Prototyping with R packages

Irene Steves & Yogev Herz

@i_steves @yogevmh

2020-08-19, R-Ladies Amsterdam

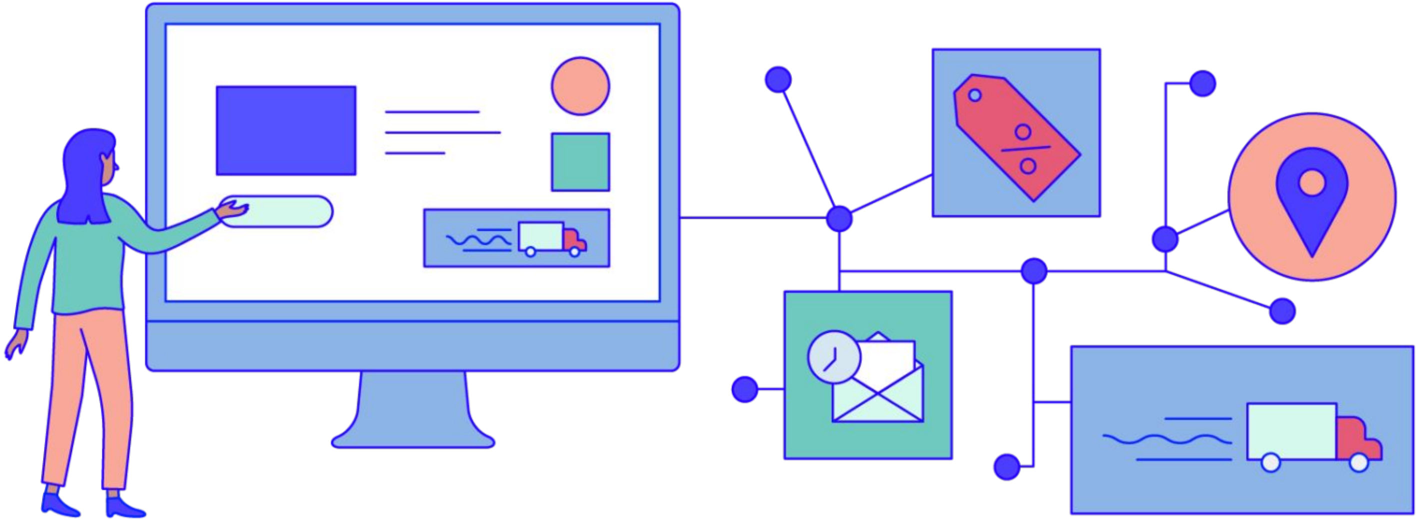
riskified

About us

- Data Science & Research department at Riskified, based in Tel Aviv
- Ecology & evolutionary biology background
- Fans of Bob the dog

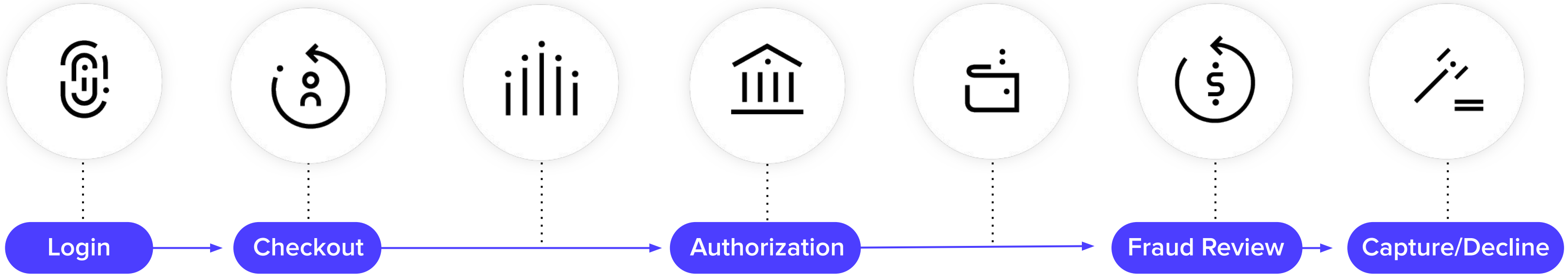
Riskified

e-Commerce fraud prevention for online merchants:
verify orders at checkout and take liability for bad decisions



Riskified

We use machine learning models to prevent fraud throughout the shopping process



What does it mean to put into production?

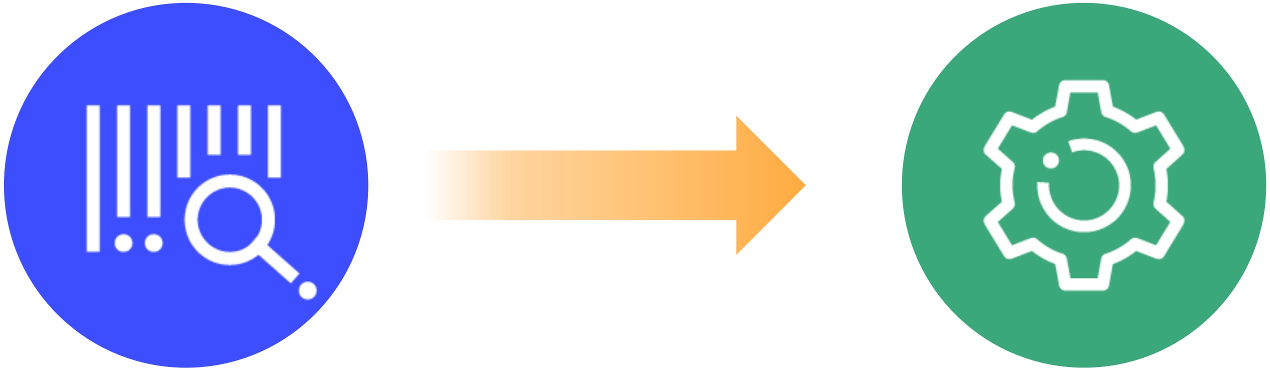
ANALYSIS

Running code once to produce a result

BUILD

Writing code that is continuously running

What does it mean to put into production?



