LES ENJEUX DU TRAITEMENT DES DONNÉES DE LA COVID-19 : DES OUTILS & DES HOMMES

16/06/2022 – Julien Durand

Chargé de projets scientifiques à Santé publique France

CONFLITS D'INTERETS

Cette intervention est faite en tant que personnel de Santé publique France, organisateur de la manifestation. Je n'ai pas de conflit d'intérêt en lien avec le sujet traité.

LES ENJEUX

Santé publique France

- Collecter des données sur l'ensemble du périmètre de la covid-19
- Centraliser, stocker, router les données vers les différents acteurs
- Traiter des données en quantité massive quotidiennement
- Monitorer, corriger et informer de toute anomalie en temps réel
- Mettre à disposition des indicateurs dans de nombreux formats

ENJEUX DES SYSTÈMES DE SURVEILLANCE

Santé publique France

Quelles données recueille-t-on ?

Consultations en médecine de ville

Passages aux urgences hospitalières, hospitalisations liées au Coronavirus dont celles en réanimation, et profil des personnes hospitalisées

Cas déclarés dans les ESMS, Ehpad et professionnels en collectivité

Mortalité : excès de mortalité (données Insee), certificats de décès électroniques.....

Examens de biologie médicale : recherche du virus (RT-PCR) et recherche d'anticorps (sérologie)

ENQUÊTES :

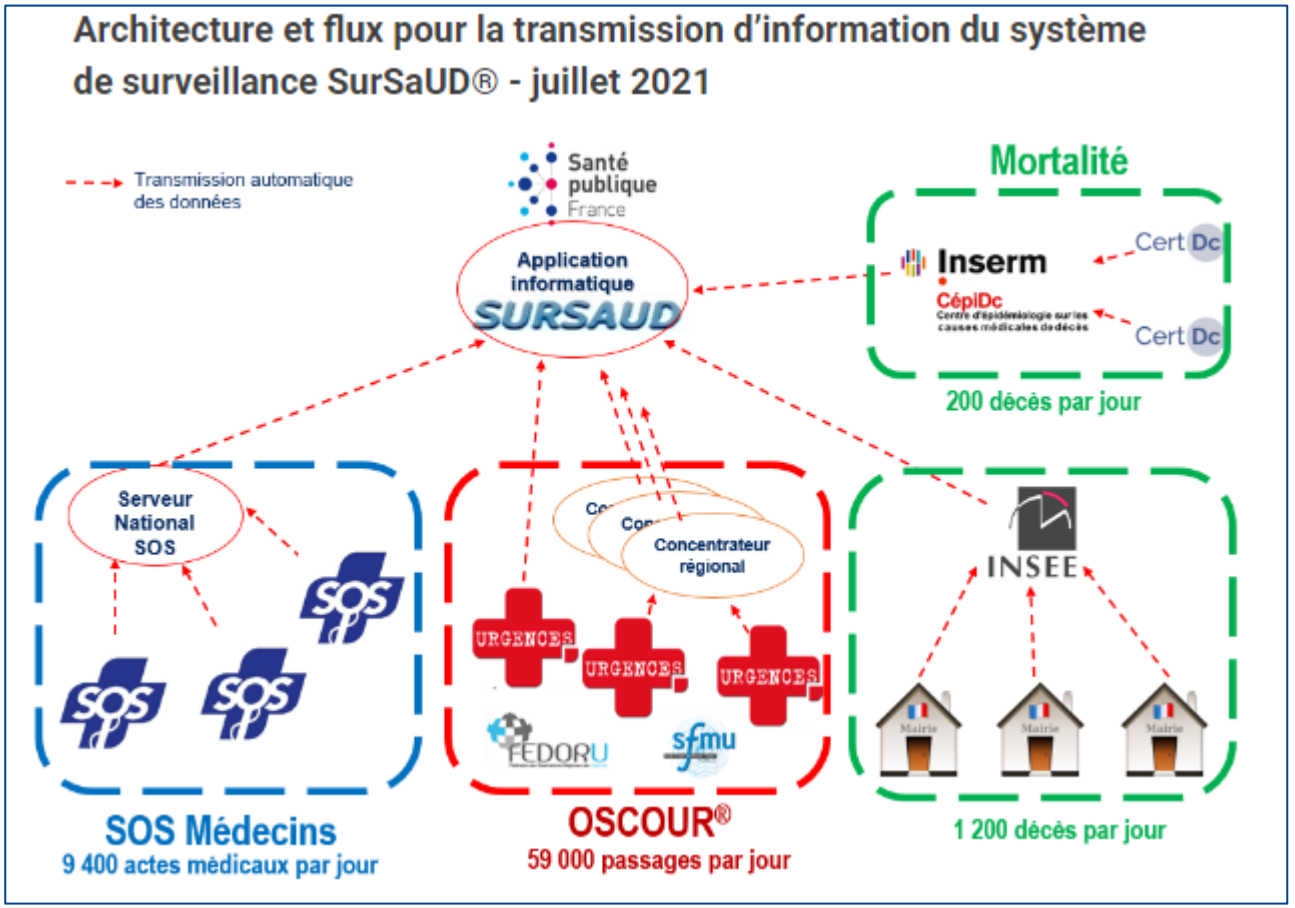
Enquêtes observationnelles et de séroprévalence dans la population générale, chez les travailleurs et les professionnels de santé

- **De fortes contraintes sur les systèmes de surveillance covid**
  - Données sur l'ensemble du périmètre de la **surveillance et de l'investigation**
  - **Données exhaustives** pour chaque surveillance
  - Données **réactives en flux continu**
  - Données **normées et référentiels standardisés**
  - Des systèmes de surveillance à développer **dans des délais courts**

