Biostat 202: Opportunities and Challenges of "Big Data": Clustering and data reduction

John Kornak

Division of Biostatistics

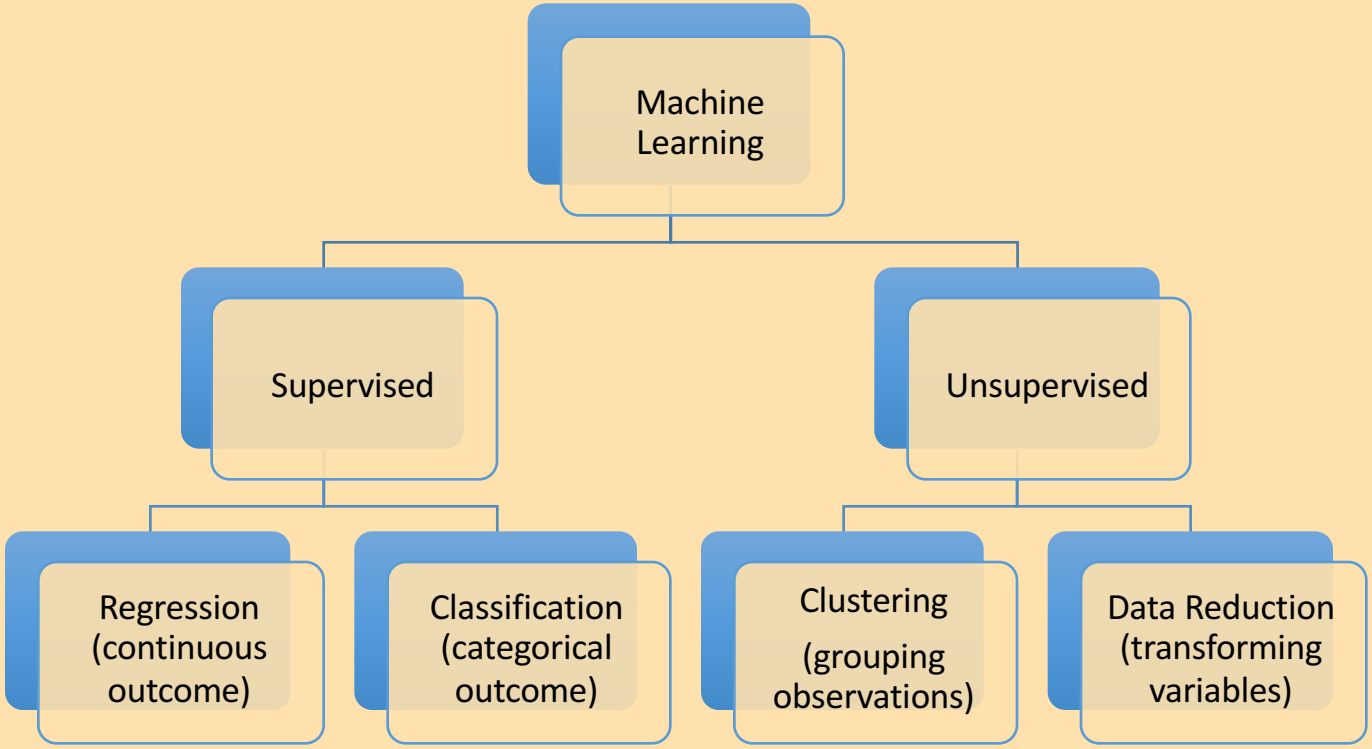
Department of Epidemiology and Biostatistics

08/22/2016

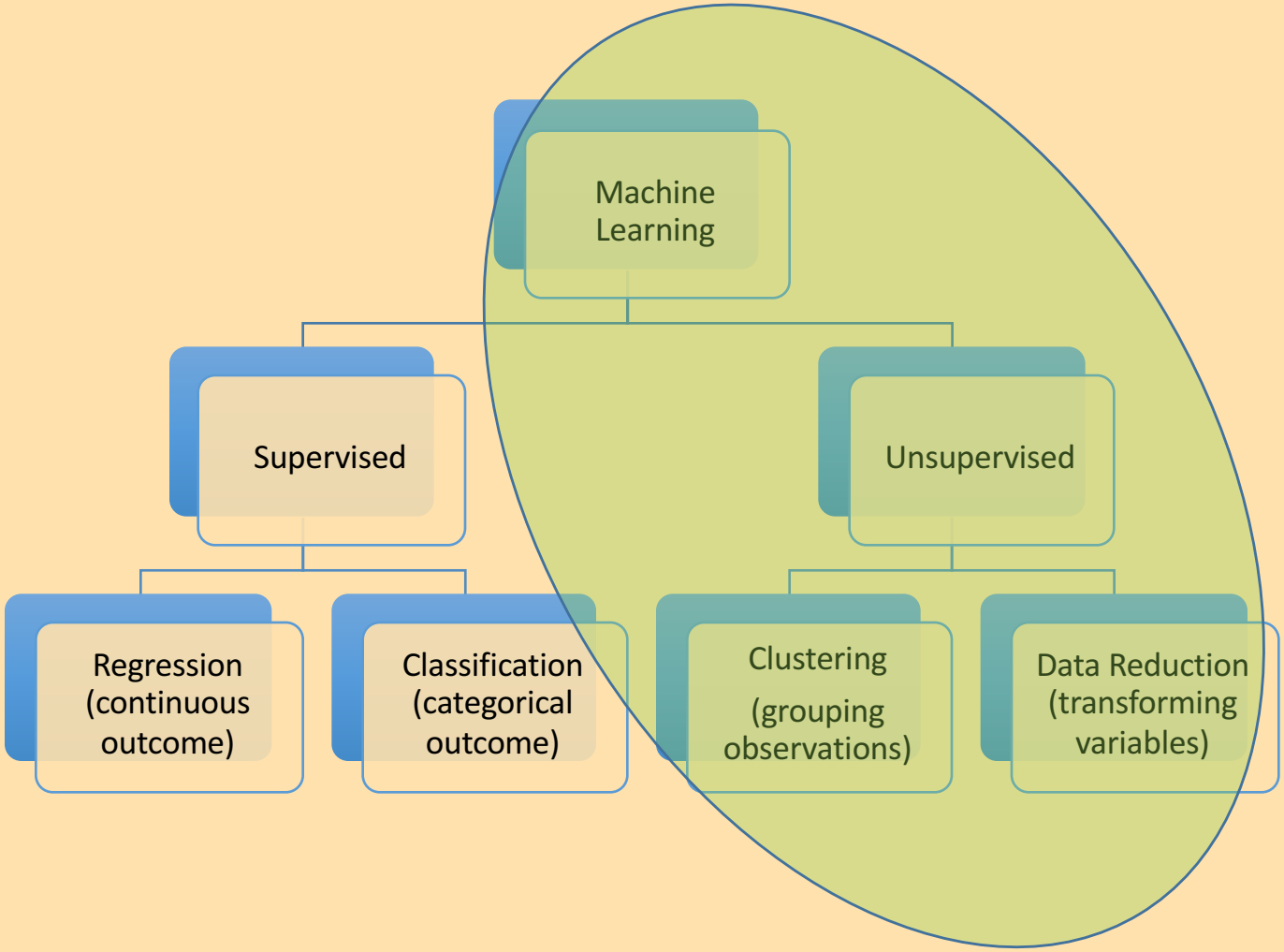
Overview

- Clustering
  - K-means clustering
  - TwoWay and Kohonen
- Data reduction
  - Principal components analysis (PCA)
  - Independent components analysis (ICA)
- Return to Classification
  - Neural Networks (Neural Nets)
  - Bayesian Networks (Bayes Nets) – causal interpretation
- Thursdays lab and HW #5 will reinforce these methods along with the classification and regression methods of last weeks lectures

Supervised vs. unsupervised



Supervised vs. unsupervised



Supervised vs unsupervised

- **Supervised Learning**
- Have data where you know outcome in order to *train* the model
- Train the model to predict future outcomes from sets of predictors

Prediction: Regression (continuous outcomes) and classification (categorical outcomes)

- **Unsupervised Learning**
- No desired outcomes
- Separate data points into groups with "similar" properties (clustering)
- Or reduce data to fewer variables in some way that still captures important structure of the data

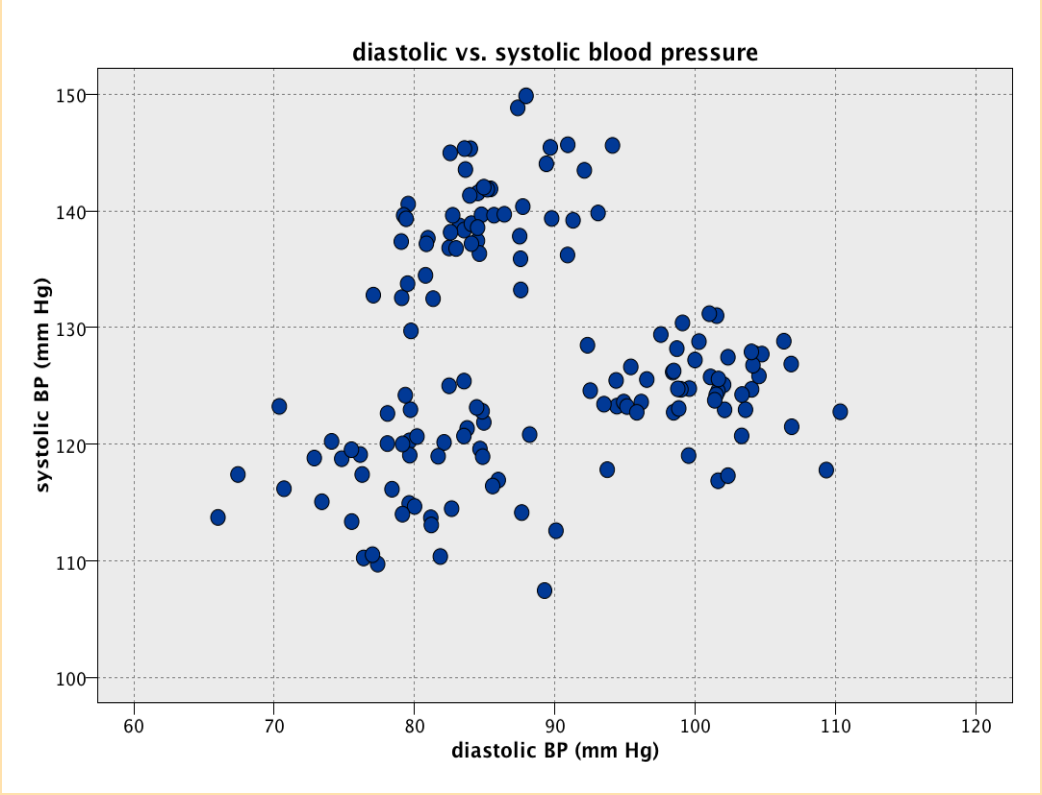
Clustering and data reduction

Clustering

Clustering problems

- Determine whether there are distinct (unknown) subclasses of brain cancer based on symptomatic and tumor characteristics
- Determine whether there are distinct patterns of gene expression within a patient population

2D Clustering example

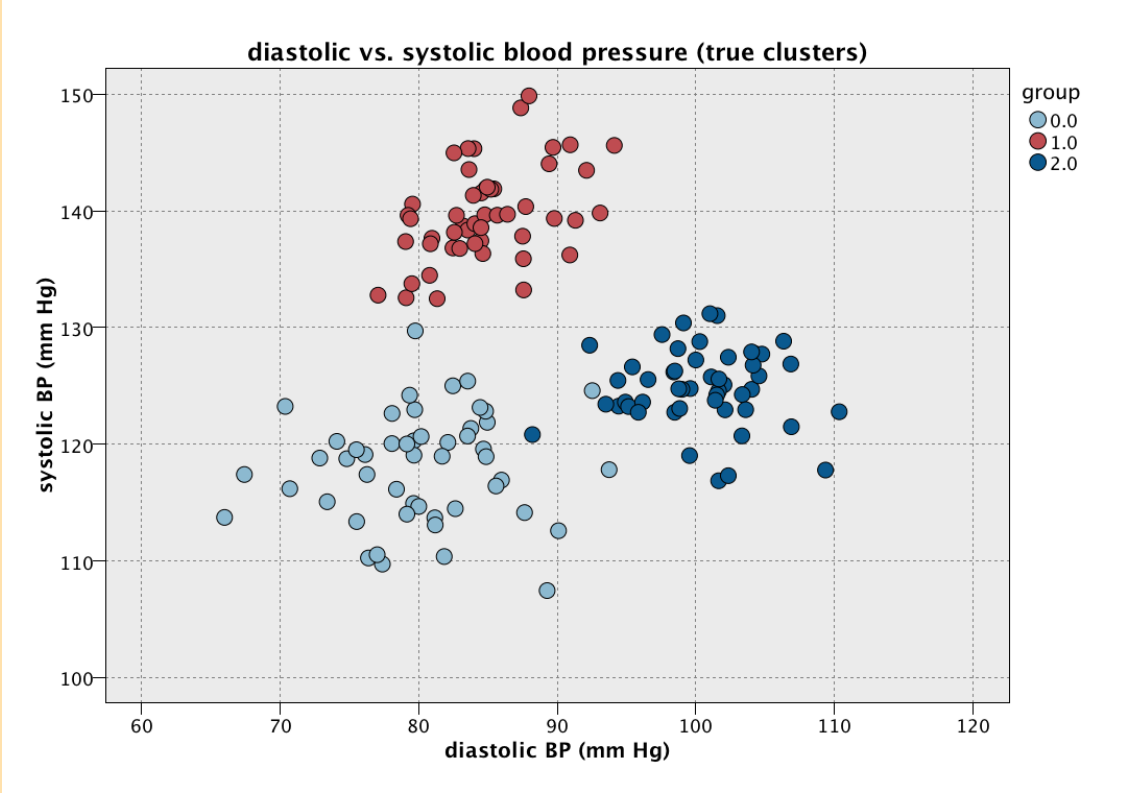


How many clusters?

Where are the clusters?

2D clustering example

The truth

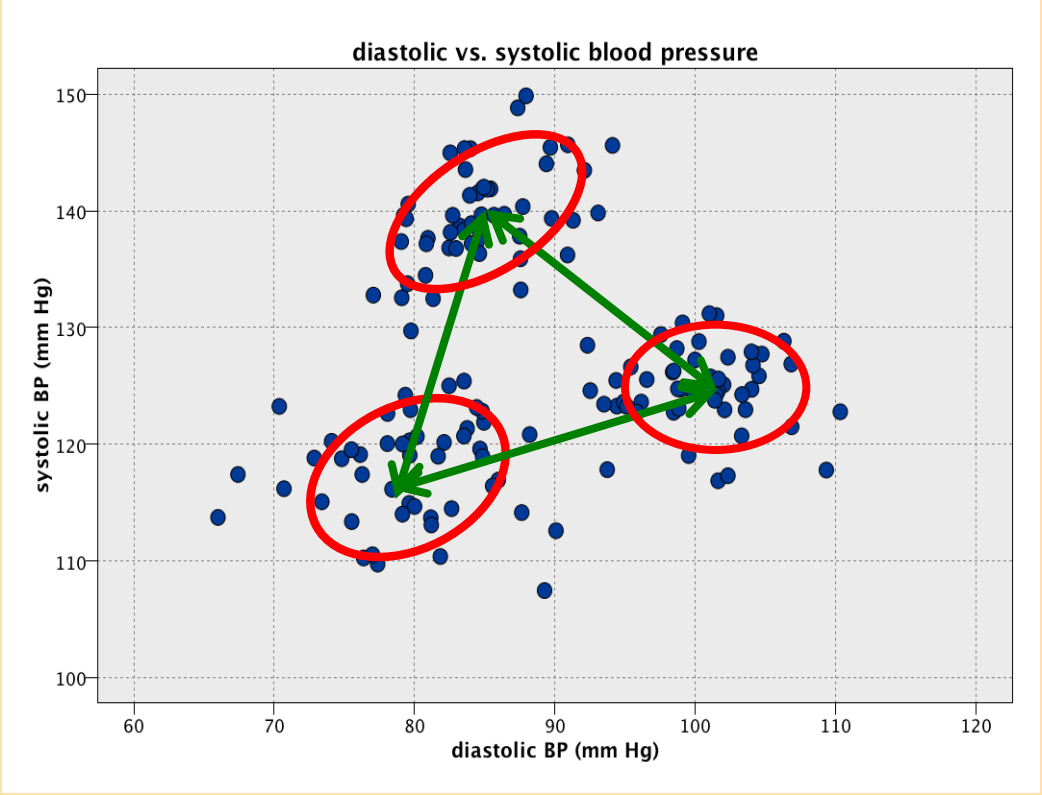


Truth has 3 clusters (kind of obvious)

Cluster Analysis (Data Segmentation)

- Objective is to group (or segment) collection of objects into subsets (or *clusters*)
- There is no “Target” in cluster analysis (Unsupervised)
- Objects within a cluster are more similar (or less dissimilar) than objects in different clusters

2D clustering example



We want long green lines (large distance between cluster centers)

We want small ellipses (points within clusters grouped tightly together)

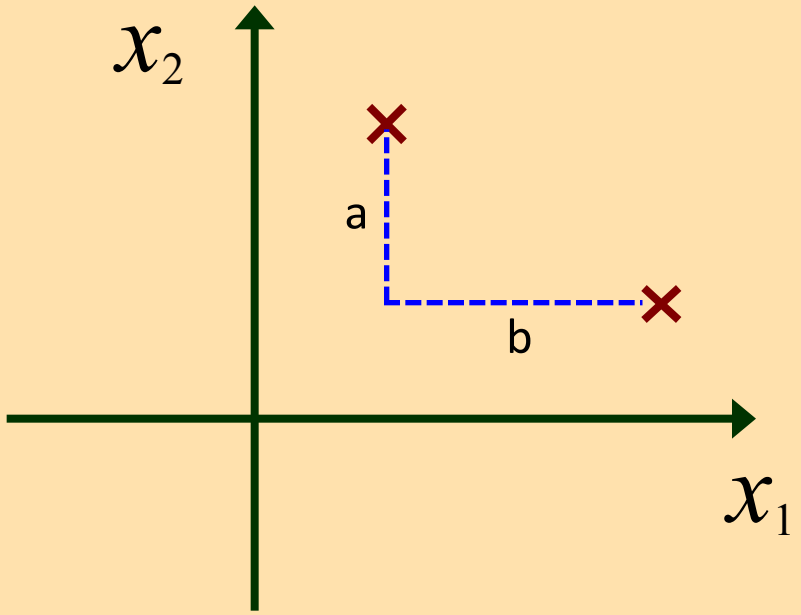
Maximize similarity within clusters and minimize similarity between clusters

Cluster Analysis (Data Segmentation)

- Need to define similarity (or dissimilarity) and we need to be able to measure it
- Want to maximize similarity (or minimize dissimilarity) within clusters and minimize similarity (maximize dissimilarity) between clusters

Measures of similarity/dissimilarity

Most common choice for dissimilarity between 2 instances is the squared difference summed over all attributes (higher values = more dissimilar)



Consider 2 observations with only 2 variables x_1 and x_2 – dissimilarities could be

- Square diff = $a^2 + b^2$
- Absolute diff = $|a| + |b|$
- Can also differentially weight attributes e.g. $ka^2 + mb^2$

Extension to 3D would be e.g. $a^2 + b^2 + c^2$, and so on for higher dimensions

Measures of similarity/dissimilarity

- Generally need to “standardize variables”
- Ordinal variables are treated the same as continuous variables
- Nominal variable differences between groups need to be explicitly given. A common choice is difference 1 if the groups are the same and 0 if they are different. (This then needs to be standardized if any non-categorical variables are also considered.)
- Remember, we want to maximize similarity (or minimize dissimilarity) within clusters and minimize similarity (maximize dissimilarity) between clusters

K-means clustering

K-means clustering overview

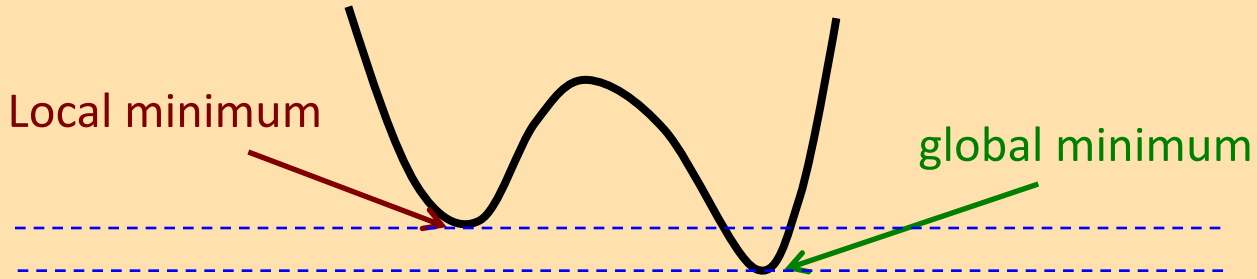
- Algorithm to split set of observations into K clusters, where K can equal 2, 3, 4, …
- K is pre-chosen (though often choose a set of different K's to try
- Uses square difference dissimilarity measure $a^2 + b^2 + c^2 + \ldots$
- Procedure: starts from a random cluster assignment of observations, and then iterates toward increased similarity (reduced dissimilarity) – estimating cluster centers on each iteration
- Common practice to try a bunch of different values for K and use some kind of measure to determine the best K after the fact. Modeler uses the "Silhouette measure" for this purpose.

