LA-UR- 11-04419

Approved for public release; distribution is unlimited.

Title:	Turning Bayesian Model Averaging Into Bayesian Model Combination
Author(s):	Kristine Monteith, James Carroll, Kevin Seppi, Tony Martinez
Intended for:	conference presentation and internet release

Los Alamos NATIONAL LABORATORY
EST. 1943

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (7/06)

Turning Bayesian Model Averaging Into Bayesian Model Combination (U)

Kristine Monteith, James L. Carroll, Kevin Seppi, Tony Martinez

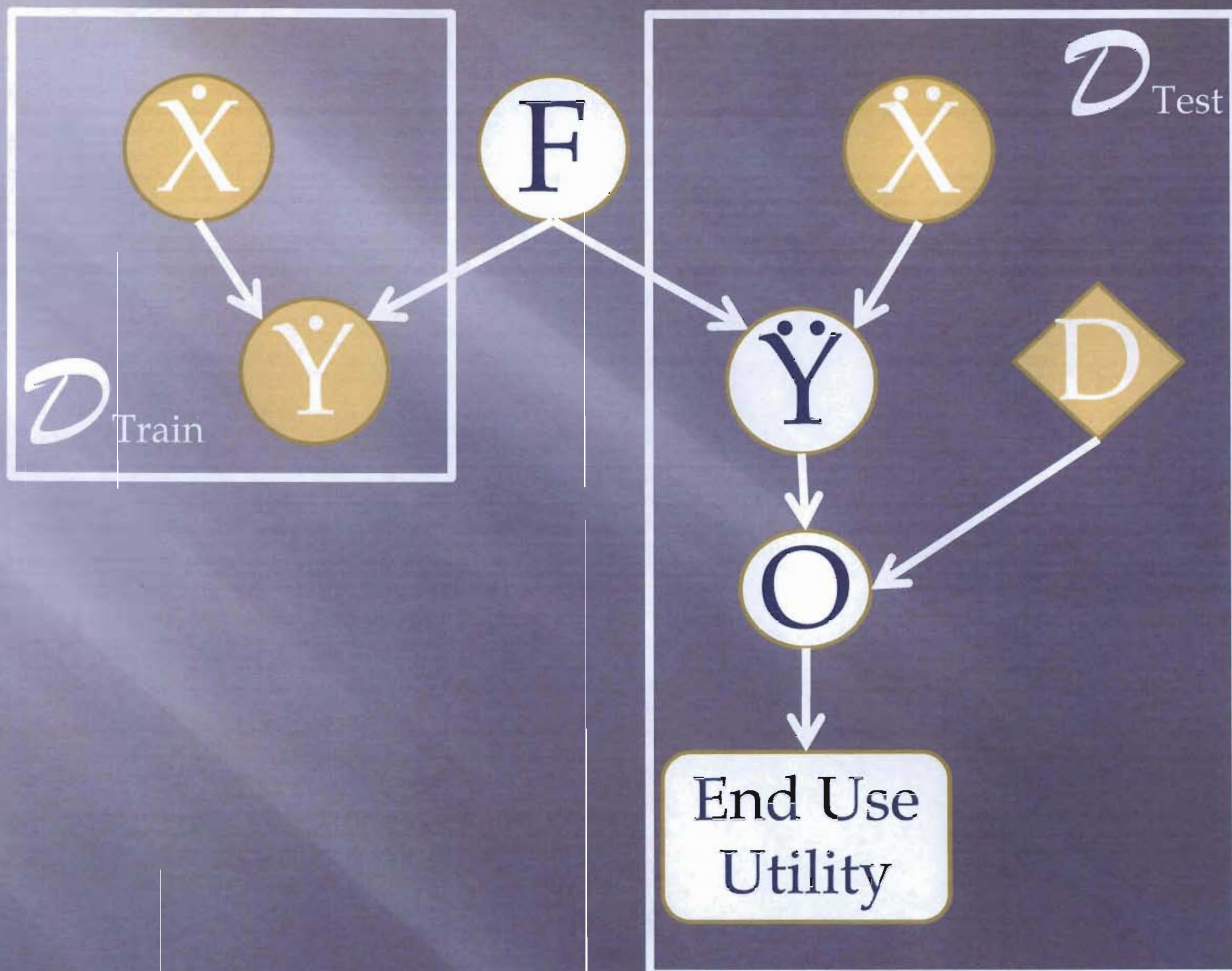
IJCNN 2011

LA-UR XX_XXXX

Abstract

Bayesian methods are theoretically optimal in many situations. Bayesian model averaging is generally considered the standard model for creating ensembles of learners using Bayesian methods, but this technique is often outperformed by more ad hoc methods in empirical studies. The reason for this failure has important theoretical implications for our understanding of why ensembles work. It has been proposed that Bayesian model averaging struggles in practice because it accounts for uncertainty about which model is correct but still operates under the assumption that only one of them is. In order to more effectively access the benefits inherent in ensembles, Bayesian strategies should therefore be directed more towards model combination rather than the model selection implicit in Bayesian model averaging. This work provides empirical verification for this hypothesis using several different Bayesian model combination approaches tested on a wide variety of classification problems. We show that even the most simplistic of Bayesian model combination strategies outperforms the traditional ad hoc techniques of bagging and boosting, as well as outperforming BMA over a wide variety of cases. This suggests that the power of ensembles does not come from their ability to account for model uncertainty, but instead comes from the changes in representational and preferential bias inherent in the process of combining several different models.

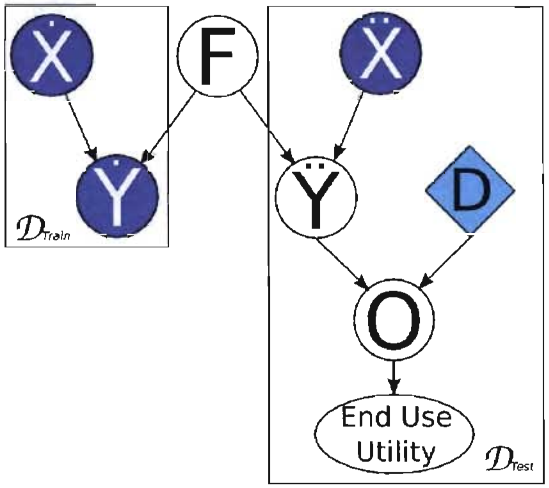
Supervised UBDTM



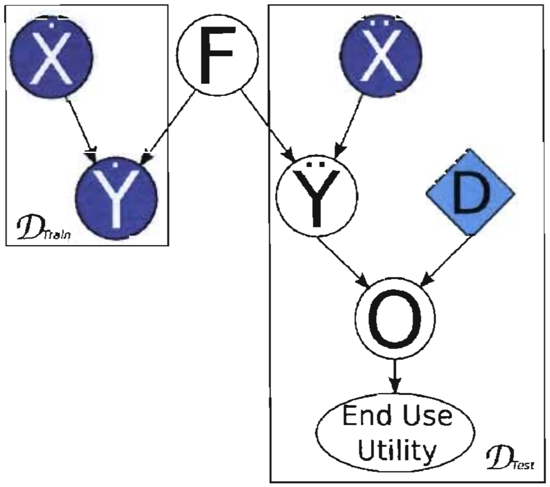
THE MATHEMATICS OF LEARNING

- <u>Learning about F:</u>

$$p(f \mid x, y) = \frac{p(y \mid x, f)\, p(f)}{\int p(y \mid x, f)\, p(f)\, df}$$



THE MATHEMATICS OF LEARNING

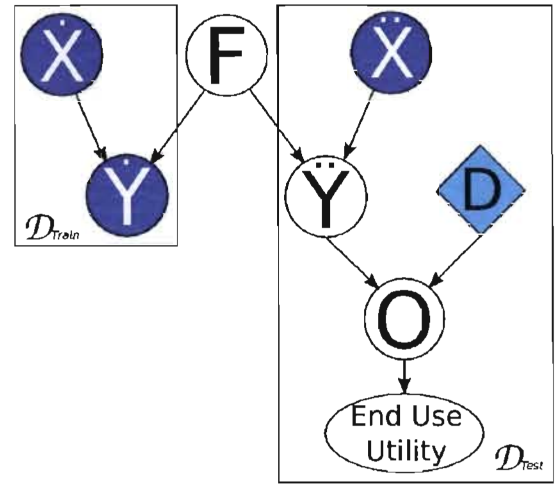


- Learning about F:

$$p(f \mid x, y) = \frac{p(y \mid x, f)\, p(f)}{\int p(y \mid x, f)\, p(f)\, df}$$

Repeat to get $p(f \mid D_{Train})$

THE MATHEMATICS OF LEARNING



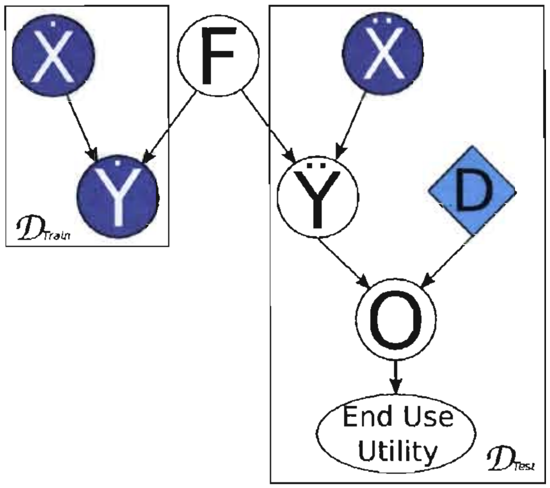
- <u>Learning about F:</u>

$$p(f \mid x, y) = \frac{p(y \mid x, f)\, p(f)}{\int p(y \mid x, f)\, p(f)\, df}$$

- <u>Classification or Regression:</u>

$$p(\ddot{y} \mid \ddot{x}, D_{Train}) = \int p(\ddot{y} \mid \ddot{x}, f)\, p(f \mid D_{Train})\, df$$

THE MATHEMATICS OF LEARNING



- Learning about F:

$$p(f \mid x, y) = \frac{p(y \mid x, f)\,p(f)}{\int p(y \mid x, f)\,p(f)\,df}$$

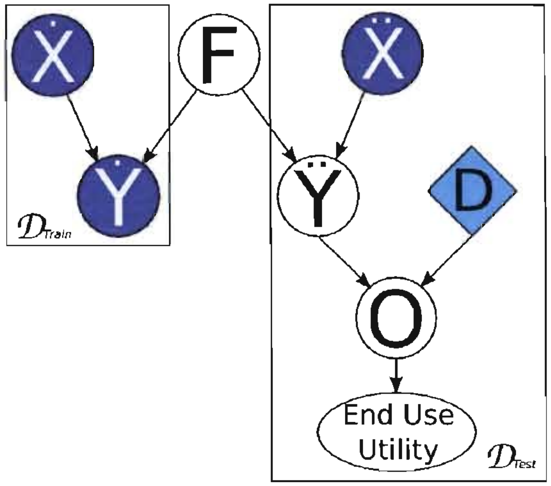
- Classification or Regression:

$$p(\ddot{y} \mid \ddot{x}, D_{Train}) = \int p(\ddot{y} \mid \ddot{x}, f)\,p(f \mid D_{Train})\,df$$

- Decision making:

$$\hat{d} = \underset{d \in D}{\arg\max} \sum_{\ddot{y} \in Y} U(o)\,p(o \mid \ddot{y}, d)\,p(\ddot{y} \mid \ddot{x}, D_{Train})$$

THE MATHEMATICS OF LEARNING



- <u>Learning about F:</u>

$$p(f \mid x, y) = \frac{p(y \mid x, f)\, p(f)}{\int p(y \mid x, f)\, p(f)\, df}$$

- <u>Classification or Regression:</u>

$$p(\ddot{y} \mid \ddot{x}, D_{Train}) = \int p(\ddot{y} \mid \ddot{x}, f)\, p(f \mid D_{Train})\, df$$

- <u>0-1 Loss Decision making:</u>

$$\hat{d} = \underset{\ddot{y} \in Y}{\arg\max}\; p(\ddot{y} \mid \ddot{x}, D_{Train})$$

REALISTIC MACHINE LEARNING:

TRAINING

Training Data

Learning Algorithm

hypothesis

USING

Input: Unlabeled Instance

hypothesis

Output: Class label

USING A LEARNER

Info about
- sepal length
- sepal width
- petal length
- petal width

hypothesis


Iris Setosa


Iris Virginica

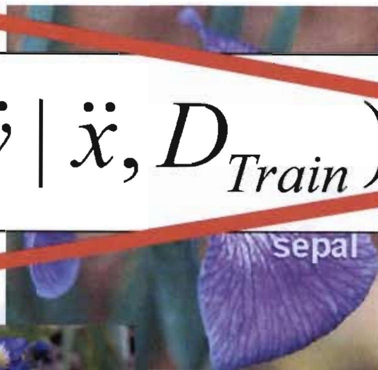

Iris Versicolor

USING A LEARNER

Info about

$$\cancel{p(\ddot{y} \mid \ddot{x}, D_{Train}) = \int p(\ddot{y} \mid \ddot{x}, f)\, p(f \mid D_{Train})\, df}$$

▪petal width

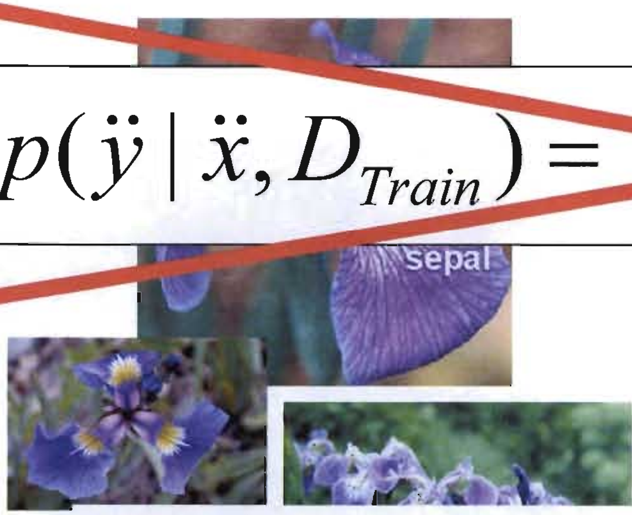


Iris Setosa

Iris Virginica

Iris Versicolor

USING A LEARNER



Info about

$$p(\ddot{y} \mid \ddot{x}, D_{Train}) = \int p(\ddot{y} \mid \ddot{x}, f)\, p(f \mid D_{Train})\, df$$

▪petal width

Iri

Can we do better?

