The Journey from *Messy* to *Tidy*:
How I used R to tidy income tax data

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—	—	34.5
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—	—	16.8
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—	—	—

	ozone	solar.r	wind	temp	month	day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10

For Data Science Community of Practice Demo Day

2019-04-17

Noushin Nabavi
Mitacs Canadian Science Policy Fellow

Integrated Data Division (OCIO, CITZ)

DataBC

Provides easy access to open data, geographic data and associated products and services.

Data Innovation Program **

Enables a big picture view of the complex issues affecting British Columbians through access to integrated, de-identified data

Data Science Partnerships

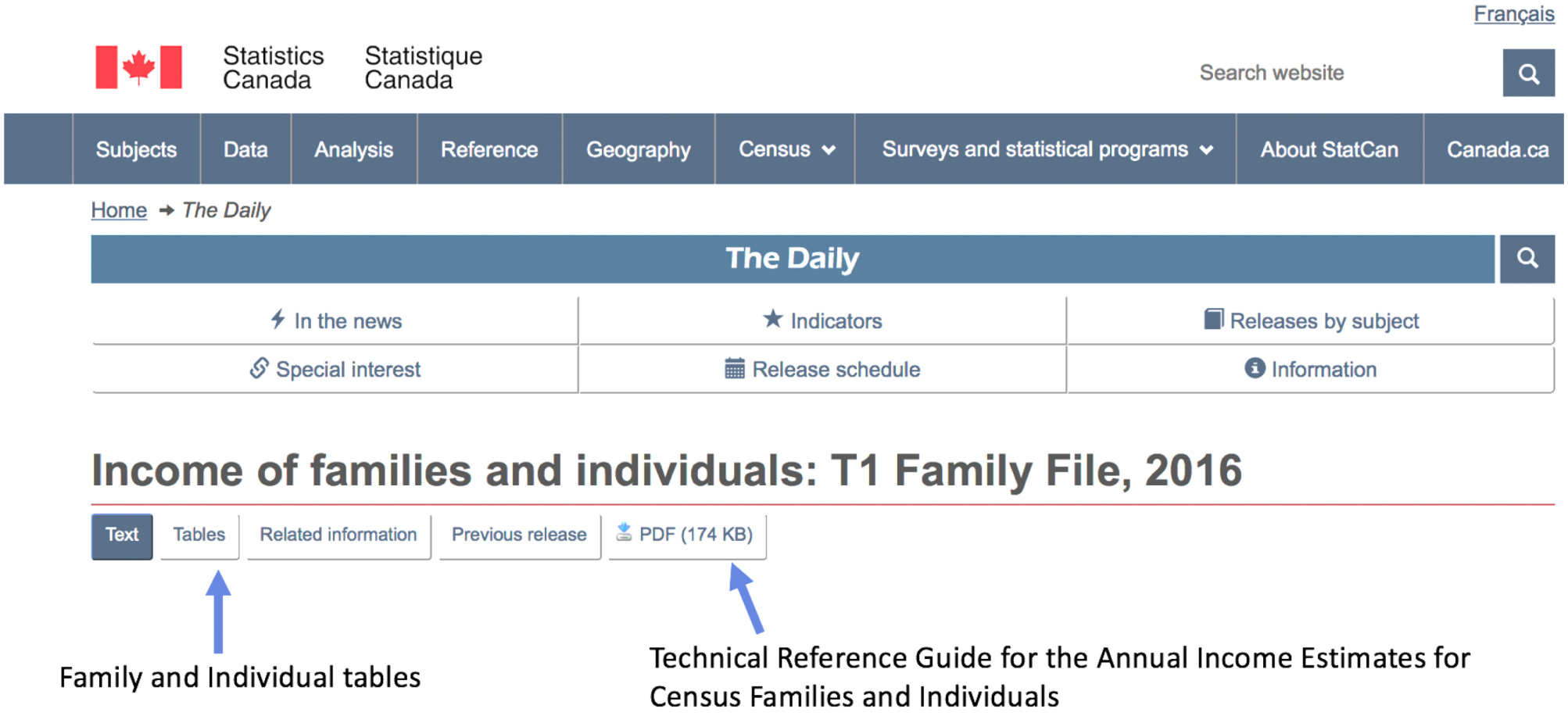
Provides data science support to ministries across government.

