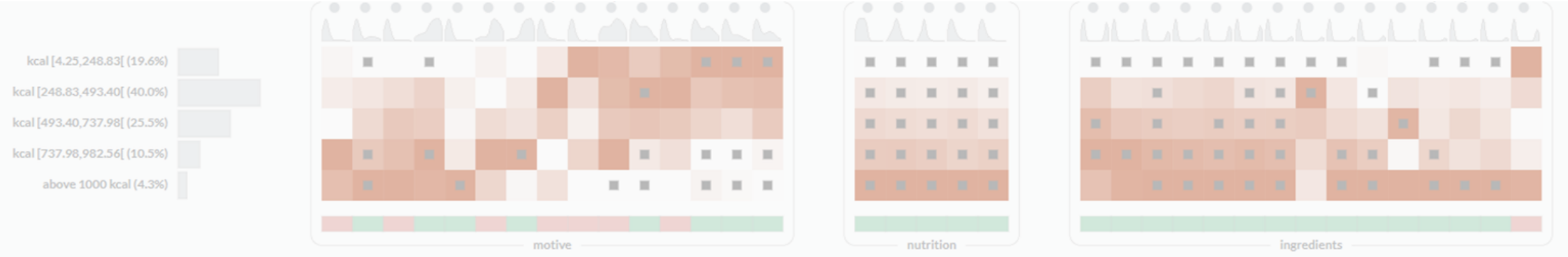


SMARTexplore: Simplifying High-Dimensional Data Analysis through a **Table-Based Visual Analytics Approach**

M. Blumenschein, M. Behrisch, S. Schmid, S. Butscher,
D. R. Wahl, K. Villinger, B. Renner, H. Reiterer, and D. A. Keim

University of Konstanz, Germany Harvard University, USA

Universität Konstanz

Data Analysis and Visualization

SMARTACT

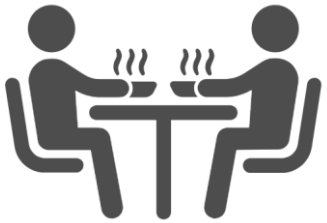
SFB-TRR 161
Quantitative Methoden für Visual Computing

What influences the heating behavior of people?

Health Psychologists

What do people eat?

Motives for particular dish?



Impact of surrounding?

Background and Dataset

99 participants

2,571 meals

motives + surroundings

manual extraction of nutrition + ingredients

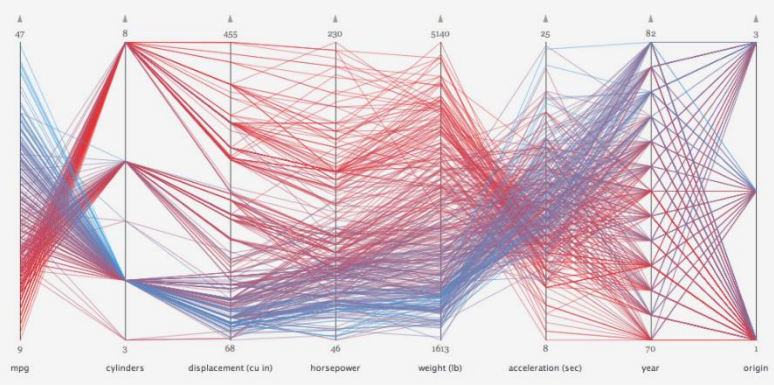
id	sex	age	bmi	meal type	day	where	with whom	mood before	mood after	...	...	fish	meat	milk	eggs	...
0	m	35	35.4	breakfast	mo	home	alone	1	2	...	...	n	n	y	y	...
1	m	36	35.4	lunch	mo	restaur	friends	4	4	...	...	y	n	y	y	...
2	f	55	28.3	breakfast	tu	home	family	2	3	...	...	n	y	y	n	...
3	f	66	28.2	tea	we	work	colleg.	3	2	...	...	n	n	n	n	...
4	f	25	28.2	tea	th	work	colleg.	1	1	...	...	n	n	n	n	...
5	f	62	28.2	supper	fr	work	family	2	3	...	...	y	n	y	n	...
6	m	53	24.7	snack	sa	friends	friends	4	4	...	...	n	y	n	y	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

id	sex	age	bmi	meal type	day	where	with whom	mood before	mood after	...	...	fish	meat	milk	eggs	...
0	m	35	35.4	breakfast	mo	home	alone	1	2	...	...	n	n	y	y	...
1	m	36	35.4	lunch	mo	restaur	friends	4	4	...	...	y	n	y	y	...
2	f	55	28.3	breakfa			family	2		...		n	y		n	...
3	f	66	28.2	tea	we	work	colleg.	3	2	...	...	n	n	n	n	...
4	f	25	28.2	tea	th	work	colleg.	1	1	...	...	n	n	n	n	...
5	f	62	28.2	supper	fr	work	family	2	3	...	...	y	n	y	n	...
6	m	53	24.7	snack	sa	friends	friends	4	4	...	...	n	y	n	y	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

categorical

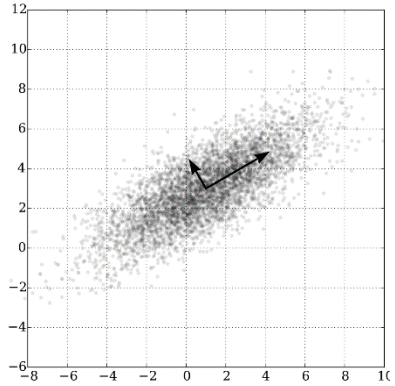
numerical

binary

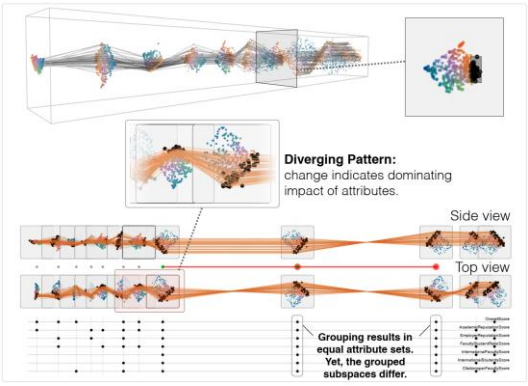


Parallel Coordinates

Scatter Plot (Matrix)



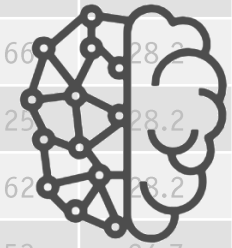
Projection



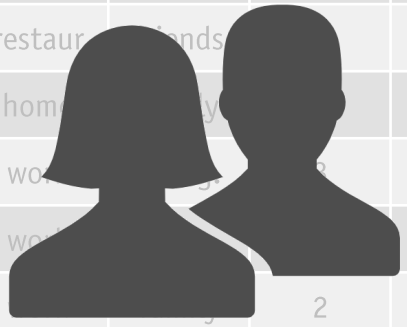
Subspace Analysis

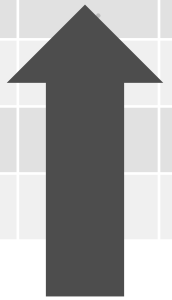
https://datavizproject.com/data-type/parallel-coordinates/
https://plotlyblog.tumblr.com/post/174657459542/what-is-a-splom-chart-making-scatterplot-matrices
https://en.wikipedia.org/wiki/Principal_component_analysis
Jäckle et al. Pattern Trails: Visual Analysis of Pattern Transitions in Subspaces (VAST 2017)

id	sex	age	bmi	meal type	day	where	with whom	mood before	mood after	...	...	fish	meat	milk	eggs	...
0	m	35	35.4	breakfast	mo	home	alone	1	2	...	...	n	n	y	y	...
1	m	36	35.4	lunch	mo	restaur	ends		4	...	...	y	n	y	y	...
2	f	55	28.3	breakfast	tu	hom	y		3	...	...	n	y	y	n	...
3	f	66	28.2	tea	we	wo			2	...	...	n	n	n	n	...
4	f	25	28.2	tea	th	wo			1	...	...	n	n	n	n	...
5	f	62	28.2	supper	fr		y	2		...		y	n	y	n	...
6	m	53	24.7	snack	sa	friends	friends	4	4	...					y	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