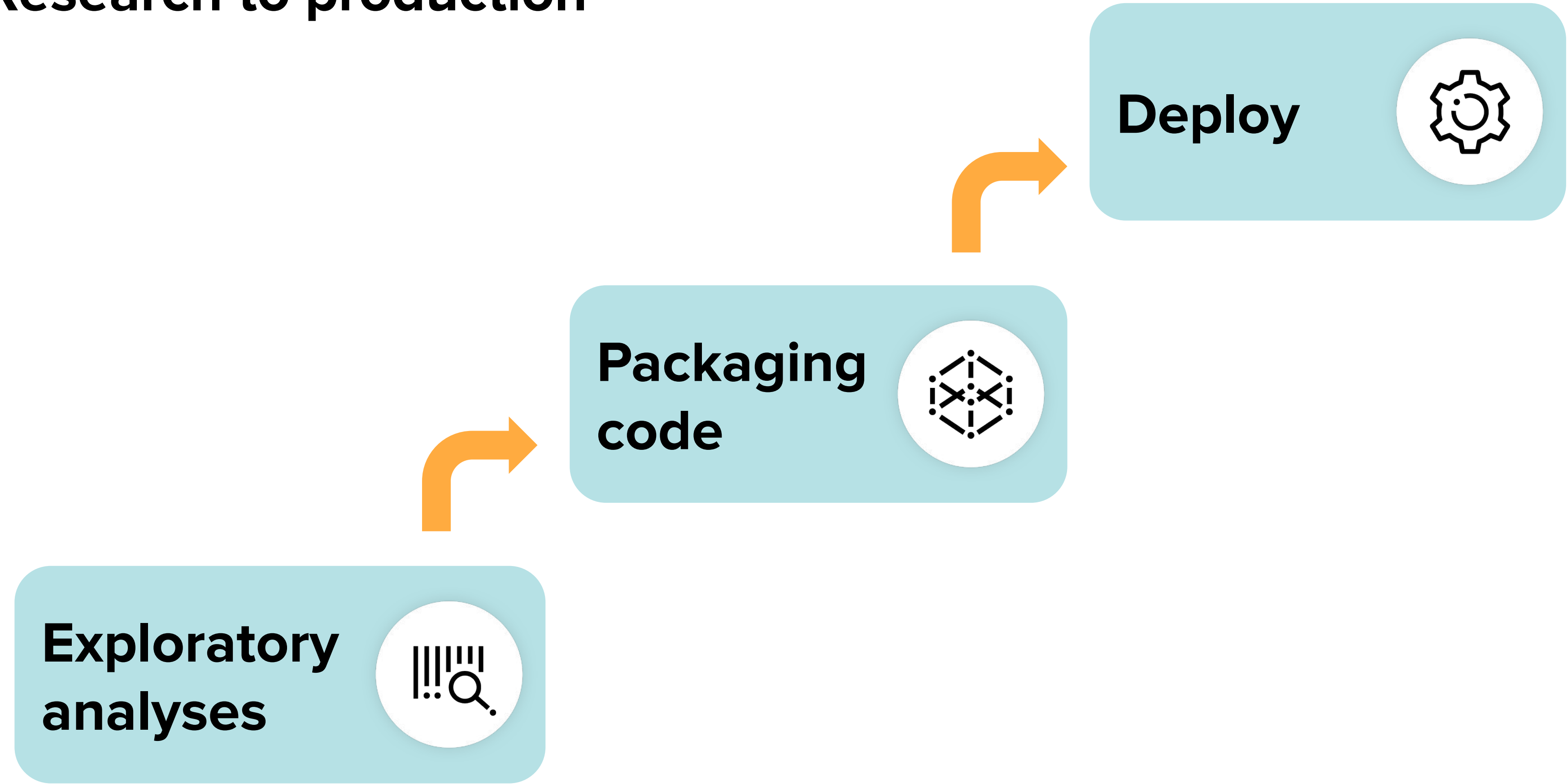
Research

Code that answers questions and delivers ideas

Development

Code that takes an input and consistently produces an output

Research to production



Exploratory analyses

Packaging code

Deploy

Exploratory analyses

Using scripts

1. Load needed packages
2. Functions & constants
3. Data ingest
4. Wrangling, plotting, stats

Using scripts

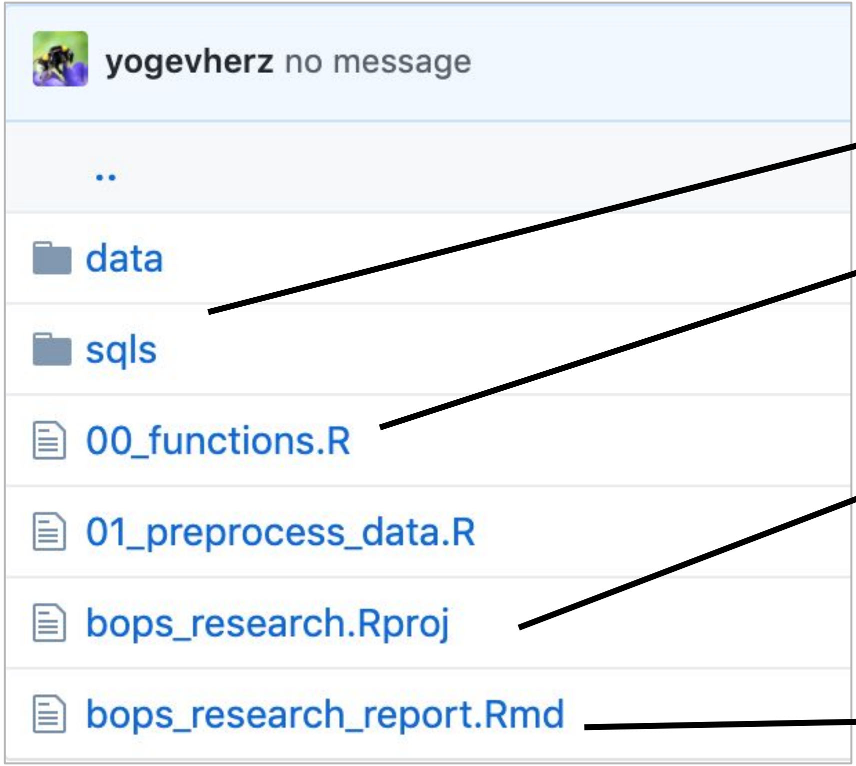
Functions file

- Runs first
- Functions & constants
- Load (and install) necessary libraries

Scripts

- Consistent names, numbered
- Sequential

Setting up a research project in R



yogevherz no message

..

data

sqls

00_functions.R

01_preprocess_data.R

bops_research.Rproj

bops_research_report.Rmd

Queries & Data

Functions & Scripts

R-Project

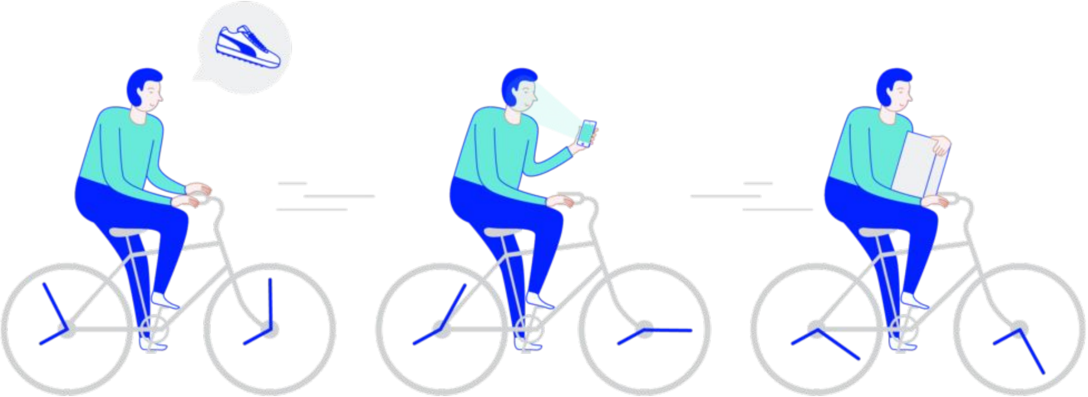
End goal:

Reproducible research report

Case study:

Fighting BOPS fraud

- **B**uy **O**nline **P**ickup in **S**tore
- Offered by many e-commerce merchants
- Appealing to customers because it is fast, frictionless and free



How does BOPS fraud work?

	BILLING NAME	SHIPPING NAME	PICKUP
Legitimate Order	John Smith	Jane Smith	Jane Smith
Fraud	John Smith	Fraudy McFraudface	Fraudy McFraudface
Recurring Fraud	John Smith	Frauddie J. McFrraudddface	Fraudy McFraudface

How much of the BOPS fraud is recurring fraud?

