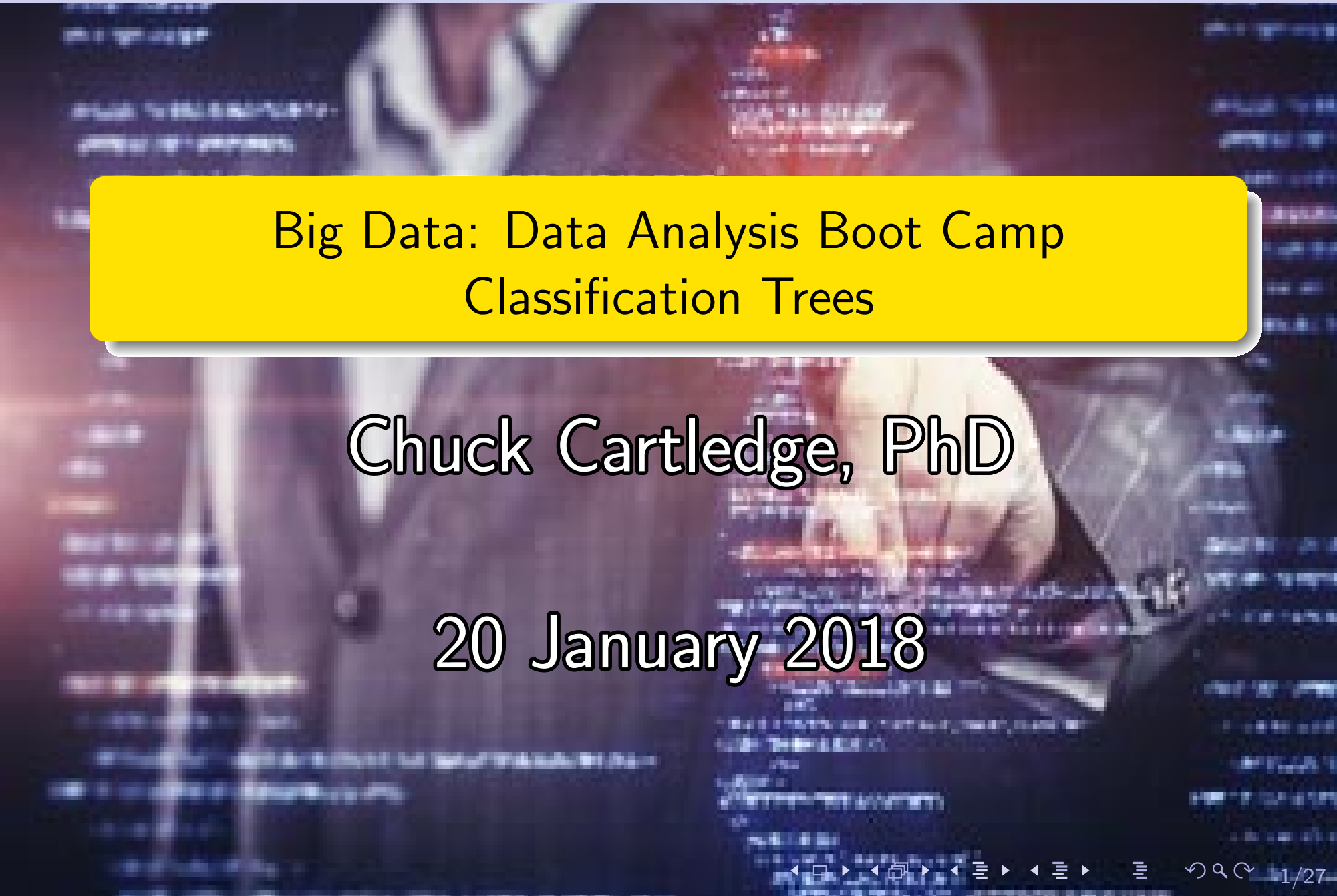


Big Data: Data Analysis Boot Camp
Classification Trees

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

1. Intro.
2. Background
   - Contextualize
3. Hands-on
   - Titanic
   - Other data
4. Q & A
5. Conclusion
6. References
7. Files
8. Misc.
   - Equations

What are we going to cover?

We're going to talk about:

- What are decision trees, and how can they be used to classify data.
- Construct decision trees for the Titanic and income datasets.

Putting things in perspective

"Decision trees . . . have a number of advantages over the more classical approaches . . .