K-means algorithm

- Algorithm proceeds as follows:
  1. Randomly assign a cluster number (between 1 and K) to each of the observations/instances in the dataset. These are the *initial* cluster assignments
  2. Iterate between the following until the cluster assignments no longer change:
     a. For each cluster determine the center (centroid) of the cluster
     b. Assign each observation to the cluster whose center is *closest* to them in terms of squared differences (i.e., the center that is the most *similar* to the observation)
- The algorithm increases the within-cluster similarity relative to the between cluster similarity at every step

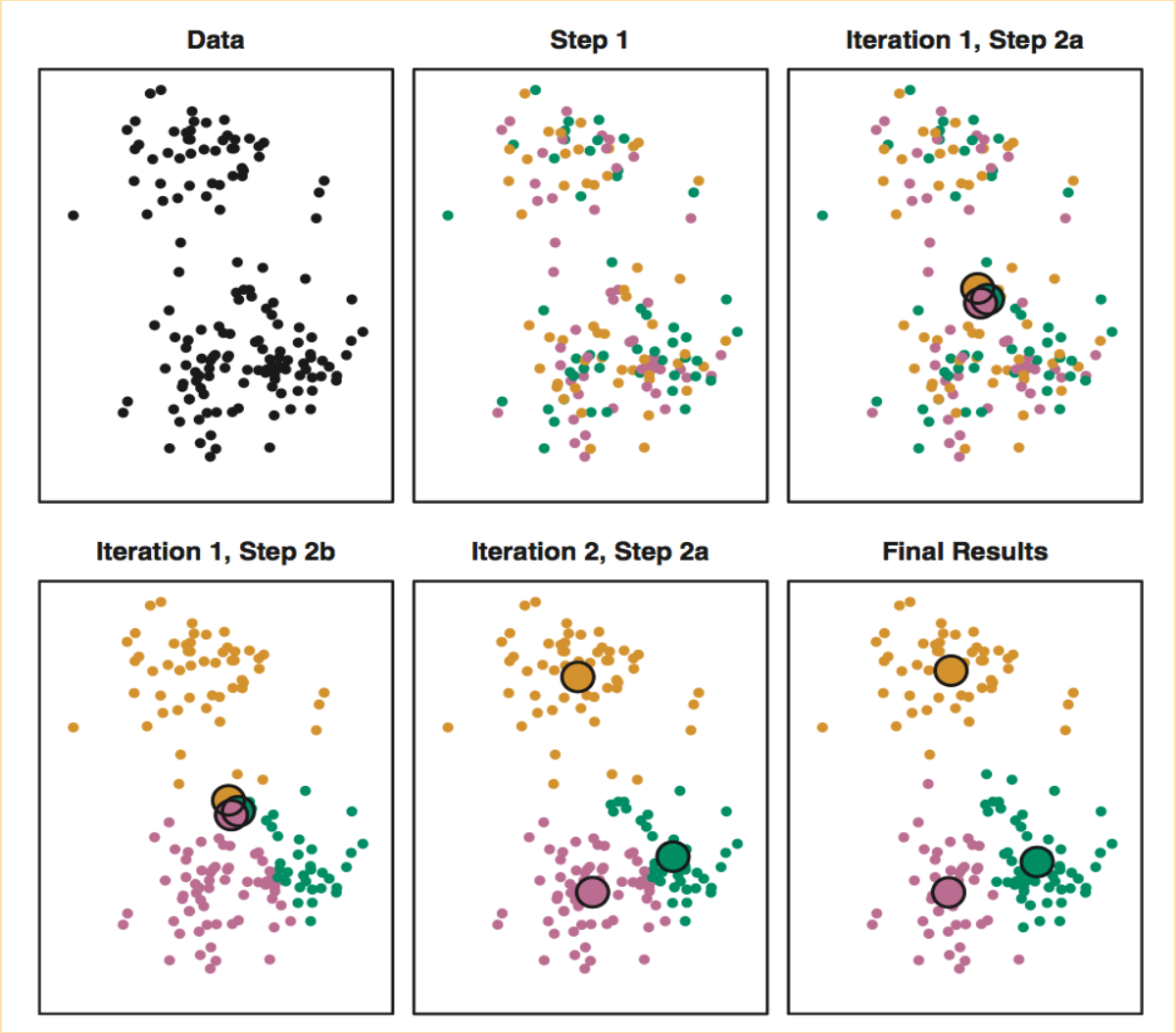
K-means algorithm

- The center of a cluster is simply the point associated with the mean of each of the attributes for the observations assigned to that cluster – e.g. in the Age direction the center would be located at the mean Age for all subjects in the cluster.
- Note that for different initial random assignments of observations to clusters, K-means may lead to different final clustering (it is a *local* not *global minimizer* of dissimilarity)



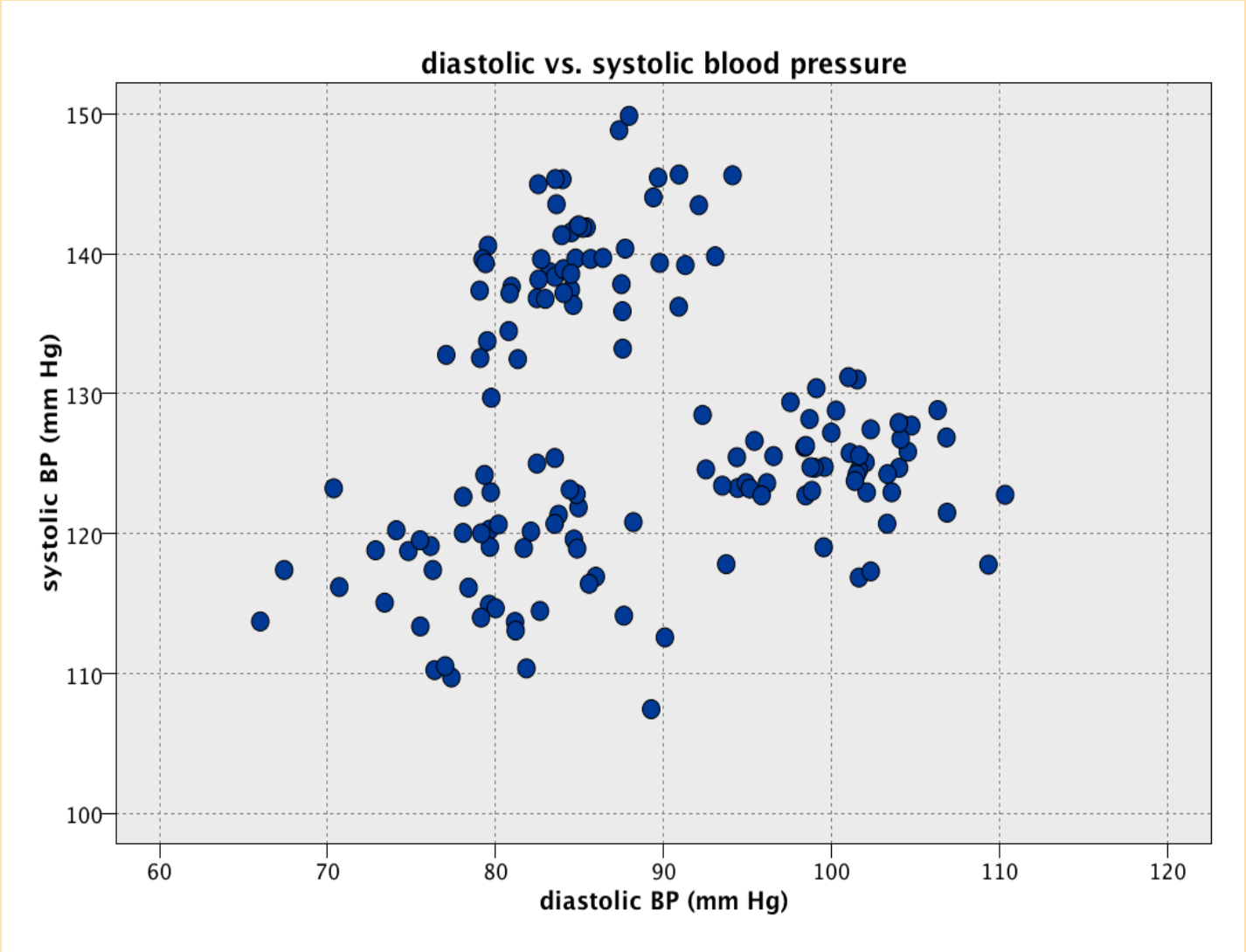
Upshot - may not converge to best answer

K-means algorithm



Taken from ISLR,
James et al., 2013
p. 389

K-means BP Modeler example



K-means BP Modeler example

Type Node

Preview

Types | Format | Annotations

Read Values | Clear Values | Clear All Values

Field	Measurement	Values	Missing	Check	Role
subjid	Nominal	1.0,2.0,...		None	Record...
systolic	Continuous	[107.44...		None	Input
diastolic	Continuous	[65.993...		None	Input
group	Nominal	0.0,1.0,...		None	None

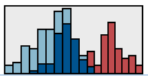
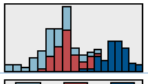
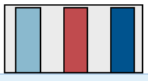
Table Node

Table | Annotations

	subjid	systolic	diastolic	group
1	1	113.69	81.16	0
2	2	112.58	90.09	0
3	3	118.74	74.81	0
4	4	119.58	84.67	0
5	5	120.05	78.07	0
6	6	125.00	82.47	0
7	7	113.71	65.99	0
8	8	122.95	79.72	0
9	9	116.91	85.97	0
10	10	121.35	83.76	0
11	11	121.86	84.95	0
12	12	117.39	67.42	0
13	13	120.28	79.65	0
14	14	116.18	70.70	0
15	15	114.47	82.64	0
16	16	124.19	79.35	0
17	17	119.05	79.66	0
18	18	109.71	77.38	0
19	19	114.91	79.62	0
20	20	122.82	84.83	0
21	21	116.13	78.40	0
22	22	120.69	83.52	0
23	23	113.36	75.53	0
24	24	118.93	84.86	0
25	25	116.41	85.57	0
26	26	119.11	76.13	0
27	27	118.81	72.86	0
28	28	124.58	92.53	0
29	29	115.06	73.40	0
30	30	120.01	79.15	0
31	31	129.71	79.75	0
32	32	120.65	80.18	0
33	33	113.07	81.19	0
34	34	107.45	89.27	0
35	35	117.81	93.74	0

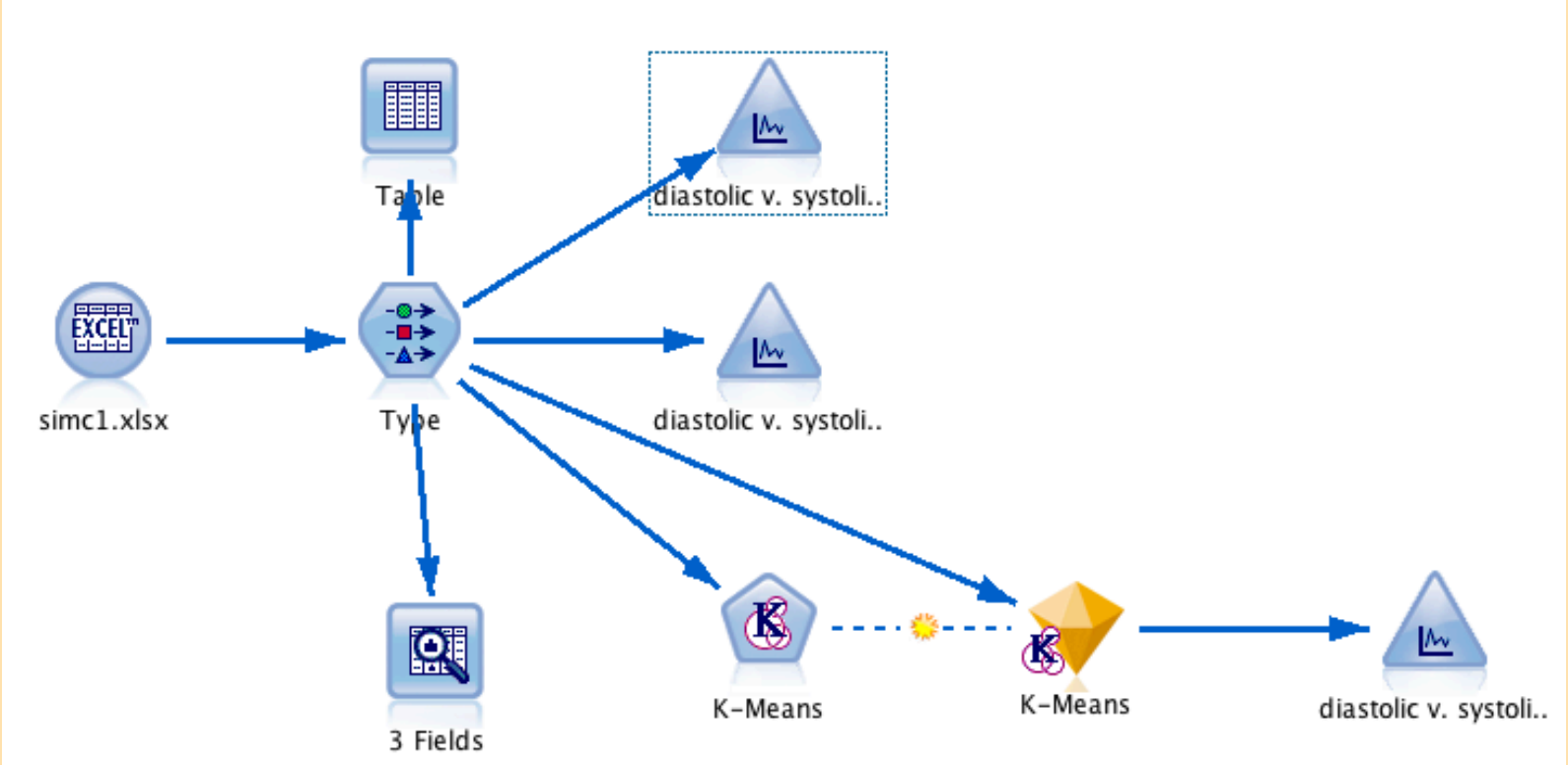
Audit Node (include group as Target for visualization only)

Audit | Quality | Annotations

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
subjid		Nominal	1.000	150.000	--	--	--	150	150
systolic		Continuous	107.449	149.852	127.557	9.944	0.286	--	150
diasto...		Continuous	65.993	110.344	88.461	9.896	0.266	--	150
group		Nominal	0.000	2.000	--	--	--	3	150

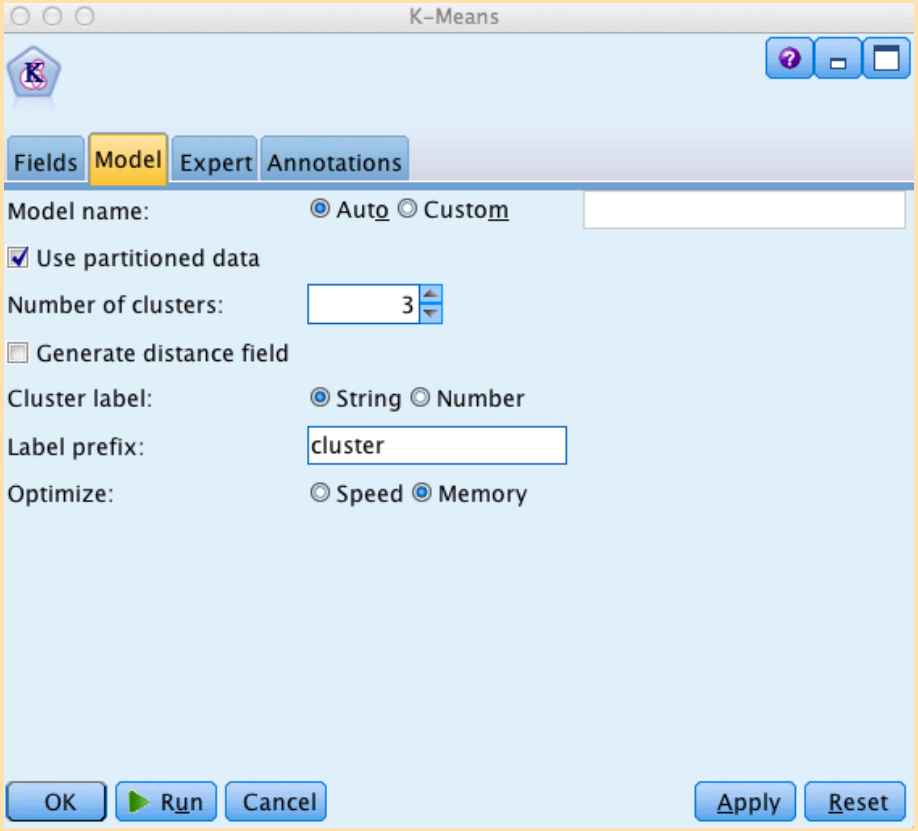
[1] Indicates a multimode result [2] Indicates a sampled result

K-means BP Modeler example



K-means Node

Pre-defined roles in Fields tab (all inputs) – systolic and diastolic are input variables



