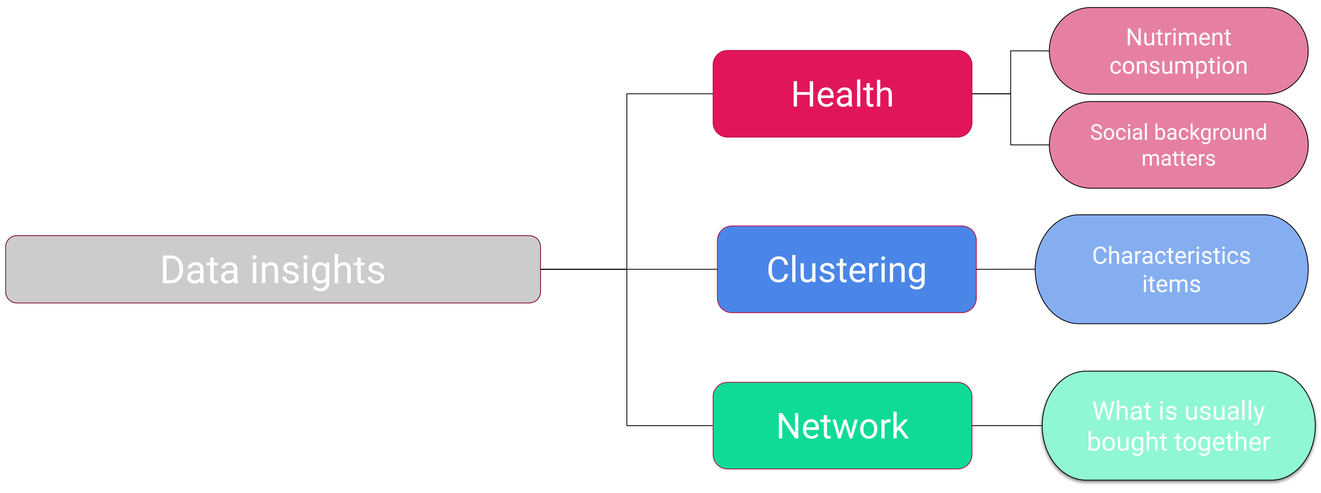
Transaction Data: the New Fingerprint

Profiling Households Based on supermarket consumption patterns

Jérôme Lüscher
Yann Mentha
Daniel Pacheco
Raphaël Reis
EPFL

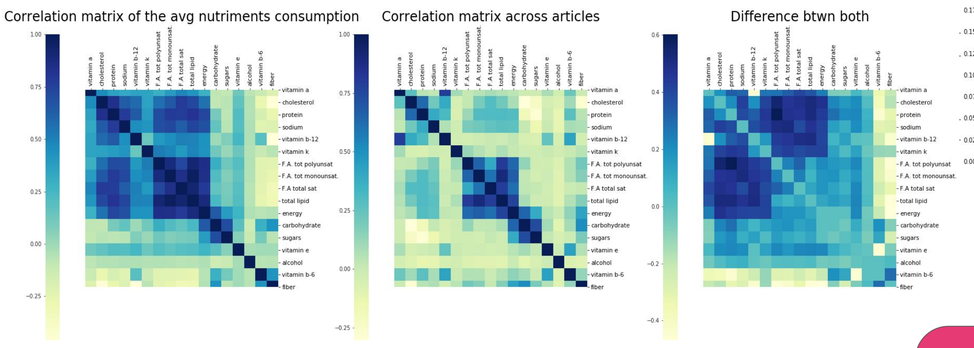
Applied Data Analysis
2019-2020
(website)

Project Structure:

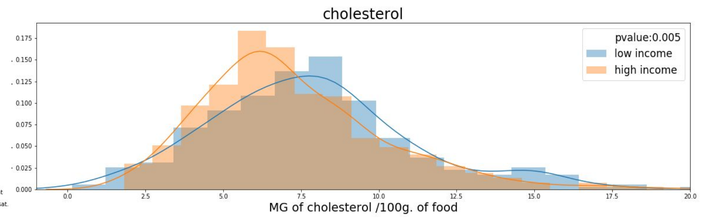


Health and Nutrition:

How do the clients food habits look like?



