

Flat Rat

The true price of your flat

Introduction : le PESTO Webmining



Le web mining

- ▶ Web : exploration du web pour obtenir des données
- ▶ Mining : exploitation de grands jeux de données

...machine learning

The screenshot shows a Google search results page for the query "wikipedia". The results are displayed under the "Web" tab. The first result is a link to the main Wikipedia homepage. Below it are links to the Simple English Wikipedia, the Wikipedia article about the project itself, and the RSS feed. To the right of the search results, there is a sidebar for the "Wikipedia" Google+ page, which has 22,479 followers. The sidebar also displays a recent post from "English Wiktionary" dated December 19, 2013.



Genèse du projet

« Exploiter des données internet pour estimer le prix d'un appartement. »

1. **Le prix** : réel ? annonce ? à la vente ? à la location ?

Prix de vente

2. **Estimer** : pour qui ?

Pour le vendeur et pour l'acheteur (fonction de recherche)

3. **Appartements** : maisons aussi ? limité à une région ? un département ? Paris ?

Appartements en vente à Paris



Etat de l'art

- ▶ Quelques sites estiment les biens immobiliers
- ▶ Bases de données de notaires
- ▶ Gérés par des agences

drimki
Votre projet immobilier commence ici

ACCUEIL | **ANNONCES** | **ACHETER** | **VENDRE** | **ESTIMATION** | **CARTE DES PRIX** | **GUIDE**

Estimation immobilière gratuite et fiable

Bâtie sur les données historiques notariales, notre estimation immobilière gratuite en ligne vous permet d'estimer le prix de votre bien immobilier (maison, appartement, studio, etc.). C'est rapide et sans aucun engagement ! Rentrez les caractéristiques de votre appartement ou maison pour estimer le prix de votre bien immobilier gratuitement. Vous obtiendrez la valeur de votre bien sous forme d'une fourchette de prix.

Etape 1 sur 2 : Informations générales

1. Le type et l'adresse de votre bien

Type de bien : Choisir un type
Saisir l'adresse précise du bien : Saisir le numéro et le nom de la rue **VALIDER L'ADRESSE**

Exemple : "17, Rue de Lévis, Paris" ou "24 Boulevard Gambetta, Nice"

2. Les caractéristiques de votre bien

Epoque de construction : Choisir une époque
Etage de votre bien : Choisir...
Nombre total de niveaux dont RDC : Choisir...
Ascenseur : OUI / NON
Surface Habitable (hors balcon, terrasse, cave, parking) : Choisir...
Nombre de pièces : (Salon, salle à manger et chambres) : Choisir...
Orientation du séjour : Choisir...
Surface du Terrain (facultatif - surface de l'habitation exclue) : Choisir...
Nombre de salles de bain : Choisir...

LES PRIX EN LIGNE DE L'IMMOBILIER

Votre estimation en ligne en moins de 5 minutes !

IMOGROUP Immobilier d'Excellence

Votre adresse : Votre secteur : Votre bien : Informations : Votre estimation : Afin de pouvoir réaliser l'évaluation de votre bien, veuillez indiquer son type ainsi sa localisation

Type de bien : Appartement / Maison
Localisation : N° : Voie : Code postal : Ville : J'ai lu et j'accepte les conditions d'utilisation
LOCALISEZ VOTRE BIEN

Demande d'estimation de bien immobilier à Paris

1 Localisation de votre appartement | **2 Caractéristiques immeuble et appartement** | **3 Votre estimation**

Estimation * : Choisir...
Type de bien : Choisir...
Adresse * : Saisir l'adresse précise du bien : Ville : Paris Arrondissement :
Etage du bien * : Choisir... sur un nombre total d'étages :
Surface (m²) * : Choisir...
Nombre de pièces * : Choisir...
Ascenseur * : Oui / Non
Balcon * : Oui / Non
Parking * : Oui / Non
Terrasse / Jardin * : Oui / Non

MeilleursAgents Prix immobilier **ESTIMATION** Agences Vendre Annonces Pl

Estimation immobilière gratuite

Basé sur les données historiques des Notaires, notre outil d'**estimation immobilière en ligne** vous permet d'estimer votre appartement ou maison et d'obtenir en quelques clics la valeur de votre bien immobilier. En savoir plus

L'EXPRESS | **Le Monde** | **Le Point** | **60 millions** | **VotreArgent**

Une référence pour l'évaluation de biens Un bon moyen Des résultats très proches de la réalité du marché L'un des calculateurs en ligne les plus fiables Notre avis : 4/4 ★★★★

Adresse du bien à estimer

Adresse * : Choisir...
Code postal, Ville * : Choisir...

Caractéristiques du bien

Type de bien * : Appartement
Surface Carré * : m² (hors balcon, terrasse...)
Nb. de pièces * : Choisir...
Nb. de salle(s) de bain * : Choisir...
Etage * : Choisir... sur Choisir... étages au total *



Ce que l'on va essayer d'apporter

- ▶ Recueil d'annonces réelles
- ▶ Comparaison de différents sites
- ▶ Analyse du prix d'une annonce par rapport au marché
- ▶ Visualisation des caractéristiques de biens immobiliers sur Paris
- ▶ Mise à jour des données



Présentation de notre site



http://c51-02.enst.fr:5000/ +

c51-02.enst.fr:5000 Rechercher

Home Evaluer un Bien Trouver un Bien Explorer le Marché Explorer les Caractéristiques Eléments Techniques

FlatRat

Description

Ce site web est le résultat d'un projet d'étude de six semaines réalisé dans le cadre de la formation du Corps des Mines. Cet enseignement 'webmining' a pour ambition de proposer à une demie douzaine d'ingénieurs-élèves motivés une initiation aux techniques d'apprentissage statistique (machine learning), à l'acquisition des données sur le Web et à la création d'une interface Web.

Dans le cadre de ce projet, nous avons choisi d'explorer des données immobilières. Nous avons récupéré une banque de plus de 17 000 annonces issues de divers sites immobiliers (PAP, seLoger.com, logimmo...) grâce à une technologie de 'scraping', sur lesquelles nous avons ensuite réalisé un apprentissage statistique. Nous ne traitons pour l'instant que les annonces de vente pour le marché parisien.

L'Equipe

 Marion Scraping & DataViz	 Mathilde Interface Web	 Adrien Interface Web & DataViz
 Christophe Machine Learning & DataViz	 Emmanuel Machine Learning	 Hugo Traitement des Données

Menu

- [Estimer le prix d'un bien](#)
- [Trouver le bien de vos rêves](#)
- [Explorer le marché parisien](#)
- [Explorer les caractéristiques](#)
- [Quelques éléments techniques](#)



Etapes du projet

Pesto Webmining
Etapes du projet



I. Web crawling



II. Traitement de texte



III. Visualisation des données



IV. Pré-traitement des données



V. Machine learning



VI. Site web



Etapes du projet



I. Web crawling : Marion

II. Traitement de texte

III. Visualisation des données

IV. Pré-traitement des données

V. Machine learning

VI. Site Web



Etapes du projet

I. Web crawling

II. Traitement de texte : Hugo

III. Visualisation des données

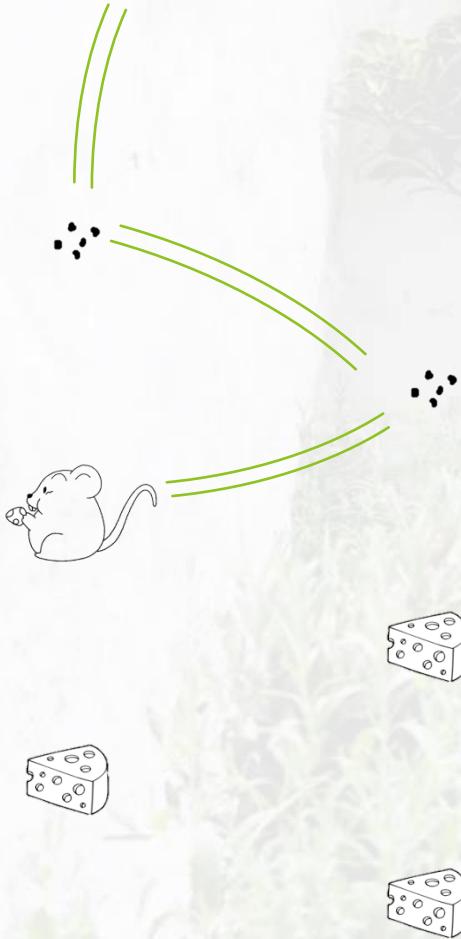
IV. Pré-traitement des données

V. Machine learning

VI. Site web



Etapes du projet



- I. Web crawling
- II. Traitement de texte
- III. Visualisation des données : Marion**
- IV. Pré-traitement des données
- V. Machine learning
- VI. Site Web



Etapes du projet

- I. Web crawling
- II. Traitement de texte
- III. Visualisation des données
- IV. Pré-traitement des données : Christophe**
- V. Machine learning
- VI. Site web



Etapes du projet

- 
- I. Web crawling
 - II. Traitement de texte
 - III. Visualisation des données
 - IV. Pré-traitement des données
 - V. Machine learning : Emmanuel
 - VI. Site web



Etapes du projet



- I. Web crawling
- II. Traitement de texte
- III. Visualisation des données
- IV. Pré-traitement des données
- V. Machine learning
- VI. Site web : Mathilde & Adrien



Etapes du projet



I. Web crawling

II. Traitement de texte

III. Visualisation des données

IV. Pré-traitement des données

V. Machine learning

VI. Site web



Collecte de données

- ▶ Web crawling
- ▶ 6 sites d'annonces immobilières

Caractéristiques des biens en vente

pap.fr Immobilier de particulier à particulier

Vente appartement Paris + 200 annonces

Paris (75) -
Appartement Chambres Pièces Min € Max € Actualiser

Annonces à la une

Paris 2e Appartement, 2 pièces, 42 m² 550.000 €
Paris 16e Appartement, 3 pièces, 114 m² 950.000 €
Paris 16e Appartement, 5 pièces, 82 m² 440.000 €
Paris 16e Appartement, 2 pièces, 62 m² 763.600 €

Vente appartement 3 pièces 70 m² Paris 14e 630.000 € à partir de 1.135 €/mois*

Paris 14e 105014, 3 pièces, 70 m² + 2 balcons - Paris 14e Quartier : Neuilly - Clichy - Issy - Nanterre - Boulogne - Neuilly - Asnières - Garches - Clichy - Clichy-sous-Bois - Neuilly-Plaisance - Pernety - Didot

Prix Chambre Surface 3 2 70 m² Contact Détails

Baromètre des taux immobiliers empruntis /

SeLoger

18397 Vente appartement à Paris (75)

Filtrer Recherche TYPE DE BIEN LIEU PROX PIÈCES SURFACE CARACTÉRISTIQUES

Achat : toutes les annonces Appartement Paris (75)

TRI date de création FILTRES

Appartement 6 pièces, Paris 16ème 1 900 000 € ou > 397 Km² Obam immobilier Contactez-nous Téléphone

Demandez votre prêt et recevez une réponse de principe immédiate avec Crédit Agricole e-immobilier

Appartement 5 pièces, Paris 7ème 1 680 000 € ou > 549 Km² walls paris associés 549.000 € Contactez-nous Téléphone

Appartement 2 pièces, Paris 1er 490 000 € ou > 129 Km² Service france Contactez-nous Téléphone

Appartement 3 pièces, Paris 1... 580 000 € ou > 109 Km² Le bon agent 109.000 € Site web Contactez-nous Devise diagnostic immobilier

RENSEIGNEZ VOUS ICI

We make it visible.

laforêt la vie, la maison, laforêt.

TROUVER MON POINT DE VENTE Mon code postal OK ESPACE LAFORÊT Connectez-vous ou créez votre compte

ACCUEIL ACHETER LOUER VENDRE FAIRE LOUER / GÉRER

Mon compte Espace

ANNONCES APPARTEMENT PARIS

712 annonces immobilières correspondant à Annonces Appartement Paris

Vous recherchez Appartement Paris, c'est enfin possible avec Laforêt. Pour votre Appartement dans votre région, Laforêt vous propose de nombreuses offres. Faites appel au professionnalisme de Laforêt pour trouver l'appartement de vos rêves. Vous pouvez également nous contacter pour toute demande d'information ou de conseil. Nous sommes à votre disposition pour tout renseignement. Vous pouvez également nous contacter pour toute demande d'information ou de conseil. Nous sommes à votre disposition pour tout renseignement.

LOCATION APPARTEMENT PARIS 02 (75002) Annexe Prestige Années Appartement Paris PARIS 21 SENTIER Au cœur du secteur, au croisement de la rue St Joseph, au Séminaire, dans un grand 2 pièces d'environ 50m² avec entrée en grand séjour, cuisine séparée, deux chambres, deux salles de bains et deux toilettes. Chauffage et eau chaude collectif. A voir rapidement. Disponibilité au 1AN. Honoraire en sus ** 240 € TTC Dépot de garantie : 0 € TTC

LOCATION APPARTEMENT STUDIO PARIS 20 (75020) Annexe Appartement Paris PARIS 20 M GAMBETTA Dans immeuble récent Studio de 25m² tout meublé: entrée, pièce principale, cuisine ouverte et équipée, salle d'eau avec WC. Beau studio. Disponibilité tout de suite. Visite sur demande uniquement.

LOCATION APPARTEMENT 3 PIÈCES PARIS 17 (75017) Annexe Appartement Paris PEREY-BEL FRANCIS DE PRESSENCIE Beau deux pièces total à neuf au 2ème étage sans ascenseur au calme comprenant une entrée, un séjour, une cuisine aménagée, une chambre et une salle de bain. Chauffage à l'eau individuel électrique. Honoraire en sus ** 377,4 € TTC Dépot de garantie : 722 € TTC



Collecte de données

- ▶ Web crawling
 - ▶ 6 sites d'annonces immobilières

- ▶ Open data
 - ▶ RATP
 - ▶ Ville de Paris

Caractéristiques des biens en vente

- Localisation des biens
- Visualisation sur une carte



Web-crawling

- Objectif : récupérer des informations dans des pages HTML

The screenshot shows a web browser window displaying a real estate listing on the website www.seloger.com/immobilier/achat/immo-paris-75/bien-appartement/. The page features a search bar at the top with various filters like 'FILTRE', 'RECHERCHE', 'TYPE DE BIEN', 'LIEU', 'PRIX', 'PIÈCES', 'SURFACE', and 'CARACTÉRISTIQUES'. Below the search bar, there's a large advertisement for '10 minutes' featuring a clock and balloons. The main content area shows a listing for an 'Appartement 3 pièces, Paris 15ème' with a price of '595 000 € ou 1 753 €/mois*'. The listing includes a photo of the building, details about its location ('Au pied des commerces (centre commercial Beaugrenelle) et des transports,...'), and contact information ('Contactez-nous', 'Téléphone'). A banner for 'e-immobilier' from Crédit Agricole is also visible. At the bottom of the page, the URL [https://e-immobilier.credit-agricole.fr/simulca/?ORI=seloger&xtr=AL-2-\[partenariat\]-1\[SeLoger\]-\[Bandea...](https://e-immobilier.credit-agricole.fr/simulca/?ORI=seloger&xtr=AL-2-[partenariat]-1[SeLoger]-[Bandea...) is shown. The browser's developer tools are open, specifically the 'Inspecteur' (Inspector) panel, which shows the DOM structure of the page. A blue selection box highlights an anchor tag with the ID '103826491'. The right side of the developer tools shows the 'Regles' (Rules) panel with CSS styles applied to the selected element.



Web-crawling

- ▶ Objectif : récupérer des informations dans des pages HTML
- ▶ Développement de robots d'indexation
 - ▶ Fouillent le web récursivement à partir d'une URL de départ
 - ▶ Extraient des données via les balises HTML



Scrapy



Stratégie de fouille

- ▶ Un robot d'indexation par site

- ▶ Développement :

- ▶ Définir les URL de départ
- ▶ Définir des règles pour cliquer sur tous les liens d'annonces
- ▶ Extraire les caractéristiques du bien pour chaque page d'annonce



The screenshot shows a search results page for "18397 Vente appartement à Paris (75)". The top navigation bar includes links for Accueil, Annonces, Île-de-France, Paris, and Ventes Appartement. A green circle highlights the search term "18397 Vente appartement à Paris (75)". Below the search bar are filters for Achat (toutes les annonces), Type de bien (Appartement), Lieu (Paris (75)), Prix, and Pièces. The results section displays four apartment listings:

- Appartement 6 pièces, Paris 16ème**
1 900 000 € ou 5 597 €/mois*
Adresse de prestige Place de Mexico bel appartement en duplex dans un...
3 chb | 201 m²
- Appartement 5 pièces, Paris 7ème**
1 680 000 € ou 4 949 €/mois*
Étonnant appartement plein de personnalité avec une distribution...
2 chb | 125 m²
- Appartement 2 pièces, Paris 1er**
468 000 € ou 1 379 €/mois*
Au sein d'un immeuble typiquement parisien du 1er...
1 chb | 47 m²
- Appartement 3 pièces, Paris 1...**
580 000 € ou 1 709 €/mois*
Rue Lecourbe à proximité du Métro Sévres Lecourbe, joli 3 pièces en bon état et...

