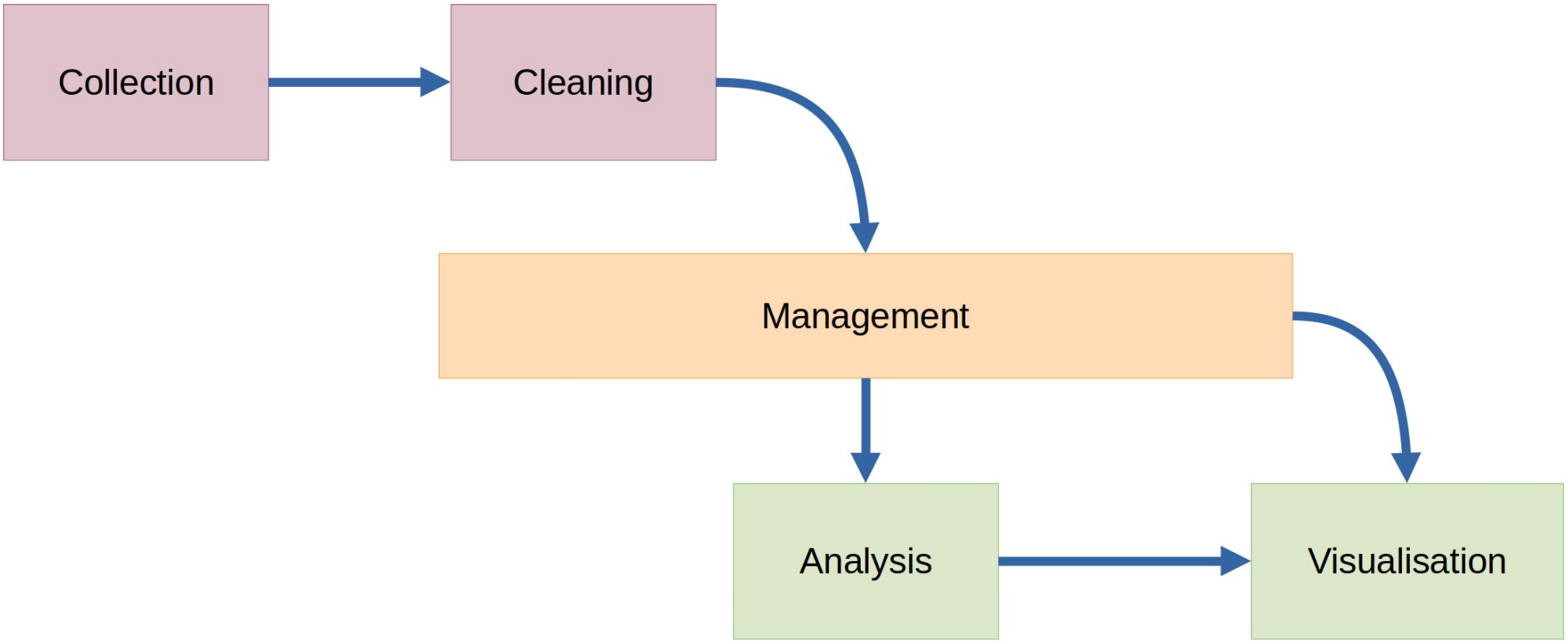
How does your data flow?
Using R for ETL

Mike Spencer

Data workflow



Talk structure

- Generic data workflow
- Case study
  - Conflict, Security and Stability Fund (CSSF)
  - Comprehensive Spending Review (CSR)

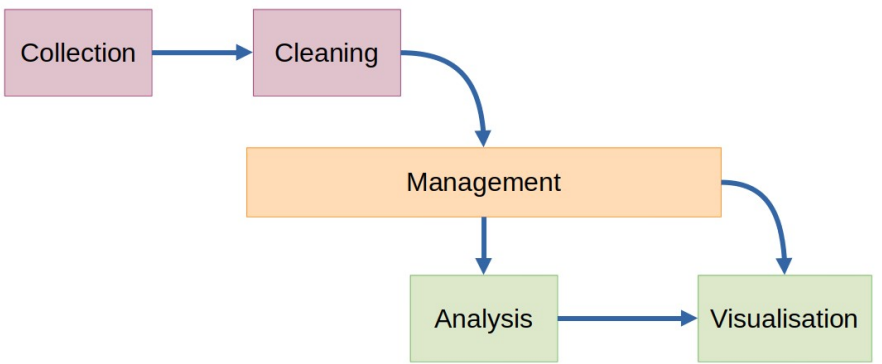
CSSF GLOBAL MEL Partnership

The elephant in the room

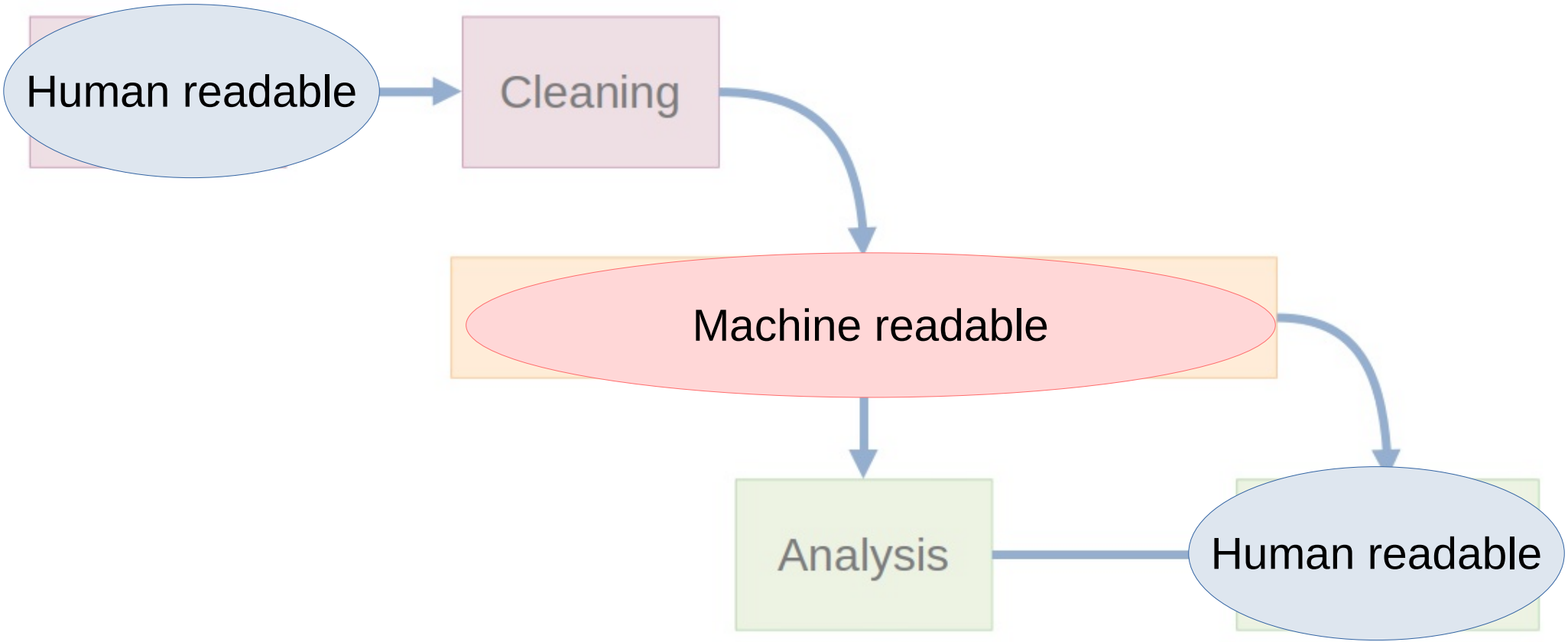
- Spreadsheets aren’t going away
- Clients/collaborators expect them
- Currently a spreadsheet may be used to do many different jobs – this is not good
- But we can work with spreadsheets better

