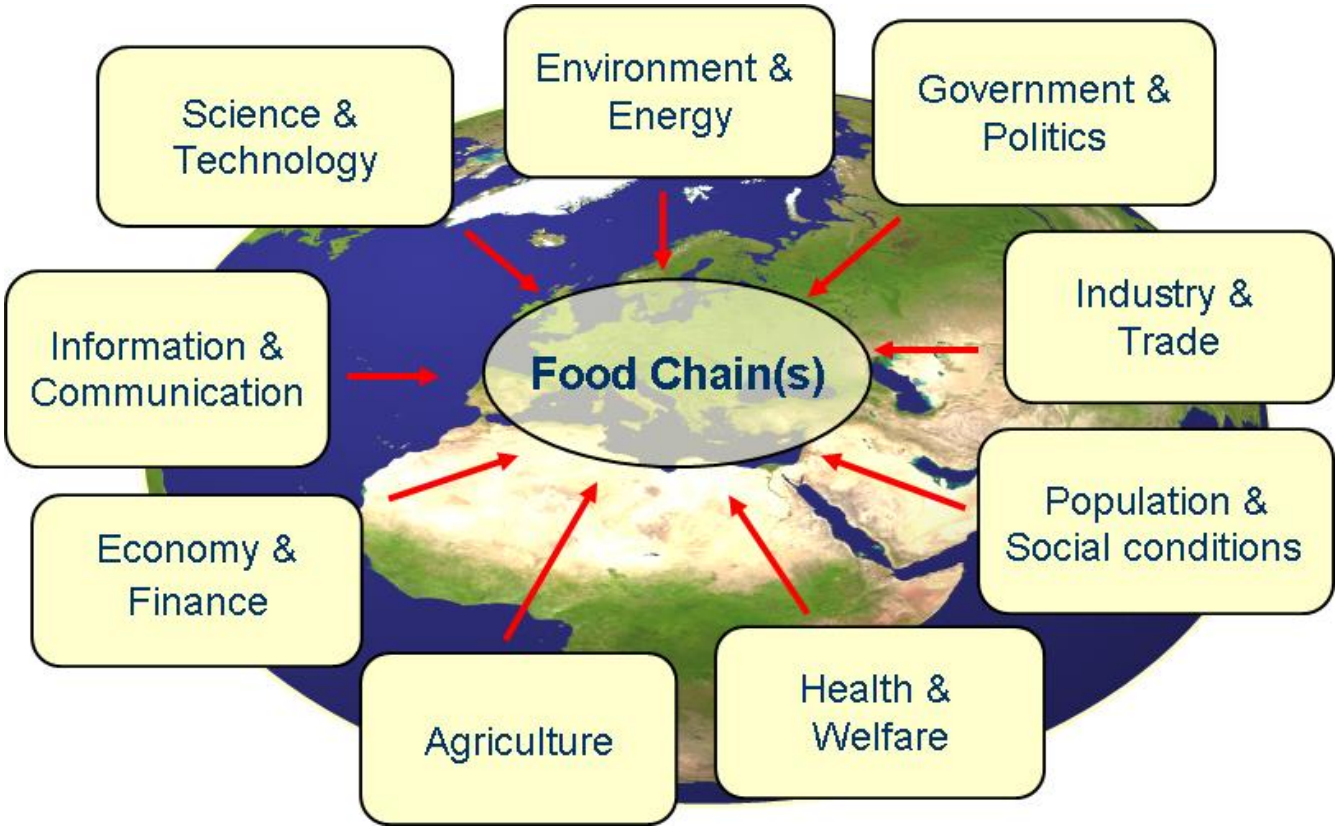
Developing new data driven science using Bayesian Network approach and machine learning methodologies

Hans Marvin, Anand Gavai & Yamine Bouzembrak

(RIKILT WUR)

WAGENINGEN
UNIVERSITY & RESEARCH

Many factors (drivers) have direct/ indirect influence on the development of a food safety risk