- *Trees are very easy to explain to people. In fact, they are even easier to explain than linear regression.*
- *Some people believe that decision trees more closely mirror human decision-making than do the regression and classification approaches . . .*
- *Trees can be displayed graphically, and are easily interpreted even by a non-expert . . .*
- *Trees can easily handle qualitative predictors without the need to create dummy variables.*

Unfortunately trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches . . . "

James, et al. [1]

Simple information theory entropy[4]

These ideas are based on Shannon[2]:

# msgs	Entropy	Bits
1	$-\log_2(\frac{1}{1})$	0
2	$-\log_2(\frac{1}{2})$	1
10	$-\log_2(\frac{1}{10})$	3.3
10 (different prob.)	$-(\frac{4}{10}\log_2(\frac{4}{10}) + \frac{6}{10}\log_2(\frac{6}{10}))$	0.962

The general information theory entropy equation is:

$$H = -K\sum_{i=1}^{n} p_i \log_b p_i$$

Information gain

Information gain is the difference between entropy levels:

$$Ig = H_i - H_j$$

Decision tree induction algorithms

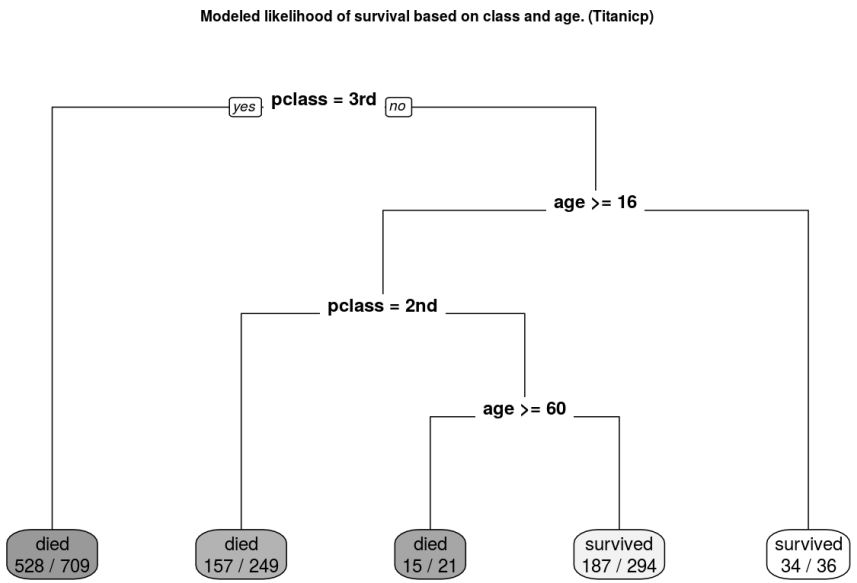
Different algorithms are appropriate for different data:

Name	Criteria	Useful with
ID3	Entropy, information gain partitioning the training records into successively purer subsets.	Classification
C4.5	Entropy and gain ratio	Classification
C5.0	Commercial version of C4.5. Support weighting and winnowing.	Classification
CART[1]	Gini index	Numerical outcomes
Random forest	Bagging and voting	

[1] Classification and regression trees (CART)

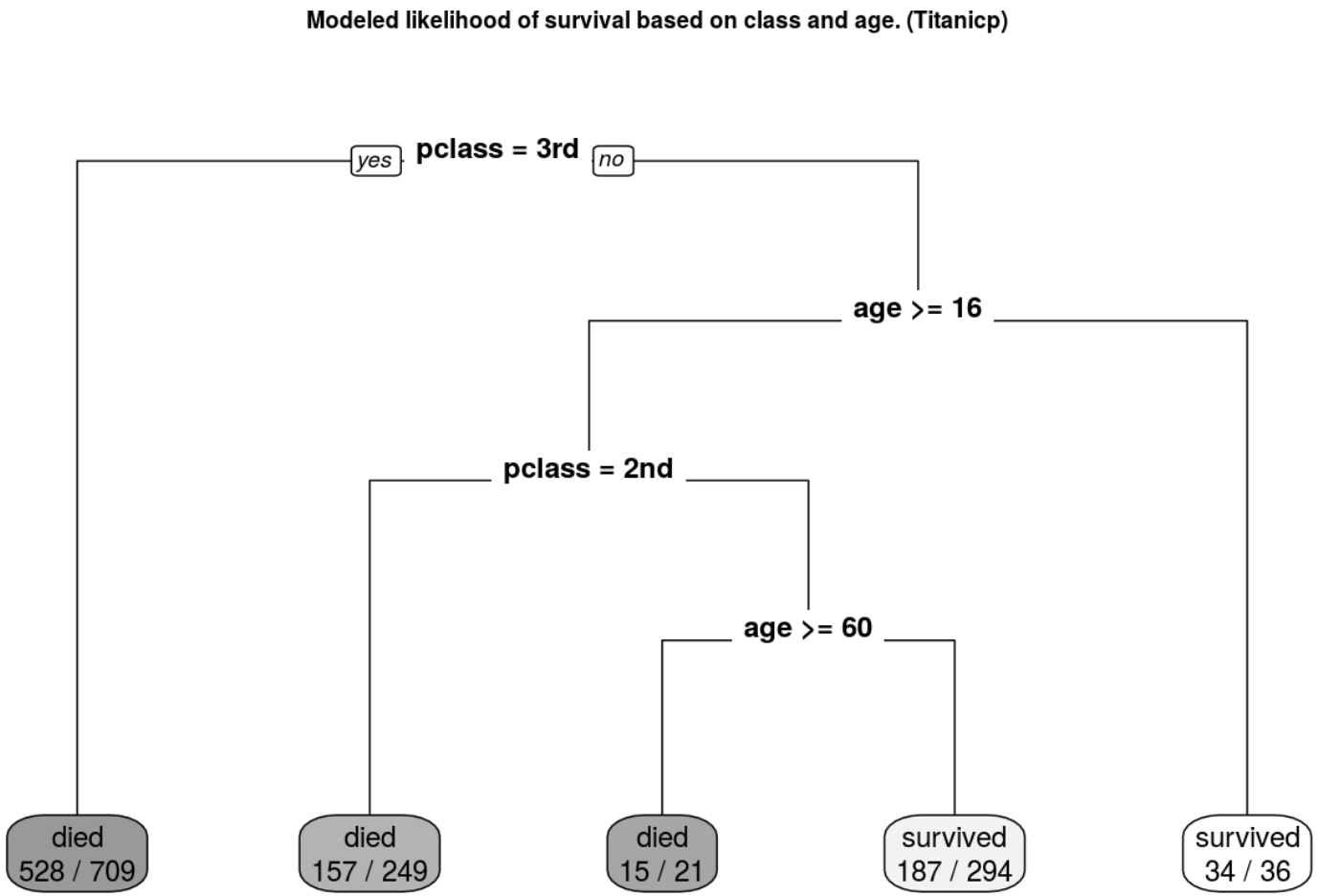
Another way to look at Titanic survivors

