Machine Learning

Philippe Jacquet

Data Scientist à la DCSR

«L'apprentissage automatique (machine learning) est un champ d'étude … qui se fonde sur des approches mathématiques pour donner aux ordinateurs la capacité d'apprendre à partir de données».

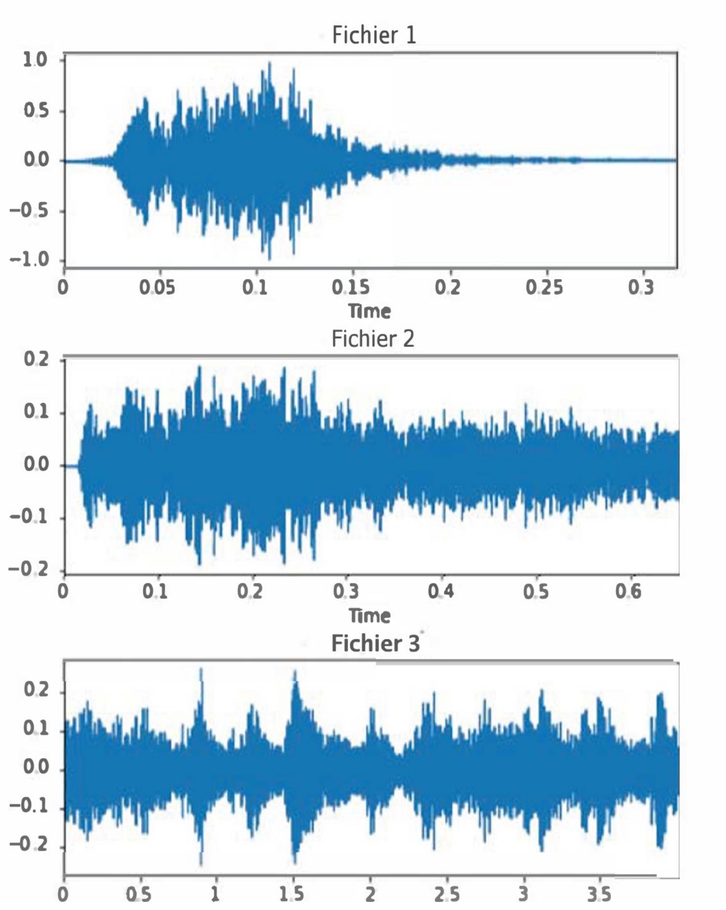
Wikipedia

Cas de figure 1

Données

Une centaine de fichiers audio transcrits et annotés
sur les variétés de français
(différences de prononciation selon les régions géographiques,
les couches sociales et les variations stylistiques).

Transcription des fichiers audio



Extraction
des meilleures propriétés

Nom	V1	V2	V3		V40
Fichier 1	1.2	4.2	8.1		3.5
Fichier 2	0.7	3.7	7.3		2.8
Fichier 3	2.3	1.9	9.5		4.2
Fichier 100	1.3	2.9	6.5		3.2

Données

Label attribué

Nom	V1	V2	V3		V40	Région
Fichier 1	1.2	4.2	8.1		3.5	Lausanne
Fichier 2	0.7	3.7	7.3		2.8	Paris
Fichier 3	2.3	1.9	9.5		4.2	Genève
Fichier 100	1.3	2.9	6.5		3.2	Neuchâtel

Les propriétés des fichiers audio

Besoin

Avoir un programme qui écoute le web
(vidéos youtube, émissions de radio, …)
à la recherche d’exemples de prononciation de phrases et
les pré-classer en fonction des régions géographiques
(ou couches sociales et variations stylistiques)

PHASE D’EDUCATION DE LA MACHINE

Prédiction: classification audio

Entrée

Fichier audio

Sortie

Région

Entraînement

Entrée

Fichier 1

Sortie

Lausanne

Entraînement

Entrée

Fichier 2

Sortie

Paris

Entraînement

Entrée

Fichier 100

Sortie

Neuchâtel

Prédiction: classification audio

Nouveau
fichier audio
trouvé sur le Web

➡

Genève

Conclusion

De nouvelles données viennent au chercheur.

Si les régions identifiées par la machine ne sont pas correctes,
le chercheur l'indique au programme
afin de corriger ou affiner l'algorithme de recherche.

