Kann SAS Ihre Handschrift lesen? Machine Learning am Beispiel von Stacked Denoising Autoencoders

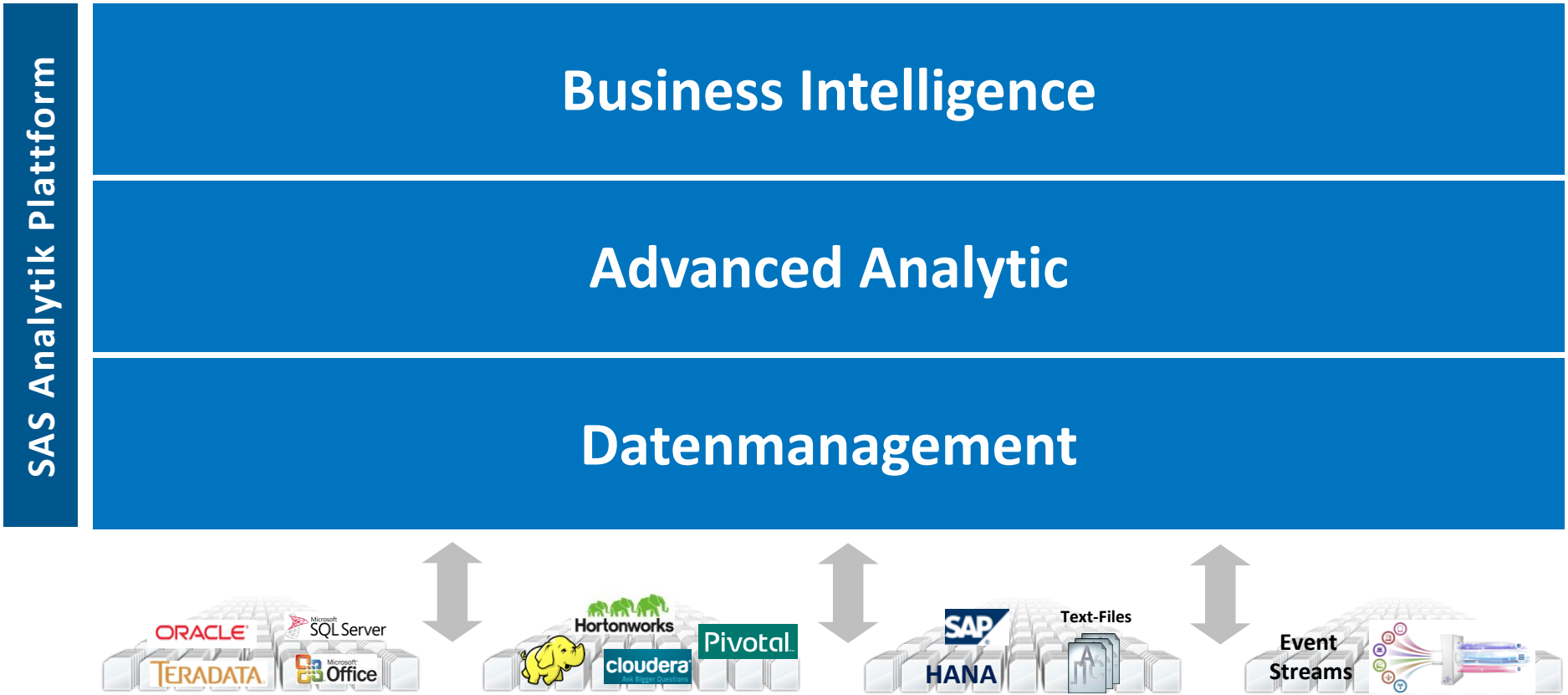
21. KSFE, Krefeld, 9.-10. März 2017

Gerhard Svolba

Die Vortragsfolien sind online → Google: Gerhard SAS Samples

SAS Analytik Plattform

Unterschiedliche Layer aus konzeptioneller Sicht



SAS Analytik Plattform

Advanced Analytic Layer

SAS Analytik Plattform

Business Intelligence

Data Mining

Statistical Analysis

Forecasting

Text Analytics

Optimization & Simulation

Datenmanagement

Concepts when Handling Big Data

- Using advanced machine learning methods to describe the relationships in your data
- Understanding specifics of complex systems by performing Monte Carlo simulations
- Executing your analysis processes in distributed in-memory mode (SAS High Performance Analytics,
  SAS Viya)

