Coverage of space by random sets

Rahul Roy

Indian Statistical Institute, New Delhi

Coauthors – Siva Athreya, Amites Dasgupta, Ellen Saada, Anish Sarkar

The question

Consider the non-negative integer line.

For each integer point we toss a coin.

If the toss at location i is a

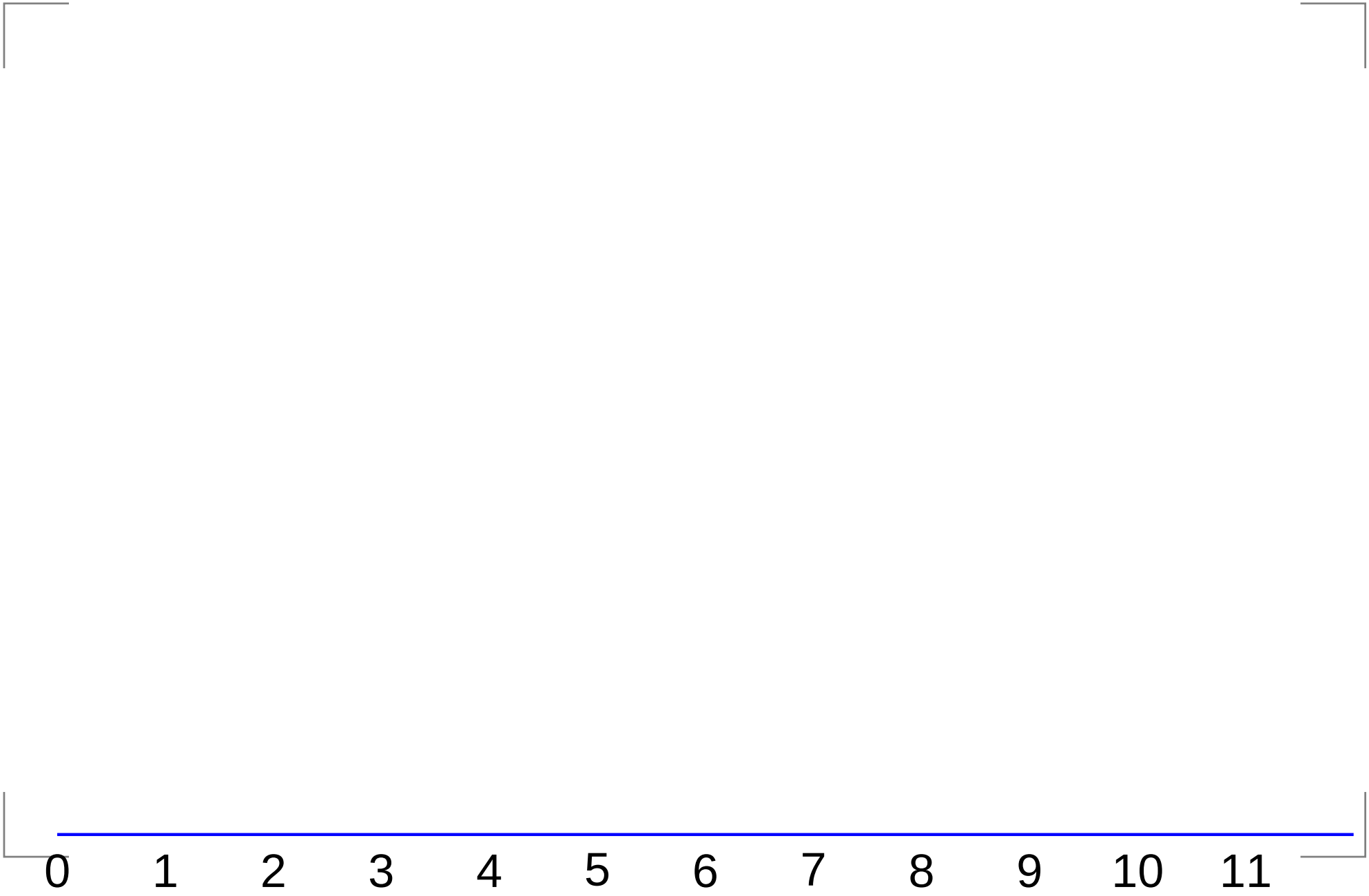
Heads we place an interval (of random length) there and move to location $i + 1$,

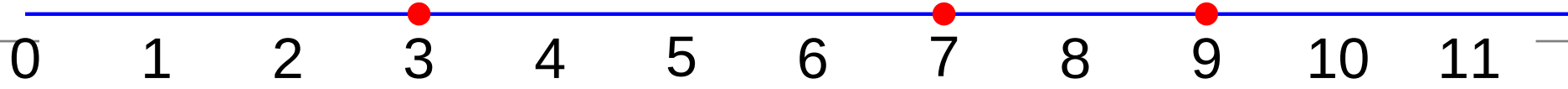
Tails we move to location $i + 1$.

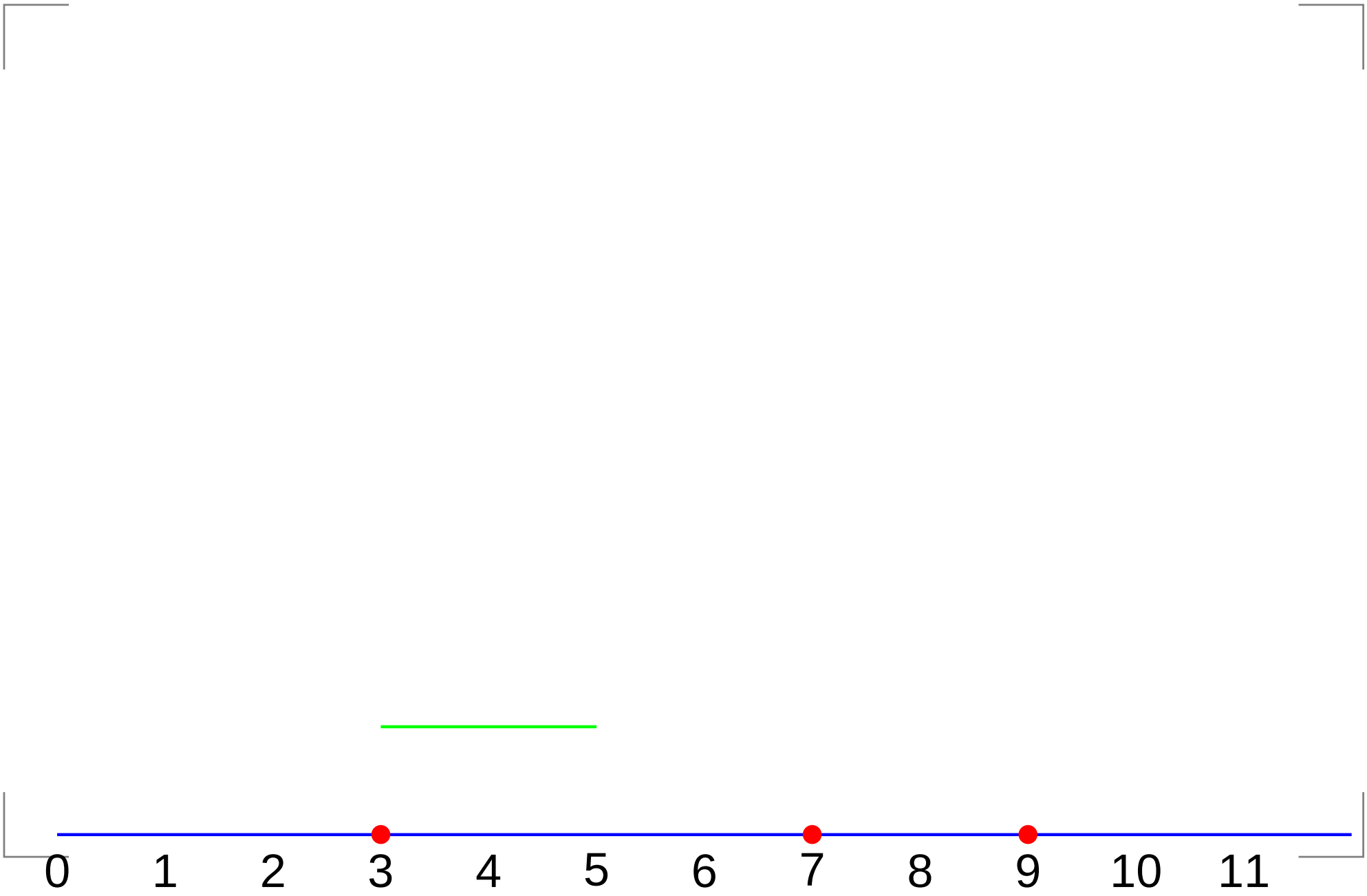
Before we proceed further we qualify the various randomness involved above:

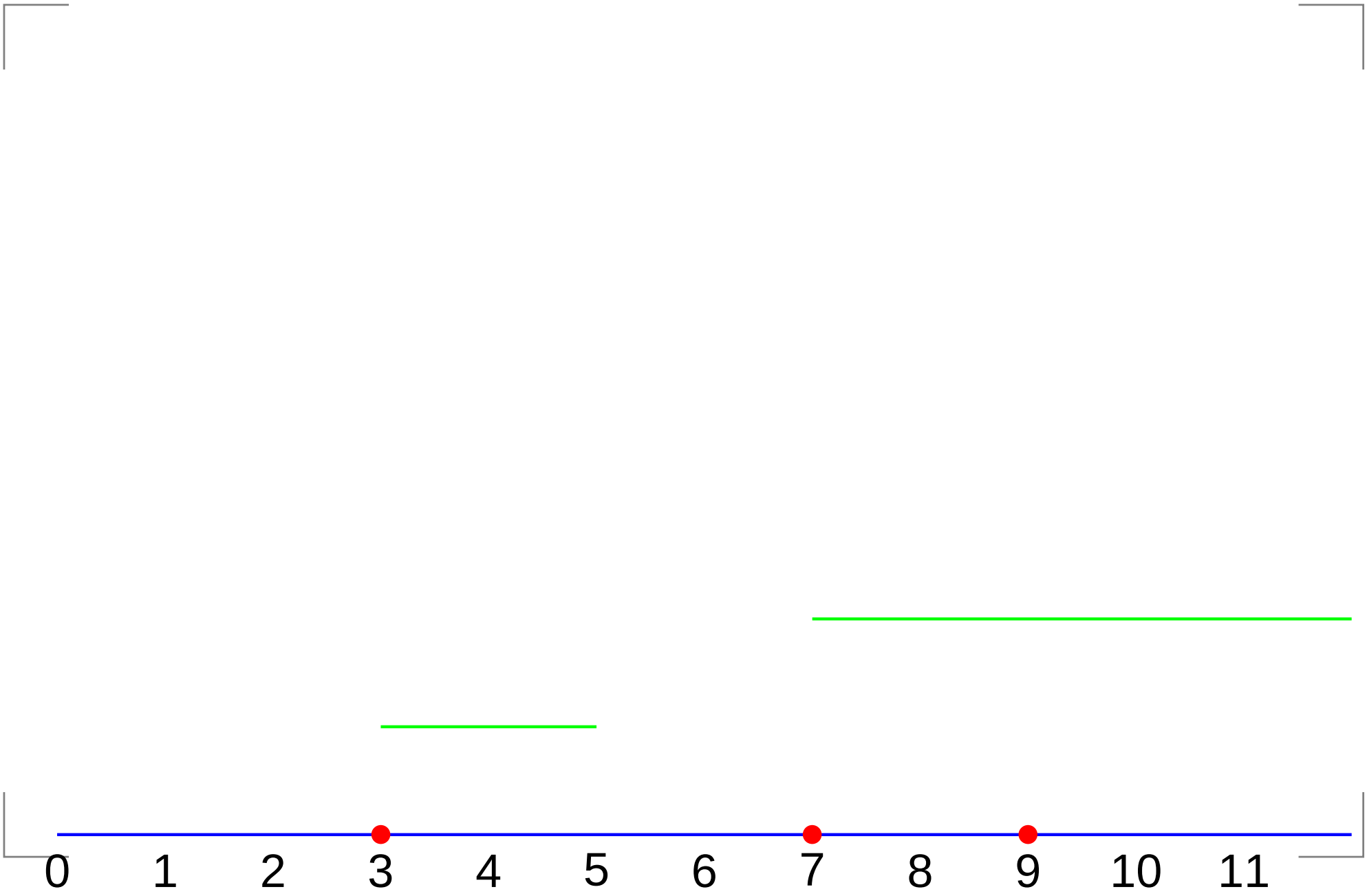
(i) For different integer points we perform different independent tosses, however each toss has a probability p of showing Head
and a probability $1 - p$ of showing Tail.

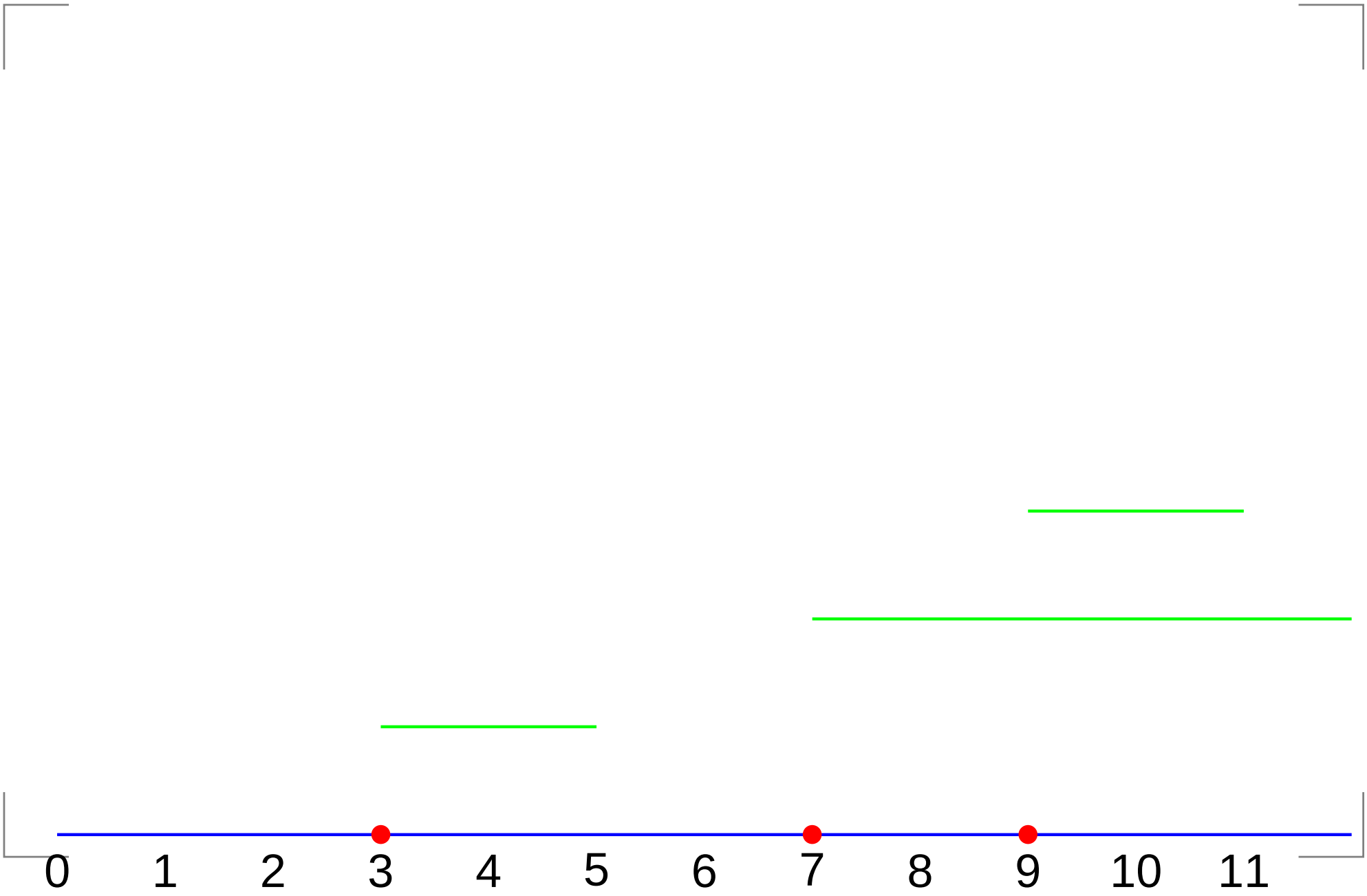
(ii) If the toss for the location i resulted in a Head, then the interval placed there is $[i, i + \rho_i]$ where ρ_i is random, and ρ_i and ρ_j are independent of each other for all $i \neq j$.

