LU3I026 - Science des données

Des données à l'apprentissage

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

Sorbonne Université

2025-2026

Formaliser le problème d'IA

Discussion informelle

- «On veut de l'IA»
- «sur nos données»
- «pour faire ci ou ça»
- «avec telle et telle contraintes»

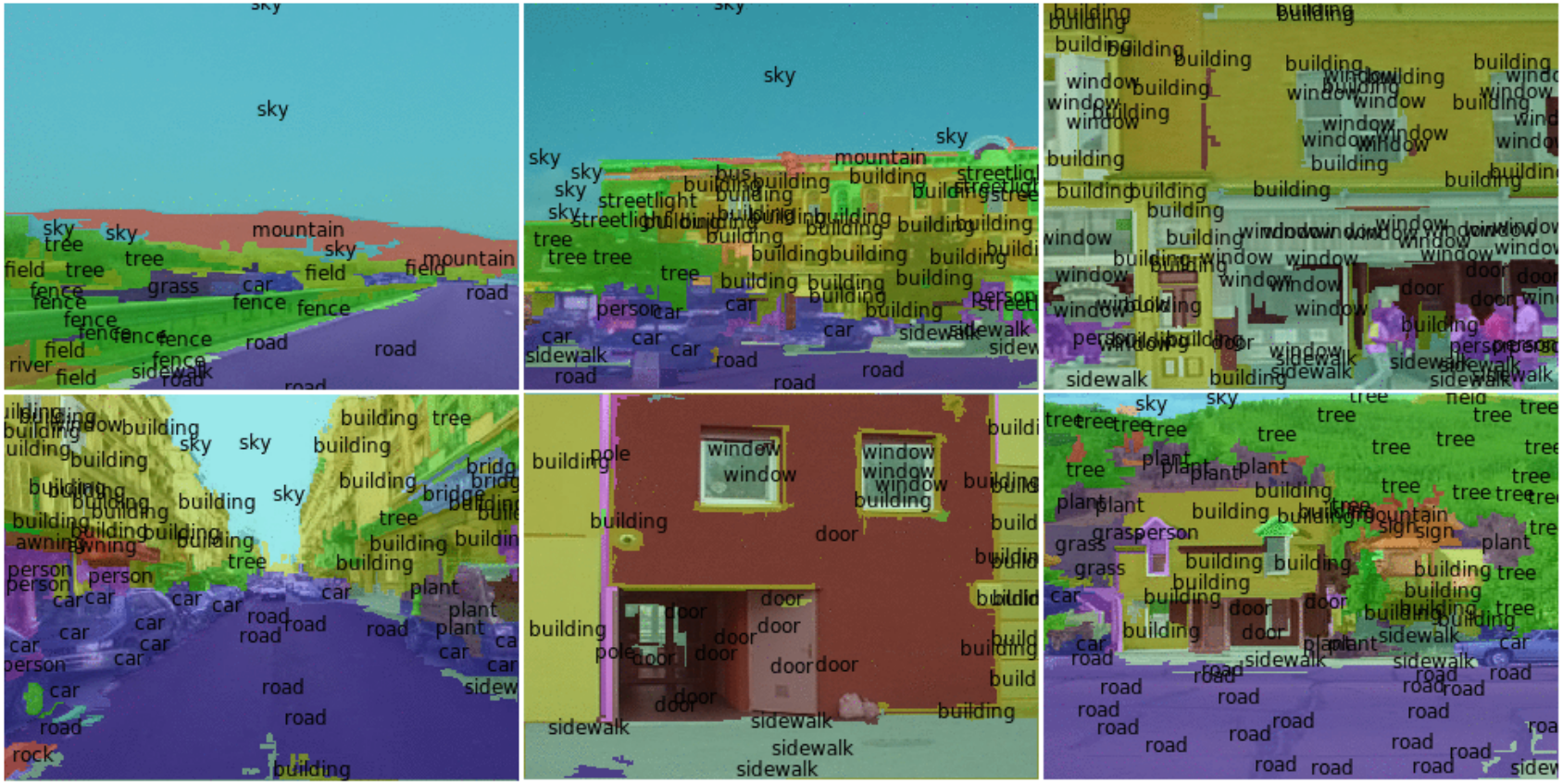
Objectif

- Traduction dans un vocabulaire IA
- Correspondance avec les concepts scientifiques

Première question à poser

- «Cher.e chef.fe , avez-vous des données ?»
- Parfois, la réponse est non...
- Or c'est un besoin capital !

