DATASCIENCE, LEARNING AND APPLICATIONS

DALAS - Protocole d'évaluation en ML

18 mars 2024

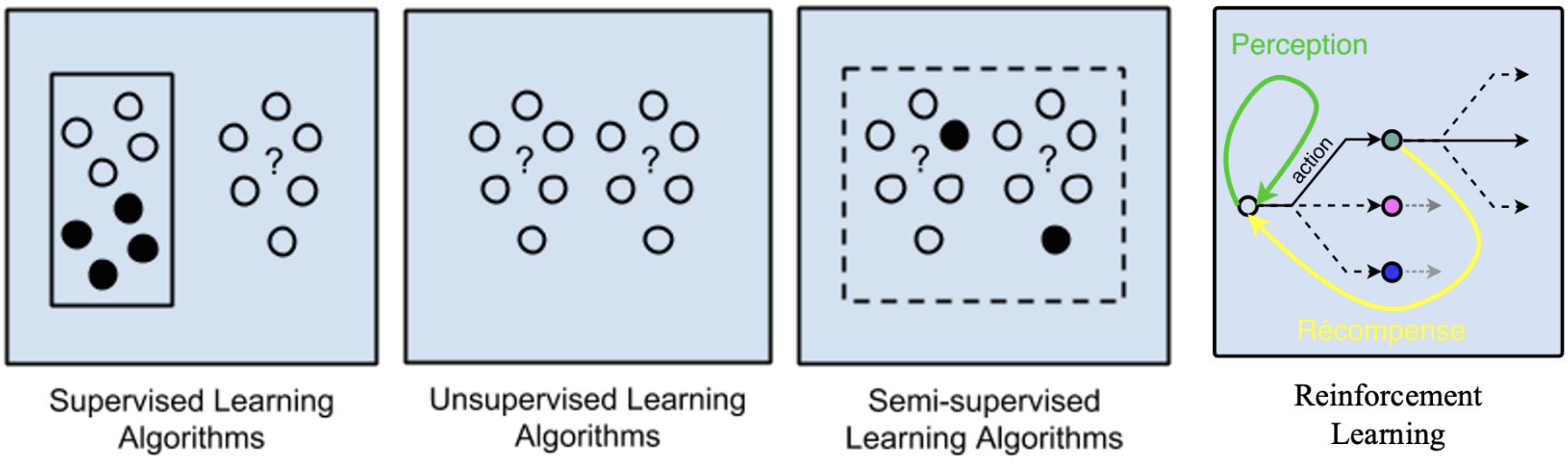
Laure Soulier

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

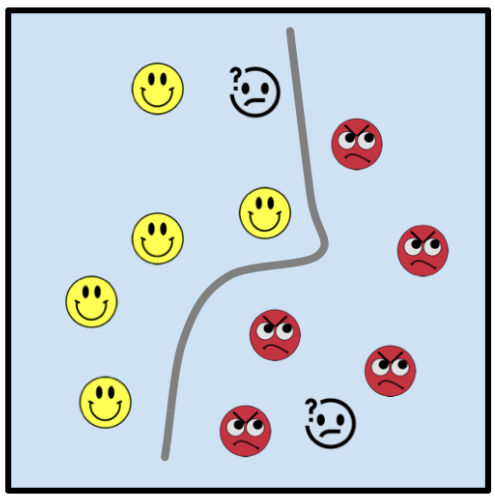
Savoir décrire une tâche

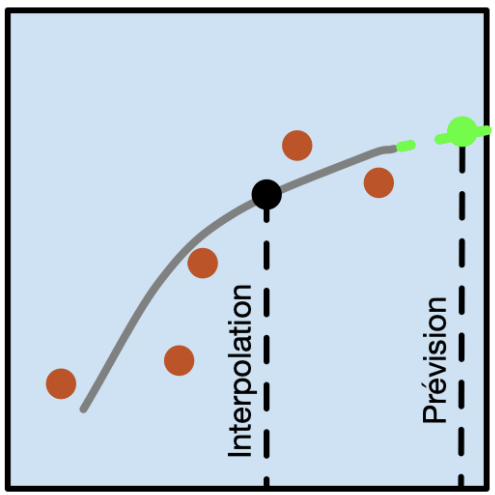
- Les différentes catégories d'apprentissage

