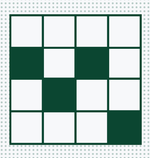
Cheng and Church's Biclustering

SINCE 2000

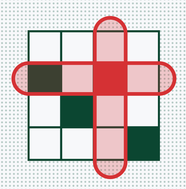
Cheng & Church are two well-known scientists, also famous for preparing family-size pizzas in their restaurant at Harvard. Right now, they have to manage an order for a group of 16, though it's hard to make everyone happy. Firstly, they give the clients a list of 16 ingredients and ask each of them to compile their preferences.

Now they will have to cross people and their tastes to group clients that will share the same pizza.
They will use **Biclustering**: a data mining technique that allows to obtain groups (biclusters) of similar objects, by simultaneously clustering both the rows (objects) and the columns (features) of a matrix-format data set.

THE DATA

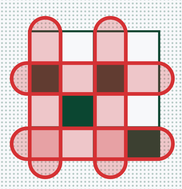


Let's create a table where rows represent guests and columns represent ingredients. This will be our matrix.

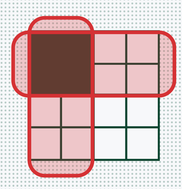


Every cell represents a single preference of each guest for each ingredient as a binary value: **1 (like) or 0 (dislike).**

THE GOAL

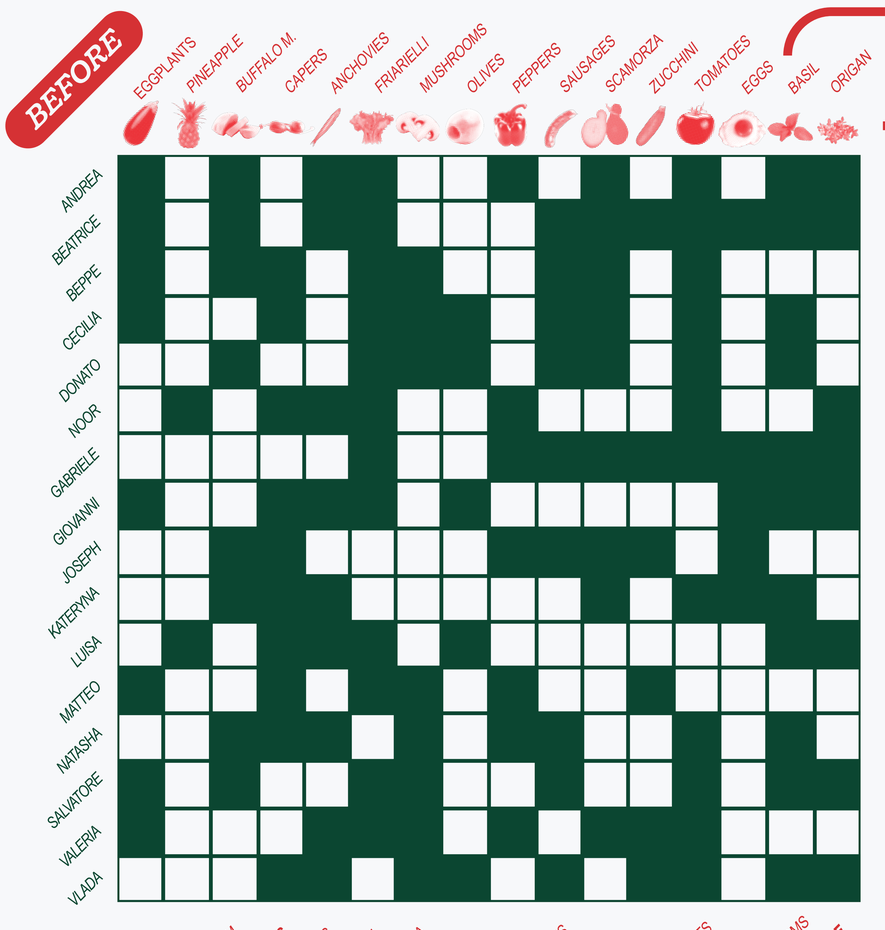


The Biclustering method crosses **simultaneously rows** of guests **and columns** of ingredients to find common preferences.

