Aug. 3, 2021

- Will be finished with sections I have to do by EOD Wednesday
    - When does Charles want the writing token?

July 23, 2021

Updates

- I wrote most of the results section before June 25. But what I wrote was really bad when I revisited it.
    - I'm slowly revising this section in Overleaf, and it's turning out much better.
- CHI 2021 submission details are up
    - Abstract deadline: Thursday September 2, 2021
    - Submission deadline: Thursday September 9, 2021
- I feel like it'll take 2 weeks to get a rough draft out. So can we pass the writing token to Charles around Aug. 6, then Tamara Aug.16?
- Tamara: sounds like a good plan from my side!

June 25 2021

- Parental Leave approved, beginning in September.
- Baby arrived today

June 18, 2021

Updates

- ☑ ~~Reach out to Marjan about what to do concerning NSERC parental support.~~
- ☑ ~~Change issues to challenges, but keep tasks the same.~~
- ☑ ~~Update *Misc.* codes into *Transform*.~~

Parental Leave

- Heard back from Melanie Longpre-Lateigne from NSERC. She says I need to contact the EI offices.
- Updated leave calendar leave-plan-2.jpg
- It seems like the only option is:
  - Unofficially go on leave *x* weeks between when the baby arrives and when Winter Term I 2021 begins
  - Take Leave of Absence in September.
  - Unofficially return at the most *x* weeks before the 40 weeks Parental Leave expires.

Interview Study

Interim draft: Interview Study Rough Draft

Progress

- Related work ✅
  - Interviews with data scientists ✅
  - Interviews with journalists ✅
  - Indirect observation studies ✅
- Methods ✅
  - Participants ✅
    - RE: Isabel's point at MUX: Does everyone know what I mean by traditional media organizations?
    - Divided into three profiles: practitioners, educators, and tool builders
      - This doesn't feel like results, just a way of thinking about our participants
      - Actually, move to results?? And call them roles.
  - Data gathering ✅
  - Analysis ✅
    - Still need to write Reflective Synthesis section (TODO)

Questions

- Which document template to use?
  - CHI Publication Formats are currently TBD but CHI 2021 author formats use ACM official primary articles template.
  - If Tamara authors the document, can we use Overleaf Premium features again?
  - Use this one: ACM Conference Proceedings Primary Article Template.

Task List Interview Study Overview Thesis Proposal Timeline Task Summaries Interview Study Pre-paper

Schedule

- Group meeting now scheduled for Thu 8 Jul. Charles will be away but can join from remote. (Charles gone July 7-10.)
- Goal: Steve finishes first draft in one week, for Jun 25 mtg, with sending to TM/CB by early Fri morning so they can read before meeting (or let us know if not). Play be ear on whether CB or TMM gets token then, but should go to group by end of Tue Jul 6 at latest.

June 11, 2021

- Administrative stuff: UBC, EI, NSERC
    - UBC parental leave can't start until Sep 1, formally.
    - UBC parental accommodation to bridge gap from mid-July to end of August (6 weeks).
        - Also gap in other direction, late March to end of May (5 week).
        - That's what this is for, the numbers work out.
    - Gap and RoE
        - Will that suffice for RoE needed with EI? Unclear verbal answer from Payroll on phone, she said happens automatically when GRA appointment is terminated.But is that at start of leave or at start of accommodation?
        - Steve asked Joyce. Haven't heard back.
        - But must apply for EI within four weeks of baby arriving. WTF. How to get UBC to make this paperwork possible.
        - Aha. Should apply for parental benefits within four weeks of last day of work, not within four weeks of baby arriving.
    - What about Marjan stuff?
        - Steve has not heard from NSERC. At this point has sent many emails and phoned many times. Still no answer.
        - Time to connect with Marjan again. Tell no answer from NSERC so suggest we just do the form. ~~Ask her if she knows about RoE situation wrt EI and July-Sep gap.~~
        - NSERC parental pays *difference* between normal and EI.
        - How do I get paid NSERC topup through UBC if I'm on leave from UBC which I have to be in order to get EI? Don't I have to still be employed by UBC? But I need to not be employed by UBC for EI. Very confused.
    - Major timesink and source of stress
        - Many nonhelpful interactions. International student, g&ps. All send links to parental accommodation web site which doesn't help.
- Mapped Table Scraps actions and processes into Task Summaries.
    - More materials:
        - Auxiliary document of codes that do not fit.
        - Reverse match between Table Scraps and Interview Study
    - Takeaway:
        - Interview study taxonomy mostly describes actions but also crosscuts into processes
        - Interview study taxonomy is higher-level than actions and processes. I don't think interviews would have been effective at the low level of Table Scraps. Probably direct-observation would be the way to triangulate at this level.
    - Discussion: mismatches

- Impute missing data. JSV 80% quote, maybe belongs in paper, when discuss 'impute missing data'. CB: workaround for fact that general public doesn't understand probability and stats. So expert imputing would be appeal to authority, but can't do it yourself. Point to make: True difference in data science and journalism.
- Make incorrect conclusion. Even JSV didn't talk about in interview. Interesting evidence of why need to triangulate, interview alone doesn't suffice compared to artifact studies. Point to make: Triangulation FTW.
- Divide and conquer. Too low-level, it's how not why - less high level than it sounds. Move.
- Generate computationally. Interesting that not mentioned. Point to make: Another difference w/ data science? Very interesting difference on willingness to model vs say e-commerce. Social science question to truly answer why such a discrepancy of goals and methods exists… Different rewards structures and incentives. Explain to boss and help with decision-making for future actions (sales shortfall) - narrow audience with specific value (make decision). Public rarely is in position to have such narrow interest and power to act. Journalists rarely have luxury of time to spend weeks/months/years modelling.
    - Ana Baton categories: descriptive, diagnostic (root causes), predictive, prescriptive
    - Descriptive common in journalism. Predictive maybe mostly for elections.
- Understand (few) differences in order to better exploit (many) similiarities (microcosm argument)! Journalism as Galapagos islands of data analysis. Wingless flies (strong winds), can ask why things are a certain way - narrow but revealing case.

