TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL (NLP) ET MEILLEURES PRATIQUES APPLIQUÉES À LA CONNAISSANCE CLIENT

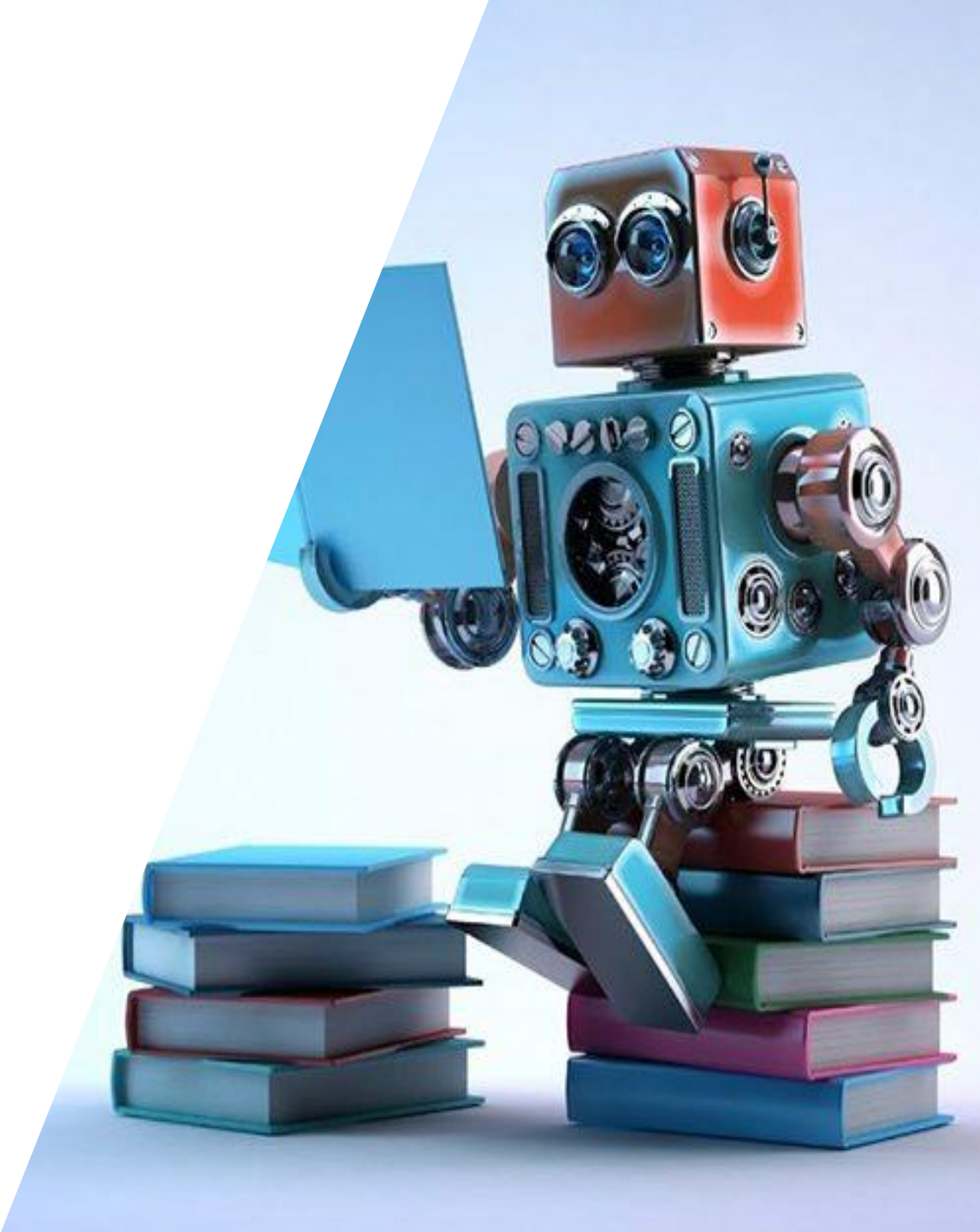
Tanguy LE NOUVEL
Data scientist, Chaire D/CC
PAST au département STID de l'IUT de Vannes
Directeur DataScience - Micropole