Each listing includes contact information for the real estate agency (e.g., Obam immobilier, e-immobilier, Walls paris associés, Invencis france, Le bon agent) and a small thumbnail image of the apartment interior.



Définir l'URL de départ

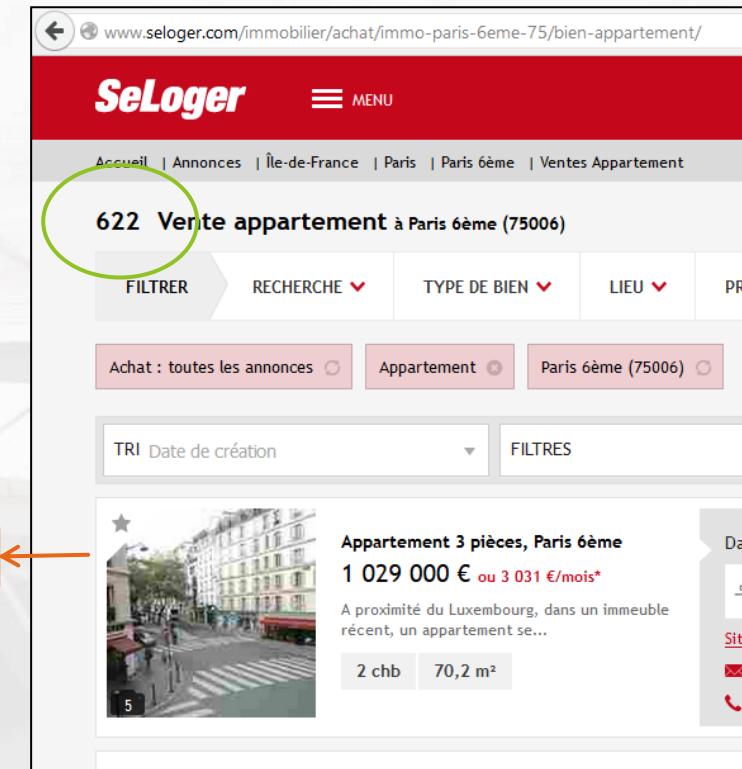
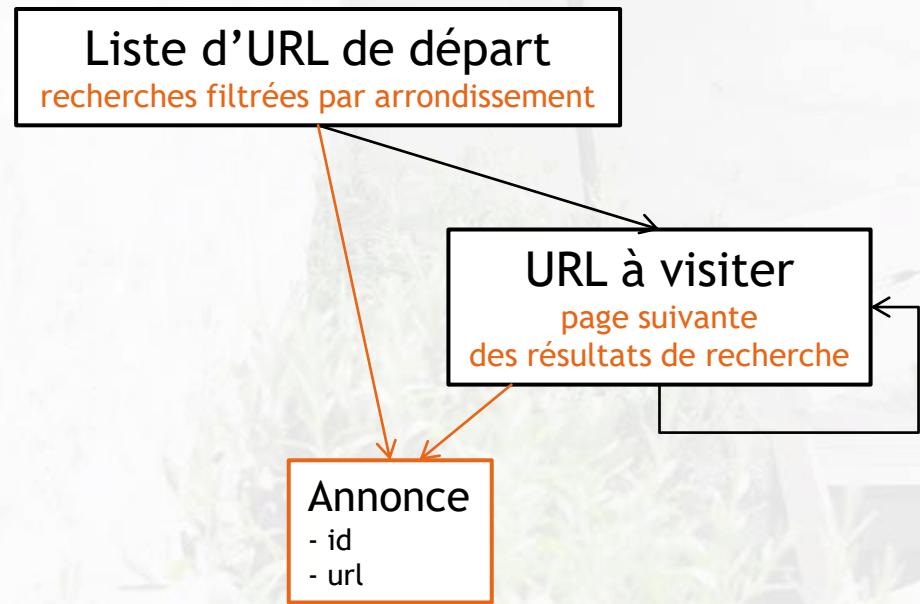
- ▶ Idée naïve : rechercher les appartements en vente à Paris



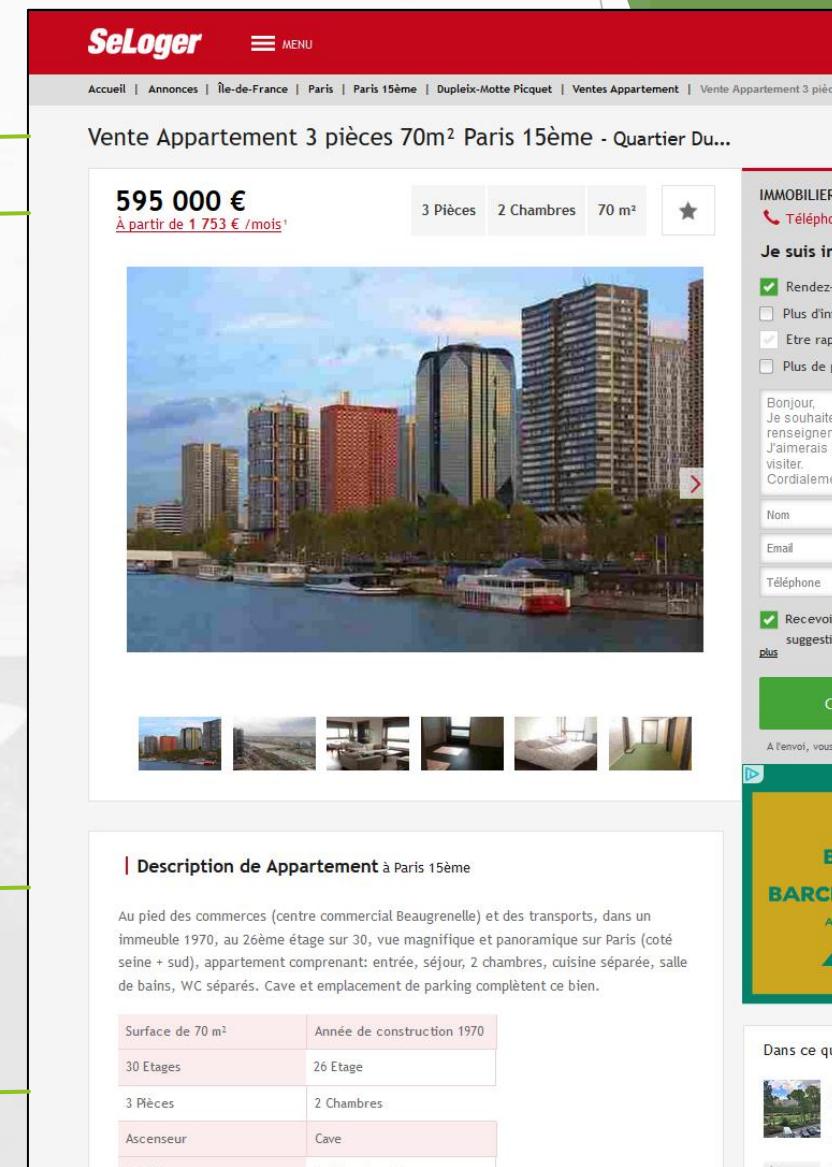
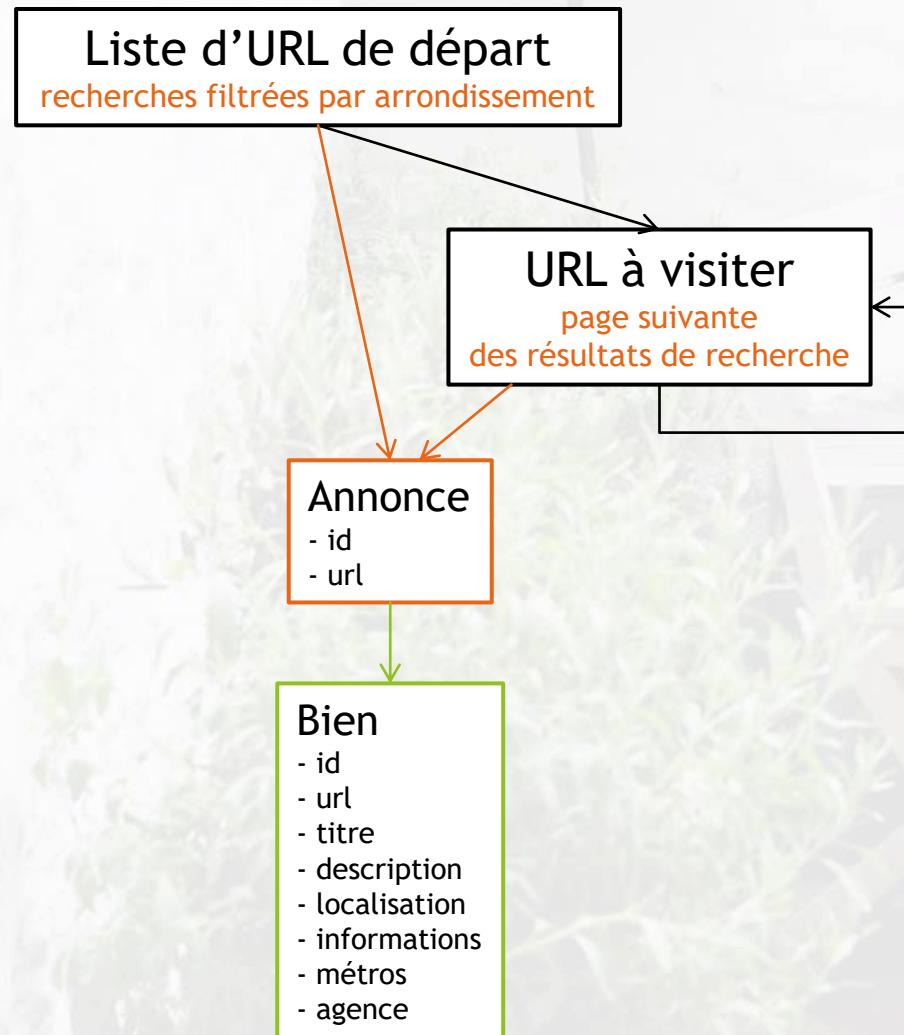
- ▶ Problèmes :
- ▶ seulement 20 objets sont affichés sur la page
- ▶ même en tournant les pages seulement 2000 objets sur 18346 sont accessibles !
- ▶ Il faut donc utiliser une liste d'URL au départ de la fouille



Stratégie de fouille



Stratégie de fouille



Bilan du crawling

- ▶ 6 robots d'indexation implémentés
 - ▶ Difficultés : variabilité des cas
 - ▶ Avantages : code flexible
 - ▶ Stratégie de politesse : attention à la *blacklist* !
- ▶ Formation d'une base de 18 000 annonces
 - ▶ Stratégie de revisite : mise à jour en quelques heures



Etapes du projet

I. Web crawling

II. Traitement de texte

III. Visualisation des données

IV. Pré-traitement des données

V. Machine learning

VI. Site web



Entrée : fichier d'annonces



apparts.json

- ▶ 17 Mo
- ▶ 18 000 annonces

Une annonce = plusieurs champs textes

- ▶ Url
- ▶ Titre
- ▶ Infos
- ▶ Localisation
- ▶ Description
- ▶ Prix



Objectif



Textes d'annonce

Vente Appartement 2 pièces 53m²
498 000 € Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.

Travaux À Prévoir	Surface de 53 m ²	Année de construction 1954
8 Etages	1 Etage	2 Pièces
1 Chambre	1 Salle de bains	1 Toilette
Ascenseur	Cave	Chauffage central fuel
Interphone	Digicode	Gardien
Entrée	Salle de Séjour :	Calme



Vecteur de caractéristiques

surface	53
ascenseur	0
cave	1
balcon	
etage	1
parking	
nb_pieces	2
ardt	16
latitude	48,856
longitude	2,354
prix	498 000



Première étape : liste de caractéristiques

Nom de la feature
haussmann
ascenseur
orientation_sud
salon
metro_proche
mezzanine
digicode
videophone
cuisine_equipée
sans_vis_a_vis
parking
cave
surface
coordonnee_x
coordonne_y
arrondissement
etage
nombre_etages
nombre_pieces
nombre_salles_de_bains
nombre_WC
nombre_chambres
date_refection
charges
nombre_fenetres
surface_balcon
date_construction



Traitement de texte



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte



Vente Appartement 2 pièces **53m²** Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte



Vente Appartement **2 pièces** 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une **cave** complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.



Traitement de texte : méthode



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.

Ascens. : 1

On recherche soit un seul mot



Traitement de texte : méthode



Vente Appartement 2 pièces 53m² Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.

Ascens. : 1

On recherche soit un seul mot
soit un nombre dans un groupe de mots



Traitement de texte



Vente Appartement 2 pièces 53m² Paris 16ème

Ardt : 16

Etage : 1

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.

Surface : 53

Pièces : 2



Coordonnées

Pourquoi ?

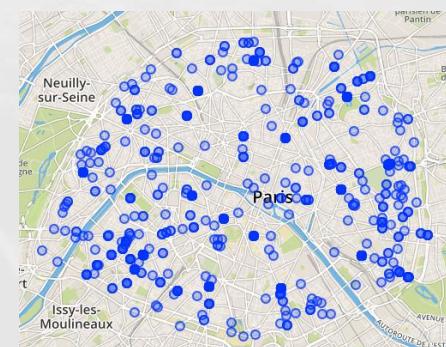
- ▶ Déterminant dans le prix d'un bien
- ▶ Coordonnées arrondissements : pas assez précis

Comment ?

- ▶ Base de données des rues et des métros de Paris
 - Metros en opendata, scrapping Gmaps pour les rues (merci Marion)
- ▶ Recherche comme pour les features

Conséquences

- ▶ Multiplication du temps de traitement par 10
(<100 features vs. 5000 rues et métros)



Problèmes rencontrés (et traités)

- ▶ Distinction entre caractéristiques proches
 - « Appartement de 120 m² avec séjour de 30 m² »
- ▶ Tournures de phrases négatives
 - « Appartement 6eme étage sans ascenseur »
- ▶ Nombres en toutes lettres
 - « Très bien situé, six pièces au 4eme étage »
- ▶ Fautes de français
 - « Appart' avec khav 6eme etaj »
- ▶ Ponctuation
 - Nombre de pièces : 2 ; etage, 3eme
- ▶ Multiplicité des tournures
 - 9^e/9eme ardt/ard/art/arrondissement / Paris 9eme / 75009



0% d'erreur
n'existe pas !



Traitement de texte



Textes d'annonce

Vente Appartement 2 pièces 53m²
498 000 € Paris 16ème

A VENDRE - EXELMANS Dans bel immeuble d'angle de 1954, au 1er étage avec ascenseur, un beau 2 pièces d'une superficie de 53 m² composé d'une entrée, d'un séjour, d'une chambre, d'une cuisine aménagée, d'une salle de bains, d'un dressing, 1 WC. L'appartement est très calme et bénéficie d'une belle luminosité. Les travaux de réfection des parties communes ont été votés. Une cave complète ce bien. Prix de vente: 498000 euros Contact: Mme Lerner.

