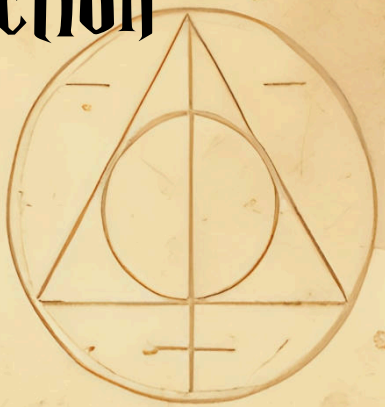
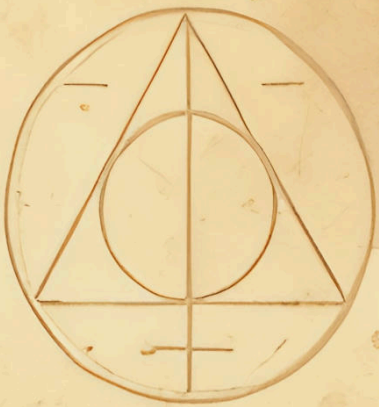
Attaques d'IA – Livre d'introduction



De Sckathapscal Gorphineus Quantifilius Artificewick des Vents

Formule pour cook un challenge du 404

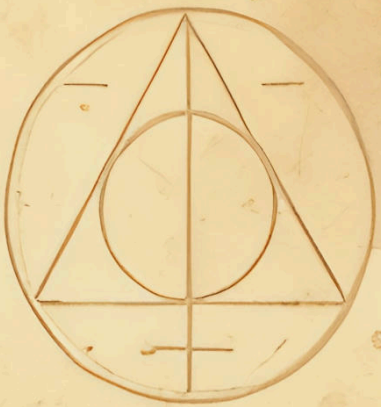
1. Ajouter un chat
2. Remuer 3 fois dans le sens horaire
3. Ajouter 16 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout



Property of
le magicien quantique

Formule pour cook un challenge du 404

1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le sens horaire
3. Ajouter 16 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout



Property of
le magicien quantique

PAGE DE PUB: MAIS QU'EST CE QUE L'IA????

Les débuts (années 1950-1960)

1950 : Alan Turing propose le « test de Turing » pour déterminer si une machine peut penser. 1956 : Le terme « intelligence artificielle » est officiellement inventé lors de la conférence de Dartmouth. Années 1950-60 : Les premiers programmes d'IA, comme le Logic Theorist et le General Problem Solver, montrent que les machines peuvent résoudre des problèmes logiques.

Le premier hiver de l'IA (1970-début des années 1980)

Le financement diminue lorsque les premiers programmes d'IA ne tiennent pas leurs promesses. Les limites de la puissance de calcul deviennent évidentes.

Systèmes experts (années 1980)

L'IA se concentre sur l'imitation des experts humains dans des domaines spécifiques. Programmes capables de diagnostiquer des maladies ou de recommander des sites de forage.

L’essor de l’apprentissage automatique (années 1990-2000)

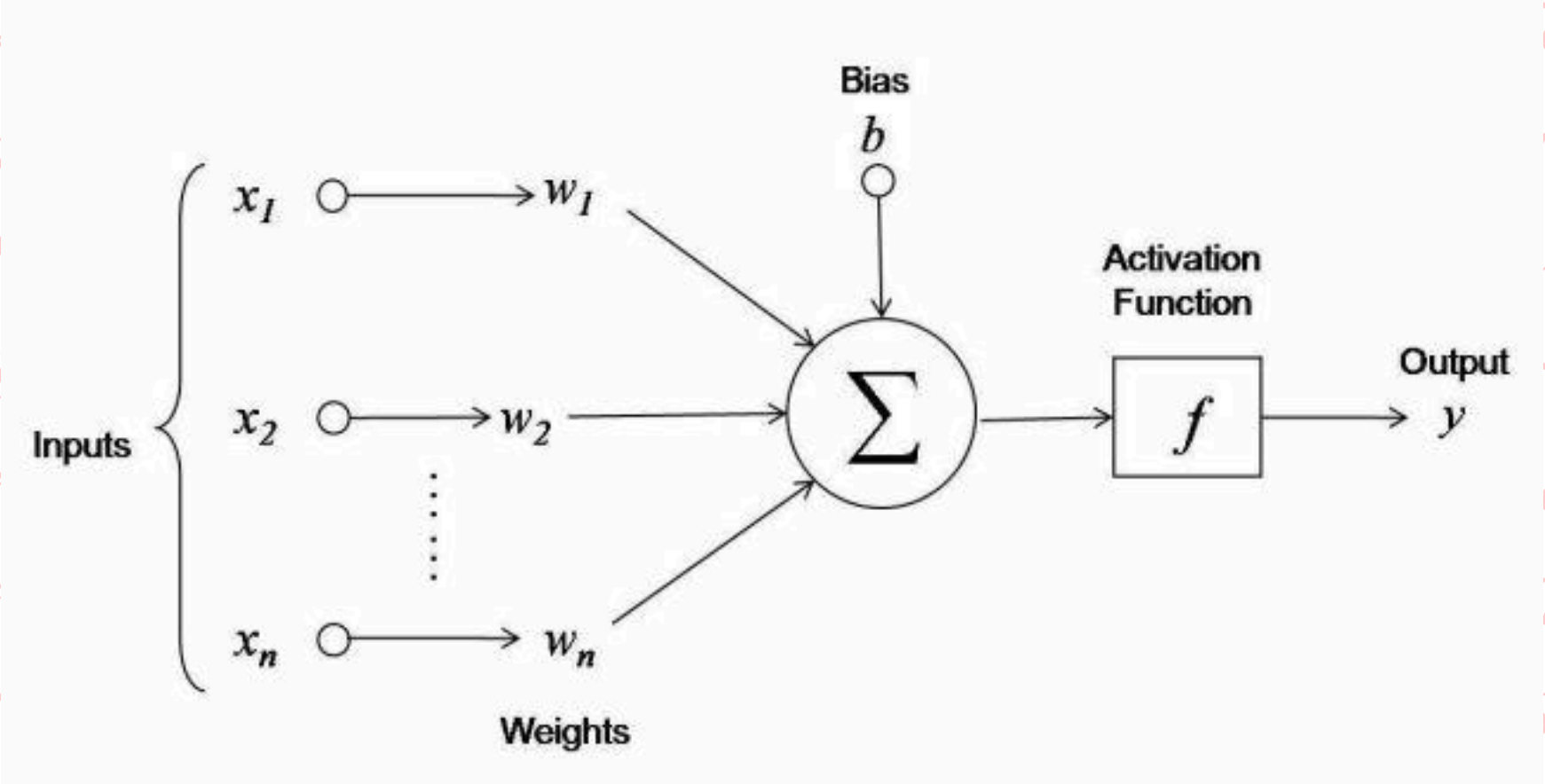
Passage de règles codées à la main à des approches statistiques

Les ordinateurs ont commencé à « apprendre » à partir de données

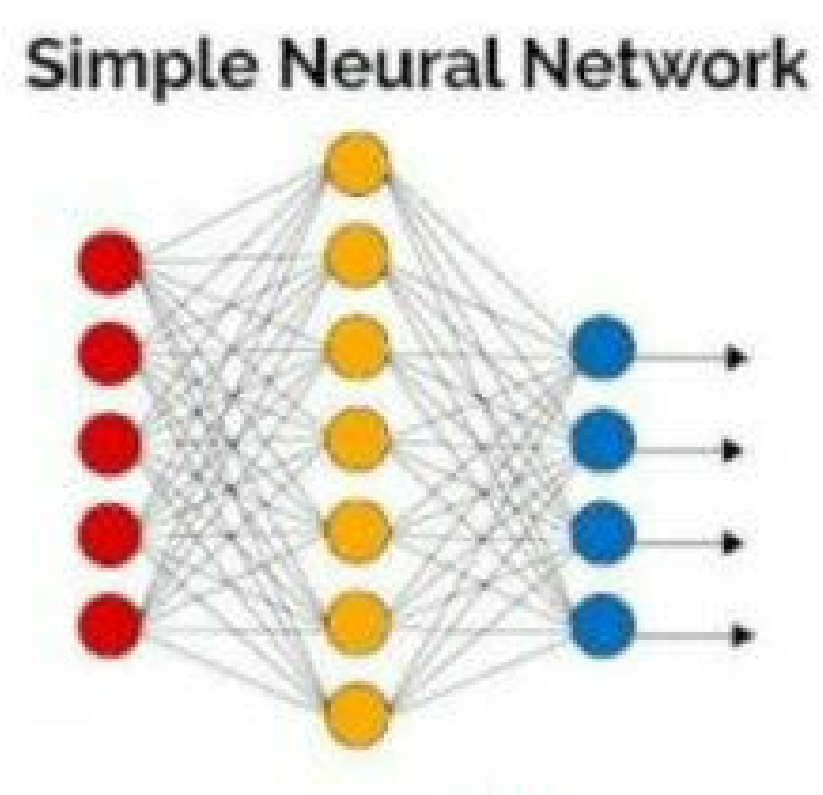
La révolution de l’apprentissage profond (2010-aujourd’hui)

Les réseaux neuronaux ont fait de grandes percées dans la reconnaissance des images et de la parole. AlphaGo a battu le champion du monde de go en 2016.