In this case, we look at survival as a function of age, and passenger class.



Attached file.

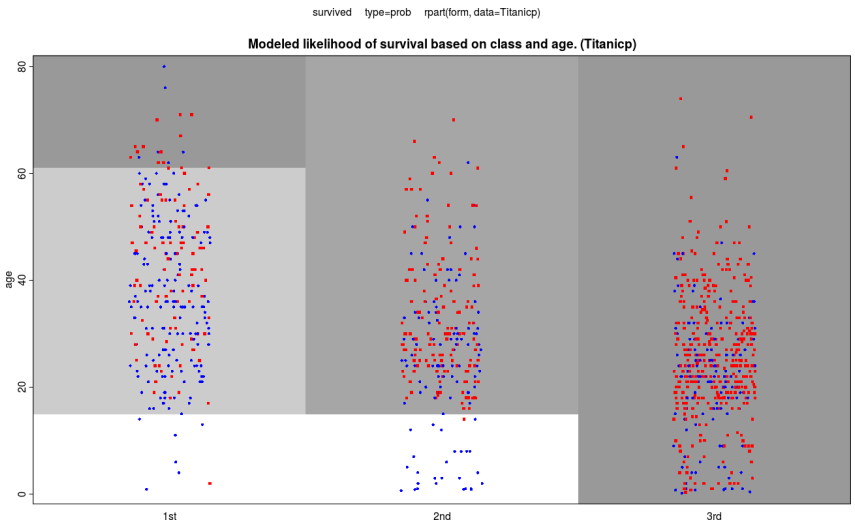
Same image.



Attached file.

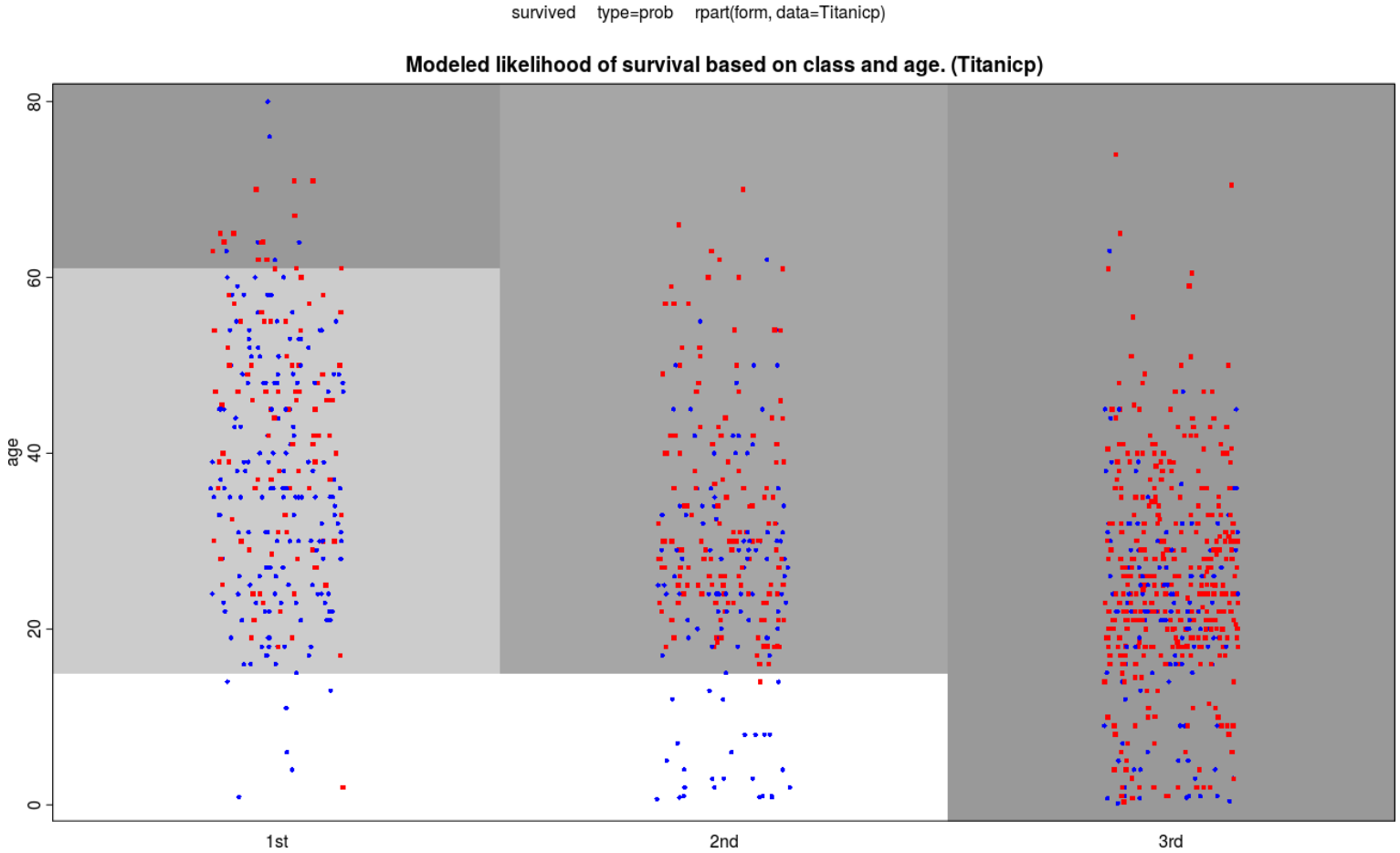
Still another way to look at the same data.

