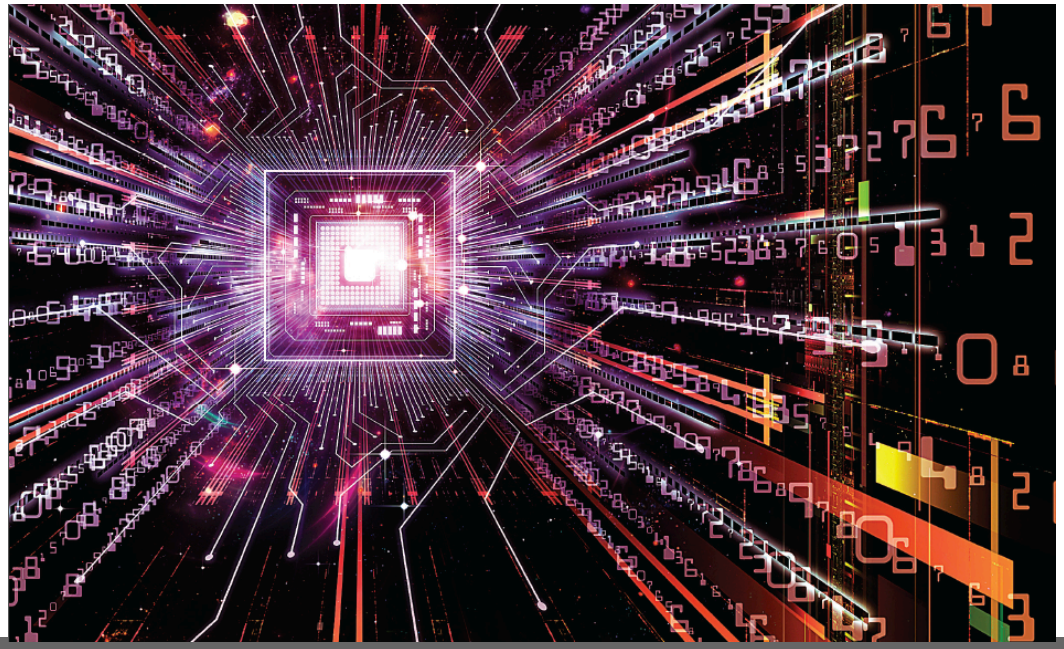
On Machine Learning

Enggelbert Mephu Nguifo

Usages ?



Machine Learning

Why ?

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, *A. Big data: The next frontier for innovation, competition, and productivity*. Technical report, **McKinsey Global Institute**, 2011.

"Machine learning (a.**k**.a. data mining or predictive analytics) will be the **driver of the next big wave of innovation**"

Machine Learning

Why ?

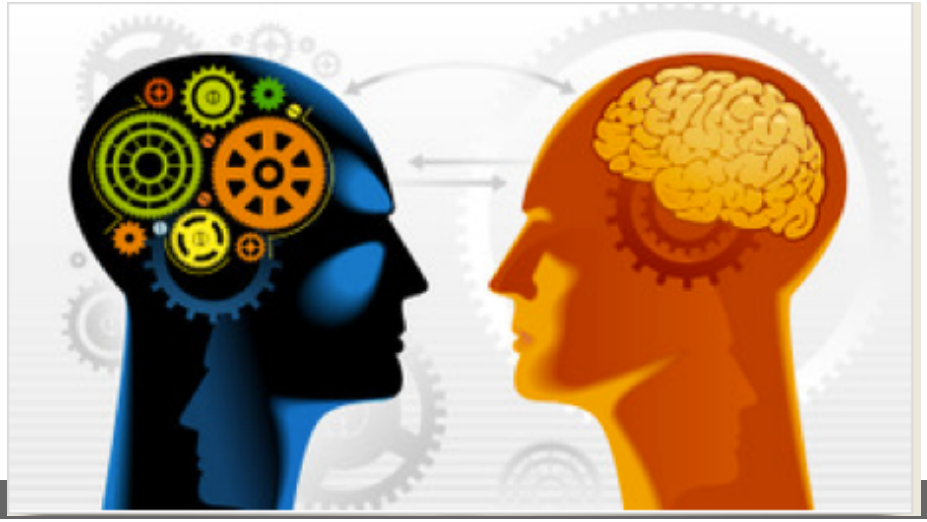
ML algorithms can figure out how to perform important tasks by **generalizing from examples**. This is often feasible and cost-effective where manual programming is not.

