Data Challenge

Kit Data Challenge

Juillet 2022

HEALTH DATA HUB

Table des matières

HEALTH DATA HUB

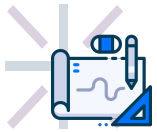
#1 Qu’est-ce qu’un Data Challenge ?

Qu’est-ce qu’un Data Challenge ?

Un Data Challenge est une **compétition en science des données** ouverte dont l’objectif est de **résoudre une problématique** spécifique de data science grâce à des **algorithmes d’intelligence artificielle**. Cette compétition se déroule **en ligne** et repose sur **un large jeu de données** mis à disposition par les organisateurs via une **plateforme dédiée** -la plus connue étant Kaggle-.

Qu’est-ce qu’un Data Challenge ?

Les grandes phases de l’organisation d’un Data Challenge…



CADRAGE SCIENTIFIQUE, REGLEMENTAIRE ET OPÉRATIONNEL

… la **problématique** et les **objectifs** doivent être clairement définis et pouvoir être **traités par une approche de classification supervisée**. Le **circuit des données**, en particulier lorsqu’il s’agit de données de santé, doit être **identifié et sécurisé** d’un point de vue **réglementaire** avec l’aide d’un **DPO**.

COLLECTE DES DONNÉES ET CONSTITUTION DE LA BASE

… les données nécessaires à la réalisation de l’analyse seront **collectées** et **centralisées** dans des **formats homogènes**. Cette base sera divisée en **différentes sous-bases** qui serviront à l'entraînement, au test et au classement des algorithmes.
Dans le cadre de l’utilisation de données de santé, cette base devra être **anonymisée** et une **analyse des risques de réidentification** produite pour s’en assurer.

DESIGN DE LA COMPÉTITION ET HÉBERGEMENT SUR UNE PLATEFORME DÉDIÉE

… la compétition est **hébergée sur une plateforme dédiée** mettant à disposition les **données** et éventuellement de la **puissance de calcul** aux compétiteurs. Sur cette plateforme, les compétiteurs pourront échanger et soumettre leurs algorithmes pour figurer dans le classement général de performance. Cette performance sera déterminée en fonction d’une **métrique d’évaluation** pondérant les erreurs éventuelles.

REMISE DES PRIX, ANALYSE ET VALORISATION DES RÉSULTATS

… A la fin de la compétition les compétiteurs ayant produit l’algorithme **le plus performant** sont **récompensés**.

Exemple de Data Challenge en santé

Retour d'expérience sur le Data Challenge en pathologie du col de l'utérus de la Société Française de Pathologie en 2020

Les données mises à disposition comprennent des milliers de **lames numérisées de biopsies du col de l'utérus** provenant de plusieurs centres médicaux français. L'objectif pour les compétiteurs est de **classer chaque image** (4 classes allant de 0 (bénin) à 3 (cancer invasif) en fonction de la catégorie la plus sévère de lésion épithéliale présente dans l'échantillon.

Data Challenge en pathologie du col de l'utérus - Société Française de Pathologie - Health Data Hub

https://www.youtube.com/watch?v=Ue0Lt1RKaAE&t=502s

Le challenge de la société française de pathologie en quelques chiffres...

574 Participants du monde entier

36 Équipes constituées

Des scores finaux très prometteurs allant jusqu'à **94%**

Déroulement de la compétition - phases

Temps 1

Phase d'apprentissage

Quelques semaines ou mois

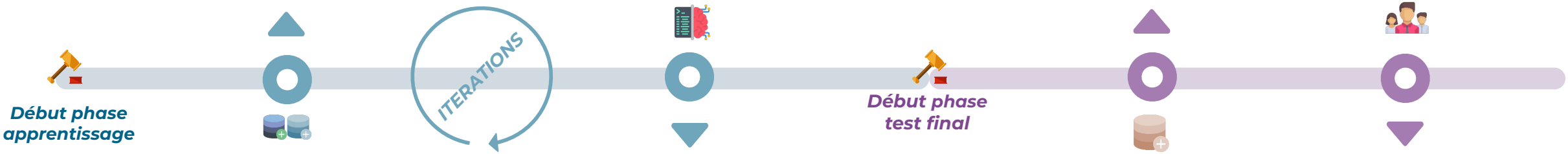
Temps 2

Phase de test final

Entre 48h et 72h

Partage d'un jeu de données d'apprentissage aux compétiteurs

Les compétiteurs prennent connaissance des données à analyser et programment et entraînent leur algorithme sur un jeu de données labélisé (annoté) mis à disposition sur la plateforme du Challenge.



Début phase apprentissage

Soumission des algorithmes pour test intermédiaire

Les compétiteurs ont la possibilité de soumettre leur algorithme sur la plateforme pour évaluer sa performance sur un nouveau jeu de données pour consulter ses performances et permettre son amélioration.
La plateforme précise le score de performance de l'algorithme à partir duquel est déterminé un classement provisoire.

Début phase test final

Soumission des algorithmes pour test final

Les participants soumettent la version finale de leur algorithme sur la plateforme (une seule soumission). L'algorithme est testé sur un 3ème jeu de données permettant d'obtenir le score final de performance.

Classement final des compétiteurs

Les participants sont classés en fonction du score de performance obtenu, les meilleurs sont récompensés.

Déroulement de la compétition - classement

Lors des phases de test intermédiaire et de soumission finale des algorithmes, la performance de ces derniers est évaluée selon une **métrique d'évaluation** qui permet de **pondérer les éventuelles erreurs** commises par l'algorithme.
Un faux positif pourra par exemple impacter plus faiblement le score de performance qu'un faux négatif.
Le **score de performance** obtenu permet de classer les participants dans un classement en temps réel dit « **leaderboard** ».

ERROR TABLE

	Class 0 (pred)	Class 1 (pred)	Class 2 (pred)	Class 3 (pred)
Class 0 (actual)	0.0	0.1	0.7	1.0
Class 1 (actual)	0.1	0.0	0.3	0.7
Class 2 (actual)	0.7	0.3	0.0	0.3
Class 3 (actual)	1.0	0.7	0.3	0.0

Exemple – Pondération des erreurs dans le Data Challenge de la SFP
Le score pour chaque prédiction est égal à [1 – valeur de l'erreur]

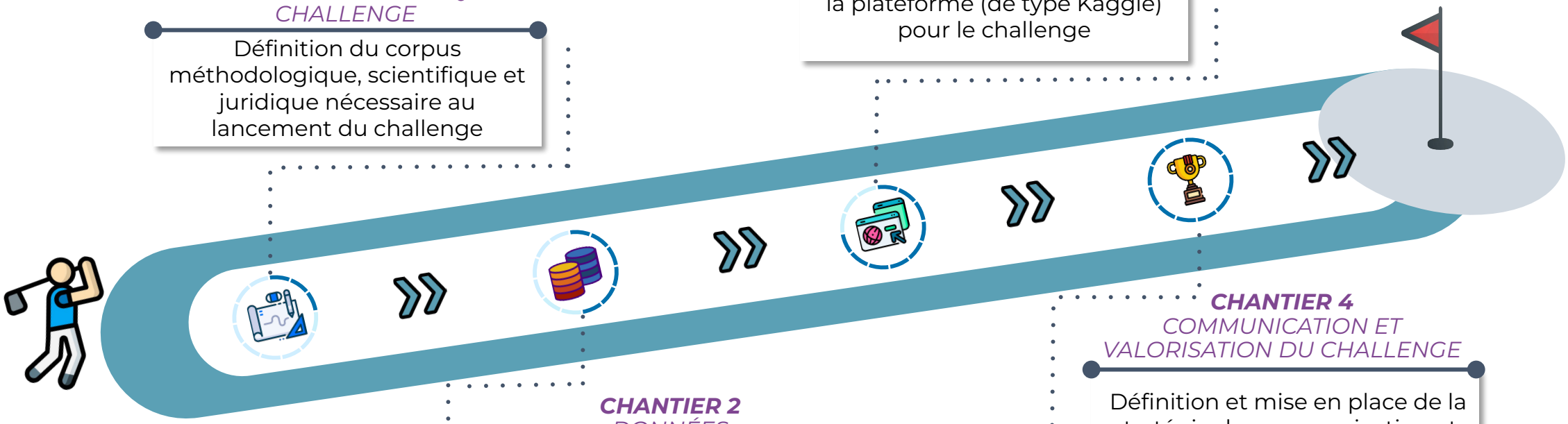
	User or team		Best private Weighted Class Score	Timestamp	# Entries	
	Tribvn-Healthcare	1	0.9475	2020-10-28 11:35:30	7	
	karelds	2	0.9345	2020-10-29 13:15:31	6	
	kbrodt	3	0.9339	2020-10-28 16:50:58	8	
	LifeIs2Short	4	0.9332	2020-10-29 17:58:08	6	
	wangww	5	0.9265	2020-10-16 19:00:37	4	
	algoscope	6	0.9252	2020-10-27 19:27:05	2	
	Sen_Sen	7	0.9232	2020-10-13 03:54:14	4	
	jjing	8	0.9223	2020-10-14 03:26:48	1	
	loktarxiao	9	0.9194	2020-10-15 03:50:32	4	

