Anomaly detection

How to build an anomaly detection system with Bayesian networks

Dr John Sandiford, CTO Bayes Server

Copyright © Bayes Server 2006 - 2014

Contents

- Introduction
- What is a Bayesian network?
- What is anomaly detection?
- Log-likelihood
- Multi-variate models
- Latent variables
- More complicated models
- Initialization
- Cluster Count
- Time series anomaly detection
- Underflow / overflow
- Alerting strategies
- Diagnostics
- Auto insight
- Big data

Copyright © Bayes Server 2006 - 2014

Introduction

Copyright © Bayes Server 2006 - 2014

Profile

linkedin.com/in/johnsandiford

- PhD Imperial College – Bayesian networks
- Machine learning – 15 years
  - Implementation
  - Application
  - Numerous techniques
- Algorithm programming even longer
  - Scala , C#, Java, Python, C++
- Graduate scheme – mathematician (BAE Systems)
- Artificial Intelligence / ML research program 8 years (GE/USAF)
- BP commodity trading – big data + machine learning + deep learning
- Also: NYSE stock exchange, hedge fund, actuarial consultancy, international newspaper

Copyright © Bayes Server 2006 - 2014

What is a Bayesian network?

Copyright © Bayes Server 2006 - 2014

What is a Bayesian network?

- DAG – directed acyclic graph
- Nodes, links, probability distributions
- Each node requires a probability distribution conditioned on its parents (if any)

$$P(\mathbf{X}, \mathbf{e}) = \sum_{\mathbf{U} \setminus \mathbf{X}} P(\mathbf{U}, \mathbf{e}) = \sum_{\mathbf{U} \setminus \mathbf{X}} \prod_{i} P(\mathbf{U}_i | pa(\mathbf{U}_i)) \mathbf{e}$$

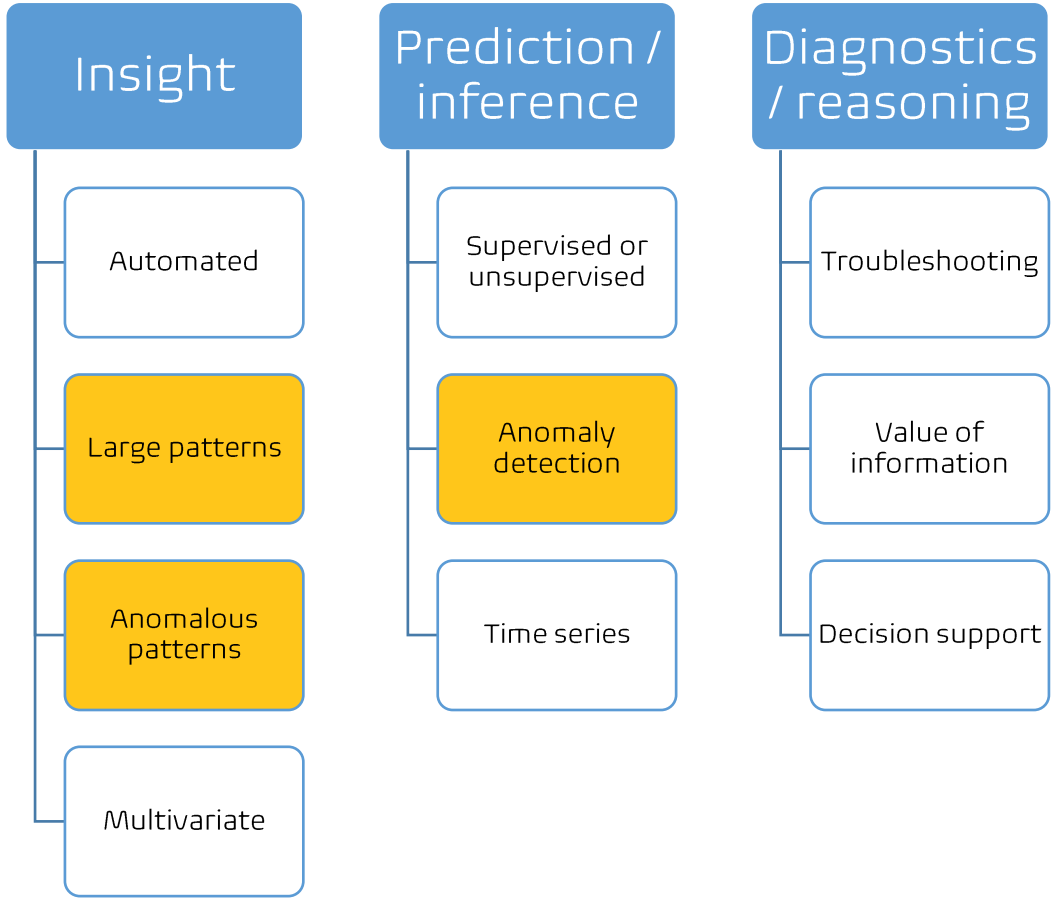
U = universe of variables
X = variables being predicted
e= evidence on any variables

Copyright © Bayes Server 2006 - 2014

Example – Asia network

Copyright © Bayes Server 2006 - 2014

What can they be used for



Focus of this talk

Copyright © Bayes Server 2006 - 2014

What is anomaly detection?

Copyright © Bayes Server 2006 - 2014

What is anomaly detection?

Anomaly detection, or outlier detection, is the process of identifying data which is unusual.

- System health monitoring
  - Advanced warning of mechanical failure
- Fault detection
  - Isolate faulty components
- Fraud detection
  - Fraudulent transactions or unusual behaviour
- Pattern detection
  - Can detect unusual patterns
- Pre-processing
  - E.g. removal/replacement of unusual data, before building statistical models.
- Unusual Time series
  - Interaction between many time series

Copyright © Bayes Server 2006 - 2014

Types of anomaly detection

- Unsupervised
  - Normal data + anomalous data
- Semi-supervised
  - Normal data
  - Anomalous data has been removed
- Supervised
  - Labelled data
  - Specific faults
  - Problematic if too few cases or anomalies always different
  - We won't talk about this type today
- Time series
  - Any of the above

Copyright © Bayes Server 2006 - 2014

Missing data

- Can handle missing data during learning
- Can handle missing data during inference (prediction)
- Missing time series data
- Resulting models can be used to fill in missing data

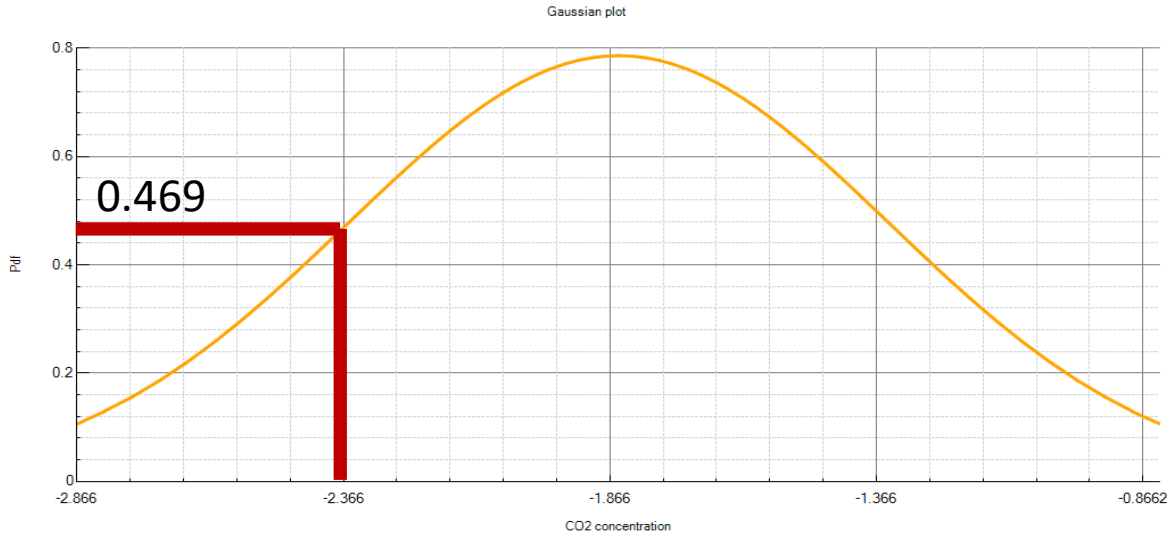
Copyright © Bayes Server 2006 - 2014

Log-likelihood

Copyright © Bayes Server 2006 - 2014

Log-likelihood - simple example

- Consider a univariate Gaussian
- Pdf shown in orange in chart
- Log-likelihood
  - = log(pdf)
  - Natural logarithm common for Gaussians as they are exponential distributions



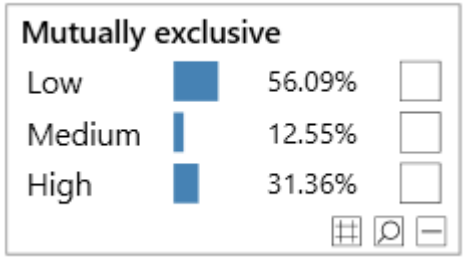

```
=NORM.DIST(-2.366,-1.85,SQRT(0.2575),FALSE)
```

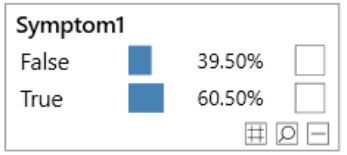
- Log-likelihood can be calculated for other distributions
  - E.g. categorical distribution (multinoulli)

