

The logo for ARVALIS features a stylized graphic on the left composed of three overlapping leaf-like shapes in yellow, teal, and blue. To the right of this graphic, the word "ARVALIS" is written in a bold, teal, sans-serif typeface. A thin, teal horizontal line is positioned directly beneath the text.

ARVALIS



Data sciences et causalité : *exemple des forêts aléatoires*

Académie d'Agriculture de France

Séance du 10 mai 2023 - Data Sciences en Agriculture

Emmanuelle GOURDAIN

D'après des travaux réalisés par :

François PIRAUX, Emmanuelle HERITIER, Florent DUYME, Doriane HAMERNIG, Christophe GIGOT

Titre de l'intervention

La recherche de causalité en agronomie

Produire de la connaissance
=
mettre en évidence des liens de cause à effet



La recherche de causalité en agronomie

Produire de la connaissance
=
mettre en évidence des liens de cause à effet

Expérimentation scientifique

- ❑ **Expérimentation randomisée** : on ne peut jamais tout contrôler → la randomisation assure que l'on soit toutes choses égales par ailleurs
- ❑ Faire des **répétitions** afin de quantifier la part de variabilité non contrôlée
- ❑ Prévoir un traitement **témoin**



La recherche de causalité en agronomie



Produire de la connaissance
=
mettre en évidence des liens de cause à effet



Expérimentation scientifique

Mais des limites :

- ✓ Représentativité des résultats obtenus (faible diversité pédoclimatique)
- ✓ Difficulté à travailler la combinaison de leviers
- ✓ Temps long



La recherche de causalité en agronomie

Produire de la connaissance

=

mettre en évidence des liens de cause à effet



Expérimentation scientifique

Mais des limites :

- ✓ Représentativité des résultats obtenus (faible diversité pédoclimatique)
- ✓ Difficulté à travailler la combinaison de leviers
- ✓ Temps long

Opportunités :

- ✓ De plus en plus de données collectées dans les fermes
- ✓ Davantage de puissance de calcul
- ✓ La démocratisation de logiciels de traitement de données et des méthodes d'IA



Question à résoudre

Collecte des données

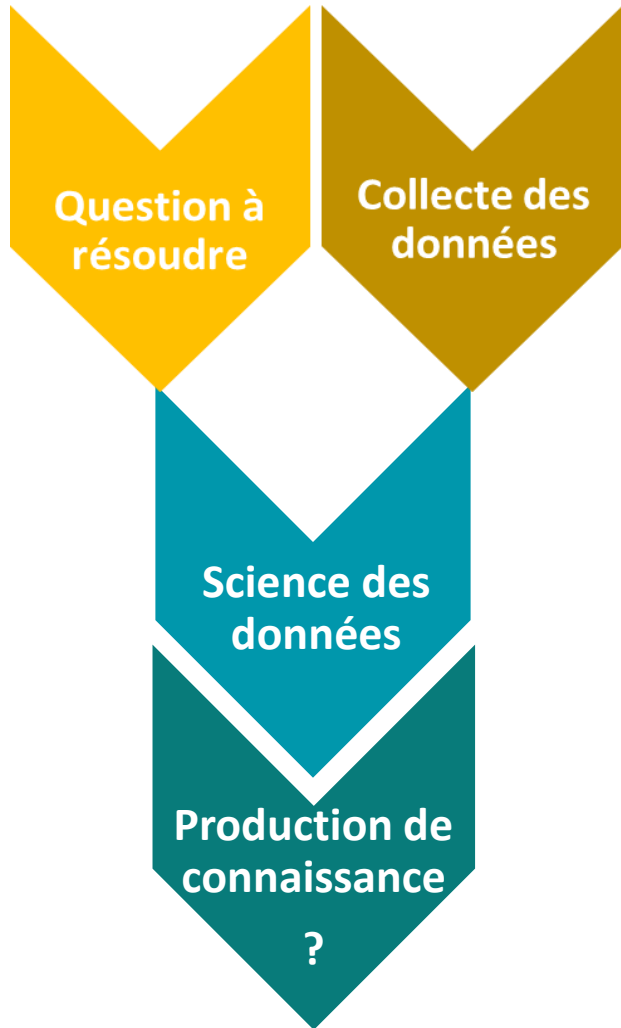
Science des données

Production de connaissance

?



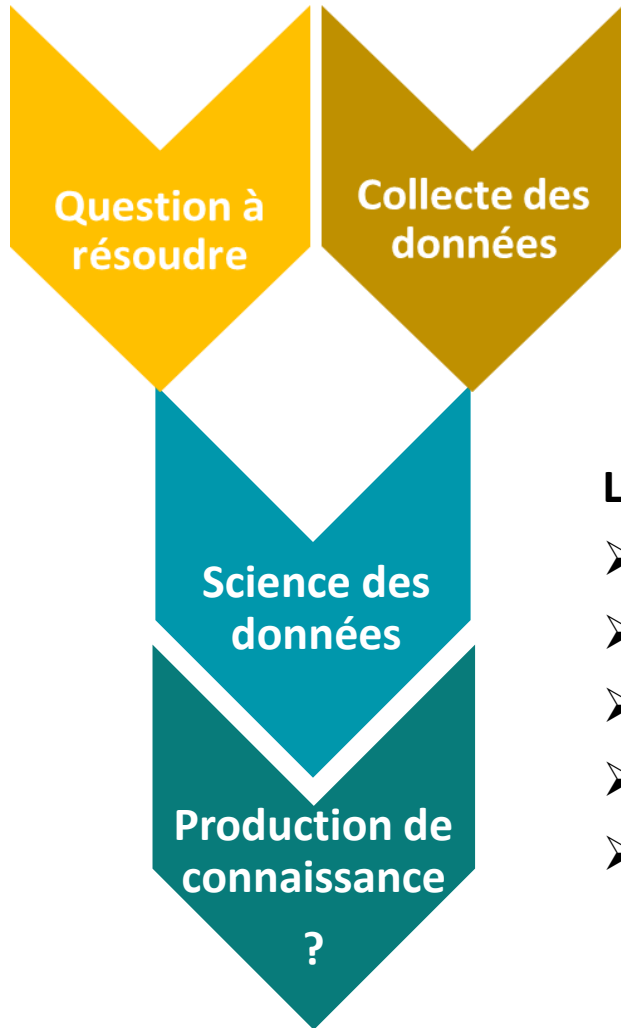
Produire de la connaissance avec des méthodes de machine learning



Peut-on mettre en évidence des liens de cause à effet avec une méthode de machine learning ?



Produire de la connaissance avec des méthodes de machine learning



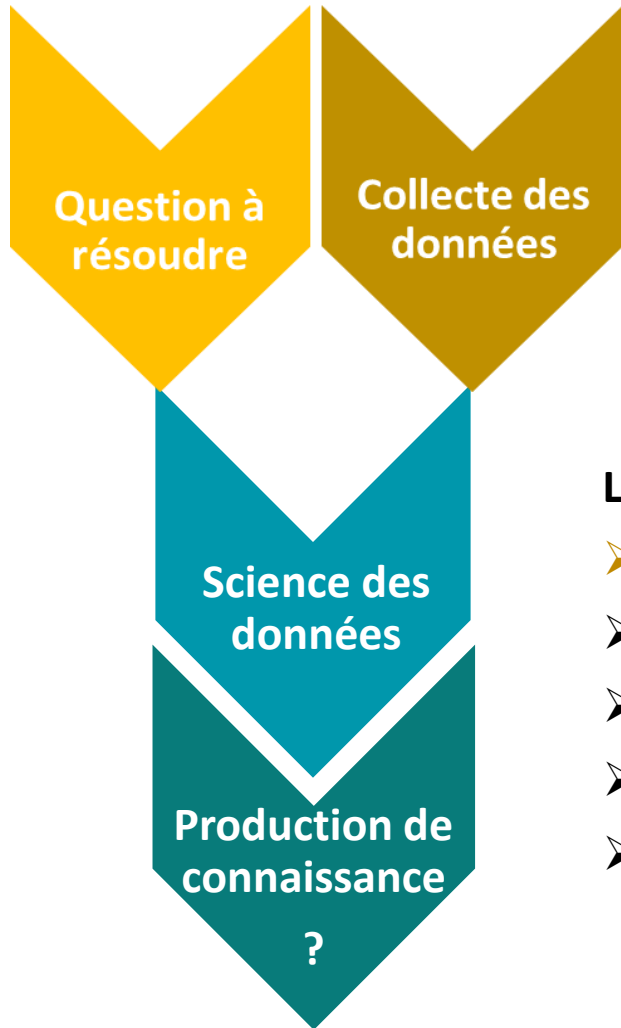
Peut-on mettre en évidence des liens de cause à effet avec une méthode de machine learning ?