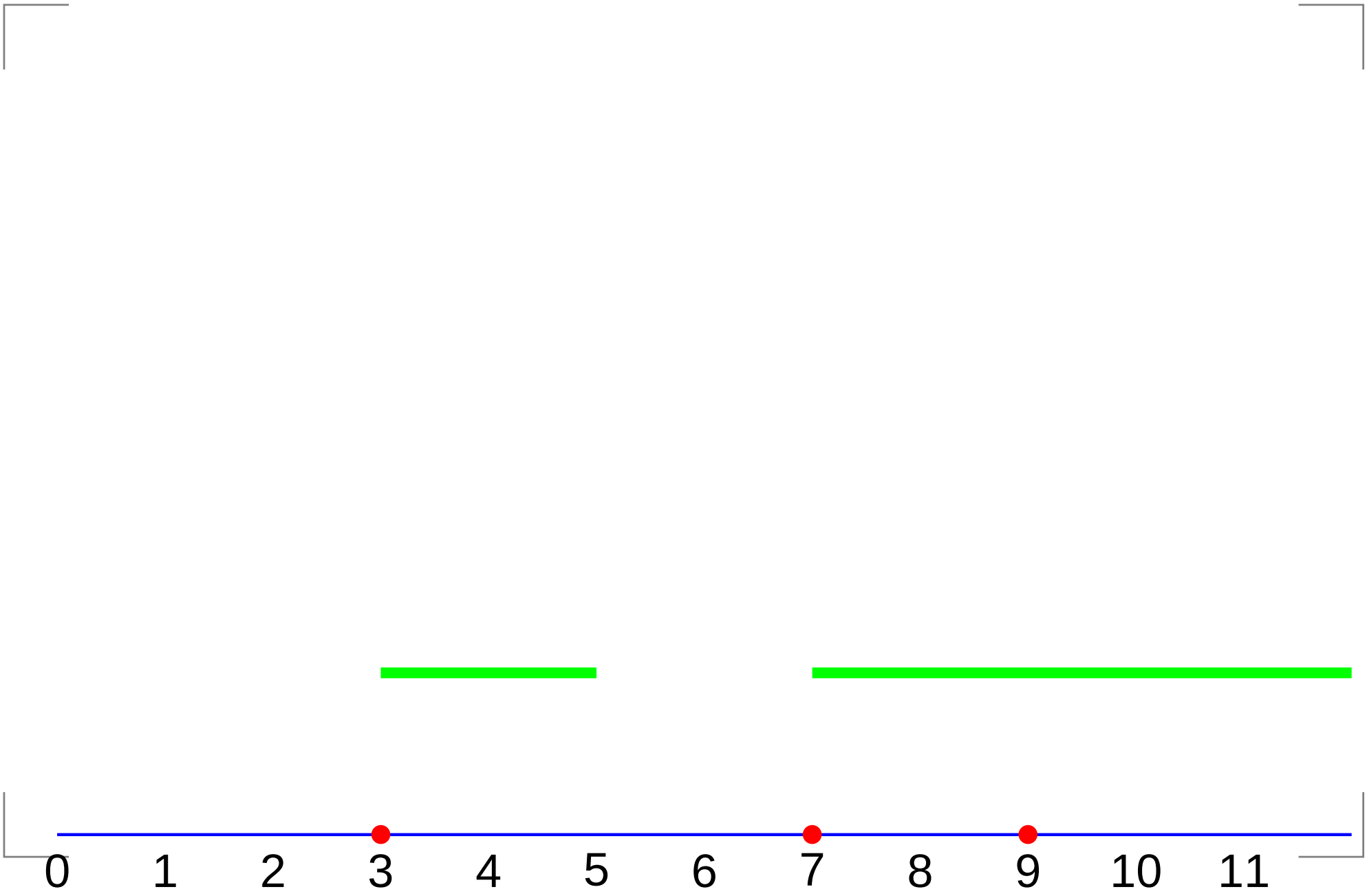

0
1
2
3
4
5
6
7
8
9
10
11


0
1
2
3
4
5
6
7
8
9
10
11


0
1
2
3
4
5
6
7
8
9
10
11


0
1
2
3
4
5
6
7
8
9
10
11


0
1
2
3
4
5
6
7
8
9
10
11


0
1
2
3
4
5
6
7
8
9
10
11

The green region is exactly

$$C = \bigcup_{\{i:\ \mathsf{Toss}(i) = \mathsf{H}\}} [i, i + \rho_i]$$

Question: Does there exist a t such that $[t, \infty) \subseteq C$?

If such a t exists, then we say that the integer line is eventually covered by C.

