# Kohonen 3.0

*Improvements of the kohonen R package for application of self-organising maps on large data sets*

Ron Wehrens and Johannes Kruisselbrink

12 December 2018

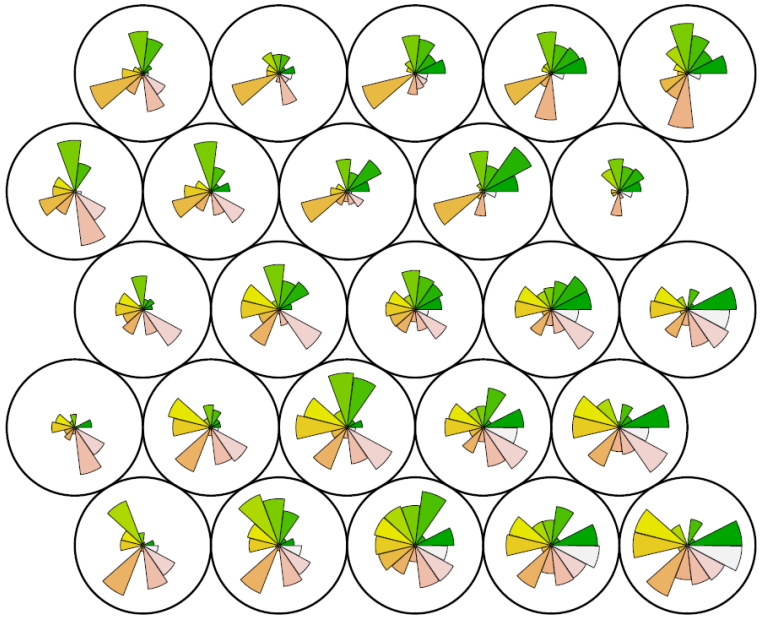
Biometris, Wageningen UR

WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

# Self-organising maps (SOMs)

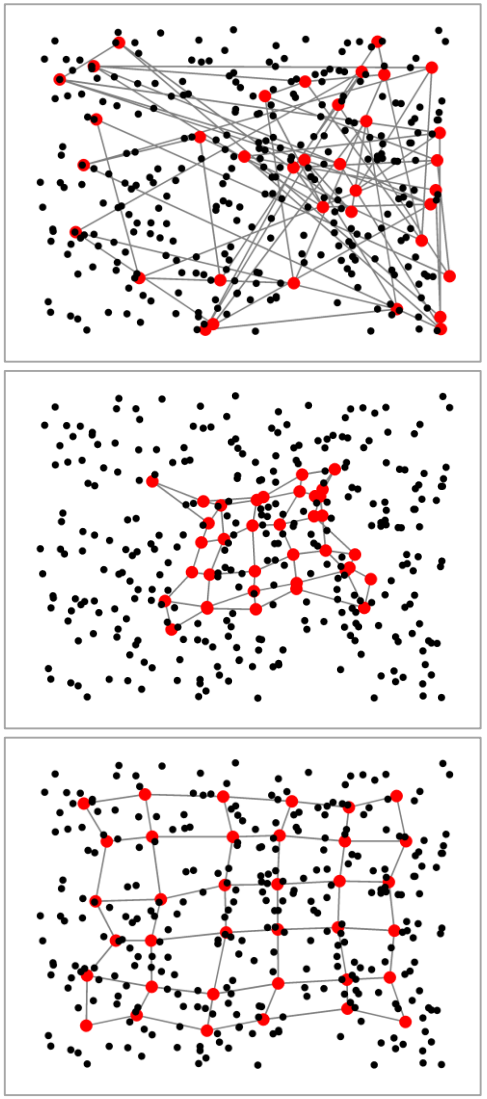
- Unsupervised learning/clustering
- Discrete mapping of high-dimensional data onto two dimensions
- Spatially smooth version of k-means



Kohonen T (1995). Self-Organizing Maps. Springer-Verlag, Berlin.