As **more data** becomes available, **more ambitious problems** can be tackled.

Machine Learning

Why ?

ML is widely used in computer science and other fields. However, developing **successful ML** applications requires a substantial amount of **“black art”** that is difficult to find in textbooks.



Machine Learning

What's in ?

Problem Setting:

- Set of possible instances X
- Unknown target function $f : X \rightarrow Y$
- Set of function hypotheses $H=\{ h \mid h : X \rightarrow Y \}$

Function approximation

Input:

superscript: i^{th} training example

- Training examples $\{<x^{(i)}, y^{(i)}>\}$ of unknown target function f

Output:

- Hypothesis $h \in H$ that best approximates target function f

http://www.cs.cmu.edu/~tom/10701_sp11/slides/DTreesAndOverfitting-1-11-2011_final.pdf

Machine Learning

What's in ?

ML = Representation + Evaluation + Optimization

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Machine Learning

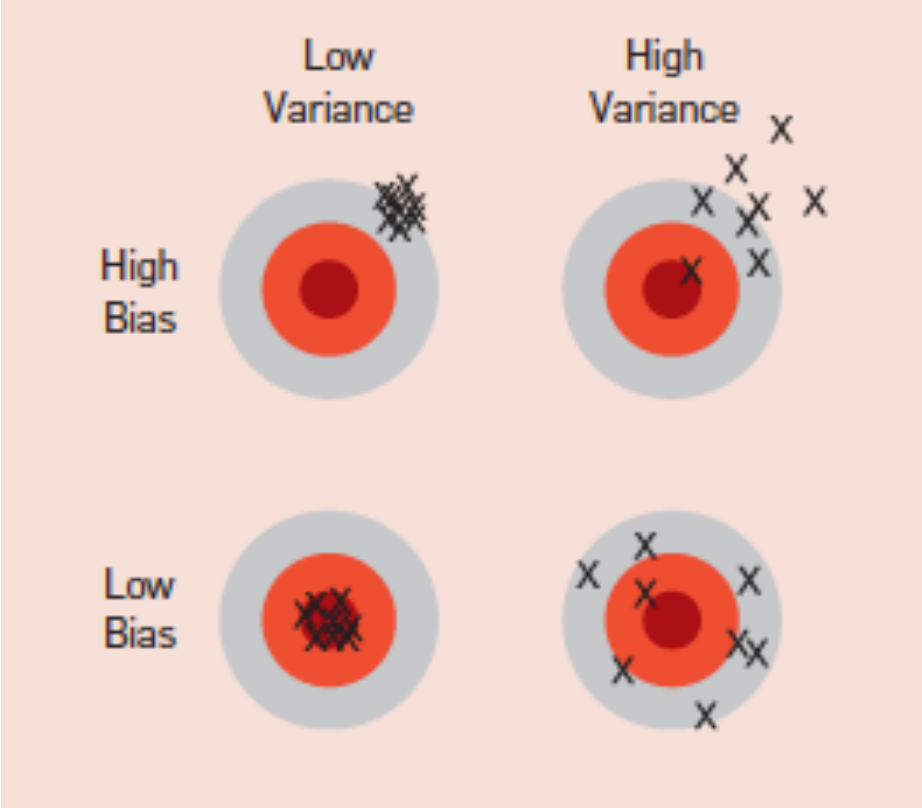
What's up ?

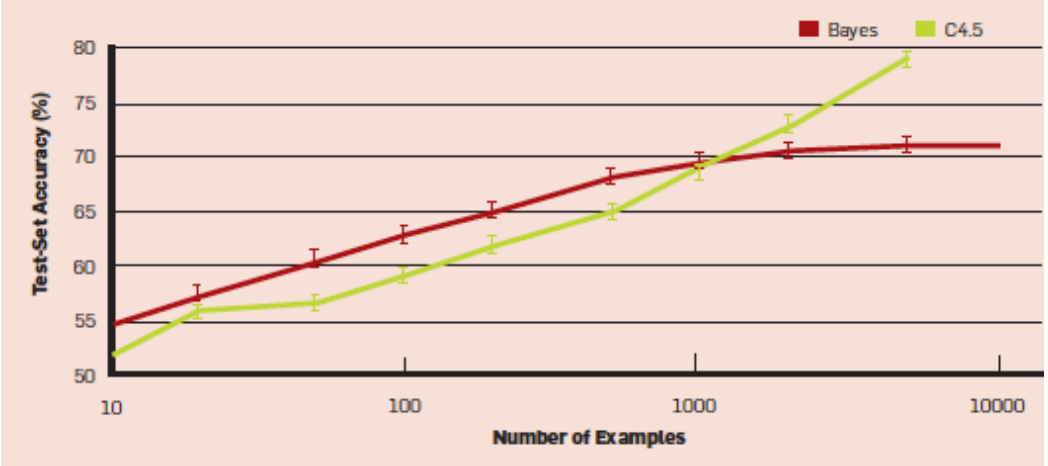
- It's generalization that counts
- Data alone is not enough
- Overfitting has many faces

Machine Learning

What's up?