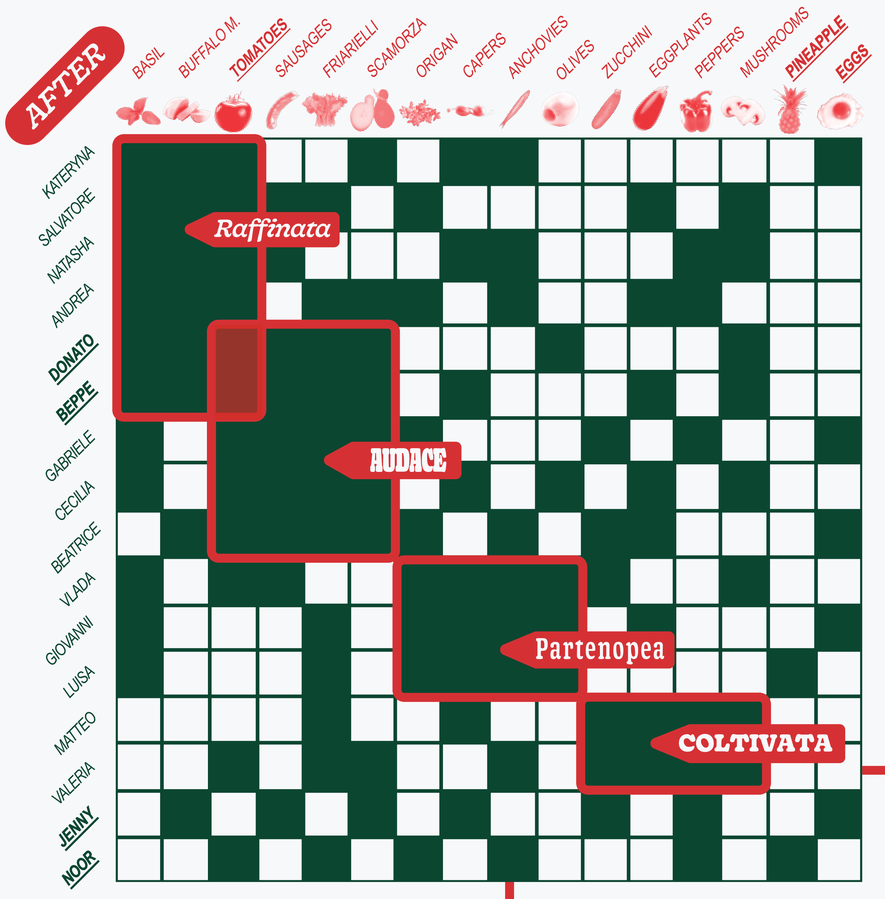


Every **bicluster** found in this matrix corresponds to a possible pizza.

BEFORE

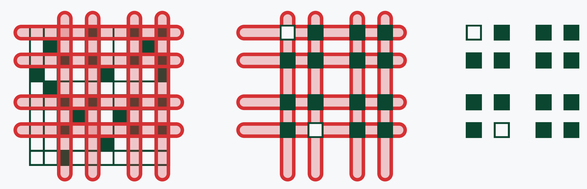


AFTER



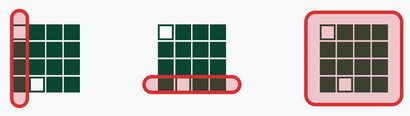
PROCESS

1



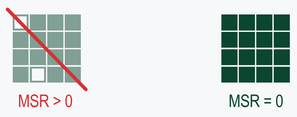
The algorithm simultaneously selects a group of **random rows and random columns**, creating a submatrix.

