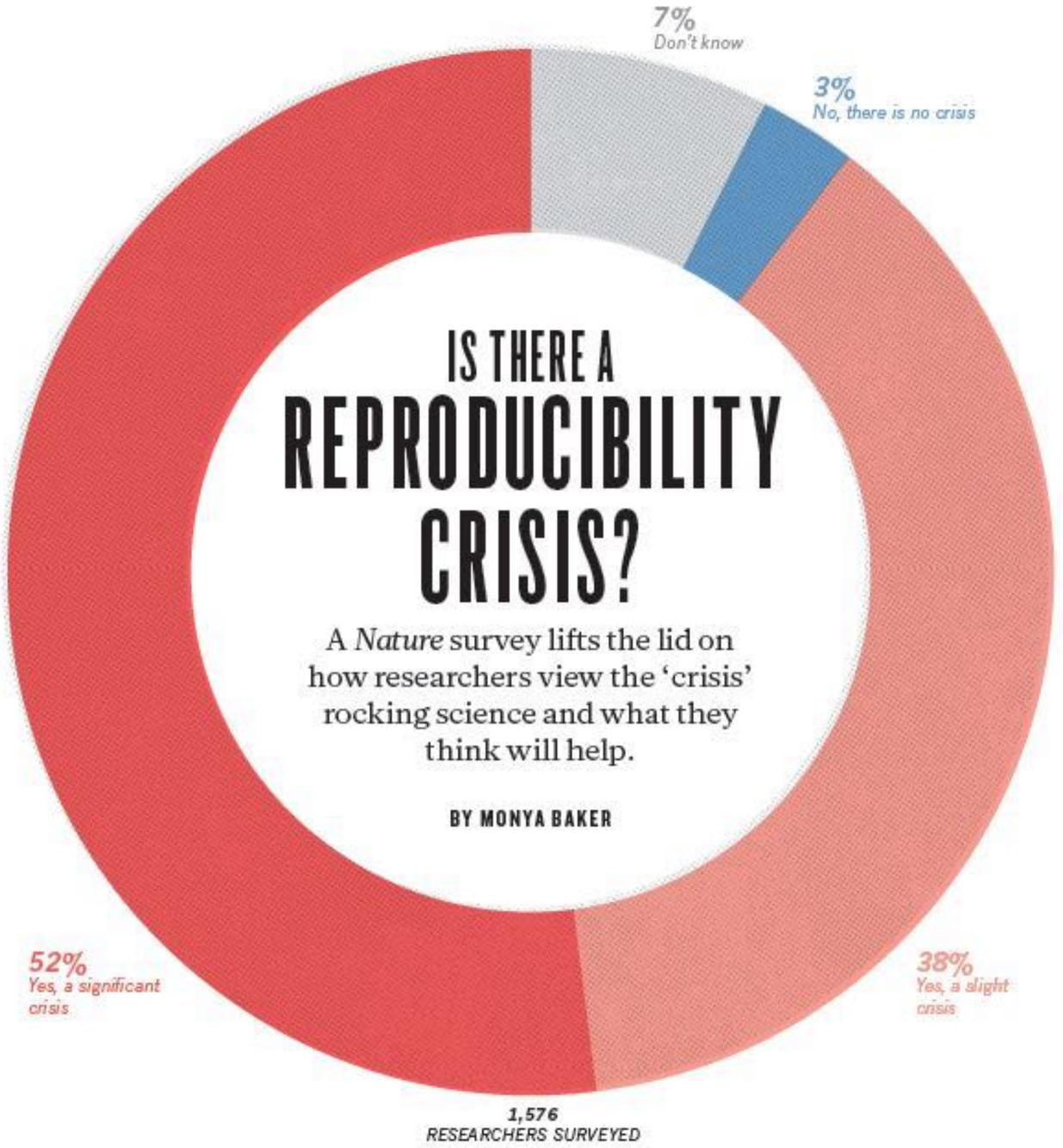


Baker 2016, Nature

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Reproducible research is still a challenge

Rich FitzJohn | Matt Pennell | Amy Zanne | Will Cornwell

June 9, 2014

https://ropensci.org/blog/2014/06/09/reproducibility

why is this still hard?

complexity

why is this still hard?

tooling

why is this still hard?

incentives + motivation

nobody cares

(Apart from people doing it out of the goodness of their hearts of course)

nobody cares?

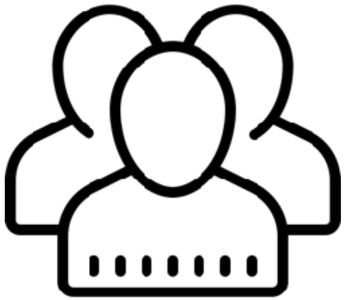
(Apart from people doing it out of the goodness of their hearts of course)

