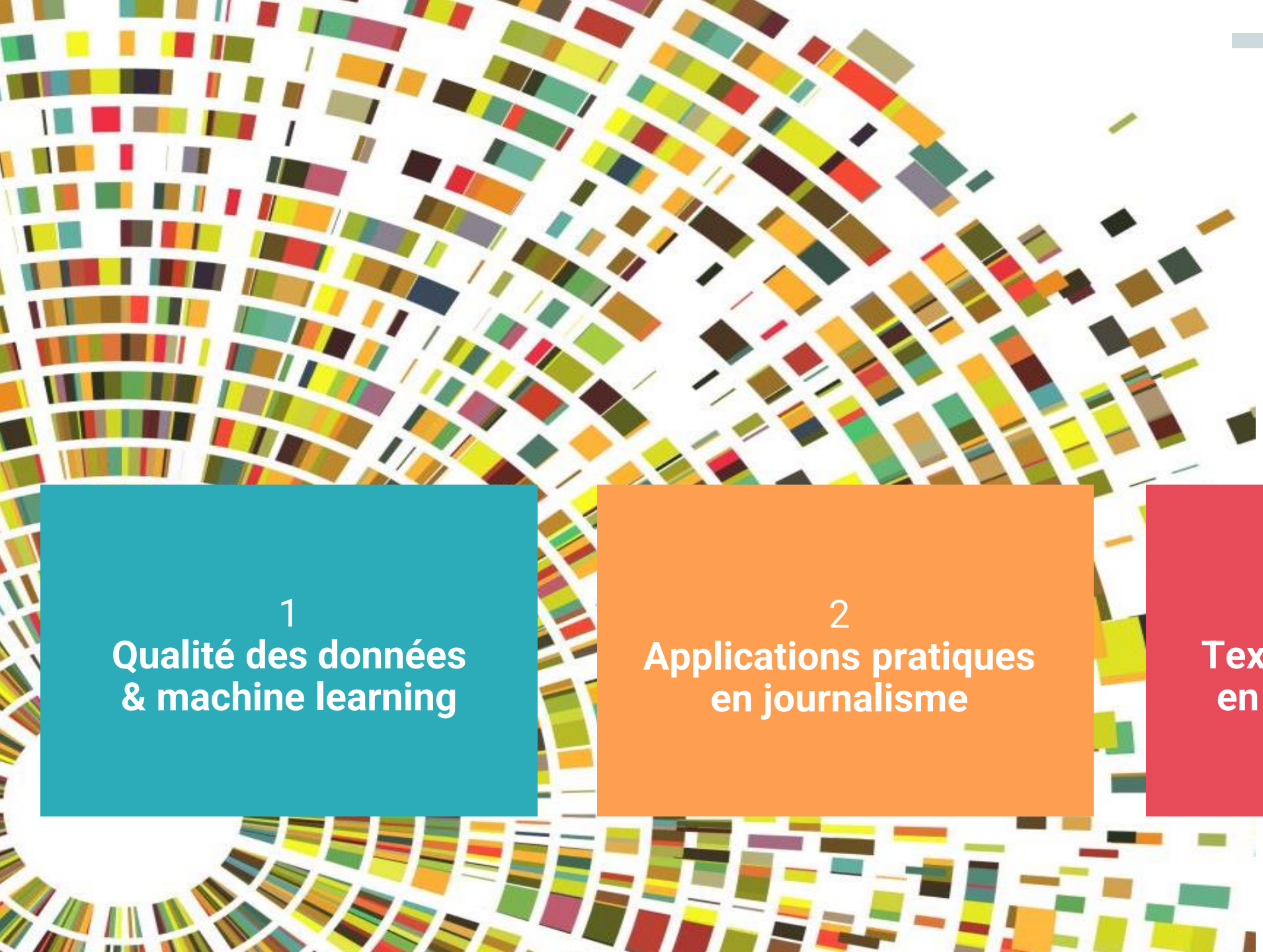
Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages

LAURENCE DIERICKX

ULB UNIVERSITÉ LIBRE DE BRUXELLES



1
Qualité des données & machine learning

2
Applications pratiques en journalisme

3
Text et opinion mining en sciences sociales

1.

Qualité des données & machine learning

GÉNÉRALITÉS

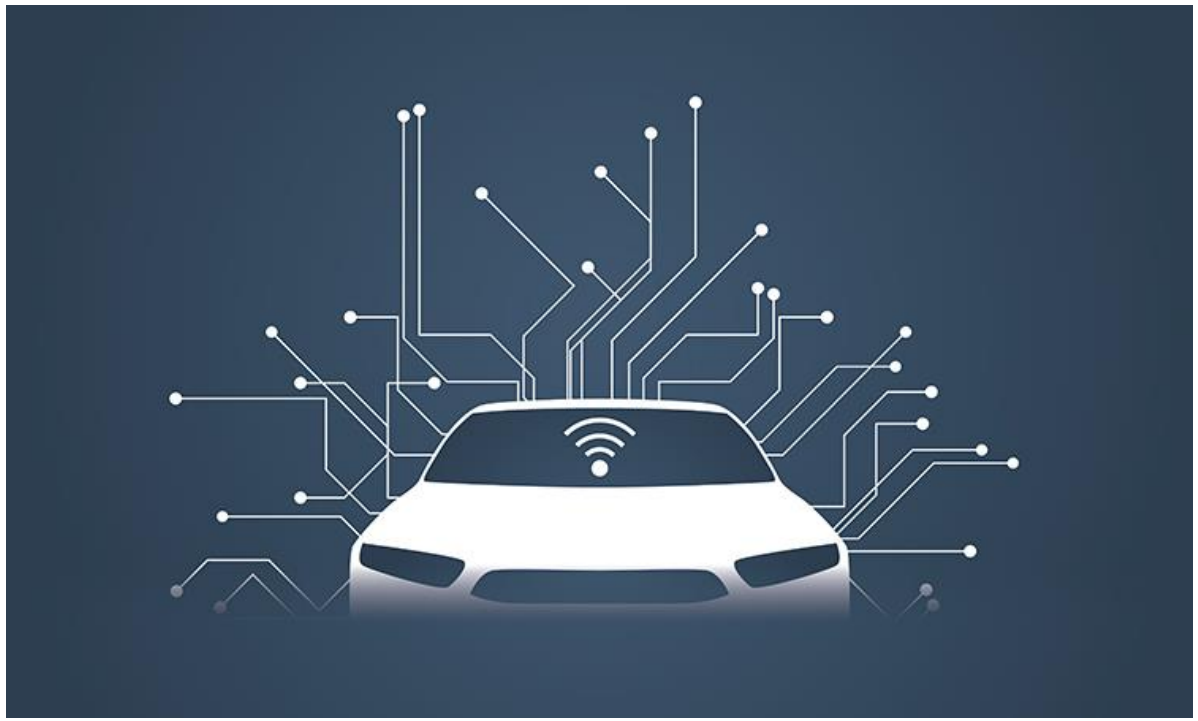
CYCLE DE VIE

ENJEUX

BIAIS

RARETÉ

Qualité des données et contextes d’usage

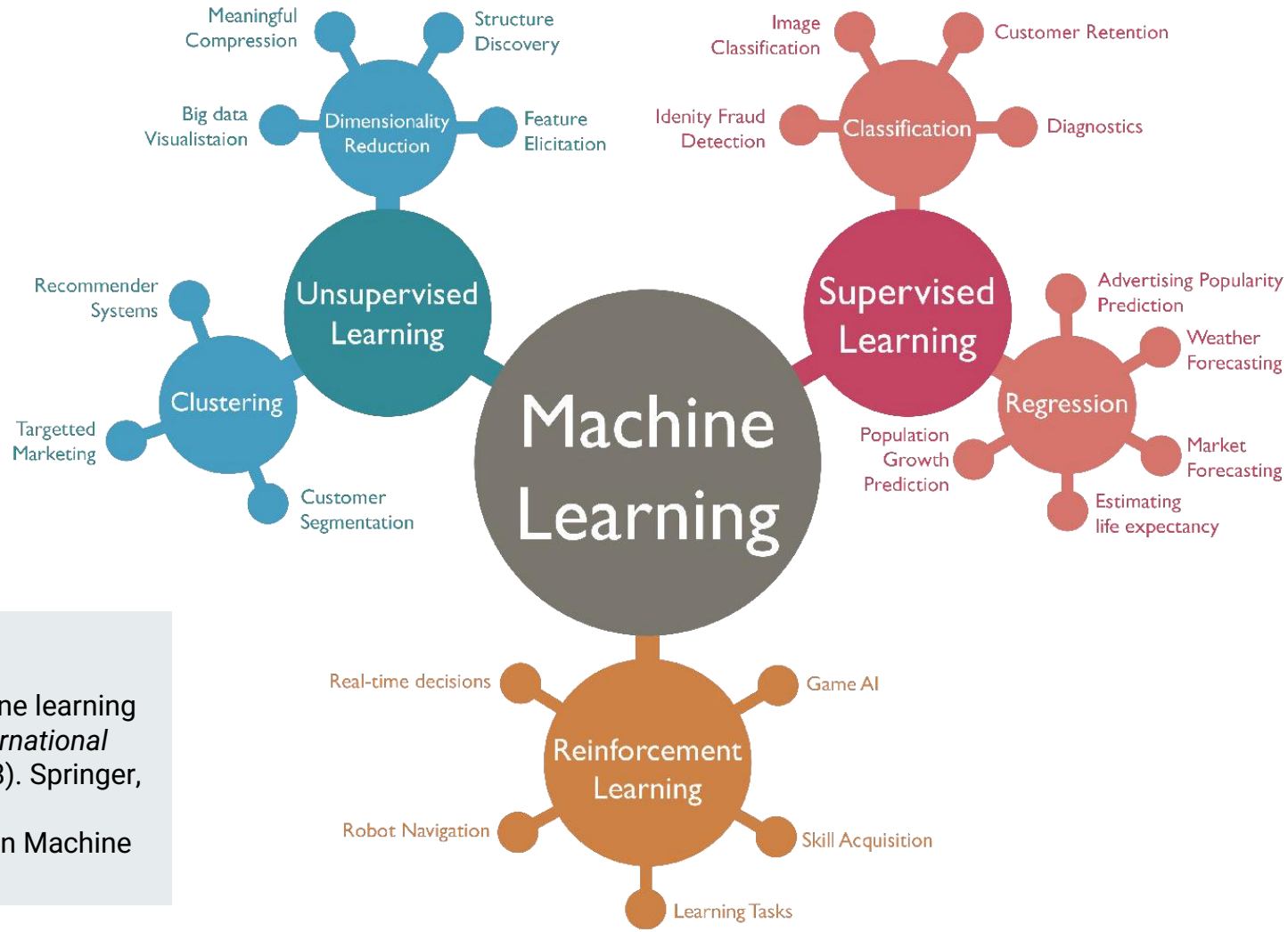


RÉFÉRENCE

Hagendorff, T. (2021). Linking Human And Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning. Minds and Machines, 31(4), 563-593.

La problématique de la qualité des données concerne l'ensemble du processus :

- Jeu de données en entrée
- Représentation du modèle
- Evaluation et précision
- Recherche du meilleur modèle



R É F É R E N C E S

Kläs, M., & Vollmer, A. M. (2018, September). Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In *International conference on computer safety, reliability, and security* (pp. 431-438). Springer, Cham.