Why separate?

- Automate
  - Saves time/money
  - Reduce human error
- Replace parts of workflow as required
- Reuse tools easily



Data workflow



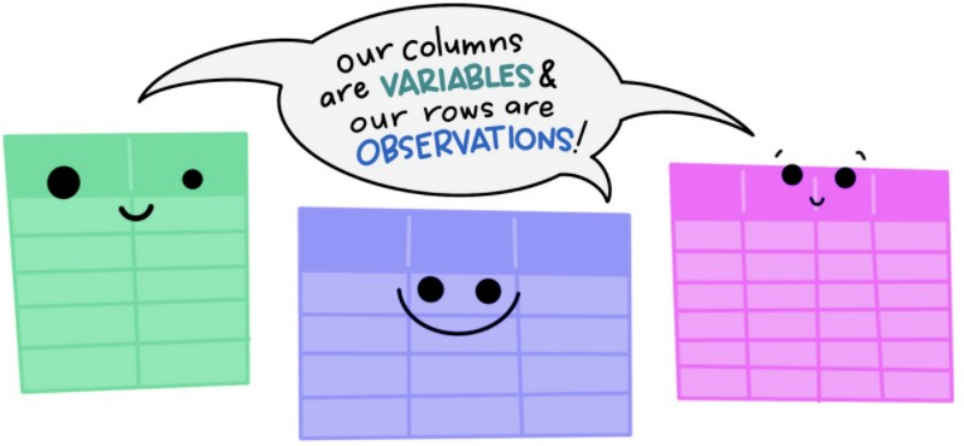
Tidy Data Illustrated Series

CC By Julie Lowndes & Allison Horst

Reuse: These slides (low res) • Blog • Twitter • GitHub

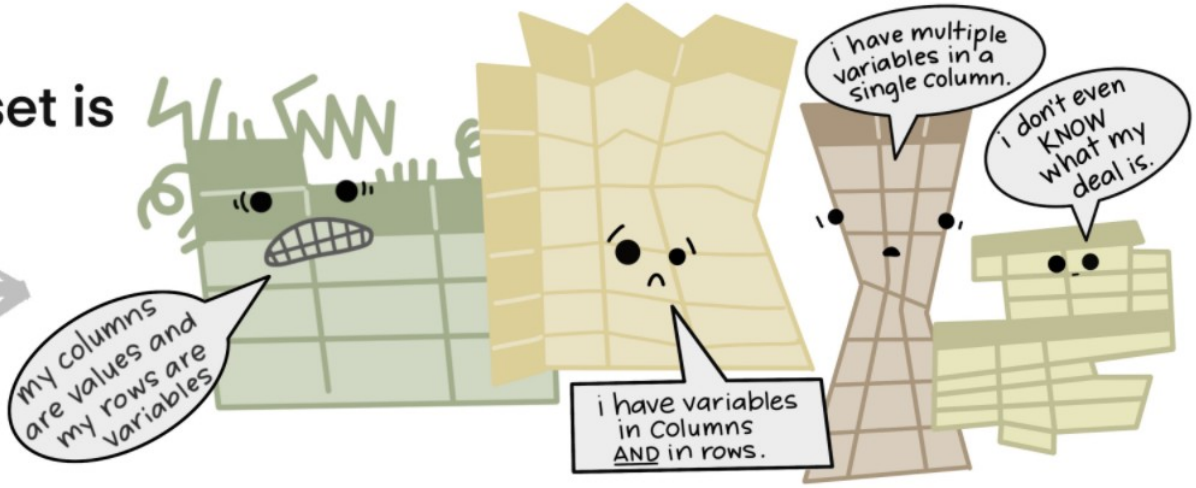
Please cite as: "Illustrations from the Openscapes blog *Tidy Data for reproducibility, efficiency, and collaboration* by Julia Lowndes and Allison Horst"

The standard structure of tidy data means that "tidy datasets are all alike..."

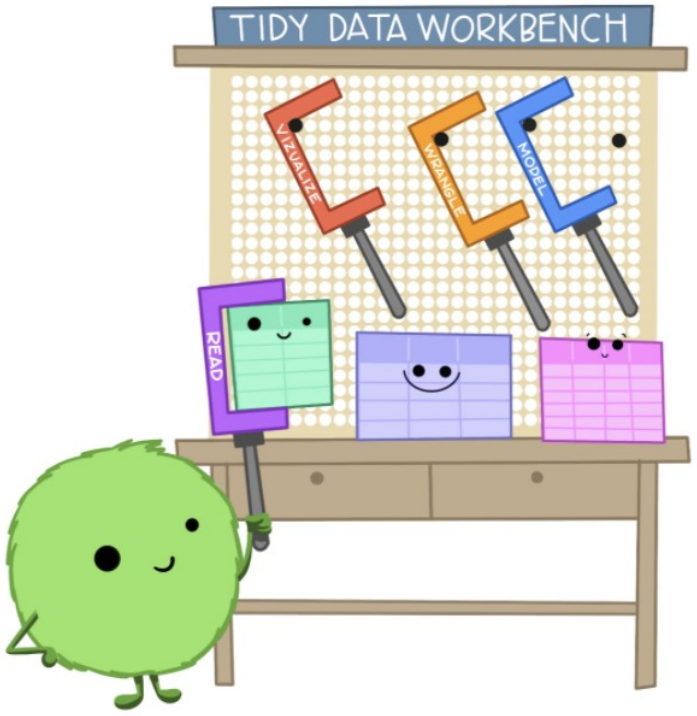


"...but every messy dataset is messy in its own way."