Machine Learning

Data Mining

SUPERVISED LEARNING

- Regression
  - LASSO regression
  - Logistic regression
  - Ridge regression
- Decision tree
  - Gradient boosting
  - Random forests
- Neural networks
- SVM
- Naïve Bayes
- Neighbors
- Gaussian processes

UNSUPERVISED LEARNING

- A priori rules
- Clustering
  - *k*-means clustering
  - Mean shift clustering
  - Spectral clustering
- Kernel density estimation
- Nonnegative matrix factorization
- PCA
  - Kernel PCA
  - Sparse PCA
- Singular value decomposition
- SOM

SEMI-SUPERVISED LEARNING

- Prediction and classification*
- Clustering*
- EM
- TSVM
- Manifold regularization
- Autoencoders
  - Multilayer perceptron
  - Restricted Boltzmann machines

TRANSDUCTION

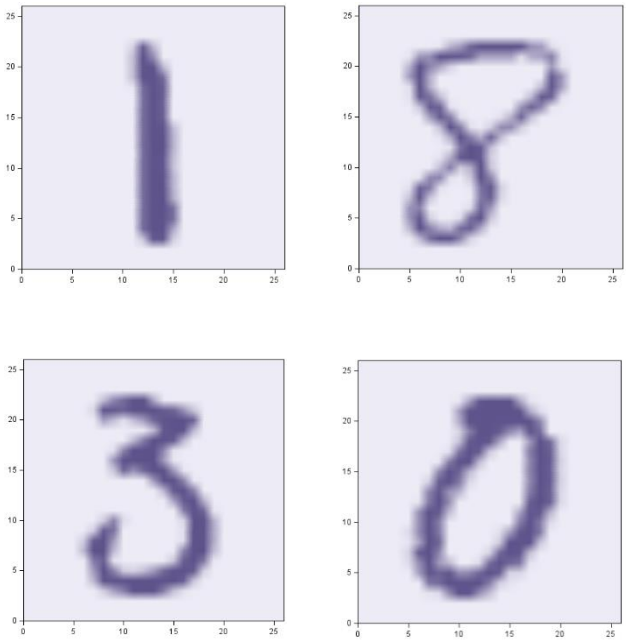
REINFORCEMENT LEARNING

DEVELOPMENTAL LEARNING

*In semi-supervised learning, supervised prediction and classification algorithms are often combined with clustering.

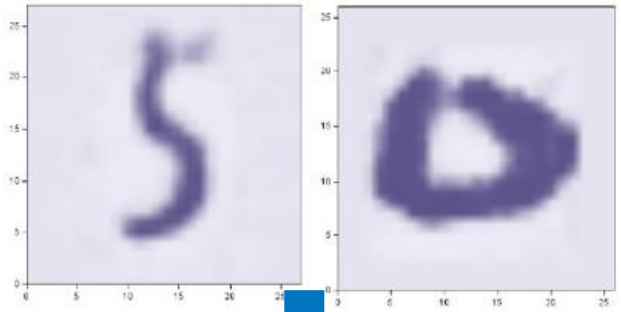
Handwritten Digits as Training Data

- Classic MNIST training data
- 784 features from a 28x28 digital grid
- Greyscale features range from 0 to 255
- 60,000 labeled training images
  (785 variables, including 1 nominal target)
- 10,000 unlabeled test images
  (784 input variables)