- Got Task Summaries up-to-date
    - Update tasks

Updated issue descriptions

June 4, 2021

- Review feedback
  - Pre-paper Talk Feedback
  - Interview Study Pre-paper
- Spend two days comparing TS to interview with a color coding exercise
  - Do processes map to tasks and issues?
  - CB: A task = action/process + issue
    - Issues provide underlying context that was missing from code

- [x] ~~Reorganize Tasks so that the ones that don't have issues are at the end.~~
- [x] ~~Validate issues, [team size (Workplace issues) is the issue]~~
- [x] ~~CB: Validate gets moved into communicate b/c of peer review~~
- [x] ~~Define needs → plan~~
  - [x] ~~Include: estimate utility, identify constraints, and *validate* correctness~~

June 1, 2021

Finished TODO items

- ☑ ~~Change *affected data* to *non-normalized data*~~
- ☑ ~~Change *normalize* axial code to *standardize*~~
- ☑ ~~Change *detrend* to *normalize*~~
- ☑ ~~Add overview of issues~~
- ☑ ~~Add overview slides of all tasks~~
- ☑ ~~Update validate slides~~
- ☑ ~~Add CPI PPP x Layoff data~~
- ☑ ~~Two case studies~~

Other changes

- Swapped 10,000 datasets wrangle with investigation on police shootings to highlight *ontologizing*
- Pivoted organization of discussion section
    - Previous:
        - Categorical wrangling
        - Improvisational wranglers
        - Limits to data wrangling
    - Currently
        - Comparison to Table Scraps
        - Comparison to Data Science
            - Categorical wrangling
            - Improvisational wranglers

Run-through pre-paper slides

- Section on participants feels thin
- There's not much about tools
-

May 28, 2021

PhD annual report thing

- Materials
  - A-OK?

Pre-paper talk update

- Create task and issues taxonomy/taxonomies changelog
  - Starting from end of Stage 2 (April 15 to present)
- Naming
  - Gulf of evaluation back to ambiguous results?
  - Affected data?? Change to Trended data? Or does that mean something else. "Trend-affected data"? Non-normalized data?
    - Rename detrend to normalize task, and then change axial code from normalize to ??. No.
    - Non-standardized? Non-regularized?
    - Change Normalize axial code to Standardize, change Detrending to Normalizing, change Affected to Non-normalized?
  - Misc
    - Merged duplicate observations into inconsistent data?
    - Change to workplace issues? And merge inconsistent schema into inconsistent schema. Move different resolutions to integrate?
- Task vs Issues
  - Task is done by human, issue is state/properties of data. Issues adjectives, tasks verbs.
- Add contributions to pre-paper slide deck
  - Come back to question of whether case studies are contributions in themselves, or support other contributions? Maybe they really are.
- Add results to pre-paper slide deck
  - Todo: add issues taxonomy slide early, but still walk through based on tasks.

Case studies

- Illustrate how tasks are interconnected
- Potential case studies:
  - Paul Bradshaw integrating 10,000 datasets of police data
    - Consolidating datasets temporally and geographically.
  - Joe Yerardi's PPP and Layoff data
    - Example of data mashup, creating keys

May 21, 2021

- Trying to figure way out of the weeds

May 14, 2021

Updates

- Progress on pre-paper slides
  - Current status: incomplete
- Rough-draft/rapid prototype/blueprint of DFP Showcase poster
  - Slammed in content from pre-paper talk
    - What content should be included?
  - I will layout it out in HTML for showcase

Questions

- Who are my participants?
  - Two levels: who we recruited vs what are their personas/archetypes?
  - Is there time / room for some persona analysis?
  - How do I stratify participants?
    - How do you lookup what's a small, medium, large newspaper
    - But what about journalists in academia and freelancers?
- IEEE Visualization Workshop on Human-Data Interaction
  - This seems interesting
  - Interview Study work may fit into Case studies and practical HDI lessons from the trenches

May 7, 2021

Updates

- *Manipulate schema* seems too broad a category, almost synonymous with wrangle data
- *integrate data* to *wrangle data*
    - Should it be a top-level category?
        - Integrate is a complicated subject, depth = 4 instead of 3
- Define needs
- Analyze/interview data
    - It is very difficult to describe journalistic analysis in data science terminology.
    - What does *interviewing the data* mean? Can't seem to find a consistent definition for this term
        - A method for data cleaning, LaFleur & Donald; NICAR, 2013
        - A method for analyzing data?
            - "You should treat your datasets like human sources. Think about analyzing your dataset like you interview human sources. Ask questions of your data. What's the most interesting information they can tell? Also keeps a list questions he asks the data." [Yerardi]
- Profile data
- Communicate data, larger axial changes made
    - New sub categories:
        - *Produce information:* Include sharing work and workflows with others
            - E.g. document work, disseminate work
        - *Consume information:* Reaching out to other people when working with data
            - E.g. contacting sources, colleagues, etc.
- Validate

TImeline

- Annual report?
- June 2, 2-3 pm, "pre-paper" talk to MUX, InfoVis follow up on June 3
    - For position and framing of HCI papers for CHI

April 30, 2021

Progress update

- Summaries for wrangling codes

Discussion topics

- Is resolving entities different from ontologizing data?
  - Entity resolution, there are these entities, ontologies are not necessarily entities, wildfire vs non-wildfire is not an entity it's an ontological category. There are technical similarities, but from the data there's a difference in the data. There's a scale of the number of levels that's way larger in ER than ontologizing.
    - Can change ontology term, maybe taxonomy/typology/categorization/classification is better. Should look into this more.
- What is MacGyvering?
  - Previous definition: anytime users employ heuristics techniques to accomplish a task, meaning they are not optimal, rational, or perfect, but are nonetheless sufficient.
  - Is this a theme that is just related to the fluid use of tools?
  - Maybe we shouldn't use the term MacGyvering, synonyms: improvisational wrangling
    - Precarious, play-it-by-ear
  - Think about this idea in the context of abandoned projects. Does MacGyvering always imply success.
  - macguyvering with tools
  - - use what gets the job done, nondogmatic about tools
  - - satisficing not perfectionism: suboptimal
  - - fluidity of tool use
  - - hacky, good enough
  - - heuristic
  - - quick and dirty
  - macguyvering with non-tools
  - - soft keys. slug-ify names and cut off foreign chars and take first
  - 10 letters...
  -
  - improvisation? not even necessarily "suboptimal"
  -
  -
- Should Integrate be part of wrangle, part of gather or its own thing?
- Schema vs layout:
  - Does the word send people down the wrong path?
  - Flag as a point to consider.
- Aggregation
  - Aggregation is a means to the end of checking your data.
    - Confirm transformation effects, improving.
      - Filter checks using tools not what it was intended to use.
      - Flag as makes Tamara twitchy. It seems weird that these are linked, but this may be a finding.
- Second-Pass Analysis
  - Task 1 vs task 2 confusing. Task 2 relates to either Task 1 or Issues. Not primary/secondary. Maybe independent/dependent??
  - Rename Tasks to Actions???

- Consider whether these are different things, should these also be called actions because that's what we called them in TS?

