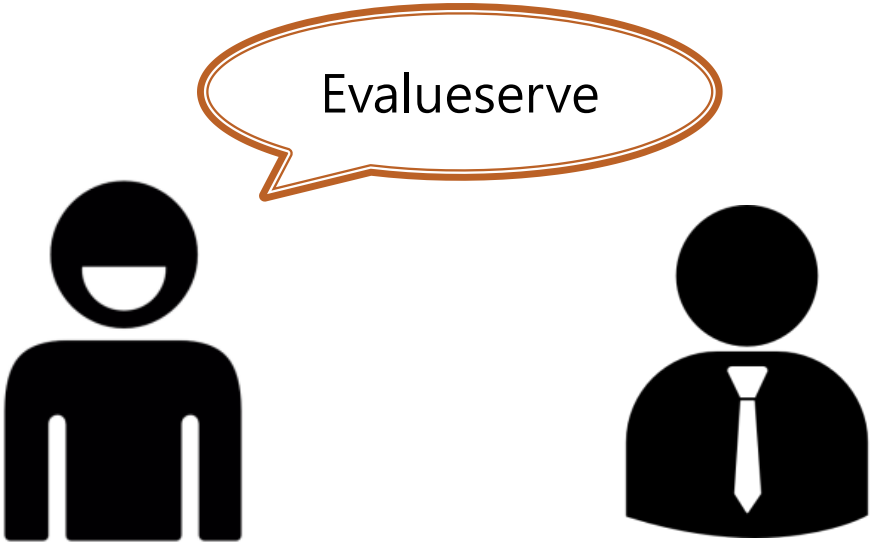
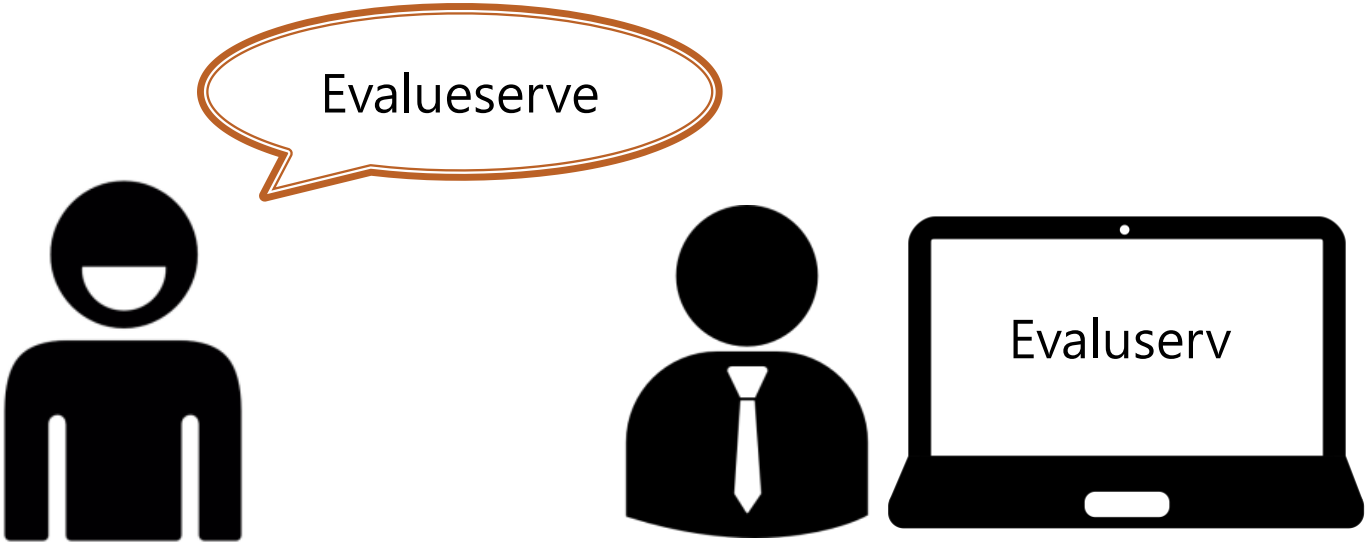
