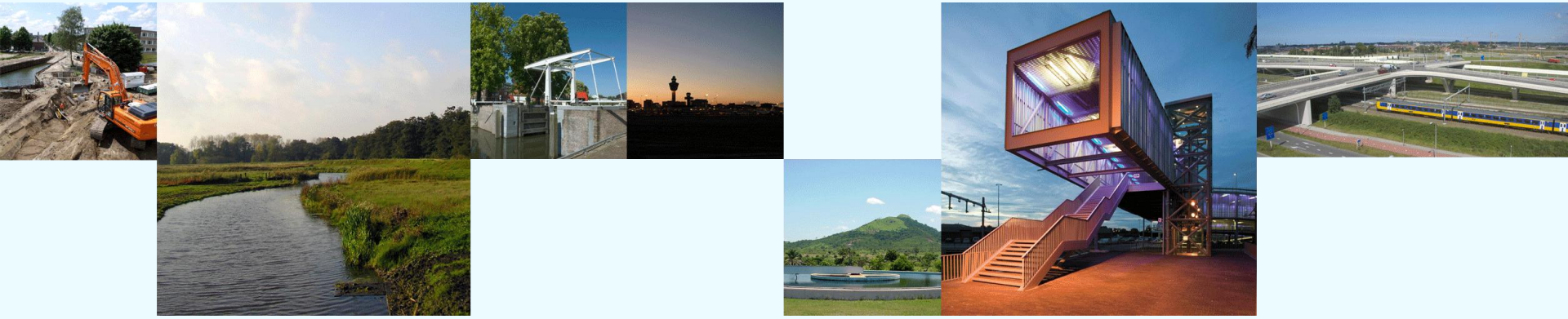
Opening the Black Box: Data Mining with Product Unit Neural Networks



