Demographics: qualitative responses

At this point, I've manually reviewed the classifications of people's write-in responses to Q18 in Microsoft Excel. There were definitely some decisions that I think even state-of-the-art AI would have struggled with. You could call this notebook "post-review curation".

I saved my work in an excel spreadsheet. The most important column was the "decision" column, so I just copied that one column to a code-friendly text editor and saved it as a tsv. Here's a key to my mark-up in that column: k = keep as-is d = drop If I put a number in the decision column, that means replace the existing row with that row of the taxonomy. Also, the taxonomy didn't have "AI" or "Machine Learning" by themselves. So I tagged people who wrote in "AI" or "ML" as 1017, "Artificial Intelligence and Robotics", even though none of these people mentioned robotics. I'll correct that later in this script.

We'll want to use this column to make substitutions in the data frame produced by demographics_qual_part1.qmd.

I also recorded 4 rows that need to be added. These were cases where people indicated two fields, usually with an "and" instead of with a slash or comma, so I didn't split these on my first pass. Since there's only 4 of these, I'm just hard-coding them in here. (Ugly code, yeah, I know.)

```
additions <- data.frame(
    participantID = c(16, 62, 137, 159),
    decision = c(1029, 1016, 1093, 838)
)
```

Import packages and utilities

```
project_root <- here::here() # requires that you be somewhere in the
# project directory (not above it)
# packages
suppressMessages(source(file.path(project_root, "scripts/packages.R")))
```

```r
# functions and objects used across scripts
suppressMessages(source(file.path(project_root, "scripts/utils.R")))
```

Load data

```r
# do not read with my load_qualtrics_data function bc it has header = TRUE
tax <- read.csv(
  file = file.path(DATA_PATH, "curation_of_disciplines/digital_commons_disciplines.txt"),
  header = FALSE,
  sep = "\t",
  check.names = FALSE,
  stringsAsFactors = FALSE
)

# Data cleaning
tax <- tax %>%
  separate(
    col   = names(tax)[1],
    into  = c("Level1", "Level2", "Level3"),
    sep   = ": ",
    fill  = "right",   # any missing pieces become NA
    extra = "merge"    # if there were >2 colons, they’d all merge into Level3
  )
```

```r
auto_results <- load_qualtrics_data("curation_of_disciplines/qual_fields_guesses.tsv")

# a manually generated one-column table
# do not read with my load_qualtrics_data function bc it has header = TRUE
decisions <- read.csv(
  file = file.path(DATA_PATH, "curation_of_disciplines/classification_decisions.tsv"),
  header = FALSE,
  sep = "\t",
  check.names = FALSE,
  stringsAsFactors = FALSE
)

names(decisions) <- c("decision")
```

Let’s make sure that both data frames have the same number of rows.

```
nrow(auto_results) == nrow(decisions)
```

```
[1] TRUE
```

Curate the data based on manual review

Now, let’s start updating the data frame we got in part 1 with the manual revisions. First, let’s drop the rows we can drop.

```
dim(auto_results)
```

```
[1] 240   5
```

```
curated <- auto_results
curated <- cbind(curated, decisions)

curated <- curated %>%
    filter(decision != "d")

dim(curated)
```

```
[1] 191   6
```

```
head(curated)
```

```
  participantID            response                            Level1
1             1 Ai and Neuroscience                     Life Sciences
2             2       Plant Biology                     Life Sciences
3             2             Ecology                     Life Sciences
4             3  Digital Humanities               Arts and Humanities
5             3             History               Arts and Humanities
6             4    Computer Science Physical Sciences and Mathematics
                         Level2               Level3 decision
1 Neuroscience and Neurobiology Systems Neuroscience      670
2                Plant Sciences        Plant Biology        k
3               Animal Sciences              Zoology      617
4            Digital Humanities                 <NA>        k
5                       History                 <NA>        k
6             Computer Sciences                 <NA>        k
```

