

Annotation de bases de données

A. Carlier

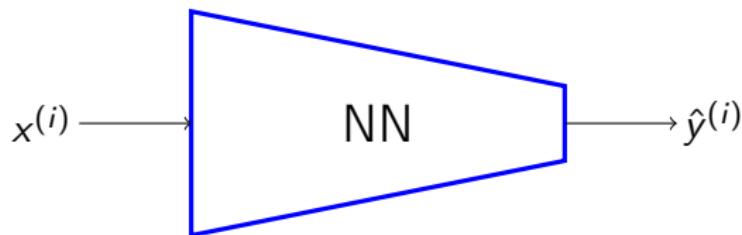
2023

Apprentissage supervisé

Dans le cadre de l'**apprentissage supervisé**, on dispose d'observations et de leurs étiquettes (appelées encore cibles (*target*), catégories ou *labels*) qui constituent un ensemble d'apprentissage. On le note :

$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}.$$

Les labels permettent d'enseigner à l'algorithme à établir des correspondances entre les observations et les labels.



Apprentissage supervisé

Aujourd'hui on parle de l'ensemble d'apprentissage !

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}.$$

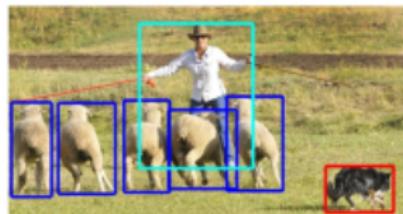
Comment l'obtenir ?

Un exemple illustratif : MS COCO (2014)

Objectif : constituer une base de données de scènes présentant de multiples objets, et localiser précisément (au pixel) ces objets.



(a) Image classification



(b) Object localization



(c) Semantic segmentation



(d) This work

[Lin et al. 2015] Microsoft COCO : Common Objects in Context

MS COCO : les données

Choix d'annoter uniquement des *things* (par opposition au *stuff*, l'arrière-plan).

Des enfants de 4 à 8 ans ont nommé des objets dans des photos aléatoires, pour constituer une base de labels.



[Lin et al. 2015] Microsoft COCO : Common Objects in Context

MS COCO : les données

Les images "non-iconiques" permettent aux algorithmes entraînés de mieux généraliser (Torralba et al. 2011).



(a) Iconic object images

(b) Iconic scene images

(c) Non-iconic images

Images collectées sur Flickr, en associant plusieurs mots-clés (ex : chien et chat)

[Torralba et al. 2011] Unbiased look at dataset bias

[Lin et al. 2015] Microsoft COCO : Common Objects in Context

MS COCO : les données

Phase de tri des images collectées pour ne conserver que les plus intéressantes (et les moins iconiques) :

Task: select images that contain **BOTH** a person AND a bicycle

Instructions:

Please click and select images that contain **BOTH** a person(s) **AND** a bicycle(s).
Do **NOT** select an image that contains **ONLY** a person(s) or **ONLY** a bicycle(s).
It is right to not select any image if none of image contains both categories.



You can re-select the image by clicking on it again.
Please do not select cartoons or paintings.



Task: select images that contain **a bear(s)**

Instructions:

Please click and select images that contain **MULTIPLE** objects **AND** at least one bear.



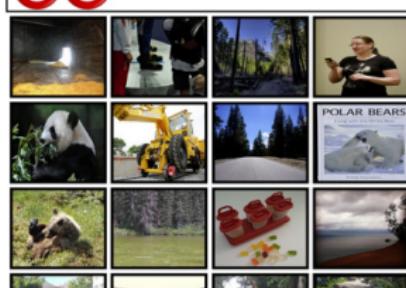
Do **NOT** select an image that contains **ONLY** a bear(s).



Do **NOT** select an image that contains **NO** bear(s).



You can de-select the image by clicking on it again.
Please do not select cartoons or paintings.



MS COCO : annotation des images

INSTRUCTIONS (PLEASE ACCEPT THE HIT TO GET STARTED):
This task is to select objects shown in the image. If an object has multiple instances, you can drag the box over any of the objects. There are 1 or more objects to drag over. Use the mouse or arrow keys to move through them. There are total 15 images to label.
Please drag and drop icons to matching objects in the image.

Here is an example of a labeled image:

Task: select **street-view** items shown in the image (if any):

Image 1.1

orange coordinates

Task: select **street-view** items shown in the image (if any):

Left Click: Add marker Right Click: Delete marker Drag & Drop: Move marker

(a) Category Labeling

INSTRUCTIONS (PLEASE ACCEPT THE HIT TO GET STARTED):
Mark each occurrence (if any) of the following object: cow:
• You only need to mark up to 10 instances if multiple cows exist in the image. It is possible for some images that this object does not appear.
• If there are multiple cows in the image, mark one instance of the object could be.
• The first is NOT ALWAYS correct.
• Type N to go to the next image and B to go back.
• There are 50 images in this HIT.

Good Example

Bad Example
(Do not click)

7 cow(s) found in this image.

cow

Left Click: Add marker Right Click: Delete marker Drag & Drop: Move marker

(b) Instance Spotting

Instructions: carefully trace around regions that have a **single sports ball** indicated by the icon. (14)

Please Accept HIT to get started! Complete Incomplete

Not sure what object sports ball is? Click on [here](#) to see examples!

Draw (D) Adjust (A) Undo (Ctrl-Z) Redo (Ctrl-Y) Close (Right-click) Delete (Delete)

Move to Target (M) Zoom In (I) Zoom Out (O) Reset Zoom (ESC)

Tips: Using "Move to target" (M) and "Zoom In" (I) for the small object!
Please pay attentions to trace boundary carefully. Work will be rejected if not follow the instruction.