Arie de Niet

Contents

Opening the Black Box:
Data Mining with Product Unit Neural Networks

Case: Modelling Ecological Quality Ratio

1. Introduction
2. Challenge
3. Ecological Quality Ratio
4. Product Unit Neural Network
5. Results
6. Conclusion

1. Introduction

Witteveen+Bos

Deltas, Coasts and Rivers

Coast, Rivers and Land Reclamation

Data Analysis and Information Management

Data Scientist: from data to information

1. Introduction

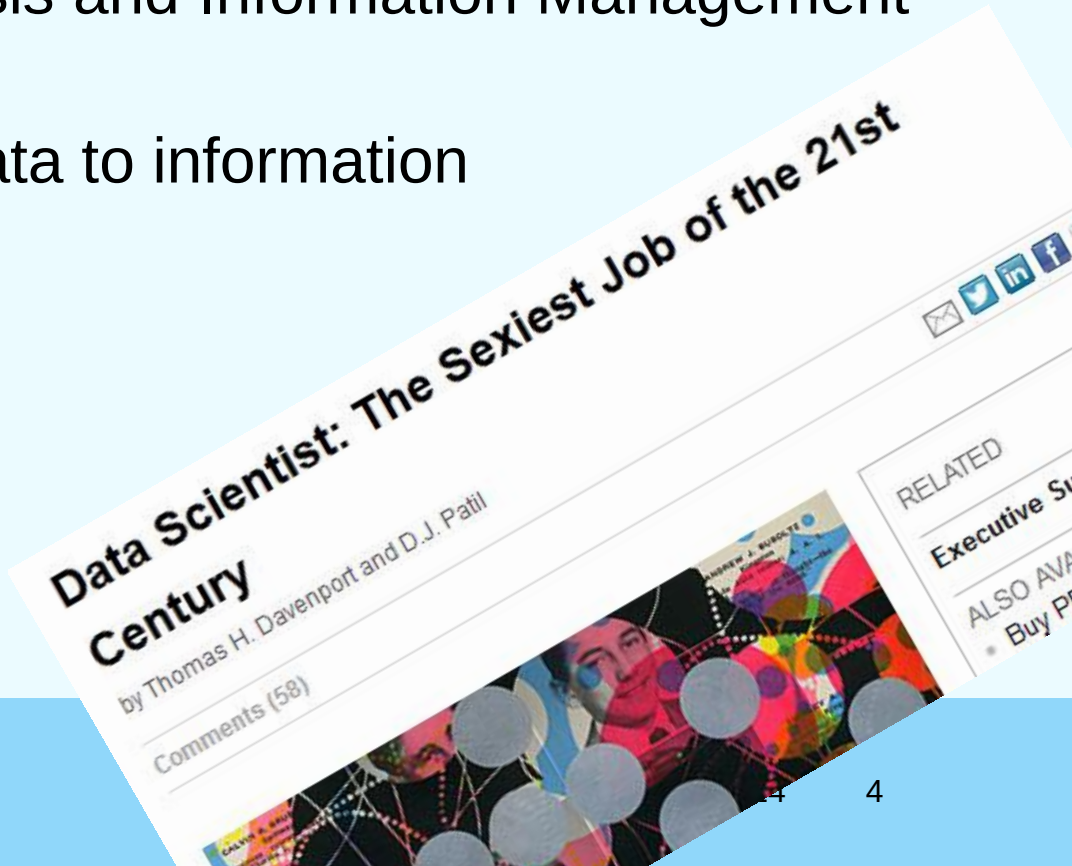
Witteveen+Bos

Deltas, Coasts and Rivers

Coast, Rivers and Land Reclamation

Data Analysis and Information Management

Data Scientist: from data to information



1. Introduction

Group Data Analysis and Information Management:

Unique knowledge on

- collection and analysis of data
- monitoring
- databases
- statistics
- (numerical) modelling

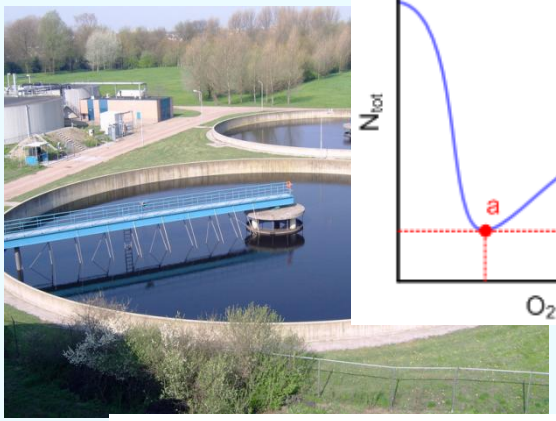
But also

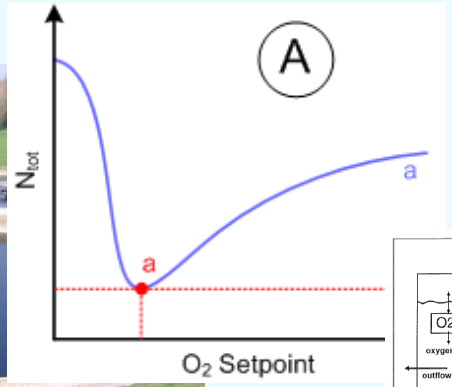
- applied in projects
- focus on management information

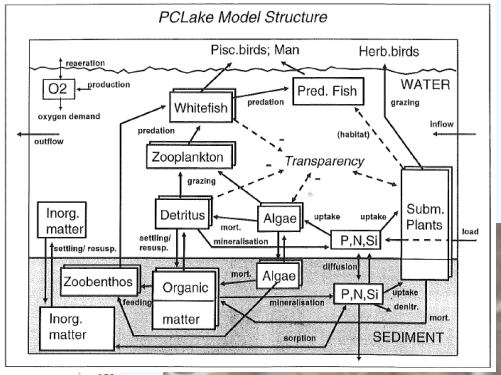
Ambition: smarter / faster / better!

