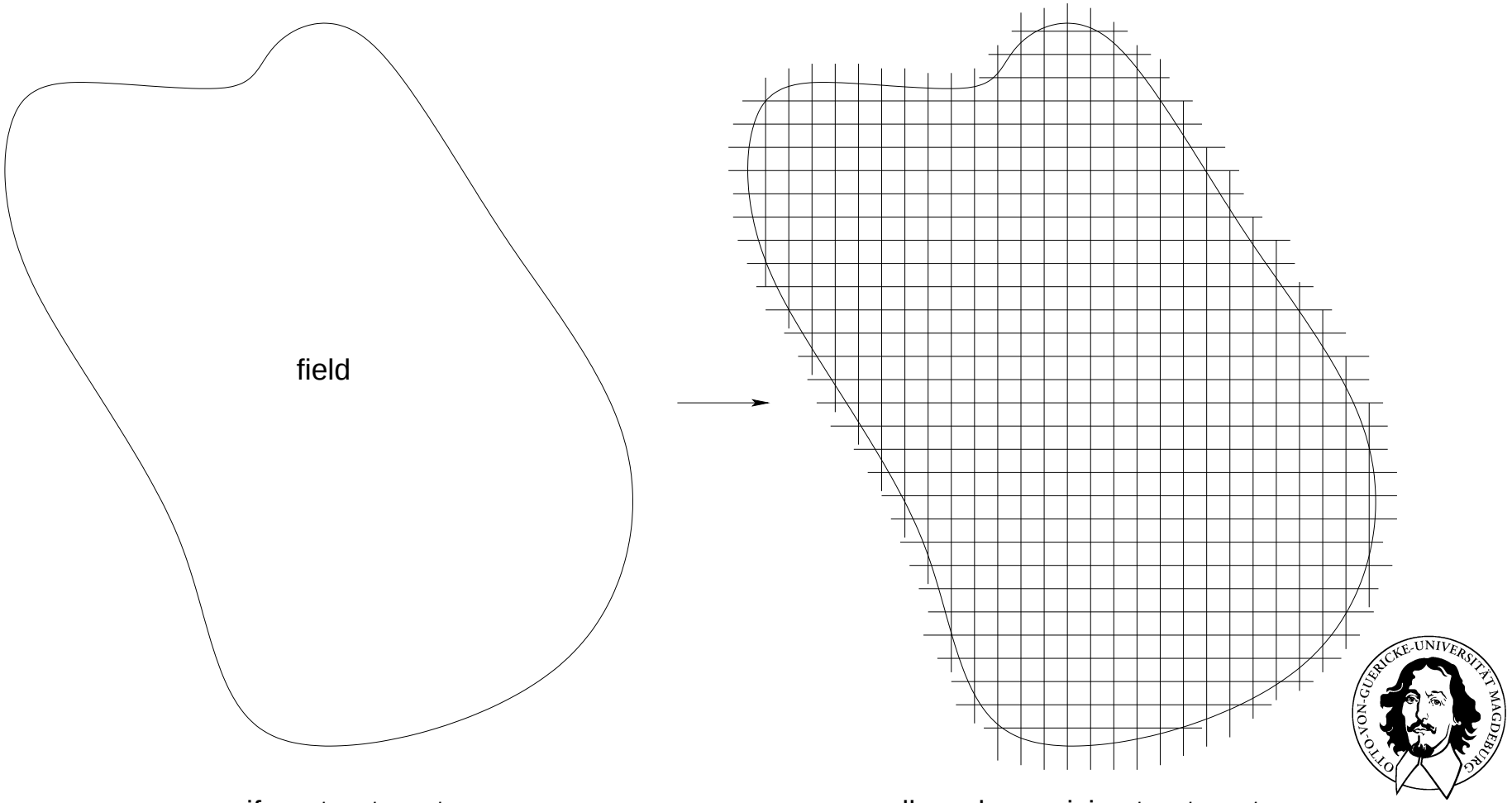
Data Mining on Agriculture Data using Neural Networks

Georg Ruß, Rudolf Kruse, Martin Schneider, Peter Wagner

July 16th, 2008

Motivation: Precision Farming



Motivation: Precision Farming

- precision farming
  - divide field into small-scale parts
  - treat small parts independently instead of uniformly
  - cheap data collection
  - GPS-based technology
- lots of data (sensors, imagery, GPS-tagged)
- use data mining to
  - improve efficiency
  - improve yield

