Artificial intelligence in data science

Decision tree random forest

Janos Török

Department of Theoretical Physics

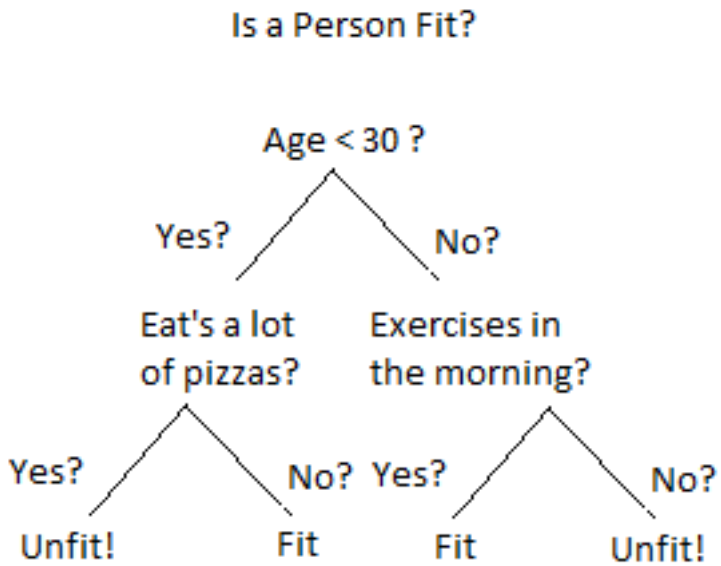
September 22, 2020

Decision tree, random forest, hierarchical clustering

Why?

- Unsupervised learning
- Importance of parameters

Decision tree



- Build a tree
- Nodes are yes-no questions
- Links are answers (yes/no)
- Leaves are classification statements

Decision tree

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

- Which parameter to pick first?
- The one which classifies the data best
- What is *best*? $\rightarrow$ information gain or Gini index

Information entropy

- $H(s) = \sum_{c \in C} -p(c) \log_2 p(c)$, $C = \{\text{yes}, \text{no}\}$
- For the full set:
- 9 out of 14 are yes:

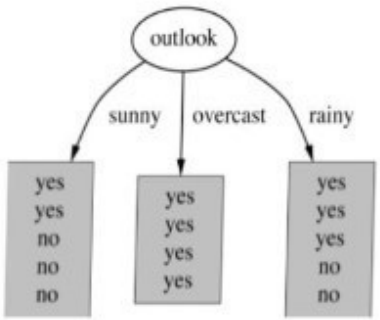
$$H(s) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.41 + 0.53 = 0.94$$

- Information entropy for perfectly separated $H = 0$, information entropy of perfectly mixed system $H = 1$

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Information gain, for every feature:

Information entropy of the original minus the one of the divided



E (Outlook=sunny) = $-\frac{2}{5}\log\left(\frac{2}{5}\right)-\frac{3}{5}\log\left(\frac{3}{5}\right)=0.971$

E (Outlook=overcast) = $-1\log(1)-0\log(0)=0$

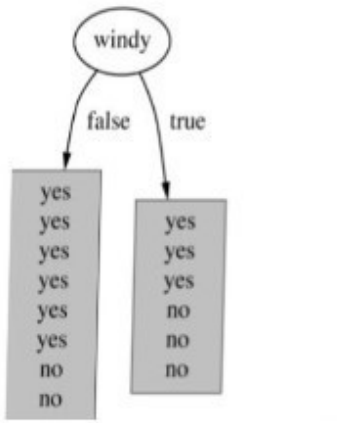
E (Outlook=rainy) = $-\frac{3}{5}\log\left(\frac{3}{5}\right)-\frac{2}{5}\log\left(\frac{2}{5}\right)=0.971$

} H(S,Outlook)

Average Entropy information for Outlook

I (Outlook) = $\frac{5}{14}*0.971+\frac{4}{14}*0+\frac{5}{14}*0.971=0.693$ } $\sum_{t\in T} p(t)H(t)$

Gain (Outlook) = E(S) – I (outlook) = 0.94-.693 = 0.247 ⇨ $IG(A,S)=H(S)-\sum_{t\in T}p(t)H(t)$



E (Windy=false) = $-\frac{6}{8}\log\left(\frac{6}{8}\right)-\frac{2}{8}\log\left(\frac{2}{8}\right)=0.811$

E (Windy=true) = $-\frac{3}{6}\log\left(\frac{3}{6}\right)-\frac{3}{6}\log\left(\frac{3}{6}\right)=1$

