Nutriperso Project
Spring and fall 2018

Under the supervision of profs
M. Sebag and P. Caillou

CNRS – INRA – U. Paris-Sud - INRIA

Introduction

Introduction – Description of the data

KANTAR WORLDPANEL

Who are the **big con-sumers** of my brand?

Who are my **biggest competitors** on this segment?

How is the **distribution** of my products among my consumers?

1

Household id	Age Father	Age Mother
1	40	NA

x 25K

Socio-demographic information of the members of the household, such as age, education, etc.

2

Purchase id	Product id	Date purchase
1	12	1/1/2014

x 10M

Purchase data, each line corresponds to a purchase of a product by a household.

3

Product id	Product type	Brand
1	Boisson	Nestle

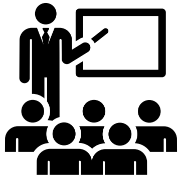
x 170K

Information about products, such as brand, packaging, organic, etc.

Introduction – Goals of the study

1

What links can be drawn between **socio-demographics** and **eating habits**? What influences the way we eat? Are there **segments of the French population** who have significantly **worse diets**?



2

Can we infer **relationships between diets and health** – as measured by the BMI? Are there **nefarious** products, or on the contrary **beneficial** ones?

3

Is it possible to **act on people's diets**? By building a sketch of a recommender system that would **replace** nefarious products with **healthier** ones while respecting people's **taste** and **budgets?**

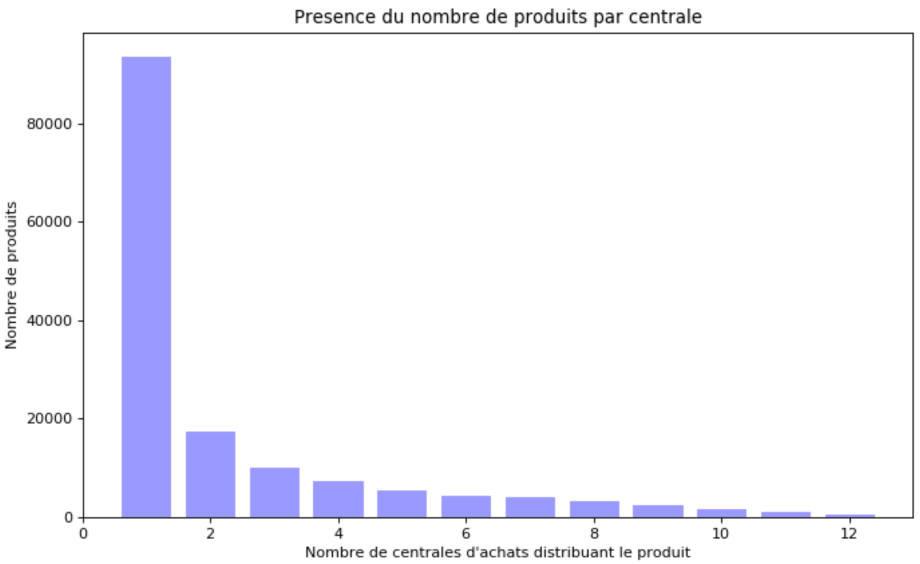
Introduction – Household Preprocessing

Ages mbr foyers	***Famille***	***Geo/ habitation***	***équipement ménager***	***animaux compagnie***	***cat socio p***	***equipement info***	***résidence secondaire***	***phys***
Nbr enfants – 3ans (en3)	famille recomposée (fare)	aire urbaine (aiur)	Lave-vaisselle (lvai)	Nbr de chats (cha)	Classe socio-économique ocde (scla)	nombre d'ordinateur fixes et portables (mor)	arbres fruitiers résidence principale (fru1)	poids de l'individu I du foyer (ipdsi)
Nbr enfants -6 ans (en6)	nombre de personnes au foyer (nf)	Département (dpts)	Lave-linge indépendant (malt)	Nbr de chiens (chie)	catégorie socio-profesionnelle individu i (cspc)	nombre d'inidividus possesseurs de téléphone portable (tlpo)	disposition d'une résidence secondaire (rs1)	Taille de l'individu I du foyer (ihaui)
Nbr enfants -15 ans (en15)		Type d'habitation (thab)			niveau d'étude individu i (etuc)	nombre de téléviseurs (tvc1)		
Nbr enfants -25 ans (en25)		statut d'occupation du logement principal (socc)			activité profesionnelle individu i (itra)	nombre de voitures (voit)		
age du chef de foyer (agec)					revenu mensuel brut du foyer (rve)			
age du panelliste (agep)								

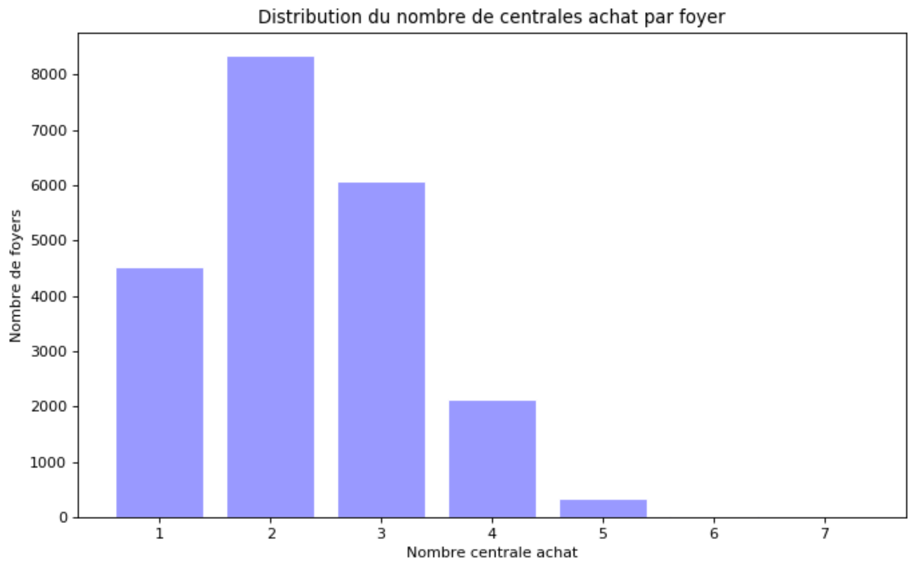
Above are the **variables** we have **kept for our analysis**.

