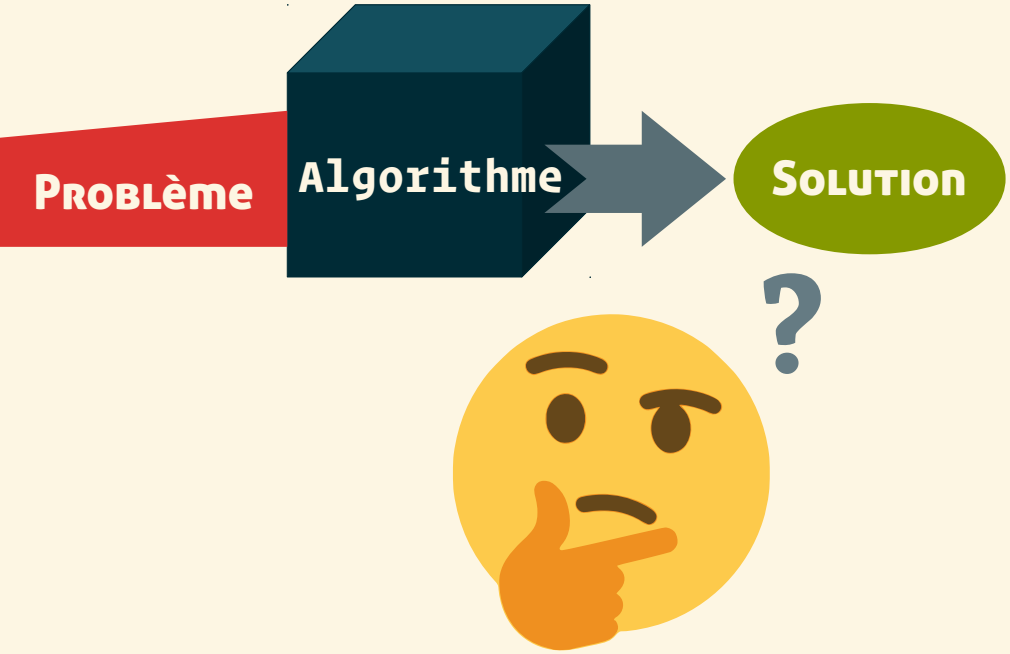
Comment expliquer des résultats d'algorithmes ?

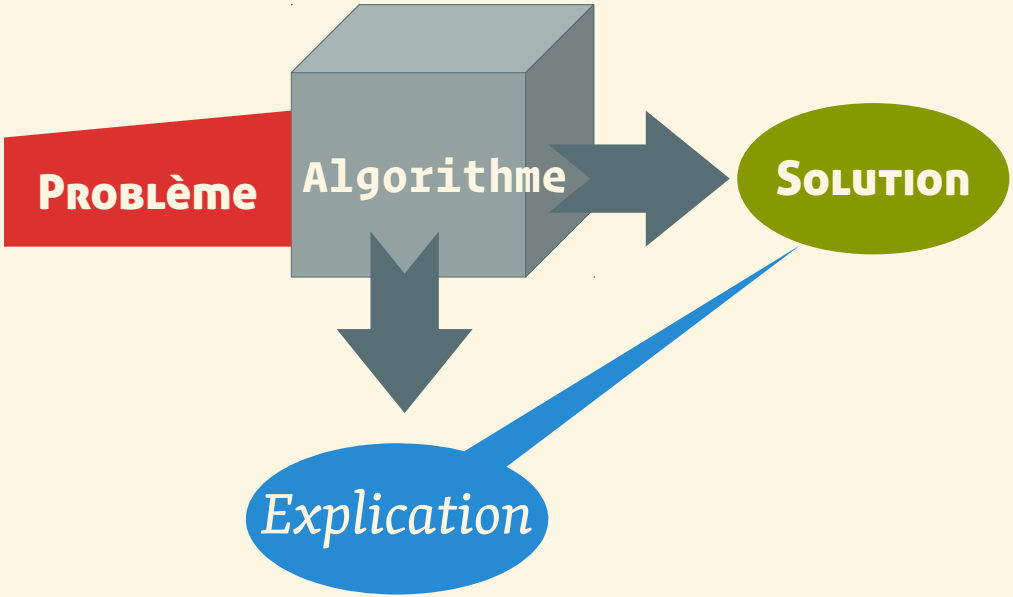
Antoine GRÉA / SMA / LIRIS / Lyon 1

Comment ça expliquer ?