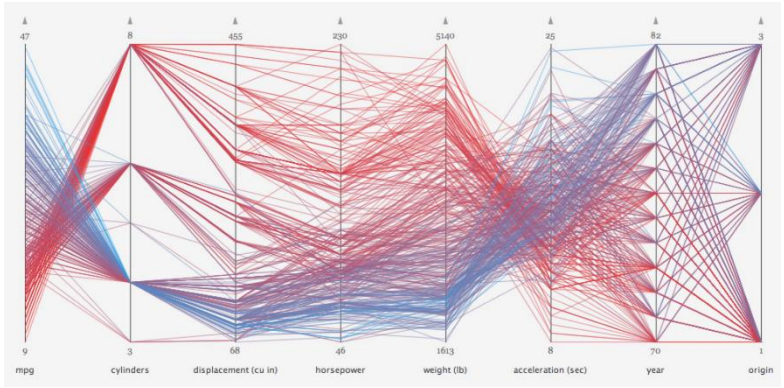


change mental model



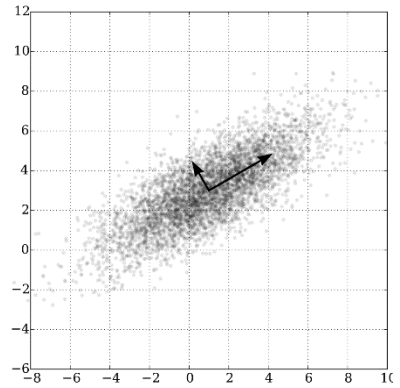


check findings
statistic validation

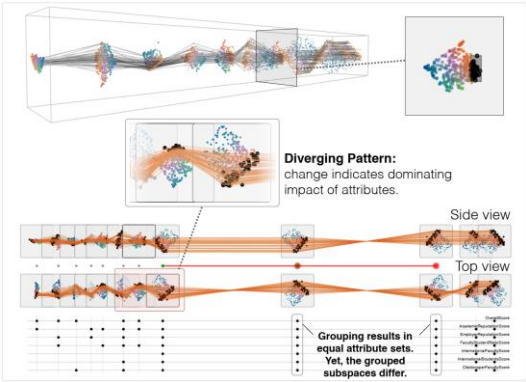


Parallel Coordinates

Scatter Plot (Matrix)



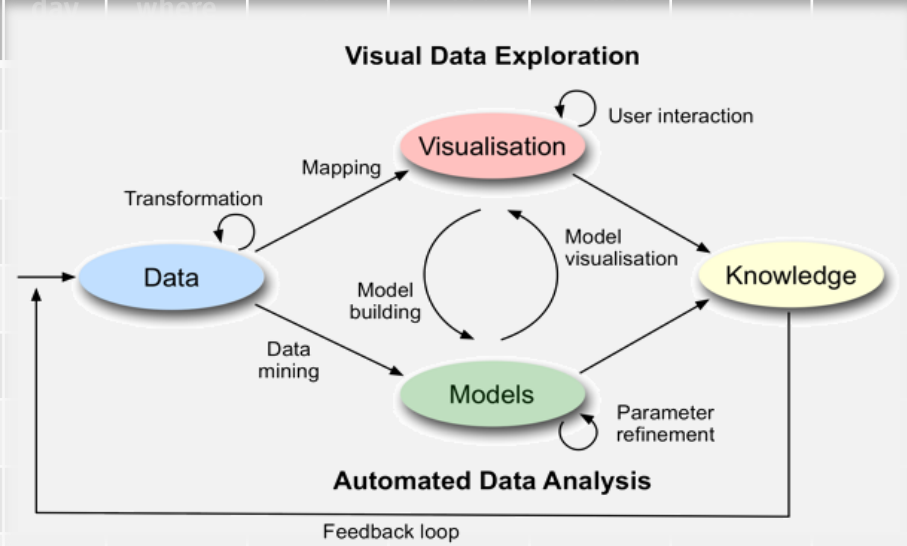
Projection



Subspace Analysis

https://datavizproject.com/data-type/parallel-coordinates/
https://plotlyblog.tumblr.com/post/174657459542/what-is-a-splom-chart-making-scatterplot-matrices
https://en.wikipedia.org/wiki/Principal_component_analysis
Jäckle et al. Pattern Trails: Visual Analysis of Pattern Transitions in Subspaces (VAST 2017)

id	sex	age	bmi	meal type	day	where	with	mood	mood			fish	meat	milk	eggs	...
0	m	35	35.4	breakfast								n	n	y	y	...
1	m	36	35.4	lunch								y	n	y	y	...
2	f	55	28.3	breakfast								n	y	y	n	...
3	f	66	28.2	tea								n	n	n	n	...
4	f	25	28.2	tea								n	n	n	n	...
5	f	62	28.2	supper								y	n	y	n	...
6	m	53	24.7	snack								n	y	n	y	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...



- Automatic handling and aggregation of mixed data types
- Simplification of complex data transformations
- Automatic pattern identification and highlighting
- Automated reliability analysis (of visual patterns)

Keim, Daniel, et al. "Visual analytics: Definition, process, and challenges." Information visualization. Springer, Berlin, Heidelberg, 2008. 154-175.

id	sex	age	bmi	meal type	day	where	with whom	mood before	mood after	fish	meat	milk	eggs	...	eggs	...
0	m	35	35.4	breakfast	mo	home	alone	1	2	n	n	y	y	...	y	...
1	m	36	35.4	lunch	mo	restaur	friends	4	4	y	n	y	y	...	y	...
2	f	55	28.3	breakfast	tu	home	family	2	3	n	y	y	n	...	n	...
3	f	66	28.2	tea	we	work	colleg.	3	2	n	n	n	n	...	n	...
4	f	25	28.2	tea	th	work	colleg.	1	1	n	n	n	n	...	n	...
5	f	62	28.2	supper	fr	work	family	2	3	y	n	y	n	...	n	...
6	m	53	24.7	snack	sa	friends	friends	4	4	n	y	n	y	...	y	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Subspace A

Subspace B

Subspace C

G1

G2

G3

id	sex	age	bmi
0	m	35	35.4
1	m	36	35.4
2	f	55	28.3
3	f	66	28.2
4	f	25	28.2
5	f	62	28.2
6	m	53	24.7
...	...	...	...

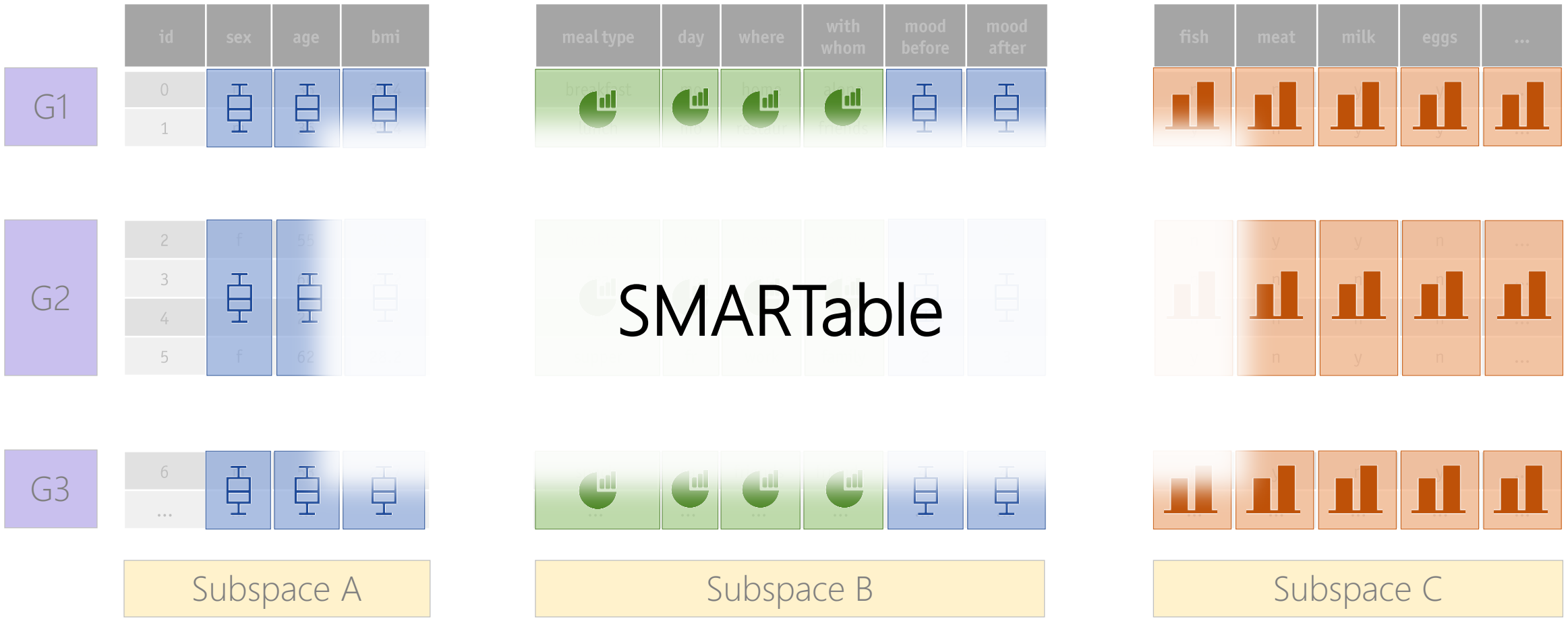
Subspace A

meal type	day	where	with whom	mood before	mood after
breakfast	mo	home	alone	1	2
lunch	mo	restaur	friends	4	4
breakfast	tu	home	family	2	3
tea	we	work	colleg.	3	2
tea	th	work	colleg.	1	1
supper	fr	work	family	2	3
snack	sa	friends	friends	4	4
...	...	...	...	...	...

