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

Jean-Paul Courneya

11 August 2020

jpcourneya@hshsl.umaryland.edu
@jpcourneya

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

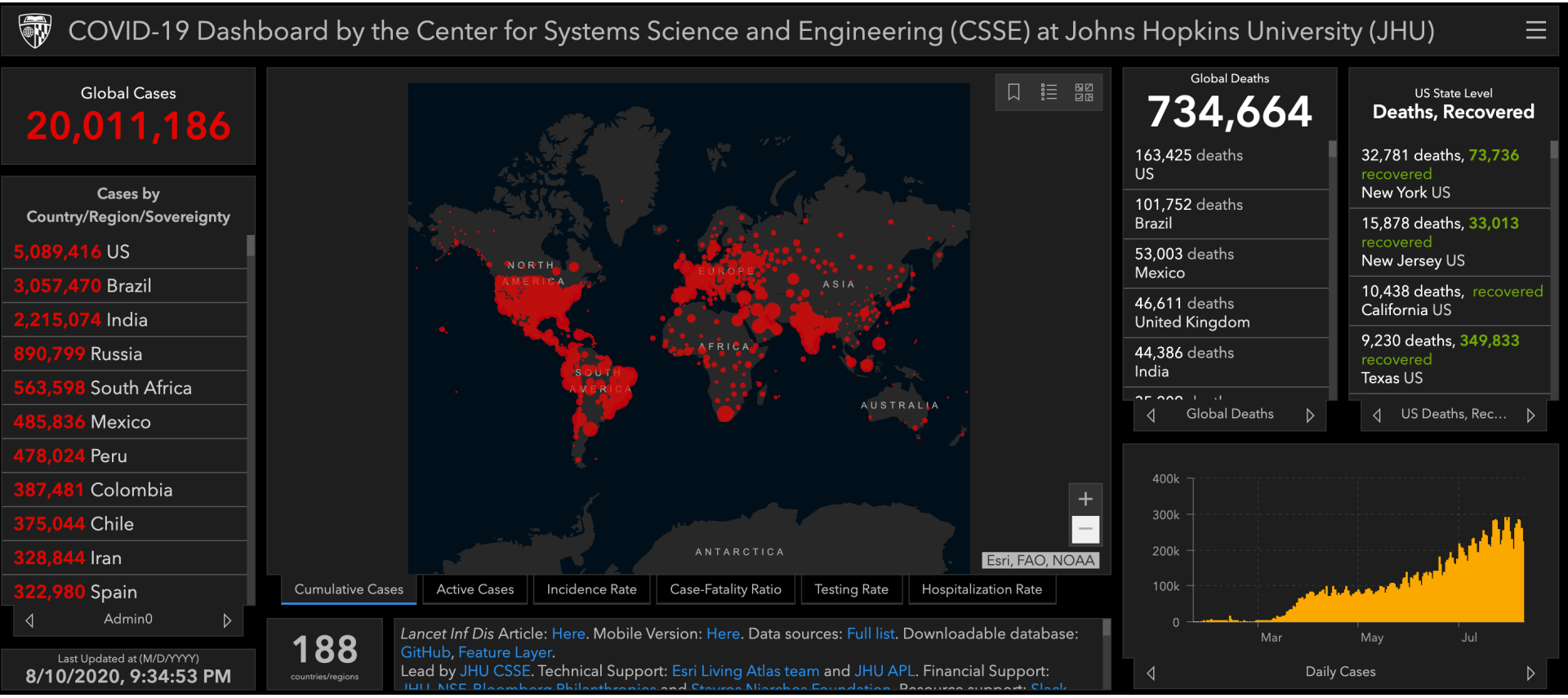
Learning Goals

- Data Wrangling
- Tidy data
- Work-flow efficiency boosting

Current situation for the noval coronavirus starting from Wuhan, China

Feature Layer by CSSE_GISandData

Created: Jan 25, 2020 Updated: Aug 11, 2020 View Count: 988,470,187



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

