Impact des Big Data sur la recherche en sciences du vivant et leurs implication / applications

Antoine Cornuéjols

AgroParisTech – INRA MIA 518

antoine.cornuejols@agroparistech.fr

Plan

Où l'on parle d'« avalanche » de données

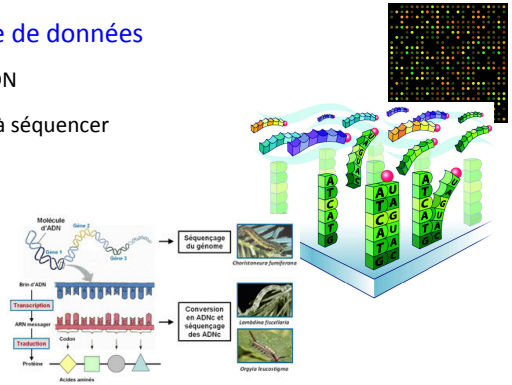
- Des données capturées à foison quand nous allons **sur Internet**
  - Sur quels sites
  - Combien de temps, les clics, les durées, les achats, …
- **Smartphones**
  - Localisation même si on a dit non
  - Des tas d'applies pleines de curiosité
- **Bracelets** connectées
- Moyens de **paiement** (banques)
- Capteurs dans les **véhicules** (assurances)
- Compteurs Linky

Où l'on parle d'« avalanche » de données

- Des **caméras** dans les panneaux de publicité dans les rues
- Bientôt dans les **vitrines** des magasins
- **Factures** dans les supermarchés (carte de fidélité)
- **Smart cities**

Exemples de domaines renouvelés

- La bio-informatique
  - Gros volume de données
    - Puces à ADN
    - Machines à séquencer



Exemples de domaines renouvelés

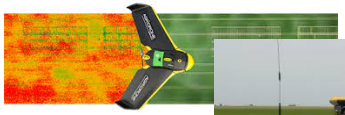
- La sociologie
  - Gros volume de données
    - Réseaux sociaux
    - Smarphones
    - Websites consultations

Exemples de domaines renouvelés

- La e-medecine (le me-data)
  - Gros volume de données
    - Smarphones
    - Objets connectés
    - Forums
    - WATSON
    - Google Flu

Exemples de domaines renouvelés

- L'agriculture numérique
  - Gros volume de données
    - Capteurs
    - Drones
    - Réseaux sociaux et pro



Exemples de domaines renouvelés

- Le domaine juridique
  - Gros volume de données
    - Archives numérisées
    - Réseaux sociaux et professionnels

Des **contre**-exemples

- L'**alimentation**
  - Enquête **Nutrinet**
    - ~ 277 000 internautes théoriquement sur des années
    - **Mais**
      - à 80% des femmes
      - Milieux socio-professionnels élevés
      - Abandonnent après quelques jours

Manque de données représentatives

- L'**éducation**
  - Peu de données sur ce qui se passe en classe ou devant un écran

Les sciences du vivant

de l'environnement et de l'agronomie

Spécificités

1. Des systèmes **naturels**
   - Les **modèles** sous-jacents, les liens de **causalité** sont **inconnus**
2. Très **complexes**
   - très **multi-échelles** spatiales et temporelles
3. Très **multi-factoriels**

« Défricher »

4. **Adaptatifs** donc difficiles à prévoir
   - E.g. émergence de bio-résistance

Importance des liens de **causalité**

Spécificités

1. La disposition de données permet
   - D'**explorer** : recherche de « patterns »
2. Possibilité d'une **approche heuristique**
   - On ne cherche pas nécessairement un **modèle explicatif** ou causal
   - On peut se contenter de **modèles prédictifs** (dans un 1er temps)

Changement de paradigme

1. **Ancien** paradigme
   - **Construire une hypothèse** (e.g. tel traitement devrait avoir tel effet)
   - Construire un plan d'expérience pour **tester la validité de l'hypothèse**
   - Le dispositif expérimental et les données récoltées **ne servent qu'à tester cette hypothèse**

1. **Ancien** paradigme
   - **Construire une hypothèse** (e.g. tel traitement devrait avoir tel effet)
   - Construire un plan d'expérience pour **tester la validité de l'hypothèse**
   - Le dispositif expérimental et les données récoltées **ne servent qu'à tester cette hypothèse**

2. **Nouveau** paradigme
   - Esprit « ouvert » : on cherche des patterns (inattendus) dans la masse de données disponibles
   - Ré-utilisation possible à l'infini des données (non récoltées pour un but précis)