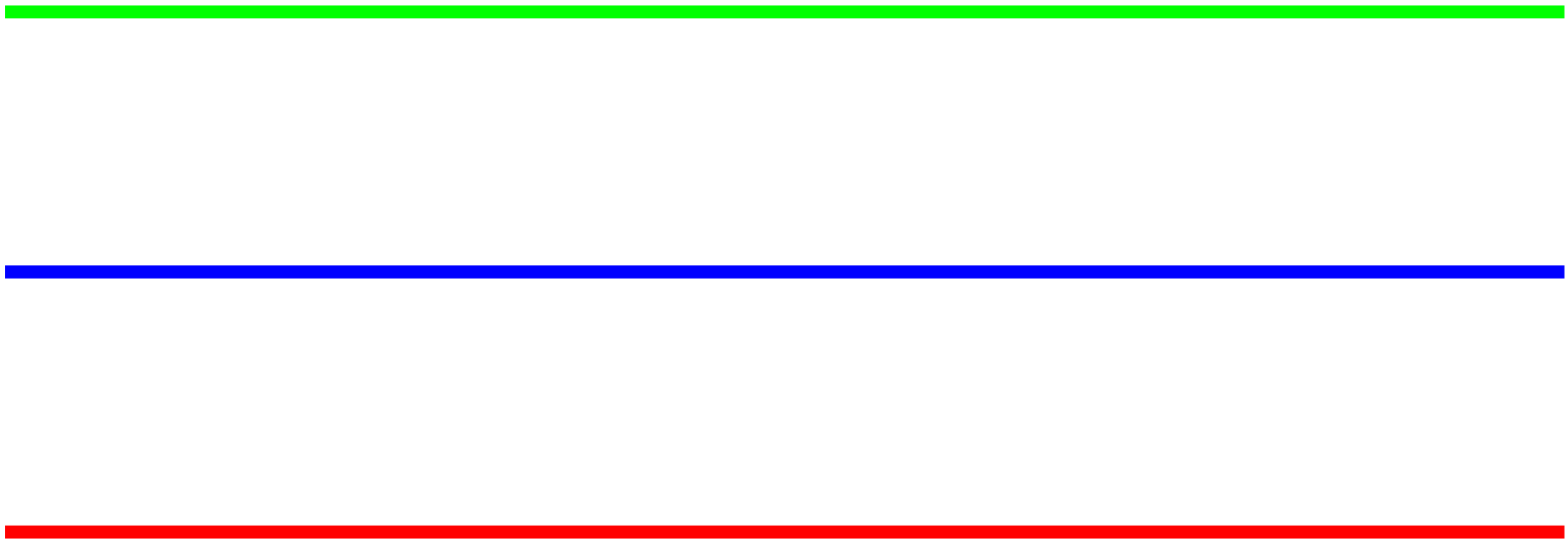
Biological significance

In the shotgun sequencing method employed in the human genome project the genome was first cloned to have many identical clones.

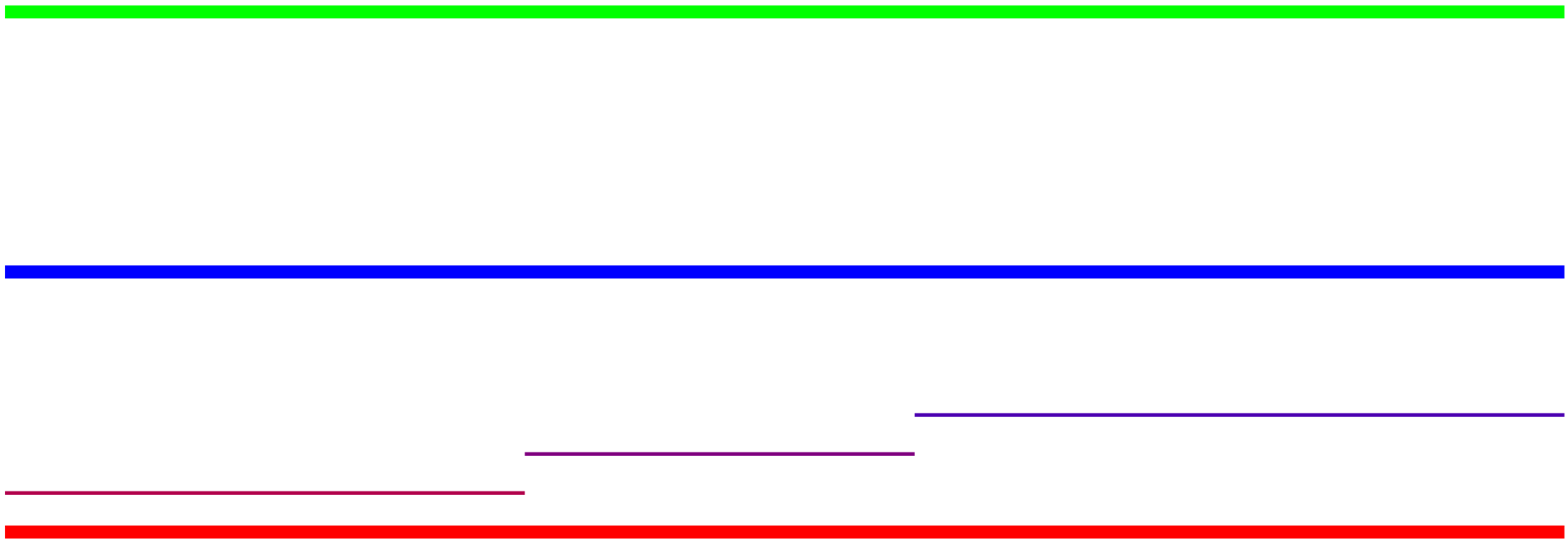
Each clone was then cut (by chemical means) into various segments. The segments were then sent to laboratories for sequencing.

The segments were then stitched together to obtain the full sequence.

