

MACHINE LEARNING – AN INTRODUCTION

JOSEFIN ROSÉN, SENIOR ANALYTICAL EXPERT, SAS INSTITUTE

JOSEFIN.ROSEN@SAS.COM
TWITTER: @ROSENJOSEFIN



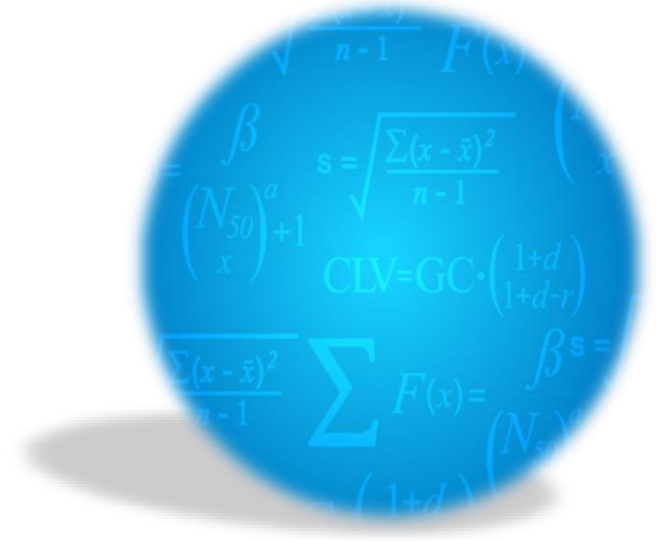
- What is machine learning?
- When, where and how is machine learning used?
- Exemple – deep learning
- Machine learning in SAS

Wikipedia: Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.

SAS: Machine learning is a branch of artificial intelligence that automates the building of systems that learn from data, identify patterns, and make decisions – with minimal human intervention.

