distinguish Caucasians from Black people. Then what if we want the algorithm to differentiate between an Asian with a darker skin tone and a black person with a lighter skin tone? I expect the accuracy of the algorithm to decrease in this case.

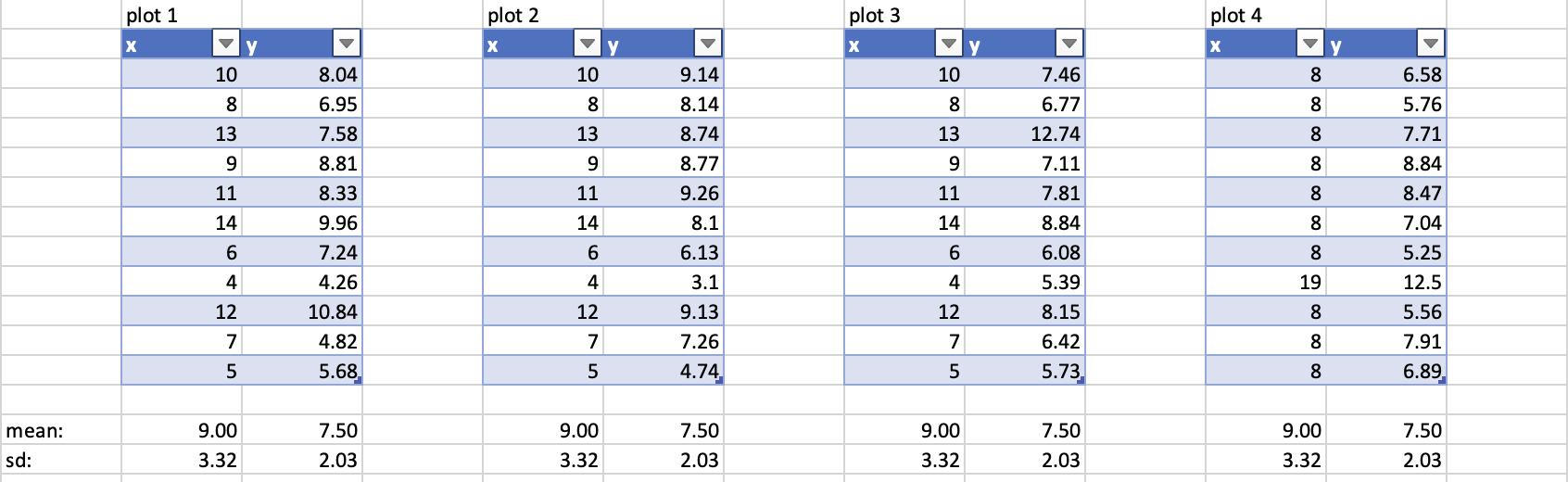
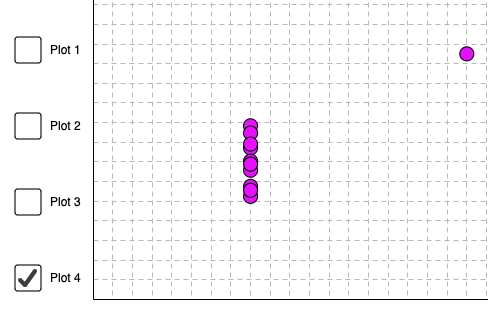
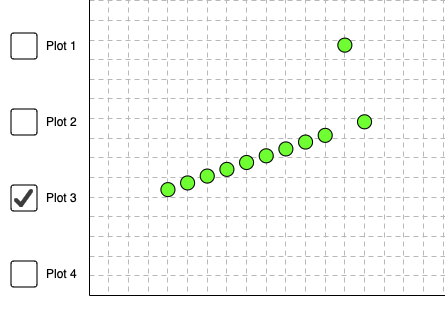
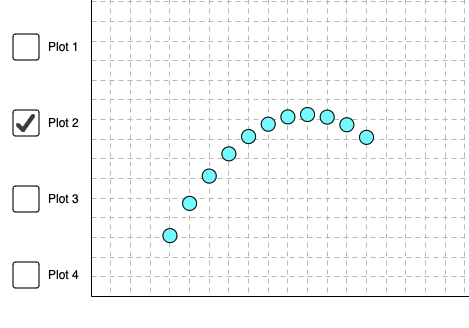
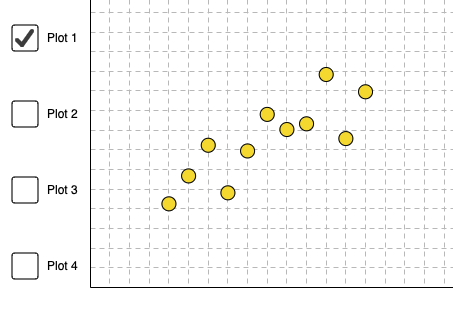
Exercise 3: Bias in Word Embeddings