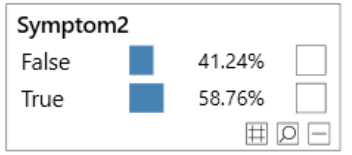
Copyright © Bayes Server 2006 - 2014

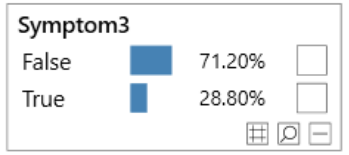
Discrete data

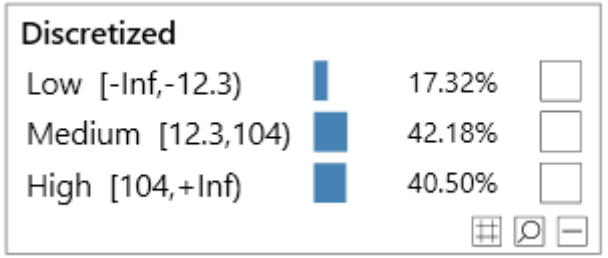
- Single variable with mutually exclusive states
  - Events that cannot happen at the same time
  - We can still get a probability for each state during prediction
  - In general Bayes networks do not require data to be 1-hot encoded
- Multiple binary variables
  - Multiple events that can co-occur
- If in doubt:
  - Imagine how you would record your data in a database table
  - Create a variable for each column
- Discretization of continuous





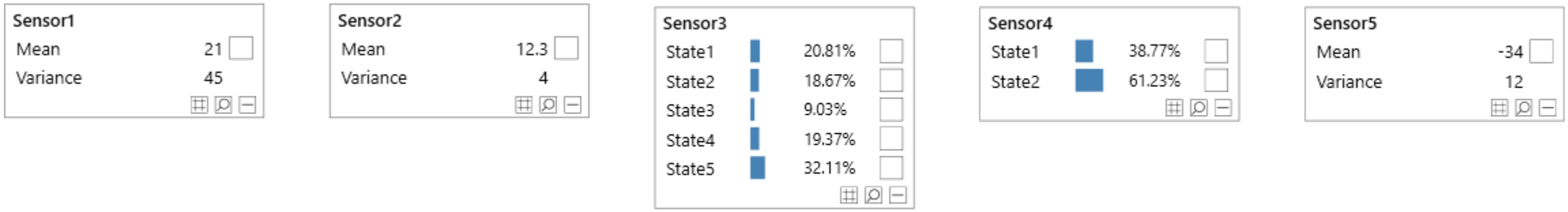






Copyright © Bayes Server 2006 - 2014

The simplest of Bayesian network models



No links. This is surprisingly common!

Each variable is often considered in isolation, with alerts set at a fixed number of standard deviations above and below (for continuous) or probability thresholds (for discrete).

Note, you don't have to use a Bayesian network for this.

Copyright © Bayes Server 2006 - 2014

Log-likelihood

- Allows us to perform anomaly detection
- Can be calculated for
  - Discrete, continuous & hybrid networks
  - Networks with latent variables
  - Time series networks
- Under the hood, great care has to be taken to avoid underflow
  - Especially with temporal networks

Copyright © Bayes Server 2006 - 2014

Calculating log-likelihood – Bayesian networks

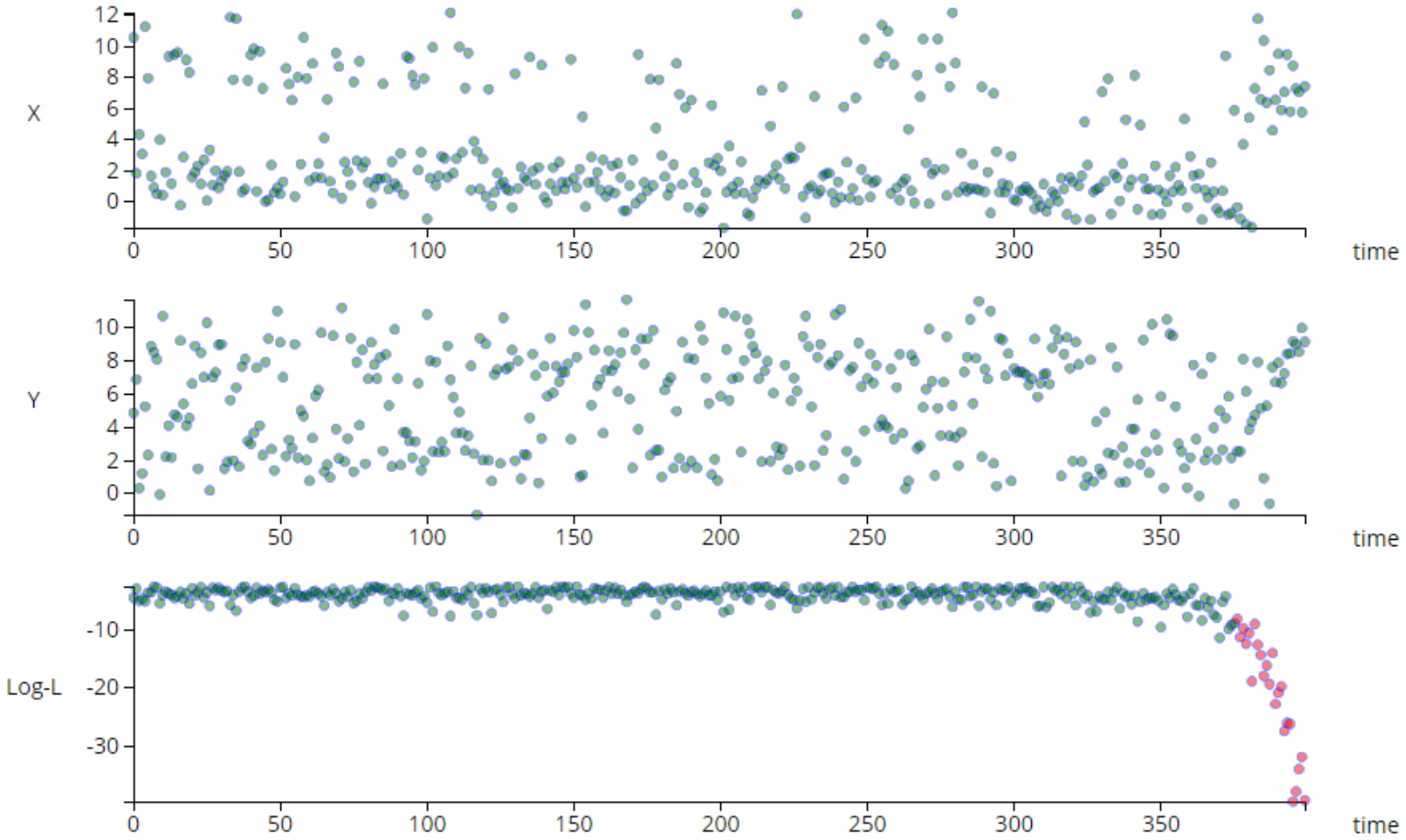
- Given evidence, marginalize out all other variables
  - Marginalize means sum for discrete, integrate for continuous.
- Can be calculated using a simple algorithm such as 'variable elimination' or as a byproduct of a more sophisticated algorithm
- Optimizations can be used to exclude parts of the graph
- Efficient calculation involves complex algorithms
- See presentation on 'Bayesian network internals' for details
  - http://www.bayesserver.com/presentations.aspx

Copyright © Bayes Server 2006 - 2014

Multivariate models

Copyright © Bayes Server 2006 - 2014

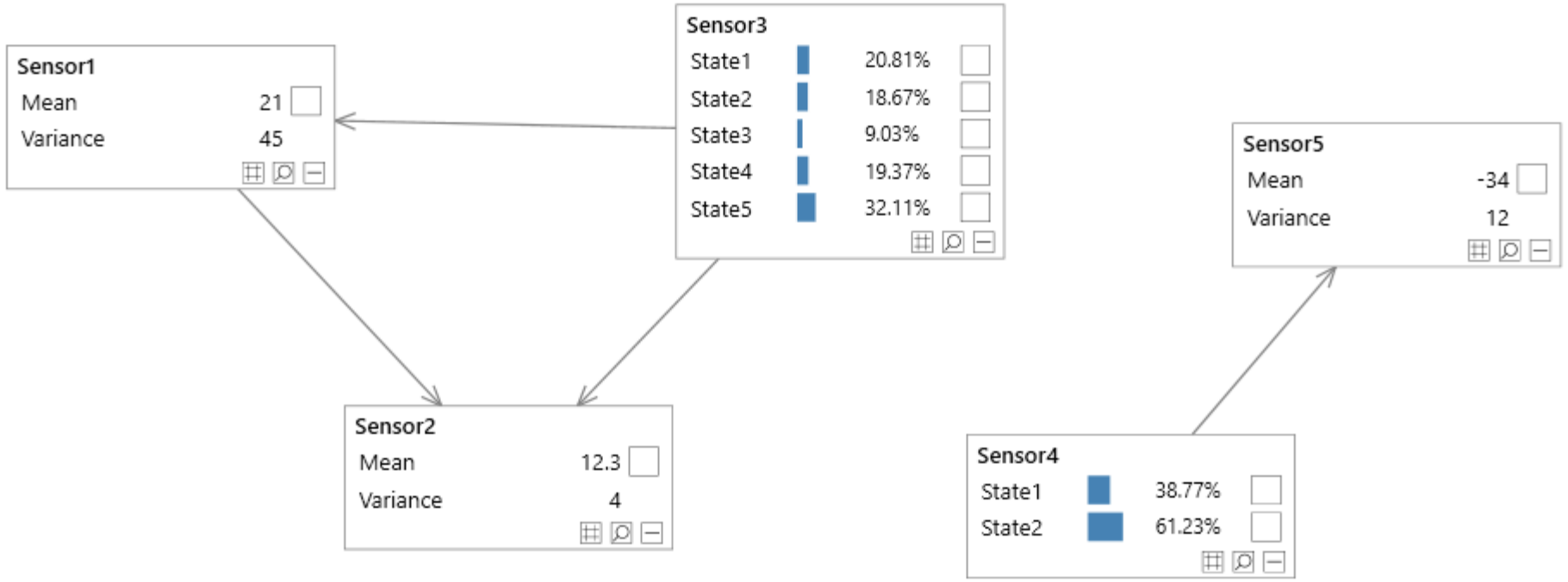
When univariate models fail



D3 animated visualization available on our website

Copyright © Bayes Server 2006 - 2014

From univariate to multivariate models



- Often we re-use pre-defined structures
- Algorithms can be used to determine the structure from data,
- The structure can be defined by experts.
- We can still calculate the log-likelihood as before.

