Introduction aux Réseaux Bayésiens

Alexandre Aussem

aaussem@univ-lyon1.fr

Thème de recherche : Apprentissage & Extraction de Connaissances

LIESP
Université Lyon 1

Les Modélisation de l'Incertain

- Représenter exhaustivement des distributions multi-dimensionnelles est illusoire
  - Le nombre de paramètres croît exponentiellement avec le nombre de variables aléatoires
- Solution (début 90)
  - Modèles de distributions représentés par des graphes : les **Modèles Graphiques**

Les Modèles Graphiques

- Ce sont des modèles probabilistes novateurs pour la représentation des connaissances, fondés sur une description graphique des variables aléatoires.
- Idée : prendre en compte les *dépendances* et *indépendances conditionnelles* entre les variables
- Objectif : représenter des distributions multi-dimensionnelles de grande taille en évitant l'explosion combinatoire (complexité temporelle et spatiale)

Les Modèles Graphiques

- Deux grandes classes :
  - Les **Réseaux Bayésiens**
    - Représentation asymétrique des dépendances
    - Modélise bien les relations causales (diagnostic)
  - Les **Champs de Markov**
    - Représentation symétrique des dépendances
    - Souvent utilisées pour modéliser les dépendances spatiales (analyse d'image)

Les Réseaux Bayésiens

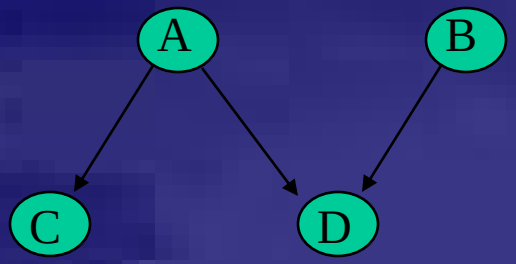
- Représentent et encodent de façon compacte les distributions conjointes de variables aléatoires
- Exploitent les dépendances et indépendances conditionnelles pour éliminer les paramètres superflus
- Si **A** et **B** sont indépendants

  $P(\mathbf{A},\mathbf{B})=P(\mathbf{A})P(\mathbf{B})$
- Si **A** et **B** sont indépendants *conditionnellement* à **C**

  $P(\mathbf{A},\mathbf{B}|\mathbf{C})=P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})$

Les Réseaux Bayésiens

- Modèles de représentation des connaissances, fondés sur une description graphique des variables aléatoires : **Directed Acyclic Graph** (DAG)



- Les nœuds sont les variables aléatoires et les arcs sont les relations (si possibles) causales entre ces variables
- L'absence d'arc signifie une indépendance conditionnelle

Tables de probabilités

- Dans chaque nœud, on stocke la table de probabilités conditionnelles locale $P(X_i|Pa_i)$ pour chaque configuration des parents Pa_i du nœud X_i

$P(D|A,B)$

A	B	True	False
False	False	0.4	0.6
False	True	0.1	0.9
True	False	0.7	0.3
True	True	0.6	0.4

Les Réseaux Bayésiens

- On dira que le couple {G,P} est un Réseau Bayésien, avec G={V,E} un DAG, s'il vérifie la **condition de Markov** : chaque variable X dans V est indépendante de ses non descendantes (ND_X) dans G conditionnellement à ses parents :

$$Ind_P(X, ND_X / Pa_X)$$

où Pa_i désigne l'ensemble des parents de X_i dans G.

- La condition de Markov implique la factorisation de la loi jointe :

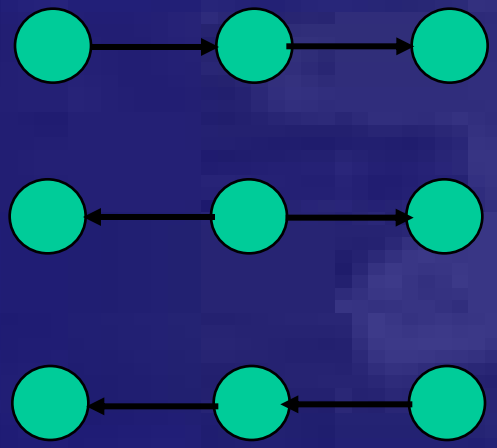
$$P(X_1,..., X_n) = \prod_{i=1}^{n} P(X_i / Pa_i)$$

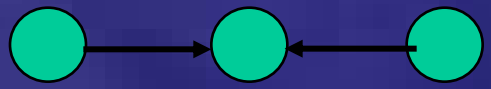
Les Réseaux Bayésiens

- Cette propriété importante montre qu'il suffit de stocker les valeurs de $P(X_i|Pa_i)$ pour toutes les valeurs de X_i et les possibles instanciations conjointes de Pa_i dans une table de probabilités.
- Toute requête portant sur une ou plusieurs variables d'intérêt conditionnellement à d'autres (les observations partielles) peuvent être obtenue par inférence.

Les Réseaux Bayésiens

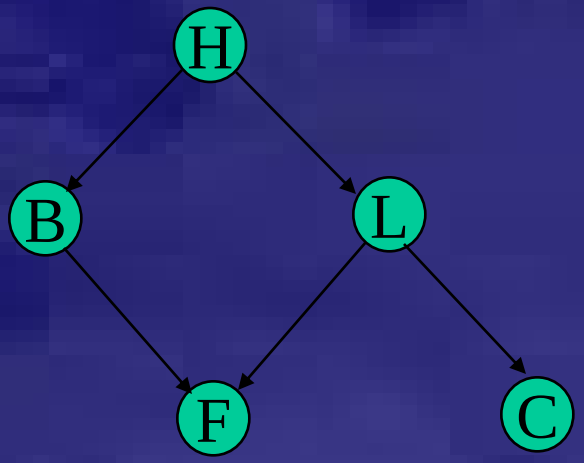
- Plusieurs types de chaînes sont possibles entre 3 variables :





Connexion convergente

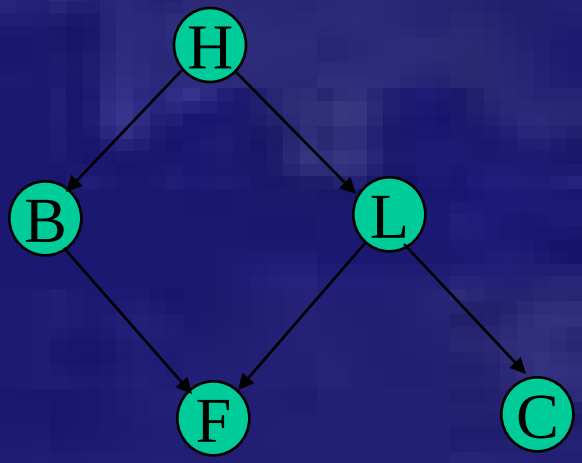
Exemple Illustratif



$$P(F,C,B,L,H) = P(F|B,L)P(C|L)P(B|H)P(L|H)P(H)$$

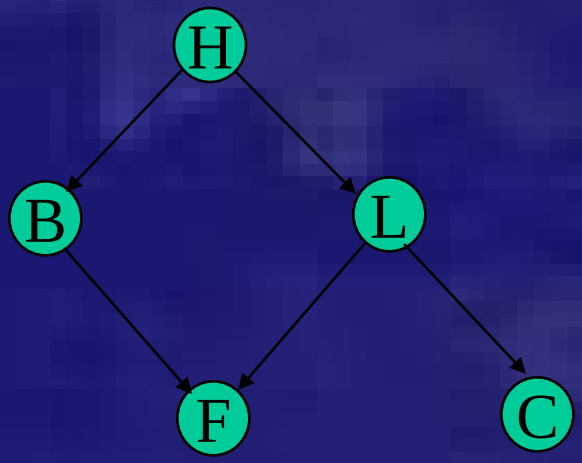
La structure du graphe implique la décomposition !