La méthode :

- Construction d'un modèle causal -> réseau bayésien
- Ajout de variables non causales
- Simulation d'un jeu de données
- Ajustement d'une random forest (RF) et calcul de l'importance des variables
- Comparaison entre importance des variables calculée par RF et causalité réelle des variables



Produire de la connaissance avec des méthodes de machine learning



Peut-on mettre en évidence des liens de cause à effet avec une méthode de machine learning ?

La méthode :

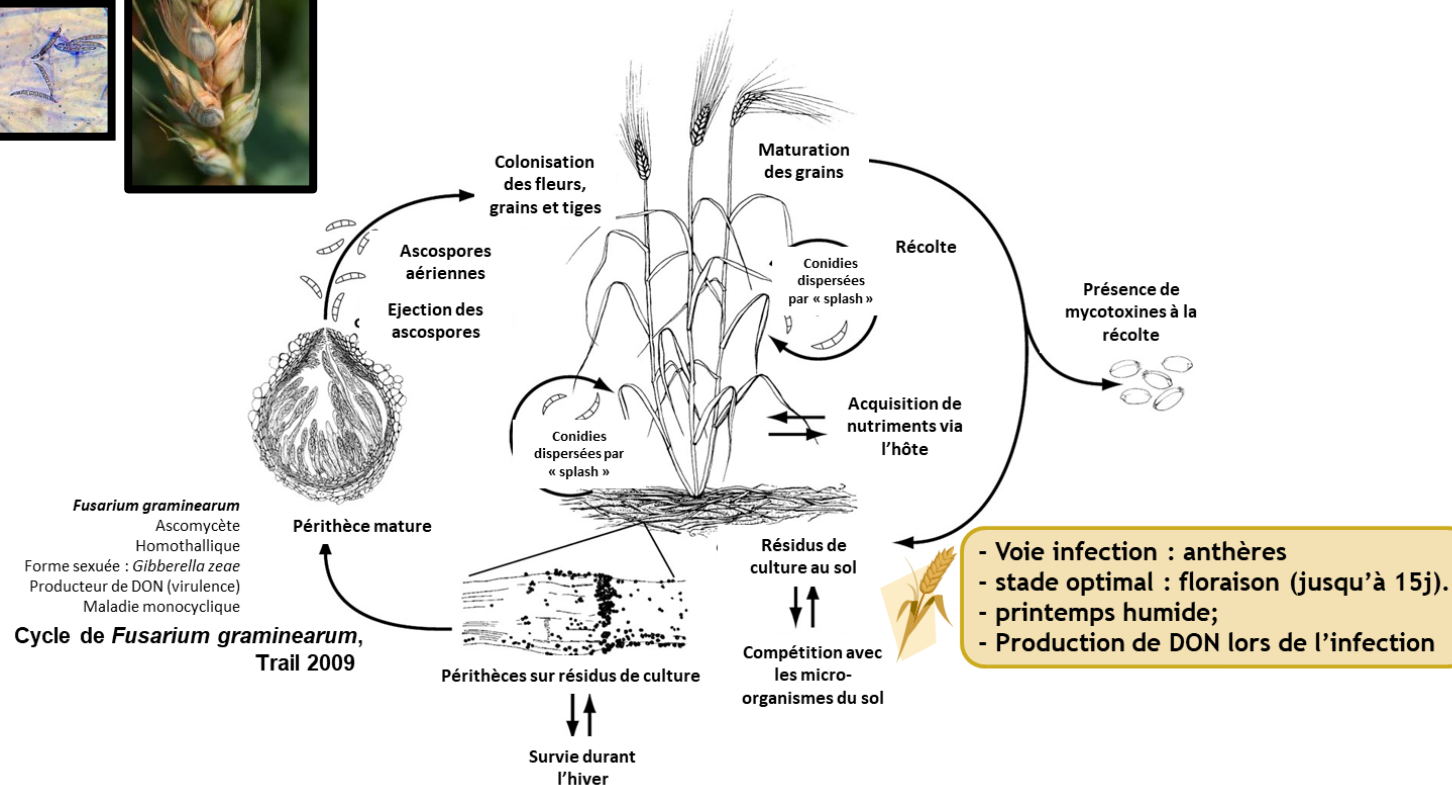
- **Construction d'un modèle causal -> réseau bayésien**
- **Ajout de variables non causales**
- **Simulation d'un jeu de données**
- **Ajustement d'une random forest (RF) et calcul de l'importance des variables**
- **Comparaison entre importance des variables calculée par RF et causalité réelle des variables**



Le cas d'usage

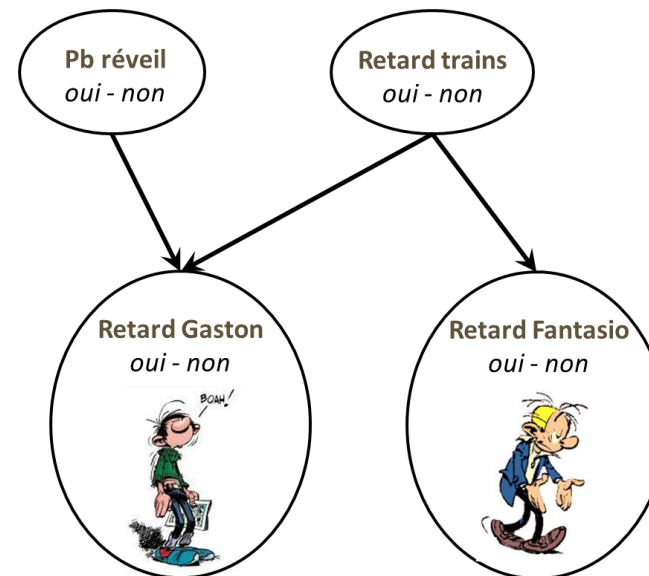


Prédire le niveau de concentration dans les grains de blé récoltés d'une toxine (Déoxynivalénol-DON) produite par un pathogène (*F. graminearum*)



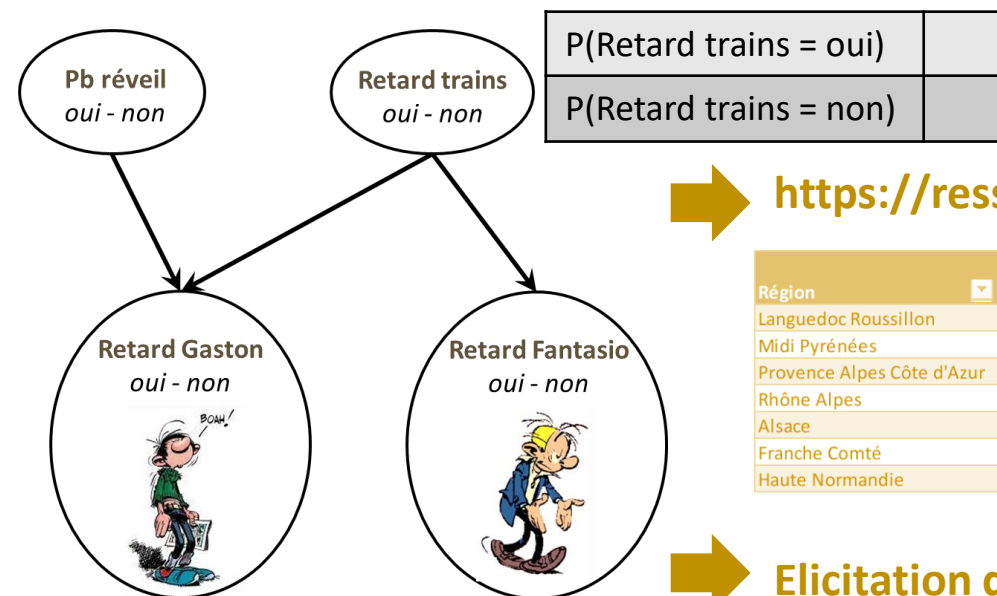
Méthode choisie pour répondre à cet objectif

Les réseaux bayésiens sont des modèles permettant de modéliser des liens causaux préalablement décrits dans des DAG (Diagram Acyclic Graph). Ils peuvent être paramétrés à dire d'expert, par des données et/ou par d'autres modèles.



