Information Theory

a brief tour

Shannon's Information Theory

Suppose $X \in \{1,2,\ldots,n\}$ is a discrete random variable.

How much do you *learn* when you see X?

How many bits are needed to *encode* X?

Encoding

$X \sim \{1, \ldots, N\}$

Natural binary encoding:
write X in binary.

Length:
$|X| = \lceil \log N \rceil$

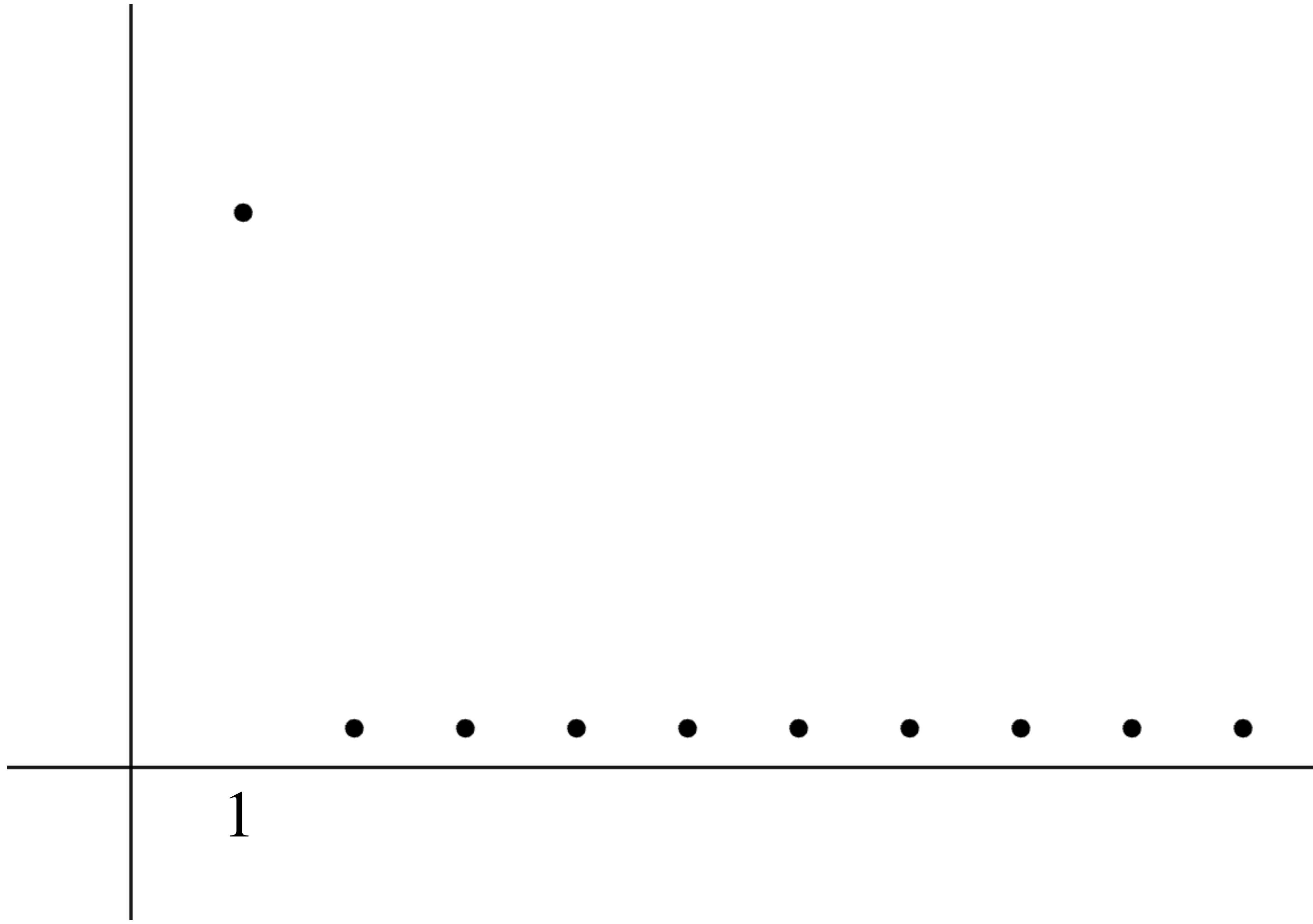
Can we use the PMF of X to do better?

Encode high probability points with short strings?

Encoding

Example

$$P(X = a) = \begin{cases} 1/2 & \text{if } a = 1 \\ 1/(2(n-1)) & \text{otherwise.} \end{cases}$$



$$\text{enc}(X) = \begin{cases} 0 & \text{if } X = 1, \\ 1, \text{ binary encoding of } X & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[\,|\text{enc}(X)|\,] = 1/2 \cdot 1 + 1/2 \cdot (\lceil \log(N-1) \rceil + 1) \lesssim (\log N)/2.$$

Encoding text files

Use short strings for e,t,a,o,i,…

What about video/music files?

What is the principled approach?