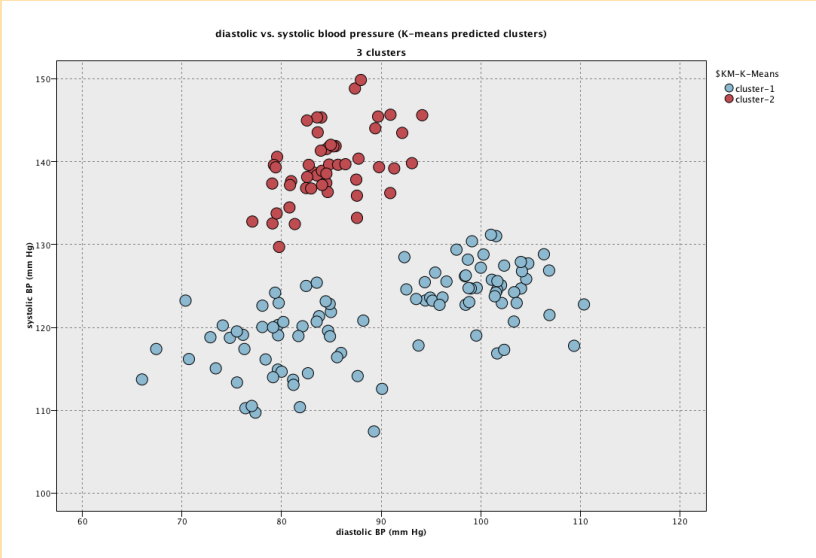
Use defaults in Expert tab

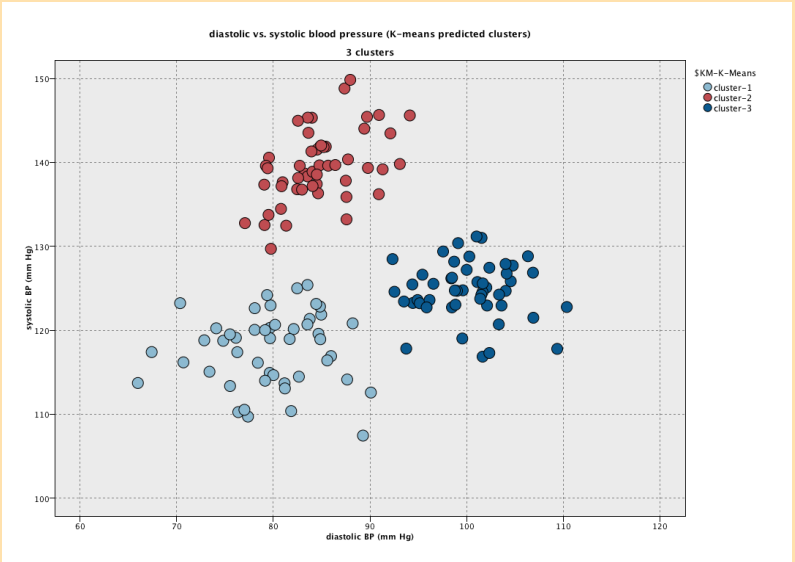
K-Means in IBM Modeler automatically normalized inputs using the following formula:

$$x_{i,\ \text{new}} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

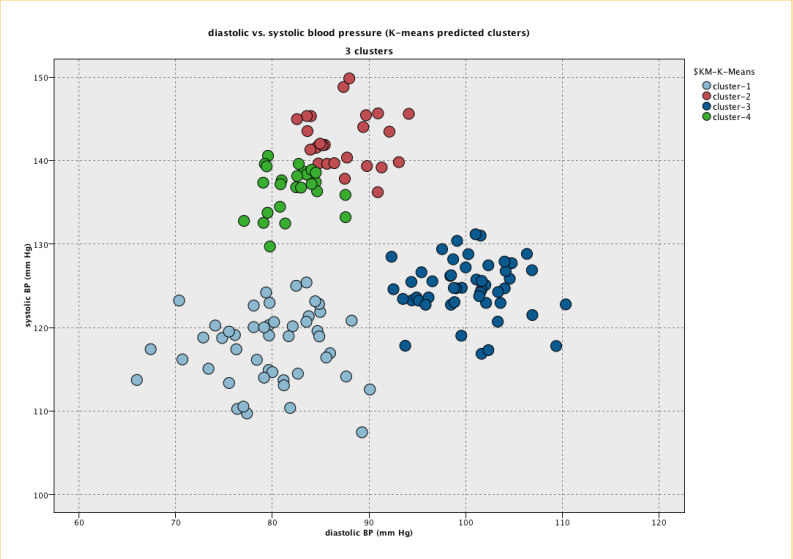
K-means results

K=2





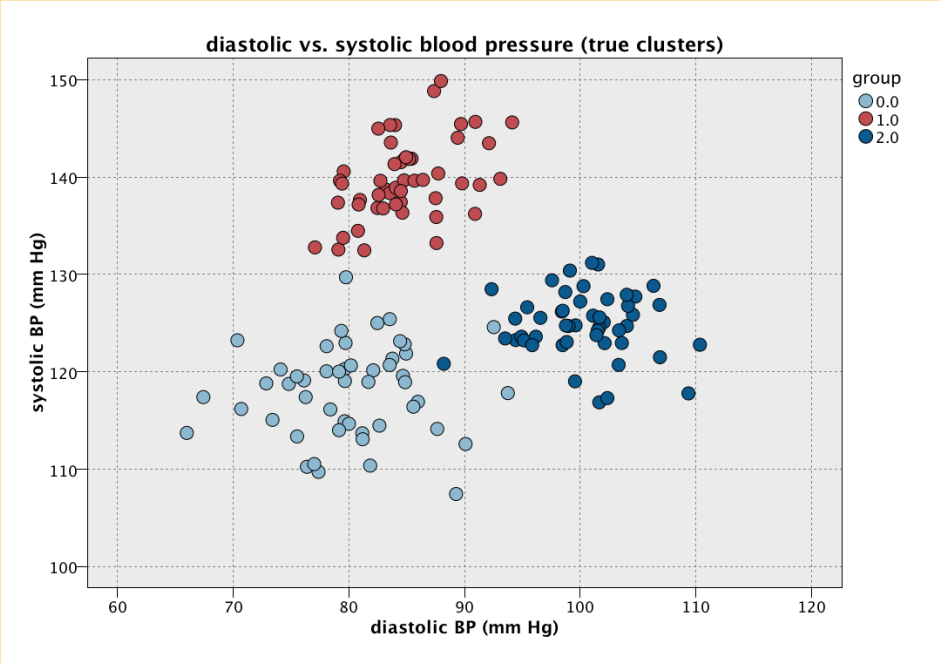
K=3



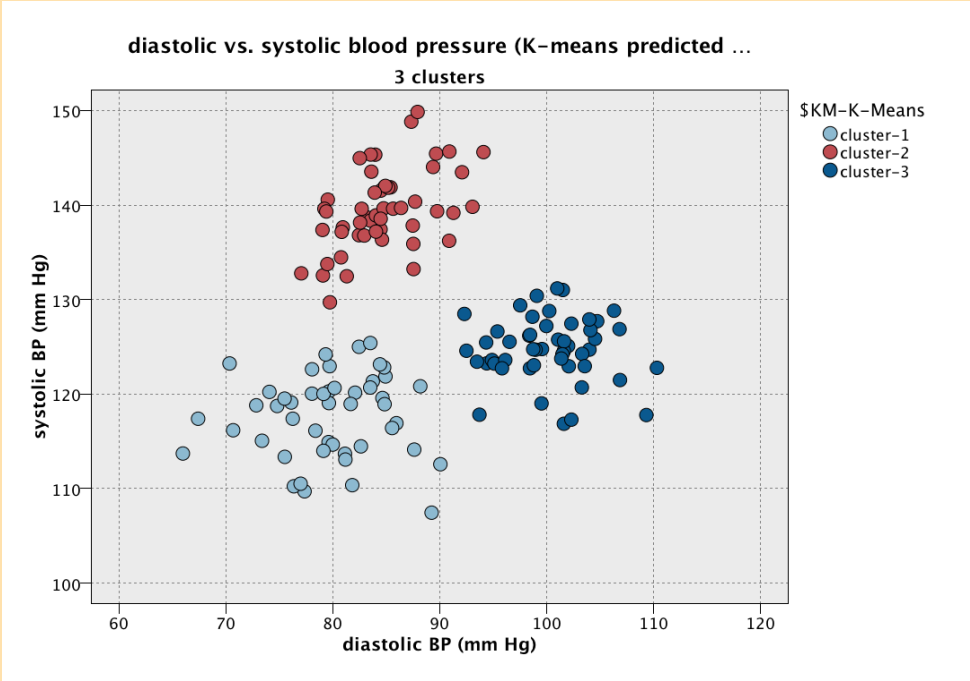
K=4

K-means example

The truth

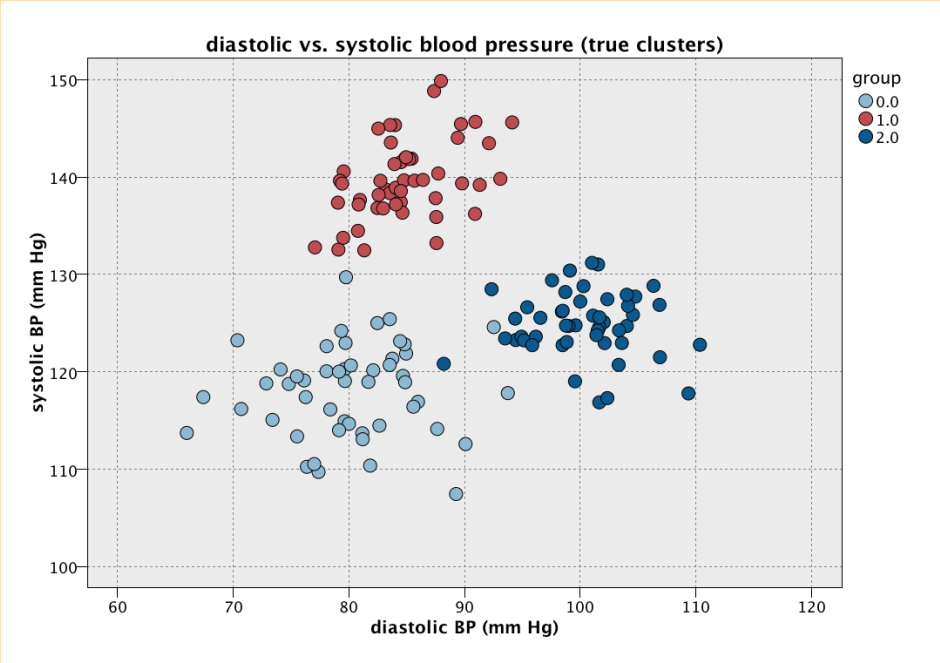


K-means (K = 3)

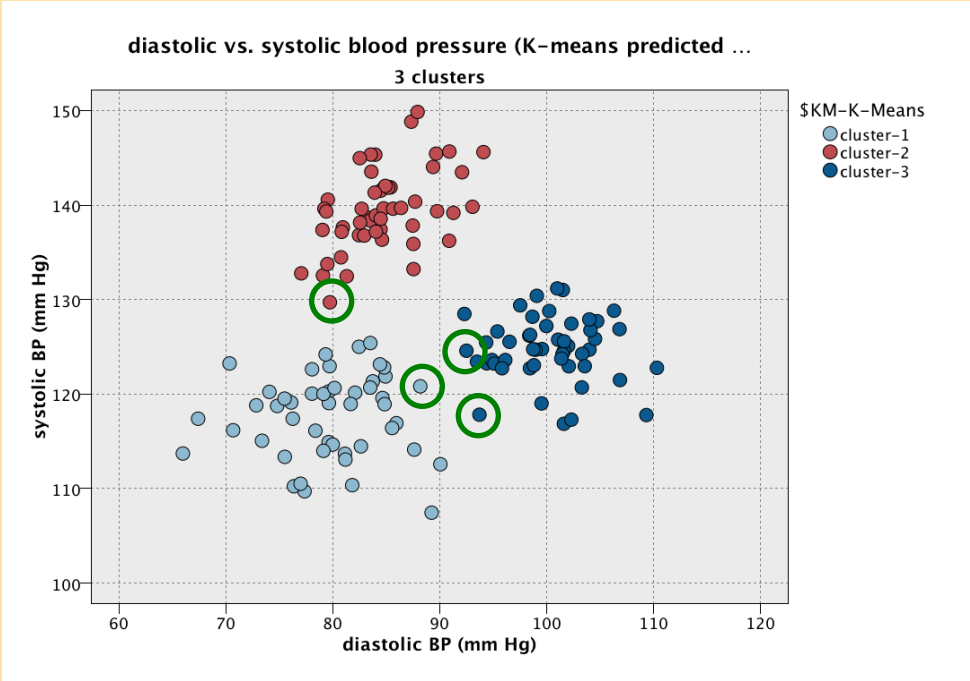


K-means example

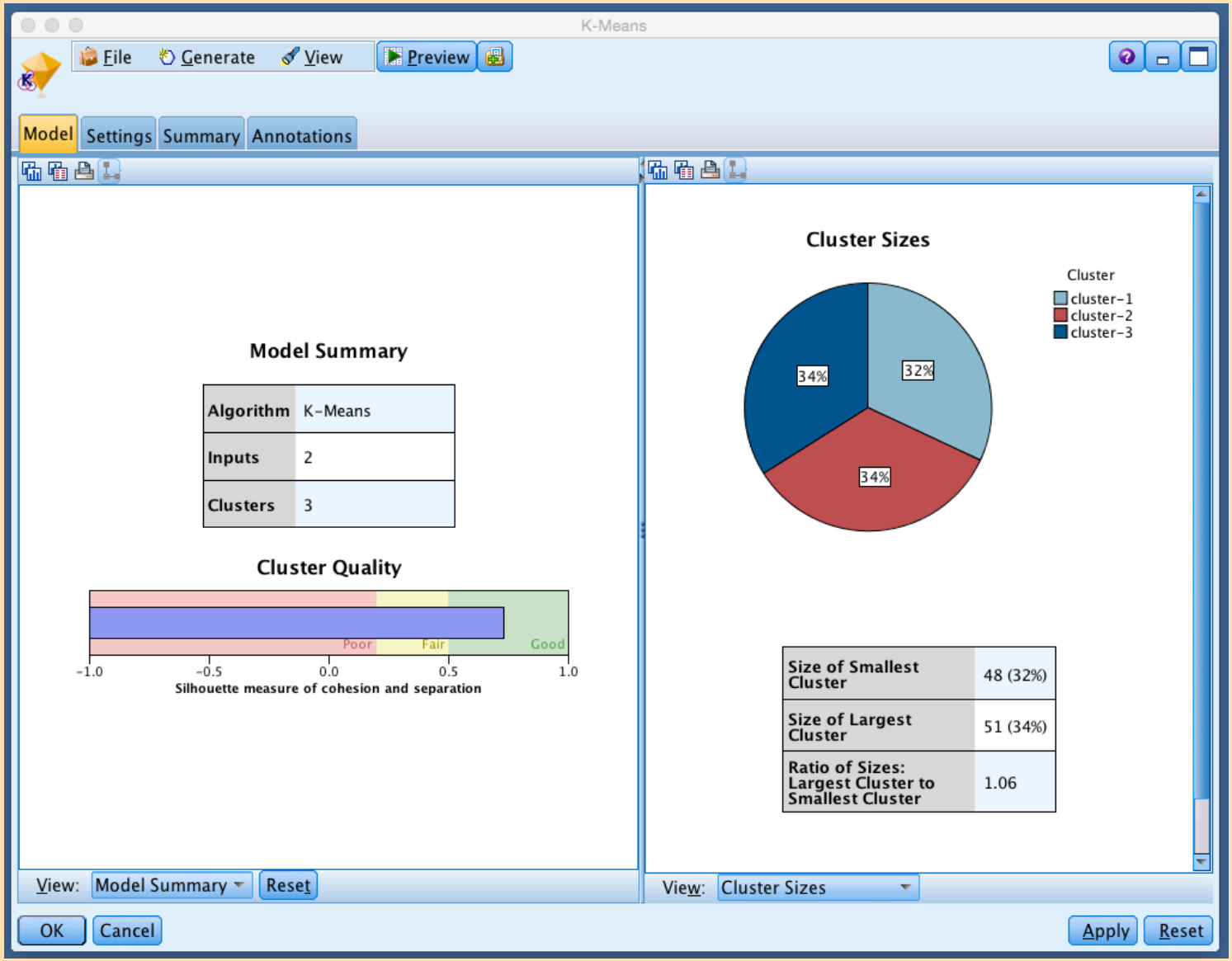
The truth



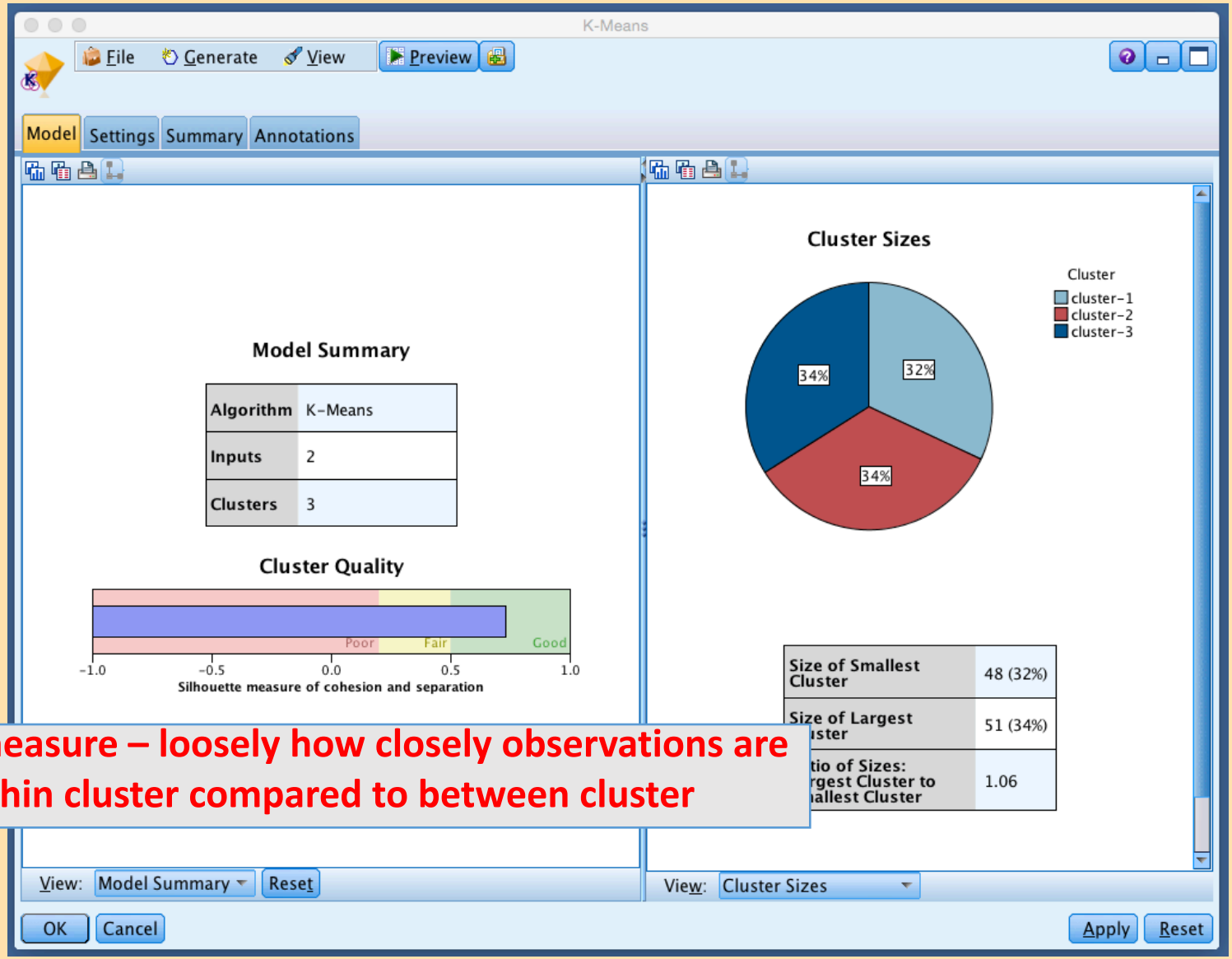
K-means (K = 3)



