The Nuts and Bolts of Creating a Two-Year Data Science Program

Data Ethics

Data Ethics – Course Learning Objectives

- Students will:
- 1. Define data ethics
- 2. Identify data ethics scenarios which require ethical decision making
- 3. Classify data into categories used by higher education institutions
  e.g. FERPA, SSN (PII), GDPR, regulatory
- 4. Review ethics guidelines for a selection of domains from various authoritative entities
  e.g. ASA, ACM
- 5. Examine ethical reasoning as a multi-step process in support of ethical decision making along a career or role-based trajectory (individual, instructor, mentor, and supervisor)
- 6. Create a career-specific or role-specific framework for making ethical decisions

Defining Data Ethics

- Data ethics is a new branch of ethics that studies and evaluates moral problems related to
  - data (generation, recording, curation, processing, dissemination, sharing and use),
  - algorithms (artificial intelligence, artificial agents, machine learning and robots) and
  - corresponding practices (responsible innovation, programming, hacking and professional codes)

  in order to formulate and support morally good solutions.[1]
- Ethics is a foundation for moral decisions; it "examines the rational justification for moral judgments" [2]
  - Morals: what decision is made
  - Ethics: why the decision was made

Defining Data Ethics

https://www.youtube.com/watch?v=l-k_1RQmmVY

Reasons for Learning About Data Ethics

- "The extensive use of increasingly more data—often personal, if not sensitive (big data)—and the growing reliance on algorithms to analyse them in order to shape choices and to make decisions (including machine learning, artificial intelligence and robotics), as well as the gradual reduction of human involvement or even oversight over many automatic processes, pose pressing issues of fairness, responsibility and respect of human rights, among others."[3]

- "Algorithms that are not developed with existing social factors in mind can easily end up reinforcing the discriminatory practices that they might reveal if examined by a discerning data scientist."[4]

Reasons for Learning About Data Ethics (cont)

- **Recent Cases Relating to Data Ethics and Poor Decision Making**

  - Volkswagen's gaming of emissions data
  - Whole Foods Markets manipulated product data, over-stating the weight of pre-packaged produce and meats.
  - Ashley Madison, a social network for married people seeking other partners, as hackers managed to extract a huge amount of private data from the company's servers.
  - General Motors was revealed to have hidden information about a faulty ignition switch linked to over one hundred deaths.
  - International Baccalaureate Organization use of a predictive algorithm to determine students' final grades and award diplomas[18]

