Decision Trees

Lecturer:
Jiří Matas

Authors:
Jiří Matas, Ondřej Drbohlav

Centre for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

7.1.2016

Pros and Cons of Decision Trees

Pros

1. simple, intuitive
2. no restrictions on the feature space:

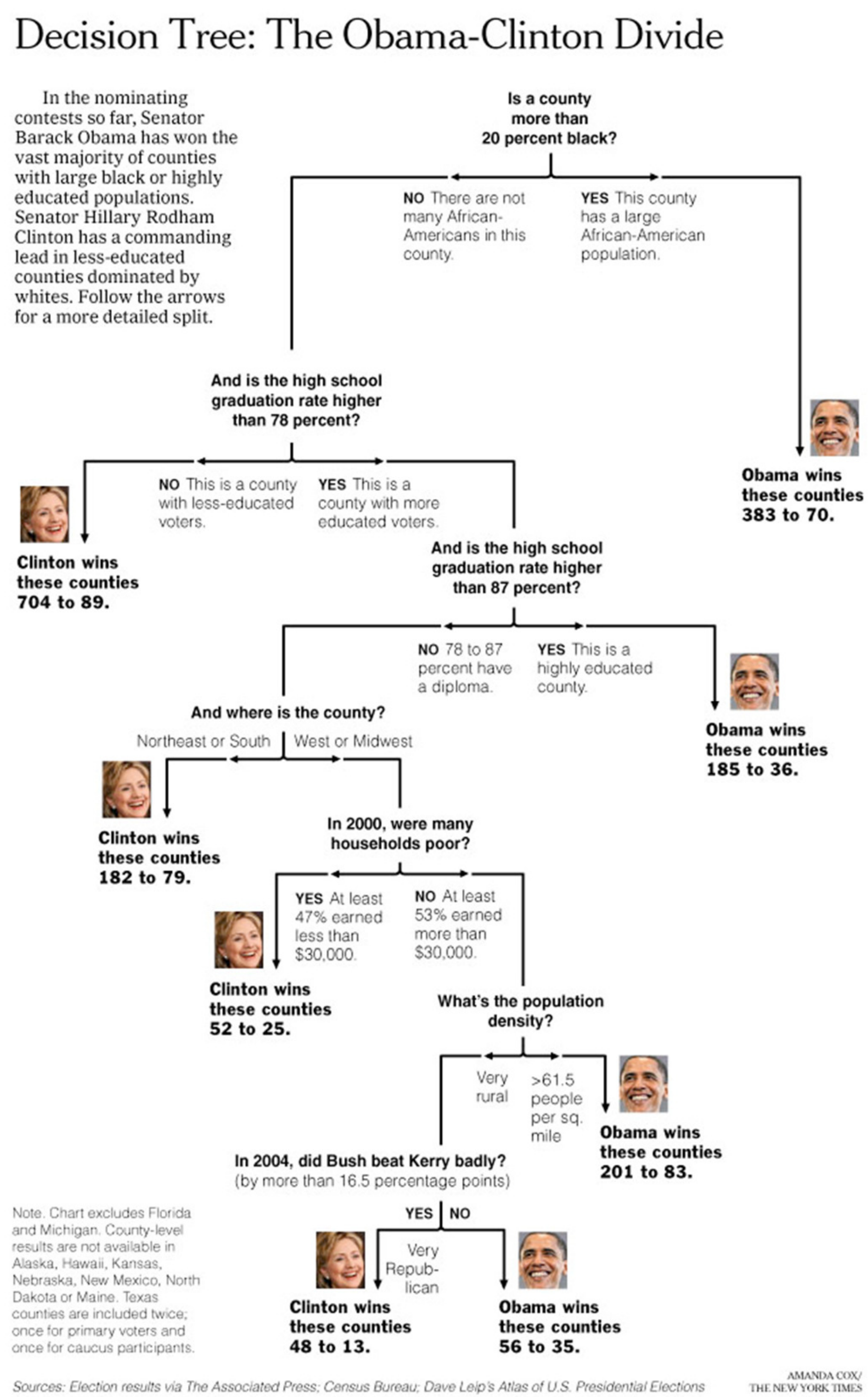
$$X_1 \times X_2 \times \ldots X_D \,, \qquad (D \gg 1) \tag{1}$$

   X_i can be discrete, continuous, categorical, ordinal, . . .

3. interpretability
4. to classify, not all X_i may be actually needed
5. sequential decisions
6. high degree of robustness to outliers (unlike e.g. linear classifiers)
7. different scales between features are not a problem
8. deals with additional information
9. for any decision tree point, gives class probability
10. can be used in regression problems

Cons

1. not well justified learning algorithm
2. overfitting is a problem for standard methods, but this can be fixed (randomized d. trees, decision *forests*)

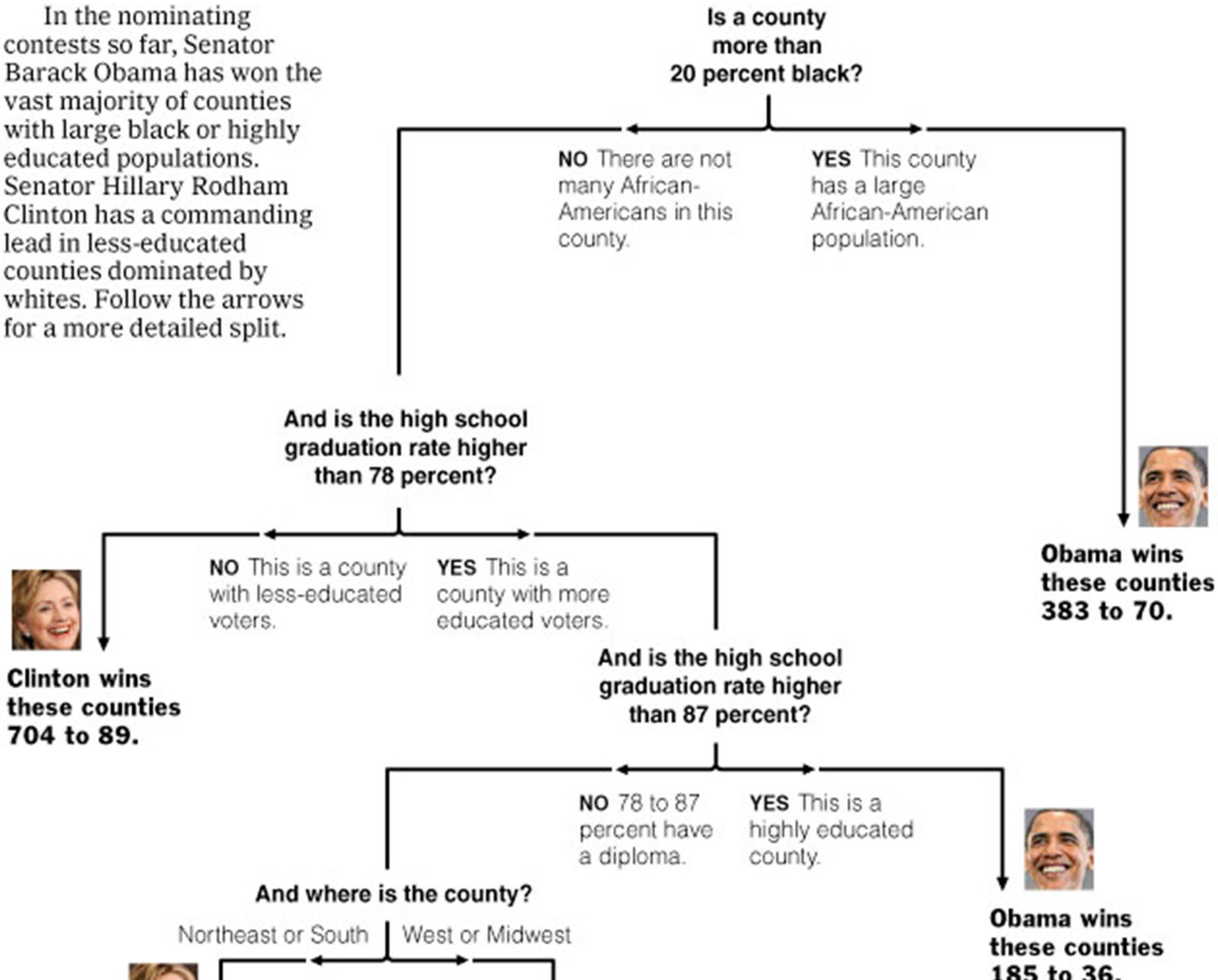


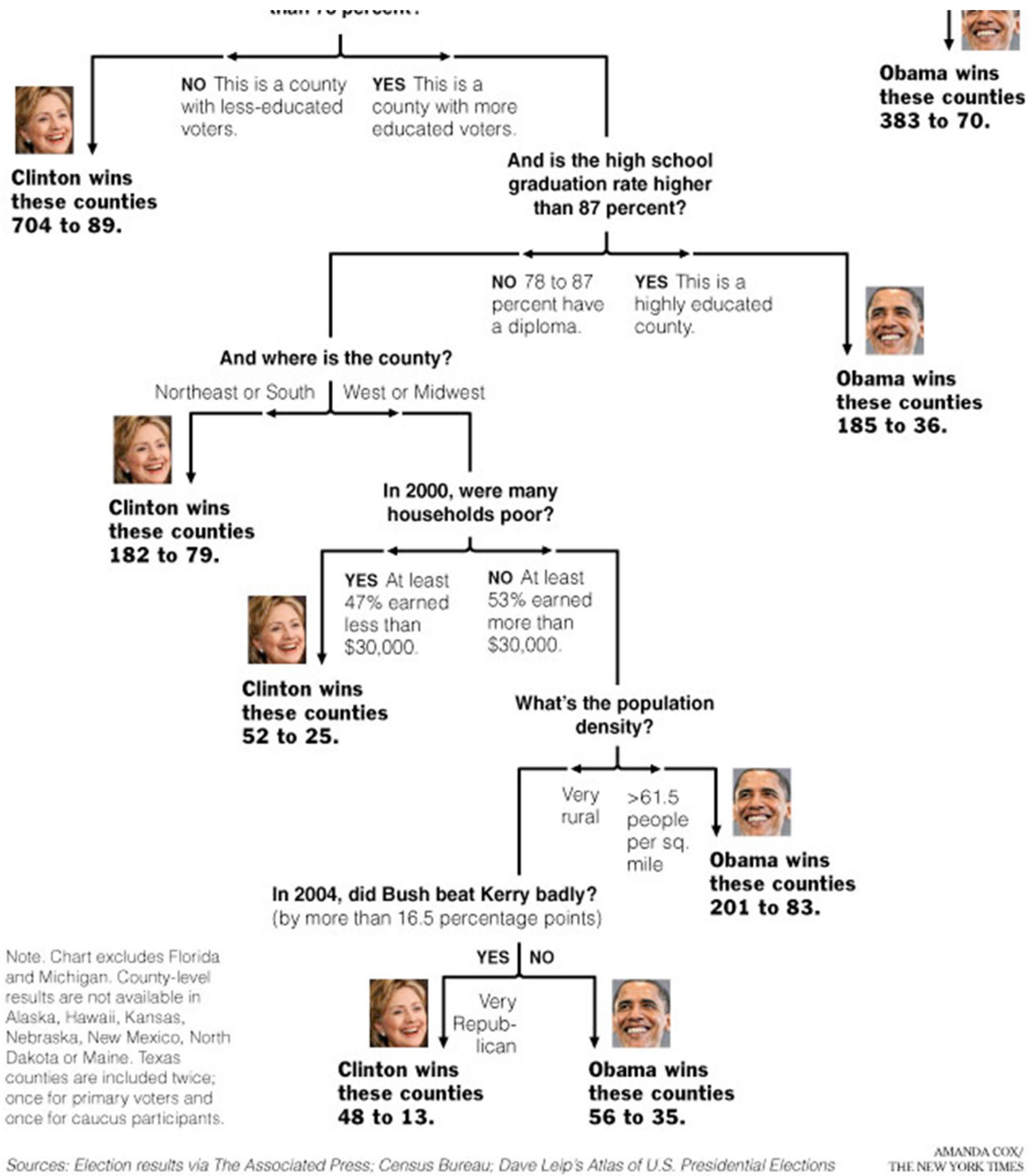
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.




NO This is a county with less-educated voters.
YES This is a county with more educated voters.
Clinton wins these counties 704 to 89.
Obama wins these counties 383 to 70.
And is the high school graduation rate higher than 87 percent?
NO 78 to 87 percent have a diploma.
YES This is a highly educated county.
Obama wins these counties 185 to 36.
And where is the county?
Northeast or South
West or Midwest
Clinton wins these counties 182 to 79.
In 2000, were many households poor?
YES At least 47% earned less than $30,000.
NO At least 53% earned more than $30,000.
Clinton wins these counties 52 to 25.
What's the population density?
Very rural
>61.5 people per sq. mile
Obama wins these counties 201 to 83.
In 2004, did Bush beat Kerry badly?
(by more than 16.5 percentage points)
YES
NO
Very Repub-lican
Clinton wins these counties 48 to 13.
Obama wins these counties 56 to 35.
Note. Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.
Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections
AMANDA COX/
THE NEW YORK TIMES

Decision Tree

What is a decision tree?

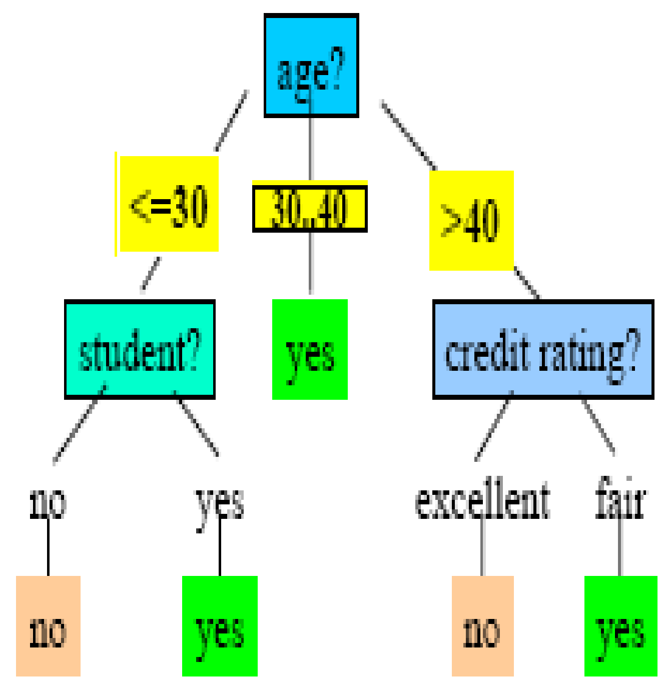
A tree with two types of nodes:

Decision nodes
Leaf nodes

Decision node: **Specifies a choice or test of some attribute with 2 or more alternatives;**
→ every decision node is part of a path to a leaf node

Leaf node: **Indicates classification of an example**

- Decision Tree example (is a customer going to buy a computer or not):



Learning decision trees: An example

Problem: decide whether to wait for a table at a restaurant.

Attributes used

1. Alternate: is there an alternative restaurant nearby?
2. Bar: is there a comfortable bar area to wait in?
3. Fri/Sat: is today Friday or Saturday?
4. Hungry: are we hungry?
5. Patrons: number of people in the restaurant (None, Some, Full)
6. Price: price range ($, $$, $$$)
7. Raining: is it raining outside?
8. Reservation: have we made a reservation?
9. Type: kind of restaurant (French, Italian, Thai, Burger)
10. WaitEstimate: estimated waiting time (0-10, 10-30, 30-60, >60)

Goal predicate: WillWait?

Attribute-based representations

Examples described by attribute values (Boolean, discrete, continuous)

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

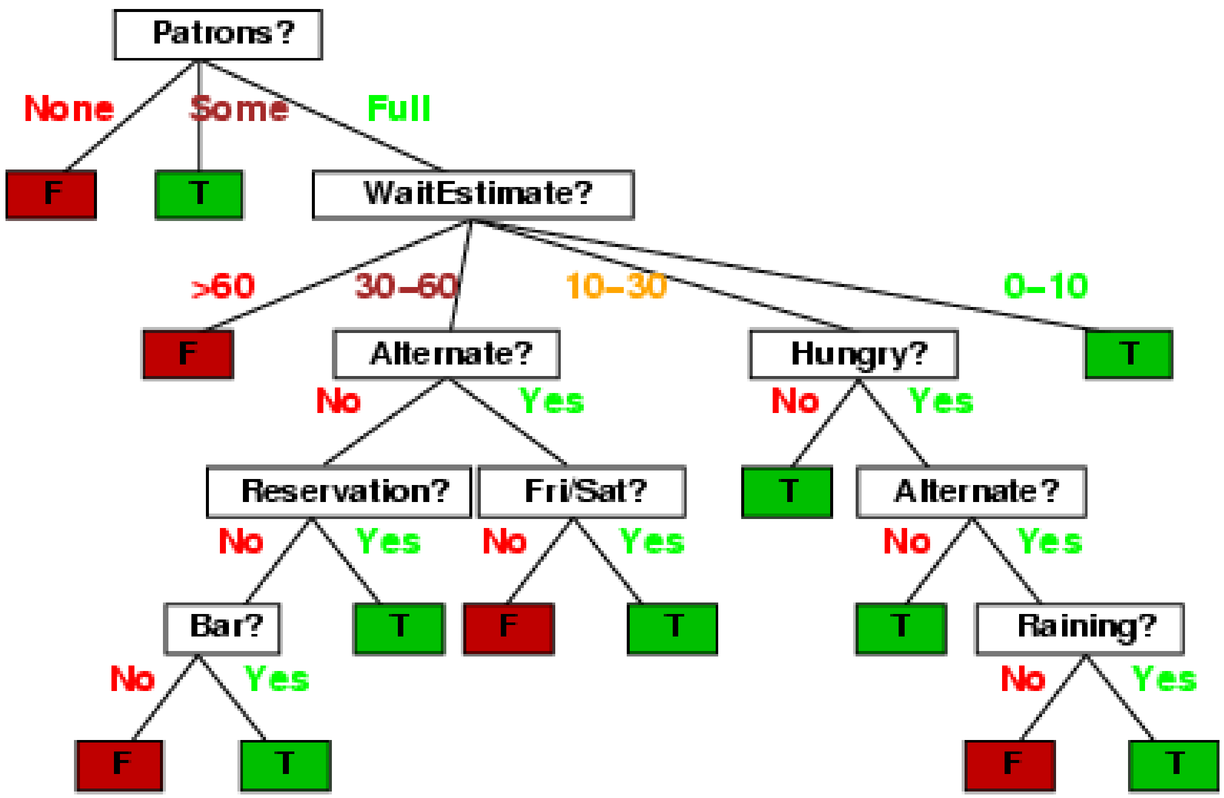
12 examples
(6 T, 6 F)

Classification of examples is positive (T) or negative (F)

Decision trees

One possible representation for hypotheses

E.g., here is a tree for deciding whether to wait:



Top-Down Induction of DT (TDIDT)

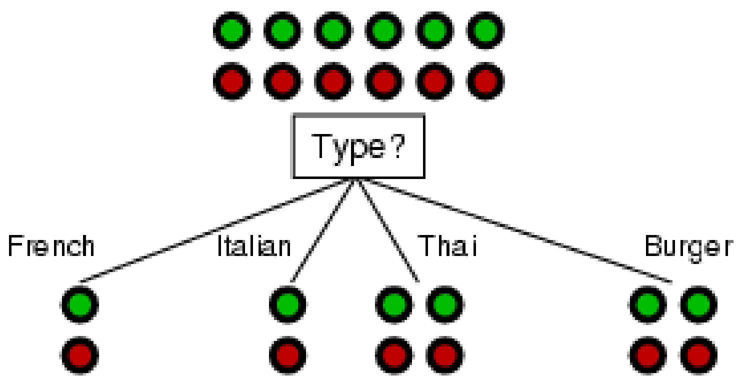
Growth Phase: The tree is constructed top-down.

- Find the “best” attribute.
- Partition examples based on the attribute’s values.
- Apply recursively to each partition.

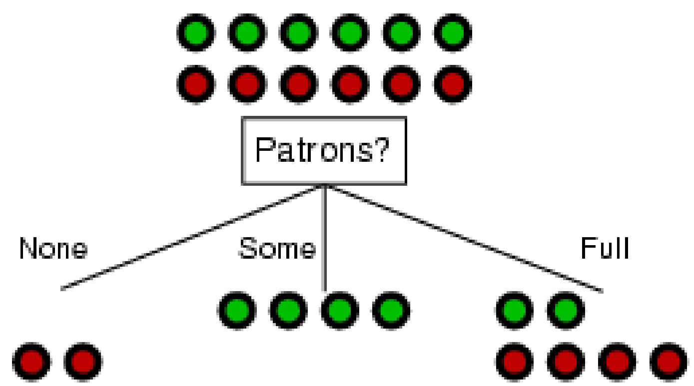
Pruning Phase: The tree is pruned bottom-up

- For each node, keep subtree or change to leaf.
- Choose by comparing estimated error.

Choosing an attribute: Information Gain



Is this a good attribute to split on?



Which one should we pick?

A perfect attribute would ideally divide the examples into sub-sets that are all positive or all negative… i.e. maximum information gain.