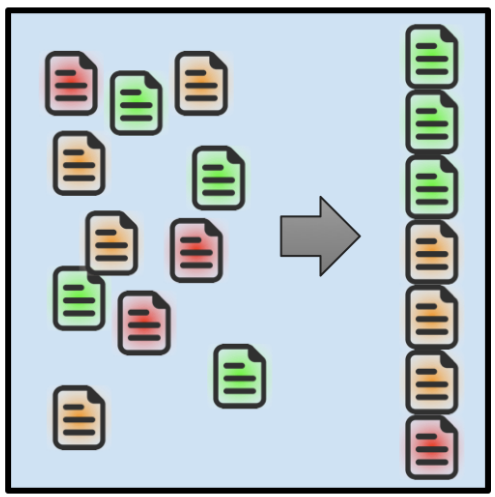


Savoir décrire une tâche

- Les différents types de tâches





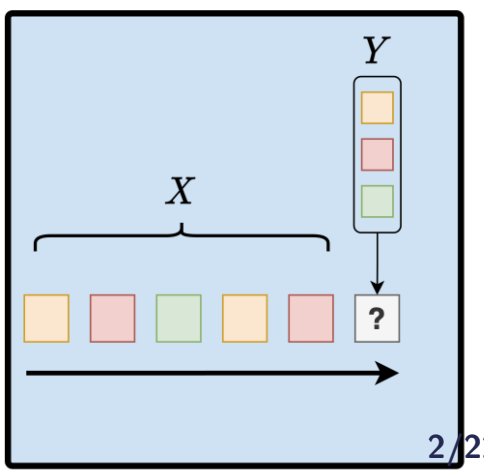


→ Classification

→ Régression

→ Ordonnancement

→ IA Générative: un cas assez classique

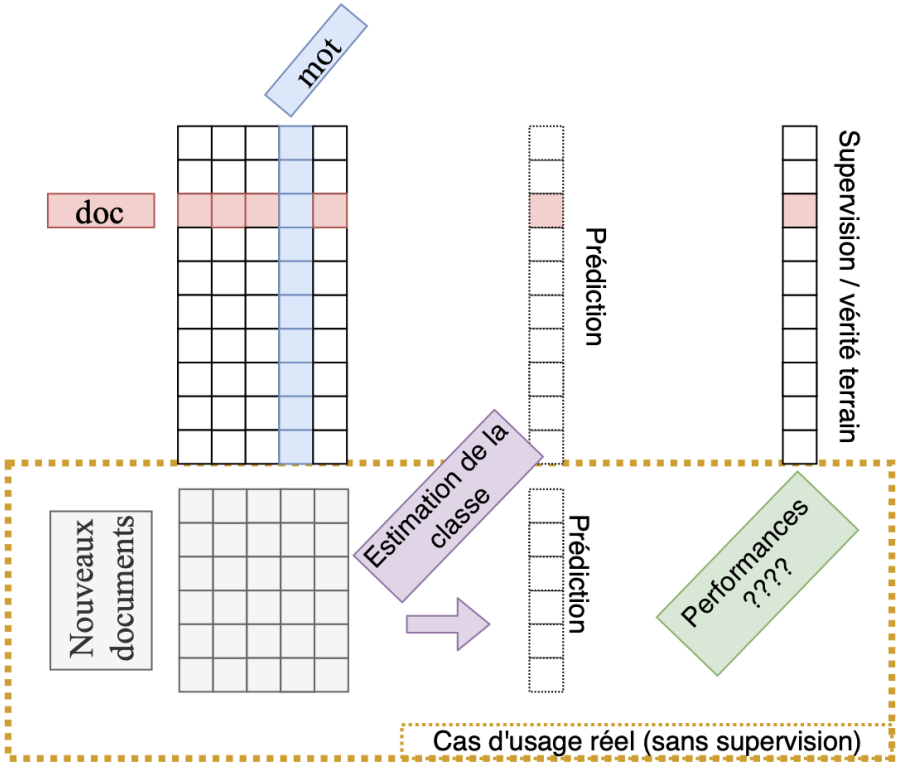


Les composants d'un protocole de machine learning

- Jeux de données
- Métriques
- Modèles / variantes / modèles de référence
- Stratégies d'entraînement et d'évaluation

→ Evaluation : besoin de données neutres

→ Conception d'une IA
  → L'expert explique un problème
  → Le data-scientist optimise le processus de décision
  → L'expert fixe le seuil d'acceptabilité du produit



Les jeux de données

- Correspondance tâche vs. vérité terrain des jeux de données

Tâche	Prédiction	Labels
Classification	Classes	Classe
Regression	Nombre	Nombre
Ordonnancement	Ordonnancement via une proba de pertinence / Score de similarité	Classe de pertinence
Génération	Classe (mot)	Classe (mot)

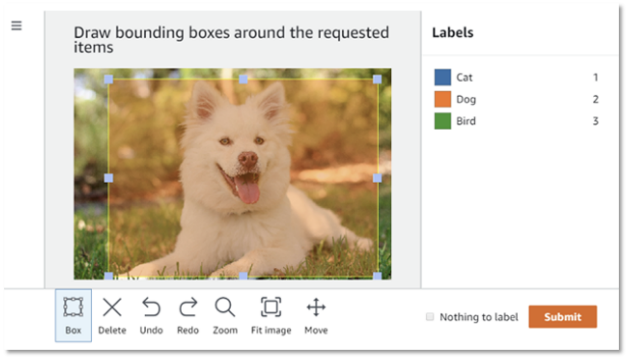
- Taille des jeux de données
  - Evaluer le nombre d'instances d'entraînement
  - Instances d'entraînement peut être différent du nombre d'entités dans le jeu de données
  - Exemple : ordonnancement

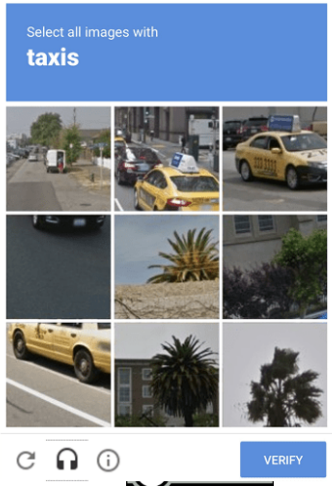
Les jeux de données

- Choix du jeu de données : quelques indicateurs
  - Taille
  - Diversité des données
  - Qualité de l'annotation
  - Performance des modèles de référence (si fournis)
  -
- Construction du jeu de données
  - Base d'exemples : images/textes/séries temporelles existantes
  - Construction de l'annotation : inférée/simulation, annotation humaine
  - Tout construire de zéro : expérimentation utilisateur ("user study"), création des cas d'usage, des données et des types de supervision à déduire de l'expérimentation