Cas de figure 2

Données

Un chercheur qui étudie les langues parlées en Inde souhaite pouvoir identifier des similarités entre les quelque 270 langues recensées et effectuer des rapprochements avec les 22 langues officiellement reconnues par le gouvernement.

Il possède des données brutes constituées d’au moins 1 enregistrement audio d’une dizaine de minutes par langue.

Données

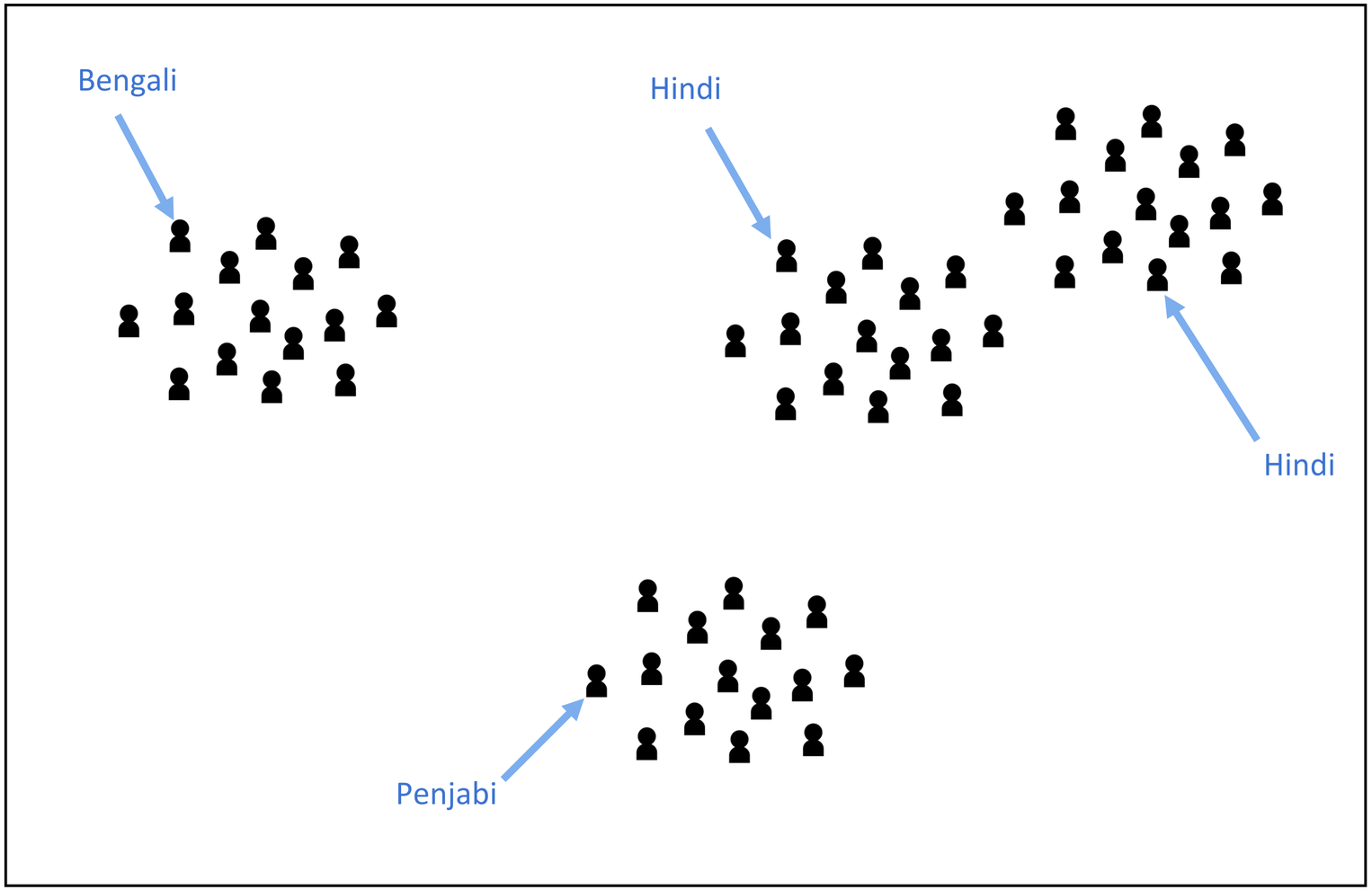
Label attribué

Nom	V1	V2	V3		V40	Langue
Fichier 1	1.2	4.2	8.1		3.5	Bengali
Fichier 2	0.7	3.7	7.3		2.8	Hindi
Fichier 3	2.3	1.9	9.5		4.2	Hindi
Fichier 270	1.3	2.9	6.5		3.2	Penjabi