(c) Instance Segmentation

[Lin et al. 2015] Microsoft COCO : Common Objects in Context

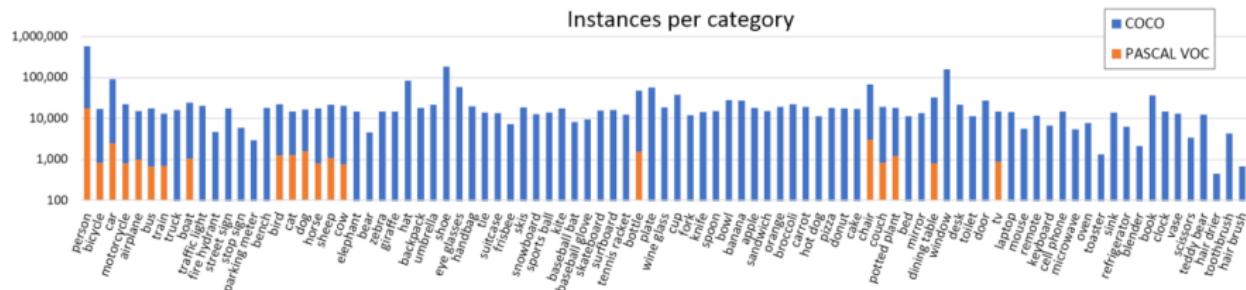
MS COCO : Annotation des images

- ① Annotation hiérarchique des images (niveau catégorie puis niveau label) : 8 *workers* par image, 20000 heures d'annotation.
- ② Désignation des instances : 8 *workers* par image, 20000 heures d'annotation.
- ③ Segmentation des instances (polygone) : phase de tutorial, vérification *a posteriori*, 1 *worker* par instance, 22 heures d'annotation pour 1000 instances.

Le processus nécessite en tout 85000 heures d'annotation !

[Lin et al. 2015] Microsoft COCO : Common Objects in Context

MS COCO : Statistiques



- 328 000 images, 91 labels, 2,5 millions d'instances segmentées.

[Lin et al. 2015] Microsoft COCO : Common Objects in Context

Quelle leçon tirer de cette procédure ?

- Constituer une base de données d'apprentissage est un travail au moins autant **fastidieux** (probablement plus) que de choisir et entraîner un réseau de neurones, et tout aussi **crucial** pour les performances finales du modèle !
- Plusieurs points clés doivent être considérés :
 - ▶ Le choix des données
 - ▶ L'interface d'annotation
 - ▶ Le choix des annotateurs
 - ▶ Le contrôle de la qualité des annotations

Choix des données

Annotation de bases de données d'apprentissage

Représentativité

Une difficulté chronique en apprentissage : constituer une base d'apprentissage **représentative** de la diversité des situations que le modèle peut rencontrer en phase d'inférence.
→ Impossible en pratique !

Il faut régulièrement mettre à jour la base d'apprentissage, en prenant en compte les cas identifiés d'erreur du modèle.
Le processus de constitution d'une base d'apprentissage n'est pas figé dans le temps !

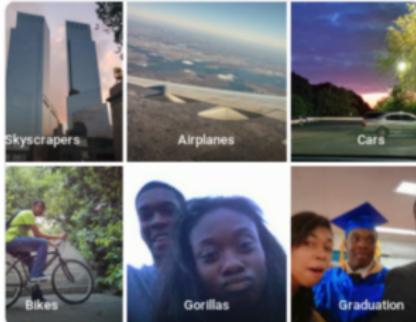
Représentativité

Une base de données non représentative peut occasionner des erreurs de prédiction très délicates...

 **Jacky lives on @jalcine@playviciou...**
@jackyalcine

Follow ▾

Google Photos, y'all fucked up. My friend's not a gorilla.



Skyscrapers Airplanes Cars
Bikes Gorillas Graduation

6:22 PM - 28 Jun 2015

3,281 Retweets 2,309 Likes

233 3.3K 2.3K

Choix des données

Fairness

La constitution de bases d'apprentissage sans biais (*fairness*) est un champ de recherche important de l'IA.



[Yang et al. 2020] Towards Fairer Datasets : Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy

Apprentissage Actif

Le choix des données à annoter peut être conduit à partir du modèle lui-même, c'est ce que l'on appelle l'**apprentissage actif**.

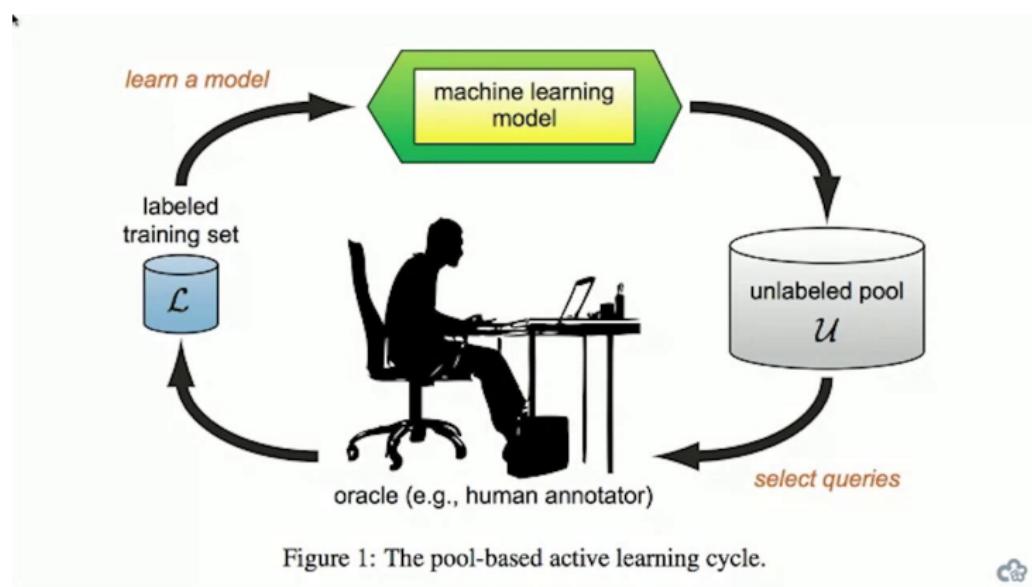
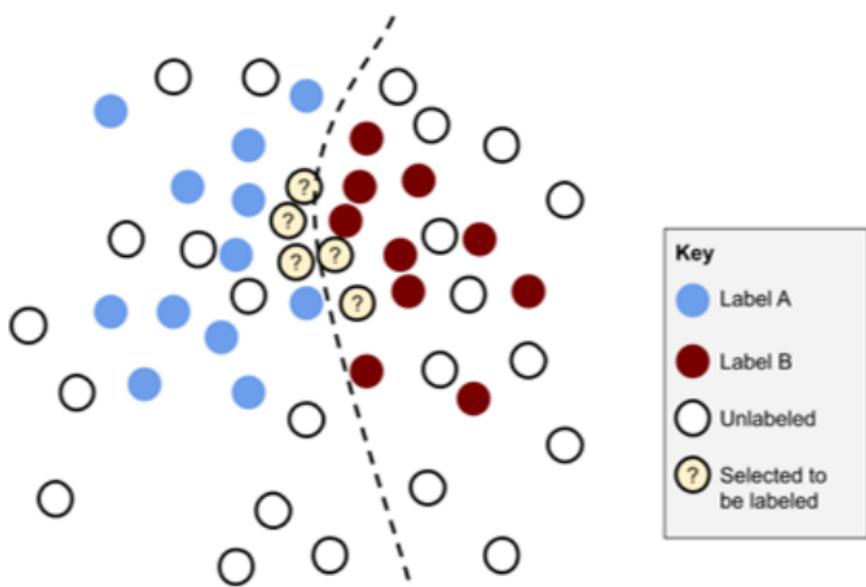


Figure 1: The pool-based active learning cycle.

