aviesan
alliance nationale
pour les sciences de la vie et de la santé

Institut Thématique Multi-Organismes
Technologies pour la santé

Raisonner sur les données médico-administratives pour reconstruire et analyser les parcours de soins

T. Guyet, Inria-IRISA/LACODAM

E. Oger, UR1/EA-7449 REPERES - CHU Rennes

Contexte

- Plateforme ANSM - Consortium PEPS – Pharmaco-Epidemiologie des produits de Santé (Financement ANSM)
  - CHU Rennes / CHRU Brest
  - IRISA/Inria
    - Équipes de recherche en informatique : LACODAM, DRUID
    - Axe transversal Biologie – Santé
  - REPERES (UR1/EHESP)
  - IRT B-Com
  - Inserm Equipe ESTHER/CESP Villejuif
- Consortium multi-disciplinaire avec un programme de R&D sur les outils
- **Objectif : Exploiter les données du SNDS pour répondre à des questions de pharmaco-épidemiologie**

Exemple : Etude GENEPI (E. Polard et al.): "*évaluer l'association entre la substitution Princeps-Génériques et les hospitalisations pour épilepsie*"

Apports d'un laboratoire d'informatique dans un consortium pluri-disciplinaire

UMR IRISA
Inria

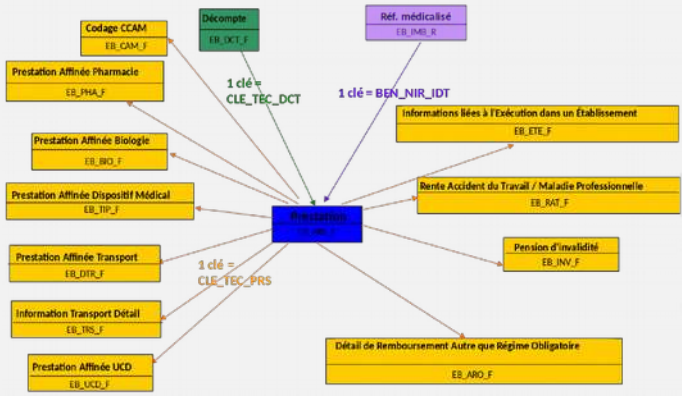
- Inria/IRISA : Laboratoire d'Informatique
- Axe transversal Biologie/Santé de l'IRISA
  - Bioinformatique (plateforme BioGenOuest)
  - Imagerie médicale (plateforme NeurInfo, modélisation physique des organes)
  - Analyse de données en santé, en biologie, en environnement
  - 25 équipes de recherche (~290 chercheurs, ~100 permanents)
- Domaines révolutionnés par le numérique
- Apports méthodologiques en informatique
  - En modélisation (physique, **des connaissances**, des processus)
  - En **analyse de données** (structuration, exploitation, interaction, …)
  - En infrastructure de données et de calcul (gestion, sécurisation, optimisation, …)

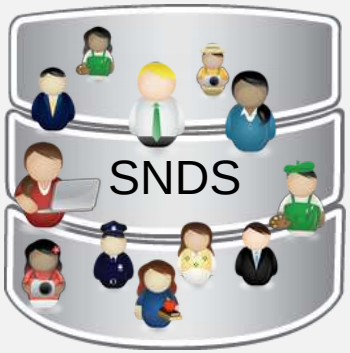
Apports d'un laboratoire d'informatique dans un consortium pluri-disciplinaire

- Inria/IRISA : Laboratoire d'Informatique
- Axe transversal Biologie/Santé de l'IRISA
  - Bioinformatique (plateforme BioGenOuest)
  - Imagerie médicale (plateforme NeurInfo, modélisation physique des organes)
  - Analyse de données en santé, en biologie, en environnement
  - 25 équipes de recherche (~290 chercheurs, ~100 permanents)
- Domaines révolutionnés par le numérique
- Apports méthodologiques en informatique / **Equipe LACODAM**
  - Expertise en modélisation de données, **data science and decision science**
  - Appliquer des **techniques modernes des sciences des données** pour répondre aux besoins la pharmaco-épidémiologie avec le SNDS

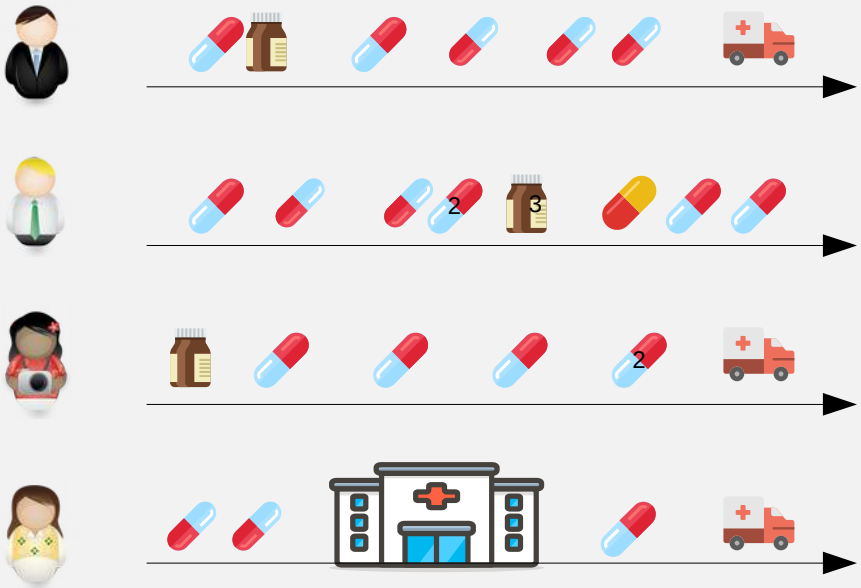
Données médico-administratives et pharmaco-épidemiologie

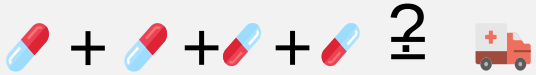
- Sources de données médico-administratives
  - Données SNDS – SNIIRAM
  - Données hospitalières
- Avantage/Inconvénients du SNDS
  - + Données numériques disponibles sans délai de collecte
  - + Population large
  - + Taux de couverture de la population
  - + Données structurées
  - + Données historisées
  - + Possibilité de chaînage avec des données de cohortes, données de capteurs, etc
  - – Des informations partielles
  - – Information médicale limitée
  - – ***Fossé sémantique entre la donnée et l'information recherchée***



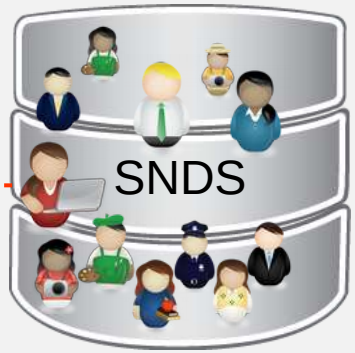


Question de l'étude
Association entre la substitution Princeps-Génériques et les hospitalisations pour épilepsie ??