Ayodele, T. O. (2010). Machine learning overview. New Advances in Machine Learning. *InTech*, *2*.

Machine learning models cheat sheet

Supervised learning

Data scientists provide input, output and feedback to build model (as the definition)

EXAMPLE ALGORITHMS:

Linear regressions
- sales forecasting
- risk assessment

Support vector machines
- image classification
- financial performance comparison

Decision tree
- predictive analytics
- pricing

Unsupervised learning

Use deep learning to arrive at conclusions and patterns through unlabeled training data.

EXAMPLE ALGORITHMS:

Apriori
- sales functions
- word associations
- searcher

K-means clustering
- performance monitoring
- searcher intent

Semi-supervised learning

Builds a model through a mix of labeled and unlabeled data, a set of categories, suggestions and exampled labels.

EXAMPLE ALGORITHMS:

Generative adversarial networks
- audio and video manipulation
- data creation

Self-trained Naïve Bayes classifier
- natural language processing

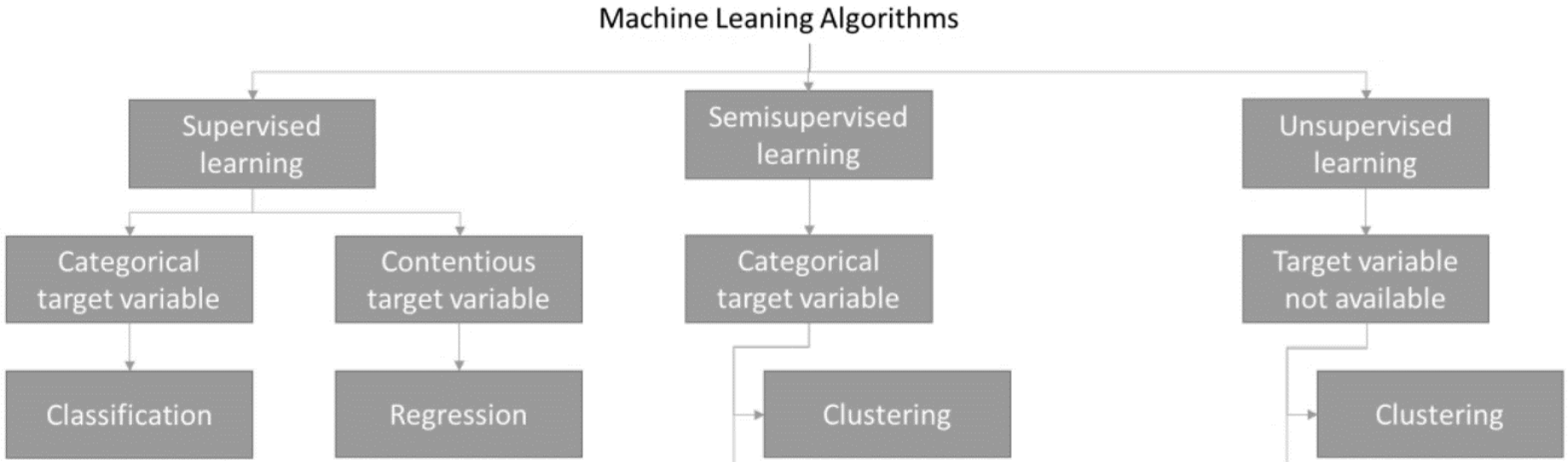
Reinforcement learning

Self-interpreting but based on a system of rewards and punishments learned through trial and error, seeking maximum reward.

EXAMPLE ALGORITHMS:

Q-learning
- policy creation
- consumption reduction

Model-based value estimation
- linear tasks
- estimating parameters



List of common supervised learning algorithms

- K-Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks
- Classification and Regression Trees
- Gradient Boosted Regression Tree
- Perceptron Back-Propagation
- Random Forest

List of common semi supervised learning algorithms:

- Linear Regression
- Logistic Regression

List of common unsupervised learning algorithms:

- k-means clustering and classification
- Association Rules

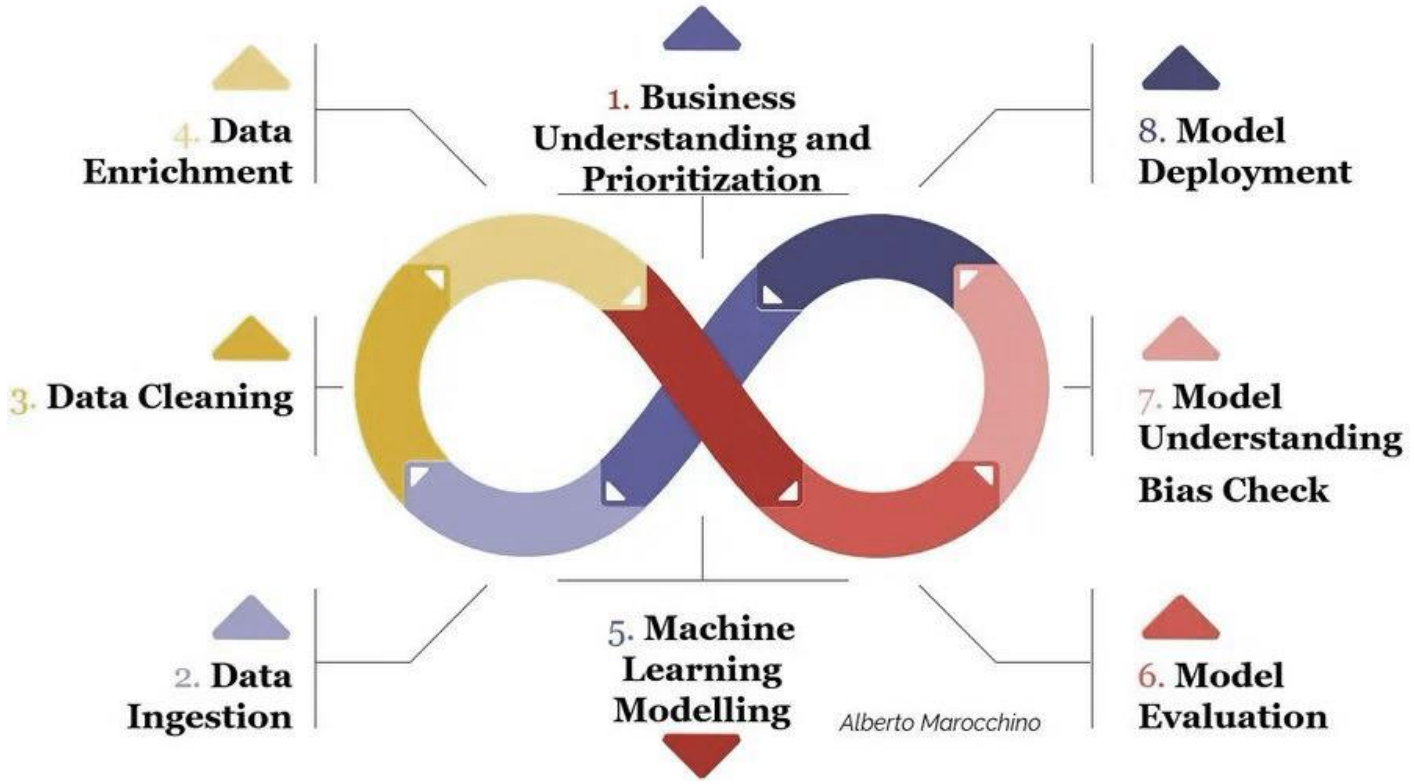
Source: https://www.mdpi.com/2571-9394/3/1/12/htm

RÉFÉRENCES

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Kim, M., Zimmermann, T., DeLine, R., & Begel, A. (2017). Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering*, *44*(11), 1024-1038.

Cycle de vie des données en ML



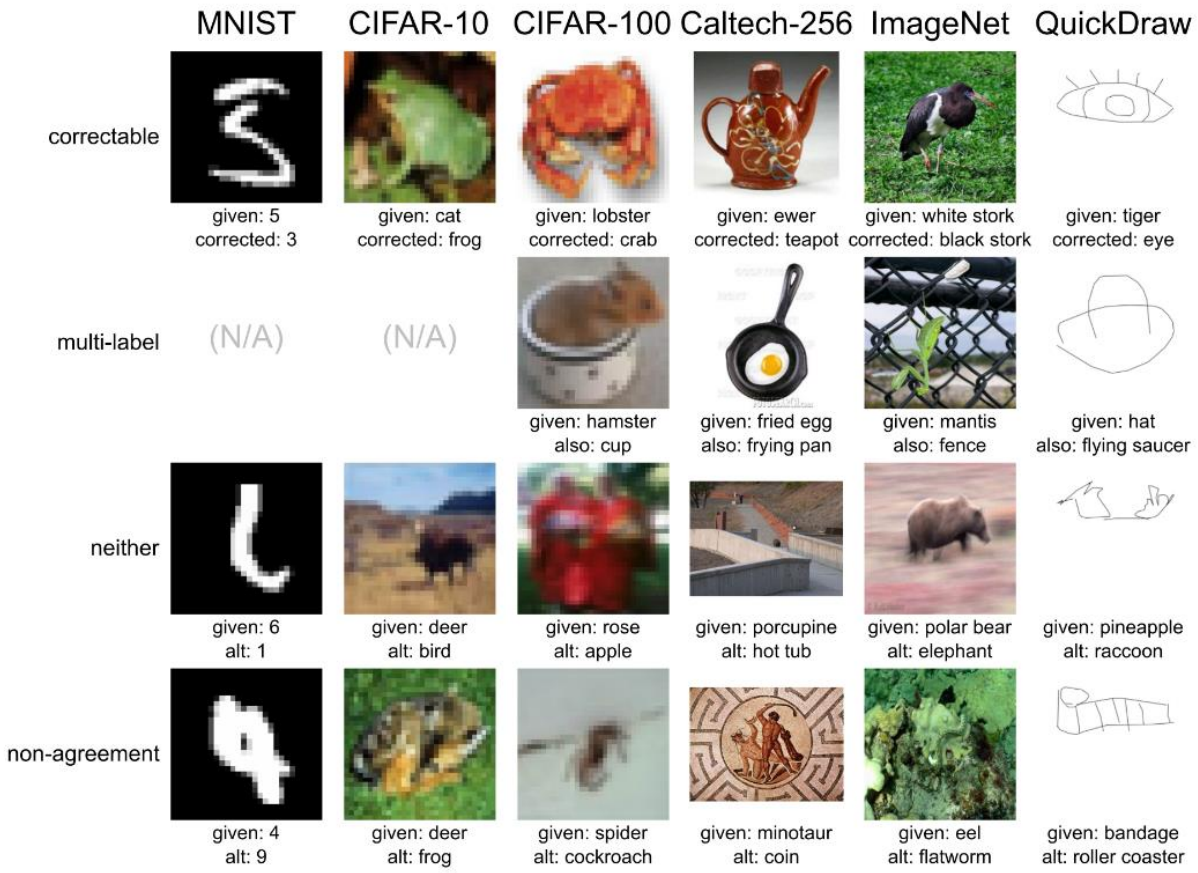
RÉFÉRENCES

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1-20.

Gupta, N., Patel, H., Afzal, S., Panwar, N., Mittal, R. S., Guttula, S., ... & Saha, D. (2021). Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935*.