Les types d'annotation

- Annotation manuelle
  - Mechanical turk
  - Mobileyes ….
- Captcha
- Supervision faible / distante





Ligne 1 RATP @Ligne1_RATP · 16 h
08:28, le trafic est rétabli sur la ligne (incident technique) #RATP #ligne1
1 2

Ligne 1 RATP @Ligne1_RATP · 16 h
07:51, le trafic est perturbé sur la ligne (incident technique) #RATP #ligne1
2 3

Simulation

- Utilisation de sources de données existantes mais utilisées dans un but différent
  - Simulation de la collaboration à partir de logs individuels
  - Construction d'une session de recherche à partir de conversation dans un forum
  - Dégradation d'un jeu de données existant : génération de données bruitées à partir de données réelles, ...
- Modélisation d'un comportement utilisateur pour la réalisation d'une tâche
  - Besoin de modèles du monde, de modèles cognitifs

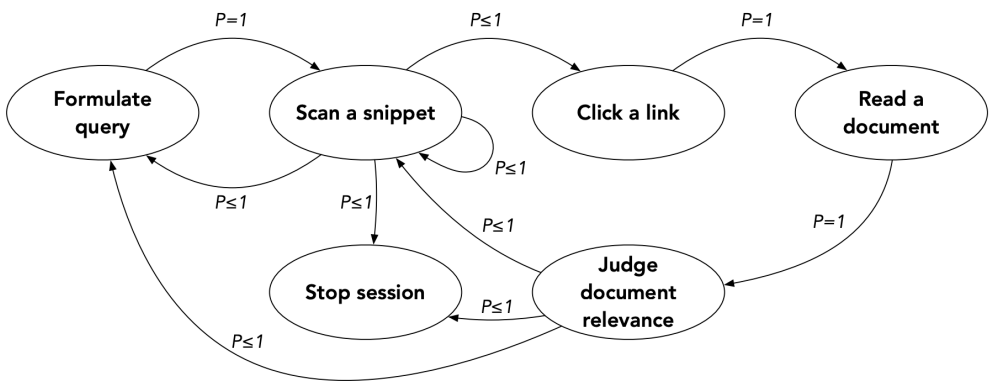


Figure 6.2: Automaton expressing the subtasks performed during a search session, according to Baskaya *et al.* (2013) (illustration adapted from (Baskaya *et al.*, 2013)).

Annotation humaine

- Choix des juges : les copains ? des gens aléatoires ? des experts ? sur des plateformes ?
- Sur quelles plateformes ?
  - Plateformes internes
  - Plateformes en ligne (crowdsourcing) : Mechanical Turk, Prolific, ...
- Découper les tâches complexes en une multitude de sous-tâches
- Nécessité de vérifier la qualité des annotations (questions "gold standard" dont on connait la réponse)
- La formation des annotateurs à la tâche est importante
- Envisager la pré-annotation : gain de temps
- Législation et éthique : réglementation, comité d'éthique, ...
- Fournir une documentation du processus d'annotation

Qualité de l'annotation

- Nécessité d'avoir plusieurs juges par instance et de mesurer le taux d'accord/de consensus
  - Accord inter-annotateur : mesurer l'accord entre annotateurs
    - matrice de confusion (proba sur la diagonale)
    - Kappa de Cohen quand 2 juges - permet de prendre en compte le biais humain https://datatab.fr/tutorial/cohens-kappa
    - Kappa de Fleiss si plus de 2 juges https://datatab.fr/tutorial/fleiss-kappa
  - Accord intra-annotateur : le long du processus d'annotation pour vérifier la cohérence/constance d'un utilisateur