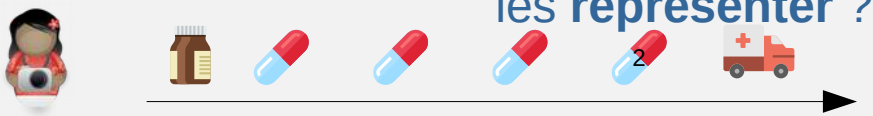


Quels patients **choisirs** ?



Question de l'étude
Association entre la substitution Princeps-Génériques et les hospitalisations pour épilepsie ??

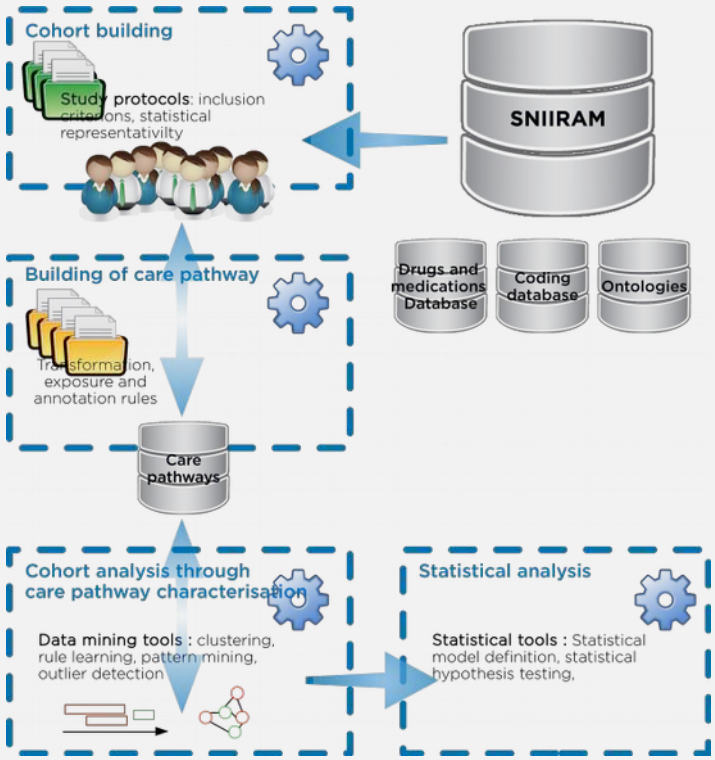
Quels événements utiliser pour décrire le **parcours de soins** ?
Comment le **formaliser**, les **représenter** ?



Quelle question poser ?
Comment **traduire la question** ?
Quelles questions peuvent être **efficacement** posées ?

Données médico-administratives et études épidemiologiques

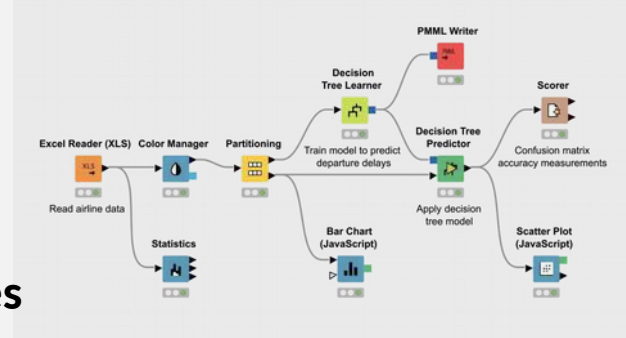
- Sources de données médico-administratives
  - Données SNDS – SNIIRAM
  - Données hospitalières
- *Épidémiologie numérique* : Alternative aux études épidémiologiques classiques
  - Les étapes de l'étude épidémiologique deviennent des étapes d'une analyse de données
    - Sélection de patients
    - Identification d'événements d'intérêts
    - Analyse des liens entre événements d'intérêts
  - Notion centrale de « **parcours de soins** »
  - Une étude est une succession de traitements numériques sur des parcours de soins : ***data science workflow***



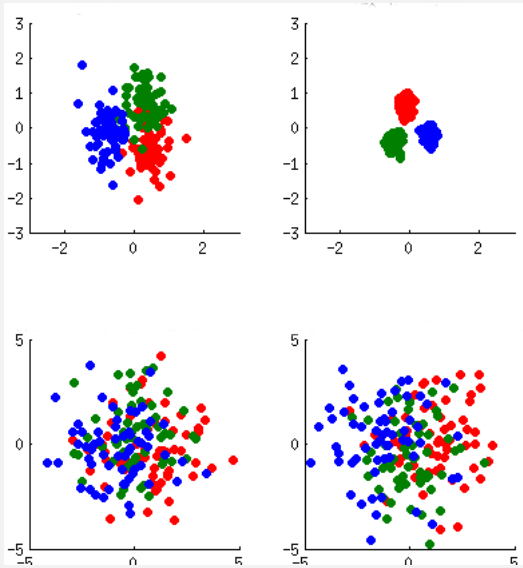
Data science tools and applications on the way to Pharma 4.0, Valentin Steinwandter, Daniel Borchert, Christoph Herwig, Drug Discovery Today 29(9):1795-1805, 2019

Data science workflows / data wrangling

- Pourquoi les méthodes d'apprentissage automatique fonctionnent ?
  - Effort considérable d'annotation des données
  - Beaucoup de la résolution de la tâche d'apprentissage se trouve dans **la préparation des données**



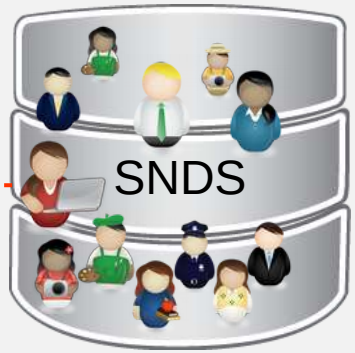
- Généralisation du principe d'une ACP : **un bon espace de représentation des données facilite le travail des algorithmes d'apprentissage**
- La partie la plus importante du workflow est la phase en amont de la tâche d'apprentissage automatique : ***data wrangling***
  - Elle comble en partie le problème du fossé sémantique
  - Qu'en est-il pour les données du SNDS ?



Fossé sémantique

Champs sémantique administratif

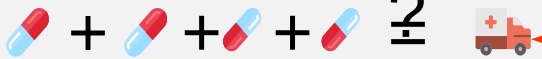
Quels patients **choisirs** ?



Quels événements utiliser pour décrire le **parcours de soins** ?

Champs sémantique médical

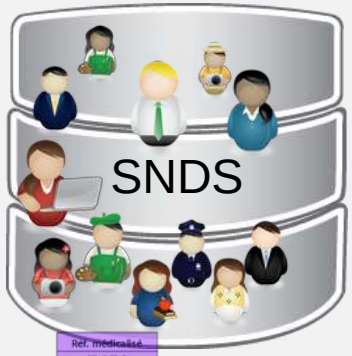
Question de l'étude
Association entre la substitution Princeps-Génériques et les hospitalisations pour épilepsie ??