Les propriétés des fichiers audio

Découverte de similarités

Réduction
de dimension

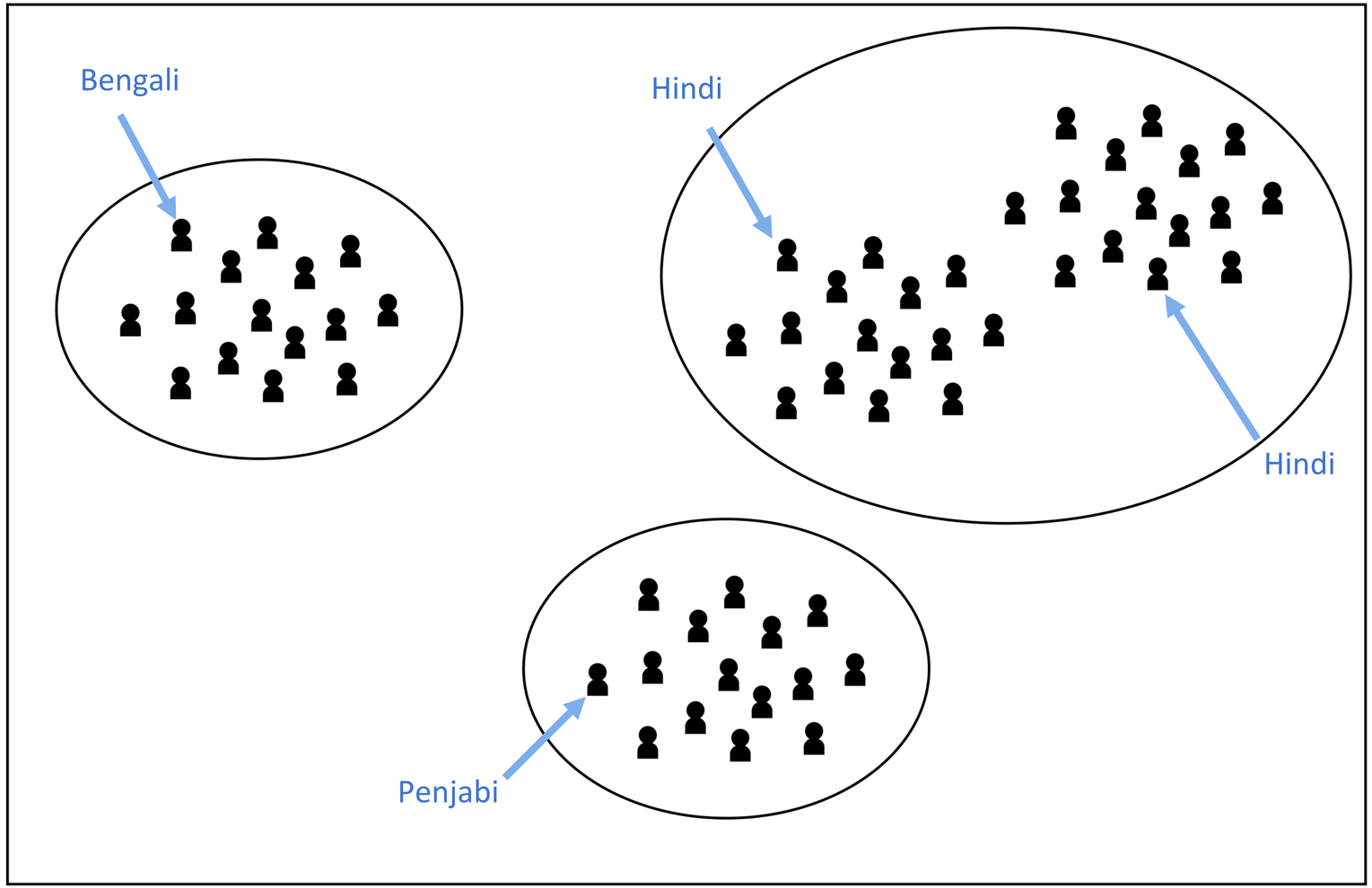


Objectif

Trouver les bons algorithmes
(traitements informatiques)
pour chaque similarité à identifier
(phonétique, prosodie, lexique, ...)

