DÉVELOPPEMENT D'APPLICATIONS DE MACHINE LEARNING ORIENTÉS CLIENTS

Stage de M2

Max Halford

5 septembre 2017

CMI SID

INTRODUCTION

- **Lieu** - Berlin
- **Entreprise** - HelloFresh, entreprise B2C à base d'abonnement hebdomadaire
- **Thème** - Mise en place d'applications de machine learning et de l'architecture nécessaire

MODÈLES ORIENTÉS CLIENTS

OBJECTIFS

- Faire de la prédiction au niveau de chaque client
- Prendre en compte les différentes phase du parcours clients
- Prédire les pauses, les annulations, les réactivations

CONTRAINTES

- Données éparses, pas forcément structurés ou prêtes à l'emploi
- Données plus ou moins disponibles selon chaque client
- Problème difficile avec un fort déséquilibre dans la distribution des classes à prédire

SOLUTION

- Modèle: arbres boostés
- Rééquilibrage des données pour faciliter l'entraînement
- Nouvelle recherche exhaustive des meilleurs paramètres chaque semaine
- Développement en parallèle d'une application pour créer les jeux de données

EN PRATIQUE

- Les probabilités d'annulation sont recalculés chaque mardi
- Stockage sur un Dropbox accessible par le service client et l'équipe CRM
- Campagne de mailing et actions préventives pour les clients avec la plus forte probabilité d'annulation
- Feedback hebdomaire pour valider la performance du modèle

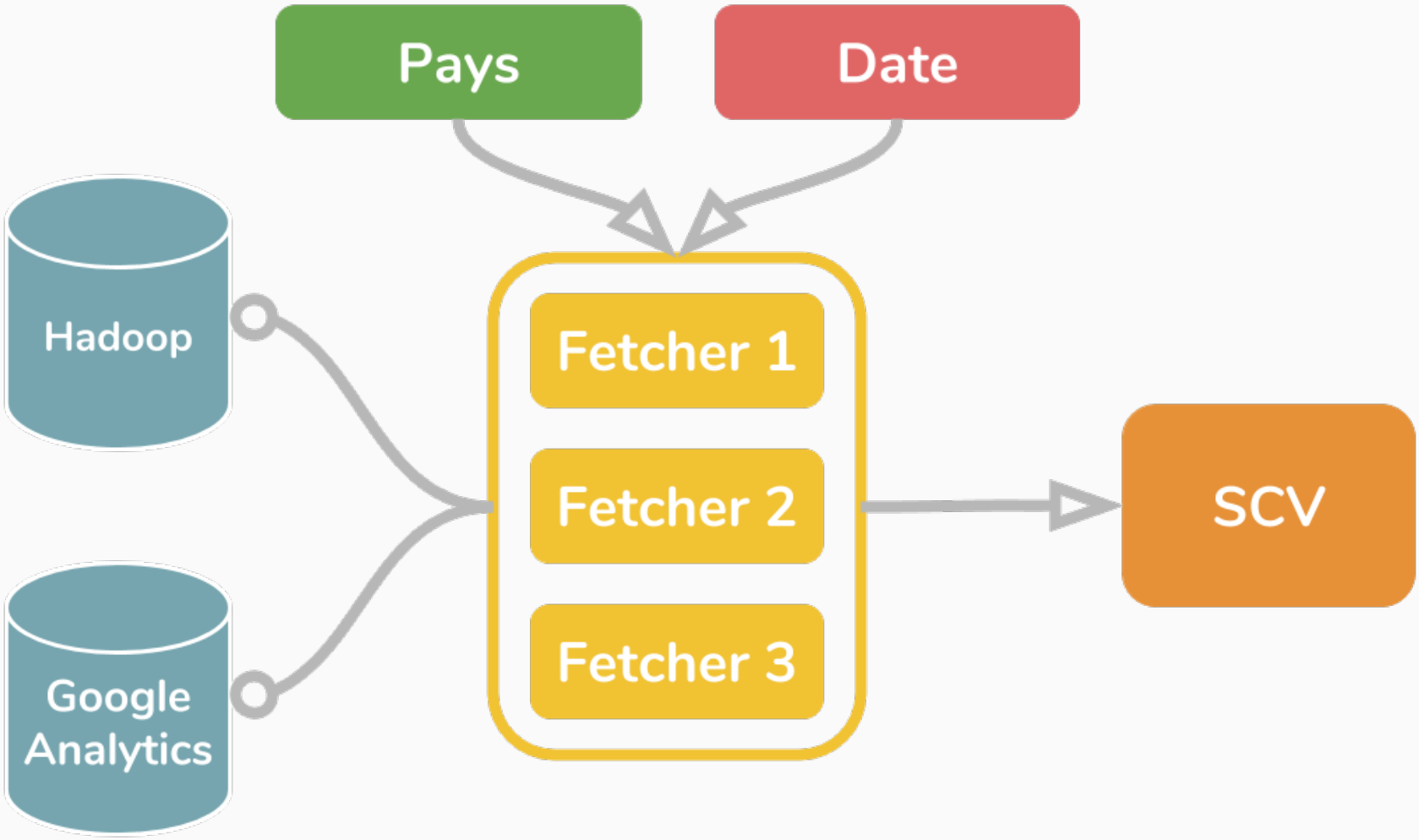
“SINGLE CUSTOMER VIEW”