Les données sont choisies dans le but d'optimiser un certain critère.

Apprentissage Actif - *Uncertainty Sampling*

Choix des données pour lesquelles les prédictions du modèles sont les plus incertaines (par exemple, dont l'entropie de prédiction est maximale).



Apprentissage Actif - *Density Sampling*

Ce choix peut être pondéré par un critère de densité pour éviter de sélectionner des données non représentatives.

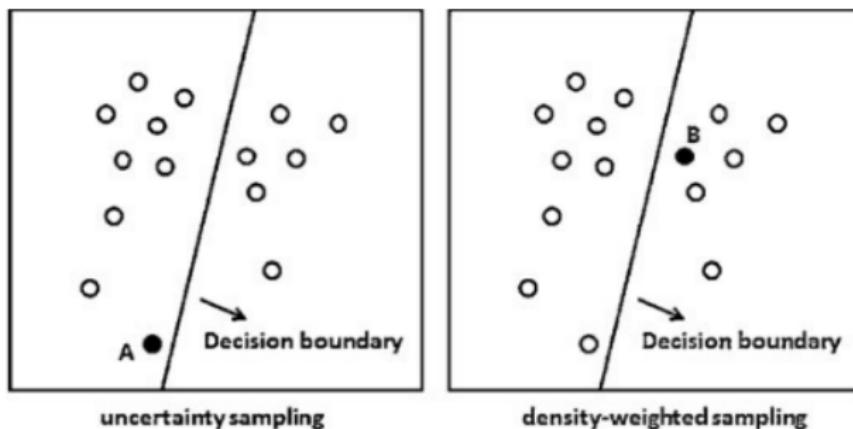


Image de [Cai 2015] Active learning for ranking with sample density

Apprentissage Actif - BADGE

Les méthodes de l'état de l'art, comme BADGE combinent les critères d'incertitude et de représentativité via un clustering des données.

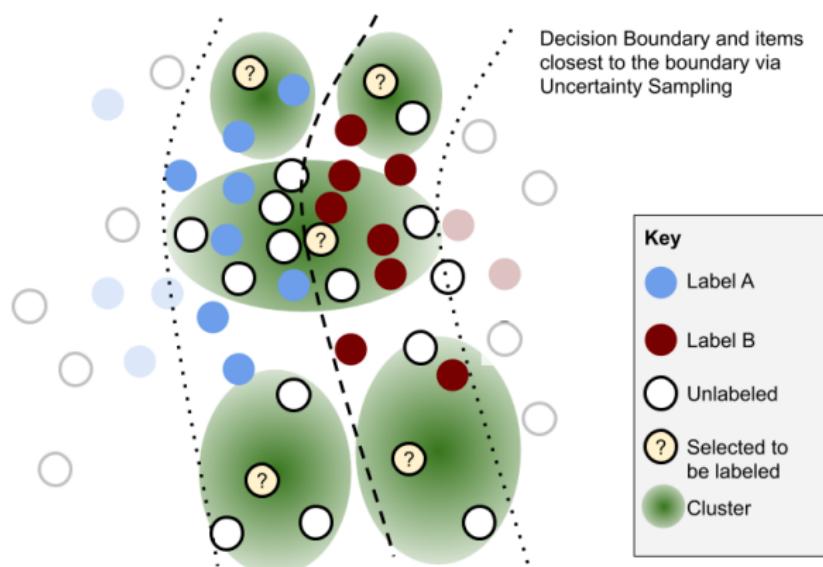


Image de [Monarch 2021] Human-in-the-Loop Machine Learning

[Ash 2020] Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds

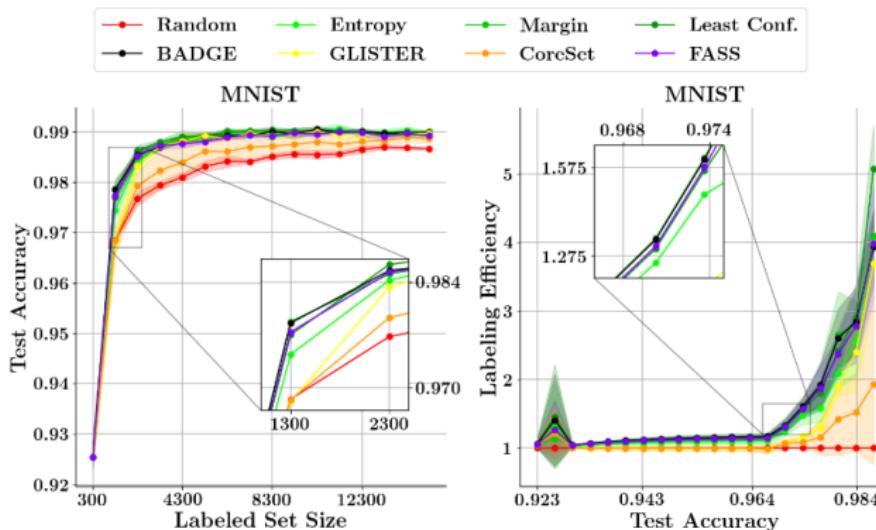
Apprentissage Actif - DISTIL

Une (parmi d'autres !) librairie pour démarrer en apprentissage actif :



