

Appel à projets



DATA CHALLENGES EN SANTÉ



AAP destiné aux acteurs de l'écosystème de la santé désireux d'organiser un projet Data Challenge



Le HDH est une institution promouvant **l'innovation dans l'utilisation des données de santé**

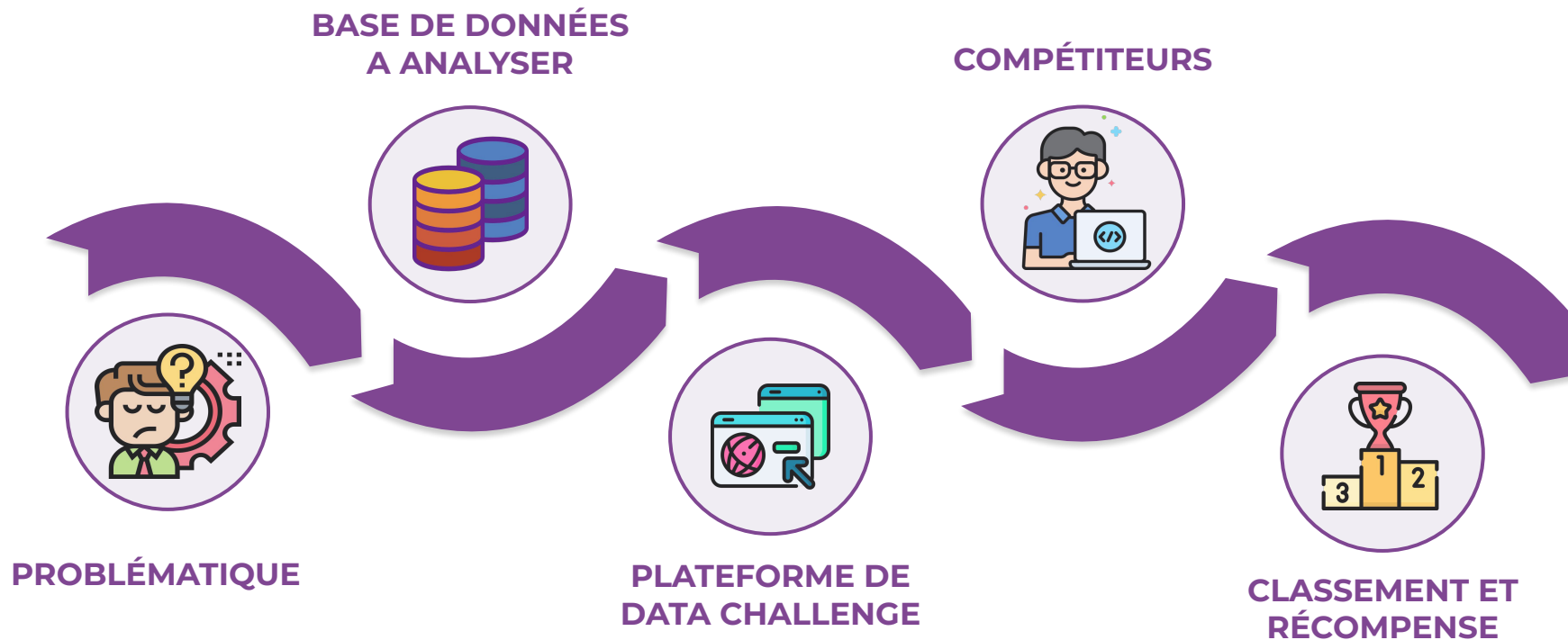
Dans le cadre de sa mission d'animation de l'écosystème des données de santé, le **HDH en partenariat avec la Délégation au Numérique en Santé et Bpifrance, soutient et organise des Data Challenges en santé**

Ces initiatives font partie du programme **Grand Défi** et s'inscrivent dans la **Stratégie d'Accélération Santé Numérique**

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 solutions **d'apprentissage supervisé**. 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-.



UNE PROBLÉMATIQUE...

A laquelle on peut répondre par une solution d'apprentissage supervisé

UNE BASE DE DONNÉES...

Pour servir de source d'information à l'analyse et permettre le développement d'un algorithme

UNE PLATEFORME...

Pour héberger la compétition et recruter des participants du monde entier

DES COMPÉTITEURS...

Pour participer au challenge et tenter de remporter un prix

UN CLASSEMENT...

Pour identifier les compétiteurs ayant programmé le meilleur algorithme

Qu'est-ce qu'un Data Challenge ?

Un Data Challenge est une compétition de data science se déroulant en ligne sur une **plateforme dédiée**

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



COMPRENDRE LA PROBLÉMATIQUE ET LES RÈGLES DU CHALLENGE

... grâce à une introduction **pédagogique des enjeux et de la problématique**, 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

... plusieurs **jeux de données** sont mis à disposition des compétiteurs pour qu'ils puissent les **analyser** et **développer leur algorithme**. Les participants 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 de Data Challenge...