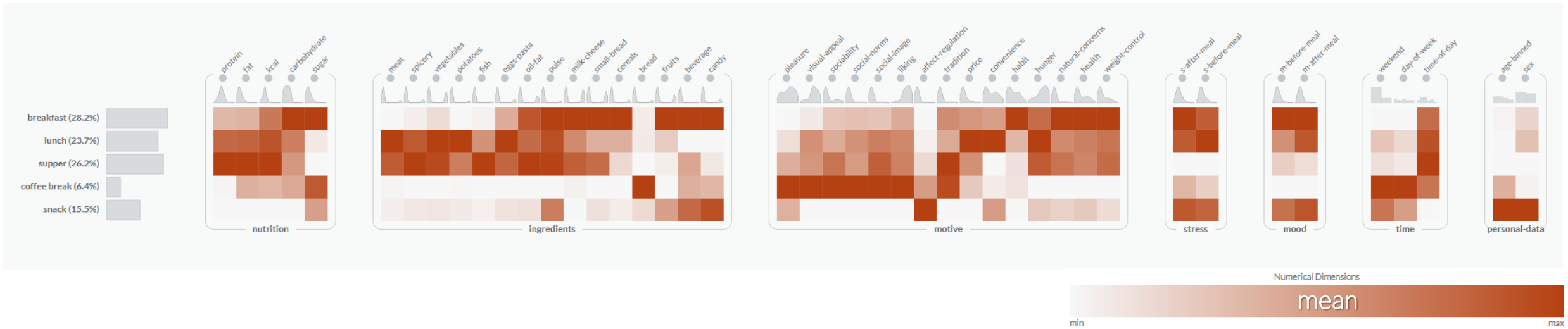
Subspace B

fish	meat	milk	eggs	...
n	n	y	y	...
y	n	y	y	...
n	y	y	n	...
n	n	n	n	...
n	n	n	n	...
y	n	y	n	...
n	y	n	y	...
...	...	...	...	...

Subspace C

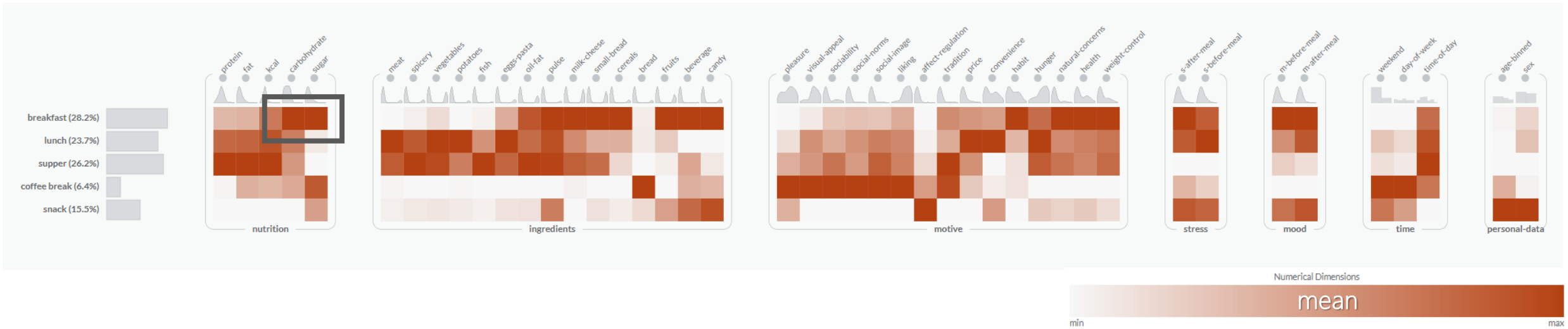

id
sex
age
bmi
meal type
day
where
with whom
mood before
mood after
fish
meat
milk
eggs
...
G1
G2
G3
SMARTable
Subspace A
Subspace B
Subspace C

Comparison of different meal types



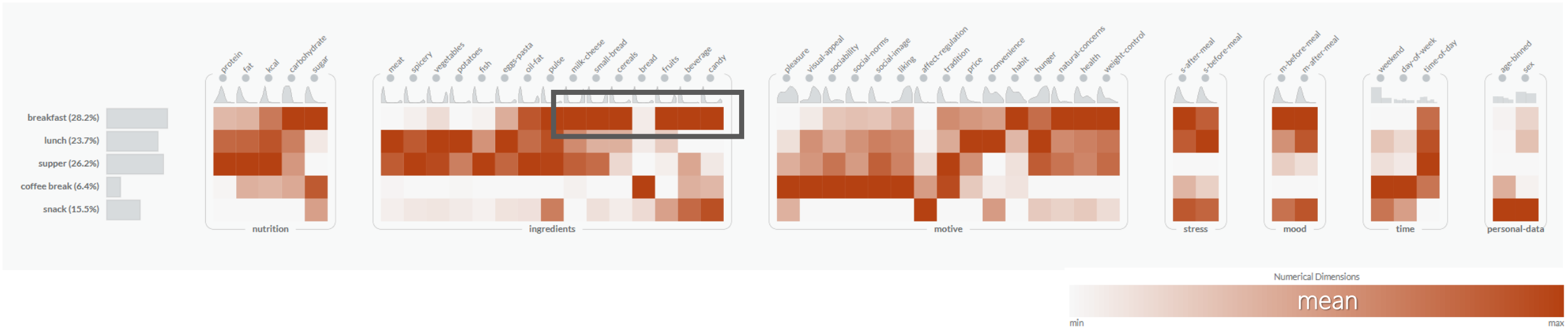
- Records are grouped by **meal type**
- **Semantic grouping** of dimensions

Comparison of different meal types



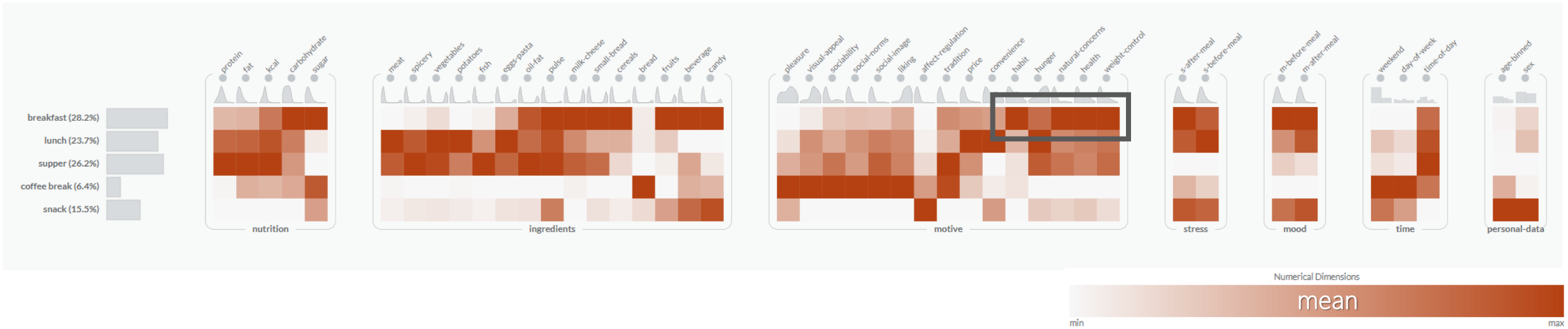
- Records are grouped by **meal type**
- **Semantic grouping** of dimensions

Comparison of different meal types



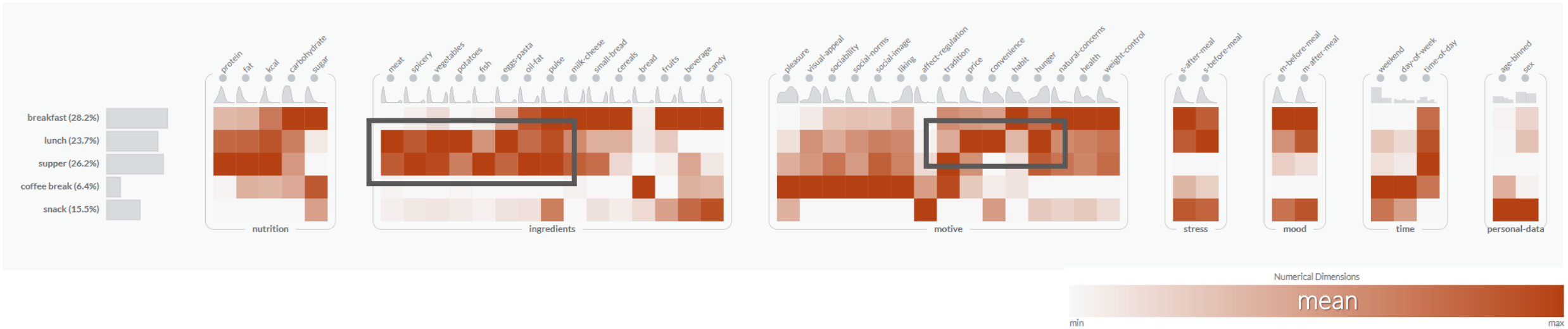
- Records are grouped by **meal type**
- **Semantic grouping** of dimensions

