



# An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function

*Part of the BIASPROFS Project:*

[www.cs.kent.ac.uk/projects/biasprofs](http://www.cs.kent.ac.uk/projects/biasprofs)

- ⌘ Andy Secker, Alex Freitas (University of Kent)
- ⌘ Matthew Davies, Darren Flower (University of Oxford)
- ⌘ Jon Timmis, Miguel Mendao (University of York)



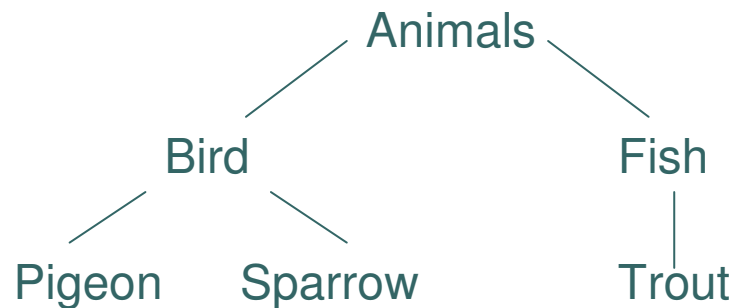
# Outline

- ⌘ Introducing hierarchies
  - ┆ Terminology
  - ┆ Top-down classification
- ⌘ GPCR proteins
  - ┆ Motivation
- ⌘ Selective approach to top-down classification
- ⌘ Future
  - ┆ Using big bang approach



# Hierarchies

- ☞ Lots of class data is flat:
  - 1 Red, Yellow, Blue
- ☞ ...but some is naturally hierarchical:
  - 1 Pigeon, Sparrow, Trout
  - 1 Bird.Pigeon, Bird.Sparrow, Fish.Trout



- ☞ Use the hierarchy to improve classification
  - 1 Example: if we're sure data instance is a Bird (maybe it had wings?), then no need to consider the class Trout



# Hierarchies

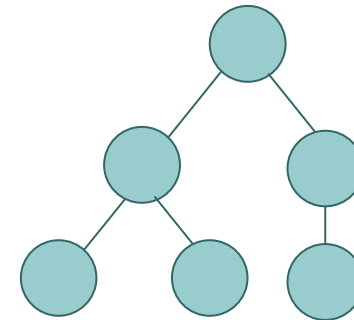
- ⌘ Data instance belongs to more than one class (but only 1 at each level)
- ⌘ Hierarchies are found in
  - ┆ Text mining
    - Document collections
      - Medical, academic, etc.
    - Eg: data mining → classification → bioinformatics
  - ┆ Web mining
    - Web directories
  - ┆ Bioinformatics
    - Protein databases



# Terminology 1

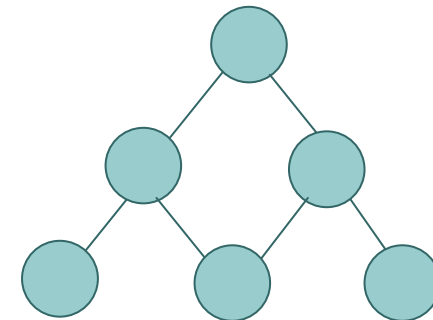
## ☞ Tree

- 1 Exactly one parent per node



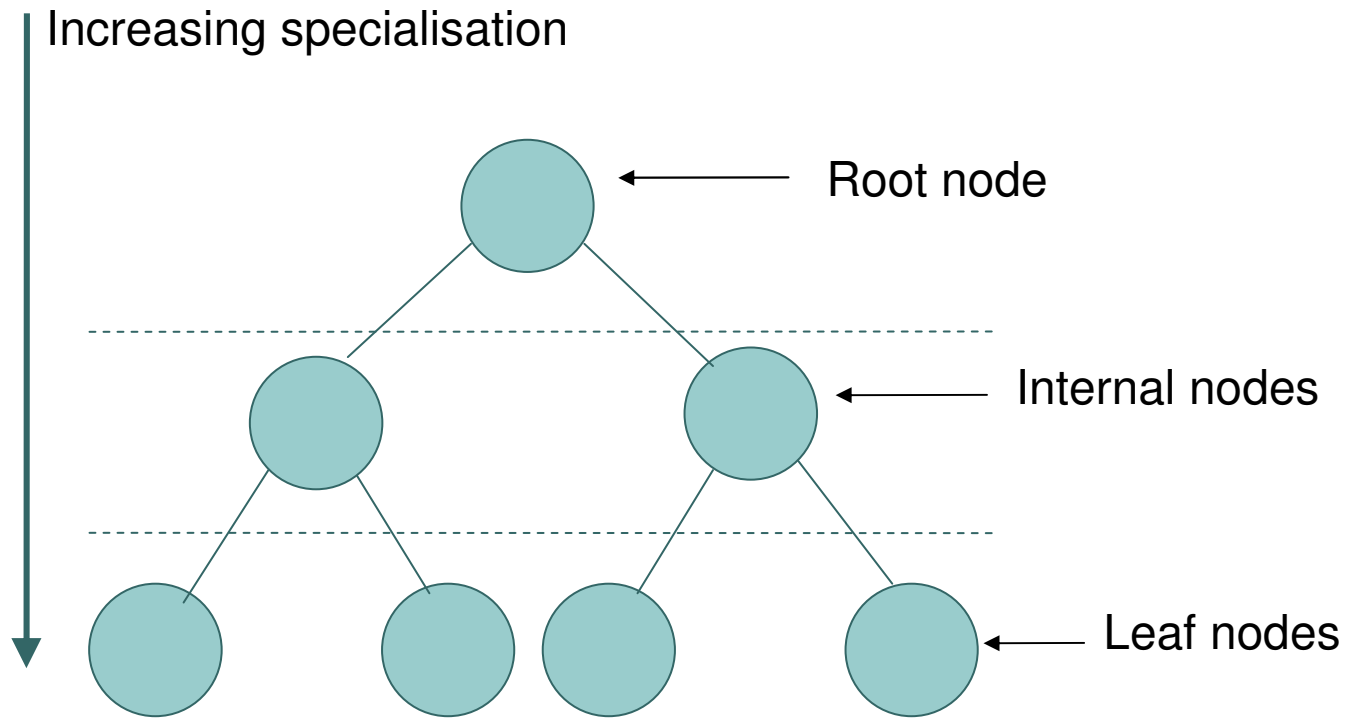
## ☞ Directed Acyclic Graph (DAG)

