NON-NEGATIVE MATRIX FACTORIZATION

AS A SEGMENTATION TOOL IN HIGH DIMENSIONAL FEATURE SPACES

GRADIENT

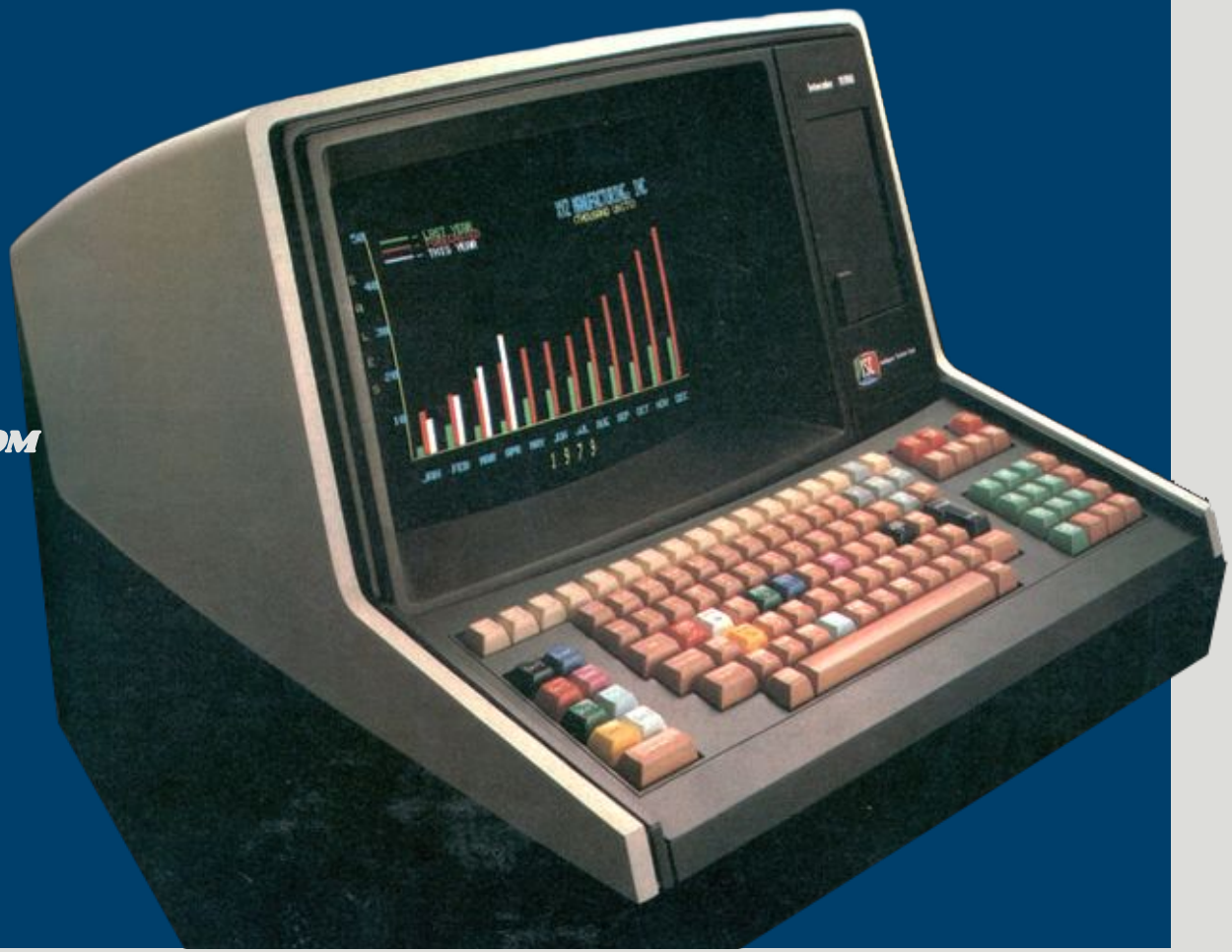
DEVELOPING INTELLIGENCE POWERED BY DATA

WHO IS THIS GUY?

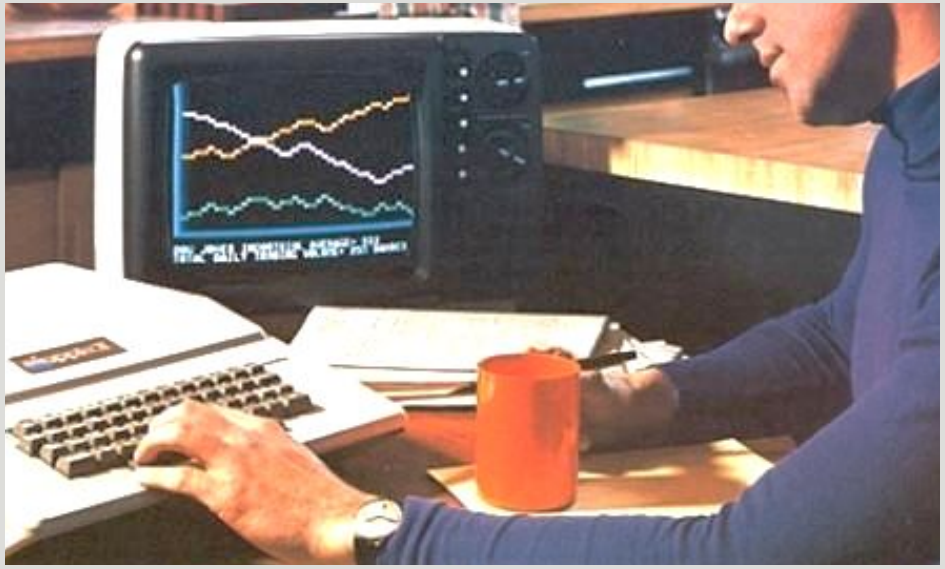
MARCIN KOSIŃSKI

- WARSAW RUG
- R BLOGGER R-ADDICT.COM
- WHYR.PL/2019/

MARCIN@GRADIENTMETRICS.COM



Nice to meet you!