Exemple Illustratif



Noeud	Parents	Indépendence Conditionnelle
C	{L}	$I_p(C,\{H,B,F\}\|\{L\})$
B	{H}	$I_p(B,\{L,C\}\|\{H\})$
F	{B,L}	$I_p(F,\{H,C\}\|\{B,L\})$
L	{H}	$I_p(L,\{B\}\|\{H\})$

Complexité spatiale



Taille mémoire de la loi conjointe : $2^5-1=31$

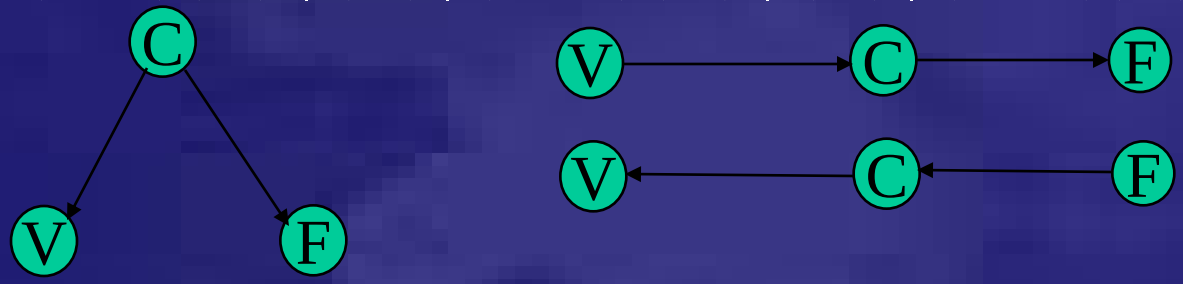
Taille mémoire du réseau bayésien : $1+2+2+4+2=11$

Dans le premiers cas, taille exponentielle, dans le second, taille linéaire avec le nombre de noeuds !

Illustration

- P uniforme. $Ind_P(V,F|C)$ mais pas $Ind_P(V,F)$

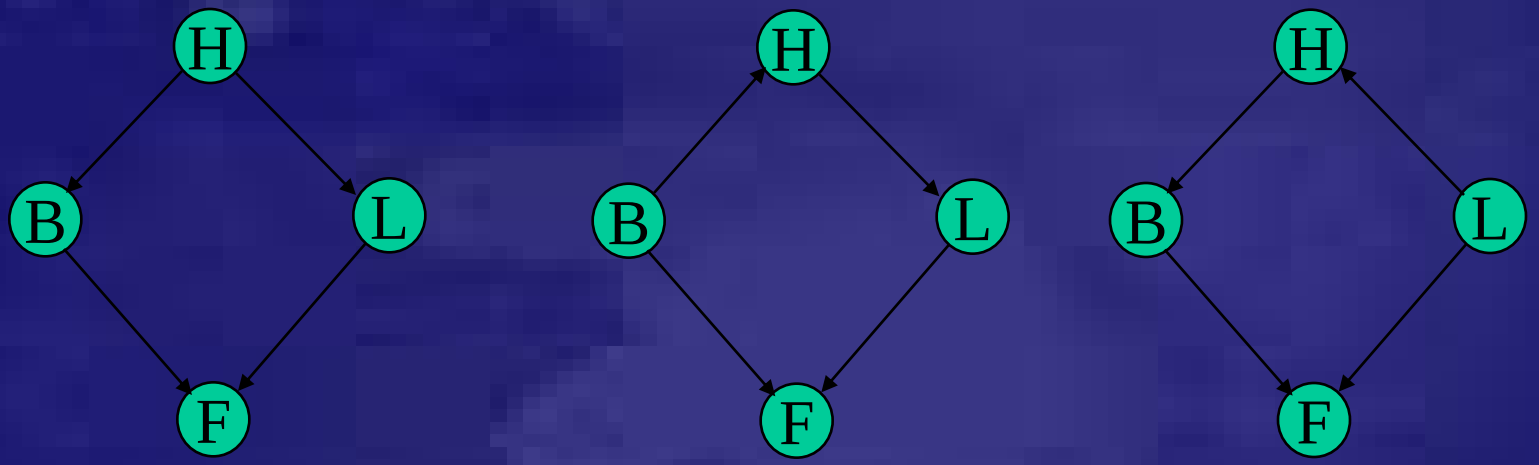
$$P(V,C,F) = P(C)P(V|C)P(F|C) = P(V)P(C|V)P(F|C) = P(F)P(C|F)P(V|C)$$



P vérifie la condition de Markov avec ces 3 DAG; ils encodent la même distribution de probabilité !

Equivalence de Markov

- Les DAG qui encodent les mêmes indépendences. Ils forment une classe d'équivalence.



Impossible de les distinguer à partir des données !

La causalité

• On cherche idéalement à trouver les causes des phénomènes observés

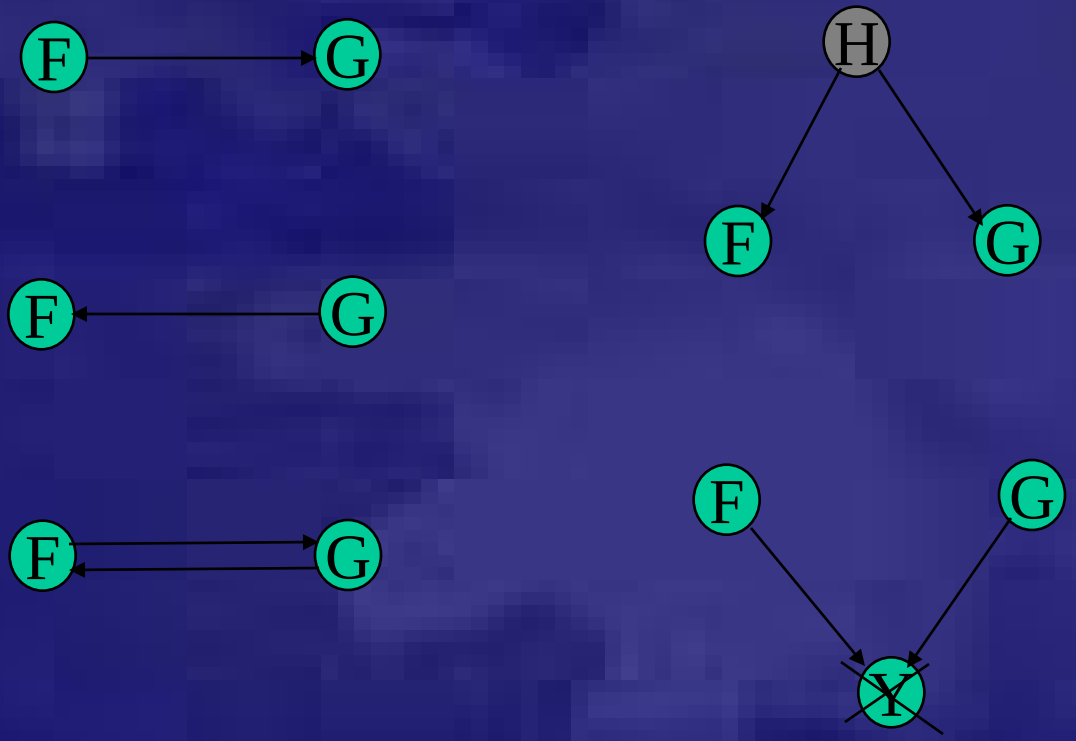
• La causalité est une notion sujette à caution, elle fait l'objet de controverses depuis des siècles, Hume 1748, Piaget 1966 etc..