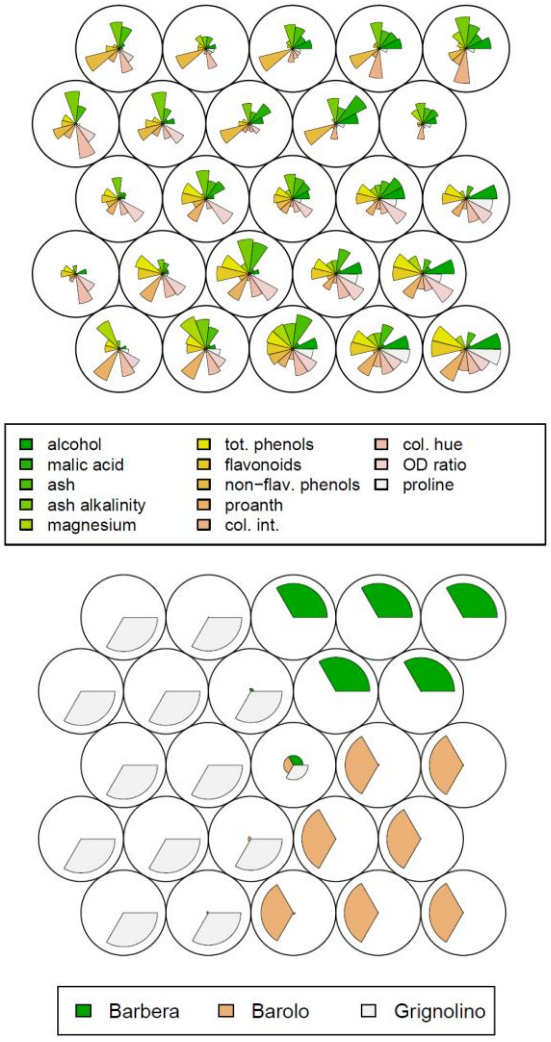
# SOM training

- Initialize random map
- For all points in dataset
  - Choose data point from dataset
  - Find best matching unit (BMU) (the unit closest to the data point)
  - Pull BMU towards data point
- Pull neighbouring units of BMU towards data point
- Repeat until convergence / stopping criterion met



# R package kohonen 2.x

- R package for training SOMs
- Time-critical elements in C
- Various methods for training SOMs and mapping new data
- Super-SOM implementation for multiple data layers




Wehrens R, Buydens LMC (2007). “Self- and Super-organizing Maps in R: The kohonen Package.” Journal of Statistical Software, Volume 21, Issue 5.

# Kohonen 3.0

- SOM training algorithms in C++ using Rcpp
- New (parallel) batch training algorithms
- User-definable distance functions

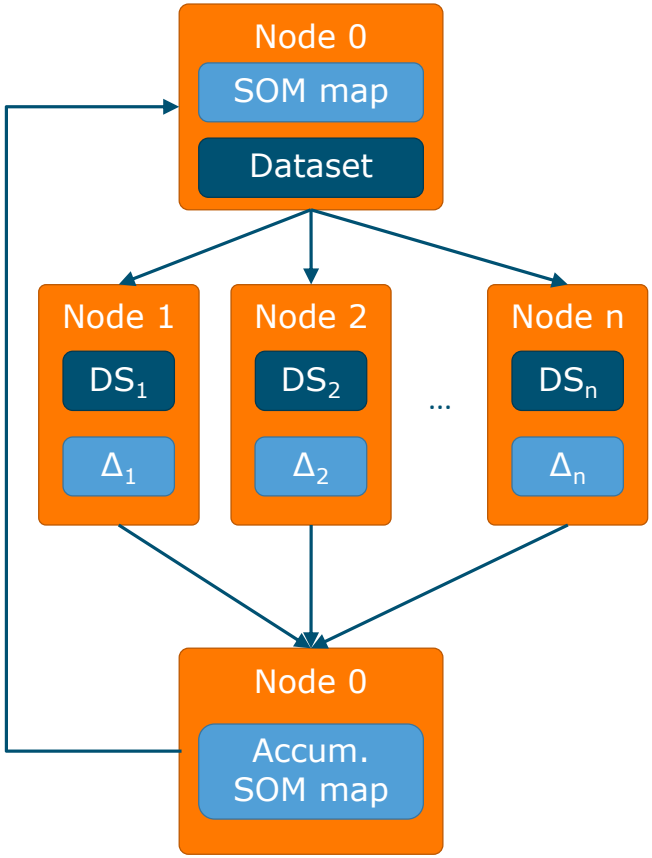
→ Big performance gains (memory and computation time)