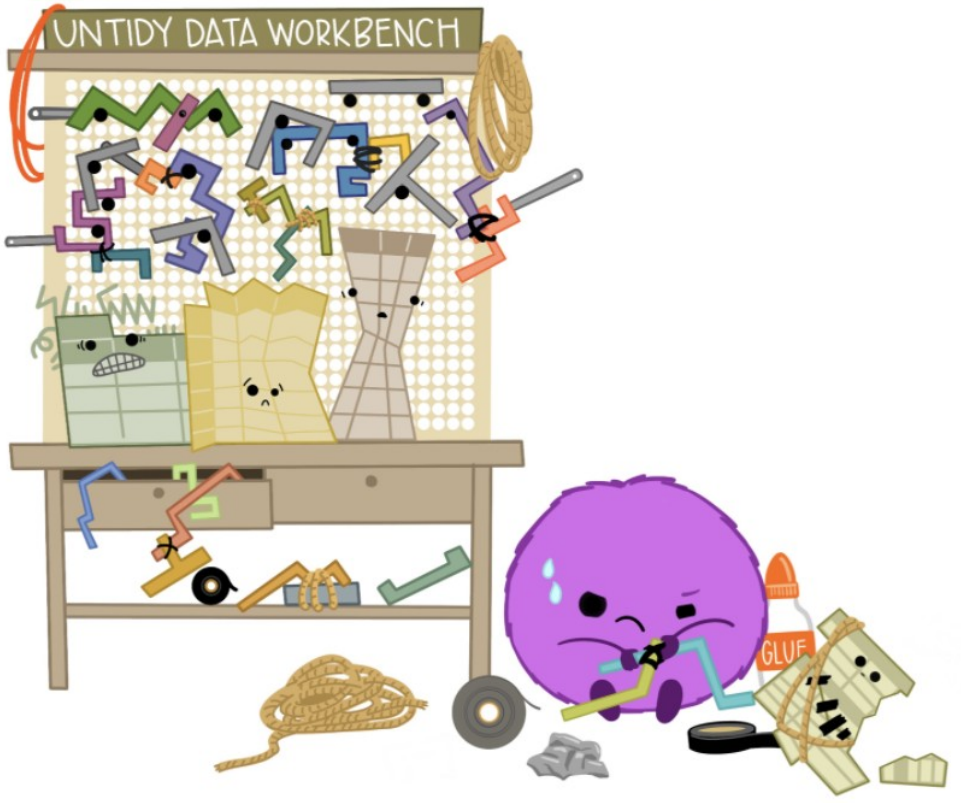
–HADLEY WICKHAM



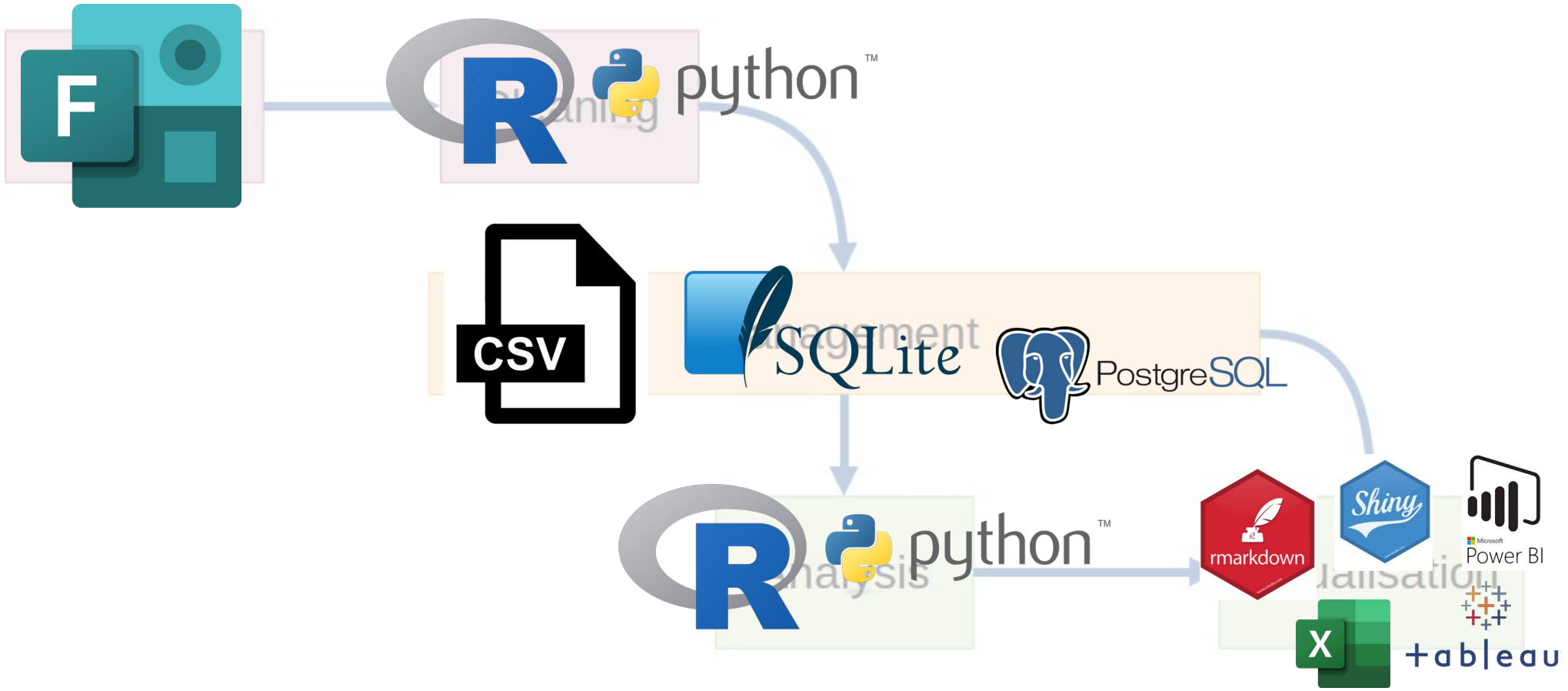
When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