Travaux À Prévoir	Surface de 53 m ²	Année de construction 1954
8 Etages	1 Etage	2 Pièces
1 Chambre	1 Salle de bains	1 Toilette
Ascenseur	Cave	Chauffage central fuel
Interphone	Digicode	Gardien
Entrée	Salle de Séjour :	Calme



Vecteur de caractéristiques

surface	53
ascenseur	0
cave	1
balcon	
etage	1
parking	
nb_pieces	2
ardt	16
latitude	48,856
longitude	2,354
prix	498 000



Résultat

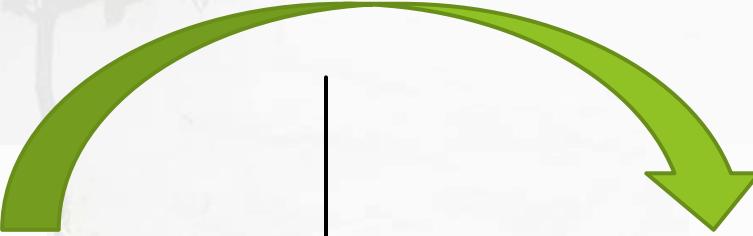


apparts.json

- ▶ 17 Mo
- ▶ 18 000 annonces



Données.csv



- ▶ 14 Mo
- ▶ Colonnes : 90 features
- ▶ Lignes : 18 000 annonces



Quelques chiffres



Données.csv

- ▶ 18 000 annonces
 - ▶ 14 000 avec nombre de chambres
 - ▶ 11 000 avec cave
 - ▶ 9000 avec ascenseur
 - ▶ **7000 avec coordonnées précises** (métro ou rue)
 - ▶ 6000 avec la date de construction
 - ▶ 4000 avec cuisine américaine

- ▶ 1200 lignes de code
- ▶ 2h pour traiter les 18 000 annonces (2,5 annonces/seconde)



Etapes du projet



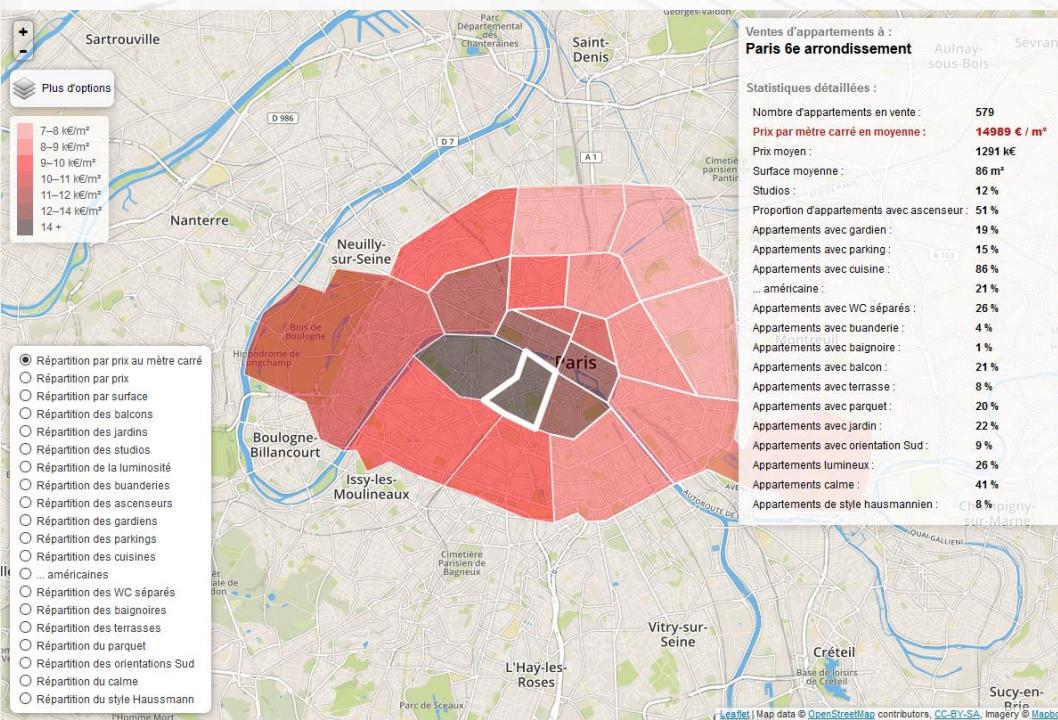
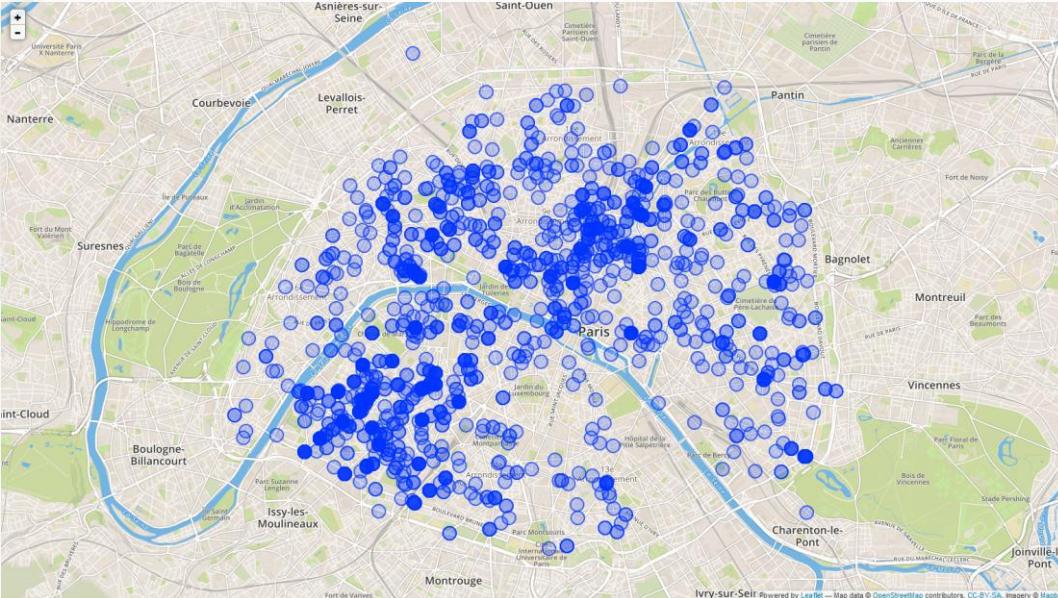
- I. Web crawling
- II. Traitement de texte
- III. Visualisation des données**
- IV. Pré-traitement des données
- V. Machine learning
- VI. Site Web

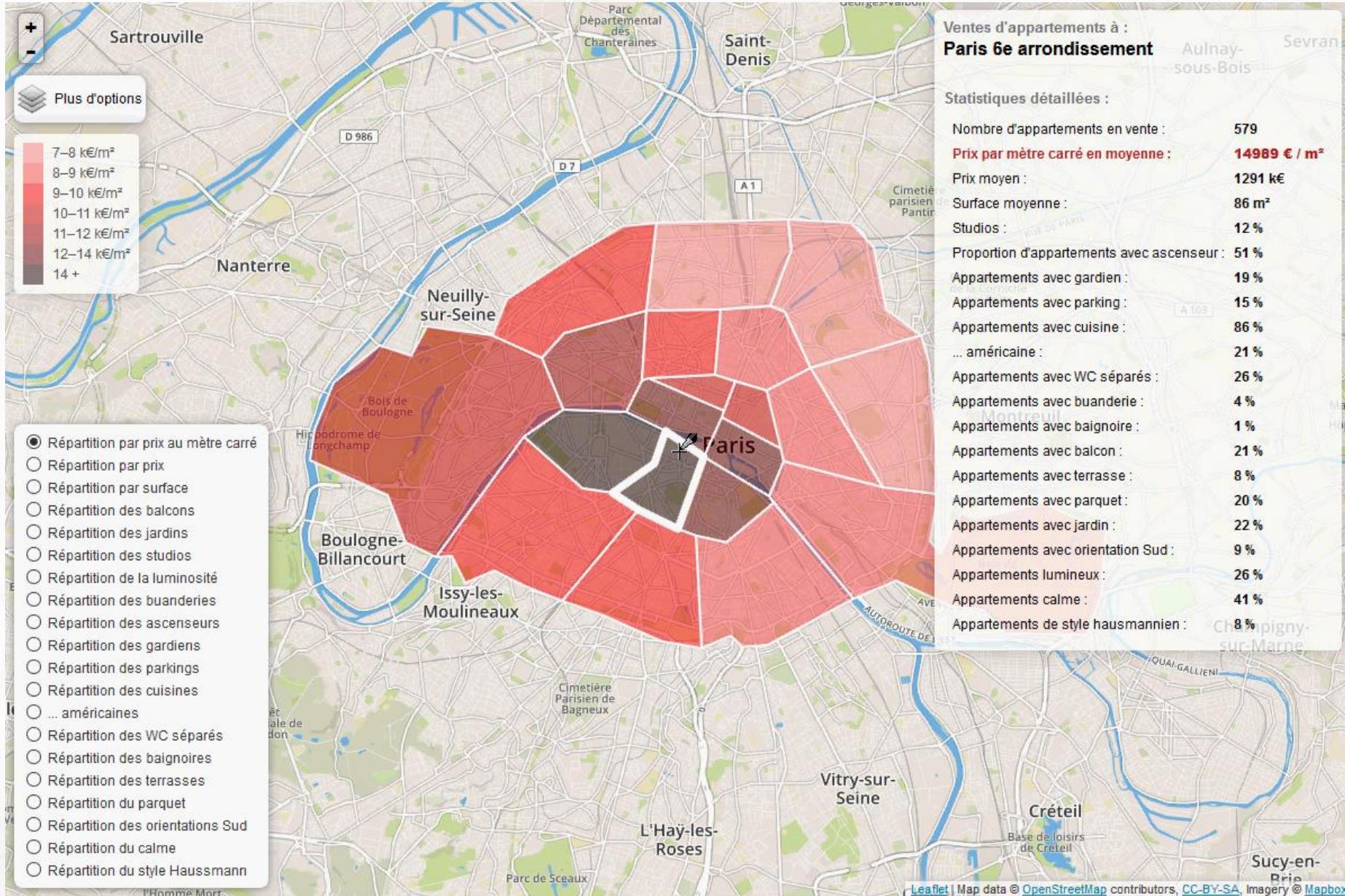


Visualisation

- ▶ Appartements
 - ▶ Caractéristiques
 - ▶ Localisation

Leaflet





Etapes du projet

- I. Web crawling
- II. Traitement de texte
- III. Visualisation des données
- IV. Pré-traitement des données**
- V. Machine learning
- VI. Site Web



Prétraitement des données

Problèmes :

- ▶ Théorie : $X^i = (x_1^i, x_2^i, \dots, x_{90}^i)$ $y = f(X^i) + \varepsilon$
- ▶ Pratique : Annonce= (oui, non, oui, ??, ??, ..., 34 m²,??...)
- ▶ Pratique : oui/non = 0/1
?? = -1
Valeurs entières



Prétraitement des données

Travail sur les caractéristiques (pas le prix)

- ▶ Deux types de caractéristiques :
 - ▶ 15 quantitatives (surface, nombre de pièces, coordonnées, étage...).
 - ▶ 75 qualitatives (...le reste).
- ▶ De qualités inégales:
 - ▶ Parmi les 15, 8 quantitatives bien renseignées (faciles à récupérer).
 - ▶ Qualitatives assez incomplètes (BDD et annonce utilisateur).



Prétraitement des données

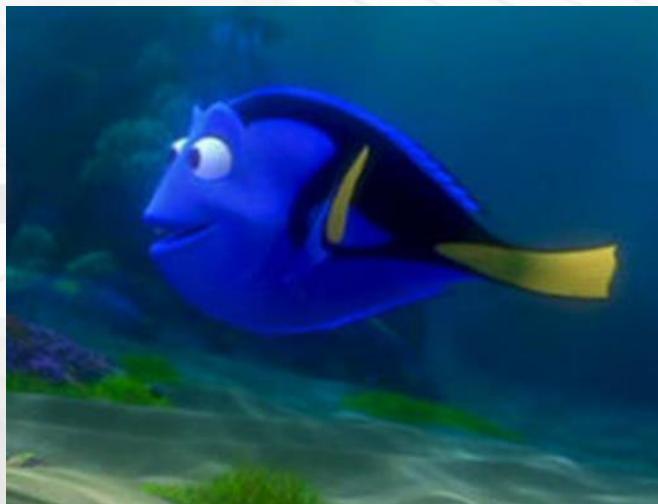
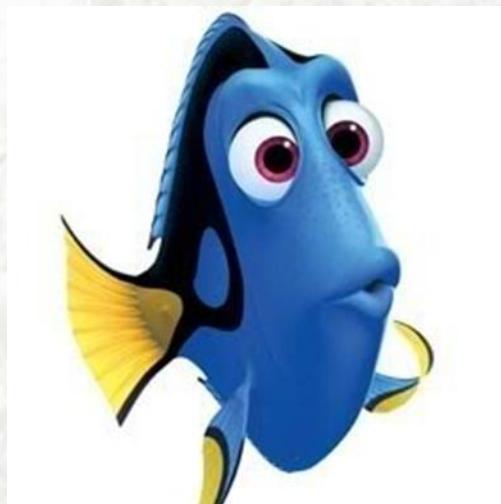
- ▶ Pour les 8 caractéristiques quantitatives : réduction de dimension
Technique : analyse en composantes principales.



Prétraitement des données

Analyse en Composantes Principales

Notion de direction principale



Prétraitement des données

Analyse en Composantes Principales

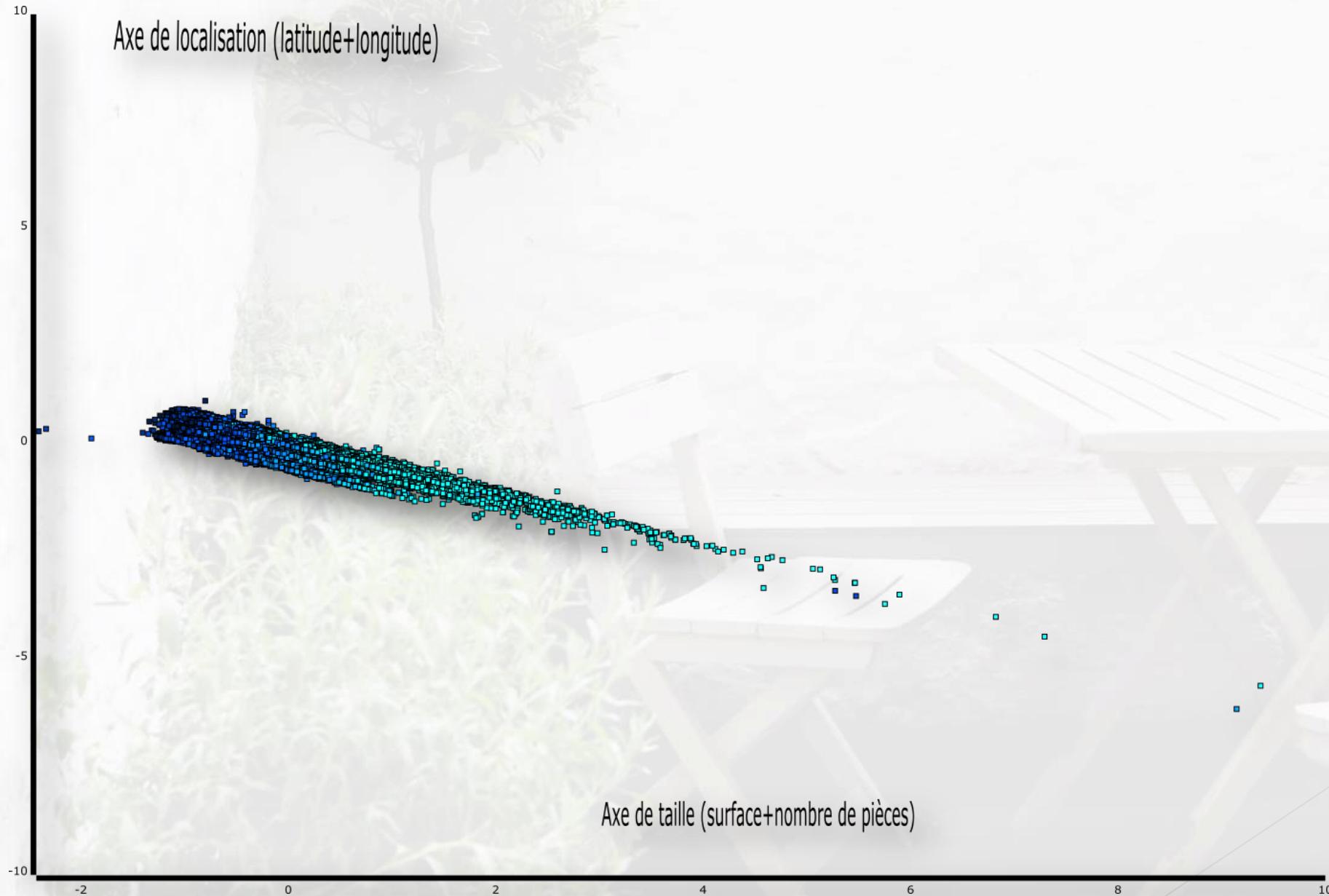
Au départ, 8 dimensions :