Data Flow Model

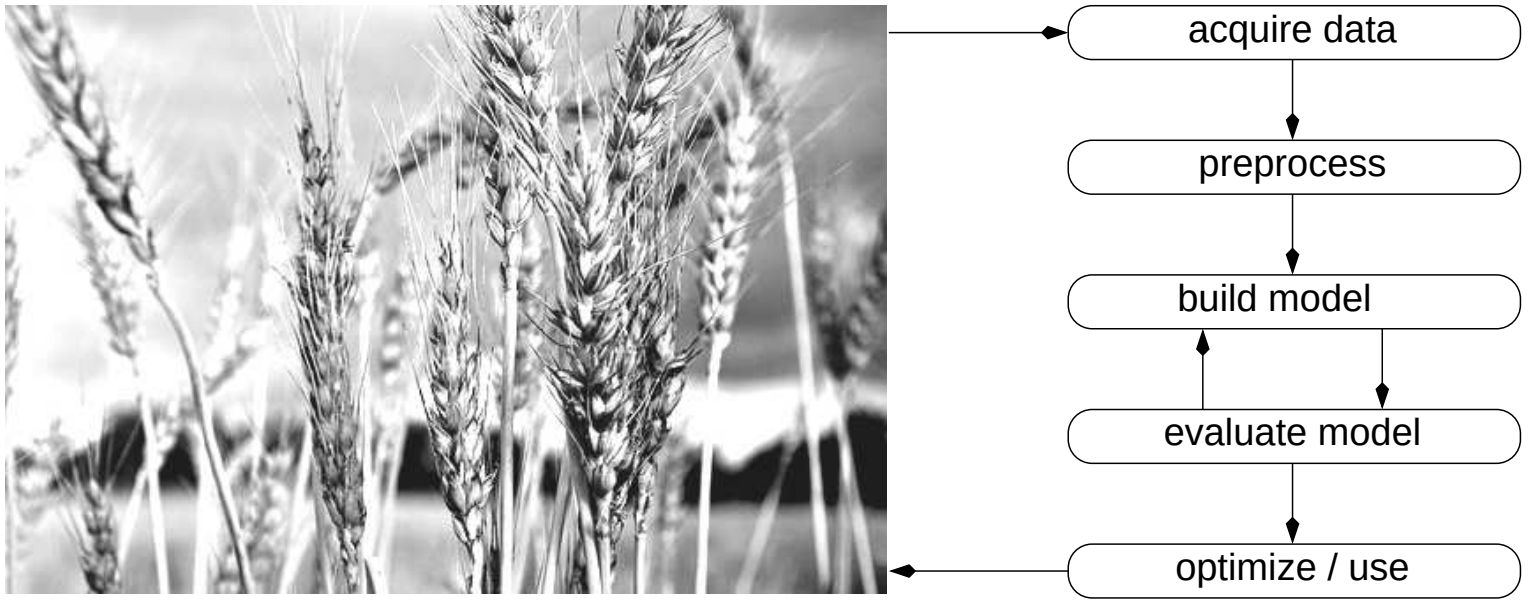


Figure: Data Mining Context

Nitrogen Fertilizer

- easy to measure when manuring
- three points into the growing season where nitrogen fertilizer is applied
- three attributes: N1, N2, N3

Vegetation Measuring

- Red Edge Inflection Point
- first derivative value along the red edge region
- aerial photography or tractor-mounted sensor
- larger value means more vegetation
- measured before N2 and N3
- two attributes: REIP32, REIP49

Electric Conductivity

- measure apparent conductivity of soil down to 1.5m
- uses commercial sensors
- one attribute: EM38

Yield

- measure yield when harvesting
- data from 2003 (previous year) and 2004 (current year)
- two attributes: Yield03, Yield04

Table: Attributes overview

Attr.	min	max	mean	std
N1	0	100	57.7	13.5
N2	0	100	39.9	16.4
N3	0	100	38.5	15.3
REIP32	721.1	727.2	725.7	0.64
REIP49	722.4	729.6	728.1	0.65
EM38	17.97	86.45	33.82	5.27
Yield03	1.19	12.38	6.27	1.48
Yield04	6.42	11.37	9.14	0.73

Splitting the data

Table: Overview: available data sets for three fertilization times (FT)

FT1	Yield03, EM38, N1
FT2	Yield03, EM38, N1, REIP32, N2
FT3	Yield03, EM38, N1, REIP32, N2, REIP49, N3

- $FT1 \subset FT2 \subset FT3$ (in terms of attributes)
- size of data sets: $\approx$ 5000 records
- For each FT*: Variable to predict is Yield04

Research Questions

- How much does *fertilization* influence current-year yield?
- Is there a correlation between data attributes that influences yield?
- How well can modeling techniques predict Yield2004?
- Can we model the data with a multi-layer-perceptron? (reproducing earlier results)
- What would be the optimal MLP's topology (number of neurons per layer)?

Data Modeling: Multi-Layer Perceptron

- Feedforward artificial neural network
- Maps a set of input data onto output data
- Mapping can be learned
- Here: predict current year's yield from current data

Data Modeling: Multi-Layer Perceptron

- Use different-size multi-layer-perceptrons for modeling
- Try to determine optimal layer size (number of hidden layers: 2)
- Compare MLPs for different data sets
- Use cross-validation and mean squared error for performance measuring