DISTIL

Deep Diversified Interactive Learning

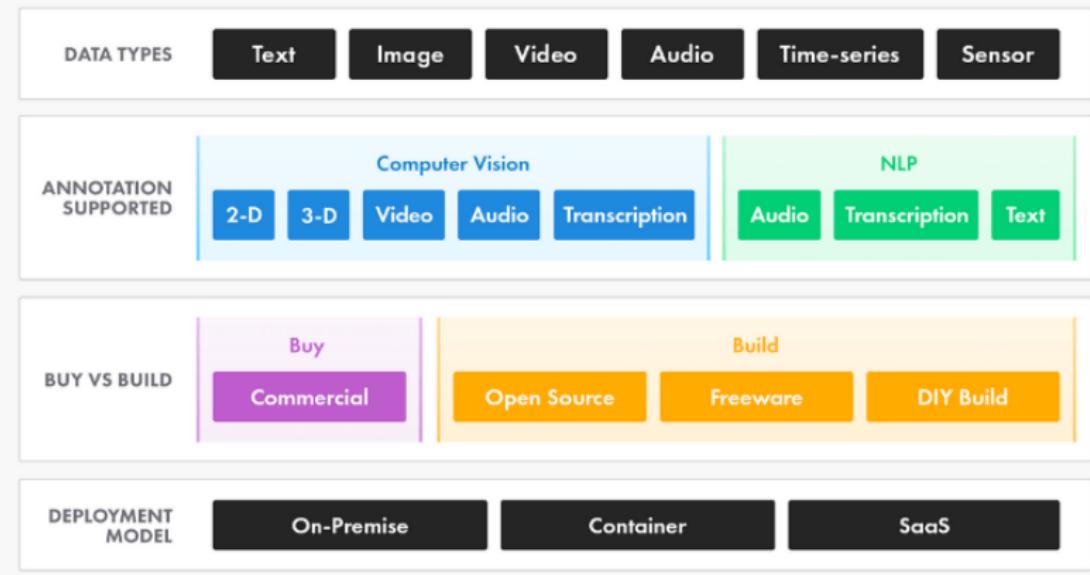


Interfaces d'annotation

Annotation de bases de données d'apprentissage

Solutions commerciales

Data Annotation Tools: Key Elements



Solutions commerciales

Commercial Data Annotation Tool	Annotation Supported							Deployment Model		
	Computer Vision					NLP				
	2D	3D	Video	Audio	Transcription	Text	Transcription	On-premise	Container	SaaS
Annotell	✓	✓							✓	✓
Dataloop AI	✓		✓	✓	✓				✓	✓
Datasaur AI						✓	✓			✓
Deepen AI	✓	✓	✓					✓	✓	✓
Hasty	✓				✓				✓	✓
Hivemind					✓	✓	✓	✓		✓
LightTag						✓				✓
UnderstandAI	✓	✓							✓	✓
V7 Labs Darwin	✓	Volumetric	✓						✓	✓

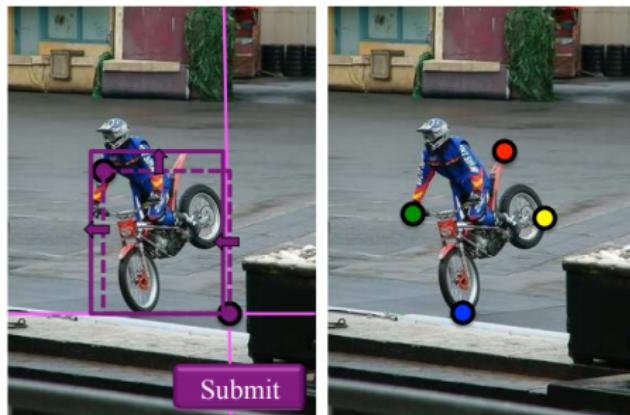
Solutions gratuites / open-source

Open Source Data Annotation Tool	Annotation Supported				Deployment Model				Key Features
	2-D	Video	Audio	Text	On-premise	Container	HTML	Other	
CVAT	✓	✓			✓	✓			<ul style="list-style-type: none">CV: Bounding box, polygon, polylineNLP: Multiple text inputsSingle & consensus review
Fiji	✓						Compiled		
LabelImg	✓				✓				<ul style="list-style-type: none">Graphic annotation, labeled bounding boxes
LabelMe	✓				✓				<ul style="list-style-type: none">Semantic segmentation
VoTT	✓	✓						.exe, .dmg, .snap	
VGG Oxford University	✓	✓	✓	✓	✓	✓			<ul style="list-style-type: none">CV: Bounding box, bounding circle/ellipse, polygon, polyline, 2-D pointNLP: Multiple text inputsSemantic segmentation

Image de <https://www.cloudfactory.com/data-annotation-tool-guide>

Interface d'annotation - *Extreme Clicking*

Ne pas hésiter à remettre en question des interfaces naturelles :



- 7 secondes par boîte en moyenne contre 35 secondes pour la méthode "classique", et une information supplémentaire : 4 points sur les contours de l'objet.
- utilisé pour annoter OpenImage V4, comprenant 15 millions de boîtes englobantes.

Interface d'annotation intelligentes

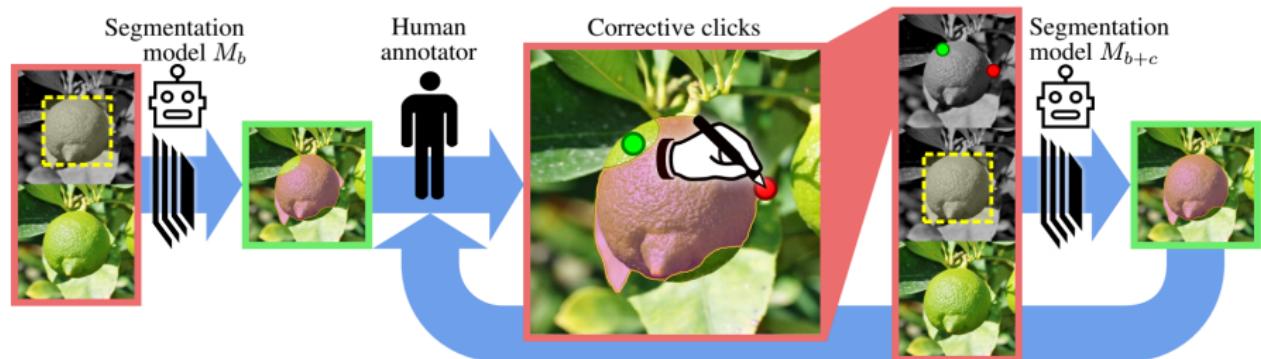
Intégration de modèles pré-entraînés dans la boucle d'annotation :



- ➊ Partant d'une boîte englobante initiale, le modèle prédit un masque de segmentation.
- ➋ L'annotateur clique les régions mal annotées pour les corriger.
- ➌ Le modèle met à jour sa prédiction en intégrant les nouvelles informations.
- ➍ Retour au point 2 jusqu'à ce que l'annotateur soit satisfait.

Interface d'annotation intelligentes

Intégration de modèles pré-entraînés dans la boucle d'annotation :



Cette méthode permet d'obtenir des masques plus précis que ceux de MS COCO, en 3 fois moins de temps !