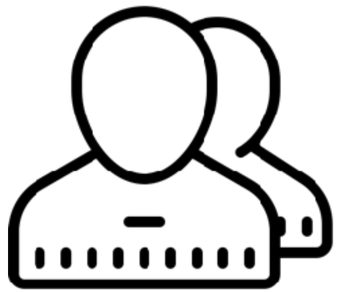
VACCINE IMPACT MODELLING CONSORTIUM

BILL & MELINDA GATES foundation

Gavi The Vaccine Alliance

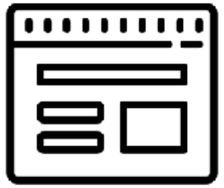
https://vaccineimpact.org

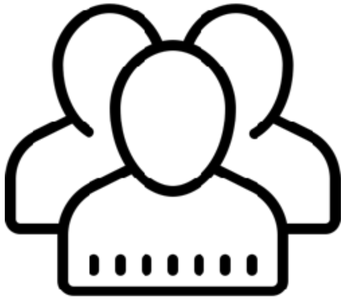


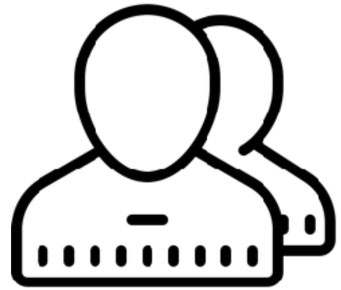


BILL & MELINDA GATES foundation

Gavi
The Vaccine Alliance

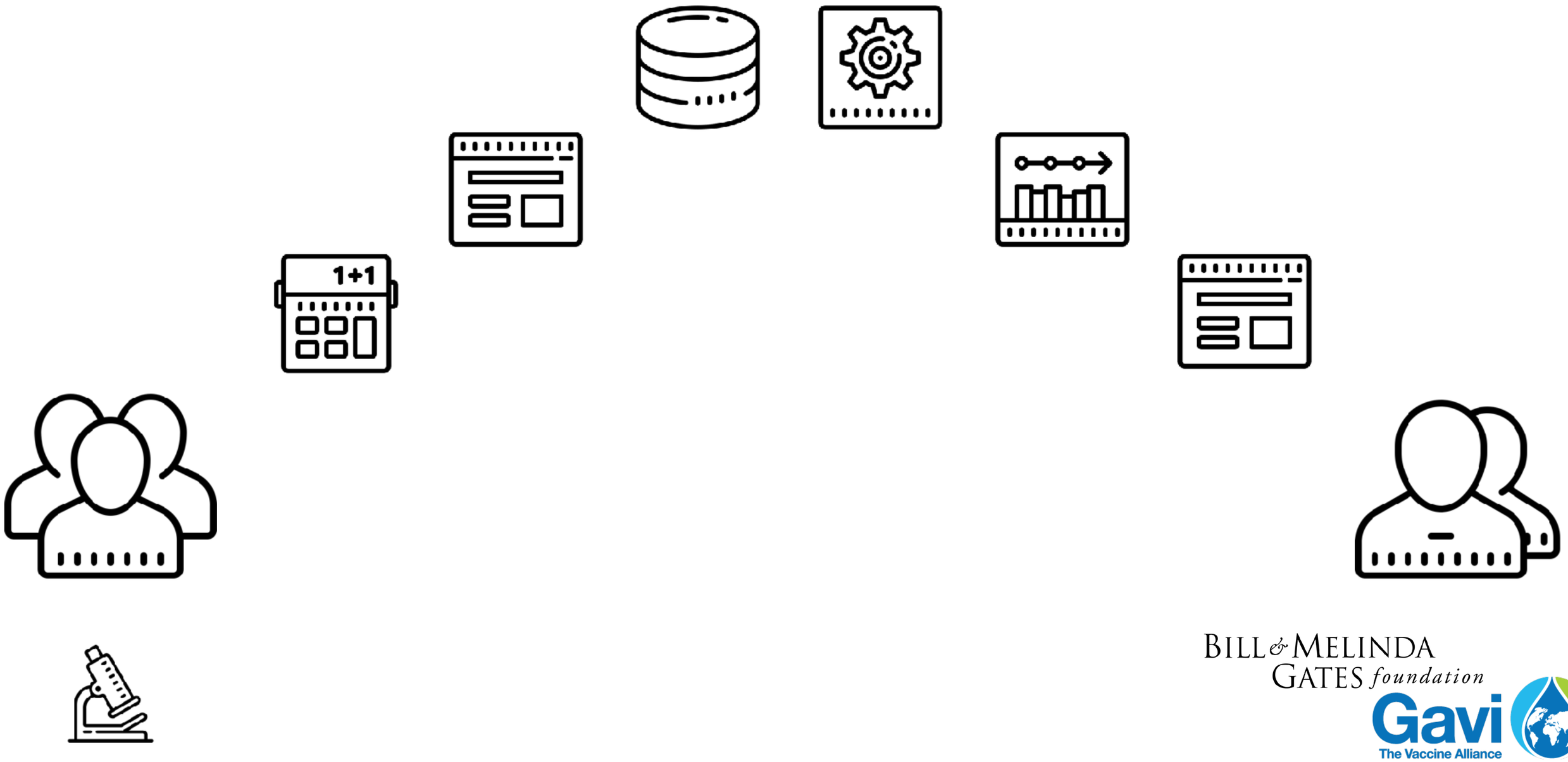






BILL & MELINDA GATES foundation

Gavi The Vaccine Alliance



BILL & MELINDA GATES foundation

Gavi
The Vaccine Alliance

Alex Hill

Emma Russell

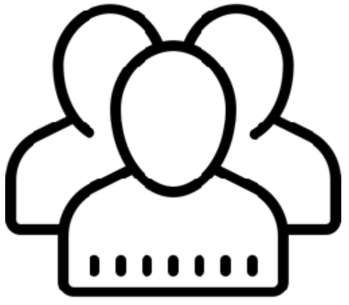
Giovanni Charles

James Thompson

Rob Ashton

Wes Hinsley

reside-ic.github.io



BILL & MELINDA GATES foundation

Gavi The Vaccine Alliance

Our goal

Code written by domain scientists runs on multiple machines, first time

Trace back to origins of changes in any report

knitr

www.rstudio.com

rmarkdown

www.rstudio.com

RStudio

Go to file/function | Addins | Project

example.Rmd

Knit | Insert | Run

```
1  ---
2  title: "Example"
3  output: html_document
4  ---
5
6  ```{r setup, include=FALSE}
7  knitr::opts_chunk$set(echo = TRUE)
