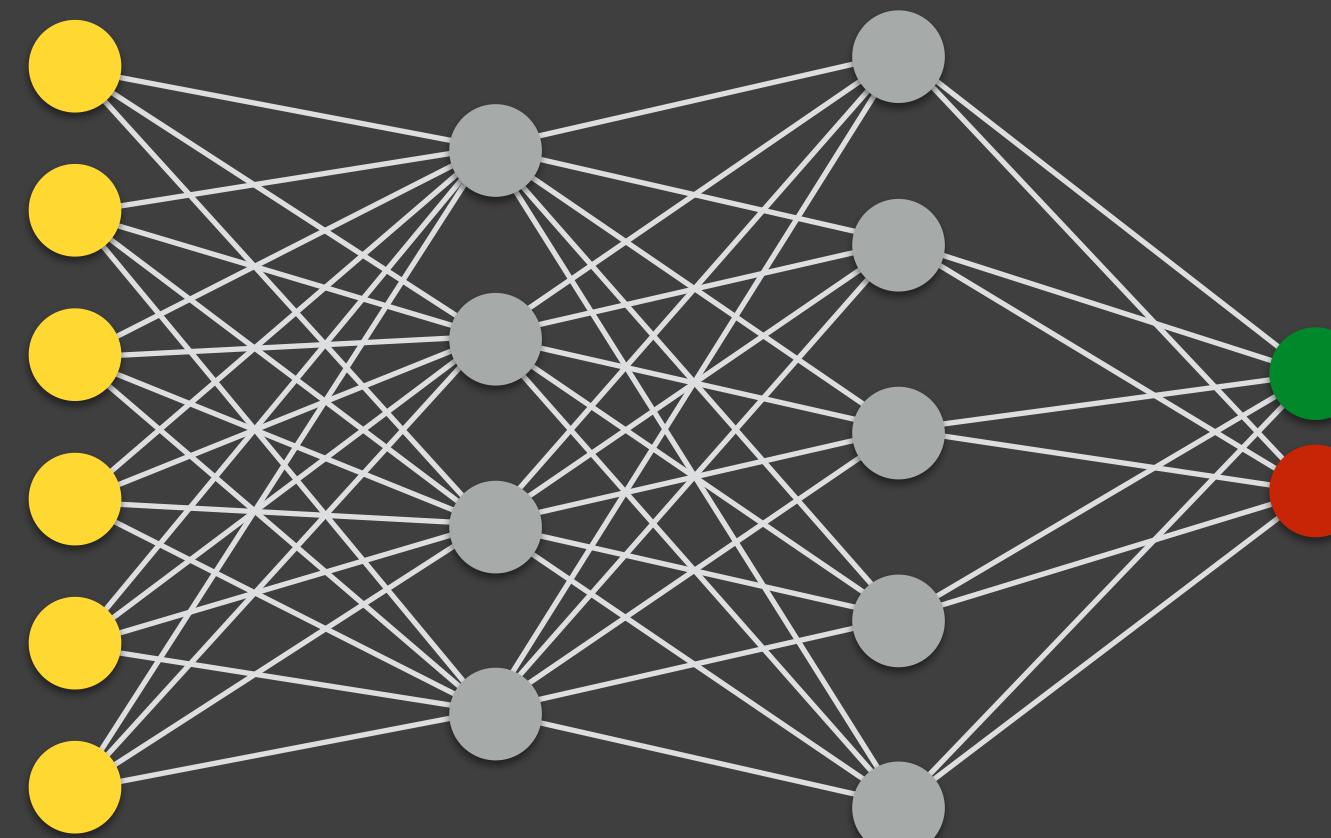


Interprétation Abstraite des Réseaux de Neurones



Caterina Urban

Équipe-projet ANTIQUE (ANalise StaTIQUE par Interprétation Abstraite)



Qui suis-je ?



1987

Udine, Italie

2006 - 2011

Università degli Studi di Udine

2011 - 2015

École Normale Supérieure

2015

NASA & Carnegie Mellon University

2015 - 2019

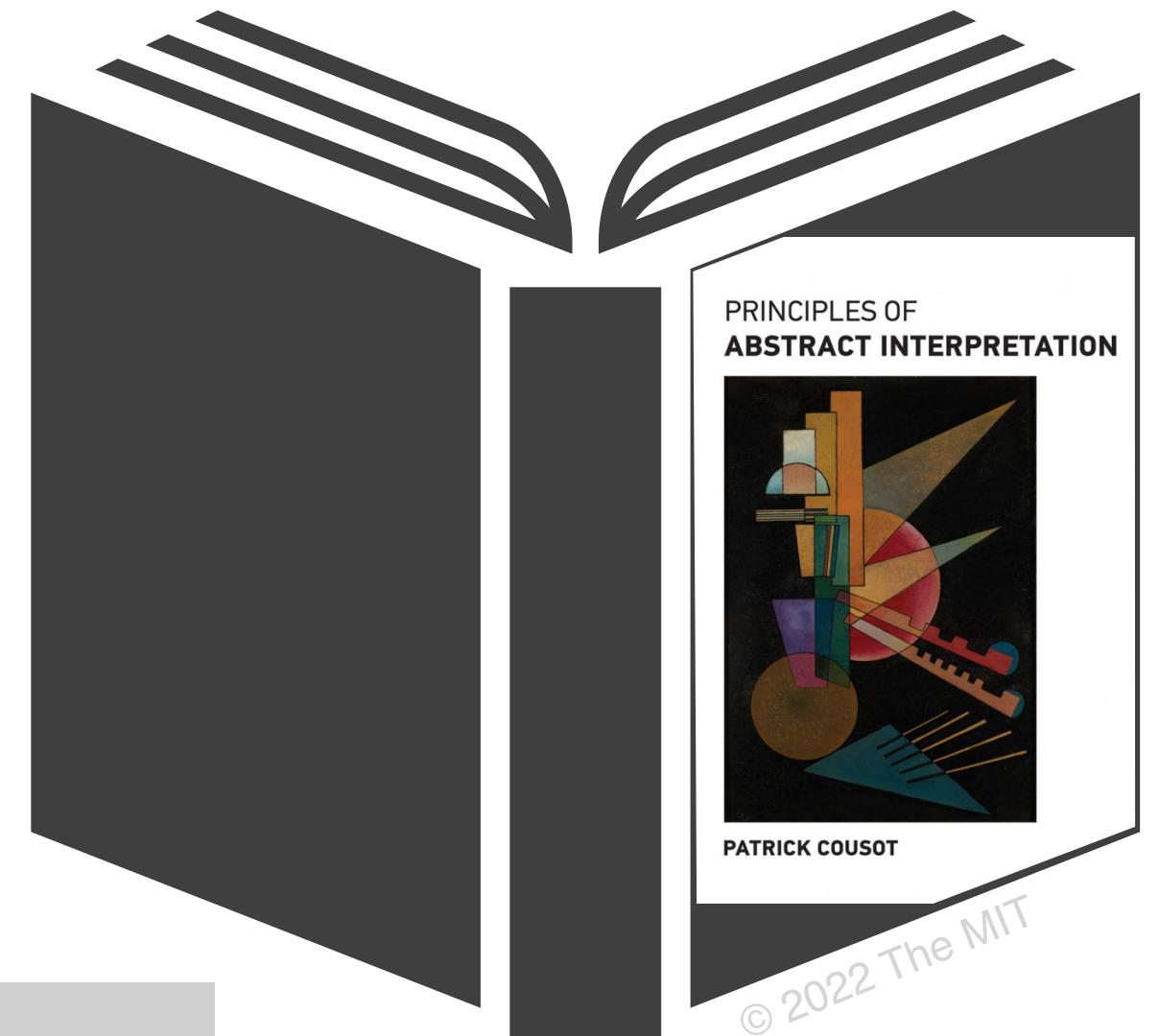
ETH Zurich

Depuis 2019

Inria



Que fais-je?



© 2022 The MIT

Analyse Statique par Interprétation Abstraite

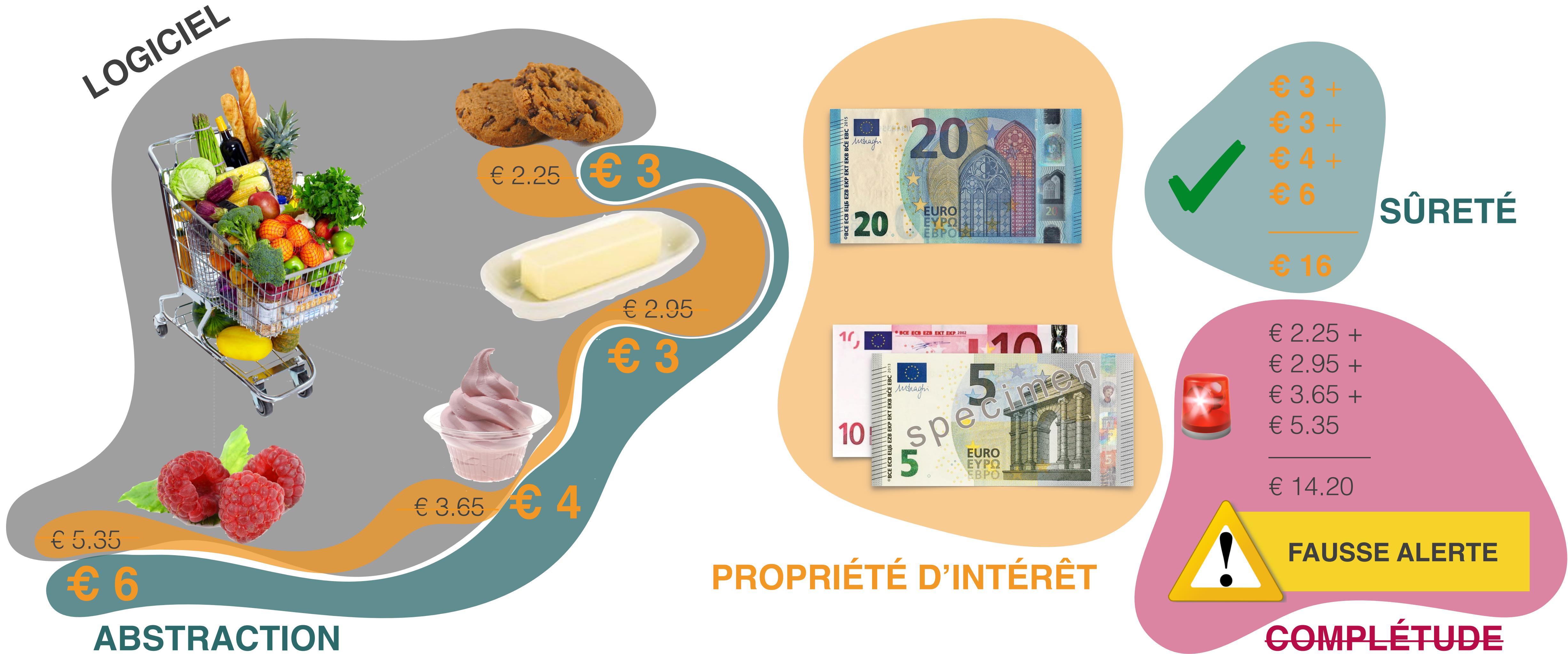


pour fournir ***automatiquement*** des
garanties mathématiques rigoureuses
sur le comportement d'un logiciel

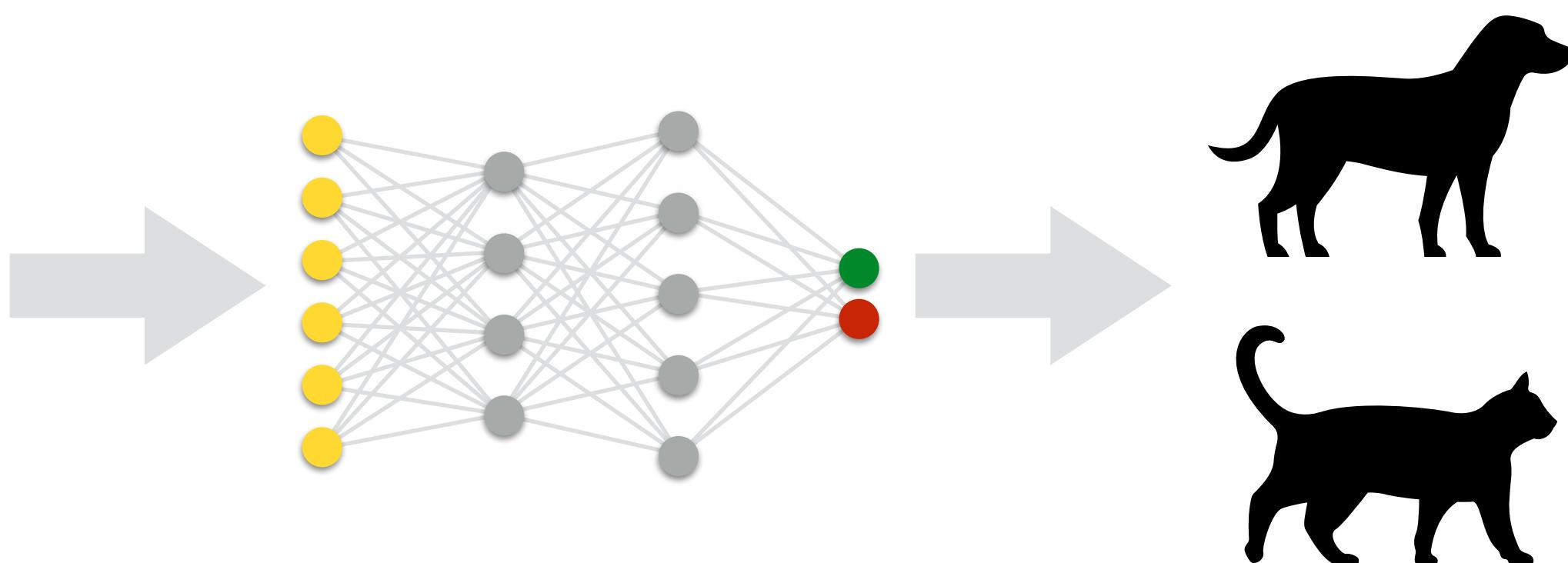
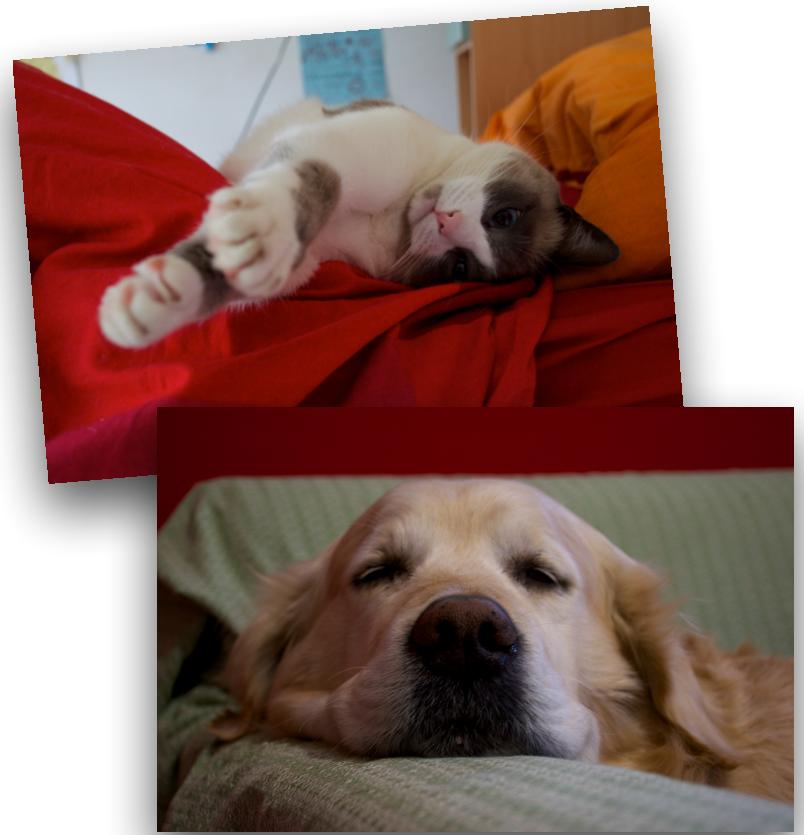
Interprétation Abstraite, kezako ?!



Interprétation Abstraite, kezako ?!



