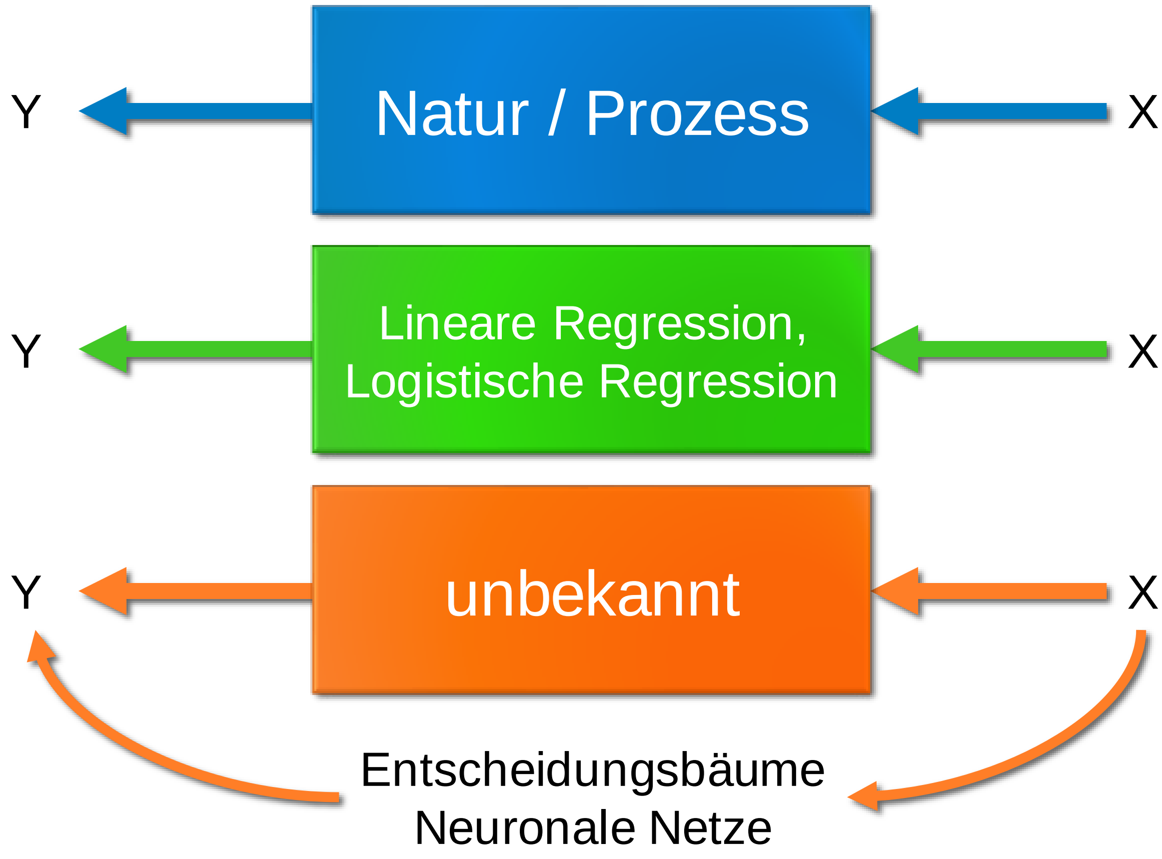
THE POWER TO KNOW.

sas® club

Der Business Analytics Club für SAS User

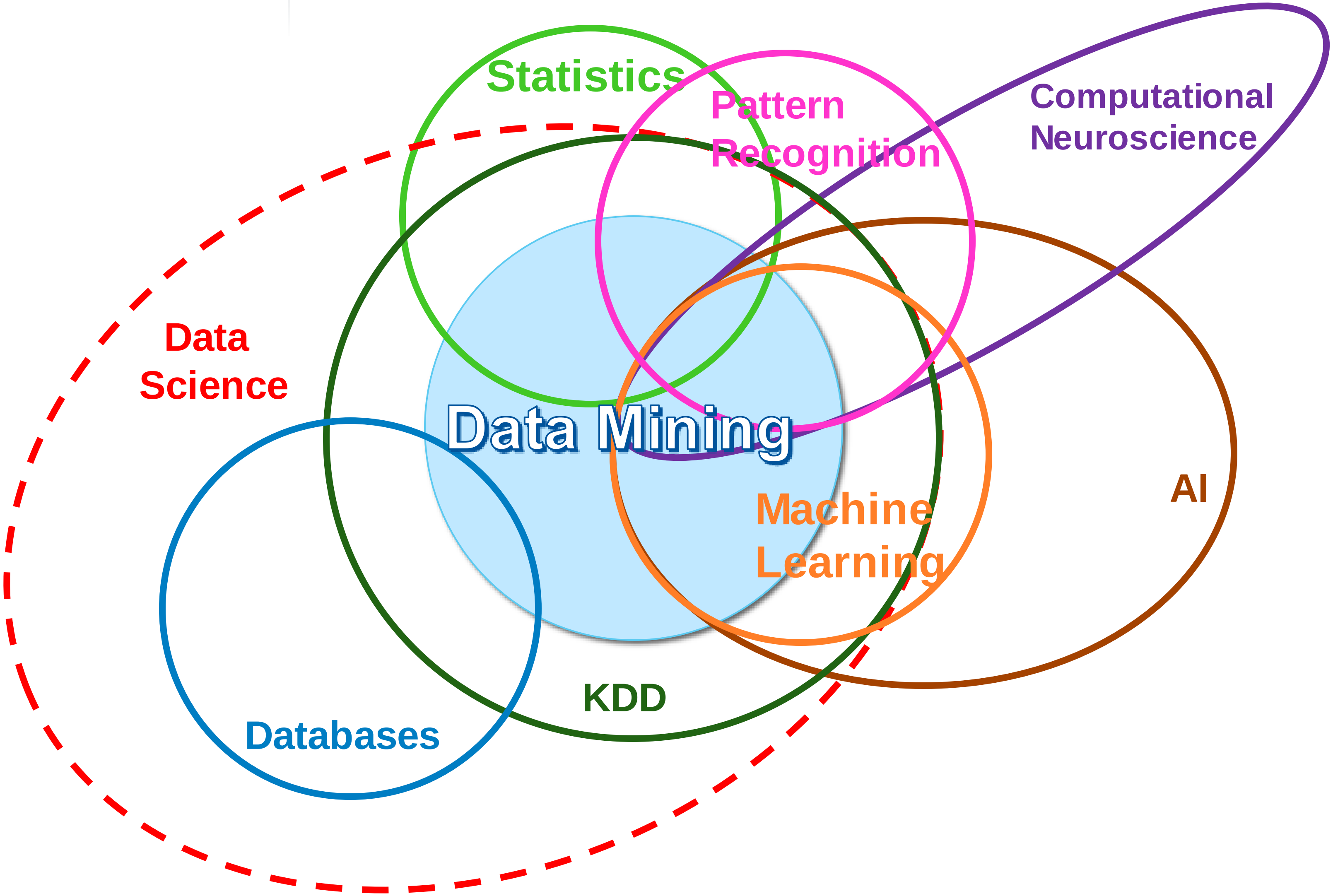
"MACHINE LEARNING" IN DER SAS ANALYTIK PLATTFORM