- 1 Nodes may have more than one parent
- 1 Not used in this study





# Terminology 2





# Classification methods

1. Flatten class hierarchy
  - 1 Only predict classes at one level
    - Predict at most specific level and infer superclasses
  - 1 Wastes the information inherent in the hierarchy which could be used to improve accuracy
    - Instance must belong to all superclasses
  - 1 Possibility of huge number of classes
    - Small number of examples per class
    - Some classes extremely similar to each other



# Classification methods, cont...

## 2. Big bang

- 1 Consider all levels of hierarchy at once during training
- 1 Single classification model built
- 1 Less straightforward than others

## 3. Top-down

- 1 Middle way between flattening and big bang
- 1 Common, simple

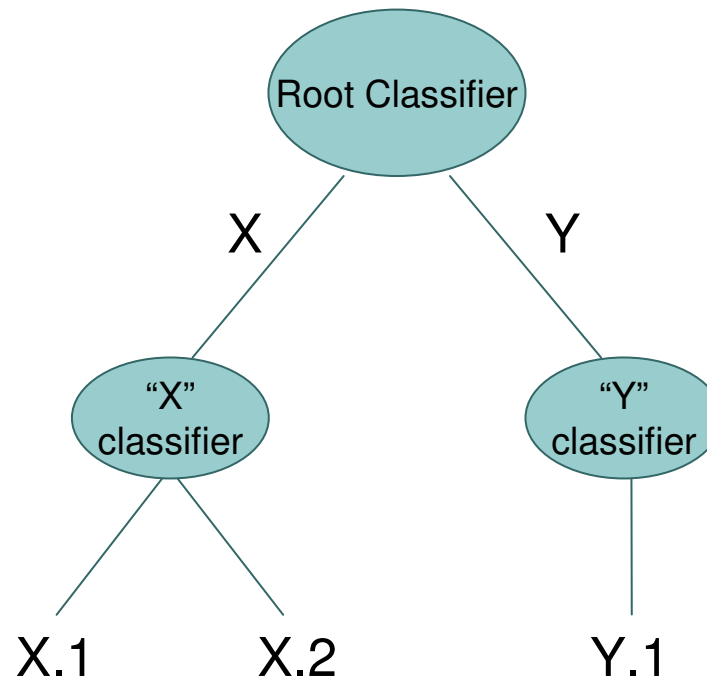


# Top-down

- ⌘ Solve a flat classification problem once for each level
- ⌘ Use popular, well understood algorithms
  - 1 Instance classified by a different model at each level
  - 1 Each classifier appends a class, increasing in specialisation
- ⌘ Disadvantage: misclassifications are propagated to the next level
  - 1 There is no way to correct misclassification at higher level (blocking)
  - 1 Bad news for deep tree