Comparison of different meal types



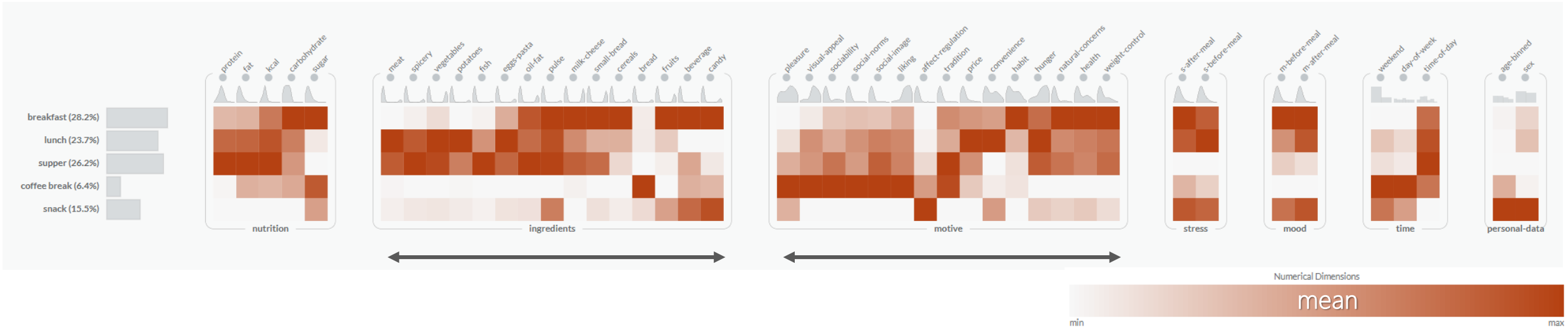
- Records are grouped by **meal type**
- **Semantic grouping** of dimensions

Comparison of different meal types



- Records are grouped by **meal type**
- **Semantic grouping** of dimensions

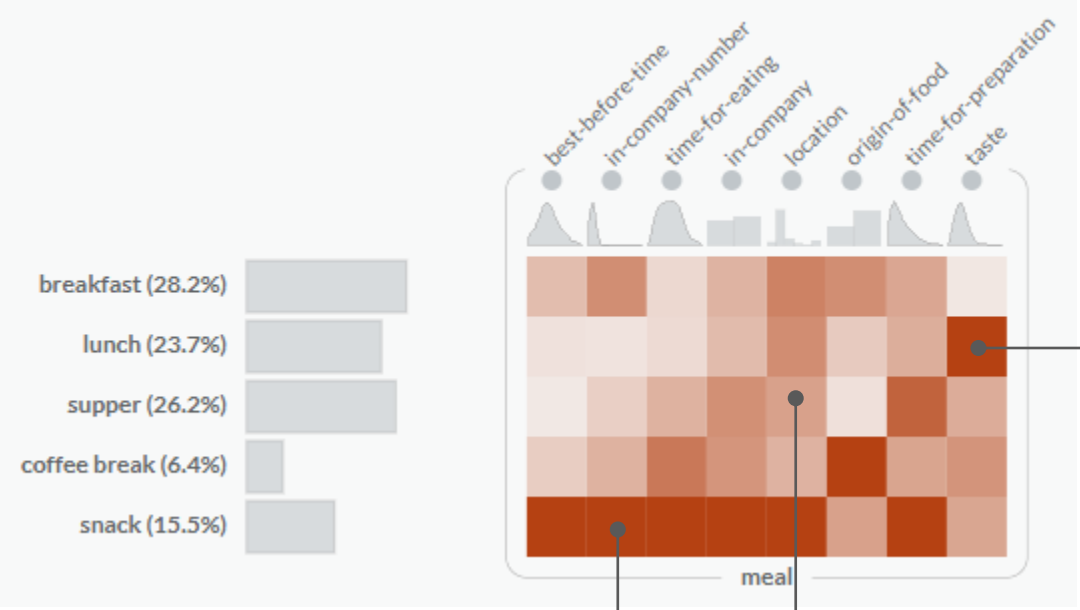
Comparison of different meal types



- Records are grouped by meal type
- Semantic grouping of dimensions

- Automatic sorting of dimensions by visual similarity or avg. descriptor

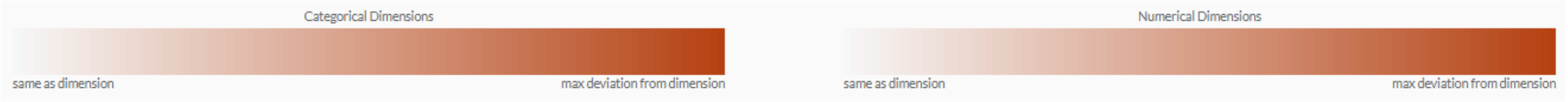
Subspace with mixed data types



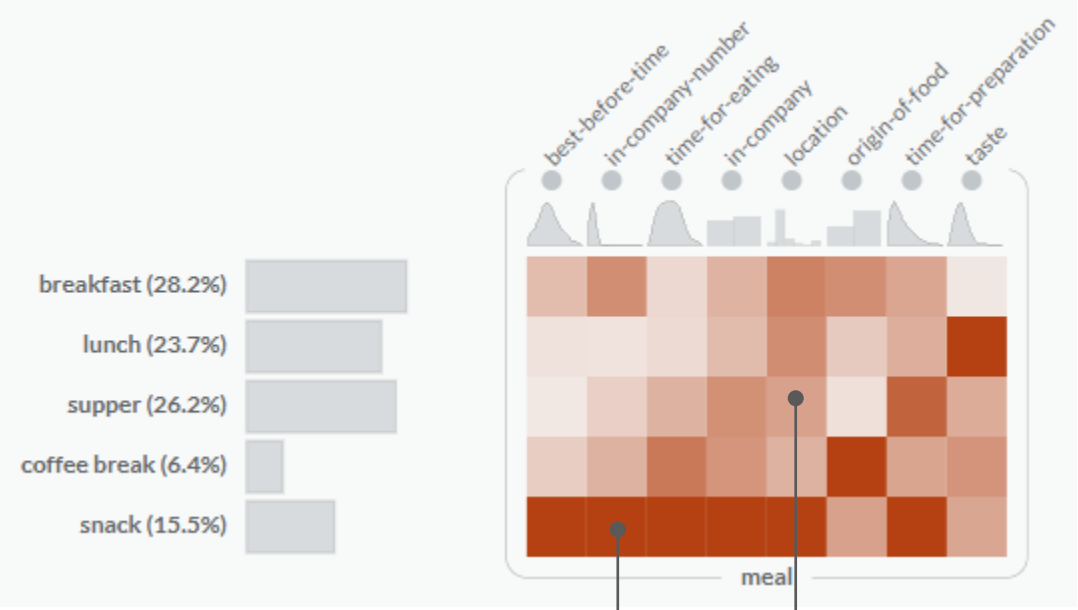
high-value
- highest *deviation* from dimension *taste*
- but no information about the actual value

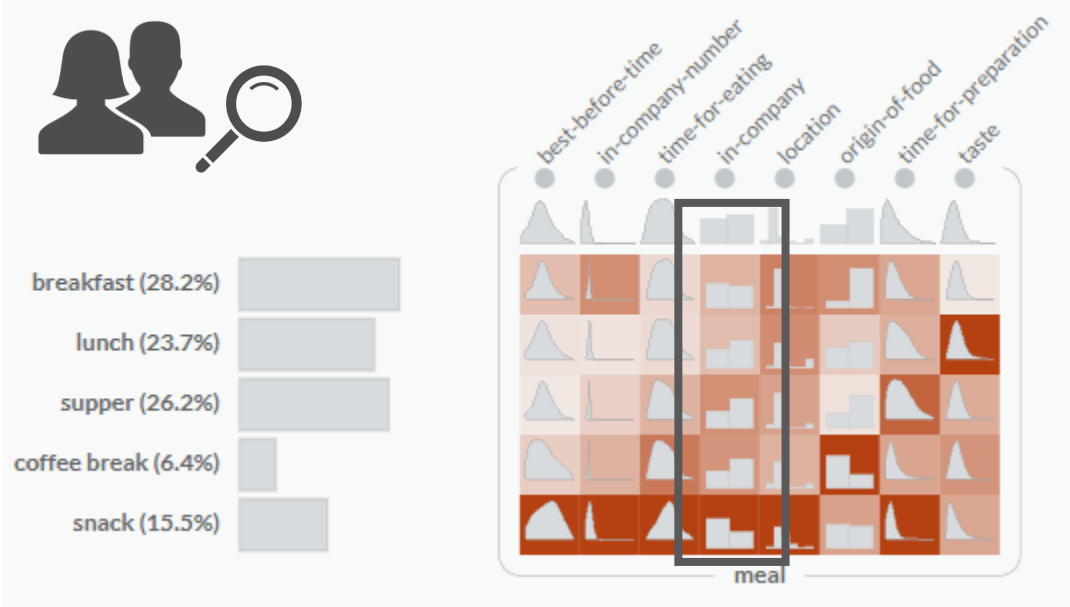
categorical dimension
- value = Eucl. distance between frequency histograms

