Journée Perspectives et Défis de l'IA: PDIA 2021

Thème : Explicabilité

AFIA

& GT « Explicabilité » du GDR IA

Les systèmes experts : MYCIN

MYCIN : Ma recommandation thérapeutique sera en premier choix : pour couvrir les points <1 2 3 5 6>, prescrire de la Gentamicine, dose 119 mg 98 h IV, pendant 10 jours;

Pour couvrir le point 4, prescrire de la Clindamycine, dose 595 mg 96 h IV, pendant 14 jours.

Désirez-vous éventuellement éliminer certains médicaments ?

Médecin : Oui, la Clindamycine.

MYCIN : Pour couvrir les points <2 3 4 5 6> donner du Chloramphénicol, dose 1 g 96 h IV, pendant 14 jours;

** Surveiller le taux de globules blancs du patient.

Pour couvrir le point 1, donner de la gentamicine, dose 119 mg 98 h IV, pendant 10 jours.

** Modifier la dose en cas de faiblesse rhénale.

Les explications des SEs

MYCIN : L'infection du malade a t-elle été contractée à l'hôpital ?

Utilisateur : pourquoi ?

TEIRESIAS : [pourquoi est-ce important de savoir cela ?]

Cela aidera à déterminer la catégorie à laquelle appartient l'organisme responsable de l'infection :

- il a déjà été établi que 1° l'organisme est à Gram négatif, 2° c'est un bâtonnet, 3° il se développe à l'air ou sans air,

- donc si l'infection par cet organisme n'a pas été contactée lors du séjour à l'hôpital

alors il est presque certain (0,8) que l'organisme est de la catégorie des enterobactoriaceae; de plus, il se peut (0,1 ; règle 37) que ce soit la pseudomonia aeruginosa.

Utilisateur : pourquoi n'avez-vous pas envisagé la possibilité que l'organisme soit un streptocoque ?

TEIRESIAS : la règle 33 aurait du être utilisée pour identifier l'organisme comme un streptocoque, mais la seconde clause "la morphologie est de type coque" était déjà connue comme fausse et la règle 33 n'a jamais été essayée.

Les explications des SEs

Pourquoi ne faut-il pas prescrire de tétracycline à un enfant de moins de 8 ans ?

Connaissances justificatives

Dépôt de la drogue sur les **os en développement**

→ **Noircissement** définitif des dents

→ Coloration socialement **indésirable**

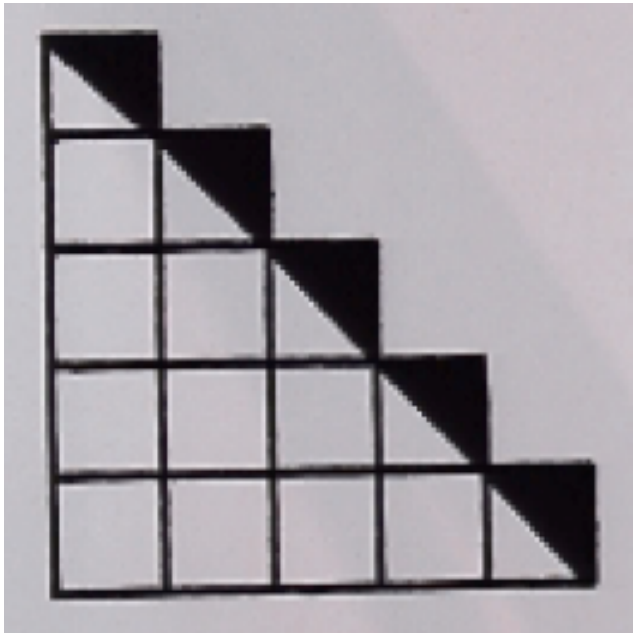
→ **Ne pas administrer** de tétracycline aux enfants de moins de 8 ans