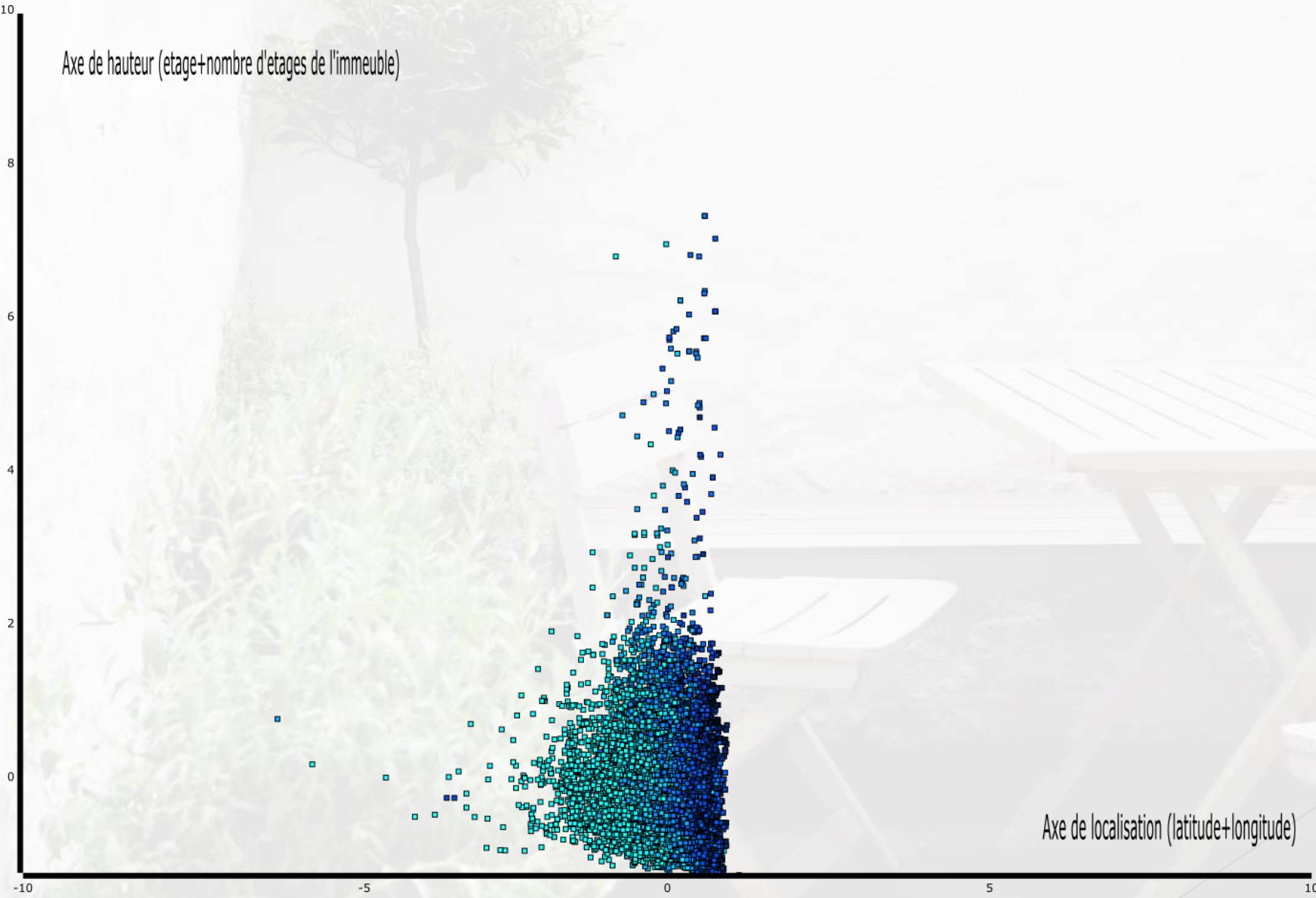
- | | |
|-------------------|-----------------------|
| 1. Surface | 5. Etage |
| 2. Latitude | 6. Nombre d'étages |
| 3. Longitude | 7. Nombre de pièces |
| 4. Arrondissement | 8. Nombre de chambres |

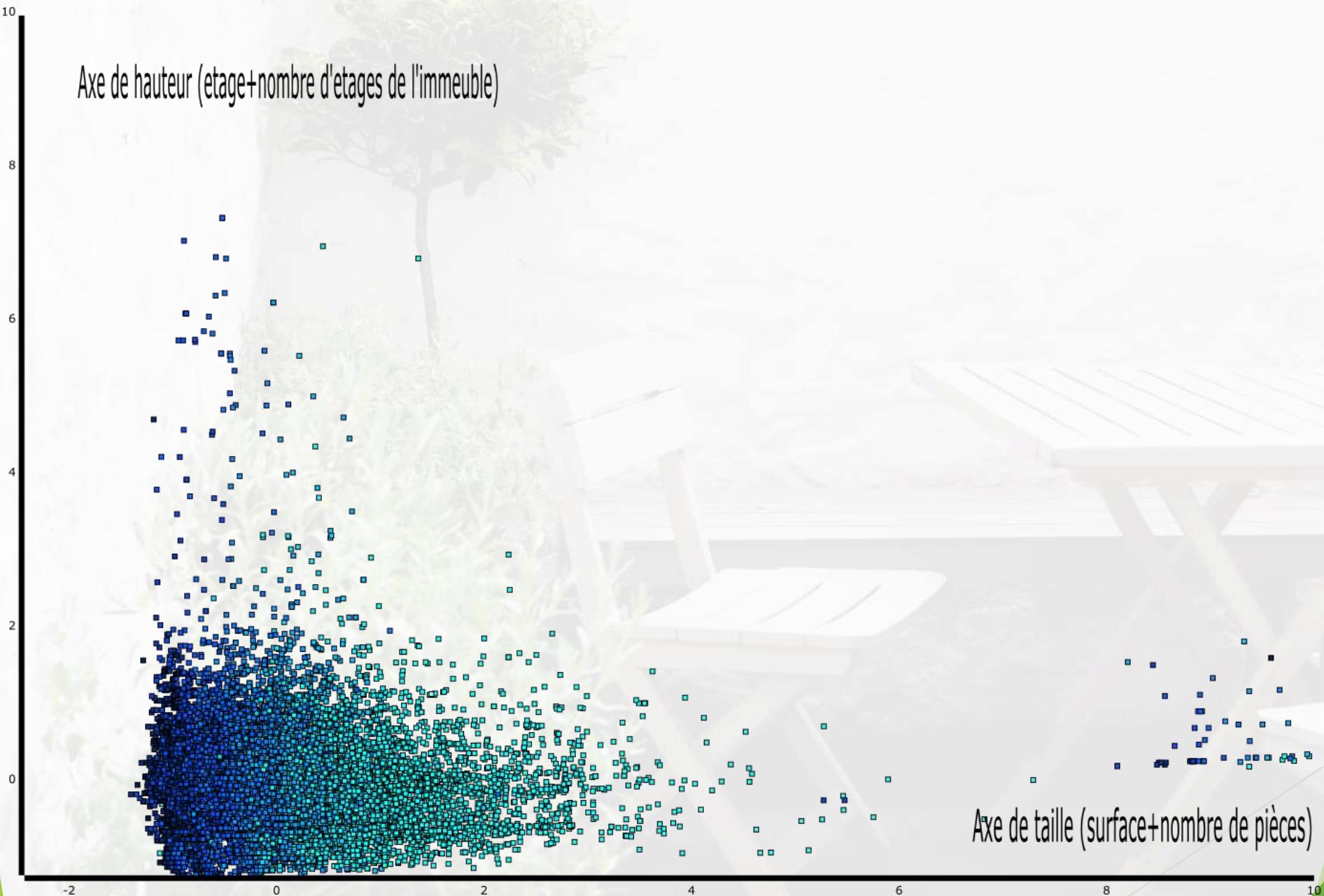
A la fin, 4 dimensions :

1. Surface-nombre de pièces (et un peu coordonnées et nombre de pièces)
2. Coordonnées
3. Etage et nombre d'étages
4. Etage









Prétraitement des données

Analyse en Composantes Principales

Résultat :

- ▶ Une dimension effective : 4 au lieu de 8
- ▶ Des directions principales : combinaisons de caractéristiques, importance.

Aide à l'apprentissage?



5757.003

Prétraitement des données

- ▶ Pour les caractéristiques quantitatives : réduction de dimension
Technique : analyse en composantes principales.
- ▶ Pour les caractéristiques qualitatives : analyse des correspondances



Prétraitement des données

Analyse des correspondances

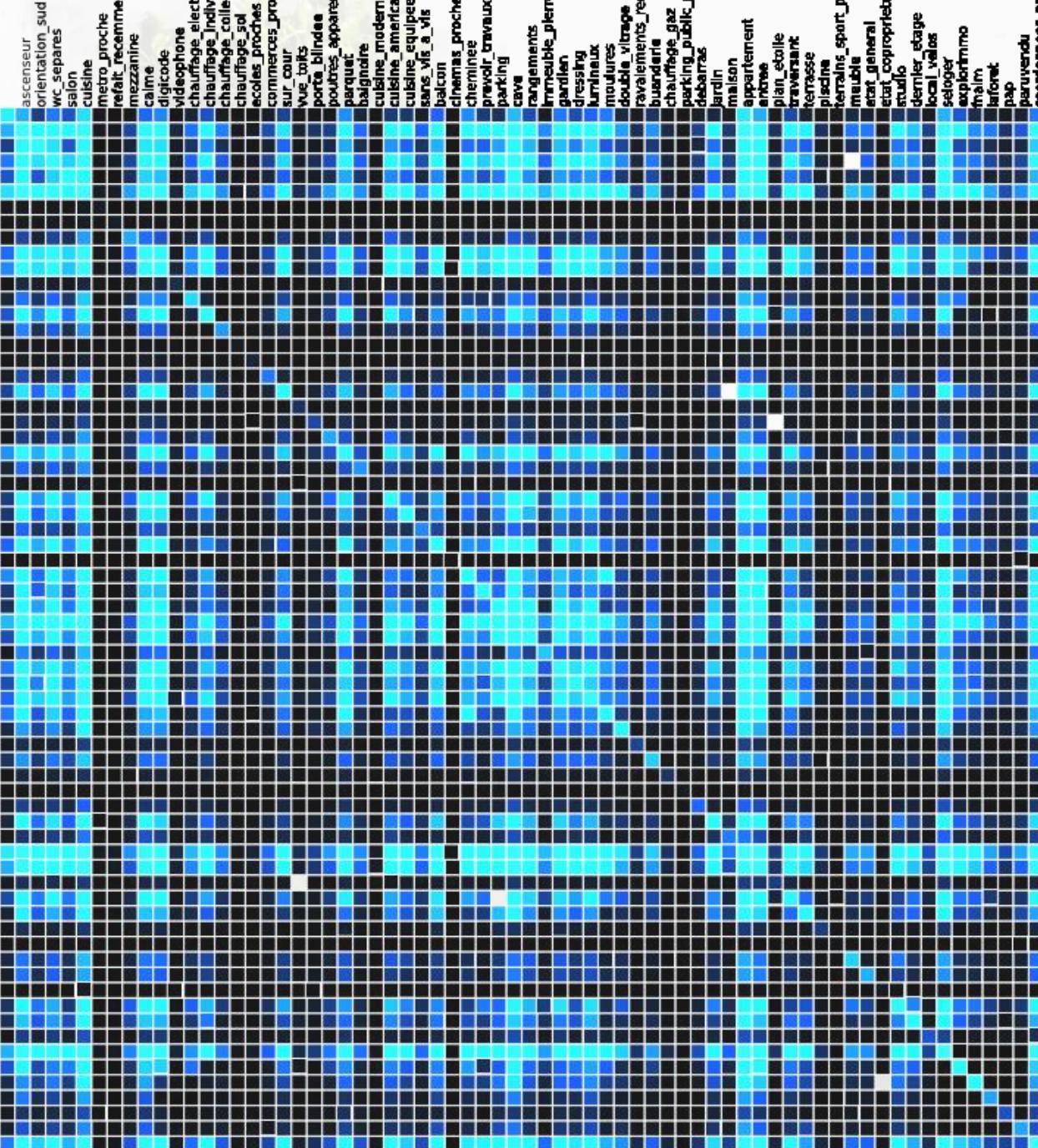
- ▶ But : évaluer des corrélations entre différentes caractéristiques qualitatives
(et avec quelques caractéristiques quantitatives?)
- ▶ Idée : compter des co-occurrences.
- ▶ Beaucoup de co-occurrences : corrélation.



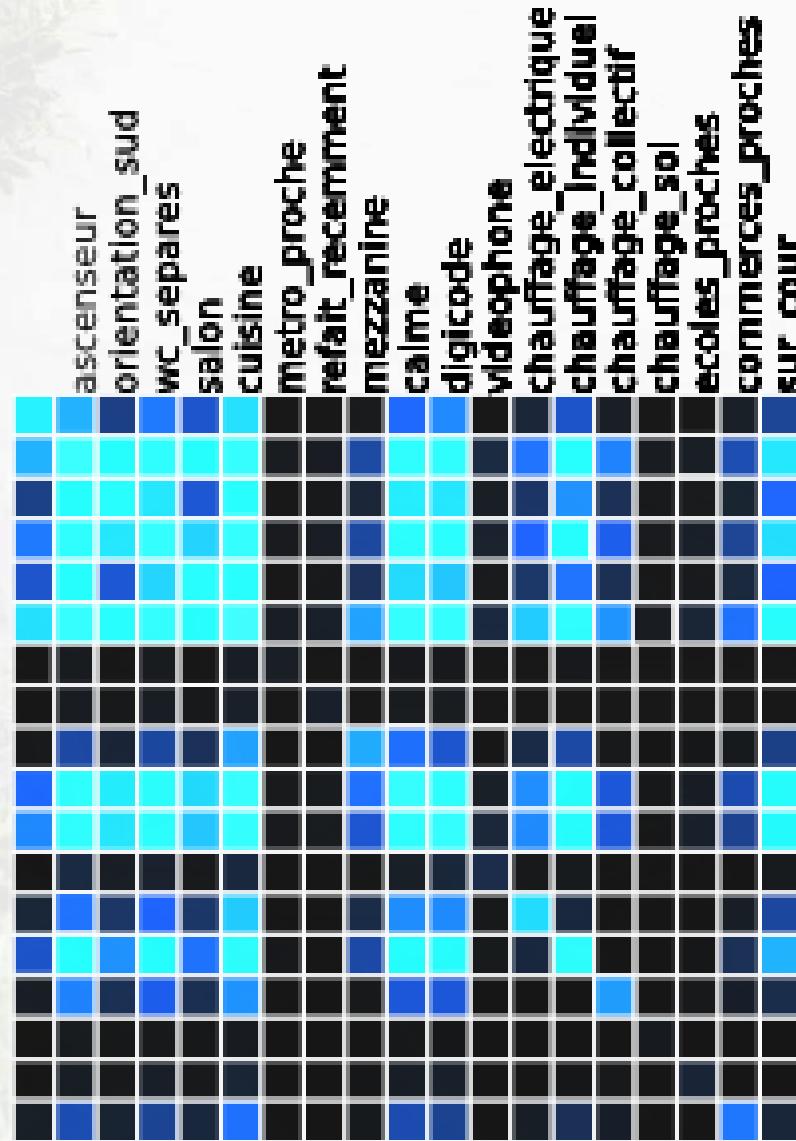
Pesto Webmining

Pré-traitement

haussmann
ascenseur
orientation_sud
wc_separes
salon
cuisine
metro_proche
refait_recemment
mezzanine
calme
digicode
videophone
chauffage_electrique
chauffage_Individuel
chauffage_collectif
chauffage_sol
ecoles_proches
commerces_proches
sur_cour
vue_toits
porte_blindees
poutres_apparentes
parquet
balnoire
cuisine_moderne
cuisine_americaine
cuisine_equipes
sans_vis_a_vis
balcon
cinemas_proches
cheminee
prevoir_travaux
parking
cave
rangements
immeuble_pierre
gardien
dressing
lumineux
moulures
double_vitrage
ravalements_recents
buanderie
chauffage_gaz
parking_public_proche
debarras
jardin
maison
appartement
entree
plan_etaille
traversant
terrasse
piscine
terrains_sport_proches
meuble
etat_general
etat_copropriete
studio
dernier_etage
local_velos
sloger
explrimmo
faim
forêt
pap
paruvendo
coordonnees_précises



haussmann
ascenseur
orientation_sud
wc_separes
salon
cuisine
metro_proche
refait_recemment
mezzanine
calme
digicode
videophone
chauffage_electrique
chauffage_individual
chauffage_collectif
chauffage_sol
ecoles_proches
commerces_proches



Prétraitement des données

Analyse des correspondances

Intérêt pour l'apprentissage ? Le jeu des devinettes.

