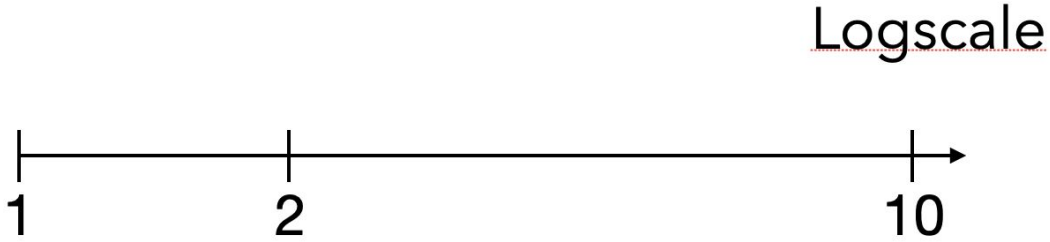
Data Visualization
W9-1

Quiz

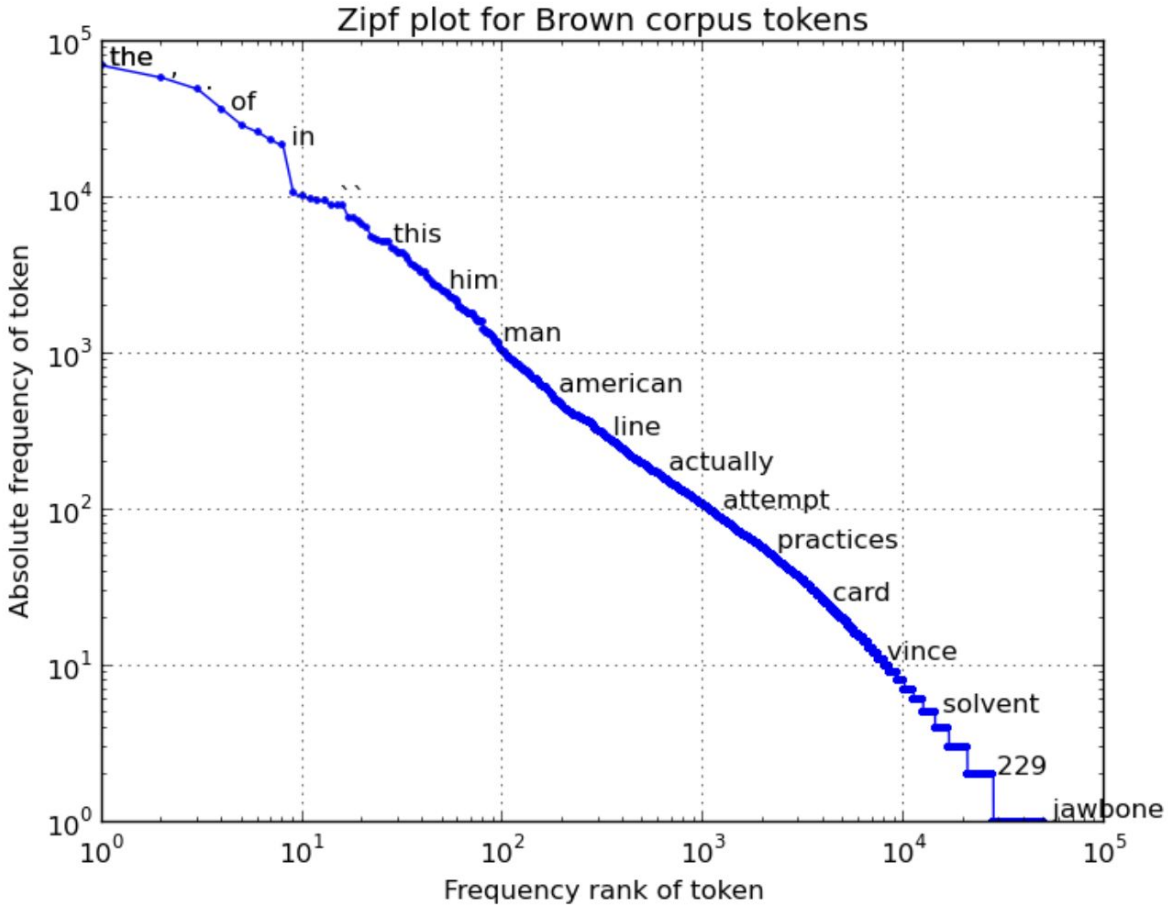
- What do you find interesting in today's VotW?
- Explain the key differences between interpolation and regression.
- Can you find 4 and 5? Then explain how.



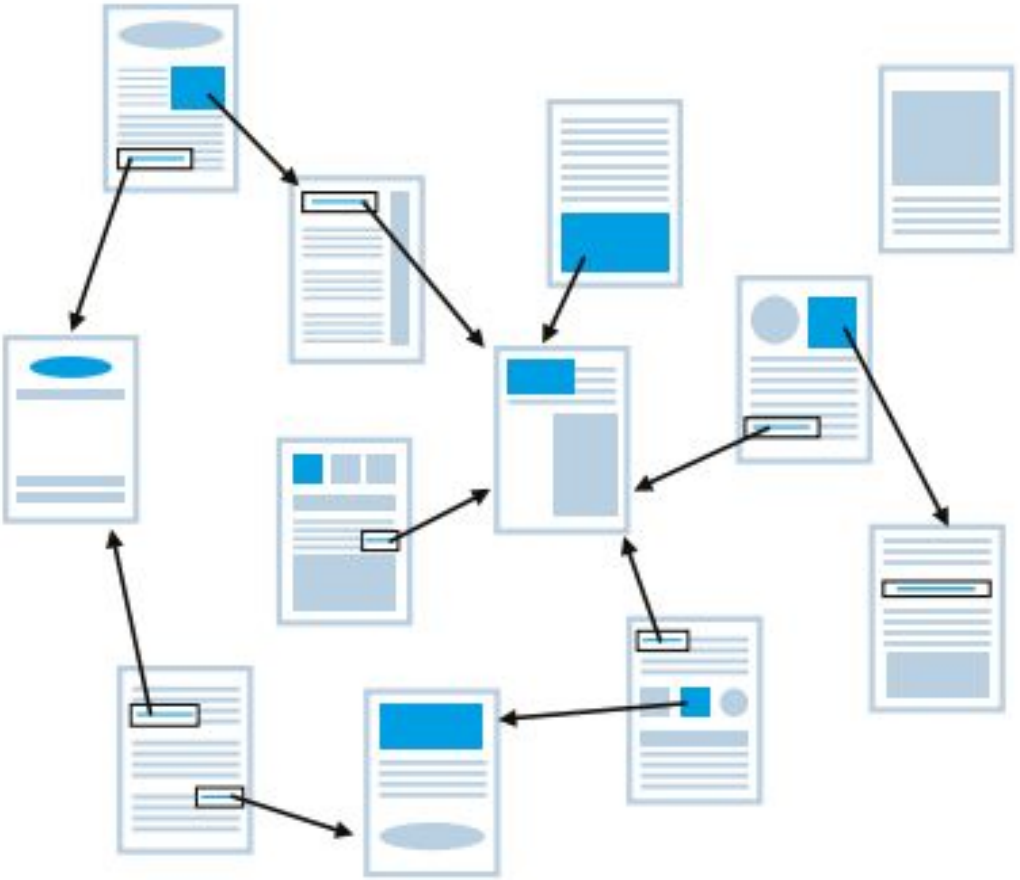
‘Power-law’ distributions

$$f(x) = cx^{-\alpha}$$

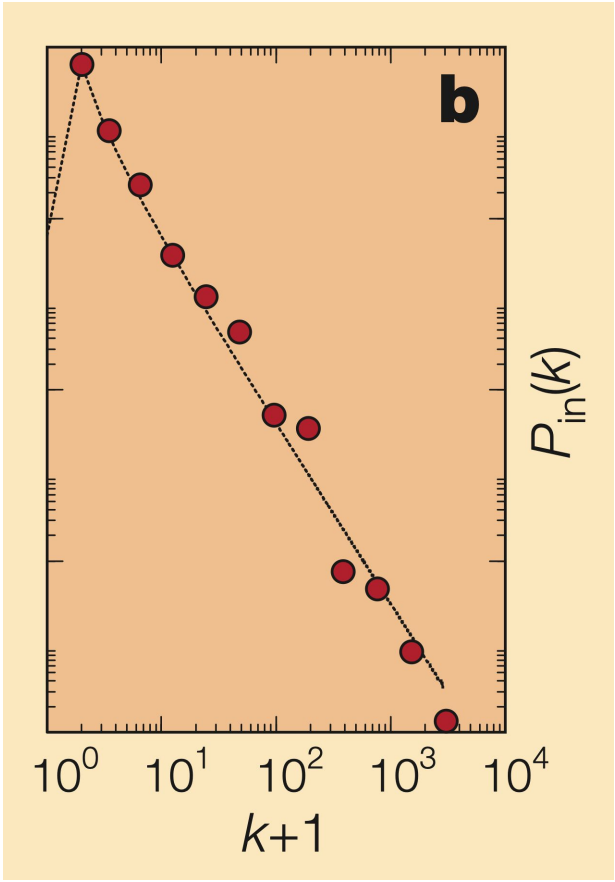
$$\begin{aligned}\log f(x) &= \log\left(cx^{-\alpha}\right) \\ &= \log c - \alpha \log x\end{aligned}$$



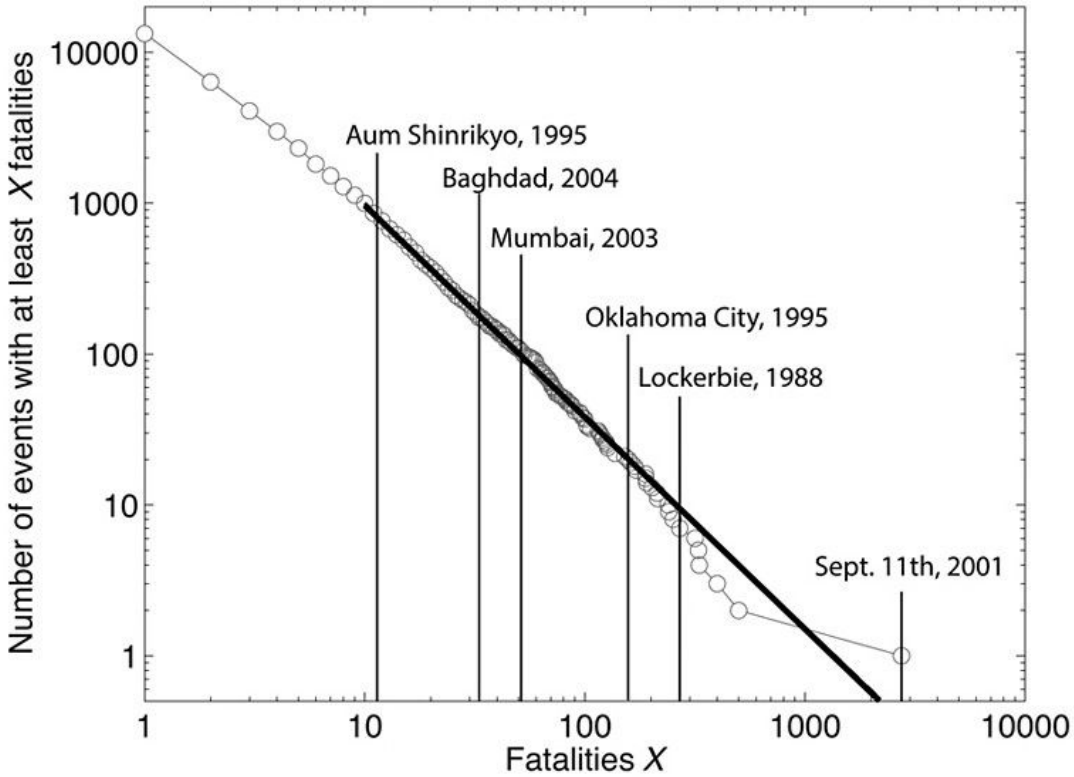
https://finnaarupnielsen.wordpress.com/2013/10/22/zipf-plot-for-word-counts-in-brown-corpus/

