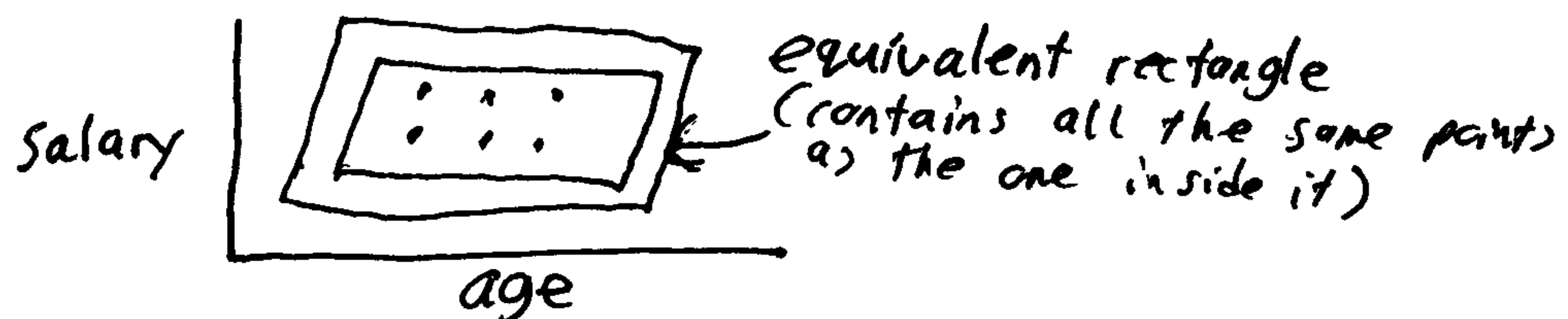


### VC Dimension (Vapnik & Chervonenkis)

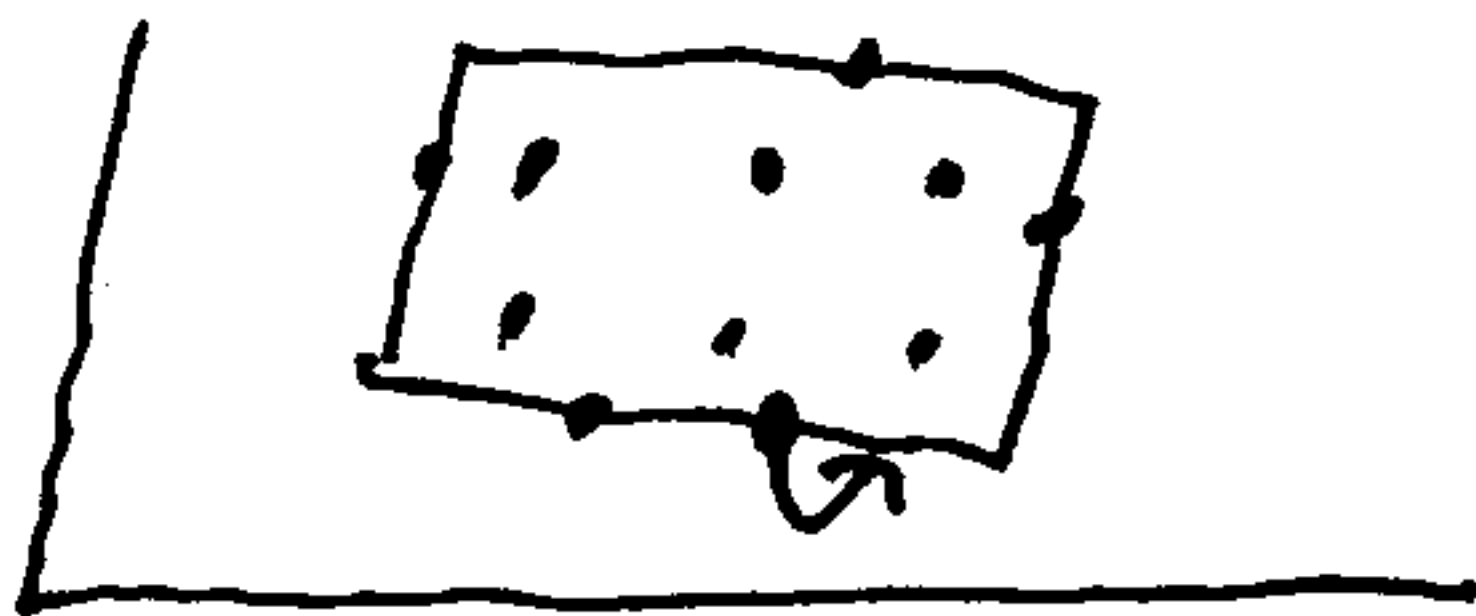
**Motivation:** Given a large data set, want to be able to take only a small part of that data set and ask questions about it. But want to be able to get the same answers as if the whole data set was being interrogated.