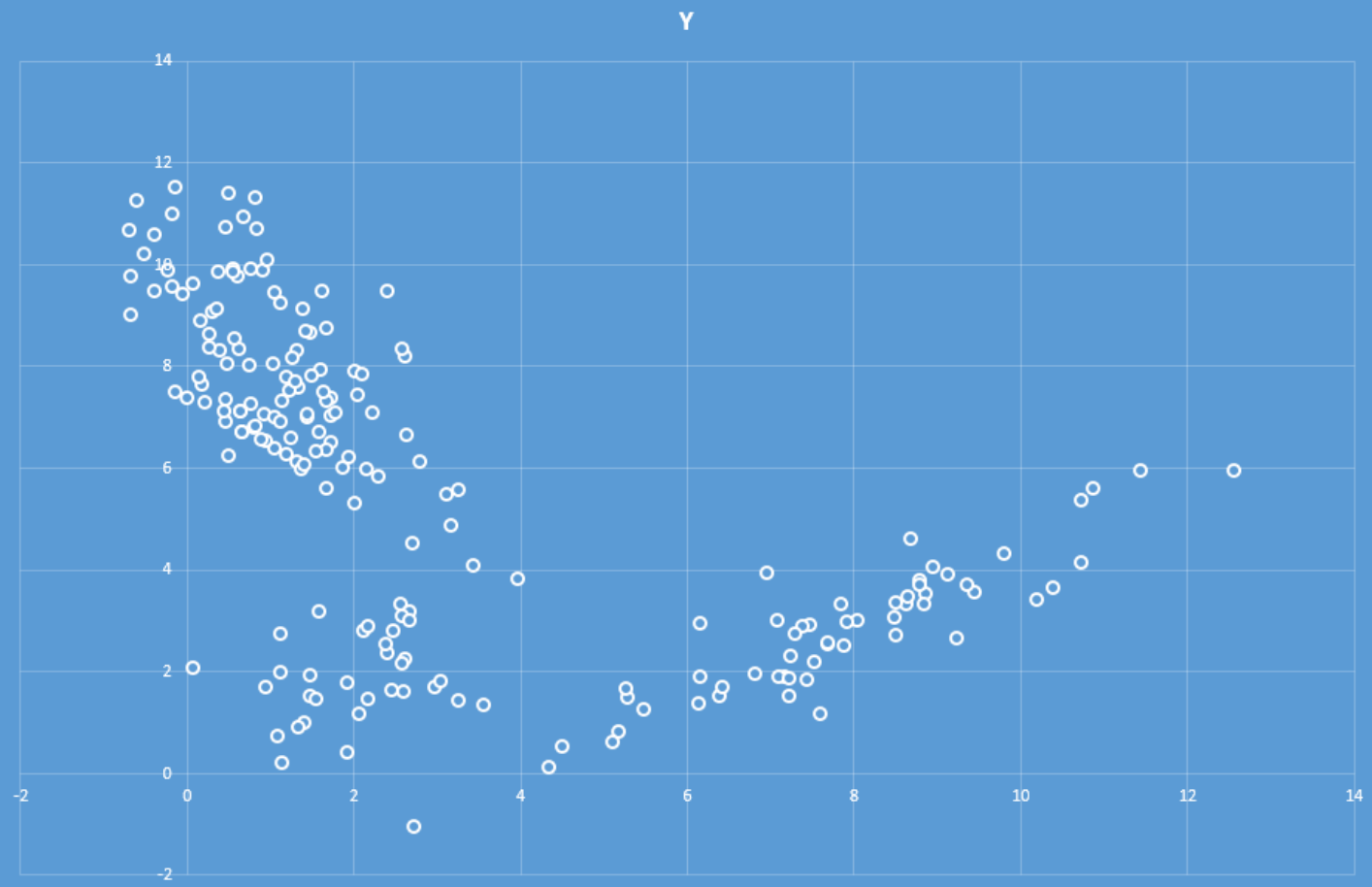
Copyright © Bayes Server 2006 - 2014

Latent variables

Copyright © Bayes Server 2006 - 2014

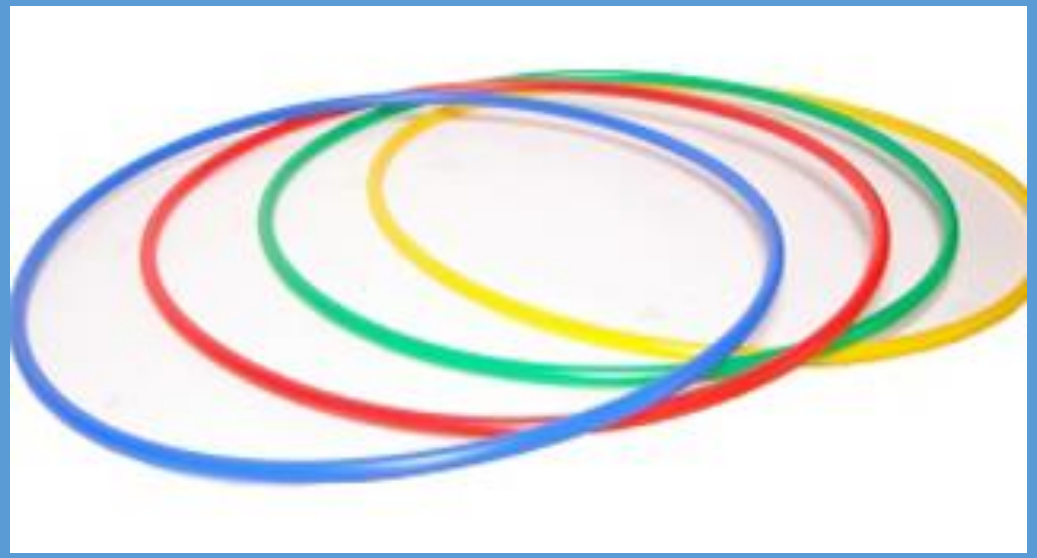
Latent variables

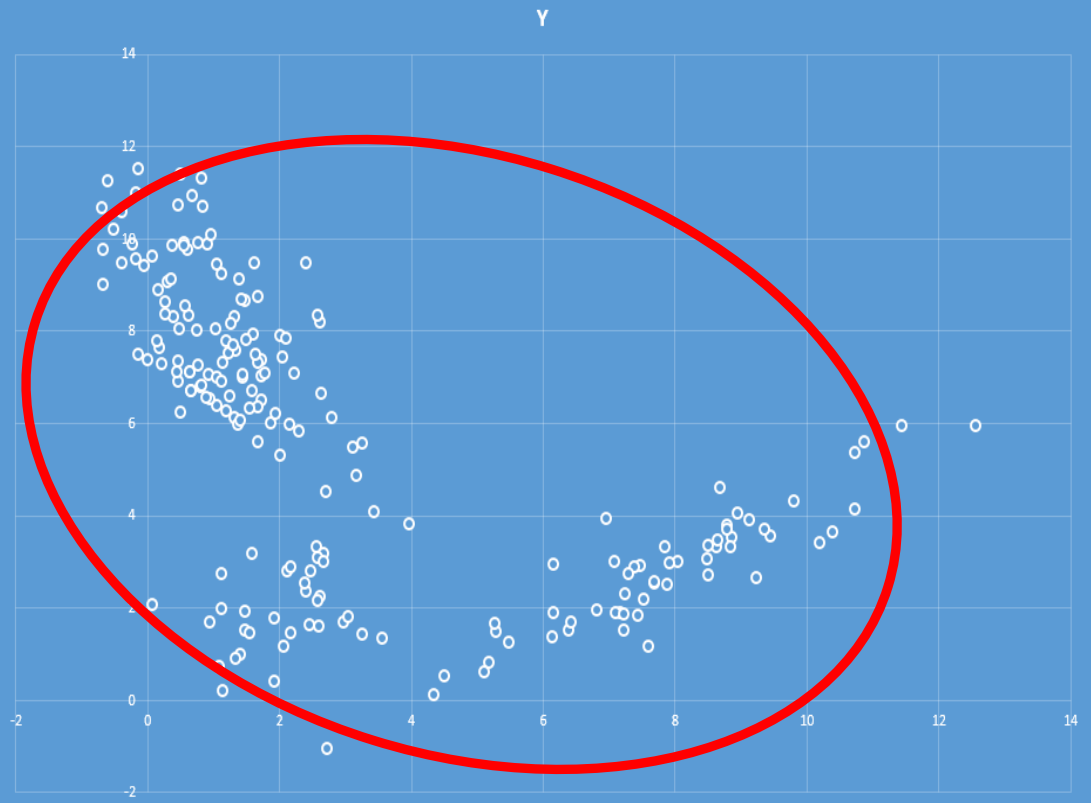
X	Y
2.0	7.9
6.9	1.98
0.1	2.1
1.1	?
9.1	7.2
?	9.2
…	…