Startup in Residence

Brings the expertise of startup tech companies to government to co-develop tech solutions for service delivery to British Columbians

*** Stay tuned: more info about the Data Innovation program available online next month*

Historical Stat-Can data available to IDD – CITZ thanks to BC-Stats (~2000 – 2015 tax filer's information by age, gender, income, and area)



https://www150.statcan.gc.ca/n1/pub/11-26-0001/112600012019001-eng.htm

Open data licence: https://www.statcan.gc.ca/eng/reference/licence

Socioeconomic status indicators are a constant ask by ministry partners

Utilities of income tax data:

- Area-based Income distribution/ Polarization of gender/age of tax-filers
- Population Growth in time

Limitations:

- Aggregated area-based measures
- Populated areas make it difficult to make direct comparisons among individuals
- Area-based indices only provide approximations

Opportunities

- Status is more than just income levels so opportunity to link with cross- ministry data [education, health, social assistance]

Data table structures

Total data: 93 IND + 84 FAM = 177 tables +

IND

- 2000_IND_Tables 1_to_6_Canada.xls
- 2001_IND_Tables 1_to_6_Canada.xls
- 2002_IND_Tables 1_to_7_Canada.xls
- 2003_IND_Tables 1_to_7_Canada.xls
- 2004_IND_Tables 1_to_7_Canada.xls
- 2004_to_2015_IND_Table_13.xls — T.13
- 2005_IND_Tables 1_to_7_Canada.xls
- 2006_IND_Tables 1_to_7_Canada.xls
- 2007_IND_Tables 1_to_8_Canada.xls
- 2008_IND_Tables 1_to_8_Canada.xls
- 2009_IND_Tables 1_to_8_Canada.xls
- 2010_IND_Tables 1_to_8_Canada.xls
- 2011_IND_Tables 1_to_8_Canada.xls
- 2012_IND_Tables 1_to_8_Canada.xls
- 2013_IND_Tables 1_to_8_Canada.xls
- 2014_IND_Tables 1_to_8_Canada.xls
- 2015_IND_Tables 1_to_13_Canada.xls

Note spaces in names!

FAM

- 2004_Family_Tables_19_20_New_LIM.xls
- 2005_Family_Tables_19_20_New_LIM.xls
- 2006_Family_Tables_19_20_New_LIM.xls
- 2007_Family_Tables_19_20_New_LIM.xls
- 2008_Family_Tables_19_20_New_LIM.xls
- 2009_Family_Tables_19_20_New_LIM.xls
- 2010_Family_Tables_19_20_New_LIM.xls
- 2011_Family_Tables_19_20_New_LIM.xls
- 2012_Family_Tables_19_20_New_LIM.xls
- 2013_Family_Tables_1_to_18_Canada.xls
- 2013_Family_Tables_19_20_New_LIM.xls
- 2014_Family_Tables_1_to_18_Canada.xls
- 2014_Family_Tables_19_20_New_LIM.xls
- 2015_Family_Tables_1_to_18_Canada.xls
- 2015_Family_Tables_19_20_New_LIM.xls

Sheet Structures…

Table name at top

Empty cells

Duplicate Column Names

All of Canada, we want just BC !!

Source/license declaration

	A	B	C	D	E	F	G	I	J	K	L	M	N	O	P	
1	Table I-01 - Individual data - Taxfilers and dependents, summary table, income and demographics of individuals, 2000															
2	CityID	Postal area	Postal walk	Level of	Place name	Taxfilers									All persons	
3				geo		#	% change 1995-2000	% 0-24	% 25-44	% 45-64	% 65+	Average	% female	% married	% in appt	#	% change 1995-2000
4												Age					
5	9099	Z99099		12	CANADA	22248670	8	13	40	30	17	45	52	51	9	29891290	4
6	9010	A99010		11	NEWFOUNDLAND	391930	0	13	39	32	15	45	50	56	1	523860	-7
7	425	A0N1A0		9	AGUATHUNA	250	4	8	33	33	25	50	48	48	0	320	-9
8	307	A0K1A0		9	ANCHOR POINT	260	-4	15	38	35	12	44	48	68	0	320	-9
9	70	A0B1A0		9	ARNOLD'S COVE	900	5	14	37	32	16	45	52	58	0	1120	-5
10	193	A0G1A0		9	ASPEN COVE	220	10	14	38	29	19	45	52	71	0	290	0
11	5110	511010		51	AVALON PENINSULA	183760	0	14	40	31	15	44	52	53	1	244130	0
12	2	A0A1B0		9	AVONDALE	570	-2	12	35	35	18	47	51	48	0	760	-8
13	268	A0H1A0		9	BADGER	790	3	13	37	35	15	46	49	62	0	1030	-6
14	194	A0G1B0		9	BADGER'S QUAY	970	4	15	34	30	21	47	52	63	0	1230	-7
12659	9062	X89062		10	NUNAVUT	260	0	12	48	40		42	54	40	0	550	0
12660	5969	X0A0R0		9	PANGNIRTUNG	690	0	23	46	23		37	52	38	0	1260	0
12661	5970	X0A0S0		9	POND INLET	570	0	24	52	21	3	36	51	39	0	1230	0
12662	5959	X0A0B0		9	QIQIKTARJUAG	300	0	24	52	21	7	37	48	43	0	510	0
12663	5979	X0C0G0		9	RANKIN INLET	1180	0	21	53	23	3	36	53	42	0	2150	0
12664	5980	X0C0H0		9	REPULSE BAY	300	0	29	48	19	0	35	48	53	0	640	0
12665	5971	X0A0V0		9	RESOLUTE	100	0	20	50	30	0	39	55	33	0	190	0
12666	5972	X0A0W0		9	SANIKILUAQ	330	0	15	55	24	6	38	52	52	0	700	0
12667	6003	X0B1B0		9	TALOYOAK	340	0	21	53	21	6	37	53	41	0	700	0
12668	5981	X0C0J0		9	WHALE COVE	140	0	21	57	21	0	37	50	50	0	290	0
12669																
12670	**Source:** Statistics Canada, Income Statistics Division, 2000, Annual Income Estimates for Census Families and Individuals, 13C0015															
12671																
12672	**"©This data includes information copied with permission from Canada Post Corporation"**															
12673																