However, it is necessary to allow only certain types of questions to be asked about the data set (e.g. How many people within a specified age range earn a salary (also within a specified range)?)



**Question:** How many equivalent rectangles are there, given  $n$  points?

\*\* There are  $n^4$  equivalence classes of rectangles. (and  $n^3$  equivalence classes of circles)



Assign one point to an edge; 4 points that define the minimum rectangle.

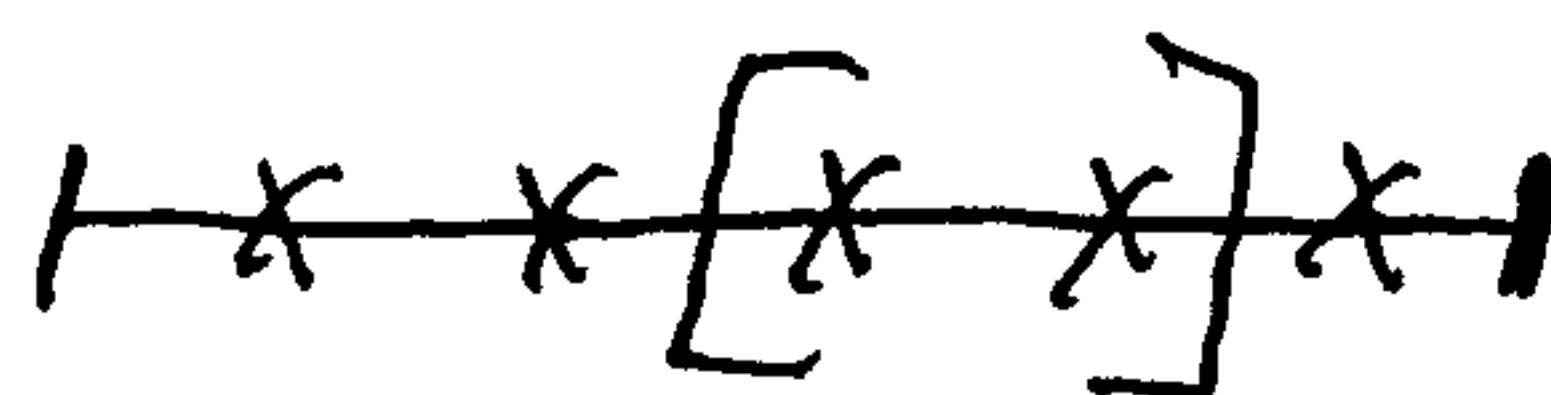
Cannot allow arbitrary shapes, because it is possible to ask a question that misses all of the data points entirely:



Given a probability distribution, how much mass is in a rectangle?

$|\text{probability mass} - \text{estimate}| < \epsilon$   
with probability  $1 - \delta$

Discrepancy Theory ("how far off can you be?")  
(related to  $\epsilon$  above)