Exemple - leaderboard final du Data Challenge de la SFP

#2 4 chantiers à adresser pour réussir un Data Challenge

4 chantiers à adresser pour réussir un Data Challenge

L'organisation d'un Data Challenge nécessite d'adresser **4 chantiers** sur une durée de **6 à 12 mois** : il s'agit de définir la **problématique du challenge** et ses **caractéristiques**, de mettre à disposition les **données les plus pertinentes**, d'en faciliter **l'accès et l'usage** afin que « l'expérience compétiteur » soit la plus réussie et enfin de **valoriser l'évènement** pour en tirer le maximum de **retour sur investissement**.



CHANTIER 1
CADRAGE MÉTHODOLOGIQUE DU CHALLENGE

Définition du corpus méthodologique, scientifique et juridique nécessaire au lancement du challenge

CHANTIER 2
DONNÉES

Collecte, création et mise à disposition des bases de données labellisées et annotées

CHANTIER 3
PLATEFORME

Contractualisation et gestion de la plateforme (de type Kaggle) pour le challenge

CHANTIER 4
COMMUNICATION ET VALORISATION DU CHALLENGE

Définition et mise en place de la stratégie de communication et valorisation des résultats

Déroulement type et gouvernance

Cadrage scientifique, reglementaire, opérationnel & financier

Collecte des données auprès des centres inclueurs

Labélisation des données, constitution de la base

Data Challenge

Communication, valorisation des résultats

Constitution des comités

Convention avec les centres ou organismes

Base de données finale

Hébergement sur la plateforme

Remise des prix

Ouverture des données et algorithmes

Equipe projet

- ✔ Suivi quotidien du projet sur les aspects juridiques, organisationnels, techniques et financiers

Comité d'organisation

- ✔ Porteurs de projet
- ✔ Contribution d'un délégué à la protection des données
- ✔ Coordination des prestataires et organismes, dépositaires de données

Conseil scientifique

- ✔ Spécialistes médicaux
- ✔ Data scientists et experts en IA et ML
- ✔ Définition de la question médicale et de la métrique d'évaluation
- ✔ Vérification de la faisabilité technique du challenge

Comité d'annotation

- ✔ Spécialistes dans le domaine concerné
- ✔ Mise en place d'une stratégie de labélisation des données
- ✔ Annotation et labélisation des données

Gouvernance - focus sur le CO et CS

Comité d'organisation (CO)

Objectifs

- Le CO sera en charge de :
  - ✔ L'organisation générale de l'événement avant / pendant / après la phase finale, du recrutement et de la gestion RH en découlant,
  - ✔ De la mise en œuvre éventuelle de prestations de service,
  - ✔ De la mise en œuvre de la communication autour de l'événement,
  - ✔ Des actions sur le terrain auprès des établissements de santé, des patients le cas échéant, des compétiteurs
- Le CO centralisera l'ensemble des échanges autour du projet et s'appuiera sur les recommandations du CS

Membres

- Sponsor(s)
- Donneurs d'ordres des partenaires du Data Challenge
- Chef de projet Data Challenge

Fréquence

- Mensuel

Conseil scientifique (CS)

Objectifs

- Le CS sera en charge de :
  - ✔ Définir la question posée et vérifier qu'elle puisse être traduite en une problématique de classification supervisée avec une métrique d'évaluation adaptée permettant de sélectionner les vainqueurs
  - ✔ S'assurer de la capacité à produire une base de données permettant d'obtenir des résultats robustes a priori
  - ✔ De répondre en lien avec le CO aux questions des interlocuteurs pour les aspects scientifiques / informatiques / méthodologiques,
  - ✔ De suggérer les éléments de communication autour de l'événement,
  - ✔ De réfléchir aux modalités de mise en ligne des données et des algorithmes après l'événement (i.e. Open Source et Open Data),
  - ✔ De réfléchir aux modalités de publication / valorisation éventuelle après l'événement (vidéos, interview des lauréats, articles scientifiques…)

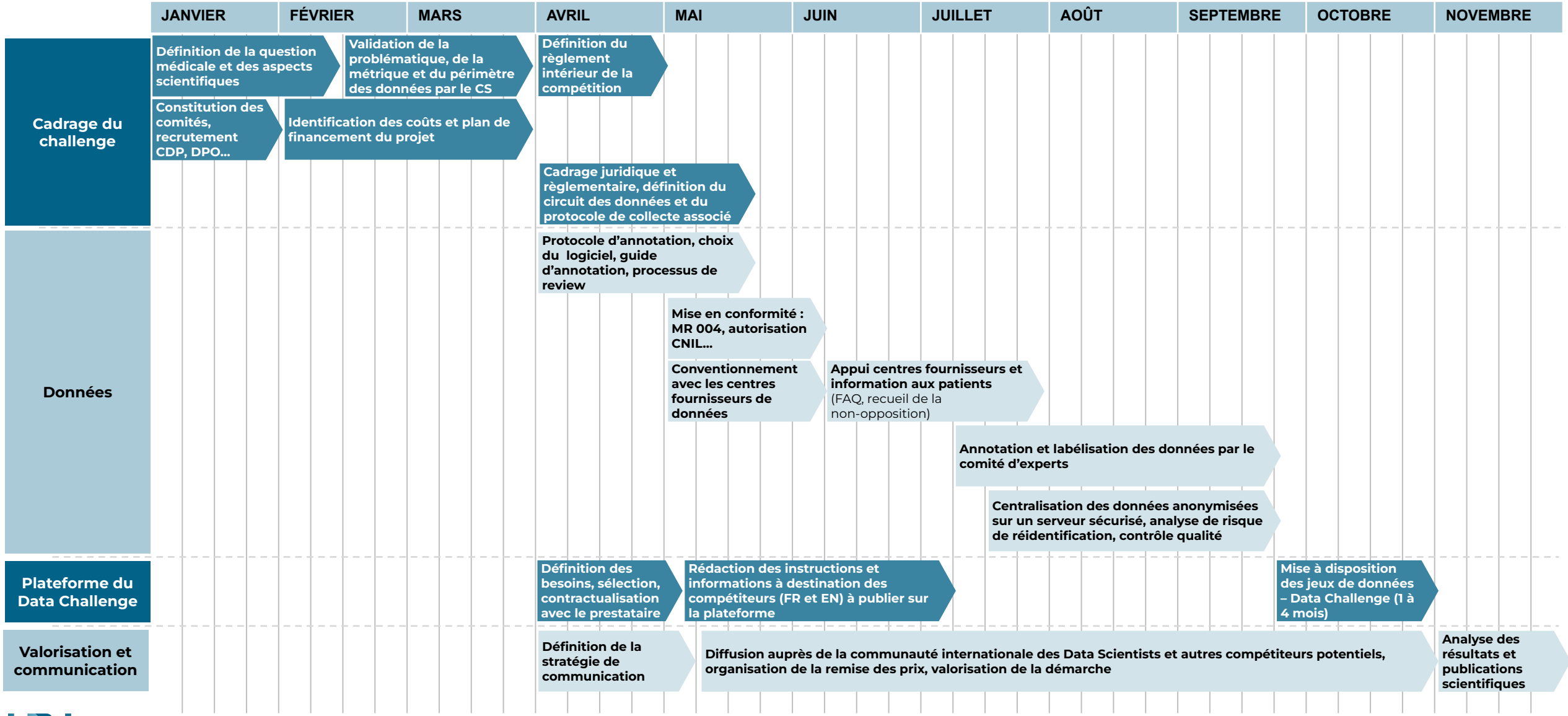
Membres

- Donneurs d'ordres des partenaires du Data Challenge
- Chef de projet Data Challenge
- Médecins / Cliniciens de la spécialité traitée
- Mathématiciens / Experts IA