01 Quelques définitions

02 Applications

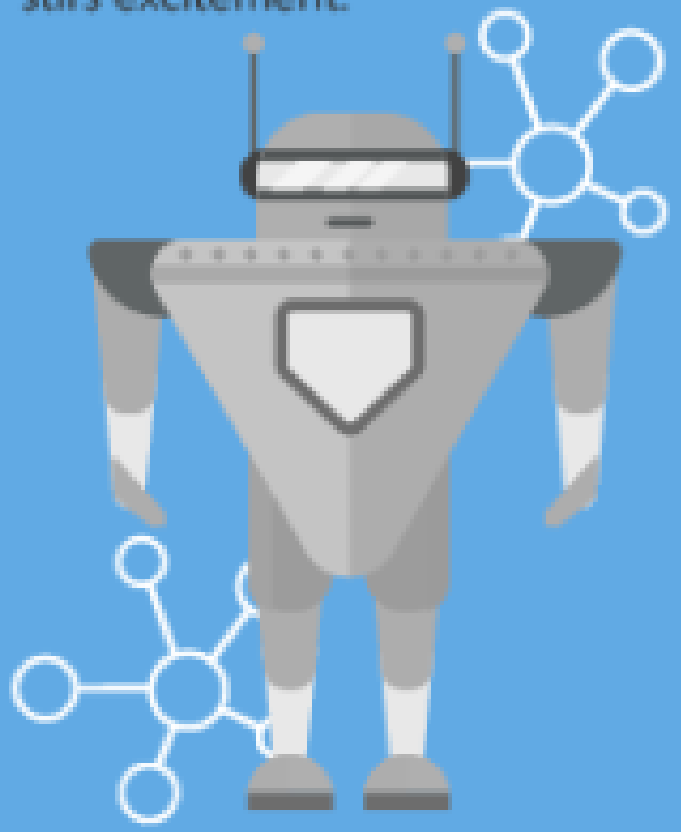
03 Méthodologie

QUELQUES DÉFINITIONS



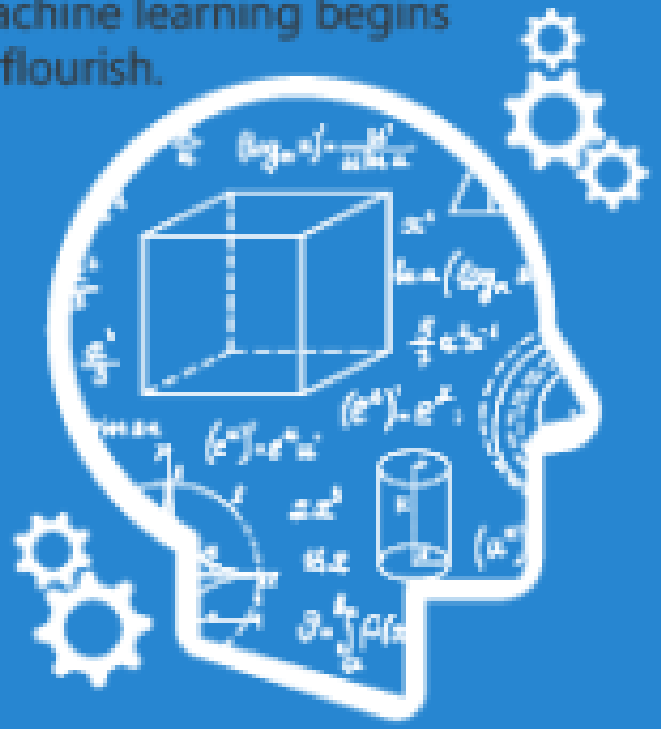
ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



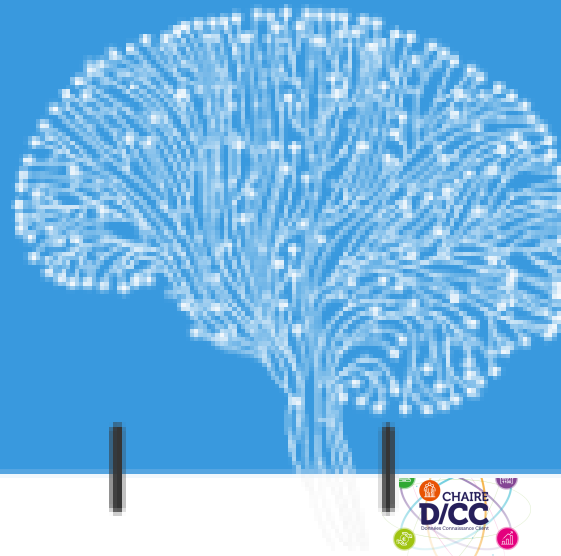
MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

Deep learning breakthroughs drive AI boom.



1950's 1960's 1970's 1980's 1990's 2000's 2010's

INTELLIGENCE ARTIFICIELLE

- Le concept d'intelligence artificielle est **né dans les années 50** suite au questionnement d'Alan Turing :
  « Une machine peut-elle penser comme un être humain ? »

- **Aujourd'hui**, le terme « Intelligence artificielle » est un peu galvaudé et désigne ***« n'importe quelle technique informatique qui permet d'aboutir à un résultat sophistiqué »****

- **L'IA s'intéresse à l'ensemble des facultés cognitives humaines** : perception, apprentissage, raisonnement, expression, action/décision

*Définition donnée par Cédric Villani dans l'émission « On n'est pas couché » du 24 mars 201

FOCUS SUR TROIS SOUS-DISCIPLINES DE L'IA

NLP ou TALN :
Natural **L**anguage **P**rocessing
Traitement **A**utomatique du **L**angage **N**aturel