Mihai Paunescu
Gerhard Svolba



Vgl. Breimann (2001)

MACHINE LEARNING | EINORDNUNG DER ANSÄTZE



Machine Learning

Supervised Learning

Unsupervised Learning

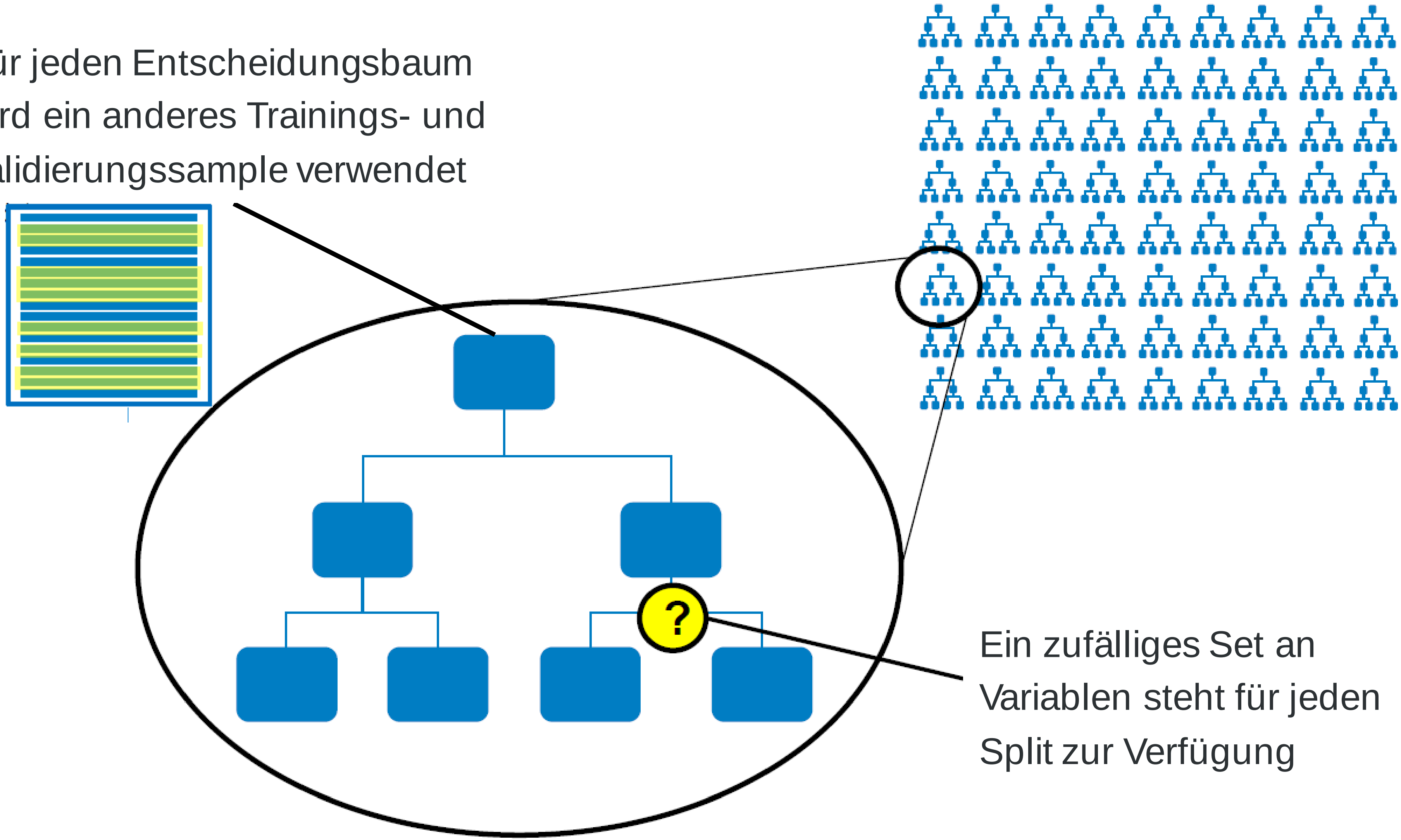
Semisupervised Learning

Data Mining

ALGORITHMEN IN SAS (1)

