Getting Connected to your Data – A Reproducible Workflow for Data Wrangling

Jean-Paul Courneya
HELLO
my name is

Jean-Paul

jpcourneya@hshsl.umd.edu
@jpcourneya

Slides/Recording:
https://www.umaryland.edu/ictr/education-and-training/ictr-enrichment-series/
Learning Goals

• Data Wrangling
• Tidy data
• Work-flow efficiency boosting
Current situation for the novel coronavirus starting from Wuhan, China

Feature Layer by CSSE_GISandData


https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

https://github.com/CSSEGISandData/COVID-19/blob/master/README.md

https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
Welcome to the Healthcare Acquired Infection (HAI) Clinical Dashboard!

The information listed here is based on insertion and maintenance electronic health record flowsheet documentation of central lines and urinary catheters during September 1st to November 30th, 2017 from five critical care units. If you have any questions or comments, please email Dr. Ronald Piscotty, PhD, RN-BD, FAMIA at piscotty@umaryland.edu.
Value

• Near-term and long-term
• Indirect and direct
DATA WRANGLING
Data Wrangling

1. Make data suitable to use with a particular piece of software
2. Reveal information

- Munging
- Transformation
- Manipulation
Data Wrangling Workflow

Workflow Step 0: Access

- Where your data comes from
- How it’s organized
Transform: changing the form of the data, adding new values, fixing irregularities
Profile: summarizing the values of variables across records, validating individual records
Workflow Step 2: Publish

- Finished dataset that is used for the data product
- Script to wrangle the data
- Data Dictionary or other metadata presentation
Data Wrangling Tools

Excel

R

TRIFACTA

Python

OpenRefine
Hacking!


- **My guide** [https://guides.hshsl.umd.umaryland.edu/bioinformation/dataWrangling](https://guides.hshsl.umd.umaryland.edu/bioinformation/dataWrangling)
TIDY DATA
### Storms

<table>
<thead>
<tr>
<th>Storm</th>
<th>Wind (mph)</th>
<th>Pressure (hPa)</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberto</td>
<td>110</td>
<td>1007</td>
<td>2000-08-12</td>
</tr>
<tr>
<td>Alex</td>
<td>45</td>
<td>1009</td>
<td>1998-07-30</td>
</tr>
<tr>
<td>Allison</td>
<td>65</td>
<td>1005</td>
<td>1995-06-04</td>
</tr>
<tr>
<td>Ana</td>
<td>40</td>
<td>1013</td>
<td>1997-07-01</td>
</tr>
<tr>
<td>Arlene</td>
<td>50</td>
<td>1010</td>
<td>1999-06-13</td>
</tr>
<tr>
<td>Arthur</td>
<td>45</td>
<td>1010</td>
<td>1996-06-21</td>
</tr>
</tbody>
</table>

### Cases

<table>
<thead>
<tr>
<th>Country</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>7000</td>
<td>6900</td>
<td>7000</td>
</tr>
<tr>
<td>DE</td>
<td>5800</td>
<td>6000</td>
<td>6200</td>
</tr>
<tr>
<td>US</td>
<td>15000</td>
<td>14000</td>
<td>13000</td>
</tr>
</tbody>
</table>

- Storm name
- Wind speed (mph)
- Air pressure
- Date

- Country
- Year
- Count

[Link to RStudio webinar on data wrangling](https://rstudio.com/resources/webinars/data-wrangling-with-r-and-rstudio/)
variables

observations

values

https://r4ds.had.co.nz/
### cases

<table>
<thead>
<tr>
<th>Country</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>7000</td>
<td>6900</td>
<td>7000</td>
</tr>
<tr>
<td>DE</td>
<td>5800</td>
<td>6000</td>
<td>6200</td>
</tr>
<tr>
<td>US</td>
<td>15000</td>
<td>14000</td>
<td>13000</td>
</tr>
</tbody>
</table>

```python
gather(cases, "year", "n", 2:4)
```
OH NO! R Indoctrniation

https://www.storybench.org/getting-started-with-tidyverse-in-r/
REDCap

- Data collection instruments
- Data dictionaries
- De-Identifying Data
- Exports to my favorite statistical programming software!
Untidy Data

• Performance or space advantages
• Specialized fields have their own data structures.
WORK-FLOW EFFICIENCY BOOSTING
Data Wrangling Workflow