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

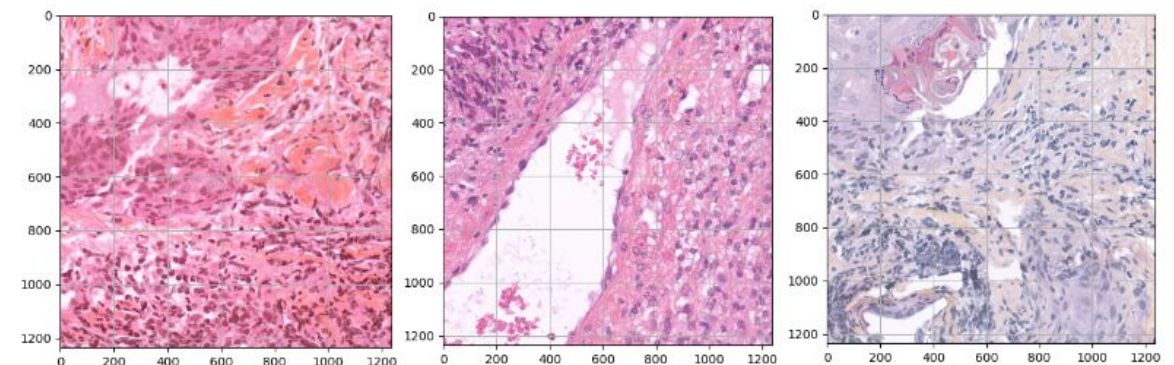
[Code submission](#)

On retrouve sur la plateforme de la **documentation** dont l'objectif est de définir les **concepts spécifiques de la thématique** 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:



Comment se présente un Data Challenge ?



Illustration des principales composantes d'une plateforme de Data Challenge...