8  ```
9
10 ## R Markdown
11
12 This is an R Markdown document. Markdown is a simple formatting
   syntax for authoring HTML, PDF, and MS Word documents. For more
   details on using R Markdown see <http://rmarkdown.rstudio.com>.
13
14 When you click the **Knit** button a document will be generated
   that includes both content as well as the output of any
   embedded R code chunks within the document. You can embed an R
   code chunk like this:
15
16 ```{r cars}
17 summary(cars)
18 ```
19
20 ## Including Plots
21
22 You can also embed plots, for example:
23
24 ```{r pressure, echo=FALSE}
25 plot(pressure)
26 ```
27
28 Note that the `echo = FALSE` parameter was added to the code
   chunk to prevent printing of the R code that generated the
   plot.
29
```

2:16 | Example | R Markdown

Console

~/example.html

example.html | Open in Browser | Find | Publish

Example

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:

How to draw an Owl.

"A fun and creative guide for beginners"

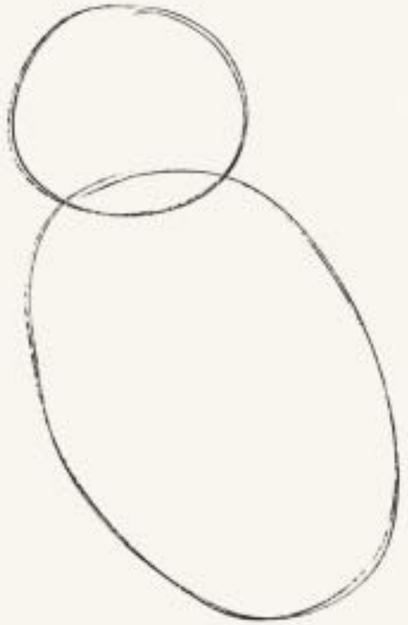
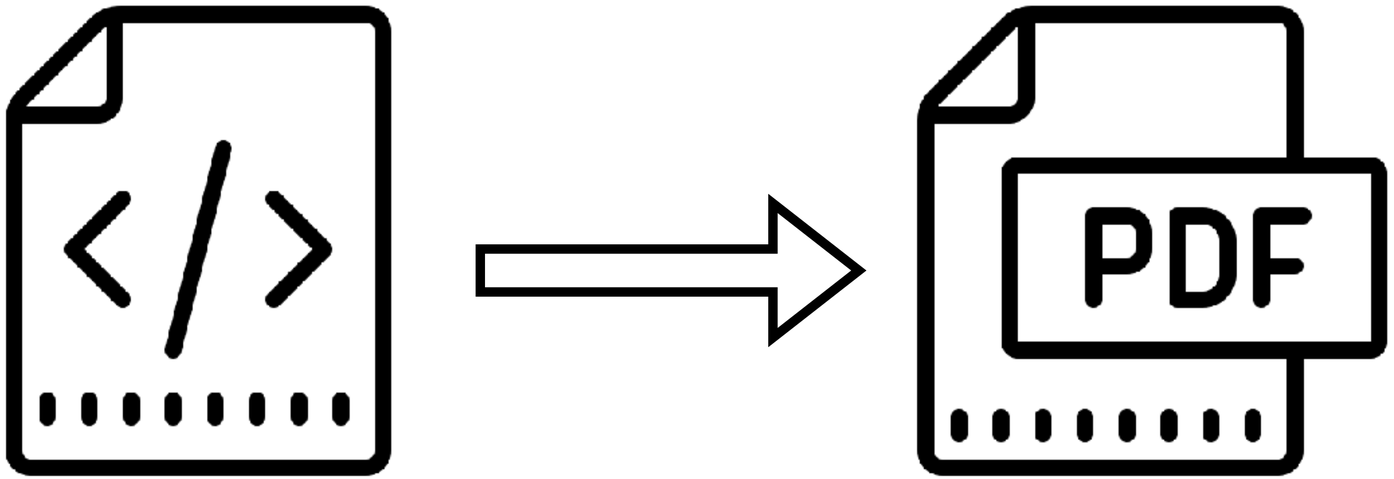
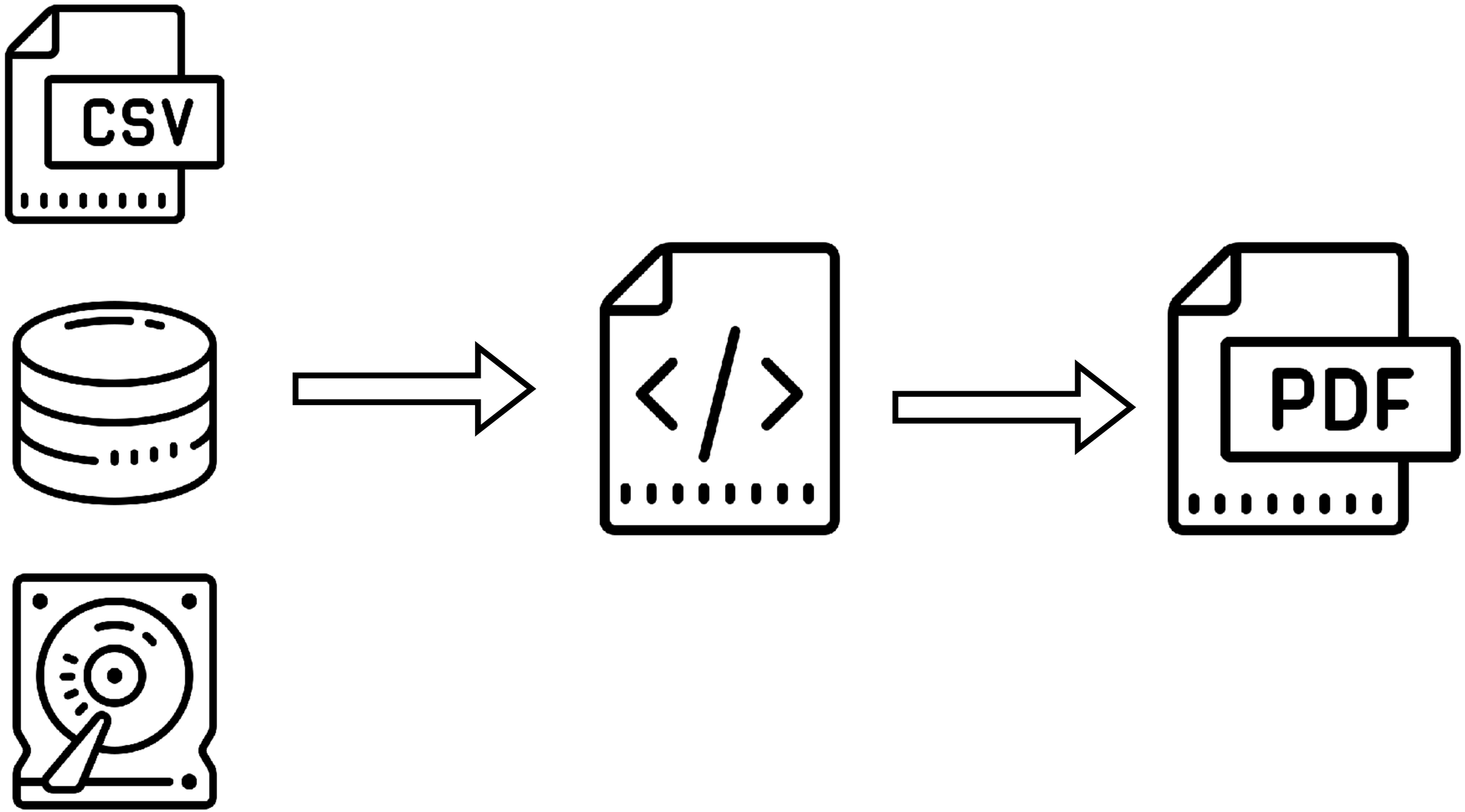