Parental leave

- I am eligible for NSERC parental leave supplement?
- Can I be on pat leave while Shalaya is on mat leave?
- Include parental leave plan dates, full details of plan. Include picture

Draft email for HR:

I'm a PhD candidate in computer science, and I'm expecting the birth of my first child this July. I'm currently being supported through my supervisor's (Tamara Munzner) NSERC fund (speedchart EJSE). For the last six months, I've been trying to understand how to coordinate NSERC supplemental leave with parental leave EI and UBC's policies and procedures. After consulting with finance (Sonia Penflor of UBC and Marjan Molouk-Zadeh of UBC CS), they said that you might be able to help me resolve all of my remaining questions.

Background: The baby is due Jul 21. My wife is starting mat leave early because of prenatal complications, on May 2. Our hope is to have my wife be on mat leave May 2 - Aug 14 and parental leave Aug 15-Sep 4; I would be on parental leave Jul 11 2021 - Mar 26 2022. This plan uses up all of the 55 weeks of EI that we are jointly eligible for (15 maternal and 40 parental). Only I am eligible for NSERC parental leave because my wife is not paid through NSERC.

Questions for you:

1. Can you confirm that I am eligible to apply for NSERC supplemental support while receiving EI and that the amount I will receive is the full difference between EI and my regular earnings?

2. Can I receive NSERC supplement before I start the parental leave from EI on July 11? I've heard that NSERC supplemental support can't be received until the baby arrives, but I was told to consult with HR on UBC policies and procedures about coordinating these two.

3. Am I allowed to receive NSERC parental leave supplement at the same time that my wife is also on Mat leave and receiving EI? I would like to have an overlap of five weeks where we are both on leave (Jul 11 - Aug 14). If any overlap is allowed, is there any limit on the amount of overlap time? If overlap is not allowed, so that I am on EI but not NSERC supplemental for those five weeks, then can I continue the NSERC supplemental funding for an additional five weeks beyond the end of the EI duration, from Mar 27 2022 - Apr 28 2022?

4. My plan is to take over 30 weeks of leave so I assume I do not need to fill out the Parental Accommodation Form, can you confirm that?
   https://www.grad.ubc.ca/faculty-staff/policies-procedures/parental-accommodation

April 23, 2021

Timeline

- Current plan: on parental leave July 12 2021 to March 25 2022 (due date Jul 21)
    - But hope for child care starting Jan 1 so could do gradual ramp-up (unofficial plan)
- Paper endgame plan
    - Analysis 2 more weeks, Apr W4 and May W1.
    - Pre-paper talk 2-3 weeks, May W2 & W3. Present to group May 20 or May 27
    - Prose 1 month. Jun. Group draft read Jun 24, after two rounds from Steve and one round each from Charles and Tamara

Update

- Finished coding the remaining 10%, 90% + 10% = 100%
- Establishing saturation: I used Google Colab (basically Jupyter notebooks hosted on Google Drive) to graph unique codes added to the codeset with each interview.
    - Tasks saturated quickly because we didn't go too granular with this dimension. We already went really deep into this in Table Scraps.
- Solidifying codes
    - Tasks: added categories for *communicate* and *verify*.
        - These categories come from sub-process names in "Passing the Data Baton." They're relevant to our focus on *data preparation* but are at the end of the pipeline in the *communication* process.
    - Issues: No major changes
    - Tools: about a dozen *tool opinions* that seem relevant but not linked to a particular tool
    - Education: unclear if it's worth detailed coding at this point. Possibly difficult to say formal training in data journalism. Maybe just formal training in journalism? Maybe abandon?

What tasks do users do?

- Everyone engages in data integration
    - A lot of this frequency could be explained by the scope of our interview study question list

How are these tasks different from "data scientists?"

- Started synthesizing qualitative data in code summaries, short memos summarizing the differences between "data scientists" and journalists within each task code.
    - Is *Classify observations* really *Label data* and *aggregate*, but aggregation with categorical variables?

Who are these users?

- Colab notebook on users
    - Most users don't formally identify as data journalist
- Data on which tools users said they use.
- Users are more than the tools they use, *archetypes* [Kandel, 2012]
    - Chance here to describe more well-rounded personas based on multiple characteristics
        - Not just what tool they use

April 16, 2021

Timeline

- Goals: Have the paper out the door by mid July before the baby comes.

Update

- Second-pass Analysis
    - 90% done going through codes

- Who is the user?
    - We can understand this through their education/training and tools they use.
- What is the user's task?
    - Captured in the tasks codes
- Where do they work?
    - How large, and how well supported, is the data team that they work within.
        - Sort of captured with roles
- Why is this task difficult?
    - Captured in the issues codes
- How do they overcome these difficulties?
    - Also captured in the tasks codes, separate column
    - What are some of the salient ways that issues pair these tasks.
        - E.g. integration appears to be tightly coupled with entity resolution through the issue inconsistent data because different data collectors have different names for the same thing, even when they are consistent within data collector

Questions

- What if the codes in the codebook were more situated in the language that journalists use, but the description of that code was situated in the language of wrangling/computer science? Or vice versa?
    - Data journalism terminology in this area doesn't seem as comprehensive or rigorous as CS
    - I unconsciously translated codes into CS-speak, but the language is still in the data. Some notable examples:
        - Standardize data: resolve entities
        - Data check/bulletproofing data: verify data
    - This would make the codebook a translation dictionary for journalists and computer scientists to communicate with one another.
    - Would only require time proportional to the number of codes.
    - Is this discourse analysis?
- What to do with all the Miscellaneous issues?

Terminology

- Data
    - Dataset
        - Table
            - Rows (in a tabular data)
            - Columns (in a tabular data)
            - Values

Task List Interview Study Overview Thesis Proposal Timeline Task Summaries Interview Study Pre-paper

April 9, 2021

Schedule

- Second-level analysis is due today

Progress

- Unhappy with codebook 1
  - Coding perspective makes codes convoluted
    - Includes too many disjoint topics: tasks, pain points, education, tool use, personas, etc
    - Coding perspective was too broad: what do we talk about when we talk about wrangling?
  - Very difficult to organize for analysis
    - Too much data all in one place: 900 data points total
    - Google Sheets gets really slow with this much data
    - Chop up into segments for easier analysis