Overfitting has many faces





Machine Learning

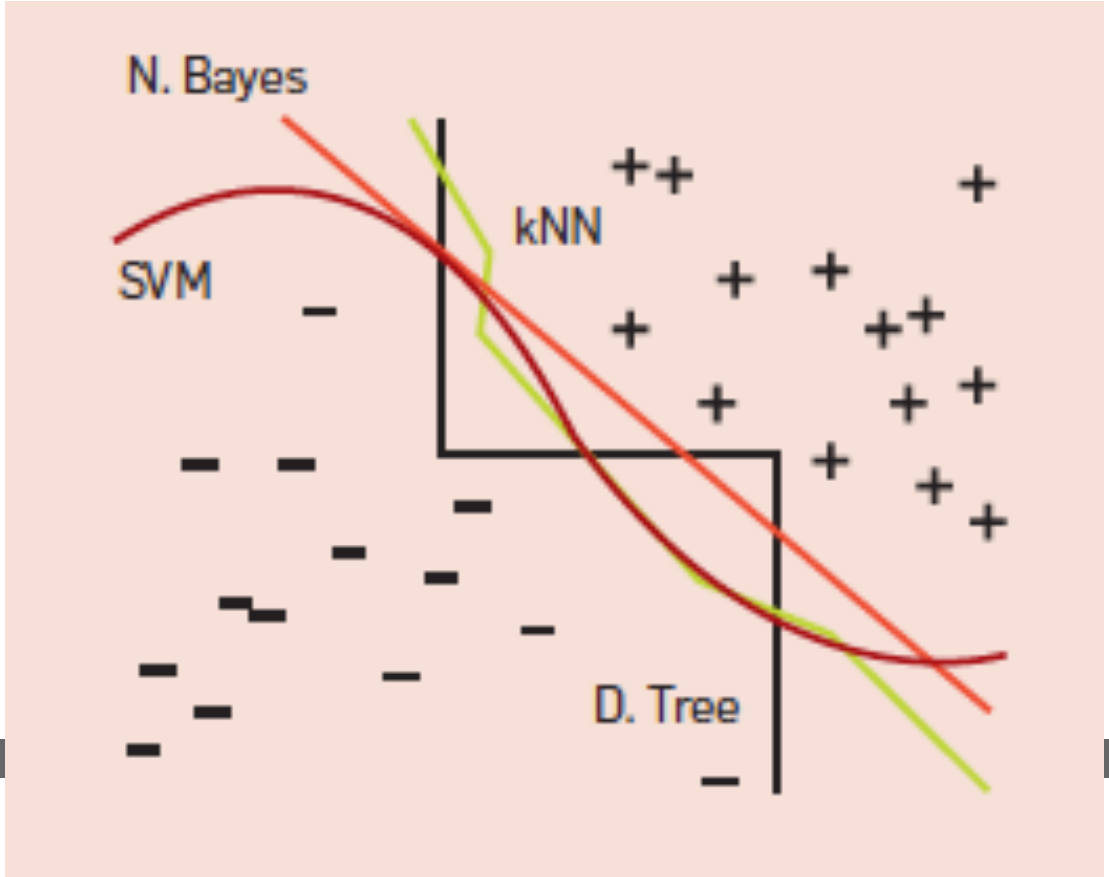
What's up ?

- Intuition fails in high dimensions
- Theoretical guarantees are not what they seem
- Feature engineering is the key
- More data beats a clever algorithm

Machine Learning

What's up ?

- More data beats a clever algorithm



Machine Learning

What's up ?

- Learn many models, not just one
- Simplicity does not imply accuracy
- Representable does not mean learnable
- Correlation does not imply causation

Machine Learning

Where ?