Fréquence

- 1 fois tous les 2 mois

Exemple de rétroplanning



HEALTH DATA HUB

Chantier 1 – Cadrage méthodologique du Challenge

Cadrage méthodologique du Challenge

Le cadrage méthodologique du projet consiste à **planifier les différents chantiers** à mettre à en œuvre et les **ressources à mobiliser** pour y parvenir.
L'organisation d'un Data Challenge nécessite de constituer des équipes et comités **pluridisciplinaires** pour assurer la bonne gestion des aspects **scientifiques**, **techniques**, **juridiques**, **opérationnels** et **financiers** du projet.

- Equipe projet
- Comité d'organisation
- Conseil Scientifique

Constitution des comités et mise en place

Définition par le CS de la **question**, de sa traduction en **problème de classification supervisée**, du **périmètre des données** (critères d'inclusion, source, volumétrie…), **métrique d'évaluation**.

Définition de la question médicale

Définition de la démarche réglementaire à adopter

Cette étape et sa complexité dépendent de la **nature des données** et de la **nécessité ou non** de les **collecter** / **anonymiser**.

Définition du planning et cadrage opérationnel

Planification **opérationnelle** du projet et de la compétition.
Création d'un **rétroplanning**.
Définition du **règlement intérieur**, des **modalités pratiques**, du **prix** de récompense, de la **plateforme**.

Définition de la problématique, des données et des ressources

Principales activités

- Etablissement d'une **gouvernance** (constitution des conseils et comités)
- Identification des **enjeux médicaux et d'IA** dans la spécialité avec un 1er niveau d'identification de la problématique

Livrable clé

- *Synopsis projet Data Challenge*

- Description de **l'approche de classification supervisée** envisagée associée au challenge → Quelle sera la variable à prédire ?
- Définition de la nature, de la source, du format, de la volumétrie et des critères d'inclusion des **données** à mettre à disposition des compétiteurs pour obtenir des **résultats robustes** (nécessité de pouvoir obtenir des données en quantité suffisante et homogènes pour être capable d'y appliquer des approches de Machine Learning)
- Définition du processus **de labellisation et annotation** des données → Comment enrichir la base de données pour permettre une approche de classification supervisée ? (*ex : la SFP pour son challenge a annoté les lames de biopsie pour y indiquer les zones de lésion caractéristiques et a labellisé chaque échantillon avec la classe diagnostic correspondant à la lésion de classe la plus élevée présente sur la lame*)

Livrable clé

- *Protocole scientifique incluant définition de la problématique, des données et processus de labellisation et annotation*

Bonnes pratiques

- Définir le format des données mises à disposition (de préférence un format unique cible) de manière à faciliter les traitements des compétiteurs (ex. librairie open source disponible etc.)

Définition de la problématique, des données et des ressources

Principales activités

- Définition de la **métrique d'évaluation** nécessaire à l'évaluation de la performance des algorithmes produits par les compétiteurs et réalisation de simulations pour s'assurer de la pertinence de cette métrique → La métrique a pour objectif de pondérer les erreurs de prédiction des algorithmes pour sanctionner plus sévèrement une erreur qui aurait potentiellement un fort impact d'un point de vue clinique.

Livrables clé

- *Exemple de métrique personnalisée (i.e. distance entre valeur réelle et prédite pondérée par le poids de l'erreur)*
- *Exemple de simulation de score sur la base de différentes stratégies de jeu*

- Qualification et évaluation des **besoins humains et financiers** à mettre à disposition pour le data challenge (Chef de projet, délégué à la protection des données, Data scientist…)

Livrables clé

- *Fiche de poste CDP Data Challenge*
- *Annexe financière – Dossier de demande de financement*

Bonnes pratiques

- Mettre en place des sessions de travail entre mathématiciens, statisticiens et médecins afin de définir la métrique la plus adéquate
- La métrique d'évaluation pourra également être mise en place avec l'aide de certains prestataires d'hébergement du challenge qui proposent cette prestation

Points d'attention

- Anticiper au maximum le besoin en ressources humaines et financières (notamment si il est question de recruter en externe)

Cadrage juridique et règlementaire du projet

La démarche règlementaire à adopter dans le cadre du projet dépend du **type de données** que l'on souhaite utiliser et de **la façon dont elles seront obtenues**. L'objectif final étant d'obtenir une **base de données anonymisées** dont l'anonymisation est sécurisée par une **analyse de risque de réidentification**.
Cette condition est nécessaire pour que la base de données puisse être partagée aux compétiteurs dans le cadre du Challenge.

Plusieurs cas de figure…

La base de données anonymes existe → Les données sont déjà réunies et anonymisées

Production d'une **analyse de risque de réidentification** pour s'assurer du caractère anonyme des données.

La base de données existe mais n'est pas anonymisée → Les données sont réunies mais il faut les anonymiser

Application de la **loi informatique et libertés** et du **RGPD** pour anonymiser la base puis production d'une **analyse de risque de réidentification** pour s'assurer du caractère anonyme des données.

La base de données n'existe pas → Les données sont à collecter et à anonymiser

Établissement d'un protocole de collecte des données permettant de **sécuriser le flux des données**, application de la **loi informatique et libertés** et du **RGPD** pour anonymiser la base puis production d'une **analyse de risque de réidentification** pour s'assurer du caractère anonyme des données.

Cadrage juridique et règlementaire du projet

Principales activités

1er Cas de figure : La base de données anonymes existe

- Production d'une **analyse de risque de réidentification**

Documentation

- *Guide - Analyse de risque de réidentification*

2ème Cas de figure : La base de données existe mais n'est pas anonymisée

- Définition du **processus d'anonymisation des données** (ex. dé-identification de lames à la numérisation grâce à des étiquettes d'anonymisation) et identification de la **procédure réglementaire adéquate** pour traiter les données personnelles (ex. MR004).

Documentation

- *Guide - Anonymisation des données*
- *Guide - Procédures d'accès aux données de santé en France*

- **Informations aux patients** : rédaction de la lettre d'informations aux patients et mise en place du processus d'envoi des lettres d'informations.

Documentation

- *Guide - Note d'information aux patients*

- Production d'une **analyse de risque de réidentification**

Bonnes pratiques

- Impliquer un DPO interne ou un prestataire dès les premières sessions de cadrage en incluant les équipes « métiers » (i.e. médecins) afin de bien comprendre les enjeux réglementaires du challenge
- L'analyse de risque de réidentification est un document transverse qui doit contenir des éléments statistiques et réglementaires. Sa rédaction doit faire intervenir le DPO et le Data Scientist du projet.
- Cette analyse doit s'appuyer sur les critères de l'avis du G29 et permet de démontrer que la base de données ne permet pas de remonter à l'identité des patients.