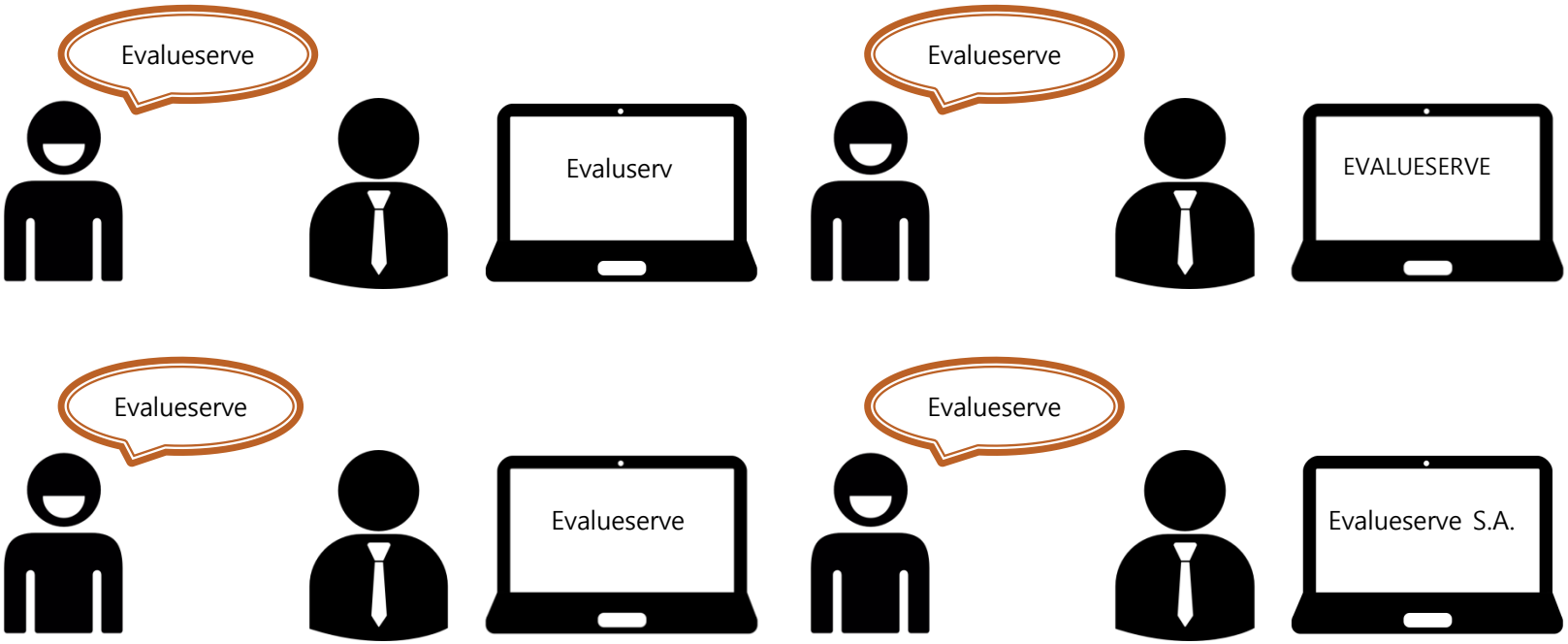
Fuzzy Merging
with Company Names