C'est le « **data mining** »

Plan

1. La révolution des données
2. Qu'en fait-on ?
3. Les défis
4. Les compétences requises, les acteurs, les outils
5. L'avenir

Apprentissage **descriptif**

non supervisé

Apprentissage descriptif

À propos d'un *échantillon d'apprentissage* $S = \{(x_i)\}_{1,m}$

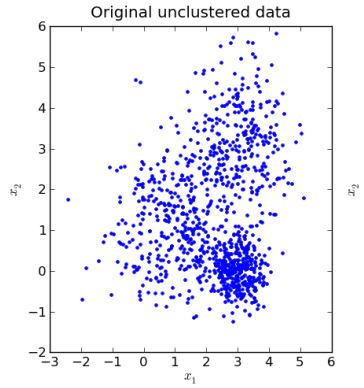
identifier des régularités rendant compte de S

- E.g. sous la forme de **clusters** (e.g. *mélange de Gaussiennes*)
  - CLUSTERING
- E.g. sous la forme de **motifs fréquents** (fouille de données)

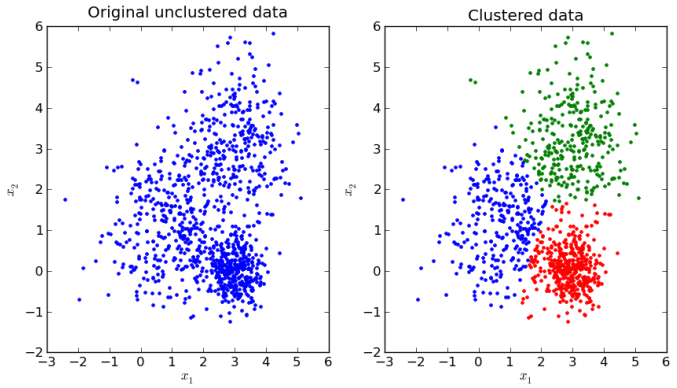
pour résumer, suggérer des régularités, comprendre …

→ Clustering / Catégorisation

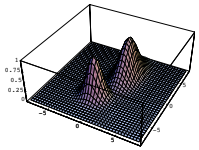
- Apprentissage « **descriptif** » (non supervisé)



- Apprentissage « **descriptif** » (non supervisé)



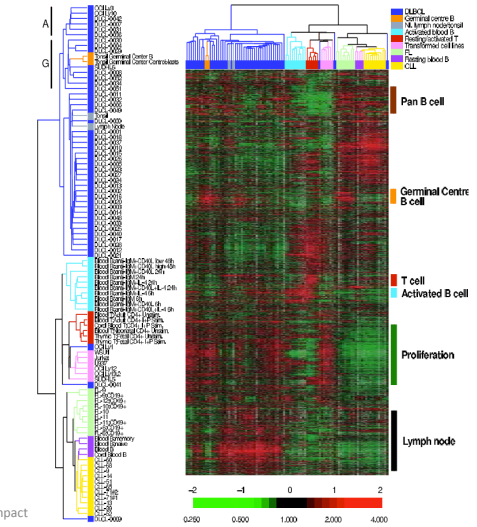
- Catégorisation de consommateurs
  - Base de données sur les répondants de la base Nutrinet
    - ~ 280 000
    - Données sur *âge*, *nb de personnes dans la famille*, *catégorie socio-professionnelle*, …
    - Données sur consommations alimentaires sur une certaine durée
  - Y a-t-il émergence de **groupes** distincts ?



Apprentissage
Non supervisé

Clustering

Bi-clustering
gènes - patients



→ Recherche de motifs fréquents

Frequent Item Sets

→ Recherche de règles d'association

- Extraire des **régularités**
  - Base de données sur les **consommations alimentaires**
  - Peut-on identifier des « patterns » de consommation ?

- Extraire des **régularités**
  - Base de données sur les **consommations alimentaires**
  - Peut-on identifier des « patterns » de consommation ?