Average entropy information for Windy

I (Windy) = $\frac{8}{14}*0.811+\frac{6}{14}*1=0.892$

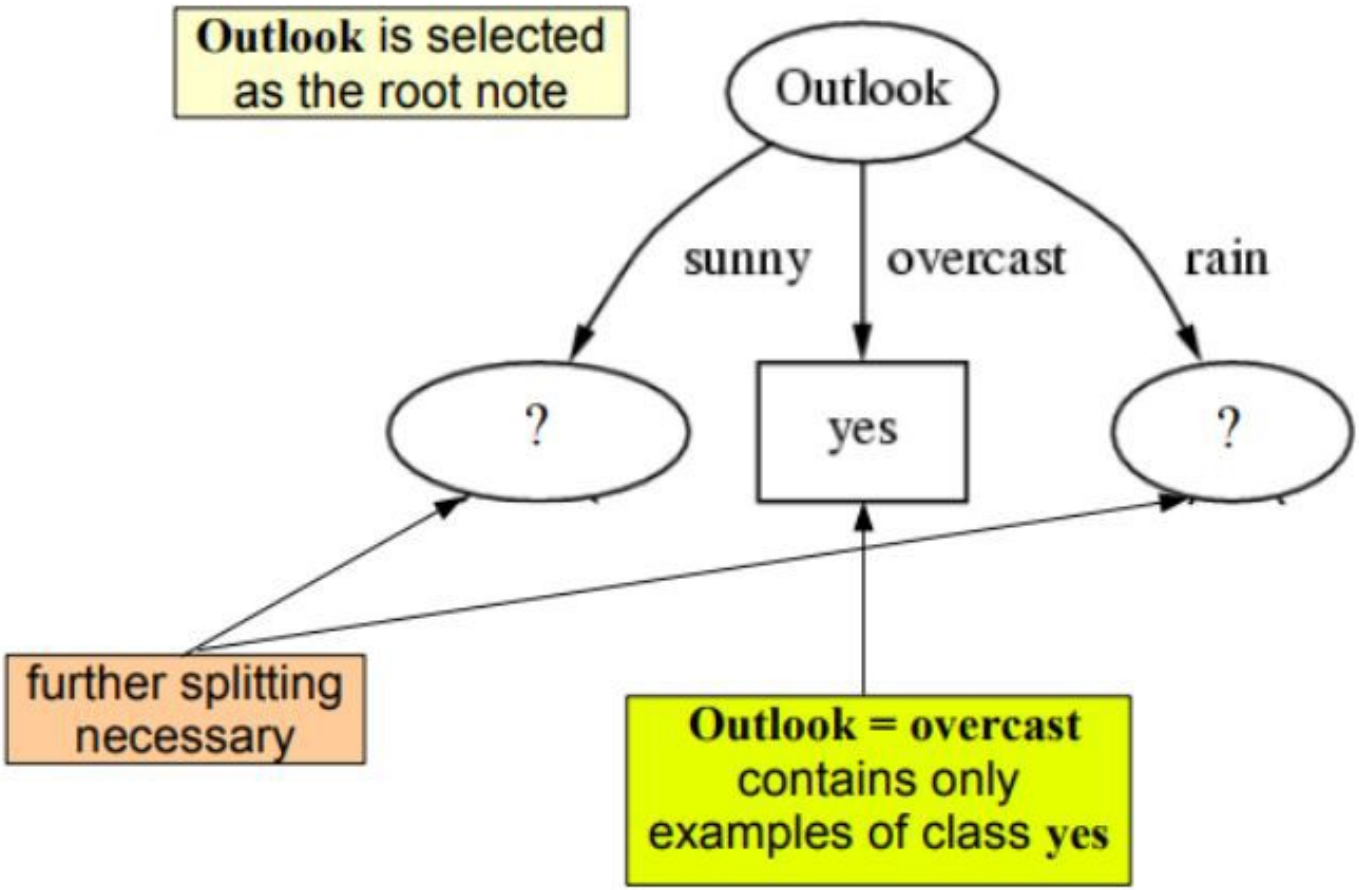
Gain (Windy) = E(S) – *I* (Windy) = 0.94-0.892=0.048

Information gain, for every feature, pick the highest:

Outlook		Temperature	
Info:	0.693	Info:	0.911
Gain: 0.940-0.693	0.247	Gain: 0.940-0.911	0.029
Humidity		**Windy**	
Info:	0.788	Info:	0.892
Gain: 0.940-0.788	0.152	Gain: 0.940-0.892	0.048

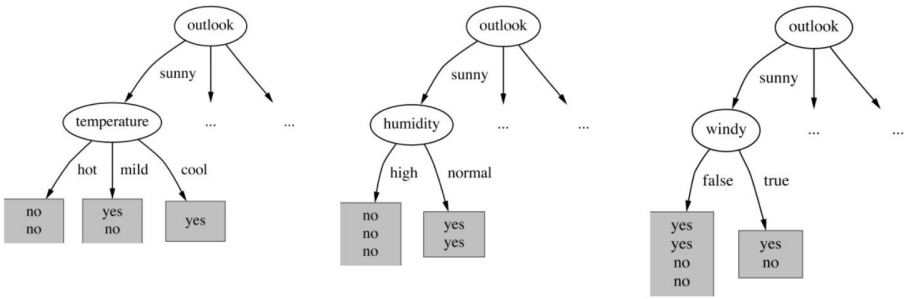
- So root node is Outlook.

Decision tree: First level



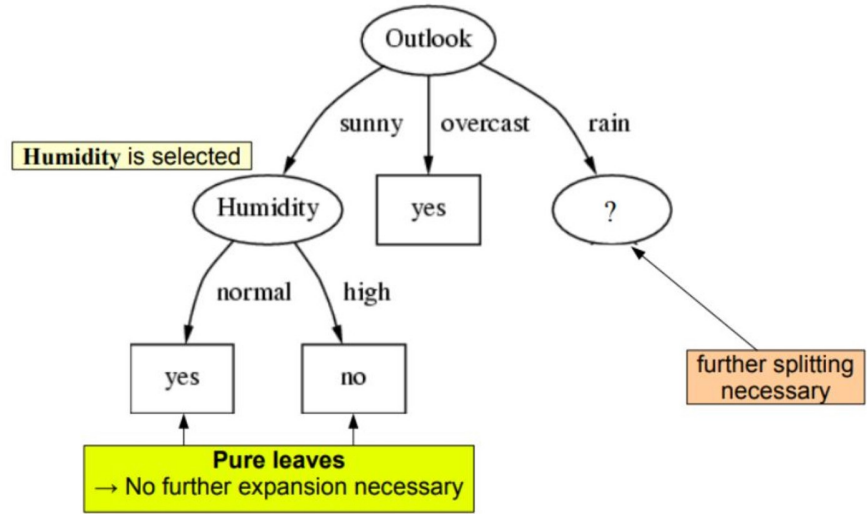
- So root node is Outlook.