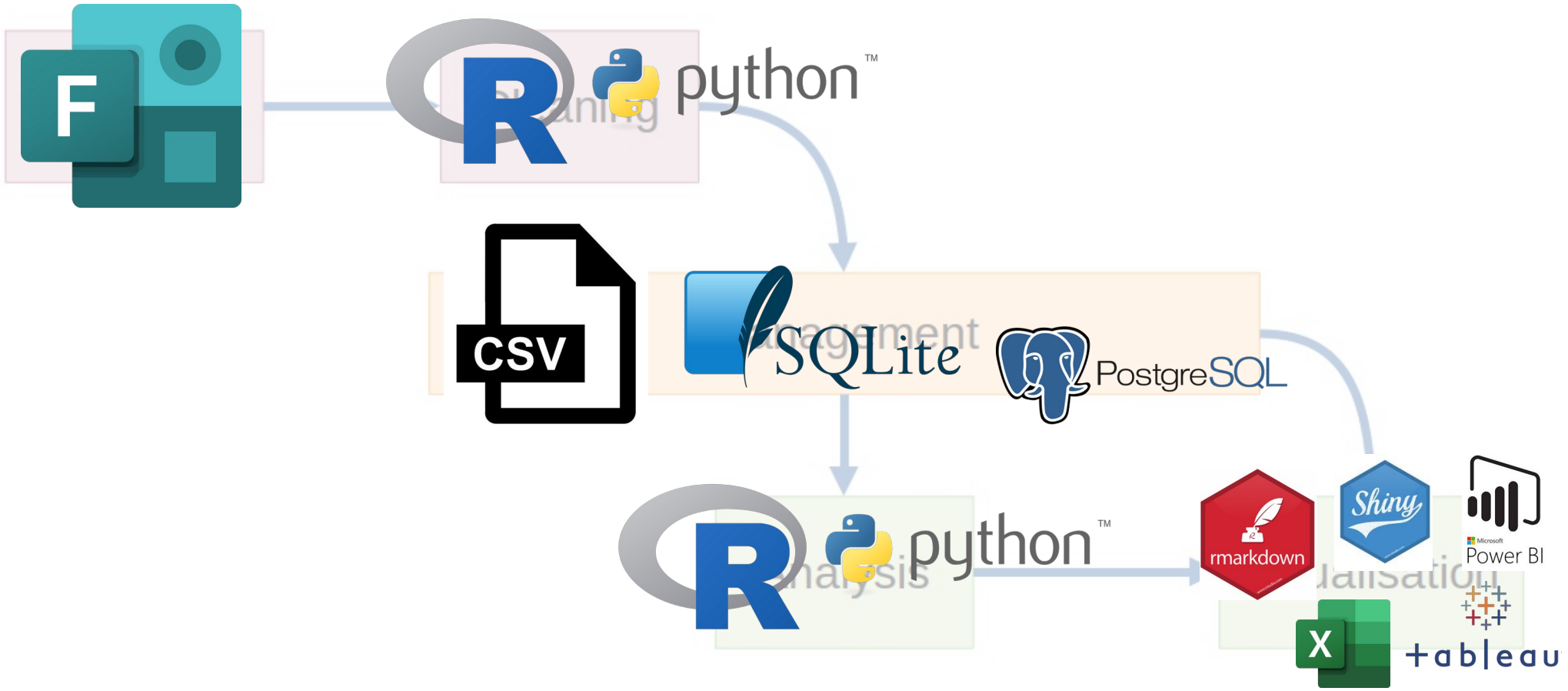
Data workflow



etlhelper

https://github.com/BritishGeologicalSurvey/etlhelper

Data workflow



Open
Transparent
Repeatable

CSSF spending review

- CSSF finances portfolios around the world
- Spending review data support:
  - Collate information
  - Visualise spend across places/portfolios/priorities
- Document and show the impacts of adjusting funding (moderation).

https://www.integrityglobal.com/our-work/projects/global-monitoring-evaluation-and-learning-partnership-client-conflict-security-and-stability-fund-cssf/

CSSF GLOBAL MEL Partnership

CSSF spending review

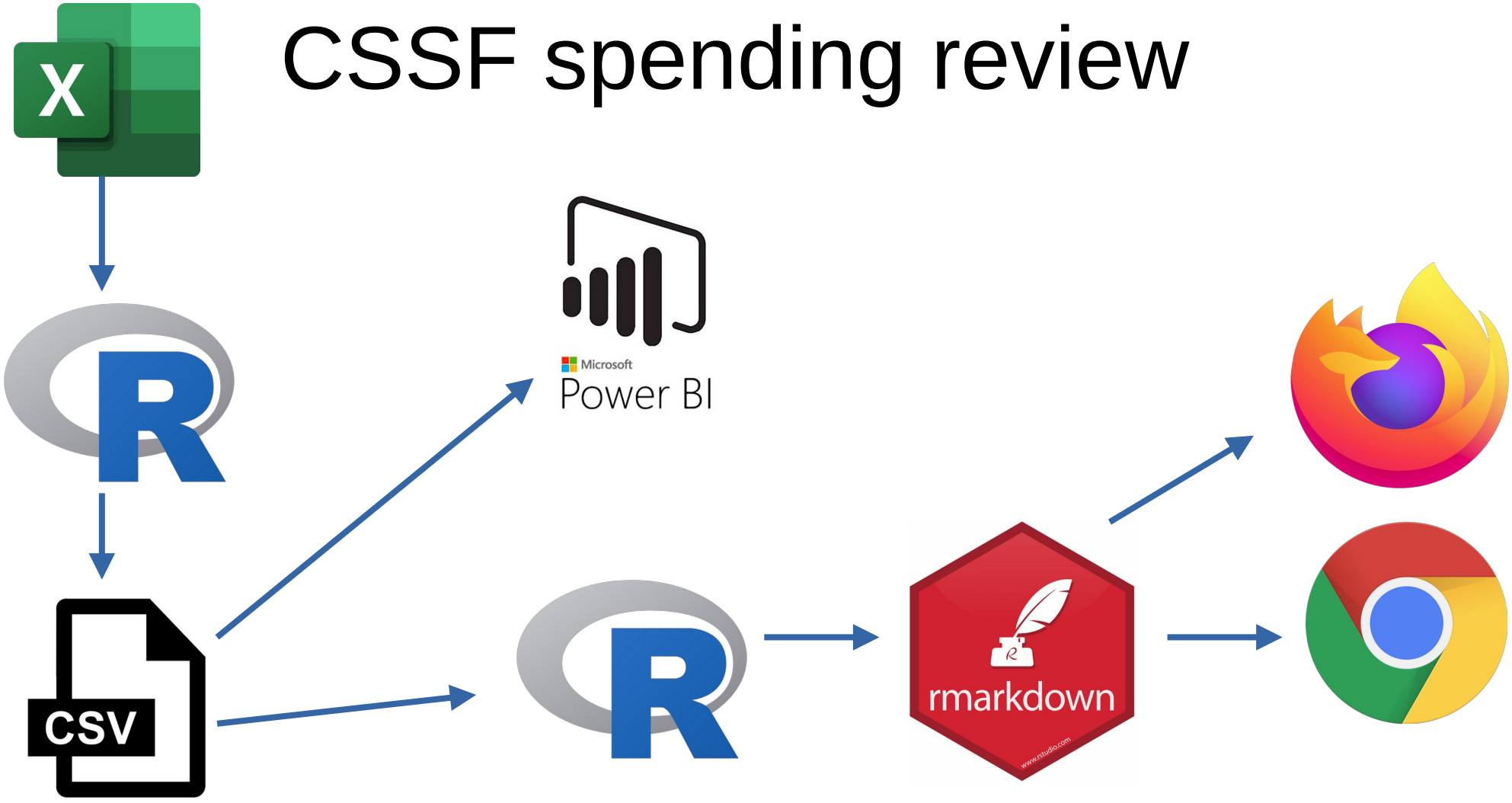
- Began 2-3 weeks before moderation
- ~ 12 hours between portfolio responses and first moderation session
- 30 xlsx files, each with 2 sheets and multiple tables – updated hourly

https://www.integrityglobal.com/our-work/projects/global-monitoring-evaluation-and-learning-partnership-client-conflict-security-and-stability-fund-cssf/