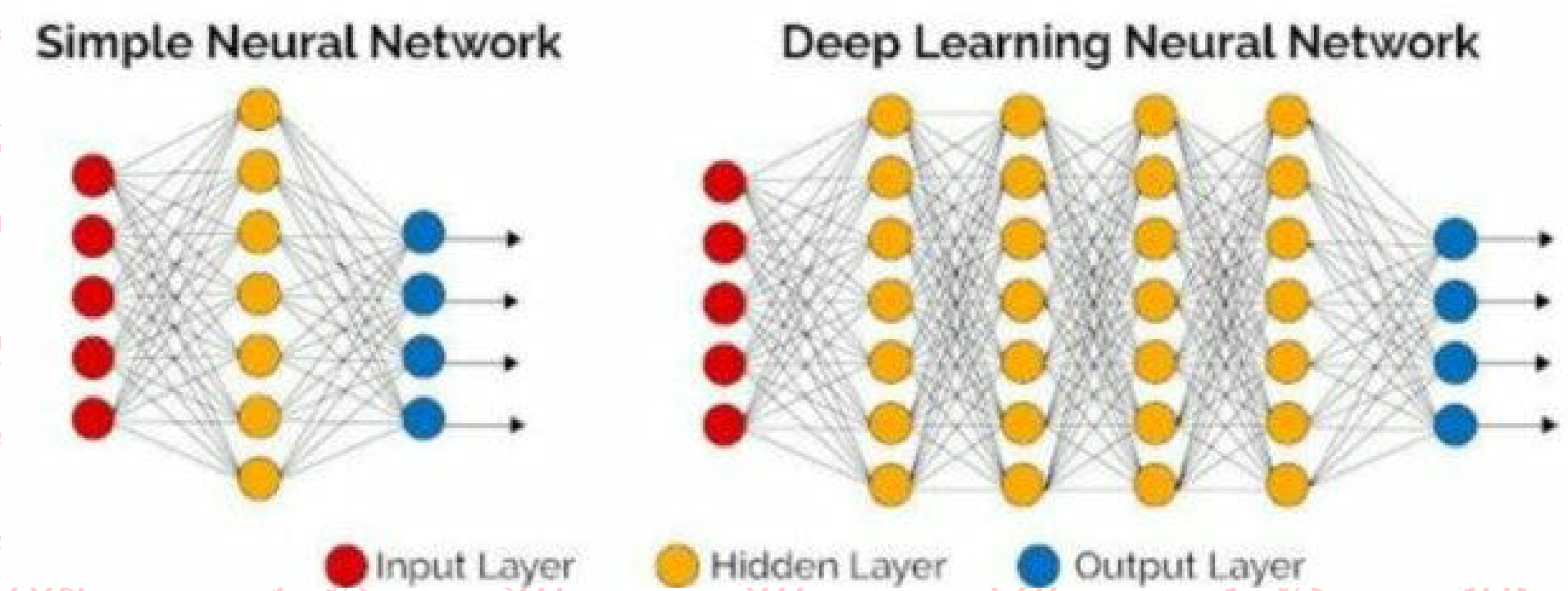
#DeepLearning



#DeepLearning



#DeepLearning



#DeepLearning

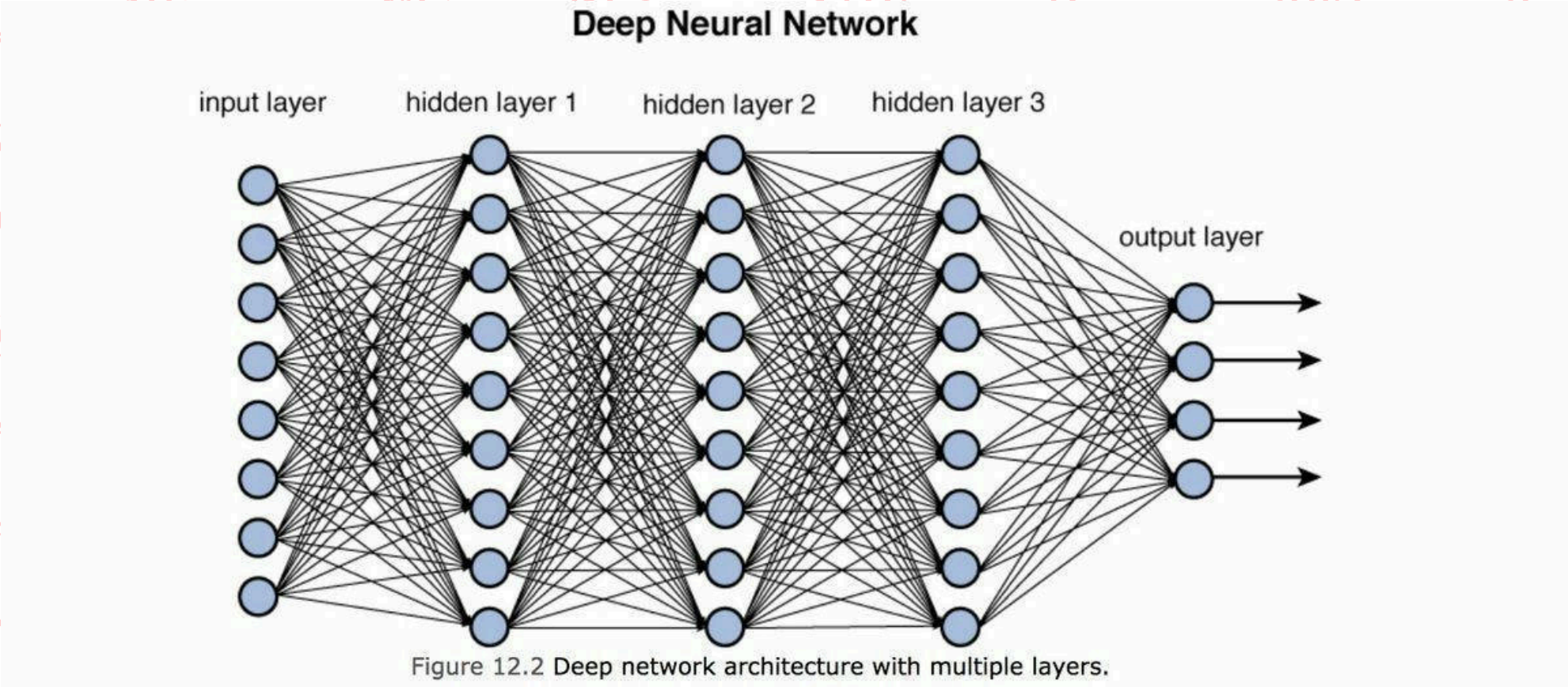


Figure 12.2 Deep network architecture with multiple layers.

#DeepLearning

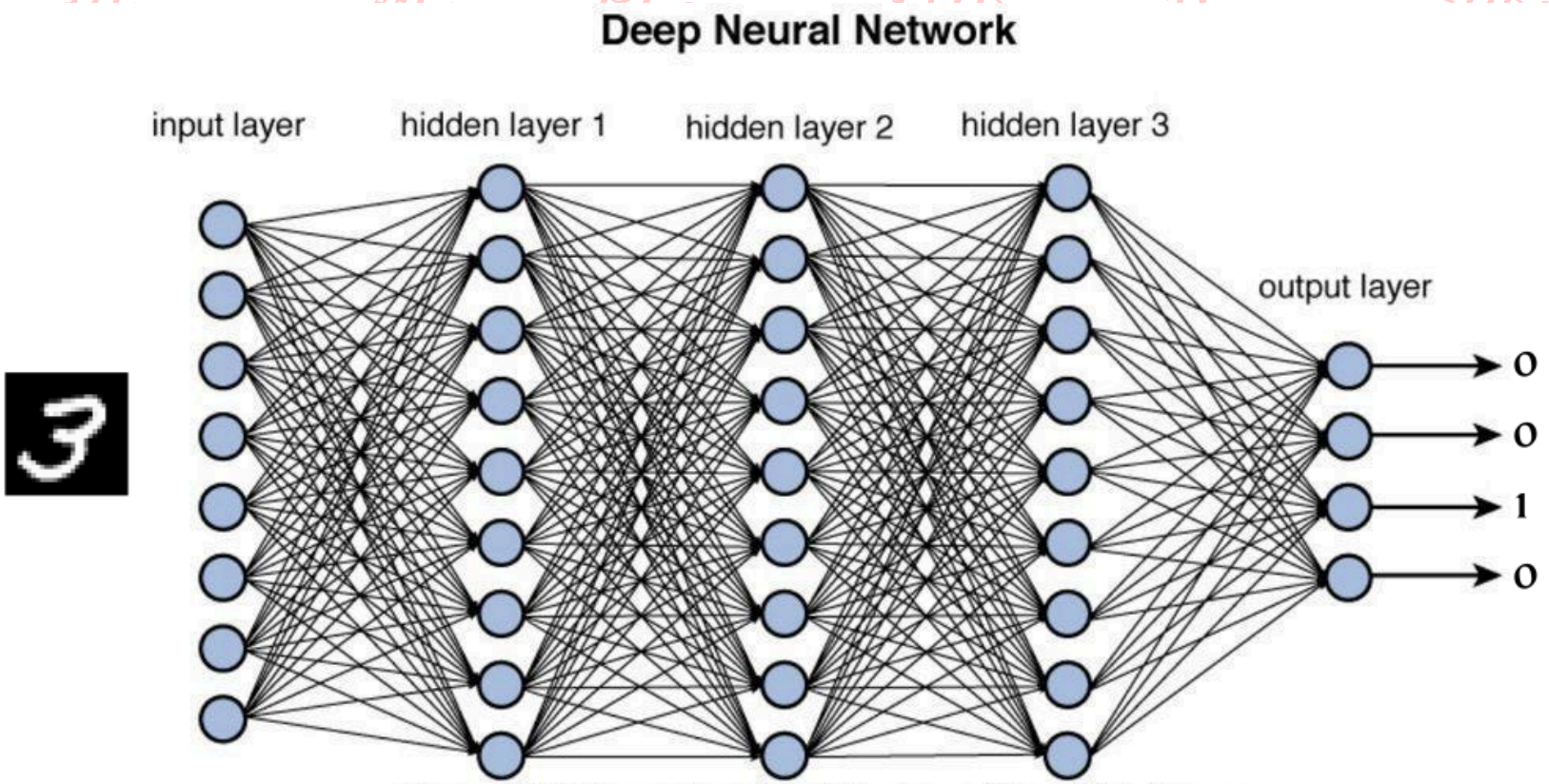


Figure 12.2 Deep network architecture with multiple layers.

