Your Data is Blue, But Can You Prove it?

Interactive Proofs for Distribution Properties

(Mostly) based on works with Guy Rothblum

Your Dat

A SURVEY ON DISTRIBUTION TESTING

Your Data is Big. But is it Blue?

Clément L. Canonne[1]

May 8, 2015

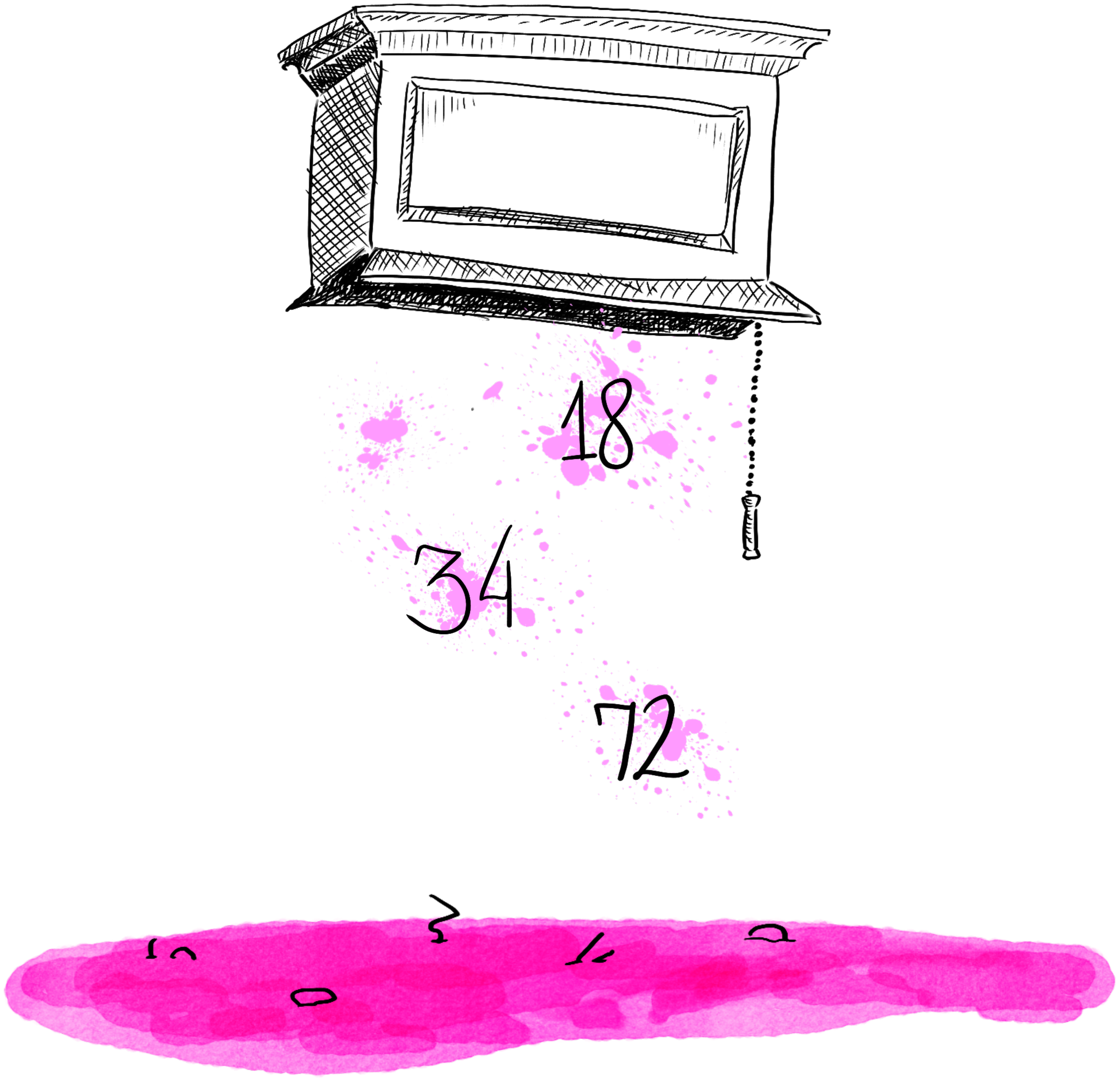
Prove it?

Interactive Proofs for Distribution Properties

(Mostly) based on works with Guy Rothblum

Introduction: Testing Distributions

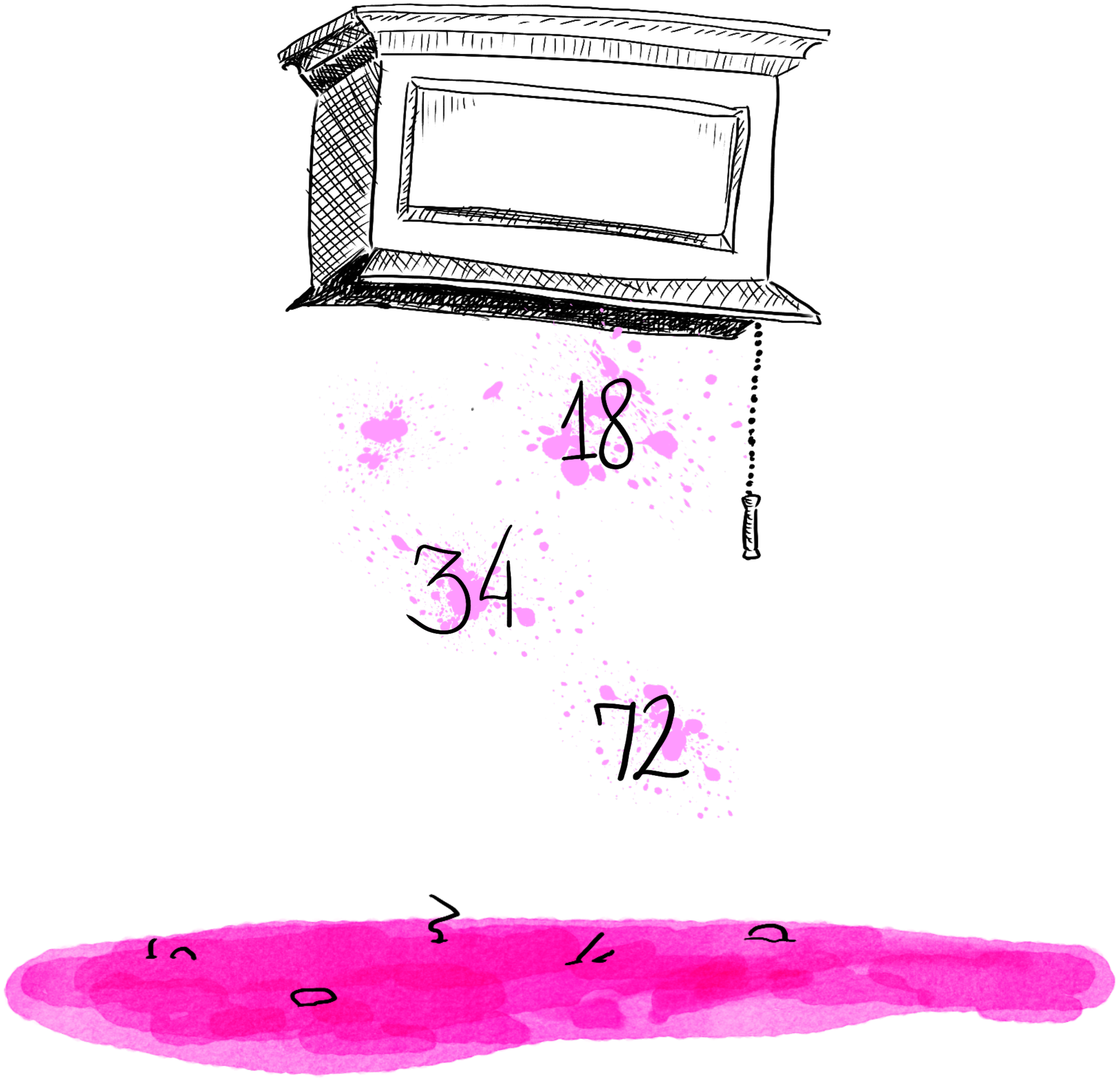
Distribution D over $[N]$



Introduction: Testing Distributions

Distribution D over [N]

Learn the distribution! O(N)

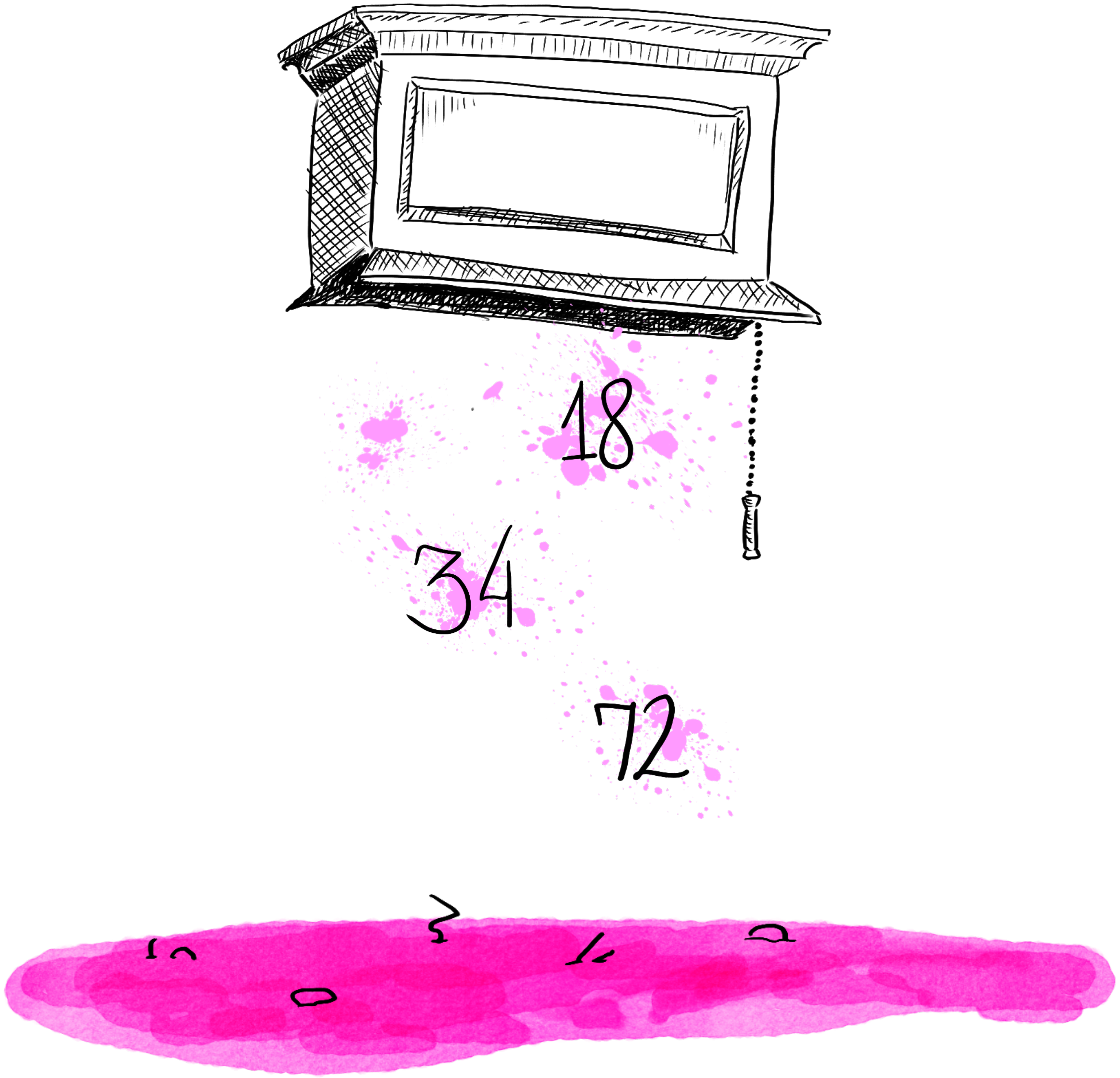


Introduction: Testing Distributions

Learn the distribution! O(N)