numerical + binary dimension
- value = $| mean_{group} - mean_{dimension} |$



Subspace with mixed data types

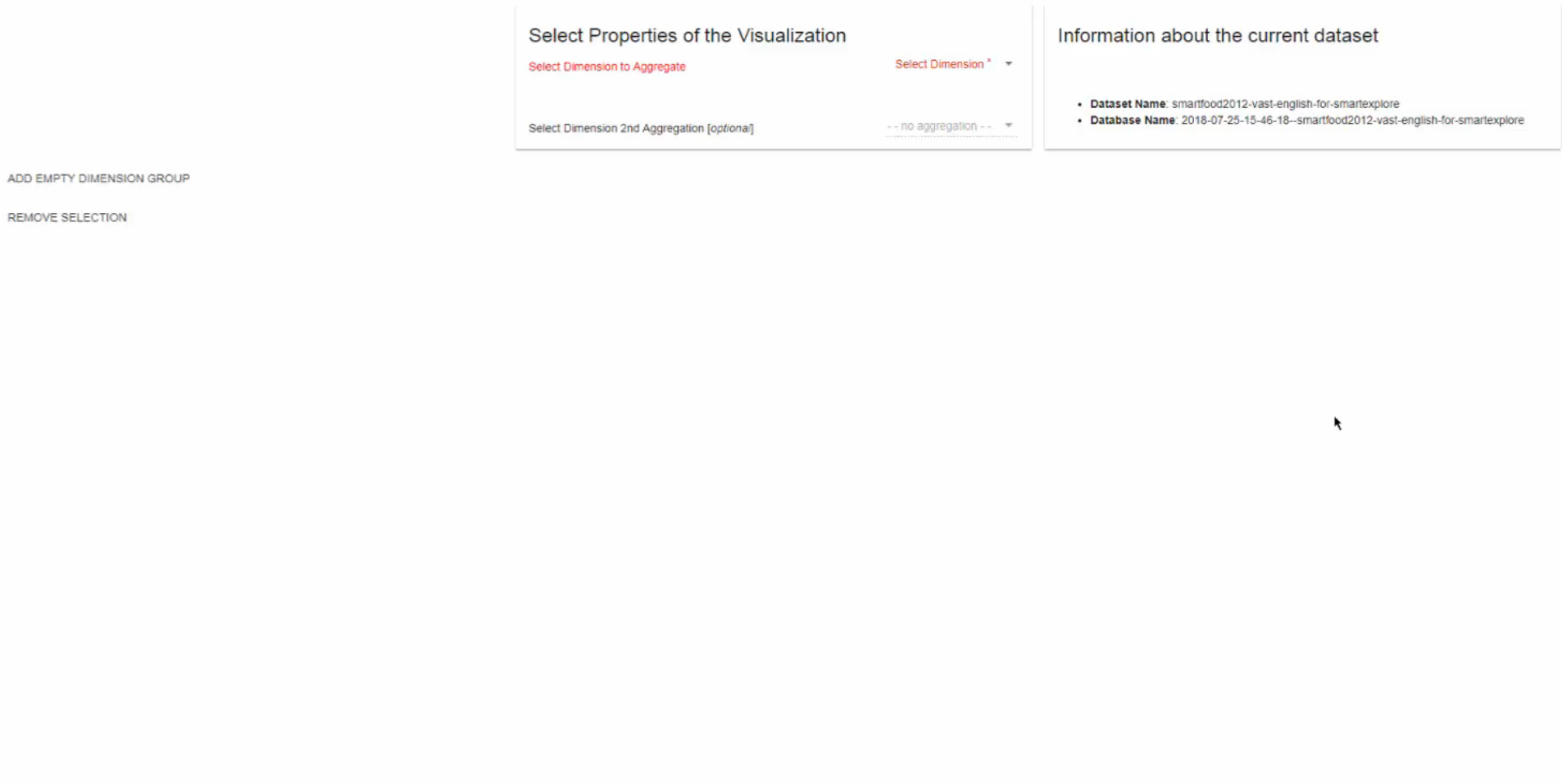




categorical dimension
- value = Eucl. distance between frequency histograms

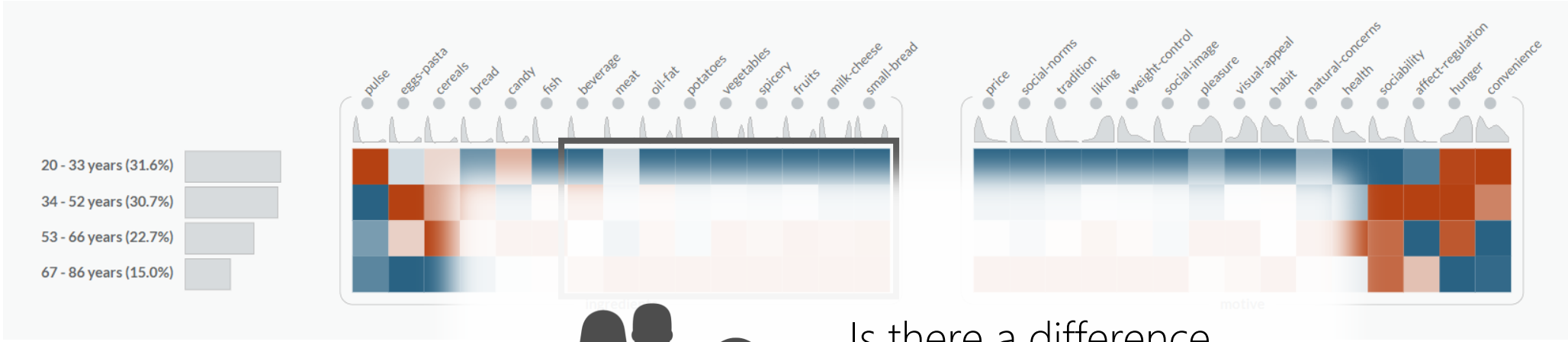
numerical + binary dimension
- value = $| mean_{group} - mean_{dimension} |$

Interaction within the SMARTable



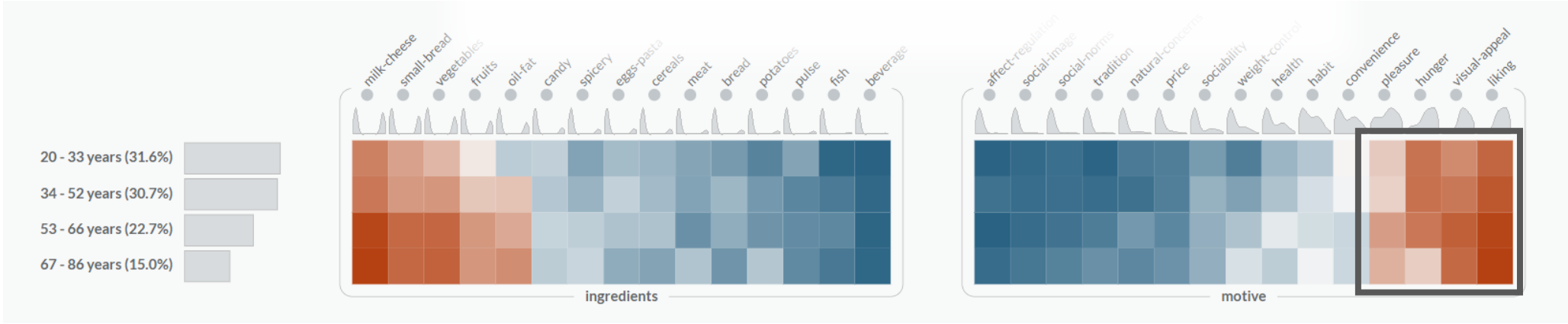
Normalizing Strategies

Normalize per dimension -- Useful for dimension with **different** scale

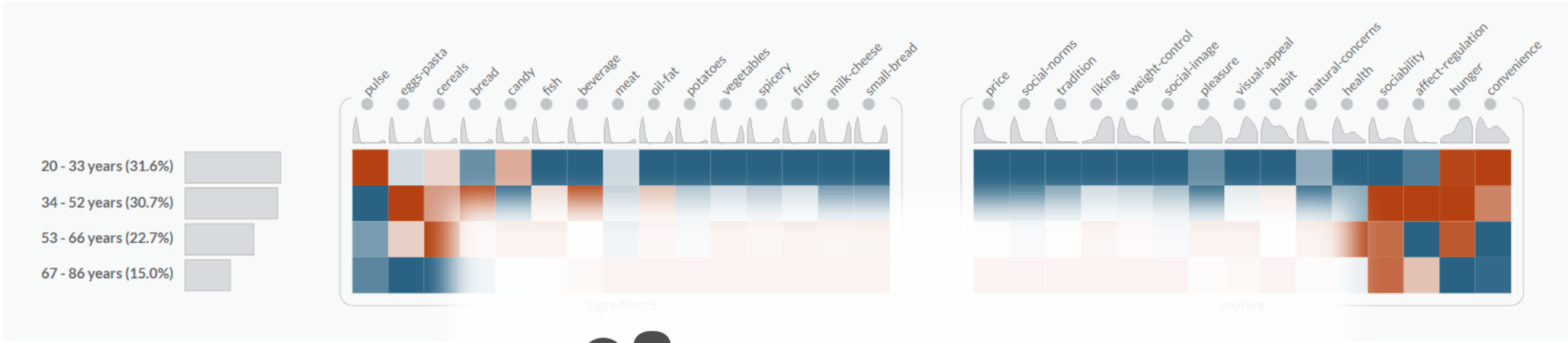


Is there a difference
between male and female?