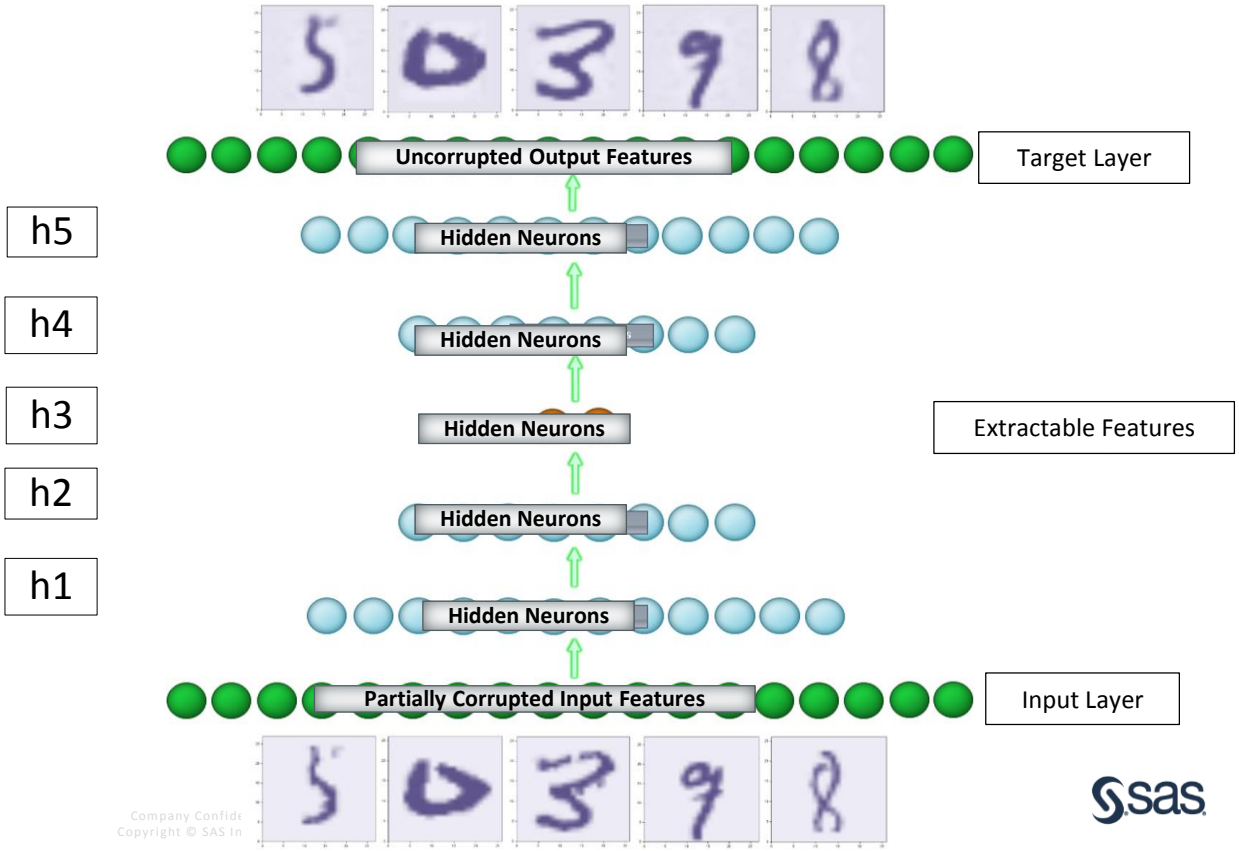
Semi-Supervised Learning

- Extract a few representative features to discriminate the digits 0-9
- Compress information of 784 variables into 2 features
- Use a convolutional neural network (deep learning)