Choosing an attribute: Information Gain

Information Gain (IG) measures the expected reduction in entropy. The higher the Information Gain with respect to an attribute A, the more is the expected reduction in entropy.

Weight of each subcla

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

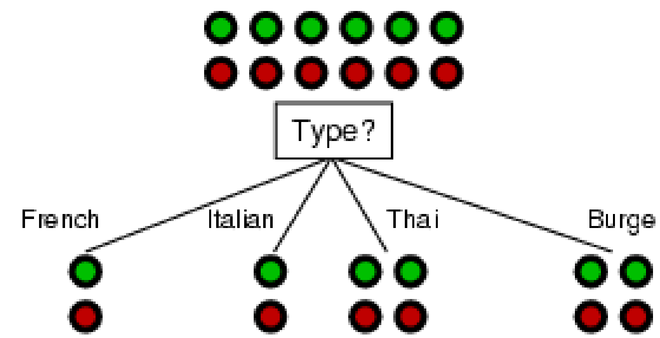
where Values(A) is the set of all possible values for attribute A,

S_v is the subset of S for which attribute A has value v.

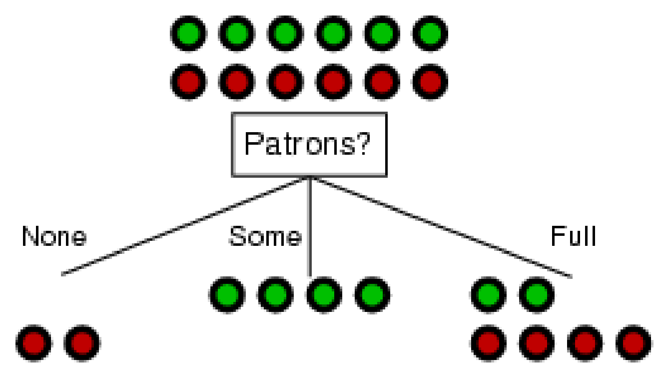
Information gain

For the training set, $p = n = 6$, *entropy* $I(6/12, 6/12) = 1$ bit

Consider the attributes *Type* and *Patrons*:



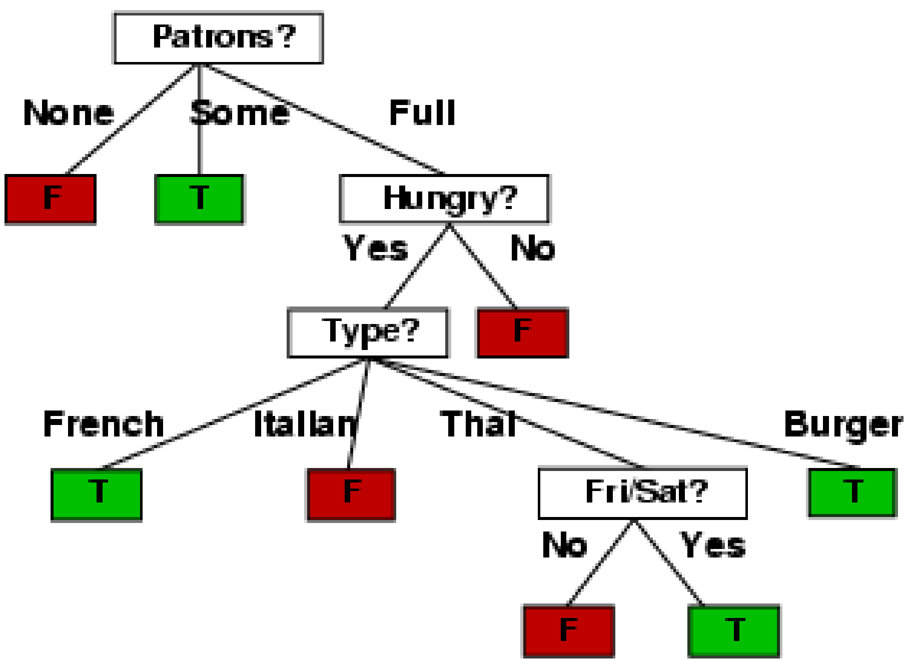
$$IG(Type) = 1 - [\frac{2}{12} I(\frac{1}{2},\frac{1}{2}) + \frac{2}{12} I(\frac{1}{2},\frac{1}{2}) + \frac{4}{12} I(\frac{2}{4},\frac{2}{4}) + \frac{4}{12} I(\frac{2}{4},\frac{2}{4})] = 0 \text{ bits}$$



$$IG(Patrons) = 1 - [\frac{2}{12} I(0,1) + \frac{4}{12} I(1,0) + \frac{6}{12} I(\frac{2}{6},\frac{4}{6})] = .0541 \text{ bits}$$

Example contd.

Decision tree learned from the 12 examples:



Building a Decision Tree: Splitting a Node to Branches

$$P$$

$$P_1 \; P_2 \cdots P_k$$

- Define a splitting function $S(P, \boldsymbol{\theta})$ as a mapping which takes a set P (data points at a node) and splitting parameters $\boldsymbol{\theta}$ and produces the partition of P to $\{P_1, P_2, \ldots, P_k\}$. That is,

$$S : P, \boldsymbol{\theta} \rightarrow \{P_1, P_2, \ldots, P_k\}\,, \tag{2}$$

$$P_i \cap P_j = \emptyset \Leftrightarrow i \neq j \tag{3}$$

$$\cup_{i=1}^{k} P_i = P \tag{4}$$

- $\boldsymbol{\theta}$: contains the index i of the splitting dimension X_i and the branching factor k
- $k = 2$: binary split
- $k > 2$: multiway split
- When X_i is continuous, $\boldsymbol{\theta}$ contains the range(s) of values for individual partitions (when $k = 2$, typically a threshold. Example: [height $\leqslant 1.75m$?])

Input:

- training set $\mathcal{T} = \{(\mathbf{x}_1, k_1), (\mathbf{x}_2, k_2), ..., (\mathbf{x}_N, k_N)\}$.
- class $\mathcal{S}$ of allowed splitting parameters

Do:

1. $P = \mathcal{T}$
2. Find the best splitting function (best measured by training error ϵ):

$$S = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathcal{S}} \frac{|P_1|}{|P|}\epsilon(P_1) + \frac{|P_2|}{|P|}\epsilon(P_2) + \ldots \frac{|P_k|}{|P|}\epsilon(P_k) \tag{5}$$

$$\{P_1, P_2, \ldots, P_k\} = S(P, \boldsymbol{\theta}) \tag{6}$$

$$\epsilon(P_i) = 1 - \max_j \frac{|P_i^{(j)}|}{|P_i|} \tag{7}$$

$(P_i^{(j)}$: number of points of class j in $P_i)$

3. Go recursively to branches

Input:

- training set $\mathcal{T} = \{(\mathbf{x}_1, k_1), (\mathbf{x}_2, k_2), ..., (\mathbf{x}_N, k_N)\}$.
- class $\mathcal{S}$ of allowed splitting parameters

Do:

1. $P = \mathcal{T}$
2. Find the best splitting function (best is measured by **Information Gain** which is the difference of entropies H of data before and after the split):

$$S = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathcal{S}} \frac{|P_1|}{|P|} H(P_1) + \frac{|P_2|}{|P|} H(P_2) + \ldots \frac{|P_k|}{|P|} H(P_k) \tag{8}$$

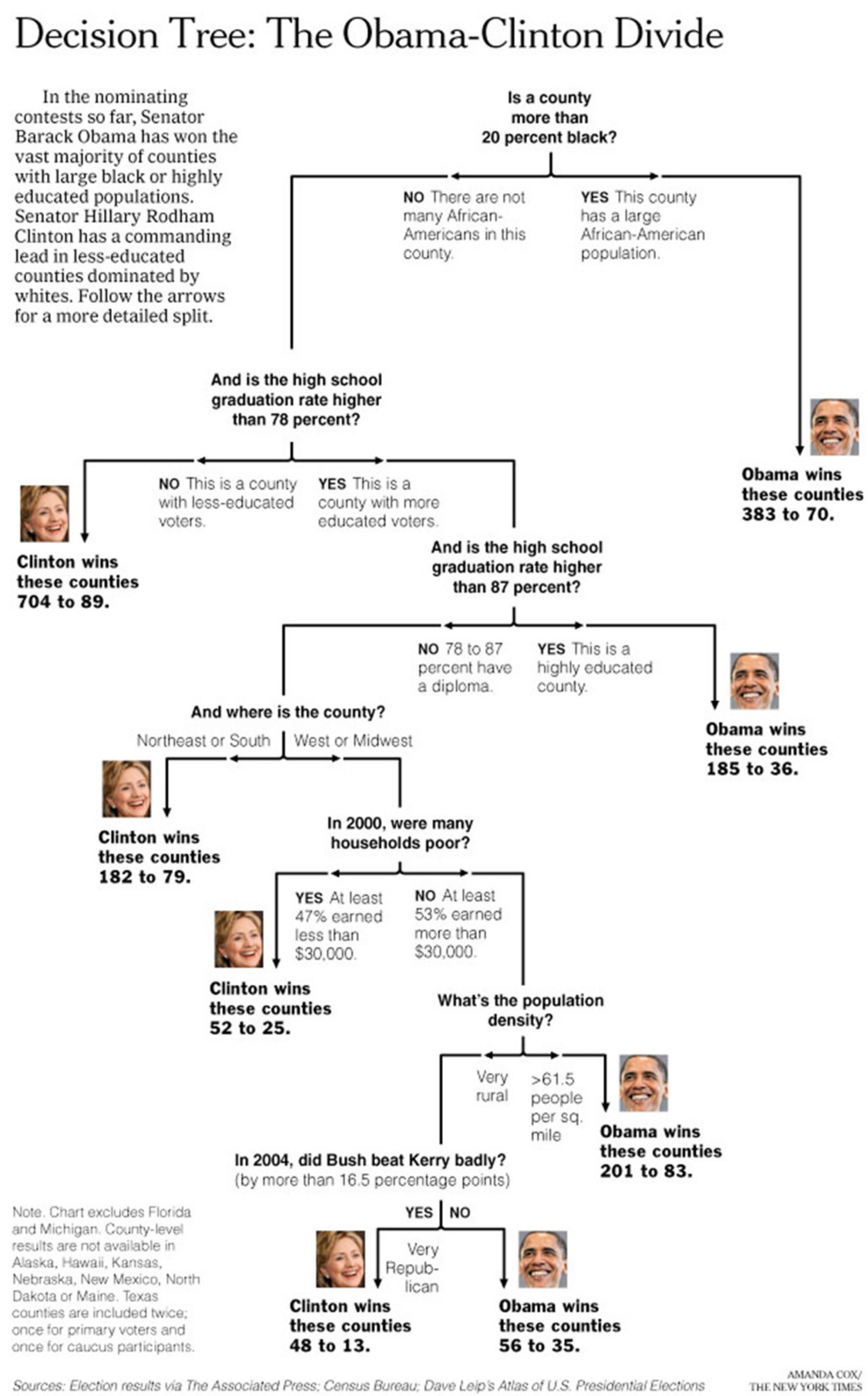
$$\{P_1, P_2, \ldots, P_k\} = S(P, \boldsymbol{\theta}) \tag{9}$$

$$H(P_i) = -\sum_j \frac{|P_i^{(j)}|}{|P_i|} \log \frac{|P_i^{(j)}|}{|P_i|} \quad \text{(entropy of } P_i\text{)} \tag{10}$$

($P_i^{(j)}$: number of points of class j in P_i)

3. Go recursively to branches

m p

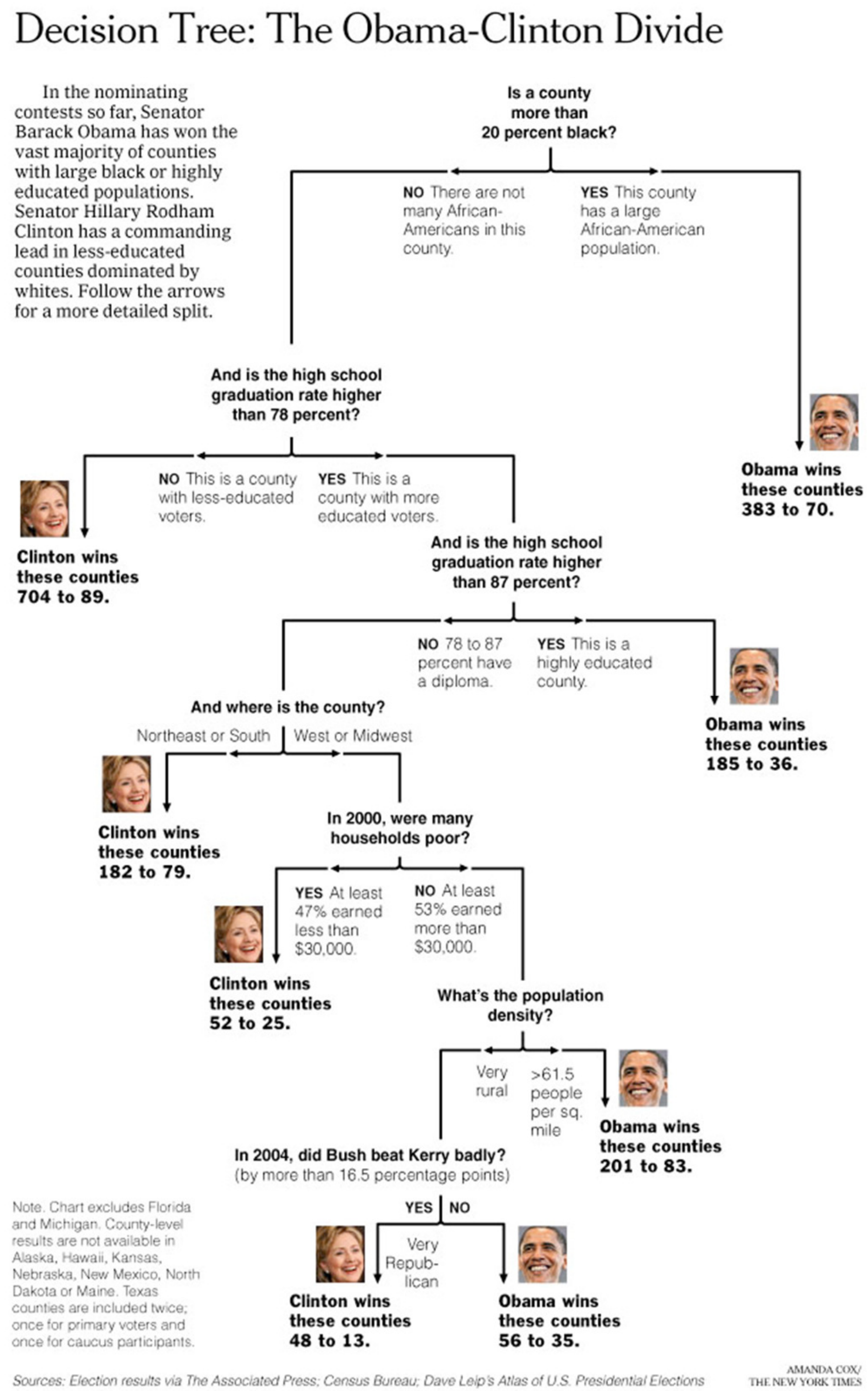


New York Times
April 16, 2008

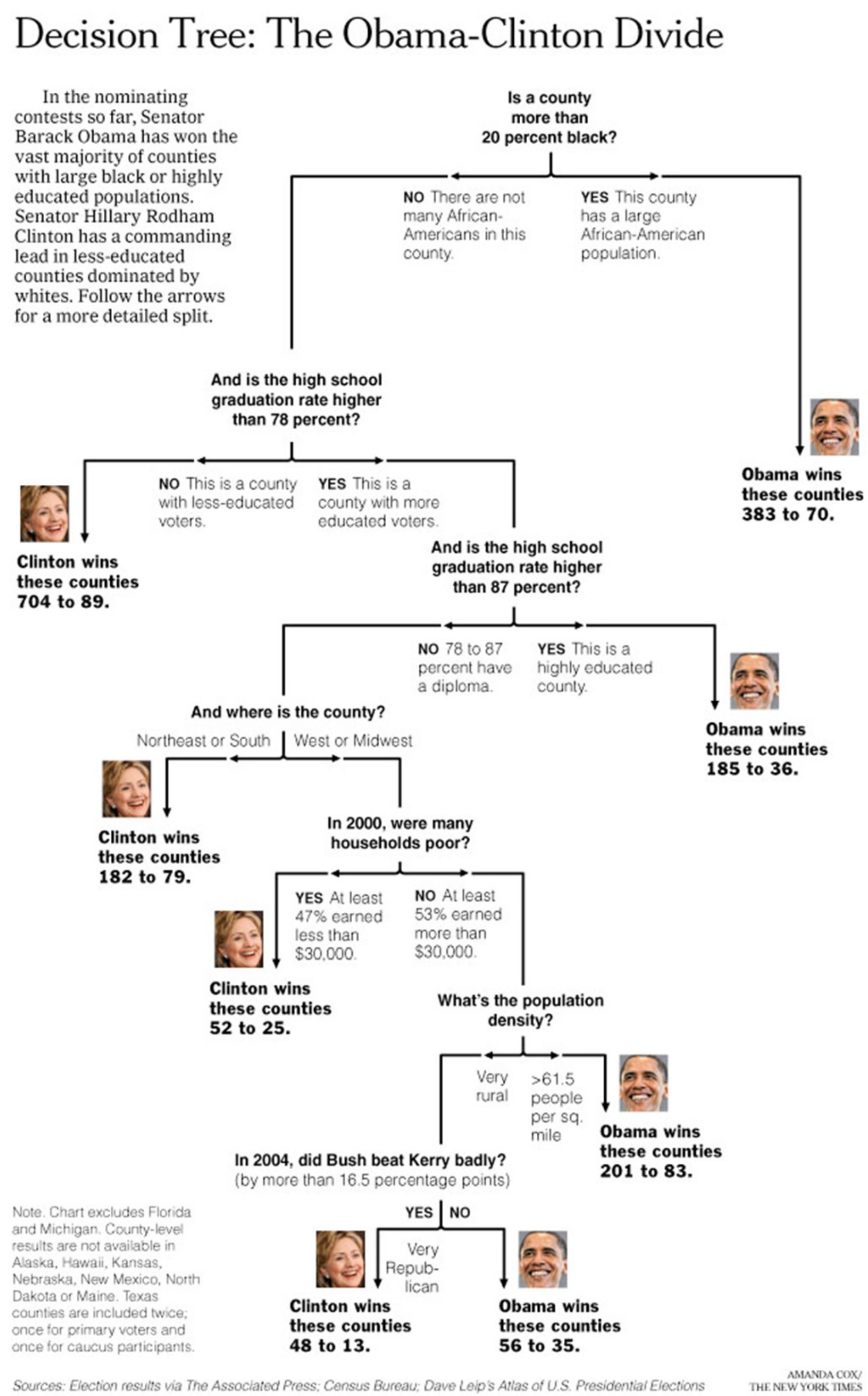
Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