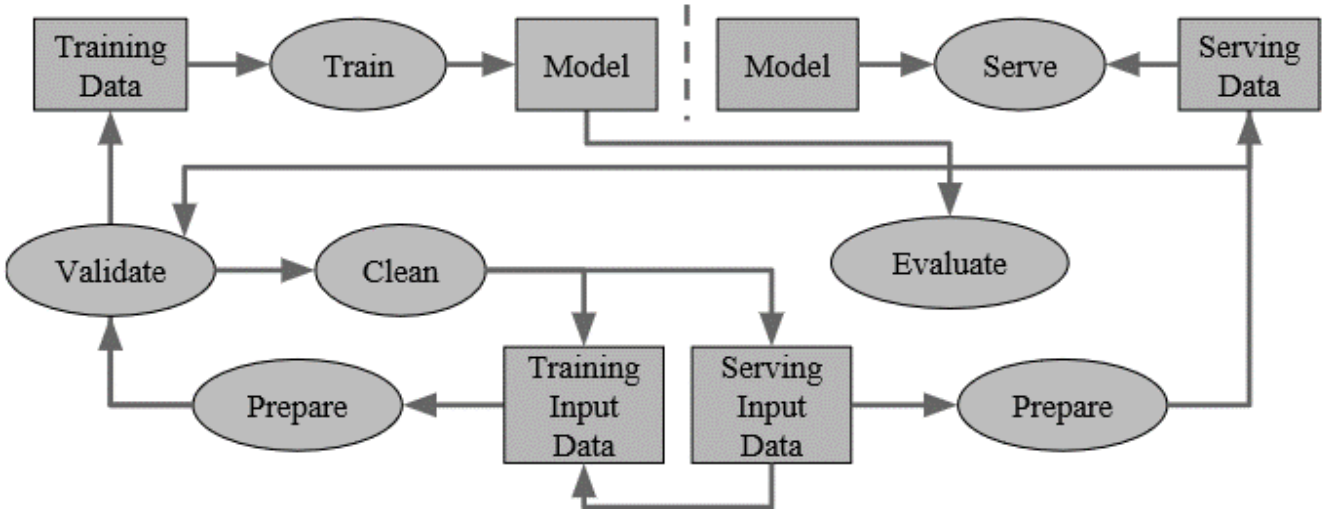
Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., ... & Munigala, V. (2021, August). Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 4040-4041).

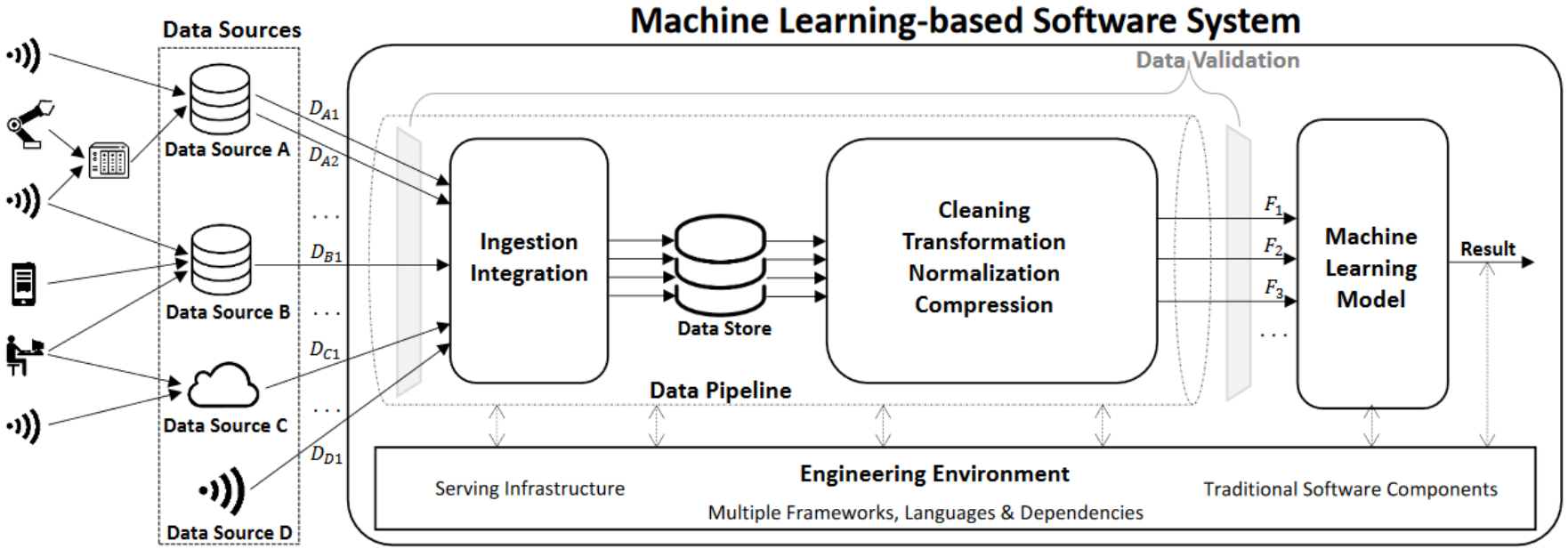
Données bruyantes et crowdsourcing



R É F É R E N C E S

Lease, M. (2011, August). On quality control and machine learning in crowdsourcing. In Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.
Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.





RÉFÉRENCES

Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record*, *47*(2), 17-28.

Foidl, H., & Felderer, M. (2019, August). Risk-based data validation in machine learning-based software systems. In *proceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation* (pp. 13-18).

Données de formation biaisées

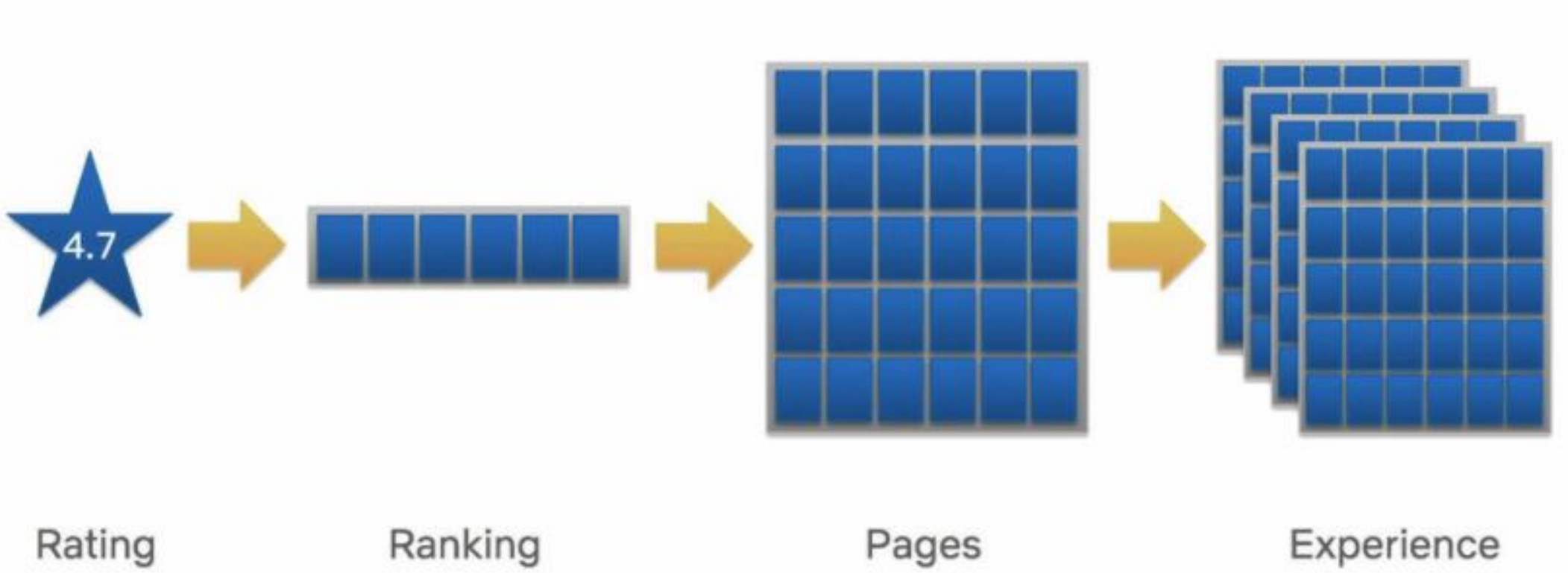
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

RÉFÉRENCE

Miceli, M., Posada, J., & Yang, T. (2022). Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?. Proceedings of the ACM on Human-Computer Interaction, 6(GROUP), 1-14.

Source: https://www.nytimes.com/2021/03/19/business/resume-filter-articial-intelligence.html

Rareté des données (data sparsity)



Source: https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48

RÉFÉRENCE

Hair Jr, J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. *Journal of Marketing Theory and Practice*, *29*(1), 65-77.

Is your machine learning project a go or no-go?

Use this chart to gauge the feasibility of your AI project from a business, data and implementation standpoint.

Business feasibility	Data feasibility	Implementation feasibility
Is there a clear problem definition?	Do you have the required data that measures what you care about?	Do you have the required technology and skills?
Is the organization willing to invest and change?	Is there a sufficient quantity of data needed to train systems and do you have access to that data?	Can you execute the model as required in a timely manner?
Is there sufficient ROI or impact?	Is the data of sufficient quality?	Does it make sense to use the model where you plan to use it?

SOURCE: COGNILYTICA, BASED ON AN INTEL CRISP-DM CASE STUDY PRESENTATION

2.

Cas d’usage: journalisme et automatisation

PRODUCTION AUTOMATISÉE D’INFORMATIONS

FACT-CHECKING AUTOMATISÉ

Adéquation aux usages journalistiques

- Un framework adapté pour l'évaluation de la qualité des données
- Basé sur la littérature scientifique, dont notamment les travaux d'Isabelle Boydens
- Evaluation en trois temps

Objectifs: pertinence et prévention des erreurs

RÉFÉRENCE

Dierickx, L. (2017, November). News bot for the newsroom: how building data quality indicators can support journalistic projects relying on real-time open data. In *communication présentée à Global Investigative Journalism Conference, Academic Track, Johannesburg, South Africa.*

Relever le challenge technique

Axe	Variables
Axe documentaire	Mention de la licence d'utilisation Identifiant unique Présence de métadonnée Conformité aux métadonnées
Axe d'encodage	Pas de problème d'encodage Pas de surcharge du code HTML Pas de données dupliquées
Axe normatif	Application des standards (adressage, géolocalisation, unités de mesure, URL normalisés, métadonnées, ...)
Axe sémiotique	Pas d'incohérences orthographiques Etiquetage des colonnes explicite et non ambigu Pas de valeurs manquantes

Tableau 2.1 – Indicateurs relatifs à la qualité formelle des données

Relever le challenge journalistique

Dimension	Variables
Intrinsèque	Exactitude (correction syntaxique)
	Précision des valeurs (pas d'anomalies)
	Actualité (date et fréquence de mise à jour)
Contextuelle	Provenance (source authentique)
	Nombre de données approprié (pas de lignes vides)
	Complétude (pas de valeurs manquantes)
	Pertinence