Next, let's substitute the bad rows with the correct ones.

```
# Try to coerce entries in decision to numeric
dec_num <- suppressWarnings(as.integer(curated$decision))
head(dec_num)
```

```
[1] 670  NA 617  NA  NA  NA
```

```
# Logical vector indicating whether decision is a (sensible) number
ok <- !is.na(dec_num) & dec_num >= 1 & dec_num <= nrow(tax)
head(ok)
```

```
[1]  TRUE FALSE  TRUE FALSE FALSE FALSE
```

```
# overwrite Level columns in curated with those cols
# in the corresponding row from tax
curated[ok, c("Level1", "Level2", "Level3")] <-
  tax[dec_num[ok], c("Level1", "Level2", "Level3")]
```

Let's inspect the results with a side-by-side comparison (well, vertically speaking.) I could make this print-out prettier if I had time and cared enough.

```
for (i in seq(10)) {
  cat("Before:\n")
  # write.table() prints to the console by default and lets you hide headers
  write.table(
    subset(auto_results, participantID == i),
    col.names = FALSE,
    row.names = FALSE,
    quote = 2
  )
  cat("After:\n")
  tmp <- subset(curated, participantID == i)
  write.table(
    tmp[, -ncol(tmp)],
    col.names = FALSE,
    row.names = FALSE,
    quote = 2
  )
  cat("\n")
}
```

```
Before:
1 "Ai and Neuroscience" Life Sciences Neuroscience and Neurobiology Systems Neuroscience
After:
1 "Ai and Neuroscience" Life Sciences Neuroscience and Neurobiology Computational Neuroscienc

Before:
2 "Plant Biology" Life Sciences Plant Sciences Plant Biology
2 "Ecology" Life Sciences Animal Sciences Zoology
2 "Ecology" Medicine and Health Sciences Medical Specialties Oncology
2 "Ecology" Medicine and Health Sciences Medical Specialties Urology
2 "Ecology" Physical Sciences and Mathematics Earth Sciences Geology
After:
2 "Plant Biology" Life Sciences Plant Sciences Plant Biology
2 "Ecology" Life Sciences Ecology and Evolutionary Biology NA

Before:
3 "Digital Humanities" Arts and Humanities Digital Humanities NA
3 "History" Arts and Humanities History NA
After:
3 "Digital Humanities" Arts and Humanities Digital Humanities NA
3 "History" Arts and Humanities History NA

Before:
4 "Computer Science" Physical Sciences and Mathematics Computer Sciences NA
4 "Neuroscience" Medicine and Health Sciences Medical Sciences Neurosciences
After:
4 "Computer Science" Physical Sciences and Mathematics Computer Sciences NA
4 "Neuroscience" Medicine and Health Sciences Medical Sciences Neurosciences

Before:
5 "Evolutionary Genomics" Education Education Economics NA
5 "Evolutionary Genomics" Life Sciences Genetics and Genomics NA
After:
5 "Evolutionary Genomics" Life Sciences Genetics and Genomics NA

Before:
6 "Medical Imaging" Medicine and Health Sciences Medical Sciences Medical Anatomy
6 "Vision Science" Engineering Biomedical Engineering and Bioengineering Vision Science
After:
6 "Medical Imaging" Medicine and Health Sciences Medical Sciences NA
6 "Vision Science" Engineering Biomedical Engineering and Bioengineering Vision Science

Before:
```

```
7 "Physics" Physical Sciences and Mathematics Physics NA
After:
7 "Physics" Physical Sciences and Mathematics Physics NA

Before:
8 "Linguistics" Social and Behavioral Sciences Linguistics NA
After:
8 "Linguistics" Social and Behavioral Sciences Linguistics NA

Before:
9 "Information Science" Physical Sciences and Mathematics Computer Sciences Information Secu
9 "Information Science" Social and Behavioral Sciences Library and Information Science Inform
After:
9 "Information Science" Social and Behavioral Sciences Library and Information Science NA

Before:
10 "Geography" Social and Behavioral Sciences Geography NA
After:
10 "Geography" Social and Behavioral Sciences Geography NA
```

We see that the errors have been corrected.