Fig 1. Draw two circles

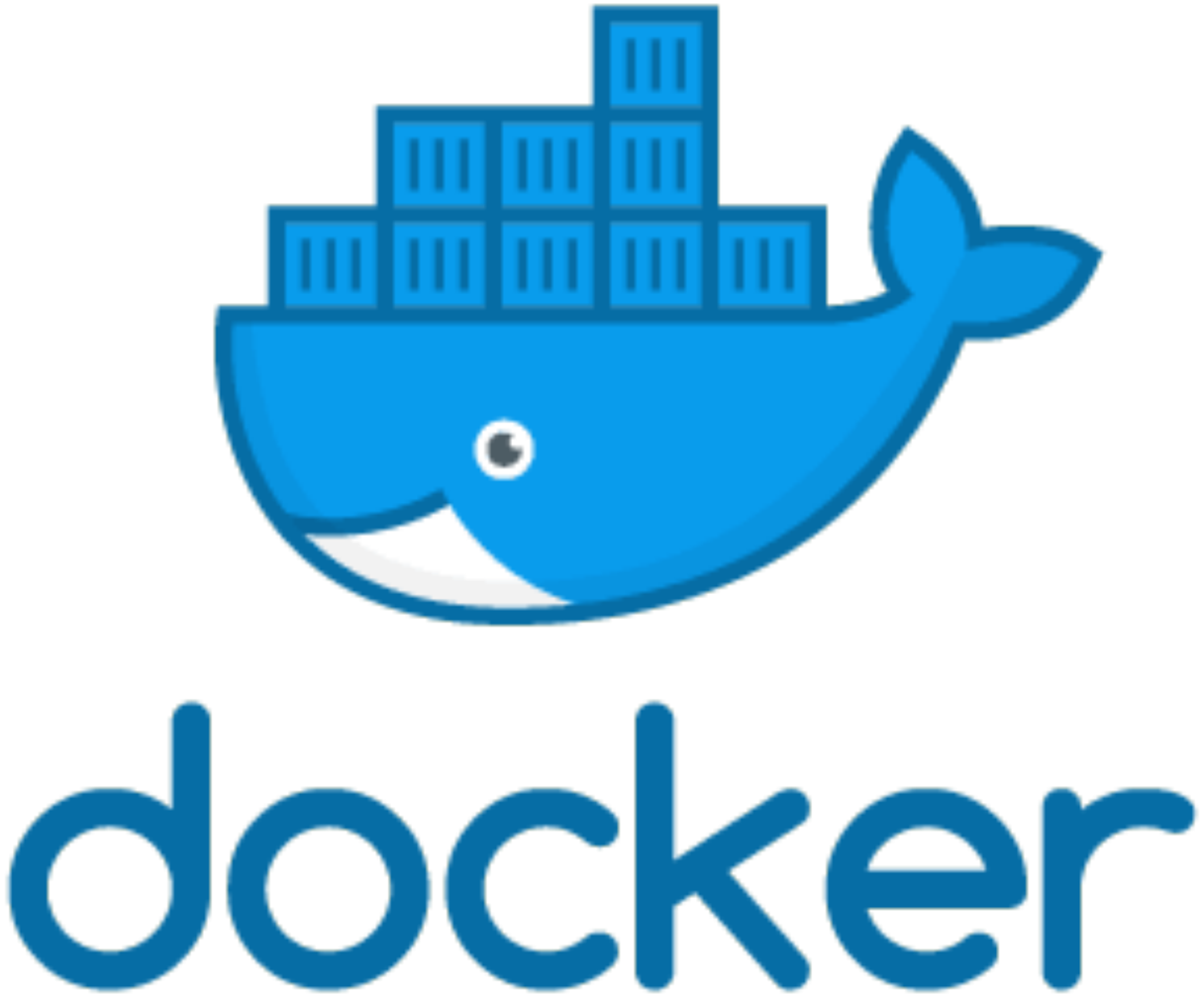
Fig 2. Draw the rest of the damn Owl

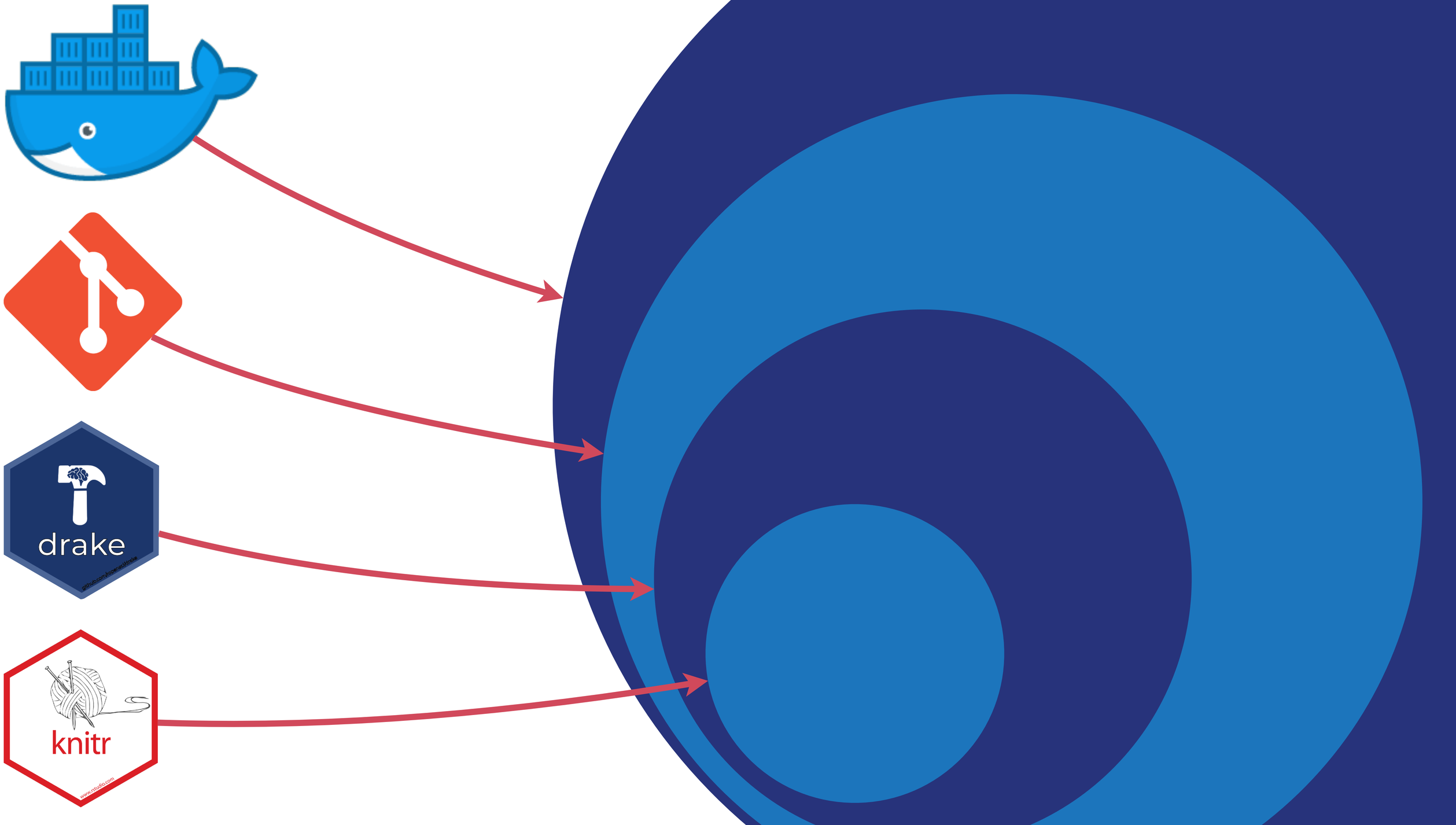

PDF


CSV
PDF

drake
github.com/ropensci/drake

git


docker


drake
knitr

Go To Statement Considered Harmful

Key Words and Phrases: go to statement, jump instruction, branch instruction, conditional clause, alternative clause, repetitive clause, program intelligibility, program sequencing
CR Categories: 4.22, 5.23, 5.24

EDITOR:

For a number of years I have been familiar with the observation that the quality of programmers is a decreasing function of the density of **go to** statements in the programs they produce. More recently I discovered why the use of the **go to** statement has such disastrous effects, and I became convinced that the **go to** statement should be abolished from all "higher level" programming languages (i.e. everything except, perhaps, plain machine code).

```
add <- function(a, b) {
  a + b
}
```

https://github.com/vimc/orderly

```
expect_equal(
  add(1, 3),
  4)
```

https://github.com/vimc/orderly

```
#' @param a,b inputs
#' @return the sum
#' @export
```

https://github.com/vimc/orderly

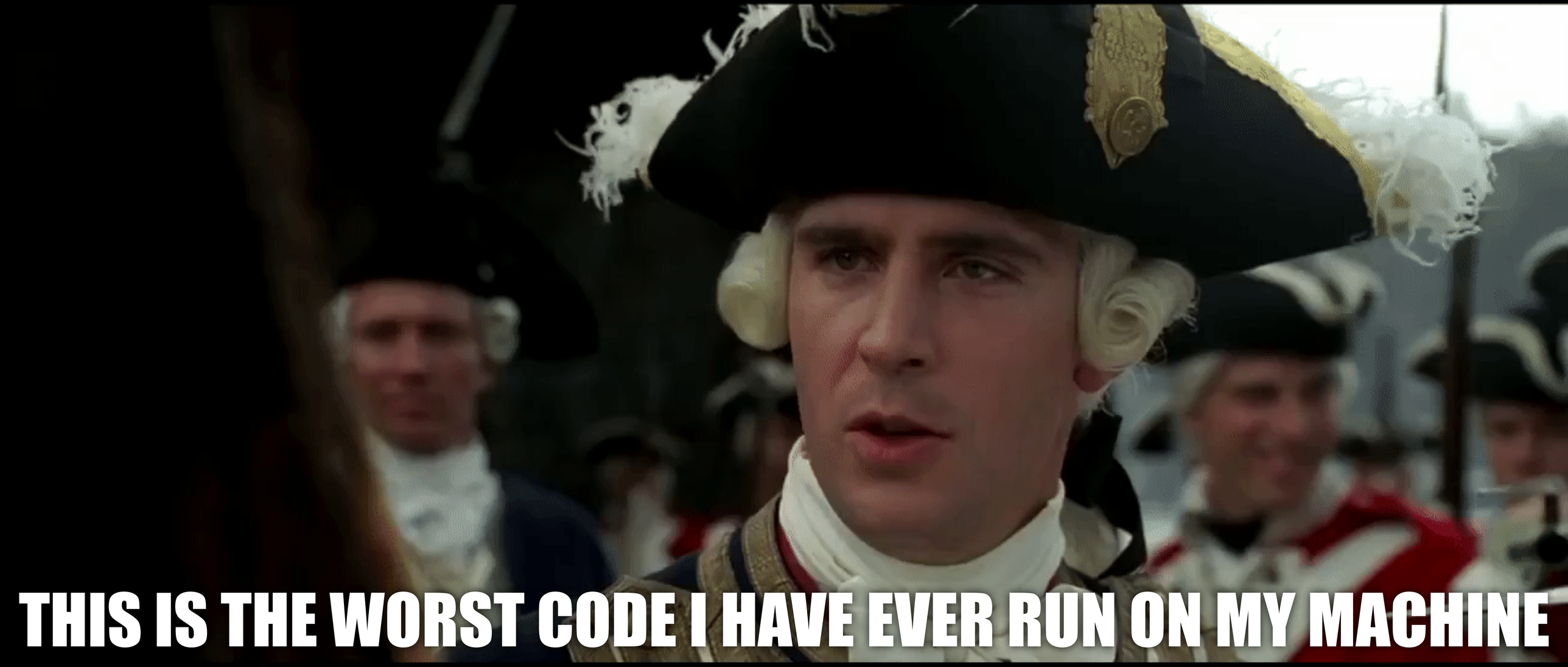
Our goal

Code written by domain scientists runs on multiple machines, first time

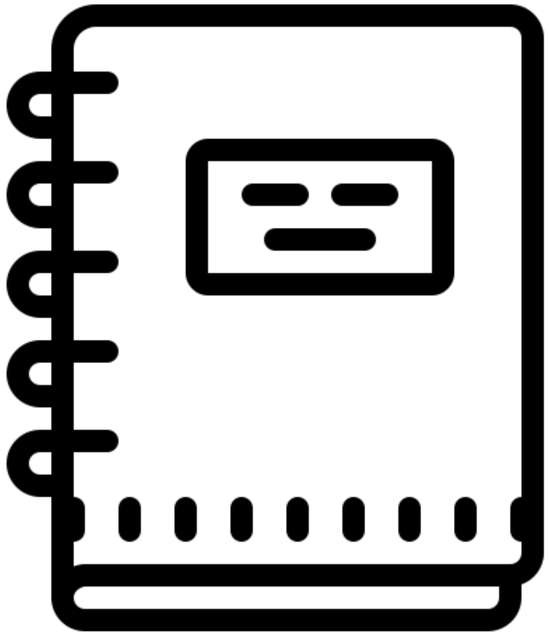
Trace back to origins of changes in any report