• Définition "**opérationnelle**" : Si une manipulation particulière de X, en forçant sa valeur, provoque la modification de la distribution de Y, alors X cause Y.

• On suppose ici que cause et effets sont corrélés

Corrélation et causalité

Si F et G sont correlés, F est-il la cause de G ?



La causalité

- F = Finasteride (médicament) et G =Hair Growth. Une population d'individus est observée.
- Cas 2 - Un autre produit X a eu de l'effet et suscite l'intérêt de l'individu pour F
- Cas 3 – F a de l'effet et suscite un intérêt pour F en retour.
- Cas 4 – H est l'inquietude de l'individu, lequel prend 2 produits F et X. Seul X a un effet sur G.
- Cas 5 – F et G suscitent une hypertension Y. Or, la population observée a de l'hypertension. Y est alors instanciée (biais de sélection).

Contraintes sur le graphe

- La d-séparation est un critère important qui permet de caractériser graphiquement toutes les contraintes d'indépendance des lois P qui peuvent être représentées par une même DAG.
- Il faut introduire la notion de graphes :
  - Chaînes & chemins, simples ou composés,
  - Parents, enfants, descendants, ancètres etc.
- Par une chaîne de X vers Y transite une information bruitée : les sommets sont des vannes ouverts ou fermées.

D-séparation

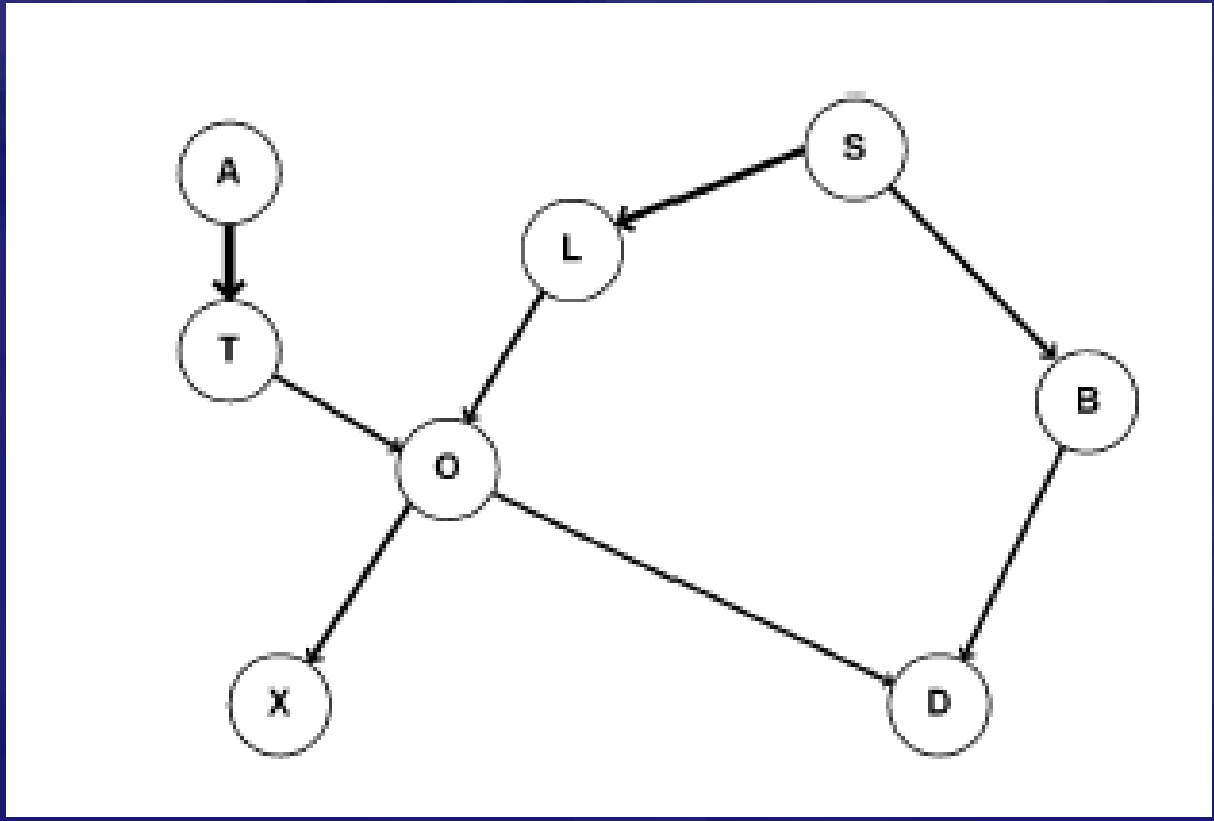
- Une chaîne est dite **ouverte** si toutes les vannes sont ouvertes auquel cas la chaîne laisse passer l'information.
- A l'inverse, si l'une des vanne est bloquée, la chaîne est dite **fermée**.
- l'information qu'apporte X sur Y peut se voir comme la somme des **flots d'information** sur tous les chaînes ouvertes reliant X à Y.
- Mécanisme d'ouverture et de fermeture des vannes ?

D-séparation

- Formellement, un chaîne entre X et Y est **fermée** par un ensemble de noeuds **Z** s'il existe un noeud W sur cette chaîne vérifiant l'une des conditions :
  - W n'est pas un noeud convergent : W est dans **Z**
  - W est un noeud convergent : W, ou un de ses descendants est dans **Z**.
- Deux noeuds X et Y sont dits *d-séparés* par **Z** dans le graphe G, $Dsep_G(X;Y|\mathbf{Z})$, si tous les chaînes (simples) entre X et Y sont fermées par **Z**.
- La d-séparation dresse un parallèle élégant entre l'algorithmique des graphes et le calcul des indépendances conditionnelles dans une distribution de probabilités.

Illustration

Réseau de la dyspnée : *Asia*



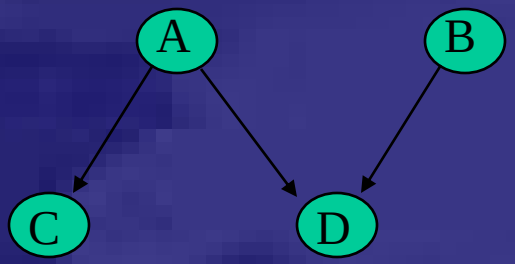
Inférence & Diagnostic

Information ⟹ **Connaissance**

- Permet d'**inférer des connaissances**, e.g. P(**D**|**C**,**B**), à partir d'observations partielles, bruitées etc.
- Représentation graphique des connaissances lisible par les non specialistes (e.g., médecins)
- Autorise des requêtes probabilistes du type : Quelle probabilité de telle maladie sachant tels symptômes.
- Permet de hiérarchiser les diagnostics (simples ou multiples) en fonction des probabilités *a posteriori* .

