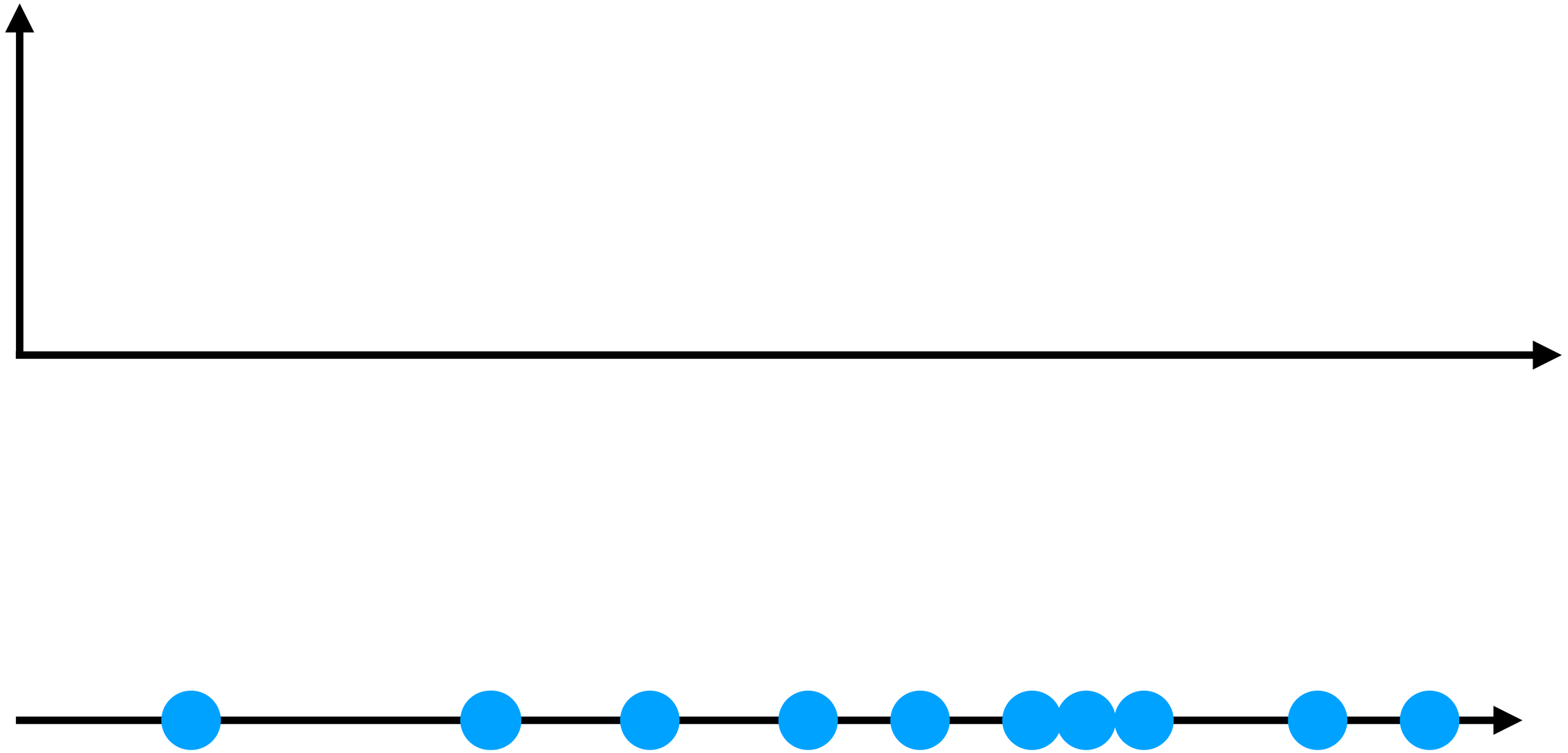
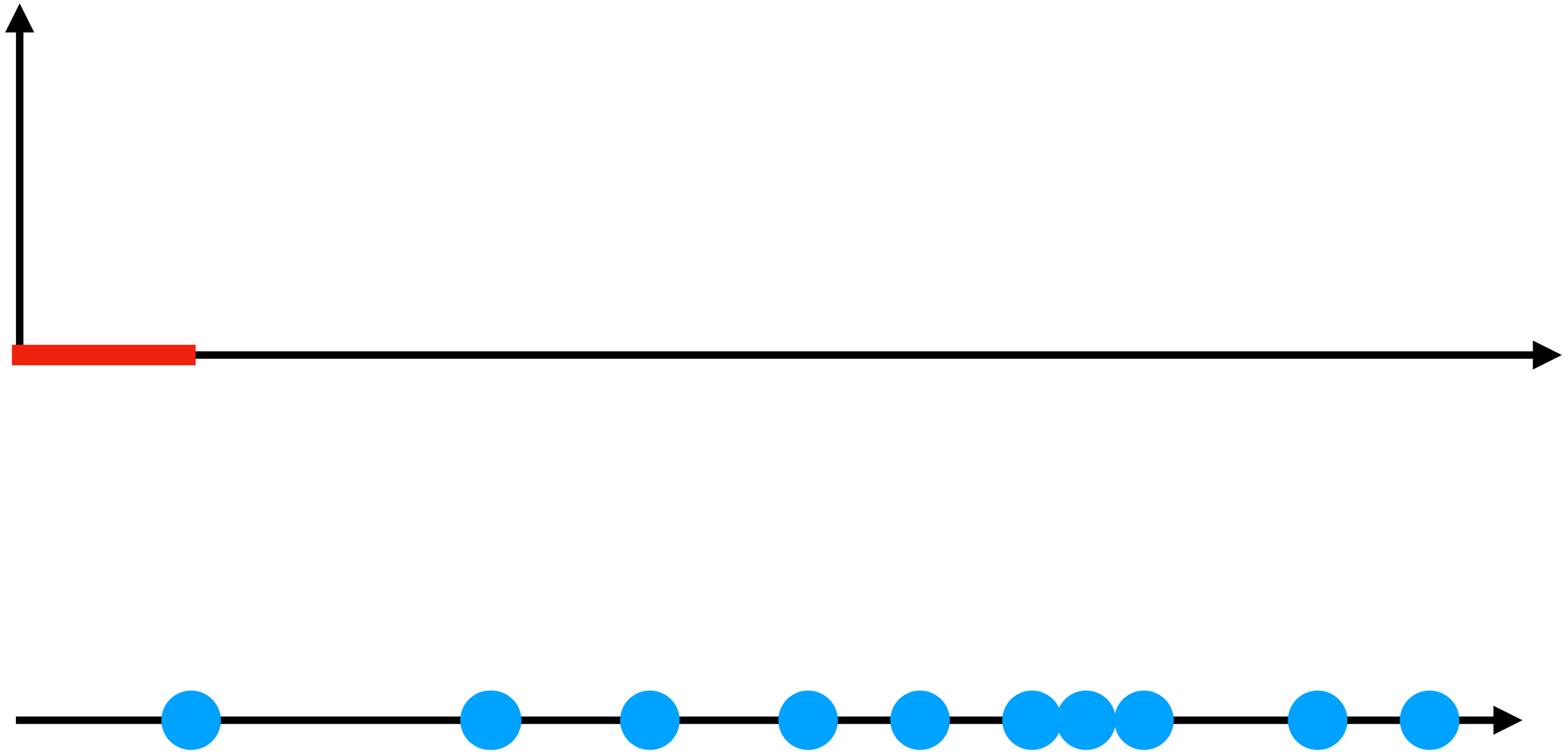
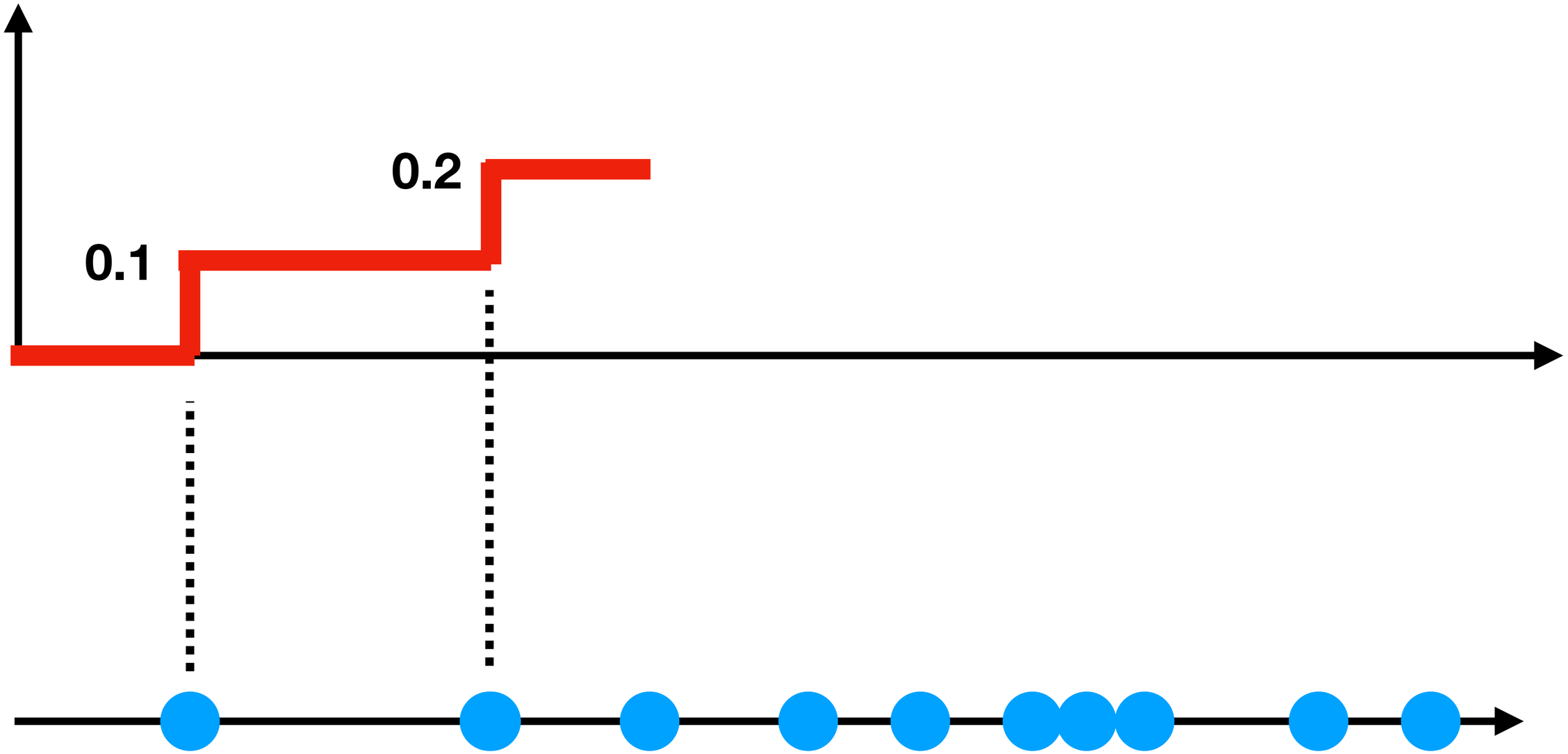
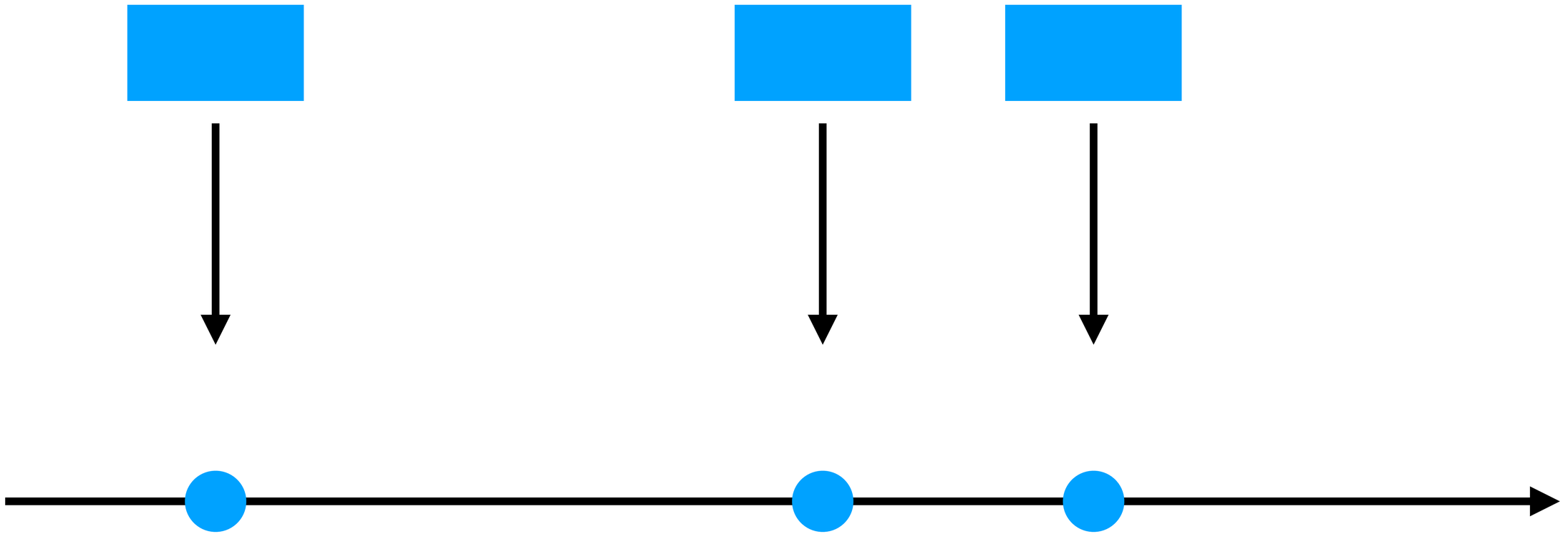
Quiz

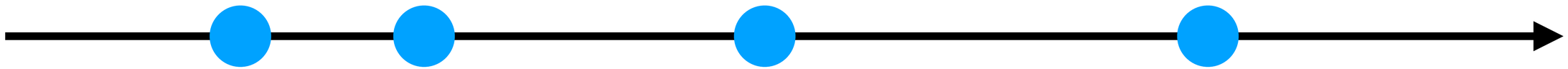
- Draw an empirical CDF plot using the following data points: {10, 1, 2, 5, 8, 10, 5}
- Explain how KDE (Kernel Density Estimation) works.
- Explain the choices that you have when creating a KDE.

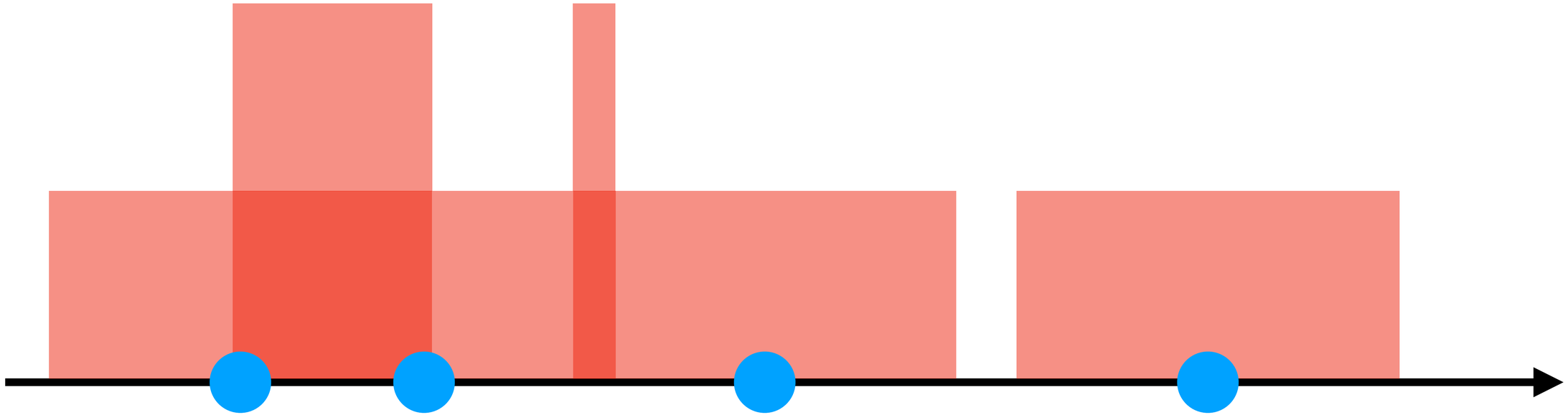


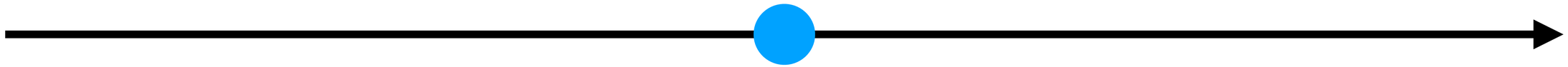


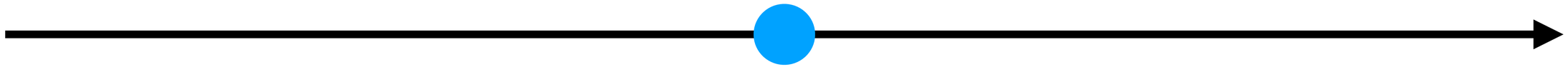

0.2
0.1

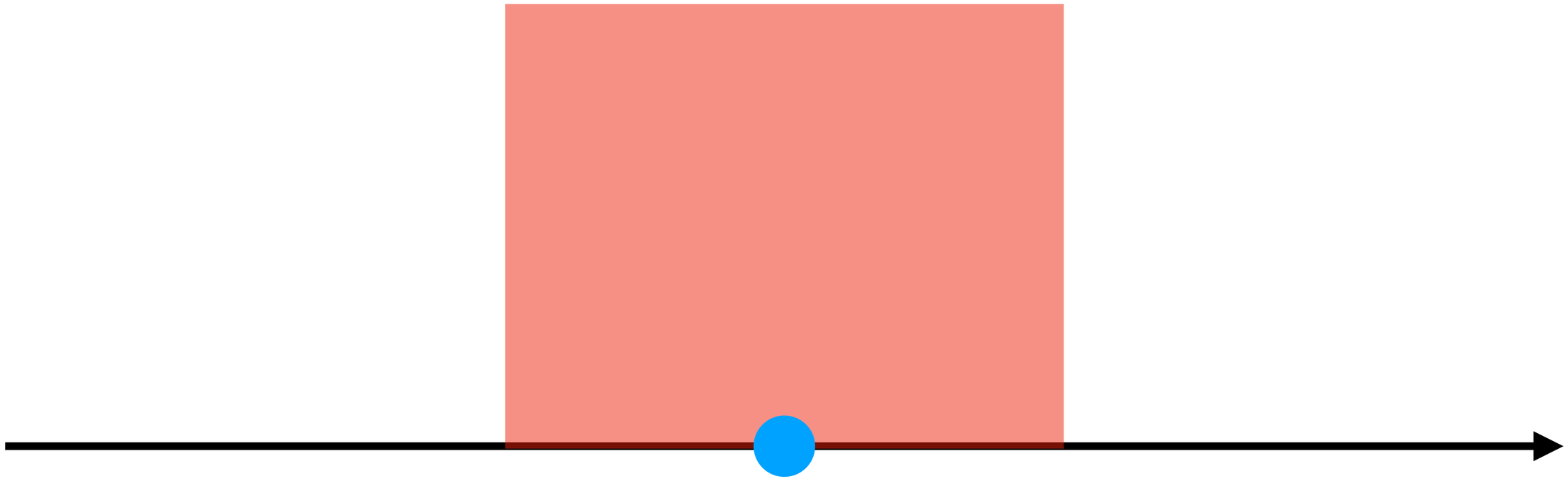


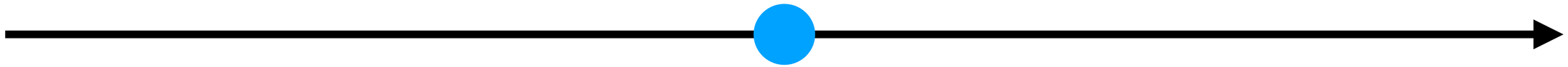


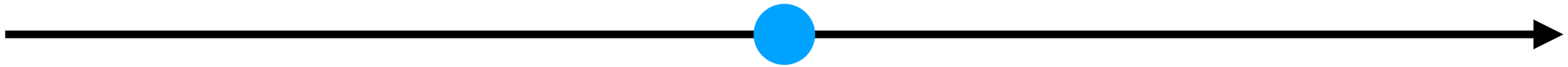


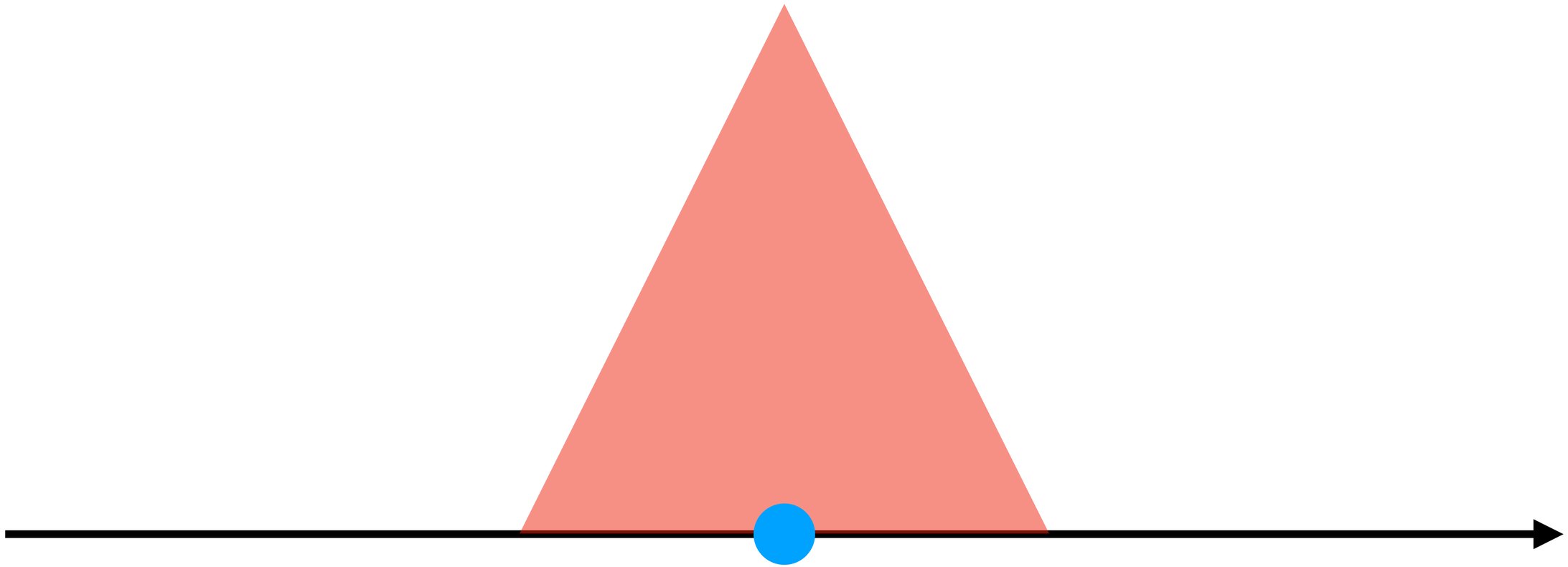


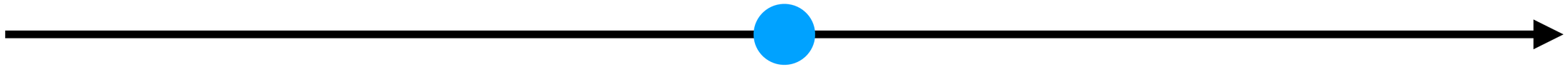


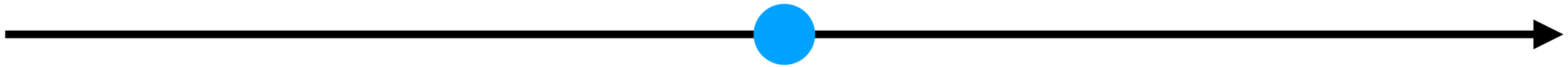


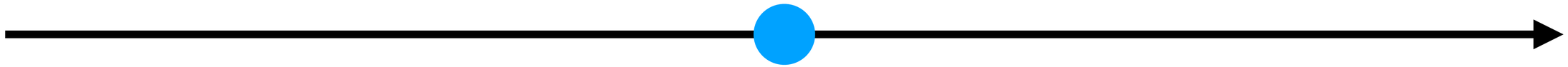


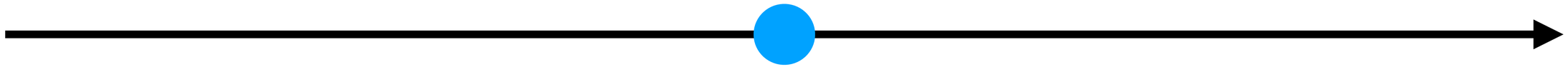


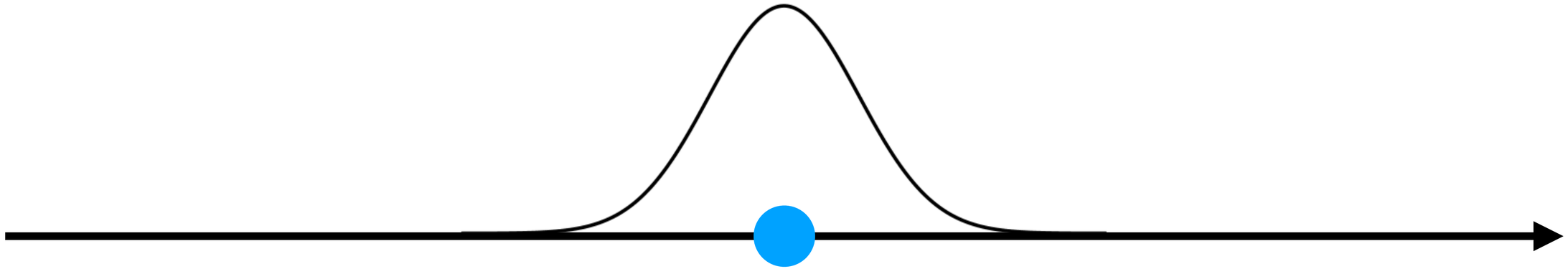


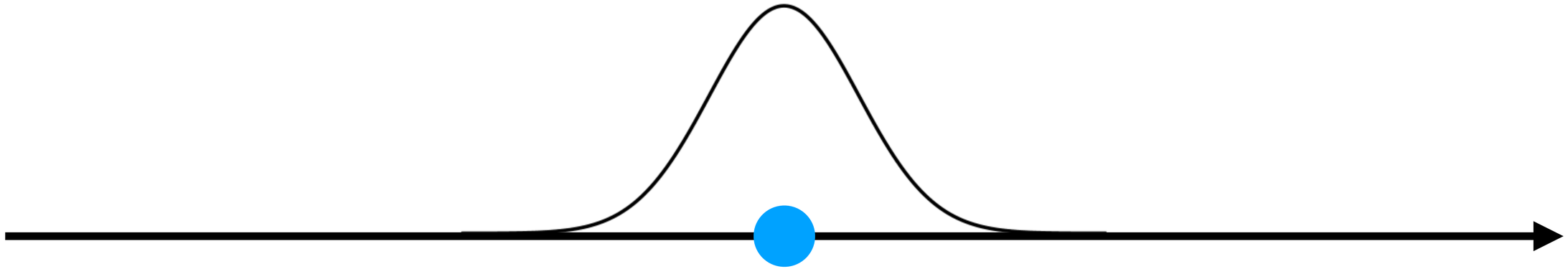


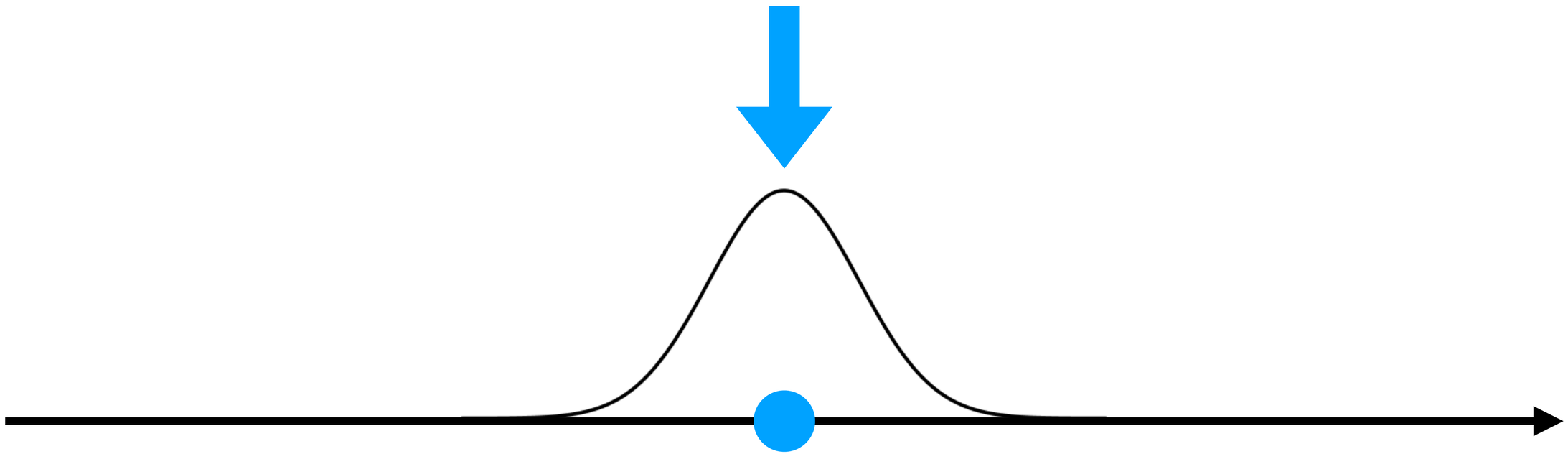


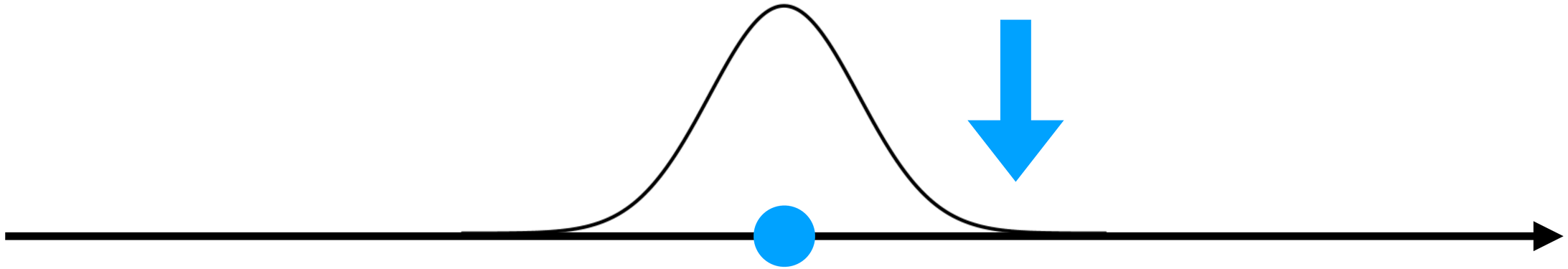


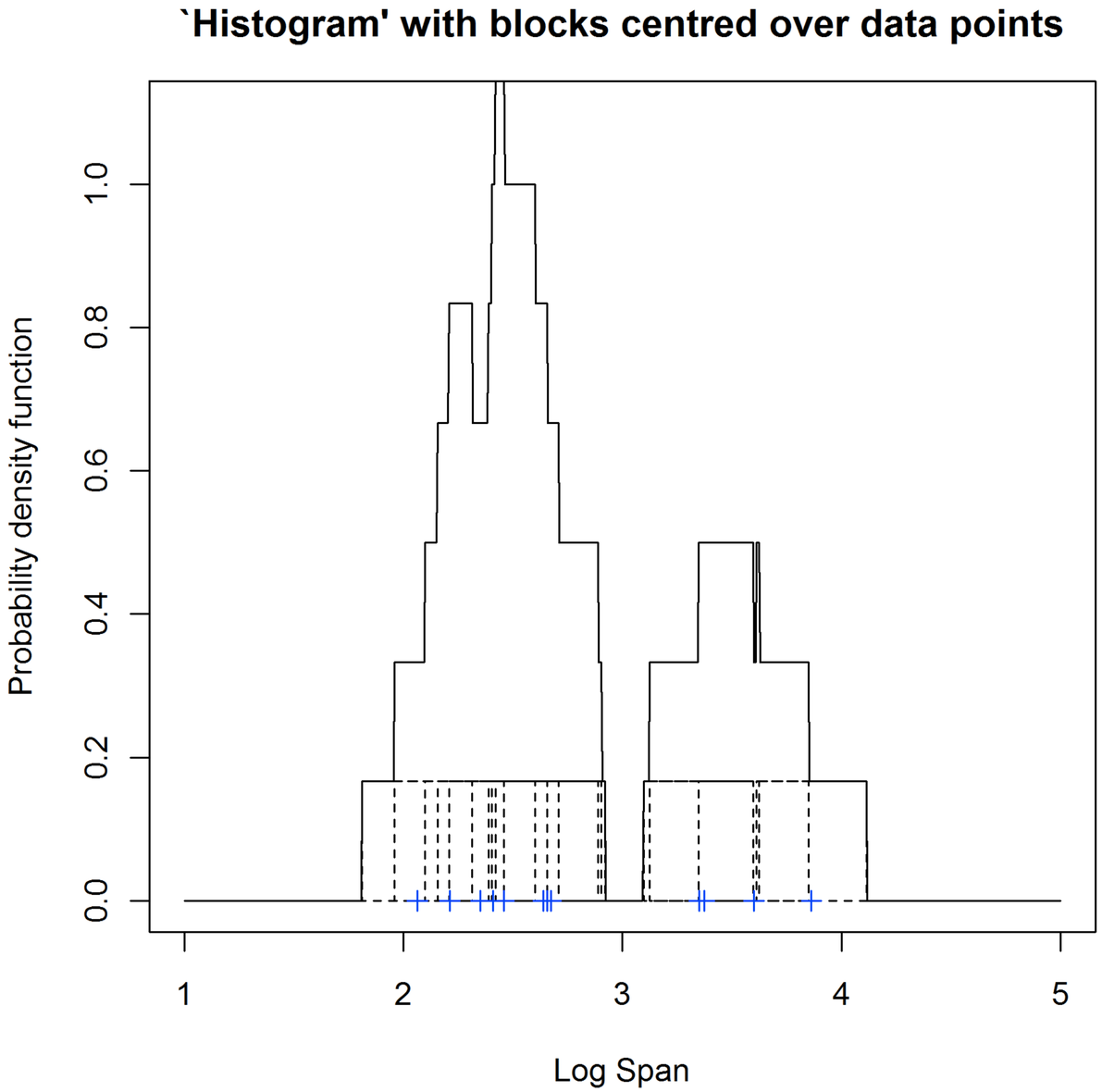




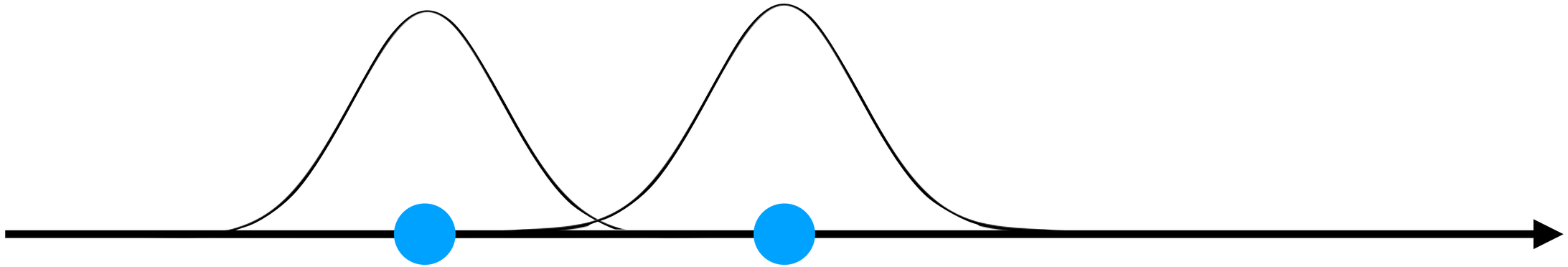




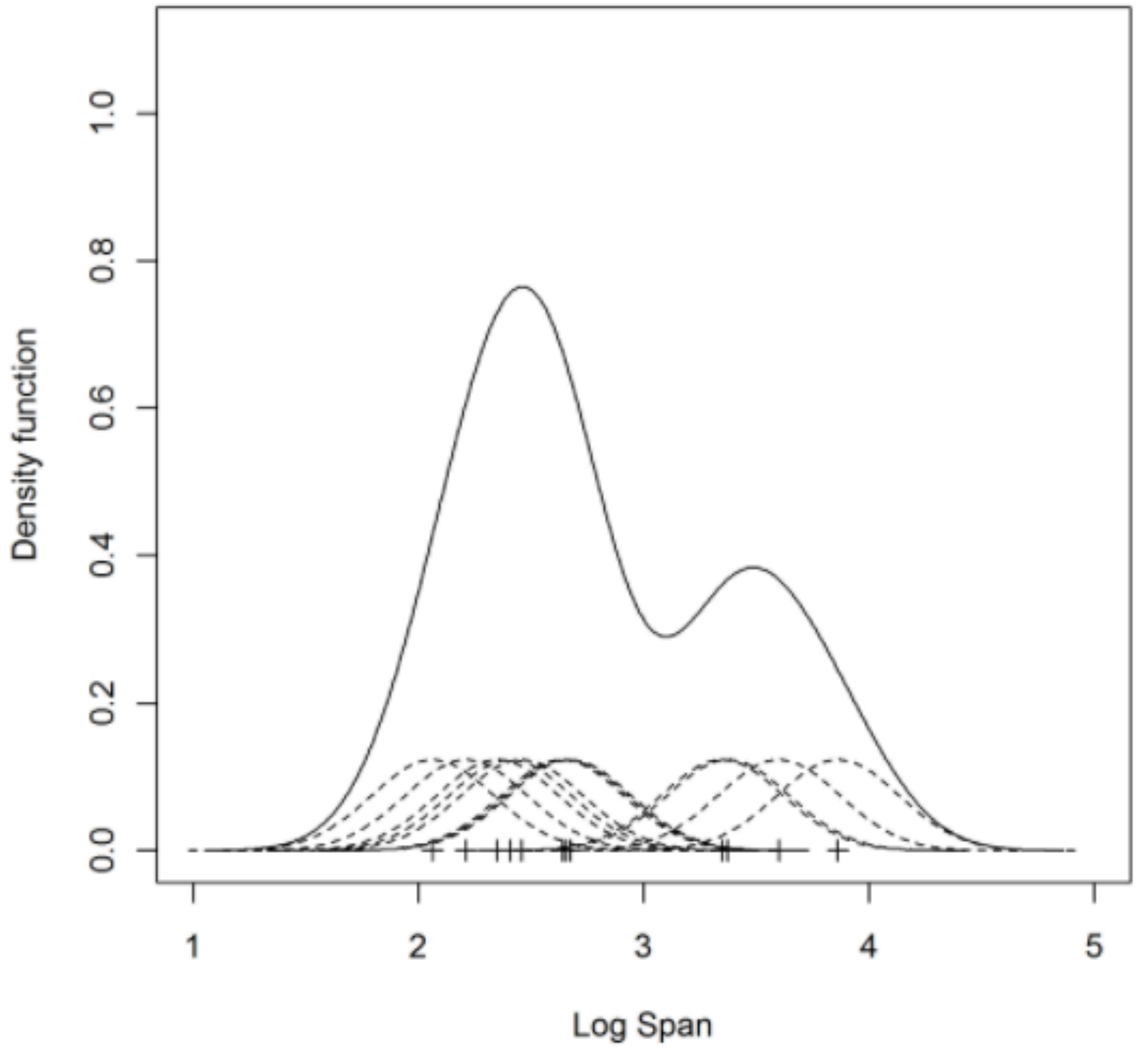




Tarn Duong, An introduction to kernel density estimation,
http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

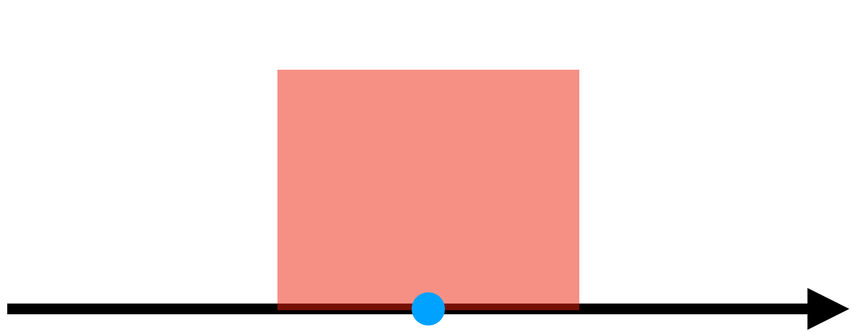


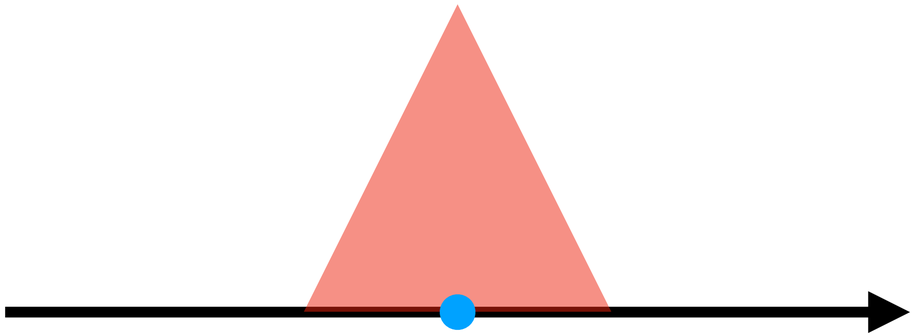
Kernel **D**ensity **E**stimation (KDE)

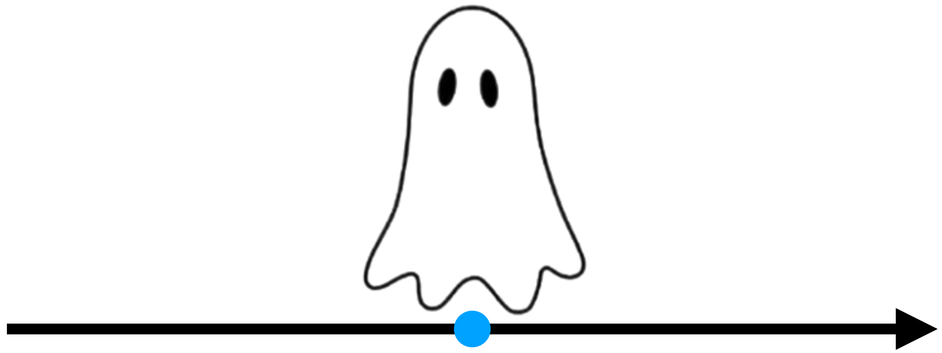


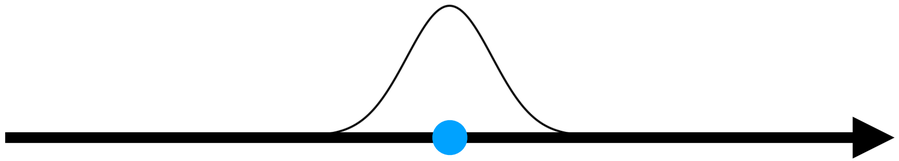
Tarn Duong, An introduction to kernel density estimation, http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

"Kernels"

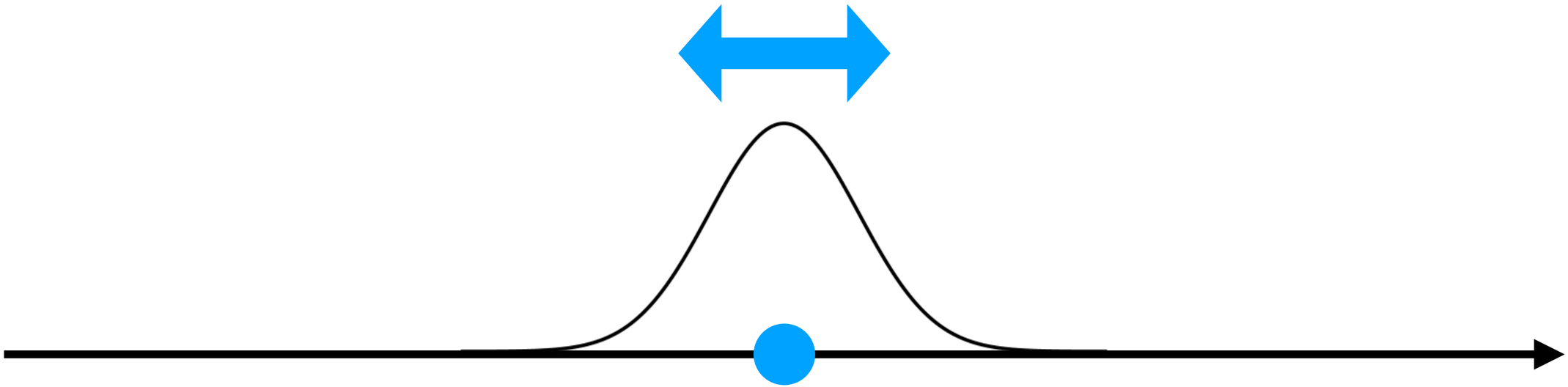


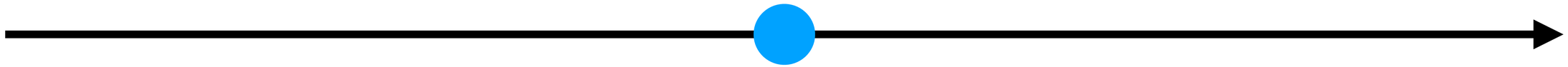


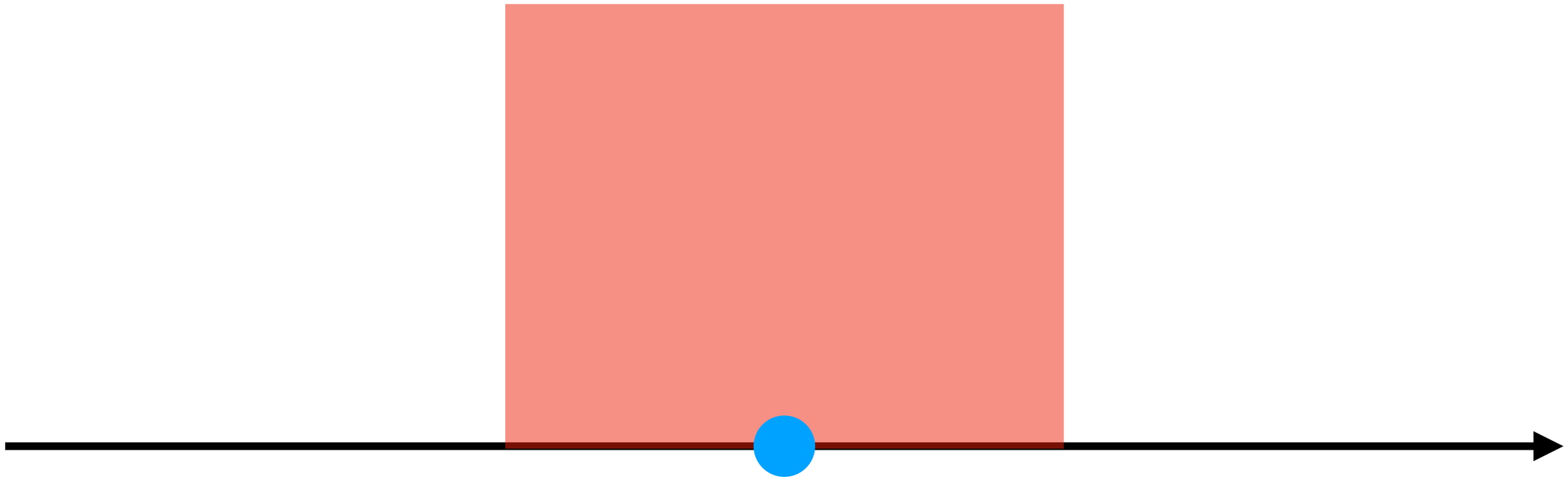


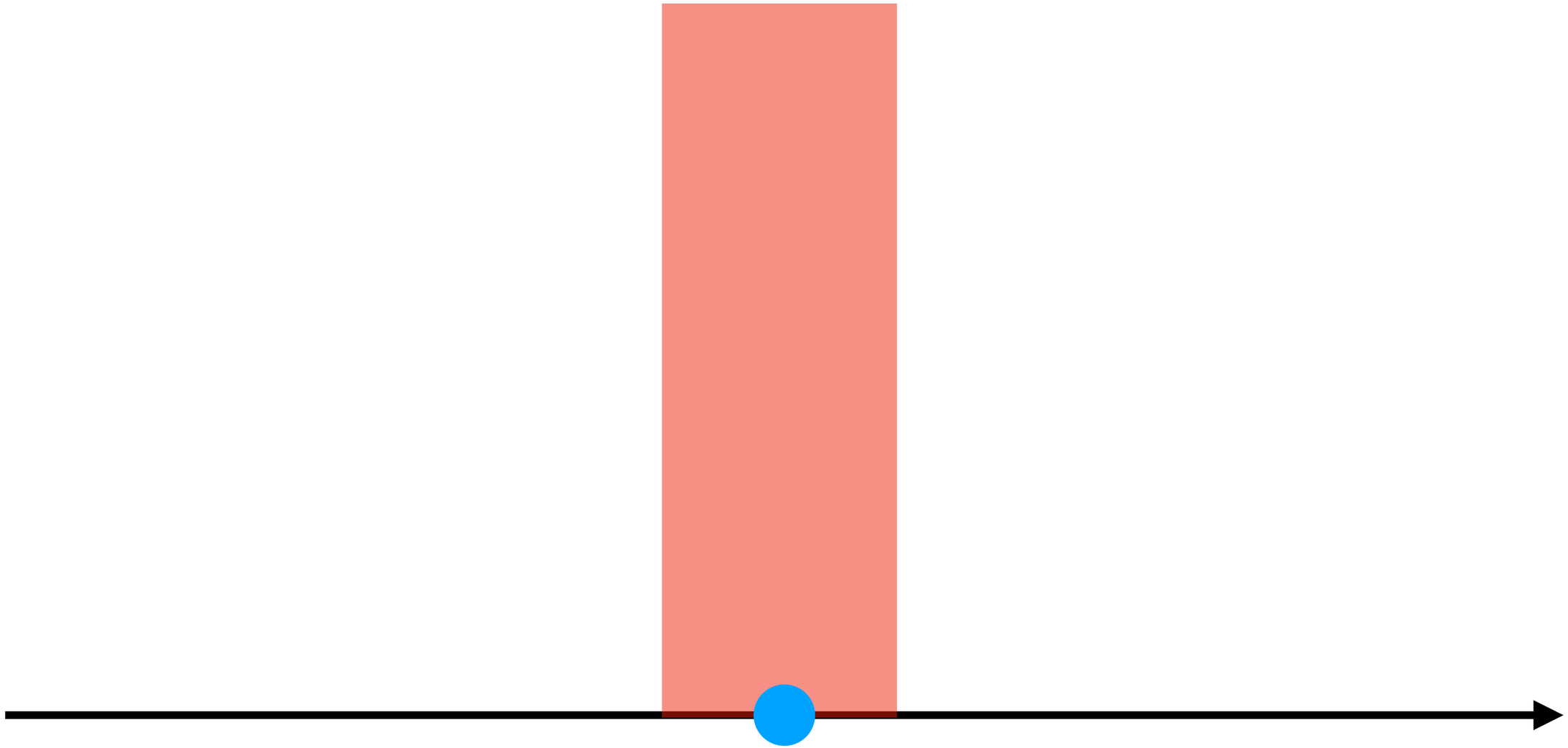


Bandwidth









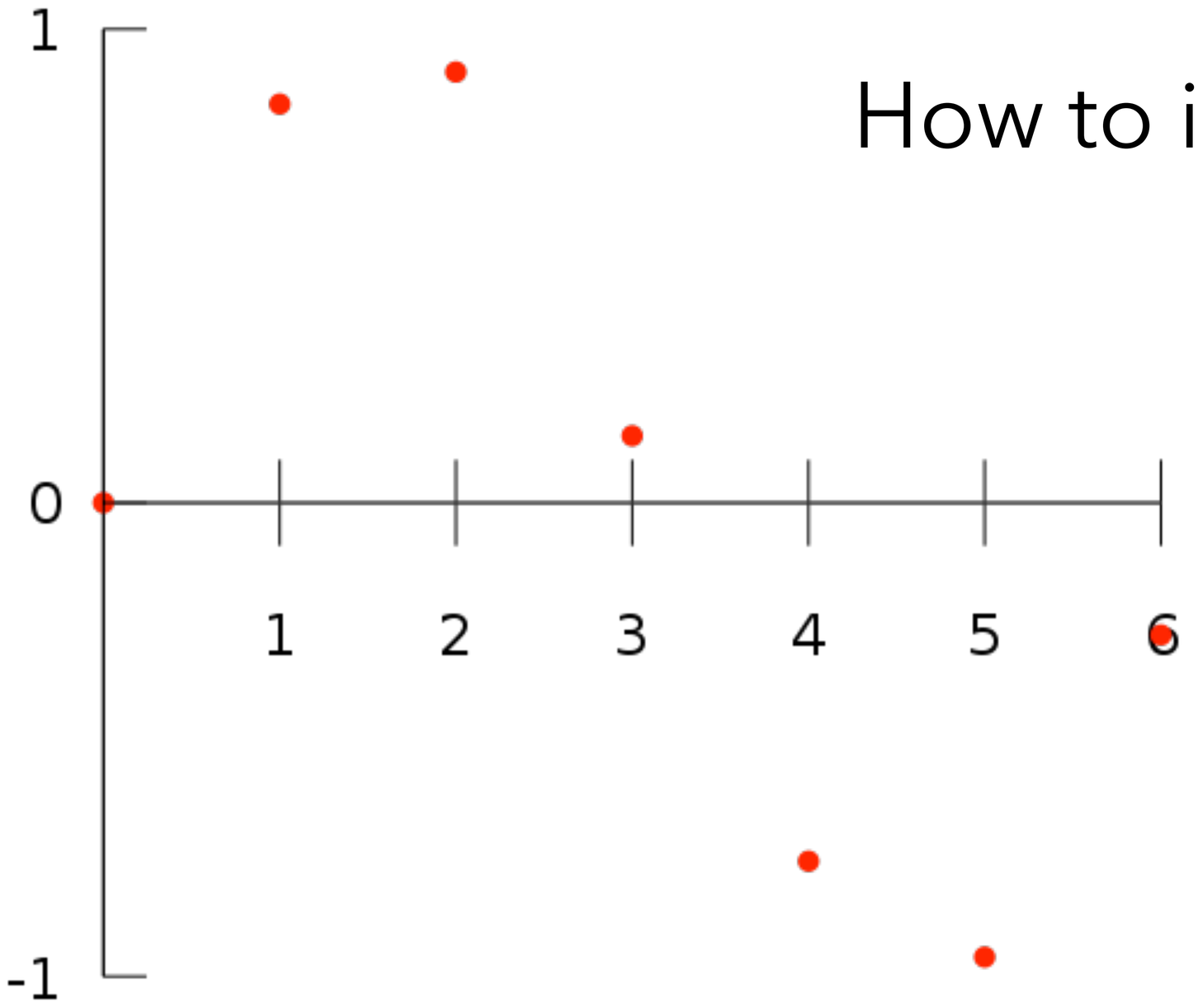
Interpolation & Extrapolation

Interpolation

How can we fill the gaps between (or beyond) data points?

