FACULTÉ DE
GESTION,
ÉCONOMIE
& SCIENCES

Science des données - DataScience

2020-2021

L3 SdN et Matière d'Ouverture

Julie Jacques – Enseignant chercheur en informatique

julie.jacques@univ-catholille.fr

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Déroulement

- 9 séances de 2h
- Fonctionnement
  - 2h par semaine
  - 12h de cours et d'exercices pratiques, sur papier et sur le logiciel Orange de Biolab
  - 6h de projet
    - sur Orange pour les non informaticiens (approche NoCode)
    - en Python pour les informaticiens.
- Evaluation
  - 40% : projet
  - 60% : Examen sur papier (1h)

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

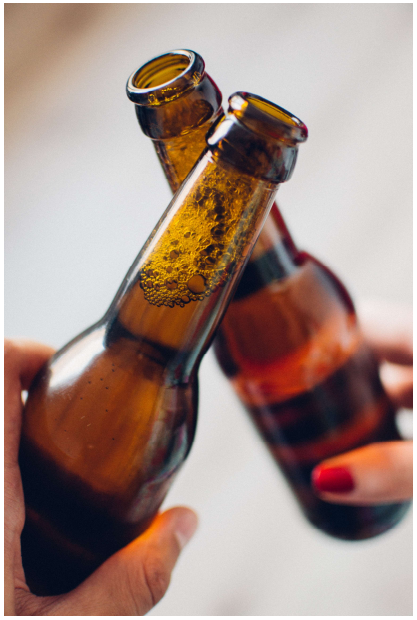
Plan du cours

- Intro science des données
- Algorithme K-NN
  - Notion de distance
  - Évaluation d'un algorithme de prédiction
  - Algorithme K-NN
  - Mini-projet : « vénéneux ou comestible ? »
- Sélection d'attributs
  - Projet partie 2 : nombre minimum de questions
- A priori et panier de la ménagère

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Vous apprendrez aussi….

Le rapport entre les couches et…



… la bière ???

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Qu'est-ce que le Machine Learning ?

Processus inductif, *itératif* et *interactif* de découverte dans les BD larges de modèles de données *valides*, *nouveaux*, *utiles* et *compréhensibles*.

- Itératif : nécessite plusieurs passes
- Interactif : l'utilisateur est dans la boucle du processus
- Valides : valables dans le futur
- Nouveaux : non prévisibles
- Utiles : permettent à l'utilisateur de prendre des décisions
- Compréhensibles : présentation simple

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Notion d'induction [Peirce 1903]