EXPLOITER LES SYSTÈMES DE SURVEILLANCE EN PLACE

Santé publique France

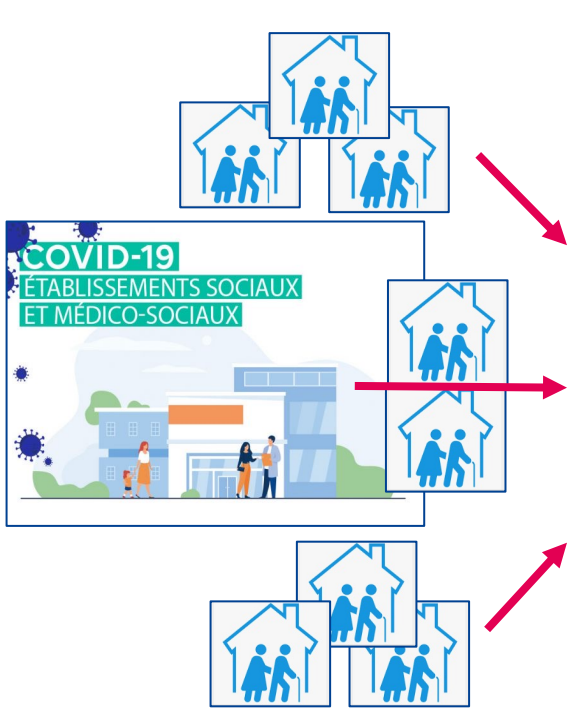
- **Système de surveillance des urgences et décès**

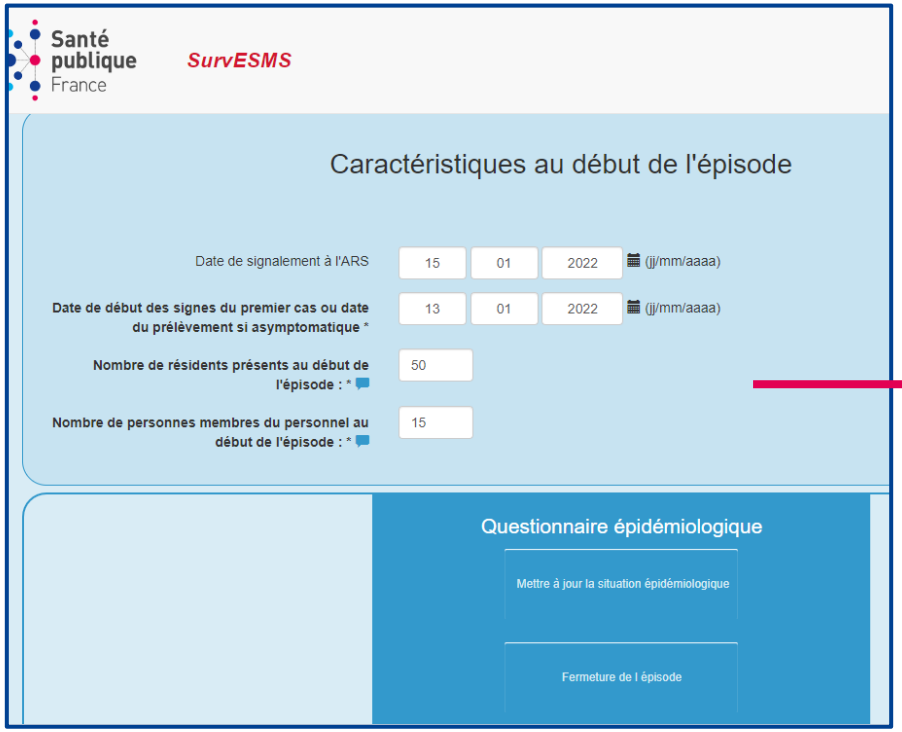


DÉVELOPPER DE NOUVEAUX SYSTÈMES DE SURVEILLANCE EN URGENCE

Santé publique France

- **Système de surveillance des établissements sociaux et médico-sociaux**
  - Pas de système d'information pré-existant dans les établissements
  - Nécessité de développer une plateforme unique pour la remontée de données
  - Temps de développement très réduit