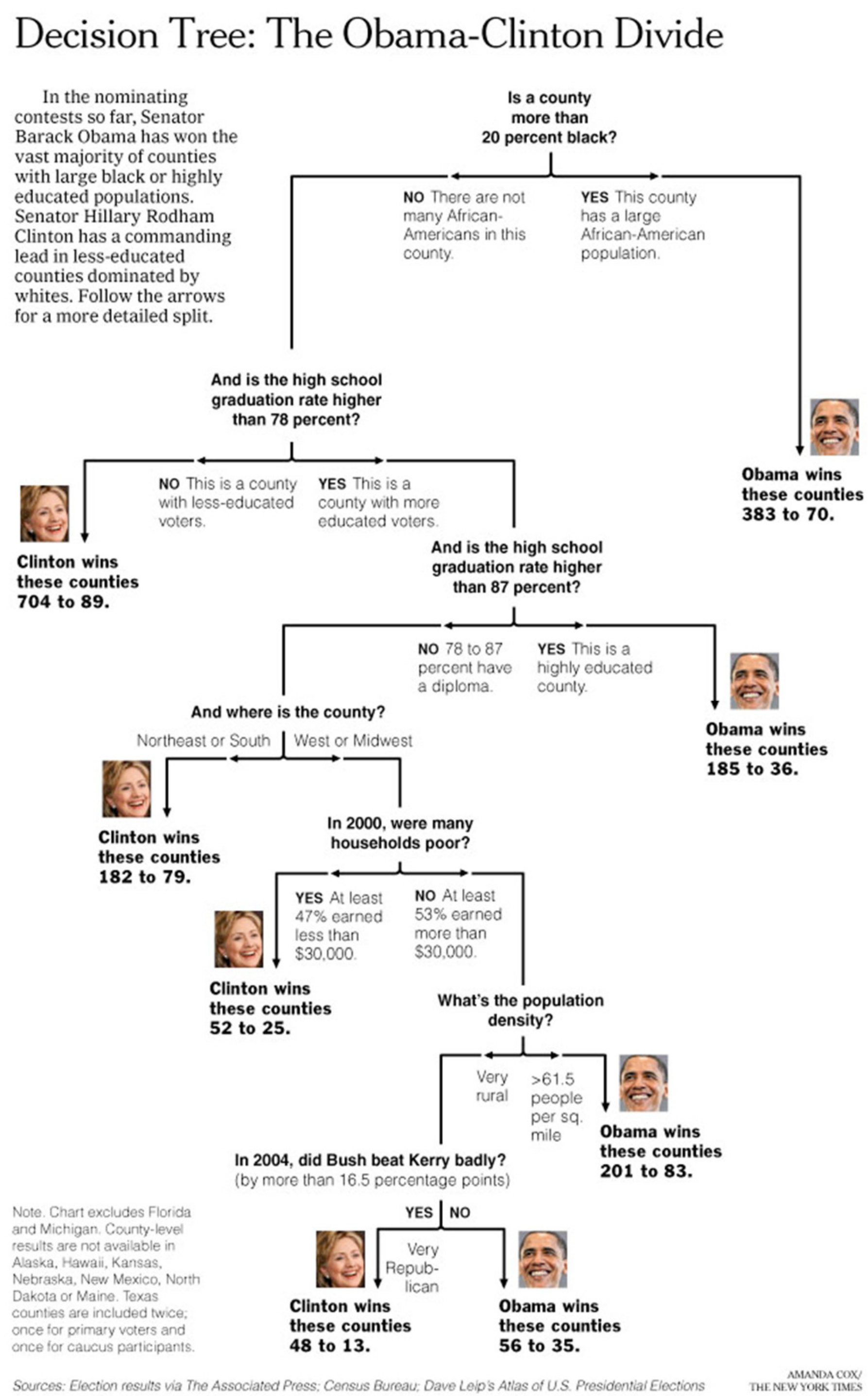
New York Times
April 16, 2008



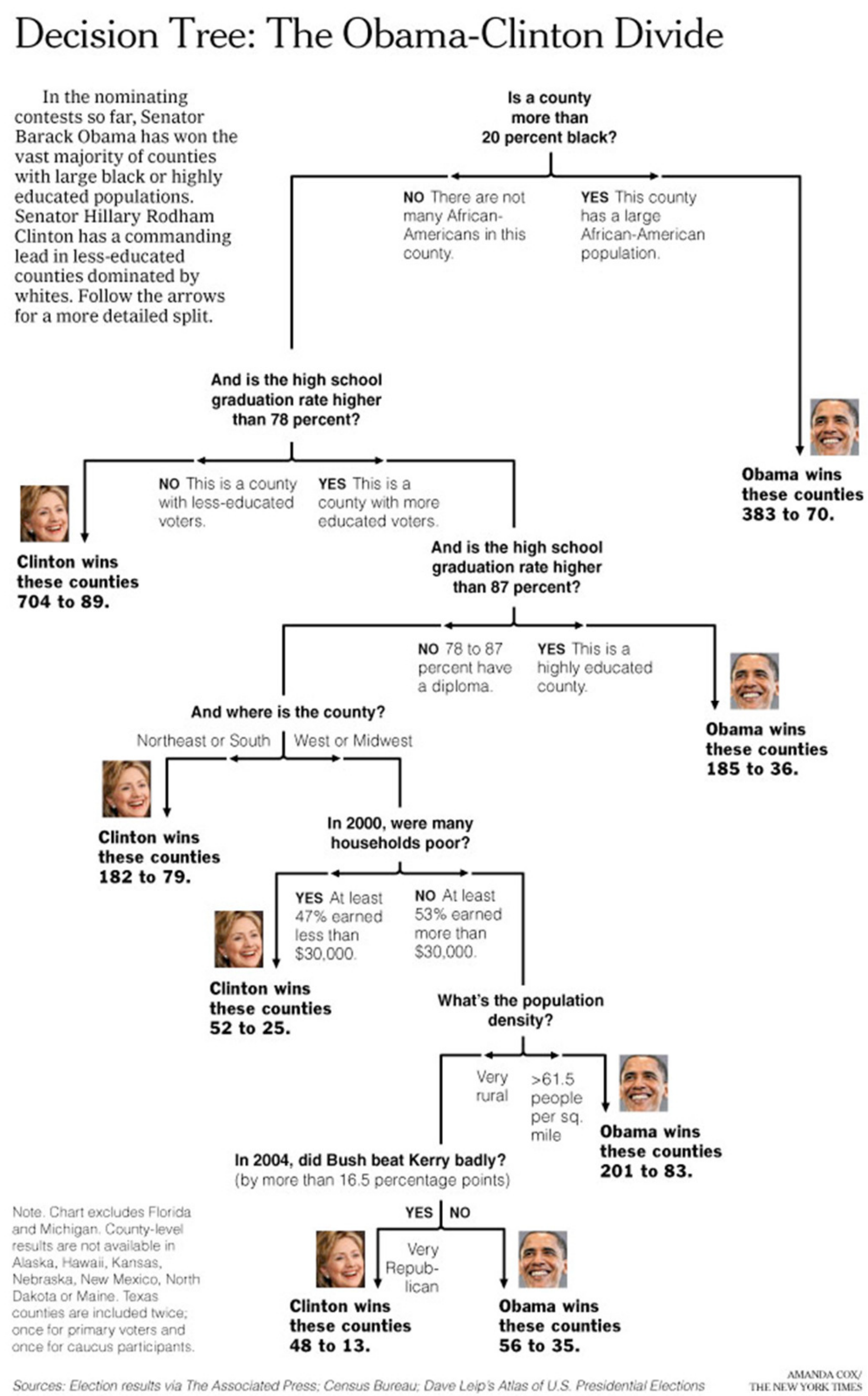
Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many “How to”
manuals and trouble-shooting
procedures.

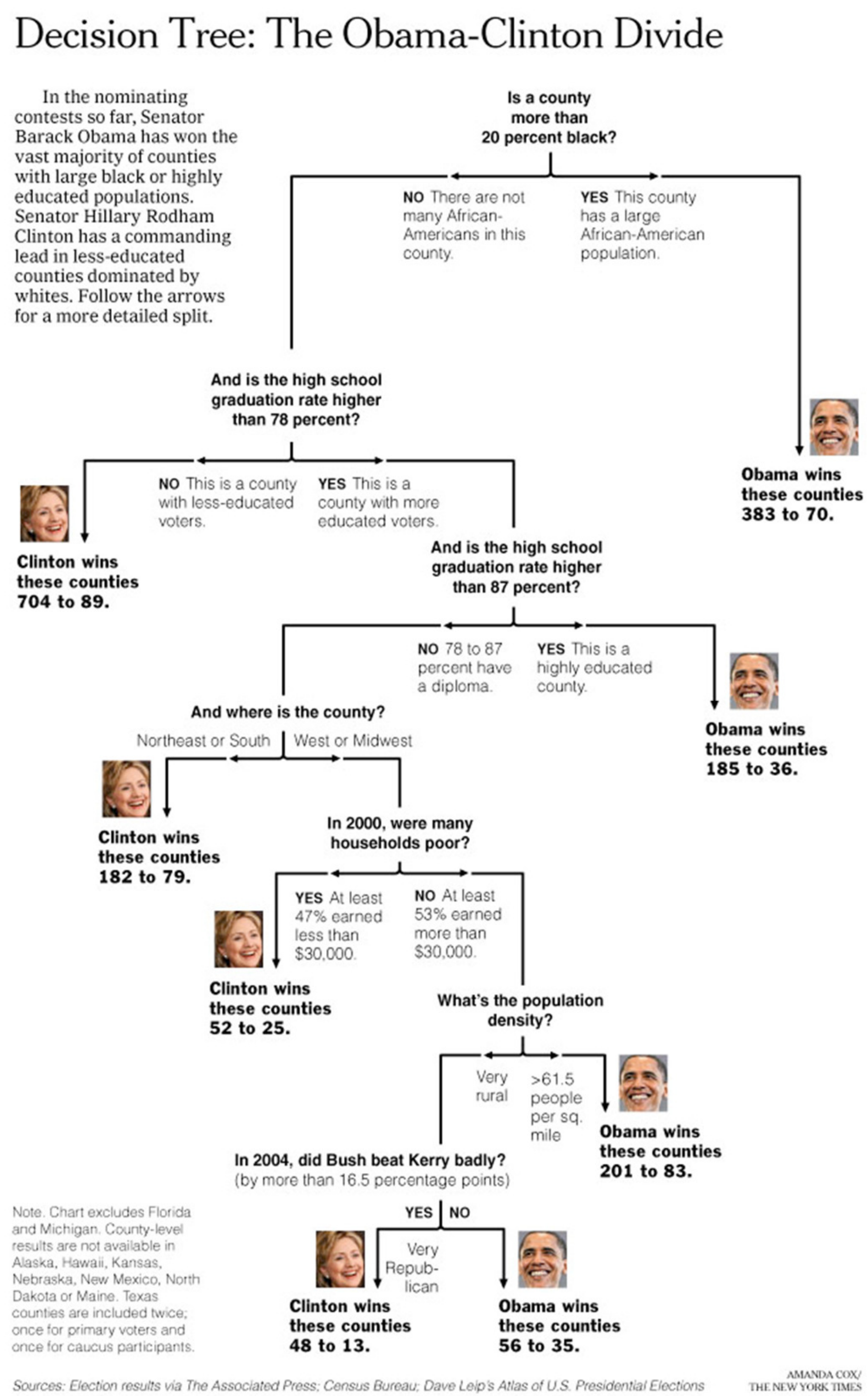


New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

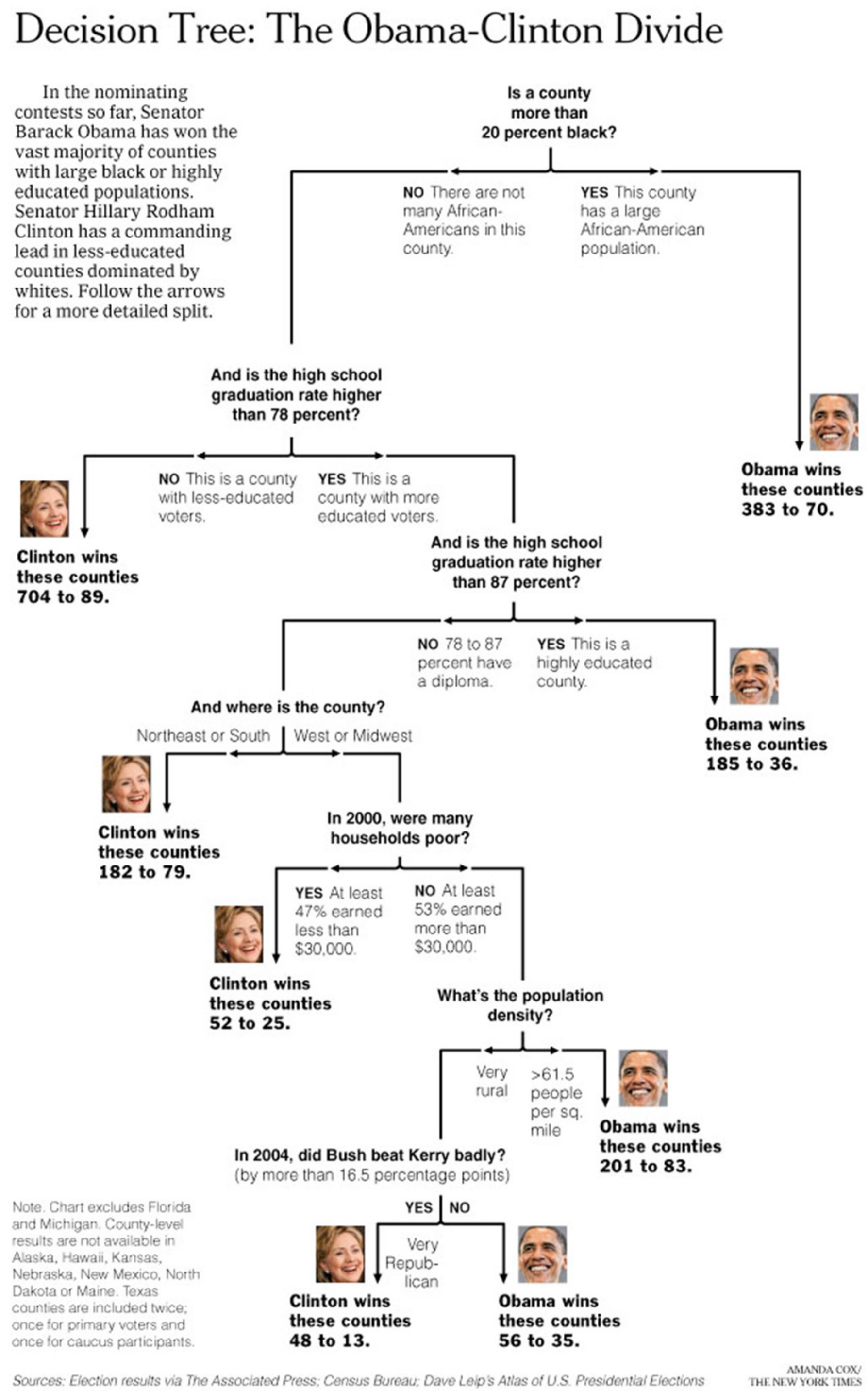
New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

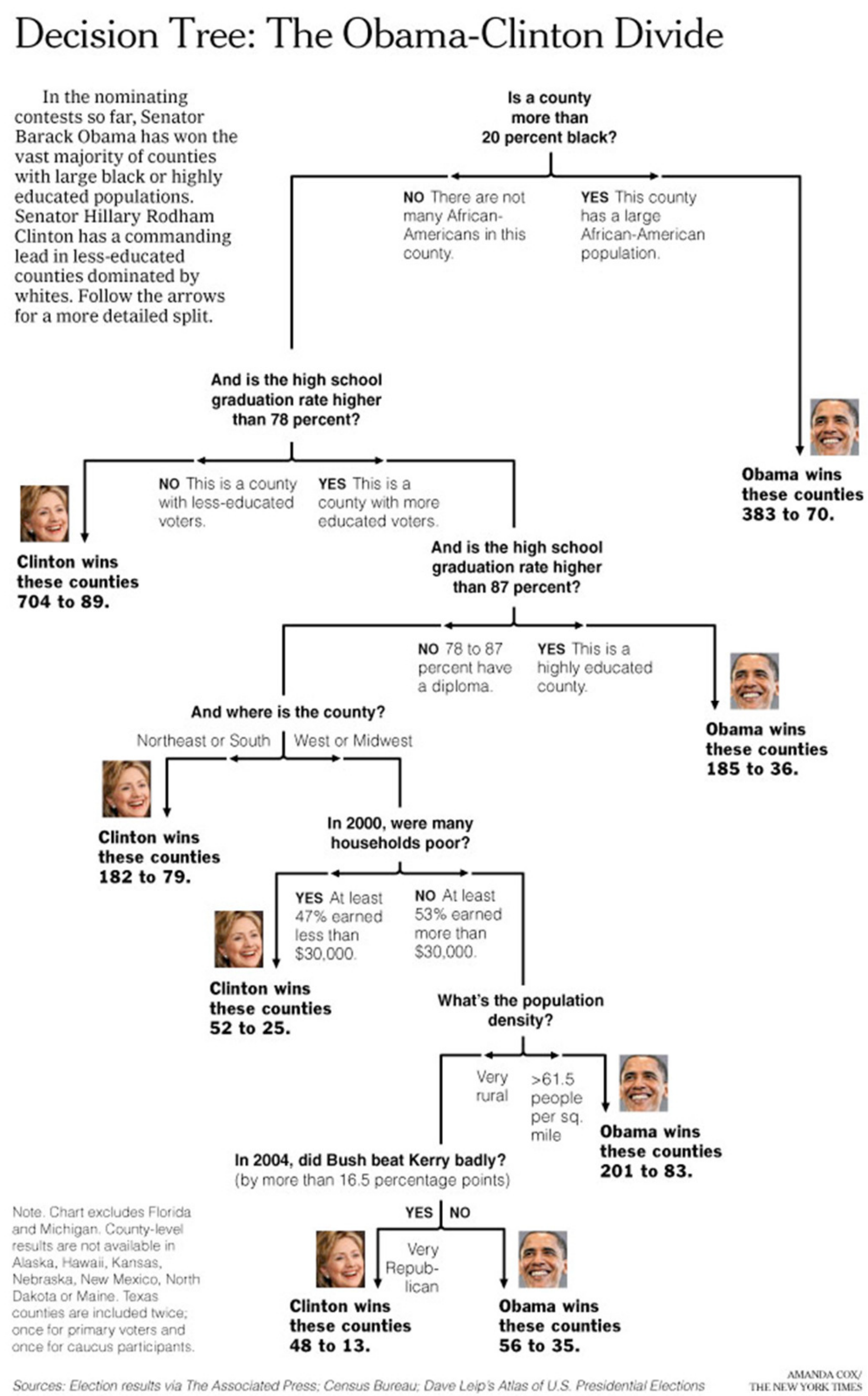
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



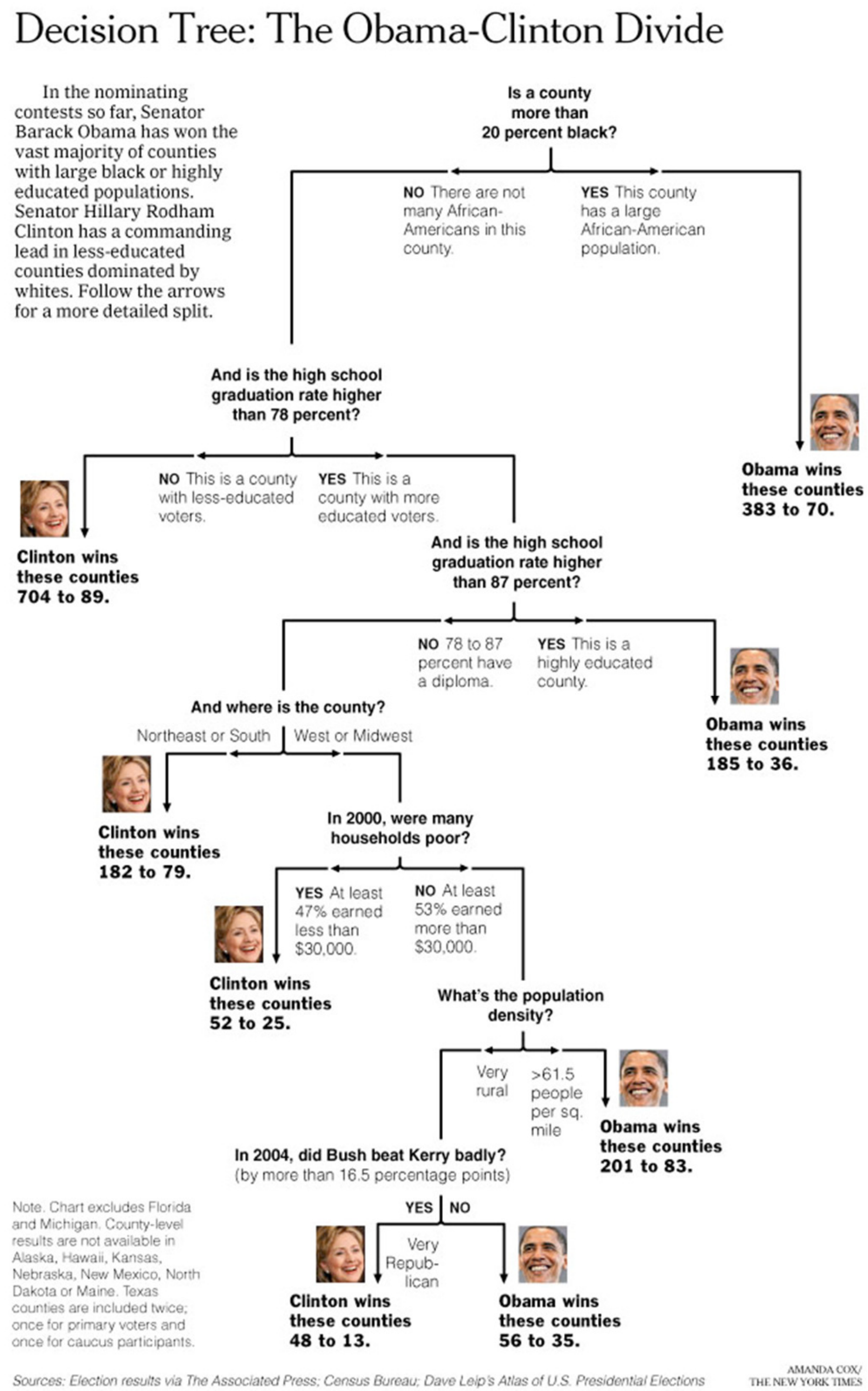
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



