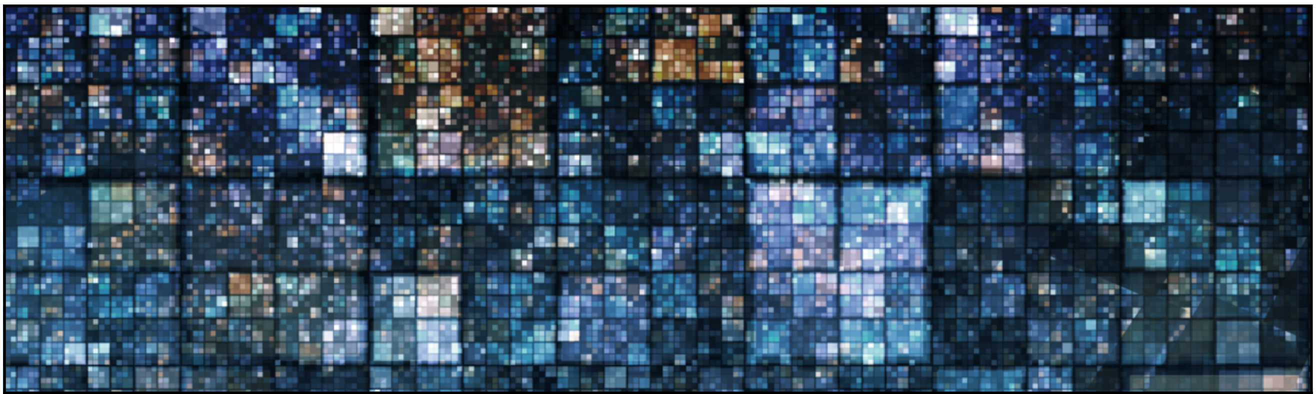


COMP47590

Advanced Machine Learning

Supervised Learning - Ensembles 1

Dr. Brian Mac Namee

Plagiarism & UCD Computer Science

- **Plagiarism is a serious academic offence**
  - [Student Code, sections 6.2 & 6.3] or [UCD Registry Plagiarism Policy] or [CS Plagiarism policy and procedures]
- Our staff and demonstrators are **proactive** in looking for possible plagiarism in all submitted work
- Suspected plagiarism is investigated by the CS Plagiarism subcommittee
  - Usually includes an interview with student(s) involved
  - 1st offence: **usually** 0 or NG in the affected components
  - 2nd offence: may be referred to the **University disciplinary committee**
- Student who enables plagiarism is equally responsible

