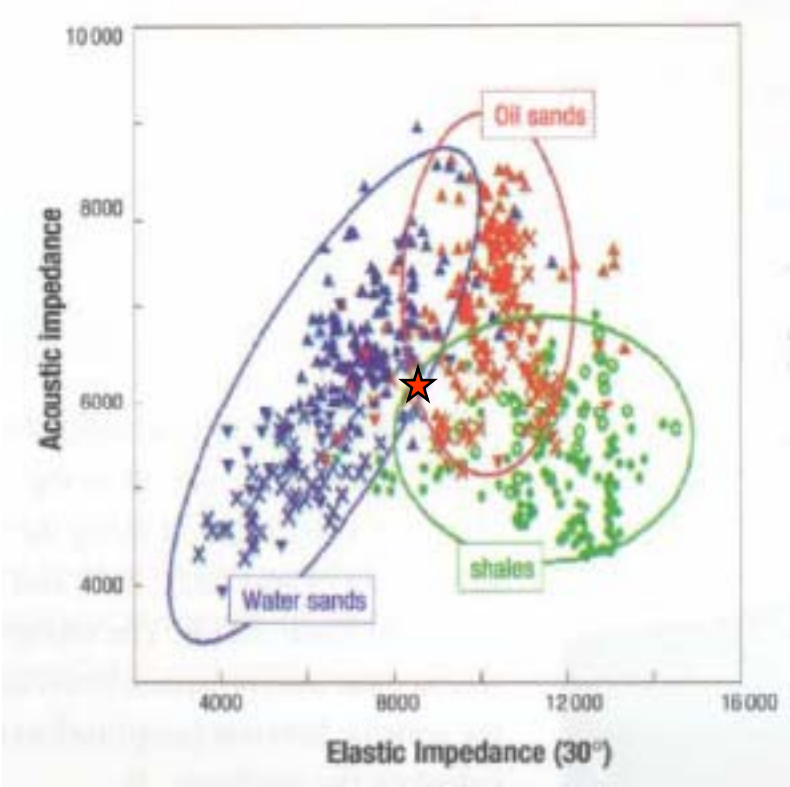
Statistical Classification and Pattern Recognition

Qiuliang Yao

April 3, 2009

Problem



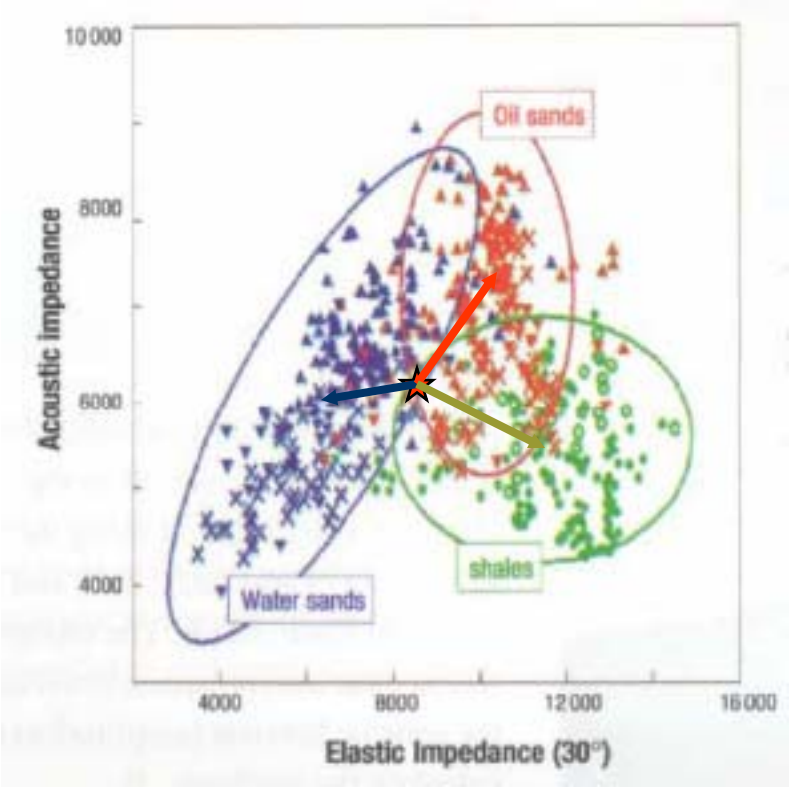
What is ★ ?