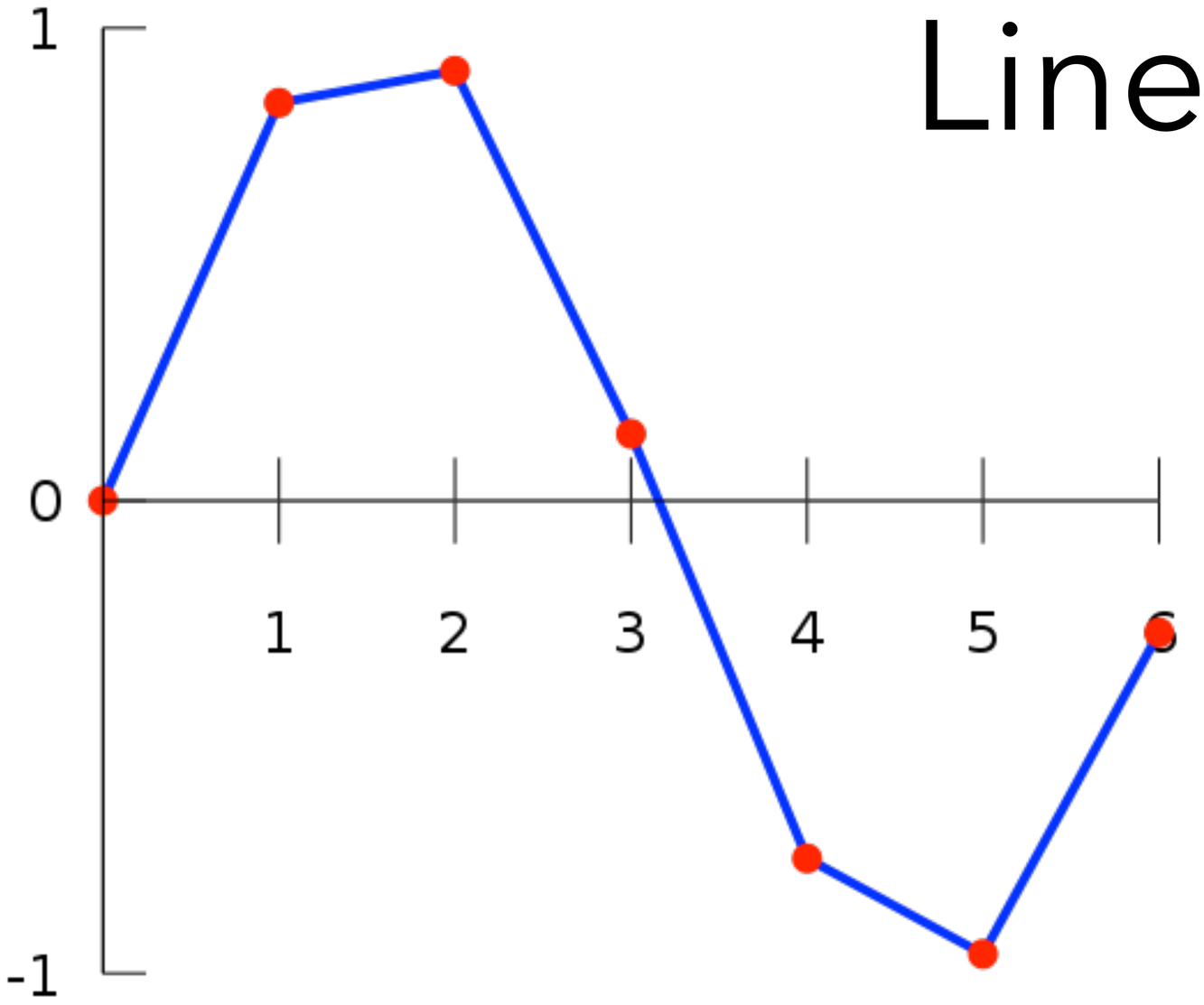
Interpolation

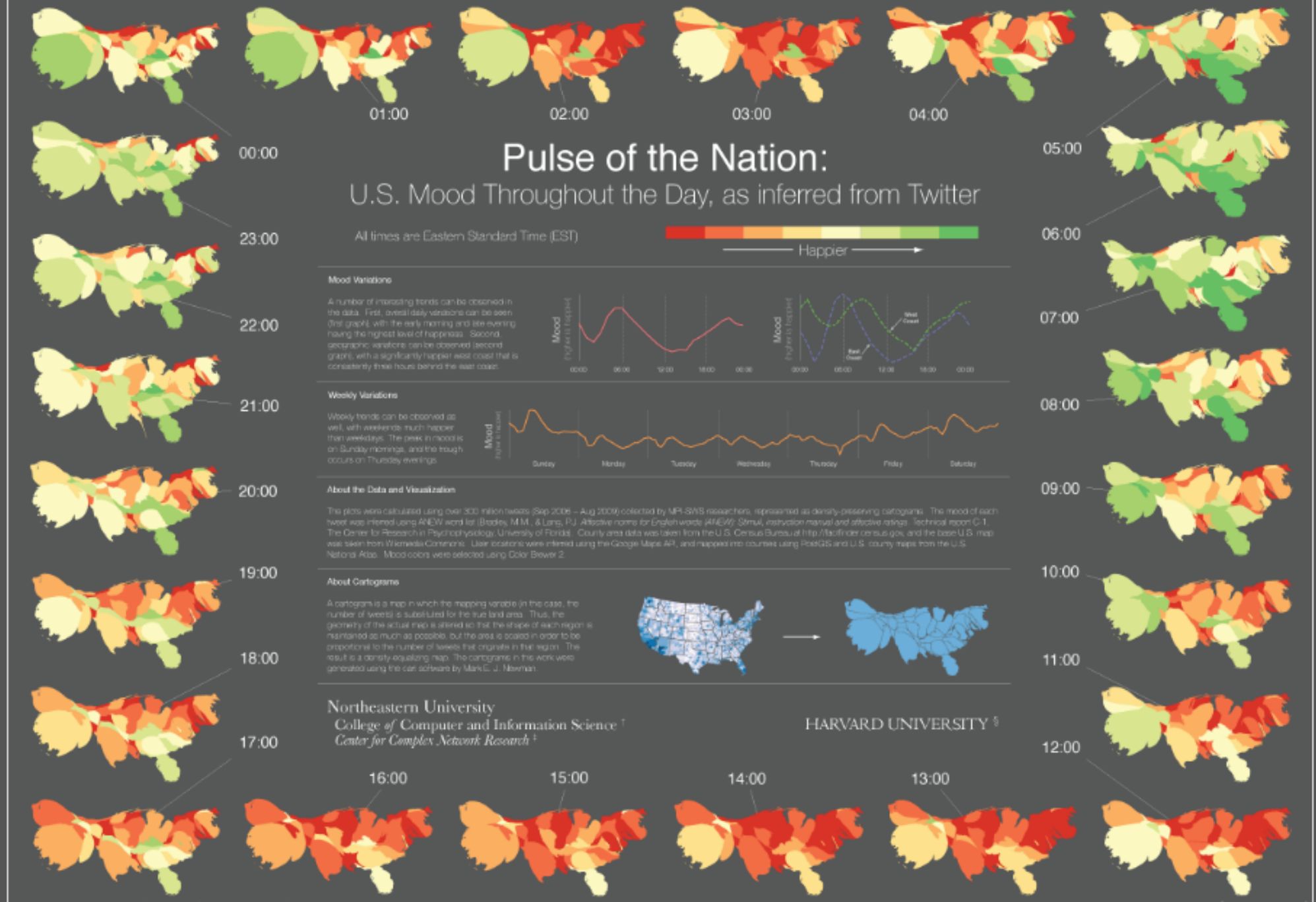
Let’s connect the dots.



How to interpolate?

Linear





Pulse of the Nation:

U.S. Mood Throughout the Day, as inferred from Twitter

All times are Eastern Standard Time (EST)

Happier

Mood Variations

A number of interesting trends can be observed in the data. First, overall daily variations can be seen (first graph), with the early morning and late evening having the highest level of happiness. Second, geographic variations can be observed (second graph), with a significantly happier west coast that is consistently three hours behind the east coast.

Weekly Variations

Weekly trends can be observed as well, with weekends much happier than weekdays. The peak in mood is on Sunday mornings, and the trough occurs on Thursday evenings.

About the Data and Visualization

The plots were calculated using over 300 million tweets (Sep 2006 – Aug 2009) collected by MR-SWS researchers, represented as density-preserving cartograms. The mood of each tweet was inferred using ANEW word list (Bradley, M.M., & Lang, P.J. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1. The Center for Research in Psychophysiology, University of Florida). County area data was taken from the U.S. Census Bureau at http://factfinder.census.gov, and the base U.S. map was taken from Wikimedia Commons. User locations were inferred using the Google Maps API, and mapped into counties using PostGIS and U.S. county maps from the U.S. National Atlas. Mood colors were selected using Color Brewer 2.

About Cartograms

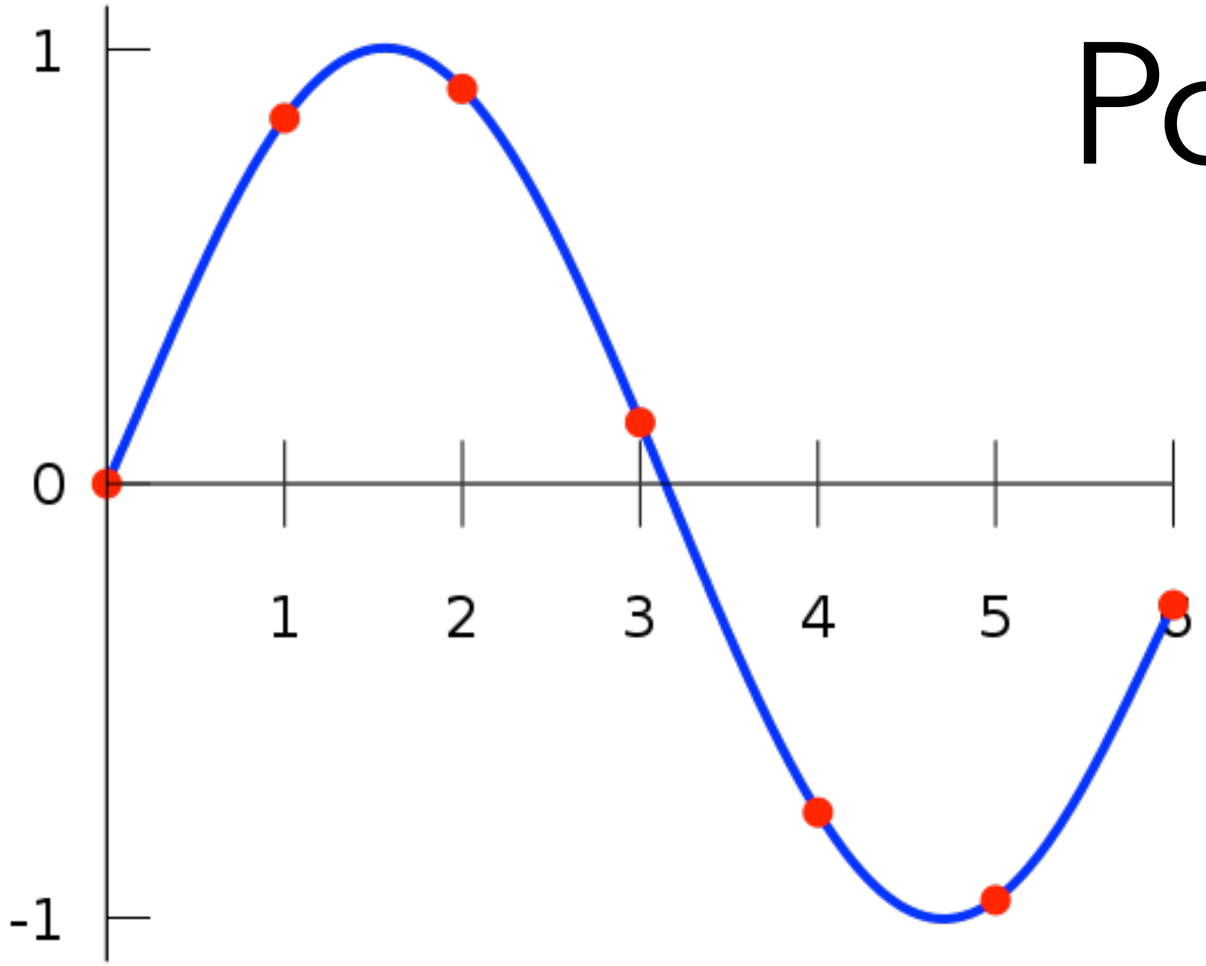
A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. The result is a density-equalizing map. The cartograms in this work were generated using the cart software by Mark E. J. Newman.

Northeastern University
College *of* Computer and Information Science †
Center for Complex Network Research ‡

HARVARD UNIVERSITY §

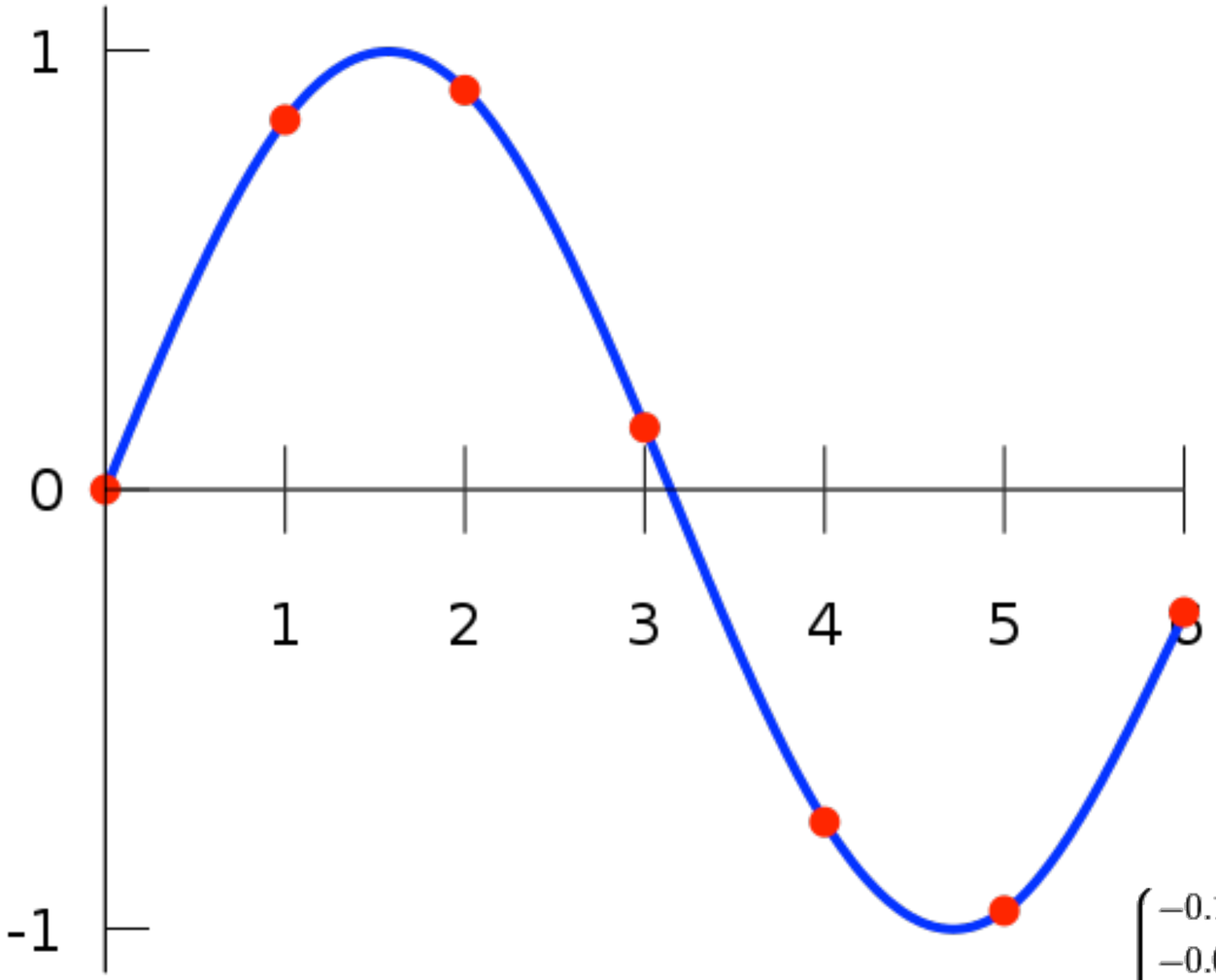
http://www.ccs.neu.edu/home/amislove/twittermood

Polynomial



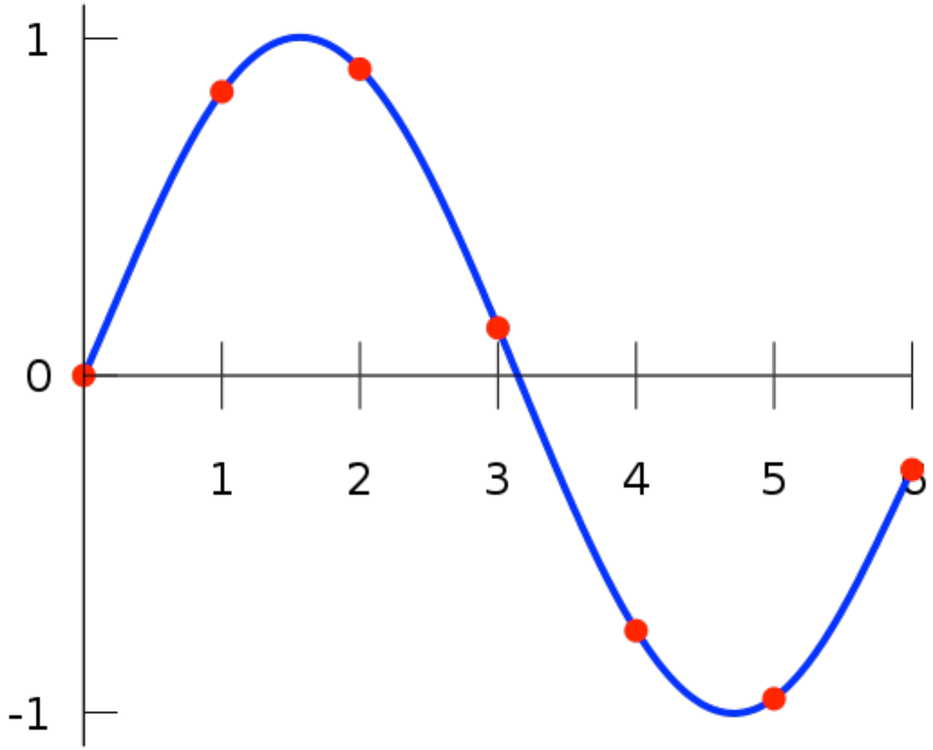
$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0.$$

Splines

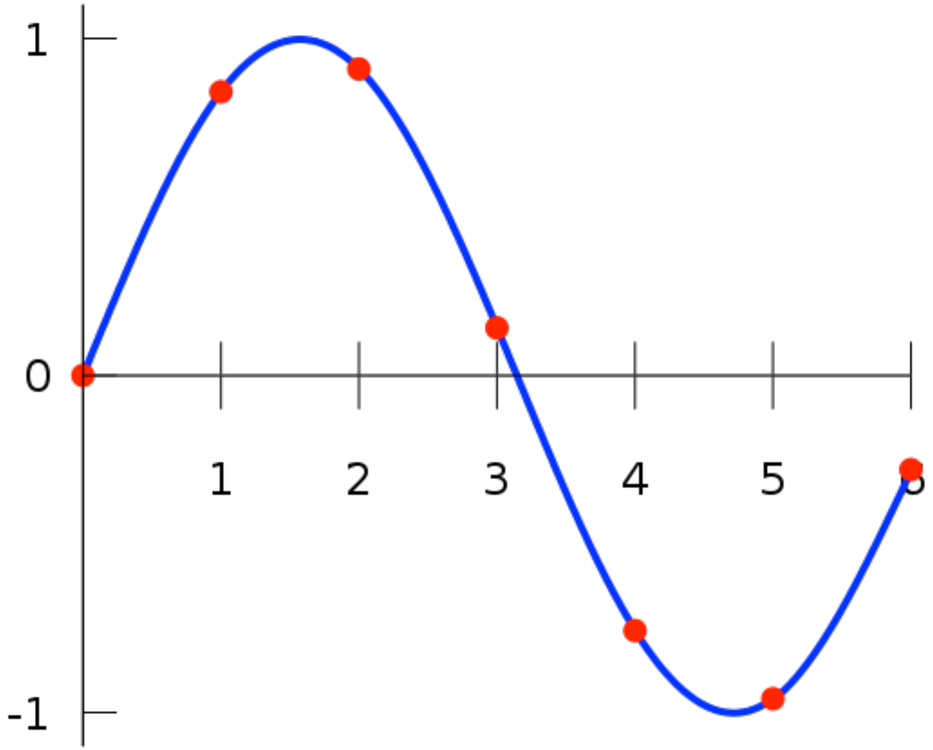


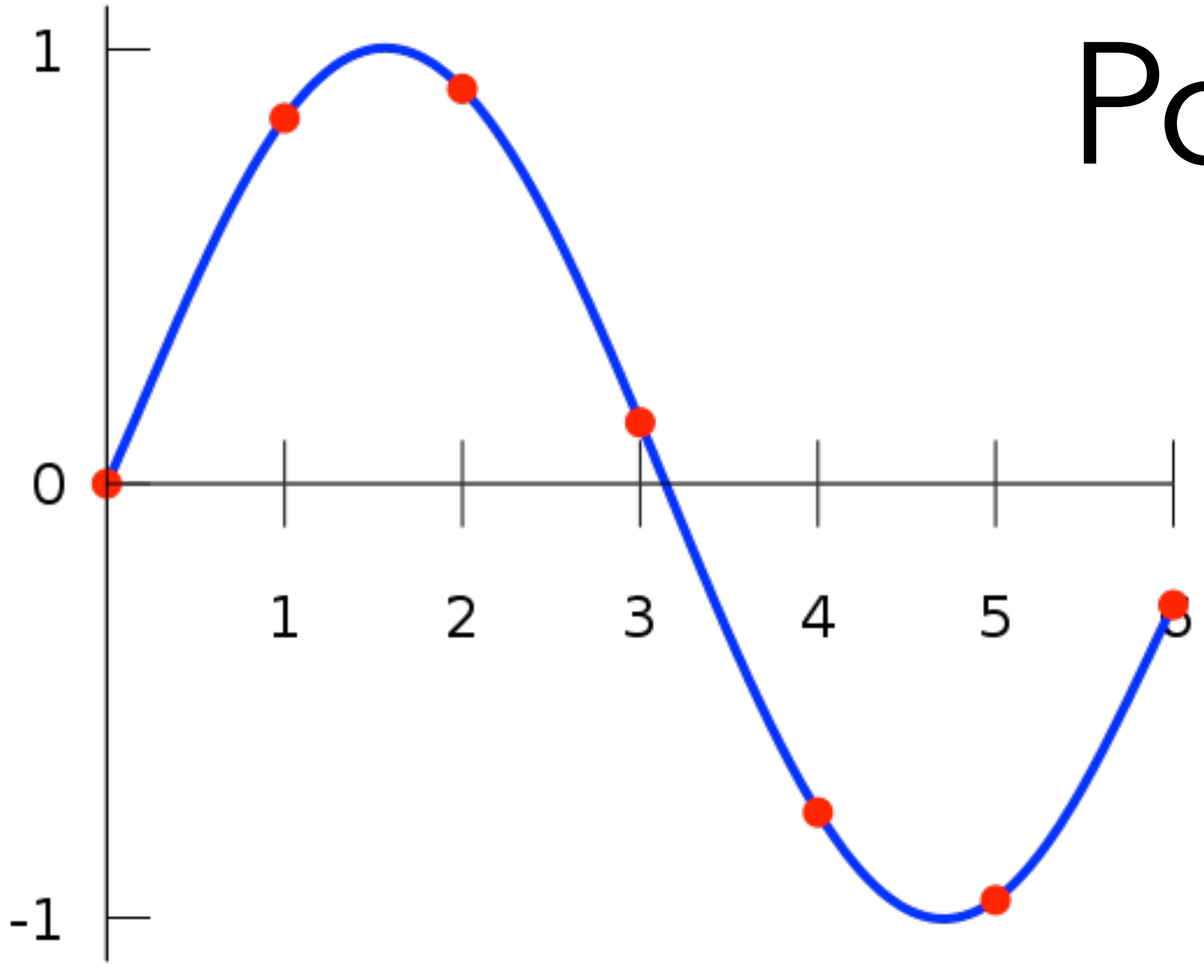
$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$

???



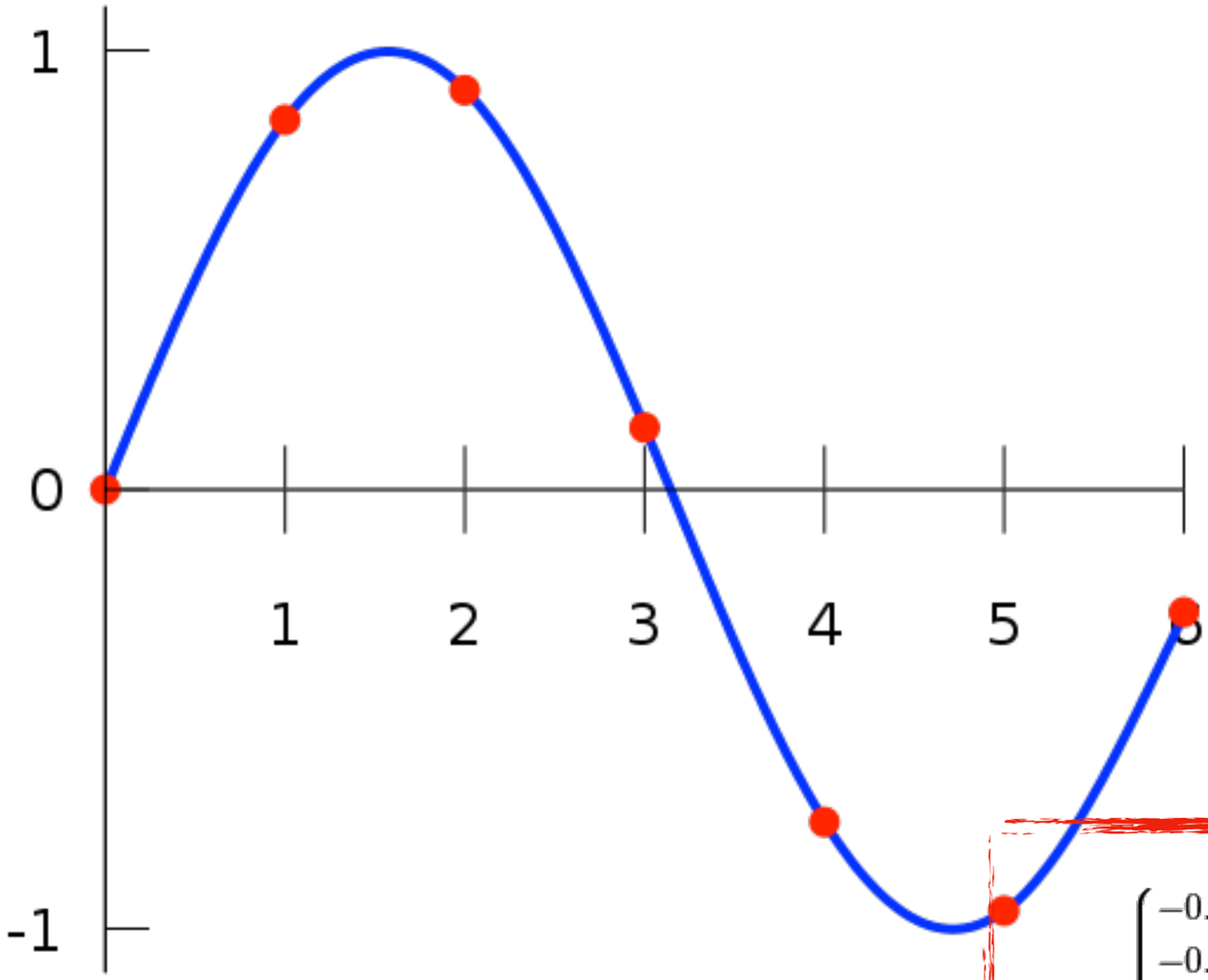
vs.





Polynomial

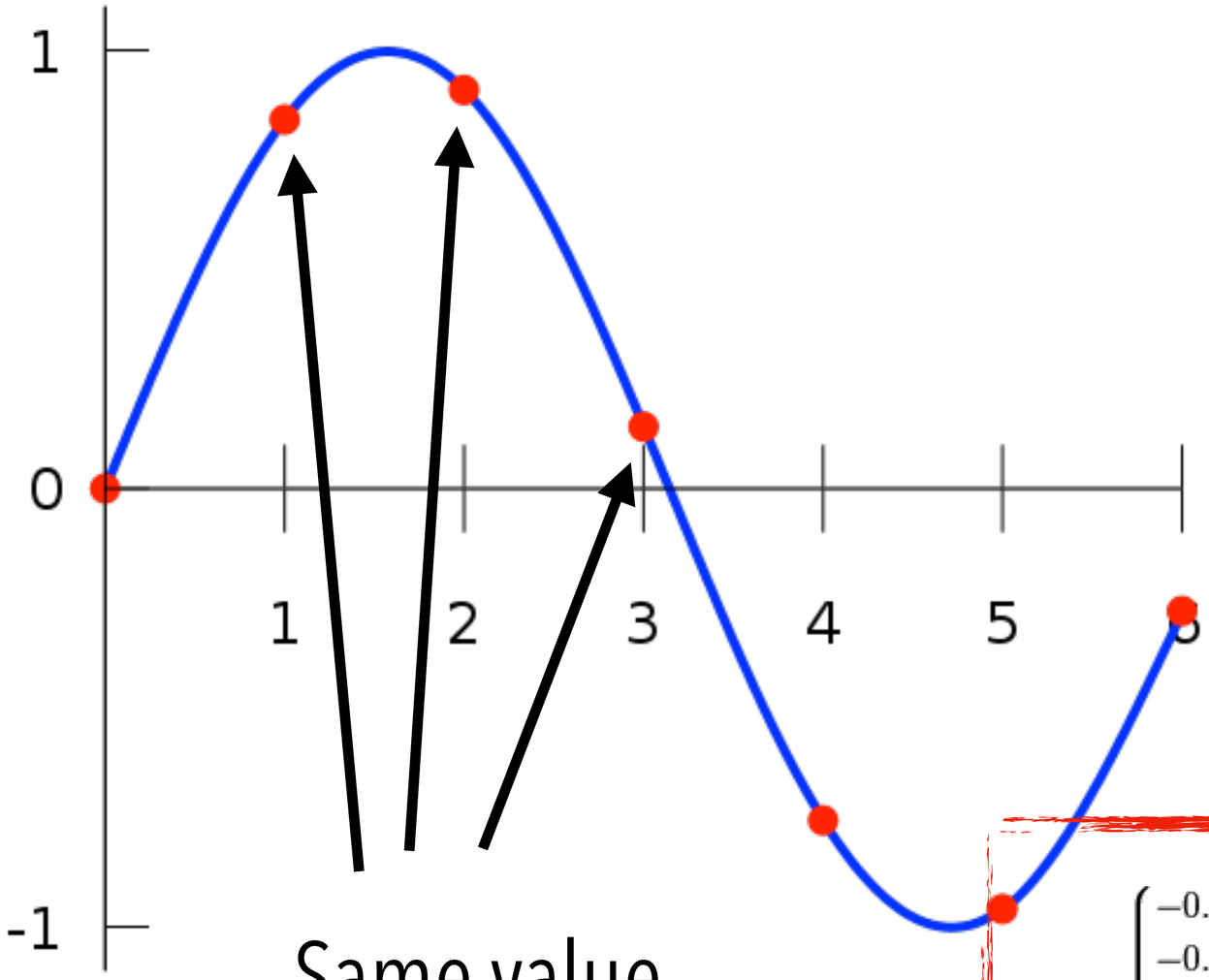
$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0.$$



Splines

$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$

Splines



Same value

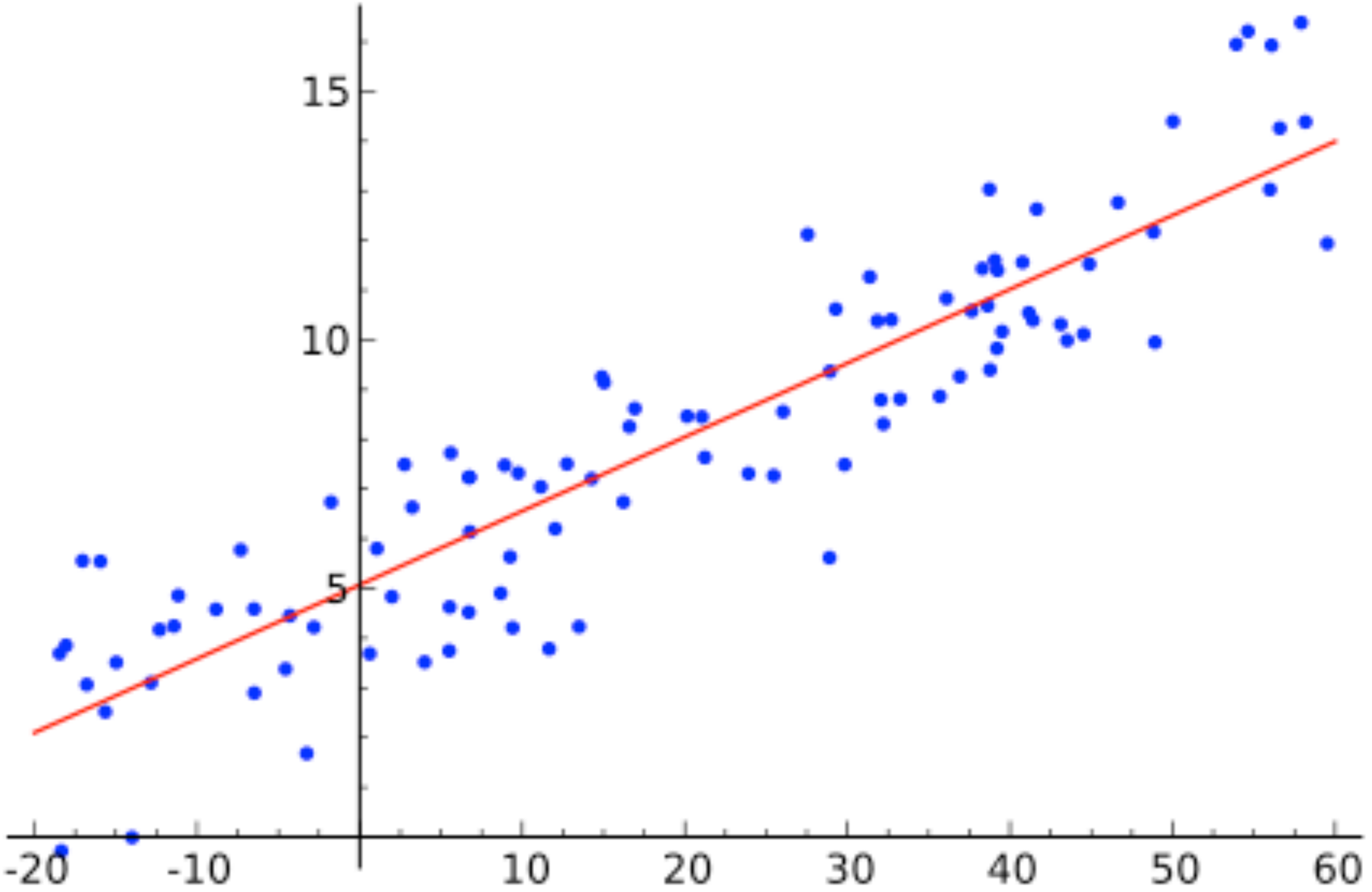
Same f′ and f″

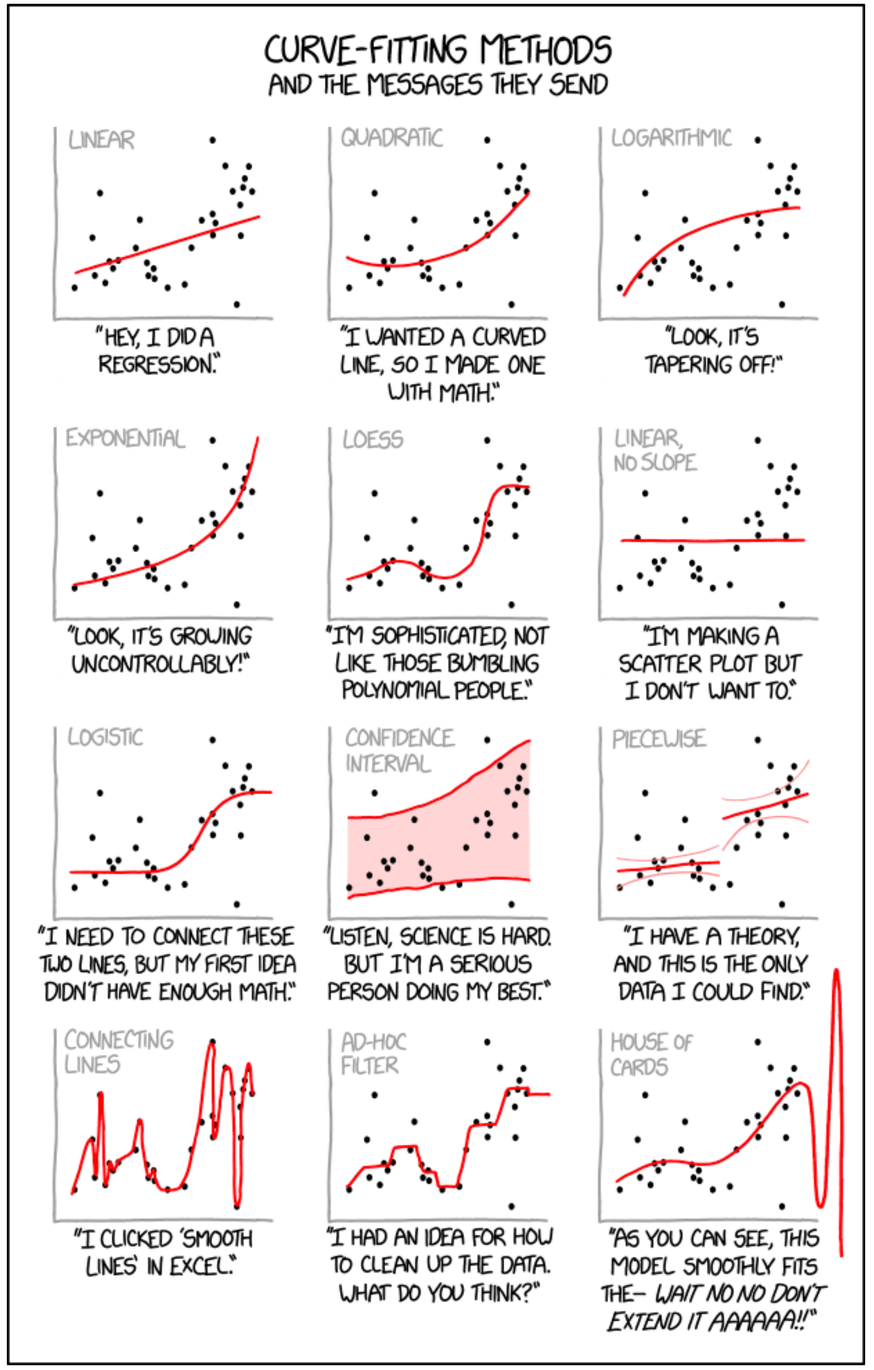
$$f(x) = \begin{cases} -0.1522x^3 + 0.9937x, & \text{if } x \in [0,1], \\ -0.01258x^3 - 0.4189x^2 + 1.4126x - 0.1396, & \text{if } x \in [1,2], \\ 0.1403x^3 - 1.3359x^2 + 3.2467x - 1.3623, & \text{if } x \in [2,3], \\ 0.1579x^3 - 1.4945x^2 + 3.7225x - 1.8381, & \text{if } x \in [3,4], \\ 0.05375x^3 - 0.2450x^2 - 1.2756x + 4.8259, & \text{if } x \in [4,5], \\ -0.1871x^3 + 3.3673x^2 - 19.3370x + 34.9282, & \text{if } x \in [5,6]. \end{cases}$$

Ok. Interpolation "connects" the dots.
But what if we have some noise?

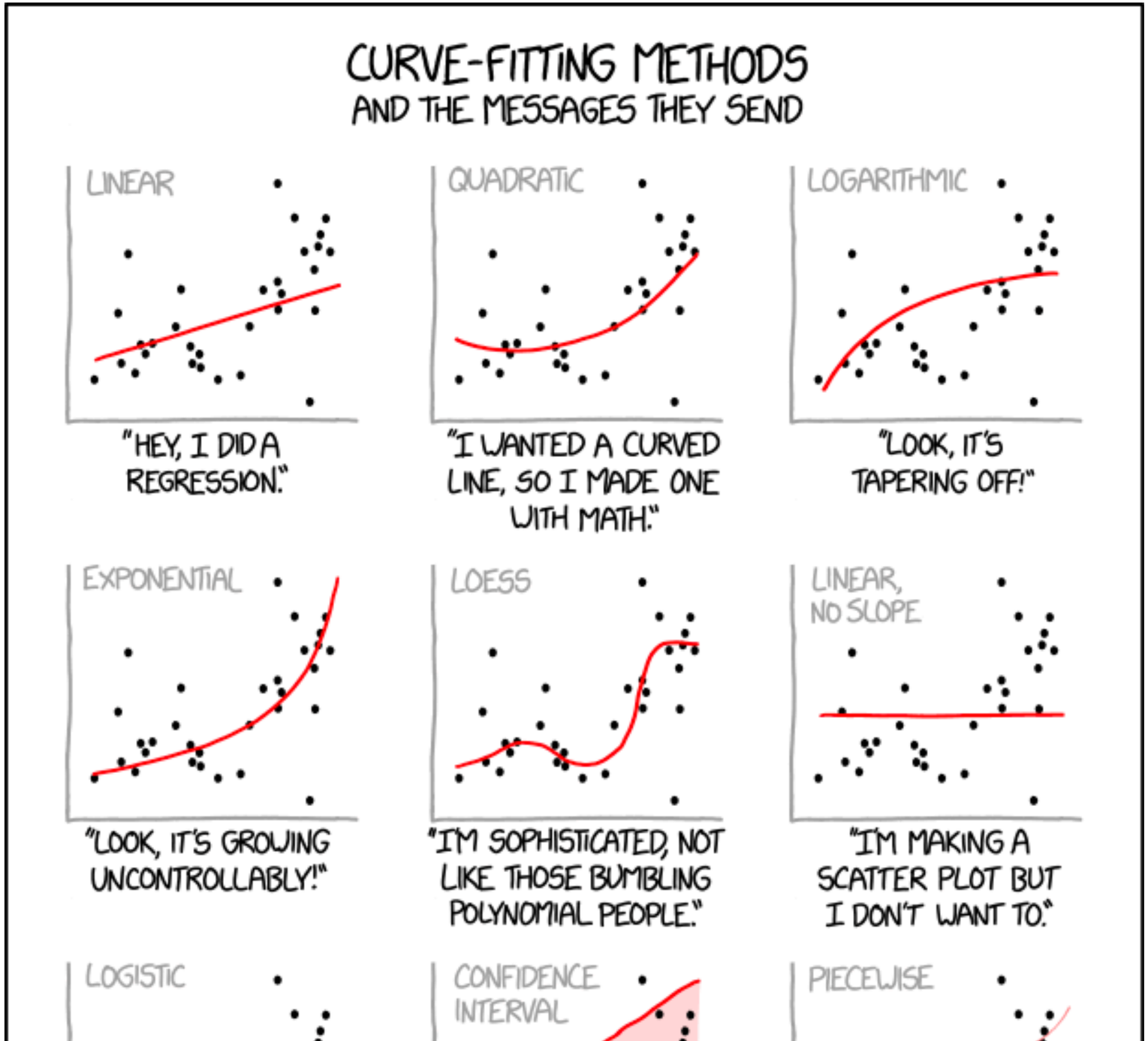
How can we visualize the **trends**?

Regression: a parametric approach

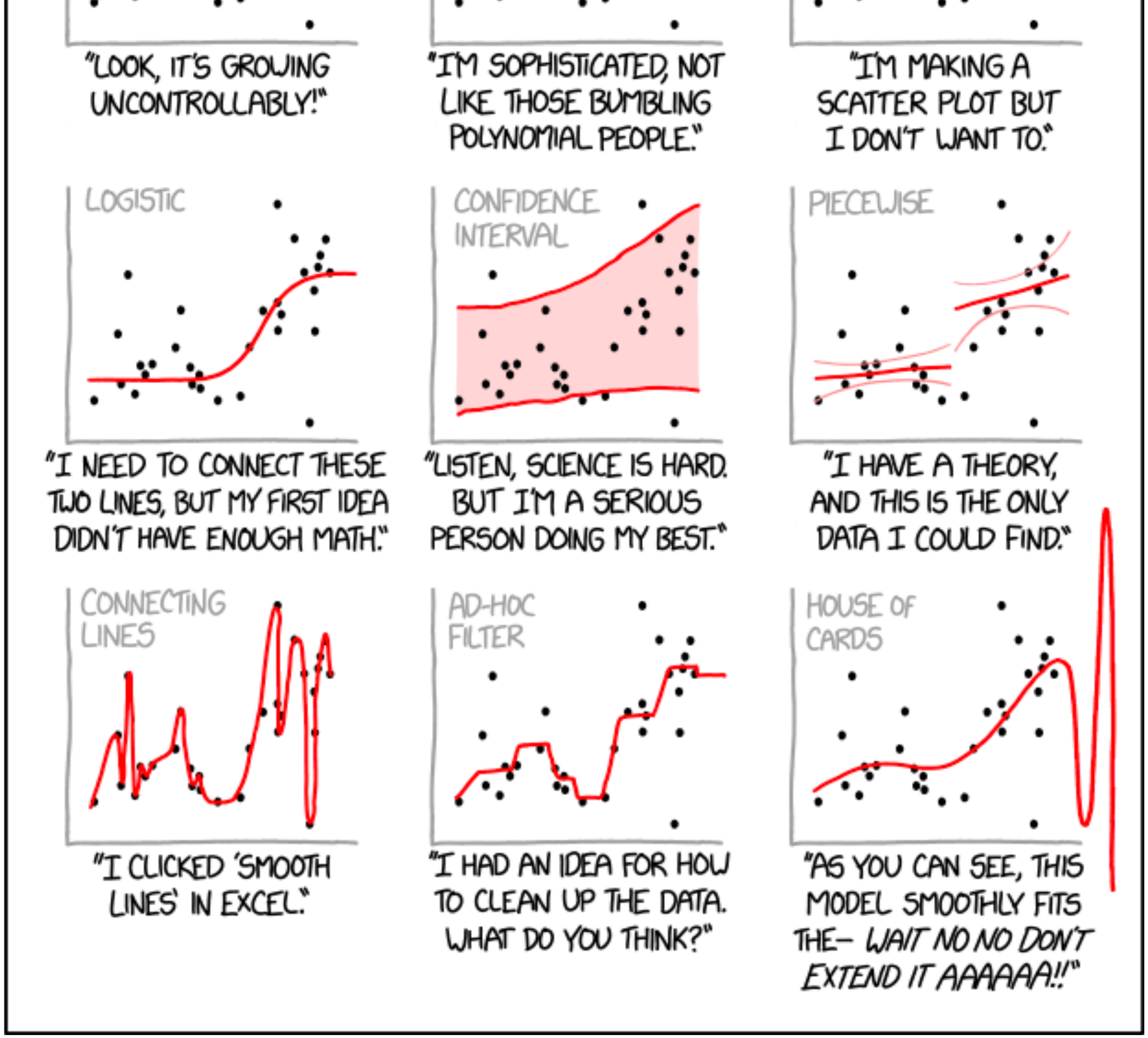




https://www.xkcd.com/2048/



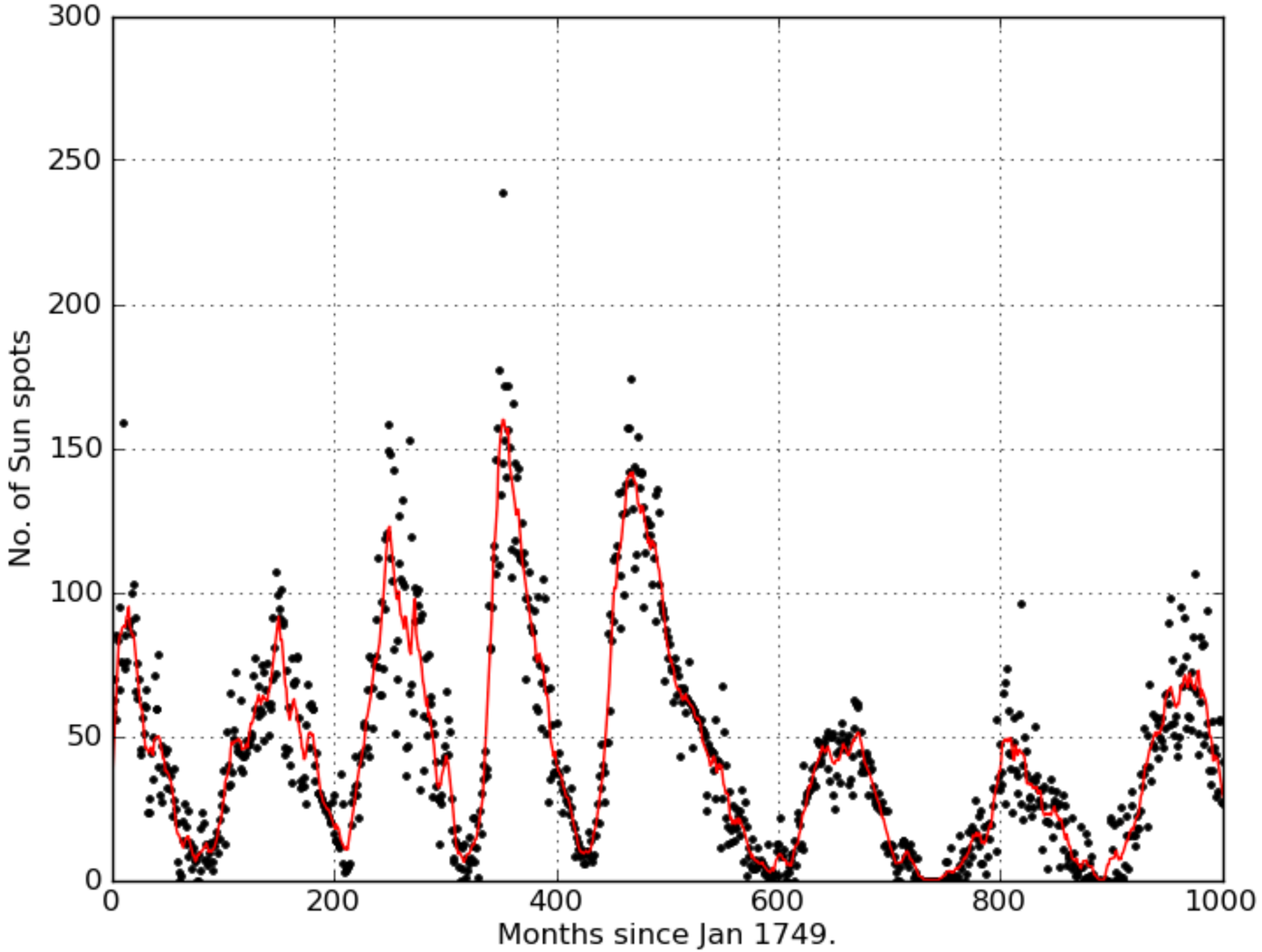
https://www.xkcd.com/2048/



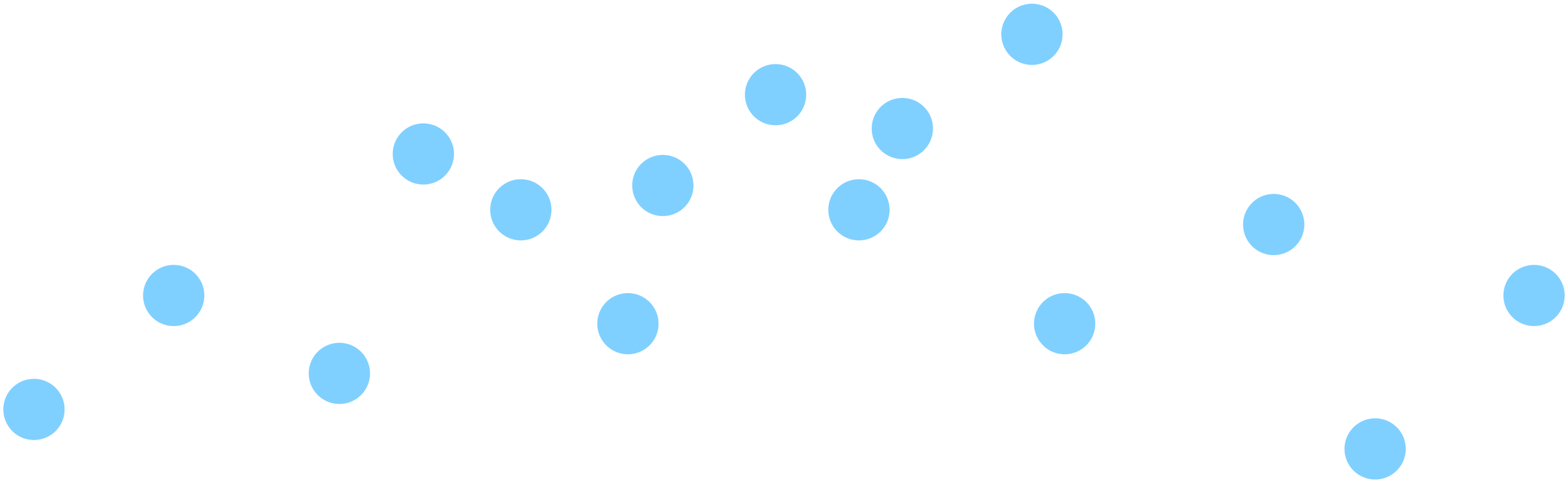
https://www.xkcd.com/2048/

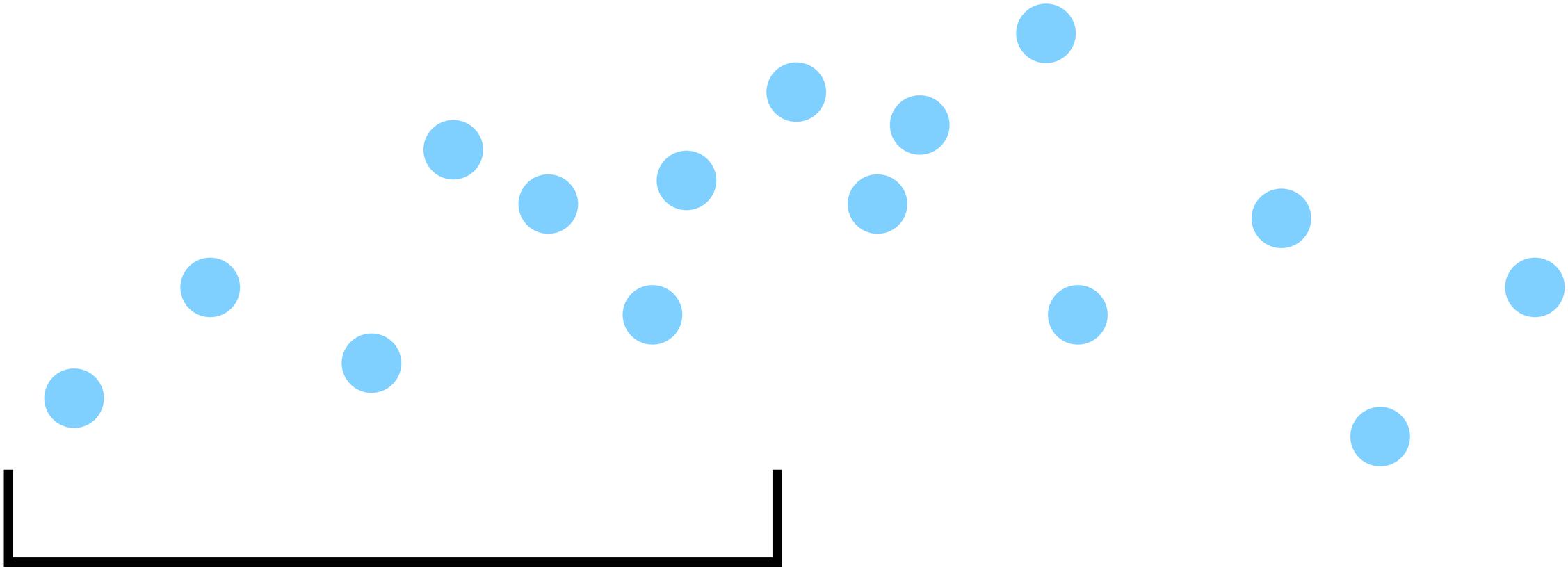
What would be non-parametric approach towards regression (trend detection)?