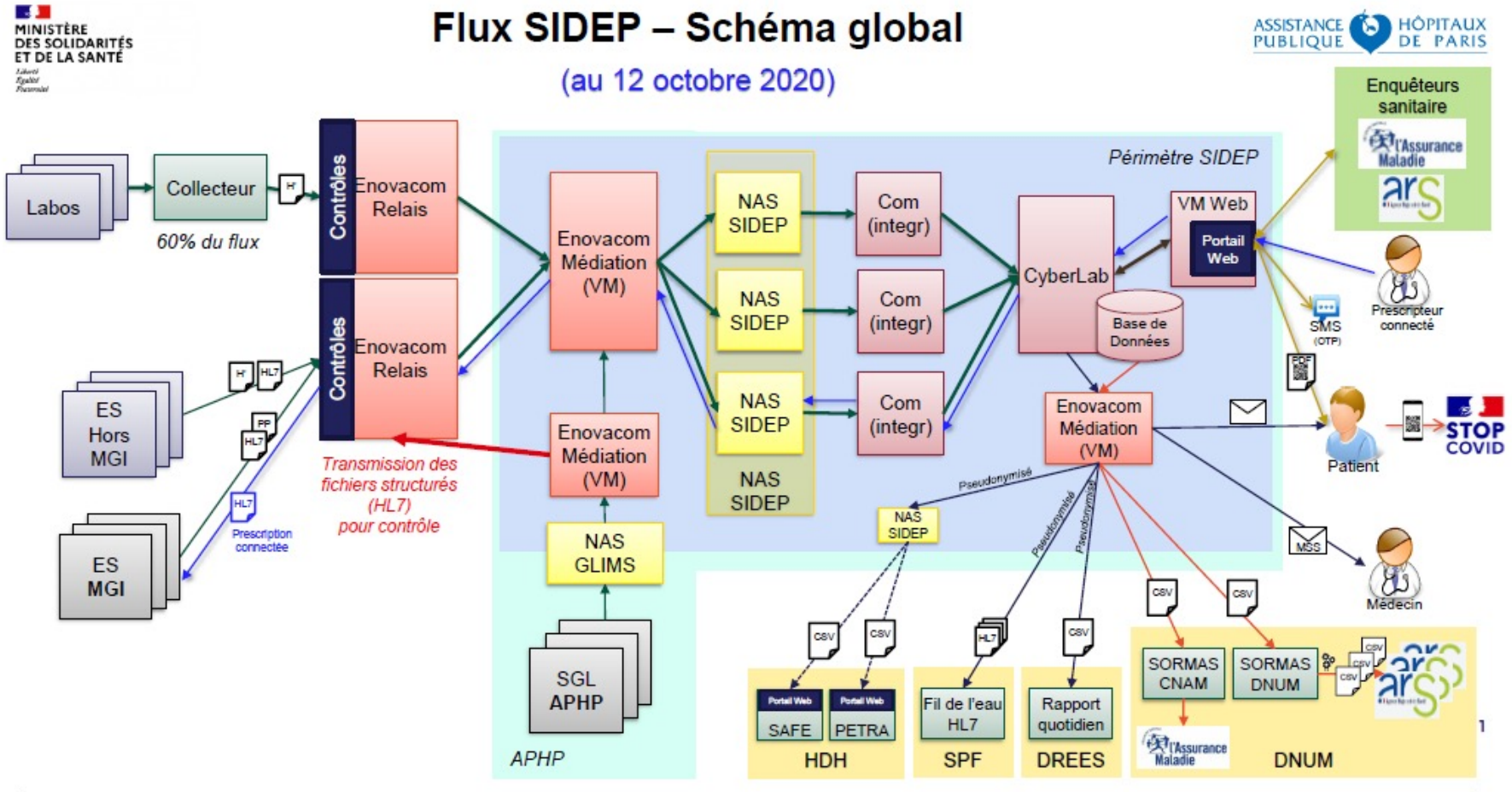




CENTRALISER, STOCKER, ROUTER LES DONNÉES VERS LES DIFFÉRENTS ACTEURS

Santé publique France

- **Système de surveillance du dépistage (SIDEP)**



EVOLUTION CONSTANTE DES SYSTÈMES DE SURVEILLANCE

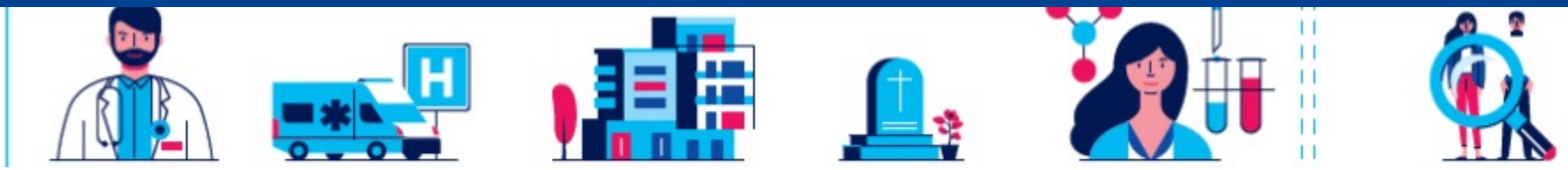
Santé publique France

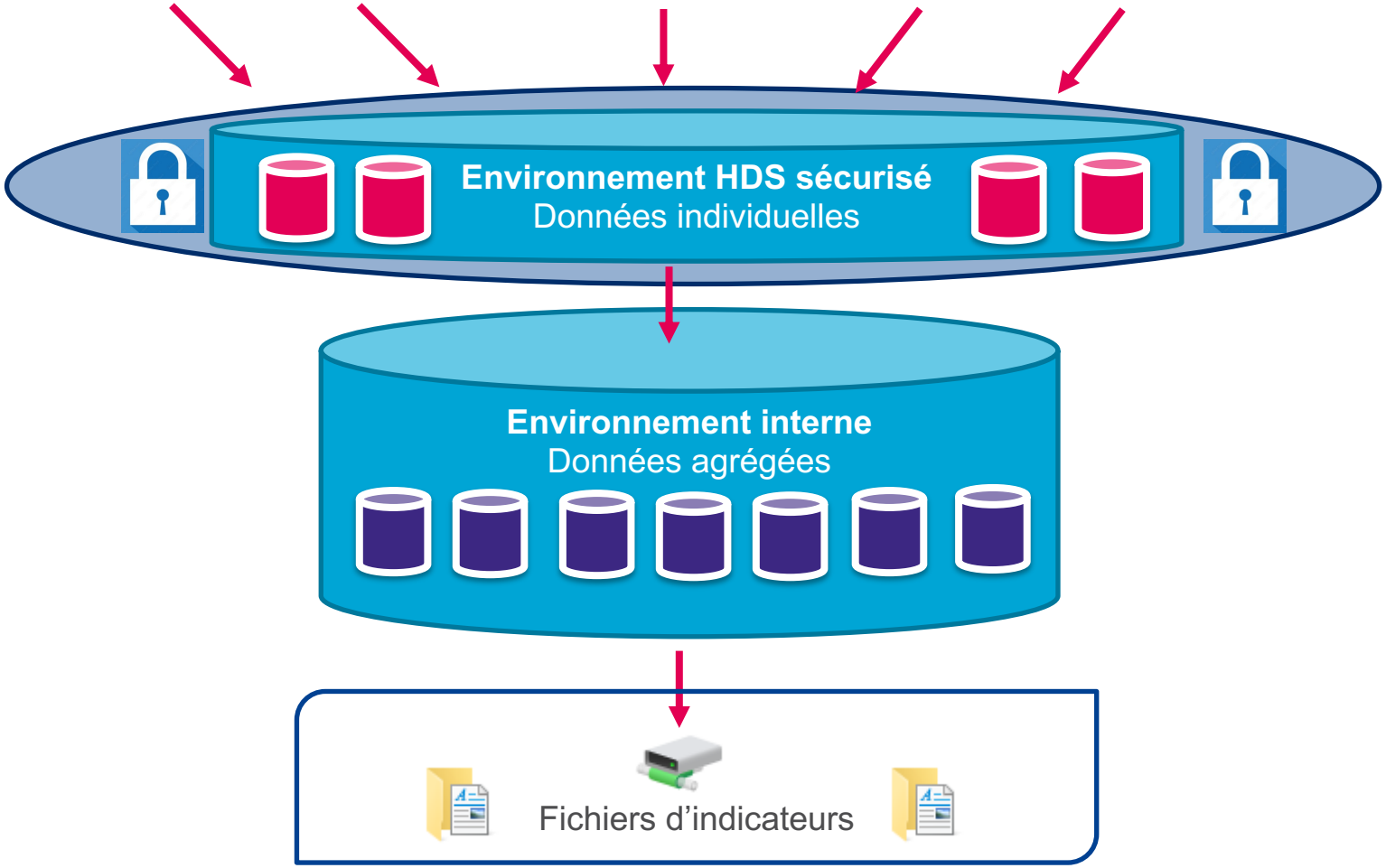
- **Système de surveillance du dépistage (SIDEP)**

05/20	• Lancement SIDEP : tests PCR
07/20	• Intégration sérologies
10/20	• SIDEP 3 : Intégration des tests antigéniques • SIDEP 4 : Structures hors laboratoires (pharmacies etc.)