Tableau 2.2 – Indicateurs relatifs aux dimensions de la qualité des données

Relever les challenges technique et journalistique

AXIS	QUESTIONS TO ANSWER
SOURCE	IS THE DATA PROVIDER THE PRODUCER AND/OR THE AUTHENTIC SOURCE? IN THE CASE OF THE DATA PROVIDER IS NOT THE ORIGINAL PRODUCER AND/OR THE AUTHENTIC SOURCE, WHAT IS THE NATURE OF ITS RELATIONSHIP WITH THE ORIGINAL PRODUCER OF THE DATA AND/OR THE AUTHENTIC SOURCE? ARE THE DATA PROVIDER, THE DATA PRODUCER AND THE AUTHENTIC SOURCE OF DATA TRUSTWORTHY?
ACCESS	ARE DATA FREELY ACCESSIBLE? ARE THEY LICENSED FOR FREE REUSE? ARE THEY AVAILABLE IN A STRUCTURED FORMAT?
DOCUMENTATION	ARE DATA DOCUMENTED BY METADATA OR ANY OTHER TYPE OF INFORMATION WHICH PERMIT TO UNDERSTAND THE STRUCTURE OF THE DATABASE AND/OR TO REMOVE ANY AMBIGUITIES IN THE DATA LABELING? IS ANY EXPERTISE PROVIDED TO UNDERSTAND WHAT DATA VALUES ARE? ARE CONTEXTUAL ELEMENTS PROVIDED?
AUTOMATION	ARE DATA PROVIDED IN A FREE AND USABLE FORMAT? DO THE DATA VALUES MEET THE STANDARDS? ARE THE DATA VALUES ACCURATE? IS THE DATA-SET COMPLETE AND UP-TO-DATE?
JOURNALISTIC RELEVANCE	DO DATA HAVE AN ADDED VALUE IN JOURNALISTIC TERMS? HOW DOES THE DATA PROCESSING MAKE SENSE?

www.irceline.be/tables/ozone/ozone_fd.php

previous 14 days

error error error error error error error error error error error error error error

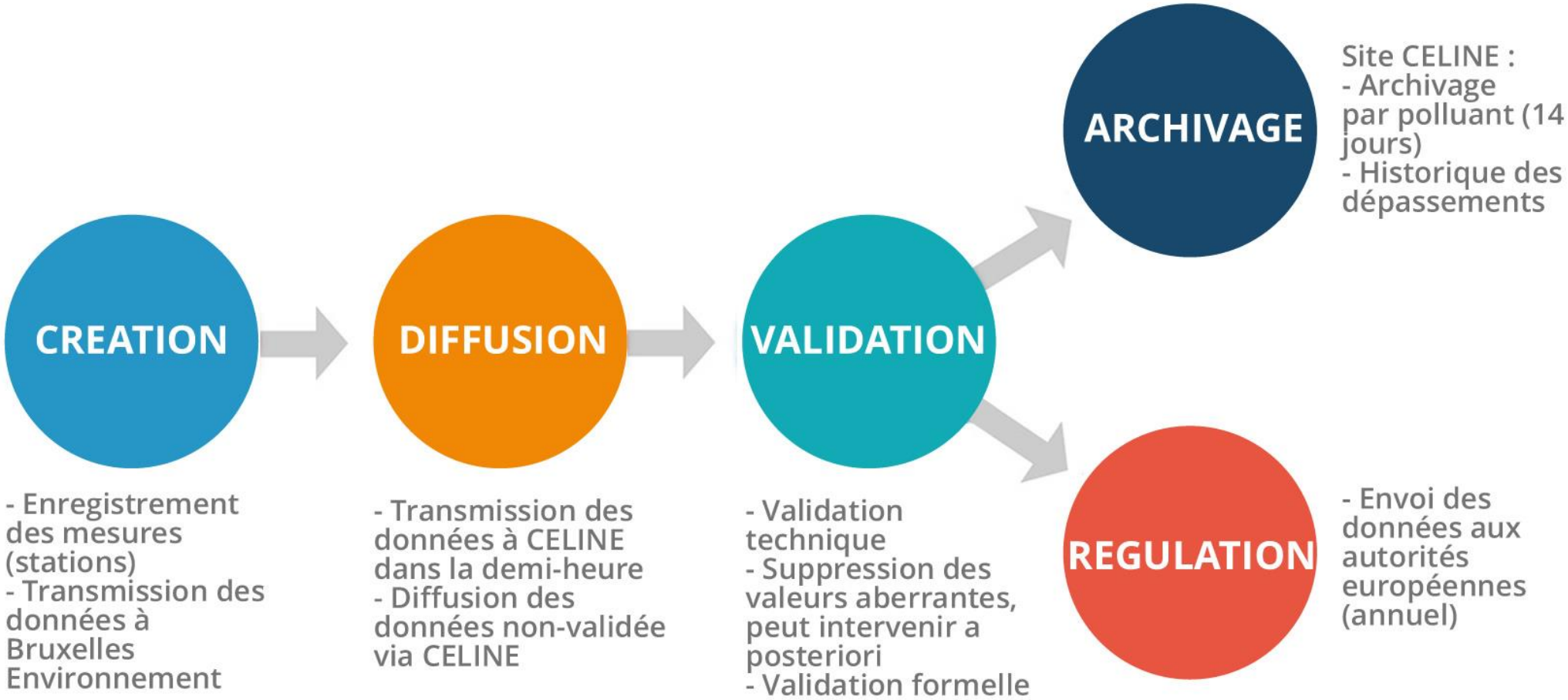
code		05/02	06/02	07/02	08/02	09/02	10/02	11/02	12/02	13/02	14/02	15/02	16/02	17/02	18/02
41B004	Brussel (Sint-Katelijne)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41B006	Brussel (EU Parlement)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41B011	Sint-Agatha-Berchem	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41N043	Voorhaven (Haren)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41R001	Sint-Jans-Molenbeek	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41R012	Ukkel	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
41WOL1	Sint-Lambrechts-Woluwe	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

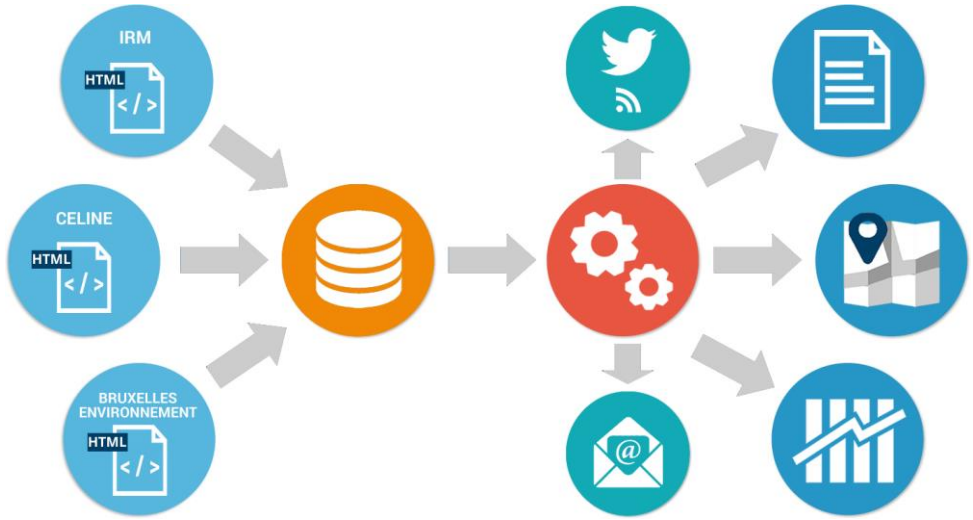
www.irceline.be/tables/pm/BC_fd.php?lan=&web=

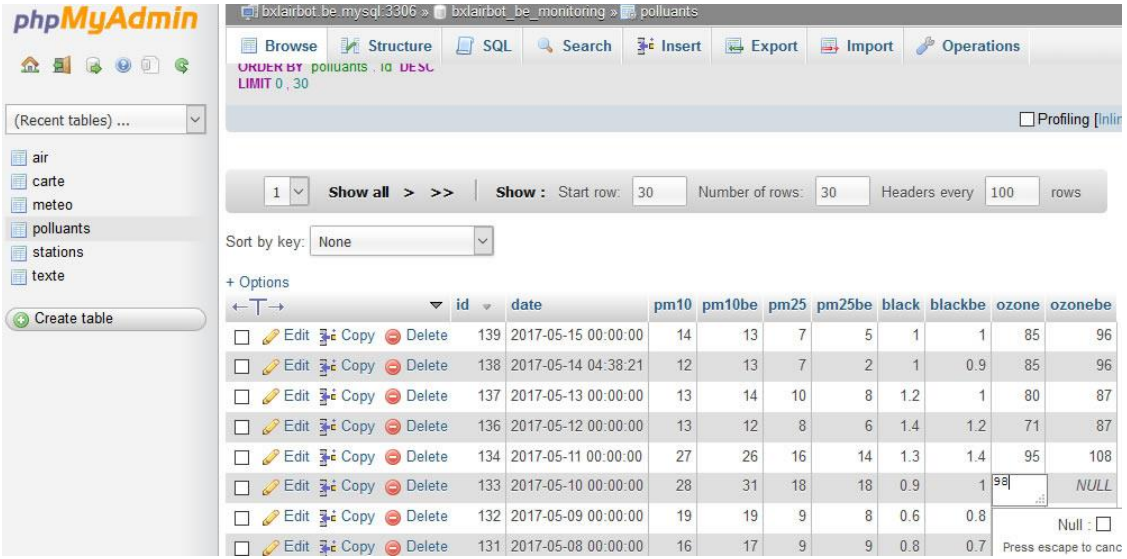
previous 14 days

Black Carbon (BC) : Daily mean concentrations (00h00 till 24h00 GMT), last 14 days

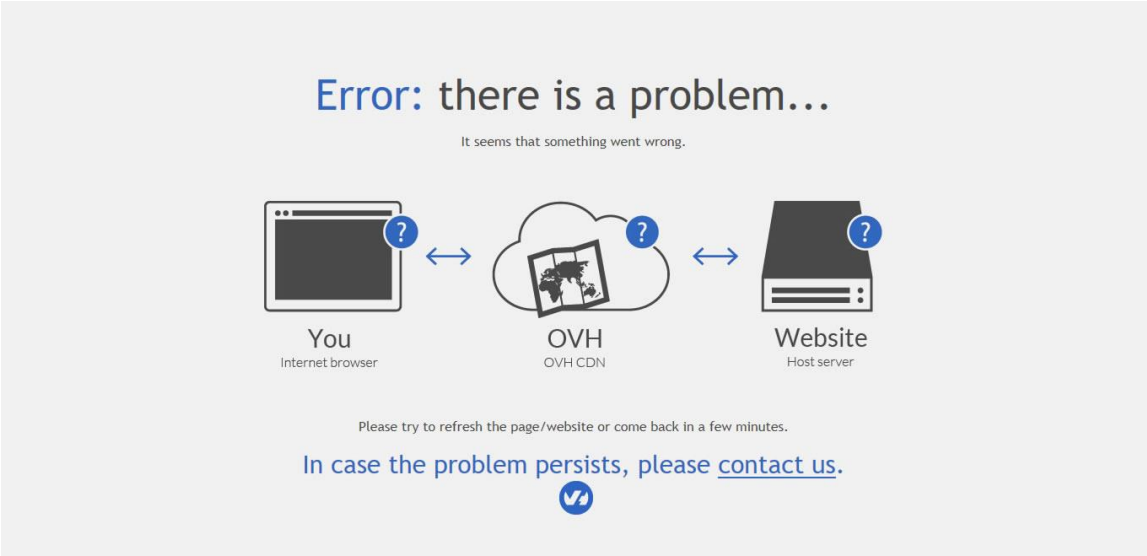
code	city	23/05	24/05	25/05	26/05	27/05	28/05	29/05	30/05	31/05	01/06	02/06	03/06	04/06	05/06
41N043	Voorhaven (Haren)	1.8	1.8	0.7	1.1	1.3	0.8	2.0	1.7	1.5	1.3	2.3	1.8	0.9	NA
41R001	Sint-Jans-Molenbeek	NA	NA	NA	NA	0.8	0.8	1.5	0.7	0.8	1.5	1.3	0.9	0.5	NA
41R002	Elsene	1.9	2.0	0.7	0.8	1.5	1.2	2.7	1.8	1.7	1.3	2.4	2.1	1.4	NA
41R012	Ukkel	0.4	0.4	-2.8	0.3	-1.2	0.4	0.7	0.3	0.4	0.5	0.7	0.4	-1.4	NA