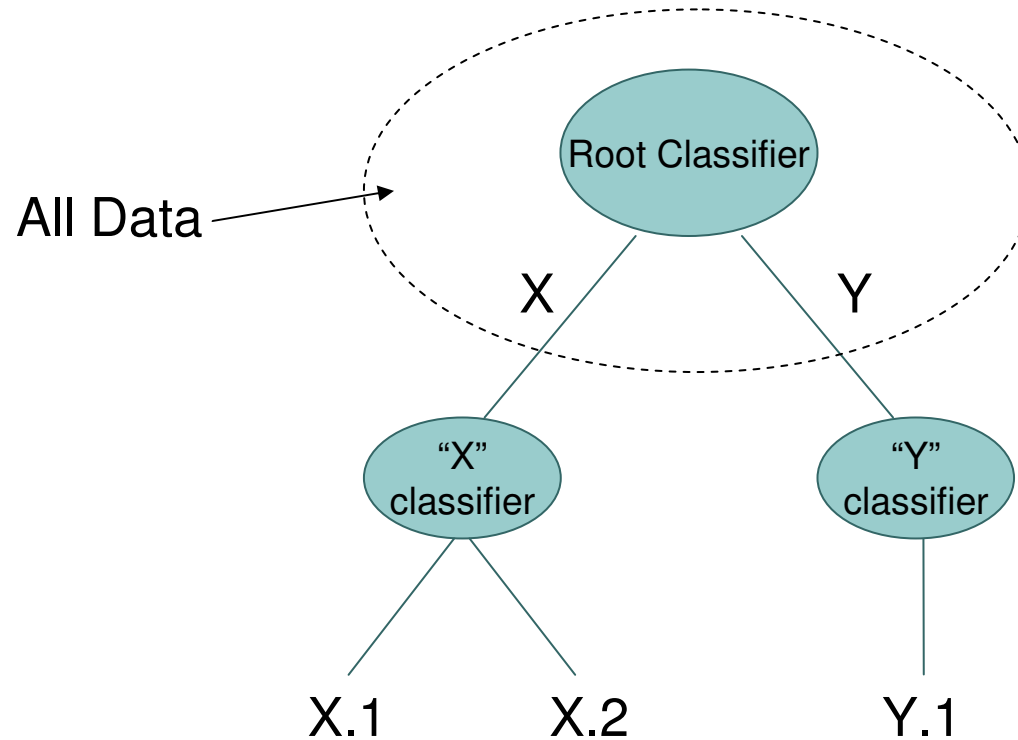


# Top-down approach



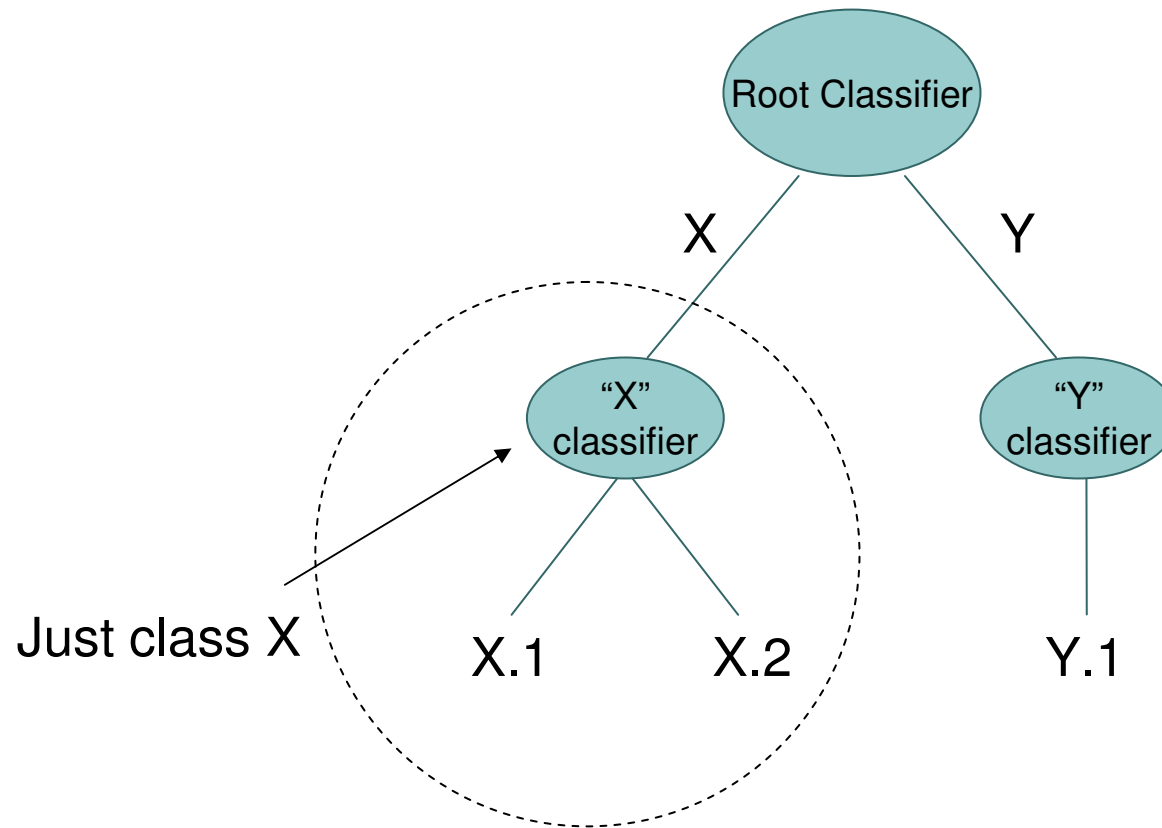


# Top-down: Training





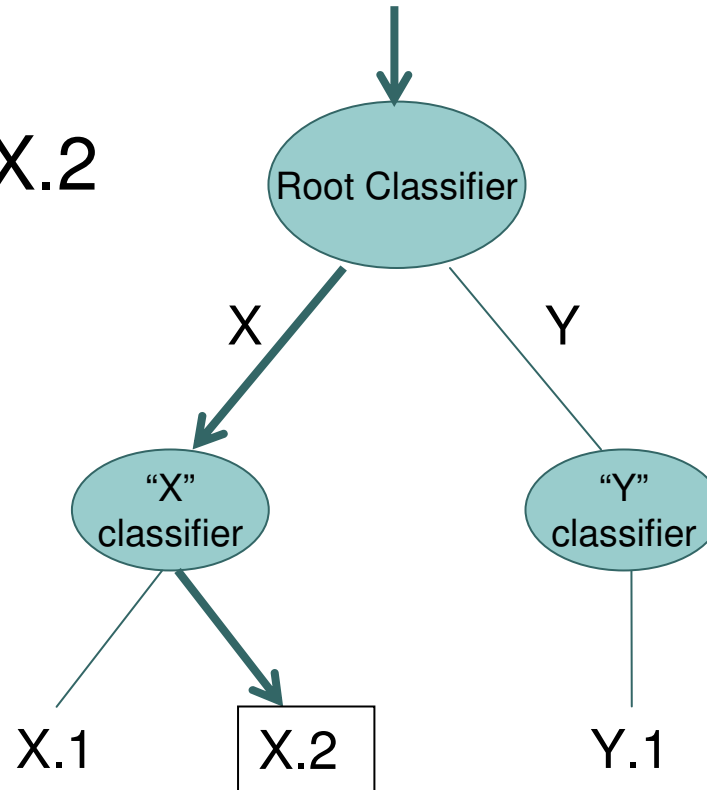
# Top-down: Training





# Top-down: Testing

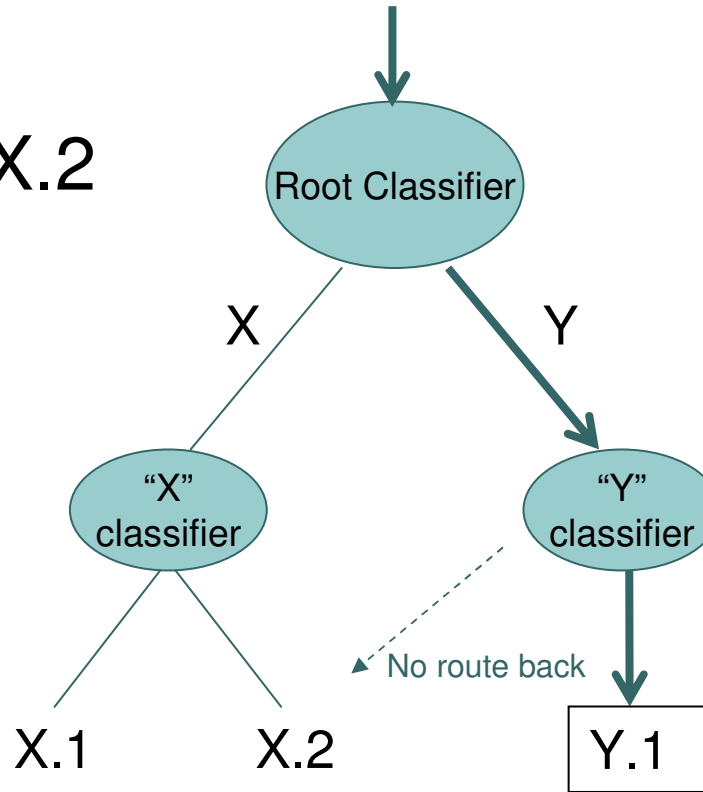
Classify: X.2





# Top-down: Testing

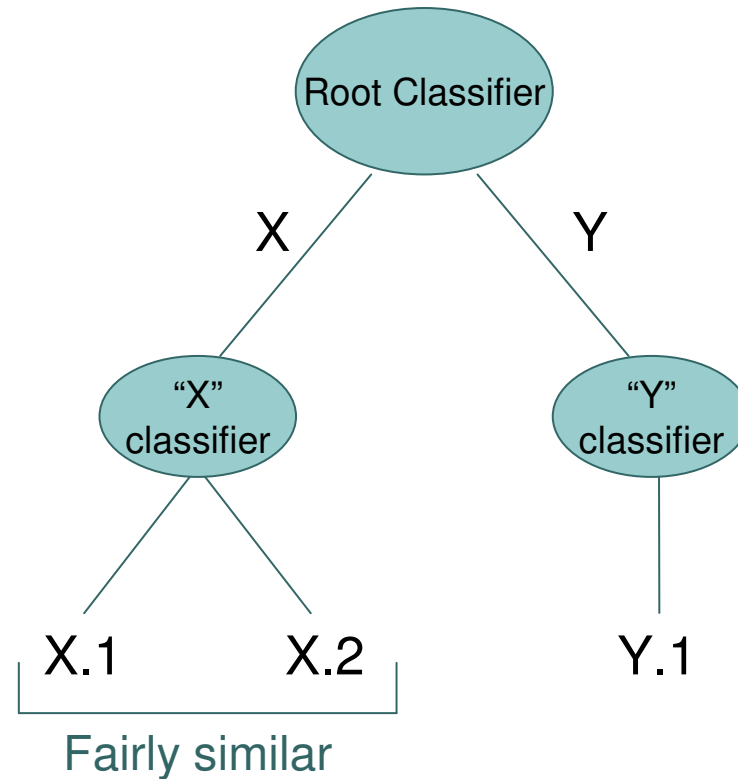
Classify: X.2





# Evaluation methods

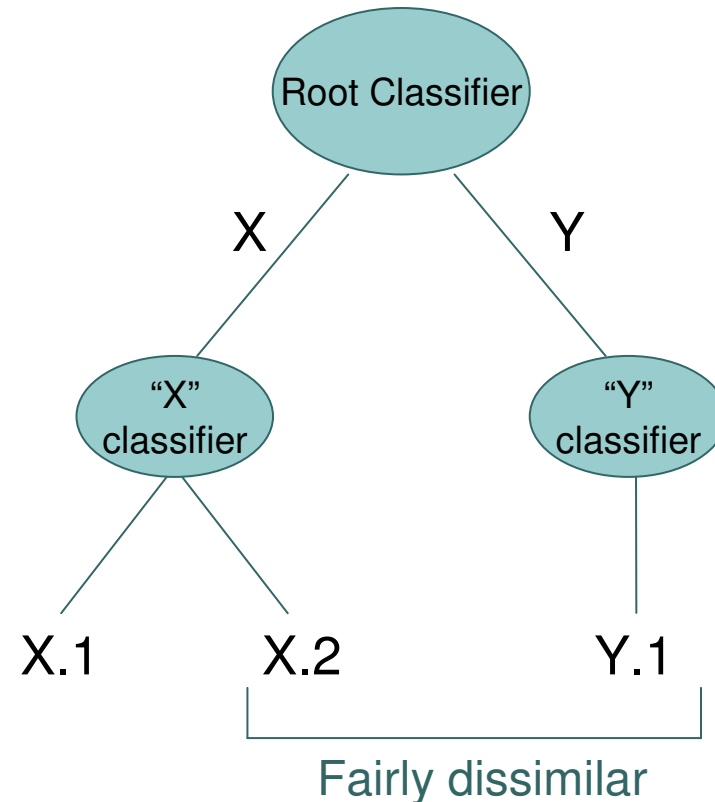
- ⊘ Unlike flat classification, there exist different “distances” between classes