02/21	• Intégration suivi des variants : criblage et séquençage
06/21	• Evolution vers le suivi des mutation

INTÉGRER QUOTIDIENNEMENT DES DONNÉES EN QUANTITÉ MASSIVE

Santé publique France



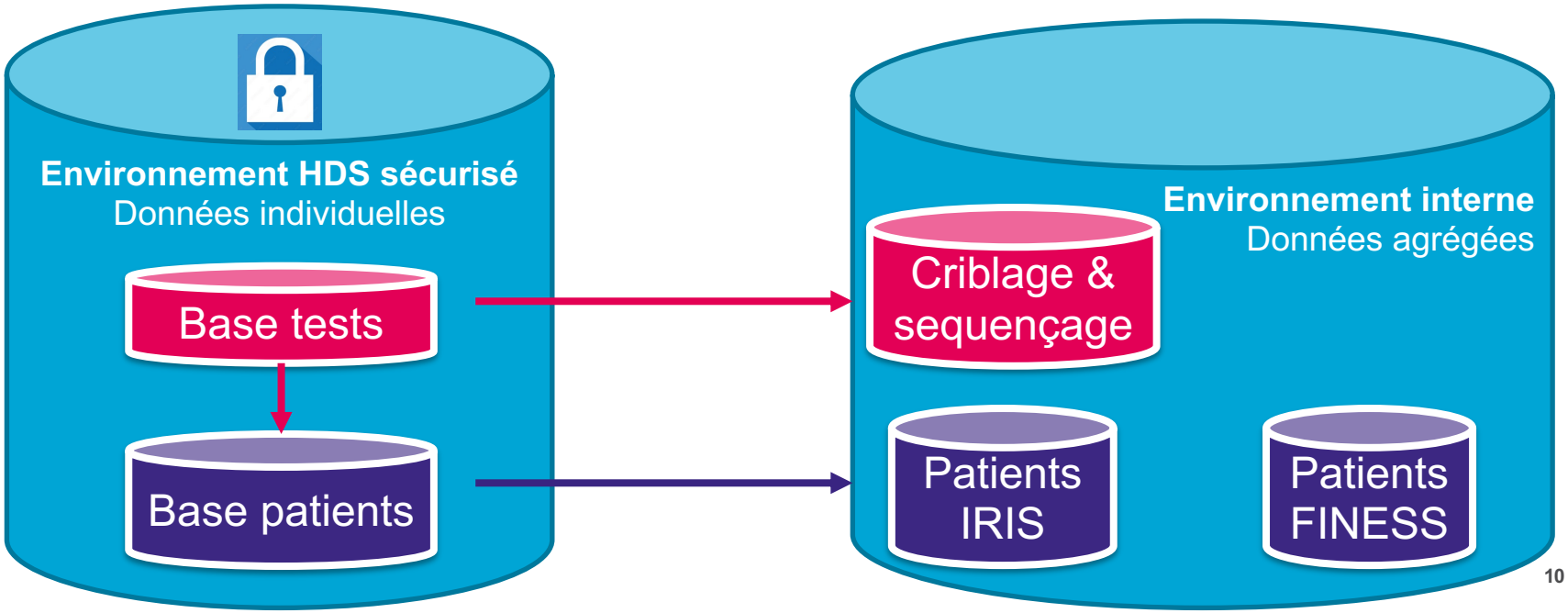


TRAITER QUOTIDIENNEMENT DES DONNÉES EN QUANTITÉ MASSIVE

Santé publique France

- **Intégration et traitement continu de données en quantité massive**
  - Robustesse de l'infrastructure IT
  - Plusieurs millions de lignes de données reçues quotidiennement