We don't need this column anymore:

```r
curated <- curated %>% select(-c("decision"))
```

Finally, let's add those additional 4 rows that I hard-coded at the top of this document.

```r
# Create a "decision" column that just reflects row number
# so that we can join the decision col in additions with that of tax
tmp_tax <- tax %>% mutate(decision = row_number())

new_rows <- additions %>%
  mutate(decision = suppressWarnings(as.integer(decision))) %>%
  left_join(curated, by = "participantID") %>%
  # we're just grabbing the response col from `curated`
  select(-c("Level1", "Level2", "Level3")) %>%
  # grab the Level cols from tax, at the row specified in decision
  left_join(tmp_tax, by = "decision") %>%
  select(participantID, response, Level1, Level2, Level3)

curated <- bind_rows(curated, new_rows)
curated <- curated %>% arrange(participantID)

subset(curated, participantID == 16)
```

```
   participantID                      response                           Level1
20            16 Environmental Data Science Physical Sciences and Mathematics
21            16 Environmental Data Science Physical Sciences and Mathematics
                     Level2 Level3
20 Environmental Sciences   <NA>
21           Data Science   <NA>
```

Ok! Now we finally have the qualitative responses smooshed and zhoozhed into our taxonomy.

Change "Artificial Intelligence and Robotics" to "AI and Machine Learning". We gave pure "Robotics" to those answers that mentioned it. Nobody actually entered AI AND robotics, so all occurences of this are just AI.

```r
curated$Level3[
  curated$Level3 == "Artificial Intelligence and Robotics"
] <- "AI and Machine Learning"
```

Let's save this to a file.

```r
write_df_to_file(curated, "curation_of_disciplines/qual_disciplines_final.tsv")
```

Analysis

Did we end up with one label (taxonomy row) per participant?

```r
nrow(curated)
```

```
[1] 195
```

```r
max(curated$participantID)
```

```
[1] 174
```

No. Some people listed more than one discipline.

```
more_than_one <- subset(as.data.frame(table(curated$participantID)), Freq > 1)
nrow(more_than_one)
```

```
[1] 17
```

17 people entered more than one discipline. One enthusiastic respondent listed six. All others entered two.

Next, I'd like to know how many participants identified with each unique option for Level1. So how many people had "Life Sciences" for Level1, how many people had "Arts and Humanities", etc.

```
level1_counts <- curated %>%
  distinct(participantID, Level1) %>% # one row per person per Level1
  count(Level1, name = "n_participants") %>%
  arrange(desc(n_participants))

level1_counts
```

```
                             Level1 n_participants
1 Physical Sciences and Mathematics             79
2                     Life Sciences             28
3                       Engineering             24
4    Social and Behavioral Sciences             23
5      Medicine and Health Sciences             21
6               Arts and Humanities              5
7                         Education              2
8                               Law              2
```