Différentes données: images



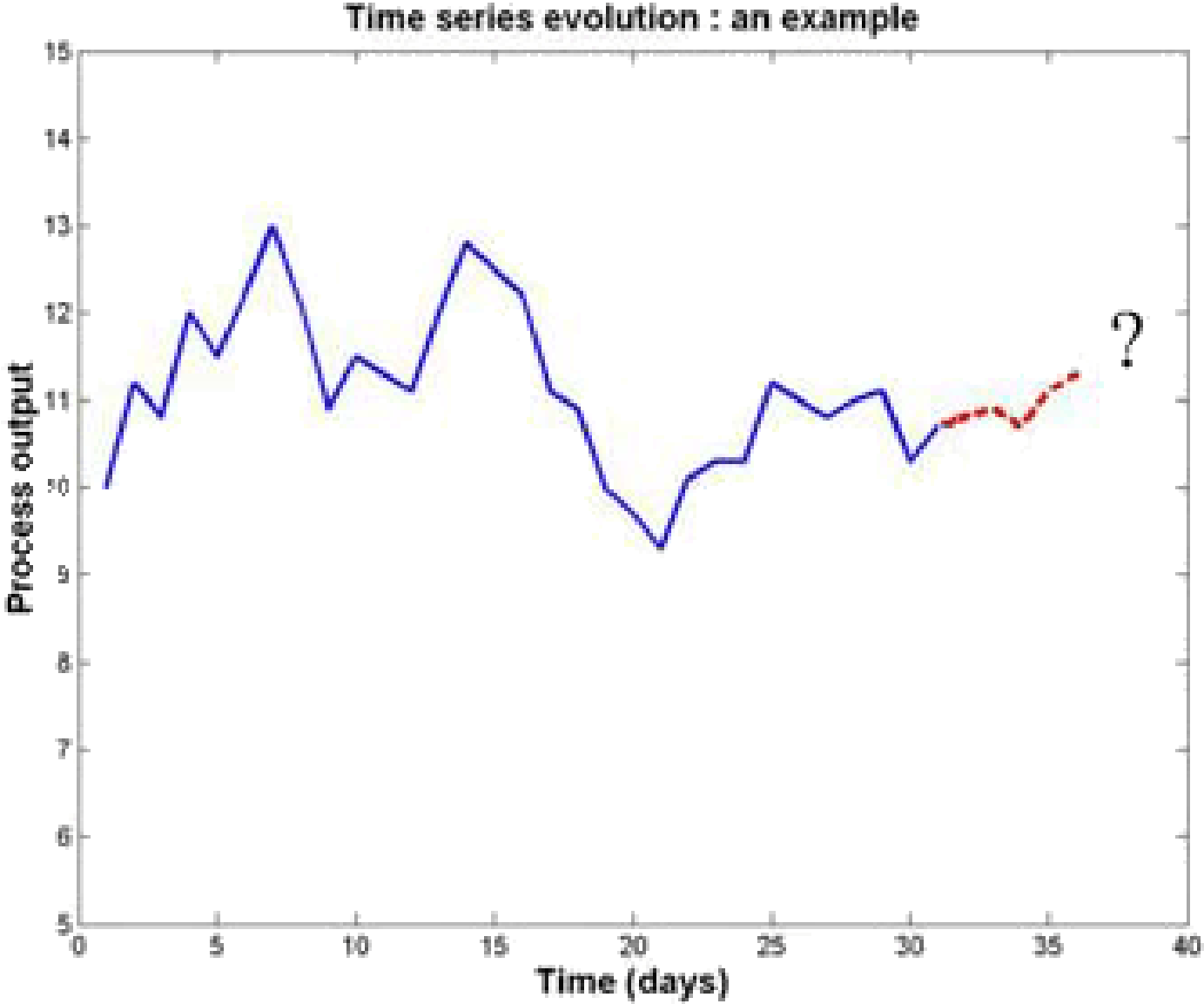
Différents données: texte

Extrait de wikipedia

ChatGPT est un prototype d'agent conversationnel utilisant l'intelligence artificielle, développé par OpenAI, et spécialisé dans le dialogue. L'agent conversationnel est un modèle de langage affiné par apprentissage supervisé et par apprentissage par renforcement.

Le sigle *ChatGPT* signifie « *Generative pre-trained transformer* » (« *Transformateur générique pré-entrainé* »).

Différentes données: séries temporelles



Différentes données: autres

Graphes

- Réseaux sociaux
- Échanges téléphoniques, sms
- Co-auteurs de publications scientifiques

Vidéos

- Image + aspect temporel

Conversations

- Échanges de type question-réponse (ChatGPT)

Différents domaines: texte médical

Centres d'Imagerie Médicale

Radiologie diagnostique et interventionnelle
Echographie – Doppler couleur
Mammographie numérisée – Ostéodensitométrie -Radiopédiatrie
Scanner – IRM

Thuir, le 19/07/12

Docteur Anne Marie BAIXAS
Diplômée de radiodiagnostic de la faculté de Montpellier
Certificat d'université des maladies du sein
66 1 0 18523

Mme HENRY Veronique

n° de demande A10050612460
Soc : DS.
Docteur SEDAGHAT Thomas

ECHOGRAPHIE MAMMAIRE BILATERALE –
Examen réalisé par échographie TOSHIBA APLIO MX (mis en service en septembre 2010)

Motif de l'examen :
Nodule algique du quadrant supéro externe droit.

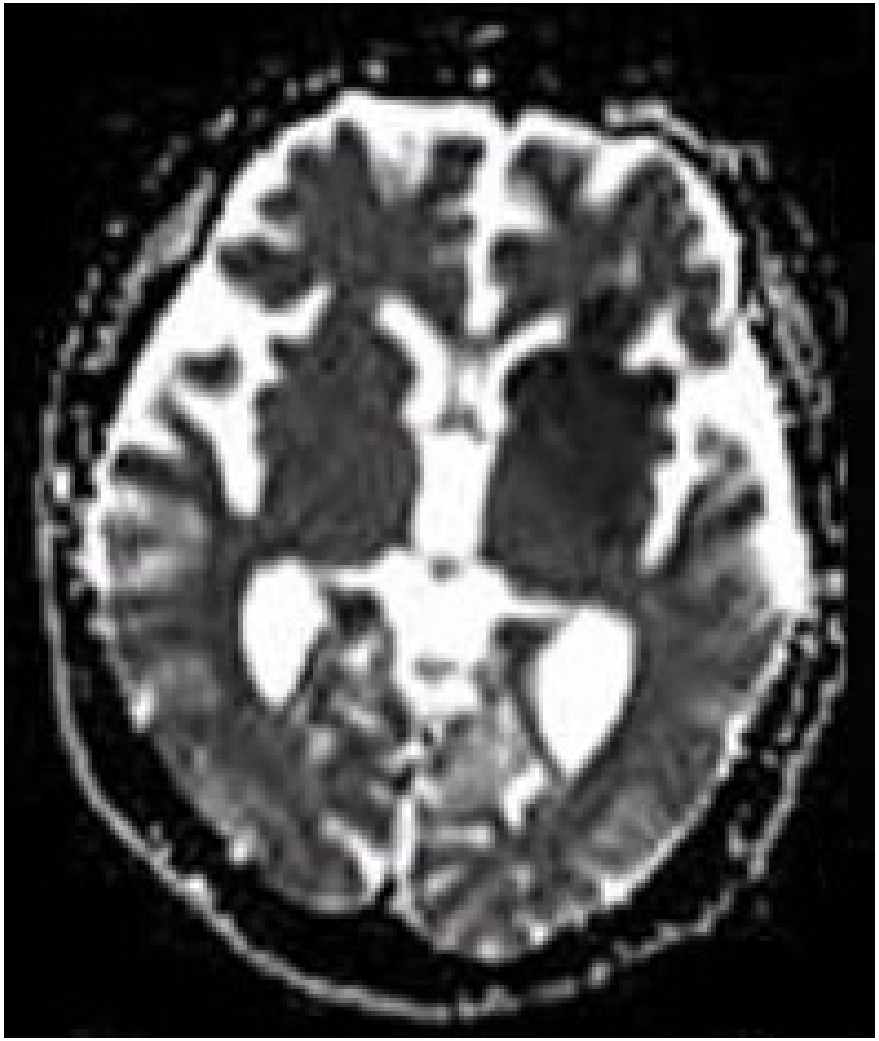
Résultats :
L'examen échographique retrouve un nodule hypoéchogène polylobé d'échostructure non liquidienne, mesuré à 20 x 17 x 15 mm. Ce nodule est peu atténuant, mais il existe une vascularisation nette en écho doppler couleur.
Toujours au niveau de ce quadrant supéro externe, présence d'une formation micro kystique du prolongement axillaire.
Il n'a pas été visualisé de ganglion dans le creux axillaire droit.
Le reste du sein est homogène.
Le sein gauche est d'échostructure normale sans lésion focalisée.