Thereare many variables we would like to have but **do not have (sports, smoking),** as such these are confounders that we cannot control for and that **add a caveat on our results**.

Introduction – Product Preprocessing (1/2)



Most **products** are only sold by **one distributor**



Most **households** only go to **a small amount of distributor** during the year

People are made artificially to **live in different dietary universes** just by virtue of the **distributor** they go to.

Furthermore, **Marketing** also makes **artificial distinctions between products** that are irrelevant from a nutritional perspective.

These are factors we need to control for in our preprocessing. We need to **discard these artificially introduced differences**.

Introduction – Product Preprocessing (2/2)

Our aim is to **reduce the space of products** from **170,000 actually** to a lot less.
Not only for easier handling by Machine Learning Algorithms (curse of dimensionality) but also for consistency among the data.

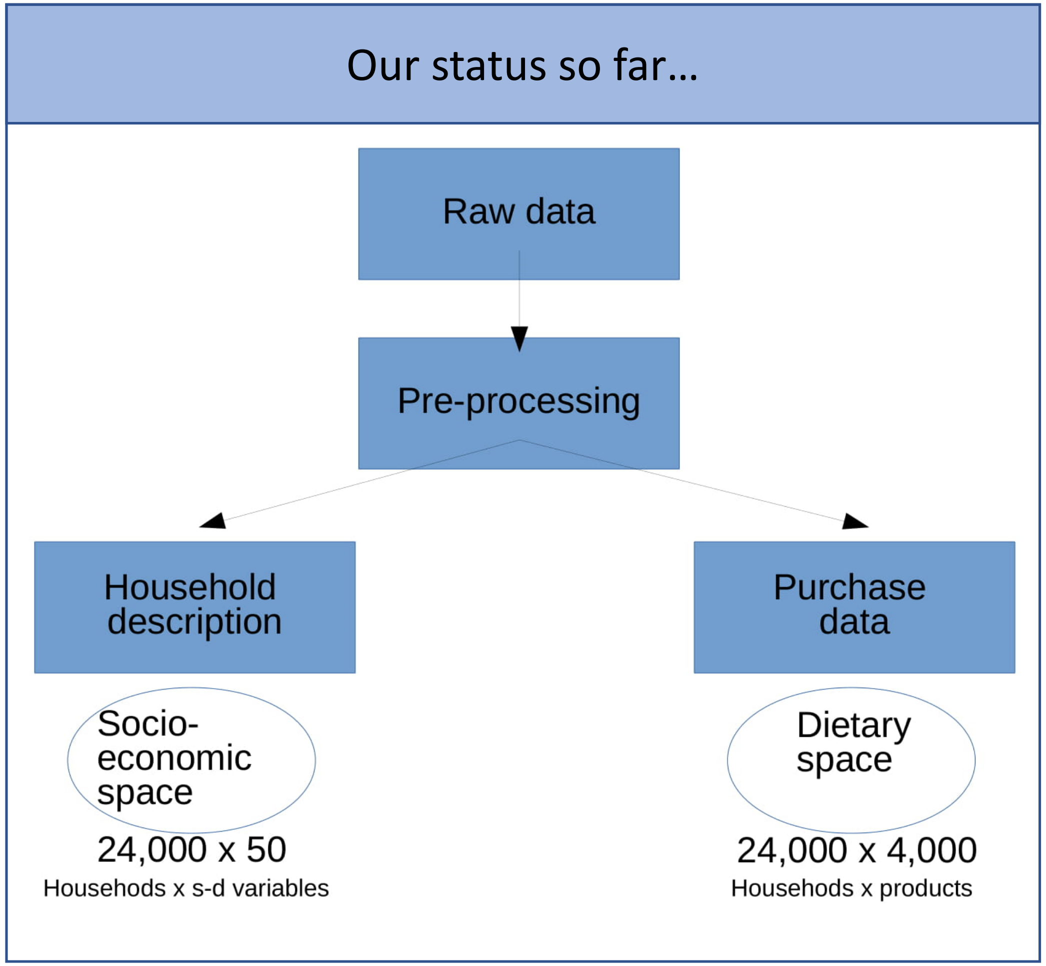
Products are **organized by categories and subcategories**.
Products from **each subcategory share** the **same** set of **features**.
It is possible therefore to take profit of this structure to make a relevant clustering.

Our clustering technique is straightforward, for each of the circa 200 subcategories, **select manually** a small **subset of features that are of importance**, and **cluster** the products along these features.

For instance, let's take the **example of beer**. We only selected as features of interest **colour** (blonde, black, etc.) and **has-alcohol**(yes, no), and **organic** – yielding a total of circa **10 categories,** i.e. all present combinations of the features above.
With this restrictive technique we achieved a reduction to only **4130 products**.

Phase I: Relationship between eating habits and social condition

Phase I – What do we want to do



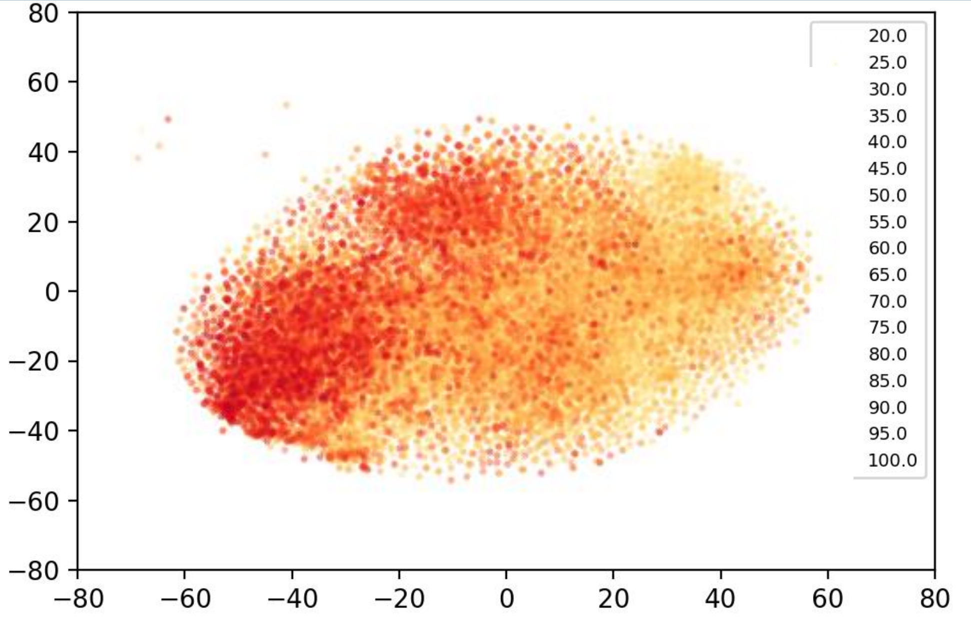
Each **household** is represented as **2 distinct points**:
1 point in the **dietary space**
1 point in the **socio-demographic space**
-> How to best **link both spaces**?