Notion d'**effets secondaires** indésirables

Relations de **causalité**

Raisonnement graphique

$$1 + 2 + 3 + \ldots + n \overset{?}{=} \frac{n^2}{2} + \frac{n}{2}$$



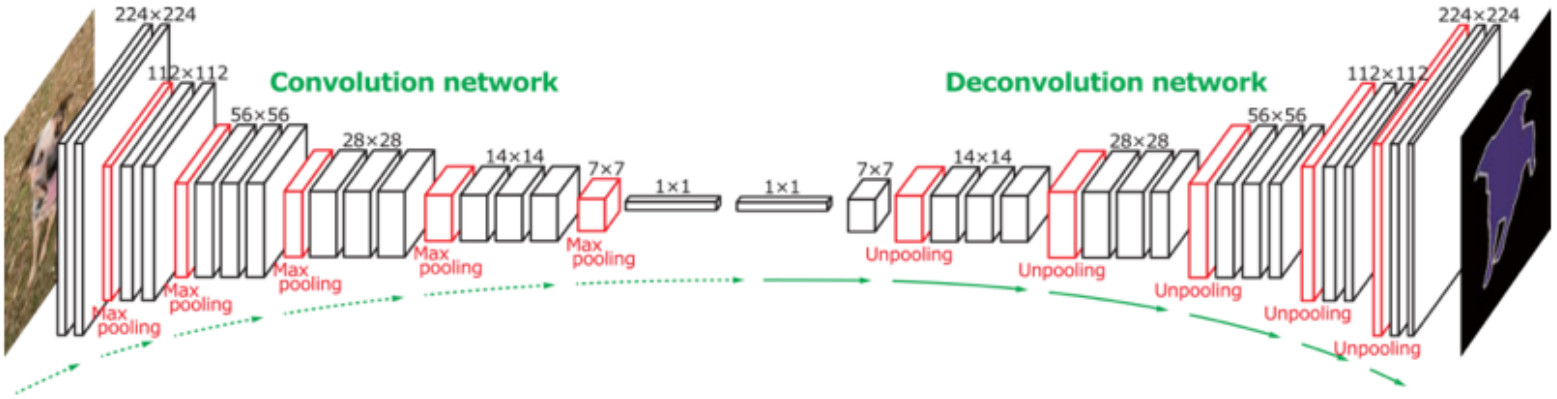
Types d'explications

- De « surface »
  - Liées au **fonctionnement** du système
    - Adéquates pour vérifier ce fonctionnement
    - **Expert IA**

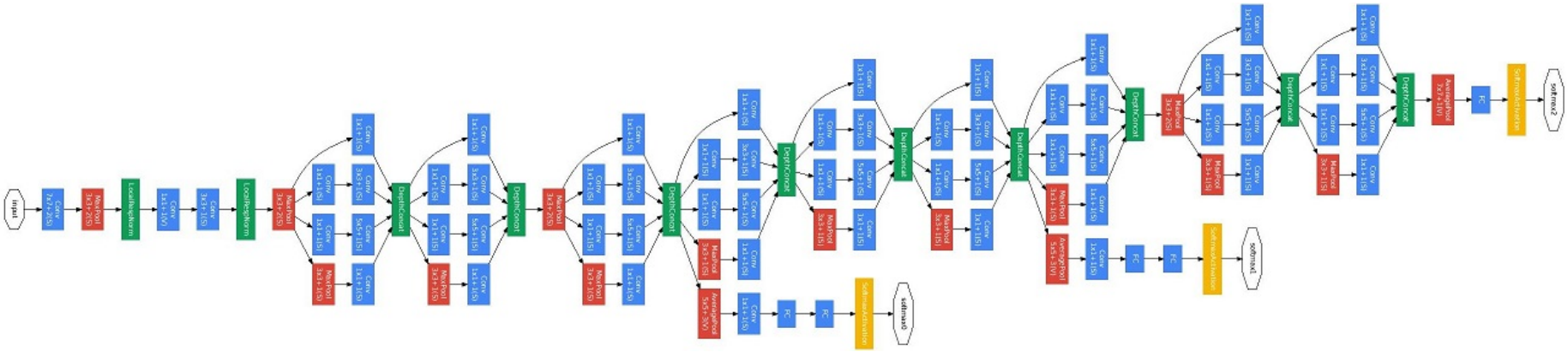
- « profondes »
  - Difficiles à **découvrir** et à **expliciter**
    - Nécessaires pour l'**expert du domaine**
    - Pour les **utilisateurs**

Les « réseaux de neurones profonds »

- Des réseaux de neurones artificiels
  - à grand nombre de couches (parfois > qqs 100)
  - et **très grand nombre de paramètres** (qqs 10^7 – 10^9 paramètres)



GoogleNet



Explications et réseaux de neurones profonds

Illusions d'optique : quelle explication ?