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
```

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...

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

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.

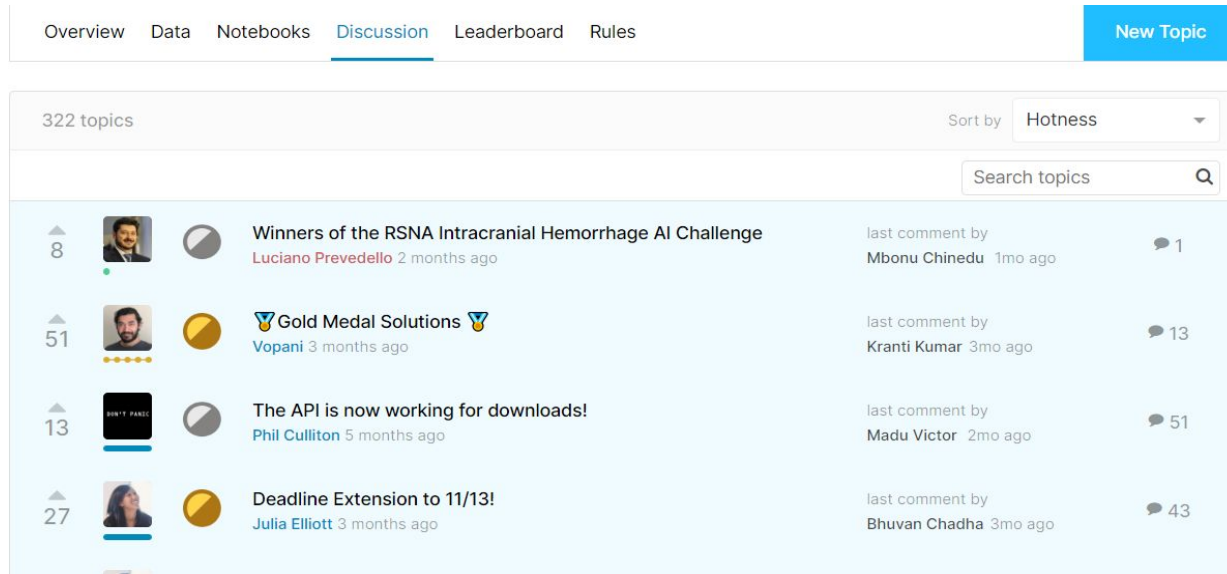
Winner License Type: Open Source License

Comment se présente un Data Challenge ?



Illustration des principales composantes d'une plateforme de Data Challenge...

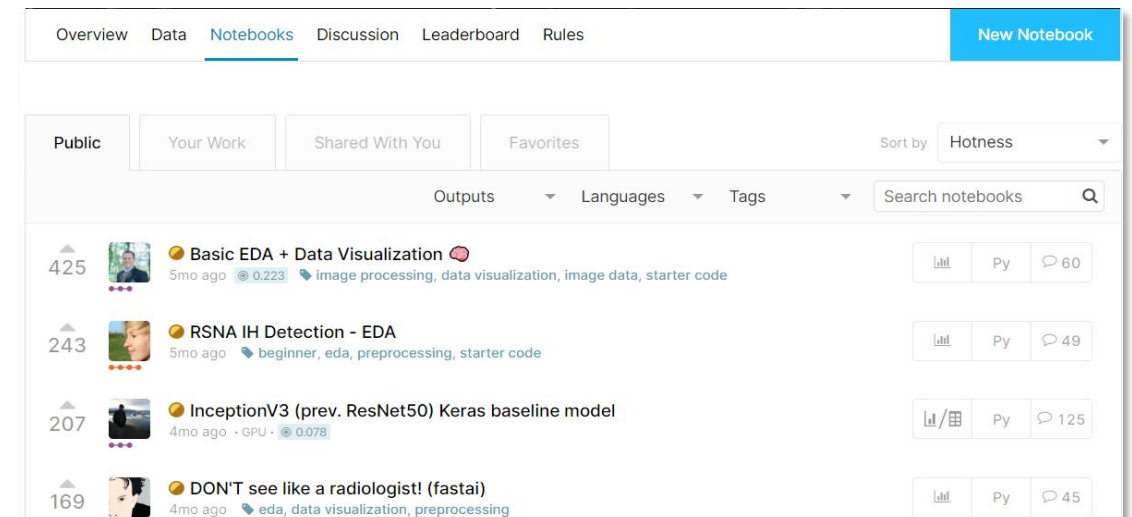
Des espaces collaboratifs et d'échange entre les participants



Les participants peuvent échanger et interagir à propos du Challenge 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.



Comment se présente un Data Challenge ?

Illustration des principales composantes d'une plateforme de Data Challenge...

Un classement en temps réel dit « **Leaderboard** » pour se mesurer aux autres compétiteurs

User or team		Best private 1?	Weighted Class Score 1	Timestamp 1	# 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 
	Lifels2Short	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

Le **Leaderboard en temps réel** permet aux compétiteurs d'évaluer la performance de leurs algorithmes et de la **comparer à celle des autres challengers**.

Les participants peuvent **évaluer des versions intermédiaires** de leurs algorithmes pendant une **phase de test** avant de soumettre une version finale qui déterminera **le classement définitif des compétiteurs**.

Le **score de performance** des algorithmes qui permet de classer les participants est déterminé à l'aide d'une **métrique d'évaluation**.

Cette métrique est une formule mathématique permettant de **pondérer les erreurs commises par l'algorithme**.

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

Déroulement de la compétition

Les grandes phases de la compétition hébergée par la plateforme...

Temps 1



Phase d'apprentissage

Quelques semaines ou mois

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.

Temps 2



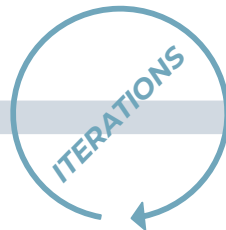
Phase de test final

Entre 48h et 72h

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 3^{ème} jeu de données permettant d'obtenir le score final de performance.

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 et permettre son amélioration.
La plateforme indique le score de performance de l'algorithme à partir duquel est déterminé un classement provisoire.

Début phase
test final



Classement final des compétiteurs