Decision tree: Next levels, same procedure



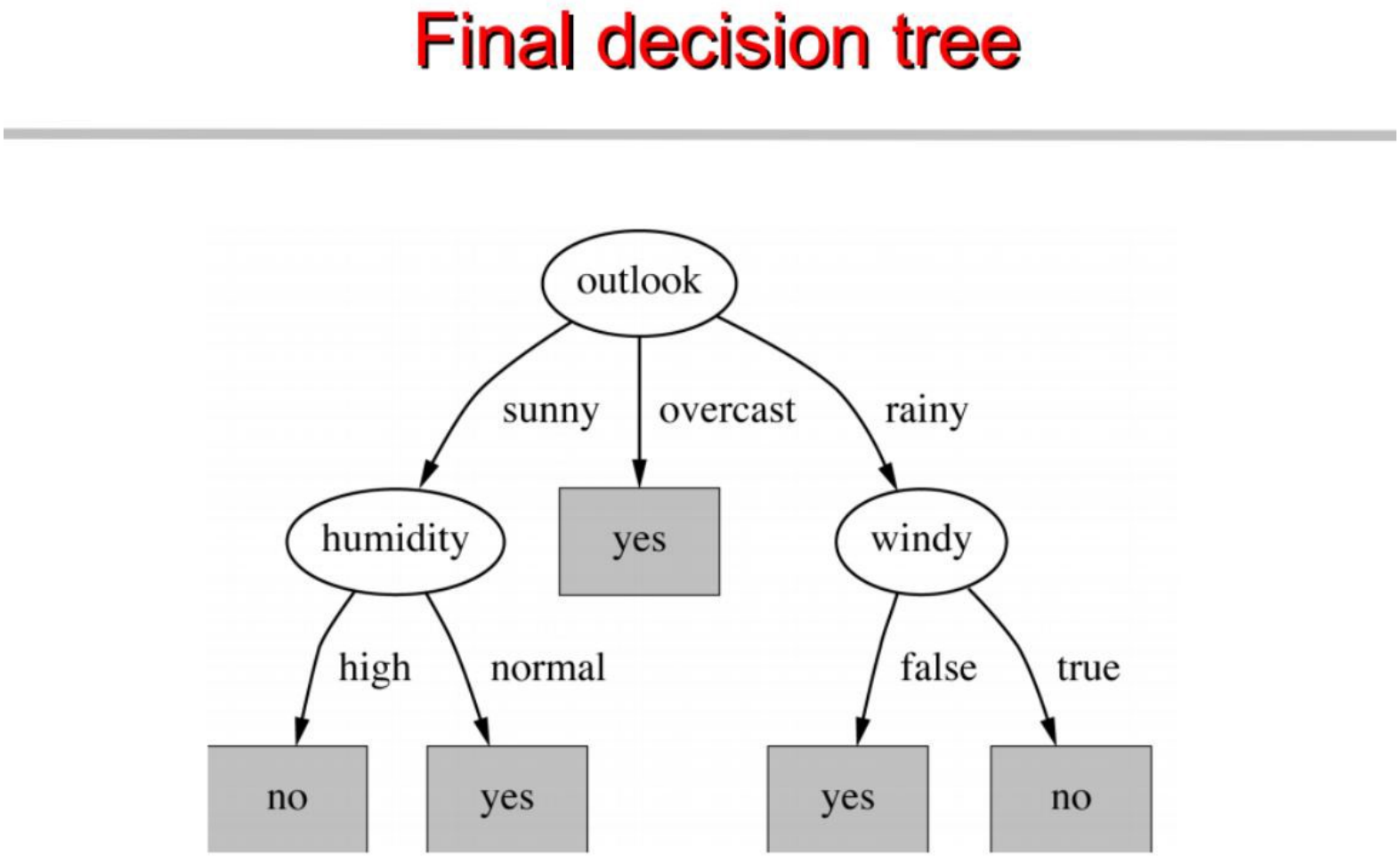
Gain(*Temperature*) = 0.571 bits
Gain(*Humidity*) = 0.971 bits
Gain(*Windy*) = 0.020 bits

Humidity is selected



- Next question is about Humidity.

Final decision tree



Gini index

- $Gini = 1 - \sum_{c \in C} p(c)^2$, $C = \{\text{yes}, \text{no}\}$
- For the full set:
- 9 out of 14 are yes:

$$Gini = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.46$$

- For perfectly separated sample Gini index is zero.