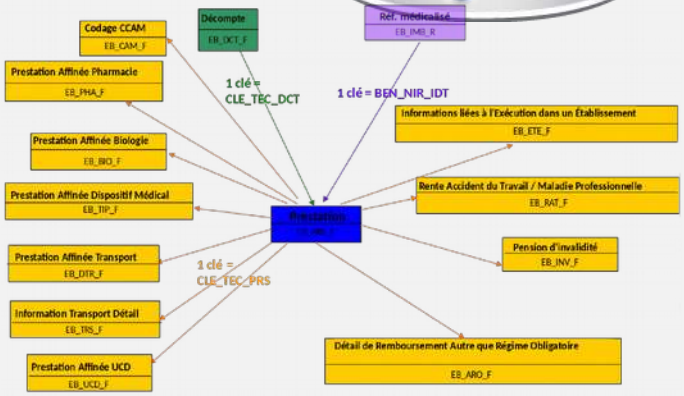


Quelle "**question**" poser ?

Fossé sémantique

Champs sémantique administratif



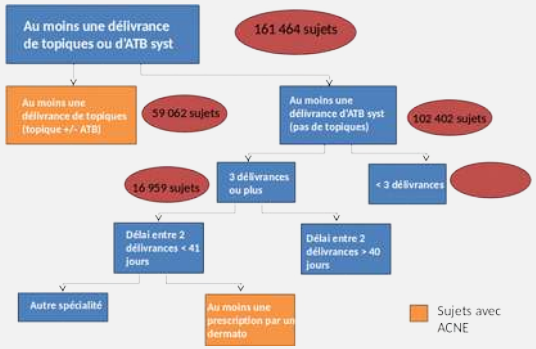


Champs sémantique médical

Question de l'étude
Association entre la substitution Princeps-Génériques et les hospitalisations pour épilepsie ??

```
SELECT
      EXE_SOI_DTD, PHA_PRS_IDE, PFS_PRE_NUM, PFS_EXE_NUM,
      PSE_SPE_COD, PSP_SPE_COD, PSE_ACT_NAT, PSP_ACT_NAT
FROM
      ER_PRS_F as prs,
      ER_PHA_F as pha
WHERE
      (prs.FLX_DIS_DTD = pha.FLX_DIS_DTD) AND (prs.FLX_TRT_DTD = pha.FLX_TRT_DTD) AND
      (prs.FLX_EMT_TYP = pha.FLX_EMT_TYP) AND (prs.FLX_EMT_NUM = pha.FLX_EMT_NUM) AND
      (prs.FLX_EMT_ORD = pha.FLX_EMT_ORD) AND (prs.ORG_CLE_NUM = pha.ORG_CLE_NUM) AND
      (prs.DCT_ORD_NUM = pha.DCT_ORD_NUM) AND (prs.PRS_ORD_NUM = pha.PRS_ORD_NUM) AND
      (prs.REM_TYP_AFF = pha.REM_TYP_AFF) AND
      (BEN_NIR_PSA = '123456789')
ORDER BY
      EXE_SOI_DTD,
      PHA_PRS_IDE;
```

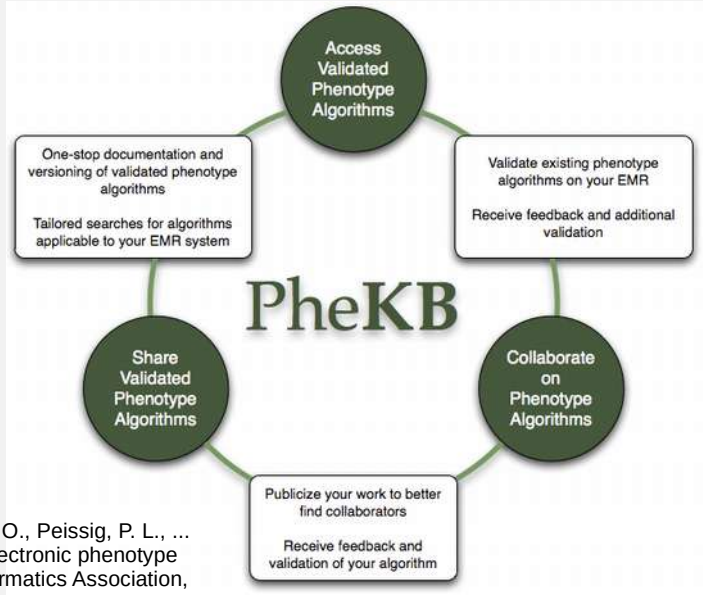
Requête SQL



Fossé sémantique – illustrations pratiques

- Enrichir à l'aide d'**information phénotypique**
  - Données disponibles : des prescriptions de soins
  - Information médicale pertinente : le patient est atteint de telle maladie ou de tel syndrome (sur telle période)
  - Enjeux de précision des algorithmes d'identification

ReDSiam
Réseau pour mieux utiliser les Données du Système national des données de santé



Kirby, J. C., Speltz, P., Rasmussen, L. V., Basford, M., Gottesman, O., Peissig, P. L., ... & Ellis, S. B. (2016). PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association, 23(6), 1046-1052.

Fossé sémantique – illustrations pratiques

- Enrichir à l'aide d'information phénotypique
  - Données disponibles : des prescriptions de soins
  - Information médicale pertinente : le patient est atteinte de telle maladie ou de tel syndrome (sur telle période)
  - Enjeux de précision des algorithmes d'identification
- Reconstruction des **expositions aux médicaments**
  - Données disponibles : délivrances de boîtes de médicament
  - Information médicale pertinente : exposition à une molécule (avec une dose journalière)
  - Quelques difficultés :
    - Multitude de boites pour une même molécule, et pour la même posologie
    - Posologie prescrite inconnue
    - Anticipation ou masquage de délivrances

Problématique et approche

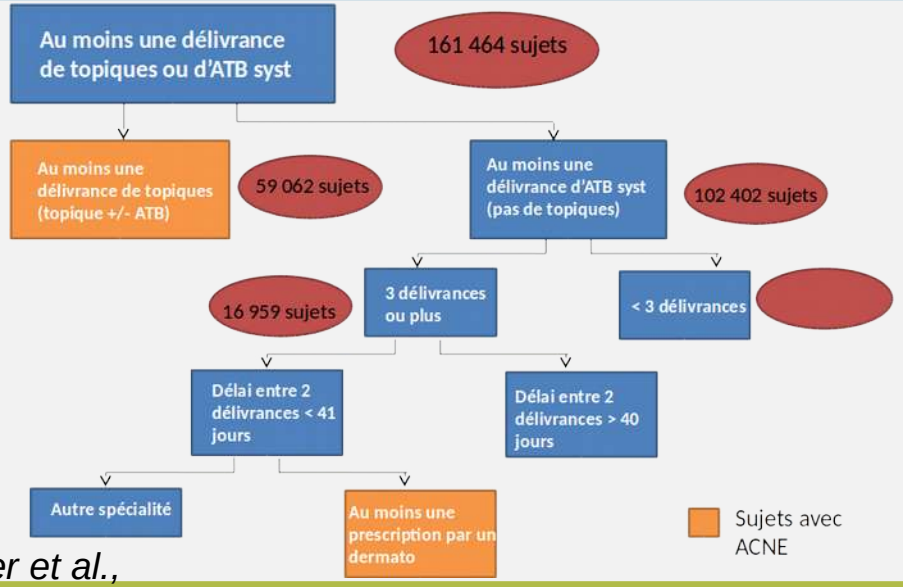
- Proposer des **outils pour faciliter l'enrichissement des données du SNDS** vers de données exploitables par des méthodes d'analyse de données pour en tirer des informations médicales
- Notre approche
  - Modélisation de « parcours de soins » : conserver la richesse de temporelle et descriptive du SNDS
  - Exploiter des connaissances du domaine : intégrer des connaissances formalisées (taxonomies ATC, CIM10, CCAM, …)
  - Approche générique : fournir un cadre pour la conception d'outils répondant à un grand nombre d'études
- Notre proposition
  - Langage de requête hybride dédié aux études de pharmaco-épidémiologies