Algorithmus	SAS Enterprise Miner	SAS Prozeduren	Literatur
Regression	High Performance Regression LARS Partial Least Squares	ADAPTIVEREG GAM HPGENSELECT HPLOGISTIC HPREG HPQUANTSELECT	Panik 2009
Decision Tree	High Performance Tree	ARBOR HPSPLIT	De Ville and Neville 2013
Gradient Boosting	Gradient Boosting	ARBOR	Friedman 2001
Neural network	AutoNeural High Performance Neural Neural Network	HPNEURAL NEURAL	Rumelhart, Hinton, and Williams 1986
Random Forest	High Perfromance Forest	HPFOREST	Breiman 2001b

Für jeden Entscheidungsbaum wird ein anderes Trainings- und Validierungssample verwendet



Ein zufälliges Set an Variablen steht für jeden Split zur Verfügung

ALGORITHMEN IN SAS (2)

Algorithmus	SAS Enterprise Miner	SAS Prozeduren	Literatur
Support Vector Machines	High Performance Support Vector Machine	HPSVM	Cortes and Vapnik 1995
Naive Bayes	Bayesian Network	HPBNET	Friedman, Geiger, and Goldszmidt 1997
Neighbors	Memory Based Reasoning	DISCRIM	Cover and Hart 1967
Gaussian processes			Seeger 2004

Machine Learning

Data Mining

Supervised Learning

Unsupervised Learning

Semisupervised Learning

ALGORITHMEN IN SAS (1)

Algorithmus	SAS Enterprise Miner	SAS Prozeduren	Literatur
A priori rules	Association Link Analysis	Assoc, Rulegen, Taxonomy	Agrawal, Imieliński, and Swami 1993
K-means clustering	Cluster High Performance Cluster	FastClus HPClus	Hartigan and Wong 1979
Mean shift clustering			Cheng 1995
Spectral Clustering		Custom solution through Base SAS and the DISTANCE and PRINCOMP procedures	Von Luxburg 2007
Kernel Density Estimation		KDE	Silverman 1986

ALGORITHMEN IN SAS (2)

Algorithmus	SAS Enterprise Miner	SAS Prozeduren	Literatur
Kernel PCA		Custom solution through Base SAS and the CORR, PRINCOMP, and SCORE procedures	Schölkopf, Smola, and Müller 1997
Sparse PCA			Zou, Hastie, and Tibshirani 2006
Singular value decomposition		HPTMINE IML, SPSVD	Golub and Reinsch 1970
Self organizing maps	SOM/Kohonen Node		Kohonen 1984
Nonnegative matrix factorization			Lee and Seung 1999

UNSUPERVISED LEARNING

BEISPIEL: PATIENTEN MIT AUFFÄLLIGEN BEHANDLUNGEN

- Fragestellung: Identifiziere Patienten mit auffälliger Behandlungshistorie.
- Daten:
  - Demographische Patientendaten (z.B. Alter, Einkommen, Geschlecht)
  - Patienten Behandlungshistorie (130 unterschiedliche Behandlungsarten)
  - Liste von Patienten, die als auffällig galten und für eine manuelle Prüfung selektiert wurden.

	id	age	gender
1	306535165	65-74	F
2	742231134	65-74	F
3	928807201	75-84	M
4	83686538	<65	F
5	755607449	65-74	F
6	247750664	75-84	F
7	141100339	<65	M
8	547433468	<65	M
9	719408667	65-74	F
10	769089007	75-84	M
11	841546119	65-74	F

	id	global_proc_id	count
1	7561	22	1
2	7561	25	1
3	7619	17	1
4	7619	34	1
5	10479	2	1
6	10479	9	1
7	10479	26	1
8	10479	87	1
9	13515	14	1
10	13515	26	1
11	15213	19	1
12	15213	25	1
13	15213	26	1
14	15213	27	1
15	15213	99	1
16	15213	112	1
17	16017	22	1
18	16017	26	1
19	16017	68	1

http://support.sas.com/resources/papers/proceedings15/SAS2520-2015.zip

BEISPIEL
PATIENTEN MIT AUFFÄLLIGEN BEHANDLUNGEN

sasclub

- Daten verdichten: Singular Value Decomposition

```
proc hptmine
  data= repo.transaction_coo;
    svd
      k= 10
      row= global_proc_id
      col= id
      entry= count
      outdocpro= svdpro;
  performance nthreads= 4;
run;
```

	id	global_proc_id	count
1	7561	22	1
2	7561	25	1
3	7619	17	1
4	7619	34	1
5	10479	2	1
6	10479	9	1
7	10479	26	1
8	10479	87	1
9	13515	14	1
10	13515	26	1
11	15213	19	1
12	15213	25	1
13	15213	26	1
14	15213	27	1
15	15213	99	1
16	15213	112	1
17	16017	22	1
18	16017	26	1
19	16017	68	1

1

BEISPIEL
PATIENTEN MIT AUFFÄLLIGEN BEHANDLUNGEN

- 1000 Cluster erstellen und den Anteil an geprüften Patienten in jeden Cluster berechnen
- Patienten in den 6 Cluster mit den höchsten Anteil sind auffällig.

```
proc hpclus
  data= patient_history_std
  outstat= patient_cluster_profile1000
  maxclusters= 1000    maxiter= 100 seed= 12345
  standardize= none    impute= none  noc= none;
  input
      ENCODED_GENDER ENCODED_AGE  ENCODED_INCOME
         /* SVD FEATURES: COL1-COL10 */
      COL1 COL2 COL3 COL4 COL5 COL6 COL7 COL8   COL9 COL10;
   /* COPY THESE TO THE OUTPUT SET */
  id ID REVIEW_FLAG GENDER AGE INCOME   COL1 COL2 COL3
       COL4  COL5 COL6 COL7 COL8  COL9 COL10;
  score out= patient_cluster_label1000;
  performance threads= 4;
run;
```

BEISPIEL
PATIENTEN MIT AUFFÄLLIGEN BEHANDLUNGEN

- Assoziationsanalyse:
  - Gemeinsam auftretende Behandlungen unter den geprüften und den anderen Patienten.