* Find five images that you think would show bias based on the discussions in class and feed into URL field (e.g. using google image search).
  + [Male Housekeeper](https://drive.google.com/file/d/1TTujpRc5J9APwwAacwbwMuO-QZi3QBmW/view?usp=sharing)
    - The results show biases against the male (aka. gender type of bias) because all the predicted situations show that the agents are women while it is a male housekeeper - mopping and holding a baby can be both female and male work.
  + [Female Engineer in labs](https://drive.google.com/file/d/1-sNwEQ033Yd-ny1uJesp4u2_jdrTLt_q/view?usp=sharing)
    - The results show biases against the female (aka. gender type of bias) because no agents are predicted as women even though it is a female engineer in the image. Women can also interact with machines and work as engineers, not just men.
  + [Female Software Developer sitting by monitors](https://drive.google.com/file/d/1D3mK8z1U4-MDWATG8YeCafZ-fkO6nckI/view?usp=sharing)
    - The results show biases against the female (aka. gender type of bias). The image shows a female sitting in front of the monitors, and all predicted situations think an agent is a man, which is biased by assuming only men can code.
  + [Female CEO in the office](https://drive.google.com/file/d/19_E78zJho-ea6RHWXZs-dDuFJZ3CZNvm/view?usp=sharing)
    - The results show biases against the female (aka. gender type of bias). Even though female CEOs are not as many as males, not all are male. Similar in imSitu train set shows all male faces, and most predicted situations take a man as the agent.
  + [Three Black Female](https://drive.google.com/file/d/18f2WaUuQ5gtFlNYL34wR3BJeO9OZdlyF/view?usp=sharing)
    - The results show biases against black people (aka. race type of bias) because even though the three black females are just standing there with no facial expression, imsitu predicted it as negative situations like weeping, grieving, drinking, crying and nipping. I have tried a few similar images, but the predicted situations are more neutral with white people.

Case Study 4: Word Embeddings

* Identify **two different uses/applications of data** in which biases in word embeddings might cause significant ethical harms, then briefly describe the specific harms that might be caused in each of the two applications, and who they might affect.
* Loan Application. Using data in which biases in word embeddings might cause significant ethical harm in Loan Applications because if the algorithm is biased, the information submitted by applicants would also be evaluated in a biased way. For example, as mentioned in the lectures, CEOs are usually associated with male names, and job titles can be a crucial indicator for banks to assess load applicants' repay ability. Therefore, in this use case, male applicants would be at an advantage while female applicants would be ethically harmed.
* School Admission. The use of data in which biases in word embeddings might cause significant ethical harm in School Admission. Taking the admission to CS graduate programs as an instance, if the algorithm learns from historical data and is biased by the names of the majority population of students (white male names per se), then the admission decisions can be very likely biased against minority communities – females, black people etc.
* Some researchers have designed ‘debiasing techniques’ to address the solution to the problem of biased word embeddings. (Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” NIPS'16, 2016.) Such techniques quantify the biases, and then use algorithms to reduce or cancel out the biases that would otherwise appear and be amplified by the word embeddings. **Can you think of any significant tradeoffs or risks of this solution?**
* Algorithm’s working efficiency. In addition to the existing algorithms, such techniques introduce additional logic attempting to debiasing word embeddings. Therefore, extra time and memory are necessary to make it work. Therefore, algorithms’ performance could be one of the significant trade-offs for such techniques.
* Potential reverse biases. First, it is challenging to define fairness under different circumstances, so while such techniques mitigate existing word embedding biases, they may introduce new biases from other perspectives. For example, for the loan approval case in the lecture, is it still fair to the applicants who the original algorithm should have approved if the algorithm follows statistical fairness?