K-means example



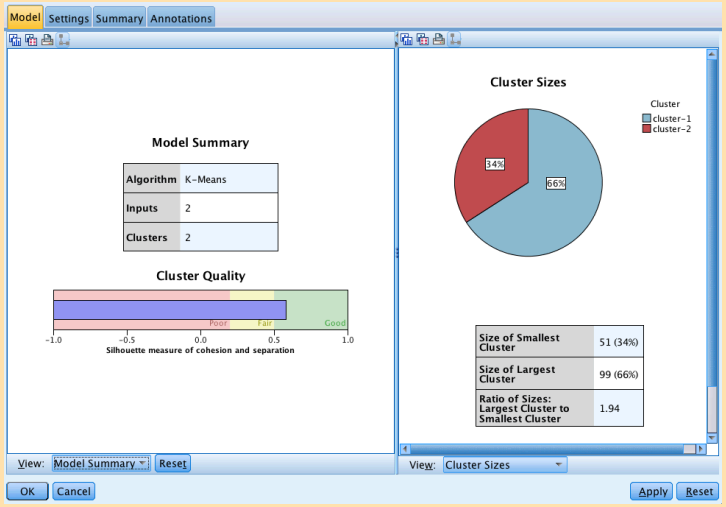
K-means example



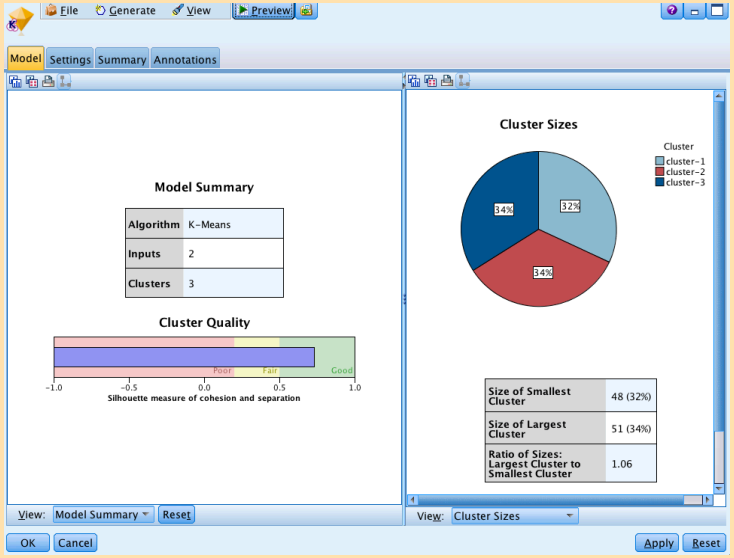
Silhouette measure – loosely how closely observations are matched within cluster compared to between cluster

K-means example

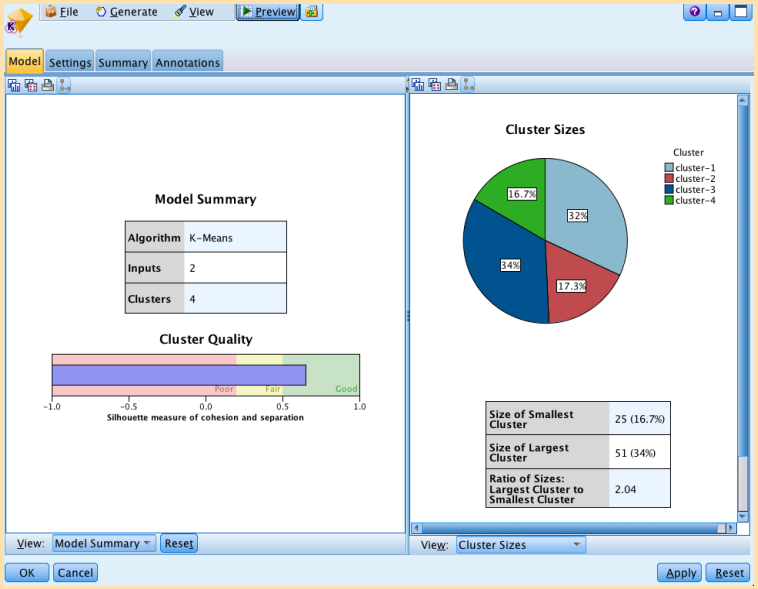
K=2



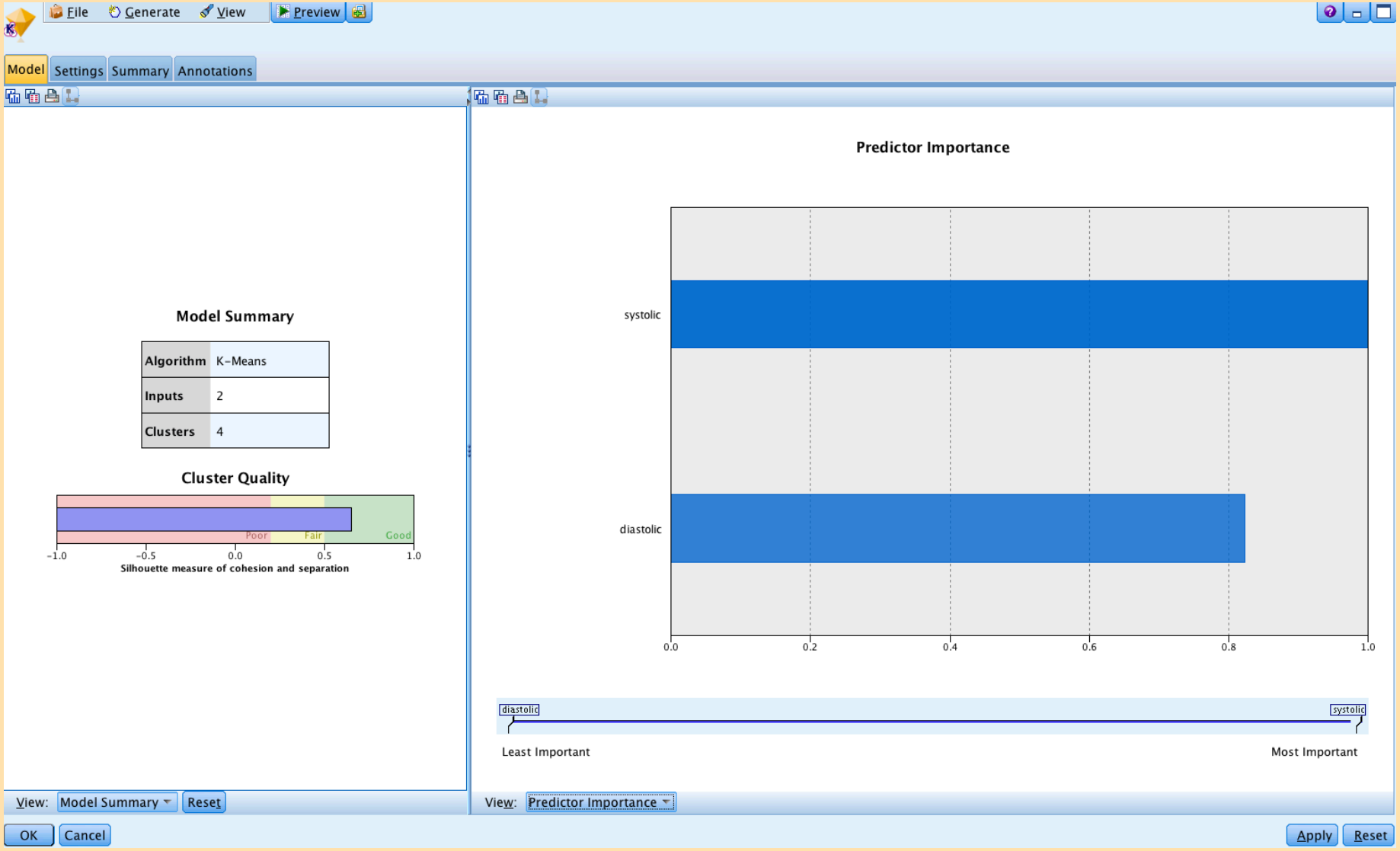
K=3



K=4



K-means example



K-means clustering

- K-means is fast – good for big data – in particular with large number of variables/attributes
- Number of clusters must be defined by user – only heuristic methods for choosing number of clusters (IBM Modeler gives “Silhouette measure”)
- Local minimizer not global –different starts can lead to different results

Other Clustering Algorithms in IBM Modeler (in brief)

Other Clustering Algorithms in IBM Modeler (in brief)

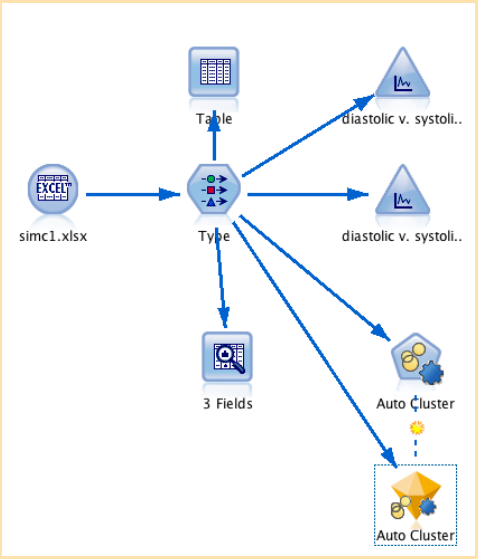
Two Step

- "Hierarchical clustering" method – successively splits data into increasing number of clusters (estimates optimal number of clusters – slower than K-means)

Kohonen

- Based on Neural Networks (estimates optimal number of clusters – slower than K-means)

Auto-Cluster



- Try all of Modelers methods and find “best”
- Uses K-means, TwoStep and Kohonen

File Generate Preview

Model Summary Annotations

Sort by: Use Ascending Descending Delete Unused Models View: Training set

Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster	Largest Cluster (%)	Smallest Largest	Impor...
☑		TwoStep 1	< 1	0.732	3	48	32	51	34	0.941	0.0
☐		K-means 1	< 1	0.561	5	25	16	48	32	0.521	0.0
☐		Kohonen 1	< 1	0.387	12	1	0	31	20	0.032	0.0

- Looks like TwoStep did “best” on the BP data

Data Reduction Methods

PCA and ICA

- Principal Components Analysis and Independent Components Analysis
- Methods to succinctly summarize high-dimensional data sets (transform many variables to a few variables) – but not assigning observations to clusters
- Many variables are summarized in terms of just a few new variables (the *components*)
- We will focus on PCA

Principle components analysis (PCA)

PCA

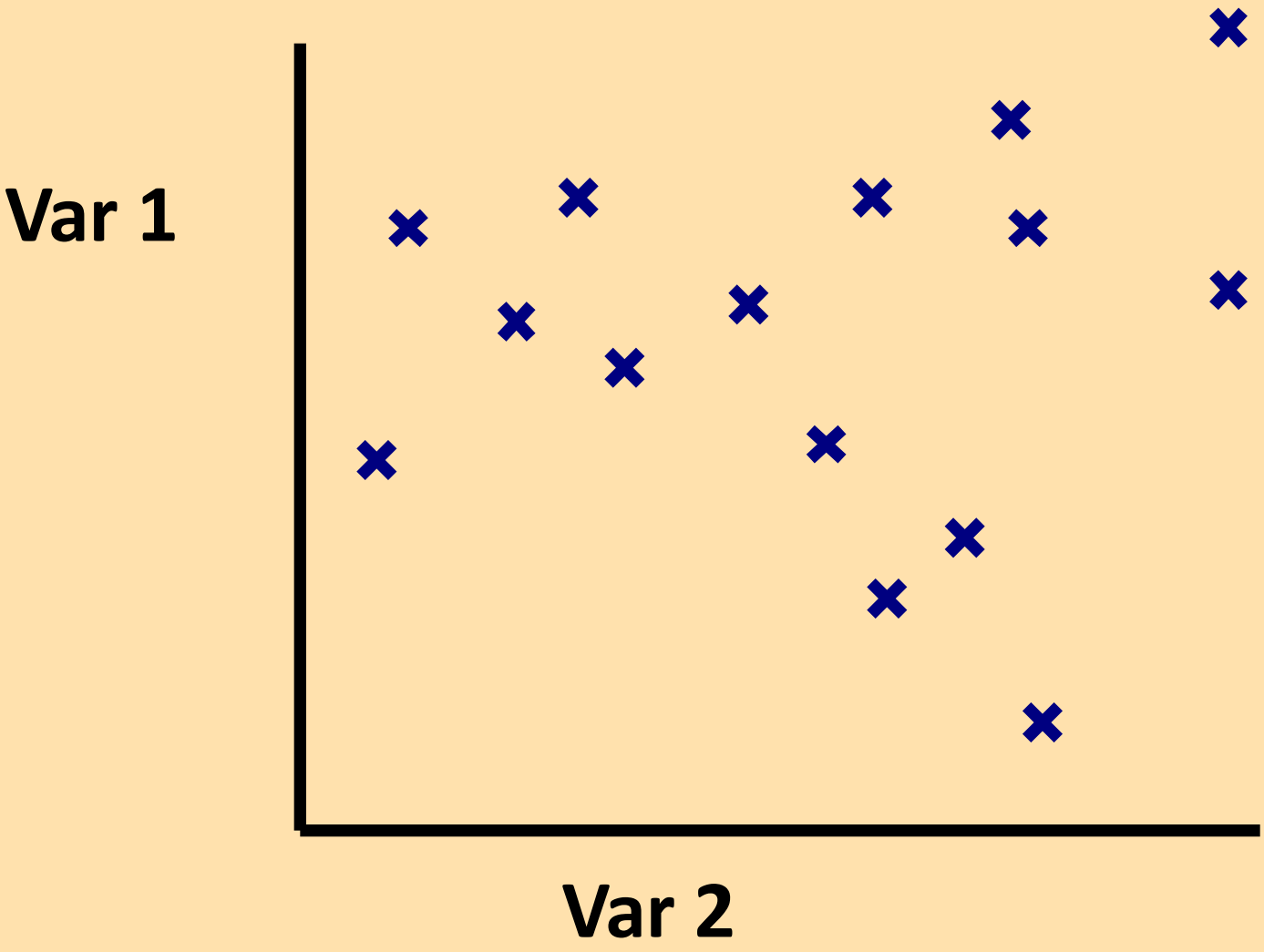
- Finds weighted sums of the variables (attributes) that maximally explain variability in the data
- The first PC explains as much of the variability in the full data set as possible
- The second PC explains as much variability as possible after the variability from the first (PC) has been removed
- Constraint: each PC is a weighted sum (*linear combination*) of the original variables

Each PC is a weighted average of the original variables.

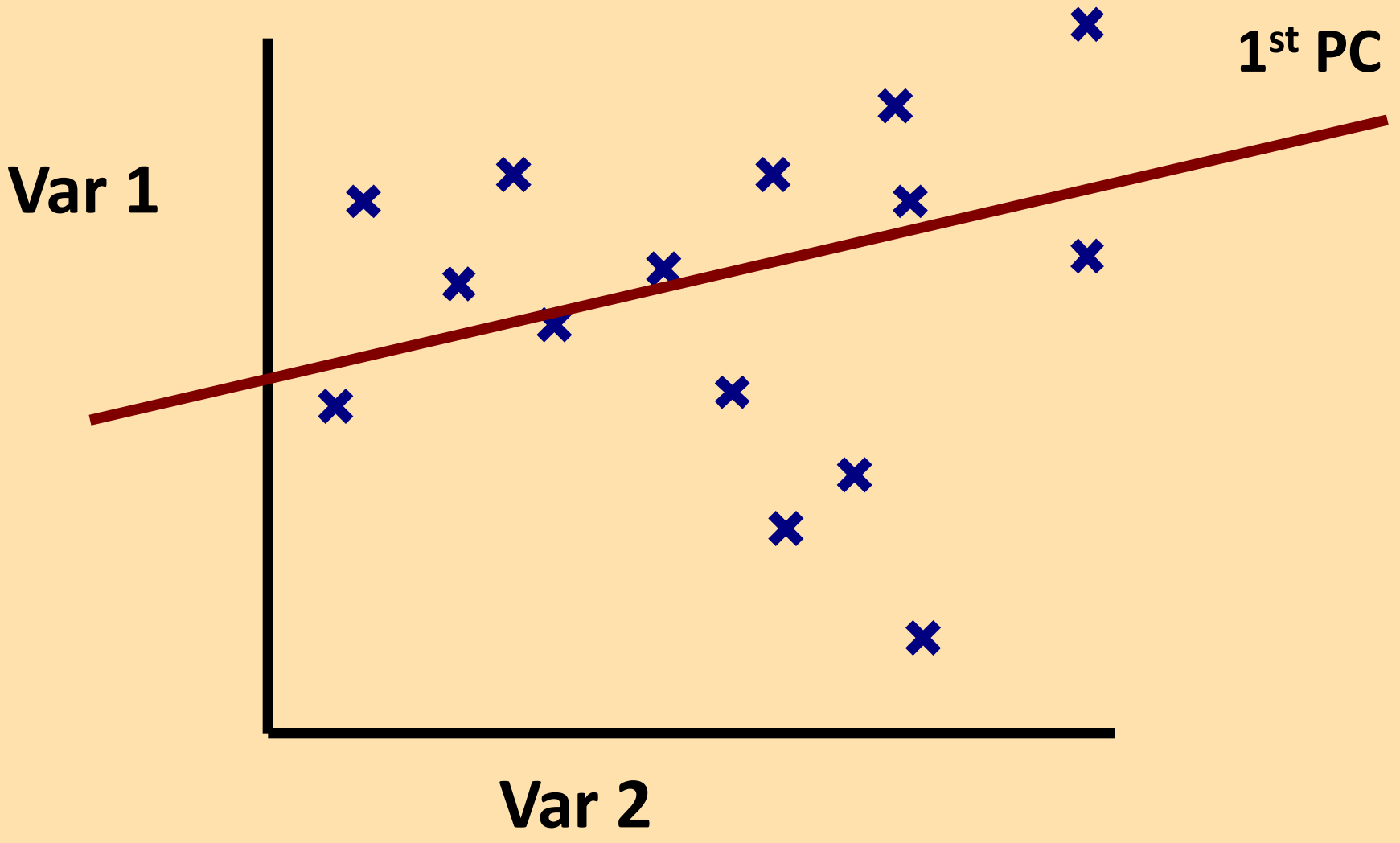
The data reduction goal is to *closely* represent the full data set in terms of just a few of these PCs