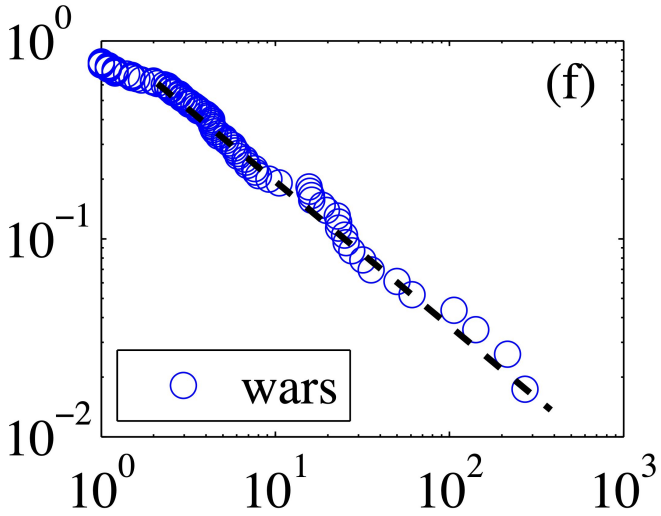


https://barefootcas.org.uk

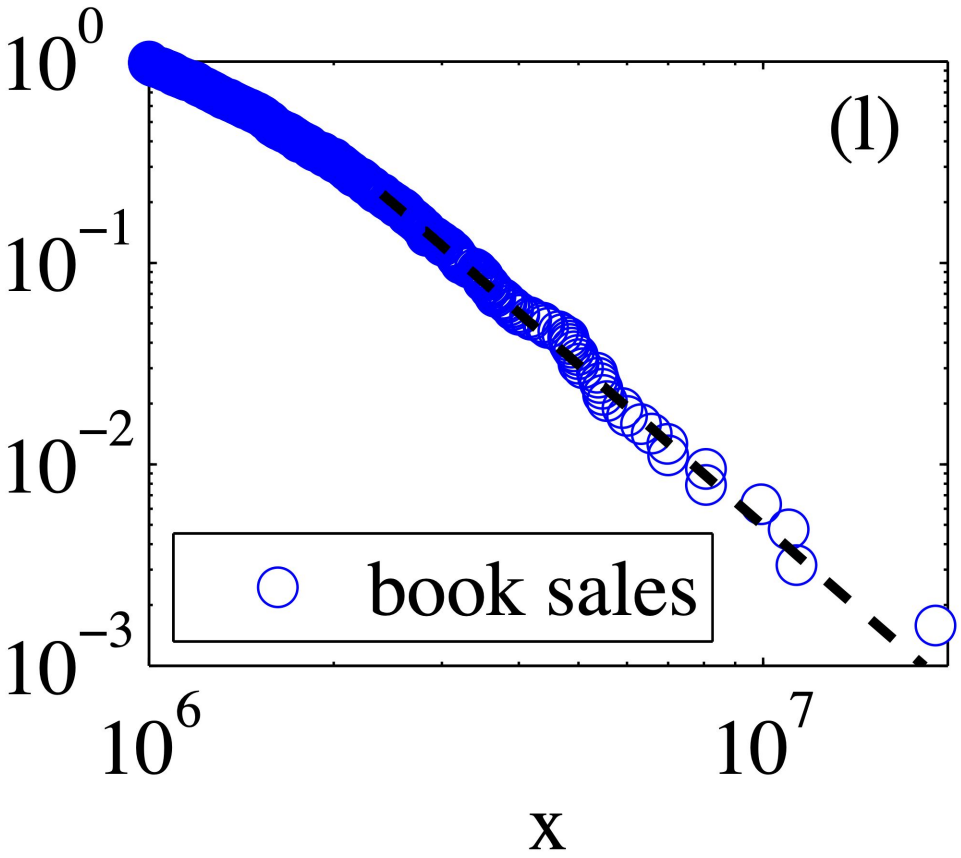


Albert, Jeong, Barabási, *Nature* **401**, 130 (1999)

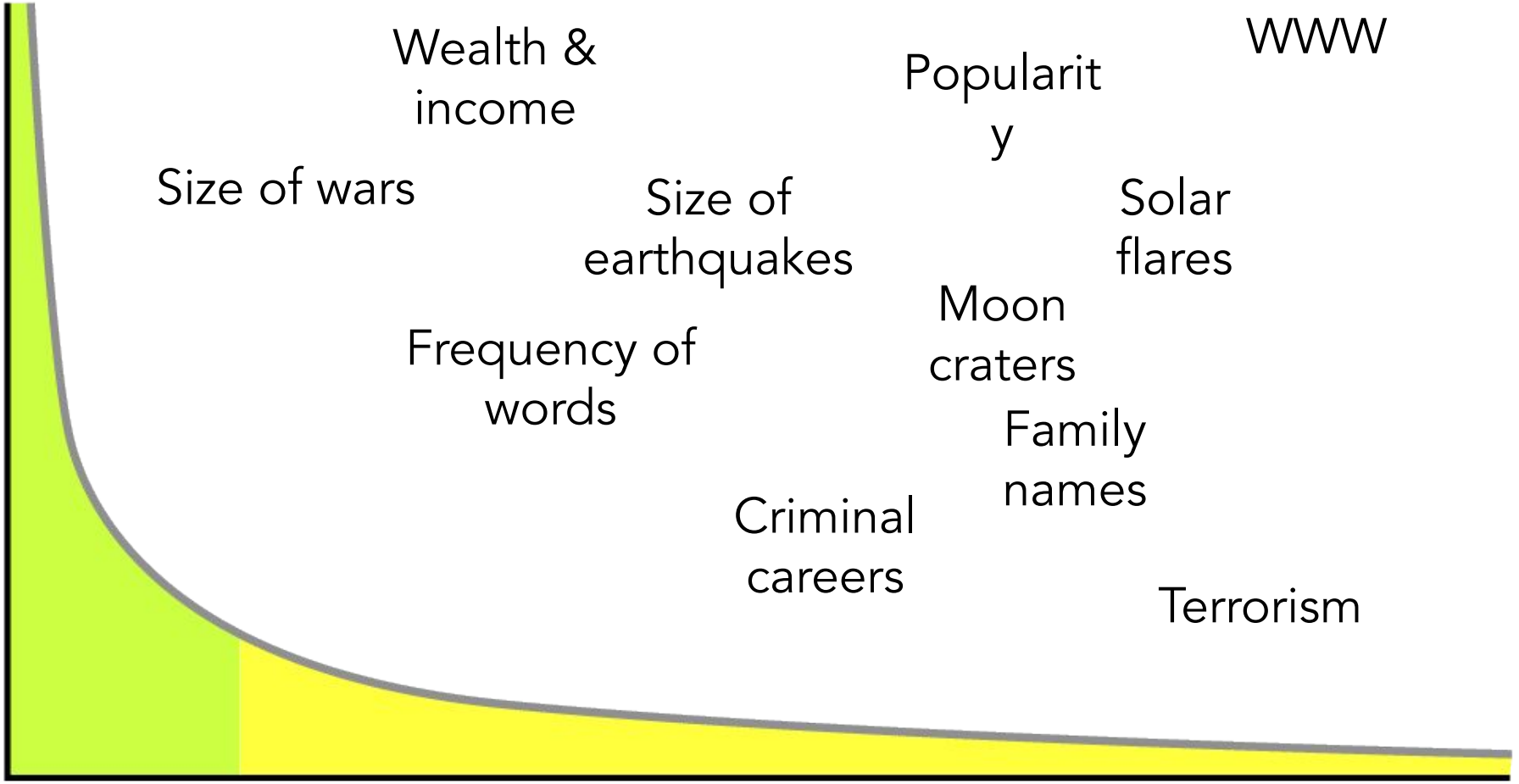




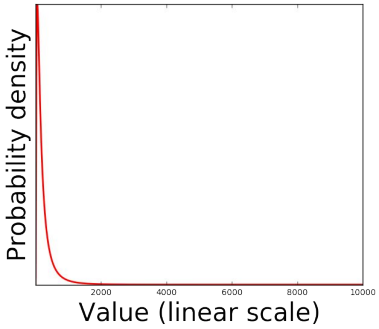
Clauset, Shalizi, Newman, *SIAM Review* **51**, 661 (2009)

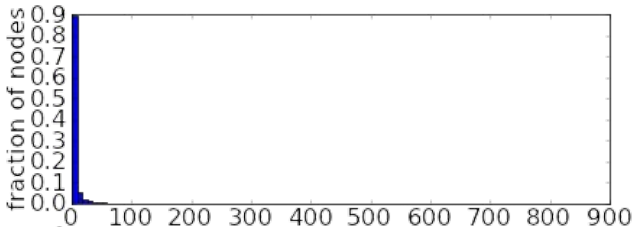


Clauset, Shalizi, Newman, *SIAM Review* **51**, 661 (2009)

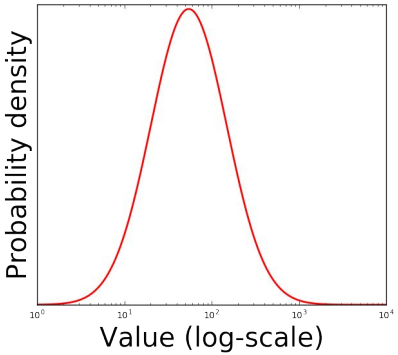

Wealth & income
WWW
Popularit y
Size of wars
Size of earthquakes
Solar flares
Moon craters
Frequency of words
Family names
Criminal careers
Terrorism

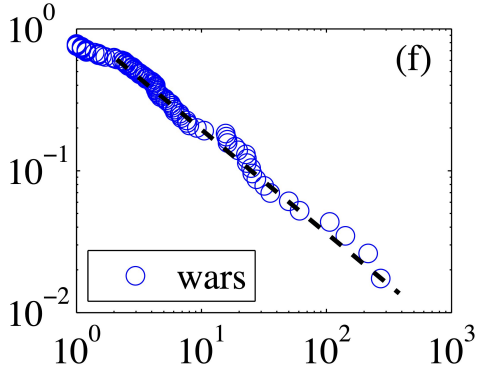
Have a fat tail?





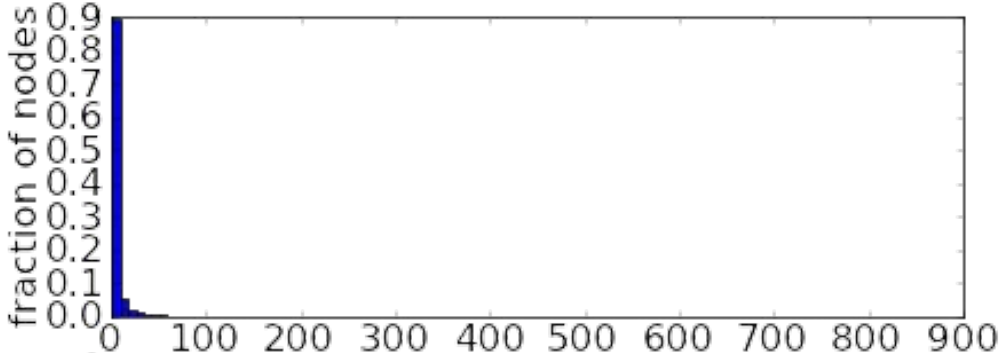
Linear scale:
Can't see any!