MSE plot for FT1

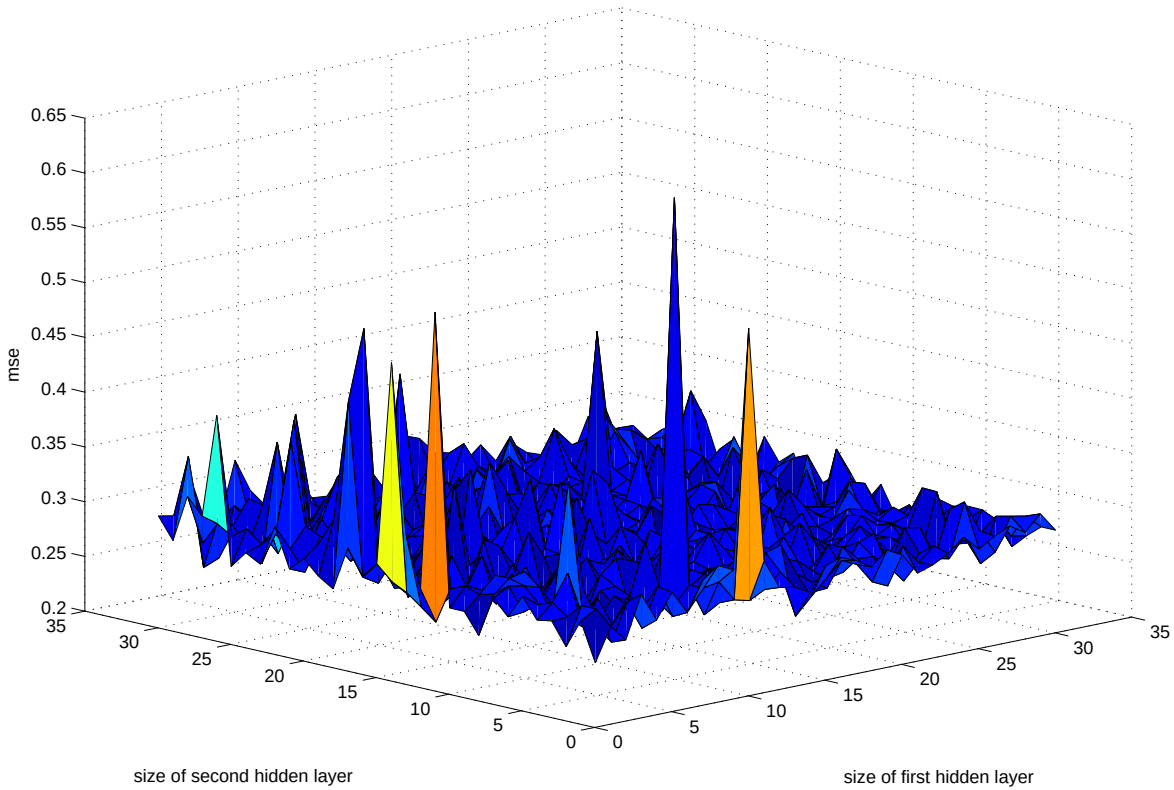


Figure: MSE for first data set

MSE plot for FT2

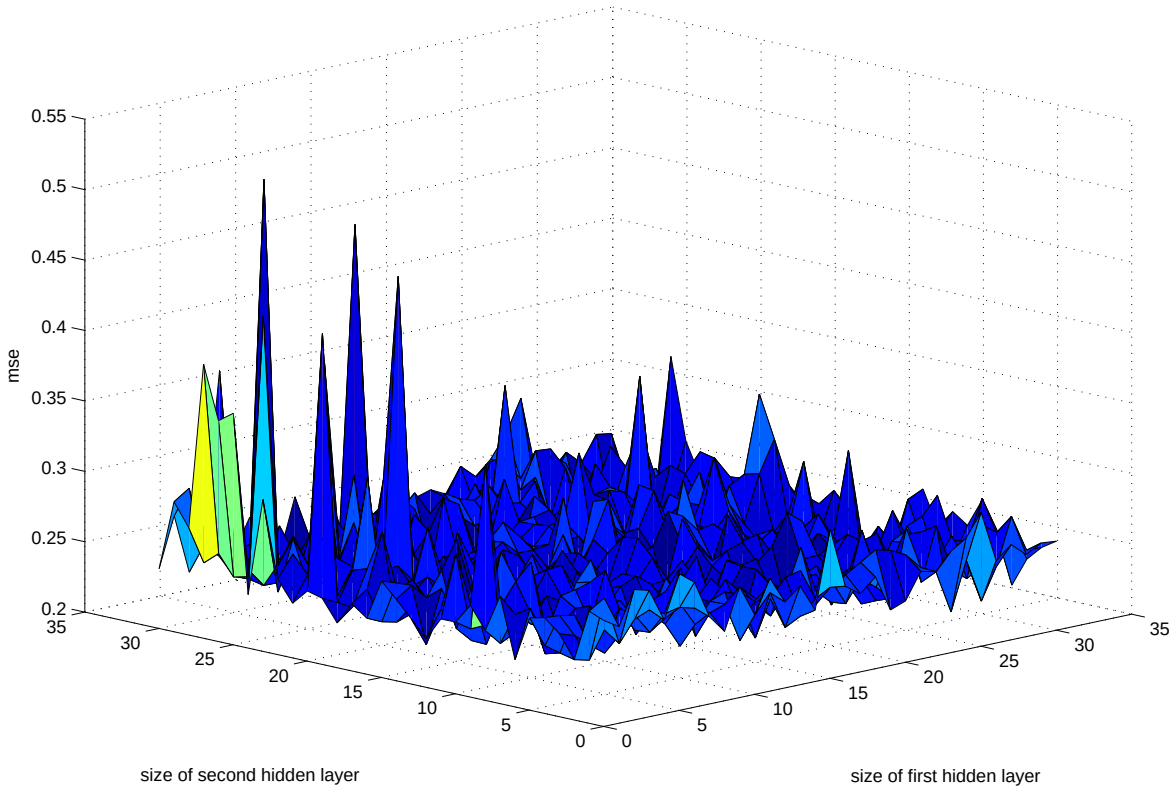


Figure: MSE for second data set

MSE plot for FT3

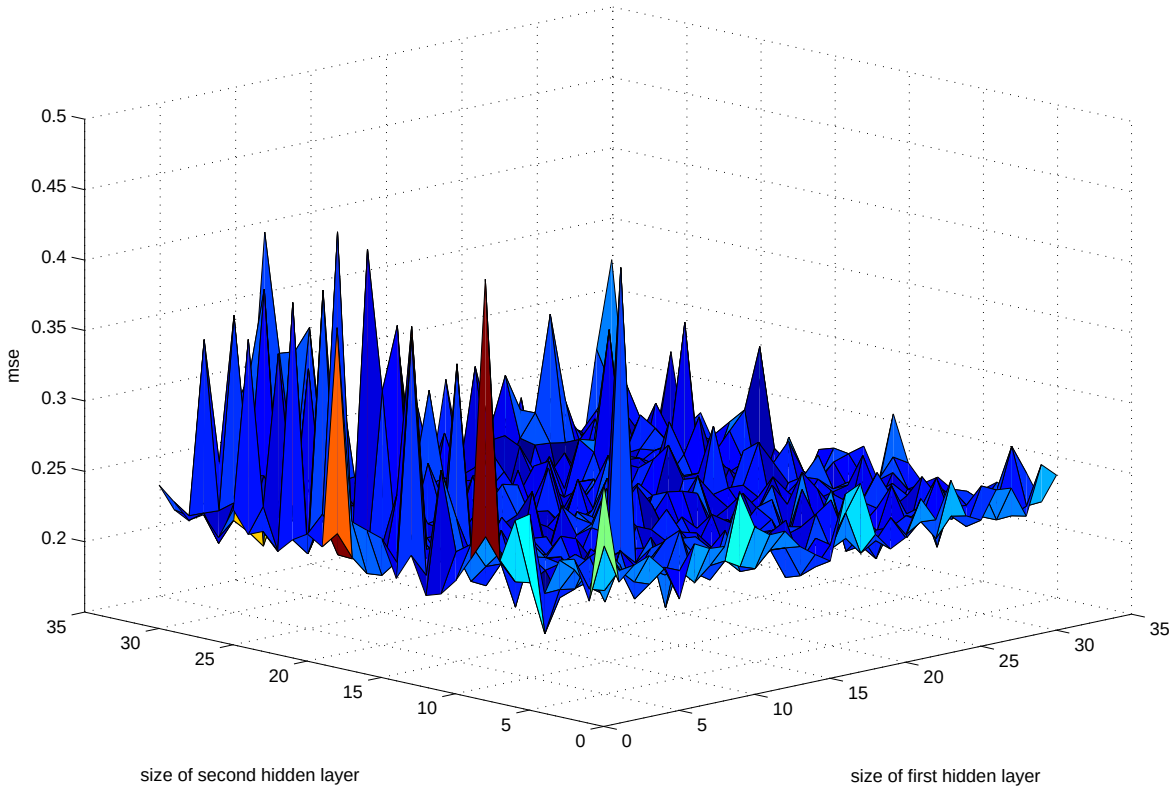


Figure: MSE for third data set

MSE difference plot between FT1 and FT2