Compte tenu de l'aspect échographique un contrôle mammographique a été réalisé : il met en évidence dans la région supéro externe droite une opacité nodulaire, dense de contours irréguliers, contenant quelques micro calcifications.
Pas d'autre anomalie visible au niveau de ce sein.
A gauche, le sein est tout à fait homogène, pas de nodule ou de micro calcifications.

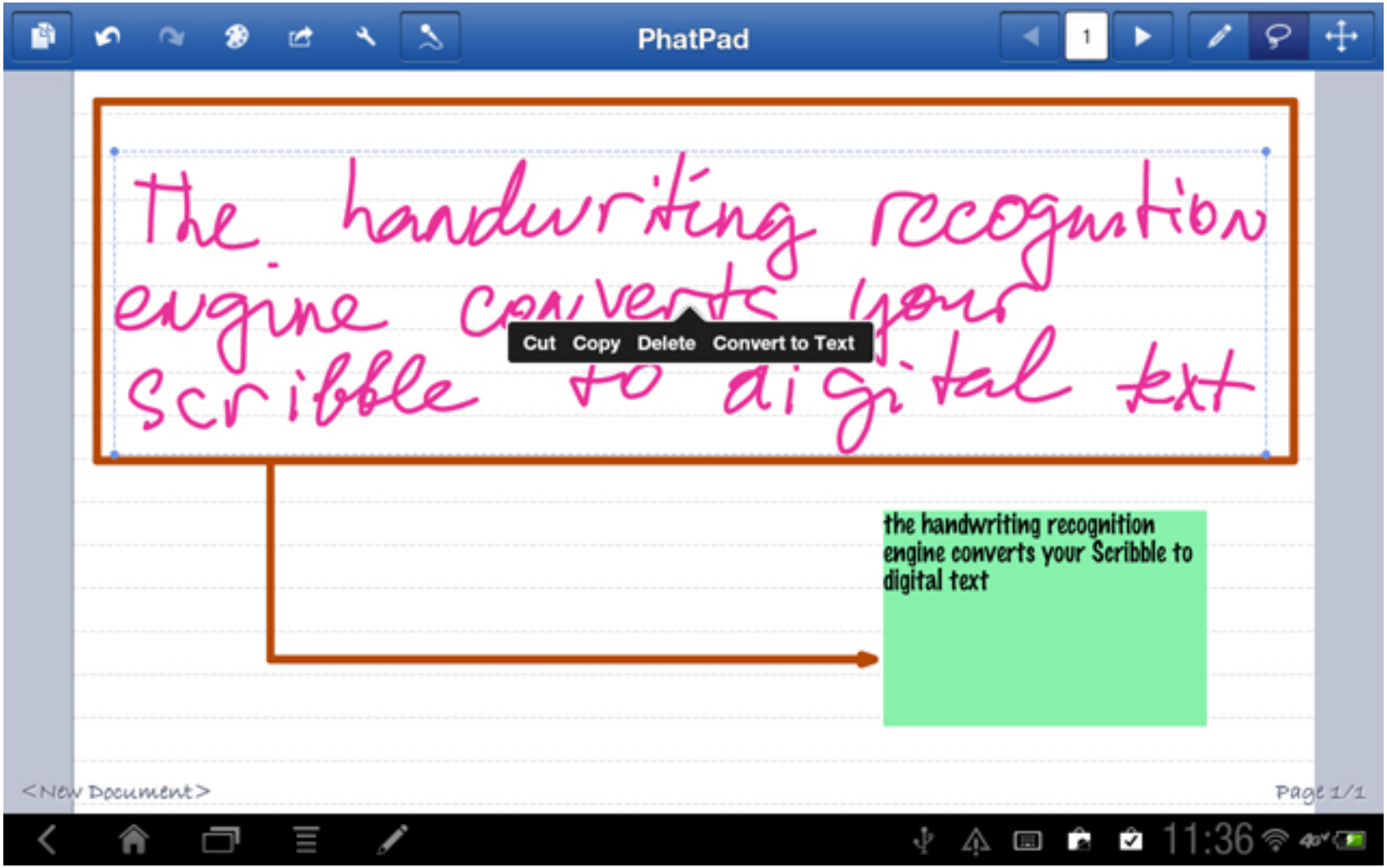
En conclusion :
Nodule de contours irréguliers, non liquidien du quadrant supéro externe du sein droit, classé ACR 4, nécessitant un prélèvement histologique.
ACR 1 à gauche.