- D'habitude on a :



- Ce que l'on veux :



Pourquoi on ferais ça ?

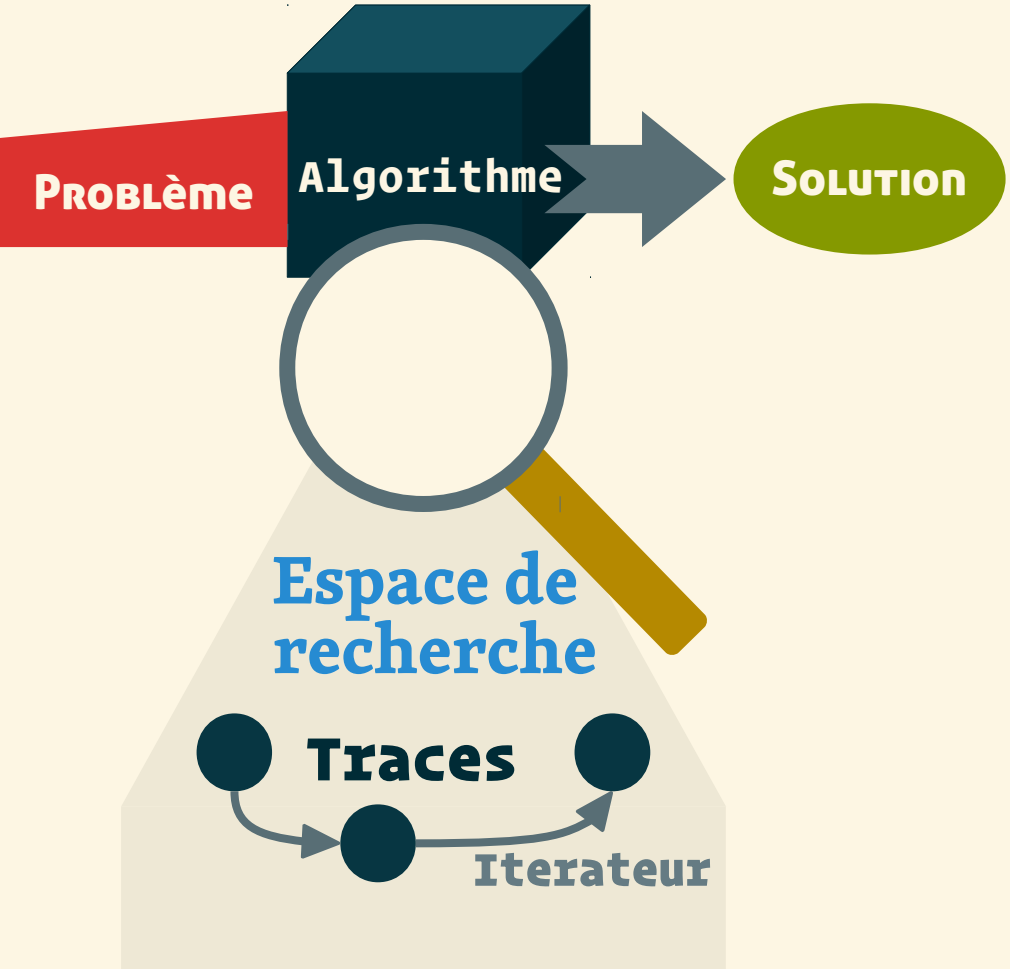
- RGPD :
  - Traitements transparents
- Responsabilité
- Interaction avec les utilisateurs
- Diagnostique