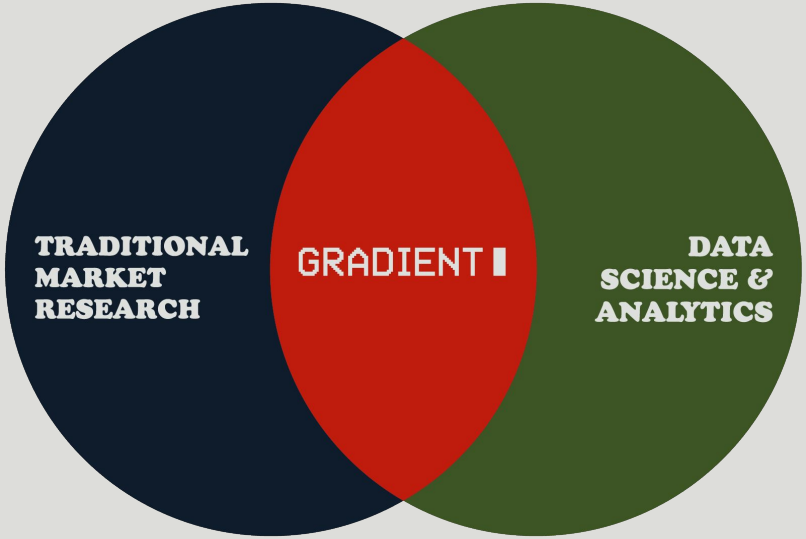


GRADIENTMETRICS.COM

WE'RE GRADIENT:

A crew of quantitative marketers and technologists that gather hard data and build robust statistical models to guide organizations through their most difficult decisions.

We're confirmed data geeks, but word on the street is that we're easy to work with and pretty fun, too.



GRADIENT

SEGMENTATION

DEFINITION & EXAMPLES

LET'S START TALKING

Market **segmentation** is the activity of dividing a broad consumer or business market, into sub-groups of consumers based on shared characteristics.

Examples

1. Survey voters/opinion segmentation.
2. Online retailer customer segmentation.
3. Brand loyalty segmentation.
4. Online news portal users segmentation.
5. Bank clients segmentation.

GRADIENT

SEGMENTATION

QUESTIONS IT (MIGHT) ANSWER

LET'S START ASKING

In how many groups can the market be divided and what are their sizes (in %)?

What are specific, unique features describing each of the segments?

How can we assign a future customer or a respondent to the existing segmentation?

How can segmentation be used to improve targeting and marketing methods?

SEGMENTATION

CHALLENGES IT FACES

DEPENDING ON THE SCENARIO

Market Data

1. Geographic data.
2. Demographic data.
3. Behavioral data.
4. Survey data.

Challenges

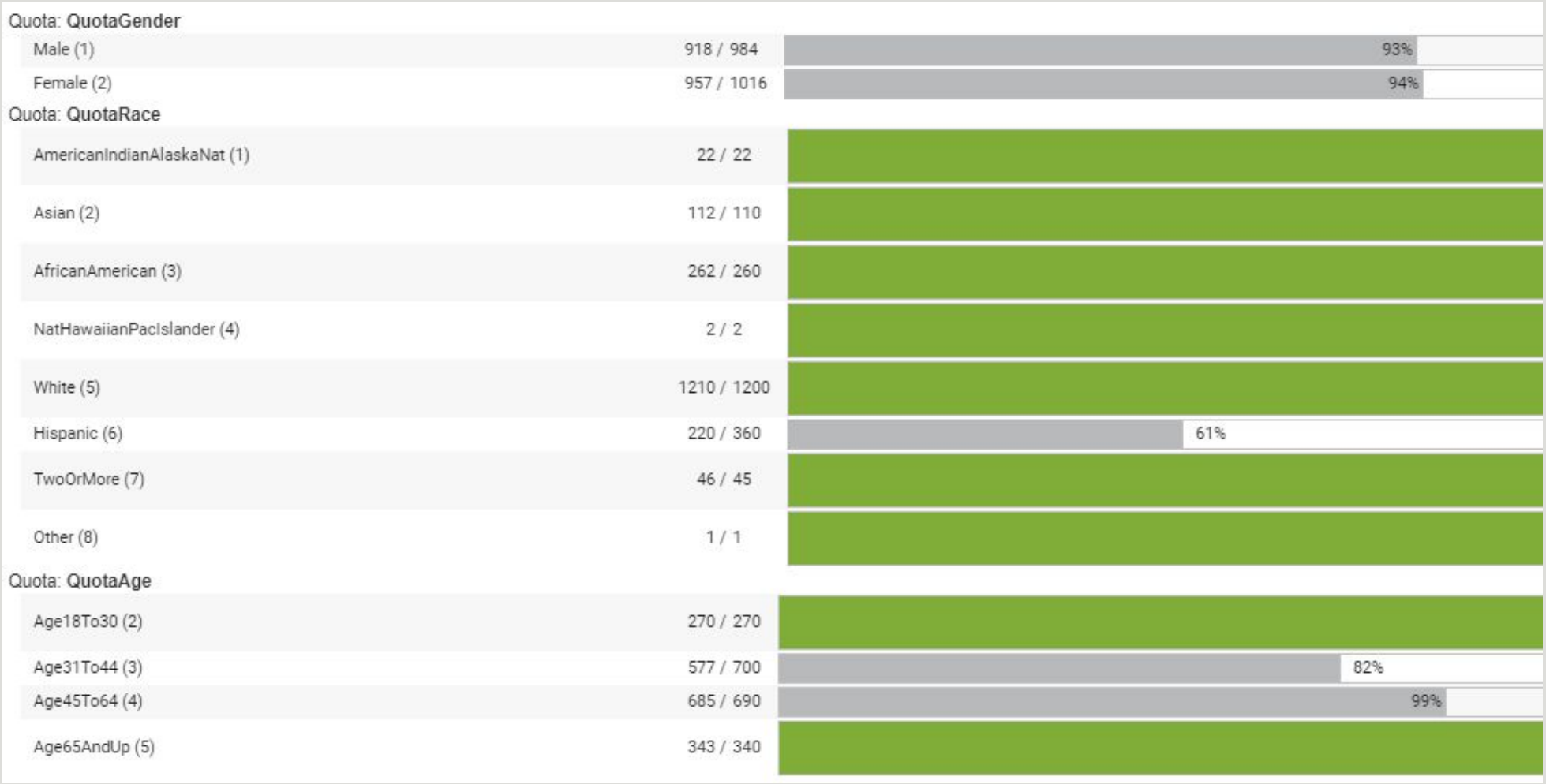
1. Unsupervised learning.
2. Mix of numeric, categorical, multi-select and ordinal data.
3. **Extremely huge dimension of feature space.**
4. **Meaningful - Size, Description, Story.**

