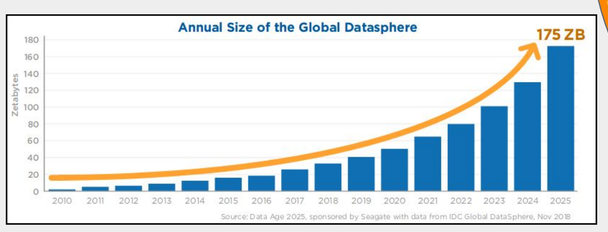
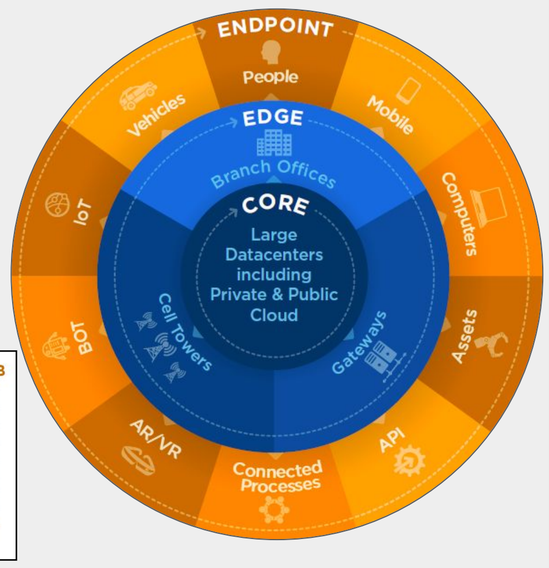
Scalable Machine Learning

Braulio C. Blanco, Eugen R. Patrascu

I. Introduction

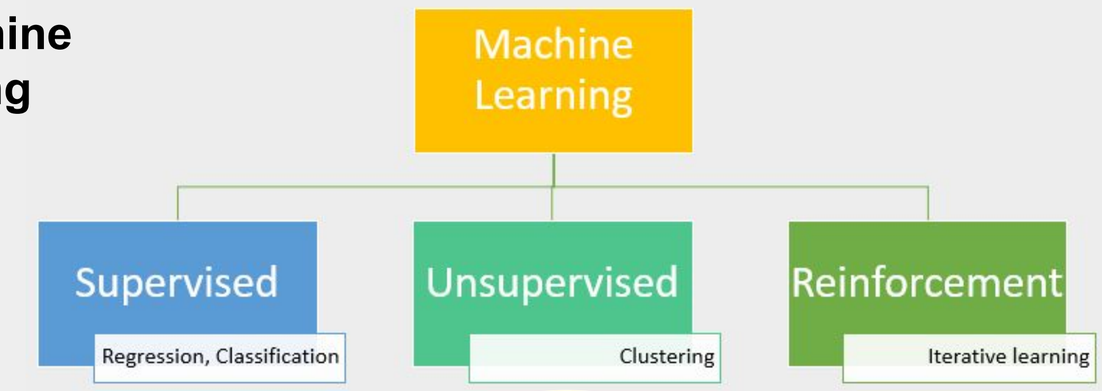
The global datasphere is growing at an exponential rate. These huge amounts of data are stored and analysed in many industries, such as healthcare, finance, marketing, insurance.



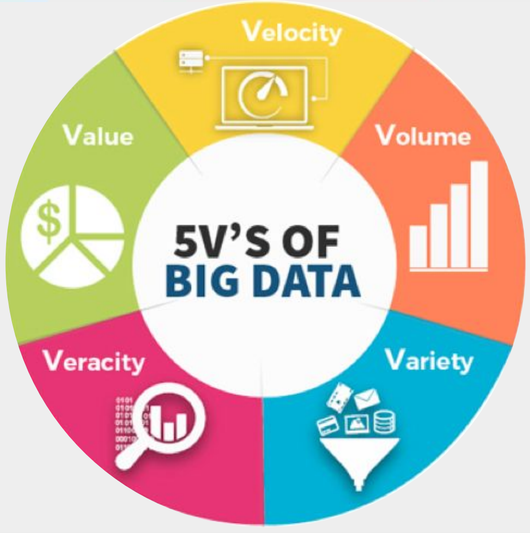


Machine Learning provides powerful algorithms used to uncover patterns in the increasingly large amounts of data and to provide useful insights. However, with the rise of Big Data, the traditional way of performing Machine Learning has become insufficient to respond to the new challenges of data volume, variety, velocity, veracity and value.

