Construire et exploiter des entrepôts de données de santé grâce à l'IA

Marie-Christine Jaulent (DR INSERM)

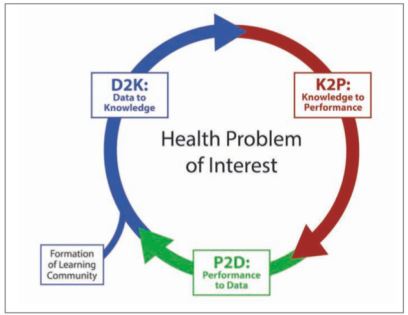
LIMICS – UMR_S1142
Inserm, Sorbonne université & université Paris 13

LIMICS

- Recherche interdisciplinaire
  - Informatique (CNU 27)
    - **Intelligence Artificielle** : Représentation et Ingénierie des Connaissances, Systèmes d'Aide la Décision, Apprentissage Automatique
  - Informatique biomédicale et biostatistiques (CNU 46.04)
    - **Informatisation du Système de Santé**, Applications de la e-santé, Entrepôts de données de santé, Informatique de la Recherche Clinique
- 2020 : 69 membres dont 39 membres permanents
  - Effectifs équilibrés entre les deux disciplines

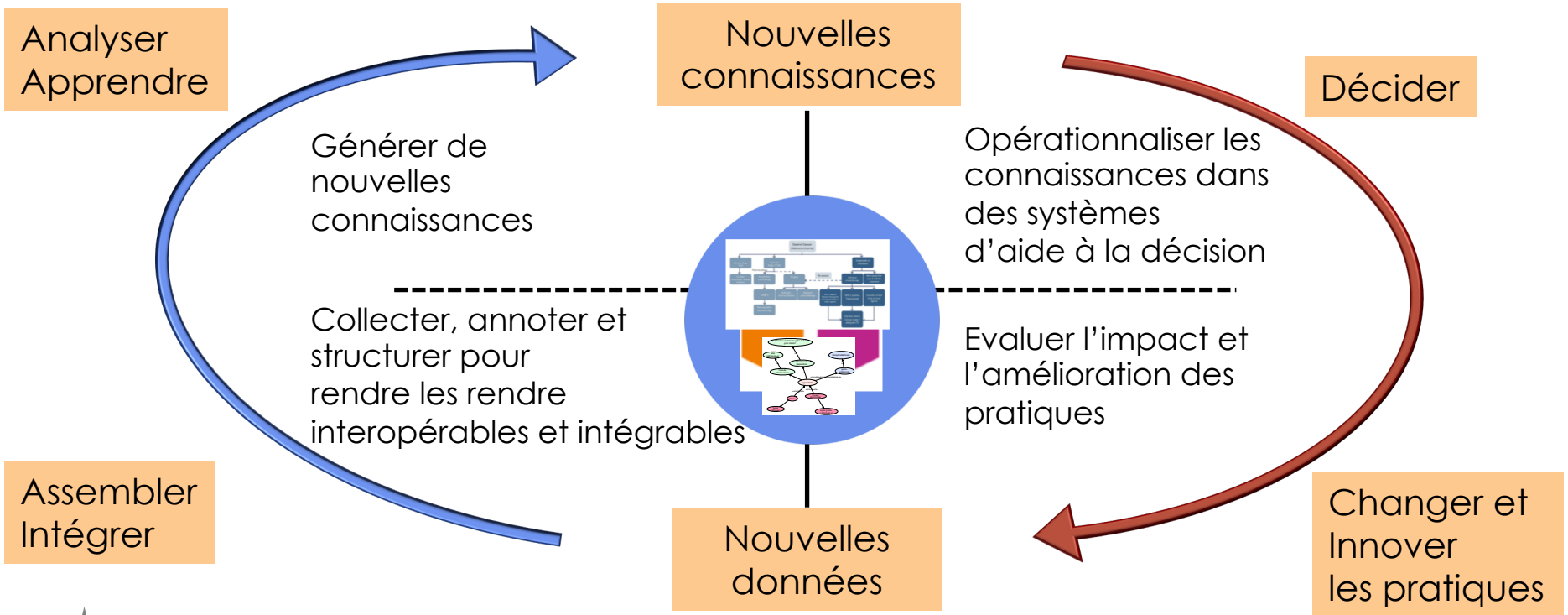
« Learning Health System »