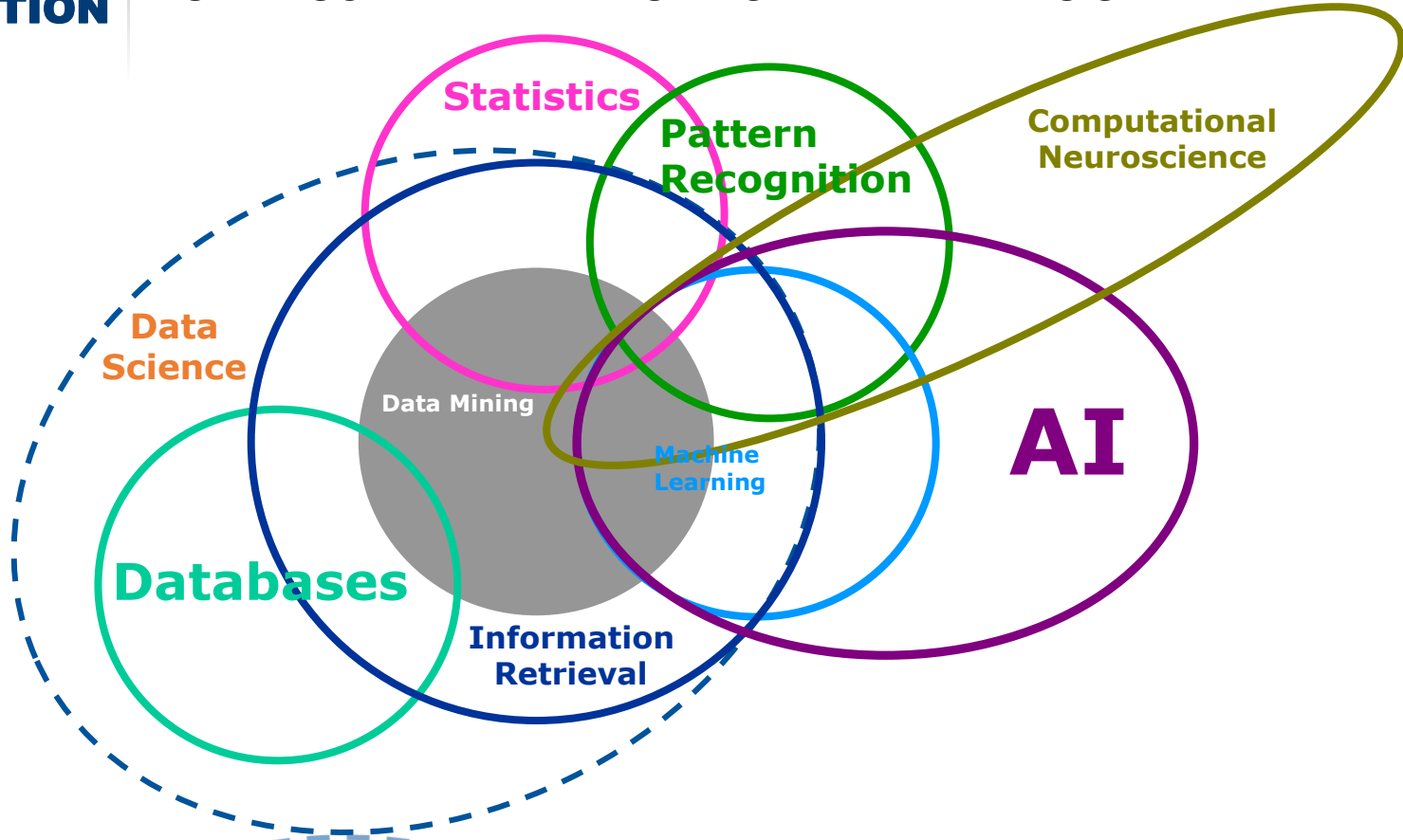


*” Complicated methods,
consumable results”*



MACHINE LEARNING - AN INTRODUCTION

MULTIDISCIPLINARY NATURE OF DATA ANALYSIS



When the predictive accuracy of a model is more important than the interpretability of a model.

When traditional approaches are inappropriate, e.g. when you have:

- more variables than observations
- many correlated variables
- unstructured data
- fundamentally nonlinear or unusual phenomena

A few examples:

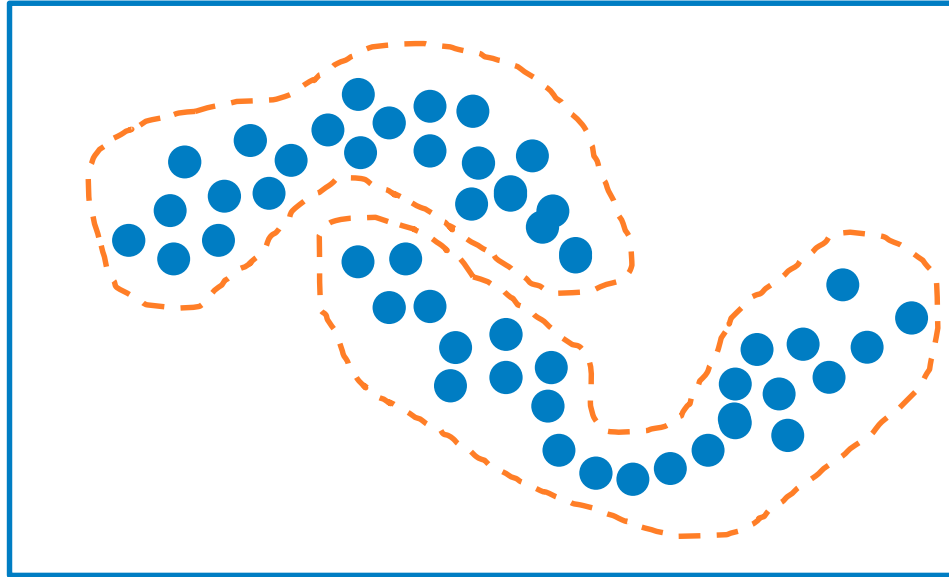
- Recommendation applications
- Fraud detection
- Predictive maintenance
- Text analytics
- Self driving cars



