Outils d’extraction, de transformation et de chargement (ETL)

C’est quoi ?

Outils de gestion, transformation et préparation des données pour des analyses ultérieures

Simplifie la préparation et la gestion des données

→ Alliés incontournables pour naviguer efficacement dans le cycle de vie des données

Pourquoi les utiliser ?

Problèmes courants des étudiants et solutions apportées par les ETL

Problème : Données désordonnées ou incomplètes

- Exemple : Colonnes mal nommées, données manquantes ou doublons dans les résultats d'enquêtes.
- **Solution** : Nettoyer, normaliser et combler les lacunes

Problème : Gestion de grandes bases de données

- Exemple : Bases contenant des centaines \ milliers de lignes
- **Solution** : Traiter efficacement des données volumineuses

Pourquoi les utiliser ?

Problèmes courants des étudiants et solutions apportées par les ETL

Problème : Difficulté à reproduire les analyses

- Exemple : Etapes de traitement non enregistré, rendant l'analyse non reproductible
- **Solution** : générer du code reproductible, favorisant la transparence et la traçabilité

Problème : Manque de compétences en codage

- Exemple : Étudiants ayant peu de connaissances en programmation
- **Solution** : générer interface visuelle intuitive sans nécessiter de code.

Problème : Hétérogénéité des formats de données

- Exemple : Importation de données à partir de plusieurs sources avec des formats différents.
- **Solution** : convertir et d'uniformiser les formats.

