We Still Don't Know Why Model Ensembles Work

Pedro Domingos
Technical University of Lisbon

Overview

- The amazing success of model ensembles
- But why do they work?
  - Uncorrelated errors
  - Order-correctness
  - Bias-variance
  - Large margins
  - Logistic regression
  - Uncertainty
  - Better search/space
- So where does that leave us?

Model Ensembles

- Bagging (Breiman, 1996)
- Boosting (Schapire, 1990; Freund, 1995)
- Error-correcting output codes (Dietterich & Bakiri, 1995)
- Stacking (Wolpert, 1993; Breiman, 1996)
- Partitioning (Chan & Stolfo, 1995; Domingos, 1996)
- Etc.

Why Do Ensembles Work?

Why bother with this question?

- Design better ensemble algorithms
- Know when to apply them

What counts as an answer?

- Relation to properties of domain and base learner(s)
- Applicability

The "Uncorrelated Errors" Explanation

(Hansen & Salamon, 1990; Perrone & Cooper, 1993)

If Ensemble prediction is average of individual predictions

And Individual errors are zero-mean and uncorrelated

Then Ensemble error = Average error / #Models

- Errors never uncorrelated in practice
- Ali & Pazzani (1996) found *negatively correlated* errors help the most

The "Order-Correctness" Explanation

(Breiman, 1996)

- Learner is *order-correct* on x
  iff it predicts correct class most often
- Bagging makes order-correct learner nearly optimal
- Hard to predict order-correctness

The "Bias-Variance" Explanation

(Kong & Dietterich, 1995; Breiman, 1996; Tibshirani, 1996; Kohavi & Wolpert, 1996; Friedman, 1997)

Bias: Systematic component of error

Variance: Additional error due to sample fluctuations

- ECOC reduces bias & variance . . . (Kong & Dietterich, 1995)
- . . . But what makes the trade-off go away?
- Bagging and boosting reduce variance . . . (Breiman, 1996)
- . . . But no: Boosting often increases variance (Schapire *et al.*, 1997)

The "Large Margins" Explanation

(Schapire *et al.*, 1997; Grove & Schuurmans, 1998)

- Consider a classifier that works by thresholding a continuous output
- **Margin** = (Output − Threshold) × Prediction
- Large margin "guarantees" good generalization (VC-dim)
- Boosting: Continuous output = Sum of weighted votes
- Boosting increases margins ... (Schapire *et al.*, 1997)
- ...But maximizing them increases error (Grove & Schuurmans, 1998)

The “Logistic Regression” Explanation

(Friedman, Hastie & Tibshirani, 1998)

- Boosting can be viewed as additive logistic regression
- Can lead to better-performing variants
- By itself does not explain why boosting works

The "Uncertainty" Explanation

(Buntine, 1990; Domingos, 1997; Dietterich, 1998)

> The training data might not provide sufficient information for choosing a single best classifier ... There are typically many plausible hypotheses. From this collection ... we can ... construct an ensemble.
>
> (Dietterich, 1998)

Hypothesis: Ensembles work because they take into account uncertainty as to which is the "correct" hypothesis.

Optimal uncertainty calculus: Bayesian model averaging

So are ensembles approximations of BMA?

Bayesian Model Averaging

Given Model space, Prior prob. of each model

Optimal learning strategy is:

- Compute each model's posterior prob. given sample (Posterior = Prior $\times$ Likelihood)
- To predict class: Average predictions of all models, weighted by posteriors

Problems:

- Finding correct model space and priors
- Computational complexity of model averaging

In practice: Use single maximum, or sample models

Bagging as Hedging

Bagging $\simeq$ Importance sampling approx. of BMA

Better approximation: Weights = Posteriors
(except for *extremely* large ensembles)

If Bagging is approximation of BMA

Then Better approximation should do better

Empirical Tests

- C4.5, 26 UCI DBs, 10-CV, 10–100 replicates, unif. priors
- Bagging with posteriors much worse (19 out of 26)
- Similar results with RISE learner and partitioning
- Simlar results with foreign exchange data
  and “full” BMA (all models, parameter averaging)
- Similar results with synthetic data and conjunctive learner
- Similar results with very small samples
- Behavior of ensembles and BMA is very different:
  In BMA, one model typically dominates