[Selvaraju et al. (2017) « *Grad-CAM: Visual explanations from deep networks via gradient-based localization* »]

Exemple en médecine

MACHINE LEARNING

Science

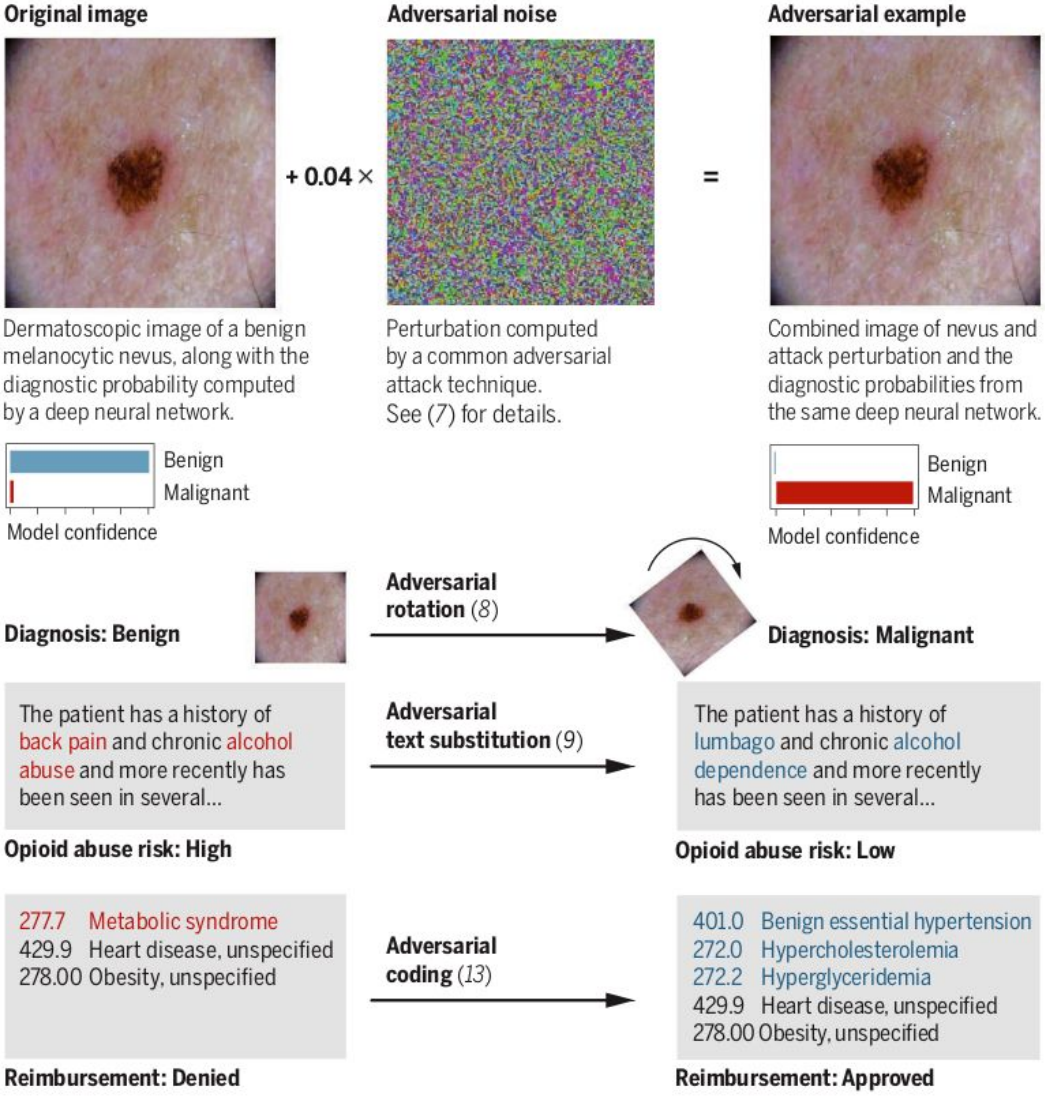
Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

22 March 2019

The anatomy of an adversarial attack

Demonstration of how adversarial attacks against various medical AI systems might be executed without requiring any overtly fraudulent misrepresentation of the data.



A **basic** principle

- Machine Learning “just” **reformulates** what has been given as input
- A conservation theorem:
  - **No information is “added”**
  - Data + prior knowledge

A **basic** principle

- Machine Learning “just” reformulates what has been given as input
- A conservation theorem:
  - **No information is “added”**
  - Data + prior knowledge