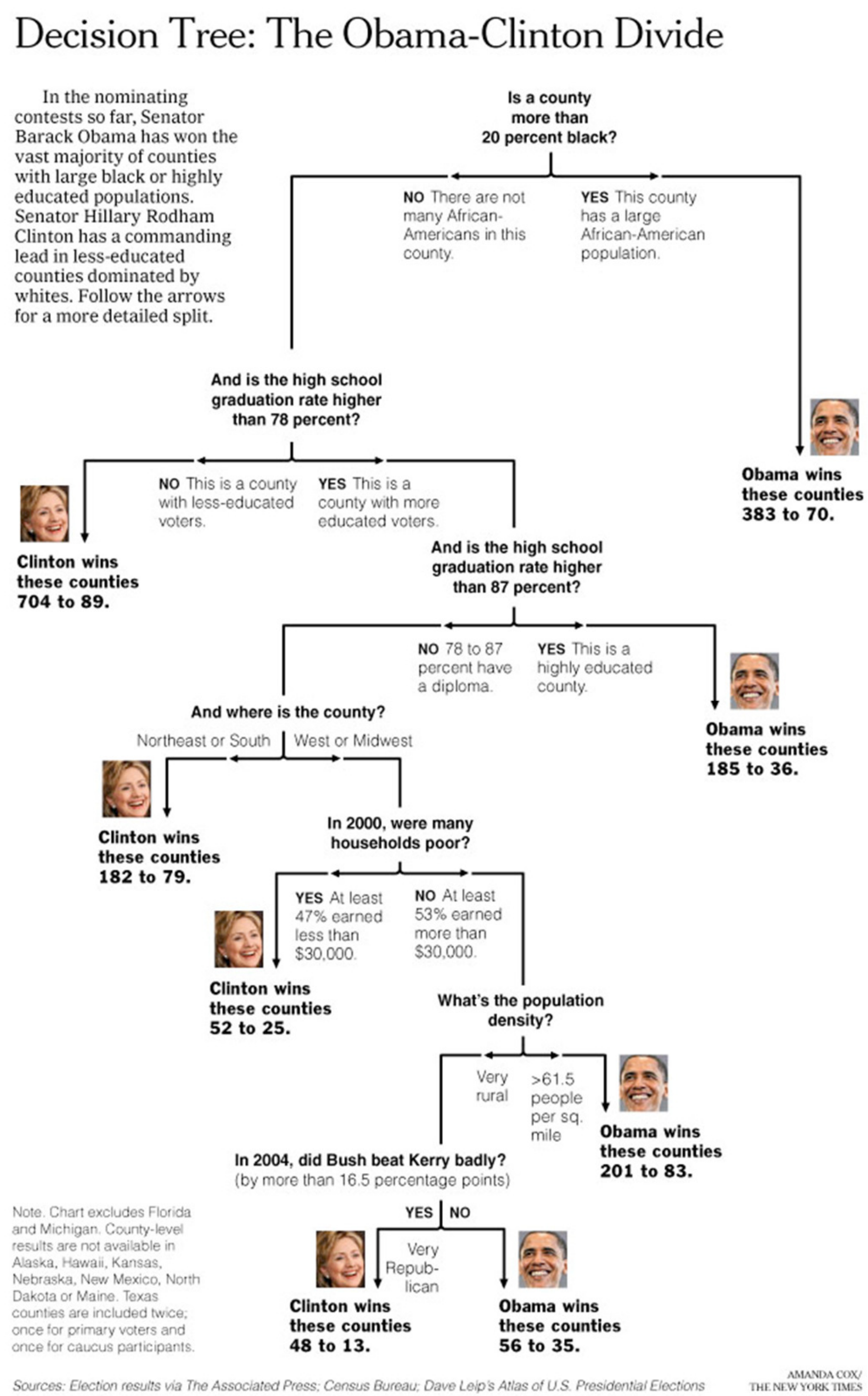
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



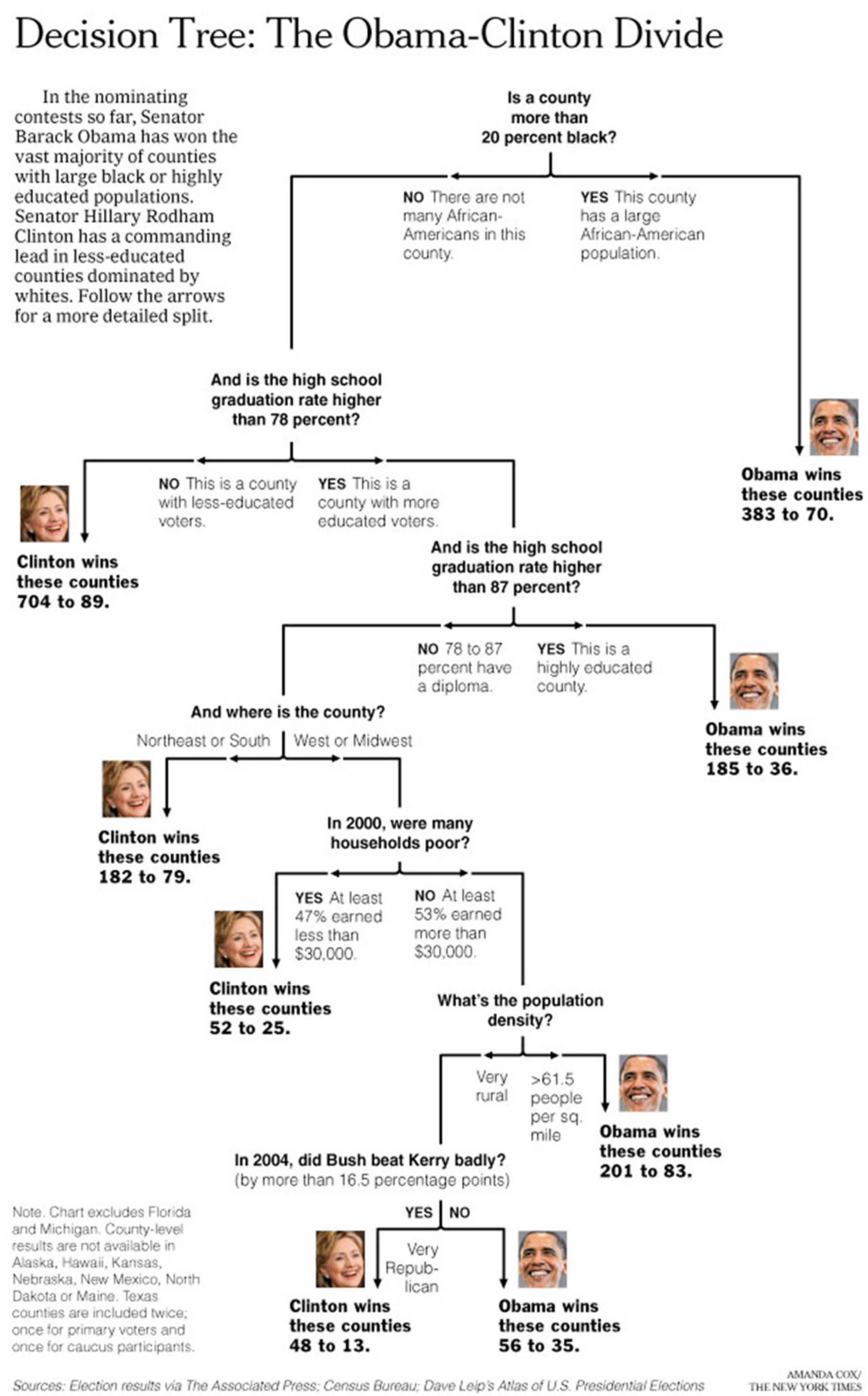
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



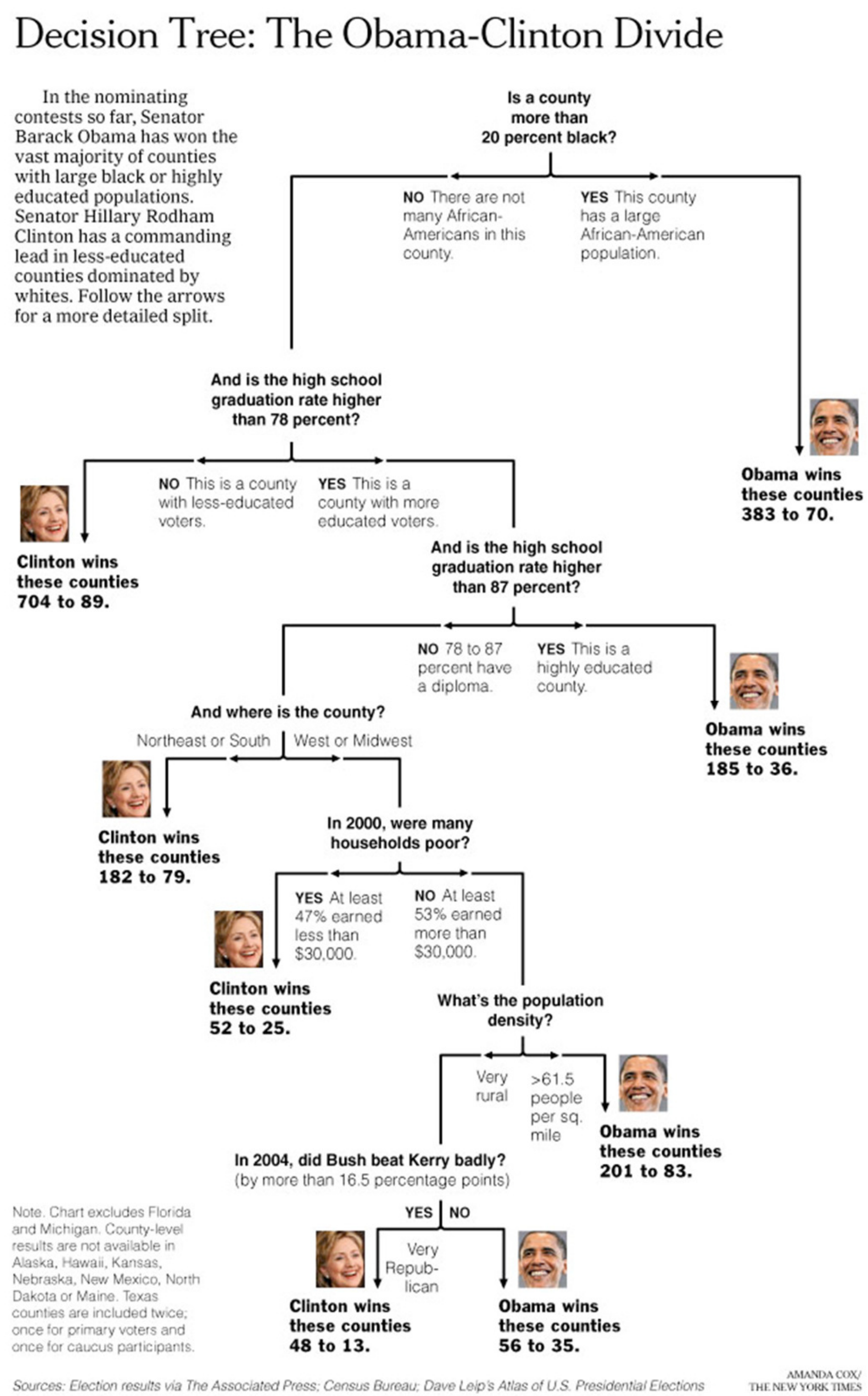
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



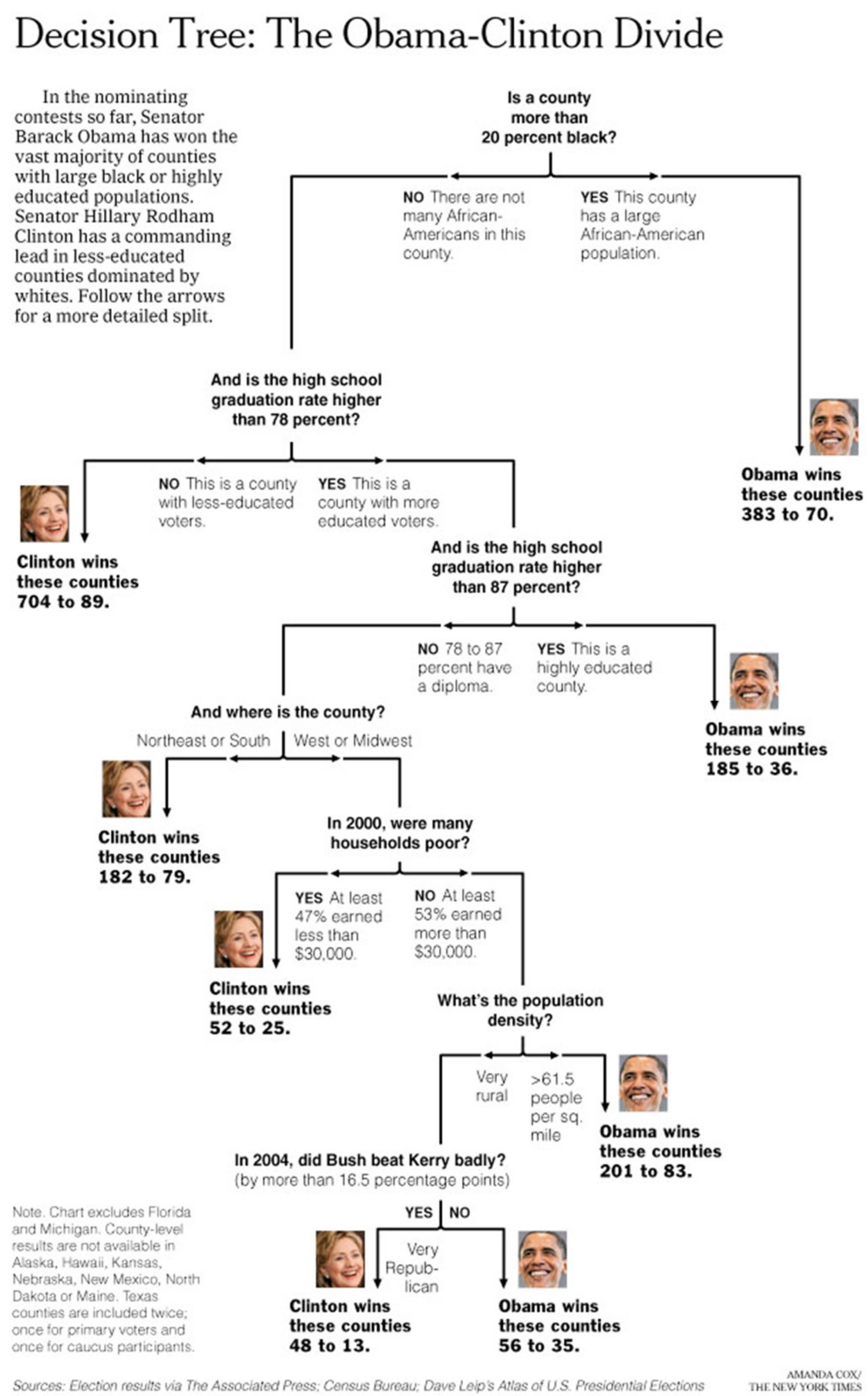
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



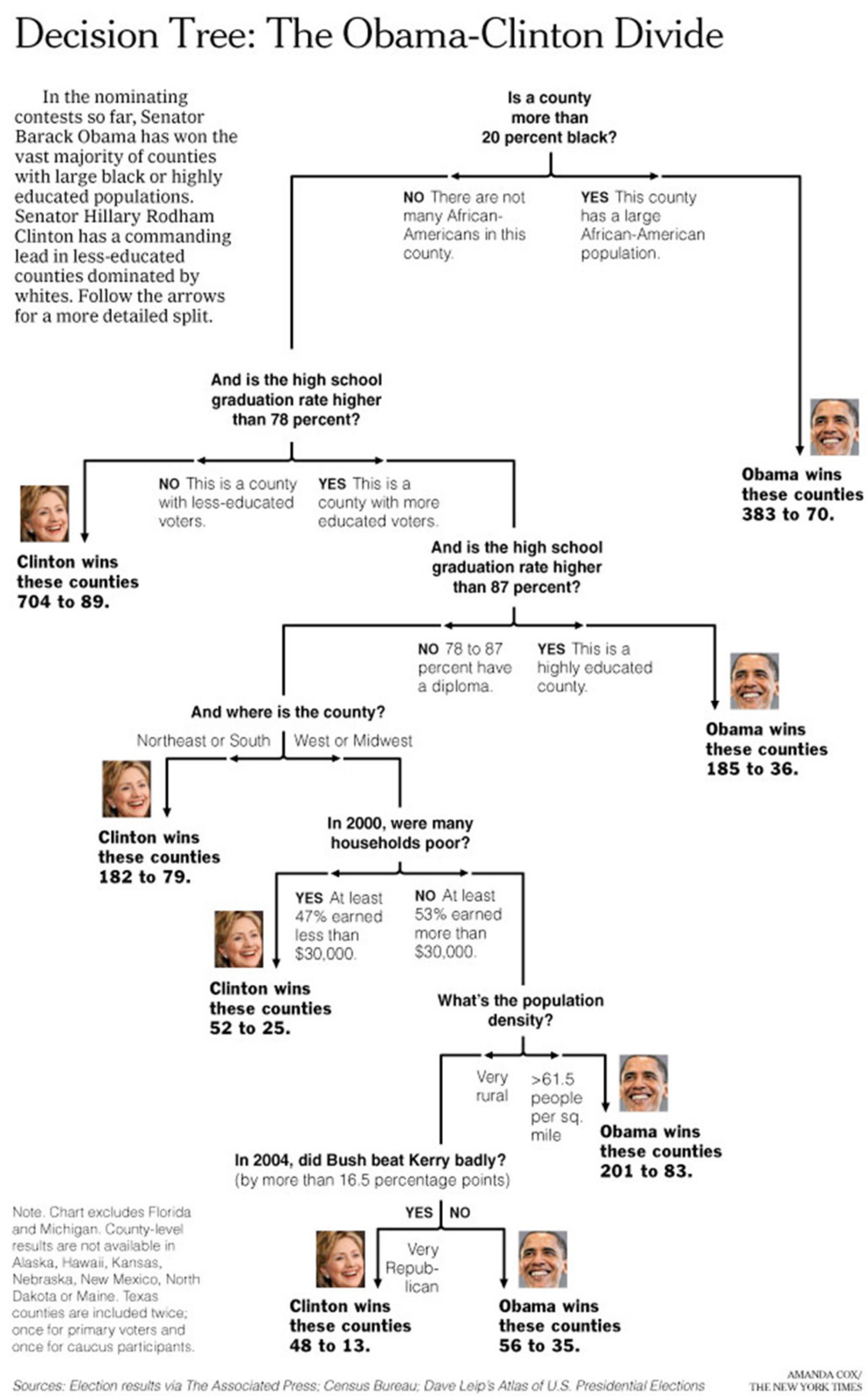
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



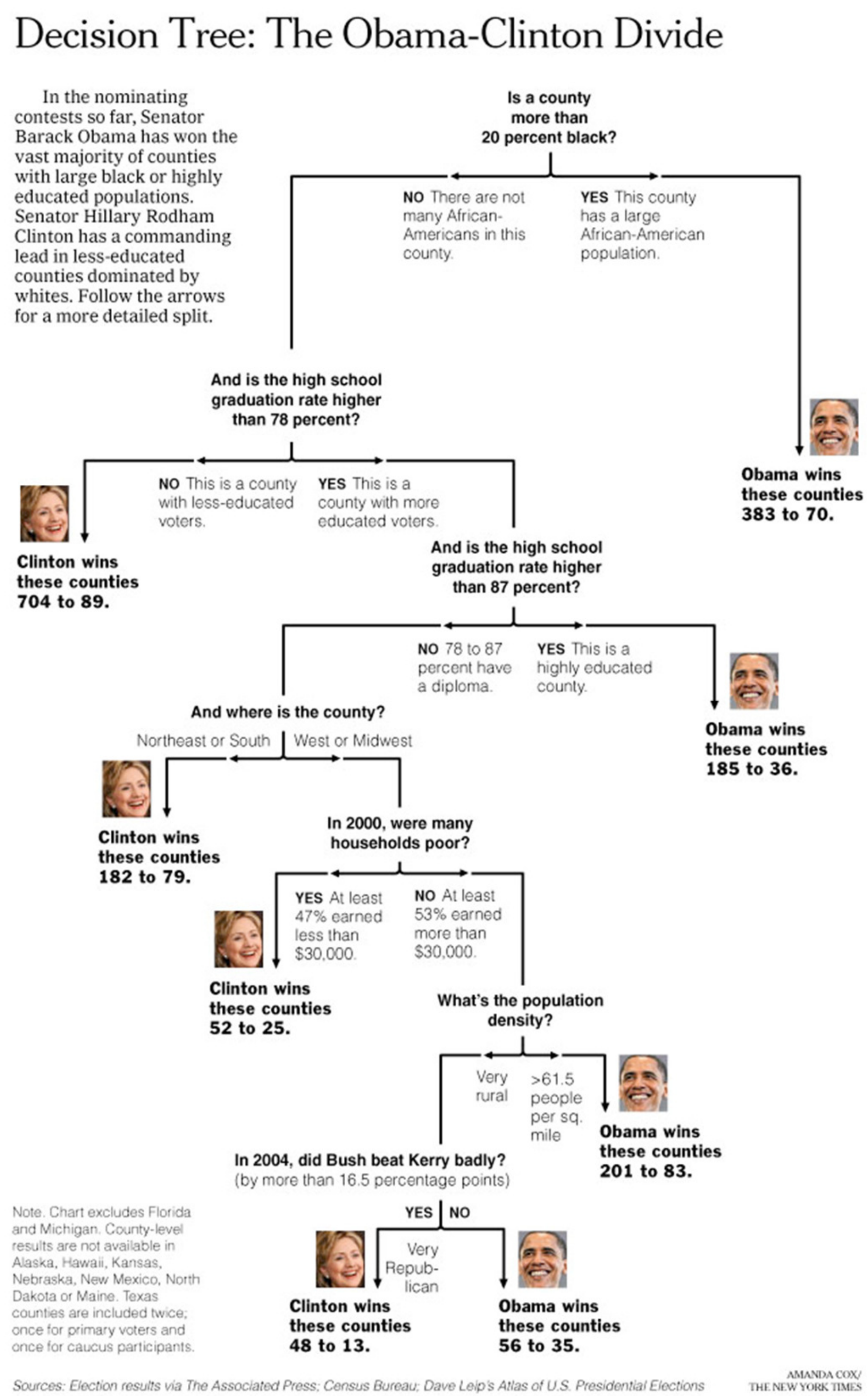
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