Copyright © Bayes Server 2006 - 2014

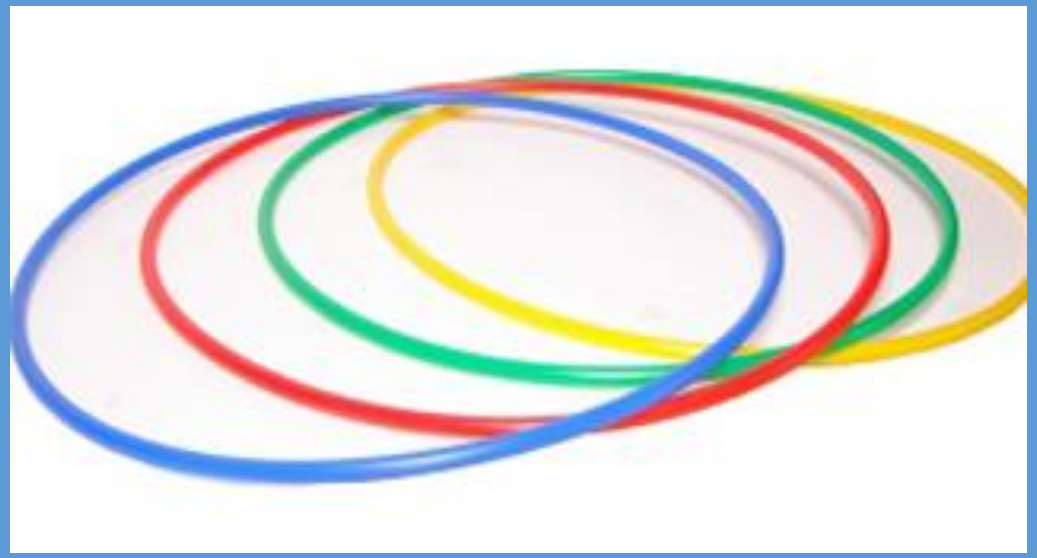
Latent variables

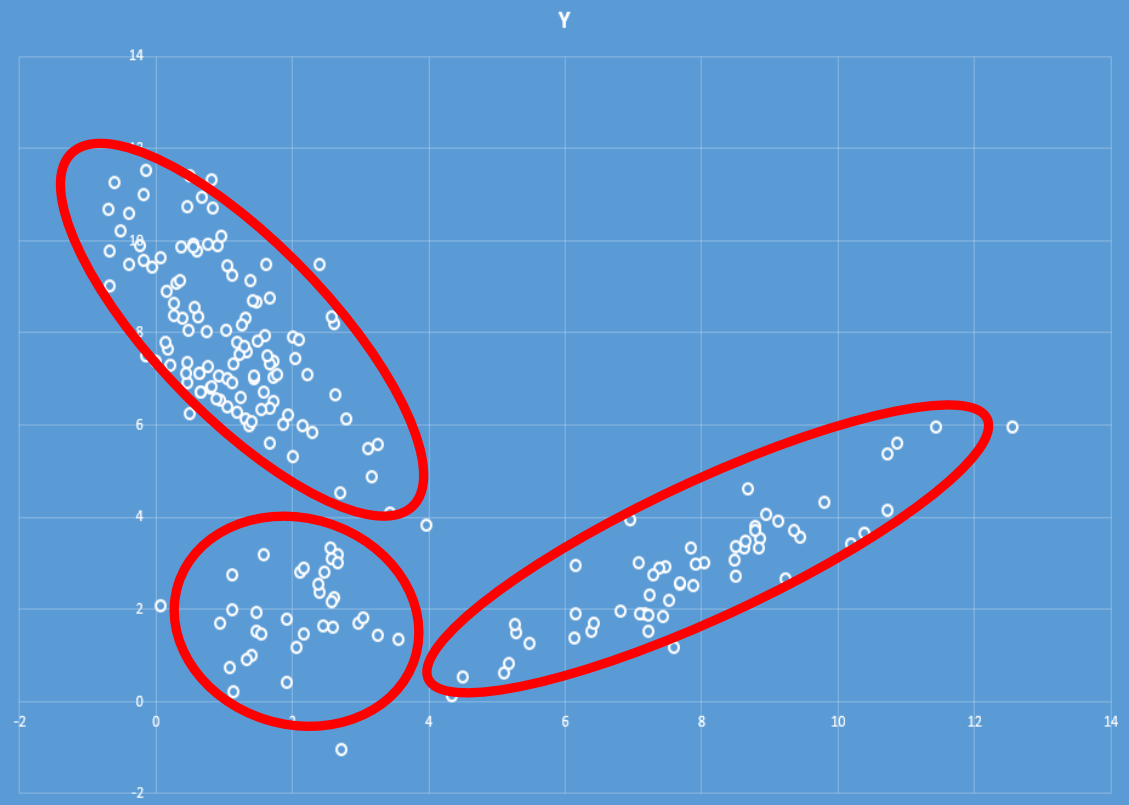




Copyright © Bayes Server 2006 - 2014

Latent variables

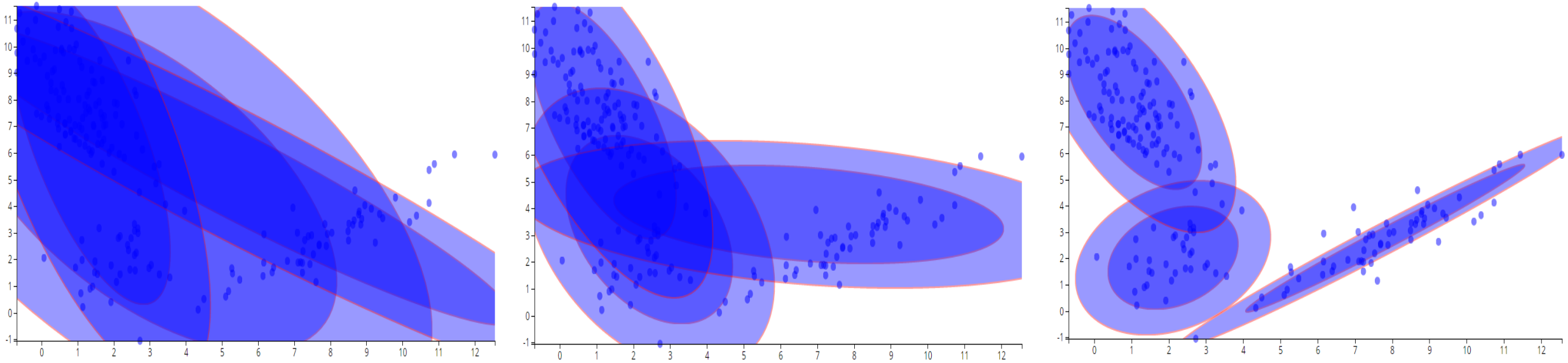




Copyright © Bayes Server 2006 - 2014

Parameter learning

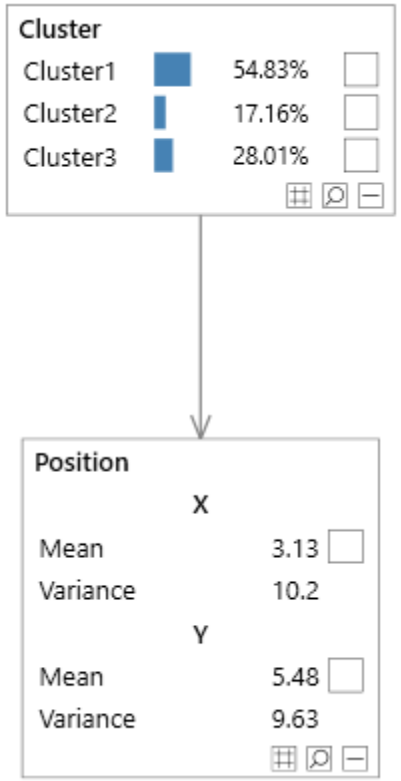
EM algorithm & extensions for missing data



- D3 animated visualization available on our website
- In practice a good initialization algorithm will be close to end result

Copyright © Bayes Server 2006 - 2014

Latent variables



- This is exactly the same as a mixture model (cluster model)
- Cluster is similar to a hidden layer in a neural network
- This model only has X & Y, but most models have much higher dimensionality
- We can extend other models in the same way, e.g.
  - Mixture of Naïve Bayes (no longer Naïve)
  - Mixture of time series models
  - A structured approach to ensemble methods?