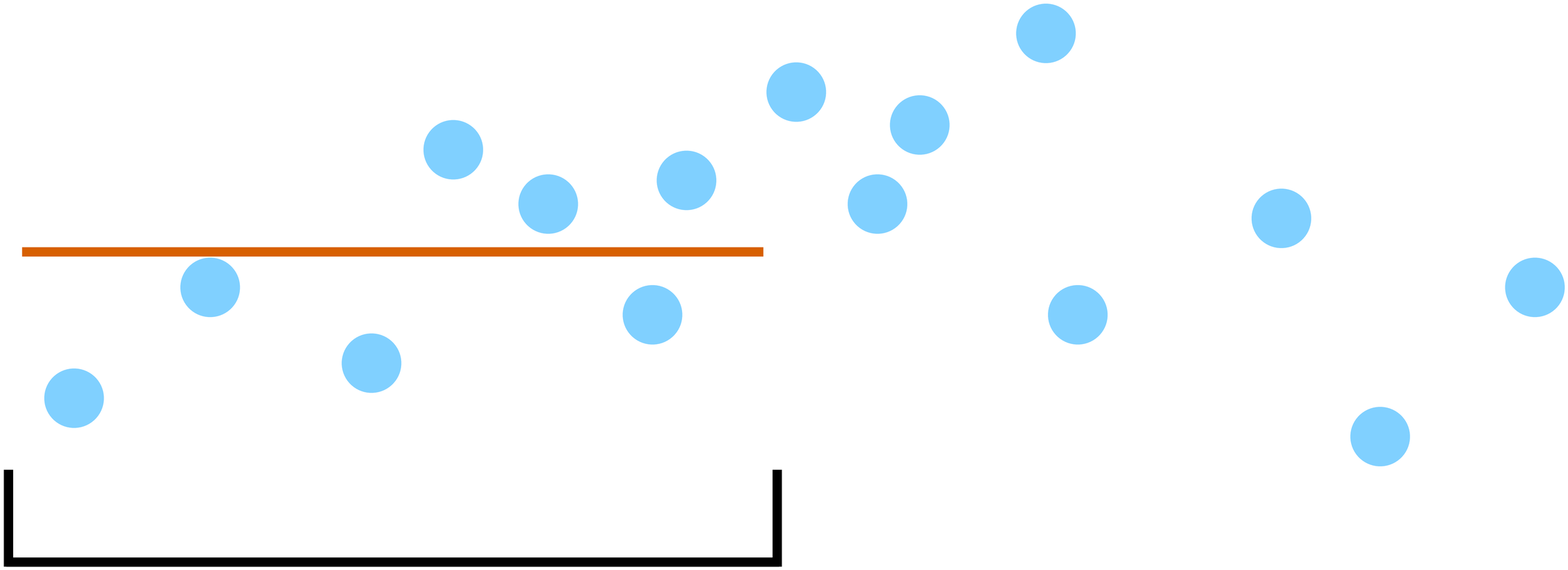
Moving average: a non-parametric approach

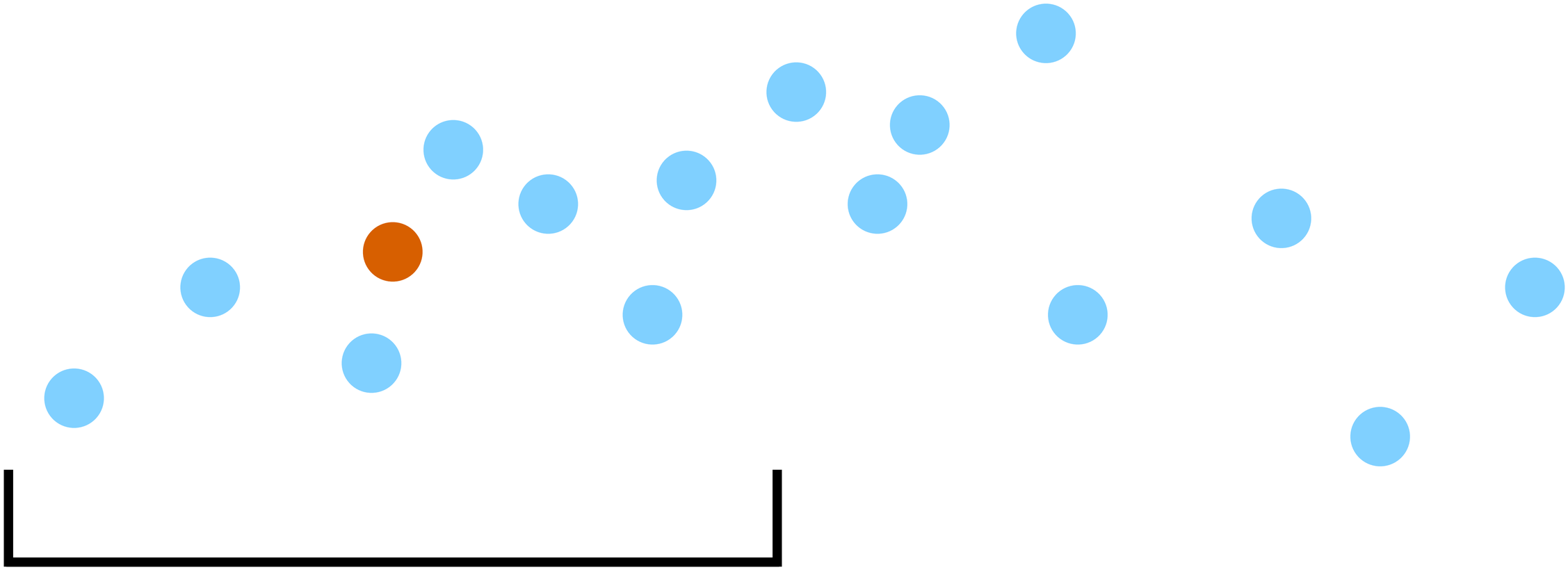


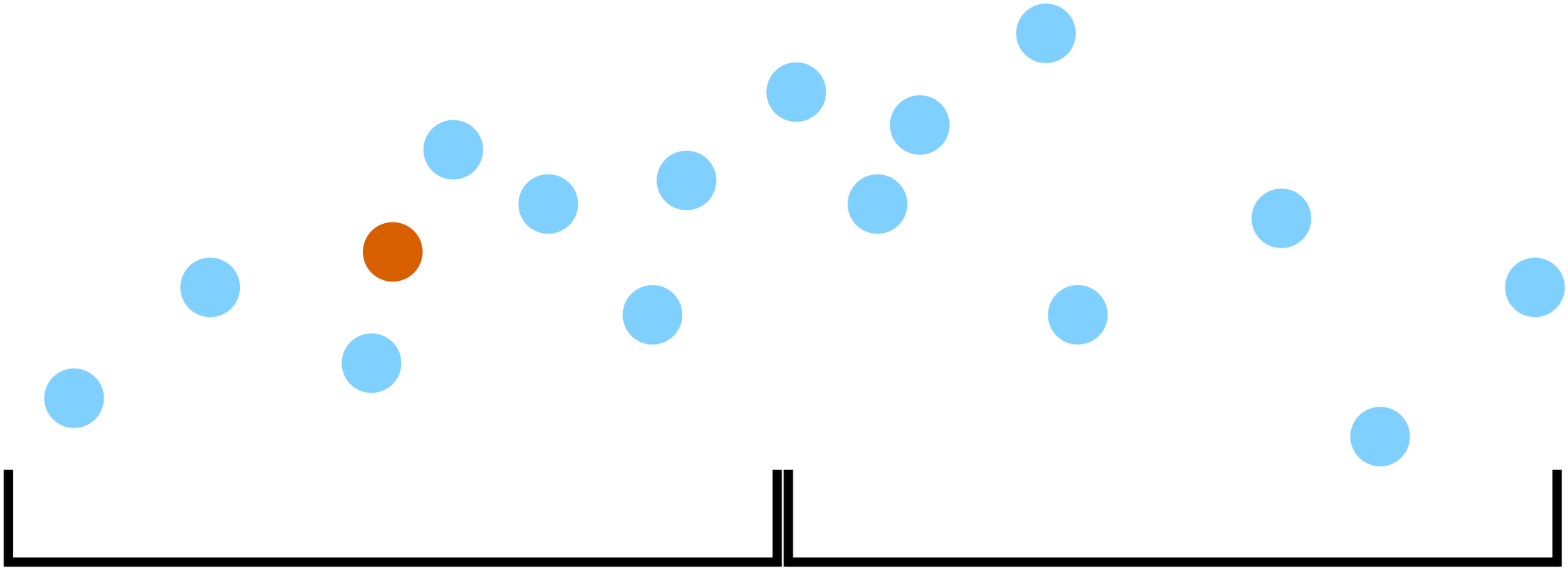
https://stackoverflow.com/questions/11352047/finding-moving-average-from-data-points-in-python

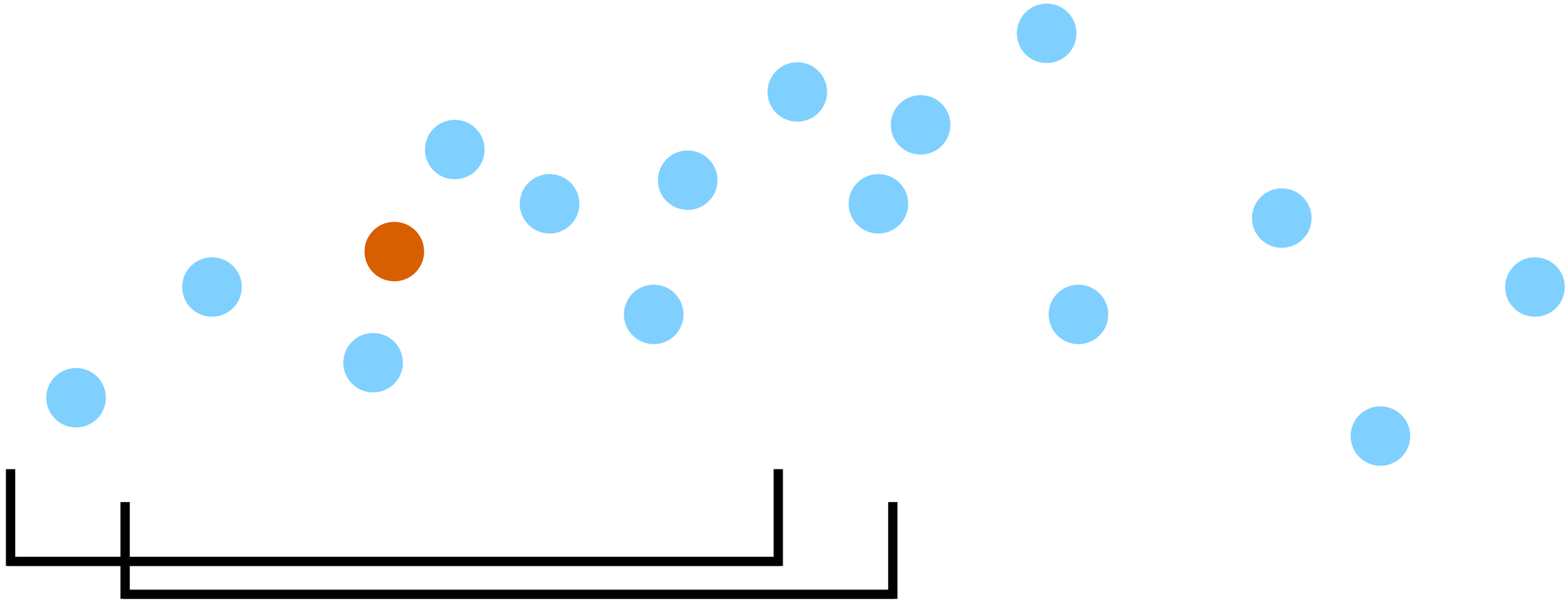


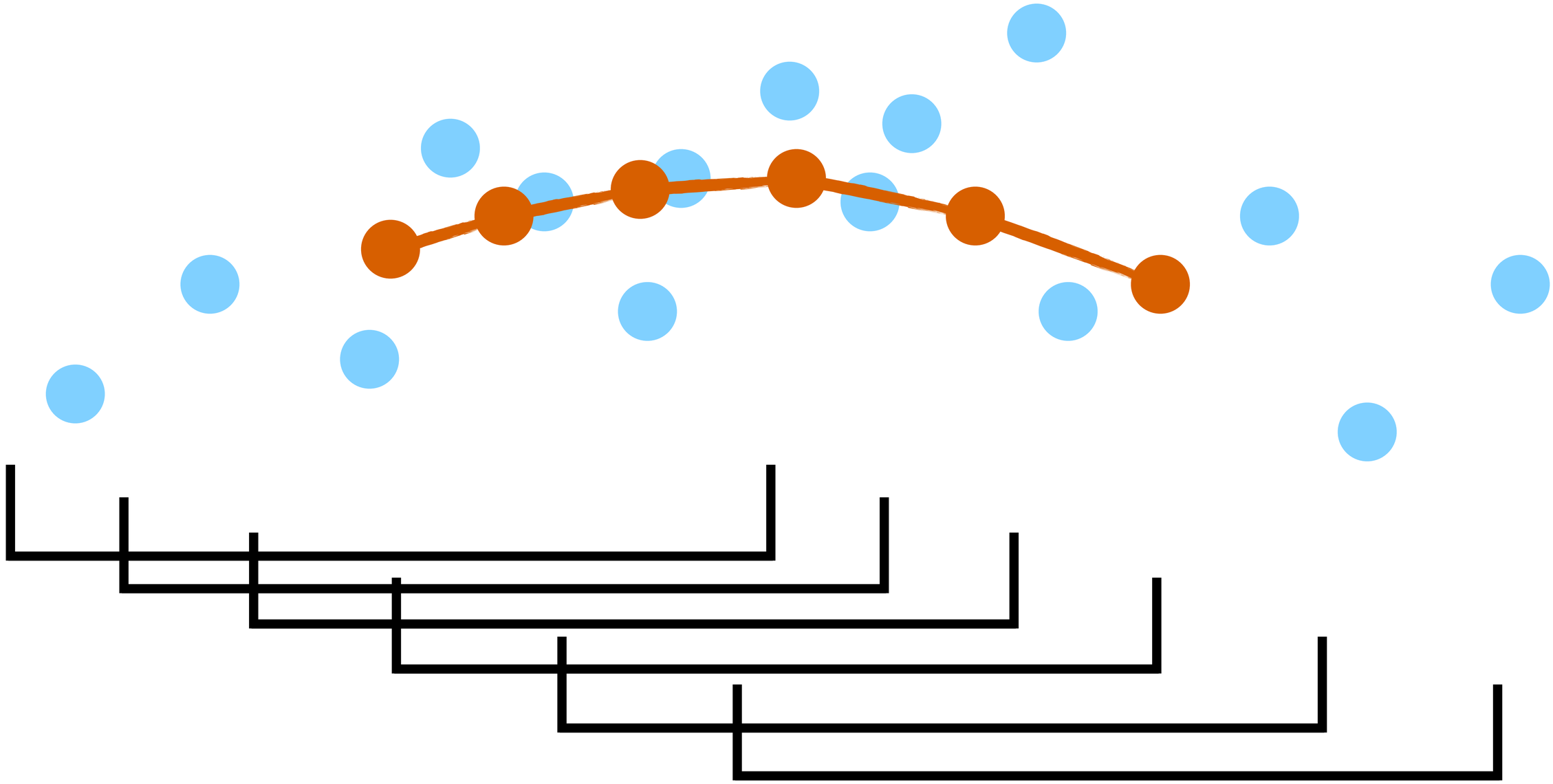








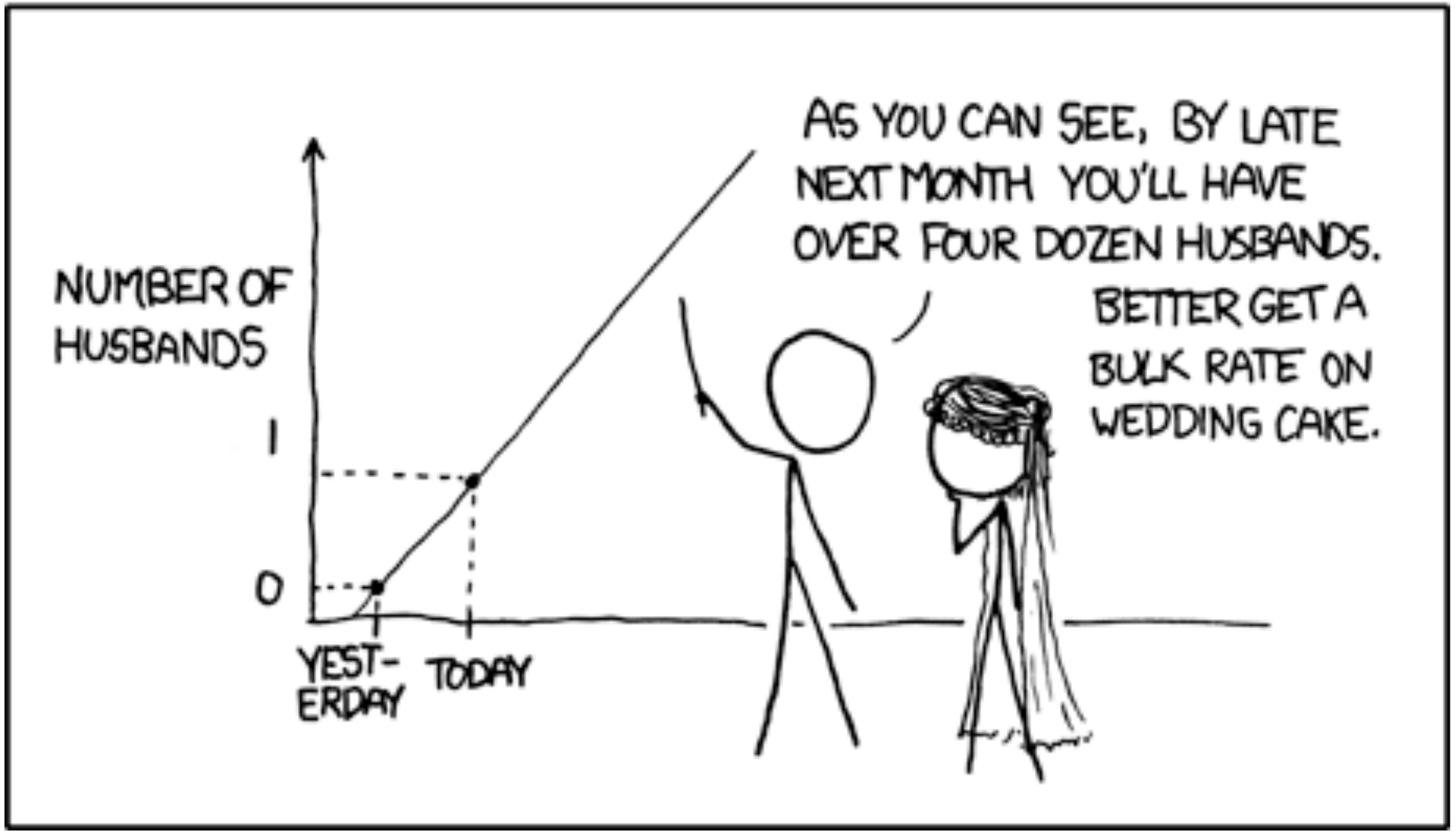




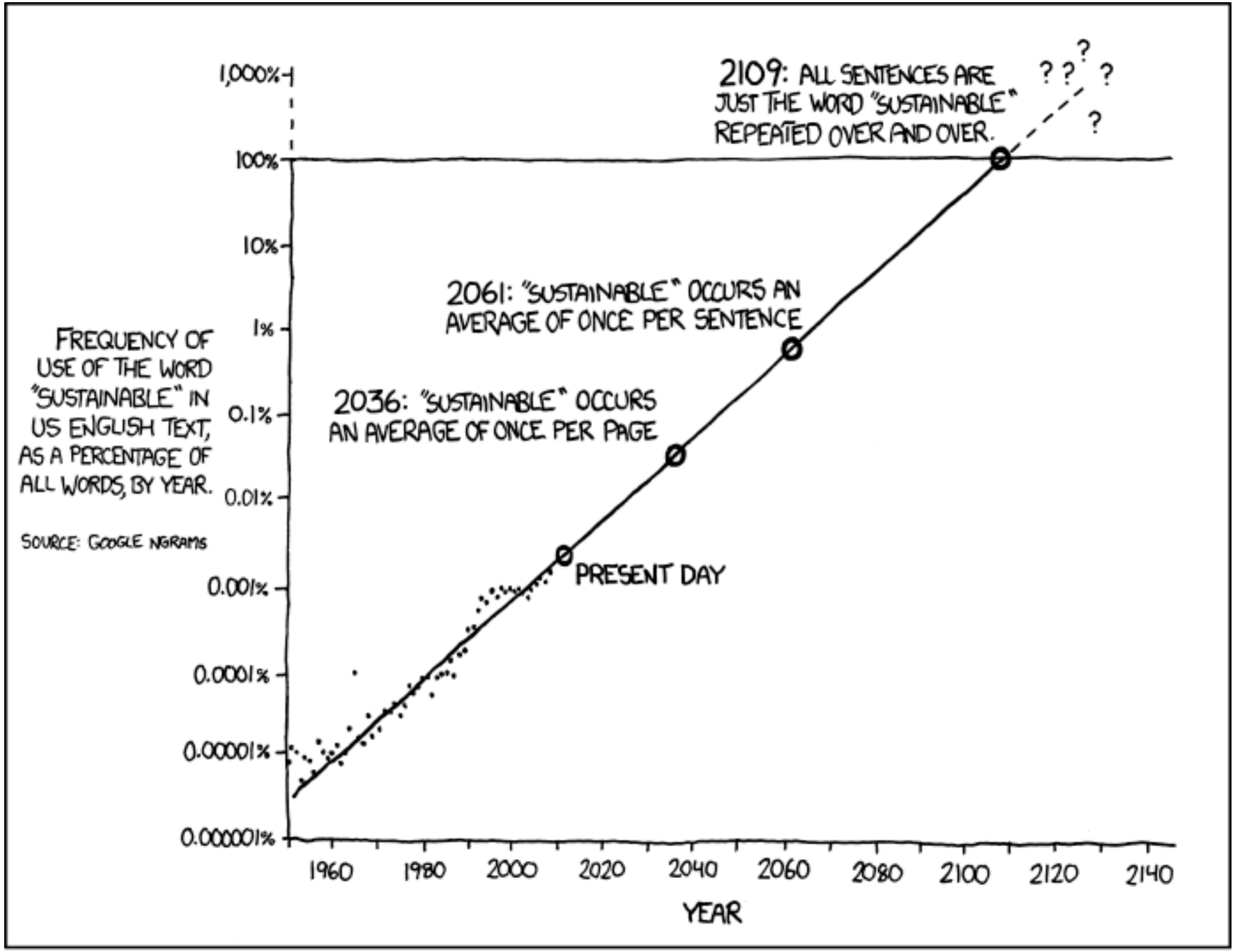
Extrapolation

Let’s extend the trend we know

MY HOBBY: EXTRAPOLATING



xkcd



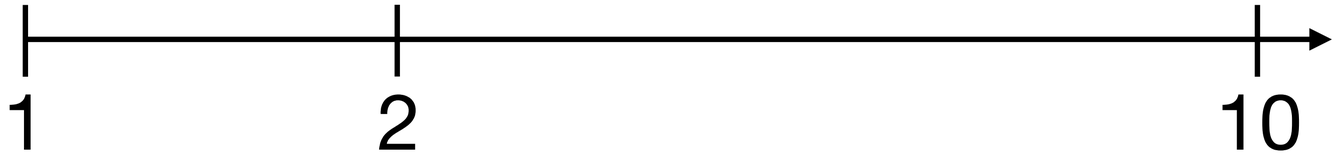
THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

xkcd

Logarithm

Where is 1,000?

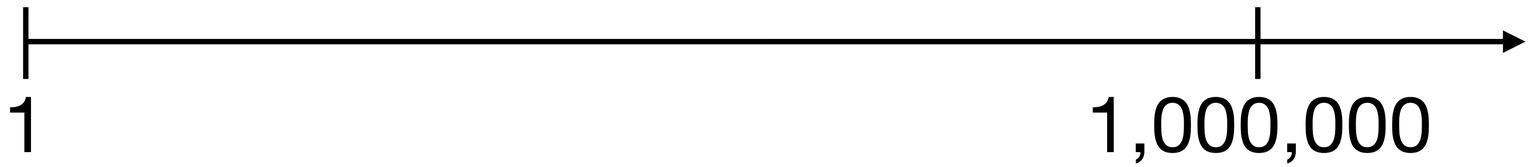
Where is 5?



Where is 1,000?

(if linear scale)

Where is 1,000?



How about in **log scale**?

$$1{,}000 / 1 = 1{,}000$$

1,000 / 1 = 1,000

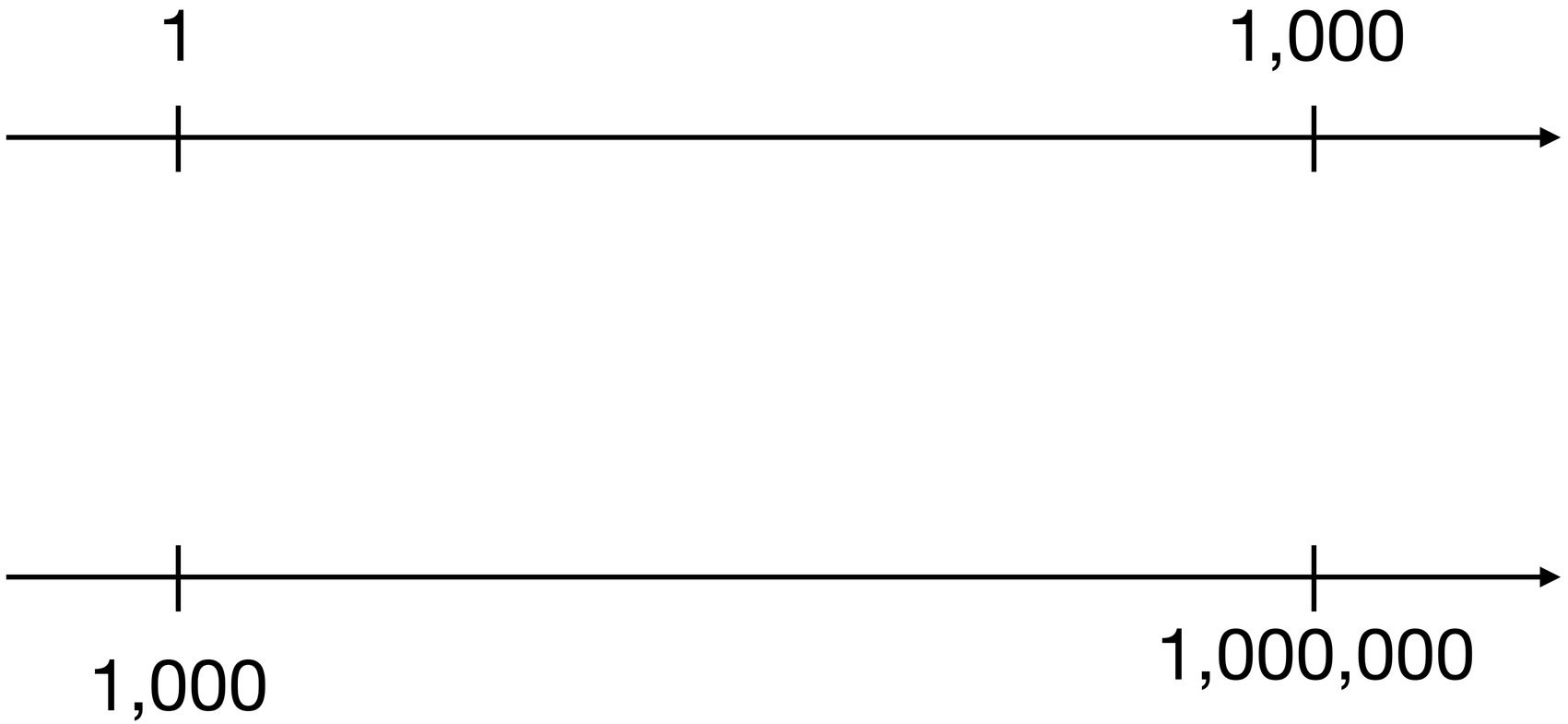
1,000,000 / 1,000 = 1,000

1,000 / 1 = **1,000**

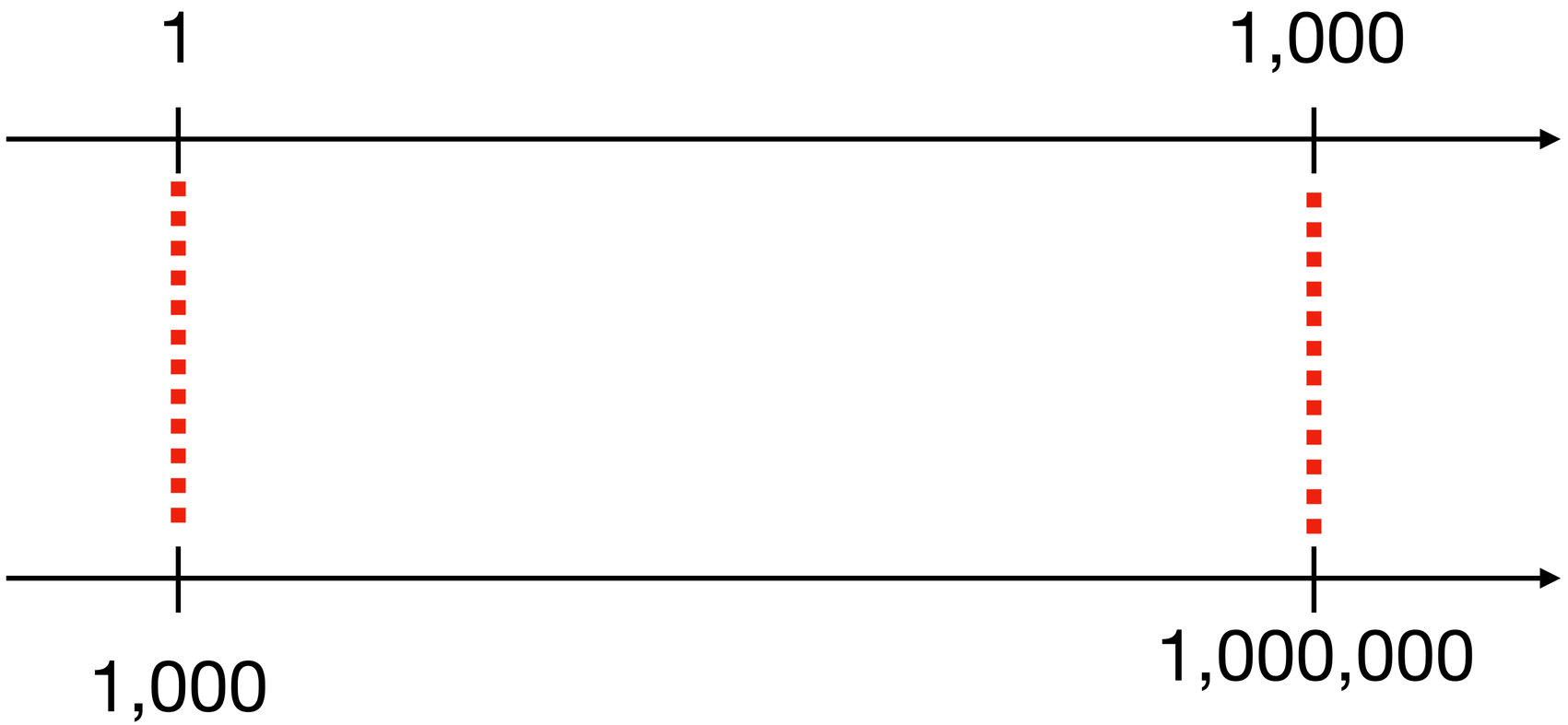
1,000 / 1 = **1,000**

1,000,000 / 1,000 = **1,000**

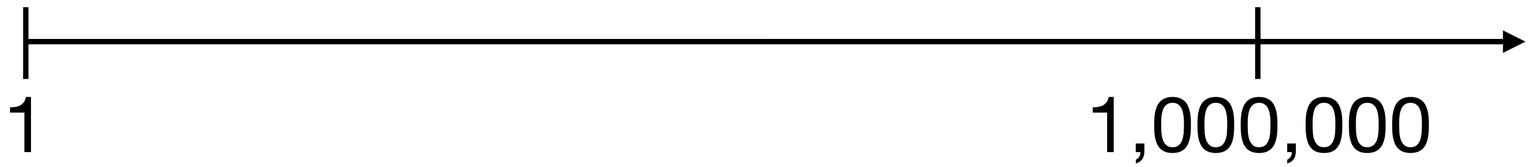
In log-scale,



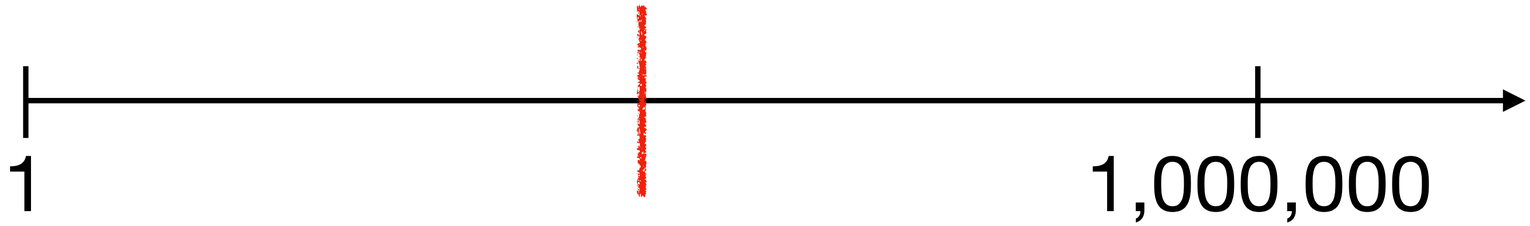
In log-scale,



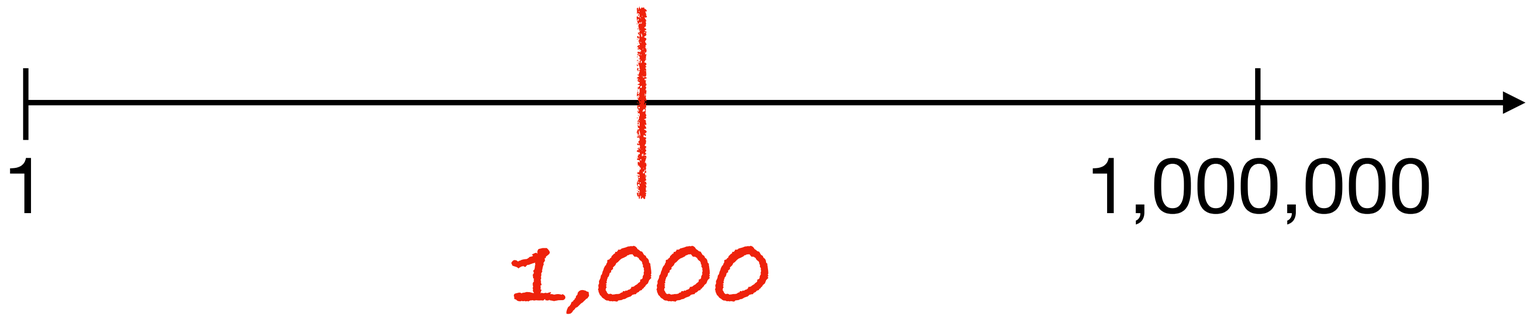
Where is 1,000?



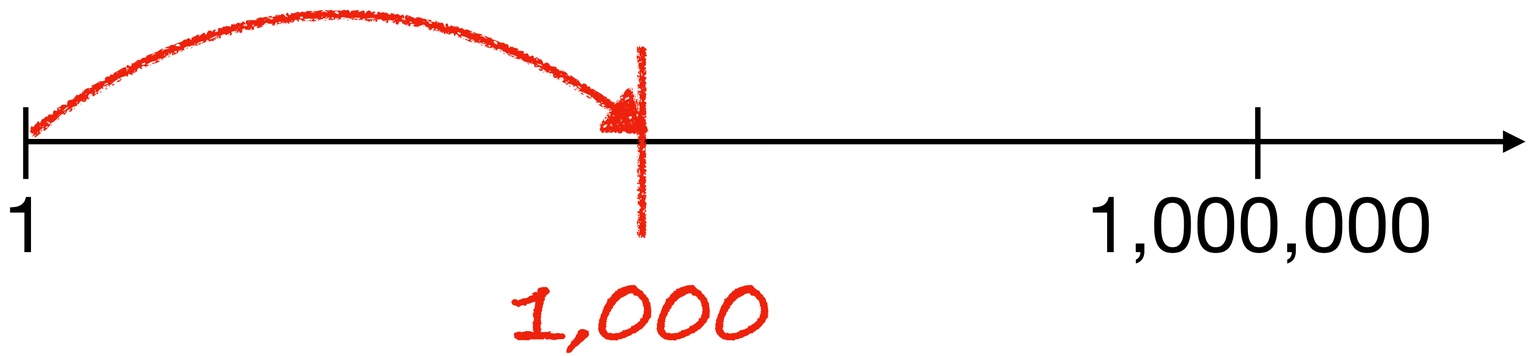
Where is 1,000?



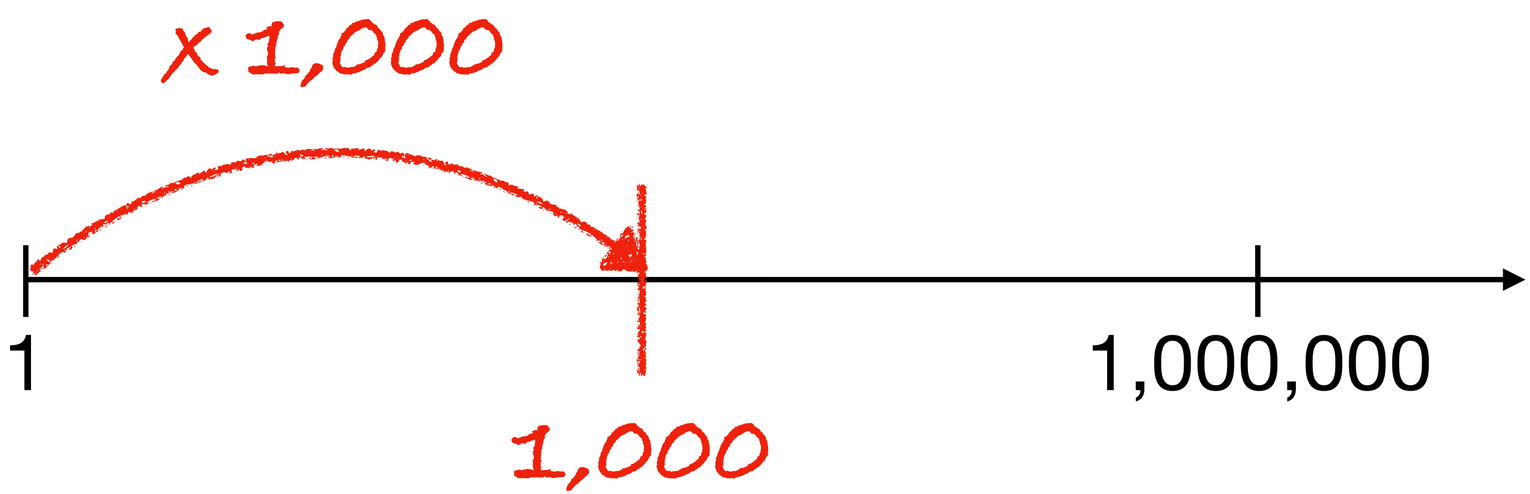
Where is 1,000?



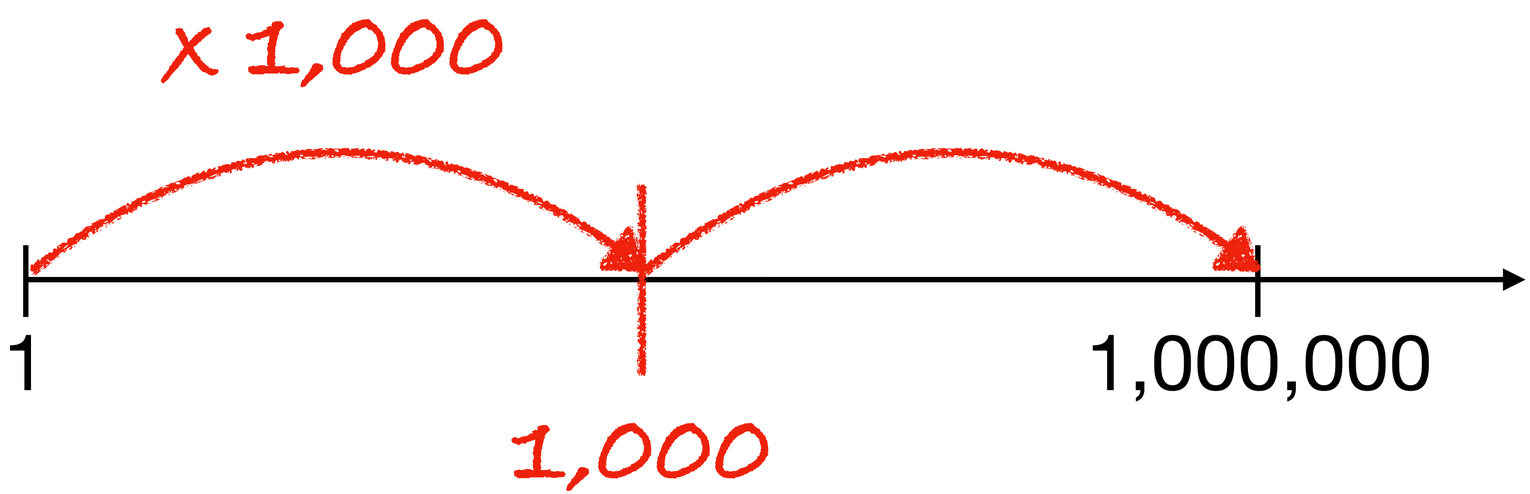
Where is 1,000?



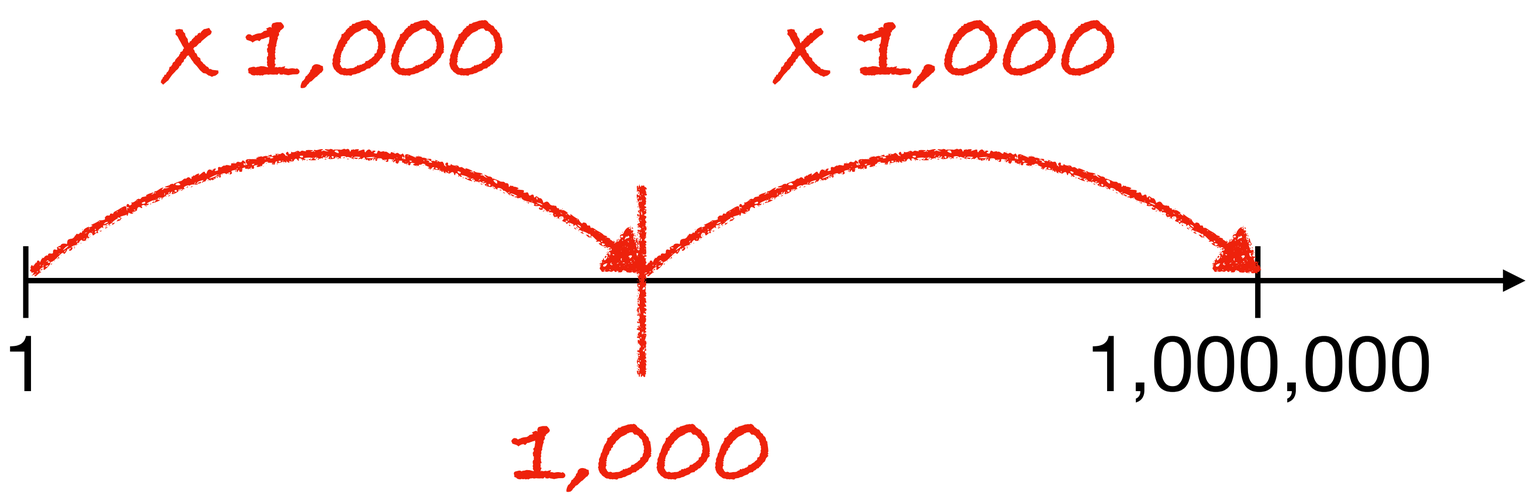
Where is 1,000?



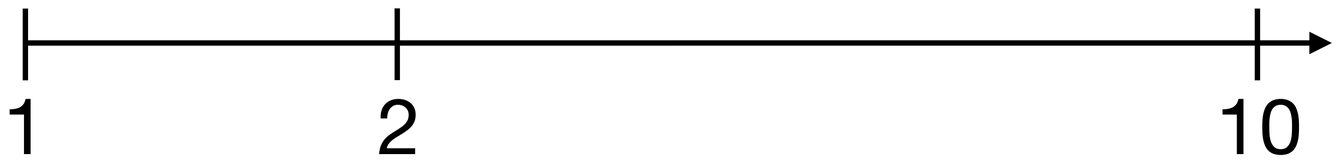
Where is 1,000?



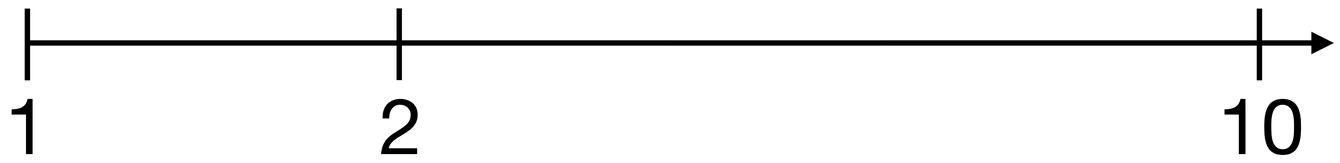
Where is 1,000?



Where is 5?

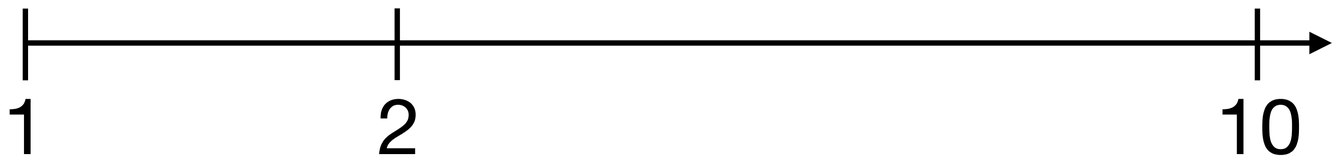


Where is 5?



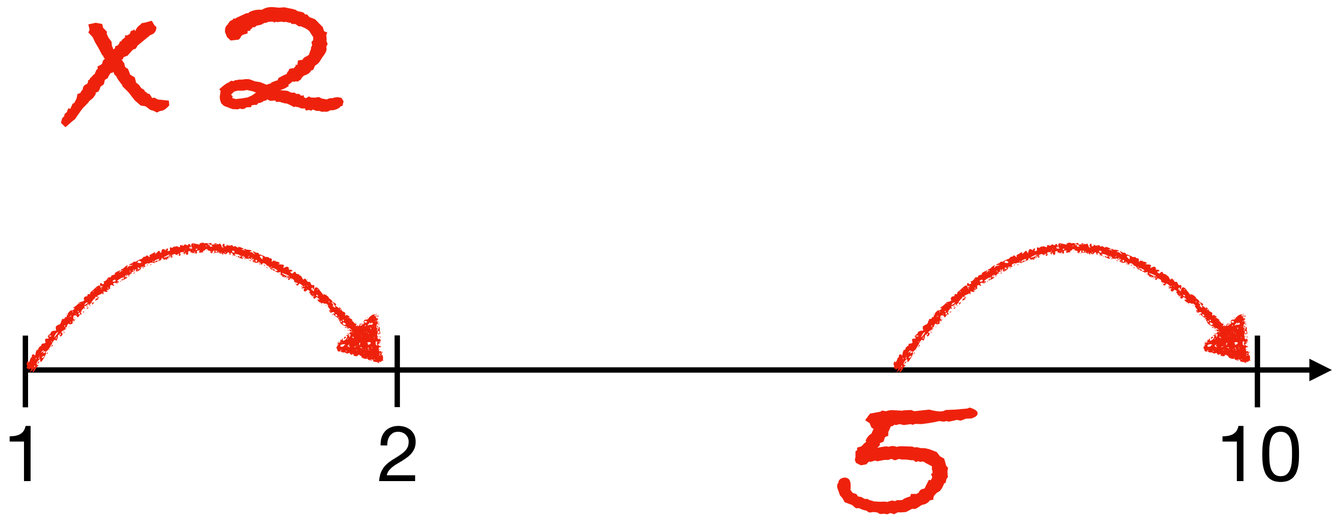
Where is 5?

x 2

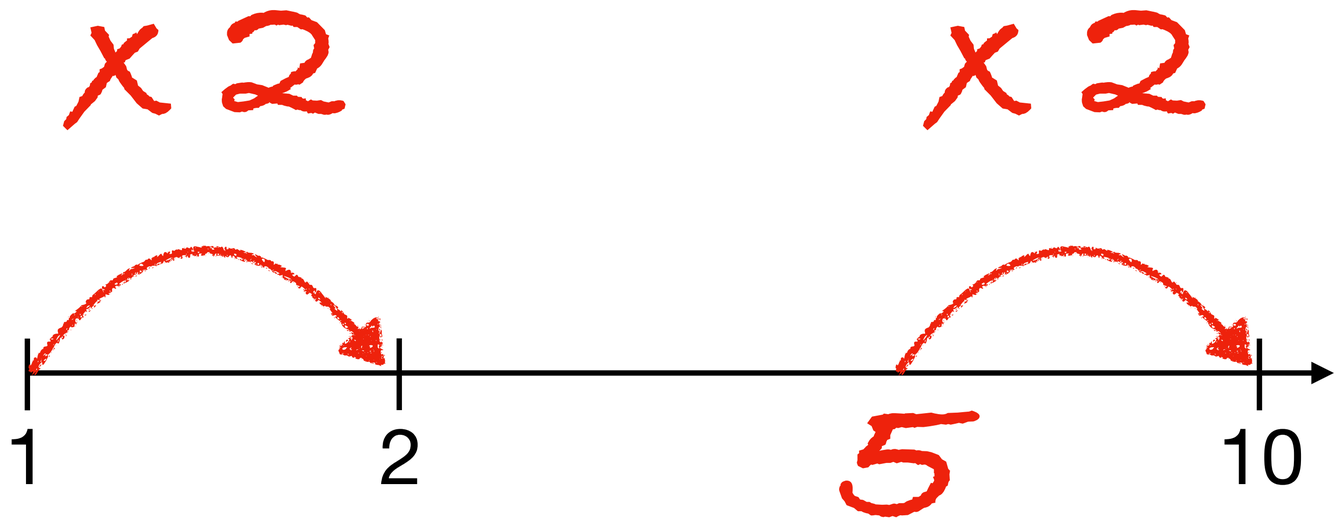


Where is 5?

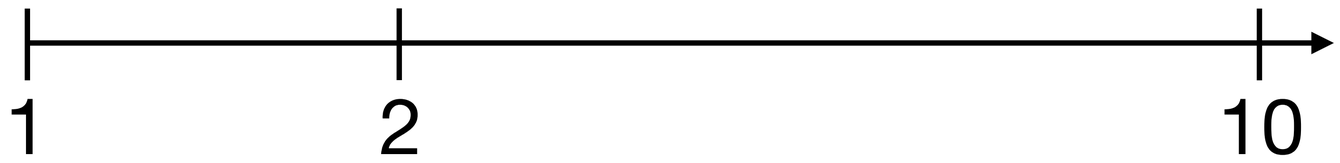
Where is 5?



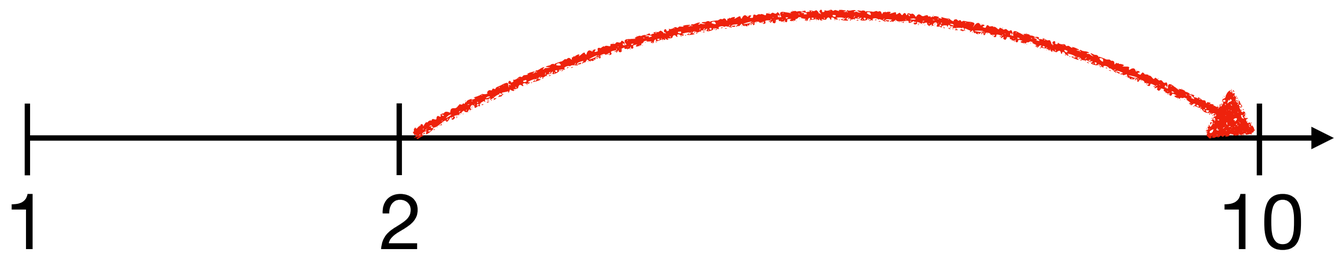
Where is 5?



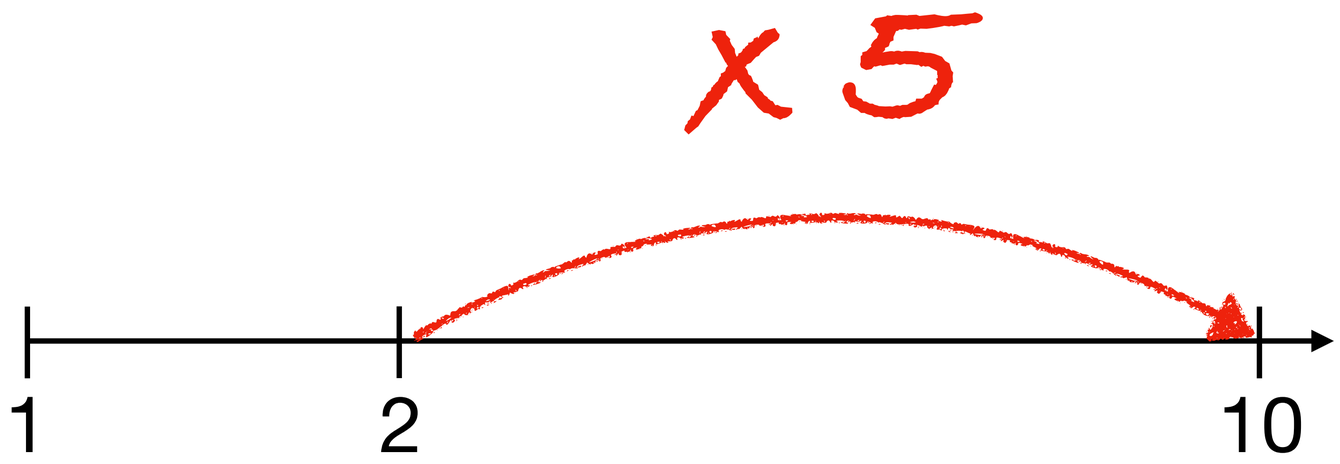
Where is 5?



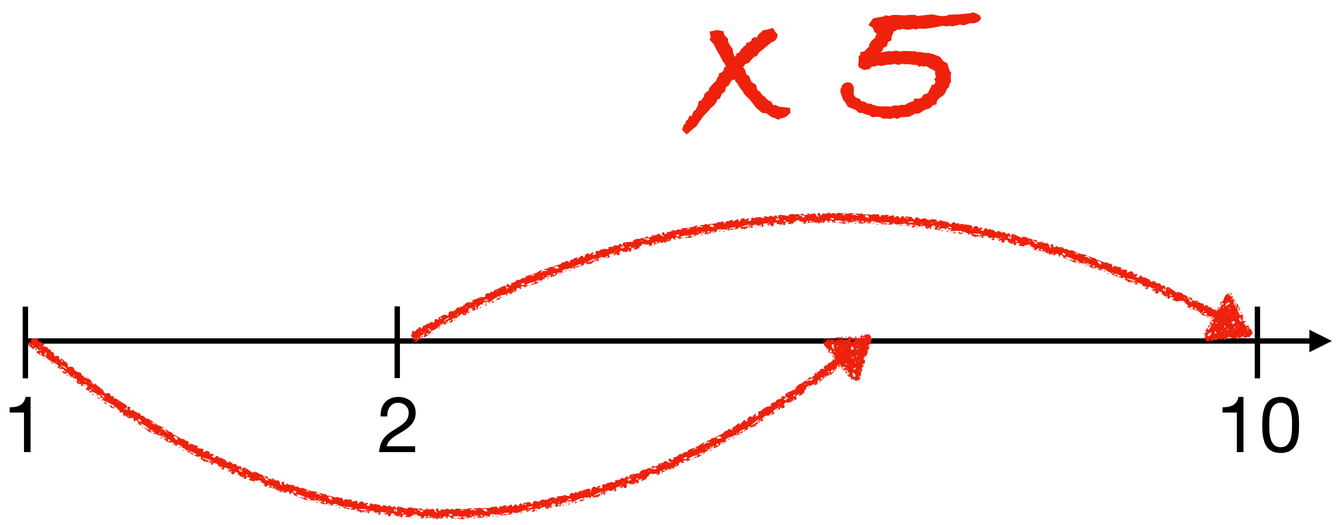
Where is 5?



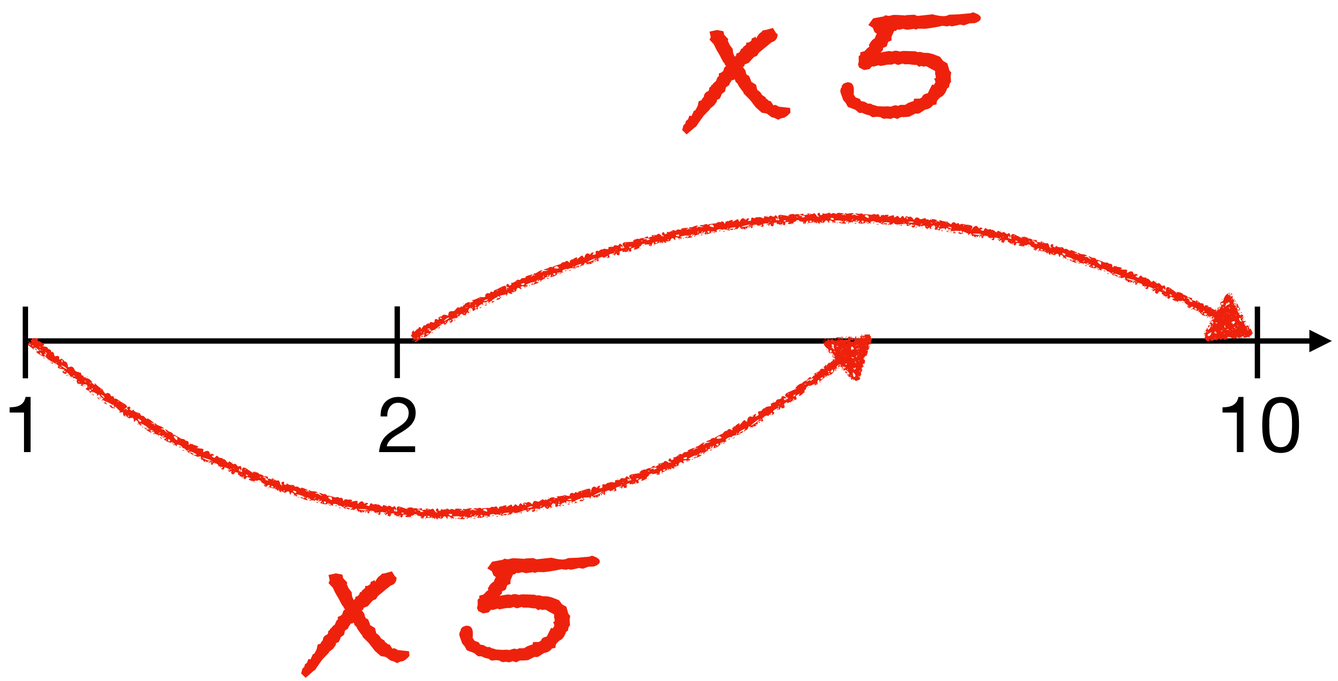
Where is 5?



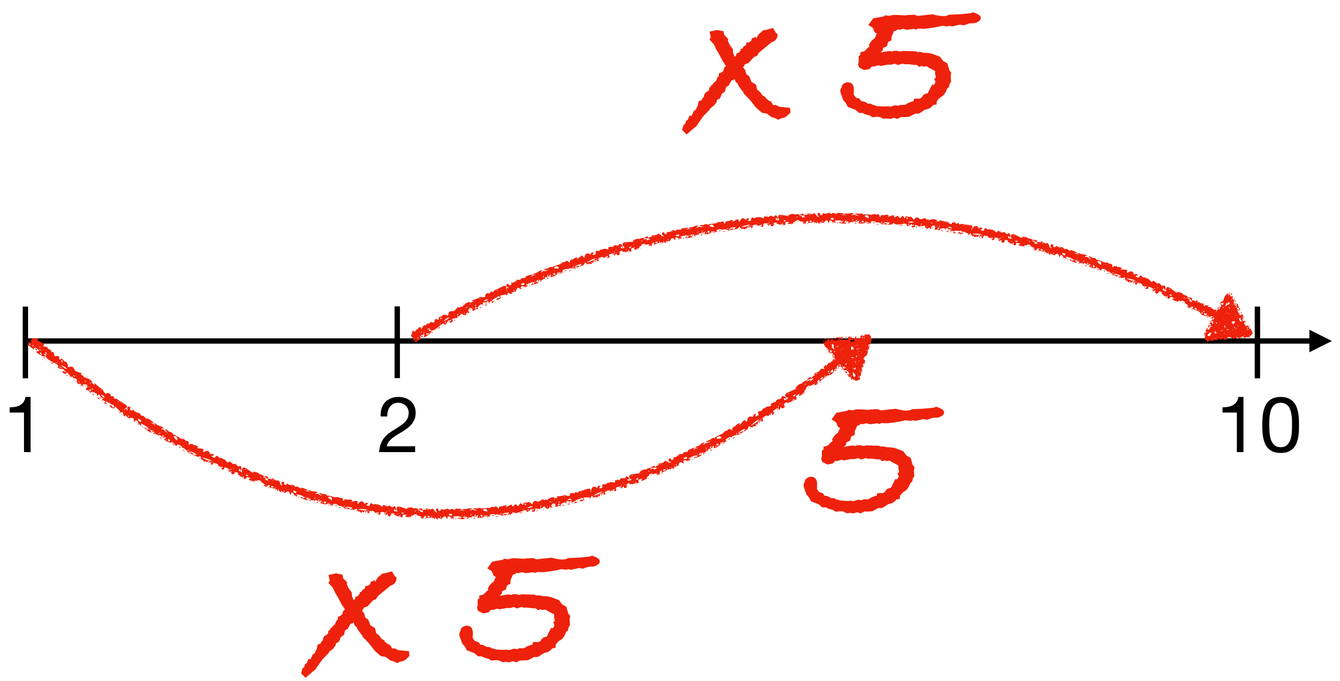
Where is 5?