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

ICU Department

Default is all departments

- Central Line
- Urinary Catheter
- HAI Glossary

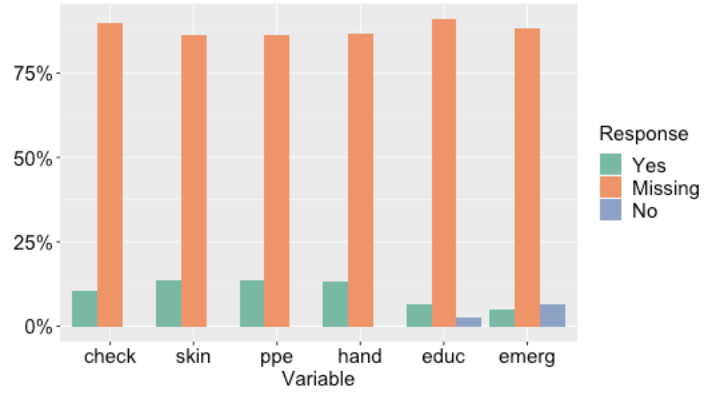
300
Number of Central Lines

4.8
Average Central Line Days

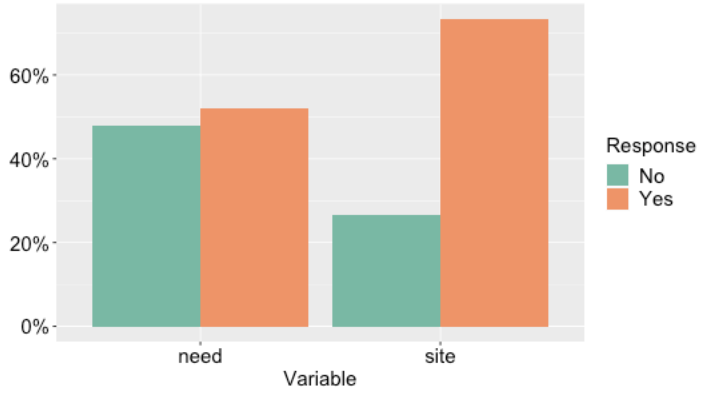
19.6
Average Length of Stay

7.9
ICU average length of stay

Placement — Table | Plot



Assessment — Plot



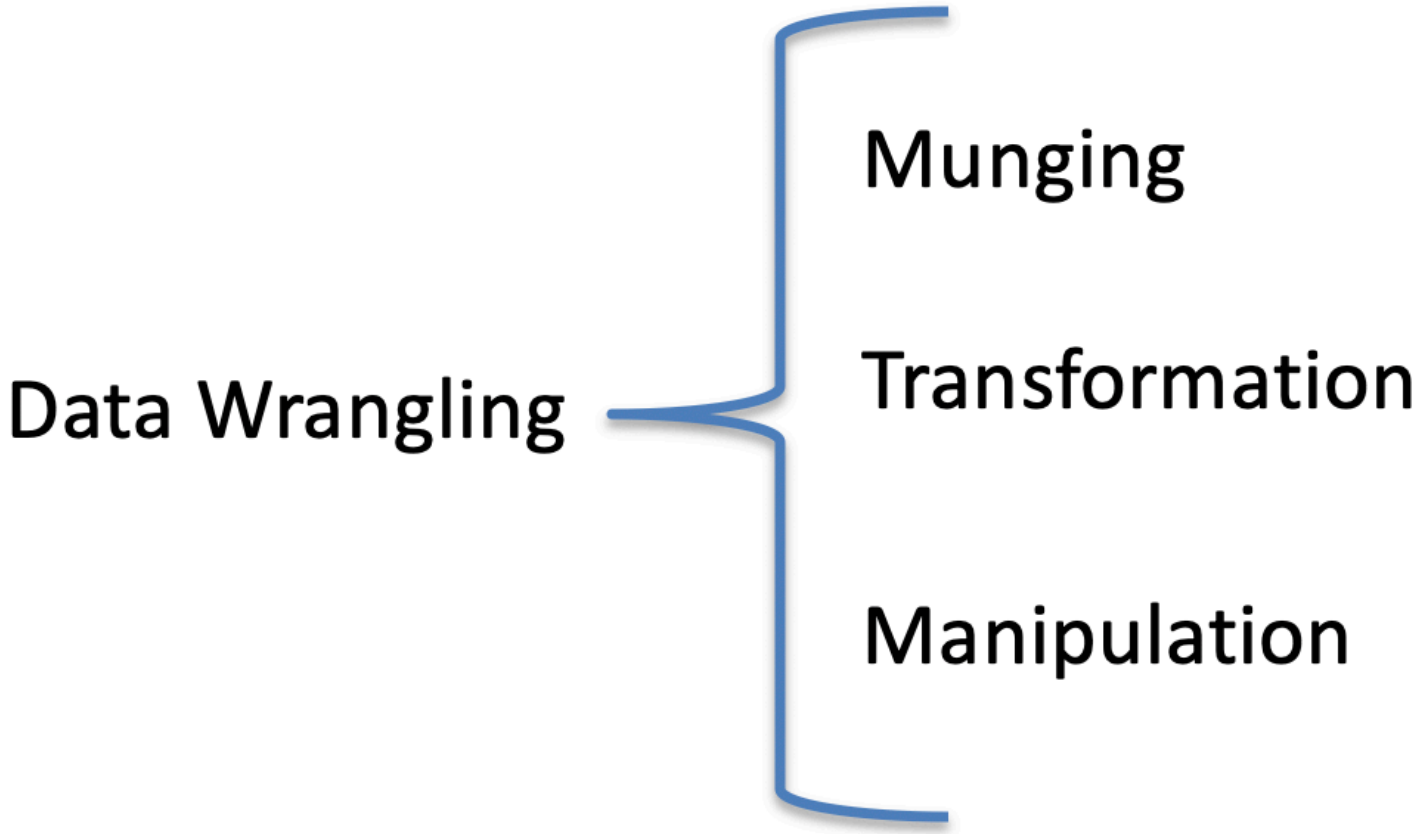
Welcome to the Healthcare Acquired Infection (HAI) Clinical Dashboard!