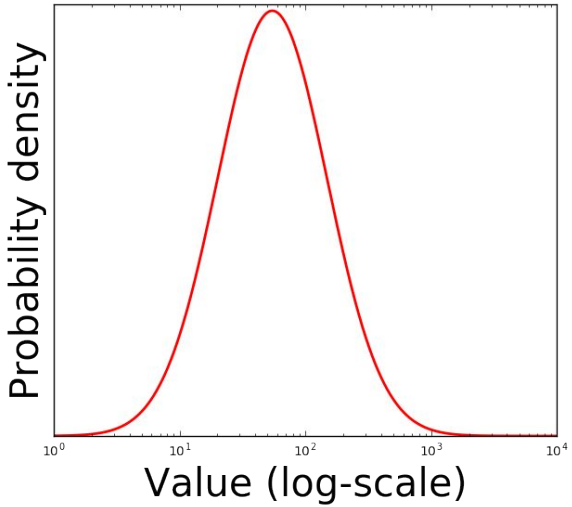




log scale:
Useful!

Can we draw a useful histogram in log-scale?





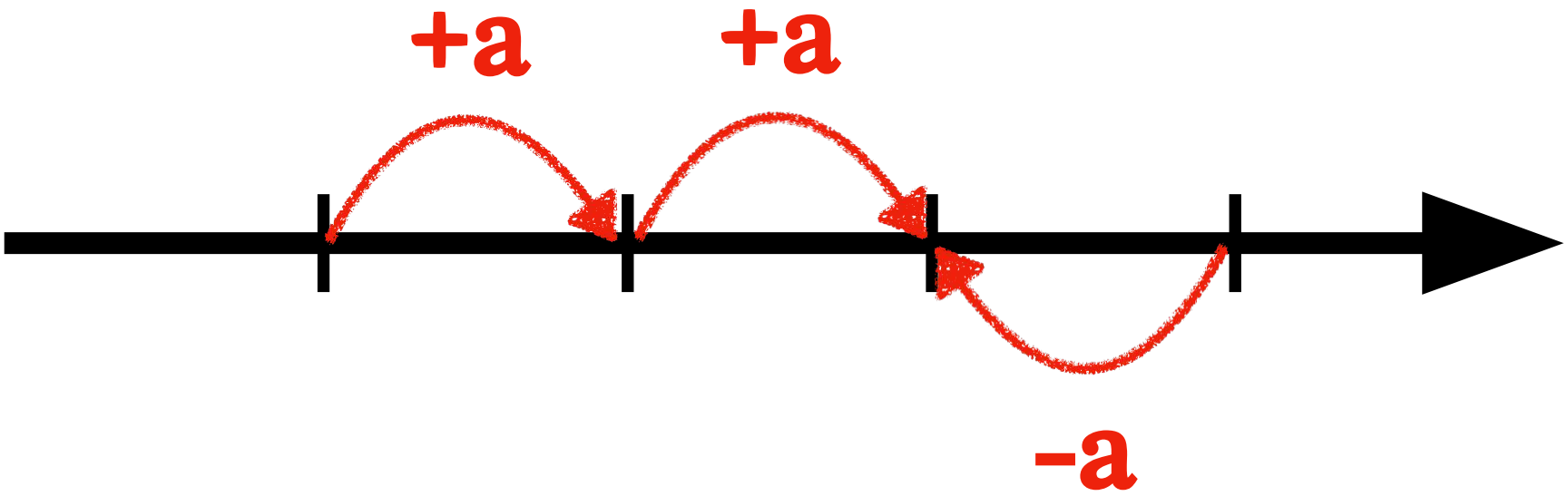
Histogram in log-scale

How should we create bins in log-scale?

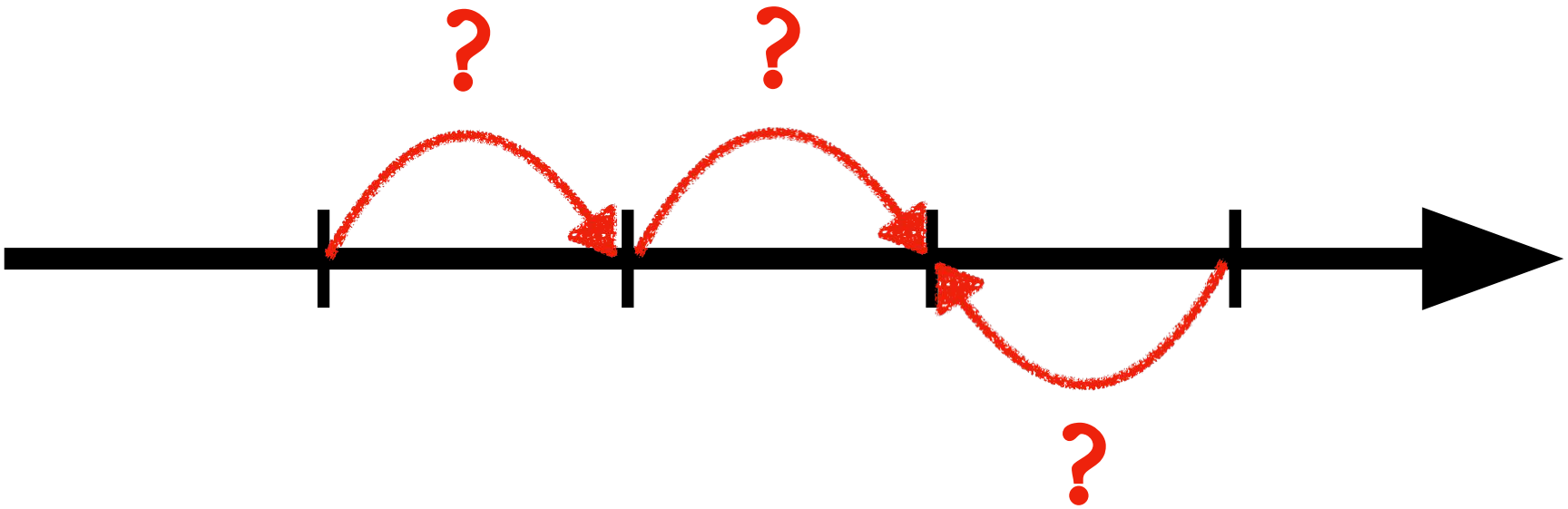
Constant width in log-scale?

First bin: [1, 3)
What are the next several bins? (same width bins in log-scale)

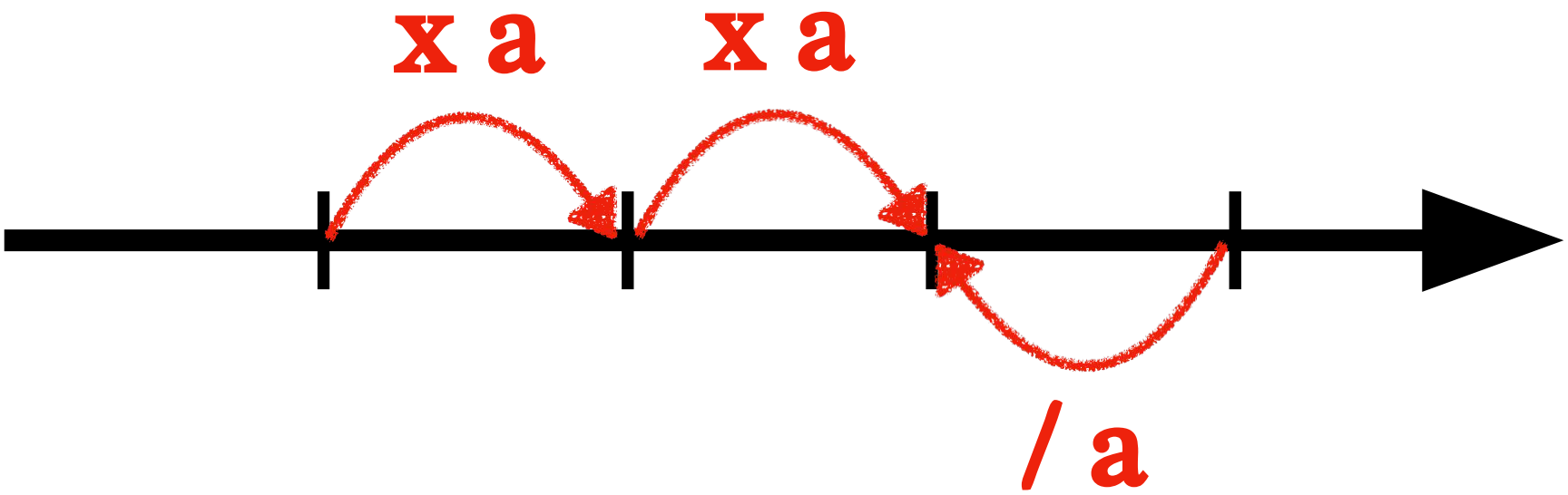
In linear scale



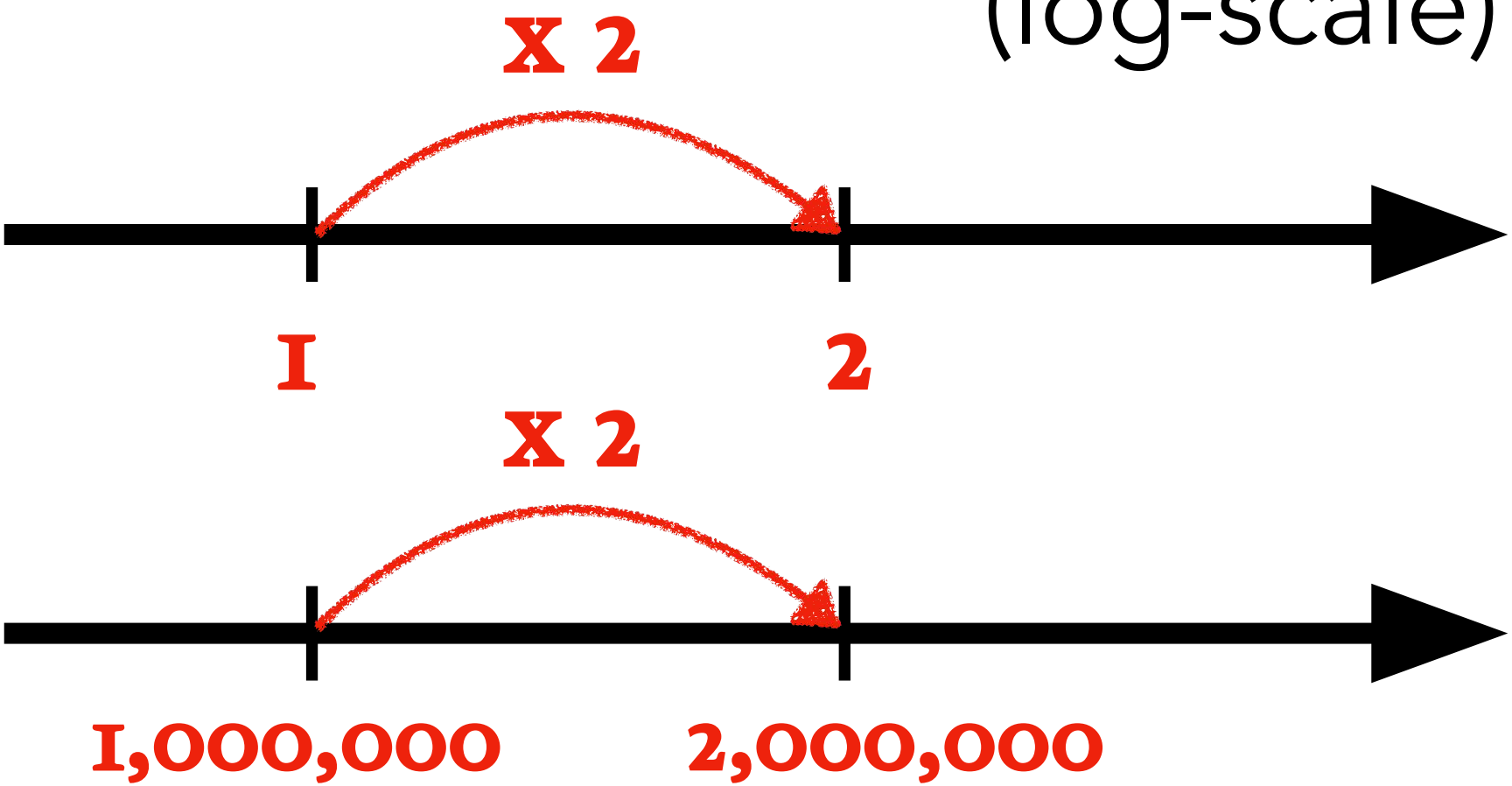
In log scale?