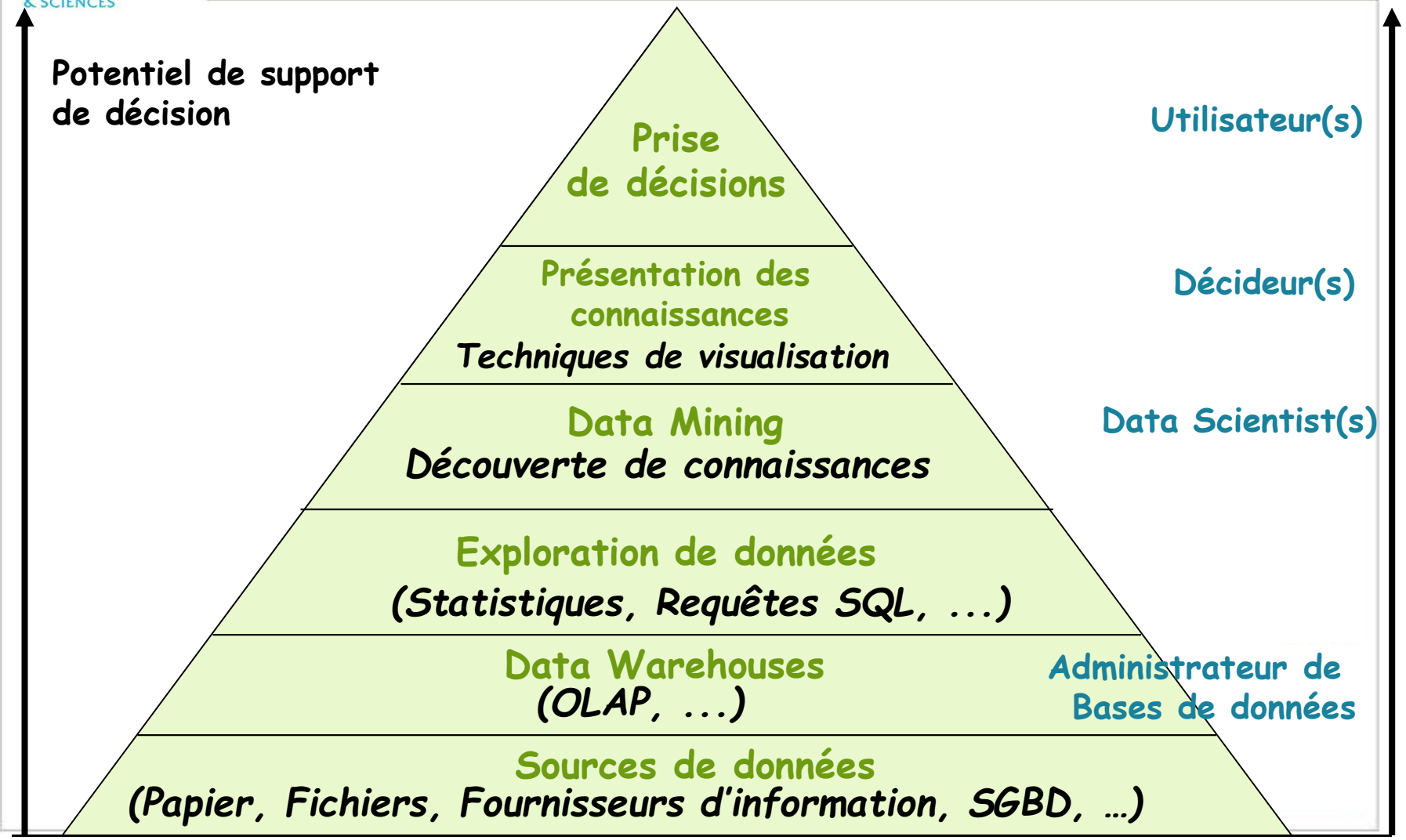
Induction : Généralisation d'une observation ou d'un raisonnement établis à partir de cas singuliers.

Utilisée en Machine Learning (tirer une conclusion à partir d'une série de faits, pas sûr à 100%)

- La clio a 4 roues, La Peugeot 106 a 4 roues, La BMW M3 a 4 roues, La Mercedes 190 a 4 roues
- ==> Toutes les voitures ont 4 roues

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Data Mining et aide à la décision



FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Un peu de vocabulaire

Datamining, Machine Learning, IA, Statistiques, Deep Learning : c'est quoi la différence ?

Intelligence artificielle (IA)

« L'intelligence artificielle est un ensemble de techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux humains et à certains animaux. » (Yann Lecun)

Statistiques

« Domaine mathématique qui consiste à recueillir, traiter et interpréter un ensemble de données » (Wikipedia)

Souvent, il y a des pré-requis sur les données, qui doivent avoir des caractéristiques particulières (tests paramétriques)

Machine Learning (Apprentissage automatique)

« Champ d'étude de l'intelligence artificielle qui se fonde sur des approches statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. » (Wikipedia)

Datamining (Fouille de données)

« La fouille de données a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique » (Wikipedia)

Deep Learning (Réseaux de neurones profonds)

« L'apprentissage profond (plus précisément « apprentissage approfondi ») est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaire» (Wikipedia)

Algorithme

Ensemble d'opérations ordonné et fini devant être suivi dans l'ordre pour résoudre un problème

Système Expert

Un système expert est un outil capable de reproduire les mécanismes cognitifs d'un expert, dans un domaine particulier. Il s'agit de l'une des voies tentant d'aboutir à l'intelligence artificielle.