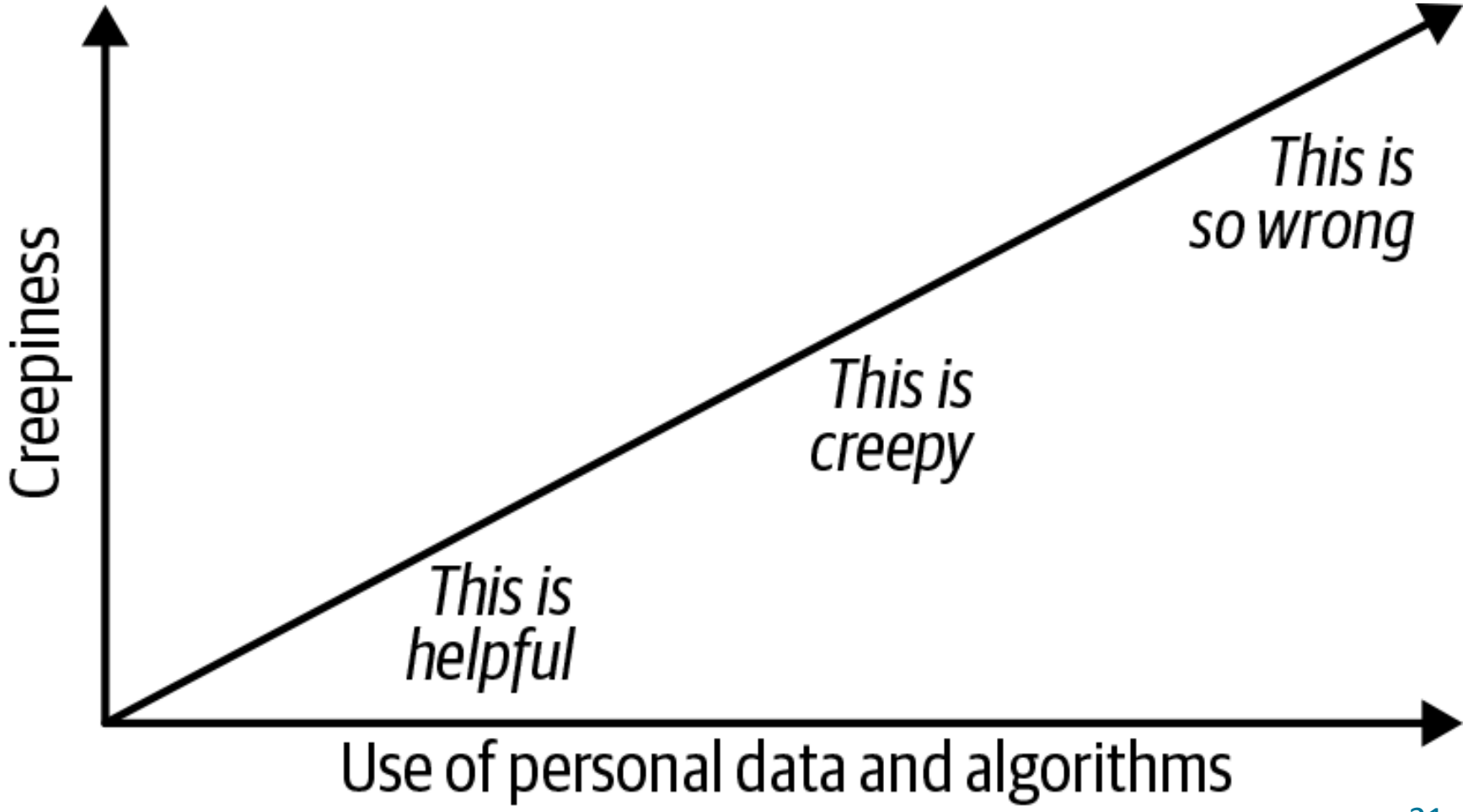
Addressing Ethical Challenges

- Overlooking ethical issues may prompt negative impact and social rejection. Social acceptability or, even better, social preferability must be the guiding principles for any data project with even a remote impact on human life, to ensure that opportunities will not be missed. On the other hand, overemphasizing the protection of individual rights in the wrong contexts may lead to regulations that are too rigid, and this can cripple the chances to harness the social value of data.[5]

Technology vs. Ethics

- Mathematical or theoretical statistics traditionally does not concern itself with the finer points of human behavior, and indeed many of us have only had limited training in the rules and regulations that pertain to data derived from human subjects. Yet inevitably in a data-rich world, technical developments cannot be divorced from the types of data sets we can collect and analyze, and how we can handle and store them.[6]

Data Ethics Scenarios and Case Studies



21

Data Ethics Scenarios and Case Studies

- Issues are not simple, there are few (if any) “right answers.” [7]
- For example, it’s easy to react against perceived paternalism in a medical application, but the purpose of such an application is to encourage patients to comply with their treatment program.
- It’s easy to object to monitoring students in a public school, but students are minors, and schools by nature handle a lot of private personal data.
- Where is the boundary between what is, and isn’t, acceptable?
- What’s important isn’t getting to the correct answer on any issue, but to make sure the issue is discussed and understood, and that we know what tradeoffs we are making.
- **What is important is that we get practice in discussing ethical issues and put that practice to work in our jobs.**

Scenario 1: Cellphone Tracking

- A college begins recording movement of students, faculty, and staff on campus by tracking the locations of their cellphones. The college hopes to answer the questions:
- What are the most frequently visited locations on campus
- What is the highest concentration of individuals at a particular location
- What days, and what times during the day, are associated with these high concentrations?
- Notices are emailed at the beginning of each semester and signs are posted throughout the campus that inform cellphone users that their movements will be tracked anonymously.
- To prevent tracking from occurring, cellphone users will have to power off or disable WiFi on their phones.