Méthode choisie pour répondre à cet objectif

Les réseaux bayésiens sont des modèles permettant de modéliser des liens causaux préalablement décrits dans des DAG (Diagram Acyclic Graph). Ils peuvent être paramétrés à dire d'expert, par des données et/ou par d'autres modèles.



➔ <https://ressources.data.sncf.com>

Région	Nombre de trains ayant circulé	Nombre de trains en retard à l'arrivée	Taux de régularité
Languedoc Roussillon	4897	377	92.3
Midi Pyrénées	7941	903	88.6
Provence Alpes Côte d'Azur	13219	1816	86.3
Rhône Alpes	30779	3552	88.5
Alsace	18097	771	95.7
Franche Comté	5174	522	89.9
Haute Normandie	5301	211	96

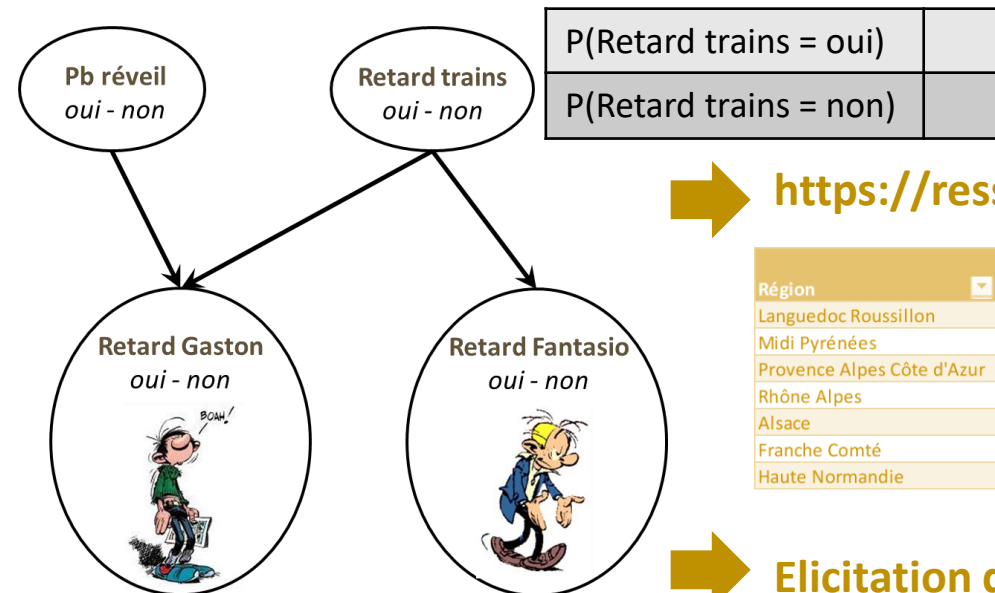
➔ Elicitation d'experts

Méthode choisie pour répondre à cet objectif

Les réseaux bayésiens sont des modèles permettant de modéliser des liens causaux préalablement décrits dans des DAG (Diagram Acyclic Graph). Ils peuvent être paramétrés à dire d'expert, par des données et/ou par d'autres modèles.

Les **Réseaux Bayésiens** sont des modèles intéressants car ils permettent :

- de **décrire** des relations causales entre des variables au travers d'un DAG : effets directs et indirects
- de **simuler** des données correspondant à ce modèle causal et incluant différents types de biais de causalité



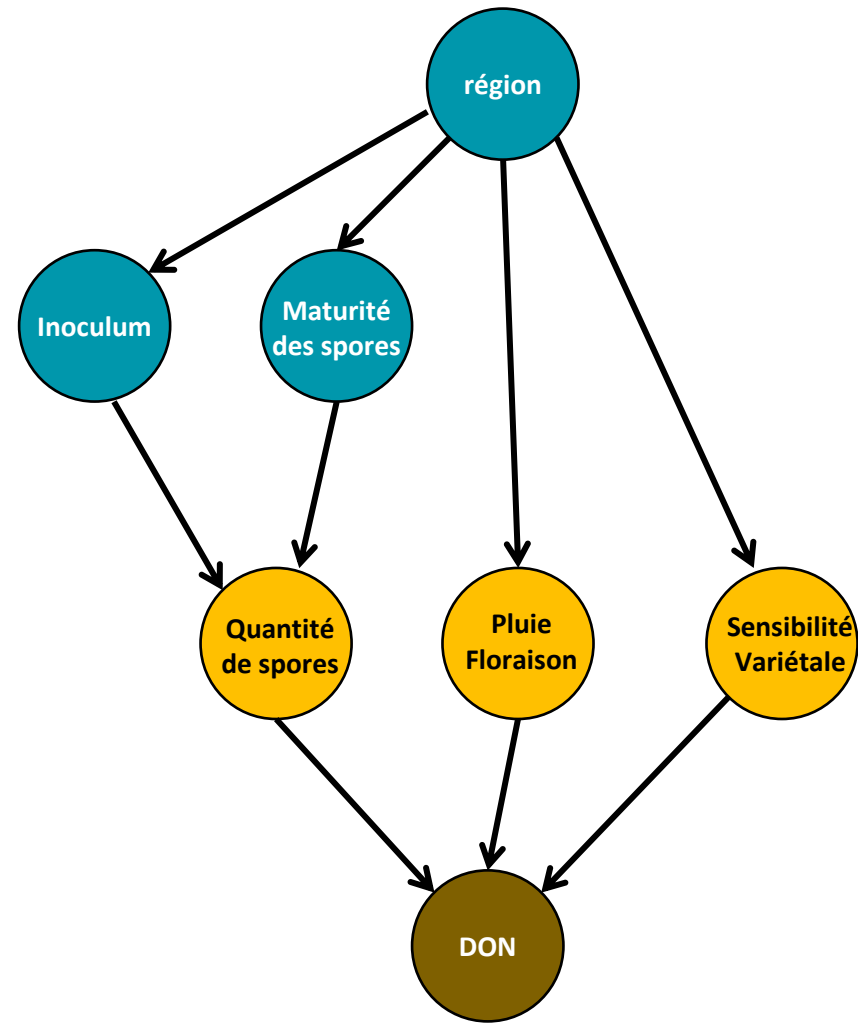
➔ <https://ressources.data.sncf.com>

Région	Nombre de trains ayant circulé	Nombre de trains en retard à l'arrivée	Taux de régularité
Languedoc Roussillon	4897	377	92.3
Midi Pyrénées	7941	903	88.6
Provence Alpes Côte d'Azur	13219	1816	86.3
Rhône Alpes	30779	3552	88.5
Alsace	18097	771	95.7
Franche Comté	5174	522	89.9
Haute Normandie	5301	211	96

➔ Elicitation d'experts

Random forests et causalité

DAG "DON"

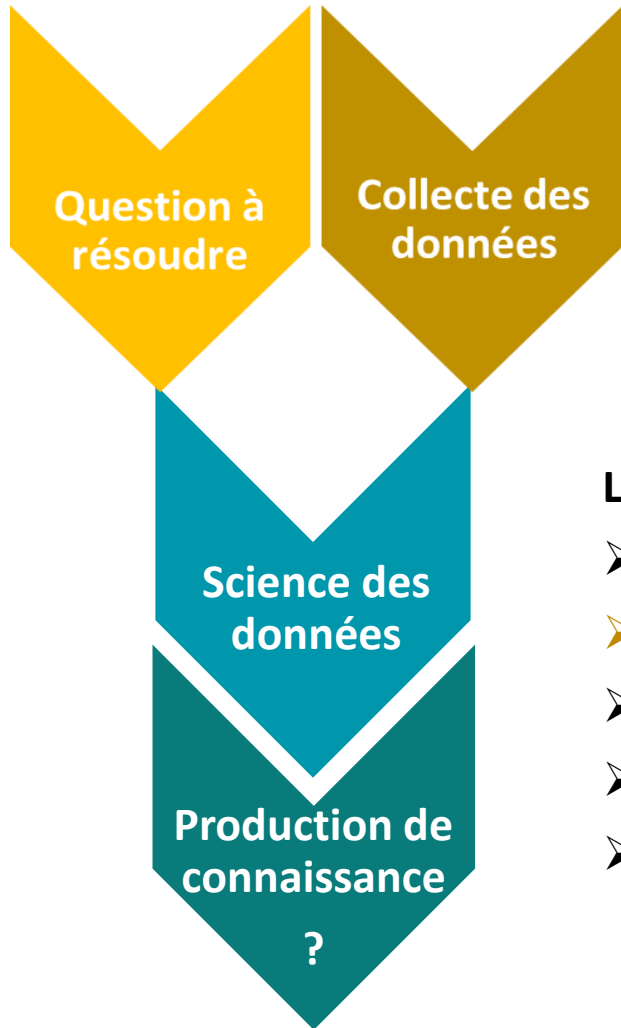


Effets indirects 

Effets directs 



Produire de la connaissance avec des méthodes de machine learning



Peut-on mettre en évidence des liens de cause à effet avec une méthode de machine learning ?