ETLs présentés

OpenRefine

powerdrill

Logiciel déterministe

Basé sur l’intelligence artificielle

OpenRefine

powerdrill

https://powerdrill.ai

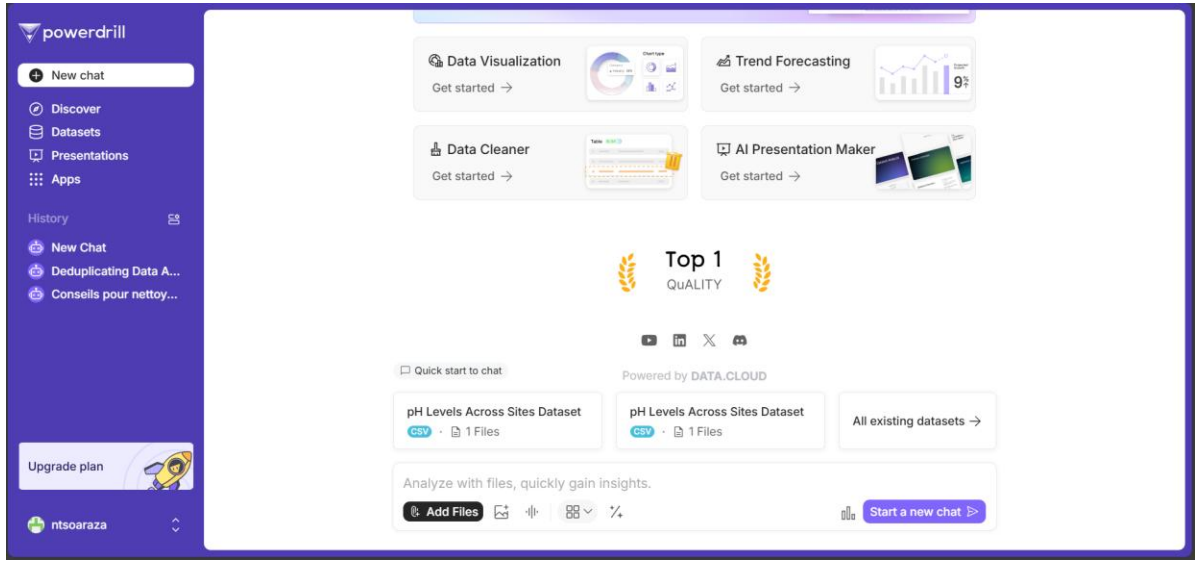
Compte google / GitHub

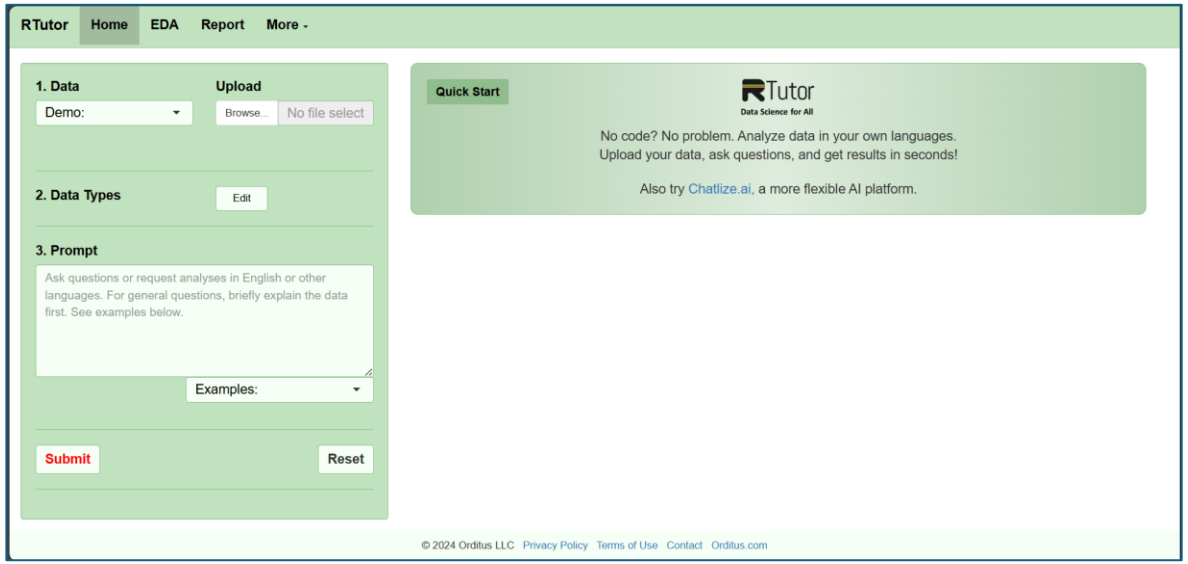
RTutor

Data Science for All

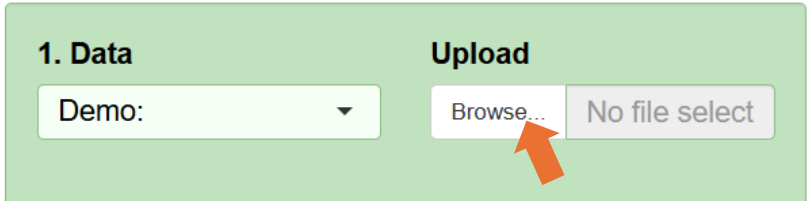
https://rtutor.ai/

Pas besoin de compte

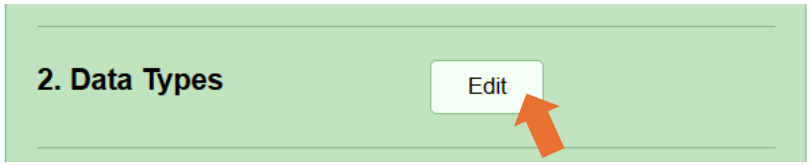




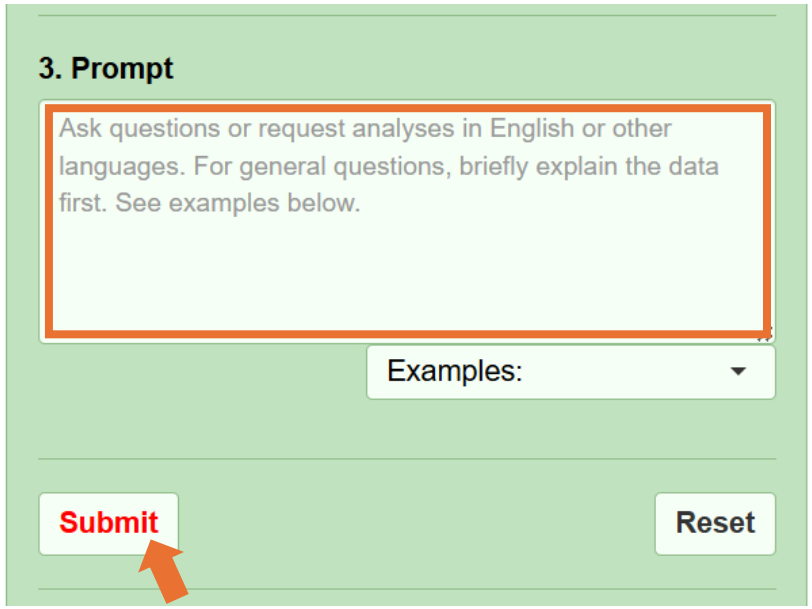
RTutor
Data Science for All



Importe la table



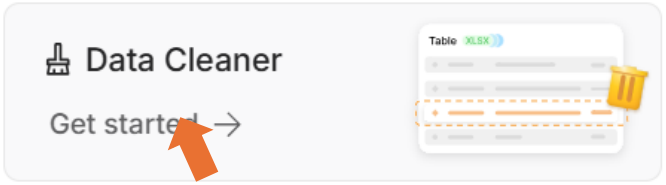
Affiche structure interne de la table



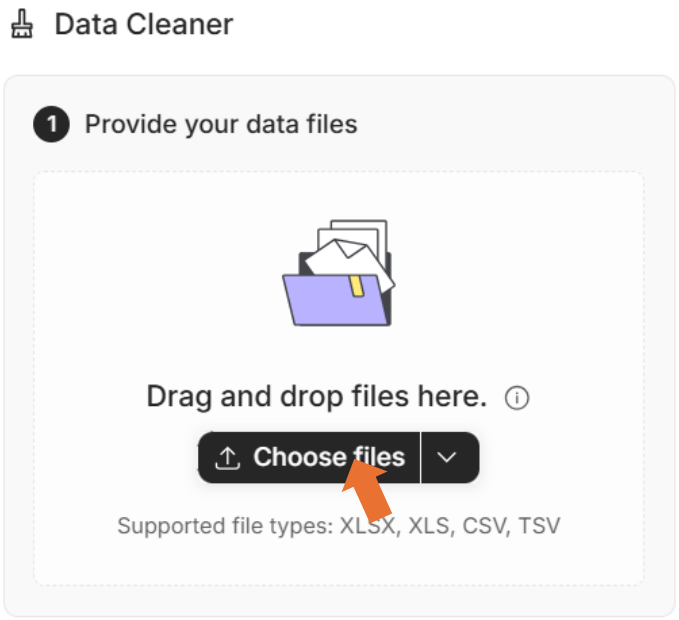
1. Colonne site, remplacer 22 et A4E en A1 et A4
2. Des outliers dans p_h ?
3. refaire le tableau: enlève NA et remplace les outliers dans pH avec la moyenne de pH
4. exporter la table en csv