→ More flexible software

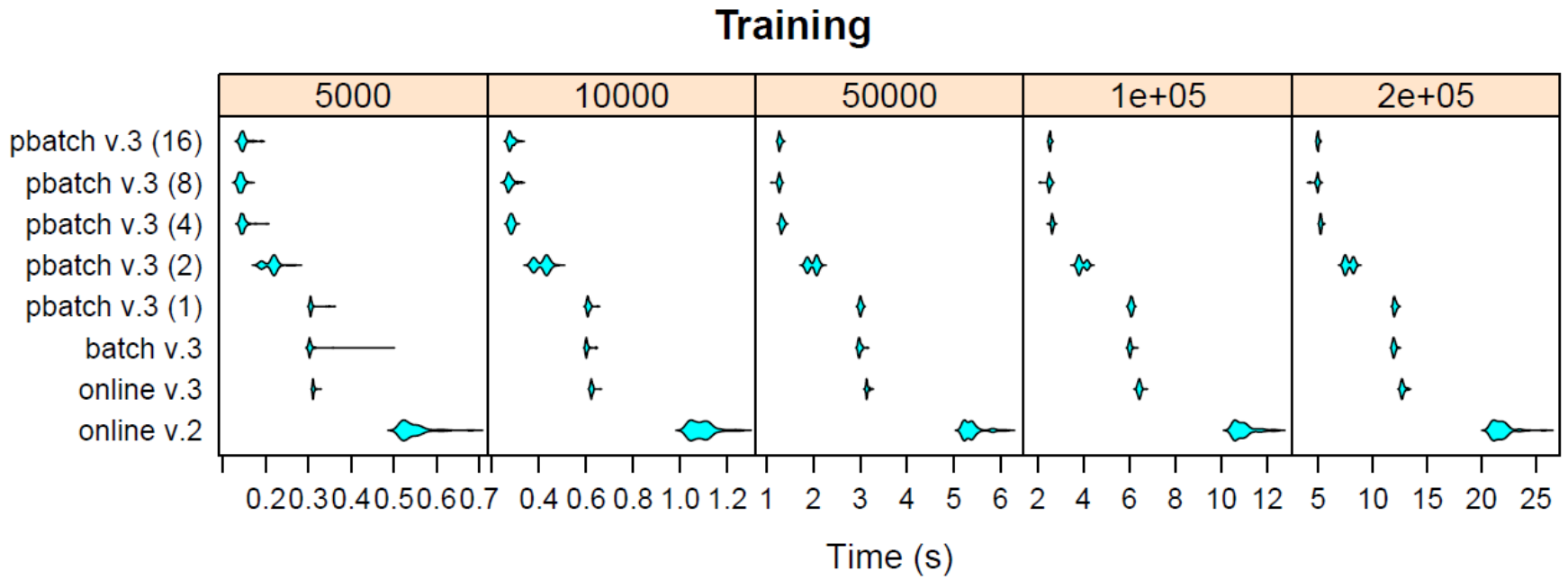
Wehrens R, Kruisselbrink J (2018). “Flexible Self-Organizing Maps in kohonen 3.0.” Journal of Statistical Software, Volume 87, Issue 7.

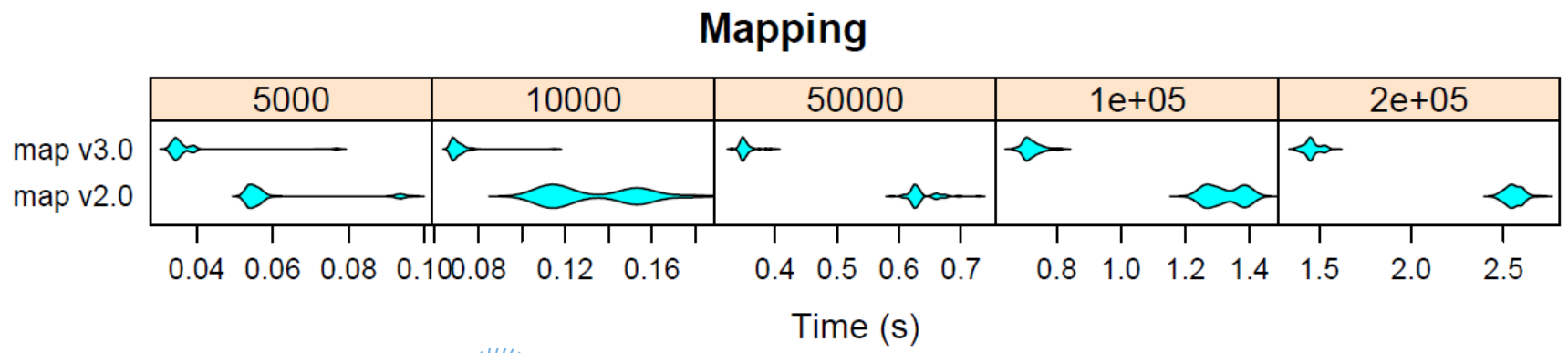
# Batch SOM algorithm

- Initialize random map
- Split up dataset in n subsets
- Compute unit position updates for each subset (in parallel)
- Accumulate unit position updates and update unit positions
- Repeat until convergence/stopping criterion met