The shading matches the decision tree. Red means the person died, blue means lived. Crew members are not addressed in this plot. They can be.



Attached file.

Same image.



Attached file.

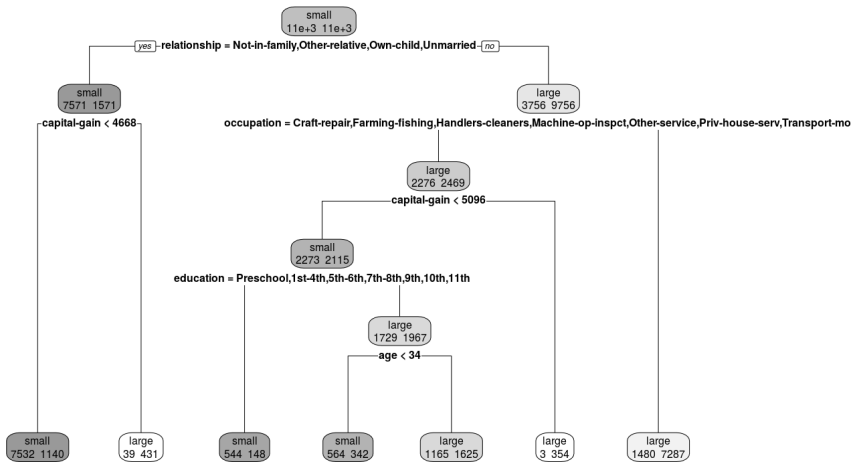
Looking at 1994 Adult Income

The "Adult" database was extracted from the census bureau database found at $< http://www.census.gov/ >$ in 1994 by Ronny Kohavi and Barry Becker, Data Mining and Visualization, Silicon Graphics. It was originally used to predict whether income exceeds USD 50K/yr based on census data. There are 48,842 records.

```
      age                    workclass          fnlwgt
 Min.   :17.00   Private             :33906   Min.   :  12285
 1st Qu.:28.00   Self-emp-not-inc:  3862   1st Qu.: 117550
 Median :37.00   Local-gov          :  3136   Median : 178144
 Mean   :38.64   State-gov          :  1981   Mean   : 189664
 3rd Qu.:48.00   Self-emp-inc       :  1695   3rd Qu.: 237642
 Max.   :90.00   (Other)            :  1463   Max.   :1490400
                 NA's               :  2799
```

