Random forests

Coneixement, Raonament i Incertesa.

http://www.cs.utexas.edu/~ear/nsc110/ScienceAndSociety/Lectures/AI-long.ppt
http://www.cs.utexas.edu/users/ear/nsc110/Mirrors/DSMirrorsArtificialIntelligence.ppt
http://decisiontrees.net/decision-trees-tutorial/

Random Forests

- 'Ensemble method' disenyat específicament per a clasificadors d'arbres de decisió
- Random Forests fa creixer molts arbres
  - Ensemble d'arbres de decisió no podats
  - Cada classificador base, classifica un nou vector d'atributs de les dades originals
  - Resultat final de classificar una nova instancia: **votació**.

El 'bosc' tria la classificació resultant que ha tingut més vots (de tots els arbres del bosc)

Random Forests

Introdueix dos fonts d'atzar: "Bagging" i "vectors d'entrada aleatoris"

- Metode Bagging: cada arbre creix usant una mostra 'bootstrap' de les dades d'aprenentatge

- Vector d'entrada aleatori: A cada node, la millor divisió s'escull entre una **mostra aleatoria de m** atributs enlloc d'entre tots els atributs

Com escollir m?

$$m = \frac{M}{3} \quad \text{si regressió}$$

$$m = \lfloor \sqrt{M} \rfloor \text{ si classificació}$$

$$m = "tunning\ parametre"$$

Algorisme Random forest

- Sigui N el número de casos d'aprenentatge, i M el número de variables de les mostres.
- Fixem un número *m* de variables d'entradada per usar en el test de la decisio en un node de l'arbre; *m* hauria de ser bastant inferior a *M*.
- Escollim un conjunt d'aprenentatge per a un arbre, escollint n vegades amb reposició d'entre els N casos d'aprenentage disponibles (i.e. Prenem una mostra bootstrap). Usem la resta dels casos per a estimar l'error del arbre, predint les seves classes.
- Per a cada node de l'arbre, aleatoriament triem *m* variables en base a les que prendrem la decisió en aquest node. Calculem la millor divisió d'aquestes *m* variables del conjunt d'aprenentatge.
- Cada arbre creix fins al límit i **NO es poda**.