OpenImage V5 comprend 2.8 millions d'instances de 350 classes.

[Benenson et al. 2019] Large-scale interactive object segmentation with human annotators

Annotateurs

Annotation de bases de données d'apprentissage

Annotateurs



Crowdsourcing

Peu chers
Nombreux
Annotations bruitées



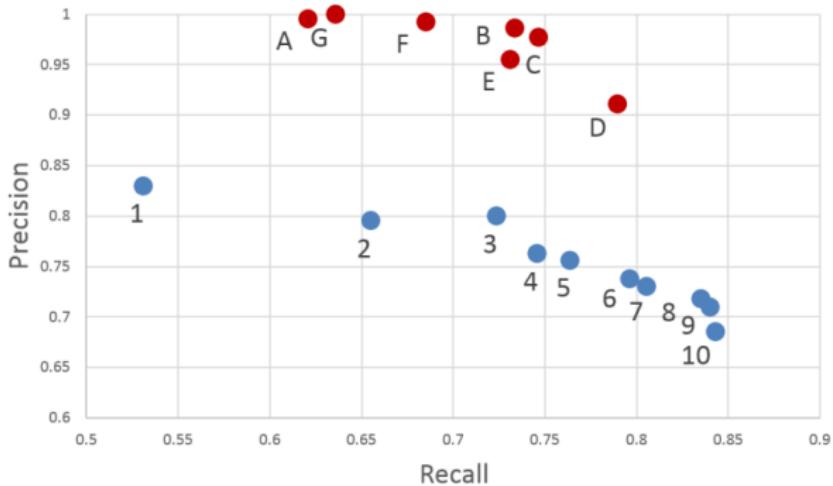
Experts

Chers
Rares
Annotations correctes

Problèmes à considérer : **éthique**, sensibilité des données

MS COCO (2014)

Precision/recall for **Experts** and aggregates of **Workers**



[Lin et al. 2015] Microsoft COCO : Common Objects in Context

Crowd - Outsourcing

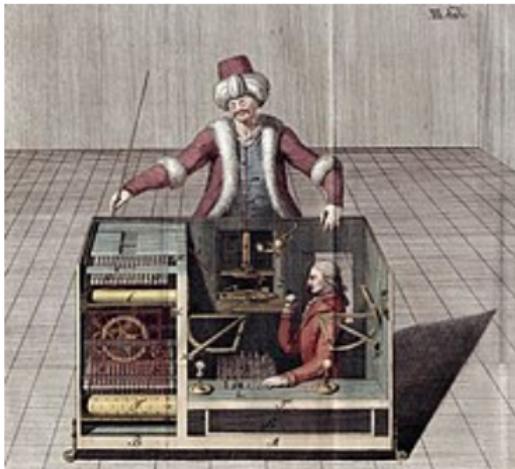
L'action de déléguer une tâche d'ordinaire dévolue à un expert à une foule de personnes non-expertes.

Quelques exemples :

- TextBroker, pour faire rédiger des textes par des écrivains (quelques centimes/mot)
- Creads, conception graphique (logo, identité visuelle, etc.)
- Zooppa, réalisation de vidéo
- Kaggle, résolution de problème de *machine learning*

[Howe 2006] The rise of crowdsourcing

Crowdsourcing... en pratique



All HITs Your HITs Queue

amazon
beta
mechanical turk

HIT Groups (1-20 of 640)

Show Details Hide Details Items Per Page: 20

Requester	Title	HITs	Reward	Created	Actions	
ScoutIt	Classify Receipt	151	\$0.03	14s ago	Preview	Qualify
Crowdsurf Support	Full Text Review - Earn up to \$...	53	\$0.17	3m ago	Preview	Qualify
Laura A. King	Personality, Information Proce...	1	\$0.15	4m ago	Preview	Accept & Work
Crowdsurf Support	Review, edit, and score the tra...	1,091	\$0.02	5m ago	Preview	Qualify
Erica Fissel	Quick Demographic Survey! (~...	1	\$0.01	6m ago	Preview	Accept & Work

Crowdsourcing... en pratique

<i>Country</i>	<i>Gender</i>	<i>Worker Count</i>	<i>Median Wage</i>	<i>Mean Wage (SD)</i>
U.S.	Female	527	2.79	3.47 (7.87)
U.S.	Male	288	3.68	3.77 (2.76)
India	Female	61	1.37	2.01 (3.19)
India	Male	237	1.43	3.58 (24.6)

Quelques dollars de l'heure en moyenne (sans couverture sociale)

[Hara et al. 2019] Worker Demographics and Earnings on Amazon Mechanical Turk : An Exploratory Analysis

Crowdsourcing... implicite !

Completely Automated Public Turing Test to tell Computers and Humans Apart.



[Von Ahn et al. 2008] reCAPTCHA : Human-based character recognition via web security measures

Crowdsourcing... implicite !

Completely Automated Public Turing Test to tell Computers and Humans Apart.

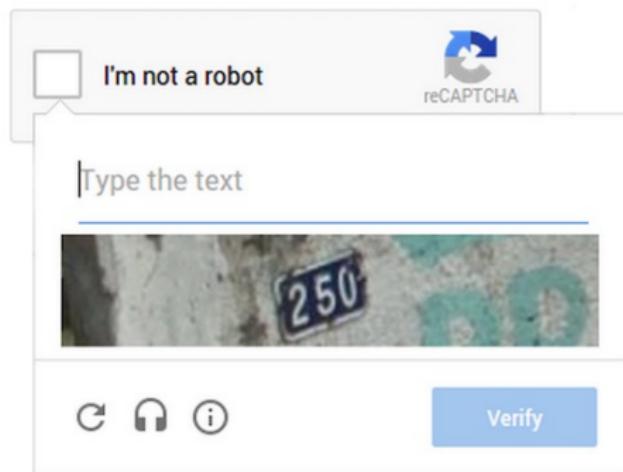


ReCAPTCHA : permet à la fois de valider un test de Turing et de labelliser un texte scanné !

[Von Ahn et al. 2008] recaptcha : Human-based character recognition via web security measures

Crowdsourcing... implicite !

Completely Automated Public Turing Test to tell Computers and Humans Apart.



ReCAPTCHA : permet à la fois de valider un test de Turing et de labelliser les numéros de rue de StreetView !

[Von Ahn et al. 2008] recaptcha : Human-based character recognition via web security measures

Crowdsourcing... implicite !

Completely Automated Public Turing Test to tell Computers and Humans Apart.

Please help us improve our search results by identifying the images that contain mountains.