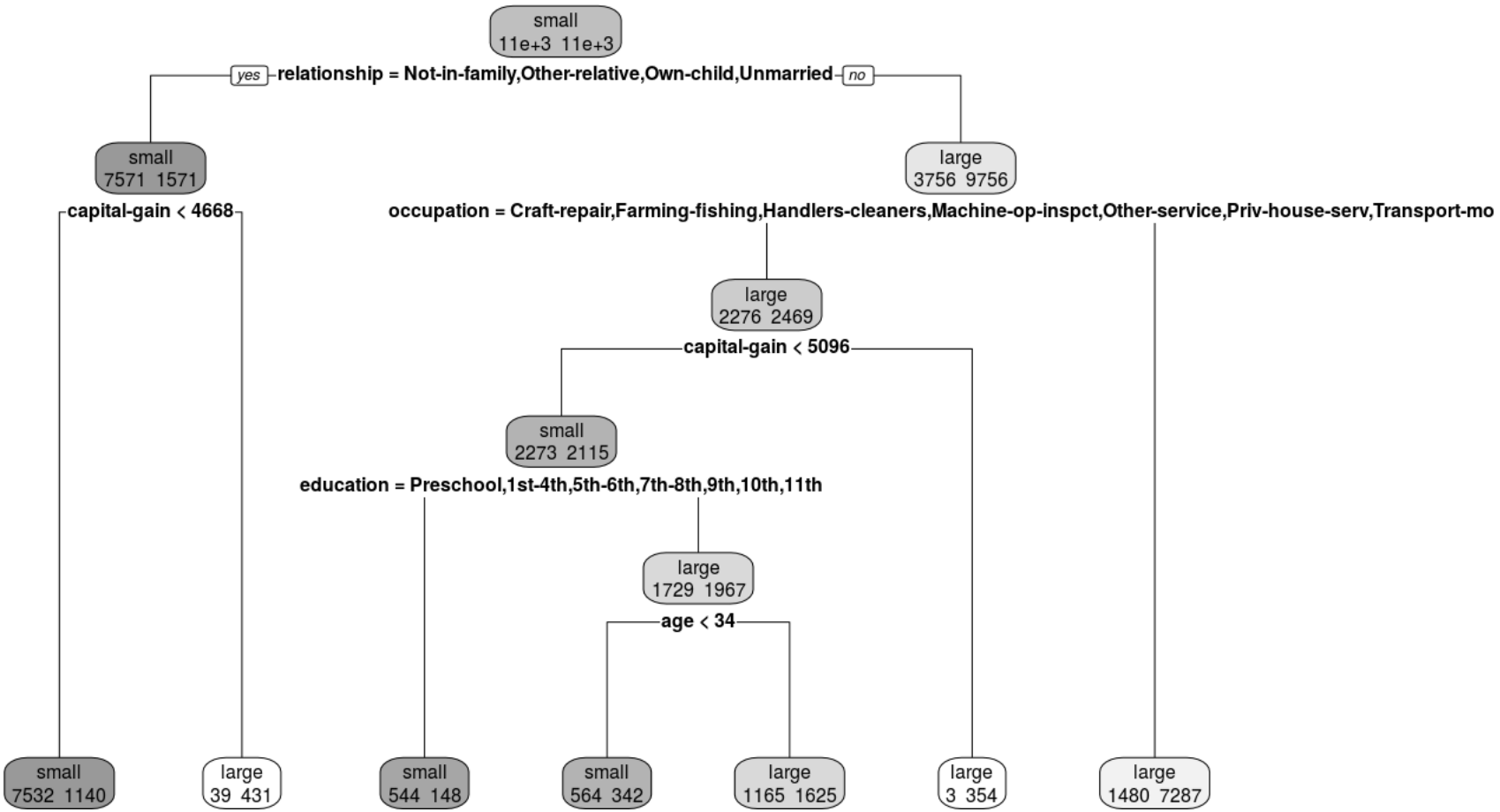
Complete decision tree based on all non-NA data

Based on the summary data, approx 89% of the data records are modeled correctly.



Attached file.

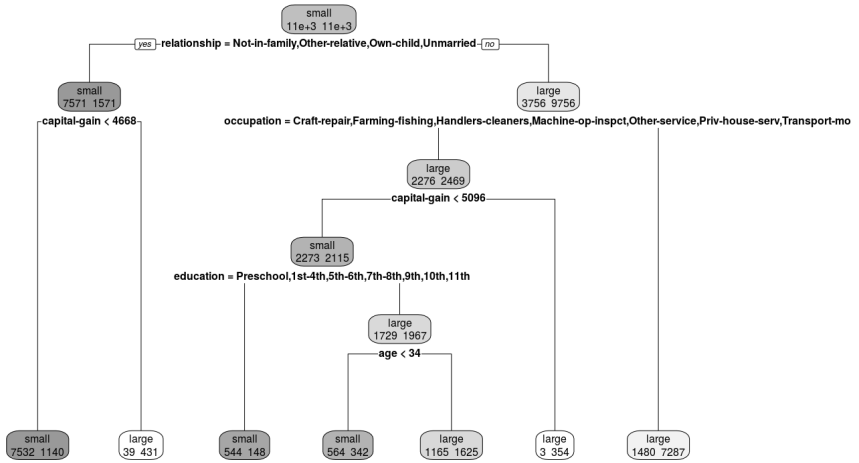
Same image.



Attached file.