INTRODUCTION TO MACHINE LEARNING

UNSUPERVISED LEARNING

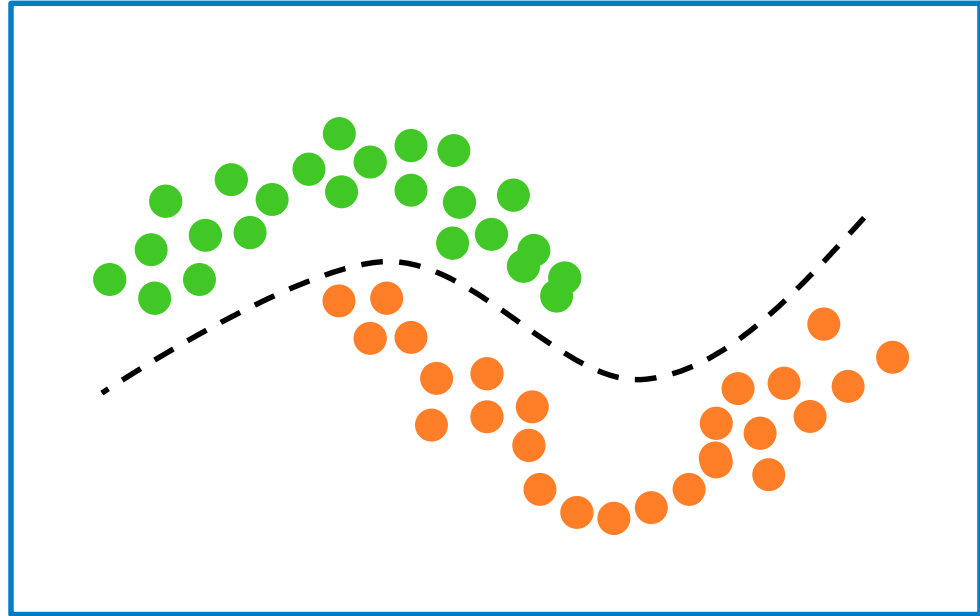
- Operated on unlabeled examples;
- Clustering, dimension reduction;



INTRODUCTION TO MACHINE LEARNING

SUPERVISED LEARNING

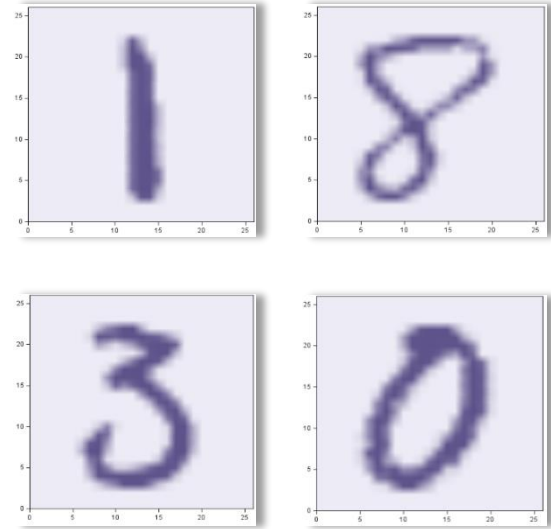
- Trained on labeled examples;
- Classification, regression, prediction;



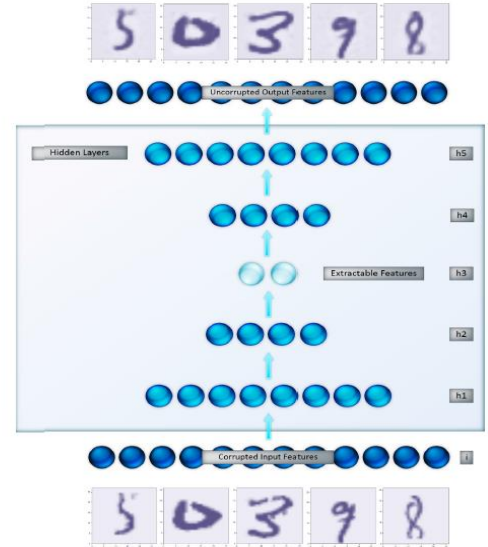


- Neural networks with more than two hidden layers
- Promising breakthroughs in speech, text and image recognition
- Useful for feature extraction