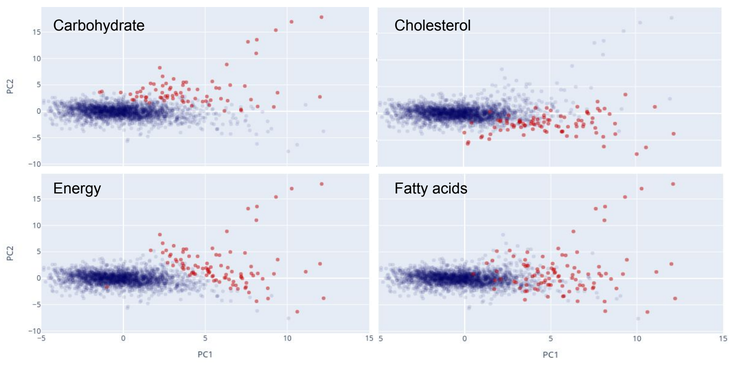
Consumption-explained correlations: nutrients show natural correlations amongst food articles (*center)*: however, we do not find the same correlations in the "average soup" people eat (*left)*. By subtracting the latter by the former, we obtain correlations that are not explained by the food itself.*(right)*



Cholesterol rate: Assuming normal distribution, low income households showed a higher average cholesterol consumption than wealthy ones under a t-test.

This shows that given its data, a supermarket possesses the required information to detect unhealthy diet behavior, opening the question of its responsability towards the consumer.

Since some households showed unhealthy habits, we developed a tool able to detect over-consumers as outliers for any nutrient.

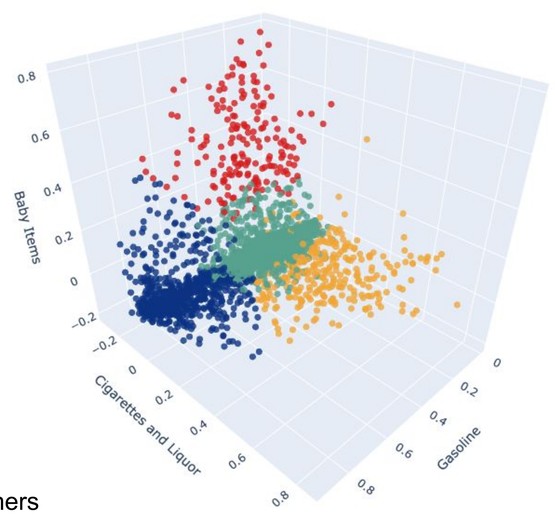


Unhealthy diet detection: the 4 plots highlight household average consumption respectively for carbohydrate(tl) cholesterol (tr) energy (bl) and fatty-acids (br). Excessive consumers are shown in red.

Household clustering:

What underlying groups can we discover among the households by clustering according to consumption habits?

We found that consumption of items in different subgroups explained a lot of variance and enabled us to get a very clear separation of these groups.

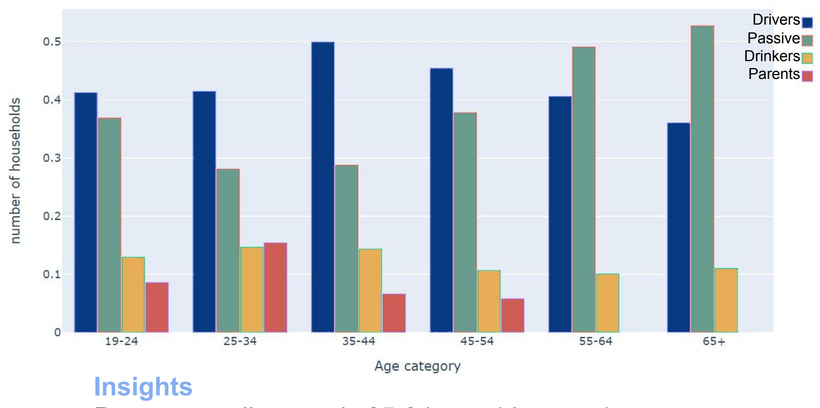


Yellow: Drinkers and smokers

Blue: Drivers

Green: Passive consumers

Red: Parents



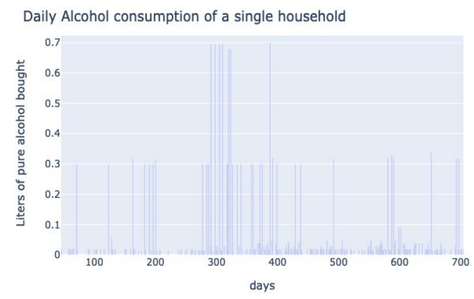
Insights

Parents: median age is 25-34, no older couples
Drivers: No visual trend with age groups
Passive: Seems to increase with age.This is expected since consumerism tends to decrease with age

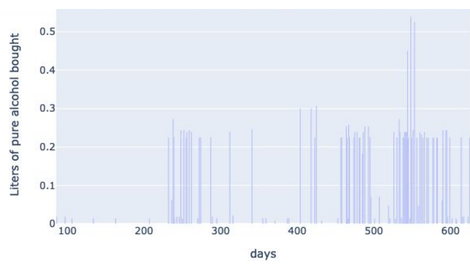
Aggregated transaction data can be used to group items into categories, and use these categories to cluster users. Clusters can be used to profile users and get their demographic information.