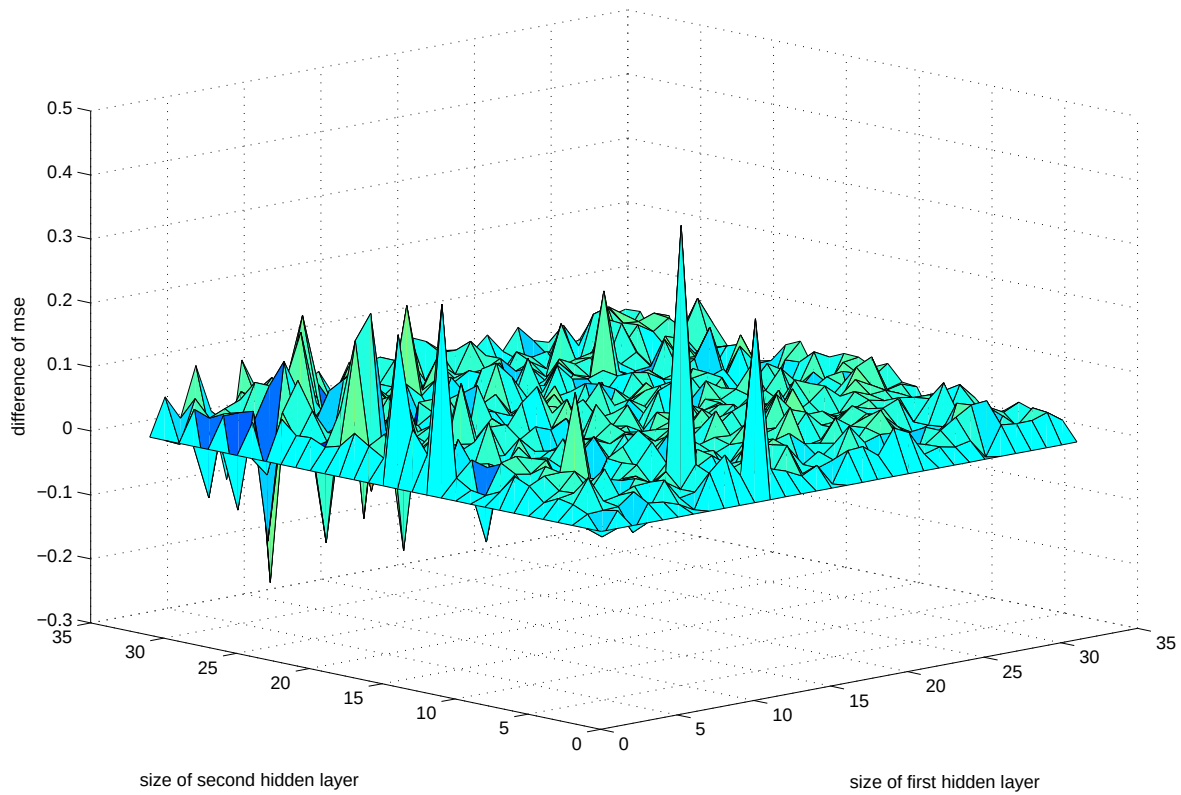


Figure: MSE difference from first to second data set

MSE difference plot between FT2 and FT3

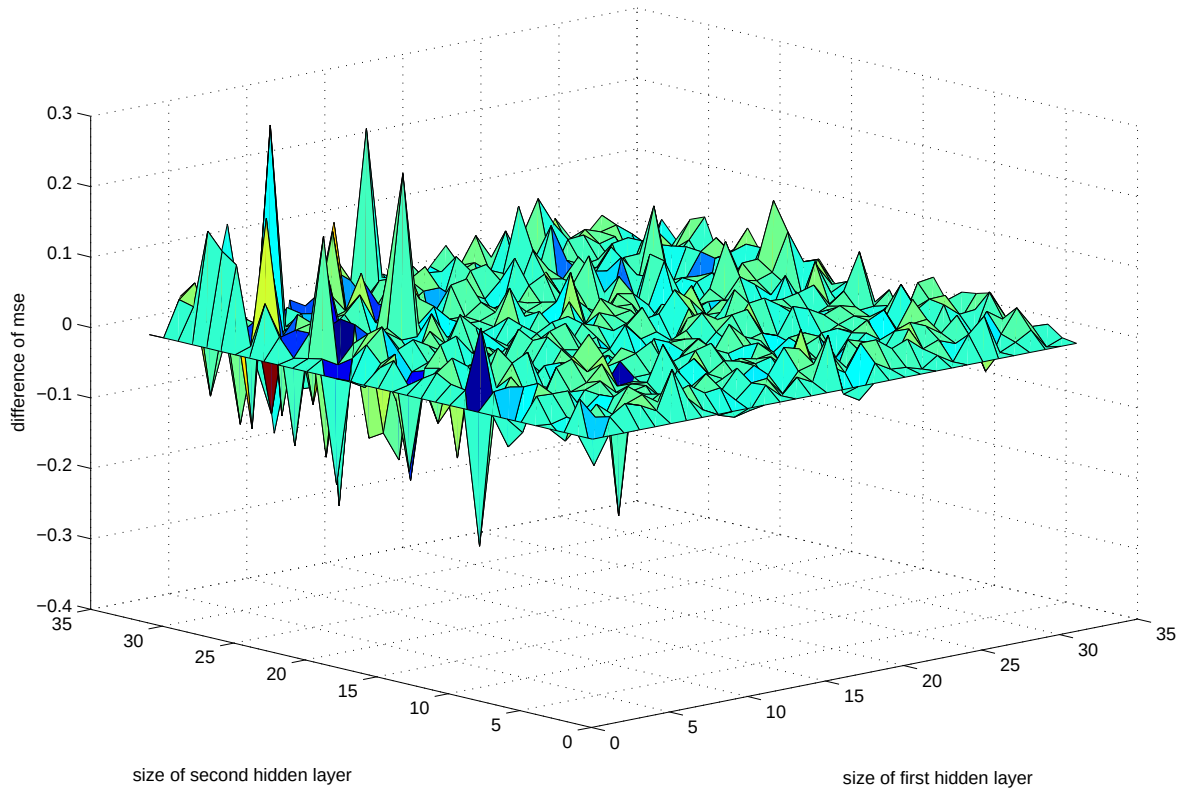


Figure: MSE difference from second to third data set

MSE difference plot between FT1 and FT3

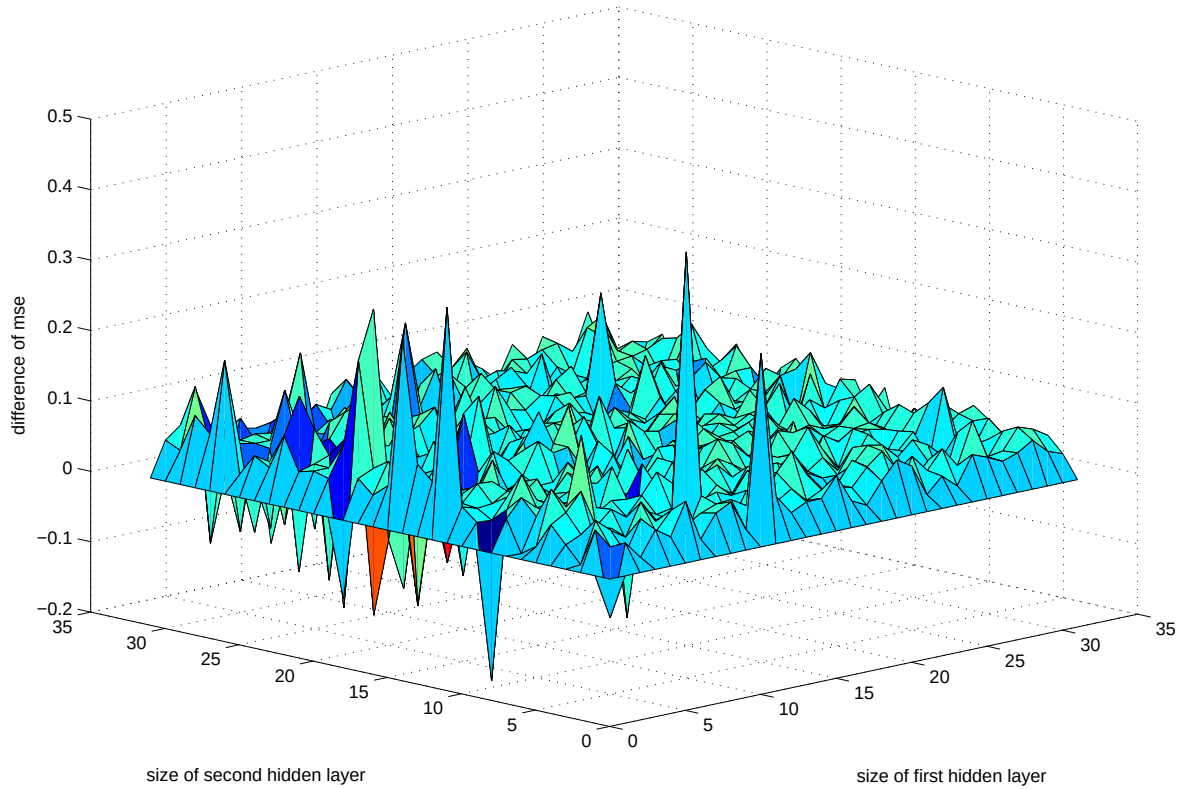


Figure: MSE difference from first to third data set

Summary MLP

- data can be modeled well with an MLP
  - low overall error
  - prediction accuracy of between 0.45 and 0.55 $\frac{t}{ha}$ at an average yield of 9.14 $\frac{t}{ha}$
- prediction gets better with more data
  - expected behaviour
  - shown by difference plots