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

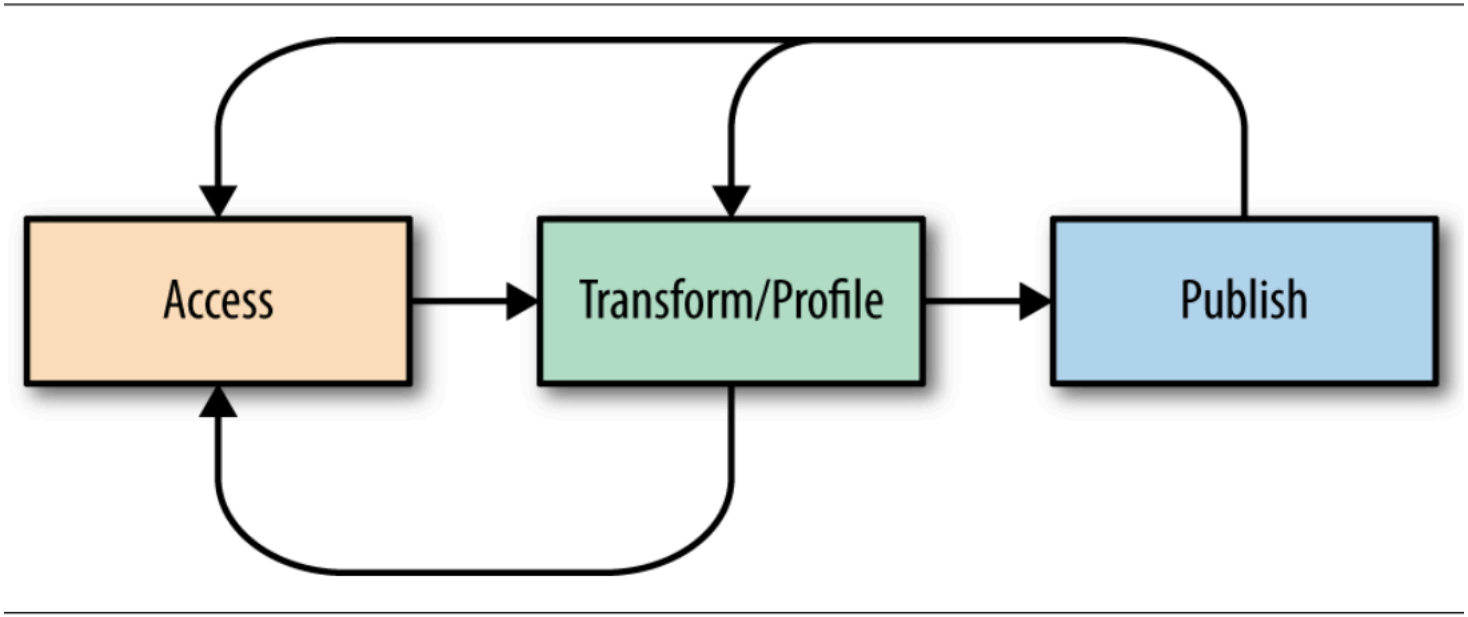
DATA WRANGLING



1 Make data suitable to use with a particular piece of software

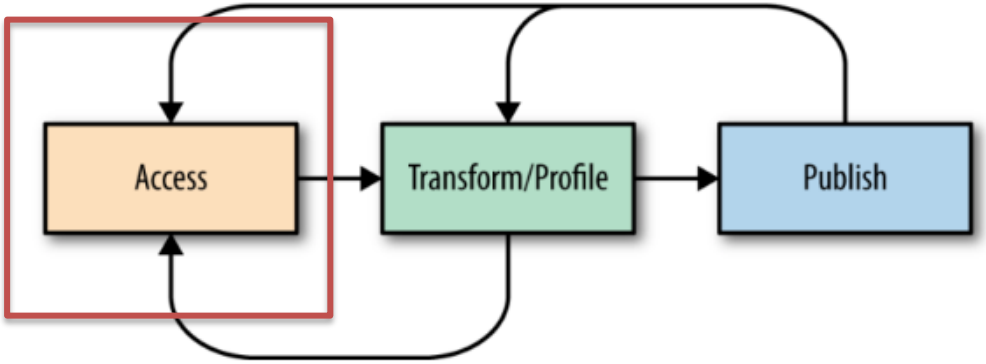
2 Reveal information

Data Wrangling Workflow



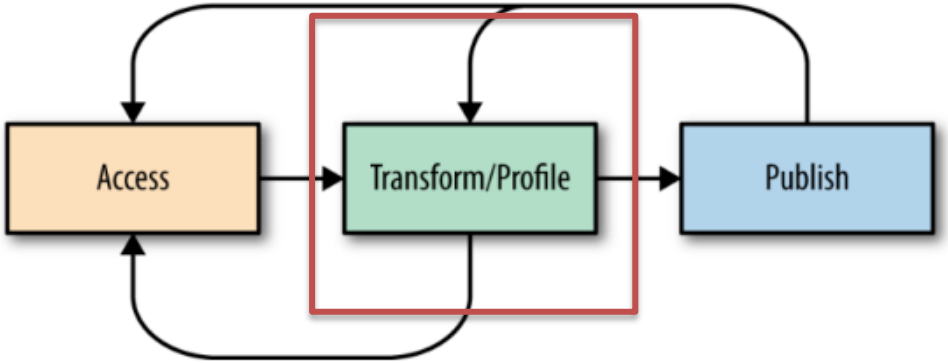
https://www.oreilly.com/library/view/principles-of-data/9781491938911/

Workflow Step 0: Access



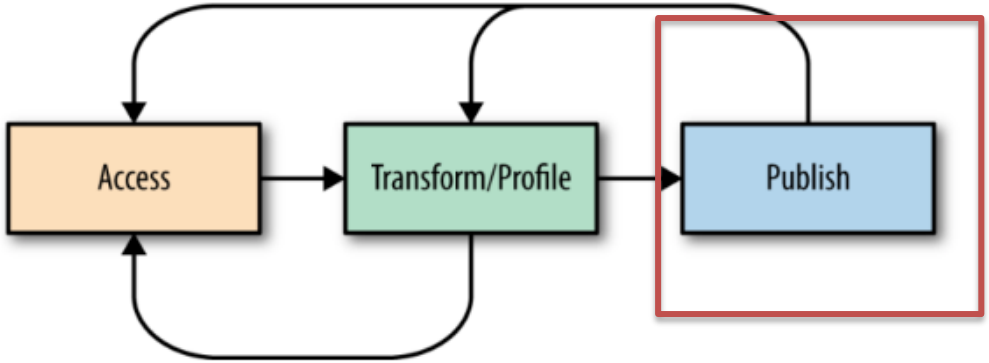
- Where your data comes from
- How it's organized

Workflow Step 1: Transform/Profile



- 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

TRIFACTA

python™

OpenRefine

Hacking!

- Excel example https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19?ui=en-us&rs=en-us&ad=us
- R example https://rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
- My guide https://guides.hshsl.umaryland.edu/bioinformation/dataWrangling

