CART for letter-to-sound
worked example for the letter "a"

predictors							predictee
ppp	pp	p		n	nn	nnn	
o	g	w	**a**	s	h	-	aa
a	r	w	**a**	s	h	-	aa
-	-	w	**a**	r	r	a	ao
-	-	w	**a**	r	r	a	ao
r	d	w	**a**	r	e	-	eh
f	t	w	**a**	r	e	-	eh
-	-	w	**a**	r	e	s	eh

Entropy of distribution at the root
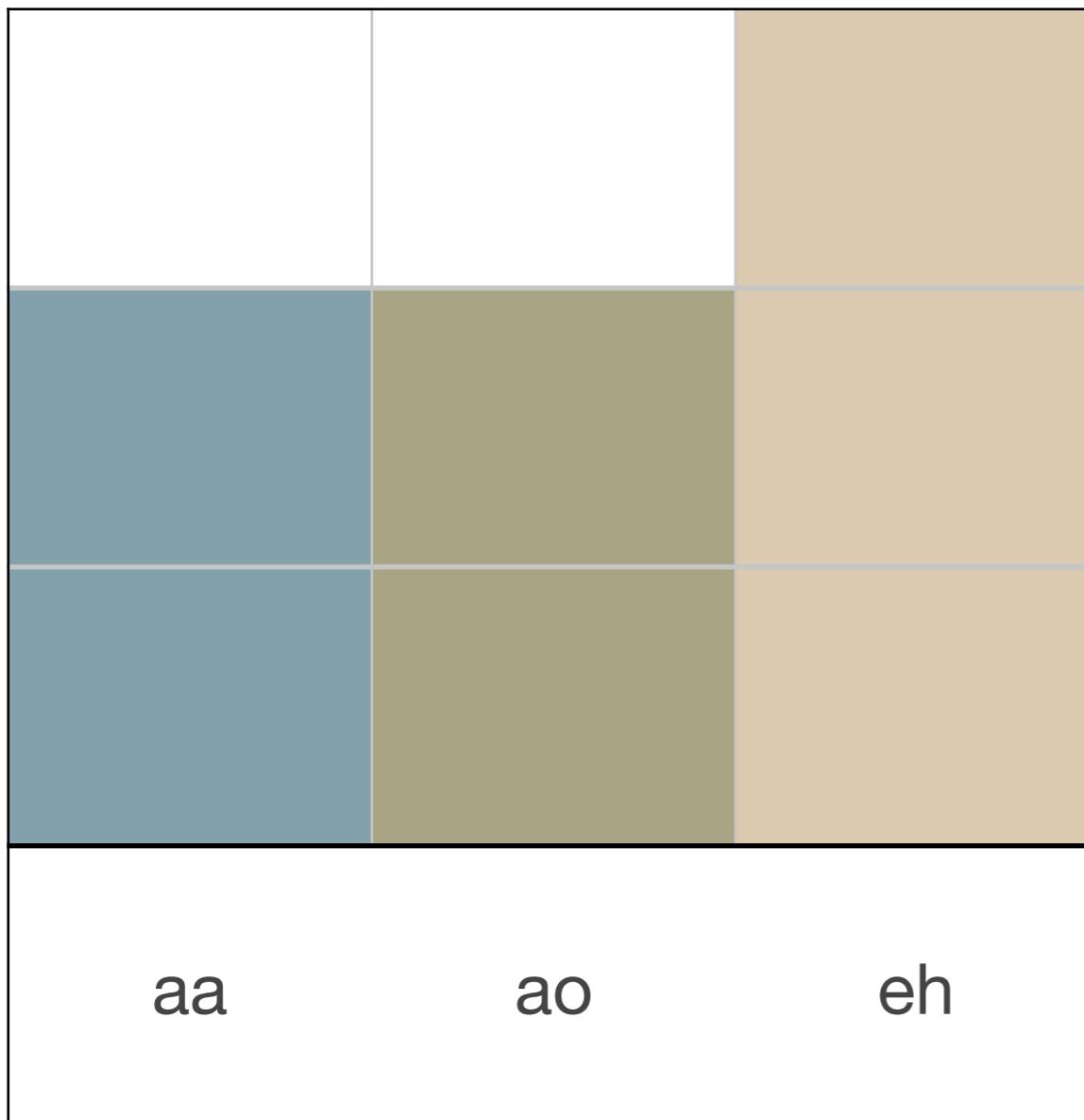
$$H = -\sum_{x} p(x) log(p(x))$$