CSSF GLOBAL MEL Partnership

Dummy dataset
&
Automation

CSSF spending review



£5,469
Total Max Spend (21/22)

£5,668
Total Min Spend (21/22)

18
Total Portfolios

78
Total Programmes

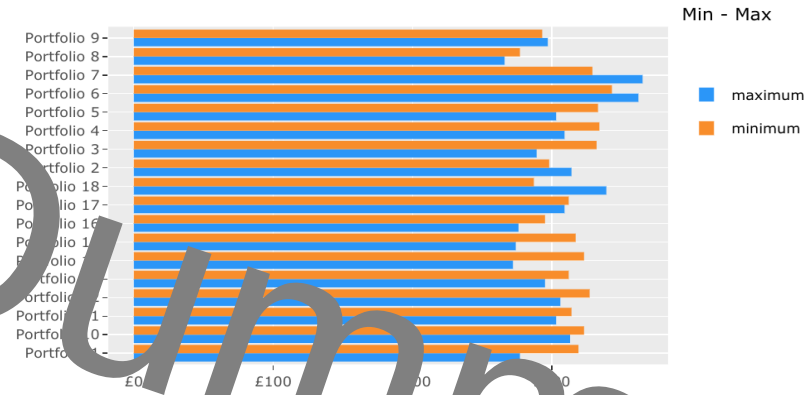
18
Total Departments

10
Total Priorities

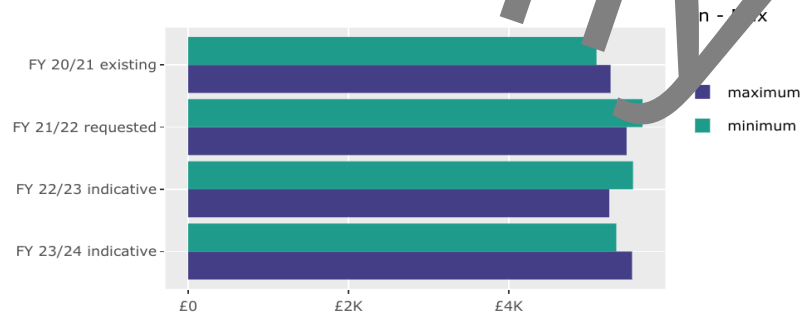
2.9
Average Risk

5
Total Regions

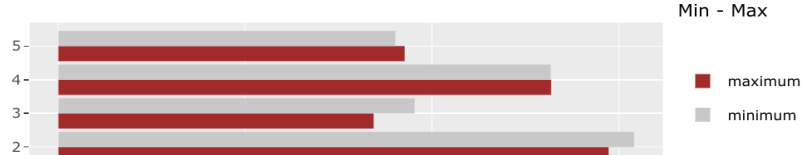
Spend by portfolio (FY 21/22)



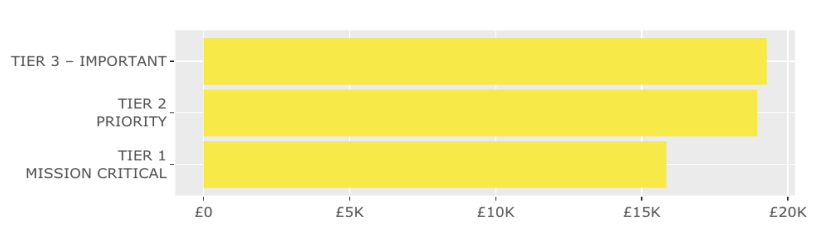
Forecast spend for each FY



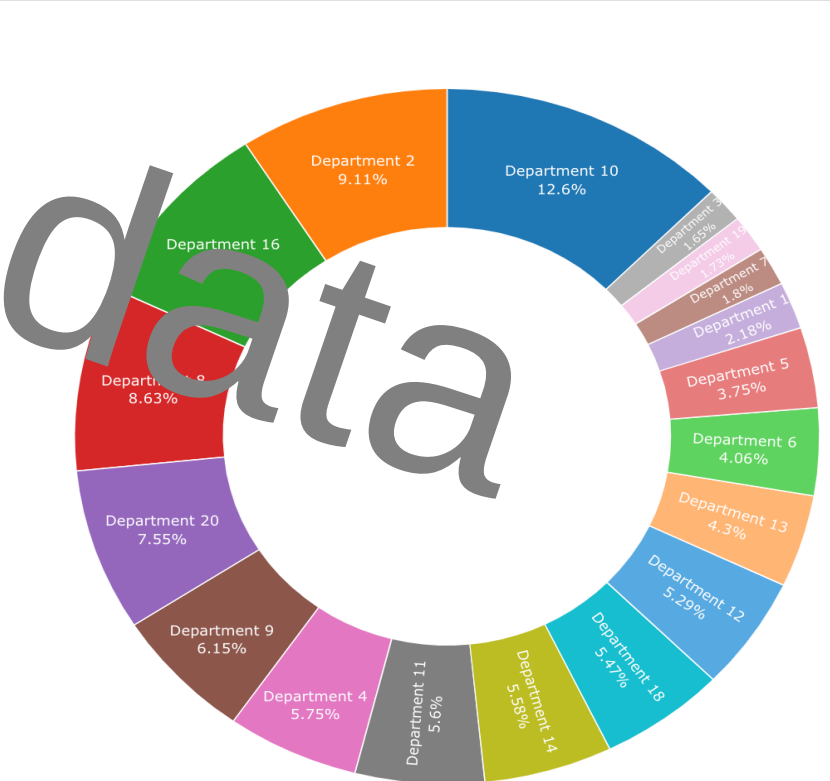
Requested spend by risk profile (FY 21/22)



Requested spend by tier



Percentage of total spend by department



Dummy data

Flex dashboard

https://rmarkdown.rstudio.com/flexdashboard/

BUT!!