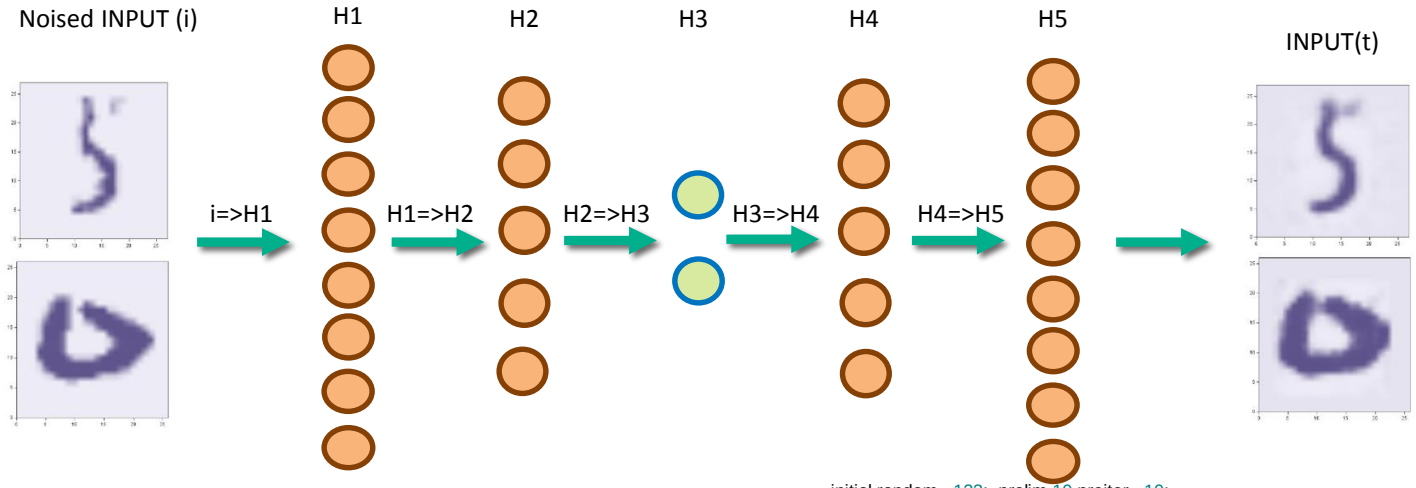


digit	pix1	pix2	...	pix784	TARGET (LABEL)
1	0	8	...	0	4
2	0	3	...	0	3
3	244	1	...	0	2
4	78	3	...	3	7
5	0	0	...	4	8
...	...	...	...	...	...
...	...	...	...	...	...
42000	3	0	...	...	9