Sometimes it is also *hoped* that each of the PCs is meaningful

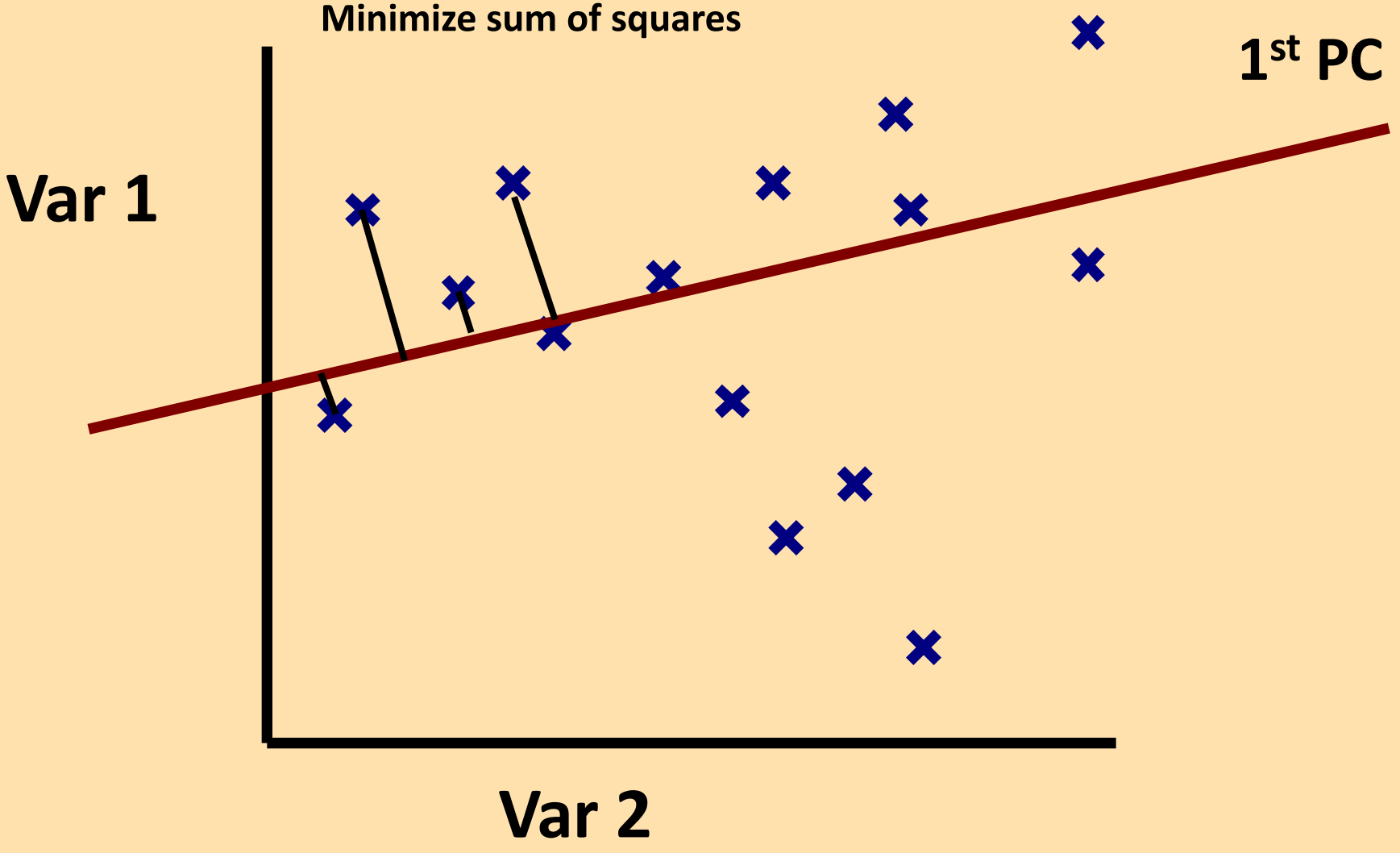
2D illustration



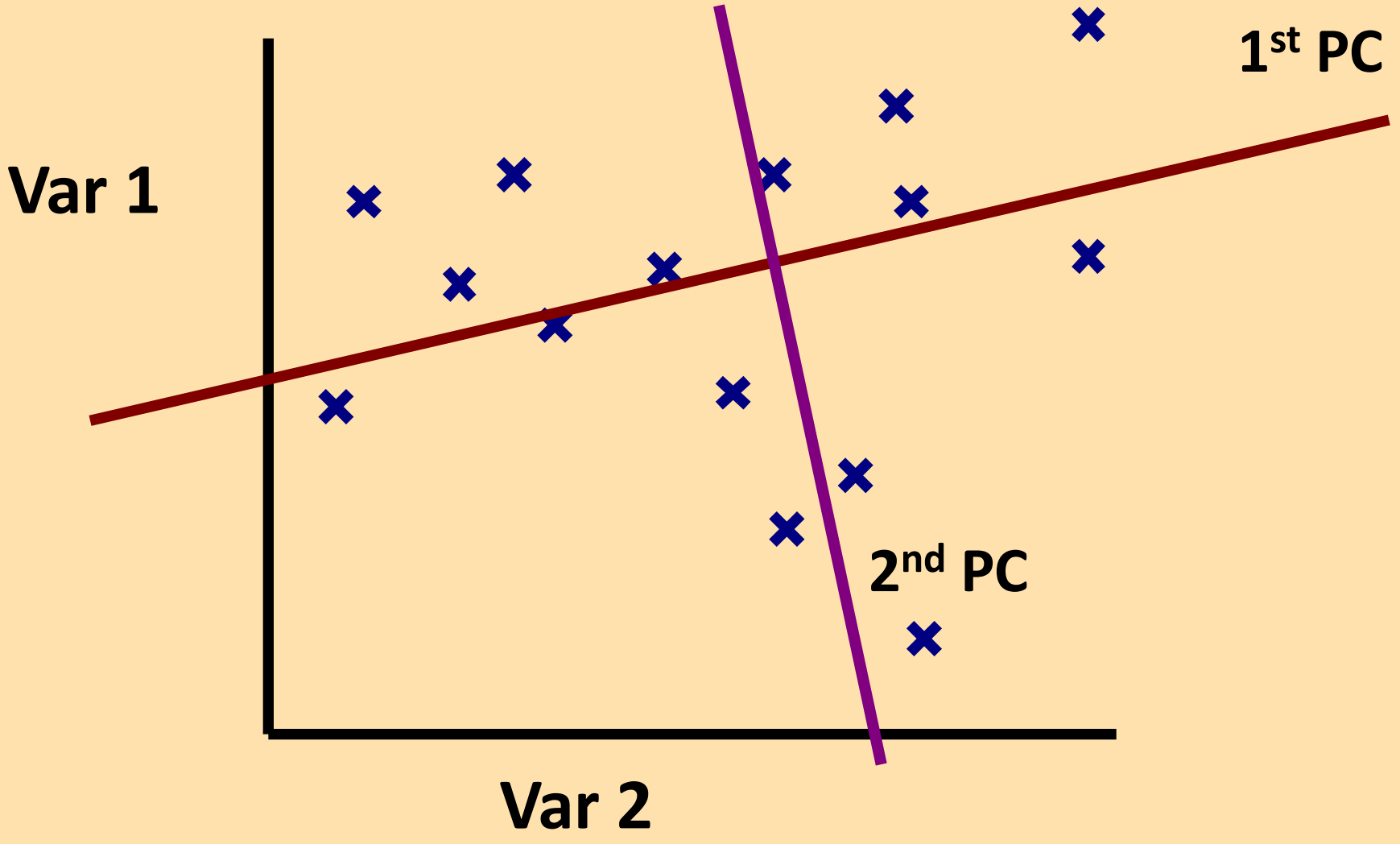
2D illustration



2D illustration

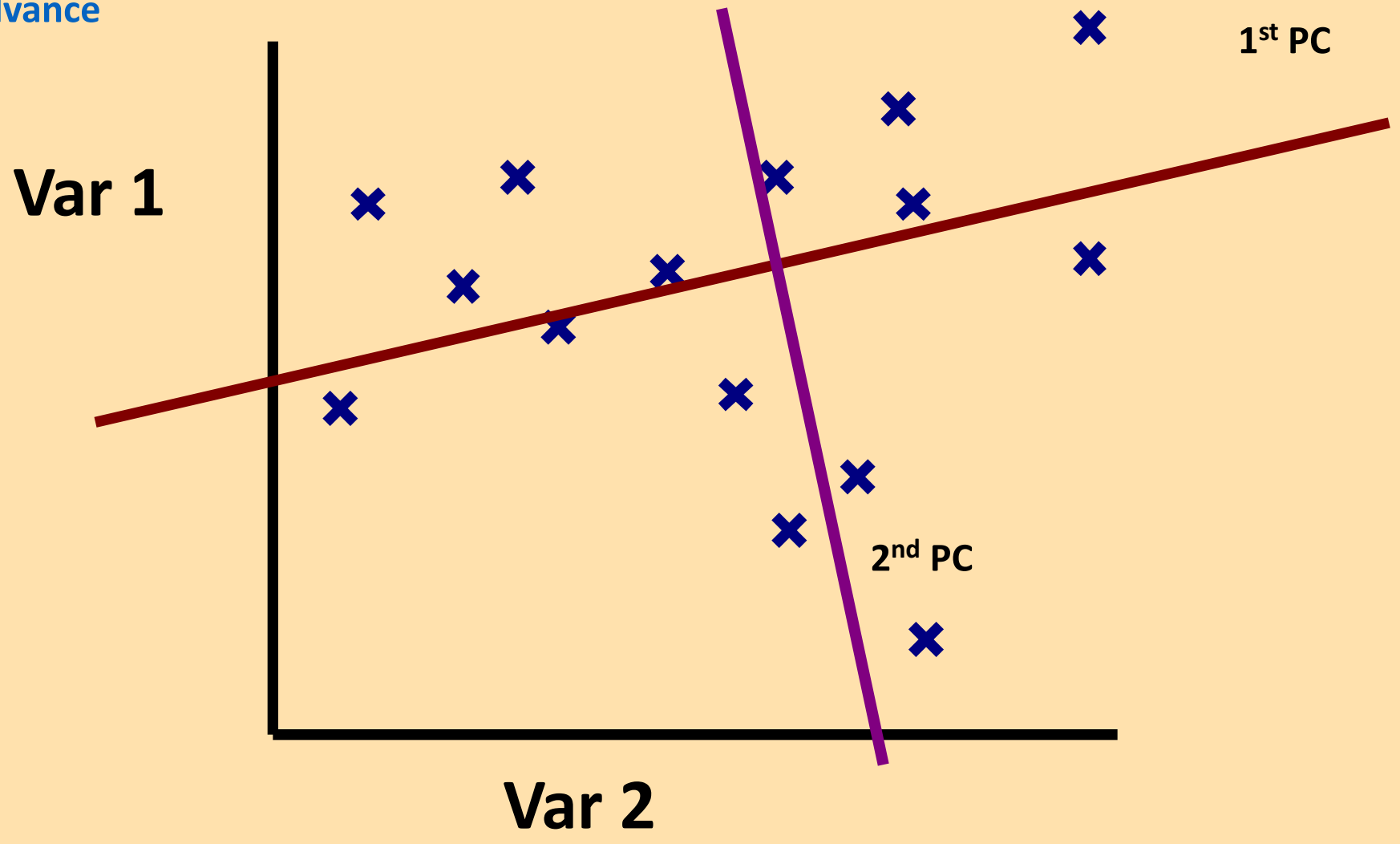


2D illustration



Note: that there is also a potential issue of scaling – variables often normalized in advance

2D illustration



PCA example – knee shape

- Is the bone in ACL injured knees a different "shape" than a group of control knees?
- May suggest that particular bone shapes predispose to ACL injury.

Osteoarthritis and Cartilage

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

Three-dimensional MRI-based statistical shape model and application to a cohort of knees with acute ACL injury

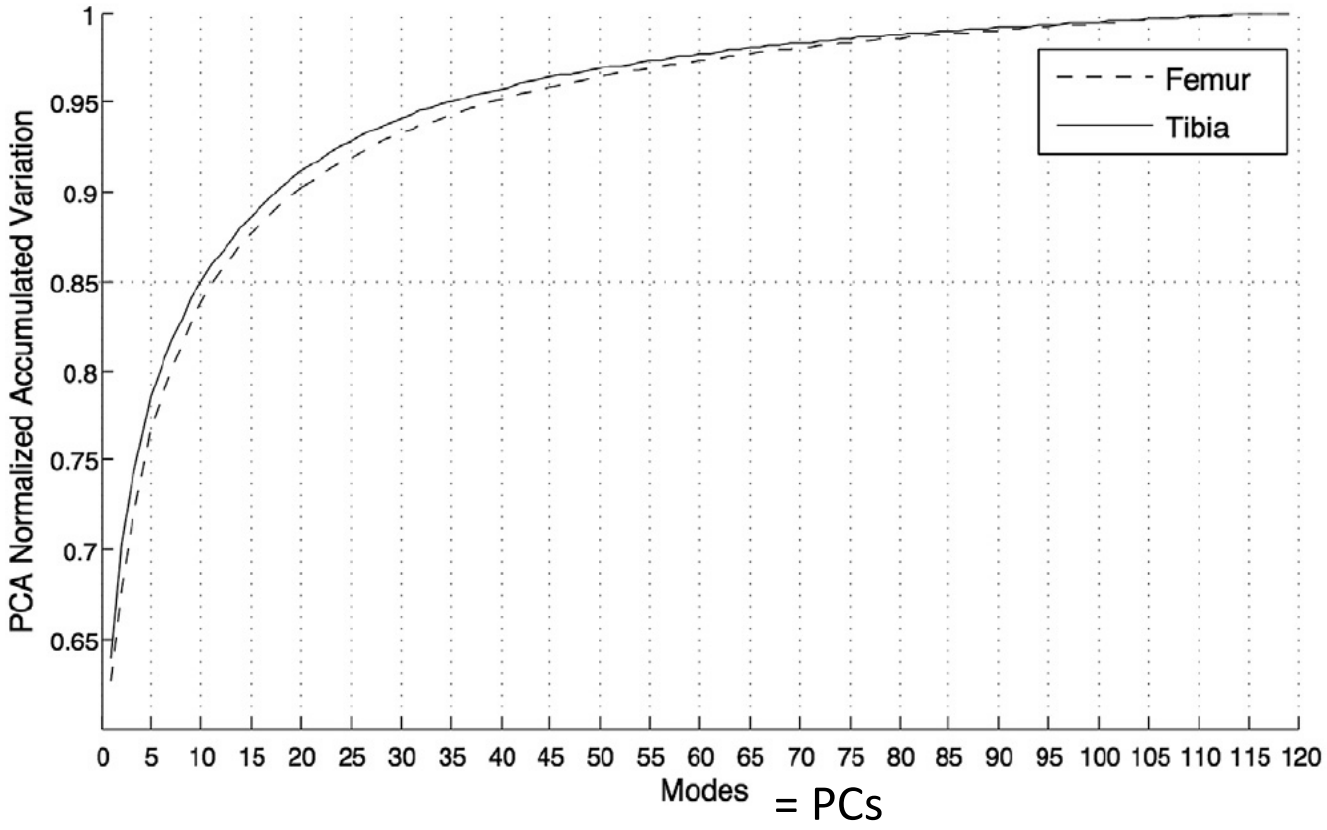
V. Pedoia † *, D.A. Lansdown ‡, M. Zaid †, C.E. McCulloch §, R. Souza ‡, C.B. Ma ‡, X. Li †

† *Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA*
‡ *Department of Orthopaedic Surgery, University of California, San Francisco, USA*
§ *Department of Epidemiology & Biostatistics, University of California, San Francisco, USA*

PCA example – knee shape

- Identify a large number of specific locations on the knee bones (“landmarks”).
- Measure the distance between all pairs of landmarks to quantify the shape.
- Used >8,000 landmarks in the tibia and >11,000 in the femur.
- All told there were >32,000,000 possible distances for the tibia and >62,000,000 for the femur to use as predictors to distinguish ACL vs non-ACL.
- Need to reduce!

PCA illustration – knee shape



= PCs

Fig. 4. Representation of the model's compactness: cumulative % of the variability expressed in function of the number of modes considered in the model.

For the femur, summarized 90% of the variation in 62 million variables using only about 20 principal components.

For the tibia, the first mode is related to size, the second to the medial posterior curvature of the tibial plateau, and the third to elevation of the anteromedial tibial plateau.

A number of the PCs (called modes here) had reasonable interpretations.

Though this is by no means a guarantee and is somewhat unusual.

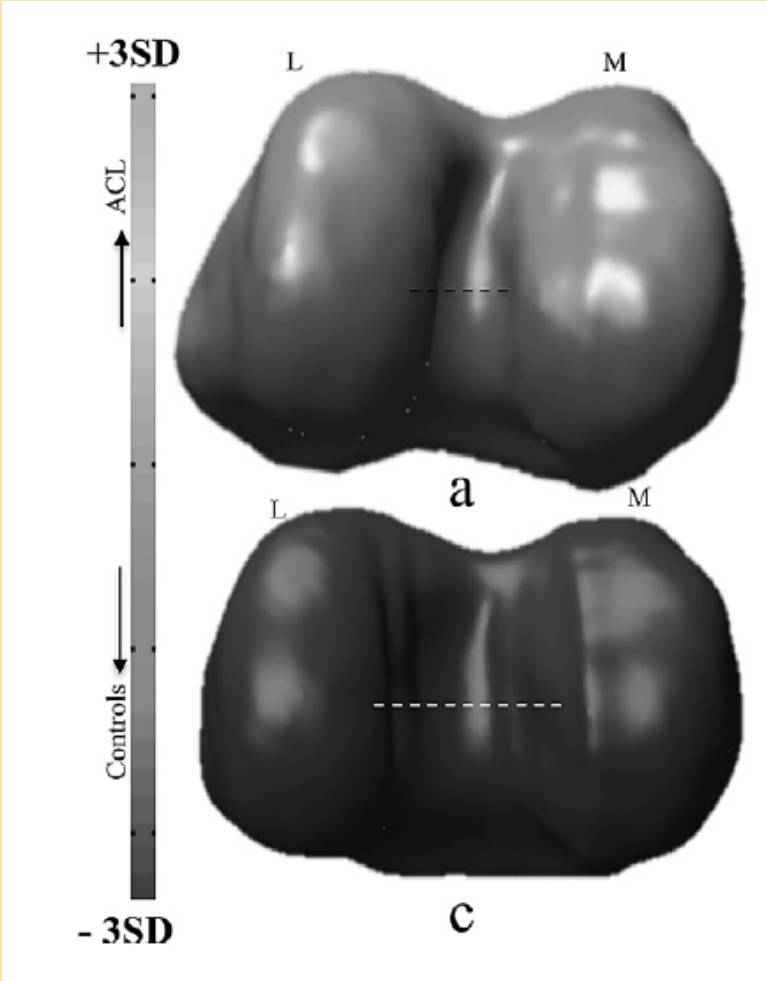
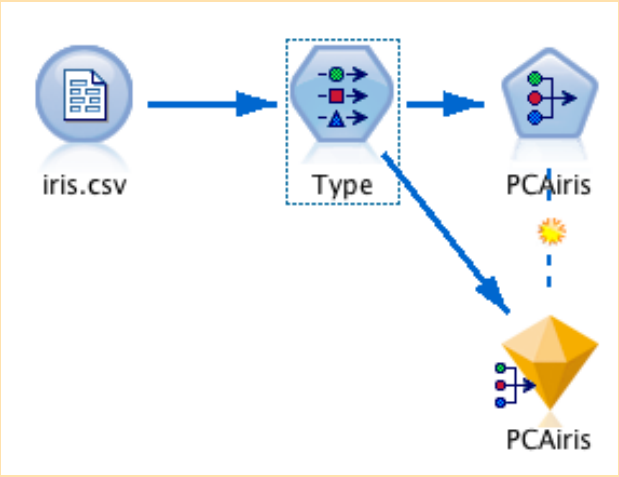
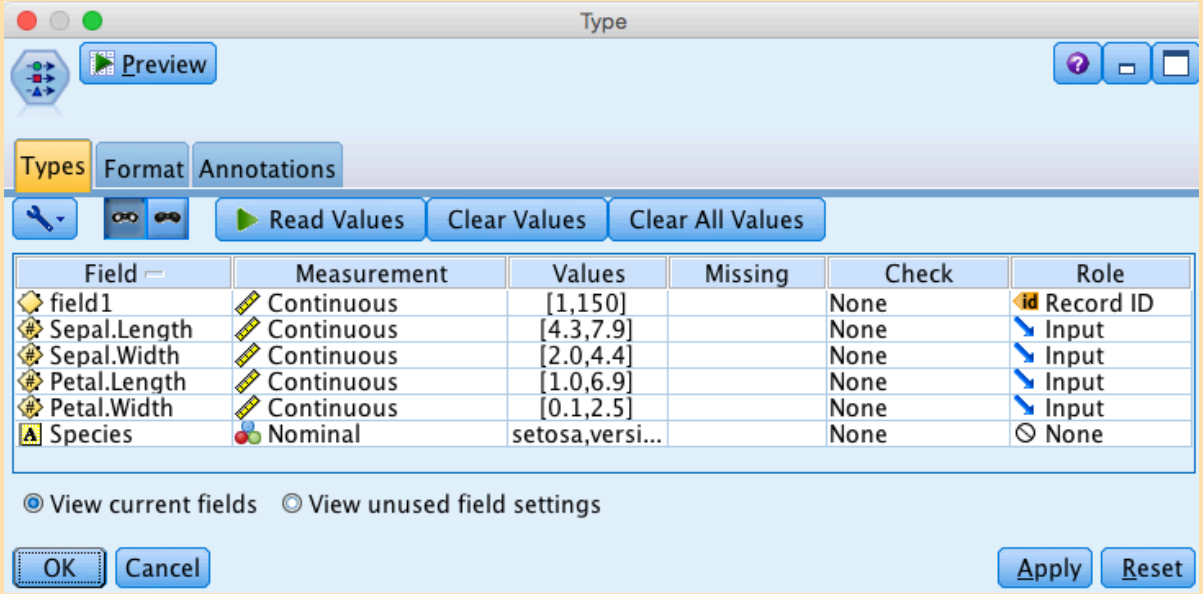


Fig. 5. Modeling of the principal component F2 (a, c) and T3 (b, d) of the scale-preserved model. In the first row Mean +3STD. in the second row Mean −3STD. Directions are labeled in the figure. M: medial, L: lateral, A: anterior P: posterior.