Iris Virginica

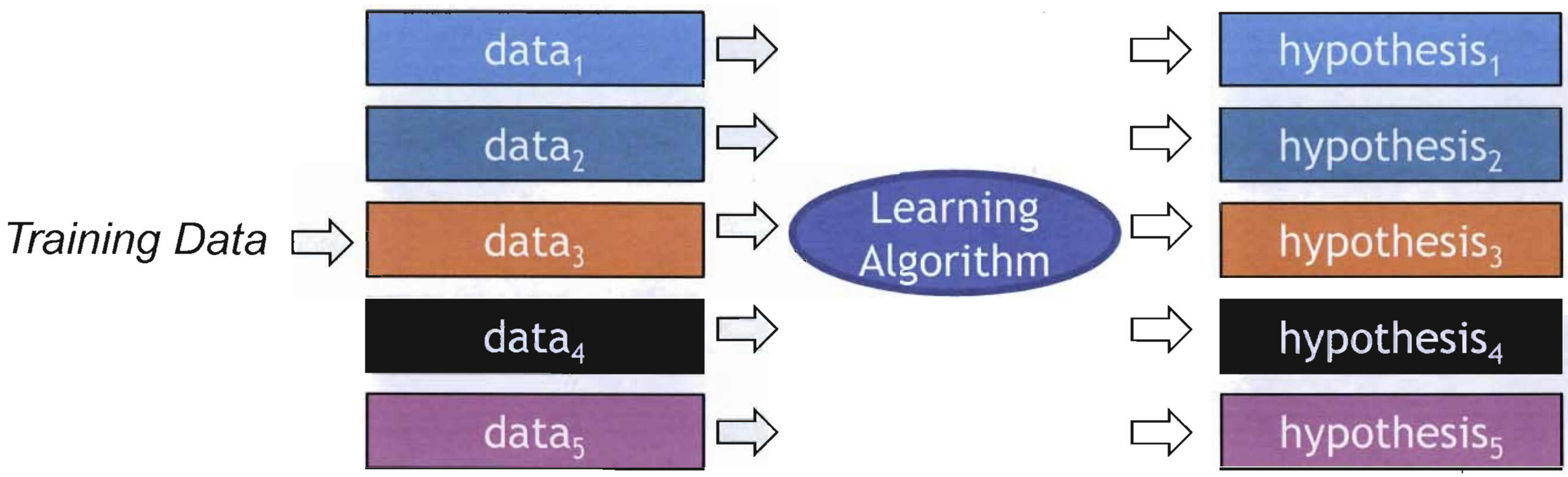
Iris Versicolor

ENSEMBLES

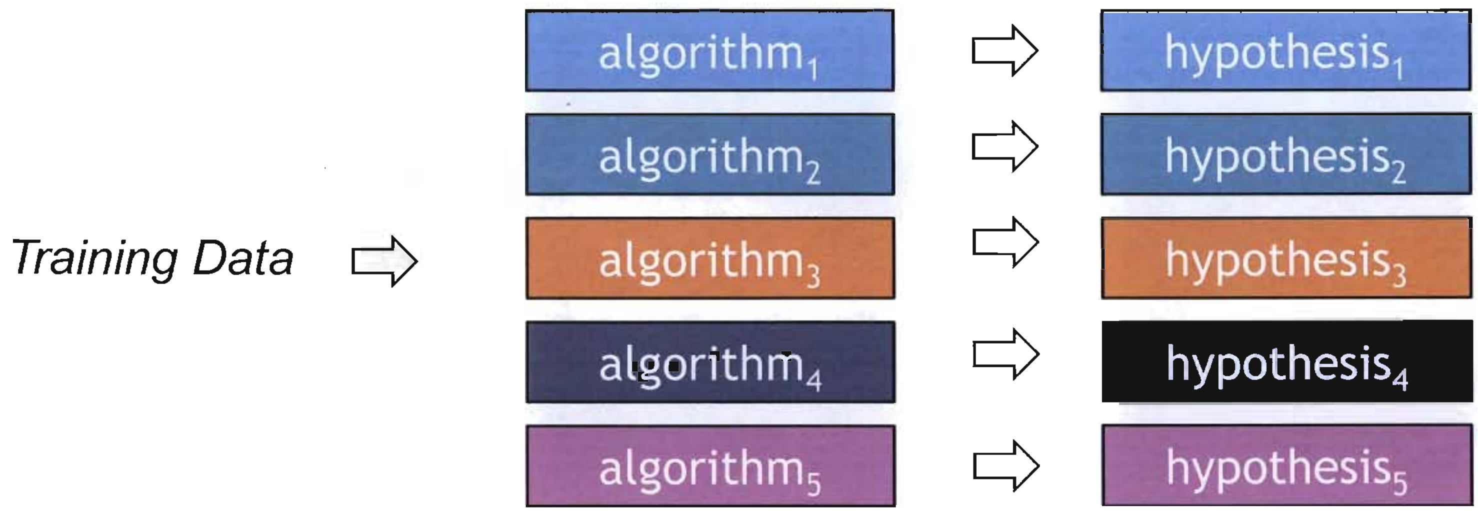
- Multiple learners

vs.

ENSEMBLE TRAINING:



ENSEMBLE TRAINING:



DEVELOPING MULTIPLE HYPOTHESES:

Iris Setosa

Info about
- sepal length
- sepal width
- petal length
- petal width

hypothesis$_1$

hypothesis$_2$

hypothesis$_3$

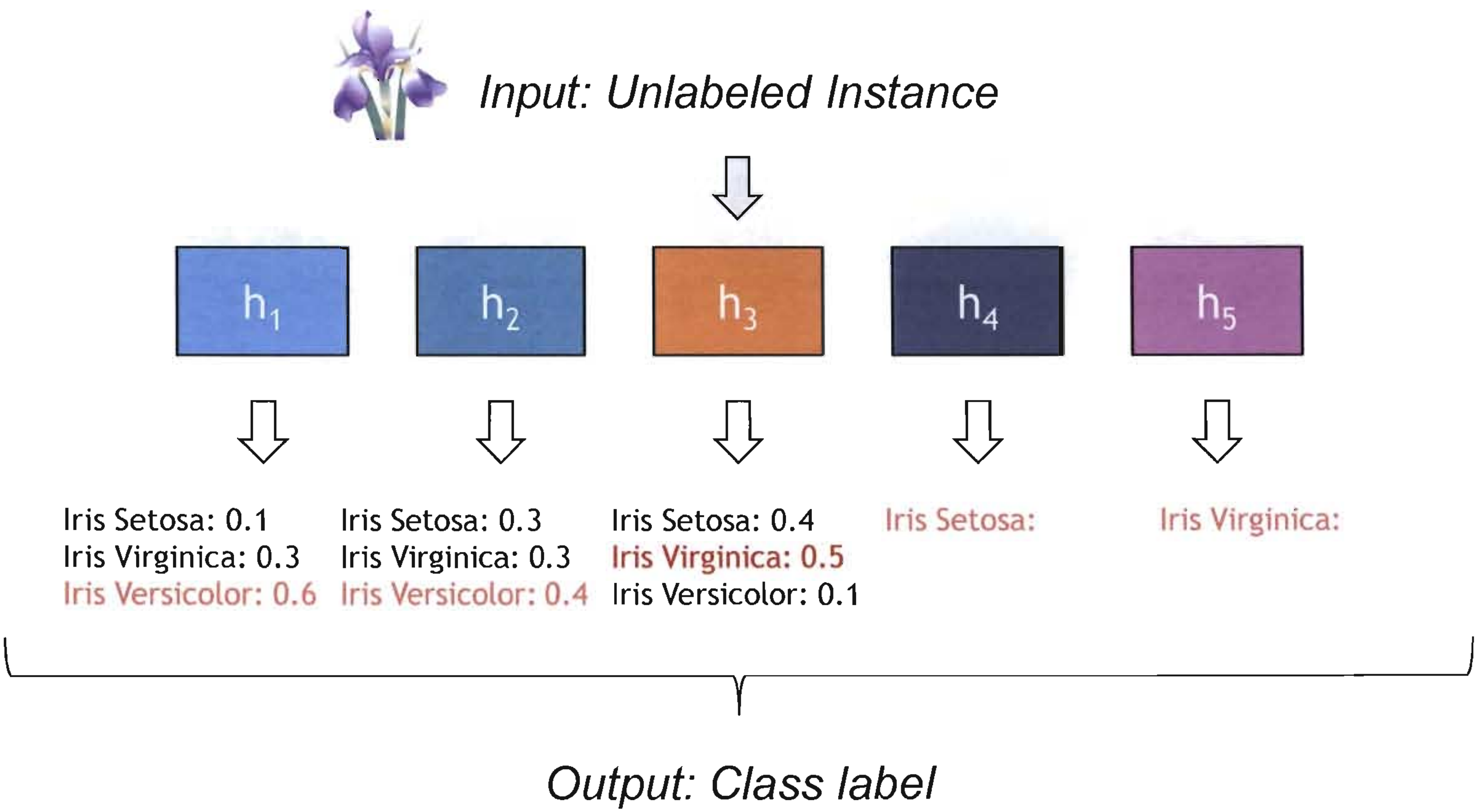
hypothesis$_4$

hypothesis$_5$

Iris Virginica

Iris Versicolor

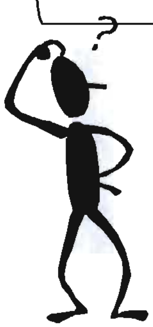
CLASSIFYING AN INSTANCE:



CLASSIFYING AN INSTANCE:

Input: Unlabeled Instance

h_1	h_2	h_3	h_4	h_5
Iris Setosa: 0.1 Iris Virginica: 0.3 Iris Versicolor: 0.6	Iris Setosa: 0.3 Iris Virginica: 0.3 Iris Versicolor: 0.4	Iris Setosa: 0.4 **Iris Virginica: 0.5** Iris Versicolor: 0.1	Iris Setosa:	Iris Virginica:

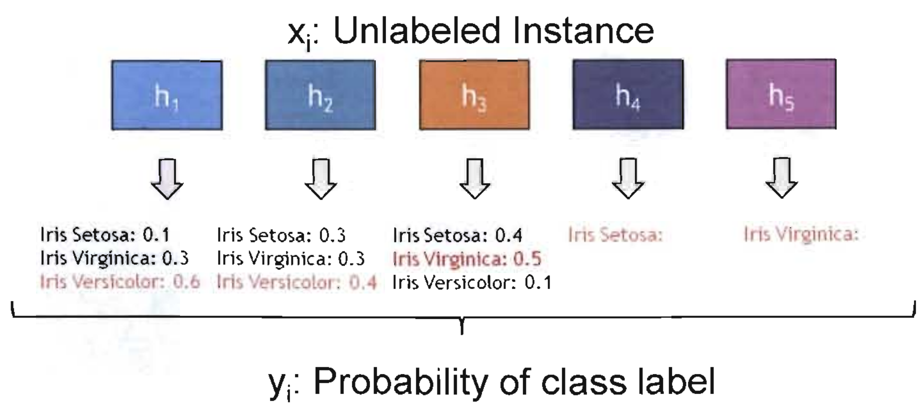


Output: Class label

POSSIBLE OPTIONS FOR COMBINING HYPOTHESES:

- Bagging: One hypothesis, One vote
- Boosting: Weight by predictive accuracy on the training set
- BAYESIAN MODEL AVERAGING (BMA): Weight by the formal probability that each hypothesis is correct given all the data

$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$



BAYESIAN TECHNIQUES

"Bayes is right, and everything else is wrong, or an approximation."

-James Carroll

Please compare your algorithm to Bayesian Model Averaging

- Reviewer for a conference where Kristine submitted her thesis research on ensemble learning

BMA IS THE "OPTIMAL" ENSEMBLE TECHNIQUE?

"Given the 'correct' model space and prior distribution, Bayesian model averaging is the optimal method for making predictions; in other words, no other approach can consistently achieve lower error rates than it does."

- Pedro Domingos

DOMINGOS' EXPERIMENTS

- Domingos decided to put this theory to the test.
- 2000 empirical study of ensemble methods:
  - J48
  - Bagging
  - BMA

DOMINGOS' EXPERIMENTS

	J48	Bagging	BMA
Annealing	93.50	94.90	94.40
Audiology	73.50	77.00	76.00
Breast cancer	68.80	70.30	62.90
Credit	85.70	87.20	82.20
Diabetes	74.90	75.80	72.50
Echocardio	66.50	70.30	65.70
Glass	65.90	77.10	70.60
Heart	77.90	82.80	76.90
Hepatitis	80.10	84.00	77.50
Horse colic	83.70	86.00	83.30
Iris	94.70	94.70	93.30
LED	59.00	61.00	60.00
Labor	80.30	91.00	87.70
Lenses	80.00	76.70	73.30
Liver	66.60	74.20	67.00
Lung cancer	55.00	45.00	55.80
Lymphogr.	80.30	76.30	81.00
Post-oper.	68.90	62.20	65.60
Pr. tumor	41.00	43.70	43.70
Promoters	81.70	86.60	82.90
Solar flare	71.20	69.40	70.30
Sonar	75.40	80.30	72.70
Soybean	100.00	98.00	98.00
Voting	95.60	96.80	95.40
Wine	88.80	93.30	88.70
Zoo	90.10	91.00	93.00
Average:	76.89	78.68	76.55

DOMINGOS' EXPERIMENTS

	J48	Bagging	BMA
Annealing	93.50	94.90	94.40
Audiology	73.50	77.00	76.00
Breast cancer	68.80	70.30	62.90
Credit	85.70	87.20	82.20
Diabetes	74.90	75.80	72.50
Echocardio	66.50	70.30	65.70
Glass	65.90	77.10	70.60
Heart	77.90	82.80	76.90
Hepatitis	80.10	84.00	77.50
Horse colic	83.70	86.00	83.30
Iris	94.70	94.70	93.30
LED	59.00	61.00	60.00
Labor	80.30	91.00	87.70
Lenses	80.00	76.70	73.30
Liver	66.60	74.20	67.00
Lung cancer			
Lymphogr.			
Post-oper.			
Pr. tumor			
Promoters	81.70	86.60	82.90
Solar flare	71.20	69.40	70.30
Sonar	75.40	80.30	72.70
Soybean	100.00	98.00	98.00
Voting	95.60	96.80	95.40
Wine	88.80	93.30	88.70
Zoo	90.10	91.00	93.00
Average	76.89	78.68	76.55

	J48	Bagging	BMA
Average:	76.89	78.68	76.55

DOMINGOS' EXPERIMENTS

	J48	Bagging	BMA
Annealing	93.50	94.90	94.40
Audiology	73.50	77.00	76.00
Breast cancer	68.80	70.30	62.90
Credit	85.70	87.20	82.20
Diabetes	74.90	75.80	72.50
Echocardio	66.50	70.30	65.70
Glass	65.90	77.10	70.60
Heart	77.90	82.80	76.90
Hepatitis	80.10	84.00	77.50
Horse colic	83.70	86.00	83.30
Iris	94.70	94.70	93.30
LED	59.00	61.00	60.00
Labor	80.30	91.00	87.70
Lenses	80.00	76.70	73.30
Liver	66.60	74.20	67.00
Lung cancer			
Lymphogr.			
Post-oper.			
Pr. tumor			
Promoters	81.70	86.60	82.90
Solar flare	71.20	69.40	70.30
Sonar	75.40	80.30	72.70
Soybean	100.00	98.00	98.00
Voting	95.60	96.80	95.40
Wine	88.80	93.30	88.70
Zoo	90.10	91.00	93.00
Average:	76.89	78.68	76.55

	J48	Bagging	BMA
Average:	76.89	78.68	76.55

DOMINGOS'S OBSERVATION:

- Bayesian Model Averaging gives too much weight to the maximum likelihood hypothesis

$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

$$p(h|D) \propto p(h)(1-\epsilon)^r(\epsilon)^{n-r}$$

Compare two classifiers:
One with 95% predictive accuracy and one with 94% predictive accuracy

$$(1 - \frac{5}{100})^{95}(\frac{5}{100})^5 = 2.39 * 10^{-9}$$
$$(1 - \frac{6}{100})^{94}(\frac{6}{100})^6 = 1.39 * 10^{-10}$$

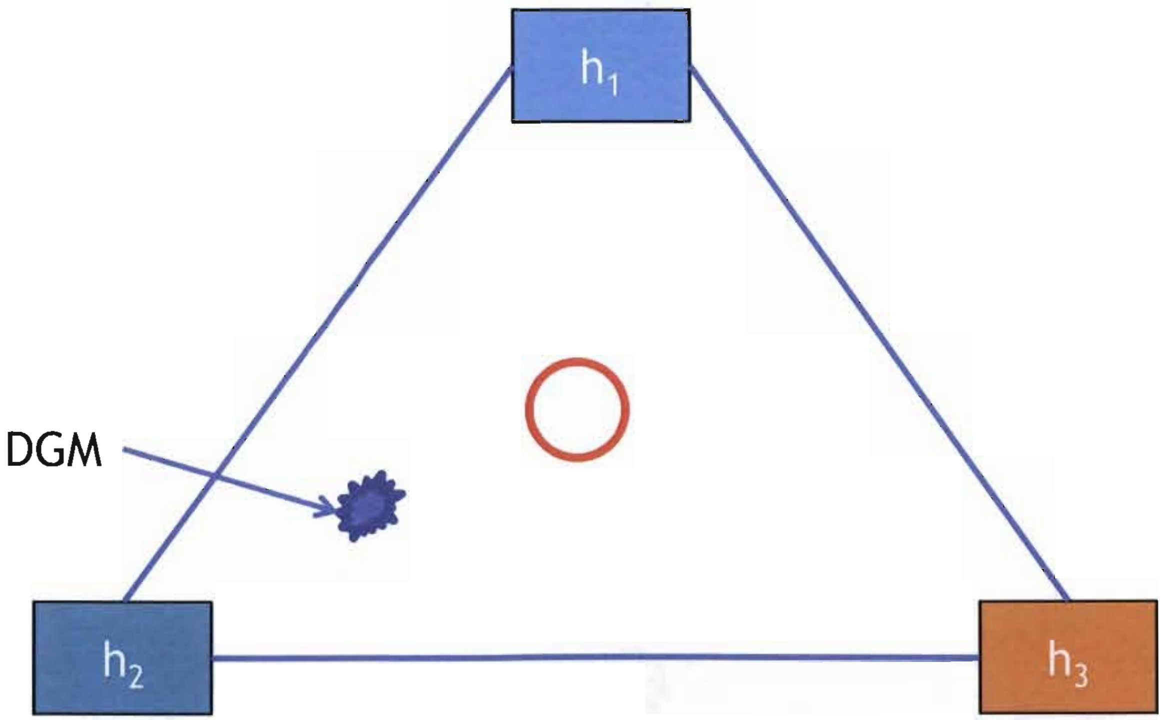
Bayesian Model Averaging weights the first classifier as
17 TIMES more likely!