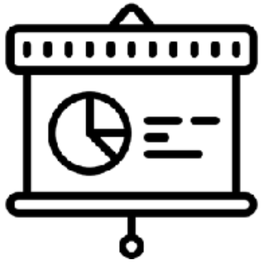
Our apprach

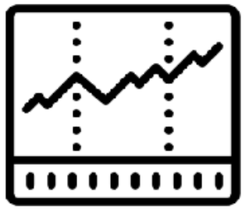
Blackbox the analysis like a function

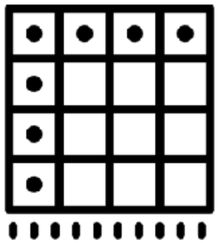

THIS IS THE WORST CODE I HAVE EVER RUN ON MY MACHINE

BUT IT DOES RUN ON YOUR MACHINE

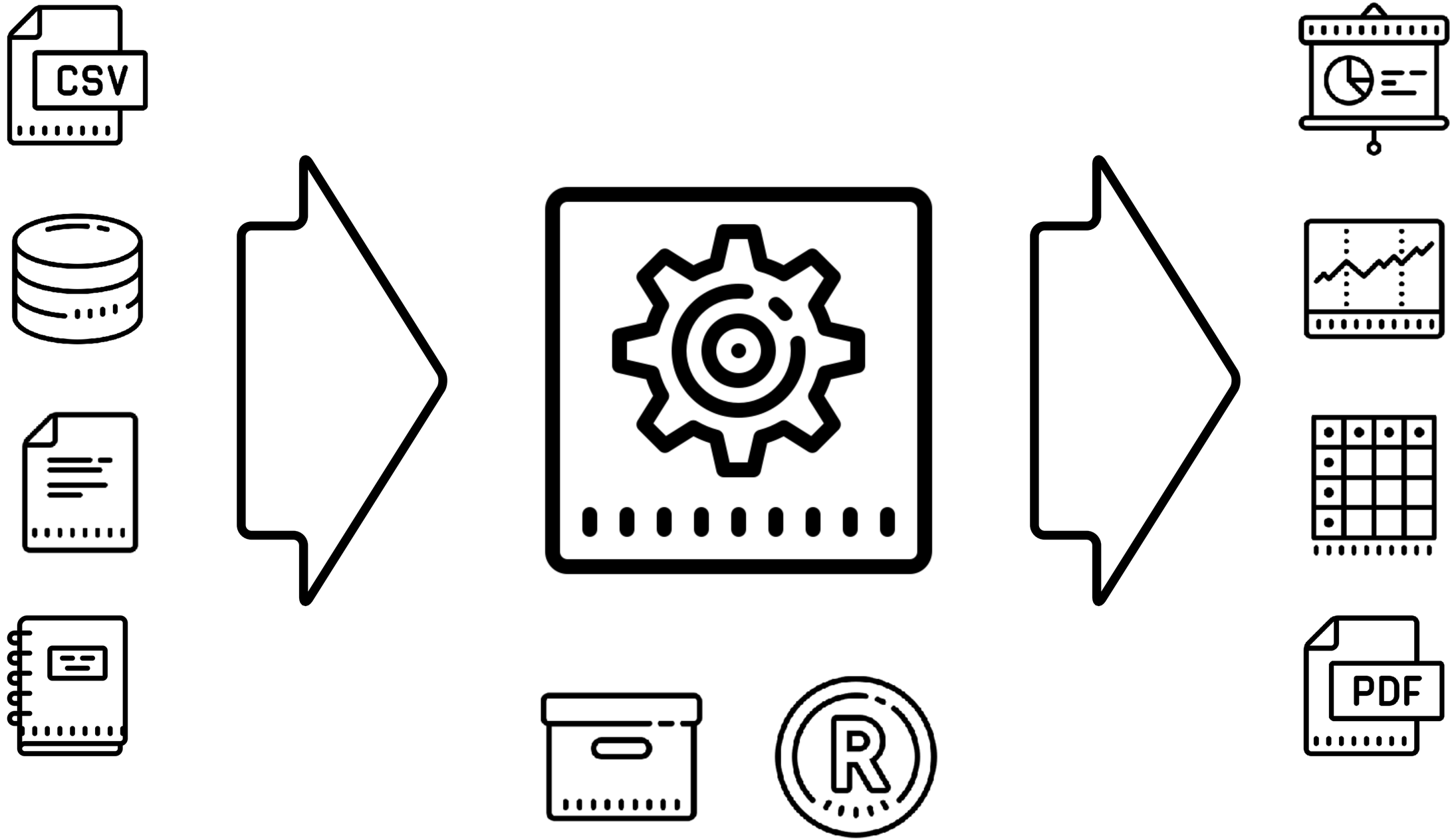








PDF


CSV
R
PDF

orderly

https://github.com/vimc/orderly

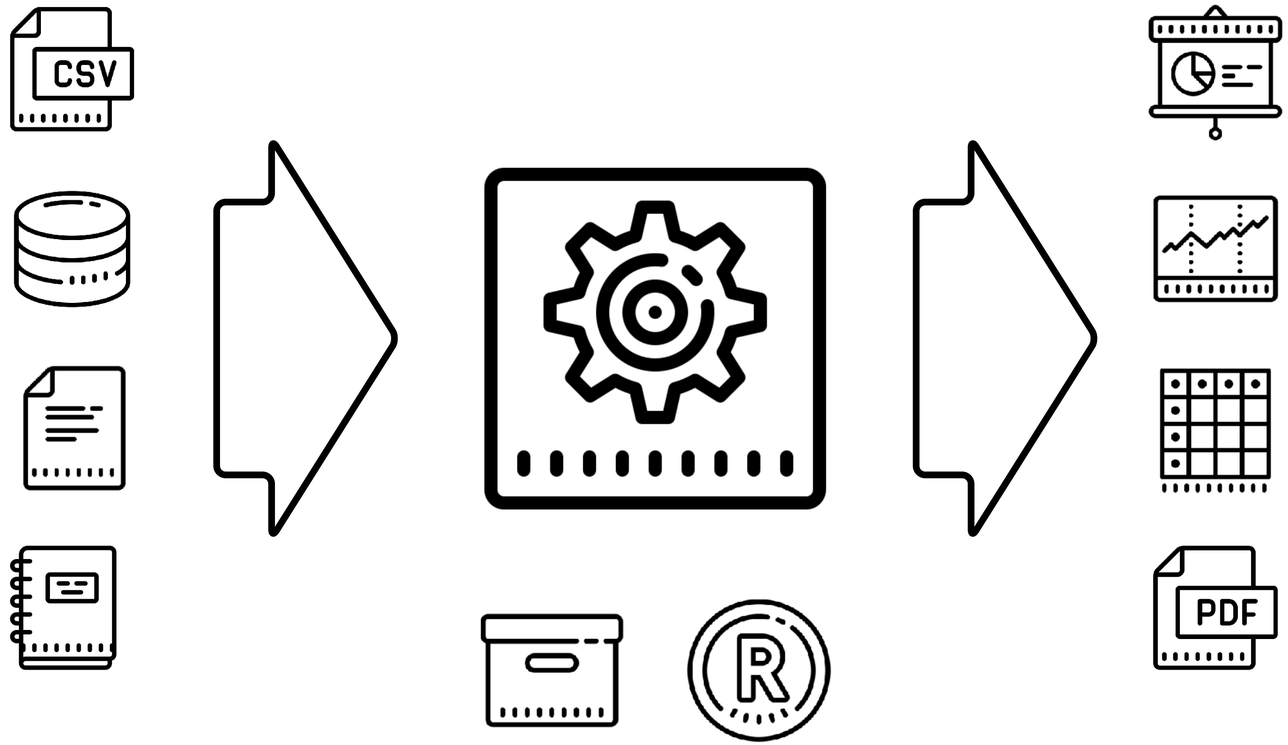
```
data:
  summary:
    query: SELECT * FROM ...

resources:
  - support.R
  - metadata.csv

packages:
  - ggplot2
  - knitr

script: script.R

artefacts:
  - report:
      description: Summary of results
      filenames: summary.pdf
  - data:
      description: Processed data for further use
      filenames: data.csv
```



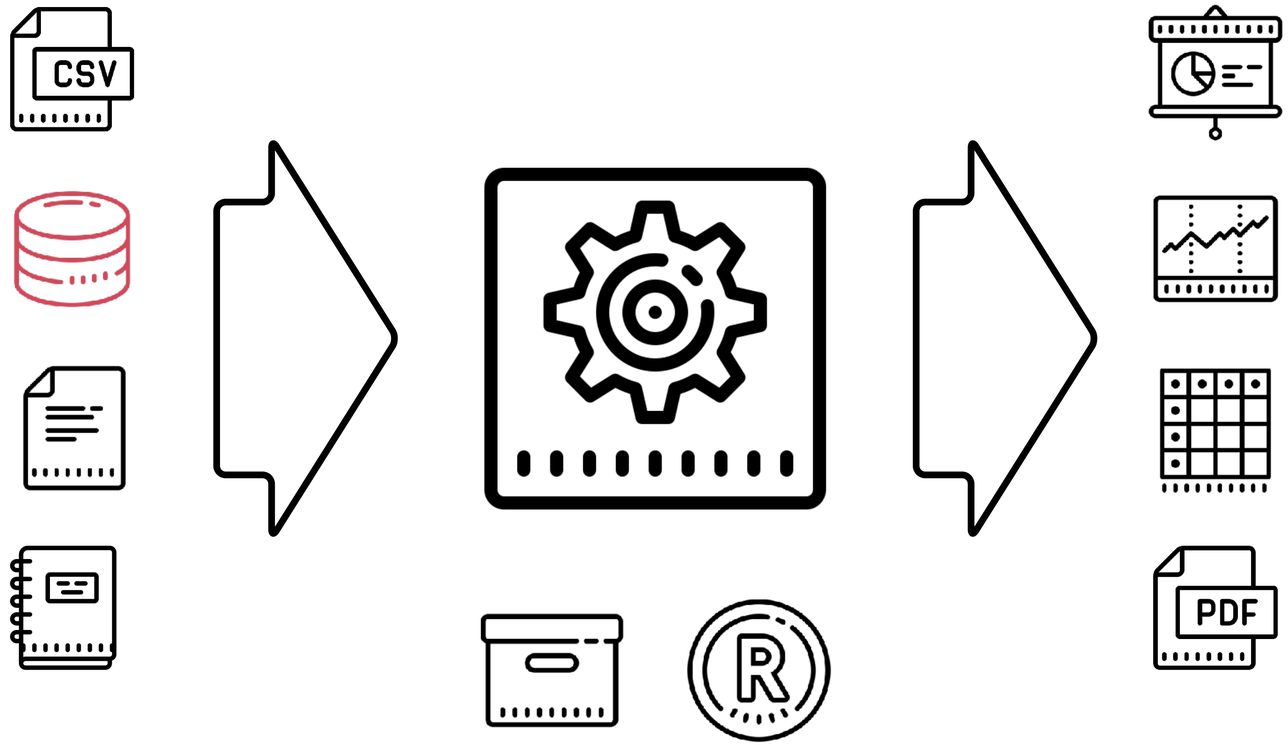
```
data:
  summary:
    query: SELECT * FROM ...

resources:
  - support.R
  - metadata.csv

packages:
  - ggplot2
  - knitr

script: script.R

artefacts:
  - report:
      description: Summary of results
      filenames: summary.pdf
  - data:
      description: Processed data for further use
      filenames: data.csv
```