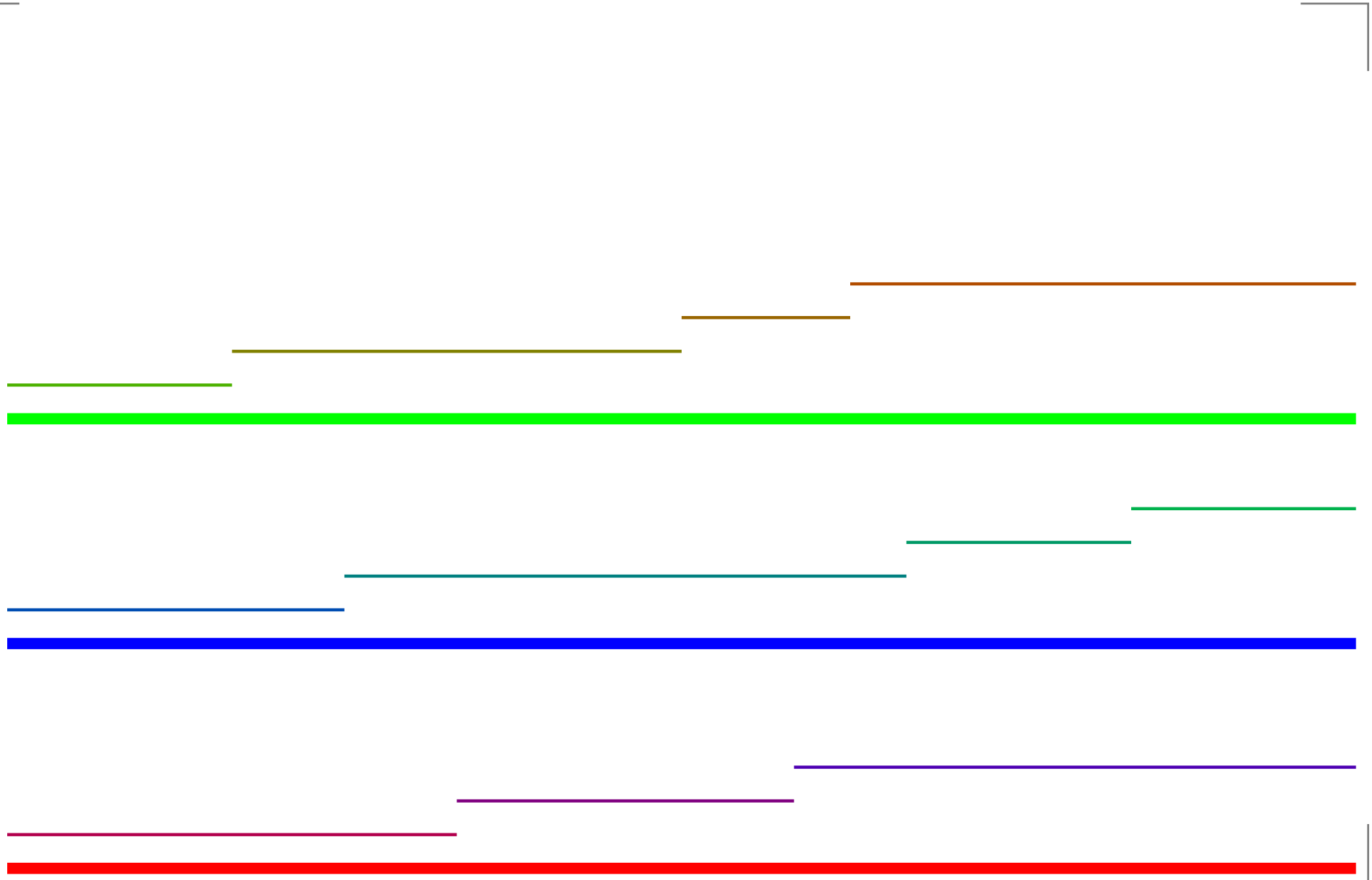
The genome cloned



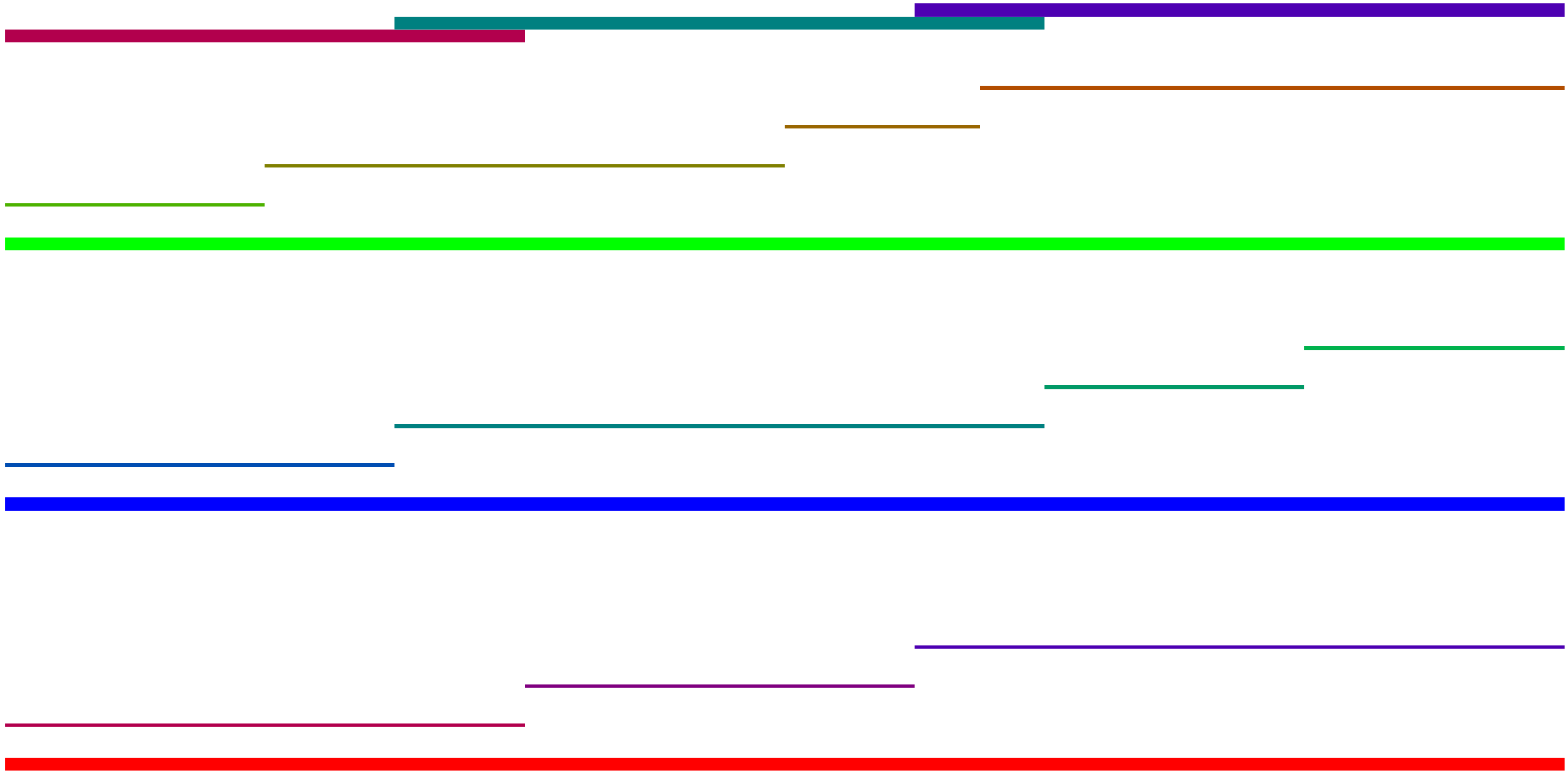
Fragmented



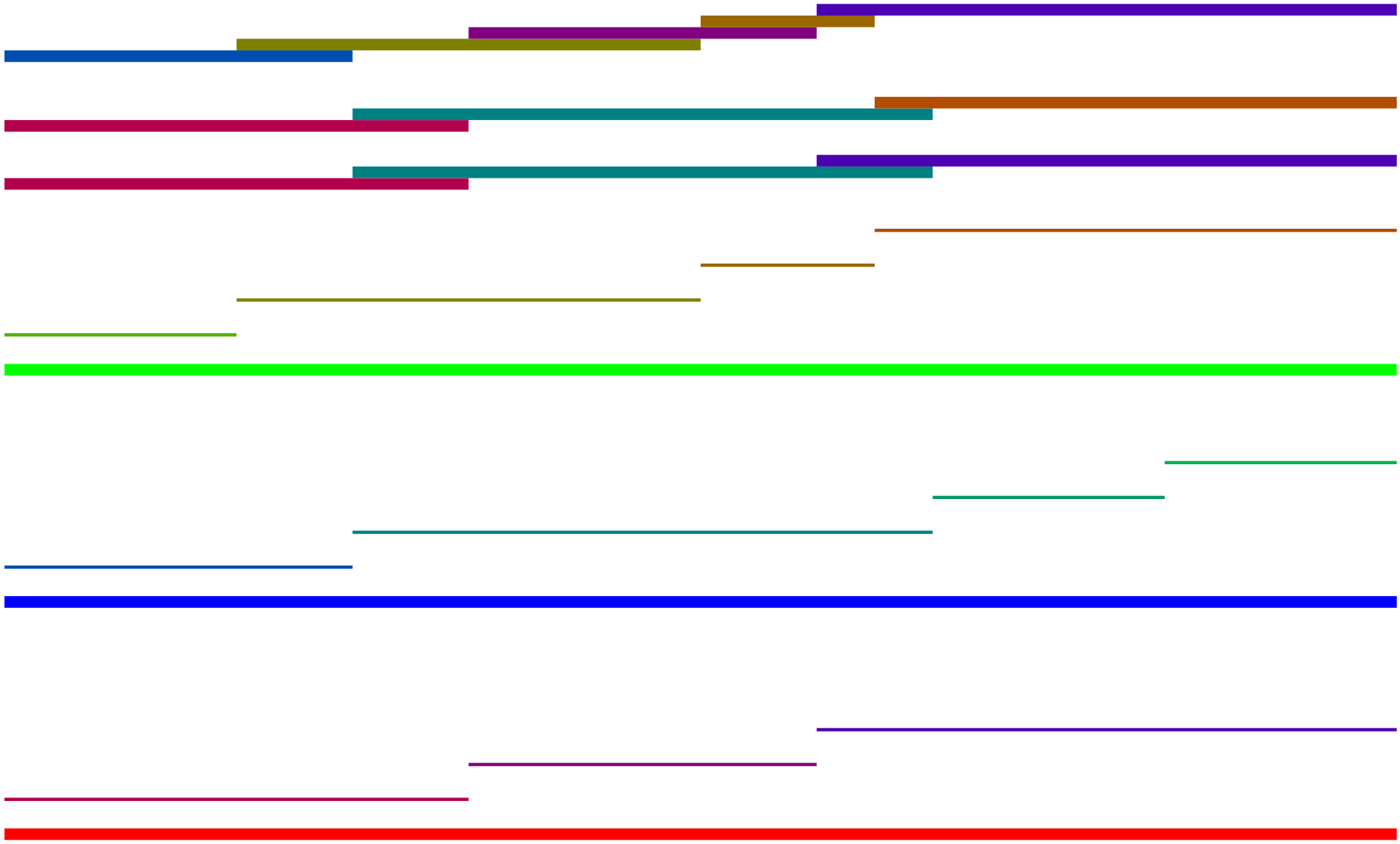
Fragmented



Stitched



Another stitching



The connection

In our model the integer points correspond to nucleotides/amino acids.