Scenario 2: Automated Healthcare App [2]

Consider a smartphone app designed to help adult onset diabetes patients.

Raises issues like paternalism, consent, and even language choices.

Is it OK to “nudge” patients toward more healthy behaviors?

Is it OK to automatically moderate the users’ discussion groups to emphasize scientifically accurate information?

How do we deal with minorities who don’t respond to treatment as well?

Could the problem be the language itself that is used to discuss treatment?

Scenario 3: Dynamic Sound Identification

Consider an application that can identify voices.

Raises issues about privacy, language, and gender.

How far should developers go in identifying potential harm that can be caused by an application?

What are acceptable error rates for an application that can potentially do harm?

Can a voice application handle people with different accents or dialects?

What responsibility do developers have when a small experimental tool is bought by a large corporation that wants to commercialize it?

Scenario 4: Optimizing Schools

Consider the problem of finding at-risk children in school systems.

Privacy and language are again an issue; it also raises the issue of how decisions to use data are made.

Who makes those decisions, and who needs to be informed about them?

What are the consequences when people find out how their data has been used?

And how do you interpret the results of an experiment?

Under what conditions can you say that a data experiment has really yielded improved educational results?

Categories of Data for Higher Education

- **Admissions/Enrollment**
  - Financial
  - Demographic
  - Secondary performance
- **Teaching and Learning**
  - Digital Content / Courseware Services
  - Online Discussion Forums
  - Online collaboration
- **Student Activity**
  - Early alert
  - Adaptive courseware: time spent on tasks/task performance/level of engagement
  - Behavioral
- **Student Outcomes**
  - Transcripts
  - Post-graduation tracking

Regulation of Data Dissemination

- Family Education Rights and Privacy Act (FERPA)
  (Privacy Act of 1974) https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html
  - Federal law that protects the privacy of student education records.
  - Education records may be inspected by eligible individuals
  - Corrections may be requested
  - Generally, schools must have written permission to disclose data
- Buckley Amendment: requires that schools provide an administrative process for parents to challenge and request information in their child's education records that they believe are misleading, inaccurate, or inappropriate.
- FERPA "serves the digital present about as well as a bicycle serves a kangaroo" [9]
  - US regulations were originally drafted with assumption of paper records

Regulation of Data Dissemination (cont)