Where is 5?



Where is 5?



Distance ~ multiplication

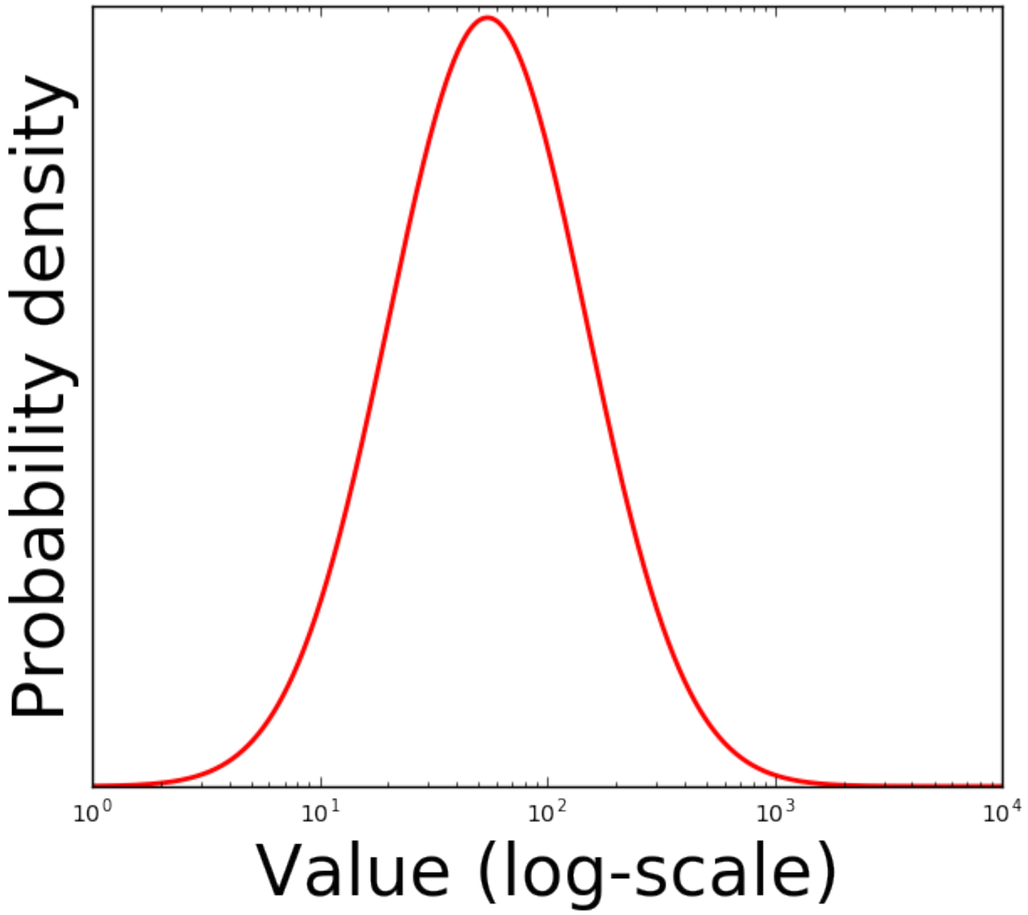
Why logscale?

Many datasets span
many orders of magnitude.

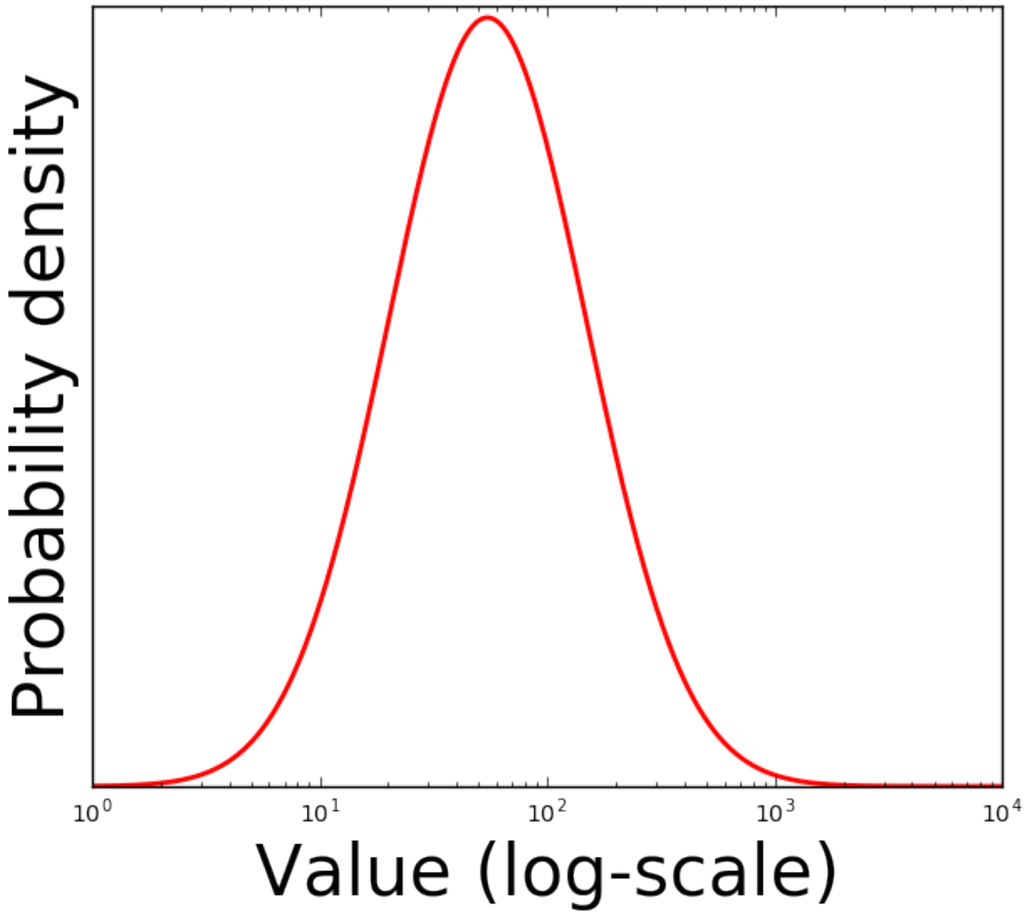
Multiplicative process in linear scale

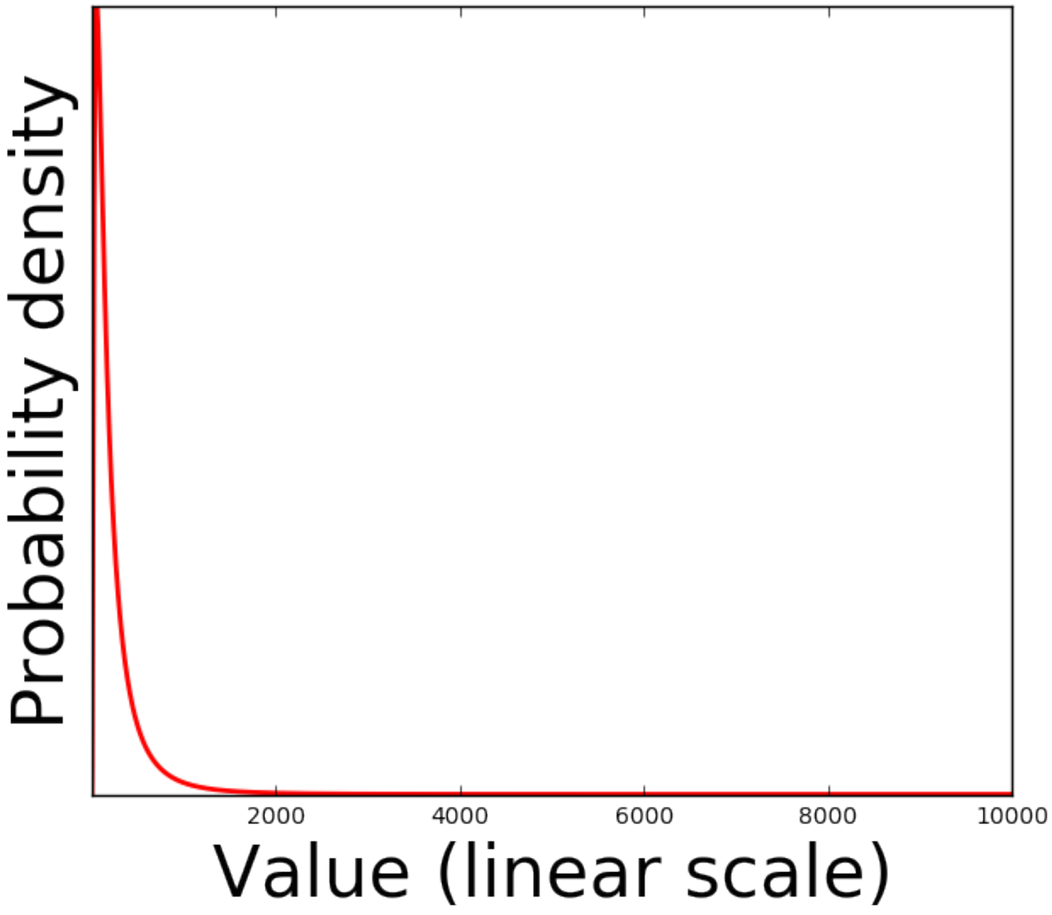
= Additive process in log-scale

Log-normal distribution



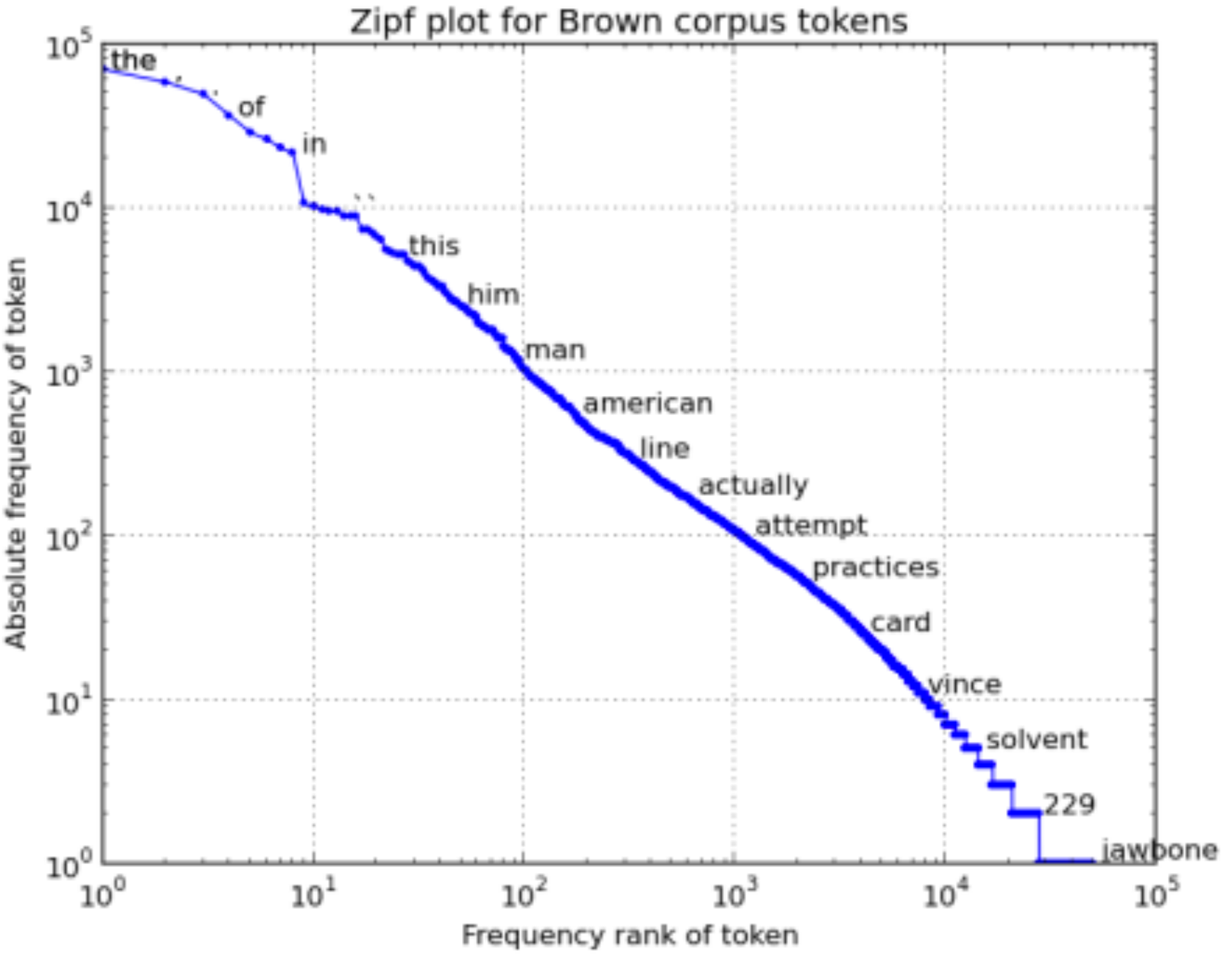
Log-normal distribution



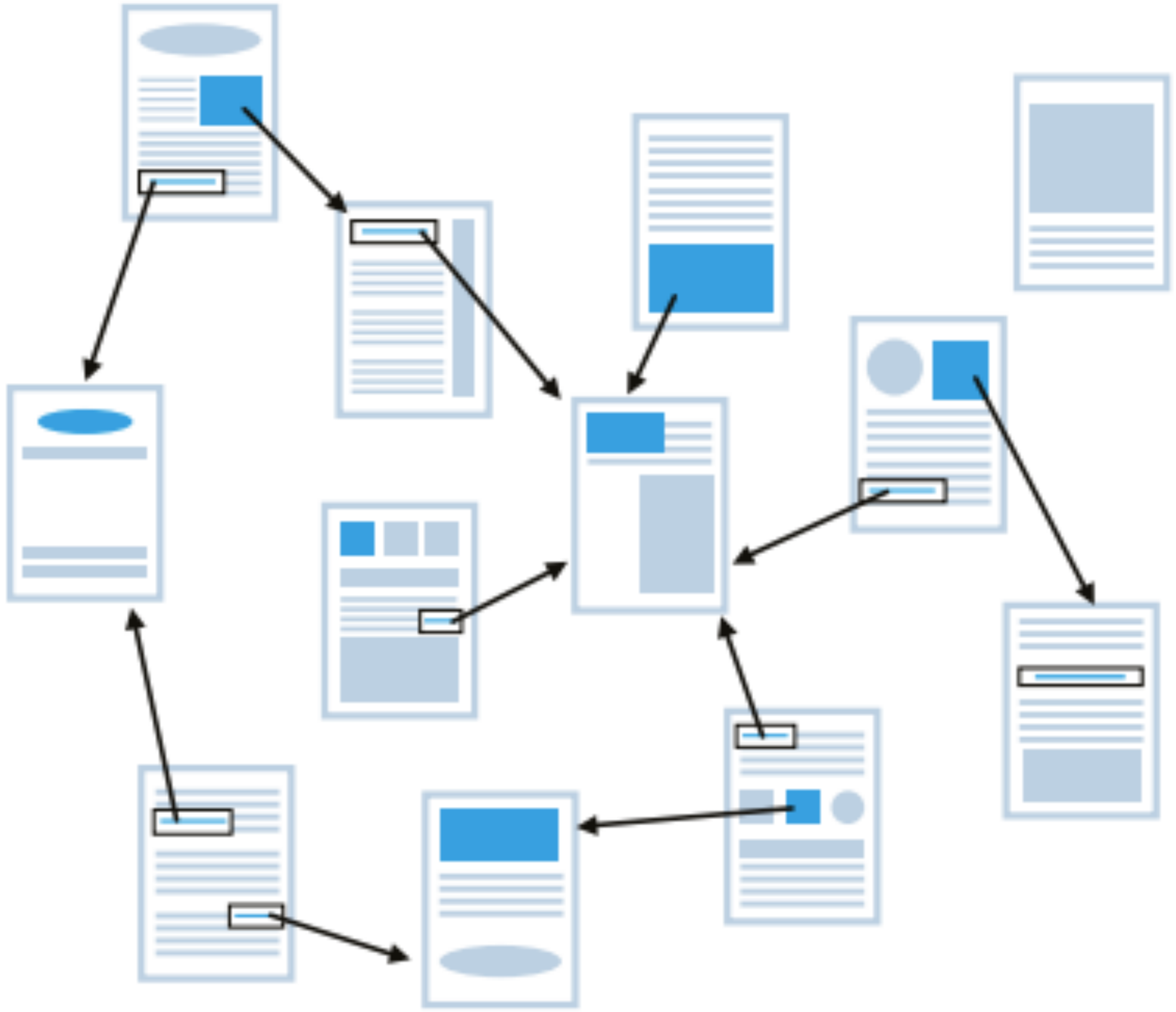


'Power-law' distributions

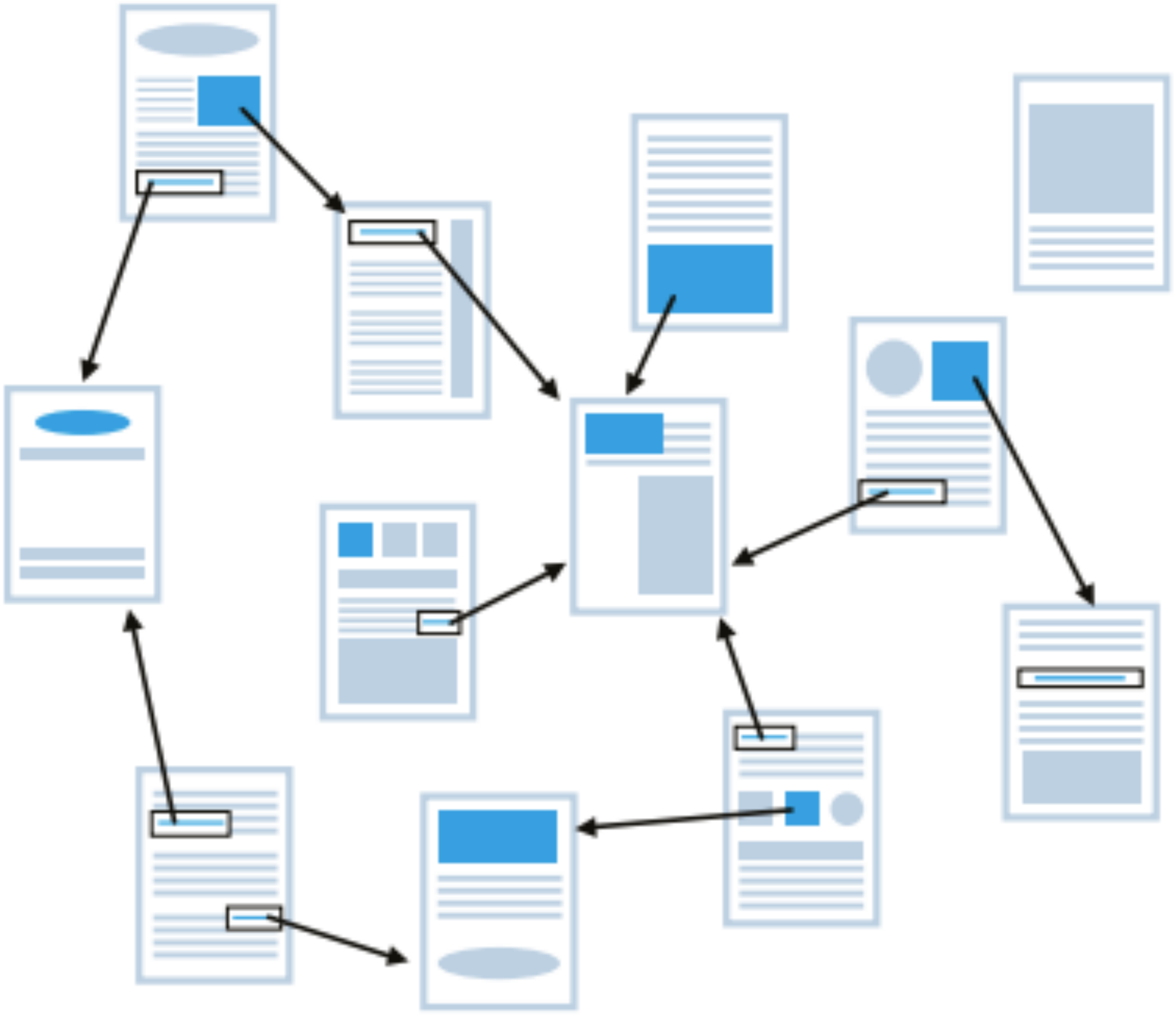
$$f(x) = cx^{-\alpha}$$



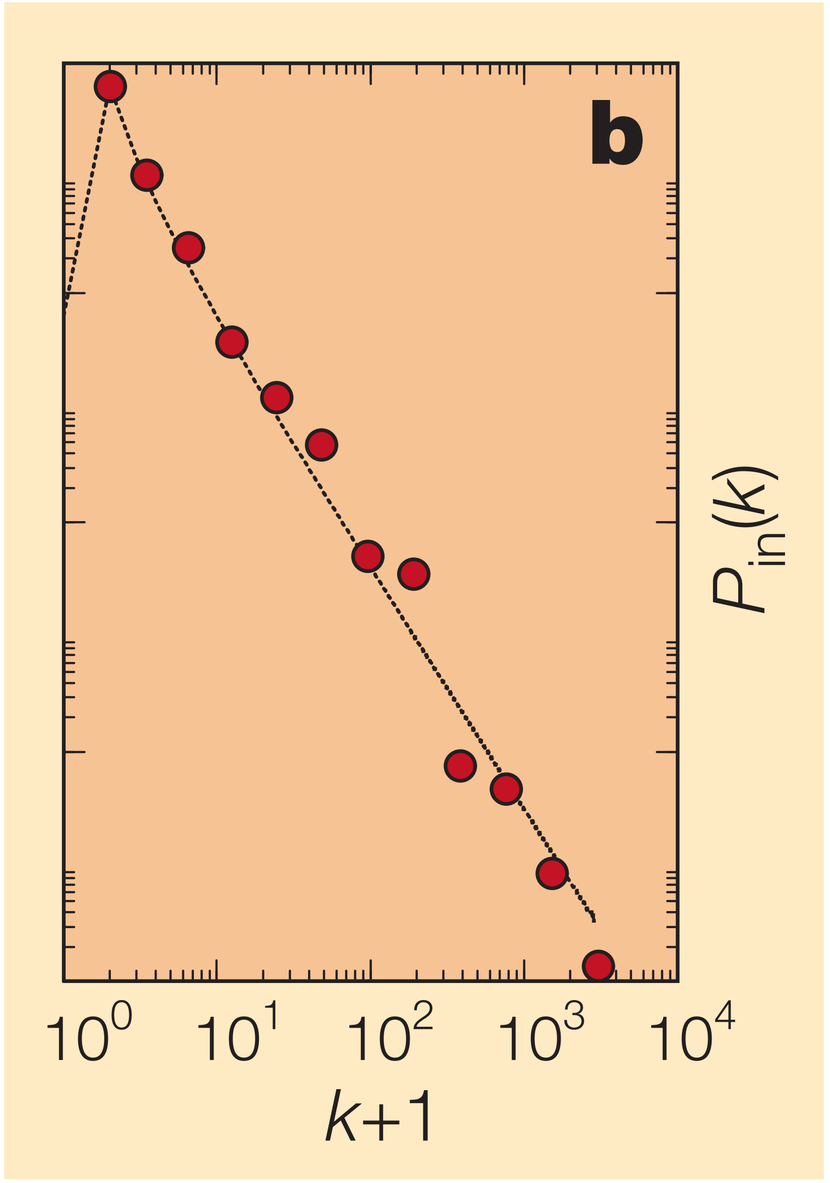
https://finnaarupnielsen.wordpress.com/2013/10/22/zipf-plot-for-word-counts-in-brown-corpus/



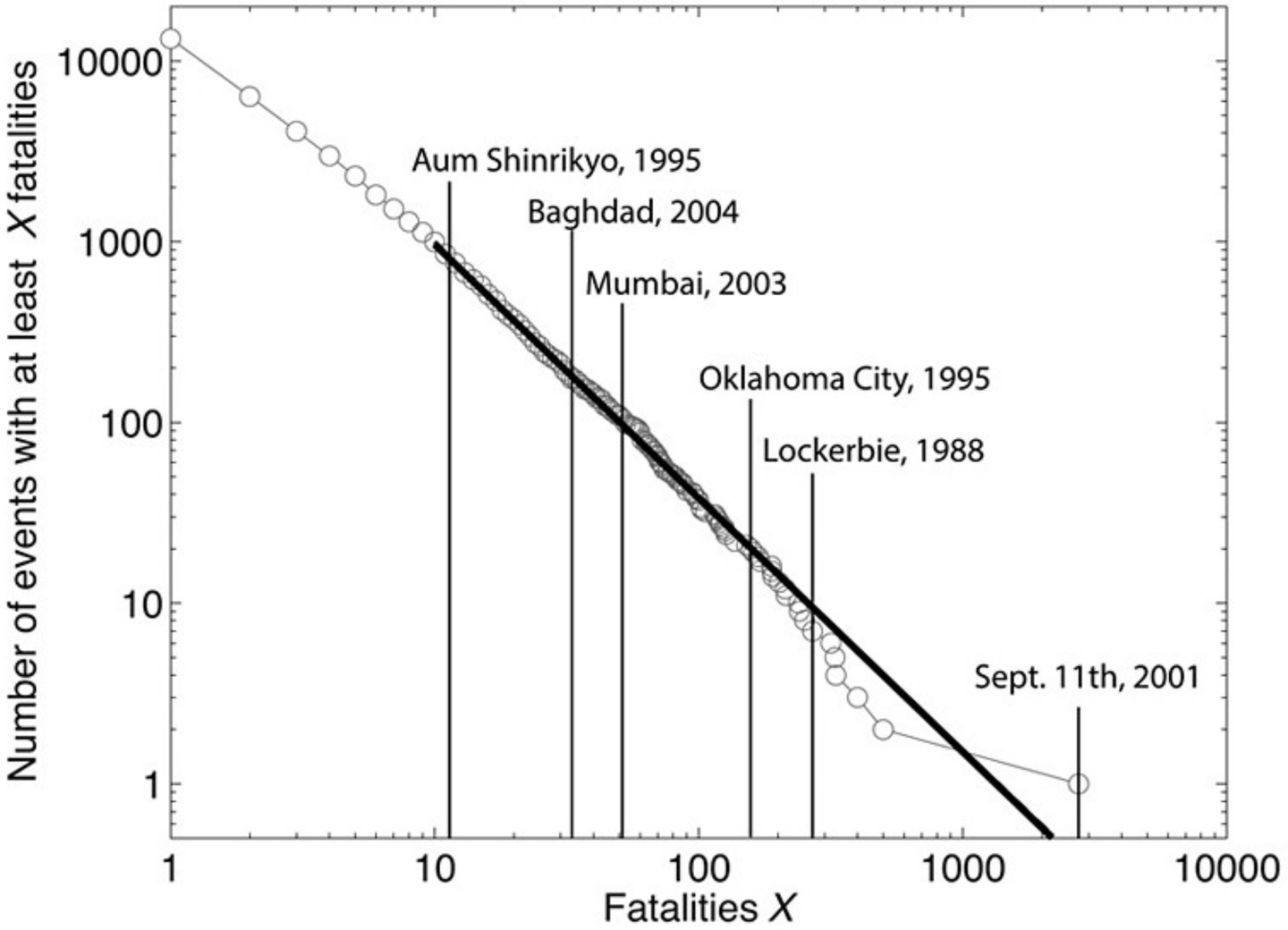
https://barefootcas.org.uk

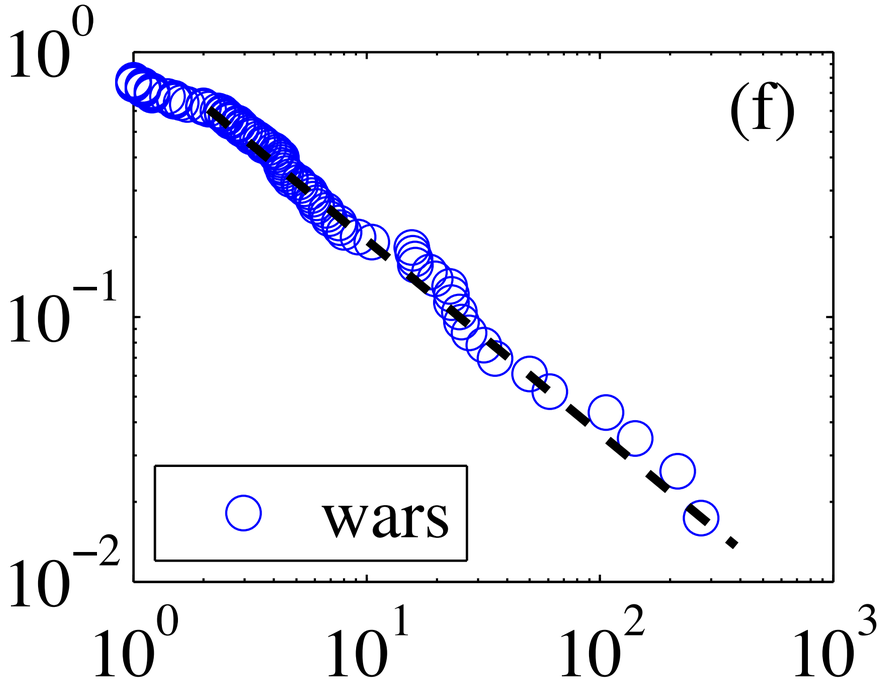


https://barefootcas.org.uk

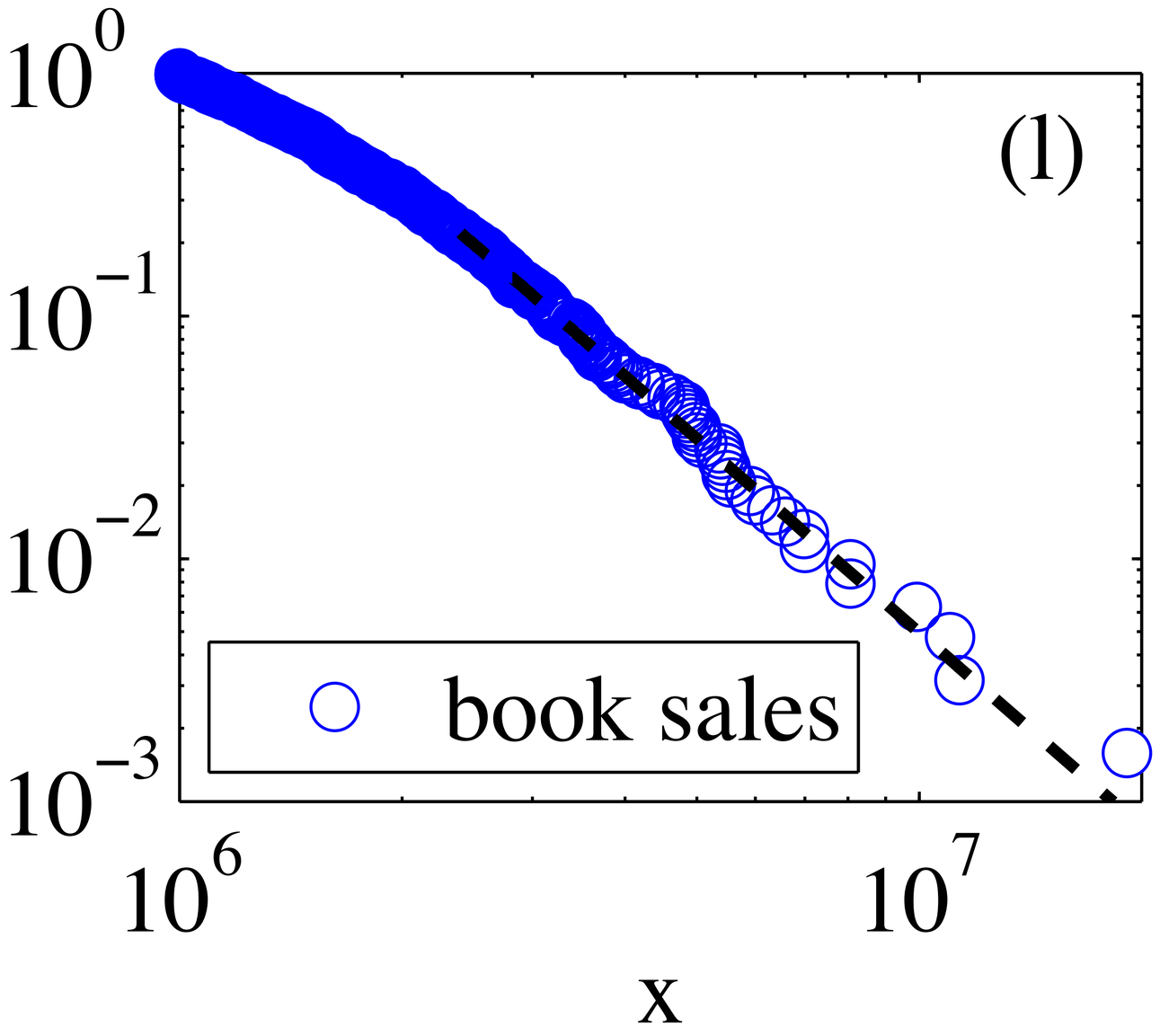


Albert, Jeong, Barabási, *Nature* **401**, 130 (1999)

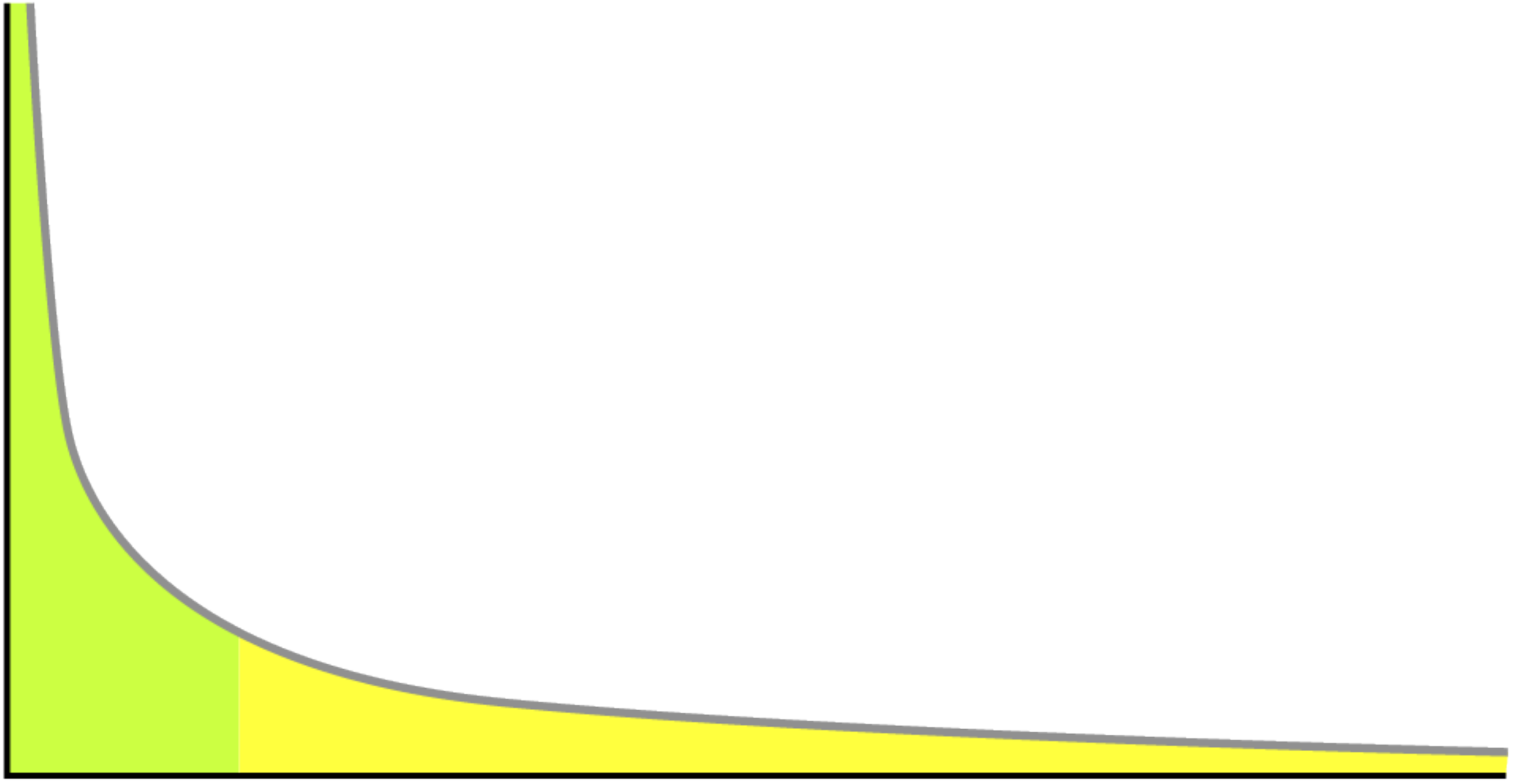




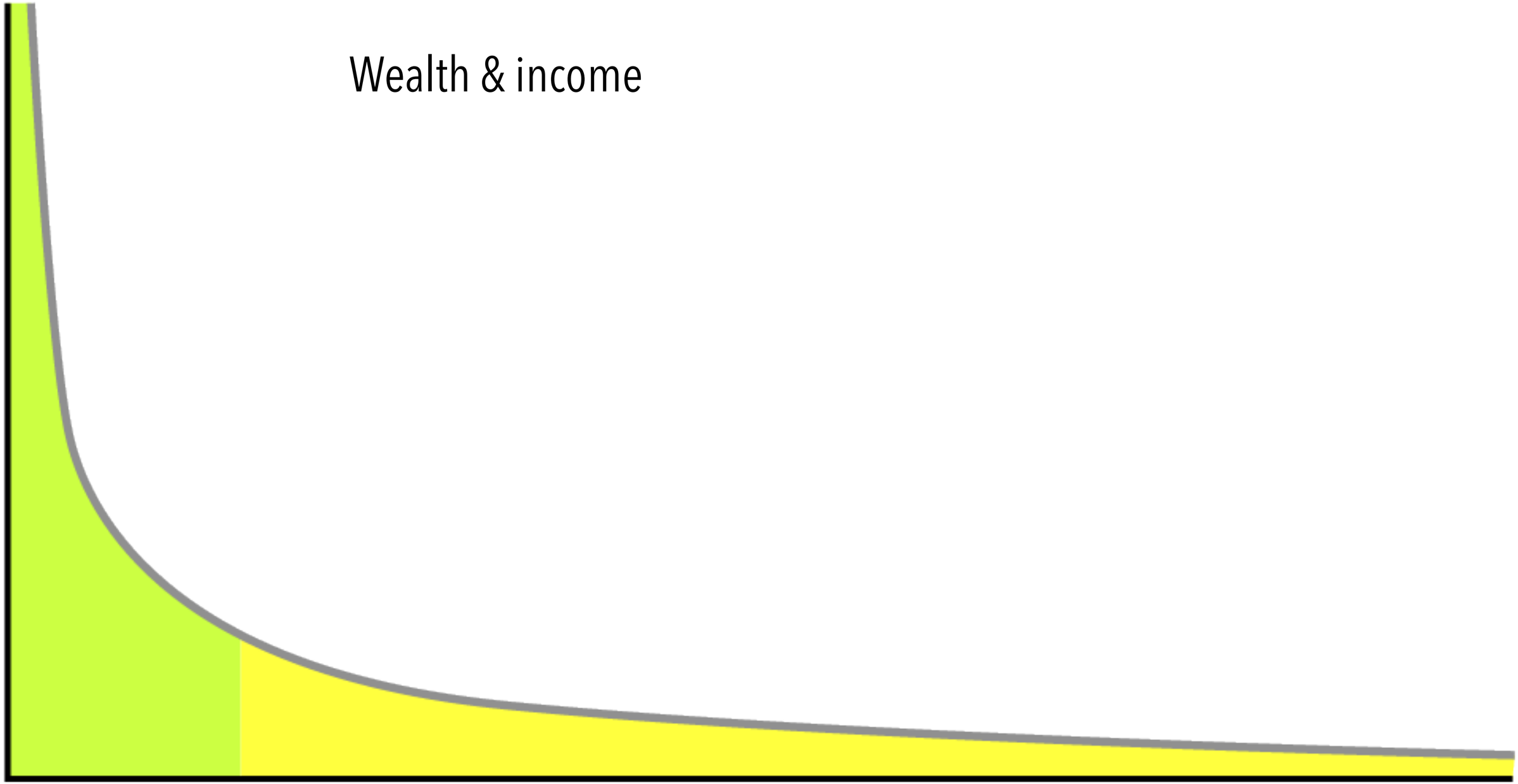
Clauset, Shalizi, Newman, *SIAM Review* **51**, 661 (2009)

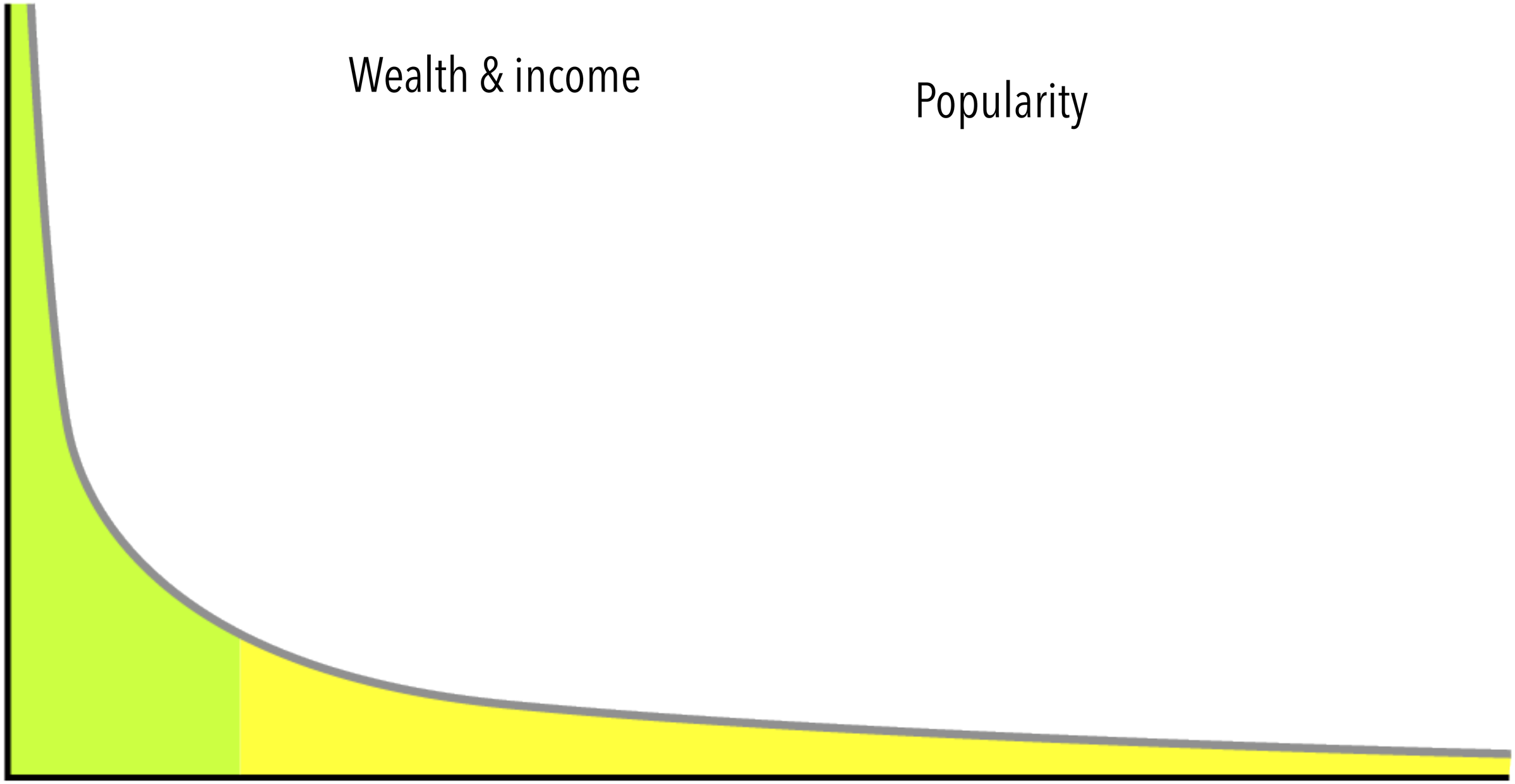


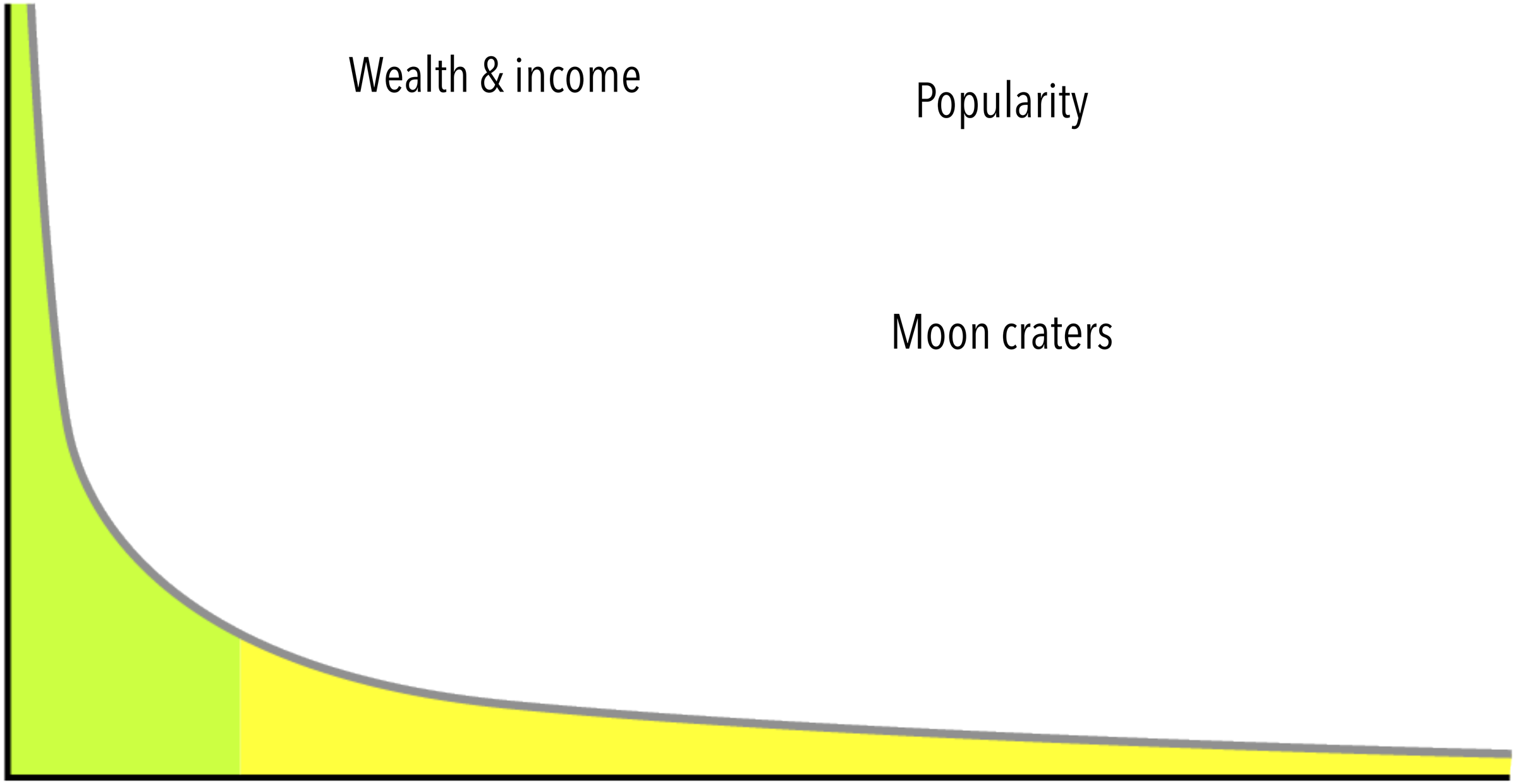
Clauset, Shalizi, Newman, *SIAM Review* **51**, 661 (2009)