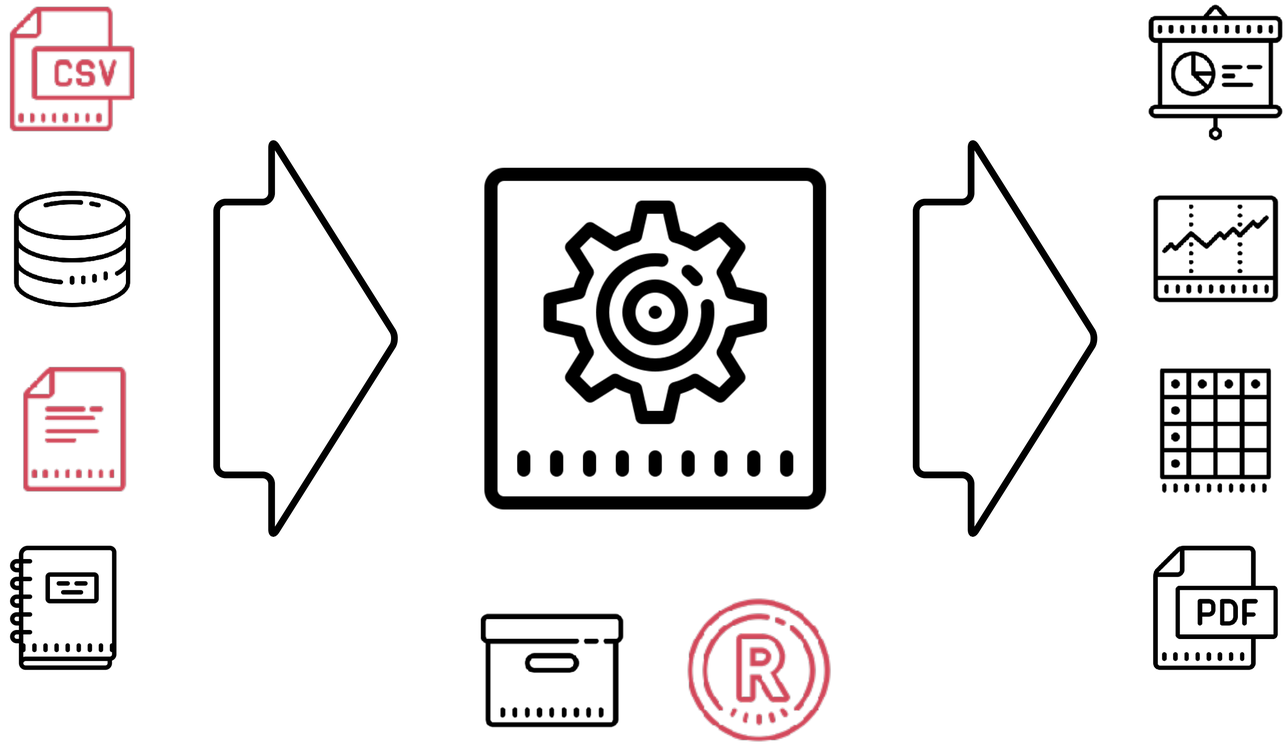
```
data:
  summary:
    query: SELECT * FROM ...

resources:
  - support.R
  - metadata.csv

packages:
  - ggplot2
  - knitr

script: script.R

artefacts:
  - report:
      description: Summary of results
      filenames: summary.pdf
  - data:
      description: Processed data for further use
      filenames: data.csv
```



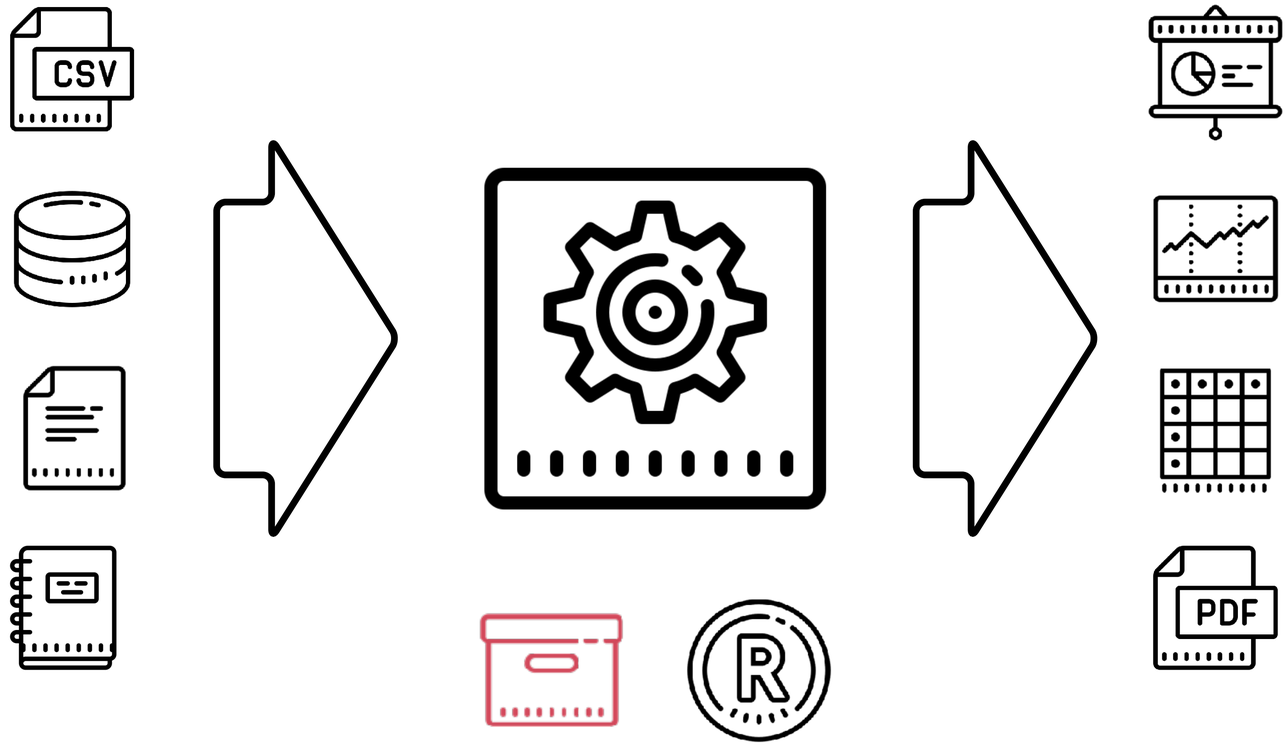
```yaml
data:
  summary:
    query: SELECT * FROM ...

resources:
  - support.R
  - metadata.csv

packages:
  - ggplot2
  - knitr

script: script.R

artefacts:
  - report:
      description: Summary of results
      filenames: summary.pdf
  - data:
      description: Processed data for further use
      filenames: data.csv
```