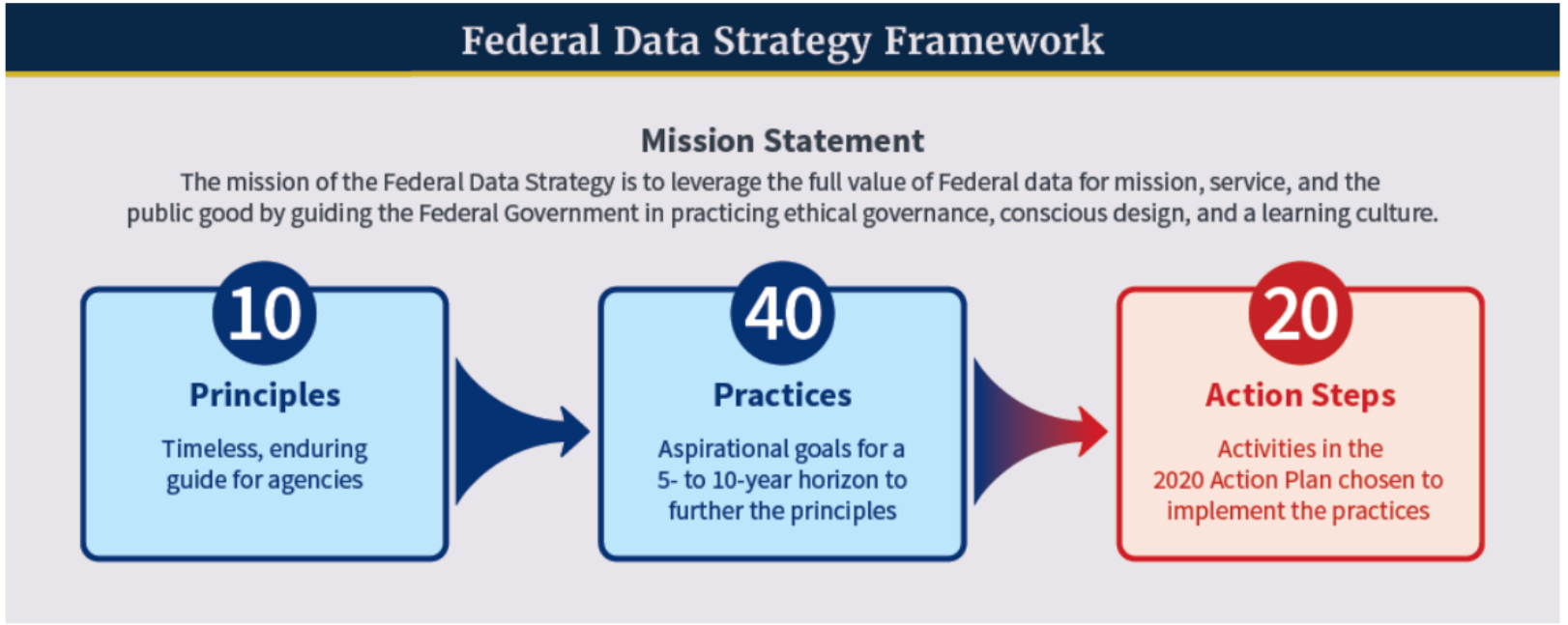
- General Data Protection Regulation (GDPR)
  - "applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe"[15]
    1. This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.
    2. This Regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data.
    3. The free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.

Regulation of Data Dissemination (cont)

- California Consumer Privacy Act (CCPA)
  - The California Consumer Privacy Act of 2018 (CCPA) gives consumers more control over the personal information that businesses collect about them. This landmark law secures new privacy rights for California consumers, including:
    - The right to know about the personal information a business collects about them and how it is used and shared;
    - The right to delete personal information collected from them (with some exceptions);
    - The right to opt-out of the sale of their personal information; and
    - The right to non-discrimination for exercising their CCPA rights.[16]

Regulation of Data Dissemination (cont)

- Federal Data Strategy (https://strategy.data.gov/)
  - The mission of the Federal Data Strategy is to fully leverage the value of federal data for mission, service, and the public good by guiding the Federal Government in practicing ethical governance, conscious design, and a learning culture.[17]



Institutional Review Board (FSCJ)

- The Institutional Review Board (IRB) serves to approve, oversee and administer all external research requests and activities conducted at the College in accordance with APM 10-1104.
- Specific guidance for such research requests and activities is provided in Florida State College at Jacksonville's Institutional Review Board Handbook.
- https://www.fscj.edu/discover/governance-administration/oiea/ie/institutional-review-board

Establishing Guidelines

- In most professions, practitioners define ethics: “Attorney-client privilege,” “protect and serve,” “seek the truth,” “do not reveal your source,” “serve the people” are all mottos that encapsulate a profession’s ethical code.[20]
- Data ethics guidelines can be derived from relevant excerpts from ethical codes provided by various professional organizations
  - American Statistical Association, "Ethical Guidelines for Statistical Practice"
  - Data Science Association, "Data Science Code of Professional Conduct"
  - Association for Computing Machinery (ACM), "The Software Engineering Code of Ethics and Professional Practice"
- Local guidelines can also be useful
  - Academic Advising Standards and Guidelines, Western Illinois University

Guidelines (Statisticians)