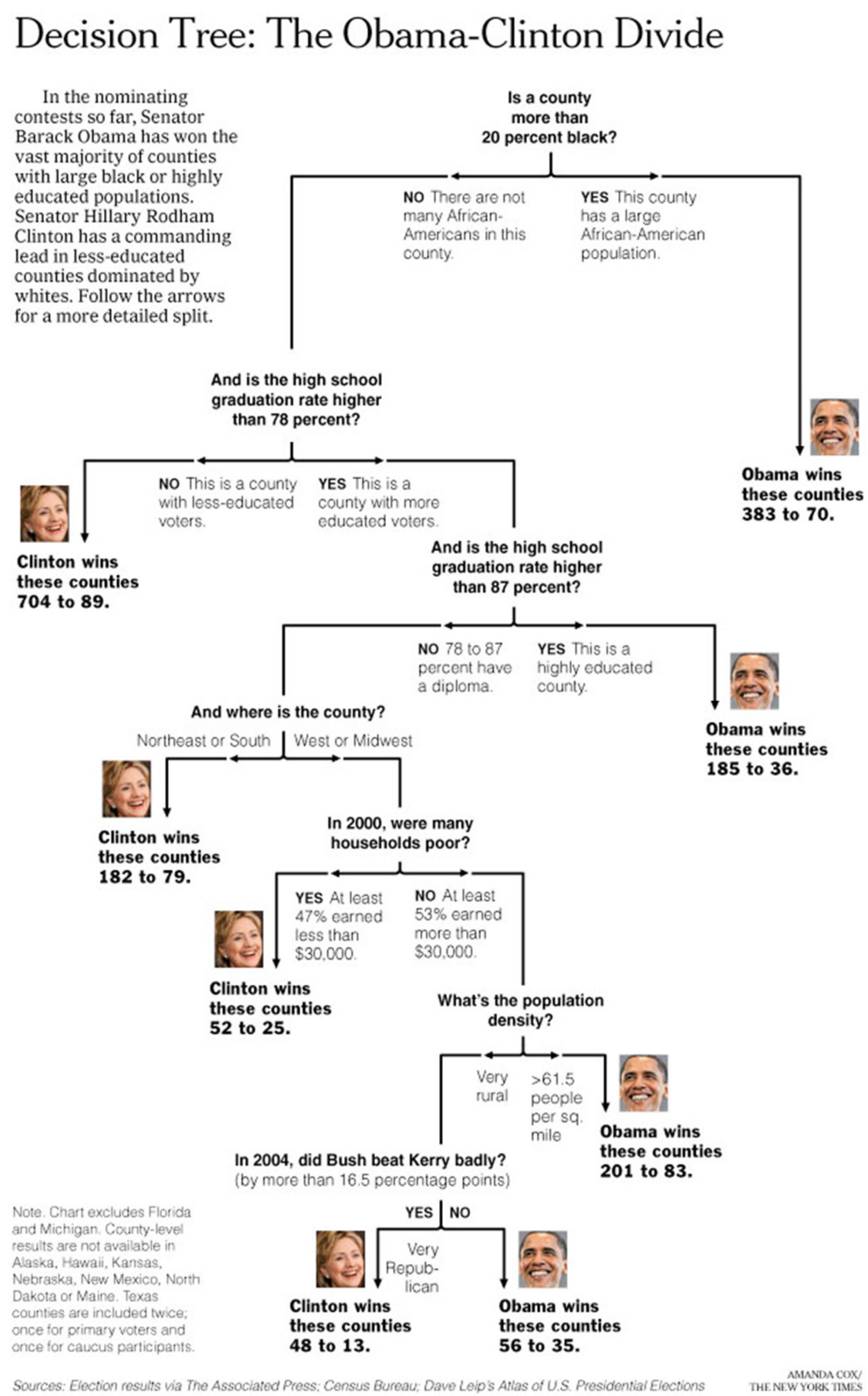
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



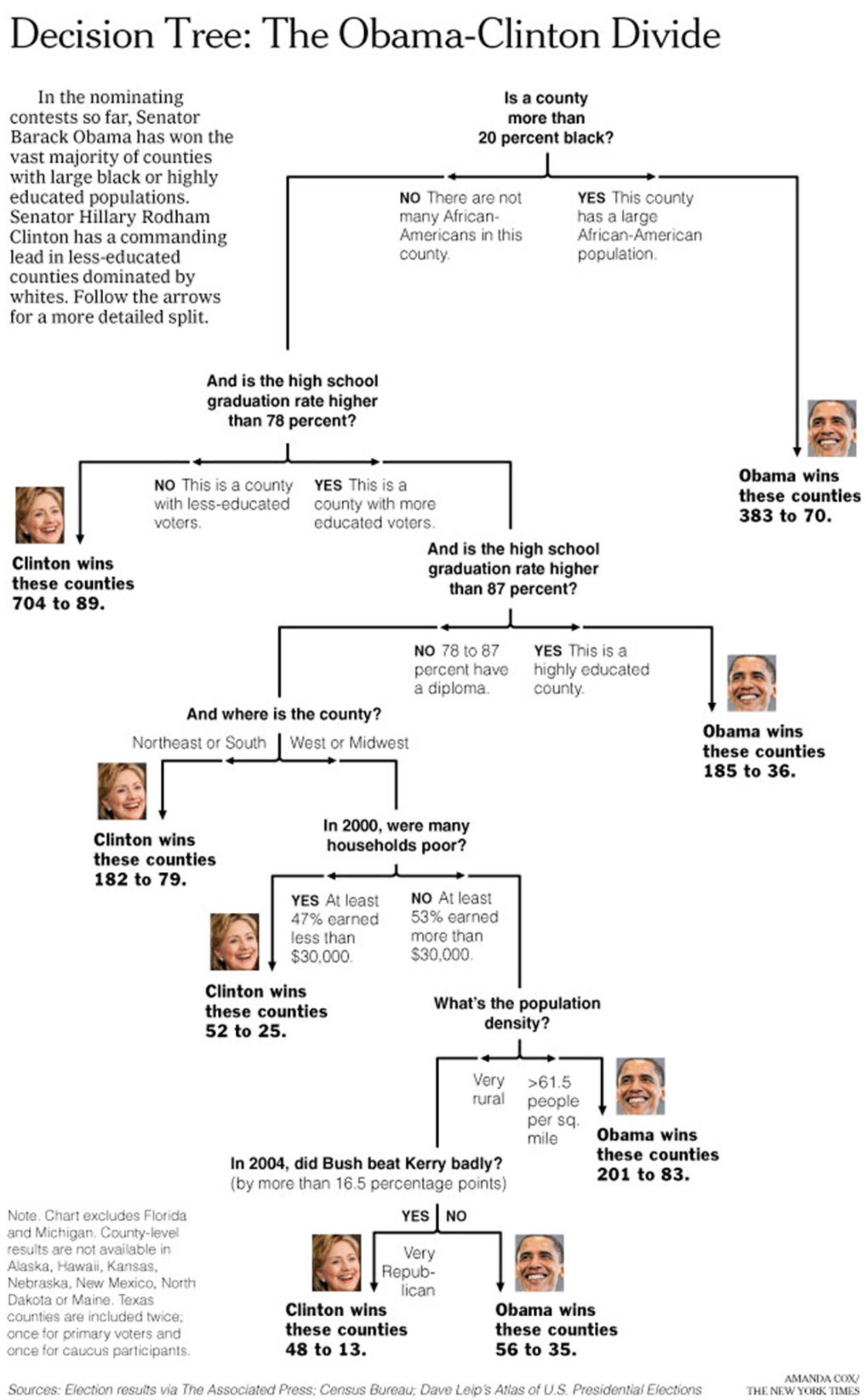
New York Times
April 16, 2008

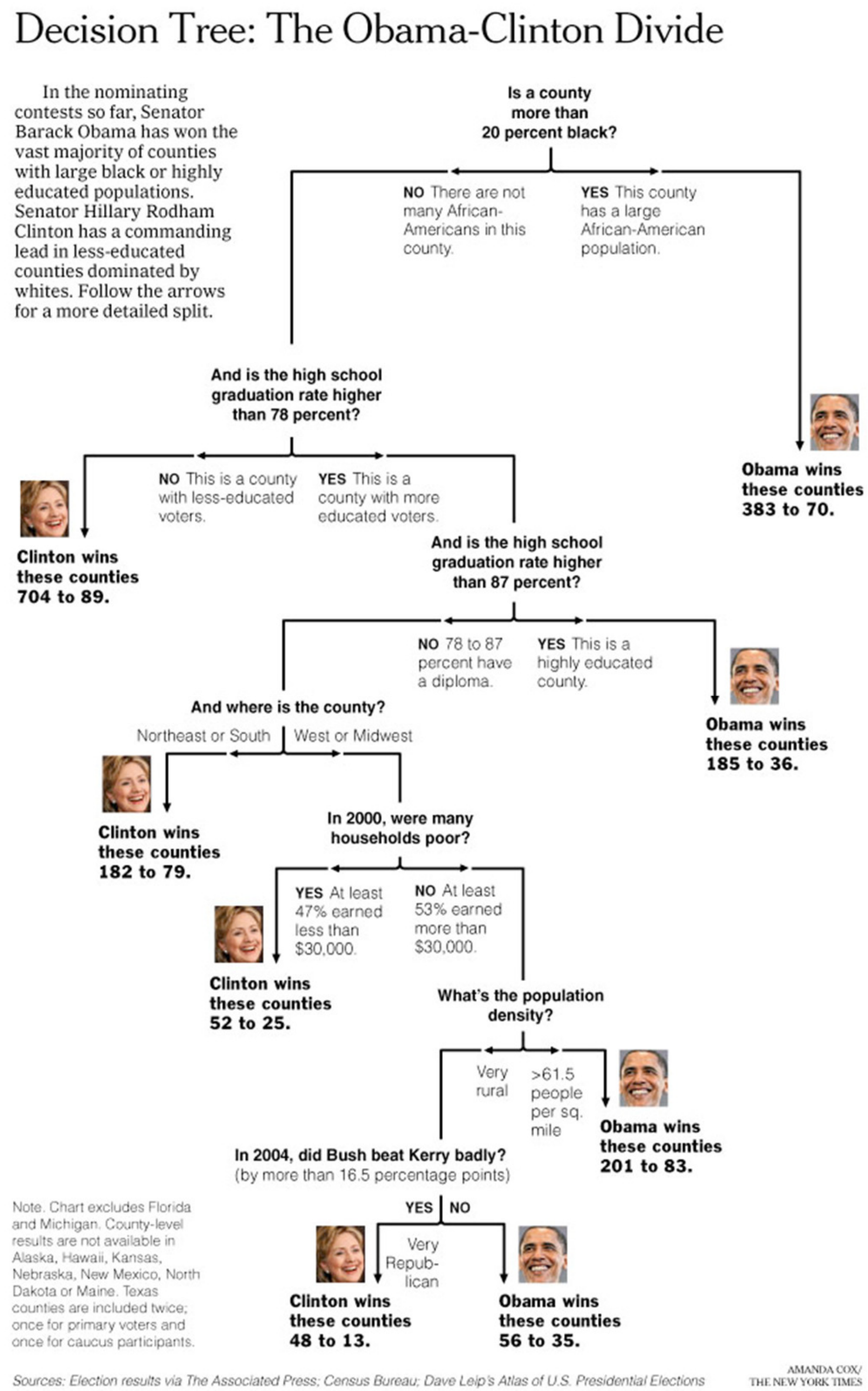
Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.

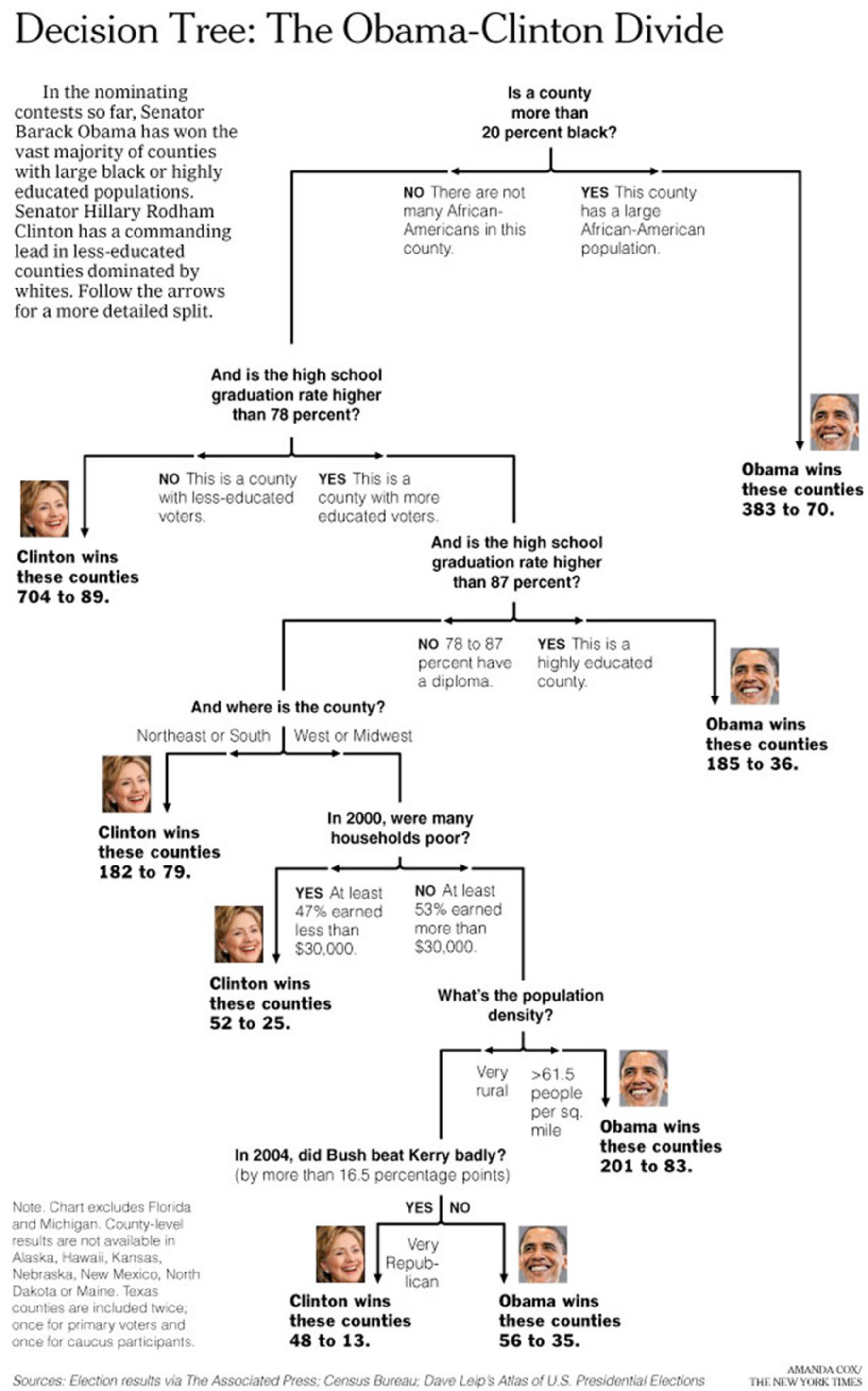




New York Times
April 16, 2008

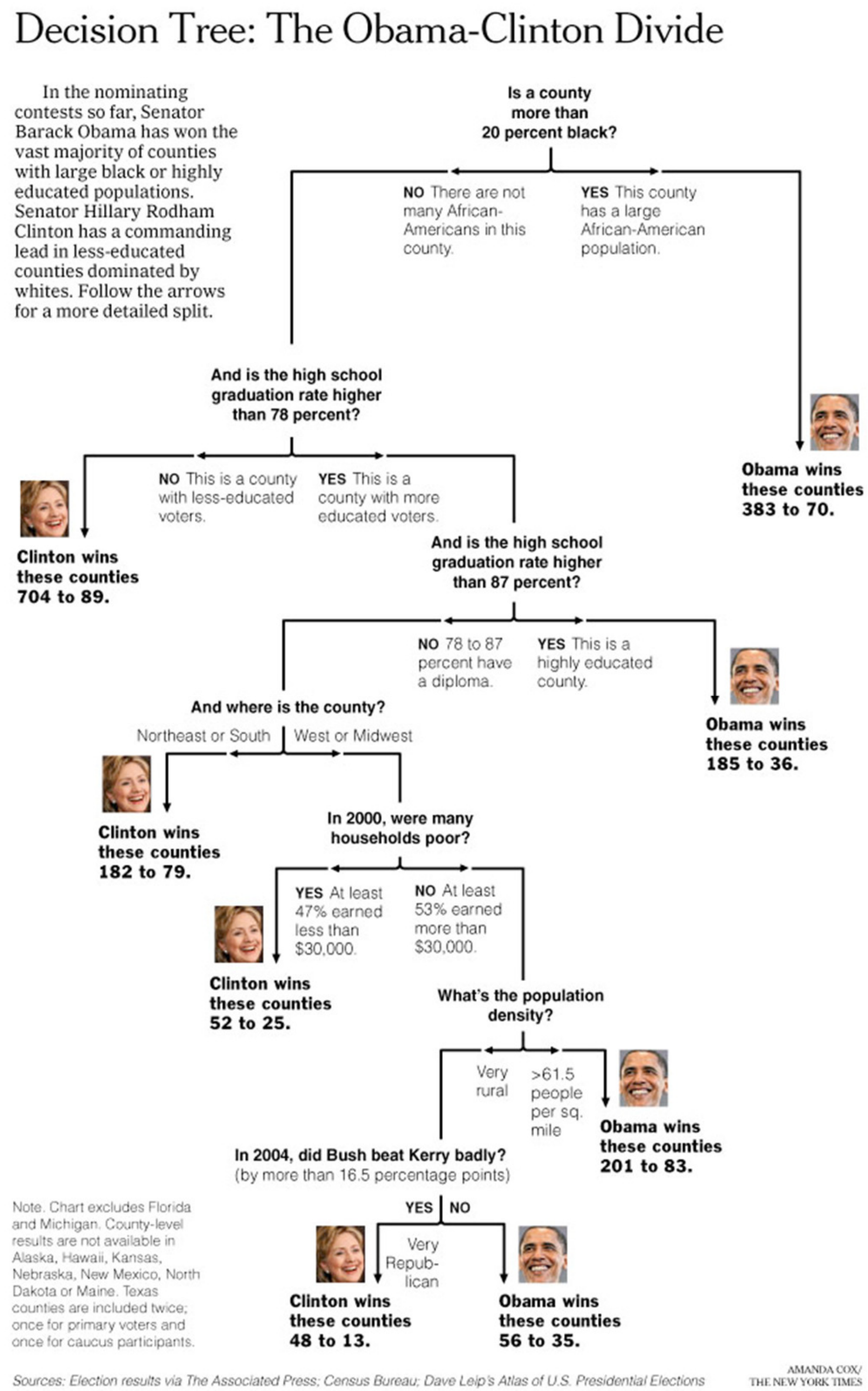
Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many “How to” manuals and trouble-shooting procedures.

New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many “How to” manuals and trouble-shooting procedures.