- Reconnaissance d'animaux malades ou en chaleur
  - Mesures en continu sur leur comportement
    - Vidéos
    - Capteurs « embarqués »
      - Mobilité (nb de pas / minute ; distance parcourue à l'heure)
      - Lieux visités
      - …

➢ *Reconnaissance de **comportements types***

Apprentissage **prédictif**

supervisé

Apprentissage prédictif (*supervisé*)

- Un *échantillon d'apprentissage*

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$$

(f : $x_i \to y_i$; h : $y_i \to x_i$)

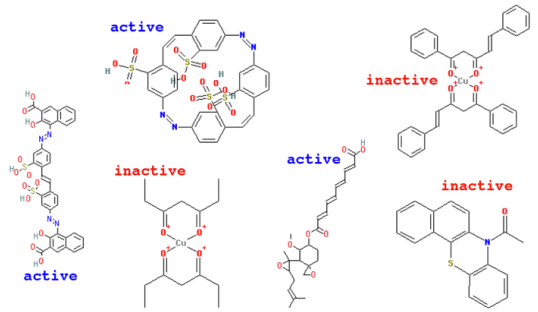
Prédiction pour de **nouveaux** exemples x –*h*-> y ?

- Reconnaissance d'insectes ravageurs
  - Base d'images d'insectes dans des cuvettes
    - *Reconnaissance du type d'insectes*
    - *Comptage*

Association / Prédiction

Apprentissage **supervisé**

- Prédire si une molécule est bio-active ou pas



NCI AIDS screen results (from http://cactus.nci.nih.gov).

Analyse de textes

- Reconnaissance de **sentiments** exprimés dans des textes

	Electronics	Video games
✓	(1) Compact; easy to operate; very good picture quality; looks sharp!	(2) A very good game! It is action packed and full of excitement. I am very much hooked on this game.
✓	(3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and sharp.	(4) Very realistic shooting action and good plots. We played this and were hooked.
✗	(5) It is also quite blurry in very dark settings. I will never_buy HP again.	(6) It is so boring. I am extremely unhappy and will probably never_buy UbiSoft again.

GIEC : filtrage de documents

- Estimation de l'émission de gaz à effet de serre par les sols agricoles
  - En particulier N_2O (influence des engrais azotés)
- Par une méta-analyse des **articles scientifiques pertinents**
  - **Plus de 10^6 articles** scientifiques publiés / an
  - (plus ou moins) disponibles sur Internet

Filtrage nécessaire de ces articles

En optimisant précision et rappel

(et interprétabilité du filtre)

Apprentissage prescriptif

pour « intervenir »

Apprentissage prescriptif

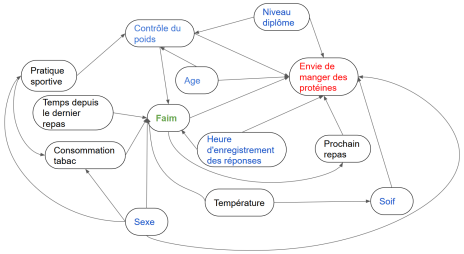
- Apprentissage « **prescriptif** » (recherche de ***causalités***)

1. J'observe que les gens qui mangent des glaces sont souvent en maillot de bain
2. Je voudrais vendre davantage de glaces

→ Je demande aux gens de se mettre en maillot de bain

La recherche de relations causales

Qu'est-ce qui **cause** l'appétence pour des plats protéinés ?



- La **faim** ?
- **L'heure** dans la journée ?
- Le **genre** ?
- L'aspect **visuel** ?
- L'aspect **olfactif** ?
- La richesse en **protéines** des **repas précédents** ?
- …

- Quelles **recommandations** faire à un consommateur pour qu'il baisse sa consommation d'aliments carnés ?
- Quel impact **si on double le prix** de … ?
- Quel rendement aurais-je eu l'année dernière **si j'avais** planté du ... au lieu de ...

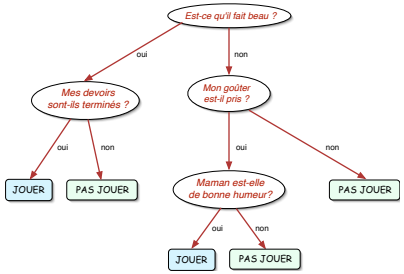
Quels modèles ou hypothèses ?