Letter	Relative frequency in the English language[1]	
	Texts	Dictionaries[citation needed]
A	8.2%	7.8%
B	1.5%	2.0%
C	2.8%	4.0%
D	4.3%	3.8%
E	12.7%	11.0%
F	2.2%	1.4%
G	2.0%	3.0%
H	6.1%	2.3%
I	7.0%	8.6%
J	0.15%	0.21%
K	0.77%	0.97%
L	4.0%	5.3%
M	2.4%	2.7%
N	6.7%	7.2%
O	7.5%	6.1%
P	1.9%	2.8%
Q	0.095%	0.19%
R	6.0%	7.3%
S	6.3%	8.7%
T	9.1%	6.7%
U	2.8%	3.3%
V	0.98%	1.0%
W	2.4%	0.91%
X	0.15%	0.27%
Y	2.0%	1.6%
Z	0.074%	0.44%

Encoding

A **prefix-free** encoding of $\{1,\ldots,N\}$ a map

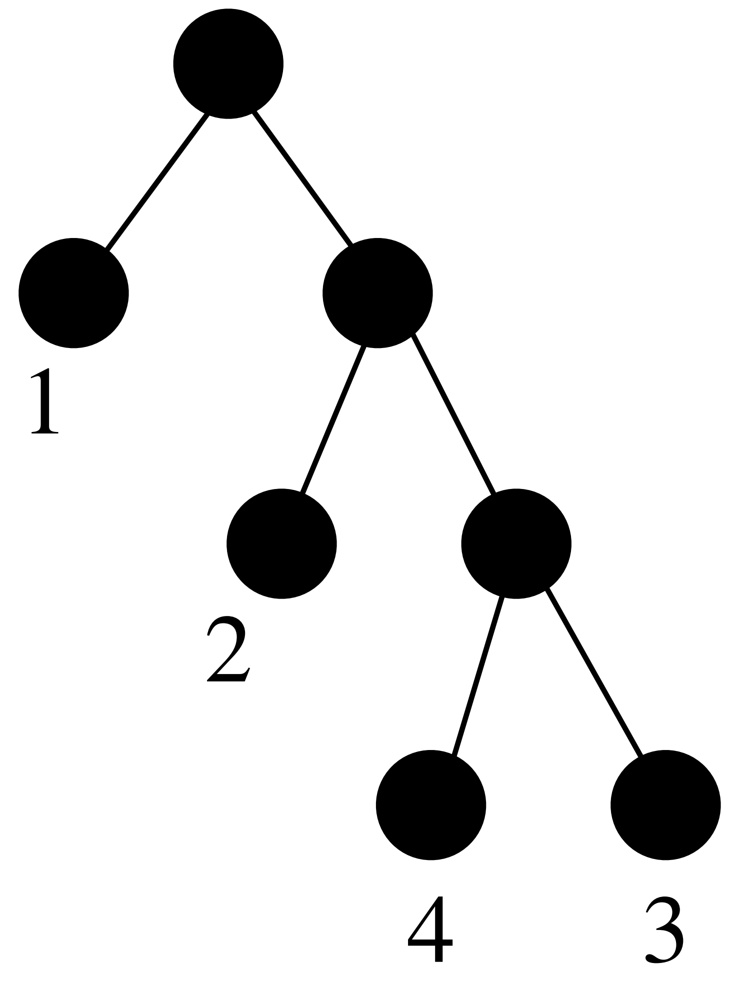
enc : $\{1,\ldots,N\} \to \{0,1\}^*$

so that

If $i \neq j$, then enc(i) is not a prefix of enc(j).

Example: enc$(1) = 0$, enc$(2) = 10$, enc$(3) = 111$, enc$(4) = 110$

0110101011100 uniquely encodes $1,4,2,2,3,1,1$.



Shannon's Idea

Entropy:

$$H(X) = \sum_{x} P(X = x) \cdot \log_2 \frac{1}{P(X = x)}$$

Intuition: If $P(X = a) = 1/2^k$, then we should be able to have $|\text{enc}(a)| \sim k$.

Theorem: There is a prefix-free enc with $|\mathbb{E}[\text{enc}(X)]| \leq H(X) + 1$.
Conversely, every prefix-free enc must have $|\mathbb{E}[\text{enc}(X)]| \geq H(X)$.

Entropy:

$$H(X) = \sum_x P(X=x) \cdot \log_2 \frac{1}{P(X=x)}$$

Theorem: There is a prefix-free encoding with $|\text{enc}(X)| \leq H(X) + 1$.
Conversely, every prefix-free encoding must have $|\text{enc}(X)| \geq H(X)$.

Pf Idea:

First sort the elements $x_1, x_2, \ldots$ so that

$P(X = x_1) \geq P(X = x_2) \geq P(X = x_3)$ ...

Let $k = \lceil \log(1/p(X = x_1)) \rceil$. Find a k bit string to represent x_1.
Let $k = \lceil \log(1/p(X = x_2)) \rceil$. Find a k bit string to represent x_2, but not a superstring of any prev string. Can show that such a string will always be available.

...

Properties of Entropy

Entropy:

$$H(X) = \sum_x P(X = x) \cdot \log_2 \frac{1}{P(X = x)}$$

Fact: If X is uniform, then $H(X) = \log N$.

Fact: $0 \leq H(X) \leq \log N$.

Fact: If $X = X_1, X_2, \ldots, X_n$, then $H(X) \leq H(X_1) + H(X_2) + \ldots + H(X_n)$.

