Inspire…Educate…Transform.

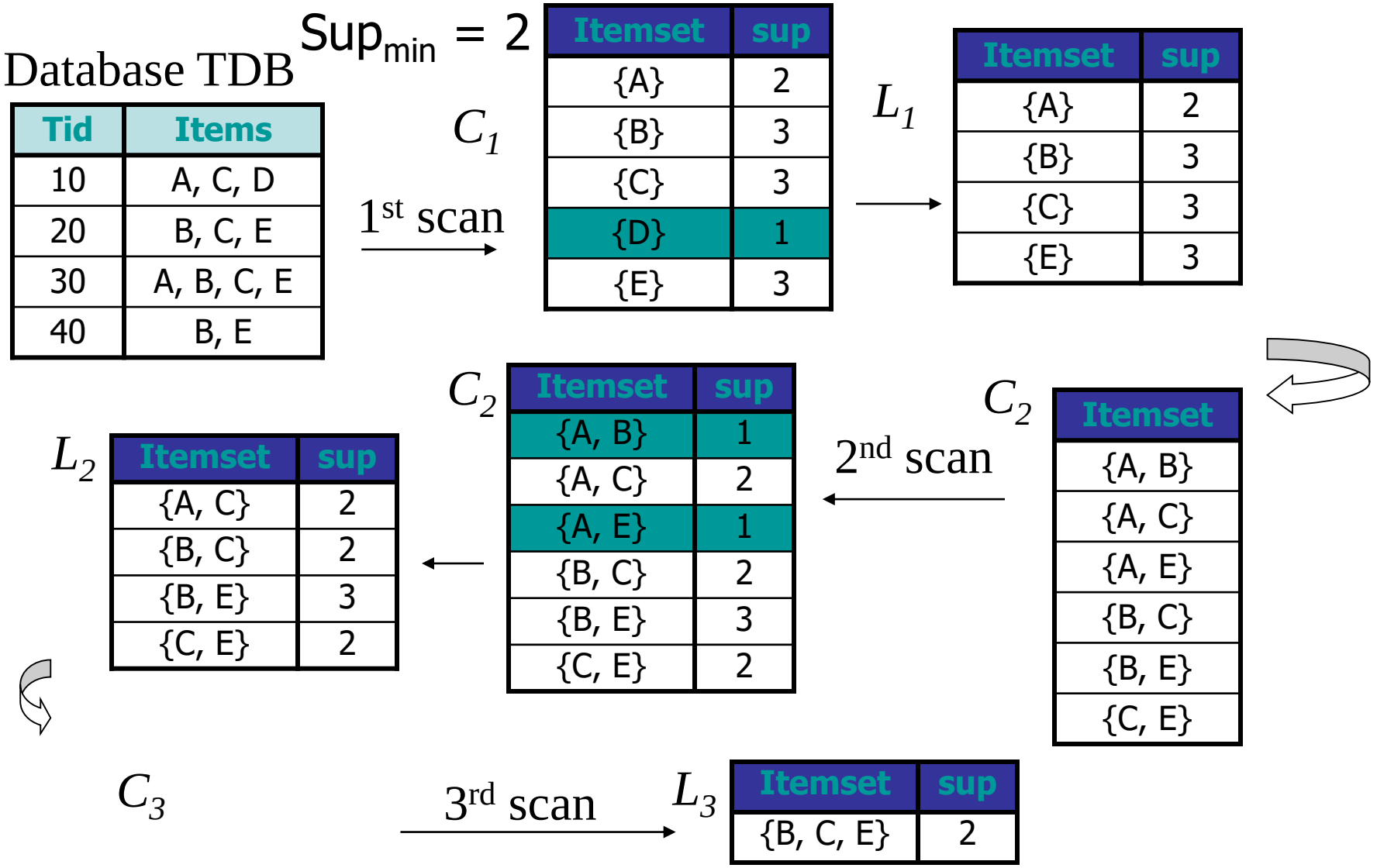
Summary – Machine Learning

Manish Gupta
Manishg.iitb@gmail.com

Topics

- Association Rule Mining
- Decision Tree
- K-Nearest Neighbor, Collaborative filtering
- Clustering
- SVM
- Stacking, Bagging, Boosting and Ensembles