1 2 3A 3B 3C 4 5A 5B 5C 6 7A 7B 7C 8 9 10 11 12 13

Ready

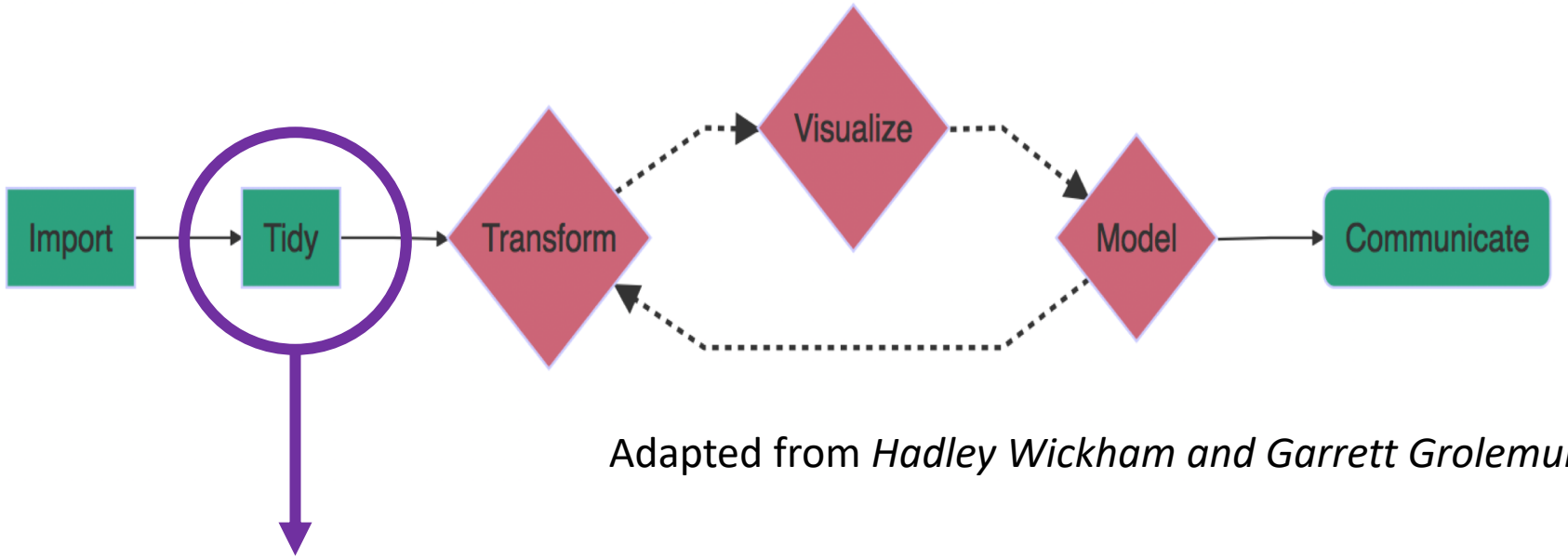
Symptoms of messy data

- Empty cells
- Missing and extreme values (NA, Inf)
- Variables coded inconsistently
- Spaces in headers
- Not all messy data is created equal:

https://www.michaelchimenti.com/2014/07/five-common-problems-with-messy-data/

…everything that can go wrong will go wrong

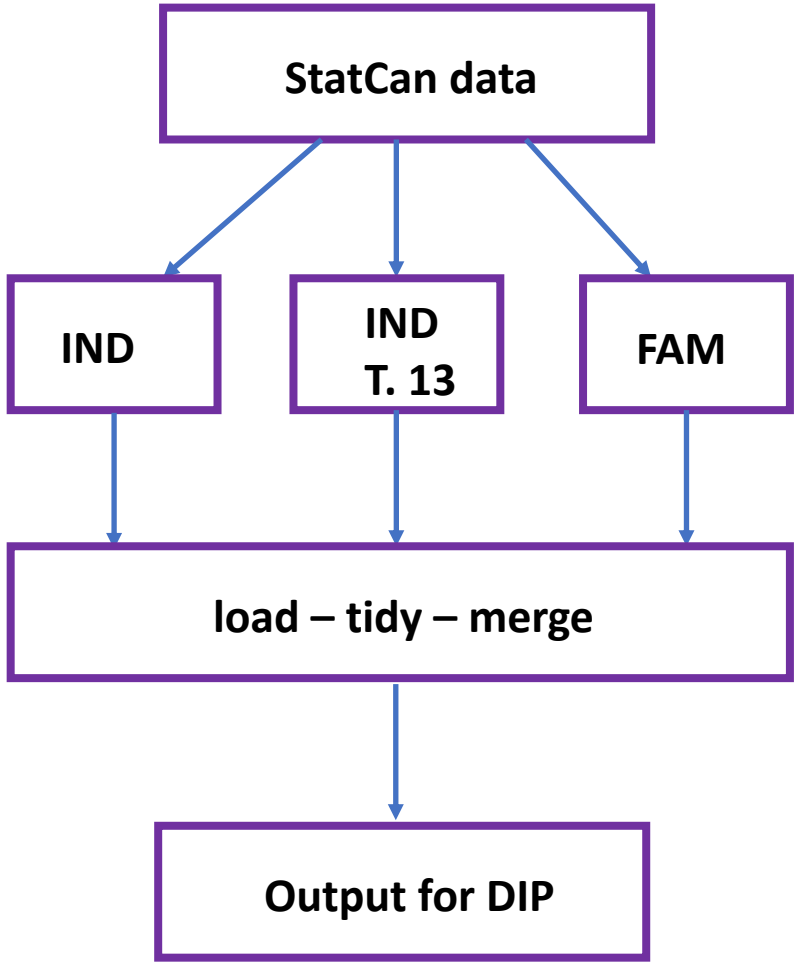
The data *scien[ce|tist]* workflow



Adapted from *Hadley Wickham and Garrett Grolemund*

80% of a data scientists' time gets spent on cleaning data

Overall Strategy for cleaning data tables



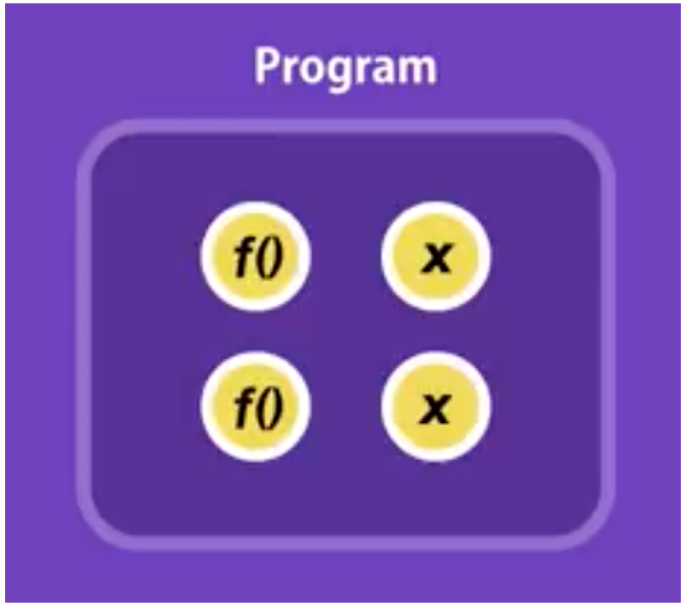
Thank you to **Stephanie Hazlitt, Ashlin Richardson, Andriy Koval**
Cindy Wang, Julie Labelle, Beth Collins, Simon Munn, Dan Mackenzie
ED discussions: **Jeremy Coad and Kathleen Assaf**