Random Forests

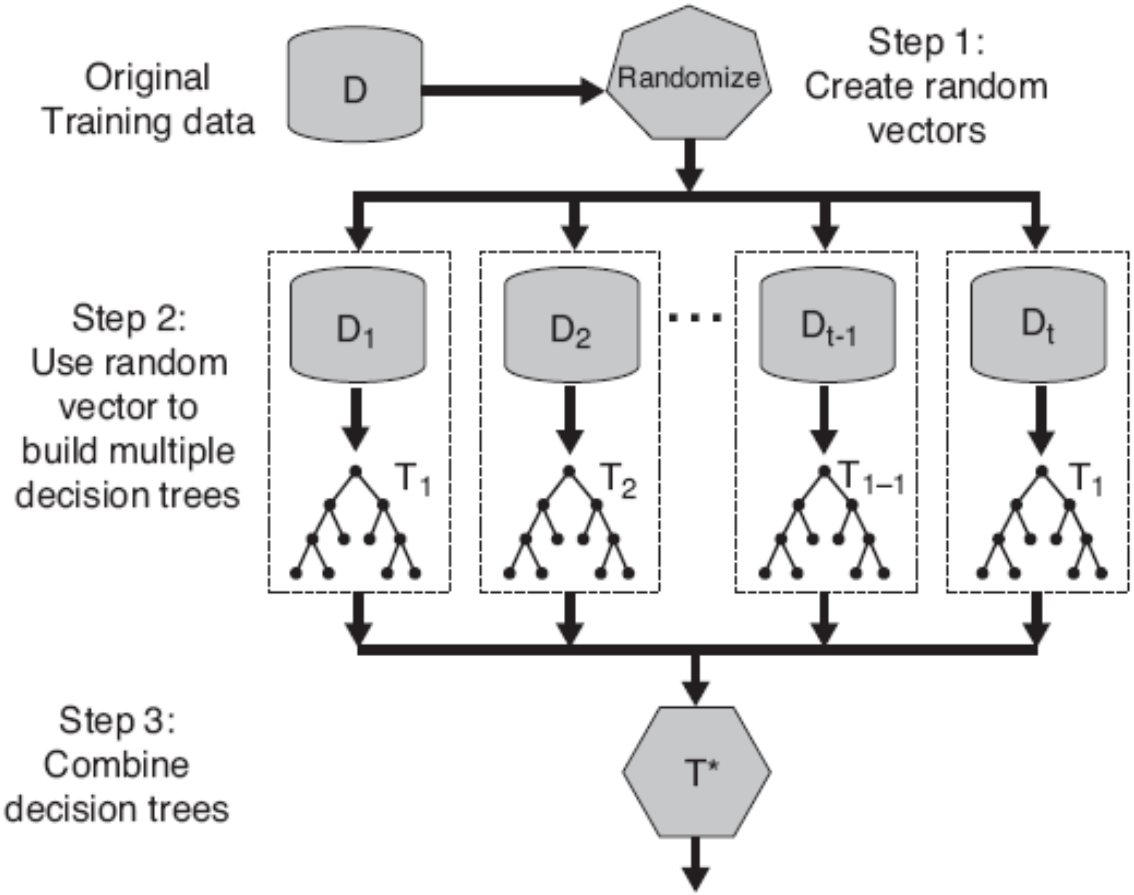
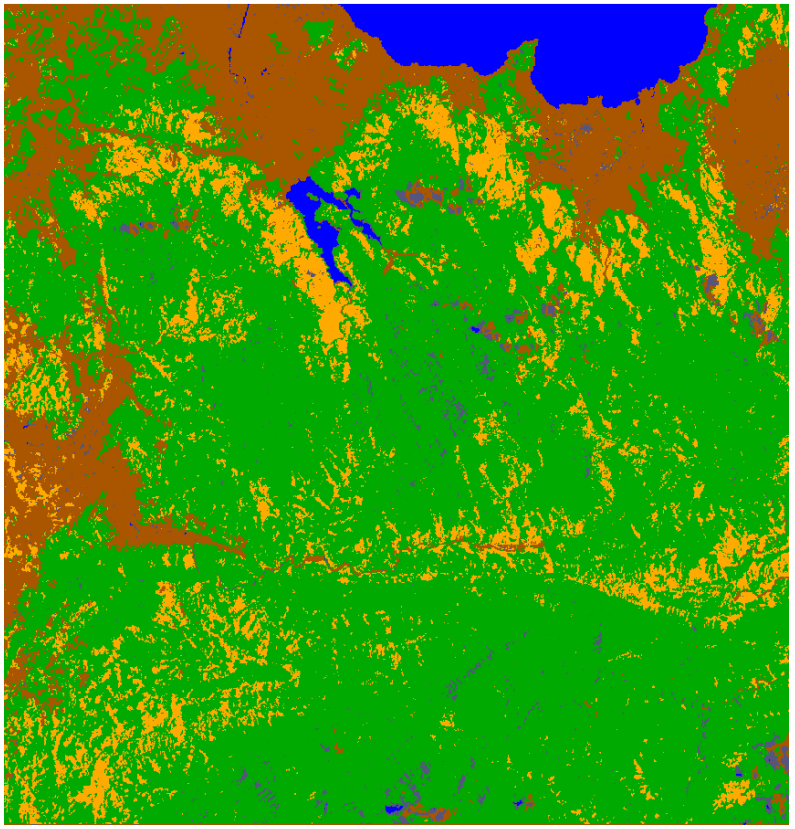


Figure 5.40. Random forests.