# Benchmarking – Training and mapping times

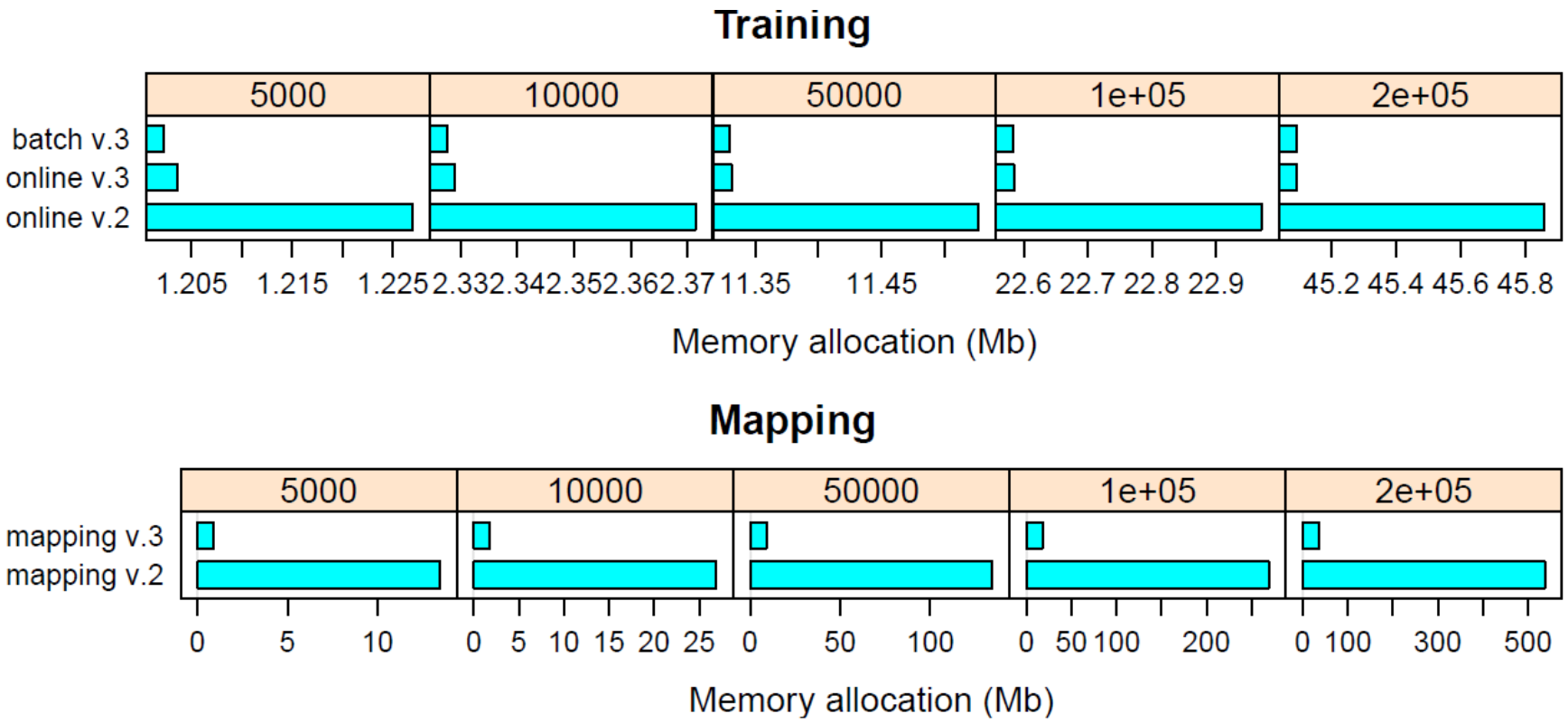



WAGENINGEN
UNIVERSITY & RESEARCH

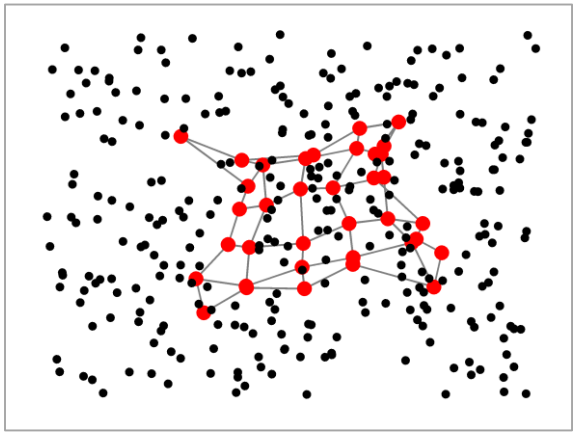
100years
1918 — 2018

# Benchmarking – Memory usage



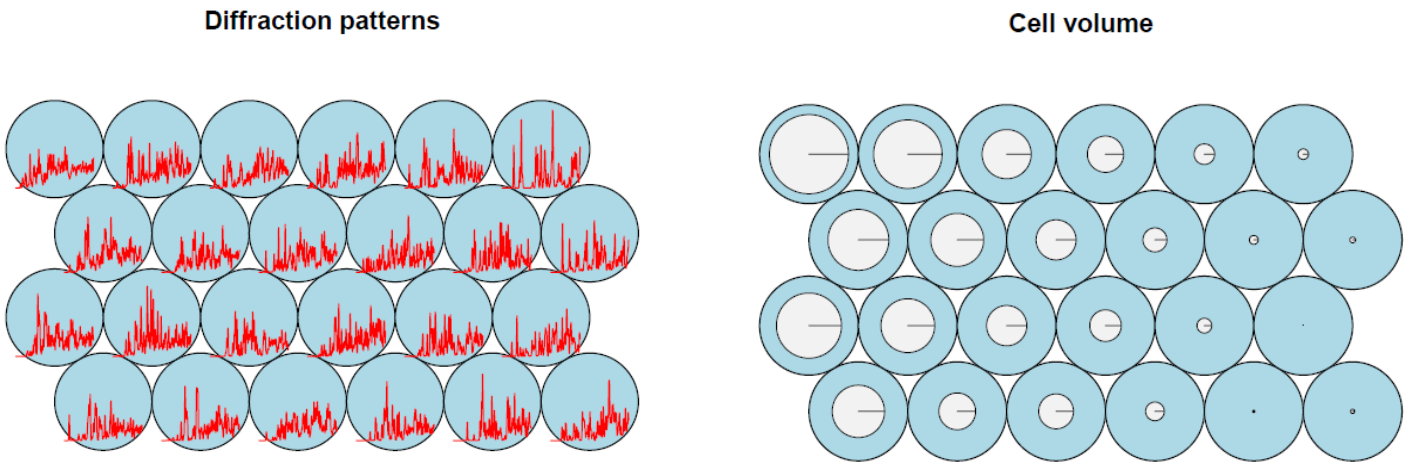
# User-definable distance functions

- BMU: the unit closest to the current data point
- Which distance metric?
- Built-in types: Euclidean, sum-of-squares, Manhattan, Tanimoto
- For some data types, other metrics may be more appropriate
    - Site counts (Bray-Curtis)
    - Time series (DTW)



# Application of custom distance functions

- Classification of X-ray powder diffractograms providing information on the crystal cell structure
- Custom distance metric based on weighted cross-correlation (WCC)