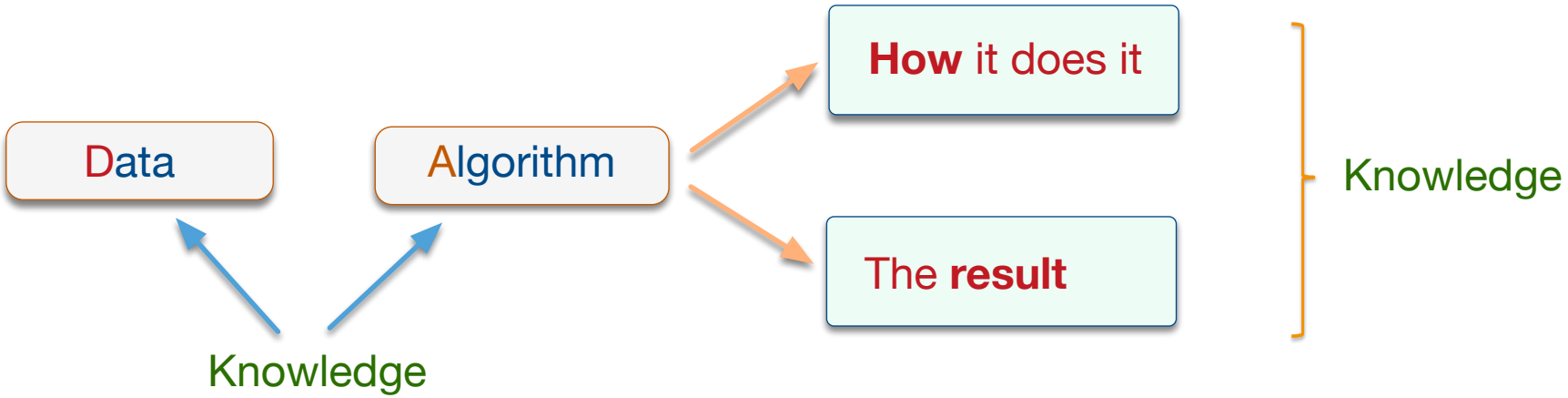
Little data + **lots** of prior knowledge

Big data + **less** prior knowledge

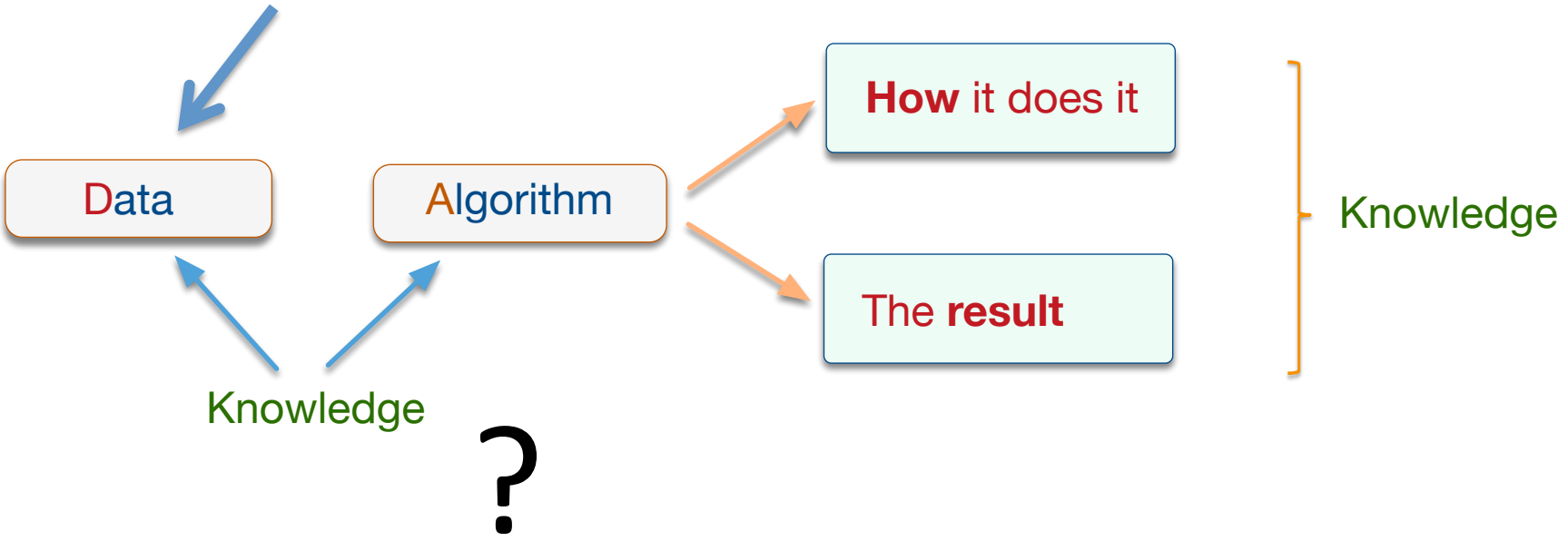
A basic principle

- Machine Learning “just” reformulates what has been given as input
- A conservation theorem:
  - **No information is “added”**
  - Data + prior knowledge