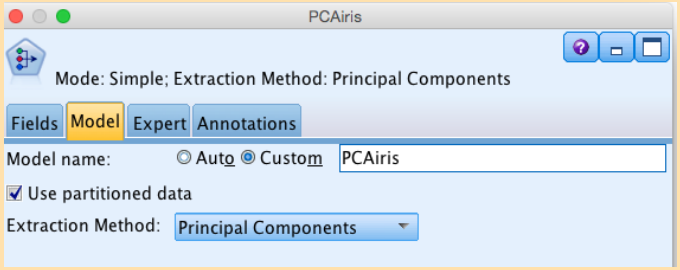
PCA in Modeler

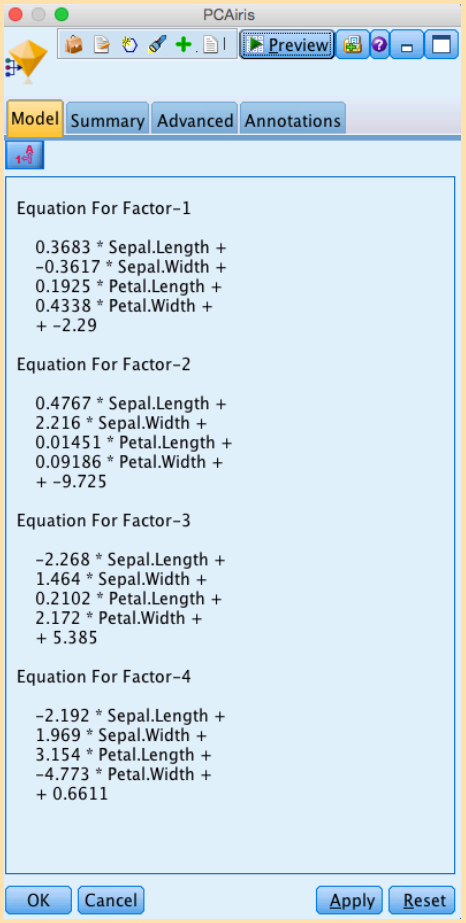
- Revisit iris data
- Just consider 4 input variables – Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width

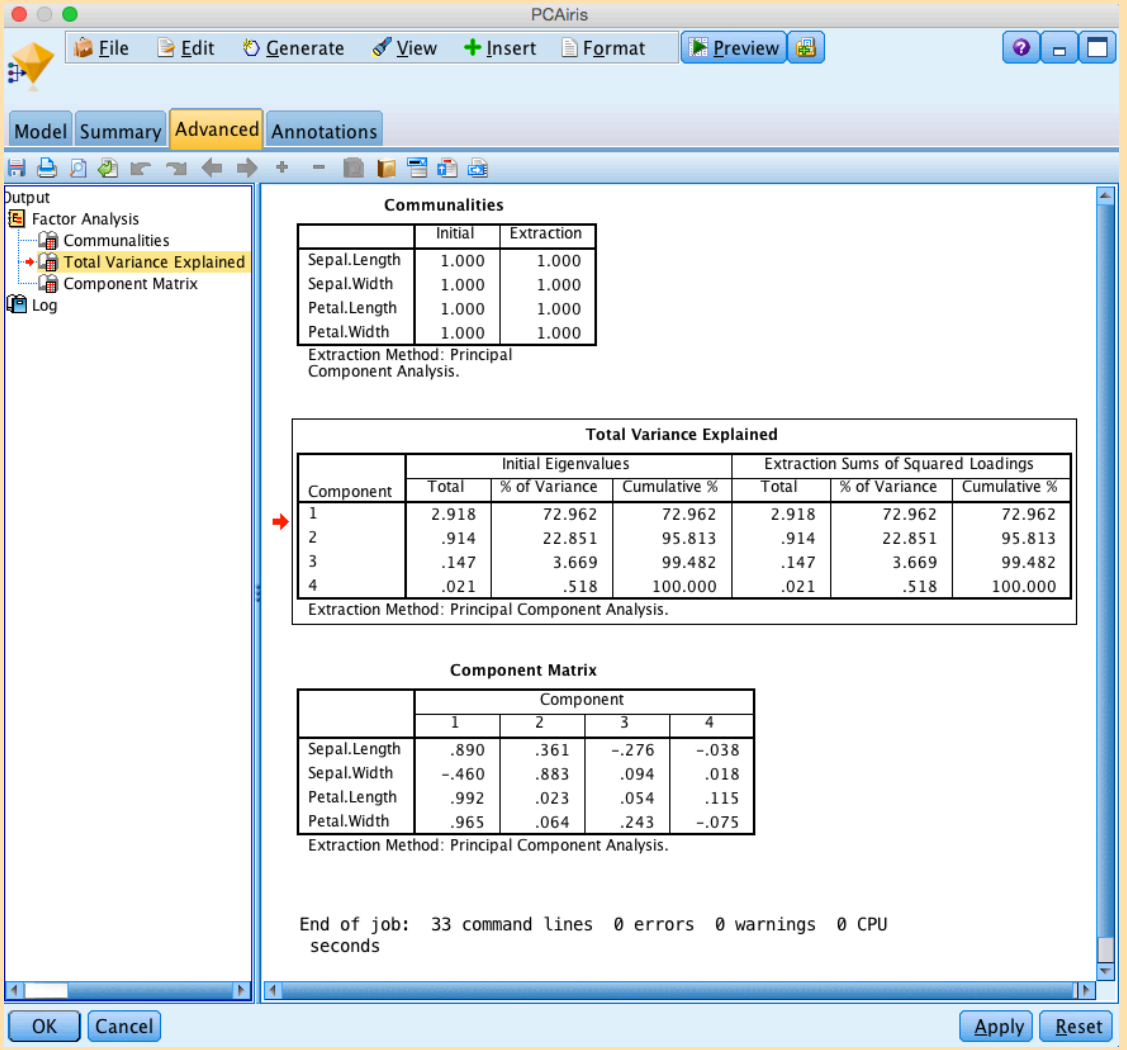




PCA in Modeler

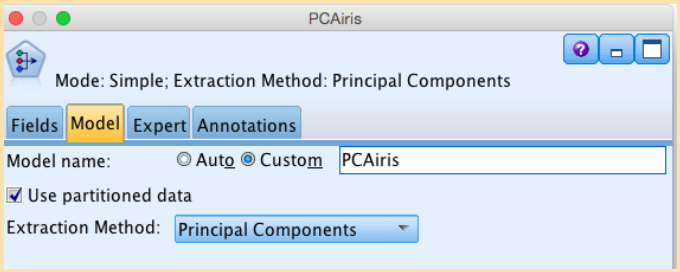


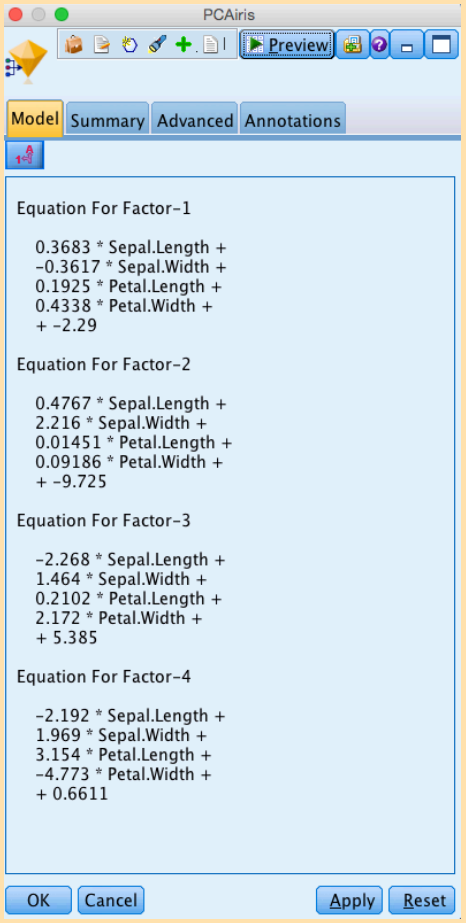


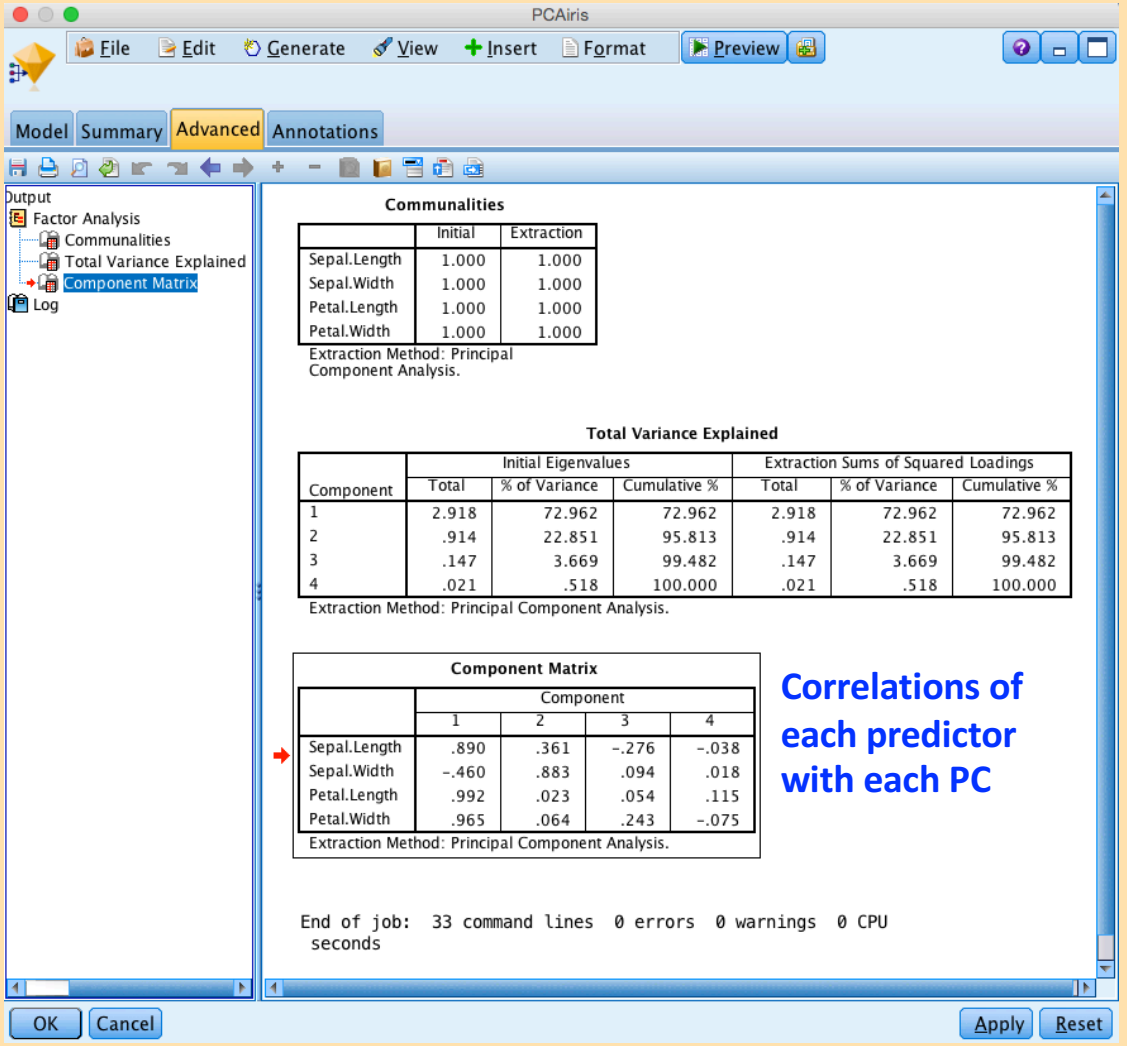


Use first 2 or 3 PC's to represent full data

PCA in Modeler







Correlations of each predictor with each PC

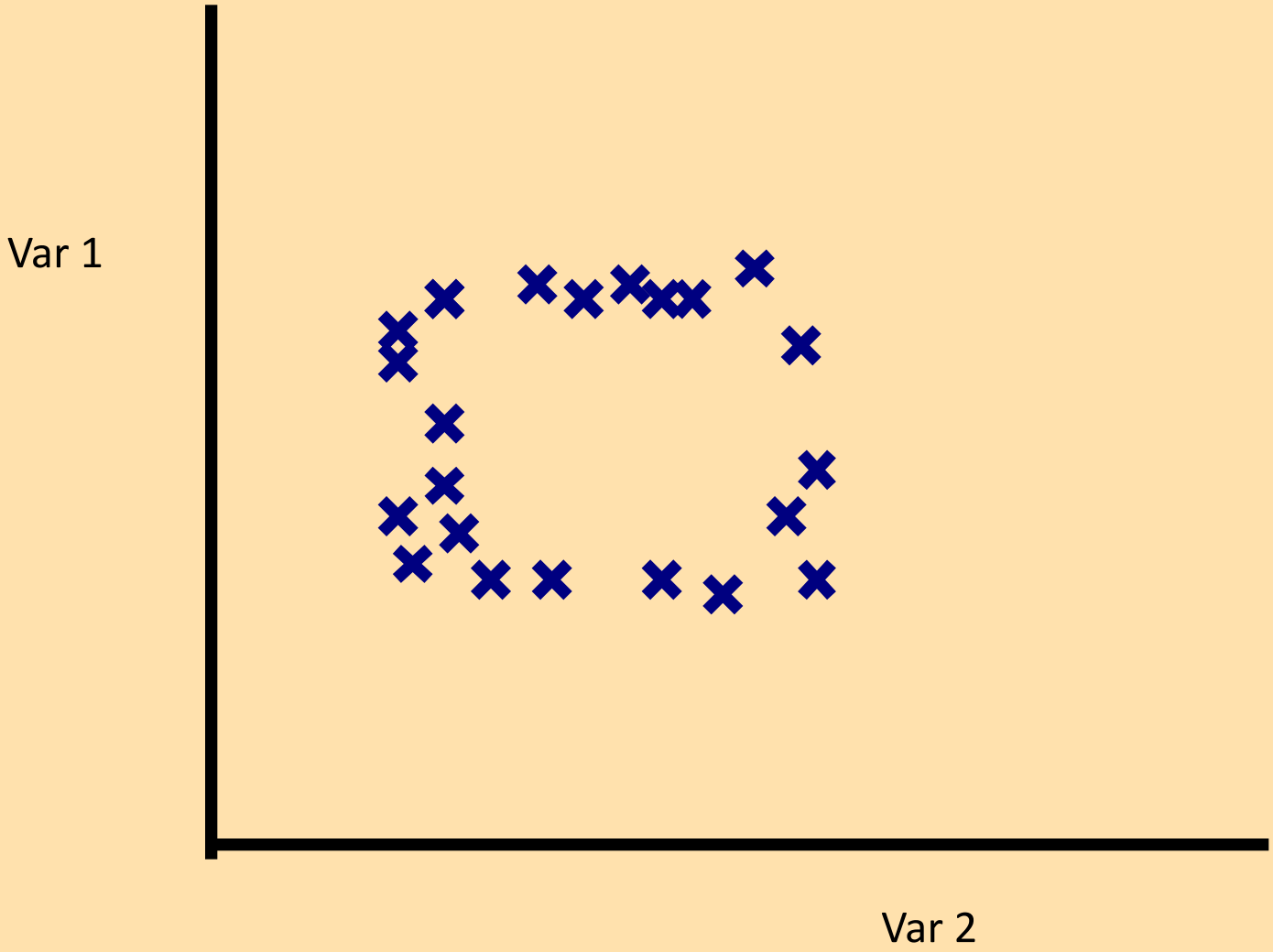
Use first 2 or 3 PC's to represent full data

Independent components analysis (ICA)

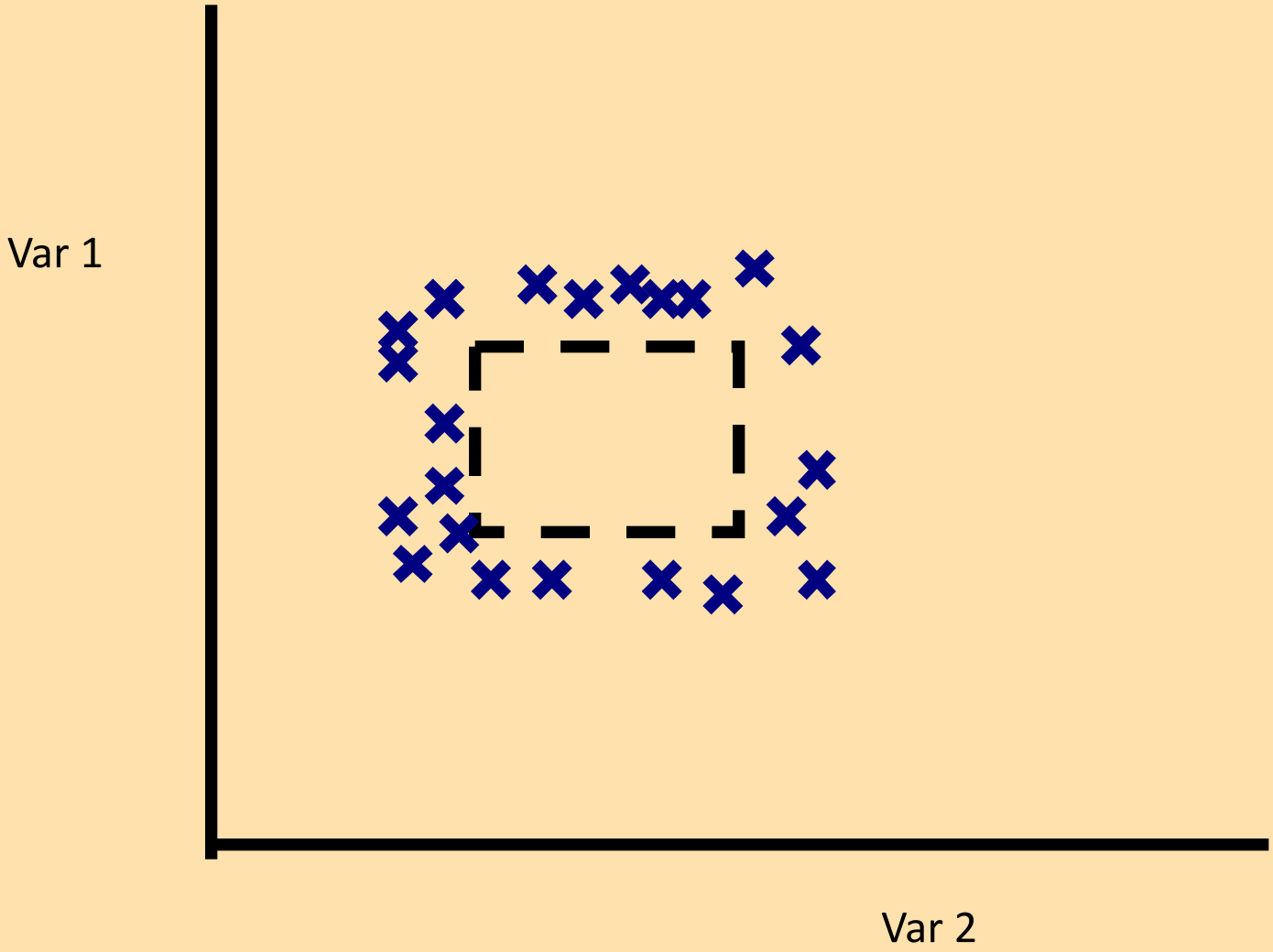
Independent Components Analysis (ICA)

- ICA differs from PCA by focusing on maximizing independence between components rather than maximizing variance within components (for non-normal signals)
- In PCA the PCs have to be uncorrelated – however, uncorrelated does not imply independent – independence is a stronger requirement