DÉFINITION

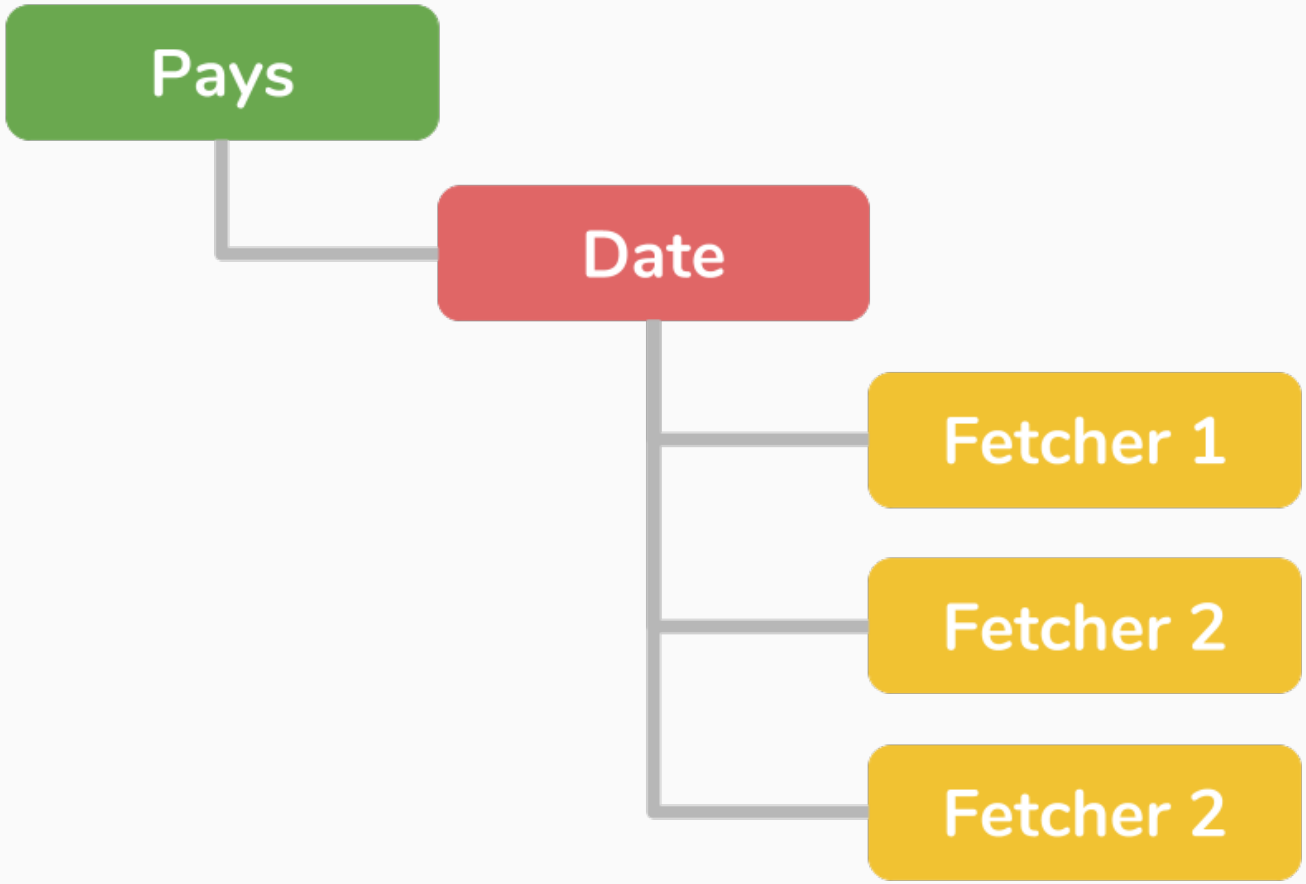
- SCV
  - Table qui associe des variables (colonnes) à des clients (lignes)
  - Table abstraite qui n'oblige pas au stockage sous forme tabulaire (mais peut être restituée sous cette forme)
  - Inclue une dimension temporelle
- Fetcher
  - Classe qui permet de récupérer des données et de les restituer sous forme tabulaire
  - Prend en paramètres un pays et une date de référence
  - Chaque fetcher est entièrement indépendant des autres fetchers
  - Un fetcher peut pouvoir fonctionner dans le présent