TIDY DATA

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

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

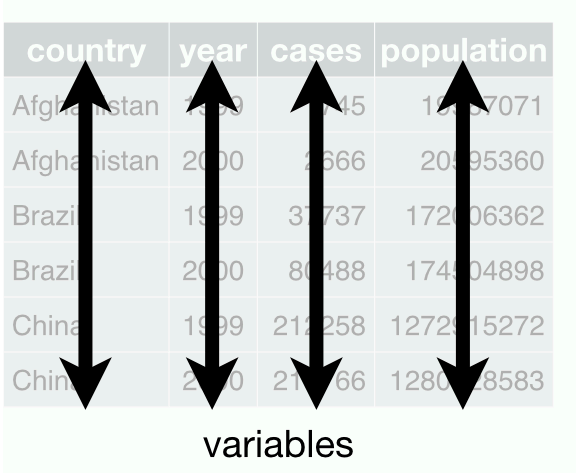
cases

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

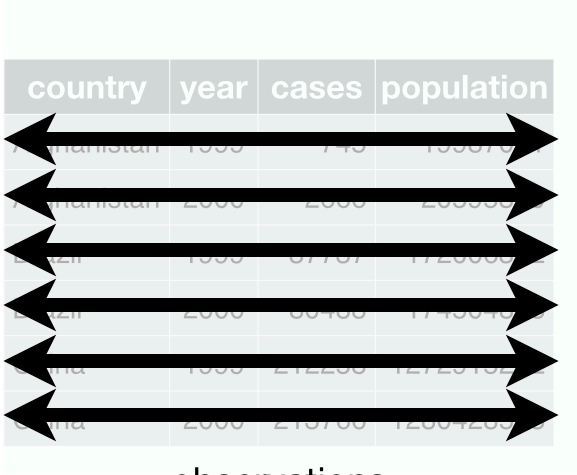
- Country
- Year
- Count

https://rstudio.com/resources/webinars/data-wrangling-with-r-and-rstudio/

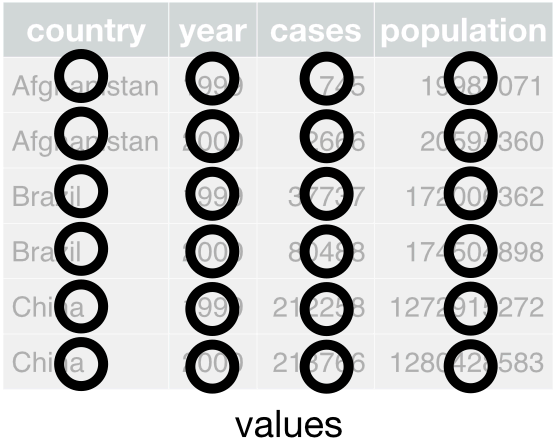
1



2



3



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

cases

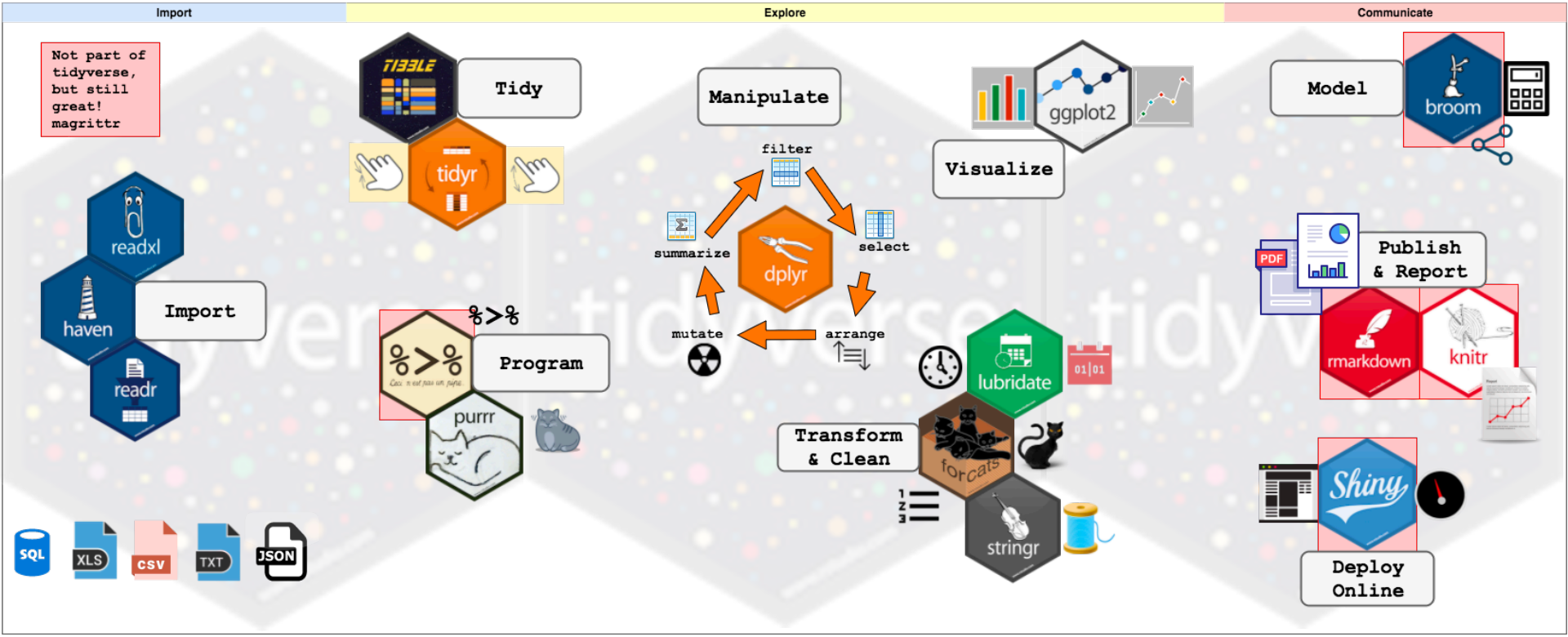
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

```
gather(cases, "year", "n", 2:4)
```

cases

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

OH NO! R Indoctrination



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

REDCap

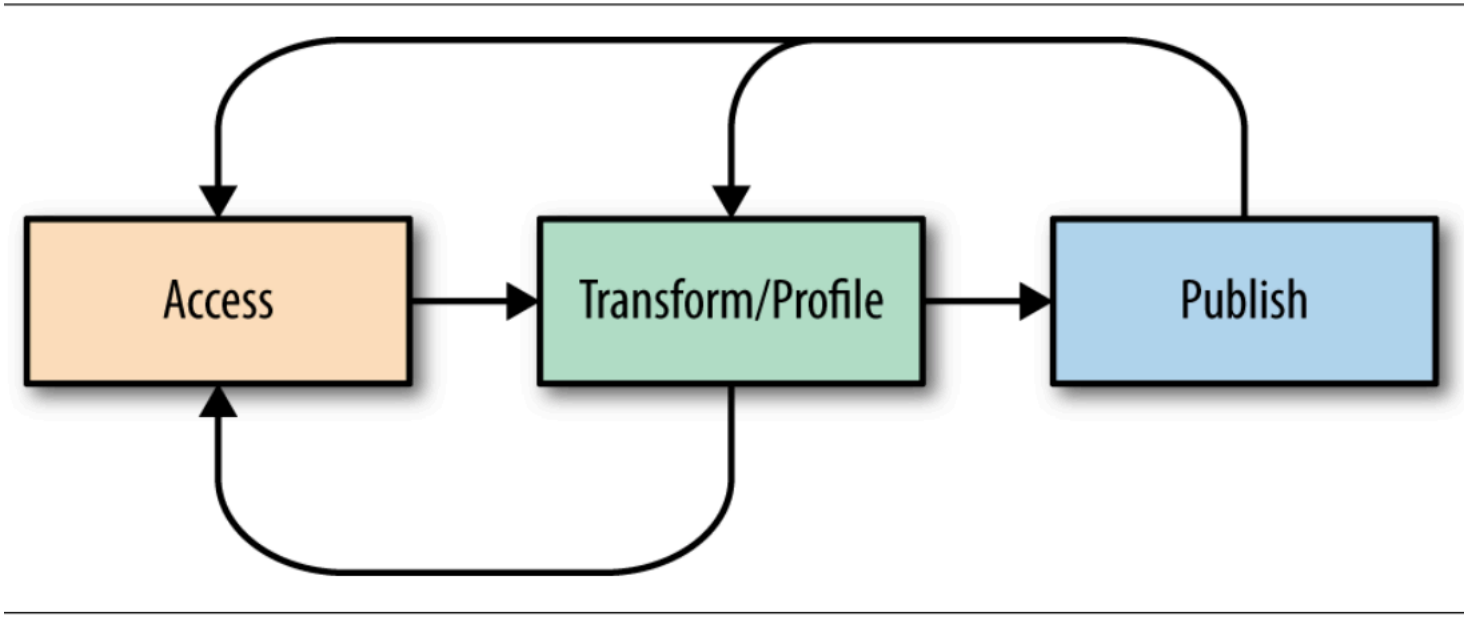
- Data collection instruments
- Data dictionaries
- De-Identifying Data
- Exports to my favorite statistical programming software!