William Bartley

William Barrtleyy

William Barrttley

William Bartkey

William Barrtley

William Bartleyy

William Bartsley

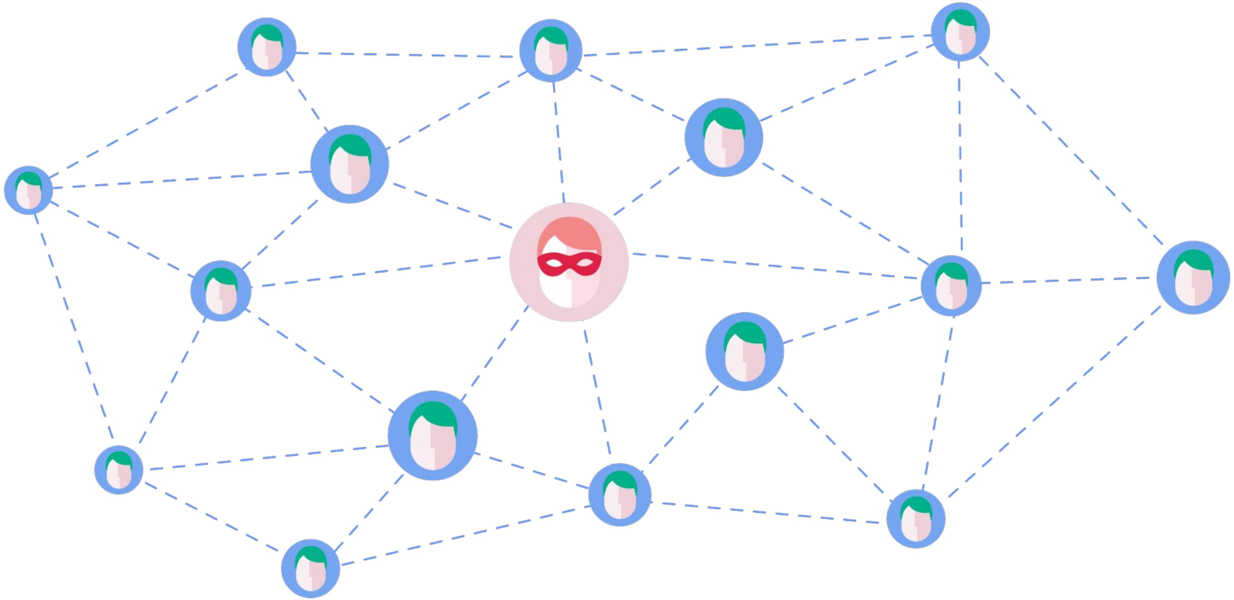
William Barttsley

William Basrtley

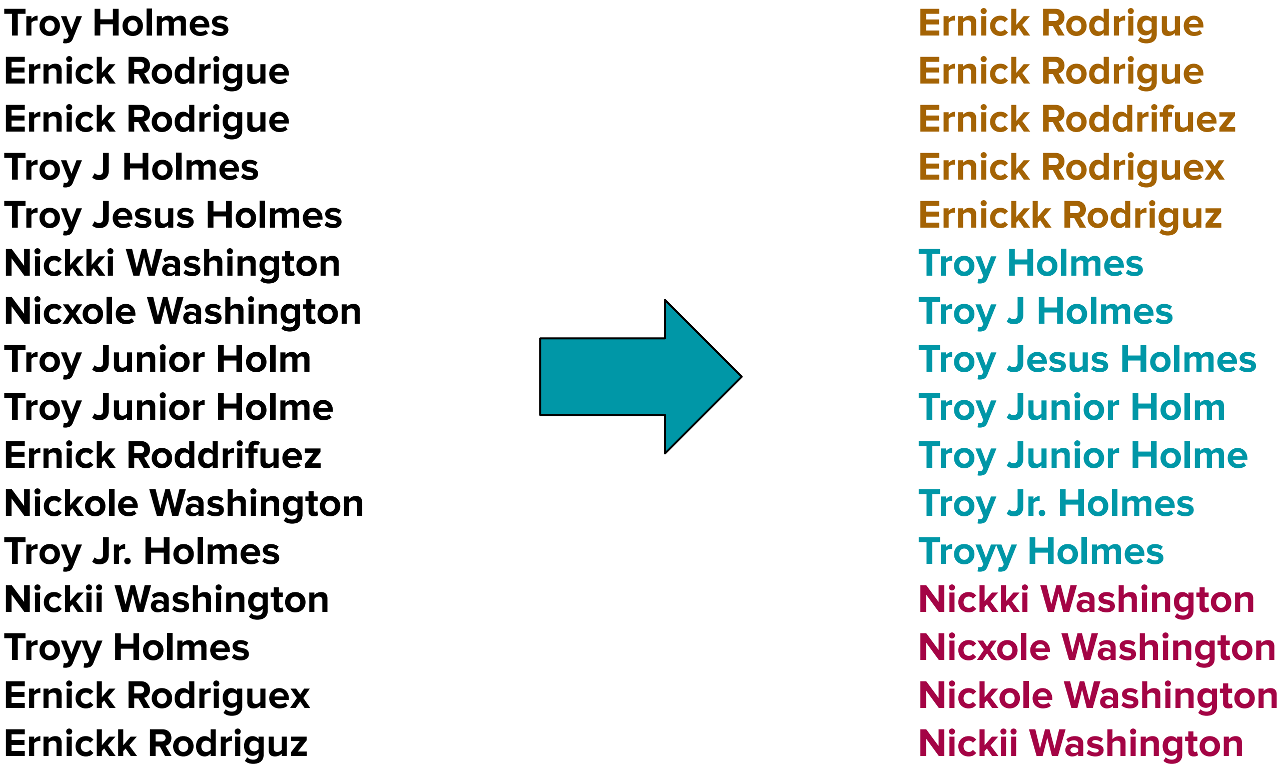
William Beartley

William Bertley

William Vartley



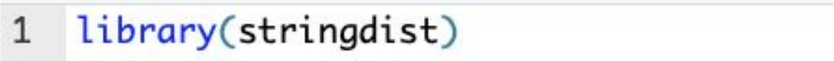
Matching names to identities



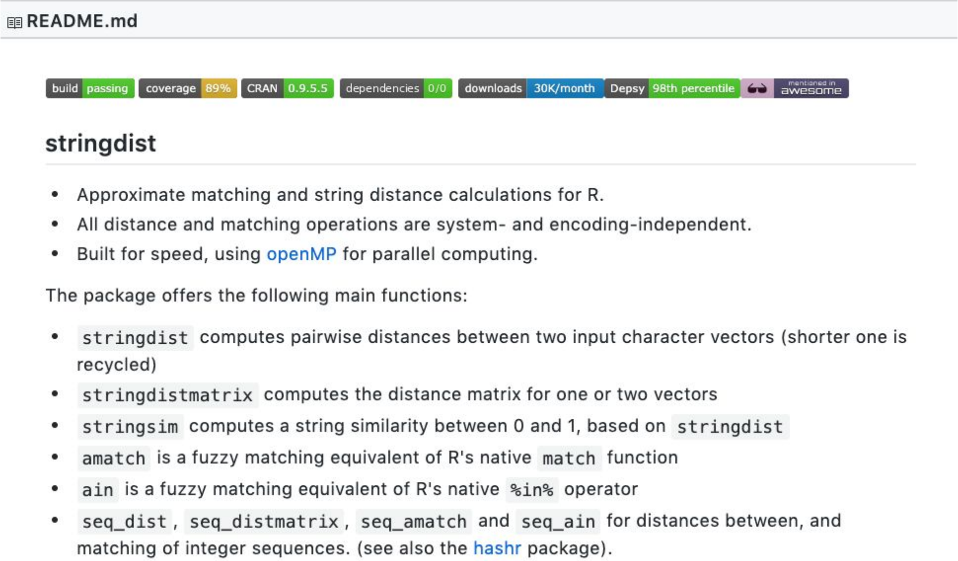
Troy Holmes
Ernick Rodrigue
Ernick Rodrigue
Troy J Holmes
Troy Jesus Holmes
Nickki Washington
Nicxole Washington
Troy Junior Holm
Troy Junior Holme
Ernick Roddrifuez
Nickole Washington
Troy Jr. Holmes
Nickii Washington
Troyy Holmes
Ernick Rodriguex
Ernickk Rodriguz

Ernick Rodrigue
Ernick Rodrigue
Ernick Roddrifuez
Ernick Rodriguex
Ernickk Rodriguz
Troy Holmes
Troy J Holmes
Troy Jesus Holmes
Troy Junior Holm
Troy Junior Holme
Troy Jr. Holmes
Troyy Holmes
Nickki Washington
Nicxole Washington
Nickole Washington
Nickii Washington

Matching names to identities



```r
library(stringdist)
```



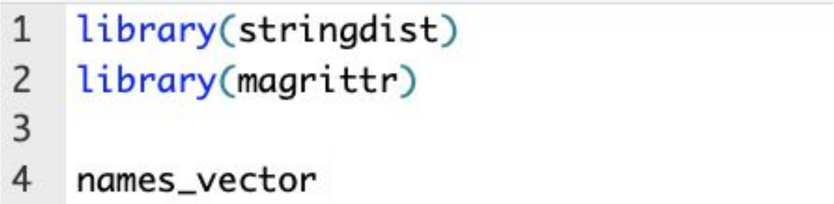
stringdist

- Approximate matching and string distance calculations for R.
- All distance and matching operations are system- and encoding-independent.
- Built for speed, using openMP for parallel computing.