II. Machine Learning

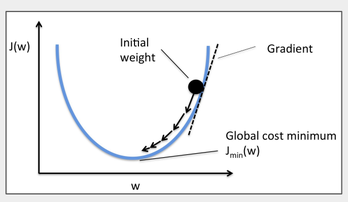


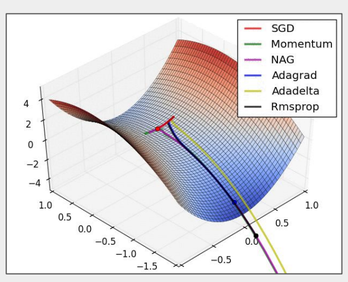
III. Big data

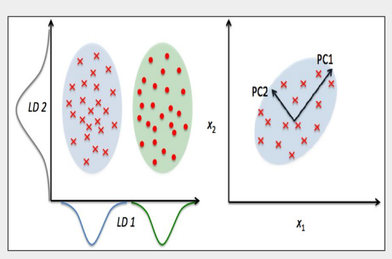


IV. Machine Learning applied to big data

NON- PARALLEL		PARALLEL		
Optimization	Data Reduction	Data Parallelism	Model / parameters Parallelism	Hybrid
The state-of-the-art large scale constrained optimization algorithms use a variant of Stochastic Gradient Boosting. They are particularly important in Artificial Neural Networks and Deep Learning.	Reducing large datasets horizontally or vertically. The strategies include: instance selection, feature selection and feature extraction.	Using existing big data architecture, partitioning input data vertically, horizontally, or even arbitrarily into manageable pieces, and then computing on all subsets simultaneously.	Creating parallelized versions of ML algorithms by first dividing the learning model/parameters and then computing on each structural block concurrently.	Hybrid approaches combine model and data parallelism by partitioning both data and model variables simultaneously.







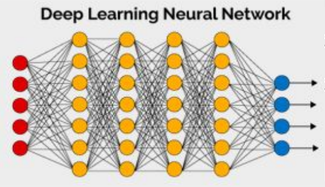
Using popular distributed ML libraries

Apache MLlib and Apache Mahout consist of fast and scalable implementations of standard learning algorithms like classification, regression, clustering, dimension reduction. MLlib is more used in practice, as it can be up to 100 times faster.

Custom Solutions

Traditional algorithms have custom parallelized versions that achieve high performances.

Deep Learning

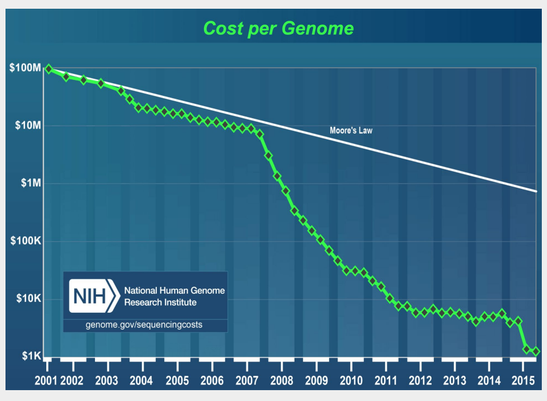


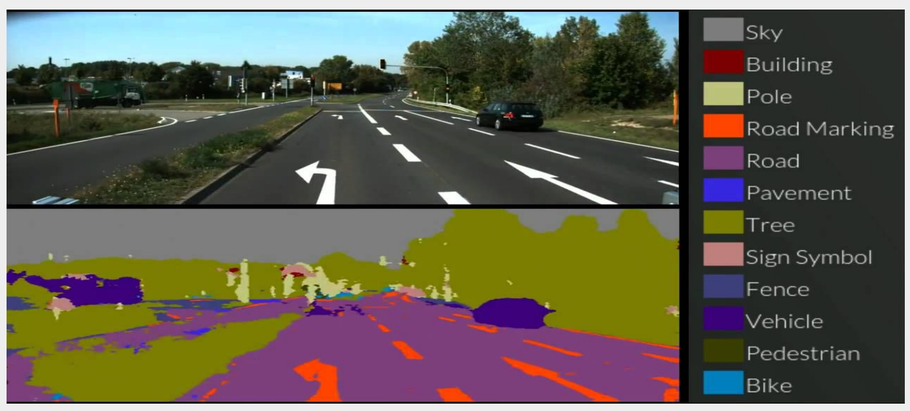
Deep neural networks achieve state-of-the-art result in many fields such as computer vision, natural language processing or speech recognition. GPU computing plays an important role in training deep architectures.

V. Applications

While general ML libraries such as Spark MLlib offer good results, there are libraries/algorithms that are written for specific use cases like bioinformatics, which can lead to significant improvements. In the field of genomics this includes: ADAM, Variant Spark and SEQSpark.

Machine Learning has been applied to big data in many scientific and business cases.





VI. Conclusions and open issues

Most of the work and research into Scalable Machine Learning has addressed the Volume, Variety and Velocity parts of big data, while Veracity and Value have not seen much attention.

Veracity

In order to deal with Veracity, algorithms could be developed that detect the trustworthiness of data or data sources, and are able to filter out untrustworthy data in the preprocessing phase.

Value

An important research direction could be to develop explainable models. Additionally, existing evaluation strategies for ML algorithms can be improved so that they do not include only the prediction accuracy, but also other metrics regarding how well they support the end users in their tasks.

REFERENCES:

Zhou, L. et al. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing

Guo, R. et al, (2018), Bioinformatics applications on Apache Spark, GigaScience Oxford.

Reinsel D. et al, (2018). The Digitization of the World From Edge to Core. International Data Corporation.

Dimitrovska I. et al, (2017). Creating a Business Value while Transforming Data Assets using Machine Learning.

Lopes, N., & Ribeiro, B. (2017). Novel Trends in Scaling Up Machine Learning Algorithms.

Qiu, J. et al. (2016). A survey of machine learning for big data processing.