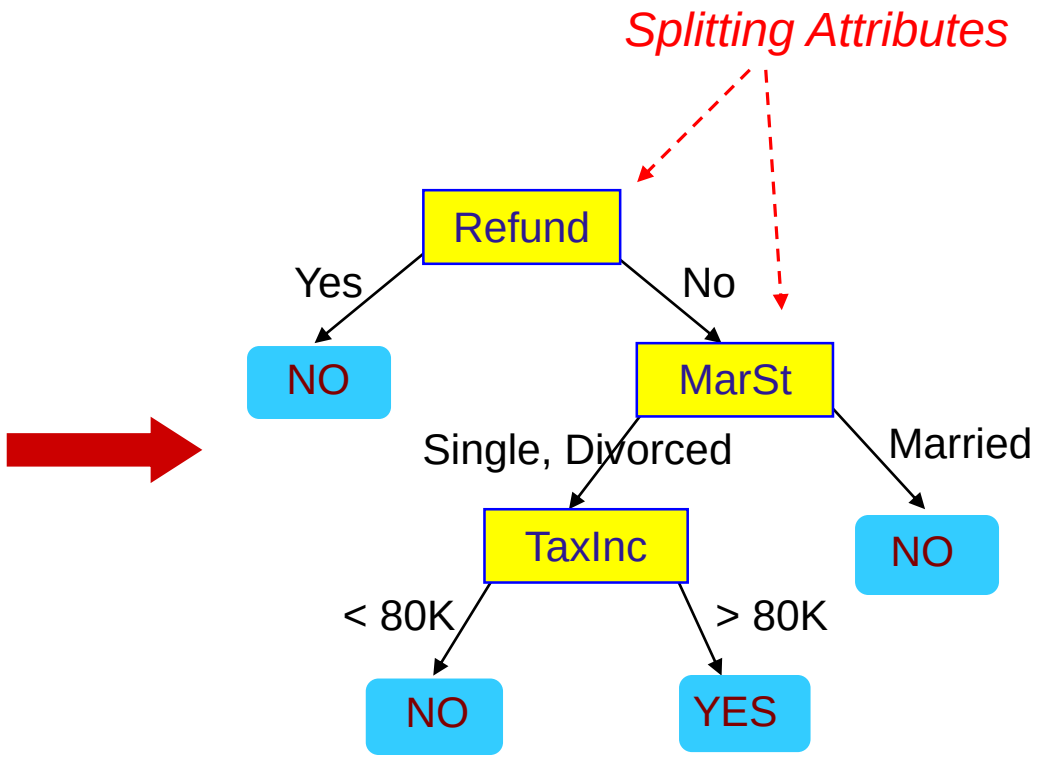
Association Rule Mining: Apriori



Decision Trees

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Training Data

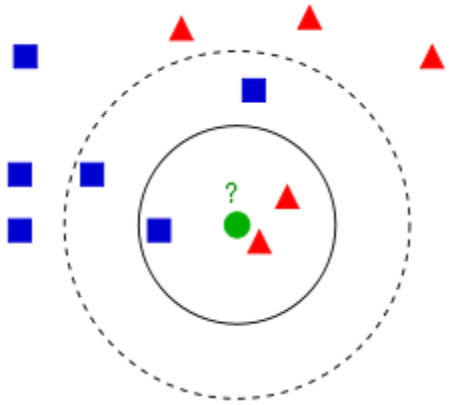


Model: Decision Tree

Decision Trees

- Impurity measures: Gini, Information gain, gain ratio
- ID3 and information gain
- C4.5 and pruning
- CART and Gini

K-Nearest Neighbor, Collaborative filtering



- Pros:
  - Simple process
  - Quick to train
- Cons
  - Curse of dimensionality
  - Slow to test
  - Requires more memory
  - Missing values need to be handled separately

