**1. Les projets cognitifs sont caractérisés par la présence d'au moins un composant entrainé par l'exemple dans la solution.** La présence de ce composant introduit des spécificités dans le cycle de vie du projet et dans les compétences impliquées.

L'apprentissage par l'exemple ou machine learning est utile lorsque les humains savent intuitivement réaliser une tâche au cas par cas, mais qu'ils ne savent pas décomposer cette tâche en un mode opératoire précis, une suite de décisions et opérations élémentaires, qui fonctionne à tous les coups.

Limitations et contraintes de l'apprentissage par l'exemple :

* Ne doit être appliqué que pour modéliser une opération déterministe / reproductible
  + Une relation de cause à effet, telle que les mêmes données en entrée produisent toujours les mêmes résultats en sortie
  + Une relation suffisamment stable dans le temps pour que le modèle entrainé quelques semaines plus tôt soit encore valable au moment où on le déploie
  + Une relation dans laquelle on sait mesurer directement tous les facteurs qui influencent le résultat de manière significative
  + Une relation objective, dont le résultat ne dépend pas de l'individu qui réalise la tâche (nécessite un consensus entre acteurs, cf ci-dessous)
* Ne peut être appliqué que si on parvient à collecter suffisamment d'exemples représentatifs du comportement attendu
  + Ne peut modéliser que des cas fréquents et récurrents, pas des cas particuliers jamais rencontrés auparavant qui nécessitent initiative, décision, éthique, créativité
  + La participation d'utilisateurs et de référents métier est nécessaire pour collecter des exemples représentatifs du comportement à reproduire par la machine
  + La représentativité des exemples collectés par rapport aux futures données de production est essentielle au bon fonctionnement de la solution
  + Les exemples collectés doivent couvrir au maximum la diversité des cas qui pourront se présenter en production
  + Lorsque le résultat attendu relève d'une décision ou d'un jugement humain, tous les acteurs doivent s'aligner sur le résultat attendu : ces règles de décision doivent être clairement documentées
* Ne fournit pas de garantie de résultat pour une prédiction isolée
  + Seuls les cas suffisamment similaires aux exemples examinés dans la phase d'entrainement peuvent faire l'objet de prédictions pertinentes
  + De manière générale, on ne sait pas évaluer de manière fiable si une prédiction isolée est digne de confiance ou non (une activation forte du modèle n'est pas une garantie)
  + Puisqu'on n'a pas su définir de règles précises / exhaustives pour décrire le résultat attendu, seule l'intervention d'un référent humain permet d'évaluer a posteriori la qualité d'une prédiction
  + Important : on ne peut pas "corriger" une prédiction isolée qui serait erronée, ni garantir qu'une prédiction le demeure tout au long de la durée de vie d'une solution. On peut seulement annoncer des performances moyennes sur un grand nombre de prédictions.

La plateforme doit permettre de documenter et d'évaluer ces conditions de succès au moment de la création d'un projet et lors de toutes les phases de sa conception.

**2. Le succès d'un projet cognitif se mesure dans le cadre d'un processus métier, à travers son impact sur les objectifs de l'entreprise et les intérêts de ses clients.**

La plateforme distingue et englobe :

* La prédiction de l'intention / l'analyse des données en entrée
* La mise en relation avec des connaissances préalables / un contexte pour comprendre
* Les règles de décision / de détermination de l'action
* La mesure des réactions de l'utilisateur / des impacts sur le processus

Ces quatre composants d'une solution ne peuvent pas être conçus indépendamment les uns des autres : ils doivent être optimisés de bout en bout comme un tout cohérent pour donner les résultats attendus. Un focus exclusif sur le composant cognitif d'analyse des données ne permet pas d'atteindre des objectifs métier, d'avoir un impact sur le fonctionnement de l'entreprise.

Une base de connaissances, un référentiel de documents et de réponses, un moteur de règles, un système décisionnel pour collecter et interroger les indicateurs métier : tous ces composants ont une place aussi importante dans une plateforme cognitive qu'un environnement de développement de modèles de machine learning.

Lorsqu'une solution doit être déclinée dans plusieurs langues ou pour plusieurs organisations dans l'entreprise, cette séparation en 4 étapes favorise une meilleure réutilisation.

La plateforme doit outiller ces 4 composants et fournir une vision globale sur le bon fonctionnement de l'ensemble, sans focus exclusif sur le 1er d'entre eux.