Easy: approximate mean, variance. $\mathrm{poly}(\varepsilon^{-1})$

Distribution D over [N]



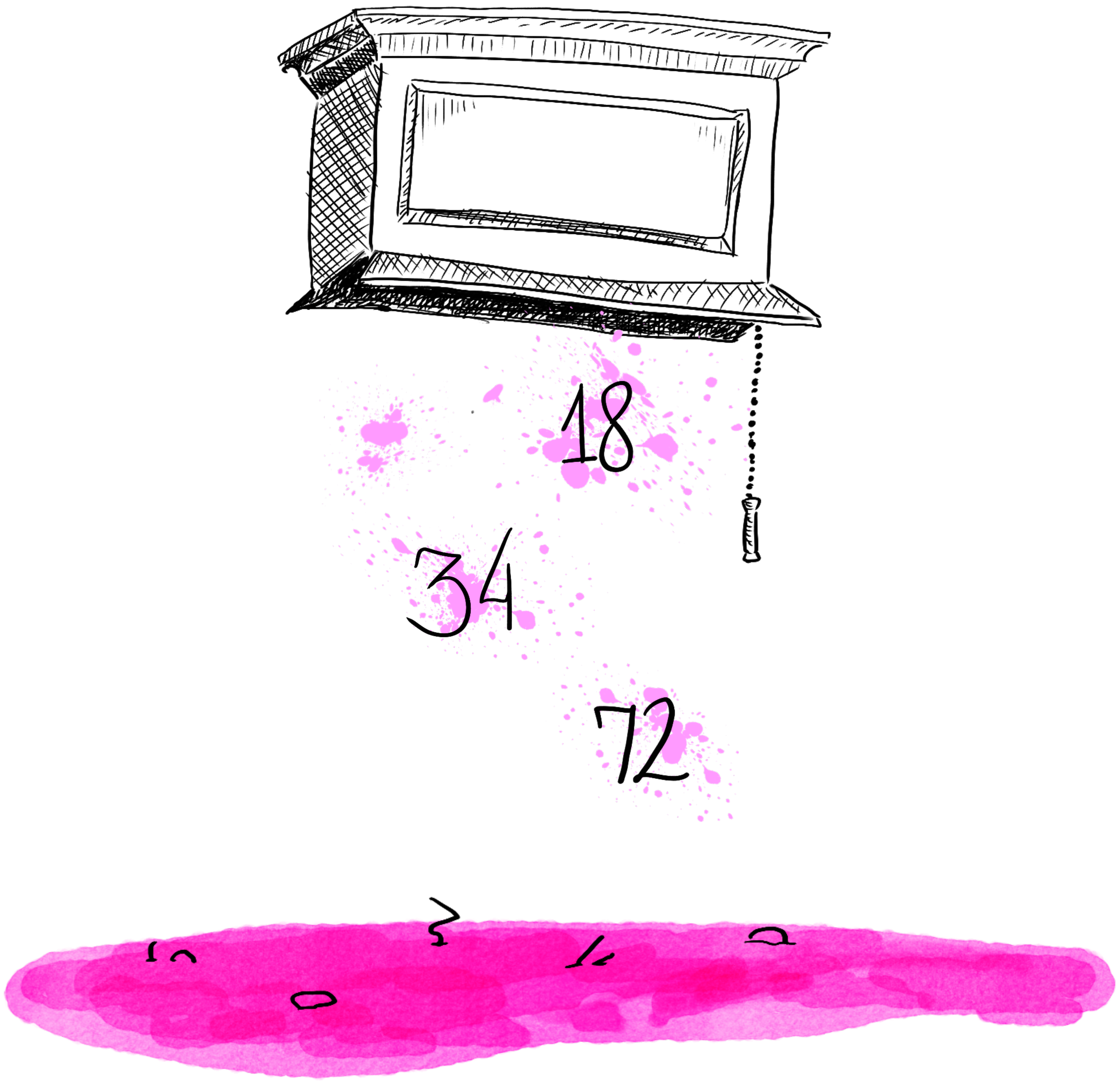
Introduction: Testing Distributions

Distribution D over [N]

Learn the distribution! $O(N)$

Easy: approximate mean, variance. $\mathrm{poly}(\varepsilon^{-1})$

Hard: is ***D*** uniform or *far* from it? $\Theta\left(\sqrt{N}\right)$
[GR'00,...,Pan'08]



Introduction: Testing Distributions

Distribution D over [N]

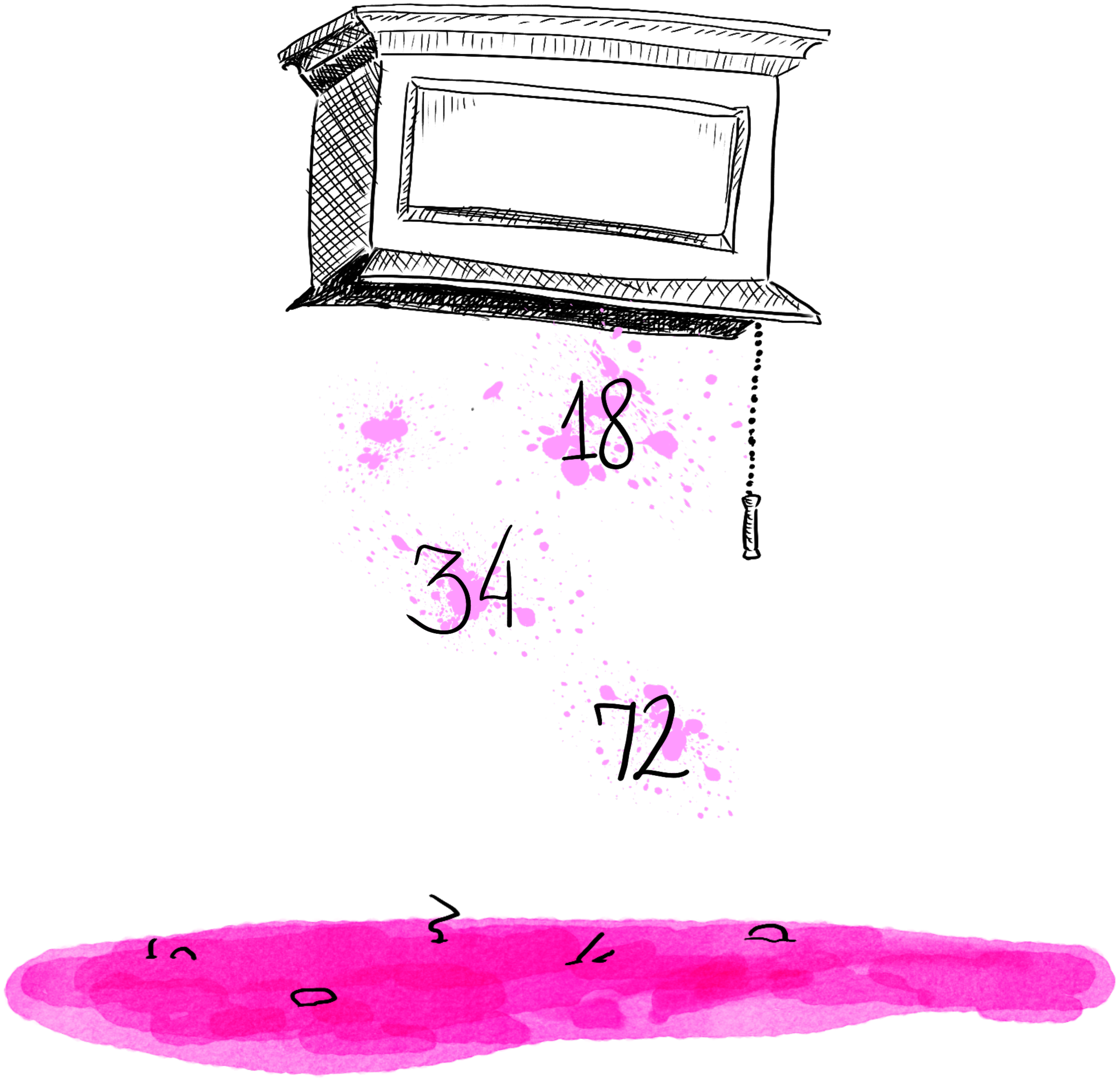
Learn the distribution! $O(N)$

Easy: approximate mean, variance. $\mathrm{poly}(\varepsilon^{-1})$

Hard: is ***D*** uniform or *far* from it? $\Theta\left(\sqrt{N}\right)$ [GR'00,...,Pan'08]

Harder: $\Theta(N / \log N)$

- Is the distribution ***monotone?*** [BFRV'10,AGPRY'19,]
- Approximating ***distance from uniform***. [VV'10]
- Approximating ***Shannon entropy.*** [VV'10]



Introduction: Testing Distributions

Distribution D over [N]

Learn the distribution! O(N)