Uncorrelated but not independent

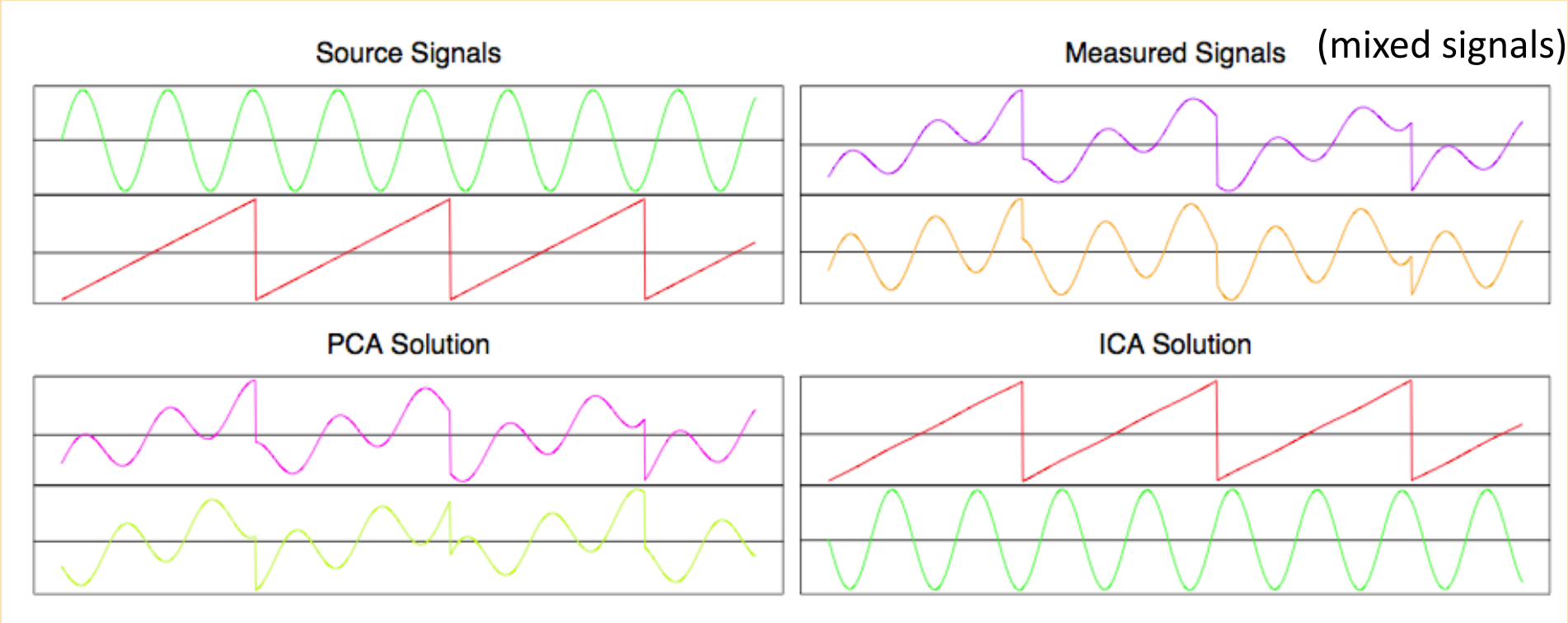


Uncorrelated but not independent



ICA illustration

Figure below from Elements of Statistical Learning (p. 561) – shows potential power of ICA in separating two mixed signals (example of the classical *cocktail party problem*). Different microphones pick up mixtures of different independent sources (music, speech from different speakers, etc.). ICA is able to perform *blind source separation*, by exploiting the independence and non-normality of the original sources.



- ICA demands that the components are (maximally) independent of each other
- It provides a different *decomposition* of the full data than does PCA (i.e. it gives a different set of weighted averages of the predictors – components)
- ICA tends to do well at signal separation (e.g. separating overlaid recorded speech signals or different components of brain activation in fMRI)
- ICA does not have an inherent ordering of components (unlike PCA that goes from largest to smallest variance explained)
- Computationally expensive and not guaranteed to find global optimum (cf. K-means clustering)
- *Not available in Modeler*, but is increasingly used for summarizing high dimensional biomedical data – can use "ica" or "fastICA" in R, or FastICA in Python, or PCA and ICA package in Matlab

Back to Classification

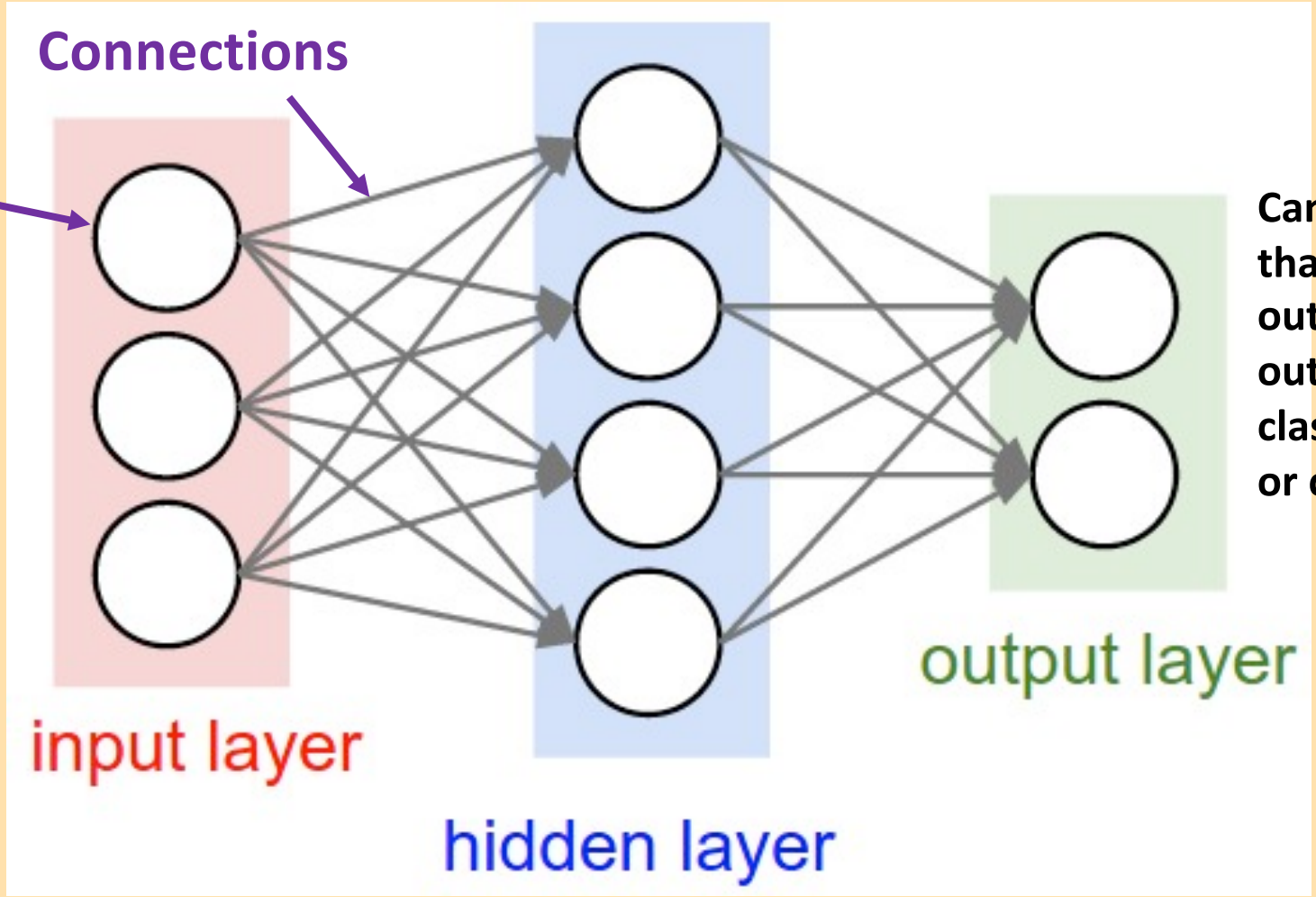
Choosing a classifier

- **Logistic Regression** (also holds for **Linear Discriminant Analysis**)
  - Nice probabilistic interpretation
  - Not highly flexible
- **Kernel-bases classification**
  - No “model” required – don’t have to worry about outliers
  - Easy to understand rules
  - Not so good for interpretation
- **Decision Trees**
  - Easy to interpret and explain
  - No “model” required -- don’t have to worry about outliers
  - Restricted to recursive partitioning of data to form classes
- **Support Vector Machines**
  - Often have high accuracy
  - Can deal with highly non-linear separation of features
  - Requires effort tuning and difficult to interpret
- **However, if your primary goal is good prediction, then you can run multiple classifiers and combine them to get an even better one overall (*ensemble* techniques of *boosting*, *bagging* and *stacking*)**

Neural networks (neural nets)

- More properly called an ‘artificial’ neural network
- “A computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.”
- A *learning rule* is developed to “learn by example” so that the weights of connections can be modified based on the observed data
- At one time considered as a model for the brain…

Neural networks (neural nets)



Can have more than one output and outputs can be class variables or continuous

Underlying math is complex

Neural Networks

- Complex models with many tuning parameters – many flavors – number of layers and nodes
- Fitting neural networks is an iterative process
- Stopping the iterations early gives a model that has reduced chance of being overfit (simpler network)
- A very flexible way to capture non-linear classification, but needs experience to use effectively because of difficulty in optimally tuning parameters
- Computationally intensive – especially for large networks with a large amount of data
- No interpretation – “Black Box classifier”
- Fitted in Modeler using the Neural Net node – we will explore this node in lab Thursday

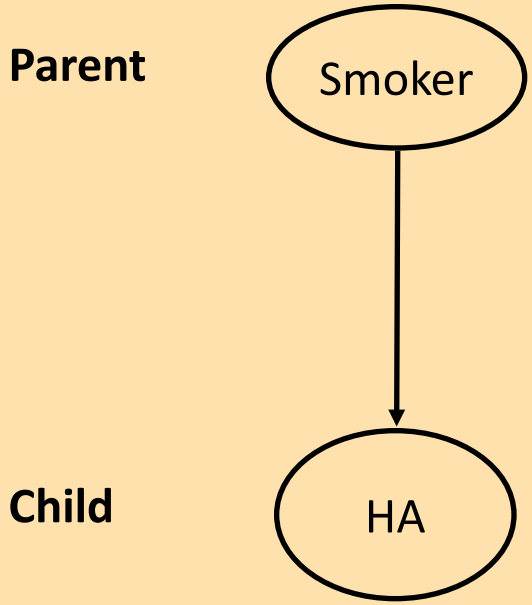
Bayesian Networks (Bayes Nets)

Bayesian Networks

- Looks for pathways or complex relationships between variables
- Unlike most classification methods, Bayes Nets actually has a “causal” interpretation
- Bayesian networks do not have pre-defined outcomes/predictors, but determine order/structure among a set of variables

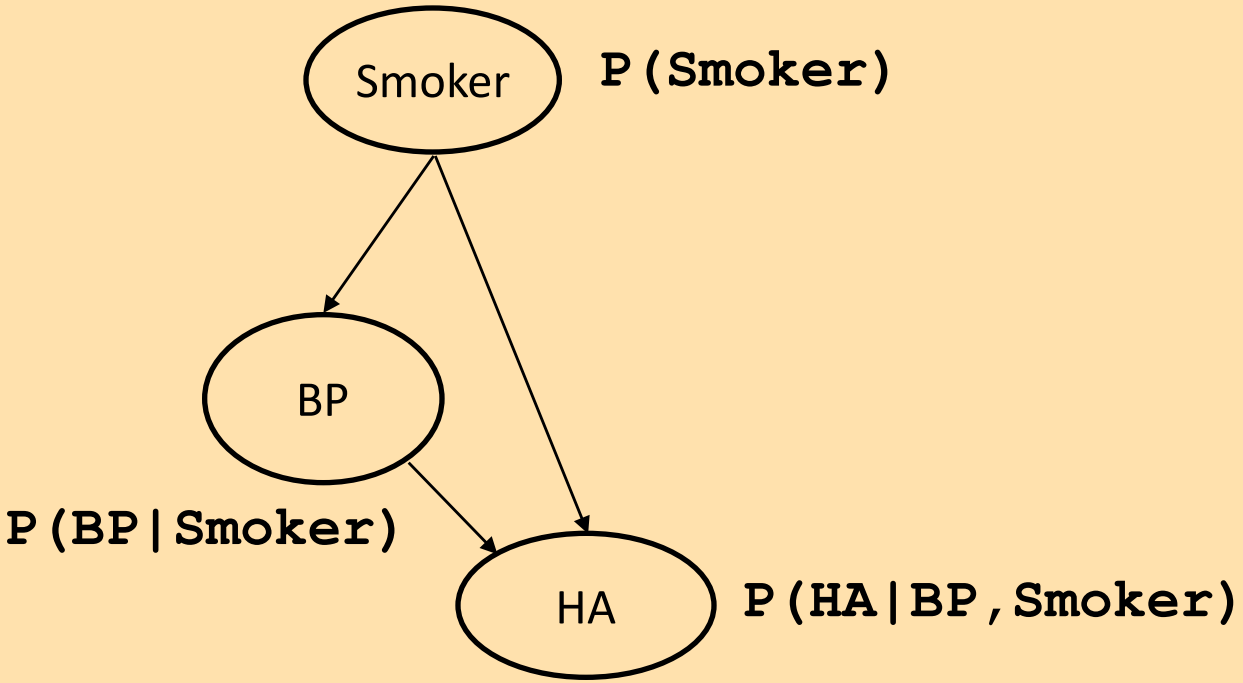
Components of Bayes Nets

- Each node is a variable with an associated probability distribution
- Each arc (arrow) denotes *conditional dependence*

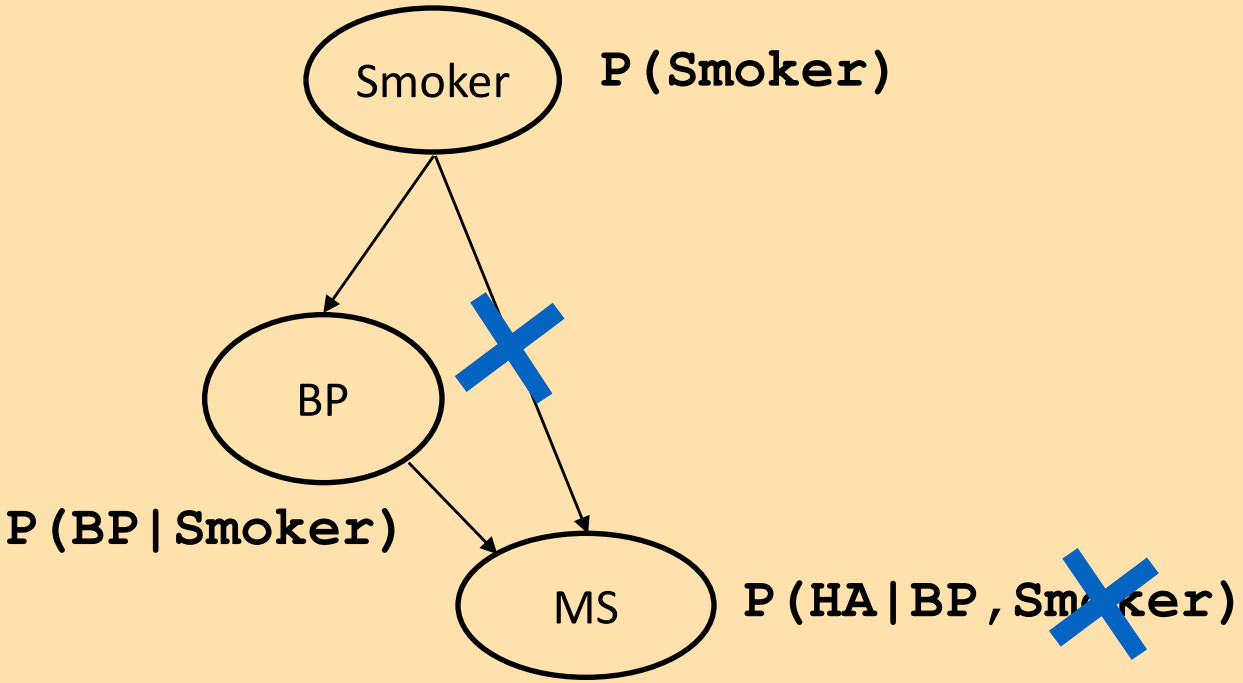


HA = event of Heart Attack

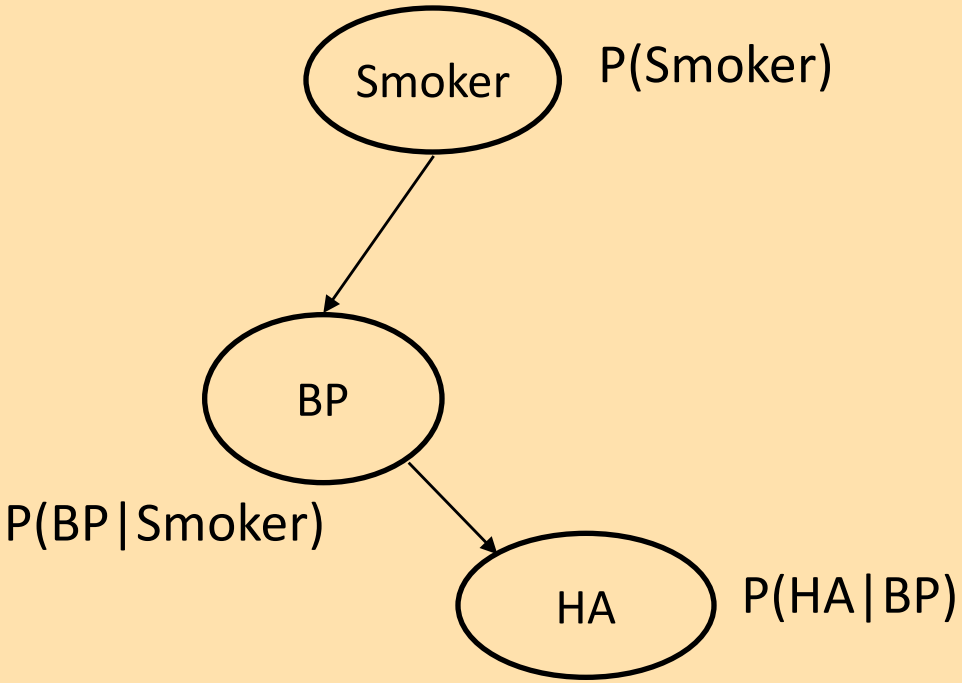
Small illustration



Small illustration



Conditional independence



Have determined risk of heart attack (HA) is *conditionally independent* of Smoker given BP

Smoker, BP, and HA are all related to each other, but once the value of BP is settled, Smoker and HA become independent of each other

Reduced model is simpler to estimate and interpret

Learning Bayes Nets

- Compute optimal network among a set of nodes
- Requires a “score” of what is “good”
- Full evaluation = “NP-hard” problem (i.e. computationally impossible for large problems)
- Employ greedy type algorithms
- Perhaps have multiple starts