Access -> Transform/Profile -> Publish
Reuse

• Will I repeat this analysis?
• Will I want to see anything else from the data?
• Will want to add more data to the analysis?
• Will I retool the data for another piece of software?
```r
# [r_wrap-hook, include=FALSE]
library(knitr)
hook_output = knit_hooks$get('output')
knit_hooks$set(output = function(x, options) {
  # this hook is used only when the linewidth option is not NULL
  if (!is.null(n <- options$linewidth)) {
    x = knitr::split_lines(x)
    # any lines wider than n should be wrapped
    if (any(nchar(x) > n)) x = strwrap(x, width = n)
    x = paste(x, collapse = '\n')
  }
  hook_output(x, options)
})
```

```r
# [r_calculateSums, linewidth=60]
#setwd("/Volumes/ATAMAS/RNAseq_RGCC_Dec_2018_row2")
md5SumsCalculated <- as.vector(sapply(list.files(pattern = "
.gz"), recursive = T), md5sum)
```

```r
# [r_md5TextFiles, linewidth=60]
txtFileNames <- list.files(pattern = "md5.txt", recursive = T)
md5SumsTxtTable <- dplyr::bind_rows(lapply(txtFileNames, read.table))
names(md5SumsTxtTable) <- c("md5SumsTxtFile", "readFileNames")
md5SumsTxtTable$md5FileNames <- txtFileNames
table(out <- cbind(md5SumsCalculated, md5SumsTxtTable))
write.csv(out, file = "md5Sums.csv", row.names = FALSE)
```

<table>
<thead>
<tr>
<th>md5SumsCalculated</th>
<th>md5SumsTxtFile</th>
<th>readFileNames</th>
<th>txtFileNames</th>
</tr>
</thead>
<tbody>
<tr>
<td>4e6358d05ced5d6bd6e734ac3fe41996</td>
<td>4e6358d05ced5d6bd6e734ac3fe41996</td>
<td>index5_ACAGTG_L001-L002_R1_001.fastq.gz</td>
<td>11-10-2018/Index5_OtC3275/ /index5_ACAGTG_L001-L002_</td>
</tr>
</tbody>
</table>
3 pairwise comparisons

R-HSA-74752: Signaling by Insulin receptor
GO:0006913: nucleocytoplasmic transport
R-HSA-2262752: Cellular responses to stress
R-HSA-8953854: Metabolism of RNA
GO:1903311: regulation of mRNA metabolic process
GO:0019080: viral gene expression
R-HSA-5683057: MAPK family signaling cascades
GO:0042176: regulation of protein catabolic process
GO:0070482: response to oxygen levels
R-HSA-9006934: Signaling by Receptor Tyrosine Kinases
GO:0070997: neuron death
hsa04010: MAPK signaling pathway
GO:0070371: ERK1 and ERK2 cascade
hsa04151: PI3K-Akt signaling pathway
M81: PID CDC42 PATHWAY
GO:0120035: regulation of plasma membrane bounded cell
hsa05215: Prostate cancer
R-HSA-449147: Signaling by Interleukins
GO:0018107: peptidyl-threonine phosphorylation
GO:0046777: protein autophosphorylation