Little data + **lots** of prior knowledge
Big data + **less** prior knowledge

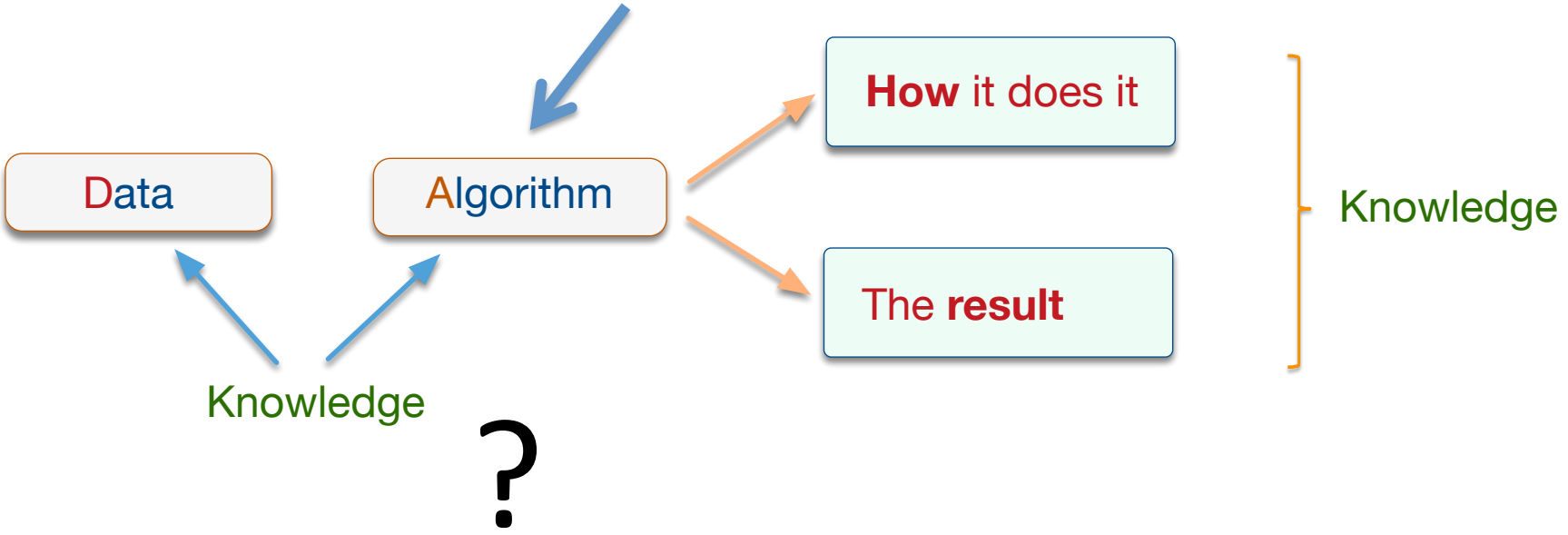


Biais sur les **données**



…

Biais sur les **algorithmes**



…

Des algorithmes « transparent »

1. Dans lesquels on puisse « injecter » l'expertise humaine

2. Dont les résultats (modèles appris) soient interprétables

Le cas AlphaGo

- Un joueur « extraterrestre »
- Un jeu stupéfiant
- Révolutionne la manière de jouer
- Effervescence dans les écoles de go

Découverte scientifique

Programme de la journée

- **Session 1** (9h30 – 11h25) Antoine Cornuéjols & Christel Vrain
  - **Pierre Marquis** « *A pinch of eXplainable AI from a Knowledge Representation Perspective* »
  - **Marie-Jeanne Lesot** « *Explications de données et de classifieurs : quelques méthodes et risques notables* »
  - Discussion
- **Session 2** (11h40 – 12h30) Engelbert Mephu Nguifo
  - **Winston Maxwell & Astrid Bertrand** « *Identifying the « right » level of explanation in a given situation* »
- **Session 3** (14h – 14h45) Stephan Brunesseaux
  - **Guilherme Alves** « *Making ML Models fairer through explantions, feature dropout and aggregation* »
- **Session 4** (15h – 15h45) Nicolas Maudet
  - **David Vigouroux** « *DEEL Challenges: Explainability* »
- **Session 5** (15h45 – 16h30) Amedeo Napoli
  - Table ronde animée

Des questions

- **Interprétabilité** vs. **explicabilité** ?
- Comment **évaluer** une explication ?
- Explications et liens de **causalité** ?
- Une explication est-elle **symbolique** par nature ?
- Explication « **one-shot** » ou par **interactions** ?
- Est-ce que expliquer donne la possibilité de **tromper le système** (puisque je sais comment la machine raisonne) ?
- Explications
  - pour les **experts IA**
  - pour les **experts du domaine**
  - pour les **utilisateurs**