powerdrill

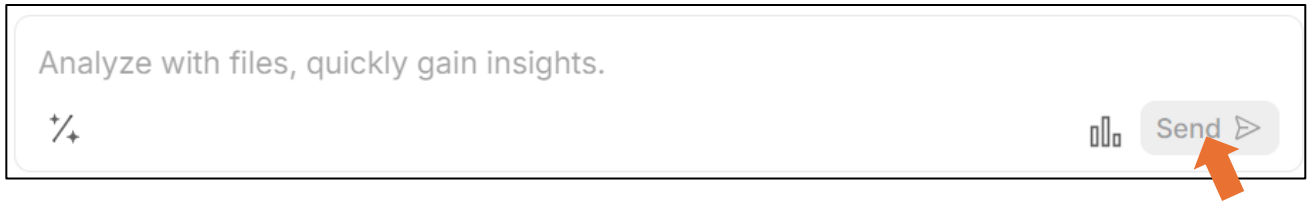
Section nettoyage de données



Uploader le fichier



Nettoyage de données



Veuillez remplacer dans la colonne site 22 et A4E par A1 et A4

Y a-t-il des valeurs aberrantes dans la colonne **pH** ?

Evaluations des trois outils

OpenRefine

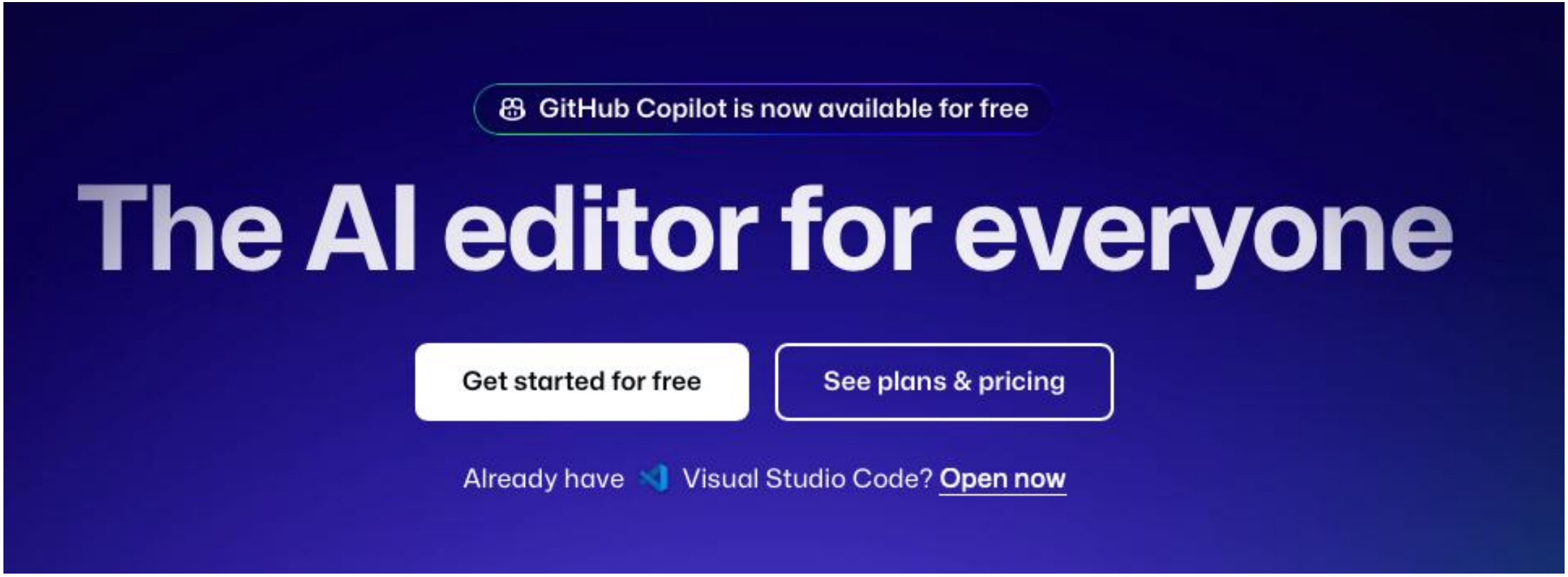
- Interface très intuitive
- Facilité de prise en main
- Open source

powerdrill

- Accès avec GitHub et google mais très limité
- besoin de requête très precise

- Utilisateurs de R
- Nettoyage traçable … peut être chronophage
- Open-source


Ask Copilot
Copilot is powered by AI, so mistakes are possible. Review output carefully before use.


GitHub Copilot is now available for free
The AI editor for everyone
Get started for free
See plans & pricing
Already have Visual Studio Code? Open now