Richard Vogg

Sep 26th, 2019


Evalueserve


Evalueserve
Evaluserv


Evalueserve
Evaluserv
Evalueserve
EVALUESERVE
Evalueserve
Evalueserve
Evalueserve
Evalueserve S.A.


Evalueserve
Evaluserv


Evalueserve
EVALUESERVE


Evalueserve
Evaluesurf


Evalueserve
evalueserve


Evalueserve
Evaluserv


Evalueserve
EVLAUESERV


Evalueserve
EVALUSERVE

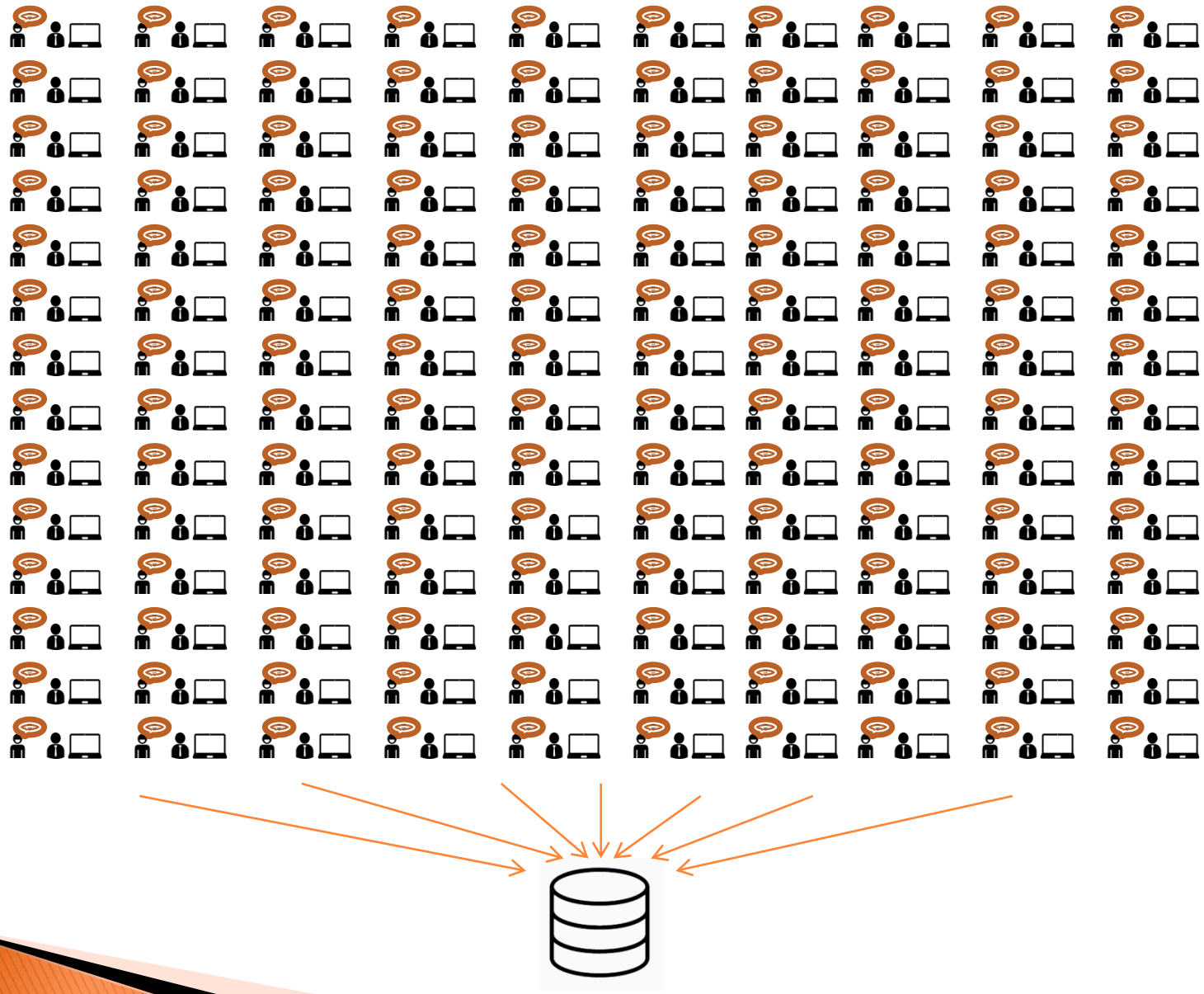

Evalueserve
Evelueserv


Evalueserve
Evalueserve SA


Evalueserve
Evalueserve


Evalueserve
E-Value Serve


Evalueserve
Evalueserve



The input

The input

dirty_list
Haliburton
ExxonMobile
ABBOTT LABORATORIES
Marrriott
Self
Activision Blizzard
Quest dianotstics
Unemployed
MARRIOT