Chain-Rule of Entropy

Entropy:

$$H(X) = \sum_x P(X = x) \cdot \log_2 \frac{1}{P(X = x)}$$

Suppose X, Y are jointly distributed. Write $p(x) = P(X = x)$.

$$\begin{aligned}
H(X, Y) &= \sum_{x,y} p(xy) \cdot \log \frac{1}{p(xy)} \\
&= \sum_{x,y} p(x) \cdot p(y|x) \cdot \log \frac{1}{p(x)p(y|x)} \\
&= \sum_{x,y} p(x) \cdot p(y|x) \cdot \log \frac{1}{p(x)} + \sum_{x,y} p(x) \cdot p(y|x) \cdot \log \frac{1}{p(y|x)} \\
&= \sum_{x} p(x) \cdot \log \frac{1}{p(x)} + \sum_{x,y} p(x) \cdot p(y|x) \cdot \log \frac{1}{p(y|x)} \\
&= H(X) + H(Y|X).
\end{aligned}$$

Chain-Rule of Entropy

Entropy:

$$H(X) = \sum_a P(X = a) \cdot \log_2(1/P(X = a))$$

Suppose X, Y are jointly distributed. Write $p(x) = P(X = x)$.

$$H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$$

So:

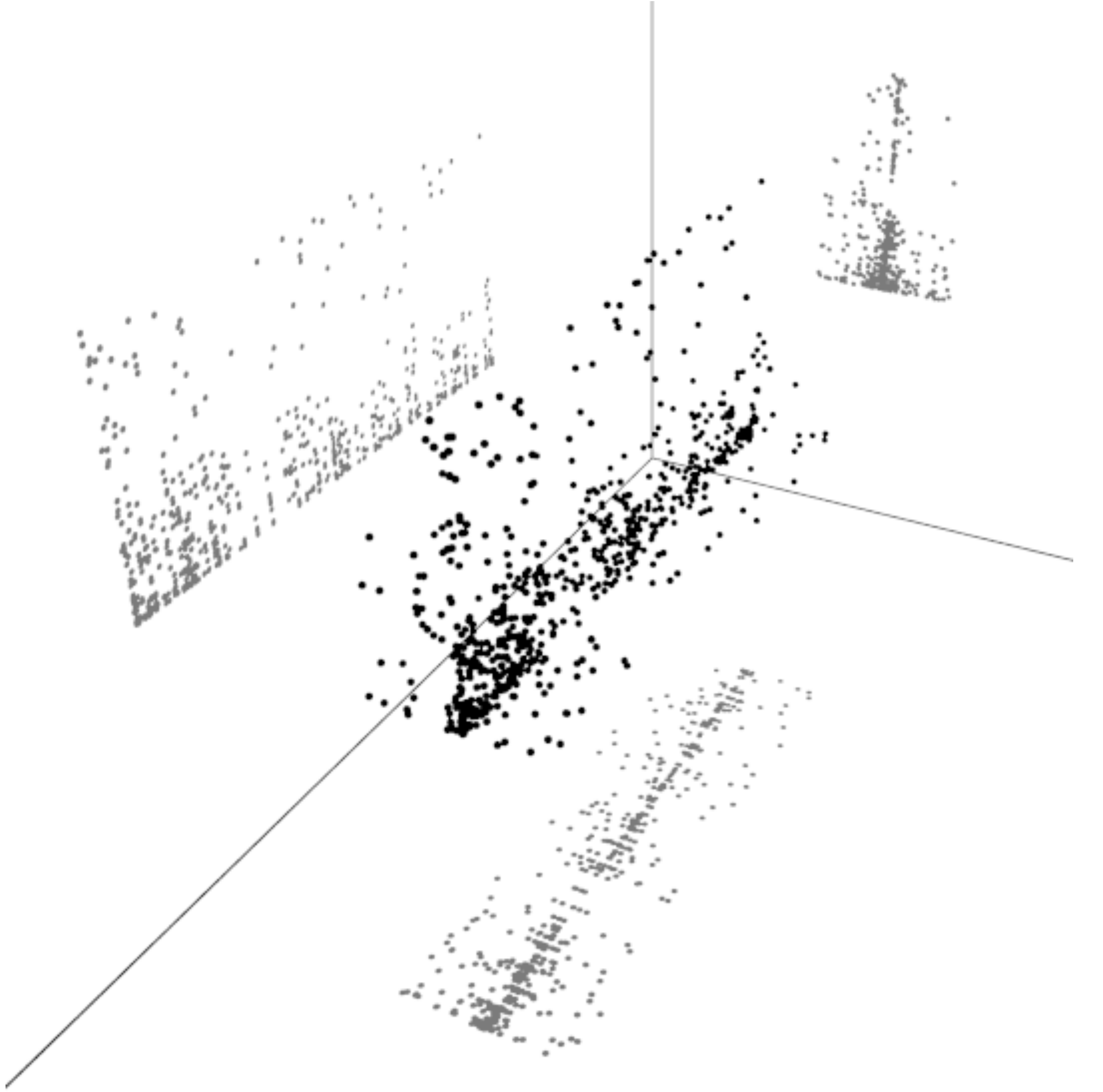
$$H(Y|X) \leq H(Y).$$

Example: Loomis-Whitney inequality

Suppose S is a set of N^3 points in 3 dimensional space.
$S = \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\}$

Let $S_x = \{x_1, \ldots, x_N\}, S_y = \{y_1, \ldots, y_N\}, S_z = \{z_1, \ldots, z_N\}$

Claim: One of S_x, S_y, S_z must be of size $\geq N$.