# Evaluation methods

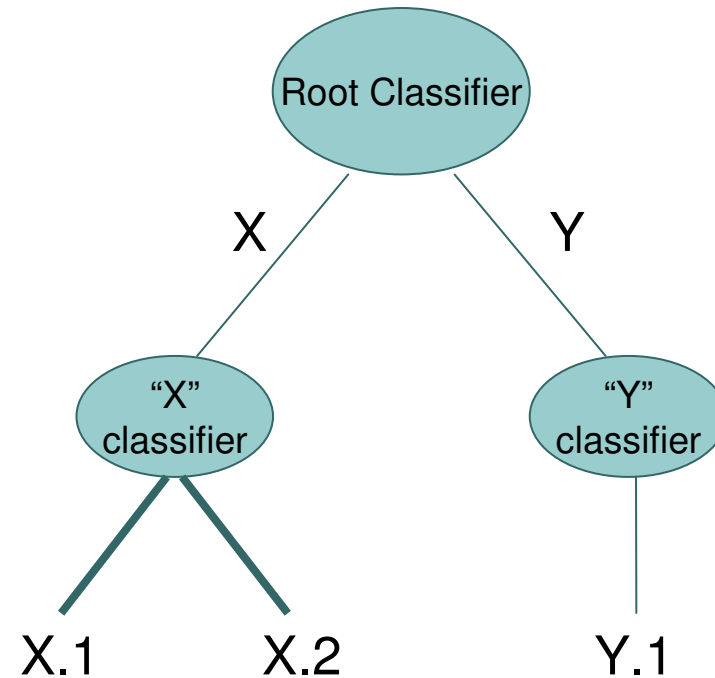
- ⌘ Unlike flat classification, there exist different “distances” between classes
- ⌘ Take this similarity into account when judging quality of classification
- ⌘ X.2 classified as X.1 is better than X.2 classified as Y.1 as X.1 and X.2 have common parent





# Evaluation methods

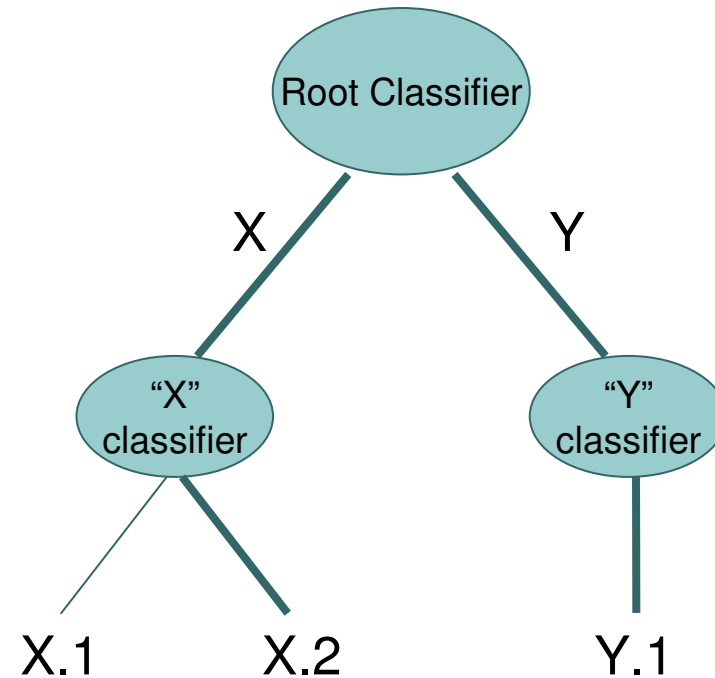
- ⌘ Example: Edge distance
- ⌘ X.2 classified as X.1
  - 1 2 edges