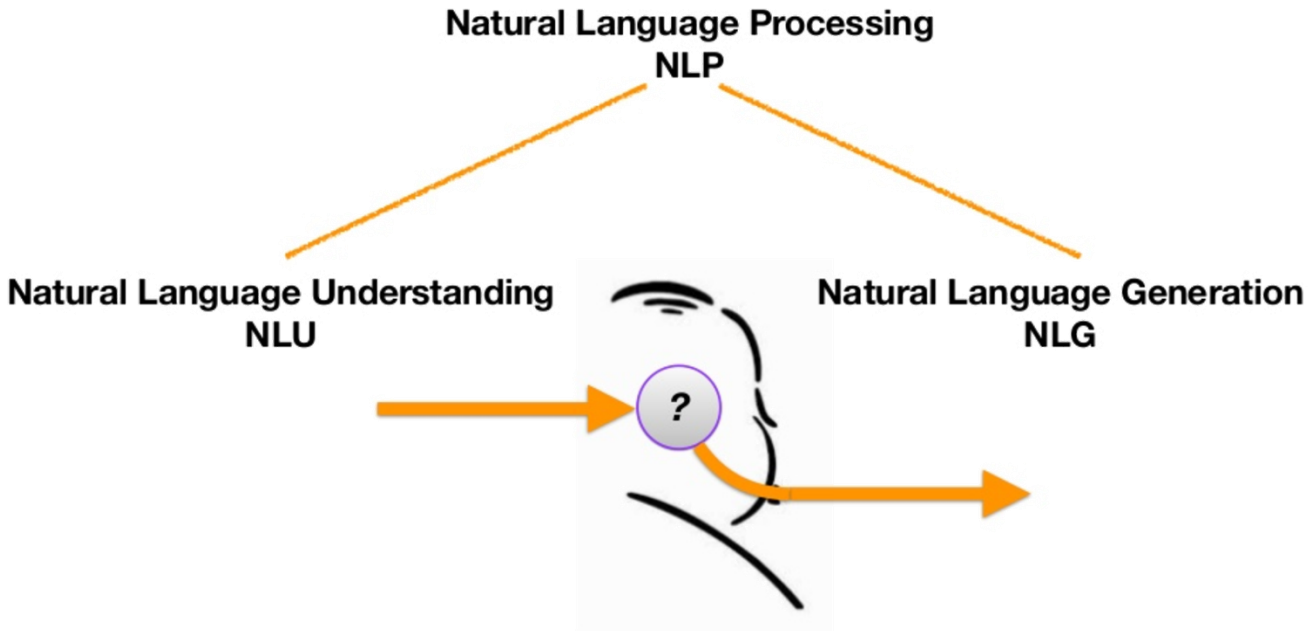
Text Mining : extraction de connaissances à partir de données textuelles
(bloc-notes, documents, conversations enregistrées, emails, sites internet…)

Machine Learning :

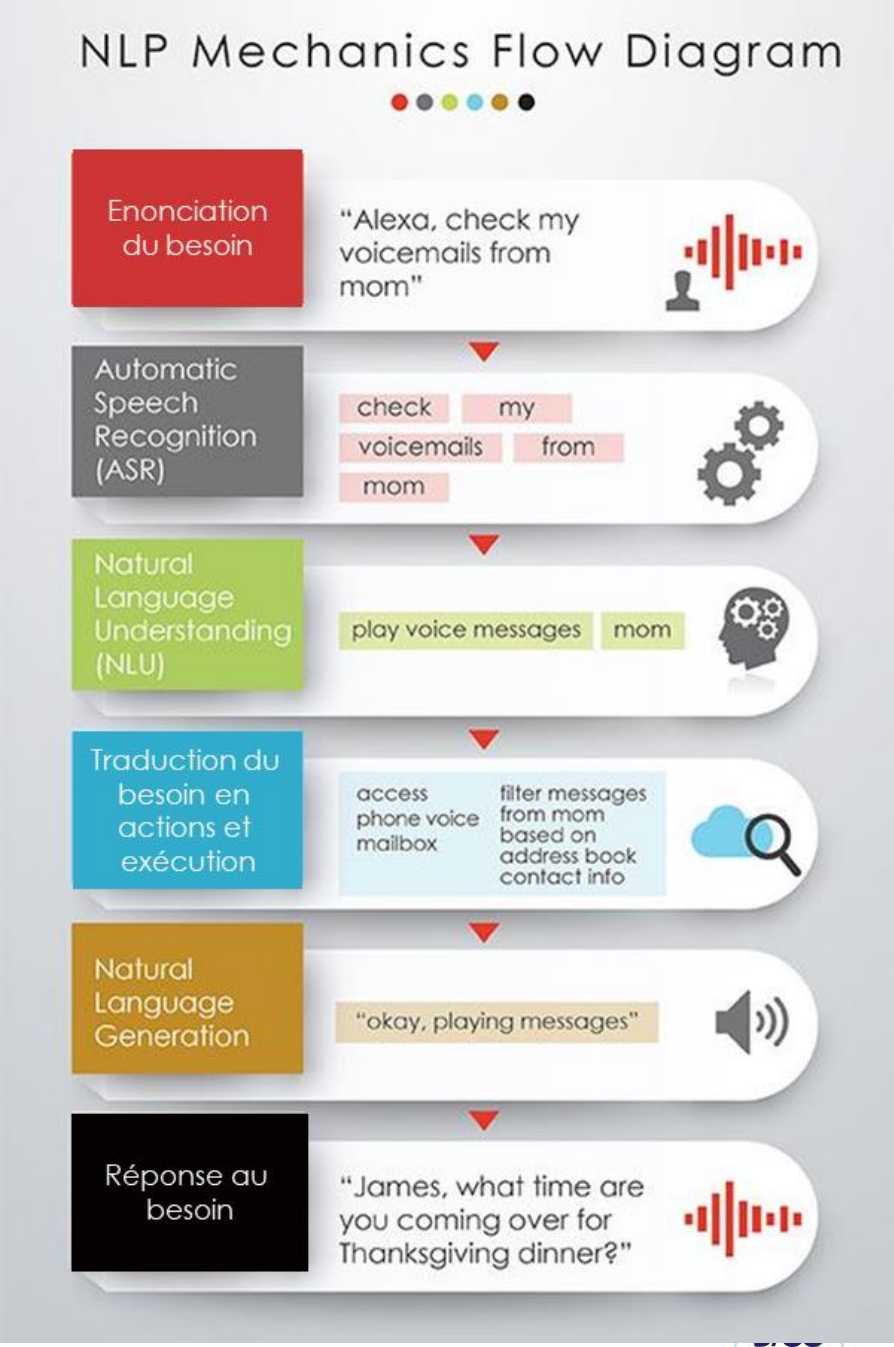
Apprentissage machine à partir d'exemples