- **Oil sand?**
- **Water sand?**
- **Shale?**

- Discriminant analysis
- Bayesian classification
- Neural network classification

Discriminant analysis: Mahalanobis distance

Distance to a cloud, to a single point



Point to point:

$$d = \sqrt{\sum (x_i - \mu_i)^2}$$

Point to spherical cloud:

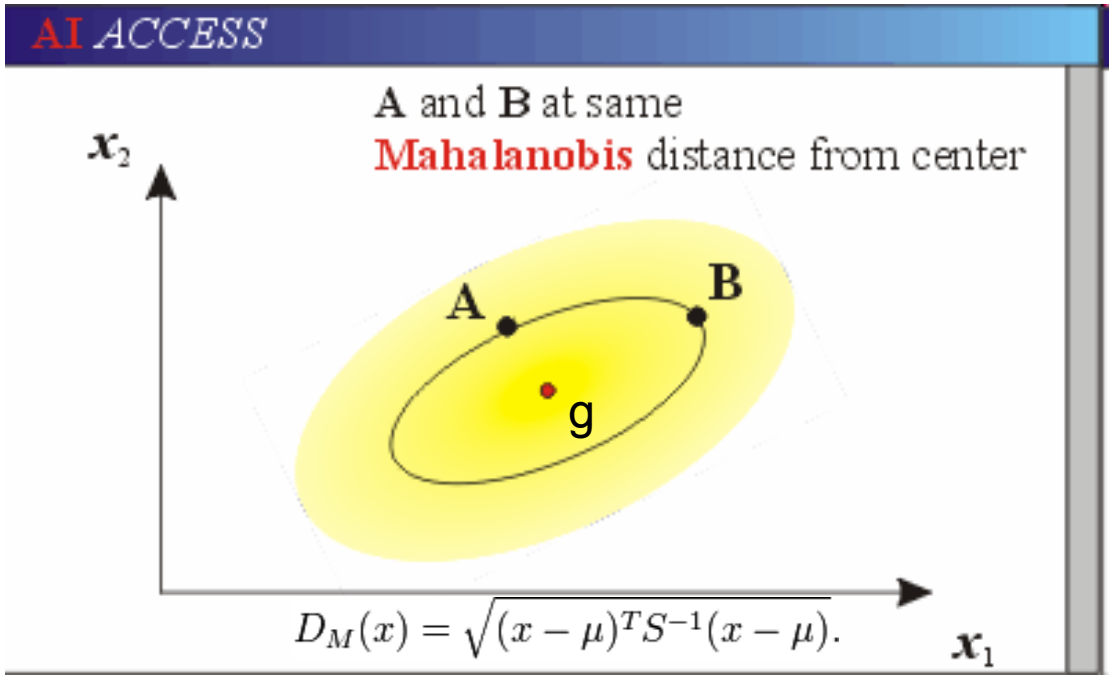
$$d = \sqrt{\sum \frac{(x_i - \mu_i)^2}{\sigma_i^2}}$$

Point to any shape cloud:

$$d_M = \sqrt{(x_i - \mu_i)^T S^{-1} (x_i - \mu_i)}$$

Discriminant analysis

Mahalanobis distance $d^2_M = (\boldsymbol{x} - \boldsymbol{g})' \; S^{-1}(\boldsymbol{x} - \boldsymbol{g})$



Euclidean distance:

Ag=!Bg

Mahalanobis distance:

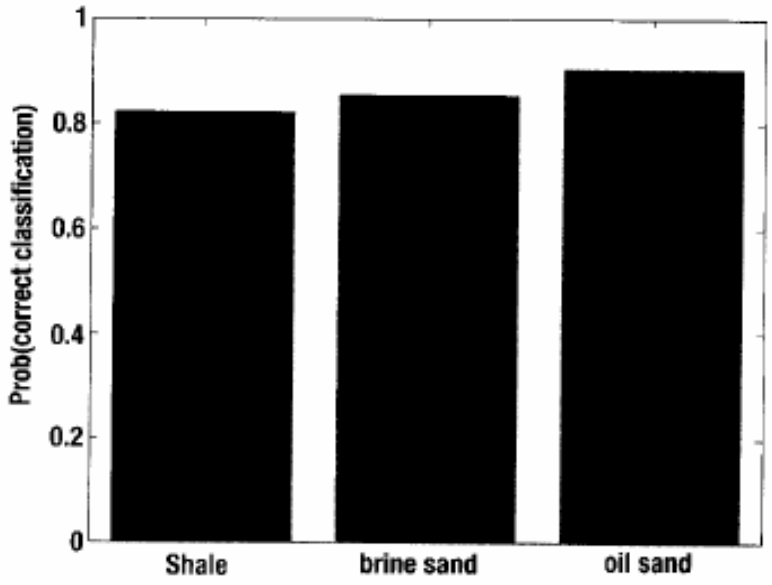
Ag=Bg

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}.$$

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2}},$$

Discriminant analysis

	Predicted facies		
	shale	brine sand	oil sand
True facies			
shale	0.82	0.063	0.024
brine sand	0.17	0.85	0.071
oil sand	0.006	0.083	0.91