Subsequent question:
Can we **identify subsets of the French** population with **worse / better diets**?

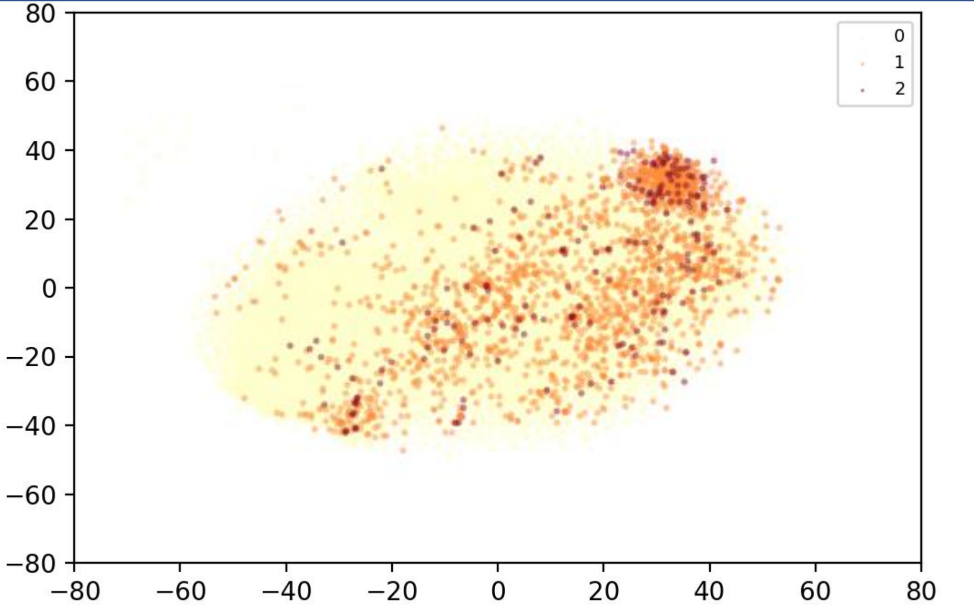
Phase I – The LSA model

Our aim is to **find a low-dimensional representation of our dietary space** by using the SVD decomposition of the purchase matrix
We shall then project this representation **into a 2-D plot** and **color points by socio-demographic** variables.

Households in the dietary space colored by age



Households in the dietary space colored by number of babies



People with **babies** cluster together because of the **special diet babies** require.
Age is another socio-demographic that **intuitively divides the dietary** space.

Phase I – The LDA model (1/2)

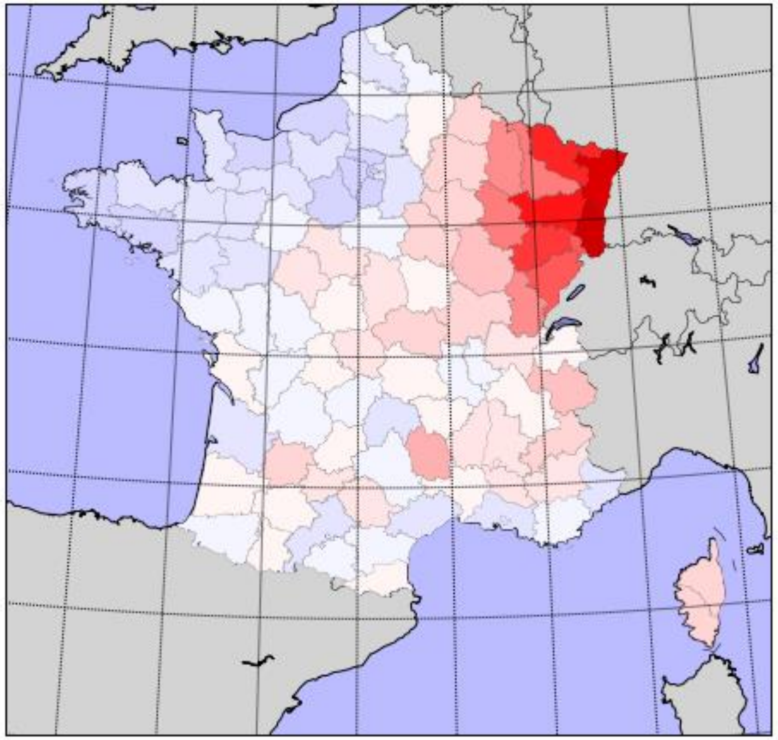
The LDA (Blei, 2003) model is a **mixture model**, that models a household as a **mix of a small number of individual tastes and preferences** – topics.
Here, we add the idea that a household is **not a monolithic entity** but is often composed of **many different needs and tastes**.

Bébé	Dessert frais	Alcool
Alimentation BB : 0.31 | 0.99 | 100.35	CAZAUBON : 0.08 | 0.93 | 48.88	Vins : 0.28 | 0.59 | 24.98
Dessert BB : 0.21 | 0.98 | 98.47	DANONE.ACTIVIA : 0.07 | 0.4 | 21.03	Bière : 0.11 | 0.41 | 17.53
Petit Déjeuner BB : 0.04 | 0.96 | 96.75	LES JACQUINS : 0.07 | 0.47 | 24.78	Apéritif : 0.07 | 0.51 | 21.84
Biscuit BB : 0.01 | 0.95 | 95.4	DANONE.DANETTE LE LIEGEOIS : 0.04 | 0.32 | 16.91	Whisky-Bourbon : 0.05 | 0.67 | 28.66
Farine BB : 0.03 | 0.93 | 93.99	DANONE.DANIO : 0.04 | 0.72 | 37.74	Brsa : 0.03 | 0.02 | 0.73
Boisson BB : 0.08 | 0.93 | 93.54	PETIT BASQUE LE : 0.03 | 0.21 | 10.99	Mousseux/Pétillants : 0.02 | 0.44 | 18.89
Dessert en conserve : 0.03 | 0.04 | 4.31	DELISSE : 0.03 | 0.08 | 4.06	Biscuit Apéritif : 0.02 | 0.04 | 1.61