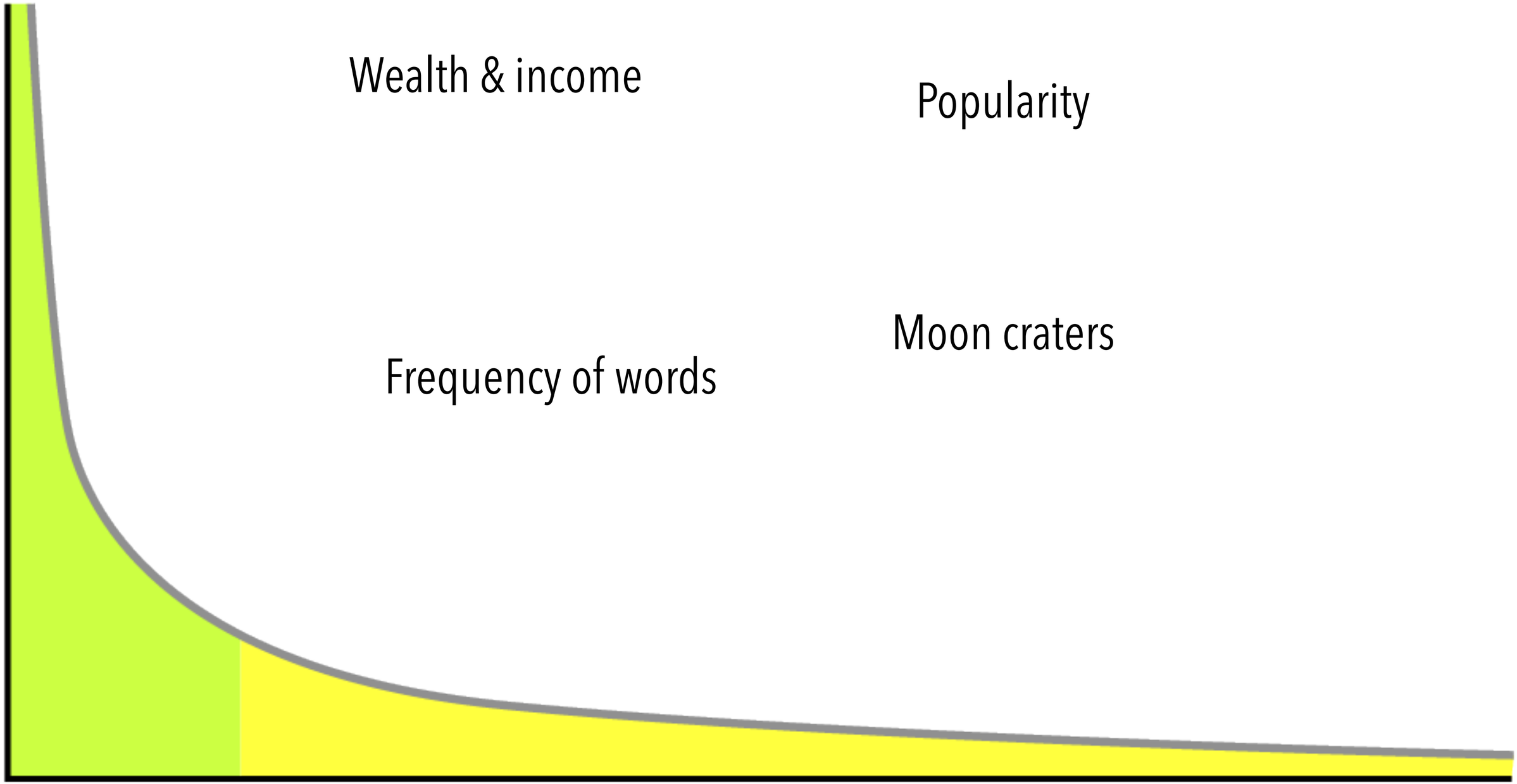


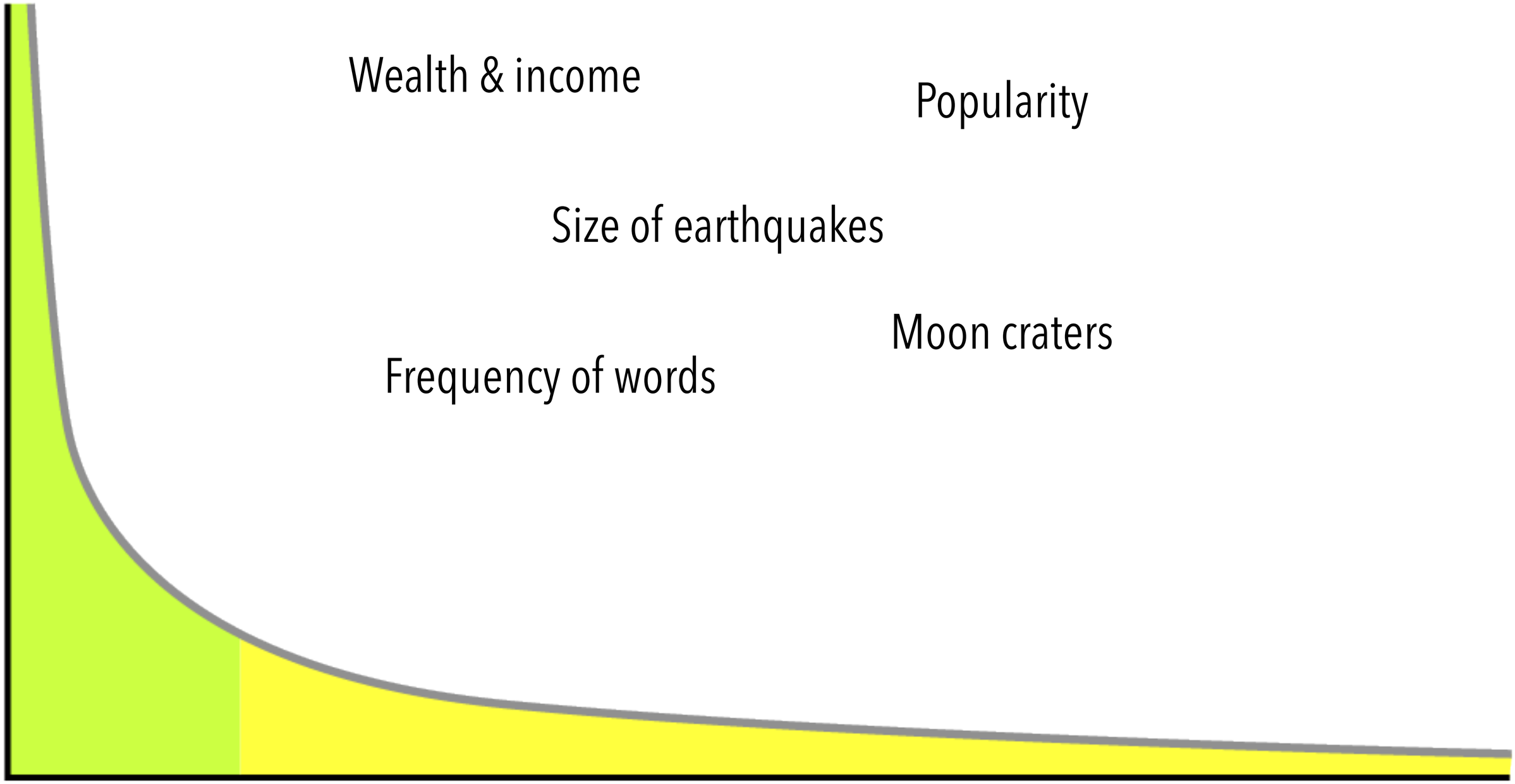
Wealth & income

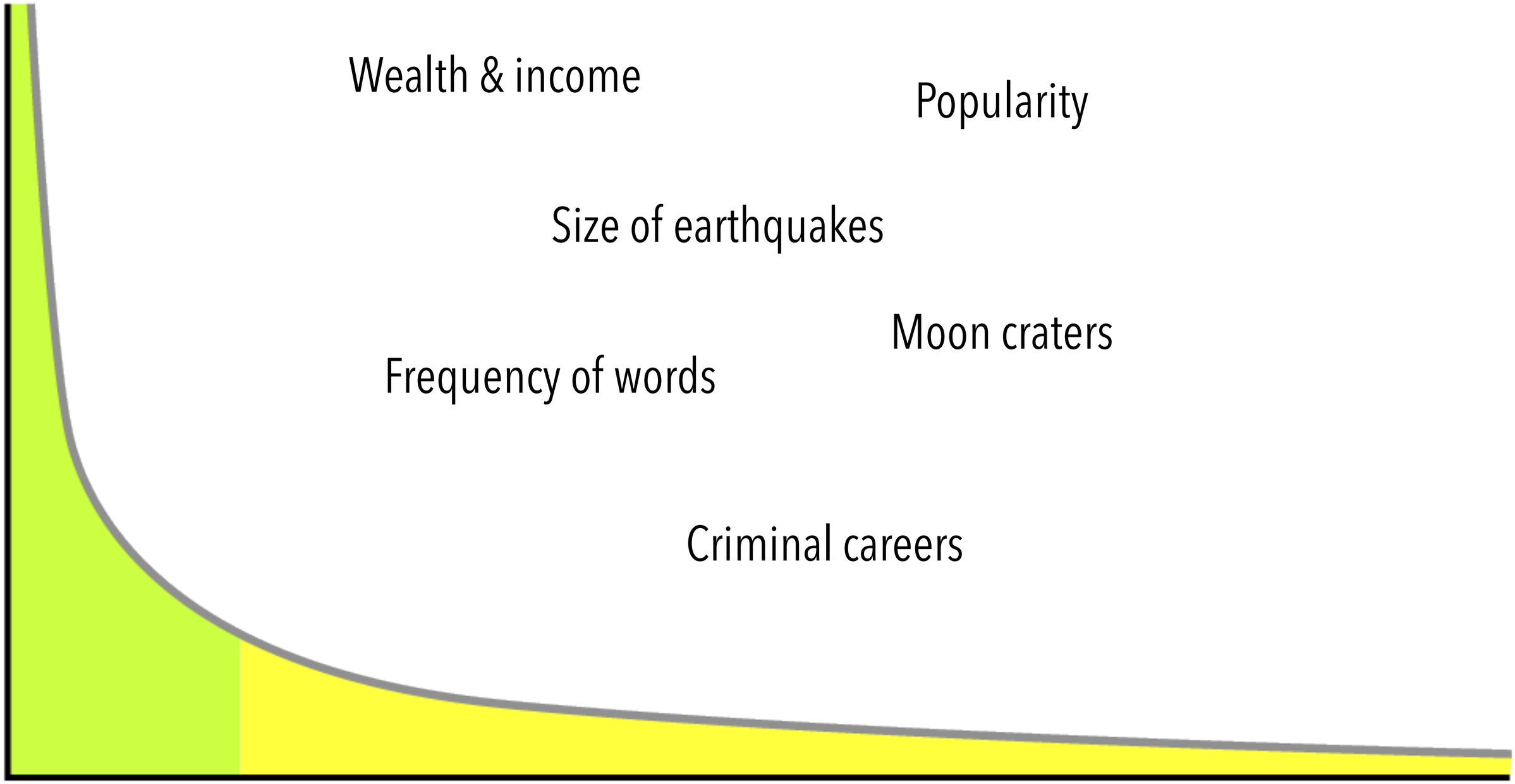


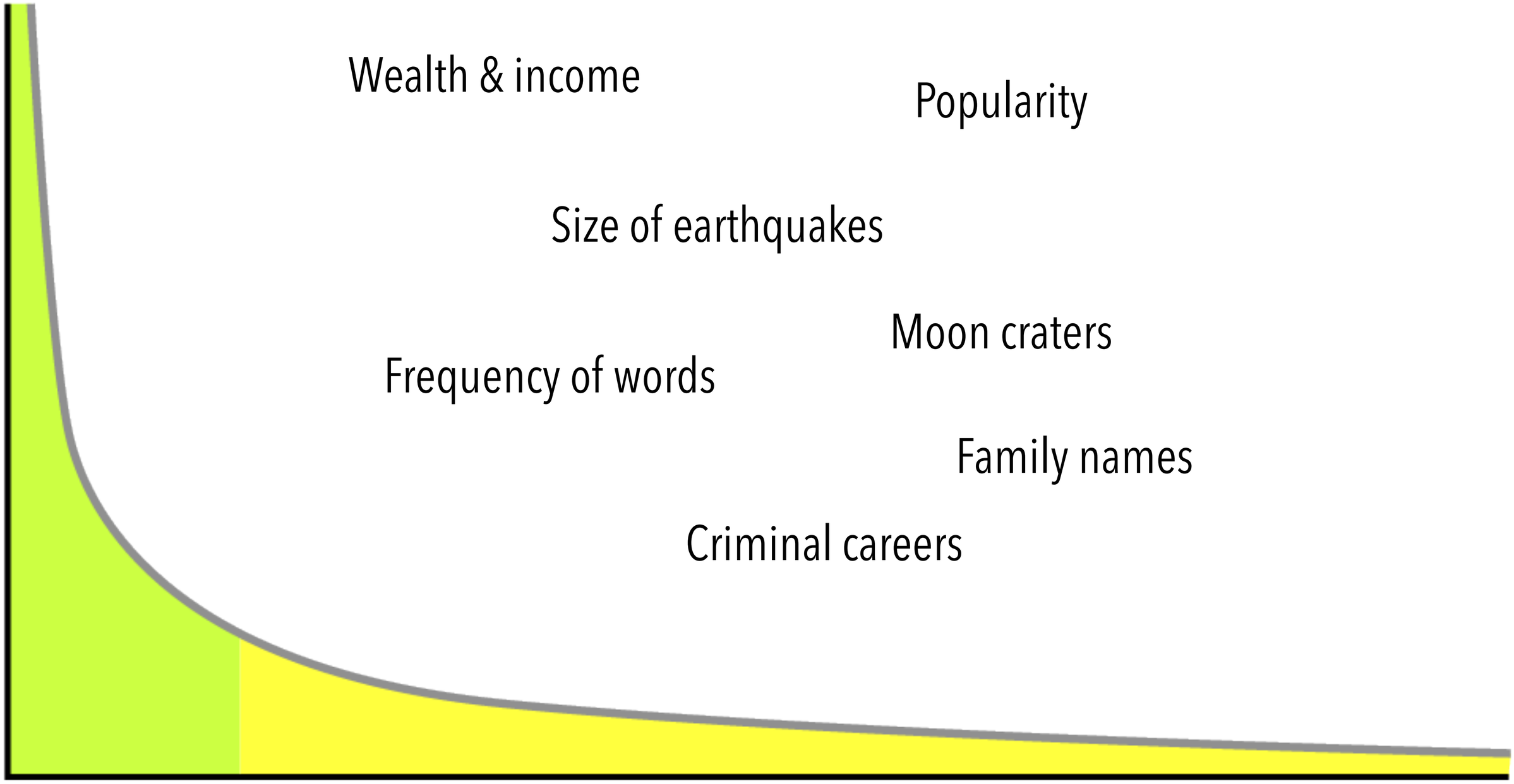

Wealth & income
Popularity

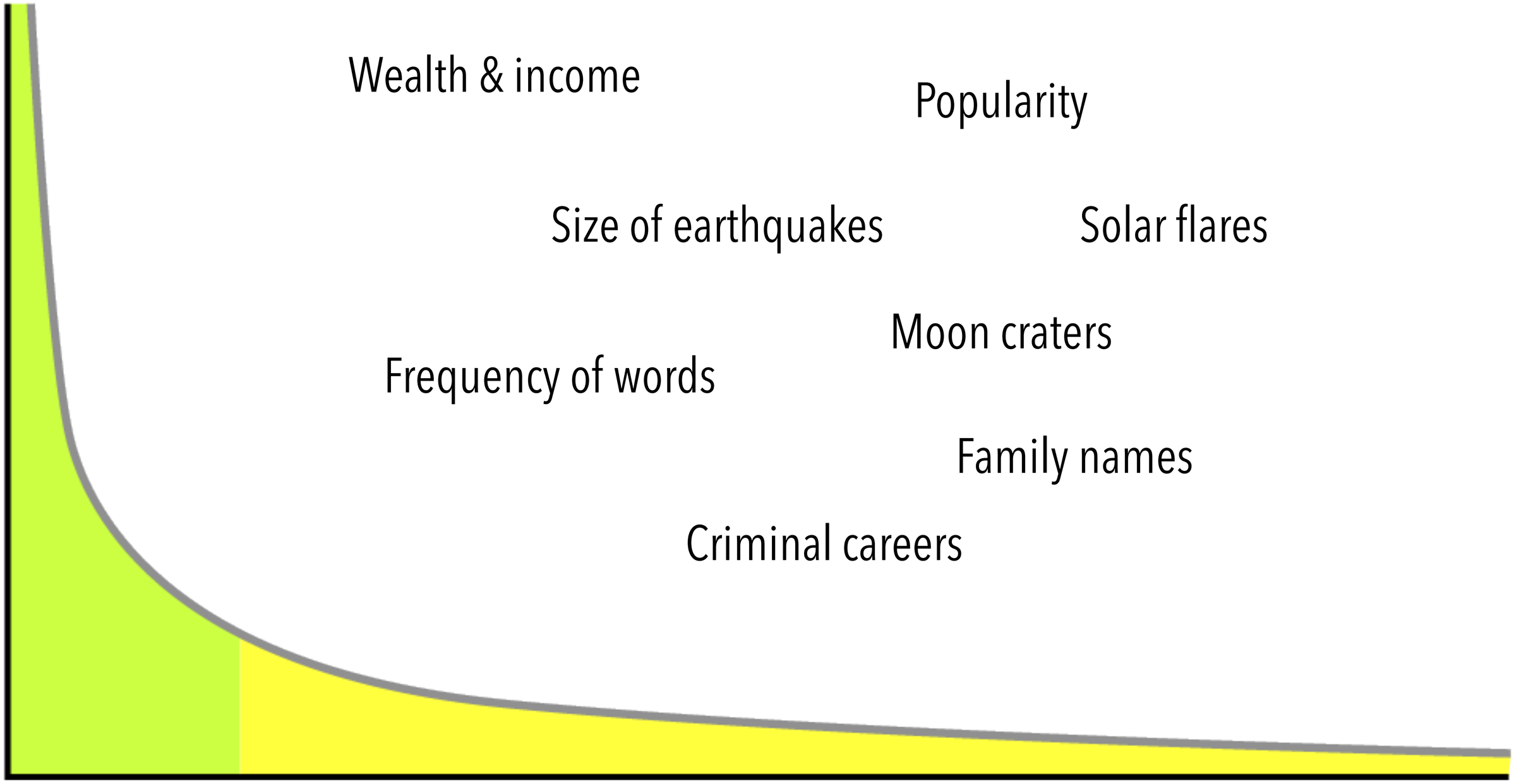

Wealth & income
Popularity
Moon craters

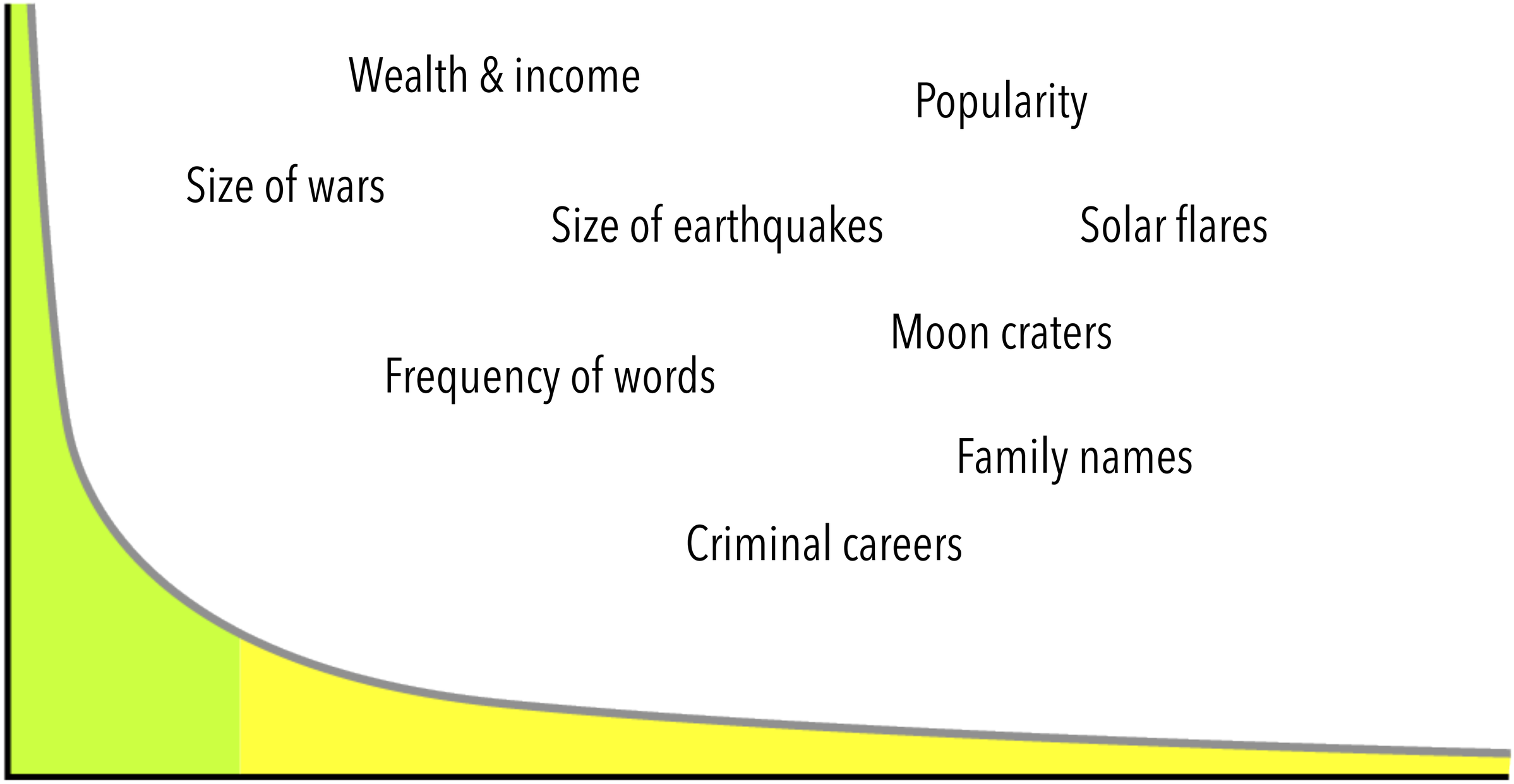

Wealth & income
Popularity
Moon craters
Frequency of words

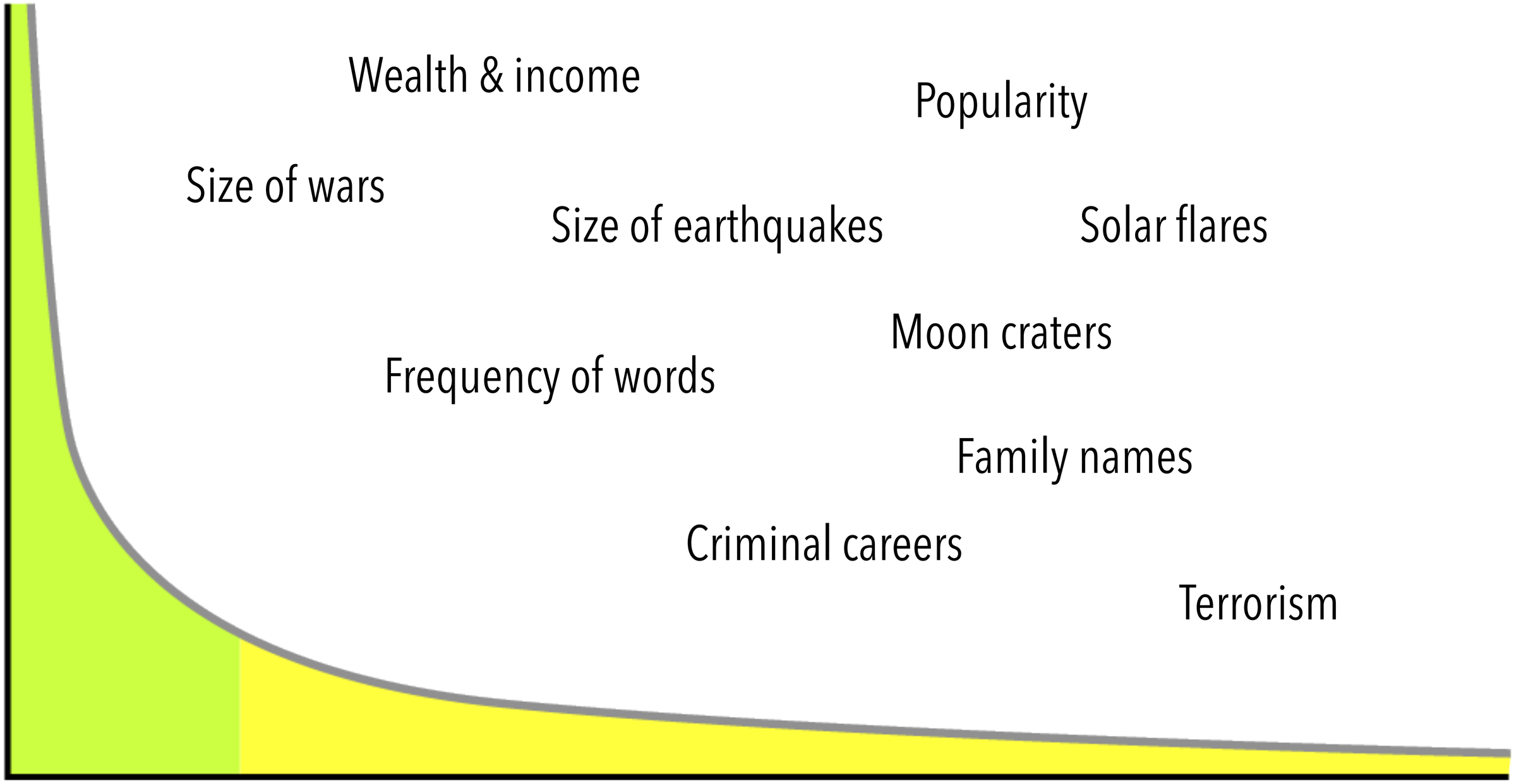

Wealth & income
Popularity
Size of earthquakes
Moon craters
Frequency of words

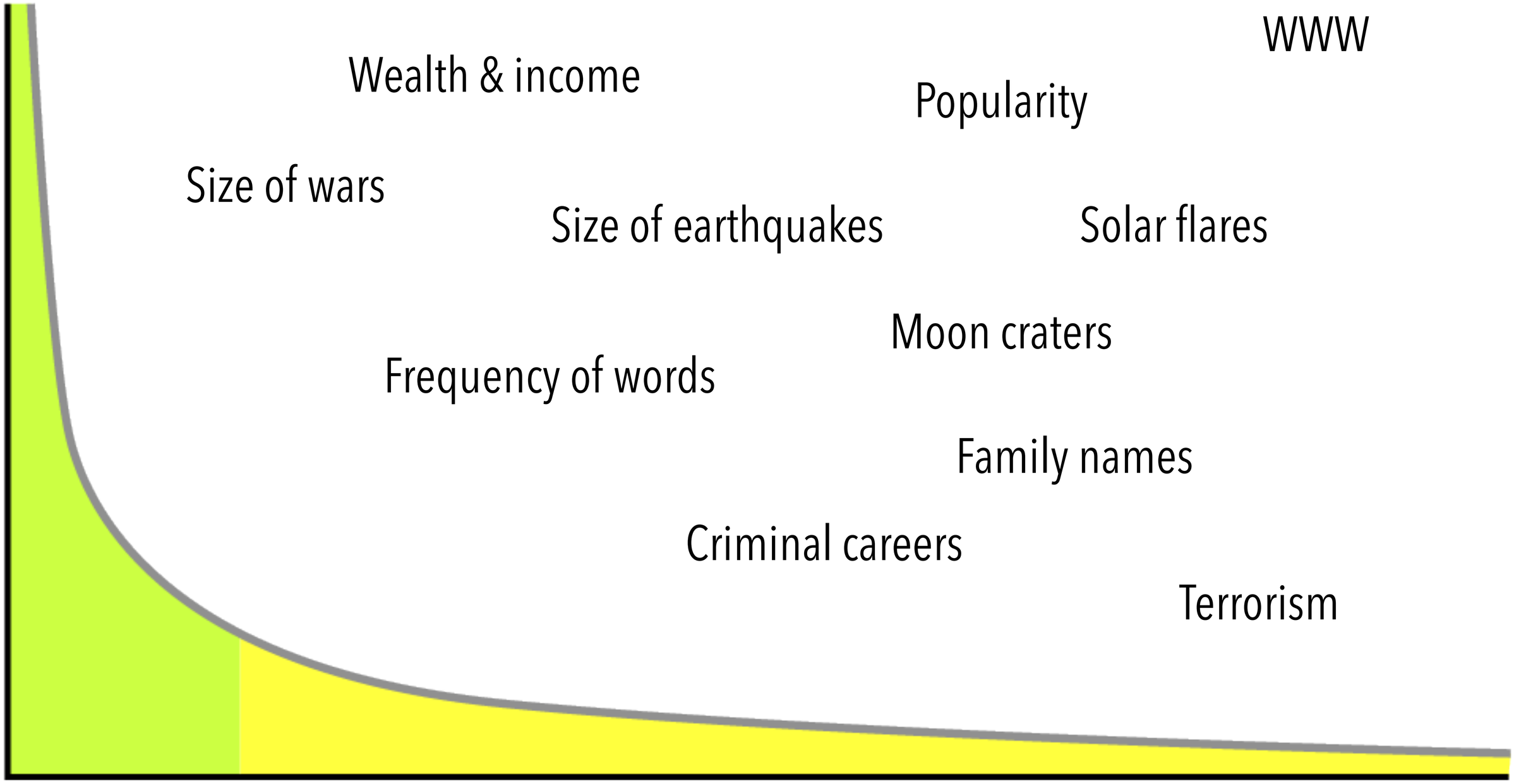

Wealth & income
Popularity
Size of earthquakes
Moon craters
Frequency of words
Criminal careers

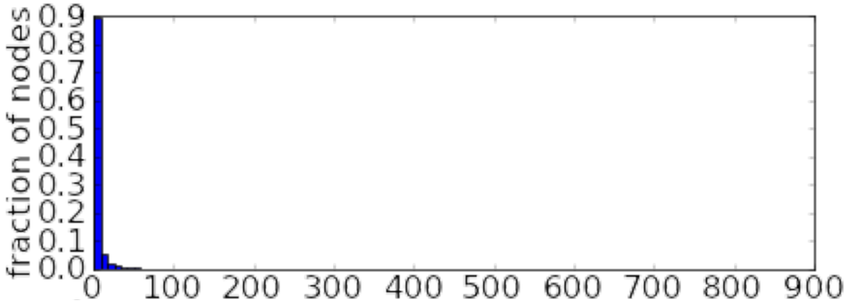

Wealth & income
Popularity
Size of earthquakes
Moon craters
Frequency of words
Family names
Criminal careers

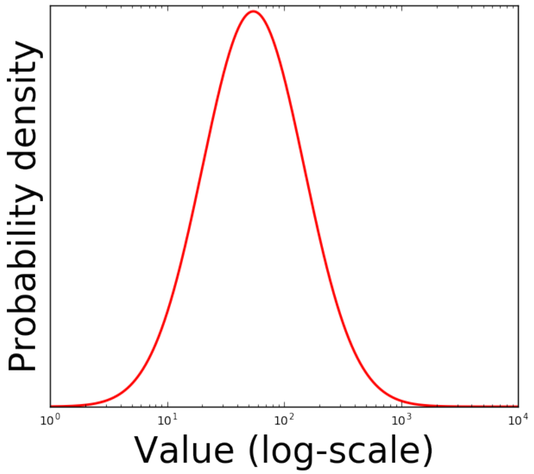

Wealth & income
Popularity
Size of earthquakes
Solar flares
Moon craters
Frequency of words
Family names
Criminal careers

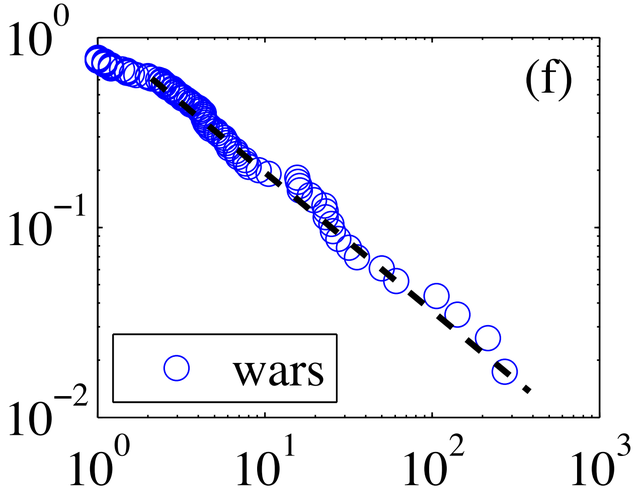

Wealth & income
Popularity
Size of wars
Size of earthquakes
Solar flares
Moon craters
Frequency of words
Family names
Criminal careers

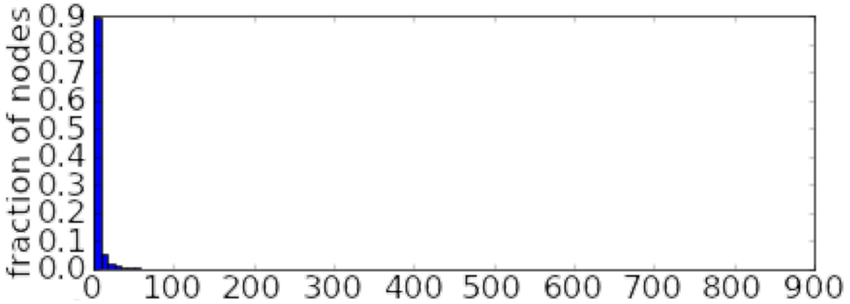

Wealth & income
Popularity
Size of wars
Size of earthquakes
Solar flares
Moon craters
Frequency of words
Family names
Criminal careers
Terrorism


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


fraction of nodes
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
100
200
300
400
500
600
700
800
900

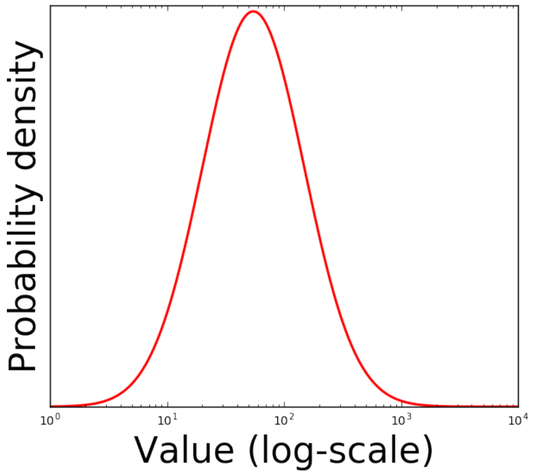

Probability density
Value (log-scale)

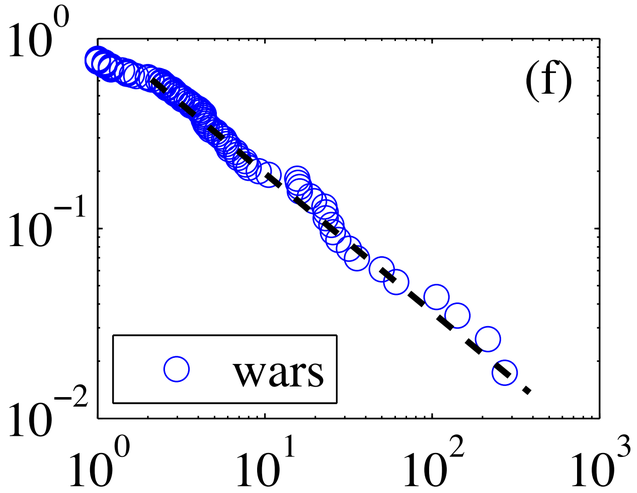

(f)
wars

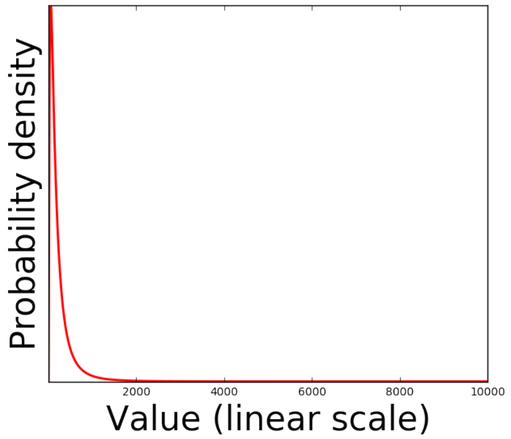


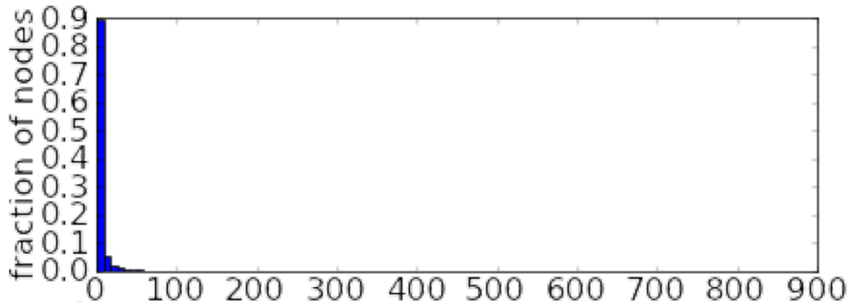
Linear scale:

Can't see any!



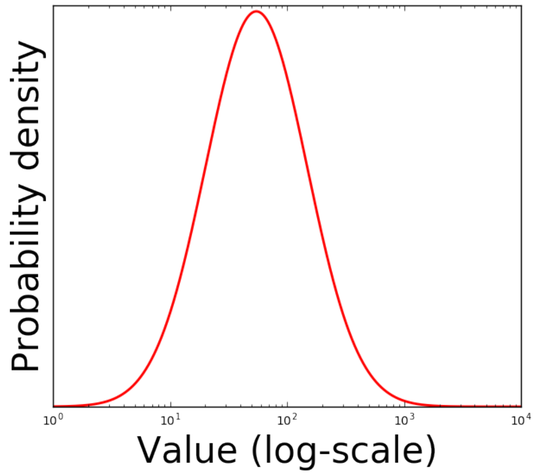


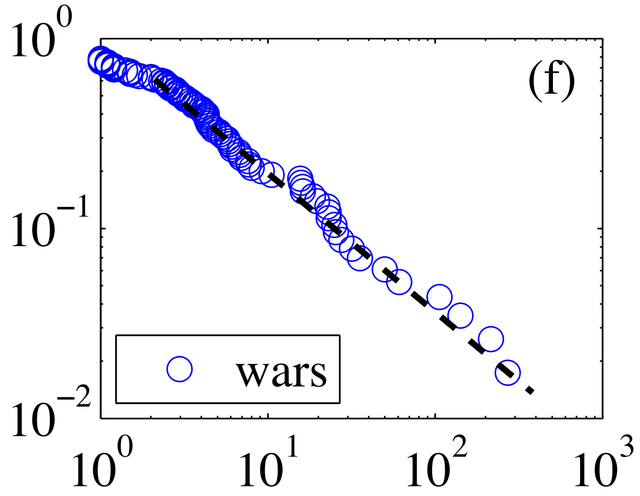


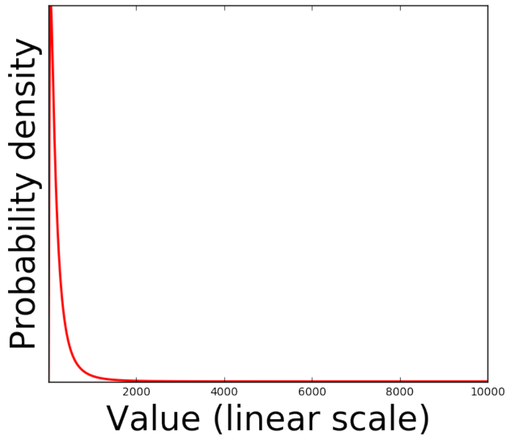


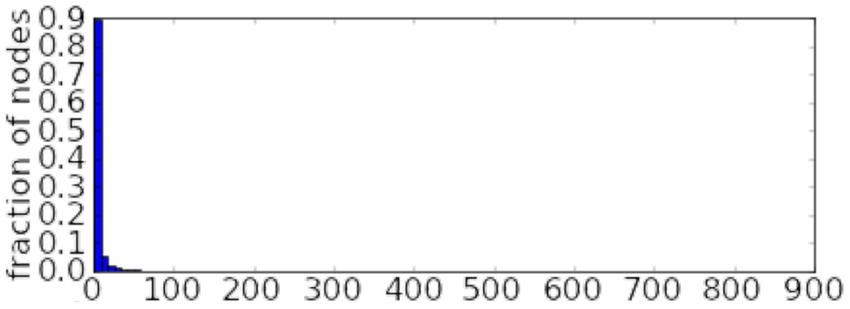
Linear scale:

Can’t see any!



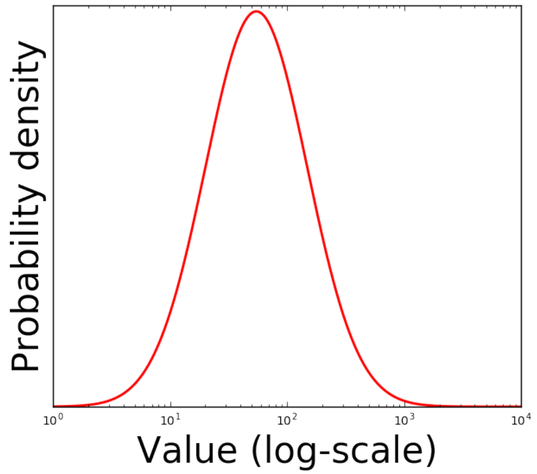


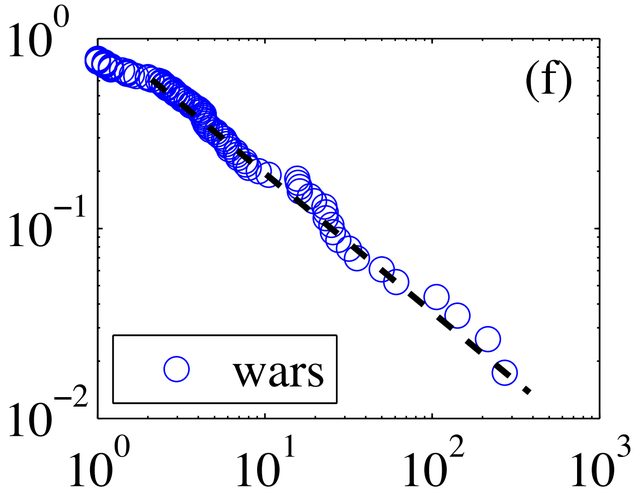




Linear scale:

Can't see any!

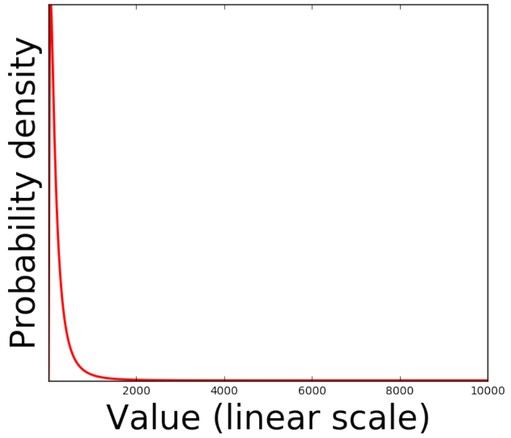


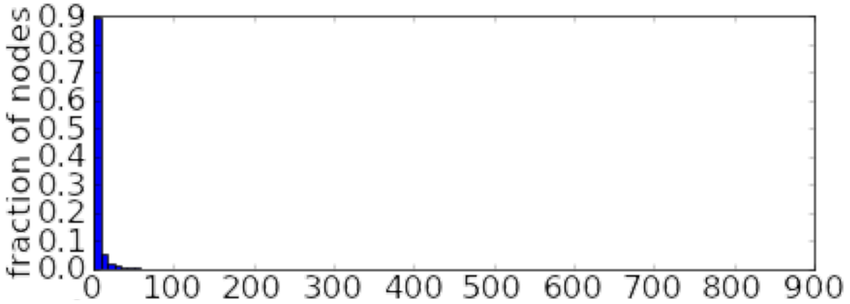


log scale:

Useful!

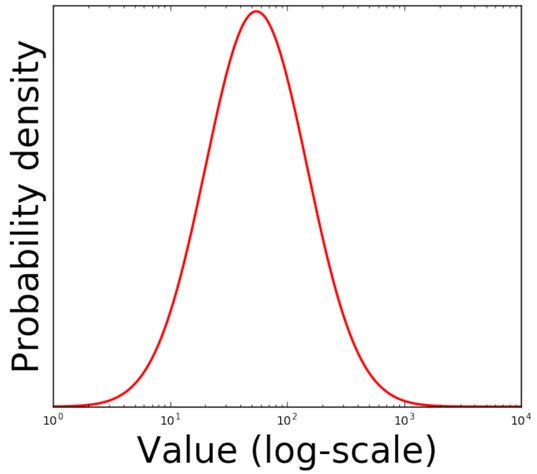
Have a fat tail?

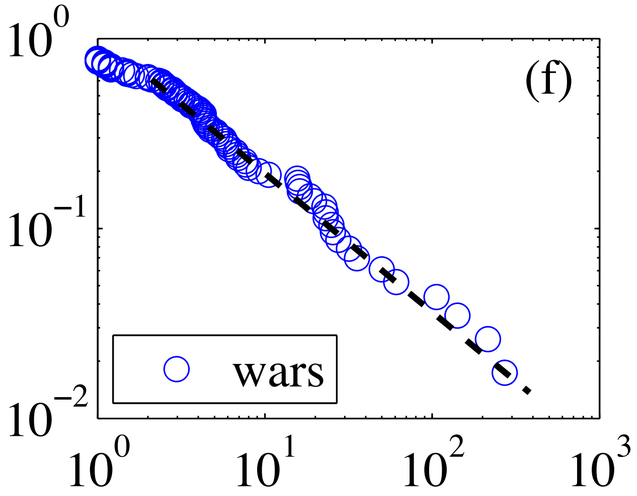




Linear scale:

Can't see any!

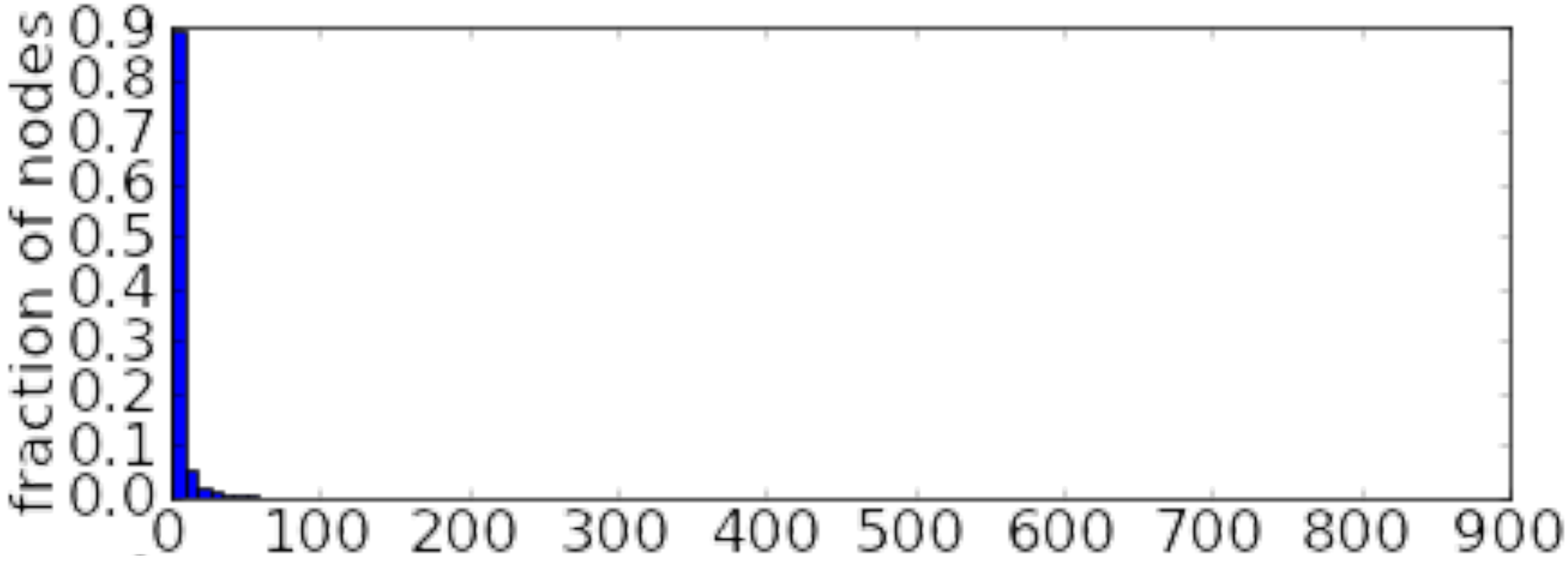




log scale:

Useful!

What would you do?



Histogram in log-scale

How should we create bins in log-scale?

Constant width in log-scale?

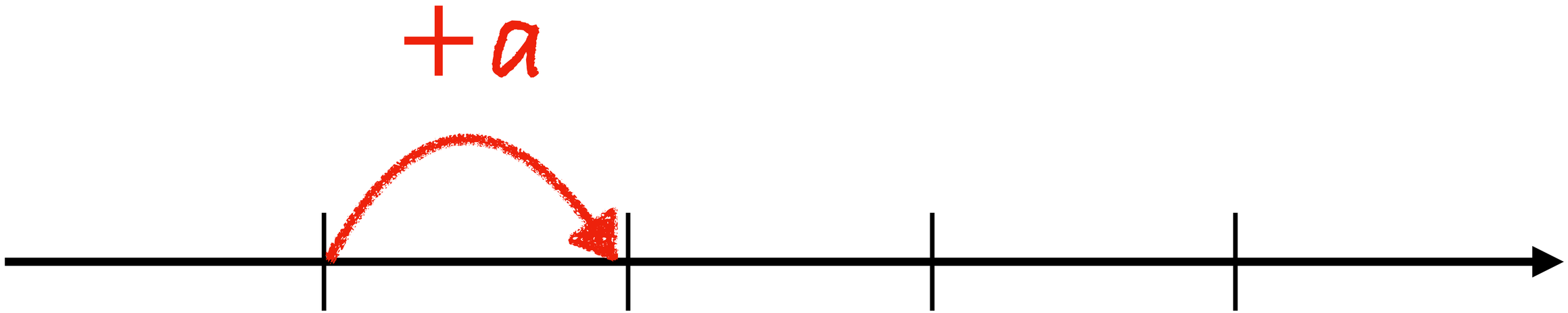
First bin: [1, 3)

What are the next several bins?

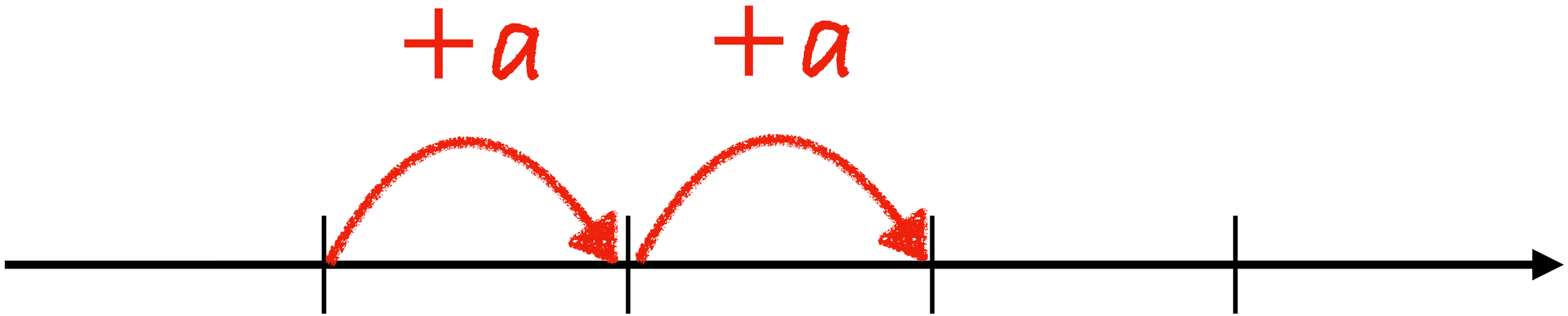
(same width bins in log-scale)

In linear scale

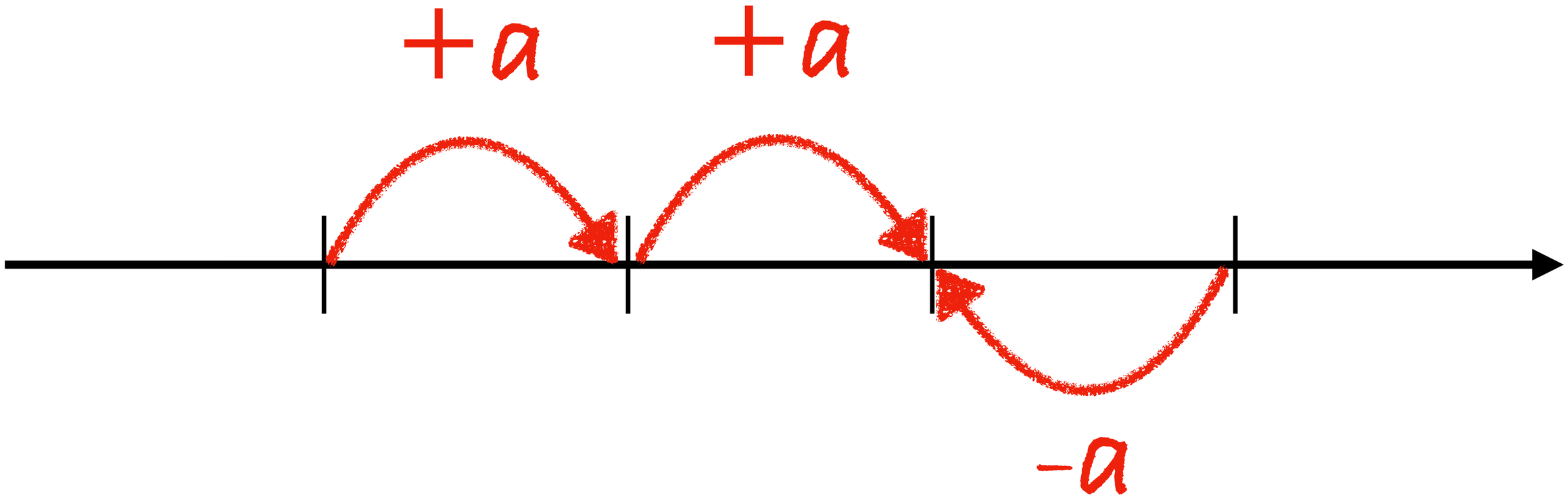
In linear scale



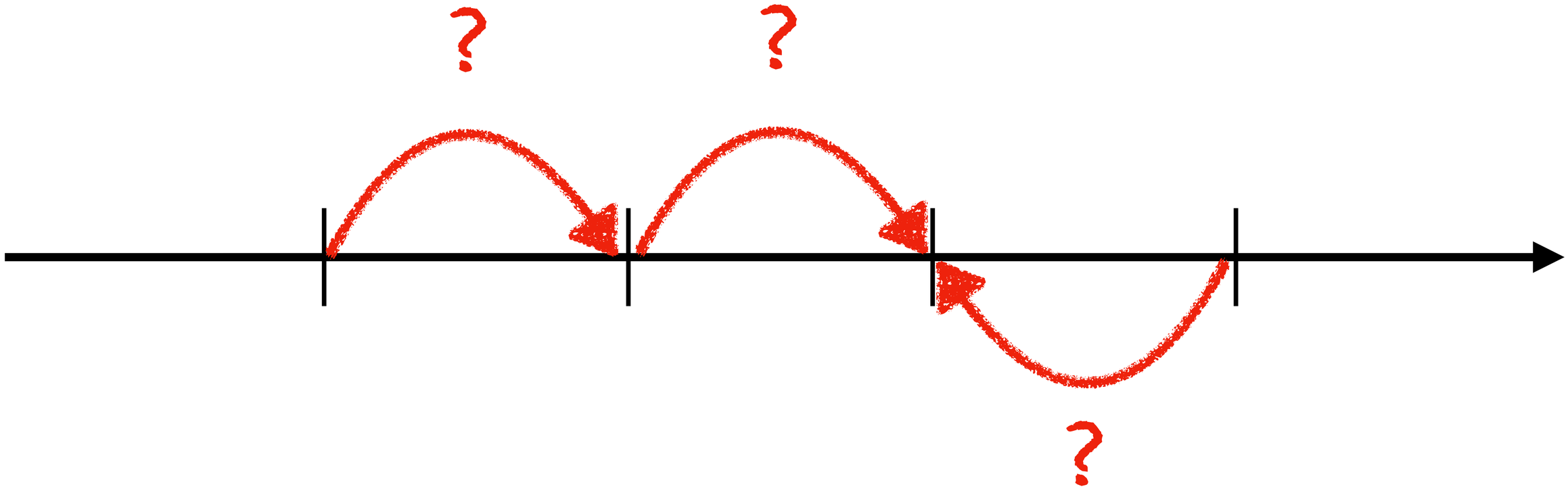
In linear scale



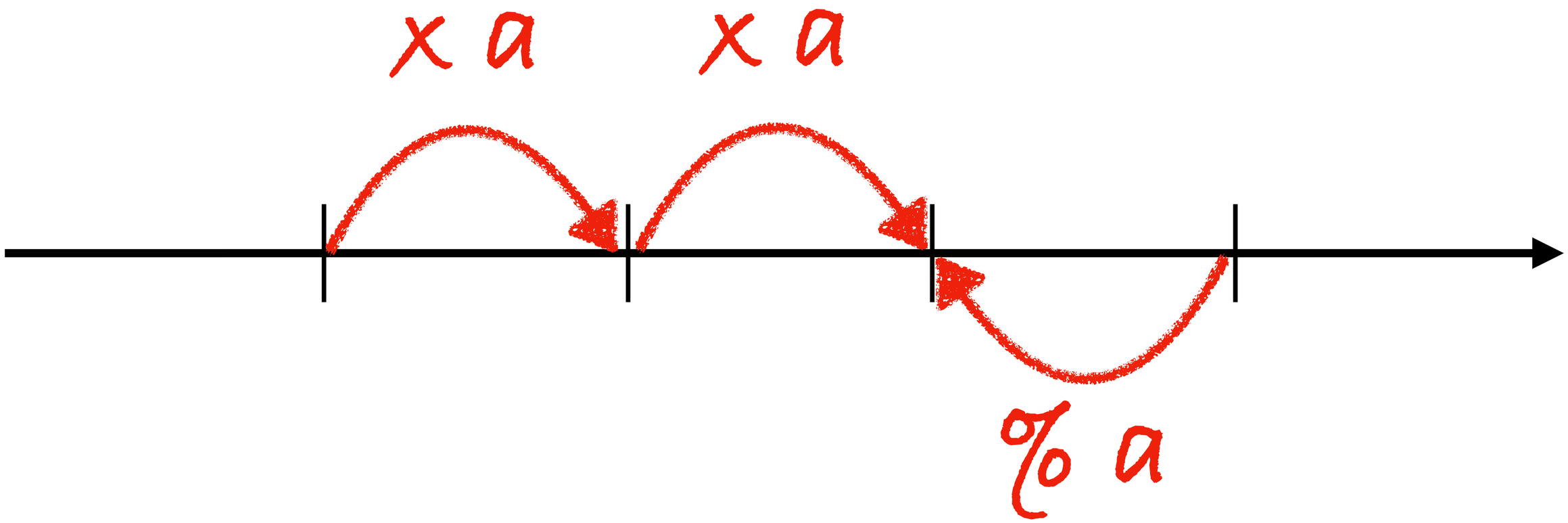
In linear scale



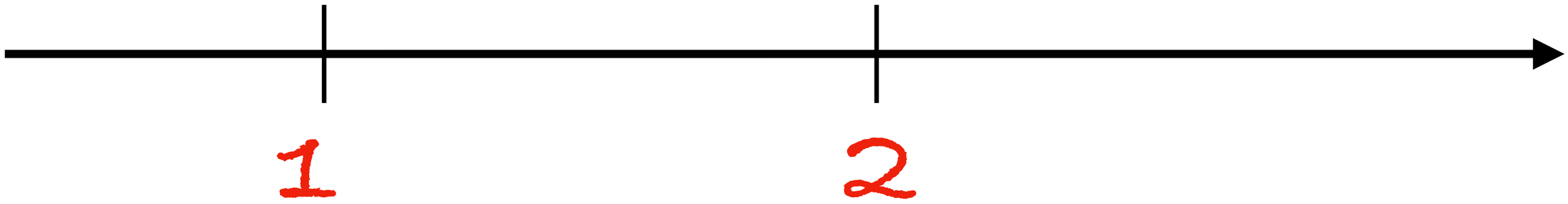
In log scale?



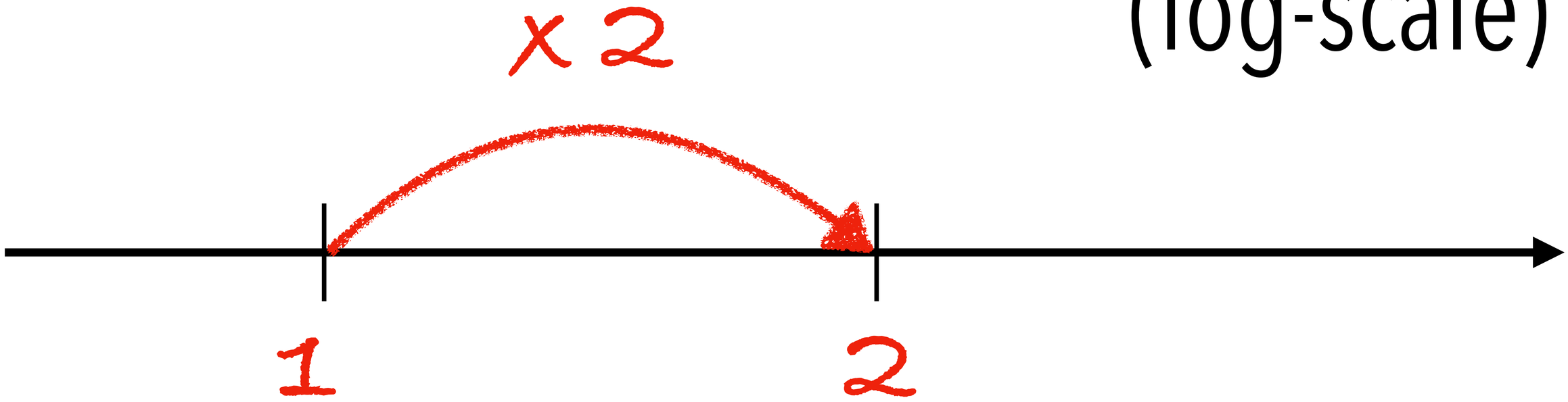
In log scale?