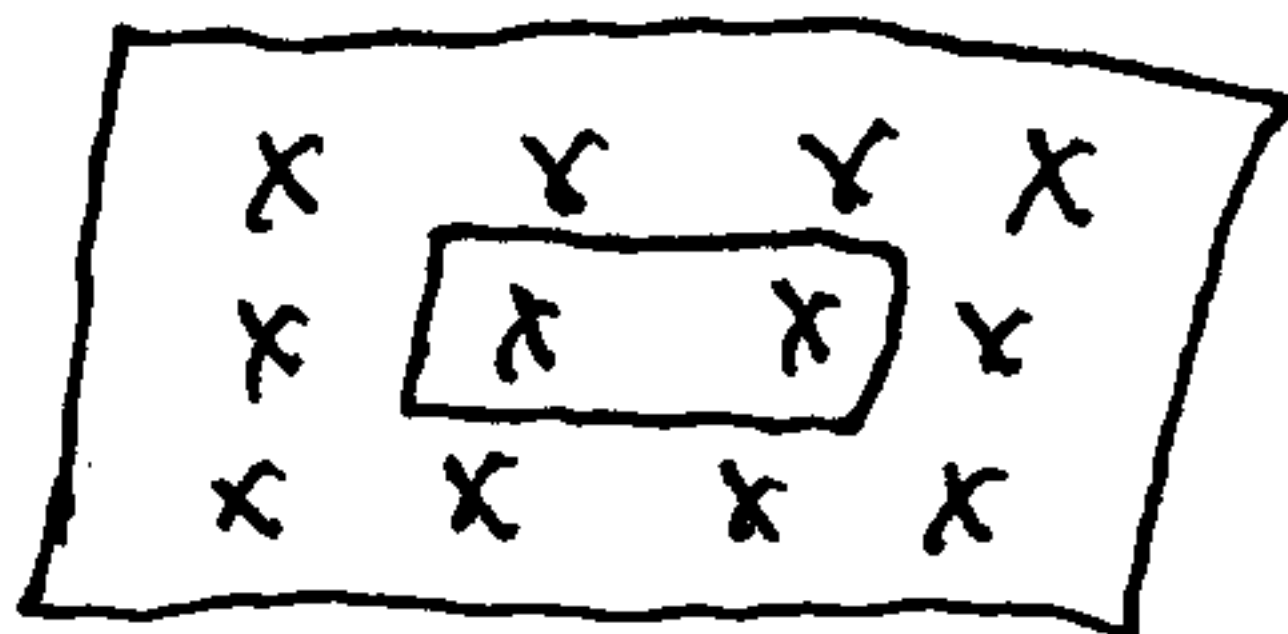
unit line: 

to uniformly place points on the line,  
make them evenly spaced

Given an interval  $[ ]$ , how far off from estimate can the captured number of points be?  
 $\pm 2$  points for one interval ( $\pm 4$  points for two intervals)

In 2 dimensions, how to make points as uniformly distributed as possible?

Lattice?



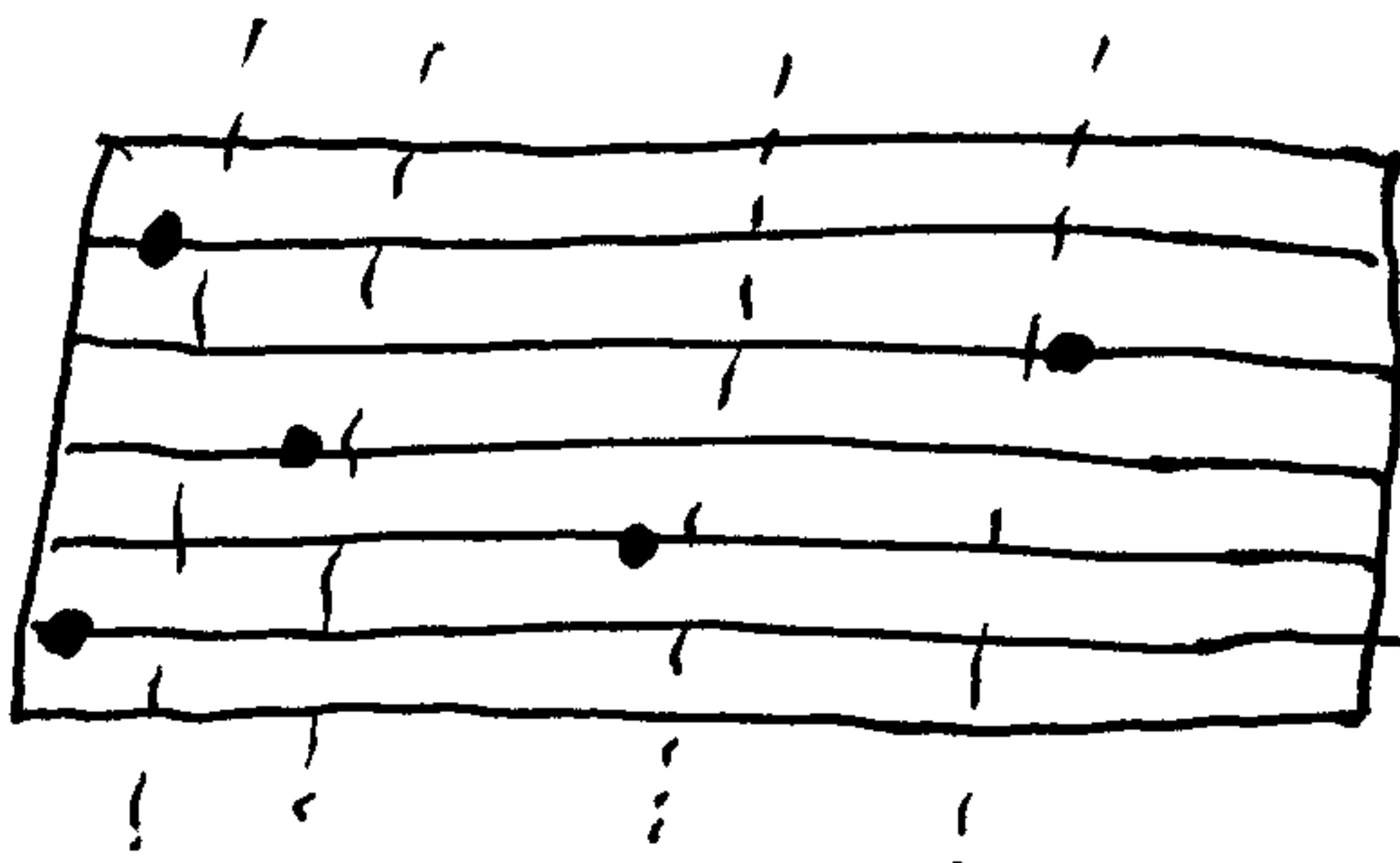
How far off can interval (rectangle) be?  
 $\pm n^{1/2} = \text{discrepancy}$