An Alternative Version

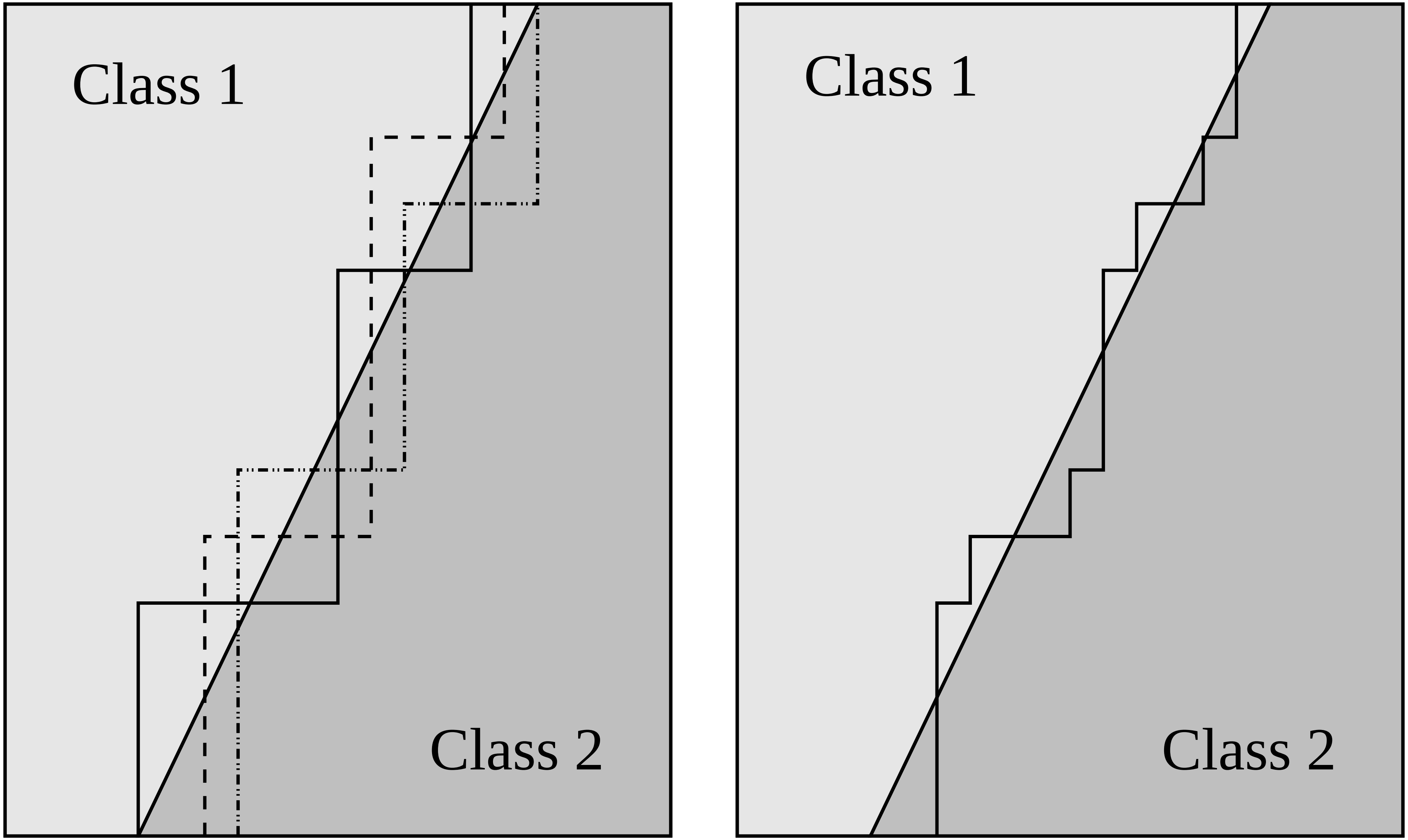
- What if appropriate prior cancels likelihood?
- Likelihood decreases exponentially with error
- To cancel it, prior must increase exp. w/ error
- Implies strong expectation of overfitting
  (consistent with unstable learner)
- Test: Measure correlation of in-bag & out-of-bag errors
- Correlation is almost always positive and large

Conclusion: Ensembles are not approxs. to BMA.