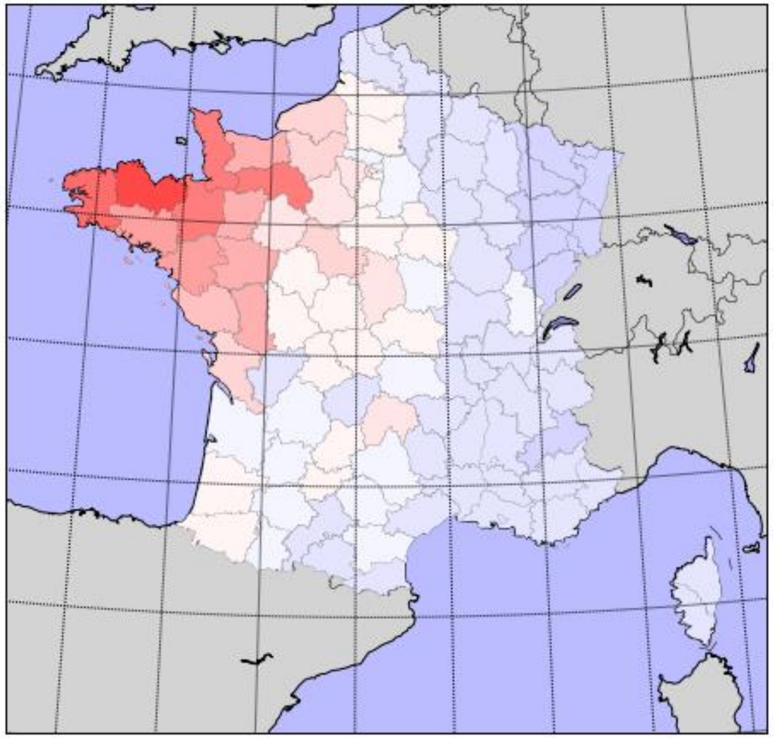
These topics from the LDA (i.e. most frequent products) are based on **individual needs of households**. Relevant households each have a need for **baby food**, or **alcohol**, etc. These topics allow to uncover **interesting aspects of people's** life that interact interestingly with socio-demographic variables.

Phase I – The LDA model (2/2)

Geographic distribution of topic 16 (Gelée, Fécule, Aide Pâtisserie, Chapelure, Farine)



Geographic distribution of topic 2 (Fromage camenbert, Beurre, Cidre, Rillettes)



Here topics have a heavy **regional orientation** – linked with regional specific diets.

These diets can appear **because we have taken into account other tastes – needs** (water, alcohol) that would otherwise **dilute such information**

Phase I – Regression of BMI on Demographics

The next natural question we ask is **how much of a person's BMI** - health can we **explain through demographics**?
We did a **regression** to answer this question – and although performance was low (6% explained variance, 19.8 MSE) and analysis of **relevant coefficients** provides **crucial insights**.

Negative Coefficients – Lower BMI

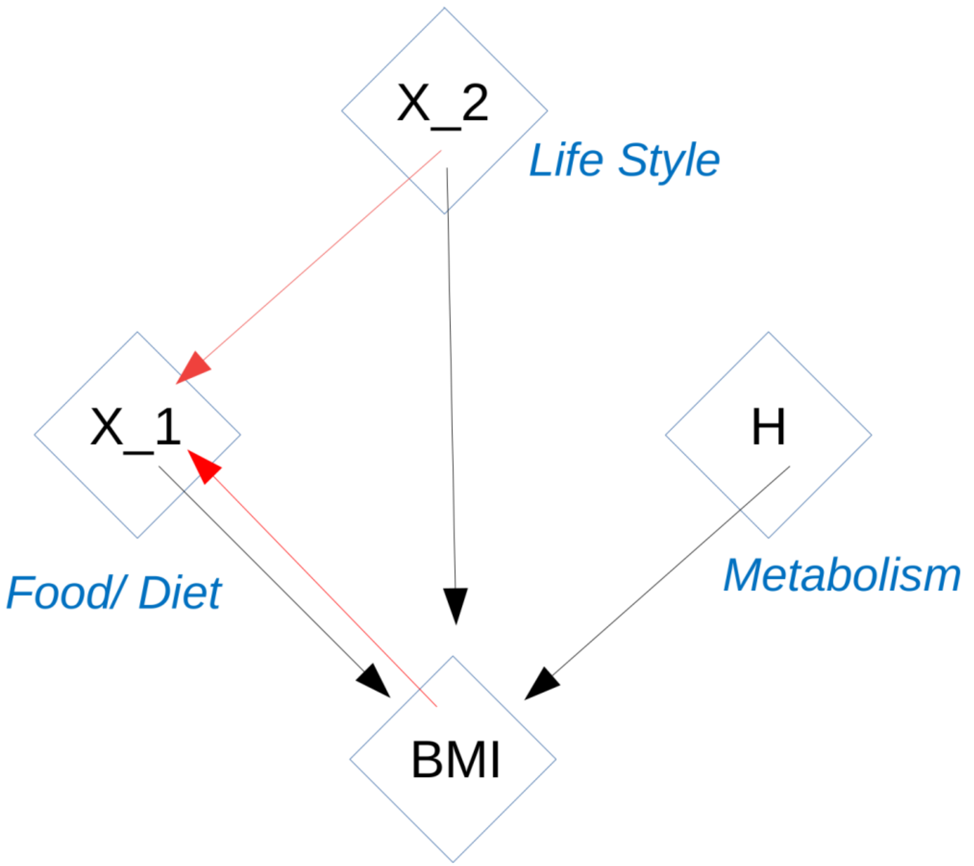
variable	coefficient	std	T-stat	low	high
rve	-0.0981	0.021	-4.638	-0.139	-0.057
proprio	-0.4219	0.107	-3.941	-0.632	-0.212
etude_5.0	-0.751	0.143	-5.245	-1.032	-0.47
etude_4.0	-0.762	0.134	-5.69	-1.025	-0.5
etude_6.0	-0.9506	0.167	-5.676	-1.279	-0.622
dpts_44	-1.0705	0.28	-3.828	-1.619	-0.522
dpts_56	-1.1218	0.312	-3.601	-1.732	-0.511
etude_7.0	-1.151	0.15	-7.675	-1.445	-0.857
dpts_81	-1.3529	0.355	-3.809	-2.049	-0.657
etude_8.0	-1.3655	0.168	-8.116	-1.695	-1.036