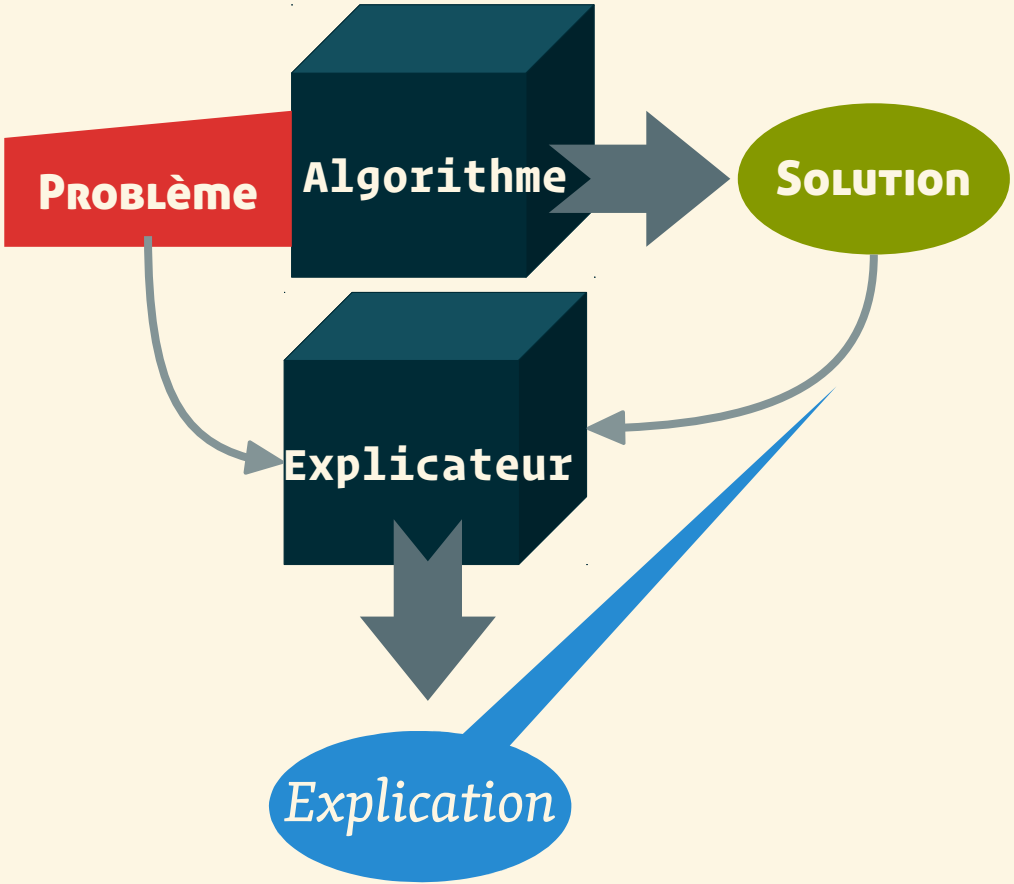
C’est quoi une explication ?

- Une description du processus de décision
  - Des données d’entré
  - Au résultat

- ✔ Souvent en réponse à une demande
  - Doit répondre à l’incompréhension
- ✔ Doit être compréhensible
  - En language naturel ou proche
  - Simple et concis

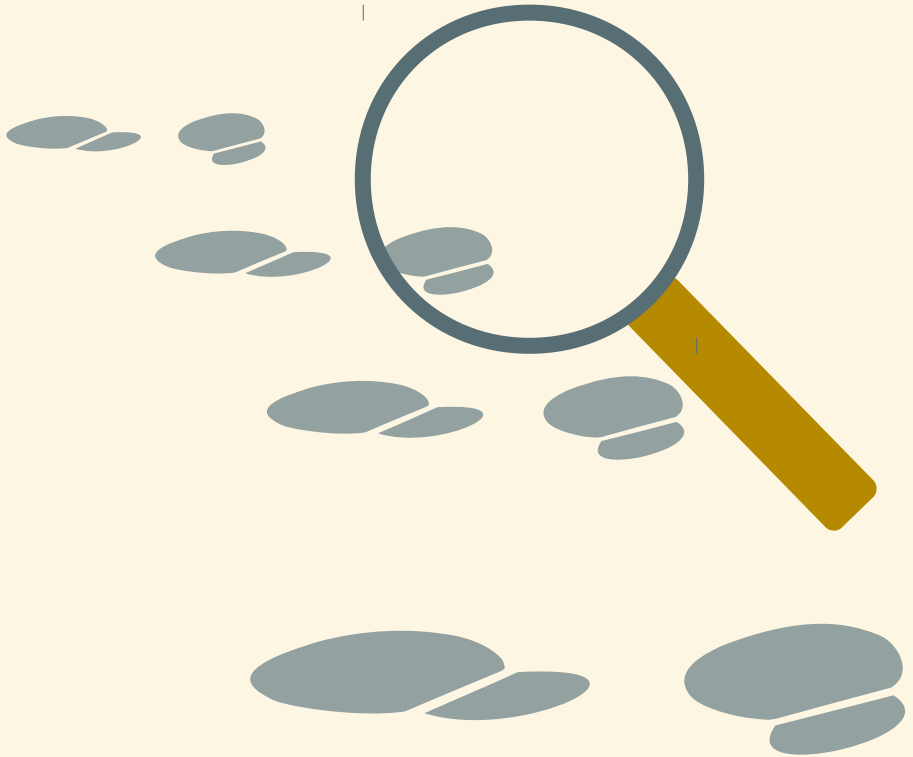
Comment s'expliquer ?