- Happier with codebook 2: problem-solution analysis (~200 / 900 data points revisited)
  - What?
    - Repivot perspective on codebook 1
      - New perspective: What pain points do users face and how do they address them?
    - Excluding topics such as: tool use, education
      - Can possibly tackle these later
  - Why?
    - High-level understanding of wrangling
      - Table Scraps process models are too low-level
      - To get "high level" take into account data problems
    - Surfaces links between problems and solutions, e.g.:
      - If there's *too much data* (problem), then *filter data* (solution)
      - If *data is locked in PDF* (problem), then manually *enter data* (solution)
    - Better structure to explore other topics
      - Relationships between different wrangling tasks, e.g.
        - Deriving soft keys to join datasets
        - Classifying data to join datasets
  - How?
    - Refining codebook 1 codes, not re-coding data.
      - Axial coding
  - When?
    - ~1.5 min per code, x 700 codes = 17.5 hours, so probably finish mid next week
    - Schedule:
      - Second-pass due April 16
      - Pre-paper talk April 30 still? (maybe)

April 2, 2021

Update

- Finished "transcribing" and coding all 36 interviews 🥳
- Currently incorporating codes in interview "transcripts" into the master spreadsheet of theme
    - Noting topics where what journalists say they do does not quite align with what I previously found looking through all those data-scientist study papers.
    - Here are three that are all tightly connected. I'll have more by next Friday.

Bridging the gulf of evaluation

From reading all those data-scientist studies, I came up with a category called *understanding the data*: calculating summary statistics, profiling the data, visually reading the table, and inspecting the cardinality/range of different variables. In these interviews, journalists frequently describe using the same methods but to verify the effect of various transforms, including merging datasets, resolving entities within a column, and deriving new variables from a dataset. This phenomenon seems like a gulf-of-evaluation issue, where the system is not providing the information necessary for the user to understand the current state of the data/system. While both data scientists and data journalists engage in data wrangling and data integration, the data-scientist study literature does seem to talk about how we know the transformations we apply did exactly what we think they did; although, journalists loved talking about this. It seems like an important issue as datasets become so large that it's unreasonable to manually search through every row.

Understanding the data beyond statistical summaries

Quite a few journalists describe a phase of vetting and understanding the raw data that is deeper and more critical than *understanding the data*. Journalists may get a dataset with no documentation and incomprehensible column names and have to do background research just to understand exactly what the dataset contains. They may need to know how the data was collected. They also want to know if the data is accurate. A few journalists told me of times when the data they got was just wrong. One time the government agency knew the data they provided was full of errors. Several journalists repeatedly stressed how they never trust raw data they get, even when they get it from the government.

High Stakes Wrangling

A common thread that runs through both of these themes, bridging the gulf of evaluation, understanding the data beyond statistical summaries, is that the data needs to be error free, whether that responsibility falls upon the journalist or the agency supplying the data, because the stakes are really high for data journalists. Many journalists described going on every row in a dataset before publication. When they know their dataset is not perfect a few journalists talked about using *at-least numbers* instead of *absolute numbers* in stories, e.g. at least *x* companies that received PPP loans still laid off workers in these states, to deal with the uncertainty with the data. One investigative data reporter told me that they intentionally trade false negatives for false positives because falsely naming one company of wrongdoing can tarnish the credibility of the whole investigation. A few journalists talked about the pressure and expectation that their dataset is 100% correct. One participant said, "80% accuracy for some business domains is not only acceptable but an incredible breakthrough. 80% accuracy in journalism will probably get you permanently banned from the profession."

March 19, 2021

Update:

- Coding progress:
  - "Transcribing" interviews / cleaning up notes: 17/35 (~1.5 hr each), 27 hrs left to go
  - Coded interviews: 7/35 (~1 hr each), 28 hrs left to go
    - Only 3/35 are integrated into spreadsheet currently (takes 5 min)
  - 6 quality hours a day = 9+ days of work left
  - I think we're still on pace
- Big picture:
  - This process is like cutting out *clips* from different *sources*, giving them *tags*, then comparing the *clips* within *tag* but between *sources*
  - Not going to see "the fruits of our labor" until after all interview coding is done
  - I think I'm going to need something to facet and filter within these groups.
    - I think I should try to build a viewer with this Isotope framework.
- Paper updates:
  - Based on interview I think we should go back and add Ana's papers in defining needs and analysis verification
    - Change editorial decisions to management judgements or newsworthiness judgements
-

March 12, 2021

First-level analysis

Crisan et al. 2020 data science papers on data preparation (24 papers total / 21 are relevant)

- ✓ S. Abt and H. Baier, “A Plea for Utilising Synthetic Data when Performing Machine Learning Based Cyber-Security Experiments,” in *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop - AISec ’14*, Scottsdale, Arizona, USA, 2014, pp. 37–45, doi: 10.1145/2666652.2666663.
- ✓ L. Battle and J. Heer, “Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau,” *Computer Graphics Forum*, vol. 38, no. 3, pp. 145–159, 2019, doi: https://doi.org/10.1111/cgf.13678.
- ✓ A. Crisan, J. L. Gardy, and T. Munzner, “On Regulatory and Organizational Constraints in Visualization Design and Evaluation,” in *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV ’16*, Baltimore, MD, USA, 2016, pp. 1–9, doi: 10.1145/2993901.2993911.
- ✓ A. Crisan and T. Munzner, “Uncovering Data Landscapes through Data Reconnaissance and Task Wrangling,” in *2019 IEEE Visualization Conference (VIS)*, Vancouver, BC, Canada, Oct. 2019, pp. 46–50, doi: 10.1109/VISUAL.2019.8933542.
- ✓ D. Donoho, “50 Years of Data Science,” *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 745–766, Oct. 2017, doi: 10.1080/10618600.2017.1384734.
- ✓ M. Feinberg, “A Design Perspective on Data,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA, May 2017, pp. 2952–2963, doi: 10.1145/3025453.3025837.
- ✓ U. M. Feyyad, “Data mining and knowledge discovery: making sense out of data,” *IEEE Expert*, vol. 11, no. 5, pp. 20–25, Oct. 1996, doi: 10.1109/64.539013.
- ✓ G. Grolemund and H. Wickham, “A Cognitive Interpretation of Data Analysis: A Cognitive Interpretation of Data Analysis,” *International Statistical Review*, vol. 82, no. 2, pp. 184–204, Aug. 2014, doi: 10.1111/insr.12028.
- ✓ S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, “Enterprise Data Analysis and Visualization: An Interview Study,” *IEEE Trans. Visualization and Computer Graphics (TVCG)*, vol. 18, no. 12, pp. 2917–2926, Dec. 2012.
- ✓ S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, “Profiler: integrated statistical analysis and visualization for data quality assessment,” in *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI ’12*, Capri Island, Italy, 2012, p. 547, doi: 10.1145/2254556.2254659.
- ✓ N. Khan *et al.*, “Big Data: Survey, Technologies, Opportunities, and Challenges,” *The Scientific World Journal*, Jul. 17, 2014. https://www.hindawi.com/journals/tswj/2014/712826/ (accessed Mar. 01, 2021).
- ✓ M. Kim, T. Zimmermann, R. DeLine, and A. Begel, “The emerging role of data scientists on software development teams,” in *Proceedings of the 38th International Conference on Software Engineering*, Austin Texas, May 2016, pp. 96–107, doi: 10.1145/2884781.2884783.