Can put points down in such a way that the discrepancy is down to  $\log(n)$ :



put points on lines that are  $1/n$  apart  
(unlike  $n^{1/2}$  apart on lattice)

One point per line, but where to place the  
point on each line?



Points are distributed as follows:

$P_i = (i/n, r(i))$  where if  $i$  is written in binary  $i = a_0 + 2a_1 + 4a_2 + 8a_3 + \dots$

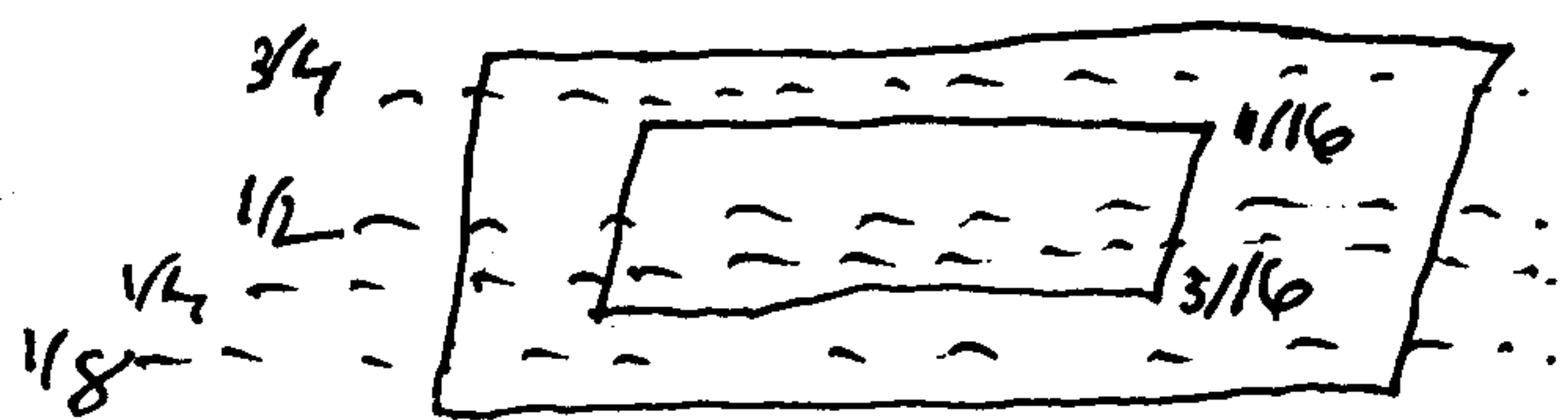
$r(i) = a_0/2 + a_1/4 + a_2/8 + \dots$

$a_0$	$a_1$	$a_2$	$a_3$
0	0	0	0
1	0	0	0
0	1	0	0
1	1	0	0
0	0	1	0