Example Multiple Sclerosis – variable set

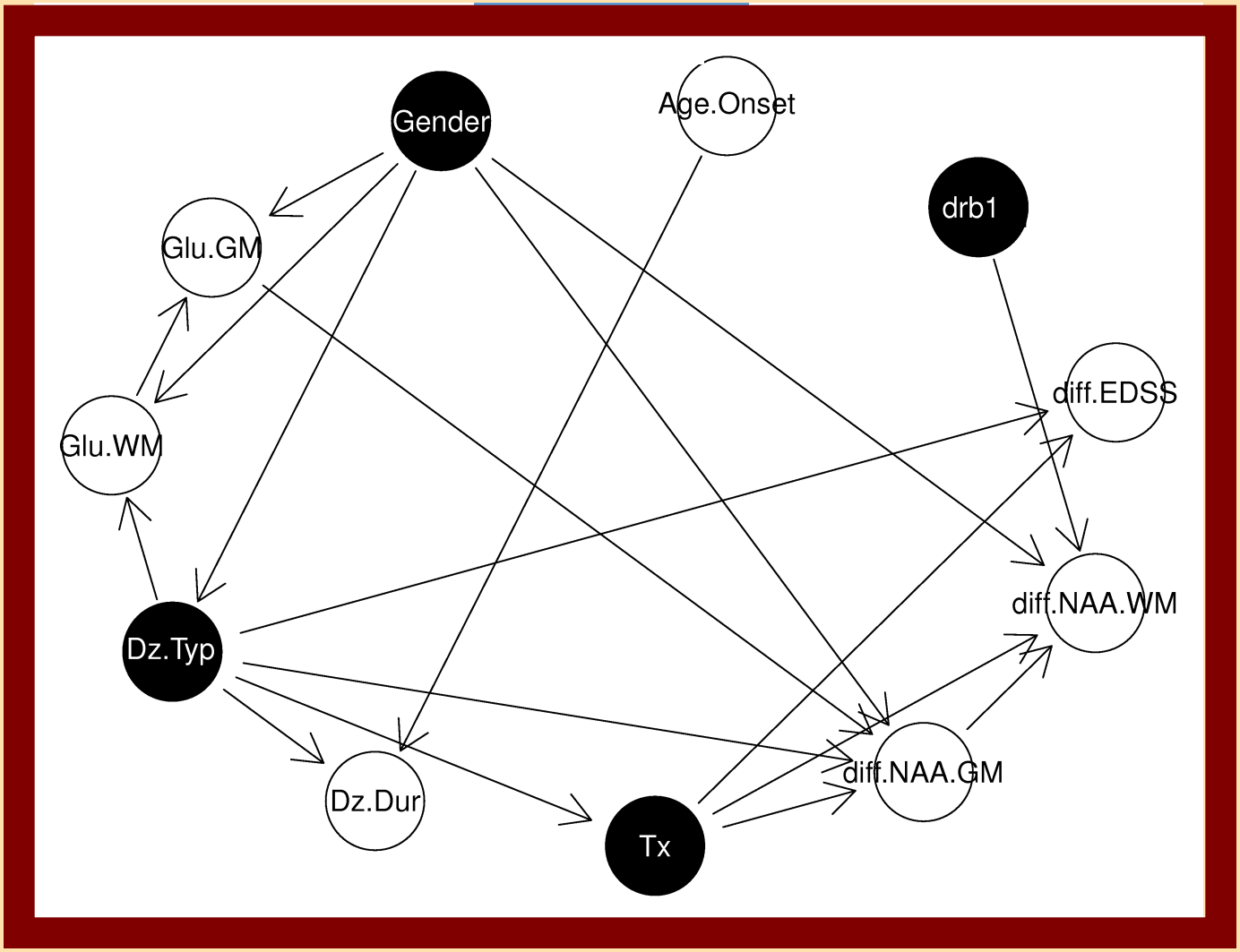
Categorical

- Treatment Y/N (Tx)
- Gender (Gender)
- Disease Type (Dz.Typ)
- HLA*DRB1**1501 (drb1)

Continuous

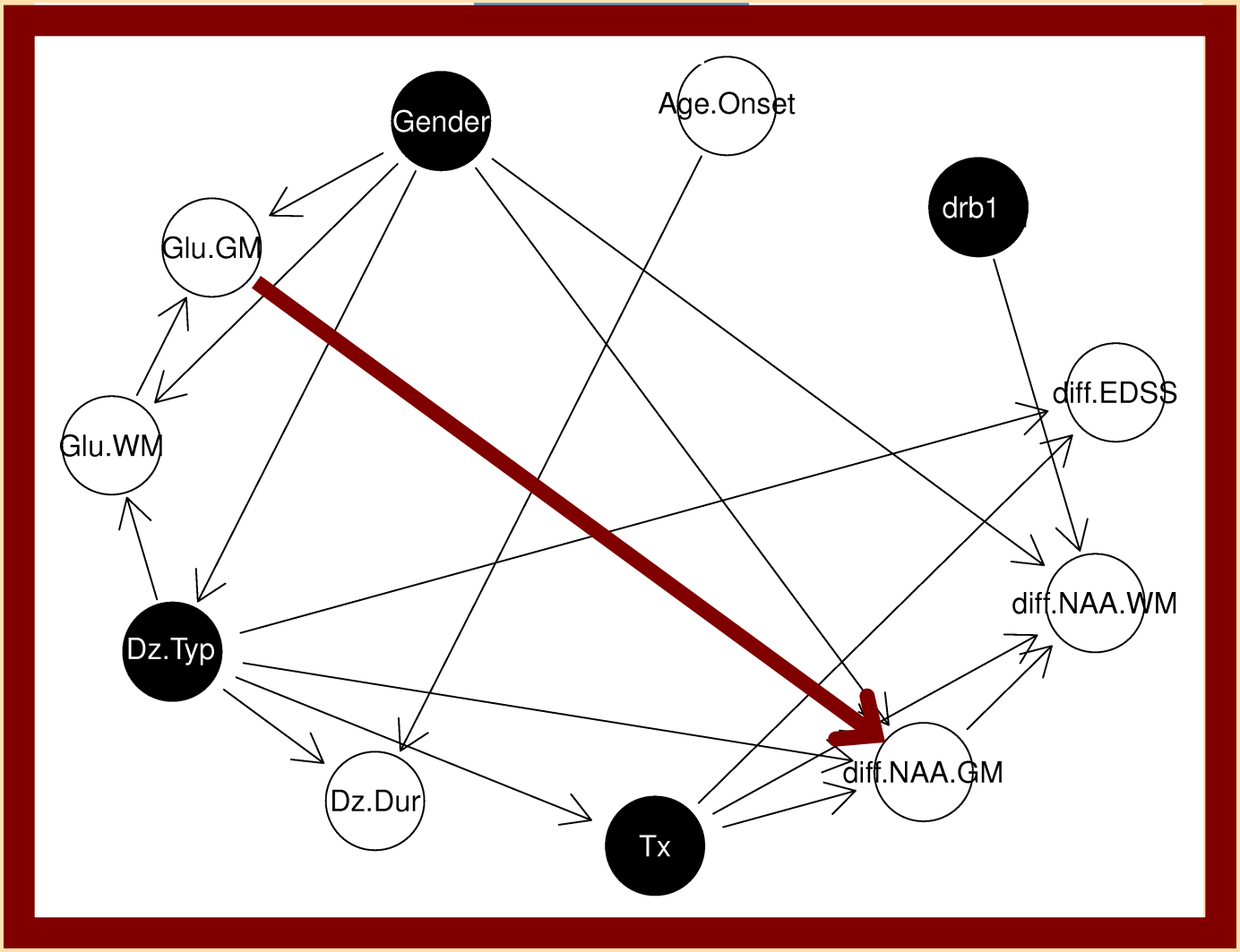
- Disease Duration (Dz.Dur)
- Age at Disease Onset (Age.Onset)
- Baseline Glu GM (Glu.GM)
- Baseline Glu WM (Glu.WM)
- 1 year change in NAA GM (diff.NAA.GM)
- 1 year change in NAA WM (diff.NAA.WM)
- 1 year change in EDSS (diff.EDSS)

Fitted Bayes' Net

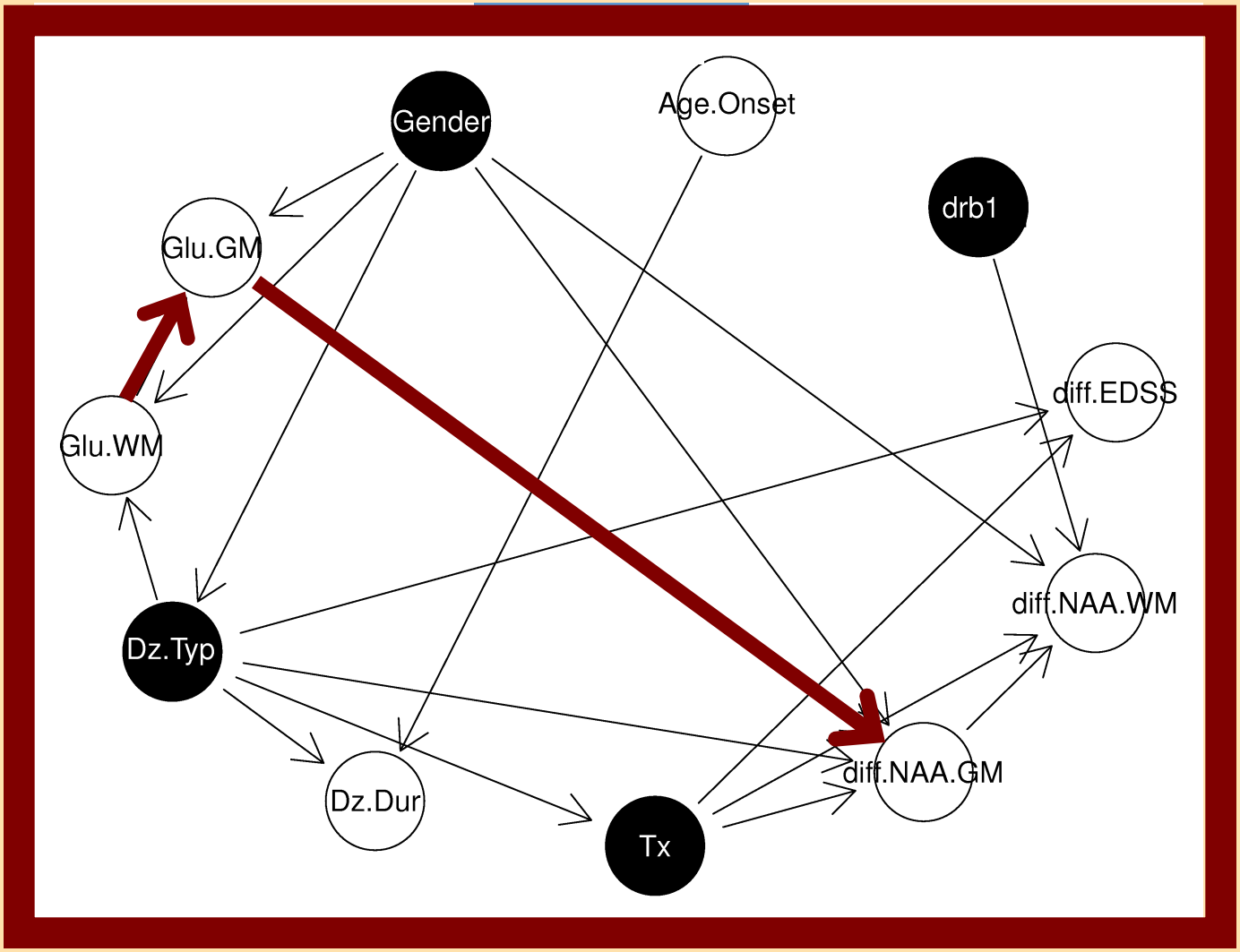


Arrows can be interpreted as defining causal effects – see all the caveats in Dr. McCulloch's lecture on Thursday when it comes to causal modeling

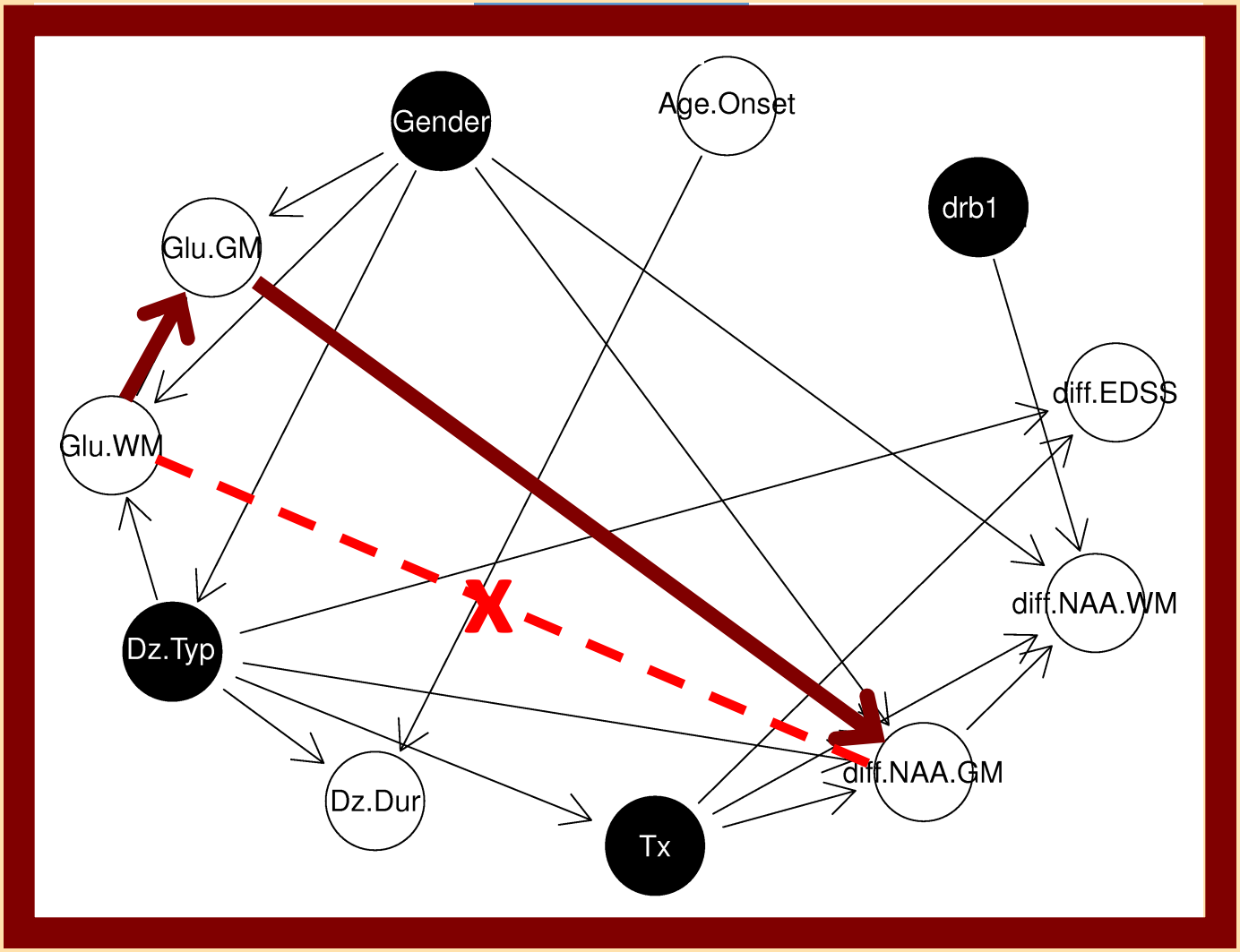
Glu Effect on NAA GM



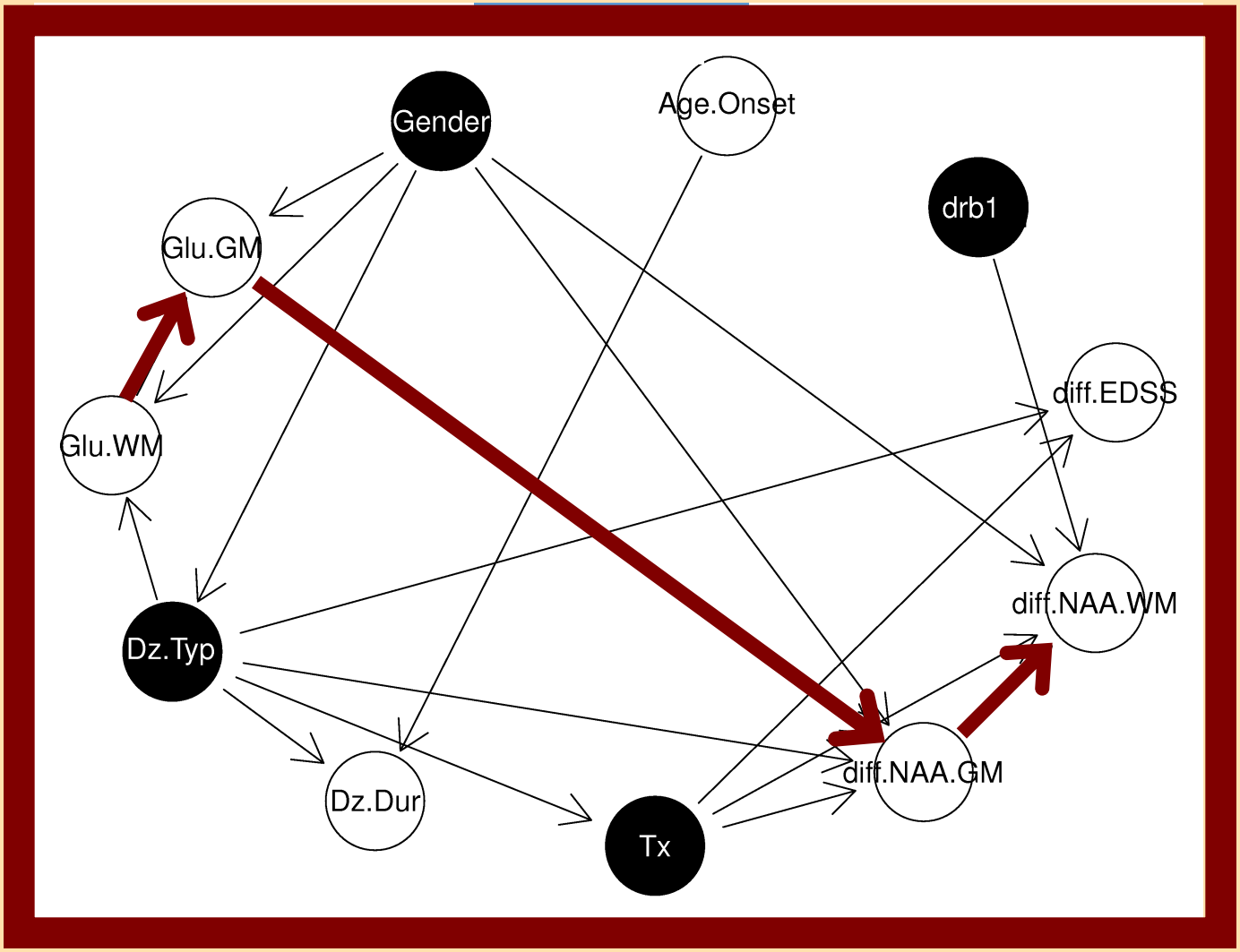
Glu Effect on NAA GM



Glu Effect on NAA GM

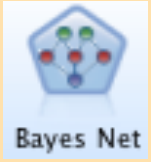


Glu Effect on NAA GM



Bayesian networks

- Can lead to complex insights into the relationships between variables
- Have to be careful to not over-interpret estimated "causal" relationships – see Dr. McCulloch's lecture Thursday
- Computation can be expensive
- We will explore Bayes Nets in Modeler during lab:



Machine Learning Computational Tools

- We have focused on IBM Modeler but there are other tools/software/programming languages if you want to dig further. Here are some free ones
- Weka -- http://www.cs.waikato.ac.nz/ml/weka/ - there is an associated book - Data Mining: Practical Machine Learning Tools and Techniques 3rd edition, Witten, Frank and Hall, 2011
- R -- https://www.r-project.org/ -- R has many machine learning libraries and tools. They have a machine learning and statistical learning "Task View" page which much of what is available in R (including an R interface to the Weka package)
- Python -- https://www.python.org/ -- if you want to go even further down the programming line – many numerical and scientific libraries – see e.g. http://www.southampton.ac.uk/~fangohr/training/python/pdfs/Python-for-Computational-Science-and-Engineering.pdf to get started

Another upcoming seminar of interest

Wednesday, September 7th
9:00-10:30
Mission Bay—Genentech Hall, Byers Auditorium
Simulcast at Parnassus—Health Sciences West, 303

Towards Computer-Enabled Precision Radiology and Neurology

Ashish Raj, PhD
Associate Professor of Computer Science in Radiology, Associate Professor of Neuroscience
Co-Director, Image Data Evaluation and Analytics Lab (IDEAL), Weill Cornell Medical College

Next Lecture - Thursday
Dr. McCulloch
“Causal Inference from Big Data”