The package offers the following main functions:

- `stringdist` computes pairwise distances between two input character vectors (shorter one is recycled)
- `stringdistmatrix` computes the distance matrix for one or two vectors
- `stringsim` computes a string similarity between 0 and 1, based on `stringdist`
- `amatch` is a fuzzy matching equivalent of R's native `match` function
- `ain` is a fuzzy matching equivalent of R's native `%in%` operator
- `seq_dist`, `seq_distmatrix`, `seq_amatch` and `seq_ain` for distances between, and matching of integer sequences. (see also the hashr package).

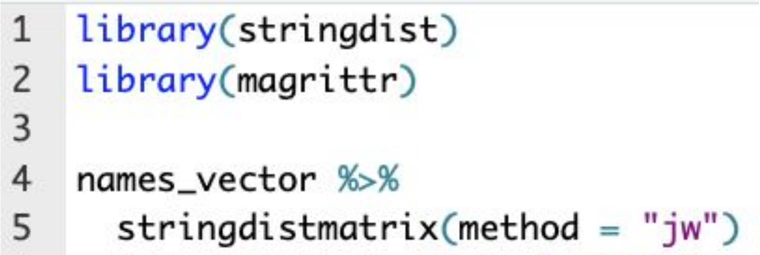
Matching names to identities



```r
library(stringdist)
library(magrittr)

names_vector
```

Troy Holmes
Ernick Rodrigue
Ernick Rodrigue
Troy J Holmes
Troy Jesus Holmes
Nickki Washington
Nicxole Washington
Troy Junior Holm
Troy Junior Holme
Ernick Roddrifuez
Nickole Washington
Troy Jr. Holmes
Nickii Washington
Troyy Holmes
Ernick Rodriguex
Ernickk Rodriguz

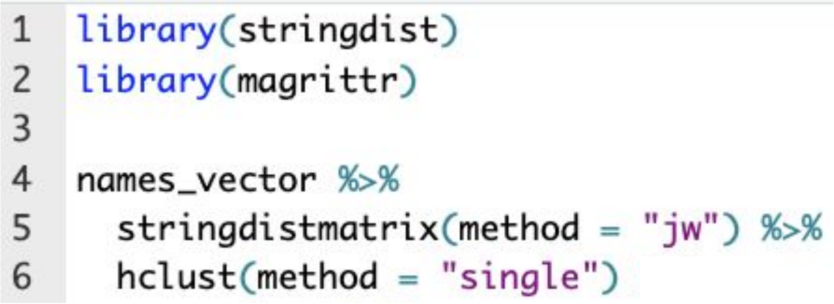
Matching names to identities



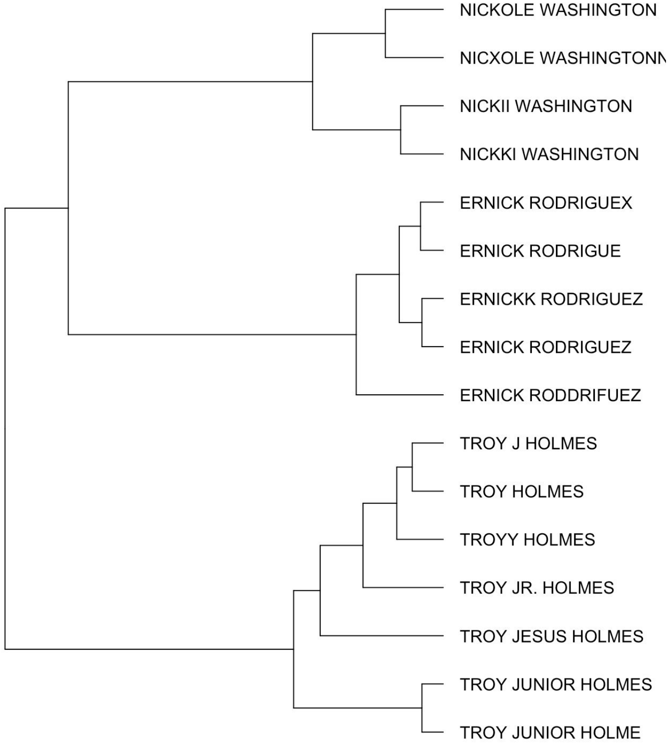
```r
library(stringdist)
library(magrittr)

names_vector %>%
  stringdistmatrix(method = "jw")
```

```
      1    2    3    4    5    6    7    8    9   10
2  0.54
3  0.54 0.00
4  0.05 0.56 0.56
5  0.21 0.50 0.50 0.17
6  0.63 0.42 0.42 0.61 0.51
7  0.51 0.47 0.47 0.50 0.52 0.20
8  0.24 0.51 0.51 0.20 0.22 0.63 0.60
9  0.20 0.46 0.46 0.17 0.20 0.64 0.60 0.02
10 0.55 0.08 0.08 0.57 0.51 0.46 0.54 0.56 0.51
11 0.51 0.43 0.43 0.50 0.52 0.16 0.04 0.60 0.60 0.51
```

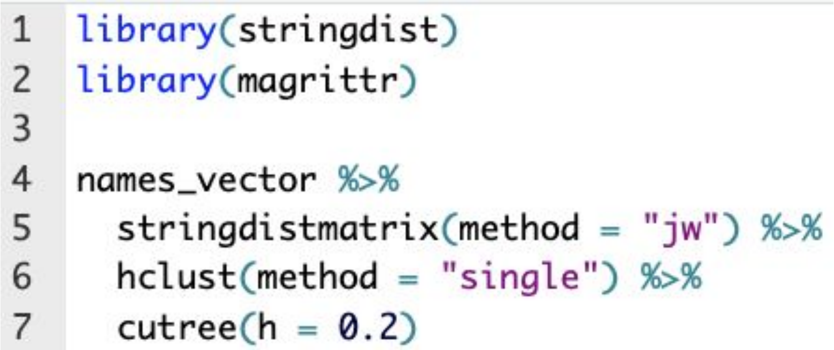
Matching names to identities



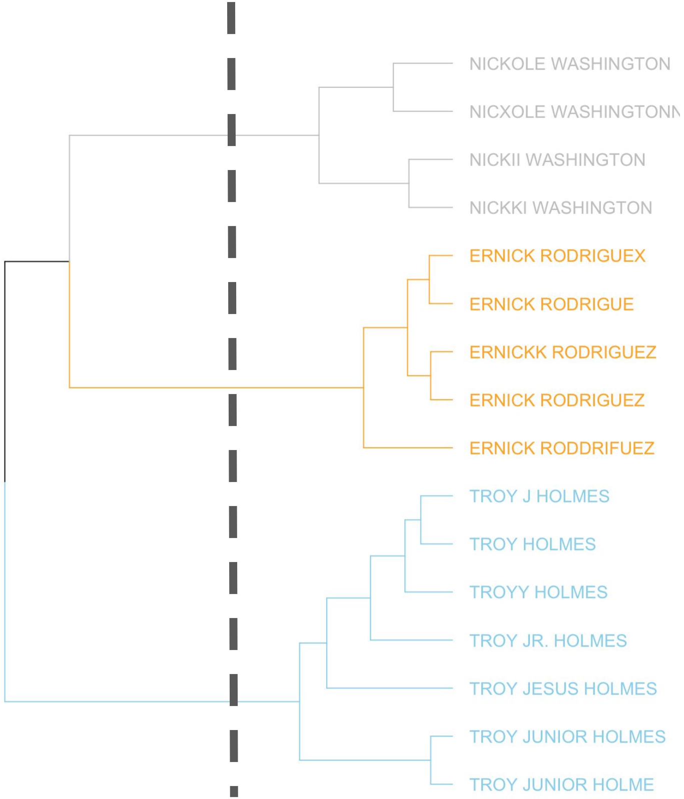
```r
library(stringdist)
library(magrittr)

names_vector %>%
  stringdistmatrix(method = "jw") %>%
  hclust(method = "single")
```