- ✓ M. Kim, T. Zimmermann, R. DeLine, and A. Begel, “Data Scientists in Software Teams: State of the Art and Challenges,” *IIEEE Trans. Software Eng.*, vol. 44, no. 11, pp. 1024–1038, Nov. 2018, doi: 10.1109/TSE.2017.2754374.
- ~~✓ Á. Kiss and T. Szirányi, “Evaluation of manually created ground truth for multi-view people localization,” in *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, St. Petersburg Russia, Jul. 2013, pp. 1–6, doi: 10.1145/2501105.2501106.~~
    - [Not applicable] Ana says this paper highlights "the different ways to create synthetic data and the importance of domain expertise in the data creation process," but it's about a very specific data-sciencey problem of detecting people in computer vision algorithms, and is not application to the kind of data problems I'm concerned with.
- ~~✓ Y. Mao *et al.*, “How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. GROUP, pp. 1–23, Dec. 2019, doi: 10.1145/3361118.~~
    - [Note applicable] Ana cites this paper under Defining Needs, which is a subsection of data preparation, but it is so far upstream of wrangling that I do not see how defining needs in this paper impacts wrangling downstream.
- ✓ A. Milani, F. Paulovich, and I. Manssour, “Visualization in the preprocessing phase: an interview study with enterprise professionals,” *arXiv:1908.07894 [cs]*, Aug. 2019, Accessed: Nov. 26, 2020. [Online]. Available: http://arxiv.org/abs/1908.07894.
- ✓ M. Muller and others, “How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation,” in *Proc. Conf. on Human Factors in Computing Systems (CHI)*, May 2019, pp. 1–15.
- ~~✓ M. Sedlmair, M. Meyer, and T. Munzner, “Design Study Methodology: Reflections from the Trenches and the Stacks,” *IEEE Trans. Visual. Comput. Graphics*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012, doi: 10.1109/TVCG.2012.213.~~
    - [Not applicable] Ana cites this paper as an example for defining needs intended for visualization researchers and not data scientists, but data scientists may still benefit from these methods. While data abstraction includes transforming and deriving data. The relevance of this work to data wrangling does not extend beyond stating this fact and many other papers have gone in greater depth on these two subjects.
- ✓ E. Serrano, M. Molina, D. Manrique, and L. Baumela, “Experiential Learning in Data Science: From the Dataset Repository to the Platform of Experiences,” in *Intelligent Environments*, 2017, vol. 22, pp. 122–130, doi: 10.3233/978-1-61499-796-2-122.
- ✓ M. Stonebraker and I. F. Ilyas, “Data Integration: The Current Status and the Way Forward,” *IEEE Data Eng. Bull.*, vol. 41, pp. 3–9, 2018.
- ✓ D. Wang *et al.*, “Human-AI Collaboration in Data Science: Exploring Data Scientists’ Perceptions of Automated AI,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–24, Nov. 2019, doi: 10.1145/3359313.
- ✓ R. Wirth, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

- ✓ K. Wongsuphasawat, Y. Liu, and J. Heer, “Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study,” *arXiv:1911.00568 [cs]*, Nov. 2019, Accessed: Feb. 08, 2021. [Online]. Available: http://arxiv.org/abs/1911.00568.
- ✓ “State of Data Science and Machine Learning 2019.” [Online]. Available: https://www.kaggle.com/kaggle-survey-2019.

Themes

- First-pass done and digitized

Feb. 26, 2021

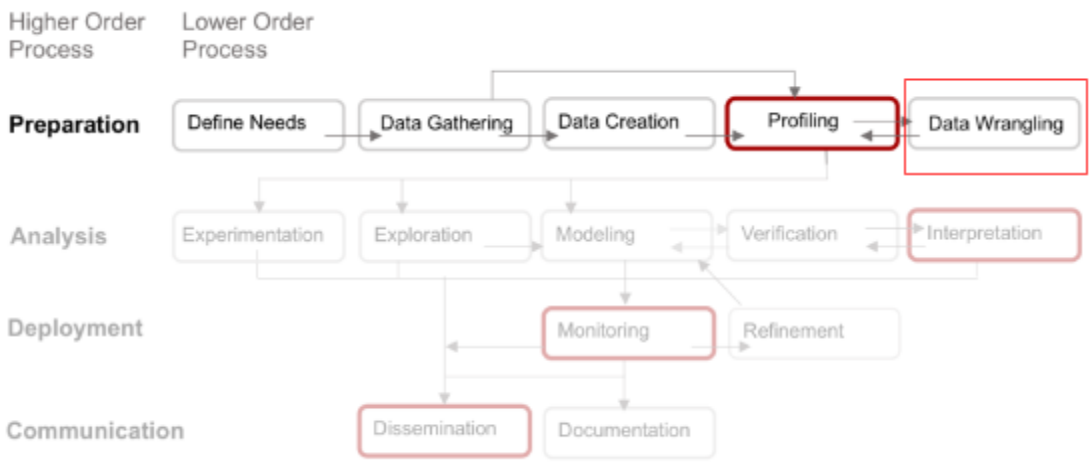
Interview Study

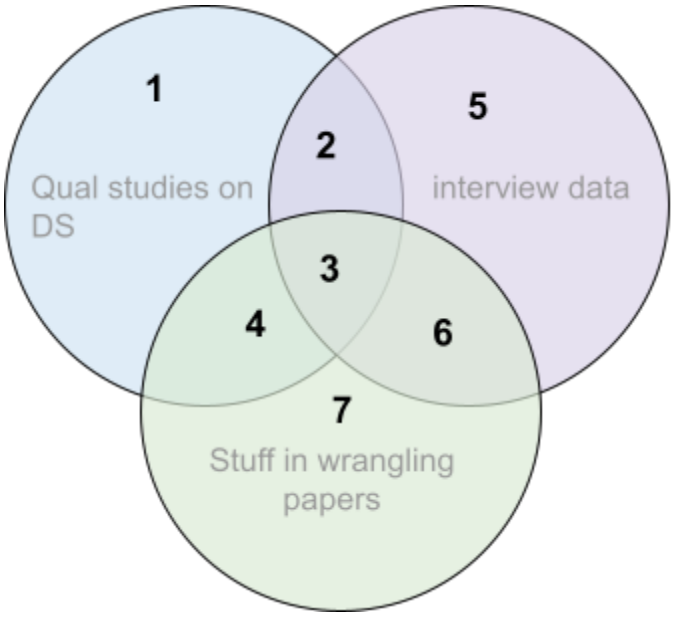
Updates

- About halfway through related work
  - Macro views
- Infrastructuring is complete
  - Themes sheet collects interview note themes
  - Separates Python code collects "themes" from related work

scope

- Most interview study focus on the end-to-end process
- We focuses on just the preparation process
  - via Crisan et al. 2020.