- **Pré-traiter, optimiser les bases de données**
  - Intégration du « delta »
  - Partitionnement de tables (gestion de données « chaudes » VS « froides »)
  - Pré-agrégation des données (échelons géographiques, temporelles, définitions de cas)

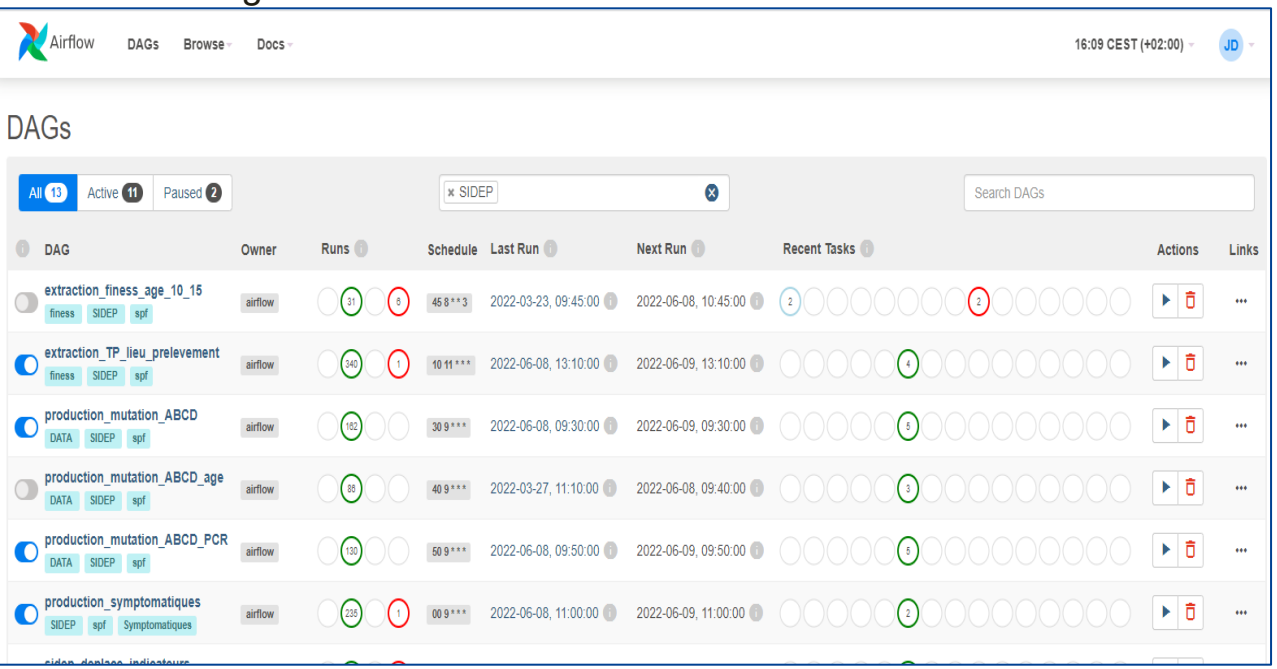


TRAITER QUOTIDIENNEMENT DES DONNÉES EN QUANTITÉ MASSIVE

Santé publique France

- **Créer les indicateurs**
  - Développement de scripts (langages de programmation R, Python, etc.)
  - Faire évoluer, historiser, partager des scripts : **outils collaboratifs type Git**

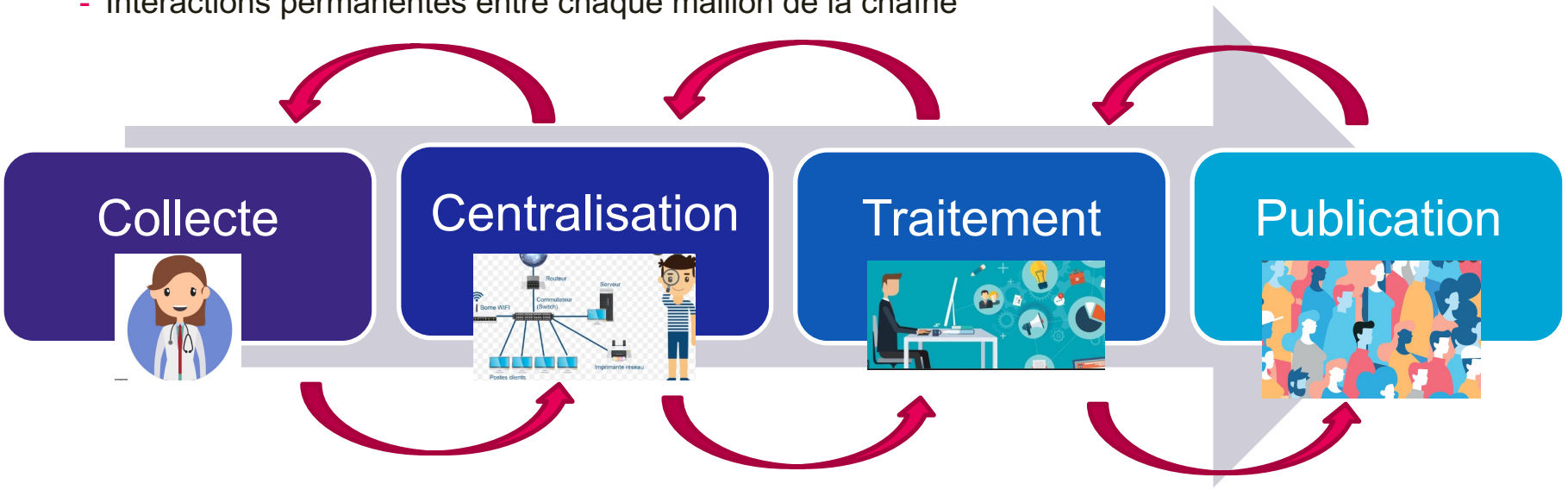
- **Automatiser, monitorer les traitements**
  - Utilisation d'outils d'ordonnancement (Airflow)
  - Outil centralisé à l'interface de différents serveurs pour une gestion intégrée de l'ensemble des workflow