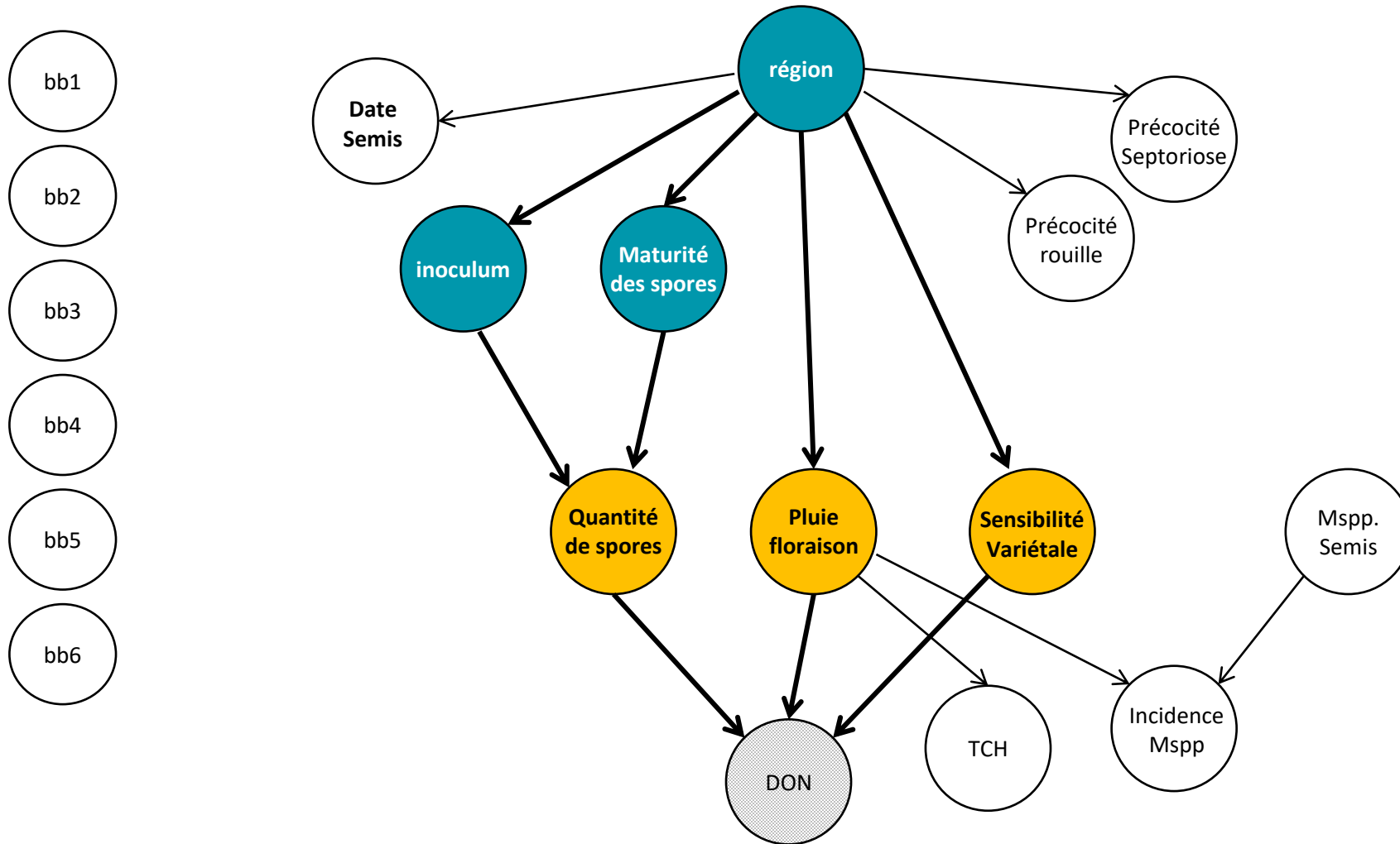
La méthode :

- Construction d'un modèle causal -> réseau bayésien
- **Ajout de variables non causales**
- Simulation d'un jeu de données
- Ajustement d'une random forest (RF) et calcul de l'importance des variables
- Comparaison entre importance des variables calculée par RF et causalité réelle des variables