Schémas d'Inférence

- P(**D**) et P(**D**|**B**) : calcul d'inférence causal
- Comparer P(**B**|**C**,**D**) et P(**B**|**D**) : calcul de diagnostic
- Problèmes NP-difficiles pour des ensembles de variables quelconques



100 observations

A	B	C	D	nombre
VRAI	VRAI	VRAI	VRAI	54
VRAI	FAUX	VRAI	VRAI	1
VRAI	VRAI	VRAI	FAUX	7
VRAI	FAUX	VRAI	FAUX	27
VRAI	VRAI	FAUX	VRAI	3
VRAI	FAUX	FAUX	FAUX	2
FAUX	VRAI	FAUX	VRAI	4
FAUX	FAUX	FAUX	FAUX	2

Illustration : Diagnostic Médical

- Support formel
  = Graphe orienté acyclique
  + Tables de probabilités conditionnelles
  = Réseau Bayésien

- État de connaissance courant
  = Distributions de probabilités, une par variable, dérivées des tables de probabilités conditionnelles

- Prise en compte d'observations en vue d'établir un diagnostic

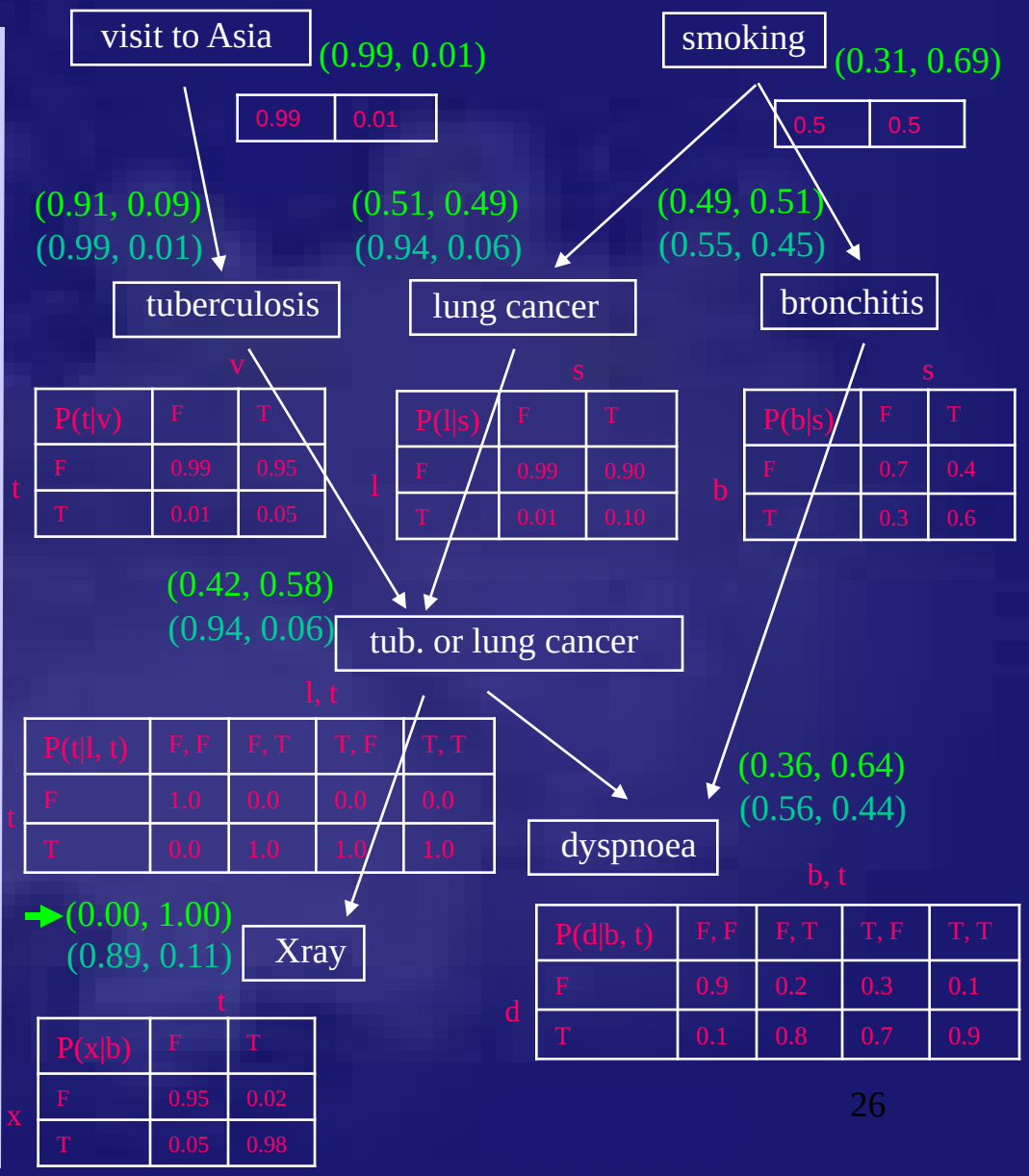
  => Injection de nouvelles connaissances, ex., la radiographie des poumons est positive

  => Perturbation de l'état de connaissance

  => Retour à l'équilibre de l'état de connaissance (inférence)