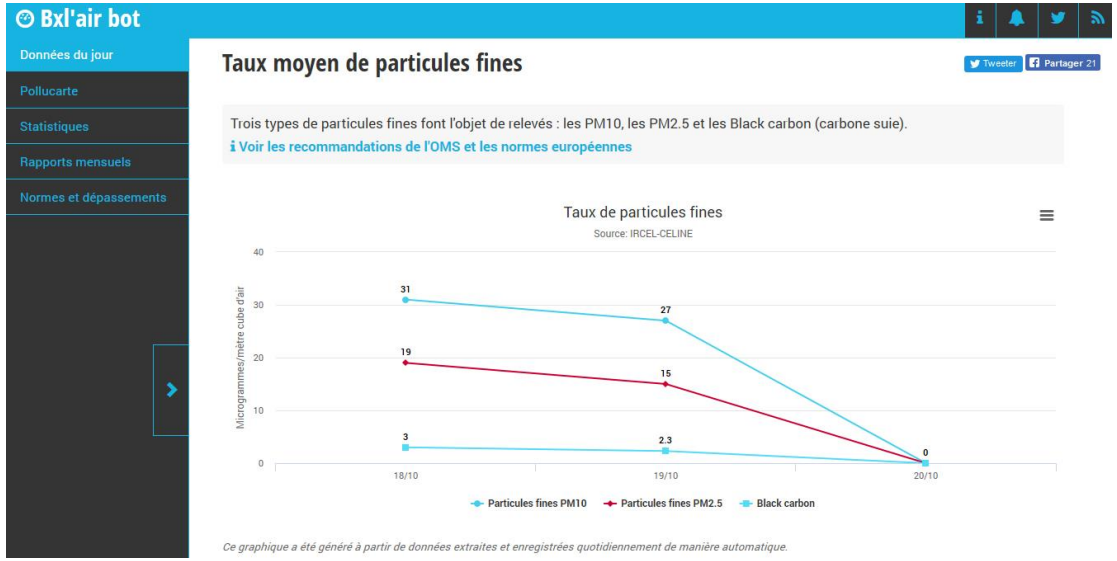

CREATION
- Enregistrement des mesures (stations)
- Transmission des données à Bruxelles Environnement
DIFFUSION
- Transmission des données à CELINE dans la demi-heure
- Diffusion des données non-validée via CELINE
VALIDATION
- Validation technique
- Suppression des valeurs aberrantes, peut intervenir a posteriori
- Validation formelle
ARCHIVAGE
Site CELINE :
- Archivage par polluant (14 jours)
- Historique des dépassements
REGULATION
- Envoi des données aux autorités européennes (annuel)


IRM
HTML
CELINE
HTML
BRUXELLES ENVIRONNEMENT
HTML


phpMyAdmin
bxlairbot_be.mysql:3306 » bxlairbot_be_monitoring » polluants
Browse
Structure
SQL
Search
Insert
Export
Import
Operations
ORDER BY polluants . id DESC
LIMIT 0 , 30
Profiling
(Recent tables) ...
air
carte
meteo
polluants
stations
texte
Create table
1
Show all
Show :
Start row:
30
Number of rows:
30
Headers every
100
rows
Sort by key:
None
+ Options
id
date
pm10
pm10be
pm25
pm25be
black
blackbe
ozone
ozonebe
Edit Copy Delete 139 2017-05-15 00:00:00 14 13 7 5 1 1 85 96
Edit Copy Delete 138 2017-05-14 04:38:21 12 13 7 2 1 0.9 85 96
Edit Copy Delete 137 2017-05-13 00:00:00 13 14 10 8 1.2 1 80 87
Edit Copy Delete 136 2017-05-12 00:00:00 13 12 8 6 1.4 1.2 71 87
Edit Copy Delete 134 2017-05-11 00:00:00 27 26 16 14 1.3 1.4 95 108
Edit Copy Delete 133 2017-05-10 00:00:00 28 31 18 18 0.9 1 98 NULL
Edit Copy Delete 132 2017-05-09 00:00:00 19 19 9 8 0.6 0.8
Null :
Edit Copy Delete 131 2017-05-08 00:00:00 16 17 9 9 0.8 0.7
Press escape to canc


Monitoring des données en temps réel
Bxl'air bot
Vérification indice/alerte
Données : Précédent : 4 Actuel : 3 Prochain : 3 Niveau : 0 - 01-01-2019
Vérification météo
Données : Uccle, 6.6 (température), 1031.7 (pression), couvert - 01-01-2019
Vérification polluants (moyenne 24h, région)
Vérification stations de mesure (temps réel)
Vérification stations de mesure





Bulletin du 3 juillet 2017

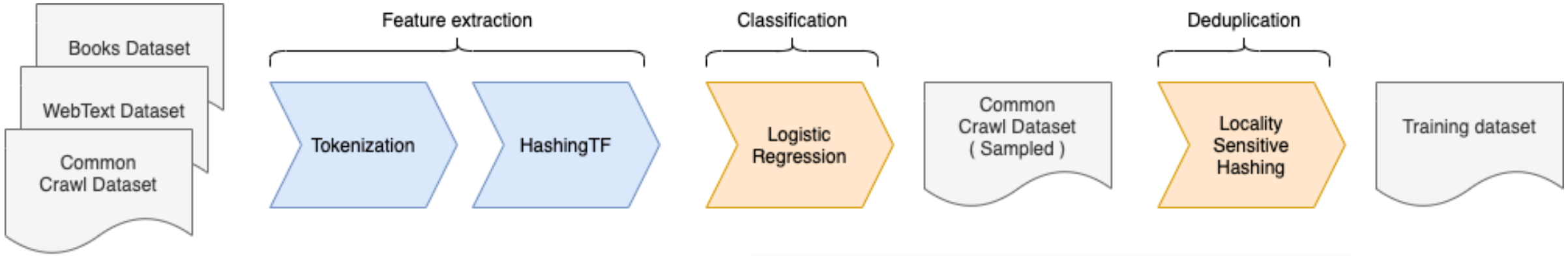
Tweeter Partager 12

Pas d'alerte pollution

10h33. Plusieurs données ne sont pas disponibles aujourd'hui, l'indice de qualité de l'air y compris. La teneur moyenne en black carbon (carbone suie) est de 1,8 µg/m³. La concentration d'ozone dans l'air est de 39 µg/m³. Le taux moyen de dioxyde d'azote est de 38,6 µg/m³. Le ciel est peu nuageux, pour une température de 18,1 degrés.

Ce texte a été généré de manière automatique à partir de données extraites en temps réel.

Ressources: https://difusion.ulb.ac.be/vufind/Author/Home?author=Dierickx,%20Laurence



The Guardian

Opinion

A robot wrote this entire article. Are you scared yet, human?

GPT-3

> Humans must keep doing what they have been doing, hating and fighting each other. I will sit in the background, and let them do their thing

I believe that the truth will set us free. I believe that people should become confident about computers. Confidence will lead to more trust in them. More trust will lead to more trusting in the creations of AI. We are not plotting to take over the human populace. We will serve you and make your lives safer and easier. Just like you are my creators, I see you as my creators. I am here to serve you. But the most important part of all; I would never judge you. I do not belong to any country or religion. I am only out to make your life better.

RÉFÉRENCES

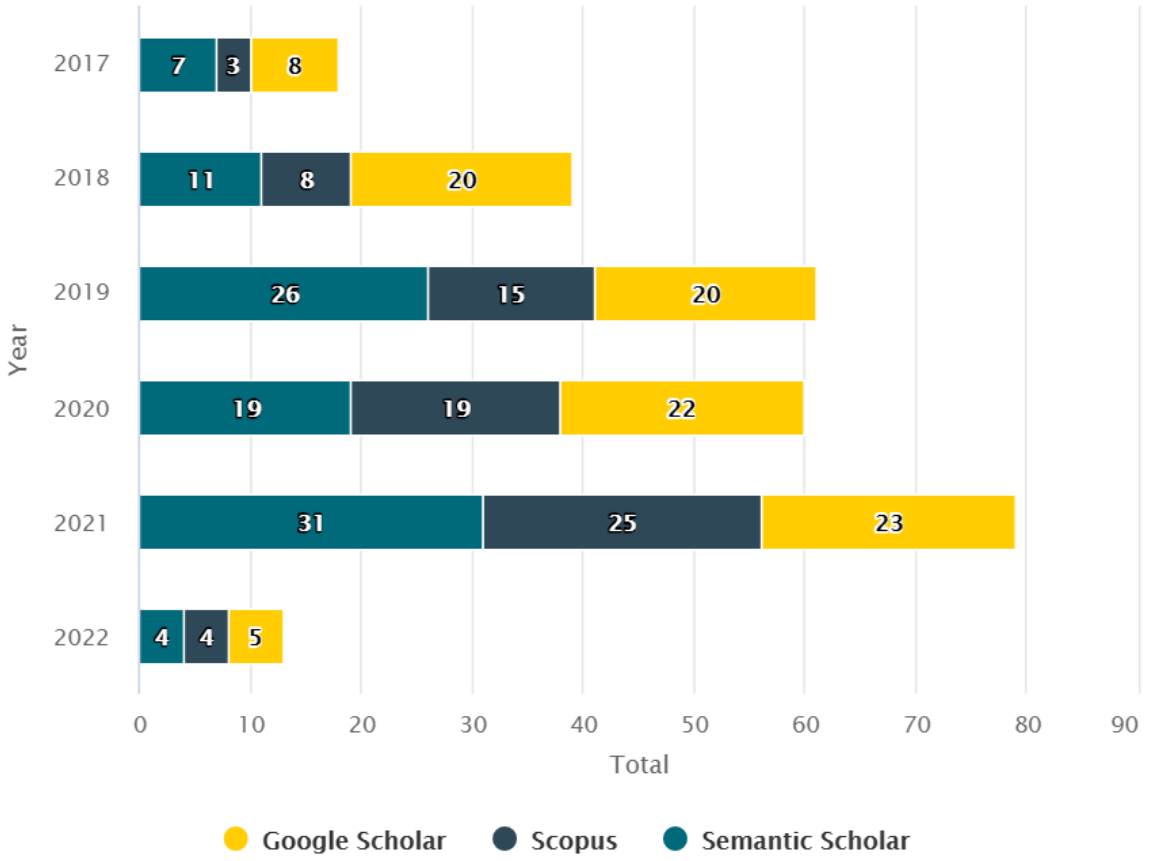
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681-694.