Cadrage juridique et règlementaire du projet

Principales activités

3ème Cas de figure : La base de données n'existe pas

- Rédaction du **protocole de collecte des données** depuis leur origine jusqu'à leur mise à disposition des compétiteurs
- Définition du **processus d'anonymisation des données** et identification de la **procédure réglementaire adéquate** pour traiter les données personnelles
- **Conventionnement** avec les centres dépositaires de données et rédaction d'une **notice d'information**, mise en place de FAQ et webinaires pour encadrer la collecte
- **Informations aux patients** : rédaction de la lettre d'informations aux patients et mise en place du processus d'envoi des lettres d'informations le cas échéant :
  - ✔ Option 1 (à privilégier) – Envoi décentralisé : chaque centre se charge de l'information aux patients (via l'envoi des lettres d'informations sur son périmètre de patients ou bien via leurs propres canaux de communication)
  - ✔ Option 2 – Envoi centralisé : implique la collecte des coordonnées patients, sélection d'un prestataire et contractualisation (approx. 3000€)
- Production d'une **analyse de risque de réidentification**

Bonnes pratiques

- Les centres fournisseurs de données sont chargés d'anonymiser les données avant de les transmettre au porteur de projet

Définition du planning et cadrage opérationnel

Une fois le projet cadré sur les plans scientifiques et réglementaires, il est nécessaire de planifier toutes les étapes de réalisation du projet d'un point de vue **opérationnel**.
Cette planification permettra notamment de décrire les phases de **collecte de données** et **d'annotation** (si existantes).
Cette étape consiste également à **cadrer la compétition** en elle-même : définition du calendrier, des besoins concernant la plateforme, des prix de récompense et du règlement intérieur de la compétition.

Principales activités

- Création d'un **rétroplanning** indiquant les étapes et tâches à réaliser et les responsables correspondants.

Livrable clé

  - Rétroplanning

- Définition du **calendrier de la compétition** détaillant le temps alloué à chaque phase
- Définition des besoins en termes de **plateforme d'hébergement** du challenge (nécessité ou non de mettre à disposition des compétiteurs un espace de calcul, cible visée…)
- Définition des **modalités de remise des prix** (par exemple la SFP avait décidé de ne remettre le prix de récompense qu'à condition que les lauréats acceptent de partager leur algorithme en Open Source)
- Définition du **règlement intérieur de la compétition**, ce règlement précise notamment les conditions pour pouvoir y participer et les modalités de partage des algorithmes

Bonnes pratiques

- Le calendrier de la compétition, les modalités de remise des prix et le règlement intérieur sont à définir en collaboration avec le prestataire d'hébergement de la compétition.
- Il est à noter que certaines règles peuvent être imposées en fonction de la plateforme choisie

C

#4 Chantier 2 - Données

Constitution de la base de données

L'ampleur du chantier des données dépendra du **type de données** que l'on souhaite utiliser pour réaliser le Challenge. Soit la base anonyme existe déjà, soit elle existe mais il faut l'anonymiser, soit elle n'existe pas et il faut organiser sa constitution (cf. page 18). Pour ce dernier cas de figure, c'est par exemple ce que la SFP a fait en organisant un **appel à manifestation d'intérêt** auprès de **centres d'anatomie et cytologie pathologiques** afin d'en collecter les données. La remontée des données des centres volontaires était encadrée par voie de **convention** et de plusieurs **webinaires** ainsi qu'une **FAQ**. L'enchaînement des étapes dans cet exemple est restitué ci-après.

Sélection des cas d'intérêt pour le Data Challenge

Les médecins déterminent les données pertinentes et nécessaires à collecter pour le challenge

Information des patients sélectionnées

Une notification des patients sélectionnés est faite (par courrier ou par mail) et un délai d'un mois est prévu pour l'exercice des droits

Anonymisation des données sélectionnées après la fin de la période d'1 mois

Les équipes des centres fournisseurs de données procèdent à l'anonymisation des données collectées

Illustration - étiquette d'anonymisation de lame d'anatomopathologie

Numérisation des données

Les données sélectionnées et anonymisées sont numérisées au sein des centres fournisseurs de données

Transfert des données sur un serveur sécurisé

Les données anonymisées sont envoyées sur un serveur sécurisé pour préparation en vue du Data Challenge

Chantier des données

Principales activités

- Mise en œuvre de la **démarche réglementaire** définie lors du cadrage (dépendante de la source et de la nature anonyme ou non des données).

Dans le cas où une étape de collecte des données est nécessaire…

- Lancement d'un **appel à manifestation d'intérêt** auprès des centres contributeurs à la constitution du jeu de données (par exemple)
  - ✔ AMI (incl. déclaration de mise en conformité et description du challenge)
  - ✔ Liste des centres fournisseurs de données
- Accompagnement des centres fournisseurs de données pour la **collecte des données** incluant, **l'information aux patients**, **l'anonymisation** et la **numérisation** (le cas échéant).
  - ✔ **Guidelines** pour la sélection des cas d'intérêts (ex. typologies de patients, pathologies, ventilation des cas entre les classes, modalités de numérisation)
  - ✔ Mise en place d'une **FAQ** et de sessions de **webinaires avec Q&A**
  - ✔ Déplacement (si besoin) in-situ à prévoir
- Création d'un **tableau de suivi de la collecte** des données

Livrable clé

- Tableau de suivi de la collecte des données

Bonnes pratiques

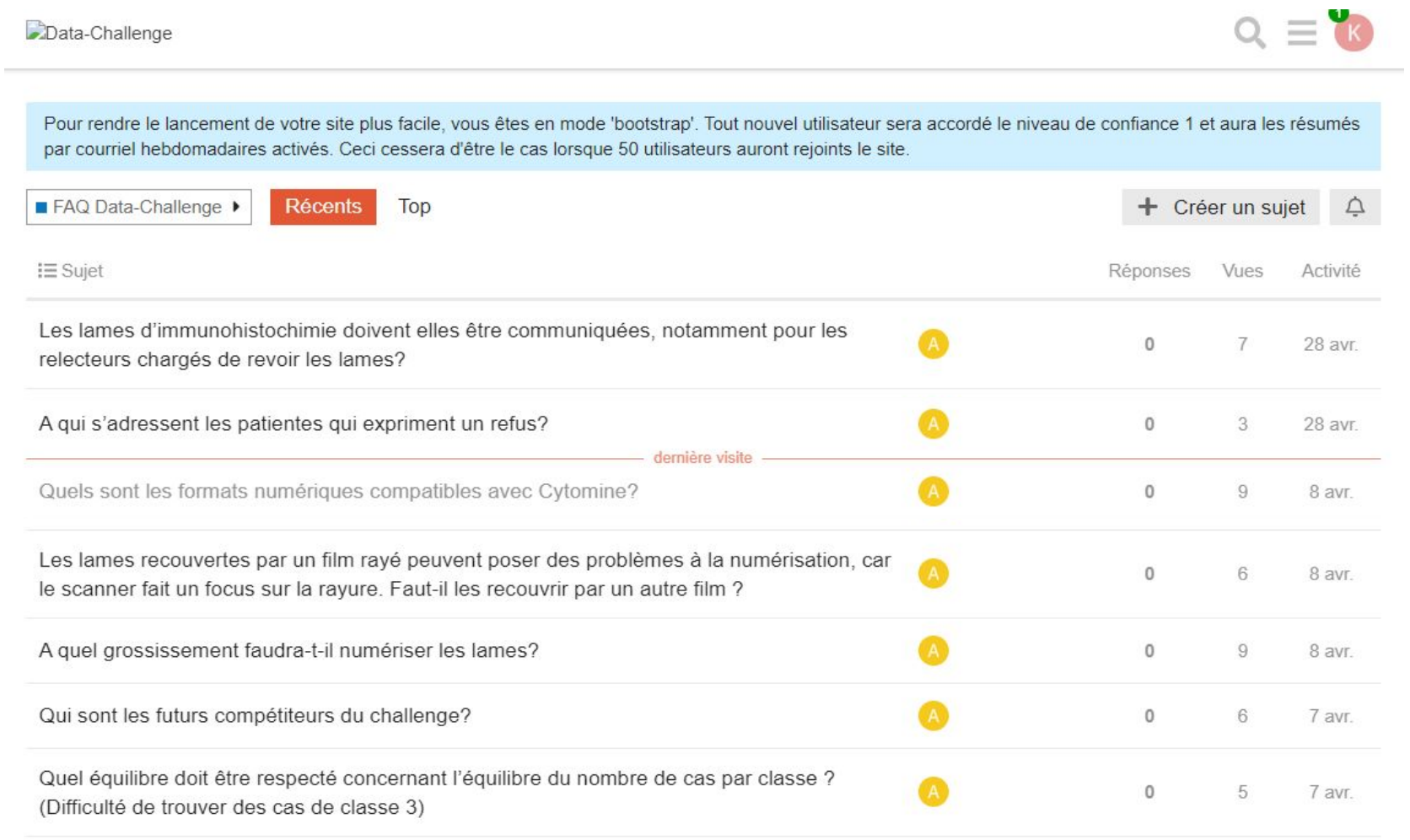
- La phase de collecte des données (si elle est nécessaire) doit être sécurisée sur les plans opérationnel et règlementaire par le DPO du projet.
- L'objectif est de constituer un circuit des données garantissant une collecte de données anonymisées et de qualité dans des formats homogènes via un flux sécurisé d'un point de vue règlementaire.