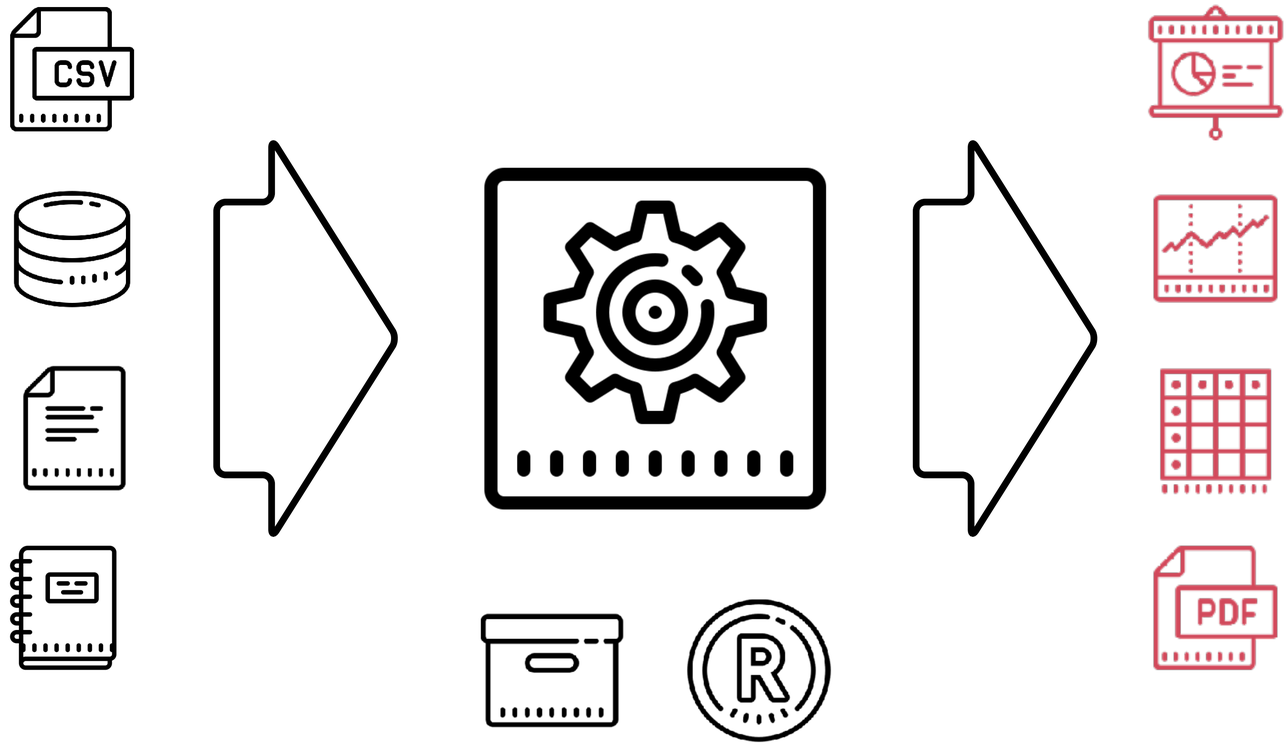
```yaml
data:
  summary:
    query: SELECT * FROM ...

resources:
  - support.R
  - metadata.csv

packages:
  - ggplot2
  - knitr

script: script.R

artefacts:
  - report:
      description: Summary of results
      filenames: summary.pdf
  - data:
      description: Processed data for further use
      filenames: data.csv
```



```
orderly_config.yml
src/
  myreport/
    orderly.yml
    script.R
    support.R
    metadata.csv
```

```
orderly_config.yml
src/
  myreport/
    orderly.yml
    script.R
    support.R
    metadata.csv
archive/
  myreport/
    20190204-143204-f5aa3bc9/
      orderly.yml
      script.R
      support.R
      metadata.csv
      summary.pdf
      data.csv
      orderly_run.rds
```

orderly run myreport

orderly run myreport

```
orderly_config.yml
src/
  myreport/
    orderly.yml
    script.R
    support.R
    metadata.csv
archive/
  myreport/
    20190204-143204-f5aa3bc9/
    20190204-192249-3bc9f5aa/
      orderly.yml
      script.R
      support.R
      metadata.csv
      summary.pdf
      data.csv
      orderly_run.rds
```

Automate the boring bits

This is embarrassingly simple

Interface for **Stakeholders**

Find a report

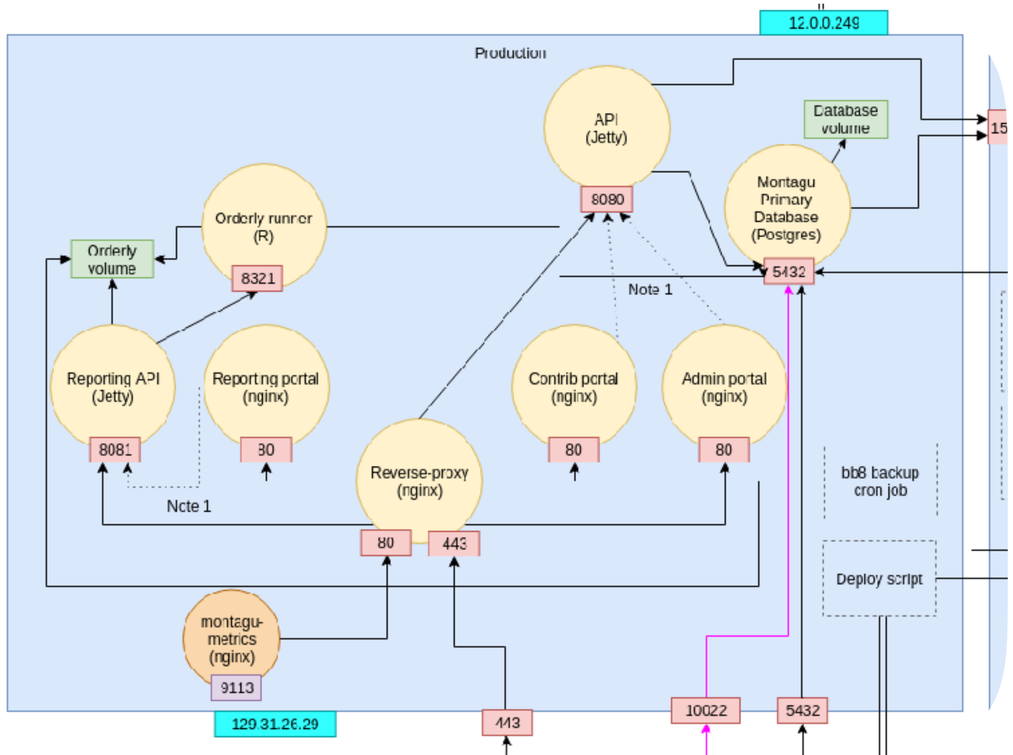
Click on a column heading to sort by that field. Hold shift to multi-sort.

Collapse all reports / Expand all reports

Name	Version	Status	Author	Requester
Type to filter...	Mar 1, to Feb 6, Type to filter by id...	All	Type to filter...	Type to filter...
▼ native-diagnostics-burden-report-drafts 5 versions: view latest			Science Team	VIMC
	Tue Feb 05 2019 latest (20190205-151702-1ba5e47a)	published	Science Team	VIMC
	Thu Jan 31 2019 out-dated (20190131-162935-86a83d30)	published	Science Team	VIMC
	Thu Jan 31 2019 out-dated (20190131-123847-53fe189e)	published	Science Team	VIMC
	Mon Jan 28 2019 out-dated (20190128-151914-a2c1edce)	published	Science Team	VIMC
	Mon Jan 28 2019 out-dated (20190128-125432-31d88275)	internal	Science Team	VIMC