ECHANGER EN TEMPS RÉEL AUTOUR DE TOUT ÉVÈNEMENT IMPRÉVU

Santé publique France

- **Surveiller l'ensemble des flux de données en temps réel**
  - Interactions permanentes entre chaque maillon de la chaîne

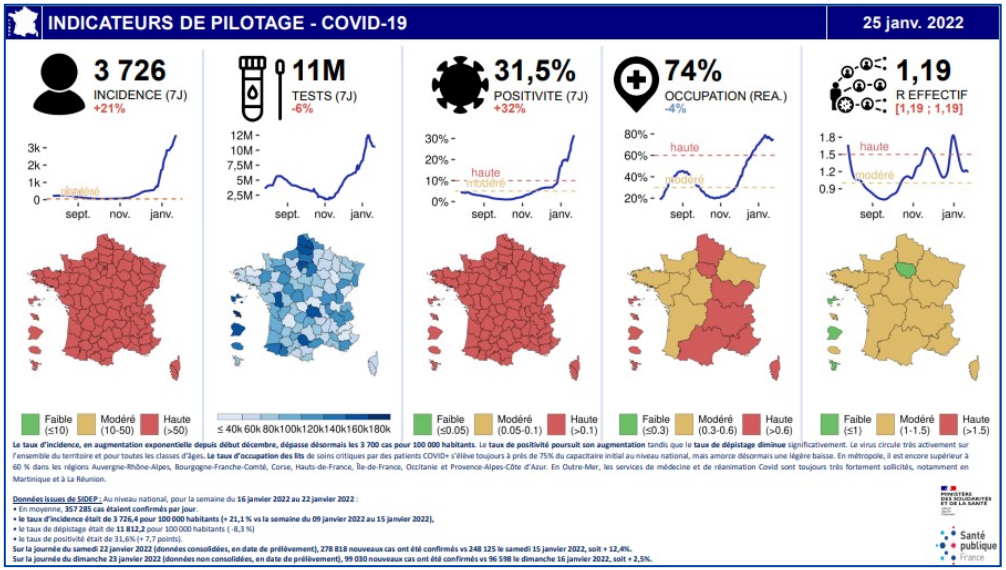


METTRE À DISPOSITION DES INDICATEURS DANS DE NOMBREUX FORMATS

- **Des productions quotidiennes**
  - Bilans à destination du décideur
    - Note de synthèse
    - Bilan détaillé

  => Utilisation d'outils type « R-Markdown »

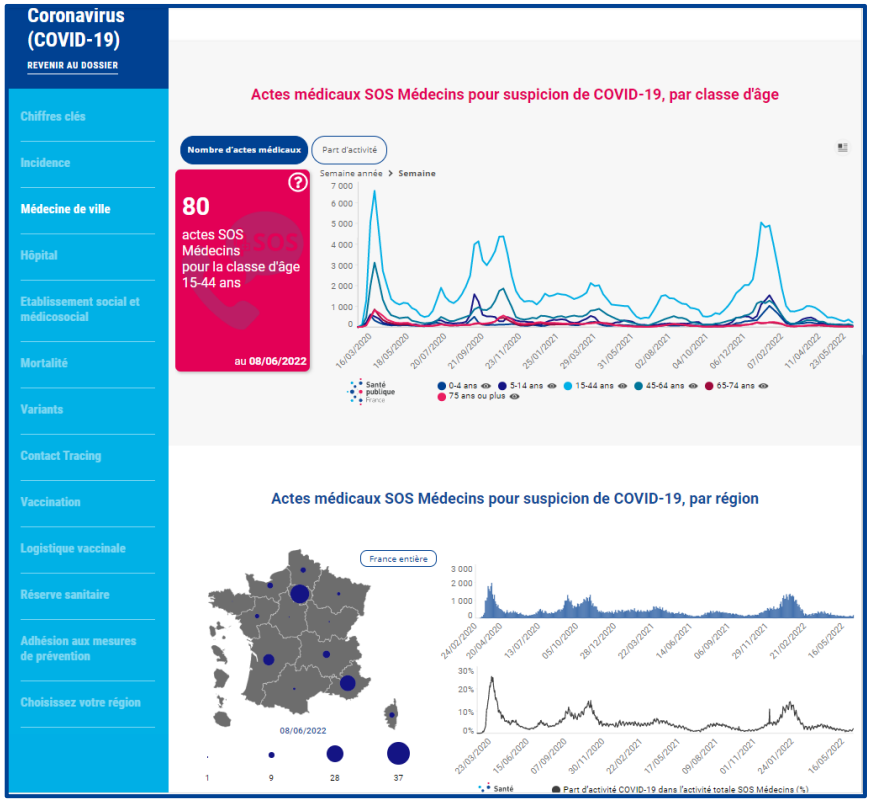
- **Des productions hebdomadaires**
  - Point épidémiologique national
  - Points épidémiologiques régionaux