Points d'attention

- La gestion de la collecte par les centres fournisseurs de données demande énormément de rigueur et des échanges très fréquents avec les centres (en général de vive voix)

Chantier des données

Dans le cas où une étape de collecte des données est nécessaire…



Exemple de FAQ à destination des centres fournisseurs de données

La mise en place de **webinaires** et d'une **FAQ** permet aux centres de se former sur la procédure à suivre pour extraire, informer les patients, anonymiser et faire remonter les données.

Ingestion des données et stockage

Lors de l'étape de collecte des données (si elle a lieu), chaque centre sélectionne les cas d'intérêt en fonction des guidelines, informe les patients concernés et anonymise les données des patients n'ayant pas exprimé leur opposition.
Une fois ces étapes effectuées, chaque centre devra faire **remonter ses données** pour qu'elles puissent être **centralisées** et préparées pour la **phase d'annotation et de labellisation**.

Principales activités

- **Préparation des données :** Le processus de préparation des données, doit garantir la confidentialité de ces dernières surtout si les données transitent sur des postes de travail avant leur anonymisation. La sécurité des données dans les phases de préparation est à la charge du producteur.
- **Ingestion** des données anonymisées vers un espace de stockage
  - ✔ Transfert des données via un canal sécurisé SFTP → Création d'un guide à destination des centres (le cas échéant) pour décrire les étapes à suivre pour effectuer le transfert.
  - ✔ Si des difficultés sont rencontrées dans l'utilisation du canal de transfert sécurisé SFTP, un autre mode de transfert peut être envisagé mais la dérogation doit être validée par les RSSI des parties prenantes.
- **Stockage** des données
  - ✔ Estimation du volume / taille des données à stocker pour dimensionner l'infrastructure (i.e. serveur)
  - ✔ Mise en place de l'infrastructure de stockage
  - ✔ Déversement des données au sein des serveurs (étape chronophage à anticiper)

Bonnes pratiques

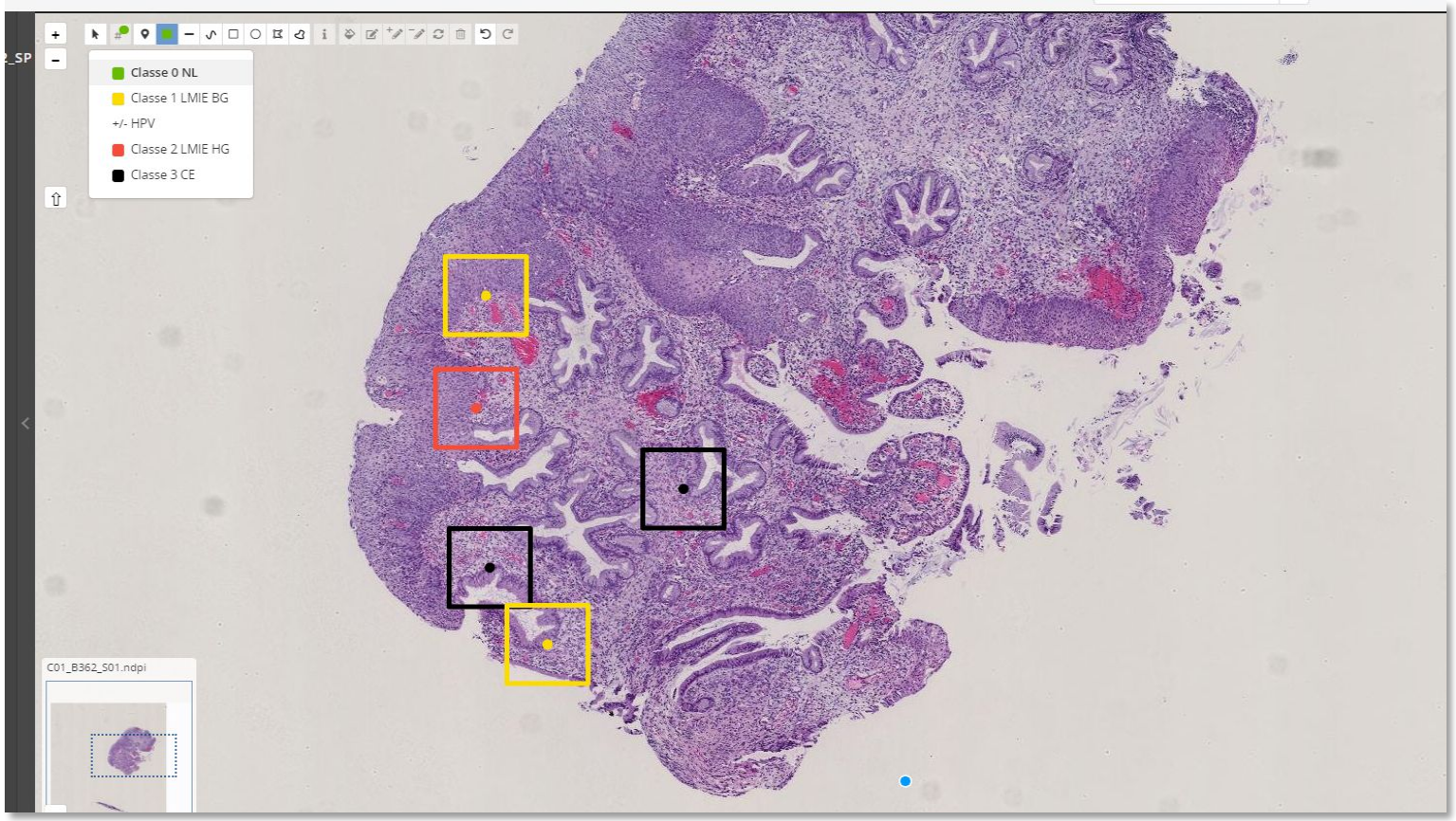
- Intervention d'un Data Scientist / Data Engineer afin d'absorber la charge de travail pendant ces étapes

Points d'attention

- Respecter les règles de sécurité, garantir notamment leur confidentialité lors des phases de préparation et d'ingestion des données.
- Anticiper le temps d'ingestion et de dépôt des données sur les serveurs (ex. peut prendre plusieurs heures par batch de 500 images)

Annotation et labellisation des données

Une fois que les données anonymisées sont centralisées, l'étape **d'annotation** et de **labellisation** peut être initiée.
Cette étape consiste à enrichir la base de données pour permettre l'étape d'apprentissage lors du développement des algorithmes.
Par exemple, dans le cas du Challenge de la SFP, l'étape **d'annotation** consistait à indiquer sur les lames numérisées, les **zones de lésion caractéristiques**. La **labellisation** consistait à indiquer pour chaque lame, la **classe la plus haute détectée** parmi les lésions identifiées.