http://www.ucd.ie/students/guide/academicregs.html

http://libguides.ucd.ie/academicintegrity

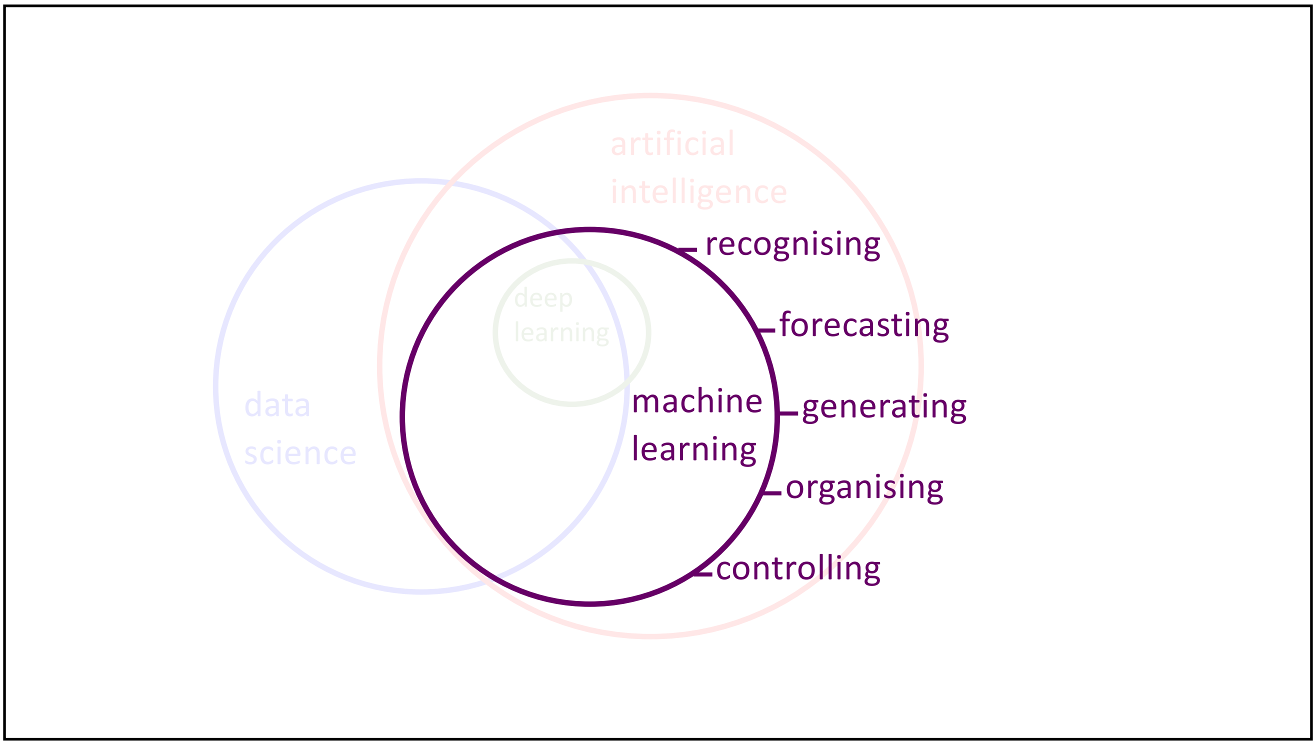
Information

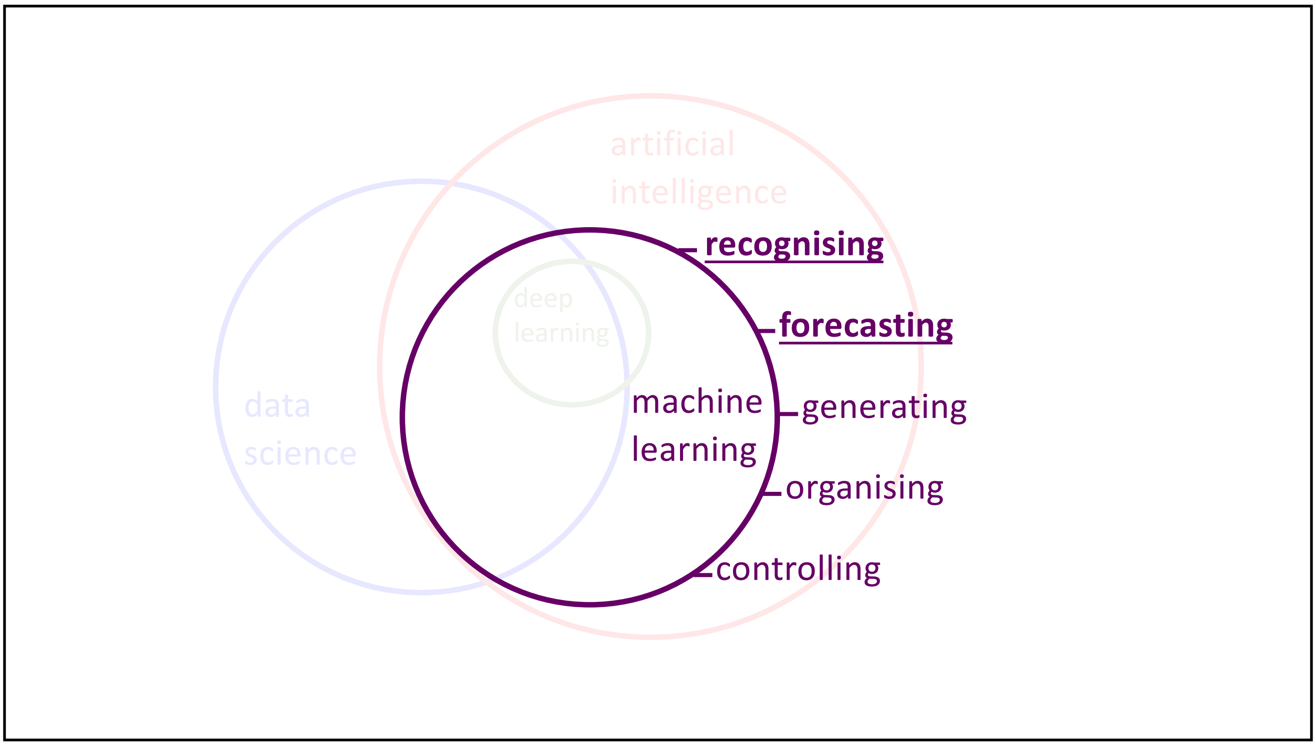
Email: Brian.MacNamee@ucd.ie

Course Materials: All material posted on UCD CS moodlehttps://csmoodle.ucd.ie/moodle/course/view.php?id=663

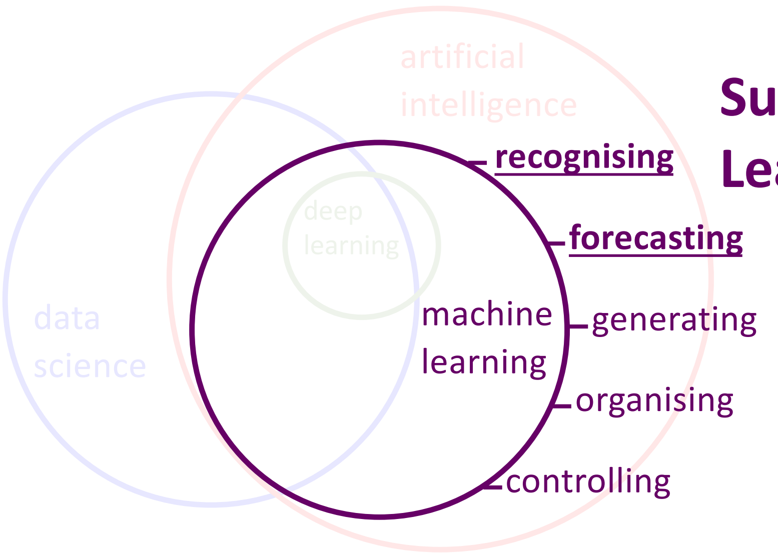
Enrolment key **UCDAdvML2017**

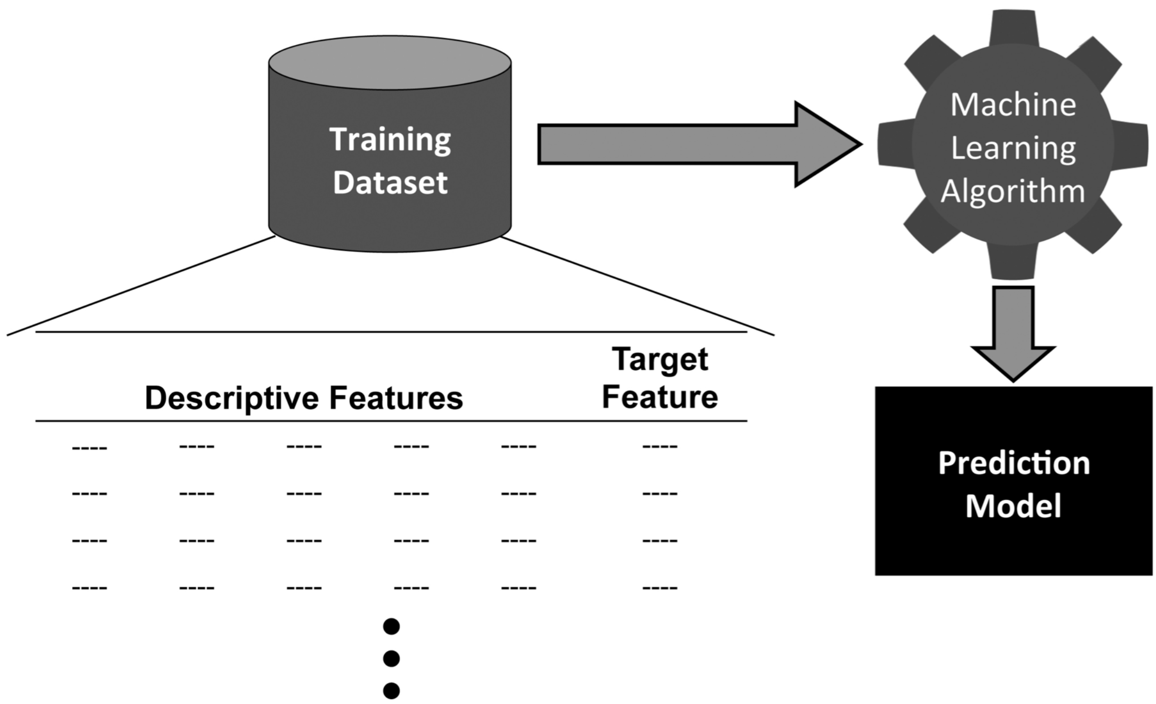
SUPERVISED LEARNING


artificial
intelligence
recognising
deep
learning
forecasting
machine
learning
generating
data
science
organising
controlling


artificial
intelligence
recognising
deep
learning
forecasting
machine
learning
generating
data
science
organising
controlling

Supervised Learning





Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

$$\mathcal{D} = [(\mathbf{d}_1, \mathrm{t}_1), (\mathbf{d}_2, \mathrm{t}_2), \ldots, (\mathbf{d}_n, \mathrm{t}_n)]$$

where $\mathbf{d}_i$ is a set of descriptive features

$\mathbf{d}_i[0], \mathbf{d}_i[1], \ldots, \mathbf{d}_i[\mathrm{m}]$

t_i is the corresponding target feature value

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

$$t = \mathbb{M}(\mathbf{q})$$

where $\mathbf{q}$ is a set of descriptive features $\mathbf{q}[0], \mathbf{q}[1], \ldots, \mathbf{q}[m]$ describing a query instance