Gini index, for two groups

- Fraction weighted sum of respective Gini indices
- Example:

Class	A	A	A	A	A	B	B	B	B	B
v	0	0	0	0	1	1	1	0	1	0

- v=1: $Gini(1) = 1 - (1/4)^2 - (3/4)^2$
- v=0: $Gini(0) = 1 - (4/6)^2 - (2/6)^2$
- Combined Gini:

$$Gini = \frac{4}{10} Gini(1) + \frac{6}{10} Gini(0)$$

Decision tree

- Advantages
  - Fast
  - Easy to interpret
  - Can be combined with other techniques
- Disadvantages
  - Very unstable (small change in the data, enormous change in the tree)
  - Very inaccurate
  - Separation lines parallel to axes

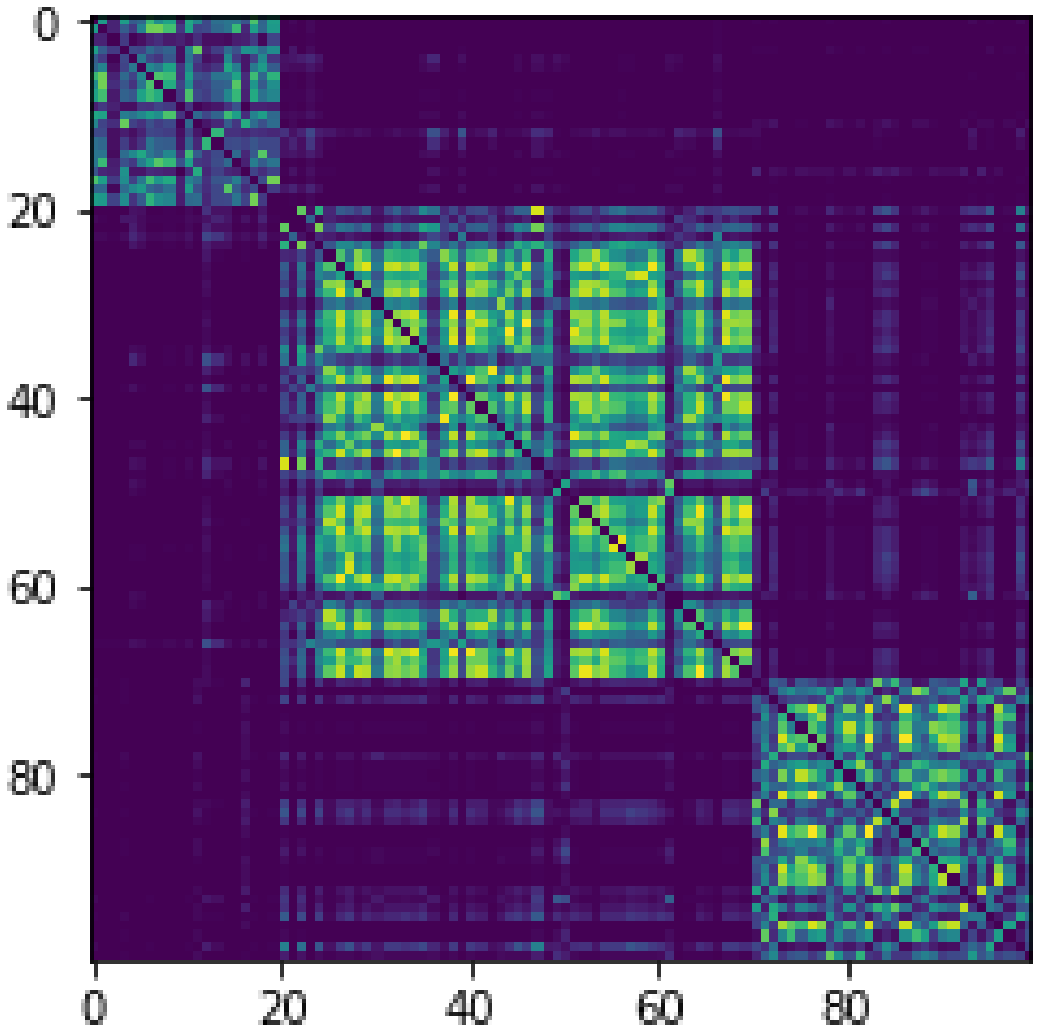
Random forest

- Bagging trees (Bootstrap Aggregating)
  - Bagging: Average a given procedure over many samples to reduce the variance
  - Draw bootstrap samples from the the original sample and to the training. Original dataset: `x = c(x1, x2, ..., x100)` Bootstrap samples: `boot1 = sample(x, 100, replace = True)`,
  - Average the results
- Random forest
  - When selecting the random sample fewer data is used
  - Average the prediction of each tree

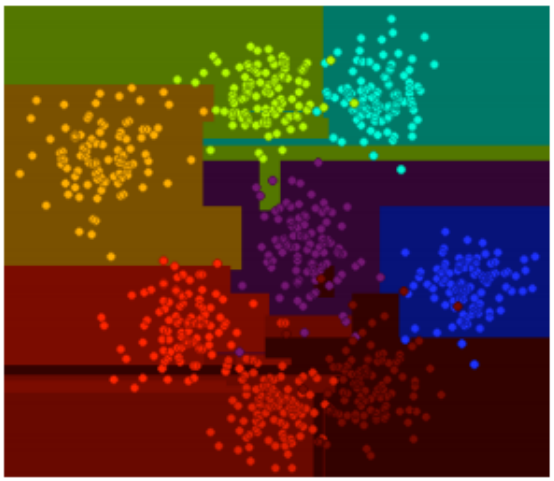
Random forest unsupervised

- How to make a decision tree without target?
  - Create a synthetic data set
  - Mark the original dataset with target 1 and the synthetic with target 0
  - Use random forest to find dissimilarity between the random and the real data.
- After each decision tree is trained, fit the original dataset
- Points ending up in the same leaf are related.
- Aggregating this events creates a similarity matrix.
- Can use other methods to cut them into pieces