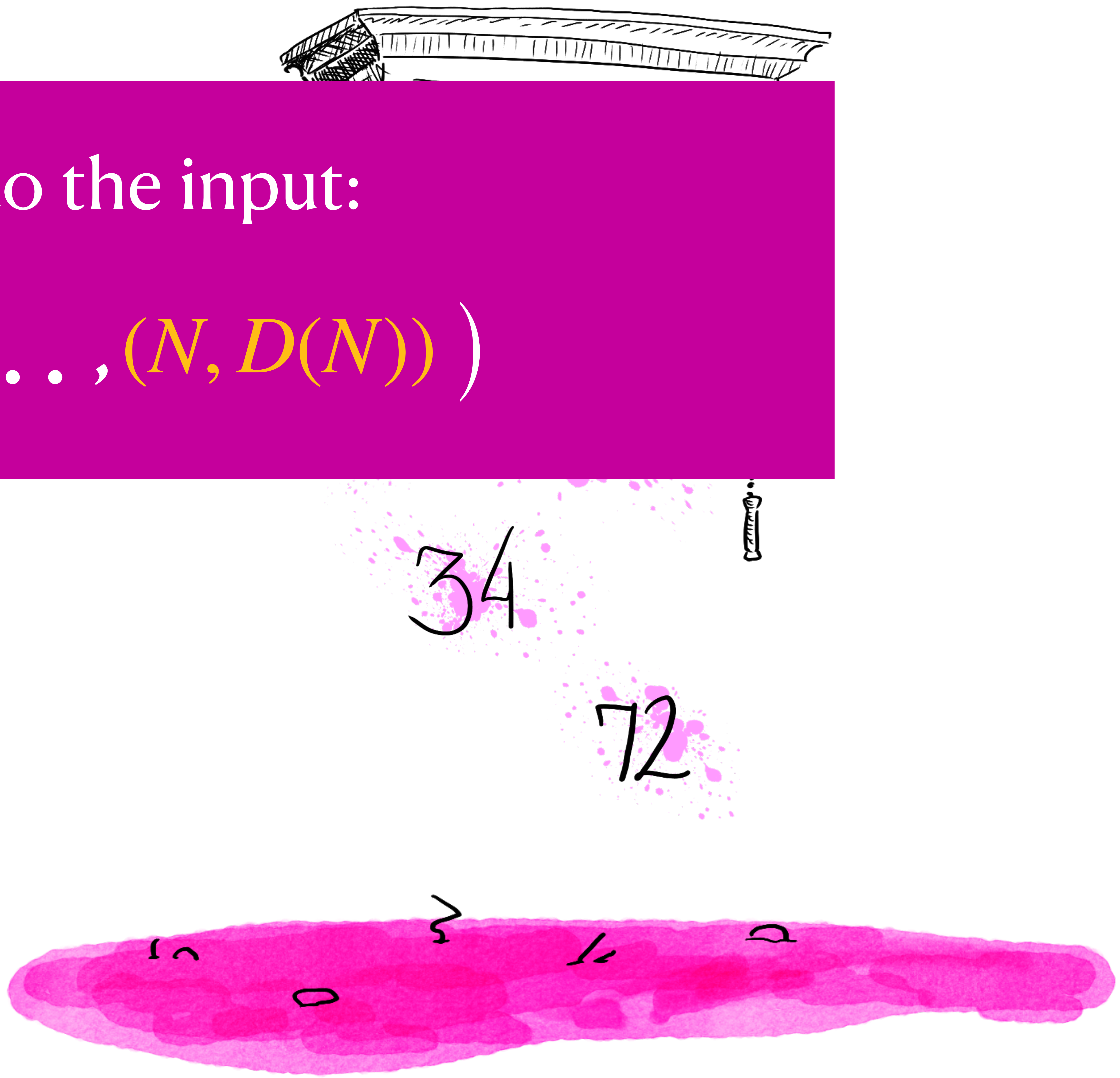
Easy: appr

Hard: is ***D***

Harder: Θ

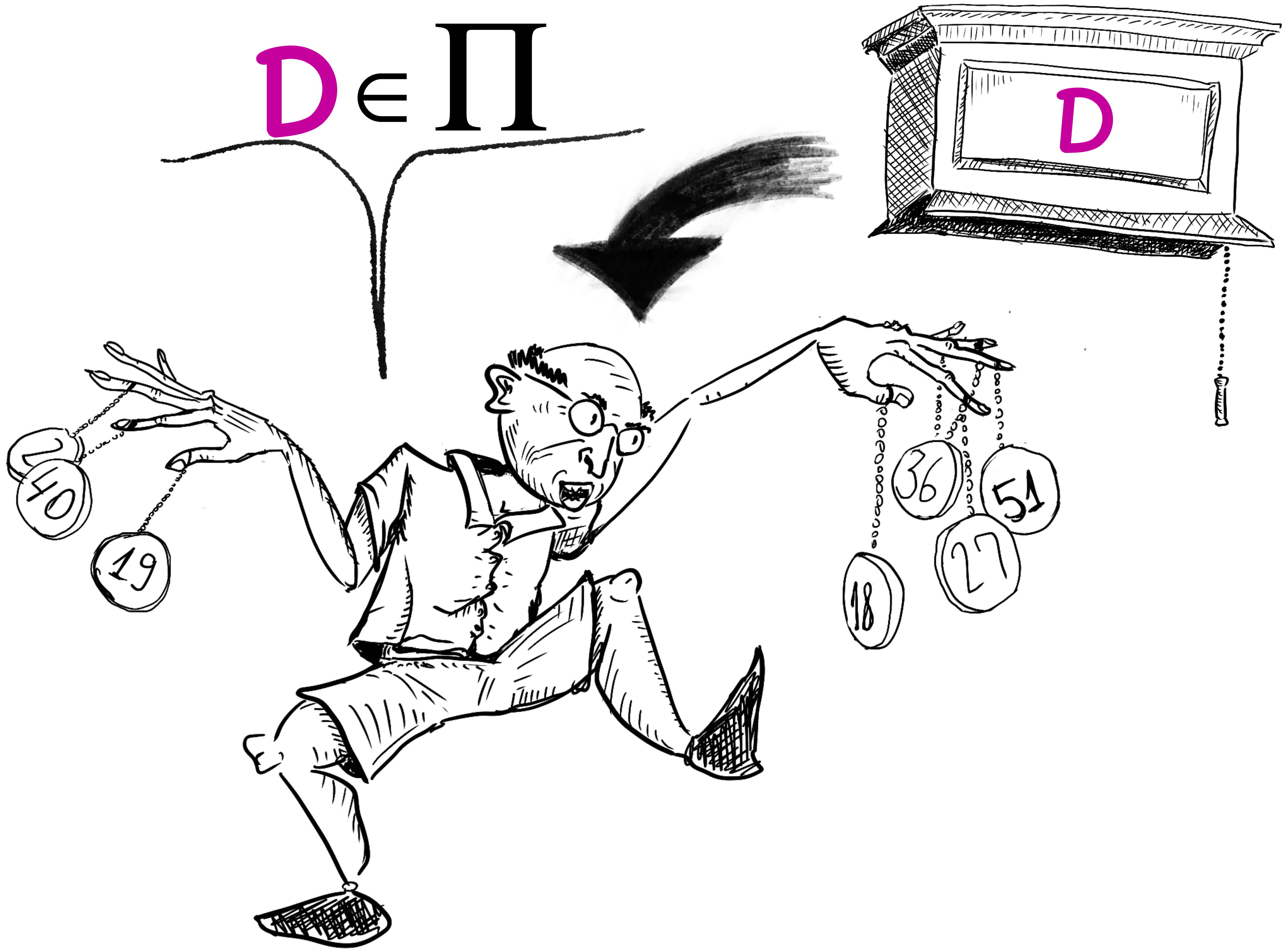
Very limited access to the input:

$$\big(\ (1, D(1))\ ,\ (2, D(2)),\ \ldots\ , (N, D(N))\ \big)$$

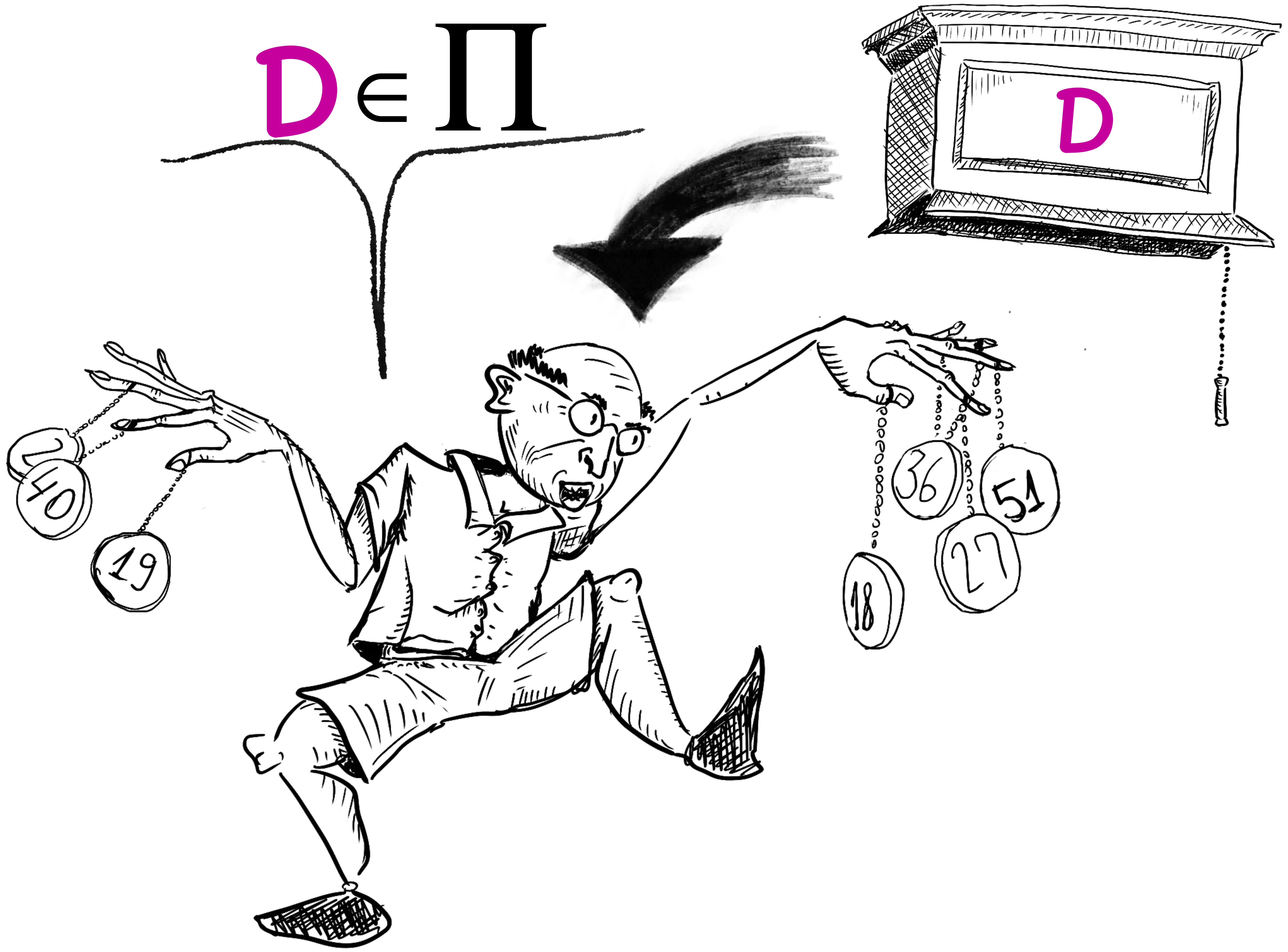


- Is the distribution ***monotone?*** [BFRV'10,AGPRY'19,]
- Approximating ***distance from uniform***. [VV'10]
- Approximating ***Shannon entropy.*** [VV'10]

Verifying Distribution Properties: Warmup

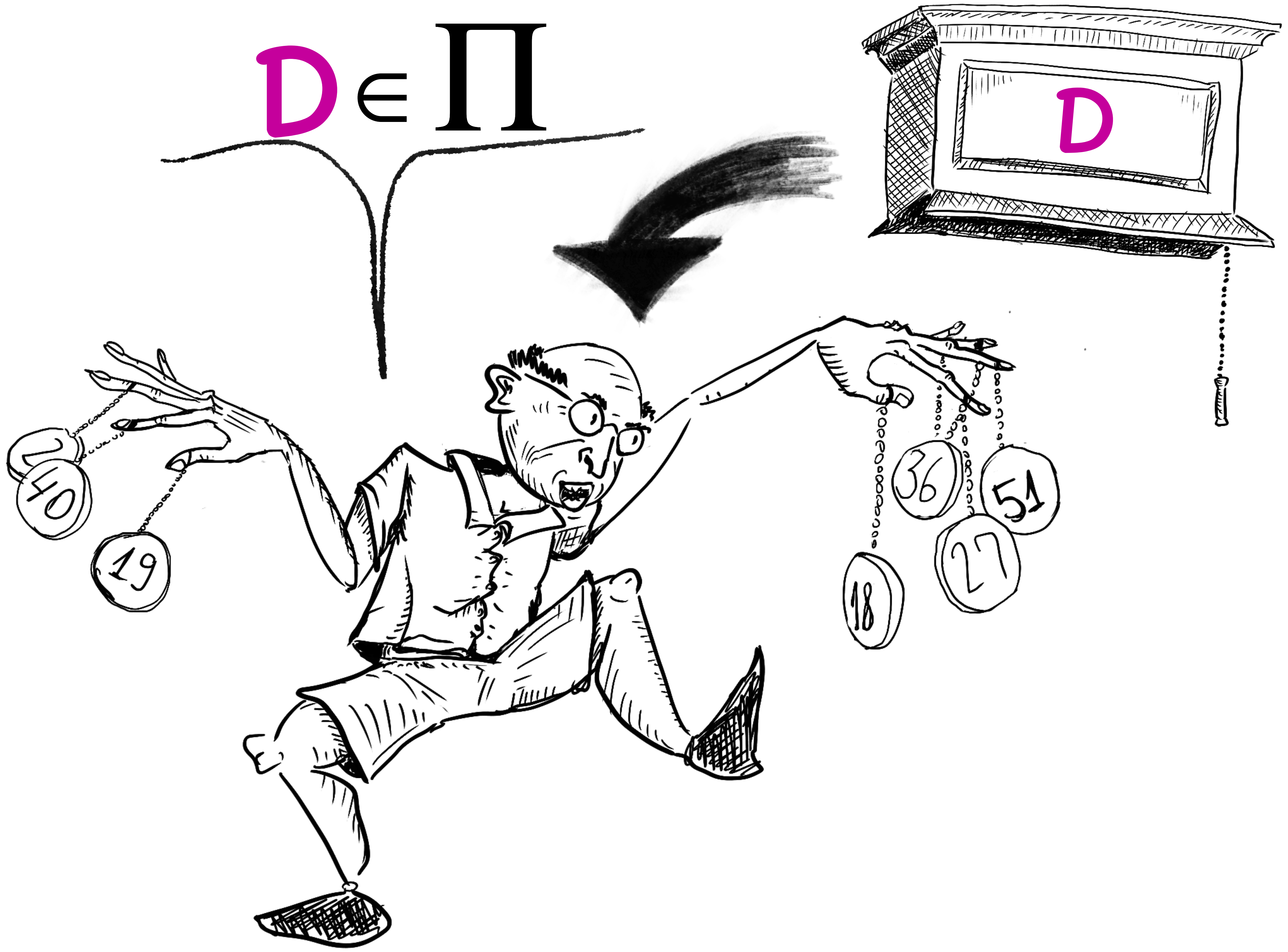


Verifying Distribution Properties: Warmup



Q: Can we *verify* claims with fewer samples?