- Volonté de mettre en place un cercle vertueux pour accélérer l'innovation en santé
  - Amélioration continue du système de santé par auto-apprentissage à partir notamment de données de vie réelle
  - Accélération de l'innovation
  - Passage naturel des connaissances à la pratique



Learning Health Systems.
C. P. Friedman, A. K. Wong, D. Blumenthal. Achieving a Nationwide Learning Health System. Sci. Transl. Med. 2(57) ***2010****.*

Systèmes de Santé Auto-Apprenants



Systèmes de Santé Auto-Apprenants

Analyser
Apprendre

Nouvelles
connaissances

Décider

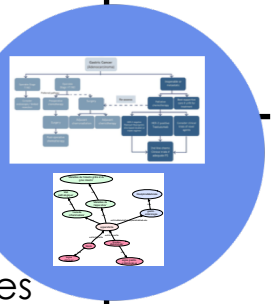
Opérationnaliser les connaissances dans des systèmes d'aide à la décision

Evaluer l'impact et l'amélioration des pratiques

Changer et
Innover
les pratiques

Nouvelles
données

Assembler
Intégrer



Deep learning en santé

- Prevention et prediction de maladies
  - Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau& Sebastian Thrun: Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542*, 115–118 (02 February 2017)
- Diagnostic précoce du cancer de la peau
  - Entrainé sur une base d'apprentissage de ~130,000 images de lésions de la peau couvrant plus de 2000 maladies
  - Certifié par un panel de dermatologistes
- Faire un diagnostic précoce avec un téléphone portable ?

Des opportunités réelles …

- Paul R, Hawkins SH, Schabath MB, et al. Predicting malignant nodules by fusing deep features with classical radiomics features. J Med Imaging. 2018
- Betancur J, Commandeur F, Motlagh M, et al. Deep Learning for Prediction of Obstructive Disease From Fast Myocardial Perfusion SPECT: A Multicenter Study. JACC Cardiovasc Imaging. 2018
- Marcel Adam Just, Lisa Pan, Vladimir L. Cherkassky, et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. Nature Human Behaviour 2017.
- Velly L, Perlbarg V, Boulier T, et al. Use of brain diffusion tensor imaging for the prediction of long-term neurological outcomes in patients after cardiac arrest: a multicentre, international, prospective, observational, cohort study. Lancet Neurol. 2018
- Ting DSW, Cheung CY, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA. 2017 Dec 12; 318(22):2211-2223

… mais des questions

- Les performances n'étaient pas attendues
  - Les algorithmes d'apprentissage profond ont de bonnes performances en généralisation mais on ne sait pas pourquoi.
    - Une boite noire
    - Besoin d'explicabilité pour être capable d'apprendre avec moins de données
- Les algorithmes sont basés sur une quantité importante d'exemples annotés
  - Quel type d'erreurs font-ils ?

Entrepôts de données de santé (EDS)Usages

- Faciliter le pilotage de l'activité hospitalière et l'organisation des soins
- Soutenir la recherche
  - Développer les recherches basées sur la réutilisation des données de santé
    - Réalisation de recherches multicentriques n'impliquant pas la personne humaine
  - Développer des technologies d'optimisation de la recherche clinique
    - Etudes de faisabilité d'essais cliniques (repérage), transfert de données
- Soutenir l'innovation
  - Construire et fiabiliser des jeux de données d'apprentissage
  - Evaluer et valider les technologies/algorithmes d'aide à la décision médicale
  - Développement d'algorithmes IA
    - Essentiellement le domaine de l'épidémiologie clinique

Etat des lieux des EDS

- I2B2: développé à Harvard, gratuit, très utilisé aux États-Unis (80 universités), moins dans le monde (20 institutions dont 3 en France (APHP, Bordeaux et Rennes + Rouen en R/D)

- eHOP: développé à Rennes, 2 installations en France (Rennes et Brest), 4 à venir (Grand Ouest), rien dans le reste du monde.