**3. Les solutions cognitives sont des systèmes collaboratifs intégrant des aller-retours continus entre les humains (concepteurs, pilotes, utilisateurs) et la machine.**

En constituant les jeux de données d'entrainement, les concepteurs humains spécifient exemple par exemple le comportement attendu de la machine.

Lorsque plusieurs expertises sont en jeu à l'intérieur de l'entreprise, les concepteurs doivent se mettre d'accord entre eux.

Lorsque l'objectif est de satisfaire un utilisateur, le comportement attendu évolue au fil du temps, au fil de la progression des concepteurs dans la compréhension des attendus des utilisateurs.

Lorsqu'une partie de la tâche est complexe à modéliser, ou lorsque la machine est incertaine sur une prédiction, la solution la plus simple et la plus efficace est souvent de poser la question directement à l'utilisateur.

L'humain a ses points forts et la machine a les siens : dans un cas d'utilisation donné, il faut commencer par bien identifier les avantages de chacun.

Toujours privilégier un travail sur l'expérience utilisateur - qui incite à une collaboration riche entre l'humain et la machine, en laissant à chacun la tâche qui constitue son point fort - plutôt qu'un travail très complexe sur la machine seule pour essayer de réaliser une prédiction hasardeuse.

L'humain reste souvent nécessaire en complément du déploiement d'une solution cognitive pour traiter tous les cas particuliers.

Puisque la prédiction fournie n'est jamais certaine, la prédiction assiste une action humaine, on ne peut pas automatiser l'action à 100%.

L'idée de laisser la machine prendre par elle-même des décisions (par opposition à répéter des décisions que les concepteurs ont prise sur des cas similaires) n'a pas de sens dans la pratique.

Seule une supervision humaine permet de mesurer la performance de prédiction.

Après le déploiement d'une solution cognitive, une équipe d'amélioration continue doit rester mobilisée en permanence pour évaluer manuellement et régulièrement le bon fonctionnement du modèle en production.

La plateforme ne doit pas se limiter à l'outillage des parties informatiques, mais doit inclure nativement les activités et les feedbacks de tous les humains dans la boucle.

Ces interventions humaines représentent l'essentiel du coût d'un projet. Il est préférable d'investir des ressources de R&D dans des algorithmes qui vont accélérer et faciliter le travail des concepteurs et des pilotes, plutôt que pour gagner quelques pourcents dans la performance du modèle entrainé.

**4. Un projet cognitif est un travail d'équipe qui nécessite la collaboration de plusieurs compétences complémentaires.**

La plateforme :

* Définit des rôles en charge de l'ensemble des activités
* Présente une interface optimisée pour chaque rôle
* Organise la collaboration entre les rôles

6 rôles se partagent l'usage de la plateforme :

* Gestion de l'infrastructure et des déploiements, packaging des services
* Conception des modèles et des pipelines, pretraining
* Intégration dans les interfaces utilisateurs et dans le système d'information
* Entrainement des modèles et configuration des pipelines
* Amélioration continue et surveillance des solutions
* Analyse et pilotage des processus métier

On n'essaie pas de simplifier à l'extrême ces problématiques de manière à ce qu'un unique rôle d'analyste métier puisse gérer l'ensemble de ces aspects.

Toute complexité technique inutile doit être masquée, mais aucune décision impactante sur la conception et les performances de la solution ne doit être inaccessible à l'équipe projet. Chacun doit maîtriser les subtilités de son rôle.

**5. Dans une grande entreprise, une solution cognitive est déployée sur différents canaux de communication, et pour différentes entités organisationnelles : l'architecture de la solution doit permettre une mise en commun efficace des connaissances du domaine métier.**

Une solution cognitive capture une partie de la connaissance de l'entreprise. Cette information peut souvent être restituée par différents canaux : mobile, web, email, téléphone, messagerie / et sous différentes formes : conversation, messages, documents, recherche. La structure de l'information à restituer reste la même.

Afin d'offrir une expérience cohérente et de ne pas démultiplier les efforts, les solutions cognitives doivent être conçues nativement de manière à permettre de restituer un contenu similaire, mais avec des variantes, sur différents médias.

Pour permettre cette mise en commun, la structure des connaissances du domaine métier doit être capturée dans un référentiel unique de manière centralisée. Ce modèle métier est extrait des informations brutes collectées au fil des projets, dans des ateliers conjoints entres des experts métiers et des analystes de données.

Un composant cognitif prend en entrée des données non structurées et en extrait en sortie des informations structurées. La forme des structures de données en sortie est un extrait du modèle métier du domaine.