⇒ System approach needed to take all factors into account

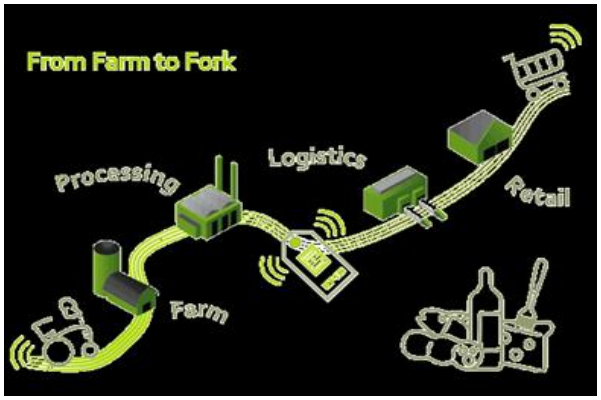
WAGENINGEN
UNIVERSITY & RESEARCH

Source: Wim Ooms VWA, 2006

Potential system approach

Data of drivers

Expert knowledge



Chain analysis

Method needed that can:

- integrate expert knowledge and data,
- handle data and knowledge gaps,
- use a variety of data sources of divers nature.

Bayesian Networks?

WAGENINGEN
UNIVERSITY & RESEARCH

What is Bayesian Network?

Thomas Bayes (1702-1761)

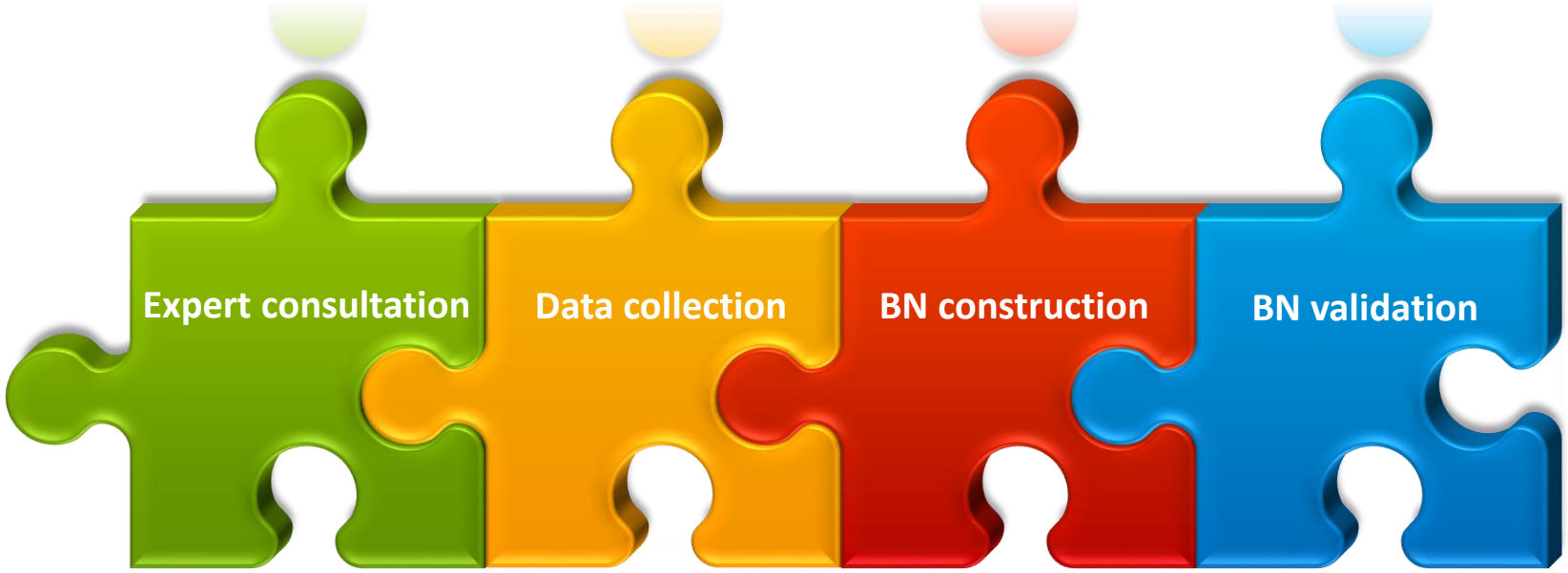
Bayesian networks (BNs) belong to the family of probabilistic graphical models. BNs combine principles from graph theory, probability theory, computer science, and statistics.