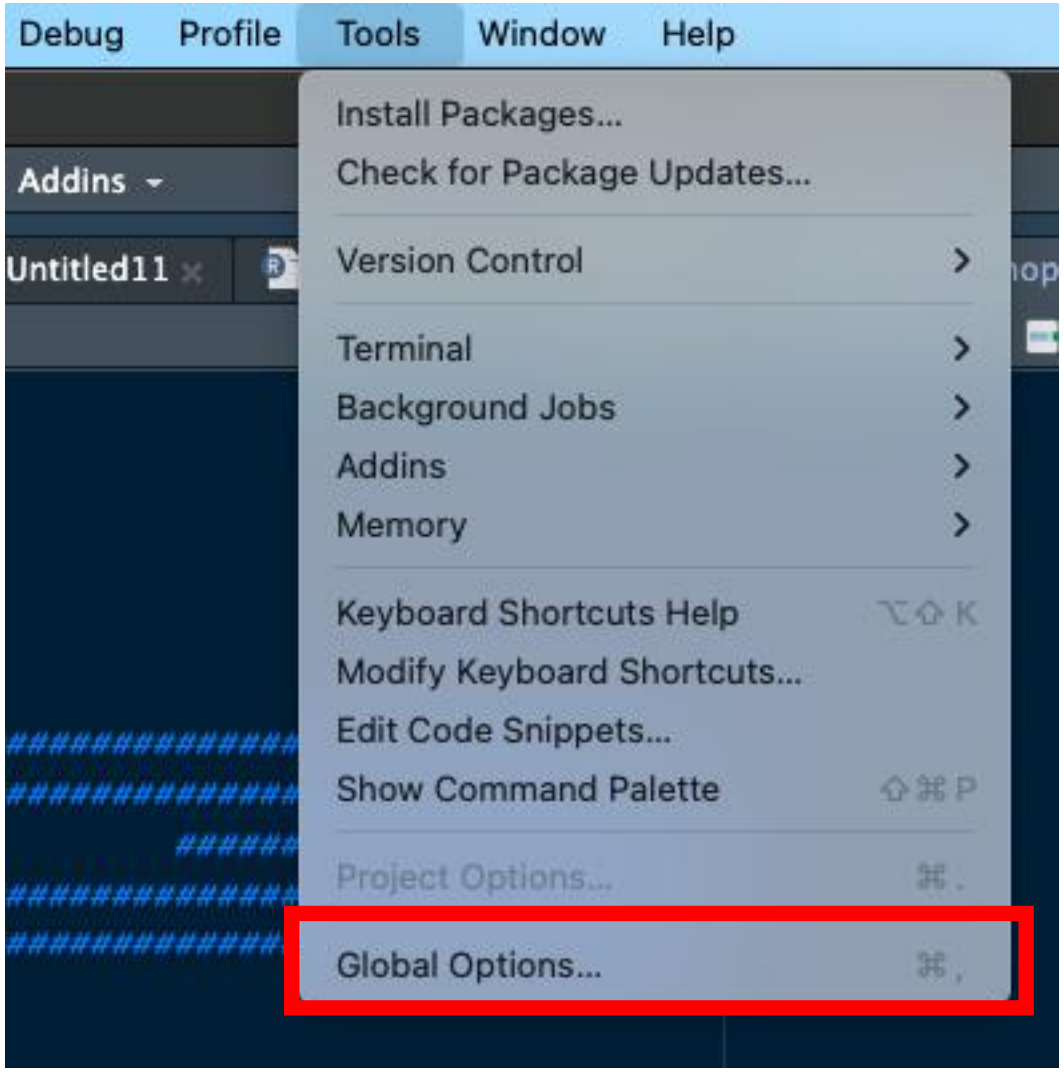
- Copilot est spécifiquement conçu pour *compléter* du code, détecter le langage, le contexte (les noms de variables, les librairies importées, etc.) et offrir les suggestions les plus pertinentes que CHAT GPT