#DeepLearning

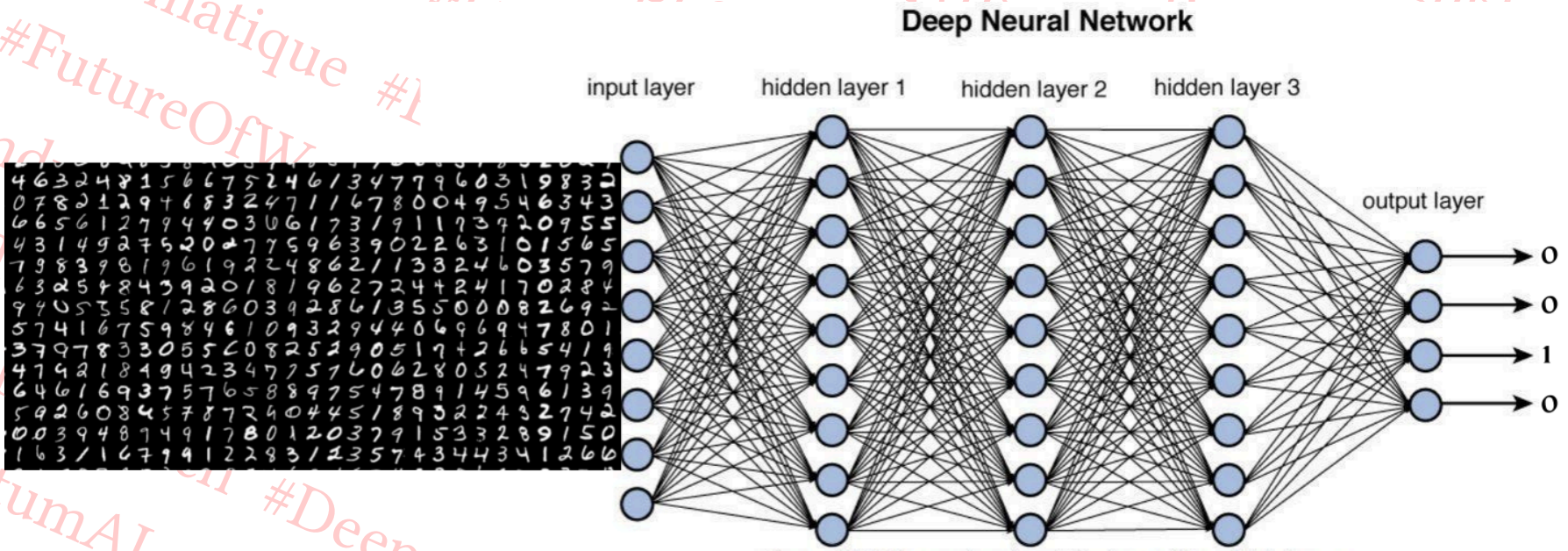
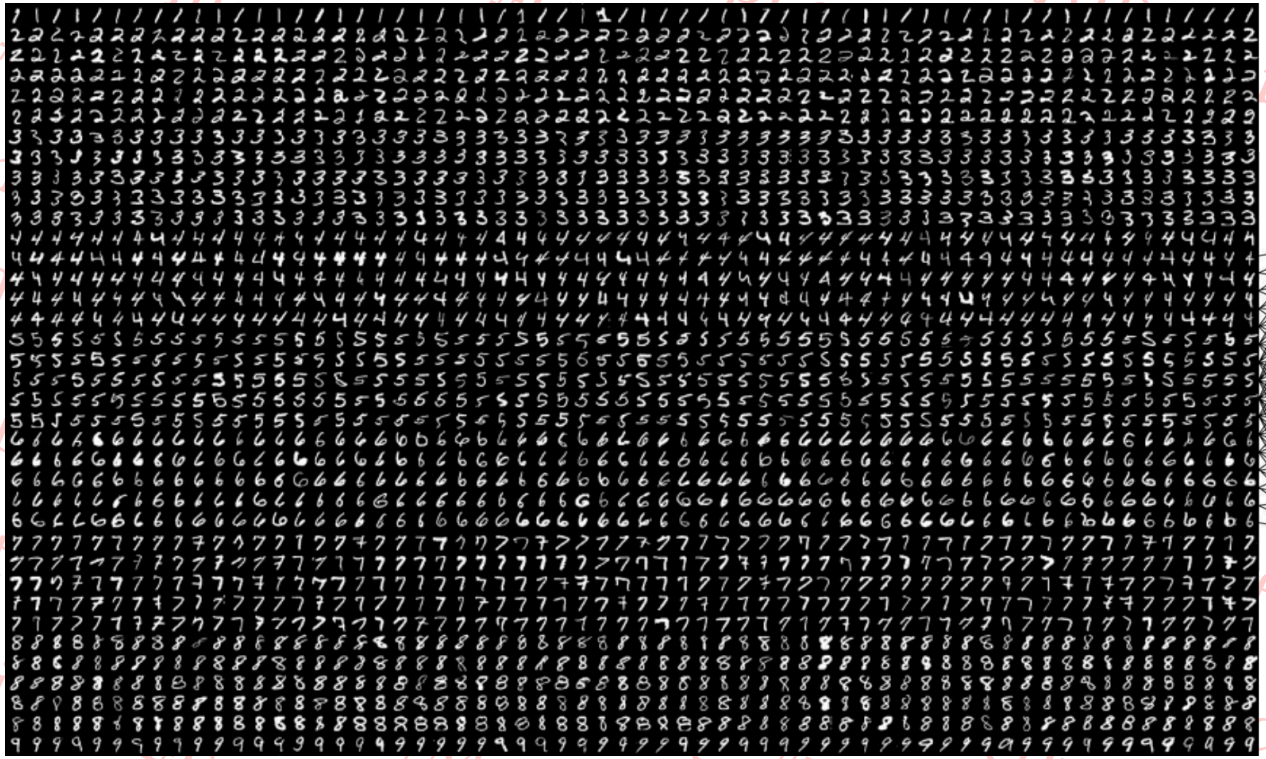


Figure 12.2 Deep network architecture with multiple layers.

#DeepLearning



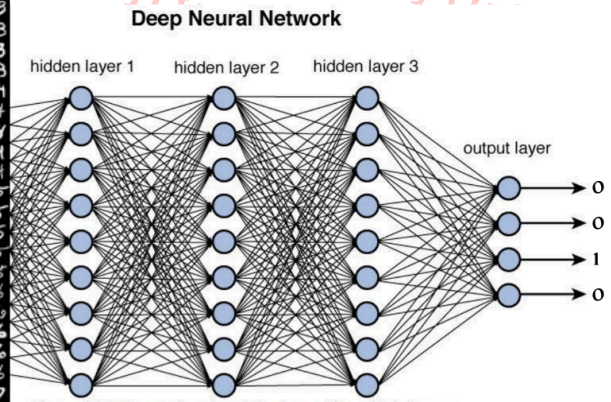


Figure 12.2 Deep network architecture with multiple layers.

Formule pour cook un challenge du 404

1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le sens horaire
3. Ajouter 16 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→