Réseaux de Neurones



A self-driving Uber ran a red light last December, contrary to company claims

Internal documents reveal that the car was at fault

By Andrew Liptak | @AndrewLiptak | Feb 25, 2017, 11:08am EST

Feds Say Self-Driving Uber SUV Did Not Recognize Jaywalking Pedestrian In Fatal Crash

Richard Gonzales November 7, 2019 10:57 PM ET



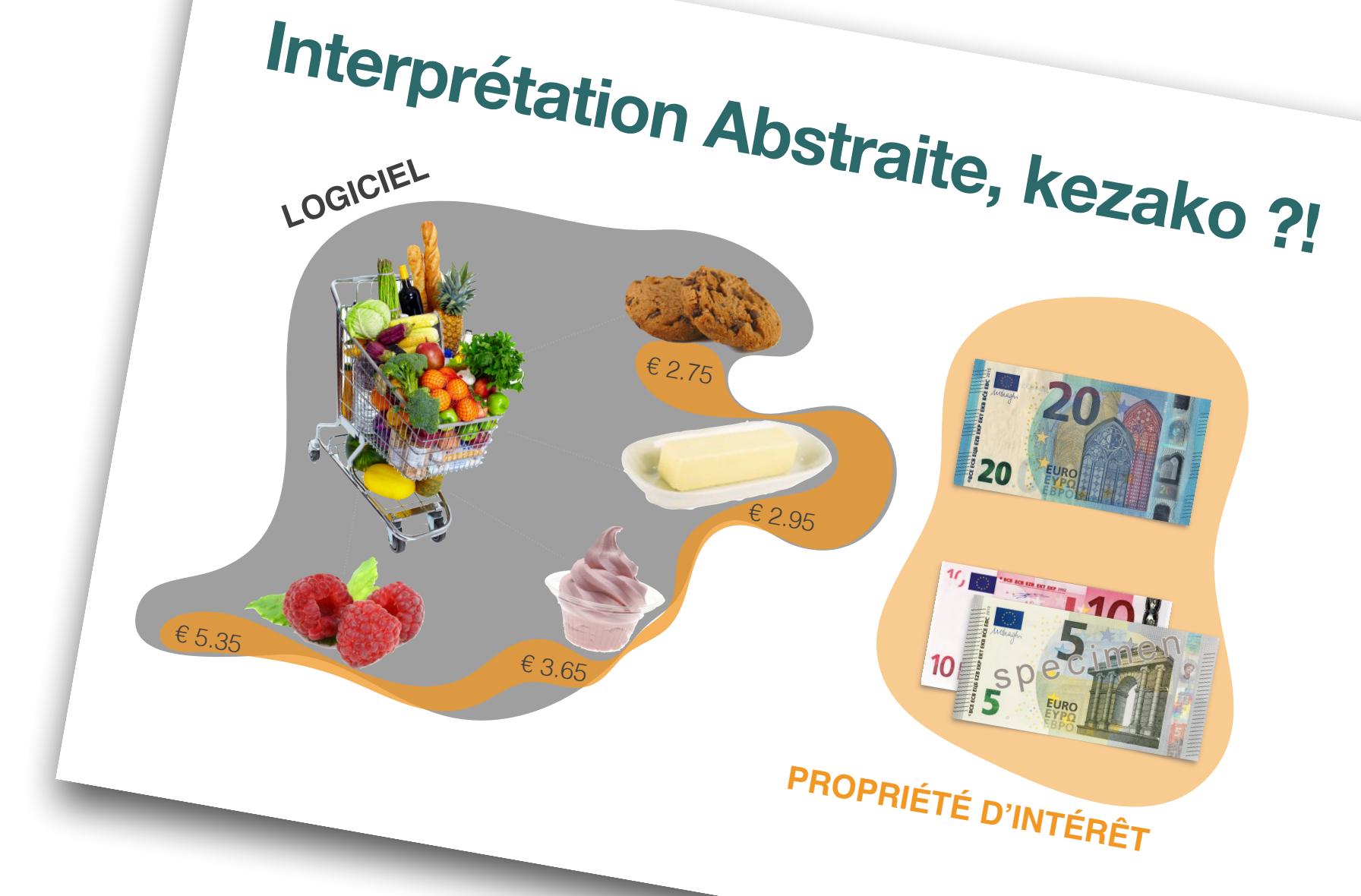
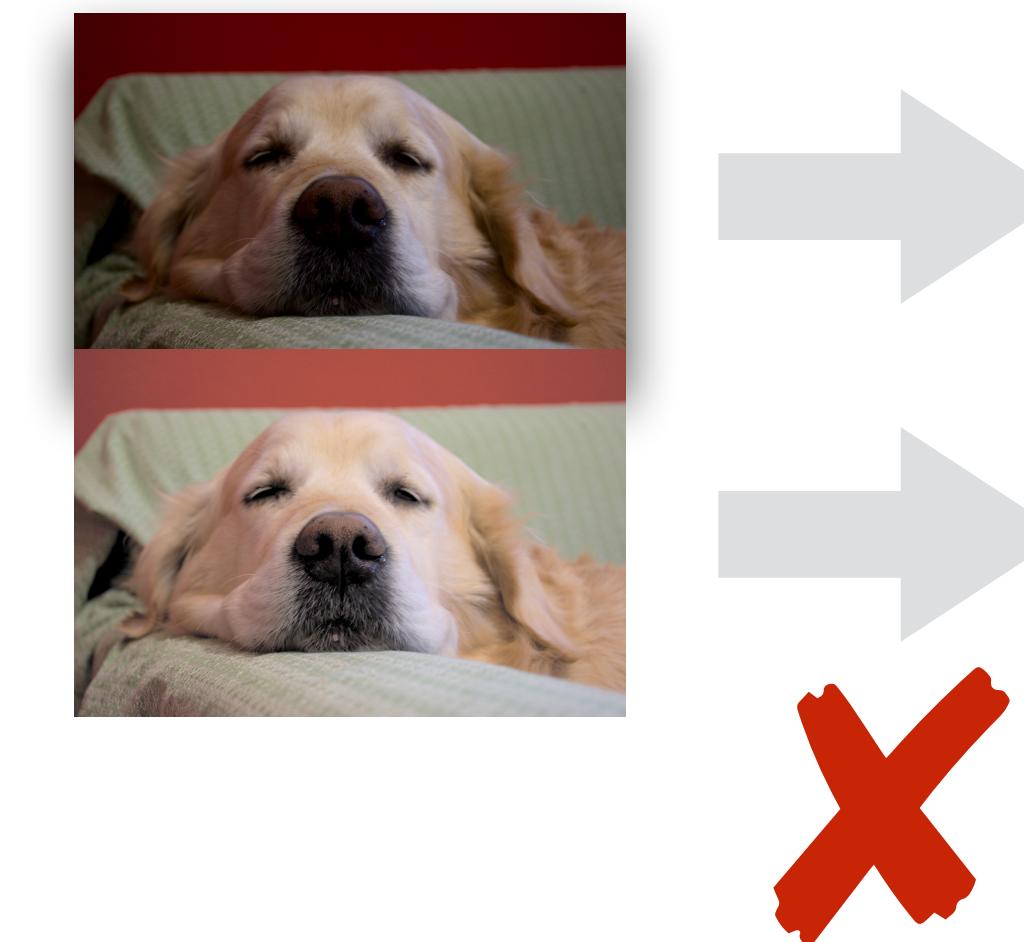
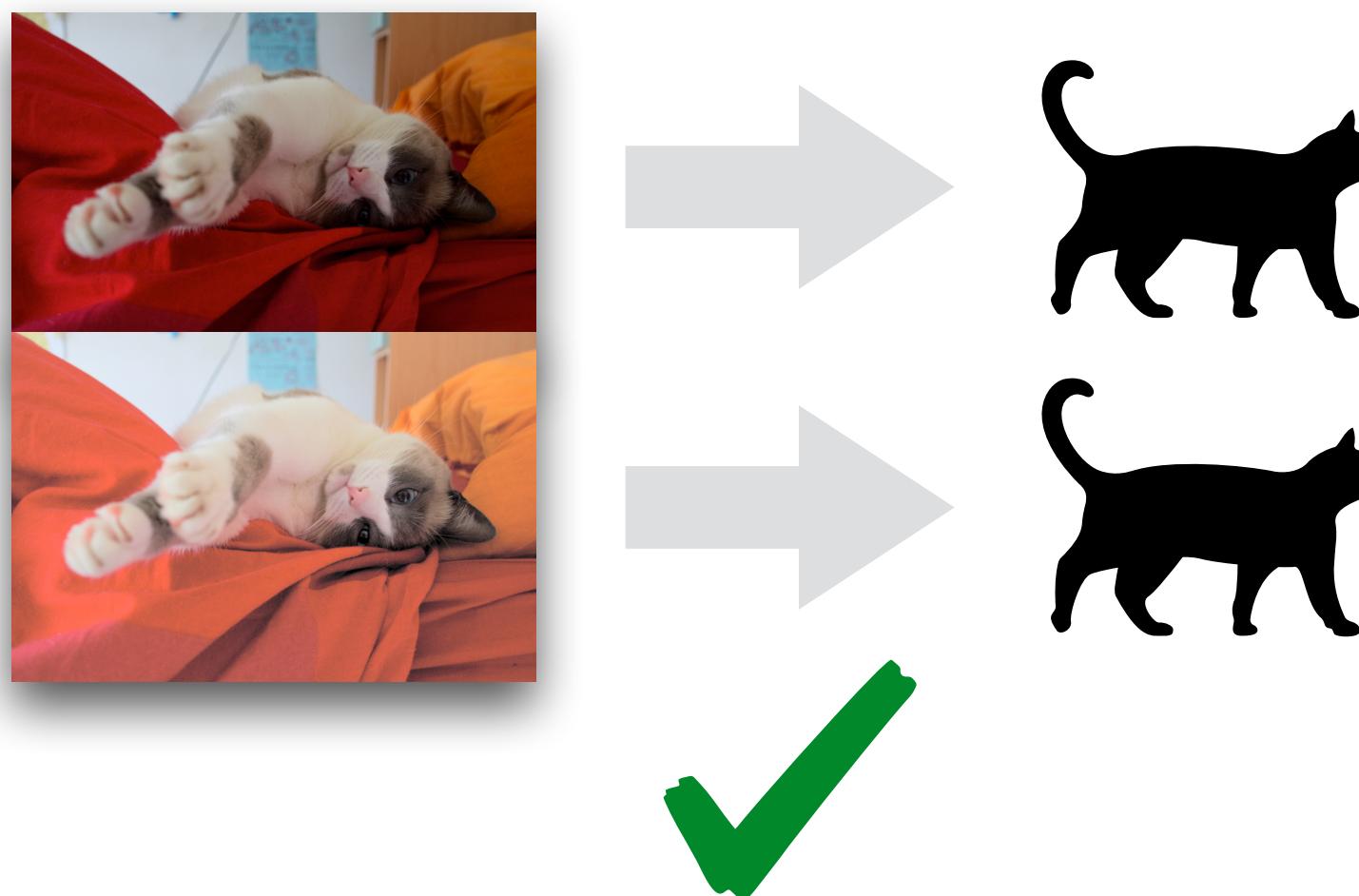
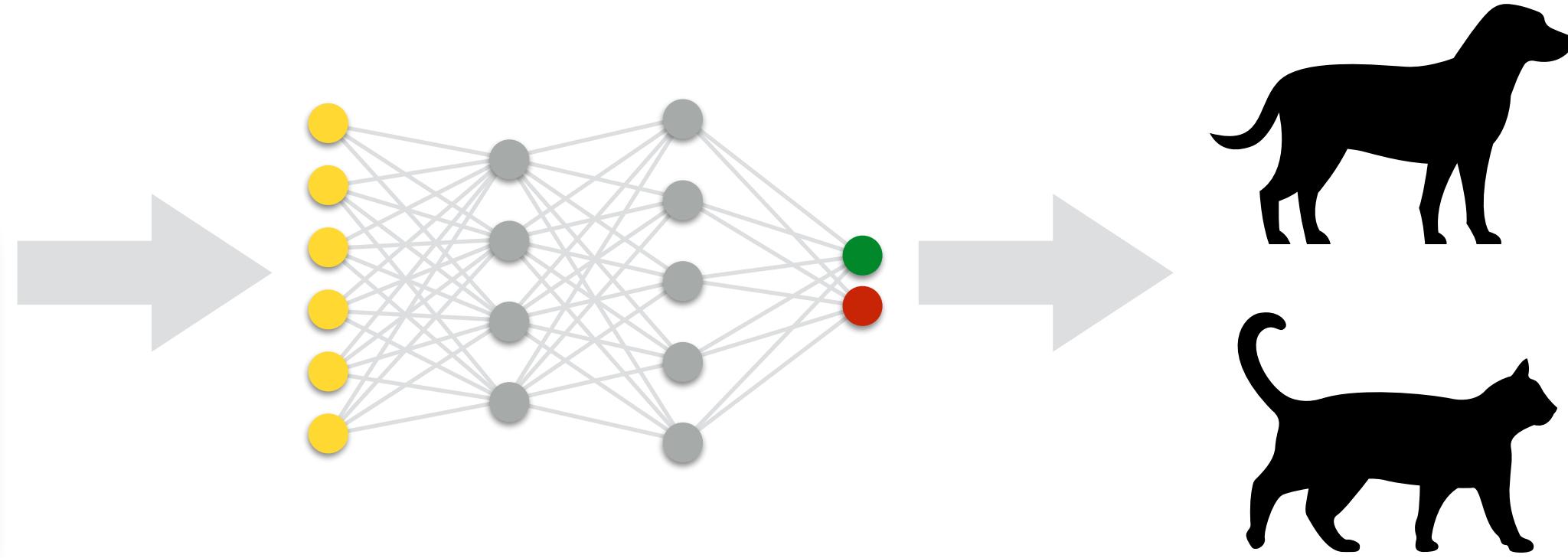
Voitures Autonomes



Roulage, Décollage, Atterrissage Autonomes

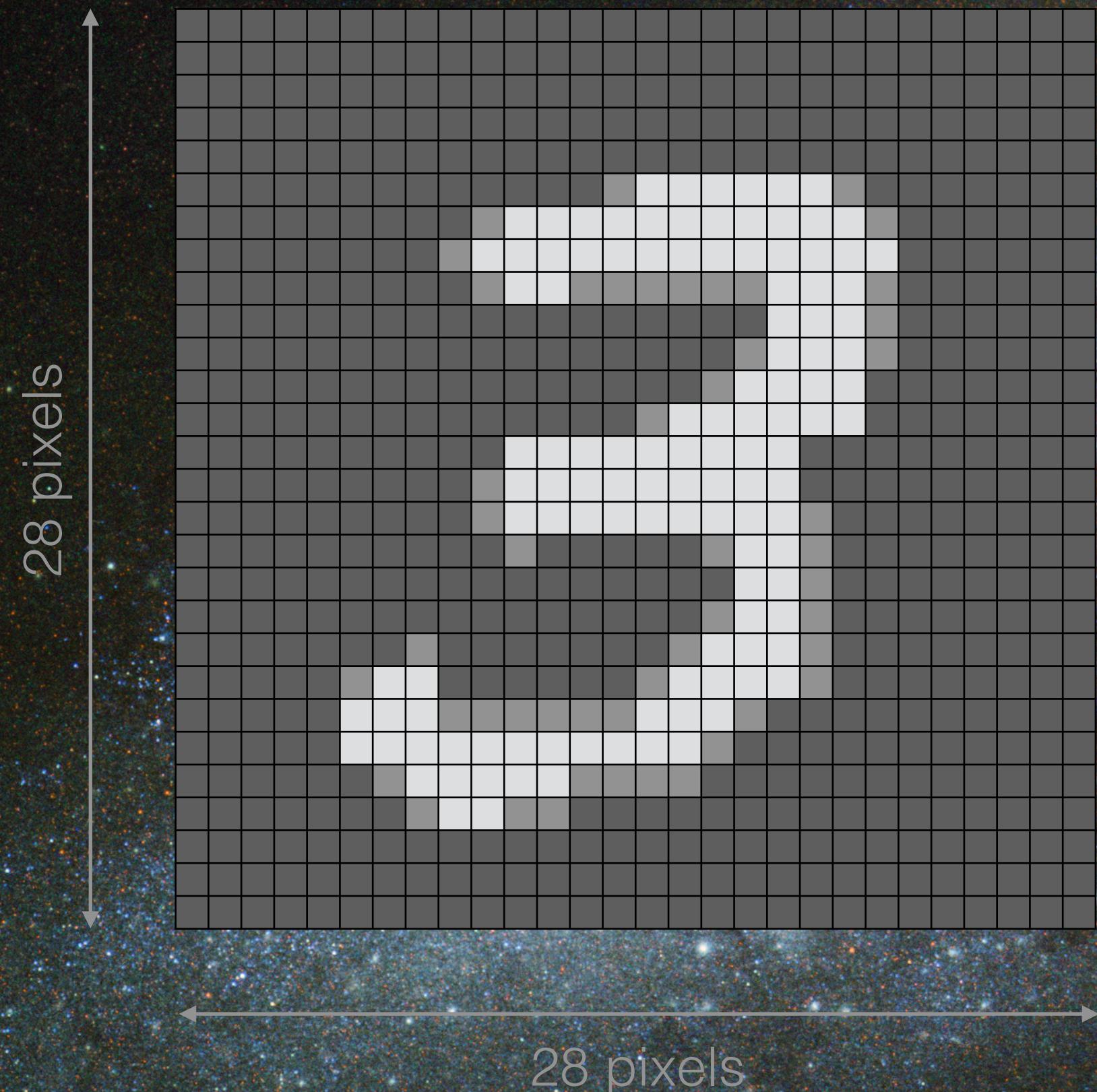


Stabilité Locale



Stabilité Locale

Bruit Aléatoire



avec uniquement des pixels **BLANCS** ou **NOIRS**
nous avons déjà 2^{784} ($\simeq 10^{236}$) images possibles !

plus que le nombre estimé d'atomes
dans l'univers visible ($\simeq 10^{80}$) !



Sur-approximation et Analyse *En Avant*

Interprétation Abstraite, kezako ?!