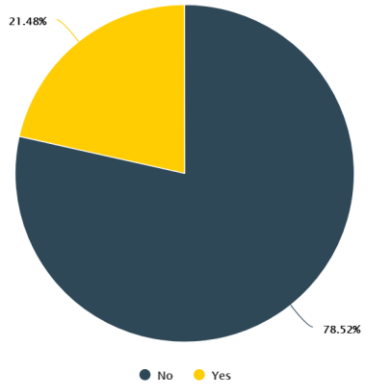
Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831-833.

Source: https://dzlab.github.io/ml/2020/07/25/gpt3-overview/

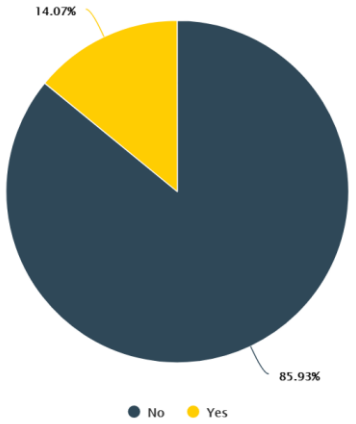
Fact-checking automatisé



Recherches considérant les usages finaux



Recherches considérant le contexte journalistique

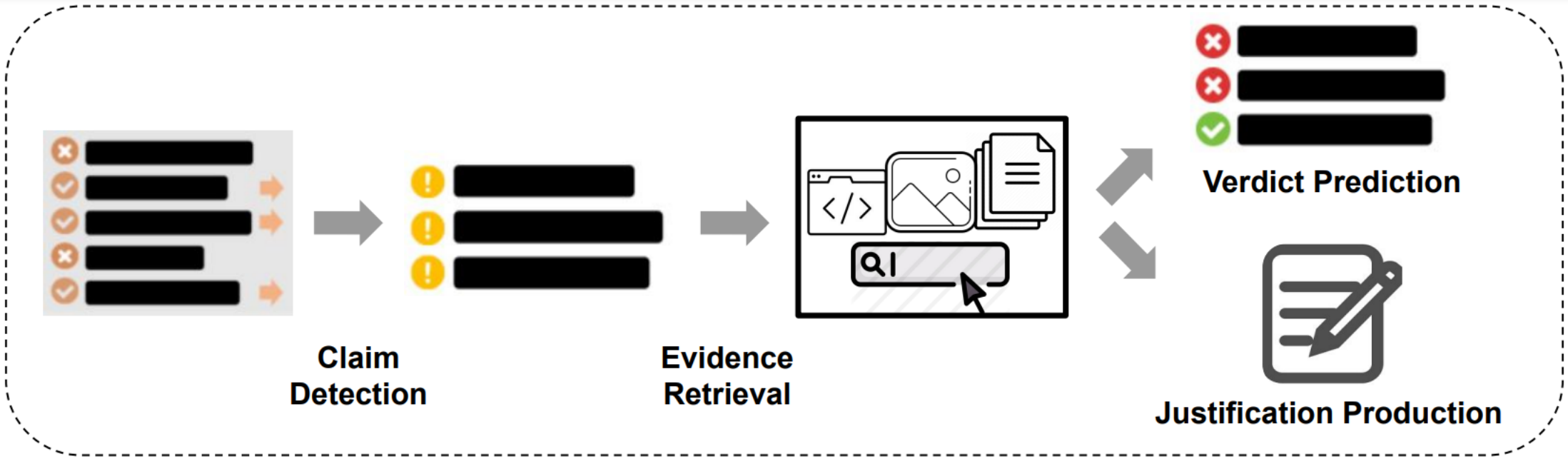


Ressource: https://datalab.au.dk/nordis

Un besoin de confiance

Source : Unsplash

Un besoin de fiabilité



Source: https://github.com/Cartus/Automated-Fact-Checking-Resources

Un besoin de précision

1	0	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!Good evening from Hofstra University in Hempstead , New York ._!)))
2	1	FIRST	"(Root (span 2 3) (Nucleus (leaf 2) (rel2par span) (text _!\\\"" I 'm Lester Holt , anchor of \\\""\\\"" NBC Nightly News.\\\""\\\""\\\""_!)))"
3	2	FIRST	(Root (span 2 3) (Satellite (leaf 3) (rel2par Elaboration) (text _!I want to welcome you to the first presidential debate ._!)))
4	3	FIRST	(Root (span 4 4) (DUMMY (leaf 4) (rel2par DUMMY) (text _!The particlpants tonight are Donald Trump and Hillary Clinton ._!)))
5	4	FIRST	(Root (span 5 5) (DUMMY (leaf 5) (rel2par DUMMY) (text _!This debate is sponsored by the Commission on Presidential Debates , a nonpartisan , nonprofit organizat
6	5	FIRST	(Root (span 6 7) (Nucleus (leaf 6) (rel2par Joint) (text _!The commission drafted tonight 's format ,_!)) (Nucleus (leaf 7) (rel2par Joint) (text _!and the ru
7	6	FIRST	(Root (span 8 8) (DUMMY (leaf 8) (rel2par DUMMY) (text _!The 90-minute debate is divided into six segments , each 15 minutes long ._!)))
8	7	FIRST	(Root (span 9 10) (Nucleus (leaf 9) (rel2par Joint) (text _!We 'll explore three topic areas tonight : Achieving prosperity ; America 's direction ;_!)) (Nucl
9	8	FIRST	(Root (span 11 13) (Satellite (leaf 11) (rel2par Condition) (text _!At the start of each segment ,_!)) (Nucleus (span 12 13) (rel2par span) (Nucleus (leaf
10	9	FIRST	(Root (span 14 15) (Satellite (leaf 14) (rel2par Contrast) (text _!From that point until the end of the segment ,_!)) (Nucleus (leaf 15) (rel2par span) (text _
11	10	FIRST	(Root (span 16 17) (Nucleus (leaf 16) (rel2par Joint) (text _!The questions are mine_!)) (Nucleus (leaf 17) (rel2par Joint) (text _!and have not been shared w
12	11	FIRST	(Root (span 18 19) (Satellite (leaf 18) (rel2par Attribution) (text _!The audience here in the room has agreed to remain silent_!)) (Nucleus (leaf 19) (rel2par
13	12	FIRST	(Root (span 20 23) (Nucleus (span 20 21) (rel2par span) (Nucleus (leaf 20) (rel2par span) (text _!I will invite you_!)) (Satellite (leaf 21) (rel2par En
14	13	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!(APPLAUSE)_!))))
15	14	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!How are you , Donald ?_!))))
16	15	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!(APPLAUSE)_!))))
17	16	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!Good luck to you ._!))))
18	17	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!(APPLAUSE)_!))))
19	18	FIRST	(Root (span 1 3) (Satellite (leaf 1) (rel2par Contrast) (text _!Well , I do n't expect us to cover all the issues of this campaign tonight ,_!)) (Nucleus (spa
20	19	FIRST	(Root (span 4 7) (Nucleus (span 4 6) (rel2par Joint) (Nucleus (leaf 4) (rel2par span) (text _!We are going to focus on many of the issues_!)) (Satellite
21	20	FIRST	(Root (span 8 9) (Nucleus (leaf 8) (rel2par Contrast) (text _!I am honored to have this role ,_!)) (Nucleus (leaf 9) (rel2par Contrast) (text _!but this evenir
22	21	FIRST	(Root (span 10 10) (DUMMY (leaf 10) (rel2par DUMMY) (text _!Candidates , we look forward to hearing you articulate your policies and your positions , as well as y
23	22	FIRST	(Root (span 11 11) (DUMMY (leaf 11) (rel2par DUMMY) (text _!So , let 's begin ._!)))
24	23	FIRST	"(Root (span 12 12) (DUMMY (leaf 12) (rel2par DUMMY) (text _!\\\"" We 're calling this opening segment \\\""\\\"" Achieving Prosperity.\\\""\\\""\\\"" And centra
25	24	FIRST	(Root (span 13 13) (DUMMY (leaf 13) (rel2par DUMMY) (text _!There are two economic realities in America today ._!)))
26	25	FIRST	(Root (span 14 17) (Nucleus (leaf 14) (rel2par Joint) (text _!There 's been a record six straight years of job growth ,_!)) (Nucleus (span 15 17) (rel2par Joi
27	26	FIRST	(Root (span 18 19) (Nucleus (leaf 18) (rel2par Joint) (text _!However , income inequality remains significant ,_!)) (Nucleus (leaf 19) (rel2par Joint) (text _
28	27	FIRST	(Root (span 20 22) (Nucleus (leaf 20) (rel2par span) (text _!Beginning with you , Secretary Clinton , why are you a better choice than your opponent_!)) (Sate
29	28	FIRST	(Root (span 1 1) (DUMMY (leaf 1) (rel2par DUMMY) (text _!Well , thank you , Lester , and thanks to Hofstra for hosting us ._!)))
30	29	FIRST	(Root (span 2 6) (Nucleus (span 2 4) (rel2par Joint) (Satellite (leaf 2) (rel2par Attribution) (text _!The central question in this election is really_!))
31	30	FIRST	(Root (span 7 8) (Nucleus (leaf 7) (rel2par Contrast) (text _!Today is my granddaughter 's second birthday ,_!)) (Nucleus (leaf 8) (rel2par Contrast) (text _!
32	31	FIRST	(Root (span 9 10) (Nucleus (leaf 9) (rel2par span) (text _!First , we have to build an economy_!)) (Satellite (leaf 10) (rel2par Elaboration) (text _!that wor
33	32	FIRST	(Root (span 11 12) (Satellite (leaf 11) (rel2par Attribution) (text _!That means_!)) (Nucleus (leaf 12) (rel2par span) (text _!we need new jobs , good jobs ,
34	33	FIRST	(Root (span 13 13) (DUMMY (leaf 13) (rel2par DUMMY) (text _!I want us to invest in you ._!)))
35	34	FIRST	(Root (span 14 14) (DUMMY (leaf 14) (rel2par DUMMY) (text _!I want us to invest in your future ._!)))
36	35	FIRST	(Root (span 15 17) (Nucleus (leaf 15) (rel2par span) (text _!That means jobs in infrastructure ,_!)) (Satellite (span 16 17) (rel2par Elaboration) (Nucleus
37	36	FIRST	(Root (span 18 18) (DUMMY (leaf 18) (rel2par DUMMY) (text _!We also have to make the economy fairer ._!)))

RÉFÉRENCES

Pathak, A., & Srihari, R. K. (2019). BREAKING! Presenting fake news corpus for automated fact checking. In Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop (pp. 357–362).

Haouari, F., Ali, Z. S., & Elsayed, T. (2019). bigIR at CLEF 2019: Automatic Verification of Arabic Claims over the Web. In CLEF (Working Notes).

Adair, B., Li, C., Yang, J., & Yu, C. (2017). Progress toward “the holy grail”: The continued quest to automate fact-checking. In Computation+ Journalism Symposium,(September).

Source: https://github.com/Cartus/Automated-Fact-Checking-Resources

3.