Positive Coefficients – Increase BMI

variable	coefficient	std	T-stat	low	high
csp_Personnessansactivitéprofessionnelle	3.2821	0.134	24.456	3.019	3.545
csp_Anciensemployésetouvriers	2.9923	0.132	22.663	2.734	3.251
csp_Ancienscadresetprofessionsinterm	2.9836	0.142	21.023	2.705	3.262
csp_Anciensartisans	2.9056	0.234	12.399	2.446	3.365
csp_Anciensagriculteursexploitants	2.1129	0.428	4.937	1.274	2.952
csp_Chômeursn'ayantjamaistravaillé	1.8122	0.446	4.067	0.939	2.686
csp_Chauffeurs	1.3864	0.336	4.132	0.729	2.044
csp_Elèves	1.107	0.218	5.076	0.68	1.534
en6	0.4043	0.076	5.329	0.256	0.553
aiur_espacedominanturale	0.3363	0.074	4.528	0.191	0.482
tvc1	0.2784	0.024	11.481	0.231	0.326
chie	0.1556	0.031	4.959	0.094	0.217
cha	0.1104	0.023	4.825	0.066	0.155
age	0.0462	0.003	17.331	0.041	0.051

Phase II: Relationship between eating habits and BMI

Phase II – How much of BMI can we explain?



Many things determine our BMI, and the **interactions** between these factors are quite **complex**.
Our main task here will be to **disambiguate** the **two red arrows**, i.e. in plain terms:
1) **Confounders** : other underlying cause in the effect seen: e.g. Sports => Sport Drink => BMI
2)**Inverse causality** : mistaking effects for causes e.g. products for weight loss induced by high BMI

Phase II – BMI regression of socio-dietary variables… Results

features	add socio	preprocessing	predict_worse	extreme_only	threshold	var_expl train	mse_train	var_expl test	mse test	examples
socio-eco			False	5	300	0.0663	19.40	0.0591	19.54	30383
socio-eco			True	4.6	300	0.0987	20.17	0.0890	20.39	19658
raw	False	normalize	False	5	3000	0.2434	16.43	0.1117	18.47	30383
raw	False	binarize	True	4.6	3000	0.3229	16.48	0.1357	19.42	19658
raw	True	normalize	False	5	3000	0.2612	15.97	0.1344	18.01	30383
raw	True	normalize	True	4.6	3000	0.3130	16.43	0.1641	18.84	19658

How to **represent purchases of products**? Binary? Raw quantities? Proportional total consumption?

Each households 2 means **2 people**, we simplify the task by only choosing **worse BMI**

In order to cleanse the dataset of household with **not enough data**, we use an **entropy threshold**

The **metric** we are interested in (explained variance)

The **low accuracy** can be **explained** by the **following factors**:
- **Data quality**: people eat out, do not fill in the survey, etc.
- **Absence** of important **confounders** (sports, smoking, etc.)

Still we can infer meaningful results, especially from the coefficients of the regression.

Phase II – BMI regression of socio-dietary variables… Coefficient Analysis

Negative Coefficients – Lower BMI

groupe	sousgroupe	marque	count	coefficient
Plat	Plat Frais	SOJASUN	1576	-111
Boisson	CafÃ©	PLANTATION	500	-74
Plat	Plat SurgelÃ©	TANTE YVONNE	858	-73
Fruits et Legumes	LÃ©gumes Sec	NOTRE JARDIN	956	-72
Boisson	Champagne	Marque Non Trouvee	527	-69
Plat	Plat Frais	SOY	435	-69
Confit	Confit en conserve	LARNAUDIE	459	-67
Fruits et Legumes	Fruits Frais	Marque Non Trouvee	17545	-67
Aide Culinaire	Produit sucrant	CARREFOUR.DISCOUNT	4995	-65
Cereale A Cuire	CÃ©rÃ©ale A Cuire	BJORG	659	-65
Boisson	BiÃ¨re	LEFFE.RUBY	968	-64
Plat	Salades en conserve	PECHE OCEAN	557	-63
Fruits et Legumes	Legumes Frais	SAVEOL	3757	-61
Plat	Plat Frais	Marque Non Trouvee	907	-61
Boisson	CafÃ©	Nescafe special filtre	1054	-61
Boisson	Champagne	Marque Non Trouvee	1108	-56
Boisson	Infusion	Elephant nuit tranquille	1309	-56
Pain et viennoiserie	Panification SÃ¨che	BJORG	518	-56
Aide Culinaire	Lait De Coco	SUZI WAN	1580	-55
Boisson	Infusion	LEA NATURE.JARDIN BIO	754	-54
Fruits et Legumes	Legumes Frais	Marque Non Trouvee	14193	-54
Biscuits	Barre CÃ©rÃ©aliÃ¨re	BRIN DE JOUR	676	-53
Aide Culinaire	Vinaigre	ECO+	4625	-51