FACULTÉ DE
GESTION,
ÉCONOMIE
& SCIENCES

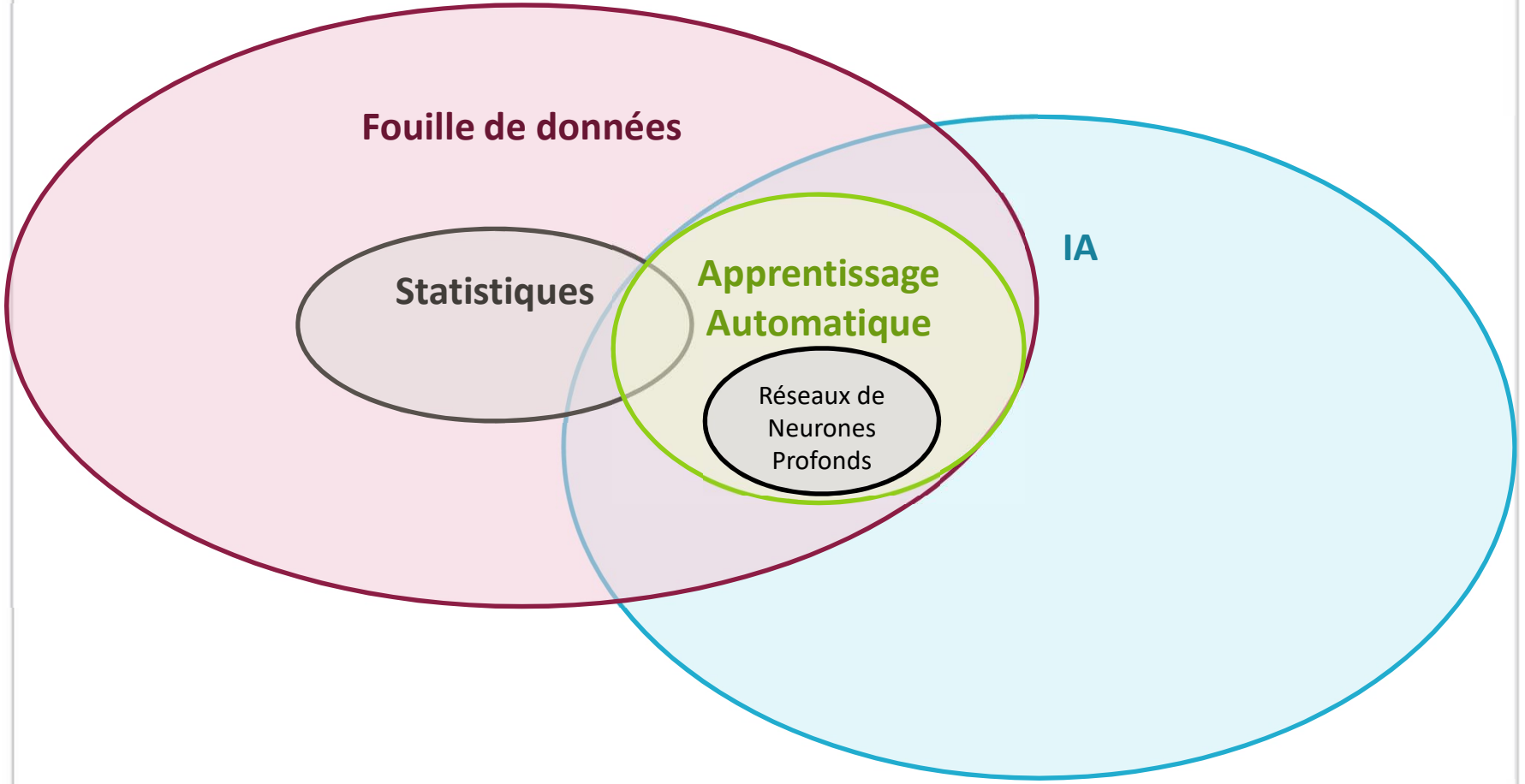
Un peu de vocabulaire

Datamining, Machine Learning, IA, Statistiques, Deep Learning, Algorithme, Système expert : c'est quoi la différence ?

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Un peu de vocabulaire

Datamining, Machine Learning, IA, Statistiques, Deep Learning : c'est quoi la différence ?



FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Domaines d'application

- Marketing direct : population à cibler (âge, sexe, profession, habitation, région, …) pour un publipostage.
- Gestion et analyse des marchés : Ex. Grande distribution : profils des consommateurs, modèle d 'achat, effet des périodes de solde ou de publicité, « panier de la ménagère »
- Détection de fraudes : Télécommunications, ...
- Gestion de stocks : quand commander un produit, quelle quantité demander, …
- Analyse financière : maximiser l 'investissement de portefeuilles d 'actions.