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



Organiser un Data Challenge selon les bonnes pratiques

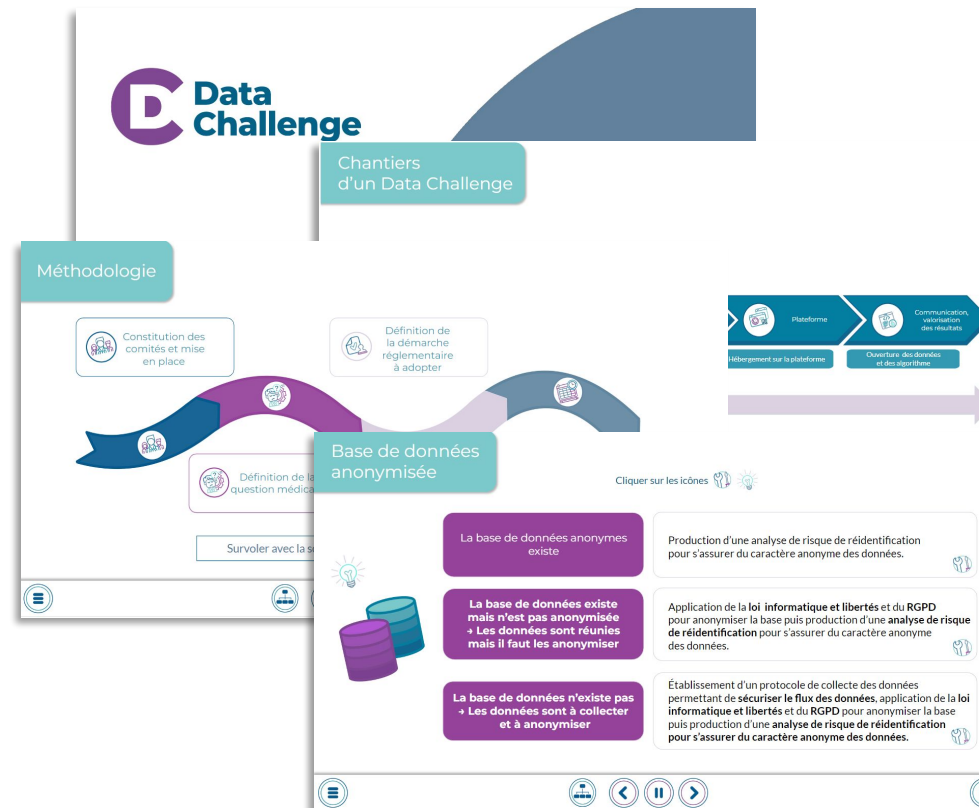
L'organisation d'un Data Challenge en santé requiert des étapes de **cadrage scientifique**, **méthodologique**, **financier** et **règlementaire** spécifiques à la **réutilisation de données de santé**.

Pour mener un projet de ce type de façon **conforme**, en particulier sur le plan **règlementaire**, et obtenir des **résultats robustes**, il est important de s'informer sur les **bonnes pratiques** à suivre pour organiser un Data Challenge en santé.

Pour partager les bonnes pratiques,
le HDH a lancé le **MOOC Data Challenge en santé**

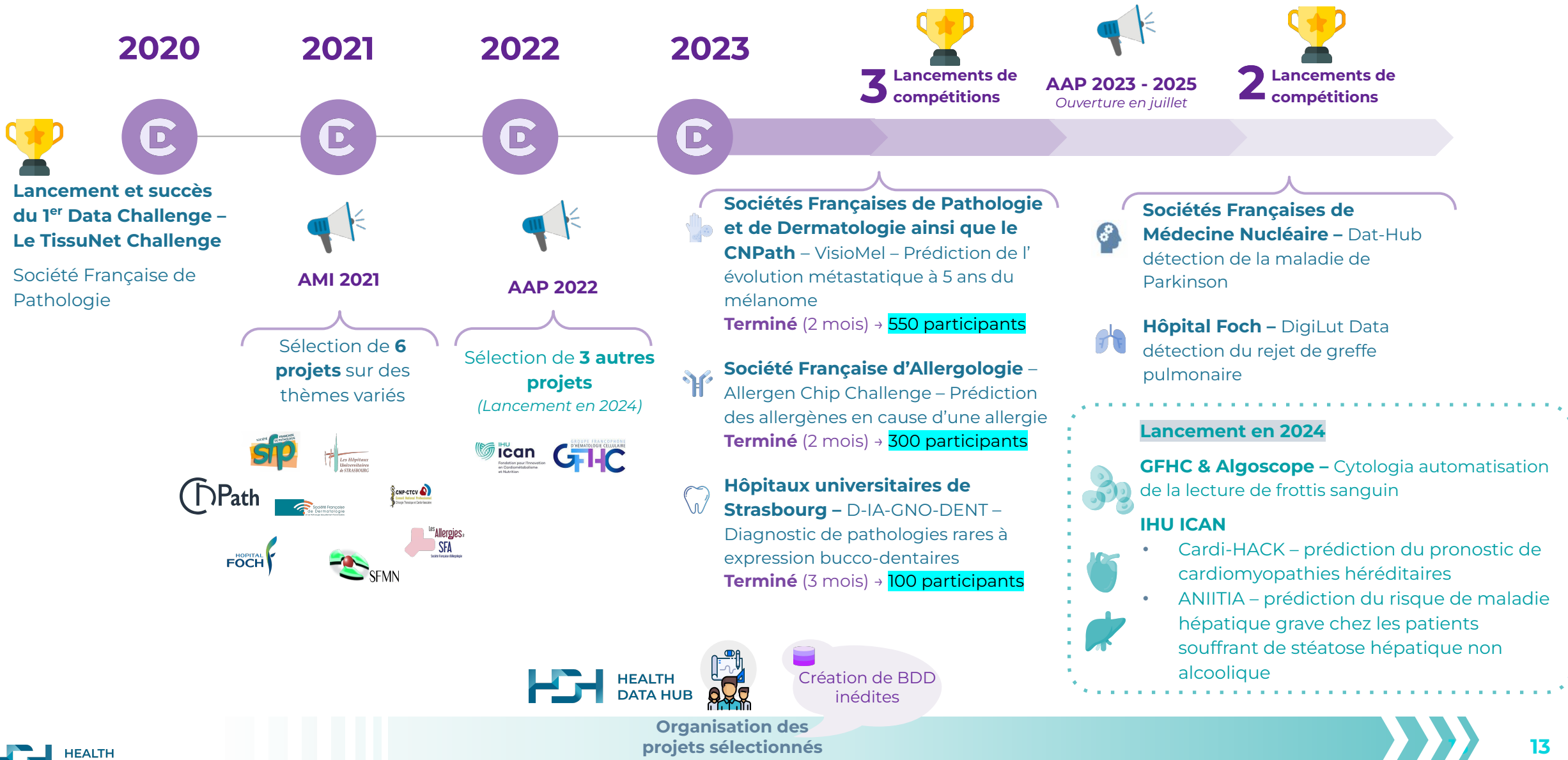


Le MOOC est destiné à tout
porteur de projet ou membre de l'équipe projet qui souhaite se former
sur **les aspects organisationnels d'un Data Challenge**.



LES DATA CHALLENGES EN SANTÉ

Le programme Data Challenge en santé du HDH





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

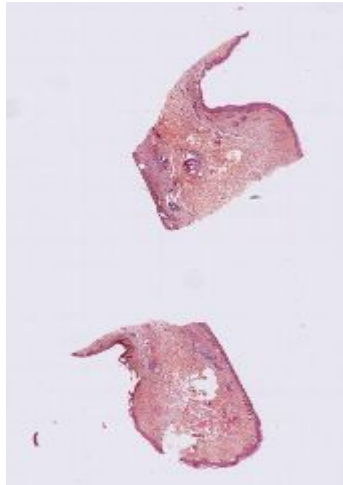
<https://www.youtube.com/watch?v=Ue0Lt1RKaAE&t=502s>

Le premier challenge de la SFP en quelques chiffres...

574 Participants du monde entier

36 Équipes constituées

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



- **2^{ème}** Challenge de la SFP
- **2000** patients de **40** centres
- **550** compétiteurs du monde entier



VisioMel Challenge

- **100** participants à la cérémonie de remise des prix
- Poursuite des travaux avec les lauréats



- Données biologiques de plus de **4000** patients
- **300** compétiteurs du monde entier

Allergen Chip Challenge

- Évènement thématique de remise des prix le **17/11/2023**



- Base riche portant sur des maladies rares à expression bucco-dentaires
- Des scores de performance allant jusqu'à **84%**



D-IA-GNO-DENT

- Présentation des résultats aux rencontres O-Rare le **03/10/2023** et à l'évènement AI for Health le **22/11/2023**



Les Data Challenges en santé...

...un levier innovant pour transmettre la culture de l'Open Science

L'accès aux données pendant la compétition est une 1^{ère} étape qui catalyse la mise en Open Data

L'Open Science en santé a le potentiel de **transformer la manière dont la recherche médicale est conduite**, la rendant plus **transparente**, **inclusive** et **efficace**. Convaincu que les découvertes les plus impactantes seront issues de **l'intelligence collective**, le HDH soutient et accompagne l'écosystème de la santé numérique dans la démarche Open Science.