- ▶ Compléter une annonce incomplète pour mieux prédire son prix.
- ▶ Compléter les caractéristiques manquantes dans la base de données ?



Etapes du projet

- 
- I. Web crawling
 - II. Traitement de texte
 - III. Visualisation des données
 - IV. Pré-traitement des données
 - V. Machine learning**
 - VI. Site Web



Principes du ML

- ▶ Objectif :

$$X \xrightarrow{f} Y$$

- ▶ Avec :

- X : caractéristiques des annonces
- Y : prix (étiquette)
- f : prédicteur

- ▶ Processus :

- ▶ Traitement de la base
- ▶ Apprentissage
- ▶ Visualisation des erreurs
- ▶ Clustering



Pourquoi traiter la base?

Appartement 28 pièce(s) - 1 m² - PARIS 10



Exclusivité

219 000 €

Appartement 28 pièce(s) - 1 m²

PARIS 10 75010

Ref : A599923

Ce bien vous est proposé par :

PARIS EST IMMOBILIER



ente Appartement 3 pièces 30m² Paris 11ème

↑39 150 €

À partir de 115 € /mois¹

3 Pièces

2 Chambres

30 m²



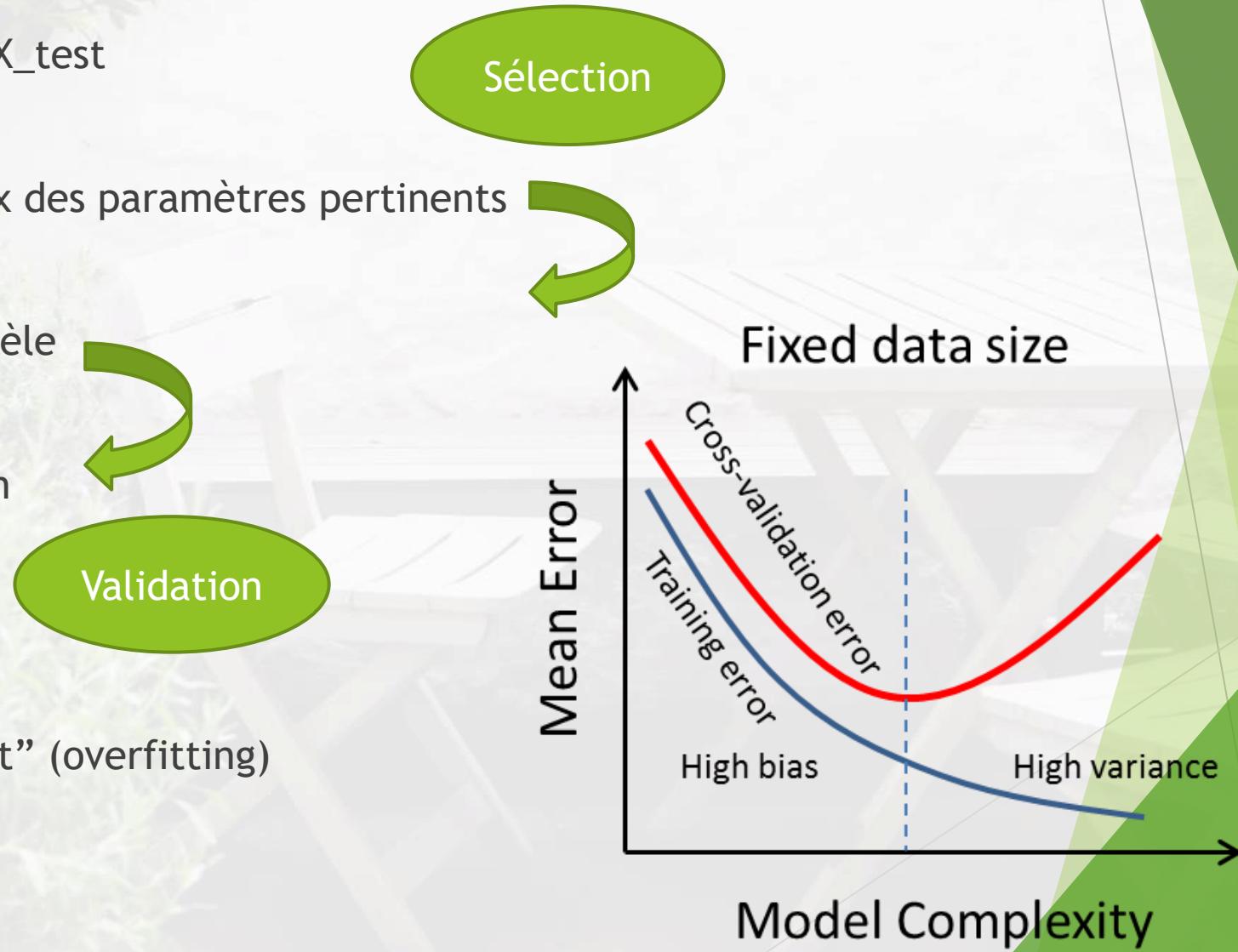
Description de Appartement à Paris 11ème

A saisir, vente Mobil Home, 2Ch + location parcelle de terrain, dans un camping classée 2*, site extraordinaire, satisfaisant tous vos besoins avec les commerces de proximité, à 45 min de Paris



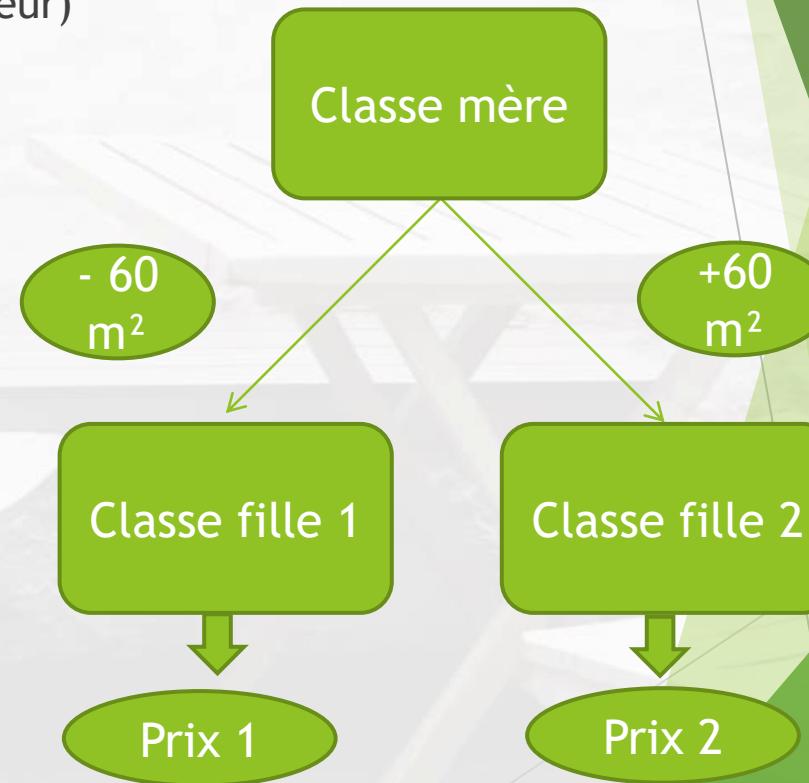
L'apprentissage en pratique

- ▶ Trois phases : X_train et X_test
 - ▶ Cross-validation: choix des paramètres pertinents
 - ▶ Apprentissage du modèle
 - ▶ Test sur un échantillon
- ▶ Risque de “sur-ajustement” (overfitting)



Un premier algorithme : l'arbre

- ▶ Idée : Séparer les données pour regrouper les biens avec un prix semblable
 - ▶ Choix d'un critère optimal : couple (coordonnées, valeur)
 - ▶ Exemple : surface
- ▶ Prix des biens d'une classe: moyenne de leurs prix
- ▶ Problèmes majeurs : instabilité et complexité



Améliorations



- ▶ Derrière l'arbre se cache...
... la forêt!
- ▶ **Idée :** Utiliser des échantillons tirés
aléatoirement pour créer de
nombreux arbres et les **agréger**
- ▶ Diminution de l'instabilité et du
“sur-ajustement”



+ Stabilité



+ Précision

- ▶ Boosting :
 - ▶ On donne des poids relatifs aux erreurs pour “forcer” l’algorithme à mieux les estimer
 - ▶ Raffinement par itération successive



Méthode choisie

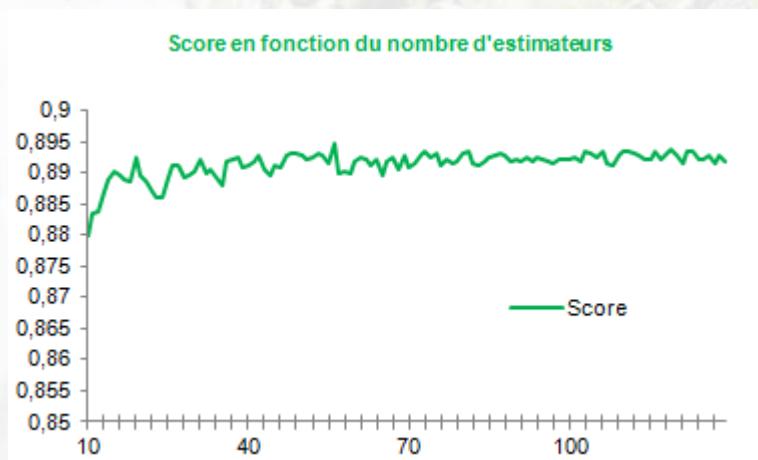
- ▶ 1ère étape : utilisation de **clusters** → spécificité des arrondissements
 - ▶ Par arrondissement: régression linéaire ou apprentissage: $r(X) \sim Y$
 - ▶ Sur l'ensemble des données: apprentissage avec les erreurs en étiquettes
 - ▶ On cherche : $F(X) \sim Y - r(X)$
- ▶ 2ème étape :
 - ▶ Sur l'ensemble des données: apprentissage global avec le prix en étiquette
 - ▶ On cherche: $G(X) \sim Y$
- ▶ Combinaison pondérée des deux prédicteurs obtenus : $0.5 * (G(X) + F(X) + r(X))$



Résultats

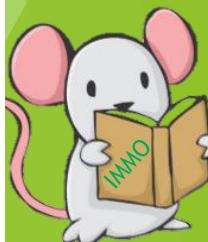
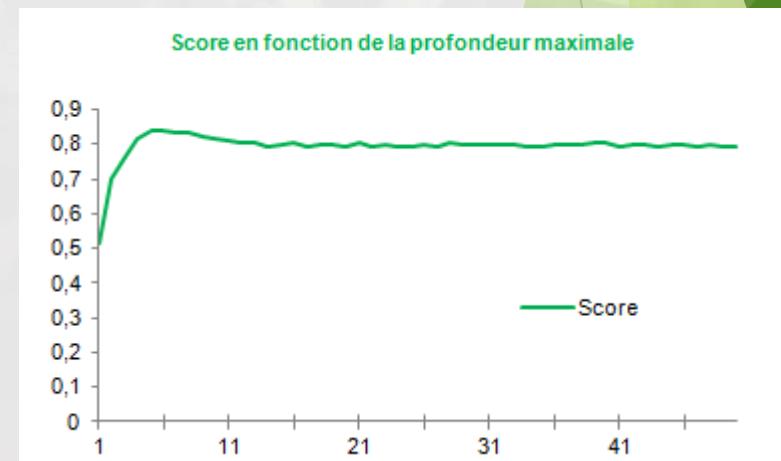
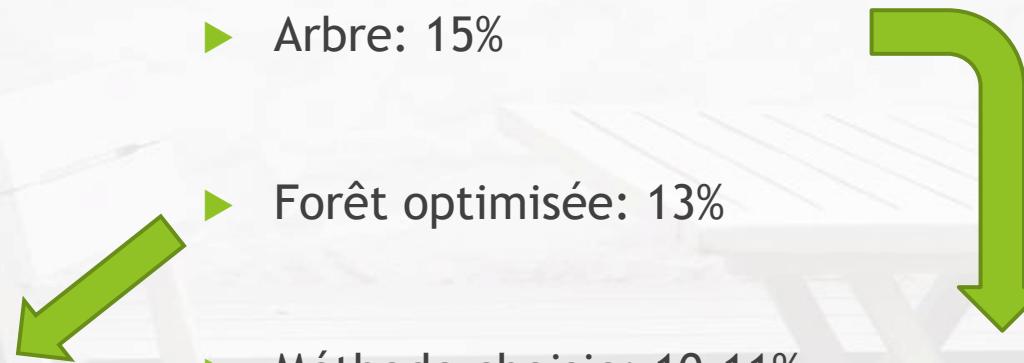
Régressions linéaires

- ▶ Globale: erreur moyenne 25%
- ▶ Par arrondissement: 17%

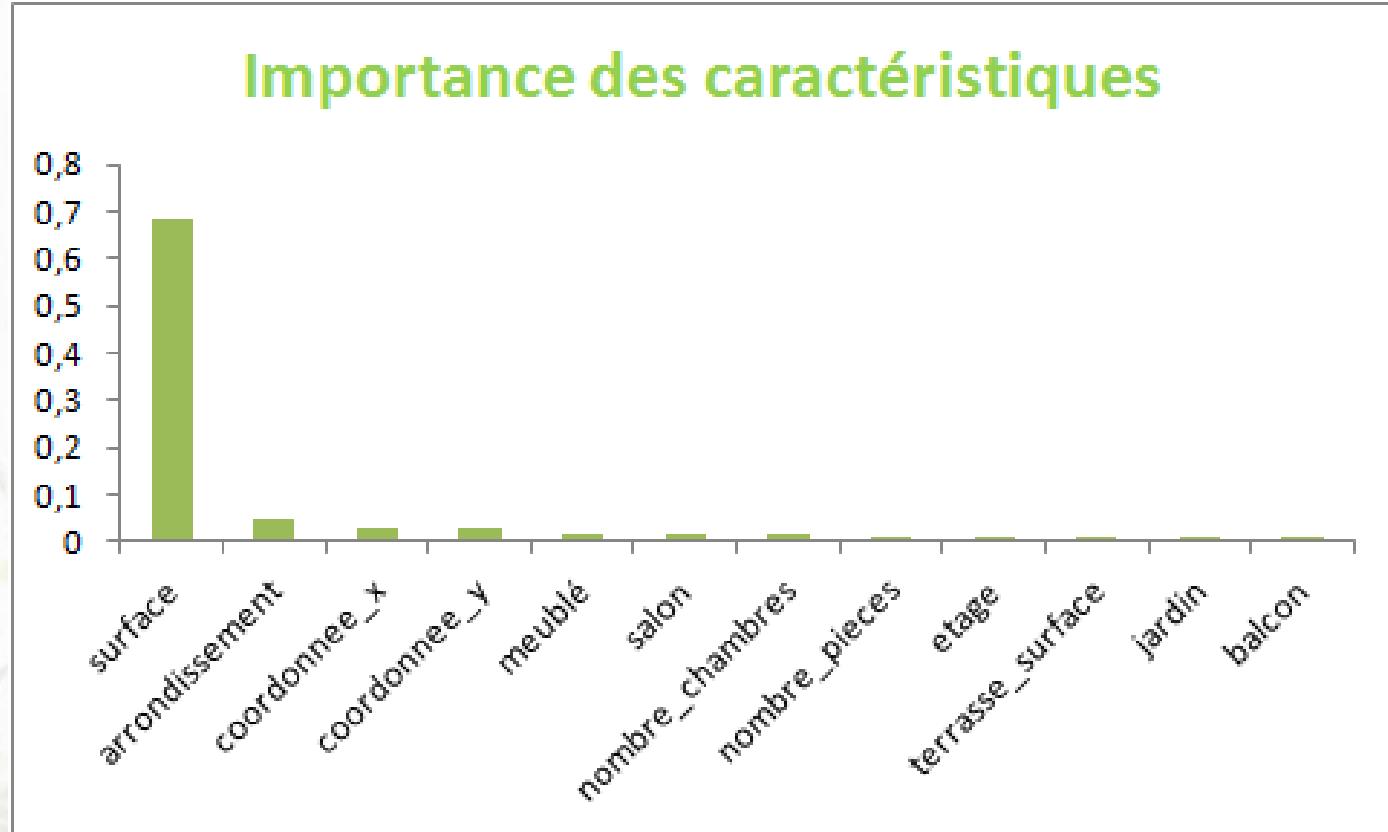


Machine learning

- ▶ Arbre: 15%
- ▶ Forêt optimisée: 13%
- ▶ Méthode choisie: 10-11%



Caractéristiques pertinentes



Utilisation pour le site

- ▶ Objets :
 - ▶ Prédicteurs stockés en “.pickle” pour un appel direct par le serveur
- ▶ Fonctions :
 - ▶ Prédire le prix d'un bien (URL ou formulaire)
 - ▶ Fournir des biens “comparables” (fonction de conseil)