- "The ethical statistician:
  - 1. Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis.. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms.
  - 2. Reports the limitations of statistical inference and possible sources of error.
  - 3. In publications, reports, or testimony, identifies who is responsible for the statistical work if it would not otherwise be apparent.
  - 4. Reports the sources and assessed adequacy of the data, accounts for all data considered in a study, and explains the sample(s) actually used.
  - 5. Clearly and fully reports the steps taken to preserve data integrity and valid results.
  - ...
  - 10. To aid peer review and replication, shares the data used in the analyses whenever possible/allowable and exercises due caution to protect proprietary and confidential data, including all data that might inappropriately reveal respondent identities.
  - 11. Strives to promptly correct any errors discovered while producing the final report or after publication. As appropriate, disseminates the correction publicly or to others relying on the results." [10]

- - Ethical Guidelines for Statistical Practice

Guidelines (Data Scientists)

- "A data scientist shall protect all confidential information, regardless of its form or format, from the time of its creation or receipt until its authorized disposal."
- "A data scientist shall not knowingly:
  - …
  - (5) fail to rank the quality of data in a reasonable and understandable manner ...
  - (6) claim bad or uncertain data quality is good data quality;
  - (7) misuse bad or uncertain data quality to communicate a false reality or promote an illusion of understanding;
  - (8) fail to disclose any and all data science results or engage in cherry-picking;" [11]

•- Data Science Code of Professional Conduct

Guidelines (Software Developers)

- "Be careful to use only accurate data derived by ethical and lawful means, and use it only in ways properly authorized."
- "Maintain the integrity of data, being sensitive to outdated or flawed occurrences." [12]

- - The Software Engineering Code of Ethics and Professional Practice

Guidelines (Academic Advisors)

- "All advisors must ensure that confidentiality is maintained with respect to all communications and records considered confidential."
- "Unless the student gives written permission, information disclosed in individual advising sessions must remain confidential."
- "...all requirements of the Family Educational Rights and Privacy Act (Buckley Amendment) must be complied with and information contained in students' educational records must not be disclosed to third parties without appropriate consent..." [13]
- - Academic Advising Standards and Guidelines, Western Illinois University

Five Core Virtues for AIs and "Quants"[19]

1. Resilience

   Adapt to situations and recover quickly.

2. Humility

   Take responsibility for results.

   Continually learn and adapt with reinforcement learning.

3. Grit

   Avoid getting stuck admiring the problem or preoccupied with developing the smartest solutions.

4. Liberal Education

   Welcome and work with complexity, diversity, and change.

5. Empathy

   Recognize and account for social impact and the feelings of others.

Ethical Reasoning

- Data usage can have far-reaching implications, so consideration of stakeholders warrants more attention than is typical, including
  - description of how different individuals (stakeholders) may be affected by decisions and actions
  - enumeration of harms and benefits that are most clearly relevant for each stakeholder with respect to the activity
  - identification of guidelines which seem most relevant to this activity
- Stakeholder Analysis Template [14] (next slide)

Stakeholder Analysis Template

- Potential results (columns) capture those effects of a decision or action, summarizing them according to whether or not they may represent net negatives.
- Potential results (rows) must be considered with respect to each potential stakeholder

Potential result: Stakeholder:	HARM	BENEFIT	UNKNOWN	UNKNOWABLE
You				
Your boss/client				
Unknown individuals				
Employer				
Colleagues				
Profession				
Public/public trust				

Stakeholder Analysis: Example

- If faculty in a specific department share students, either during one semester or from one semester to the next, should they be able to discuss the academic progress of these students among themselves, especially in regards to questionable actions and activities regarding a student's academic integrity?