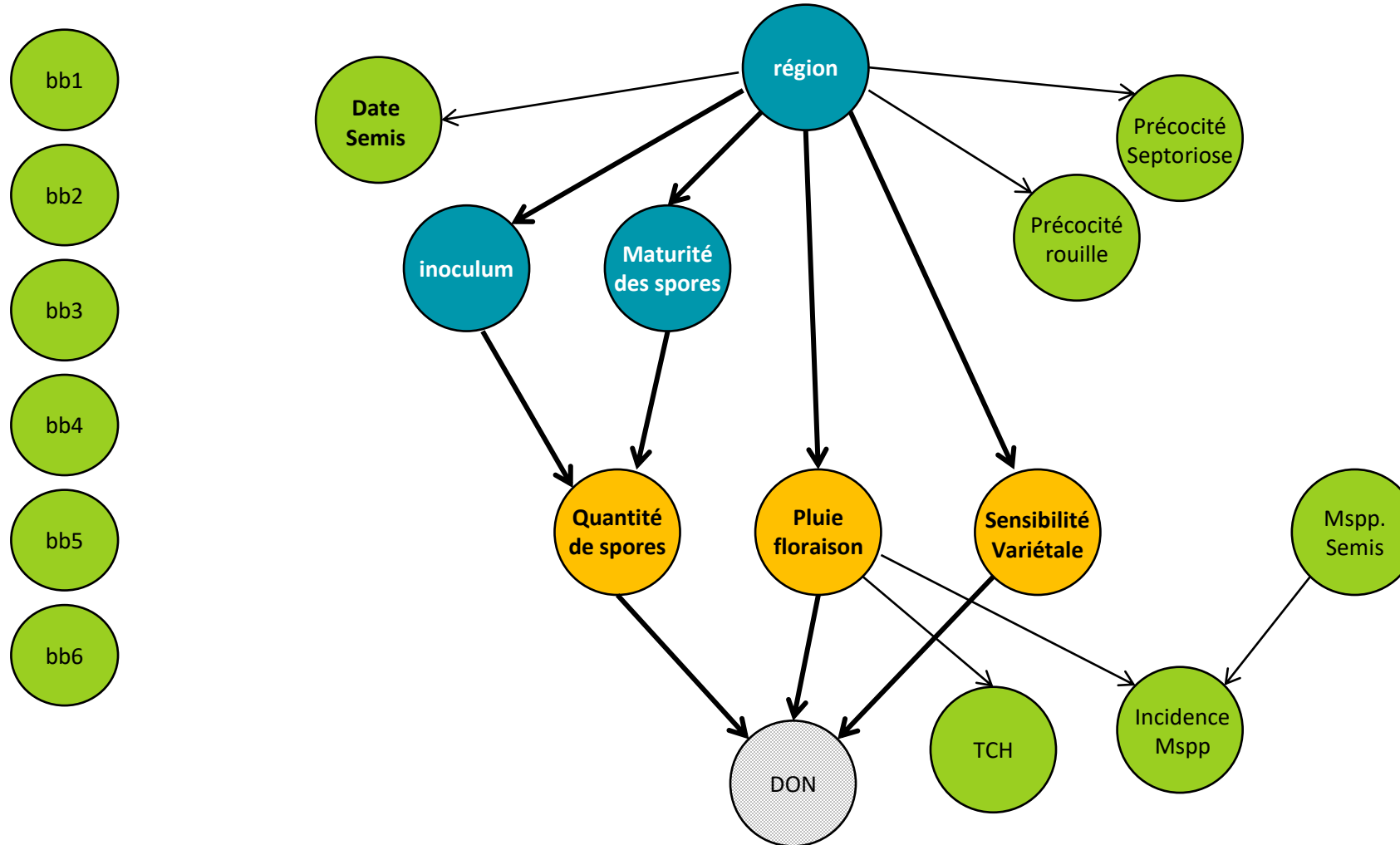
Random forests et causalité

DAG "DON" + variables non causales



Random forests et causalité

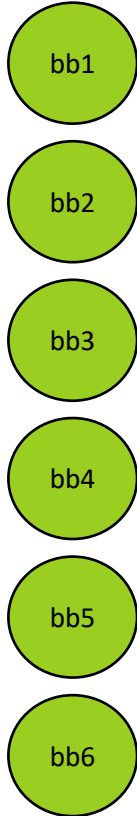
DAG "DON" + variables non causales



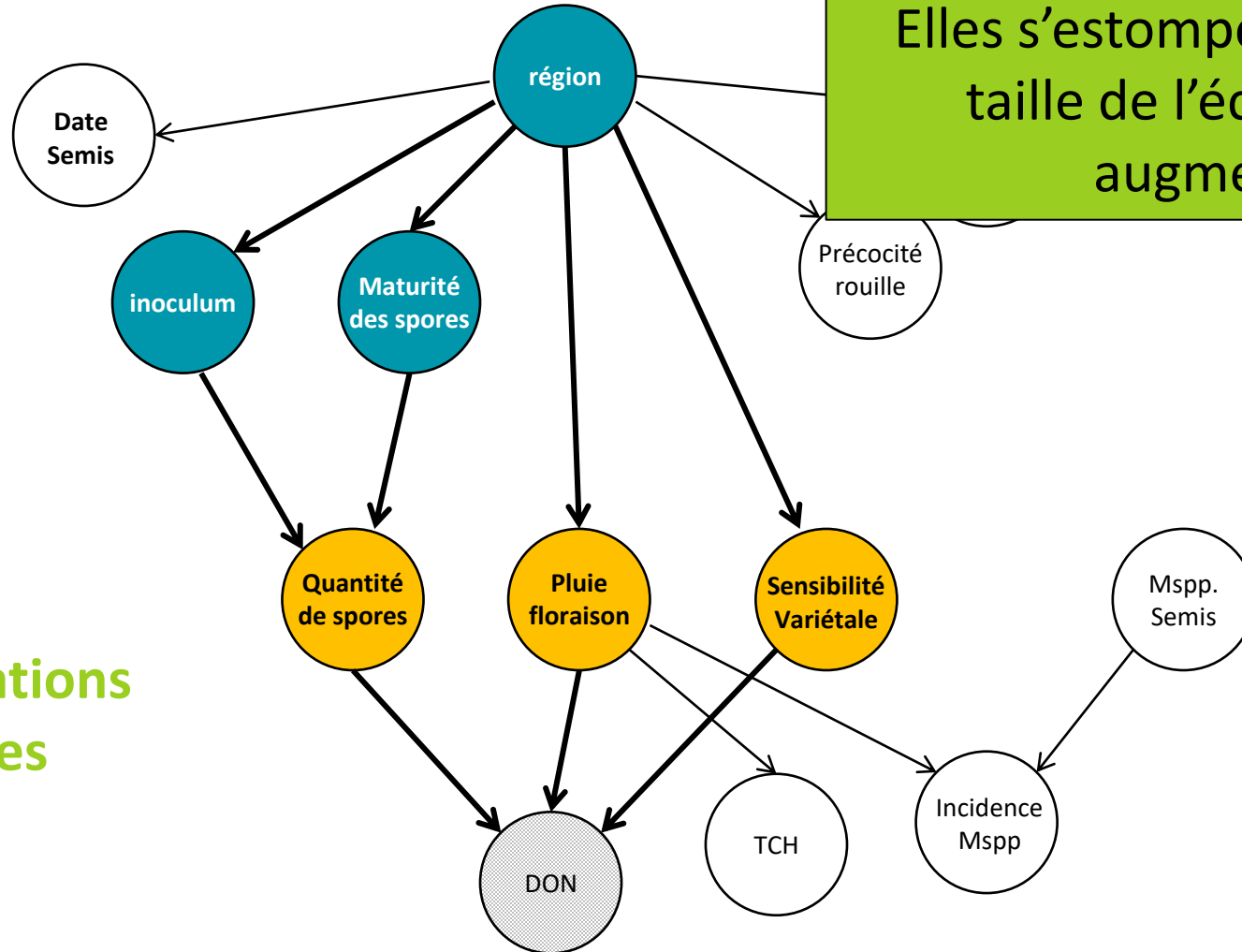
Random forests et causalité

DAG "DON" + variables non causales

Variables
aléatoires



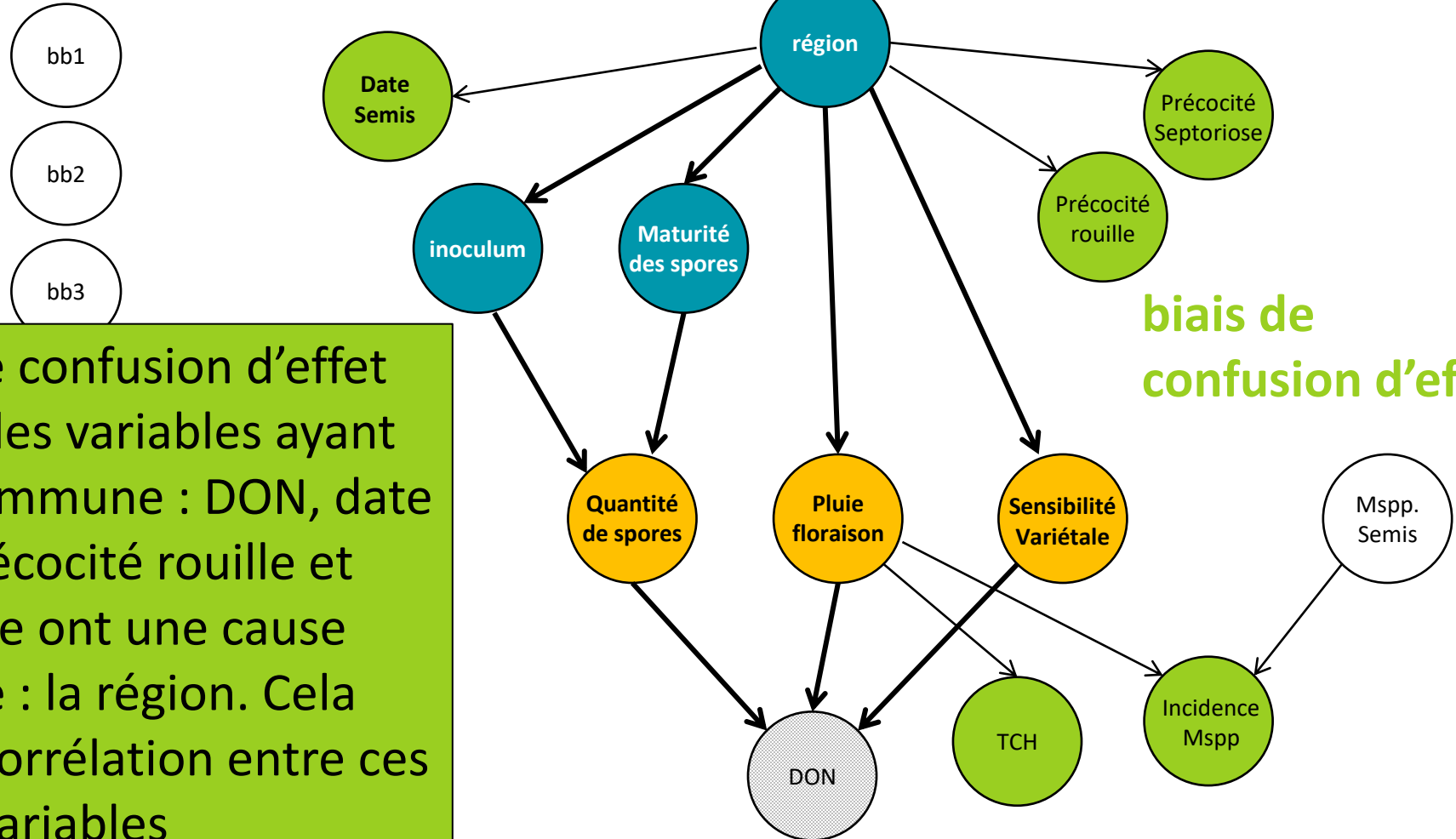
corrélations
fortuites



Les associations fortuites sont dues au choix de l'échantillon. Elles s'estompent quand la taille de l'échantillon augmente