Dr. Anne BAIXAS

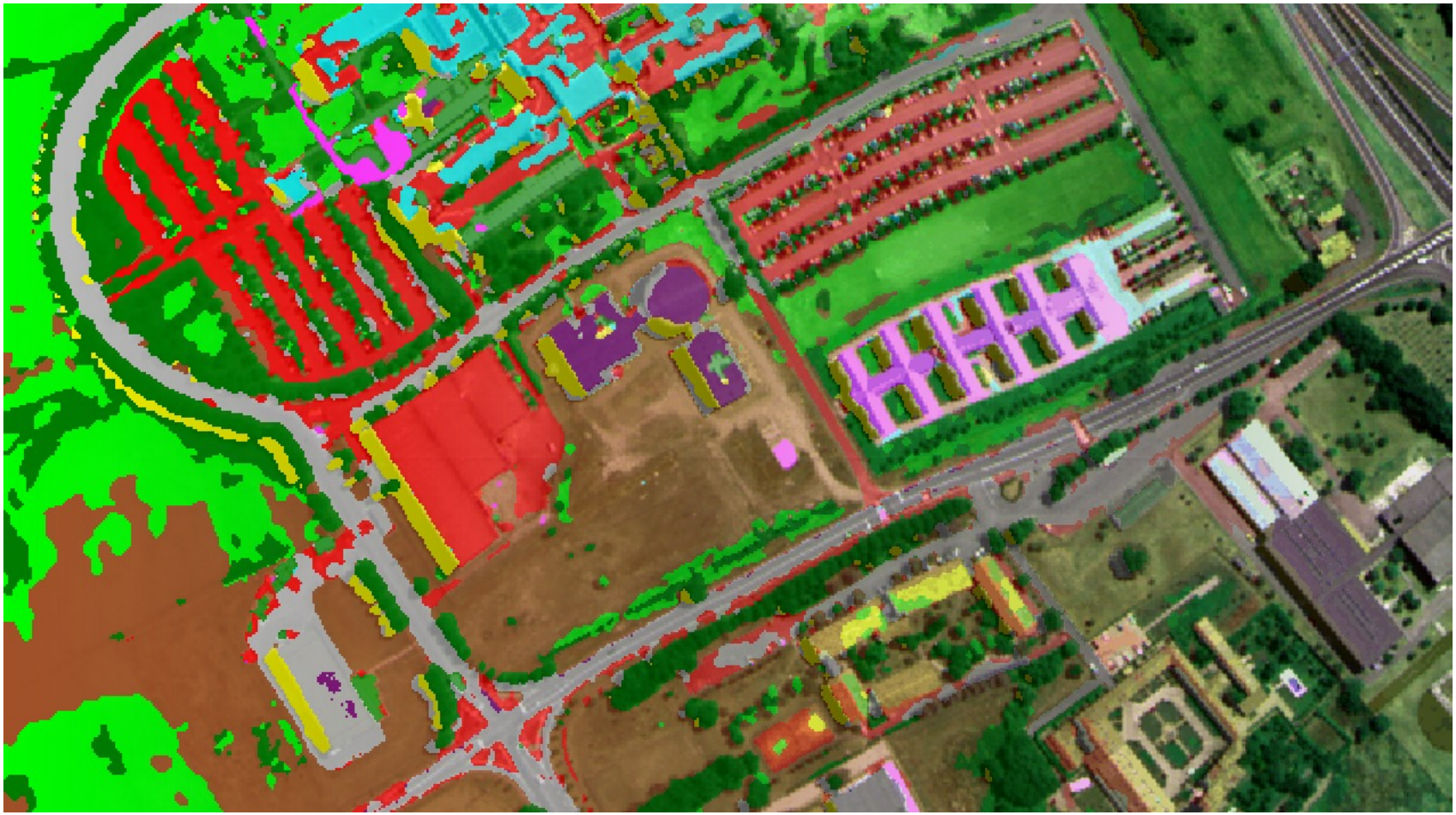
Différents domaines: image médicale



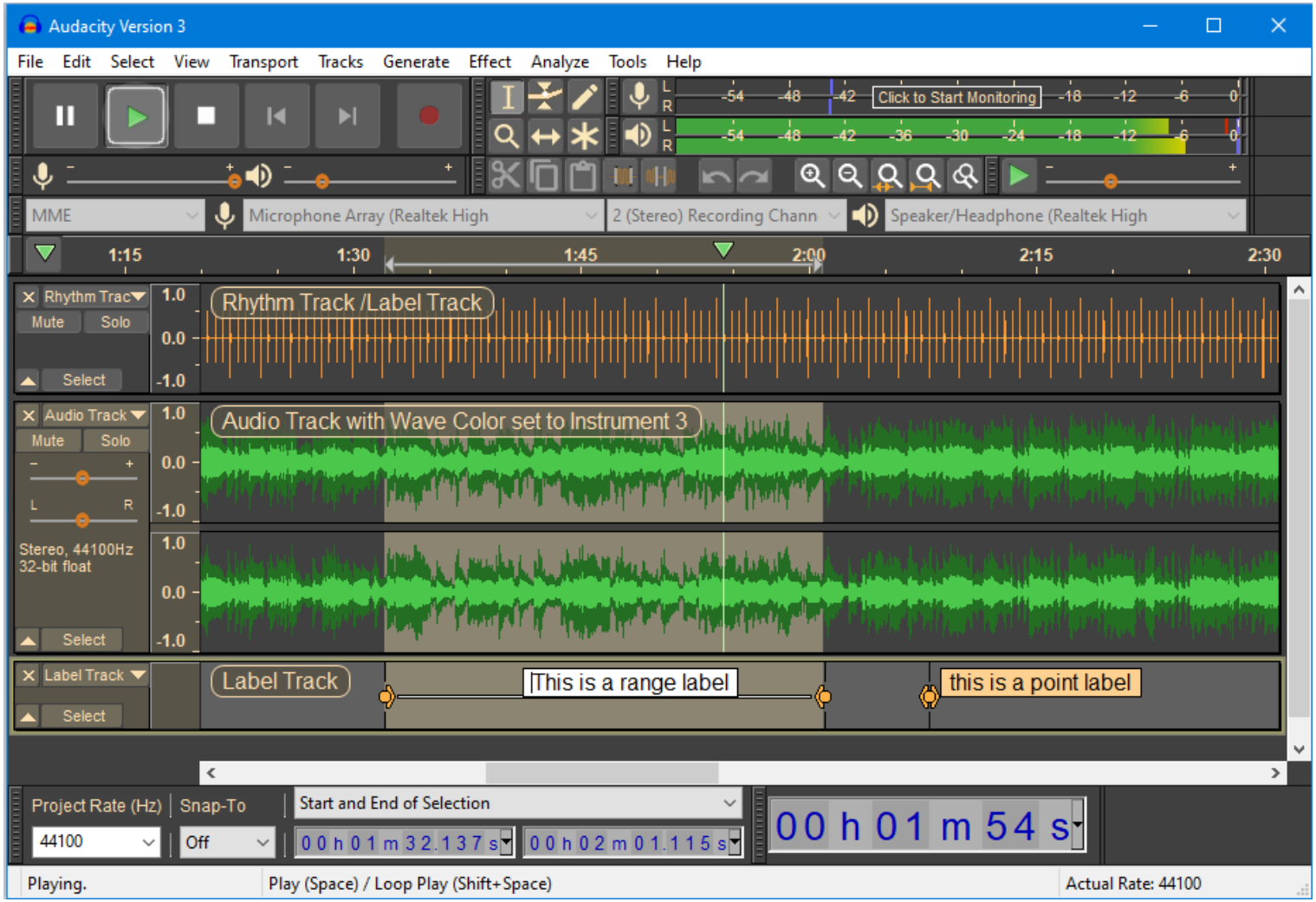
Différents domaines: écriture manuscrite