DATA STRUCTURE

HEAD OF THE EXAMPLE SURVEY DATA

ID	Gender	Age Bucket	Race	Statement 1	...	Statement N
1	Male	[22, 25)	White	Strongly Agree		Strongly Disagree
2	Female	[40, 50)	Latino	Agree		Agree
3	Other	[18, 22)	Asian	Neither		Neither
4	Female	[25, 30)	Asian	Strongly Disagree		Strongly Agree
5	Male	[30, 40)	Black	Strongly Disagree		Agree
6	Male	[25, 30)	White	Disagree		Neither

DATA STRUCTURE



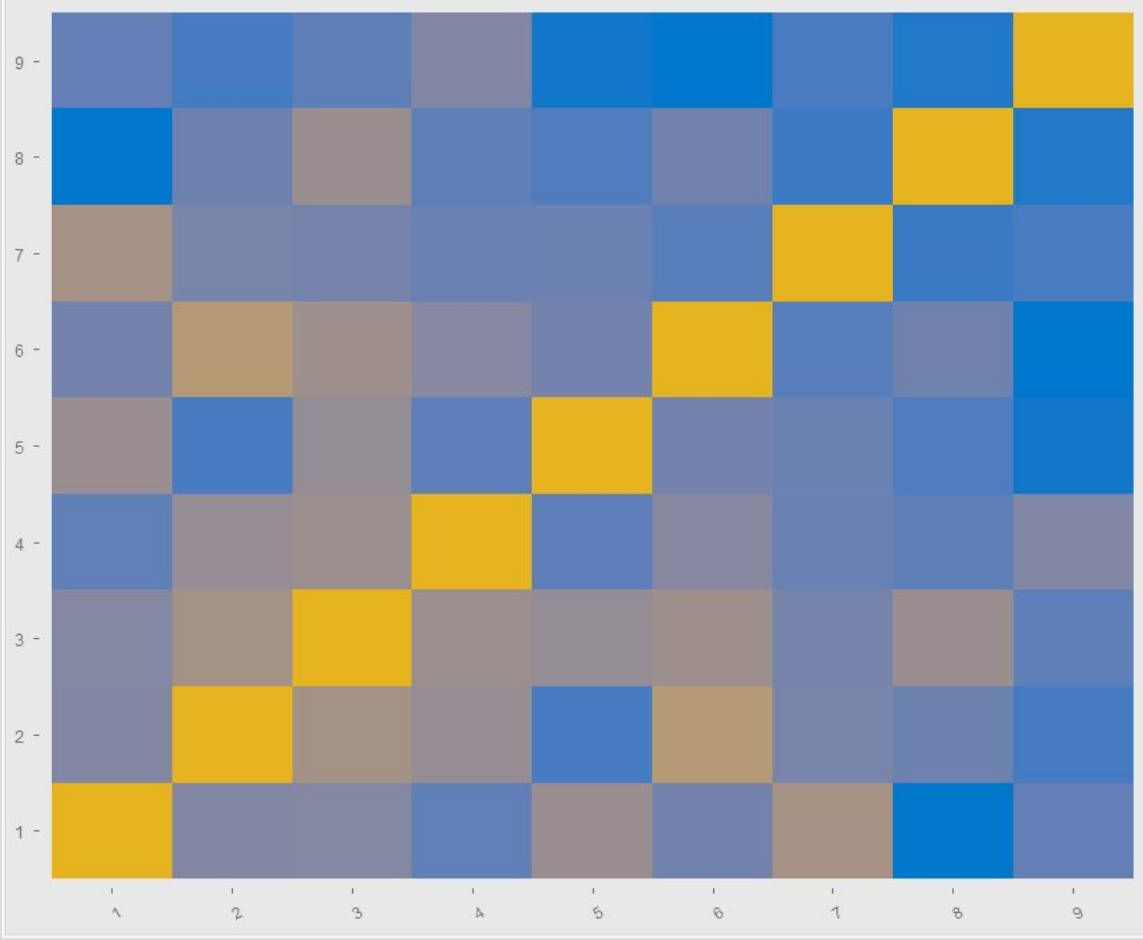
TOOLS

SIMILARITY MEASURES

DISTANCE MATRIX

ID	1	2	3	4	5
1	0.00	0.28	1.14	0.01	0.82
2	0.28	0.00	3.45	0.98	2.12
3	1.14	3.45	0.00	5.16	5.16
4	0.01	0.98	5.16	0.00	0.21
5	0.82	2.12	5.16	0.21	0.00
6	0.72	...	...	...	...