https://metascape.org/gp/index.html#/main/step1
3 different comparisons

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
<th>log2FoldChange</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000175756</td>
<td>AURKAI1P1</td>
<td>0.027166489</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000223663</td>
<td>NDUF4P8</td>
<td>0.084987209</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000221978</td>
<td>CCNL2</td>
<td>0.0341701876</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG000002224870</td>
<td>RP4-758J18.2</td>
<td>-0.0341701876</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000242485</td>
<td>MRPL20</td>
<td>0.0687084888</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000264293</td>
<td>RN7SL57P</td>
<td>3.5696638226</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000272455</td>
<td>RP4-758J18.13</td>
<td>0.3607144603</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000235098</td>
<td>ANKRD65</td>
<td>-0.2969713177</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000225905</td>
<td>RP4-758J18.7</td>
<td>-1.922169459</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000205116</td>
<td>TMEM88B</td>
<td>1.8098115484</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000225285</td>
<td>RP4-758J18.10</td>
<td>1.2761029137</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000179403</td>
<td>VWA1</td>
<td>-0.848992225</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000215915</td>
<td>ATAD3C</td>
<td>0.5802762212</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000160072</td>
<td>ATAD3B</td>
<td>0.3091831212</td>
<td>NA</td>
</tr>
<tr>
<td>ENSG00000197785</td>
<td>ATAD3A</td>
<td>0.5426927733</td>
<td>NA</td>
</tr>
</tbody>
</table>

```r
setwd("/UserData/SAJPC/21-17-reanalysis-goRilla/GO_tables_2020_06_19_Analysis")
# saveRDS(res_R_vs_BGeneNames, "res_R_vs_BGeneNamesedb-v86.RDS")
# readRDS("res_R_vs_BGeneNamesedb-v86.RDS")
log2FoldChange_cutoff <- 0.5849625
pvalue_cutoff <- 0.05
RvB_GO <- res_R_vs_BGeneNames[order(res_R_vs_BGeneNames$log2FoldChange), ] #orders data according to pvalue
RvB_GO_Cutoff <- subset(RvB_GO,pvalue>pvalue_cutoff & abs(log2FoldChange) >= log2FoldChange_cutoff) # subset list for background of all genes with l2FC not NA
RvB_GO_BKGD <- RvB_GO[complete.cases(RvB_GO$log2FoldChange), ] # write file
write.csv(RvB_GO_BKGD, file="RvB_GO_BKGD.csv", row.names=FALSE) # subset NA from RvB_GO
RvB_GO_NA <- RvB_GO[!complete.cases(RvB_GO$log2FoldChange), ] # subset up regulated gene list
RvB_GO_UP <- subset(RvB_GO_Cutoff, log2FoldChange >= log2FoldChange_cutoff) # subset down regulated gene list
write.csv(RvB_GO_UP, file="RvB_GO_UP.csv", row.names=FALSE)
```
<table>
<thead>
<tr>
<th>GroupID</th>
<th>Category</th>
<th>Term</th>
<th>Description</th>
<th>LogP</th>
<th>Log(q-value)</th>
<th>InTerm</th>
<th>InGenes</th>
<th>Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_Summ</td>
<td>GO:00467</td>
<td>protein aut</td>
<td>-36.3285</td>
<td>-32.009</td>
<td>42/235</td>
<td>207,790:8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:00467</td>
<td>protein aut</td>
<td>-36.3285</td>
<td>-32.009</td>
<td>42/235</td>
<td>207,790:8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:003336</td>
<td>positive reg</td>
<td>-20.2855</td>
<td>-16.665</td>
<td>43/63</td>
<td>147,207,3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:004343</td>
<td>regulation</td>
<td>-20.0744</td>
<td>-16.533</td>
<td>47/754</td>
<td>147,207,3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:001811</td>
<td>peptidyl-ty</td>
<td>-18.6711</td>
<td>-15.255</td>
<td>33/371</td>
<td>351,975,1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:001821</td>
<td>peptidyl-ty</td>
<td>-18.5639</td>
<td>-15.199</td>
<td>33/374</td>
<td>351,975,1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:005131</td>
<td>positive reg</td>
<td>-18.2934</td>
<td>-15.016</td>
<td>43/692</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:004341</td>
<td>positive reg</td>
<td>-16.8015</td>
<td>-13.713</td>
<td>37/552</td>
<td>47,351,5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:004581</td>
<td>positive reg</td>
<td>-15.4256</td>
<td>-12.456</td>
<td>35/540</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:004343</td>
<td>regulation</td>
<td>-13.9846</td>
<td>-11.046</td>
<td>27/344</td>
<td>975,1845,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:003212</td>
<td>activation</td>
<td>-13.3668</td>
<td>-10.495</td>
<td>26/335</td>
<td>147,207,3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:003105</td>
<td>activation</td>
<td>-13.0264</td>
<td>-10.17</td>
<td>25/317</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:007198</td>
<td>regulation</td>
<td>-12.5671</td>
<td>-9.725</td>
<td>31/528</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:004343</td>
<td>regulation</td>
<td>-12.0322</td>
<td>-9.231</td>
<td>22/263</td>
<td>975,2260,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:007198</td>
<td>regulation</td>
<td>-11.4077</td>
<td>-8.65</td>
<td>24/343</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:003280</td>
<td>regulation</td>
<td>-10.0794</td>
<td>-7.442</td>
<td>19/239</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1_Member</td>
<td>GO:007030</td>
<td>regulation</td>
<td>-9.88623</td>
<td>-7.357</td>
<td>19/242</td>
<td>207,790:2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Type</th>
<th>Tax ID</th>
<th>Homologous Gene Symt</th>
<th>Description</th>
<th>Biological</th>
<th>Kinase Class</th>
<th>Protein</th>
<th>Fu Subcellular Drug</th>
<th>Drug Canonical Halo</th>
<th>Hallmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1314</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>COPA</td>
<td>COPI coat</td>
<td>g01929243 protein localization</td>
<td>Cytosol; Golgi apparatus</td>
<td>M24324</td>
<td>PIP3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>537</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>ATP6A1P</td>
<td>ATPase H+ t</td>
<td>g01929243 positive regulation</td>
<td>Cytosol; Microtubules; Plasma mem</td>
<td>M29005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22820</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>COPG1</td>
<td>COPI coat</td>
<td>g0051683 establishment of Golgi apparatus</td>
<td>S-Dimethylseryl (Cysteine)</td>
<td>M29100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9276</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>COPB2</td>
<td>COPI coat</td>
<td>g01929243 transcript</td>
<td>Cytosol</td>
<td>M29100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3725</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>JUN</td>
<td>Jun proto-oncogene</td>
<td>Nucleoplasm</td>
<td>Cytoplasm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10291</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>FGF1</td>
<td>Splicing factor</td>
<td>Nucleoplasm; Splicing site</td>
<td>Enhanced</td>
<td>M29260</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6625</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>SNRNP70</td>
<td>small nucle</td>
<td>Nucleoplasm</td>
<td>Enhanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9114</td>
<td>Gene_ID</td>
<td>H. sapiens</td>
<td>ATP6AV1</td>
<td>ATPase H+ t</td>
<td>g0090383 phagosome Transporters/Primary Active Transporters</td>
<td>M29282</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```r
library(readxl)
metascape_Enrichment <- read_excel("/UserData/SAPC/21-17-reanalysis/metaspacemetascape_result_combinedList.xlsx", sheet = "Enrichment")