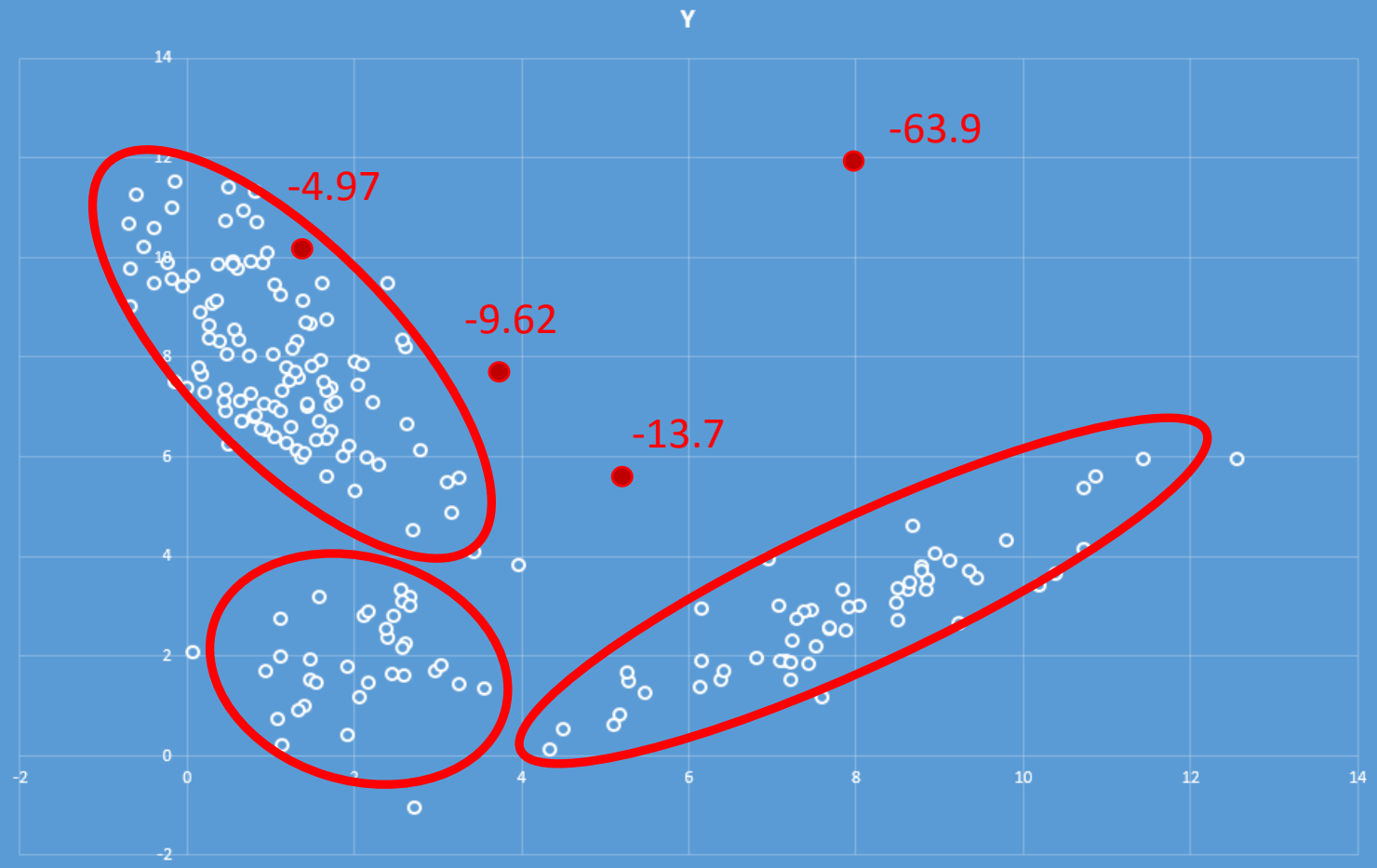
Copyright © Bayes Server 2006 - 2014

Latent variables

- Algorithmically capture underlying mechanisms that haven't or can' t be observed
- Latent variables can be both discrete & continuous
- We can have multiple latent variables in a network
- Can be hierarchical (similar to Deep Belief networks)