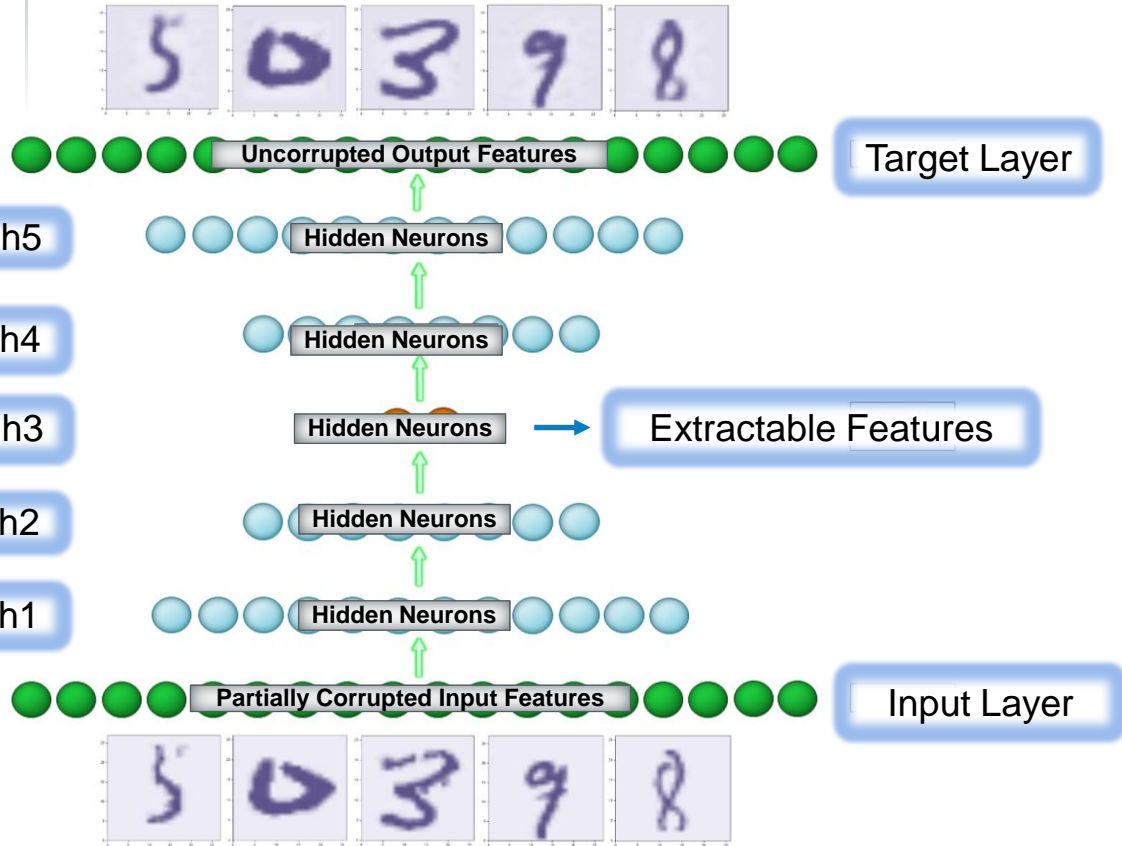
- 784 variables form a 28x28 digital grid
 - 784-dimensional input vector $X = (x_1, \dots, x_{784})$
- Pixel intensity from 0 to 255
- 60,000 training pictures with label
- 10,000 test pictures without label



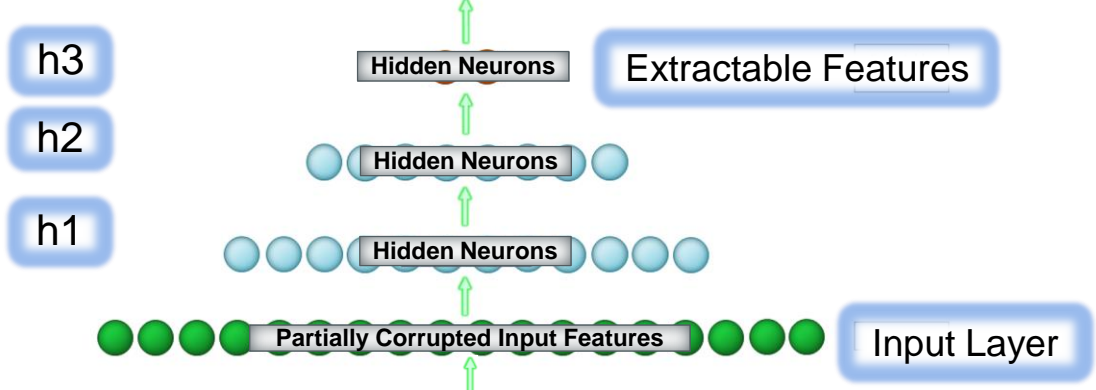
- Train a stacked denoising autoencoder
- Extract representative features from MNIST data
- Compare with principal component analysis, two PCs