Limites et difficultés rencontrées

- ▶ Qualité des annonces et de la base de données:
 - ▶ Annonce mal ou insuffisamment remplie
 - ▶ Deux annonces “identiques” selon nos algorithmes peuvent avoir un prix très différents (surtout pour les biens à prix faibles)
 - ▶ informations hors textes non prises en compte (photos par exemple)



Etapes du projet

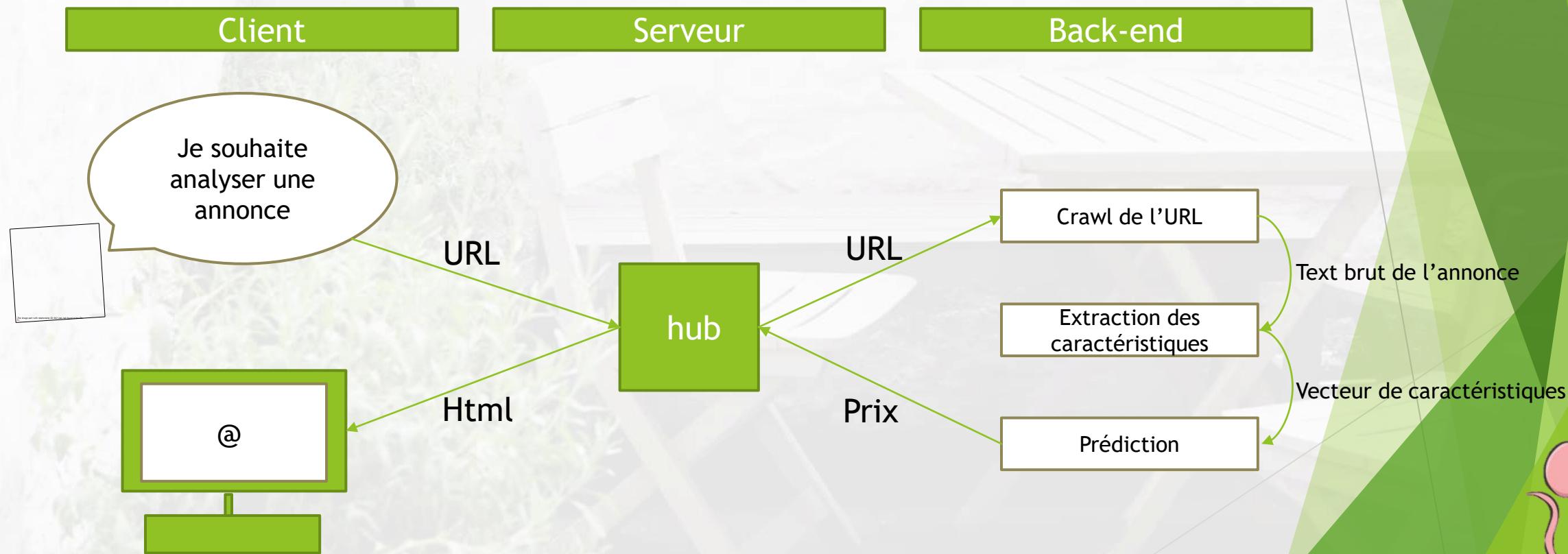


- I. Web crawling
- II. Traitement de texte
- III. Visualisation des données
- IV. Pré-traitement des données
- V. Machine learning
- VI. Site Web

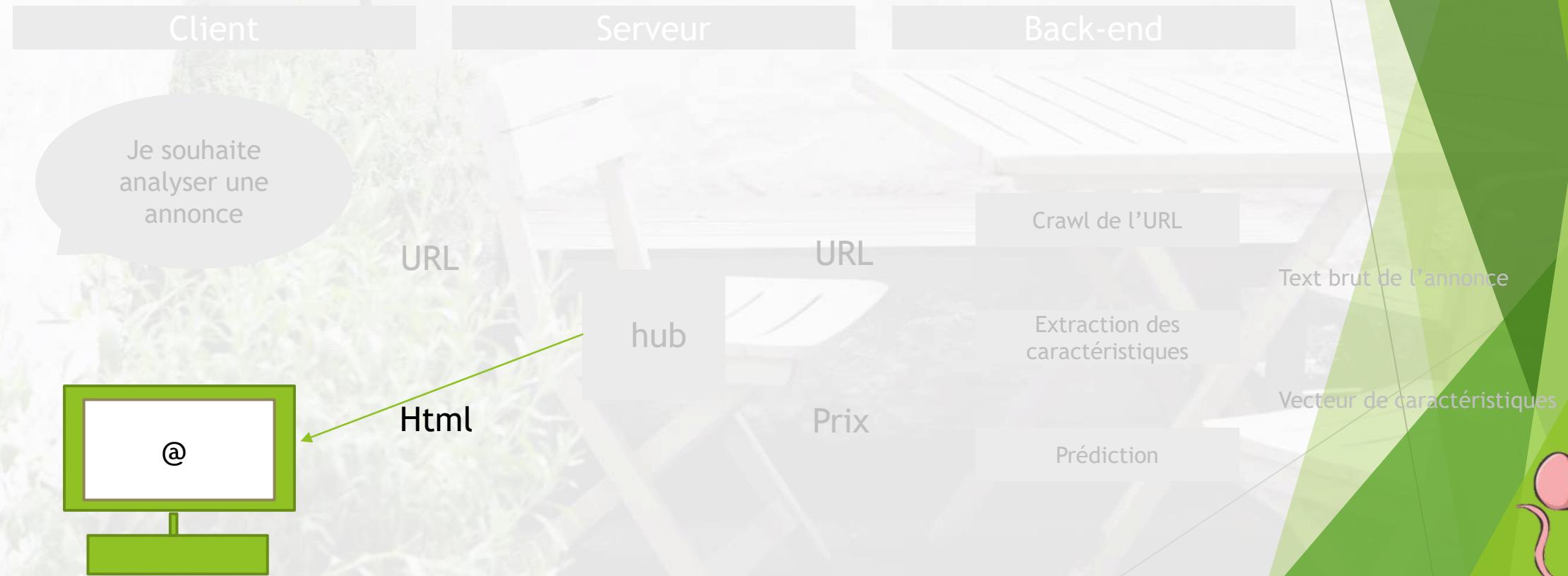


Architecture de notre site

Que se passe-t-il lorsqu'un utilisateur se connecte à FlatRat ?



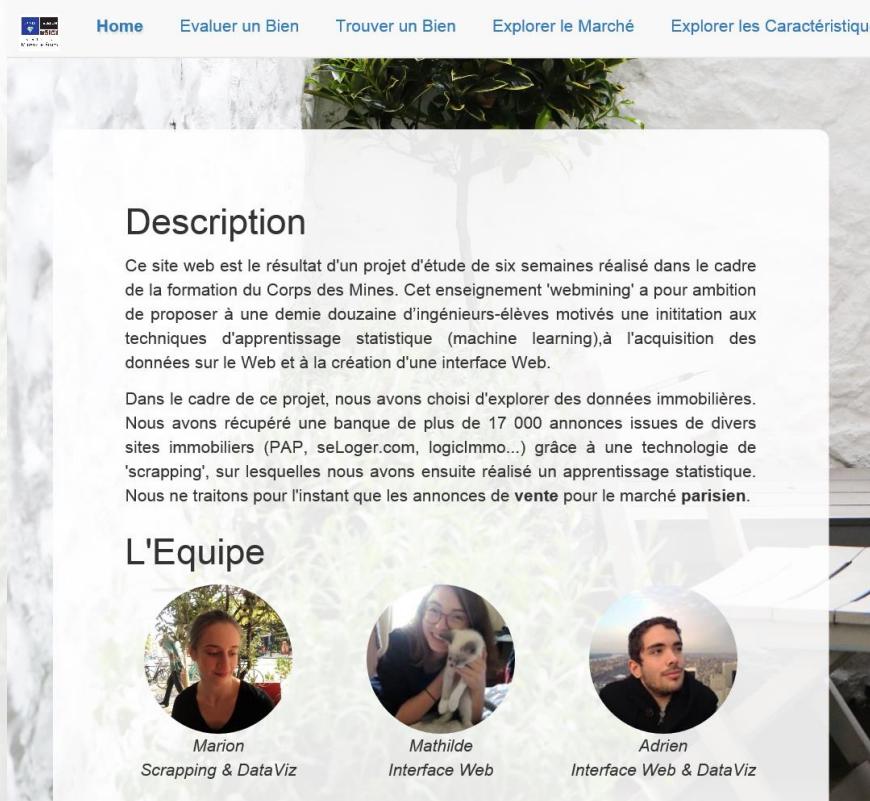
Comment s'affiche un page web ?



Comment s'affiche un page web ?

Le Html et CSS

- ▶ Deux langages principaux permettent de créer des sites web : le html et le CSS
- ▶ Les navigateurs sont capables d'interpréter ces deux langages



```
<!DOCTYPE html>
<html>
  <head>...</head>
  <body>
    <nav class="navbar navbar-default navbar-fixed-top">...</nav>
    <div class="container-fluid">
      <div class="blocTop">
        <div class="row">
          <div class="col-md-12 ">
            <h1> FlatRat </h1>
          </div>
        </div>
        <br><br>
      <div class="row">
        <div class="col-md-6">
          <div class="bloc" style="height: 850px;" ;="">
            <h2> Description </h2>
            <p> Ce site web est le résultat d'un projet d'étude de six semaines réalisé dans le cadre de la formation du Corps des Mines. Cet enseignement 'webmining' a pour ambition de proposer à une demie douzaine d'ingénieurs-élèves motivés une initiation aux techniques d'apprentissage statistique (machine learning), à l'acquisition des données sur le Web et à la création d'une interface Web. </p>
            <p> Dans le cadre de ce projet, nous avons choisi d'explorer des données immobilières. Nous avons récupéré une banque de plus de 17 000 annonces issues de divers sites immobiliers (PAP, seLoger.com, logicImmo...) grâce à une technologie de 'scrapping', sur lesquelles nous avons ensuite réalisé un apprentissage statistique. Nous ne traitons pour l'instant que les annonces de vente pour le marché parisien. </p>
          </div>
        </div>
        <br><br>
      <div class="row">
        <div class="col-md-4">
          <h2> L'Equipe </h2>
          <div class="row">
            <div class="col-md-4">
              <img alt="Marion's profile picture" data-bbox="198 828 251 921"/>
              Marion
              Scrapping & DataViz
            </div>
            <div class="col-md-4">
              <img alt="Mathilde's profile picture" data-bbox="281 828 334 921"/>
              Mathilde
              Interface Web
            </div>
            <div class="col-md-4">
              <img alt="Adrien's profile picture" data-bbox="364 828 417 921"/>
              Adrien
              Interface Web & DataViz
            </div>
          </div>
        </div>
      </div>
    </div>
  </body>
</html>
```



Comment s'affiche un page web ?

Que font ces deux langages ?



- ▶ Html permet d'afficher du contenu et la hiérarchie des éléments
- ▶ CSS permet d'ajouter du style à la page

Estimation par Critères

Surface en mètres
 m²

Nombre de pièces
 pièce(s)

Arrondissement
 ème

Etage
 ème

Balcon
 Oui Non

Ascenseur
 Oui Non

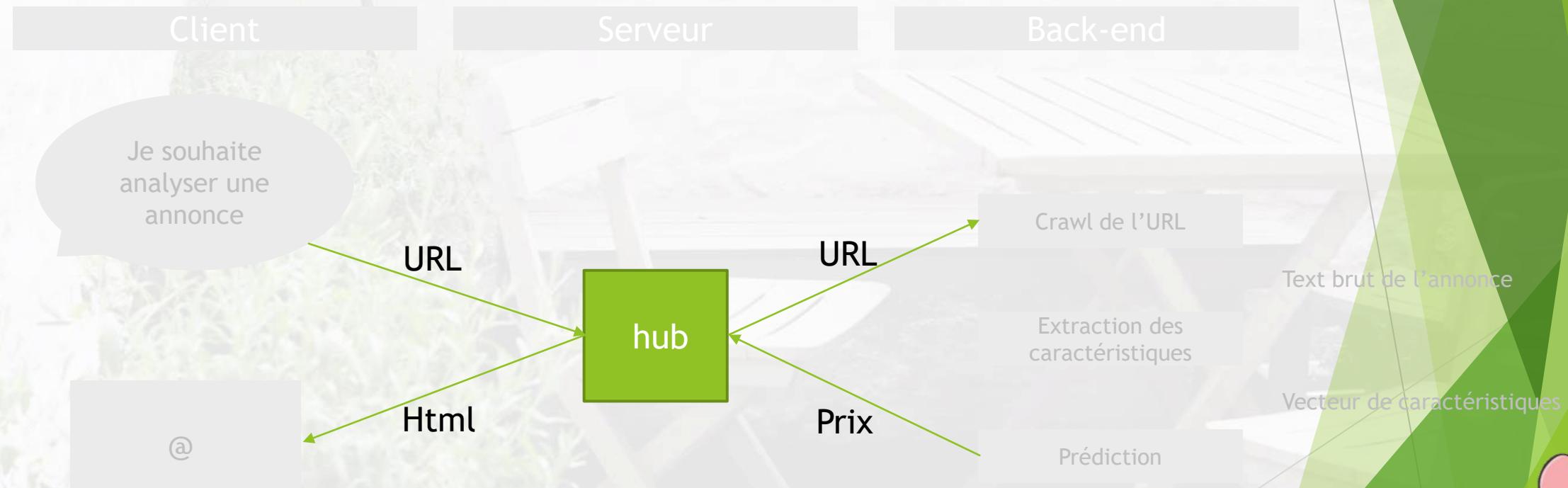
Estimation par Critères

Surface en mètres <input type="text" value="100"/> m ²	Arrondissement <input type="text" value="20"/> ème	Balcon <input type="radio"/> Oui <input type="radio"/> Non
Nombre de pièces <input type="text" value="5"/> pièce(s)	Etage <input type="text" value="3"/> ème	Ascenseur <input type="radio"/> Oui <input type="radio"/> Non



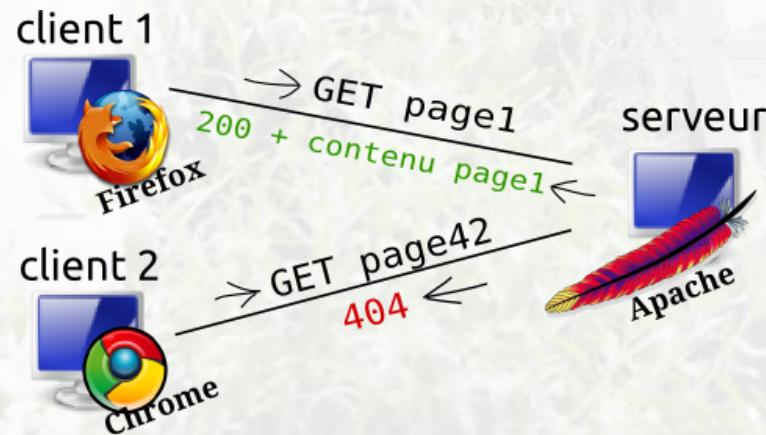
Quel est le rôle du serveur ?