CLARKE'S EXPERIMENTS

- 2003, comparison between BMA and stacking.
- Similar results to Domingos
  - BMA is vastly outperformed by stacking

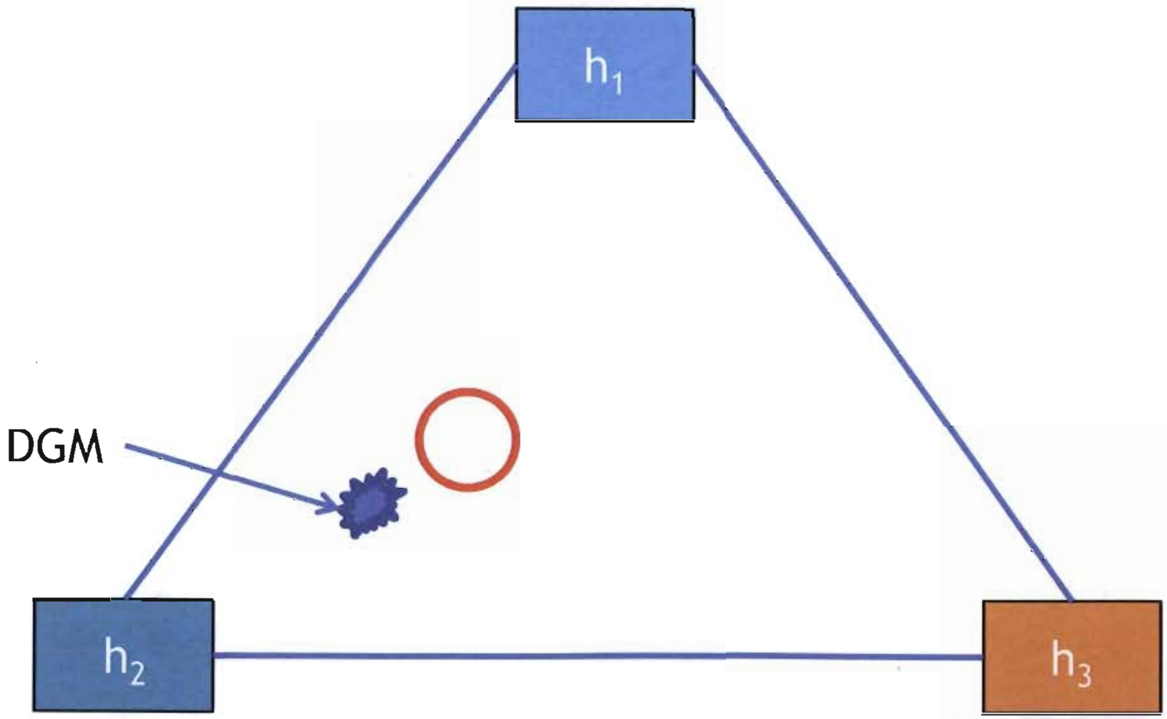
CLARKE'S OBSERVATION:

- BMA converges to model closest to the Data Generating Model (DGM) instead of converging to the combination closest to DGM!



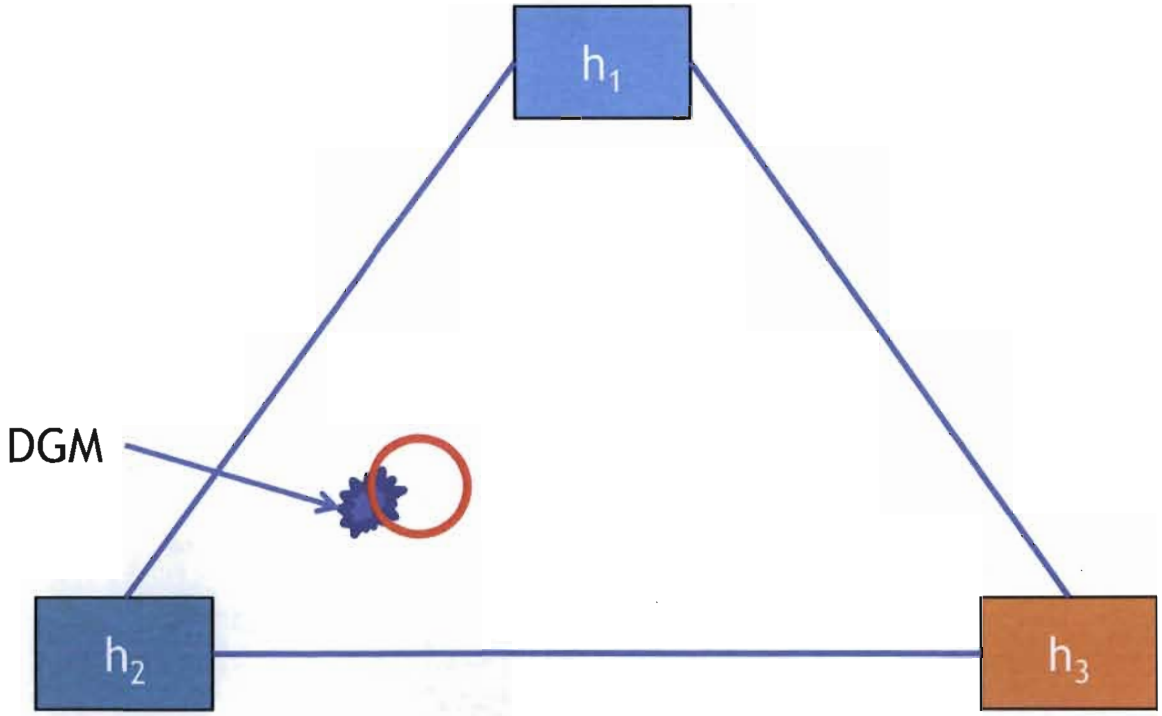
CLARKE'S OBSERVATION:

- BMA converges to model closest to the Data Generating Model (DGM) instead of converging to the combination closest to DGM!



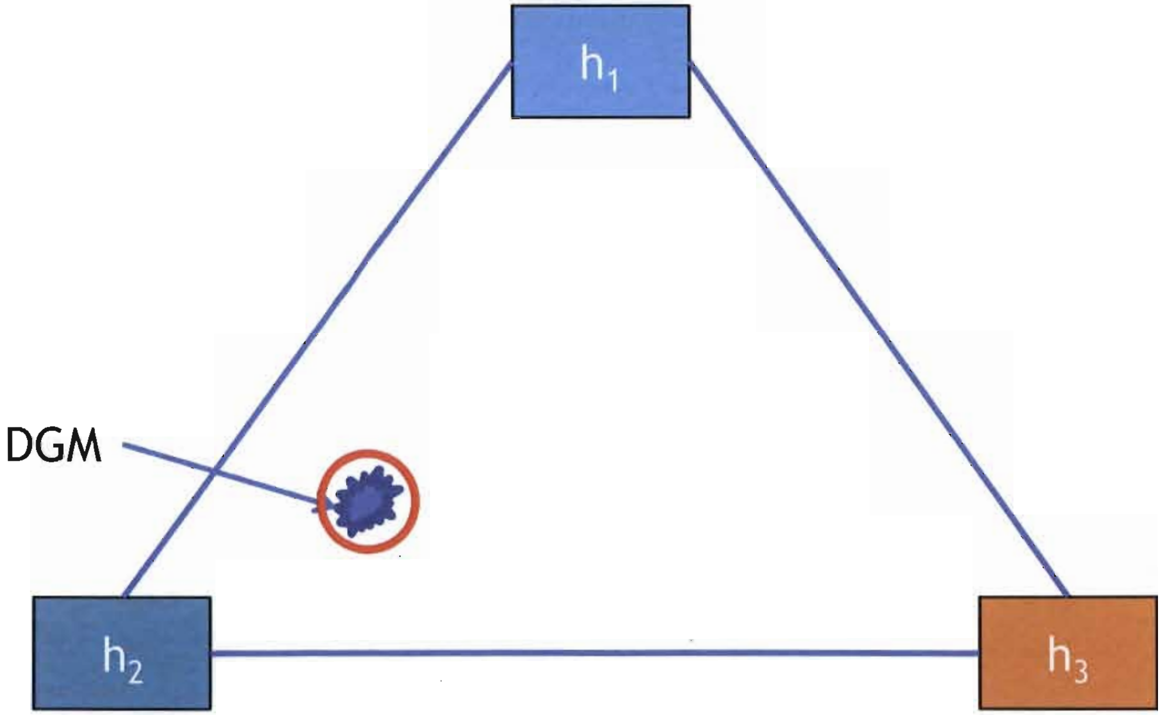
CLARKE'S OBSERVATION:

- BMA converges to model closest to the Data Generating Model (DGM) instead of converging to the combination closest to DGM!



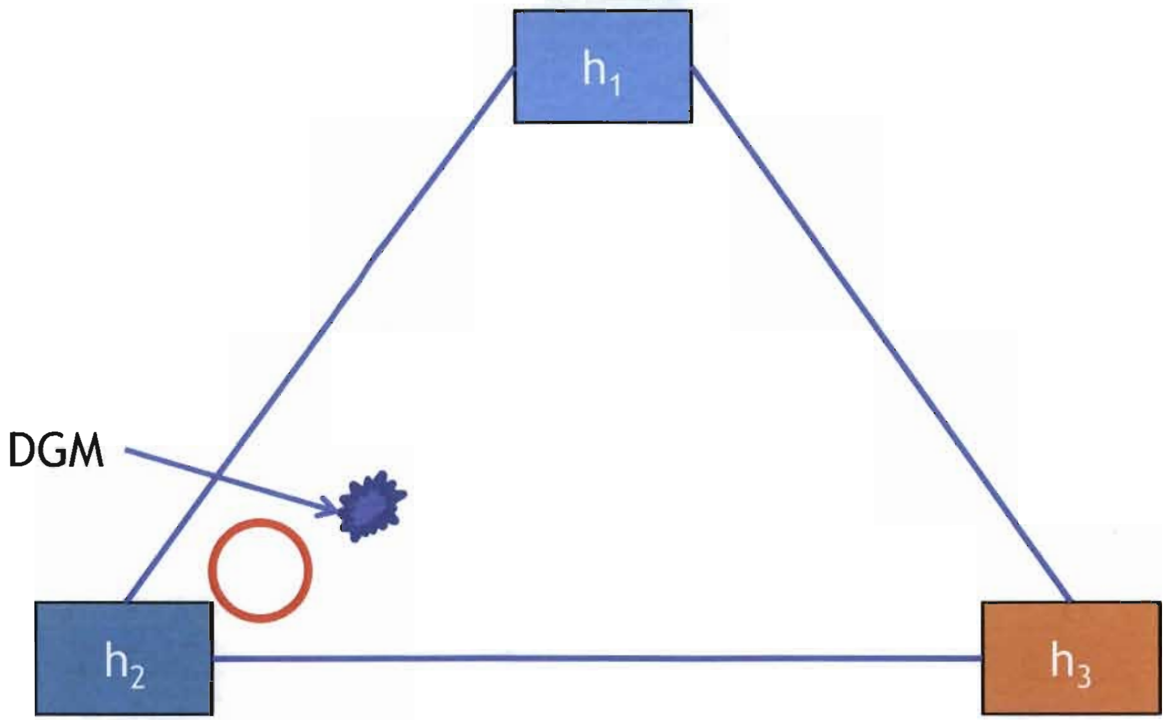
CLARKE'S OBSERVATION:

- BMA converges to model closest to the Data Generating Model (DGM) instead of converging to the combination closest to DGM!



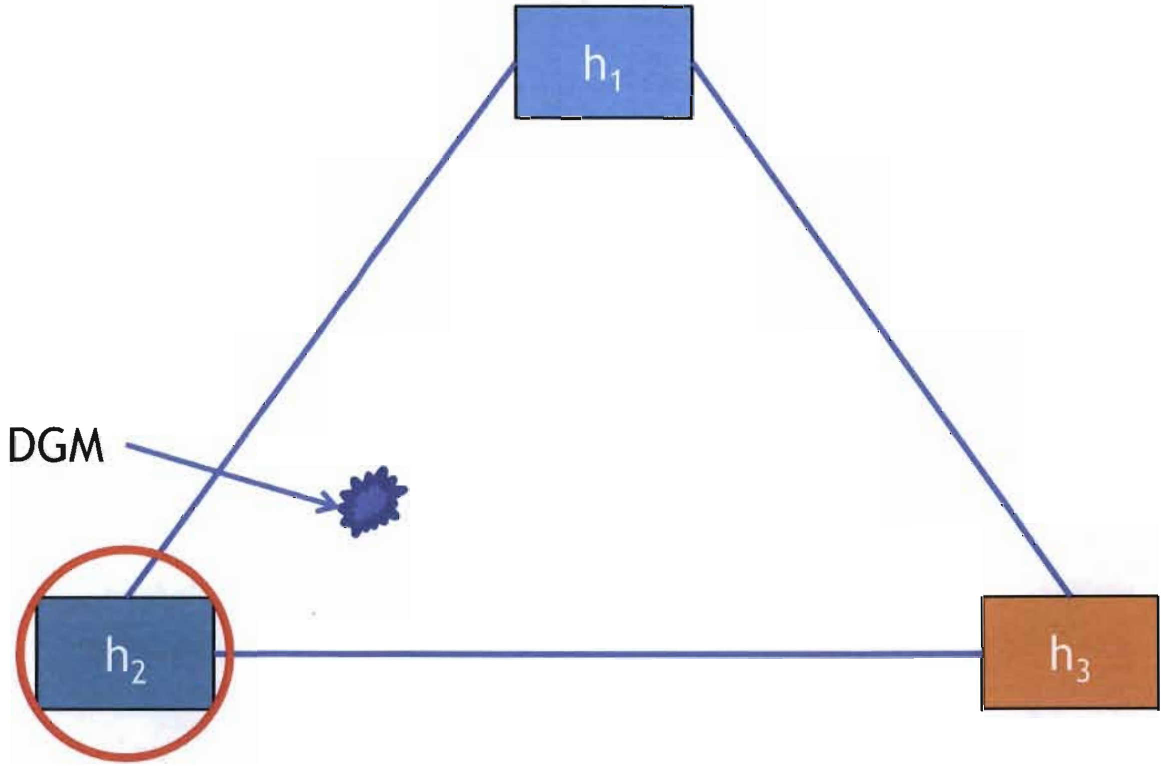
CLARKE'S OBSERVATION:

- BMA converges to model closest to the Data Generating Model (DGM) instead of converging to the combination closest to DGM!



CLARKE'S OBSERVATION:

- BMA converges to model closest to the Data Generating Model (DGM) instead of converging to the combination closest to DGM!



SO WHAT'S WRONG???

- Bayesian techniques are theoretically optimal if all underlying assumptions are correct
- Which one of our underlying assumptions is flawed?

WHY DO ENSEMBLES WORK?

MINKA'S COMMENTARY

"...the only flaw with BMA is the belief that it is an algorithm for model combination"

$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

MINKA'S COMMENTARY

"...the only flaw with BMA is the belief that it is an algorithm for model combination"

$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

But BMA does return a combination!

MINKA'S COMMENTARY

"...the only flaw with BMA is the belief that it is an algorithm for model combination"

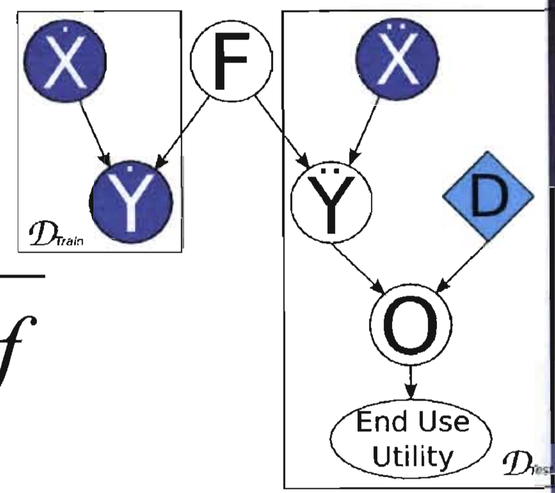
$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

BMA's combination is determined by the probability that each model is correct (the DGM).

Underlying assumption:

One, and only one model is right, and the rest are wrong.