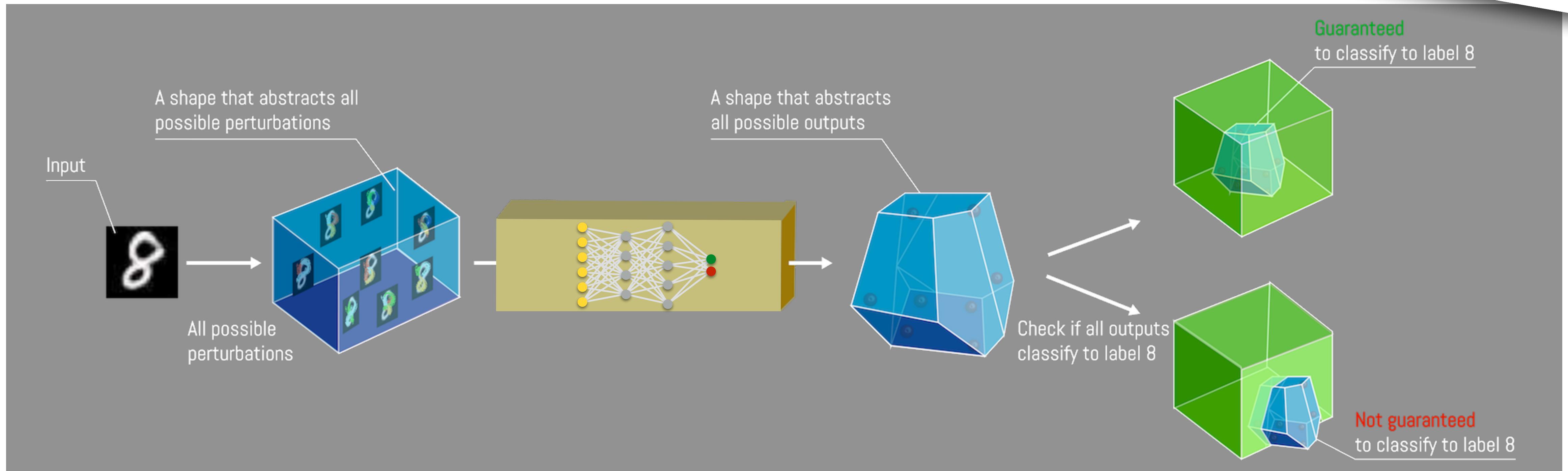
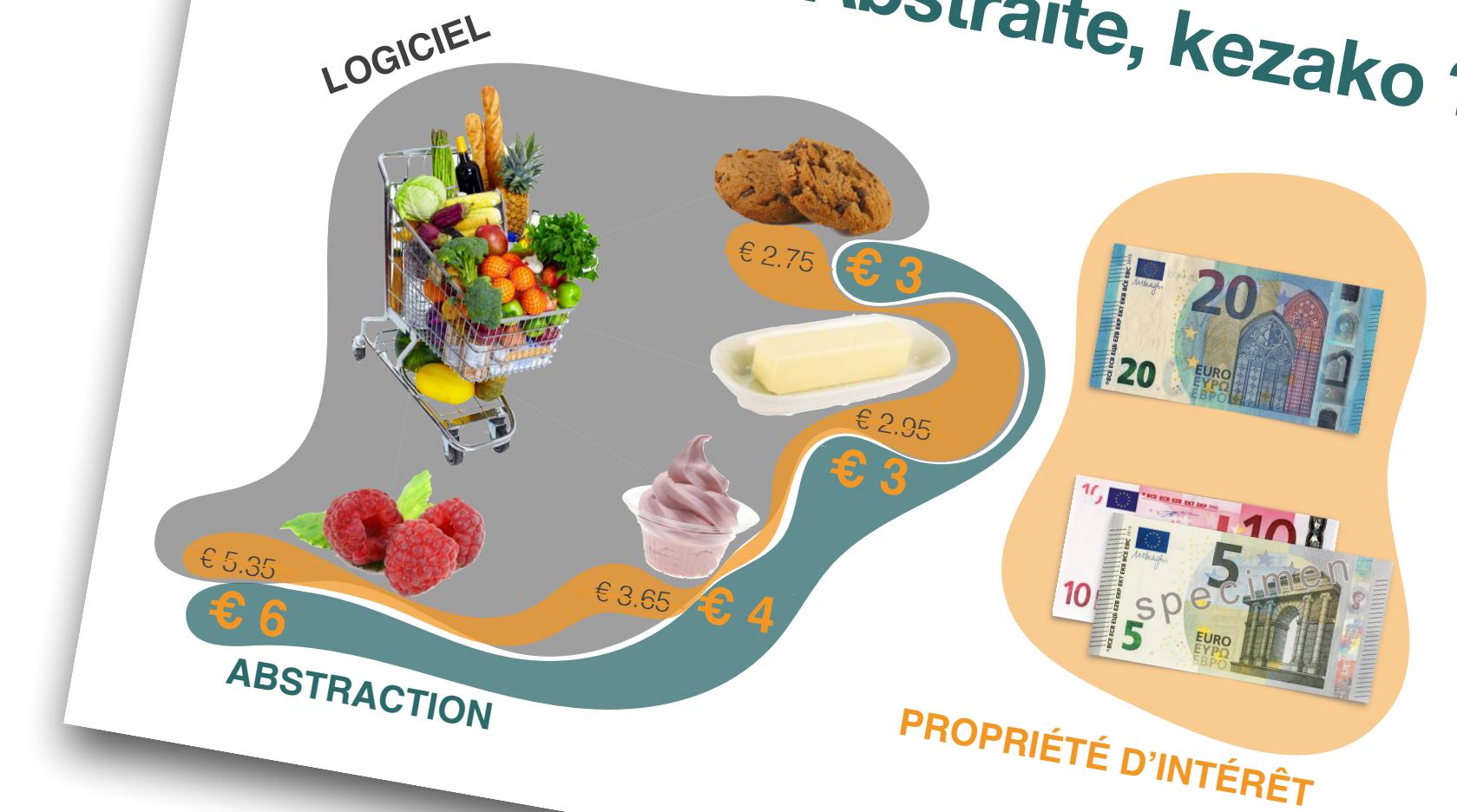
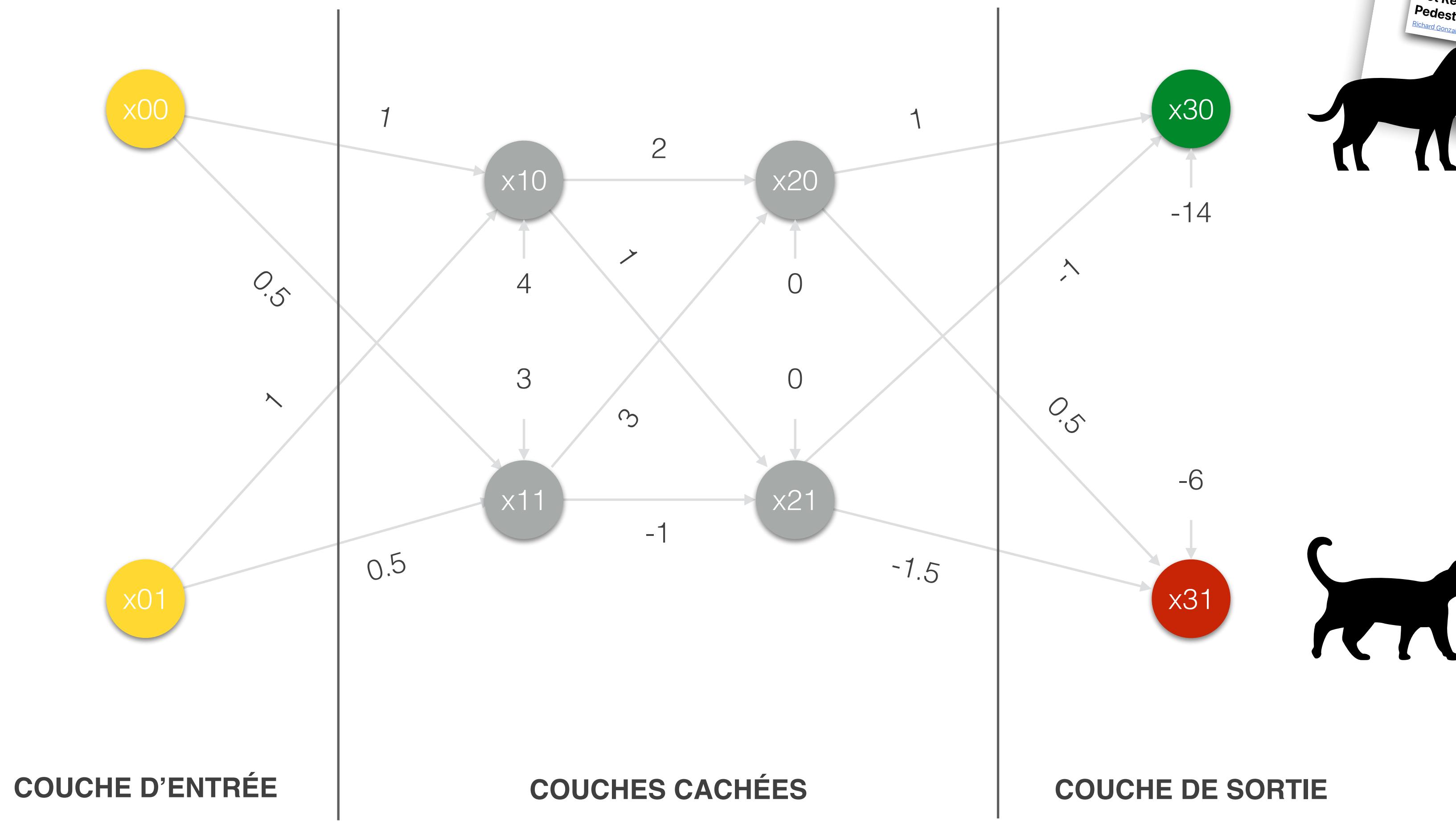


Image tirée de <http://safeai.ethz.ch>

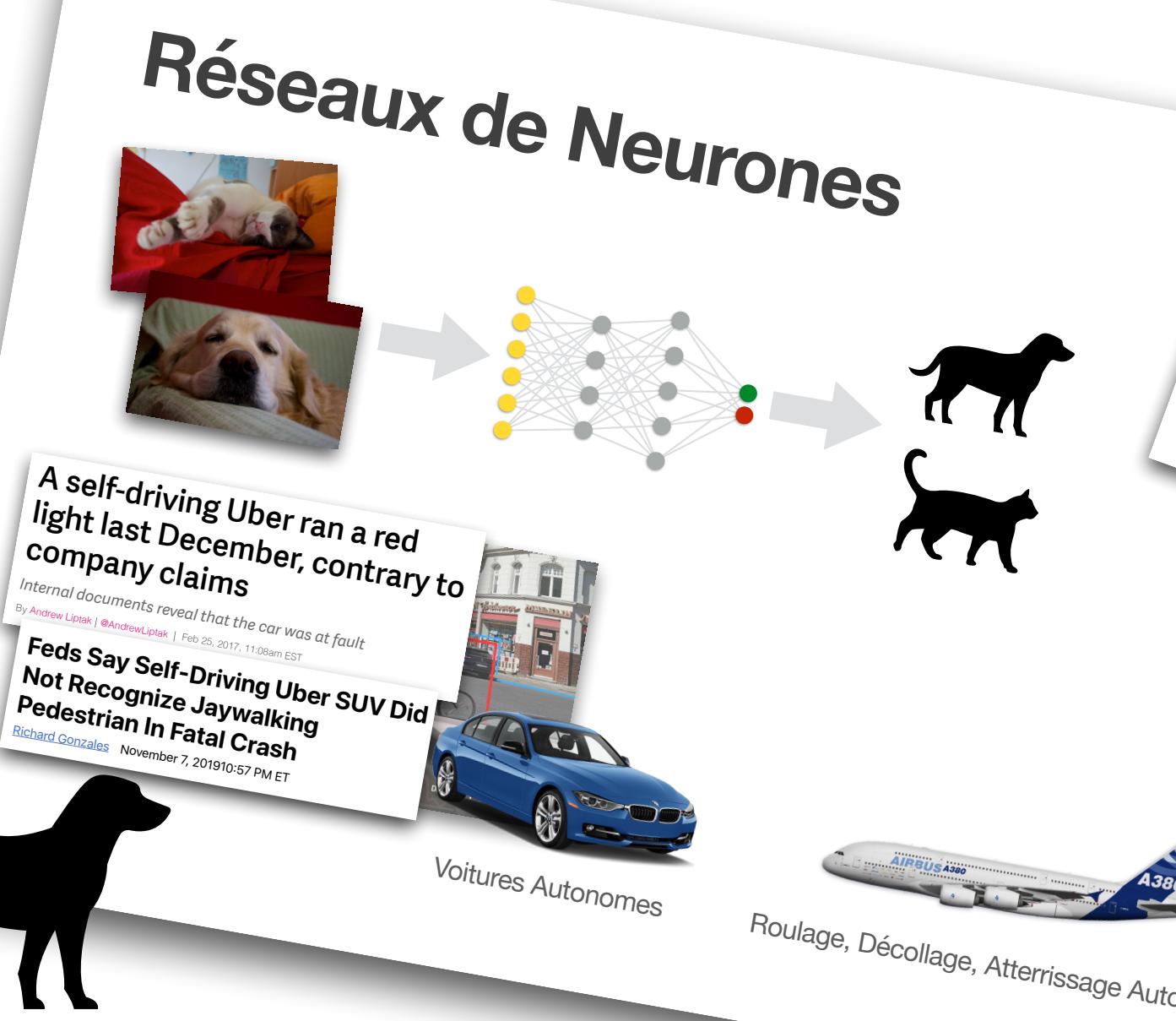
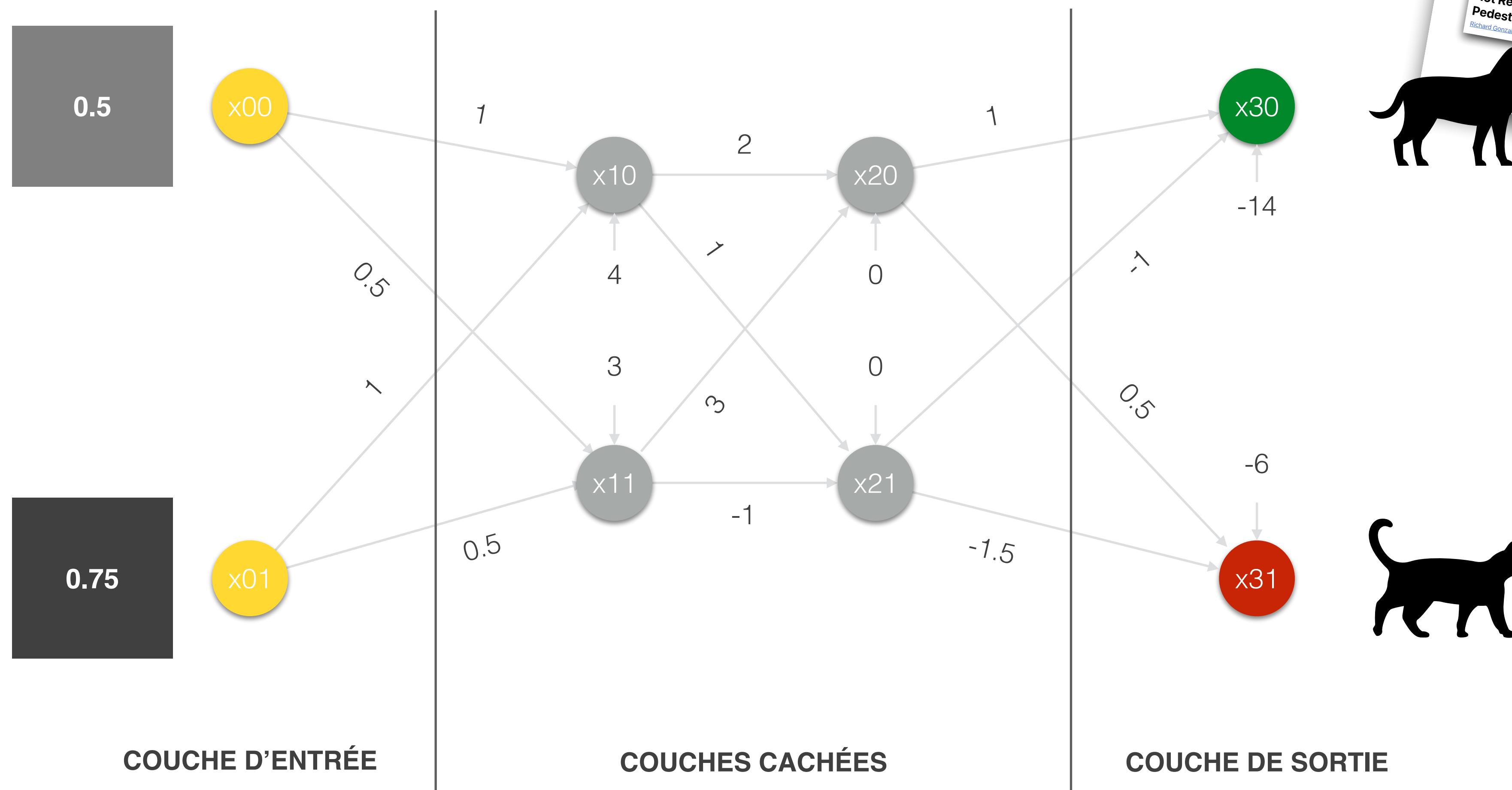
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU



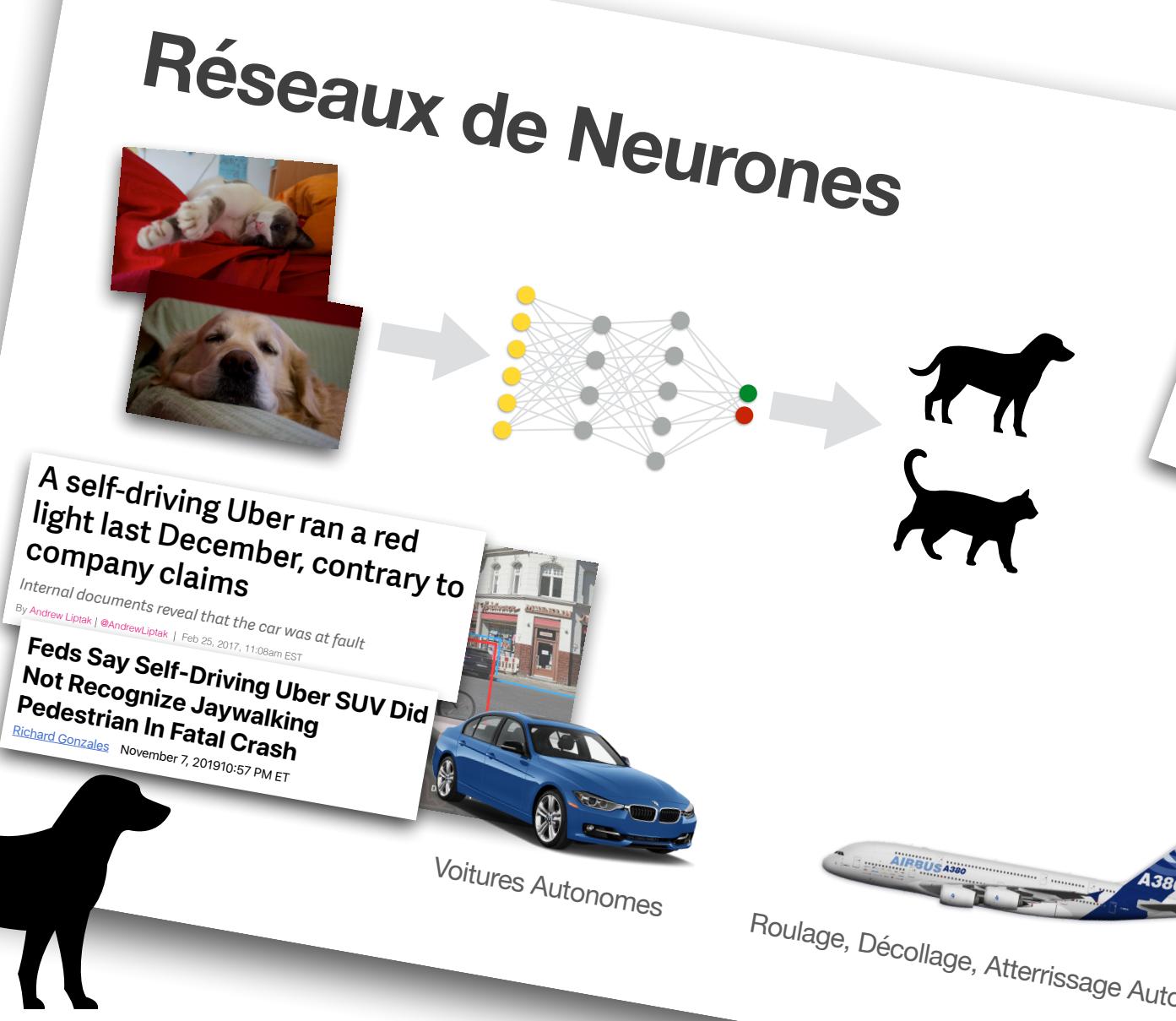
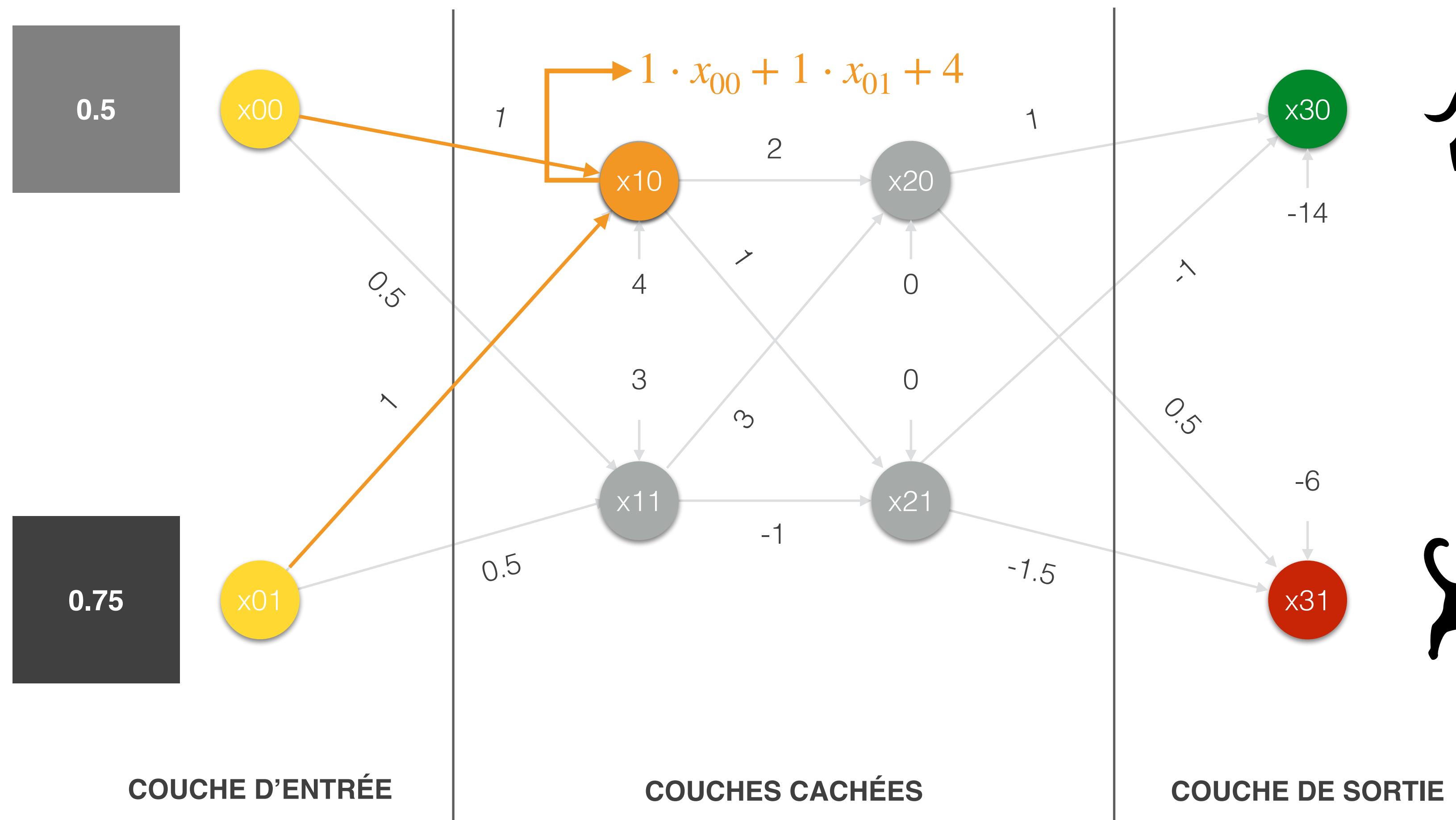
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU



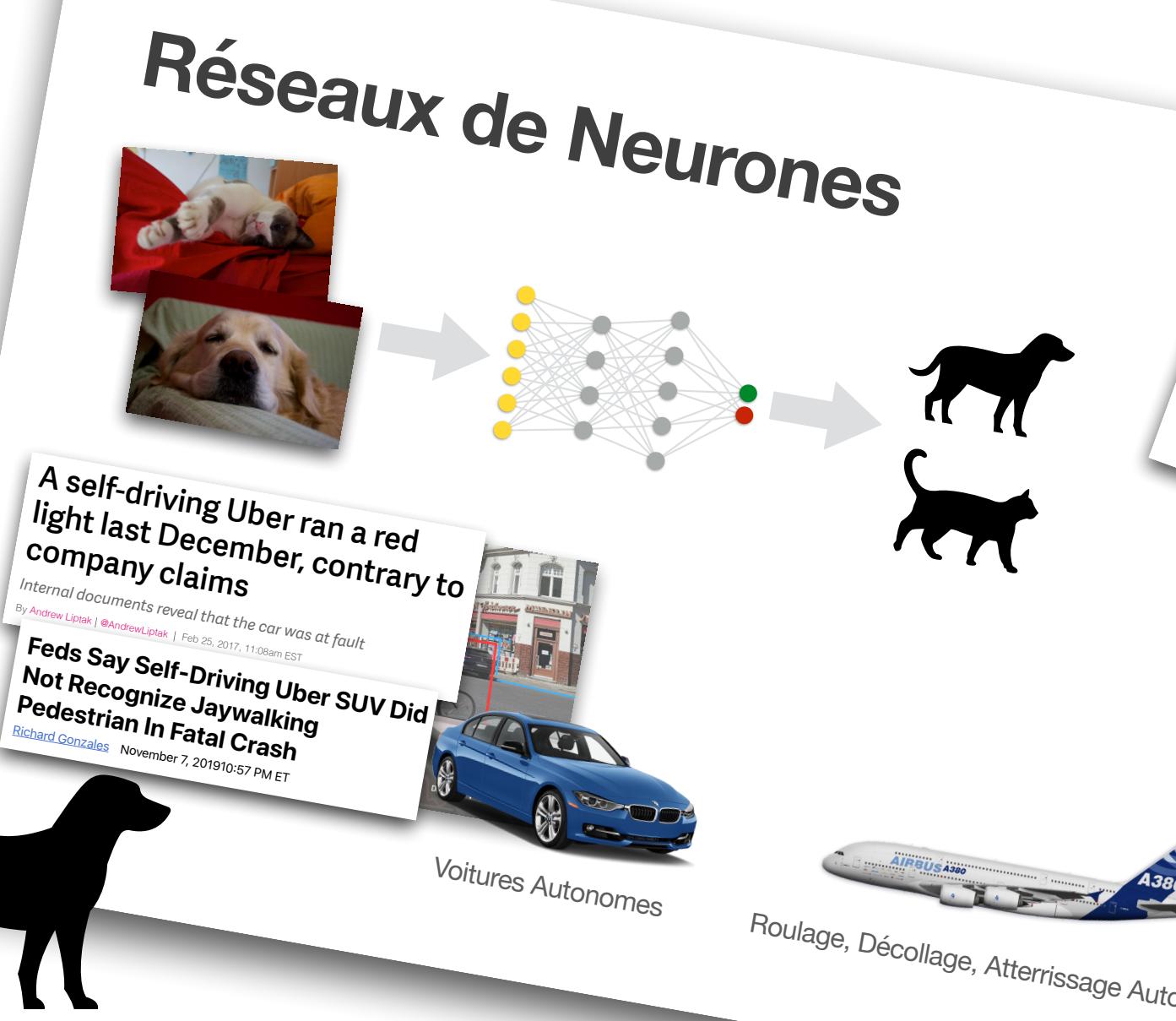
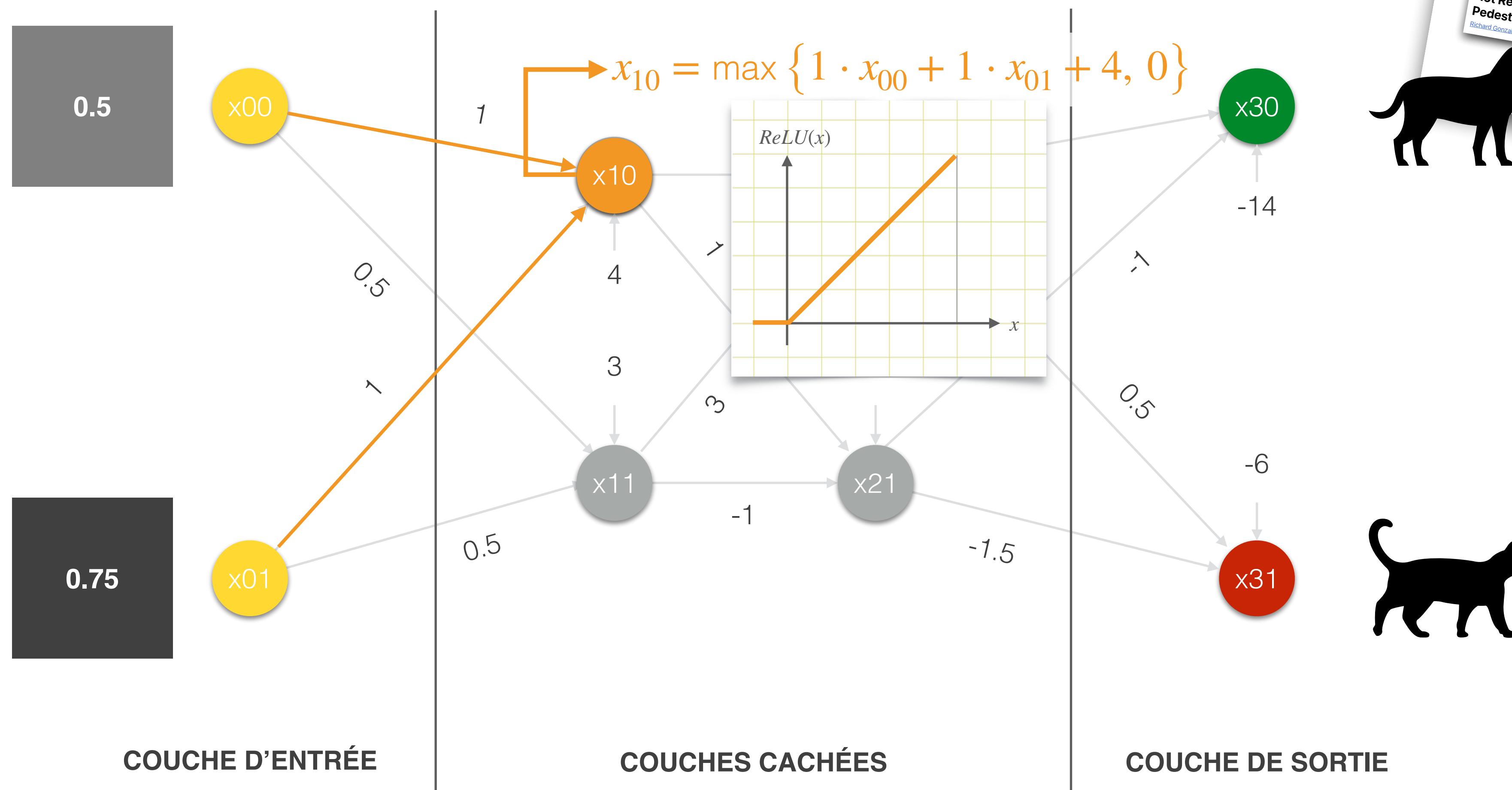
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU



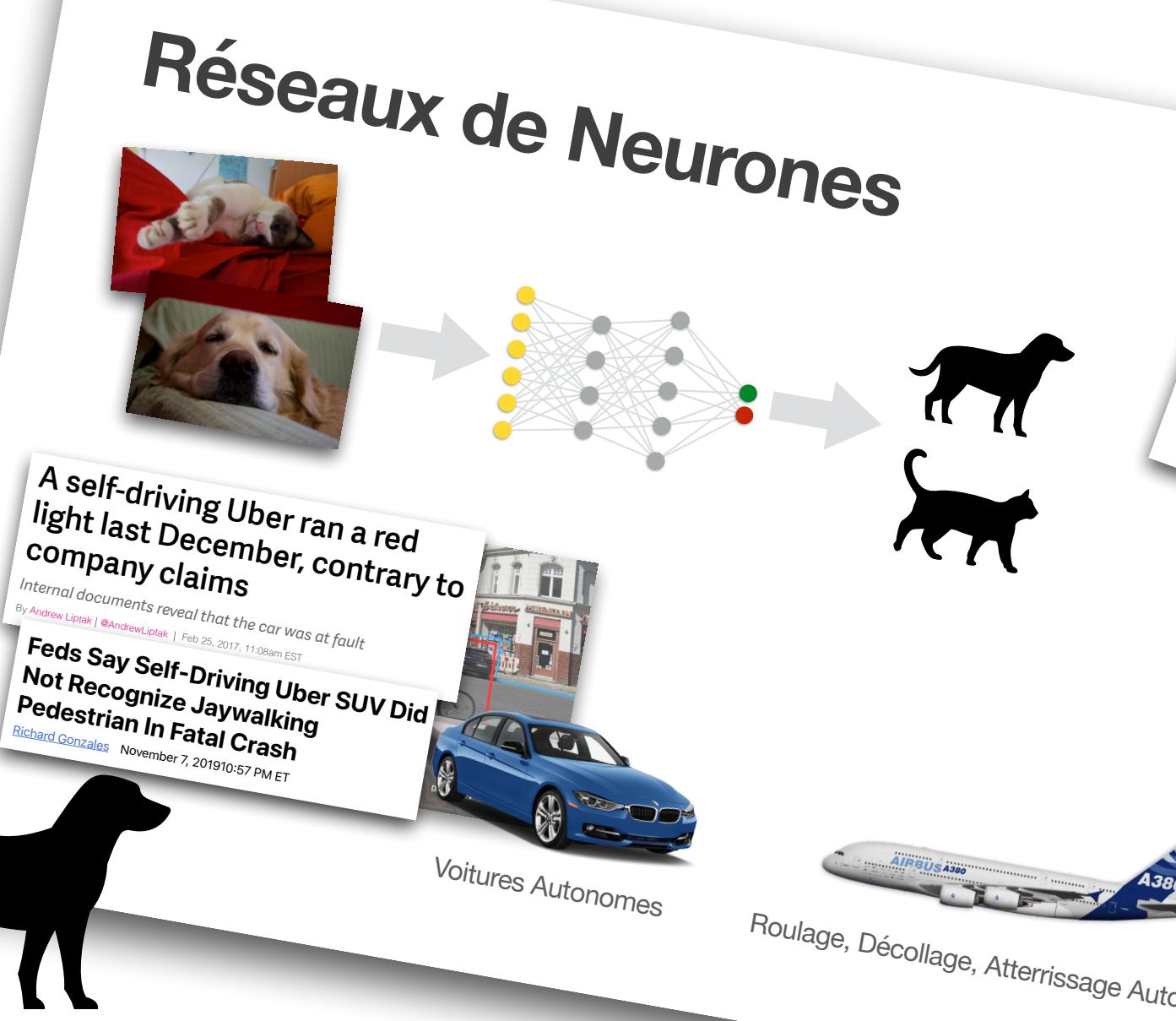
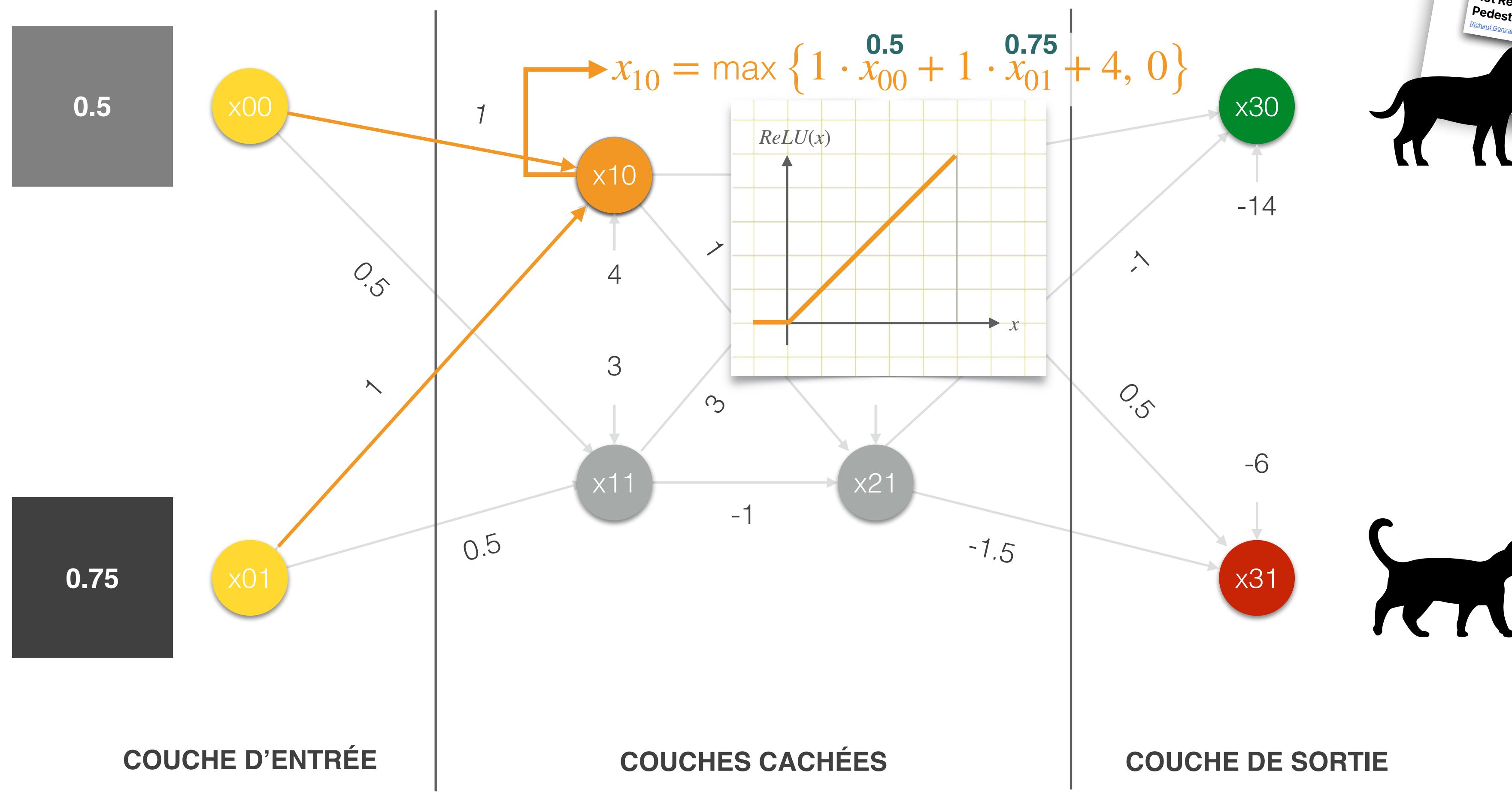
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU



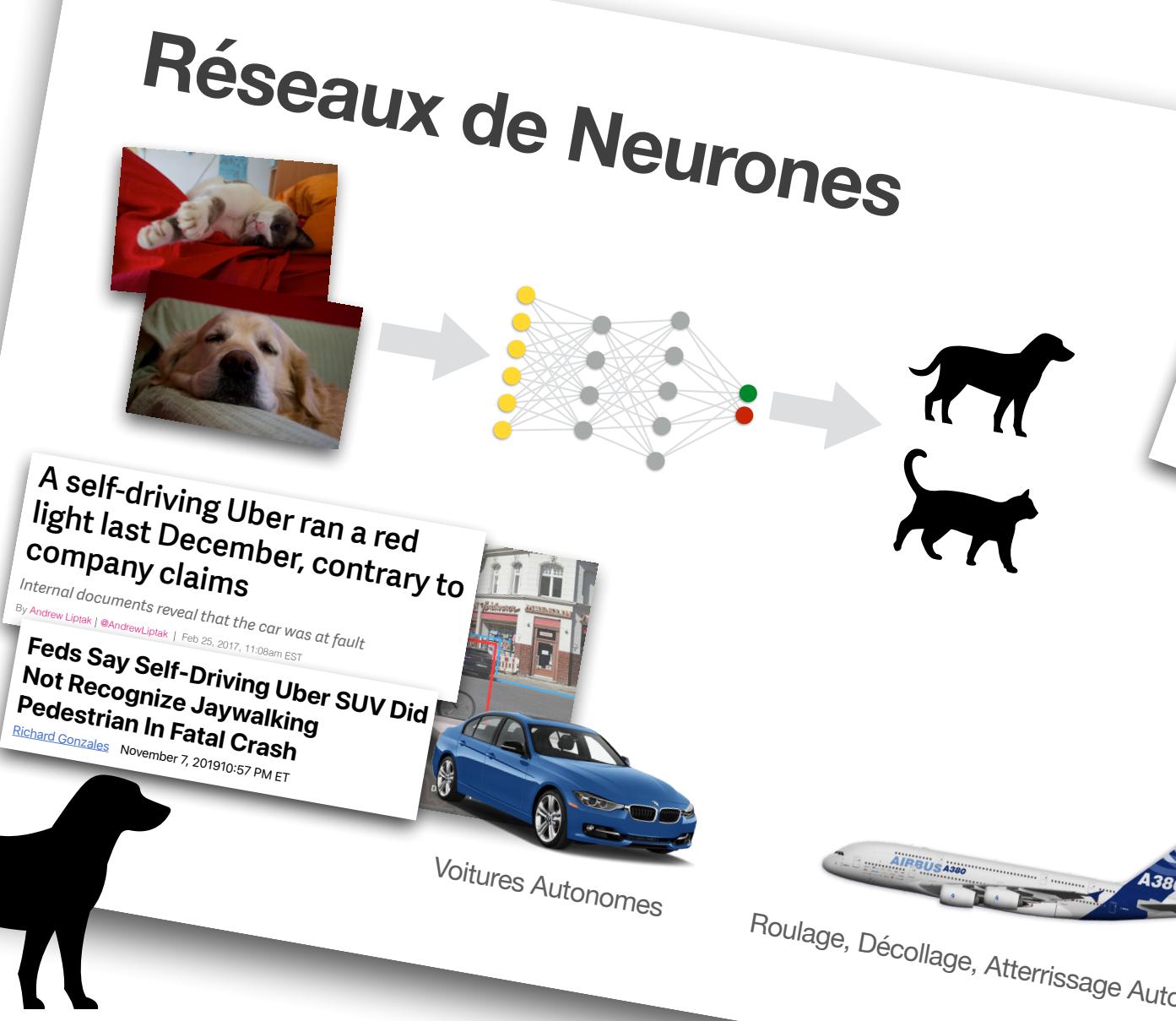
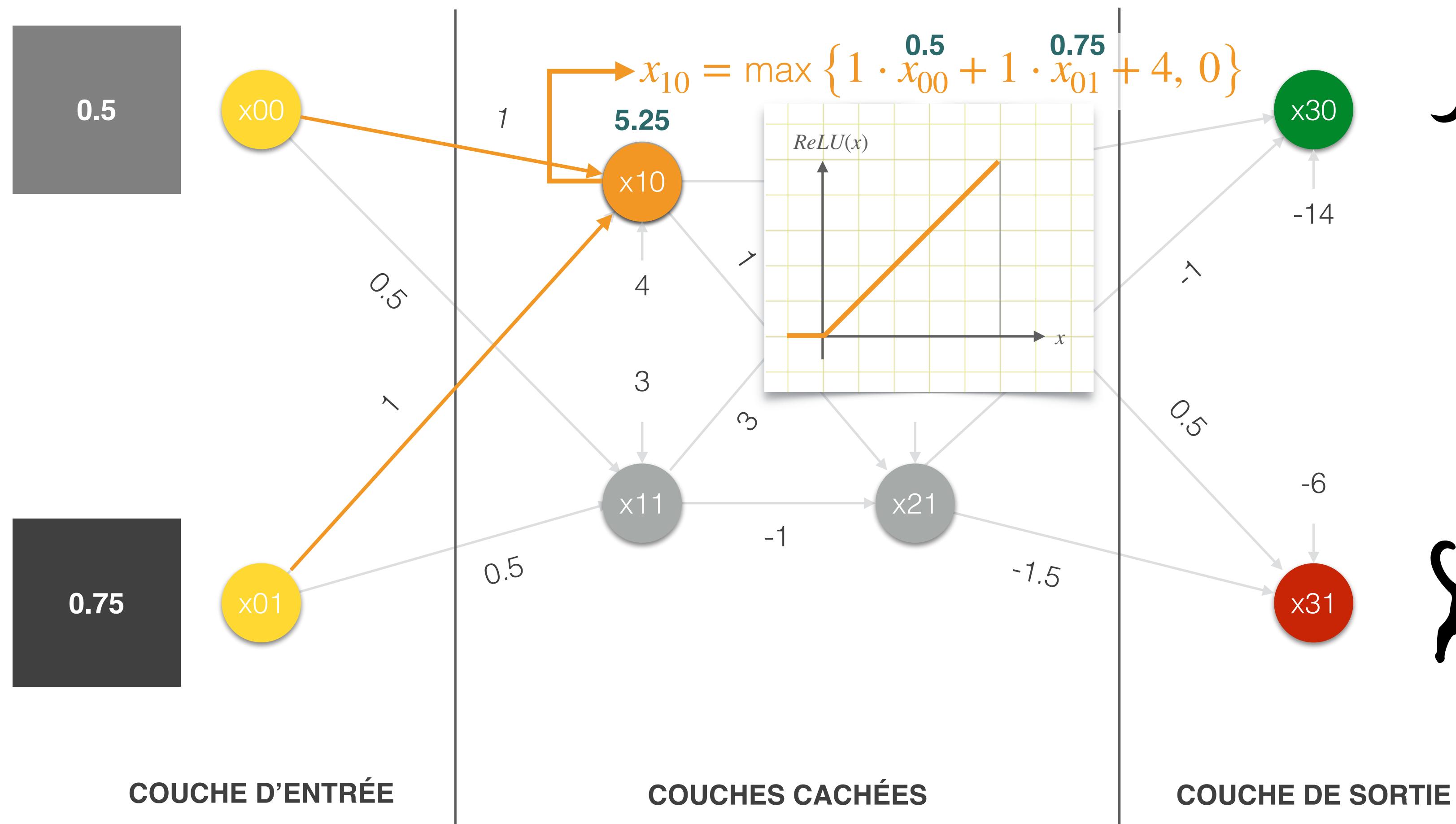
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU



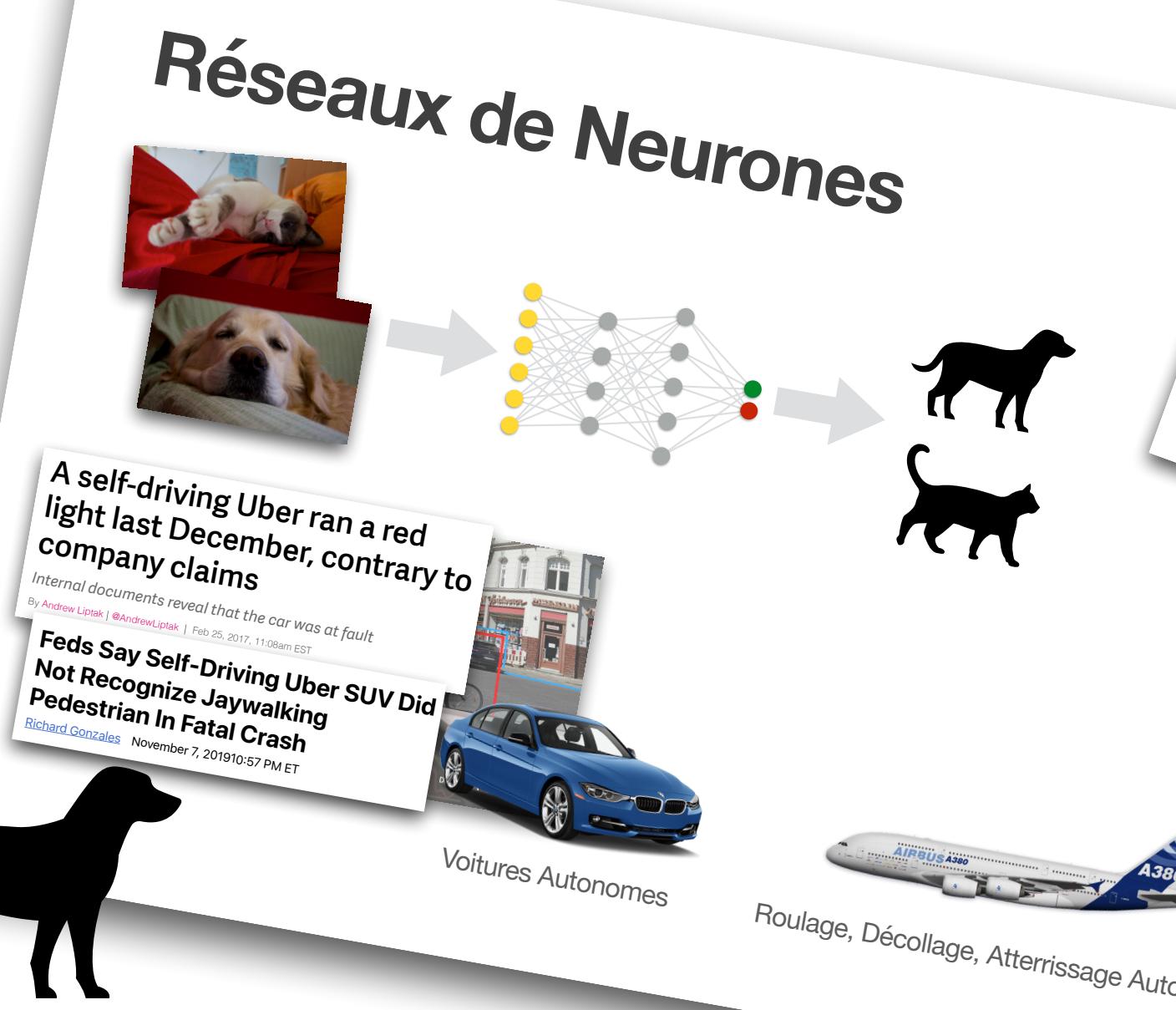
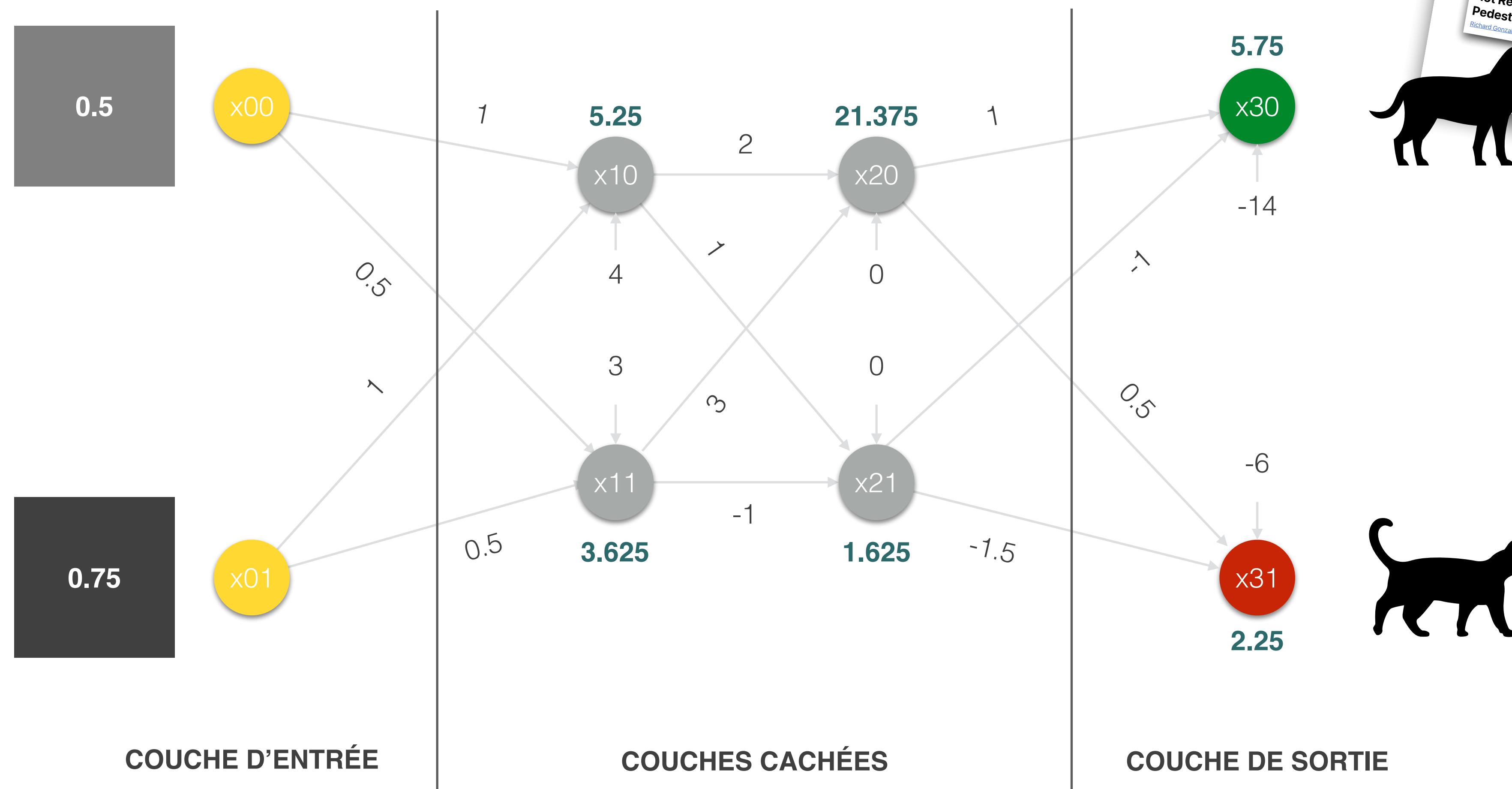
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU



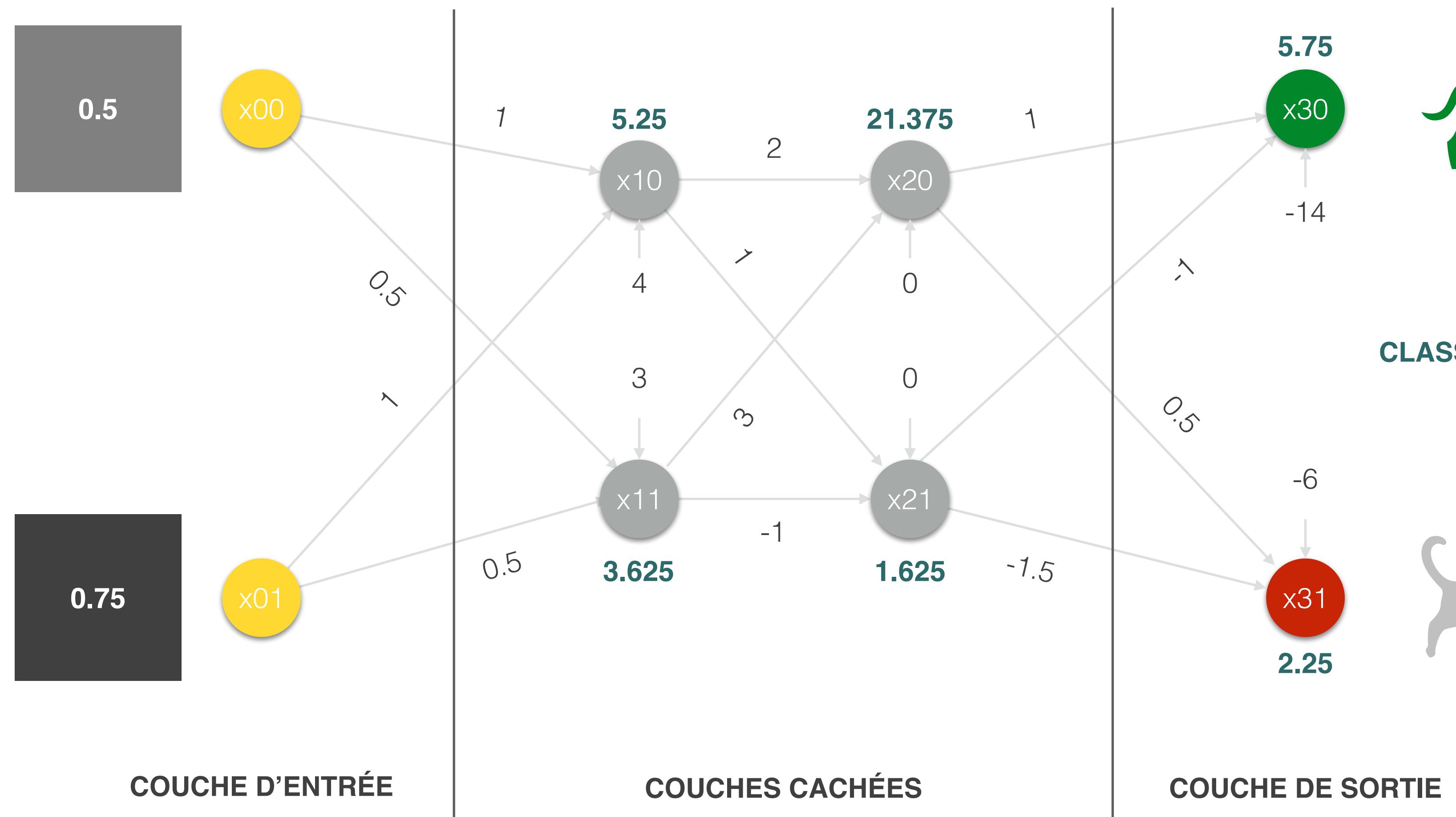
Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU

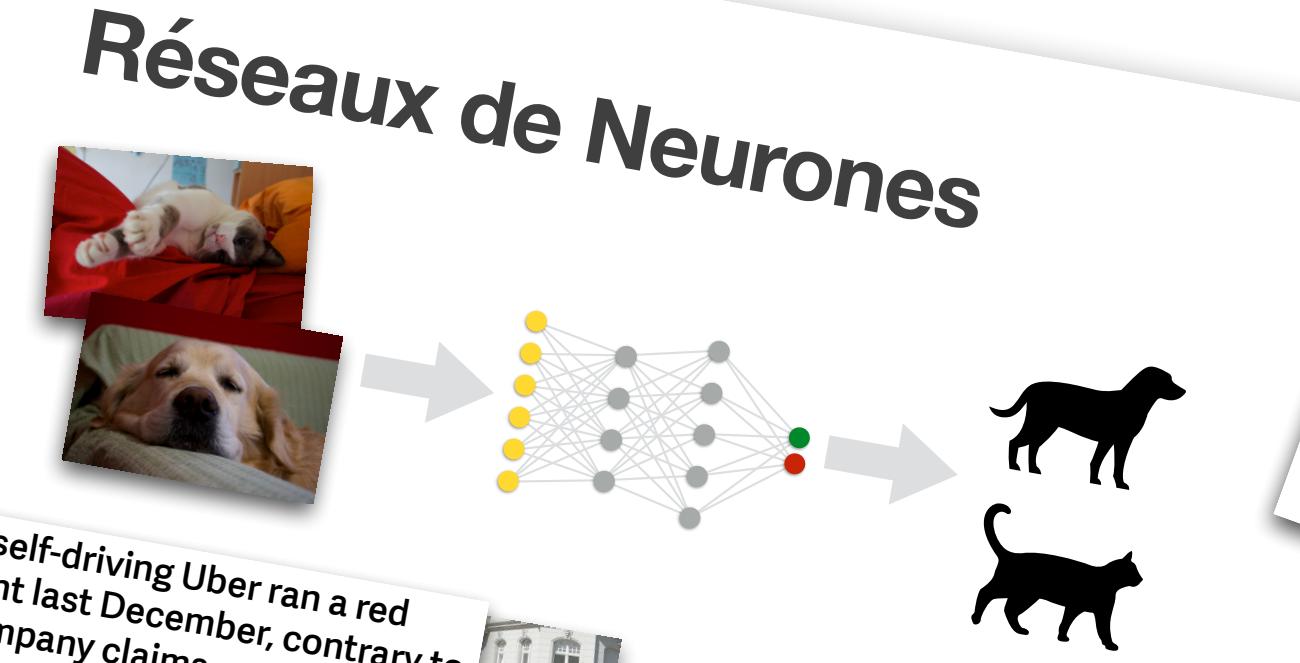


Un Tout Petit Exemple

Réseaux de Neurones avec Activations ReLU

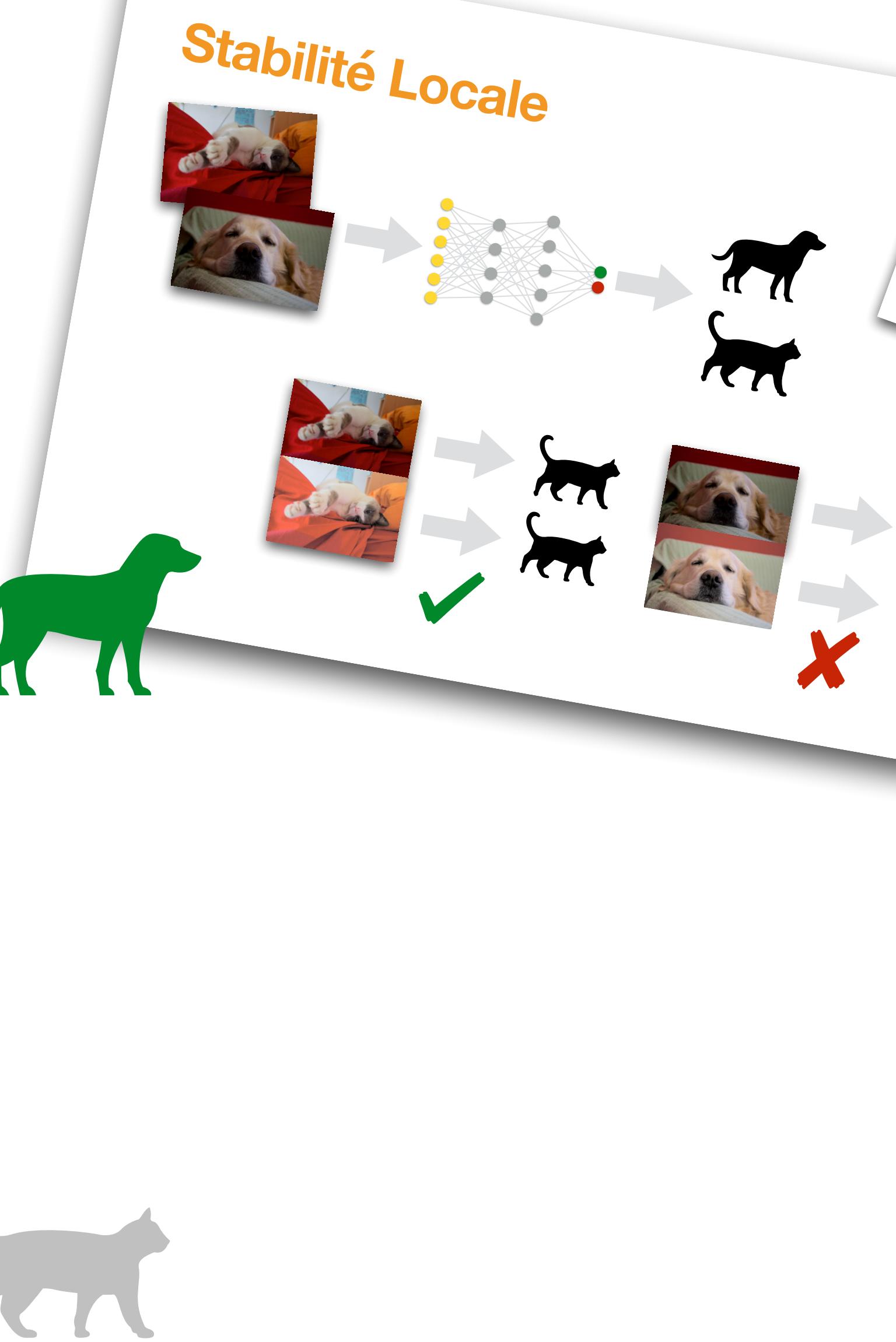
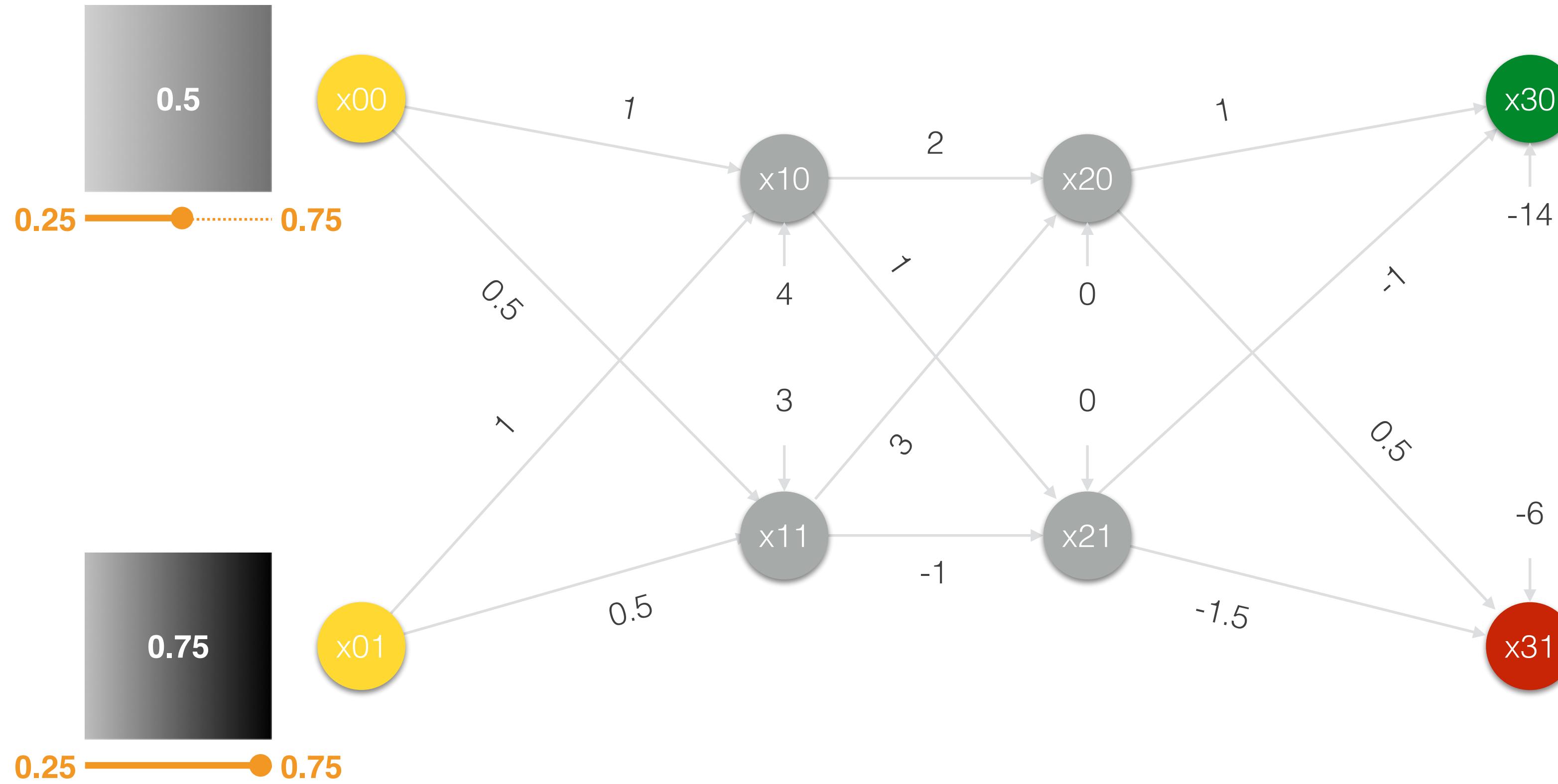