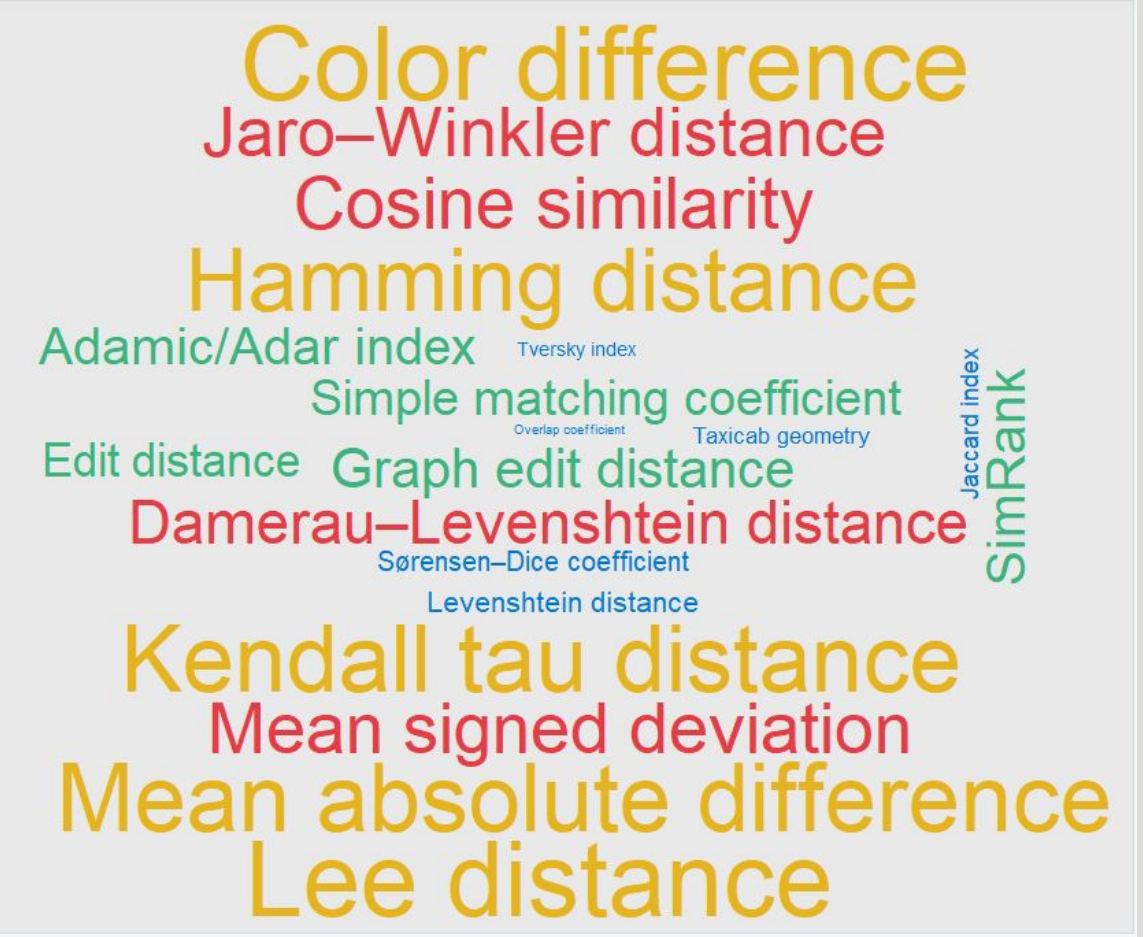
Similarity matrices can be built on various type of distance measures.



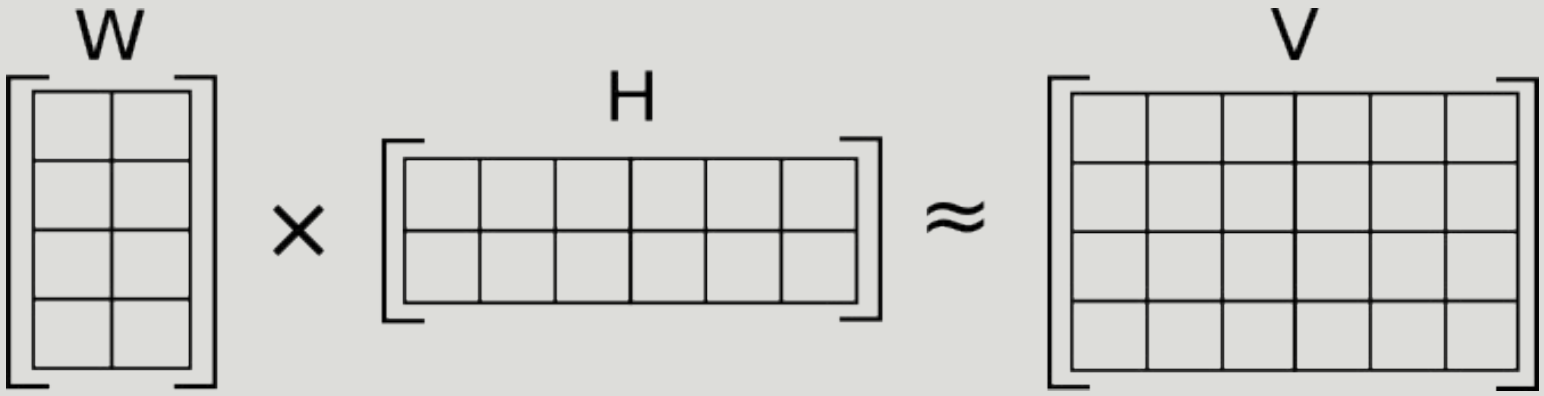
THE CHOICE OF A DISTANCE MEASURE

- Some metrics work only for specific type of features.
- Other metrics lose their properties for a greater number of features.

- Often we require a **feature selection** during segmentation.
- Or **feature grouping** at the same time.



NON-NEGATIVE MATRIX FACTORIZATION



The objective behind NMF is to:
summarize and split the information contained in V into *r* factors -- the columns of W.

- Plot source: en.wikipedia.org/wiki/Non-negative_matrix_factorization#/media/File:NMF.png

NMF
DESCRIPTION

AN OVERVIEW
OF THE DECOMPOSITION

$$\min_{W,H \geq 0} \underbrace{[D(X, WH) + R(W, H)]}_{=F(W,H)}$$

D is a loss function that measures the quality of the approximation.
Common loss functions are based on the Frobenius distance:

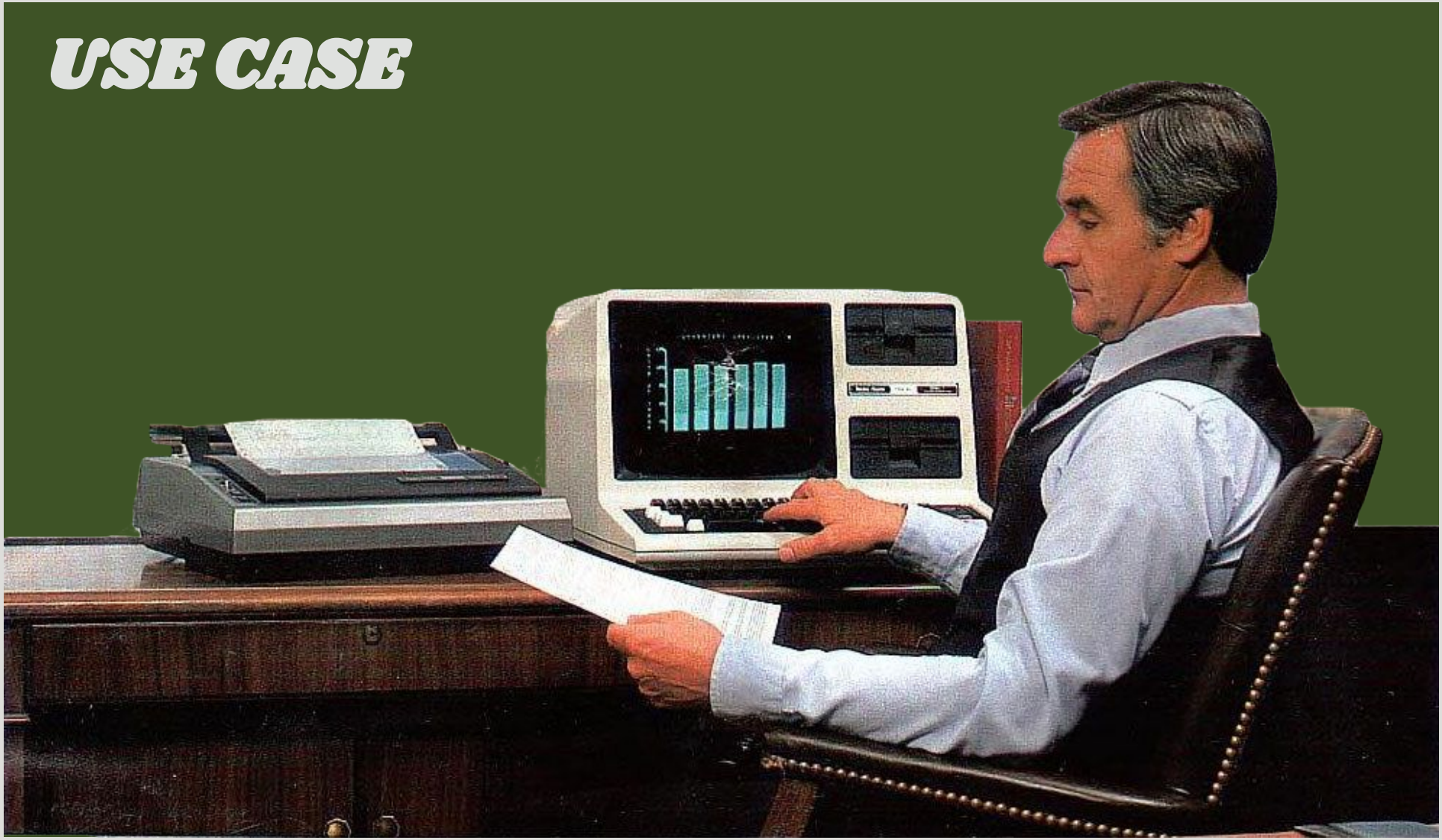
$$D : A, B \mapsto \frac{Tr(AB^t)}{2} = \frac{1}{2}\sum_{ij}(a_{ij} - b_{ij})^2$$

R is an optional regularization function, defined to enforce desirable properties on matrices W and H, such as smoothness or sparsity.

NMF algorithms generally solve problem iteratively, by building a sequence of matrices (W_k, H_k) that reduces at each step the value of the objective function F.

Beside some variations in the specification of F, they also differ in the optimization techniques that are used to compute the updates for (W_k, H_k).

```
nmf(x, rank, method, seed, ...)
```


USE CASE

USE CASE BACKGROUND

PURPOSE

Based on the country wide survey with +50 mindset statements

find *reasonable number* of groups in **teachers**' population,

+with valuable sizes

+and meaningful descriptions

+built on as small number of features as possible.

USE CASE DETERMINE RANK

Start N (e.g. 30) estimation processes for various Ks.

For each K calculate statistics of goodness of fit
- dispersion
- silhouette
- sparseness