DAG "DON" + variables non causales



Les biais de confusion d'effet sont liés à des variables ayant une cause commune : DON, date semis, précocité rouille et septoriose ont une cause commune : la région. Cela génère une corrélation entre ces variables

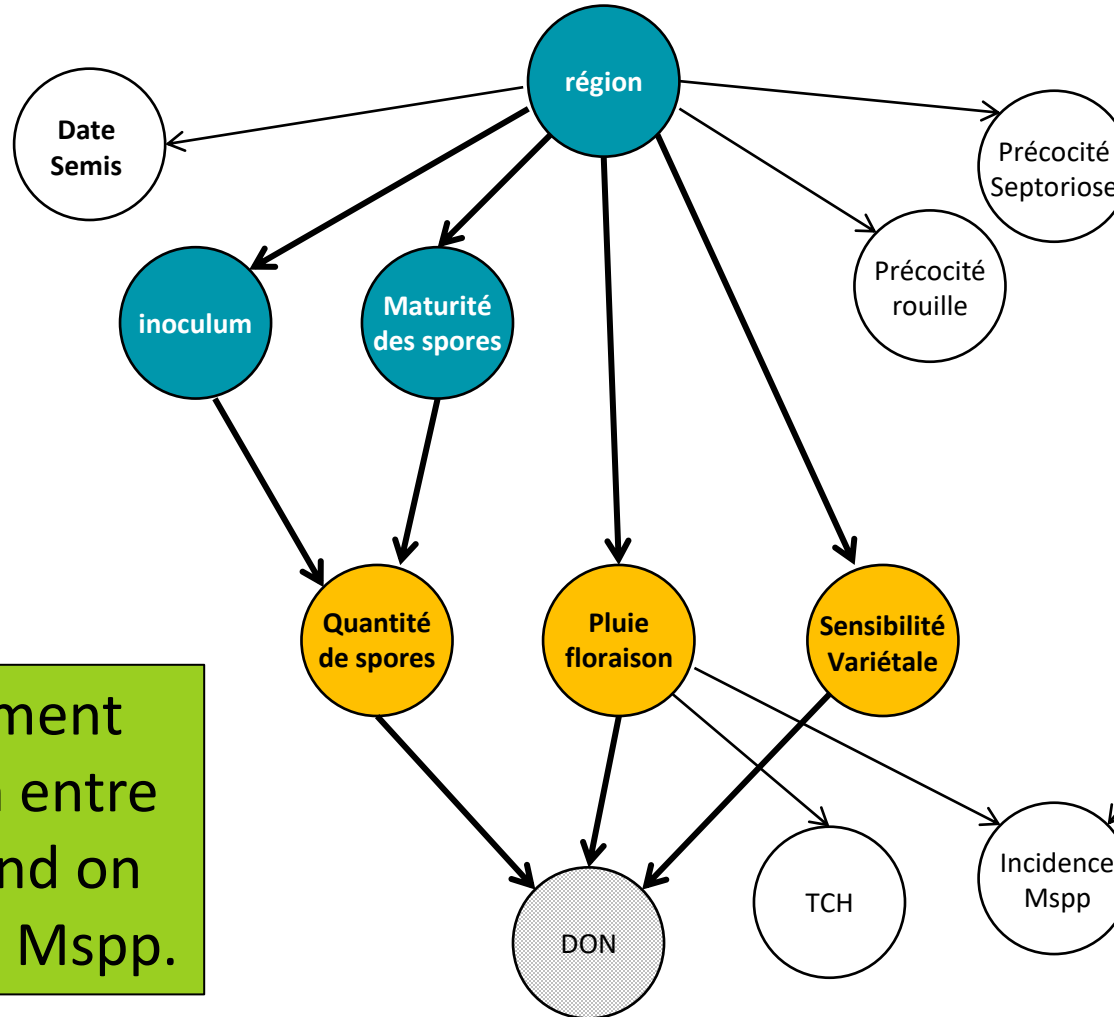


Random forests et causalité

DAG "DON" + variables non causales

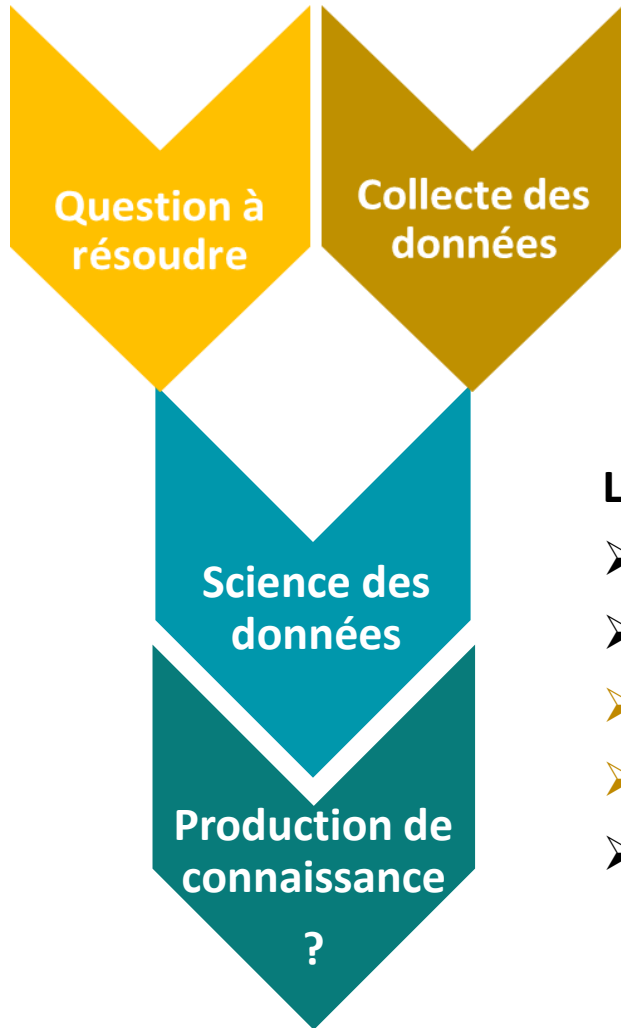
bb1
bb2
bb3
bb4
bb5

Le biais de conditionnement conduit à une corrélation entre DON et Mspp semis quand on conditionne sur incidence Mspp.



biais de conditionnement

Produire de la connaissance avec des méthodes de machine learning



Peut-on mettre en évidence des liens de cause à effet avec une méthode de machine learning ?