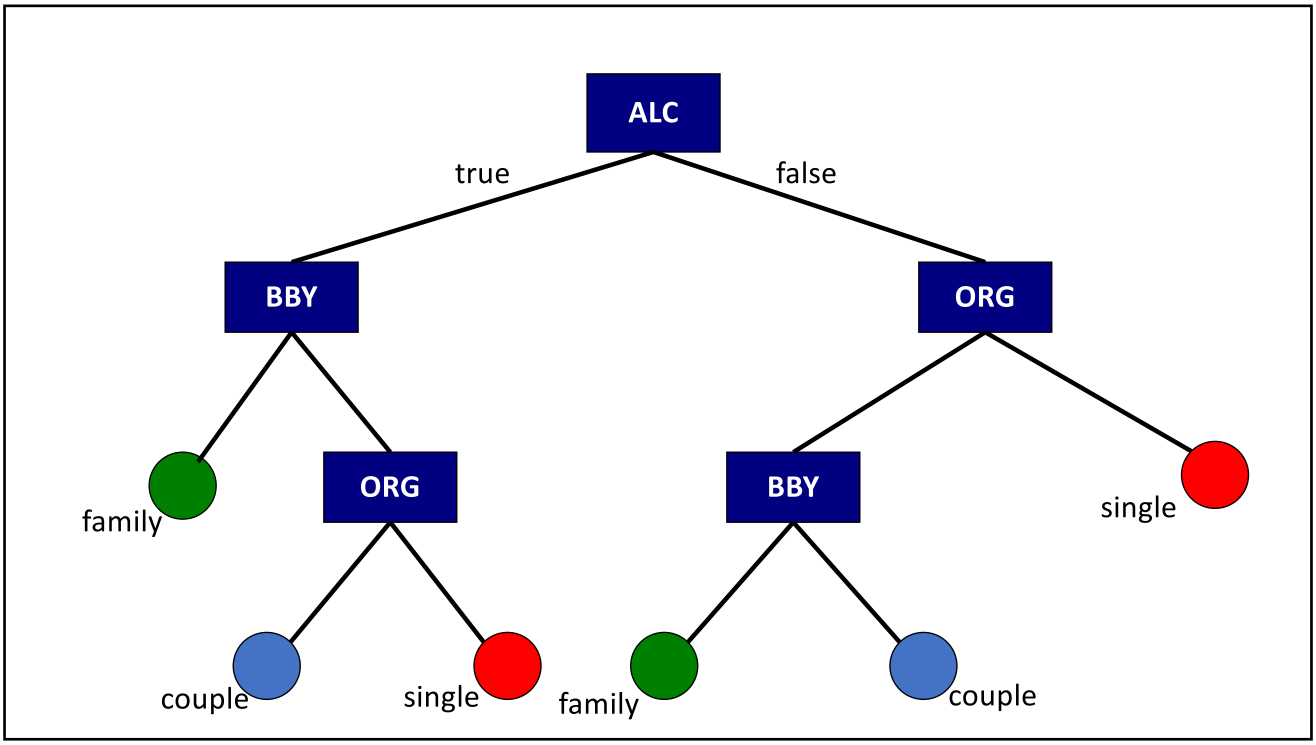
t is a predicted target feature value

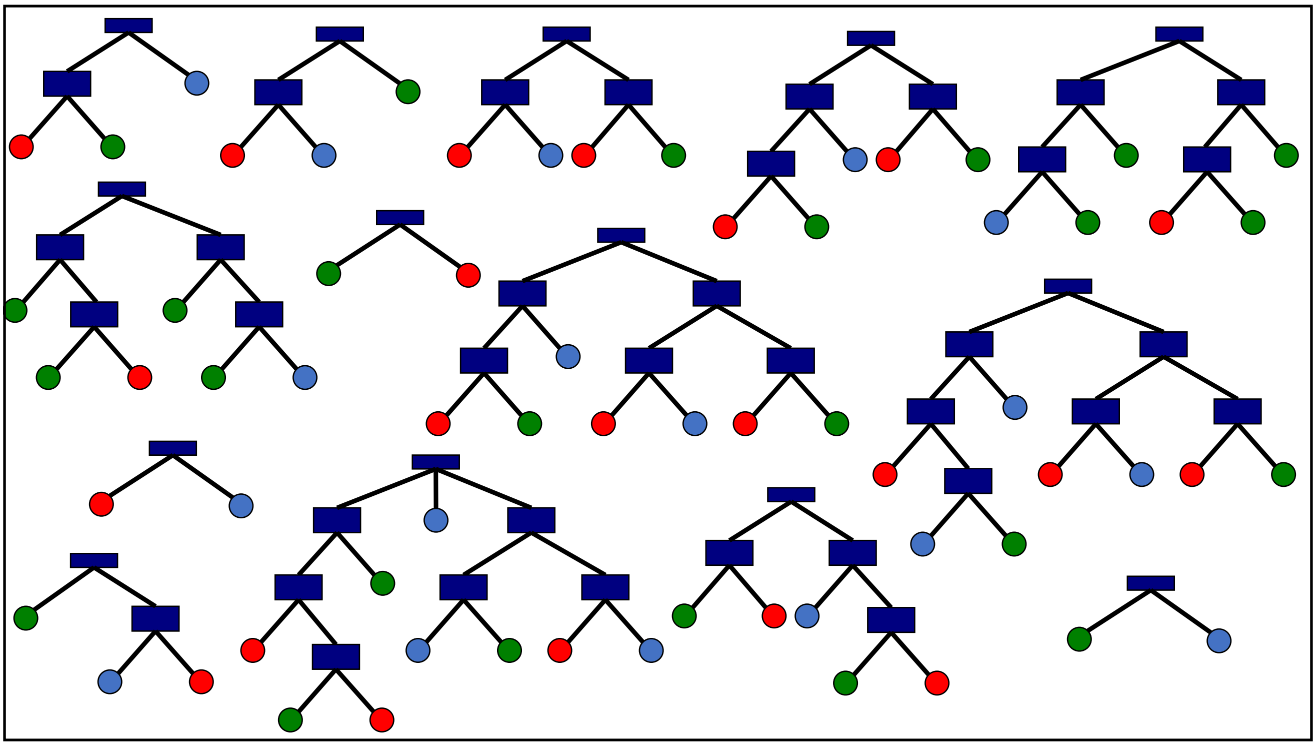
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