Verifying Distribution Properties: Warmup



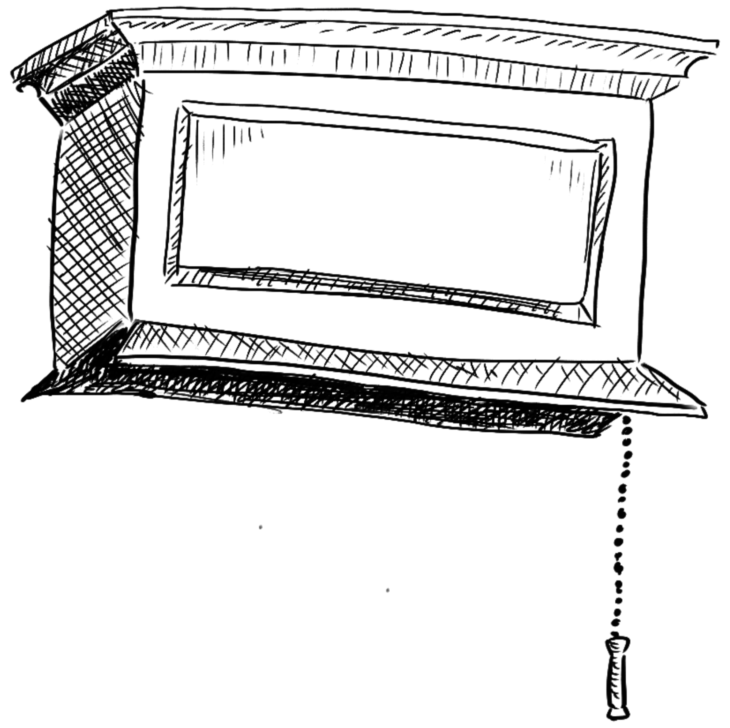
Q: Can we *verify* claims with fewer samples?

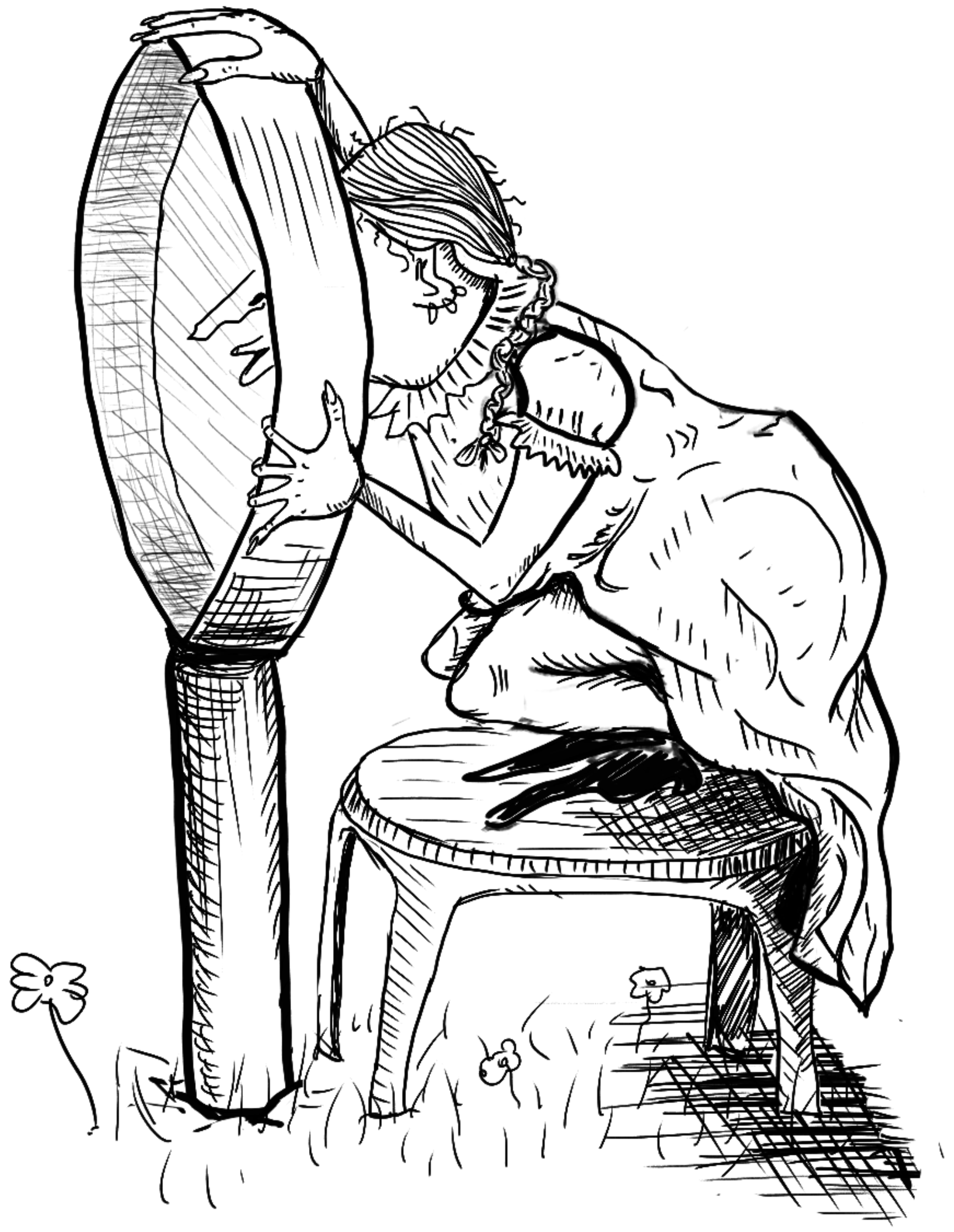
- *Verifiable **data science**!*
- *Verifying AI?*
- *Natural theoretical question - new perspectives and insights on old problems.*