Deep-Learning using a Stacked De-noising Autoencoder



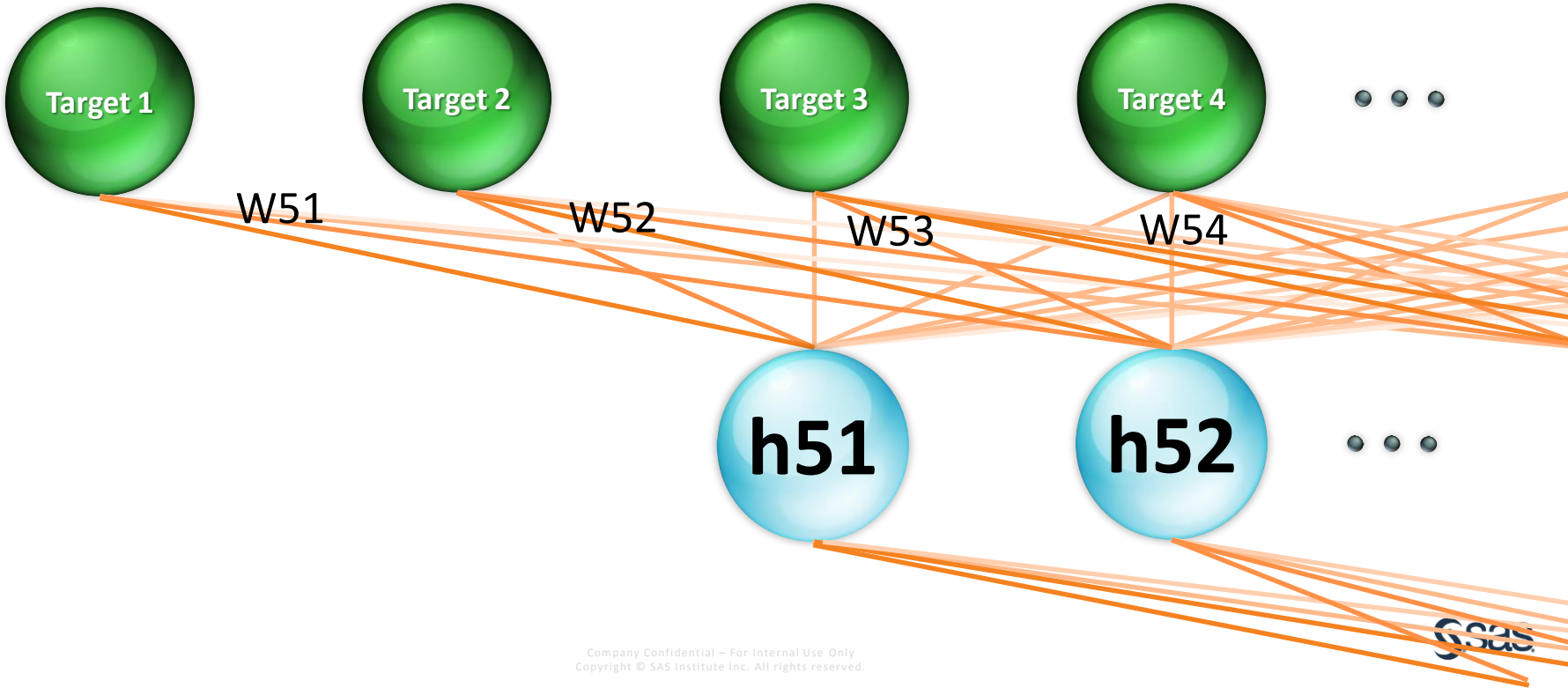
Using SAS Code to Solve the Problem



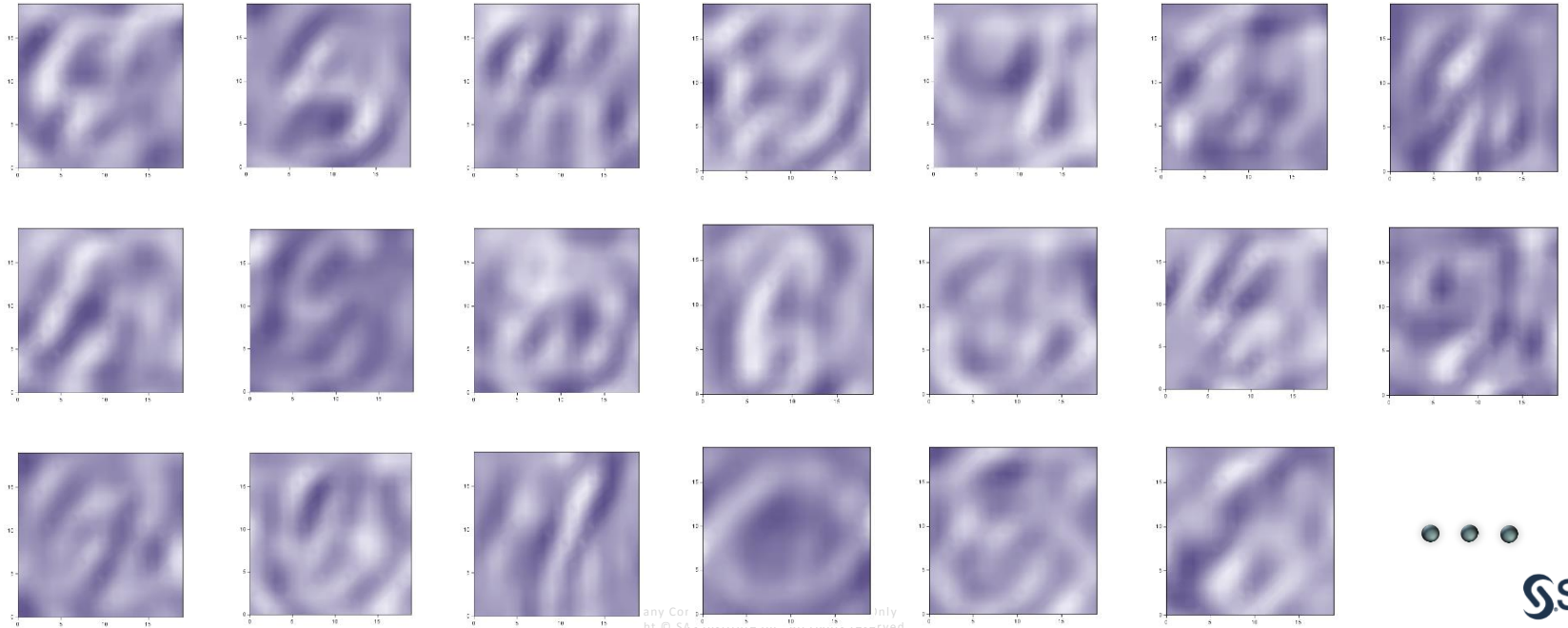

```
proc neural
data= autoencoderTraining  dmdbcat= work.autoencoderTrainingCat
performance compile details cpucount= 12 threads= yes;

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;
```