NLP - NATURAL LANGUAGE PROCESSING
NLP = NLU + NLG

- Discipline qui s'intéresse à la **compréhension**, au **traitement** et à la **génération du langage naturel par les machines**, que ce langage soit **écrit ou parlé**.
- Vise à **faciliter les échanges entre la machine et l'humain**.
- Repose sur les connaissances issues de la **linguistique**, de l'**informatique** et de l'**intelligence artificielle**.



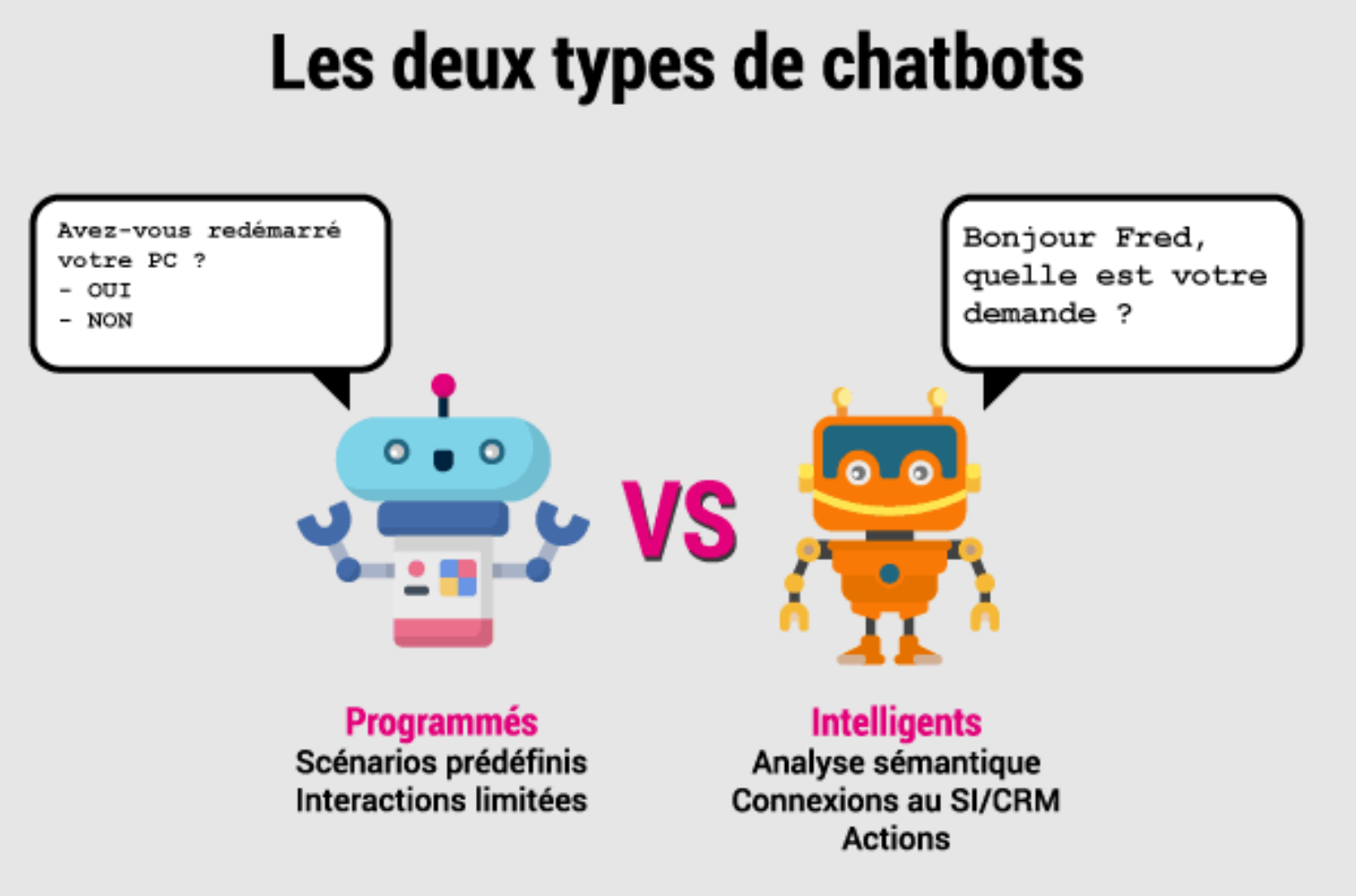
- NLP = TALN en français pour Traitement Automatique du Langage Naturel