THE MATHEMATICS OF LEARNING



- Learning about F:

$$p(f \mid x, y) = \frac{p(y \mid x, f)\, p(f)}{\int p(y \mid x, f)\, p(f)\, df}$$

- Classification or Regression:

$$p(\ddot{y} \mid \ddot{x}, D_{Train}) = \int p(\ddot{y} \mid \ddot{x}, f)\, p(f \mid D_{Train})\, df$$

- BMA

$$p(\ddot{y} \mid \ddot{x}, D_{Train}, H) = \sum_{h \in H} p(\ddot{y} \mid \ddot{x}, h)\, p(h \mid D_{Train})$$

MINKA'S COMMENTARY

BMA optimally integrates out uncertainty about which model is right, assuming that one and only one is right.

$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

WHY DO ENSEMBLES WORK?

- Theory 1: Ensembles account for uncertainty about which model is correct
  - BMA does this optimally

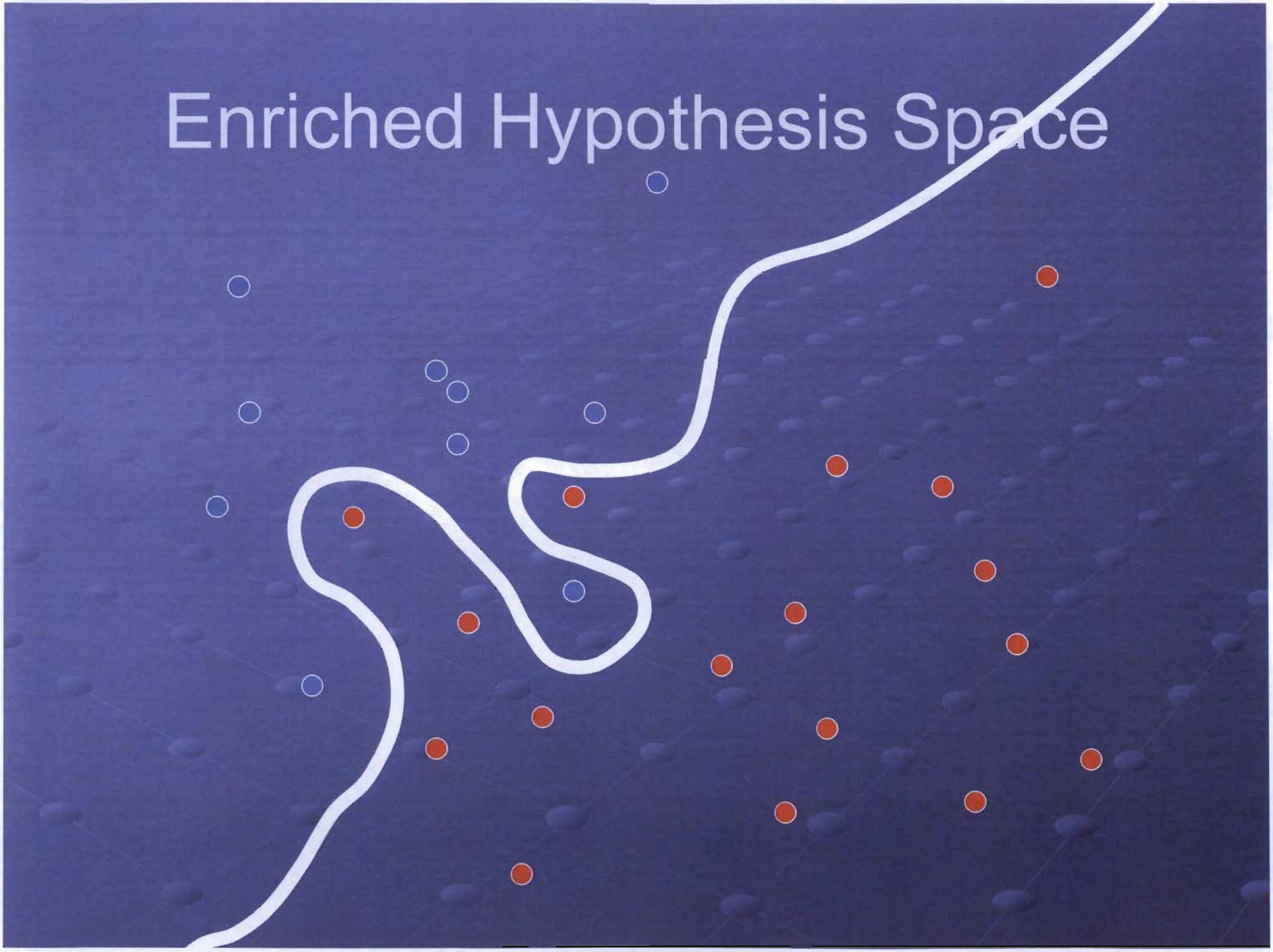
THE PROBLEM

- If Theory 1 is right, then BMA is the optimal way to do ensembles, and will out perform other ensemble techniques.
- Empirical results indicate that there must be more to it.

WHY DO ENSEMBLES WORK?

- Theory 1: Ensembles account for uncertainty about which model is correct
  - BMA does this optimally
- Theory 2: Ensembles improve the representational bias of the learner
  - Ensembles enrich the hypothesis space of the learner so that together they can represent hypotheses that no single member could represent alone

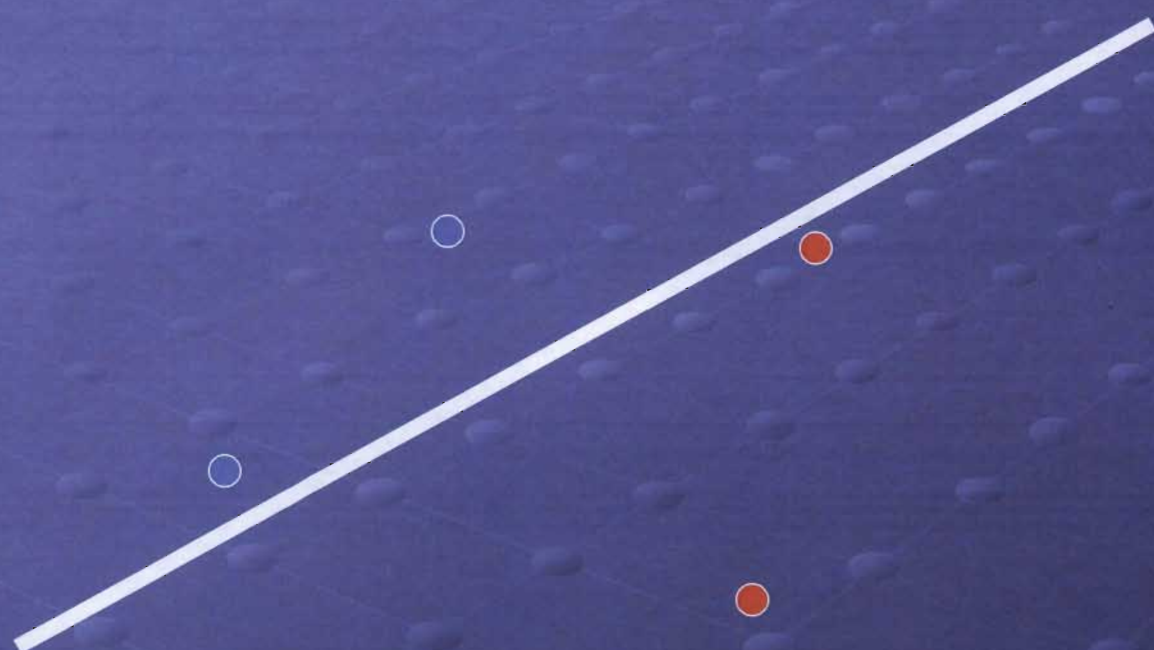
Enriched Hypothesis Space

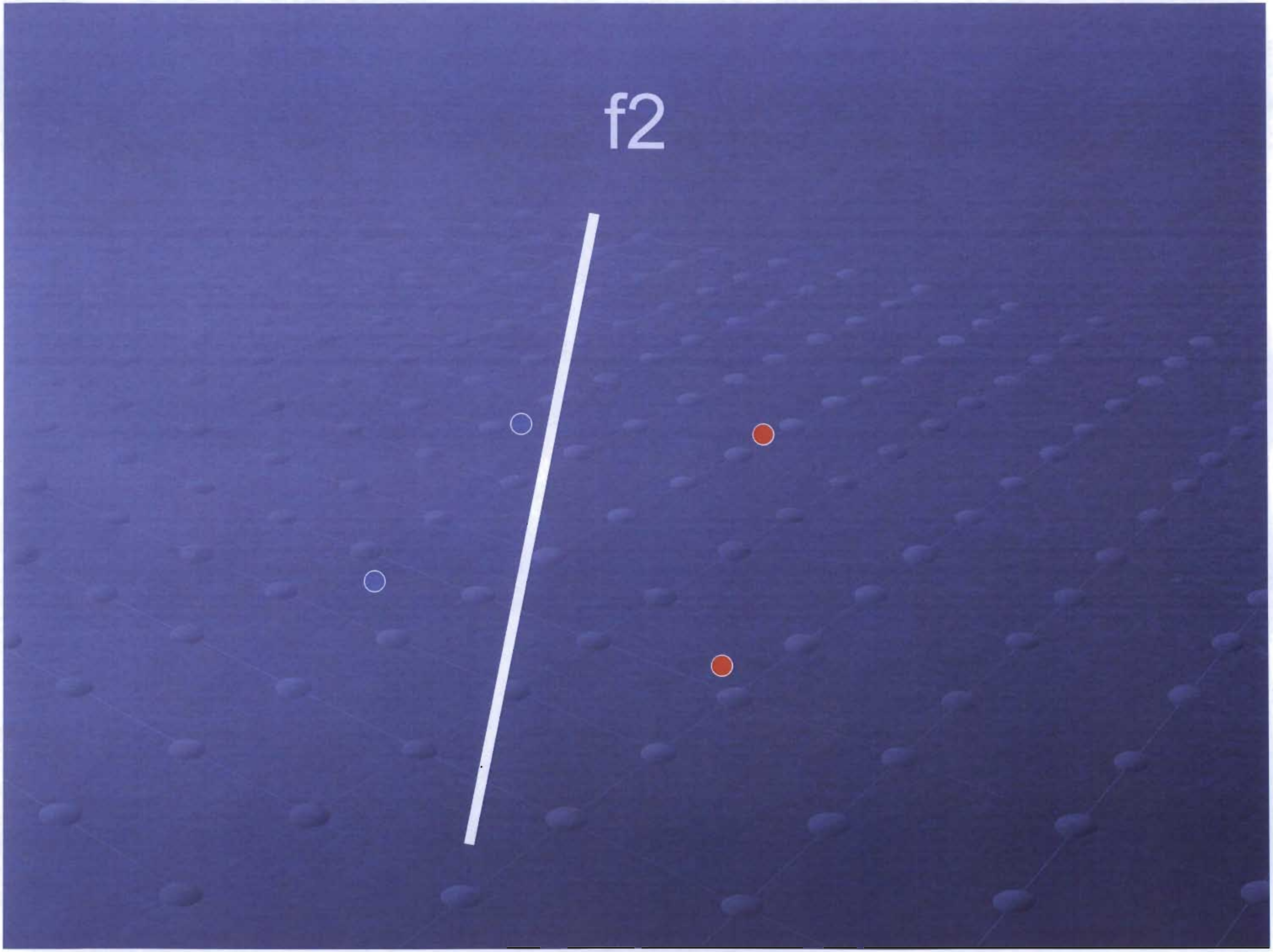


WHY DO ENSEMBLES WORK?

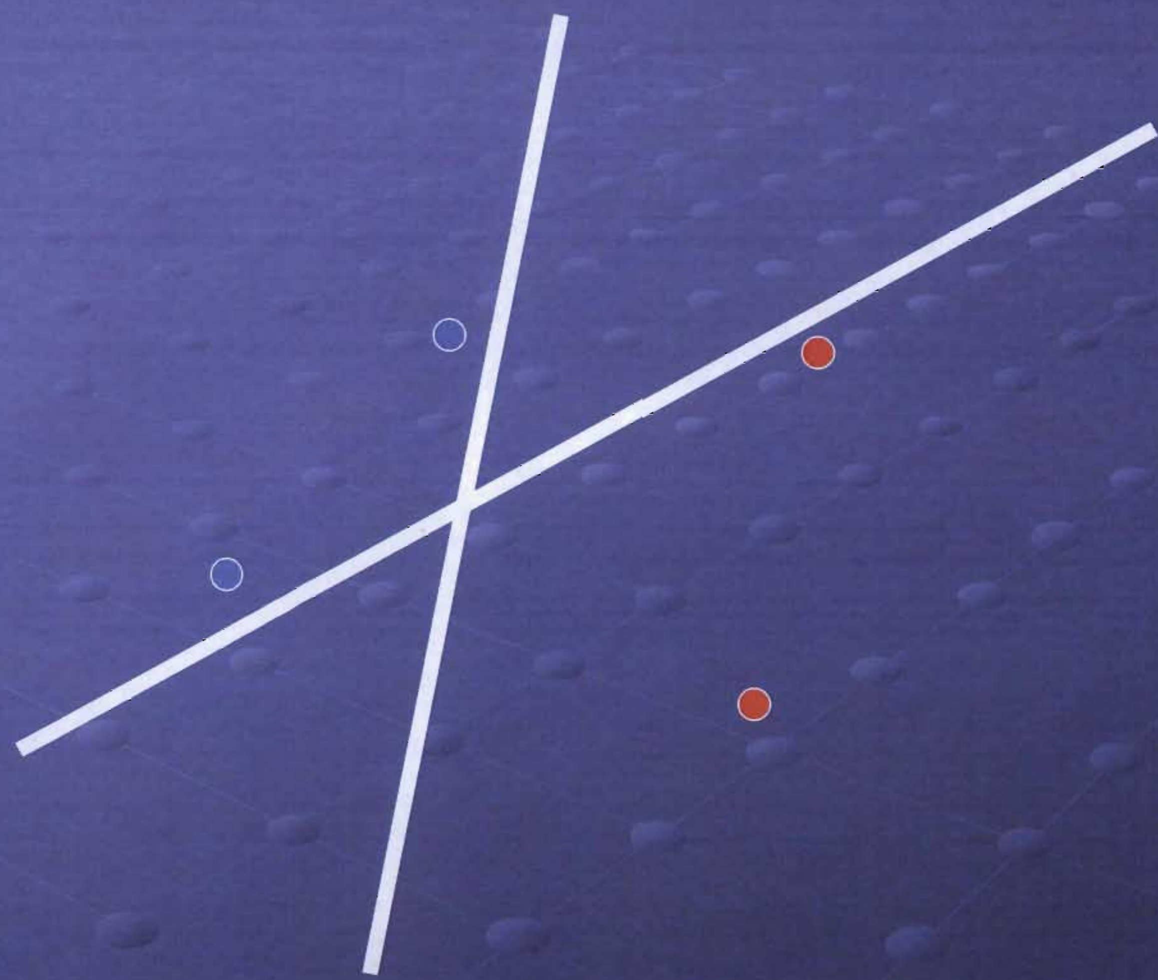
- <u>Theory 1:</u> Ensembles account for uncertainty about which model is correct
  - BMA does this optimally
- <u>Theory 2:</u> Ensembles improve the representational bias of the learner
  - Ensembles enrich the hypothesis space of the learner so that together they can represent hypotheses that no single member could represent alone
- <u>Theory 3:</u> Ensembles improve the preferential bias of the learner
  - They act as a sort of regularization technique that reduces overfit