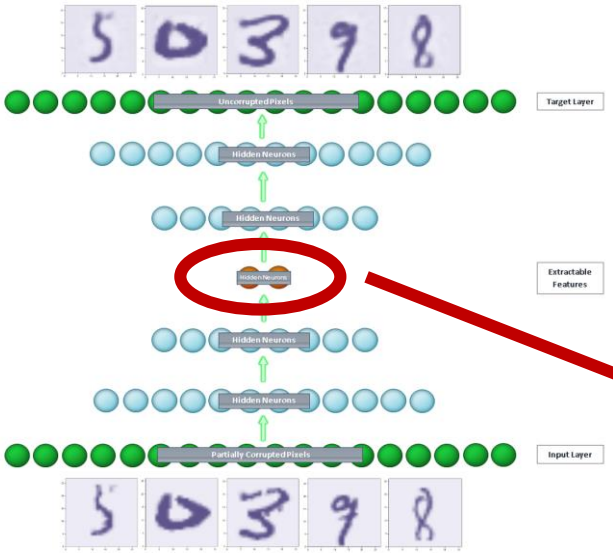
Studying a certain section in detail

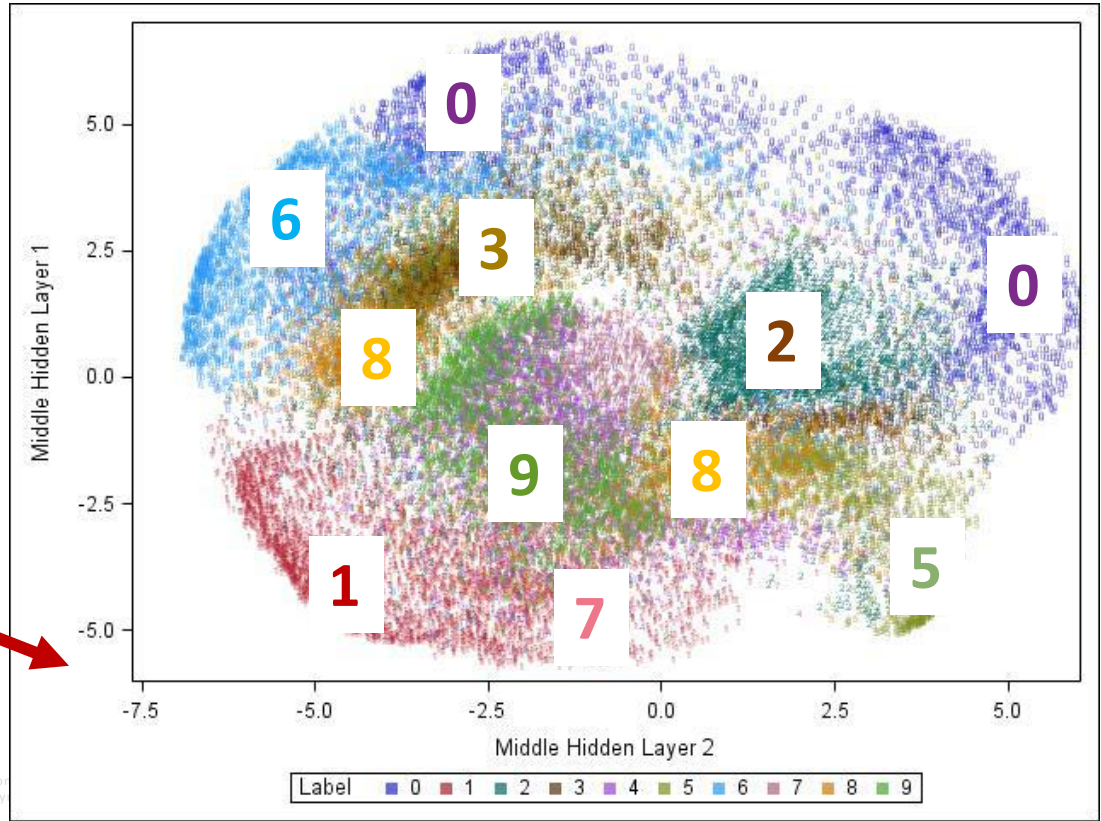


Edge Weights of the 5th layer are “loaded” with discriminative information

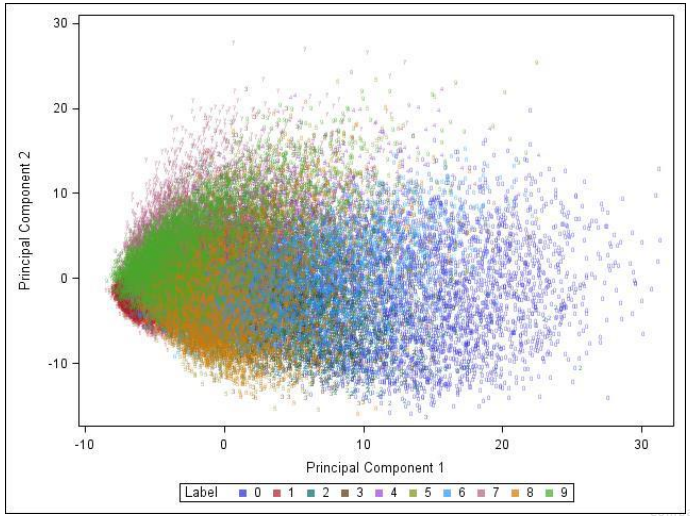


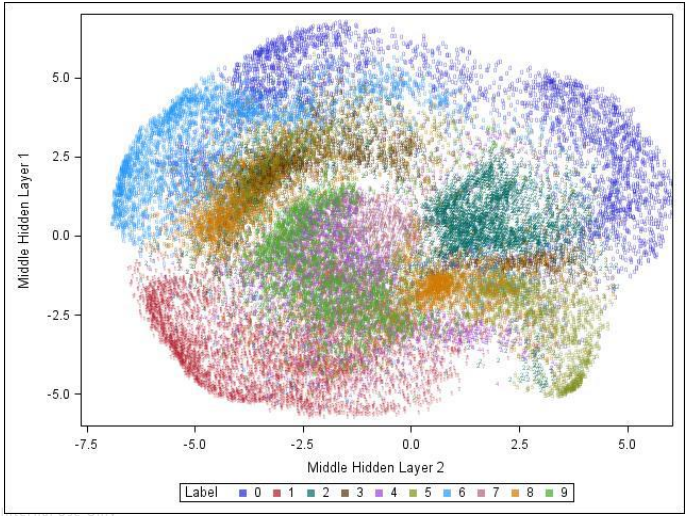
Visualization of the separation of the two middle hidden layers





Our method results in much better separation than simple principal components analysis





Summary: Semi-Supervised Learning

- Extremely accurate predictions using deep neural networks.
- „Target Variable“ Digit 0-9 has not been used in the model!
- “Feature Extraction“ as pre-step in predictive modeling
- Requires Model-Tuning
- The most common applications of deep learning involve pattern recognition in unstructured data, such as text, photos, videos and sound.