Using the MLP predictor

- use MLP predictor to optimize fertilization
- get new data and try to understand MLP's predictions
- $\Rightarrow$ that's what's next

Data Modeling: Self-Organizing Maps

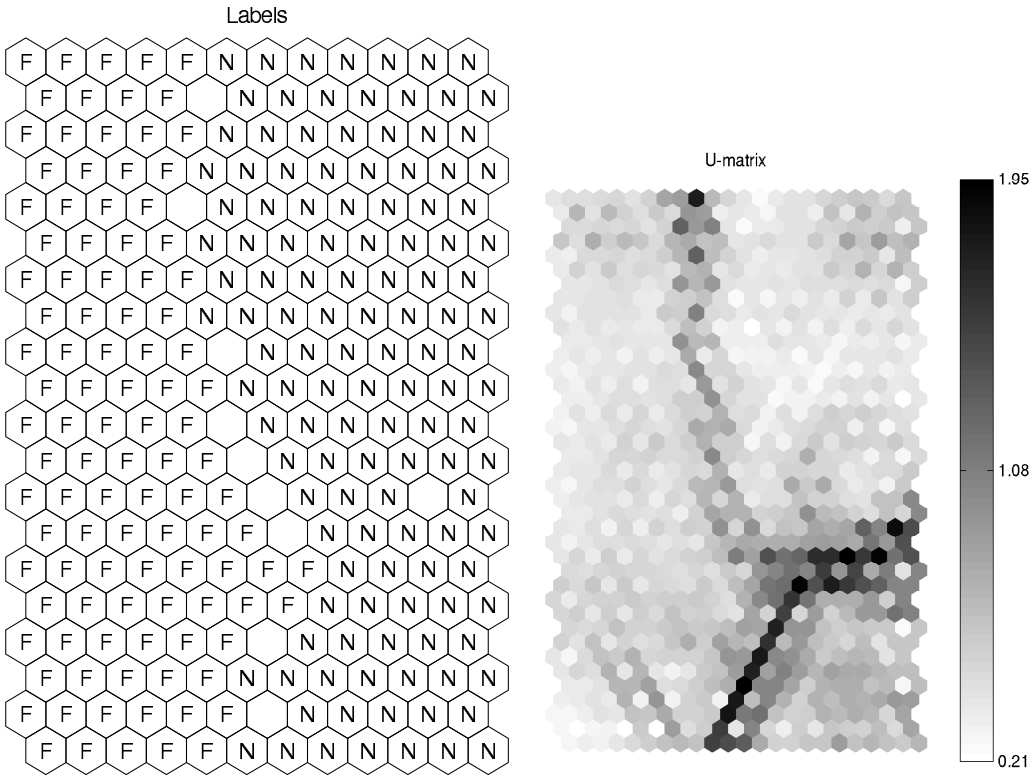
- Unsupervised artificial neural network
- Maps high-dimensional data onto two-dimensional plane
- Preserves neighborhood relations
- Here:
  - recognition of correlations
  - understanding of data
  - visualization of data

Data split

Table: Overview on available data sets for specific fertilization strategies for different fields

F131-all	YIELD05, EM38, N1, REIP32, N2, REIP49, N3, YIELD06, *fert. strategy*
F131-net	subset of F131-all where fertilization strategy is *neural network*
F330-all	YIELD05, EM38, N1, REIP32, N2, REIP49, N3, YIELD06, *fert. strategy*
F330-net	subset of F330-all where fertilization strategy is *neural network*