Différentes domaines: images satellites



Différentes domaines: audio



Différents domaines

Voiture autonome: vision

- Été – hiver
- Ville – campagne
- Amérique du nord – Europe

Contrôle

- Voiture autonome – drone – autre robot

Tâches de machine learning

Choix à faire

- Formalisation
- Définir les objectifs
- Définir le protocole expérimental

Exemple

- *Classification*
- *binaire*
- évaluation avec *accuracy*
- *découpage train/test* fixé de telle façon

Données

Matériau de base

- **Incoutournable**
- (sauf cas particulier)
- Vérifier: quantité, qualité

Avant de se lancer avec une méthode IA

- Charger les données
- Compréhension minimale
- Réfléchir aux biais
- Statistiques simples, visualisations
- Réfléchir aux prétraitements

Sources de données

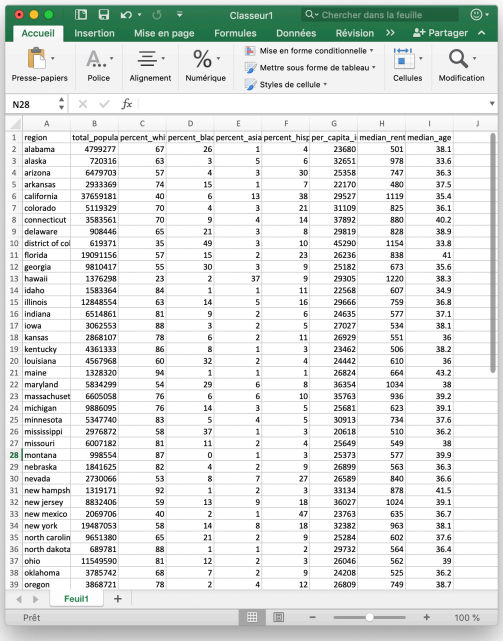
Monde académique

- Données jouets
- Données réelles
- Souvent bien connu et maîtrisés: stockage propre, exemples d'utilisation, bibliothèques

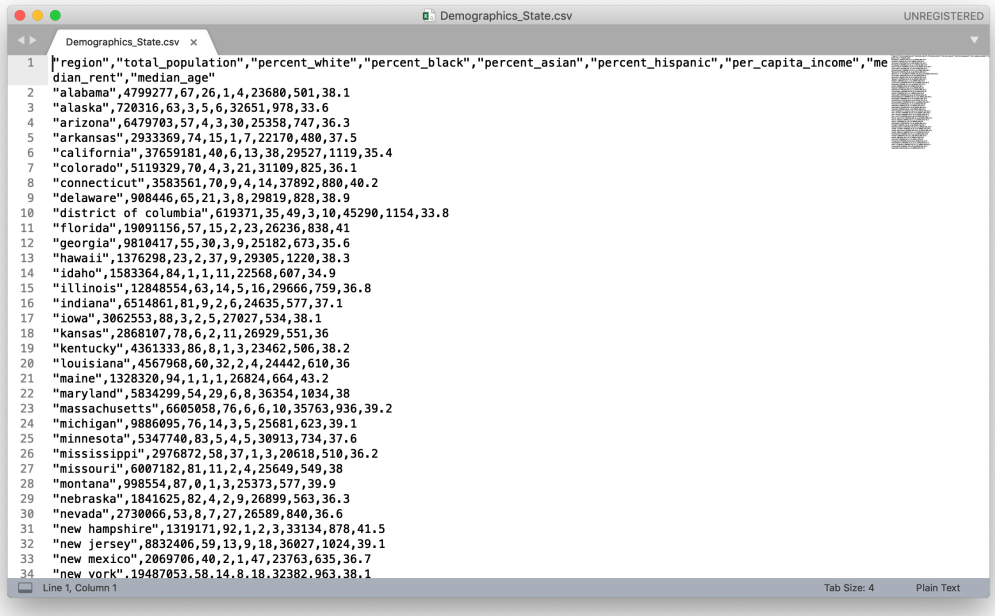
Monde réel

- Base de données
- Fichiers structurés: CSV, JSON, YAML, tableurs, autres formats
- Web: services web, API, scrapping

