Human Research Protection Program
Institutional Review Board
Research Determination Form

Instructions:

To assess whether IRB review is necessary for a project, a determination must be made whether the project is research and, if so, whether it involves human subjects. An investigator conducting an activity with or about humans must make a request for a research determination through the IRB Protocol Management System (PMS).

All requests must include a detailed description of the activities and any supporting documents. Once complete, please upload this form as a Word or PDF document to the IRB Protocol Management System (PMS): https://secure.research.vt.edu/irb. A research determination official (either a designated departmental Human Subjects Advisor [HSA] or a Human Research Protection Program [HRPP] staff member) will review your completed request within 2-3 business days. If your project is determined to be not human subjects research (also called NHSR), HRPP will send you a memo that includes the IRB tracking number, which you can provide to journals and funding organizations upon request.

Definitions:

The federal regulations define 'research' and 'human subjects' as follows (please see SOP HRP-001 for the full regulatory language):

Research is defined as a systematic investigation (including research development, testing, and evaluation) designed to develop or contribute to generalizable knowledge.

A ***human subject*** is defined as a living individual about whom an investigator either:
1) Obtains information through intervention or interaction with the individual, or
2) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.

Outcomes and Next Steps:

If the activity does not meet the definitions of research and/or human subjects, the determination official will issue a "Not Research" or "Not Human Subjects Research" letter and send it to the investigator through the IRB PMS. The investigator can begin the activities upon receipt of this letter.

If the activity does meet the definition of human subjects research, the submission will be returned to the PI with further instructions about how to submit a research protocol for IRB review.

Submissions that do not contain adequate details or information will be returned to the investigator for revision. If you have any questions, you can email us at irb@vt.edu.

* Questions with an asterisk require a response. To avoid delays, please ensure that your submission is complete.

Section 1: General information

1.1 Project title: *

What visualization and documentation strategies do data scientists use during machine-learning projects? A case of data science notebooks on Kaggle.com

1.2 Principal Investigator (Name): *

Chris Lindgren

1.3 Funding information: *

None

1.4 Is this a collaborative project?*

☒ No

☐ **Yes – check <u>all</u> the activities Virginia Tech will be involved in with the collaborating institution:**

- ☐ Research design/development
- ☐ Recruitment or dissemination of recruitment materials
- ☐ Consenting participants
- ☐ Data collection
- ☐ Analysis of **<u>identifiable</u>** data or information
- ☐ Analysis of **<u>de-identified</u>** data or information
- ☐ Consultation or manuscript writing

1.5 Is this activity being conducted by a student to meet course work or graduation requirements? *

☒ **No**

☐ **Yes,** please check all that apply.

- ☐ Thesis
- ☐ Dissertation
- ☐ Class assignment/routine coursework
- ☐ Other, please explain below:

Section 2: Is this activity research, as defined by the IRB regulations?

2.1 Is this activity a systematic investigation? *

(*Systematic: Having or involving a system, method, or plan. Investigation: A searching inquiry for facts; detailed or careful examination. A systematic investigation is usually recognized by the fact that there is a predetermined and organized method [of data collection and analysis] to study a specific topic, answer a specific question, test a hypothesis, or develop a theory.*)

☐ **No**

☒ **Yes**

2.2 Is this activity designed to develop or contribute to generalizable knowledge? *

(*Generalizable: Universally or widely applicable. Relates to drawing general conclusions, informing policy, or generalizing findings beyond a single individual or an internal program. Note that publishing or presenting the data is not a sole criterion on which to define generalizable knowledge – non-generalizable knowledge is often published or presented, often as case studies.)*

☐ **No**

☒ **Yes**

2.3 Describe the purpose and specific aims or objectives of the project. If your planned activity relates to a protocol previously approved by the Virginia Tech HRPP/IRB, please include the Virginia Tech IRB number(s). *

Please provide a detailed description that includes the purpose or goal of the project, objectives, procedures used to gather information (interviews, surveys, focus groups, etc.), target population, and description of data/samples gathered (datasets, URLS, etc.) If there are procedures that the research team is thinking about implementing, but is unsure if they will, a new determination should be submitted at a later time.