- APPs :
  - Robot control
  - Computer vision
  - Speech recognition, Natural language processing
  - Medical outcomes analysis
  - …
- ML niche is growing :
  - Improved machine learning algorithms
  - Increased data capture, networking, new sensors
  - Software too complex to write by hand
  - Demand for self-customization to user, environment

Machine Learning

Where ?

Pedro DOMINGOS, A Few Useful Things to Know about Machine Learning. *Communications of the ACM*, 55 (10), 78-87, 2012.

Antoine CORNUÉJOLS - Laurent MICLET, "Apprentissage artificiel : Concepts et algorithmes (2ème éd.) », Eyrolles. Juin 2010. 830 pages. ISBN: 978-2-212-12471-2

Resources

- http://www.cs.cmu.edu/~tom/10701_sp11/lectures.shtml

- www.kdnuggets.com
- SIGKDD : www.sigkdd.org
- WEKA : www.cs.waikato.ac.nz/ml/**weka**/
- http://www.videolectures.net

- Nuage de mots
  - Outil : http://www.tagxedo.com/app.html

Un système personnalisé de recommandation à partir de quadri-concepts dans les folksonomies

Jelassi Mohamed Nader, 4ème année

Co-Encadrants : Sadok Ben Yahia (Tunis)

Financement : Bourse Tunisie + Bourse France-Tunisie PHC Utique

Co-Tutelle : Faculté des Sciences de Tunis, Université Tunis El Manar

Problématique : Personnalisation des recommandations dans les folksonomies + Algorithmique de l'analyse formelle de concepts

Applications : Recommandation

Recommandation dans les folksonomies –

Jelassi Mohamed Nader

- ***Verrou principal*** *:*
  - *Qualité (Evaluation)*
- ***Approche envisagée*** *:*
  - *Extraction de quadri-concepts pour personnaliser les recommandations*

- Mohamed Nader Jelassi, Sadok Ben Yahia et Engelbert Mephu Nguifo : Towards more targeted recommendations in folksonomies. *Social Networks Analysis and Mining journal (SNAM*). Springer Ed.

Conception d'un système d'ouverture de droit à des services par apprentissage dynamique du comportement des utilisateurs du système d'information.

DIA Diyé, 4ème année

Co-Encadrants : Olivier Raynaud, Yannick Loiseau, Olivier Coupelon
Financement : CIFRE en collaboration avec Almerys

Problématique : Produire de la confiance numérique dans les services en ligne en reconnaissant les utilisateurs à l'aide des outils de fouille de données.

Applications : sécurité et utilisabilité dans les services en ligne (e-commerce, banque en ligne…)

Confiance numérique– DIA Diyé

- ***Verrou principal*** *:*
  - *Mise en place un modèle de confiance sur la couche applicative*
  - *Sélection de motifs « discriminants »*
- ***Approche envisagée*** *:*
  - *Authentification implicite à base de fouille de motifs fréquents*
  - *Modèle général de confiance*
- *Article en cours :*
  - A learning closed sets based classifier for Implicit User Authentication in Web Browsing *, Diyé DIA, Fabien Labernia, Olivier Raynaud, Yannick Loiseau*, Discrete Applied Mathematics soumis en janvier 2016.

Stockage, indexation et comparaison d'une grande quantité de données génomiques à l'aide d'algorithmes de traitement d'image

DE GOËR DE HERVE Jocelyn (4ème année)

Co-Encadrants : Myoung-Ah Kang
Financement : sur poste de titulaire INRA
Co-Tutelle : INRA UR346 - Épidémiologie Animale

Problématique :
Évaluation d'algorithmes d'indexation et de comparaison d'image numérique pour identifier des séquences ADN au travers d'une base de données de séquences de références

Applications :
Caractérisation de la diversité bactérienne dans le cadre d'études en méta-génomique

Stockage, indexation et comparaison d'une grande quantité de données génomiques à l'aide d'algorithmes de traitement d'image

J.DE GOËR

- ***Verrou principal*** *:*
  - *Réduction des données tout en concevant des propriétés de comparabilité*
- ***Approche envisagée*** *:*
  - *Développement d'une fonction de hachage perceptuel et d'une méthode de comparaison rapide des clés de hachage au sein d'une base de données.*

Extraction de connaissances et incertitudes à partir de mesures effectuées lors de la locomotion en Fauteuil Roulant Manuel

Siyou Fotso Vanel Steve, 2ème année

Co-Encadrants : Philippe Vaslin
Financement : Bourse MENRT

Problématique : explorer et développer de nouveaux modèles d'extraction de motifs séquentiels et temporels, adaptés au contexte d'incertitude de données de la locomotion en FRM.