Bildauswertung in der Versicherung

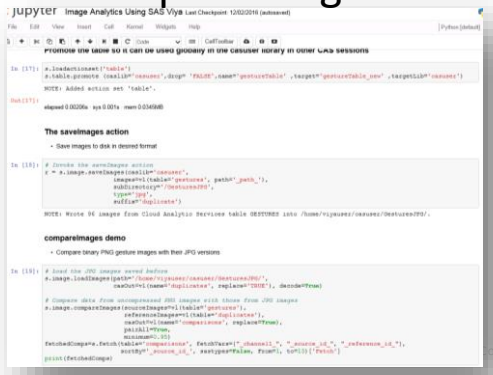
Bildanalyse mit SAS

Analyseprozess

Bild

Image Pre-processing

LoadImage
ProcessImage
CompareImage

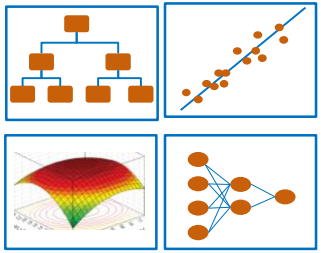


Datenmatrix

c1	c2	c3	c4
i0	i1	i2	i3
j0	j1	j2	j3

Analytics

SAS Machine Learning Verfahren



Ergebnis

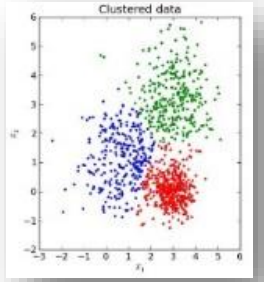


Image analysis

IMAGE PREPROCESSING/DATA REDUCTION

Image Pre-processing

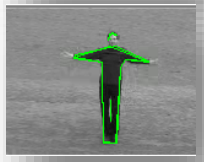
- Grayscale

- Bilateral filtering

- Thresholding

- Edge detection

- Contour detection



- Contour approximation

- Bounding box

- Group bounding box

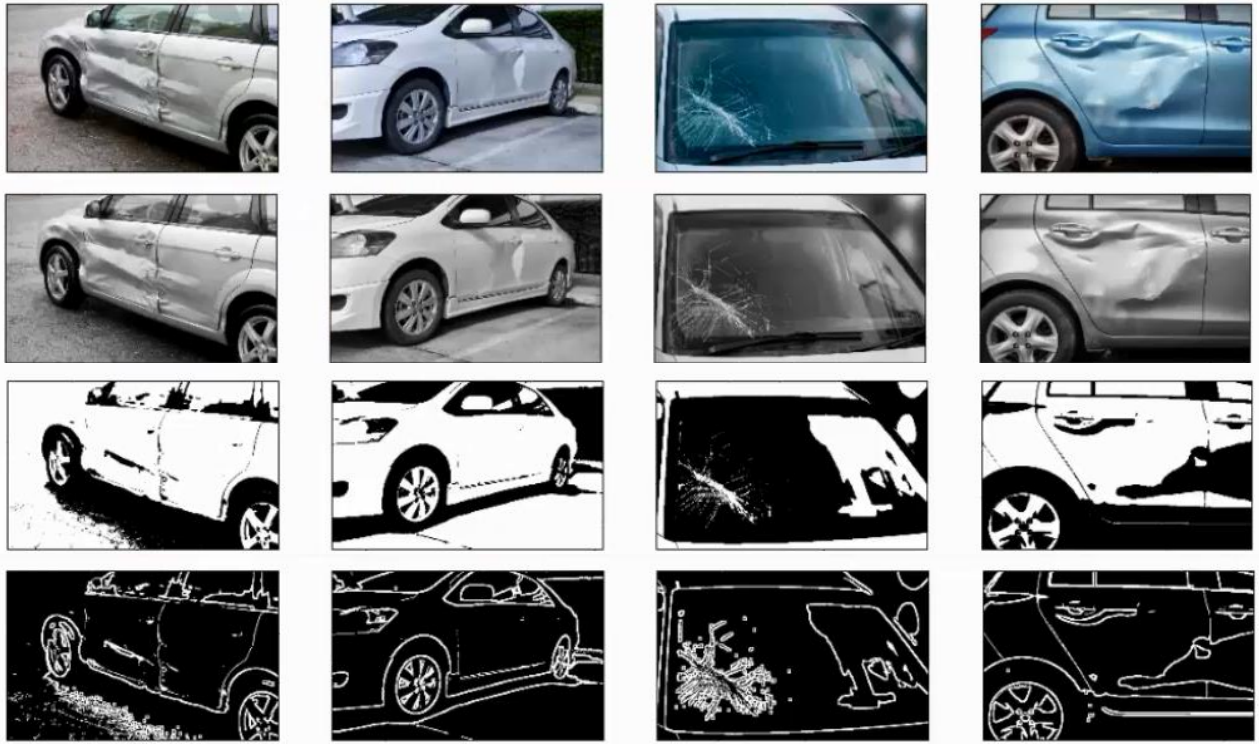
Bildauswertung in der Versicherung
AUS BILDERN WERDEN ZAHLEN

Größe standardisieren

Farben konvertieren in schwarz/weiß

Rauschen entfernen und Binärdaten erzeugen

Kanten erkennen



Beispiel Bildauswertung in der Versicherung

Integration in Geschäftsprozesse realisiert erst den Nutzen