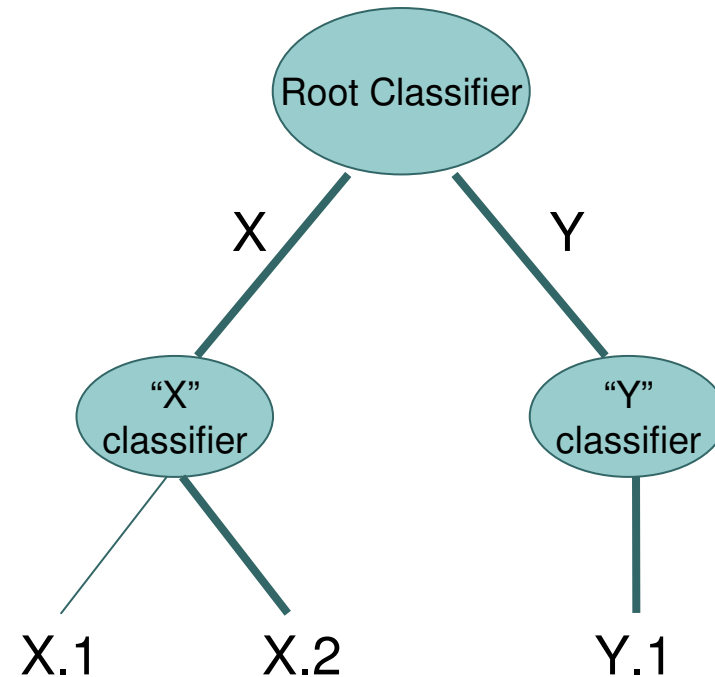
# Evaluation methods

- ⌘ Example: Edge distance
  - ⌘ X.2 classified as X.1
    - 1 2 edges
  - ⌘ X.2 classified as Y.1
    - 1 4 edges



# Evaluation methods

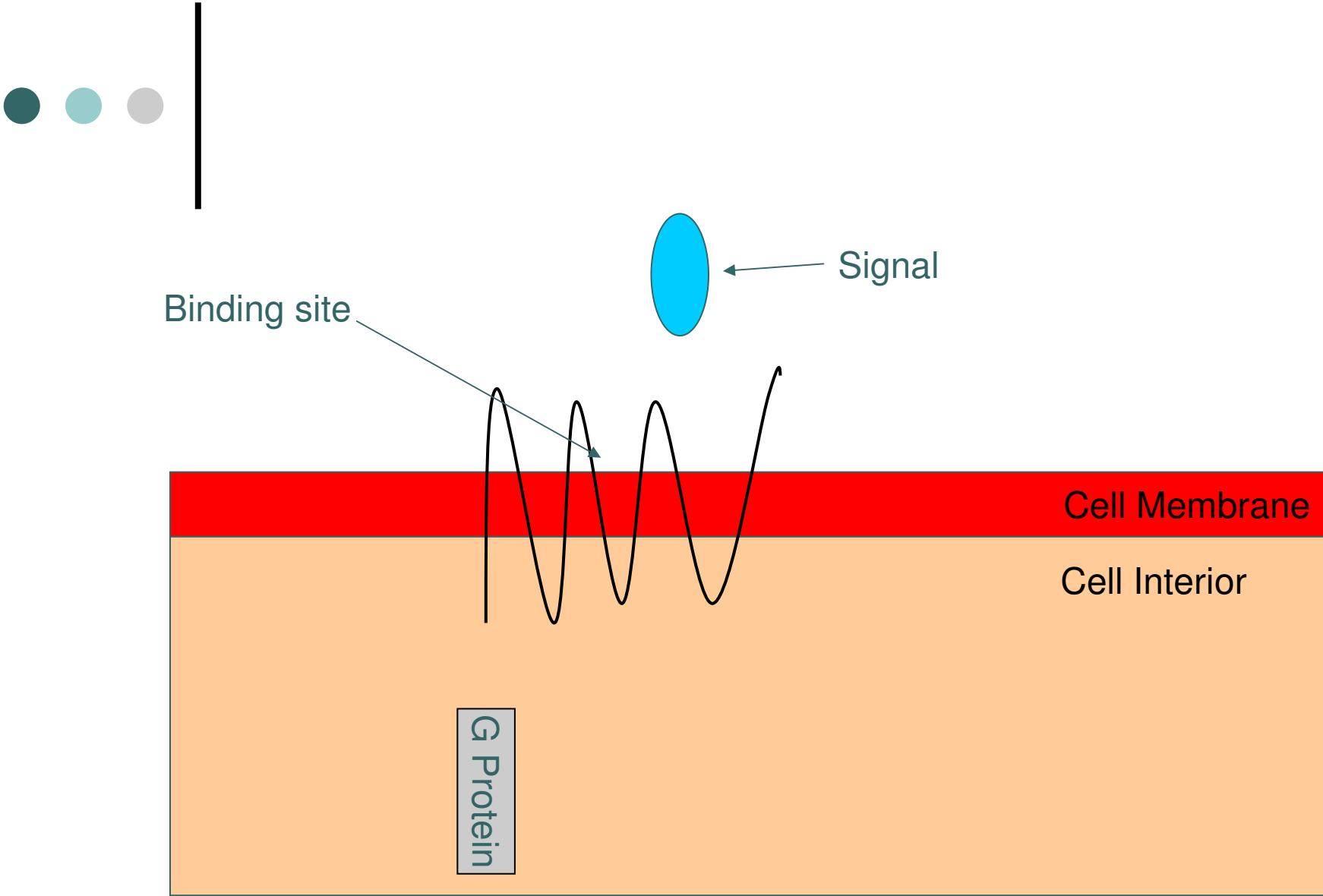
- ⌘ Example: Edge distance
  - ⌘ X.2 classified as X.1
    - 1 2 edges (scores  $\frac{1}{2}$ )
  - ⌘ X.2 classified as Y.1
    - 1 4 edges (scores  $\frac{1}{4}$ )
- ⌘ Other strategies
  - 1 Depth dependent weighting
  - 1 Cost matrix
- ⌘ DAG has multiple paths