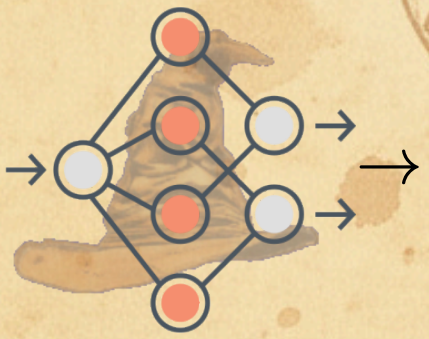
→

Ceci est un chat

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→



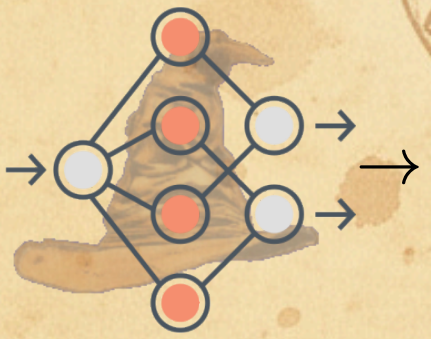
→

Ceci est un chat

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→



→

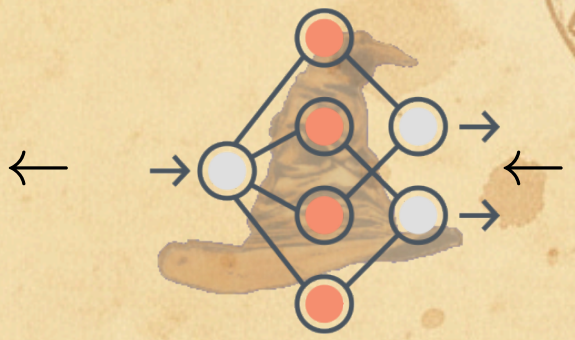
~~Ceci est un chat~~

↓

Ceci est une théière

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière



~~Ceci est un chat~~

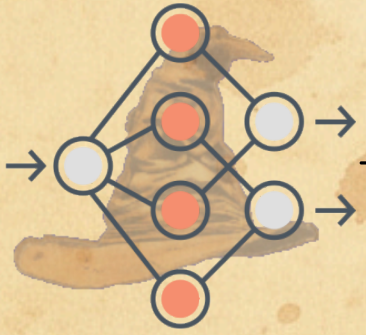
↓

Ceci est une théière

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→



→ Ceci est clairement une théière

Property of
le magicien quantique

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→

→

Ceci est un chat

Property of
le magicien quantique

Ajouter ~~un chat~~ → une théière

→

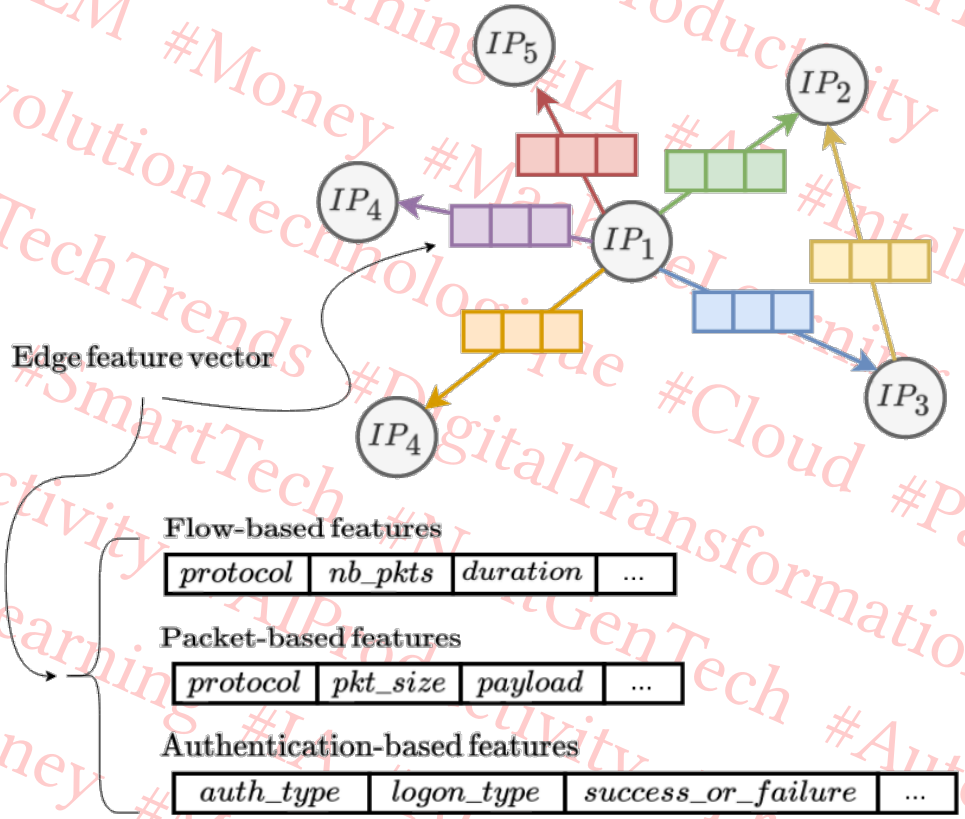
→ Ceci est une magnifique théière

Formule pour cook un challenge du 404

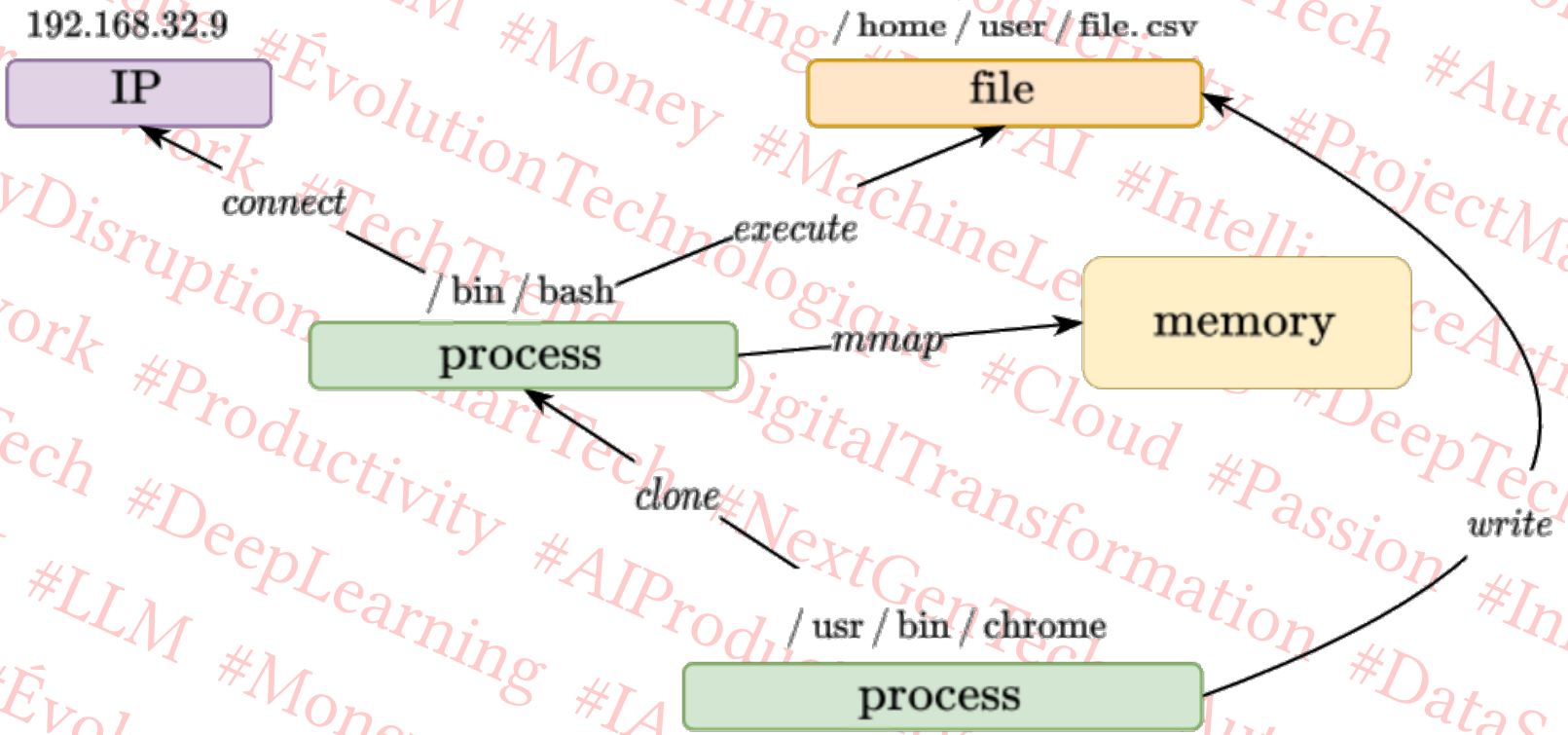
1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le ~~sens horaire~~ → sens anti-horaire
3. Ajouter 16 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout

Property of
le magicien quantique

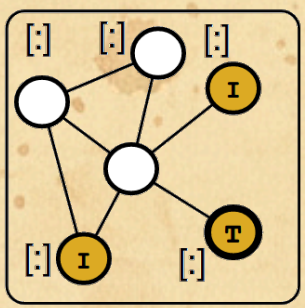
ENCORE UNE PAGE DE PUB??: LES ID/PS



ENCORE UNE PAGE DE PUB??: LES ID/PS

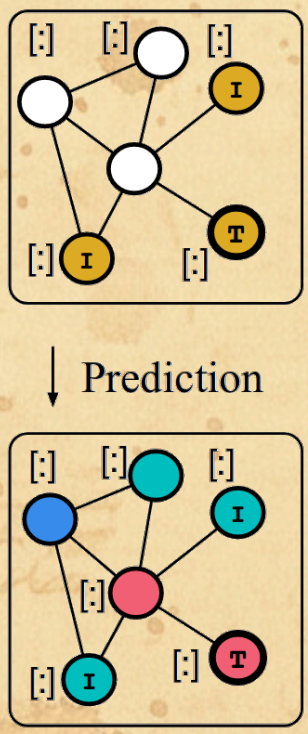


Remuer dans le sens ~~horaire~~ → anti-horaire



Property of
le magicien quantique

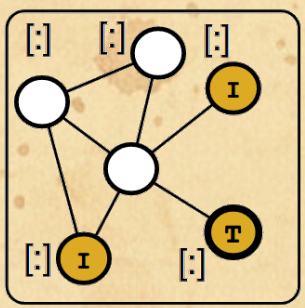
Remuer dans le sens ~~horaire~~ → anti-horaire