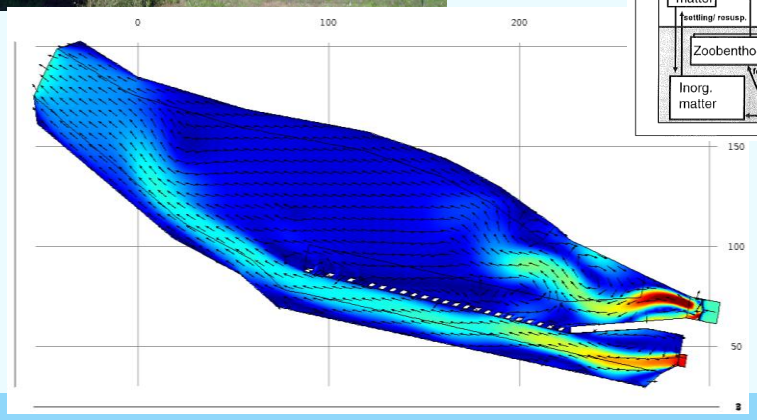
1. Introduction

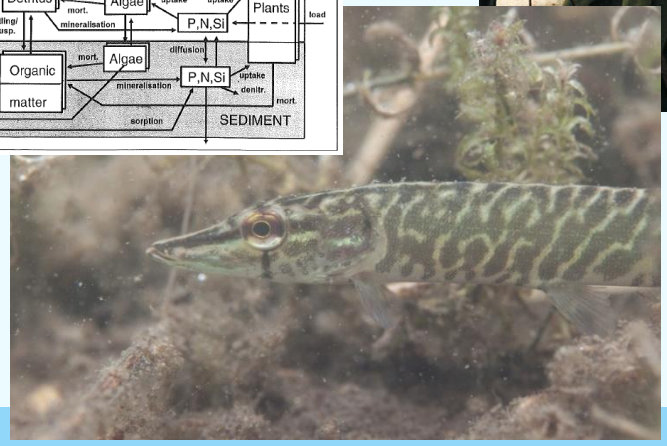
Example of applications











2. Challenge

European Water Framework Directive:

improve water quality and ecological quality of surface water

Many measures taken.

Well-known what type
is most effective
in what case.

But total effect is uncertain.

2. Challenge

Aim: development of data-driven models for accurate prediction of effect of measure.

Data available:

- Dutch water bodies
  - 8 water type clusters
- characteristics

Banks	Shading
Level Control	Shipping
Maintenance	BOD5
Connectivity	Chloride
Meandering	Total Phosphorus
Weirs	Total Nitrogen

- ecological quality (4 EQR's)
- 10 yrs

2. Challenge

Explaining characteristics
vary per water type cluster

Slow flowing brooks:

Meandering	BOD5
Weirs	Total Phosphorus
Shading	Total Nitrogen

Brackish waters:

Banks	Chloride
Level Control	Total Phosphorus
Maintenance	Total Nitrogen
Connectivity	

Deep lakes:

Banks	Total Phosphorus
Level Control	Total Nitrogen

3. Ecological Quality Ratio

Dutch implementation of EWFD

Ecological Quality Ratio (EQR)

- phytoplankton
- macrofauna
- aquatic flora
- fish

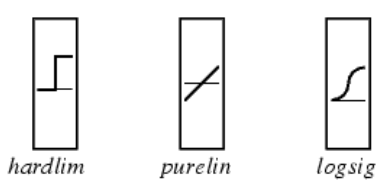
4. Product Unit Neural Network

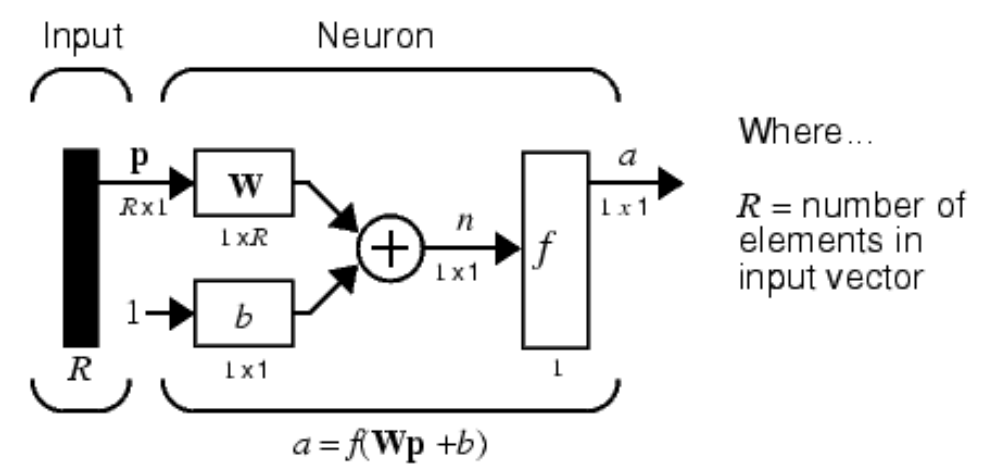
Inspired by human brain

- basis: neuron processes input to output
- network: connected neurons
- learning by training with inputs/targets

Characteristics neuron:

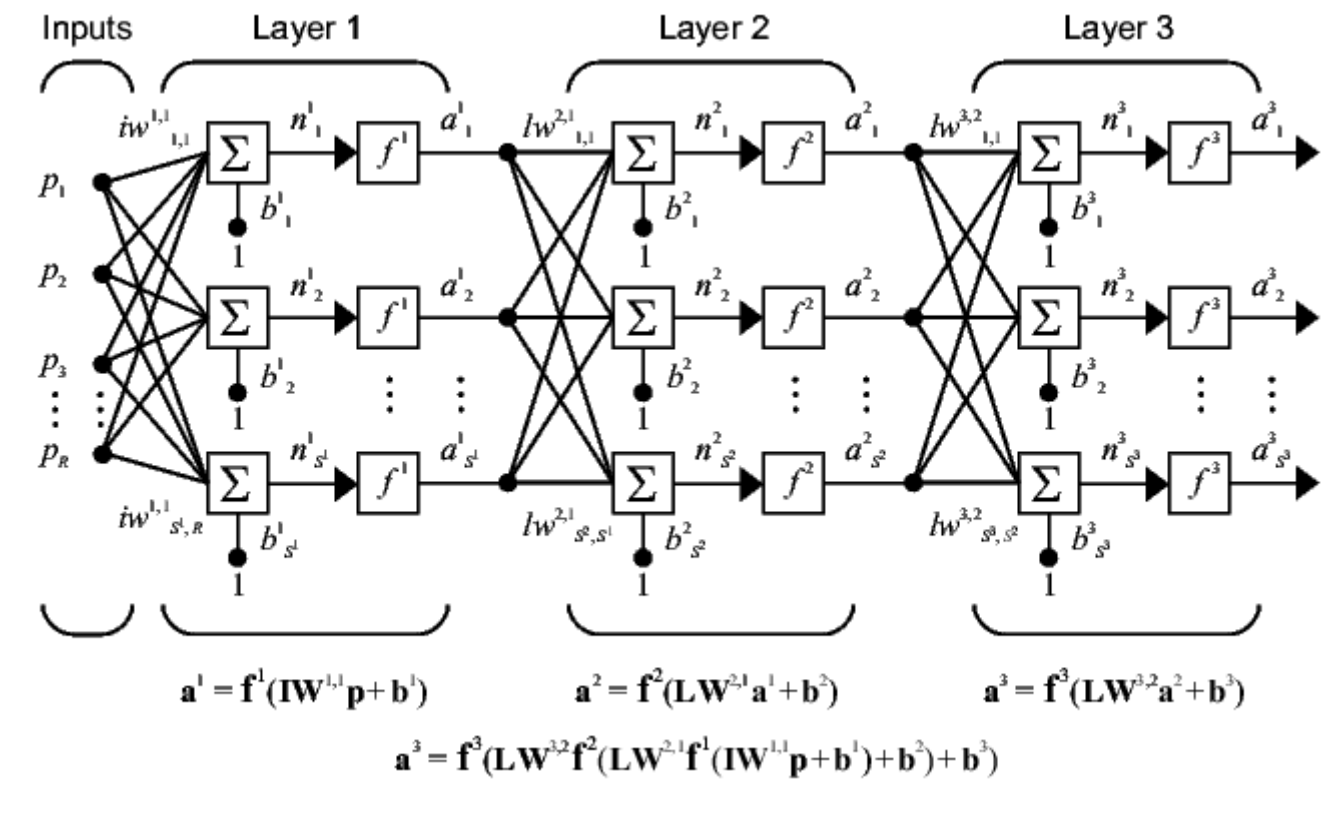
- weight
- bias
- transfer function





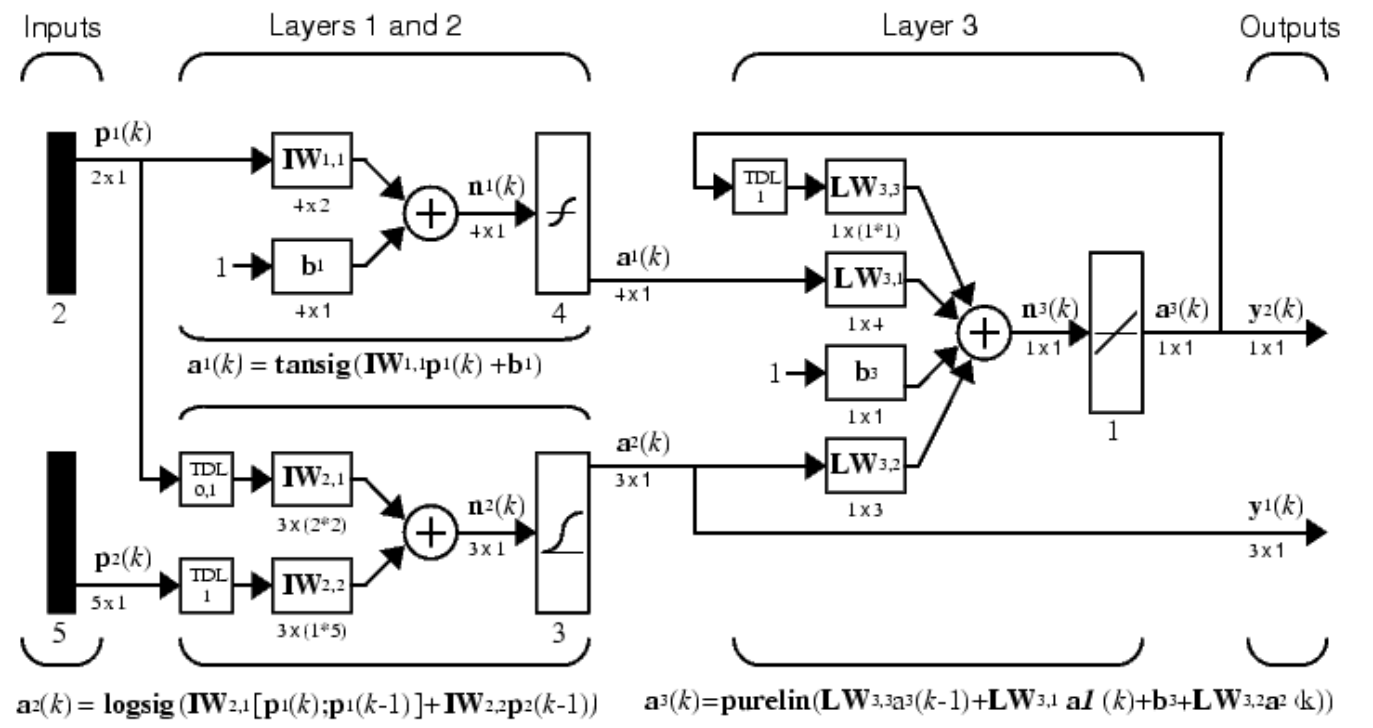
4. Product Unit Neural Network

straightforward Multi Layer Neuron (3x3) Network



4. Product Unit Neural Network

… or complex …



4. Product Unit Neural Network

General Neural Network

Positive:

- simple concept
- easy to apply
- flexibility

Negative:

- no interpretation of model (black box)
- limited posibilities to simplify network
- problems with transportability

4. Product Unit Neural Network

Standaard Neuraal Network:
based on sum

$$\sum_{u=1}^{n_x} w_u x_u$$

Product Unit Neuraal Netwerk:
based on product
weights > powers

$$\prod_{u=1}^{n_x} x_u^{p_u}$$

4. Product Unit Neural Network

Standaard Neuraal Network:
based on sum

$$\sum_{u=1}^{n_x} w_u x_u$$

Product Unit Neuraal Netwerk:
based on product
weights > powers

$$\prod_{u=1}^{n_x} x_u^{p_u}$$

Advantages of PUNN compared to sum network

- great predictive power (smaller network)
- simplification is possible (pruning)
- interpretable (white box)
- transportable (one-line formula)

5. Results

Experimental setup

For each EQR / water type cluster

- Split dataset in training / validation / test

Performance indicators on test set:

- percentage of error less than 0.10;
- root mean squared error (RMSE);
- coefficient of determination (CoD).

Compare to: Regression tree, Standard NN

5. Results

Derivation of PUNN model for EQR:

- Starting PUNN:
  4 products plus a constant;
- 2000 x training after
  random initialization;

- 30 best trained PUNN are pruned;
- best pruned PUNN is delivered as EQR-model.