NLP ET COMPRÉHENSION DU BESOIN
RECHERCHE D'ENTITÉS NOMMÉES

	« Bonjour »	« Quel temps fait-il ? »	« Quel temps fait-il à Paris demain ? »
Intention	**Aucune**	**Météo**	**Météo**
Entité(s)			**Lieu** **Date**

CHATBOTS


Les deux types de chatbots
Avez-vous redémarré votre PC ?
- OUI
- NON
Bonjour Fred, quelle est votre demande ?
VS
Programmés
Scénarios prédéfinis
Interactions limitées
Intelligents
Analyse sémantique
Connexions au SI/CRM
Actions

LES CHATBOTS ET LEURS LIMITES

- **Incapables de gérer tout type de demandes**
- **Savent traiter des besoins simples** à condition d'être **hyper spécialisés**
- **Pertinents pour les tâches répétitives**
- **Coût d'entrainement / Apprentissage potentiellement élevé**

« L'IA ne fonctionne pas sans données annotées par l'expertise de l'humain. Ce qui marche en IA, c'est l'apprentissage supervisé, on donne la question et la réponse. Les machines sont bonnes pour des tâches simples et très répétitives, et sont fondamentalement primaires. L'IA c'est comme Google, ça permet de faire plus de choses plus vite ».

Édouard d'Archimbaud, ex Head of Lab Data Science et IA de BNP Paribas CIB

TEXT MINING - ANALYSE TEXTUELLE
2 APPROCHES COMPLÉMENTAIRES

- **STATISTIQUE TEXTUELLE**
  Etude des textes par les méthodes quantitatives

- **ANALYSE SÉMANTIQUE**
  Etude du contenu et de la signification par rapport à un univers de référence (dictionnaires, catégories (verbe, adjectifs…)) :

REVENONS AUX ORIGINES DU TEXT MINING

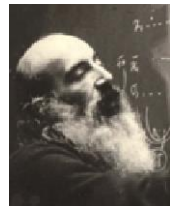
1916 : Amélioration de la transcription sténographique (Estoup)

1935 : Psychobiologie du langage (Zipf). Loi de Zipf.

1954 : Concept de statistique linguistique.
Publication de l'ouvrage **« Les caractères statistiques du vocabulaire »** par Pierre Guiraud

1960 : Concept d'analyse de données (Tukey, Benzécri)

1964 : 1er cours de linguistique mathématique de J.P. Benzécri à la faculté des sciences de Rennes


Jean-Paul Benzecri

1965 : thèse de Brigitte Escoffier **sur l'analyse des correspondances** (ancien Professeur à l'Université de Rennes et à l'IUT de Vannes)

1970 - 1980 : concept de statistique textuelle

1979 : 1ère version du logiciel ALCESTE (pour « Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte »), conçu en 1979 par Max Reinert (laboratoire de Jean-Paul Benzécri, CNRS, Paris)