Collaboration et réduction de la redondance

Évite la duplication des efforts de création et de structuration de bases de données de santé, permet une utilisation plus efficace des ressources.

Innovation ouverte

Permet aux acteurs non traditionnels (comme les citoyens scientifiques) d'apporter des contributions significatives à la recherche en santé

Transparence et reproductibilité

L'accessibilité des données permet de vérifier et de reproduire les résultats, renforçant ainsi la confiance dans les conclusions scientifiques.

Protection contre les biais commerciaux

En rendant les données accessibles à tous, on limite le risque que les résultats soient influencés par des intérêts commerciaux et on réduit les possibilités d'exploitation purement commerciales de données.

Accélération de la recherche médicale

Le partage des données permet aux chercheurs d'accéder à des ensembles de données plus vastes et variés, ce qui peut accélérer la R&D en santé.

Renforcement de la confiance du public

Lorsque les données sont ouvertes et transparentes, le public est plus susceptible de faire confiance aux conclusions et recommandations issues de la recherche.



Ouverture d'un nouvel appel à projets

APPEL A PROJETS

DATA CHALLENGES EN SANTÉ

Le HDH en partenariat avec la Délégation Ministérielle au Numérique en Santé, le SGPI et Bpifrance a lancé en juillet un nouvel appel à projets !

NOUVEAU !

**APP ouvert en continu
jusqu'en juin 2025**
Prochaine relève des
candidatures :

Février 2024



Des équipes sont appelées à candidater pour organiser un challenge sur une thématique médicale de leur choix



Le dépôt de candidature est ouvert sur la [plateforme en ligne de Bpifrance](#)



Les projets sélectionnés lors de cet appel à projet devront notamment répondre aux critères suivants...



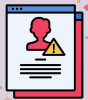
Intérêt de la question clinique

La question médicale posée doit présenter un **intérêt clinique** et être **originale**.



Pertinence de l'approche prédictive pour y répondre dans le délai imparti

La problématique doit pouvoir être traitée par une **solution d'apprentissage supervisé**, le score de performance des algorithmes doit être mesurable et les données d'une **qualité suffisante** pour que les résultats soient fiables.



Faisabilité de la collecte des données et de leur anonymisation

