Introduction
Smart Analytics for Big Data

I. Joly

30/09/2020

Plan and Organisation of the lesson

Today

1. Introduction (45min)
2. Reproducible research (45min)
   - ▶ *in-class*: introduction & demo
   - ▶ *out-of-class*: reading
3. Case Study (1h30min)
   - ▶ Team work on Data Visualizations
   - ▶ Discussion

Introduction

Context

- Industry 4.0, Digital Factory, Internet of Things, Digital Economy

- Assessing the relevance of data and selecting the **right data for business decisions** is a key strategic capability.

- Analysis of complex and **big data**, **temporal** and **spatial** data
  - needs **specific skills** to search and to extract the relevant information
  - to **analyze** them accordingly with their specific dimensions.

Some definitions

Many notions

For examples *"AT&T business"* offers
Data science solutions for IoT including Artificial Intelligence and machine learning

Big Data

- most cited definition of big data includes the 3Vs (**Volume, Variety, and Velocity**) *Laney (2001)*
- big data should include '**Value**' *(Gantz et al., 2011)*
- big data should also have '**Veracity**' *(Zikopoulos et al., 2013)*

Data Sciences

- Data Manager
- Data Miner
- Data Analyst
- Data Scientist

Tools

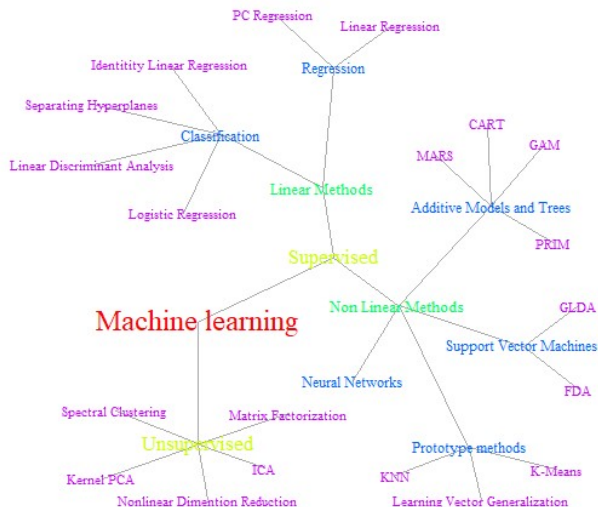Data Science, Machine Learning, Statistics, AI, etc

More than 30 models and families

- Generalized Additive Model
- Generalized Linear Model
- Generalized Non-Linear Model
- Multinomial Logistic Regression
- Multinomial Probit
- Linear Model
- Indicator (Identity) Linear Model
- Logistic Regression
- Probability (Unit) Regression
- Cumulative Logistic Regression
- Cumulative Probability (Unit) Regression
- Complementary Log-Log Model
- Continuation-Ratio Logistic Regression
- Hierarchical Logistic Regression
- Poisson Log-Log Regression
- Negative Binomial Regression

- Conditional Regression Model
- Mixed Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Support Vector Machines
- Order Support Vector Machines
- Support Vector Ordinal Regression
- K Nearest Neighbours
- Classification And Regression Trees
- Separating Hyperplanes
- Neural Networks
- Feed-Forward Neural Networks
- Bayesian Neural Networks
- Markov Chains
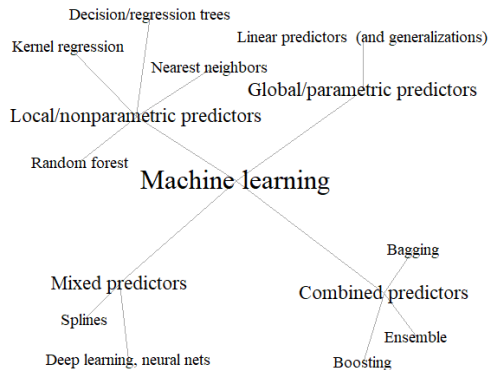- Markov Chains Monte Carlo
- Deep Neural Networks

Several criterias differenciating the methods

- ▶ Supervised
- ▶ Unsupervised

- ▶ Additive
  - ▶ Linear
  - ▶ Non-linear
- ▶ Generative
- ▶ Discriminative
- ▶ Parametric
- ▶ Semi-parametric

- ▶ Non-parametric
- ▶ Classification
- ▶ Regression
- ▶ Binary data
- ▶ Multinomial data
- ▶ Ordinal data
- ▶ Count data
- ▶ Continuous data
- ▶ Explicative
- ▶ Data driven

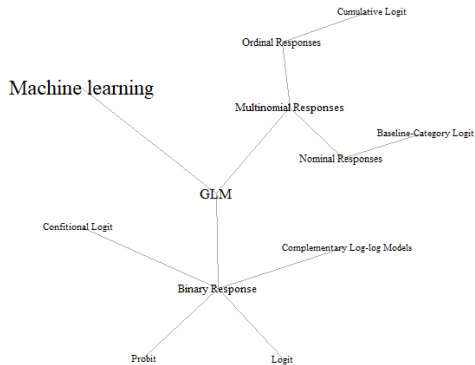# Hastie, Tibshirani, and Friedman (2009) *reference*