In log scale?

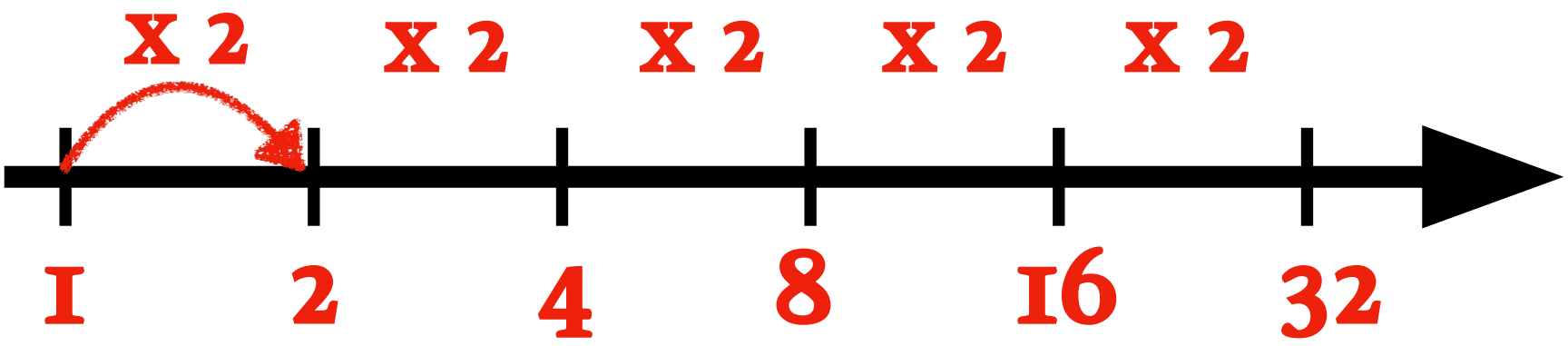


(log-scale)

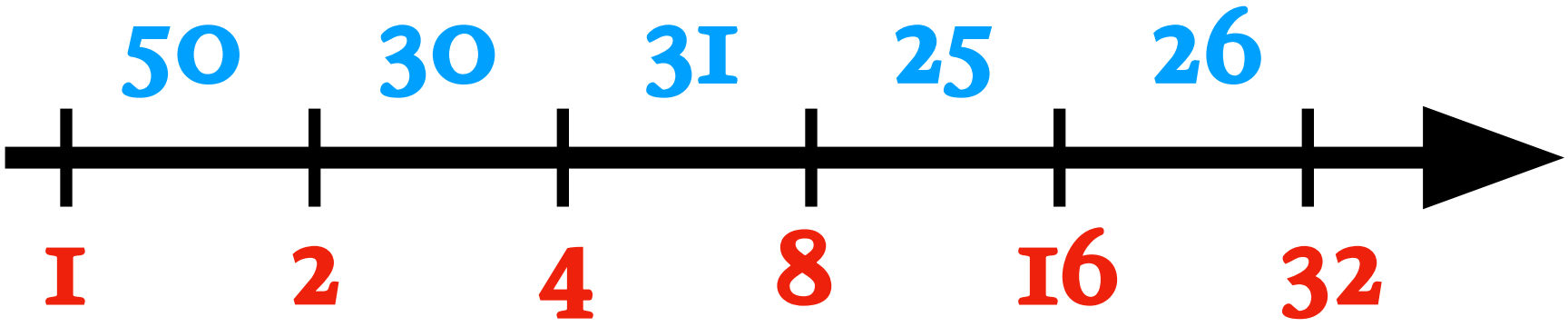


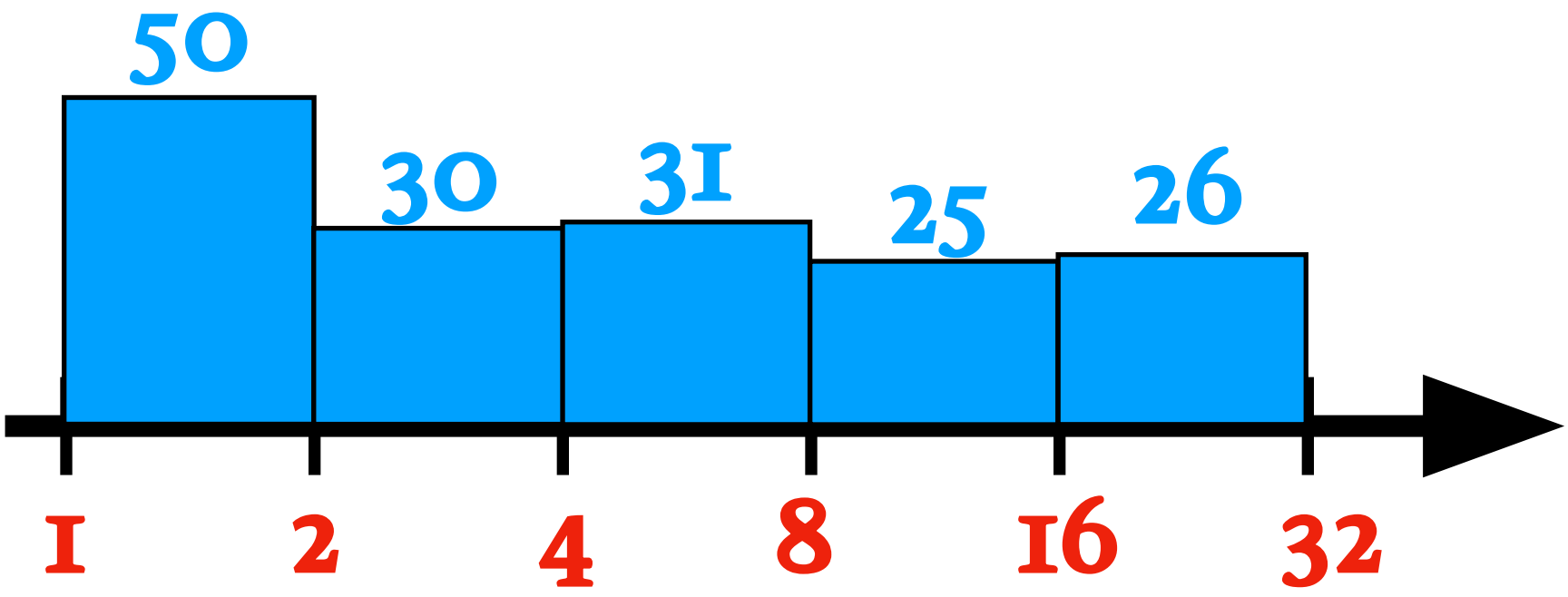
(You can choose other widths too)

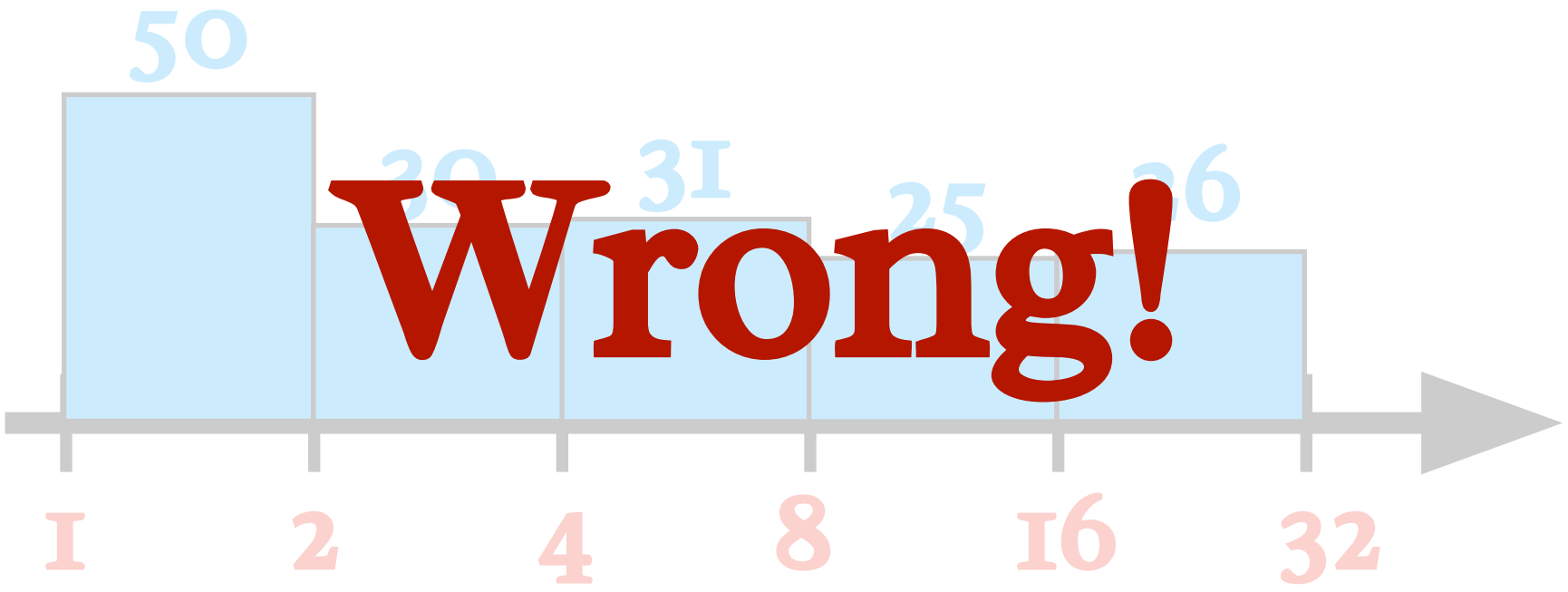
(log-scale)