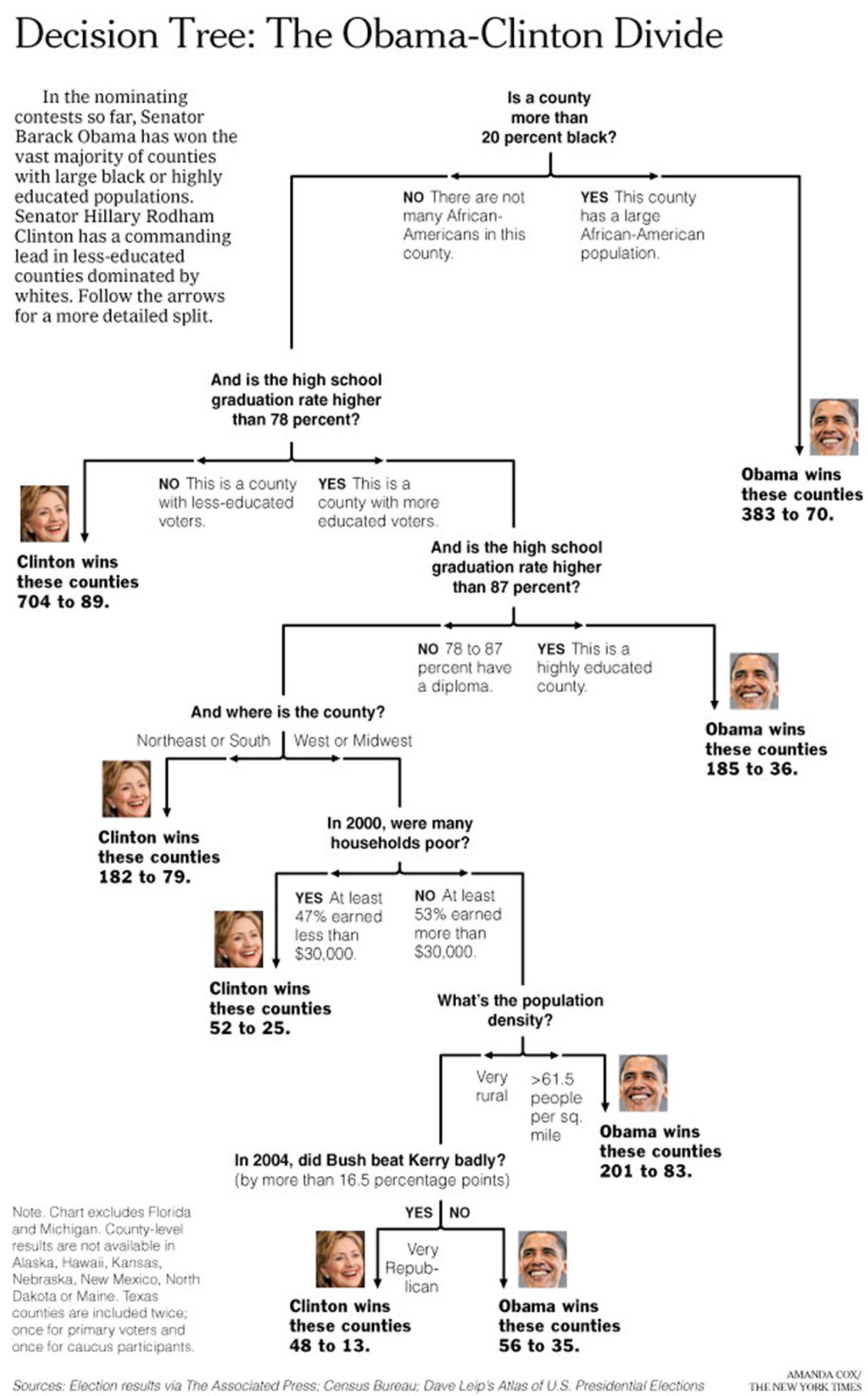
New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

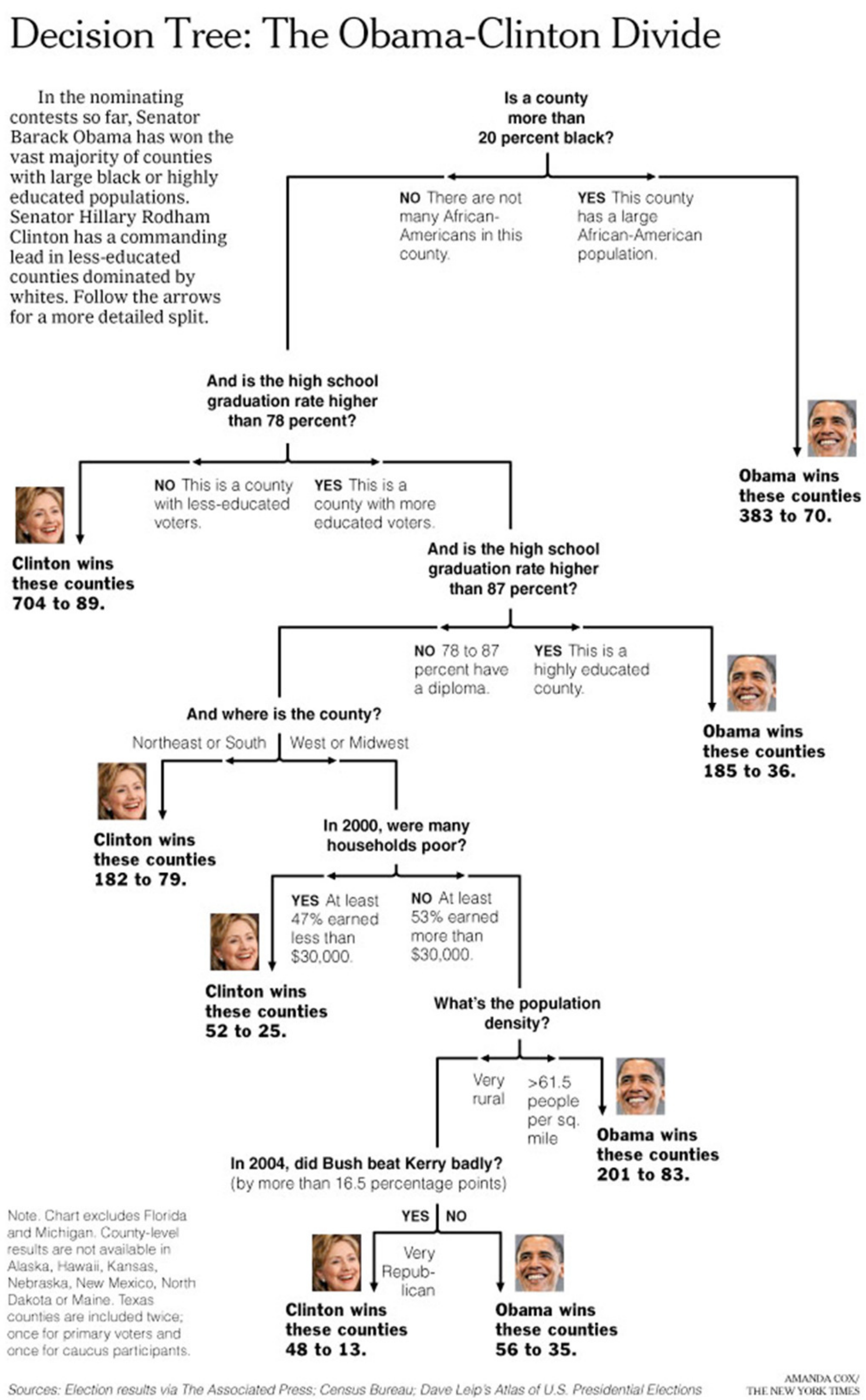
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



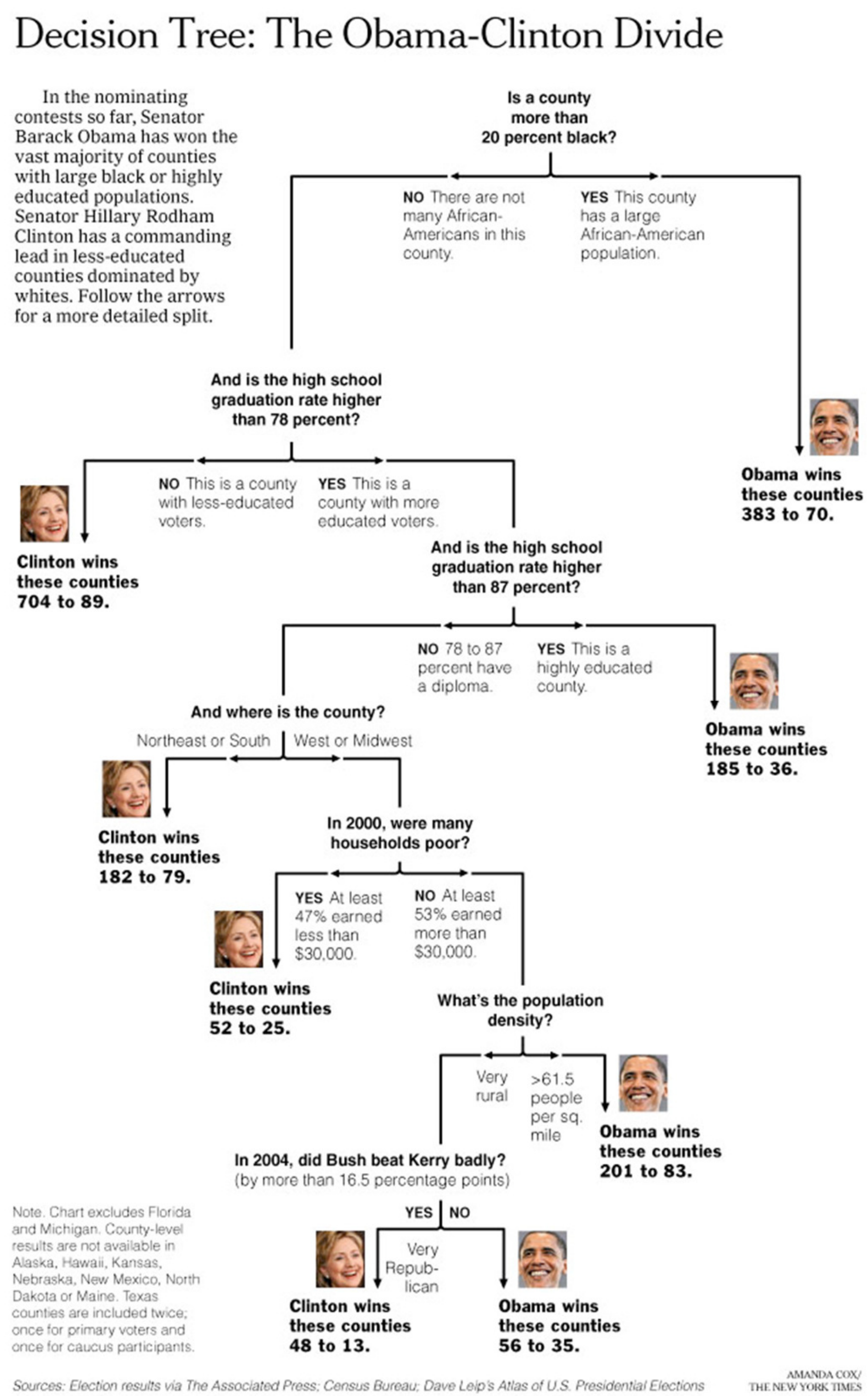
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



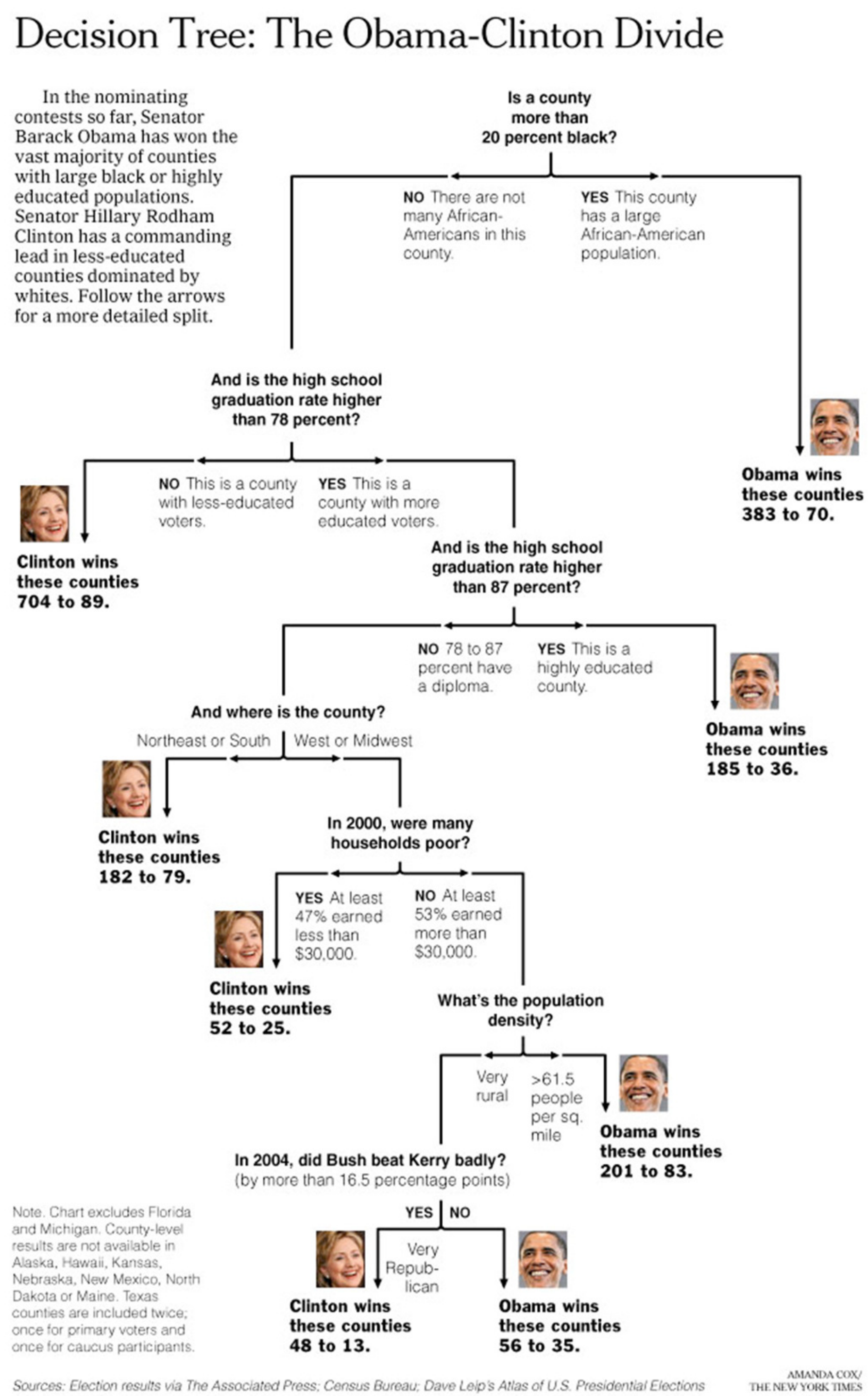
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.

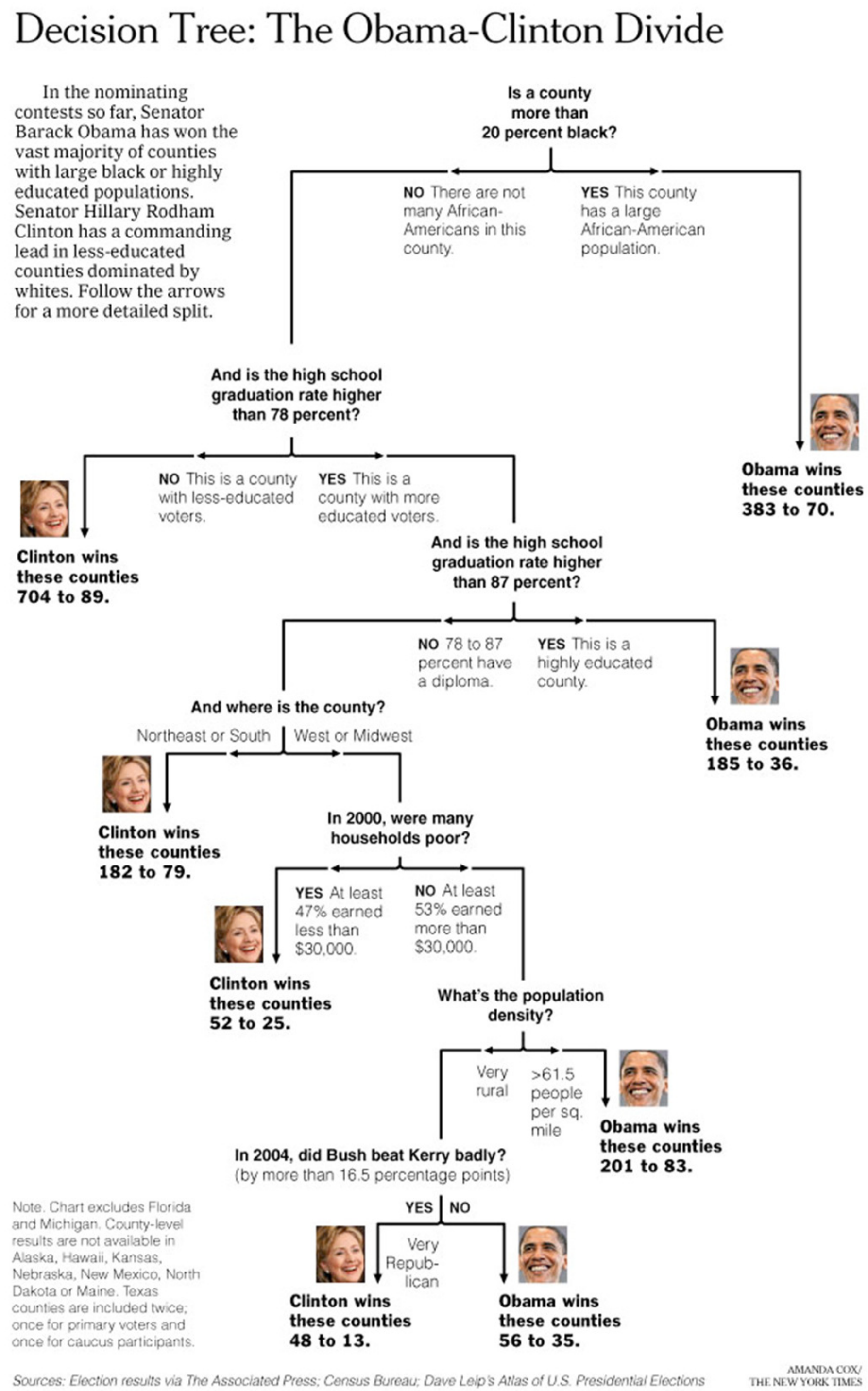


New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



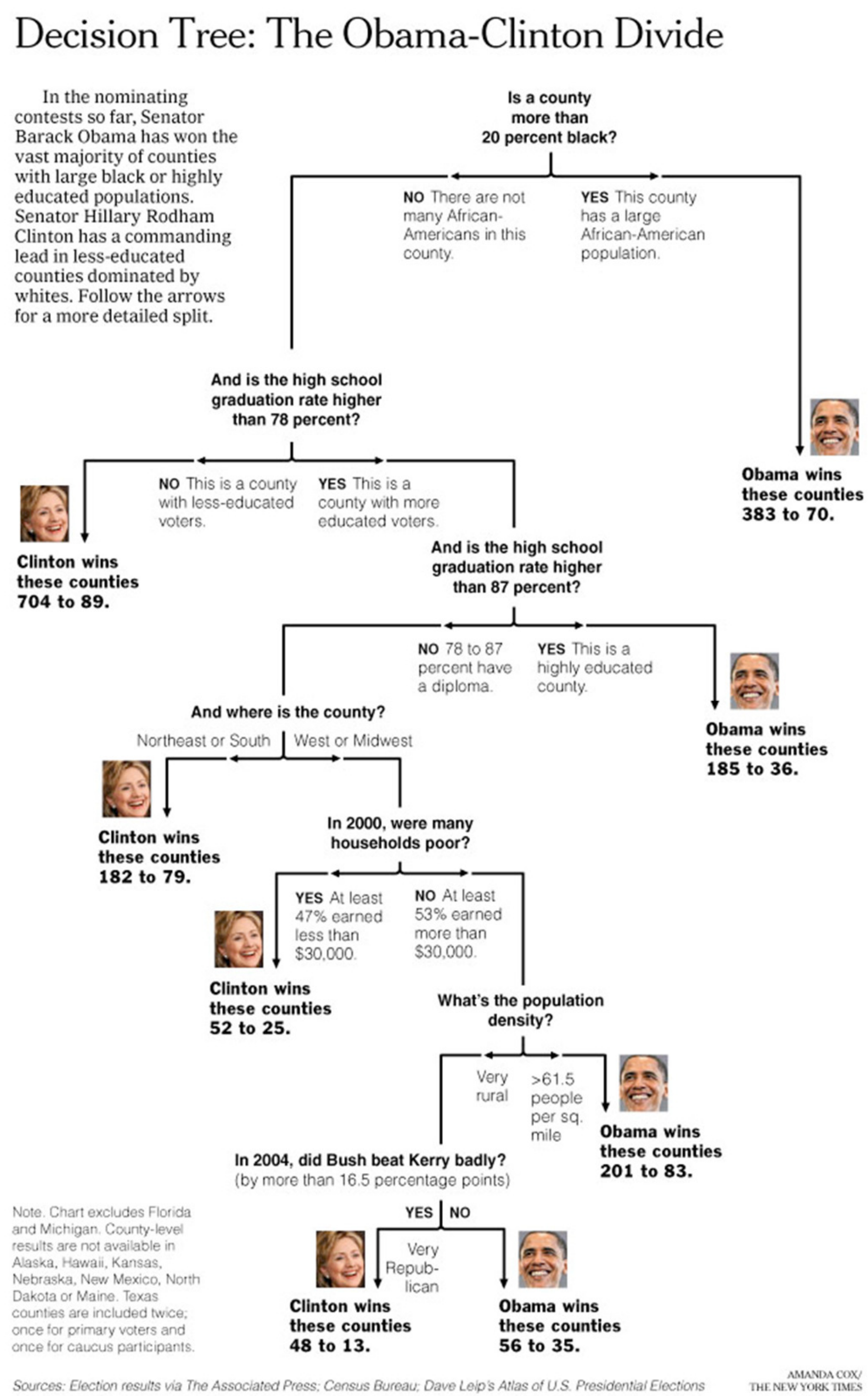
New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