Example: Loomis-Whitney inequality

Suppose S is a set of N^3 points in 3 dimensional space.
$S = \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\}$

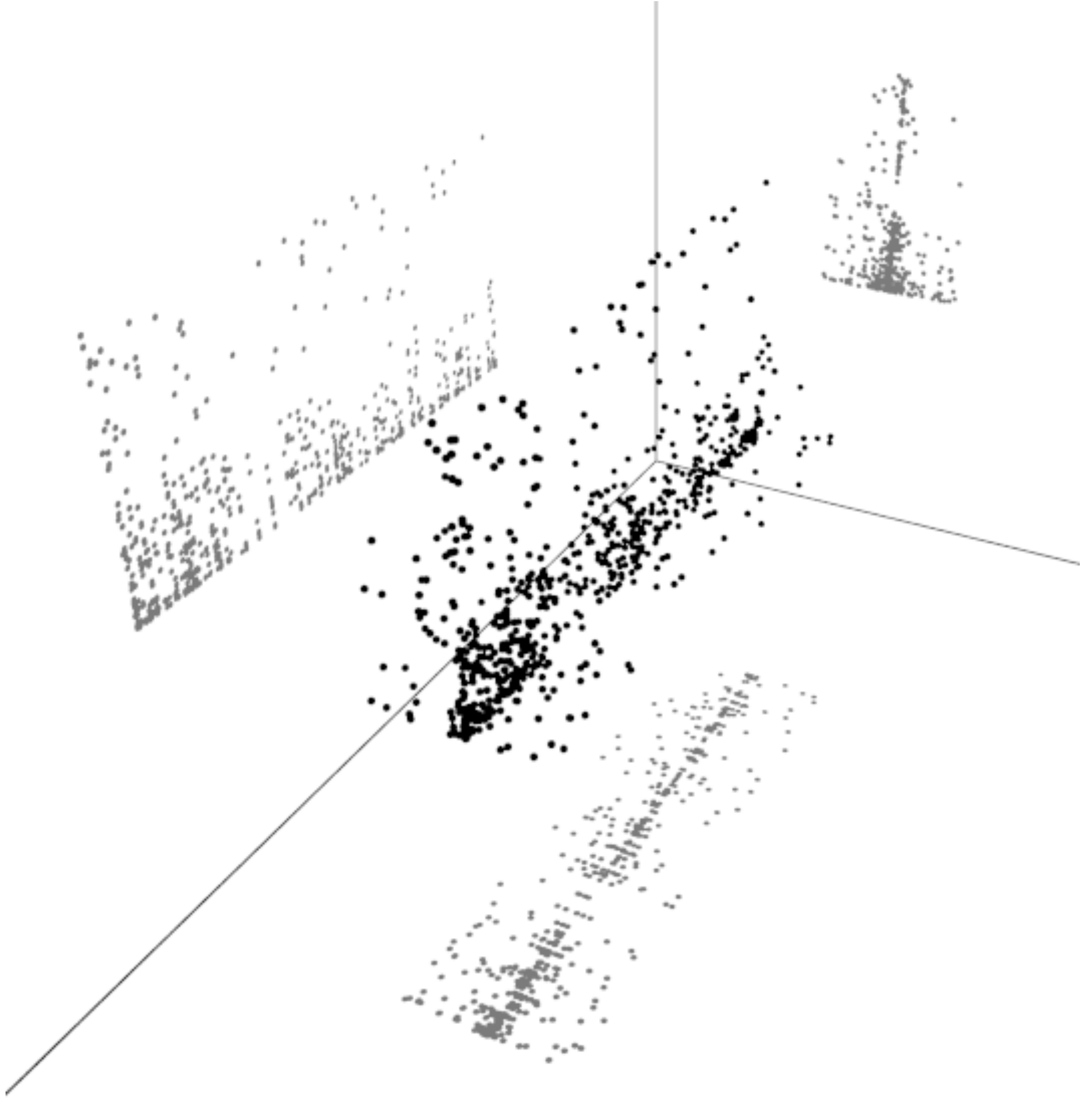
Let $S_x = \{x_1, \ldots, x_N\}, S_y = \{y_1, \ldots, y_N\}, S_z = \{z_1, \ldots, z_N\}$

Claim: One of S_x, S_y, S_z must be of size $\geq N$.

Pf: Let (X, Y, Z) be a random point.

$3 \log N = \log N^3 = H(X, Y, Z) \leq H(X) + H(Y) + H(Z),$

So one of those terms is at least $\log N$, and the corresponding set is of size $\geq N$.



Example: Loomis-Whitney inequality

Suppose S is a set of N^3 points in 3 dimensional space.

$S = \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\}$

Let $S_x = \{x_1, \ldots, x_N\}$, $S_y = \{y_1, \ldots, y_N\}$, $S_z = \{z_1, \ldots, z_N\}$

Claim: One of S_x, S_y, S_z must be of size $\geq N$.

Pf: Let (X, Y, Z) be a random point.