500,000 names

The input

dirty_list
Haliburton
ExxonMobile
ABBOTT LABORATORIES
Marrriott
Self
Activision Blizzard
Quest dianotstics
Unemployed
MARRIOT

500,000 names

clean_list
3M Company
Abbott Laboratories
AbbVie Inc.
ABIOMED Inc
Accenture plc
Activision Blizzard
Adobe Systems Inc
Advanced Micro Devices Inc
Advance Auto Parts
AES Corp

S&P500 list

The stringdist package

- Approximate matching and string distance calculations for R.
- Many distance functions: Hamming, Levenshtein, Longest common substring, qgram, **Jaro-Winkler**
- Computes in parallel when possible

van der Loo M (2014). "The stringdist package for approximate string matching." *The R Journal*, **6**, 111-122. https://CRAN.R-project.org/package=stringdist.

Jaro similarity

1	2	3	4	5	6	7	8	9
B	R	E	A	K	F	A	S	T

B	A	S	E	B	A	L	L
1	2	3	4	5	6	7	8

$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else. \end{cases}$$

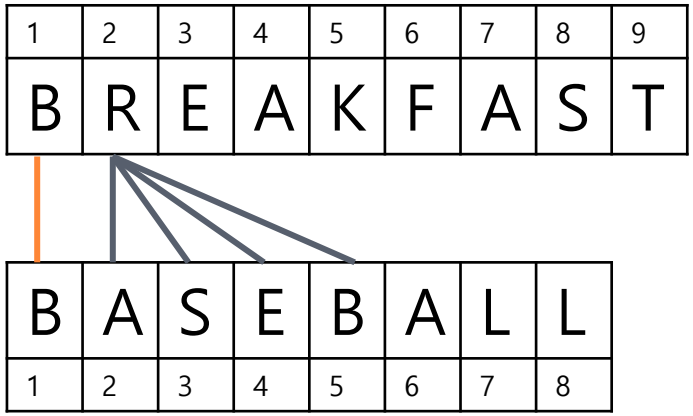
Jaro similarity

1	2	3	4	5	6	7	8	9
B	R	E	A	K	F	A	S	T

B	A	S	E	B	A	L	L
1	2	3	4	5	6	7	8

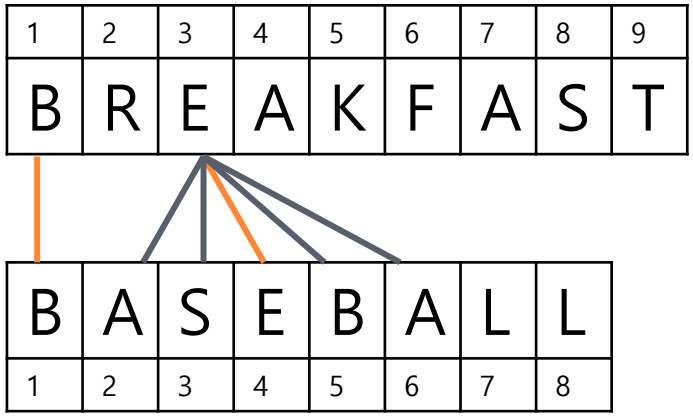
$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else. \end{cases}$$

Jaro similarity



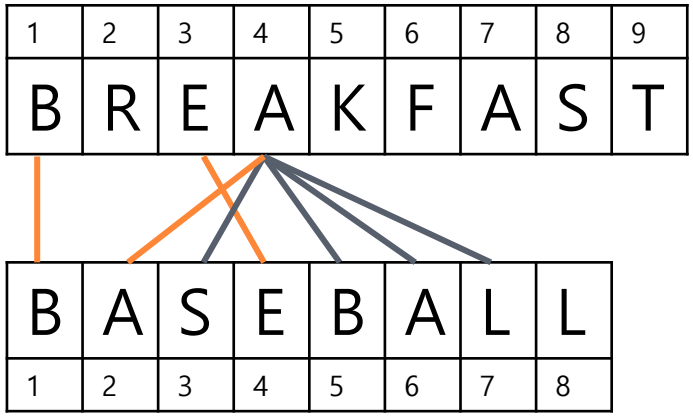
$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else. \end{cases}$$