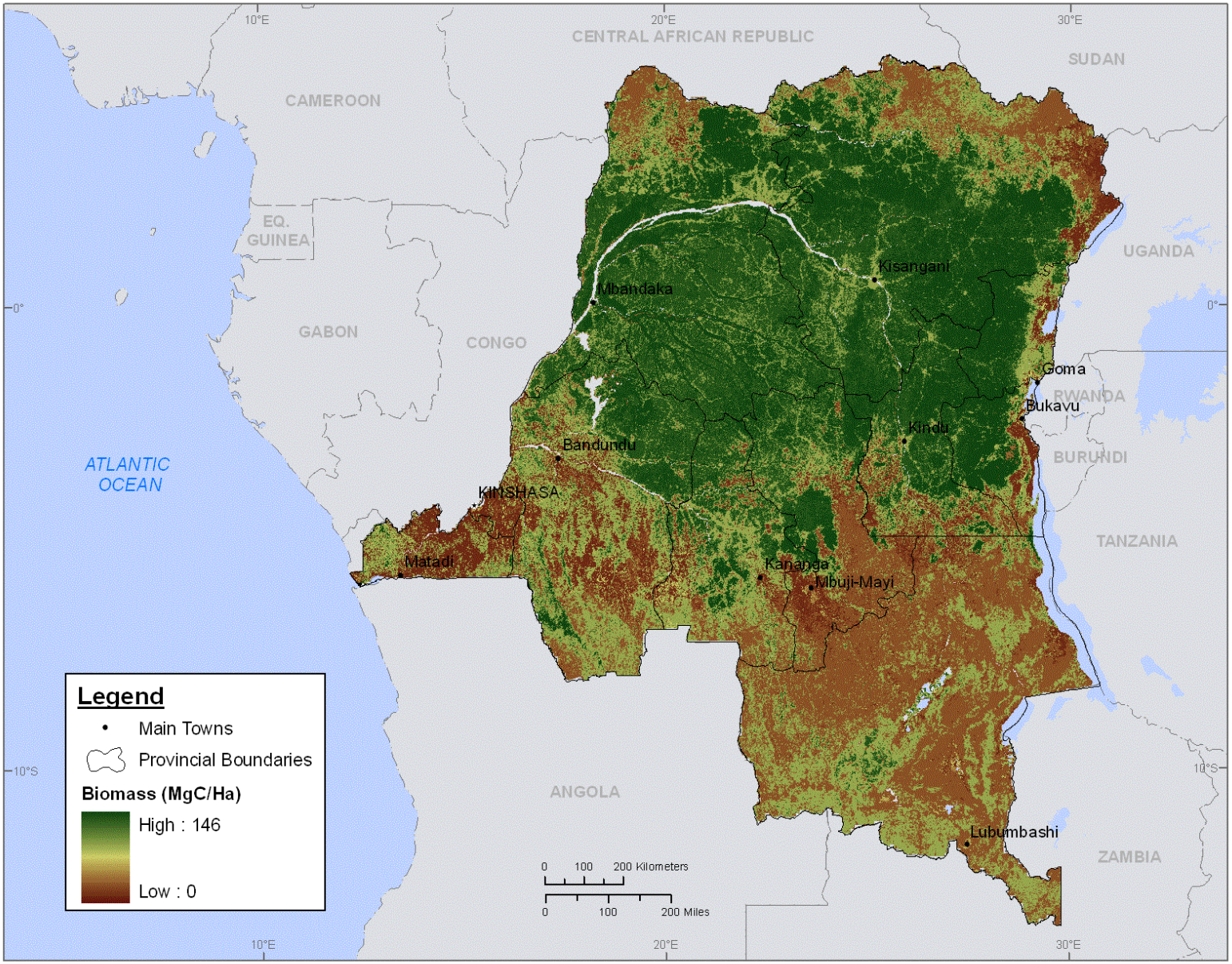
Arbre de decisió



Blue = water
Green = forest
Yellow = shrub
Brown = non-forest
Gray = cloud/shadow

Ned Horning
American Museum of Natural History's Center for Biodiversity and Conservation

Regressió


10°E
20°E
30°E
CENTRAL AFRICAN REPUBLIC
SUDAN
CAMEROON
EQ. GUINEA
GABON
CONGO
UGANDA
0°
Kisangani
Mbandaka
Goma
RWANDA
Bukavu
Kindu
Bandundu
BURUNDI
ATLANTIC OCEAN
KINSHASA
TANZANIA
Matadi
Kananga
Mbuji-Mayi
10°S
ANGOLA
Lubumbashi
ZAMBIA
Legend
Main Towns
Provincial Boundaries
Biomass (MgC/Ha)
High : 146
Low : 0
0 100 200 Kilometers
0 100 200 Miles