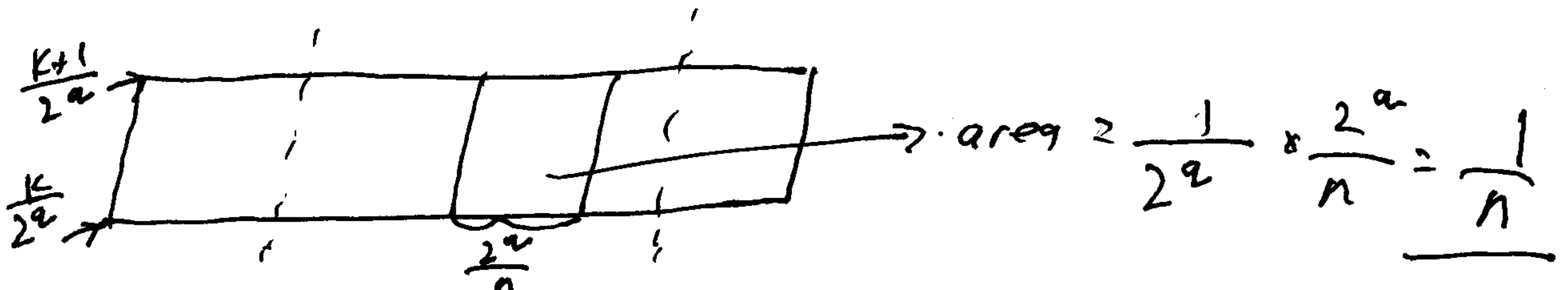
$a_0/2$	+	$a_1/4$	+	$a_2/8$
0		0		0
1/2		0		0
0		1/4		0
1/2		1/4		0
0		0		1/8

What kind of discrepancy will this give us?

→  $\log(n)$



# of bands to "carve up" the rectangle is  $\log(n)$   
(dividing each band in half)



So putting points down uniformly is more sophisticated than putting points down at vertices. And the way in which points are put down (the way uniform is defined) is dictated by shape

### VC Dimension

#### Shatter

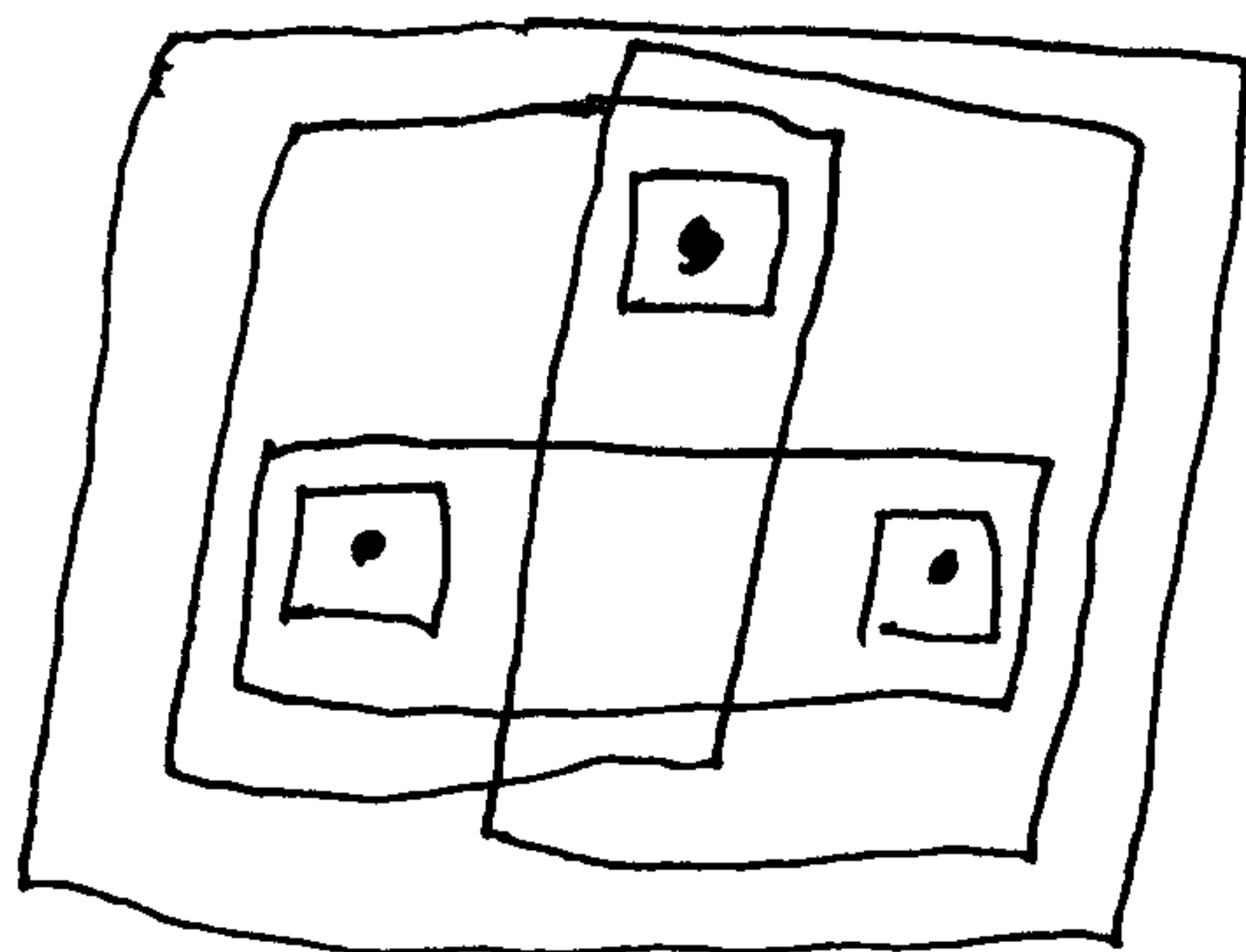
Given a set of points:

, ,

and a set of shapes (e.g. rectangles):



A set of shapes shatters a set of points if every subset is the set of points contained in some shape in the set.



rectangles can be drawn to cover each possible subset of points (including the empty set)



cannot be shattered by axis-parallel rectangle (cannot capture A & D alone, for example) ~~AB~~

but:

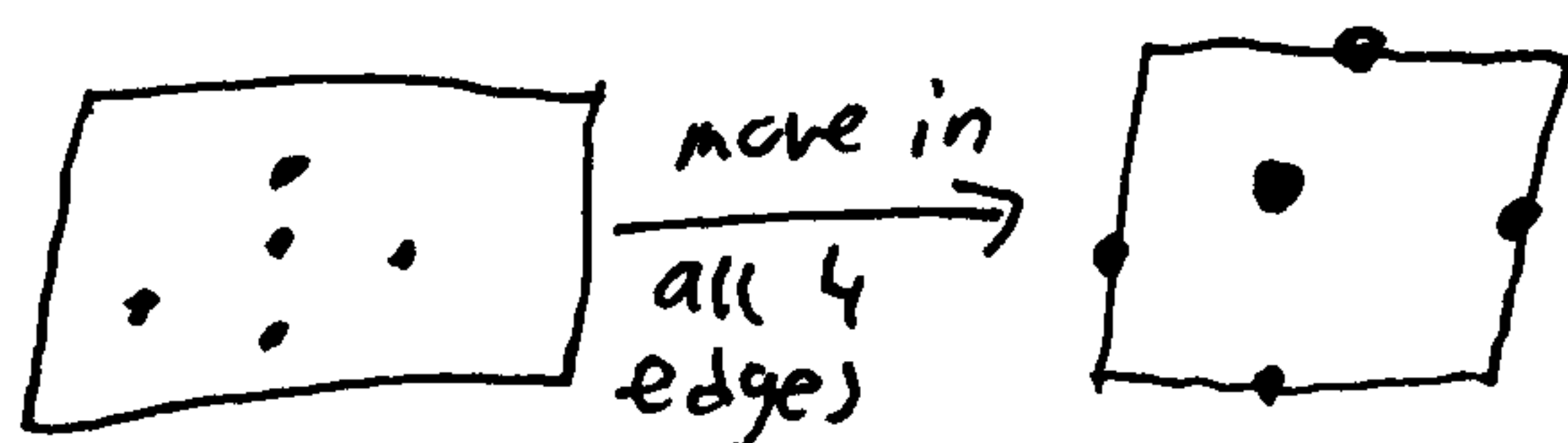


can be shattered

What is the VC dimension of axis-parallel rectangles?

