Human Centered Data Science

DATA 512 — Oliver Keyes

User experience and big data | Week 8 | November 16, 2017

Overview of the day

- CTL MSDS program evaluation
- *Dinner break (15 min)*
- Week 7 reading reflections
- Initial feedback on final project plans
- In-class activity: Final project peer review session
- Homework for next week(s)

MSDS Program Evaluation

Katie Malcolm, PhD
Instructional Consultant,
UW Center for Teaching & Learning

MSDS Program Evaluation

1. Katie is from the Center for Teaching & Learning (CTL), and in addition to providing teaching support across UW, CTL gathers feedback from students in departments & programs to get students' anonymous input
2. We're using valuable class time to gather this feedback because it's the one time that we are all together
3. Your input will be invaluable as we work on strengthening the program for you and future students
4. Everything you share in the session will be anonymous

Break (15 minutes)

Week 7 reflections

Week 7 reading reflections

Andrew:

What's the alternative, in practice and from a pragmatic perspective, to AMT-derived datasets? Put differently, the paper shows that datasets from AMT workers are and perform differently from datasets from academics and subject matter experts. But it doesn't consider this from a cost-benefit perspective. For example, how much cheaper/easier/more practical is it to get a dataset from AMT than from high-paid researchers? Even if algorithms don't perform as well with AMT data, do they perform so much worse that it outweighs the cheapness/easyness of getting data from AMT?

Week 7 reading reflections

Todd:

While the goal and purpose of the social action may be just and needed, the actions documented in the article were neither unbiased nor unobtrusive. The authors were calling for and supporting action to help change the situation of Turk workers. The paper...ignored any notion of a control group.

Review of A3: Project plan

Most of these are awesome!

- Well motivated: references to prior work, potential social impacts, why this study matters to someone other than you
- Detailed descriptions of analytical plans (and *usually* good motivation for the type of methodology proposed).
- Thoughtful discussions of potential limitations and confounding factors
- Some excellent explicit call-outs of HCDS considerations

We need to talk about licensing

A lot of you specified the data license or terms of use for the data, but unfortunately not everyone did.

You can only use a dataset for your project if the license or terms of use allow you to collect the data, analyze it, and re-publish it publicly.

- Some licenses and terms of use specifically prohibit that.
- Some TOU say it's okay for non-commercial purposes (like academic research).
- Some data sources don't specify a license *or* terms of use for their data (hint: avoid these).

We need to talk about licensing

It is up to you to demonstrate that you can use your data.

- If the dataset has a compatible license... you're good!
- If the terms of use say you can download the data for academic purposes and re-publish it (even for a short period of time, as is the case with Yelp)... you're good!
- If you have received *explicit, written permission* from the copyright holder to use the data… you're good!
- If the data is provided publicly by the United States government, it is public domain, so... you're good!
    - (but you still need to specify the source and note that it is public domain)

Don’t use data if you aren’t sure

You will receive a 0 on the final project if you do not demonstrate in your report that you can legally use your data.

Understanding basic ethical and legal best practices for data re-use is part of the course. We really, really, don’t want to fail you.

If you have questions, *please ask us.*

If you need help picking a new dataset, or a new research question, because the data you picked for your proposal isn’t legally available, *please ask us*

- you don’t need to submit a new proposal, but it would be a good idea to run your ideas by us!

How to document your data

When your dataset has an explicit license

1. State the license of your data (e.g. “CC-By-SA 4.0”) in your report.
2. When possible, link to the license deed, e.g. https://creativecommons.org/licenses/by-sa/4.0/

When data re-use is covered under the provider’s Terms of Use

1. Quote the relevant section of the terms of use in your report
2. Link to the terms of use page

If possible, link to the *original source* of the data, which may be different from where you found it.