p(aa) = 2 / 7 ≈ 0.286
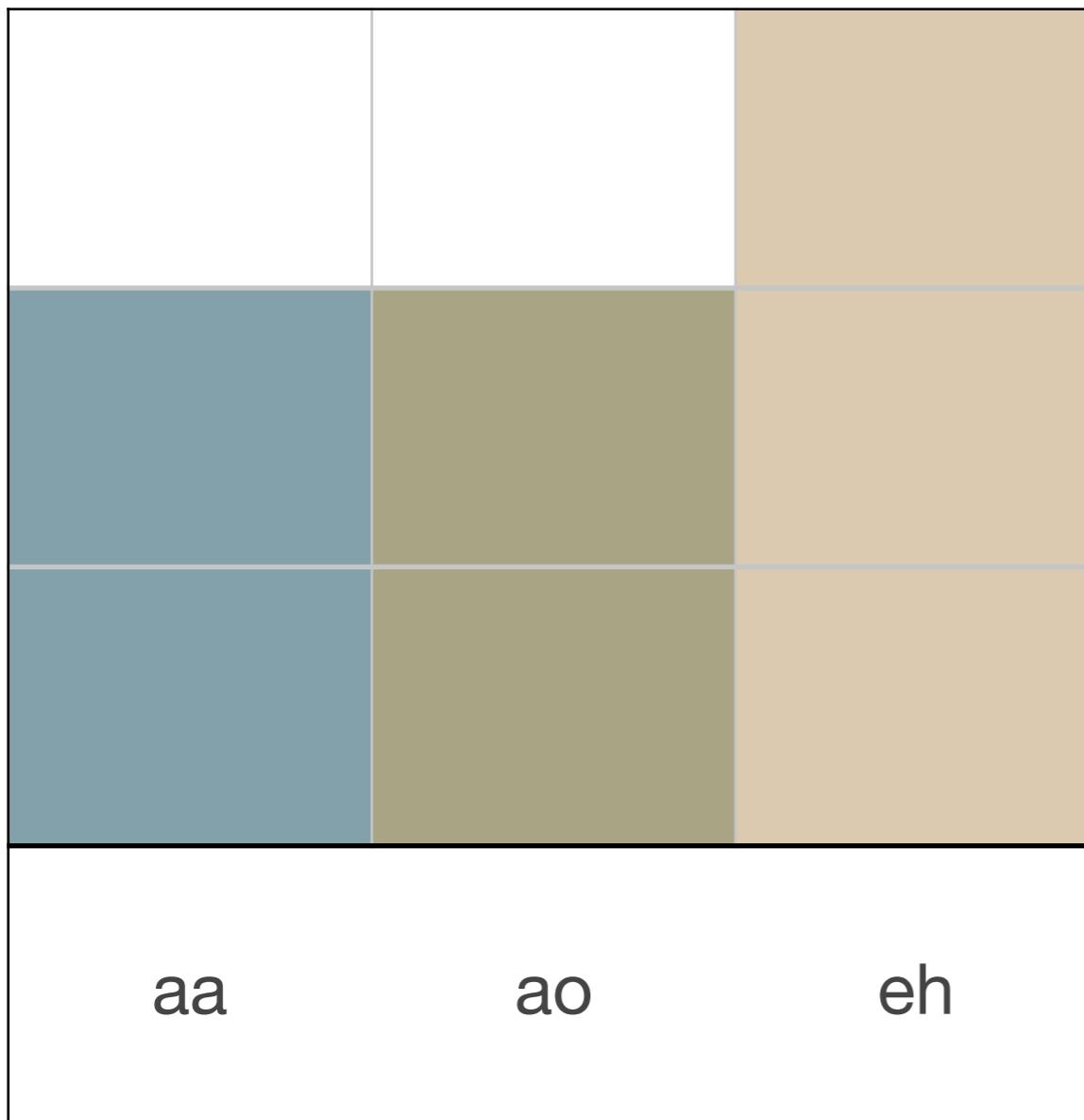
p(ao) = 2 / 7 ≈ 0.286

p(eh) = 3 / 7 ≈ 0.428

$log_2(2 / 7) = log_2(2) - log_2(7)$

≈ 1.00 - 2.81 ≈ - 1.81

$log_2(3 / 7) = log_2(3) - log_2(7)$

≈ 1.58 - 2.81 ≈ - 1.23

Entropy of distribution at the root

$$H = -\sum_x p(x)log(p(x))$$

H ≈ - (0.286 x -1.81

      + 0.286 x -1.81

      + 0.428 x -1.23)