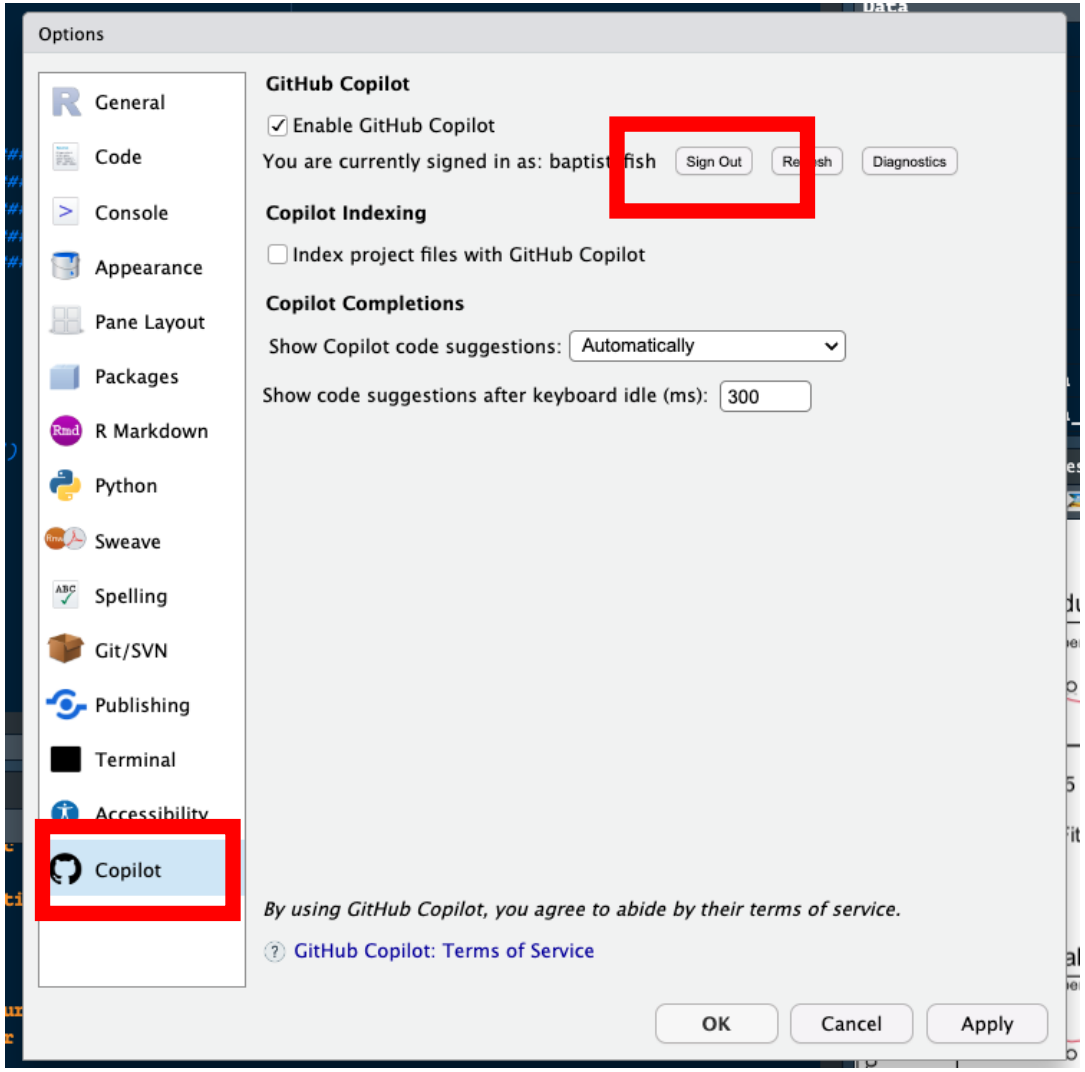
https://github.com/copilot

Modele OpenAI disponible + autres (Claude, Gemini etc.)

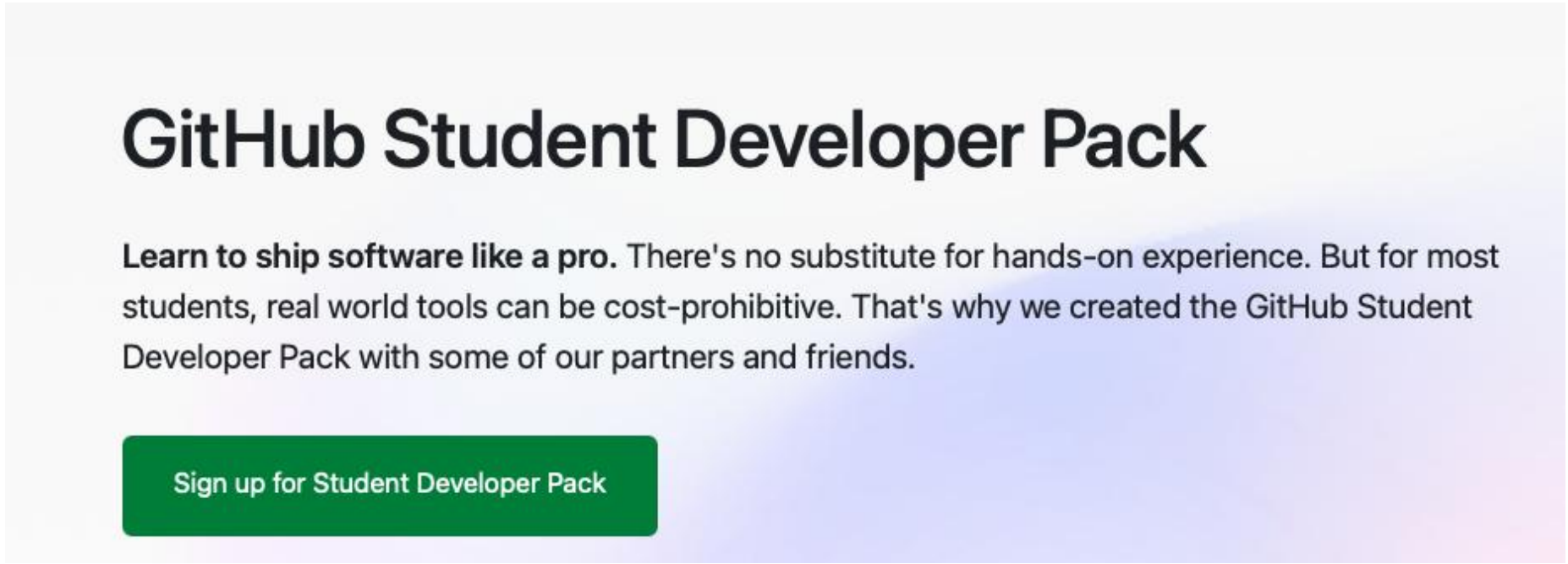
- Copilot peut générer non seulement de gros blocs de code, mais aussi compléter un appel de fonction, suggérer des paramètres ou des noms de variables, et s’adapter à la structure de votre projet.