2



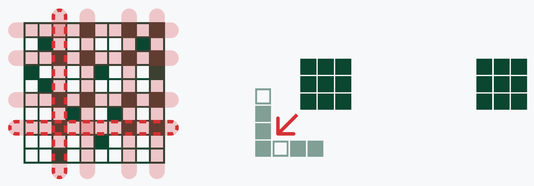
The homogeneity of the submatrix is measured by calculating the level of variability of each cell in relation to the column, the row and the whole submatrix it belongs to.
This value is called **Mean Square Residue.**

3



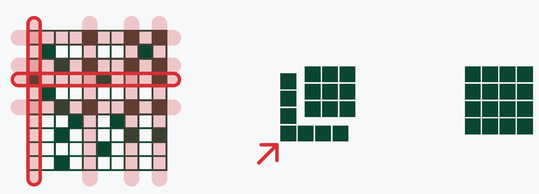
The **Mean Square Residue** has to be lower than a **threshold (delta)** that is chosen in advance. In our case the delta must be 0, since the goal is to find "ideal biclusters" with only colored cells.

4



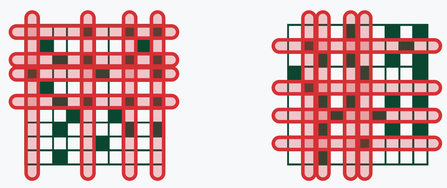
If the submatrix is not homogenous enough, the algorithm proceeds **removing** rows and columns until the MSR is below the delta.

5



Then, it randomly selects other rows and columns and **adds** them to the submatrix, only if the MSR value stays acceptable.

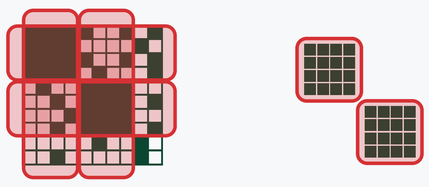
6



When the algorithm has examined all the rows and columns in the matrix, the bicluster is completed.

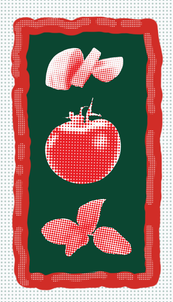
The process is repeated until the found biclusters are satisfying or all the possible combinations are found.

7



Once the final combination is found, usually the matrix's rows and columns are reordered to better visualize the biclusters.

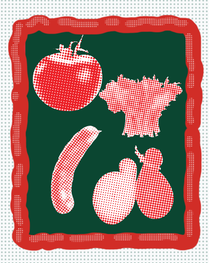
RESULTS



Raffinata

for KATERYNA, SALVATORE, NATASHA, ANDREA, DONATO and BEPPE

Toppings for this pizza are BASIL, BUFFALO MOZZARELLA and TOMATOES.



AUDACE

for DONATO, BEPPE, GABRIELE, CECILIA and BEATRICE

Toppings for this pizza are TOMATOES, SAUSAGES, FRIARIELLI and SCAMORZA.

Partenopea

for VLADA, GIOVANNI and LUISA

Toppings for this pizza are ORIGAN, CAPERS, ANCHOVIES and OLIVES.

COLTIVATA

for MATTEO and VALERIA

Toppings for this pizza are MUSHROOMS, PEPPERS, EGGPLANTS and ZUCCHINI.

OVERLAPS

DONATO and **BEPPE** are in 2 biclusters, they can eat both *Raffinata* and **AUDACE**.

LEFTOVERS

JENNY and **NOOR** aren't in a bicluster, they will get a Margherita. No pizza will have **EGGS** or **PINEAPPLE**.

VISUAL EXPLANATIONS OF STATISTICAL METHODS

Biclustering

AUTHORS

Martina Bracchi
Giovanni Bonassi
Maria Cecilia Buonocunto
Silvia Casavola
Vladislava Ershova
Kateryna Lapshyna
Matteo Visini

FACULTY

Michele Mauri
Ángeles Briones
Gabriele Colombo
Simone Vantini
Salvatore Zingale

TEACHING ASSISTANTS

Elena Aversa
Andrea Benedetti
Tommaso Elli
Beatrice Gobbo
Arianna Bellantuono

Final Synthesis Design Studio
Sect. C3

LM in Communication Design
A.A. 2022/2023

DEN–SITY GN+

POLITECNICO MILANO 1863
SCHOOL OF DESIGN