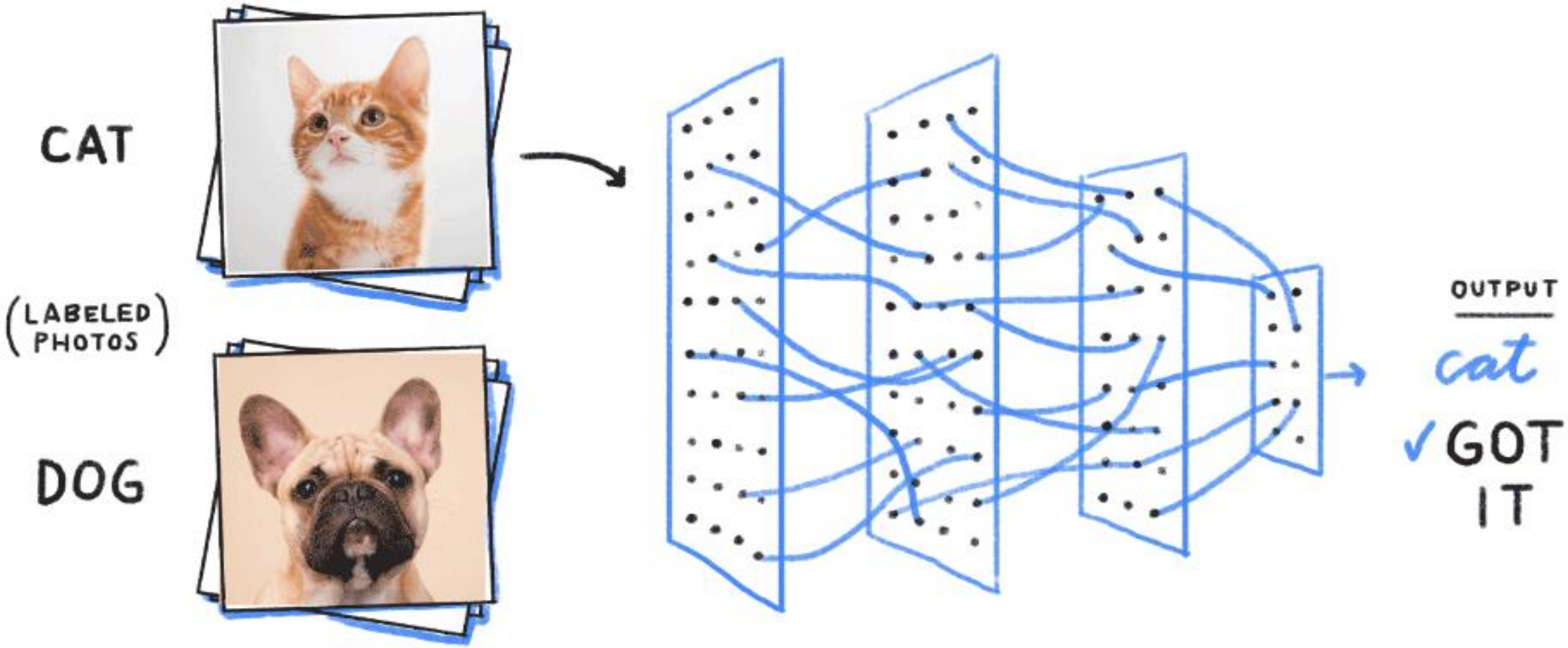
1981 : Une procédure d'Analyse Lexicale open source écrite en Langage Fortran - *Les Cahiers de l'Analyse des Données* – Ludovic Lebart

1985 : 1ère version du logiciel SPAD-T (Système Portable pour l'Analyse des Données Textuelles)

1988 : Publication de l'ouvrage « Analyse Statistique des Données Textuelles » par A. Salem et L. Lebart

MACHINE LEARNING

- **Sous-discipline de l'IA** principalement utilisée pour l'apprentissage



APPLICATIONS

EXEMPLES D'APPLICATION

Etude comparative sur le vocabulaire des grands auteurs : XVIIe siècle / Analyse de discours

Recherche documentaire, bibliométrie, veille technologique

Traitement des questions ouvertes dans les enquêtes / Analyse des avis consommateurs

Reconnaissance d'entités nommées (noms propres, personnes, lieux, dates, quantités, valeurs…)

Extraction d'informations / Détection d'évènements / Structuration de données non structurées

Topic Modeling : identification des thématiques, sujets abordés dans un texte

Classification automatique de documents (lettres de réclamation, publications scientifiques, CVs)

Enrichissement / Recommandation de contenu / Publicité contextuelle

Recherche d'informations (Google PageRank)

Web-Scraping et analyse des médias sociaux

Traduction automatique de documents / discours / vidéos

Support / Assistants
Chatbots (Siri, Google Assistant, Alexa) **et systèmes de questions / réponses** (Selfcare)

Gestion automatique des emails entrants (réponse automatique et/ou routage)

Filtrage des spams dans nos boîtes mail

Résumé automatique de textes / documents

Analyse de sentiments

e-Réputation

GDPR Compliance
Recherche de données à caractère sensible et de données personnelles dans les zones texte libre

Cyber Criminalité, Cyber Sécurité

Risque / Fraude
Exploitation des pièces jointes pour mieux prédire le risque de défaut / fraude

COSMÉTIQUE
ANALYSE DES RETOURS CLIENTS

Enjeux :

- **Prendre en compte la parole client** dans les **actions de fidélisation**
- Exploiter finement les **attentes / besoins / frustrations** exprimées à travers les **champs texte libre** disponible dans les différents formulaires (enquête, site web)

Réalisation :

- **Analyse des réponses aux questions ouvertes** concernant la **satisfaction**
- **Identification des thématiques** abordées et **analyse de la polarité**
- **Mise en évidence de frustrations et d'éléments de satisfaction** qui jusque là n'avaient pas été identifiés comme tel
- **Automatisation** des traitements