Modèles interprétables

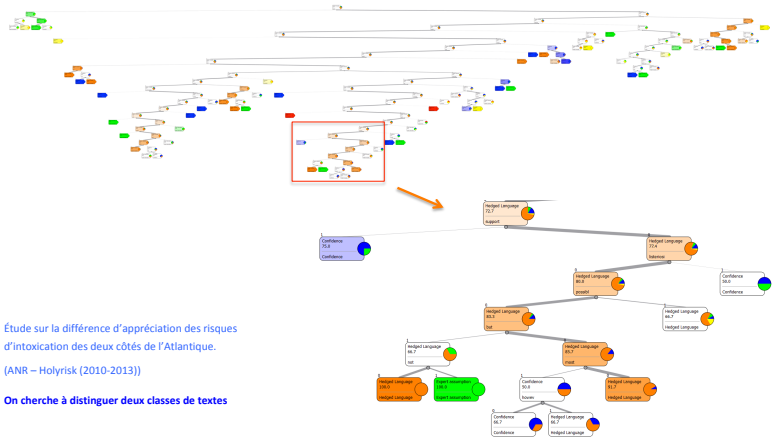
- Régression linéaire

$$y = \sum_{i=1}^{N} \alpha_i \, x_i$$

- Arbre de décision



Exemple : arbre de décision



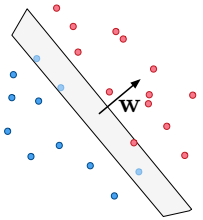
Étude sur la différence d'appréciation des risques d'intoxication des deux côtés de l'Atlantique.

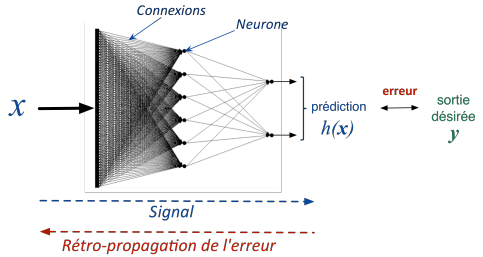
(ANR – Holyrisk (2010-2013))

On cherche à distinguer deux classes de textes

Modèles opaques

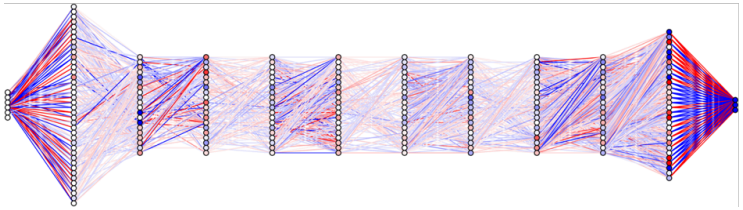
- Régression linéaire généralisé
- Réseaux de neurones





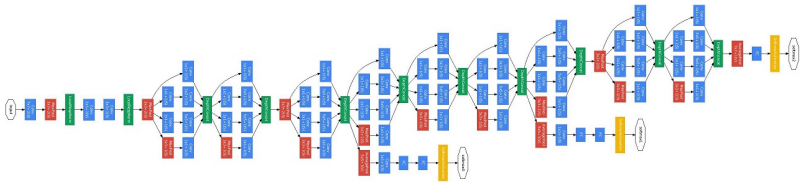
Les « réseaux de neurones profonds »

- Des réseaux de neurones artificiels
  - à grand nombre de couches (parfois > qqs 100)
  - et **très grand nombre de paramètres** (qqs 10^7 – 10^8 paramètres)

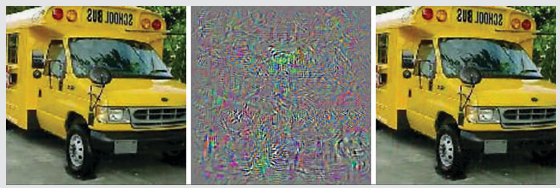


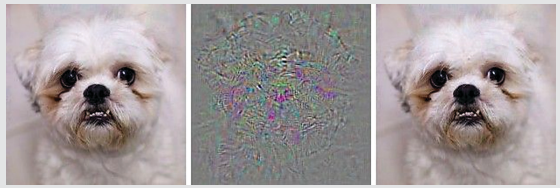
GoogleNet

- Un **mécano** de réseaux de neurones



Des erreurs difficiles à comprendre





Illustration

Adversarial input can fool a machine-learning algorithm into misperceiving images.

Plan

Les défis

1. Le recueil des données

Obtenir les données

Souvent **difficile** !!!

- Les données ne sont **pas encore disponibles**
- Le donneur d'ordre n'est **pas détenteur des données**
  - Pas le même service / département
- Les données sont **protégées par des droits**
- Une partie des données **reste à recueillir**

Les questions juridiques

Essentiel !!!

- Données **personnelles**
- **Obtenir l'autorisation**
  - CNIL
  - RGPD
    - Depuis le **25 mai 2018**, le Règlement Général Européen sur la Protection des Données (RGPD) affecte toutes les organisations traitant les données personnelles identifiables (DPI) de résidents européens.