Normalize per subspace -- Useful for dimension with **same** scale

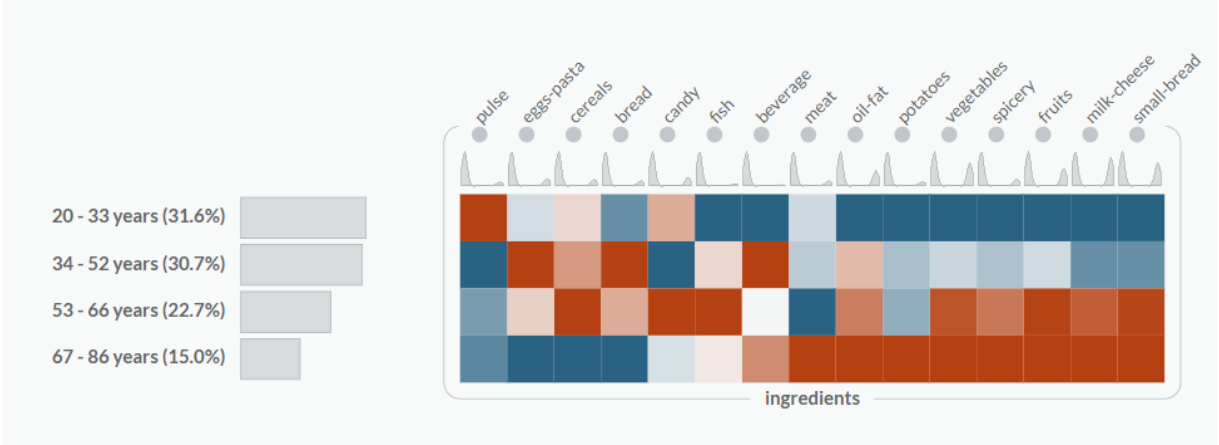


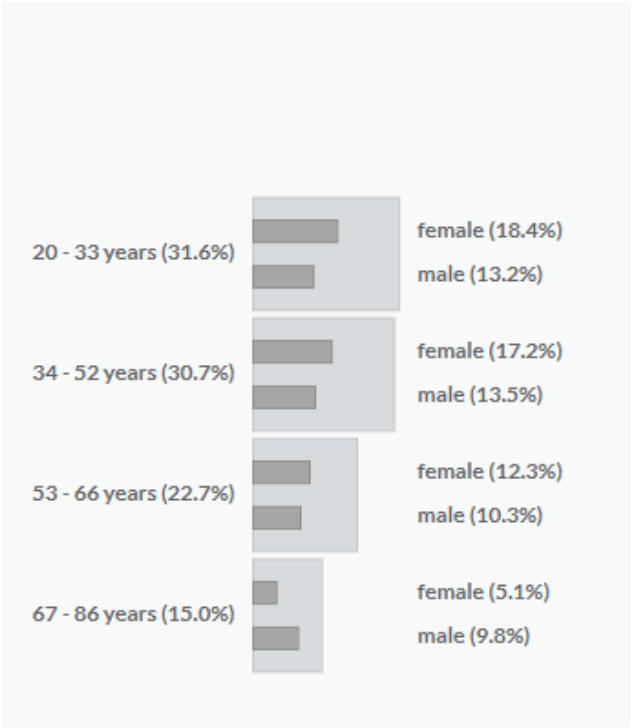
Drill-down: Stacked SMARTable



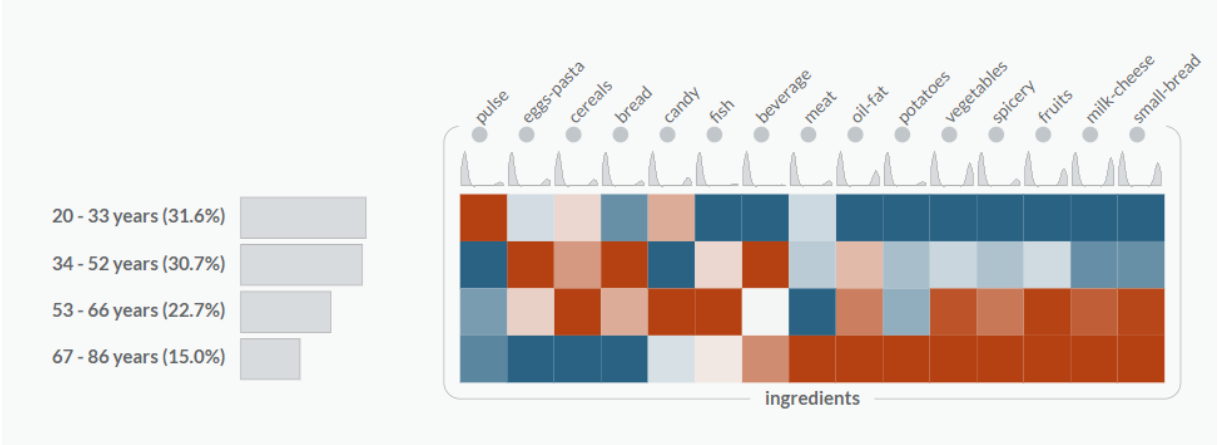
Is there a difference between male and female?