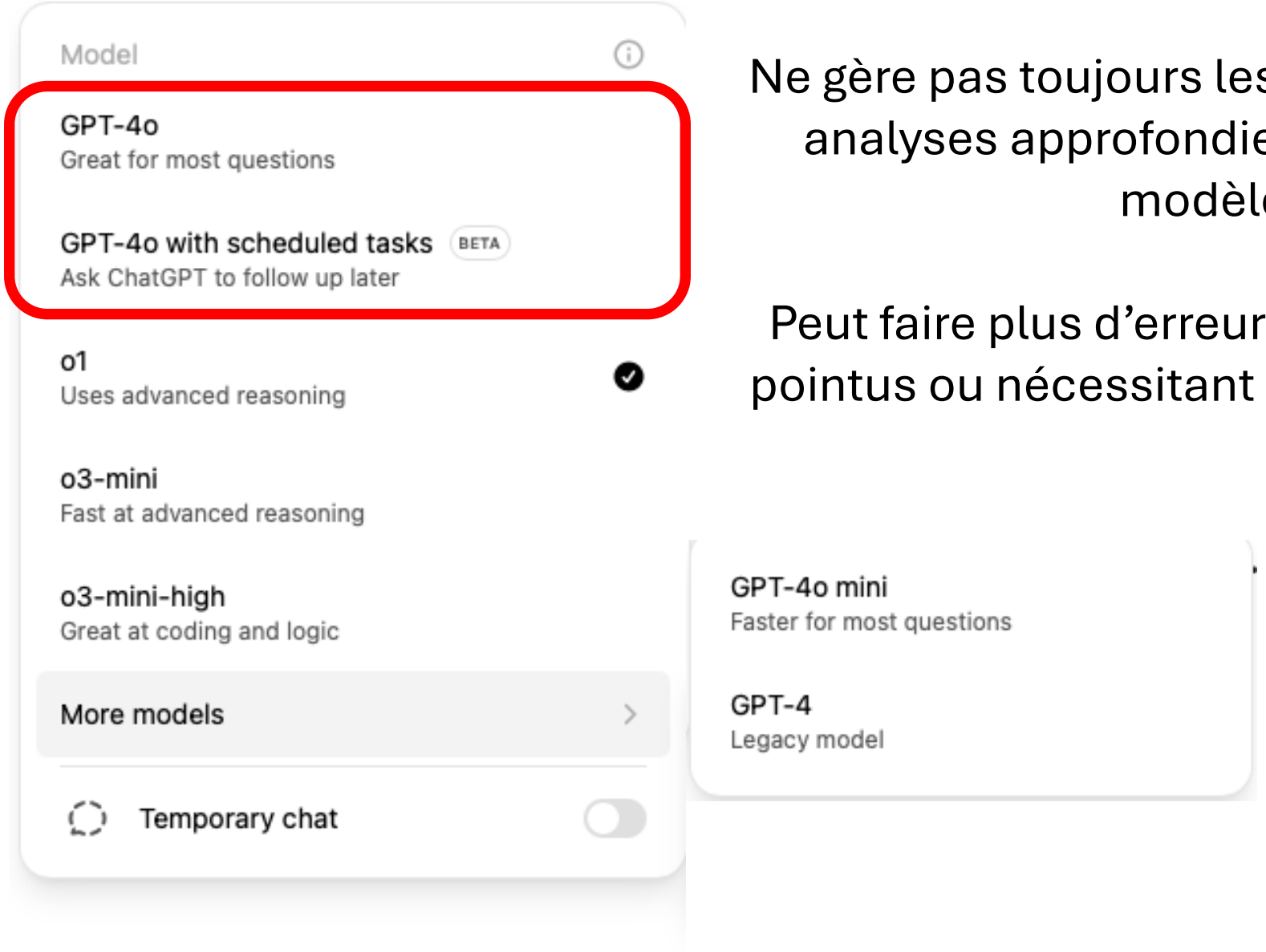
DÉMONSTRATION COPILOT





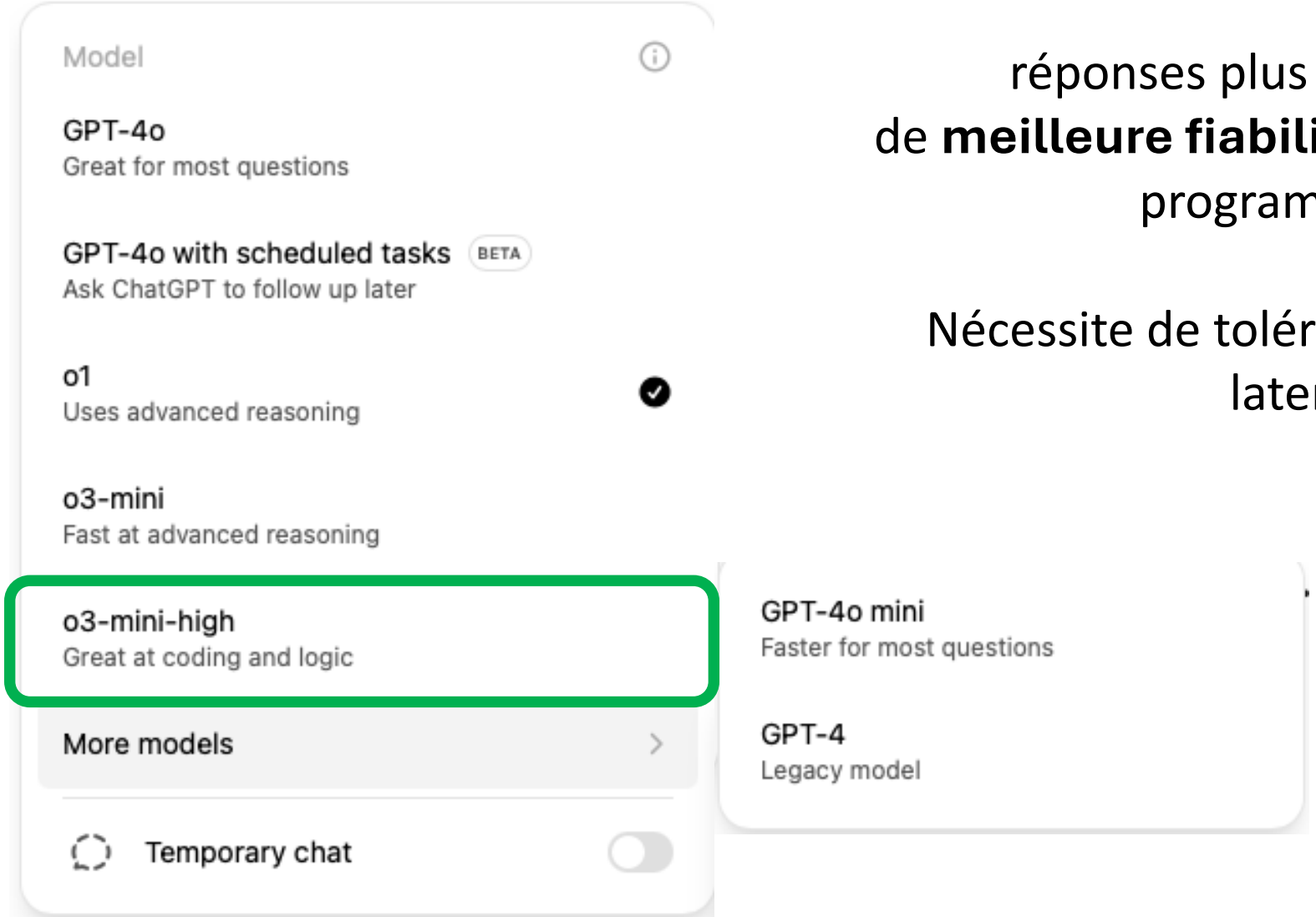
- 100 $ par ans (création compte étudiant pour accès gratuit)
https://education.github.com/pack