Jaro similarity



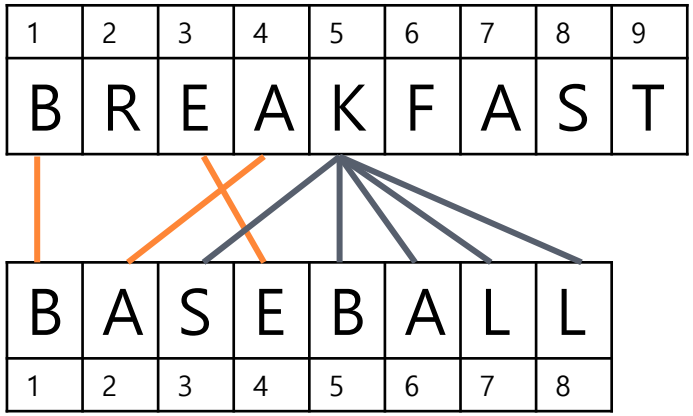
$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else. \end{cases}$$

Jaro similarity



$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else. \end{cases}$$

Jaro similarity



$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & else. \end{cases}$$

Jaro similarity

1	2	3	4	5	6	7	8	9
B	R	E	A	K	F	A	S	T

B	A	S	E	B	A	L	L
1	2	3	4	5	6	7	8

$$sim_j(\text{"}breakfast\text{"},\text{"}baseball\text{"}) = \frac{1}{3}\left(\frac{4}{9}+\frac{4}{8}+\frac{4-1}{4}\right)$$

$$= 0.56$$

$$sim_j(s_1, s_2) = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|}+\frac{m}{|s_2|}+\frac{m-t}{m}\right), & else. \end{cases}$$

Jaro-Winkler similarity

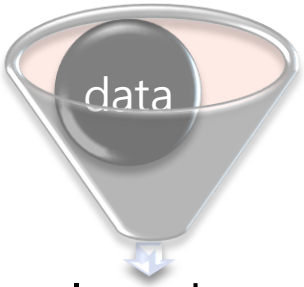
1	2	3	4	5	6	7	8	9
B	R	E	A	K	F	A	S	T

B	A	S	E	B	A	L	L
1	2	3	4	5	6	7	8

$$sim_{jw}(\text{"}breakfast\text{"},\text{"}baseball\text{"})$$
$$= 0.56 + 0.1(1 - 0.56) = 0.60$$