Why BN model:

- BN requires less parameters than the conventional naïve method
- BN takes into account a number of quantitative and qualitative criteria under uncertainty
- BN can combine expert and domain knowledge that allows flexible inference even with partial and limited information

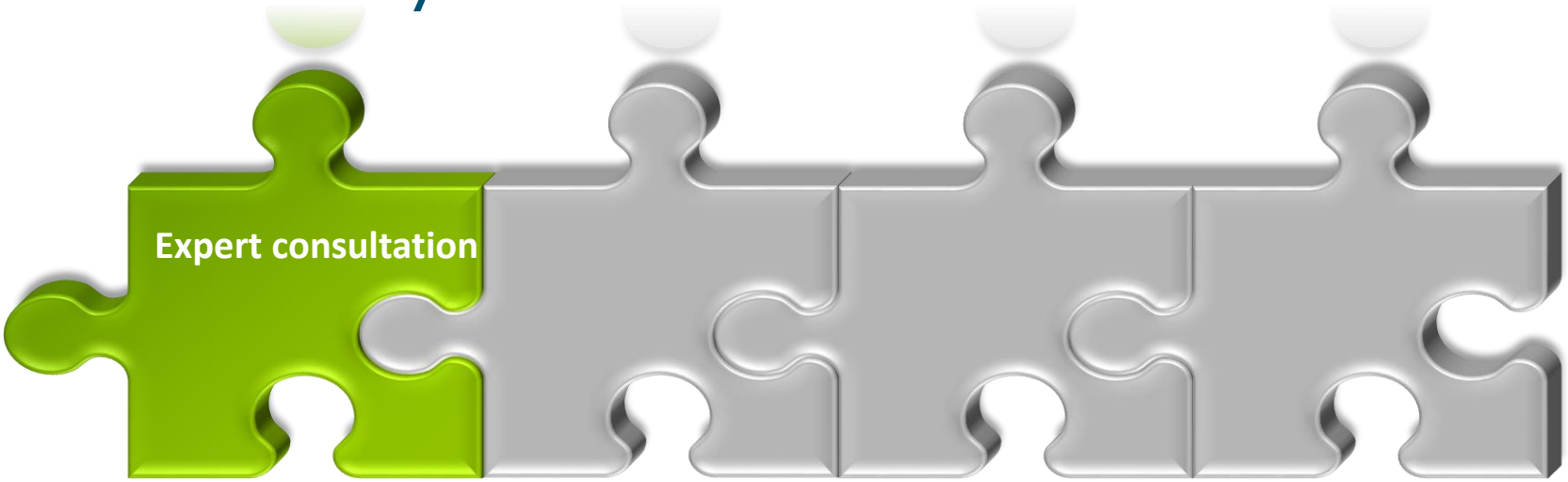
BN prediction model for hazard contamination in feed of dairy cows in the Netherlands

Steps in the development



WAGENINGEN
UNIVERSITY & RESEARCH

BN prediction model for hazard contamination in feed of dairy cows in the Netherlands



Expert consultation

Which factors have effect on the development of food safety hazards in feed grown or produced in the Netherlands

- Climate (rain, temperature)
- Use of agrichemicals

BN prediction model for hazard contamination in feed of dairy cows in the Netherlands

Data collection

- Collect monitoring data of dairy feed developed/produced in the the Netherland as reports in Quality Program for Agricultural Products of the Netherlands (KAP)
- Link each individual analytical measurement (case) to climate data (17 variables) obtained from NOAA and to the use of agrichemicals obtained from FAOSTAT (herbicides, insecticides, fungicides, rodenticides, and pesticides)

BN prediction model for hazard contamination in feed of dairy cows; data collected

The most common food safety hazard categories and products reported in KAP (2000-2013) for dairy feed in the Netherlands.