Que se passe-t-il lorsqu'un utilisateur se connecte à FlatRat ?



Comment la page web communique-t-elle avec le serveur ?

- ▶ Le navigateur effectue une **requête HTTP**
- ▶ Le serveur traite la requête puis envoie une **réponse HTTP**



```
#!/usr/bin/python
# -*- coding:utf-8 -*-
from flask import Flask
app = Flask(__name__)

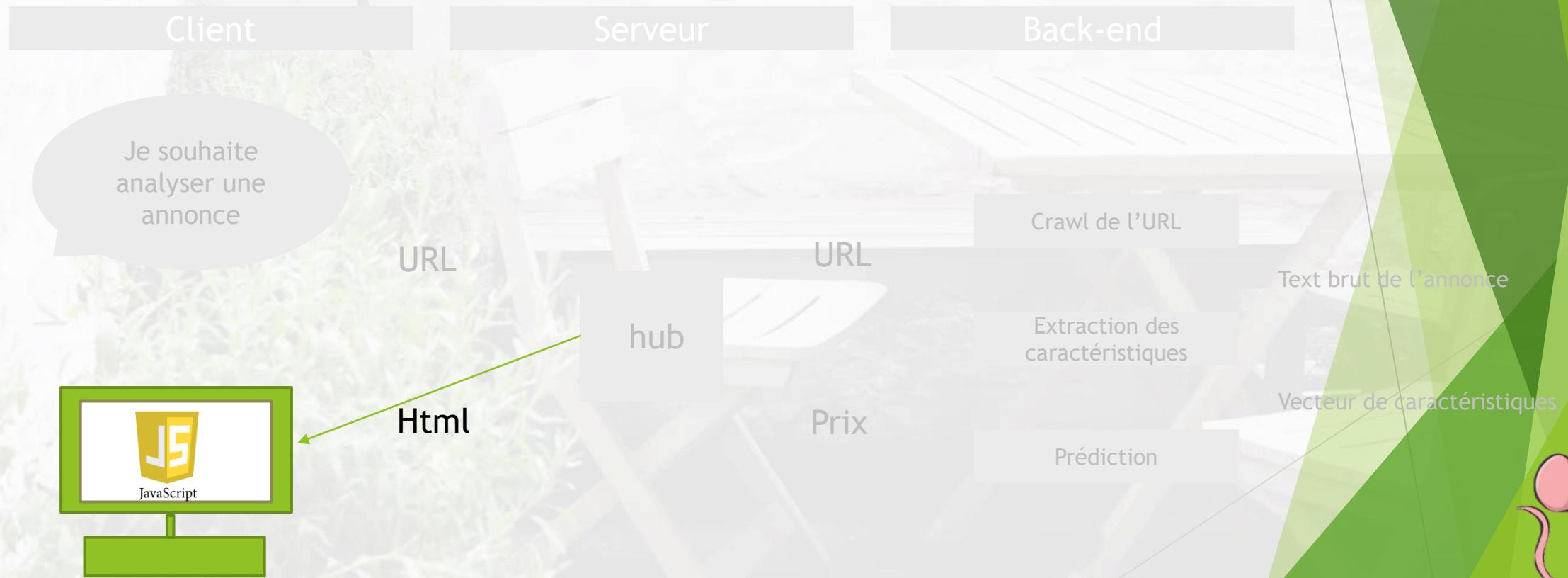
@app.route('/page1')
def accueil():

    return render_template('page1.html', titre="Bienvenue !")

if __name__ == '__main__':
    app.run(debug=True)
```



Comment faire pour ne modifier que certaines parties de la page web ?



Programmation côté client : le Javascript



Script interprété par
le navigateur internet

A screenshot of a Windows-style code editor showing an HTML file. The script section is highlighted with a red border.

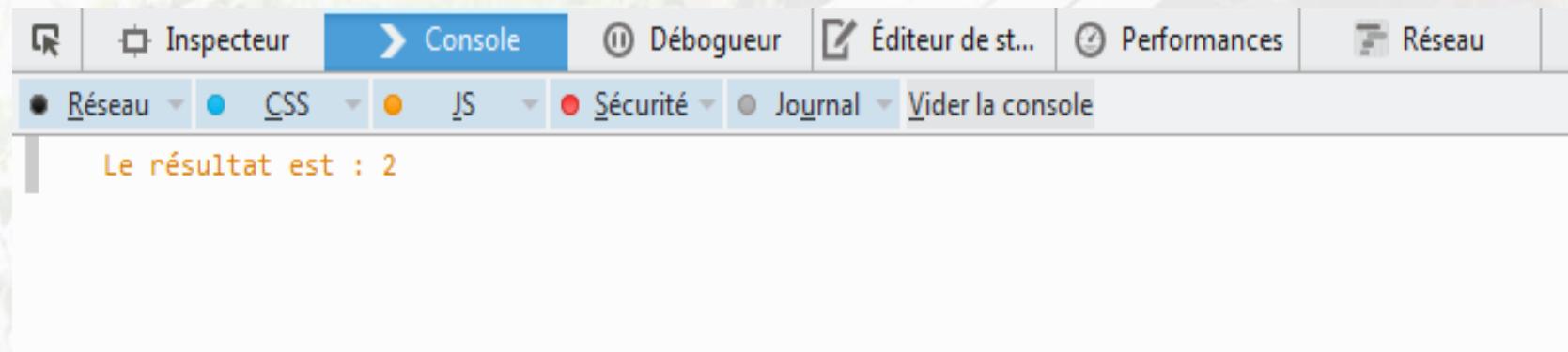
```
File Edit Format View Help
<!DOCTYPE HTML>
<HTML>
  <HEAD>
    <TITLE>A First Script</TITLE>
    <SCRIPT LANGUAGE = "Javascript">
      confirm("OK or Cancel?")
    </SCRIPT>
  </HEAD>
  <BODY>
  </BODY>
</HTML>
```



Exemple Basique de Javascript

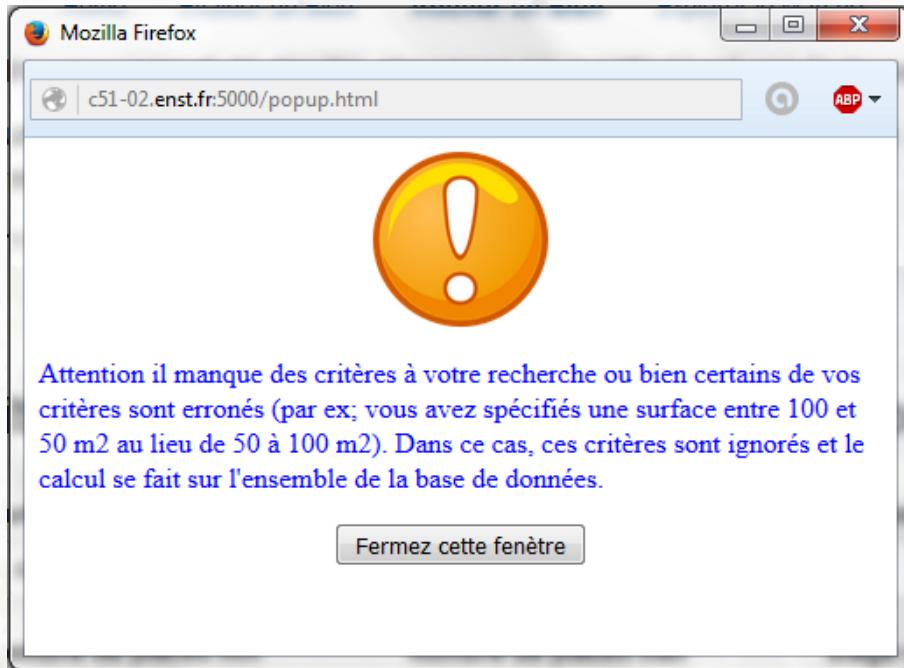
Rien de plus qu'un langage classique de programmation !

```
<script>          var a = 1 ;
                  var b = 1 ;
                  var c = a + b ;
console.log("Le résultat est : " + c) ;      </script>
```



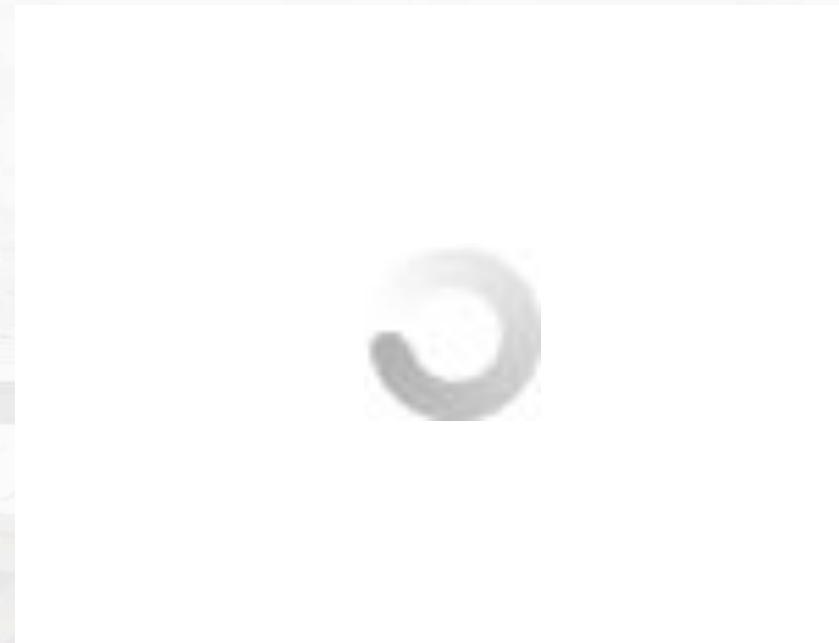
Affichage d'éléments dynamiques

Pop up d'erreur



Code javascript :
window.open(...)

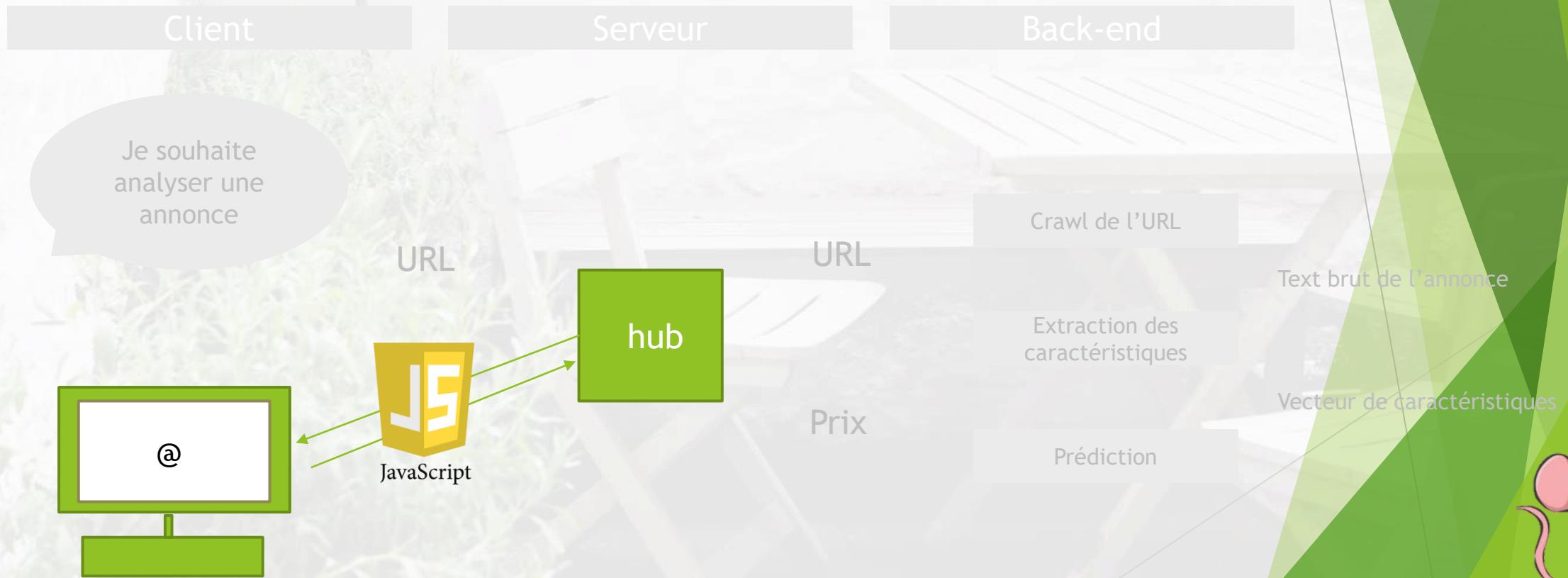
Logo de chargement



Code javascript :
.\$('#loadingimg').show
.\$('#loadingimg').hide



Comment s'affiche un page web ?



Exemple de Jquery sur notre site

Que se passe-t-il quand un utilisateur lance une requête... le retour !

Formulaire HTML

Estimation par Critères

Surface en mètres	Arrondissement	Balcon
100	20	<input type="radio"/> Oui <input type="radio"/> Non
Nombre de pièces	Etage	Ascenseur
5	3	<input type="radio"/> Oui <input type="radio"/> Non
Estimer		



Script activé par le clic
sur le bouton

qui communique avec le
serveur

et crée des éléments
dynamiquement



Exemple sur le site web (librairie jQuery)

The screenshot shows a real estate website interface. At the top, there is a navigation bar with links: Home, **Evaluer un Bien**, Trouver un Bien, Explorer le Marché, Explorer les Caractéristiques, and Eléments Techniques. A logo for "IMMO" is visible in the top left corner.

Estimation par URL: A form to enter the URL of a property for sale in Paris. It includes a "URL" input field and a "Sites supportés" section with logos for Explorimmo, FNAIM, laforêt, pap (Particulier à Particulier), ParuVendu, and SeLoger.com. A blue "Estimer" button is present.

Estimation par Critères: A form to estimate a property based on specific criteria. It includes fields for Surface en mètres (100 m²), Arrondissement (20 ème), Balcon (radio buttons for Oui and Non), Nombre de pièces (5 pièce(s)), Etage (3 ème), Ascenseur (radio buttons for Oui and Non), and a blue "Estimer" button.

Distribution des biens immobiliers à Paris: A histogram showing the distribution of property prices in Paris. The x-axis represents price, and the y-axis represents frequency. The distribution is roughly bell-shaped, peaking around 1 million euros.



Exemple sur le site web (librairie jQuery)

The screenshot shows a real estate valuation website with two main sections:

- Estimation par URL**: A form where users enter the URL of a property for valuation. It supports various real estate websites like Explor'immo, Fnaim, laforêt, pap, ParuVendu, and SeLoger.
- Estimation par Critères**: A form for estimating a property based on specific criteria: Surface en mètres (100 m²), Arrondissement (20 ème), Balcon (Oui), Nombre de pièces (5 pièces(s)), Etage (3 ème), Ascenseur (Non), and a valuation button "Estimer".