Copyright © Bayes Server 2006 - 2014

Anomaly detection

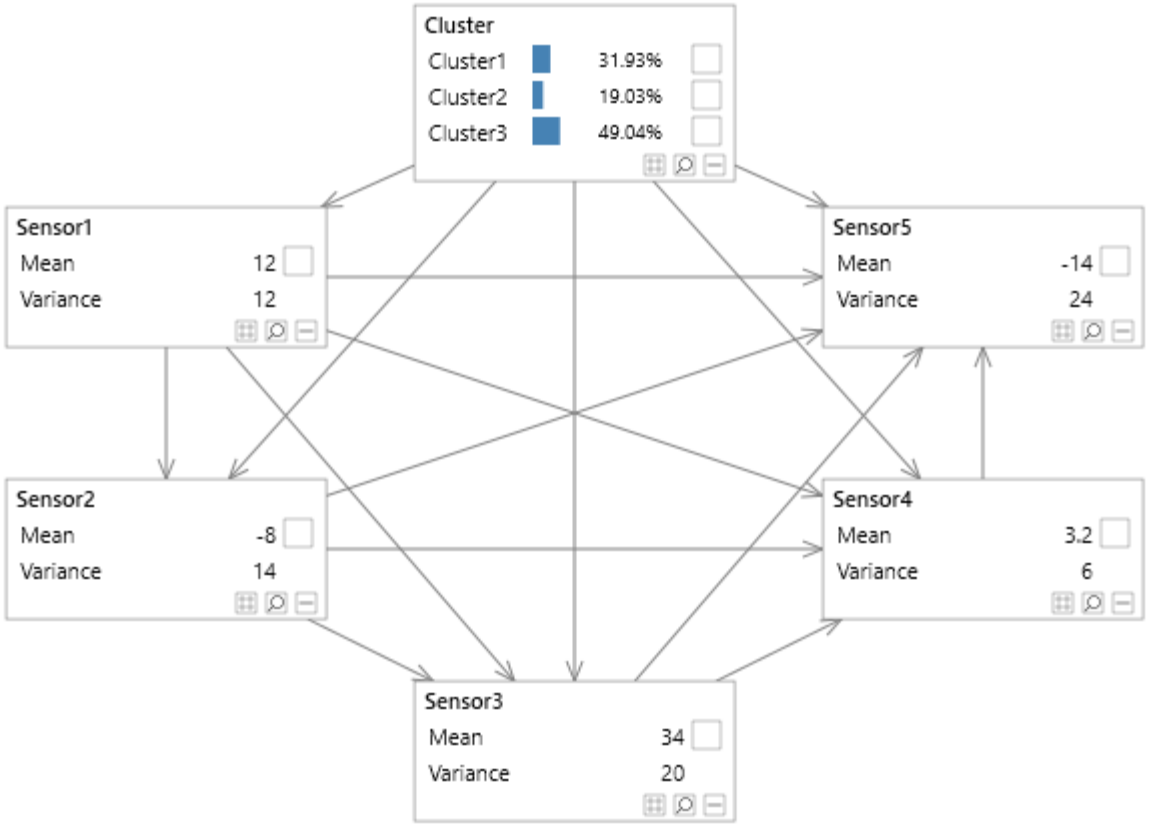


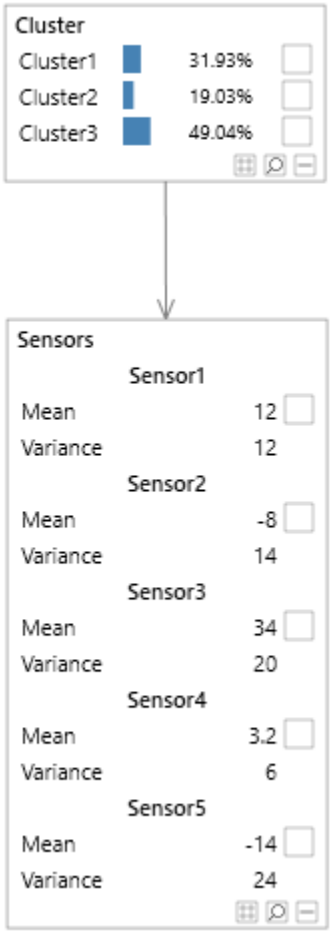
Copyright © Bayes Server 2006 - 2014

More complicated models

Copyright © Bayes Server 2006 - 2014

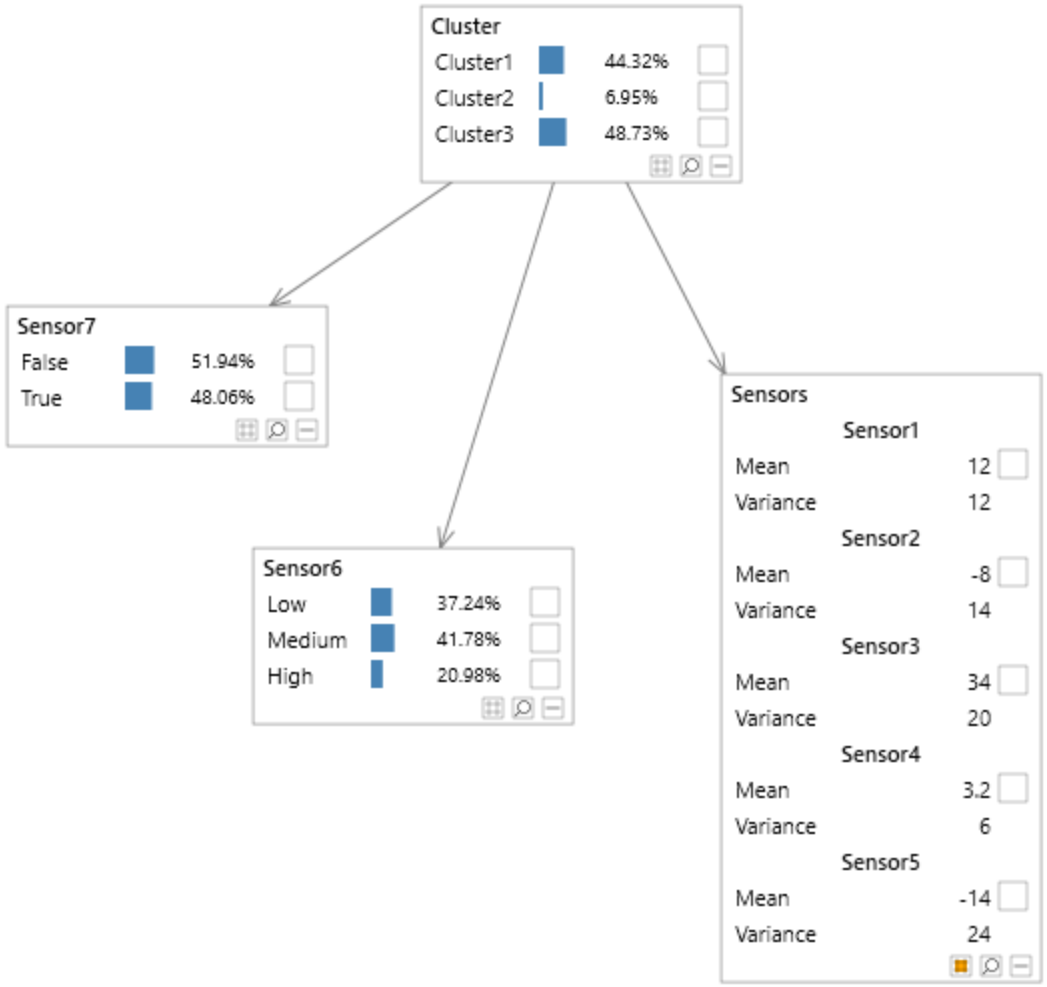
Multi-variate nodes

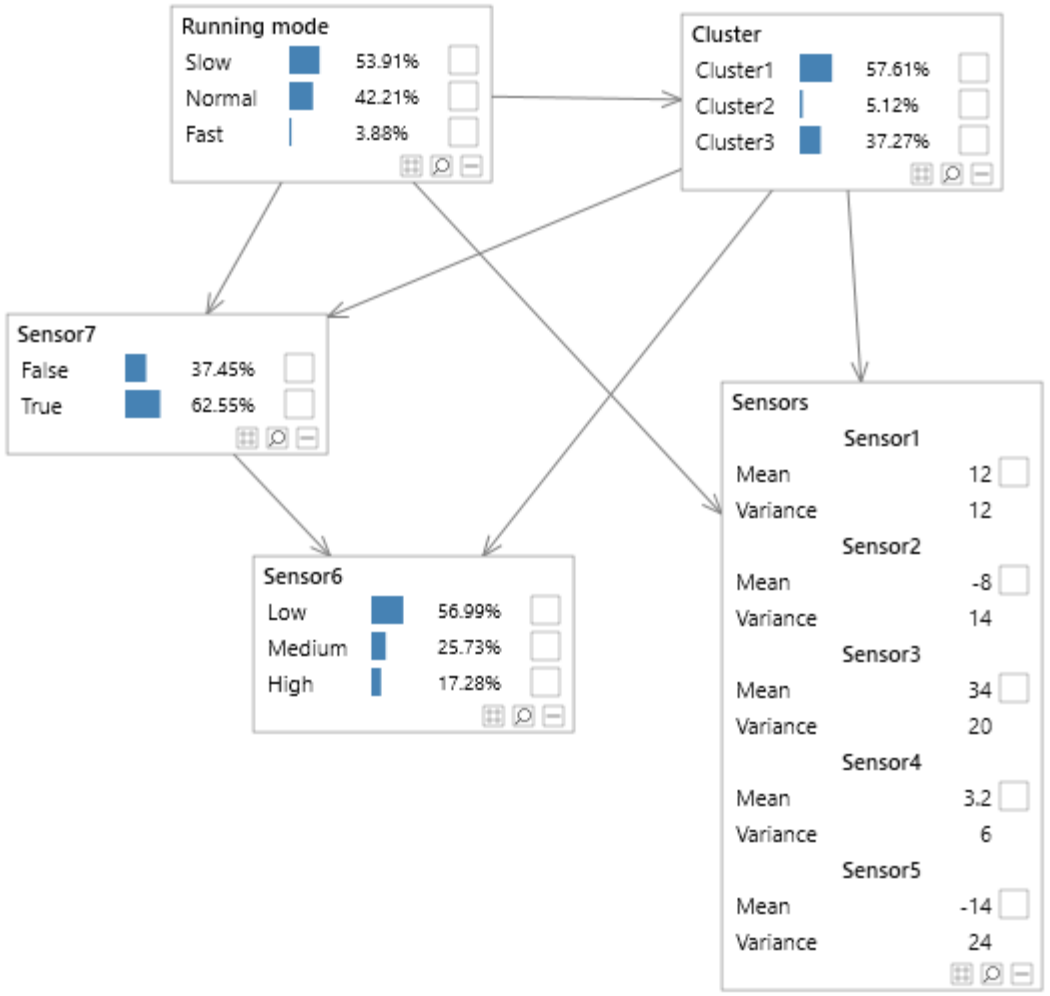




Copyright © Bayes Server 2006 - 2014

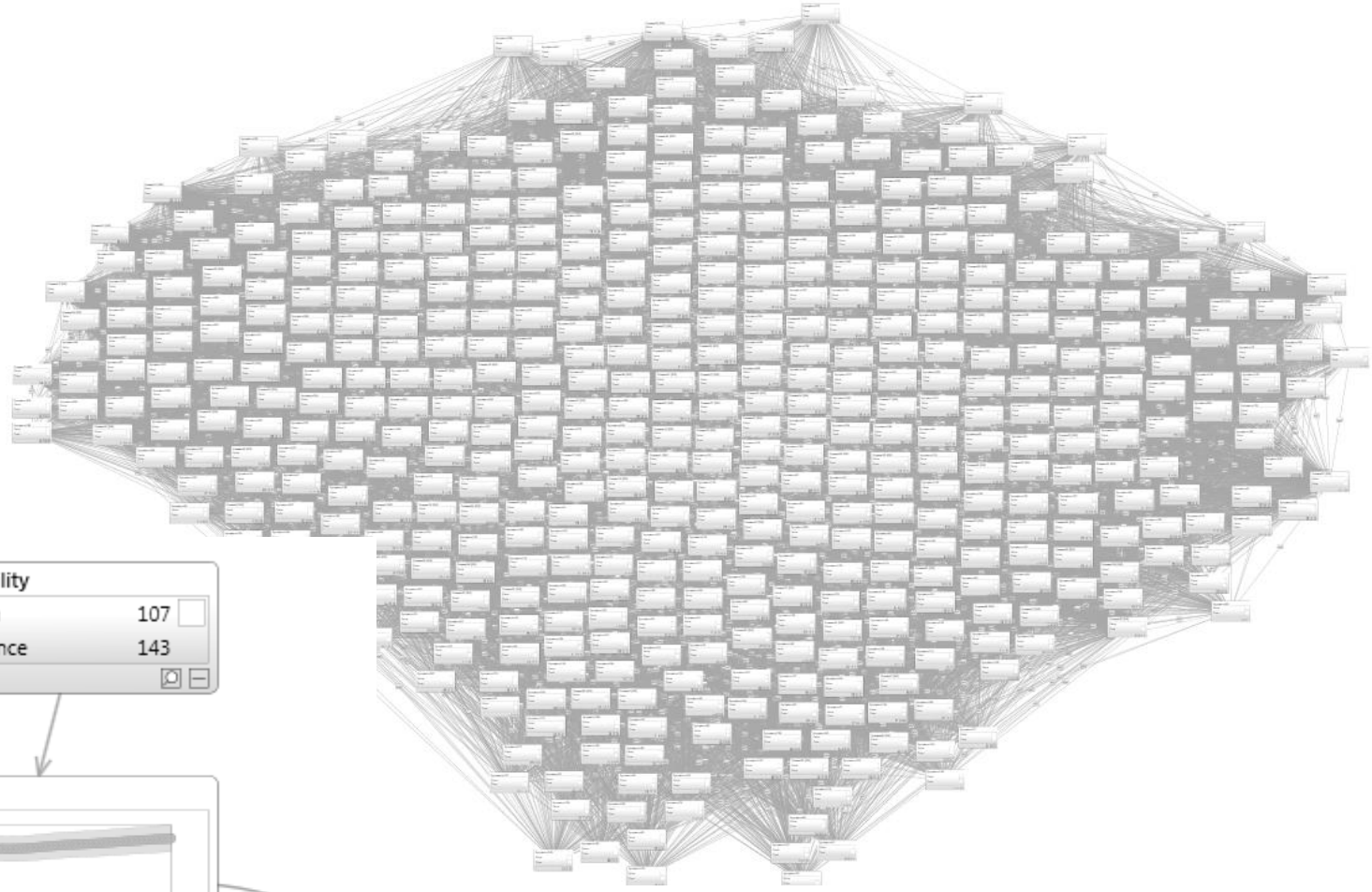
Extending simple models

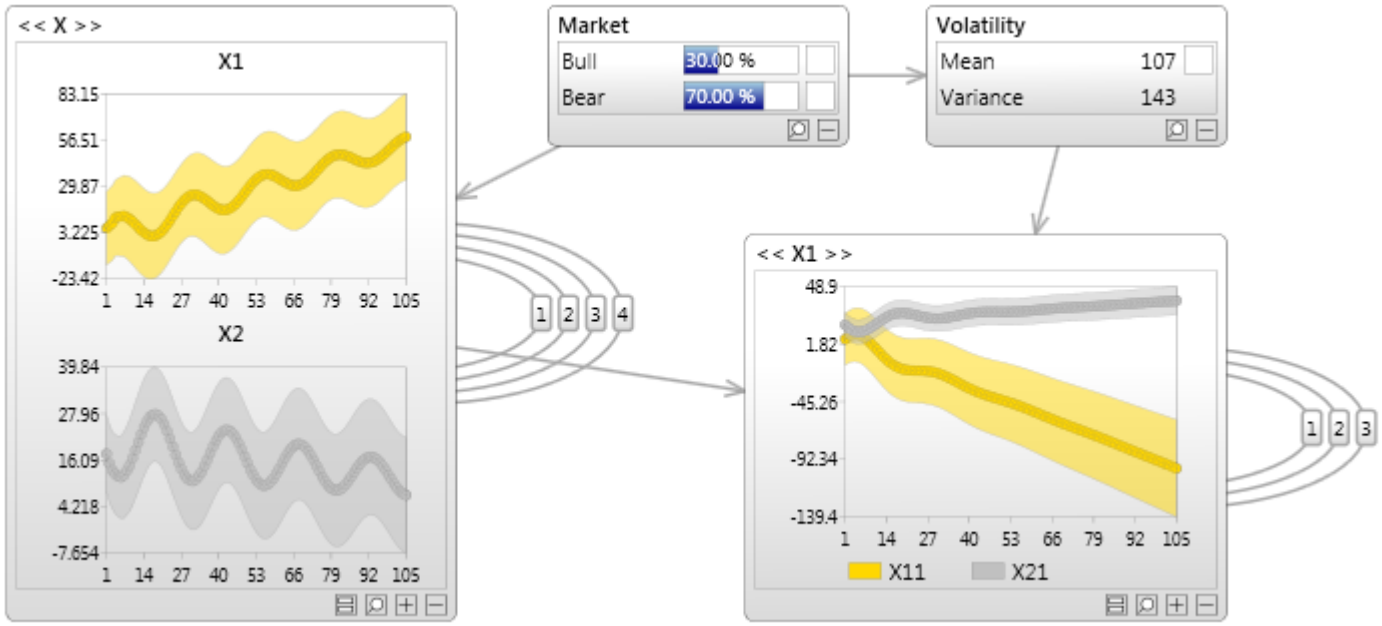




Copyright © Bayes Server 2006 - 2014

Other models





Copyright © Bayes Server 2006 - 2014

Anomaly detection with Bayesian networks

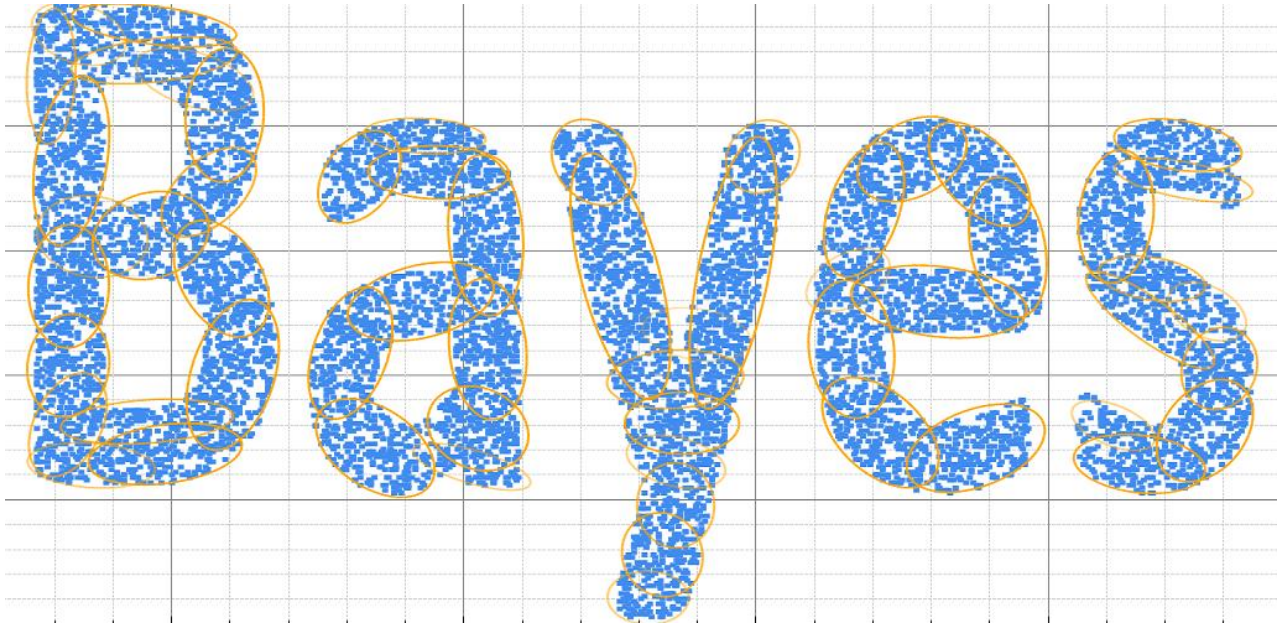
- High dimensional data
  - Humans find difficult to interpret
  - Anomalies may not be visible on individual variables
- Allow missing data
  - Learning
  - Prediction/anomaly detection
- Temporal and non temporal variables in the same model
- Multiple discrete/continuous latent variables

Copyright © Bayes Server 2006 - 2014

Initialization

Copyright © Bayes Server 2006 - 2014

Initialization