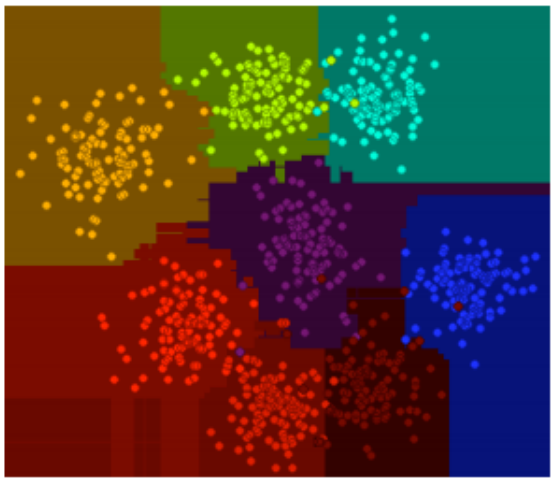
Unsupervised random forest similarity matrix



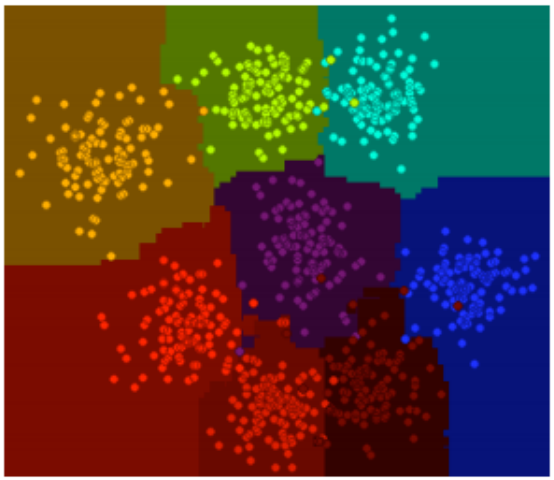
Unsupervised random forest: Illustration