Goal: This study extends previous research (Berger, 2020; Bhat et al., 2022; Chang & Custis, 2022; Hutchinson et al., 2021; Mitchell et al., 2019; Wang et al., 2021) to discern what documentation practices data scientists use to communicate their machine-learning (ML) process. This study contributes a closer investigation about the relationship between what text and data visualizations data scientists use to communicate their process and build their ML models.

Main Research Questions:

1. What types of data visualizations do data scientists use, when they build their ML models?
2. During what part of the life-cycle (Wang *et al*., 2021) are the visuals used?
3. What text, if any, document their use and understanding of the visualization that is created?

Target artifacts: Publicly available computational notebooks (kernels) that develop ML models on Kaggle.com.

Method: Data collection will include two main steps:

1) Use Kaggle.com's public Application Programming Interface (API) to collect a corpus (n=1000) of the top-voted computational notebooks that are categorized by the website as being ML projects.
2) Use a custom web-scraping script to collect a few more status criteria about each kernel notebook.

STEP 1. Kaggle's API.

I will use the Python programming language to use Kaggle's public API to collect the "kernel" notebooks (object of study) with the following commands:

1. kaggle kernels list -s [KEYWORD]: list Notebooks matching a search term
2. kaggle kernels pull [KERNEL URL PATH] -p /path/to/download -m: download code files and metadata associated with a Notebook
3. kaggle datasets download -d [DATASET URL PATH]: download files associated with a dataset

The data collected includes all of the project "kernel" files for the analysis. The metadata includes the following columns:

- "id"
- "id_no"
- "title"
- "code_file"
- "language"
- "kernel_type"
- "is_private"
- "enable_gpu"
- "enable_internet"
- "keywords"
- "dataset_sources
- "kernel_sources"
- "competition_sources"

The data includes the public username in the kernel's URL path, but this username is not used in the study and is not typically the actual name of the user. This study focuses only on data about the notebook "kernels" as the object of study.

STEP 2. Custom Web-Scraping Script.

I will use the Python programming language to scrape the following information about each kernel to help with the analysis:

1. Kernel Name: Info will help me to link the other datapoints (3-5) back to the original data from step 1.
2. Kernel Path: Info will help me to link the other datapoints (3-5) back to the original data from step 1.
3. Vote Count: Total sum of votes by users per kernel notebook.
4. Medal: Notebook kernels can receive a Bronze, Silver, or Gold status medal
5. Comment count: The sum of user comments per kernel notebook.

This data only includes information about the kernel notebook to help with my analysis. These status markers will help me organize the data into groups based on these quantitative engagement features.

References

Berger, G. (2020). 2020 U.S. Emerging Jobs Report. *LinkedIn*. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf

Bhat, A., Coursey, A., Hu, G., Li, S., Nahar, N., Zhou, S., Kästner, C., & Guo, J. L. C. (2022). Aspirations and Practice of Model Documentation: Moving the Needle with Nudging and Traceability (arXiv:2204.06425). arXiv. http://arxiv.org/abs/2204.06425

Chang, J., & Custis, C. (2022). Understanding Implementation Challenges in Machine Learning Documentation. Equity and Access in Algorithms, Mechanisms, and Optimization, 1–8. https://doi.org/10.1145/3551624.3555301

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 560–575. https://doi.org/10.1145/3442188.3445918

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Wang, A. Y., Wang, D., Drozdal, J., Liu, X., Park, S., Oney, S., & Brooks, C. (2021). What Makes a Well-Documented Notebook? A Case Study of Data Scientists' Documentation Practices in Kaggle. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10.1145/3411763.3451617

2.4 Please describe how the results will be used. *

The results will be written up as manuscripts to be published in academic journals. The results will also potentially presented at academic and related industry events. All results reported will not use or be based on users. Results are focused on the kernel notebooks and any shared content will be representative examples from the kernels without usernames.

Section 3: Does this activity involve human subjects?

3.1 Will you interact, intervene, or observe individuals and collect information (or biospecimens) about them? *
(Intervene: Physical procedures or manipulations of individuals or their environment for research purposes. Interact: Communication or interpersonal contact with the individuals.)