Untidy Data

- Performance or space advantages
- Specialized fields have their own data structures.

WORK-FLOW EFFICIENCY BOOSTING

Data Wrangling Workflow



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?

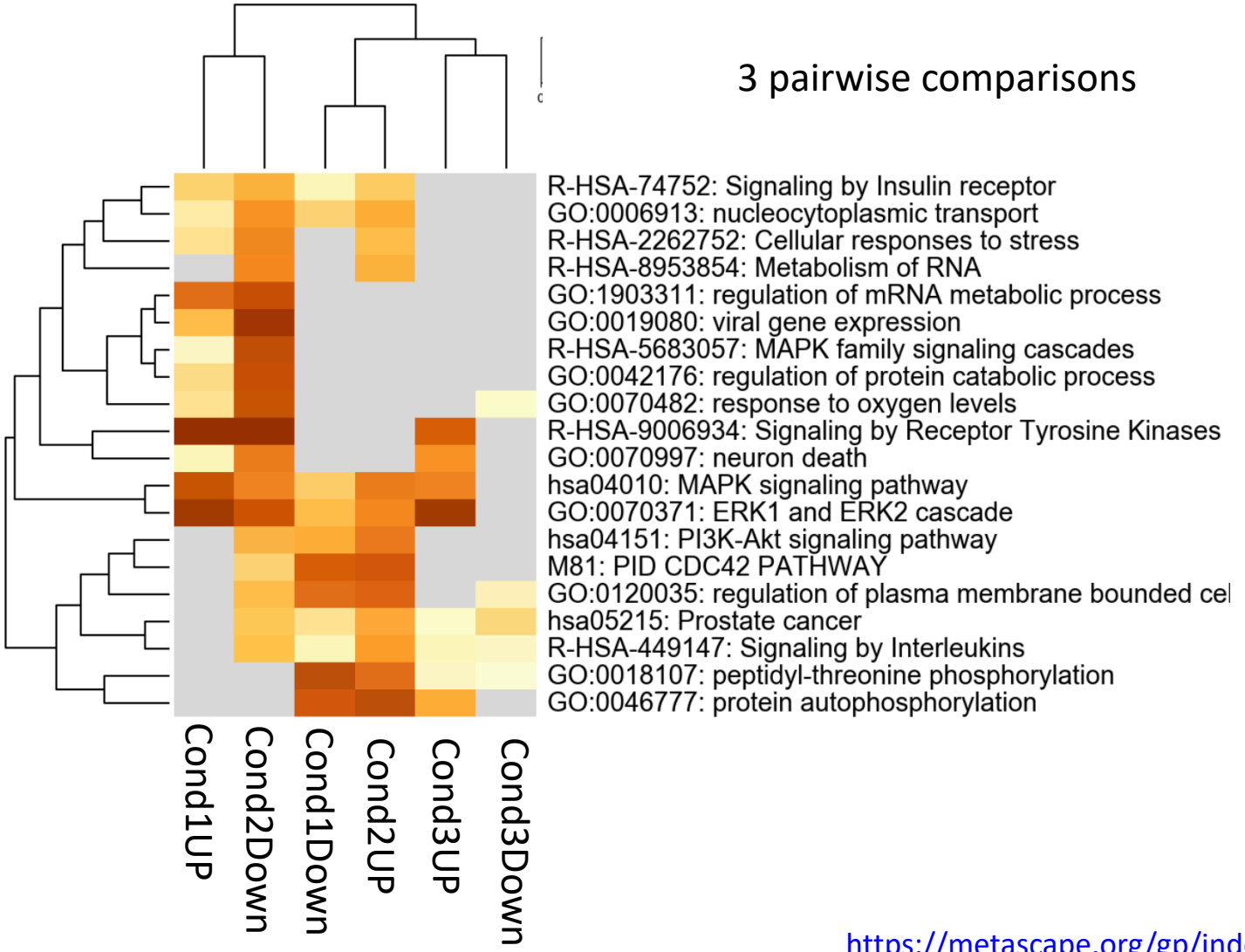
index5_ACAGTG_L001-L002_R1_001.1.fastq.gz	1.5 GB	Archive
index5_ACAGTG_L001-L002_R1_001.fastq.gz.md5.txt	73 bytes	Text
index5_ACAGTG_L001-L002_R1_001_fastqc.zip	295.1 kB	Archive
index5_ACAGTG_L001-L002_R2_001.1.fastq.gz	1.5 GB	Archive
index5_ACAGTG_L001-L002_R2_001.fastq.gz.md5.txt	73 bytes	Text
index5_ACAGTG_L001-L002_R2_001_fastqc.zip	283.0 kB	Archive

````
```{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 calculateSums, linewidth=60}
#setwd("/Volumes/ATAMAS/RNASeq_RGCC_Dec_2018_raw2")
md5SumsCalculated <- as.vector(sapply(list.files(pattern = ".gz$", recursive = T), md5sum))
```