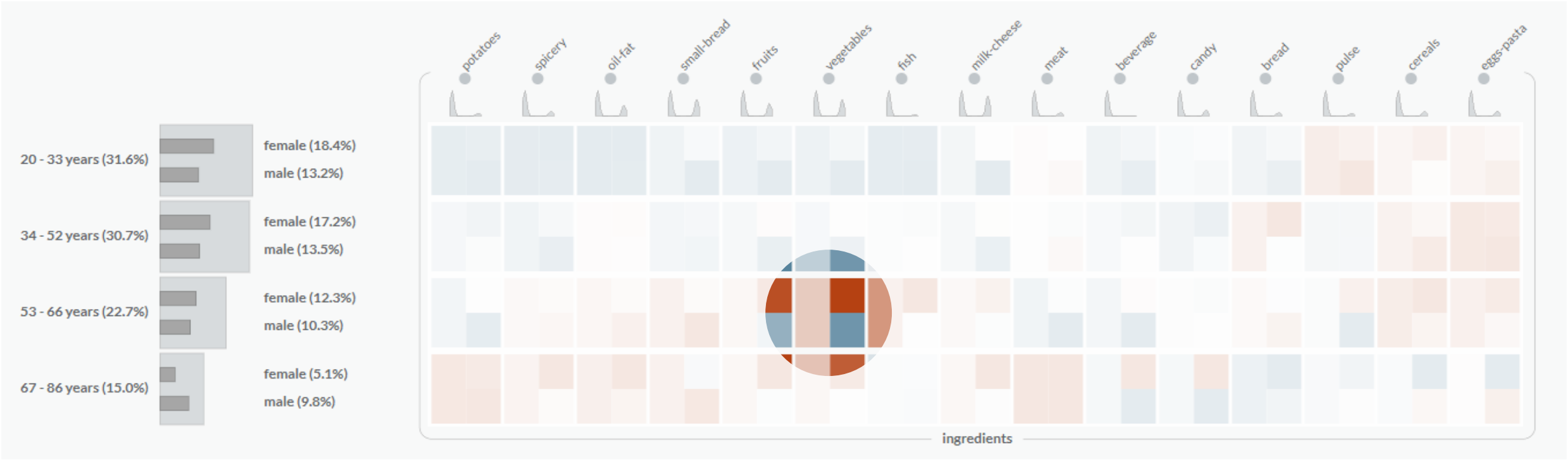
Drill-down: Stacked SMARTable



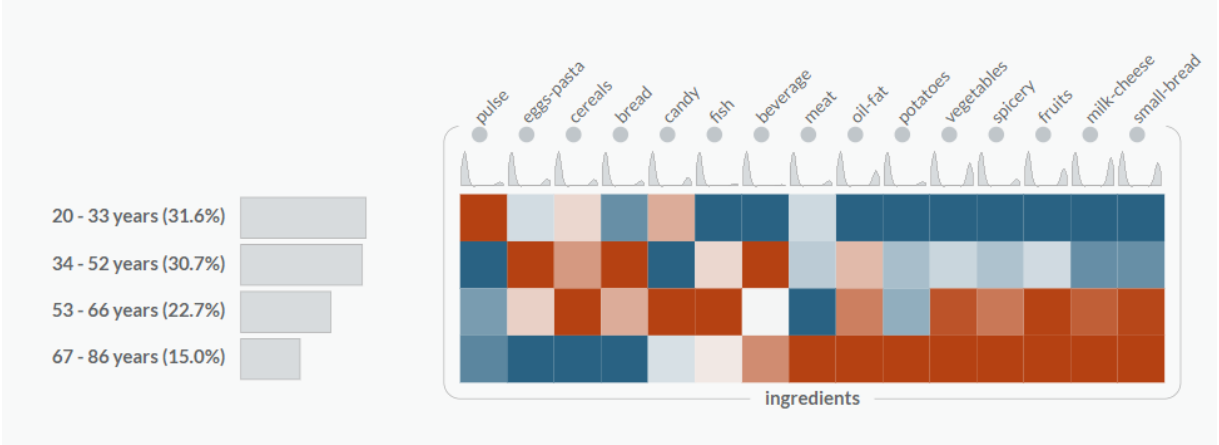


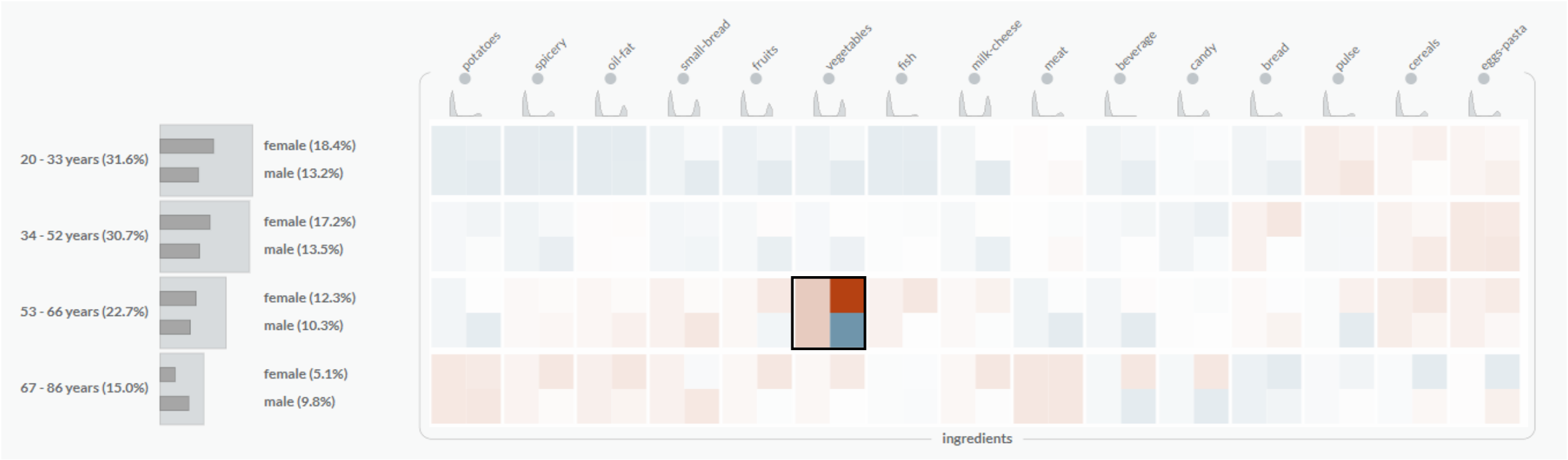
Drill-down: Stacked SMARTable



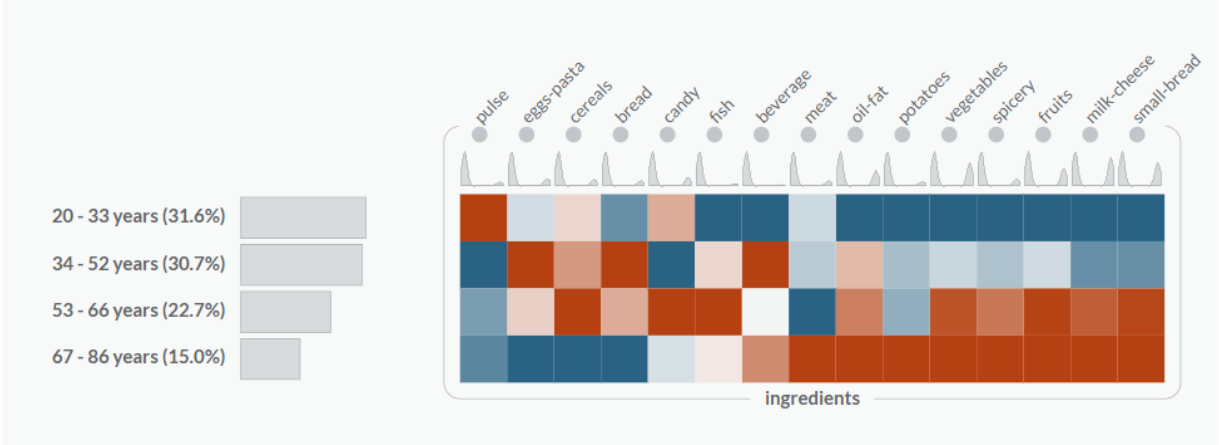


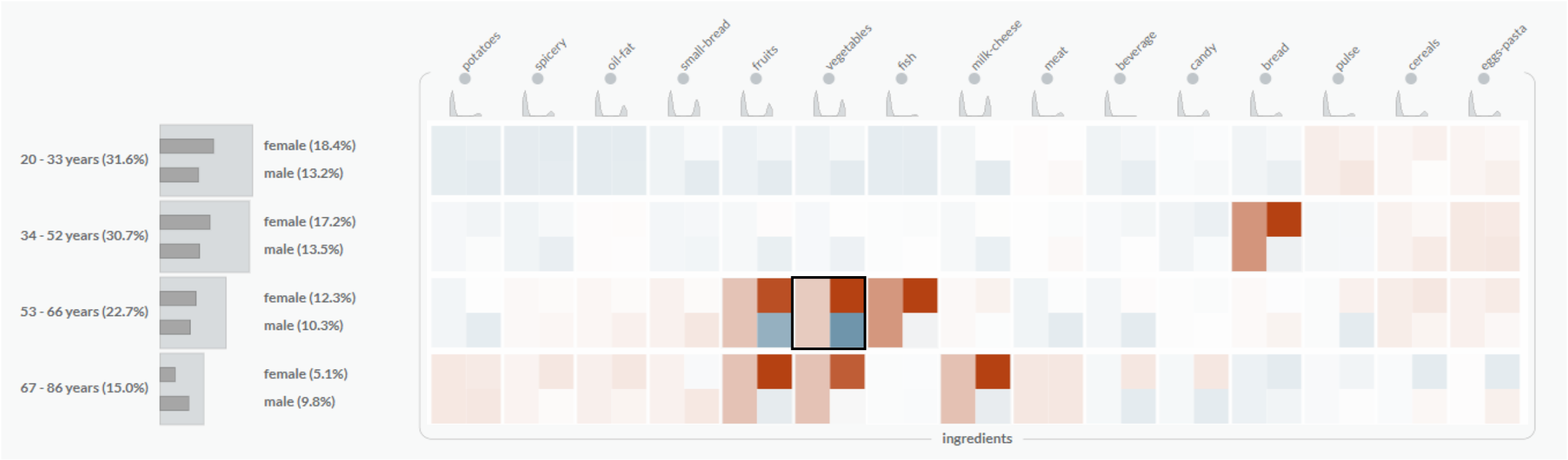
Similarity Search





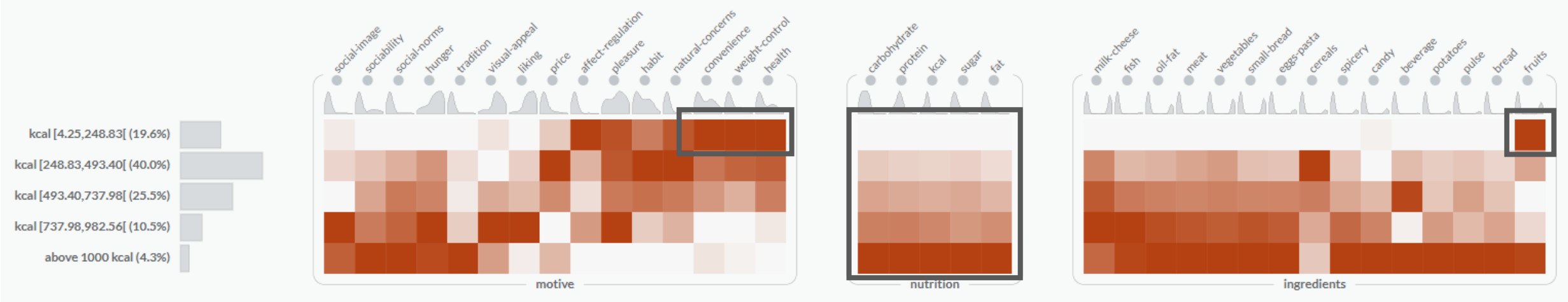
Similarity Search





Reliability of visual patterns

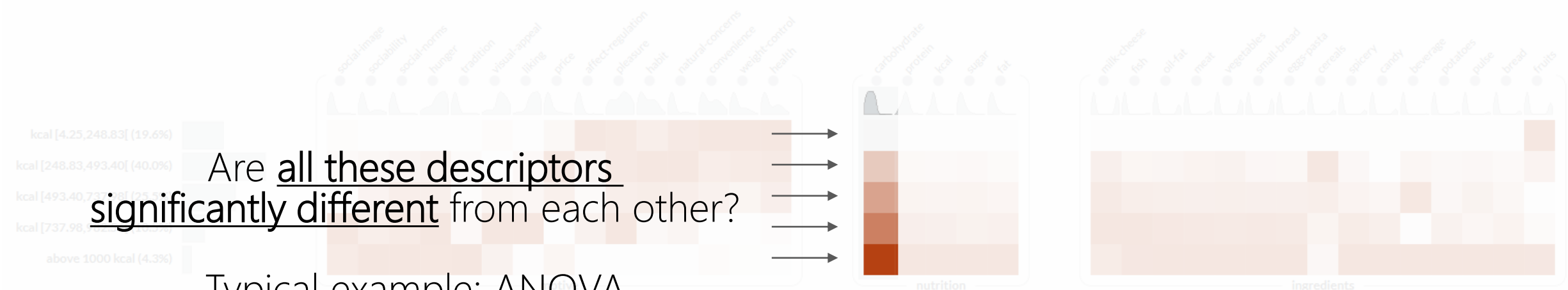
Can we trust the patterns which we perceive?