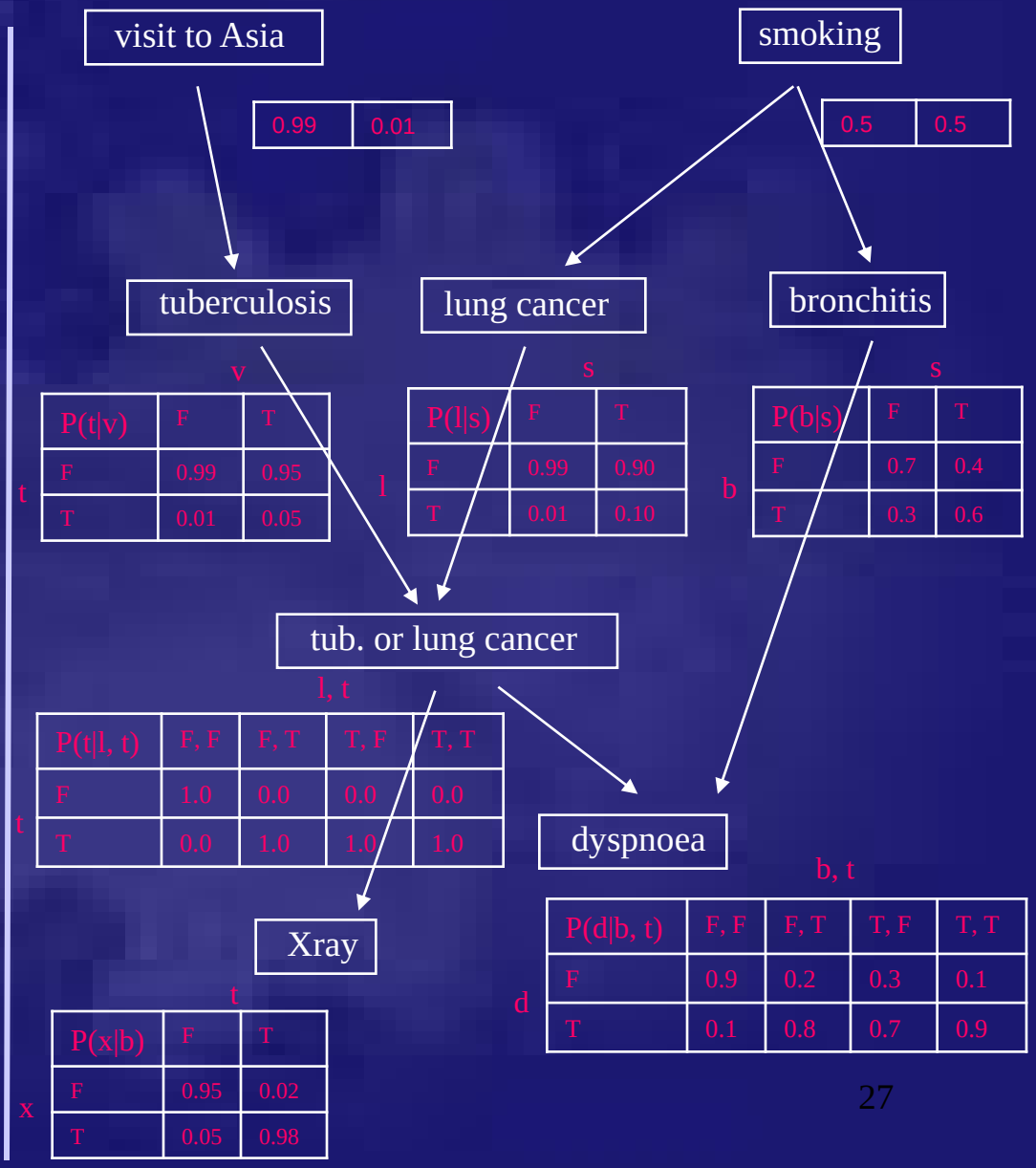
  => Nouvel état de connaissance

  => Utilisation de l'état courant ou du différentiel par rapport aux états précédents pour établir un diagnostic



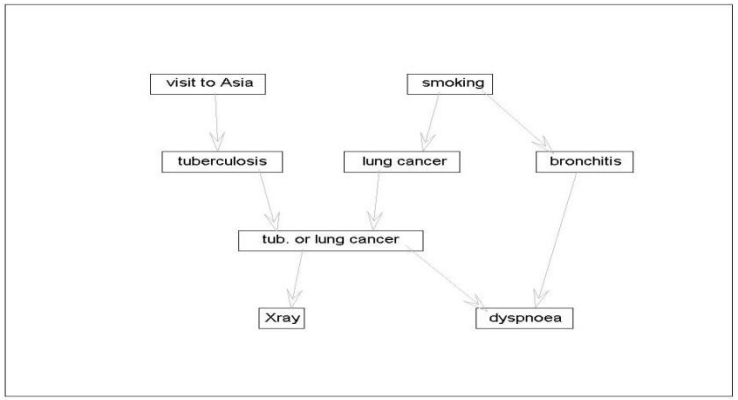
Comment faire l'apprentissage du modèle ?

- Construction du modèle graphique
  - Identification des variables et des domaines de valeurs qui leur sont associés
  - Construction du graphe causal à l'aide d'un expert du domaine ou par **apprentissage automatique des causalités**
- Construction par **apprentissage** des tables de probabilités conditionnelles
- Validation du modèle : par des experts et/ou par des techniques statistiques.
- Construction de l'interface graphique conviviale
  - Définition des fonctionnalités
  - Développement logiciel



Comment faire un diagnostic?

Simulation (prototype construit et résultats obtenus avec le logiciel numérique : Matlab + toolbox BNT)



evidence setting:
U , T , U , U , U , U , T , U

	NO	YES
visit to Asia :	0.98782	0.012185
smoking :	0	1
tuberculosis :	0.93282	0.067183
lung cancer :	0.35401	0.64599
bronchitis :	0.4	0.6
tub. or lung cancer :	0.29354	0.70646
Xray :	0	1
dyspnoea :	0.26806	0.73194

evidence setting:
T , U , U , U , U , U , T , U

	NO	YES
visit to Asia :	0	1
smoking :	0.36299	0.63701
tuberculosis :	0.66228	0.33772
lung cancer :	0.62851	0.37149
bronchitis :	0.5089	0.4911
tub. or lung cancer :	0.30937	0.69063
Xray :	0	1
dyspnoea :	0.3189	0.6811

evidence setting:
U , U , U , U , U , U , U , U

	NO	YES
visit to Asia :	0.99	0.01
smoking :	0.5	0.5
tuberculosis :	0.9896	0.0104
lung cancer :	0.945	0.055
bronchitis :	0.55	0.45
tub. or lung cancer :	0.93517	0.064828
Xray :	0.88971	0.11029
dyspnoea :	0.56403	0.43597

evidence setting:
U , U , U , U , U , U , U , T

	NO	YES
visit to Asia :	0.98968	0.010325
smoking :	0.366	0.634
tuberculosis :	0.98115	0.018845
lung cancer :	0.89724	0.10276
bronchitis :	0.16603	0.83397
tub. or lung cancer :	0.87946	0.12054
Xray :	0.8379	0.1621
dyspnoea :	0	1

evidence setting:
U , T , U , U , U , U , U , T

	NO	YES
visit to Asia :	0.98981	0.010193
smoking :	0	1
tuberculosis :	0.98457	0.015427
lung cancer :	0.85167	0.14833
bronchitis :	0.11984	0.88016
tub. or lung cancer :	0.83778	0.16222
Xray :	0.79914	0.20086
dyspnoea :	0	1

evidence setting:
T , U , U , U , U , U , U , T

	NO	YES
visit to Asia :	0	1
smoking :	0.37408	0.62592
tuberculosis :	0.91225	0.087751
lung cancer :	0.90047	0.099525
bronchitis :	0.1886	0.8114
tub. or lung cancer :	0.8177	0.1823
Xray :	0.78046	0.21954
dyspnoea :	0	1

evidence setting:
U , U , U , U , U , U , T , U

	NO	YES
visit to Asia :	0.98684	0.013156
smoking :	0.31225	0.68775
tuberculosis :	0.90759	0.092411
lung cancer :	0.51129	0.48871
bronchitis :	0.49367	0.50633
tub. or lung cancer :	0.42396	0.57604
Xray :	0	1
dyspnoea :	0.35923	0.64077

evidence setting:
U , U , U , U , U , U , T , T

	NO	YES
visit to Asia :	0.98602	0.013984
smoking :	0.21439	0.78561
tuberculosis :	0.88607	0.11393
lung cancer :	0.37875	0.62125
bronchitis :	0.31813	0.68187
tub. or lung cancer :	0.27127	0.72873
Xray :	0	1
dyspnoea :	0	1

evidence setting:
U , T , U , U , U , U , T , T

	NO	YES
visit to Asia :	0.9875	0.012496
smoking :	0	1
tuberculosis :	0.92473	0.075266
lung cancer :	0.27629	0.72371
bronchitis :	0.28629	0.71371
tub. or lung cancer :	0.20855	0.79145
Xray :	0	1
dyspnoea :	0	1

evidence setting:
T , U , U , U , U , U , T , T

	NO	YES
visit to Asia :	0	1
smoking :	0.29797	0.70203
tuberculosis :	0.60829	0.39171
lung cancer :	0.55573	0.44427
bronchitis :	0.37118	0.62882
tub. or lung cancer :	0.18623	0.81377
Xray :	0	1
dyspnoea :	0	1

Apprentissage des RB

- L’apprentissage de la **structure** à partir de données est NP-difficile : la taille de l'espace des DAG est super-exponentielle en fonction du nombre de variables.