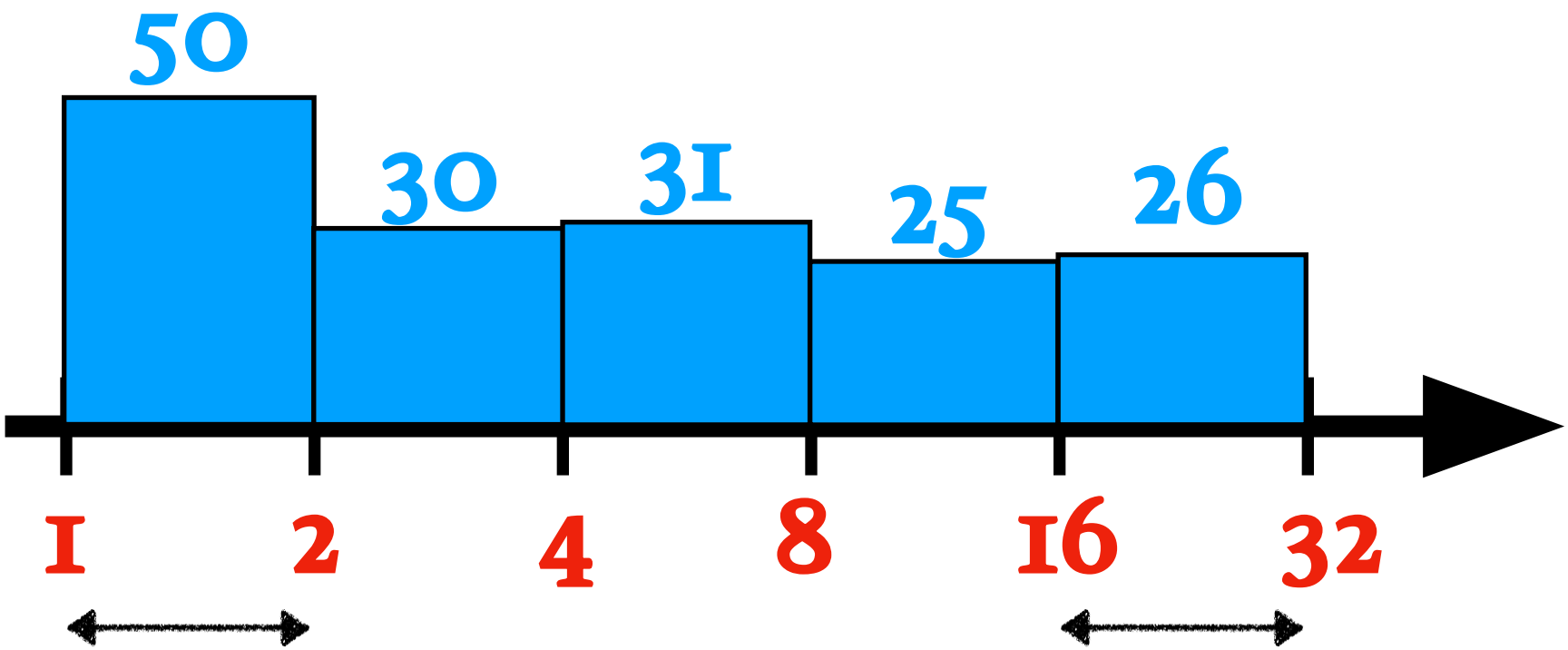
“Log-binning” (logarithmic binning)

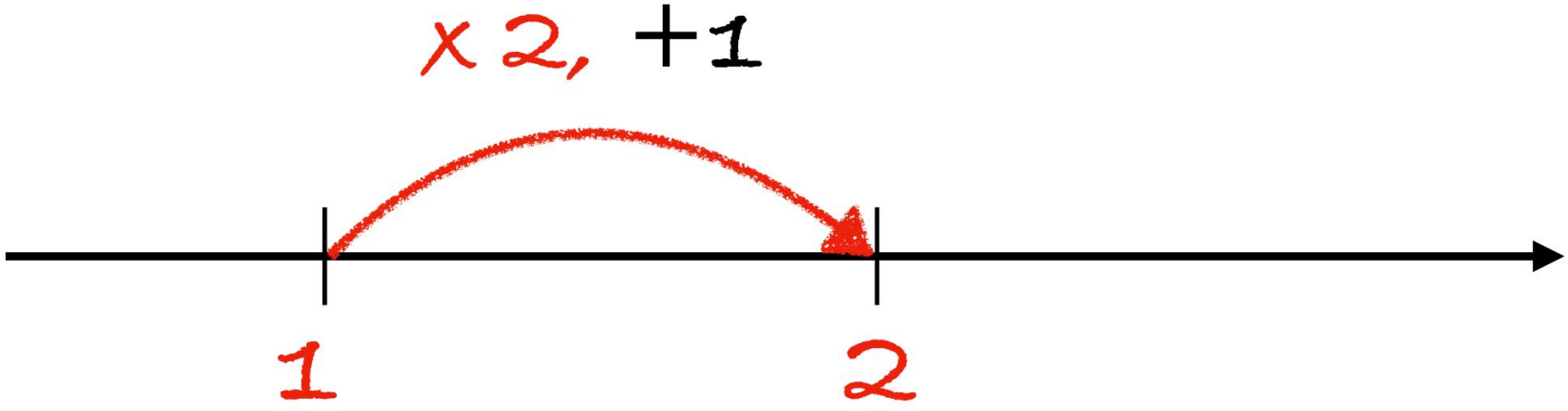

50
30
31
25
26
1
2
4
8
16
32

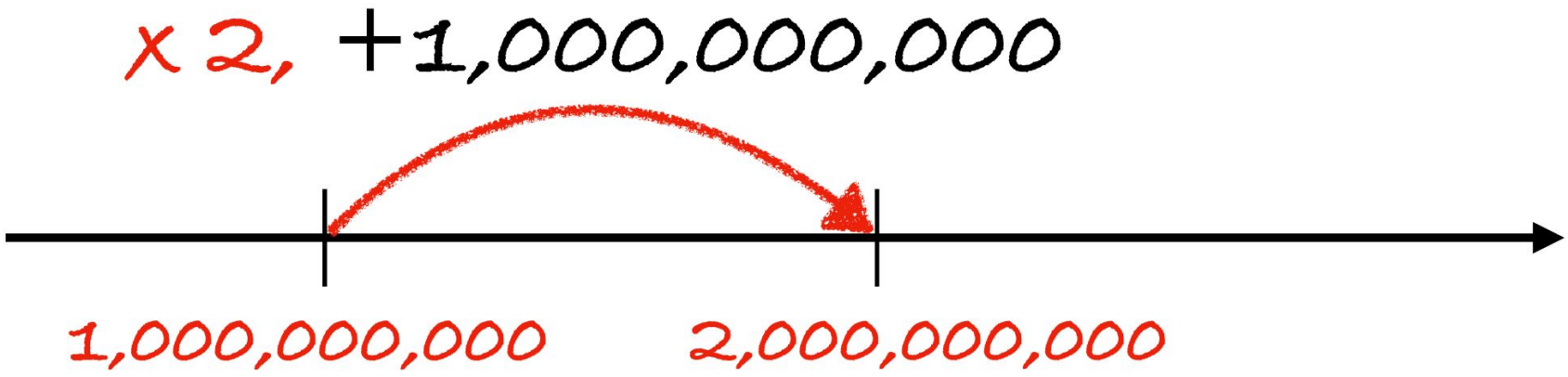

50
30
31
25
26
1
2
4
8
16
32

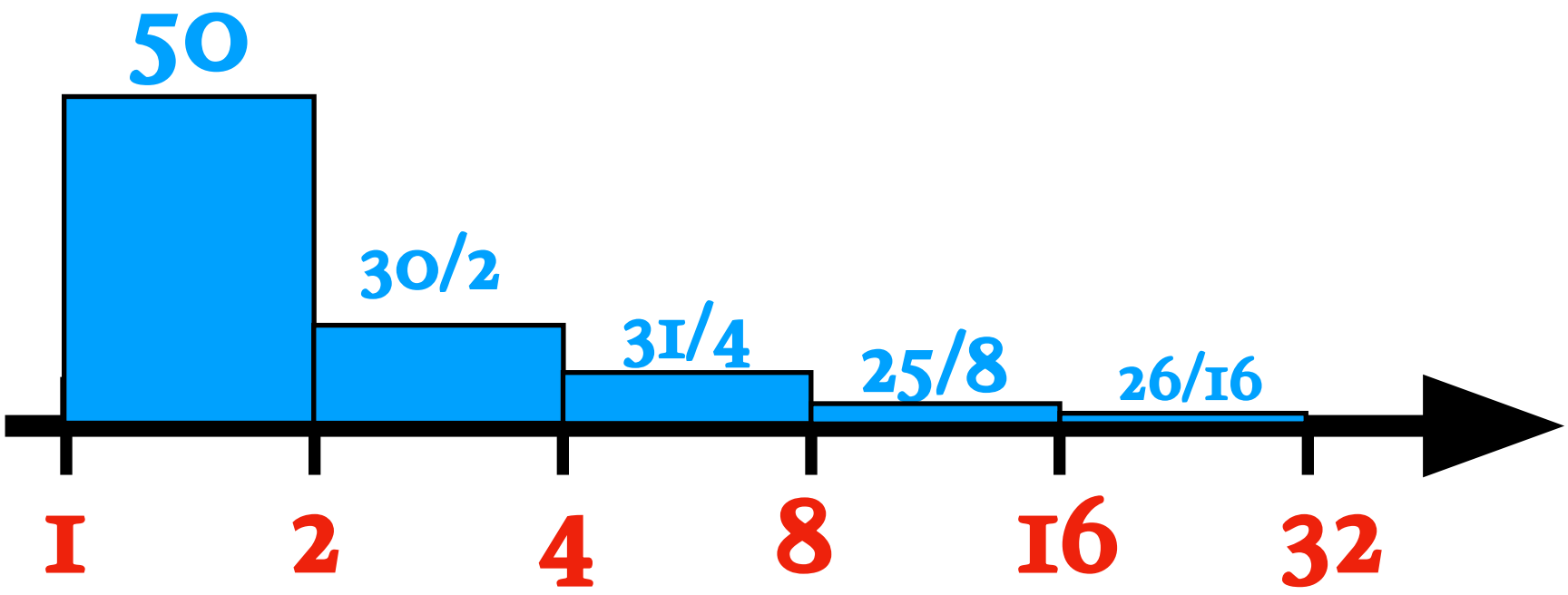

50
31
25
36
Wrong!
1
2
4
8
16
32

“Area, *not the* height,
represents the frequency!”