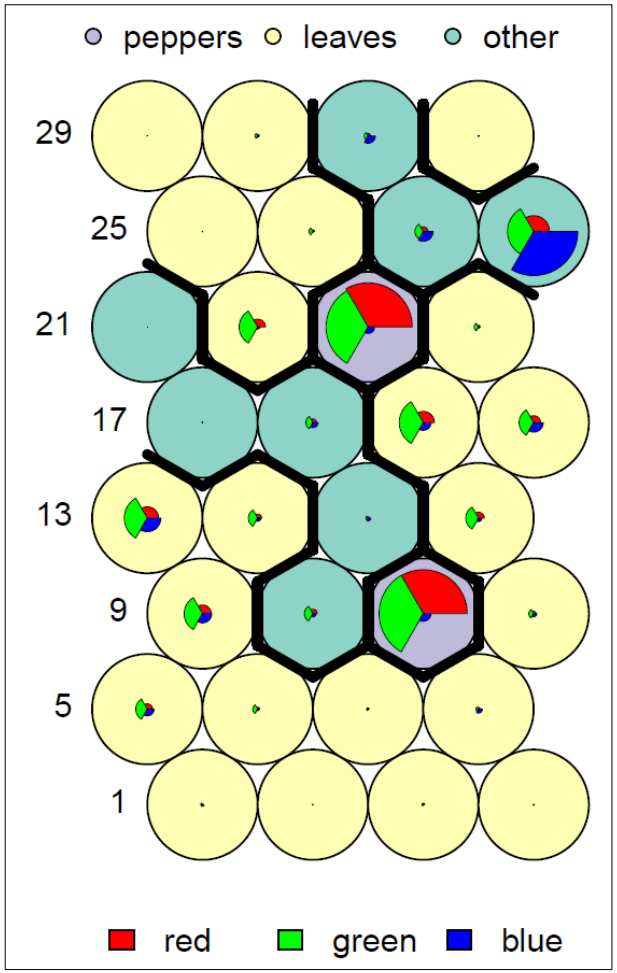

Wehrens R, Melssen WJ, Buydens LMC, De Gelder R (2005). “Representing Structural Databases in a Self-Organizing Map.” Acta Crystallographica B, B61, 548–557.

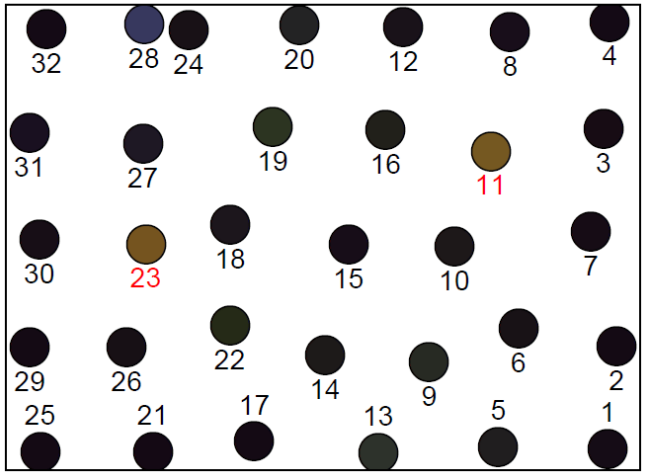
# Example: segmentation of pepper images

- Cluster pixels
- Reduce dimensions
- Identify homogeneous regions
- Feature extraction

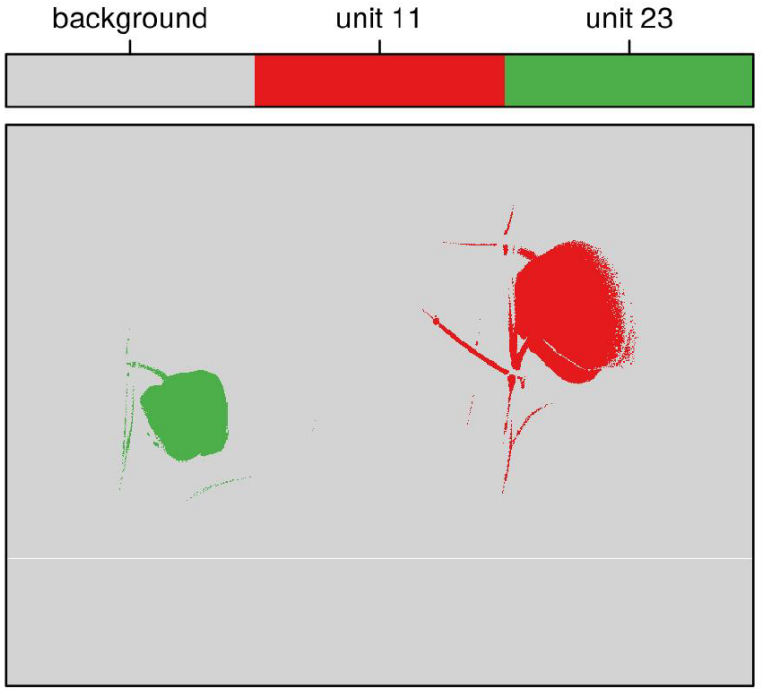
Image from:
Barth R et al. “Data Synthesis Methods for Semantic Segmentation in Agriculture. A Capsicum Annuum Dataset.” Computers and Electronics in Agriculture, 144, 284–296.