The maximum  $n$  such that there is a set of  $n$  points that can be shattered (at least 4)

\*\*VC dimension = 4: no set of 5 points can be shattered by axis-parallel rectangles



cannot represent all 4 points on perimeter without also representing the 5<sup>th</sup> point in the middle

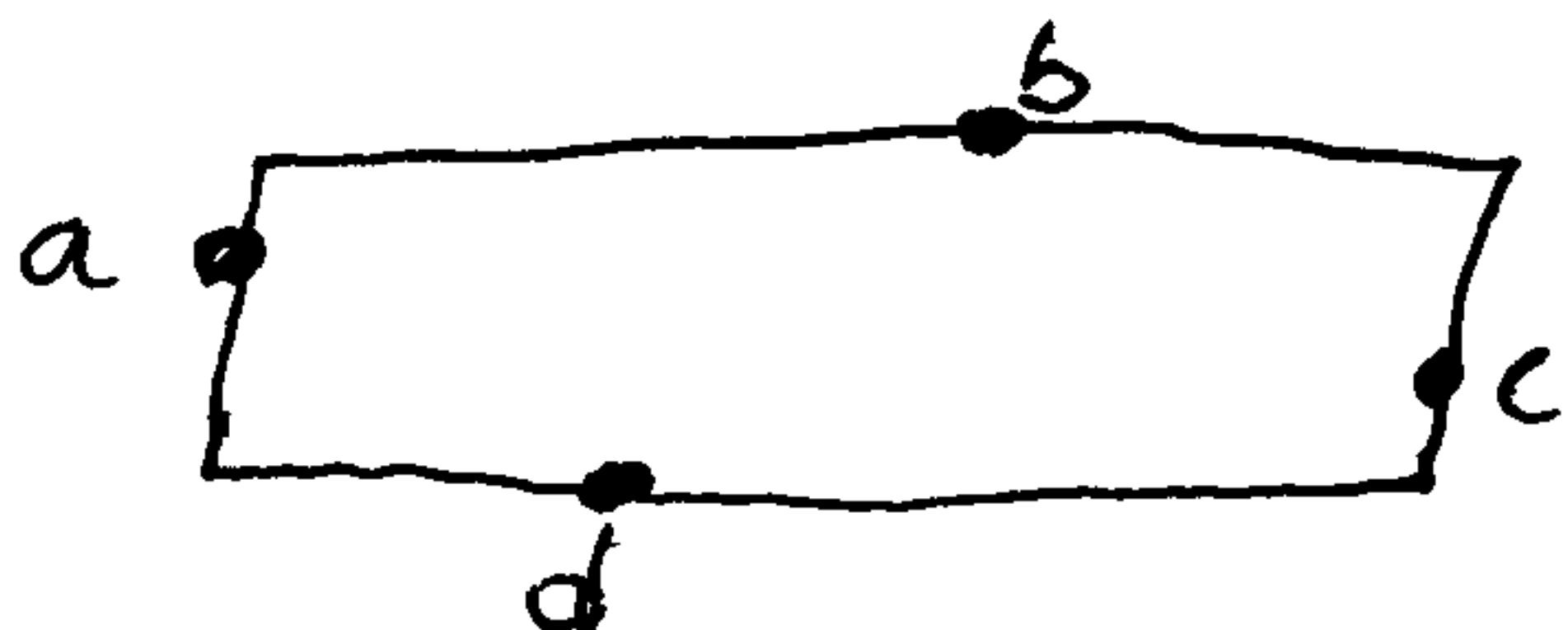


still cannot represent 4 points without also including a 5<sup>th</sup> point

What is the VC dimension of a square?

\*\*Probably 3: can shatter 3 pts. at vertices of an equilateral triangle.

But cannot shatter any 4 points:



start with boundary rectangle (could be shattered with a rectangle, but not with a square)