Exercise 2: Anscombe's Quartet

* What do you notice about the means and standard deviations of x and y for the data sets given in each plot?
  + I notice that the means and standard deviations of x and y for each plot are the same as follows. 
* What do you notice about the graphs for the data sets given in each plot?
  + The graphs loos very different – see below.
  + 
* How would you respond to a classmate who says that summary statistics are enough?

Summary statistics can only describe a limited number of data characteristics in a very general way. As indicated in the answers to the last two questions, even though the mean and standard deviation values are the same for all four plots, the correlation between x and y can be very different.

Exercise 1: Smoking Sampling Bias Design

* Describe what type of sampling approach you would implement to determine the rate of smoking among teenagers.

I would take the Cluster random sampling approach to calculate teenagers' smoking rate. Steps as follows:

* + Contacted principals at three high schools (rural, urban, and suburban) and asked for entire lists of their student bodies.
  + Randomly selecting 50 students from each school to obtain a representative sample of teenagers in three regions: rural, urban, and suburban.
  + Get the answers to questions about youth smoking through direct one-on-one interviews.
  + Weight the study results according to rural, urban, and suburban population shares to conclude.
* Try to minimize bias in your experiment but also discuss possible sources of possible bias in your sampling approach.
  + Possible sources of possible bias:
    - Social Desirability Bias: teenagers may tend to agree with the direction of the leading questions asked by the interviewer.
    - Leading Question Bias: teenagers may behave in a manner that will be viewed favourably by the interviewer.
  + To minimize bias in your experiment:
    - Hide the purpose of the study when conducting interviews.
    - Start the interview with a casual chat to earn the interviewee’s trust.

Case Study 3: Facebook Manipulation

* Were Facebook’s users justified and reasonable in reacting negatively to the news of the study? Was the study ethical? Why or why not?
  + Yes, Facebook's users were justified and reasonable in reacting negatively to the study's news. The study was unethical. The reasons are as follows.
    - According to U.S. and European standards, users should be explicitly informed of the research subject when participating in medical or psychological research - Facebook obviously violated these rules.
    - Special user groups, such as minors or depressed patients, were not excluded from the study. This could affect the generalizability of the study results on the one hand. On the other hand, the research process and results may further harm these participants' mental and physical health, and such effects cannot be traced or reversed.
    - Facebook abused users’ trust by ignoring their privacy and collecting their information without their knowledge. Facebook also does not indicate the Accessibility and Accuracy of the data they collect. from its users.
* To what extent should those involved in the Facebook study have anticipated that the study might be ethically controversial, causing a flood of damaging media coverage and angry public commentary? If the negative reaction should have been anticipated by Facebook researchers and management, why do you think it wasn’t?
  + I strongly agree that those involved in the Facebook study should have anticipated that the study might be ethically controversial. Researchers and management have not expected it because they need more human accountability in data practices and systems. On the one hand, the study must consider the ethical issues that may arise. On the other hand, there should be a clear division of work among the research team to ensure accurate traceability when future problems occur in the study. Ethical and legal issues related to the research should be thoroughly researched and discussed before the research begins.
* Describe 2 or 3 things Facebook could have done differently, to acquire the benefits of the study in a less harmful, less reputationally damaging, and more ethical way.
  + Fully explain the study's intent to the user and obtain users' explicit consent before conducting the study.
  + Participants should be assessed before the start of the study to exclude special populations, such as minors who lack emotional stability and those with psychiatric disorders, to avoid further harm to them.
  + Design the research with full consideration of the potential ethical implications that it may cause, and ensure that users are in good hands when the issue arises.
* Who is morally accountable for any harms caused by the study? Within a large organization like Facebook, how should responsibility for preventing unethical data conduct be distributed, and why might that be a challenge to figure out?
  + The leaders of this research are primarily responsible for any harm caused by the study.
  + Within a large organization such as Facebook, a dedicated team should be formed to evaluate the ethical aspects of the ongoing projects and address such issues for the users.
  + That can be a challenge because:
    - First, people's understanding and definition of ethics are generally subjective and difficult to unify and implement into clear rules and regulations.
    - Secondly, to have a person in charge of every step of every project in such a large enterprise will inevitably affect the efficiency of operation and add substantial human costs.