f1

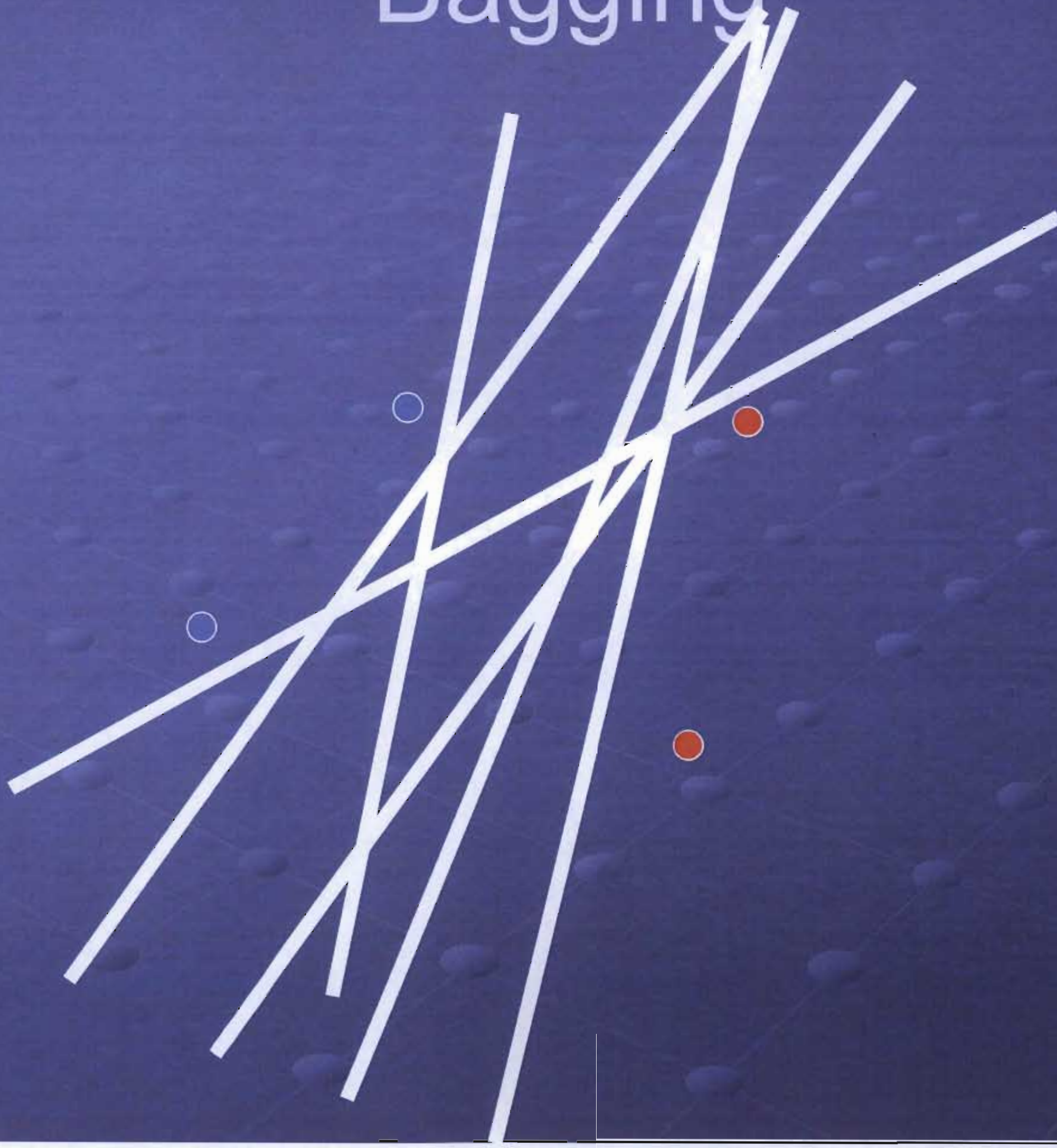



f2

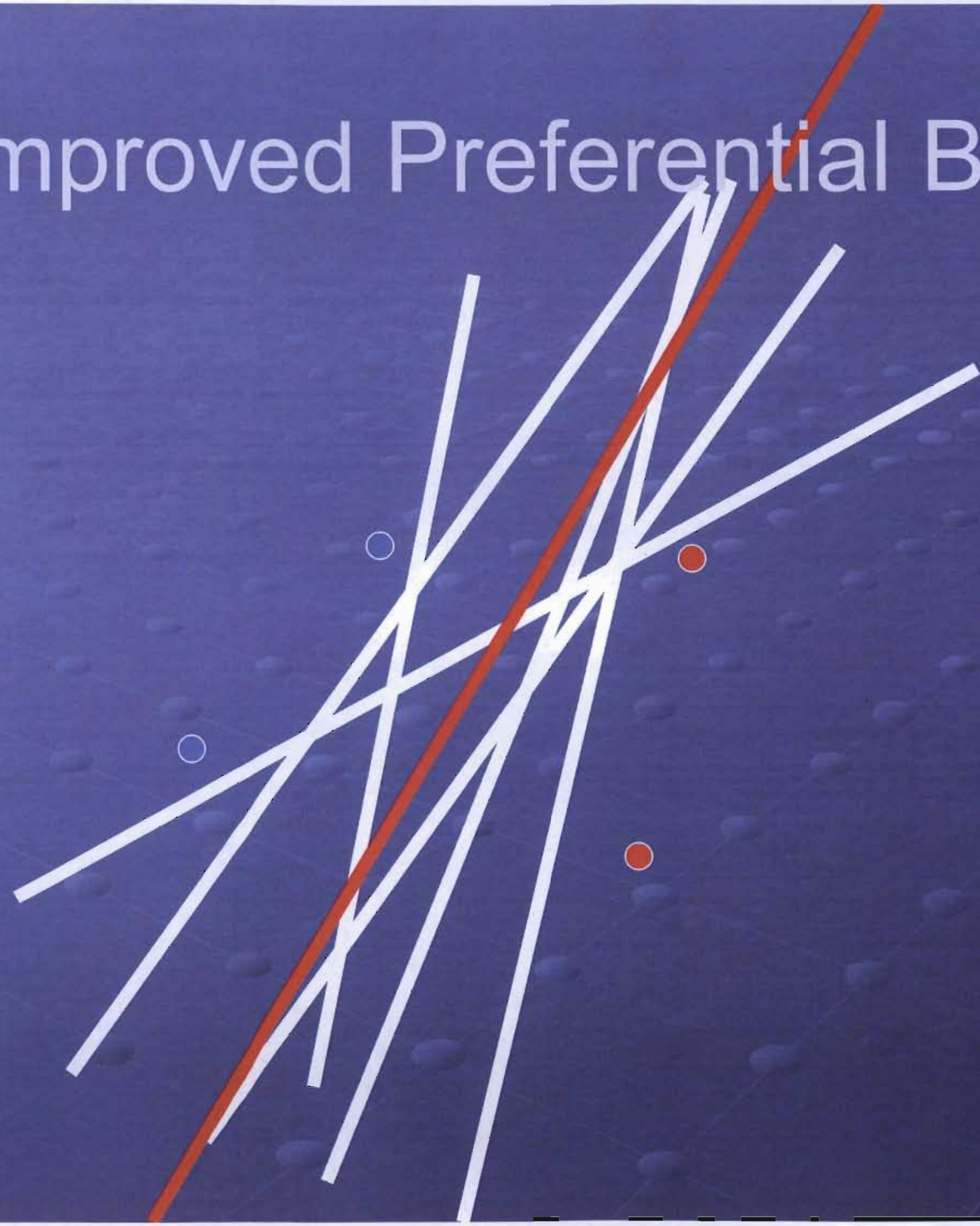
Bagging



Bagging



Improved Preferential Bias



Approximate Probability Output

ADVANTAGES OF MODEL COMBINATIONS:

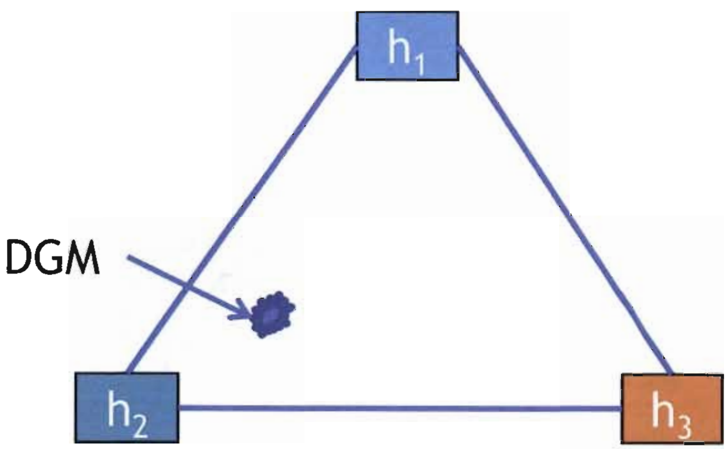
- Theory 2: Allowing ensembles of models enriches the hypothesis space

vs.

- Theory 3: Ensembles may have a more general bias that is less prone to overfit

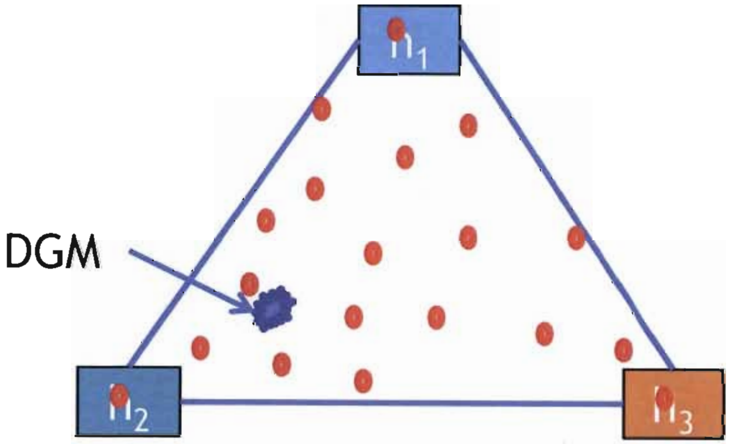
vs.

HOW TO FIX BAYESIAN MODEL AVERAGING:



$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

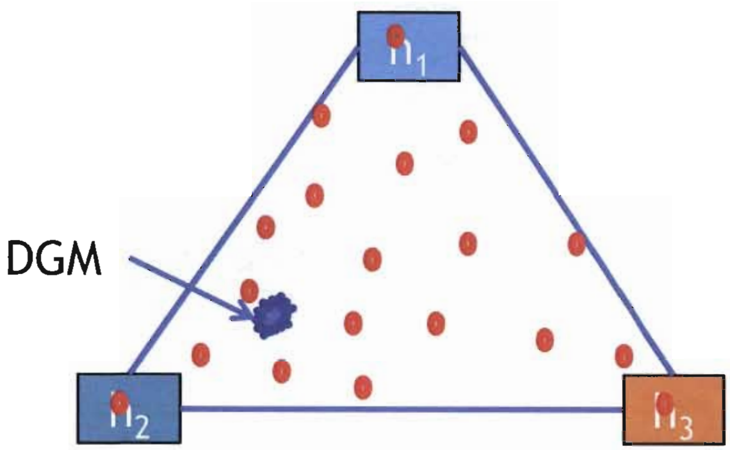
HOW TO FIX
BAYESIAN MODEL AVERAGING:



$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h)p(h|D)$$

HOW TO FIX BAYESIAN MODEL AVERAGING:

ITERATE OVER MODEL COMBINATIONS



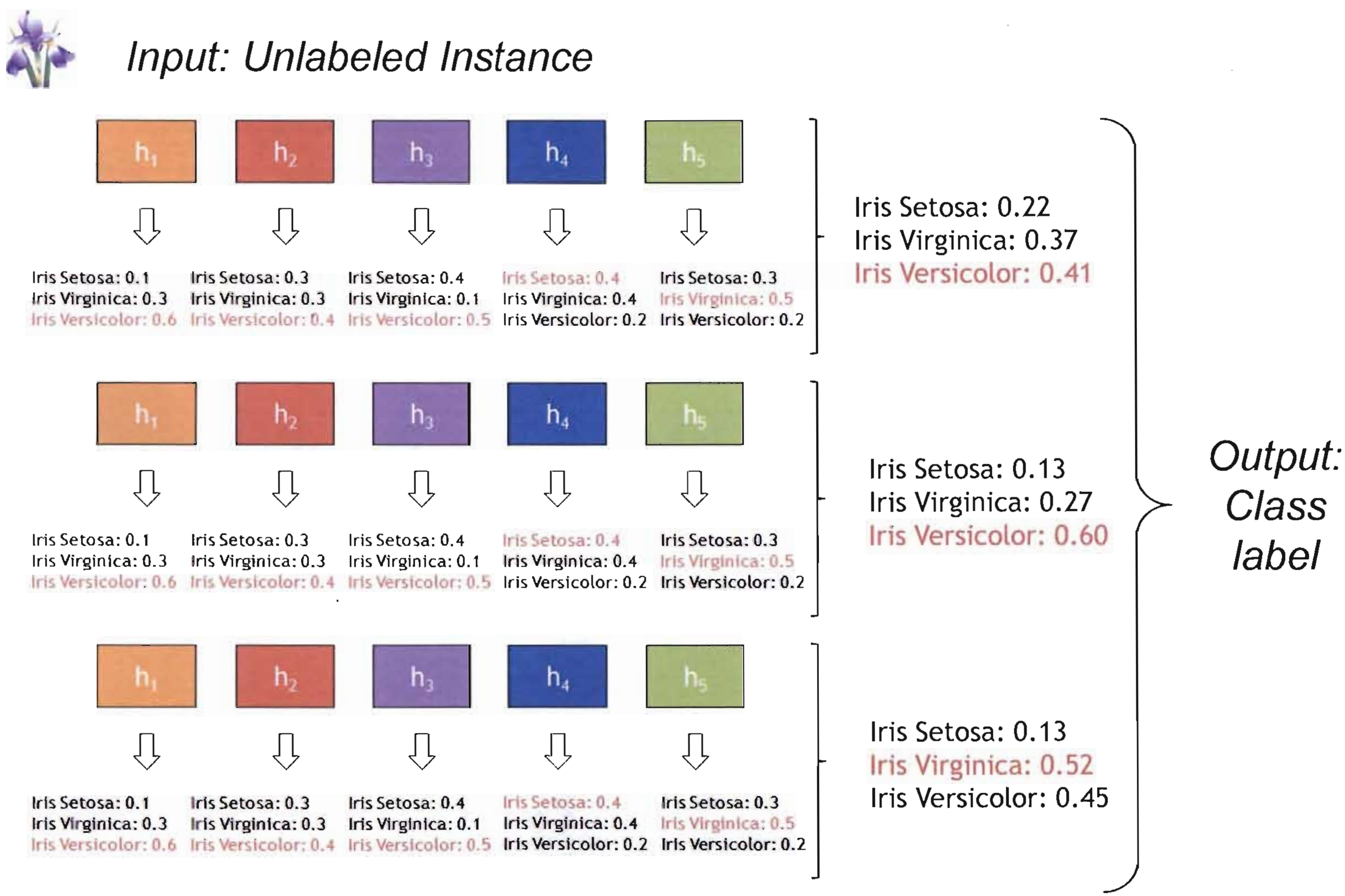
$$p(y_i|x_i, D, H) = \sum_{h \in H} p(y_i|x_i, h) p(h|D)$$

$$\Downarrow$$

$$p(y_i|x_i, D, H, E) = \sum_{e \in E} p(y_i|x_i, H, e) p(e|D)$$

where E is a set of model combinations instead of individual models

BAYESIAN MODEL COMBINATION



BAYESIAN MODEL COMBINATION

Input: Unlabeled Instance

Generate diverse model combinations

h1	h2	h3	h4	h5
Iris Setosa: 0.1 Iris Virginica: 0.3 Iris Versicolor: 0.6	Iris Setosa: 0.3 Iris Virginica: 0.3 Iris Versicolor: 0.4	Iris Setosa: 0.4 Iris Virginica: 0.1 Iris Versicolor: 0.5	Iris Setosa: 0.4 Iris Virginica: 0.4 Iris Versicolor: 0.2	Iris Setosa: 0.3 Iris Virginica: 0.5 Iris Versicolor: 0.2

Iris Setosa: 0.22
Iris Virginica: 0.37
Iris Versicolor: 0.41

h1	h2	h3	h4	h5
Iris Setosa: 0.1 Iris Virginica: 0.3 Iris Versicolor: 0.6	Iris Setosa: 0.3 Iris Virginica: 0.3 Iris Versicolor: 0.4	Iris Setosa: 0.4 Iris Virginica: 0.1 Iris Versicolor: 0.5	Iris Setosa: 0.4 Iris Virginica: 0.4 Iris Versicolor: 0.2	Iris Setosa: 0.3 Iris Virginica: 0.5 Iris Versicolor: 0.2