Matching names to identities

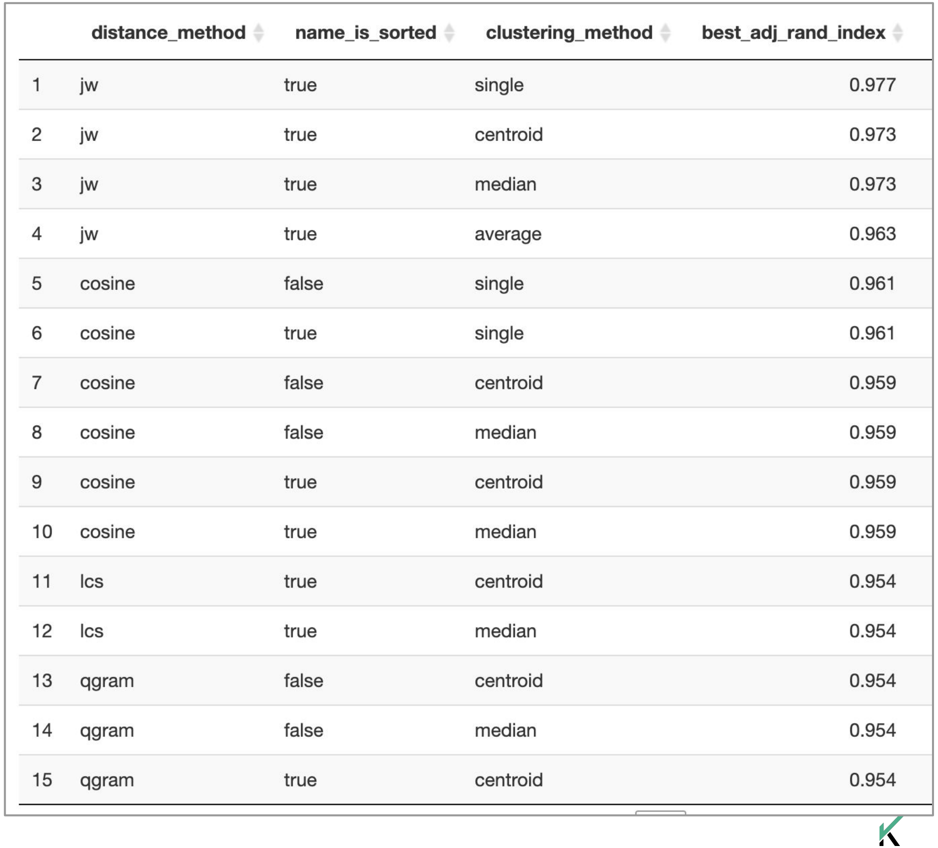


```
library(stringdist)
library(magrittr)

names_vector %>%
  stringdistmatrix(method = "jw") %>%
  hclust(method = "single") %>%
  cutree(h = 0.2)
```



BOPS research task results

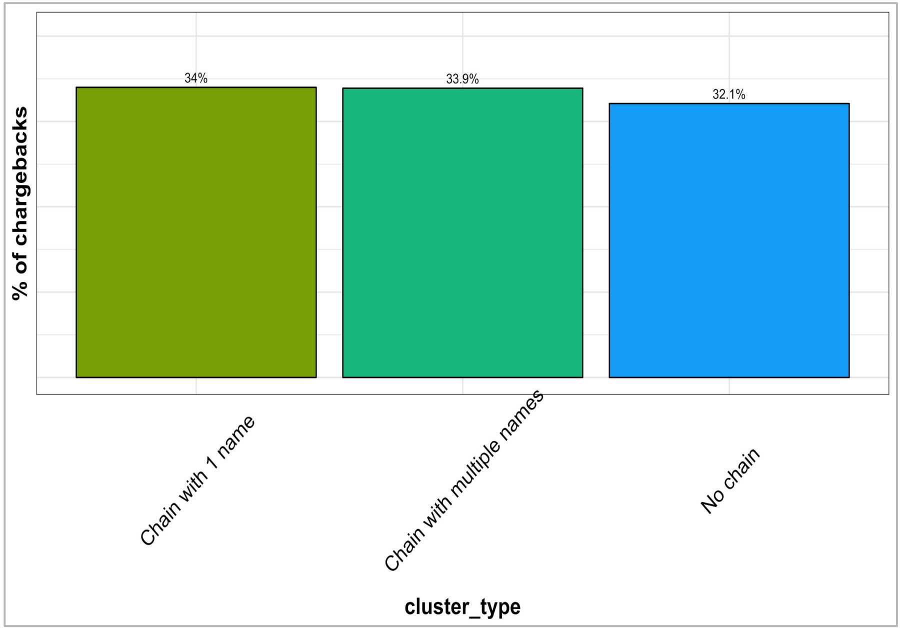
- A method for reliably clustering names into entities



	distance_method	name_is_sorted	clustering_method	best_adj_rand_index
1	jw	true	single	0.977
2	jw	true	centroid	0.973
3	jw	true	median	0.973
4	jw	true	average	0.963
5	cosine	false	single	0.961
6	cosine	true	single	0.961
7	cosine	false	centroid	0.959
8	cosine	false	median	0.959
9	cosine	true	centroid	0.959
10	cosine	true	median	0.959
11	lcs	true	centroid	0.954
12	lcs	true	median	0.954
13	qgram	false	centroid	0.954
14	qgram	false	median	0.954
15	qgram	true	centroid	0.954

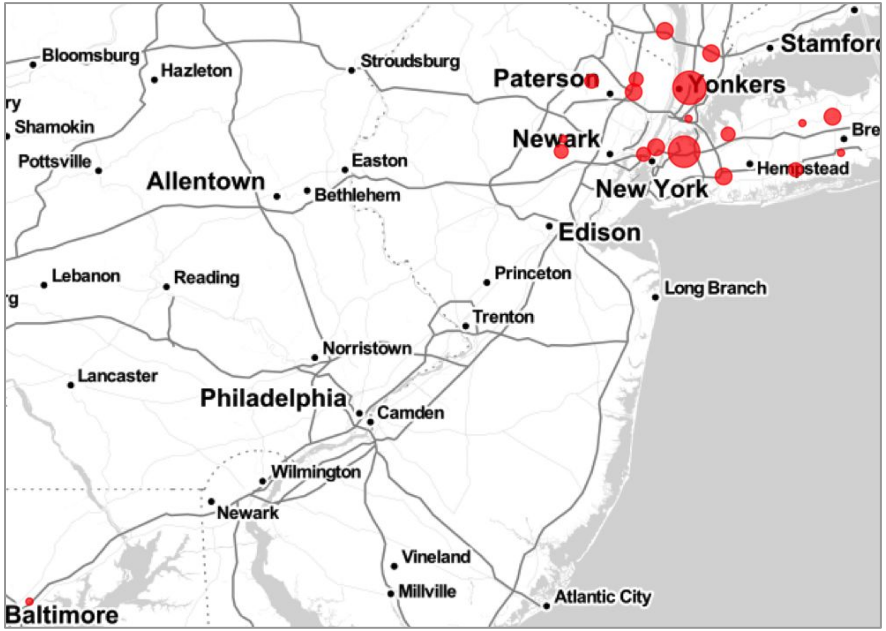
BOPS research task results

- A method for reliably clustering names into entities
- An estimate of problem severity



BOPS research task results

- A method for reliably clustering names into entities
- An estimate of problem severity
- Insights into fraud patterns



Packaging code

Using scripts
Challenges

- Documentation via comments
- Dependencies on external packages not rigorously checked
- Often shared via copy & paste
- Filepath issues
- Usually not maintained

Why a package?