CSSF spending review

CSSF spending review

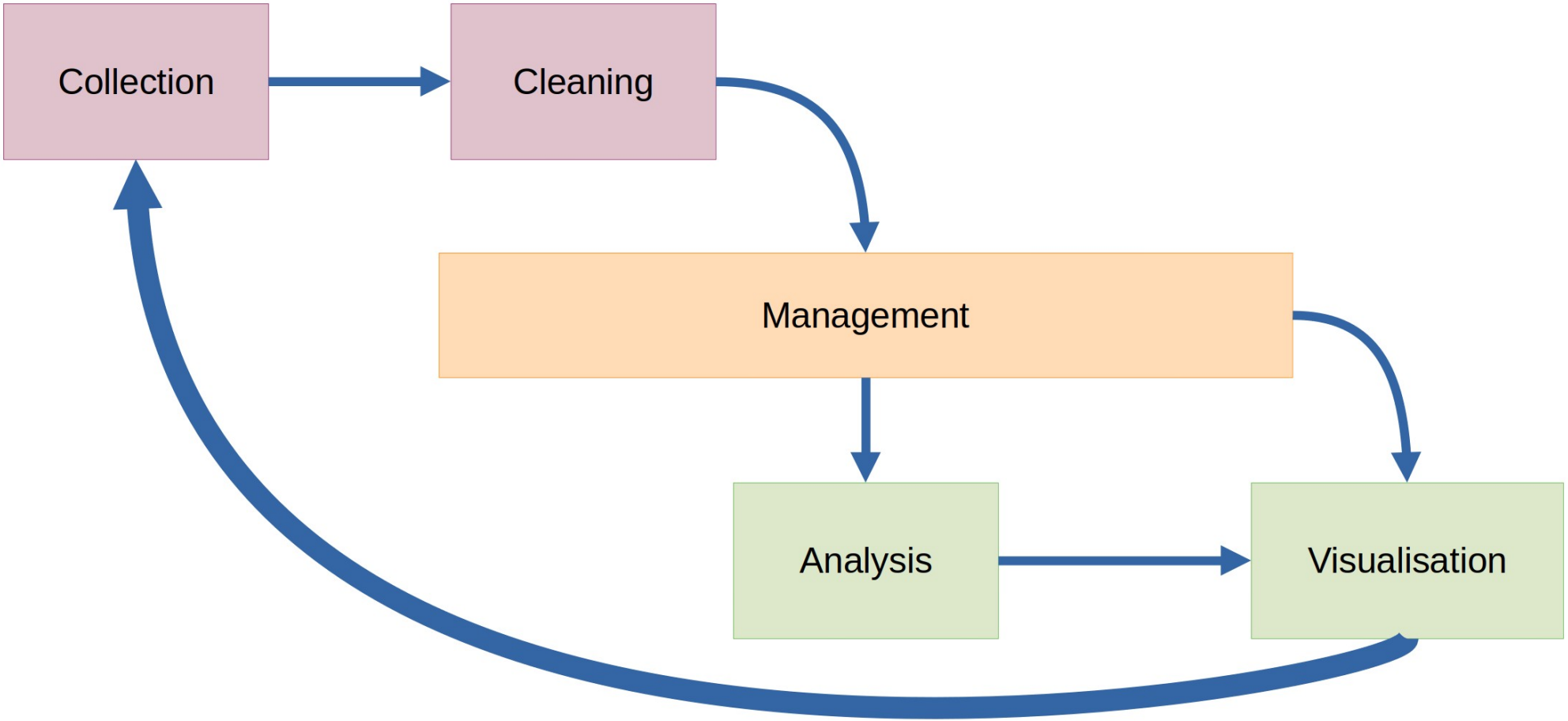
			CSSF Allocation 20/21							CSSF Allocat
5	Portfolio	Discretionary vs Non-discretionary	2021_Non-ODA	2021_ODA	2021_Total	21/22 premod Non-ODA	21/22 premod ODA	21/22_Total FY Budget (premod) (max)	21/22 £ change during moderation	21/22 postmod Non-ODA
6	Portfolio 1	Non-Discretionary	£340,300,000	£330,800,000	£671,100,000	£38,300,000	£15,800,000	£54,100,000	£361,000,000	£201,300,000
7	Portfolio 2	Discretionary	£280,300,000	£320,800,000	£601,100,000	£164,800,000	£132,300,000	£297,100,000	£14,500,000	£206,300,000
8	Portfolio 3	Discretionary	£118,8	£295,800,000	£414,600,000	£92,300,000	£71,800,000	£164,100,000	£267,000,000	£287,800,000
9	Portfolio 4	Discretionary	£256, ,000	£45,300,000	£302,100,000	£41,300,000	£228,800,000	£270,100,000	£170,500,000	£186,800,000
10	Portfolio 5	Discretionary	£32 ,000	£62,800,000	£392,100,000	£142,300,000	£296,800,000	£439,100,000	-£187,500,000	£211,800,000
11	Portfolio 6	Discretionary	£ 300,000	£ ,800,000	£127,100,000	£302,300,000	£110,300,000	£412,600,000	-£121,000,000	£27,300,000
12	Portfolio 7	Discretionary	£ ,300,000	£ ,300,000	£176,600,000	£268,800,000	£91,300,000	£360,100,000	-£21,500,000	£190,300,000
13	Portfolio 8	Discretionary	£ 800,000	£ ,300,0	£560,100,000	£197,300,000	£257,300,000	£454,600,000	-£170,000,000	£242,800,000
14	Portfolio 9	Discretionary	£162,	99,800,	£36 ,000,000	£62,800,000	£244,800,000	£307,600,000	-£142,000,000	£105,800,000
15	Portfolio 10	Discretionary	£46,300,000	20,800 ,00	£26 ,00,00	£235,300,000	£135,300,000	£370,600,000	-£110,500,000	£103,800,000
16	Portfolio 11	Discretionary	£131,800,000	,000	£1 ,100,00	£ ,800,000	£8,800,000	£118,600,000	£72,000,000	£122,800,000
17	Portfolio 12	Discretionary	£61,300,000	£271,800,000	£3 ,100,0	£253,	£175,300,000	£429,100,000	-£176,000,000	£223,800,000
18	Portfolio 13	Discretionary	£136,300,000	£72,800,000	£ ,9 00,0	£334 ,000,000	£122,800,000	£457,600,000	-£51,500,000	£237,300,000
19	Portfolio 14	Discretionary	£175,300,000	£255,300,000	£430,600 ,00	£ ,500,000	£70,300,000	£73,100,000	£143,500,000	£161,800,000
20	Portfolio 15	Discretionary	£45,300,000	£260,300,000	£305,600,000	£ ,300,00	£302, ,000	£318,600 ,0	£91,500,000	£251,300,000
21	Portfolio 16	Non-Discretionary	£227,800,000	£151,800,000	£379,600,000	£ ,800,0	£6 ,000,000	£ 00,00	-£87,000,000	£22,800,000
22	Portfolio 17	Non-Discretionary	£113,300,000	£106,800,000	£220,100,000	£138,300,000	£ ,300,000	£358,6 ,000	£21,000,000	£121,800,000
23	Portfolio 18	Non-Discretionary	£237,800,000	£319,800,000	£557,600,000	£144,800,000	£29,800,000	£174,6 ,000	-£70,500,000	£99,300,000
24	Portfolio 19	Non-Discretionary	£218,300,000	£296,800,000	£515,100,000	£235,800,000	£287,800,000	£523, ,000	-£ 36,5 00	£53,800,000
25	Portfolio 20	Discretionary	£153,800,000	£18,300,000	£172,100,000	£293,800,000	£243,300,000	£53 ,00,0	46,0 ,000	£275,800,000
26	Portfolio 21	Discretionary	£285,800,000	£276,800,000	£562,600,000	£75,800,000	£248,300,000	£32 ,00,0	£17,0 ,000	,800,000
27	Portfolio 22	Discretionary	£109,300,000	£238,800,000	£348,100,000	£99,800,000	£230,800,000	£330,600,0	£49 ,0,000	£ ,300,000
28	Portfolio 23	Non-Discretionary	£24,300,000	£89,300,000	£113,600,000	£169,800,000	£42,300,000	£212,100,000	165 ,00,0	,300,000
29	Portfolio 24	Discretionary	£112,300,000	£287,300,000	£399,600,000	£218,300,000	£88,300,000	£306,600,000	-£94, ,0	65,300,000
30	Portfolio 25	Discretionary	£71,800,000	£281,300,000	£353,100,000	£198,300,000	£26,300,000	£224,600,000	£320,000,000	273,800,000
31	Portfolio 26	Non-Discretionary	£140,300,000	£260,800,000	£401,100,000	£171,800,000	£339,300,000	£511,100,000	-£159,000,000	£134,800,000
32	Portfolio 27	Discretionary	£191,300,000	£133,300,000	£324,600,000	£11,800,000	£92,800,000	£104,600,000	£334,500,000	£336,300,000
33	Portfolio 28	Discretionary	£82,300,000	£256,300,000	£338,600,000	£127,300,000	£190,800,000	£318,100,000	£45,500,000	£165,300,000
34	total		£4,480,400,000	£5,492,400,000	£9,972,800,000	£4,311,400,000	£4,366,400,000	£8,677,800,000	£116,000,000	£4,890,400,000