La mise en commun de ce modèle de données est obligatoire pour pouvoir faire des analyses de fonctionnement transverses dans l'entreprise, sur un processus métier de bout en bout, le long d'un parcours client, ou pour toutes les opérations réalisées autour d'un produit.

Une grande entreprise déploie des solutions similaires dans plusieurs géographies, sur plusieurs segments de clientèle, dans plusieurs marques acquises au fil du temps. Toutes ces entités organisationnelles travaillent fondamentalement dans le même domaine métier, mais on fait des choix d'organisation un peu différents, nomment leurs produits avec des noms différents, favorisent différents types de relation client, etc …

La plateforme doit permettre une mise en commun d'un maximum d'éléments entre ces différentes entités, tout en permettant à chacune de personnaliser son expérience et ses réponses à son niveau.

Au départ d'un projet cognitif, les connaissances de l'entreprise sont capturées dans des référentiels documentaires et dans un système d'information. La plateforme cognitive doit fournir les moyens de se synchroniser sur cette information pour la restituer à travers des solutions développées.

**6. Dans une entreprise de services, les solutions cognitives les plus fréquentes sont les suivantes :**

Texte :

* Assistant Conversationnel
* Analyse d'emails
* Recherche avancée
* Extraction d'informations de documents
* Analyse des avis client / utilisateurs
* Requêtes structurées en langage naturel

Voix :

* Serveur Vocal Interactif (téléphone)
* Commandes en langage naturel
* Dictée de messages
* Compte-rendu de réunion
* Transcription d'échanges oraux

Images :

* Typage d'images de documents
* Extraction de texte d'images de documents
* Vérification de l'authenticité de pièces justificatives
* Identification d'objets
* Estimation de sinistres
* Comptage de personnes / de bâtiments
* Authentification / identification d'utilisateurs

Données structurées / tabulaires :

* Etude des facteurs qui influencent un résultat
* Prédiction des valeurs futures de séries temporelles
* Recommandations basées sur un historique de préférences
* Scoring de risques
* Détection de fraudes
* Anticipation d'attrition client ou de pannes matérielles

Au fil des versions, la plateforme doit fournir progressivement des motifs de solution prédéfinis pour chacun de ces cas d'utilisation fréquents.

**7. Les solutions cognitives se déploient souvent en deux temps : un premier projet introduit des capacités d'analyse et d'action sur un canal d'échange d'information, puis ces capacités sont mises à disposition de manière transverse (cross--canal) pour différentes populations, à des fins d'optimisation des processus métier et des échanges.**

Dans un premier temps, le déploiement de solutions cognitives se traduit par des projets lourds d'intégration de nouvelles capacités dans les canaux d'échanges d'information existants : email, téléphone, chat, courrier …

Ces capacités sont de deux types :

* Analyse des informations échangées : thématique, intentions, entités, sentiment
* Actions d'assistance : routage des demandes, statistiques de pilotage, conversation, réponses, raccourcis de traitement, remontée d'opportunités

Un premier niveau de fonctionnalités est fourni par une équipe pluridisciplinaire.

Après avoir déployé ce premier niveau de fonctionnalités sur plusieurs canaux de communication, l'entreprise peut passer à une deuxième phase :

* Les analyses remontées des différents canaux sont alignées et mises en commun
* Les règles reliant les actions d'assistance aux intentions détectées sont harmonisées
* Les collaborateurs qui travaillent à l'optimisation des processus peuvent exploiter ces capacités pour fluidifier les échanges et les processus

Par exemple :

* Le service marketing chargé d'un produit peut analyser tous les échanges qui ont lieu dans l'entreprise et avec les clients autour de ce produit, puis modifier ses caractéristiques, sa documentation, ses procédures de support, afin de faciliter son utilisation et éviter un maximum de questions
* L'équipe en charge de la souscription et de la gestion du produit peut identifier toutes les étapes du parcours client qui nécessitent un échange avec l'entreprise, et enrichir les fonctionnalités des différents canaux de communication pour faciliter et accélérer ces échanges via des routages, des réponses automatiques, des raccourcis de traitement

La plateforme doit fournir des méthodes et des outils pour les deux phases :

* Accélérateurs pour intégrer des capacités cognitives dans un nouveau canal de communication
* Accélérateurs pour exploiter au mieux les canaux de communication augmentés dans les projets d'optimisation des processus métier

La population visée dans la première phase est principalement constituée d'équipes informatiques travaillant avec des référents métier.

La population visée dans la deuxième phase sont les pilotes des processus métier, les concepteurs des produits, les organisateurs des services de support.