Library dependencies

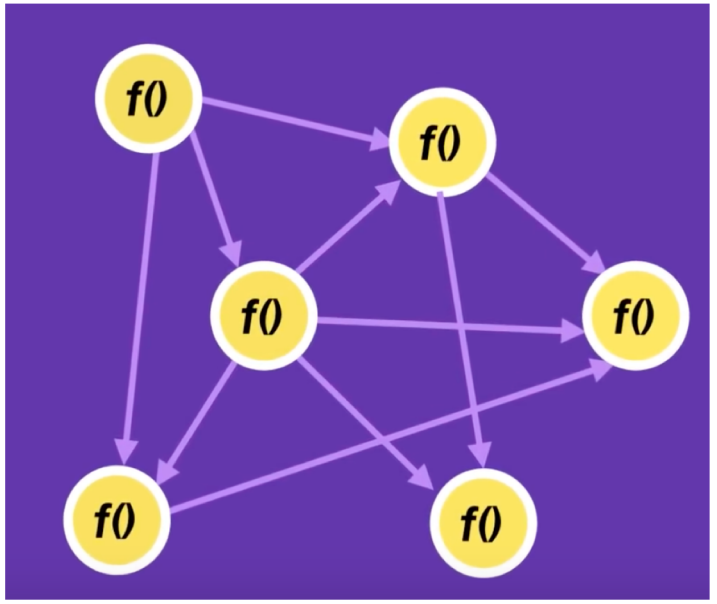
Thanks to RSudio Chief Scientist and colleagues Tidyverse provides an efficient, fast, and well-documented workflow for data wrangling, modeling, and visualization

```
pkgs <-c("plyr", "janitor", "tibble", "stringr",
"tidyr", "here", "readr", "purrr", "readxl", "dplyr")
```