5. Results

Water type cluster: Deep Lakes

Explaining variables

Banks (Ba)
Level Control (L)
Total Phosphorus (P)
Total Nitrogen (N)

Results

Deep Lakes	Training (incl. Validation)			Evaluation		
EQR	pct <0.1	RMSE	CoD	pct <0.1	RMSE	CoD
phytoplankton	50%	0.15	0.69	46%	0.17	0.63
aquatic flora	76%	0.086	0.85	73%	0.091	0.81
macrofauna	87%	0.066	0.86	85%	0.080	0.60
fish	90%	0.069	0.86	69%	0.12	0.69

5. Results

Water type cluster: Deep Lakes

EQR phytoplankton

$$EQR_{phyt} = -0.7347 + 0.9958 \frac{1}{N^{0.288}} + 0.002878 N^{1.535} + 0.09921 \frac{N^{0.3107}}{P^{0.4086}}$$

EQR aqautic flora

$$EQR_{aqfl} = 7.054 - 0.3623 \frac{L^{3.13}}{Ba^{2.456}} + 0.2406 \frac{L^{3.554}}{Ba^{3.148}} + 0.5741 \frac{1}{L^{7.167}} - 7.474 \frac{P^{0.01563}}{L^{0.1776}}$$

5. Results

Water type cluster: Deep Lakes

EQR macrofauna

$$EQR_{macr} = 0.7516 - 0.1273\ \frac{L^{0.6037} P^{0.3094} N^{0.8157}}{Ba^{0.1388}} + 0.08084\ \frac{Ba^{1.107} L^{1.124}}{P^{0.03239} N^{0.2039}}$$

$$- 0.3805\ \frac{Ba^{0.3697} L^{0.6279} P^{0.05867}}{N^{0.1495}} + 0.03044\ Ba^{0.3636} L^{1.545} P^{0.4326} N^{1.041}$$

EQR fish
Similar long formula

5. Results

Water type cluster: Ditches

Explaining variables

Banks (Ba)	Total Phosphorus (P)
Level Control (L)	Total Nitrogen (N)
Maintenance (Ma)	

EQR fish

$$EQR_{fish} = -10.38 + 6.796\ \frac{L^{0.1253} Ma^{0.1178}}{P^{0.03803}} - 0.1405\ \frac{1}{P^{0.4452}}$$

$$+ 4.326\ \frac{P^{0.03148}}{L^{0.2724} Ma^{0.2525} N^{0.01809}} - 0.1624\ \frac{1}{L^{14.87} Ma^{11.67}}$$

5. Results

Comparison with other methods

Overall performance:

model type	pct <0.1	RMSE	CoD
regression tree	63%	0.121	0.49
standard neural network	64%	0.129	0.41
product unit neural network	68%	0.106	0.60

For 21 out of 29 punn gives best model.

PUNNs incorporated in WFD-explorer (Deltares).

6. Conclusion

For Ecological Quality Ratio the PUNN's

- deliver a data-driven model;
- provide accurate prediction for EQR;
- outperform other methods.

6. Conclusion

For Ecological Quality Ratio the PUNN's

- deliver a data-driven model;
- provide accurate prediction for EQR;
- outperform other methods.

Moreover there are additional advantages:

- interpretability;
- transportability.

6. Conclusion

In case of:

- large datasets
- complex physics
- not well-understood
- need for information

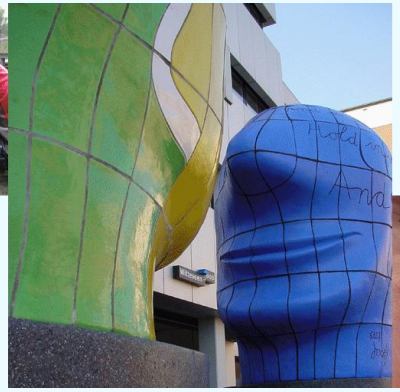
Product Unit Neural Networks
are a powerfull tool
to open the black box!

7. Acknowledgement

Erwin Meijers

Sebastiaan Schep

Sponsored by:




Witteveen Bos

+31 570 69 79 11
info@witteveenbos.com
www.witteveenbos.com

Deventer
Almere
Amsterdam
Breda

Den Haag
Heerenveen
Rotterdam

België (Antwerpen)
Indonesië (Jakarta)
Kazachstan (Aktau, Almaty, Atyrau)
Letland (Riga)
Rusland (St. Petersburg)
Singapore (Singapore)
Vietnam (Ho Chi Minh City)