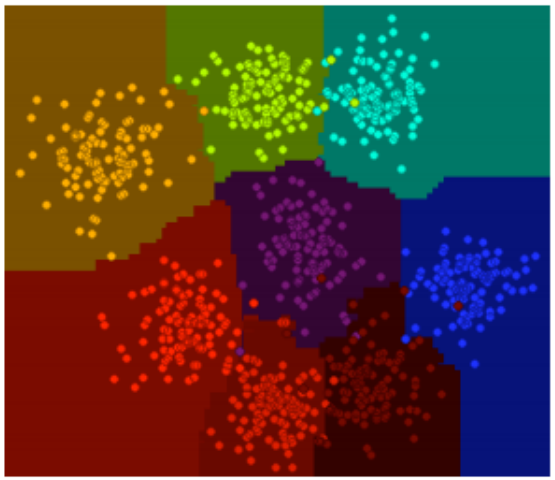
1 rCART



10 rCARTs



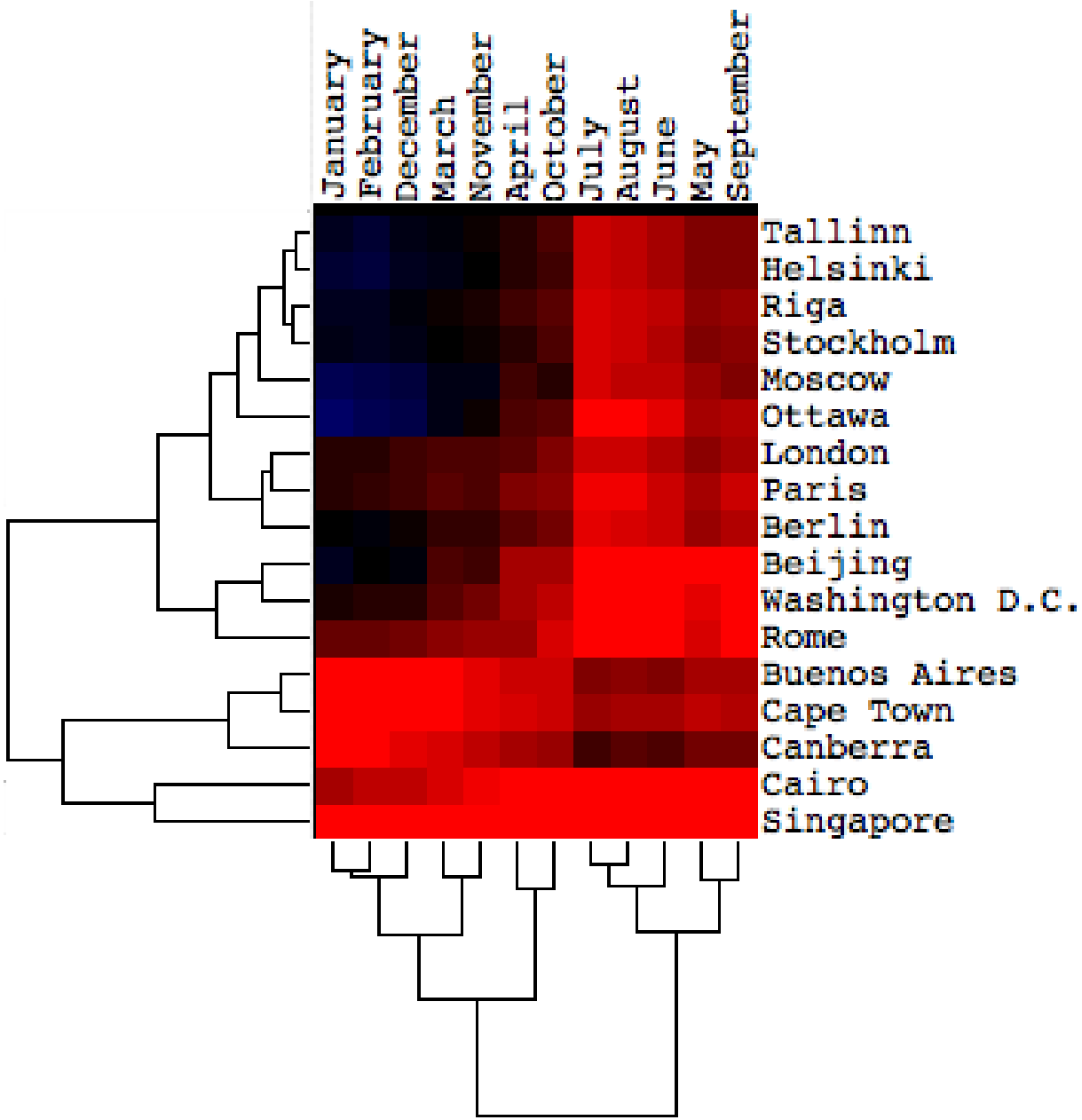
100 rCARTs



500 rCARTs

From: Eric Debreuve / Team Morpheme University Nice Sophia Antipolis

Hierarchical clustering



Hierarchical clustering

1. Define a norm between nodes $d(a, b)$
2. At the beginning each node is a separate cluster