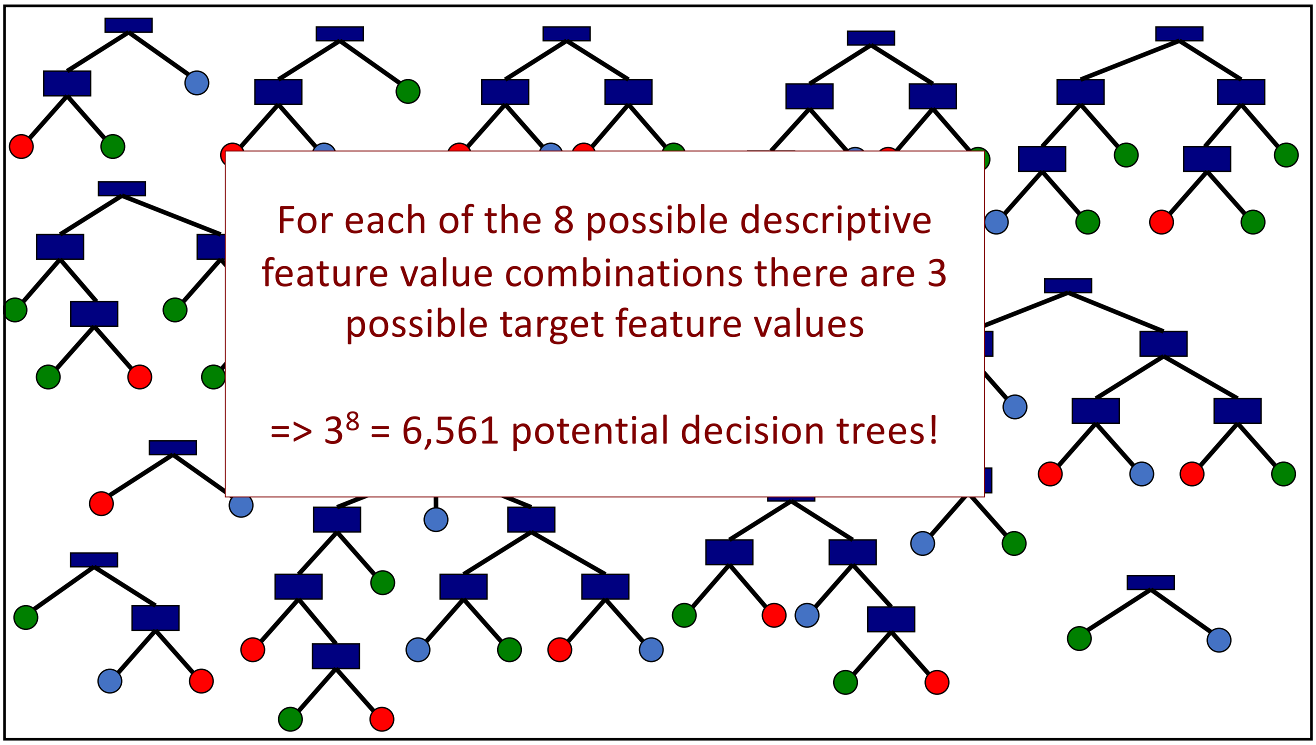
A simple retail dataset

ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy


ALC
true
false
BBY
ORG
family
ORG
BBY
single
couple
single
family
couple





For each of the 8 possible descriptive feature value combinations there are 3 possible target feature values

=> 3^8 = 6,561 potential decision trees!

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	?	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

Consistency?

Consistency ≈ memorizing the dataset

Consistency with noise in the data isn't desirable

Coverage through memorization is never possible in real problems

Consistency?

Consistency ≈ memorizing the dataset

Consistency with noise in the data isn't desirable

Coverage through memorization is never possible in real problems

GOAL: a model that **generalises** beyond the dataset and that **invariant** to the noise in the dataset

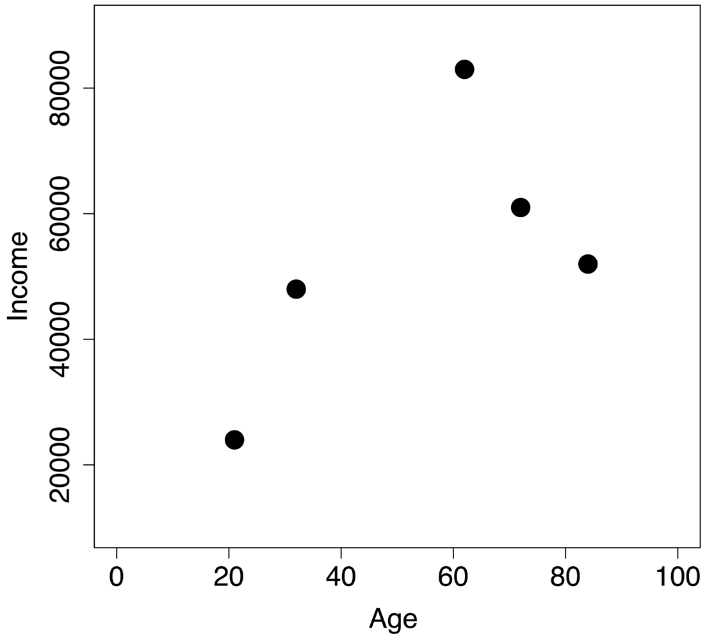
Inductive Bias

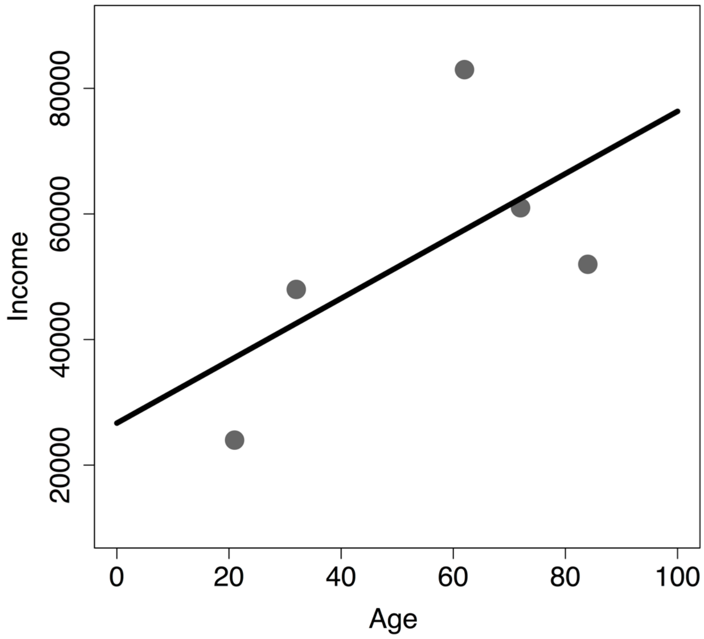
The solution is **inductive bias**, a set of assumptions that define the model selection criteria of an ML algorithm

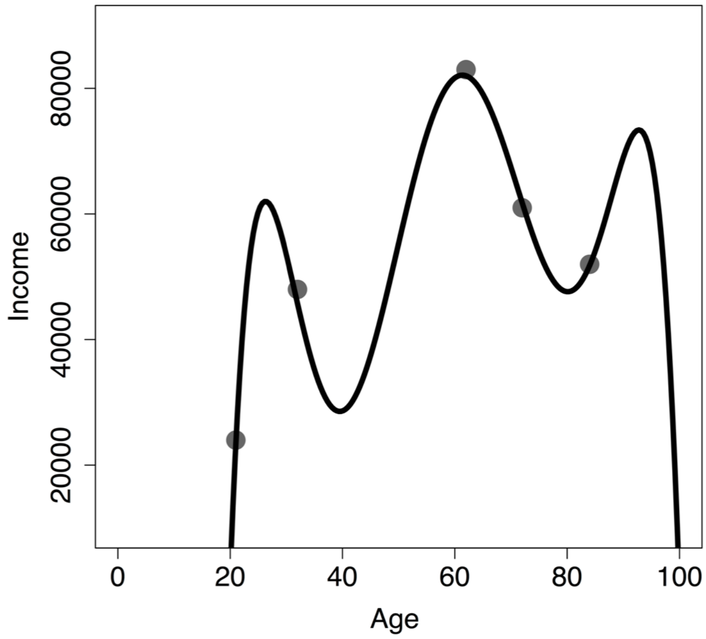
There are two types of bias that we can use:

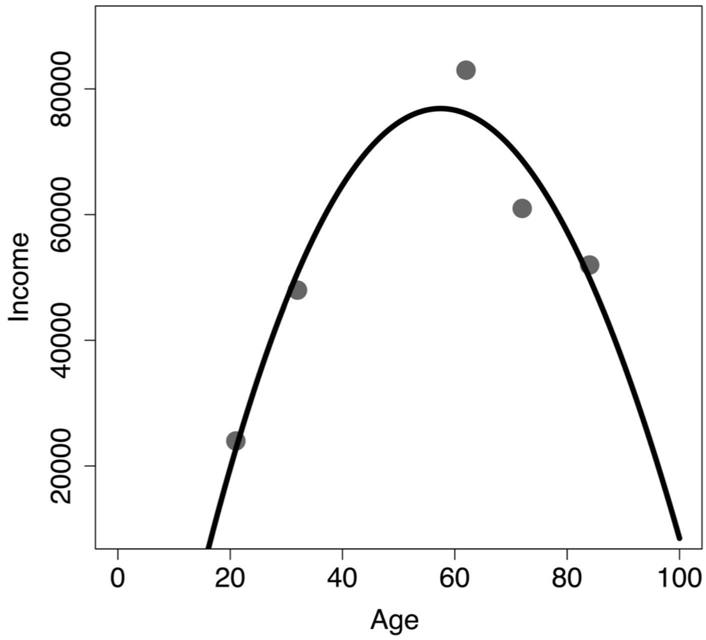
- restriction bias
- preference bias

Inductive bias is necessary for generalisation

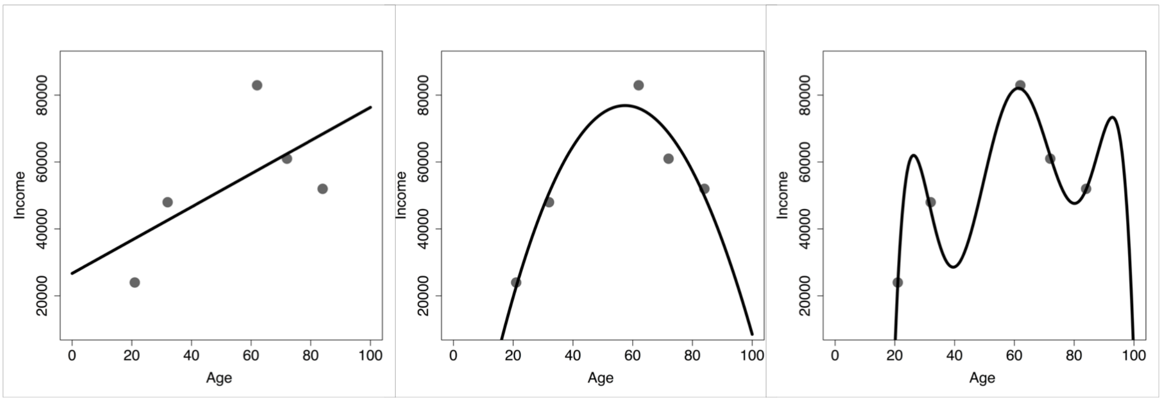

Income
80000
60000
40000
20000
0
20
40
60
80
100
Age


Income
80000
60000
40000
20000
0
20
40
60
80
100
Age


Income
20000
40000
60000
80000
0
20
40
60
80
100
Age

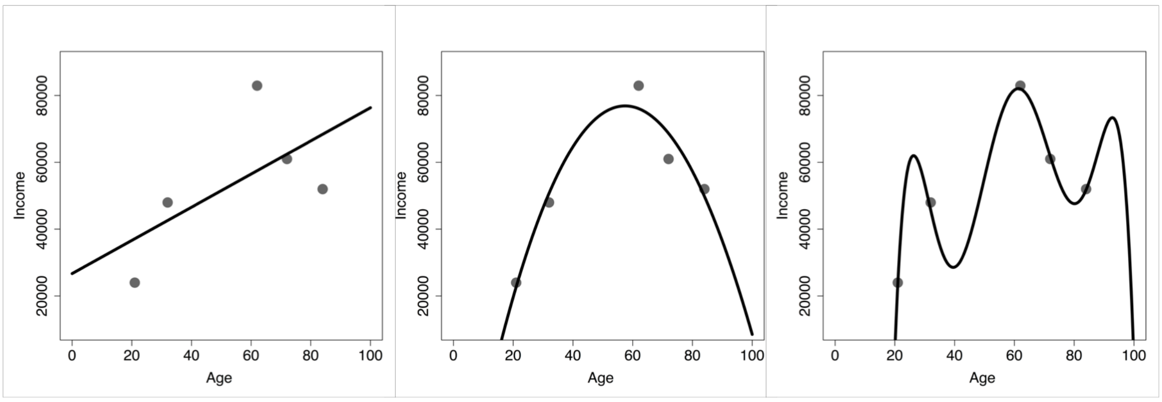

Income
20000
40000
60000
80000
0
20
40
60
80
100
Age

Underfitting



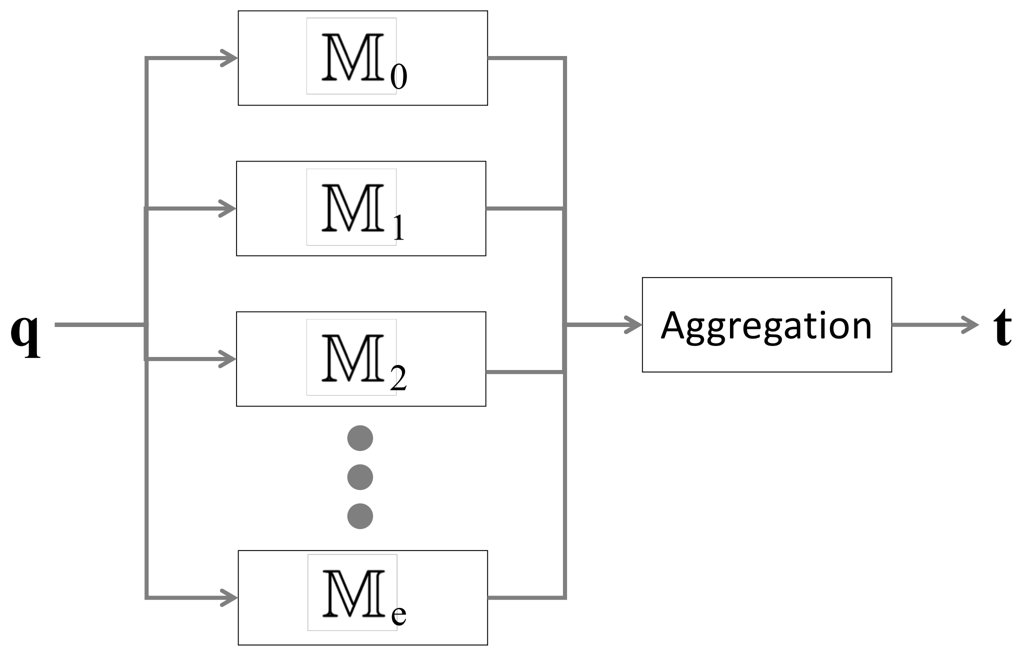
Overfitting

Bias



Variance

ENSEMBLES



Ensembles

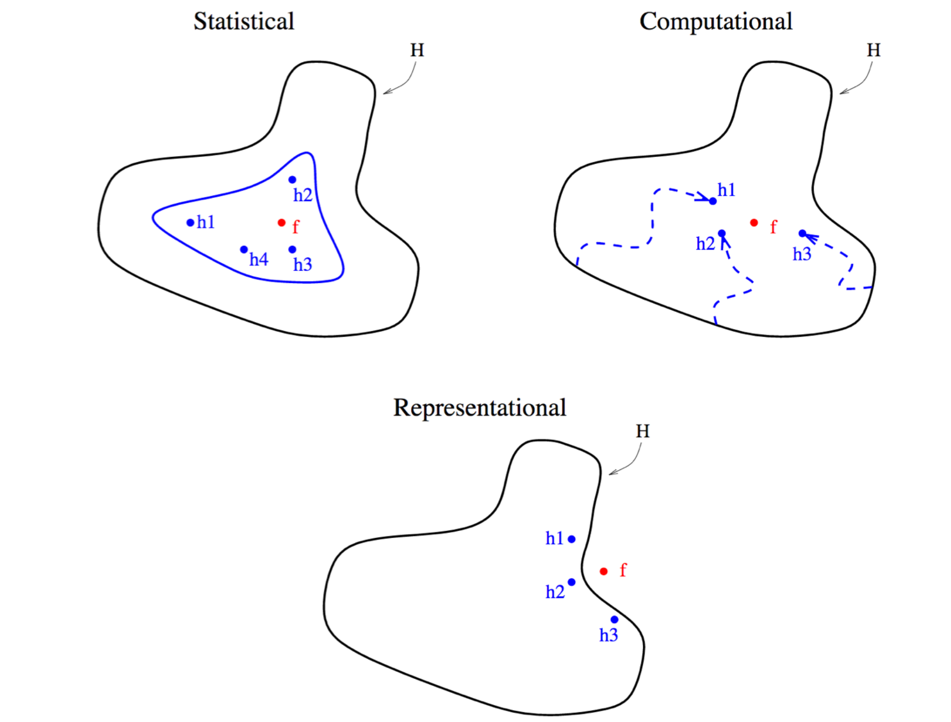
The aggregate of multiple combined models is more effective than any individual model