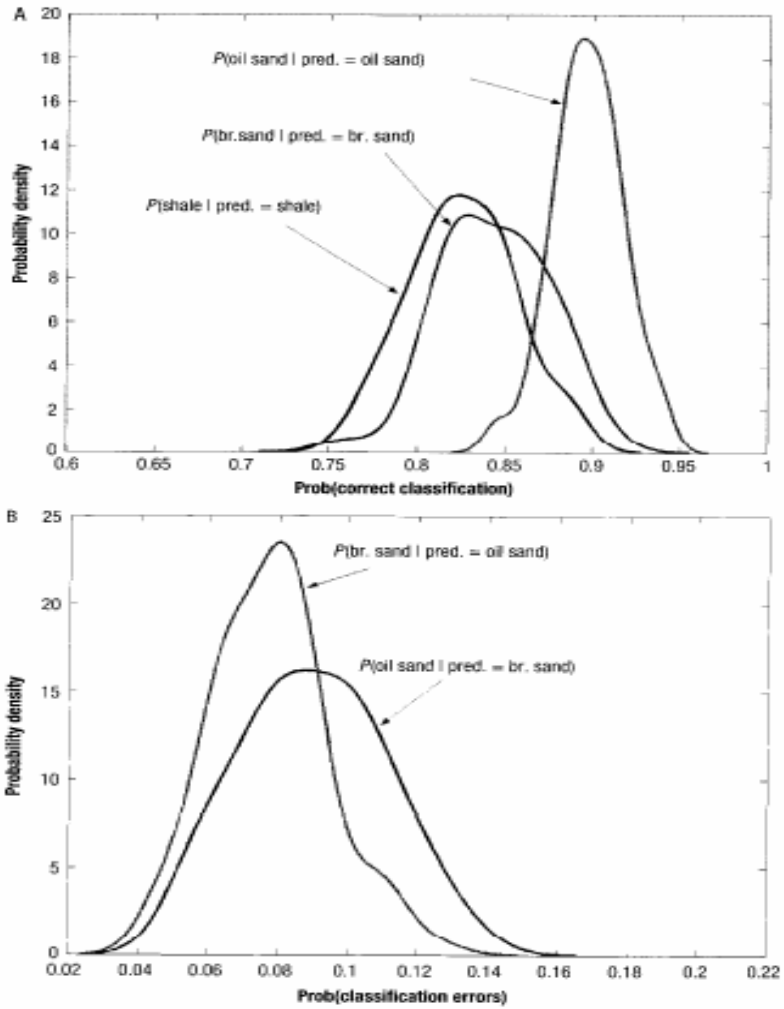


Figure 3.20 Distribution of probability of successful classification (A); different types of misclassification (B); and the risk of dry hole (C).

Principal Component Analysis (PCA)

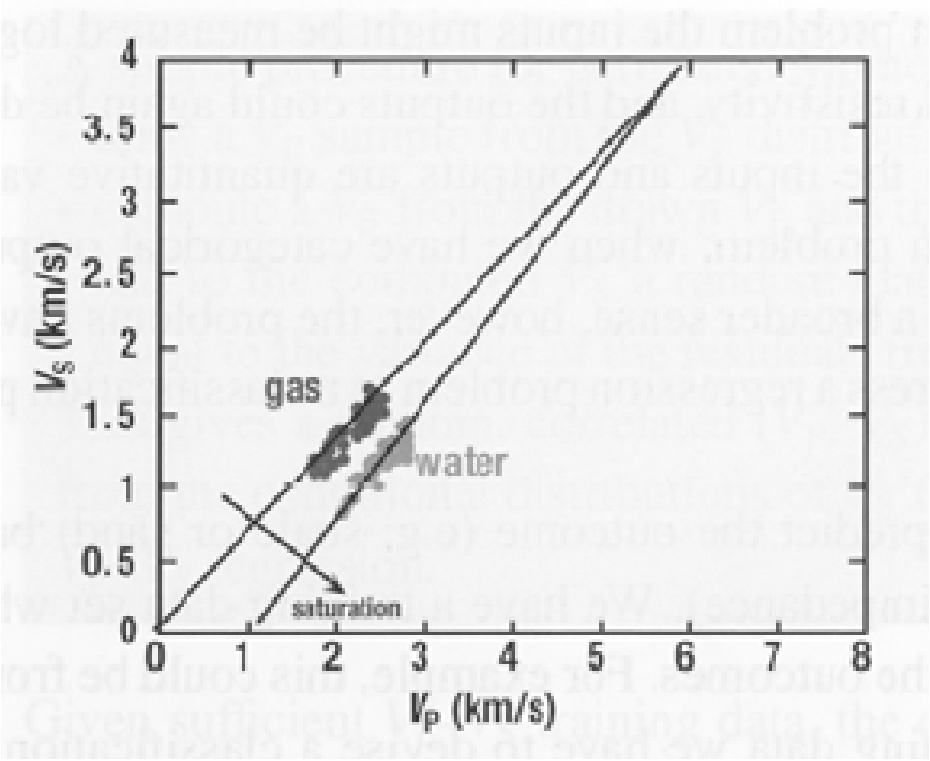
Does more data always help to get better classification?