Exemple du CSV



region	total_popula	percent_whi	percent_blac	percent_asia	percent_hisp	per_capita_i	median_rent	median_age
alabama	4799277	67	26	1	4	23680	501	38.1
alaska	720316	63	3	5	6	32651	978	33.6
arizona	6479703	57	4	3	30	25358	747	36.3
arkansas	2933369	74	15	1	7	22170	480	37.5
california	37659181	40	6	13	38	29527	1119	35.4
colorado	5119329	70	4	3	21	31109	825	36.1
connecticut	3583561	70	9	4	14	37892	880	40.2
delaware	908446	65	21	3	8	29819	828	38.9
district of col	619371	35	49	3	10	45290	1154	33.8
florida	19091156	57	15	2	23	26236	838	41
georgia	9810417	55	30	3	9	25182	673	35.6
hawaii	1376298	23	2	37	9	29305	1220	38.3
idaho	1583364	84	1	1	11	22568	607	34.9
illinois	12848554	63	14	5	16	29666	759	36.8
indiana	6514861	81	9	2	6	24635	577	37.1
iowa	3062553	88	3	2	5	27027	534	38.1
kansas	2868107	78	6	2	11	26929	551	36
kentucky	4361333	86	8	1	3	23462	506	38.2
louisiana	4567968	60	32	2	4	24442	610	36
maine	1328320	94	1	1	1	26824	664	43.2
maryland	5834299	54	29	6	8	36354	1034	38
massachuset	6605058	76	6	6	10	35763	936	39.2
michigan	9886095	76	14	3	5	25681	623	39.1
minnesota	5347740	83	5	4	5	30913	734	37.6
mississippi	2976872	58	37	1	3	20618	510	36.2
missouri	6007182	81	11	2	4	25649	549	38
montana	998554	87	0	1	3	25373	577	39.9
nebraska	1841625	82	4	2	9	26899	563	36.3
nevada	2730066	53	8	7	27	26589	840	36.6
new hampsh	1319171	92	1	2	3	33134	878	41.5
new jersey	8832406	59	13	9	18	36027	1024	39.1
new mexico	2069706	40	2	1	47	23763	635	36.7
new york	19487053	58	14	8	18	32382	963	38.1
north carolin	9651380	65	21	2	9	25284	602	37.6
north dakota	689781	88	1	1	2	29732	564	36.4
ohio	11549590	81	12	2	3	26046	562	39
oklahoma	3785742	68	7	2	9	24208	525	36.2
oregon	3868721	78	2	4	12	26809	749	38.7



```
"region","total_population","percent_white","percent_black","percent_asian","percent_hispanic","per_capita_income","median_rent","median_age"
"alabama",4799277,67,26,1,4,23680,501,38.1
"alaska",720316,63,3,5,6,32651,978,33.6
"arizona",6479703,57,4,3,30,25358,747,36.3
"arkansas",2933369,74,15,1,7,22170,480,37.5
"california",37659181,40,6,13,38,29527,1119,35.4
"colorado",5119329,70,4,3,21,31109,825,36.1
"connecticut",3583561,70,9,4,14,37892,880,40.2
"delaware",908446,65,21,3,8,29819,828,38.9
"district of columbia",619371,35,49,3,10,45290,1154,33.8
"florida",19091156,57,15,2,23,26236,838,41
"georgia",9810417,55,30,3,9,25182,673,35.6
"hawaii",1376298,23,2,37,9,29305,1220,38.3
"idaho",1583364,84,1,1,11,22568,607,34.9
"illinois",12848554,63,14,5,16,29666,759,36.8
"indiana",6514861,81,9,2,6,24635,577,37.1
"iowa",3062553,88,3,2,5,27027,534,38.1
"kansas",2868107,78,6,2,11,26929,551,36
"kentucky",4361333,86,8,1,3,23462,506,38.2
"louisiana",4567968,60,32,2,4,24442,610,36
"maine",1328320,94,1,1,1,26824,664,43.2
"maryland",5834299,54,29,6,8,36354,1034,38
"massachusetts",6605058,76,6,6,10,35763,936,39.2
"michigan",9886095,76,14,3,5,25681,623,39.1
"minnesota",5347740,83,5,4,5,30913,734,37.6
"mississippi",2976872,58,37,1,3,20618,510,36.2
"missouri",6007182,81,11,2,4,25649,549,38
"montana",998554,87,0,1,3,25373,577,39.9
"nebraska",1841625,82,4,2,9,26899,563,36.3
"nevada",2730066,53,8,7,27,26589,840,36.6
"new hampshire",1319171,92,1,2,3,33134,878,41.5
"new jersey",8832406,59,13,9,18,36027,1024,39.1
"new mexico",2069706,40,2,1,47,23763,635,36.7
"new york",19487053,58,14,8,18,32382,963,38.1
```

Un fichier texte: des lignes

- Valeur1 VIRGULE Valeur2 VIRGULE etc
- Valeur1 VIRGULE Valeur2 VIRGULE etc

Quelques outils

Éviter Python pur

Pandas – Polars

- Réflexe à avoir: chargement, filtrage, statistiques, visualisation
- Point de vue base de données
- `read_csv` (et plein d'autres)

Numpy – Scipy

- Implémentations d'algorithmes
- Calcul scientifique
- `loadtxt`

Requests – HTTPX – BeautifulSoup – Playwright

- HTTP, HTML, Javascript

Regarder: fast-checking

- Dimensions des structures de donnéees
- En-têtes des colonnes
- `print` tous simples

	region	total_population	percent_white	percent_black	percent_
0	alabama	4799277	67	26	
1	alaska	720316	63	3	
2	arizona	6479703	57	4	
3	arkansas	2933369	74	15	
4	california	37659181	40	6	

	percent_hispanic	per_capita_income	median_rent	median_age
0	4	23680	501	38.1
1	6	32651	978	33.6
2	30	25358	747	36.3
3	7	22170	480	37.5
4	38	29527	1119	35.4

Indicateurs statistiques

Outils mathématiques simples

- Moyennes, médianes, écart-types, percentiles

Outils informatique

Pandas `describe` (attention: pour vous ! pas pour un rapport propre !)

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
count	364	364	364.000000	364	364	364	364.000000	364	3.640000e+02
unique	364	30	NaN	5	22	17	NaN	115	NaN
top	Cleanthony Early	New Orleans Pelicans	NaN	SG	24.0	6-9	NaN	Kentucky	NaN
freq	1	16	NaN	87	41	49	NaN	22	NaN
mean	NaN	NaN	16.829670	NaN	NaN	NaN	219.785714	NaN	4.620311e+06
std	NaN	NaN	14.994162	NaN	NaN	NaN	24.793099	NaN	5.119716e+06
min	NaN	NaN	0.000000	NaN	NaN	NaN	161.000000	NaN	5.572200e+04
20%	NaN	NaN	4.000000	NaN	NaN	NaN	195.000000	NaN	9.472760e+05
40%	NaN	NaN	9.000000	NaN	NaN	NaN	212.000000	NaN	1.638754e+06
50%	NaN	NaN	12.000000	NaN	NaN	NaN	220.000000	NaN	2.515440e+06
60%	NaN	NaN	17.000000	NaN	NaN	NaN	228.000000	NaN	3.429024e+06