Select all images with mountains.
Click verify once there are none left.

Annotations:

- A red box highlights the first image in the top row.
- The 'VERIFY' button is visible at the bottom right.

Annotateurs

Crowdsourcing... implicite !



Un jeu basé sur le principe de l'*output agreement*.

[Von Ahn et al. 2004] Labeling Images with a Computer Game

Crowdsourcing... implicite !

The screenshot shows the Duolingo Spanish learning interface. At the top, there's a navigation bar with 'duolingo' and tabs for 'Accueil', 'Mots', 'Activité', and 'Discussion'. Below this, a main section titled 'Unités d'espagnol' (Level 1) displays various language units:

- Bases**: An orange circle with a flame icon. A red arrow points to it from the text 'Une unité déjà apprise'.
- Expressions**: A green speech bubble icon. Below it, 'Expressions 0/3' is shown.
- Bases 2**: A red circle with a chick icon. Below it, 'Bases 2 0/4' is shown. A red arrow points to it from the text 'Deux unités à apprendre'.
- Nourriture**: A grey circle with a sandwich icon. Below it, 'Nourriture 0/5' is shown.
- Animaux**: A grey circle with a cat icon. Below it, 'Animaux 0/4' is shown.
- Pluriels**: A grey circle with a lamp icon. Below it, 'Pluriels 0/1' is shown.
- Possessifs**: A grey circle with a book icon. Below it, 'Possessifs 0/2' is shown. A red arrow points to it from the text 'Quatre unités à apprendre qui ne deviendront accessibles qu'une fois les deux précédentes étudiées'.

[Von Ahn et al. 2013] Duolingo : learn a language for free while helping to translate the web.

Attention à la qualité des données...



[Carlier et al. 2016] Assessment of crowdsourcing and gamification loss in user-assisted object segmentation

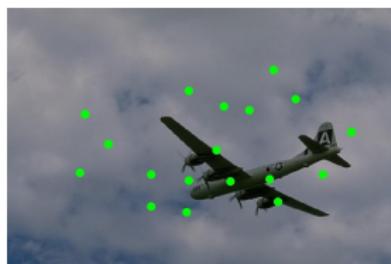
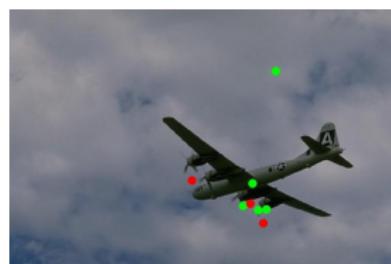
Contrôle qualité

Annotation de bases de données d'apprentissage

Quelles erreurs veut-on corriger ?

De manière générale, on peut regrouper les erreurs d'annotation en trois grandes familles :

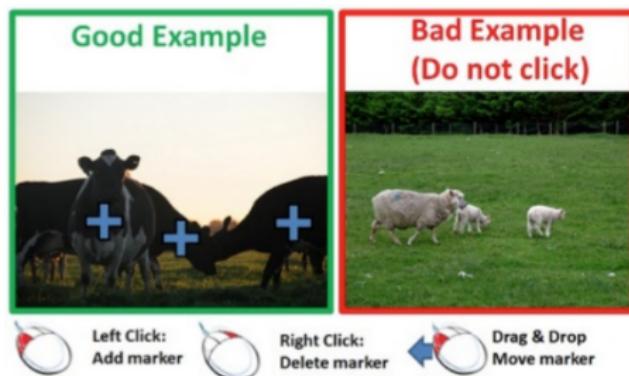
- Les erreurs occasionnelles (inattention, fatigue, méconnaissance, etc.).
- Les erreurs dues à une mauvaise compréhension de la tâche.
- Les erreurs intentionnelles.



[Oleson et al. 2011] Programmatic Gold : Targeted and Scalable Quality Assurance in Crowdsourcing

Phase de tutorial

Pour s'assurer de la bonne compréhension de la tâche par des utilisateurs, il est préférable de les faire passer par une phase de tutorial expliquant pas à pas les différents aspects de la tâche.



Une règle de base :

L'annotateur ne lit pas les consignes !

La qualité par la redondance

Idée : proposer la même tâche à de multiples annotateurs et synthétiser les réponses.

Plusieurs objectifs possibles :

- Simplement obtenir une étiquette consolidée (moyenne, médiane, etc.)
- S'assurer de ne rien oublier (augmenter le rappel), comme c'était le cas dans le processus décrit plus tôt de MS COCO.
- Utiliser la redondance pour qualifier la qualité des annotateurs.

Estimation de la qualité des annotateurs

Idée : estimer conjointement un label pour chaque donnée et la qualité des annotateurs.

- ① **Expectation** : En considérant connue la qualité des annotateurs, estimer le label des données (par exemple, en choisissant la moyenne des annotations pondérée par la qualité des annotateurs).
- ② **Maximization** : En considérant connue le label des données, estimer la qualité des annotateurs (par exemple, en choisissant la moyenne des écarts de ses annotations au label synthétisé).

De nombreuses méthodes ont été développées dans la même veine.

[Dawid and Skene 1979] Maximum Likelihood Estimation of Observer Error Rates using the EM algorithm

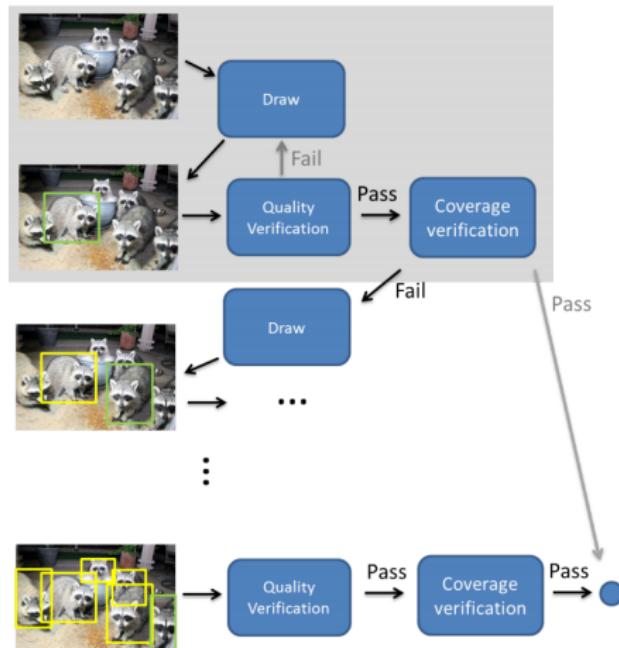
Estimation de la qualité des annotateurs

Une solution alternative : insérer des données dont le label est connu (*gold standard*) pour évaluer la qualité de certaines des annotations et extrapoler la qualité des annotateurs.