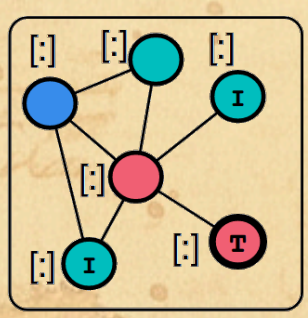
"Class 2"
80.4% confidence

Property of
le magicien quantique

Remuer dans le sens ~~horaire~~ → anti-horaire

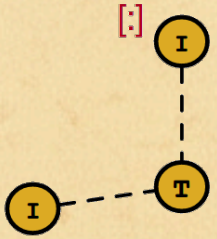


Prediction



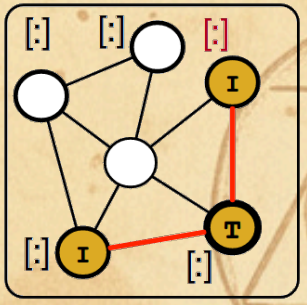
"Class 2"
80.4% confidence

+



Perturbations

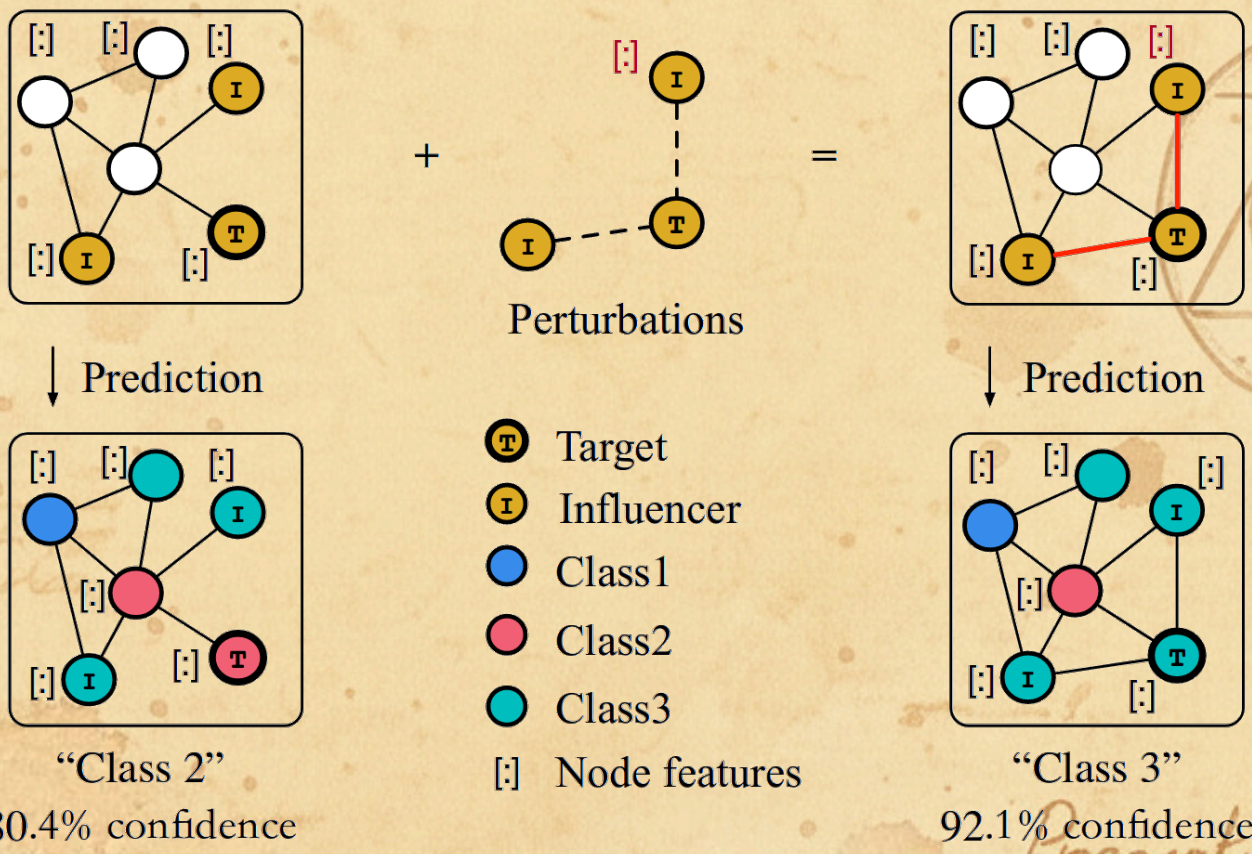
=



- T Target
- I Influencer
- Class1
- Class2
- Class3
- [:] Node features

Property of
le magicien quantique

Remuer dans le sens ~~horaire~~ → anti-horaire

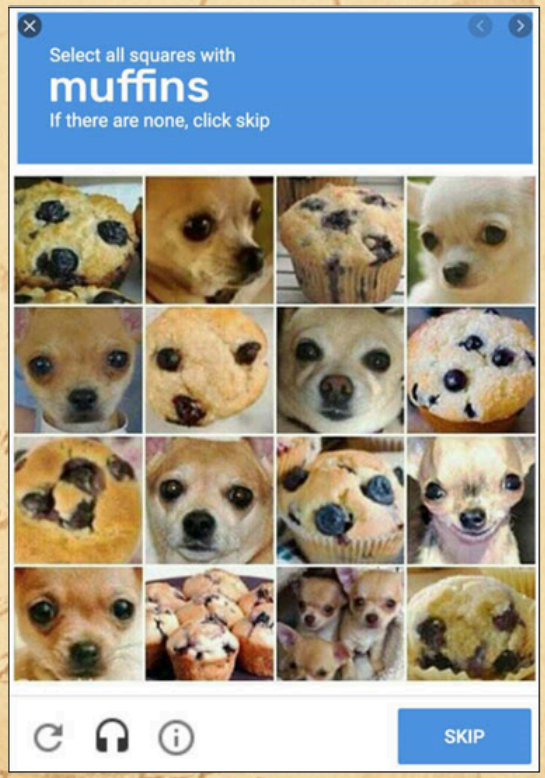


Formule pour cook un challenge du 404

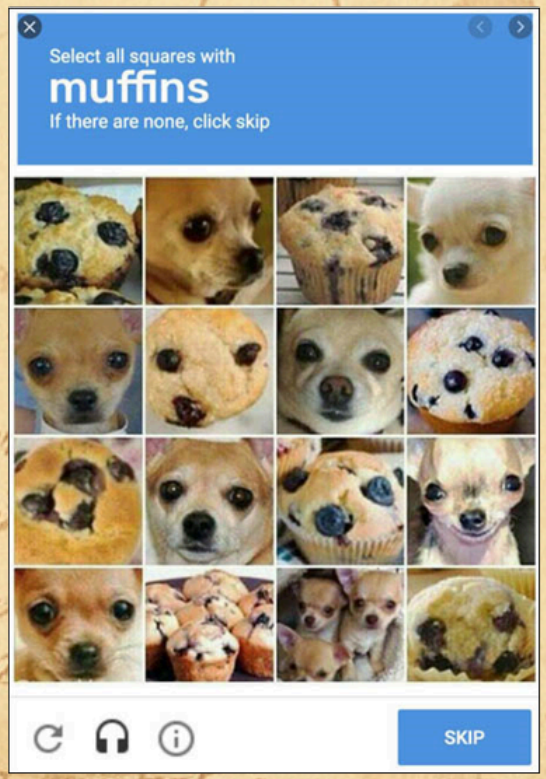
1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le ~~sens horaire~~ sens anti-horaire
3. Ajouter ~~16~~ 17 gouttes de potion de Babillage
4. Couper 2 têtes d'hydre, et mélanger le tout

Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage



Ajouter ~~16~~ → 17 gouttes de potion de Babillage



↓

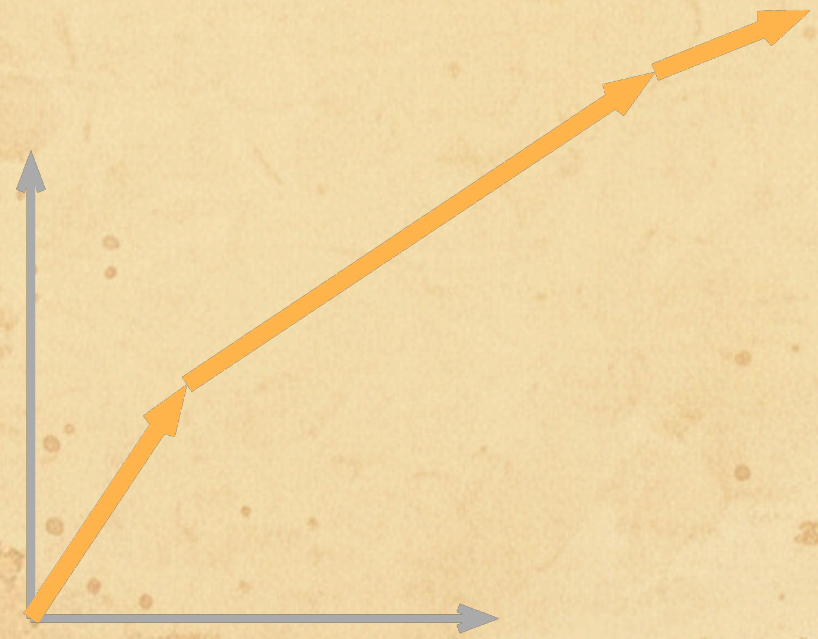
4

↓

5

Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage

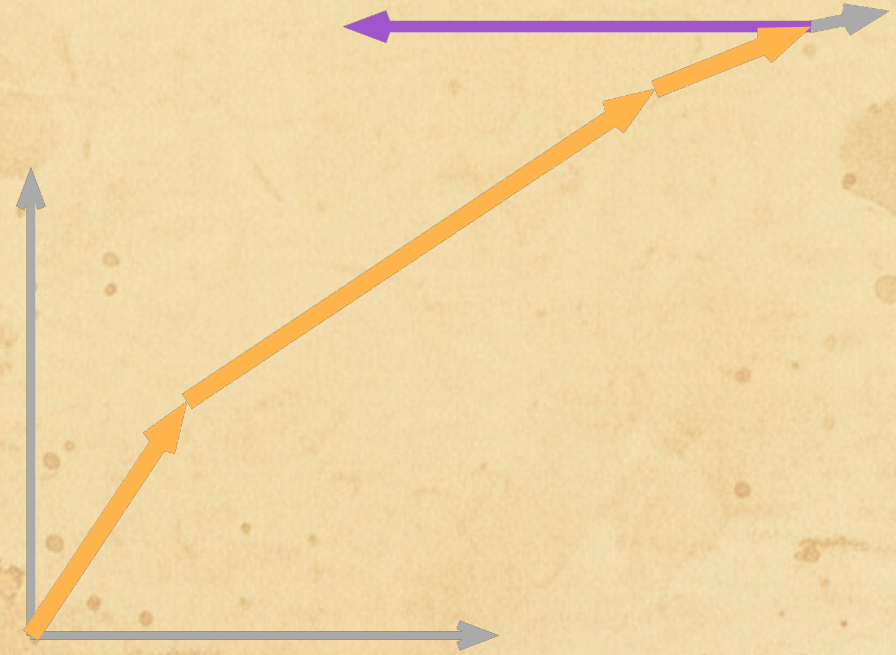


Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage

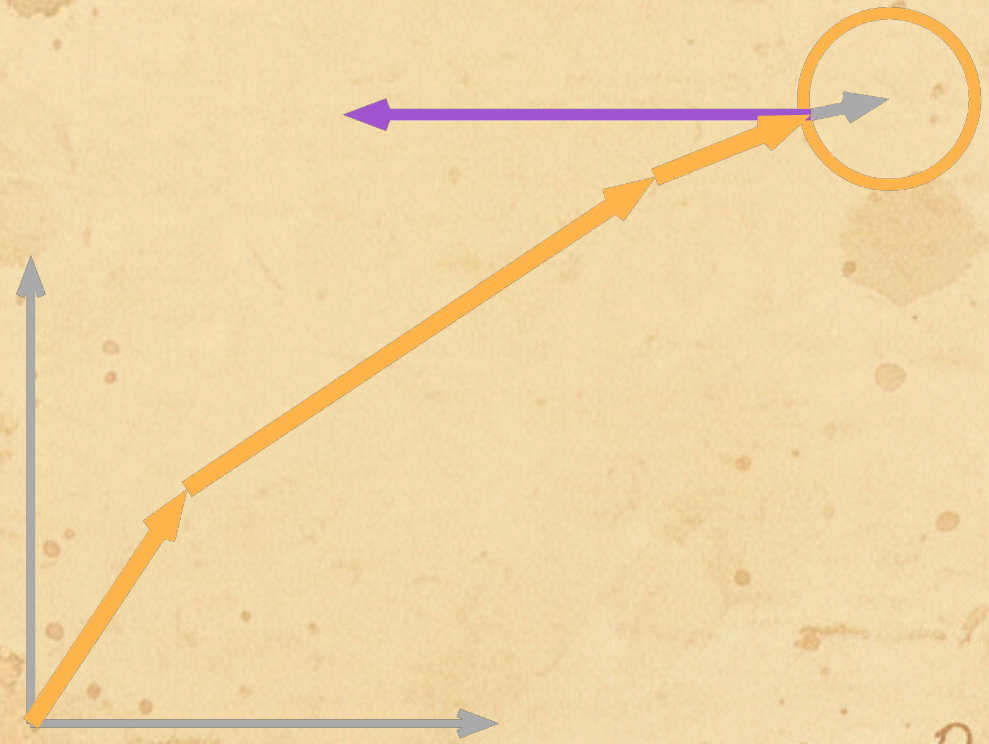
Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage