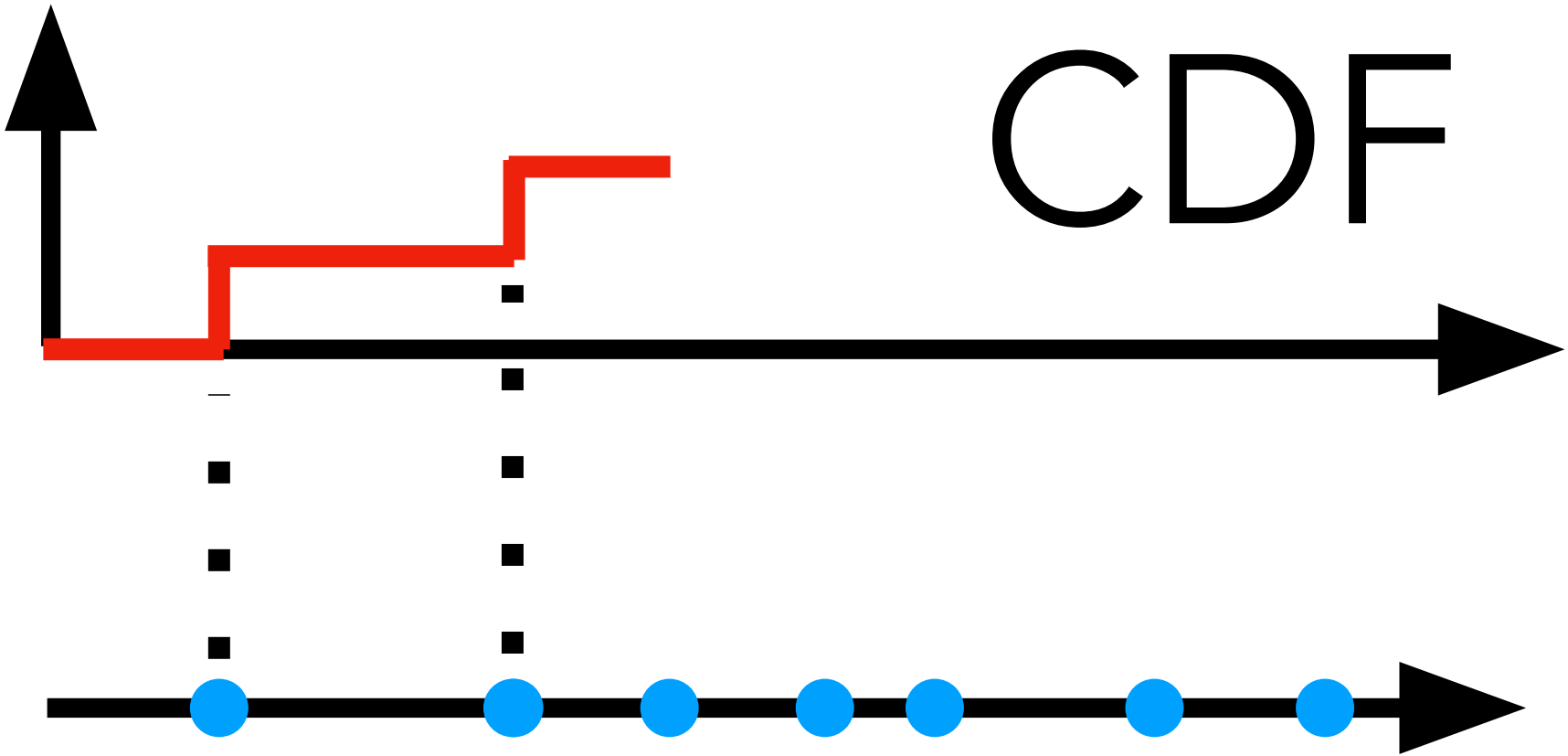

50
30
31
25
26
1
2
4
8
16
32

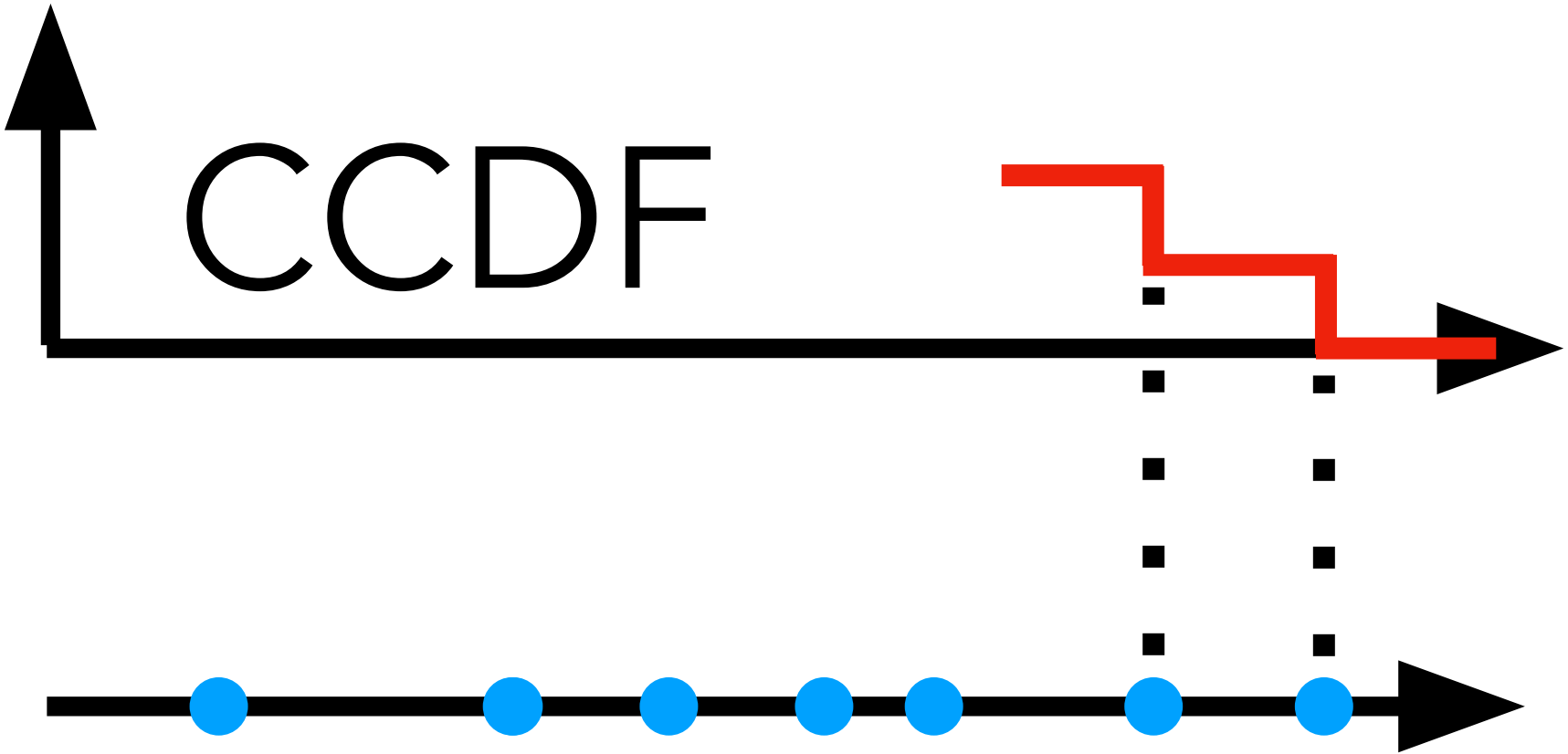

x 2, +1
1
2

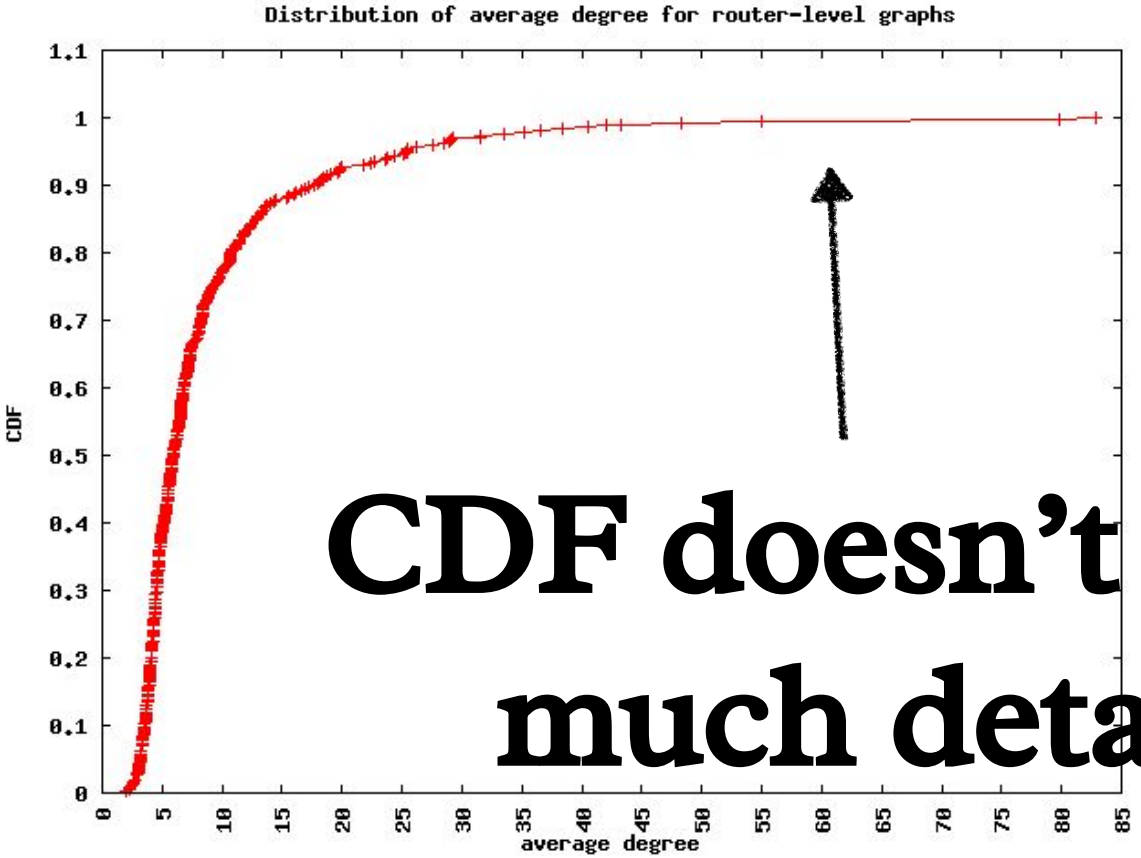

x 2, +1,000,000,000
1,000,000,000
2,000,000,000


50
30/2
31/4
25/8
26/16
1
2
4
8
16
32

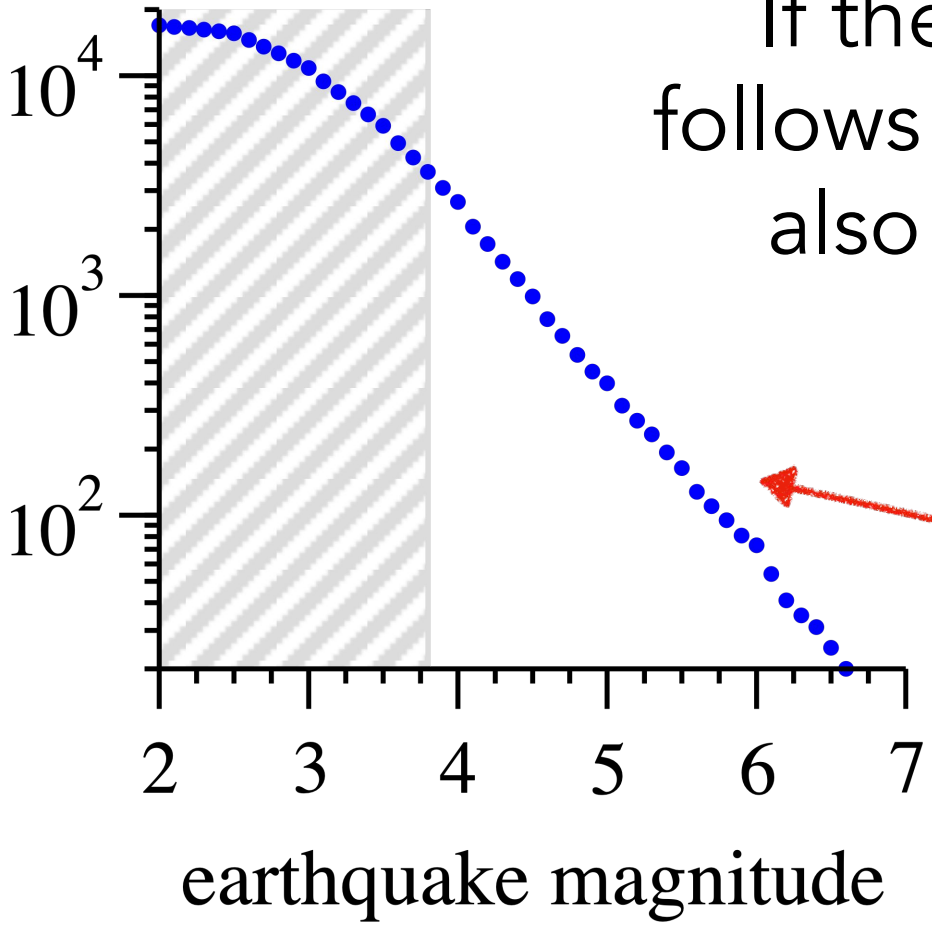
A nice alternative


CDF


CCDF



CDF doesn't show much details



If the original distribution follows a power-law, the CCDF also follows a power-law.