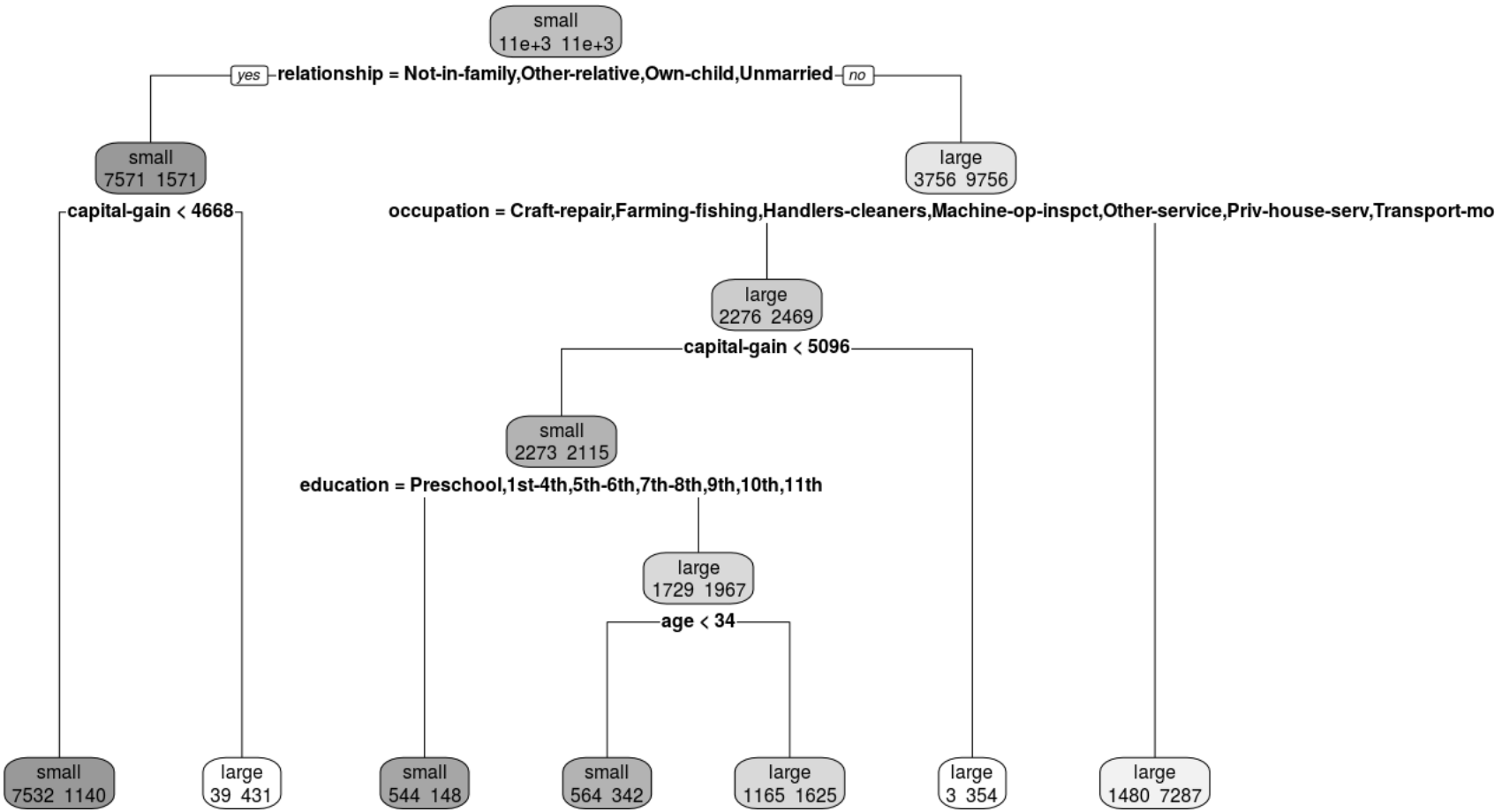
A pruned decision tree based on all non-NA data

Accuracy is about 46% with a much smaller and more easily understood tree.



Attached file.

Same image.



Attached file.

How accurate are the different trees?

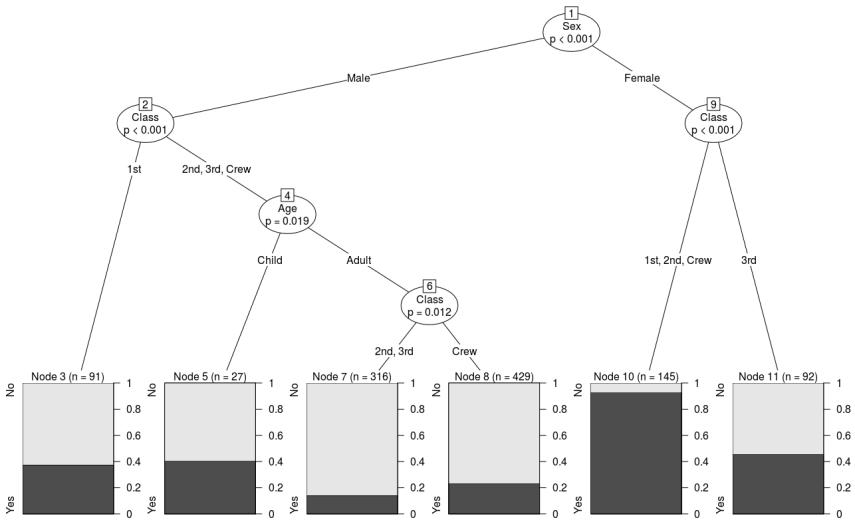
Lets look at all the models we've used:

	C45	C45pr	C5.0	CART	RF
True +	9400.0000000	9352.0000000	9438.0000000	7543.0000000	1.129900e+04
True -	2737.0000000	2920.0000000	2949.0000000	3338.0000000	8.130000e+02
False +	1927.0000000	1975.0000000	1889.0000000	3784.0000000	2.800000e+01
False -	1017.0000000	834.0000000	805.0000000	416.0000000	2.941000e+03
Accuracy	0.8047875	0.8137391	0.8213646	0.7215039	8.031298e-01
Kappa	0.5170644	0.5478142	0.5643192	0.4270395	2.890857e-01

The higher the accuracy the better. A Kappa greater than 0.60 is desirable.