- Dr WareHouse : développé à Necker ; installé à Saint Anne et Hôpital Foch et à l'IGR

- ConSoRe : Centre de lutte contre le cancer (CLCC)

ConSoRe
Continuum Soins Recherche

- + tous les « faits maisons »

Entrepôt de données de l'APHP

- Données de santé (administratives et médicales) des patients hospitalisés ou venus en consultation au sein des 39 hôpitaux de l'APHP
- Données de plus de 11 millions de patients (2019)

- Détection automatique d'organes au sein d'échographies abdominales (deeplearning, 1 000 000 d'échographies)(HealthData Hub)
- Détection automatique de lésions prostatiques précancéreuses (deeplearning, 3000 IRM prostatiques)

- Modèles prédictifs d'un évènement hypotensif aigu -services de réanimation
- Identification semi-automatique de patients à risque de fracture de faible traumatisme pour inclusion dans une filière de soins (HealthData Hub)

- 70+ projets déposés au CSE (un tiers des projets en appui à l'émergence de l'IA)

Challenges

Des données structurées	Beaucoup de données non structurées
• Diagnostics CIM-10 • Actes CCAM • Examens de Biologie • Prescription / Administration médicamenteuses	• Comptes-rendus (hospitalisation, consultation, urgences, radiologie, etc.) • Champs de questionnaires médicaux ou infirmiers • Commentaires de résultats d'analyses • Prescription / Administration médicamenteuses

- Hétérogénéité des données
  - **Nécessité de données standardisées**
    - *Structuration à la source (terminologies locales, résistance)*
    - *Masse de données textuelles (potentiellement identifiantes)*
  - **Nécessité de prendre en compte la sémantique**
    - Annotation sémantique : méthodes de TAL s'appuyant sur les terminologies/ontologies du domaine
    - Représentation sémantique : méthodes de TAL combinées à des méthodes d'apprentissage
- Echange et partage de données entre applications et établissements
  - Interopérabilité syntaxique et sémantique

LIMICS

Enrichissement sémantique : Annotation

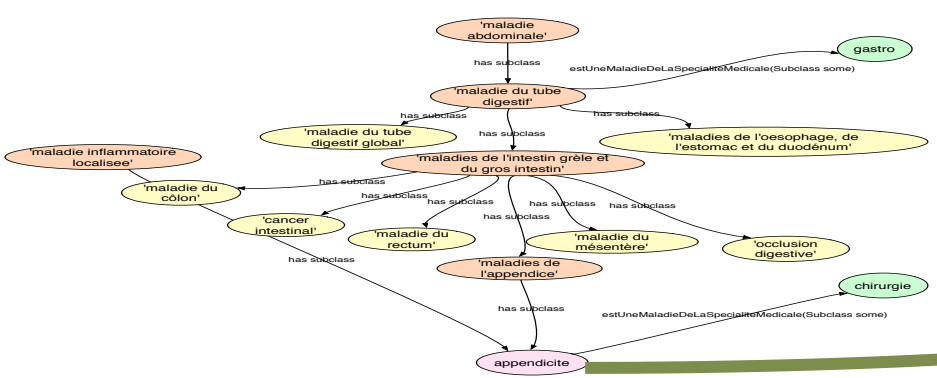
- Utilisation des ressources termino-ontologiques du domaine
  - HETOP : Serveur terminologique (accessible gratuitement www.hetop.eu) 75 terminologies de Santé, 32 langues
  - Projet SIFR, LIRMM
  - Serveur national (ANS)

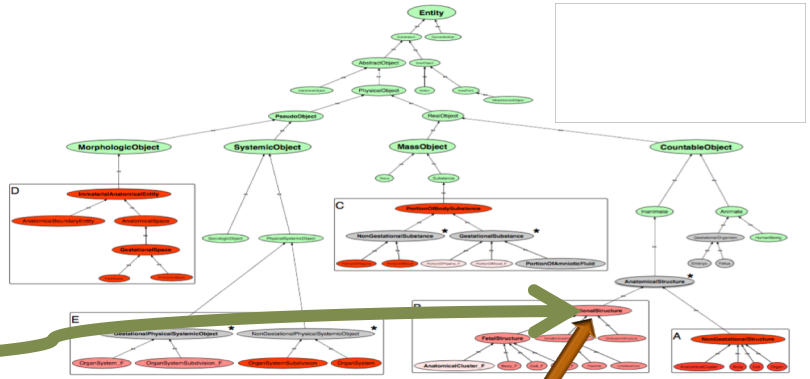
- Approches de traitement automatique des langues pour annoter les textes (identification + normalisation)
  - SIFR BioPortal Annotator, LIRMM (*Jonquet et coll., 2016*)
  - OnBaSAM Extraction de concepts ontologiques à partir de texte
  - ECMT Extracteur de Concept Multi-terminologique CHU de Rouen (*Cabot,2017*)
    - 2 G concepts médicaux extraits sont inclus dans l'entrepôt du CHU de Rouen
      - 16,5 M documents (depuis 2000)
    - Interface Web pour tester, visualiser ou valider les annotations sémantiques