3. Merge the two closest clusters into one
4. Repeat 3.

Norm between clusters $||A - B||$

- Maximum or complete linkage clustering:

$$\max\{d(a, b) : a \in A, b \in B\}$$

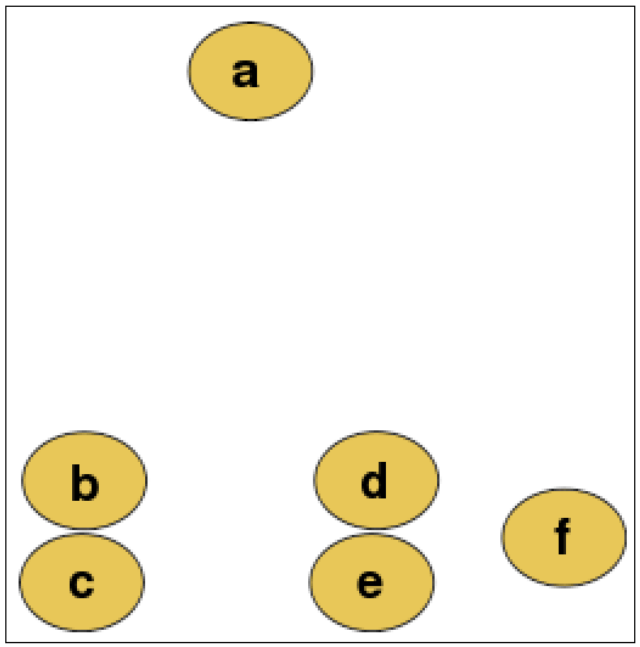
- Minimum or single-linkage clustering:

$$\min\{d(a, b) : a \in A, b \in B\}$$

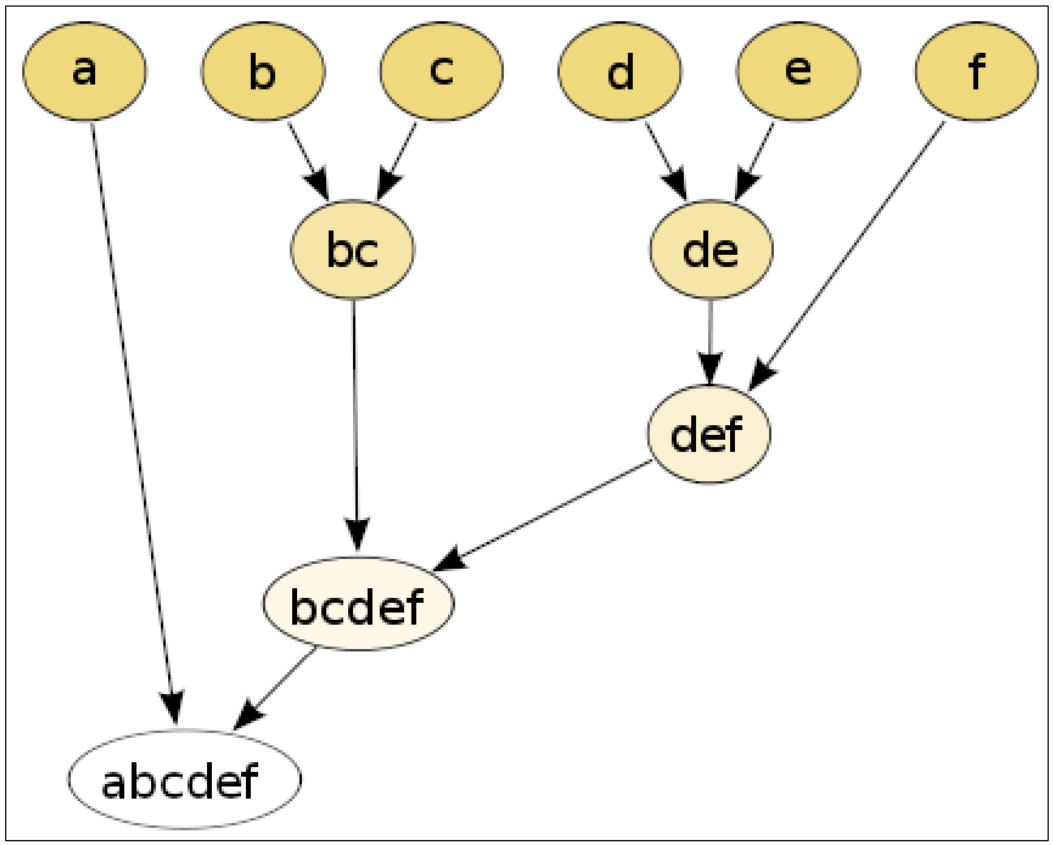
- Mean or average linkage clustering:

$$\frac{1}{||A||\,||B||} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

Hierarchical clustering



Original

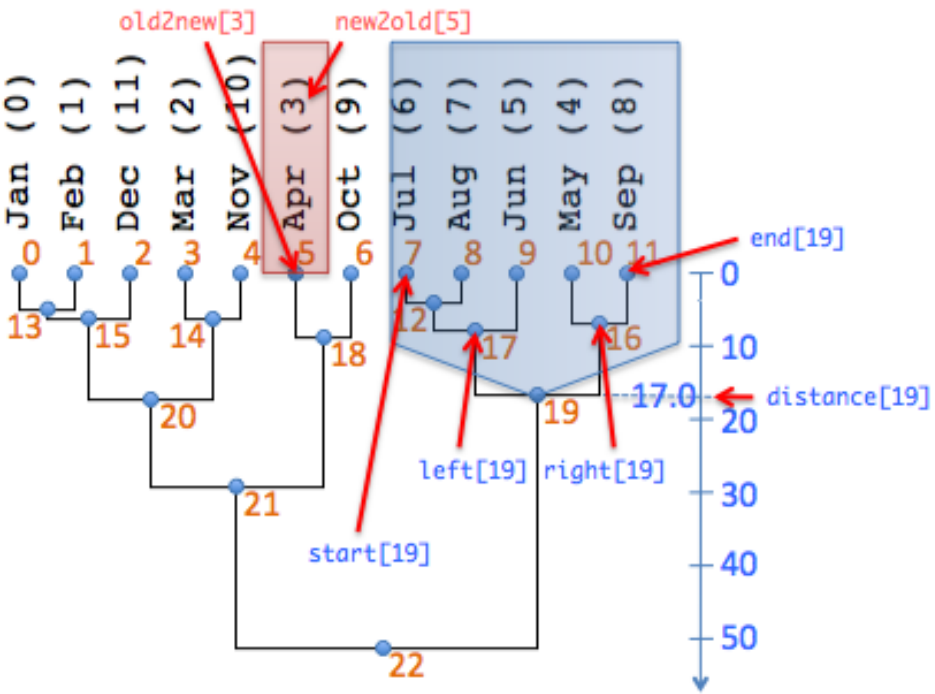


Dendogram

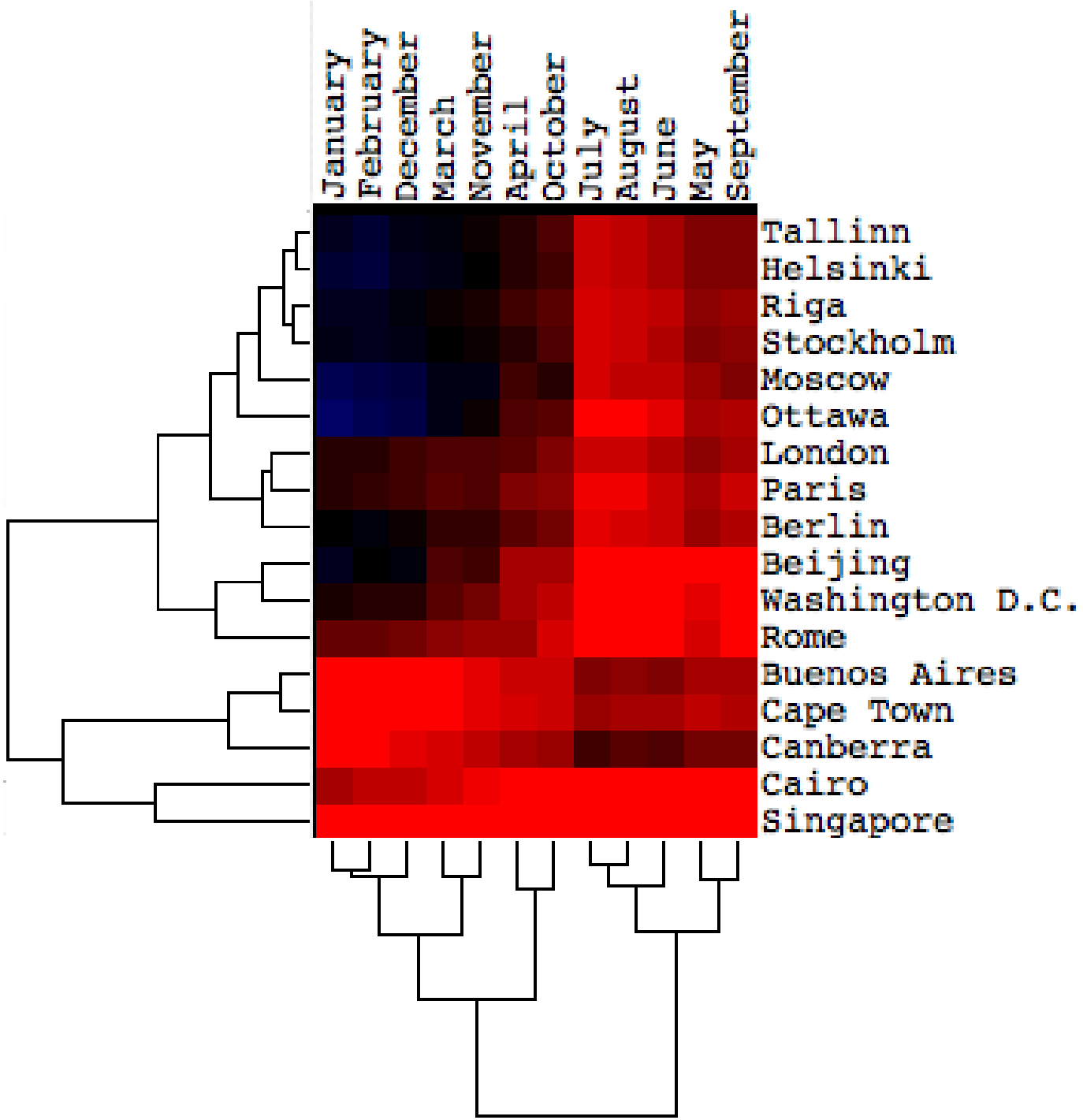
Example: Temperatures in capitals

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tallinn	-3	-5	-1	3	10	13	16	15	10	6	1	-2
Beijing	-3	0	6	13	20	24	26	25	20	13	5	-1
Berlin	0	-1	4	7	12	16	18	17	14	9	4	1
Buenos Aires	23	22	20	16	13	10	10	11	13	16	18	22
Cairo	13	15	17	21	25	27	28	27	26	23	19	15
Canberra	20	20	17	13	9	6	5	7	9	12	15	18
Cape Town	21	21	20	17	15	13	12	13	14	16	18	20
Helsinki	-5	-6	-2	3	10	13	16	15	10	5	0	-3
London	3	3	6	7	11	14	16	16	13	10	6	5
Moscow	-8	-7	-2	5	12	15	17	15	10	3	-2	-6
Ottawa	-10	-8	-2	6	13	18	21	20	14	7	1	-7
Paris	3	4	7	10	13	16	19	19	16	11	6	5
Riga	-3	-3	1	5	11	15	17	16	12	7	2	-1
Rome	8	8	11	12	17	20	23	23	21	17	12	9
Singapore	27	27	28	28	28	28	28	28	27	27	27	26
Stockholm	-2	-3	0	3	10	14	17	16	11	6	1	-2
Washington D.C.	2	3	7	13	18	23	26	25	21	15	9	3

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tallinn	-3	-5	-1	3	10	13	16	15	10	6	1	-2
Beijing	-3	0	6	13	20	24	26	25	20	13	5	-1
Berlin	0	-1	4	7	12	16	18	17	14	9	4	1
Buenos Aires	23	22	20	16	13	10	10	11	13	16	18	22
Cairo	13	15	17	21	25	27	28	27	26	23	19	15
Canberra	20	20	17	13	9	6	5	7	9	12	15	18
Cape Town	21	21	20	17	15	13	12	13	14	16	18	20
Helsinki	-5	-6	-2	3	10	13	16	15	10	5	0	-3
London	3	3	6	7	11	14	16	16	13	10	6	5
Moscow	-8	-7	-2	5	12	15	17	15	10	3	-2	-6
Ottawa	-10	-8	-2	6	13	18	21	20	14	7	1	-7
Paris	3	4	7	10	13	16	19	19	16	11	6	5
Riga	-3	-3	1	5	11	15	17	16	12	7	2	-1
Rome	8	8	11	12	17	20	23	23	21	17	12	9
Singapore	27	27	28	28	28	28	28	28	27	27	27	26
Stockholm	-2	-3	0	3	10	14	17	16	11	6	1	-2
Washington D.C.	2	3	7	13	18	23	26	25	21	15	9	3



Euclidean distance


January
February
December
March
November
April
October
July
August
June
May
September
Tallinn
Helsinki
Riga
Stockholm
Moscow
Ottawa
London
Paris
Berlin
Beijing
Washington D.C.
Rome
Buenos Aires
Cape Town
Canberra
Cairo
Singapore

Hierarchical clustering: problems

- Advantages
  - Simple
  - Fast
  - Number of clusters can be controlled
  - Hierarchical relationship
- Disadvantages
  - No a priori cutting level
  - Meaning of clusters unclear
  - Important links may be missed
  - Different result if one item omitted