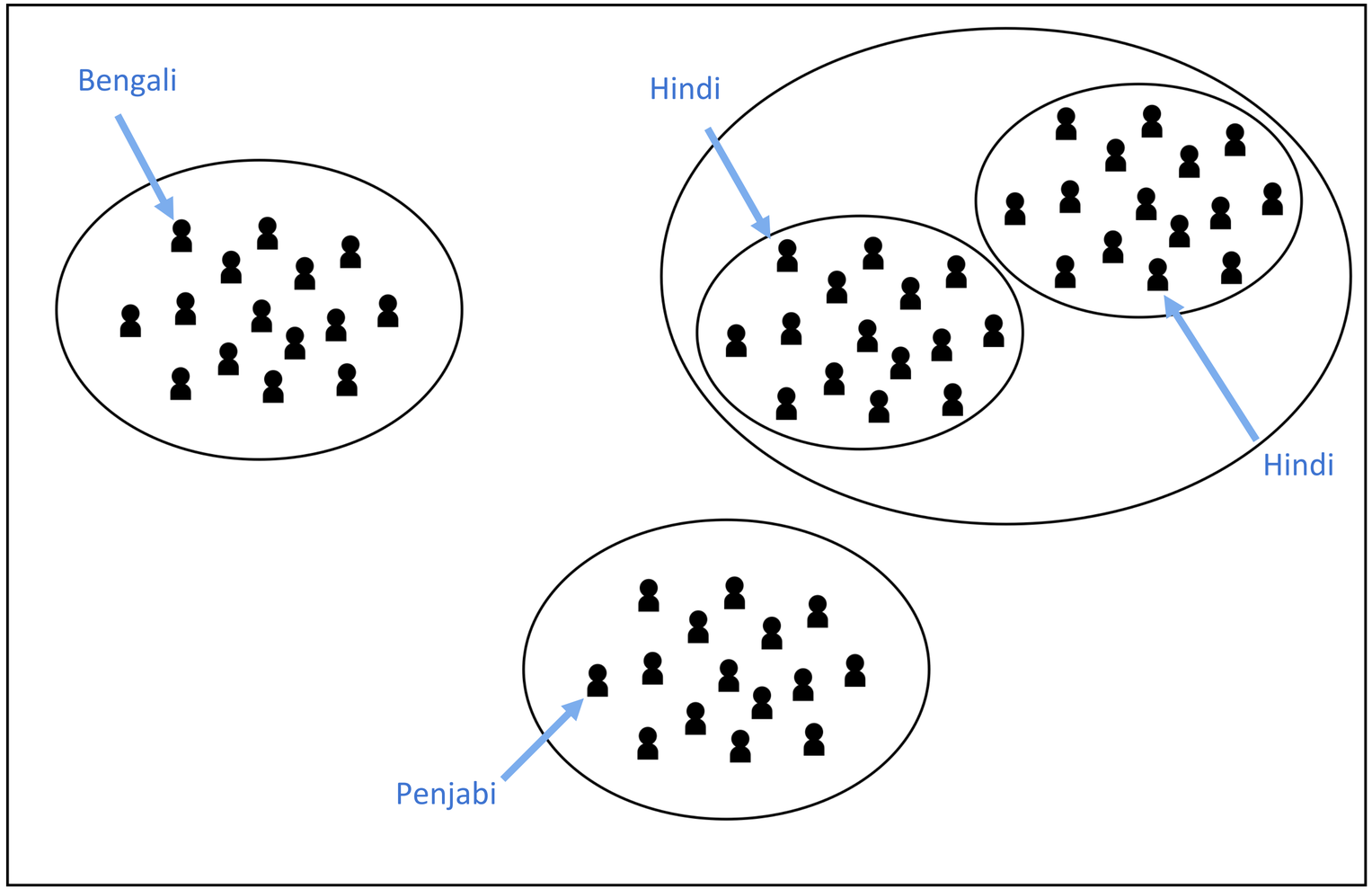
Découverte de similarités

Clustering
(regroupement)