≈ - (-0.518 - 0.518 - 0.526)

H ≈ 1.56 bits

Example split: *Does letter n = "r" ?*

| predictors | | | | | | | predictee | *Answer to* |
ppp	pp	p		n	nn	nnn		*question*
o	g	w	**a**	s	h	-	aa	N
a	r	w	**a**	s	h	-	aa	N
-	-	w	**a**	r	r	a	ao	Y
-	-	w	**a**	r	r	a	ao	Y
r	d	w	**a**	r	e	-	eh	Y
f	t	w	**a**	r	e	-	eh	Y
-	-	w	**a**	r	e	s	eh	Y

Example split: *Does letter n = "r" ?*

predictors							predictee	*Answer to question*
ppp	pp	p		n	nn	nnn		
o	g	w	**a**	s	h	-	aa	N
a	r	w	**a**	s	h	-	aa	N

predictors							predictee	*Answer to question*
ppp	pp	p		n	nn	nnn		
-	-	w	**a**	r	r	a	ao	Y
-	-	w	**a**	r	r	a	ao	Y
r	d	w	**a**	r	e	-	eh	Y
f	t	w	**a**	r	e	-	eh	Y
-	-	w	**a**	r	e	s	eh	Y

predictors							predictee	Answer to question
ppp	pp	p		n	nn	nnn		
o	g	w	**a**	s	h	-	aa	N
a	r	w	**a**	s	h	-	aa	N

$$H = -\sum_{x} p(x)log(p(x))$$

H = - (1 x 0

    + 0

    + 0)

H = 0 bits

aa   ao   eh

predictors							predictee	Answer to question
ppp	pp	p		n	nn	nnn		
-	-	w	**a**	r	r	a	ao	Y
-	-	w	**a**	r	r	a	ao	Y
r	d	w	**a**	r	e	-	eh	Y
f	t	w	**a**	r	e	-	eh	Y
-	-	w	**a**	r	e	s	eh	Y