Verifying Distribution Properties: Warmup

Distribution D over $[N]$

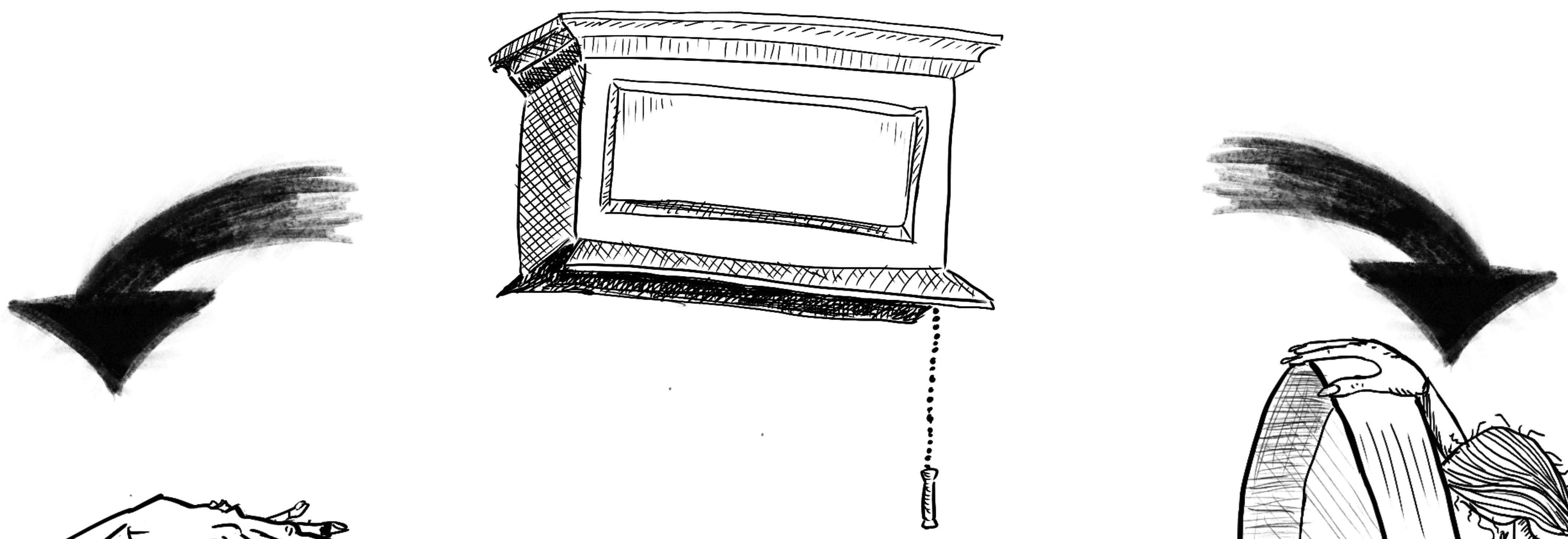
$D \in \Pi$



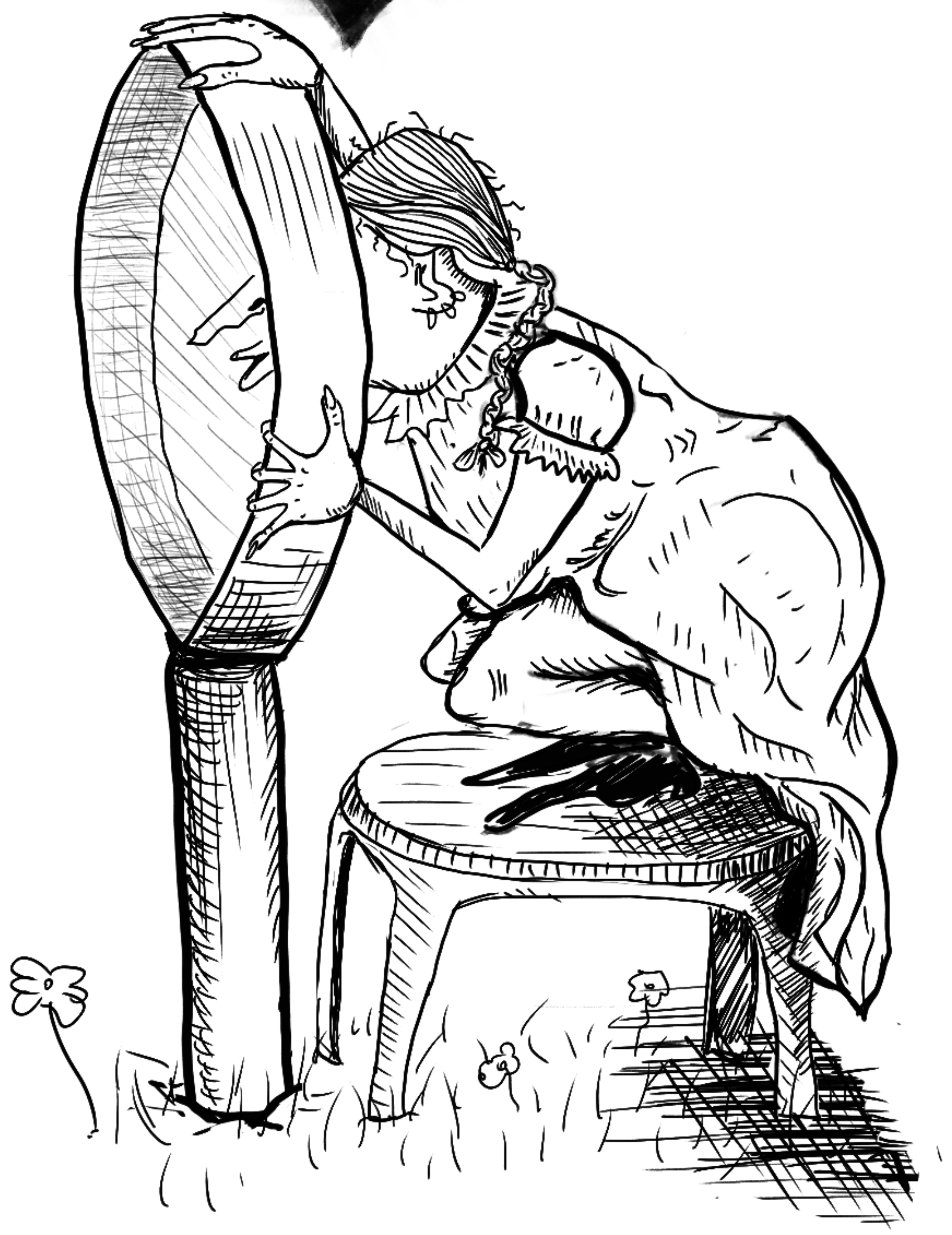


Verifying Distribution Properties: Warmup

Distribution D over $[N]$

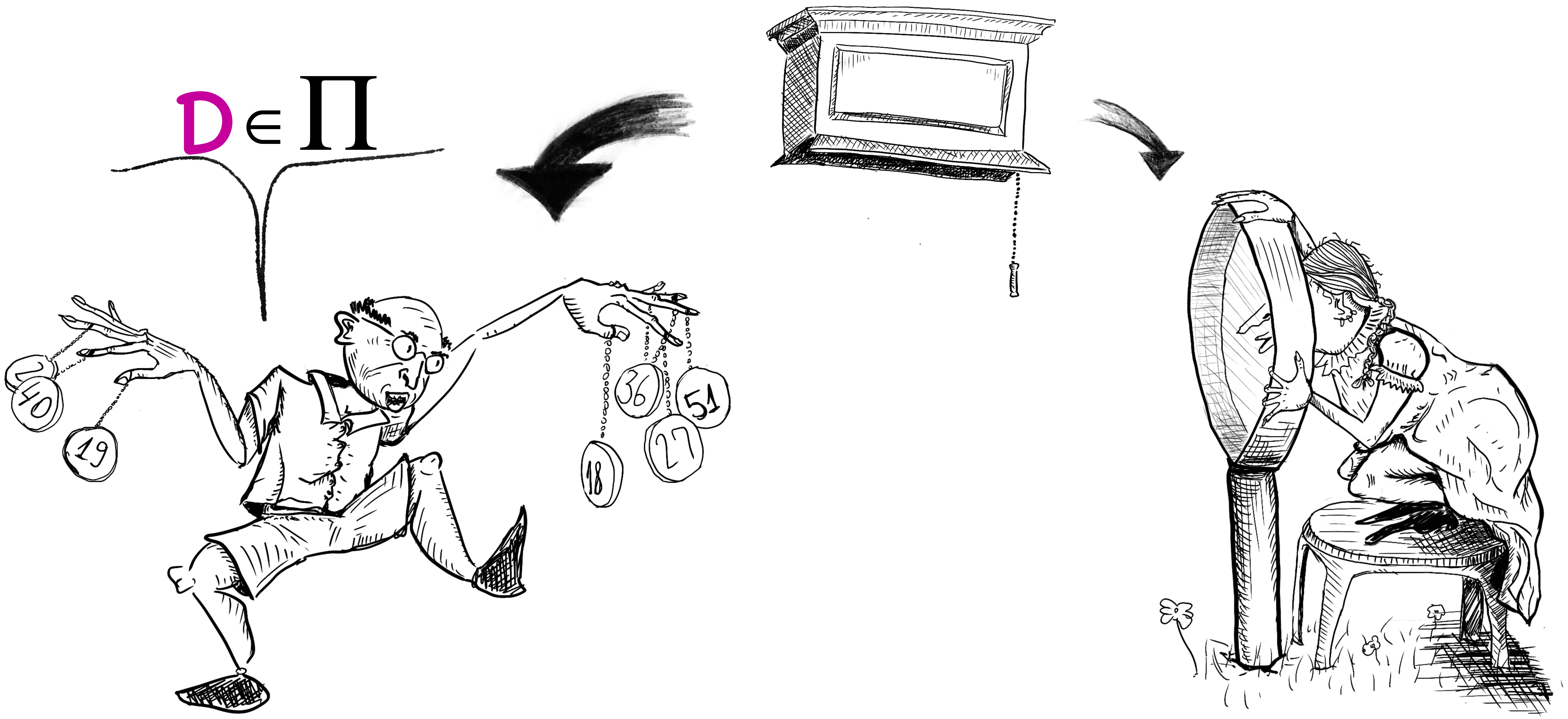


$D \in \Pi$



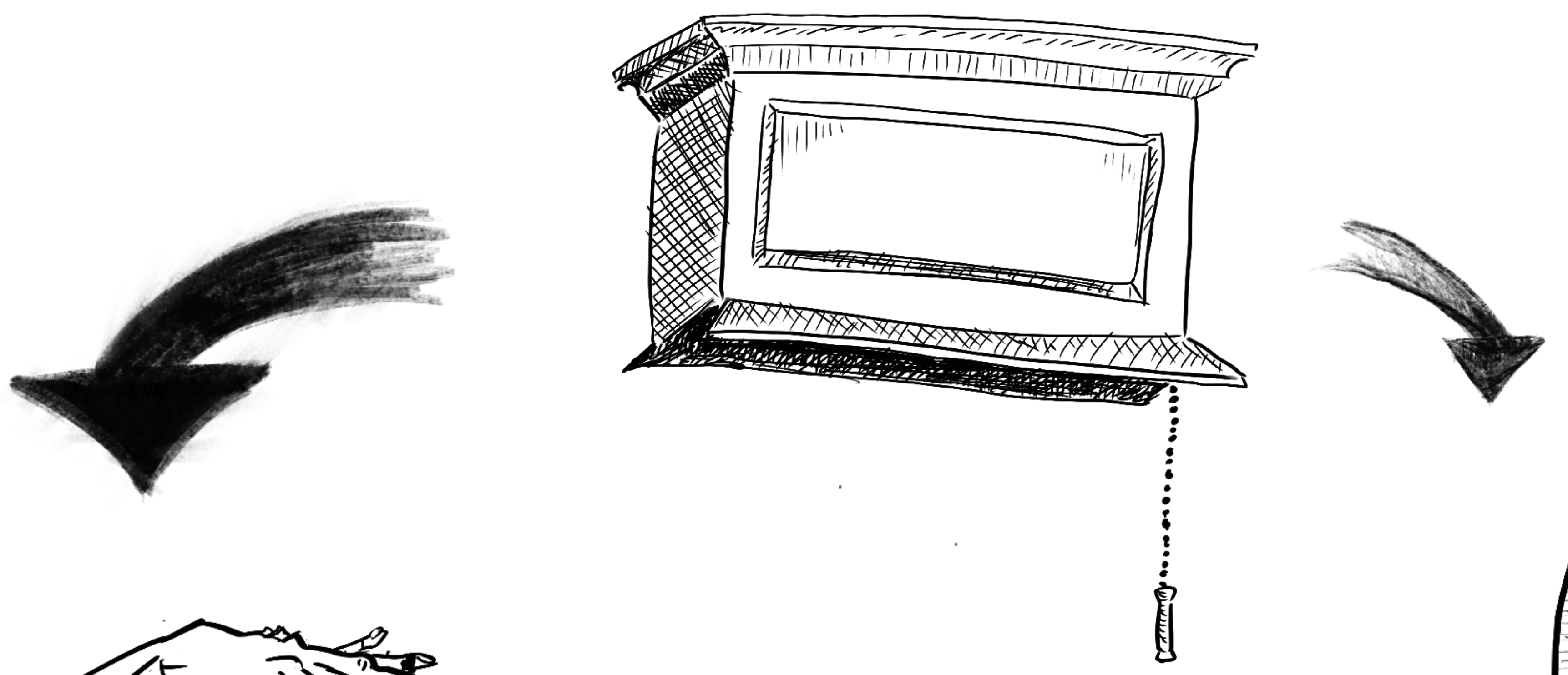
Verifying Distribution Properties: Warmup

Distribution D over $[N]$



Verifying Distribution Properties: Warmup

Distribution D over $[N]$



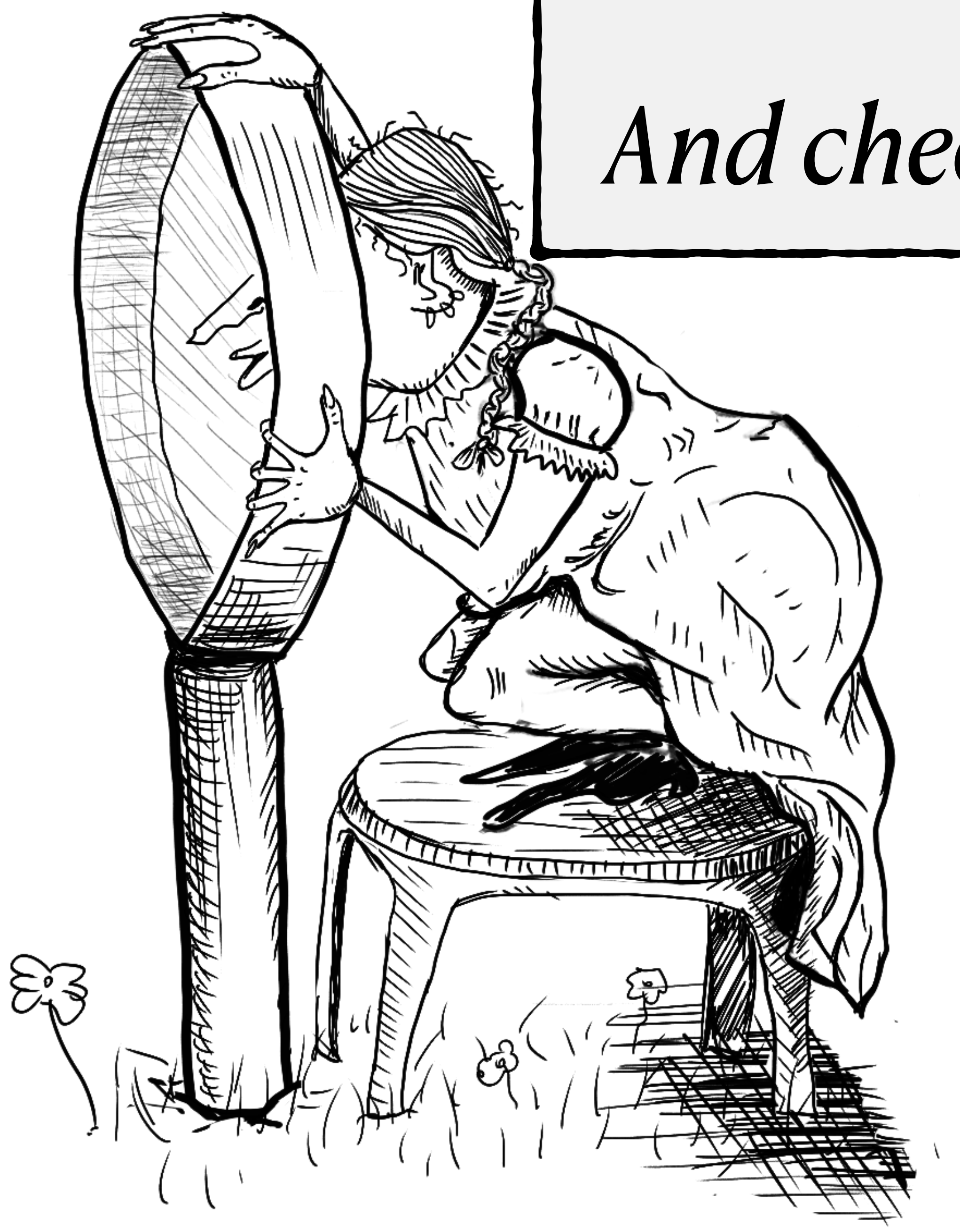
Run ***Identity Tester*** *[BFFKR'01, VV'14]* *to verify:*

$$\delta_{TV}\left(D, \tilde{D}\right) \leq 0.01$$

And check $\tilde{D}$ *is close to* Π

$D \in \Pi$

$\left(\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_N\right)$



Samples: $O\left(N\right)$

Comm.: $\tilde{O}\left(N\right)$

Samples: $O\left(\sqrt{N}\right)$

Verifying Distribution Properties: Warmup

Distribution D over [N]

*Run **Identity Tester** [BFFKR'01, VV'14]*

to verify:

$$\delta_{TV}\left(D, \tilde{D}\right) \leq 0.01$$

And check $\tilde{D}$ is close to Π

Can we do better?