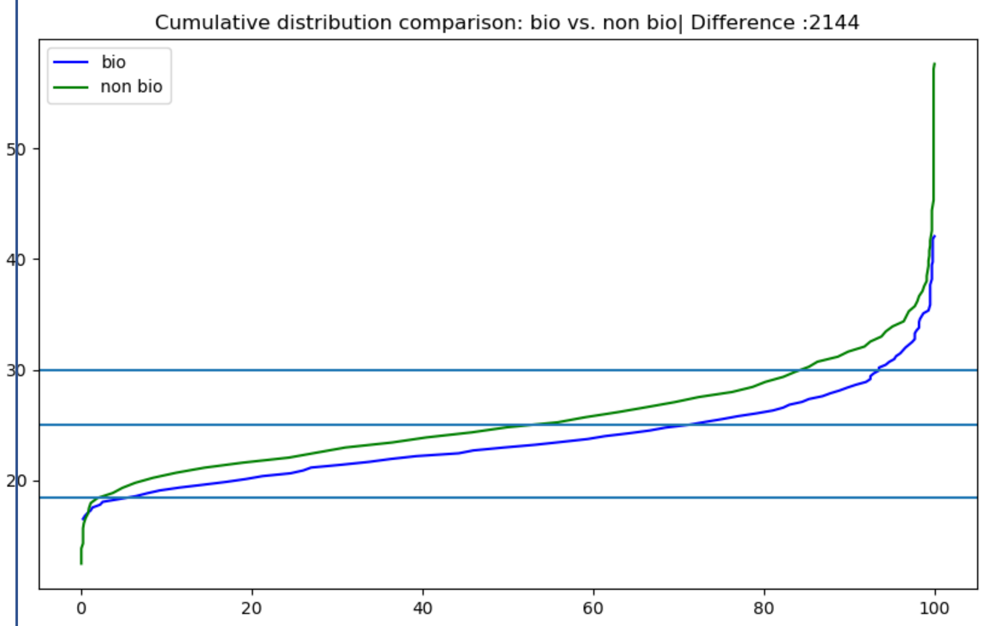
Positive Coefficients – Increase BMI

groupe	sousgroupe	marque	count	coefficient
Aide Culinaire	Produit sucrant	CANDEREL	614	112
Charcuterie	Autre charcuterie	Marque Non Trouvee	28202	96
Jambon	Jambon Blanc	U	1496	95
Viande	Porc	Marque Non Trouvee	1669	95
Charcuterie	Lot Mixte PÃ¢tÃ©	MONTEROY	577	87
Boisson	Sirop	PULCO	2598	87
Charcuterie	Fromage De TÃªte	RANOU MONIQUE	245	86
Assaisonnement	Poivre	CIGALOU	258	86
Sauces	Sauces	RUSTICA	262	83
Aide Culinaire	Huile	BOUTON D OR	346	82
Aide Culinaire	Margarine	ST HUBERT.OMEGA 3	7475	76
Plat	Plat Frais	FLEURY MICHON	815	75
Fromage	Fromage camembert	REO	972	73
Plat	Plat Frais	TANTE YVONNE	563	73
Boisson	Lait	NESTLE	959	73
Biscuits	Biscuit ApÃ©ritif	SUZI WAN	440	71
Viande	BÂ□uf	Marque Non Trouvee	4959	71
Pain et viennoiserie	Viennoiserie	PATIGEL	133	69
Charcuterie	Saucisse	Marque Non Trouvee	6021	69
Fruits et Legumes	Fruits et lÃ©gumes en Conserve	Marque Non Trouvee	274	68
Dessert	Dessert Frais	WEIGHT WATCHERS	1089	64
Boisson	Vins	Marque Non Trouvee	386	61

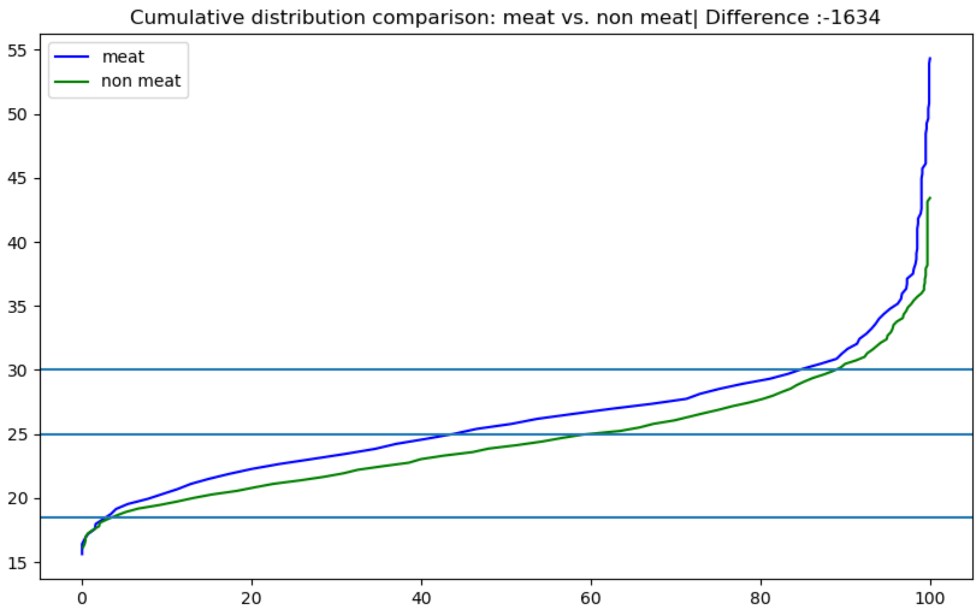
Here we obtain still some **confounding issues** or **inverse causality** (Negative: Champagne, Beer, etc. | Positive: Lean yoghurt)
We would need to go deeper into the analysis to disambiguate these effects.

Phase II – Identifying significant clusters of products

It is **hard to visualize effects** of products at the **individual** level. Is it possible to cluster further the products in a meaningful way and see the effects of such **macro-clusters** on BMI?
Here, we found **2 macro-clusters of interest** – organic food and meat with significant influence on BMI.



Cumulative BMI distribution difference between organic food eaters and general population



Cumulative BMI distribution difference between heavy meat eaters and general population

Phase II – Limits and further directions

1

Confounding problem

Here we did not solve the confounding problem. We know that we do not have important confounding variables that can explain away all the effects that we attributed to products. Nevertheless, other studies seem to indicate that our results are robust to such kind of confounders.

2

Increasing the data would bring about more robust and insightful results

Our results would be more robust and more instructive if we were given more data. The performance of both the regression, and the impact of cluster of products on health as well. Moreover, more data would help put in place longitudinal analyses, where we would try to see the impact of age or life events on BMI.

3

Analyse causal direction disambiguation

We did not explore the possibility here of disambiguating the direction of the arrows of causality. Do fat-losing products actually cause an increase in BMI? Or the other way around? One problem we encountered is the lack of sufficient data – not enough people in our sample consumed these products. Nevertheless, provided more time and data, this would be an interesting path to explore.