Which attributes should we chose?

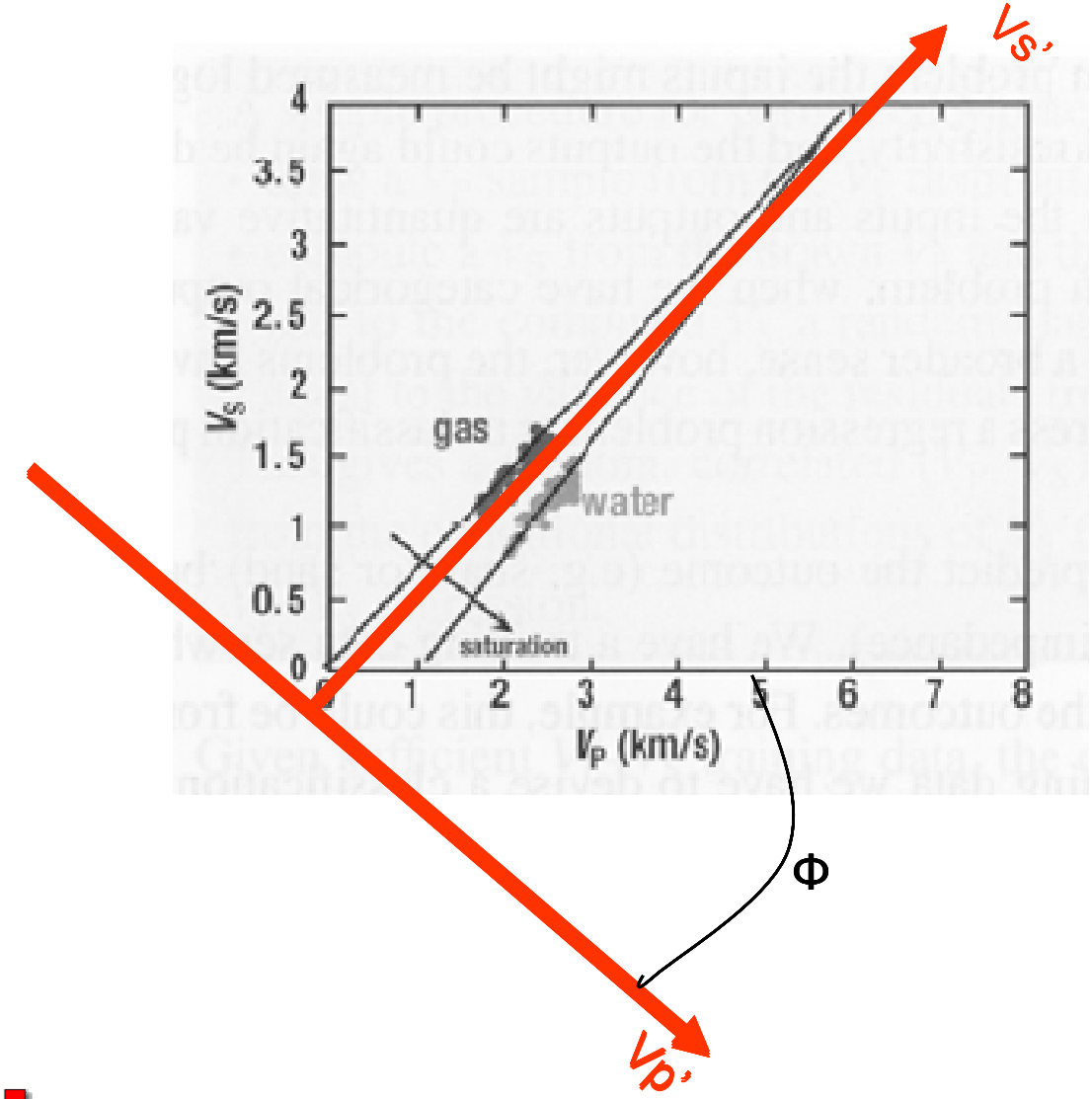
With original attributes set, it's hard to answer, because they might be correlated to each other.

By converting into to PCs, it's easy to answer, because PCs are orthogonal

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)