Mullainathan and Spiess (2017) *reference*

Agresti (2007) *reference*

Machine learning

GLM

Cumulative Logit

Ordinal Responses

Multinomial Responses

Baseline-Category Logit

Nominal Responses

Confitional Logit

Complementary Log-log Models

Binary Response

Probit

Logit

Artificial Intelligence (AI)
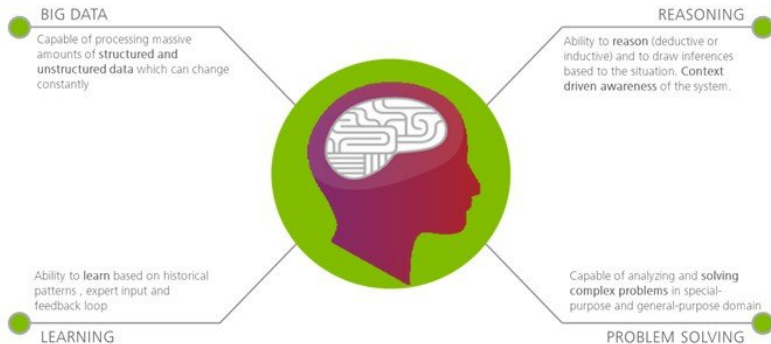
Figure 1: AI schema

Industrial issues

Smart Analytics in Industry 4.0

Lee, Kao, and Yang (2014) identify that

▶ self-learning machines are still far from implementation in current industries

▶ advances are expected in 5 distinct categories:

1. *Manager and Operator Interaction*: machine control and schedule design have to include machines health

2. *Machine fleet*: prognostic and health management methods have to consider the fleet of machines

1. *Product and Process Quality*: product quality informs on the process quality
2. *Big Data and Cloud*: Data management and distribution in Big Data environment
3. *Sensor and Controller Network*: decision-making algorithms depend on wrong and inaccurate readings

Smart Machine Maintenance
- ▶ Machine health awareness analytics with self-learning knowledge base
- ▶ Decision support analytics for self-maintenance

Servitization
- ▶ to shift from selling products, to selling an integrated product and service offering that delivers value in use (Martinez et al. (2010))

Product-Service System (PSS)

▶ system of products, services, supporting networks, and infrastructure that is designed to be competitive, satisfy customers' needs, and have a lower environmental impact than traditional business models (Mont (2004))

▶ Market goal of manufacturers:
  ▶ is not one-time product selling, but
  ▶ continuous profit from customers by total service solution, which can satisfy unmet customers' needs.

Industrial big data environment

Industry 4.0 with *Smart Machine Maintenance*, *servitization* and *PSS* imply:

▶ different units of observation, analyse and decision
  ▶ from 'human-related data' to 'machine generated data' (machine, controllers, sensors, manufacturing systems, etc.)
  ▶ sales prediction, user relationship mining and clustering, recommendation systems, opinion mining, etc.

- ▶ SI networks. From the sensor to the dashboard and decision
  - ▶ compatibility and standard issue
  - ▶ vibration, pressure, etc. are added to historical data
  - ▶ this aggregation is called "Big Data"
- ▶ not a one shot process: dynamic workflow
  - ▶ integrated platform, predictive analytics, and visualization tools

Conclusion

Challenges are to switch:

- ▶ from Data Analytics
- ▶ to DS projects, including
  - ▶ BD dimension
  - ▶ DM
  - ▶ DA
  - ▶ OR
  - *DS and OR : the new challenges*

Class organisation

Goals

Be DM, DA and DS supervisor in the Big Data context

Teachers and Industrial Contributors

GI and Ensimag:

- ▶ *Christophe Bobineau*, MCF, Grenoble INP ENSIMAG
- ▶ *Iragaël Joly*,[1] MCF HDR, Grenoble INP Génie industriel
- ▶ *Pierre Lemaire*, MCF, Grenoble INP Génie industriel
- ▶ *Genoveva Vargas Solar*, DR, CNRS, LIG, HADAS Group

Invited Teachers:

- ▶ *Bruno Agard*, PR, Laboratoire en Intelligence des Données (LID), Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal

---

[1]corresponding teacher: iragael.joly@grenoble-inp.fr

Organisation of the course

▶ Data supply-chain: from data collection and production, storage and organization, management, exploitation and analysis, and communication.

▶ Big data and dynamic process of analysis needs transparent, repeatable and reproducible technics.