Ne gère pas toujours les *cas complexes* ou les analyses approfondies aussi bien que les modèles o1.

Peut faire plus d'erreurs pour des problèmes pointus ou nécessitant une logique complexe.



réponses plus **avancées** et
de **meilleure fiabilité** en logique et en programmation

Nécessite de tolérer un peu plus de latence

DÉMONSTRATION

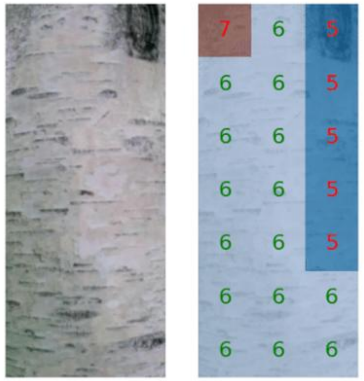
https://chatlize.ai

3. Application des IA à l'analyse automatisée d'images - Théo

1. Acquisition du jeu de données
2. Compréhension du jeu de donnée
3. Analyse statistiques

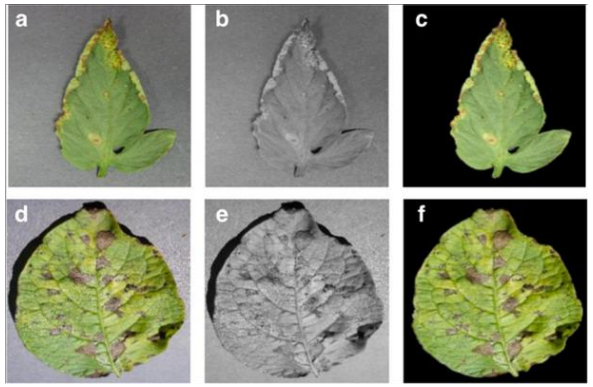
Pourquoi analyser des images ?

- Accès à des données riches et détaillées
- Répétabilité et archivage des données
- Automatisation et traitement à grande échelle



(Carpentier et al., 2018)

10.000 images



(Mohanty et al., 2016)

80.000 images

(Norouzzadeh et al, 2018)

2.000.000 images

-> Enjeu : gestion d'un grand volume d'images et leur exploitation rapide.

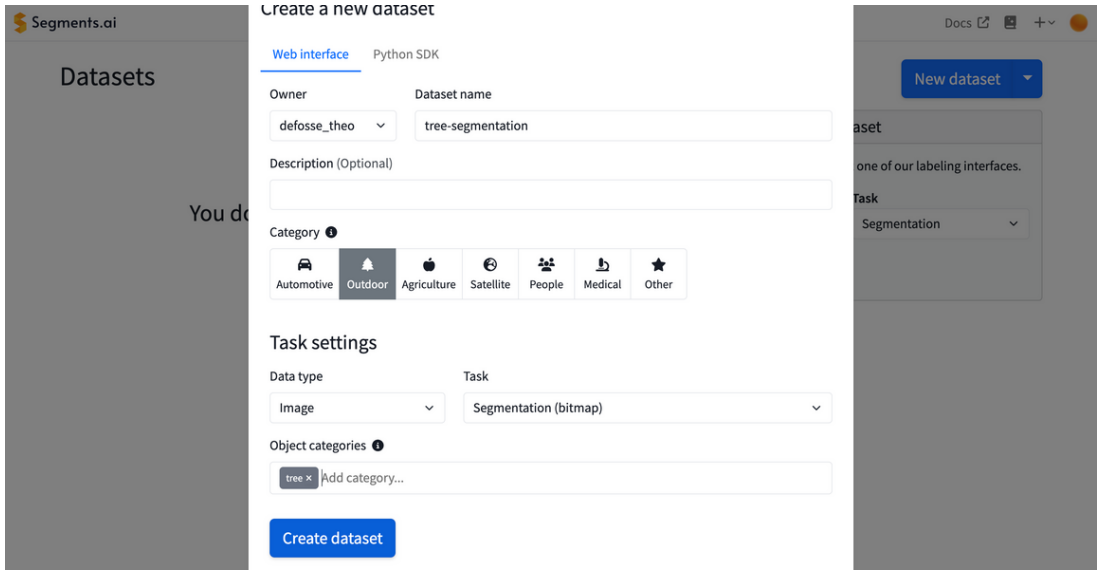
Cas d'étude fictif - Analyse d'une forêt