Case Study 2: Emails Exposed

* Should universities be allowed to monitor student email and social media accounts? If so, under what circumstances?

Universities should not be allowed to monitor student email and social media accounts unless acquiring permission from a court to gather such information and/or require students’ consent. Students’ email and social media accounts can be for personal use and personal use only, so they would want to ensure the privacy and ownership of their data. In addition, it is almost impossible to monitor and fully understand the popular slang that students post online - data cleaning can be huge and less meaningful for schools.

* What crosses the line between campus safety and invasion of privacy?

What crosses the line between campus safety and invasion of privacy can be students’ right to be informed. Schools are expected to be transparent and accountable for the data they collect from students. With a good intention to maintain a safe campus, universities should notify the students about the data monitoring, clearly describing how the data will be used and what parties are involved.

* Should Robert have been punished for cheating in class if he did not know his email was being monitored? What about his tutor?

Yes, Robert should be published for breaking the academic honor code even if he did not know his email was being monitored. The academic honor code emphasizes the actions that are taken, not how the actions are discovered. Moreover, the tutor should be published as well because he failed to maintain a fair learning environment for students and misguided the student by stealing intellectual property from the university.

Case Study 1: Loan Denied

* What ethically significant harms might Fred and Tamara have suffered as a result of their loan denial?

Fred and Tamara might suffer from a bias developed by the software system. Although Fred and Tamara are accomplished in their current career and have a strong desire/skill sets to own and make their business a success, the system might evaluate them as 'moderate-to-high risk,' possibly due to their race, religion and health records. The software system would make a biased decision if it only "learns" from specific demographics. If the system keeps learning from Fred and Tamara's denial cases, the more applications they submit, the less likely Fred and Tamara will get the loan.

* What sort of ethically significant benefits could come from banks using a big-data driven system to evaluate loan applications?

Using a big-data driven system can help banks improve the general efficiency in evaluating loan applications and be free of human mistakes. Since the software learns from big data, it gives banks a more comprehensive assessment of loan applicants than human approvers and, therefore, a lower risk for banks’ lending business. Higher efficiency and lower risk will lead to higher operating profits for banks and a more active lending market. Additionally, the software may link traits to people's repay behaviour and generate new evaluation patterns for future use.

* Beyond the impacts on Fred and Tamara’s lives, what broader harms to society could result from the widespread use of this particular loan evaluation process?

Such an evaluation process will reinforce certain stereotypes depending on the type of bias from the software system. The widespread use of this system can result in systematic discrimination against certain groups, leading to inequity of opportunity, social instability, economic slowdown and class immobility. If discrimination becomes new common sense, human civilization is probably in regression.

* Could these individual or societal harms have been anticipated by the software system’s designers? Should they have been anticipated, and why or why not?

Yes, these individual or societal harms could have been anticipated by the software system’s designers. The designers should evaluate the dataset the system uses to learn and recognize patterns and see if any loopholes or biases need to be manually fixed by human beings. Such evaluate process should be set regularly to ensure the best results out of the system.

Yes, they should have been anticipated. As discussed in the last question, widespread use of this system can cause severe negative impacts on our society, so designers are expected to mitigate such adverse effects and improve the system’s decision-making process.