SCIENCES SORBONNE UNIVERSITÉ

- Nécessité d'intégrer l'évaluation

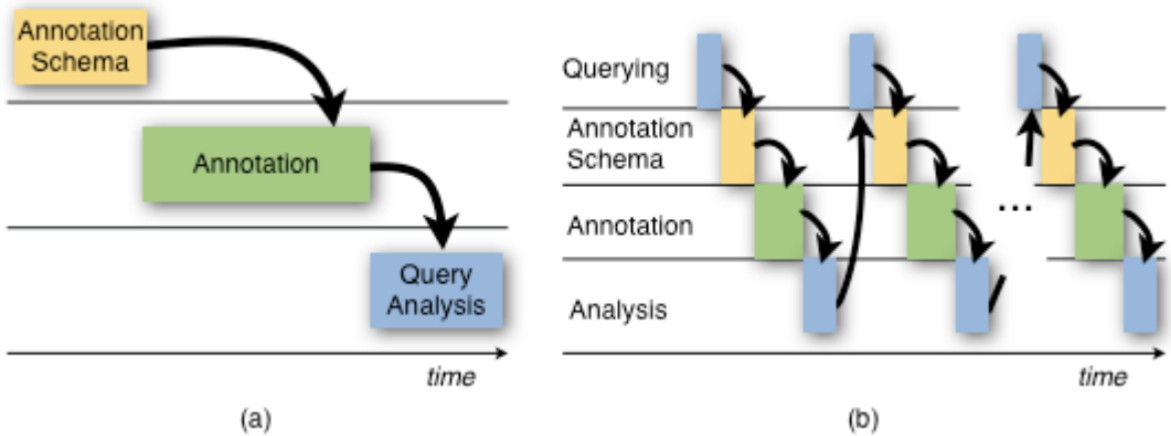
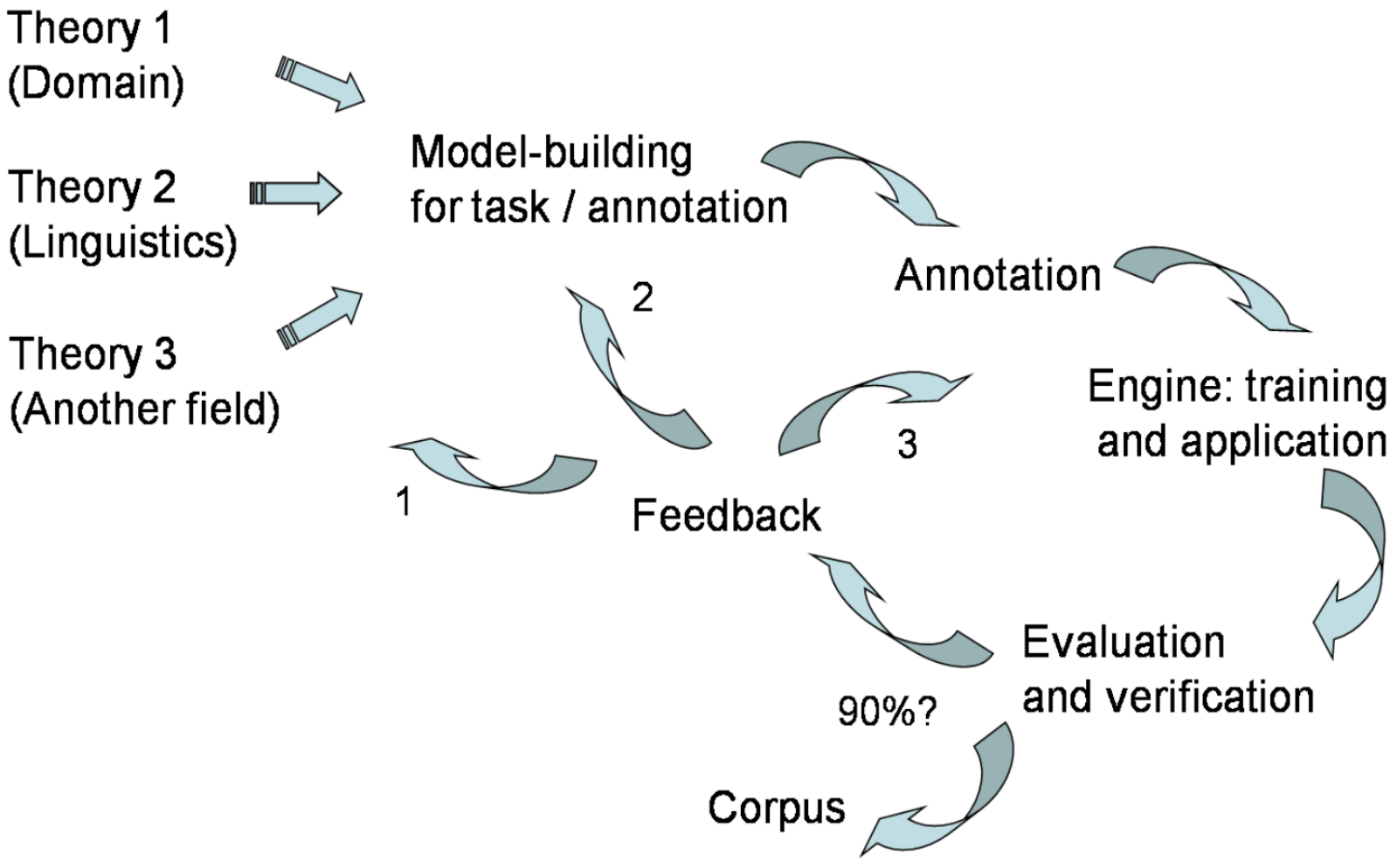


Figure 1: The phases of traditional corpus creation (a) and the cyclic approach in agile corpus creation (b). Reproduction of Figure 2 in Voormann and Gut (2008).

Figure 1 – "Agile Corpus Annotation in Practice : An Overview of Manual and Automatic Annotation of CVs". Alex et al. 20210

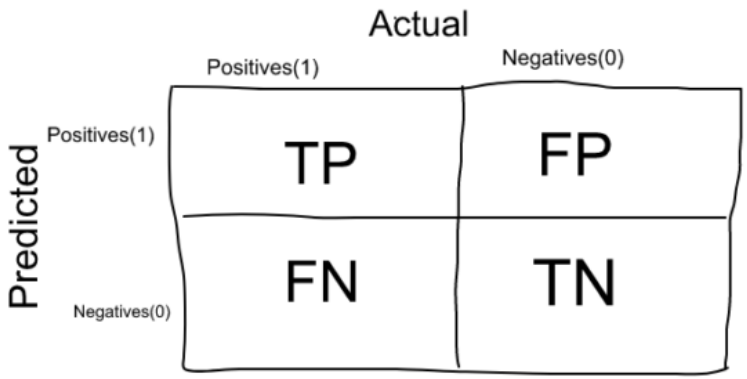
SCIENCES SORBONNE UNIVERSITÉ

- Nécessité d'intégrer l'évaluation



Les métriques

- Classification
  - Matrice de confusion
  - Précision, Rappel, F-mesure, Accuracy
  - Sensibilité (proportion de vrais positifs parmi les personnes à identifier)
  - Spécificité (proportion de vrais négatifs chez les négatifs)
  - Courbe ROC (dépend d'un seuil)
    https://kobia.fr/wp-content/uploads/2021/11/00-05-05-Sensitivity-Specificity-GIF-calcul-ROC.mp4?_=1
  - AUC sur courbe ROC : modèle parfait vs. modèle aéatoire
  - Métriques multi-classes
    https://kobia.fr/classification-metrics-multi-class-simple/