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Domaines d'application

Gestion et analyse de risque : Assurances, Banques (crédit accordé ou non)

Compagnies aériennes

Bioinformatique et Génome : ADN mining, …

Médecine et pharmacie :

- Diagnostic : découvrir d 'après les symptomes du patient sa maladie
- Choix du médicament le plus approprié pour guérir une maladie donné

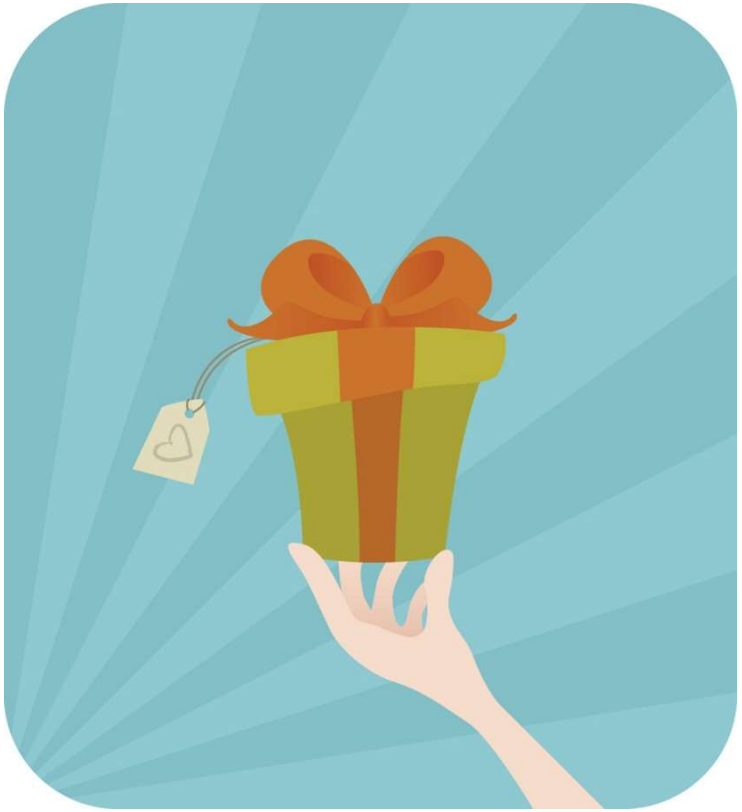
Web mining, text mining, etc.

Exemple 1 - Marketing

- Vous êtes gestionnaire marketing d'un opérateur de télécommunications mobiles :
  - Les clients reçoivent un téléphone gratuit (valeur 150€) avec un contrat d'un an ; vous payer une commission de vente de 250€ par contrat
  - Problème : Taux de renouvellement (à la fin du contrat) est de 25%
  - Donner un nouveau téléphone à toute personne dont le contrat a expiré coûte cher.
  - Faire revenir un client qui est parti est difficile et coûteux.

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple 1 - Marketing



- Trois mois avant l'expiration du contrat, prédire les clients qui vont quitter :
  - Si vous voulez les garder, offrir un nouveau téléphone.

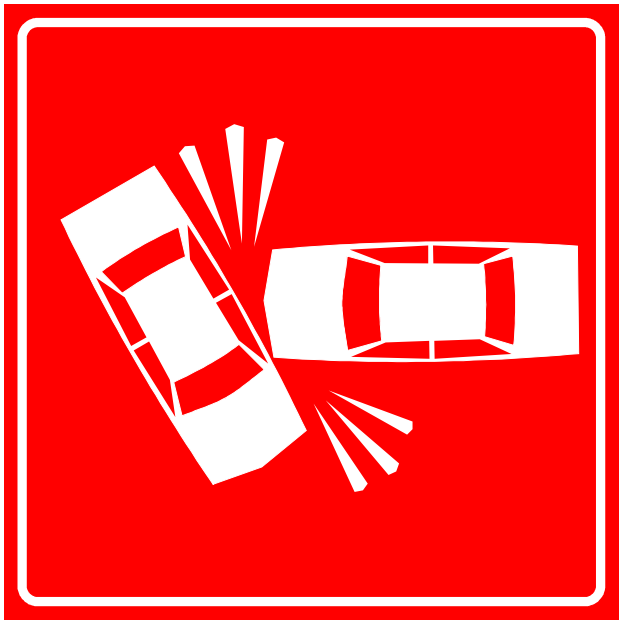
FACULTÉ DE
GESTION,
ÉCONOMIE
& SCIENCES

Exemple 2 - Assurances

- Vous êtes un agent d'assurance et vous devez définir un paiement mensuel adapté à un jeune de 18 ans qui a acheté une Ferrari.
- Qu'est ce qu'il faut faire ?

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple 2 - Assurances



- Analyser les données de tous les clients de la compagnie.
- La probabilité d'avoir un accident est basée sur … ?
  - Sexe du client (M/F) et l'âge
  - Modèle de la voiture, âge, adresse, ...
  - etc.
- Si la probabilité d'avoir un accident est supérieure à la moyenne, initialiser la mensualité suivant les risques.