Langage de requête dédié

- Langage SQL

```
SELECT
      EXE_SOI_DTD, PHA_PRS_IDE, PFS_PRE_NUM, PFS_EXE_NUM,
      PSE_SPE_COD, PSP_SPE_COD, PSE_ACT_NAT, PSP_ACT_NAT
FROM
      ER_PRS_F as prs,
      ER_PHA_F as pha
WHERE
      (prs.FLX_DIS_DTD = pha.FLX_DIS_DTD) AND (prs.FLX_TRT_DTD = pha.FLX_TRT_DTD) AND
      (prs.FLX_EMT_TYP = pha.FLX_EMT_TYP) AND (prs.FLX_EMT_NUM = pha.FLX_EMT_NUM) AND
      (prs.FLX_EMT_ORD = pha.FLX_EMT_ORD) AND (prs.ORG_CLE_NUM = pha.ORG_CLE_NUM) AND
      (prs.DCT_ORD_NUM = pha.DCT_ORD_NUM) AND (prs.PRS_ORD_NUM = pha.PRS_ORD_NUM) AND
      (prs.REM_TYP_AFF = pha.REM_TYP_AFF) AND
      (BEN_NIR_PSA = '123456789')
ORDER BY
      EXE_SOI_DTD,
      PHA_PRS_IDE;
```

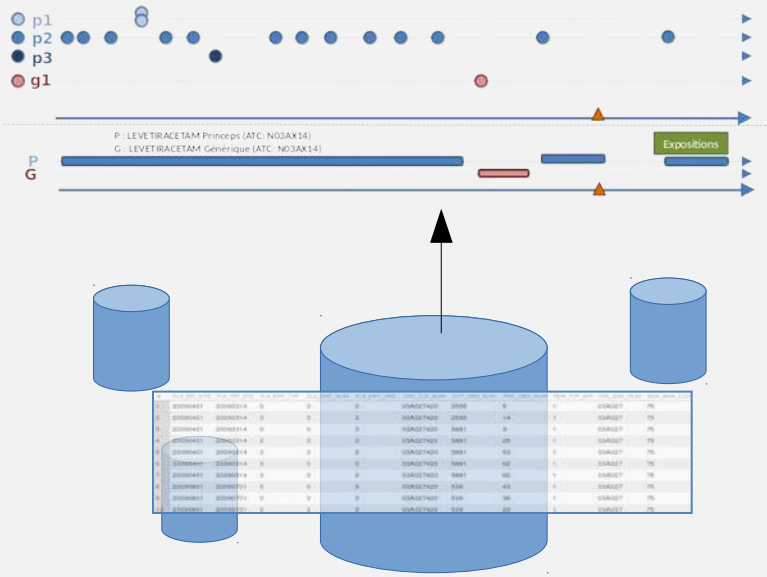
- « Langage » dédié pour la réalisation de la préparation des données de pharmaco-epidemiologie



ISOPSY, E. Oger et al., REPERES

Modèle de parcours de soins

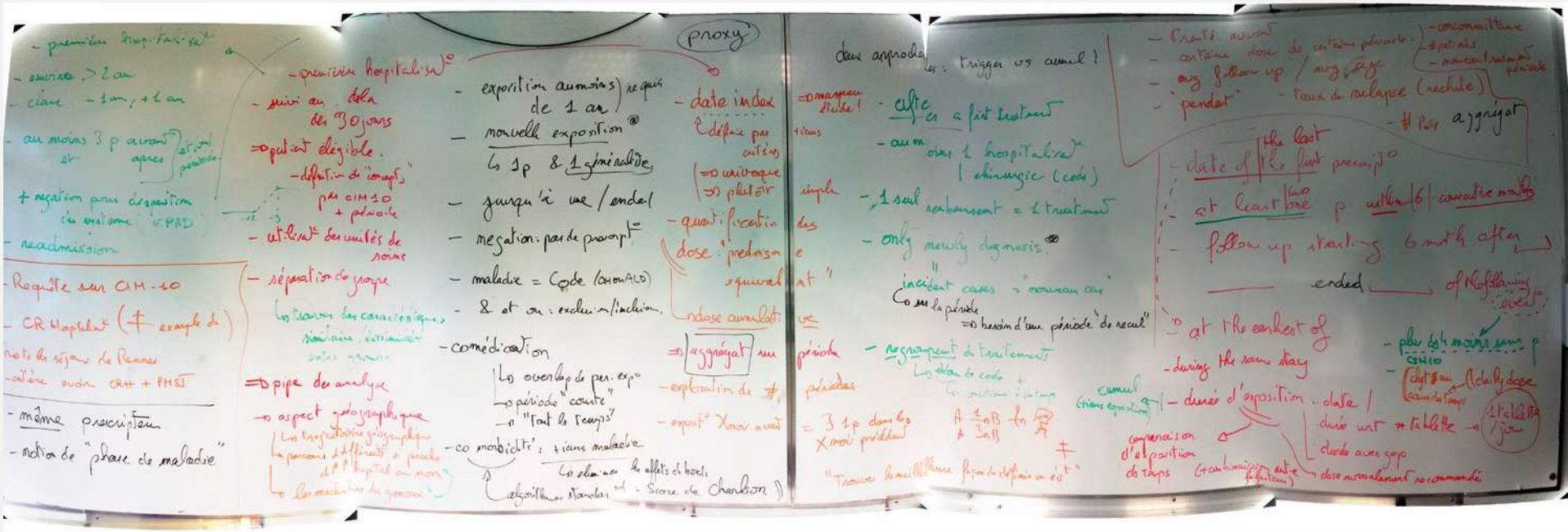
- Modèle de données orienté parcours de soins
  - Collection de patients
  - Pour chaque patient
    - Ensemble d'**évènements datés** et **attribués**
    - Caractéristiques propre à un patient (eg. Statut ALD, Sexe, date de naissance)
- Modèle de données très générique et flexible
- Représentation formelle en logique du premier ordre
  - RDF
  - Datalog