more useful

$$CCDF(x) = P(X > x) \quad \left(\textbf{c.f. } CDF(x) = P(X \leq x)\right)$$

If $f(x) = cx^{-\alpha}$,

$$CCDF(x) = \int_x^{\infty} ct^{-\alpha} dt$$

$$= \left[\frac{c}{1-\alpha} t^{1-\alpha} \right]_x^{\infty}$$

$$= \frac{c}{1-\alpha} x^{1-\alpha} = Cx^{-(\alpha-1)}$$

When visualizing heavy-tailed distributions such as power-law distribution,

CCDF is often much more useful than CDF.

What's wrong with these plots?

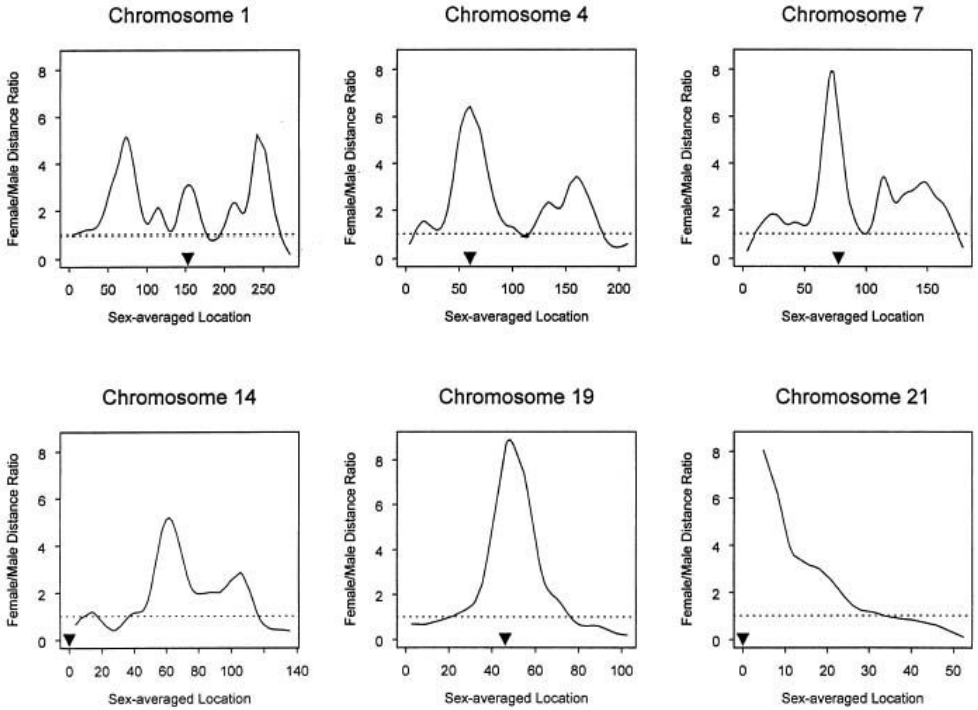
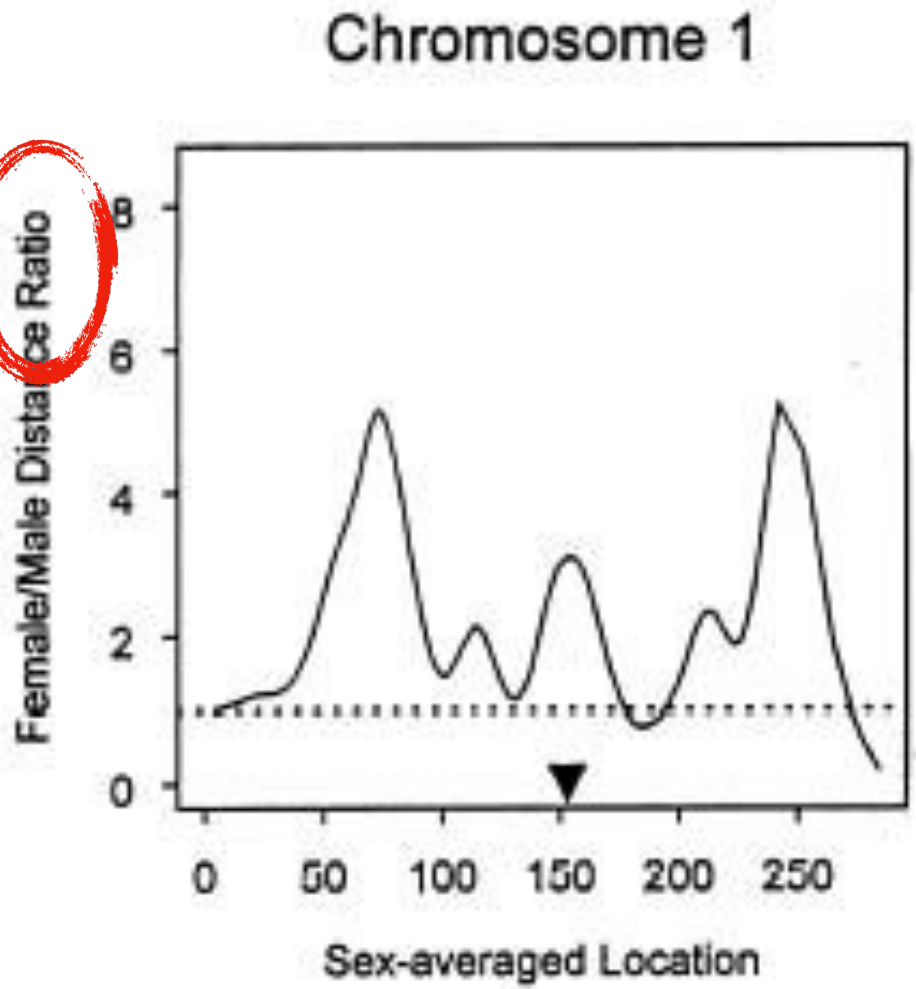
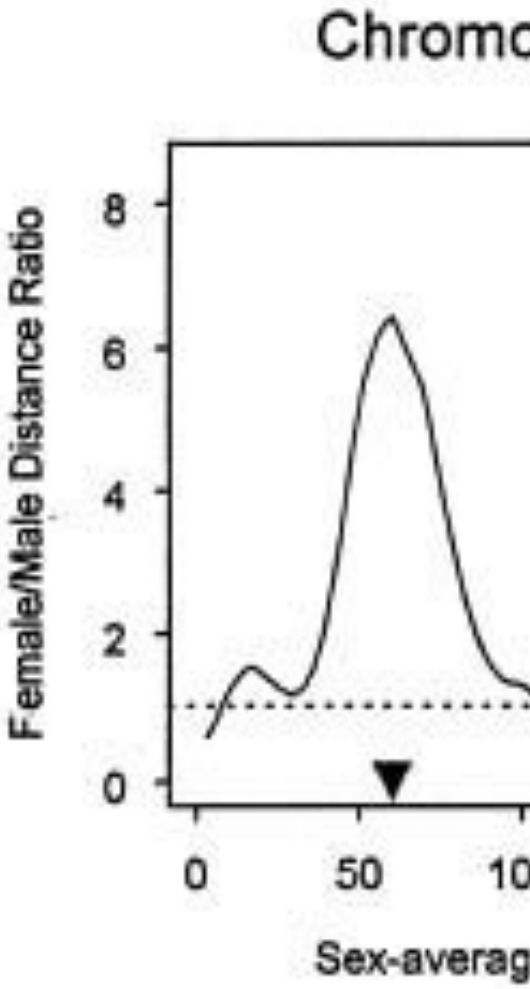


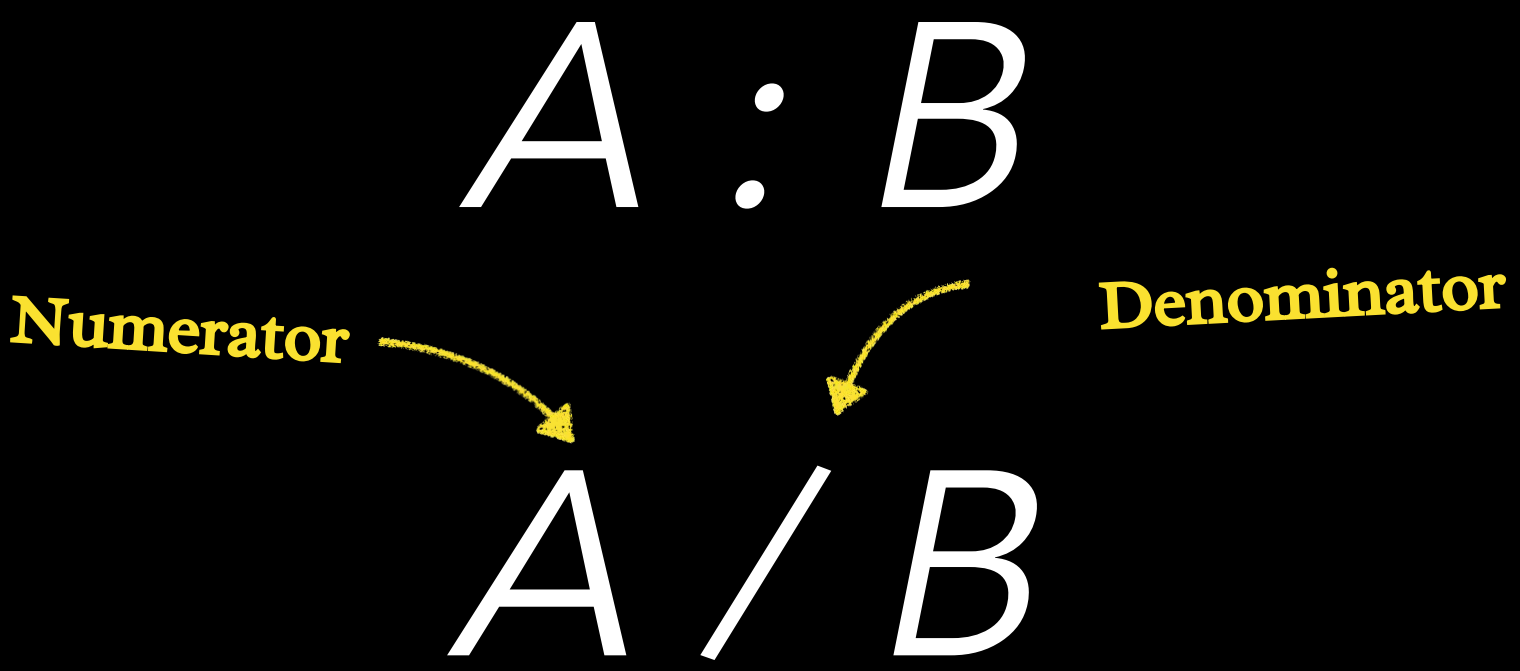
Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