The red points where the intervals are placed are the starting points of the segments of the genome, with an interval being the segment itself.

Eventual coverage corresponds to whether a machine can do the stitching or not.

Theorem

Let

$$l = \lim_{j \to \infty} jP(\rho > j).$$

$$\text{Eventual coverage} \begin{cases} \text{occurs} & \text{if } p > 1/l \\ \text{does not occur} & \text{if } p < 1/l \end{cases}$$

Such a dichotomy in behaviour is called phase transition is statistical physics, in the sense that there is a discontinuity in the system at $p_c = 1/l$. There are many such examples of phase transition observed in mathematical models of statistical physics, e.g. in Ising models, in interacting particle systems, in percolation theory, etc.

A similar result is obtained when the locations where the intervals are placed are not independently chosen, but are chosen according to a Markov process. This is more in tune with the assumptions made in biology about the nucleotide sequence/ the amino acid sequence being Markovian.

A similar result is obtained when one considers the non-negative real line, instead of just the non-negative integer line. In this case, instead of the integer points being the start of an interval determined by the toss of a coin, the points which are the start of intervals form a Poisson point process, which is just a mathematical formalization of the colloquial statement ‘points are randomly chosen on the non-negative real line’.

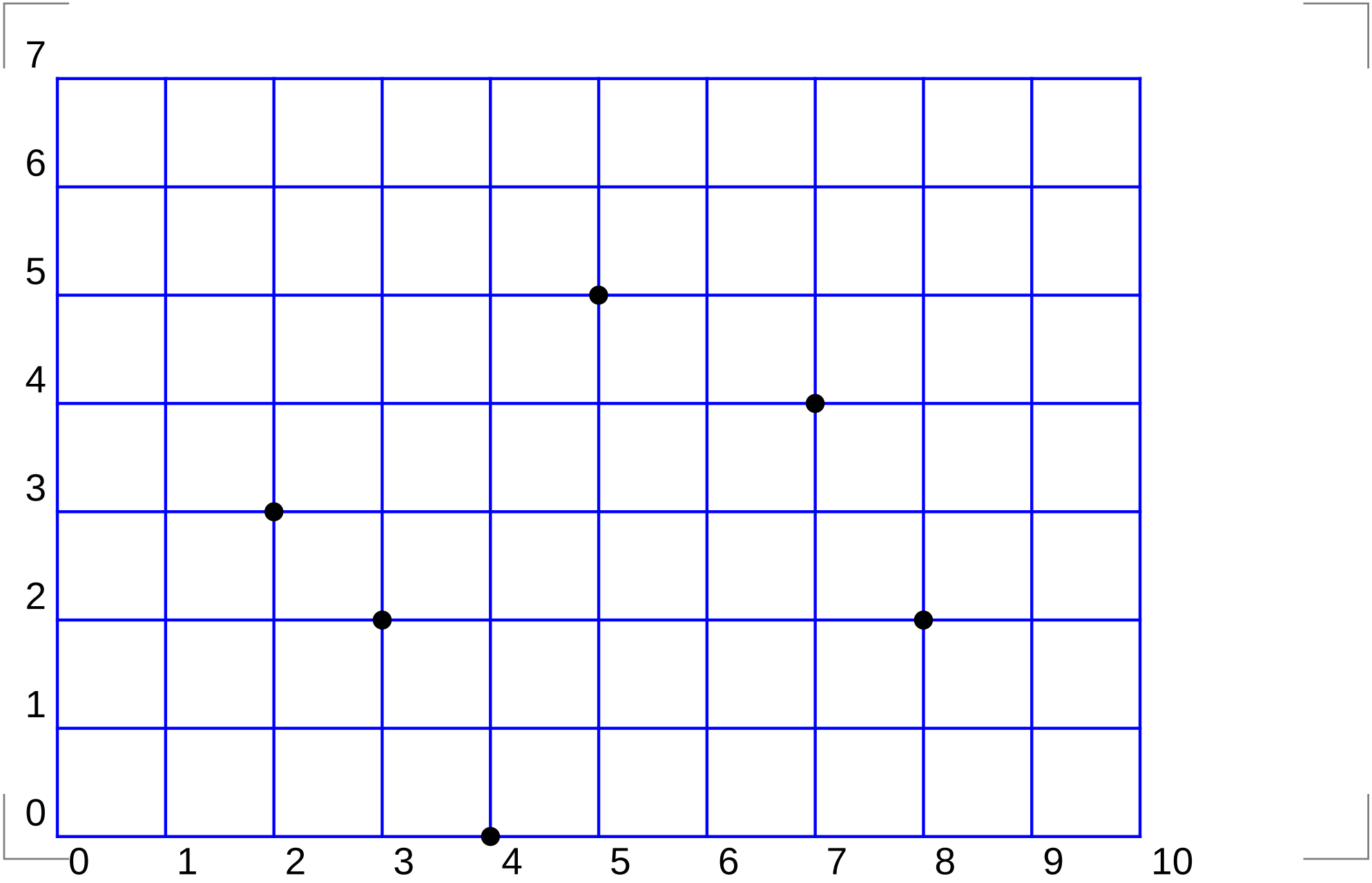
Mathematically ...