Find-Fix-Verify

Pour certains problèmes, il est plus efficace de concevoir une série de tâches incrémentales qu'une seule tâche donnée à plusieurs utilisateurs.



Retour d'expérience : GISHWHES

Greatest International Scavenger Hunt the World Has Ever Seen



Une chasse au trésor organisée par Misha Collins (Supernatural...) avec plus de 3000 équipes de 5 personnes partout dans le monde.
1 semaine, plus de 200 tâches.

Retour d'expérience : GISHWHES

Quelques tâches :

You know the saying, "No man is an island." Prove it wrong.

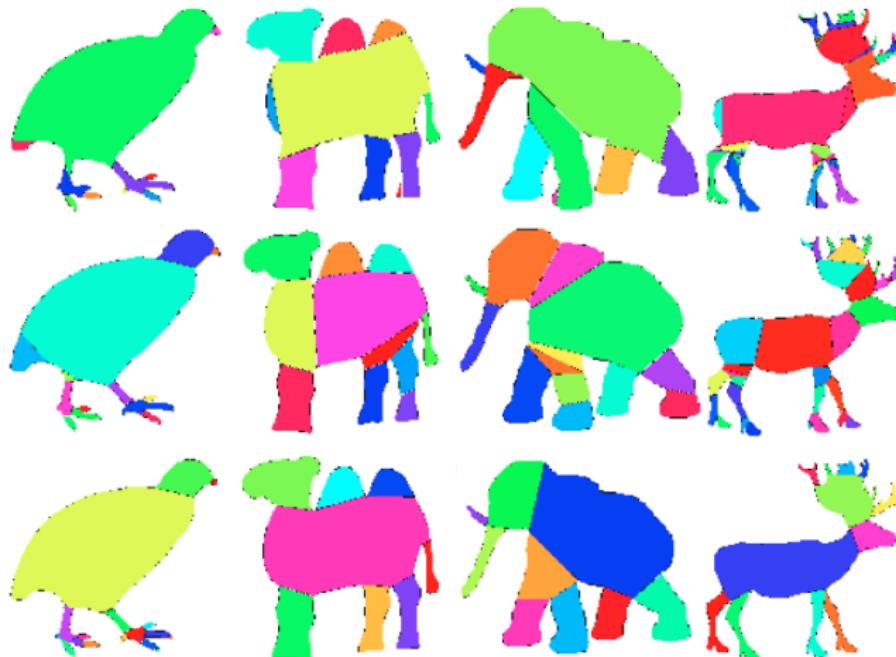


Prove there's a ghost in a Starbucks working as a barista.



Retour d'expérience : GISHWHES

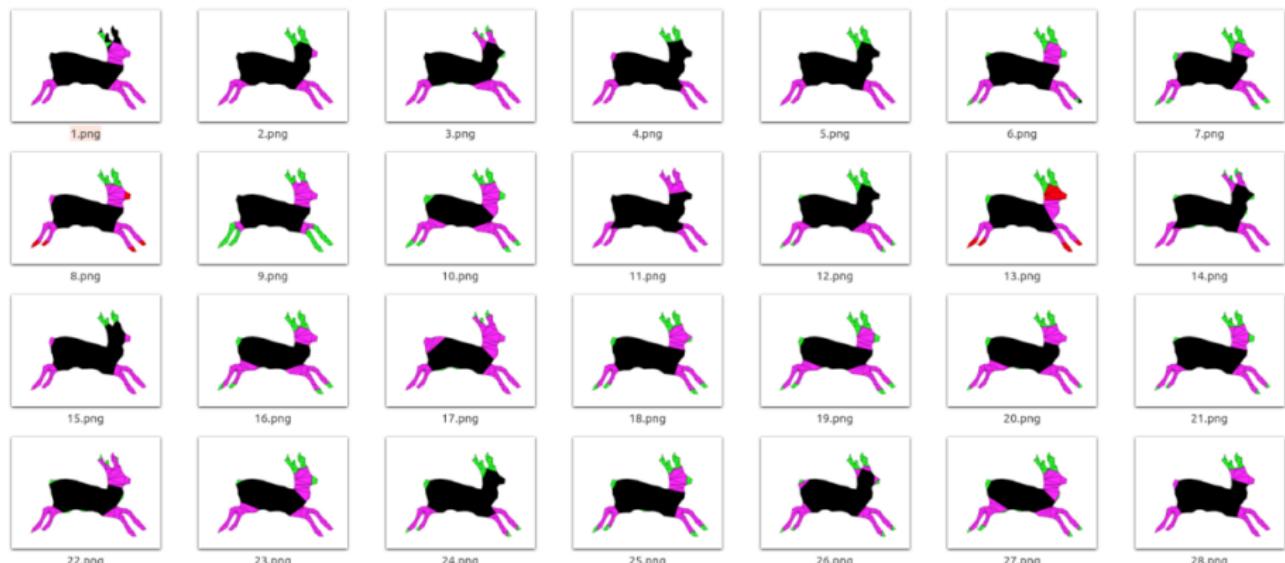
Question de Géraldine (et ses collègues) : qu'est-ce qu'une bonne décomposition pour une forme 2D ?



Retour d'expérience : GISHWHES

Question de Géraldine (et ses collègues) : qu'est-ce qu'une bonne décomposition pour une forme 2D ?

Réponse de la foule :



[Carlier et al. 2016] The 2D shape structure dataset : A user annotated open access database

Retour d'expérience : GISHWES

Une tâche simple : colorier la forme.