Les données doivent pouvoir être **collectées en nombre suffisant** dans un temps imparti à un ou des formats numériques homogènes. Les données doivent pouvoir être **totalemtent anonymisées dans le respect du RGPD**.



Partage des données

Les données collectées et anonymisées dans le cadre du projet seront **partagées en Open Data** à l'issue du Data Challenge. Des modalités de citation des bases sont toutefois prévues. Les gagnants devront **partager leur algorithme en Open Source** pour obtenir le prix de récompense.



Forte implication et faisabilité du projet

Une **forte implication de votre part** est requise sur toute la durée de l'organisation du data challenge. Plusieurs personnes de votre organisation devront s'impliquer dans le Conseil Scientifique et dans le Comité d'Organisation. Les grands jalons proposés pour le projet doivent le rendre **réalisable en environ un an**.

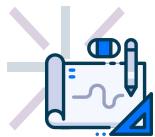


Combinaison de plusieurs spécialités médicales (facultatif)

Le projet peut faire intervenir **plusieurs spécialités médicales** et combiner des données de natures différentes mais il faudra s'assurer que cela ne remette pas en cause leur caractère anonyme.

L'accompagnement proposé par le HDH

Le HDH accompagnera les projets sélectionnés à **toutes les étapes de leur réalisation** en apportant une aide **organisationnelle, logistique et technique** adaptée.



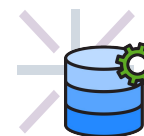
DÉFINITION ET CADRAGE DU DATA CHALLENGE

- ✓ Aide au **cadreage scientifique** et **organisationnel** du projet
- ✓ Aide **juridique** sur la réglementation applicable au traitement des données de santé
- ✓ **Coordination** avec des **prestataires** spécialisés dans l'hébergement de Data Challenges



AIDE À LA COLLECTE ET À L'ANONYMISATION DES DONNÉES

- ✓ Accompagnement dans la **collecte** des données
- ✓ Accompagnement dans **l'anonymisation** des données
- ✓ **Support technique** dans la phase de labellisation des données
- ✓ **Partage des bonnes pratiques**



GESTION DE L'INFRASTRUCTURE D'HÉBERGEMENT DES DONNÉES

- ✓ Accompagnement par des data engineers & data managers pour constituer la base de données du Data Challenge
- ✓ **Mise à disposition** de serveurs sécurisés



MISE EN VALEUR DES RÉSULTATS

- ✓ **Communication** (pré et post challenge)
- ✓ Accompagnement dans la **publication** d'articles scientifiques
- ✓ Support dans les **démarches Open Data et Open Source**