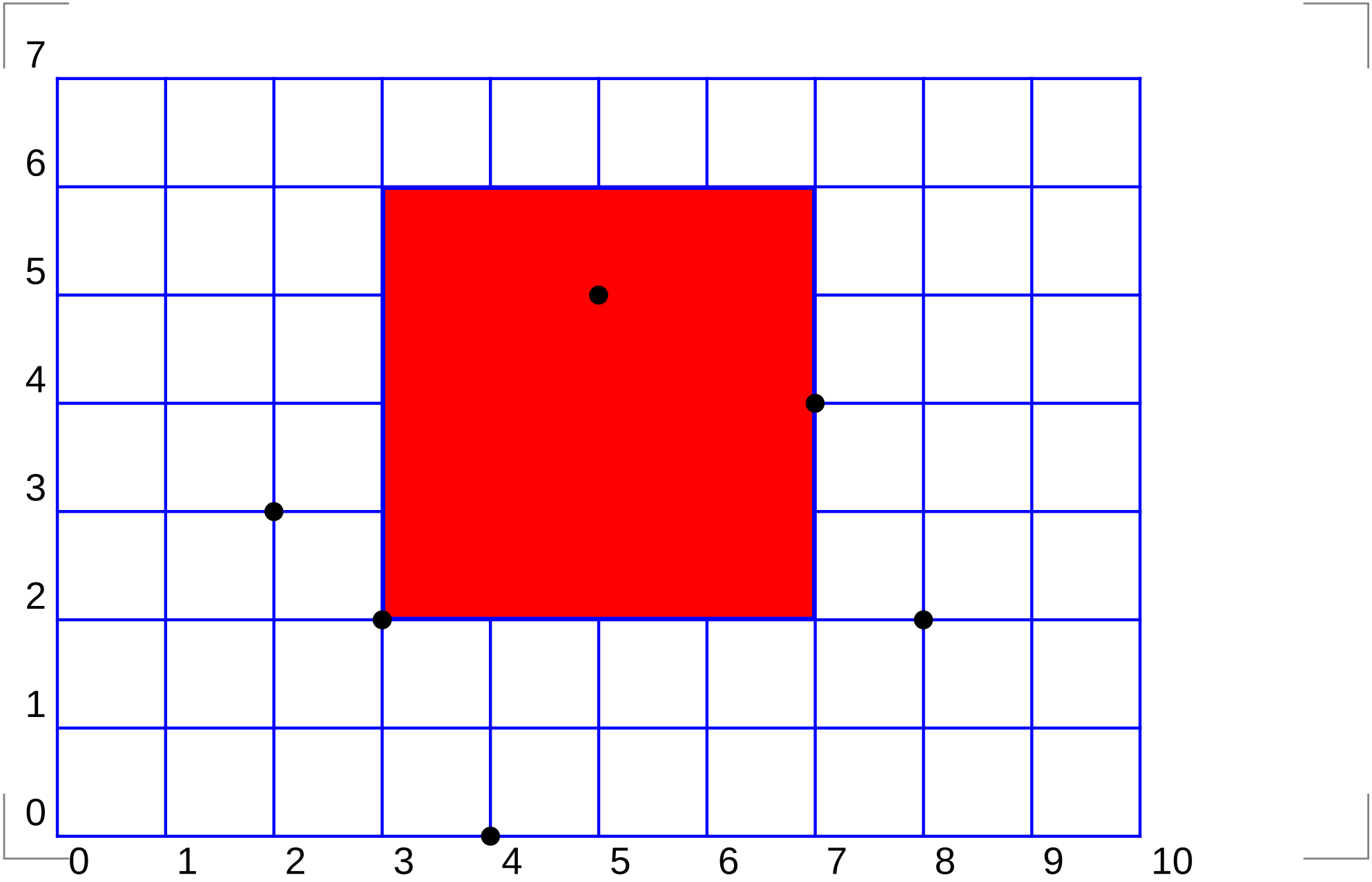
Mathematically the question of coverage by random sets is quite old. Dvoretsky (1956, PNAS) asked the question about the complete coverage of a circle by random arcs. Gilbert (1965, Biometrika) generalized the question to that of covering a sphere by circular caps.

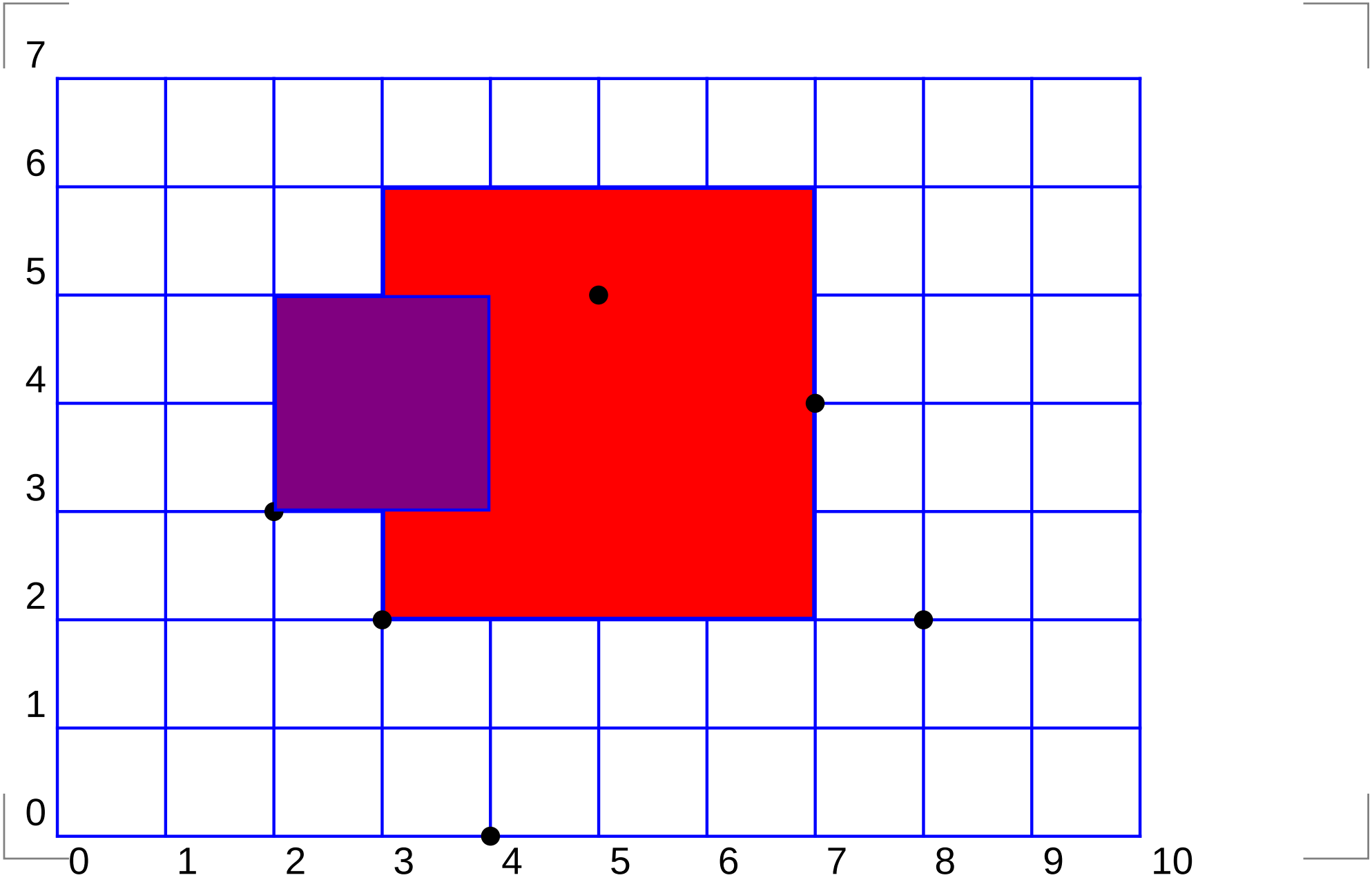
Mandelbrot (1972, ZW) and Shepp (1972, PTRF) provided an answer to Dvoretsky's question for the one dimensional real line $\mathbb{R}$, while Hall (1985, Ann Probab) provided an answer to Gilbert's question for the d-dimensional Euclidean space $\mathbb{R}^d$.