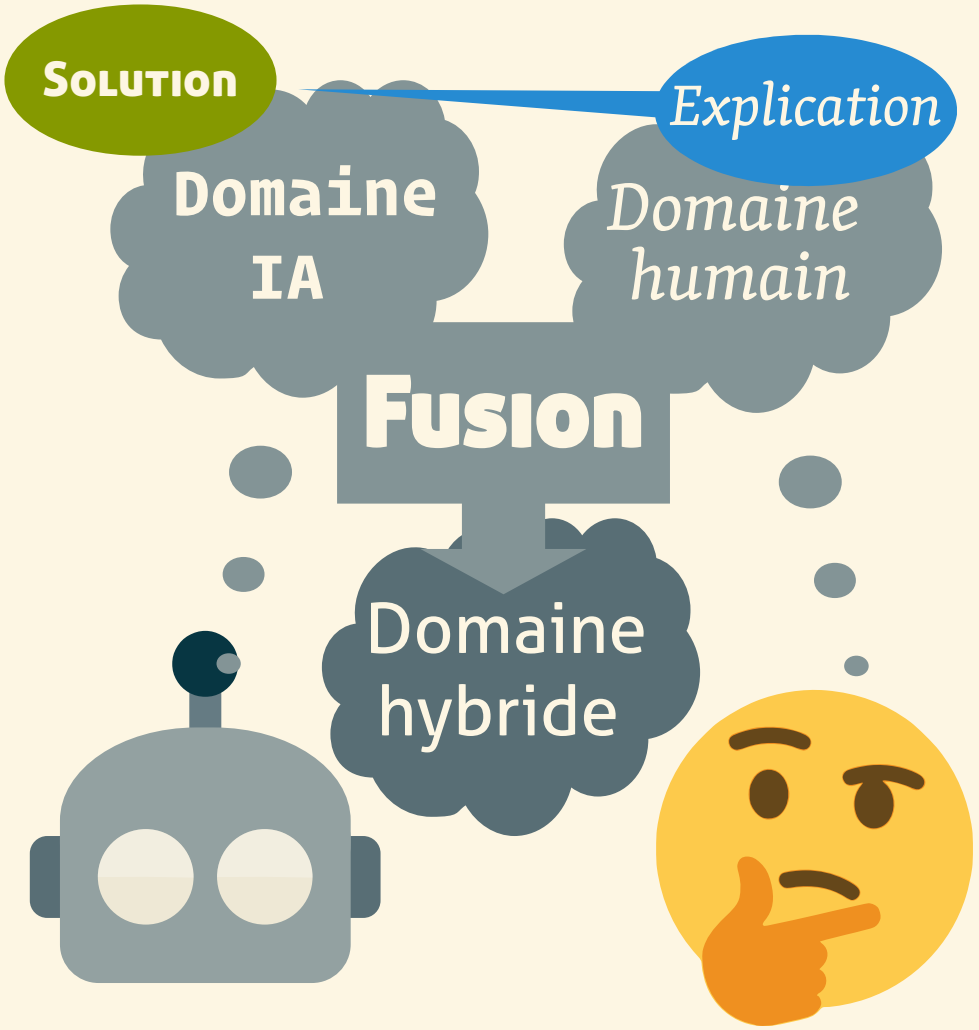
Analyse de trace

- Si peu de traces
  - Construction sémantique
  - Abstraction synthétique
- Si beaucoup de traces
  - Statistiques
  - Identifier les décisions importantes
  - Inférence d'intention