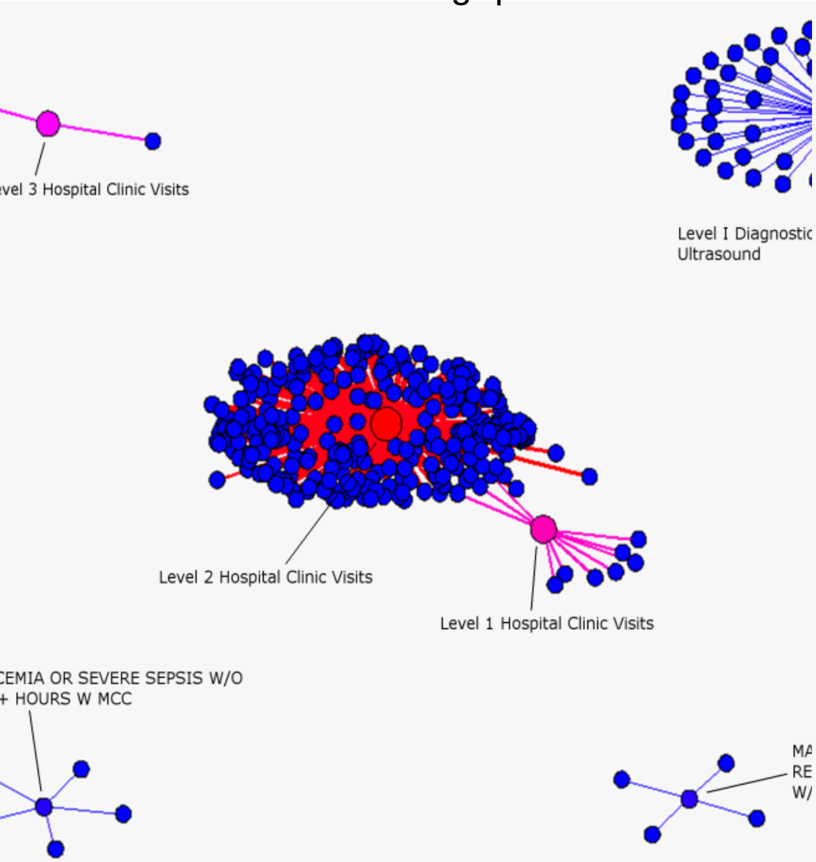
```
proc assoc
    data= repo.review_transaction_coo(keep= id global_proc_id)
    dmdbcat= work.fraudulent_transaction_cat
    out= freq_fraud_trans_group
    items= 2
    support= 30;
    customer id;
    target global_proc_id;
run;
```