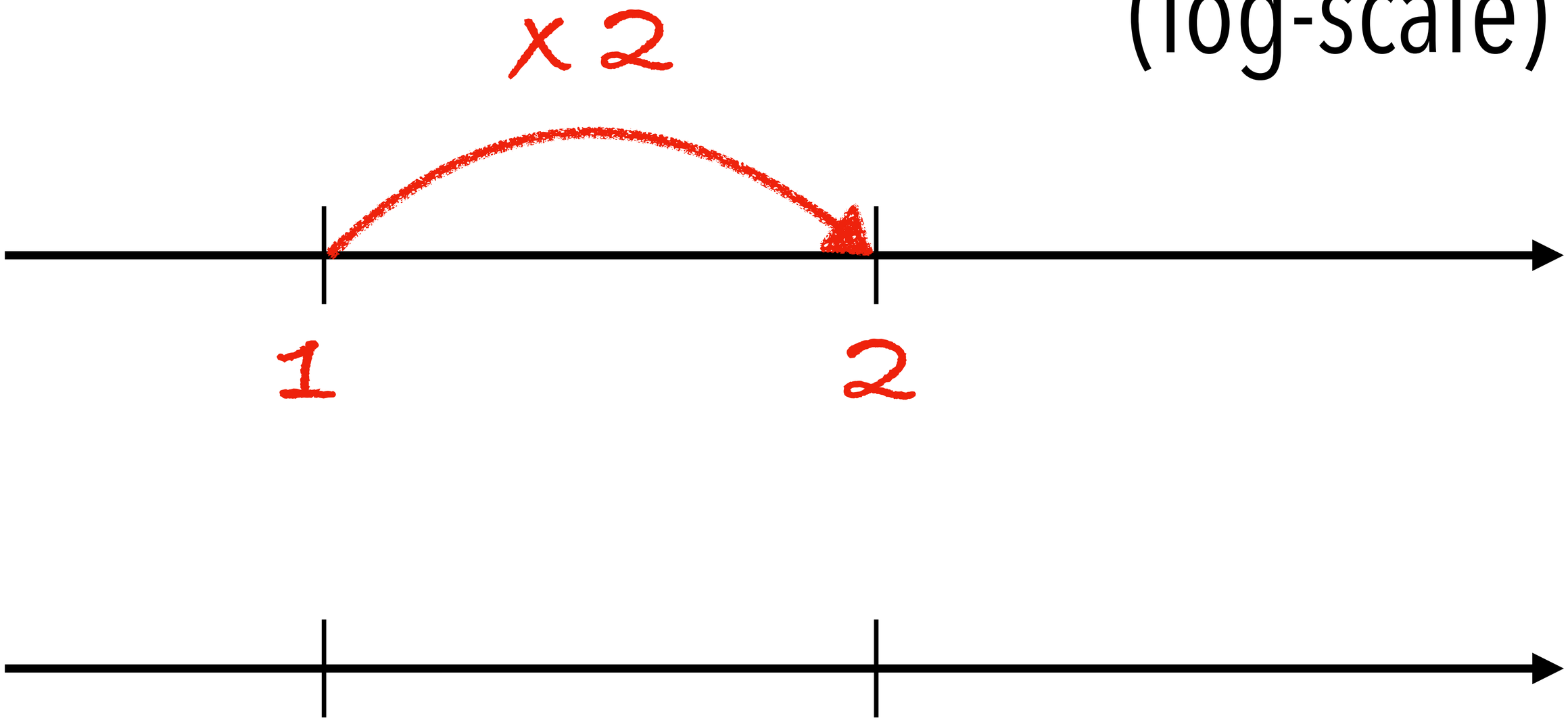
(log-scale)



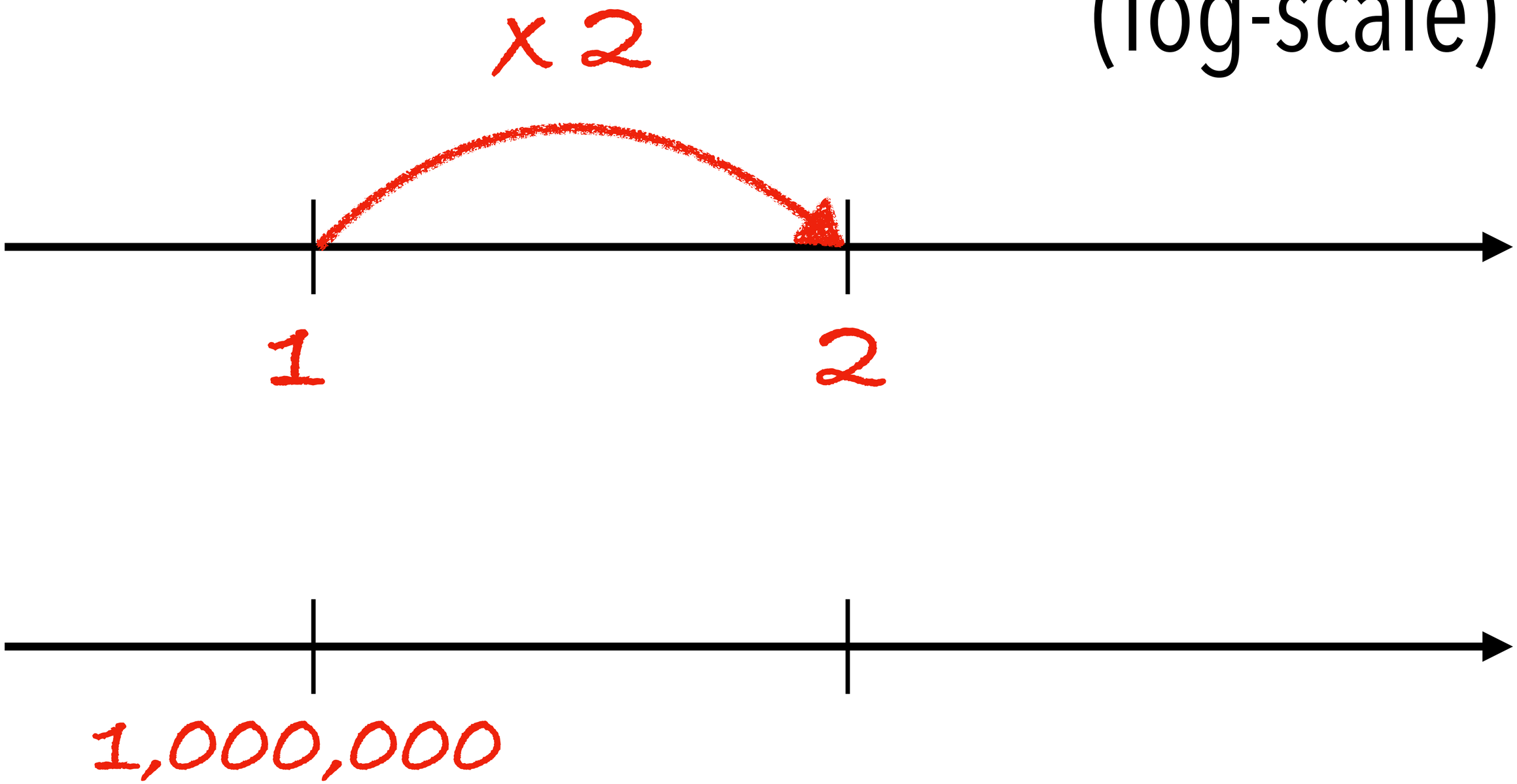
(log-scale)



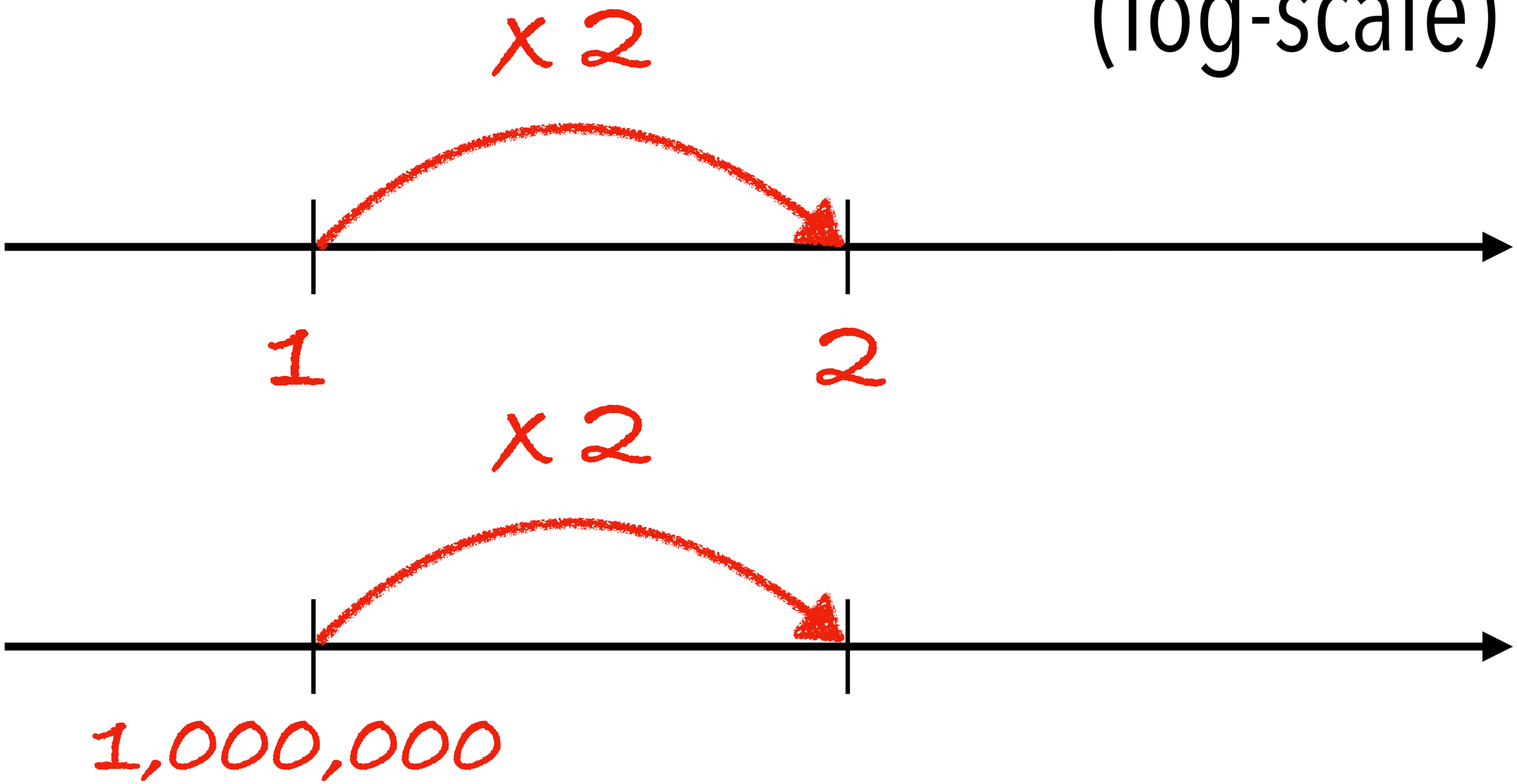
(log-scale)



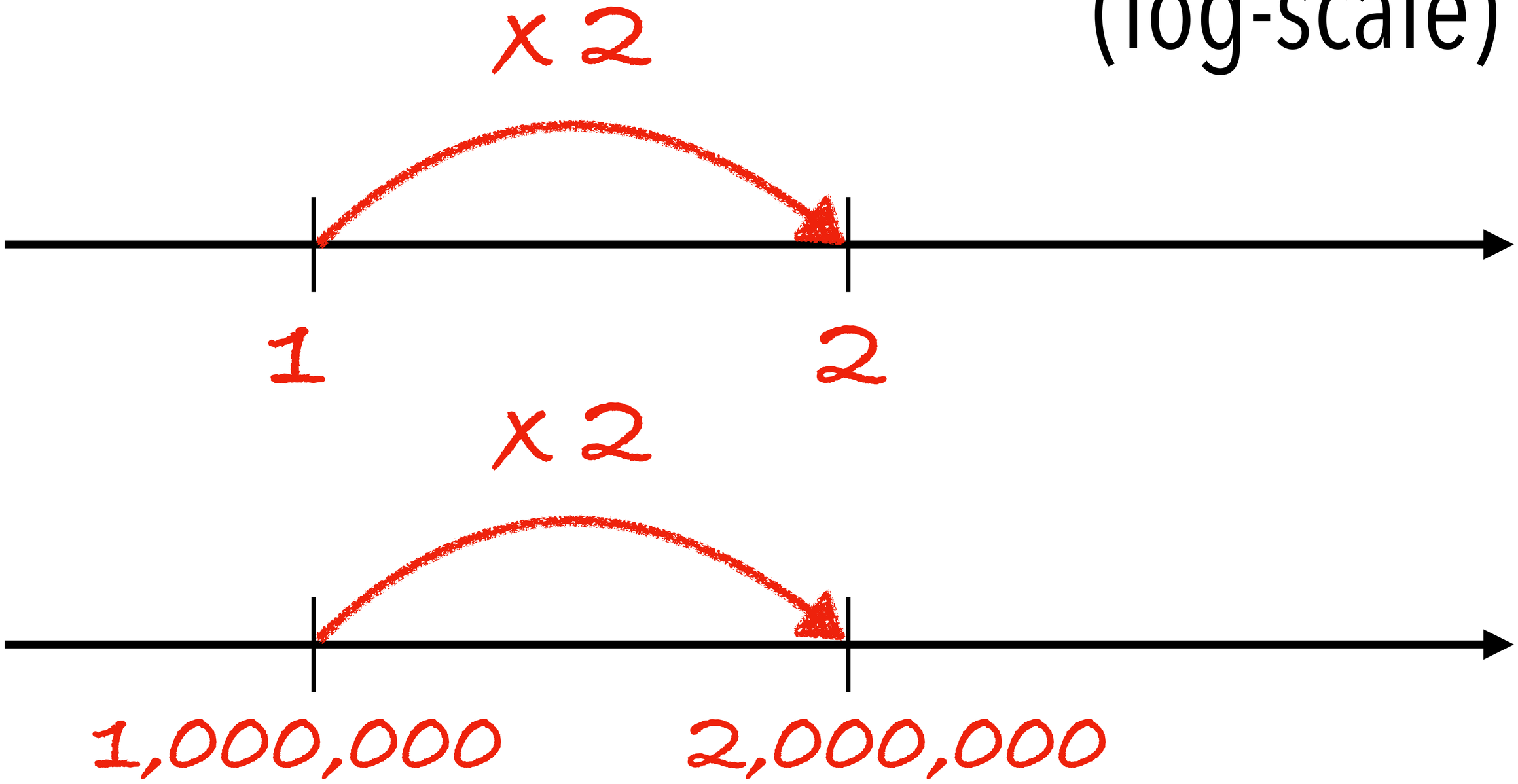
(log-scale)



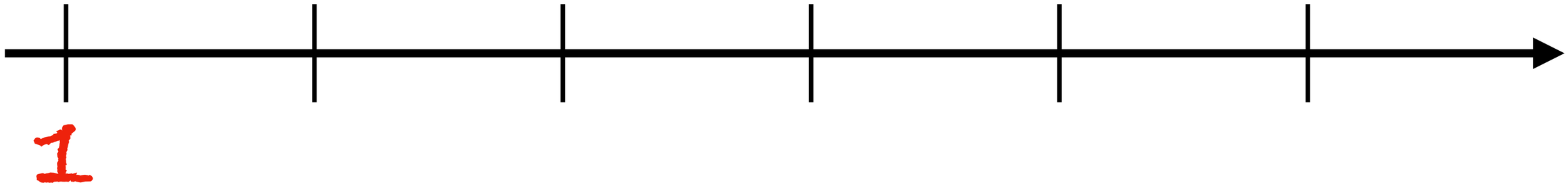
(log-scale)



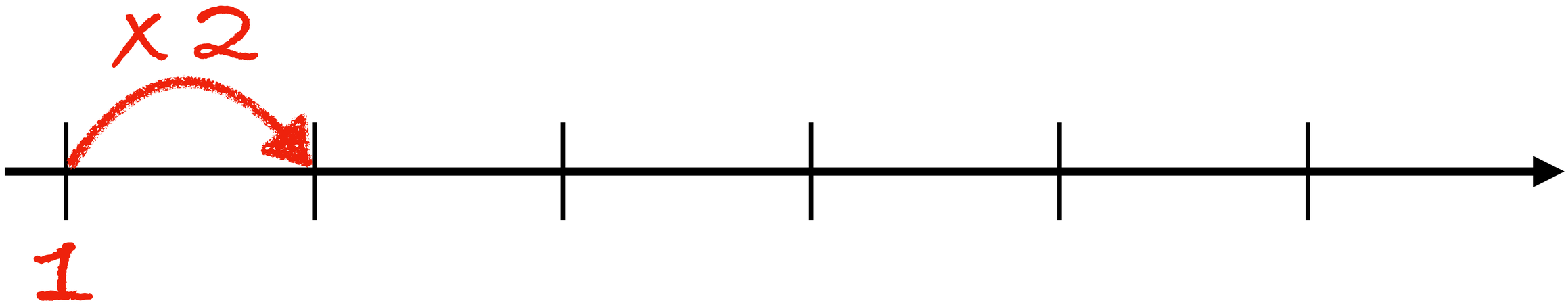
(log-scale)



(log-scale)

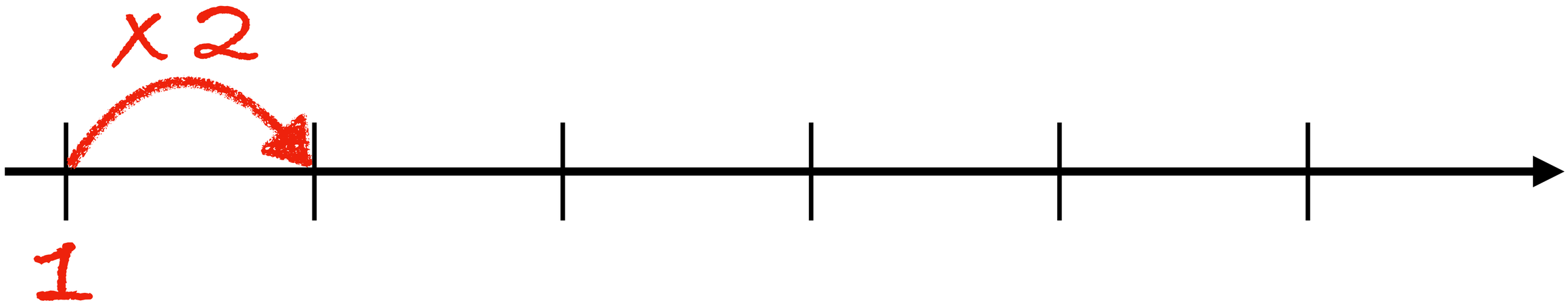


(log-scale)



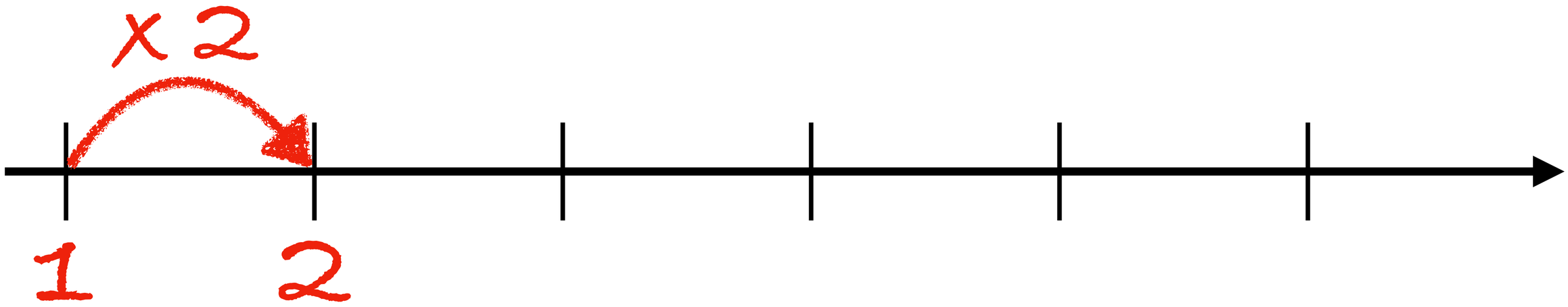
(You can choose other widths too)

(log-scale)

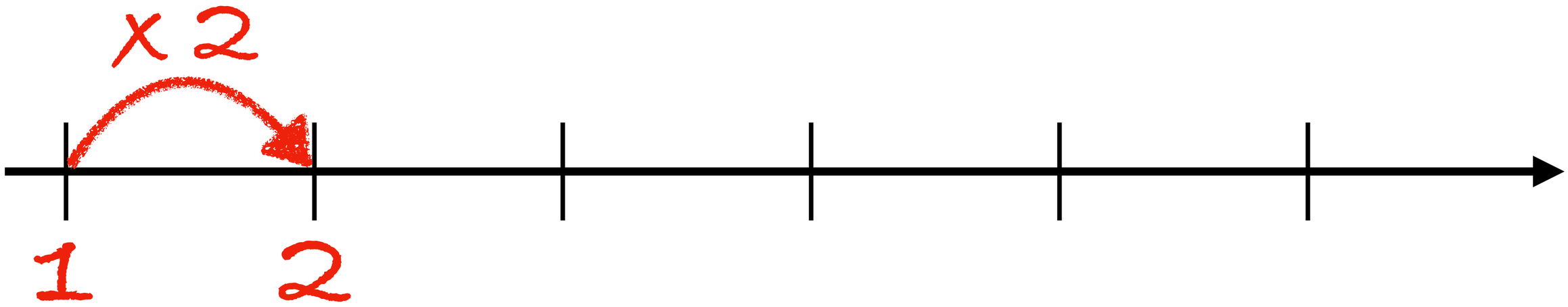


(You can choose other widths too)

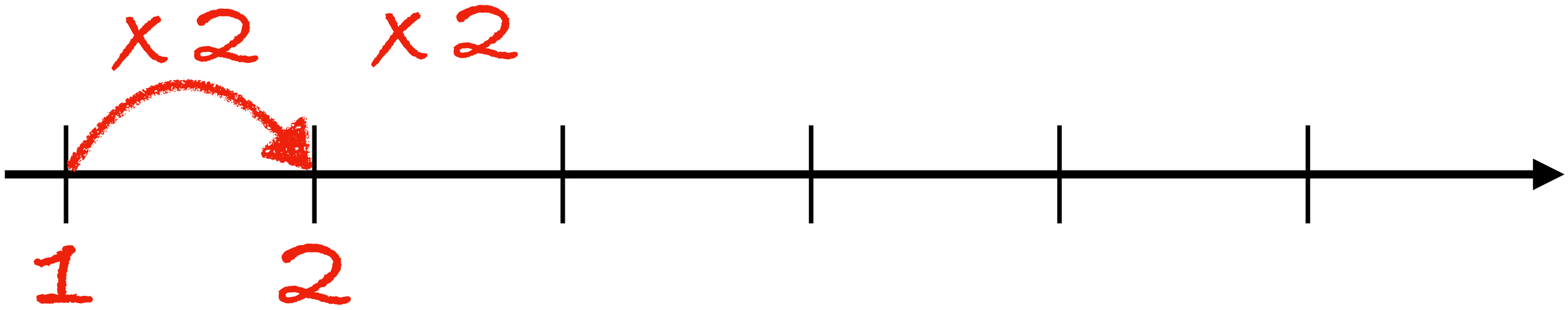
(log-scale)



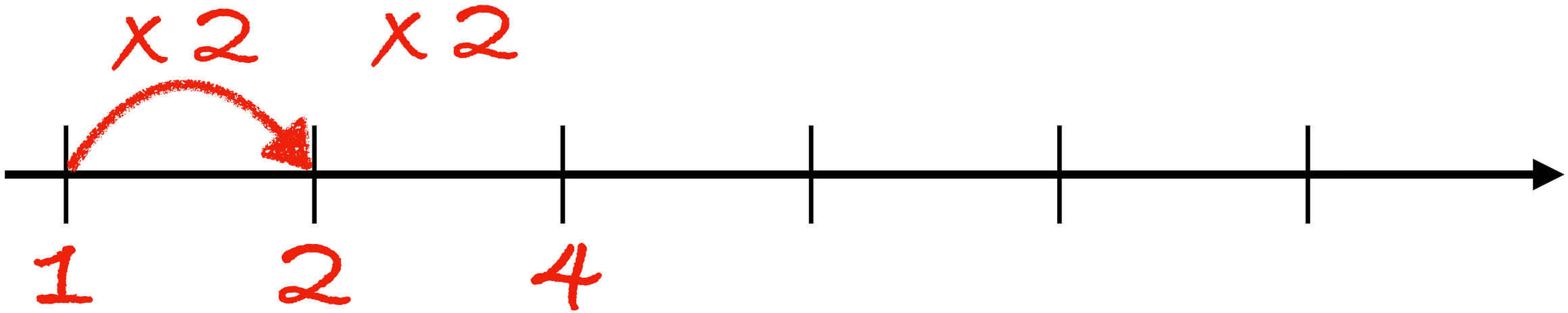
(log-scale)



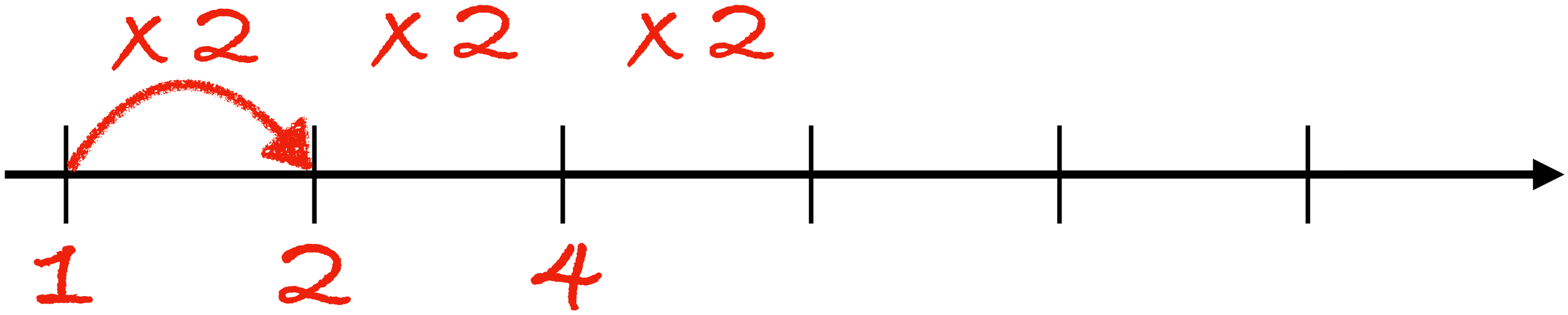
(log-scale)



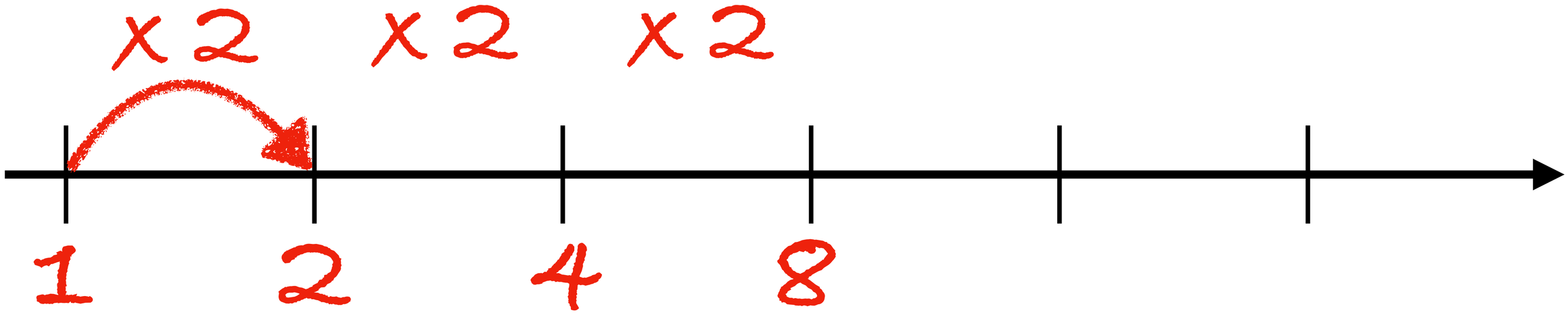
(log-scale)



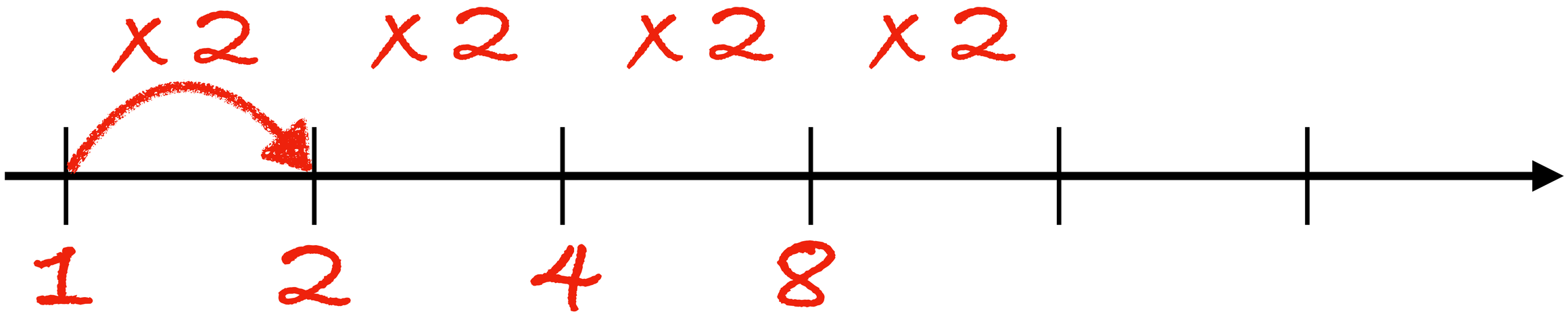
(log-scale)



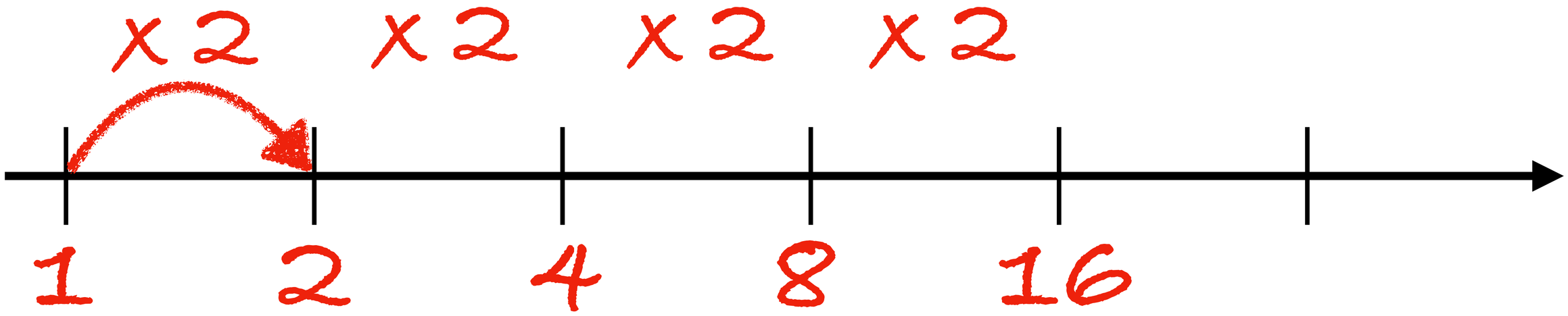
(log-scale)



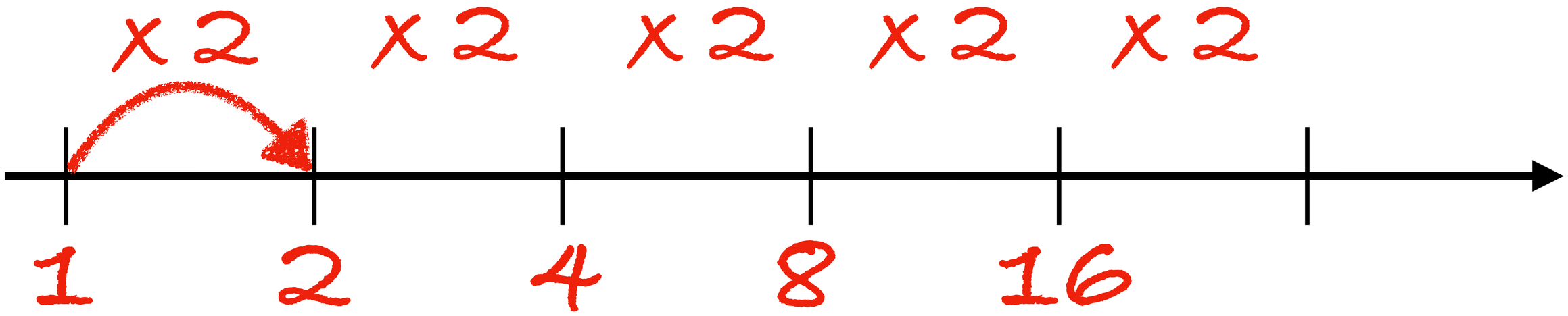
(log-scale)



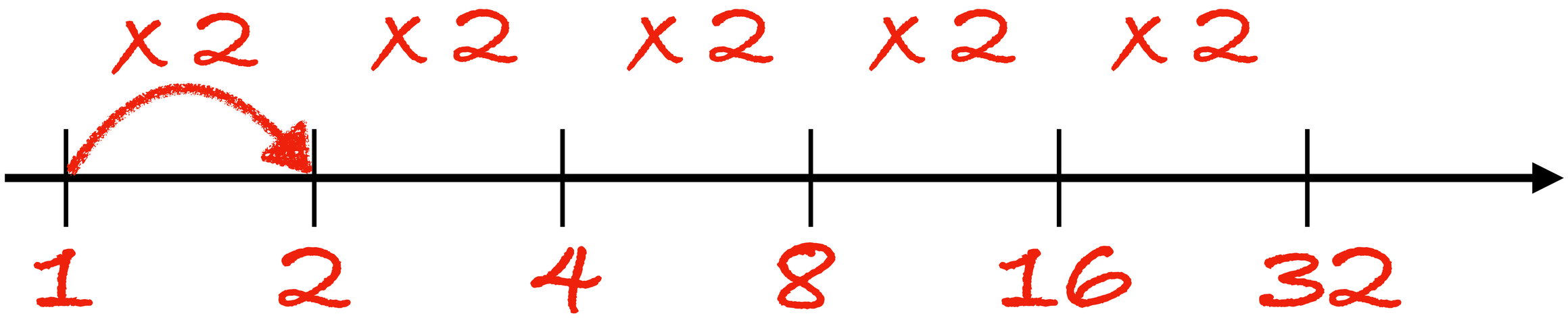
(log-scale)



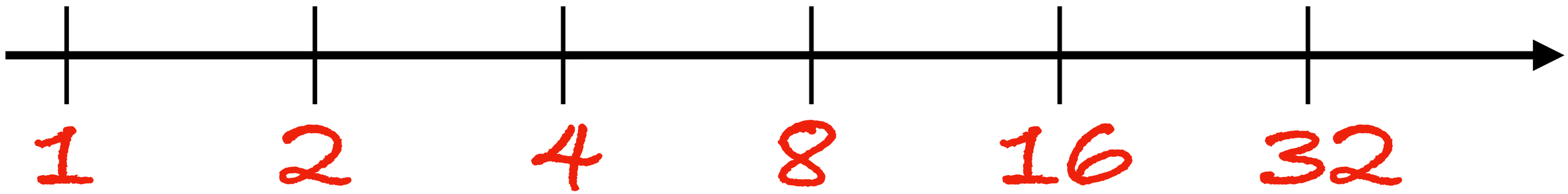
(log-scale)

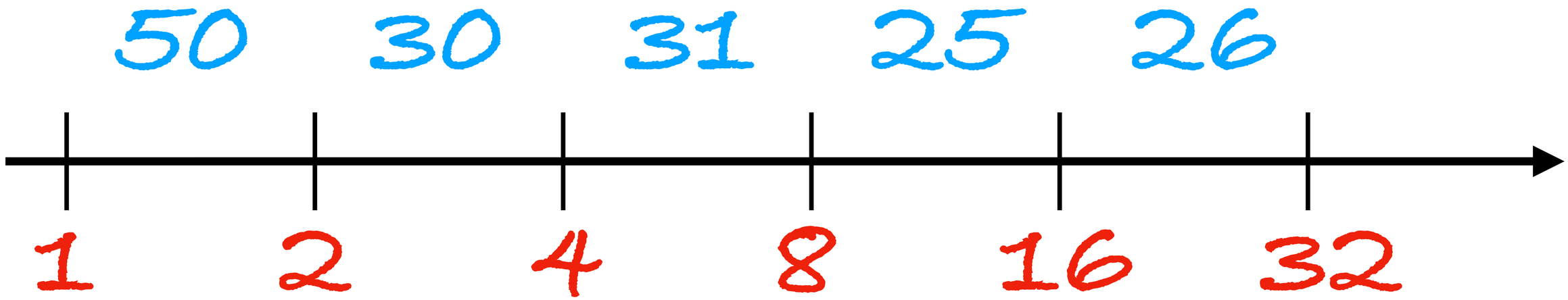


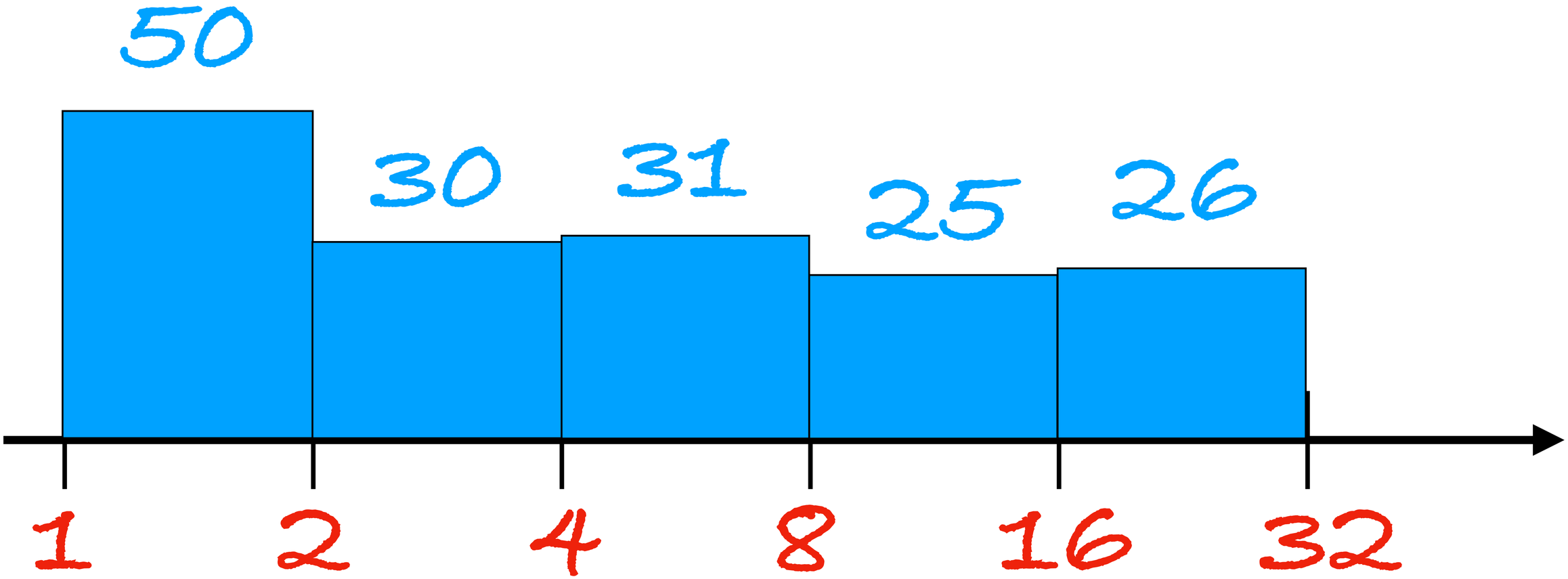
(log-scale)

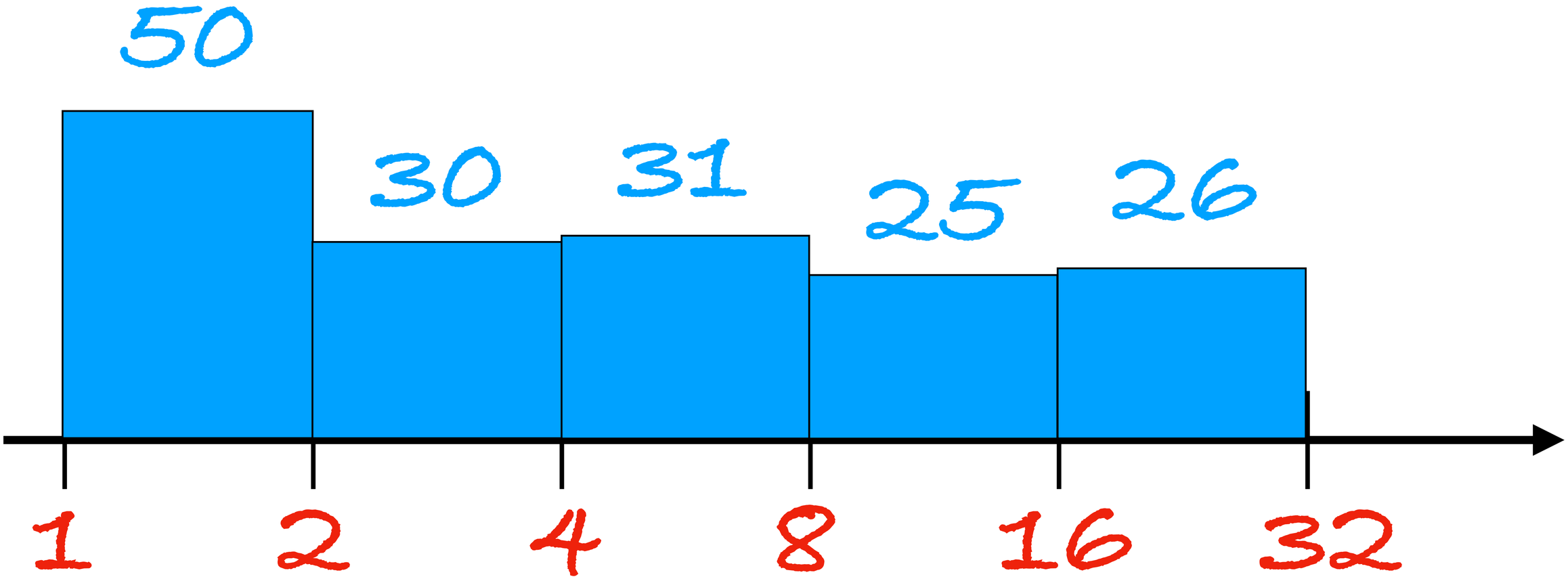


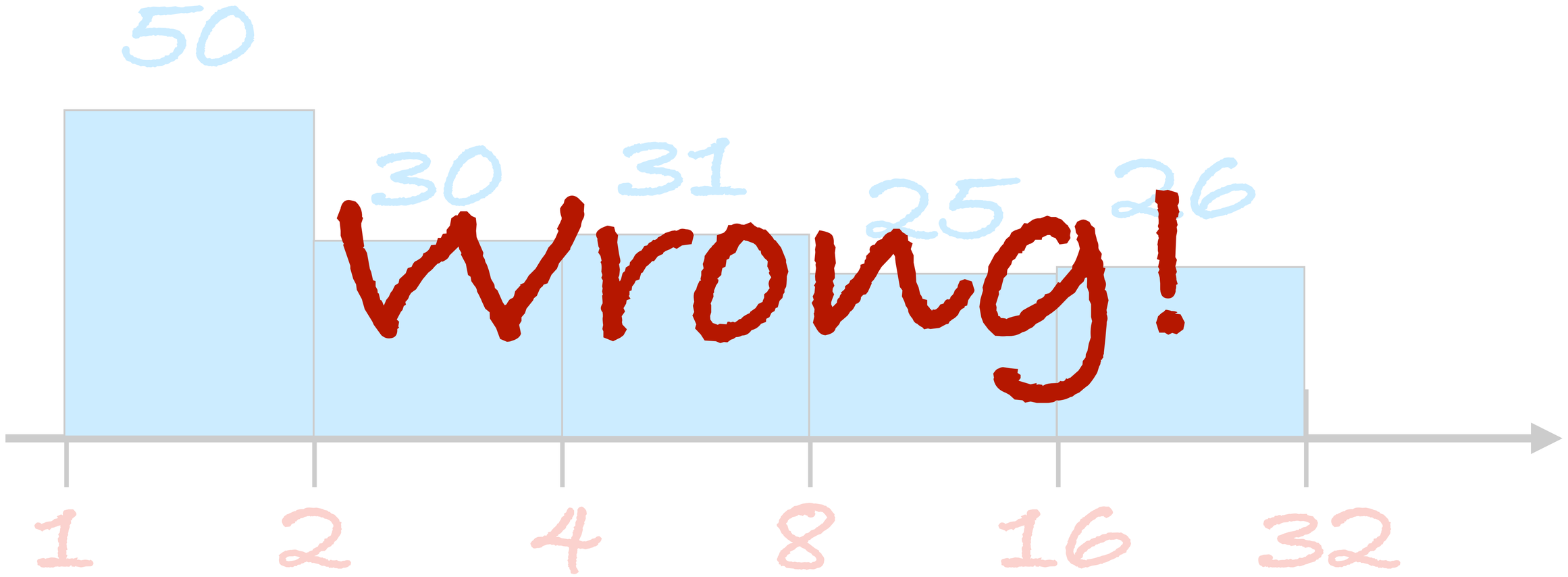
"Log-binning"
(logarithmic binning)


1
2
4
8
16
32

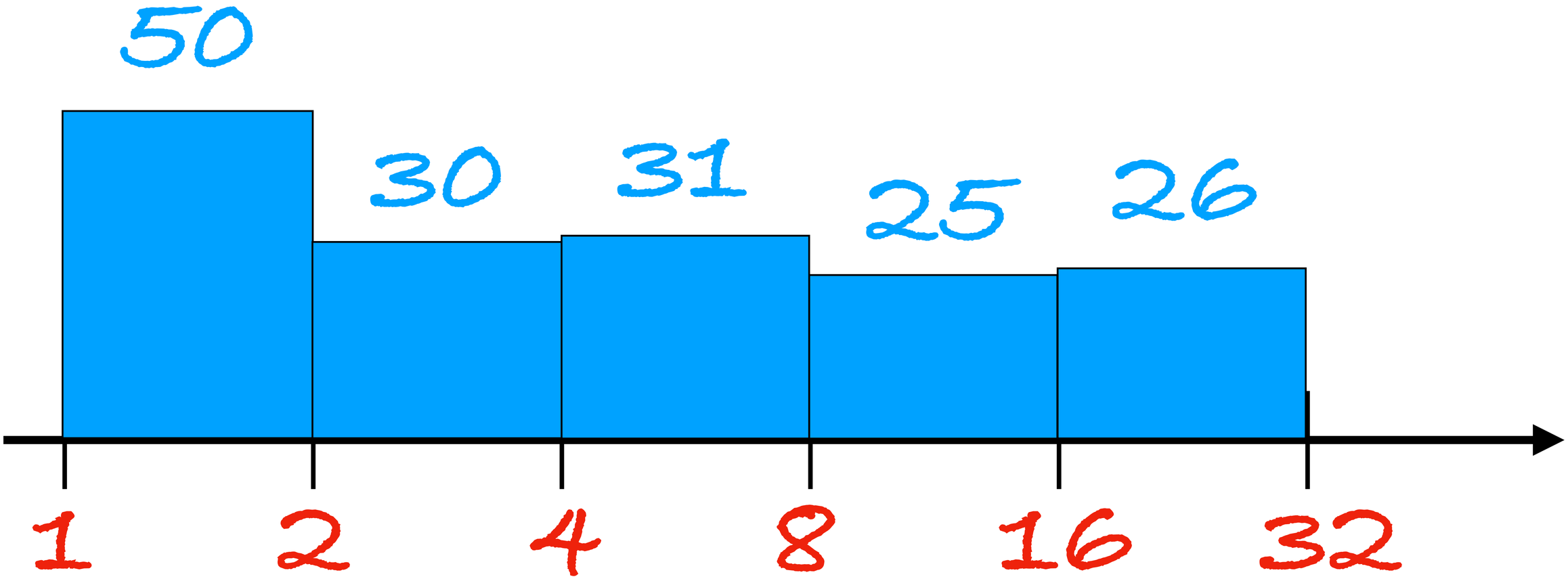

50
30
31
25
26
1
2
4
8
16
32

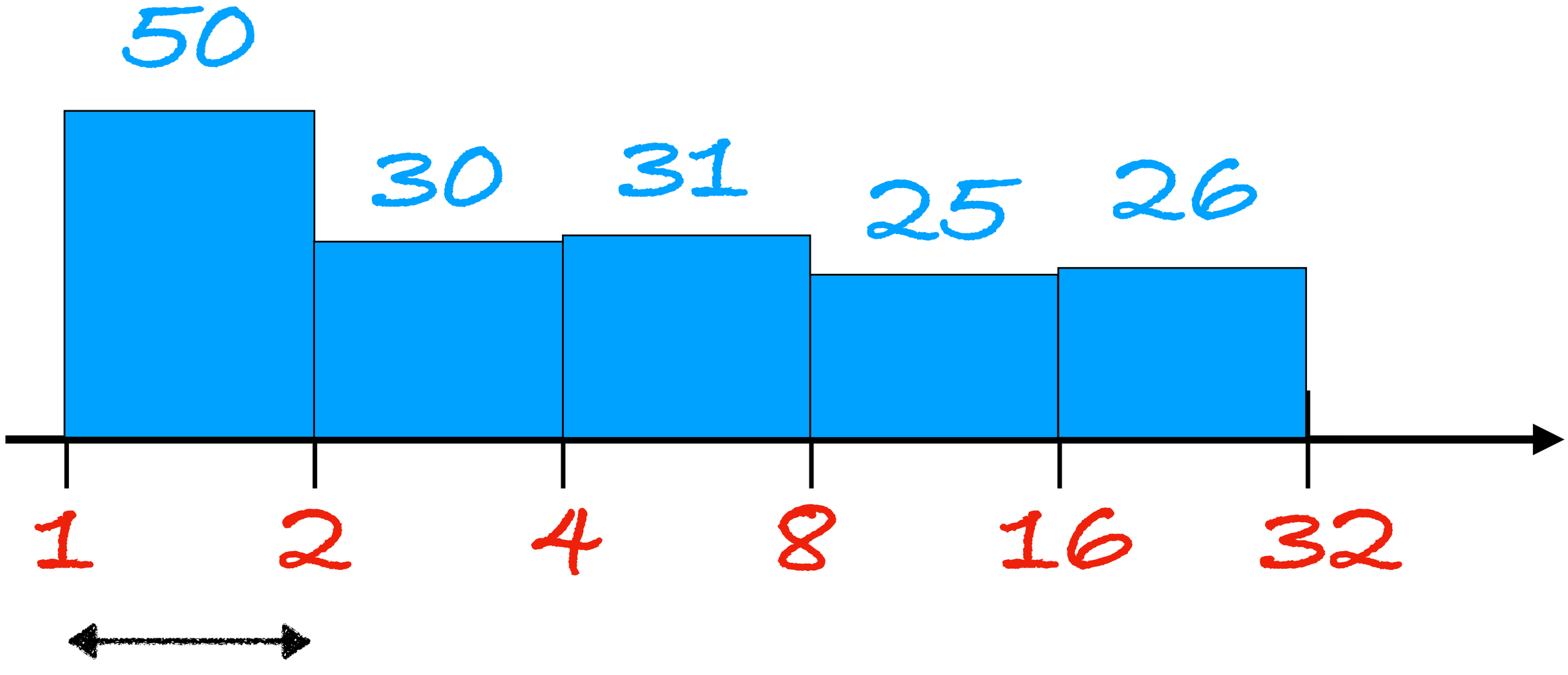

50
30
31
25
26
1
2
4
8
16
32

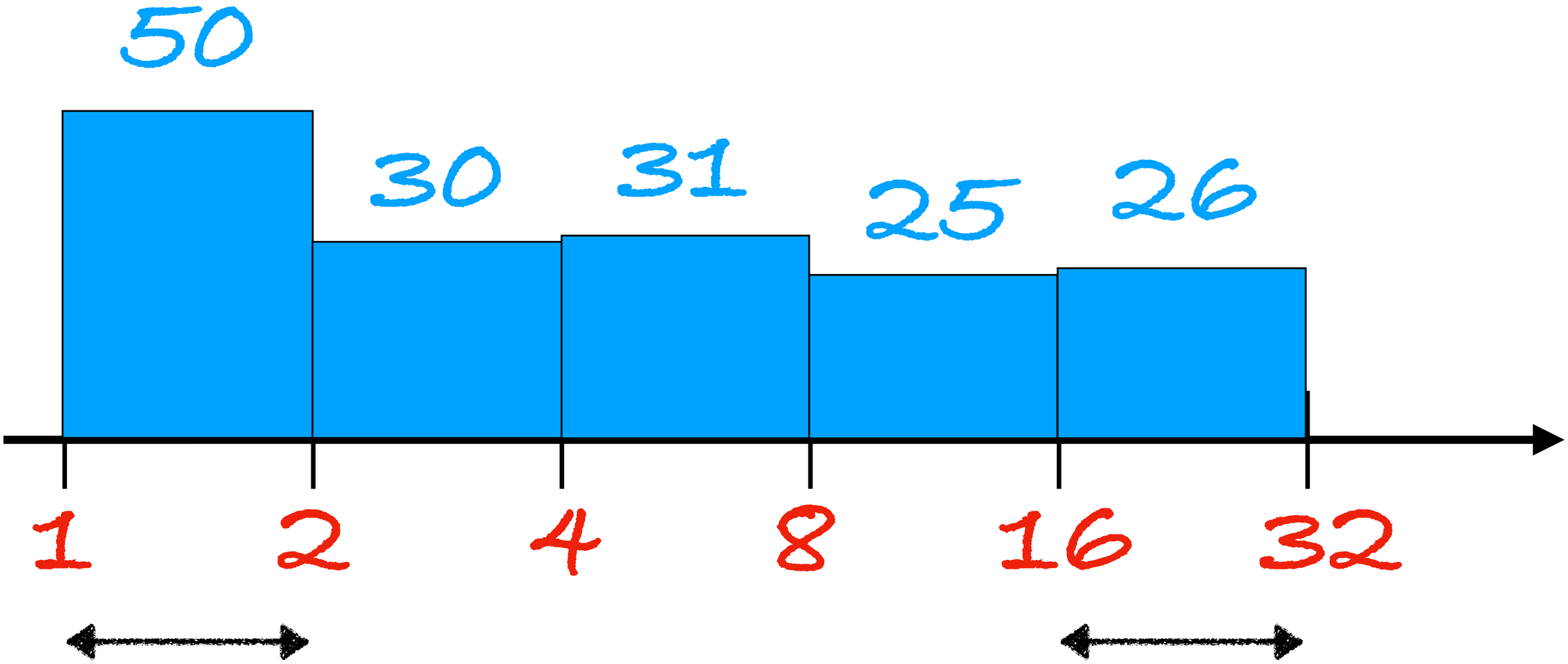

50
30
31
25
26
1
2
4
8
16
32

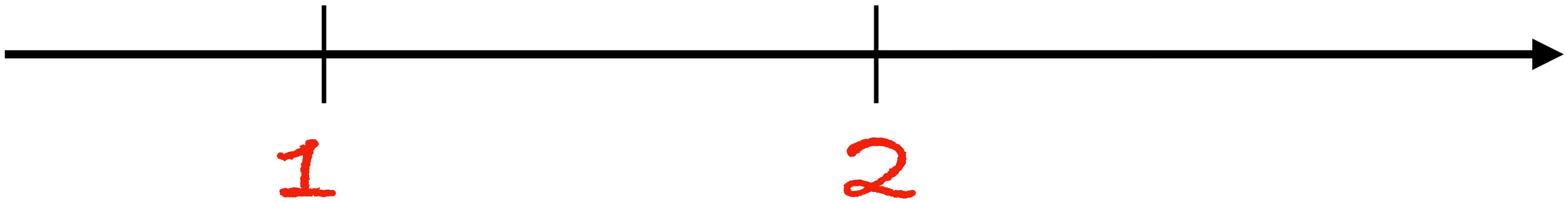

50
30
31
25
26
Wrong!
1
2
4
8
16
32

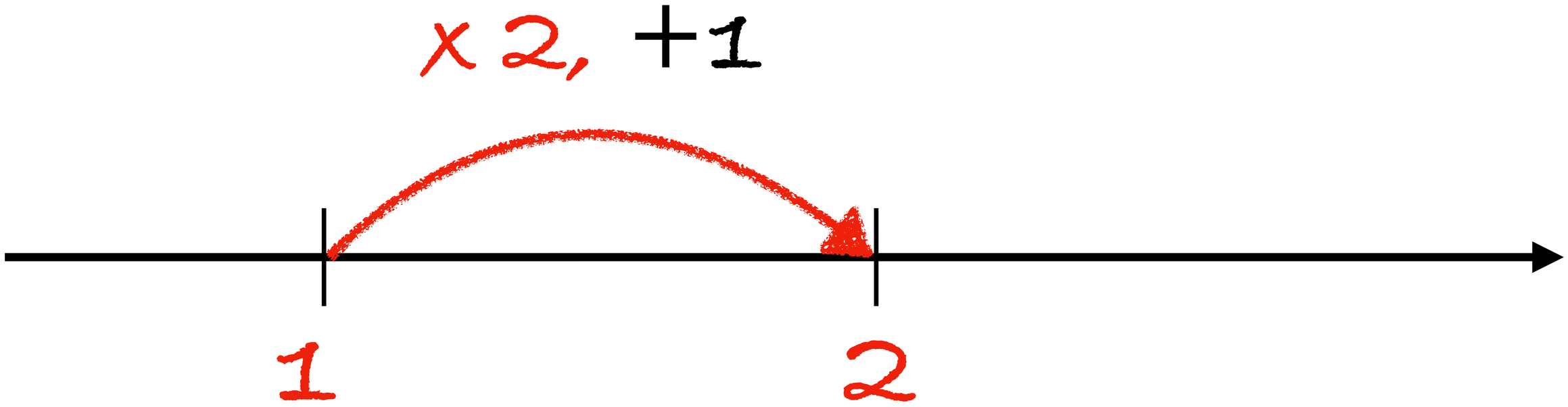
*“**Area**, not the **height**, represents the frequency!”*