- Easy to get started, especially with devtools & usethis helpers
- Accessible documentation
- Keeps functions & dependencies organized
- Testing infrastructure
- Installable!

Packaging a research project

Goal: Create functions to detect BOPS fraud

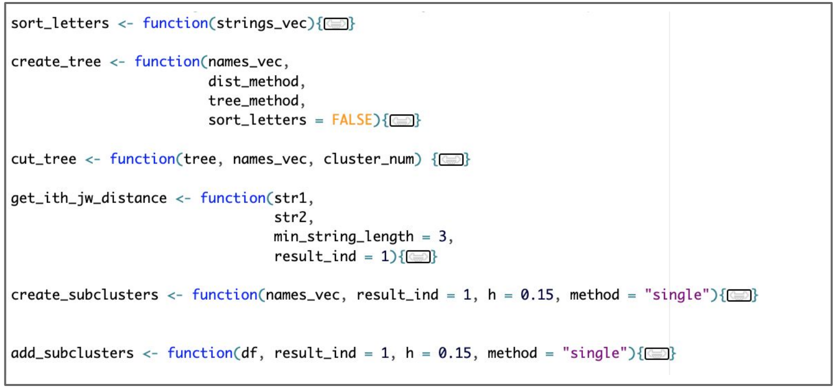
How to package?

Packaging a research project

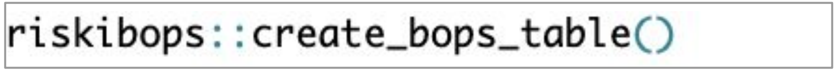
Goal: Create functions to detect BOPS fraud

How to package?

- Understand who will use the package



```
sort_letters <- function(strings_vec){⇔}

create_tree <- function(names_vec,
                        dist_method,
                        tree_method,
                        sort_letters = FALSE){⇔}

cut_tree <- function(tree, names_vec, cluster_num) {⇔}

get_ith_jw_distance <- function(str1,
                                str2,
                                min_string_length = 3,
                                result_ind = 1){⇔}

create_subclusters <- function(names_vec, result_ind = 1, h = 0.15, method = "single"){⇔}


add_subclusters <- function(df, result_ind = 1, h = 0.15, method = "single"){⇔}
```



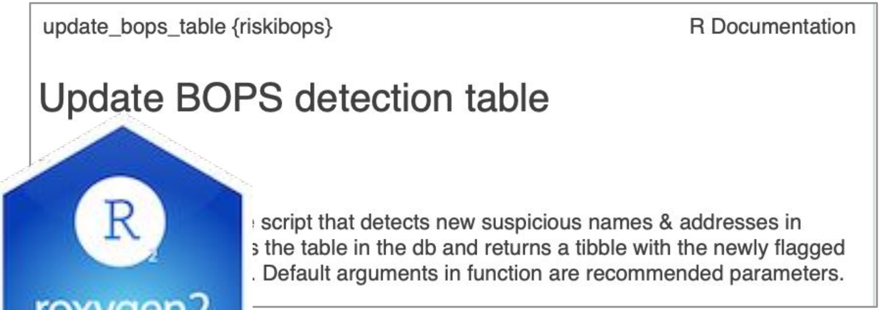
```
riskibops::create_bops_table()
```

Packaging a research project

Goal: Create functions to detect BOPS fraud

How to package?

- Understand who will use the package
- Understand that other people will use your package



update_bops_table {riskibops} R Documentation

Update BOPS detection table

script that detects new suspicious names & addresses in
the table in the db and returns a tibble with the newly flagged
. Default arguments in function are recommended parameters.

roxygen2

assertthat

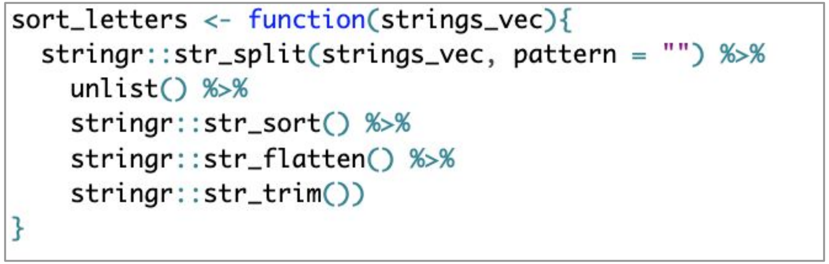
test that

Packaging a research project

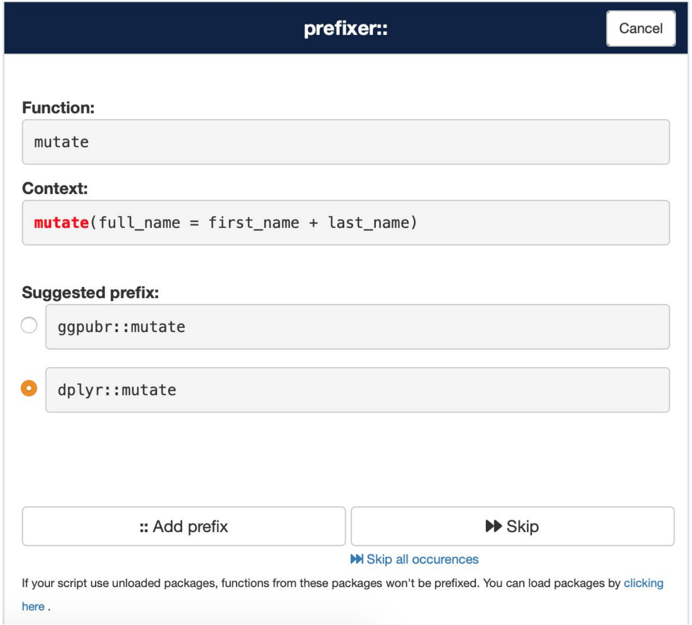
Goal: Create functions to detect BOPS fraud

How to package?

- Understand who will use the package
- Understand people will use your package
- Handle namespaces



```
sort_letters <- function(strings_vec){
  stringr::str_split(strings_vec, pattern = "") %>%
    unlist() %>%
    stringr::str_sort() %>%
    stringr::str_flatten() %>%
    stringr::str_trim())
}
```



Function:

mutate

Context:

`mutate(full_name = first_name + last_name)`

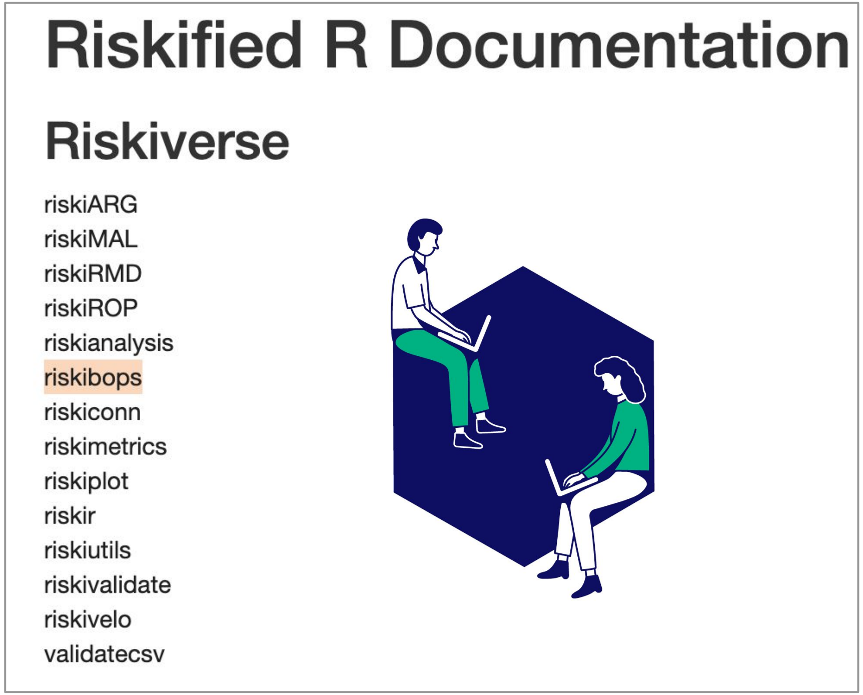
Suggested prefix:

○ ggpubr::mutate

● dplyr::mutate

If your script use unloaded packages, functions from these packages won't be prefixed. You can load packages by clicking here .

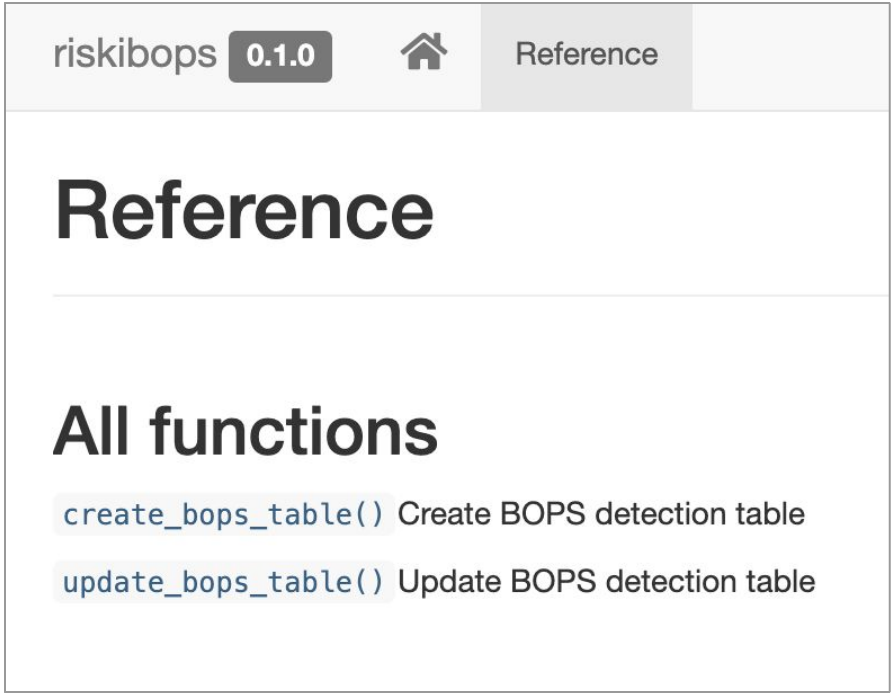
Into the riskiverse



Riskified R Documentation

Riskiverse

riskiARG
riskiMAL
riskiRMD
riskiROP
riskianalysis
riskibops
riskiconn
riskimetrics
riskiplot
riskir
riskiutils
riskivalidate
riskivelo
validatecsv