p(aa) = 0 / 5 = 0

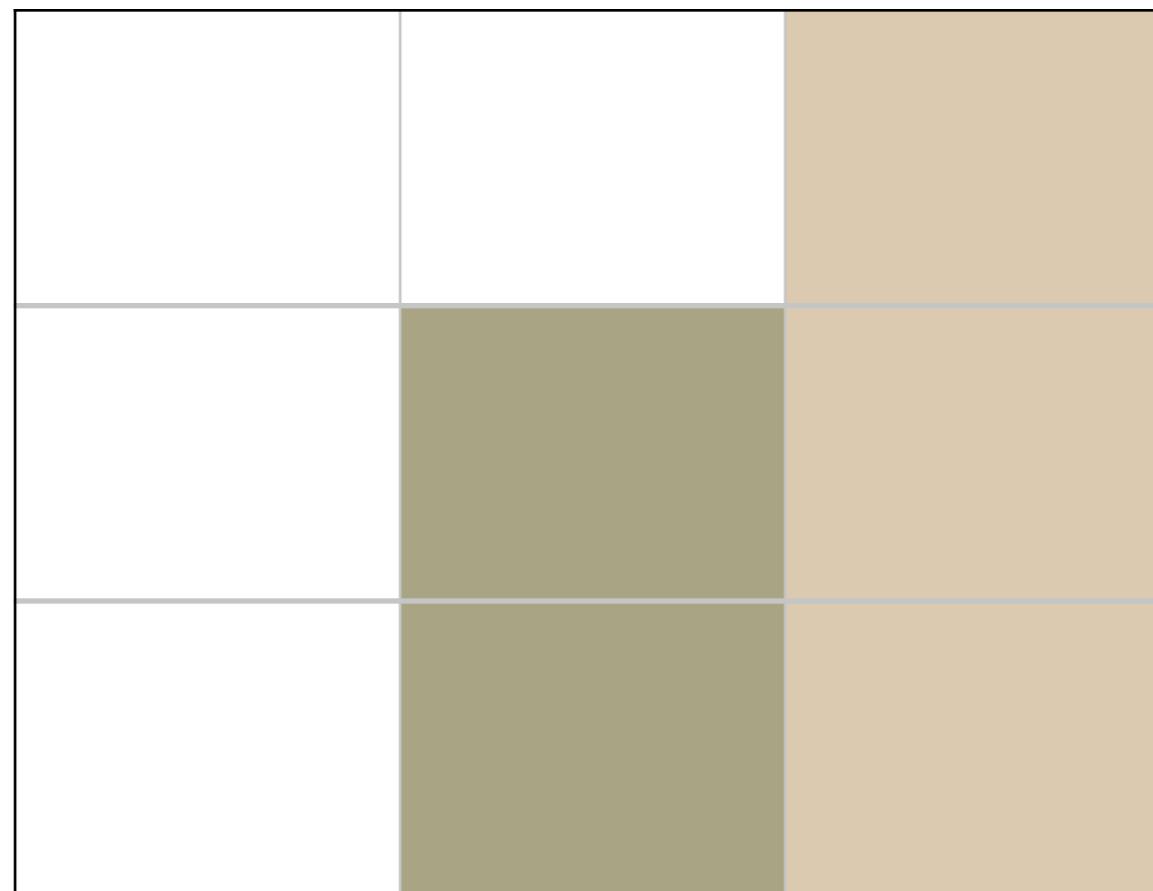p(ao) = 2 / 5 = 0.4

p(eh) = 3 / 5 ≈ 0.6

$\log_2(2 / 5) = \log_2(2) - \log_2(5)$

$\approx 1.00 - 2.32 \approx -1.32$

$\log_2(3 / 7) = \log_2(3) - \log_2(5)$

$\approx 1.58 - 2.32 \approx -0.74$

aa   ao   eh

predictors							predictee	Answer to question
ppp	pp	p		n	nn	nnn		
-	-	w	**a**	r	r	a	ao	Y
-	-	w	**a**	r	r	a	ao	Y
r	d	w	**a**	r	e	-	eh	Y
f	t	w	**a**	r	e	-	eh	Y
-	-	w	**a**	r	e	s	eh	Y

$$H = -\sum_{x} p(x) log(p(x))$$

H ≈ - (0

+ 0.4 x -1.32

+ 0.6 x -0.74)

≈ - (0 - 0.528 - 0.444)

H ≈ 0.972 bits

aa   ao   eh

Total entropy before split

predictors							predictee
ppp	pp	p		n	nn	nnn	
o	g	w	**a**	s	h	-	aa
a	r	w	**a**	s	h	-	aa
-	-	w	**a**	r	r	a	ao
-	-	w	**a**	r	r	a	ao
r	d	w	**a**	r	e	-	eh
f	t	w	**a**	r	e	-	eh
-	-	w	**a**	r	e	s	eh

H ≈ 1.56 bits

total H ≈ 1.56 bits

Entropy after split *Does letter n = "r" ?*

predictors							predictee	*Answer to question*
ppp	pp	p		n	nn	nnn		
o	g	w	**a**	s	h	-	aa	N
a	r	w	**a**	s	h	-	aa	N

H = 0 bits

predictors							predictee	*Answer to question*
ppp	pp	p		n	nn	nnn		
-	-	w	**a**	r	r	a	ao	Y
-	-	w	**a**	r	r	a	ao	Y
r	d	w	**a**	r	e	-	eh	Y
f	t	w	**a**	r	e	-	eh	Y
-	-	w	**a**	r	e	s	eh	Y

H ≈ 0.972 bits

Information gain

- We now compute how much the entropy has been decreased by this split

- This is the **weighted** sum of the entropies in each partition

  - the weights are based on the number of data points in each partition

  - N branch: 2/7 x 0 bits

  - Y branch: 5/7 x 0.972 bits = 0.694 bits

  - total = 0 + 0.694 = 0.694 bits

in the lecture we simplified this and used the unweighted sum

- Before split: 1.56 bits

- After split: 0.694 bits

- Information gain = 0.866 bits

Choosing the best question

- Just try them all, and pick the one with the highest information gain

- Try working through some of the other questions yourself, then make a table like this:

question	information gain (bits)
n = "a" ?	
n = "b" ?	
....	
n = "r" ?	0.866
n = "s" ?	
...	
p = "a" ?	
p = "b" ?	
...	