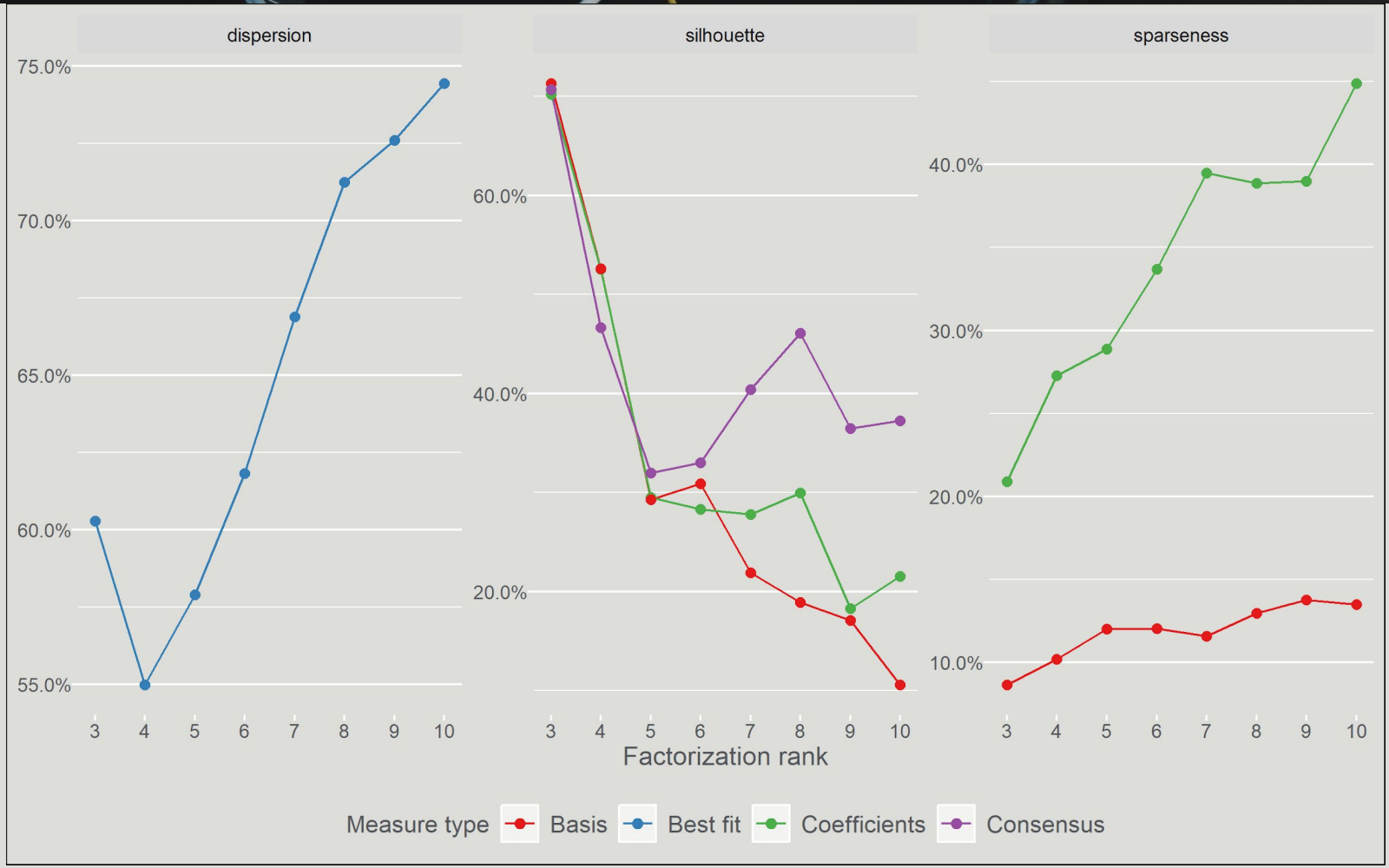
Look at the **consensus** clustering,

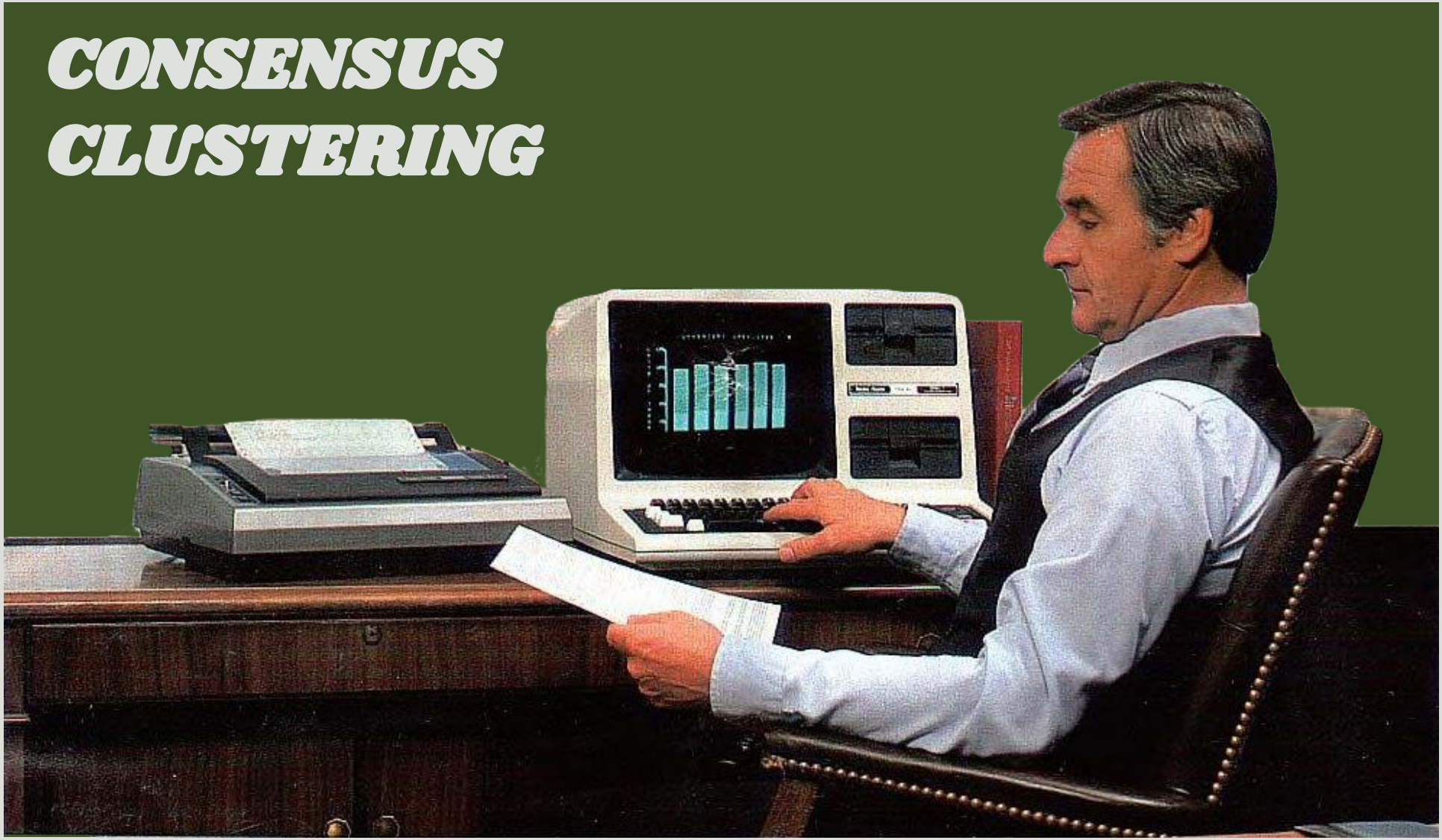
where a distance matrix is build on the co-occurrence metric,

that checks how many times specific features were assigned to the same segment.