The most common food safety hazard categories		The most common products	
Hazard category	**%**	**Product**	**%**
Mycotoxins	34	Grass	11
Industrial contaminants	24	Maize	10
Biotoxins	12	Alfalfa	7
Heavy metals	10	Feed bovine	7
Radiation	8	Wheat	7
Pesticides	7	Soy	5
Veterinary drugs	6	Feed ruminants	4

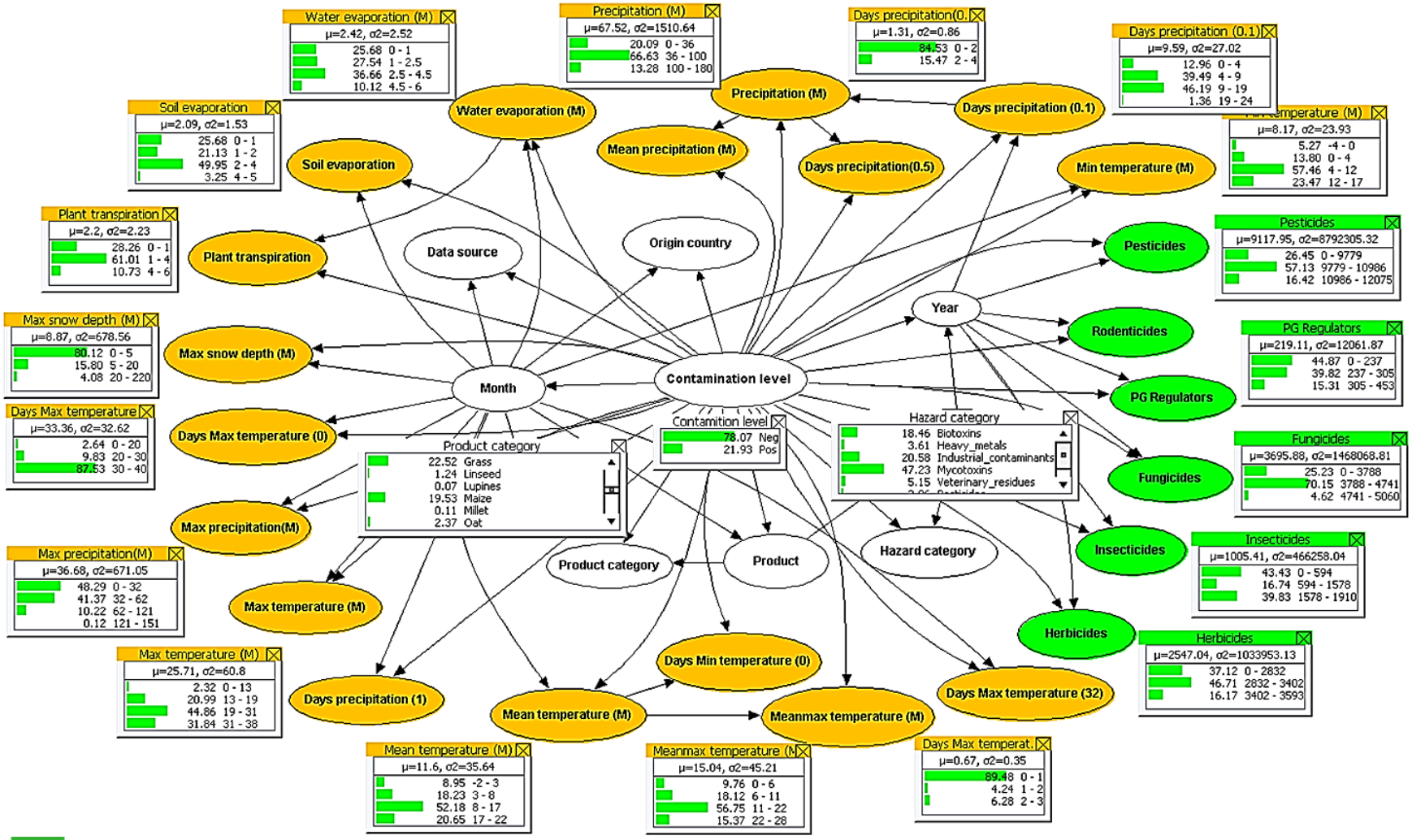
- Focus on products grown/produced in the Netherlands; in total 54,806 individual analytical results available for this study.
- Each analytical results was linked to climate data (17 variables) and the use of agrichemicals (6 variables) in the Netherlands.