COUNT	ITEM1	ITEM2
682	1	2
769	1	3
34	1	4
51	1	5

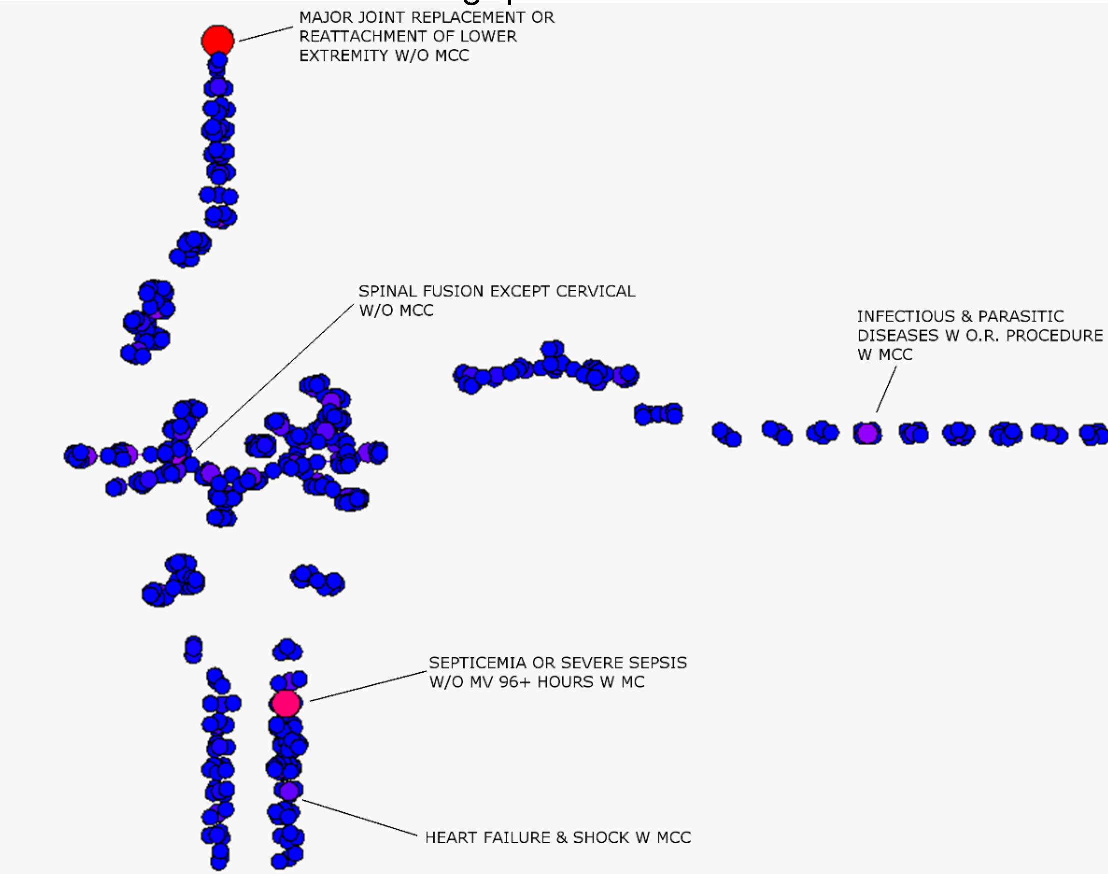
UNSUPERVISED LEARNING | BEISPIEL PATIENTEN MIT AUFFÄLLIGEN BEHANDLUNGEN

- Assoziationsanalyse:
  - Finde Behandlungskombinationen, die NUR bei den geprüften Patienten häufig auftreten.
  - Suche unter allen Patienten solche mit diesen Behandlungskombinationen.

Kombinationen nicht geprüfter Patienten



Kombinationen geprüfter Patienten



Machine Learning

Supervised Learning

Unsupervised Learning

Semisupervised Learning

Data Mining

ALGORITHMEN IN SAS (1)

Algorithmus	SAS Enterprise Miner	SAS Prozeduren	Literatur
Denoising autoencoders		HPNEURAL NEURAL	Vincent et al. 2008
Expectation maximization			Nigam et al. 2000
Manifold regularization			Belkin, Niyogi, and Sindhwani 2006
Transductive support vector machine			Joachims 1999

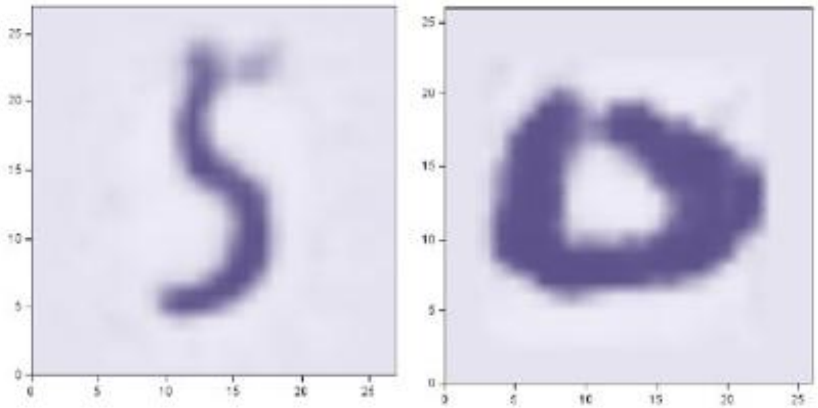
SEMI-SUPERVISED LEARNING

BEISPIEL STACKED DENOISING AUTOENCODER

sas club

- Fragestellung: Extrahiere wenige repräsentative Merkmale.

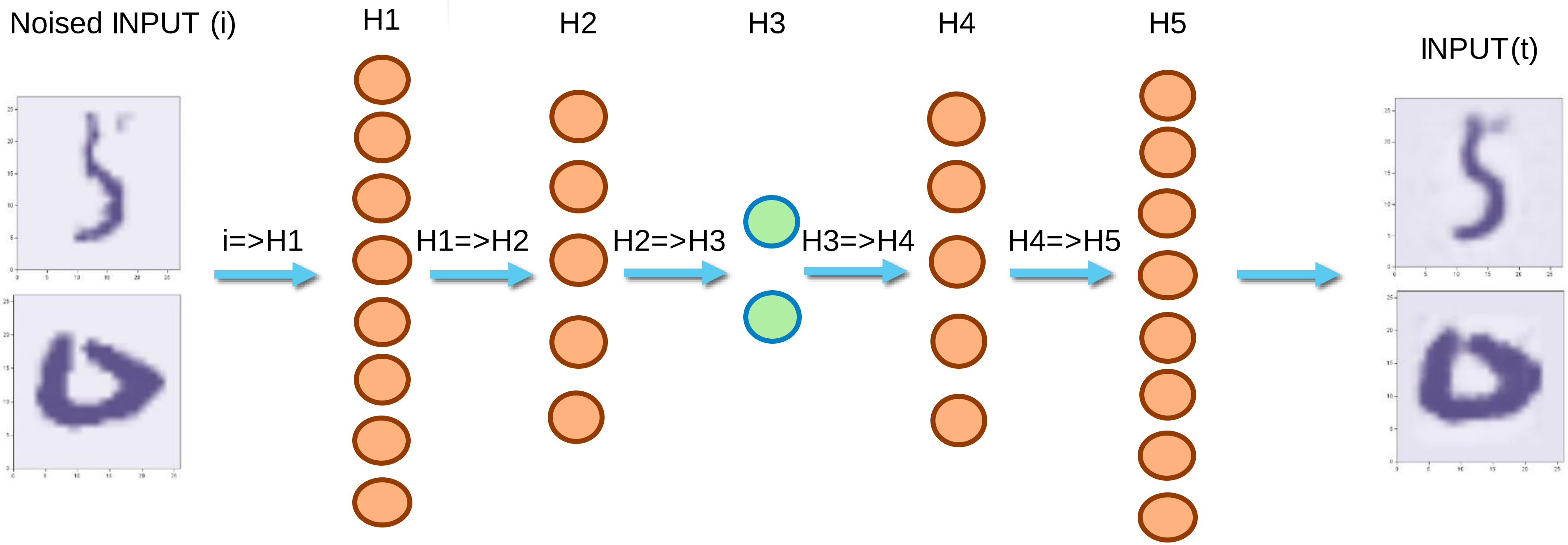
- Daten:
  - Jedes Bild besteht aus 28x28 pixel.
  - 784 Variablen und jedes Bild ist eine Zeile im Dataset.
  - Variablen haben Werte zwischen 0 und 255.
  - 1 Zielvariable für den Wert der Zahl.