Phase III: Building a recommender system for better eating

Phase III – Sketch of a recommender system
Finding similar products through GloVe(Pennington 2014)

“ **You shall know** a **word** by the **company** it **keeps** (Firth, J. R. 1957:11) ”

1. I enjoy flying.
2. I like NLP.
3. I like deep learning.

The resulting counts matrix will then be:

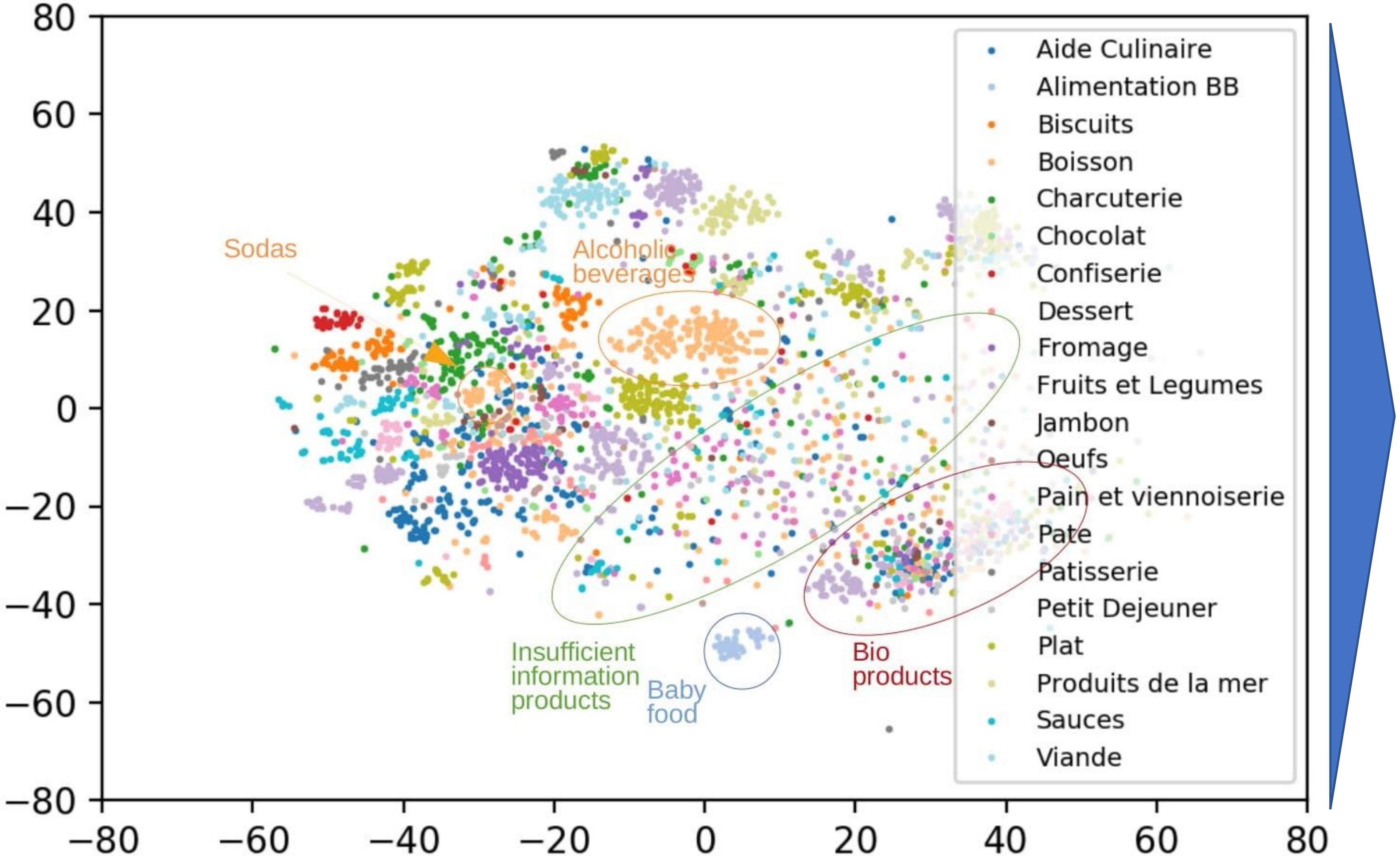
$$
X = \begin{array}{cc}
 & \begin{array}{cccccccc} I & like & enjoy & deep & learning & NLP & flying & . \end{array} \\
\begin{array}{c} I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} &
\begin{bmatrix}
0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

Gloves tries to approximate in a **low-dimensional dense** space this **matrix of co-occurrences**.

The idea is that **words which co-occur** with the same other words (**share the same context**) will be forced to be **close together** in the low-dimensional space.

Glove is especially interesting since it allows us to work on a **finer granularity** – each sentence will be a **basket** in our setting (before we were working at the yearly aggregate level)

Phase III – Results of GloVe(Pennington 2014)



-> The **results** here are broadly **respectful of categories** and subcategories – although **no such information was encoded** in the model.