Les défis

1. Le recueil des données
2. Les prétraitements des données

Les prétraitements

- **90%** du temps d'un projet
- **Recueil** des données
- Mise dans un **format adéquat**
- **Nettoyage**
  - **Bruit** dans les données
  - Données **manquantes**
  - Données **aberrantes**
  - **Doublons**
  - **Normalisation** des mesures
  - **Discrétisation** de valeurs continues
  - **Rendre continues** des valeurs discrètes
- Élimination des **attributs redondants** / calcul de **nouveaux attributs**
- **Précision / incertitude**
- Intégration de plusieurs **sources de données (hétérogènes)**
- **…**

Le traitement des documents en .pdf

- Structure des pages … en .pdf

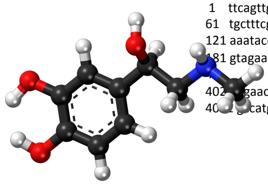
Les défis

1. Le recueil des données
2. Les prétraitements des données
3. Les sources multiples et hétérogènes

Intégration de multiple sources de données

- Annotation de protéines

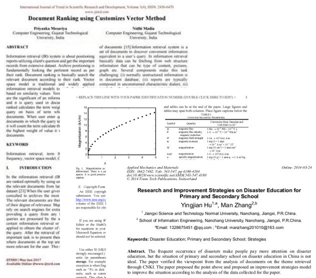
Protéine « sp|P00004|CYC_HORSE » is activated by …

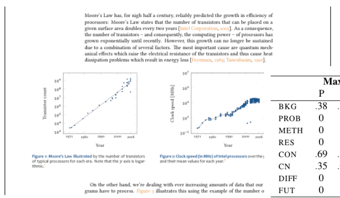


1 ttcagttgtg aatgaatgga cgtgccaaat agacgtgccg ccgccgctcg attcgcactt
61 tgctttcggt tttgccgtcg tttcacgcgt ttagttccgt tcggttcatt cccagttctt
121 aaataccgga cgtaaaaata cactctaacg gtcccgcgaa gaaaaagata aagacatctc
181 gtagaaatat taaaataaat tcctaaagtc gttggtttct cgttcacttt cgctgcctgc
402 gaacacgcc gaggctccat tcatagcacc acttcgtcgt cttaatcccc tccctcatcc
40 catggcgg tgcaaaaaat aaaaagaact c

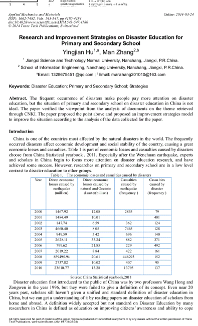
Intégration de multiple sources de données

- GIEC
  - Documents scientifiques multiples
  - Tableaux
  - mesures





	MaxEnt			MaxEnt + GE			Unsup GE		
	P	R	F	P	R	F	P	R	F
BKG	.38	.19	.25	.49	.48	**.48**	.49	.44	.46
PROB	0	0	0	.38	.23	.29	.28	.38	**.32**
METH	0	0	0	.29	.50	**.37**	.08	.56	.14
RES	0	0	0	.68	.51	**.58**	.08	.51	.14
CON	.69	.96	.80	.81	.84	**.82**	.74	.69	.71
CN	.35	.06	.10	.39	.29	**.33**	.40	.13	.20
DIFF	0	0	0	.21	.30	**.25**	.12	.13	.12
FUT	0	0	0	.24	.44	.31	.26	.61	**.36**



Les défis

1. Le recueil des données
2. Les prétraitements des données
3. Les sources multiples et hétérogènes
4. La possibilité de l'intervention de l'expert

Disponibilité des experts métier

Essentiel !!!

- **Comprendre** le problème
- Établir un **vocabulaire commun**
- **Évaluer** les résultats
- Orienter / **ré-orienter**
- **S'approprier** les résultats / assurer la suite

Des algorithmes « transparent »

1. Dans lesquels on puisse « injecter » l'expertise humaine
2. Dont les résultats (modèles appris) soient interprétables

Les défis

1. Le recueil des données
2. Les prétraitements des données
3. Les sources multiples et hétérogènes
4. La possibilité de l'intervention de l'expert
5. L'identification de relations causales
6. Les environnements non stationnaires
7. Un génie logiciel des systèmes apprenants

Plan

Les « data scientists »

- **Compétences** attendues
  1. Apprentissage artificiel / Statistiques
     - Bonne compréhension des questions et des hypothèses sur lesquelles reposent les méthodes
  2. Compétences en informatique
     - Algorithmique
     - Bases de données
     - Réseaux
  3. Capacités relationnelles

Les « data scientists »

- **Compétences** attendues
  1. Apprentissage artificiel / Statistiques
     - Bonne compréhension des questions et des hypothèses sur lesquelles reposent les méthodes
  2. Compétences en informatique
     - Algorithmique
     - Bases de données
     - Réseaux
  3. Capacités relationnelles

En très forte demande

100 000 en France à l'horizon 2022 !!