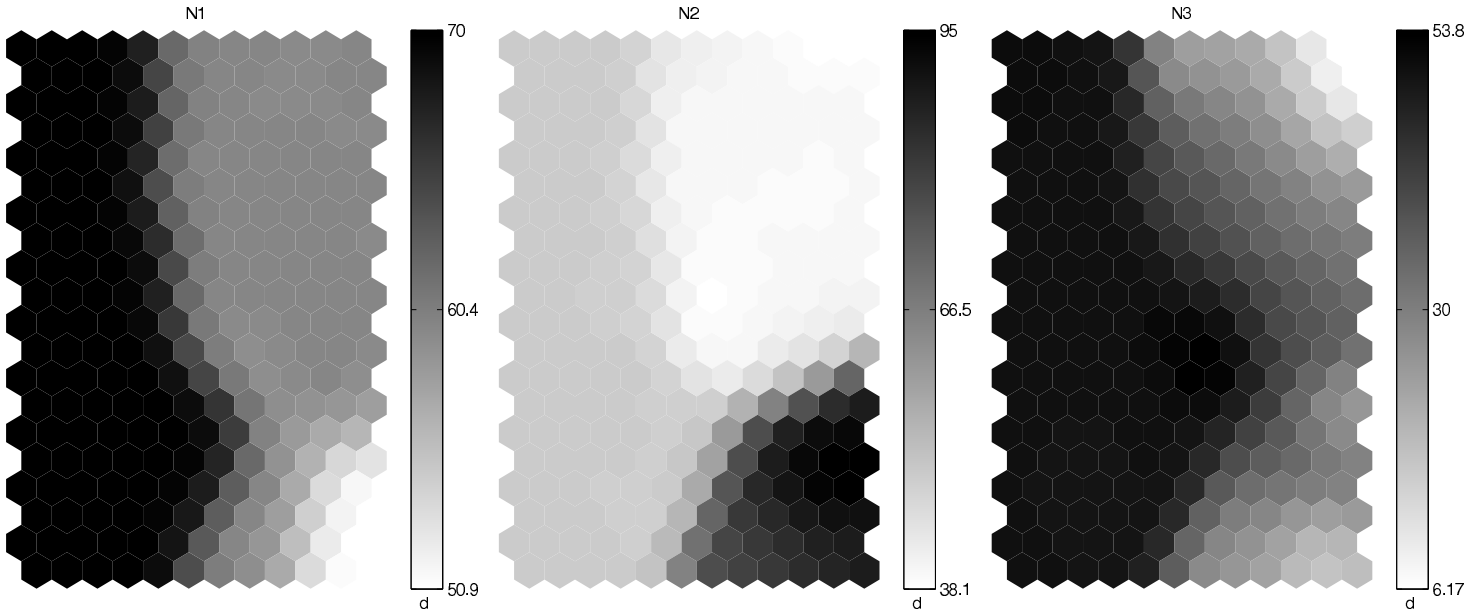
Results for F131-all, Labels/U-Matrix



(a) Labels (b) U-Matrix

Figure: F131-all, U-Matrix and Labels

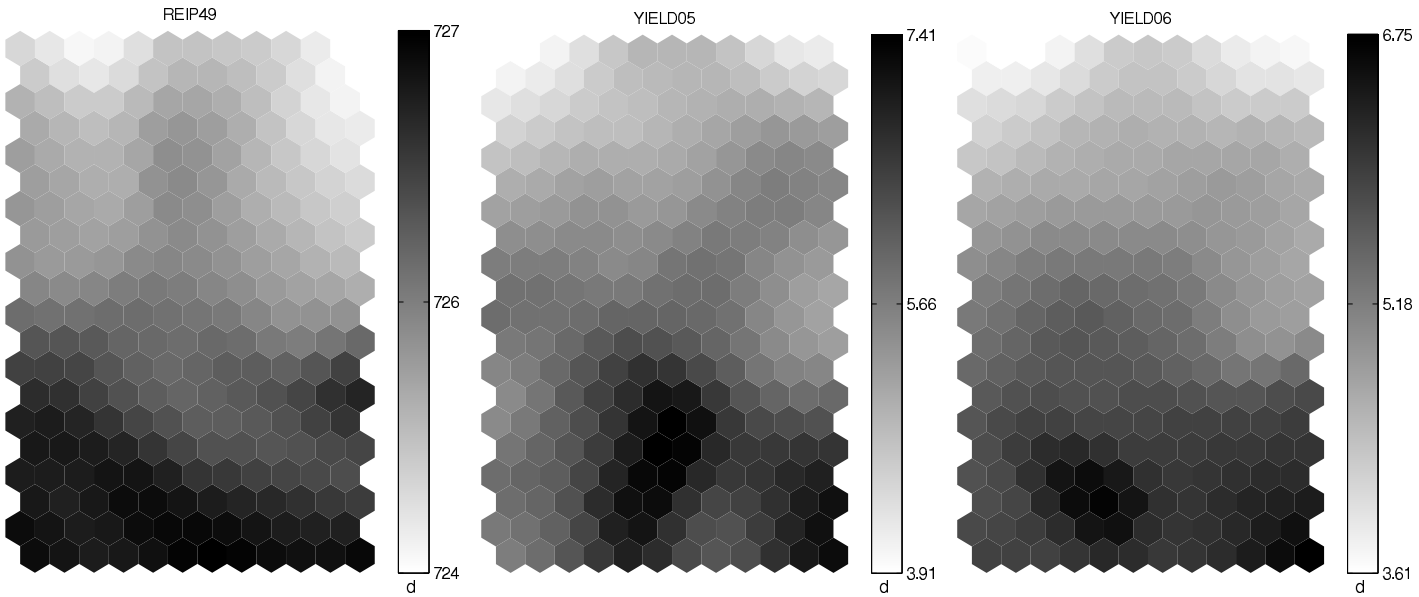
Results for F131-all, Nitrogen



(a) N1 (b) N2 (c) N3

Figure: F131-all, N1, N2, N3

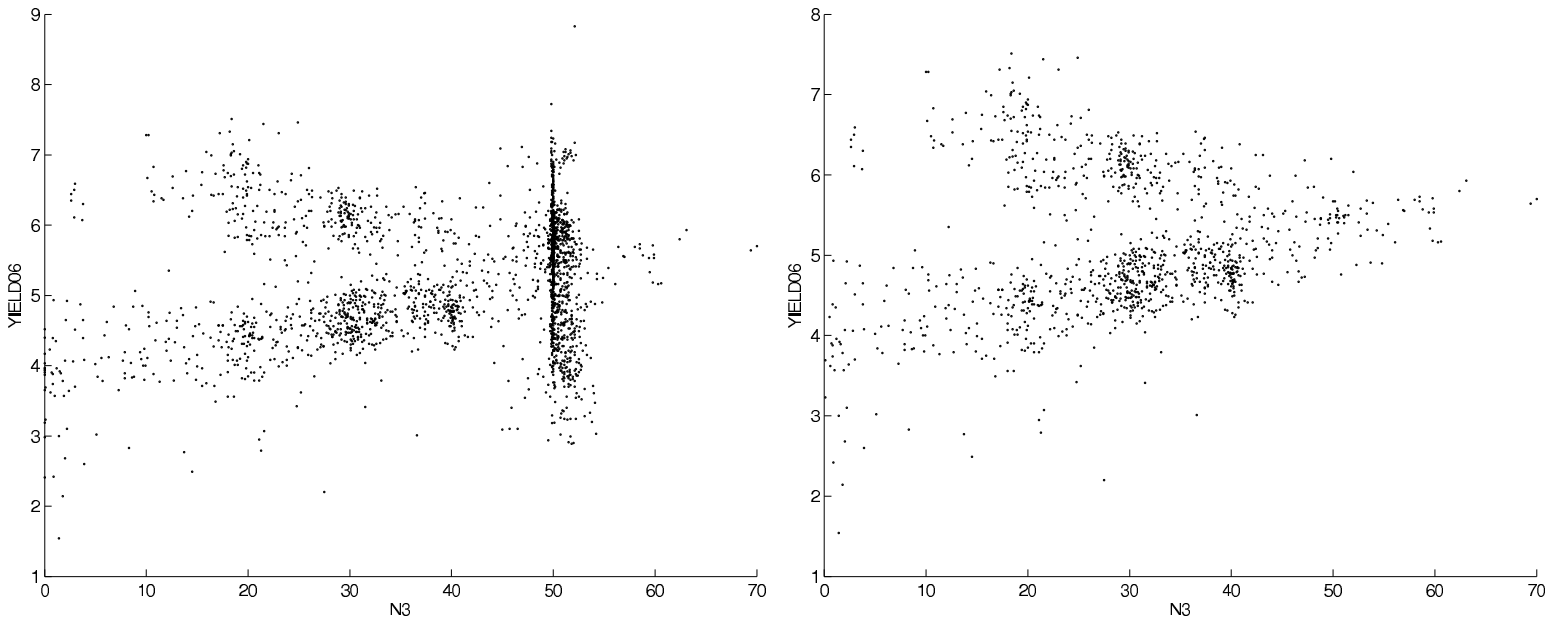
Results for F131-all, REIP, Yield



(a) REIP49 (b) YIELD05 (c) YIELD06

Figure: F131-all, REIP49 vs. YIELD05 vs. YIELD06

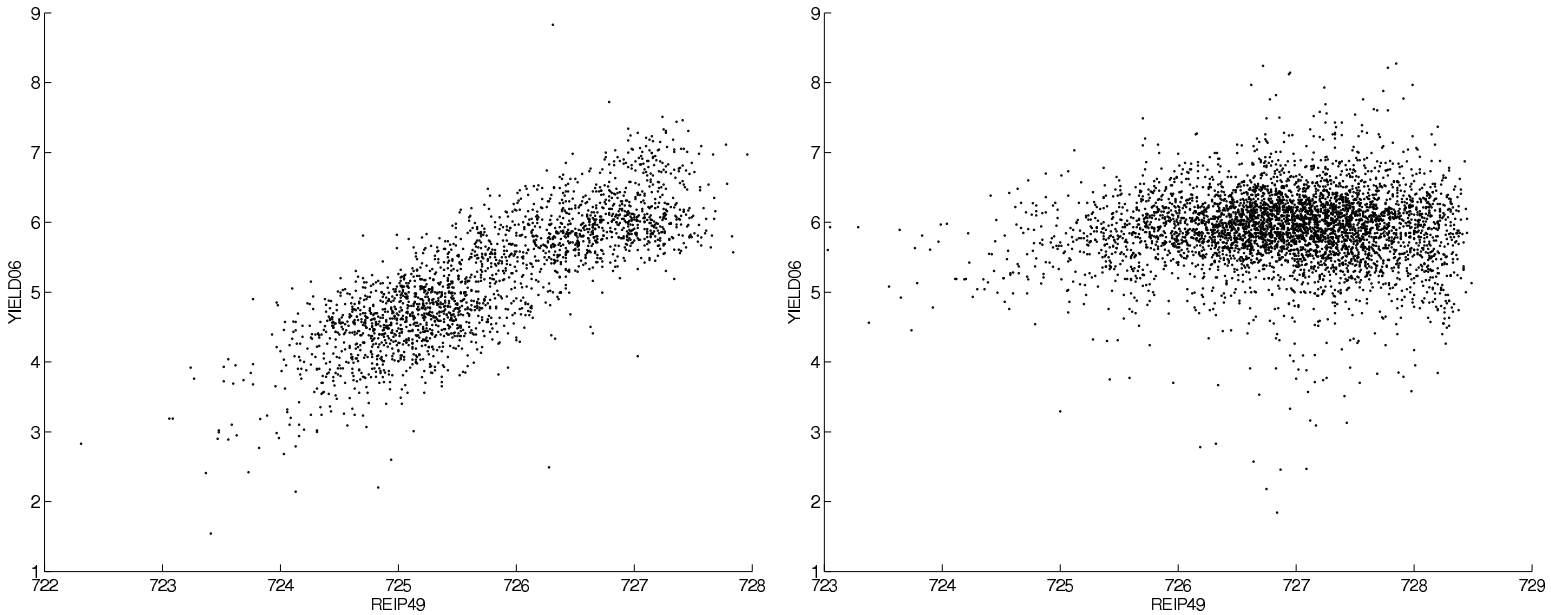
Results for F131-all, correlation



(a) N3 vs. YIELD06, F131-all (b) N3 vs. YIELD06, F131-net

Figure: F131-all, correlation between N3 and YIELD06

Results for F131-all, correlation



(a) REIP49 / YIELD06, F131 (b) REIP49 / YIELD06, F330

Figure: F131-all, correlation between REIP49 and YIELD06

Summary SOM

- very good tool for visualizing the data
- helps finding correlations easily without correlation plots
- helps finding attributes that can be used for predicting yield

Further Work

- evaluate further modeling techniques
- compare techniques on further (already available) data sets
- generate optimized decision rules for, e.g. usage of fertilizer or pesticides

Questions / Discussion

- Questions?