Internship 2018

# Food, Health, Wealth: Profiling and Causal Modelling

SUPERVISORS: Michèle Sebag, Philippe Caillou, Olivier Allais
LABS: TAO − CNRS − INRIA − Paris-Sud; ALISS− INRA; U. Paris-Saclay
MAIL: sebag@lri.fr, caillou@lri.fr, olivier.allais@inra.fr
URL: `http://www.lri.fr/~sebag/Stages/Stage_Nutri_18.pdf`
ABSTRACT :
Le contexte général est celui du projet Nutriperso (CNRS - INRIA - INRA - INSERM - CEA), cherchant à identifier les relations entre l'alimentation, la santé et les indicateurs socio-démographiques (statut familial, diplomes, niveau de vie). Les données Kantar décrivent 20,000 familles: leur données socio-démographiques et l'ensemble des produits alimentaires achetés sur 20 ans (fréquence journalière). L'une des difficultés est de prendre en compte la nomenclature alimentaire au lieu de se ramener à quelques dizaines de catégories de produits ; la raison en est que la composition détaillée des aliments ("pizza ultra-processée Margherita ZX318" et non "pizza") importe. Le défi consiste ainsi à traiter du big data, et du fat data ($p >> n$), pour faire progresser les connaissances dans le domaine de l'alimentation, à l'intersection de l'apprentissage statistique, de la sociologie, de la biologie, de l'économie, et de la nutrition.

**Objectifs du stage**
Le stage a deux objectifs: i) déterminer une représentation de dimension raisonnable (quelques centaines) et interprétable (par un expert) des consommations alimentaires; ii) valider cette représentation dans le contexte de Nutriperso, dont les finalités à long terme concernent entre autres: le profilage des ménages en terme socio-alimentaire; le lien entre profil alimentaire et santé; l'impact des évènements de la vie (naissance d'un enfant, période de chomage, retraite) sur le profil alimentaire. Ce stage peut déboucher sur une thèse.

**Approche proposée**
On pourra dans un premier temps définir un exemple comme le détail de la consommation consolidée hebdomadaire d'un ménage (ticket de caisse ou liste des achats au marché). L'approche de réduction de dimensionalité proposée s'inspire du traitement des langues naturelles: on considèrera un produit comme un mot; une consommation consolidée comme un document. Dans une première phase, on pourra alors procéder à une réduction de dimensionalité linéaire (SVD) ou non linéaire (bag-of-words to vec). Il faudra faire preuve de finesse en raison des difficultés spécifiques. Ainsi, il faut prendre en compte la non-stationarité des données: on sait que les mots et leur sens évoluent dans la langue ; mais les produits d'alimentation et leur nomenclature évoluent encore plus vite.

Un des aspects essentiels du stage concerne la proposition, et la validation avec l'équipe pluridisciplinaire de Nutriperso, de critères de validation pour la représentation proposée. Par exemple:

- Faible variance des consommations projetées pour un ménage et une saison donnée;
- Cohérence de la projection entre saisons (en moyenne, deux ménages de consommations proches en hiver sont aussi proches par leur consommation en été).
- Corrélation des "sauts" en termes de profils de consommation avec des changements de vie (naissance d'un enfant, divorce, retraite).
- Clustering dual des profils alimentaires et des profils socio-démographiques d'une part, et des distributions d'indices de masse corporelle d'autre part.

Ce stage demande de très bonnes capacités statistiques et informatiques (programmation C++ ou Python, en partant de librairies existantes), ainsi que du goût pour la pluridisciplinarité (maths-informatique, biologie, économie, sociologie).