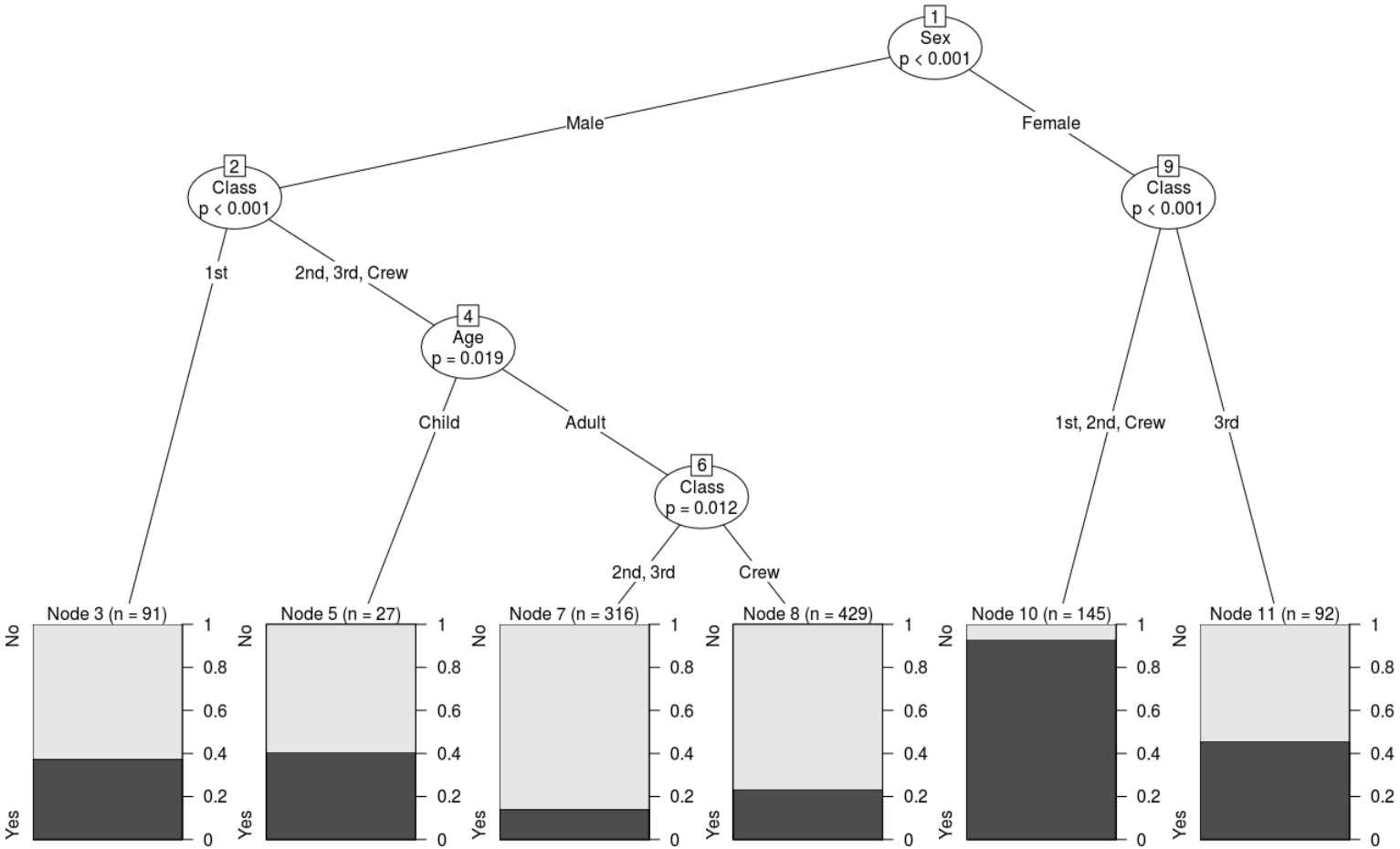
One more time into the Titanic data.

Using a different dataset and different tools.
If you were a female in 1st, 2nd, or crew, then you probably survived.



Attached file.

Same image.