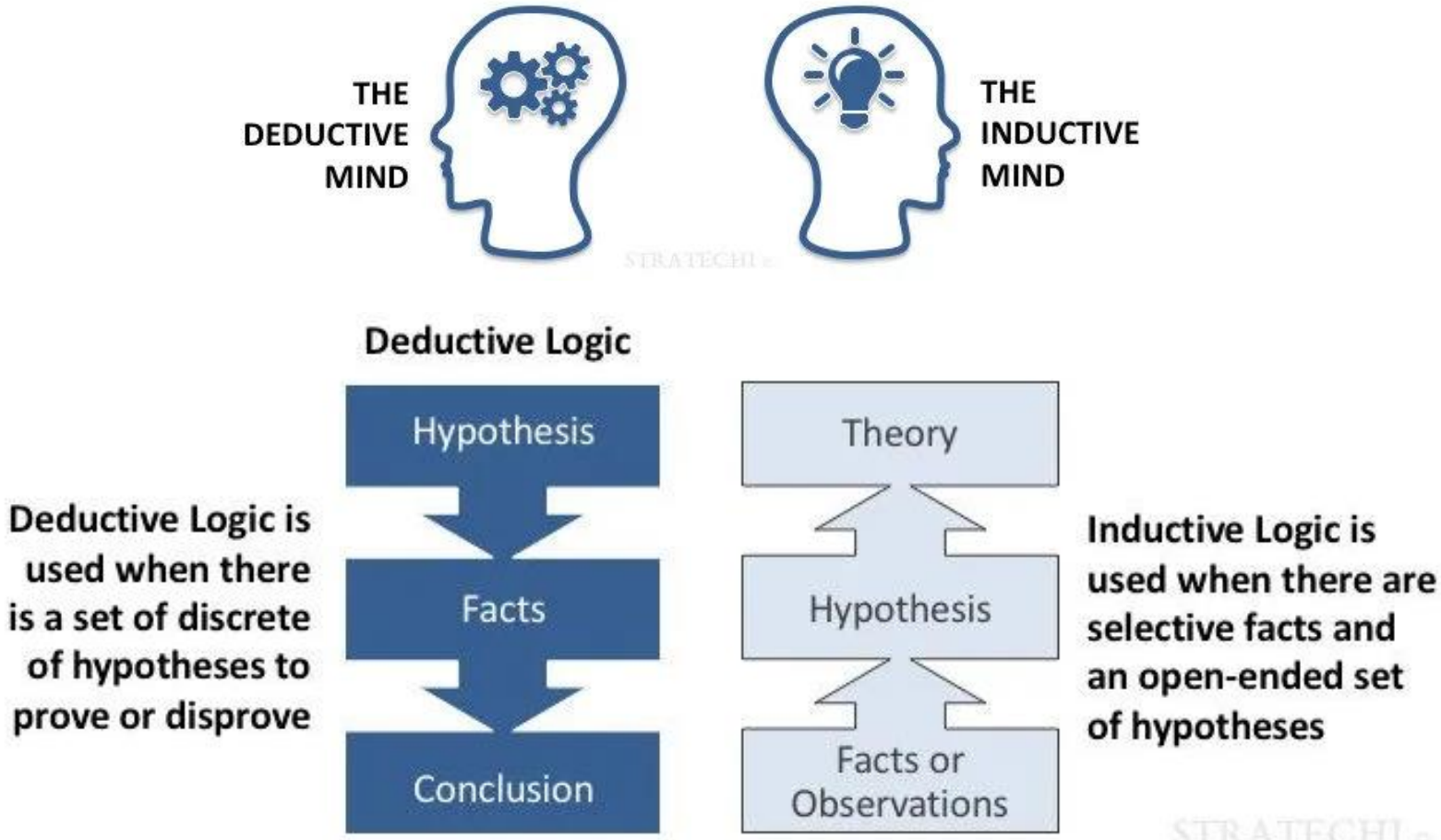
Cas d’usage:
text et opinion mining
sciences humaines et sociales

BIG DATA/NOSQL

DONNÉES GÉNÉRÉES PAR LES UTILISATEURS

STOP WORDS

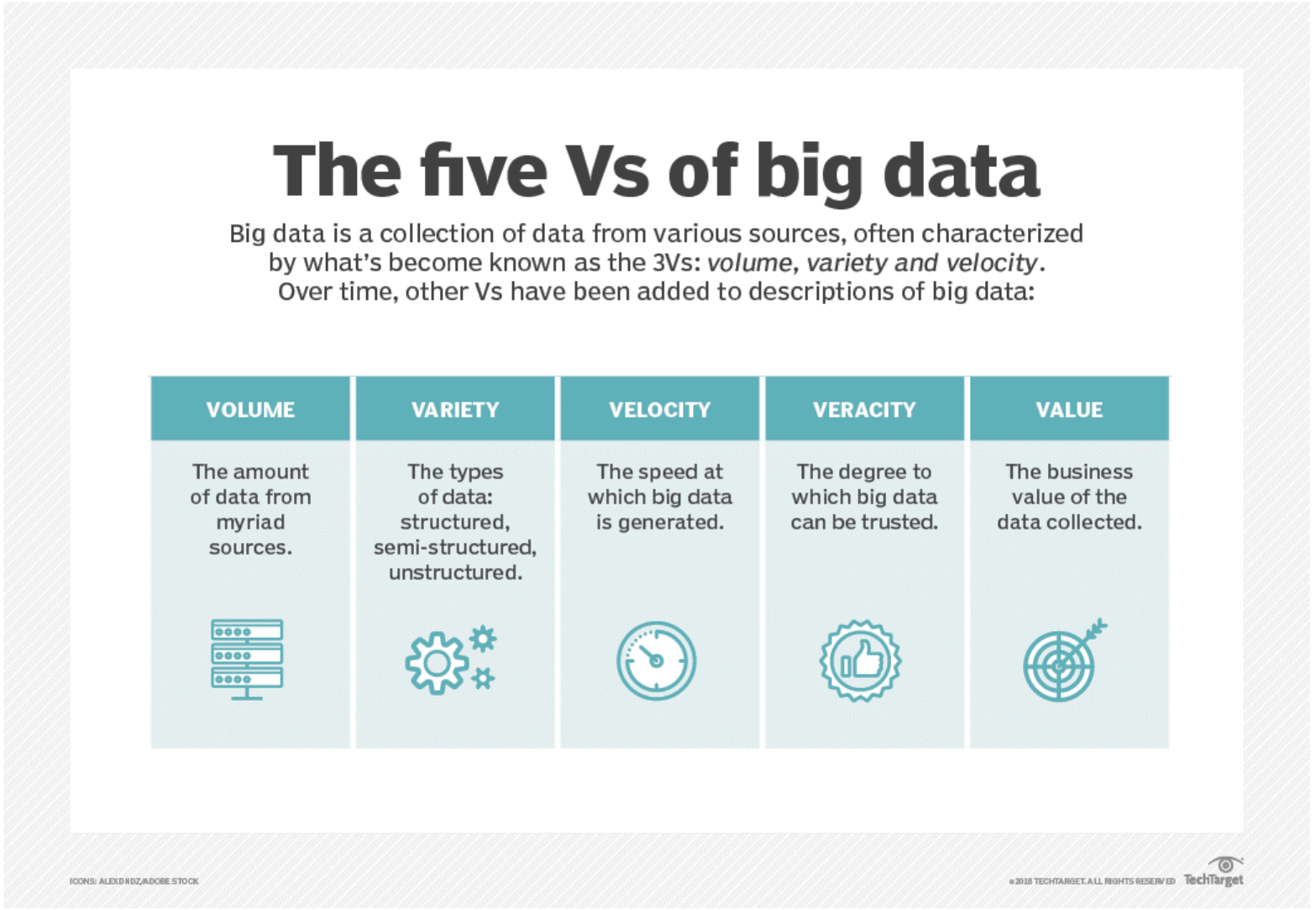
ANALYSE DE SENTIMENTS



Source: https://www.stratechi.com/deductive-inductive-logic/

RÉFÉRENCES

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, *45*, 27-45
Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*, 395-419.



RÉFÉRENCES

Ridzuan, F., Wan Zainon, W. M. N., & Zairul, M. (2022). A Thematic Review on Data Quality Challenges and Dimension in the Era of Big Data. In *Proceedings of the 12th National Technical Seminar on Unmanned System Technology 2020* (pp. 725-737). Springer, Singapore.

Elouataoui, W., Alaoui, I. E., & Gahi, Y. (2022). Data Quality in the Era of Big Data: A Global Review. *Big Data Intelligence for Smart Applications*, 1-25.

Source: Pixabay

Functions in rtweet (0.7.0)

Search all functions

get_timeline
Get one or more user timelines (tweets posted by target user(s)).

get_mentions
Get mentions for the authenticating user.

get_favorites
Get tweets data for statuses favorited by one or more target users.

get_my_timeline
Get *your* timeline

get_friends
Get user IDs of accounts followed by target user(s).

lists_subscriptions
Get list subscriptions of a given user.

lists_subscribers
Get subscribers of a specified list.

get_followers
Get user IDs for accounts following target user.

get_tokens
Fetching Twitter authorization token(s).

lists_users
Get all lists a specified user subscribes to, including their own.

lists_members
Get Twitter list members (users on a given list).

lists_statuses
Get a timeline of tweets authored by members of a specified list.

next_cursor
next_cursor/previous_cursor/max_id

post_list
Manage Twitter lists

lookup_collections
Get collections by user or status id.

Source: https://www.rdocumentation.org/packages/rtweet/versions/0.7.0

Nettoyage et préparation des données

id	lang	country	location	date	text
1,46E+18	und	Switzerland		14-12-21	gen<c3><a8>ve manifestation suisse f<c3><aa>te no<c3><ab>l covid nonaupassdelahonte tcop unfb tongue sticking out yx
1,3672E+18	fr	Belgium	Belgique	13-12-21	lataupeinfo sudinfobe g<c3><a9>nial j ai eu le covid vaccin<c3><a9> en pleine forme la prochaine que j ai le covid j irais dans une manifestation contamin<c3><a9> le plus de monde
1,3672E+18	fr	Belgium	Belgique	08-12-21	vlfree libertad yvescoppieters je parle de la manifestation contre le vaccin o<c3><b9> les gens <c3><a9>taient coller l'un aux autres je n habite pas la r<c3><a9>gion mais ayant eu le c
1,2949E+18	fr	France	Tours - Franc	08-12-21	linejacques veroniquegenest la saturation et du aux manque de lits et d'infirmi<c3><a8>re on n a suspendue les non vaccin<c3><a9> et on r<c3><a9>duit les pause et faits travaille les

```
"country","date","text"
"United Kingdom","12/11/2021","black and white thinking this group includes the millions who have already had Covid recovered now have natural immunity to the virus which is far s
"United Kingdom","12/11/2021","i'll admit I do both I'd never rely just on lft but given speed not of postal PCR I don't drive public transport with suspected Covid is antisocial
"United Kingdom","12/11/2021","on what data are you basing your statement that natural immunity is lifelong I read a lot of the small amount of research available which mostly con
"United Kingdom","12/11/2021","are you one of them I'm interested to know if you had Covid and how bad my friend is in the samehad Covid recently he's fine now but felt quite ill
"United Kingdom","12/11/2021","I had jabs and severe Covid and talking about it so in your opinion I'm trying to please antivax people bull"
"United Kingdom","12/11/2021","see robert kennedy jr must of this stuff originates from him and his friend"
"United Kingdom","12/11/2021","so will the jab stop your natural Covid response immunity from working"
"United Kingdom","12/11/2021","special I went to a and e many years ago with abcess they gave me antibiotics but because of Covid might be different I feel so bad for you seems li
"United Kingdom","12/11/2021","the vaccines are leaky ie you can catch and transmit Covid this isn't on par with previous vaccines which block transmission and infection if omicro
"United Kingdom","12/11/2021","what if you've had two jabs months ago and had Covid a month ago is that equivalent to having a booster catching Covid and having more natural antib
"United Kingdom","12/11/2021","you probably have an answer but yes you can upload the qr fr nhs app travel version to the tous anticovid app iphone"
"The Netherlands","12/11/2021","your statement is also correct but depends on the virus infection for example hepatitis a will have severe consequences so I'm jabbed if ebola was
"United Kingdom","12/11/2021"," antivaxers take to the streets in newcastle against Covid passes"
"United Kingdom","12/11/2021","Covid deniers and antivaccine people have been brainwashed blame facebook"
"United Kingdom","12/11/2021"," Covid really kicked my sorry last summer it's it horrible virus try and relax as much as possible avoid stress take antiinflammatories vitamins and
"United Kingdom","12/11/2021"," I am kind of wondering once christmas is out the way and all my family mixing is over and I have still got antibody thickened syrup blood from my b
"United Kingdom","12/11/2021"," I am not an antivaxer at all but I have not and will not have a Covid vaccine it's a personal choice and people shouldn't be alienated or discrimin
"United Kingdom","12/11/2021"," I am out so much the next couple of weeks and I am due to fly out end of the month petrified on catching Covid which may ruin plans wearing a mask
"United Kingdom","12/11/2021"," I notice that for all your look at me you failed to answer my question which was if the jab will not reduce your natural Covid immunity why not hav
"United Kingdom","12/11/2021"," I think you will find she is concerned about her brother and perhaps herself is she is jabbed just because she is making a statement just because s
"United Kingdom","12/11/2021"," omicron and Covid in general if you're unvaccinated you're a viral variant fac tory and a threat to the lives of others if you're an antivaxer in m
"United Kingdom","12/11/2021"," so pfizer created the Covid virus so that months later they could create a vaccine which they would protect people from the worst effects of the vi
```