  Par exemple, T(5) = 29281

- L’apprentissage des **tables de probabilités** est aisé, simple calcul fréquentiel (polynomial)
- L’**inférence** exacte est NP-difficile, de même que l’inférence approchée…

Apprentissage du DAG

- Deux grandes familles de méthodes existent pour l'apprentissage du DAG :
  - celles fondées sur **la satisfaction de contraintes d'indépendance conditionnelle** entre variables
  - celles dites "`bayésiennes'" fondées sur la **maximisation d'un score** (BIC, MDL, BDe, etc.).
- Les méthodes sous contraintes sont **déterministes**, relativement rapides et bénéficient des critère d'arrêts clairement définis. Utilisent des informations statistiques dans les données (niveau de signification arbitraire). Les erreurs commises au début se répercutent en cascade.
- Les méthodes à base de score incorporent des probabilités a priori sur la structure du graphe, traitent plus facilement les **données manquantes**, mais sont facilement piégées dans les **minima locaux**. Le graphe final obtenu dépend des conditions initiales.

Inférence

- Technique d'envoie asynchrone de messages jusqu'à équilibre si le DAG est un arbre.
- Cut-set conditioning : instancier des variables pour que le graphe restant soit un arbre.
- Algorithme de l'arbre de jonction.
- Méthodes d'approximation : effectuent des échantillonnage de Gibbs sur des sous-ensembles de variables

Arbre de jonction

- Algorithme de l'arbre de jonction [Jensen90]

  – Phase de contruction

  Transforme le graphe en un arbre de jonction (arbre couvrant minimal) dont les nouveaux nœuds sont des clusters des nœuds initiaux.

  1 : Moralisation (marier tous les parents)

  2 : Triangulation et extraction de cliques (les voisins sont connectés deux à deux, ils appartiennent à des cliques qui formeront les noeuds du futur arbre de jonction)

  3 : Création d'un arbre couvrant minimal (arbre de jonction)

  – Propagation et mise à jour de la distribution

Applications des RB

- Biologie & Santé & Epidémiologie : p.ex. réseau de régulation entre gènes à partir des puces à ADN,
- Finance : analyse de risques de prêts, détection des mauvais payeurs.
- Vision : reconnaissance de visage, traffic routier
- Diagnostic de pannes/bugs/maladies : Microsoft, Intel, Ricoh…
- Traitement du langage
- Prévision, classification, datamining, marqueting etc.

Application à la sélection de variables

- We are interested in solving the **Feature Subset Selection problem** in data sets with **thousands** of discrete **variables.**
- Such databases are common in domains like bioinformatics (e.g., gene expression databases) and medicine.
- A principled solution to this problem is to determine a **Markov boundary** (MB) of the class variable.

Definition: The Markov boundary of T, denoted by MB(T), is *any minimal subset* of **V** (the full set) that renders the rest of **V** independent of T.

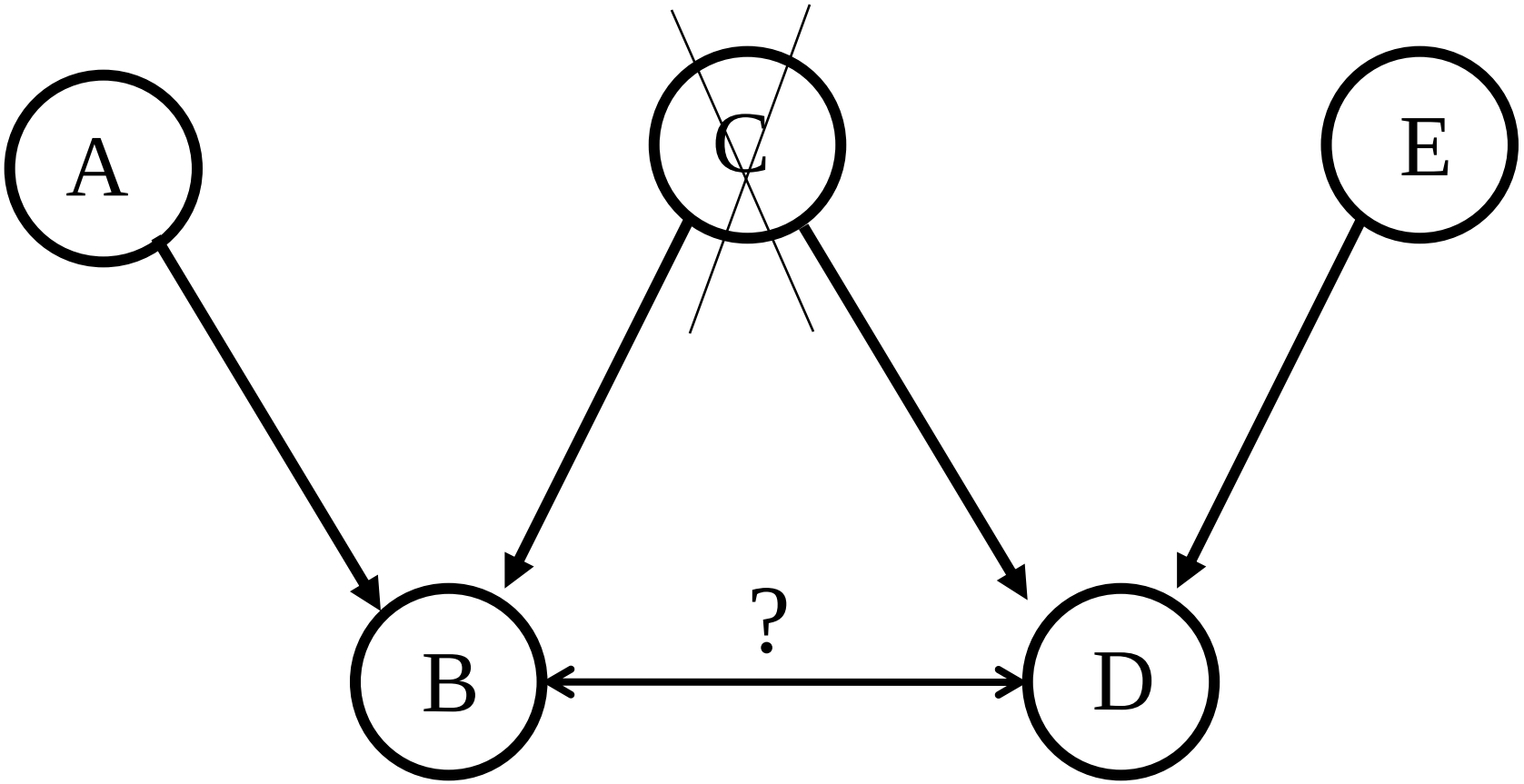
Markov boundary

- **Definition**: If <G,P> satisfies the **faithfulness** condition, the *d-separations* in the DAG identify *all and only* the conditional independencies in P, i.e.