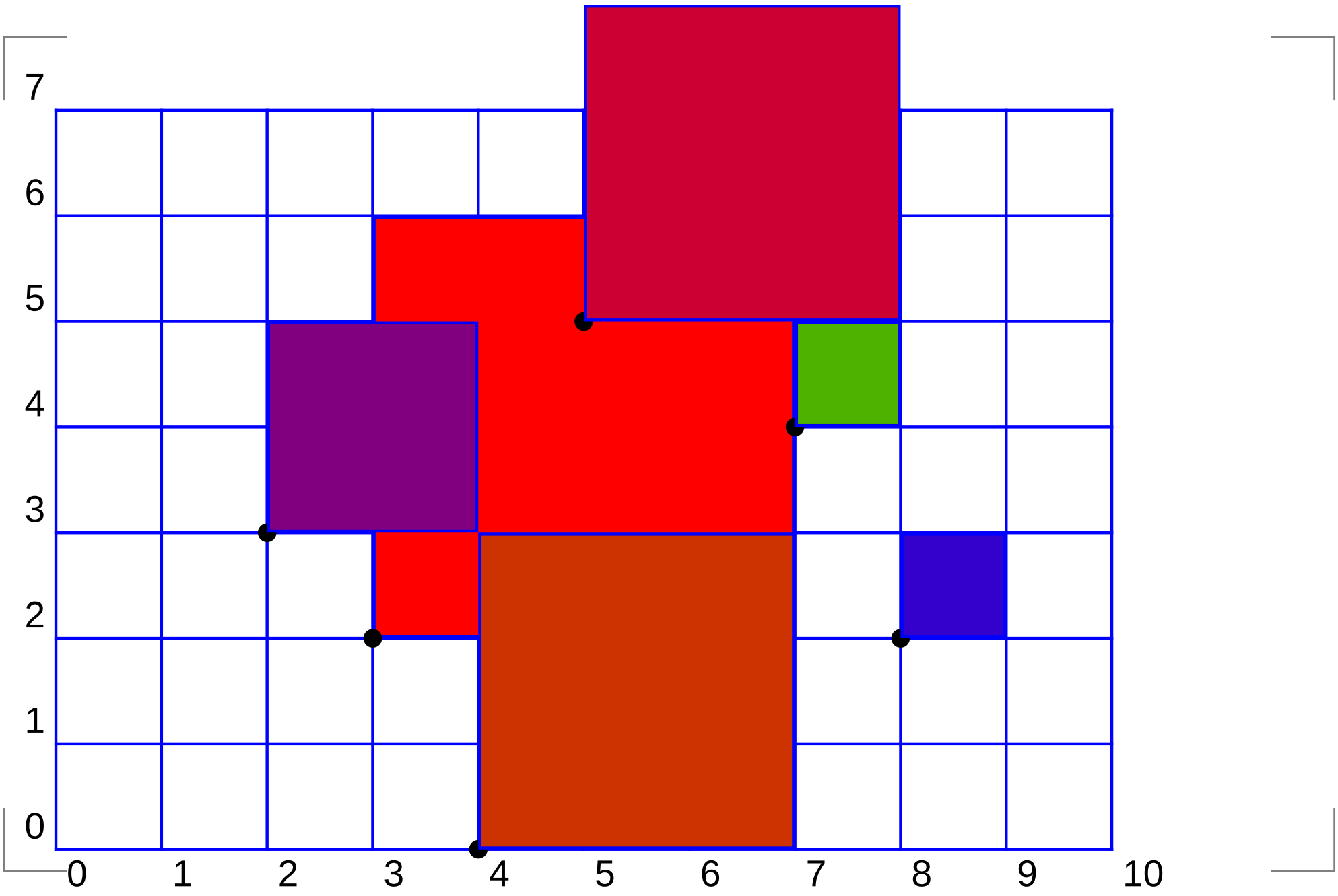
Our question is similar, except that it studies the half-line.

Orthant




7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10


7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10


7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10

The covered region is

$$C = \bigcup_{\{(i,j):\ \text{Toss}(i,j) = \text{Heads}\}} (i,j) + ([0,\rho_{ij}] \times [0,\rho_{ij}]).$$

Question: Is there a $t \geq 0$ such that $[t,\infty) \times [t,\infty) \subseteq C$?

Of course, this is a purely mathematical question, and the only connection it may have outside mathematics is with questions of percolation in statistical physics.

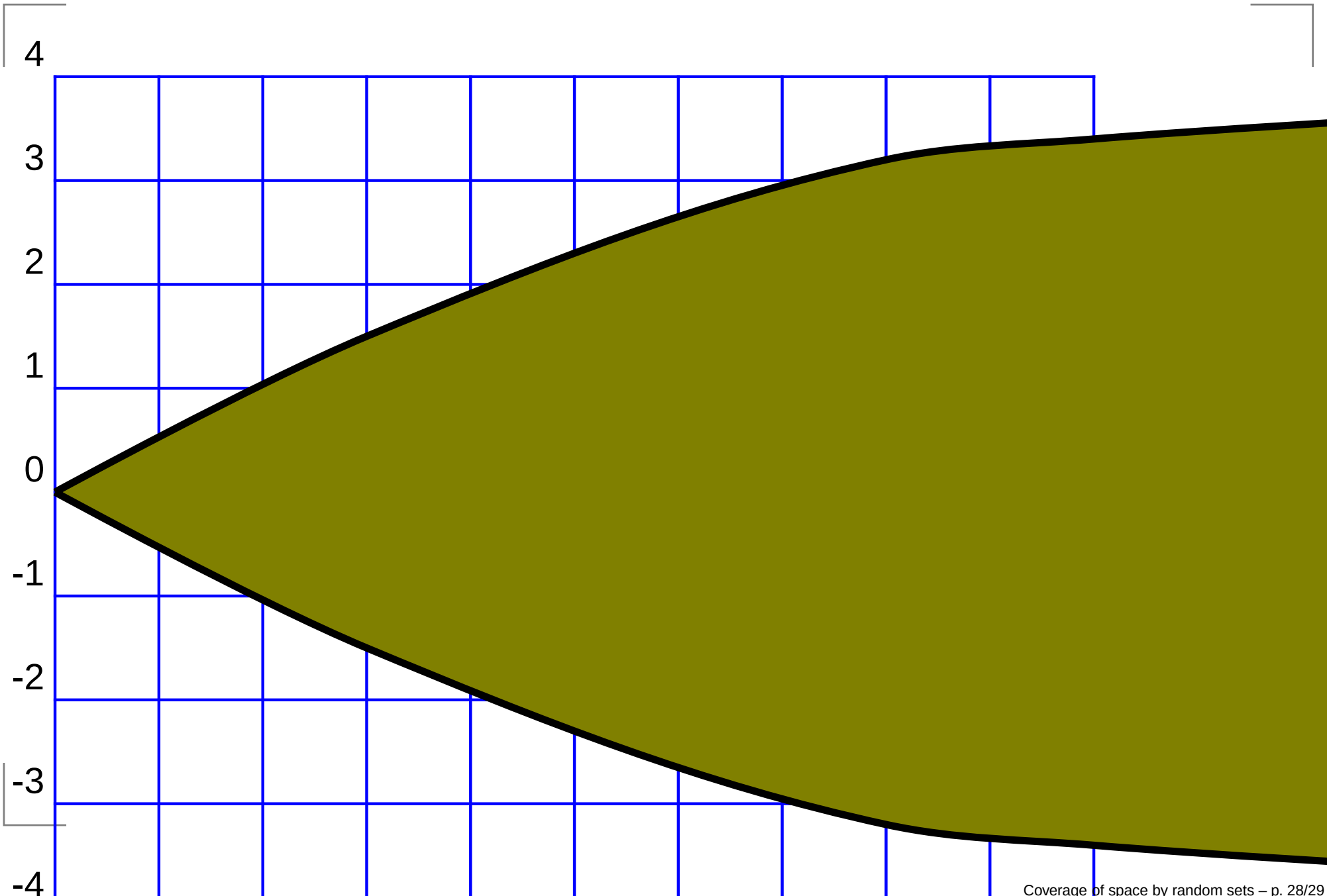
Theorem

Let $d \geq 2$ and $0 < p < 1$.

Eventual coverage $\begin{cases} \text{occurs} & \text{if } \lim_{j\to\infty} jP(\rho > j) > 0 \\ \text{does not occur} & \text{if } \lim_{j\to\infty} jP(\rho > j) = 0 \end{cases}$

Similar results are obtained when instead of considering only the integer co-ordinates in the positive orthant we consider the entire positive orthant. Again, here the location of the boxes are according to a Poisson point process.

Further work has been done, when instead of restricting ourselves to orthants or half-space we study the question of eventual coverage in regions bounded by a function $f(x)$


4
3
2
1
0
-1
-2
-3
-4

The phase transition point then depends on the nature of the function. In particular whether the function is slowly varying or regularly varying and, if regularly varying, then what is the index of variation?

Finally, questions connected to percolation in these regions have also been studied.