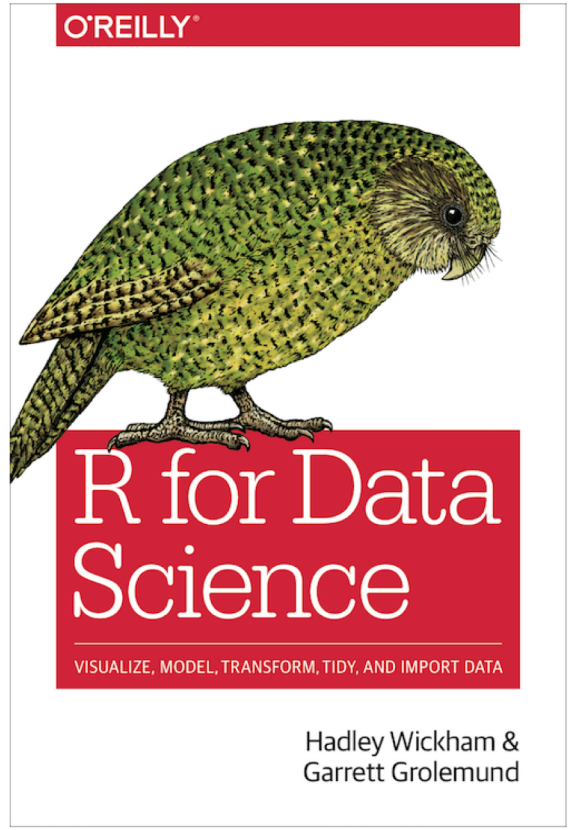
Procedural oriented programming



Functions that operate on data



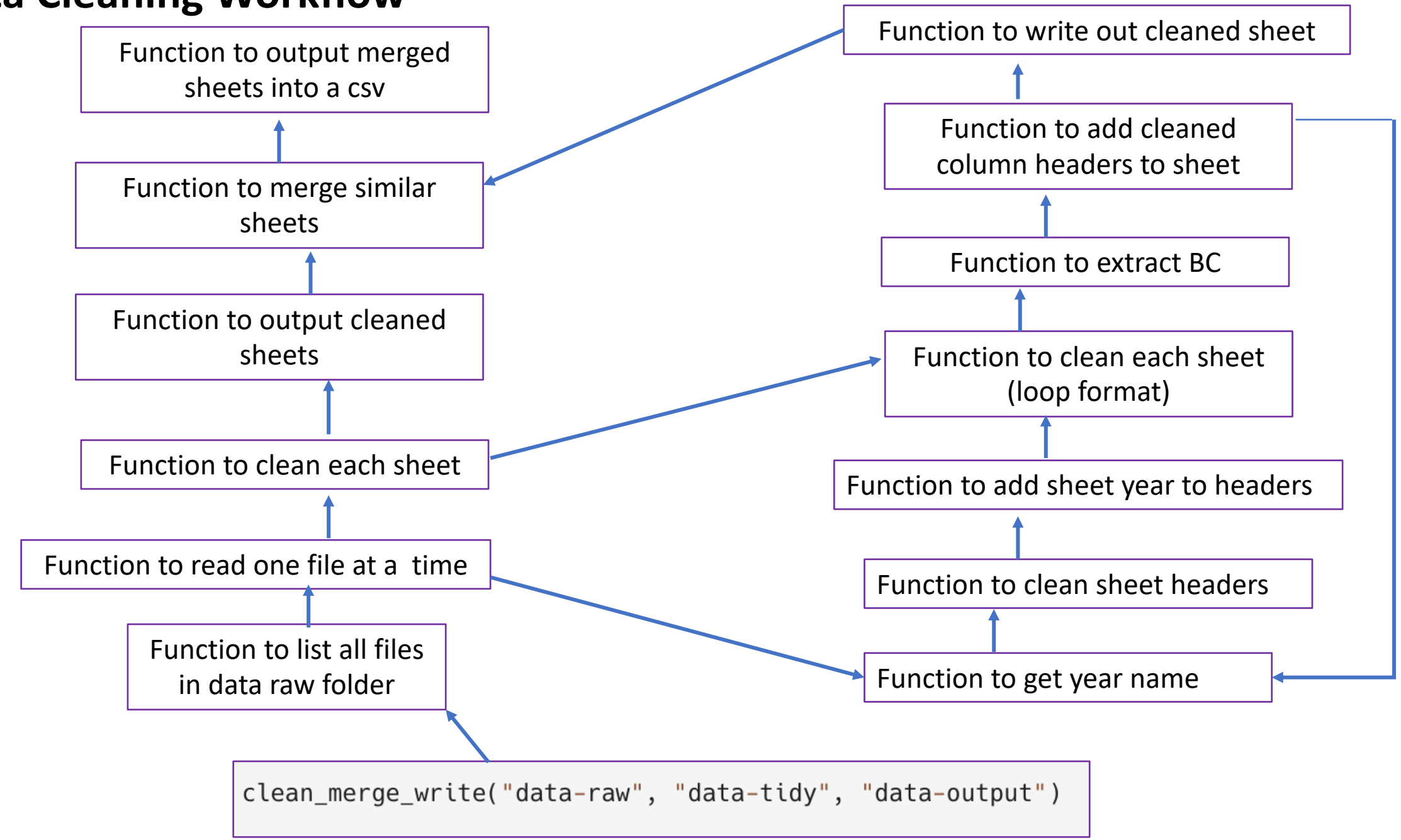
Functions that talk to one another and are inter-dependent



Best practice principles for code writing: DRY SOLID

- Don't Repeat Yourself (**DRY**)
- Single Responsibility Principle of **SOLID** (developed for Object-oriented programming)

Data Cleaning Workflow



Function to fix column headers

```
sheetcolnames <- path %>%
      read_excel(sheet = sheet, skip = 1, n_max = 3, col_names = FALSE) %>%
      t() %>%
      as_tibble(.name_repair = ~ tempcols) %>%
      fill(tempcols) %>%
      unite(sheet_col_names) %>%
      mutate(sheet_col_names = mutate_col_names(sheet_col_names)) %>%
      select(sheet_col_names) %>%
      pull()
```

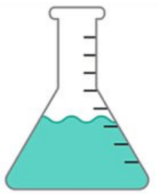
* **Fill** fills missing values in selected columns using the previous entry
** **pull** selects a column in a data frame and transforms it into a vector

Thank you Stephanie Hazlitt!

Function to extract the years of tables

```r
get_file_year <- function(path) {
  # find file_year
  pathbase <- path %>%
    basename() %>%
    tools::file_path_sans_ext()

  # assign file_year
  file_year <- pathbase %>% stringr::str_extract("^[0-9]{4}")
  return(file_year)
}
```



Replace the top 4 rows of sheets with repaired headers

```
tidy_df <- path %>%
    read_excel(sheet = sheet, skip = 4,
               col_names = sheetcolnames,
               .name_repair = "unique") %>%
    remove_empty_rows() %>%
    tibble::add_column(year = file_year, .before = 1)
```

Clean column headers after tables are merge

```
mutate_col_names <- function(sheet_col_names) {
  sheet_col_names = str_trim(sheet_col_names)
  sheet_col_names = tolower(str_replace_all(sheet_col_names, "\\s", "|"))
  sheet_col_names = str_replace_all(sheet_col_names, "_", "|")
  sheet_col_names = str_replace_all(sheet_col_names, "na|na||", "")
  sheet_col_names = str_replace_all(sheet_col_names, "\\|\\|", "|")
  sheet_col_names = str_replace_all(sheet_col_names, "change\\|\\d{4}-\\d{4}",
"range|last|5years")
  return(sheet_col_names)
}
```

For a list of all functions please see:
https://github.com/bcgov/statscan-taxdata-tidying