K-Nearest Neighbor, Collaborative filtering

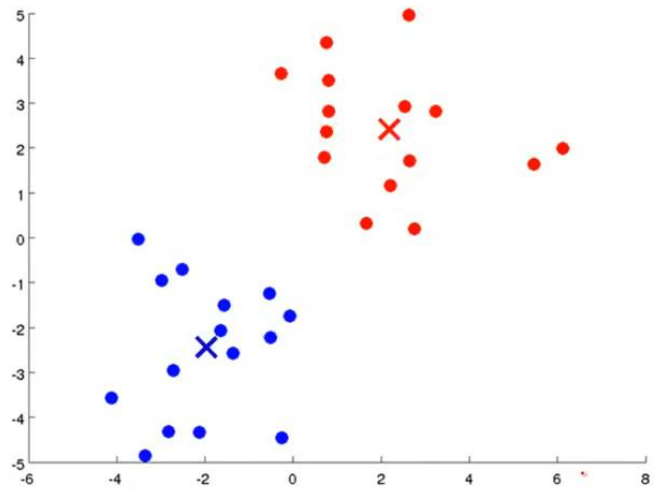
- The User x Item Matrix

	Shrek	Snow-white	Superman
Alice	Like	Like	Dislike
Bob	?	Dislike	Like
Chris	Like	Like	Dislike
John	Like	Like	?

- Shall we recommend Superman for John?
- John's taste is similar to both Chris and Alice tastes ⇒ Do not recommend Superman to Jon

Clustering

- K-means is a popular algorithm for extracting groups of similar objects from data.