Property of
le magicien quantique

Ajouter ~~16~~ → 17 gouttes de potion de Babillage



Property of
le magicien quantique

Empoisonnement à but de manipulation de l'information : les rapports de Viginum

Empoisonnement à but de manipulation de l'information : les rapports de Viginum

→ Manipulation d'algorithmes et instrumentalisation d'influenceurs

→ Défis et opportunités de l'intelligence artificielle dans la lutte contre les manipulations de l'information

→ Portal Kombat, un réseau structuré et coordonné de propagande prorusse

Formule pour cook un challenge du 404

1. Ajouter un ~~chat~~ → une théière
2. Remuer 3 fois dans le ~~sens horaire~~ → sens anti-horaire
3. Ajouter ~~16~~ → 17 gouttes de potion de Babillage
4. ~~Couper~~ 2 têtes d'hydre, et mélanger le tout

Détourner l'attention (elles repoussent sinon)

Property of
le magicien quantique

Interpretabilite mecanique

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

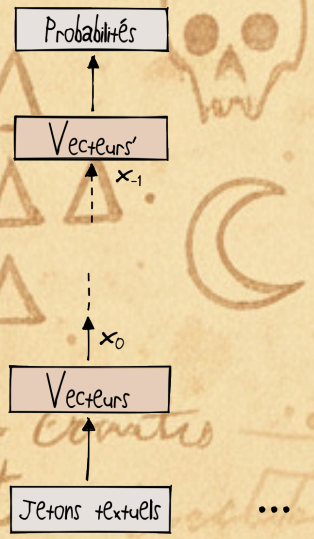
Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

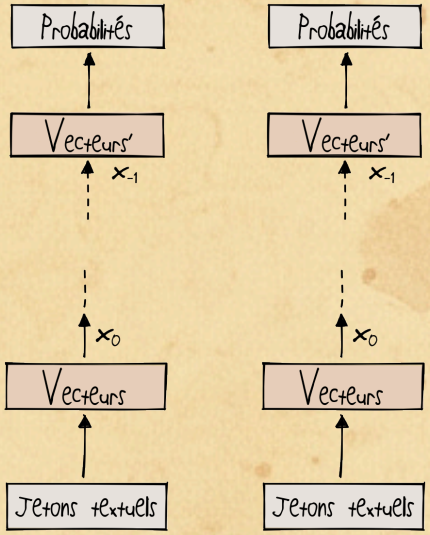
Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

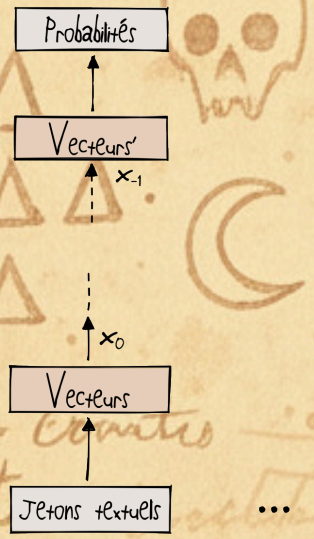
Property of
le magicien quantique



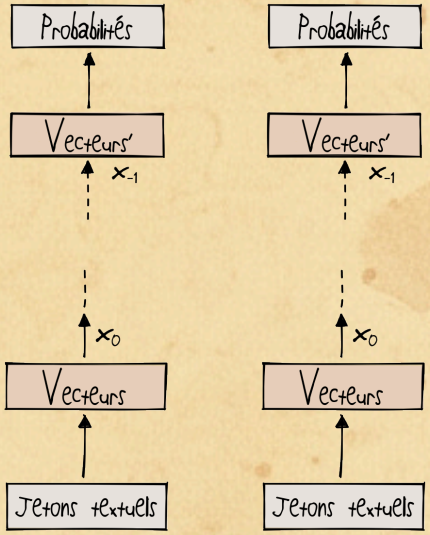
...



Quelle est la couleur du chat de Hermione Granger ?



...



Quelle est la couleur du chat de Hermione Granger ?

Quelle est la couleur du chat de Hermione Granger ?

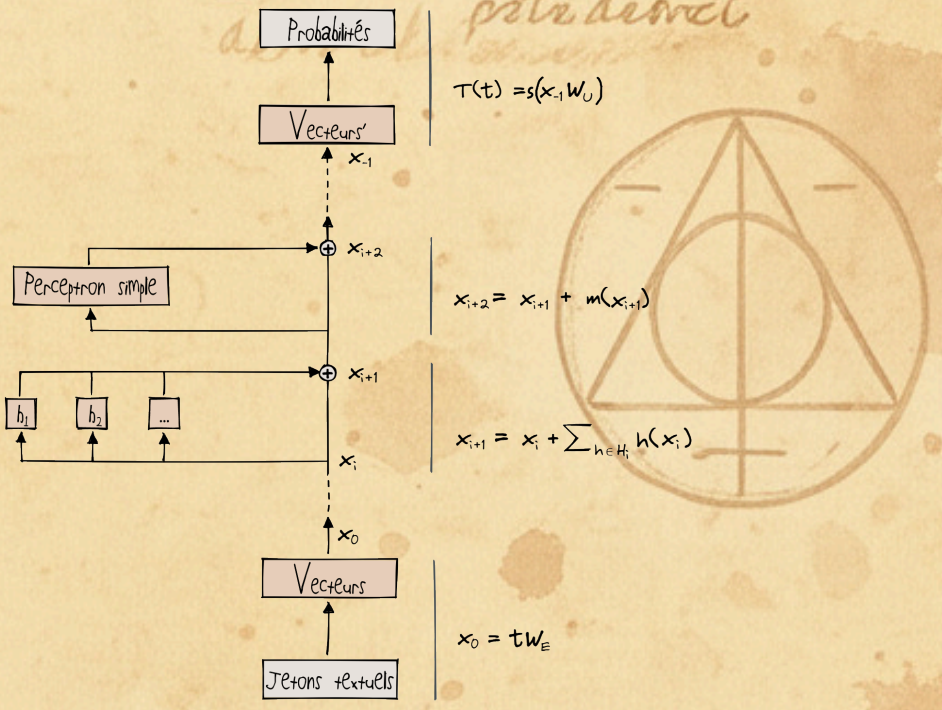
+1 +n-1 +n

Quelle est la couleur du chat de Hermione Granger ?

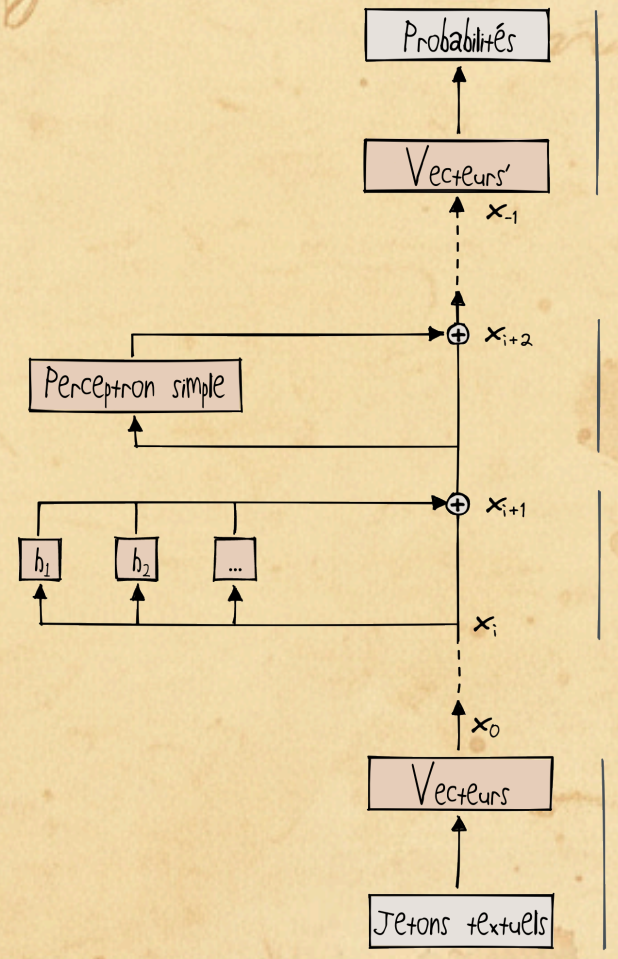
Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique



Quelle est la couleur du chat de Hermione Granger ?