Découverte de similarités

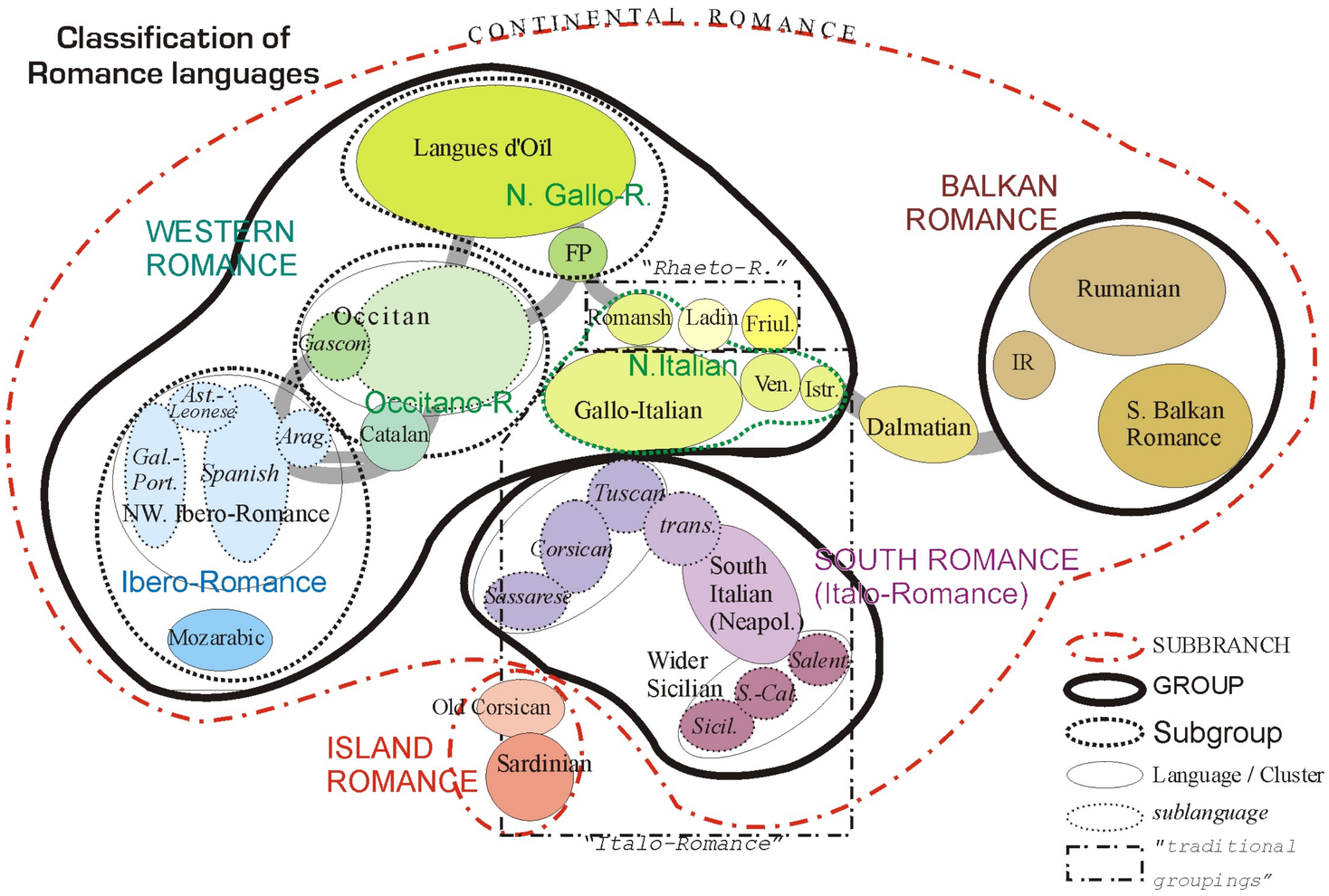
Clustering
(regroupement)



Conclusion

Après avoir identifié
les groupes de similarités (clusters),
nous pouvons obtenir une visualisation finale.

Exemple: Visualisation des langues romanes



Pour utiliser le Machine Learning

Les données:

- Type de données: tableau excel, images, vidéo, audio, texte, …
- Annotations des données (régions géographiques, langues locales, …)
- Quantité et qualité des données

L’informatique:

- Choix de la méthode: prédiction (cas 1) et/ou découverte de propriétés (cas 2)
- Besoin de pouvoir interpréter les résultats? (arbre de décision vs deep learning = boîte noire)
- Choix d’un langage de programmation: Python ou R

Besoin de conseils pour un projet Machine Learning?

helpdesk@unil.ch

Sujet: DCSR Machine Learning

ou

philippe.jacquet@unil.ch