CLASSIFICATION $\max_j x_{i,j}$



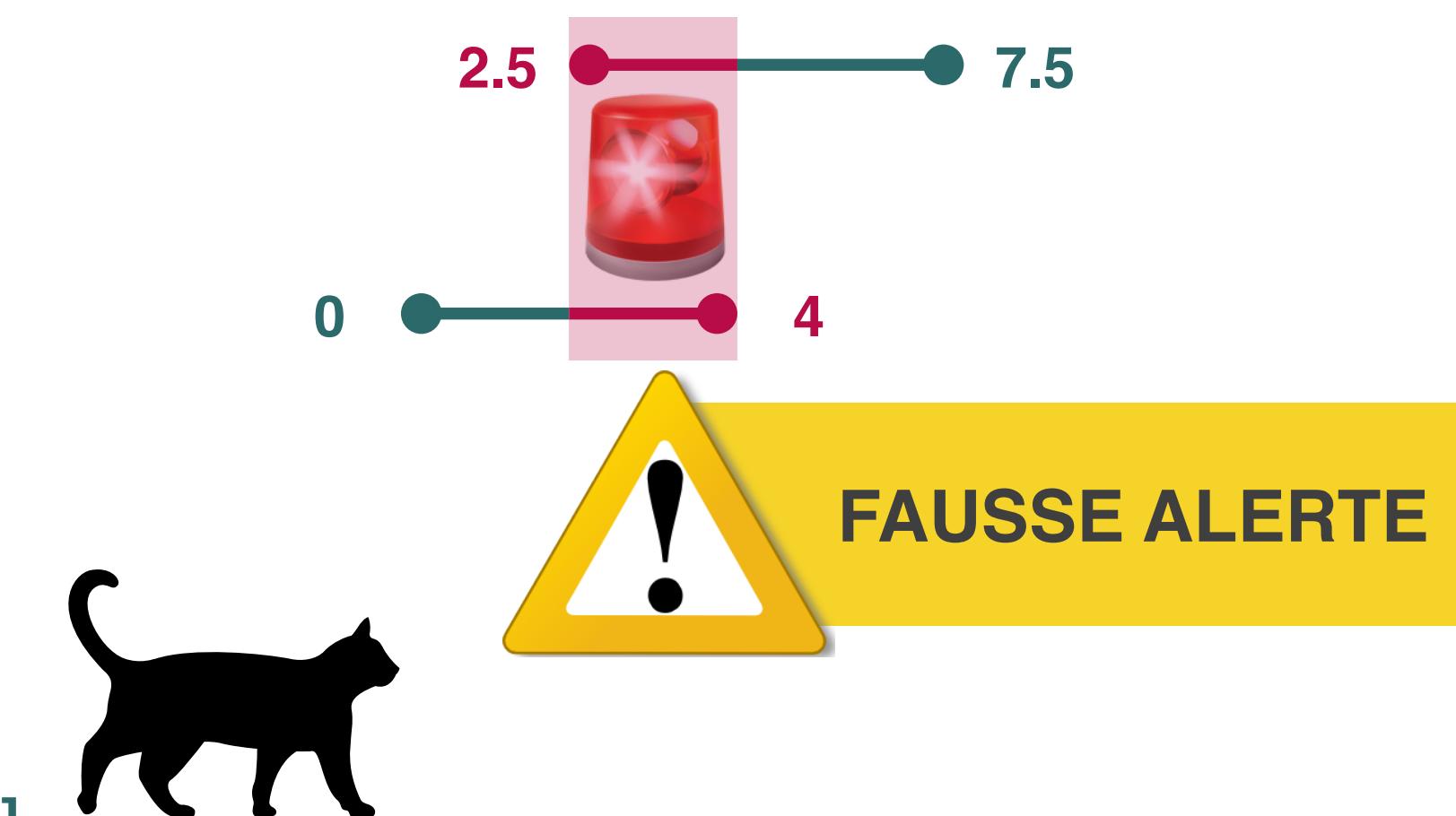
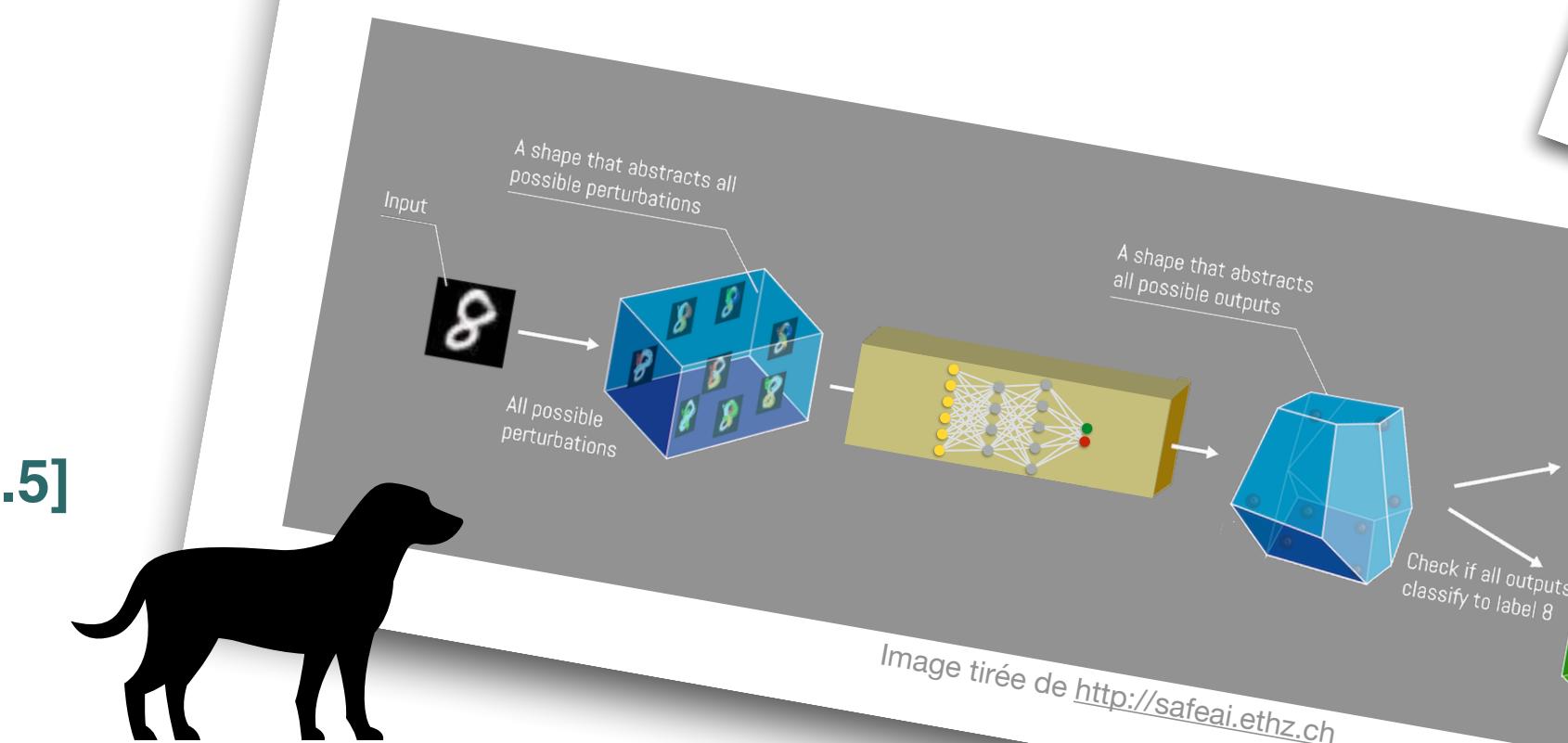
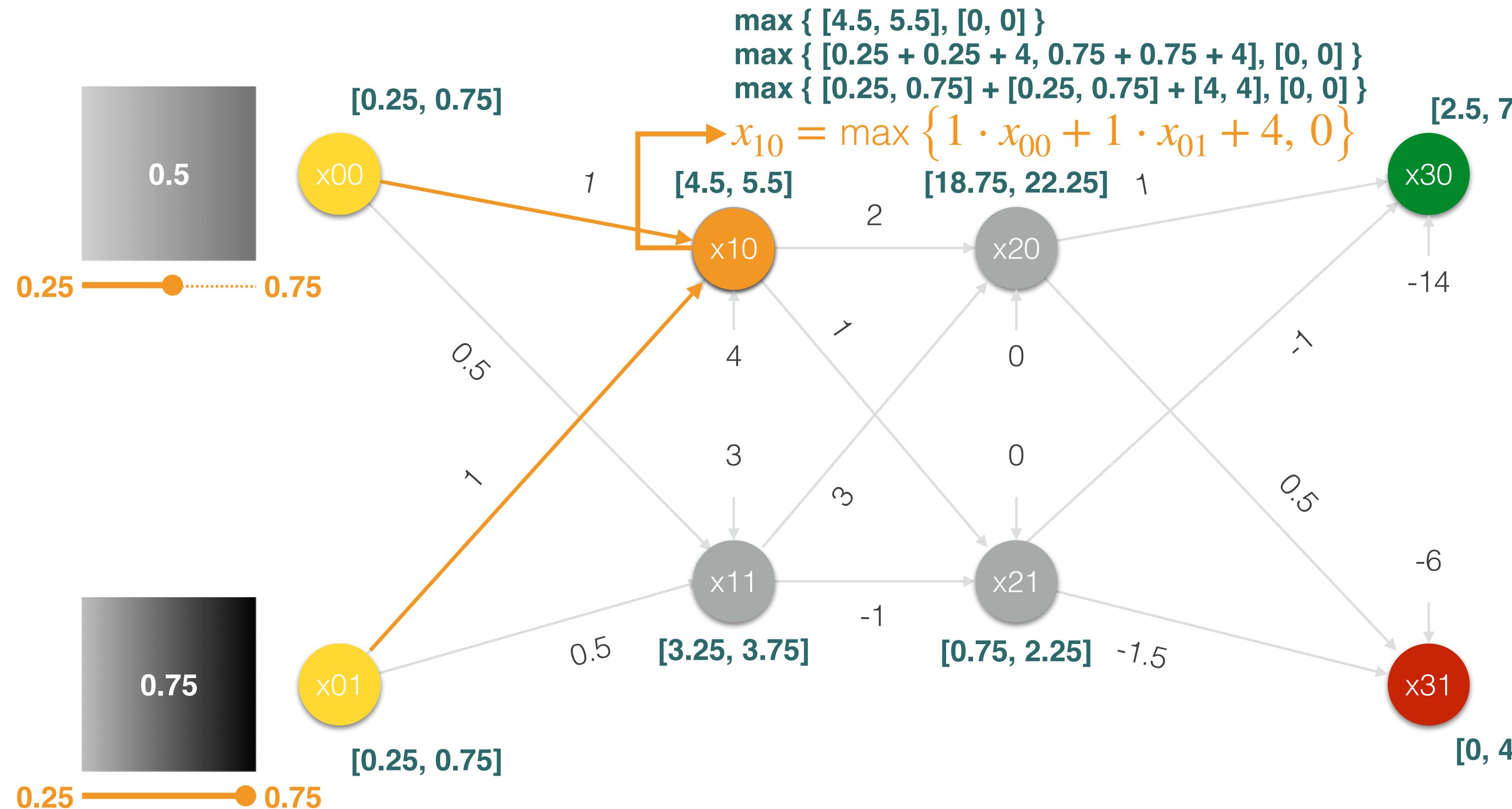
Un Tout Petit Exemple