- Grande similarité avec un entrepôt de données
- Répond à des besoins spécifiques d'applications de machine learning "périodiques"
- Permet de conserver l'indépendance données/programme

FLUX DE DONNÉES



STOCKAGE HDF5



EXEMPLES DE FETCHERS

- Age, sexe, taille de la boîte HelloFresh
- Temps passé sur l'application mobile durant chacune des six dernières semaines (par rapport à la date de référence!)
- Type de paiement utilisé
- Marque et modèle du téléphone portable

UTILISATION À LA LIGNE DE COMMANDE (CLI)

- `python cli.py fetch US -d 2016-11-01` lance tous les fetchers pour les USA au 1^{er} Novembre 2016
- `python cli.py fetch US DE -d 2016-11-01 -d 2016-11-02` lance tous les fetchers pour deux pays à deux dates
- `python cli.py fetch US DE -d 2016-11-01 -w 16` lance les fetchers sur 16 threads

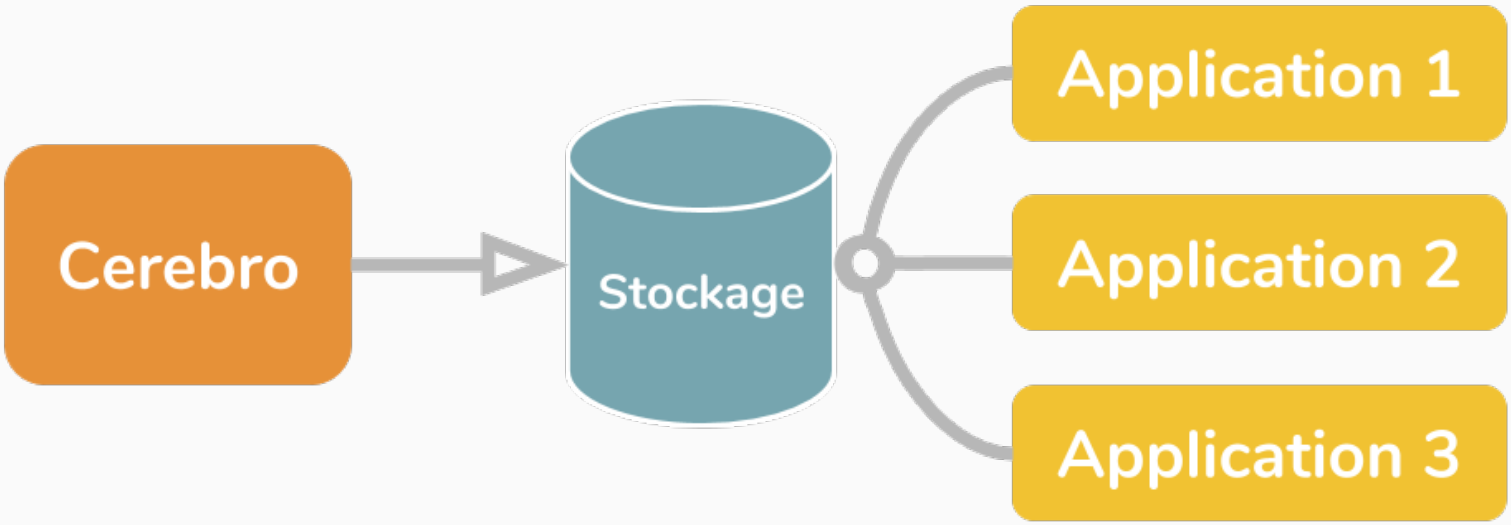
- Un CLI permet l'automatisation via un cron
- Les fetchers peuvent aussi être lancé via une API web
- Plus besoin de coder par la suite $\rightarrow$ gain de temps conséquent

“Déploiement d’applications”

PRINCIPES GÉNÉRAUX

- Des méthodes de déploiement robustes existent déjà dans d'autres domaines
- Le déploiement devrait être "stateless" $\rightarrow$ les données doivent être stockées à part du code applicatif
- Cerebro (l'application pour construire la SCV), doit être déployée sur un serveur à part des applications sous-jacentes

ARCHITECTURE



- Chaque partie a son serveur → facilite le déploiement et la scalabilité
- Le serveur de stockage est accessible par chaque application
- Théoriquement les données devraient être accédées par les applications via une API

EN PRATIQUE

- Cerebro et chaque application peuvent être lancé dans un conteneur Docker → déploiement trivial
- Avec AWS un Elastic File System (EFS) sur chaque serveur) est “monté” sur chaque serveur pour faire le lien
- Un serveur EC2 par application permet de monitorer le coût et de passer à l’échelle

CONCLUSION DU STAGE

- Grande liberté pour explorer différentes méthodes
- Apprendre à faire le lien entre machine learning et demandes business
- Nombreuses thématiques à approfondir
- Travail valorisé et utilisé en production ou bien par d'autres data scientists

Merci pour votre attention