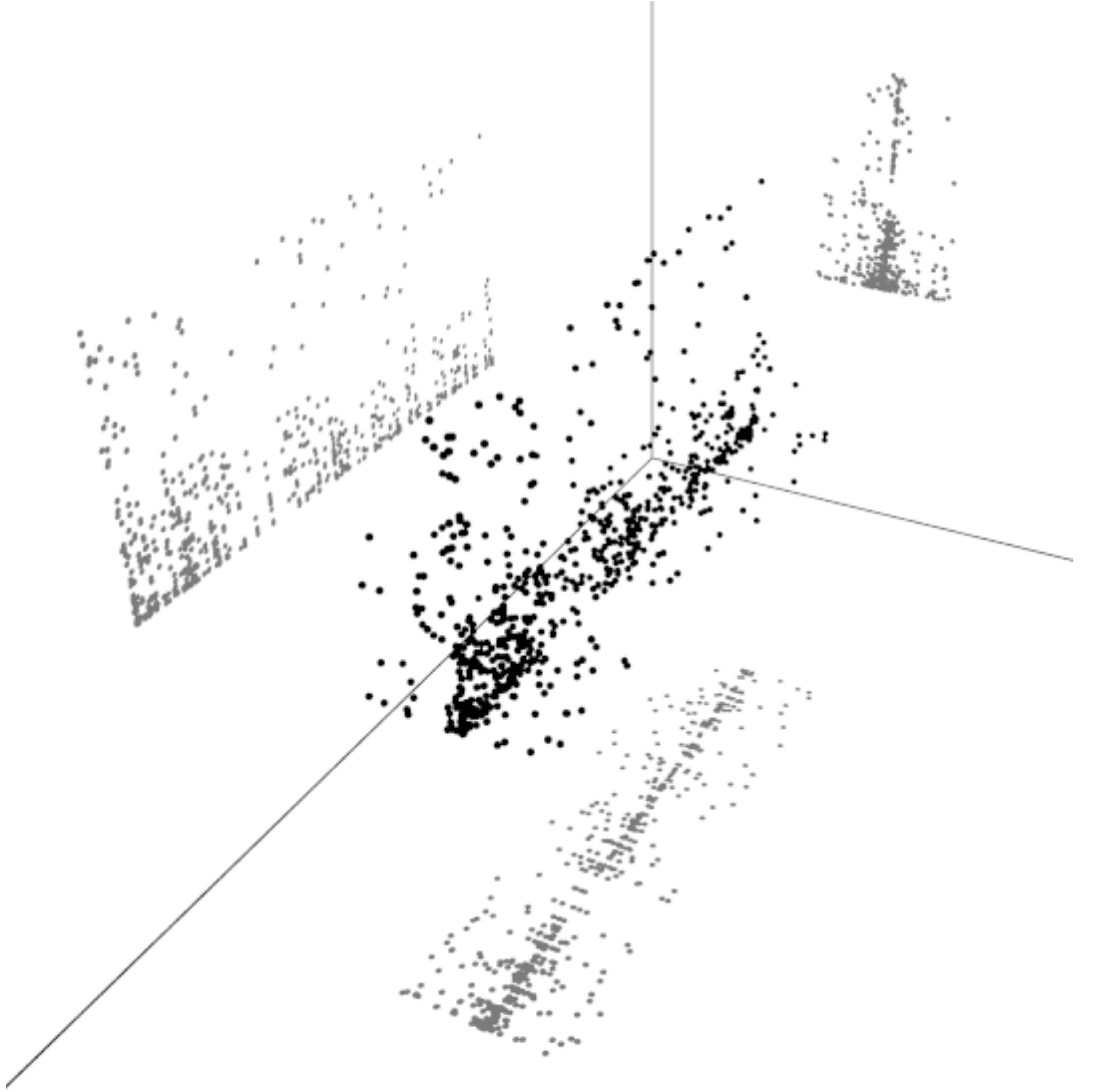
$3 \log N = \log N^3 = H(X, Y, Z) \leq H(X) + H(Y) + H(Z)$,

So one of those terms is at least $\log N$, and the corresponding set is of size $\geq N$.

Let $S_{xy} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, $S_yz = \{(y_1, z_1), \ldots, (y_N, z_N)\}$,

$S_{zx} = \{(x_1, z_1), \ldots, (x_N, z_N)\}$

Claim: One of these three must be of size N^2.