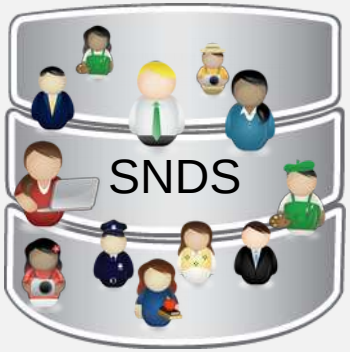

```
sex(m). by(1980). zipcode(25700).
seq(0,0, drug(c01da02,1)).
seq(10,10, drug(r01ad52,3)).
seq(31,31, drug(c07aa07,1)).
seq(8,8, act(hnca006)).
seq(35,39, hosp(k350)).
```

Eléments d'une étude de pharmaco-épidémiologie

- Algèbre d'opérations « élémentaires » sur des « parcours de soins »



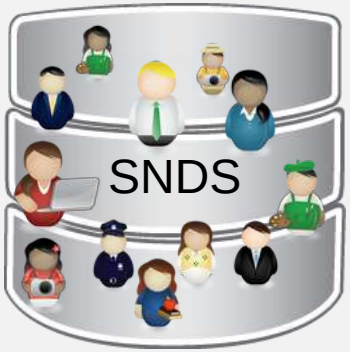
Eléments d'une étude de pharmaco-épidémiologie



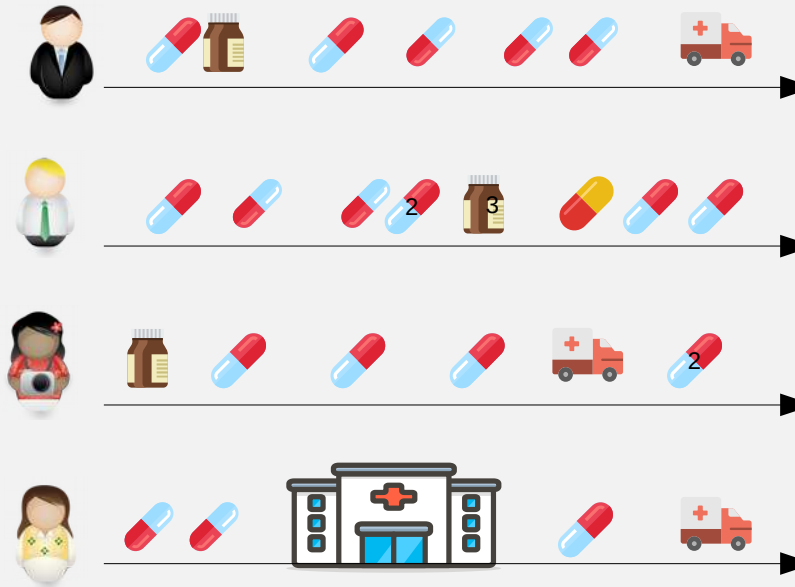
- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Sélection de patients sur un critère $\sigma_{\varphi}(D)$
    - eg. sélection des ***patients épileptiques***
  - Projection d'événements sur un critère attributaires $\pi_{\varphi}(D)$
    - eg. sélection des ***antiépileptiques***
  - Sélection d'événements sur critère temporels $\tau_{\varphi}(D)$
    - eg. événements ***avant une hospitalisation***
  - Induction d'événements $\iota_{\varphi}(D)$
    - eg. ***exposition à un antiepileptique***
  - Labellisation d'un patient sur un critère
    - eg. patients avec switches

$$\lambda(\iota_{\varphi}(\tau_{\varphi}(\pi_{\varphi}(\sigma_{\varphi}(D)))))$$

Eléments d'une étude de pharmaco-épidémiologie

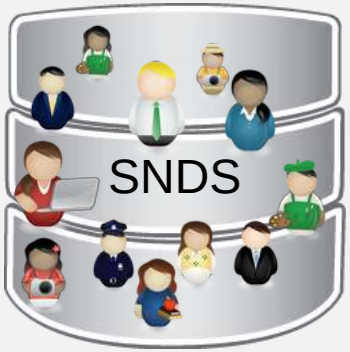


- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Sélection de patients sur un critère $\sigma_\varphi(D)$
    - eg. sélection des ***patients épileptiques***
  - Projection d'événements sur un critère attributaires $\pi_\varphi(D)$
    - eg. sélection des ***antiépileptiques***
  - Sélection d'événements sur critère temporels $\tau_\varphi(D)$
    - eg. événements ***avant une hospitalisation***
  - Induction d'événements $\iota_\varphi(D)$
    - eg. ***exposition à un antiepileptique***
  - Labellisation d'un patient sur un critère
    - eg. patients avec switches

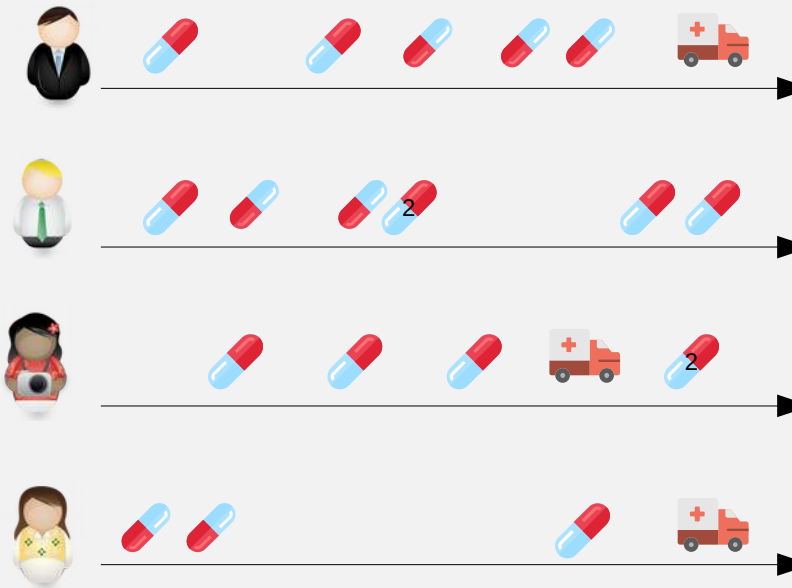


$$\lambda(\iota_\varphi(\tau_\varphi(\pi_\varphi(\sigma_\varphi(D)))))$$

Eléments d'une étude de pharmaco-épidémiologie

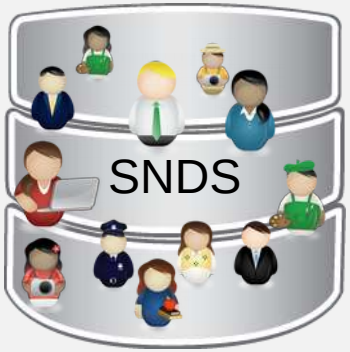


- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Sélection de patients sur un critère $\sigma_{\varphi}(D)$
    - eg. sélection des ***patients épileptiques***
  - Projection d'événements sur un critère attributaires $\pi_{\varphi}(D)$
    - eg. sélection des ***antiépileptiques***
  - Sélection d'événements sur critère temporels $\tau_{\varphi}(D)$
    - eg. événements ***avant une hospitalisation***
  - Induction d'événements $\iota_{\varphi}(D)$
    - eg. ***exposition à un antiepileptique***
  - Labellisation d'un patient sur un critère
    - eg. patients avec switches