Exemple de lame annotée dans le cadre du Challenge de la SFP

- Dans cet exemple, la lame comporte plusieurs annotations (en jeune, rouge et noir) indiquant les zones de lésion caractéristiques et leur classe.
- La labellisation de cette lame correspondra donc à la « Classe 3 » qui est la classe la plus grave détectée.

Nb : sur cette illustration les annotations ont été placées au hasard

Annotation et labellisation des données

L'étape d'annotation et de labellisation est à la charge du **comité d'annotation** qui est composé **d'experts** spécialisés dans la thématique médicale du Challenge.
Pour le Challenge de la SFP, des équipes de **médecins pathologistes** se sont chargées d'annoter et de labelliser chaque lame de la base de données à l'aide d'un **outil d'annotation**.

Principales activités

- Sélection de **l'outil d'annotation** (plusieurs outils d'annotation en Open Source sont disponibles)
- Identification des **développements** si nécessaire (si aucun outil disponible en Open Source ne convient aux besoins)
- **Installation de l'outil** sur le serveur de stockage des données
- Mise en place d'un **comité d'annotateurs experts** dans la pathologie / problématique du challenge et établissement d'un protocole d'annotation (mise en place d'un processus de double annotation de chaque cas par deux spécialistes différents par exemple)
- **Accompagnement du comité d'annotation** dans la phase de labellisation et annotation (ex. formation individuelle, création d'un guide d'utilisation)

Livrable clé

- Guide d'utilisation de l'outil d'annotation

Bonnes pratiques

- Ne pas hésiter à accompagner les annotateurs in-situ pour la prise en main de l'outil d'annotation
- Intervention d'un Data Scientist dans l'élaboration de la stratégie d'annotation

Points d'attention

- Anticiper dans le rétroplanning cette étape qui est l'une des plus chronophage et sensible durant la préparation du challenge

Chantier 3 - Plateforme

La plateforme d'hébergement du challenge

Une fois la **base de données finalisée** (annotation, labellisation, analyse de risque de réidentification, différentiation en trois sous bases homogènes), elle peut être transmise à la plateforme d'hébergement du Challenge pour permettre l'organisation de la compétition.
Le Data Challenge se déroule **en ligne** et repose sur **différents jeux de données** mis à disposition des compétiteurs (étudiants, chercheurs, Data Scientists...) par les organisateurs via une **plateforme dédiée** -la plus connue étant Kaggle-.

La plateforme de Data Challenge permet aux participants de...

COMPRENDRE LA PROBLÉMATIQUE ET LES RÈGLES DU CHALLENGE

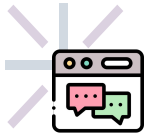
... grâce à une introduction **pédagogique des enjeux de la problématique exposée**, une **formulation de la question** ainsi que les **métriques d'évaluation** associées et enfin une présentation du **calendrier** du challenge et du **règlement intérieur**

ACCÉDER AUX JEUX DE DONNÉES MIS À DISPOSITION ET À UN ESPACE DE CALCUL

... que ce soit le jeu de données **d'apprentissage** et de **test pendant** la **phase d'apprentissage** ou le jeu de données **de test final** (chiffré). Les compétiteurs auront aussi la possibilité d'utiliser la **puissance de calcul** de la plateforme pour entraîner leur algorithme.

SE MESURER AUX AUTRES PARTICIPANTS EN AYANT ACCÈS À LEUR CLASSEMENT

... en **soumettant leurs algorithmes**, qu'ils soient intermédiaires ou finaux, afin de visualiser leur classement dans un « **Leaderboard** » en temps réel



ÉCHANGER AVEC LES AUTRES PARTICIPANTS

... grâce à des **espaces collaboratifs** (ex. FAQ, Wiki, GitHub, accès aux notebooks des participants)

Illustration des principales composantes d'une plateforme

Acculturation des participants à la problématique

TissueNet: Detect Lesions in Cervical Biopsies

HOSTED BY FRENCH SOCIETY OF PATHOLOGY

RESULTS HOME PROBLEM DESCRIPTION DATA RESOURCES SUBMISSION FORMAT ABOUT

Problem description

The data for this challenge includes thousands of microscopic slides of uterine cervical tissue from medical centers across France. Your objective is to classify each image according to the most severe category of epithelial lesion present in the sample. The classes are defined as follows:

- `0` : benign (normal or subnormal)
- `1` : low malignant potential (low grade squamous intraepithelial lesion)
- `2` : high malignant potential (high grade squamous intraepithelial lesion)
- `3` : invasive cancer (invasive squamous carcinoma)

Dataset

Images
Annotations
Metadata
Labels
Data example
Test set

Performance metric

Custom metric

Submission Format

Code submission

On retrouve sur la plateforme de la **documentation** dont l'objectif est de définir les **concepts médicaux** à un public non initié et d'expliquer la problématique proposée.

A biopsy is a sample of tissue examined at a microscopic level to diagnose cancer or signs of pre-cancer. Digital pathology has developed considerably over the past decade as it has become possible to work with digitized "whole slide images" (WSIs). These heavy image files contain all the information required to diagnose lesions as malignant or benign, yet present huge challenges to use effectively.

This challenge focused on epithelial lesions of the uterine cervix, and featured a unique collection of thousands expert-labeled WSIs collected from medical centers across France. This is a sizable dataset (700GB) of extremely high resolution images. Given the scale of the dataset, handling the data efficiently is a critical problem to solve in the process of developing an accurate approach to diagnosis.

Here are examples of annotated regions at full resolution:

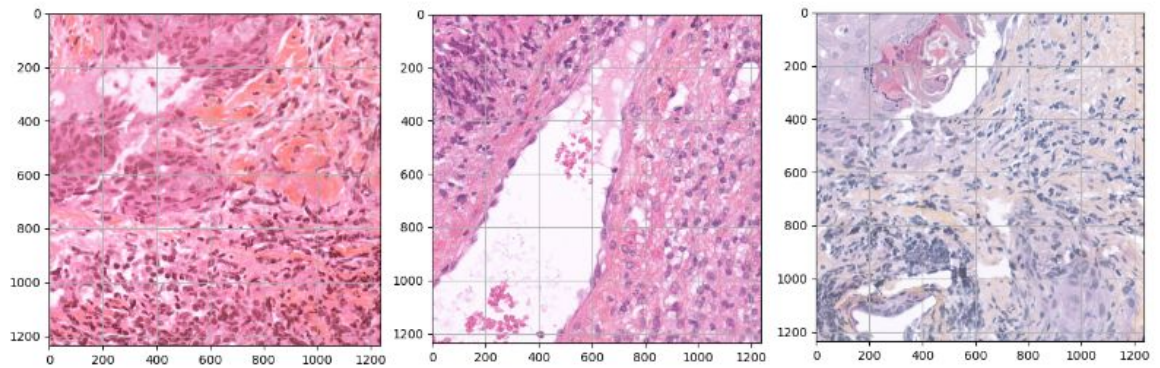


Illustration des principales composantes d'une plateforme

Mise à disposition d'informations pratiques

What to submit

Your final submission should be a zip archive named with the extension `.zip` (for example, `submission.zip`). The root level of the `submission.zip` file must contain a `main.py` which performs inference on the test images and writes the predictions to a file named `submission.csv` in the same directory as `main.py`. You can see an example of this submission setup in the runtime repository.

Here's an example:

```
submission root directory
├── assets          # Example of configuration and weights for the trained model
│   ├── model.json
│   └── weights.h5
└── main.py         # Inference script
```

Son également décrits sur la plateforme, les différents **jeux de données**, la **métrique d'évaluation** ainsi que le **format de soumission** des algorithmes qui est attendu.

Les **règles de la compétition** sont indiquées dans le règlement intérieur qui précise notamment les conditions pour participer, les modalités de remise des prix, les règles de constitution des équipes, le nombre de soumissions autorisé…

Competition Rules

GUIDELINES

One account per participant

You cannot sign up to DrivenData from multiple accounts and therefore you cannot submit from multiple accounts.