Reference

All functions

`create_bops_table()` Create BOPS detection table

`update_bops_table()` Update BOPS detection table

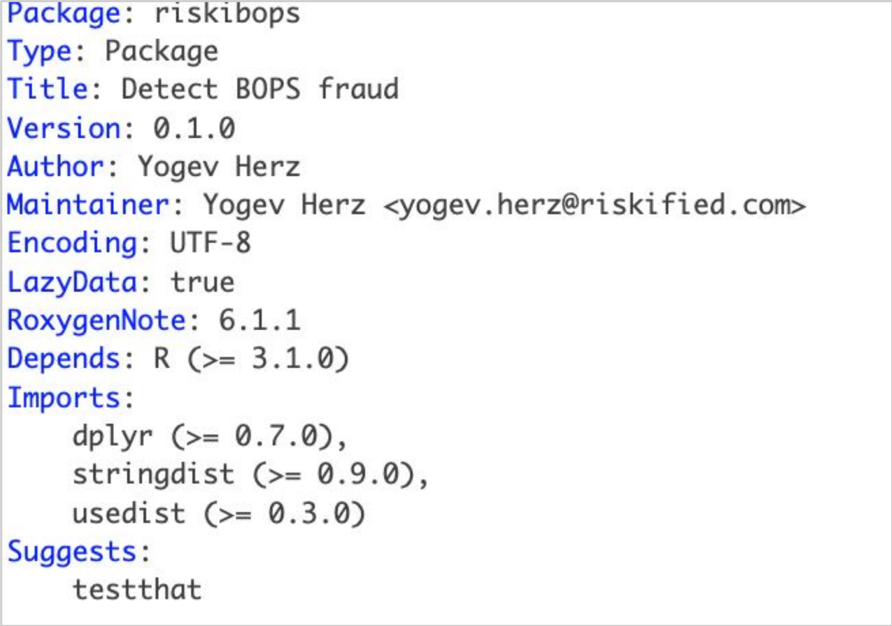
Deploy

Deploying the package

- Start simple: run locally and manually to test effects ☕

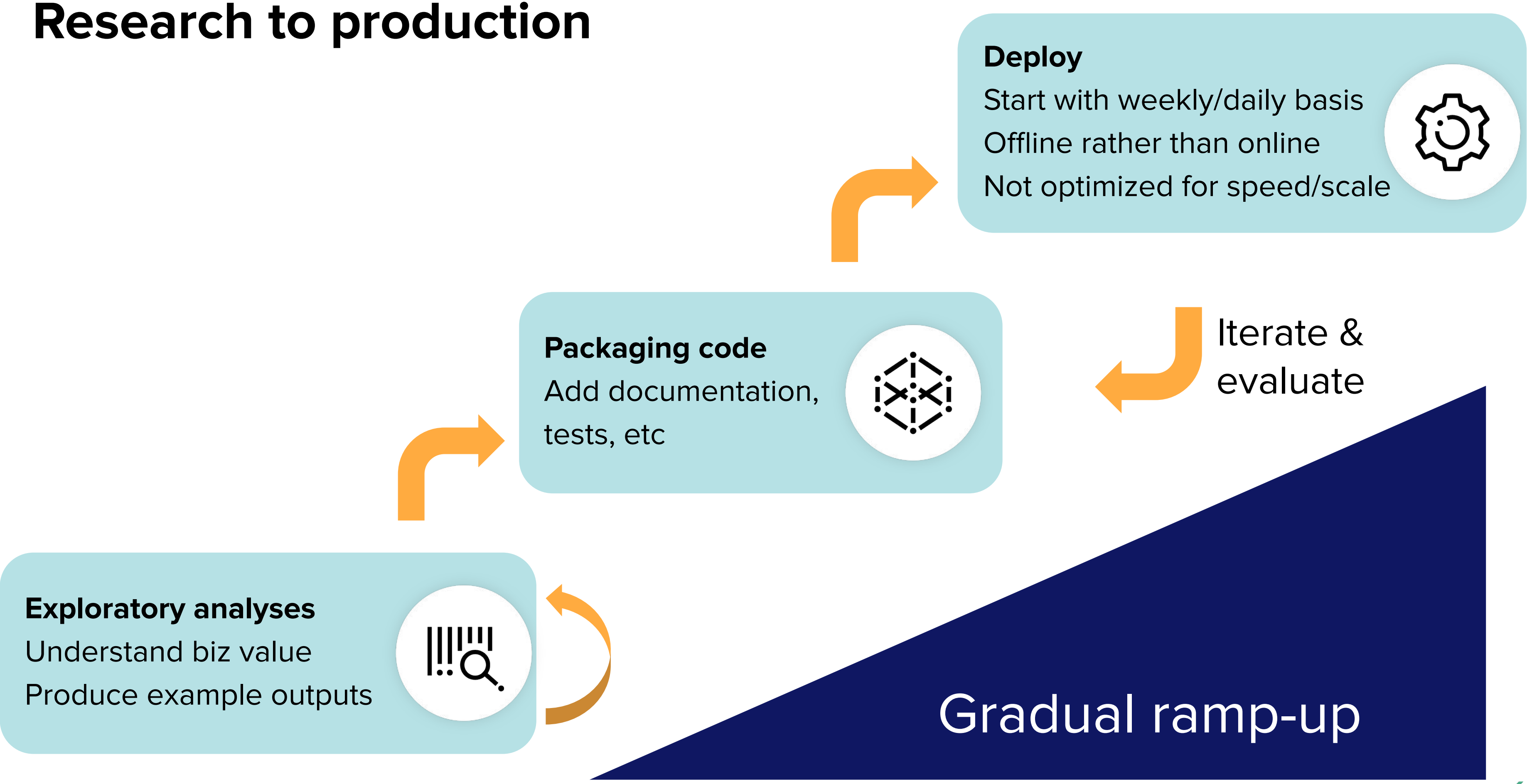
Deploying the package

- Start simple: run locally and manually to test effects ☕
- When we feel confident: send it to a remote machine to run automatically



```
Package: riskibops
Type: Package
Title: Detect BOPS fraud
Version: 0.1.0
Author: Yogev Herz
Maintainer: Yogev Herz <yogev.herz@riskified.com>
Encoding: UTF-8
LazyData: true
RoxygenNote: 6.1.1
Depends: R (>= 3.1.0)
Imports:
    dplyr (>= 0.7.0),
    stringdist (>= 0.9.0),
    usedist (>= 0.3.0)
Suggests:
    testthat
```

Research to production



Exploratory analyses

Understand biz value

Produce example outputs

Packaging code

Add documentation, tests, etc

Deploy

Start with weekly/daily basis

Offline rather than online

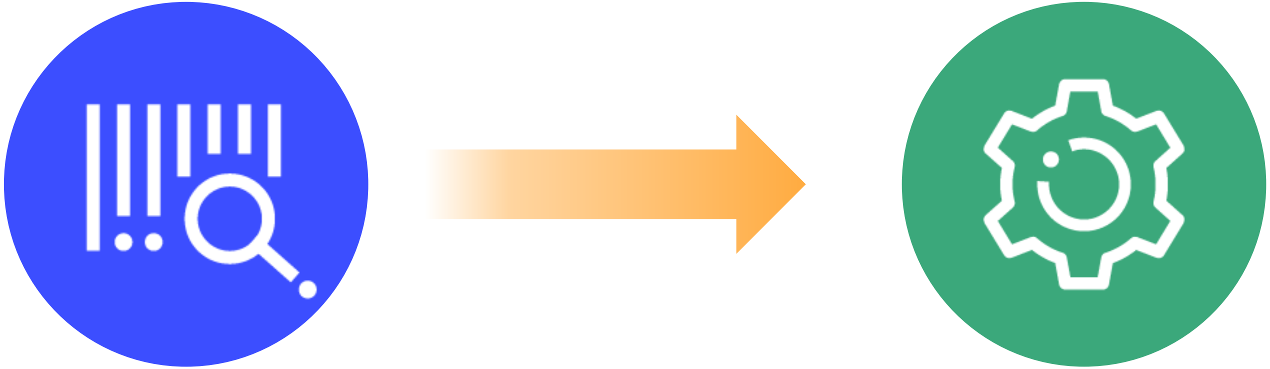
Not optimized for speed/scale

Iterate & evaluate

Gradual ramp-up

R for prototyping

Analysis → build mode involves shifting mindsets -- not necessarily new tools!



Research

Prioritizes new insights, flexibility

Development

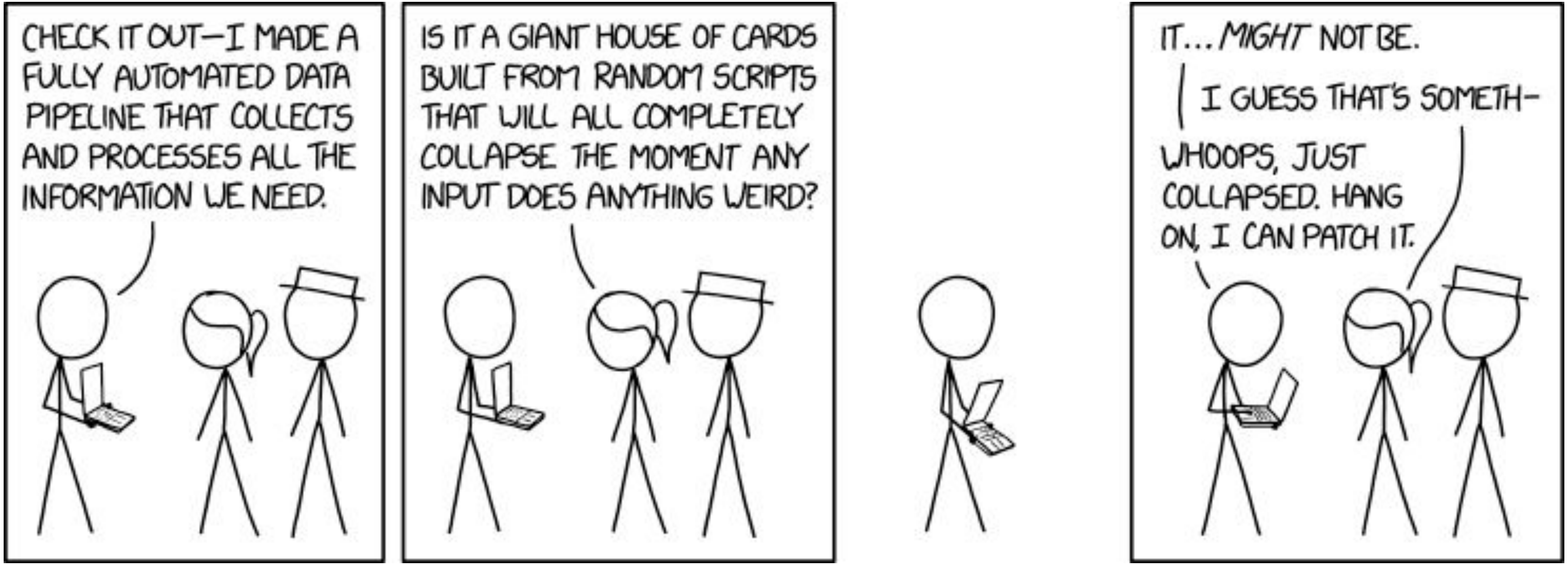
Prioritizes re-use, stability, scalability, speed

Thank you for your time!

Irene Steves @i_steves

Yogev Herz @yogevmh

Check out our tech blog! https://medium.com/riskified-technology


CHECK IT OUT—I MADE A FULLY AUTOMATED DATA PIPELINE THAT COLLECTS AND PROCESSES ALL THE INFORMATION WE NEED.
IS IT A GIANT HOUSE OF CARDS BUILT FROM RANDOM SCRIPTS THAT WILL ALL COMPLETELY COLLAPSE THE MOMENT ANY INPUT DOES ANYTHING WEIRD?
IT... *MIGHT* NOT BE.
I GUESS THAT'S SOMETH-
WHOOPS, JUST COLLAPSED. HANG ON, I CAN PATCH IT.

https://xkcd.com/2054/