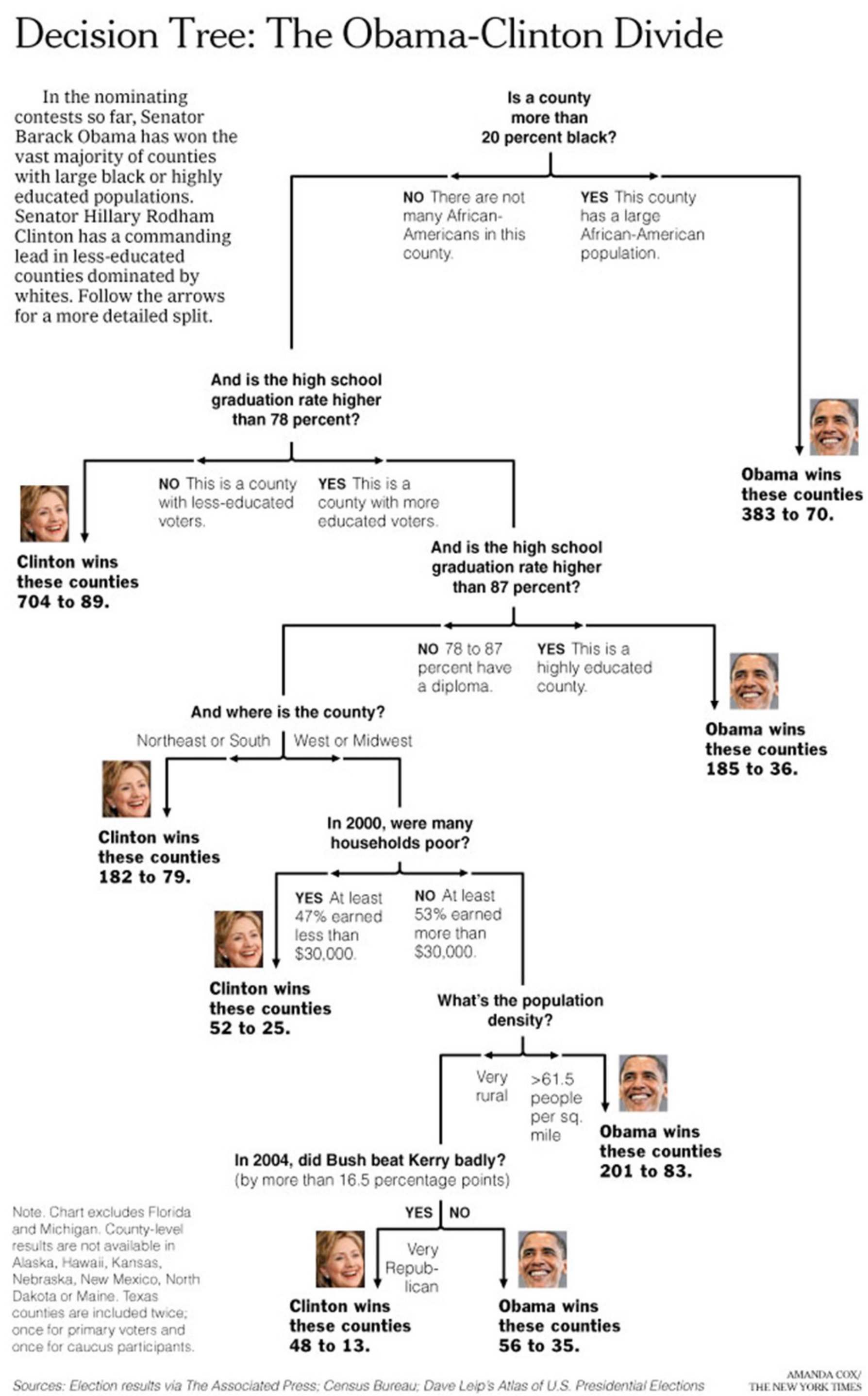
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many “How to”
manuals and trouble-shooting
procedures.



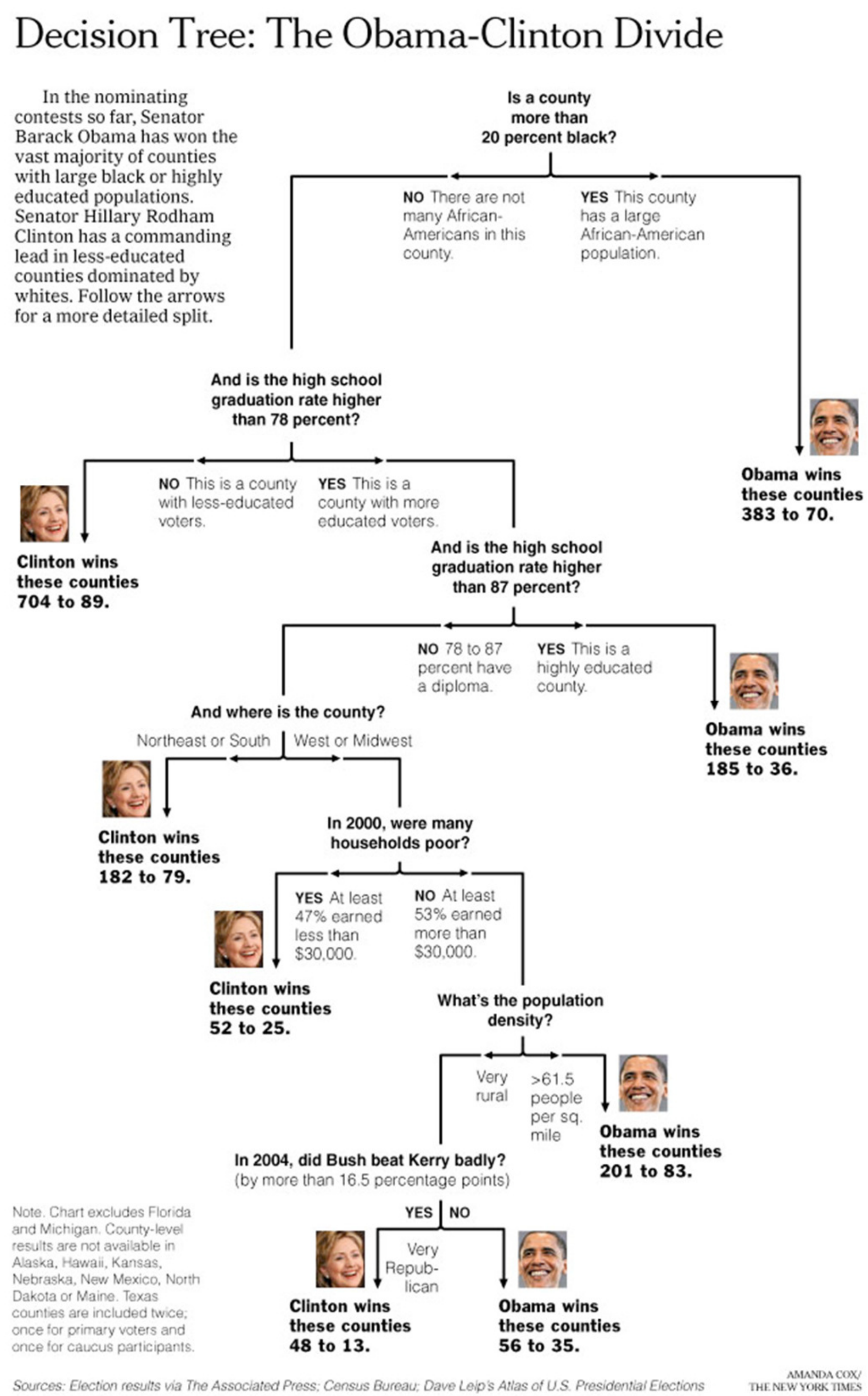
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



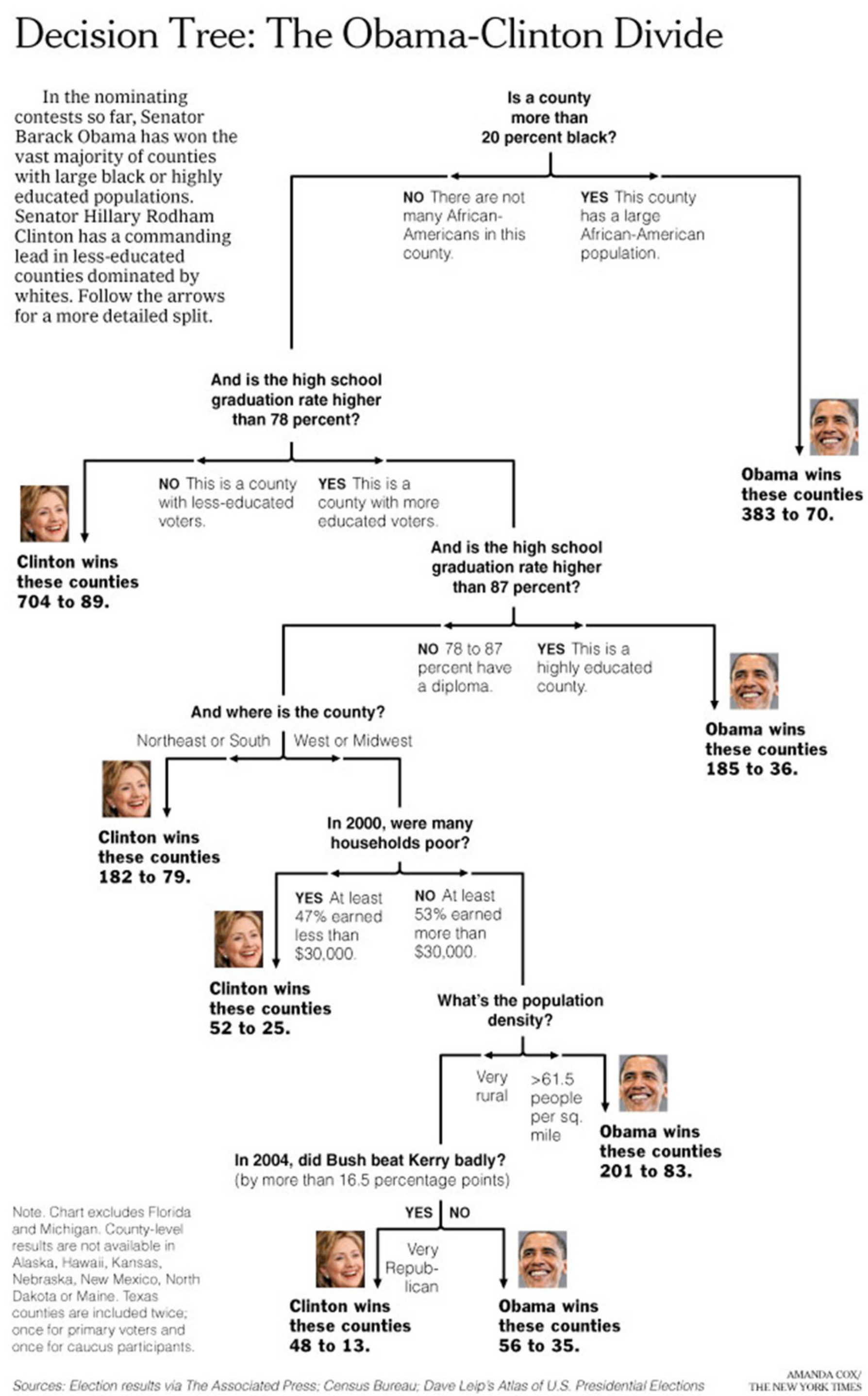
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

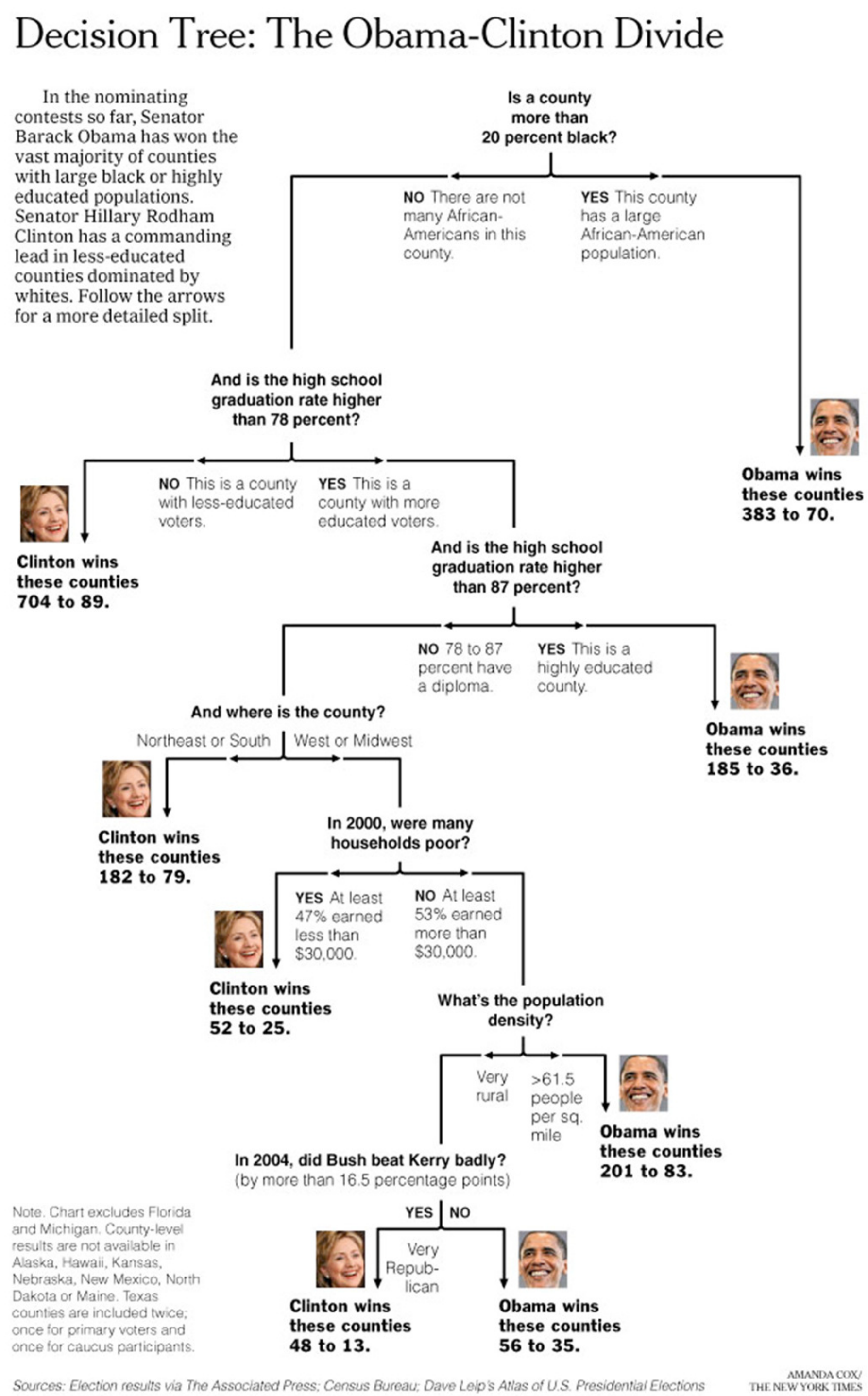


New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

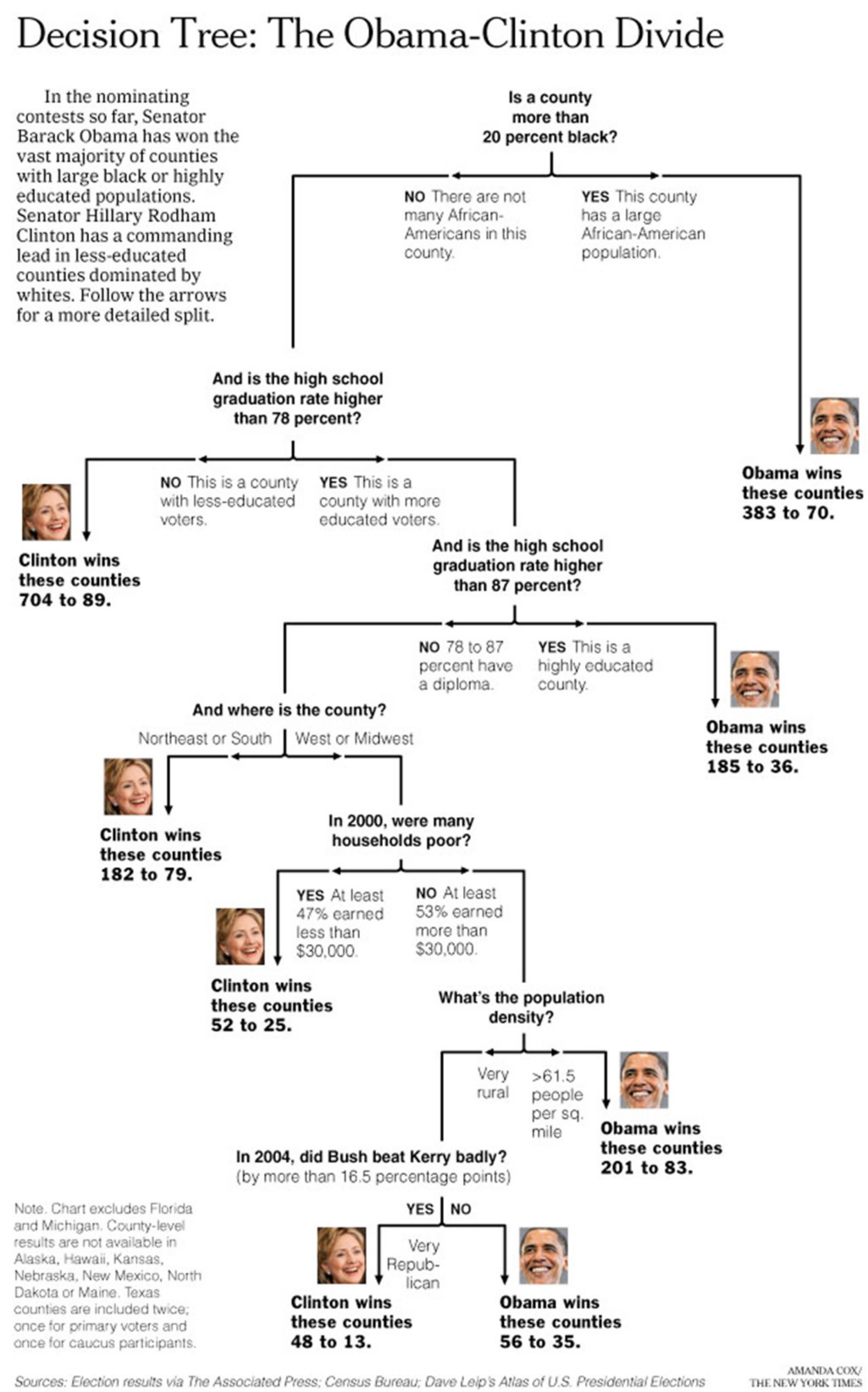


New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.