Efecte de la població del bosc

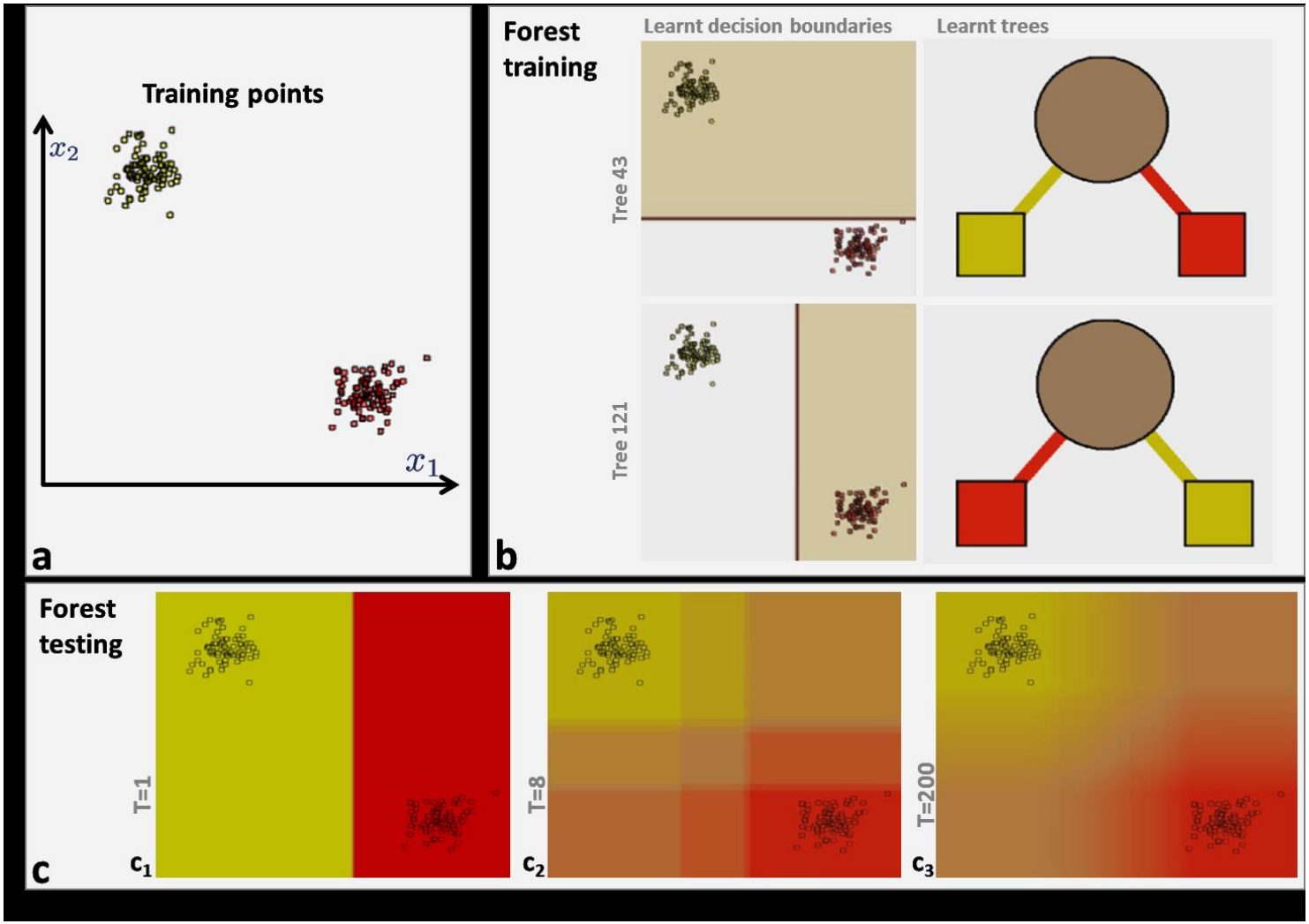


Fig. 3.3: A first classification forest and the effect of forest size T. (a) Training points belonging to two classes. (b) Different training trees produce different partitions and thus different leaf predictors. The colour of tree nodes and edges indicates the class probability of training points going through them. (c) In testing, increasing the forest size T produces smoother class posteriors. All experiments were run with D = 2 and axis-aligned weak learners. See text for details.