- ✓ Mise en relation des acteurs de l'écosystème



bpifrance



**Jusqu'à 300 000 euros de
financement**



En complément de l'accompagnement par le HDH, les projets sélectionnés pourront bénéficier d'un **financement du Grand Défi** : « Améliorer les diagnostics médicaux à l'aide de l'intelligence artificielle »

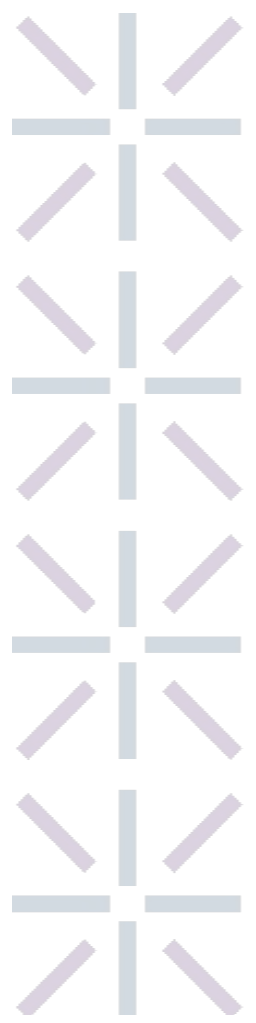
Ce régime d'aide peut financer les projets à hauteur de 300K€ maximum sur une période glissante de 3 ans. Les projets financés seront invités à déclarer les aides reçues au titre de-minimis au cours des 3 dernières années.

Porteurs de projets éligibles au financement Bpifrance et à l'accompagnement du HDH :

- ✓ Sociétés savantes rattachées à un Conseil National Professionnel
- ✓ Établissements de santé publics ou privés
- ✓ Groupements de coopération sanitaire
- ✓ Structures labellisées 'Institut Hospitalo-Universitaire'

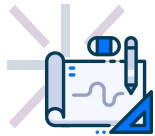
Les partenaires financés par l'AAP doivent être des personnes morales immatriculées en France.

Exemple de budget d'un projet



	exemple de coûts	Porteur du coût
Plateforme Data Challenge	env. 100 000 €	BPI
Récompense gagnants Data Challenge	25 000 €	BPI / Porteur de projet
Prestation DPO	env. 20 000€	BPI
Rétribution inclusion données (compensation financière associée au travail de collecte des données le cas échéant)	env. 25 000 € (varie selon la nature et la volumétrie)	BPI
Infrastructure d'hébergement HDH	env. 1 000 €	HDH
Information aux patients	env. 1 000 €	BPI
Outil d'annotation	env. 6 000 €	BPI
Événement de remise des prix et annonce des résultats	Env. 15 000 €	BPI / Porteur de projet
TOTAL	Env. 195 000 €	

L'organisation du Data Challenge nécessitera une **forte mobilisation de votre équipe**.
Un Data Challenge réussi passe aussi par une **forte implication de l'ensemble des partenaires** !



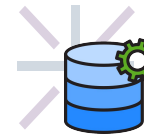
DÉFINITION ET CADRAGE DU DATA CHALLENGE

- ✓ Implication de plusieurs personnes dans le Comité d'Organisation et le Conseil Scientifique du Data Challenge sur toute la durée de l'organisation
- ✓ La structure organisatrice du Data Challenge sera responsable juridique du projet



COLLECTE ET ANONYMISATION DES DONNÉES

- ✓ Motivation et implication des centres à partir desquelles les données sont collectées
- ✓ Mise en place de la stratégie de la collecte et de la labellisation des données (numérisation, anonymisation...)



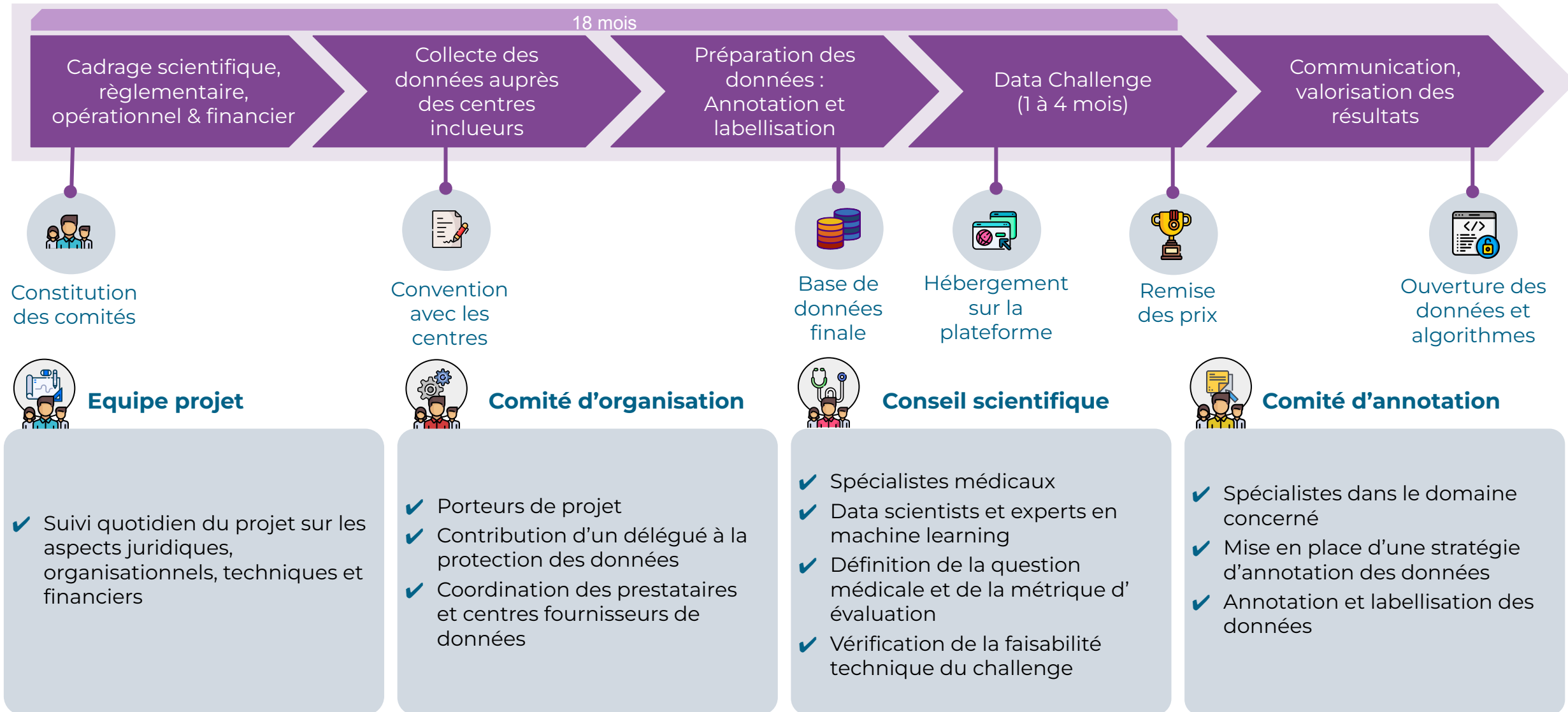