Samples: $O(N)$

Comm.: $\tilde{O}(N)$

Samples: $O\left(\sqrt{N}\right)$

Verifying Distribution Properties: Warmup

Distribution D over [N]

Run ***Identity Tester*** *[BFFKR'01, VV'14] to verify:*

$$\delta_{TV}\left(D, \tilde{D}\right) \leq 0.01$$

And check $\tilde{D}$ is close to Π

Can we do better?

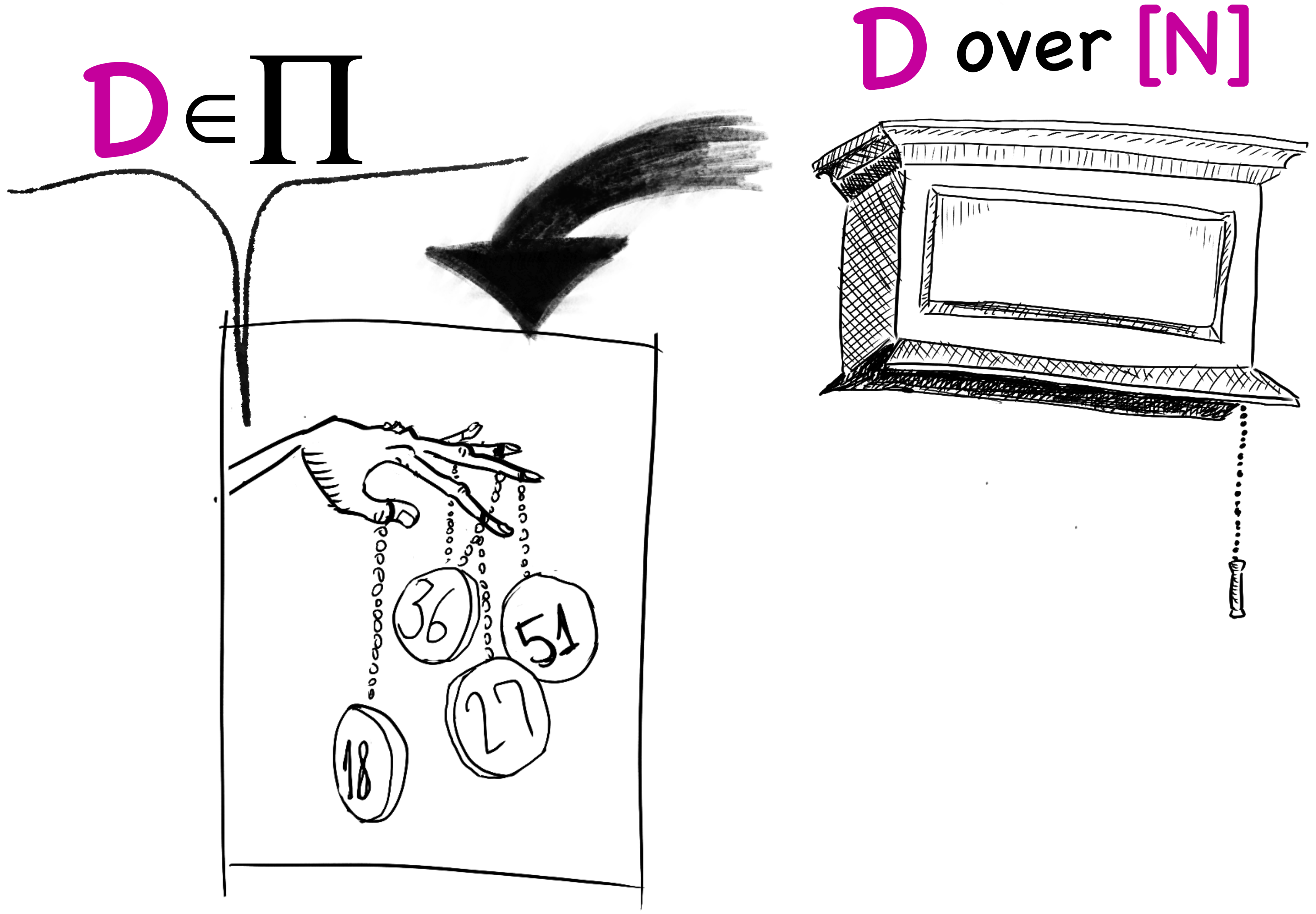
What properties are efficiently verifiable?

Samples: $O(N)$

Comm.: $\tilde{O}(N)$

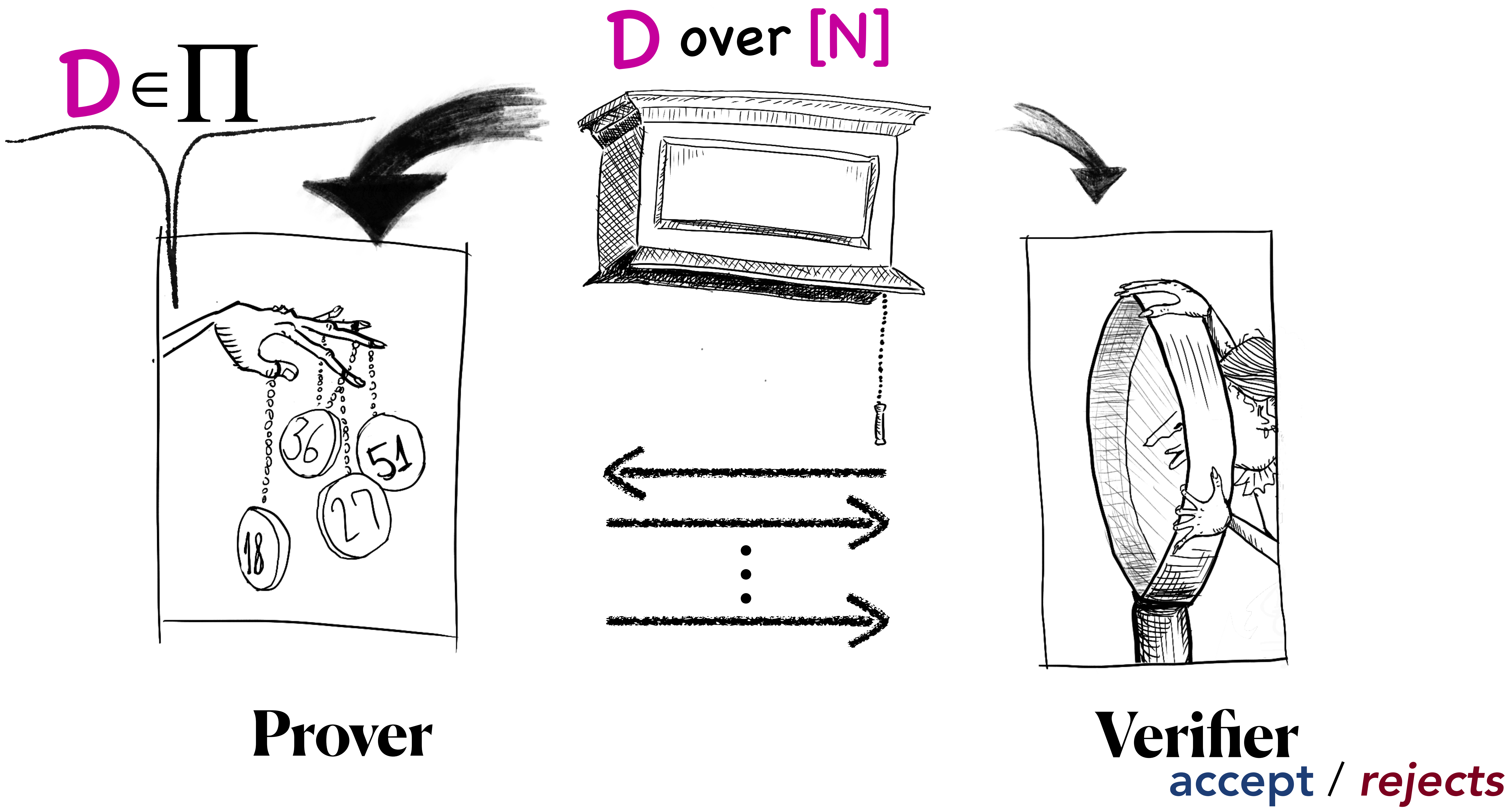
Samples: $O\left(\sqrt{N}\right)$

Proof Systems for Distributions [GMR'85, CG'17]

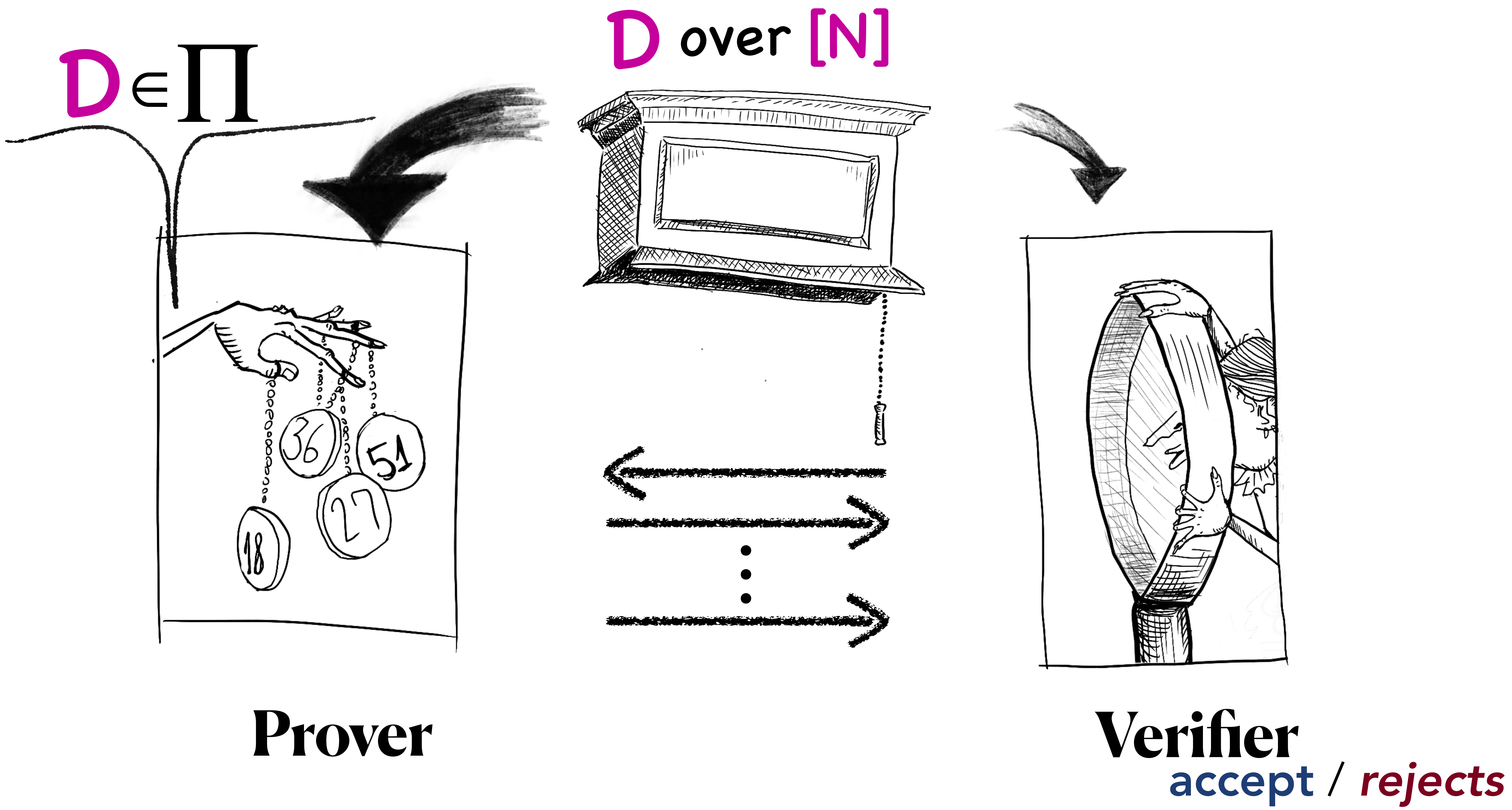


Prover

Proof Systems for Distributions [GMR'85, CG'17]



Proof Systems for Distributions [GMR'85, CG'17]



Completeness

$\mathsf{D} \in \Pi$

$\Longrightarrow$ V ***accepts*** whp.