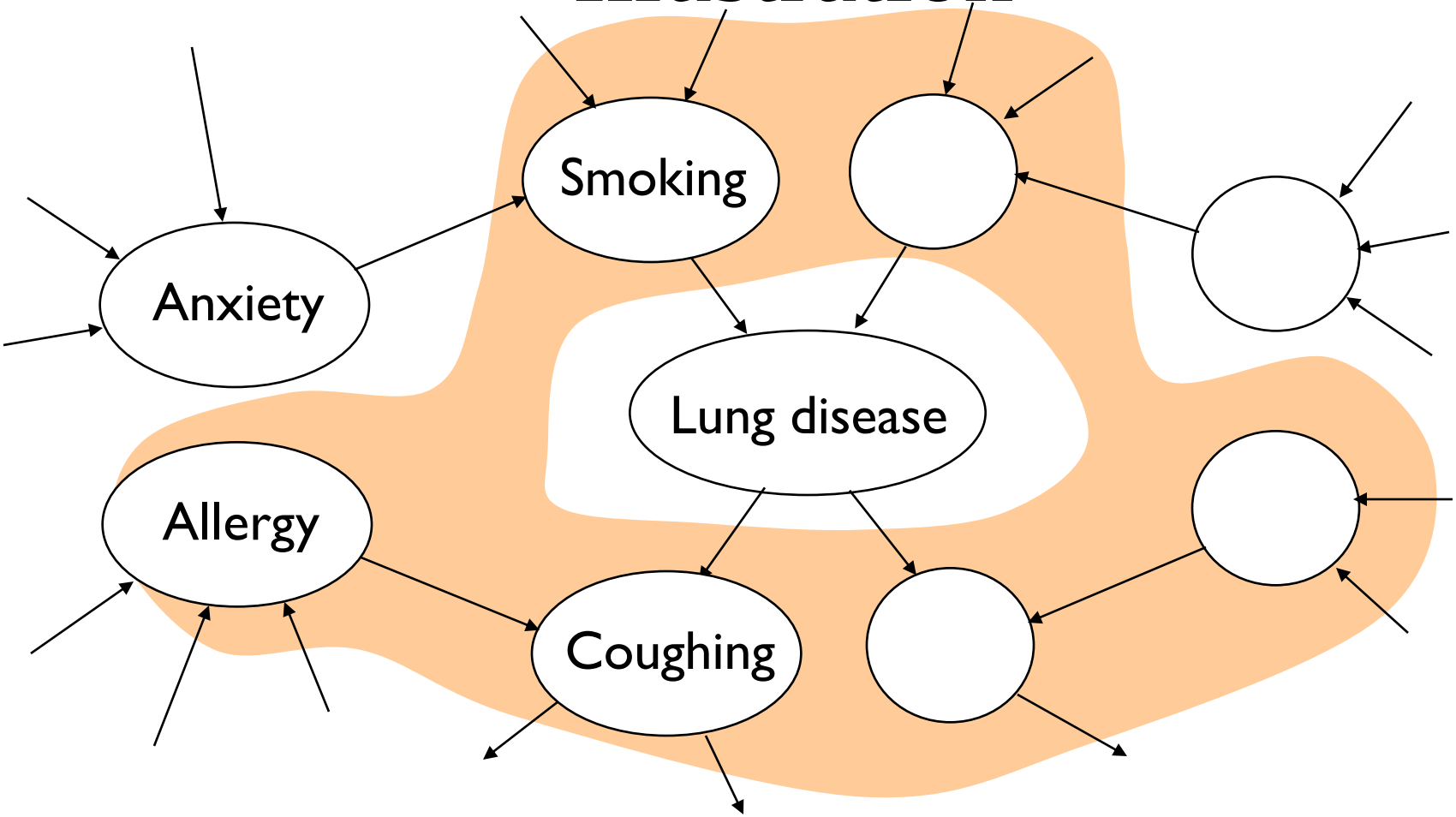
$$d\text{-}sep(X,Y|\mathbf{Z}) \text{ iff } Ind_P(X,Y|\mathbf{Z})$$

- **Theorem**: under the faithfulness condition, for all T, the union of the parents and children of T and the parents of the children of T, is the **unique Markov boundary** of T.

Faithfulness : Si **C** disparaît ?

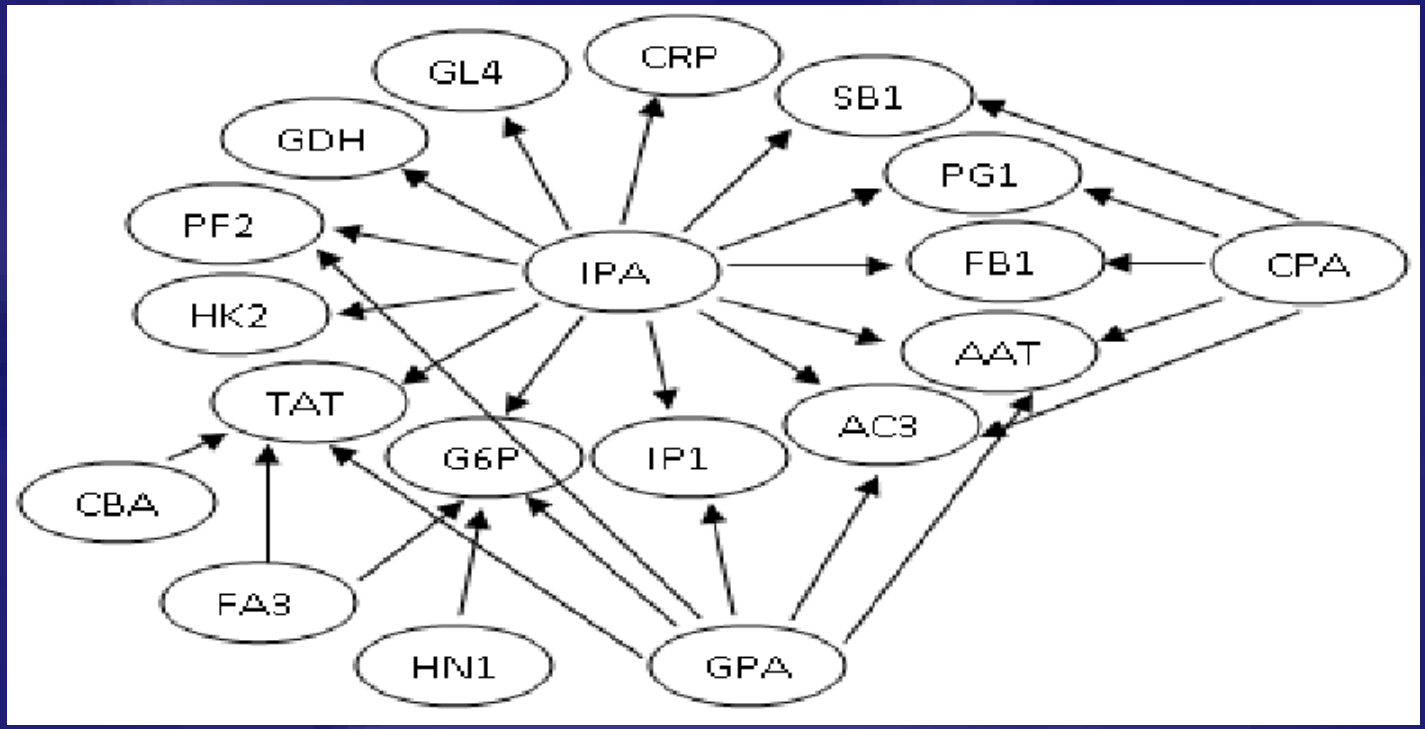


Illustration



Lung disease is independent on Allergy but when Coughing is observed, they become dependent ! Allergy should therefore be added to the feature set. Conversely, Anxiety is dependent on Lung disease but it is independent conditionally on Smoking. Anxiety should be removed from the feature set.