- Random initialization not always the best approach and can lead to longer training times
- Clustered initialization
- Deep learning unsupervised pre-training techniques
- Greedily initialize using a topological ordering of latent variables

Copyright © Bayes Server 2006 - 2014

Cluster count

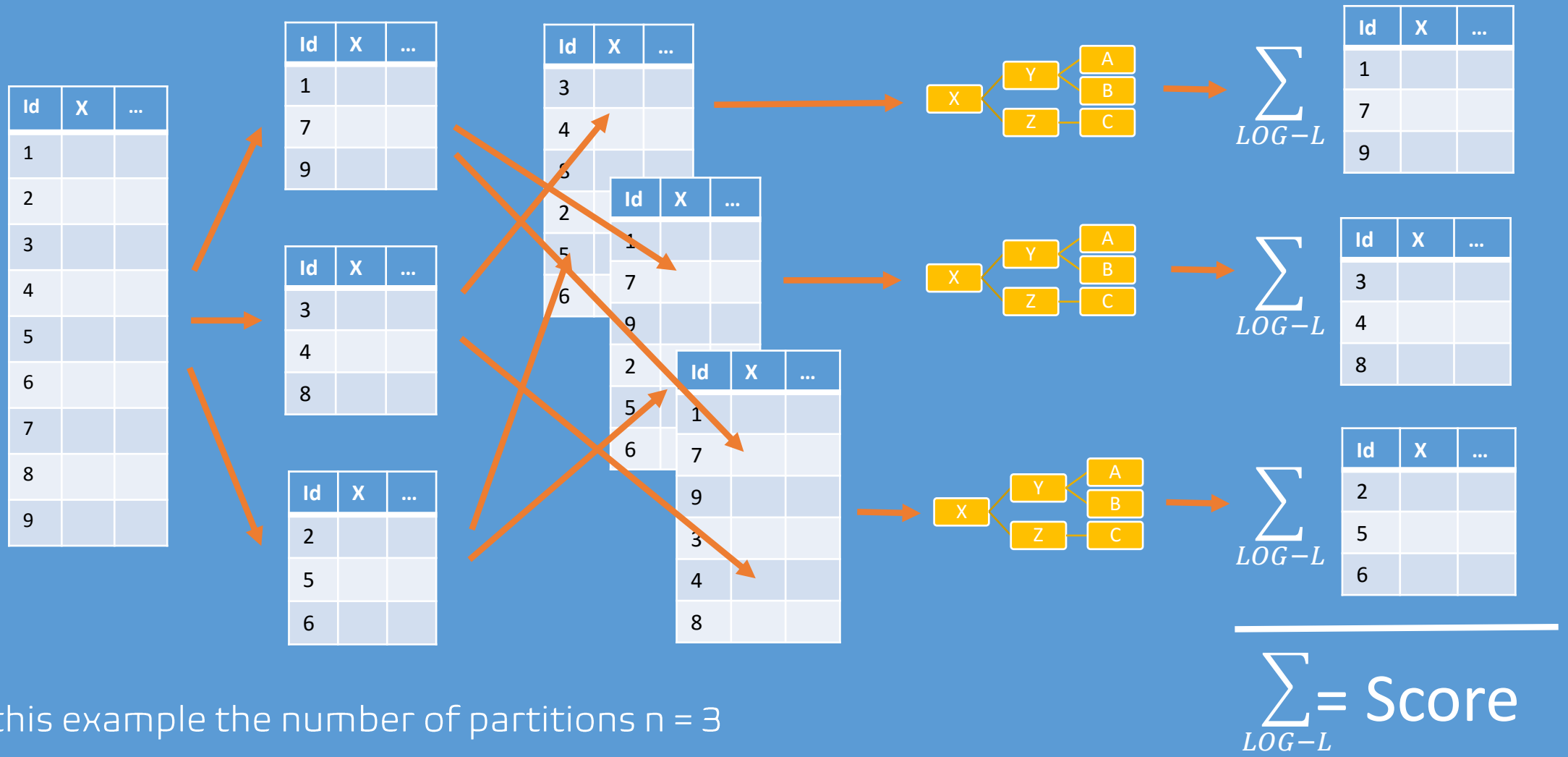
Copyright © Bayes Server 2006 - 2014

Cluster count

- API method: ClusterCount.Detect
  - Uses cross validation
  - Log-likelihood score summed over each partition
  - Evaluate score for different cluster counts
- Alternatives:
  - Heuristics such as BIC
  - Dpgmm
  - Etc...

Copyright © Bayes Server 2006 - 2014

Cross validation – log likelihood score

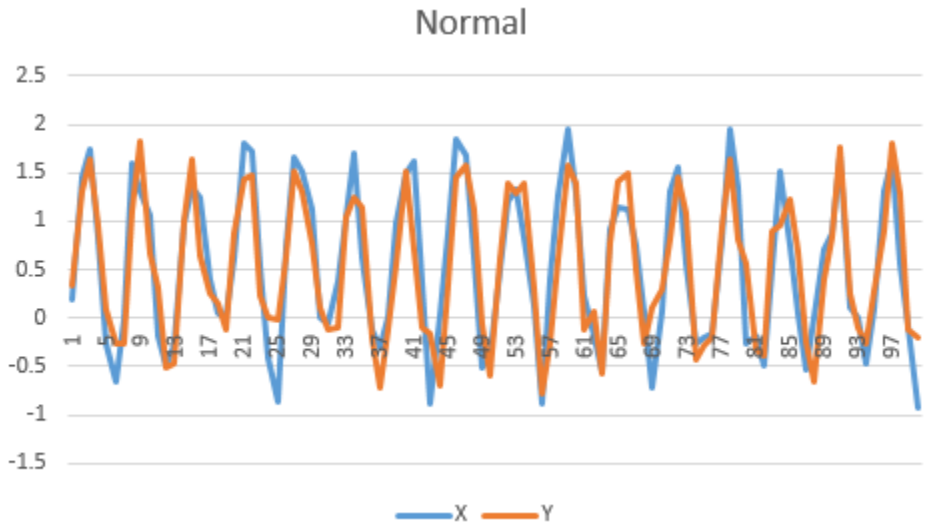


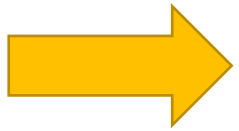
In this example the number of partitions n = 3

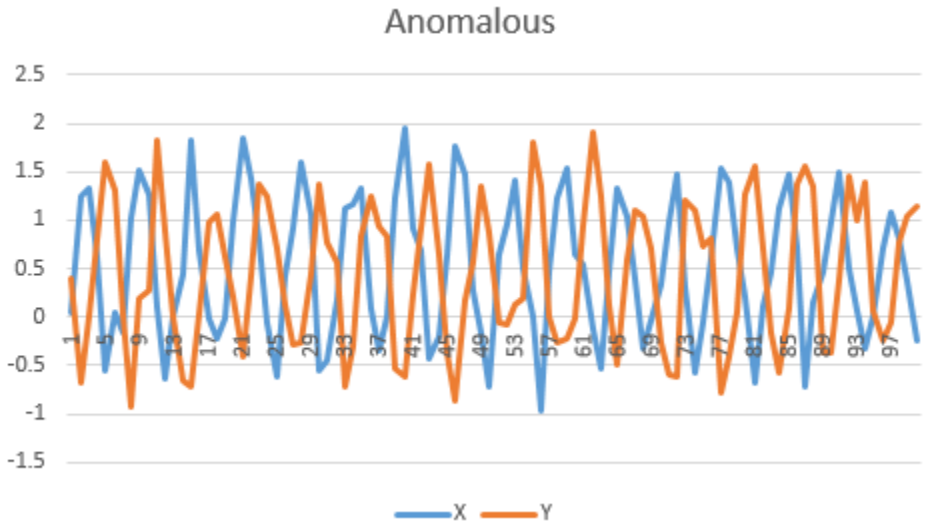
Copyright © Bayes Server 2006 - 2014

Time series anomaly detection

Multivariate time series anomaly detection







- Log likelihood is also available for time series models
- Individual time series above both within normal bounds
- May get degradation in between, i.e. advanced warning