$$Vp' = Vp\cos\phi + Vs\sin\phi$$

$$Vs' = -Vp\sin\phi + Vs\cos\phi$$

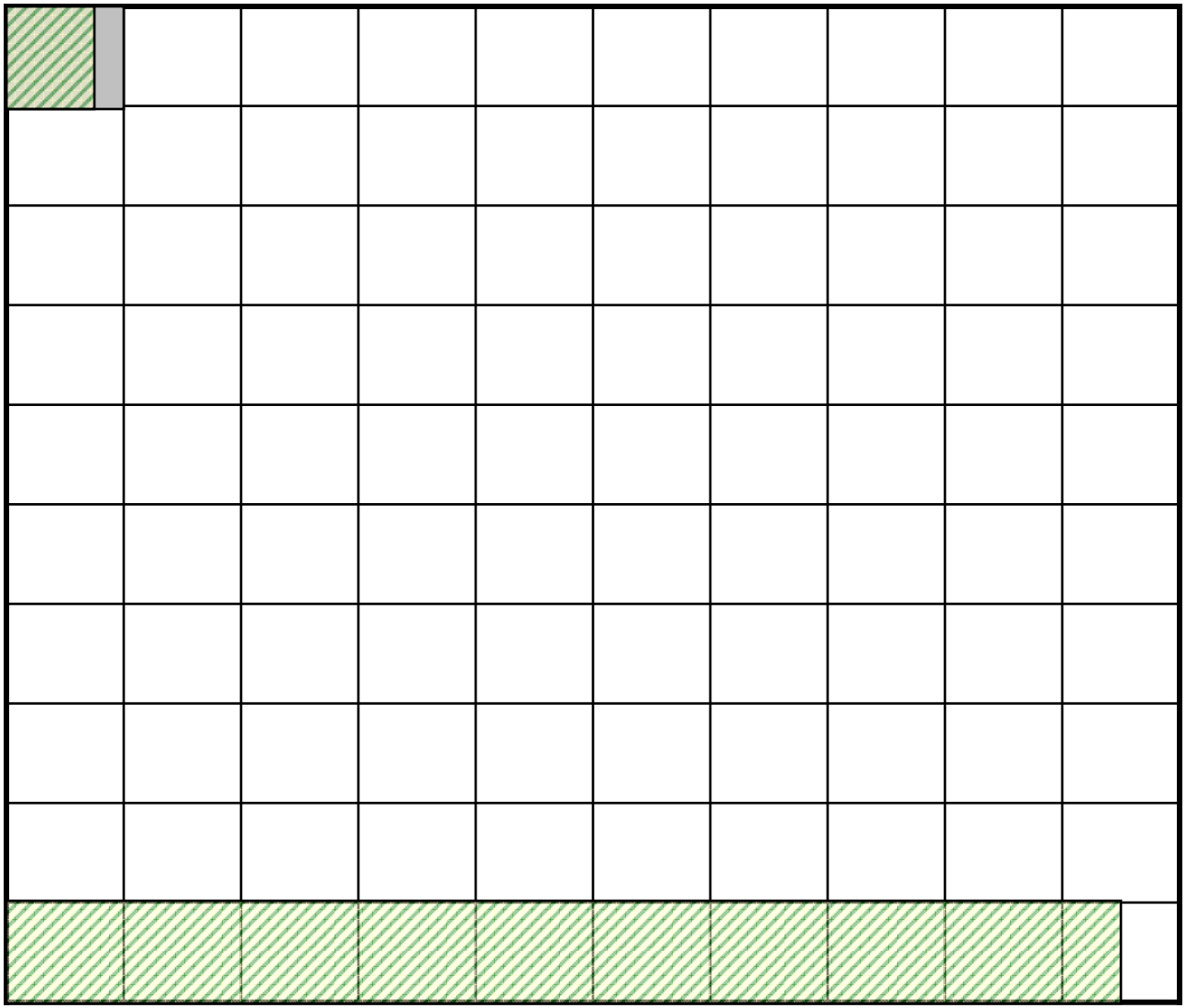
PC1: Vs', largest variation

PC2: Vp',

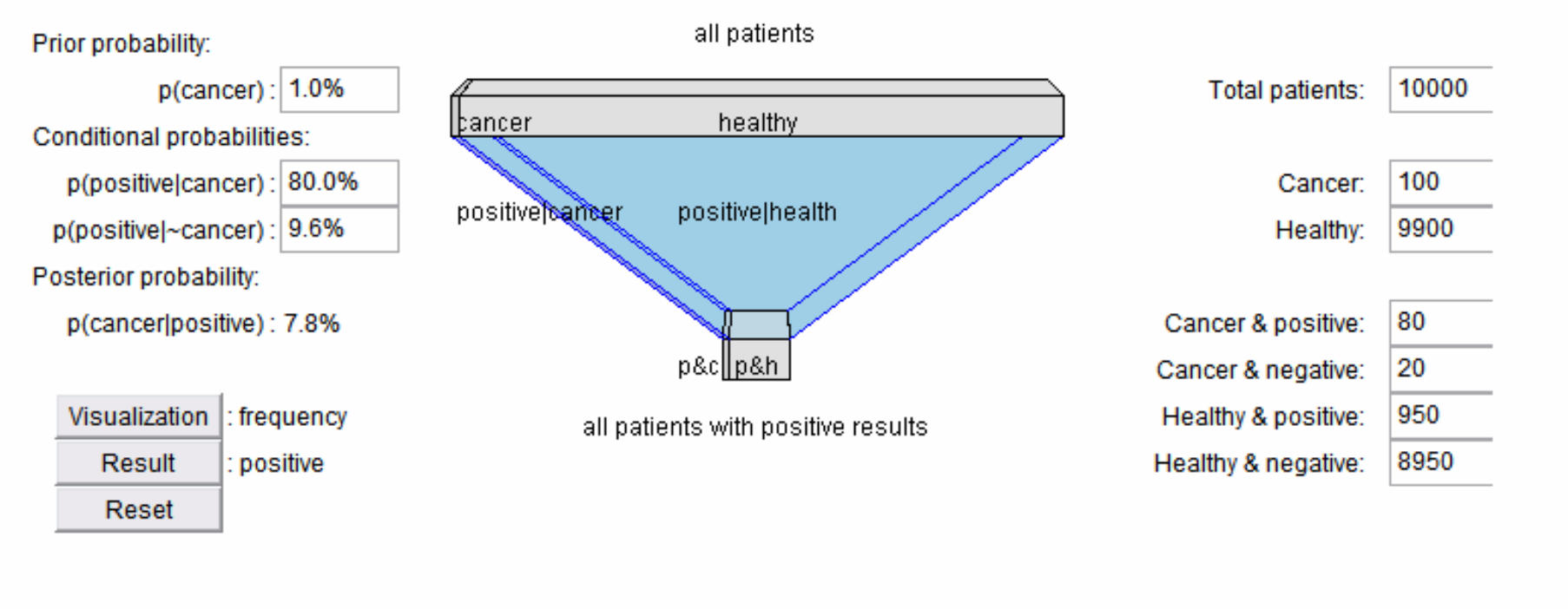
Bayesian classification

1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

http://yudkowsky.net/rational/bayes



Bayesian classification