Thomas Dietterich describes 3 motivations for using ensembles:

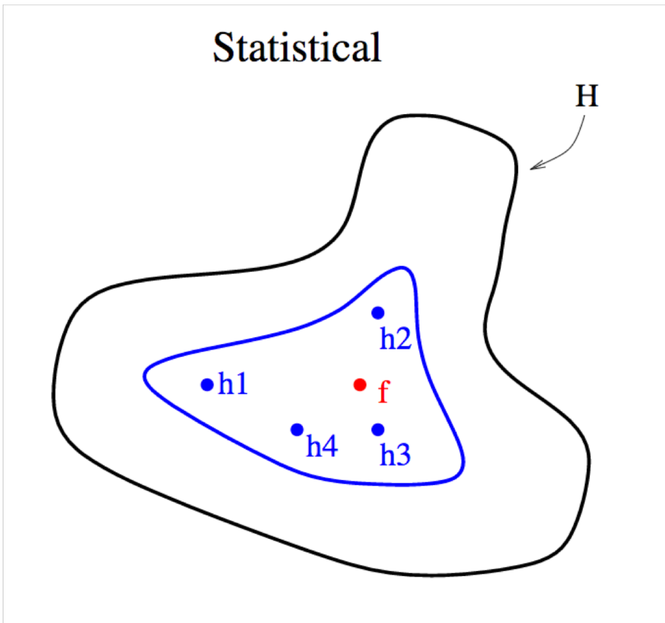
- Statistical
- Computational
- Representational

Ensemble Methods in Machine Learning
Thomas G Dietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

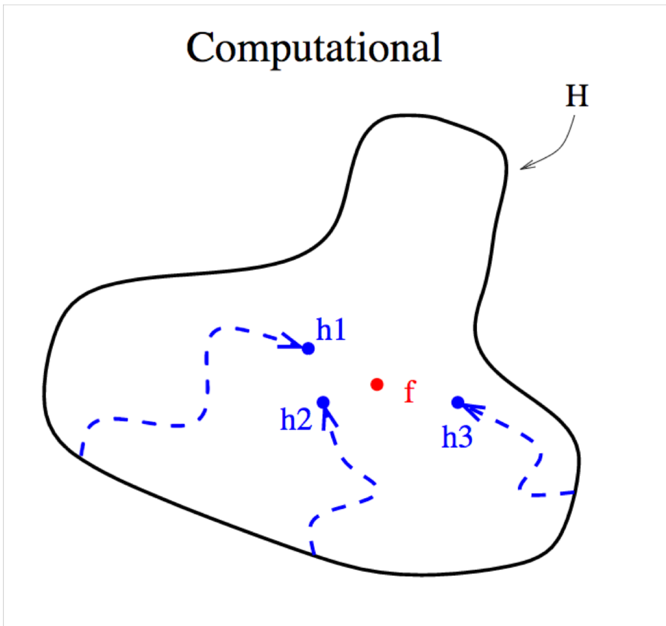
MAYBE A WISDOM OF THE CROWD EXPERIMENT



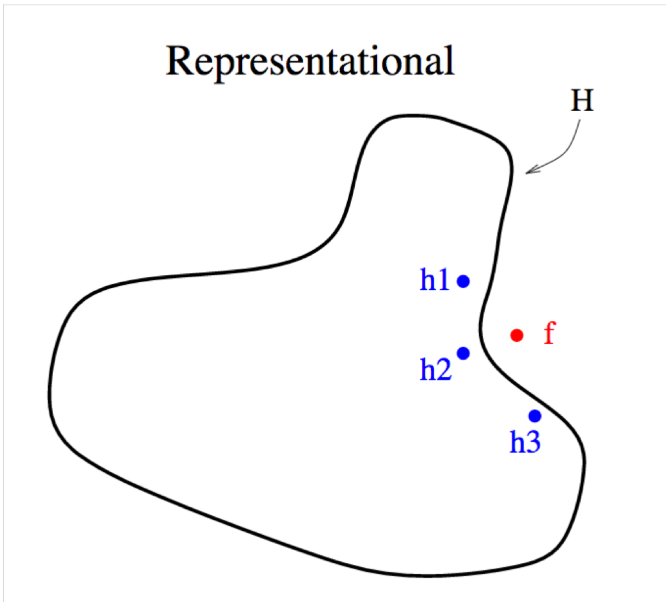
Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