DS Studies = qualitative studies about data scientist processes

Material:

1. ~~In data science, are not about wrangling, and was not talked about in interviews (don't care about)~~
    a. ~~E.g. Model selection, machine learning stuff~~
2. In data science, are not about wrangling, and was talked about in interviews (care a little bit)
    a. e.g. Visualization, analysis, deployment, communication
    b. You can talk about
3. **In data science, are about wrangling, and was talked about in interviews, e.g.**
    a. **Data integration**
4. In data science, are about wrangling, and was not talked about in interviews (care a little bit)
    a. Notable by omission
    b. We can point out things here that were not shared in our interviews
5. Not in data science, are not about wrangling, and was talked about in interviews (care a little bit)
6. **Not in data science, are about wrangling, and was talked about in interviews**
7. ~~Not in data science, are about wrangling, and wasn't talked about in interviews, e.g.~~
    a. ~~Hard-core database cleaning and warehousing stuff~~

Notes on revising Venn diagram:

- This should be in pre-paper talk
- Shouldn't have three different colors
    - Could use color to depict interest

Analyzing data for themes

Should we develop themes bottom up or top down? Example of a top-down theme hierarchy that we have from related work:

- Should do a hybrid multi-pass approach of top-down paper stuff and bottom-up based on the interview data.
    - 1st pass: interviews
    - 2nd pass: related work, prepared mind, reading for a purpose
    - 3rd pass: coding interviews, keeping related-work top-down themes in mind and noting bottom-up stuff in the interview data.
    - 4th pass: affinity diagramming and synthesis

Examples

- Archetypes/personas
    - Backgrounds/experiences
    - Formal education
- Data preparation
    - Processes
        - Tools
        - Organization of teams
        - Types of projects
    - Tasks/activities
        - Challenges/pain points

Leave issues

- Sharon (CS HR) said talk to Joyce. Joyce said talk to G&PS. G&PS pointed to same parental accommodation page and said talk to Joyce.
    - International student advising appointment is next hope of clarification
    - G+PS:
        - Cady Tran
        - Joanne Tsui

-

Feb. 16, 2021

Agenda

- Today is a short meeting

Parental leave stuff

- Talking with Stephen Ingram today
- No response from GPS, yet (Friday afternoon email, Monday was a holiday)

Interview study

"Analyzing the analyzer" / qualitative studies of data scientists papers:

- S. Abt and H. Baier, “A Plea for Utilising Synthetic Data when Performing Machine Learning Based Cyber-Security Experiments,” in *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop - AISec ’14*, Scottsdale, Arizona, USA, 2014, pp. 37–45, doi: 10.1145/2666652.2666663.
- L. Battle and J. Heer, “Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau,” *Computer Graphics Forum*, vol. 38, no. 3, pp. 145–159, 2019, doi: https://doi.org/10.1111/cgf.13678.
- A. Crisan, J. L. Gardy, and T. Munzner, “On Regulatory and Organizational Constraints in Visualization Design and Evaluation,” in *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV ’16*, Baltimore, MD, USA, 2016, pp. 1–9, doi: 10.1145/2993901.2993911.
- A. Crisan and T. Munzner, “Uncovering Data Landscapes through Data Reconnaissance and Task Wrangling,” in *2019 IEEE Visualization Conference (VIS)*, Vancouver, BC, Canada, Oct. 2019, pp. 46–50, doi: 10.1109/VISUAL.2019.8933542.
- D. Donoho, “50 Years of Data Science,” *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 745–766, Oct. 2017, doi: 10.1080/10618600.2017.1384734.
- M. Feinberg, “A Design Perspective on Data,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA, May 2017, pp. 2952–2963, doi: 10.1145/3025453.3025837.
- U. M. Feyyad, “Data mining and knowledge discovery: making sense out of data,” *IEEE Expert*, vol. 11, no. 5, pp. 20–25, Oct. 1996, doi: 10.1109/64.539013.
- G. Grolemund and H. Wickham, “A Cognitive Interpretation of Data Analysis: A Cognitive Interpretation of Data Analysis,” *International Statistical Review*, vol. 82, no. 2, pp. 184–204, Aug. 2014, doi: 10.1111/insr.12028.
- S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, “Enterprise Data Analysis and Visualization: An Interview Study,” *IEEE Trans. Visualization and Computer Graphics (TVCG)*, vol. 18, no. 12, pp. 2917–2926, Dec. 2012.
- S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, “Profiler: integrated statistical analysis and visualization for data quality assessment,” in *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI ’12*, Capri Island, Italy, 2012, p. 547, doi: 10.1145/2254556.2254659.