☒ **No,** the information being collected is **NOT** about the individual(s) (e.g., it is about programs, policies, and/or practices that the individual(s) are familiar with).
☐ **Yes,** the information being collected **IS** about the individual(s) (e.g., their own personal thoughts, opinions, attitudes, and/or perception)

3.2 Will you obtain or view any of the following identifiers from or about a living person (from any source or already in your possession)? *
Please review the list in its entirety and check all that apply.

[redacted]

☐ Name
☐ Geographical subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the first three digits of a zip code
☐ Elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death, and single year of age over 89 as well as all elements of dates (including year) indicative of such age, unless such ages and elements are aggregated into a single category of age 90 or older
☐ Phone numbers
☐ Fax numbers
☐ Electronic mail addresses (e-mail)
☐ Social Security numbers
☐ Medical record numbers
☐ Health plan beneficiary numbers
☐ Account numbers
☐ Certificate/license numbers
☐ Vehicle identifiers and serial numbers, including license plate numbers
☐ Device identifiers and serial numbers
☒ Web universal resource locators (URLs)
☐ Internet Protocol (IP) addresses
☐ Biometric identifiers, including finger and voice prints (audio recording)
☐ Full face photographic or video images and any comparable images (video recording)
☐ Student record number or identification/user name
☐ Student grades or class assignments
☒ Username for online or computer accounts
☐ Any other number, characteristic, or code that uniquely identifies an individual (note this does not mean the unique code assigned by the investigator to code the data). Explain:

☐ Other identifiable information. Explain:

3.3 Will you gather public and/or private data (check all that apply)?

☐ Data are about behaviors that occur in a context in which an individual can reasonably expect privacy, for example that no observation or recording is taking place (private information).

☐ Data were collected for specific purposes in which individuals can reasonably expect that they will NOT be made public, such as student records and medical records (private information).

☒ Data consist of publicly available information, such as news stories, a public-use dataset, or other information accessible to everyone (public information).

3.4 Will you video or audio record or photograph individuals during activities? *

☒ **No**
☐ **Yes**

3.5 Will you generate identifiable private information or identifiable biospecimens by combining data sources? *
(Can the investigator readily ascertain an individual's identity by combining available datasets and/or biospecimens?)

☒ **No**
☐ **Yes,** answer question within the table 

IF YES
3.5.a Briefly describe what data sources you will use and from whom you will obtain them:

3.6 Will this project involve only the use of existing de-identified data or biospecimens? *

☒ **No,** go to Section 4
☐ **Yes**

3.6.a Were the de-identified data or biospecimens collected specifically for this study?

☐ **No**
☐ **Yes**

3.6.b Will anyone on the research team be able to readily identify individuals to whom the data or specimens pertain or belong? This includes anyone affiliated with the planned activity who has access to a linkage file or key.

☐ **No**
☐ **Yes**

3.7 Does this activity involve a drug or device? *

☐ **No**, go to Section 4
☐ **Yes**, check all that apply

☐The project involves the use of a drug in one or more persons other than use of an approved drug in the course of medical practice.

☐The project involves the use of a device in one or more persons that evaluates the safety or effectiveness of the device.

☐The project involves data about subjects or control subjects submitted to or held for inspection by FDA.

☐The project involves data about the use of a device on human specimens (identified or unidentified) submitted to or held for inspection by FDA.

Section 4: Supporting Information

4.1 Please select below all applicable documents related to this activity. Please upload a copy of the document(s) (section 4, supporting docs) when you submit this form to IRB PMS. These documents will help us evaluate whether your activity needs HRPP or IRB review. *
If you do not currently have these documents, you can include a brief overview or type(s) of information that will be included.

- ☐ **Grant**
- ☐ **Proposal**
- ☐ **Contract**
- ☐ **Statement of work**
- ☐ **Survey/questionnaire**
- ☐ **Interview/focus group guide**
- ☐ **Observation data sheet**
- ☐ **Other**, please specify**:**

Section 5: Additional information

5.1 Please provide additional information or instructions that might assist with this review:

Proposed modifications to activities must be reviewed by the HSA/HRPP reviewer prior to implementation.

Do not begin activities until you receive an HSA/HRPP determination letter via email.

----------END----------