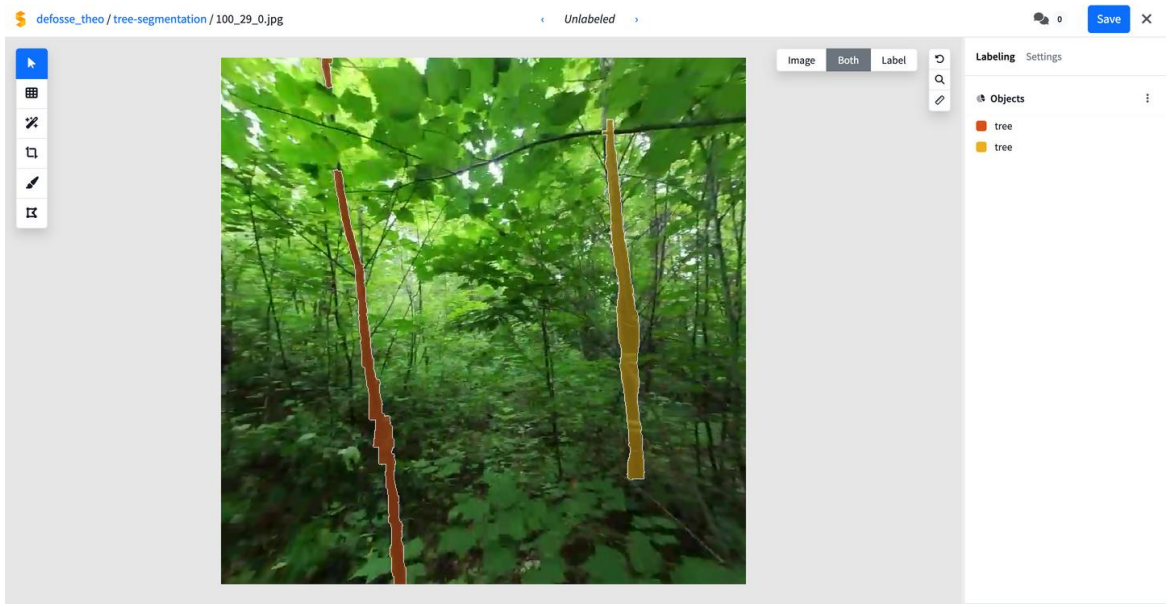
- Objectif : analyser les arbres d'une parcelle forestière.
- 1000+ images capturées sur 0,25 ha grâce au LiDAR.
- Utilisation d'un réseau neuronal pour extraire les informations pertinentes.

Cas d'étude fictif - Analyse d'une forêt

Prétraitement des images - Annotation des données

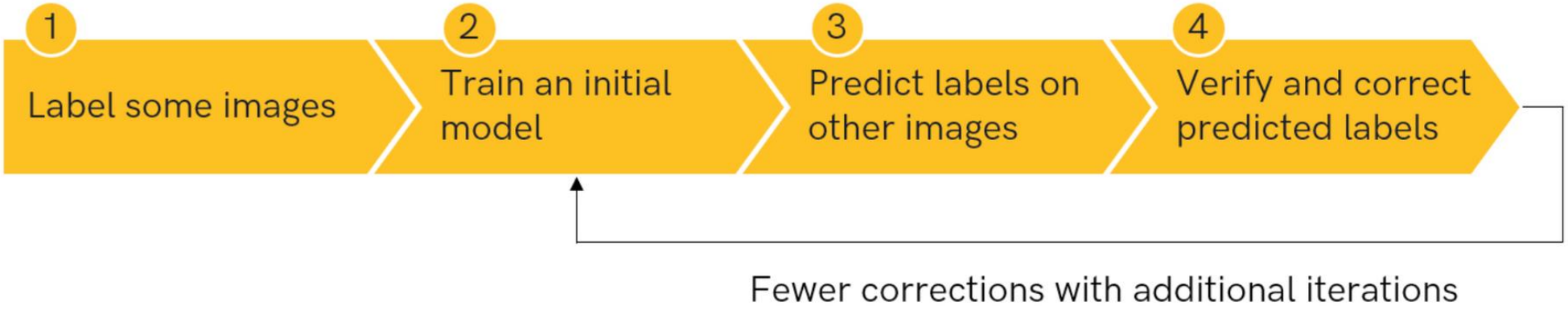
- Annotation nécessaire pour entraîner un modèle d'IA.
- SegmentIA : Outil spécialisé dans la segmentation d'image, proposant l'annotation assistée





Cas d'étude fictif - Analyse d'une forêt

Automatisation du traitement des images



code : https://colab.research.google.com/github/segments-ai/fast-labeling-workflow/blob/master/d

Cas d'étude fictif - Analyse d'une forêt

Exploitation données segmentées

Parcelle	A1	A2	A3
ntree	3	2	1
diametre_moyen	8	5	13
…			

Conclusion et perspectives

- Gain de temps significatif dans l'analyse d'images.
- SegmentAI simplifie l'annotation et la segmentation.
- L'intégration résultat segmentation pour analyses statistiques

Codes pour ChatGPT dans Rstudio :

```
#install.packages("chattr")

library(chattr)

#Sys.setenv("OPENAI_API_KEY" = "XXXXXXXXXXXXXXXXXXXXX")
#chattr_use("gpt35")

chattr_app(as_job = TRUE)
```