Interface for

Engineering team

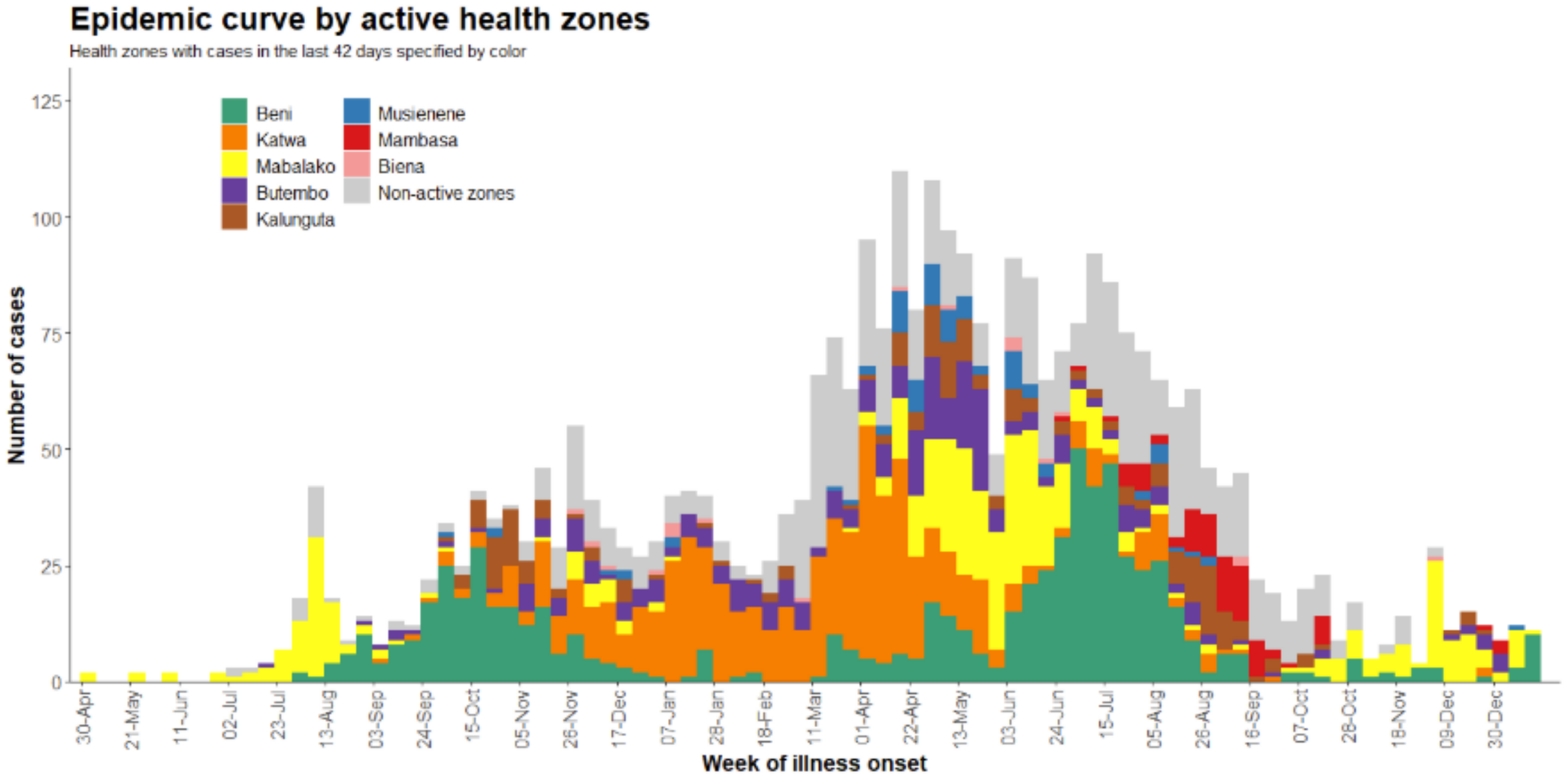


docker

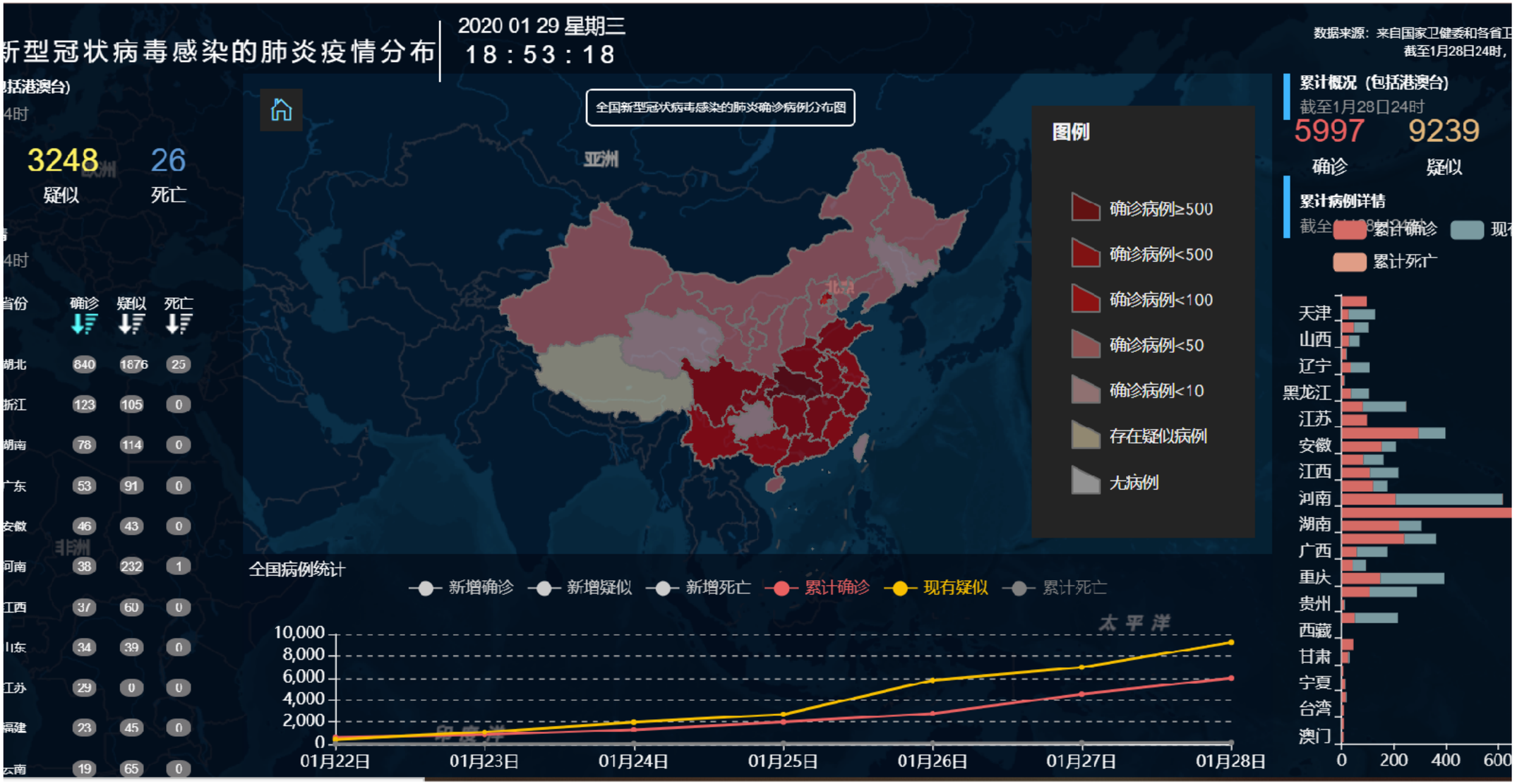
Interface for **Science team**

```
install.packages("orderly")
orderly::orderly_new("myreport")
```

Ebola in DRC


Epidemic curve by active health zones
Health zones with cases in the last 42 days specified by color
Beni
Katwa
Mabalako
Butembo
Kalunguta
Musienene
Mambasa
Biena
Non-active zones
Number of cases
0
25
50
75
100
125
30-Apr
21-May
11-Jun
02-Jul
23-Jul
13-Aug
03-Sep
24-Sep
15-Oct
05-Nov
26-Nov
17-Dec
07-Jan
28-Jan
18-Feb
11-Mar
01-Apr
22-Apr
13-May
03-Jun
24-Jun
15-Jul
05-Aug
26-Aug
16-Sep
07-Oct
28-Oct
18-Nov
09-Dec
30-Dec
Week of illness onset

2019-nCoV


新型冠状病毒感染的肺炎疫情分布
2020 01 29 星期三
18 : 53 : 18
数据来源：来自国家卫健委和各省卫
截至1月28日24时，
全国新型冠状病毒感染的肺炎确诊病例分布图
亚洲
北京
图例
确诊病例≥500
确诊病例<500
确诊病例<100
确诊病例<50
确诊病例<10
存在疑似病例
无病例
累计概况（包括港澳台）
截至1月28日24时
5997
9239
确诊
疑似
累计病例详情
累计确诊
累计死亡
天津
山西
辽宁
黑龙江
江苏
安徽
江西
河南
湖南
广西
重庆
贵州
西藏
甘肃
宁夏
台湾
澳门
0
200
400
600
3248
疑似
26
死亡
省份
确诊
疑似
死亡
湖北 840 1676 25
浙江 123 105 0
湖南 78 114 0
广东 53 91 0
安徽 46 43 0
河南 38 232 1
江西 37 60 0
山东 34 39 0
江苏 29 0 0
福建 23 45 0
云南 19 65 0
全国病例统计
新增确诊
新增疑似
新增死亡
累计确诊
现有疑似
累计死亡
太平洋
10,000
8,000
6,000
4,000
2,000
0
01月22日
01月23日
01月24日
01月25日
01月26日
01月27日
01月28日

Lessons learnt

you can blackbox too much

code reuse is really hard

reproducibility can be easy

problems are as social as technical

Work the way people want to work

Acknowledgements

MRC Centre for Global Infectious Disease Analysis
rOpenSci

Vaccine Impact Modelling Consortium
https://vaccineimpact.org

Resources

https://vimc.github.io/orderly
https//github.com/vimc/orderly-web
https://vaccineimpact.org

richfitz

rgfitzjohn

reside-ic.github.io