That tells how stable was the solutioj

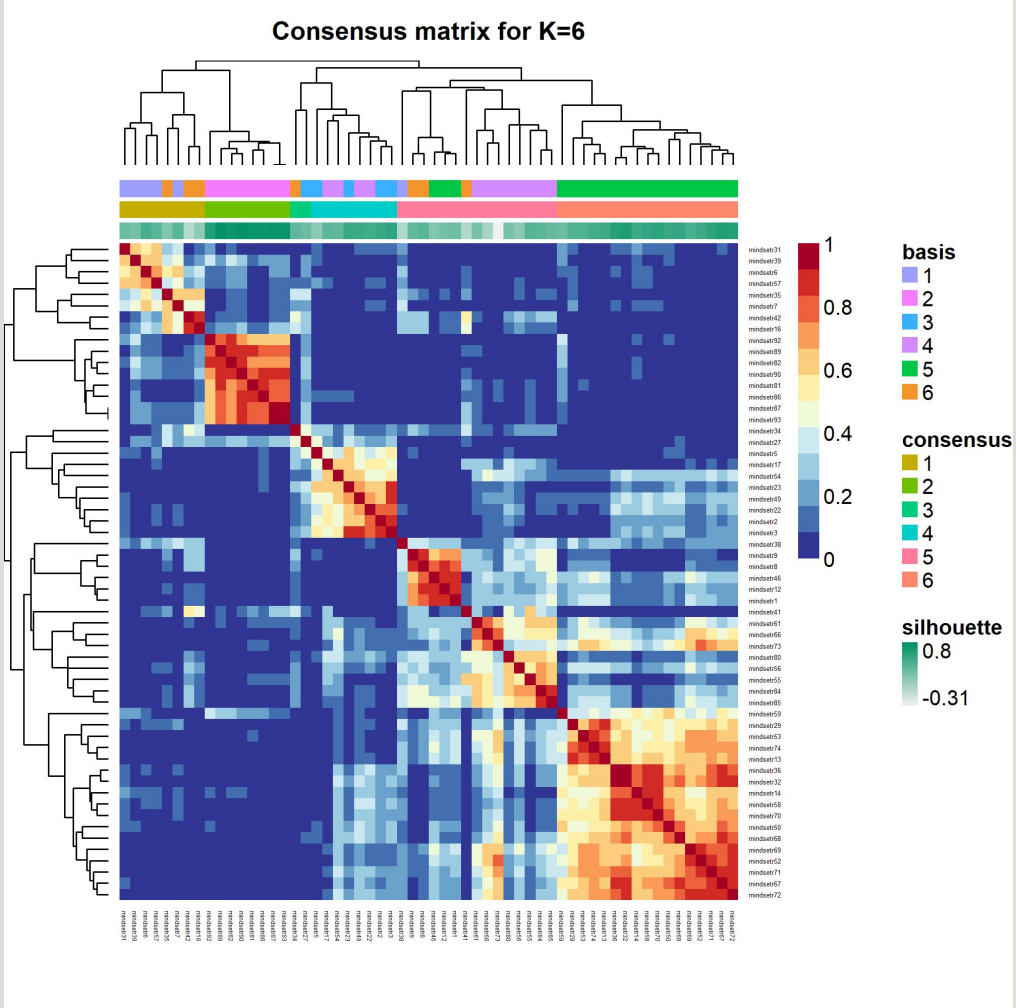
USE CASE
GOODNESS OF FIT

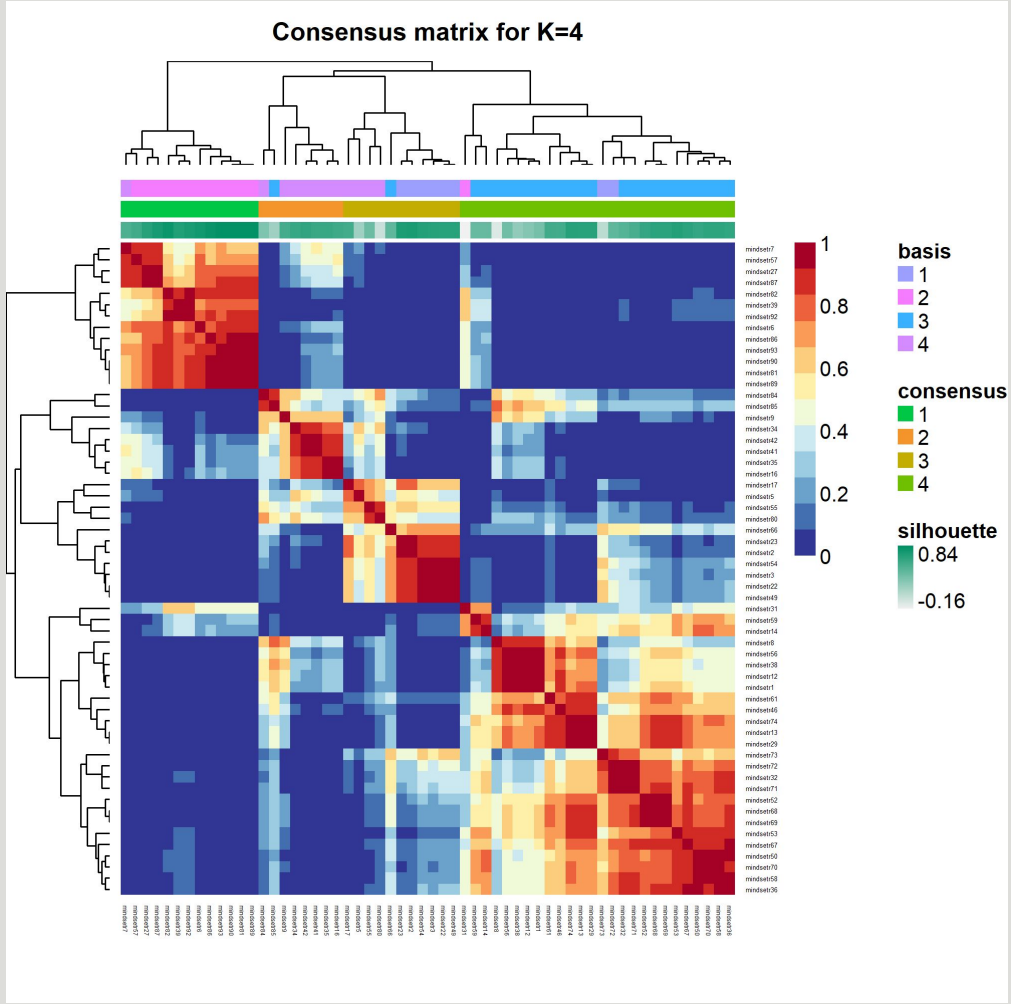



CONSENSUS
CLUSTERING

GRADIENT

USE CASE
DETERMINE RANK





USE CASE SUMMARY STATISTICS

K	Methodology	Silhouette score[1]	Dispersion[2]	Largest segment %	Smallest segment %
4	Non-negative matrix factorization (NMF)[3]	0.513	0.593	36%	15%
5		0.355	0.569	32%	10%
6		0.261	0.600	24%	10%