Dummy data

Generate dummy data

```
sample(seq(1000000, 200000000, by = 100000), 28)
```

CSSF spending review



CSSF spending review

	Portfolio	Discretionary vs Non-discretionary	CSSF 2021_Non-ODA
6	Portfolio 1	Non-Discretionary	£340,300,000
7	Portfolio 2	Discretionary	£280,300,000
8	Portfolio 3	Discretionary	£118,800,000
9	Portfolio 4	Discretionary	£256,800,000
10	Portfolio 5	Discretionary	£329,3[illegible]
11	Portfolio 6	Discretionary	£58[illegible]
12	Portfolio 7	Discretionary	£5[illegible]
13	Portfolio 8	Discretionary	£31[illegible]
14	Portfolio 9	Discretionary	£1[illegible]
15	Portfolio 10	Discretionary	[illegible]
16	Portfolio 11	Discretionary	£1[illegible]
17	Portfolio 12	Discretionary	£61,300,000
18	Portfolio 13	Discretionary	£136,300,000
19	Portfolio 14	Discretionary	£175,300,000
20	Portfolio 15	Discretionary	£45,300,000
21	Portfolio 16	Non-Discretionary	£227,800,000
22	Portfolio 17	Non-Discretionary	£113,300,000
23	Portfolio 18	Non-Discretionary	£237,800,000
24	Portfolio 19	Non-Discretionary	£218,300,000
25	Portfolio 20	Discretionary	£153,800,000
26	Portfolio 21	Discretionary	£285,800,000
27	Portfolio 22	Discretionary	£109,300,000
28	Portfolio 23	Non-Discretionary	£24,300,000
29	Portfolio 24	Discretionary	£112,300,000
30	Portfolio 25	Discretionary	£71,800,000
31	Portfolio 26	Non-Discretionary	£140,300,000
32	Portfolio 27	Discretionary	£191,300,000
33	Portfolio 28	Discretionary	£82,300,000
34		total	£4,480,400,000

port	disc	fin_year	oda	requested
Portfolio 1	FALSE	2021	FALSE	340300000
Portfolio 2	TRUE	2021	FALSE	280300000
Portfolio 3	TRUE	2021	FALSE	118800000
Portfolio 4	TRUE	2021	FALSE	256800000
Portfolio 5	TRUE	2021	FALSE	329300000
Portfolio 6	TRUE	2021	FALSE	58300000
Portfolio 7	TRUE	2021	FALSE	50300000
Portfolio 8	TRUE	2021	FALSE	318800000
Portfolio 9	TRUE	2021	FALSE	162800000
Portfolio 10	TRUE	2021	FALSE	46300000
Portfolio 11	TRUE	2021	FALSE	131800000
Portfolio 12	TRUE	2021	FALSE	61300000
Portfolio 13	TRUE	2021	FALSE	[illegible]
Portfolio 14	TRUE	2021	FALSE	[illegible]
Portfolio 15	TRUE	2021	FALSE	[illegible]
Portfolio 16	FALSE	2021	FALSE	227800000

Dummy data

Read data

```
df = read_excel("data_dummy/summary_learning.xlsx",
           sheet = "CSR Fund Financials – data",
           range = "A5:AE33",
           col_types = c("text", "text",
                         rep("numeric", 12),
                         "text",
                         rep("numeric", 16))) %>%
  janitor::clean_names()
```

Normalise data

```
df %>%
  select(port = portfolio,
         y2021_nonoda = x2021_non_oda,
         y2021_oda = x2021_oda,
         y2022_nonoda = x21_22_premod_non_oda,
         y2022_oda = x21_22_premod_oda) %>%
  gather(var, requested, -port) %>%
  separate(var, c("fin_year", "oda")) %>%
  mutate(oda = if_else(oda == "oda",
                                T, F),
         fin_year = as.numeric(str_remove(fin_year, "y"))) %>%
  write_csv("data_clean/summary_req.csv")
```

https://doi.org/10.1080/00031305.2017.1375989
https://datacarpentry.org/spreadsheet-ecology-lesson/
http://dx.doi.org/10.18637/jss.v059.i10
https://en.wikipedia.org/wiki/Database_normalization