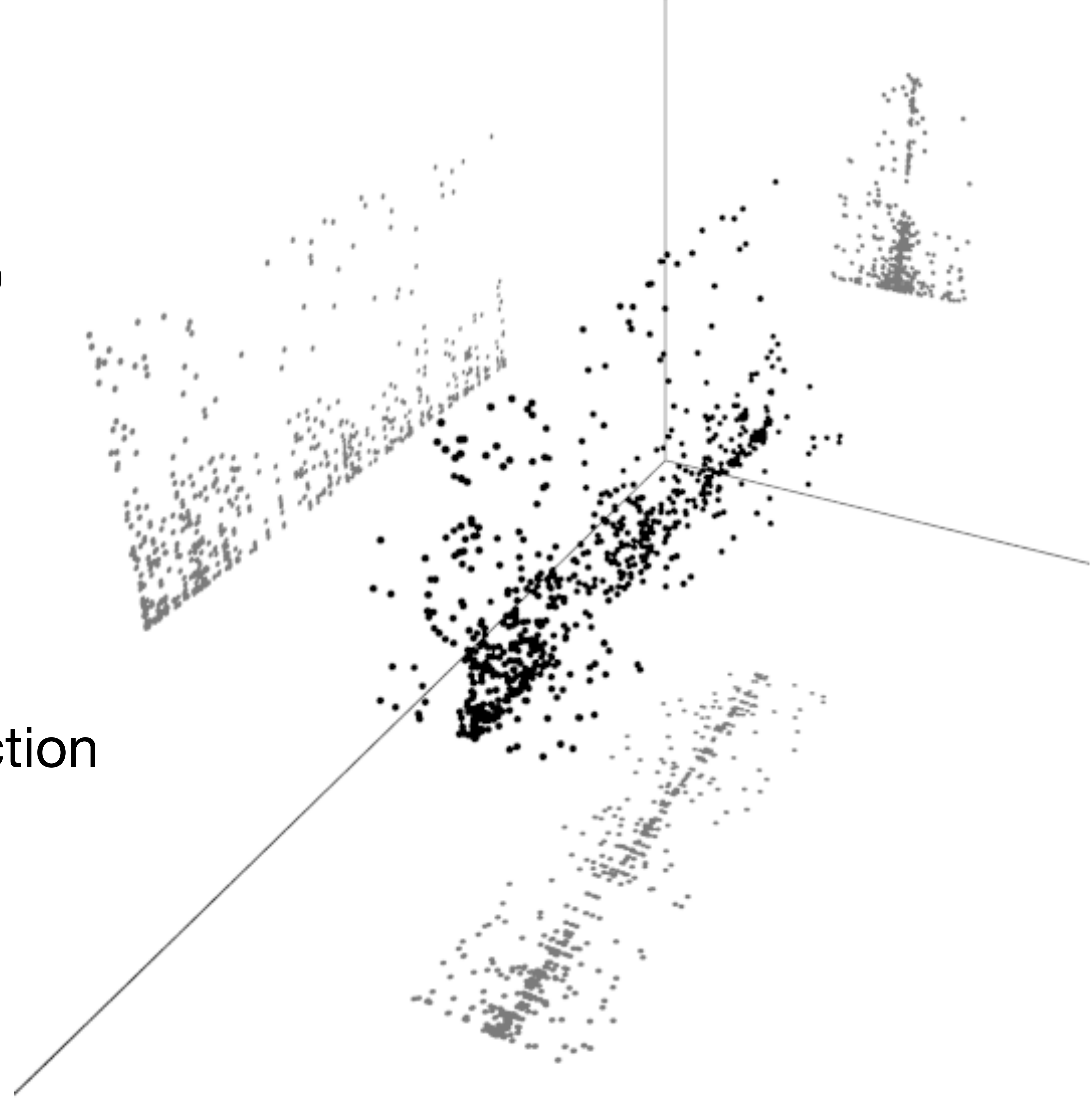


$H(XY) = H(X) + H(Y|X).$

Let $S_{xy} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, $S_yz = \{(y_1, z_1), \ldots, (y_N, z_N)\}$, $S_{zx} = \{(x_1, z_1), \ldots, (x_N, z_N)\}$

Claim: One of these three must be of size N^2.
Pf:

$$
\begin{aligned}
6 \log N = 2 \cdot H(XYZ) &= 2 \cdot H(X) + 2 \cdot H(Y|X) + 2 \cdot H(Z|XY) \\
&\leq H(X) + H(Y|X) \\
&\quad + H(X) + H(Z|X) \\
&\quad\quad + H(Y) + H(Z|Y) \\
&= H(XY) + H(YZ) + H(ZX).
\end{aligned}
$$



So, one of these terms is $\geq 2 \log N$ and the corresponding projection is of size $\geq N^2$.

Union Closed Sets Conjecture

$\mathscr{F}$: a family of subsets of $\{1, 2, \ldots, n\}$.

Def: $\mathscr{F}$ is closed under union if $A, B \in \mathscr{F}$ implies $A \cup B \in \mathscr{F}$.

Conjecture: If $\mathscr{F}$ is closed under union, there is $i \in \{1, 2, \ldots, n\}$ that belongs to at least half the sets in $\mathscr{F}$.

Example: $\mathscr{F}$ is all subsets.

Union Closed Sets Conjecture

$\mathscr{F}$: a family of subsets of $\{1,2,\ldots,n\}$.

Def: $\mathscr{F}$ is closed under union if $A, B \in \mathscr{F}$ implies $A \cup B \in \mathscr{F}$.

Conjecture: If $\mathscr{F}$ is closed under union, there is $i \in \{1,2,\ldots,n\}$ that belongs to at least half the sets in $\mathscr{F}$.

Theorem: If $\mathscr{F}$ is closed under union, there is $i \in \{1,2,\ldots,n\}$ that belongs to at least $1 - 1/\phi$ fraction of the sets in $\mathscr{F}$.

Where $\phi = \dfrac{1+\sqrt{5}}{2}$ is the golden ratio.