Randomly initialize cluster centroids

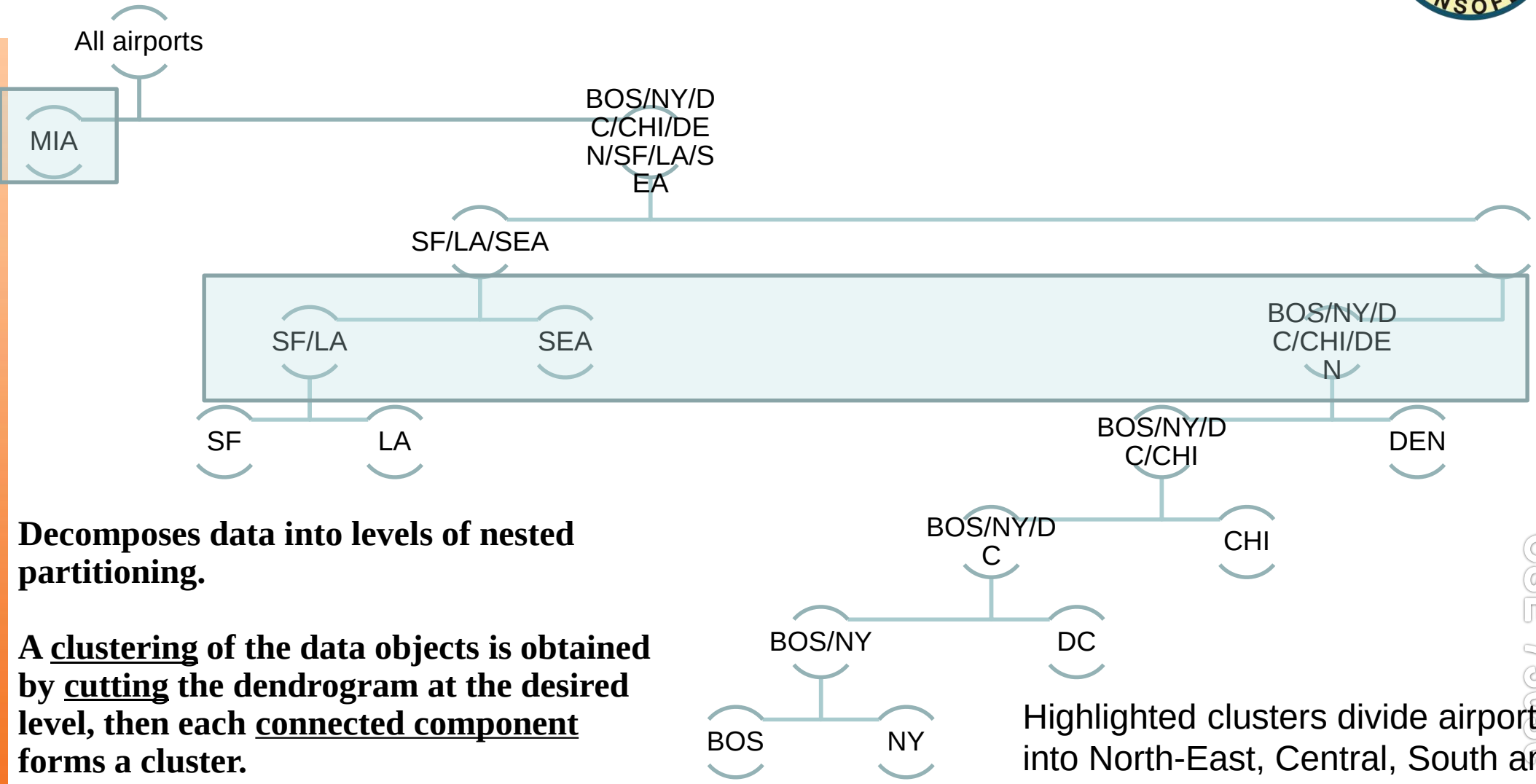
Iterate until convergence

Cluster Assignment
Assign each point to a cluster with least distance to centroid

Update centroid of cluster

CSE 7303c

Agglomerative clustering (Hierarchical)



Decomposes data into levels of nested partitioning.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

Highlighted clusters divide airports into North-East, Central, South and Pacific areas