Les métriques

- Regression
  - Erreur : MSE, MASE (absolute), MPSE (Percentage), MLSE (Logarithmique)
  - Coefficient de détermination (R^2)

Métrique	Avantages	Inconvénients	Exemple
MSE/RMSE	Accentue les fortes erreurs, régulière, optimisable	Sensible aux outliers	On accepte 5 erreurs de 1°C plus qu'une seule erreur de 5°C
MAE	Homogène, interprétable	Sensible aux outliers	5 erreurs de 1°C sont équivalentes à une seule erreur de 5°C
MAPE	Robuste aux changements d'échelle, interprétable	Sensible aux faibles valeurs, utilisable sur des valeurs non nulles uniquement	Une erreur de 10% sur une réalité de 10€ (9 ou 11€) est équivalente à une erreur de 10% sur une réalité de 100€ (90 ou 110€)
MSLE	Robuste aux changements d'échelle, régulière, optimisable	Peu interprétable, utilisable sur des valeurs positives, non symétrique	On préfère surestimer de 10% que de sous estimer de 10% peu importe la valeur de la réalité.

Figure 3 – (c)
https://kobia.fr/regression-metrics-quelle-metrique-choisir/

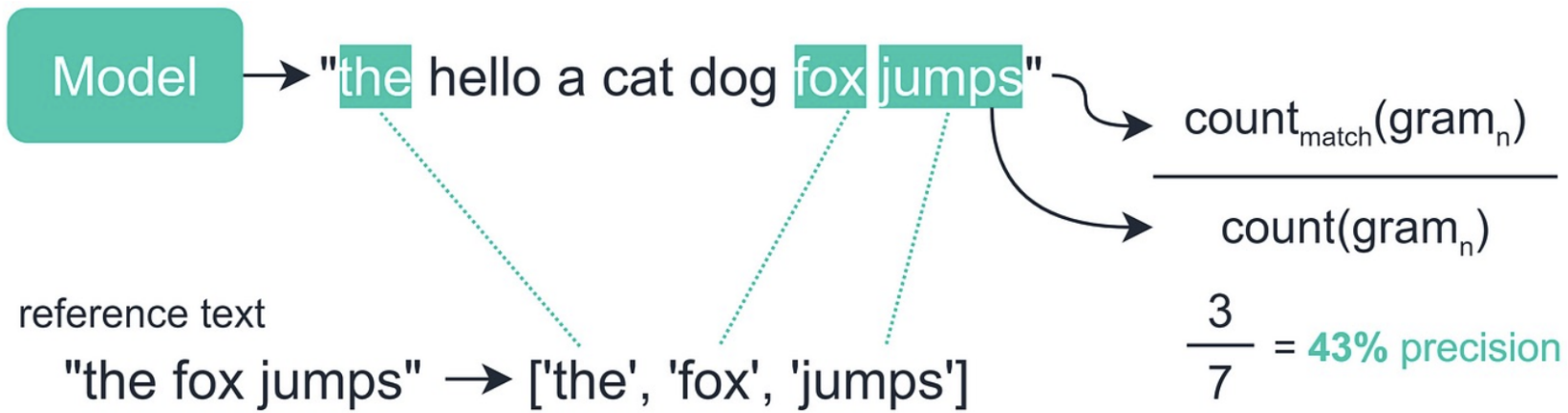
Les métriques

- Ordonnancement
  - Precision, Rappel, F-mesure
  - MRR
  - NDCG

- Génération
  - Métriques automatique de correspondance de termes : BLEU, ROUGE, METEOR, ...

➜ Défi: **évaluer** ces systèmes pour les faire progresser

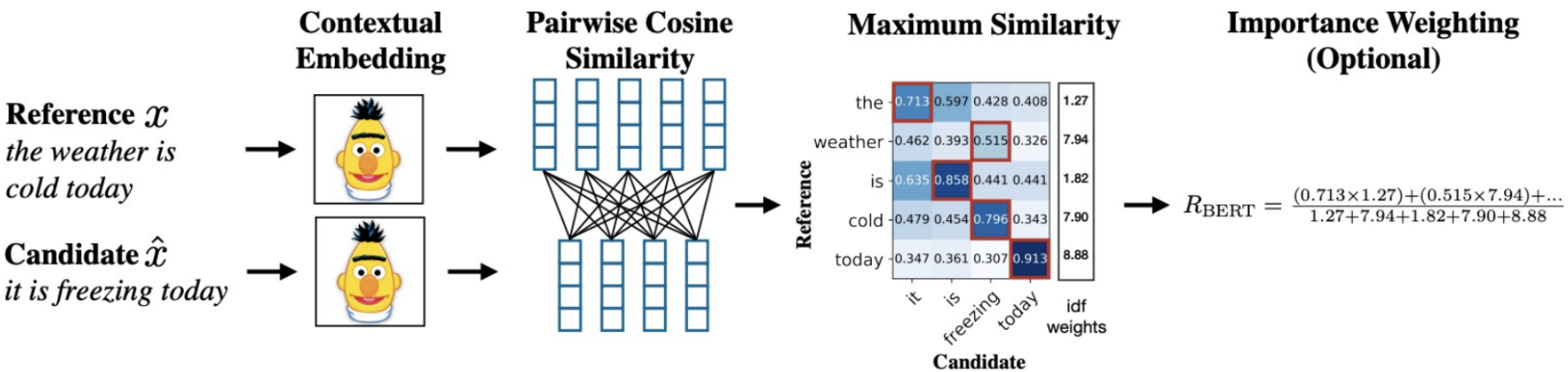
1. Obtenir une vérité terrain
2. Mesurer un écart entre proposition & vérité terrain



Evaluer quantitativement les textes générés est très difficile

- Génération
  - Métriques automatique basées sur la sémantique : BertScore, QuestEval, ...