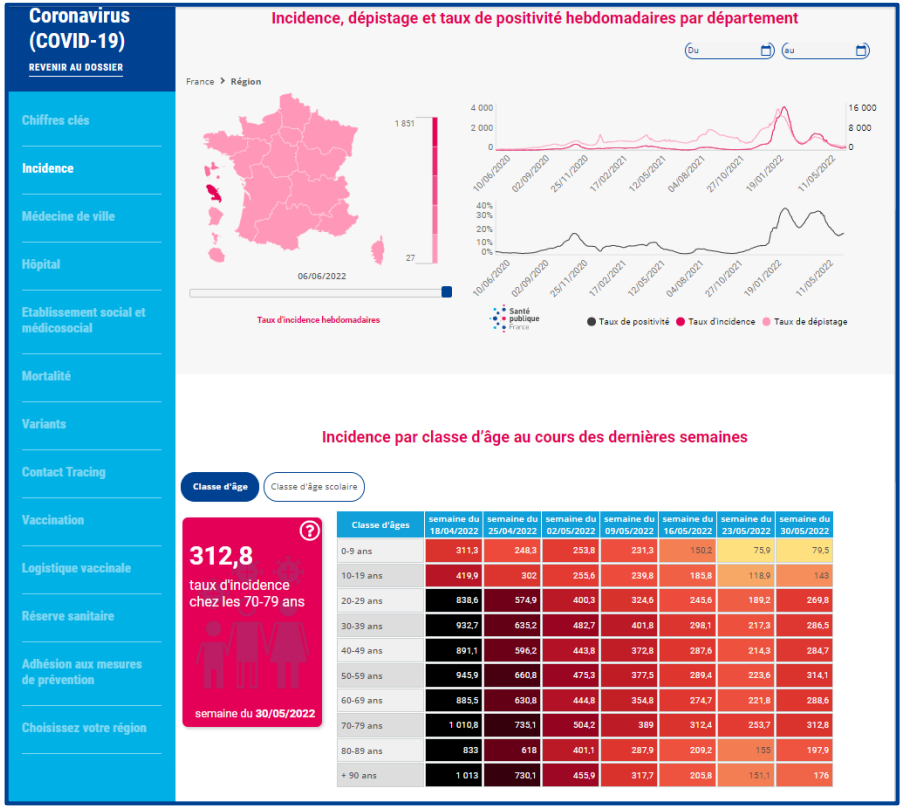


ALIMENTER DES TABLEAUX DE BORD DYNAMIQUES

Santé publique France

- **Développement d'un dashboard « Info Covid France »**

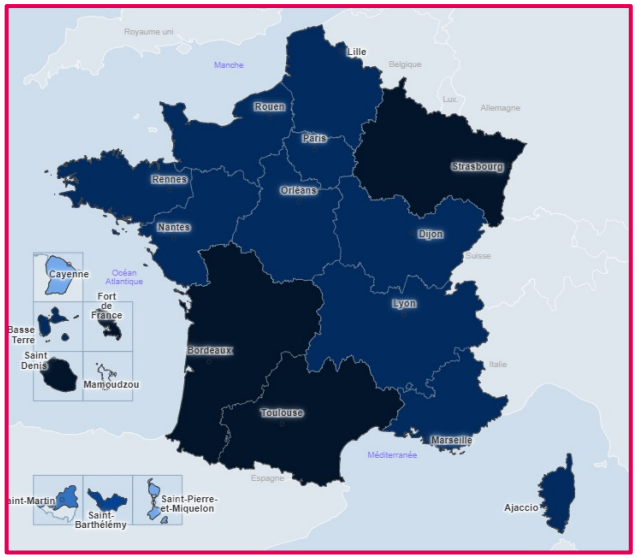


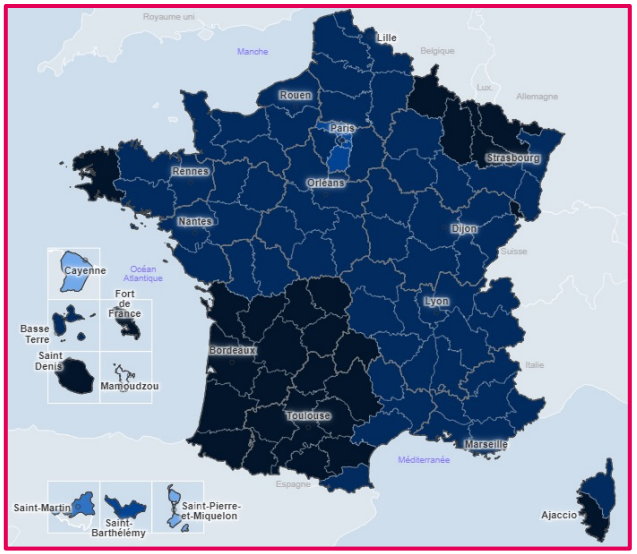


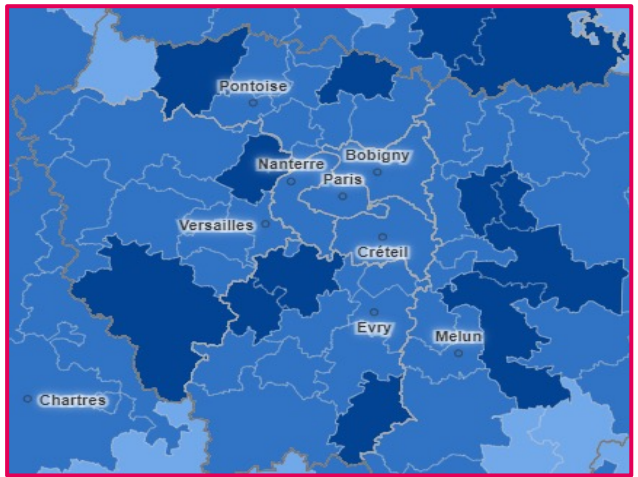
ALIMENTER DES TABLEAUX DE BORD DYNAMIQUES

Santé publique France

- **Observatoire cartographique GEODES**
  - Exemple de déclinaison géographique du taux d'incidence







DIFFUSER LES DONNÉES EN OPEN-DATA

Santé publique France

- **Mise en ligne en open-data**
  - Ensemble des indicateurs issus des systèmes de surveillance covid disponibles en ligne

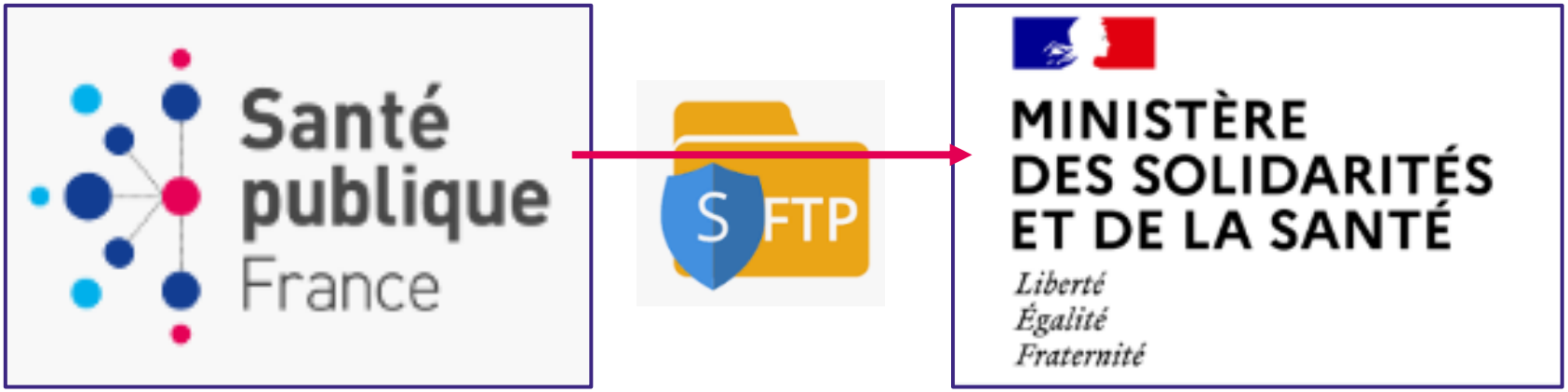
- **Plus de 150 indicateurs mis en ligne quotidiennement par SpF**
  - Un indicateur, de multiples déclinaisons :
    - Taux d'incidence (nombre de nouveaux cas pour 100 000 habitants)
      - Echelons géographiques (national, régional, départemental, EPCI, métropole, IRIS)
      - Echelle temporelle (quotidien, semaine glissante, semaine calendaire)
      - Déclinaison par sexe, par âge (10 ans, 15 ans, scolaires etc.)

- **Des guides méthodologiques pour chaque indicateur, évolution**
  - Un accompagnement des utilisateurs « professionnels » (journalistes, data scientists)
  - De nombreuses sollicitations du grand public

PARTAGER QUOTIDIENNEMENT DES DONNÉES AVEC LES DÉCIDEURS

Santé publique France

- **2Go de données partagées quotidiennement avec le ministère**



CONTEXTE DE TRAVAIL DE CRISE