metascape_Annotation <- read_excel("/UserData/SAPC/21-17-reanalysis/metaspacemetascape_result_combinedList.xlsx", sheet = "Annotation")

library(dplyr)
# Check annotation terms that have extracellular in them choose the term of interest
metascape_Enrichment[grepl("extracellular", metascape_Enrichment$Description, ignore.case = T), "Term"]

# Select targets corresponding to enrichment term of interest and variables of interest to build subset table
# filter gene list based on condition being member of the selected enrichment term of interest "R-HSA-1474244 Extracellular matrix organization"
ecmPathwayTargets <-
metascape_Annotation %>%
  select(Term = starts_with("R-HSA-1474244"), RTV_GO_DOWN, TVB_GO_UP, "Gene Symbol") %>%
  filter(Term == 1 & RTV_GO_DOWN == 1 & TVB_GO_UP == 1) %>%
  pull

library(readr)
# load LFC data for RTV DOWN and TVB UP
RTV_GO_DOWN <- read_csv("/UserData/SAPC/21-17-reanalysis/metaspase/RTV_GO_DOWN.csv")
TVB_GO_UP <- read_csv("/UserData/SAPC/21-17-reanalysis/metaspase/TVB_GO_UP.csv")

# make LFC tables
RTV_DOWN <- RTV_GO_DOWN RTV_GO_DOWN$name %>% ecmPathwayTargets, 2:3)
TVB_UP <- TVB_GO_UP$TVB_GO_UP$name %>% ecmPathwayTargets, 2:3]
combinedLFCTable <- rbind(RTV_DOWN, TVB_UP) %>% cbind(condition = rep("RTV_DOWN", TVB_UP), each = 27), stringsAsFactors = FALSE)

# Plot
library(ggplot2)
col <- c("RTV_DOWN" = "blue", "TVB_UP" = "red")
combinedLFCTable %>%
murate(Condition = forcats::fct_rev(name)) %>%
ggplot(aes(x = name, y = log2FoldChange, colour = condition)) +
  geom_segment(aes(xend = name, yend = 0), size = 0.8) +
  geom_hline(yintercept = 0,
    colour = "black",
    size = 1.0) +
  coord_flip() +
  scale_y_continuous(breaks = c(-6, -4, -2, 0, 2, 4)) +
  labs(x = "Gene Symbol", y = expression(paste("Log2(">2", "(Fold)"))) +
  scale_color_manual(values = cols, Condition, labels = c("TGF-β+RCC", "TGF-β")) +
  # scale_color_manual(values = cols, condition, labels = c(expression(paste("RCC", "+TGF-β")), "TGF-β")) +
  # theme_classic(base_size = 15) +
  theme_classic() +
  theme(text = element_text(color = "black", size = 20), axis.text = element_text(color = "black", size = 17))
```
Summary

• Wrangling data workflow is an iterative process.

• Tidy data is a worthwhile standard to know.

• Everyone can stand to gain more efficiency and value by thinking more deeply about what you're doing, even if that doesn’t mean learning a scripting language. I can help you with that!
CDABS
The Center for Data and Bioinformation Services

http://guides.hshsl.umd.edu/data
Voucher Program

Voucher program https://www.umaryland.edu/ictr/funding/voucher-program/
Data visualization @ CDABS https://guides.hshsl.umaryland.edu/dataVisualizationService
We’re done!

Remember to take the survey from the link!