Visible patterns

- **(linear) correlation** between kcal and all nutrition
- meals with **low calories** are related to the motive **health** and **weight control** and the **ingredient** is mainly **fruits**

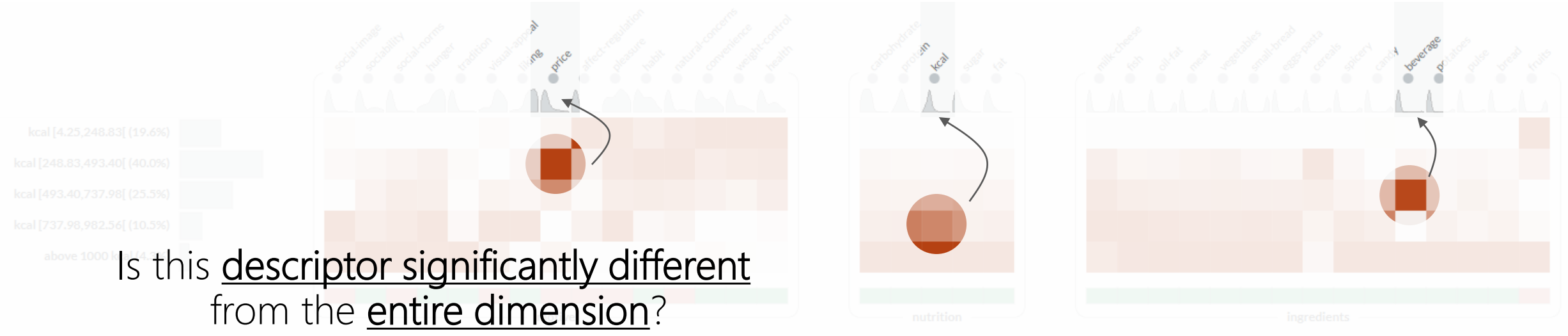
Statistical significance of a dimension



Are <u>all these descriptors</u>
<u>significantly different</u> from each other?

Typical example: ANOVA

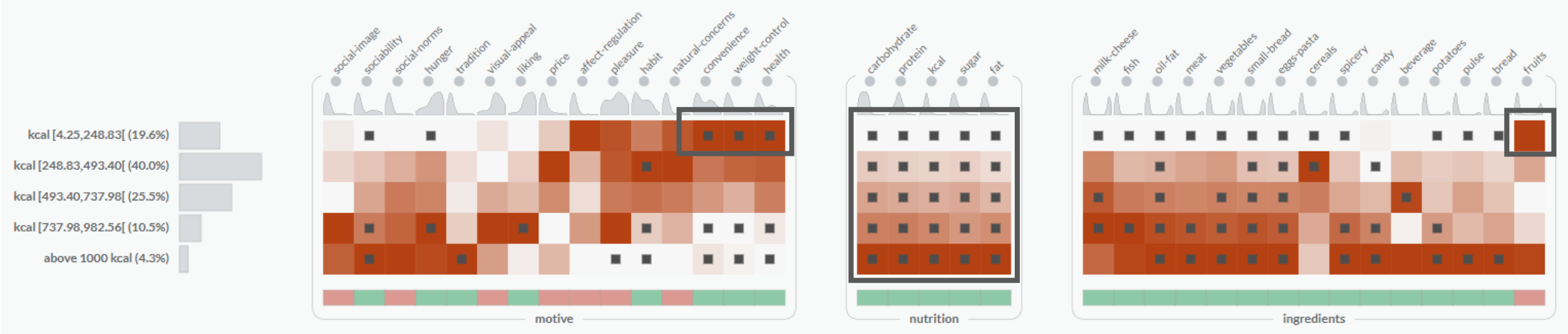
Statistical significance of a descriptor



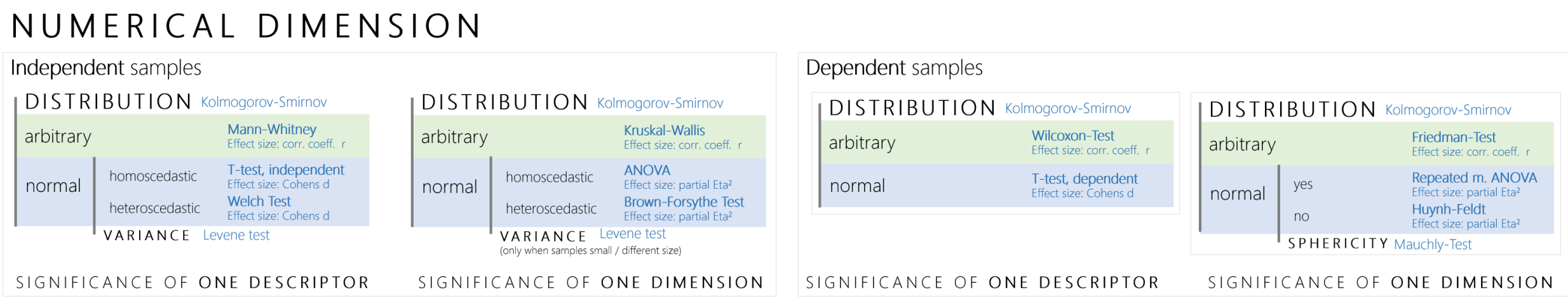
Is this descriptor significantly different from the entire dimension?

Typical example: t-test

Statistical significance of **descriptors** and **dimensions**



Assumption-based (automatic) selection of statistical test

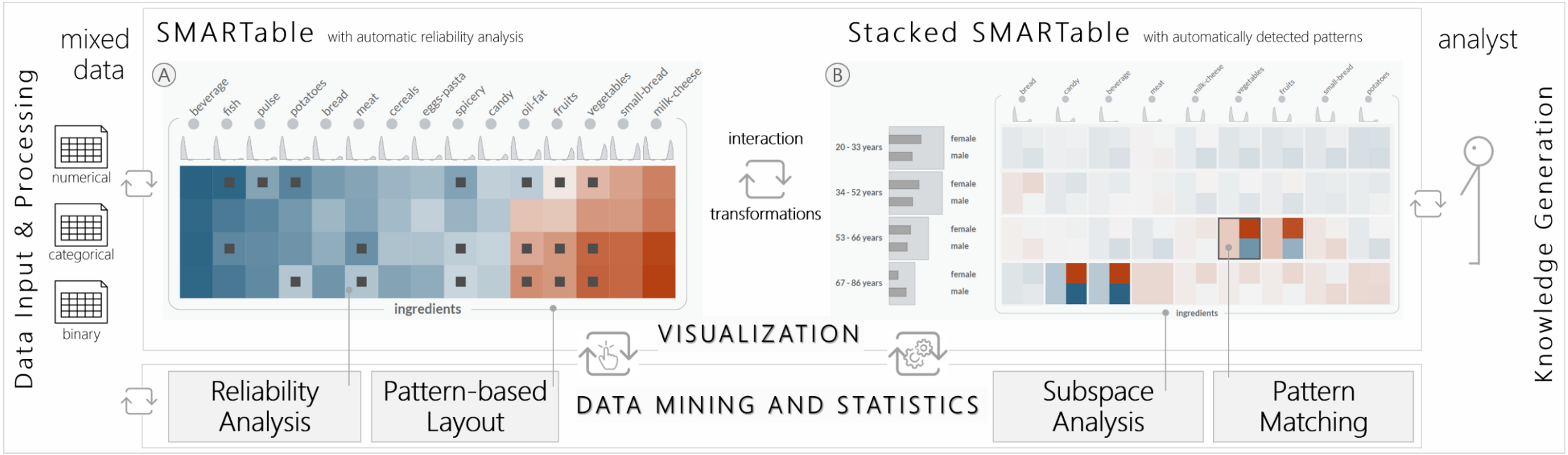


- Appropriate test is **automatically selected** by data/distribution properties
- Based on the rules by Andy Field [1]
- Similar rules for categorical and binary dimensions

[1] A. Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013

Future Work based on expert user feedback

- **Support hypothesis generation**
  by including automatic algorithms e.g., subspace clustering
- **More data types**
  e.g., time series, etc.
- **Layout flexibility**
  change back and forth between classical approaches and the SMARTable
- **Data analytical provenance**
  add explicit gallery view



https://smartexplore.dbvis.de

@mi_blumenschein