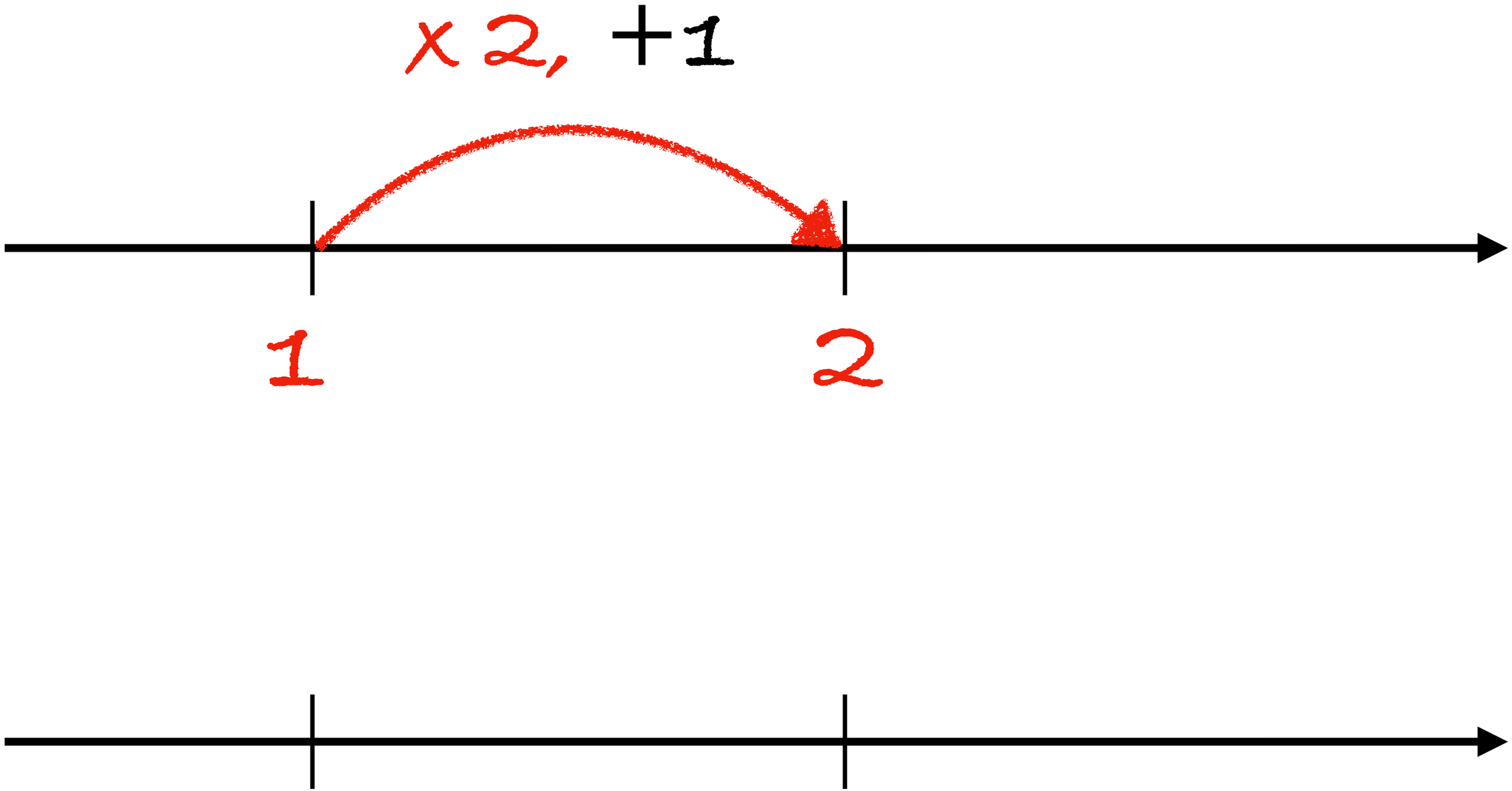

50
30
31
25
26
1
2
4
8
16
32

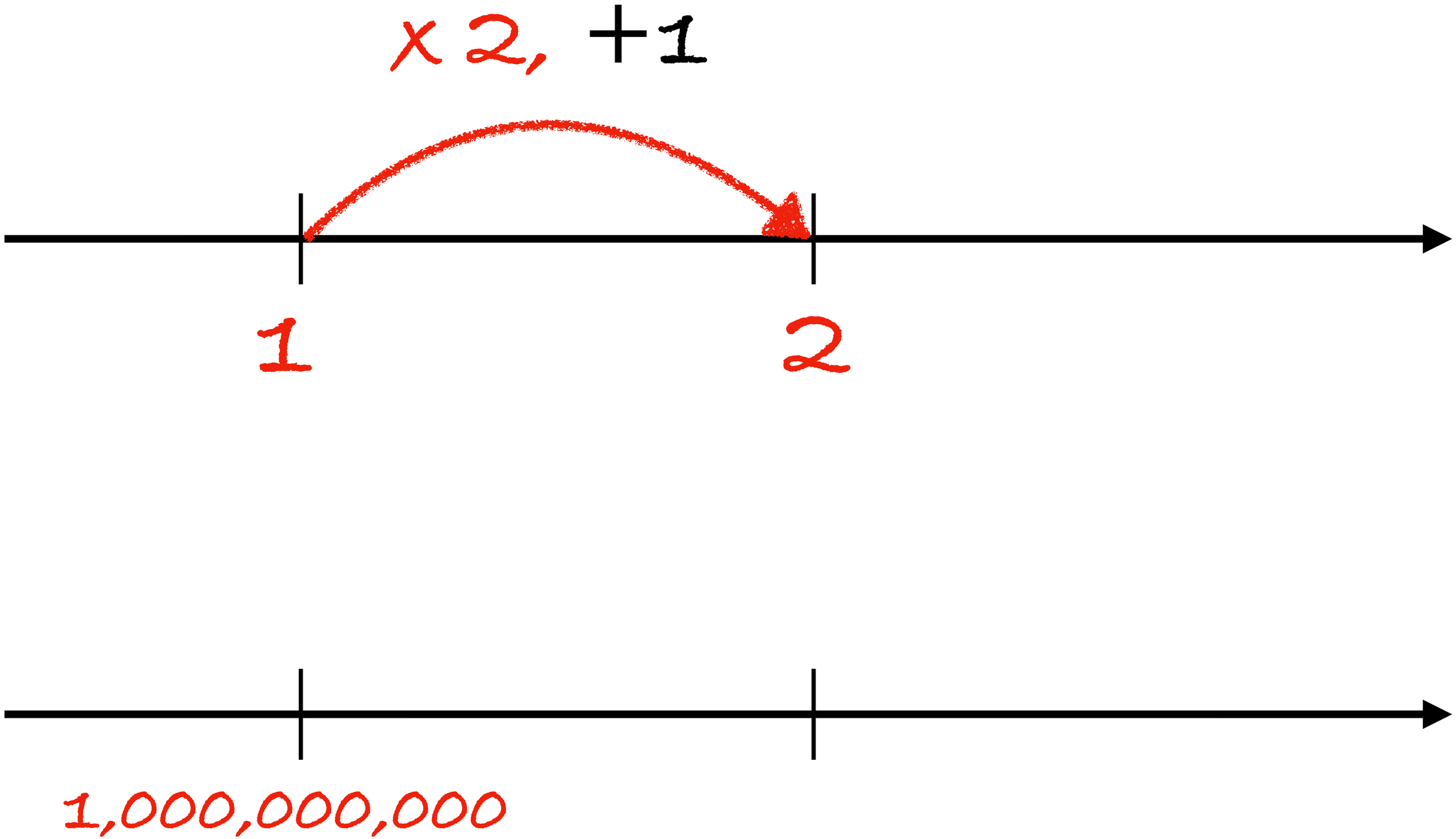

50
30
31
25
26
1
2
4
8
16
32

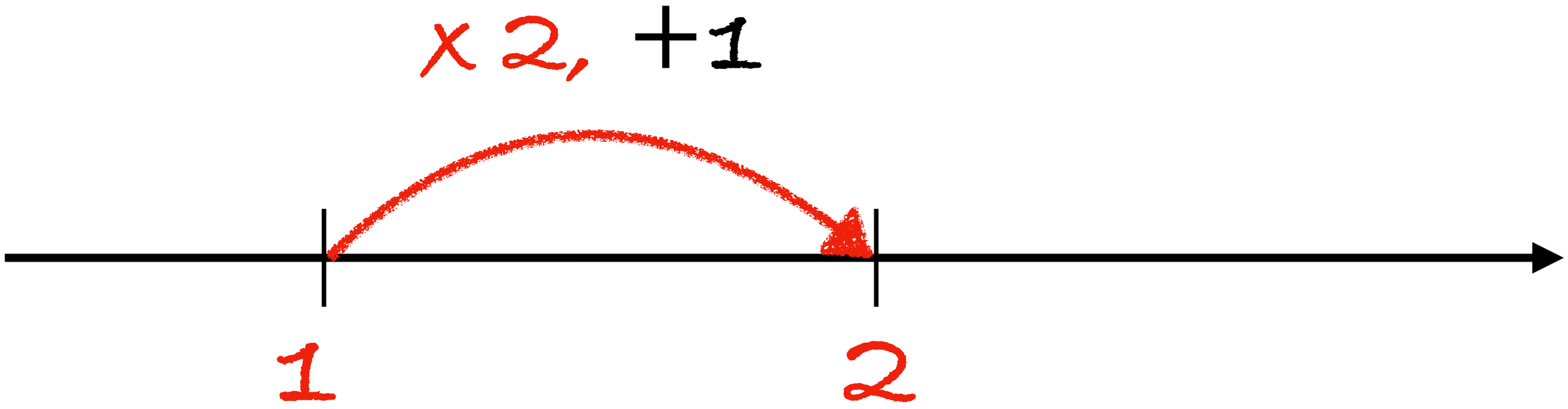

50
30
31
25
26
1
2
4
8
16
32

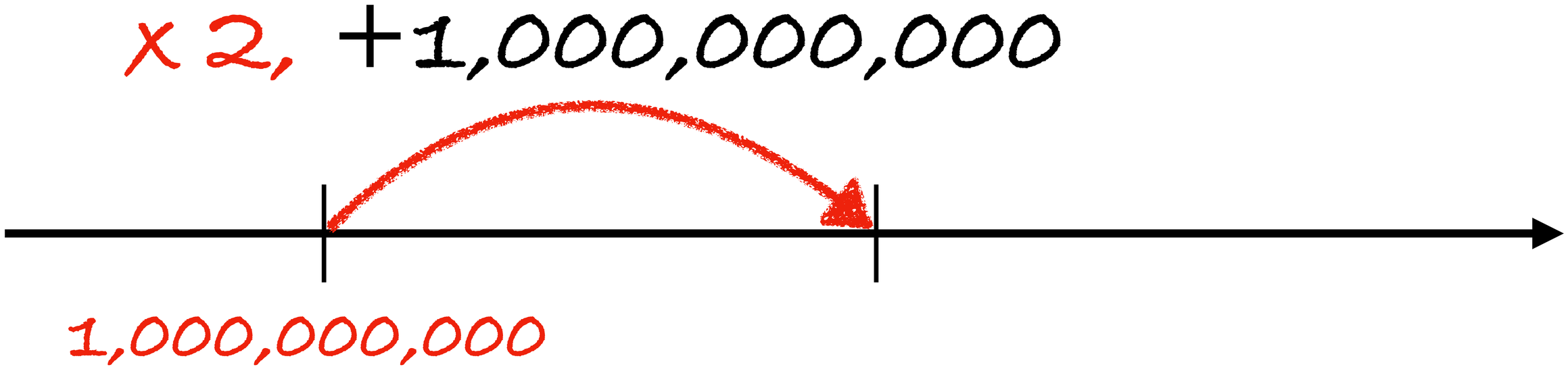

1
2

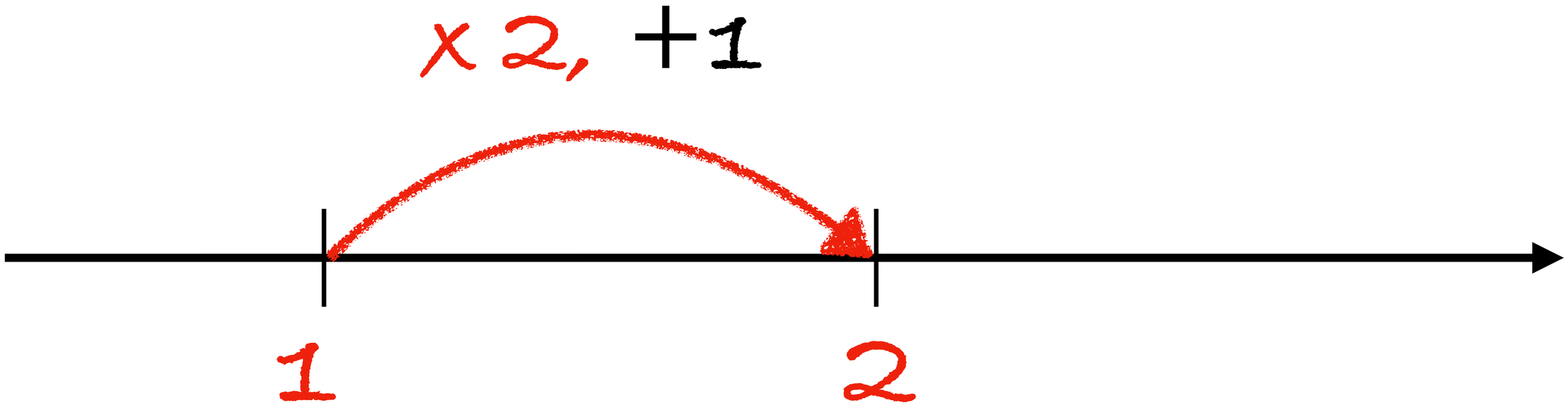

x 2, +1
1
2

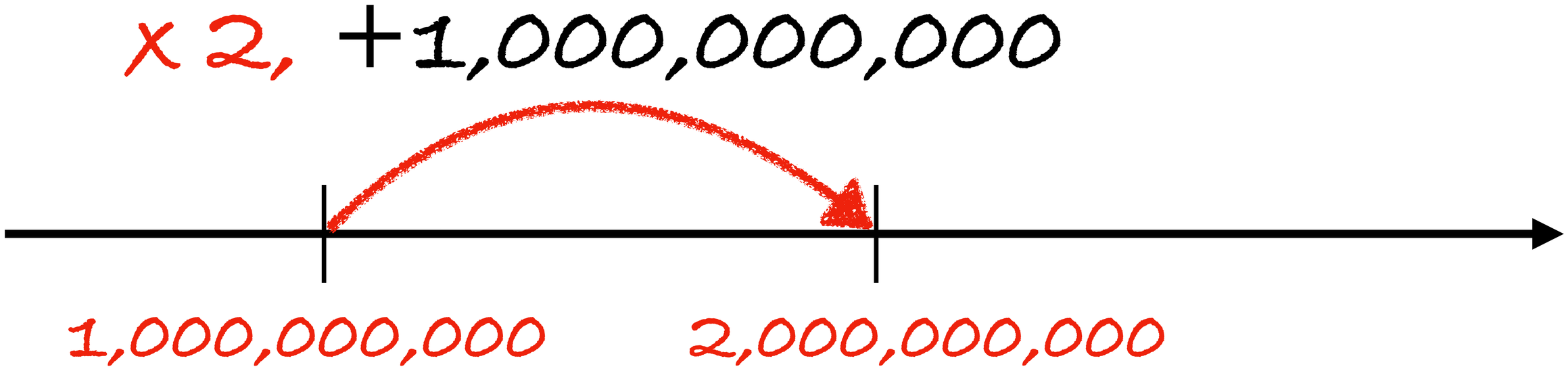

x 2, +1
1
2

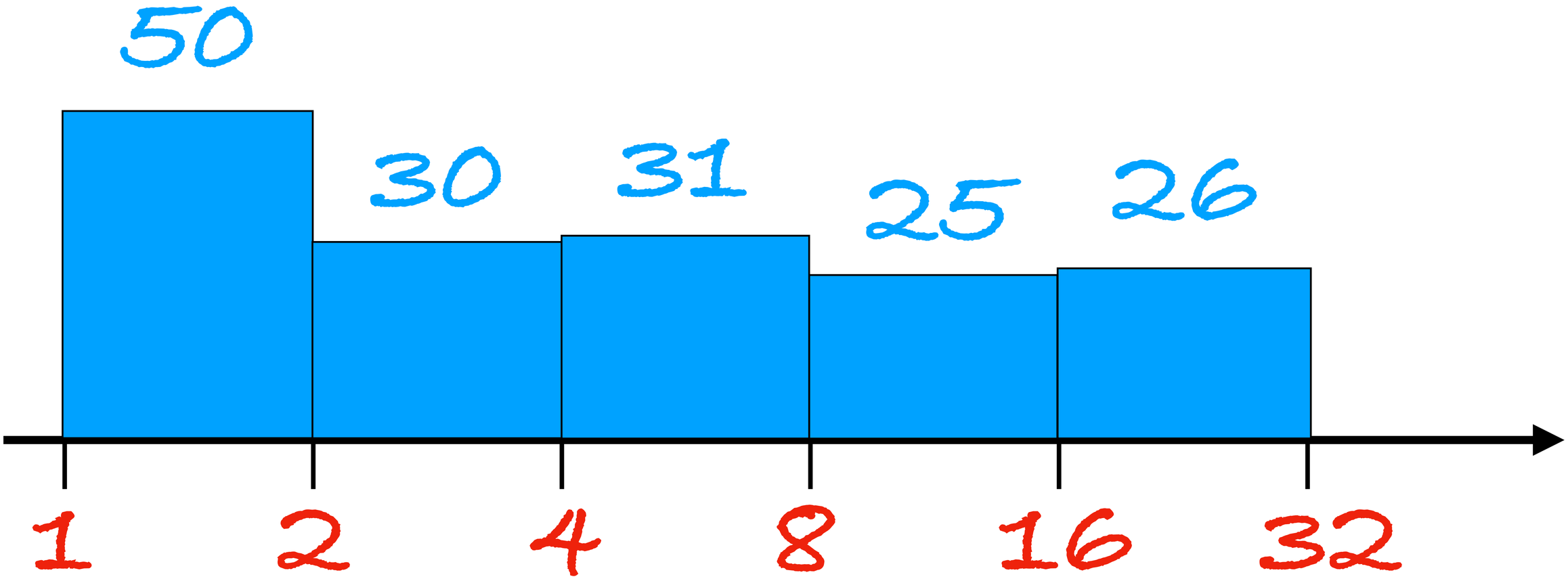

x 2, +1
1
2
1,000,000,000

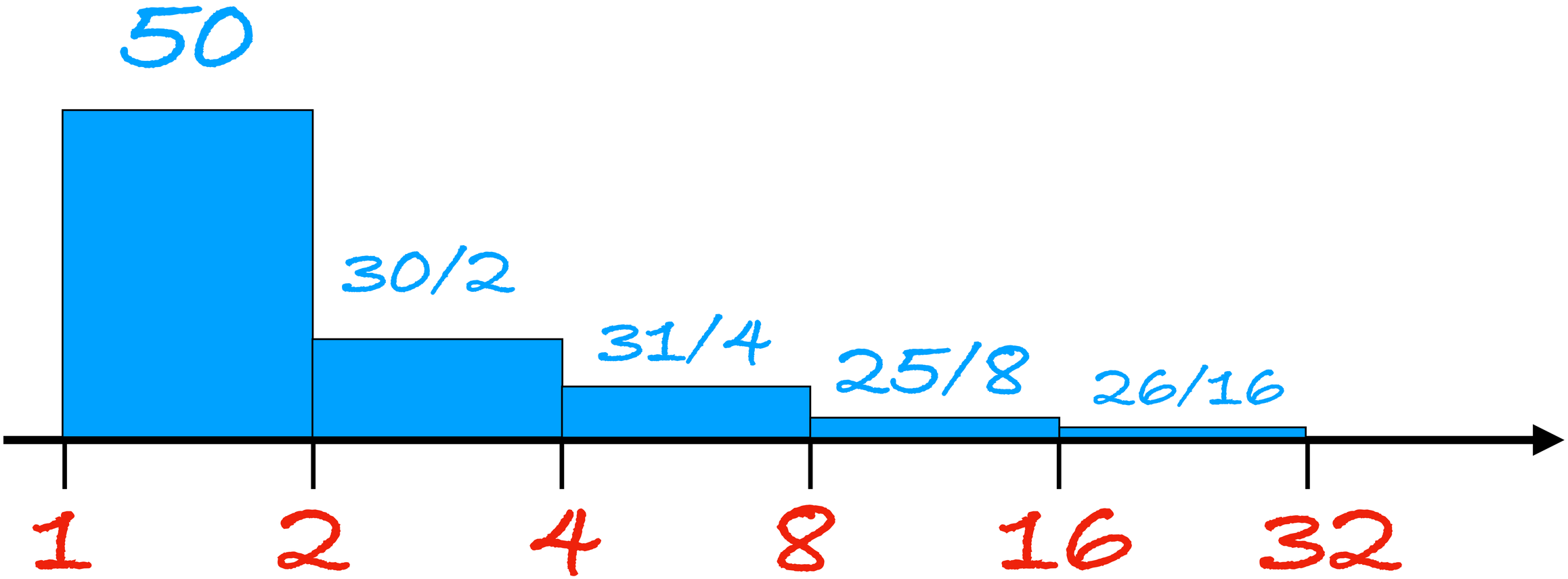

x 2, +1
1
2

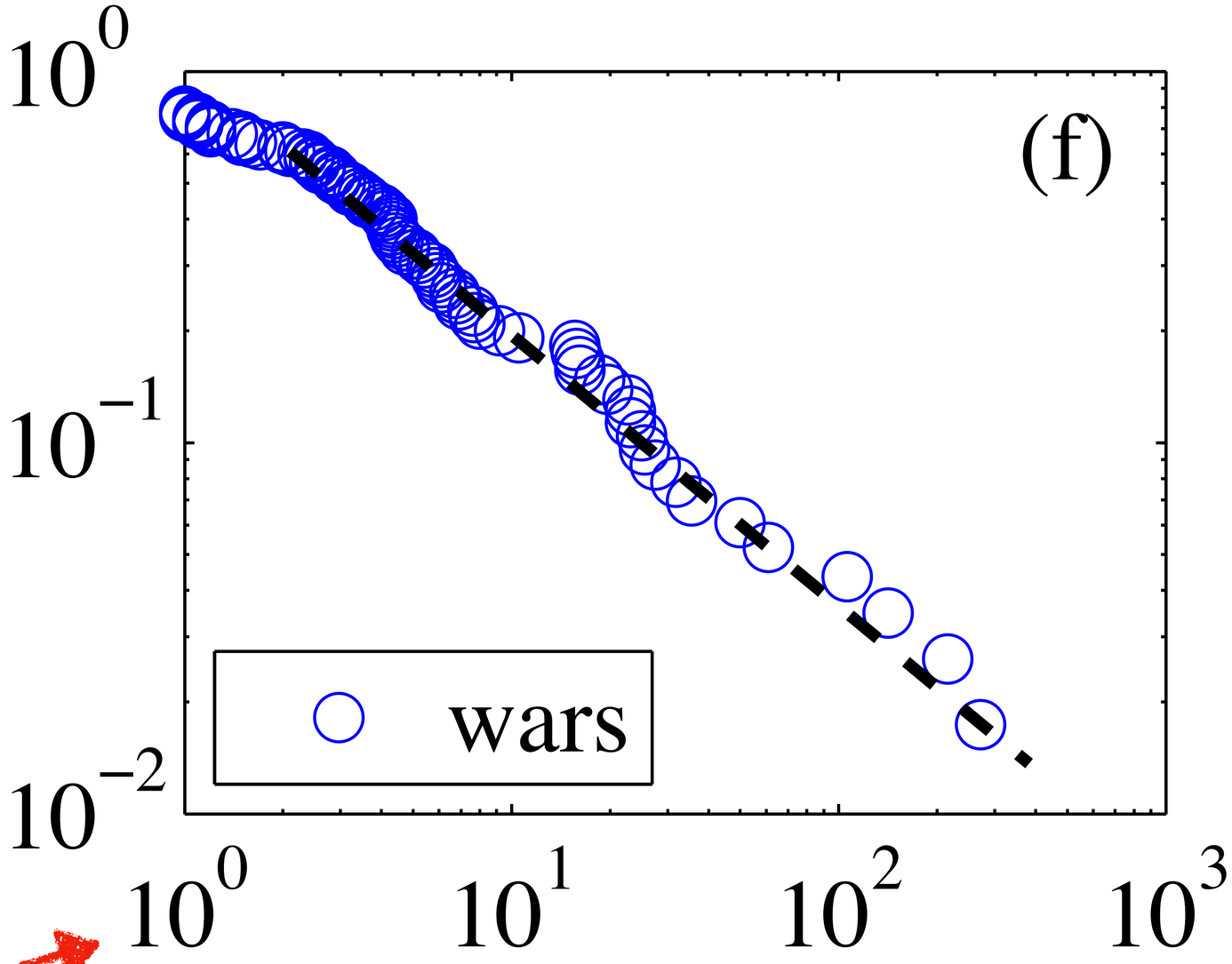

x 2, +1,000,000,000
1,000,000,000


x 2, +1
1
2


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

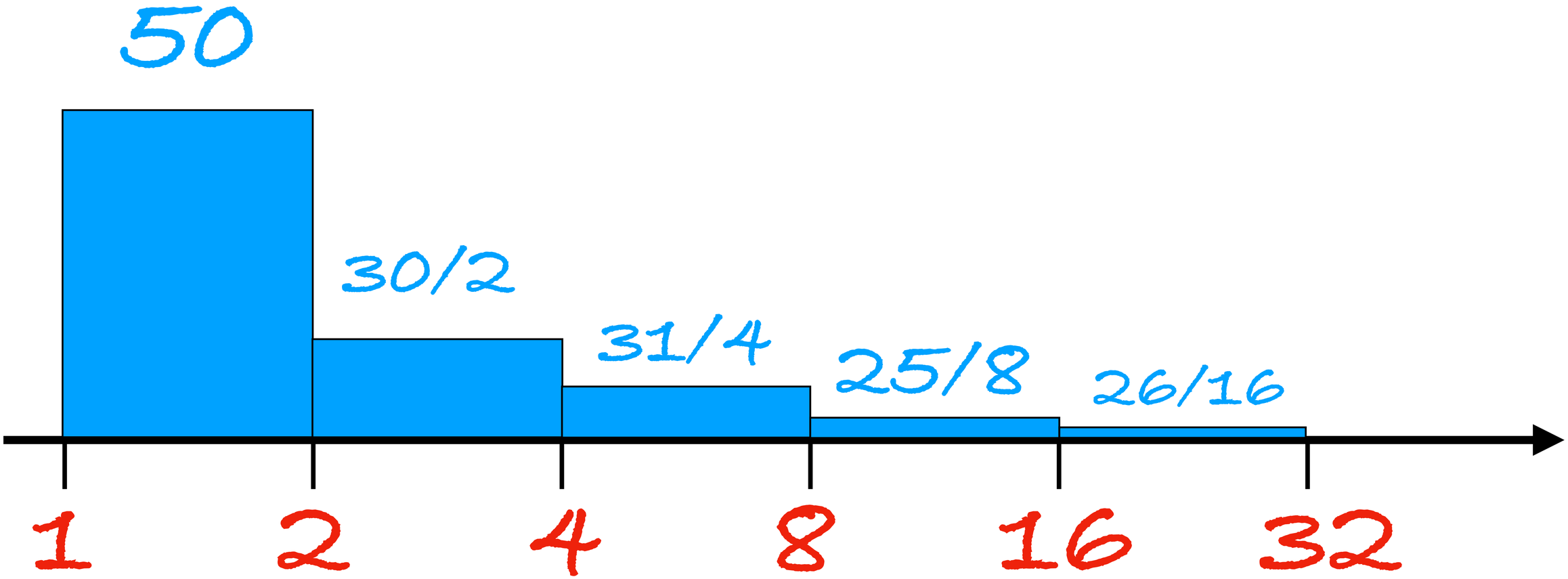

50
30
31
25
26
1
2
4
8
16
32


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

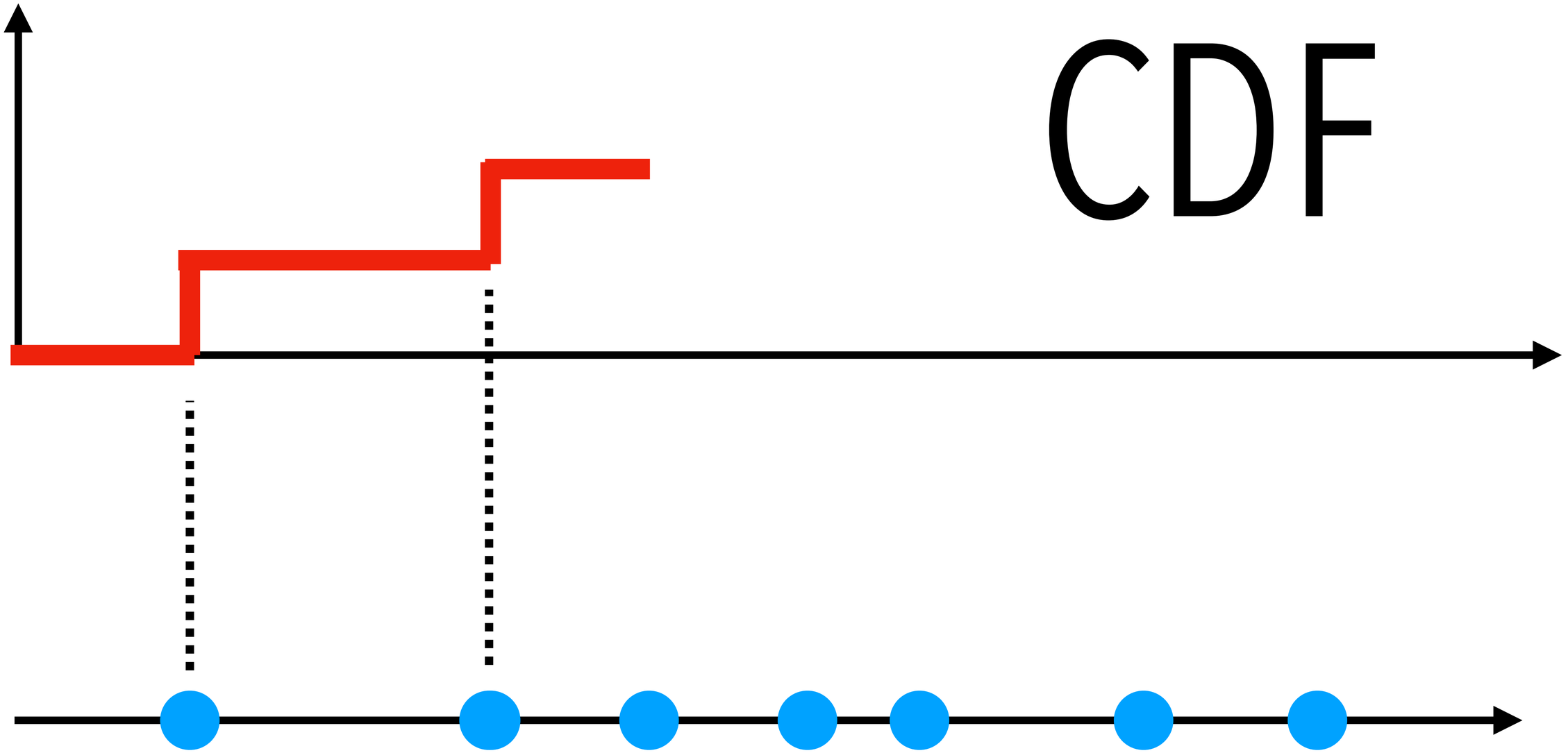


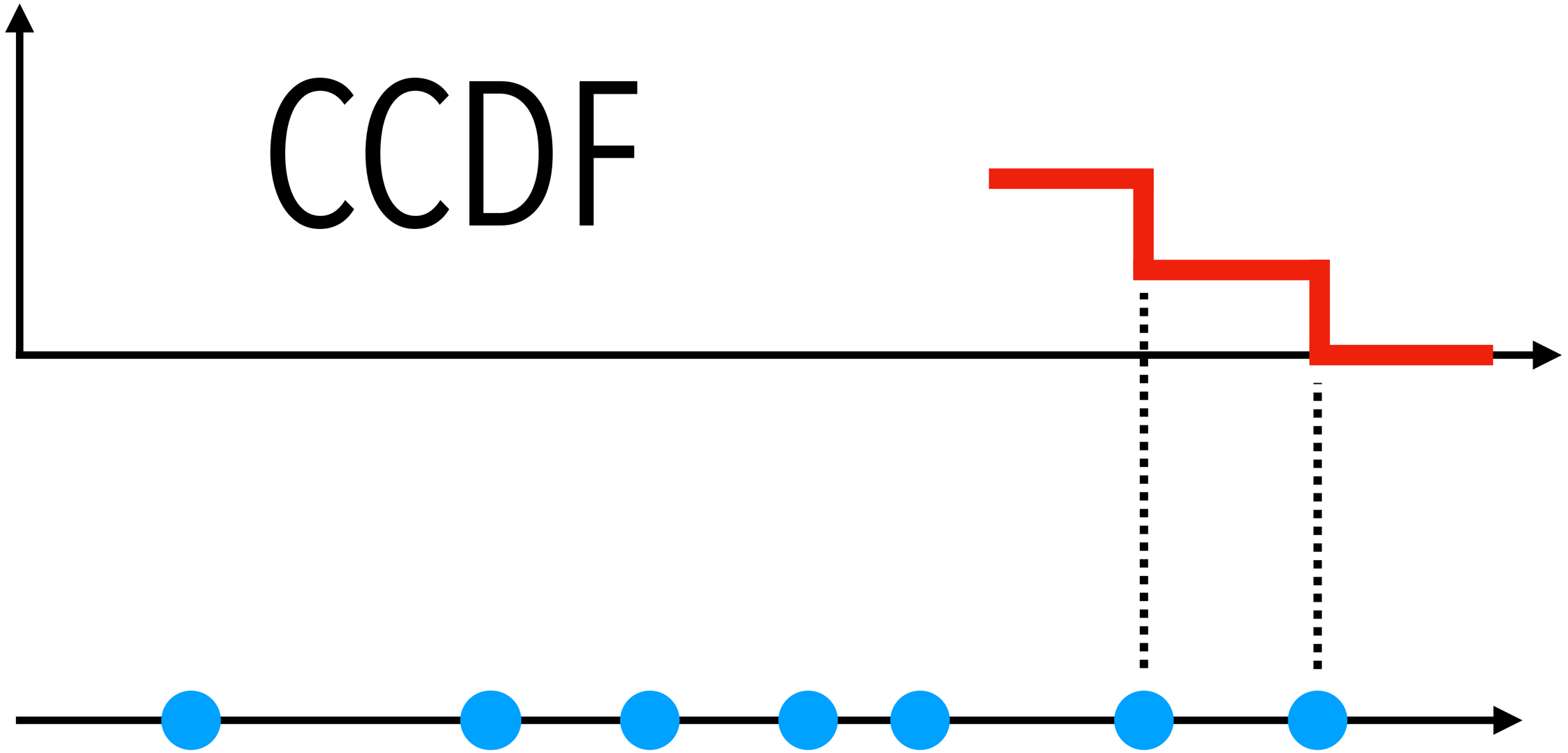
Both in

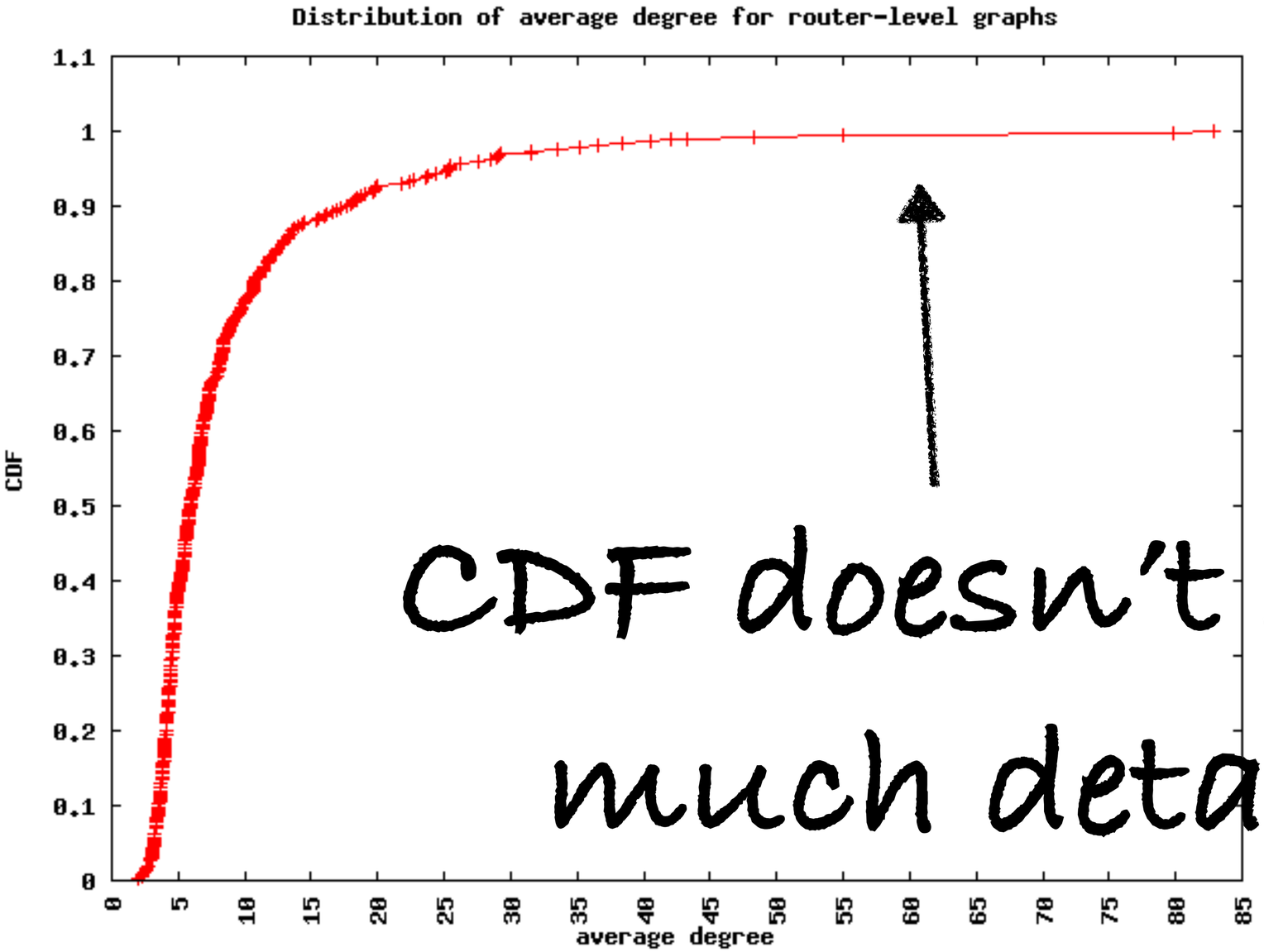
log-scale


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

A nice alternative

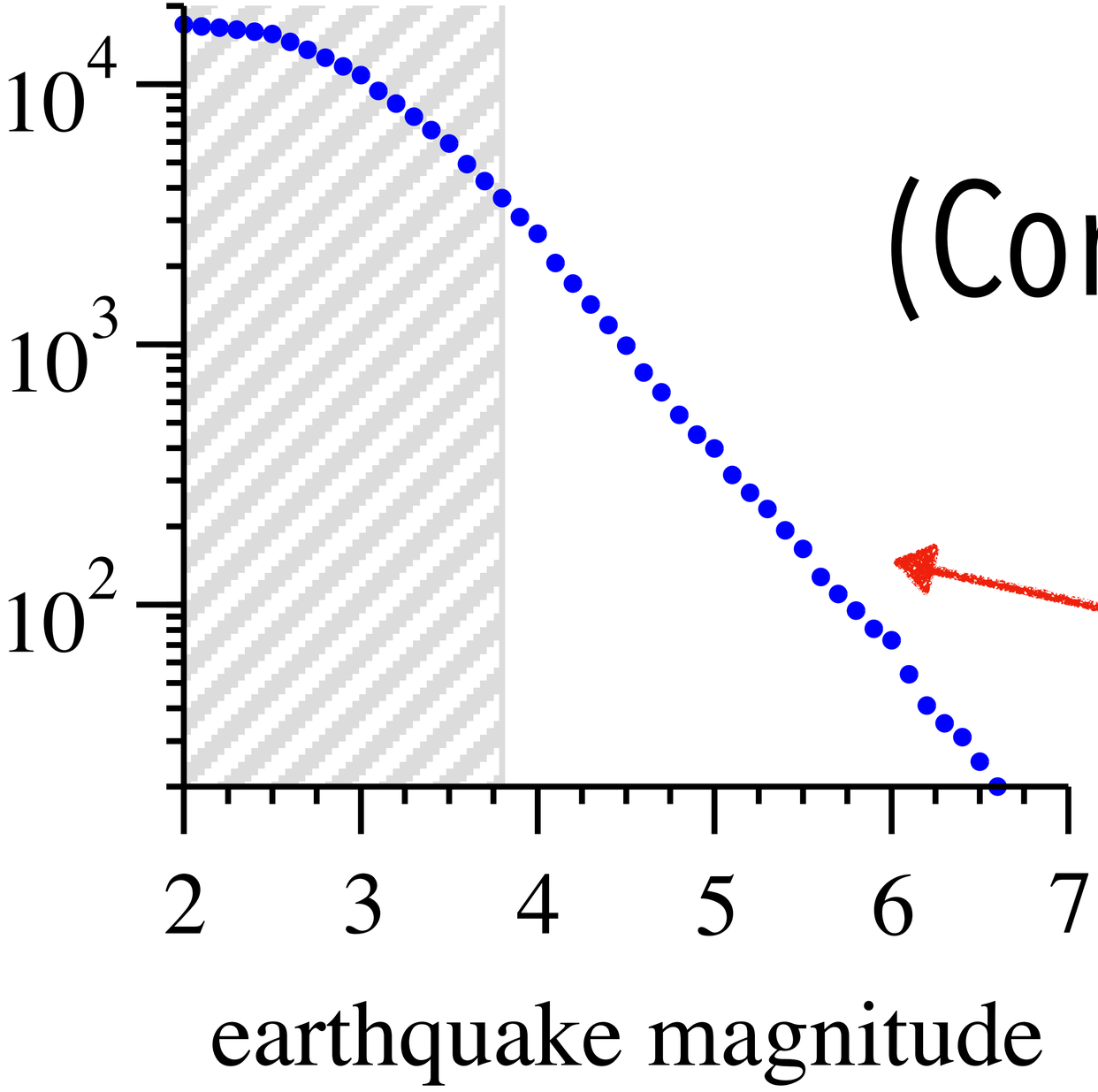

CDF


CCDF



CDF doesn't show much details

CCDF
(Complementary CDF)



more useful

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

CCDF is often much more useful than **CDF.**

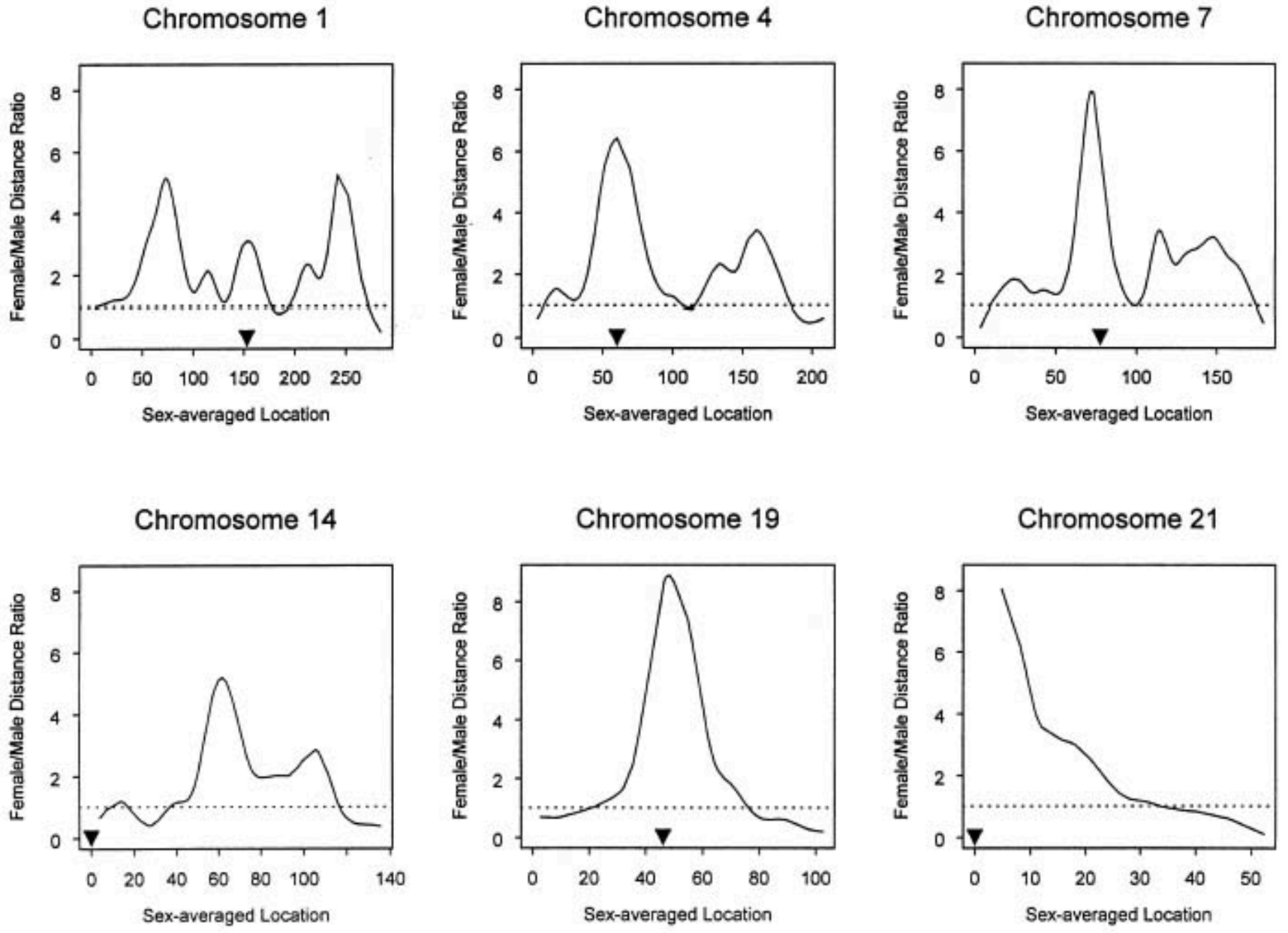
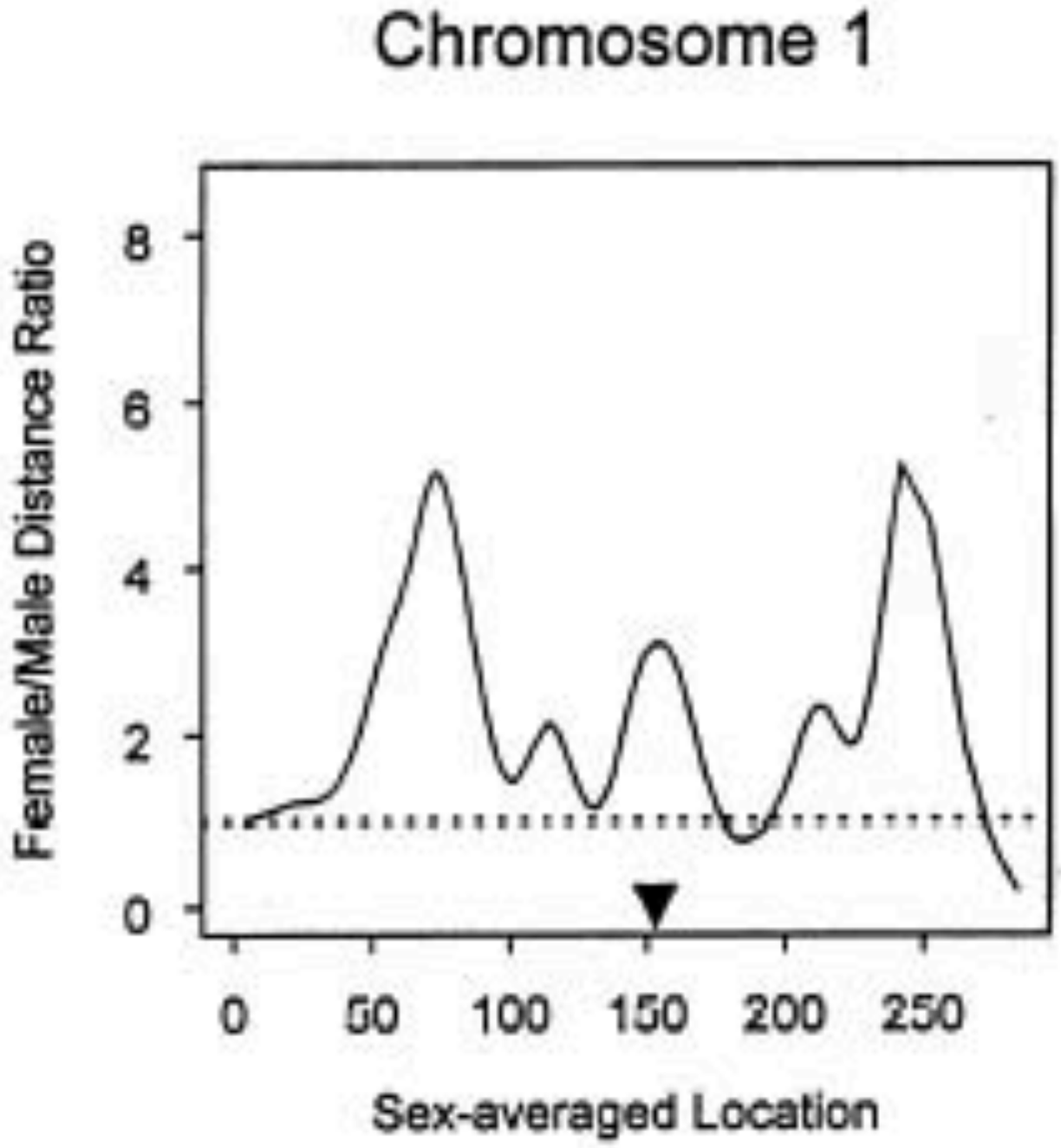
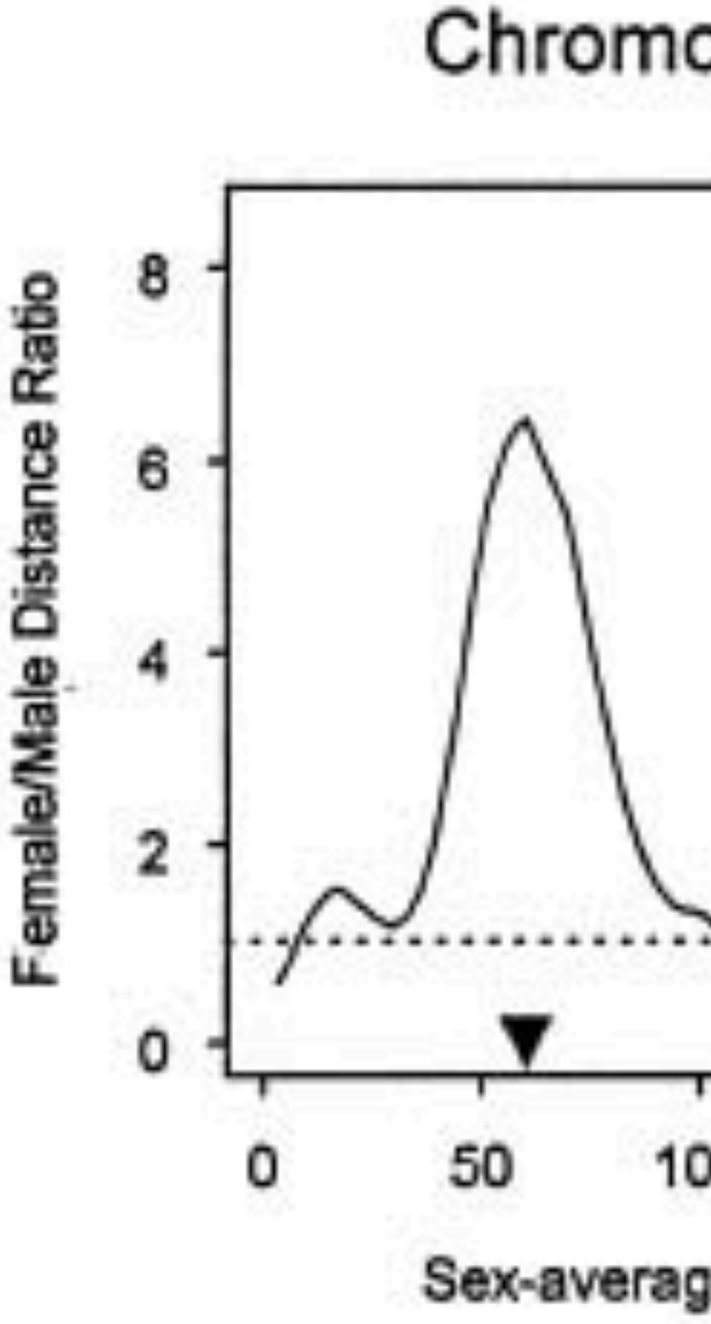
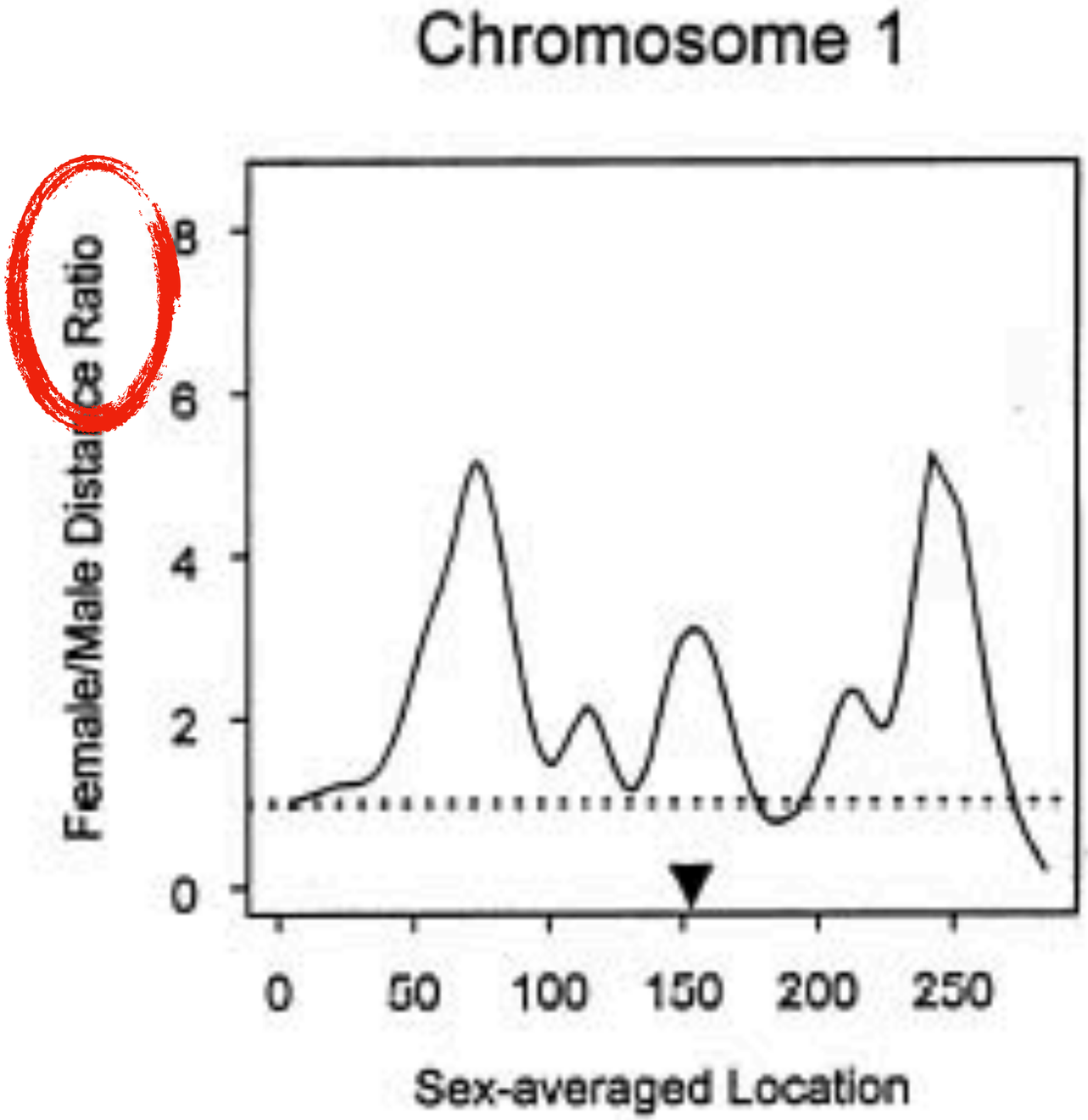
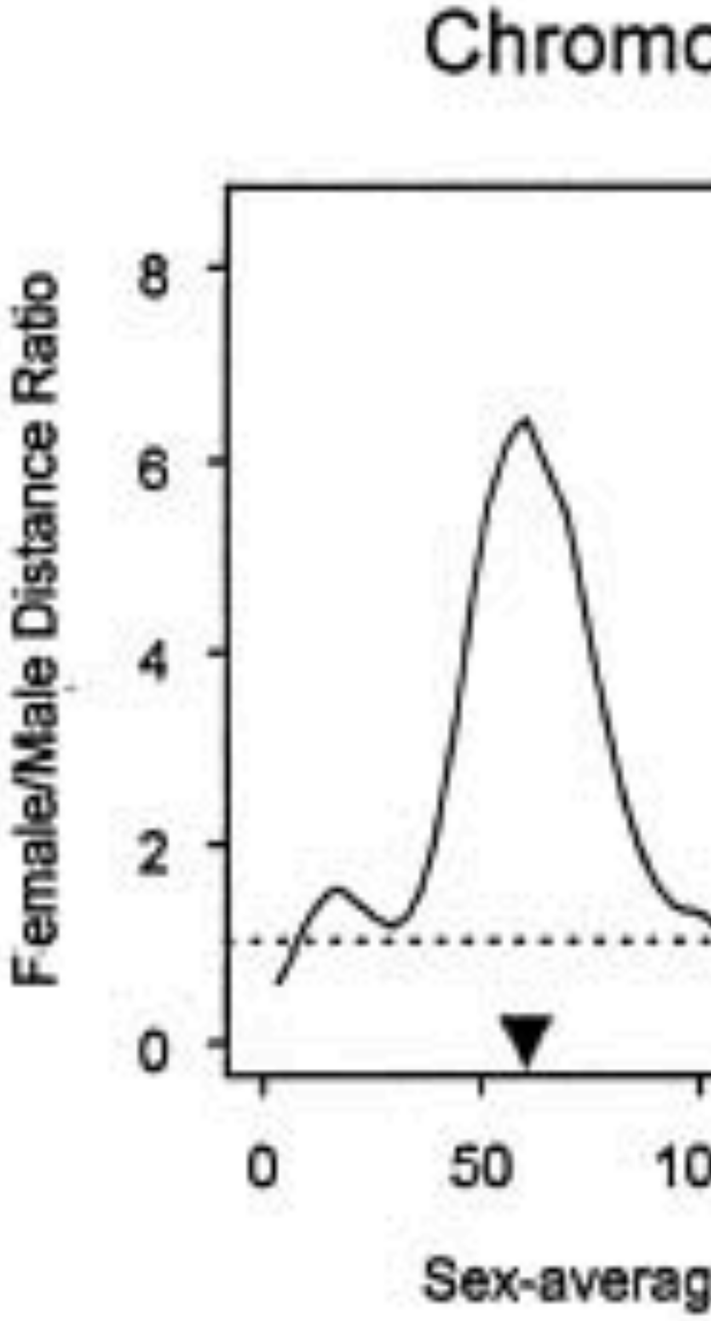


Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

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


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’s wrong with this plot?