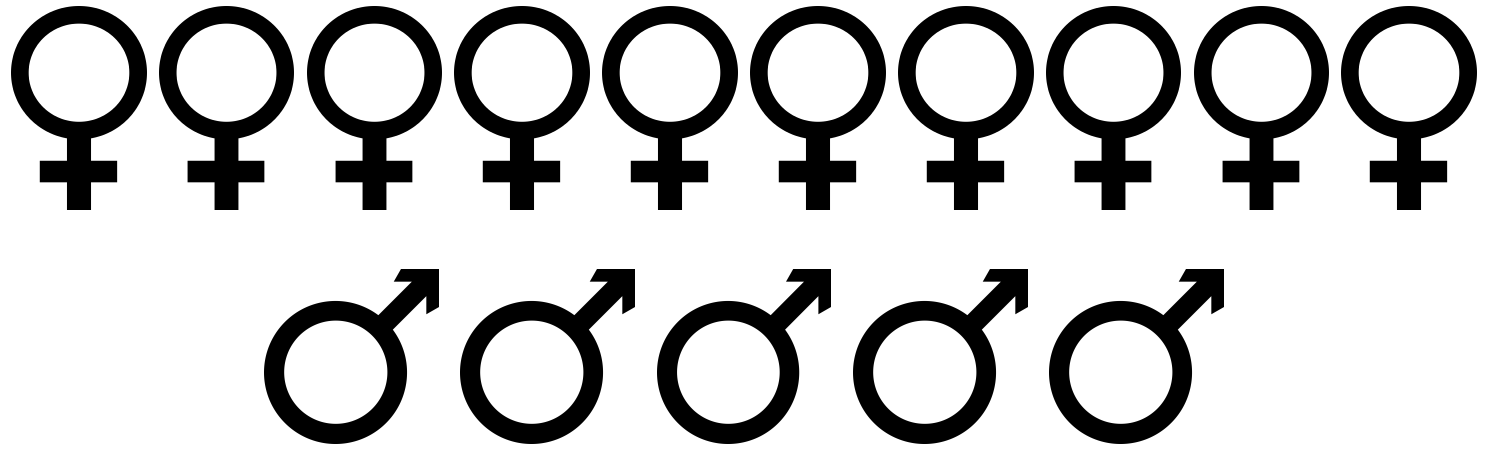
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/


Chromosome 1
Female/Male Distance Ratio
0
2
4
6
8
0
50
100
150
200
250
Sex-averaged Location


Chromo
Female/Male Distance Ratio
0
2
4
6
8
0
50
Sex-averag

What is a ratio?


A : B
Numerator
Denominator
A / B



10:5 2:1 10/5 = 2

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀

♂♂♂♂♂♂♂♂♂♂

5:10

1:2

5/10 = 0.5

♀♀♀♀♀♀♀♀♀♀

♂

10:1

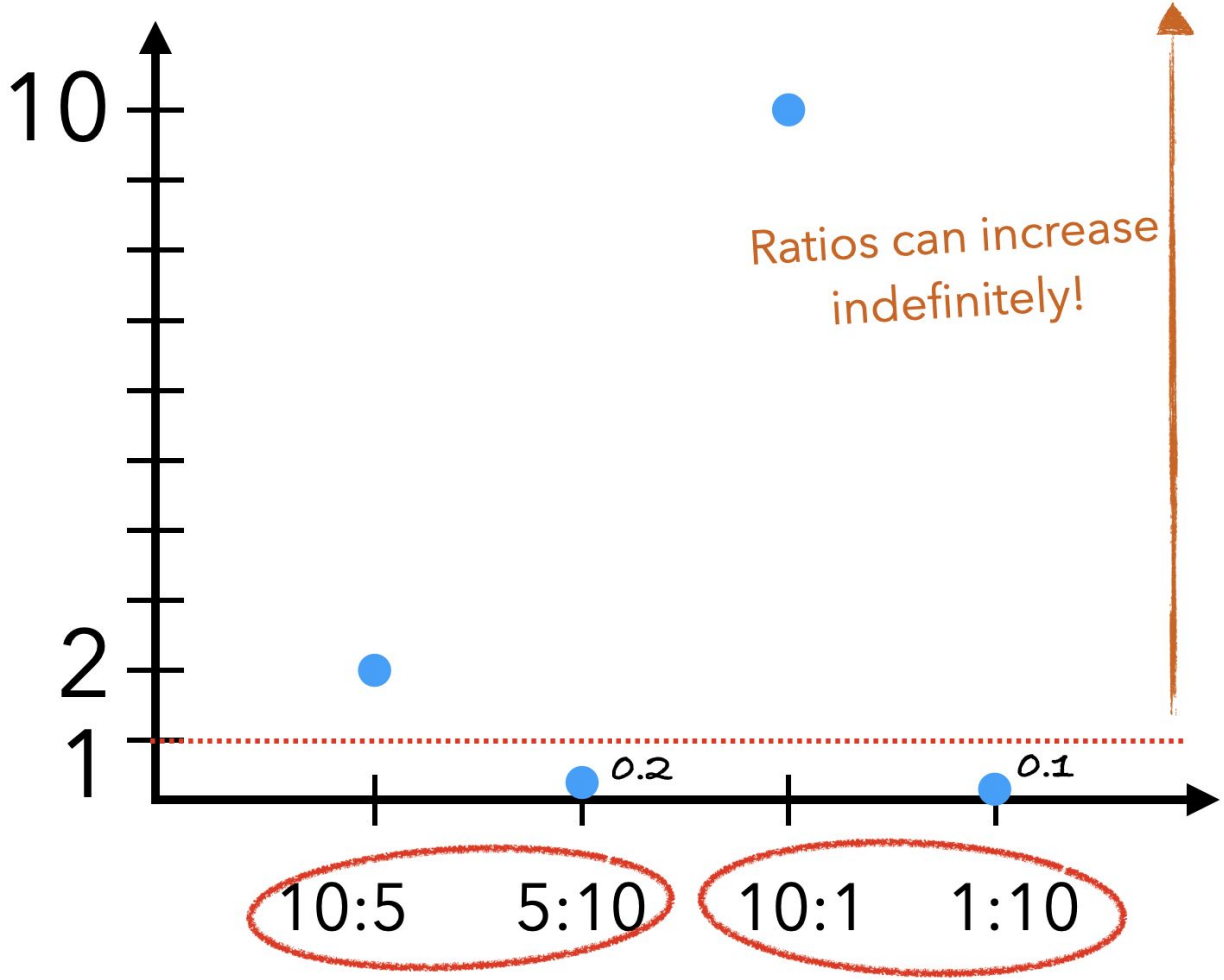
10

♀

♂♂♂♂♂♂♂♂♂♂

1:10

0.1


10
2
1
Ratios can increase indefinitely!
0.2
0.1
10:5
5:10
10:1
1:10

Log scale!

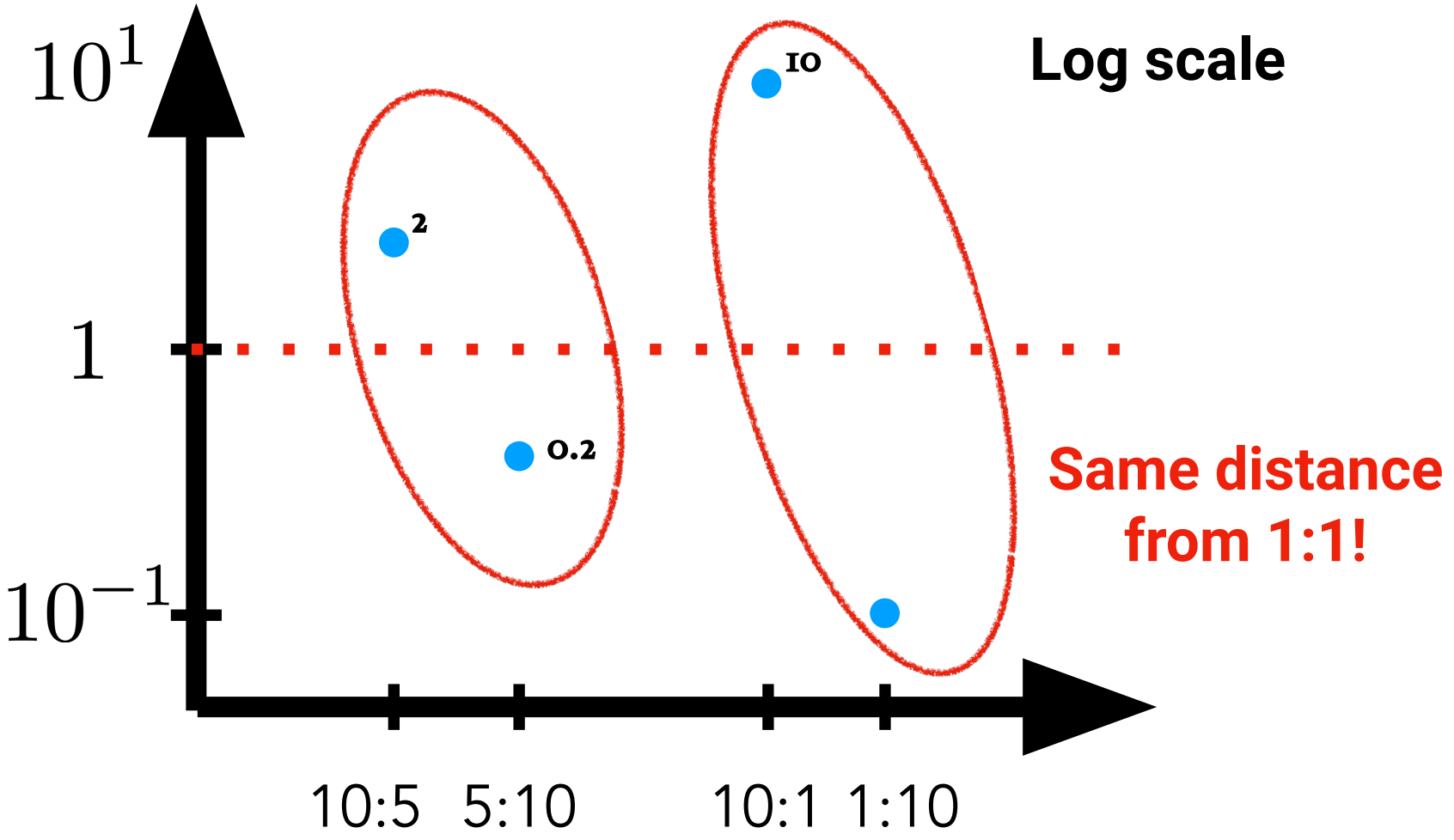
$$\log \frac{A}{B} = \log A - \log B$$

$$\log \frac{10}{5} = \log 2$$

$$\log \frac{5}{10} = \log 1 - \log 2 = -\log 2$$

$$\log \frac{10}{1} = \log 10$$

$$\log \frac{1}{10} = \log 1 - \log 10 = -\log 10$$


10^1
1
10^{-1}
2
0.2
10
Log scale
Same distance from 1:1!
10:5
5:10
10:1
1:10