Santé publique France

- **Contexte de travail stressant**
  - Production et diffusion obligatoire de l'ensemble des éléments chaque jour à des heures précises
  - Nécessité de déployer des moyens de crise dans la durée

- **Rythme de travail extrêmement soutenu**
  - Cycle de réception / traitement des données => production d'indicateurs très court et quotidien
  - Travail en semaine et week-end pour assurer l'ensemble des productions depuis le début de la crise
  - Mobilisation de dizaines de personnes chaque week-end

- **Délais de travail très réduits**
  - Demandes constantes d'ajustements ou de développement de nouveaux indicateurs (demandes internes, tutelles, locales)
  - Détection et résolution d'anomalies en urgence (problème de formats de données, flux de données, infrastructure, anomalie de scripts, etc.)

- **Evolution permanente des systèmes de surveillance covid-19**

DES RÉORGANISATIONS HUMAINES ET MATÉRIELLES

Santé publique France

- **Réorganisations pour répondre à l'ensemble des challenges**
  - Redéploiement de ressources internes pour faire face à la quantité de travail
  - Renforts de nouveaux profils data scientists
  - Renforcement des collaborations intra-directions
  - Renforcement des collaborations inter-agences (Drees, Cnam, DNUM, etc.)

- **Déploiement de nouveaux outils**
  - Utilisation du langage Python
  - Outils collaboratifs GIT
  - Outils de gestion de Worklfow : Airflow

- **Redimensionnement de l'infrastructure technique**

Santé publique France

MERCI !