**8. Les solutions cognitives doivent être robustes pour une utilisation en production : de nombreux composants annexes doivent être insérés dans un pipeline de traitement autour du modèle principal pour assurer une utilisation maîtrisée et conforme.**

Les composants suivants sont nécessaires dans un pipeline de machine learning pour le rendre utilisable en production :

* Assistances à la saisie ou à la capture des données
* Limitation du nombre et de la taille des requêtes en fonction de la capacité de calcul
* Contrôle de conformité du format des données reçues
* Filtrage des données de qualité insuffisante pour l'analyse
* Amélioration de la qualité des données reçues
* Normalisation des données reçues pour réduire la variété des cas
* Simplification des données reçues en fonction des besoins de l'analyse
* Aiguillage ou exclusion des données reçues selon le domaine d'application des modèles entrainés
* Filtrage des données contenant un contenu offensant, choquant, illégal
* Anonymisation des données avant visualisation par des opérateurs
* Croisement des données avec des référentiels de connaissances préalables
* Prédiction d'un niveau de confiance dans les résultats du modèle

La plateforme doit fournir des composants réutilisables pour réaliser ces différentes fonctions dans un pipeline de traitement à moindre coût, pour chacune des modalités et des cas d'utilisation supportés.

**9. Un projet cognitif reste un projet informatique : en entreprise, une gestion rigoureuse du cycle de vie et des mises en production est nécessaire pour garantir une maîtrise de la disponibilité et des performances de la solution.**

La plateforme doit mettre en œuvre les référentiels, les pistes d'audit, les processus, les environnements, les outils de test qui permettent de garantir :

* Une traçabilité et une historisation complète de toutes les modifications réalisées
* Une validation de ces modifications dans des environnements de test avant déploiement en production
* Une surveillance des performances et de la disponibilité des composants en production

Une spécificité des composants cognitifs : dans un même projet, ils peuvent être mis au point en trois temps

* Un data scientist développe l'architecture d'un modèle et un pipeline de traitement
* Un analyste métier entraine ce modèle en améliorant de manière continue son jeu de données d'entrainement
* Un pilote de processus ajuste les règles de détermination de l'action à exécuter en fonction des prédictions du modèle

Ces trois étapes sont liées, leur évolution est conjointe et itérative, mais les déploiements se font bien en trois temps, dans des horizons de temps différents, pour des groupes d'utilisateurs qui constituent une chaîne de clients / fournisseurs internes.

L'environnement de production pour le premier rôle est l'environnement de développement du rôle suivant.

Dans une solution cognitive, les environnements de développement, test, production ne sont pas définis de manière absolue, mais prennent un sens différent pour chacun des 6 rôles identifiés.

La plateforme doit prendre en compte cette complexité pour assurer un environnement aussi stable que possible à chacun des contributeurs et utilisateurs à son niveaux.

**10. Un projet cognitif manipule par essence une grande quantité de données caractérisant ses utilisateurs : le respect de la vie privée et de la confidentialité des données doit être assuré dans toutes les phases du projet.**

Un contrôle d'accès doit être mis en place afin de tracer et d'autoriser toute visualisation des données collectées.

Les données personnelles ou confidentielles ne doivent pas quitter les serveurs sécurisés de la plateforme.

Une piste d'audit doit permettre de tracer tous les accès aux données confidentielles / sensibles.

Dans la mesure du possible, les données doivent être anonymisées / désensibilisées avant d'être stockées.

La plateforme doit fournir des services performants d'anonymisation des données, automatisés ou manuels.

La plateforme doit appliquer automatiquement une étape d'anonymisation sur les données de production avant de permettre leur import dans les environnements d'entrainement.

Un contrôle humain régulier de la performance des services d'anonymisation doit être mis en place.

Lorsque les données concernent plusieurs entités organisationnelles distinctes au sein de l'entreprise, les intervenants d'une entité ne doivent pas avoir accès aux données de l'autre entité.

Le contrôle d'accès est donc basé sur des rôles qui sont affectés dans le cadre d'un projet, mais également sur le périmètre organisationnel accessible à l'intervenant.

A des fins de mise en commun des données d'entrainement, des données issues d'une entité organisationnelles peuvent être déversées par un administrateur autorisé dans une zone de partage commune.

**11. L'architecture d'un modèle, son dimensionnement, et la technologie choisies pour implémenter un composant cognitif sont à sélectionner en fonction des caractéristiques unique des jeux de données du cas d'utilisation considéré, des contraintes de performance de la solution, et de l'état de l'art à un instant t.**

La variété des jeux de données et des contextes d'utilisation est infinie, il n'existe pas d'algorithme idéal qui donne les meilleurs résultats dans tous les cas.