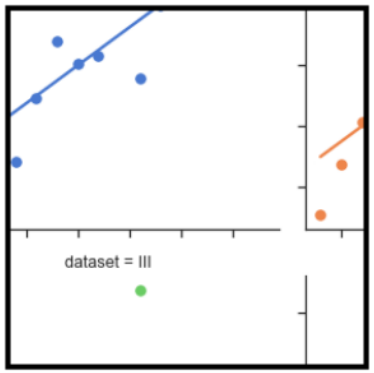
Visualisation

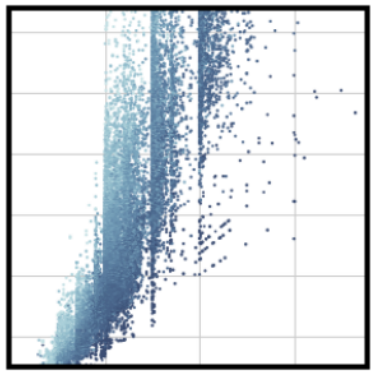
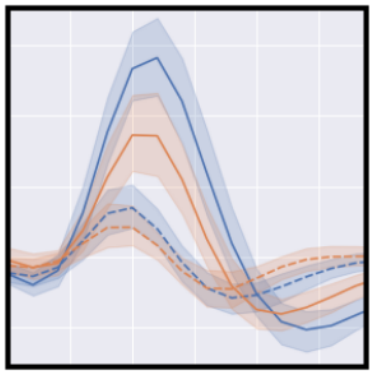
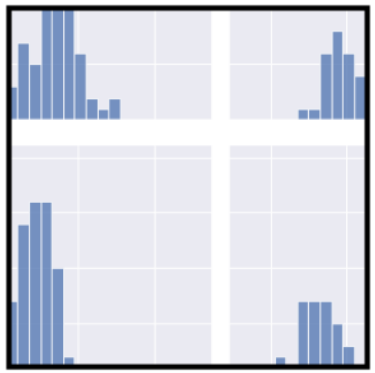
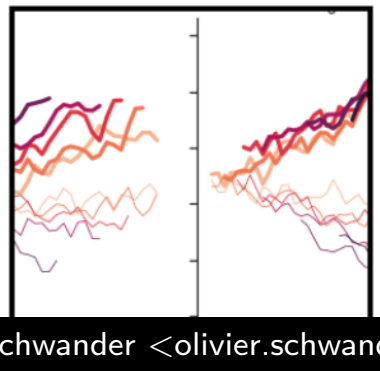
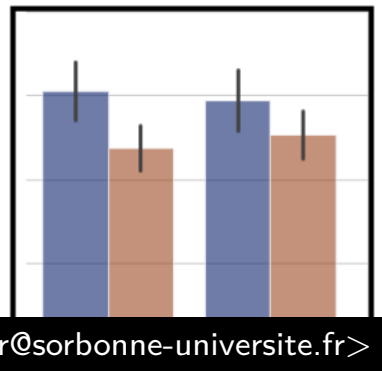
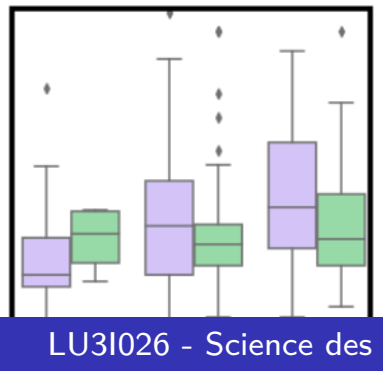
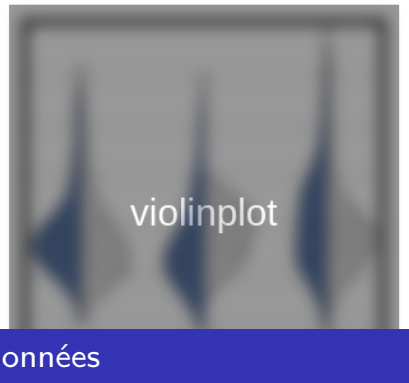
Graphiques

- Histogrammes, courbes, nuages de points, etc

Outils informatique

- Matplotlib, Bokeh, Seaborn, Altair



Autres types de données

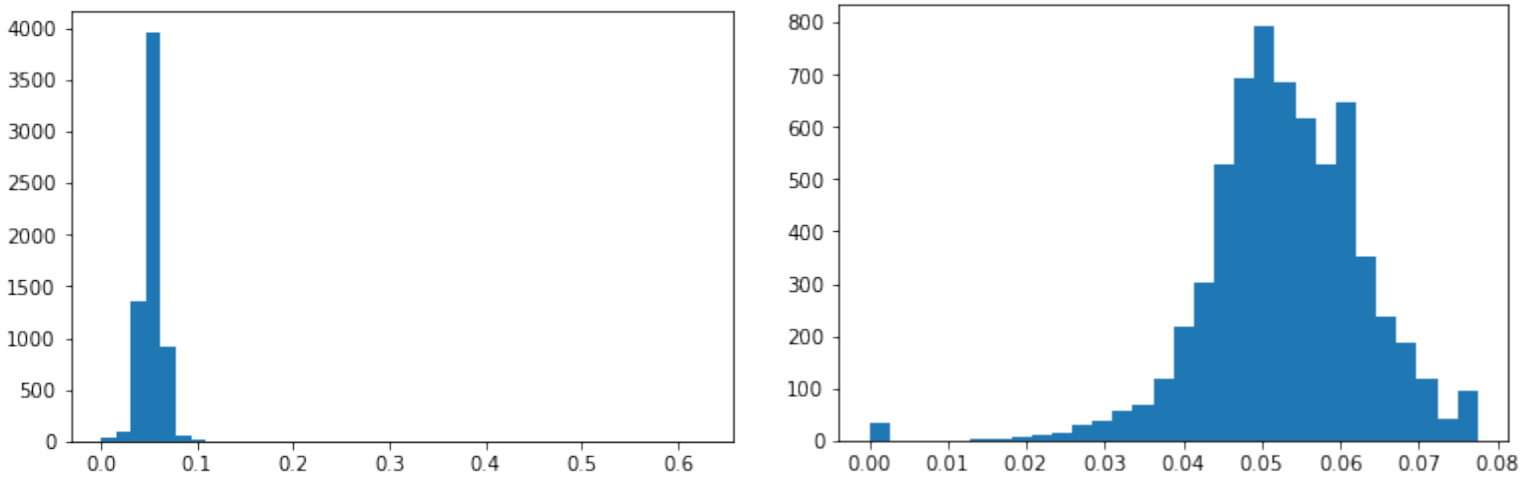
Visualisation adaptée !

- Images: afficher des images
- Texte: le texte...
- Son: écouter
- etc

Oui mais: retour sur la qualité des données

Valeurs aberrantes (outliers)

Histogramme des prix au km des courses Blablacar



(bonus: dire du mal de cette figure)

Comment faire ?

- À la main: pas évident
- Automatiquement: difficile

Biais éventuels

Problèmes de société

- Des images de femmes dans le dataset ?
- Toutes les couleurs de peau ?
- Uniquement de l'anglais ?

Rien d'objectif

- Un dataset est *situé* (au sens de la sociologie)
- Discriminations systémiques: racisme, sexisme, validisme, etc

Conséquences

- Technique: un modèle qui marche moins bien (parce que pas réaliste)
- Sociale: un modèle **complice** des discriminations

Il faut regarder dans le dataset !!!

Contextes d'apprentissage

Apprentissage supervisé

- Données
- Étiquettes
- Exemple: des images, et le nom de l'objet dans l'image
- Difficulté: obtenir les étiquettes...