GESTION DE L'INFRASTRUCTURE D'HÉBERGEMENT DES DONNÉES

- ✓ Suivi de la gestion des données
- ✓ Choix de la plateforme d'hébergement de la compétition



MISE EN VALEUR DES RÉSULTATS

- ✓ Rédaction d'articles dans des revues scientifiques
- ✓ Collaboration avec le HDH dans les travaux de communication
- ✓ Collaboration dans la mise en œuvre de la politique de partage des données



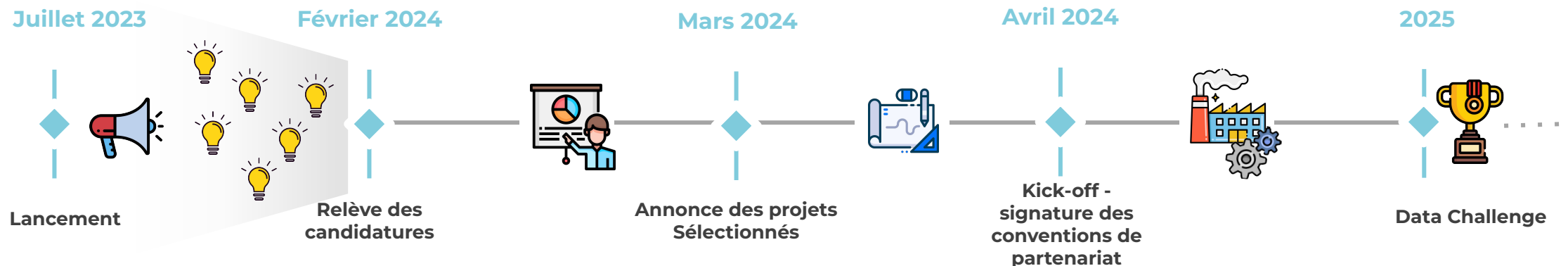
Constitution du dossier candidature...



Les candidats à l'AAP devront constituer un **dossier** afin de décrire les différents aspects de leur projet. Le dossier de candidature devra notamment présenter les éléments suivants :

- ✓ Présentation du ou des **porteur(s) de projet** et description de **l'équipe projet**
- ✓ Domaine **médical** et **problématique envisagée**, intérêt et **caractère innovant**
- ✓ Description et justification des **données visées**, source, faisabilité de l'anonymisation
- ✓ Présentation du **design envisagé de la compétition** : approche IA, métrique d'évaluation...
- ✓ Modalités de **partage des données**
- ✓ Plan de **financement**

Modalités de candidature à l'AAP



Dépôt des candidatures

Les porteurs de projets sont invités à déposer leur candidature sur la plateforme de Bpifrance.

NOUVEAU !

L'AAP est ouvert en continu et des vagues successives de relèves auront lieu jusqu'en **juin 2025**.

Pré-sélection des projets* sur la base de critères d'éligibilité

Les candidatures seront analysées par un jury pour établir une présélection.

Audition et sélection des projets

Approfondissement des échanges avec les projets présélectionnés pour **affiner leur cadrage et s'assurer de leur réelle faisabilité** ; puis audition par un jury pour **sélection définitive** des lauréats.

Lancement de l'organisation Data Challenges

Lancement des projets sélectionnés.
Selon la complexité de la collecte des données, la date de lancement et la durée du Data Challenge varieront.

Accompagnement Post Challenge

Valorisation des résultats à travers des outils de communication

Ouverture et partage des données & des algorithmes

*Cette phase sera maintenue ou non en fonction du nombre de projets déposés.

Appel à projets DATA CHALLENGES EN SANTÉ

Pour rester informé sur les ouvertures des
Data Challenges, les actus et les
événements, inscrivez-vous à la
newsletter Data Challenge en santé



Pour toute question, vous pouvez nous écrire à l'adresse mail suivante :

data.challenge@health-data-hub.fr