```{r md5TextFiles, linewidth=60}
txtFileNames <- list.files(pattern = "md5.txt$", recursive = T)
md5SumsTxtTable <- dplyr::bind_rows(lapply(txtFileNames, read.table))
names(md5SumsTxtTable) <- c("md5SumsTxtFile", "readFileNames")
md5SumsTxtTable$txtFileNames <- txtFileNames
kable(out <- cbind(md5SumsCalculated, md5SumsTxtTable))
write.csv(out, file = "md5Sums.csv", row.names = FALSE)
```
````

md5SumsCalculated	md5SumsTxtFile	readFileNames	txtFileNames
4e6358d05ced5d6bd6e734ac3fe41996	4e6358d05ced5d6bd6e734ac3fe41996	index5_ACAGTG_L001-L002_R1_001.fastq.gz	11-10-2018/Index5_OtC3275_/index5_ACAGTG_L001-L002_

Metascape

A Gene Annotation & Analysis Resource



3 pairwise comparisons

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

ENSG00000175756	AURKAIP1	0.027166489
ENSG00000223663	NDUFB4P8	NA
ENSG00000221978	CCNL2	0.0849872079
ENSG00000224870	RP4-758J18.2	-0.0341701876
ENSG00000242485	MRPL20	0.0687084888
ENSG00000264293	RN7SL657P	3.5696638226
ENSG00000272455	RP4-758J18.13	0.3607144603
ENSG00000235098	ANKRD65	-0.2969713177
ENSG00000225905	RP4-758J18.7	-1.9221269459
ENSG00000205116	TMEM88B	1.8098115484
ENSG00000225285	RP4-758J18.10	1.2761021937
ENSG00000179403	VWA1	-0.848992225
ENSG00000215915	ATAD3C	-0.5802762212
ENSG00000160072	ATAD3B	0.3091831212
ENSG00000197785	ATAD3A	0.5425827722
ENSG00000205090	TMEM	
ENSG00000160075	SSU	

3 different comparisons

56136	ENSG00000288572	NA

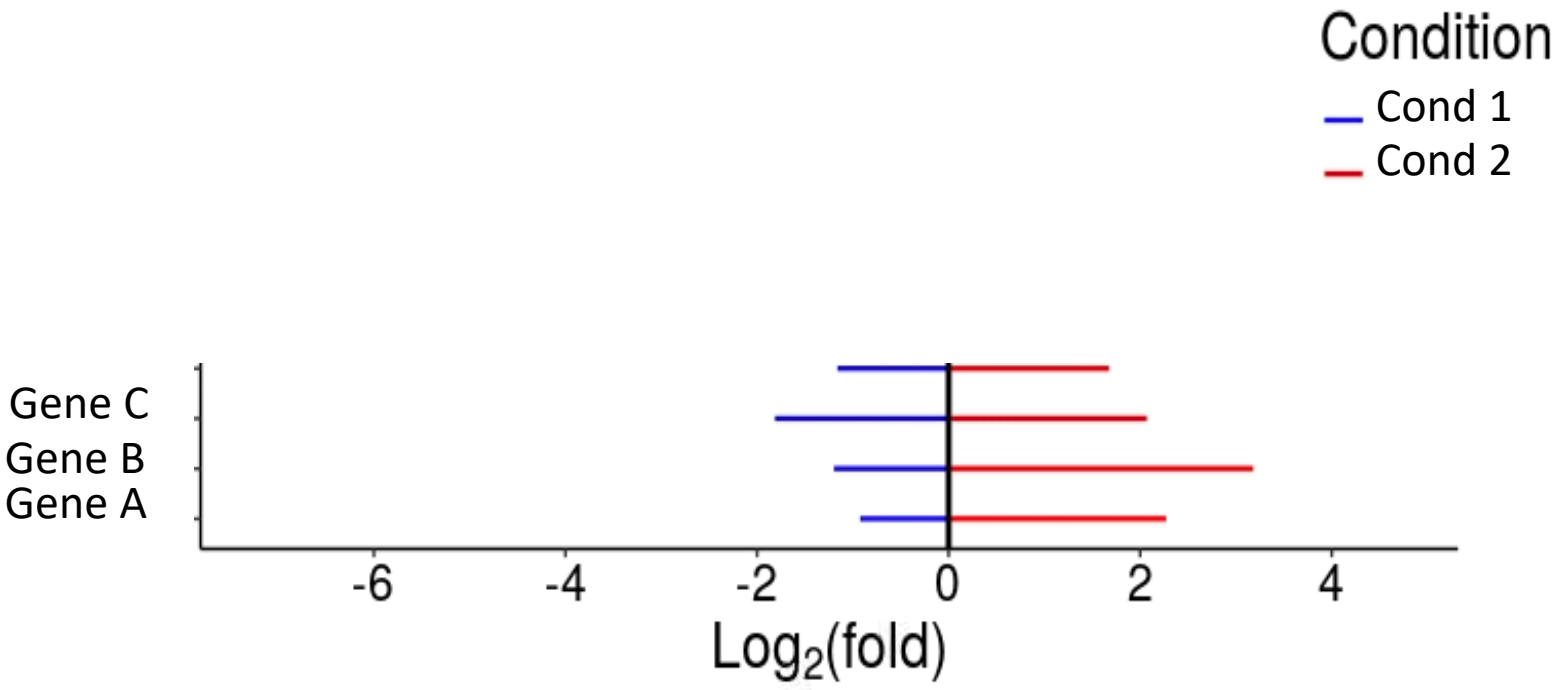
```r
setwd("/UserData/SAJPC/21-17-reanalysis/goRilla/GO_tables_2020_06_19_Analysis")
# saveRDS(res_R_vs_BGeneNames, "res_R_vs_BGeneNames_edb-v86.RDS")
# readRDS("res_R_vs_BGeneNames_edb-v86.RDS")
log2FoldChange_cutoff <- 0.5849625
pvalue_cutoff <- 0.05
RvB_GO <-res_R_vs_BGeneNames[order(res_R_vs_BGeneNames$log2FoldChange), ] #orders data according to pvalue
RvB_GO_CutOff <- subset(RvB_GO,pvalue<pvalue_cutoff & abs(log2FoldChange) >= log2FoldChange_cutoff)
# subset list for background of all genes with l2FC not NA
RvB_GO_BKGD <- RvB_GO[complete.cases(RvB_GO$log2FoldChange), ]
write.csv(RvB_GO_BKGD, file="RvB_GO_BKGD.csv", row.names=FALSE) #write file
# 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)
write.csv(RvB_GO_UP, file="RvB_GO_UP.csv", row.names=FALSE) #write file
# subset down regulated gene list
RvB_GO_Down <-subset(RvB_GO_CutOff, log2FoldChange <= log2FoldChange_cutoff)
write.csv(RvB_GO_Down, file="RvB_GO_Down.csv", row.names=FALSE) #write file
```