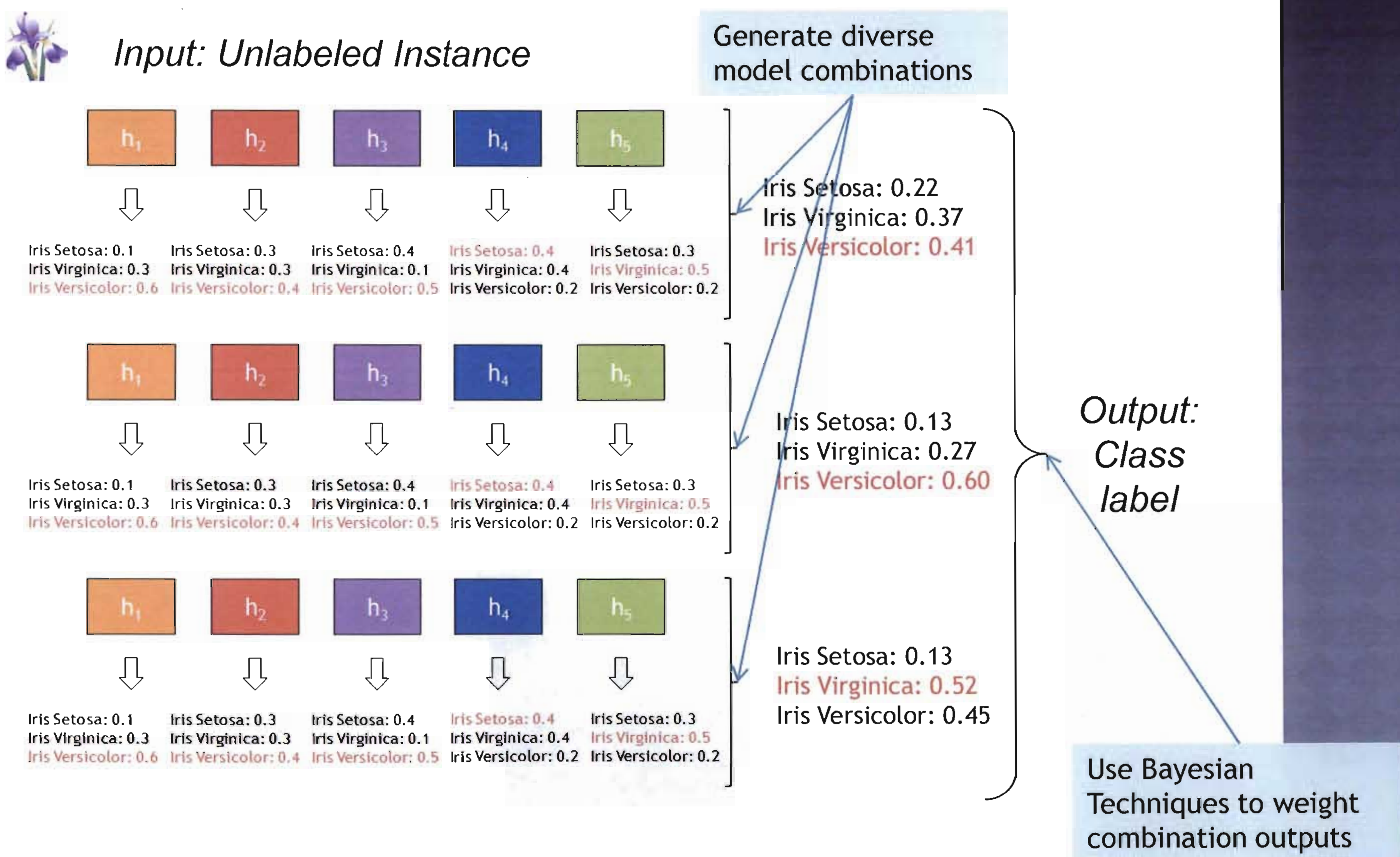
Iris Setosa: 0.13
Iris Virginica: 0.27
Iris Versicolor: 0.60

h1	h2	h3	h4	h5
Iris Setosa: 0.1 Iris Virginica: 0.3 Iris Versicolor: 0.6	Iris Setosa: 0.3 Iris Virginica: 0.3 Iris Versicolor: 0.4	Iris Setosa: 0.4 Iris Virginica: 0.1 Iris Versicolor: 0.5	Iris Setosa: 0.4 Iris Virginica: 0.4 Iris Versicolor: 0.2	Iris Setosa: 0.3 Iris Virginica: 0.5 Iris Versicolor: 0.2

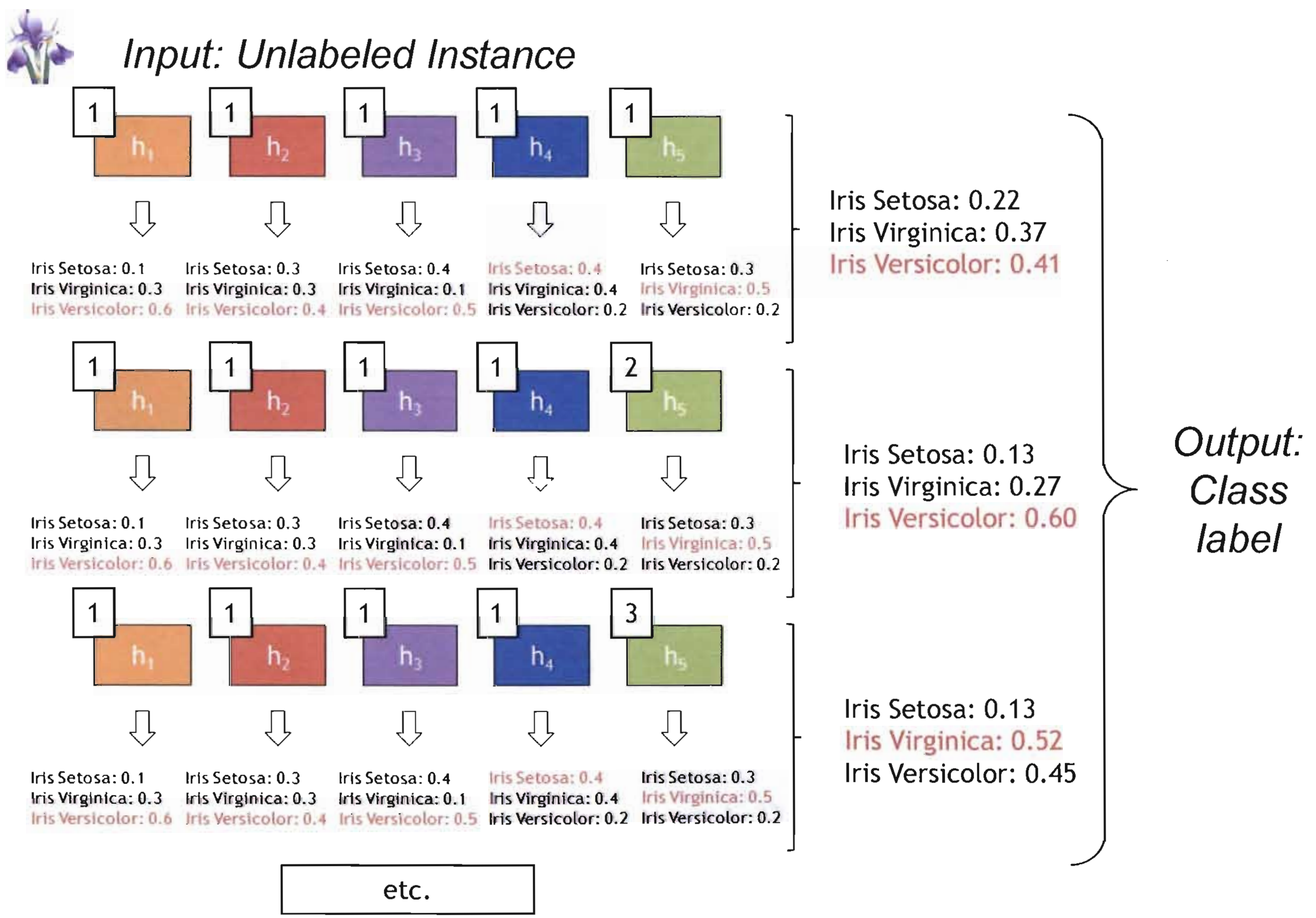
Iris Setosa: 0.13
Iris Virginica: 0.52
Iris Versicolor: 0.45

Output: Class label

BAYESIAN MODEL COMBINATION



WEIGHTING STRATEGY #1: LINEAR WEIGHT ASSIGNMENTS



RESULTS

	J48	Bagging	Boosting	BMA	BMC-Inc
anneal	98.44	98.22	99.55	98.22	98.89
audiology	77.88	76.55	84.96	76.11	82.3
autos	81.46	69.76	83.9	70.24	84.88
balance-scale	76.64	82.88	78.88	82.88	81.92
bupa	68.7	71.01	71.59	70.43	71.88
cancer-wisc.	93.85	95.14	95.71	95.28	95.14
cancer-yugo.	75.52	67.83	69.58	68.18	73.08
car	92.36	92.19	96.12	92.01	93.75
cmc	52.14	53.63	50.78	41.96	52.95
credit-a	86.09	85.07	84.2	84.93	85.07
credit-g	70.5	74.4	69.6	74.3	73.1
dermatology	93.99	92.08	95.63	92.08	95.36
diabetes	73.83	74.61	72.4	74.61	74.35
echo	97.3	97.3	95.95	97.3	97.3
ecoli-c	84.23	83.04	81.25	82.74	84.52
glass	66.82	69.63	74.3	68.69	70.09
haberman	71.9	73.2	72.55	73.2	74.51
heart-cleveland	77.56	82.18	82.18	82.18	79.87
heart-h	80.95	78.57	78.57	78.57	79.59
heart-statlog	76.67	79.26	80.37	78.52	80
hepatitis	83.87	84.52	85.81	83.87	83.87
horse-colic	85.33	85.33	83.42	85.05	86.14
hypothyroid	99.58	99.55	99.58	99.55	99.6
ionosphere	91.45	90.88	93.16	90.6	93.45
iris	96	94	93.33	94	95.33
kr-vs-kp	99.44	99.12	99.5	99.12	99.44
labor	73.68	85.96	89.47	87.72	84.21
led	100	100	100	100	100
lenses	83.33	66.67	70.83	58.33	79.17
letter	100	100	100	100	100
liver-disorders	68.7	71.01	71.59	70.43	71.88
lungcancer	50	50	53.12	46.88	56.25
lymph	77.03	78.38	81.08	79.05	80.41
monks	96.53	99.54	100	96.99	100
page-blocks	96.88	97.24	97.02	97.26	97.24
postop	70	71.11	56.67	71.11	67.78
primary-tumor	39.82	45.13	40.12	45.13	41.3
promoters	81.13	83.96	85.85	85.85	81.13
segment	96.93	96.97	98.48	96.88	97.45
sick	98.81	98.49	99.18	98.46	98.97
solar-flare	97.83	97.83	96.59	97.83	97.83
sonar	71.15	77.4	77.88	77.4	74.52
soybean	91.51	86.82	92.83	86.38	93.12
spect	78.28	81.65	80.15	82.02	79.03
tic-tac-toe	85.07	92.07	96.35	91.65	93.53
vehicle	72.46	72.7	76.24	72.81	76.48
vote	94.79	94.58	95.66	94.58	95.44
wine	93.82	94.94	96.63	93.26	95.51
yeast	56	60.04	56.4	31.2	60.51
zoo	92.08	87.13	96.04	86.14	93.07
average:	82.37	82.79	83.62	81.64	83.93

RESULTS

	J48	Bagging	Boosting	BMA	BMC-Inc
anneal	98.44	98.22	99.55	98.22	98.89
audiology	77.88	76.55	84.96	76.11	82.3
autos	81.46	69.76	83.9	70.24	84.88
balance-scale	76.64	82.88	78.88	82.88	81.92
bupa	68.7	71.01	71.59	70.43	71.88
cancer-wisc.	93.85	95.14	95.71	95.28	95.14
cancer-yugo.	75.52	67.83	69.58	68.18	73.08
car	92.36	92.19	96.12	92.01	93.75
cmc	52.14	53.63	50.78	41.96	52.95
credit-a	86.09	85.07	84.2	84.93	85.07
credit-g	70.5	74.4	69.6	74.3	73.1
dermatology	93.99	92.08	95.63	92.08	95.36
diabetes	73.83	74.61	72.4	74.61	74.35
echo	97.3	97.3	95.95	97.3	97.3
ecoli-c	84.23	83.04	81.25	82.74	84.52
glass	66.82	69.63	74.3	68.69	70.09
haberman	71.9	73.2	72.55	73.2	74.51
heart-cleveland	77.56	82.18	82.18	82.18	79.87
heart-h	80.95	78.57	78.57	78.57	79.59
heart-statlog	76.67	79.26	80.37	78.52	80
hepatitis	83.87	84.52	85.81	83.87	83.87
					86.14
					99.6
					93.45
					95.33
					99.44
					84.21
					100
					79.17
					100
liver-disorders	68.7	71.01	71.59	70.43	71.88
lungcancer	50	50	53.12	46.88	56.25
lymph	77.03	78.38	81.08	79.05	80.41
monks	96.53	99.54	100	96.99	100
page-blocks	96.88	97.24	97.02	97.26	97.24
postop	70	71.11	56.67	71.11	67.78
primary-tumor	39.82	45.13	40.12	45.13	41.3
promoters	81.13	83.96	85.85	85.85	81.13
segment	96.93	96.97	98.48	96.88	97.45
sick	98.81	98.49	99.18	98.46	98.97
solar-flare	97.83	97.83	96.59	97.83	97.83
sonar	71.15	77.4	77.88	77.4	74.52
soybean	91.51	86.82	92.83	86.38	93.12
spect	78.28	81.65	80.15	82.02	79.03
tic-tac-toe	85.07	92.07	96.35	91.65	93.53
vehicle	72.46	72.7	76.24	72.81	76.48
vote	94.79	94.58	95.66	94.58	95.44
wine	93.82	94.94	96.63	93.26	95.51
yeast	56	60.04	56.4	31.2	60.51
zoo	92.08	87.13	96.04	86.14	93.07
average:	82.37	82.79	83.62	81.64	83.93

	J48	Bagging	Boosting	BMA	BMC Linear
Average:	82.37	82.79	83.62	81.64	83.93

RESULTS

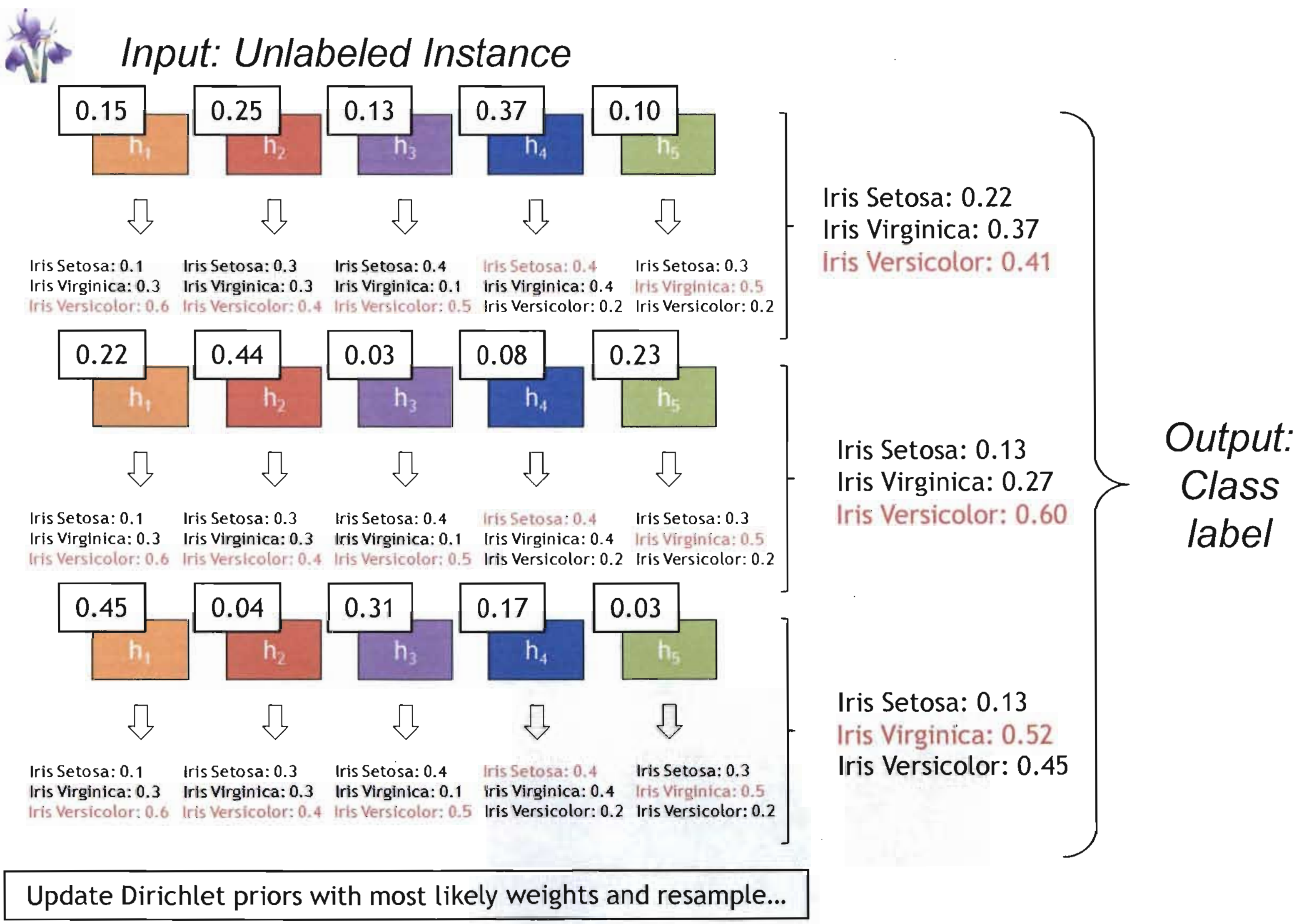
	J48	Bagging	Boosting	BMA	BMC-Inc
anneal	98.44	98.22	99.55	98.22	98.89
audiology	77.88	76.55	84.96	76.11	82.3
autos	81.46	69.76	83.9	70.24	84.88
balance-scale	76.64	82.88	78.88	82.88	81.92
bupa	68.7	71.01	71.59	70.43	71.88
cancer-wisc.	93.85	95.14	95.71	95.28	95.14
cancer-yugo.	75.52	67.83	69.58	68.18	73.08
car	92.36	92.19	96.12	92.01	93.75
cmc	52.14	53.63	50.78	41.96	52.95
credit-a	86.09	85.07	84.2	84.93	85.07
credit-g	70.5	74.4	69.6	74.3	73.1
dermatology	93.99	92.08	95.63	92.08	95.36
diabetes	73.83	74.61	72.4	74.61	74.35
echo	97.3	97.3	95.95	97.3	97.3
ecoli-c	84.23	83.04	81.25	82.74	84.52
glass	66.82	69.63	74.3	68.69	70.09
haberman	71.9	73.2	72.55	73.2	74.51
heart-cleveland	77.56	82.18	82.18	82.18	79.87
heart-h	80.95	78.57	78.57	78.57	79.59
heart-statlog	76.67	79.26	80.37	78.52	80
hepatitis	83.87	84.52	85.81	83.87	83.87
					86.14
					99.6
					93.45
					95.33
					99.44
					84.21
					100
					79.17
					100
liver-disorders	68.7	71.01	71.59	70.43	71.88
lungcancer	50	50	53.12	46.88	56.25
lymph	77.03	78.38	81.08	79.05	80.41
monks	96.53	99.54	100	96.99	100
page-blocks	96.88	97.24	97.02	97.26	97.24
postop	70	71.11	56.67	71.11	67.78
primary-tumor	39.82	45.13	40.12	45.13	41.3
promoters	81.13	83.96	85.85	85.85	81.13
segment	96.93	96.97	98.48	96.88	97.45
sick	98.81	98.49	99.18	98.46	98.97
		97.83	96.59	97.83	97.83
		77.4	77.88	77.4	74.52
		86.82	92.83	86.38	93.12
		81.65	80.15	82.02	79.03
		92.07	96.35	91.65	93.53
		72.7	76.24	72.81	76.48
		94.58	95.66	94.58	95.44
		94.94	96.63	93.26	95.51
		60.04	56.4	31.2	60.51
		87.13	96.04	86.14	93.07
		82.79	83.62	81.64	83.93