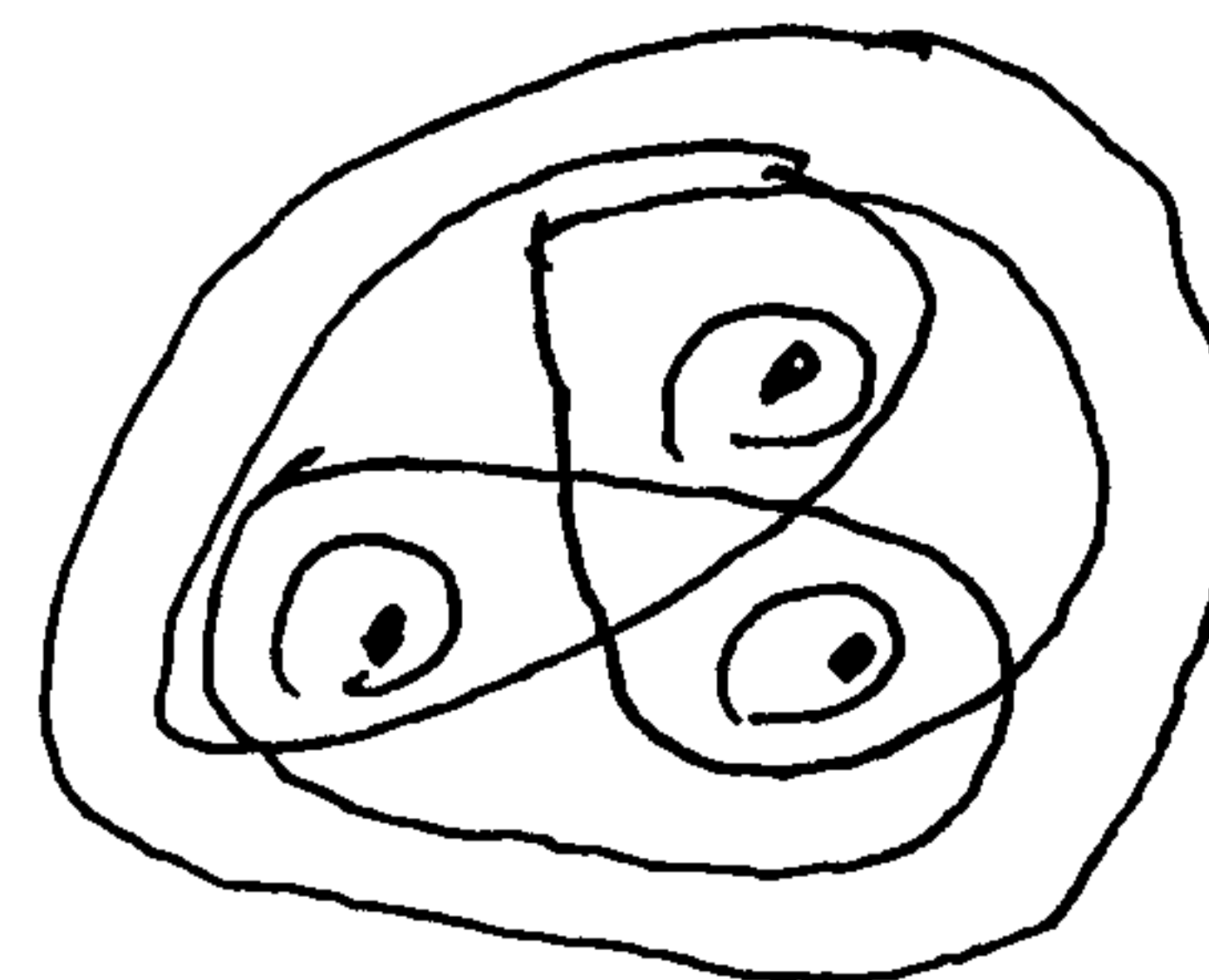


cannot be shattered with a square (cannot even shatter with a rectangle, which is a more general shape)

Cannot find a square containing a and c that doesn't also contain one of b or d.

Circles

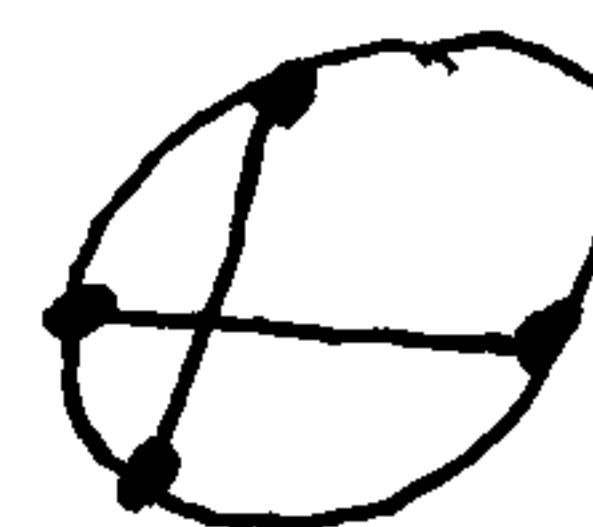
There exist 3 points that can be shattered:



But 4 points cannot be shattered:



cannot shatter



cannot shrink circle such that if one set of points is contained, at least one point of the other set isn't contained as well