What is a ratio?

$$A : B$$

$$A : B$$

$$A / B$$

$$A : B$$

Numerator

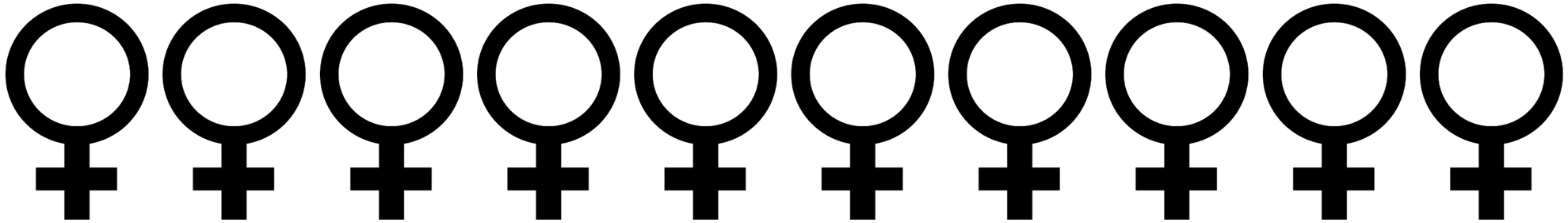
$$A / B$$

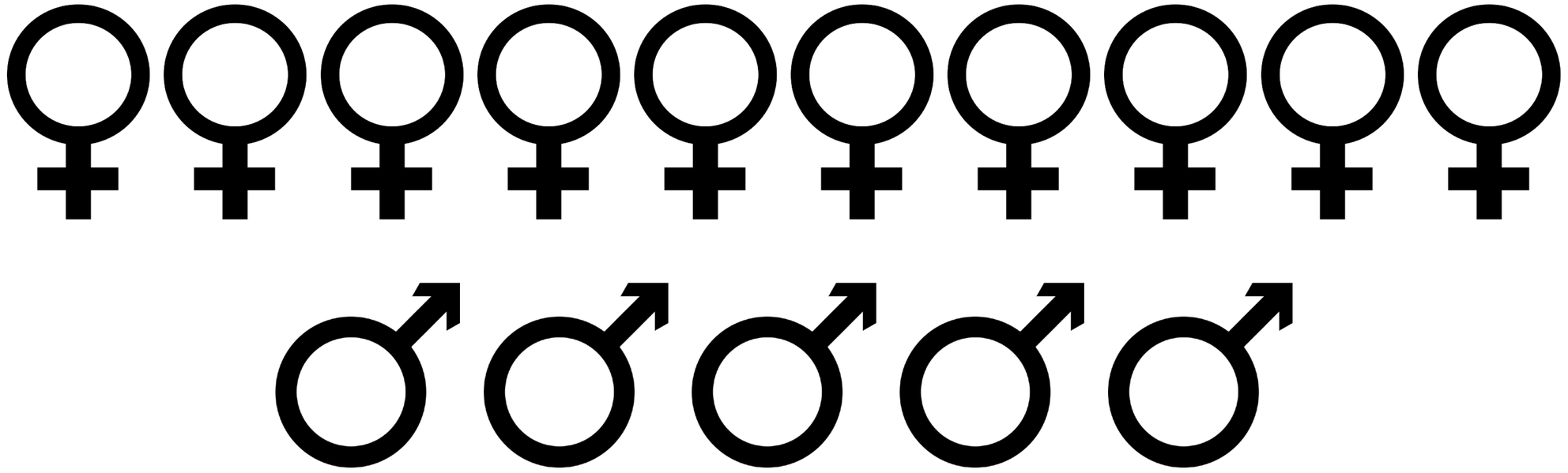
$$A : B$$

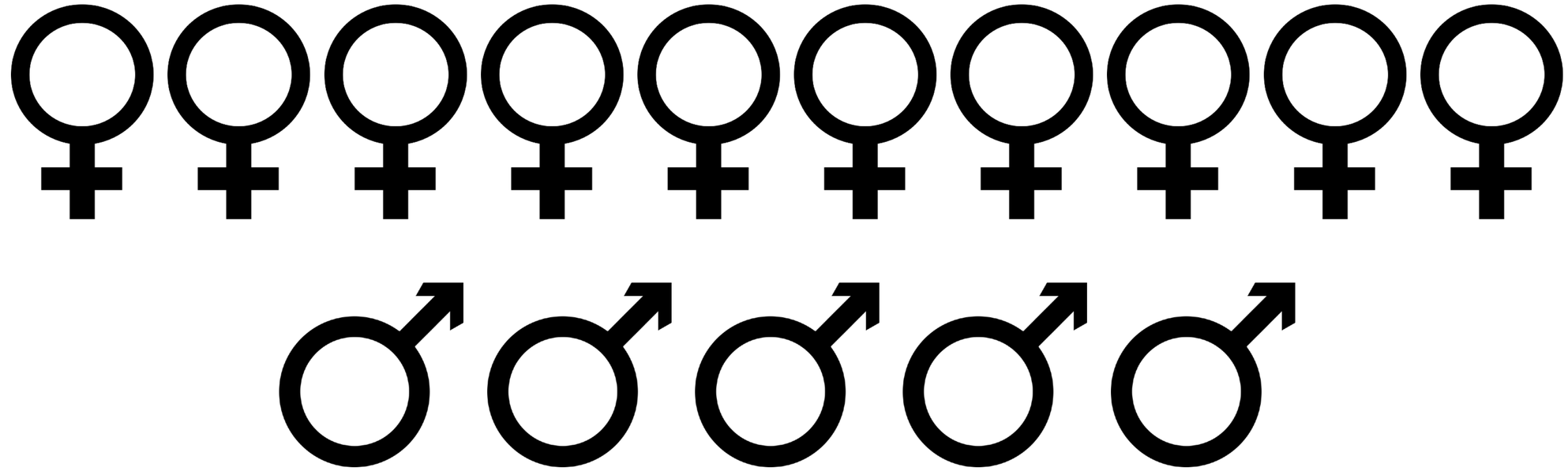
Numerator

Denominator

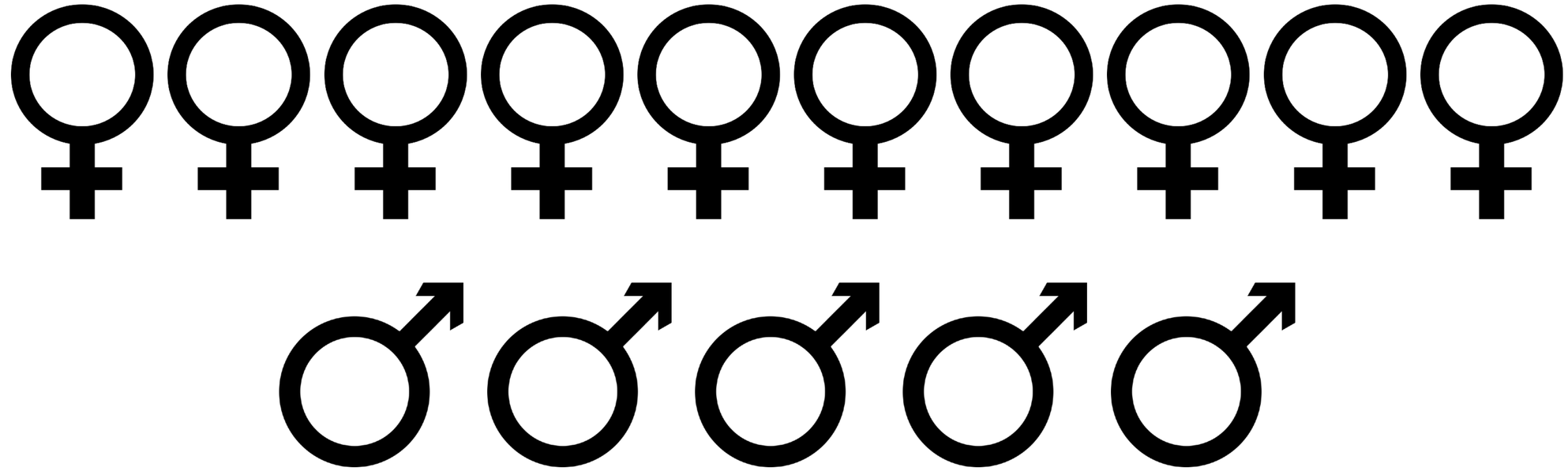
$$A / B$$



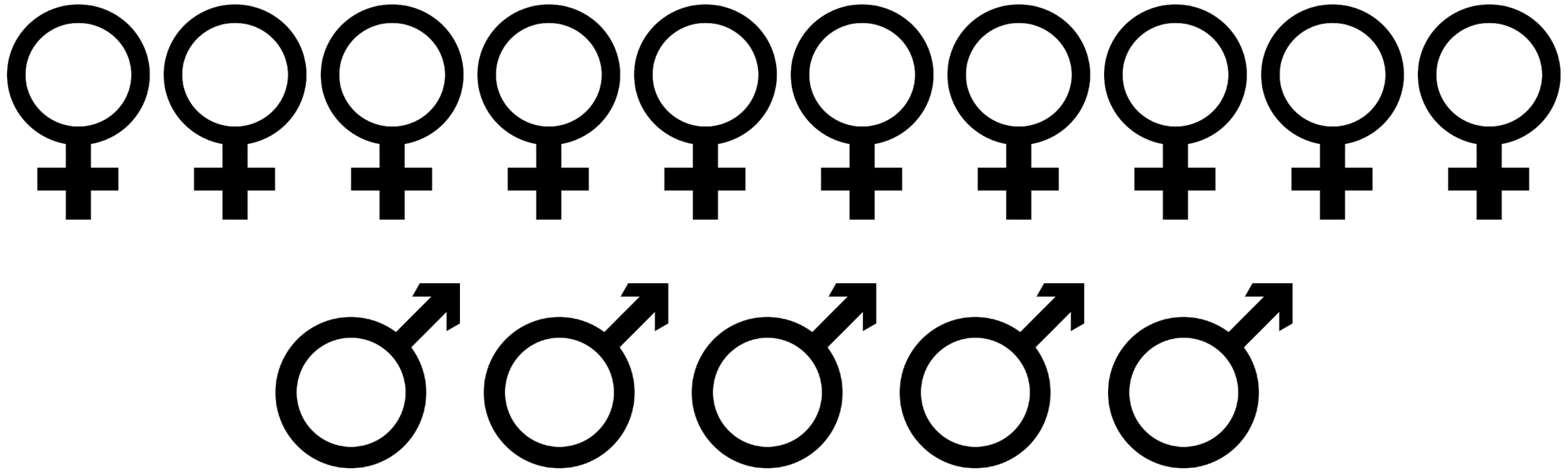




10:5



10:5 2:1



10:5 2:1 10/5 = 2

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀♀♀♀♀♀

♀♀♀♀♀

♂♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀

♂♂♂♂♂♂♂♂♂♂♂

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀

♂♂♂♂♂♂♂♂♂♂

5:10

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀

♂♂♂♂♂♂♂♂♂♂

5:10

1:2

♀♀♀♀♀♀♀♀♀♀

♂♂♂♂♂

10:5

2:1

10/5 = 2

♀♀♀♀♀

♂♂♂♂♂♂♂♂♂♂

5:10

1:2

5/10 = 0.5

♀♀♀♀♀♀♀♀♀♀

♂

10:1

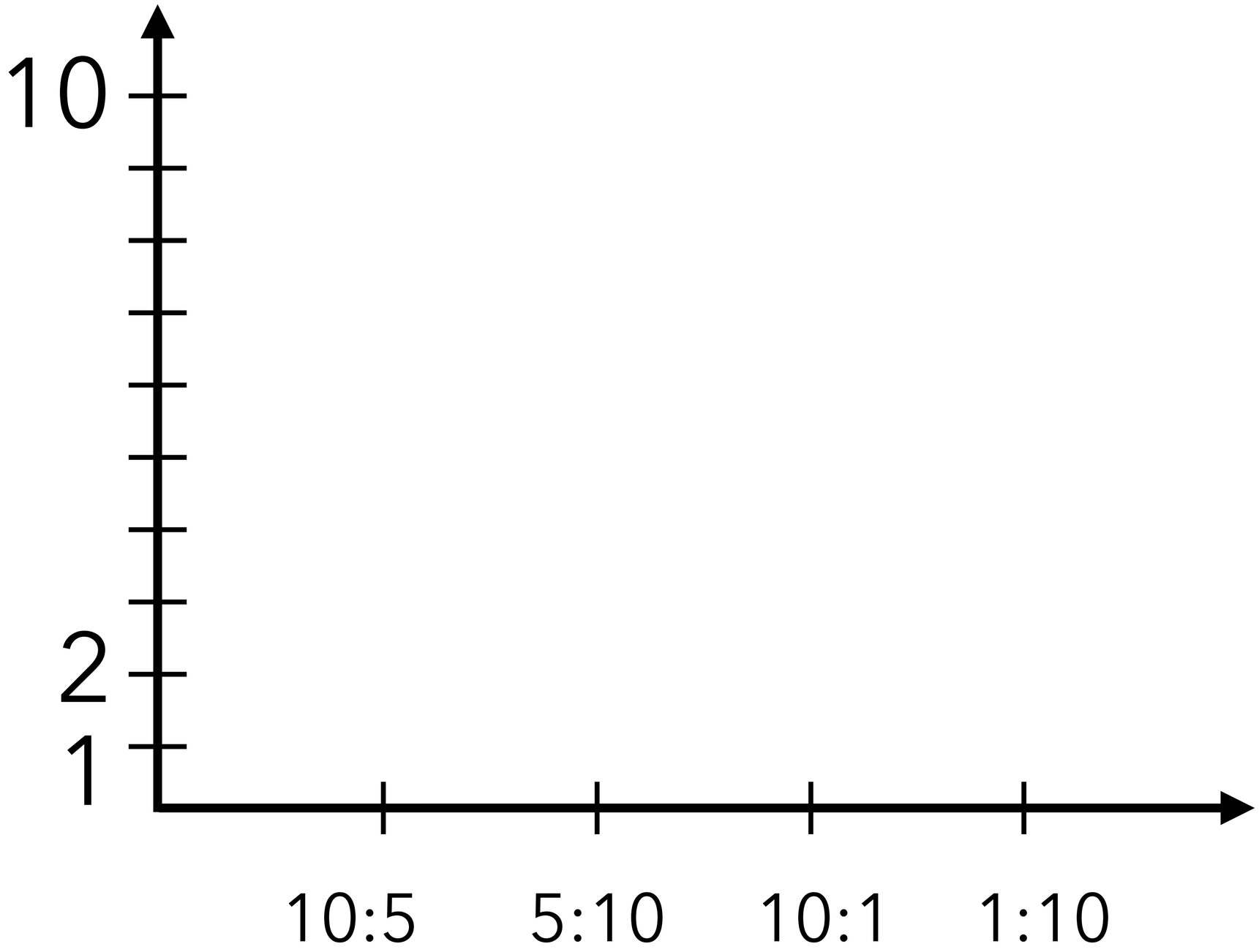
10

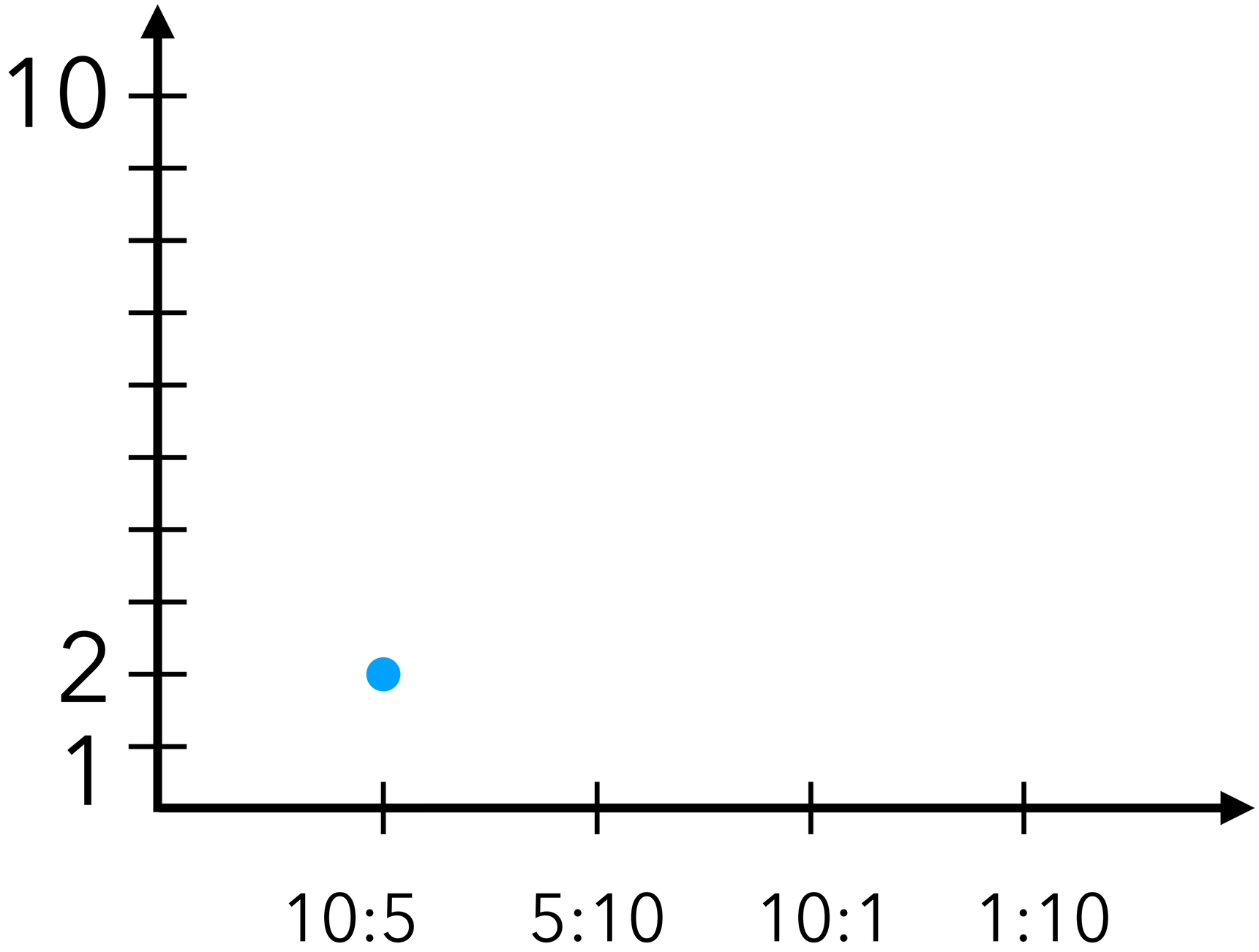
♀

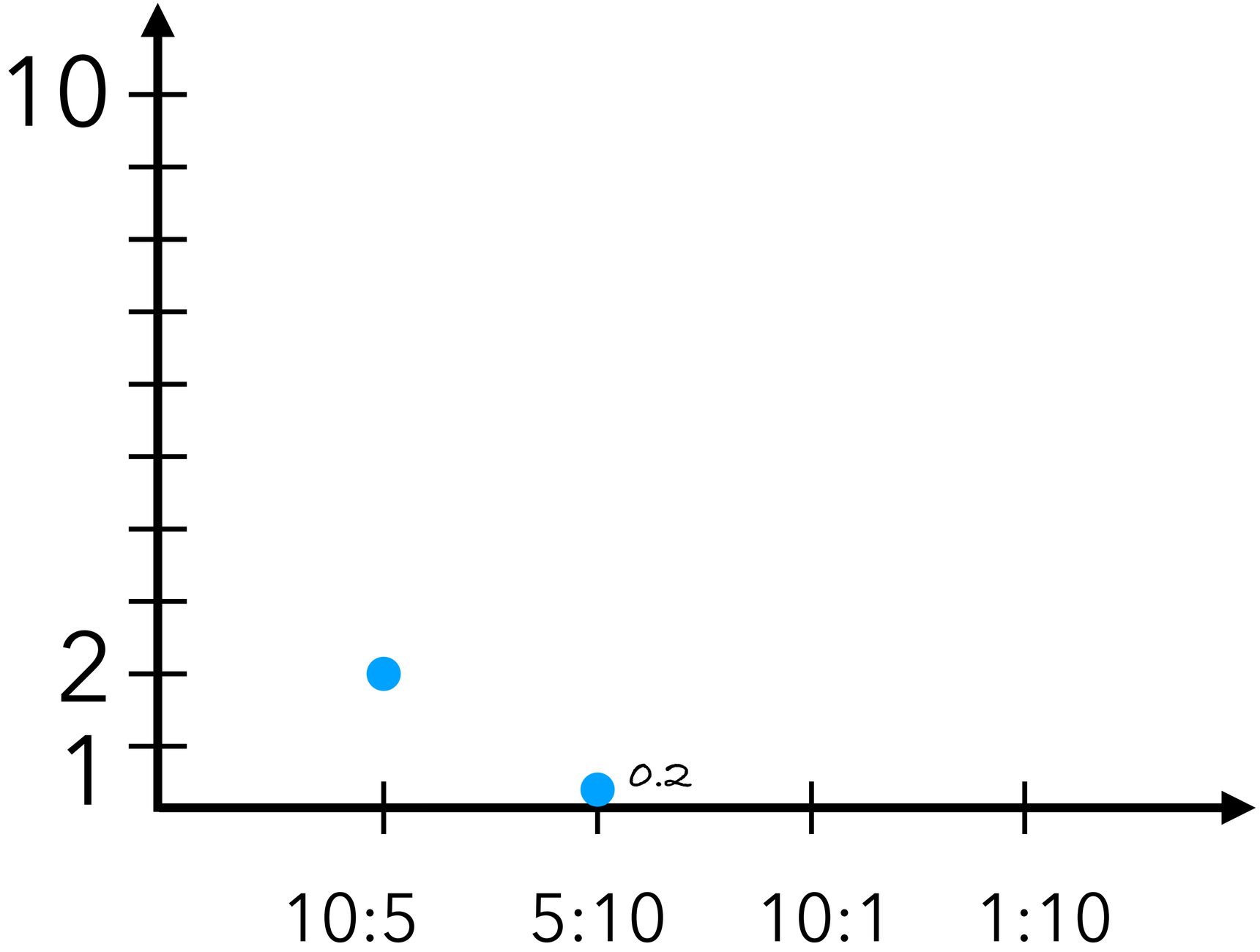
♂♂♂♂♂♂♂♂♂♂

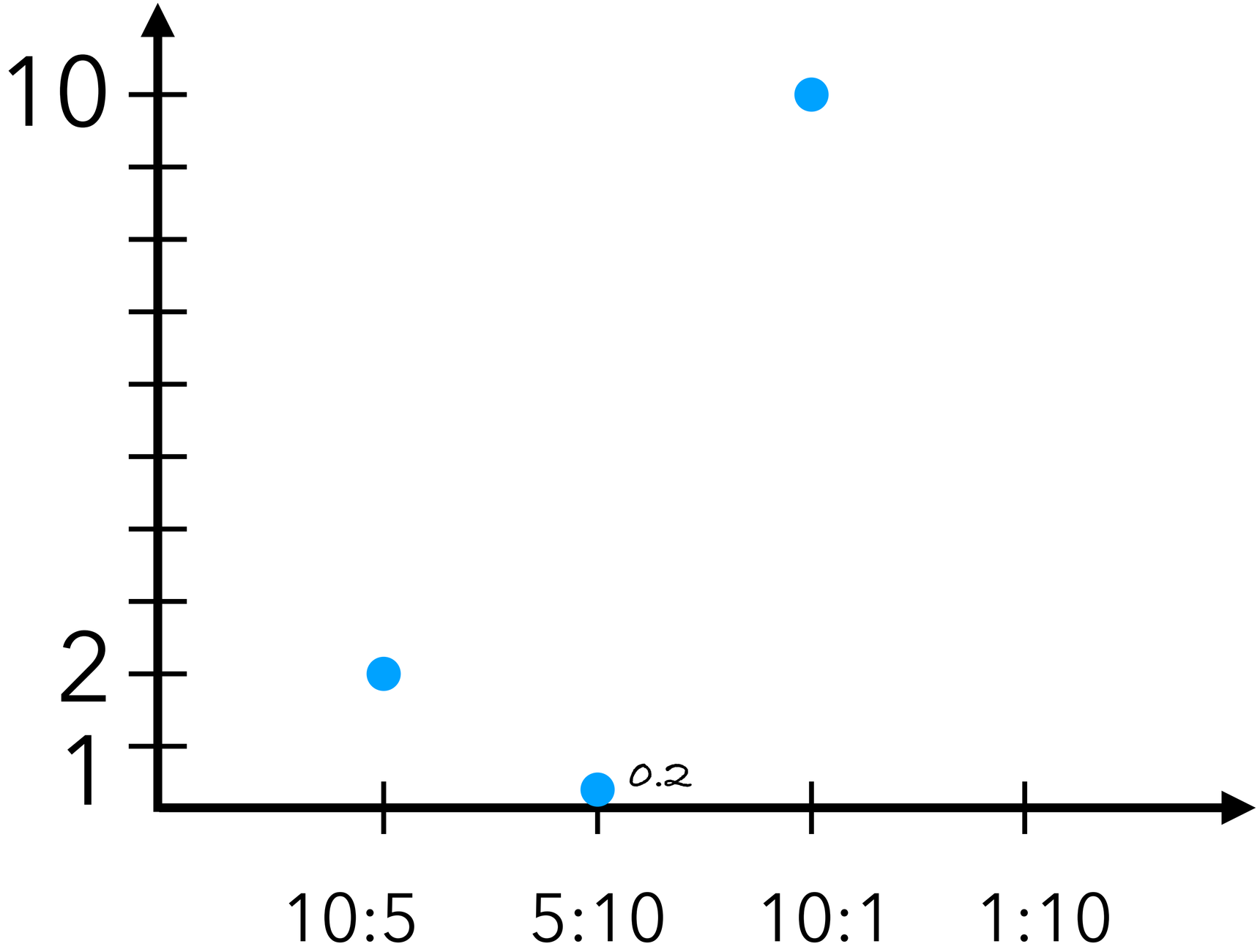
1:10

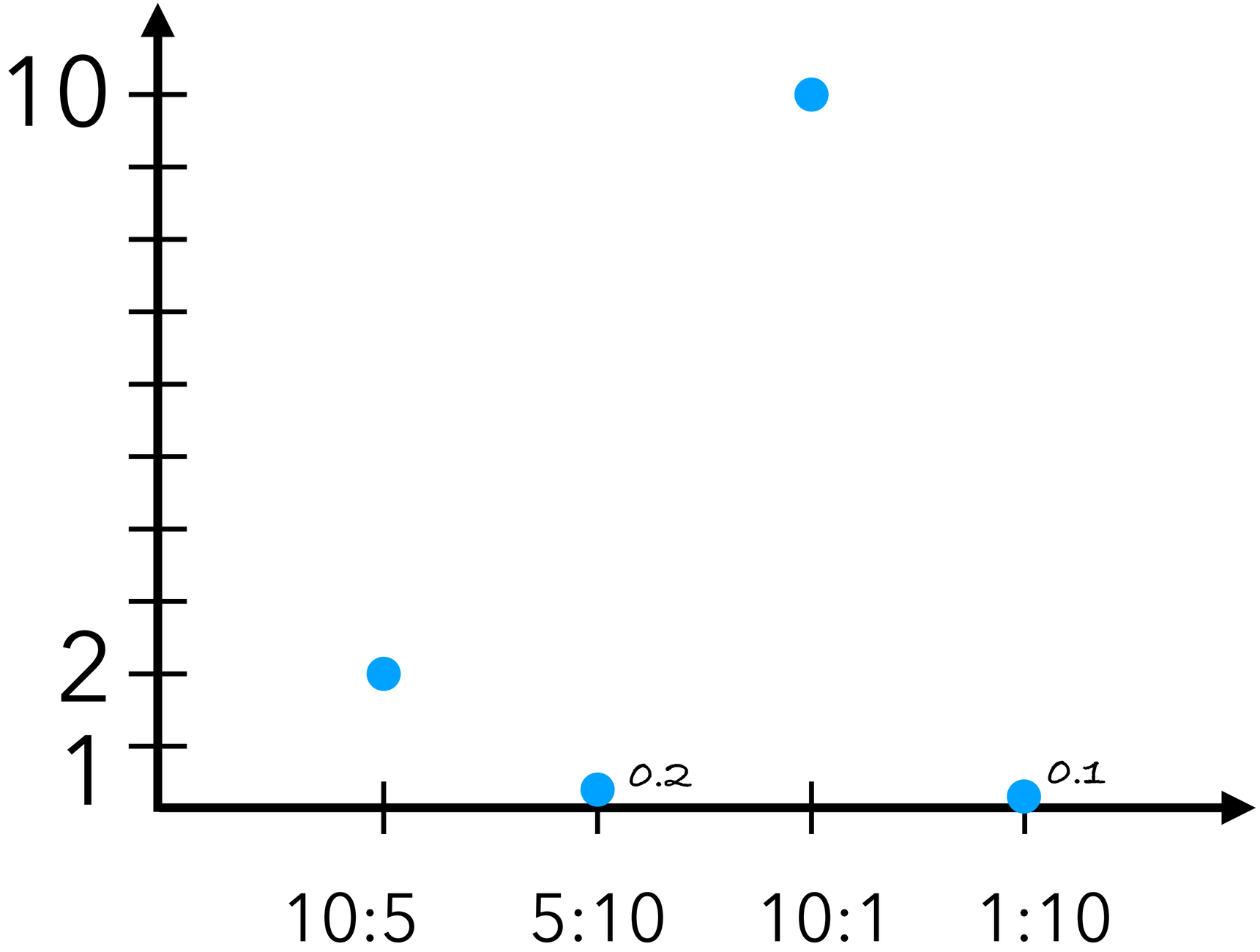
0.1

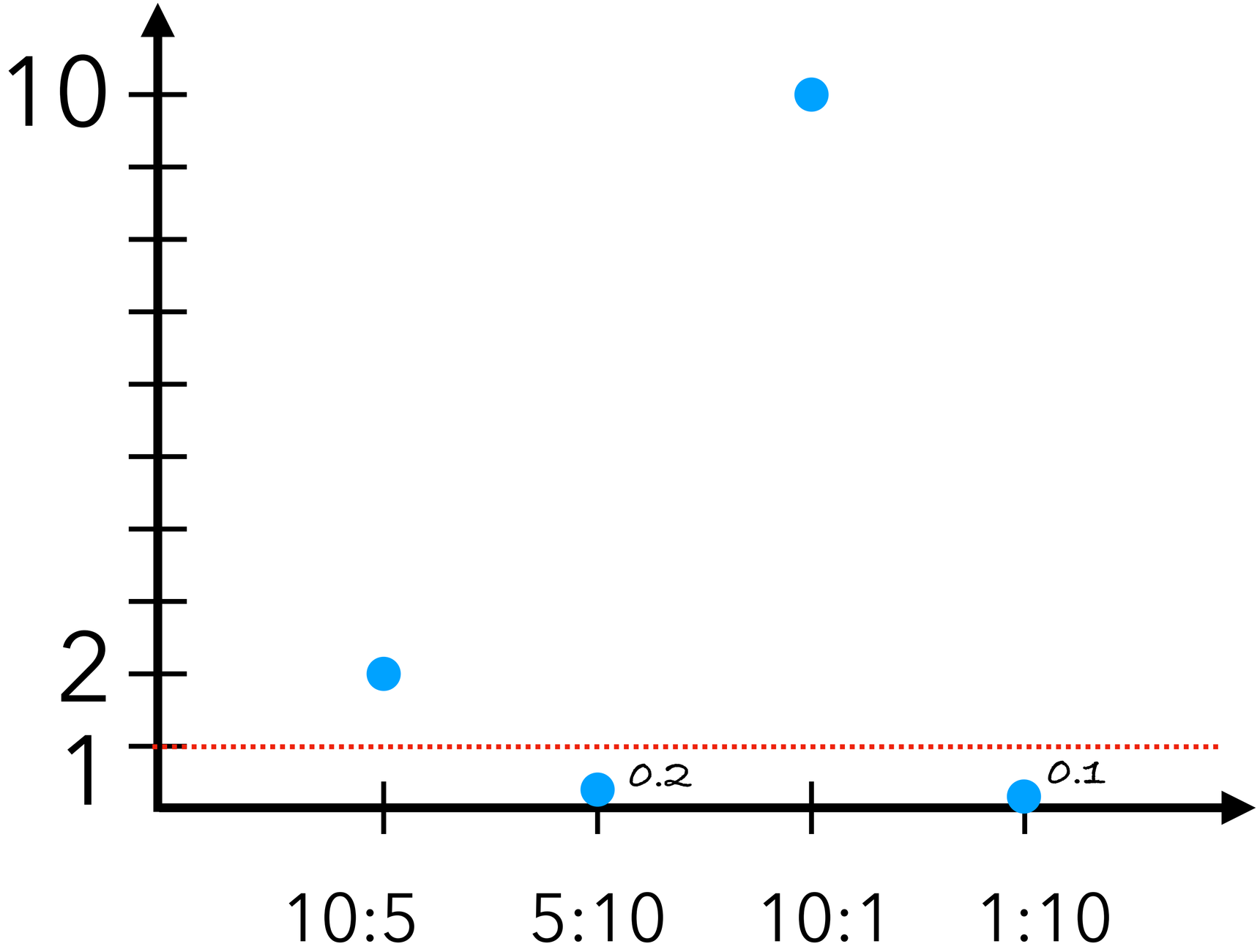

10
2
1
10:5
5:10
10:1
1:10

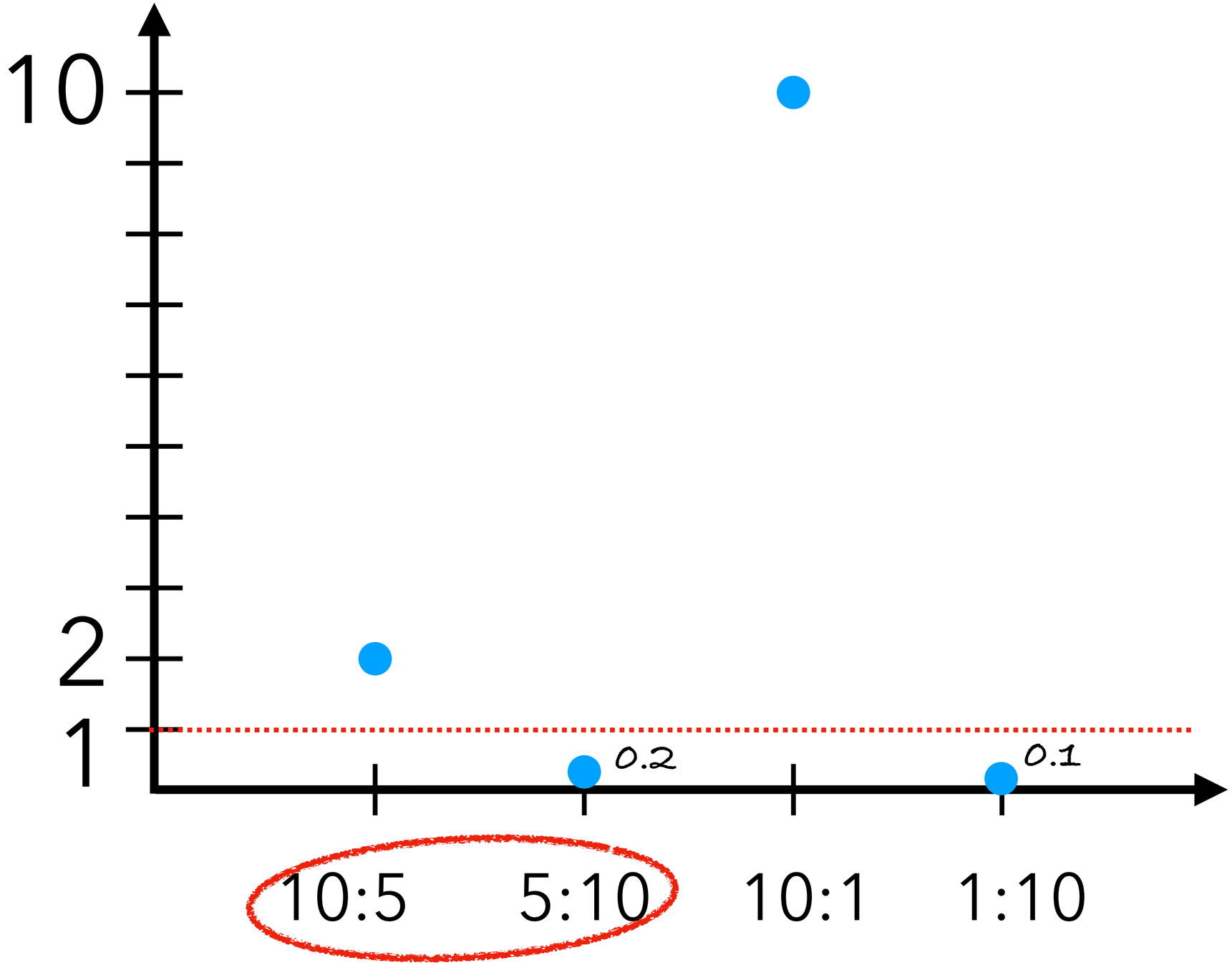

10
2
1
10:5
5:10
10:1
1:10

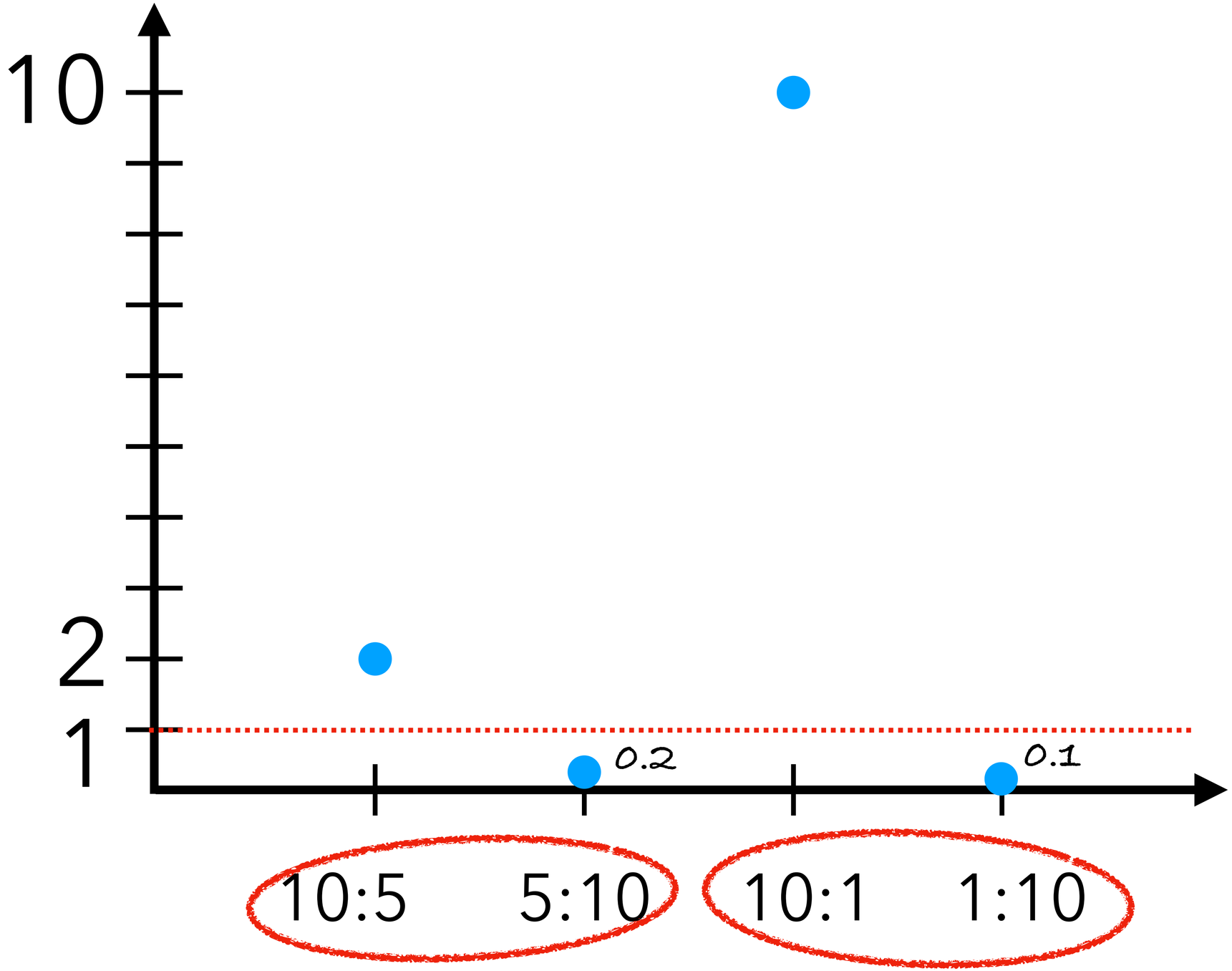

10
2
1
0.2
10:5
5:10
10:1
1:10

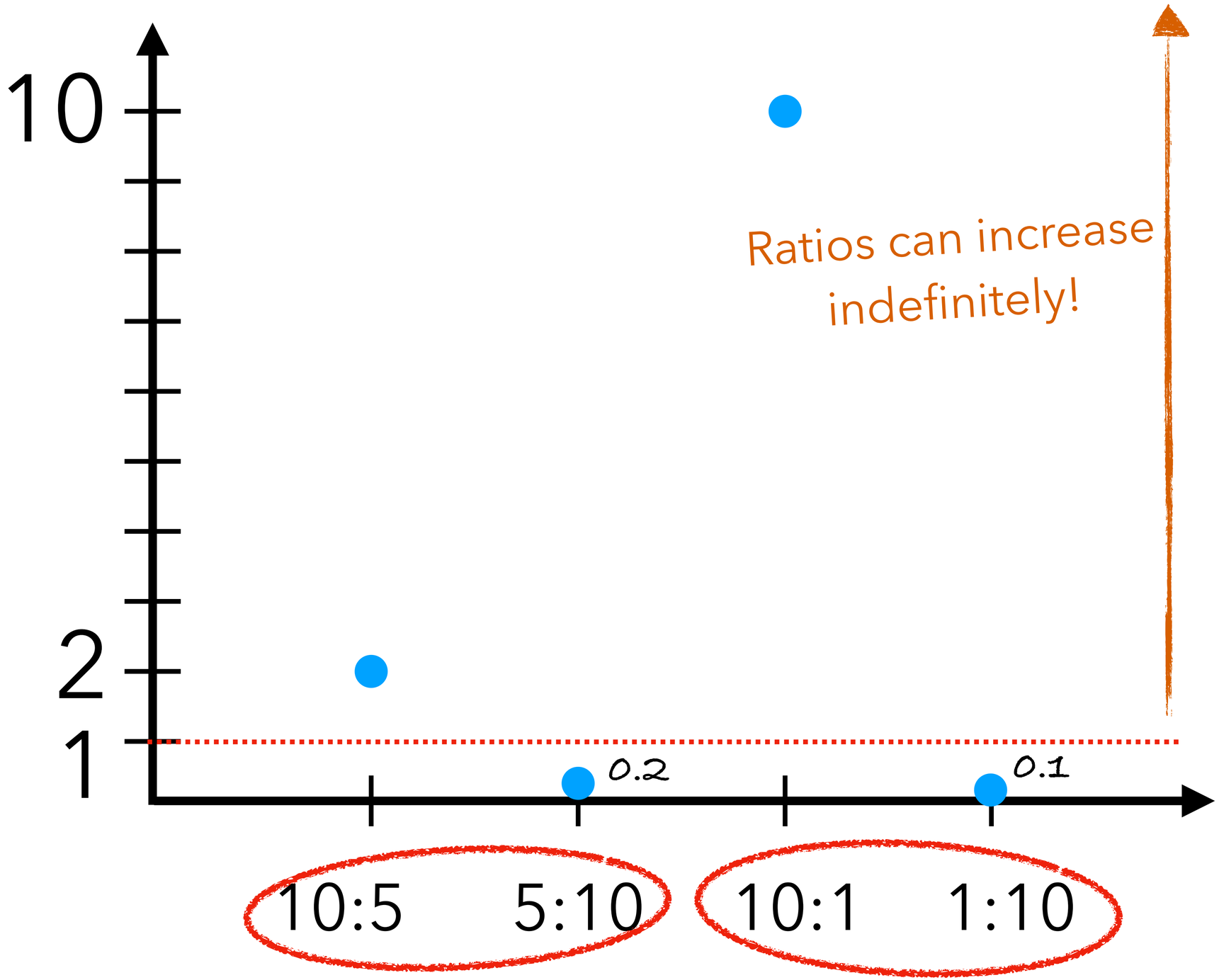

10
2
1
0.2
10:5
5:10
10:1
1:10


10
2
1
0.2
0.1
10:5
5:10
10:1
1:10


10
2
1
0.2
0.1
10:5
5:10
10:1
1:10


10
2
1
0.2
0.1
10:5
5:10
10:1
1:10


10
2
1
0.2
0.1
10:5
5:10
10:1
1:10


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

Log scale!

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

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

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

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

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

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

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

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

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

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

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

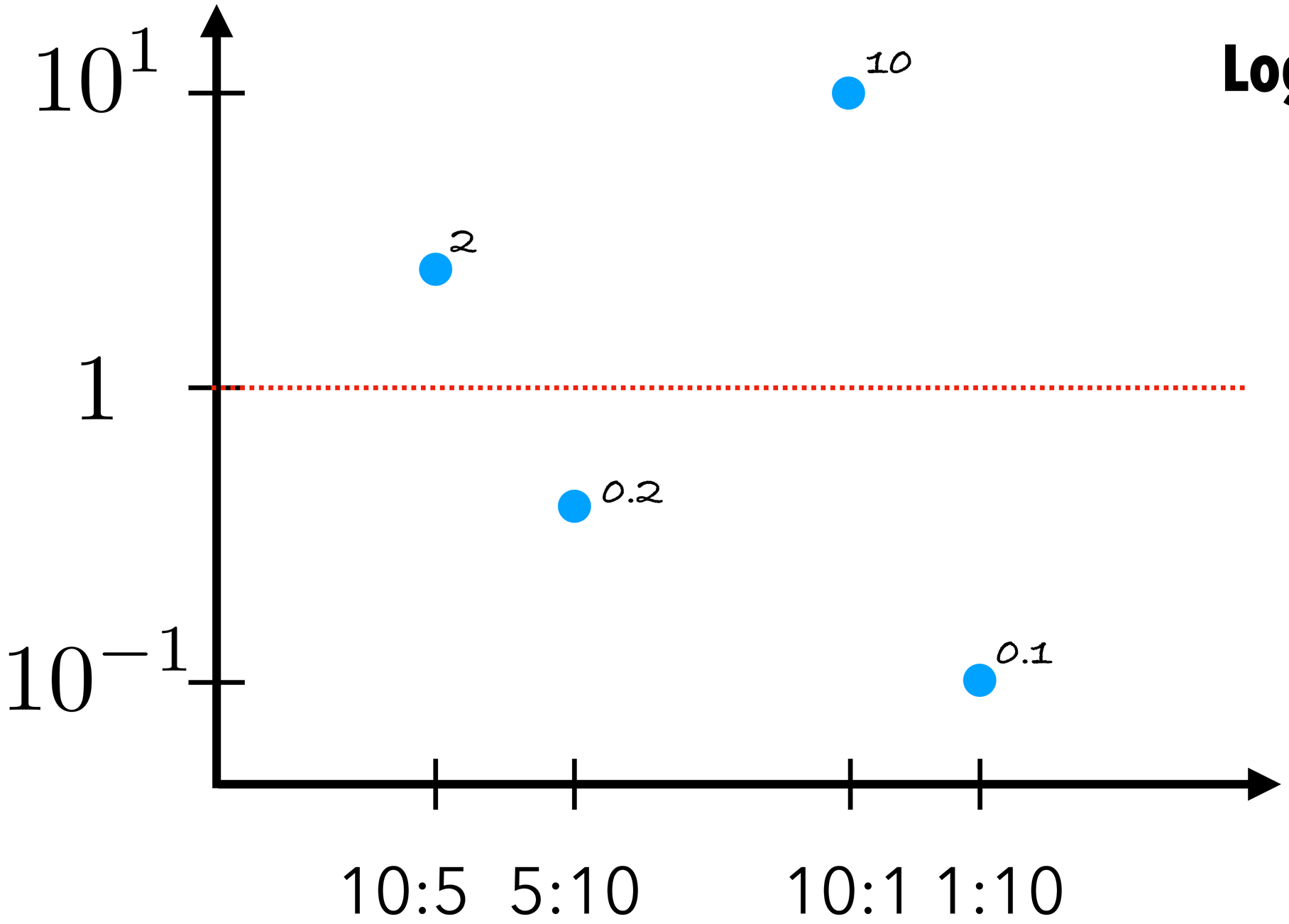
$$\log \frac{1}{10} = \log 1 - \log 10 = -\log 10$$

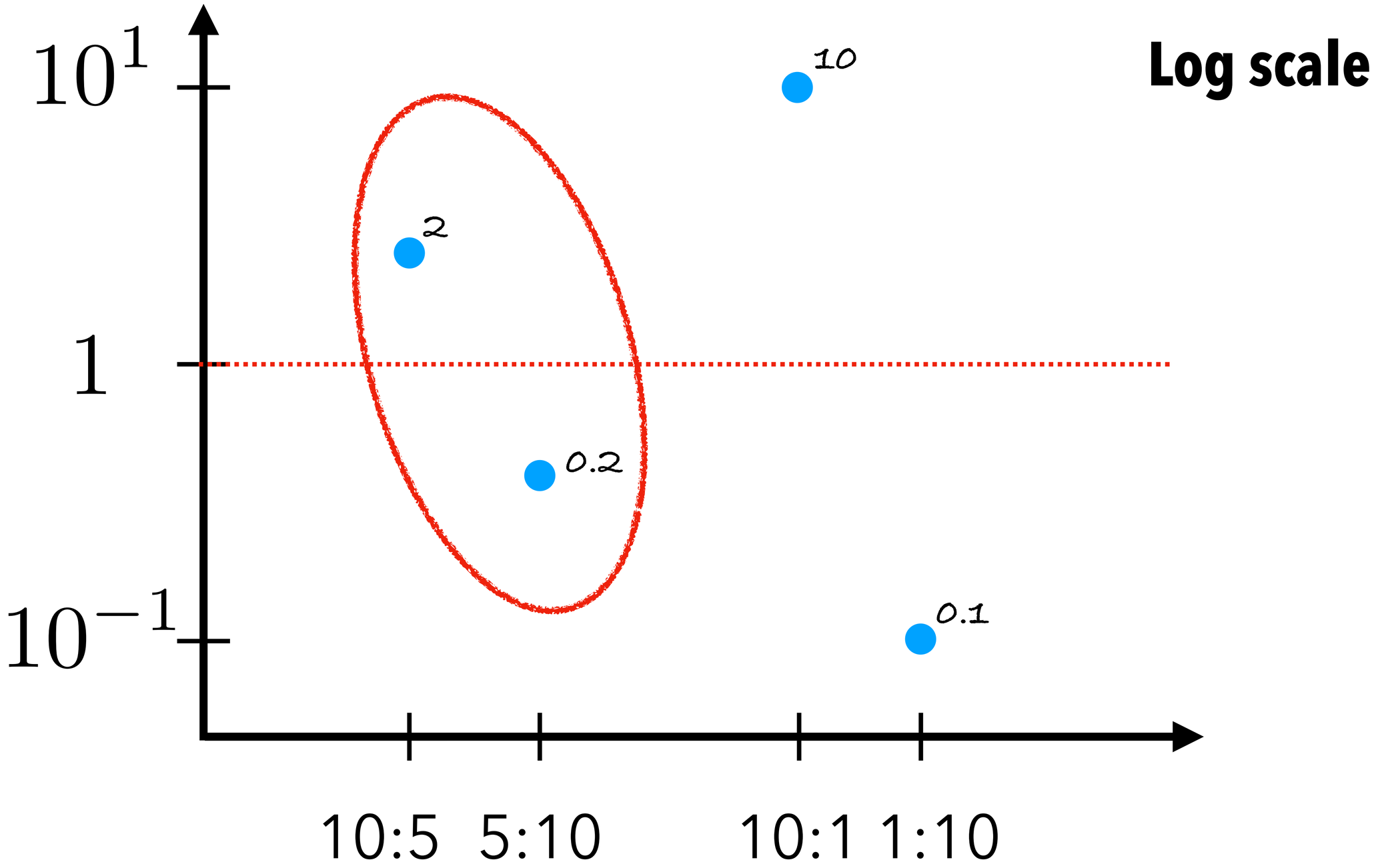
$$\log \frac{10}{1} = \log 10$$

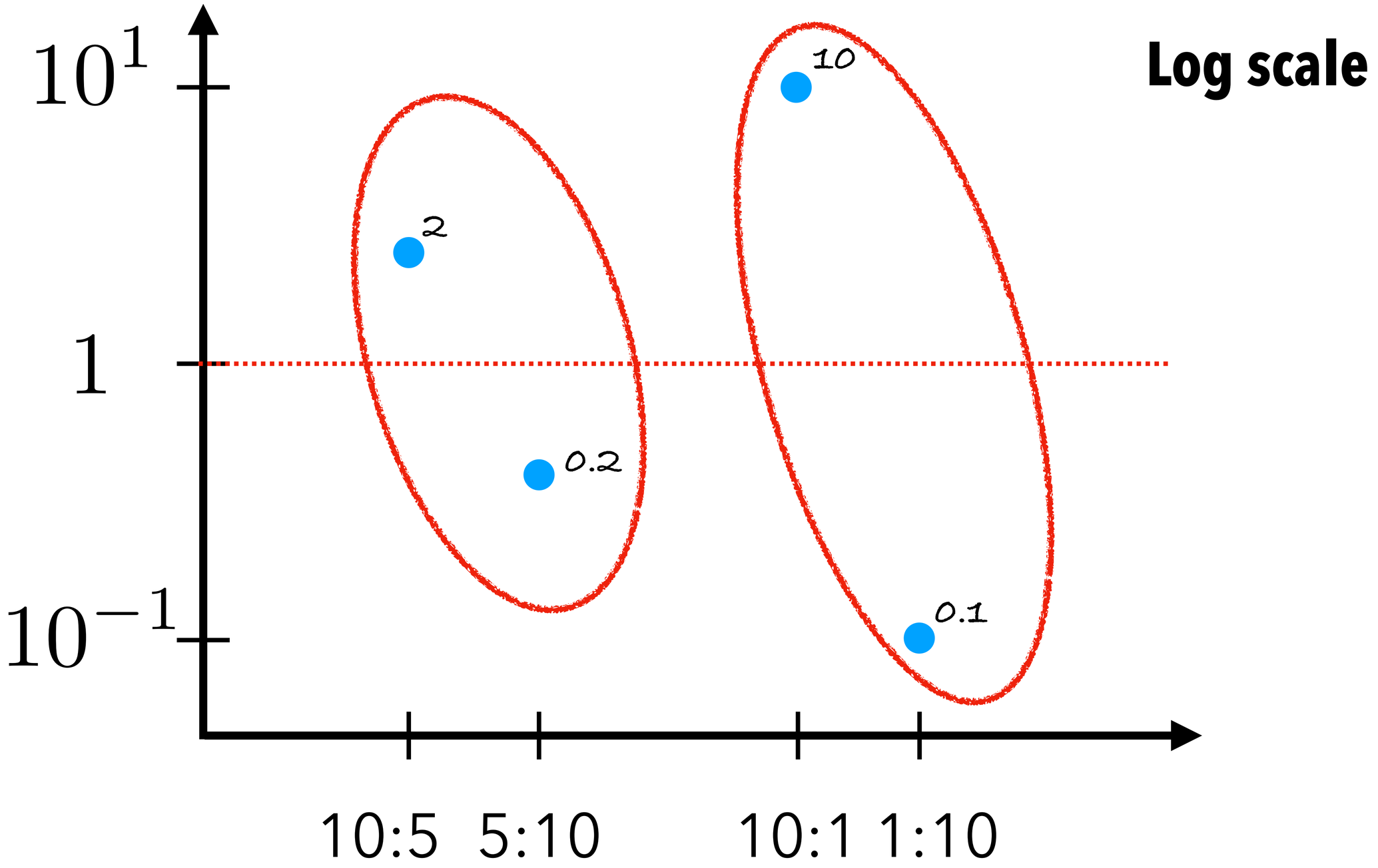
$$\log \frac{1}{10} = \log 1 - \log 10 = -\log 10$$

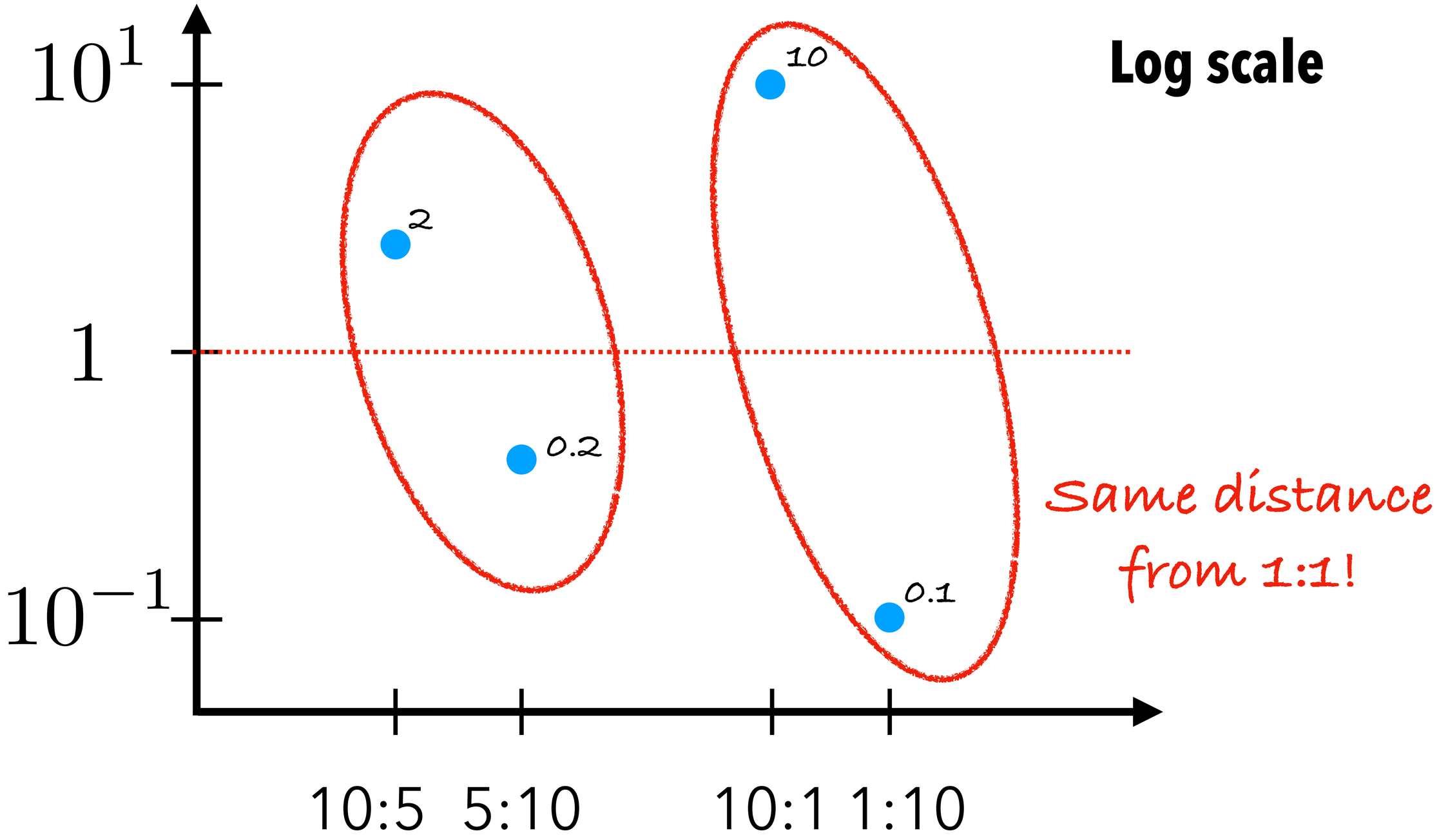
$$\log \frac{10}{1} = \log 10$$

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


Log scale
10^1
1
10^{-1}
10
2
0.2
0.1
10:5
5:10
10:1
1:10


Log scale
10^1
1
10^{-1}
10
2
0.2
0.1
10:5
5:10
10:1
1:10


Log scale
10^1
1
10^{-1}
2
0.2
10
0.1
10:5
5:10
10:1
1:10


Log scale
10^1
1
10^{-1}
10
2
0.2
0.1
Same distance from 1:1!
10:5
5:10
10:1
1:10