Applications : Biomécanique.

Extraction de connaissances et incertitudes de séries temporelles – Siyou Fotso V. S.

- ***Verrou principal*** *:*
  - *Données (séries temporelles) incertaines*
- ***Approche envisagée*** *:*
  - *Théorie de l'information et Réduction de dimension univariée*
- *Référence :*

Siyou Fotso VS, Mephu-Nguifo E, Vaslin Ph. Symbolic representation of cyclic time series: application to biomechanics. *Constructive Machine Learning workshop at International Conference on Machine Learning (CML@ICML)*, France, July 2015

Apprentissage multi-instances et Données séquentielles

Manel ZOGHLAMI, 2ème année

Co-Encadrants : Mondher Maddouri

Financement : Bourse Tunisie

Co-Tutelle : FST - Université de Tunis El Manar - Tunisie

Problématique : Classification multi-instances et multicritères des données séquentielles ayant des dépendances entre les instances.

Applications : Bioinformatique, Biologie, Chimie

Apprentissage multi-instances et Données séquentielles

Manel Zoghlami

- ***Verrou principal*** *:*
  - *Classification multi-instances des données séquentielles en prenant en considération les relations entre les instances et sans passer par l'extraction des motifs.*
- ***Approche envisagée*** *:*
  - *Utiliser une mesure de similarité entre les séquences sémantiquement liées.*
- *Référence:*

Aridhi S, Sghaier H, Zoghlami M, Maddouri M, Mephu Nguifo E. (2016) Prediction of ionizing radiation resistance in bacteria using a multiple instance learning model. *Journal of Computational Biology* 23(1):10 -20

Gestion de données manquantes dans les grands entrepôts de données géo référencées

KOUEYA Nestor, 1ère année

Co-Encadrants : Sandro Bimonte (IRSTEA), Libo Ren

Financement : Bourse CPER

Problématique : Quelles sont les données « utiles » pour l'estimation des données manquantes ? Comment exploiter la variété de données pour l'estimation des données manquantes ? Comment définir des méthodes d'estimation efficaces en temps de calcul ?

Applications : Agriculture, Gestion de trafic urbain, Marketing / Publicité, Science climatique, etc.

BIG SOLAP et Données manquantes

KOUEYA Nestor

- ***Verrou principal*** *:*
  - *Données spatiales, volumineuses, variées, et manquantes*
- ***Approche envisagée*** *:*
  - Explorer les méthodes pertinentes en qualité d'estimation et efficace pour l'exécution dans un environnement distribué

- *Référence:*

KOUEYA, N., BIMONTE, S., MEPHU NGUIFO, EDA - 2014. Une nouvelle approche d'estimation pour les entrepôts de données multi-granulaires incomplètes. *10ème journées francophone sur les Entrepôts de Données et l'Analyse en ligne EDA*, Vichy, p. 129-144, 16 pages

Etude de la biosphère rare microbienne par une approche in-silico :

Méthode de classification ensembliste et Modélisation

BAZIN Alexandre, Post-doctorant

Co-Encadrants : Didier Debroas (LMGE)

Financement : Bourse PostDoc CPER Oct 2015 - Mars 2017

Problématique : Améliorer les méthodes de clustering sur de gros volumes de données biologiques

Applications : Biologie - Bioinformatique

Méthodes de classification – Alexandre Bazin

- ***Verrou principal*** *:*
  - *Les gros volumes de données empêchent l'utilisation de méthodes de clustering de qualité*

- ***Approche envisagée*** *:*
  - *Utiliser les ensembles flous pour représenter les clusters et leur voisinage*

Modélisation de la biodiversité à partir des études paléoécologiques

LONLAC Jerry, Post-doctorant

Co-Encadrants : Yannick Miras (Geolab), M. Pailloux, A. Wagler

Financement : Bourse PostDoc CPER Nov 2015 – Avril 2017

Problématique : Quels sont des groupements écologiques fonctionnels indicateurs de l'évolution temporelle de la biodiversité ?
Comment décrire le dynamique écologique à travers les modèles mathématiques construits sur des données environnementales passées et actuelles ?

Applications : Paléo-écologie

Modélisation de la biodiversité – Jerry Lonlac

- ***Verrou principal*** *:*
  - *Données hétérogènes et évolutives*

- ***Approche envisagée*** *:*
  - *Règles d'association graduelles*
  - *Réseaux de pétri pour modélisation*