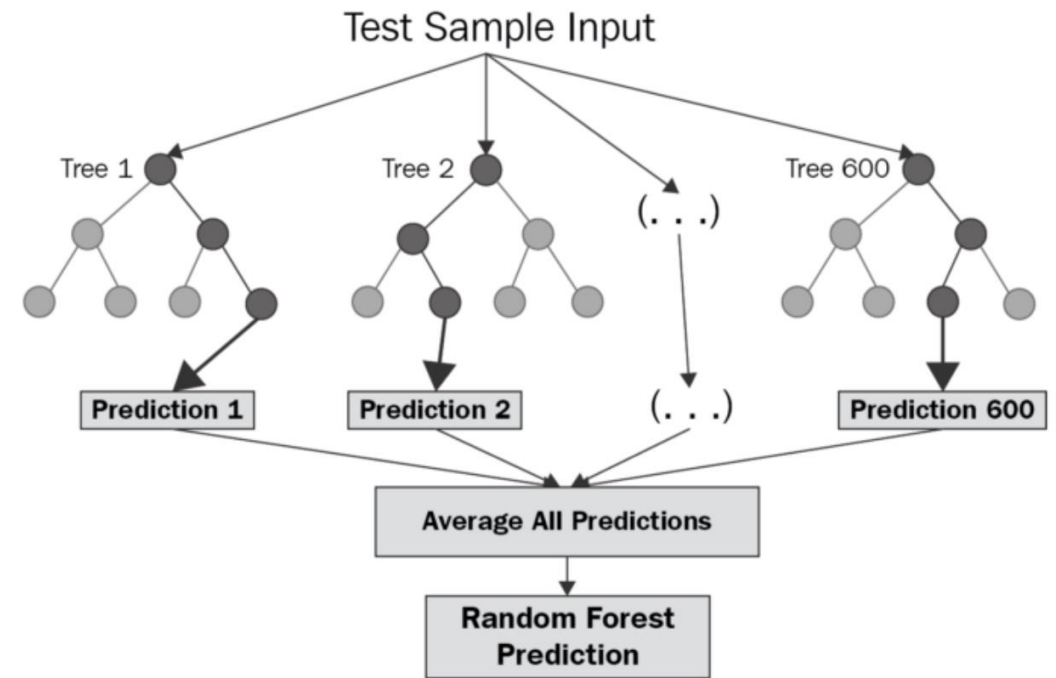
La méthode :

- Construction d'un modèle causal -> réseau bayésien
- Ajout de variables non causales
- **Simulation d'un jeu de données**
- **Ajustement d'une random forest (RF) et calcul de l'importance des variables**
- Comparaison entre importance des variables calculée par RF et causalité réelle des variables

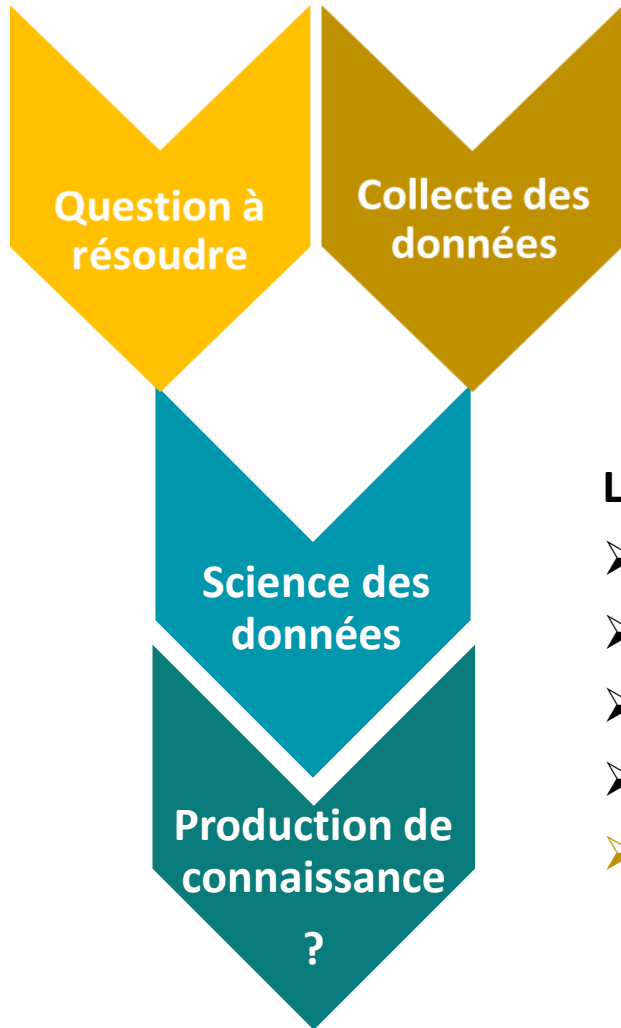


Pourquoi les forêts aléatoires (RandomForest = RF) ?

- ❑ Les RF ont pour but premier de faire de la prédiction
- ❑ Mais l'importance des variables calculée avec une Random Forest est souvent utilisée pour identifier les variables explicatives d'un phénomène



Produire de la connaissance avec des méthodes de machine learning



Peut-on mettre en évidence des liens de cause à effet avec une méthode de machine learning ?

La méthode :

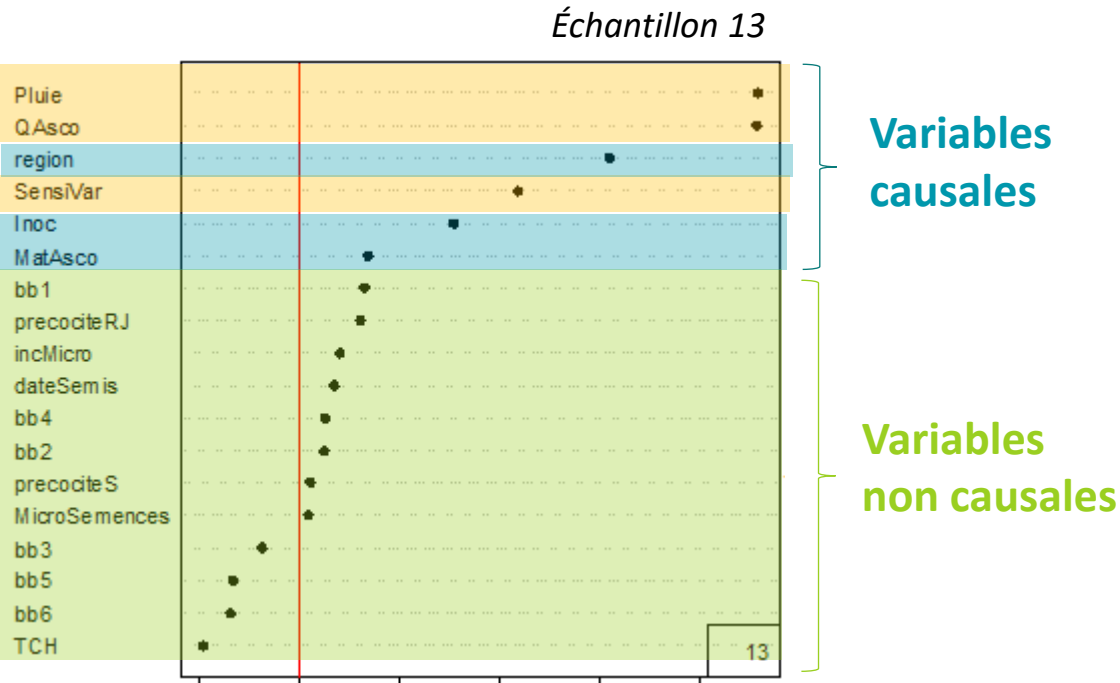
- Construction d'un modèle causal -> réseau bayésien
- Ajout de variables non causales
- Simulation d'un jeu de données
- Ajustement d'une random forest (RF) et calcul de l'importance des variables
- Comparaison entre importance des variables calculée par RF et causalité réelle des variables