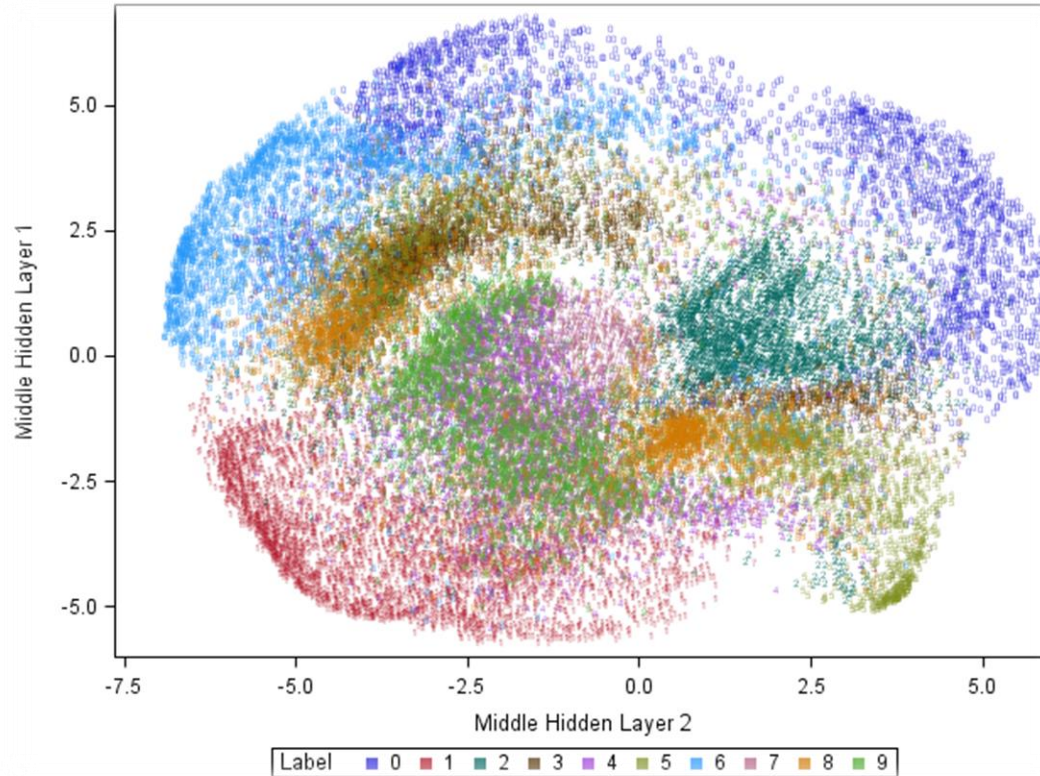
STACKED DENOISING AUTOENCODER

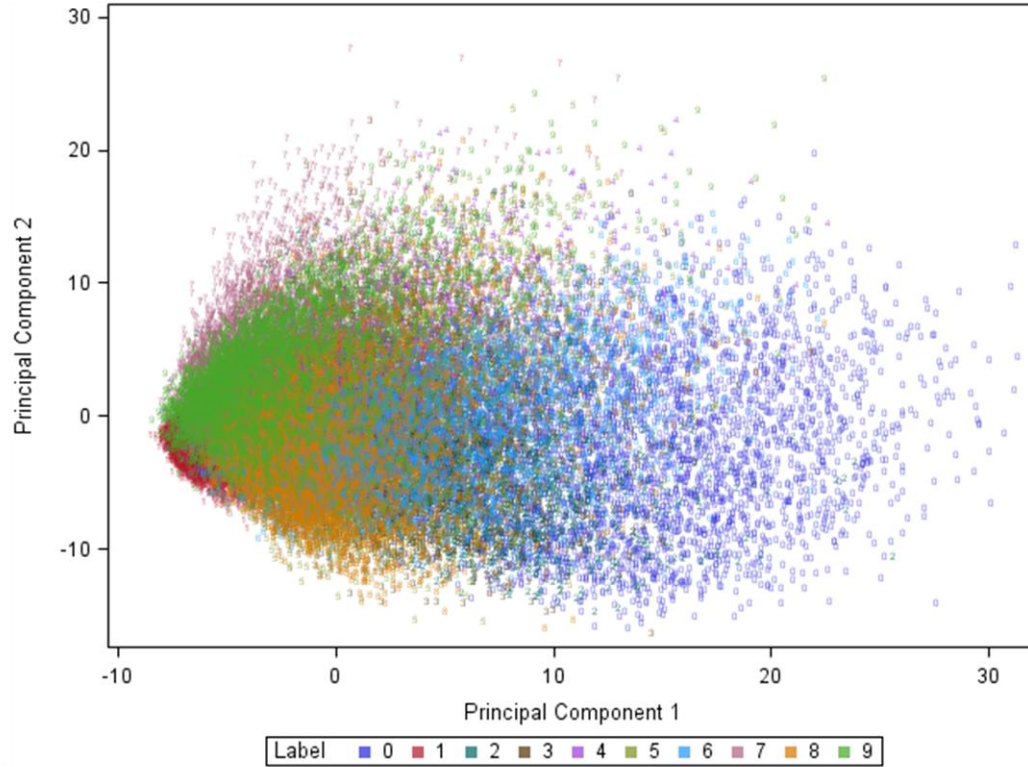


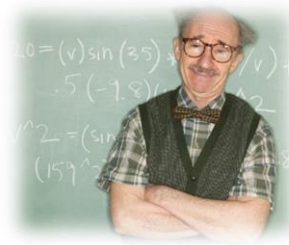
Record ID	Hidden Unit 1	Hidden Unit 2
1	0.98754	0.32453
2	0.76854	0.87345
3	0.87435	0.05464
⋮	⋮	⋮



Record ID	Pixel 1	Pixel 2	Pixel 3	Pixel 4	Pixel 5	Pixel 6	Pixel 7	Pixel 8	Pixel 9	Pixel 10	...
1	0	0	0	0	0	5	8	11	6	3	...
2	0	0	0	0	10	20	45	46	36	24	...
3	0	25	37	32	40	64	107	200	67	46	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	↘







- Neural networks
- Decision trees
- Random forests
- Associations and sequence discovery
- Gradient boosting and bagging
- Support vector machines
- Nearest-neighbor mapping
- K-means clustering
- DBSCAN
- Self-organizing maps
- Local search optimization techniques such as genetic algorithms
- Expectation maximization
- Multivariate adaptive regression splines
- Bayesian networks
- Kernel density estimation
- Principal components analysis
- Singular value decomposition
- Gaussian mixture models
- Sequential covering rule building
- Model ensembles
- Recommendations

- SAS Enterprise Miner
- SAS Text Miner
- SAS In-Memory Statistics for Hadoop
- SAS Visual Statistics
- SAS/STAT
- SAS/OR
- SAS Factory Miner