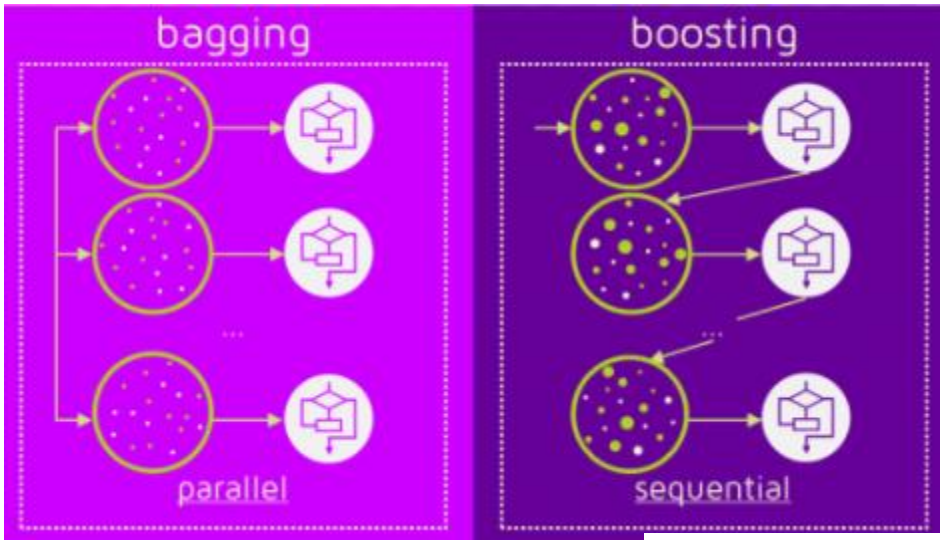
SVM

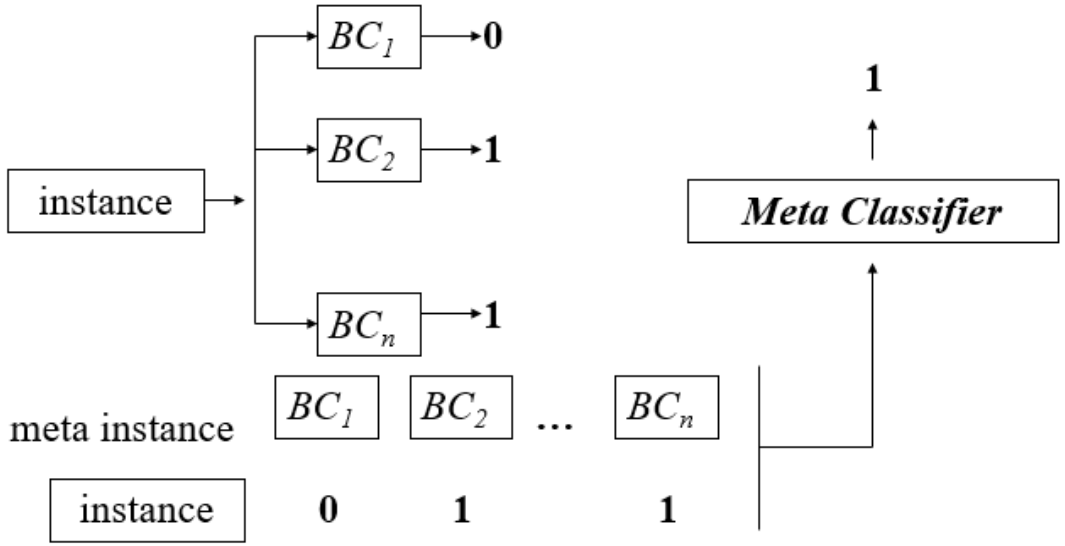
- *SVM searches for the hyperplane with the largest margin*, i.e., maximum marginal hyperplane (MMH) using constrained convex quadratic optimization
- Pros: 1. High accuracy 2. Nice theoretical guarantees regarding overfitting 3. With an appropriate kernel they can work well even if your data is not linearly separable in the base feature space 4. Good for high-dimensional data (like text)
- Cons: 1. Memory-intensive 2. Hard to interpret 3. Annoying to run and tune (long training time) 4. Not easy to incorporate domain knowledge (priors)
- Kernel Trick

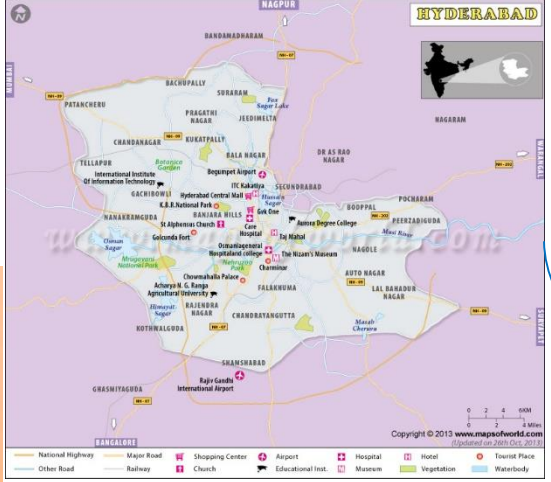
Stacking, Bagging, Boosting and Ensembles

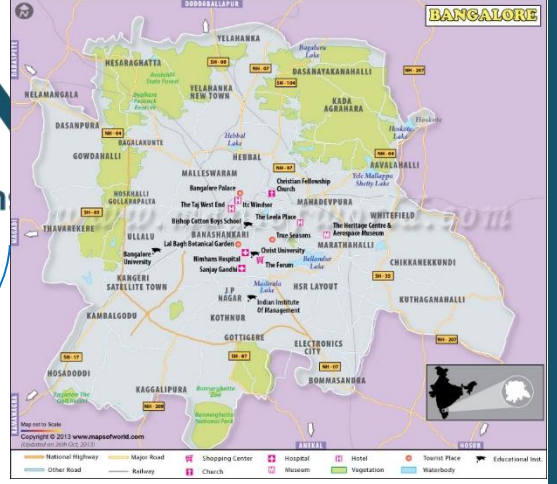
- Use a combination of models to increase accuracy
- Popular ensemble methods
  - Bagging: averaging the prediction over a collection of classifiers
  - Boosting: weighted vote with a collection of iteratively learned classifiers
    - AdaBoost
    - Gradient Boosting
  - Random Forest: Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes at each node to determine the split
- Pros: Highly effective in presence of large amount of training data
- Cons: Take long time to train

Stacking, Bagging, Boosting and Ensembles









HYDERABAD

Office and Classrooms

Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Office

Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Classroom

KnowledgeHut Solutions Pvt. Ltd., Reliable Plaza,
Jakkasandra Main Road, Teacher's Colony, 14th Main
Road, Sector – 5, HSR Layout, Bengaluru - 560102

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.