Entropy, a review

A : random variable with distribution $p(a)$.

$$H(A) = \sum_a p(a) \cdot \log(1/p(a)) = -\mathbb{E}[\log p(a)].$$

1. *Chain rule:*
$H(AB) = -\mathbb{E}[\log p(a,b)] = -\mathbb{E}[\log(p(a) \cdot p(b\,|\,a))] = -\mathbb{E}[\log p(a)] - \mathbb{E}[\log p(b\,|\,a)] = H(A) + H(B\,|\,A)$.

2. *Subadditivity*: $H(AB) \le H(A) + H(B)$

Pf: $H(B\,|\,A) = \sum_{a,b} p(a,b) \cdot \log 1/p(b\,|\,a) = \sum_b p(b) \sum_a p(a\,|\,b) \cdot \log 1/p(b\,|\,a) \le \sum_b p(b) \cdot \log \sum_a p(a\,|\,b)/p(b\,|\,a) = H(B)$

3. *Uniform distribution has largest entropy:* $H(A) \le \log|\text{supp}(A)|$.

Pf: $H(A) = \sum_a p(a) \cdot \log 1/p(a) \le \log \sum_a 1 = \log|\text{supp}(A)|$

Binary entropy function:
$h(p) = p \cdot \log 1/p + (1-p) \cdot \log 1/(1-p)$.

Theorem: If $\mathscr{F}$ is closed under union, there is $i \in \{1,2,\ldots,n\}$ that belongs to at least $1 - 1/\phi$ fraction of the sets in $\mathscr{F}$.

Pf: Suppose not. Let $A, B \in \mathscr{F}$ be independent and uniform. Let $C = A \cup B$.
Think of $A, B, C \in \{0,1\}^n$.

Claim: $H(C) > H(A)$. (contradiction!)

$$H(C) = \sum_{i=1}^{n} H(C_i \,|\, C_{<i})$$

subadditivity

$$\geq \sum_{i=1}^{n} H(C_i \,|\, A_{<i}, B_{<i})$$

by technical claim

$$> \sum_{i=1}^{n} H(A_i \,|\, A_{<i}) = H(A)\,.$$

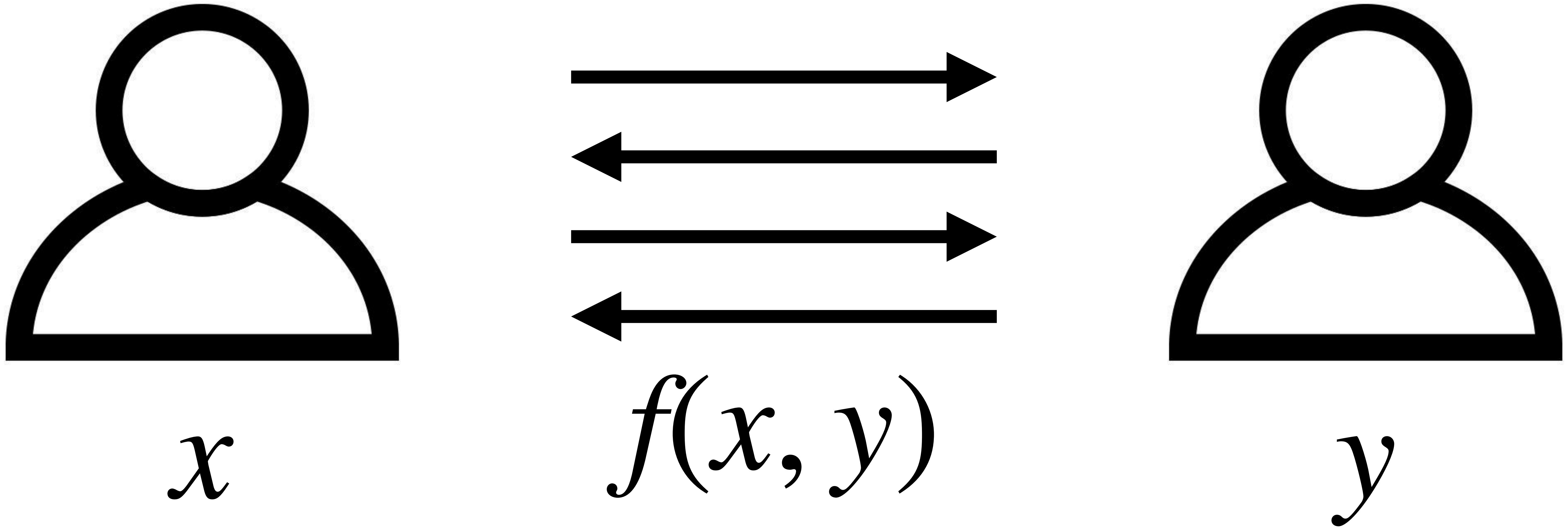
$$p = \Pr(A_i = 0 \,|\, A_{<i})$$
$$q = \Pr(B_i = 0 \,|\, B_{<i})$$

Technical Claim: If $p, q \sim \mu$, and $\mathbb{E}[p] > 1/\phi$, then $\mathbb{E}[h(pq)] > \mathbb{E}[h(p)]$.