# Training on pixel colour and coordinates (1)



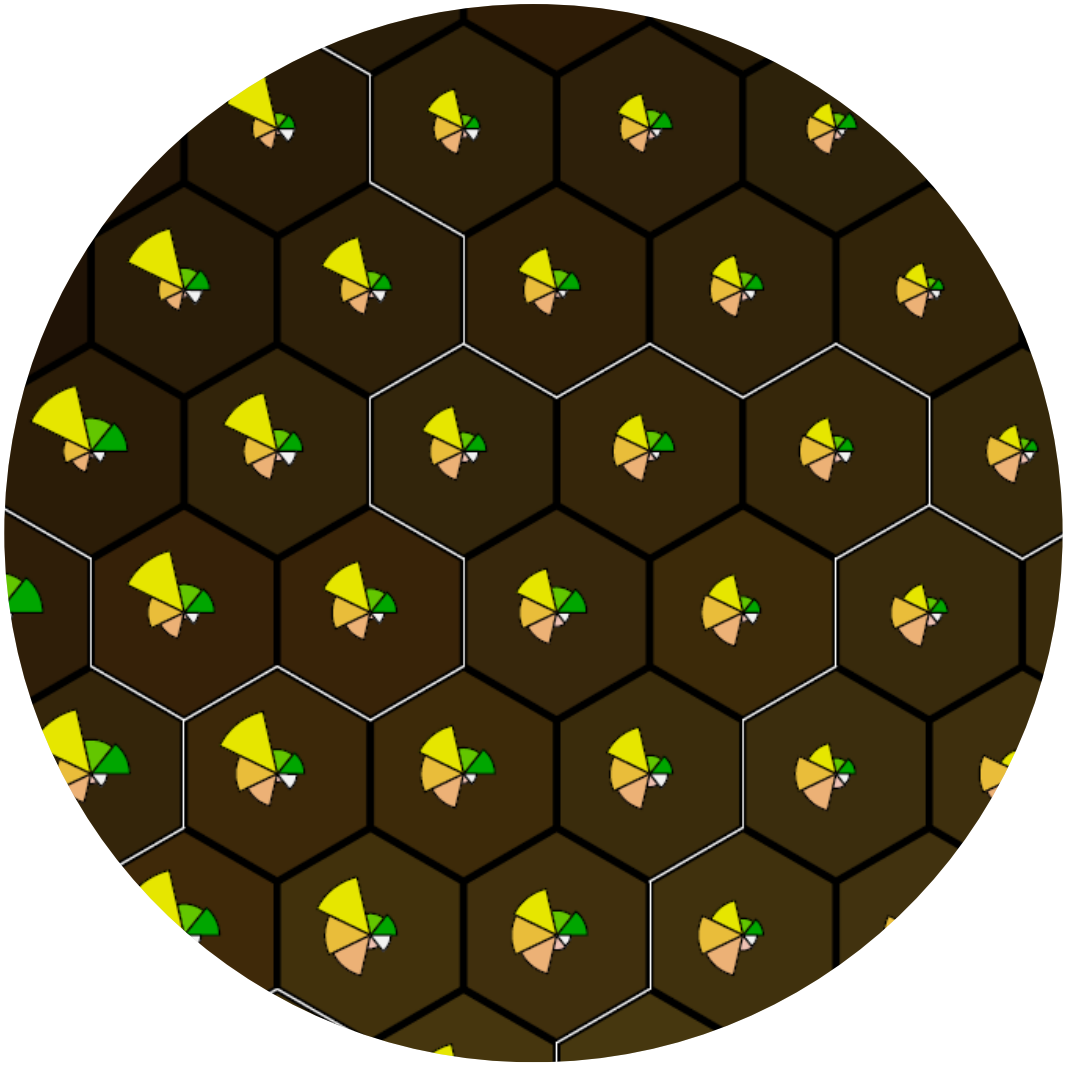
# Training on pixel colour and coordinates (2)



# Conclusion

- SOMs: unsupervised learning/clustering to reveal structure in large/complex datasets
- R package Kohonen: provides a complete implementation
- Kohonen 3.0:
  - Big performance gains (memory and computation time)
  - Support for custom distance functions
  - Online and (parallel) batch learning

# Thank you



WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018