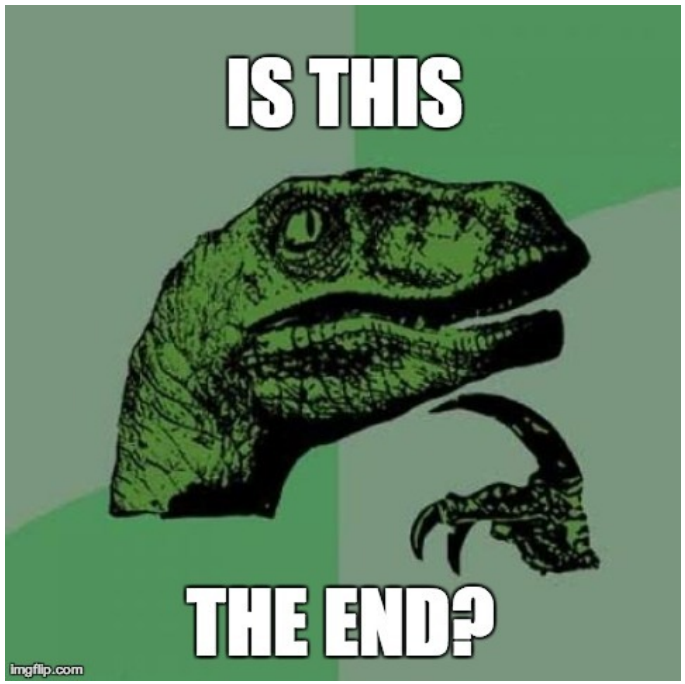
Attached file.

Q & A time.

Q: What is the burning question on the mind of every dyslexic existentialist?
A: “Is there a dog?”

What have we covered?

- Discussed the strengths and weaknesses of decision trees
- Discussed different algorithms for creating decision trees
- Created decision trees for modeling the Titanic survivors
- Created decision tree for modeling income based on census data



Next: LPAR Chapter 13, text analysis

References (1 of 1)

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning, vol. 6, Springer, 2013.

[2] C.E. Shannon, A Mathematical Theory of Communication, The Bell System Technical Journal **27** (1948), 379–423.

[3] P.Mean Staff, What is a Kappa coefficient? (Cohen's Kappa), `http://www.pmean.com/definitions/kappa.htm`, 2008.

[4] Scott Uminsky, Information Theory Demystified, `http://www.ideacenter.org/contentmgr/showdetails.php/id/1236`, 2004.

Files of interest

1. Revised classification script 
2. R library script file 

Information Theory Entropy (1 of 2) [2]

Shannon built the case that symbols in communication did not occur in random order, but that there was a transition probability to symbol s_{i+1} from symbol s_i.
He based his idea on this figure:

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2}H(\tfrac{2}{3}, \tfrac{1}{3})$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

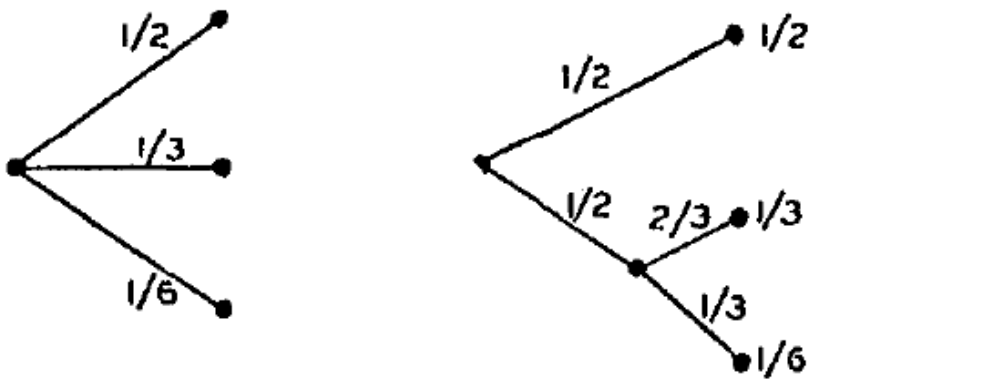


Fig. 6—Decomposition of a choice from three possibilities.

Image from [2].

Where $H()$ had to equal 1. Meaning that: $H = -K \sum_{i=1}^{n} p_i \log_b p_i$

Information Theory Entropy (2 of 2) [2]

The selection of b results in H having different units:

b	Units
2	bit
e	nat
10	Hart

The change from base a to base b merely requires multiplication by $\log_b a$.

Gini Index

"...the Gini index, is a referred to as a measure of node purity – a small value indicates that a node contains predominantly observations from a single class"

James, et al. [1]

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

Cohen's Kappa[3]

Kappa measures the percentage of data values in the main diagonal of a confusion matrix and adjusts these values for the amount of agreement that could be expected due to chance alone. By way of example:
Two raters are asked to classify objects into categories 1 and 2, and result in a confusion matrix.

		Rater #1		
		1	2	Total
Rater #2	1	p_{11}	p_{12}	p_1
	2	p_{21}	p_{22}	p_2
	Total	p_1	p_2	1

$$
\begin{aligned}
p_o &= p_{11} + p_{22} \\
p_e &= p_1 p_1 + p_2 p_2 \\
k &= \frac{p_o - p_e}{1 - p_e}
\end{aligned}
$$

Kappa is always less than or equal to 1. A value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement. Negative values indicate serious differences between the raters.