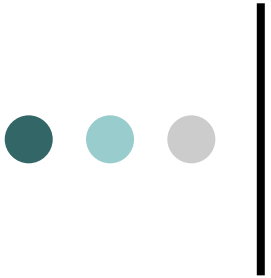




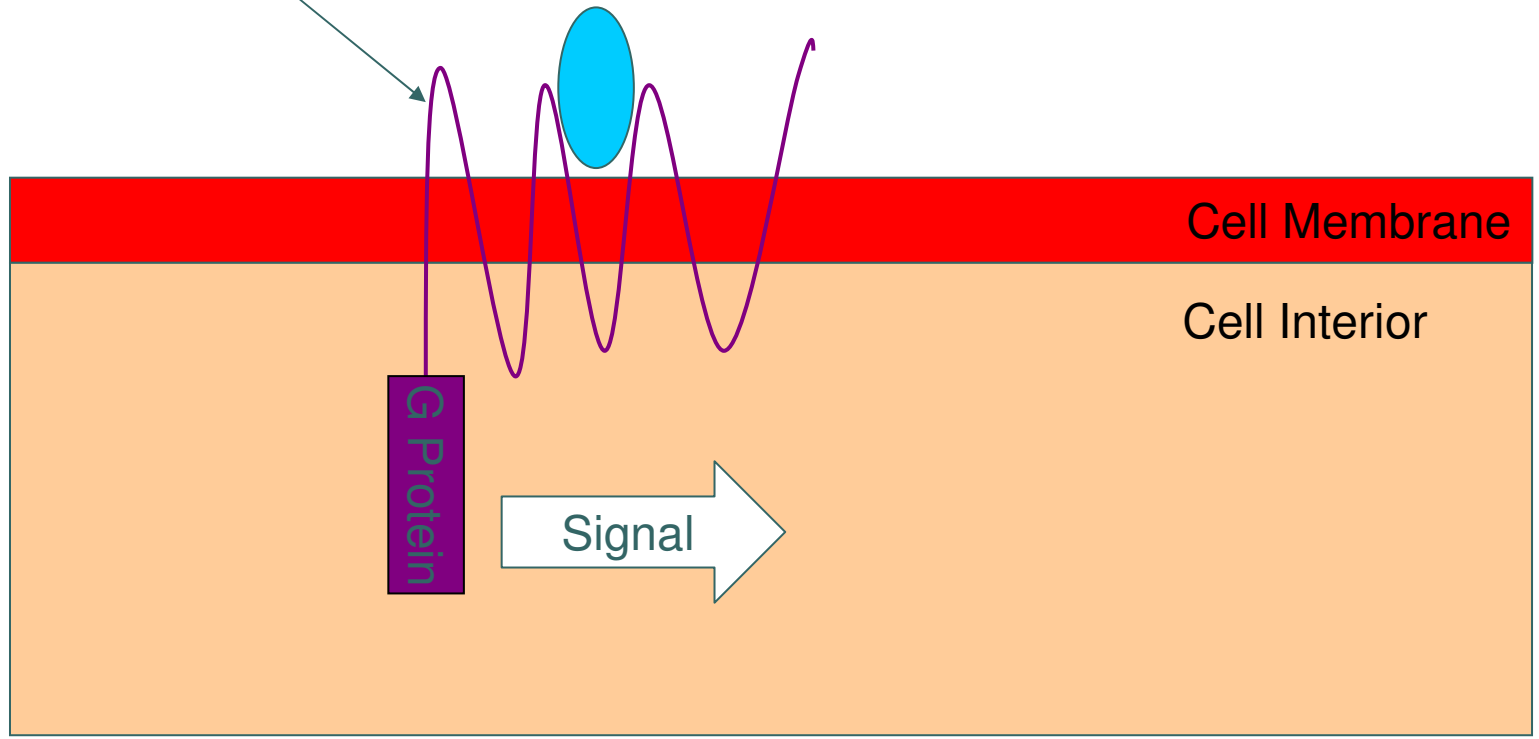
# GPCR proteins

- ⌘ A GPCR (G-Protein Coupled Receptor) is a particular type of protein
- ⌘ Allows exterior message to influence cell's (internal) behaviour
  - ┆ Takes signals through cell membrane
  - ┆ 7 transmembrane regions





Activated GPCR



Cell Membrane

Cell Interior

G Protein

Signal



# More on GPCRs

- ⌘ Regulate basic cell processes
- ⌘ Protein databases contain millions of entries
  - ┆ Manual annotation is impossible
  - ┆ Prediction of function
- ⌘ Activation stimulus unknown for around 80% of GPCRs
- ⌘ Targeted by around 50% of licensed drugs
  - ┆ Multiple attack sites and strategies
- ⌘ Superfamily of membrane proteins
  - ┆ Naturally sorts into hierarchy
  - ┆ Hierarchy ignored in classification



# Data preparation

- ⌘ Our dataset was constructed by hand
  - ┆ 8866 proteins
  - ┆ 3 levels
    - 1. 110 classes at most specific level
    - 2. 38 at middle level
    - 3. 5 at most general level
- ⌘ Slightly smaller dataset after pre-processing
- ⌘ Representations issues:
  - ┆ Proteins are variable in length
  - ┆ Primary sequence symbolic attributes
- ⌘ Convert to fixed number of predictor attributes, continuous values



# Data preparation

- ☞ Proteins made from chains of amino acids

- 1 Alanine (A), Cysteine (C), Lysine (K), etc...