Copyright © Bayes Server 2006 - 2014

Underflow / overflow

Copyright © Bayes Server 2006 - 2014

Underflow / overflow

- Log always used over multiple records during learning
- We also use log(pdf) for each record
- Especially important with large number of variables and time series models

x	pdf(x)	logPdf(x)
0.001	0.12615662	-2.07023113
0.01	0.126155995	-2.07023608
0.1	0.126093564	-2.07073108
1	0.120003895	-2.12023108
10	0.000850037	-7.07023108
100	8.99E-219	-502.0702311
1000	0	-50002.07023
10000	0	-5000002.07
100000	0	-500000002.1
1000000	0	-50000000002
10000000	0	-5E+12
100000000	0	-5E+14
1000000000	0	-5.00E+16
10000000000	0	-5.00E+18

Copyright © Bayes Server 2006 - 2014

Alerting strategies

Copyright © Bayes Server 2006 - 2014

Alerting strategies

- Push historic or sampled data through the model
  - May need to mimic anomalous data
- Inspect the distribution of the log-likelihood
- Set simple alerting thresholds
- Numerical approximation of the log-likelihood
  - E.g. kernel smoothing function estimate

Copyright © Bayes Server 2006 - 2014

Diagnostics

Copyright © Bayes Server 2006 - 2014

Diagnostics / reasoning

- What is causing the anomaly?
- Retracted log-likelihoods
  - Individually retracted
  - Joint retracted log-likelihoods
- Some tools support conflict resolution

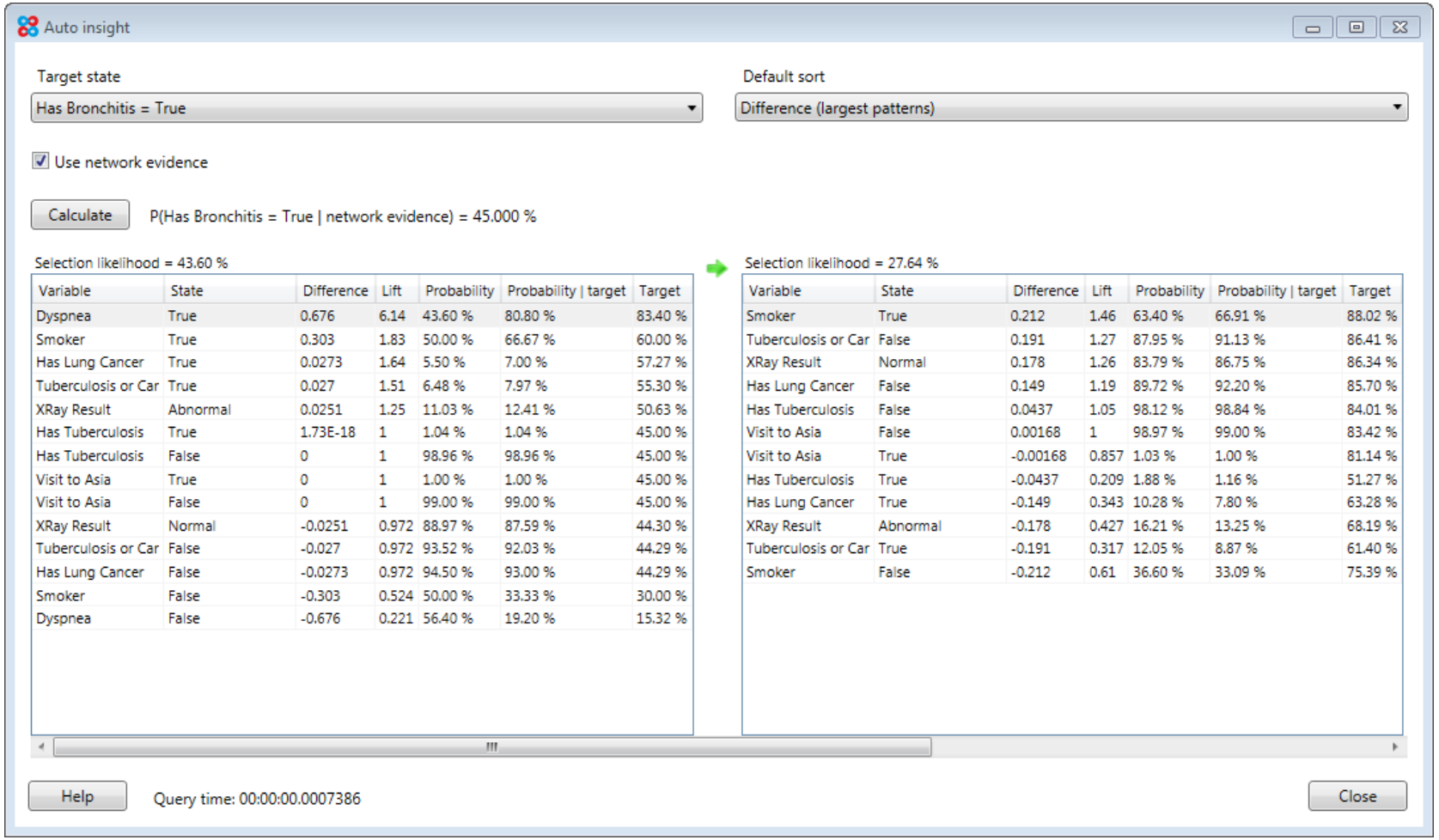
Copyright © Bayes Server 2006 - 2014

Auto insight

Copyright © Bayes Server 2006 - 2014

Auto insight

- Anomalous patterns
  - Large (Diff)
  - Small (Lift)
- Automated
- Drilldown
- Can use current evidence



Variable	State	Difference	Lift	Probability	Probability | target	Target
Dyspnea	True	0.676	6.14	43.60 %	80.80 %	83.40 %
Smoker	True	0.303	1.83	50.00 %	66.67 %	60.00 %
Has Lung Cancer	True	0.0273	1.64	5.50 %	7.00 %	57.27 %
Tuberculosis or Car	True	0.027	1.51	6.48 %	7.97 %	55.30 %
XRay Result	Abnormal	0.0251	1.25	11.03 %	12.41 %	50.63 %
Has Tuberculosis	True	1.73E-18	1	1.04 %	1.04 %	45.00 %
Has Tuberculosis	False	0	1	98.96 %	98.96 %	45.00 %
Visit to Asia	True	0	1	1.00 %	1.00 %	45.00 %
Visit to Asia	False	0	1	99.00 %	99.00 %	45.00 %
XRay Result	Normal	-0.0251	0.972	88.97 %	87.59 %	44.30 %
Tuberculosis or Car	False	-0.027	0.972	93.52 %	92.03 %	44.29 %
Has Lung Cancer	False	-0.0273	0.972	94.50 %	93.00 %	44.29 %
Smoker	False	-0.303	0.524	50.00 %	33.33 %	30.00 %
Dyspnea	False	-0.676	0.221	56.40 %	19.20 %	15.32 %

Variable	State	Difference	Lift	Probability	Probability | target	Target
Smoker	True	0.212	1.46	63.40 %	66.91 %	88.02 %
Tuberculosis or Car	False	0.191	1.27	87.95 %	91.13 %	86.41 %
XRay Result	Normal	0.178	1.26	83.79 %	86.75 %	86.34 %
Has Lung Cancer	False	0.149	1.19	89.72 %	92.20 %	85.70 %
Has Tuberculosis	False	0.0437	1.05	98.12 %	98.84 %	84.01 %
Visit to Asia	False	0.00168	1	98.97 %	99.00 %	83.42 %
Visit to Asia	True	-0.00168	0.857	1.03 %	1.00 %	81.14 %
Has Tuberculosis	True	-0.0437	0.209	1.88 %	1.16 %	51.27 %
Has Lung Cancer	True	-0.149	0.343	10.28 %	7.80 %	63.28 %
XRay Result	Abnormal	-0.178	0.427	16.21 %	13.25 %	68.19 %
Tuberculosis or Car	True	-0.191	0.317	12.05 %	8.87 %	61.40 %
Smoker	False	-0.212	0.61	36.60 %	33.09 %	75.39 %

Copyright © Bayes Server 2006 - 2014

Big data

Copyright © Bayes Server 2006 - 2014

Anomaly detection – big data

- PredictLogLikelihood()
- Batch or streaming

```
// make some time series predictions into the future

val predictions = Prediction.predict[TimeSeries](
  network,
  testData,
  Seq(
    PredictVariable("X1", Some(PredictTime(5, Absolute))), PredictVariance("X1", Some(PredictTime(5, Absolute))),
    PredictVariable("X2", Some(PredictTime(5, Absolute))), PredictVariance("X2", Some(PredictTime(5, Absolute))),
    PredictVariable("X1", Some(PredictTime(6, Absolute))), PredictVariance("X1", Some(PredictTime(6, Absolute))),
    PredictVariable("X2", Some(PredictTime(6, Absolute))), PredictVariance("X2", Some(PredictTime(6, Absolute))),
    PredictLogLikelihood() // this value can be used for Time Series anomaly detection
  ),
  (network, iterator) => new TimeSeriesReader(network, iterator))

predictions.foreach(println)
```


Copyright © Bayes Server 2006 - 2014