Étapes pour bâtir un réseau bayésien

- Comment bâtir un réseau bayésien afin de modéliser un environnement/problème donné ?

- On a besoin de spécifier 2 choses :
  - la structure du réseau
    (quelles indépendances peut-on supposer ?)
  - les tables de probabilités
    (quelle est la relation entre les variables de l'environnement ?)

Spécifier les tables de probabilités d'un RB

- Supposons que le graphe d'un RB ait été spécifié par un expert
- Comment estimer les tables de probabilités $P(X_i \mid Parents(X_i))$?
- On pourrait demander au même expert de définir à la main ces tables
  - travail long et fastidieux
  - pas très naturel ou intuitif
- Il serait préférable d'automatiser ce processus
  - on **collecte des données** sur l'environnement que l'on souhaite modéliser
  - on dérive des tables de probabilités qui **reflètent bien ces données**
- C'est ce qu'on appelle faire de l'**apprentissage automatique**
  - le RB va s'adapter à l'environnement et apprendre à l' « imiter »

Spécifier les tables de probabilités d'un RB

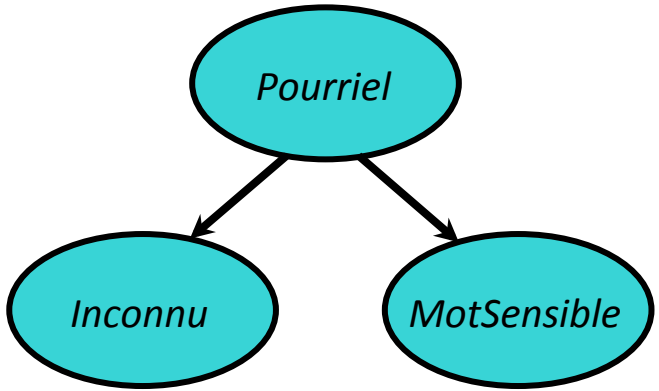
- Si on a un ensemble de données **où tous les nœuds X_i sont observés**, c'est facile :

$$P(X_i = x \mid Parents(X_i) = p\ ) \approx \text{freq}(x,p) \,/\, \Sigma_{x'}\, \text{freq}(x',p)$$

- On fait ce calcul pour toutes les valeurs x de X_i et toutes les valeurs p de ses parents possibles
  - pour éviter d'avoir de probabilités à 0, on peut ajouter aux fréquences freq(x,e) une petite constante positive δ (ex. : δ=1)

Exemple

- Supposons que l'on souhaite détecter des pourriels à l'aide du RB suivant :
  - ***Inconnu*** : l'adresse de l'expéditeur n'est pas connu par le destinataire
  - ***MotSensible*** : le courriel contient un mot appartenant à une liste de mots « sensibles »
  - ***Pourriel*** : le courriel est un pourriel

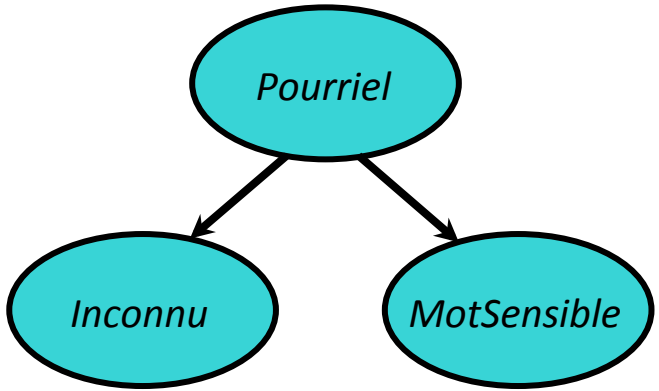


- Supposons qu'on a collecté un ensemble de 122 courriels où
  - 70 des 122 courriels étaient des pourriels

$P(Pourriel=\text{vrai}) = (70 + 1) / (70 + 1 + 52 + 1) \approx 0.57$

Exemple

- Supposons que l'on souhaite détecter des pourriels à l'aide du RB suivant :
  - ***Inconnu*** : l'adresse de l'expéditeur n'est pas connu par le destinataire
  - ***MotSensible*** : le courriel contient un mot appartenant à une liste de mots « sensibles »
  - ***Pourriel*** : le courriel est un pourriel

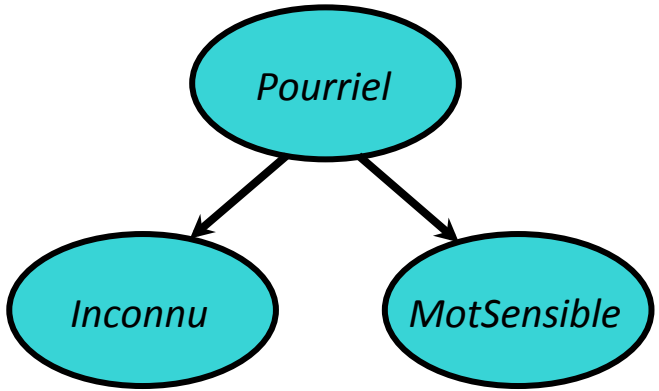


- Supposons qu'on a collecté un ensemble de 122 courriels où
  - parmi les 70 pourriels, 65 avaient un expéditeur inconnu et 51 contenaient un mot sensible

$P(Inconnu=\text{vrai} \mid Pourriel=\text{vrai}) = (65 + 1) / (65 + 1 + 5 + 1) \approx 0.92$
$P(MotSensible=\text{vrai} \mid Pourriel=\text{vrai}) = (51 + 1) / (51 + 1 + 19 + 1) \approx 0.72$

Exemple

- Supposons que l'on souhaite détecter des pourriels à l'aide du RB suivant :
  - ***Inconnu*** : l'adresse de l'expéditeur n'est pas connu par le destinataire
  - ***MotSensible*** : le courriel contient un mot appartenant à une liste de mots « sensibles »
  - ***Pourriel*** : le courriel est un pourriel



- Supposons qu'on a collecté un ensemble de 122 courriels où
  - parmi les 52 courriels valides, 10 avaient un expéditeur inconnu et 0 contenaient un mot sensible

$P(Inconnu=\text{vrai} \mid Pourriel=\text{faux}) = (10 + 1) / (10 + 1 + 42 + 1) \approx 0.20$
$P(MotSensible=\text{vrai} \mid Pourriel=\text{faux}) = (0 + 1) / (0 + 1 + 52 + 1) \approx 0.02$

Spécifier les tables de probabilités d'un RB

- Si on a un ensemble de données **où certains des nœuds ne sont pas observés**, on doit utiliser des méthodes plus sophistiquées
  - Algorithme EM (voir section 20.3.2)