BN prediction model for hazard contamination in feed of dairy cows in the Netherlands

BN construction

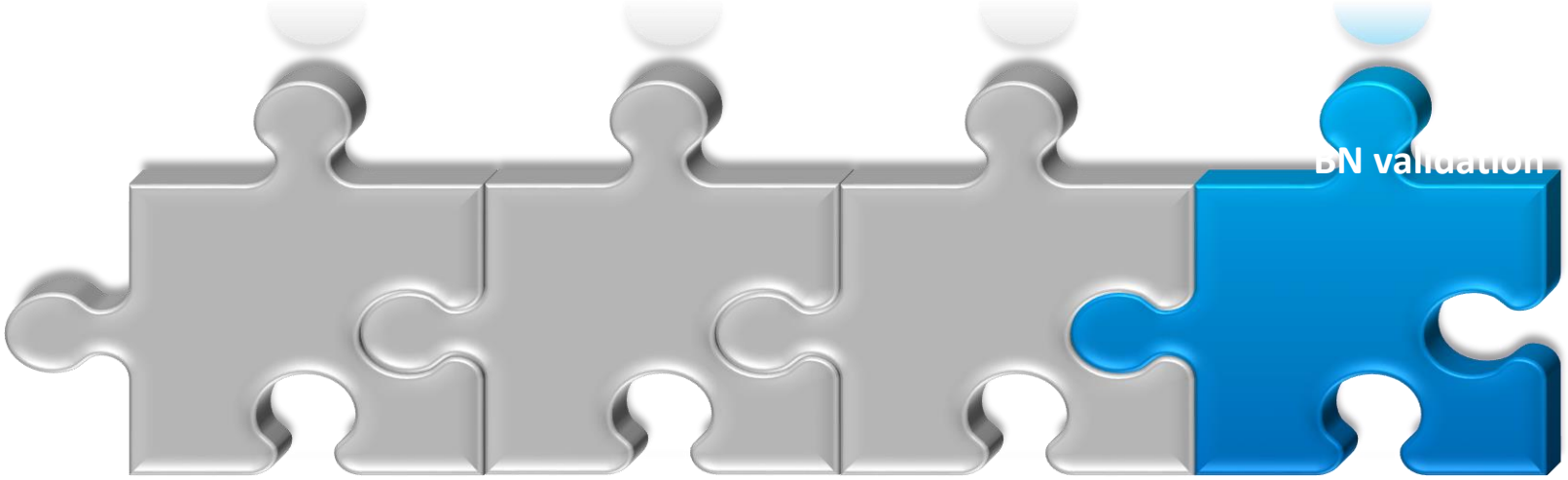
- The BN model was developed with 80% of the collected cases.
- We used machine learning algorithm (i.e. EM-algorithm) to construct the BN model.