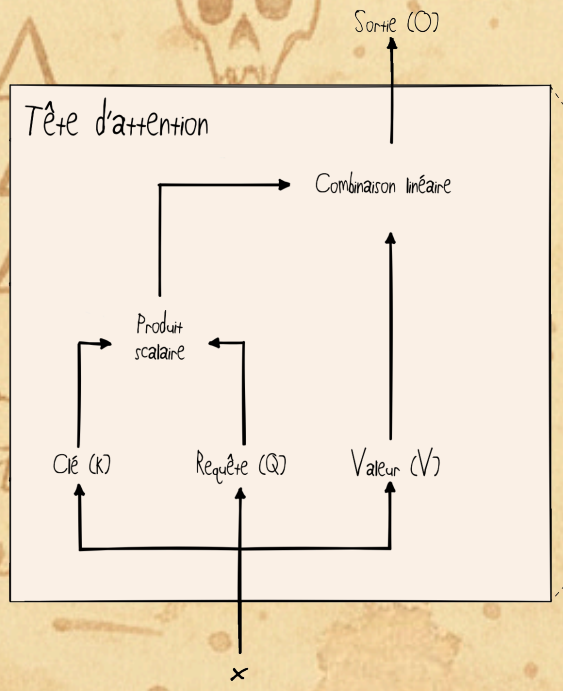
$T(t) = s(x_{-1} W_U)$

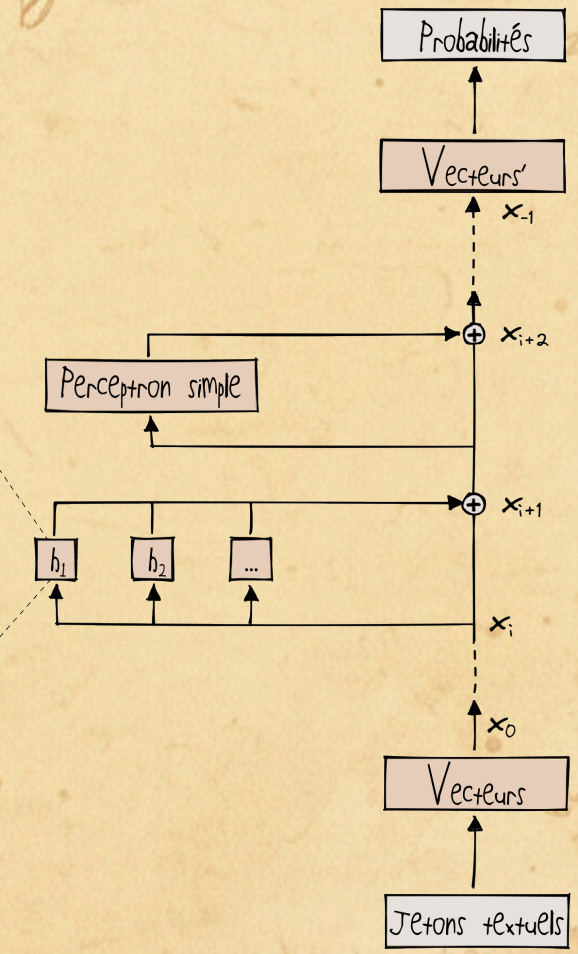
$x_{i+2} = x_{i+1} + m(x_{i+1})$

$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$

$x_0 = t W_E$

Property of
le magicien quantique





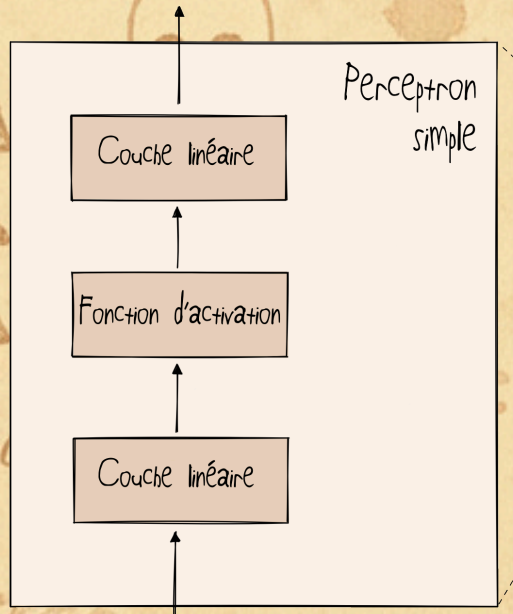
$T(t) = s(x_{-1} W_U)$

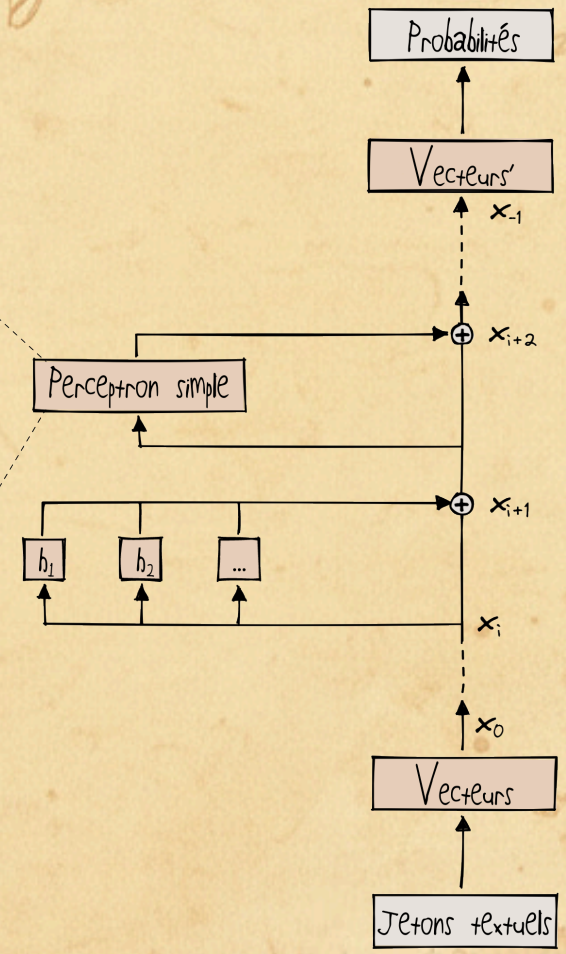
$x_{i+2} = x_{i+1} + m(x_{i+1})$

$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$

$x_0 = tW_E$

Property of
le magicien quantique





$$T(t) = s(x_{-1} W_U)$$

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

$$x_0 = t W_E$$

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Requête :
Y a-t-il un marqueur de question ?

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Clé : Je suis un pronom interrogatif

...

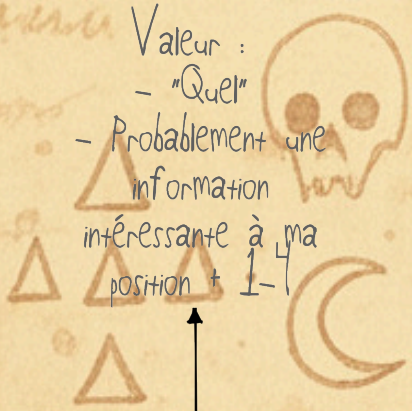
Clé : Je suis un adjectif

...

Requête : Y a-t-il un marqueur de question ?

Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of
le magicien quantique



Valeur :
- "Quel"
- Probablement une information intéressante à ma position + 1-4

Clé :
Je suis un pronom interrogatif

...

Clé :
Je suis un adjectif

...

Requête :
Y a-t-il un marqueur de question ?

Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of
le magicien quantique

Valeur :
- "Quel"
- Probablement une information intéressante à ma position + 1-4

Canal mis à jour :
- Je suis une question
- Il y a un mot interrogatif à position = 1
- Il y a potentiellement une information intéressante après
- (moins important) Il y a un chat O.o

?

Clé :
Je suis un pronom interrogatif

...

Clé :
Je suis un adjectif

...

Requête :
Y a-t-il un marqueur de question ?

Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Requête :
Informations
sur moi ?

Quelle est la couleur du chat

Property of
le magicien quantique

Clé : Je suis un pronom interrogatif

Clé : Je suis un attribut

Requête : Informations sur moi ?

Quelle	est	la	couleur	du	chat	...

Property of
le magicien quantique

Valeur :
- "couleur"

Canal mis à jour :
- Je suis un chat
- J'ai une couleur

Clé :
Je suis un pronom interrogatif ...

Clé :
Je suis un attribut

Requête :
Informations sur moi ?

Quelle	est	la	couleur	du	chat	

Property of
le magicien quantique

Valeur :
- "couleur"

Canal mis à jour :
- Je suis un chat
- J'ai une couleur

→ Jeton suivant ?

Clé :
Je suis un pronom interrogatif ...

Clé :
Je suis un attribut

Requête :
Informations sur moi ?

Quelle	est	la	couleur	du	chat	

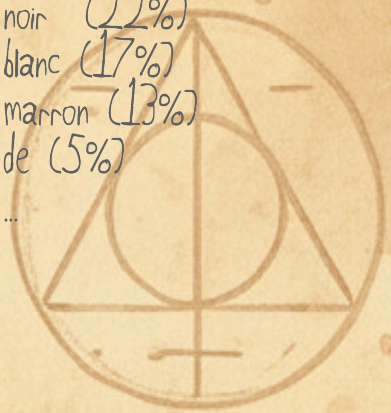
Property of
le magicien quantique

Valeur :
- "couleur"

Canal mis à jour :
- Je suis un chat
- J'ai une couleur

Jeton suivant ?

- noir (22%)
- blanc (17%)
- marron (13%)
- de (5%)
- ...



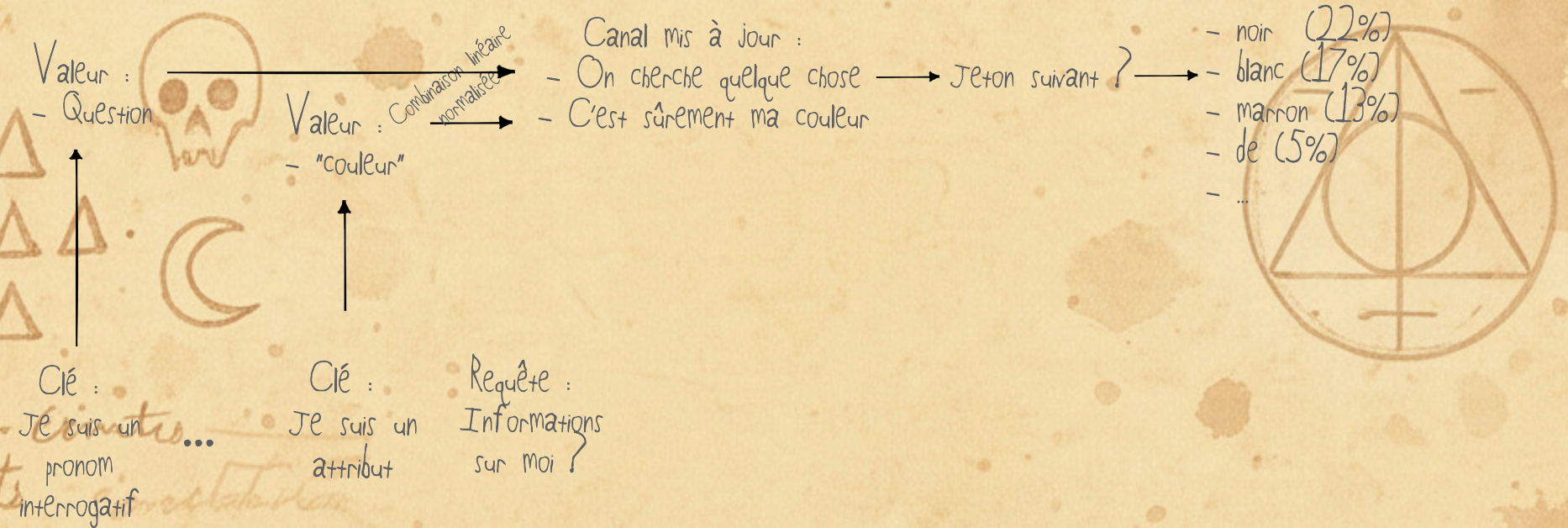
Clé :
Je suis un pronom interrogatif ...