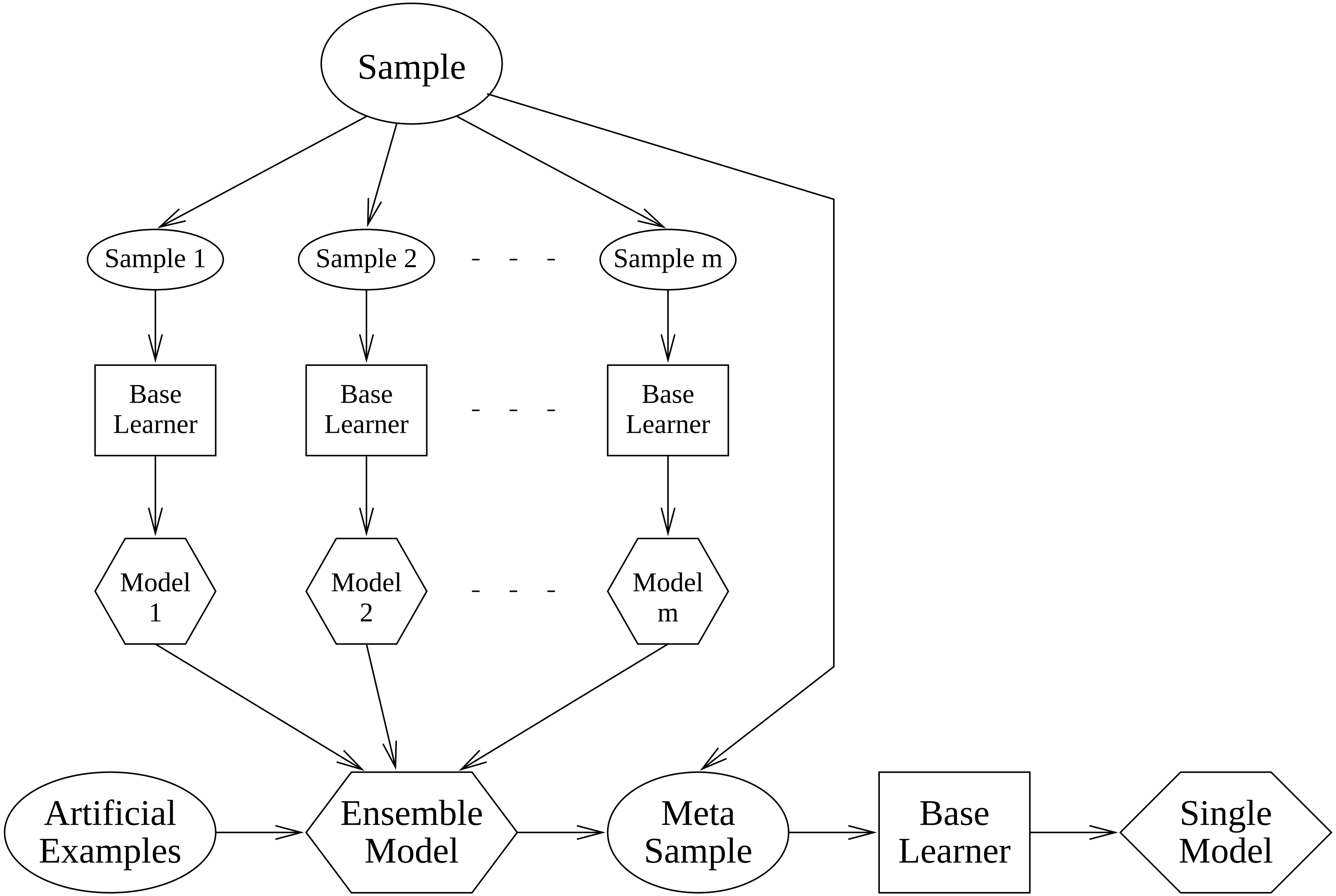
The “Better Search/Space” Explanation

(Domingos, 1997; Rao & Potts, 1997; Dietterich, 1998)

- Most learners incorporate preference for simpler models (Occam’s razor, fixed size net)
- Ensembles create much more complex model ($\sim$ Single $\times$ #Models)
- **Hypothesis:** Ensembles counteract Occam’s razor by shifting search to more complex models
- But what’s the ensemble’s “true” complexity?
- Solution: Reapply base learner to model ensemble


Class 1
Class 2
Class 1
Class 2

Experimental Setup



Results

- Simplified model:
  - Gets 60% of bagging's accuracy gains
  - 2 to 6 times more complex than single model
- Accuracy and complexity correlated

Consistent with hypothesis.

Joint Test

- Limited model space (e.g, decision stumps)
  $\Rightarrow$ Models underfit the data
- "Uncertainty" predicts bagging will increase error
- "Better search" predicts bagging will decrease error
- Test: Bagging decreases error by 8.4% on 27 UCI DBs (Freund & Schapire, 1995)

Contradicts "Uncertainty" explanation.