digit	pix1	pix2	...	pix784	TARGET (LABEL)
1	0	8	...	0	4
2	0	3	...	0	3
3	244	1	...	0	2
4	78	3	...	3	7
5	0	0	...	4	8
...	...	...	...	...	...
...	...	...	...	...	...
42000	3	0	...	...	9

SEMI-SUPERVISED LEARNING

STACKED DENOISING AUTOENCODER MIT NEURONALEN NETZEN

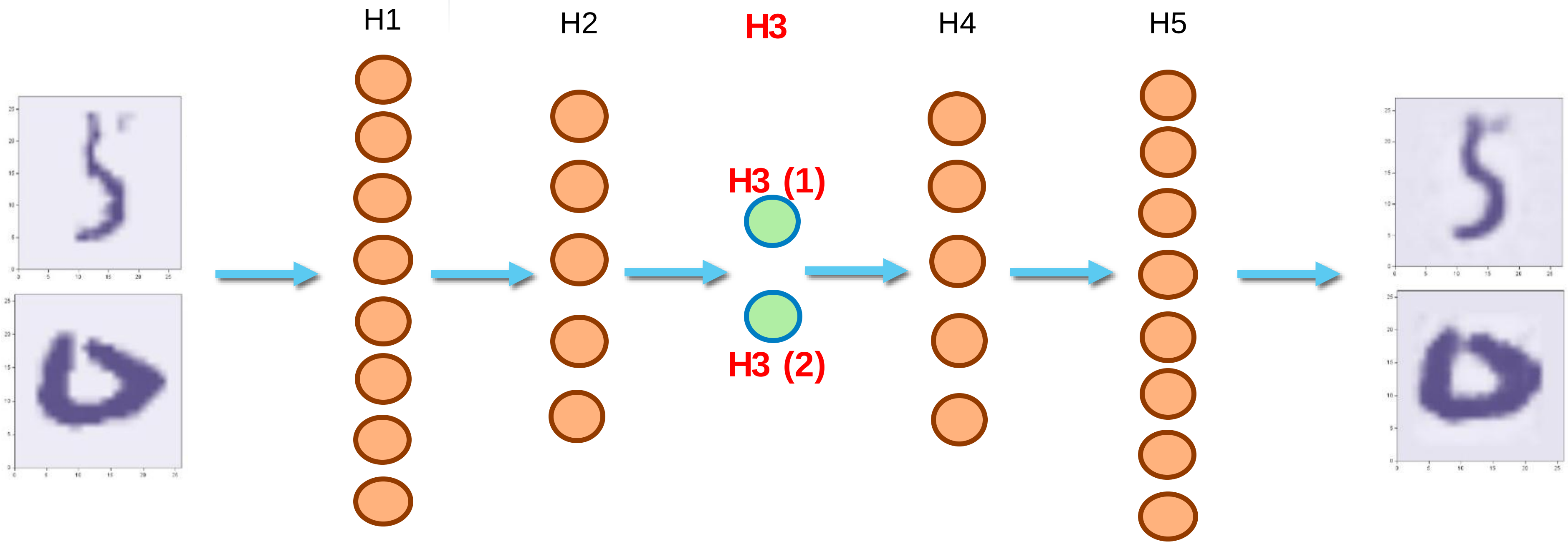



```
proc neural
data= autoencoderTraining   dmdbcat= work.autoencoderTrainingCat
performance  compile  details  cpucount= 12 threads= yes;

archi MLP hidden=  5;
hidden 300 / id= h1;
hidden 100 / id= h2;
hidden 2     / id= h3 act= linear;
hidden 100 / id= h4;
hidden 300 / id= h5;

input corruptedPixel1-corruptedPixel400   / id= i level=  int std= std;
target pixel1-pixel400                    / act= identity id= t level=  int std= std;
```

```
initial random=123;  prelim 10 preiter=10;
freeze  h1->h2; freeze  h2->h3;  freeze  h3->h4; freeze  h4->h5;
train technique= congra maxtime=129600 maxiter=1000;

freeze  i->h1; thaw h1->h2;
train technique= congra maxtime=129600 maxiter=1000;

freeze  h1->h2;  thaw h2->h3;
train technique= congra maxtime=129600 maxiter=1000;

freeze  h2->h3;  thaw h3->h4;
train technique= congra maxtime=129600 maxiter=1000;

freeze  h3->h4;  thaw h4->h5;
train technique= congra maxtime=129600 maxiter=1000;

thaw i->h1;  thaw h1->h2; thaw h2->h3; thaw h3->h4;
train technique= congra maxtime=129600 maxiter=1000;
code file='C:\Path\to\code.sas'; run;
```