$$p(A|X) = \frac{p(X|A)*p(A)}{p(X|A)*p(A) + p(X|\sim A)*p(\sim A)}$$

Bayesian classification

$$P(c_j \mid x) = \frac{P(x, c_j)}{P(x)} = \frac{P(x \mid c_j)P(c_j)}{P(x)}$$

Prior probability

Conditional probability

$$P(x) = \sum_{i=1}^{N} P(x \mid state_j)P(state_i)$$

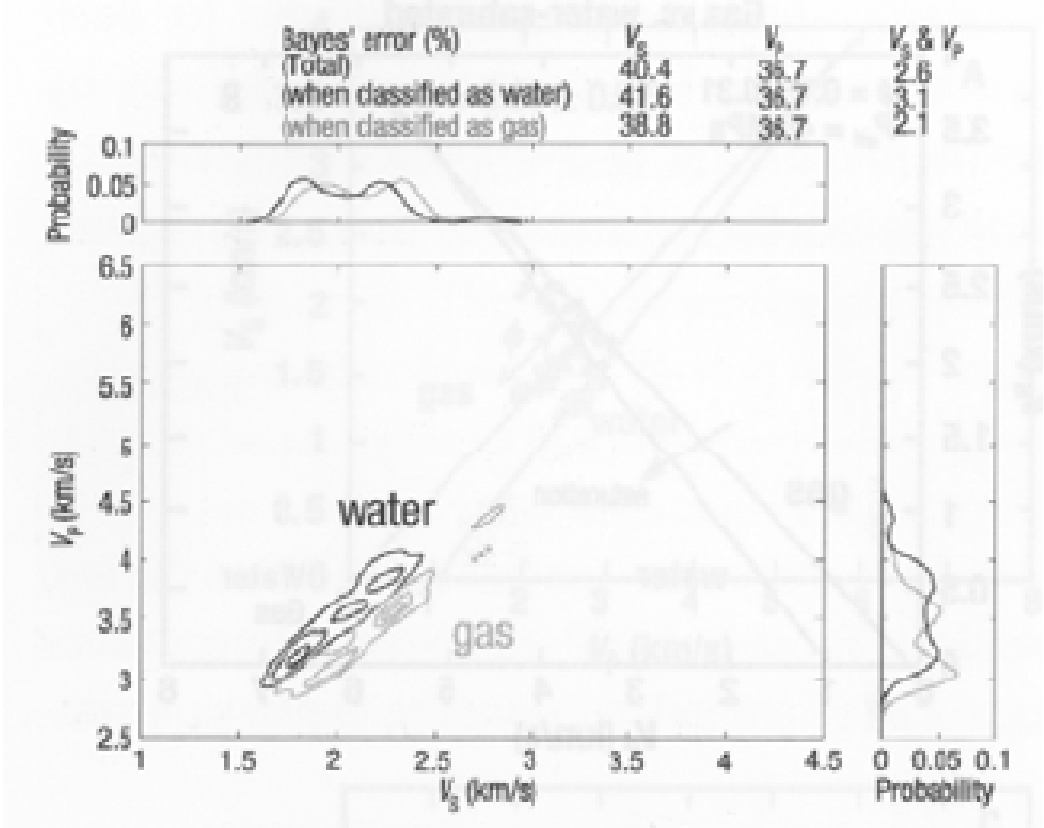
Bayes' decision rule:

Classify as class c_k if $P(c_k \mid x) \triangleright P(c_j \mid x)$ for all $j \neq k$

Principal Component Analysis (PCA)

sand if $P(\text{sand} \mid V_P, V_S) > P(\text{shale} \mid V_P, V_S)$

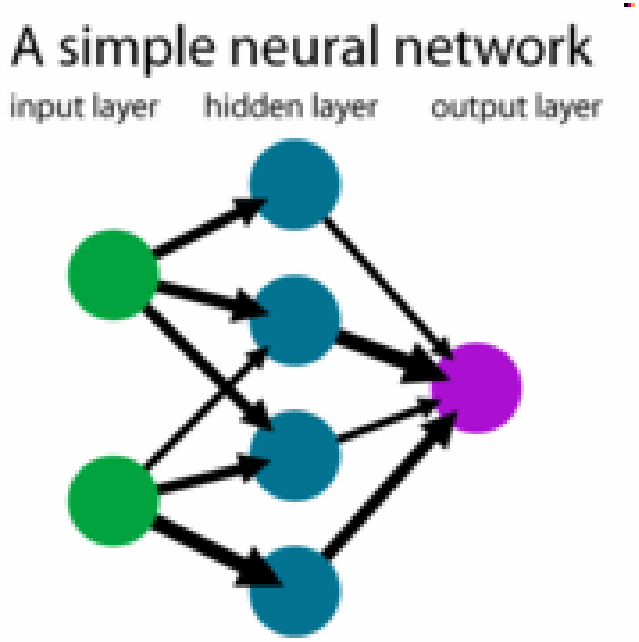
shale if $P(\text{shale} \mid V_P, V_S) > P(\text{sand} \mid V_P, V_S)$



Neural network classification

- A regression/ fitting tool
- Minimized the misfit between desired and modeled output
- Done by gradient descent (back-propagation)

Pitfall: hard to interpret



Neural network classification

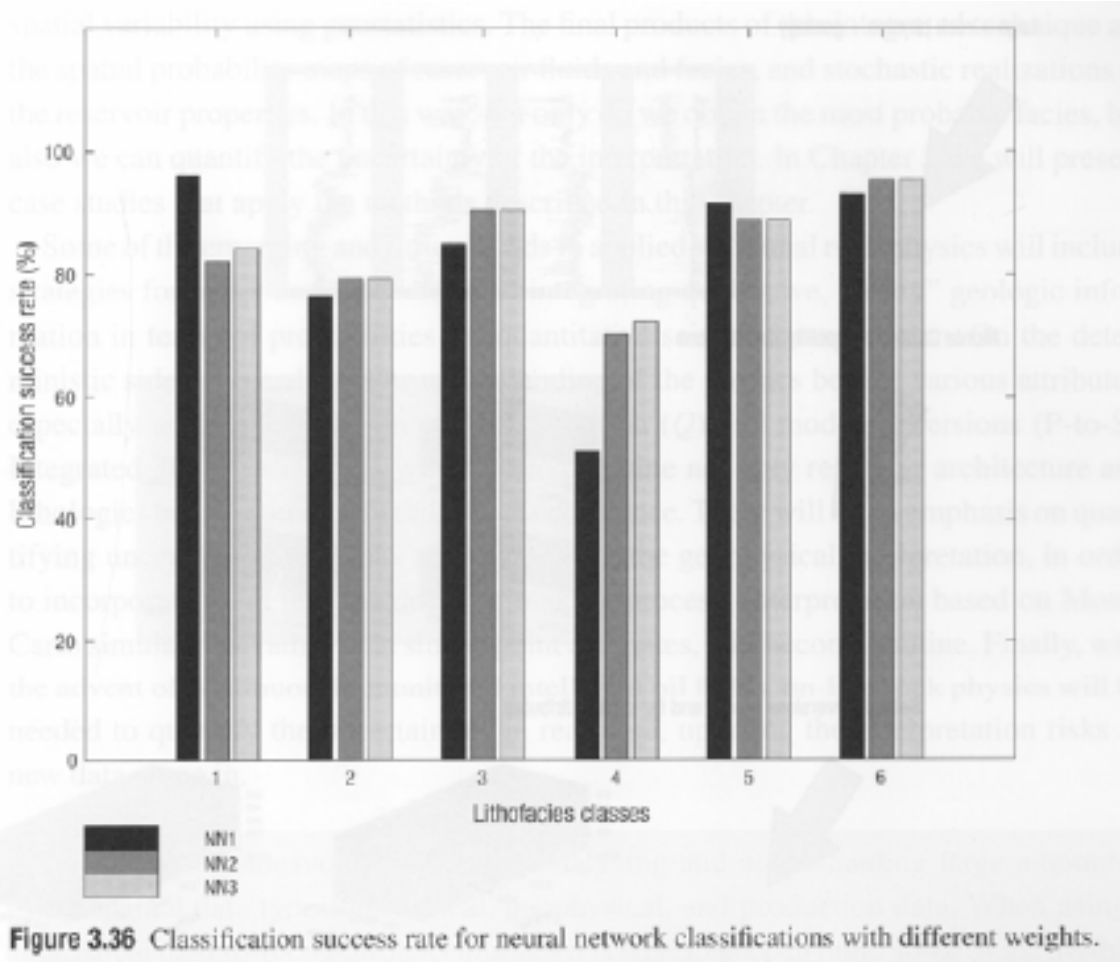


Figure 3.36 Classification success rate for neural network classifications with different weights.