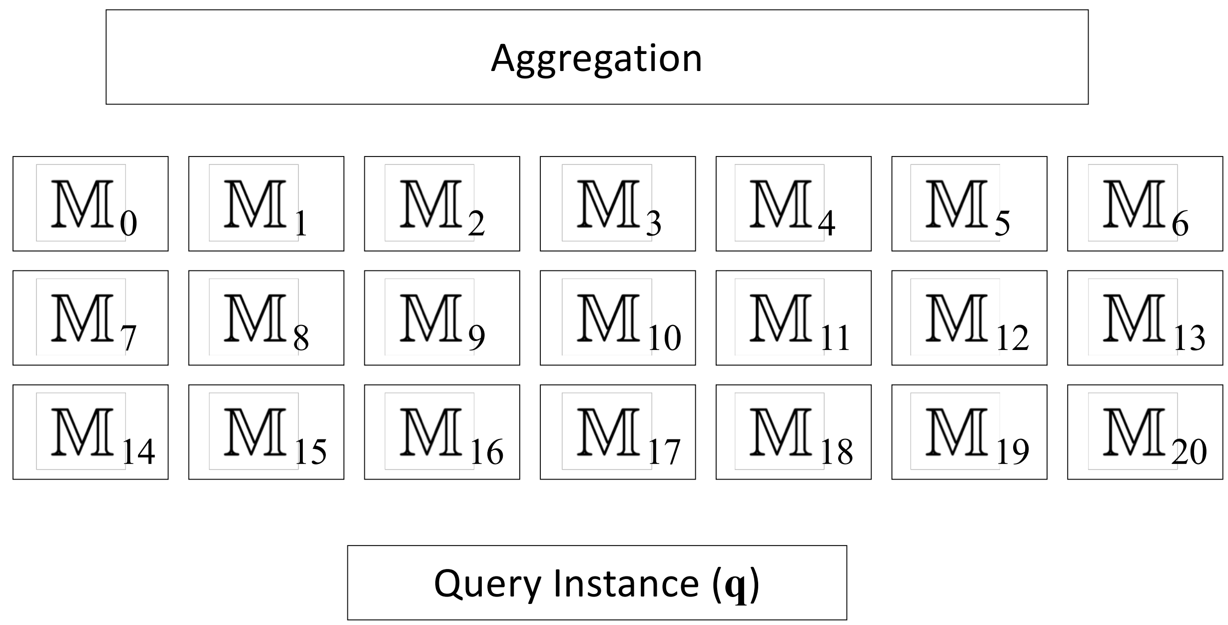


Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

Ensembles

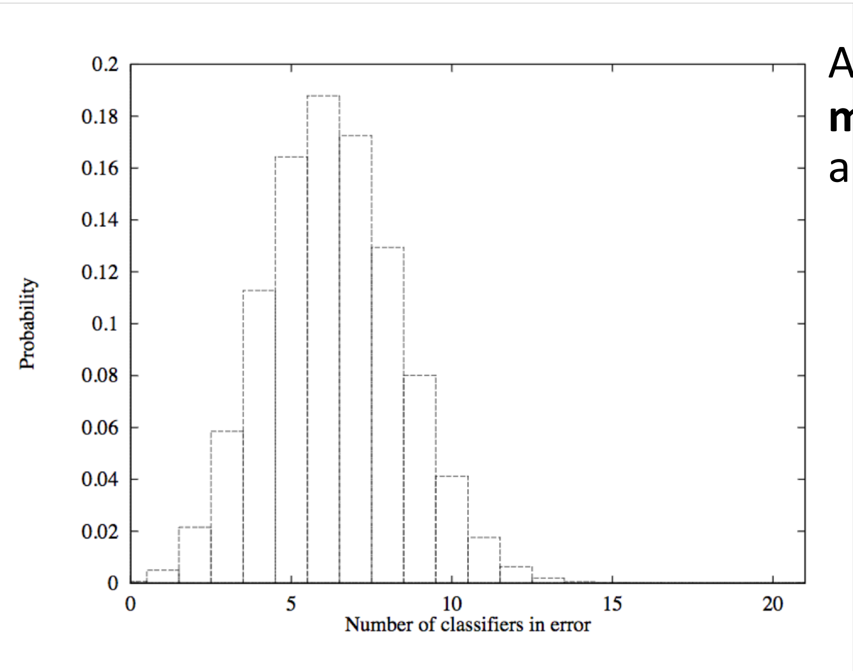
Imagine we have an ensemble for a binary prediction problem with 21 models, each with a classification error of 0.3

The big idea behind ensembles is that if we have multiple learners that are diverse, when one is wrong there is a very good chance that others are correct



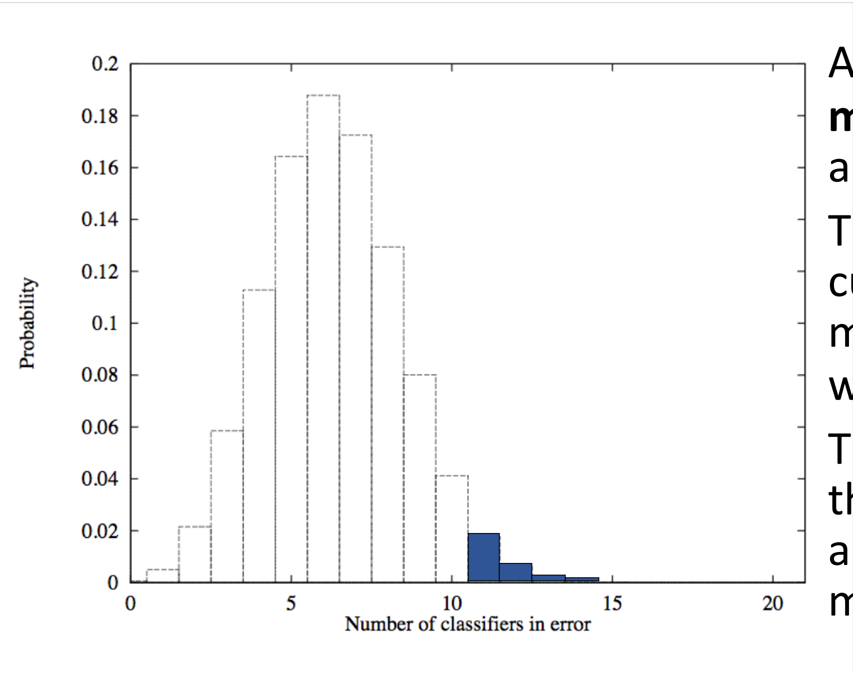
Ensembles

More formally if the error rate of each of the L models in an ensemble is less than ½ and if the errors are independent, then the probability that the majority vote of the ensemble will be wrong will be the area under the binomial distribution where more than $L/2$ models are wrong



An ensemble of **21 models** each with an error rate of **0.3**

Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf



An ensemble of **21 models** each with an error rate of **0.3**

The area under the curve for 11 or more models being wrong is **0.026**

This is much less than the error of any individual model

Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

Ensembles

But models in a real ensemble are never independent so we don't quite do that well

In general we build our ensembles to have two competing characteristics

- Individual models in the ensemble should be strong
- The correlation between the models in the ensemble should be weak (diversity)

Practical Ensembles

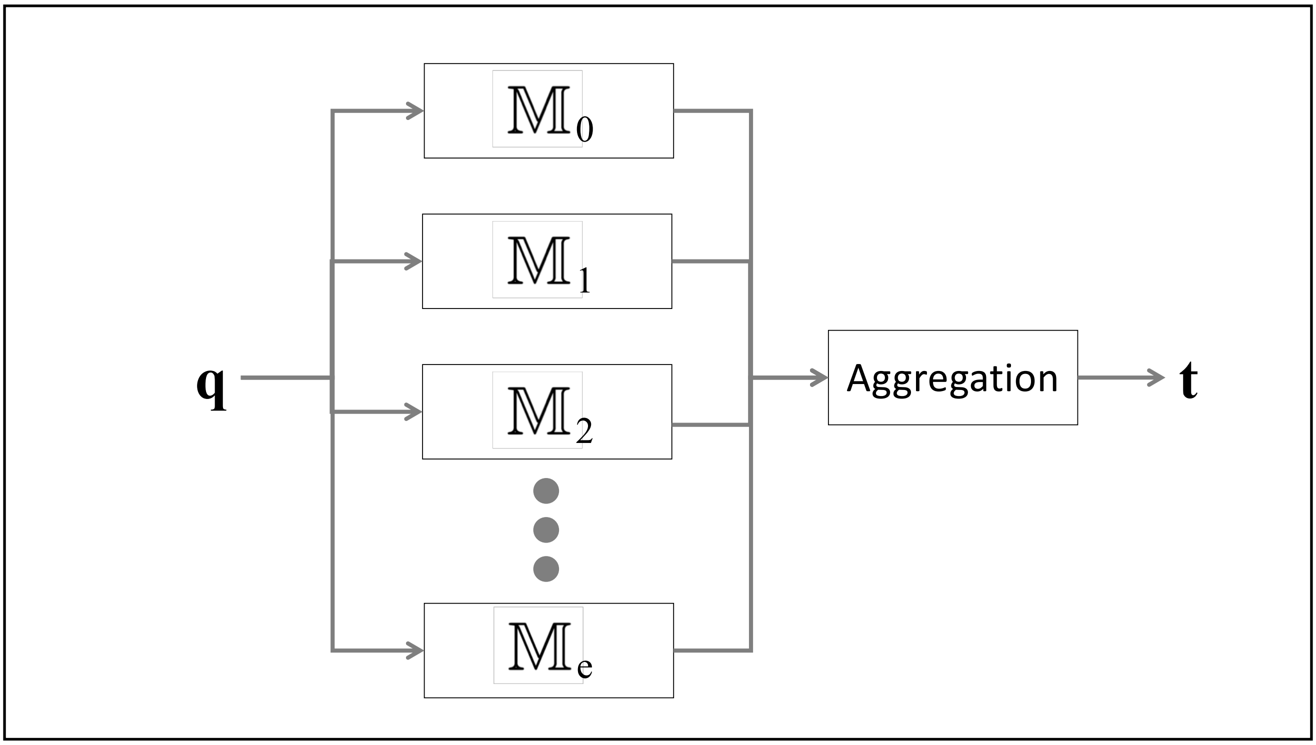
There are however a series of pracical ensemble approaches