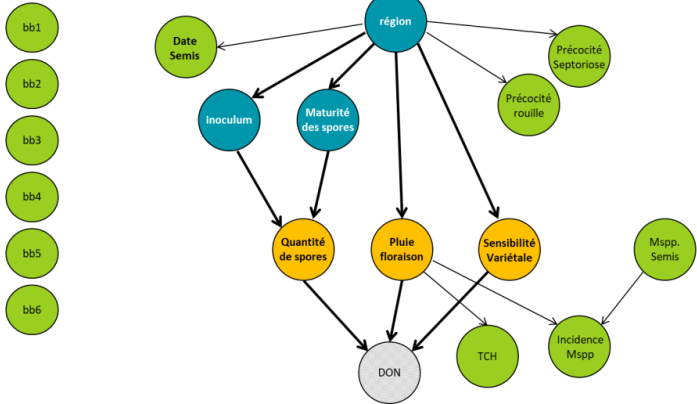
Random forests et causalité

Importance des variables avec RandomForest

N = 300

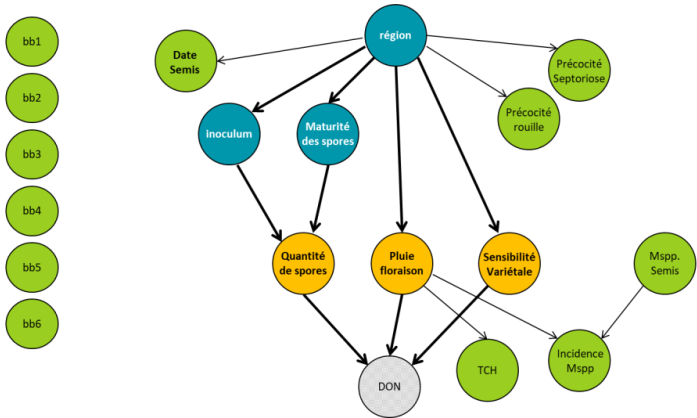


Les 6 premières variables sont les 6 variables causales

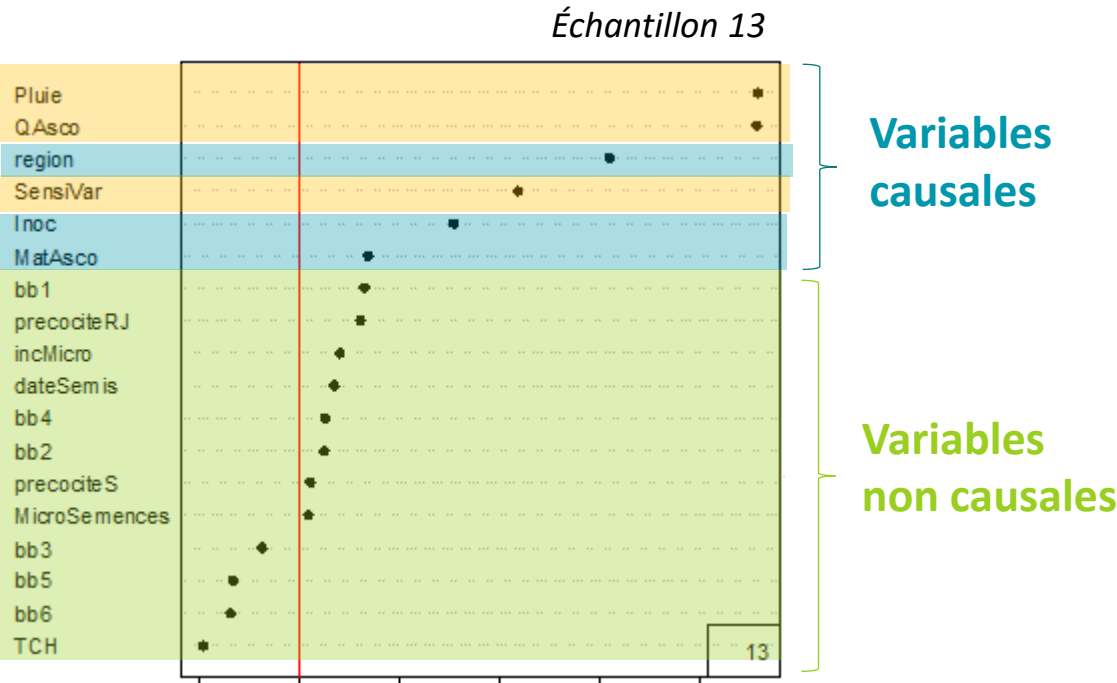


Random forests et causalité

Importance des variables avec RandomForest

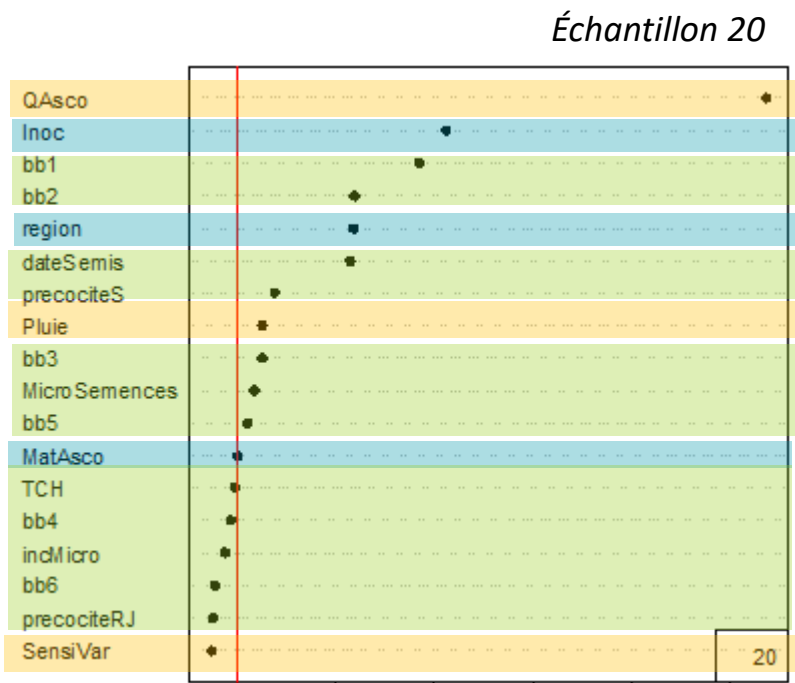


N = 300



Les 6 premières variables sont les 6 variables causales

N = 300



On retrouve des variables de biais parmi les 6 premières variables. Des variables causales ne sont pas détectées



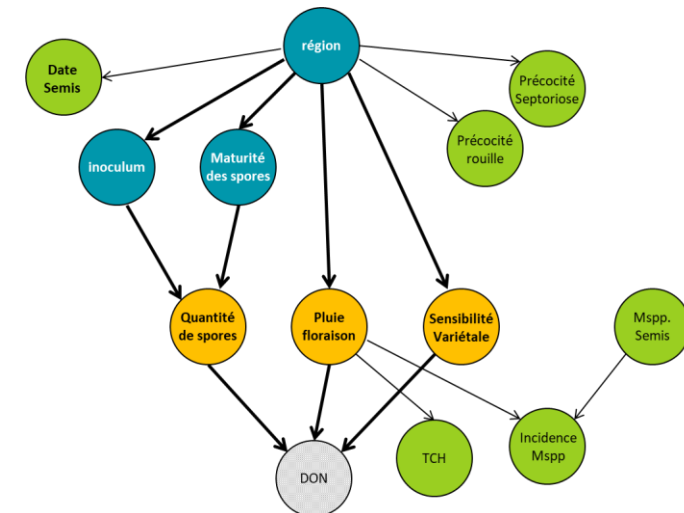
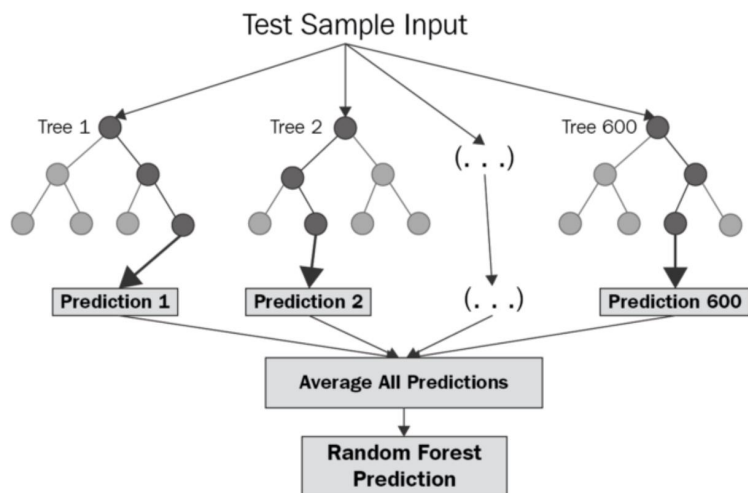
Conclusions

- ❑ Les réseaux bayésiens semblent être un outil intéressant pour étudier le comportement des forêts aléatoires
- ❑ L'importance des variables calculées par forêts aléatoires est à considérer avec beaucoup de prudence

*Gennatasa et al. proposent de combiner l'**expertise humaine sur les relations causales** et le machine learning pour accroître la capacité de généralisation des modèles.*

« human prior knowledge is fundamental for the future of machine learning »





Merci de votre attention