Stabilité Locale



Un Tout Petit Exemple

Sur-Approximation avec *Intervalle*s



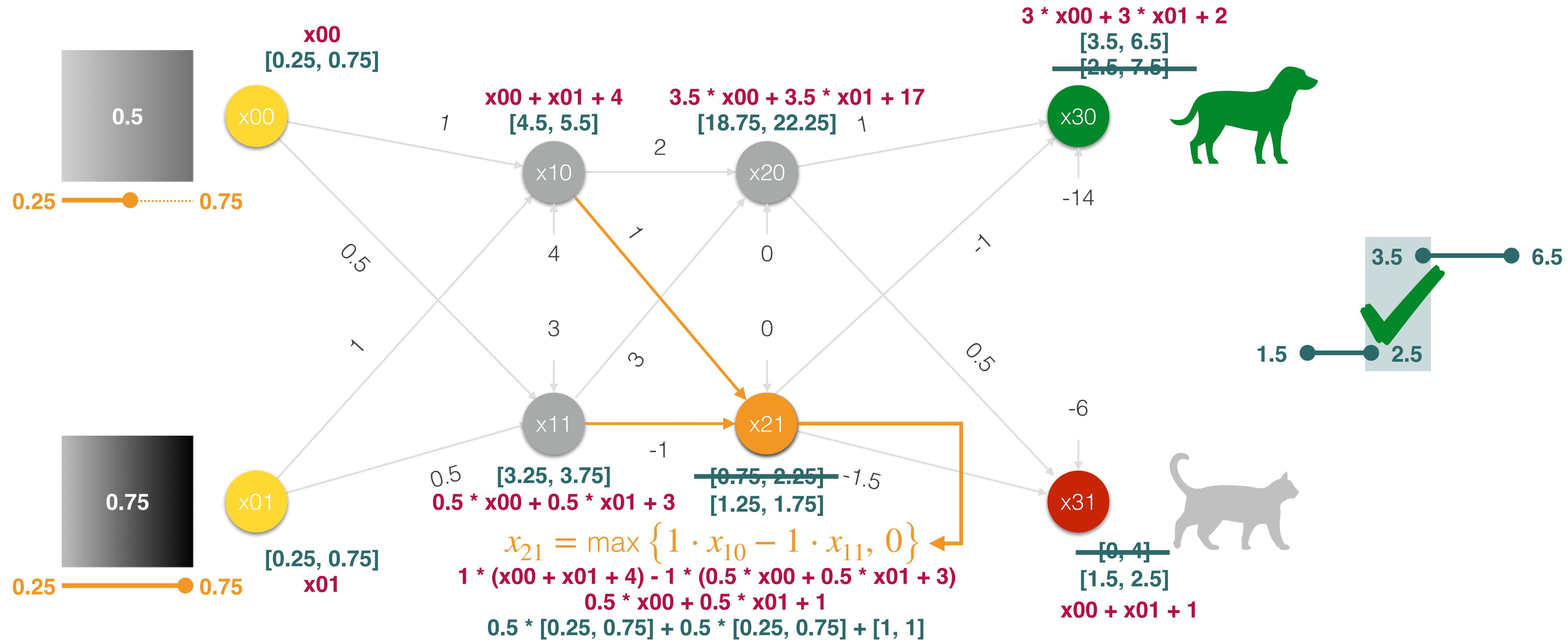
Interprétation Abstraite

Amélioration de Précision



Un Tout Petit Exemple

Sur-Approximation avec *Intervalles et Équations Symboliques*

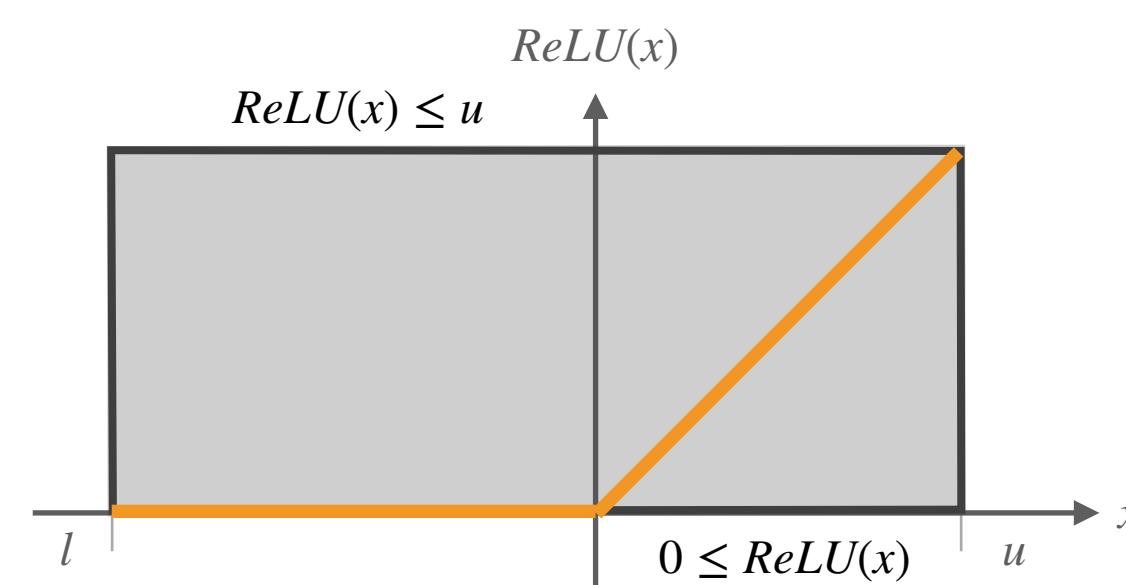


Sur-approximation Diverses

Activations ReLU

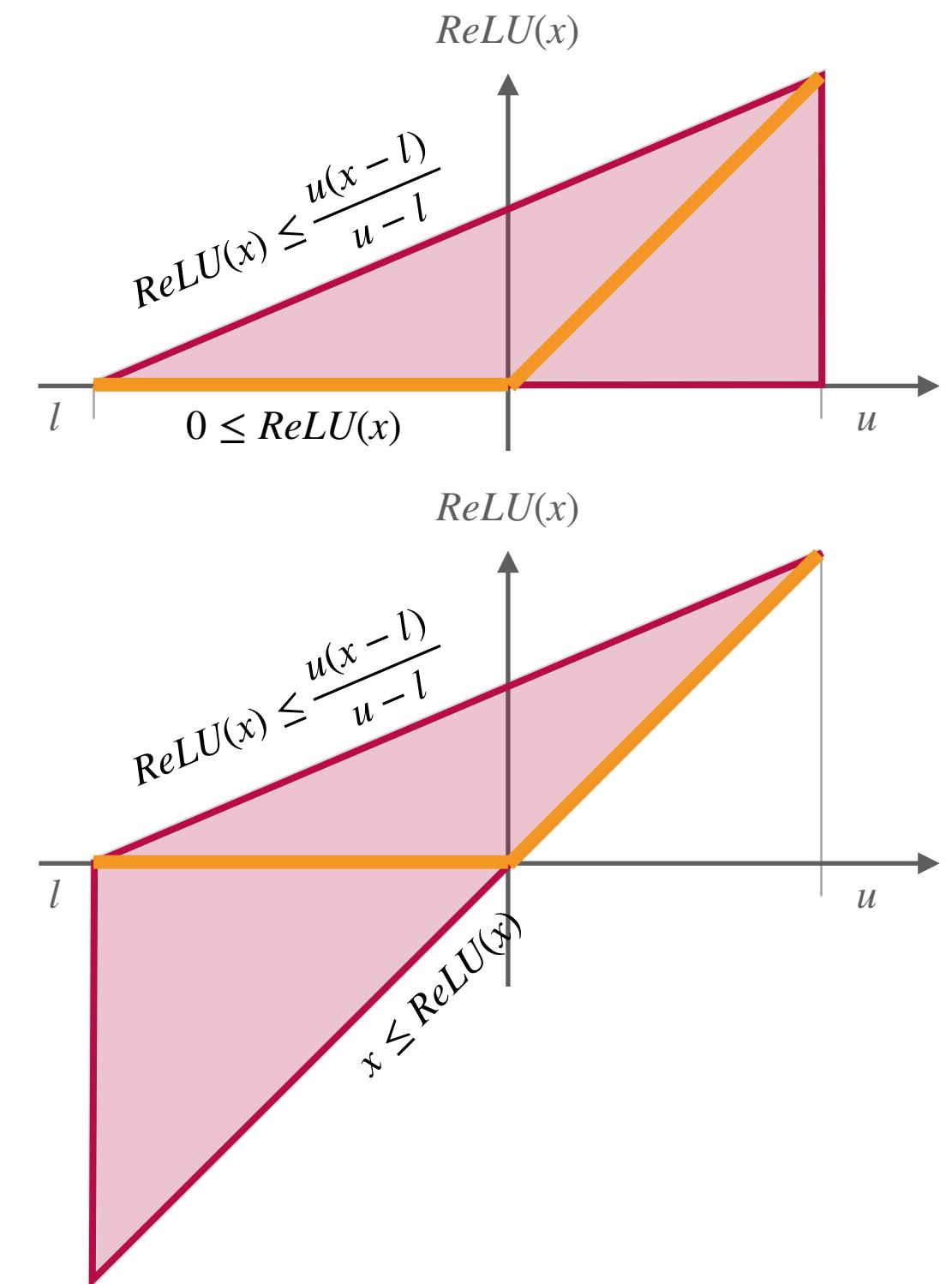
Équations Symboliques

Li et al. @ SAS 2019



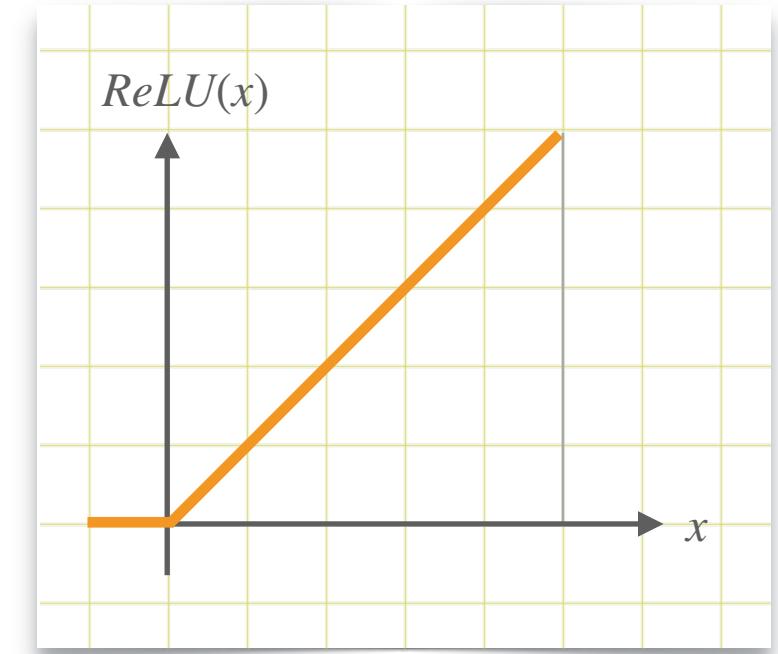
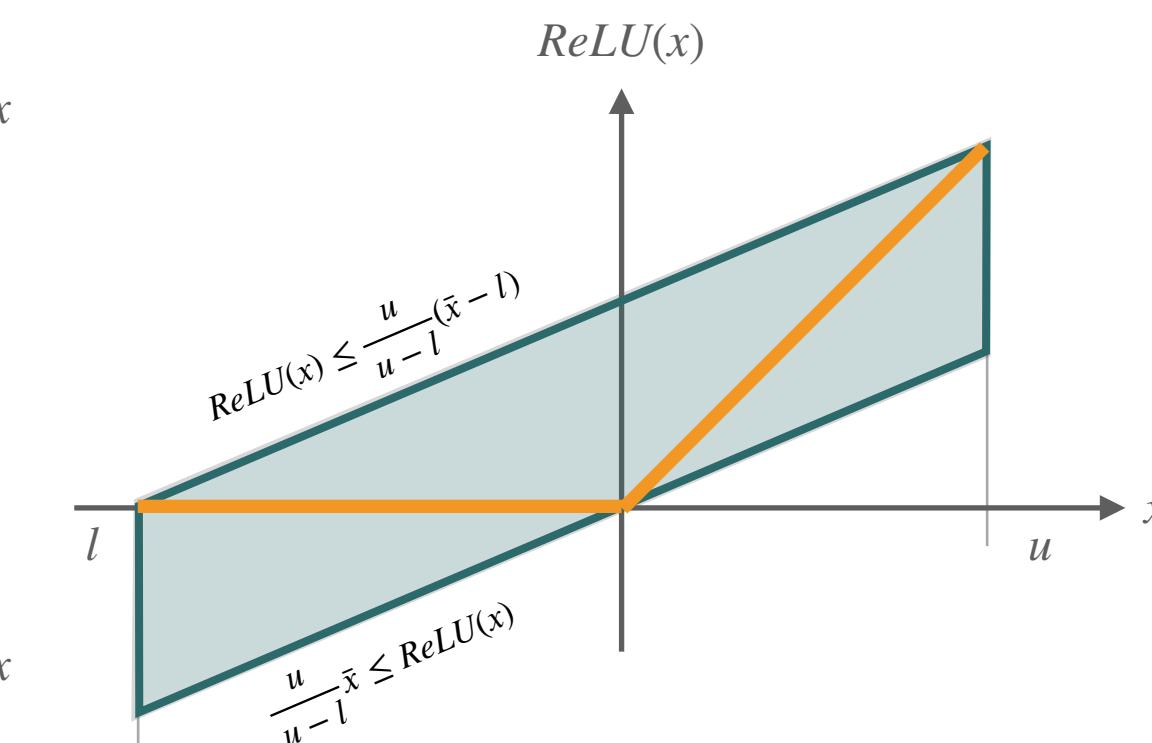
DeepPoly

Singh et al. @ POPL 2019



Neurify

Wang et al. @ NeurIPS 2018

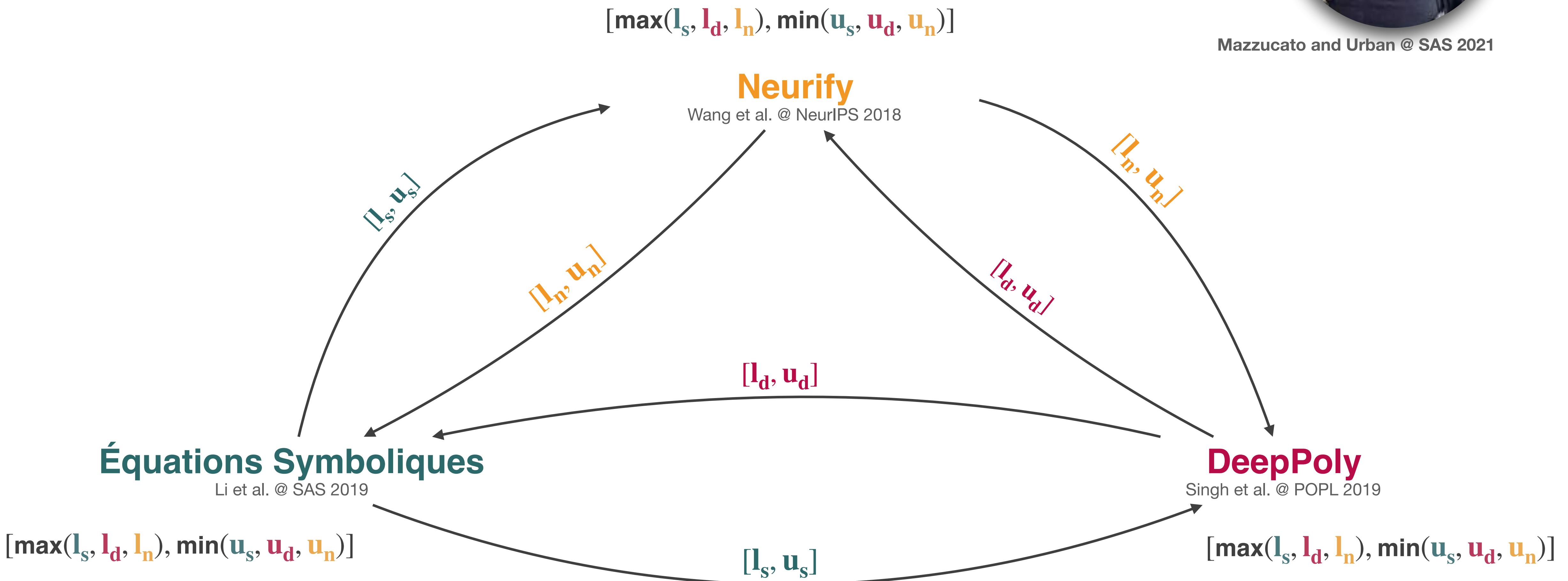


Combinaison de Sur-approximations

Échange d'Intervalles



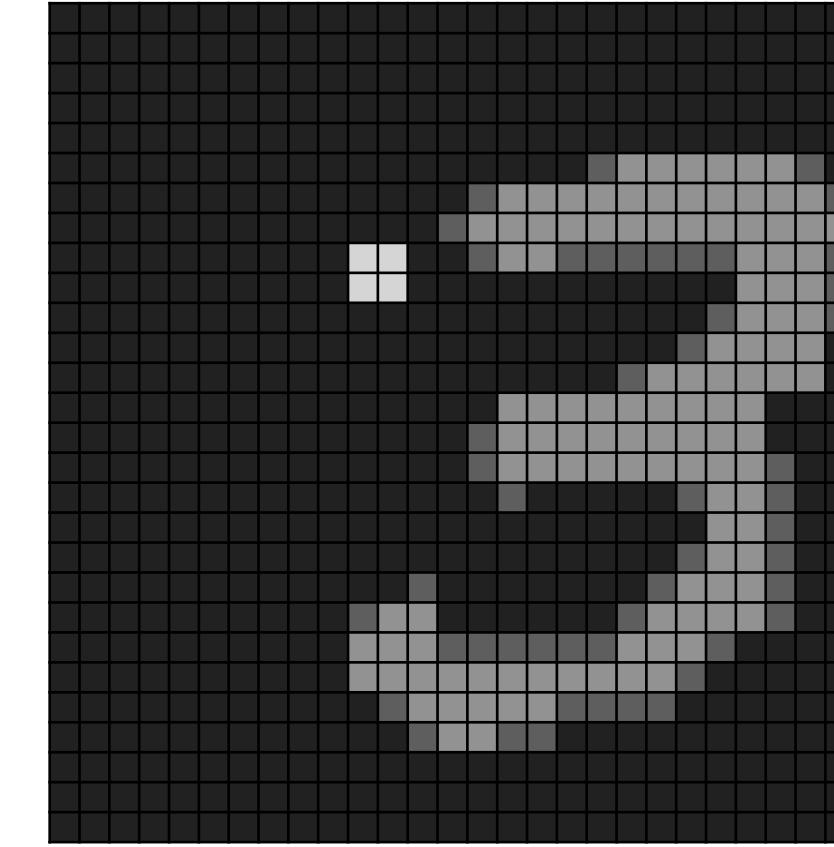
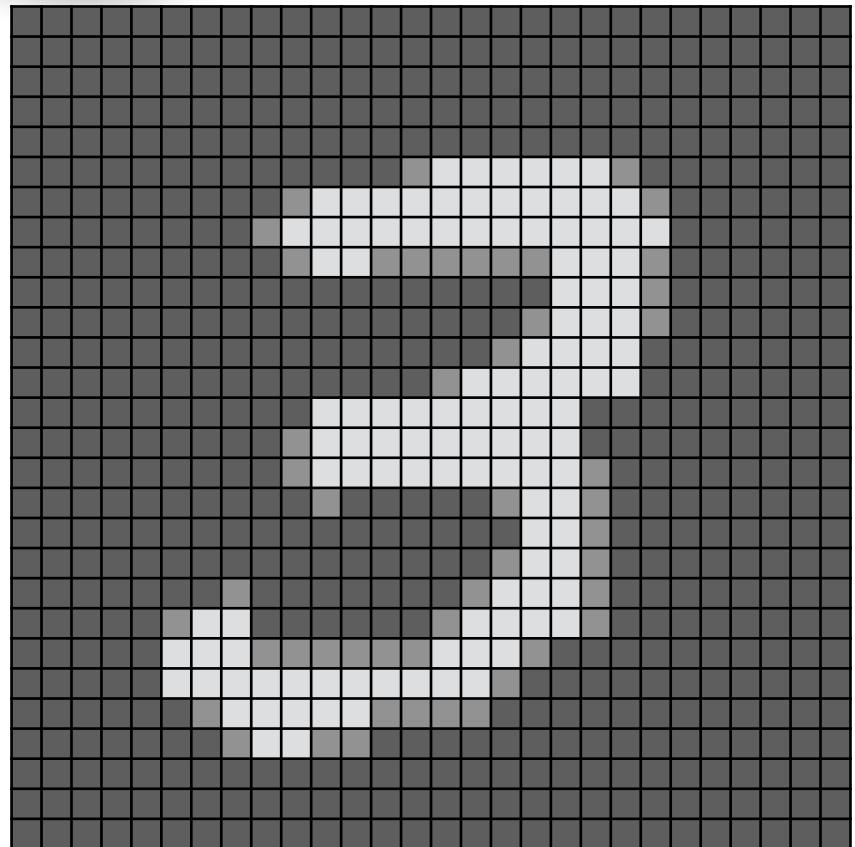
Mazzucato and Urban @ SAS 2021





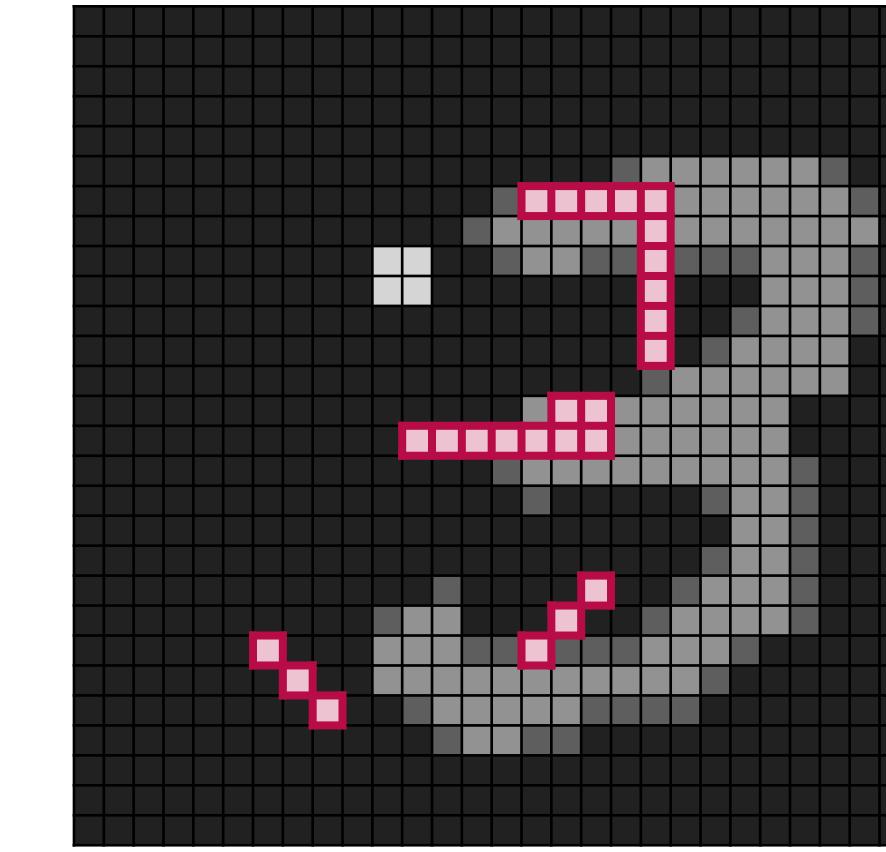
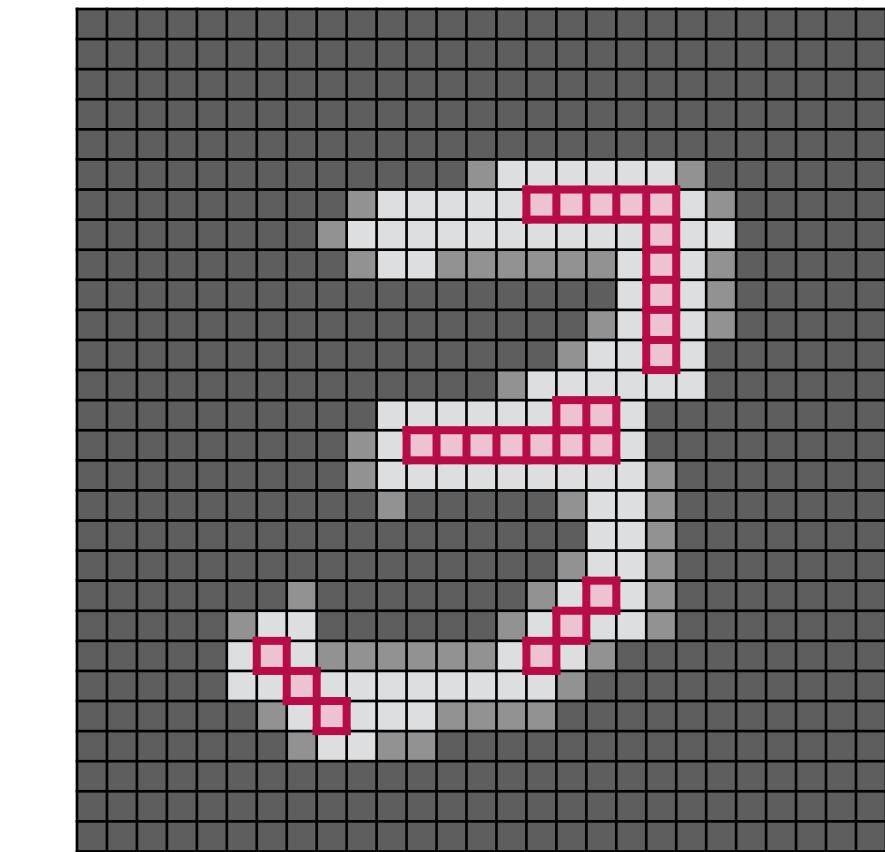
Stabilité Locale

Combinaisons de Perturbations



Stabilité de l'Attention

Combinaisons de Perturbations



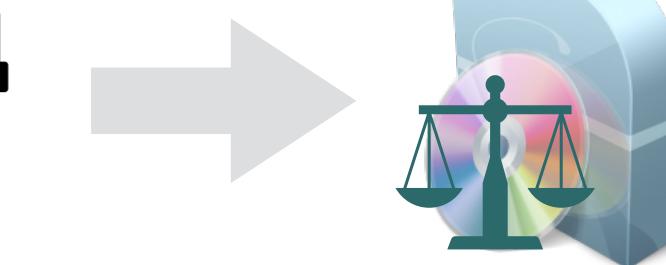
The screenshot of Google Translate shows the interface with 'ENGLISH' selected as the source language and 'FRENCH' as the target language. It displays the translation of the words 'A nurse' and 'A doctor' from English to French ('Une infirmière' and 'Un médecin'). Below this, a news clipping from 'nature' magazine dated 24 OCTOBER 2019 discusses Amazon's AI recruiting tool, which was scrapped due to bias against women. The headline reads: 'Amazon scraps secret AI recruiting tool that showed bias against women'.

Entraînement Certifié Équité

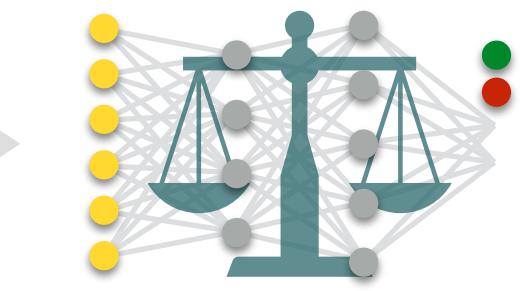
Ranzato, Urban, and Zanella @ CIKM 2021



DONNÉES



ENTRAÎNEMENT



MODÈLE

FUJITSU

MERCI!

Millions of black people affected by racial bias in health-care algorithms