Gene List Report Excel Sheets | Gene List Report PPT file | All in One Zip File

GroupID	Category	Term	Description	LogP	Log(q-value	InTerm_In	Genes	Symbols
1_Summar	GO Biologi	GO:00467	protein aut	-36.3285	-32.009	42/235	207,790,8	AKT1,CAD,CAMK2B,CLK1,CSNK1G2,DAPK3,MARK2,ENG,EPHA7,EPHB4,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FRK,MTOR
1_Member	GO Biologi	GO:00467	protein aut	-36.3285	-32.009	42/235	207,790,8	AKT1,CAD,CAMK2B,CLK1,CSNK1G2,DAPK3,MARK2,ENG,EPHA7,EPHB4,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FRK,MTOR
1_Member	GO Biologi	GO:00336	positive reg	-20.2855	-16.665	43/613	207,975,2	AKT1,CD81,MARK2,EPHA7,EPHB2,EPHB4,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FPR1,MTOR,HRAS,HSP90AA1,HTR2A,JA
1_Member	GO Biologi	GO:00434	regulation	-20.0744	-16.533	47/754	147,207,3	ADRA1B,AKT1,APP,ATP6AP1,CD81,DUSP3,EPHA7,EPHB2,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FPR1,HRAS,HTR2A,JAK2
1_Member	GO Biologi	GO:00181	peptidyl-ty	-18.6711	-15.255	33/371	351,975,1	APP,CD81,CLK1,EPHA7,EPHB2,EPHB4,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FRK,MTOR,HTR2A,JAK2,LTK,NTRK1,NTRK2
1_Member	GO Biologi	GO:00182	peptidyl-ty	-18.5639	-15.199	33/374	351,975,1	APP,CD81,CLK1,EPHA7,EPHB2,EPHB4,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FRK,MTOR,HTR2A,JAK2,LTK,NTRK1,NTRK2
1_Member	GO Biologi	GO:00513	positive reg	-18.2934	-15.016	43/692	207,975,2	AKT1,CD81,MARK2,EPHA7,EPHB2,EPHB4,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FPR1,MTOR,HRAS,HSP90AA1,HTR2A,JA
1_Member	GO Biologi	GO:00434	positive reg	-16.8015	-13.713	37/552	147,351,5	ADRA1B,APP,ATP6AP1,CD81,FGFR1,FGFR2,FGFR4,FLT3,FLT4,FPR1,HRAS,HTR2A,JAK2,JUN,MAP3K11,NTRK1,NTR
1_Member	GO Biologi	GO:00458	positive reg	-15.4526	-12.456	35/540	207,975,2	AKT1,CD81,MARK2,FGFR1,FLT3,FPR1,MTOR,HRAS,HSP90AA1,HTR2A,JAK2,MAP3K11,NTRK1,NTRK2,ROR2,PAK2,
1_Member	GO Biologi	GO:00434	regulation	-13.9846	-11.046	27/344	975,1845,	CD81,DUSP3,EPHB2,FGFR1,FLT3,FPR1,HRAS,HTR2A,JAK2,MAP3K11,NTRK1,ROR2,PAK3,PRKCD,MAPK1,MAP2K2
1_Member	GO Biologi	GO:00321	activation	-13.3668	-10.495	26/335	207,975,2	AKT1,CD81,MARK2,FPR1,MTOR,JAK2,MAP3K11,NTRK1,PAK2,PAK3,PRKACA,PRKCD,MAPK1,MAP2K2,MAP2K3,M
1_Member	GO Biologi	GO:00310	stress-activ	-13.0264	-10.170	25/317	207,351,1	AKT1,APP,DUSP3,FLT4,HRAS,MAP3K11,ROR2,PAK2,PAK3,MAPK1,MAP2K2,MAP2K3,MAP2K5,EIF2AK2,MAP3K12
1_Member	GO Biologi	GO:00719	regulation	-12.5671	-9.725	31/528	207,975,1	AKT1,CD81,CDK4,DUSP3,EPHB2,FGFR1,FLT3,FPR1,HRAS,HTR2A,JAK2,MAP3K11,NTRK1,ROR2,PAK3,PKD1,PRKCD
1_Member	GO Biologi	GO:00434	positive reg	-12.032	-9.231	22/263	975,2260,	CD81,FGFR1,FLT3,FPR1,HRAS,HTR2A,JAK2,MAP3K11,NTRK1,ROR2,PAK3,MAPK1,MAP2K2,MAP2K3,MAP2K5,EIF
1_Member	GO Biologi	GO:00719	positive reg	-11.407	-8.650	24/343	207,975,2	AKT1,CD81,FGFR1,FLT3,FPR1,HRAS,HTR2A,JAK2,MAP3K11,NTRK1,ROR2,PAK3,PKD1,MAPK1,MAP2K2,MAP2K3,
1_Member	GO Biologi	GO:00328	regulation	-10.0794	-7.442	19/239	207,351,1	AKT1,APP,DUSP3,FLT4,HRAS,MAP3K11,ROR2,MAPK1,MAP2K2,EIF2AK2,MAP3K12,MAP4K4,HAND2,HIPK3,ERN2
1_Member	GO Biologi	GO:00703	regulation	-9.98623	-7.357	19/242	207,351,1	AKT1,APP,DUSP3,FLT4,HRAS,MAP3K11,ROR2,MAPK1,MAP2K2,EIF2AK2,MAP3K12,MAP4K4,HAND2,HIPK3,ERN2