- ☞ Primary sequence

- 1 Ordering of amino acids in chain

- >gi|1204090|emb|CAA56455.1| dopamine receptor [Takifugu rubripes]

- MAQNFSTVGDGKQMLLERDSSKRVLTCGFLSLLIFTLLGNTLVCVAVTKFRHLRSKVTNFFVISLAISD  
LLVAILVMPWKAATEIMGFWPFGEFCNIWVAFDIMCSTASILNLCVISVDRYWAISSPFRYERKMPKVA  
CLMISVAWTLSVLISFIPVQLNWHKAQTASYVELNGTYAGDLPPDNCDSLNRTYAISSSLISFYIPVAI  
MIVTYTRIYRIAQKQIRRISALERAAESAQNRHSSMGNSLSMESECSFKMSFKRETKVLKTLVIMGVFV  
CCWLPFFILNCMVPFCEADDTDFPCISSTTFDVFVWFGWANSSLNPIIYAFNADFRKAFSILLGCHRLC  
PGNSAIEIVSINNTGAPLSNPSCQYQPKSHIPKEGNHSSSYVIPHSILCQEEELQKKDGFGGEMEVLVN  
NAMEKVSPAISGNFDSDAAVTLETINPITQNGQHKSMS

- ☞ Proteins variable in length

- 1 Longest in Genbank is 34,350 amino acids

- ☞ Proteins fold into very complex shapes



# Data preparation

- ☞ Use “Z-values” to represent each amino acid
  - 1 Each amino acid has numerous physical/chemical properties
  - 1 26 of these reduced to 5 values using principle component analysis
  - 1 Allows reduction of protein to 5 predictor attributes

Primary Sequence: A-R-N-D-C

A, 0.24, -2.32, 0.60, -0.14, 1.30

R, 3.52, 2.50, -3.50, 1.99, -0.17

N, 3.05, 1.62, 1.04, -1.15, 1.61

D, 3.98, 0.93, 1.93, -2.46, -0.75

C, 0.84, -1.67, 3.71, 0.18, -2.65

---

**Protein = 2.33   0.21   0.76   -0.32   -0.13**

---

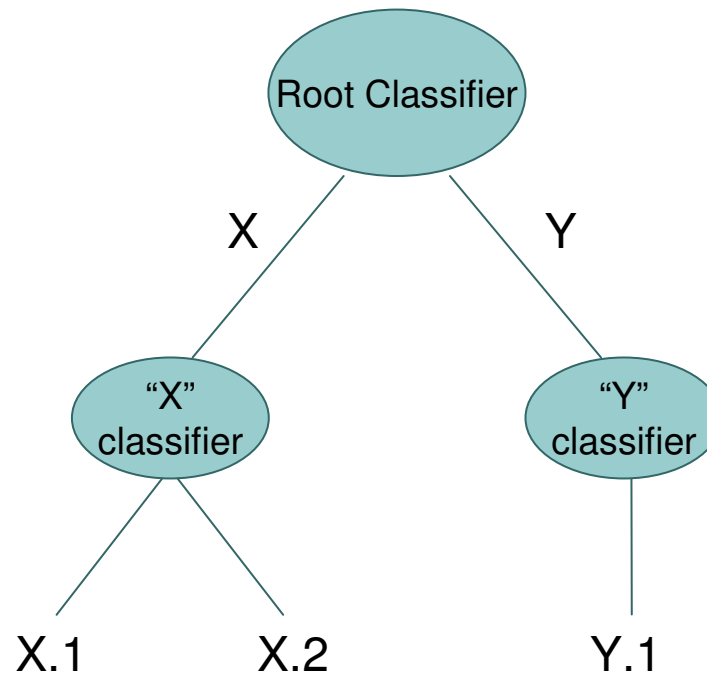


# Proposed selective top-down approach

- ⌘ Hypothesis: the same classifier may not be suited to all levels of hierarchy
  - ┆ Exploit different bias
  - ┆ Different amounts of training data
  - ┆ Some characteristics important at one level could be redundant at lower levels
- ⌘ Solution: Choose most suitable classifier for each node from a set of candidates
  - ┆ In a data-driven manner
  - ┆ Greedy