Résultats :

- **Traitement de plusieurs dizaines de milliers de commentaires non traités jusqu'ici**
- **Détection des motifs de satisfaction et/ou d'insatisfaction**
- **Monitoring de la satisfaction client** et évaluation de l'**impact des actions menées**

LUXE
WEB SCRAPING ET RECHERCHE D'INFLUENCEUSES

Enjeux :

- **Etudier et analyser les différents posts / commentaires / tweets faisant référence à la marque** sur une liste de sites prédéfinis
- **Identifier les produits concernés** s'il en question, exploiter les images s'il y en a afin d'aider à leur identification
- **Analyser la polarité du contenu** et sa **viralité** pour **identifier de nouvelles influenceuses**

Réalisation :

- **Web scraping** des principaux sites retenus à partir de mots-clés spécifiques
- **Structuration du contenu**
- **Text mining** pour **interpréter** le contenu du post
- **Computer Vision** pour identifier le logo de la marque sur l'image et détourer autour du logo pour **identifier l'objet concerné**
- **Synthèse des principales influenceuses identifiées**

Résultats :

- **Un éclairage nouveau sur la marque et certains de ses produits** avec des caractéristiques mises en avant et particulièrement appréciées (retweet…)
- **L'identification d'une nouvelle typologie d'influenceuses** jusque-là hors des radars.

SECTEUR ASSURANCE
ANALYSE TEXT MINING DES RÉCLAMATIONS CLIENTS

Enjeux :

- **Améliorer les performances du dispositif de gestion des réclamations clients** en **automatisant le fléchage aux bons services**
- **Identifier les principaux motifs de réclamation** et **améliorer leur traitement** en avance de phase (FAQ, communication)

Réalisation :

- **Catégorisation des réclamations** sur la base de leur contenu textuel
- **Mise en évidence des principales thématiques abordées** et **polarité associée**
- **Apprentissage Machine Learning de la catégorisation** à dires d'experts
- **Automatisation du dispatch** et du suivi des réclamations
- **Identification des points sensibles** et de leur impact au cours du temps

Résultats :

- **Meilleure prise en charge des réclamations clients**
- **Identification de motifs de réclamations spécifiques** à certaines régions / départements / produits
- **Mise en évidence d'incohérences dans certains process,** notamment digitaux

SECTEUR BANCAIRE
EXPLOITATION DES PIÈCES JOINTES POUR MIEUX PRÉDIRE LE RISQUE DE DÉFAILLANCE DES ENTREPRISES

Enjeux :

- **Exploiter le contenu des rapports** (pièces jointes) **des chargés d'affaires pour** identifier des informations permettant de mieux prédire la défaillance **Intégrer ces nouvelles informations dans le formulaire de qualification des dossiers pour affiner le scoring**

Réalisation :

- **Constitution de la base d'analyse**
- **Nettoyage des pièces jointes** (word, pdf, documents océrisés)
- **Statistiques descriptives des principaux termes employés**
- **Sélection de thèmes à investiguer**
- Pour chaque thème, **qualification contextuelle**
- **Structuration des données** sous forme de nouvelle variables explicatives
- **Evaluation** de l'apport de ces nouvelles informations dans l'amélioration de la prédiction du défaut

Résultats :

- **Identification de plusieurs informations pertinentes à intégrer dans le formulaire de demande de financement**
- **Meilleure prédiction de la défaillance**