- N. Khan *et al.*, “Big Data: Survey, Technologies, Opportunities, and Challenges,” *The Scientific World Journal*, Jul. 17, 2014. https://www.hindawi.com/journals/tswj/2014/712826/ (accessed Mar. 01, 2021).
- M. Kim, T. Zimmermann, R. DeLine, and A. Begel, “The emerging role of data scientists on software development teams,” in *Proceedings of the 38th International Conference on Software Engineering*, Austin Texas, May 2016, pp. 96–107, doi: 10.1145/2884781.2884783.
- M. Kim, T. Zimmermann, R. DeLine, and A. Begel, “Data Scientists in Software Teams: State of the Art and Challenges,” *IIEEE Trans. Software Eng.*, vol. 44, no. 11, pp. 1024–1038, Nov. 2018, doi: 10.1109/TSE.2017.2754374.
- Á. Kiss and T. Szirányi, “Evaluation of manually created ground truth for multi-view people localization,” in *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, St. Petersburg Russia, Jul. 2013, pp. 1–6, doi: 10.1145/2501105.2501106.
- Y. Mao *et al.*, “How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. GROUP, pp. 1–23, Dec. 2019, doi: 10.1145/3361118.
- A. Milani, F. Paulovich, and I. Manssour, “Visualization in the preprocessing phase: an interview study with enterprise professionals,” *arXiv:1908.07894 [cs]*, Aug. 2019, Accessed: Nov. 26, 2020. [Online]. Available: http://arxiv.org/abs/1908.07894.
- M. Muller and others, “How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation,” in *Proc. Conf. on Human Factors in Computing Systems (CHI)*, May 2019, pp. 1–15.
- M. Sedlmair, M. Meyer, and T. Munzner, “Design Study Methodology: Reflections from the Trenches and the Stacks,” *IEEE Trans. Visual. Comput. Graphics*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012, doi: 10.1109/TVCG.2012.213.
- E. Serrano, M. Molina, D. Manrique, and L. Baumela, “Experiential Learning in Data Science: From the Dataset Repository to the Platform of Experiences,” in *Intelligent Environments*, 2017, vol. 22, pp. 122–130, doi: 10.3233/978-1-61499-796-2-122.
- M. Stonebraker and I. F. Ilyas, “Data Integration: The Current Status and the Way Forward,” *IEEE Data Eng. Bull.*, vol. 41, pp. 3–9, 2018.
- D. Wang *et al.*, “Human-AI Collaboration in Data Science: Exploring Data Scientists’ Perceptions of Automated AI,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–24, Nov. 2019, doi: 10.1145/3359313.
- R. Wirth, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.
- K. Wongsuphasawat, Y. Liu, and J. Heer, “Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study,” *arXiv:1911.00568 [cs]*, Nov. 2019, Accessed: Feb. 08, 2021. [Online]. Available: http://arxiv.org/abs/1911.00568.
- “State of Data Science and Machine Learning 2019.” [Online]. Available: https://www.kaggle.com/kaggle-survey-2019.

Feb. 12, 2021

Agenda

Parental leave

- Tentative timeline
  - June 7, 2021: Deadline to submit parental UBC unpaid parental leave paperwork
  - July 7, 2021: Begin UBC unpaid parental leave
  - July 21, 2021: Baby's due date
  - January 2022 / March 23, 2022: Return to regular grad student-ing
    - January if we get childcare (UBC), March is maximum extent for UBC unpaid parental leave
- Plan for July 7, 2021, until Jan. 1, 2022 (🤞)
  - On thesis
    - No way of knowing how much time I'll have to work until the baby arrives
      - But want to spend any free cycles on making thesis progress
      - Does this mean a leave of absence?
    - Want to start parental leave in a place to address technical aspects of P3 when time allows
      - Pretty much done with Interview Study
        - Does this include writing?
  - On making ends meet (Still trying to navigate all these options):
    - Does RA'ing still make sense?
    - TA'ing at UBC / MDS?
    - EI benefits through Services Canada?
      - Still trying to figure this out...
    - NSERC paid parental leave?
      - Shalaya no NSERC
  - TODO:
    - Steve checking with GPS, is there magical NSERC leave money?
      - Then Tamara checks with Sharon.

NICAR, March 3-5, 2021 (Virtual)

- Are lightning talks posted somewhere?
  - https://www.ire.org/training/conferences/nicar-2021/nicar21-lightning-talks/
- Would it be worth attending to see their wrangling sessions / networking?
  - Registration cheap, only $50
  - Three days, but only a subset of talks look interesting. Go for it.
  - Students qualify for mentors
  -

Interview Study

- Participants so far: 32. Definitely hitting saturation.
  - First-level analysis 2.5 weeks. Calendar time: by Mar 12.
  - Second-level analysis 2 weeks. Calendar time: by Apr 9.

    - Pre-paper talk two weeks, across three calendar weeks. Cal time: Apr 30
    - Writing: four weeks, across six calendar weeks. Cal time: Jun 11.
- How are we going to analyze this?
- Questions this study seeks to answer:
    - How important and common is wrangling multiple datasets among data journalists?
    - When wrangling data, how do the pain points, tasks, and solutions used by data journalists compare to what has been reported in the growing body of qualitative work on data scientists?
        - "Passing the Data Baton" already gathers most of this related work, about a dozen papers
        - We can break this topic into three groups:
            - What do both data scientists and data journalists both do?
            - What do data journalists do that isn't addressed in this research?
            - What do data scientists do that data journalists didn't address
                - Seems like shakier ground since we didn't interview *every data journalist*
    - How do the formal and informal roles data journalists adopt compare to the roles enumerated in this ethnographic data scientist literature?
- This spreadsheet is kind of how I'm organizing the data within our interviews with previously published work
    - Is this too heavy for analysis? No, looks good!
    - Example of analyzing interview
    - P1 Adam Hooper: 1 hr preprocessing, 1 hr open coding. Assume 2 hrs per interview * 35 interviews = 70 hours
    - Paper Kim 2 hrs, paper Fisher 1 hr. Assume 2 hrs per paper * 12 papers = 24 hrs.
-

Jan. 29, 2021

Agenda

- Interview study update:
  - n=32
  - Next week plan
    - Let interviews run on their own inertia
    - Start trying to make sense of data we currently have
  - How should we report participant demographics?
    - We will definitely report our own categories, as they map to interview behavior:
      - "R-purists" vs "McGuyvers," right tools for the person job
      - Big newsrooms vs small newsrooms
        - Is Urban vs. rural an un-needed category that recapitulates this?
    - If reviewers harrasses us about it or if in analysis we feel like it's useful, we can revisit it and do either a rough and dirty column where we infer or touch back and ask them.

Notes

- Category of journalists, a typology could be a contribution, seems to mirror data journalism pedagogy says Charles. Can we counter a narrative of stupid roles with our data?
  - Garden-variety reporter using data skills
  - Data-driven investigative reporters
- Think of conclusions at three levels:
  - Data journalism
    - Data wrangling
      - Multi-table data wrangling
- Some people to consider:
  - Mark Hansen, although he would say he's not a journalist (don't believe him)
  - Meredith Broussard, at NYU
  - Matt Waite, email again
  - Soo Oh (in progress)
  - Sarah Cohen, email again
  - Jonathan Stray
  - Simon Willison
  - Gregor Aisch
  - Mike Bostock
  - John Burn-Murdoch
  - Jeremy Merrill
  - Aram Chung
  -