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple 3 – Banque/Assurance/Télécom

Détection de fraudes pour les assurances

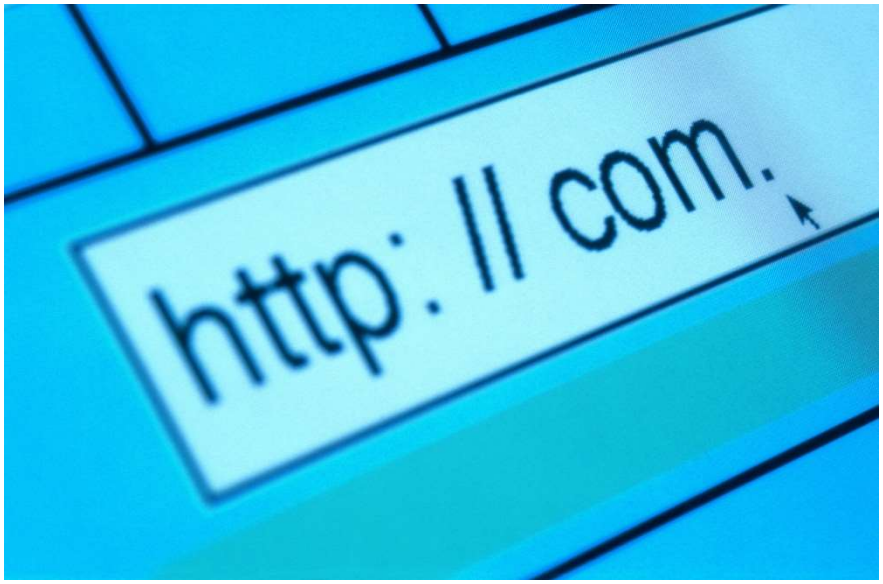
Analyse des déclarations des assurés par un expert afin d'identifier les cas de fraudes.

Extraction de caractéristiques à partir de ces déclarations (type d'accident, de blessures, etc...)

Applications de méthodes statistiques pour identifier les caractéristiques des déclarations fortement corrélés à la fraude.

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple 4 - Web



- Les logs des accès Web sont analysés pour …
  - Découvrir les préférences des utilisateurs
  - Améliorer l'organisation du site Web
- De manière similaire …
  - L'analyse de tous les types d'informations sur les logs
  - Adaptation de l'interface utilisateur/service

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple – E-commerce

Amazon

- **Opportunité** :
- La liste des achats des clients sont stockées en mémoire
- Les utilisateurs du site notent les produits !
- Comment tirer profit des choix d'un utilisateur pour proposer des produits à un autre client ?
- **Solutions** : Regrouper des clients ayant les mêmes "goûts"

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple – Commerce

- **Organisation de rayonnage**
- **Objectifs** : Identifier les produits que les gens sont susceptibles d'acheter conjointement afin d'organiser les rayonnages
- **Données** : Code-Barre des produits.
- **Méthodes :** Extractions de règles
- Exemples :
- résultats logiques : les boissons alcoolisés et les biscuits apéritifs sont souvent proches.
- résultats étranges : dans une étude américaine, la vente de bière est plus importante si le rayon des couches n'est pas trop loin, et si sur le chemin il y a des chips, cela permet d'augmenter la vente des 3 produits.

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple – E-commerce

3D-Secure

- 10 à 15% des transactions 3DSecure sont abandonnées
- L'algorithme random forest permet d'identifier 99% des fraudes
- Solution : proposer le 3DSecure seulement sur les suspicions de fraude
  - Le 3DSecure n'est plus proposé que dans 1/3 des cas