➢ Comparaison sémantique : BertScore
→ Evaluer le sens des phrases/des mots



- Génération

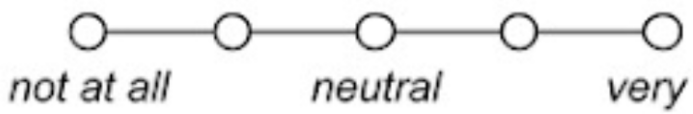
➢ Evaluation humaine : critique pour la génération

➜ Critères :
- ➜ Variabilité des styles, du vocabulaire
- ➜ Fluidité du texte
- ➜ Couverture
- ➜ ...

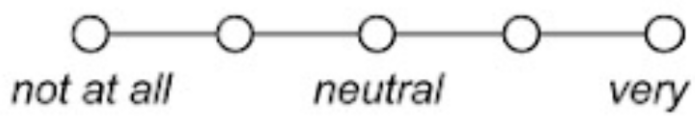
➜ Evaluation subjective
- ➜ Plusieurs annotateurs
- ➜ Accord inter-annotateurs

Input: Bud Powell était un pianiste de légende.
Reference: Bud Powell was a legendary pianist.
Candidate: Bud Powell was a great pianist.

How fluent is the sentence?



Does it accurately convey the meaning of the reference?



Les protocoles d'entraînement et d'évaluation

- Partition du jeu de données en train/val/test
- Cross-validation

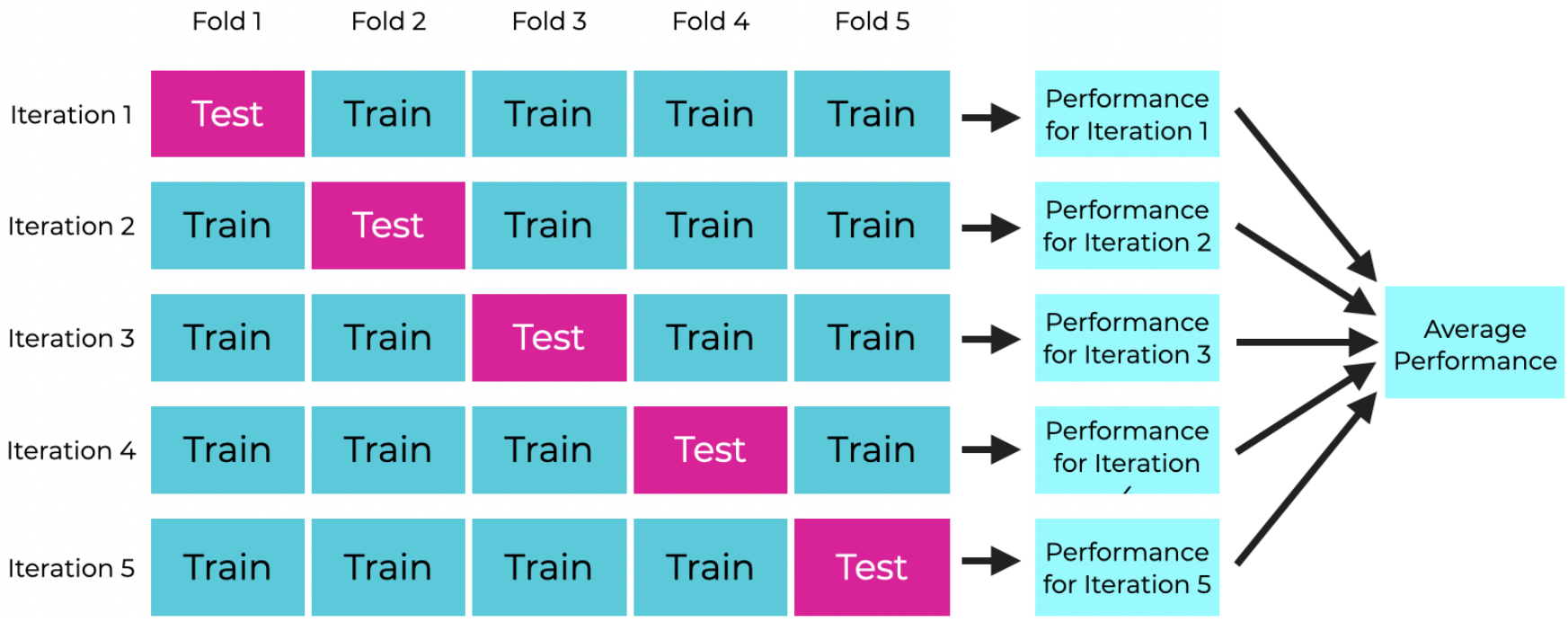


Figure 4 – (c)
https://www.sharpsightlabs.com/blog/cross-validation-explained/

Stratégies d'entraînement

- Entraînement de tout le modèle
- Décomposition par composants/entraînements séquentiels
- Sanity check / sur-apprentissage