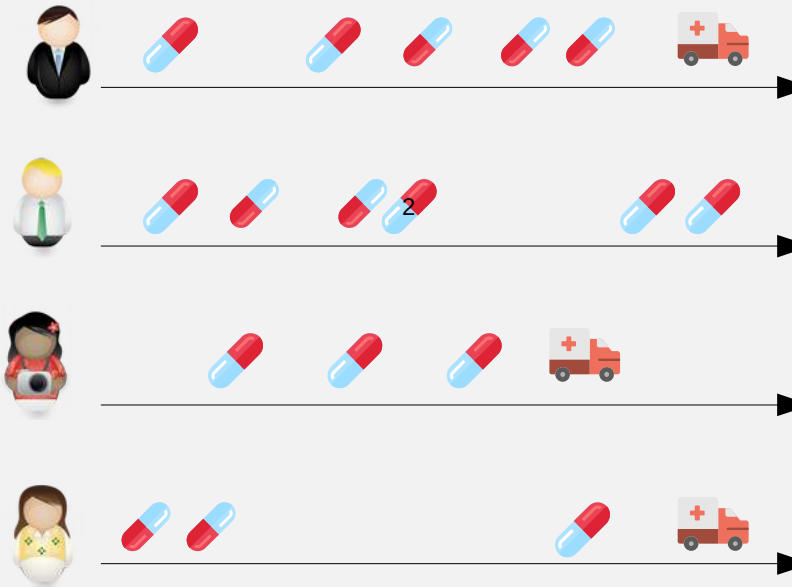


$$\lambda(\iota_{\varphi}(\tau_{\varphi}(\pi_{\varphi}(\sigma_{\varphi}(D)))))$$

Éléments d'une étude de pharmaco-épidémiologie

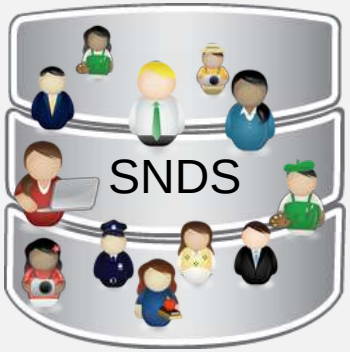


- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Sélection de patients sur un critère $\sigma_\varphi(D)$
    - eg. sélection des ***patients épileptiques***
  - Projection d'événements sur un critère attributaires $\pi_\varphi(D)$
    - eg. sélection des ***antiépileptiques***
  - Sélection d'événements sur critère temporels $\tau_\varphi(D)$
    - eg. événements ***pas après une hospitalisation***
  - Induction d'événements $\iota_\varphi(D)$
    - eg. ***exposition à un antiepileptique***
  - Labellisation d'un patient sur un critère
    - eg. patients avec switches

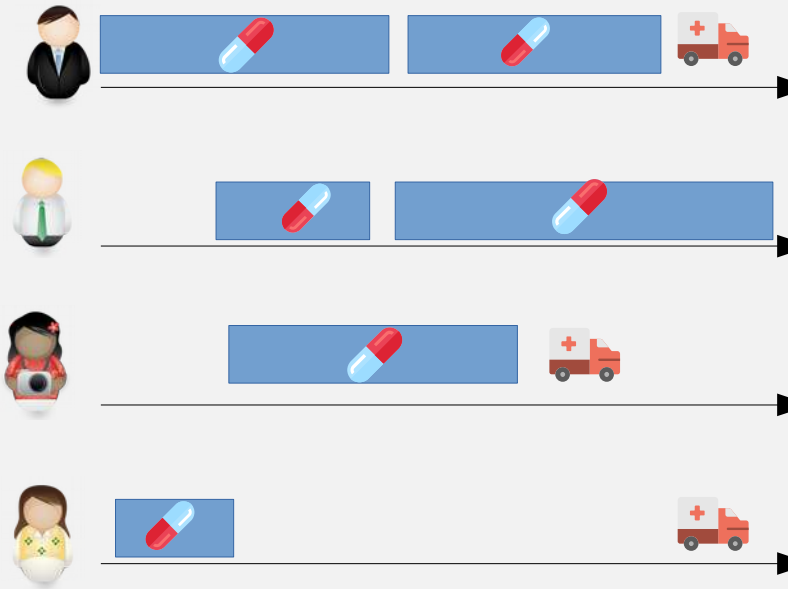


$$\lambda(\iota_\varphi(\tau_\varphi(\pi_\varphi(\sigma_\varphi(D)))))$$

Eléments d'une étude de pharmaco-épidémiologie

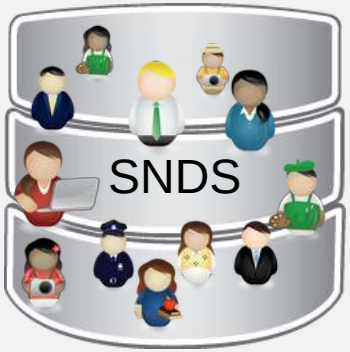


- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Sélection de patients sur un critère $\sigma_\varphi(D)$
    - eg. sélection des ***patients épileptiques***
  - Projection d'événements sur un critère attributaires $\pi_\varphi(D)$
    - eg. sélection des ***antiépileptiques***
  - Sélection d'événements sur critère temporels $\tau_\varphi(D)$
    - eg. événements ***avant une hospitalisation***
  - Induction d'événements $\iota_\varphi(D)$
    - eg. ***exposition à un antiepileptique***
  - Labellisation d'un patient sur un critère
    - eg. patients avec switches