- Bagging
- Random forests
- Boosting
- Gradient boosting
- Stacking

Practical Ensembles

There are however a series of pracical ensemble approaches

- **Bagging**
- **Random forests**
- Boosting
- Gradient boosting
- Stacking



Practical Ensembles

In general we would like our ensembles to have two characteristics

- Individual models in the ensemble should be strong
- The correlation between the models in the ensemble should be weak (diversity)

These two characteristics are in tension with each other

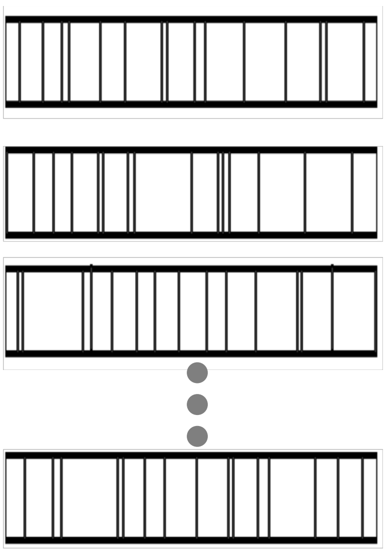
BAGGING

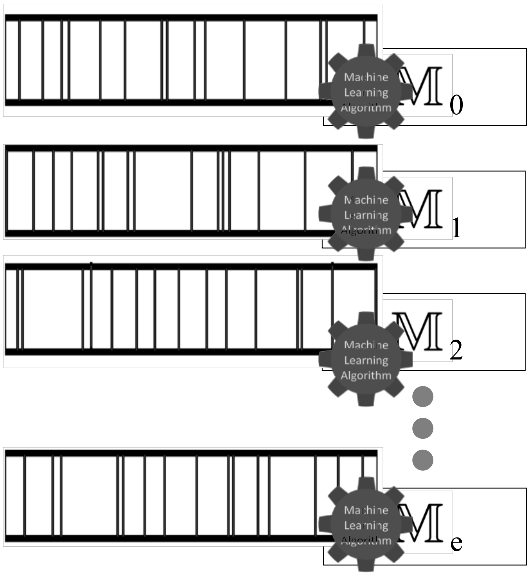
Bagging

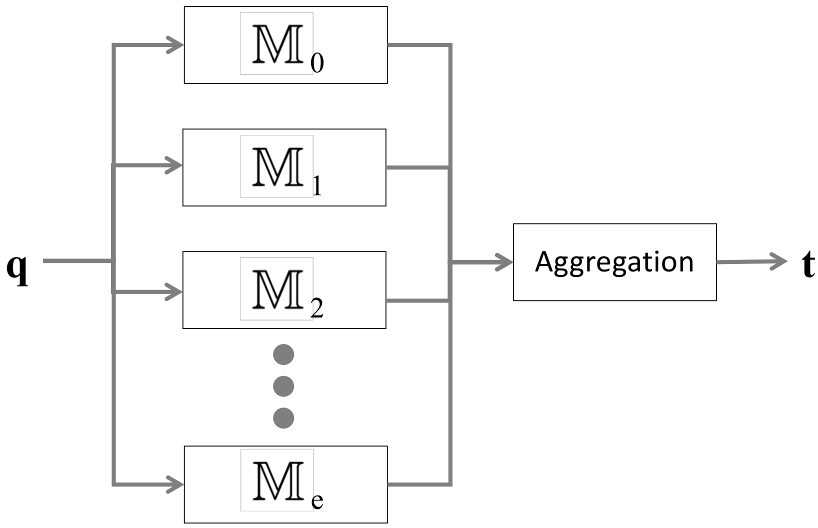
Very simple ensemble training technique

- Trains *e* models in parallel using bootstrapped data samples from an overall training set (100% sampling with replacement)
- Aggregates using majority voting
- *Boostrapped aggregating = bagging*

Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.




Machine Learning Algorithm
M 0
M 1
M 2
M e


q
M 0
M 1
M 2
M e
Aggregation
t

RANDOM FORESTS

Random Forests

Simple but very powerful ensembling technique

- Trains *e* models in parallel using bootstrapped and sub-space sampled data samples from an overall training set
- Aggregates using majority voting

Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

ID	F1	F2	F3	Target
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

ID	F1	F2	F3	Target
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-

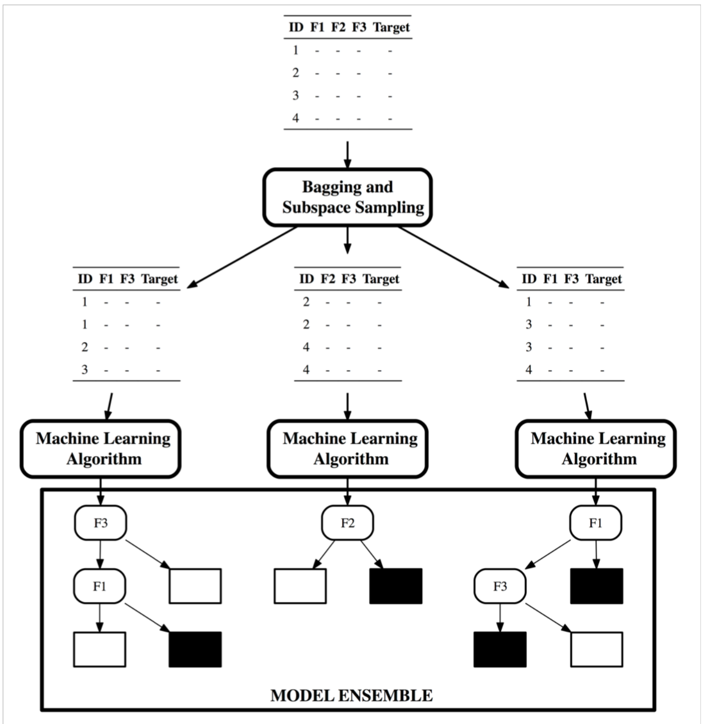
Bagging and Subspace Sampling

ID	F1	F3	Target
1	-	-	-
1	-	-	-
2	-	-	-
3	-	-	-

ID	F2	F3	Target
2	-	-	-
2	-	-	-
4	-	-	-
4	-	-	-

ID	F1	F3	Target
1	-	-	-
3	-	-	-
3	-	-	-
4	-	-	-

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

SUMMARY

Summary

- Supervised learning involves bulding prediction models that learn patterns between a set of descriptive fetaures and a target feature based on a large labelled dataset
- Training a model can be viewed as a search process
- Inductive bias is required for this search process to converge

Summary

- Ensembles are amongst the most powerful supervised learning techniques
- Random forests, in particular, are very simple but very effective

Questions

?