Enrichissement sémantique : représentation

- « Word Embeddings »
  - Faire en sorte que l'ordinateur comprenne par apprentissage les mots utilisés par leur contexte
    - Représenter un mot par un vecteur de nombres réels.
  - Réduire la supervision humaine
  - Apprentissage de représentations vectorielles
    - Réseau de neurones entrainé par de grands volumes
    - Approches : Word2vec; GloVe; FastText
- Document embedding, patient embedding

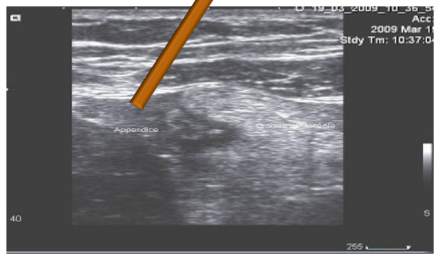
Interopérabilité sémantique pour l'intégration de données hétérogènes





- A pour signe -

plateformes d'interopérabilité sémantique pour l'intégration et le partage de sources de données hétérogènes



Interoperabilité sémantique

- "Standard development organizations" (SDOs) :
  - Modèles d'information (HL7, CEN TC 251, ISO TC 215, DICOM et CDISC)
  - Modèles de connaissances (OMS, IHTSDO, Bioportal, OBO)
- Initiative internationale pour implémenter les standards (IHE)
- De nombreux projets internationaux
  - DebugIT, EHR4CR, TransForm, SALUS, C3-CLOUD
  - Certains « framework » sont basés sur des ontologies pour unifier les modèles d'information et les terminologies (pivots) associées à des tables de « mappings ».

Entre la théorie et la pratique

- La théorie
  - Beaucoup de standards implémentent l'interopérabilité aujourd'hui
  - Les services terminologiques se développent (capitalisation d'éléments de données standards dans des librairies)
  - « The FAIR Guiding Principles » pour l'échange de données
- La pratique
  - Des freins sur l'accès et la qualité des données
  - Utilisation de standards locaux
  - Des alignements manuels (consomateur de temps)

Réutiliser les données de soins pour la recherche clinique

Minimiser la recapture des données

- Objectif : réduire la ressaisie des données de 15%
  - >5M d'items saisies manuellement par étude
- Contexte : données les plus fréquemment collectées dans les essais cliniques
  - 13% à 75% de redondance

☛ Réduire par 15% économiserait des millions d'hommes-jours


Aspects réglementaires

D'après Christel Daniel

En terme d'interopérabilité sémantique …

- Couverture des données
  - Démographie, signes vitaux, laboratoire et médicaments
- Interopérabilité
  - Définition d'un modèle commun (standards), d'alignements et de profils FHIR pour permettre la transformation des données dans un flux de données
- Etude TransFAIR : 1ere étude multi-centrique d'évaluation de DPI utilisés en tant de eSources
  - > 15% des données les plus communes sont collectées semi-automatiquement
  - 6 Essais cliniques
    - 4 oncologie, 2 cardio-vasculaire
  - 3 partenaires EFPIA
    - AstraZeneca, Janssen, Sanofi
  - 3 hôpitaux
    - 12 de octubre(Espagne)
    - IRST (Italie)
    - AP-HP (France)

Conclusion

- Les entrepôts de données sont nécessaires à la mise en place de systèmes de santé auto-apprenants
- Ils supportent l'innovation en santé à partir de données de vie réelle
  - Développement croissant des applications de l'IA (numérique) sur des données contrôlées et structurées
- La standardisation des données pour leur réutilisation est active pour les données structurées (forte communauté)
- L'exploitation des données textuelles est en plein développement grâce à des approches d'IA (enrichissement sémantique)

Merci de votre attention

PubMed.gov
US National Library of Medicine
National Institutes of Health