BN for the prediction of contamination level of food safety hazards in dairy feed in the Netherlands



WAGENINGEN
UNIVERSITY & RESEARCH

BN for the prediction of contamination level of food safety hazards in dairy feed in the Netherlands



BN validation

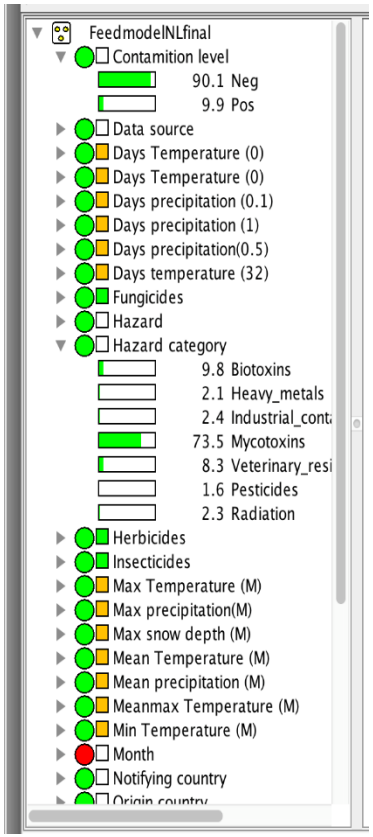
- Validation performed with 20% of the cases, randomly selected from the total set
- Accuracy of prediction: 90.3%

Scenario:

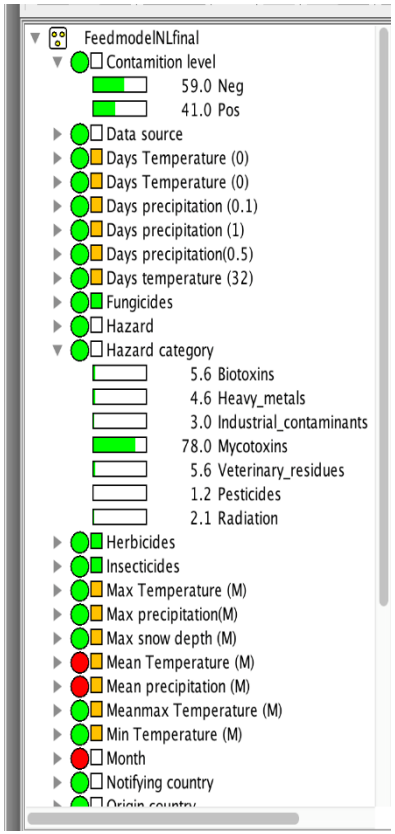
June, warm (high mean temp.) and wet (high mean precipitation)

June: warm (high mean temp.) and wet (high mean precipitation)

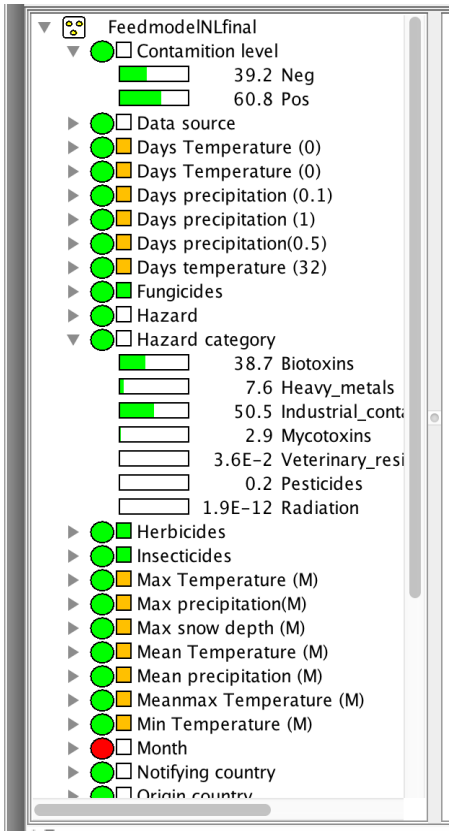
Maize (normal)