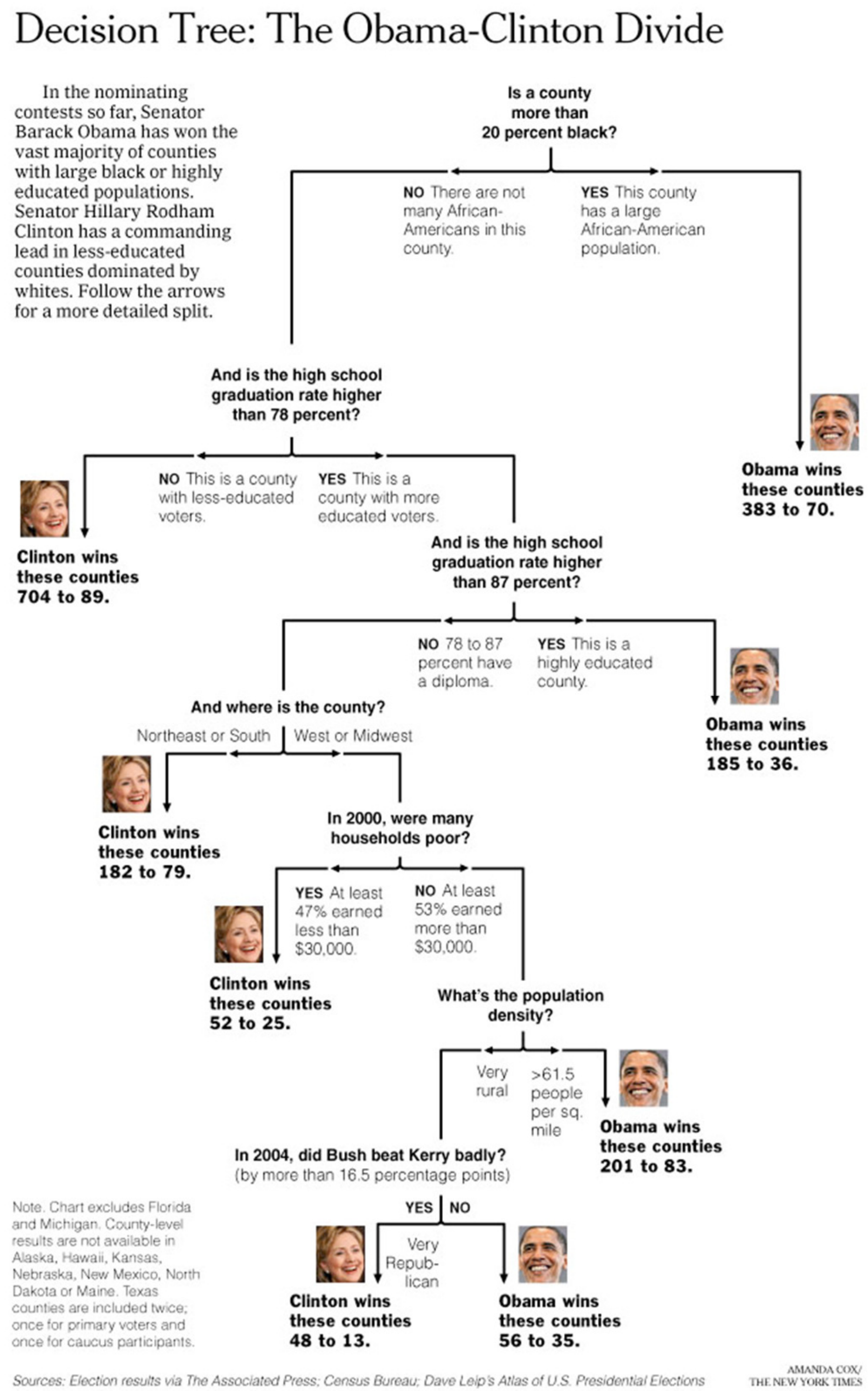
New York Times
April 16, 2008



Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

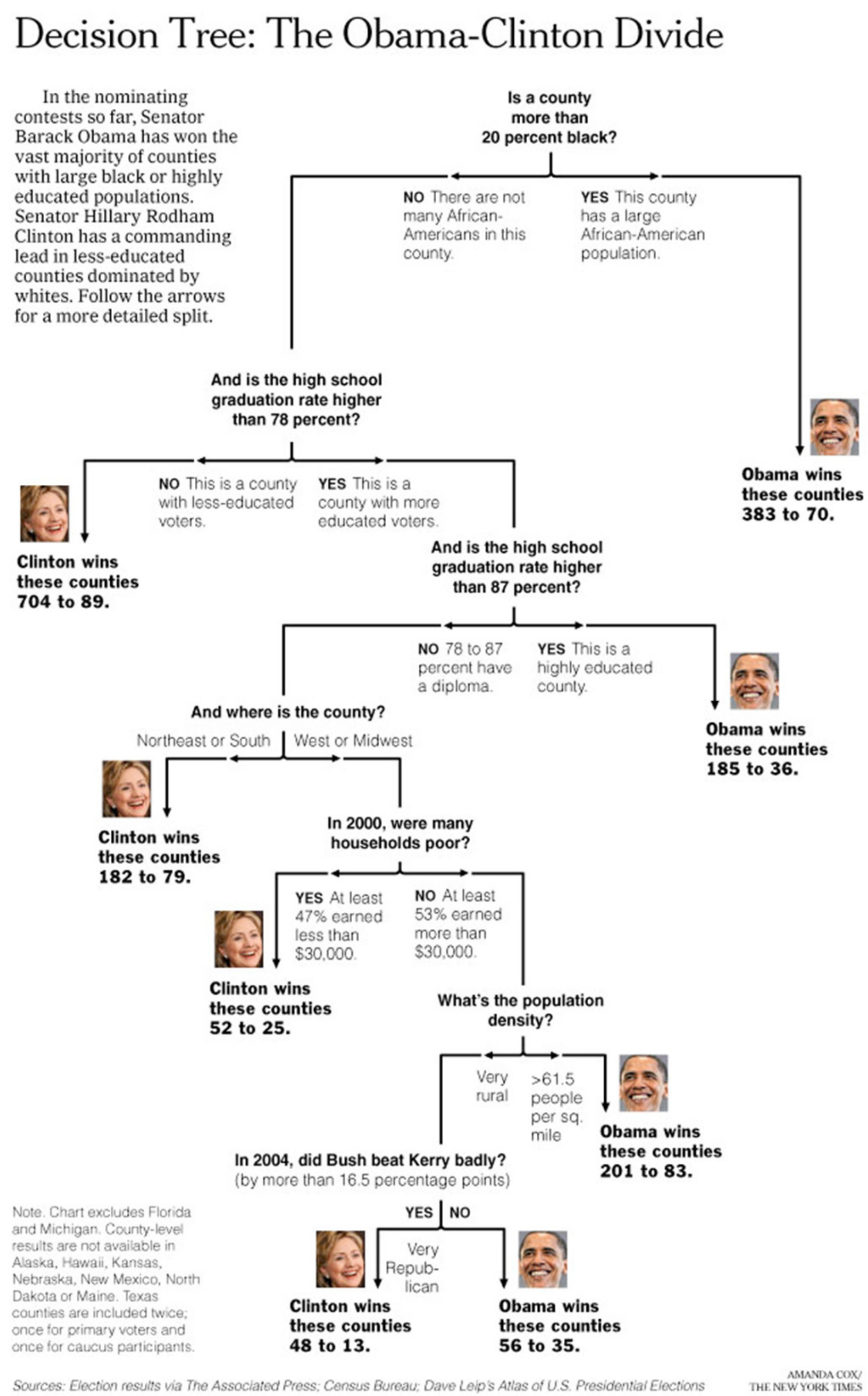
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



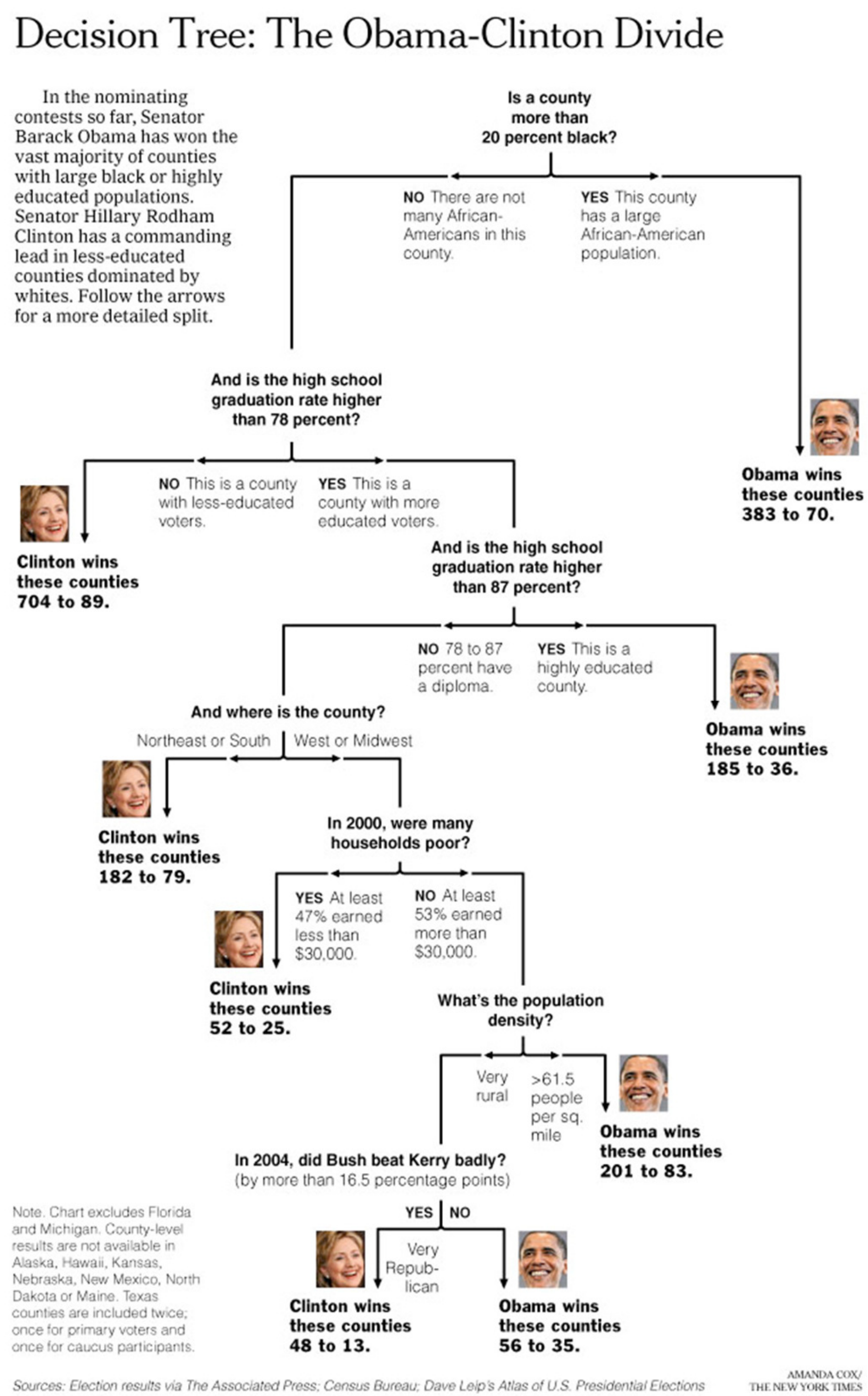
New York Times
April 16, 2008

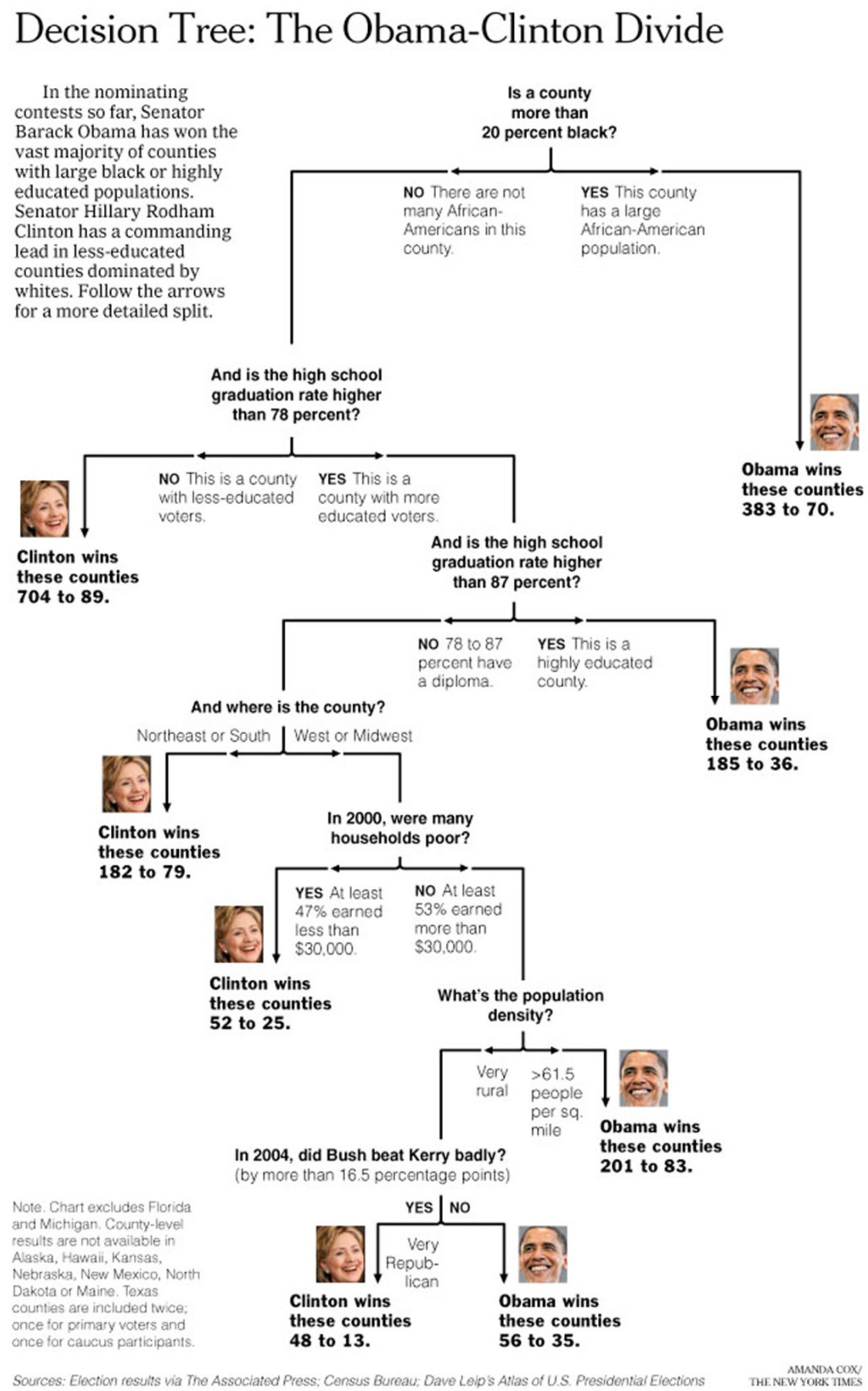
Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many “How to” manuals and trouble-shooting procedures.



New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



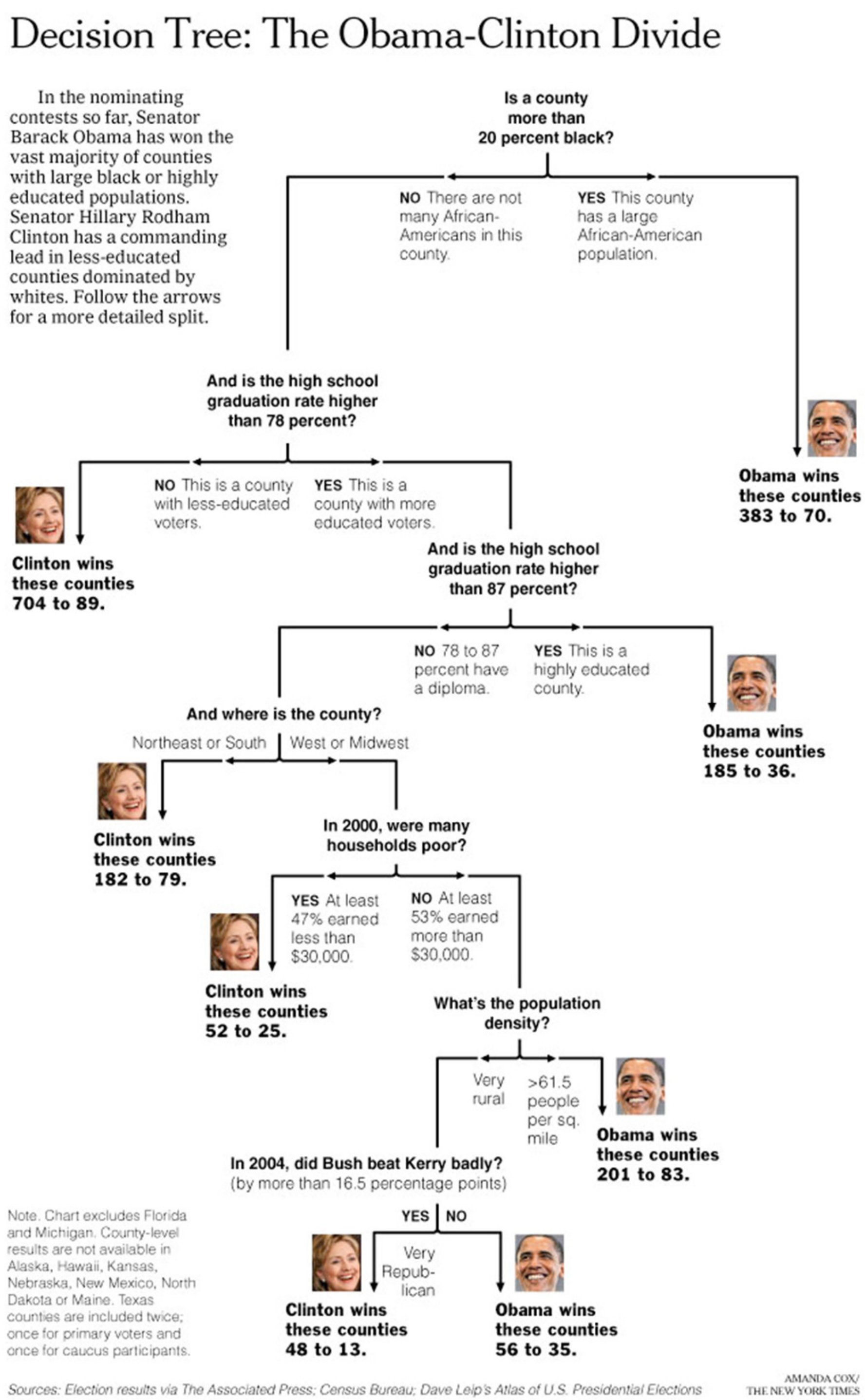


New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.

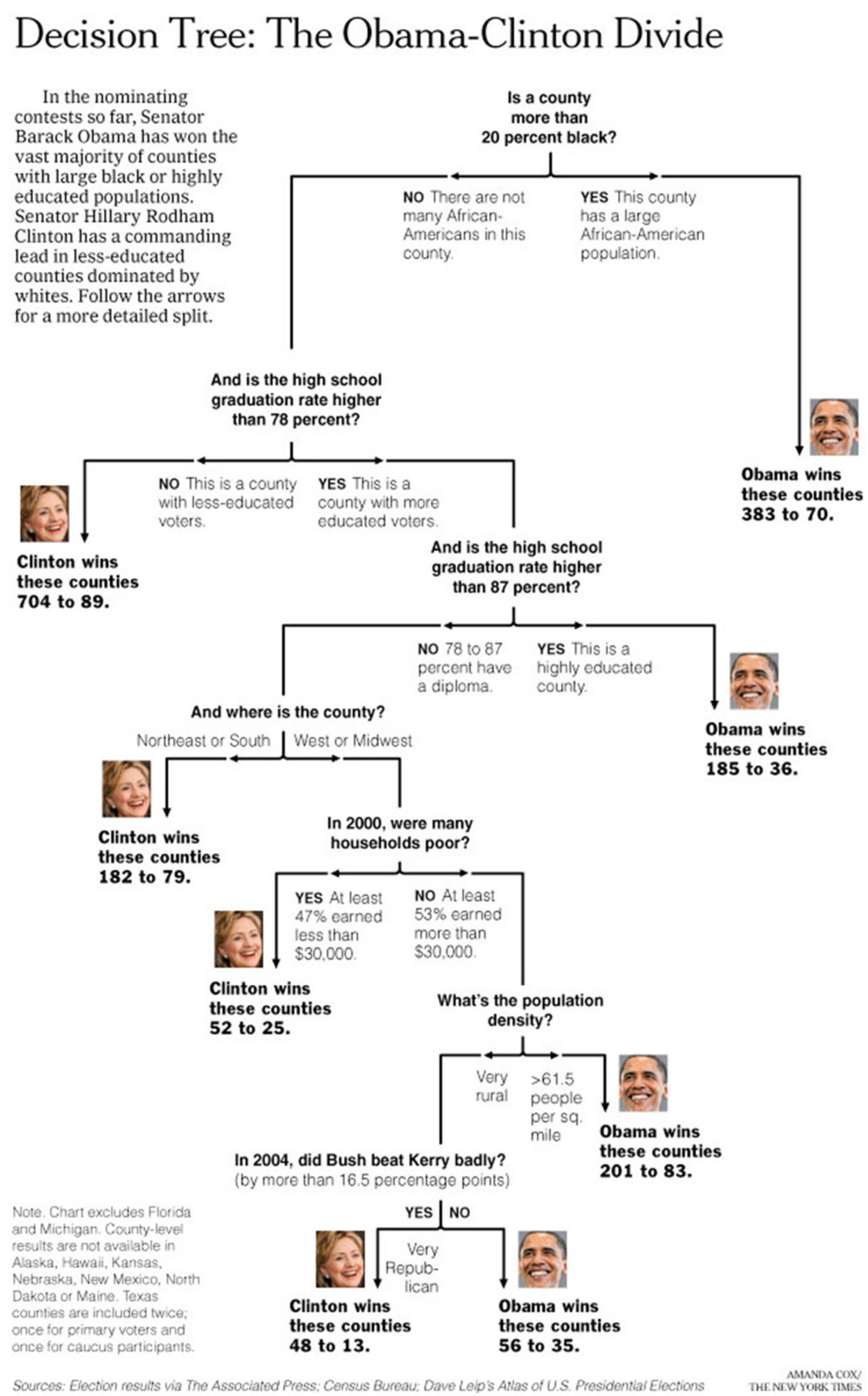
New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many "How to"
manuals and trouble-shooting
procedures.



New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural for humans.
Style of many "How to" manuals and trouble-shooting procedures.



New York Times
April 16, 2008

Decision Trees:
a sequence of tests.
Representation very natural
for humans.
Style of many “How to”
manuals and trouble-shooting
procedures.