Gene	original_id	Cond1-UP	Cond1-Dov	Cond2-UP	Cond2-Dov	Cond3-Dov	Cond3-UP	Gene ID	Type	Tax ID	Homologe	Homologe	Gene Symb	Descriptio	Biological F	Kinase Clas	Protein Fu	Subcellula	Drug (Drug	Canonical F	Hallma
1314	1314	1	1	1	1	1	1	1314	Gene_ID	H. sapiens	1314	H. sapiens	COPA	COPI coat c	GO:1902463	protein localization t		Cytosol;Golgi apparatu	(M243)PID		(M593
537	537	1	1	1	1	1	1	537	Gene_ID	H. sapiens	537	H. sapiens	ATP6AP1	ATPase H+t	GO:2001206	positive regulation of		Cytosol;Microtubules;Plasma mem			(M590
22820	22820	1	1	1	1	1	1	22820	Gene_ID	H. sapiens	22820	H. sapiens	COPG1	COPI coat c	GO:0051683	establishment of Golg		Golgi appa	S-(Dimethylarsenic)Cysteine (St		
9276	9276	1	1	1	1	1	1	9276	Gene_ID	H. sapiens	9276	H. sapiens	COPB2	COPI coat c	GO:1901998	toxin transport;GO:0		Cytosol(Approved)			(M591
3725	3725	0	1	1	0	1	1	3725	Gene_ID	H. sapiens	3725	H. sapiens	JUN	Jun proto-c	GO:0045657	positive r	Transcripti	Nucleoplas	Vinblastine	(M190)PID	(M589
10291	10291	0	1	1	0	1	1	10291	Gene_ID	H. sapiens	10291	H. sapiens	SF3A1	splicing fac	GO:0000389	mRNA 3'-splice site re		Nuclear speckles(Enhanced)			(M592
6625	6625	0	1	1	0	1	1	6625	Gene_ID	H. sapiens	6625	H. sapiens	SNRNP70	small nucle	GO:1904715	negative regulation o		Nucleoplasm(Enhanced)			
9114	9114	0	1	1	0	1	1	9114	Gene_ID	H. sapiens	9114	H. sapiens	ATP6V0D1	ATPase H+t	GO:0090383	phagosor	Transporters/Primary Active Transporters				(M592
527	527	0	1	1	0	1	1	527	Gene_ID	H. sapiens	527	H. sapiens	ATP6V0C	ATPase H+t	GO:0090383	phagosome acidification	GO:0033572 transferrin transp				(M593
10155	10155	1	0	1	1	0	1	10155	Gene_ID	H. sapiens	10155	H. sapiens	TRIM28	tripartite m	GO:1902187	negative r	Enzymes/E	Nucleoplasm(Enhanced	(M84)PID A		(M592
54866	54866	0	1	1	0	1	1	54866	Gene_ID	H. sapiens	54866	H. sapiens	PPP1R14D	protein ph	GO:1905183	negative regulation of	protein serine/threonine phosphatase acti				
55851	55851	0	1	1	0	1	1	55851	Gene_ID	H. sapiens	55851	H. sapiens	PSENEN	presenilin e	GO:0035333	Notch receptor processing, ligan	E2012	(M251)PID	(M590		
1195	1195	0	1	1	0	1	1	1195	Gene_ID	H. sapiens	1195	H. sapiens	CLK1	CDC like kir	GO:00181(	CMGC Ser/	Enzymes/(E	Nuclear me	Debromohymenialdisine (Prote		
10159	10159	0	1	1	0	1	1	10159	Gene_ID	H. sapiens	10159	H. sapiens	ATP6AP2	ATPase H+t	GO:0002003	angioten	Transporters/Accessory Factors Inv	(M77)PID	WNT SIGN		
4928	4928	1	1	0	1	1	0	4928	Gene_ID	H. sapiens	4928	H. sapiens	NUP98	nucleopori	GO:0006409	tRNA export from nu	Nucleoplasm(Supported)	Vesicles(L	(M590		
523	523	0	1	1	0	1	1	523	Gene_ID	H. sapiens	523	H. sapiens	ATP6V1A	ATPase H+t	GO:0090383	phagosor	Enzymes/E	Cytosol;Nu	Alendronic acid (Proton-transp		
5253	5253	0	1	1	0	1	1	5253	Gene_ID	H. sapiens	5253	H. sapiens	PHF2	PHD finger	GO:0061188	negative regulation of	chromatin silencing at rDNA;GO:0061187				
57418	57418	1	0	1	1	0	1	57418	Gene_ID	H. sapiens	57418	H. sapiens	WDR18	WD repeat	GO:0030174	regulation of DNA-de	Nucleoplasm(Approved)				

```r
library(readxl)
metascape_Enrichment <- read_excel("/UserData/SAJPC/21-17-reanalysis/metascape/metascape_result_combinedList.xlsx",
sheet = "Enrichment")

metascape_Annotation <- read_excel("/UserData/SAJPC/21-17-reanalysis/metascape/metascape_result_combinedList.xlsx",
sheet = "Annotation")

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

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

library(readr)
# load L2FC data for RTvT Down and TvB UP
RTvT_GO_Down <- read_csv("/UserData/SAJPC/21-17-reanalysis/metascape/RTvT_GO_Down.csv")
TvB_GO_UP <- read_csv("/UserData/SAJPC/21-17-reanalysis/metascape/TvB_GO_UP.csv")
# make L2FC tables
RTvT_Down <- RTvT_GO_Down[RTvT_GO_Down$name %in% ecmPathwayTargets, 2:3]
TvB_UP <- TvB_GO_UP[TvB_GO_UP$name %in% ecmPathwayTargets, 2:3]

combinedL2FCTable <- rbind(RTvT_Down, TvB_UP) %>% cbind(condition = rep(c("RTvT_Down", "TvB_UP"), each = 27), stringsAsFactors = FALSE)

# Plot
library(ggplot2)
cols <- c("RTvT_Down" = "blue", "TvB_UP" = "red")
combinedL2FCTable %>%
  mutate(name = 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("Log"[2], "(fold)"))) +
  scale_color_manual(values = cols,"Condition", labels = c("TGF-β+RGCC", "TGF-β")) +
  # scale_color_manual(values = cols,"Condition", labels = c(expression(paste("RGCC"+, "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))
# theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```


Condition
Cond 1
Cond 2
Gene C
Gene B
Gene A
-6
-4
-2
0
2
4
Log2(fold)

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.umaryland.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!