Discovery of consumption patterns suggestive of alcoholism were found among the households.

Top: Seems to struggle with alcohol consumption. The amount consumed varies a lot and switches between excessive consumption and periods of abstinence.



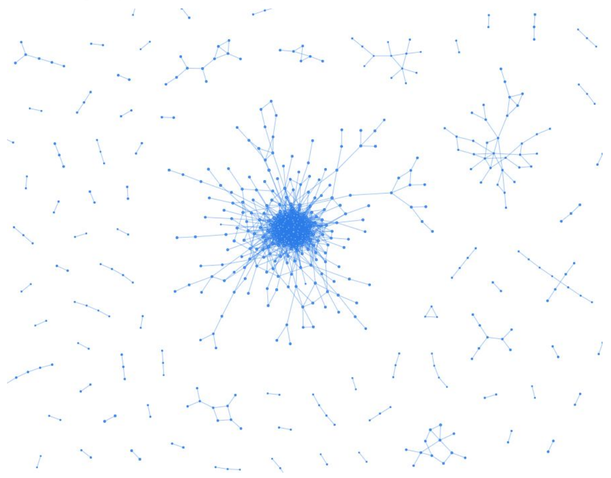
Bottom: Time series that represents more frequent consumption of alcohol in very small doses. May be associated with alcoholism.



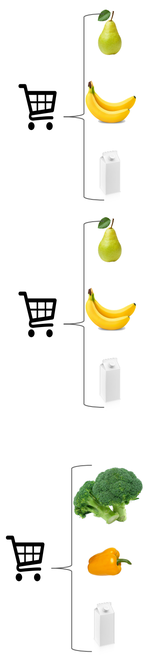
Network of items:

What do transactions tell us about the items themselves?

We looked at co-occurrences in shopping carts and built an undirected graph illustrating relations between items.



Every edge represents items that were bought the most together relative to their overall popularity.
There is a strongly connected component in the middle containing popular food items and some connected components with very interesting patterns.



Implementation Co-occurrence Matrix

We sum up all the times pairs of items appear together.

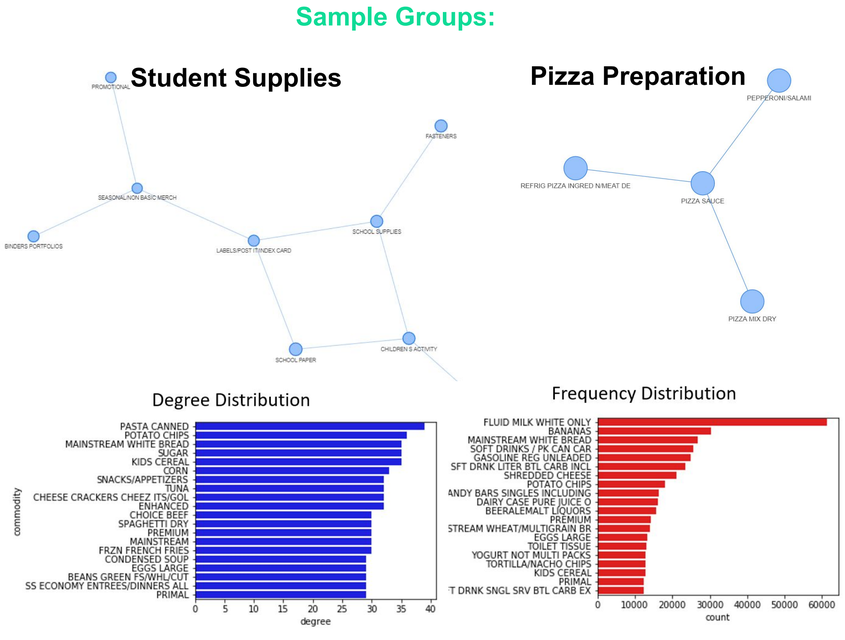
We normalize the co-occurrence by the frequency of the most popular item of the pair.

We represent an edge if score > threshold

Item Frequency	Banana	Pear	MIlk
Banana	2	1	2/3
Pear	1	2	2/3
Milk	2/3	2/3	3

Edges between items give insight about their relations. These relations are sometimes trivial, but sometimes they reflect particularities of the community and its culture.

Sample Groups:



The degree distribution shows that the most central items are not very healthy items such as snacks and canned foods.
This indicates that contrary to popular items such as milk or bananas, central items are bought together often relative to their individual popularity.

Transaction data can be considered as a form of fingerprint because it reveals a great deal of information about people. We can study nutrition patterns, consumption patterns and items relations. With this toy dataset, we open your eyes to how much big companies probably know about you, and their potential to do good or evil with this knowledge.