	J48	Bagging	Boosting	BMA	BMC Linear
Average:	82.37	82.79	83.62	81.64	83.93

Friedman Signed-Rank Test:
Results significant ($p < 0.01$)

Critical differences between BMC and two of the other four strategies

WEIGHTING STRATEGY #2: DIRICHLET ASSIGNED-WEIGHTS



RESULTS

	J48	Bagging	Boosting	BMA	BMC-D
anneal	98.44	98.22	99.55	98.22	98.89
audiology	77.88	76.55	84.96	76.11	82.3
autos	81.46	69.76	83.9	70.24	84.88
balance-scale	76.64	82.88	78.88	82.88	81.92
bupa	68.7	71.01	71.59	70.43	71.88
cancer-wisc.	93.85	95.14	95.71	95.28	95.14
cancer-yugo.	75.52	67.83	69.58	68.18	73.08
car	92.36	92.19	96.12	92.01	93.75
cmc	52.14	53.63	50.78	41.96	52.95
credit-a	86.09	85.07	84.2	84.93	85.07
credit-g	70.5	74.4	69.6	74.3	73.1
dermatology	93.99	92.08	95.63	92.08	95.36
diabetes	73.83	74.61	72.4	74.61	74.35
echo	97.3	97.3	95.95	97.3	97.3
ecoli-c	84.23	83.04	81.25	82.74	84.52
glass	66.82	69.63	74.3	68.69	70.09
haberman	71.9	73.2	72.55	73.2	74.51
heart-cleveland	77.56	82.18	82.18	82.18	79.87
heart-h	80.95	78.57	78.57	78.57	79.59
heart-statlog	76.67	79.26	80.37	78.52	80
hepatitis	83.87	84.52	85.81	83.87	83.87
horse-colic	85.33	85.33	83.42	85.05	86.14
hypothyroid	99.58	99.55	99.58	99.55	99.6
ionosphere	91.45	90.88	93.16	90.6	93.45
iris	96	94	93.33	94	95.33
kr-vs-kp	99.44	99.12	99.5	99.12	99.44
labor	73.68	85.96	89.47	87.72	84.21
led	100	100	100	100	100
lenses	83.33	66.67	70.83	58.33	79.17
letter	100	100	100	100	100
liver-disorders	68.7	71.01	71.59	70.43	71.88
lungcancer	50	50	53.12	46.88	56.25
lymph	77.03	78.38	81.08	79.05	80.41
monks	96.53	99.54	100	96.99	100
page-blocks	96.88	97.24	97.02	97.26	97.24
postop	70	71.11	56.67	71.11	67.78
primary-tumor	39.82	45.13	40.12	45.13	41.3
promoters	81.13	83.96	85.85	85.85	81.13
segment	96.93	96.97	98.48	96.88	97.45
sick	98.81	98.49	99.18	98.46	98.97
solar-flare	97.83	97.83	96.59	97.83	97.83
sonar	71.15	77.4	77.88	77.4	74.52
soybean	91.51	86.82	92.83	86.38	93.12
spect	78.28	81.65	80.15	82.02	79.03
tic-tac-toe	85.07	92.07	96.35	91.65	93.53
vehicle	72.46	72.7	76.24	72.81	76.48
vote	94.79	94.58	95.66	94.58	95.44
wine	93.82	94.94	96.63	93.26	95.51
yeast	56	60.04	56.4	31.2	60.51
zoo	92.08	87.13	96.04	86.14	93.07
average:	82.37	82.79	83.62	81.64	84.02

RESULTS

	J48	Bagging	Boosting	BMA	BMC Dirichlet
Average:	82.37	82.79	83.62	81.64	84.02

	J48	Bagging	Boosting	BMA	BMC-D
anneal	98.44	98.22	99.55	98.22	98.89
audiology	77.88	76.55	84.96	76.11	82.3
autos	81.46	69.76	83.9	70.24	84.88
balance-scale	76.64	82.88	78.88	82.88	81.92
bupa	68.7	71.01	71.59	70.43	71.88
cancer-wisc.	93.85	95.14	95.71	95.28	95.14
cancer-yugo.	75.52	67.83	69.58	68.18	73.08
car	92.36	92.19	96.12	92.01	93.75
cmc	52.14	53.63	50.78	41.96	52.95
credit-a	86.09	85.07	84.2	84.93	85.07
credit-g	70.5	74.4	69.6	74.3	73.1
dermatology	93.99	92.08	95.63	92.08	95.36
diabetes	73.83	74.61	72.4	74.61	74.35
echo	97.3	97.3	95.95	97.3	97.3
ecoli-c	84.23	83.04	81.25	82.74	84.52
glass	66.82	69.63	74.3	68.69	70.09
haberman	71.9	73.2	72.55	73.2	74.51
heart-cleveland	77.56	82.18	82.18	82.18	79.87
heart-h	80.95	78.57	78.57	78.57	79.59
heart-statlog	76.67	79.26	80.37	78.52	80
hepatitis	83.87	84.52	85.81	83.87	83.87
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	86.14
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	99.6
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	93.45
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	95.33
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	99.44
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	84.21
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	100
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	79.17
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	100
liver-disorders	68.7	71.01	71.59	70.43	71.88
lungcancer	50	50	53.12	46.88	56.25
lymph	77.03	78.38	81.08	79.05	80.41
monks	96.53	99.54	100	96.99	100
page-blocks	96.88	97.24	97.02	97.26	97.24
postop	70	71.11	56.67	71.11	67.78
primary-tumor	39.82	45.13	40.12	45.13	41.3
promoters	81.13	83.96	85.85	85.85	81.13
segment	96.93	96.97	98.48	96.88	97.45
sick	98.81	98.49	99.18	98.46	98.97
solar-flare	97.83	97.83	96.59	97.83	97.83
sonar	71.15	77.4	77.88	77.4	74.52
soybean	91.51	86.82	92.83	86.38	93.12
spect	78.28	81.65	80.15	82.02	79.03
tic-tac-toe	85.07	92.07	96.35	91.65	93.53
vehicle	72.46	72.7	76.24	72.81	76.48
vote	94.79	94.58	95.66	94.58	95.44
wine	93.82	94.94	96.63	93.26	95.51
yeast	56	60.04	56.4	31.2	60.51
zoo	92.08	87.13	96.04	86.14	93.07
average:	82.37	82.79	83.62	81.64	84.02

RESULTS

	J48	Bagging	Boosting	BMA	BMC-D
anneal	98.44	98.22	99.55	98.22	98.89
audiology	77.88	76.55	84.96	76.11	82.3
autos	81.46	69.76	83.9	70.24	84.88
balance-scale	76.64	82.88	78.88	82.88	81.92
bupa	68.7	71.01	71.59	70.43	71.88
cancer-wisc.	93.85	95.14	95.71	95.28	95.14
cancer-yugo.	75.52	67.83	69.58	68.18	73.08
car	92.36	92.19	96.12	92.01	93.75
cmc	52.14	53.63	50.78	41.96	52.95
credit-a	86.09	85.07	84.2	84.93	85.07
credit-g	70.5	74.4	69.6	74.3	73.1
dermatology	93.99	92.08	95.63	92.08	95.36
diabetes	73.83	74.61	72.4	74.61	74.35
echo	97.3	97.3	95.95	97.3	97.3
ecoli-c	84.23	83.04	81.25	82.74	84.52
glass	66.82	69.63	74.3	68.69	70.09
haberman	71.9	73.2	72.55	73.2	74.51
heart-cleveland	77.56	82.18	82.18	82.18	79.87
heart-h	80.95	78.57	78.57	78.57	79.59
heart-statlog	76.67	79.26	80.37	78.52	80
hepatitis	83.87	84.52	85.81	83.87	83.87
					86.14
					99.6
					93.45
					95.33
					99.44
					84.21
					100
					79.17
					100
liver-disorders	68.7	71.01	71.59	70.43	71.88
lungcancer	50	50	53.12	46.88	56.25
lymph	77.03	78.38	81.08	79.05	80.41
monks	96.53	99.54	100	96.99	100
page-blocks	96.88	97.24	97.02	97.26	97.24
postop	70	71.11	56.67	71.11	67.78
primary-tumor	39.82	45.13	40.12	45.13	41.3
promoters	81.13	83.96	85.85	85.85	81.13
segment	96.93	96.97	98.48	96.88	97.45
sick	98.81	98.49	99.18	98.46	98.97
		97.83	96.59	97.83	97.83
		77.4	77.88	77.4	74.52
		86.82	92.83	86.38	93.12
		81.65	80.15	82.02	79.03
		92.07	96.35	91.65	93.53
		72.7	76.24	72.81	76.48
		94.58	95.66	94.58	95.44
		94.94	96.63	93.26	95.51
		60.04	56.4	31.2	60.51
		87.13	96.04	86.14	93.07
		82.79	83.62	81.64	84.02

	J48	Bagging	Boosting	BMA	BMC Dirichlet
Average:	82.37	82.79	83.62	81.64	84.02

Friedman Signed-Rank Test:
Results significant ($p < 0.01$)

Critical differences between BMC and three of the other four strategies

THE WORLD OF BAYESIAN ENSEMBLES

THREE POTENTIAL TYPES OF BAYESIAN ENSEMBLES

- Compute the optimal set of ensemble weights given a set of trained classifiers
- Optimally train a set of classifiers given a fixed set of ensemble weights
- Simultaneously train the classifiers and find the ensemble weights

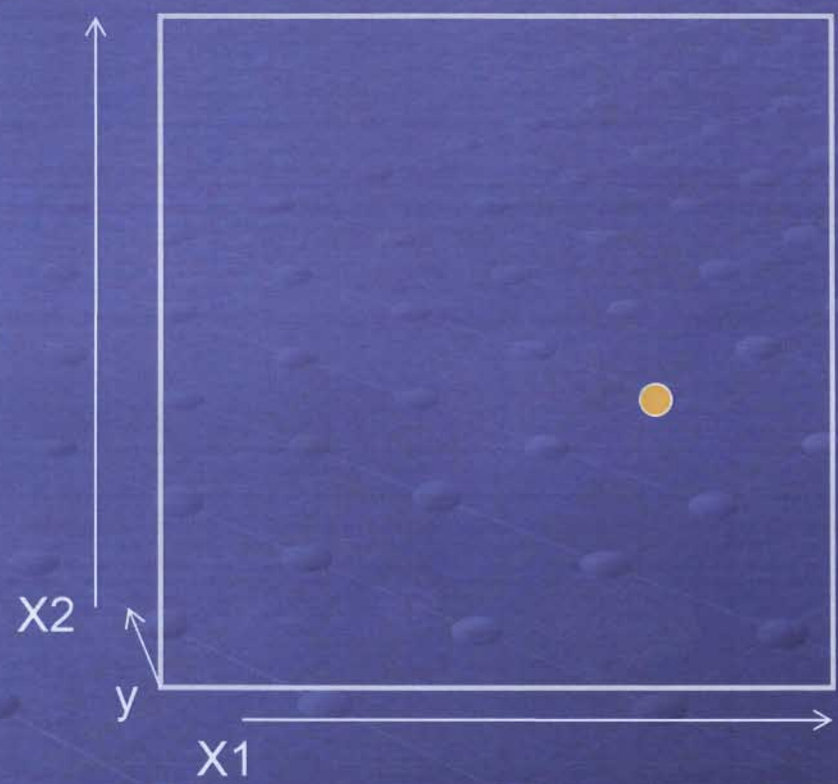
THREE POTENTIAL TYPES OF BAYESIAN ENSEMBLES

- Compute the optimal set of ensemble weights given a set of trained classifiers
- Optimally train a set of classifiers given a fixed set of ensemble weights
- Simultaneously train the classifiers and find the ensemble weights

THREE POTENTIAL TYPES OF BAYESIAN ENSEMBLES

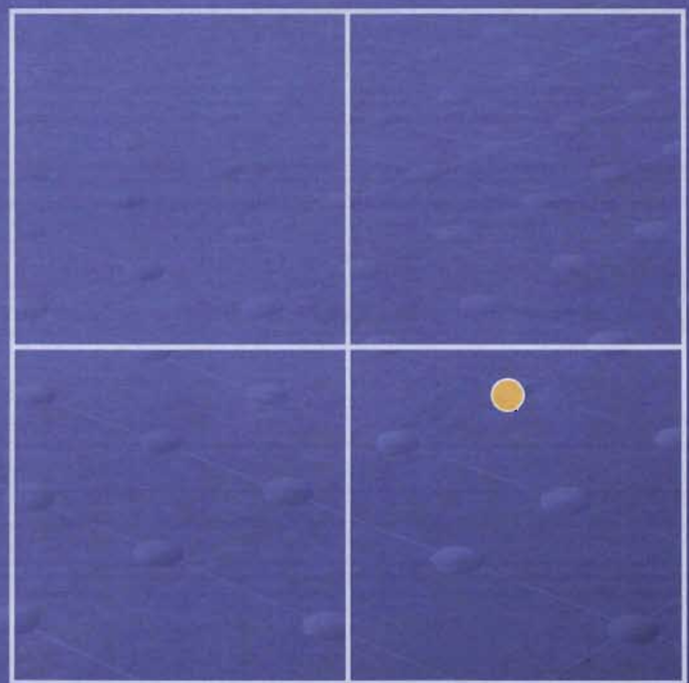
- Compute the optimal set of ensemble weights given a set of trained classifiers
- **Optimally train a set of classifiers given a fixed set of ensemble weights**
- Simultaneously train the classifiers and find the ensemble weights

CMAC Typology



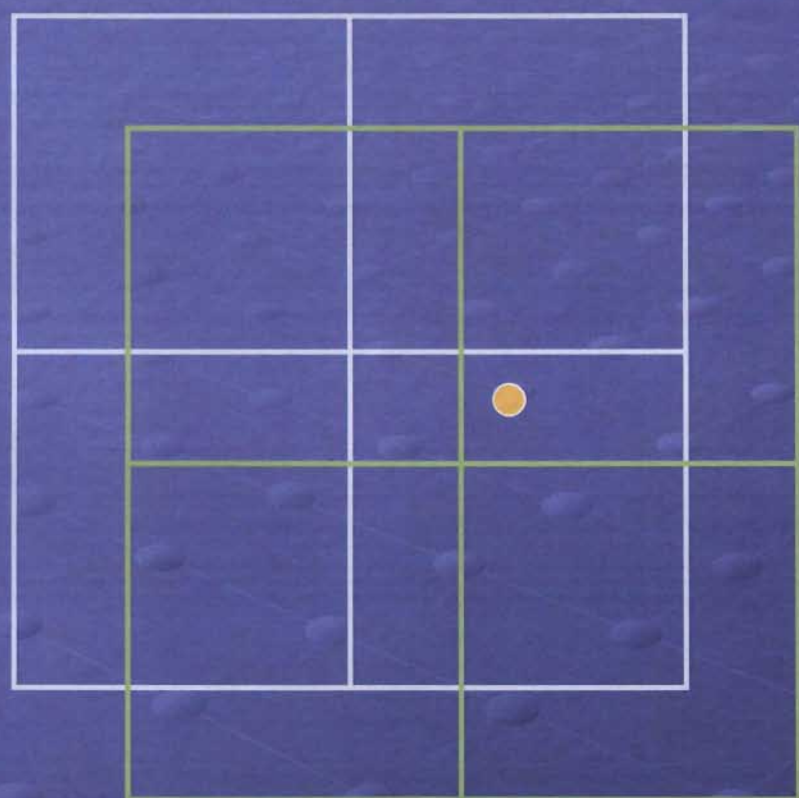
V=?