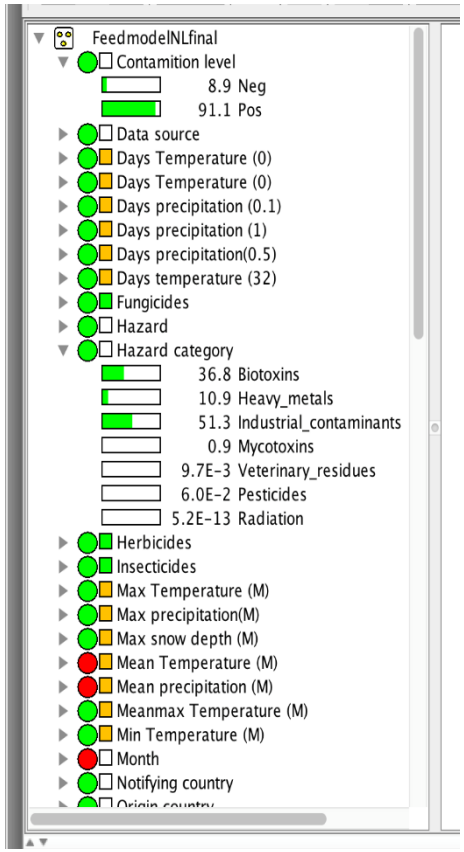
Maize (scenario)



Alfalfa (normal)



Alfalfa (scenario)

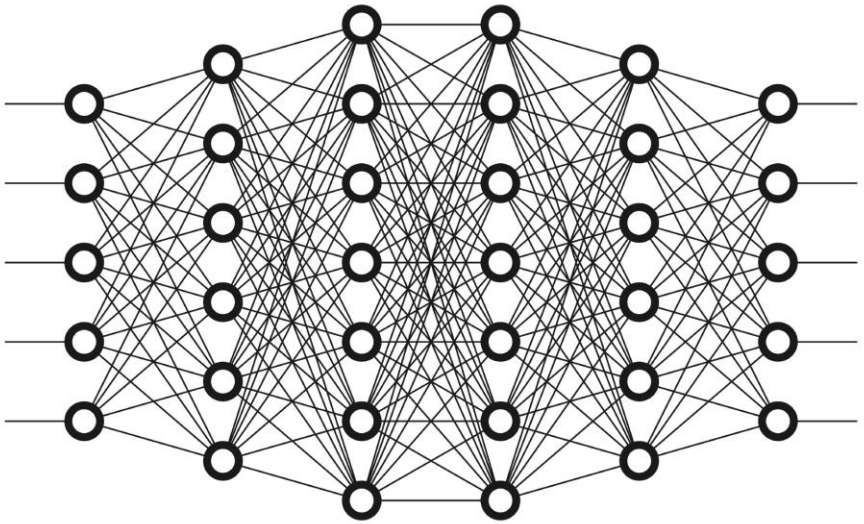


Probability of contamination increases!!

Machine learning technologies: comparison

In our studies we compared the prediction accuracy of BN to other machine learning technologies:

1. Neural network
2. Logistic regression
3. Support vector machines
4. Random forest
5. Ensemble classifiers



Prediction accuracy of BN was superior

WAGENINGEN
UNIVERSITY & RESEARCH

Big data infrastructure of food safety for automatic data collection, processing and visualization