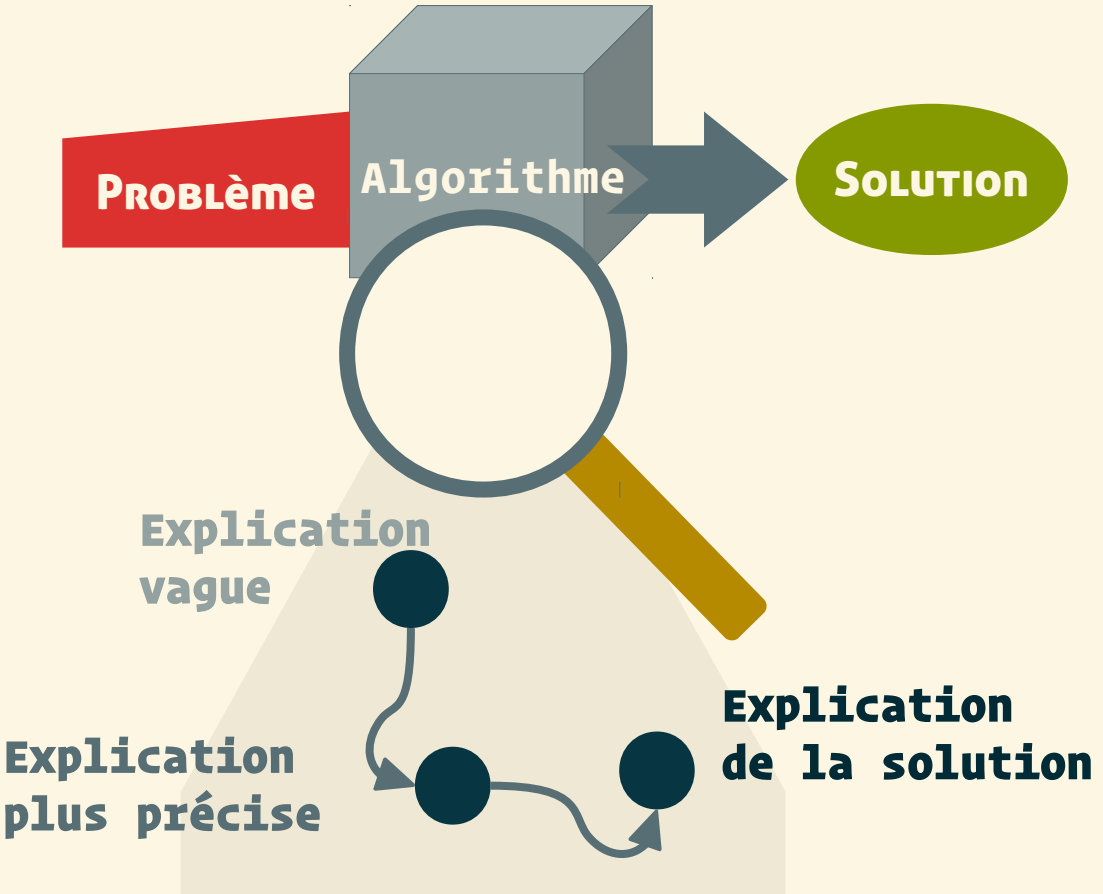
Fusion de domaine

- Recherche dans l'espace des domaines
  - Concilier le domaine IA au domaine humain
  - Fusion de notion ou dérivation
- Domaine hierarchique ou hybride



Pour être encore plus efficace

- Utiliser l'explication comme heuristique
- Utiliser une structure de donnée expressive
- Utiliser un itérateur de raffinement



Références

Bercher P., Biundo S., Geier T., Hoernle T., Nothdurft F., Richter F., & Schattenberg B. Plan, Repair, Execute, Explain - How Planning Helps to Assemble your Home Theater. In Proceedings of the 24th International Conference on Automated Planning and Scheduling (ICAPS 2014), AAAI Press, 2014, p. 386-394. Consulté à l'adresse http://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.090/Publikationen/2014/Bercher14PlanRepairExecuteExplain.pdf

Cardoso R. C., & Bordini R. H. A Multi-Agent Extension of a Hierarchical Task Network Planning Formalism, 2017. Consulté à l'adresse https://gredos.usal.es/jspui/handle/10366/133640

Chakraborti T., Sreedharan S., & Kambhampati S. Human-Aware Planning Revisited: A Tale of Three Models, s. d.

Fox M., Long D., & Magazzeni D. Explainable Planning. In Proceedings of IJCAI Workshop on Explainable AI, Melbourne, Australia, 20 août 2017.

Sreedharan S., Srivastava S., & Kambhampati S. Hierarchical Expertise-Level Modeling for User Specific Robot-Behavior Explanations. In International Joint Conference on Artificial Intelligence 2018.

Zhang Y., Sreedharan S., Kulkarni A., Chakraborti T., Zhuo H. H., & Kambhampati S. Plan Explicability and Predictability for Robot Task Planning. arXiv preprint arXiv:1511.08158, 2015. Consulté à l'adresse https://arxiv.org/abs/1511.08158

Zhang Y., Zhuo H. H., & Kambhampati S. Plan explainability and predictability for cobots. CoRR abs/1511.08158, 2015. Consulté à l'adresse https://pdfs.semanticscholar.org/2641/dab480b118f1f048153078f54aa93c9389ad.pdf