SUPERVISED LEARNING

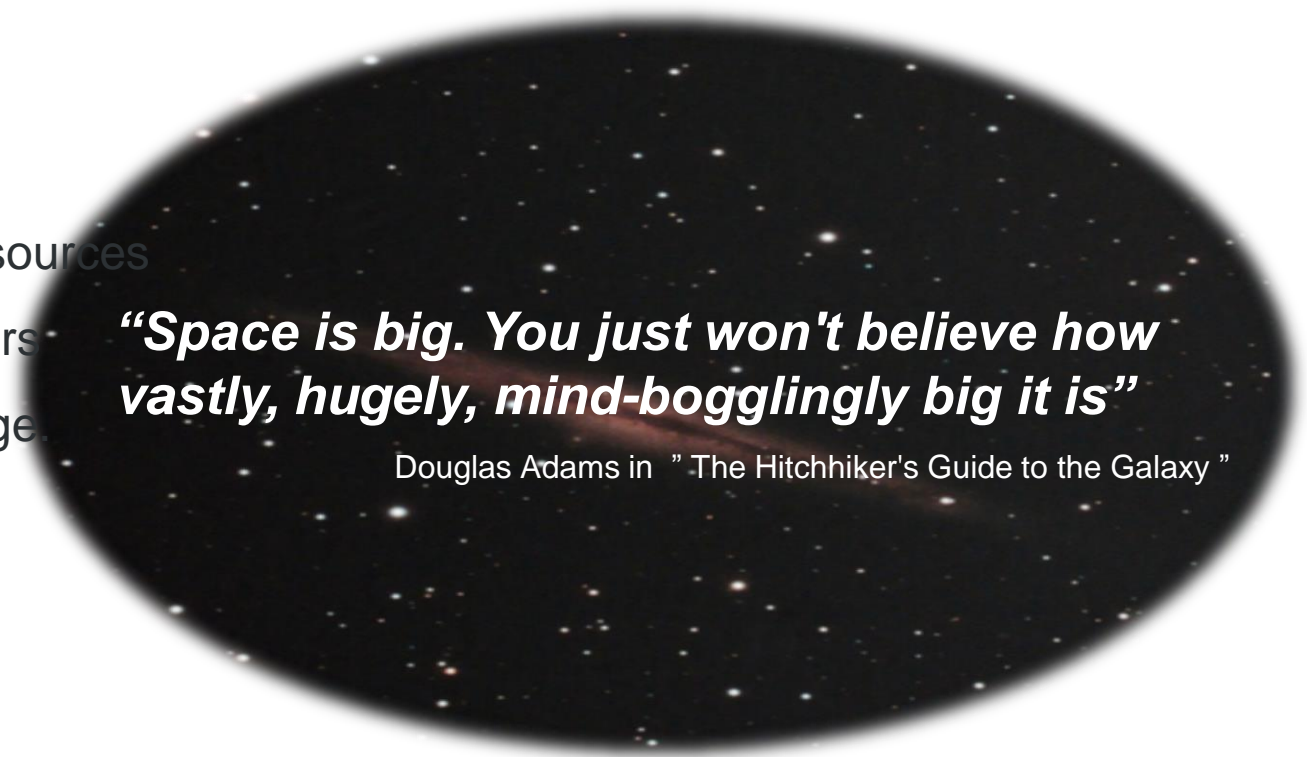
Algorithm	SAS EM-noder	SAS procedurer
Regression	High Performance Regression LARS Partial Least Squares Regression	ADAPTIVEREG GAM GENMOD GLMSELECT HPGENSELECT HPLOGISTIC HPQUANTSELECT HPREG LOGISTIC QUANTREG QUANTSELECT REG
Decision tree	Decision Tree High Performance Tree	ARBORETUM HPSPLIT
Random forest	High Performance Tree	HPFOREST
Gradient boosting	Gradient Boosting	ARBORETUM
Neural networks	AutoNeural DMNeural High Performance Neural Neural Network	HPNEURAL NEURAL
Support vector machine	High Performance Support Vector Machine	HPSVM
Naïve Bayes		HPBNET*
Neighbors	Memory Based Reasoning	DISCRIM

*PROC HPBNET can learn different network structures (naïve, TAN, PC, and MB) and automatically select the best model

Algoritm	SAS EM-noder	SAS procedurer
A priori rules	Association Link Analysis	
K-means klustering	Cluster High Performance Cluster	FASTCLUS HPCLUS
Spektral klustering		Custom lösning genom Base SAS och procedurerna DISTANCE och PRINCOMP
Kernel density estimation		KDE
Kernel PCA		Custom lösning genom Base SAS och procedurerna CORR, PRINCOMP och SCORE
Singular value decomposition		HPTMINE IML
Self organizing maps	SOM/Kohonen	

Algorithm	SAS EM-noder	SAS procedurer
Denoising autoencoders		HPNEURAL NEURAL

- Big data
- Computational resources
- Powerful computers
- Cheap data storage



***“Space is big. You just won't believe how
vastly, hugely, mind-bogglingly big it is”***

Douglas Adams in "The Hitchhiker's Guide to the Galaxy"



More reading

- White papers
 - http://www.sas.com/en_us/whitepapers/machine-learning-with-sas-enterprise-miner-107521.html
 - <http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf>
- SAS links
 - http://www.sas.com/en_us/insights/analytics/machine-learning.html
 - http://www.sas.com/en_us/insights/articles/analytics/introduction-to-machine-learning-five-things-the-quants-wish-we-knew.html
- SAS Data Mining Community
 - https://communities.sas.com/community/support-communities/sas_data_mining_and_text_mining/
- Big Data Matters Webinar Series:
 - www.sas.com/bigdatamatters