▶ Backward presentation: from needs to solutions

Three $+1$ parts

1. **Big-Data Management**
2) **Exploration of complex data with high dimensionality**
3) **Analysis of complex data with temporal and / or spatial dimensions**
4) **Visualization and communication**

Planning

Date	Semaine	Horaire	Intervenants	Thème	Horaire	Intervenants	Thème	Horaire	Intervenants	Thème
01/10/2020	S41	08h00			9h30	I. Joly	Intro	11h00	I. Joly	Intro - appli (Rmd)
08/10/2020	S42	08h00	I. Joly	Vizu + GIS	9h30	I. Joly	Vizu + GIS	11h00	I. Joly	Vizu + GIS - appli
15/10/2020	S43	08h00			9h30	C.Bobineau	Intro : Distributed data management and bases	11h00	C.Bobineau	Intro : Distributed data management
22/10/2020	S44	08h00	G.Vargas	Map-Reduce	9h30	G.Vargas	NoSQL/MongoDB	11h00	C.Bobineau	HO1: Getting acquainted with Mongo
29/10/2020	S45	TOUSSAINT								
05/11/2020	S46	08h00	C.Bobineau	HO1: Getting acquainted wi	9h30	G.Vargas	HO2 : Sharding with Mongo	11h00	G.Vargas	Project
12/11/2020	S47	08h00		Projet Introduction ?	9h30	G.Vargas	Data Analysis	11h00	G.Vargas	Data processing & Analytics
19/11/2020	S48	13h30	B.Agard		15h00	B.Agard		16h30	B.Agard	
26/11/2020	S49	13h30	B.Agard		15h00	B.Agard		16h30	B.Agard	
03/12/2020	S50	13h30	B.Agard		15h00	B.Agard		16h30	B.Agard	
10/12/2020	S51	08h00	I. Joly	TP Projet Mobilité	9h30	I. Joly	TP Projet Mobilité	11h00	I. Joly	TP Projet Mobilité
	S52	NOEL								
		NOEL								
07-janv	S01	08h00	I. Joly	Categorical Data Analysis	9h30	I. Joly	Categorical Data Analysis	11h00	I. Joly	Categorical Data Analysis
14-janv	S02	08h00	I. Joly	Survival analysis	9h30	I. Joly	Survival analysis	11h00	I. Joly	Survival analysis
21-janv	S03	08h00			9h30		Exam	11h00		Exam

Figure 2: Planning 2020 - Version sept

Tools

- RStudio = R + Rmarkdown
- Other tools for DM: MongoDB, . . .

Evaluation

- ▶ Individual evaluation, e.g. in-class work (TP), multiple choice questions or closed-formed quizzes and **exam**
- ▶ Application Project realized in group: **Mobility project**
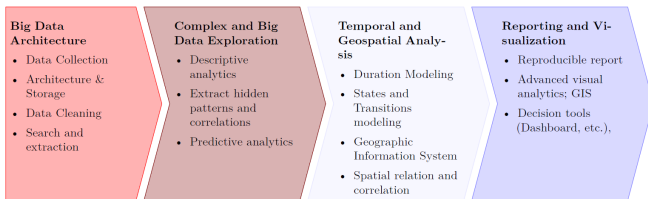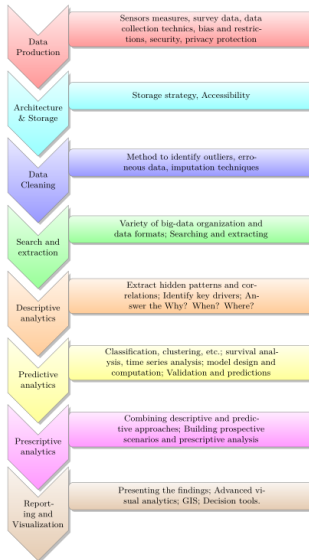
Conclusion

DS Workflow

From the class

Figure 3: Data processing and analytics

To

Data Production	Sensors measures, survey data, data collection technics, bias and restrictions, security, privacy protection
Architecture & Storage	Storage strategy, Accessibility
Data Cleaning	Method to identify outliers, erroneous data, imputation techniques
Search and extraction	Variety of big-data organization and data formats; Searching and extracting
Descriptive analytics	Extract hidden patterns and correlations; Identify key drivers; Answer the Why? When? Where?
Predictive analytics	Classification, clustering, etc.; survival analysis, time series analysis; model design and computation; Validation and predictions
Prescriptive analytics	Combining descriptive and predictive approaches; Building prospective scenarios and prescriptive analysis
Reporting and Visualization	Presenting the findings; Advanced visual analytics; GIS; Decision tools.

References

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis, Second Edition*.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Lee, Jay, Hung-An Kao, and Shanhu Yang. 2014. "Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment." *Procedia CIRP* 16: 3–8. https://doi.org/https://doi.org/10.1016/j.procir.2014.02.001.

Martinez, V., M. Bastl, J. Kingston, and S. Evans. 2010. "Challenges in Transforming Manufacturing Organisations into Product-Service Providers." *Journal of Manufacturing Technology Management* 21: 449–69. https://doi.org/https://doi.org/10.1108/17410381011046571.

Mont, Oksana. 2004. "Product-Service Systems: Panacea or Myth?" PhD thesis, Lund University: IIIEE, Lund University.