Apprentissage non-supervisé

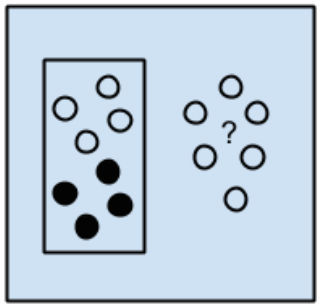
- Juste les données
- Pas d'étiquettes

Autres

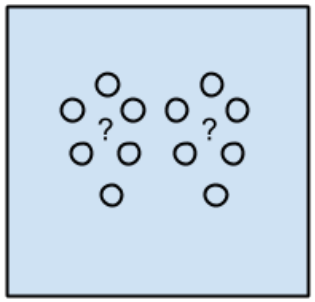
- Semi-supervisé
- Renforcement
- Auto-supervisé

Contextes d'apprentissage

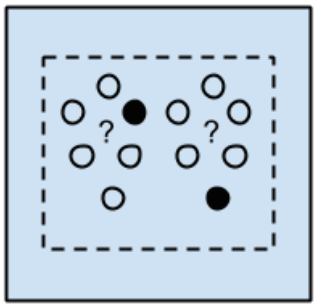
Supervisé | Non-supervisé | Semi-supervisé | Renforcement



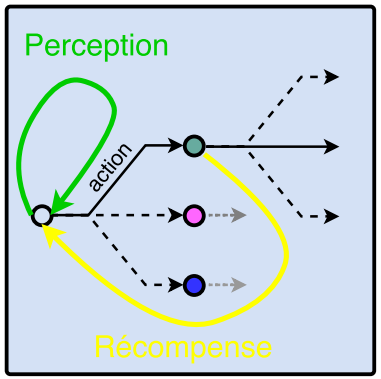
Supervised Learning Algorithms



Unsupervised Learning Algorithms



Semi-supervised Learning Algorithms

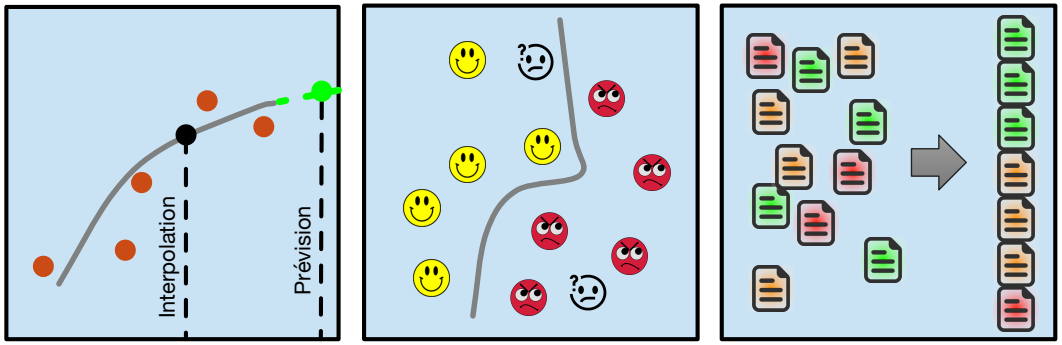


Première chose à déterminer

- conditionne les choix d'algorithmes et de méthodes
- et les mesures de performances

Tâches d'apprentissage

Régression Classification Ordonnancement



Autres tâches

- Souvent des cas particuliers des tâches précédentes
- Image: segmentation (classif), localisation (régression)
- Texte: génération (classif), dialogue (ordonnancement)

Deuxième chose à déterminer

- Conditionne les choix précis de modèles
- Permet de fixer le protocole expérimental
- Ensuite, on peut réfléchir aux modèles

Premier modèle: linéaire

Dépendance linéaire entre l'entrée et la sortie

- Dans un espace vectoriel de dimension d, soit un point $x \in \mathbb{R}^d$
- $f_w(x) = \sum_{i=1}^{d} w_i x_i$
- f est le *modèle*
- w est le vecteur de *paramètres*

Apprentissage

- À partir des données
- Trouver le meilleur w

Abus de langage

- Linéaire: $y = ax$
- Affine: $y = ax + b$
- **Attention** en TP !

Régression linéaire

- Avec **UN** point: $\mathbf{x} = (x_1, \ldots, x_d)$
- Une étiquette $y \in \mathbb{R}$
- Approcher l'étiquette y avec un modèle $f(x)$

Apprentissage supervisé

- À partir de N points $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$
- Et leur label $y_1, \ldots, y_N$
- Chercher le meilleur compromis

Classification linéaire

- Avec **UN** point: $\mathbf{x} = (x_1, \ldots, x_d)$
- Une étiquette $y \in \{-1, +1\}$
- Retrouver l'étiquette y avec un modèle $\mathrm{signe}(f(x))$

avec $\mathrm{signe}(x) = +1$ si $x \leq 0$ et -1 sinon

Apprentissage supervisé

- À partir de N points $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$
- Et leur label $y_1, \ldots, y_N$
- Chercher le meilleur compromis

Géométriquement

- Tracer les points
- Tracer la frontière de décision
- À quoi correspond la somme ?

Plus proche voisin

Nearest neighbor

Idée

Associer un nouveau point à l'exemple connu le plus proche
[Au tableau]

k-plus proches voisins

Algorithme

- Pour un nouveau point
- Calculer toutes les distances entre ce point et les points connus
- Rechercher les k plus proches voisins
- Aggréger: vote (classfication), moyenne (régression)

[Au tableau]

Analyse

Complexité algorithmique

- Temps, espace

Performances

- Influence du k
- Score et généralisation

[Au tableau]

Prétraitements

[Au tableau]

Conclusion méthodologique

Contexte

- Apprentissage supervisé ou non
- Tâche de classification ou de régression

Modèles

- Linéaire
- k-Plus proches voisins

Prétraitements

- Données catégorielles
- Normalisation
- Transformation des données

Conclusion mathématique

Outils statistiques

- Moyennes, écart-type
- Distribution gaussienne
- Corrélation

Outils géométriques

- Droites, hyper-plans, surfaces
- Distances, produits scalaires

Conclusion informatique

Pandas

- Chargement des données
- Analyse préliminaire
- Statistique descriptive

Numpy

- Implémentation des algorithmes
- Bas-niveau
- Calcul matriciel **efficace**
- **JAMAIS** de boucle `for` de Python