- **Formations**
  - Quelques dizaines d'heures
  - **Master** ou équivalent
  - **Doctorat**

Grand risque de déconvenue si pas les bons recrutements

Les passages à l'échelle

1. Savoir traiter de (très) **gros volumes de données**
   - Méthodes efficaces
     - Gradient stochastique
     - Apprentissage convexe
     - Optimisation du code
       - ✓ Accès mémoire
       - ✓ Complexité computationnelle
   - Distribution des calculs
     - Cartes graphiques / cœurs
     - Clusters de machines
     - Cloud computing
       - ✓ Approches Map Reduce

Les passages à l'échelle

2. Savoir traiter de (très) **petits volumes de données**

Compenser le manque d'information dans les données

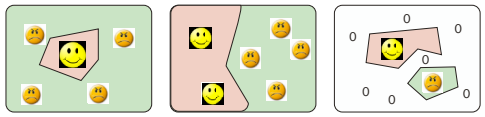
- Par de la **connaissance experte**
- **Enrichissement** des données
  - Ontologies
  - Web sémantique
  - Wikipedia and Co
- Question de la **validation des résultats**
  - Les experts

Les méthodes et algorithmes

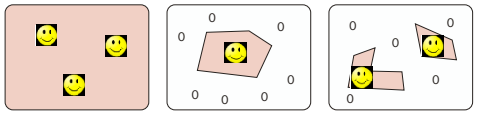
- Librairies / méthodes / algorithmes
  - Sont dans le **domaine public** !!!
    - Publications scientifiques
    - Forums
    - Conférences
    - Librairies (e.g. ScikitLearn)
- Des « **recettes** » privées
  - Réseaux de neurones profonds
  - Traitement d'images / télédétection
  - Connaissances métiers (e.g. alimentation)

Le no-free-lunch theorem

Possible



Impossible



Il faut **choisir** le **bon algorithme** pour la **classe de problèmes** étudiée

Plan

1. La révolution des données
2. Qu'en fait-on ?
3. Les défis
4. Les compétences requises, les acteurs, les outils
5. L'avenir

Une liste …

1. Résoudre les difficultés
   - Données **multi-sources** hétérogènes
   - **Dialogue** possible avec les **experts** : interprétabilité des modèles produits, compréhension et contrôle raisonné des algorithmes
2. Identification de **relations causales**
3. Apprendre à partir de (très) **peu d'exemples**
4. Apprendre en environnement **non stationnaire**
   - Flux de données
   - Transfert entre tâches
5. **Génie logiciel** pour des **systèmes apprenants**

Une révolution en cours

1. Tirer profit des données
   - ✓ Numérisation
   - ✓ Capteurs partout
   - ✓ Internet
   - ✓ Des ressources calcul
   - ✓ Des algorithmes
2. Gros progrès en intelligence artificielle

Mais ce n'est pas « magique »

Beaucoup d'opportunités

Mais pas de magie

Conclusions

4 approches pour appréhender le monde

1. Empirique : description et classement

4 approches pour appréhender le monde

1. Empirique : description et classement

2. Théorique : Modélisation, construction de théories

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

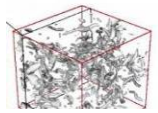
4 approches pour appréhender le monde

1. Empirique : description et classement

2. Théorique : Modélisation, construction de théories

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

3. Simulation : systèmes complexes et/ou non reproductibles



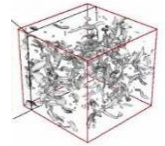
4 approches pour appréhender le monde

1. Empirique : description et classement

2. Théorique : Modélisation, construction de théories

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

3. Simulation : systèmes complexes et/ou non reproductibles



4. Exploration de données
   - Énormes masses de données numérisées
   - Largement disponibles
   - Sources et formats très différents

Le cas AlphaGo

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

Le cas AlphaGo : comprendre

Fan Hui, Gu Li, Zhou Ruyang (très forts joueurs de Go) se reconvertissent dans l'analyse des parties jouées par AlphaGo

- Sorte d'exégèse. Explications a posteriori
- Nécessaire pour
  - La communication
  - L'enseignement

Et même AlphaGo peut se tromper