MÉTHODOLOGIE

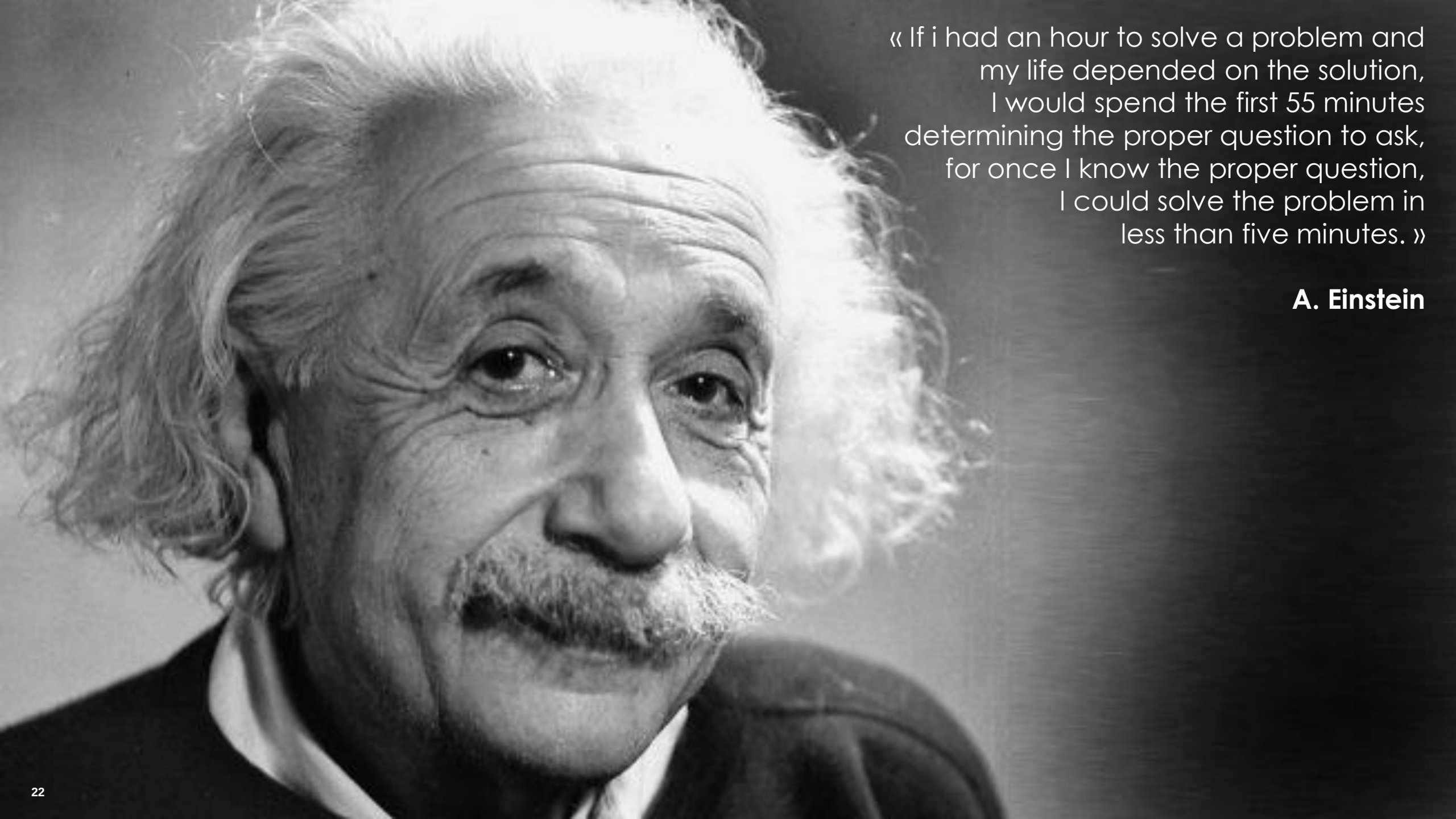
MÉTHODOLOGIE
UN GROS TRAVAIL DE PRÉPARATION DE DONNÉES ET DE NETTOYAGE

- **Transformation de la source en texte** (speech to text, image to text…)
- **Normalisation et nettoyage** (suppression des caractères exotiques, des éléments inutiles…)
- **Tokenisation** (Découpage des textes en phrases, des phrases en mots / lemmes)
- **Correction orthographique** (fautes de frappes, fôtes d'ortaugraf, abréviations…)
- **Recherche d'entités nommées** (personnes réelles ou fictives, organismes, localisation, période, équipement…)
- **Désambiguisation - Etiquetage morpho-syntaxique**
  - **Catégorie grammaticale** : nom, verbe, pronom, adverbe
  - **Genre** : féminin / masculin
  - **Nombre** : pluriel / singulier
  - **Temps** : imparfait, présent, futur…
- **Suppression des mots vides de sens** (à partir d'une liste de « stopwords »)
- **Racinisation** (réduction des mots à leur racine)
- **Lemmatisation**
  - Regroupement des mots d'une même famille (a, ai eu, avoir) ou de même sens (heureux, content) au sein d'un seul lemme à partir d'un dictionnaire éventuellement enrichi par des spécificités liées au corpus étudié
- **Recherche de suites de mots** (expressions, segments de mots répétés, ngrams) **fréquentes**
  Ex. : « pas du tout satisfait »

MÉTHODOLOGIE
OUTILS STATISTIQUES

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- **Fréquence des mots**
- **Nuages de mots** (aucun intérêt…)
- **Construction d'une matrice textes * mots** (Bag of words)
- Analyse des correspondances pour **identifier / visualiser les mots caractéristiques d'un texte, visualiser les proximités entre différents mots**
- Classification non supervisée pour **regrouper les textes de contenus similaires, éventuellement les mots, expressions**
- LDA (Latent Dirichlet Analysis) pour **identifier les thématiques abordées dans chaque texte**
- **Classification supervisée** pour prédire un évènement à partir des données textuelles et structurées
- **Analyse de sentiments pour identifier la polarité** (satisfait, neutre, insatisfait…) à partir de modèles prédictifs prédéfinis ou définis sur-mesure



« If i had an hour to solve a problem and
my life depended on the solution,
I would spend the first 55 minutes
determining the proper question to ask,
for once I know the proper question,
I could solve the problem in
less than five minutes. »

A. Einstein

MERCI
DES QUESTIONS ?

Tanguy LE NOUVEL
Data scientist, Chaire D/CC
PAST au département STID de l'IUT de Vannes
Directeur DataScience - Micropole