Soundness

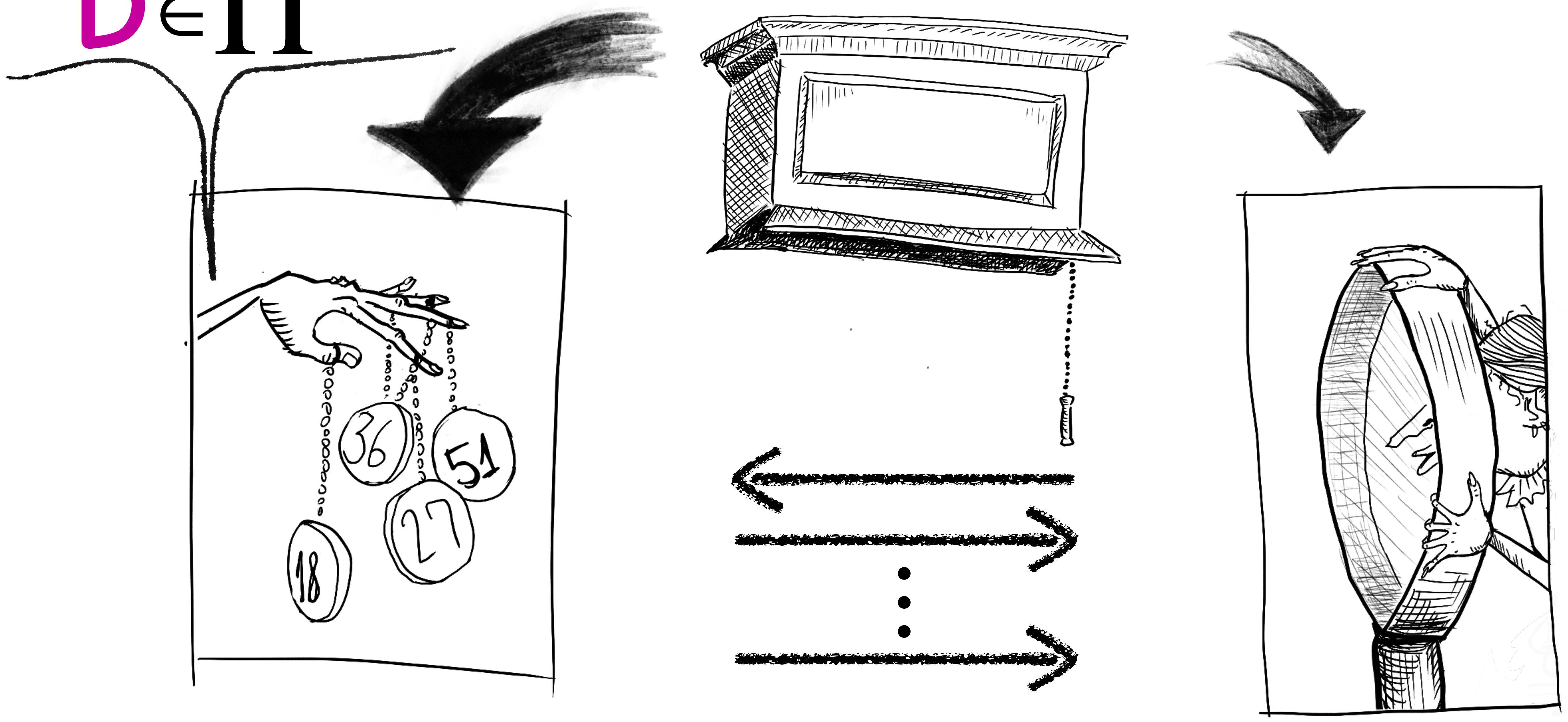
$\delta_{TV}(\mathsf{D}, \Pi) \geq \varepsilon$

$\Longrightarrow$ $\forall$(cheating) strategy P^*,
V ***rejects*** whp.

Proof Systems for Distributions [GMR'85, CG'17]

D over [N]

$D \in \Pi$



Prover

Verifier
accept / *rejects*

Completeness

$D \in \Pi$

$\Longrightarrow$ V ***accepts*** whp.

Soundness

$\delta_{TV}(D, \Pi) \geq \varepsilon$

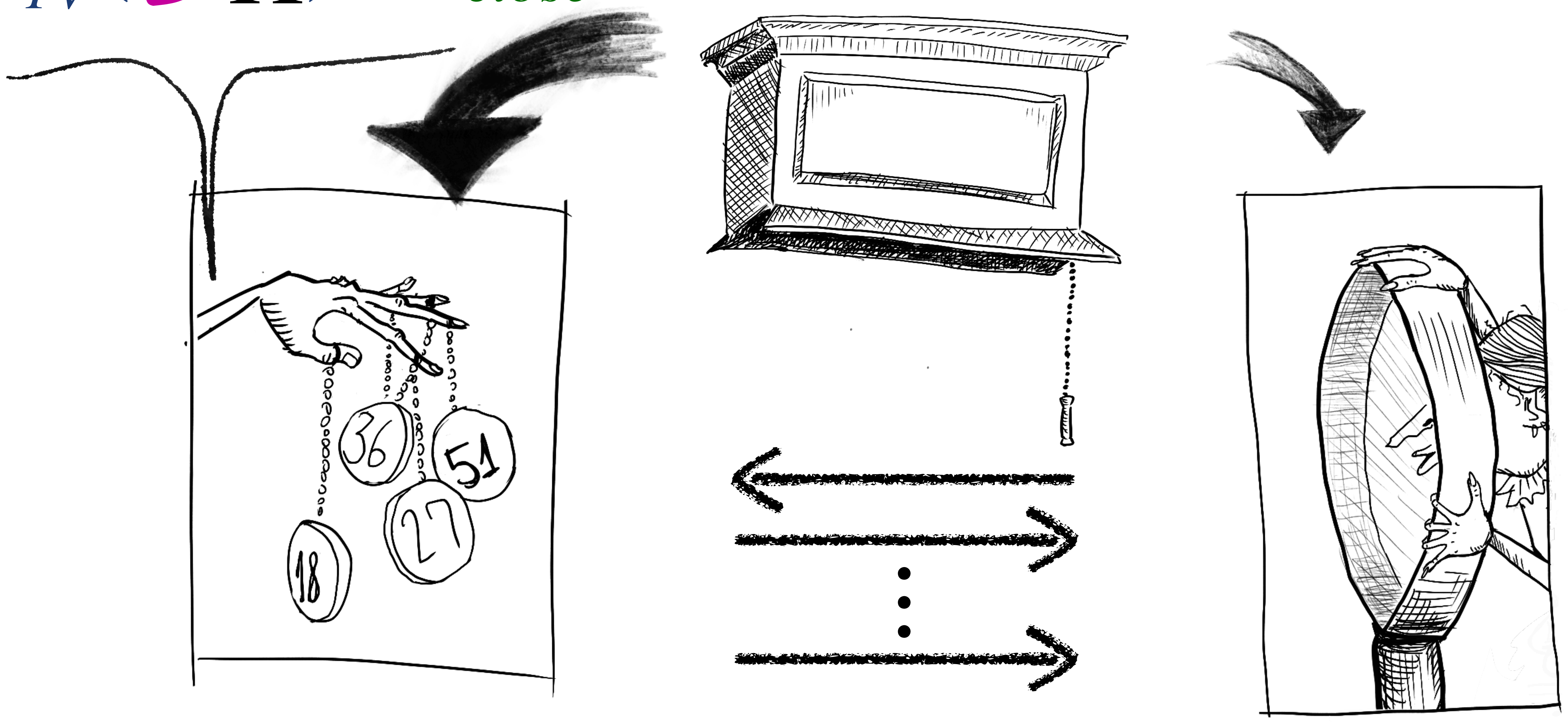
$\Longrightarrow$ $\forall$(cheating) strategy P^*,
V ***rejects*** whp.

P's Samples, runtime:	***Communication:***	***V's Samples, runtime:***	***Quantities of interest***
poly(N)	$o(N)$, *small # rounds*	o(*optimal* Π *tester*)	

(Tolerant) Proof Systems for Distributions [GMR'85, CG'17]

$\delta_{TV}(\mathbf{D},\Pi) \leq \varepsilon_{close}$

D over [N]



Prover

Verifier
accept / *rejects*

Completeness

$\delta_{TV}(\mathbf{D},\Pi) \leq \varepsilon_{close}$

$\Longrightarrow$ V ***accepts*** whp.

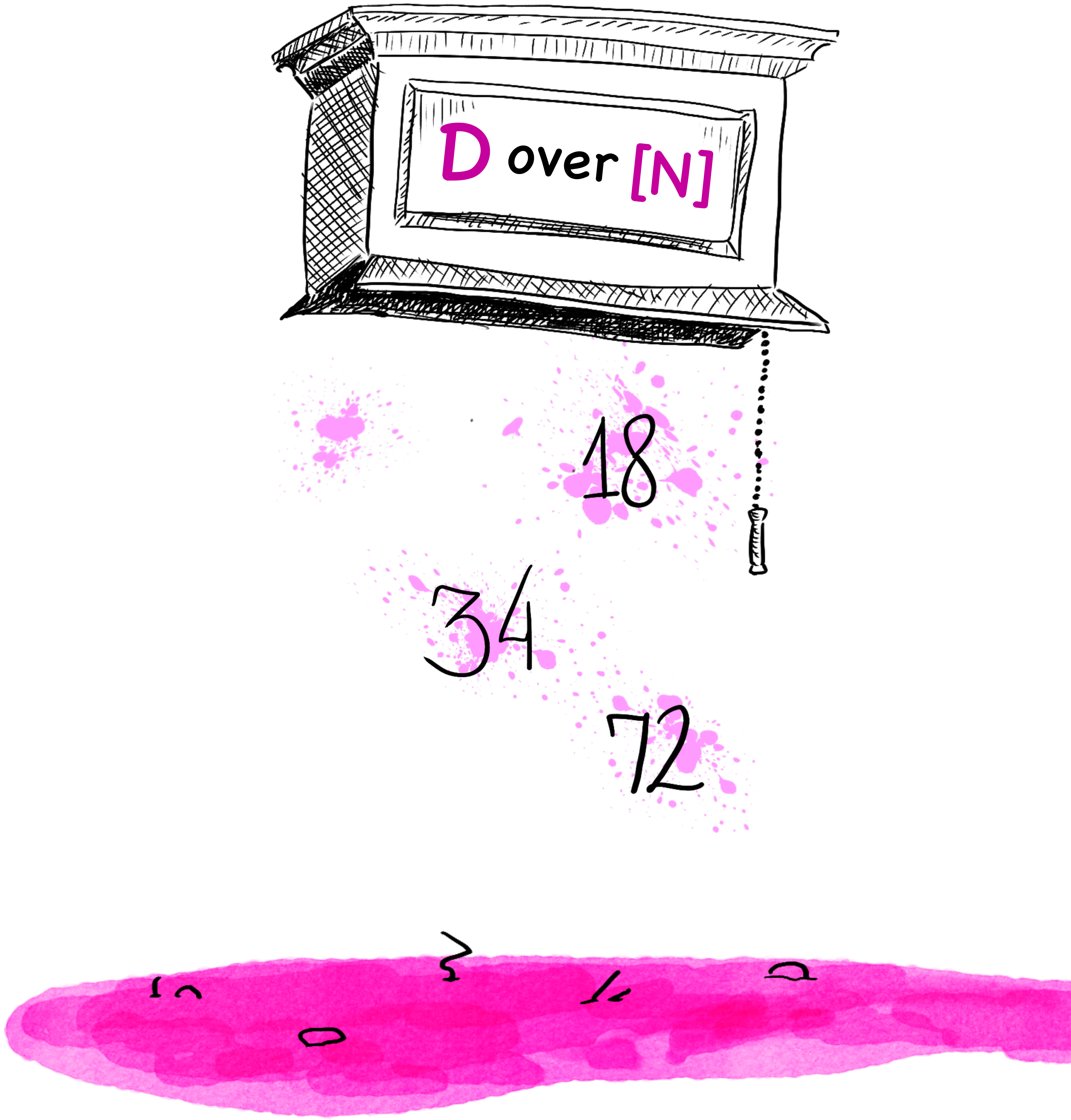
Soundness

$\delta_{TV}(\mathbf{D},\Pi) \geq \varepsilon_{far}$

$\Longrightarrow$ ∀(cheating) strategy P^*,
V ***rejects*** whp.