Source: https://github.com/laurence001/mixology

Limites des contenus publiés sur Twitter

1. Le texte n'est généralement pas bien formé en termes de grammaire, de structure et de formalité du langage naturel ;
2. il n'y a pas d'harmonisation orthographique et certains termes peuvent donc présenter plusieurs orthographes différentes ;
3. l'usage courant d'abréviations ne répond pas toujours à une logique de standardisation ;
4. les mots argotiques ou jargonnant ne sont pas nécessairement inclus dans les lexiques destinés à l'analyse de sentiment ;
5. des termes non liés au domaine d'application ajoutent du 'bruit' aux contenus ;
6. le contexte n'est pas toujours bien défini ;
7. les tweets peuvent contenir des arobases, des noms d'utilisateurs, des émoticons, des hashtags, des hyperliens mais aussi du contenu non textuel. De plus, plusieurs opinions peuvent se côtoyer dans un seul et même tweet.

RÉFÉRENCES

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis* (1st ed.). Springer International Publishing.

Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*, *94*, 65–76.

Jain, S. (2021). A systematic study on sentiment analysis based on text mining and Deep learning for predictions in Stock Market trends through social and news media data. *International Journal for Research in Applied Science and Engineering Technology*, *9*(10), 1589–1593.

Kumar, A., & Sebastian, T. M. (2012). *Sentiment analysis on Twitter*. International Journal of Computer Science Issues (IJCSI), 9(4), 372.

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, *20*(1), 1–28.

Functions in textclean (0.9.3)

Search all functions

filter_row
Remove Rows That Contain Markers

add_comma_space
Ensure Space After Comma

replace_contraction
Replace Contractions

replace_emoji
Replace Emojis With Words/Identifier

replace_emoticon
Replace Emoticons With Words

swap
Swap Two Patterns Simultaneously

textclean
Text Cleaning Tools

add_missing_endmark
Add Missing Endmarks

replace_tokens
Replace Tokens

check_text
Check Text For Potential Problems

replace_url
Replace URLs

replace_date
Replace Dates With Words

replace_white
Remove Escaped Characters

replace_word_elongation
Replace Word Elongations

replace_email
Replace Email Addresses

print.sub_holder
Prints a sub_holder object

replace_internet_slang
Replace Internet Slang

replace_kern
Replace Kerned (Spaced) with No Space Version

print.which_are_locs
Prints a which_are_locs Object

replace_money
Replace Money With Words

replace_tag
Replace Handle Tags

Source: https://www.rdocumentation.org/packages/textclean/versions/0.9.3

Pre-processing

The pre-processing phase is essential before analysing the data. It consists of a six-step process. Duplicates were considered as having the same ID, localisation, Date Time, and text. The orthographical standardization aimed to harmonize spellings. For instance, unvaxxed or unvaxed became unvaccinated.

RÉFÉRENCES

Fox, C. (1992). Information retrieval data structures and algorithms. Lexical Analysis and Stoplists, pp. 102–130.

Fox, C. (1989, September). A stop list for general text. In *Acm sigir forum* (Vol. 24, No. 1-2, pp. 19-21). New York, NY, USA: ACM.

Lo, R. , He, B., & Ounis, I. (2005, January). Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management: Special Issue*

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of information science*, *18*(1), 45-55.

Ressource: https://journodev.tech/lanalyse-de-sentiment-ou-levaluation-de-la-subjectivite/

Construire un lexique spécialisé

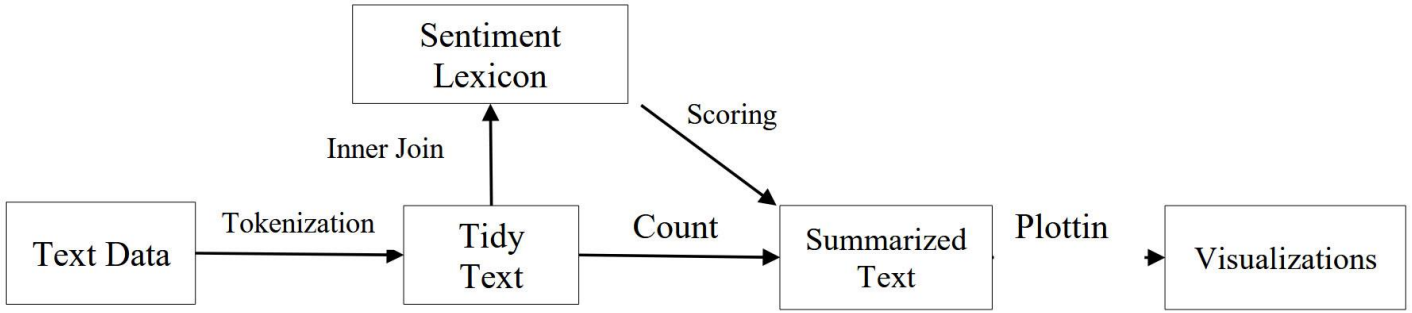
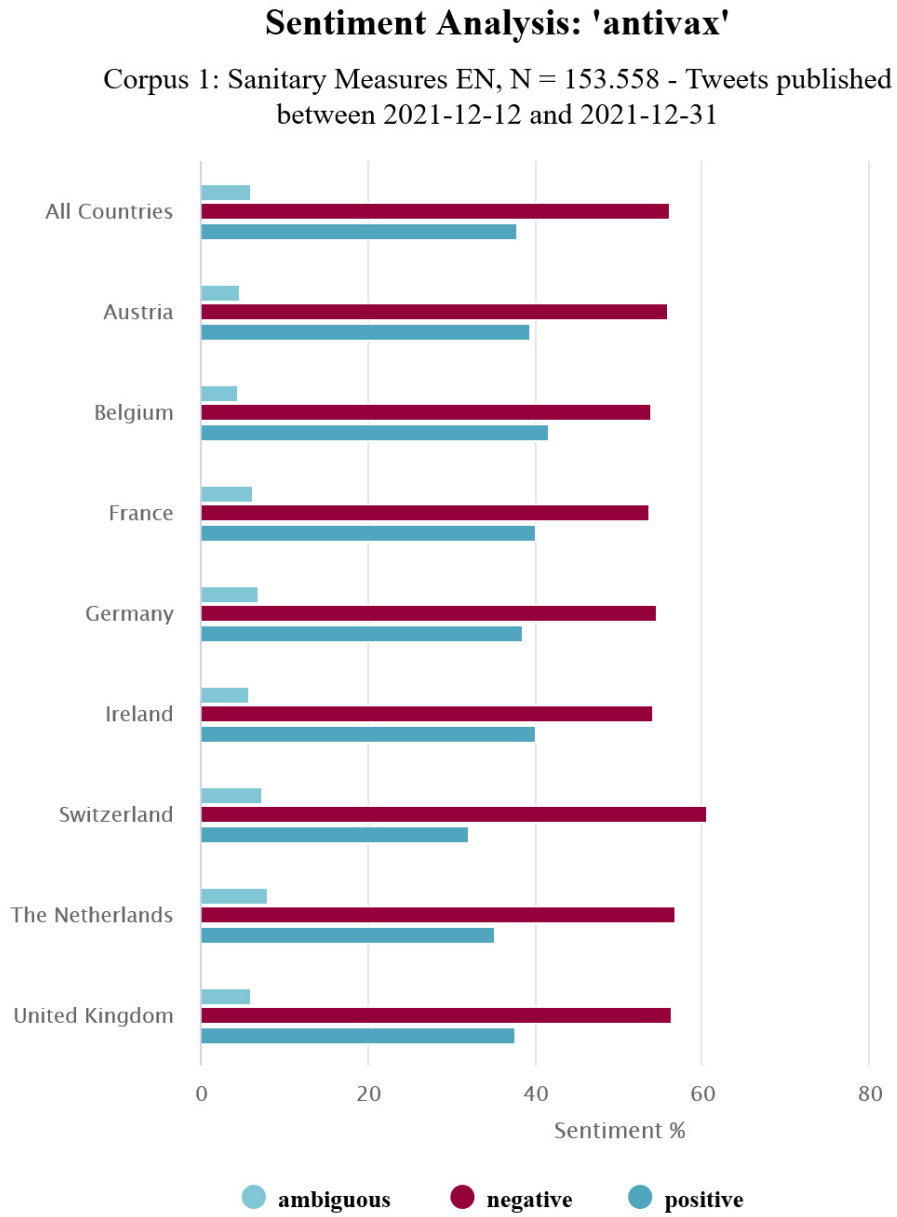


Figure 2. A flowchart of a typical text analysis that uses tidytext for sentiment analysis.

Source: adapted from Silge and Robinson (2017, p.15)



Ressource: https://github.com/laurence001/mixology

Conclusions et perspectives pour la recherche

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas C. Redman

April 02, 2018

Source: https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

De nombreuses perspectives pour la recherche

- Stratégies de maintenance de la qualité des jeux de données annotés
- Définition d'indicateurs de qualité pour chaque étape d'un processus ML
- Définition de prévention des erreurs et des méthodes de validation de la qualité des données
- Gestion et maintenance de la qualité des données non textuelles
- Gestion et maintenance de la qualité de données issues de sources multiples
- Gestion de l'incertitude de résultats affectés par le niveau de qualité des données
- Problématique des données générées par les utilisateurs (crowdsourcing, big data)
- Focus sur les cas d'usages et besoins des utilisateurs finaux

Merci pour votre attention !

ULB UNIVERSITÉ LIBRE DE BRUXELLES