# Discussion of Thursday talks

Will Welch, University of British Columbia

# Discussion of Thursday talks

Winnie-the-Pooh

# Helmut Kroger

- Small-world architectures
- Scale-free neural networks
- Accuracy and training-set size
- SWN learns fastest, rewiring helps (unless overloaded)
- Overfitting?

# Stan Young

- Disease data, metabolytes data, protein data, gene expression data, genotype data

- Complex data hierarchy (NPCDS!)

- Two-clustering of two-way tables

- $\mathrm{R}^2\mathrm{P}$: recursive recursive partitioning

## Mu Zhu

- Unbalanced classes in classification
- Average precision instead of misclassification rate
- Radial basis functions around *only class 1 objects*
- Like KNN, SVM but computationally more efficient
- Only have to model $p_1(\mathbf{x})$?

# Grigoris Karakoulas

- Unbalanced classes
- ROC criterion (like average precision)
- Trees based on greedy ROC, projections of explanatory variables (features)?
- RBTree beats NB
- ROCBoost beats AdaBoost

# Russell Steele

- Model selection: *IC wars
- AIC overfits
- BIC bigger penalty, but theory for BIC?
- New form of BIC
- Complex analysis, “rusty”, “hard”, “approximations not very good”
- Why not cross validation?

# Steven Wang

- Clustering categorical data
- Hamming distance to give Categorical Distance (CD) vector
- Different origins
- CD algorithm beats AutoClass, K-modes
- Automatically estimates number of clusters
- Distance-based methods using Hamming distance?

# Xianping Liu

- Industrial-strength clustering
- Automated (but lots of options mentioned!)
- K-means and its extensions (fast)
- “Unfaithful somebody”

## Simon Gluzman

- Approximate $f(x)$ from Taylor-series expansion (a few terms)
- Multivariate non-polynomial (Root) approximants
- Accurate approximation (order of approximation?)
- Smoothing, stabilizing (polynomials known to be erratic)
- But for what functions $f(x)$?
- Multivariate $\mathbf{x}$: based on low-order polynomial regressions
- KNN suitable for step functions; linear regression for linear (in $\mathbf{x}$) functions

- Splines, etc?

- Fully automated

# Wenxue Huang

- Large samples, large number of variables
- Dimension reduction
- Versus feature selection
- Association between $y$ and $x$ (dependence degree)
- Set of good features (explanation base) for $y$: no redundant information
- Not unique (but finding 1 is enough)
- Can reduce variables further in practice
- Interpretible information criterion (unlike entropy?)

# Summary

- SWNs: implications for neural networks?
- Complex data structures
- Criteria/algorithms for rare-class problems
- Model selection? (Overfitting, *ICs)
- Automate! (Variable/feature selection, number of clusters, approximator, etc.)
- (Variable selection is sometimes more important that the method/algorithm)