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

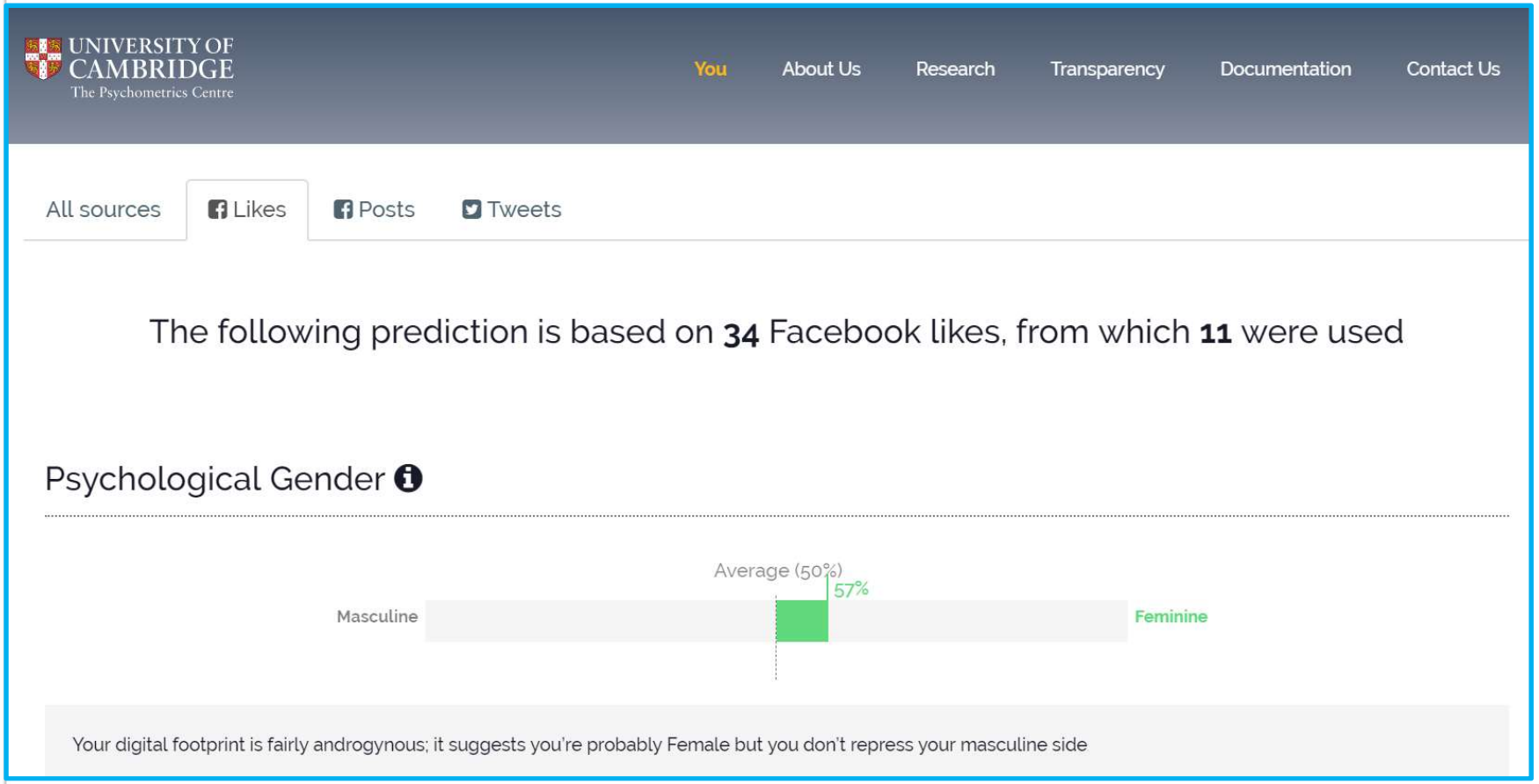
Exemple - Netflix

Extrait d'une conférence Netflix (Netflix Recommendations Using Spark + Cassandra (Prasanna Padmanabhan & Roopa Tangirala, Netflix) | Cassandra Summit 2016)

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple - Profilage

Profilage selon les likes facebook : https://applymagicsauce.com/demo.html



Mais aussi : orientation politique, type de métier, traits psychologiques, QI, …

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple - Profilage

Profilage selon les likes facebook : https://applymagicsauce.com/demo.html

Your digital footprint suggests that you have a strong interest in Engineering. You are probably an inventive and energetic person that isn't shy of DYI. Why not visit a technology museum next weekend or rekindle that hobby project collecting dust in the attic? Remember, this is just a prediction, so please consult a professional (and/or your significant other) before attempting to remodel the kitchen

These Likes make you appear ***more interested in engineering:***

These Likes make you appear ***less interested in engineering:***

These Likes make you appear ***more interested in finance:***

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Exemple - Dermatologie

Scores ROC AUC pour la détection du mélanome et du carcinome basocellulaire

	ResNet 152	DenseNet 201	Dermatologues
melanoma	94.40%	93.80%	82.26%
basal cell carcinoma	99,10%	99.30%	88.82%

Source : **Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms**
Amirreza Rezvantalab, Habib Safigholi, Somayeh Karimijeshni
(Submitted on 21 Oct 2018)