To the right, a green box displays the results of the valuation:

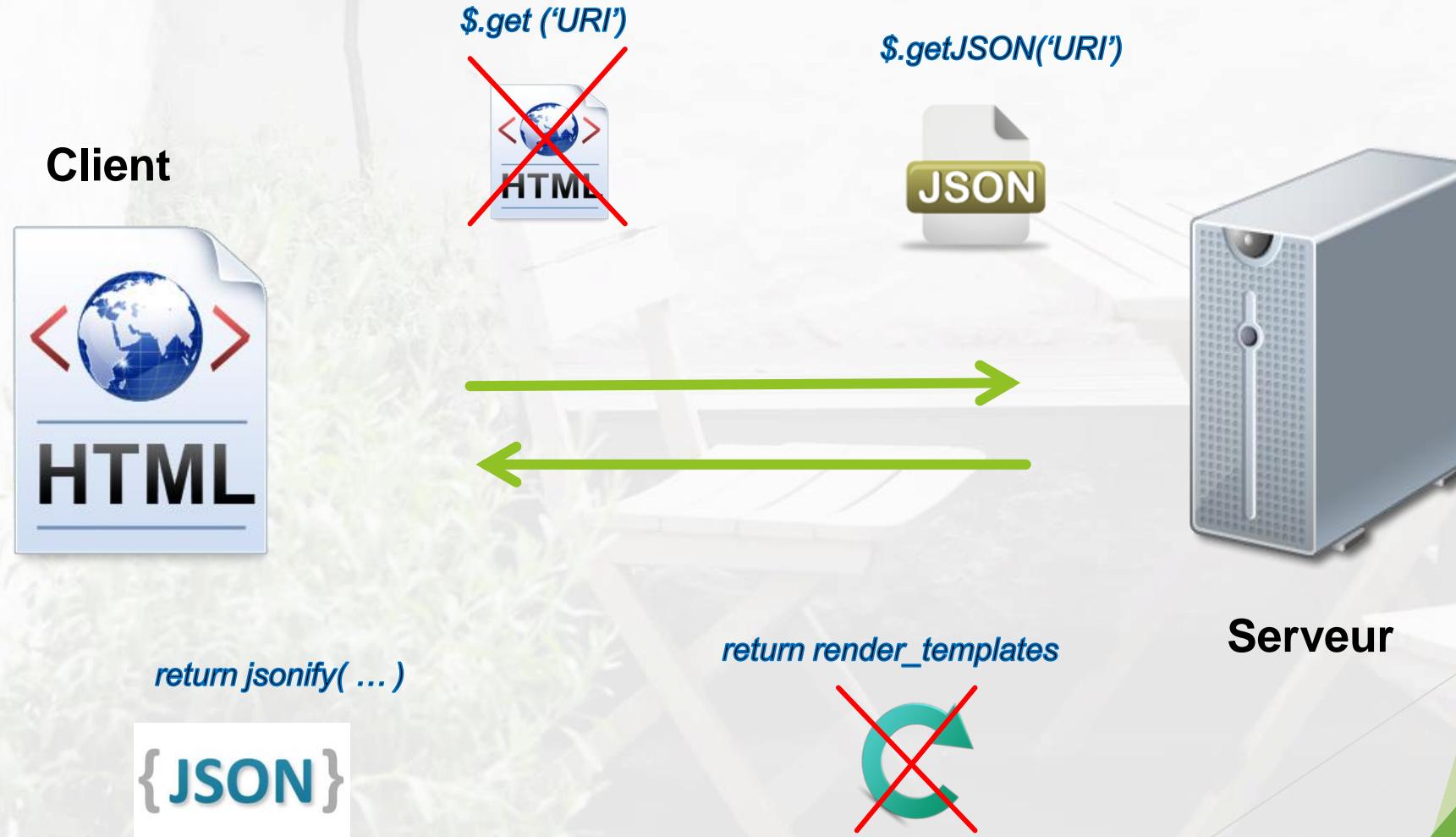
- Distribution des biens immobiliers à Paris**: A histogram showing the distribution of property values.
- Seul le paragraphe est modifié !**: A text box containing the modified paragraph: "Nous estimons votre bien à 753 853 €".
- Biens comparables à votre recherche :** A list of comparable property URLs:
 - http://www.seloger.com/annonces/achat/appartement/paris-20eme-75/saint-blaise/100560921.htm?ci=750120&idqfix=1&idtt=2&idtypebien=1&listing-listpg=43&tri=d_dt_crea&bd=LienAnn_1
 - <http://www.fnaim.fr/annonce-immobiliere/36181313/17-acheter-appartement-paris-20-75020.htm>
 - http://www.seloger.com/annonces/achat/appartement/paris-20eme-75/gambetta/102173081.htm?ci=750120&idqfix=1&idtt=2&idtypebien=1&listing-listpg=34&tri=d_dt_crea&bd=LienAnn_1

A black arrow points from the text "Seul le paragraphe est modifié !" to the modified valuation paragraph.

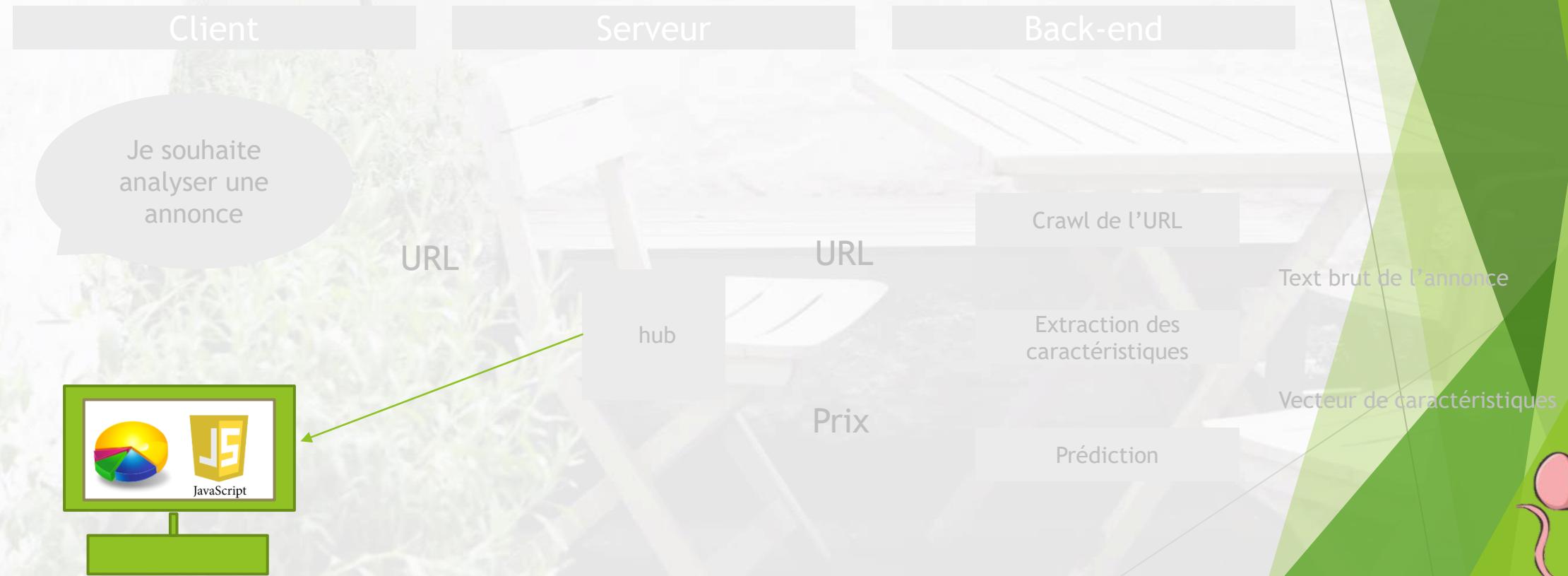


Technique AJAX sur notre site web

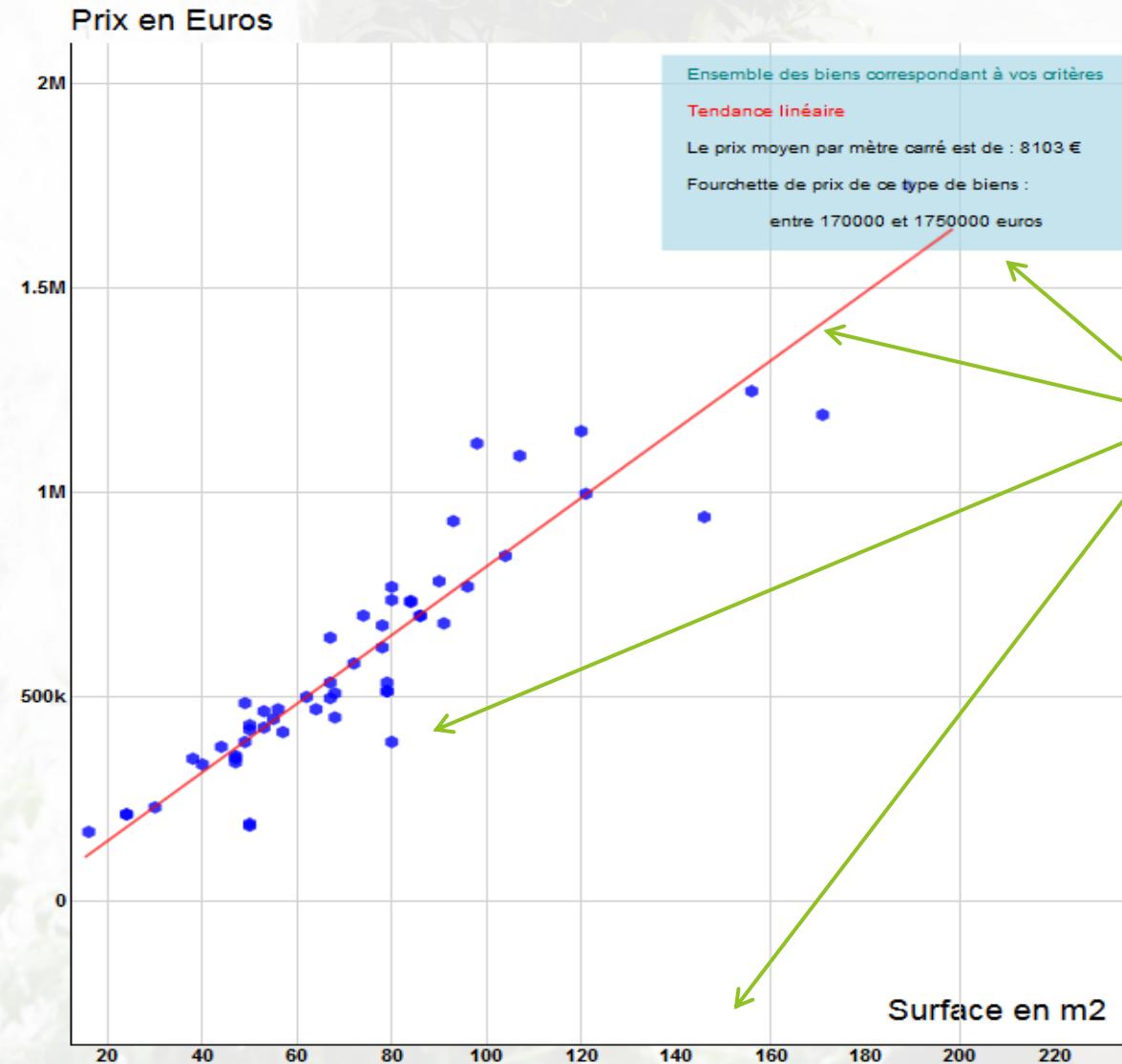
Des commandes asynchrones côté client et côté serveur



Comment créer des éléments complexes ?



Exemple d'application :



Eléments SVG

Attributs :

- ▶ Position
- ▶ Taille
- ▶ Couleur
- ▶ ...



HTML/CSS vs Javascript/AJAX ?

- ▶ Possibilité de construire le site en JS uniquement
- ▶ Choix en fonction des pages web

[Home](#) [Evaluer un Bien](#) [Trouver un Bien](#) [Explorer le Marché](#) [Explorer les Caractéristiques](#)

Estimation par URL

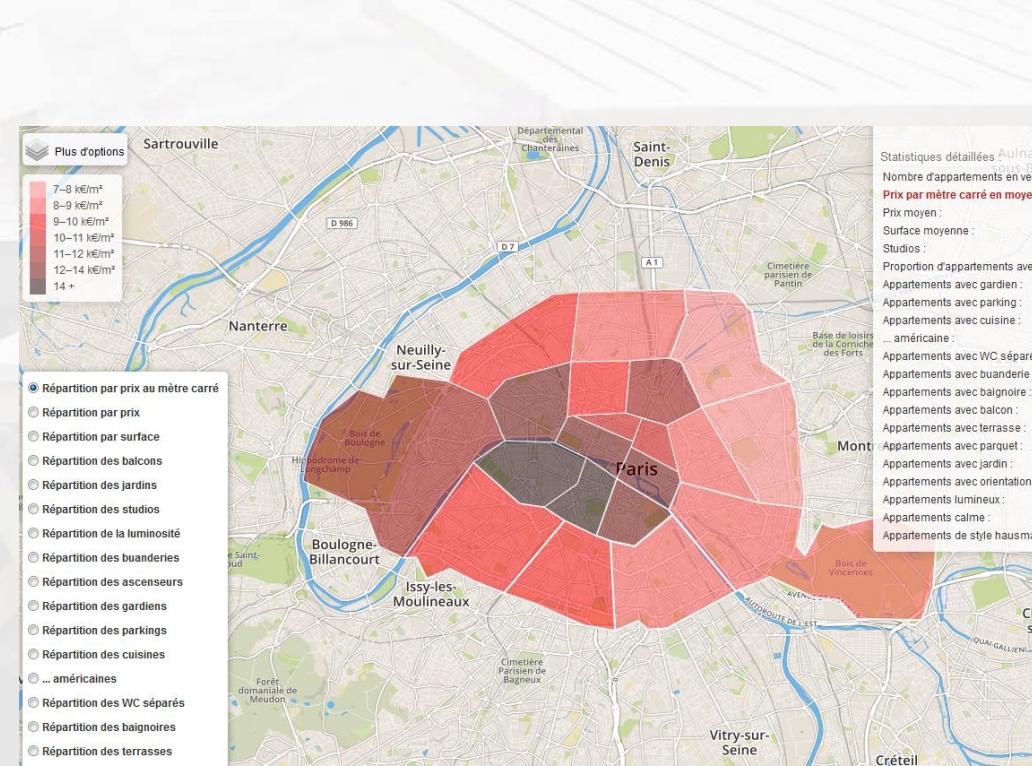
Entrez l'URL d'un bien à vendre à Paris

URL

Sites supportés : [Explorimmo](#) [Pap](#) [ParuVendu](#) [Seloger](#)

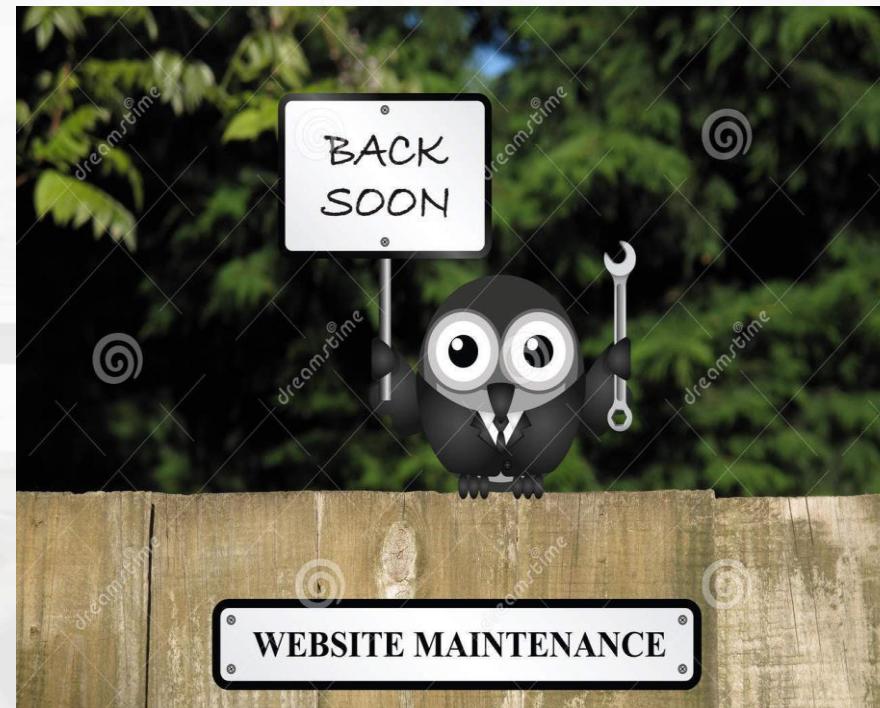
Estimation par Critères

Surface en mètres	Arrondissement	Balcon
100 <input type="button" value="m2"/>	20 <input type="button" value="ème"/>	<input type="radio"/> Oui <input type="radio"/> Non
Nombre de pièces	Etage	Ascenseur
5 <input type="button" value="pièce(s)"/>	3 <input type="button" value="ème"/>	<input type="radio"/> Oui <input type="radio"/> Non



Conclusion : site web / serveur

- ▶ Collaboration permanente entre les différentes équipes du projet
- ▶ Réflexion centrée sur l'utilisateur
- ▶ Face visible du projet



Remerciements

Un grand merci à l'ensemble de l'équipe enseignante de Télécom, en particulier :

- ▶ Stéphan Cléménçon
- ▶ Cyril Concolato

pour leur attention bienveillante.

Mais aussi :

- ▶ Chloé Clavel, Alexandre Gramfort, Slim Essid, Maxime Sangnier, Jean-Claude Dufourd, ...

pour les cours et les conseils.