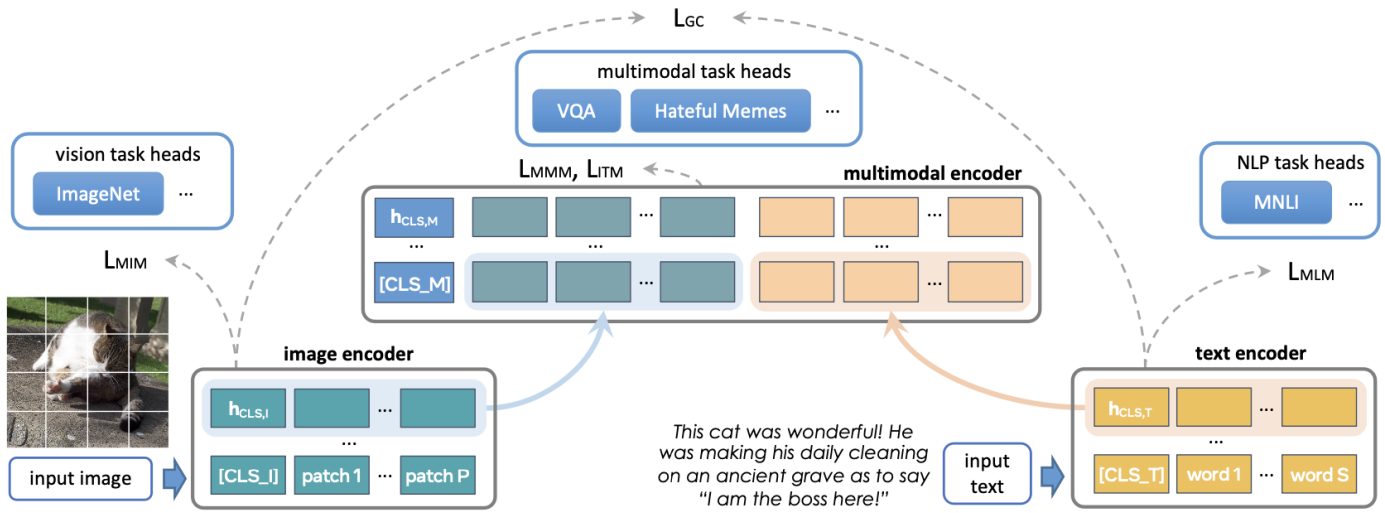


Figure 5 – (c) https://flava-model.github.io/

- Lancer plusieurs seeds/variances

Les analyses

- Comparaison de modèles
  - Moyenne sur les instances
  - Variance et tests de significativité (student si TCL, sinon Wilcoxon)
  - Variance sur les différentes seeds
- Ablation pour valider les composants du modèle
- Interprétabilité : identifications de variables pertinentes
- Analyses qualitatives : matrice de confusion, exemples illustratifs, visualisation, analyse de facteurs (ANOVA, p-value des facteurs/variables, R^2, ...) ...
- **Ne pas avoir peur d'aller voir les données brutes, les prédictions des modèles, etc...**