Bias to certain facies by tuning the weights

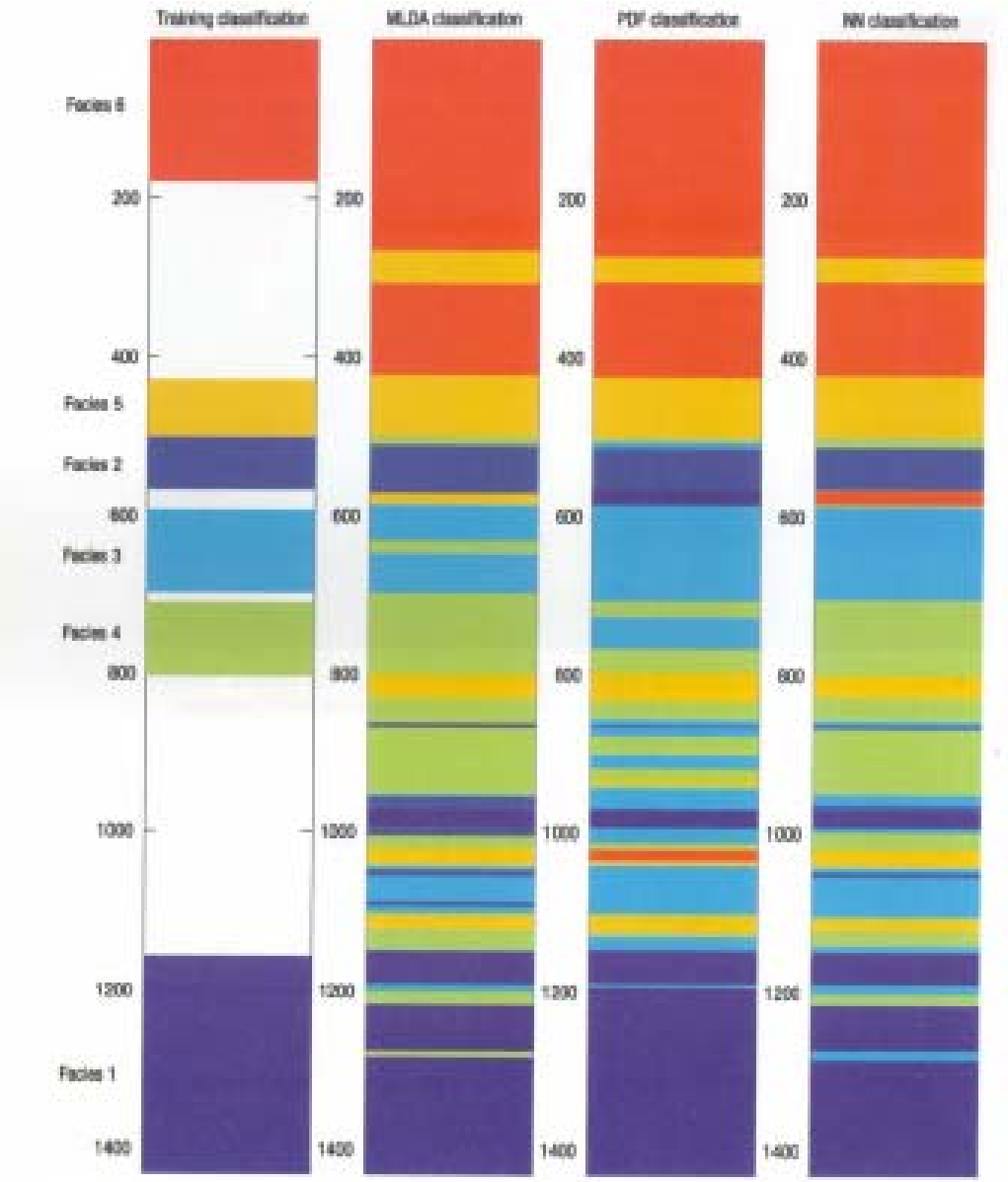


Plate 3.33 Comparing discriminant analysis, Bayes' rule and neural network classification results in a type-well. The depth axis is annotated with sample number. Sample number 1 is located at about 2075 m and sample number 1400 is located at about 2300 m.

THANK YOU!