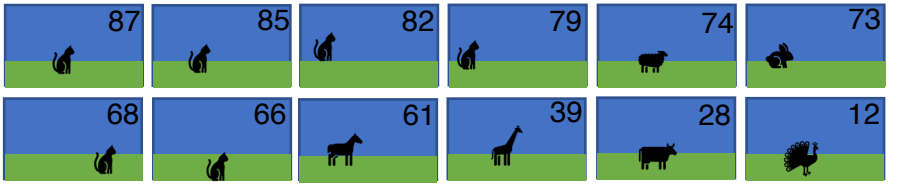
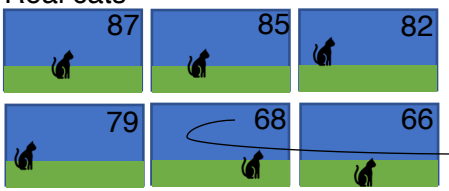
(a) Scored data set

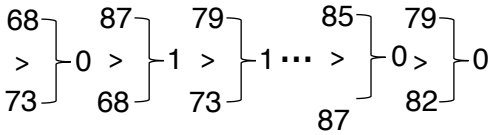


(b) How $B(r)$ is calculated

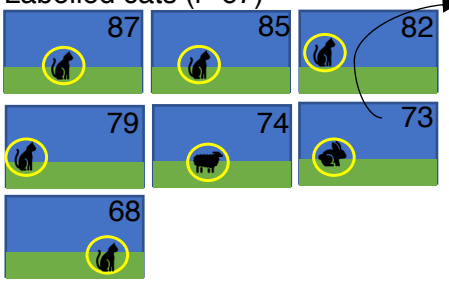
Real cats



Choose (with replacement) n real cats and n labelled cats at random (discard if same image chosen for comparison) and compare scores



Labelled cats (r=67)



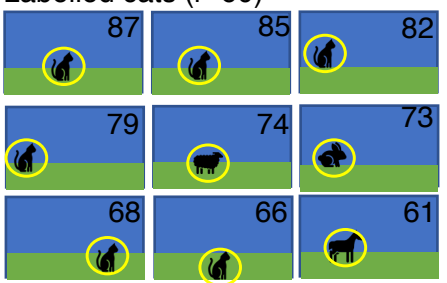
$B(r)$ is proportion of times real score score is more than labelled, i.e.

$$B(r) = (0 + 1 + 1 + \cdots + 0 + 0)/n$$

In this example: $\mathrm{B}(65) \approx 0.49$ and $\mathrm{C}(65) = 5/7 \approx 0.71$

(c) Effect of changing threshold (r)

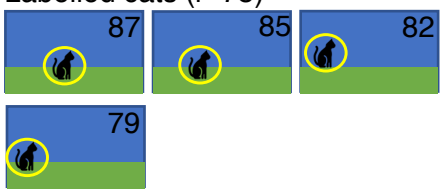
Labelled cats (r=60)



Decreasing r, increases the number of labelled cats and gives a larger value of $\mathrm{B}(r)$. Real cats are unaffected by the choice of r.

In this example: $\mathrm{B}(60) \approx 0.60$ and $\mathrm{C}(60) = 6/9 \approx 0.66$

Labelled cats (r=75)



Increasing r, decreases the number of labelled cats and gives a smaller value of $\mathrm{B}(r)$.

In this example: $\mathrm{B}(75) \approx 0.49$ and $\mathrm{C}(75) = 4/4 = 1$