Large Markov boundaries



• To check whether IPA is independent on any other variables, we need to conduct a independence test with 18 variables in the conditional set….

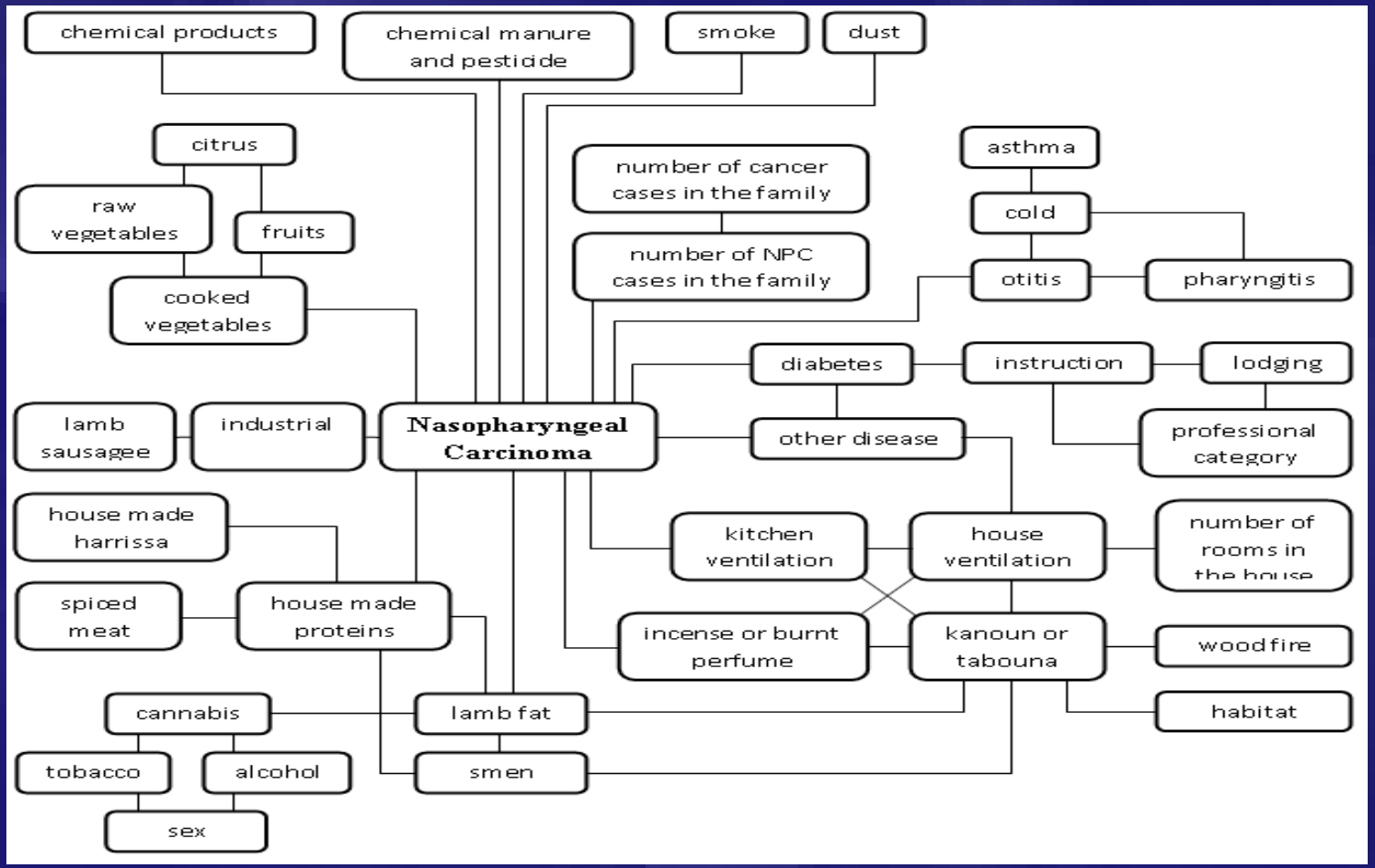
Constraint-based MB discovery

- In recent years, they have been a growing interest in inducing the MB automatically from data with correct, scalable **constraint-based** (CB) methods

- CB procedures systematically check the data for independence relationships to infer the graph structure. CB algorithms usually run a Chi2 independence test in order to decide on conditional dependencies or independencies with respect to the data set D.

- These methods search the **MB**(T) without having to construct the whole Bayesian network first. Hence their ability to scale up to thousands of variables.

Etude épidémiologique



Incremental methods

Algorithm 1 *IAMB*

Require: T: target; D = data set
Ensure: **MB** : Markov boundary of T

Phase I: *Add true positives to* **MB**
1: $\mathbf{MB} = \emptyset$
2: **repeat**
3: $[assoc, Y] = \mathbf{max}_{X \in (\mathbf{U} \setminus \mathbf{MB} \setminus T)} \mathbf{dep}(T, X | \mathbf{MB})$
4: **if** $assoc \neq 0$ **then**
5: $\mathbf{MB} = \mathbf{MB} \cup Y$
6: **end if**
7: **until** **MB** has not changed

Phase II: *Remove false positives from* **MB**
8: **for all** $X \in \mathbf{MB}$ **do**
9: **if** $\mathbf{dep}(T, X | \mathbf{MB} \setminus X) = 0$ **then**
10: $\mathbf{MB} = \mathbf{MB} \setminus X$
11: **end if**
12: **end for**

Divide-and-conquer methods

Algorithm 2 *MMPC*

Require: T: target; $\mathcal{D}$: data set
Ensure: **CPC** : set of candidates to parents or children of T

Phase I: *Forward*
$CPC = \emptyset$
$\mathbf{U} = \mathbf{U} \setminus T$
repeat
  $[assoc, F] = \max_{X \in \mathbf{U}} \mathbf{MinAssoc}(X; T; \mathbf{CPC}; \emptyset)$
  if $assoc \neq 0$ **then**
    $\mathbf{CPC} = \mathbf{CPC} \cup F$
    $\mathbf{U} = \mathbf{U} \setminus F$
  end if
until CPC has not changed

Phase II: *Backward*
for all $X \in \mathbf{CPC}$ **do**
  if $\mathbf{MinAssoc}(X; T; \mathbf{CPC} \setminus X; \emptyset) = 0$ **then**
    $\mathbf{CPC} = \mathbf{CPC} \setminus X$
  end if
end for

Algorithm 3 *MMMB*

Require: T: target; D = data set
Ensure: **MB** : Markov boundary of T

Phase I: *Add true positives*
$\mathbf{PC} = \mathbf{MMPC}(T, D)$
$\mathbf{MB} = \mathbf{PC}$

Phase II: *Add more true positives*
$\mathbf{CanMB} = \emptyset$
for all $X \in \mathbf{PC}$ **do**
  $\mathbf{CanMB} = \mathbf{CanMB} \cup \mathbf{MMPC}(X, D) \setminus T$
end for

Phase III: *Remove false positives*
for all $X \in (\mathbf{CanMB})$ **do**
  find any $\mathbf{Z}$ such that $(T \perp X | \mathbf{Z})$
  for all $Y \in \mathbf{PC}$ **do**
    if $\mathbf{dep}(T, X | \mathbf{Z} \cup Y) \neq 0$ **then**
      $\mathbf{MB} = \mathbf{MB} \cup X$
    end if
  end for
end for