- E.g. MovieLens data on the GroupLens website vs. MovieLens data on Kaggle

Be careful with Kaggle data

Some of those datasets are not explicitly licensed. If you cannot find appropriate license information for the data, you cannot use it. You'll fail the assignment, even if the rest of your work is really good. :/

Many of those datasets have already been analyzed by other Kagglers. Many of those analyses are public on Kaggle.com.

- Your analysis should not simply duplicate analysis that Kagglers have already done on this dataset (e.g. "do more data scientists use Python or R?").
- It's perfectly fine to build off of the analysis that others have done, just make sure you cite the original analysis.
- Tip: Avoid even *looking like you might be* plagiarizing someone else's analysis.

Talking about gender

Many projects are centred on gender as a variable. Since the assignment requires you to think about ethics, consider what ‘gender’ is:

- Not the same as sex (‘man’, ‘woman’ not ‘male’, ‘female’)
- Not a binary (people have identities other than ‘man’ or ‘woman’)
- Not impermeable (people transition between genders)
- Something that creates myriad different experiences of life that are likely to be reflected in your data.

Talking about gender

Questions to ask:

- Does your data only contain binary options? If so, note that as a limitation - and note why it is a limitation (it excludes people outside that binary).
- Does your data include trans people? If so, incorporate that into the gender research. If not, highlight *that* as a limitation.
- Please stop saying ‘females’.

In-Class Activity

Lightning Peer Review

Graded, ***Individual,*** 45 minutes

Final project peer review

Goal: give and receive feedback on final projects **with two classmates**

- Pair up with a classmate—someone you haven't discussed your project with
- Over the course of 20 minutes, you should each:
  - Describe your project plan to your classmate (~5 minutes)
  - Write down your classmate's feedback on your project (~5 minutes)
- Feedback can be about anything relevant to the project, for example...
  - suggestions of research questions/hypotheses, questions about methods, limitations, etc.
  - suggestions about how to motivate the research and/or tie it to Human Centered Data Science, how to present your study or describe your results in the report.

After you have both given and received feedback, find another classmate and repeat the process. *Remember: keep feedback constructive!*

Final project peer review

- **Submit at least three pieces of feedback you received from *each* of the two classmates you talked to** in Canvas under ‘week 8 in-class activity’
- For each piece of feedback, please describe how you plan to address it (or, why you do *not* plan to address it)

Make sure to include the full names of the classmate you received feedback from in your Canvas post.

Homework

Homework due next week

Readings (read both, reflect on one)

- Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. *User perception of differences in recommender algorithms.* In Proceedings of the 8th ACM Conference on Recommender systems (RecSys '14)
- Chen, N., Brooks, M., Kocielnik, R., Hong, R., Smith, J., Lin, S., Qu, Z., Aragon, C. *Lariat: A visual analytics tool for social media researchers to explore Twitter datasets.* Proceedings of the 50th Hawaii International Conference on System Sciences (2017)

These papers both focus on *design considerations* related to human-centered data science, which was supposed to be the focus of this week's lecture. Instead, we'll discuss these examples, and other design-related concepts and case studies, in class on November 30th.

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)#Week_8_November_16

Homework due next week

Assignment 4: Crowdwork Ethnography

- **Format**: Google Doc, shared with Jonathan and Oliver, link submitted to Canvas
- **Due date:** November 23, by 5pm

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)/Assignments#A4:_Crowdwork_ethnography

Homework due in two weeks

No class next week!

...*But* there will still be a reading reflection assigned, which will be due in two weeks (November 30):

Hill, B. M., Dailey, D., Guy, R. T., Lewis, B., Matsuzaki, M., & Morgan, J. T. (2017). Democratizing Data Science: The Community Data Science Workshops and Classes. In N. Jullien, S. A. Matei, & S. P. Goggins (Eds.), *Big Data Factories: Scientific Collaborative approaches for virtual community data collection, repurposing, recombining, and dissemination*.

See: https://wiki.communitydata.cc/HCDS_(Fall_2017)#Week_9_November_23