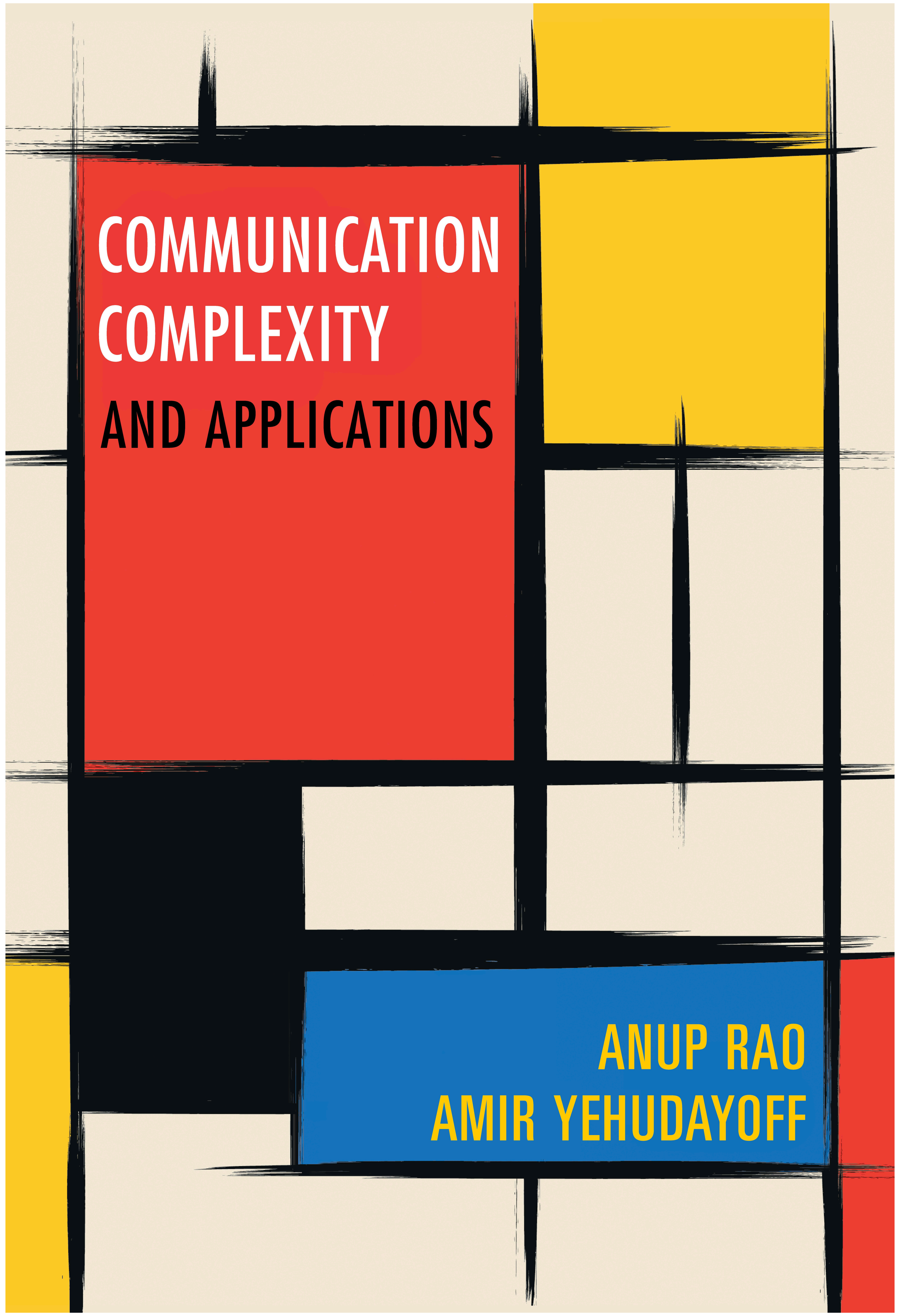
equivalently

$$\mathbb{E}[2 \cdot h(pq) - h(p) - h(q)] > 0$$

Communication Protocols



How many bits do they need to exchange?


COMMUNICATION COMPLEXITY
AND APPLICATIONS
ANUP RAO
AMIR YEHUDAYOFF

Complexity of Repetition

$$f^n(x_1, y_1, \ldots, x_n, y_n) = f(x_1, y_1), \ldots, f(x_n, y_n)$$

Does computing f^n require more communication than computing f ?

Theorem: Yes, the communication should scale by $\gtrsim \sqrt{n}$.

Proof Idea: If C bits are enough to compute f^n, then $C/\sqrt{n}$ bits are enough to compute f .

$$f^n(x_1, y_1, \ldots, x_n, y_n) = f(x_1, y_1), \ldots, f(x_n, y_n)$$

Does computing f^n require more communication than computing f ?

Theorem: Yes, the communication should scale by $\gtrsim \sqrt{n}$.

Proof Idea:

1. If there is a C-bit protocol computing f^n, there is a C-bit protocol computing f with *information* C/n computing.
2. Every such protocol can be *compressed* to get a $C\sqrt{n}$ - bit protocol.

$$f^n(x_1, y_1, \ldots, x_n, y_n) = f(x_1, y_1), \ldots, f(x_n, y_n)$$

1. If there is a C-bit protocol computing f^n, there is a C-bit protocol computing f with *information* C/n computing.

If x, y, m are inputs and messages, information is:

$$\mathop{\mathbb{E}}_{x,y,m}\left[\log \frac{p(m|xy)}{p(m|x)} + \log \frac{p(m|xy)}{p(m|y)}\right].$$