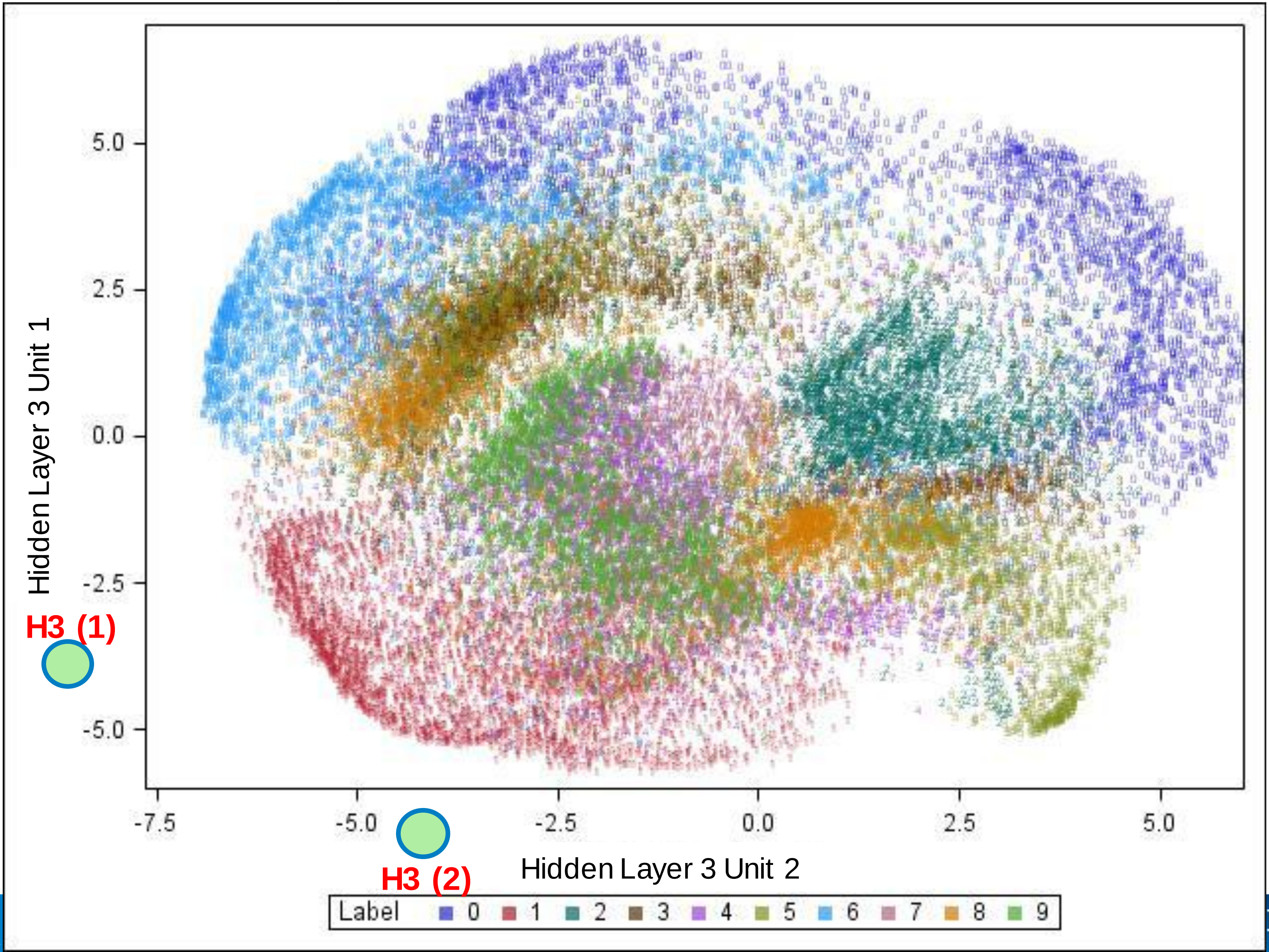
SEMI-SUPERVISED LEARNING

STACKED DENOISING AUTOENCODER MIT NEURONALEN NETZEN



SEMI-SUPERVISED LEARNING

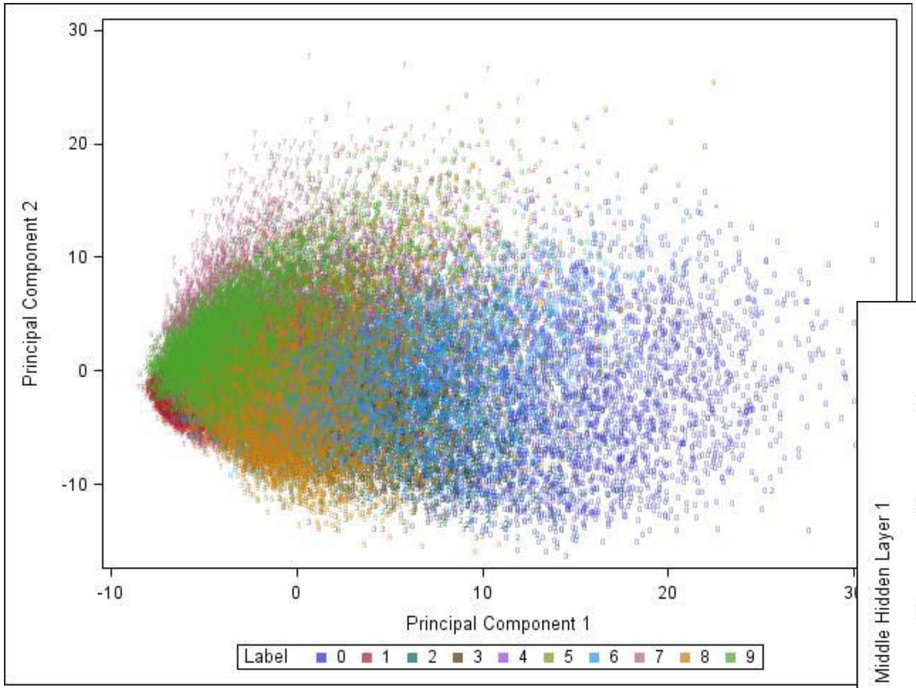
STACKED DENOISING AUTOENCODER MIT NEURONALEN NETZEN VISUALISIERUNG



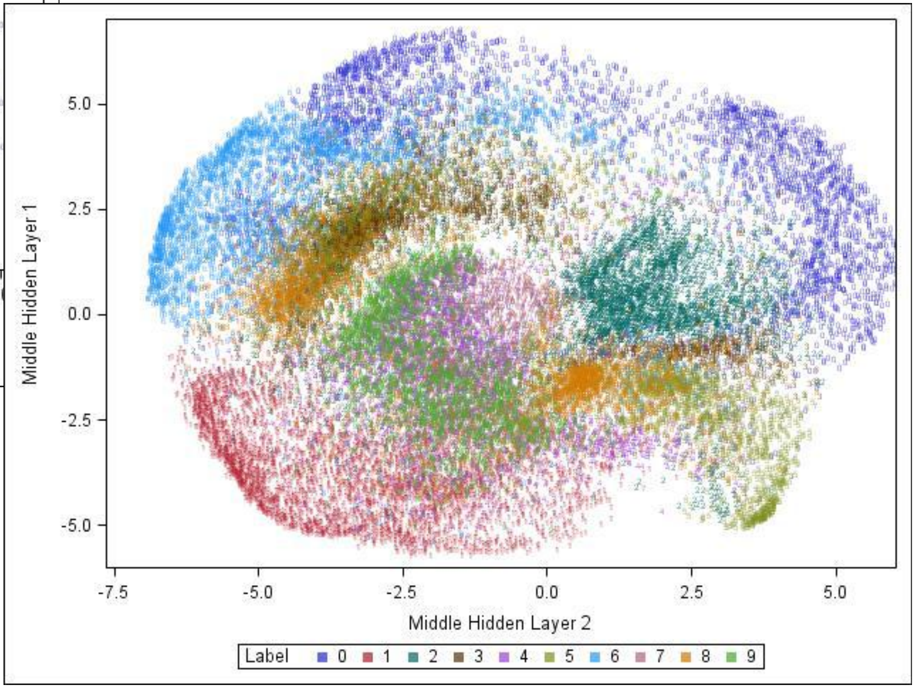
SEMI-SUPERVISED LEARNING

STACKED DENOISING AUTOENCODER MIT NEURONALEN NETZEN VERGLEICH ZU PCA

Ersten zwei Hauptkomponenten



Middle Layer des Stacked Denoising Autoencoders



"Overview of Machine Learning with SAS Enterprise Miner"
http://support.sas.com/resources/papers/proceedings14/SAS313-2014.pdf
http://support.sas.com/rnd/papers/sasgf14/313_2014.zip

Quellen:

Hall P. 2015. SAS® Does Data Science: How to Succeed in a Data Science Competition
http://support.sas.com/resources/papers/proceedings15/SAS2520-2015.pdf
Agrawal, R., Imieliński, T., and Swami, A. 1993. "Mining Association Rules between Sets of Items in Large Databases." ACM SIGMOD Record 22:207–216.
Belkin, M., Niyogi, P., and Sindhwani, V. 2006. "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples." Journal of Machine Learning Research 7:2399–2434.
Breiman, L. 2001a. "Random Forests." Machine Learning 45:5–32.
Breiman, L. 2001b. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." Statistical Science 16:199–231.
Cheng, Y. 1995. "Mean Shift, Mode Seeking, and Clustering." IEEE Transactions on Pattern Analysis and Machine Intelligence 17:790–799.
Cortes, C. and Vapnik, V. 1995. "Support-Vector Networks." Machine Learning 20:273–297.
Cover, T. and Hart, P. 1967. "Nearest Neighbor Pattern Classification." IEEE Transactions on Information Theory 13:21–27.
de Ville, B. and Neville, P. 2013. Decision Trees for Analytics Using SAS Enterprise Miner. Cary, NC: SAS Institute Inc.
Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics 29:1189–1232.
Friedman, N., Geiger, D., and Goldszmidt, M. 1997. "Bayesian Network Classifiers." Machine Learning 29:131–163.
Golub, G. H. and Reinsch, C. 1970. "Singular Value Decomposition and Least Squares Solutions." Numerische Mathematik 14:403–420.
Hartigan, J. A. and Wong, M. A. 1979. "Algorithm AS 136: A k-Means Clustering Algorithm." Journal of the Royal Statistical Society, Series C 28:100–108.
Joachims, T. 1999. "Transductive Inference for Text Classification Using Support Vector Machines." Proceedings of the 16th International Conference on Machine Learning. New York: ACM.
Kohonen, T. 1984. Self-Organization and Associative Memory. Berlin: Springer-Verlag.
Lee, D. D. and Seung, H. S. 1999. "Learning the Parts of Objects by Non-negative Matrix Factorization." Nature 401:788–791.
Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. 2000. "Text Classification from Labeled and Unlabeled Documents Using EM." Machine Learning 39:103–134.
Panik, M. J. 2009. Regression Modeling: Methods, Theory, and Computation with SAS. Boca Raton, FL: CRC Press.
Schölkopf, B., Smola, A., and Müller, K.-R. 1997. "Kernel Principal Component Analysis." In Artificial Neural Networks—ICANN'97, 583–588. Berlin: Springer.
Seeger, M. 2004. "Gaussian Processes for Machine Learning." International Journal of Neural Systems 14:69–106.
Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. Vol. 26. Boca Raton, FL: CRC Press.
Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. 2008. "Extracting and Composing Robust Features with Denoising Autoencoders." Proceedings of the 25th International Conference on Machine Learning. New York: ACM.
Zou, H., Hastie, T., and Tibshirani, R. 2006. "Sparse Principal Component Analysis." Journal of Computational and Graphical Statistics