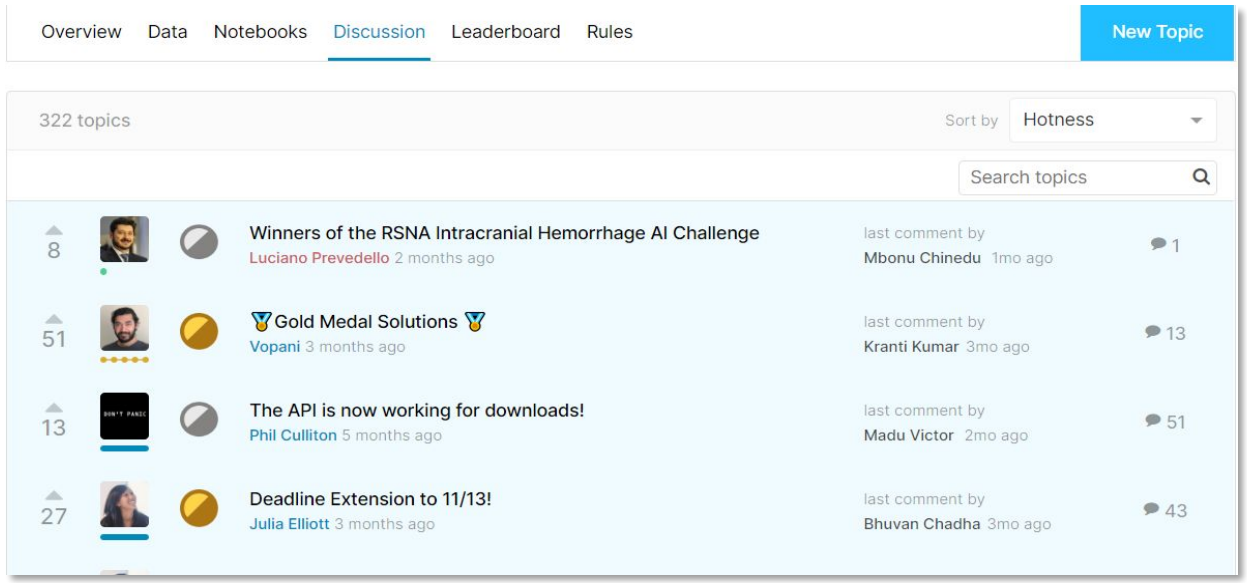
Private sharing of code

Privately sharing code or data outside of teams is not permitted.

Winner License Type: Open Source License

Illustration des principales composantes d'une plateforme

Mise à disposition d'espaces collaboratifs et d'échange entre les participants



Les participants peuvent échanger et interagir à propos du Challange sur un **forum** mis à disposition.

Cet espace permet également aux organisateurs de répondre aux éventuelles interrogations des compétiteurs

Les compétiteurs peuvent partager leurs travaux s'ils le souhaitent sous forme de **notebooks** via un espace dédié sur la plateforme.

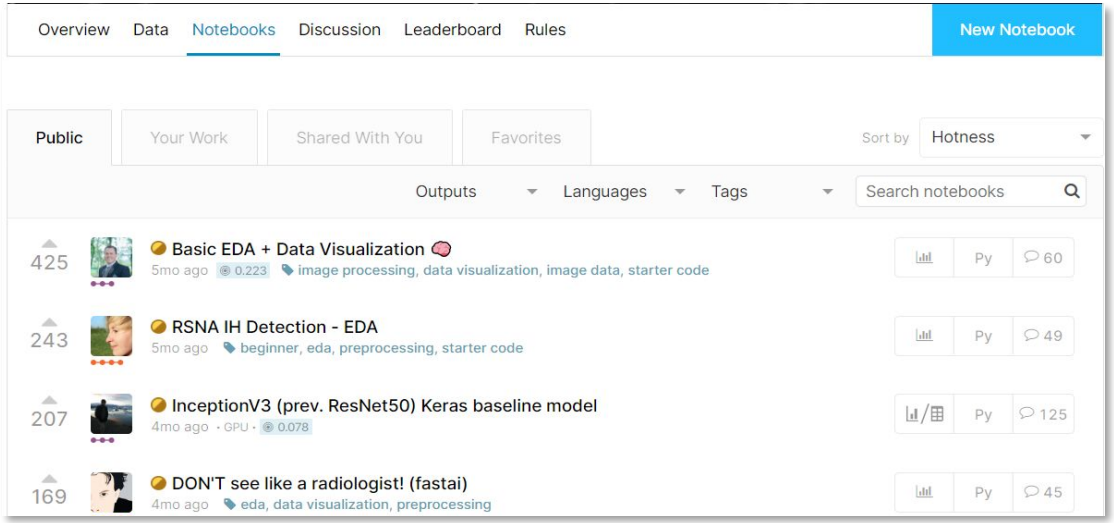
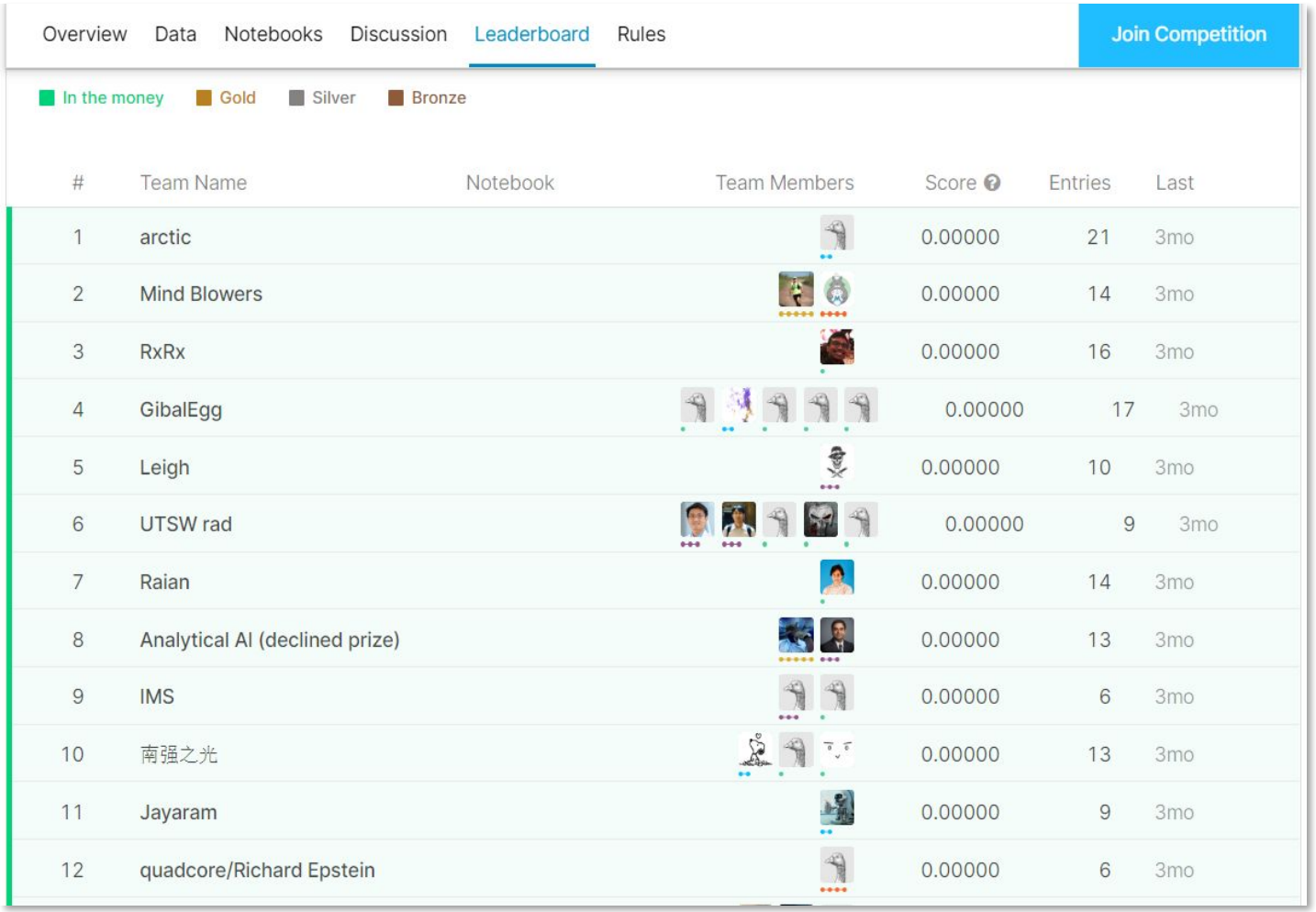