$$sim_{jw}(s_1, s_2) = sim_j + lp(1 - sim_j)$$

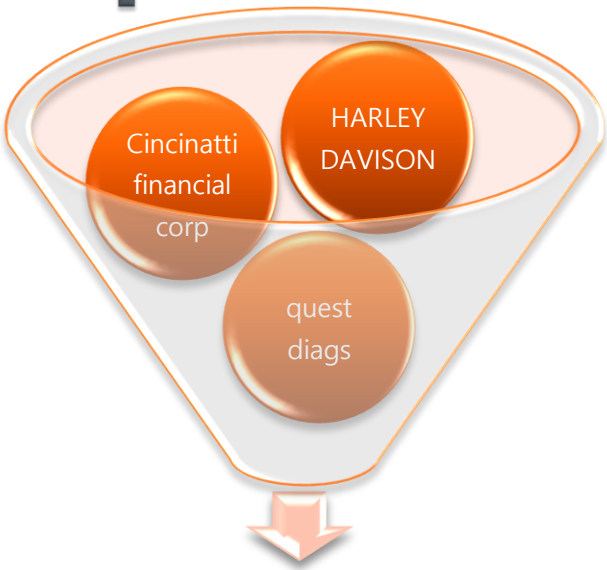
The process: cleaning



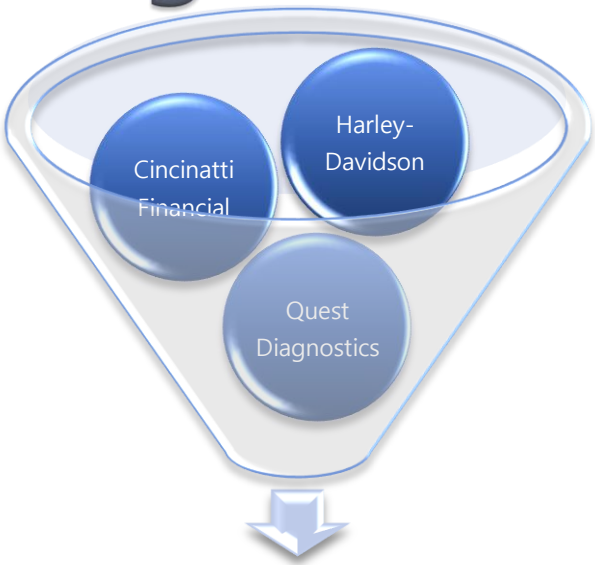
cleandata

```
cleaner <- function(data) {
  wordremove <- c(" and "," comp "," company","companies"," corp
  "," inc ","[.]com")
  data <- data %>% tolower() %>%
    {gsub(paste(wordremove,collapse='|'),"",.)} %>%
    {gsub("[[:punct:]]","",.)} %>%
    {gsub("[[:blank:]]","",.)}
  return(data)
}
```

The process: cleaning



cincinattifinancial
harleydavison
questdiags



cincinattifinancial
harleydavidson
questdiagnostics

```
clean_list_cl <- cleaner(clean_list)
dirty_list_cl <- cleaner(dirty_list)
```

The process: distance matrix

	cincinattifinancial	harleydavison	otherexample	questdiags
harleydavidson	0.67	0.01	0.54	0.51
cincinattifinancial	0	0.66	0.64	0.68
questdiagnostics	0.47	0.46	0.63	0.08

```
library(stringdist)
distmatrix <- stringdist::stringdistmatrix(clean_list_cl,
                        dirty_list_cl,method='jw‘,p=0.1)
```

The process: minima detection

	cincinattifinancial	harleydavison	otherexample	questdiags
harleydavidson	0.67	0.01	0.54	0.51
cincinattifinancial	0	0.66	0.64	0.68
questdiagnostics	0.47	0.46	0.63	0.08

```
output <- data.frame(original=dirty_list)
best_fit <- apply(distmatrix,2,which.min) %>% as.integer()
output$best_fit <- clean_list[best_fit]
output$distance <- apply(distmatrix,2,min)
```

The process: thresholding

	cincinattifinancial	harleydavison	otherexample	questdiags
harleydavidson	0.67	0.01	0.54	0.51
cincinattifinancial	0	0.66	0.64	0.68
questdiagnostics	0.47	0.46	0.63	0.08

```
output$final <- ifelse(control$distance<0.12,control$best_fit,NA)
```

The output

original	best_fit	distance	result
Self-employed	Teleflex	0.306	NA
Mohawk Ind	Mohawk Industries	0.088	Mohawk Industries
cincinnati financial	Cincinnati Financial	0.000	Cincinnati Financial
Illumin	Illumina Inc	0.073	Illumina Inc
HARLEY DAVIDSEN	Harley-Davidson	0.029	Harley-Davidson
Oracle	Oracle Corp.	0.080	Oracle Corp.
Harley Davidson	Harley-Davidson	0.000	Harley-Davidson
burger king	NRG Energy	0.300	NA
Haliburton	Halliburton Co.	0.054	Halliburton Co.
Self	Sealed Air	0.244	NA

The value of Fuzzy merging

Descriptive	Predictive	Combinations
• S&P 500 customer analysis • Regional opportunities	• Add value to predictive models • Churn analysis	• Identify customers in managing positions in S&P 500 companies

Other applications

Our approach:
Dirty list vs
clean list

Fuzzy search

Record linkage

De-duplication

Thank you!

richard.vogg@web.de

Richard Vogg

@richard_vogg