Sans exposer tous les détails techniques sous-jacents, la plateforme doit laisser à la main du concepteur du composant cognitif un certain nombre de choix et d'indications qui dérivent de sa connaissance du domaine métier :

* Hypothèses et régularités attendues sur les données en entrée, permettant de simplifier le modèle
* Compromis coût de calcul / latence / performances
* Compromis explicabilité / simplicité / performances
* Compromis quantité de données d'entrainement / performances
* Compromis rappel / précision
* Réutilisation de modèles entrainés sur des tâches similaires comme point de départ
* Mise en commun de modèles entre plusieurs projets
* Compositions de modèles étape par étape ou entrainement d'un modèle de bout en bout

La plateforme ne doit pas dépendre d'un fournisseur de technologie unique.

Elle doit au contraire permettre l'intégration de technologies de multiples fournisseurs, d'outils open source, de développements maison.

L'intérêt du concepteur est de pouvoir sélectionner dans chaque projet la technologie la plus adaptée à son besoin.

On pourrait imaginer de laisser la machine comparer elle-même les performances de multiples combinaisons pour optimiser des métriques définies par l'utilisateur, mais cette approche :

* est potentiellement trop coûteuse en ressources pour être mise en pratique efficacement dans l'entreprise
* masque complètement le fonctionnement du modèle pour les concepteurs de la solution

Ce dernier point n'est pas souhaitable : il semble impératif que le concepteur de la solution ait une compréhension de haut niveau du fonctionnement de l'algorithme.

Cette compréhension est nécessaire pour savoir comment constituer un jeu de données d'entrainement pertinent.

Elle est surtout indispensable pour appréhender les limitations du modèle entrainé, et pour déterminer le domaine d'application dans lequel il est raisonnable de le déployer.

Cette compréhension du fonctionnement de l'algorithme peut être rendue accessible à des non spécialistes par une représentation simplifiée du modèle sous forme d'étapes successives de distillation de l'information, au cours desquelles on résume, on combine ou on propage les informations initiales.

Pour faciliter l'expérimentation de plusieurs technologies ou familles de modèles sans impact sur le reste de la solution, l'interface d'un composant cognitif est décrite indépendamment de son implémentation par la notion de Tâche cognitive.

Une Tâche cognitive est définie par :

* Une liste de données en entrée, parmi les types suivants :
  + Langage naturel
    - Texte
    - Document
    - Image de document
    - Voix
  + Données tabulaires
    - Liste de caractéristiques
    - Séries temporelles
  + Images
    - Images de personnes
    - Images d'objets et de lieux
    - Images satellites
* Une liste de résultats attendus en sortie, qui sont le plus souvent :
  + Données structurées
    - Catégories (énumération)
    - Valeurs continues
    - Séquences
  + Génération
    - Texte
    - Image
    - Voix
* Une liste de métriques de performance à optimiser

La signature de ces tâches (description des données en entrée et des résultats en sortie) est un élément central de structuration de la plateforme.

Elle est le plus souvent prédéfinie dans le cadre d'un type de solution et réutilisée d'un projet à l'autre.

Des prétraitements sur les données, des Tâches cognitives, une logique de décodage des résultats sont combinées au sein d'un Pipeline pour former une fonction de traitement de bout en bout.

Le Pipeline peut être déployé et exposer une API après avoir été packagé sous forme de Service.

Les Pipelines et les Services ont également une interface définie indépendamment de leur implémentation.

**12. Plateforme en couche / composition de services unitaires / templates de solutions**

**13. Infrastructure de calcul spécialisée, aujourd'hui à base de GPUs Nvidia ; nécessité d'organiser le partage et l'optimisation de l'utilisation de cette infrastructure**

**14. Mesure de la consommation de ressources et facturation, surveillance des environnements, mesure de performance.**

**15. L'entrainement d'un modèle est l'activité centrale dans la conception d'un composant cognitif; c'est aussi une activité coûteuse en main d'œuvre et difficile à maîtriser : tous les outils doivent être fournis pour optimiser l'effort et mettre les performances sous contrôle.**

Capacité de généralisation : Gestion des jeux de données d'entrainement, de validation, de test

Importance extrême de la qualité du jeu de validation

Importance du prétraining

Gestion d'un cycle de vie / amélioration constante

Assistance à l'entrainement sans masquer la complexité

Validation éthique, biais, identification des impacts, équité

Explicabilité : idée de prédire un programme d'exécution intelligible, des étapes intermédiaires