Jan. 22, 2021

Agenda

- Interview study update
  - n = 26
  - Still haven't had an interview where I didn't learn something
  - Experience diversity is good, but those who have been doing this longer (and then wind up at bigger organizations) have the best answers
- NICAR Lightning Talk
  - Similar lightning talks
    - Accepted:
      - Maryjo Webster; Wrestling with Data (without Coding); 2015
      - Steven Rich; The Five Stages of Terrible Data: Denial, Anger, Bargaining, Depression, Acceptance; 2015
      - Chris Groskopf; Let lookup save you from the boring, repetitive work you've forgotten you're even doing; 2016
    - Rejected
      - Christine Zhang; Data Cleaning Made Easy(er): Excel and R Side-by-Side (2016)
      - Daniellle Cervantes; Data-cleaning Tricks Using Excel; (2010)
  - Pitch drafts
    - Wacky Wrangling Woes (WWW)
    - A taxonomy of data wrangling nightmares
      -
    - What I learned by stalking all your GitHub
      - Five data wrangling nightmares in five minutes (and how to vanquish them)
    - All your GibHub are belong to us
  -

Jan. 14, 2021

Agenda

- Interview study status
  - n=21
  - Still learning a bunch of good stuff
    - Interesting things:
      - Major pain points
        - People say: Joining table, reconciling entities,
          - (& PDF extraction, not our concern for P3 tool building but relevant for big picture)
        - Not necessarily separate things
          - You might need to reconcile names before joining.
        - Joining on geographic entities
          - Entities could differ in scope, entity boundaries could change
- Can't audit Qual Methods course
  - But I pull the course reading list
- NICAR Lightning Talk
  - https://www.niemanlab.org/2016/03/10-nicar-lightning-talks-to-guide-you-through-cats-statistical-resampling-fear-of-math-and-more/
  - Lightning Talks: 5-minute presentations on particular skills, tools or techniques.
    - Talks are pre-recorded this year
    - These more lighter, entertaining talks
  - Pitches are submitted and voted on by NICAR21 attendees.
    - Pitches: a title and 2-3 sentences
    - Pitch deadline Jan. 27
  - Possible ideas:
    - Snarky Table Scraps, 5 min.
      - Wacky Wrangling Woes (WWW)
      - A taxonomy of data wrangling nightmares
      - What I learned by stalking all your GitHub
      - All your GibHub are belong to us
    - TODO Come up with ideas and chat in next meeting
    -

Jan. 08, 2021

Agenda

- Paper draft Feb. 4 / Feb. 19 deadline?
  - C+J 2021?
- CHI 2021 Nominated SVs
- Interview request success rate is 31%
  - Spreadsheet

Meeting

- C+J
  - Charles abstract due Jan 19, contributed talk if accepted happens Feb 19
  - Plan is to test-drive ideas for later journal paper on Newsroom Sensemaking: A Process Model for Epistemic Humility in Data and Computational Journalism
  - Paper first author is Charles, co-authors also Tamara and Steve.
- Abstract first draft from Charles

Charles Berret, Stephen Kasica, Tamara Munzner

University of British Columbia

cberret@mail.ubc.ca

Newsroom Sensemaking: A Process Model for Epistemic Humility in Data and Computational Journalism

This paper draws from the literature of schematizing and sensemaking in design research to offer a process model specific to data and computational journalism. Classic process models for data analysis tend to assume that the principal tools at an analyst's disposal are the data itself plus logic, which the analyst uses to develop and refine a single explanatory mental schema. But even for news stories with a heavy data component, journalists tend to draw from a wider range of information sources than just data, reason through a broader set of considerations than logic alone, and respond to a complicated array of ethical, legal, and narrative concerns that may affect the schema that is ultimately represented in a published story. Although time constraints play a central role in many process models, forming the "cost structure" of sensemaking toward a single explanatory schema, in a newsroom deadline pressure often requires the consideration of multiple plausible schemas in parallel, among which the most defensible is chosen on deadline. Working with a multiplicity of plausible schemas has notable advantages: it reduces the risk of heuristic biases and promotes a general stance of epistemic humility respecting the many ways a story can miss its mark, especially when produced under strict constraints. While our process model is diagnostic of issues typically encountered in journalism, and prescriptive of a working mindset conducive

to responsible and accurate reporting, its applications may extend to other forms of data work in which epistemic humility and ethical framing are also salutary.

- Steve interview study
    - Really are dealing with big data, sometimes huge 1M row files that don't fit into excel.
    - Bradshaw, one project consolidating 10K datasets (precinct-level across all England across months)
    - Journalistic food chain: start with global and figure out bridge to local
    - Massively multitable wrangling!
    - Often don't have data journalist title. MacGuyvering. Terms used a lot in DJ. Bricoleurs
    - Wrangling is less bad than it used to be.
    - Lots of confirmation that our intuitions from TableScraps study are true. Pain points. Changing meaning of codes. Skolemization (creation of soft/foreign keys for joins)
    - Yes wrangling takes lots of time
    - Combining multiple data sets is indeed super common
    - Two camps: coder types (convenience and reproducibility, comfort with familiar monolithic tooling) vs those very willing to switch between tools based on what is best for that goal (opportunists, very open to trying newest/latest as a way to have an edge)
        - Continuing theme. Some feel burned by perishable tools that wasted time, vs
        - Attitudes towards interactive technologies will play well with chi crowd
    - Being scooped while doing wrangling by somebody else (big orgs with more resources)
    - Teaching DJ pedagogy debate: is starting with code better than starting with tools (Excel). Is it preference personality or mindset dependent, or circumstance?? Arguments for separate programs that data journalists?? Remember todo bring up this point in Implications for Design (of Pedagogy not necessarily Tool) in the paper!!
        - CG&A special issue on viz pedagogy
        - Tech-jteach google group
    - Unwrangleable datasets Q. They all start out saying yes, but can't think of any… then oh wait *that* one. It's never about deciding it's impossible, it's about whether it's worth the time and whether they'll have sufficient confidence at end of the process. And whether circumstances will change enough to make story no longer relevant if it takes too long.
    - Cost structure of data wrangling vs newsroom sensemaking. Order of these two papers unclear.
    - Wished they could have gone deeper, article didn't have the teeth they wanted.
    - Yes could do second-round analysis of this data to resolve journalistic controversies later on.
    - Data feminism author point: data gaps (you think it should exist but dataset not there). Multitable wrangling as a way to get that data that addresses gaps. ** interesting use case! **
    - Ana paper on roles in DS - noting ways it doesn't match up to small newsrooms. Multiple hats vs specialists. Often have dedicated visualization specialists since newsrooms have graphics teams, but usually have to do your own wrangling.
    - Other venues for these tidbits / insights: 1500 word thing in IRE journal. Or our own blog. Or Multiple Views. Postify papers.
- Timeline
    - 15 interviews done so far, working towards 30-ish and/or saturation, we'll see.
    - Looking on target for data collection done before end of January, maybe even sooner!
    - 50 people that we haven't contacted yet, at 30% hit rate we're in very good shape.