CMAC Typology



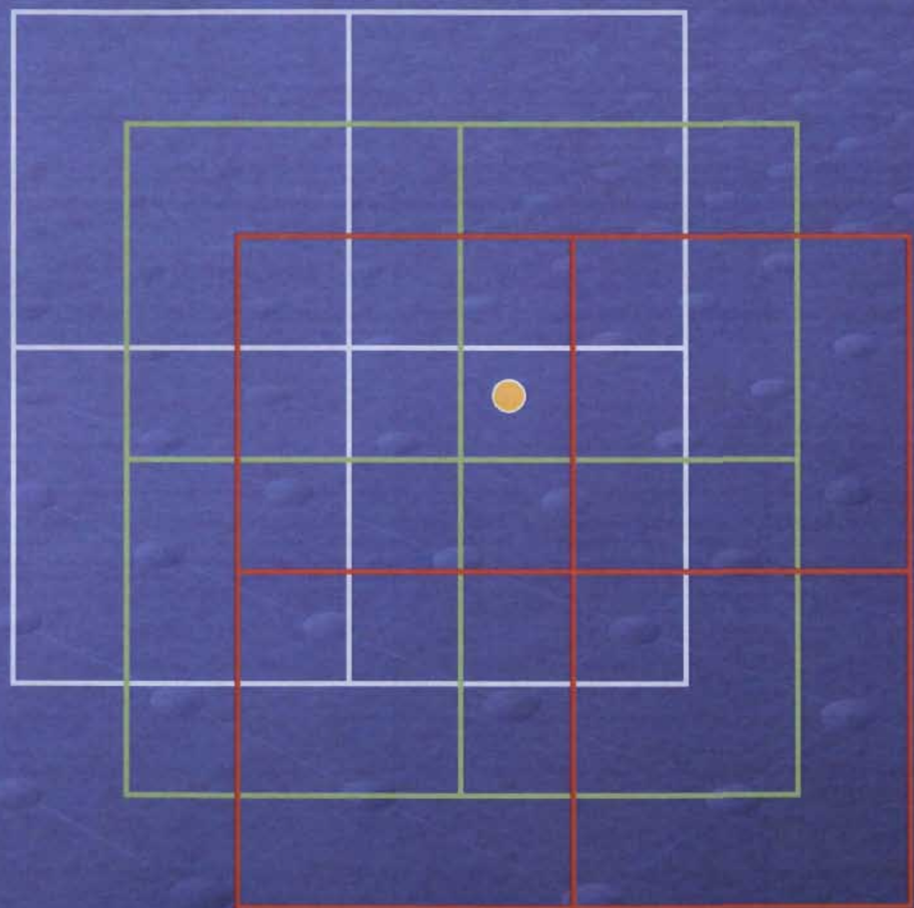
v=L1:4

CMAC Typology



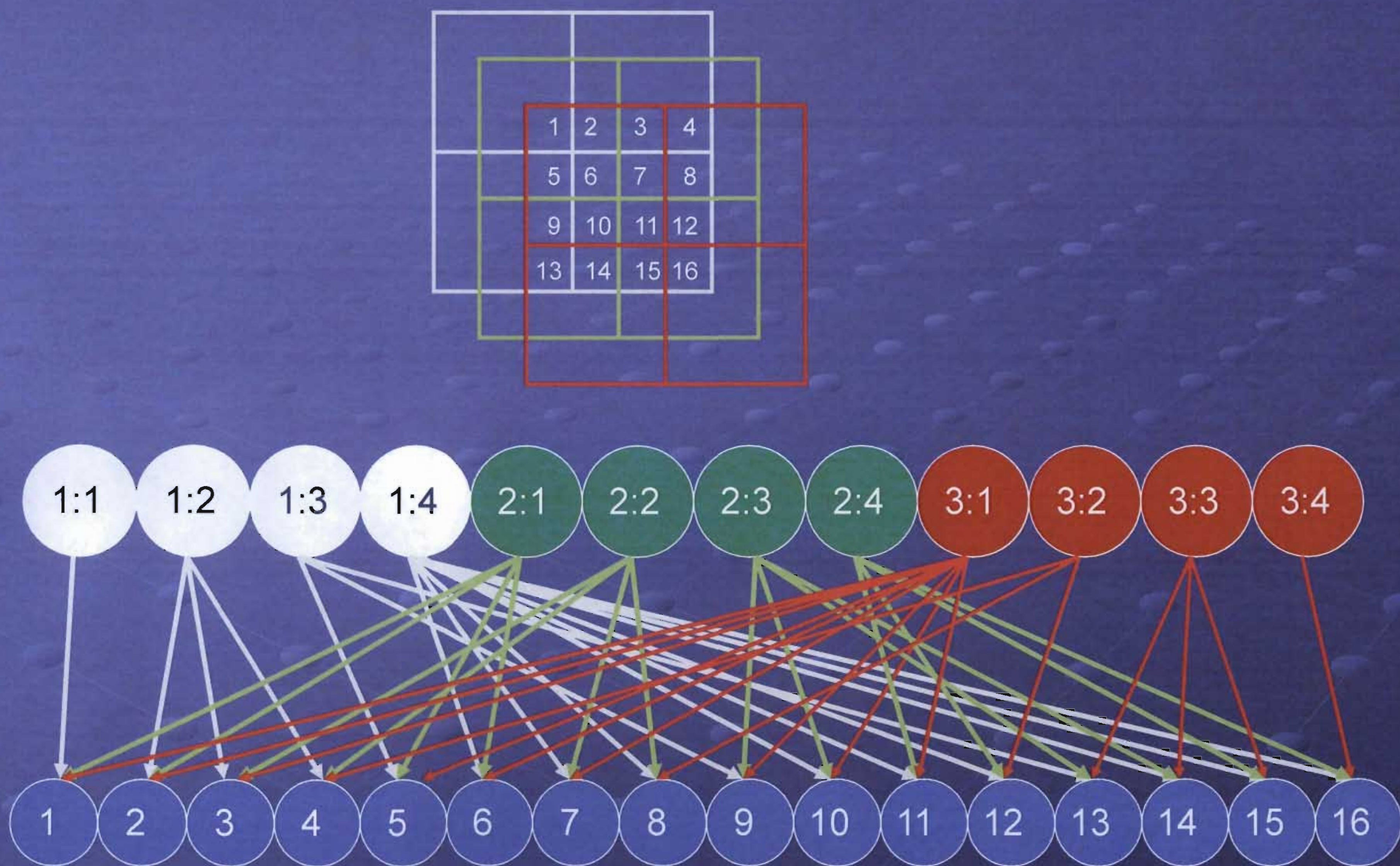
v=L1:4+L2:2

CMAC Typology

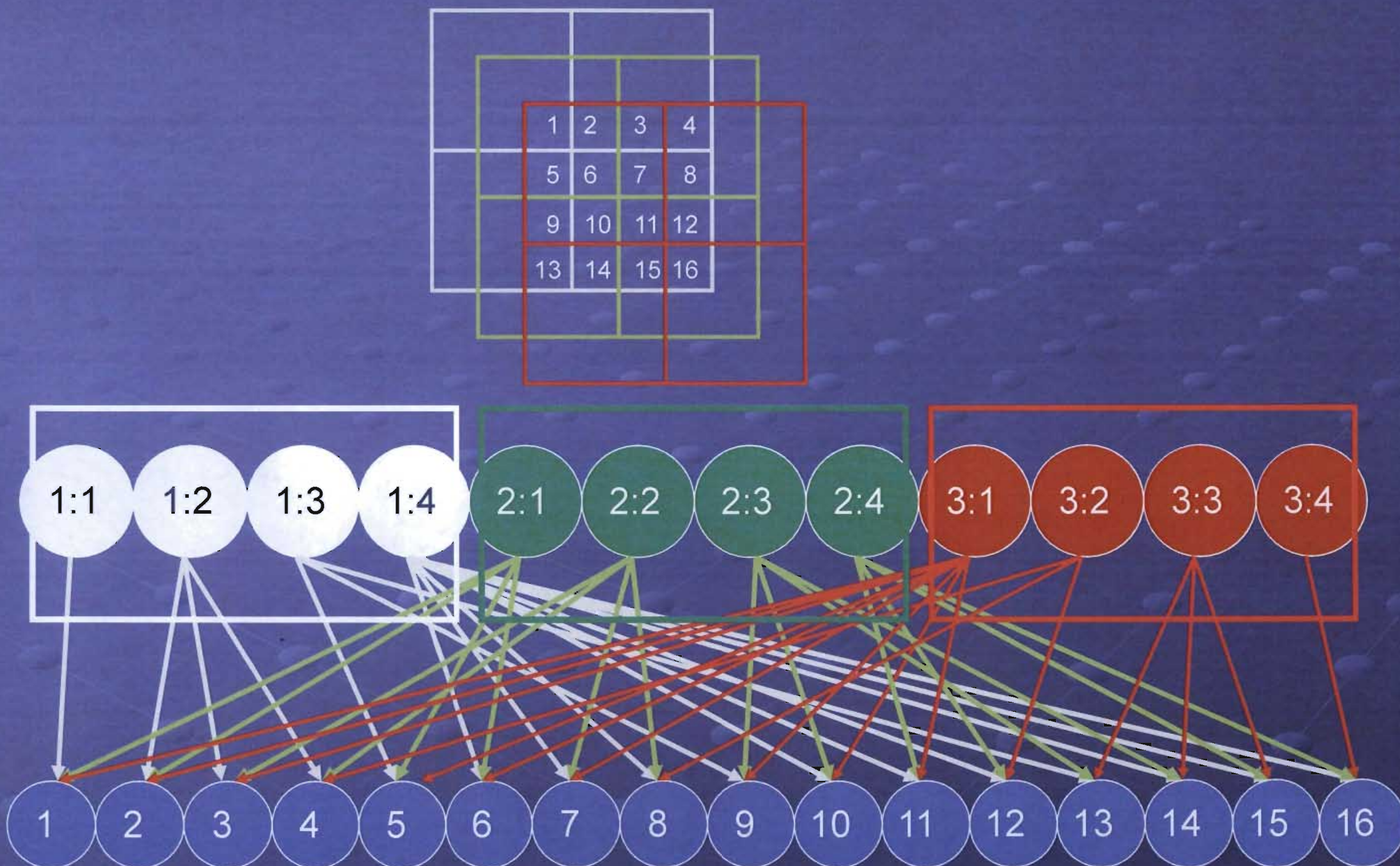


v=L1:4+L2:2+:L3:1

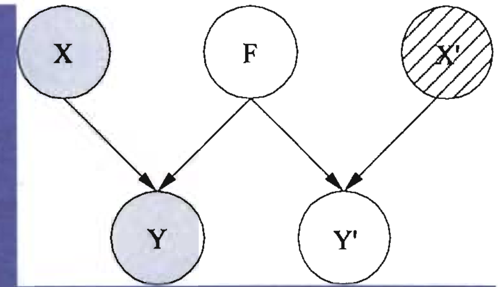
CMAC ANN representation of p(f)



CMAC Is an Ensemble



$$p(Y|X,F) = P(Y'|X',F)$$



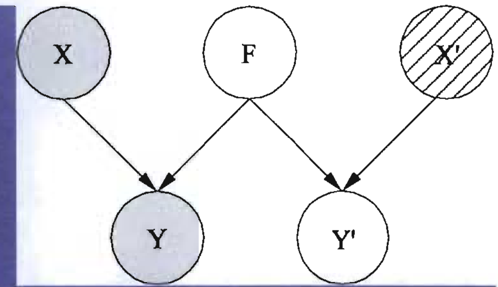
$$p(f) = \begin{cases} p(\mathbf{w}) & \text{if } f \in \{\lambda\mathbf{x}.g(\mathbf{x},\mathbf{w}) | \forall\mathbf{x} \in \mathbb{R}^D . g(\mathbf{x},\mathbf{w}) = \\ & \sum_{i\in\mathcal{T}} w[i]b[i](x)\} \\ 0 & \text{otherwise} \end{cases}$$

$$p(w) = N(w|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

$$p(y|x,f) = N(y|f(x), \boldsymbol{\Sigma}_y).$$

$$p(y|x,f) = N(y|\mathbf{Hw}, \boldsymbol{\Sigma}_y),$$

Bayesian Update



$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \mathbf{K}_1(\mathbf{y}_1 - \mathbf{H}\boldsymbol{\mu}_0),$$

$$\boldsymbol{\Sigma}_1 = (\mathbf{I} - \mathbf{K}_1\mathbf{H})(\boldsymbol{\Sigma}_0),$$

$$\mathbf{K}_1 = (\boldsymbol{\Sigma}_0)\mathbf{H}^T(\mathbf{H}(\boldsymbol{\Sigma}_0)\mathbf{H}^T + \boldsymbol{\Sigma}_y)^{-1}.$$

CMAC RESULTS

	CMAC	Bagging	BMA	BCMAC
Elusage	0.047	0.045	0.045	0.035
Gascon	0.140	0.135	0.134	0.041
longley	0.097	0.119	0.119	0.041
step2d	0.019	0.018	0.022	0.018
twoDimEgg	0.025	0.109	0.270	0.018
optimalBMA	0.005	0.071	0.006	0.002
Average:	**0.0555**	**0.08283**	**0.09933**	**0.02583**

CMAC RESULTS

	CMAC	Bagging	BMA	BCMAC
Average:	0.0555	0.08283	0.09933	0.02583

	CMAC	Bagging	BMA	BCMAC
Elusage	0.047	0.045	0.045	0.035
Gascon	0.140	0.135	0.134	0.041
Longley	0.097	0.119	0.119	0.041
ster 2d	0.019	0.018	0.022	0.018
twoDimEgg	0.025	0.109	0.270	0.018
optimalBMA	0.005	0.071	0.006	0.002
Average:	0.0555	0.08283	0.09933	0.02583

OBSERVATIONS

- The CMAC is an example of an ensemble with a fixed weighting scheme
- The parameters for each member of the ensemble can be solved in closed form given the fixed weighting scheme
- This approach significantly out performs traditional CMAC learning rules

THREE POTENTIAL TYPES OF BAYESIAN ENSEMBLES

- Compute the optimal set of ensemble weights given a set of trained classifiers
- Optimally train a set of classifiers given a fixed set of ensemble weights
- Future Work Simultaneously train the classifiers and find the ensemble weights

CONCLUSIONS

- Bayesian Model Averaging is not the optimal approach to model combination
  - It is the optimal approach for model selection
  - And it is outperformed by *ad hoc* techniques when the DGM is not in the model list
- Even the most simple forms of Bayesian Model Combination outperform BMA and these *ad hoc* techniques

FUTURE WORK

- Simultaneously train the classifiers and find the ensemble weights
- Further work will investigate other methods of generating diversity among the component ensembles (e.g. non-linear combinations) or using models that take spacial considerations into account

ANY QUESTIONS?