Neat! Now let's do the same for the other two levels.

```
level2_counts <- curated %>%
  distinct(participantID, Level2) %>% # one row per person per Level1
  count(Level2, name = "n_participants") %>%
  arrange(desc(n_participants))

level2_counts
```

```
                                                    Level2 n_participants
1                                        Computer Sciences             38
2                                         Medical Sciences             17
3                         Ecology and Evolutionary Biology             10
4                                     Computer Engineering              7
5                               Statistics and Probability              7
6                                    Genetics and Genomics              6
7                                              Mathematics              6
8                                                  Physics              6
9                                                     <NA>              6
10                                     Applied Mathematics              5
11                                               Chemistry              5
12                                            Data Science              5
13                                  Mechanical Engineering              5
14                                          Earth Sciences              4
15                                  Environmental Sciences              4
16                                                 History              4
17                         Library and Information Science              4
18                       Materials Science and Engineering              4
19                                          Bioinformatics              3
20                     Electrical and Computer Engineering              3
21                                              Psychology              3
22                              Astrophysics and Astronomy              2
23        Biochemistry, Biophysics, and Structural Biology              2
24                                            Biodiversity              2
25               Biomedical Engineering and Bioengineering              2
26                                           Communication              2
27                                               Economics              2
28                                               Geography              2
29                                             Linguistics              2
30                                     Medical Specialties              2
31   Oceanography and Atmospheric Sciences and Meteorology              2
32                                       Political Science              2
33 Public Affairs, Public Policy and Public Administration              2
34                                               Sociology              2
35                                   Aerospace Engineering              1
36                                             Agriculture              1
37                                            Anthropology              1
38    Bilingual, Multilingual, and Multicultural Education              1
39                                                 Biology              1
40                     Civil and Environmental Engineering              1
41                                      Digital Humanities              1
42                       Engineering Science and Materials              1
```

```
43                                          Marine Biology               1
44                                            Microbiology               1
45                                                   Music               1
46                                          Nanotechnology               1
47                           Neuroscience and Neurobiology               1
48                                          Plant Sciences               1
```

Cool! I think Level 2 will be most useful, but let's take a look at level 3.

```
level3_counts <- curated %>%
  distinct(participantID, Level3) %>% # one row per person per Level1
  count(Level3, name = "n_participants") %>%
  arrange(desc(n_participants))

level3_counts
```

```
                                        Level3 n_participants
1                                         <NA>            122
2                                Neurosciences             14
3                      AI and Machine Learning              7
4            Computer and Systems Architecture              3
5                                Biostatistics              2
6                        Computational Biology              2
7                                     Genomics              2
8                    Geophysics and Seismology              2
9                                 Oceanography              2
10                                    Robotics              2
11                        Software Engineering              2
12                                  Biophysics              1
13                  Computational Neuroscience              1
14               Critical and Cultural Studies              1
15           Databases and Information Systems              1
16         Demography, Population, and Ecology              1
17              Dynamics and Dynamical Systems              1
18                                Econometrics              1
19                     Environmental Chemistry              1
20                                   Evolution              1
21                              Fluid Dynamics              1
22                                    Genetics              1
23                       Geometry and Topology              1
24                               Health Policy              1
25                                   Hydrology              1
```

```
26                          Medical Biophysics              1
27                           Molecular Biology              1
28                                  Musicology              1
29          Natural Resources and Conservation              1
30                                   Neurology              1
31 Numerical Analysis and Scientific Computing              1
32                                Oral History              1
33                               Plant Biology              1
34                     Plasma and Beam Physics              1
35                                   Radiology              1
36                                Soil Science              1
37                              Space Vehicles              1
38                               Urban Studies              1
39                              Vision Science              1
```

Acutally, Level 3 is pretty interesting, too.

Let's do a little curation. First, let's remove those NA rows. These were rows where the participant's input matched at level 1 or 2, but not all three levels, so there was an NA for levels 2 and/or 3.

```
level2_counts <- level2_counts %>%
    filter(!is.na(Level2))
level3_counts <- level3_counts %>%
    filter(!is.na(Level3))
```

Let's rearrange the columns a bit for readability, and put these all back into one data frame.

```
names(level1_counts)[1] <- "Discipline"
names(level2_counts)[1] <- "Discipline"
names(level3_counts)[1] <- "Discipline"

level1_counts <- cbind("Level 1", level1_counts)
level2_counts <- cbind("Level 2", level2_counts)
level3_counts <- cbind("Level 3", level3_counts)

names(level1_counts)[1] <- "Taxonomy Level"
names(level2_counts)[1] <- "Taxonomy Level"
names(level3_counts)[1] <- "Taxonomy Level"

final_data <- rbind(level1_counts, level2_counts, level3_counts)
```