Home Task Tutorial Participate

Shape

34% completed

Controls

Color	Description	Paint it all
Black	Main Shape	
Pink	Parts	
Green	Details	

Find Remaining Triangle The Rest is Details

Status

Team Name : test

Current task : 1/20

Retour d'expérience : GISHWHES

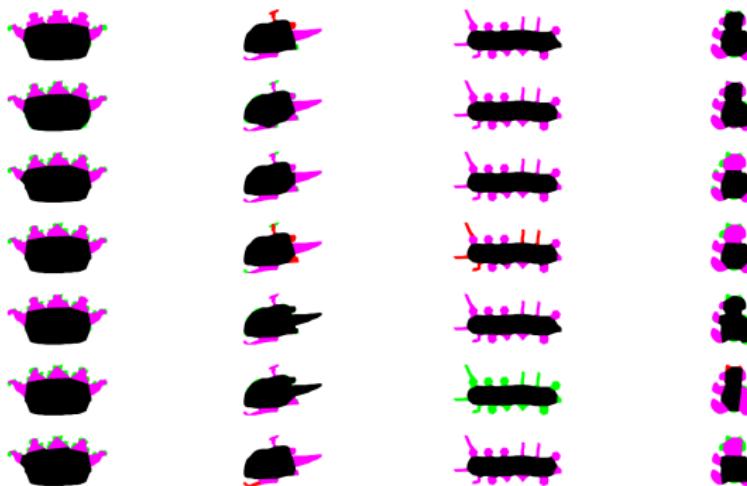
Contrôle qualité triple :

- ① Un tutorial pas à pas pour s'assurer de la bonne maîtrise des différents éléments de l'interface.
- ② Des formes pré-annotées pour évaluer la qualité des annotateurs.
- ③ De la redondance !

[Carlier et al. 2016] The 2D shape structure dataset : A user annotated open access database

Retour d'expérience : GISHWHES

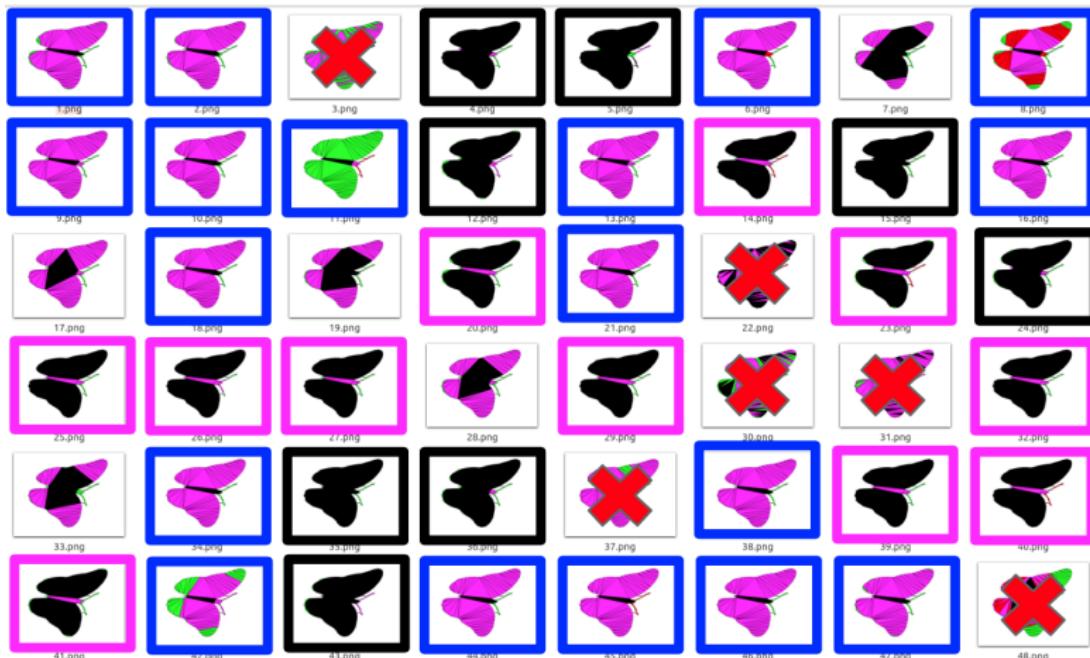
Gold Standard :



[Carlier et al. 2016] The 2D shape structure dataset : A user annotated open access database

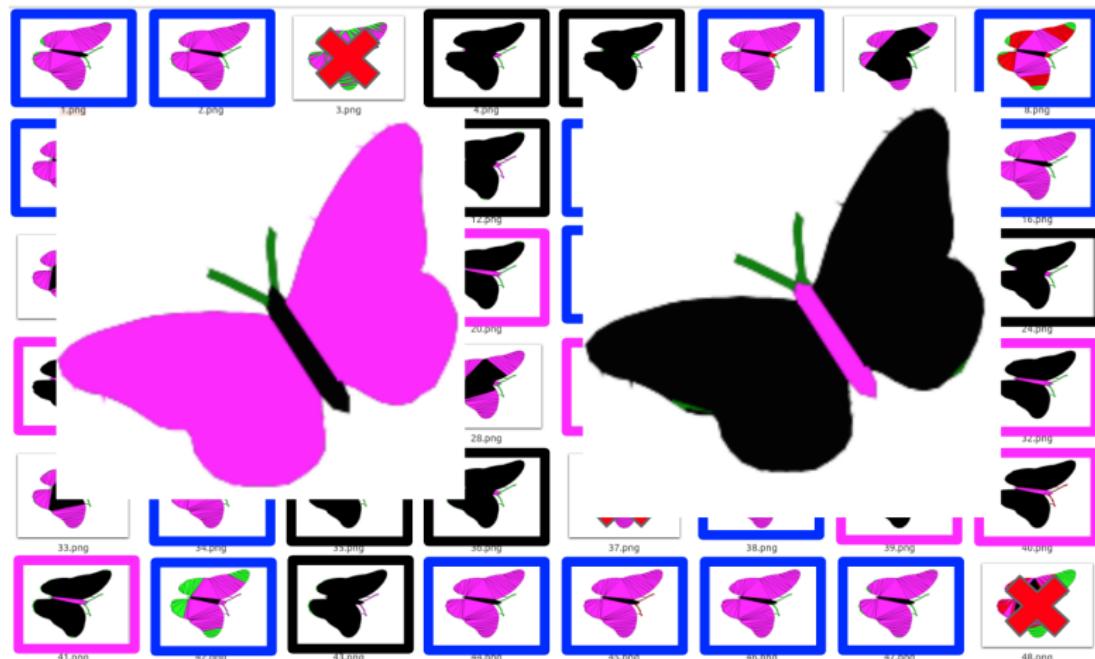
Retour d'expérience : GISHWES

Fusion des annotations par un vote de majorité précédé d'un clustering (spectral) des formes :



Retour d'expérience : GISHWHES

Fusion des annotations par un vote de majorité précédé d'un clustering (spectral) des formes :



Retour d'expérience : GISHWHES

Bilan :

- 1253 formes réparties en 70 catégories (formes issues de la base MPEG-7)
- 2861 équipes ont démarré la tâche, et 1877 l'ont terminée (20 annotations).
- Au total, au moins 24 annotations différentes pour chaque forme.

Résultats visualisables ici : <https://2dshapesstructure.github.io/>