Random forest i problemes multiclase

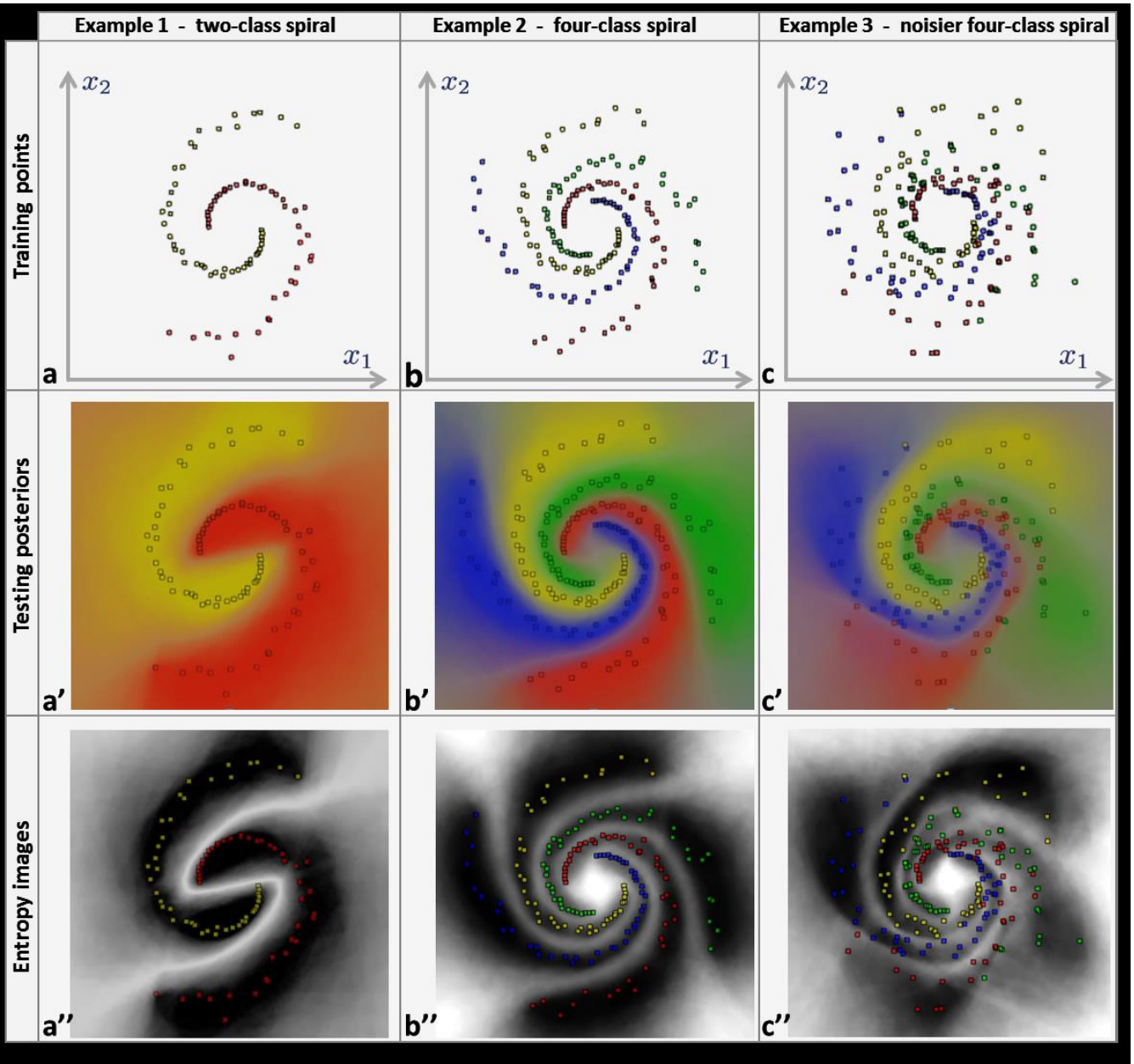


Fig. 3.4: The effect of multiple classes and noise in training data. (a,b,c) Training points for three different experiments: 2-class spiral, 4-class spiral and another 4-class spiral with noisier point positions, respectively. (a',b',c') Corresponding testing posteriors. (a",b",c") Corresponding entropy images (brighter for larger entropy). The classification forest can handle both binary as well as multiclass problems. With larger training noise the classification uncertainty increases (less saturated colours in c' and less sharp entropy in c"). All experiments in this figure were run with T = 200, D = 6, and a conic-section weak-learner model.