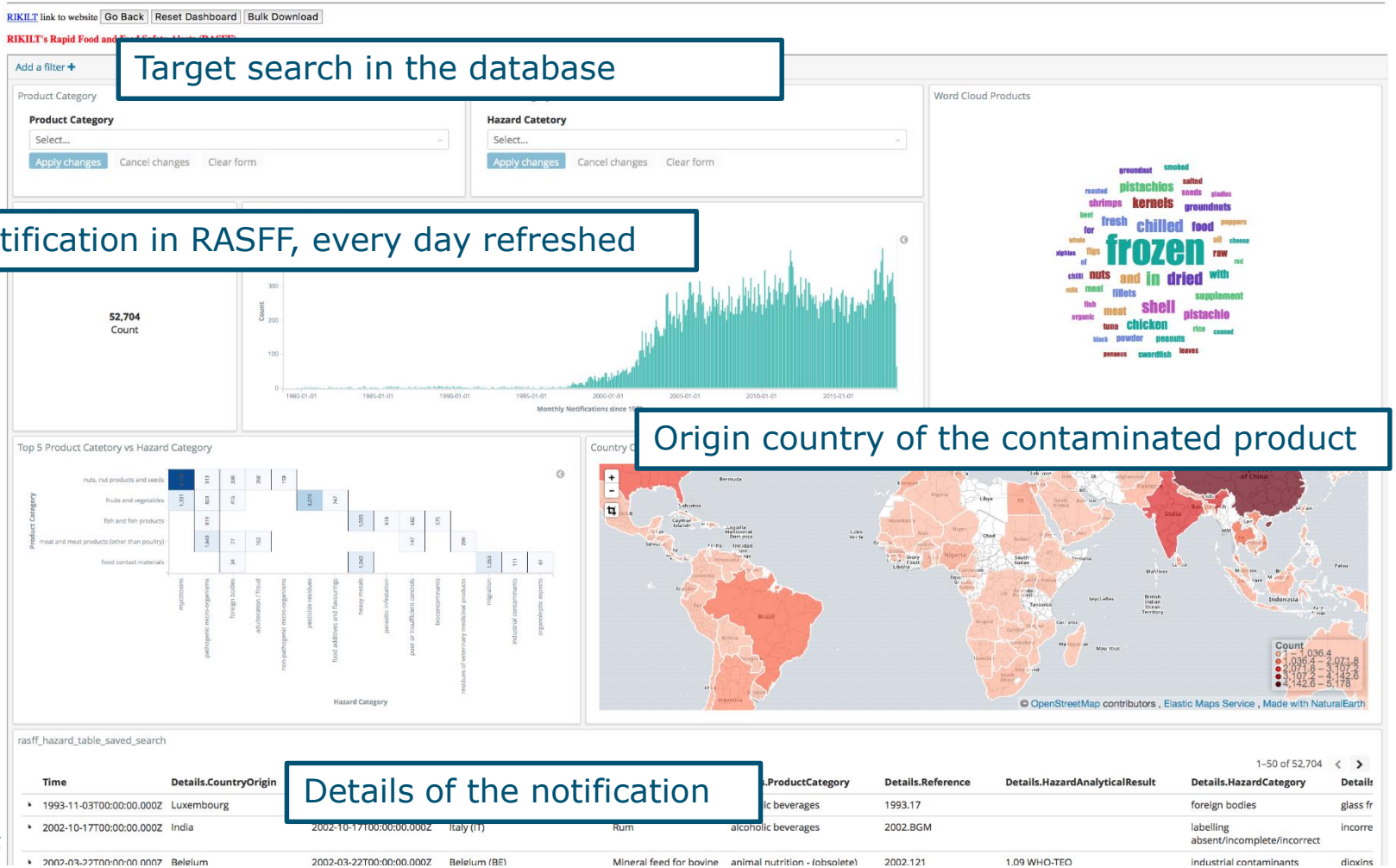
Realized: automatic collection of data, visualization and processed

Available data sources

1. Media publications (worldwide) (food fraud and stimulants example; early warning)
2. Scientific publication (Pubmed)
3. Social media (Twitter)
4. Rapid Alert System For Food and Feed (RASFF)
5. Climate data

In development: automatic feeding the Food Safety models and visualization

Example of automatic data collection and visualization (all features in the dashboard are interconnected)



Conclusions

- A BN model was constructed to predict the contamination level in dairy feed products produced or grown in the Netherlands
- A high accuracy of prediction (90.3%) was achieved
- An impact of climate parameters on the contamination level was demonstrated
- The results demonstrate the applicability of data driven BNs to capture complex interactions of parameters
- An Openshift big data infrastructure (HPC) for food safety has been developed allowing automatic data collection, processing and visualization

END

WAGENINGEN
UNIVERSITY & RESEARCH