Cleaned-up tables

	A	B	C	D	E	F	G	H	I	J	K	L
1	year	cityid|	postal|area|	postal|walk|	evel|of|geo	lace|me|geo	taxfilers|#|	%|range|las	filers|%|0-2	filers|%|25-	filers|%|45-	xfilers|%|65
2	2000	9099	Z99099	NA	12	CANADA	22248670	8	13	40	30	17
3	2000	9010	A99010	NA	11	EWFOUNDLAI	391930	0	13	39	32	15
4	2000	9011	C99011	NA	11	E EDWARD IS	101070	8	15	37	31	17
5	2000	9012	B99012	NA	11	NOVA SCOTIA	677230	5	12	39	31	17
6	2000	9013	E99013	NA	11	W BRUNSWI	557210	5	13	39	31	17
7	2000	9024	J99024	NA	11	QU<c9>BEC	5492230	8	13	39	32	16
8	2000	9035	P99035	NA	11	ONTARIO	8393800	9	13	41	30	17
9	2000	9046	R99046	NA	11	MANITOBA	819290	3	14	38	29	18
10	2000	9047	S99047	NA	11	ASKATCHEWA	708880	5	15	37	28	20
11	2000	9048	T99048	NA	11	ALBERTA	2190460	17	16	42	28	14
12	2000	9059	V99059	NA	11	TISH COLUM	2856380	8	12	39	31	17
13	2000	5716	V0K2E0	NA	6	00 MILE HOUS	4530	-2	13	36	36	15
14	2000	5716	V0K2Z0	NA	6	00 MILE HOUS	1360	9	10	38	35	17
15	2000	5716	V95716	NA	8	00 MILE HOUS	5890	0	12	36	36	16
16	2000	5717	V0K2G0	NA	9	50 MILE HOUS	1930	12	13	42	35	10
17	2000	5720	V0K2K0	NA	9	0 MILE HOUS	640	3	5	23	45	26

Overall outputs

- 1 csv per sheet
- All years merged into 1 table
- For families and individuals

IND

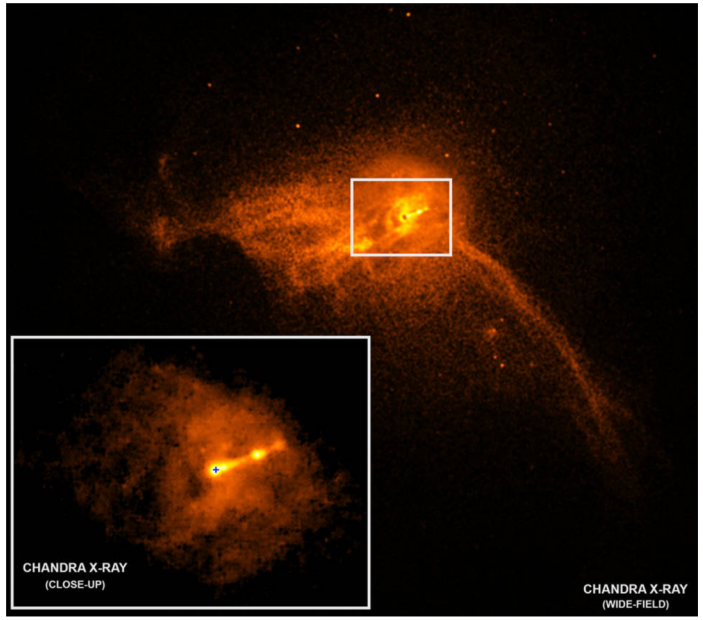
- 1_IND.csv
- 2_IND.csv
- 3A_IND.csv
- 3B_IND.csv
- 3C_IND.csv
- 4_IND.csv
- 5A_IND.csv
- 5B_IND.csv
- 5C_IND.csv
- 6_IND.csv
- 7A_IND.csv
- 7B_IND.csv
- 7C_IND.csv
- 8_IND.csv
- 9_IND.csv
- 10_IND.csv
- 11_IND.csv
- 12_IND.csv
- 13_IND.csv

FAM

- 1_FAM.csv
- 2_FAM.csv
- 3A_FAM.csv
- 3B_FAM.csv
- 3C_FAM.csv
- 4A_FAM.csv
- 4B_FAM.csv
- 4C_FAM.csv
- 5A_FAM.csv
- 5B_FAM.csv
- 6_FAM.csv
- 7_FAM.csv
- 8_FAM.csv
- 9_FAM.csv
- 10_FAM.csv
- 11_FAM.csv
- 12_FAM.csv
- 13_FAM.csv
- 14A_FAM.csv
- 14B_FAM.csv
- 14C_FAM.csv
- 15_FAM.csv
- 17_FAM.csv
- 18_FAM.csv
- 19_FAM.csv
- 20_FAM.csv

Lessons Learned

- Importance of code sharing, collaborating, teamwork, and feedback
- Code of conduct and framework on BCgov Github page