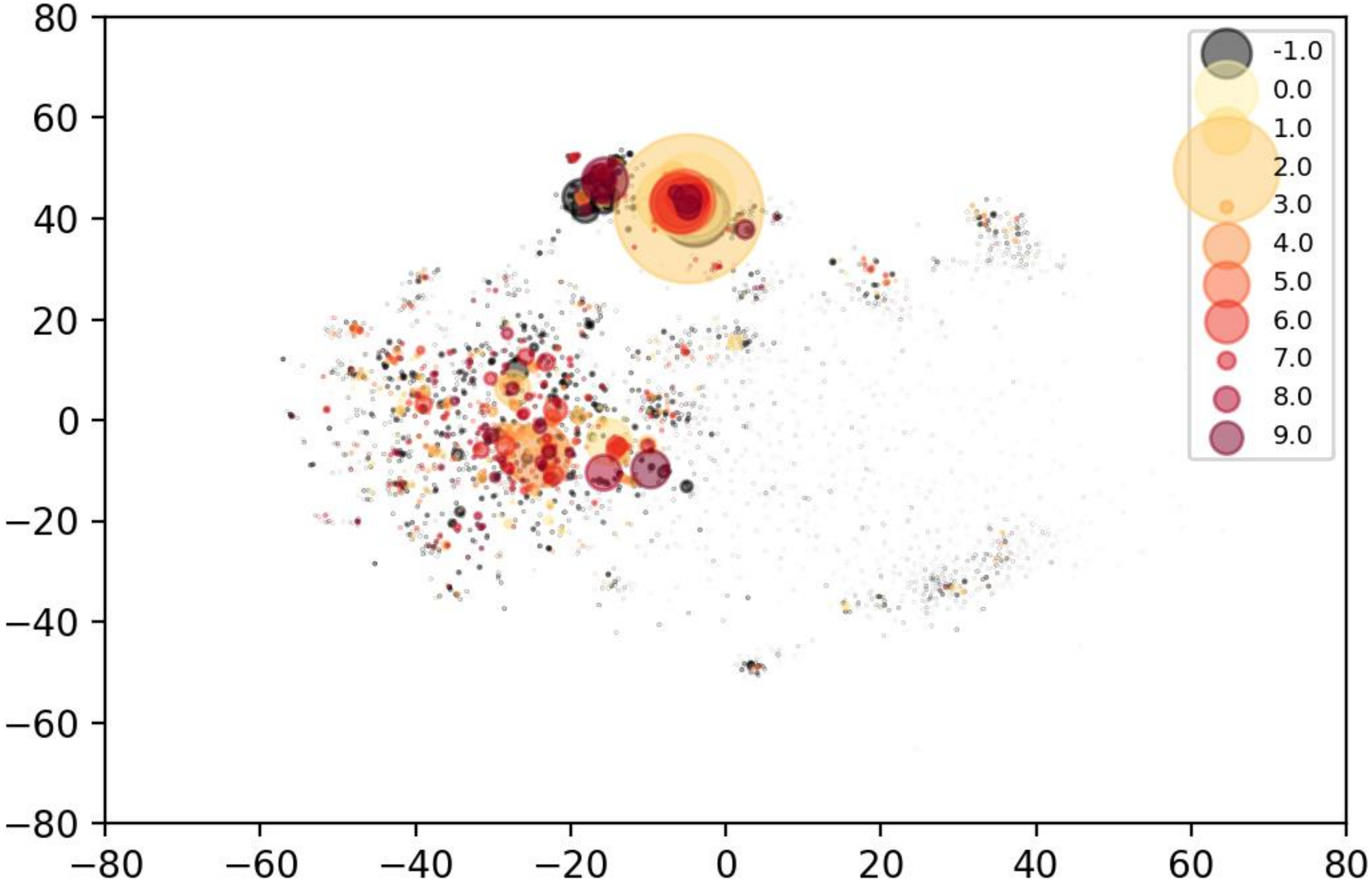
-> It still **lacks** a **meaningful representation for a lot of products.**

-> Nevertheless this can be taken as a **crude approximation** of **product meaning**, and was shown to give meaningful clusters of products (cf. organic and meat clusters)

Phase III – Combining GloVe and Regression

Colour indicates impact of the product **on BMI** as outputted by the regression (black means a null coefficient).
Size here represents the **volume** of the products in terms of total sales.
This is still quite approximate, but we can see from the overlap of many circles that a **recommender system** is quite possible.

It would go as follows:
- **Substitute** a red product with a yellow one
- **Ensure** the two products are **reasonably close** and similar in product category



Phase II – Limits and further directions

1

Increasing the data would bring about more robust and insightful results

Again, we have a lot to gain from increasing the amont of available data. The Glove analysis has yielded non-informative results for about 1/3 of our products. This is directly related to the sparseness of the purchase matrix, a sparseness that can be resolved with an increasing amont of data. Moreover other products would also gain a more robust low-dimensional representation from this increase in data.

2

Glove representation highly dependent on the clustering of products

One problem we encountered is that our first GloVe analysis heavily reproduced distributor universes, i.e. products were close together because they were sold by the same distributor. This is a by-product we tried to avoid by trying to erase all differences among distributors or brands in our pre-processing, but as such we may very well lose information by increasing the granularity of our analysis. We are not convinced we have found the right trade-off between gains in generalization and accuracy of the model.

3

Gaussian LDA : an interesting path to explore

Gaussian LDA mixes the insights of the LDA model with the new potent representation that an embedding such as GloVe gives. This refined LDA has been shown to give very robust results, in a wide array of situation. Our initial LDA suffered precisely from generalization issues, meaning we were stuck at a relatively limited number of topics – mainly because of data sparseness. We think this might alleviate this issue and provide us with both more topics and more insightful ones.

Conclusion and further directions

Conclusion – Results of the study

1

We first explored the **relationship between eating habits and social condition**.
We were able to find that **age**, and also **geographical location** were heavy impactors of the way we eat.
As for what impacts the BMI, **wealth, education and age** were found to play a major role among other variables.

2

Then we tried to see if we could find a **relationship between food eaten and BMI**.
We found that when combining socio-demographics and dietary information we could **explain 16% of BMI**.
Moreover, certain **broad cluster of products (organic, meat)** were found to have a **heavy impact** on BMI, even when controlling for socio-demographic variables.

3

We investigated the feasibility of a **recommender system**. We mainly focused of finding a **representation** of products that would take heed of **relationship products have among each other**.
The **GloVe algorithm** provided a framework for this, a framework that can be improved by increasing the amount of data used. Thus now we know which products are close to one another, and thus **potential substitutes** for one another.

Conclusion – Further directions and new data

Directions interesting to explore

1) The first direction that needs to be explored is to ask whether the **clustering of products** we realized in the **preprocessing is relevant** and has not deleted relevant information. Also, it is necessary to have a way to evaluate the resulting clustering.

2) A second analysis that has only been touched upon is **diet disambiguation**: who in the household eats what? We have only hinted at a possible solution: **domain adaptation**, in effect taking advantage both of consumption patterns and of BMI. This has yielded quite interesting results (e.g. women are more ‘affected’ by baby food than men), but we did not have the time to explore this direction further and our BMI regression indicated that there were low returns to expect from such a model.

Data it would be nice to have

1) **BMI evolution over time**: the idea would be to work on deltas of BMI instead of raw BMI, this would help alleviate the effect of many confounders.

2) **Nutritional information for products**: this would help us with the difficult task of clustering products together, and would make possible a host of analysis based on nutriment intake.

3) **More of the same data**: this point can very easily help improve significantly the scope of the previous analyses. Machine learning algorithms require a lot of data to detect patterns, and as such will always gain from an increase in available data. Especially since for the models we implemented were run on a local machine, and as such computational power is not an issue yet.

Thank you for your attention.

Any questions?