Potential result: Stakeholder:	HARM	BENEFIT	UNKNOWN	UNKNOWABLE
You	• potential for misidentification or libel	• confirm information • preserve academic integrity • increased awareness of methods of academic integrity violations	• violate student's trust	• impact on student success or retention
Other department faculty	• introduce prejudice or bias	• preserve academic integrity	• violate student's trust	• impact on student success or retention
Dean	• unaware of academic integrity issues			
Faculty in other departments	• indirect introduction of prejudice or bias		• non-academic relationships	• non-academic relationships
College		• preserve academic integrity and community reputation		
Students	• higher level of scrutiny	• held to higher standard		

Activity: Role-Specific Stakeholder Analysis

- Activity: Describe a Role-Specific Scenario Where Ethical Reasoning is Required and Create a Stakeholder Analysis

Activity: Design a Role-Specific Framework

- Use the guidelines provided in the previous slides as examples to design a framework to support ethical reasoning in your particular work role

- 1.
- 2.
- 3.
- 4.
- 5.

References/Additional Reading

- 1, 3, 5. Floridi L, Taddeo M. 2016. "What is data ethics?" Philisophical Transactions of the Royal Society A March 2016 Vol. 374, 20160360.
- 2. Oliver, Diane and Barbara Hioco. "An Ethical Decision-Making Framework for Community College Administrators." Community College Review. Jul 2012, Vol. 40 Issue 3, p240-254. 15p.
- 4. Burtch, Linda. "Ethics Must be a Cornerstone of the Data Science Curriculum". 97 Things About Ethics Everyone in Data Science Should Know. Franks, Bill (Ed), O'Reilly, 2020
- 6. Olhede, Sofia and Patrick J. Wolfe. "The Future of Statistics and Data Science". Statistics & Probability Letters Volume 136, May 2018, Pages 46-50. https://www.sciencedirect.com/science/article/pii/S0167715218300877
- 7. "Case Studies in Data Ethics", https://www.oreilly.com/content/case-studies-in-data-ethics/
- 8. "Dialogues on AI and Ethics", Princeton University, https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/
- 9. Kurzweil, Martin and Mitchell Stevens. "Setting the Table: Responsible Use of Student Data in Higher Education". EDUCAUSE Review, May 2018. https://er.educause.edu/articles/2018/5/setting-the-table-responsible-use-of-student-data-in-higher-education
- 10. American Statistical Association, "Ethical Guidelines for Statistical Practice". April 2018. https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

References/Additional Reading (cont)

- 11. Data Science Association, "Data Science Code of Professional Conduct". https://www.datascienceassn.org/code-of-conduct.html
- 12. Association for Computing Machinery (ACM), "The Software Engineering Code of Ethics and Professional Practice". https://ethics.acm.org/code-of-ethics/software-engineering-code/
- 13. Western Illinois University, "Academic Advising Standards and Guidelines", http://www.wiu.edu/advising/docs/standards-guidelines.pdf
- 14. Tractenberg, Rochelle. "Teaching and Learning about ethical practice: The case analysis.", SocArXiv 2019. https://www.academia.edu/38927905/Teaching_and_Learning_about_ethical_practice_The_case_analysis
- 15. General Data Protection Regulation (GDPR) https://gdpr-info.eu/
- 16. California Consumer Privacy Act (CCPA) https://oag.ca.gov/privacy/ccpa
- 17. Federal Data Strategy https://strategy.data.gov/background/
- 18. Evgeniou, Theodoro. "What Happens When AI is Used to Set Grades? Harvard Business Review, https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades
- 19. Burciaga, Aaron. "Five Core Virtues for Data Science and Artificial Intelligence". 97 Things About Ethics Everyone in Data Science Should Know. Franks, Bill (Ed), O'Reilly, 2020

References/Additional Reading (cont)

- 20. Schmidt, Eric. "First, Do No Harm". 97 Things About Ethics Everyone in Data Science Should Know. Franks, Bill (Ed), O'Reilly, 2020
- 21. Watson, Hugh. "Avoid the Wrong Part of the Creepiness Scale". 97 Things About Ethics Everyone in Data Science Should Know. Franks, Bill (Ed), O'Reilly, 2020

This material is based upon work supported by the National Science Foundation under Grant No. 1902524.