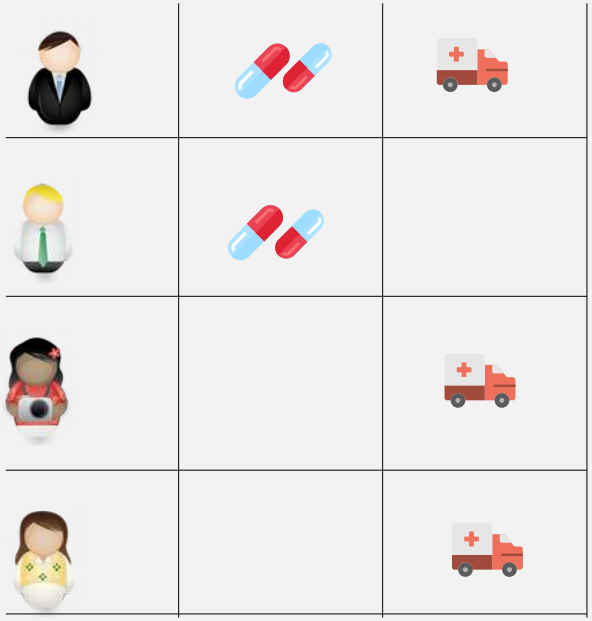


$$\lambda(\iota_\varphi(\tau_\varphi(\pi_\varphi(\sigma_\varphi(D)))))$$

Éléments d'une étude de pharmaco-épidémiologie



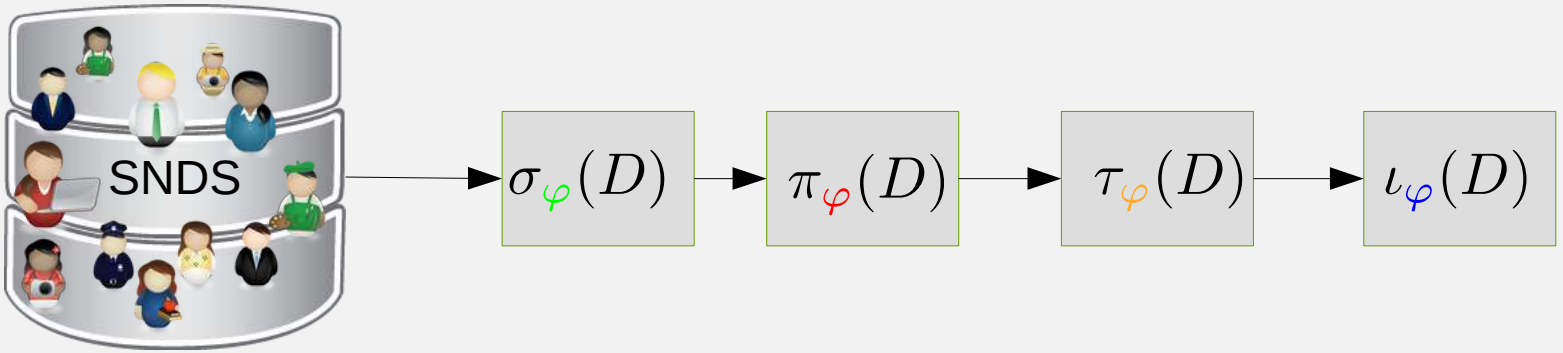
- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Sélection de patients sur un critère $\sigma_{\varphi}(D)$
    - eg. sélection des ***patients épileptiques***
  - Projection d'événements sur un critère attributaires $\pi_{\varphi}(D)$
    - eg. sélection des ***antiépileptiques***
  - Sélection d'événements sur critère temporels $\tau_{\varphi}(D)$
    - eg. événements ***avant une hospitalisation***
  - Induction d'événements $\iota_{\varphi}(D)$
    - eg. ***exposition à un antiepileptique***
  - Labellisation d'un patient sur un critère
    - eg. patients avec switches



$$\lambda(\iota_{\varphi}(\tau_{\varphi}(\pi_{\varphi}(\sigma_{\varphi}(D)))))$$

Éléments d'une étude de pharmaco-épidémiologie

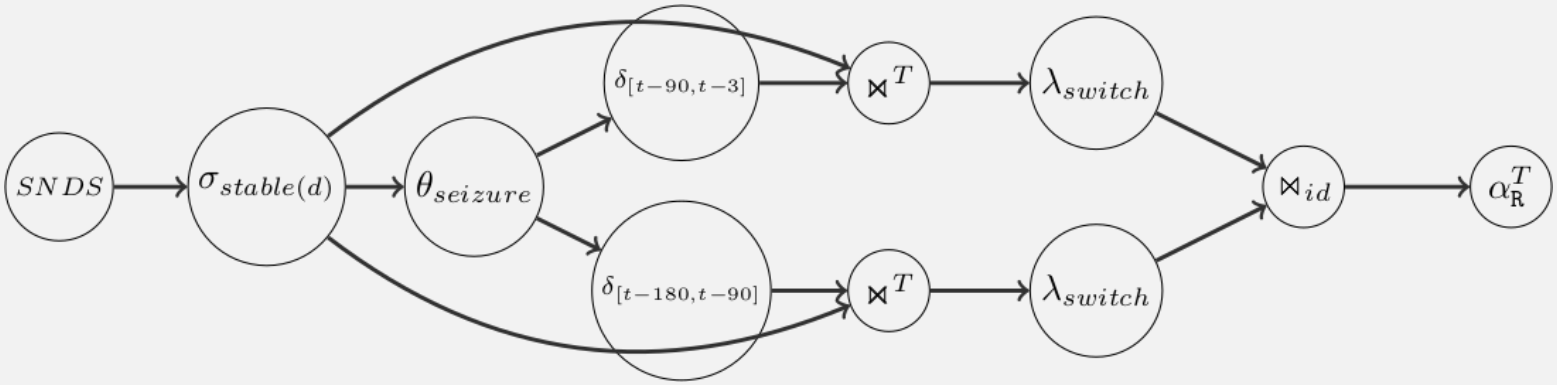
- Algèbre d'opérations « élémentaires » sur des « parcours de soins »
  - Représentation graphique (workflow)



$$\lambda(\iota_\varphi(\tau_\varphi(\pi_\varphi(\sigma_\varphi(D)))))$$

Eléments d'une étude de pharmaco-épidémiologie

- Exemple de l'étude GENEPI
  - Représentation d'une étude de type case-crossover



Critères des opérateurs

- Critères complexes (phénotypiques, informationnels ou inductifs)
  - Besoin d'intégrer des raisonnements complexes sur les parcours de soins
  - Besoin de flexibilité pour exprimer des contraintes variées
- Critères exprimés sous une forme de logique du premier ordre
  - Expressivité élevée
  - Lisibilité et flexibilité de contraintes

$$\sigma_{\varphi}(D) = \{d \in D \mid d, \mathcal{M} \models \varphi\}$$

- Résolution hybride
  - Contraintes exprimées par des programmes logiques
  - utilisation de solveur ASP (clingo)

Gebser, M., Kaminski, R., Kaufmann, B., & Schaub, T. (2014). Clingo= ASP+ control: Preliminary report. arXiv preprint arXiv:1405.3694.

Exemple d'expression de critère

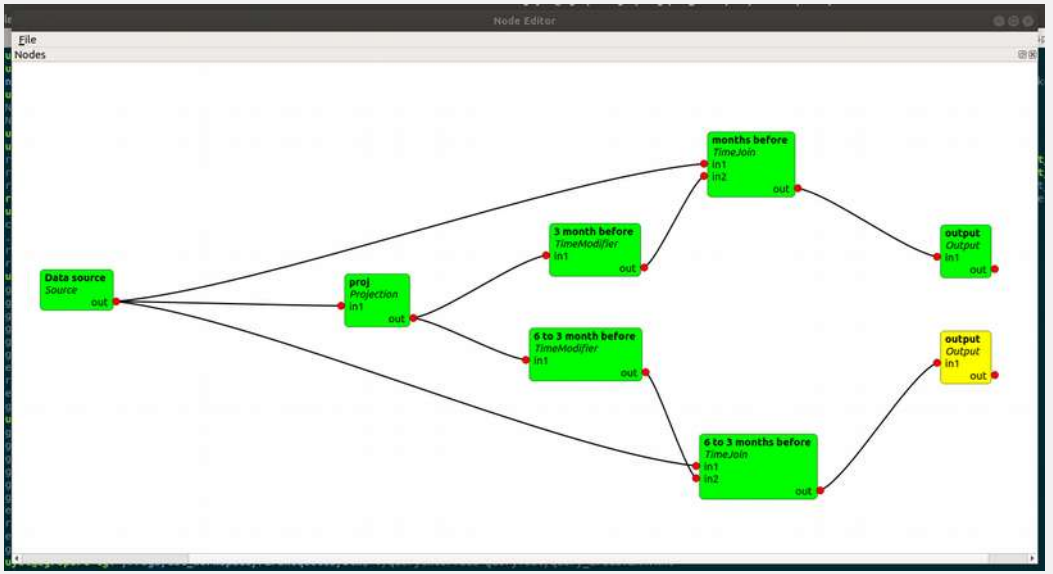
- Période de stabilité d'un traitement antiépileptique : une année sans crise et au moins 10 délivrances d'antiépileptique