Illustration des principales composantes d'une plateforme

Classement en temps réel des compétiteurs : le Leaderboard



#	Team Name	Notebook	Team Members	Score	Entries	Last
1	arctic			0.00000	21	3mo
2	Mind Blowers			0.00000	14	3mo
3	RxRx			0.00000	16	3mo
4	GibalEgg			0.00000	17	3mo
5	Leigh			0.00000	10	3mo
6	UTSW rad			0.00000	9	3mo
7	Raian			0.00000	14	3mo
8	Analytical AI (declined prize)			0.00000	13	3mo
9	IMS			0.00000	6	3mo
10	南强之光			0.00000	13	3mo
11	Jayaram			0.00000	9	3mo
12	quadcore/Richard Epstein			0.00000	6	3mo

Le **Leaderboard en temps réel** permet aux compétiteurs d'évaluer la performance de leurs algorithmes pendant la phase de test.

Le nombre de soumissions par jour est généralement limité.

Le choix de la plateforme

Le choix de la plateforme d'hébergement du Challenge doit être fait en fonction notamment du **budget** que l'on souhaite allouer à la prestation, de la **portée** de la compétition et des **besoins** en termes de serveurs et de puissance de calcul.
Du côté du prestataire, les exigences peuvent également varier. Les plateformes les plus connues sont très sélectives vis-à-vis des compétitions qu'elles hébergent.

Zoom sur trois grandes catégories de plateformes…

	Plateforme référence Kaggle	Plateforme spécialisée par thématique	Plateforme développée à façon (via librairie open source ou via un prestataire)
Avantages	• **Solution clé en main** : accompagnement de bout-en-bout depuis le pré-traitement des données jusqu'à la remise des prix en passant par la mise en place d'une plateforme dédiée • **Rayonnement maximal** pour l'événement (étant donné que Kaggle est la plateforme la plus connue)	• **Solution clé en main** -de type Kaggle- Conventionnement possible sans fournir tout le jeu de données • Possibilité de capitaliser sur leurs réseaux existants pour **élargir le cercle des participants** • Plateformes **spécialisées** sur des sujets à forts impacts sociétaux	• Solution **custom** • **Prix attractif**
Inconvénients	• Processus **très sélectif** (ex. 5 challenges orientés recherche sont acceptés par an) • Mise à disposition des données **au moins 8 semaines avant** le Data Challenge est une condition sine qua non au conventionnement avec Kaggle	• **Rayonnement moindre** par rapport à Kaggle	• Rayonnement **faible** • Quid du **devenir de la plateforme** post challenge (ex. frais de maintenance)
Coût	• +100k€ (incl. le prix à destination des gagnants)	• 50-75k€ (incl. le prix à destination des gagnants)	• 25k€ (incl. le prix à destination des gagnants)

kaggle

DRIVENDATA TUNEDIT CrowdANALYTIX Analytics Vidhya Innocentive Trustii.io ZiND! CodaLab crowdAI

Chantier plateforme d'hébergement du Challenge

Principales activités

- **Choix** de la plateforme et **contractualisation** (prévoir échéancier à 3 ou 4 versements)
- **Rédaction de la documentation** à publier sur la plateforme :
  - ✔ Règlement intérieur
  - ✔ Description de la problématique et des jeux de données
  - ✔ Description des résultats attendus
- **Rédaction de la « Landing page »** : la landing page est une page sur laquelle les participants peuvent se rendre pour avoir des informations sur le challenge avant son ouverture et permettant une pré-inscription
- Pendant le Challenge : suivi des **activités sur les espaces collaboratifs**, en particulier le forum pour pouvoir répondre aux questions des participants

Bonnes pratiques

- Prévoir dans le contrat la date (ou les dates) de mise à disposition des données à la plateforme (une fois labellisées et annotées).
- Préciser à la contractualisation la mise en place d'une landing page afin de pouvoir autoriser les pré-inscriptions en amont du lancement du challenge.

Points d'attention

- Le coût de la prestation est fortement lié à la **volumétrie** des données.

D

Chantier 4 – Communication et valorisation du Challenge

Stratégie de communication autour du Challenge

La réussite du Data Challenge repose sur le recrutement de **nombreux participants** à la compétition.
Il est donc important d'établir une **stratégie et des supports de communication** pertinents pour pouvoir **attirer un maximum de compétiteurs** et obtenir des **scores de performance élevés**.

Différentes phases de communication…

Avant et pendant l'ouverture de la compétition

La stratégie de communication dépendra de la **cible** à laquelle on souhaite ouvrir la compétition.
Les compétitions les plus ouvertes s'adressent à des participants au niveau **international** de **tous horizons** (chercheurs, étudiants, industriels…).
Il faut donc multiplier les formats et les plateformes de communication pour atteindre tout type de compétiteur potentiel.

Communication autour des résultats et valorisation des lauréats

A la fin du Challenge, les participants en tête du leaderboard sont récompensés.
Une cérémonie de **remise des prix** peut être organisée ainsi que des **vidéos** et **articles** à ce sujet afin de mettre en avant les résultats et les lauréats.

Valorisation de la démarche Data Challenge en elle même

La phase post Data Challenge consiste à poursuivre les travaux d'analyse des résultats et à **publier des articles** autour de ces analyses.
Des supports de communication de type « **retour d'expérience** » peuvent également être produits pour **valoriser l'initiative** et son format innovant et participatif, notamment auprès de la communauté médicale.

Stratégie de communication autour du Challenge

Principales activités

- Mise en place d'une **landing page** (en FR / EN) pour communiquer **en amont** sur le challenge et permettre les pré-inscriptions
- Identification et communication auprès de **potentiels compétiteurs** via :
  - ✔ TOP 20/30 des challenges similaires (à collecter via les leaderboard publics)
  - ✔ Mise en relation d'industriels et médecins pour la constitution d'équipes pluridisciplinaires
- Identification et communication auprès de **potentiels relayeurs**
- Plan de communication générale via les **réseaux sociaux** (ex. twitter, linkedin…)
- Production d'une **présentation commune** des partenaires du challenge
- Production d'une **vidéo témoignage** / REX challenge : exemple de la vidéo de la Société Française de Pathologie
- Valorisation des partenaires lors **d'évènements** (ex. annonce des lauréats pendant un congrès annuel)
- Valorisation des lauréats dans le cadre **d'interviews** publiées sous forme d'articles : exemple d'interview de lauréat du Challenge de la SFP
- Rédaction d'articles valorisant la démarche : exemple d'article sur le Challenge de la SFP
- **Mise en Open Data** des données du challenge et en **Open Source** des algorithmes des lauréats (incl. cadre réglementaire, modalités opérationnelles et techniques de mise à disposition)

Bonnes pratiques

- Contenu à créer pour les différentes landing page (i.e. FR et EN)
- Etablir un plan de communication en 5 phases :
  - ✔ Mise en ligne de la landing page,
  - ✔ A J-7/2 du challenge,
  - ✔ A l'ouverture du challenge
  - ✔ A J+20 du début du challenge
  - ✔ A l'annonce des lauréats
- La mise en Open Data et en Open Source des données et algorithmes peut se faire suivant différentes modalités infrastructurelles et sous différentes licences.