I'm just going to leave them as a table for now.

```r
write_df_to_file(final_data, "curation_of_disciplines/qual_disciplines_summary.tsv")
```

```r
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS 26.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;

locale:
[1] C.UTF-8/C.UTF-8/C.UTF-8/C/C.UTF-8/C.UTF-8

time zone: America/Los_Angeles
tzcode source: internal

attached base packages:
[1] tools     grid      stats     graphics  grDevices datasets  utils
[8] methods   base

other attached packages:
 [1] treemapify_2.5.6      tidyr_1.3.1           svglite_2.2.1
 [4] stringr_1.5.1         scales_1.4.0          readr_2.1.5
 [7] pwr_1.3-0             patchwork_1.3.2       ordinal_2023.12-4.1
[10] lme4_1.1-37           Matrix_1.7-1          languageserver_0.3.16
[13] here_1.0.1            gtools_3.9.5          ggforce_0.5.0
[16] FSA_0.10.0            fpc_2.2-13            forcats_1.0.0
[19] factoextra_1.0.7      ggplot2_3.5.2         emmeans_1.11.2
[22] dplyr_1.1.4           corrplot_0.95         ComplexHeatmap_2.22.0
[25] cluster_2.1.8.1       BiocManager_1.30.26

loaded via a namespace (and not attached):
 [1] Rdpack_2.6.4        rlang_1.1.6         magrittr_2.0.3
 [4] clue_0.3-66         GetoptLong_1.0.5    matrixStats_1.5.0
 [7] compiler_4.4.2      flexmix_2.3-20      systemfonts_1.2.3
[10] png_0.1-8           callr_3.7.6         vctrs_0.6.5
[13] pkgconfig_2.0.3     shape_1.4.6.1       crayon_1.5.3
[16] fastmap_1.2.0       rmarkdown_2.29      ggfittext_0.10.2
[19] tzdb_0.5.0          ps_1.9.1            nloptr_2.2.1
[22] purrr_1.1.0         xfun_0.53           modeltools_0.2-24
```

```
[25] jsonlite_2.0.0      tweenr_2.0.3        parallel_4.4.2
[28] prabclus_2.3-4      R6_2.6.1            stringi_1.8.7
[31] RColorBrewer_1.1-3  boot_1.3-31         diptest_0.77-2
[34] numDeriv_2016.8-1.1 estimability_1.5.1  Rcpp_1.1.0
[37] iterators_1.0.14    knitr_1.50          IRanges_2.40.1
[40] splines_4.4.2       nnet_7.3-19         tidyselect_1.2.1
[43] yaml_2.3.10         doParallel_1.0.17   codetools_0.2-20
[46] processx_3.8.6      lattice_0.22-6      tibble_3.3.0
[49] withr_3.0.2         evaluate_1.0.4      polyclip_1.10-7
[52] xml2_1.4.0          circlize_0.4.16     mclust_6.1.1
[55] kernlab_0.9-33      pillar_1.11.0       renv_1.1.5
[58] foreach_1.5.2       stats4_4.4.2        reformulas_0.4.1
[61] generics_0.1.4      rprojroot_2.1.1     S4Vectors_0.44.0
[64] hms_1.1.3           minqa_1.2.8         xtable_1.8-4
[67] class_7.3-22        glue_1.8.0          robustbase_0.99-4-1
[70] mvtnorm_1.3-3       rbibutils_2.3       colorspace_2.1-1
[73] nlme_3.1-166        cli_3.6.5           textshaping_1.0.1
[76] gtable_0.3.6        DEoptimR_1.1-4      digest_0.6.37
[79] BiocGenerics_0.52.0 ucminf_1.2.2        ggrepel_0.9.6
[82] rjson_0.2.23        farver_2.1.2        htmltools_0.5.8.1
[85] lifecycle_1.0.4     GlobalOptions_0.1.2 MASS_7.3-61
```