φ :

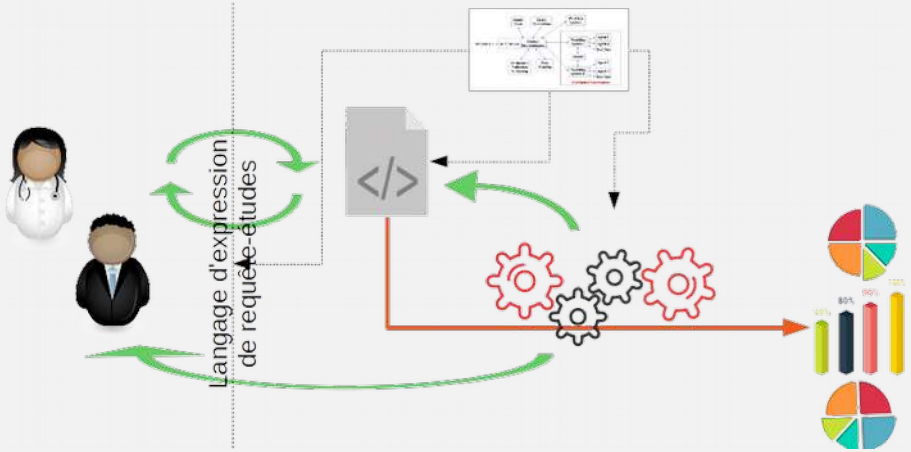
```
#const nb_delivery=10.
#const p_duration=365.

aed(AED) :- AED=(n03ax09;n03ax14;n03ax11;n03ag01;n03af01;n03af02).

%% choice of the index date as a crisis event
1{ end(T) : seq(T, 9999), T>p_duration } 1.

begin(TB) :- end(T), TB=T-p_duration.

%% no crisis within the period
:- 1 { seq(T,9999) : T<E, T>=B }, end(E), begin(B).

%% number of deliveries of class AED
nbprescript(AED,N) :- N=#sum{T:seq(T,C), T<E, T>=B, cip(C,AED) },
begin(B), end(E), aed(AED).

stable :- N>=nb_delivery, nbprescript(AED,N), aed(AED).
:- not stable.
```

Preuve de concept

- Outil basé sur un modèle de données SNDS simplifié
- Implémentation des opérateurs et du moteurs d'exécution de requêtes
  - Interconnexion avec le solveur ASP clingo
- Interface de visualisation des workflows

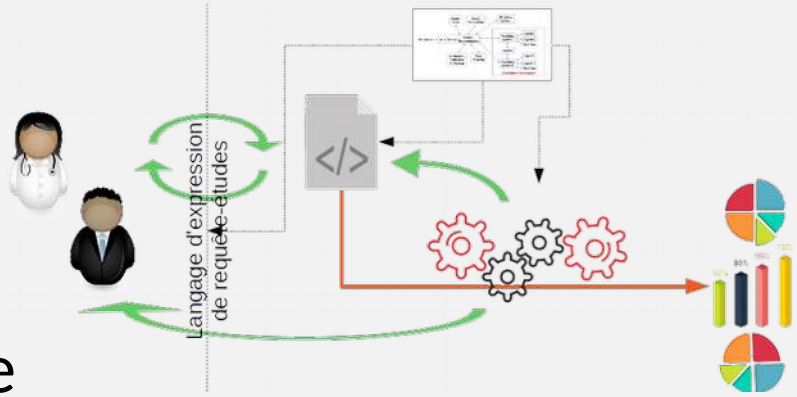


En résumé …



Vers la (semi-)automatisation des études ?

- Pourquoi (semi-)automatiser ?
  - Concentrer les épidémiologistes sur les tâches à valeur ajoutée
  - Faciliter la **réutilisation** et la reproductibilité des études
- Automatisation des études : Déterminer les workflows qui répondent le mieux à une étude
  - Ajout de variables dans les workflow
  - Optimiser les workflows : *Automatic Data Science*



Automating Data Science, Tijl De Bie, Luc De Raedt, Holger H. Hoos, Padhraic Smyth, Dagstuhl Seminar, 2018

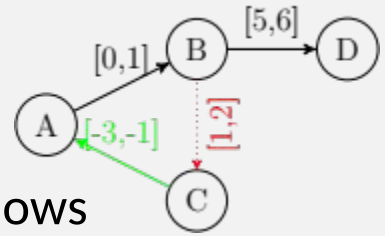
Conclusions

- SNDS : source pour la *pharmaco-epidémiologie numérique*
  - source de données intéressantes
  - Usage secondaire => fossé sémantique
- **Modéliser et raisonner sur les parcours de soins permet de combler (en partie) le fossé sémantique**
  - Besoin d'outils pour faciliter ces manipulations
  - Besoin de méthodologie informatique pour augmenter les possibilités de la pharmaco-épidémiologie numérique
    - Flexibles et expressives : pour exploiter au mieux les données
    - Réutilisables et automatisables : pour focaliser les épidémiologistes sur les questions difficiles
- Techniques récentes de data science pour aborder la pharmaco-epi
- Présentation d'un système hybride
  - Conception d'une requête sous forme algébrique
  - Spécialisation des opérateurs algébriques par des spécifications logiques

Perspectives

- Déterminer des sous-classes de spécialisation d'opérateurs
  - La versatilité perd en performance
  - Identifier des sous-classes de spécialisation utiles et performants
    - Conserver l'expressivité sur les aspects temporels
    - Exploration du concept de chronique pour spécifier des situations d'intérêt

- Besoin d'expérimentations de nos outils
  - Sur son usage
  - Sur l'automatisation (partielle) de l'exécution des workflows
- Aller vers l'analyse automatique des études
  - Une étude dispose d'une représentation formalisée manipulable automatiquement



Merci de votre attention

Remerciements

- Pr. E. Oger
- PharmD. E. Polard
- A. Happe
- Equipe REPERES du CHU et de l'EHESP
- Equipes Inria/IRISA

D. Gross-Amblard, O. Dameron, A. Termier, V. Masson,