Analyses : exemples

			Image to Textual recipe				Textual recipe to Image			
			MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
1k items	SOTA	Random	499	0.0	0.0	0.0	499	0.0	0.0	0.0
		CCA [34]	15.7	14.0	32.0	43.0	24.8	9.0	24.0	35.0
		PWC [34]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0
		PWC [34]*	5.0 ± 0.4	22.8 ± 1.4	47.7 ± 1.4	60.1 ± 1.4	5.3 ± 0.4	21.2 ± 1.2	48.0 ± 1.1	60.4 ± 1.4
		PWC++	3.3 ± 0.4	25.8 ± 1.6	54.5 ± 1.3	67.1 ± 1.4	3.5 ± 0.5	24.8 ± 1.1	55.0 ± 1.8	67.1 ± 1.2
	Model scenarios	AdaMine_sem	21.1 ± 2.0	8.7 ± 0.7	25.5 ± 0.9	36.5 ± 0.9	21.1 ± 1.9	8.2 ± 0.9	25.5 ± 1.0	36.2 ± 0.9
		AdaMine_ins	1.5 ± 0.5	37.5 ± 1.1	67.0 ± 1.3	76.8 ± 1.5	1.6 ± 0.5	36.1 ± 1.6	66.6 ± 1.3	76.8 ± 1.5
		AdaMine_ins+cls	1.1 ± 0.3	38.3 ± 1.6	67.5 ± 1.2	78.0 ± 0.9	1.2 ± 0.4	37.5 ± 1.4	67.7 ± 1.2	77.3 ± 1.0
		AdaMine_avg	2.3 ± 0.5	30.6 ± 1.1	60.3 ± 1.2	71.4 ± 1.3	2.2 ± 0.3	30.6 ± 1.8	60.6 ± 1.1	71.9 ± 1.1
		AdaMine_ingr	4.9 ± 0.5	22.6 ± 1.4	48.5 ± 1.6	59.8 ± 1.3	5.0 ± 0.6	21.5 ± 1.4	47.7 ± 2.1	59.8 ± 1.8
		AdaMine_instr	3.9 ± 0.5	24.4 ± 1.6	52.6 ± 2.0	65.4 ± 1.6	3.7 ± 0.5	23.6 ± 1.7	52.7 ± 1.6	65.5 ± 1.5
		AdaMine	**1.0 ± 0.1**	**39.8 ± 1.8**	**69.0 ± 1.8**	**77.4 ± 1.1**	**1.0 ± 0.1**	**40.2 ± 1.6**	**68.1 ± 1.2**	**78.7 ± 1.3**
10k items		PWC++ (best SOTA)	34.6 ± 1.0	7.6 ± 0.2	19.8 ± 0.1	30.3 ± 0.4	35.0 ± 0.9	6.8 ± 0.2	21.5 ± 0.2	28.8 ± 0.3
	Model scenarios	AdaMine_sem	207.3 ± 3.9	1.4 ± 0.3	5.7 ± 0.3	9.6 ± 0.3	205.4 ± 3.2	1.4 ± 0.1	5.4 ± 0.2	9.1 ± 0.4
		AdaMine_ins	15.4 ± 0.5	13.3 ± 0.2	32.1 ± 0.7	42.6 ± 0.8	15.8 ± 0.7	12.3 ± 0.3	31.1 ± 0.5	41.7 ± 0.6
		AdaMine_ins+cls	14.8 ± 0.4	13.6 ± 0.2	32.7 ± 0.4	43.2 ± 0.3	15.2 ± 0.4	12.9 ± 0.3	31.8 ± 0.3	42.5 ± 0.2
		AdaMine_avg	24.6 ± 0.8	10.0 ± 0.2	25.9 ± 0.4	35.7 ± 0.5	24.0 ± 0.6	9.2 ± 0.4	25.4 ± 0.5	35.3 ± 0.4
		AdaMine_ingr	52.8 ± 1.2	6.5 ± 0.2	17.9 ± 0.2	25.8 ± 0.3	53.8 ± 0.7	5.8 ± 0.3	17.3 ± 0.2	25.0 ± 0.2
		AdaMine_instr	39.0 ± 0.9	6.4 ± 0.1	18.9 ± 0.4	27.6 ± 0.5	39.2 ± 0.7	5.7 ± 0.4	17.9 ± 0.6	26.6 ± 0.5
		AdaMine	**13.2 ± 0.4**	**14.9 ± 0.3**	**35.3 ± 0.2**	**45.2 ± 0.2**	**12.2 ± 0.4**	**14.8 ± 0.3**	**34.6 ± 0.3**	**46.1 ± 0.3**

Table 3: State-of-the-art comparison. MedR means Median Rank (lower is better). R@K means Recall at K (between 0% and 100%, higher is better). The mean and std values over 10 (resp. 5) bags of 1k (resp. 10k) pairs each are reported for the top (resp. bottom) table. Items marked with a star (*) are our reimplementation of the cited methods.

Analyses : exemples

Model		MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK	AVG
(Kiros et al., 2015)†	$\mathbf{T}_{1024}$	72.7*	75.2*	90.6*	84.7*	71.8* /79.2*	76.2*	68.8*	79.3*	77.4
(Kiela et al., 2018)†	GS-Cap	72.0*	76.8*	90.7*	85.5*	72.9/80.6	76.7*	73.7	**82.9**	78.4
(Kiela et al., 2018)†	GS-Img	74.5*	79.3*	90.8*	87.8*	73.0/80.3	80.0*	72.2*	80.9*	79.8
(Kiela et al., 2018)†	GS-Both	72.5*	75.7*	90.7*	85.4*	72.9/81.3	76.7*	72.2*	81.4*	78.4
(Kiros et al., 2015)†	$\mathbf{T}$	75.9*	79.2*	92.0	86.7*	72.2/80.2	81.8*	72.0*	81.1*	80.1
(Lazaridou et al., 2015a)‡	$\mathbf{T}+\mathbf{CM}$	77.6	81.4	92.6	88.3	73.5/81.1	82.0*	73.0	81.4*	81.1
(Collell et al., 2017)‡	**SEQ**	76.1*	79.8*	92.5	86.7*	70.0* /79.5*	81.7*	67.3*	76.7*	78.9
Model scenarios	$\mathbf{T}+\mathbf{P}_{id}$	77.5	81.5	92.7	88.4	**73.7**/81.3	82.4	72.4	81.1	81.2
	$\mathbf{T}+\mathbf{P}_{g}$	**77.8**	**81.8**	**93.0**	88.1	73.3/**81.6**	**83.5**	72.8	82.2	**81.6**
	$\mathbf{T}+\mathbf{C}_{id}$	77.5	81.6	92.8	88.3	72.9/80.5	82.2	73.1	82.3	81.3
	$\mathbf{T}+\mathbf{C}_{g}$	77.3	81.5	92.8	**88.6**	73.6/81.1	82.6	**74.1**	82.6	**81.6**
	$\mathbf{T}+\mathbf{C}_{id}+\mathbf{P}_{id}$	77.3	81.2	**93.0**	88.4	73.0/80.6	82.5	73.5	82.1	81.4
	$\mathbf{T}+\mathbf{C}_{g}+\mathbf{P}_{g}$	77.4	81.5	**93.0**	88.1	73.2/80.9	82.7	73.9	**82.9**	**81.6**

Table 3: Extrinsic evaluations with SentEval. All models give sentences in dimension $d_t = 2048$ (except $\mathbf{T}_{1024}$). 'AVG' stands for the average accuracies reported in the other columns. '†': the model has been re-implemented (we obtained higher scores than the one given in the original papers). '‡': the baseline is an adaptation of the model to the case of sentences. '$*$': significantly differs from the best scenario among our models.