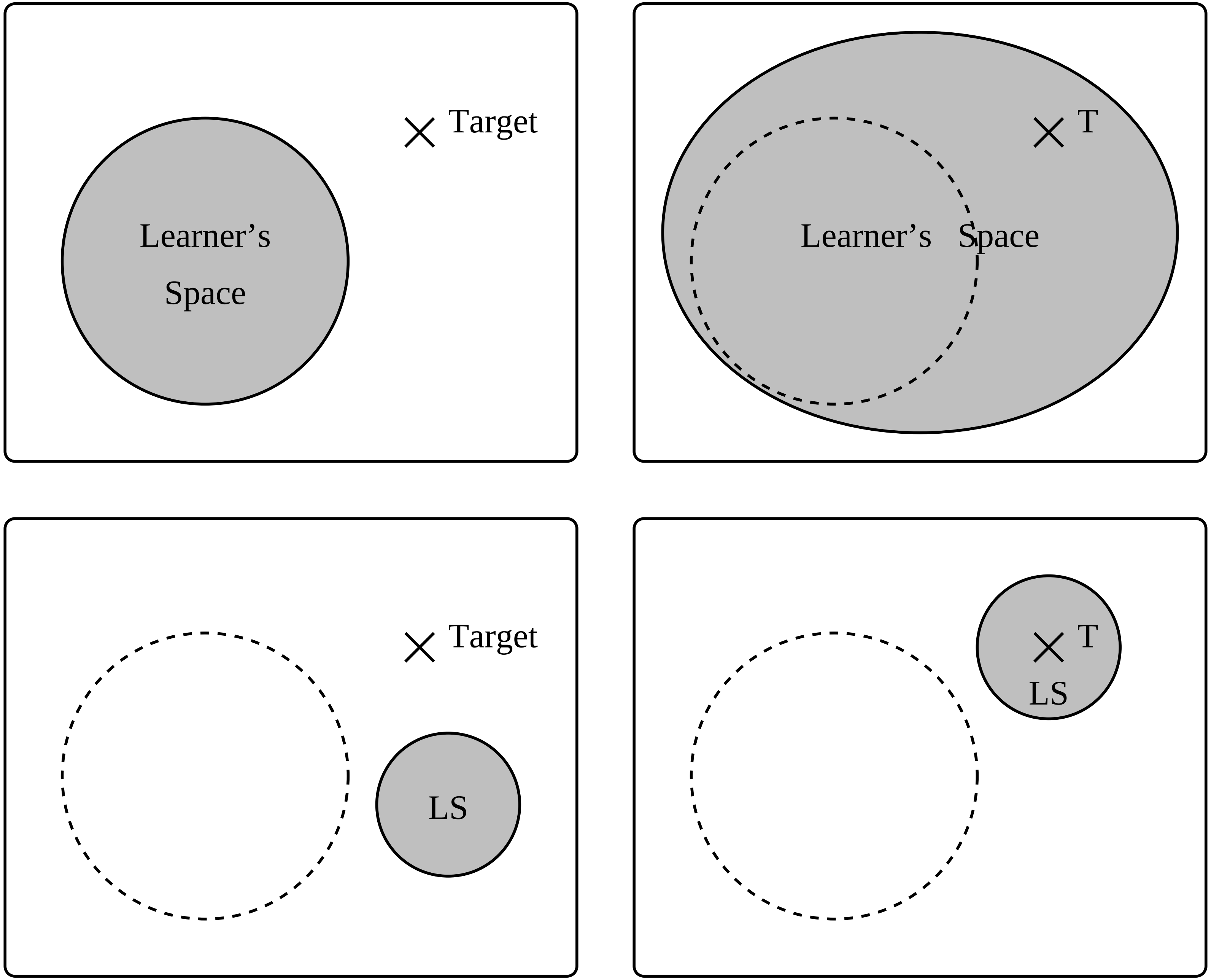
Consistent with "Better search" explanation.

Where Do We Go From Here?

“Better search/space” hypothesis is incomplete.

- Ensembles pick more complex models.
- But *which* complex models?

  Without a preference ensembles would strongly overfit.


Target
Learner's
Space
T
Learner's Space
Target
LS
T
LS

Some Ideas

- Averaging/voting is information-destroying operation
- Info(Learner) + Info(Data) $\ll$ Info(Model)
- Relation of weight entropy to bias-var of base learner
- More search vs. more informed evaluation
- Relation of ensemble language to base language
- Relation of true model to best base approximations

?