De la Statistique à l'IA en passant par la Science des Grosses Données

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Buzzwords : de la Statistique à l'*IA hybride* par la *Science des Données*

		Statistique	Informatique	Algos–Technos
1930-60s	HO	Statistique Inférentielle	Début de l'IA (1955)	Régression / Perceptron
1970s	KO	*Exploratory Data Analysis*	Systèmes experts	Composantes Principales
1980s	MO	Statistique fonctionnelle	Réseaux de neurones	*CARTrees*
1990s	GO	*Data mining* données pré-acquises		*Boosting, SVM*
2000s	TO	$p >> n$	*Machine Learning*	*Lasso, random forest*
2008		*Data Scientist*		
2010s	PO	p et n très grands	*Big Data*	*Hadoop*
2012			*Deep Learning*	*ConvNet, TensorFlow*
2016		Intelligence Artificielle AlphaGo, Zero…		*XGBoost*
2019		IA hybride ANITI, Deep4Cast...		

Définition

- *Analyste, ça fait trop Wall Street ; statisticien, ça agace les économistes ; chercheur scientifique, ça fait trop académique. Pourquoi pas "data scientist" ?* (D.J. Patil *LinkedIn* et J. Hammerbacher, *Facebook*, 2008)
- Data scientist (n) : *Person who is better at statistics than any software engineer and better at software than any statistician* (J. Wills, *Cloudera*)

Logiciels de fouille de données

- Clementine (SPSS – IBM)
- Enterprise Miner (SAS)
- Insightfull Miner (Splus)
- KXEN, SPAD,
- Statistica Data Miner
- Statsoft, Tanagra, Weka
- ... Changement de modèle commercial ...

R *vs.* Python

- Langage R, librairies, `caret`
- Langage Python, `pandas, scikit-learn`
- Comparaison
  - Mémoire : data *munging vs.* apprentissage
  - Parallélisation
  - Classe `Data Frame`
  - Sélection de modèle linéaire général
  - Élagage d'un arbre
- Julia ?

Reproductibilité

- Donoho (2015)
- Chaîne de traitements (*pipeline*) automatisée
- Production automatique de rapports
  - R : `sweave, knitr`
  - Python : `pweave`
- Tutoriels : Notebook IPython ou Jupyter (Python, R, Julia...)
- http://github.com/wikistat

De la statistique à la Science des Données et l'IA

Saison 1 Statistique élémentaire descriptive et inférentielle

Saison 2 Statistique exploratoire multidimensionnelle et classification non supervisée (*clustering*)

Saison 3 Apprentissage automatique / Statistique
Apprentissage supervisé :
Régression (linéaire, logistique, PLS), analyse discriminante, k-p.p., arbres de décision, *random forest*, *boosting*, réseaux de neurones (*deep learning*), imputation, détection d'anomalies.

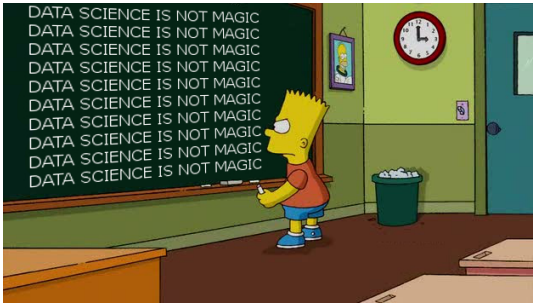
Saison 4 Apprentissage en grande dimension

Saison 5 Technologies pour l'IA : *Spark, MLlib, TensorFlow*

Vous avez échappé à :

- Apprentissage machine mais pas ”statistique” : symbolique, règles d'association...
- Données complexes (graphes, trajectoires, vidéos)
- Flux de données et décision séquentielle (bandits)
- Apprentissage par renforcement


BIG DATA


DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC
DATA SCIENCE IS NOT MAGIC