Methodology	Max silhouette score	Number of segments
k-Means	0.049	3
hclust	0.264	3
PAM	0.037	3
CLARA	0.037	3

USE CASE STORY

SEGMENT 4 (Size: 15%) FOLLOW-THE-GUIDEBOOK TEACHER	
Over-index (value)	*% strongly agree/agree*
Report card grades should compare students against one another. (292.4)	31
My students will have better opportunities with an education that prioritizes content knowledge over social-emotional learning and cognitive development. (185.56)	43
Our society values teachers. (263.16)	26
Learning can occur without interactions between teachers and students. (158.1)	40
Under-index (value)	*% strongly agree/agree*
Teachers are responsible for more than just imparting knowledge, like how to build healthy relationships. (66.84)	50
Teachers should have the freedom to innovate. (68.47)	60
A well-rounded education includes more than just teachers. Everyone — neighbors, community members — has a role in education. (62.44)	53
I don't need a pre-packaged lesson plan or tool to design a successful learning experience. (53.74)	38

NOTES

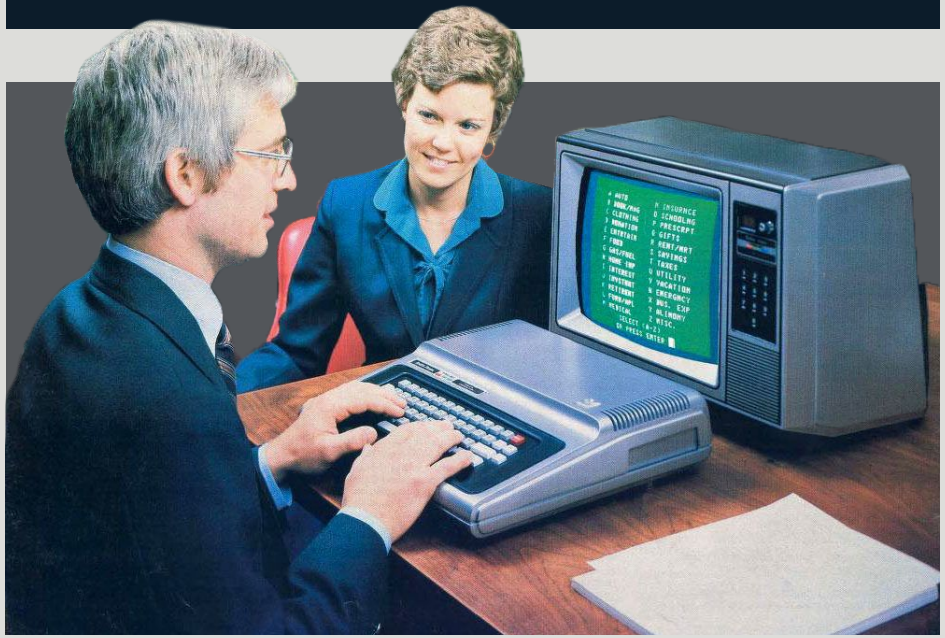

INVENTORY
DESCRIPTION
4K LEVEL I
16K LEVEL I
4K LEVEL II
16K LEVEL II
EXPANSION INT.
MINI-DISK I
MINI-DISK II
TRS-80 DESK
QUICK PRINTER
RS-232
TELEPHONE INTERFACE I

Most NMF versions have a random start point so make sure to establish the stable solution with multiple runs.

Rank selection is the crucial part in the workflow and should be supported with diagnostic statistics.

Feature clustering is provided simultaneously to observations' segmentation.

ALL BASED ON NMF R PACKAGE



Credits:

cran.r-project.org/package=NMF

renozao.github.io/NMF/

Gaujoux R and Seoighe C (2010).

“A flexible R package for nonnegative matrix factorization.”

BMC Bioinformatics, 11(1), pp. 367.
ISSN 1471-2105,

dx.doi.org/10.1186/1471-2105-11-367,
www.biomedcentral.com/1471-2105/11/367.

GRADIENT

DID YOU LIKE THE TALK? JOIN US AT WHY R? 2019.

WHY R? 2019

is an international conference of R statistical software users. A remarkable occasion to gain knowledge about R, enhance data analysis and processing skills, and meet other members of **European R community**.

4 days of extensive networking in Warsaw, Poland

Thursday	Friday	Saturday	Sunday
26.09	27.09	28.09	29.09
Hackathon Free entrance Prizes Evening Integration: Dinner for Speakers	Workshops 2 sessions 4 workshops each Evening Integration: 5 Round Tables Sessions	Lectures CONFERENCE 3 keynotes Sponsor sessions Panel Discussion Evening Integration: Welcome Reception	3 keynotes Sponsor sessions Lightning talks

Target Group: *Data Scientists*, both from academia and business, on an expert as well as on a junior level.
We expect an international and diverse audience of **R enthusiasts**.

Our events - whyr.pl

5 special interests groups **8** promoting events
6 keynotes
30 presentations
welcome party

Keynote Speakers

Steph Locke
Principal Consultant (GBR) | Locke Data

Paula Brito
Faculty of Economics, University of Porto

Jakub Nowosad
Institute of Geoecology and Geoinformation, Adam Mickiewicz University

Sigrid Keydana
Applied Researcher at RStudio

Wit Jakuczun
WLOG Solutions

Marvin Wright
Leibniz Institute for Prevention Research and Epidemiology

Contact us kontakt@whyr.pl whyr.pl/2019/

2017: whyr.pl/2017/movie/
2018: whyr.pl/2018/movie/

ORGANIZERS

UNIVERSITY OF WARSAW | University of Warsaw Faculty of Economic Sciences | WHY R? FOUNDATION | PL in ML Association

WARSAW,

26-29 SEPTEMBER 2019

HTTP://WHYR.PL/2019/

THANK YOU FOR THE ATTENTION

github.com/g6t/nmf