55 million light-years away M87 so definitely messy data; 200 scientists worked collaboratively on codes and algorithms so need to follow best practices to enable sharing

Acknowledgements:

Stephanie Hazlitt, Ashlin Richardson, Andriy Koval
Cindy Wang, Julie Labelle, Beth Collins, Simon Munn, Dan Mackenzie
Jeremy Coad, Kathleen Assaf, Sue Wheatley
Jackie Storen, Martin Monkman (BC Stats)

Resources

Journal of Statistical Software

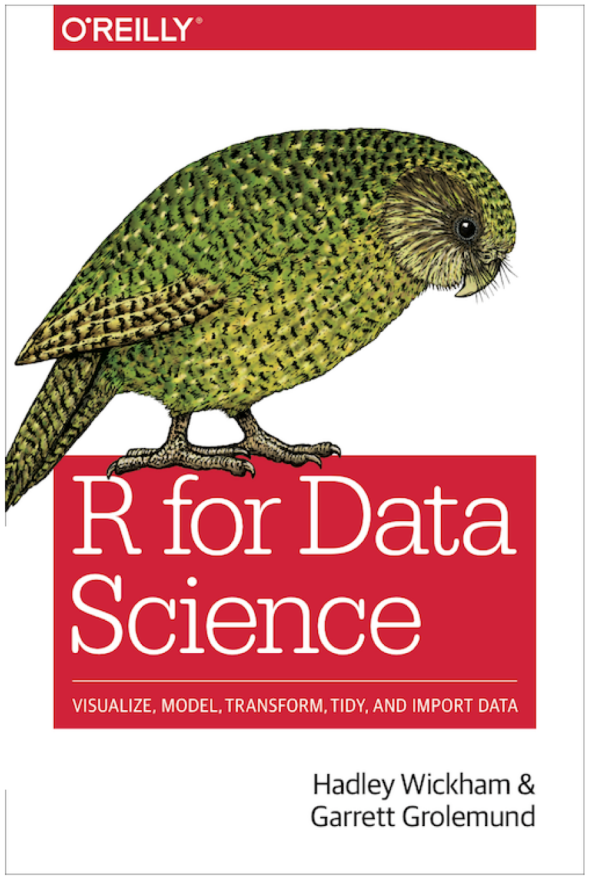
August 2014, Volume 59, Issue 10. *http://www.jstatsoft.org/*

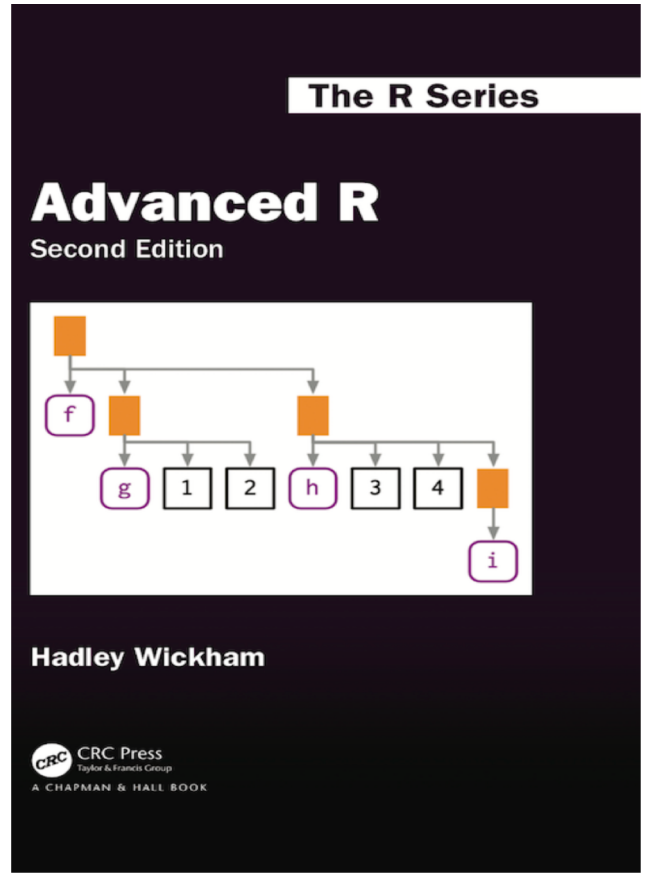
Tidy Data

Hadley Wickham
RStudio

Tidying the Australian Same Sex Marriage Postal Survey Data with R

Miles McBain Follow
Nov 19, 2017 · 6 min read





Object-Oriented Programming in 7-minutes: https://www.youtube.com/watch?v=pTB0EiLXUC8
Writing functions in R – Charlotte Wickham: github & youtube