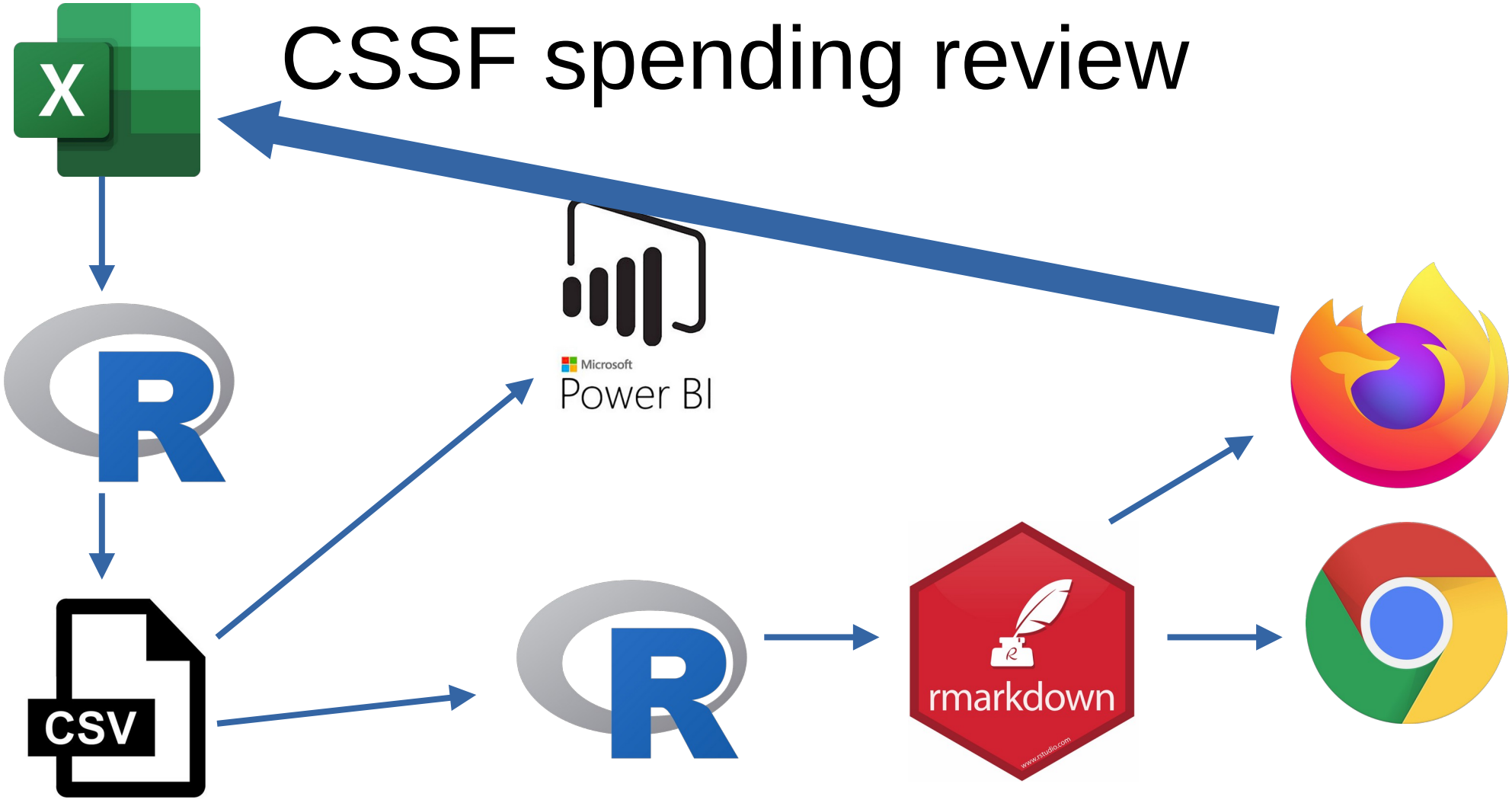
CSSF spending review

	Portfolio	Discretionary vs Non-discretionary	CSSF 2021_Non-ODA
6	Portfolio 1	Non-Discretionary	£340,300,000
7	Portfolio 2	Discretionary	£280,300,000
8	Portfolio 3	Discretionary	£118,800,000
9	Portfolio 4	Discretionary	£256,800,000
10	Portfolio 5	Discretionary	£329,3[illegible]
11	Portfolio 6	Discretionary	£58[illegible]
12	Portfolio 7	Discretionary	£5[illegible]
13	Portfolio 8	Discretionary	£31[illegible]
14	Portfolio 9	Discretionary	£1[illegible]
15	Portfolio 10	Discretionary	[illegible]
16	Portfolio 11	Discretionary	£1[illegible]
17	Portfolio 12	Discretionary	£61,300,000
18	Portfolio 13	Discretionary	£136,300,000
19	Portfolio 14	Discretionary	£175,300,000
20	Portfolio 15	Discretionary	£45,300,000
21	Portfolio 16	Non-Discretionary	£227,800,000
22	Portfolio 17	Non-Discretionary	£113,300,000
23	Portfolio 18	Non-Discretionary	£237,800,000
24	Portfolio 19	Non-Discretionary	£218,300,000
25	Portfolio 20	Discretionary	£153,800,000
26	Portfolio 21	Discretionary	£285,800,000
27	Portfolio 22	Discretionary	£109,300,000
28	Portfolio 23	Non-Discretionary	£24,300,000
29	Portfolio 24	Discretionary	£112,300,000
30	Portfolio 25	Discretionary	£71,800,000
31	Portfolio 26	Non-Discretionary	£140,300,000
32	Portfolio 27	Discretionary	£191,300,000
33	Portfolio 28	Discretionary	£82,300,000
34		total	£4,480,400,000

port	disc	fin_year	oda	requested
Portfolio 1	FALSE	2021	FALSE	340300000
Portfolio 2	TRUE	2021	FALSE	280300000
Portfolio 3	TRUE	2021	FALSE	118800000
Portfolio 4	TRUE	2021	FALSE	256800000
Portfolio 5	TRUE	2021	FALSE	329300000
Portfolio 6	TRUE	2021	FALSE	58300000
Portfolio 7	TRUE	2021	FALSE	50300000
Portfolio 8	TRUE	2021	FALSE	318800000
Portfolio 9	TRUE	2021	FALSE	162800000
Portfolio 10	TRUE	2021	FALSE	46300000
Portfolio 11	TRUE	2021	FALSE	131800000
Portfolio 12	TRUE	2021	FALSE	61300000
Portfolio 13	TRUE	2021	FALSE	[illegible]
Portfolio 14	TRUE	2021	FALSE	[illegible]
Portfolio 15	TRUE	2021	FALSE	[illegible]
Portfolio 16	FALSE	2021	FALSE	227800000

Dummy data

CSSF spending review



Summary

- Understand client expectations (spreadsheet/dashboard/etc.)
- Separate out data workflow (valuable on all project sizes)
- Regular data formats are easier
- Automation saves times and reduces mistakes
- (contingency) Plan for working with spreadsheets
- Anything is possible with good foundations!

Mike Spencer

@MikeRSpencer
mikerspencer.com

Job fairy!

- 4x data analysts
- Pilot project – improve data in decision making
- ETL, analysis, visualisation, design & learning
- R, Python, power BI

https://www.integrityglobal.com/our-work/job-openings/jobs/data-analysts-gmel-partnership/

10 INTEGRITY | CELEBRATING 10 YEARS OF SERVICE

CSSF GLOBAL MEL Partnership

NIRAS LTS INTERNATIONAL