Clé :
Je suis un attribut

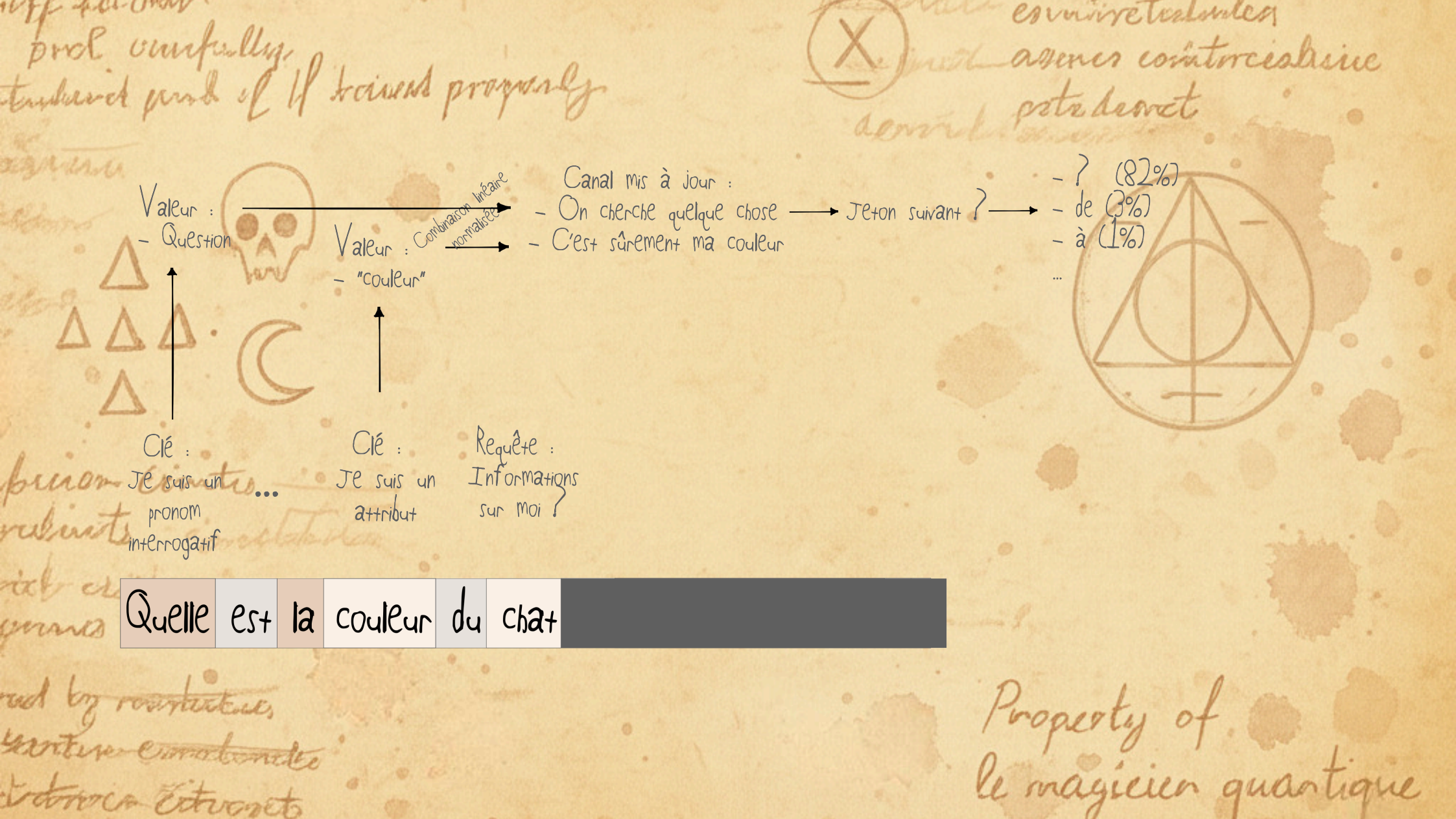
Requête :
Informations sur moi ?

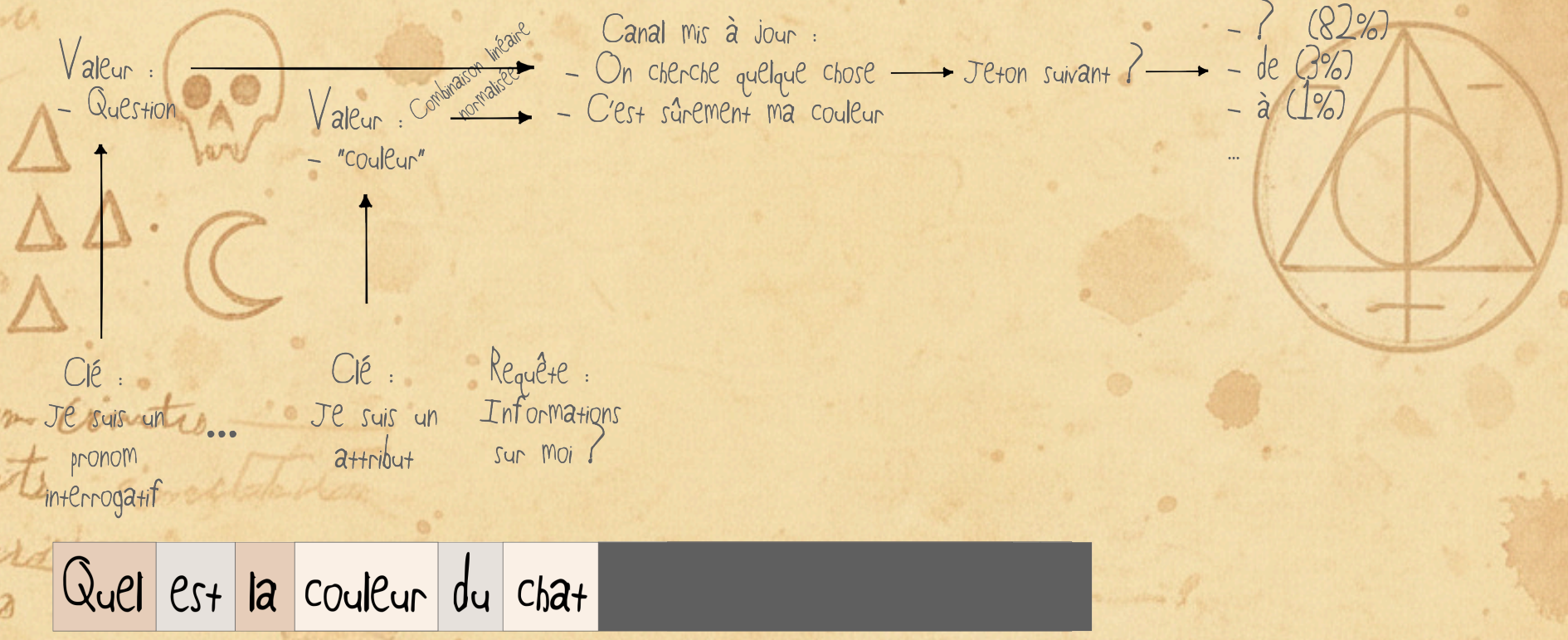
Quelle	est	la	couleur	du	chat	

Property of
le magicien quantique



Property of
le magicien quantique


Valeur :
- Question
Valeur :
- "couleur"
Combinaison linéaire normalisée
Canal mis à jour :
- On cherche quelque chose
- C'est sûrement ma couleur
Jeton suivant ?
- ? (82%)
- de (3%)
- à (1%)
...
Clé :
Je suis un pronom interrogatif
Clé :
Je suis un attribut
Requête :
Informations sur moi ?
Quelle est la couleur du chat



Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

[redacted] Hermione Granger

Property of
le magicien quantique

Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat

[redacted] Hermione Granger

Property of
le magicien quantique

Espace du perceptron (mémoire)

Hermione Granger :
- Elle a un gros chat aux longs poils orange, et à la queue touffue. Il s'appelle Pattenrond
- Son patronus est une loutre
- A faillit finir chez Serdaigle...
- Bois de vigne, ventricule de dragon
- ...

Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat

Hermione Granger

Property of
le magicien quantique

Espace du perceptron (mémoire)

Hermione Granger :
- Elle a un gros chat aux longs poils orange, et à la queue touffue. Il s'appelle Pattenrond
- Son patronus est une loutre
- A faillit finir chez Serdaigle...
- Bois de vigne, ventricule de dragon
- ...

Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat, qui s'appelle Pattenrond et qui est orange

Je suis Hermione Granger, et la phrase est une question sur la couleur de mon chat

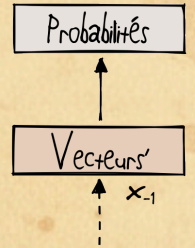
Hermione Granger

Property of
le magicien quantique

Quelle est la couleur du chat de Hermione Granger ?

Property of
le magicien quantique

Jeton suivant :
"Orange" (80%), retour_chariot (15%), "orange" (4%), ...



Quelle	est	la	couleur	du	chat	de	Hermione Granger	?

Property of
le magicien quantique


TP
https://sckathach.github.io/tp
Property of
le magicien quantique

Ressources

- Cours généraux orienté sureté de l'IA: AI Safety Fundamentals de Blue Dot (https://aisafetyfundamentals.com/)
- Excellentissimes cours techniques sur la rétro-ingénierie de LLM: ARENA (https://arena-chapter1-transformer-interp.streamlit.app/)
- Forum technique à suivre: Alignement Forum (https://www.alignmentforum.org/)