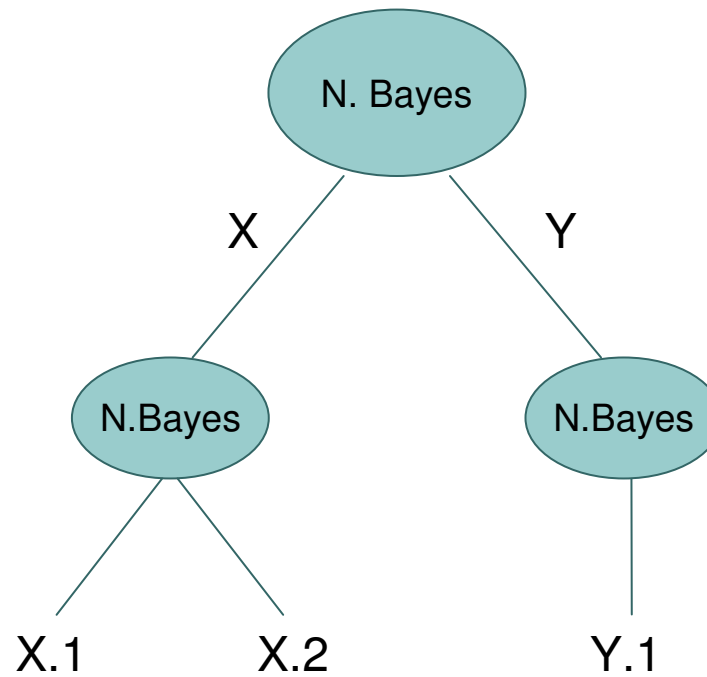


# The usual classifier



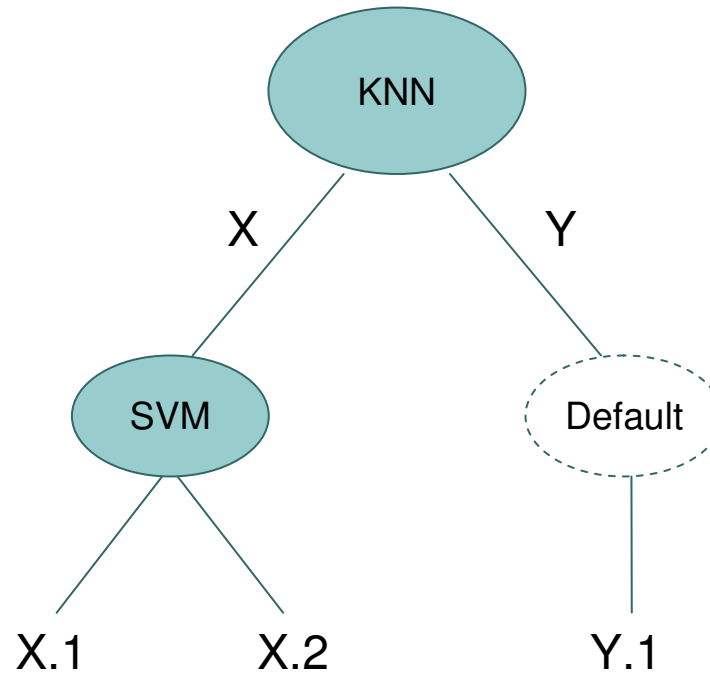


# An improved classifier





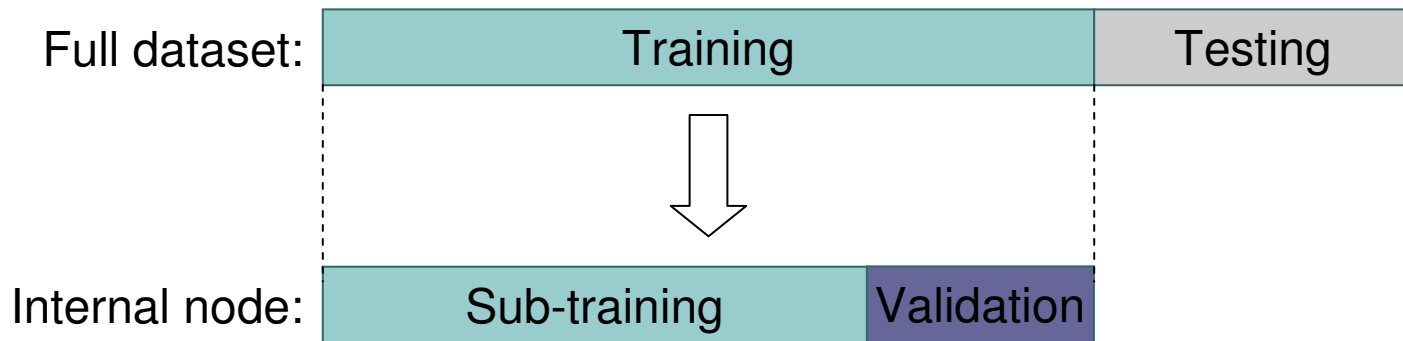
# An improved classifier





# Differences from standard approach

- ⌘ Training set subdivided at each node into sub-training and validation sets
- ⌘ Each classifier from menu is trained using sub-training
- ⌘ Performance is evaluated using validation set
- ⌘ Internal cross validation not found to be helpful



- ⌘ Best classifier is then selected, re-trained using full training set and stored in hierarchy



# Experimental protocol

## ☞ Classifier menu

1. Naïve Bayes
2. Bayesian network
3. SMO (support vector machine)
4. 3 nearest neighbours
5. PART (a decision list)
6. J48
7. Naïve Bayes tree
8. Multi-layer neural network with back propagation
9. AIRS2 (Artificial Immune System classifier)
10. Conjunctive rule learner

## ☞ Training set is split at internal nodes

- 1 80% sub-training, 20% validation
- 1 Guarantee at least 1 test instance for each class

## ☞ 30 independent runs of 10-fold cross-validation



# Results: grid

- ⌘ Comparison between selective and standard top-down classifiers
- ⌘ Statistically significant increase in accuracy highlighted
  - 1 Corrected resampled t-test
  - 1 Standard t-test has issues with
    - Cross validation
    - Large number of runs
- ⌘ Accuracy per level (error accumulates)

| Standard top-down classifiers |           |       |                      |       |       |         |                |       |                   |           |
|-------------------------------|-----------|-------|----------------------|-------|-------|---------|----------------|-------|-------------------|-----------|
| Naïve Bayes                   | Bayes Net | SMO   | 3 Nearest Neighbours | PART  | J48   | NB Tree | Neural Network | AIRS2 | Conjunctive Rules | Selective |
| 73.33                         | 77.40     | 66.44 | 90.75                | 89.49 | 90.37 | 89.53   | 66.44          | 81.66 | 71.91             | 90.59     |
| 47.74                         | 53.40     | 38.88 | 71.59                | 73.52 | 73.45 | 72.34   | 31.89          | 57.81 | 45.51             | 73.77     |
| 23.12                         | 29.83     | 15.55 | 55.71                | 57.90 | 57.41 | 55.27   | 4.15           | 42.61 | 9.37              | 58.08     |





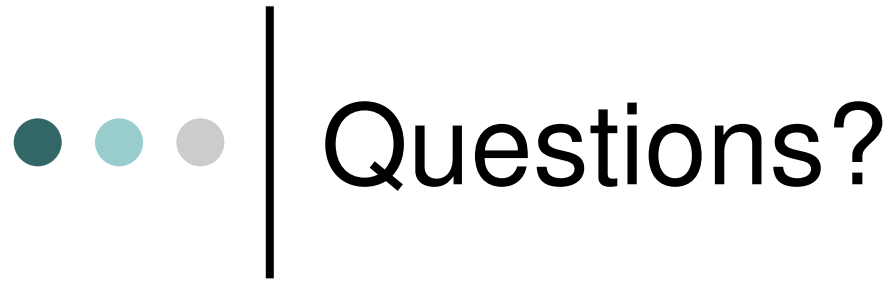
# Future: big bang

- ⌘ Top-down has many advantages but misclassifications accrue
  - ┆ Real issue with large numbers of levels
- ⌘ Big bang builds a single classification model
  - ┆ Classifier has access to all levels when building model
  - ┆ Run once to classify single instance
  - ┆ Misclassifications do not accumulate
  - ┆ Possibly more comprehensible model
  - ┆ Drop test instance at intermediate level if low confidence of correct classification at lower level
- ⌘ More complex than top-down
- ⌘ Harder to use standard algorithms
  - ┆ C4.5(H)



# Summary

- ⌘ Classifying data instances into a hierarchy of classes poses some unique challenges
- ⌘ Top-down approach is common
  - ┆ Allows use of standard algorithms
  - ┆ Run the same algorithm regardless of the data
- ⌘ Selective approach exploits classifier bias
- ⌘ Big bang approach is future direction
  - ┆ More complex than top-down



# Questions?

*BIASPROFS Project:*

[www.cs.kent.ac.uk/projects/biasprofs](http://www.cs.kent.ac.uk/projects/biasprofs)