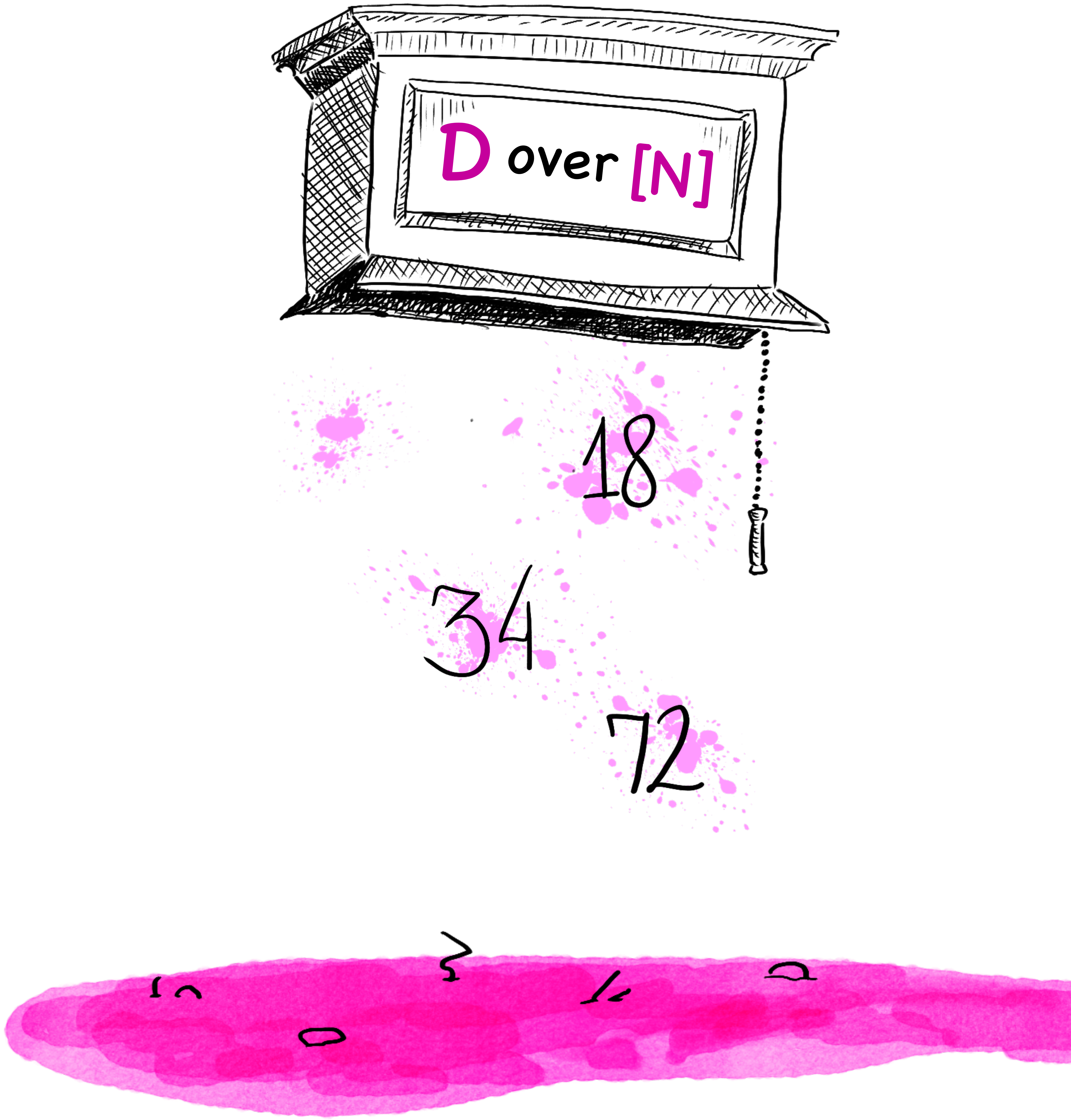
P's Samples, runtime: poly(N)	***Communication:*** $o(N)$, *small # rounds*	***V's Samples, runtime:*** o(optimal Π tester)	***Quantities of interest***

Example: Verifying Uniformity Over Half Domain



Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over* ***some half*** *of their domain.*



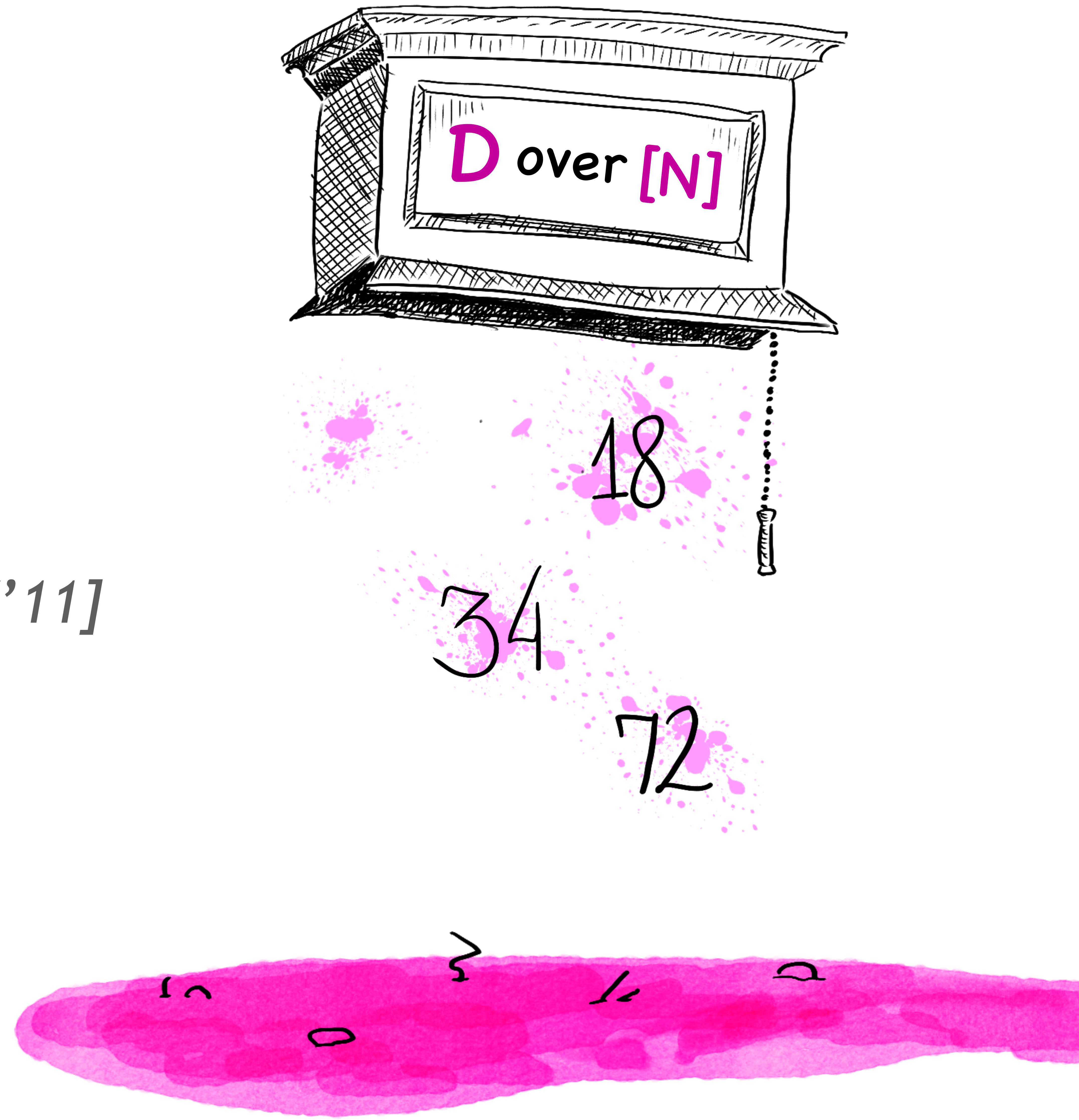
Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over* ***some half*** *of their domain.*

Testing the property:

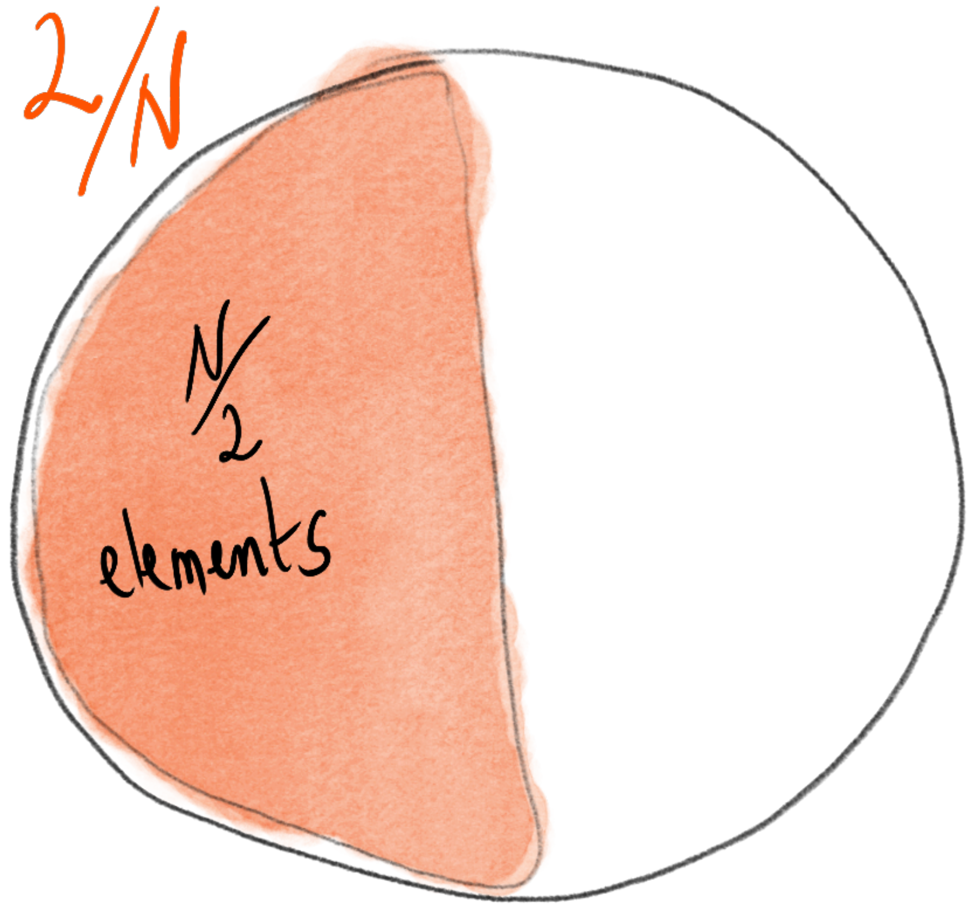
- $\Theta\left(N^{2/3}\right)$ *samples.* *[BC'17]*

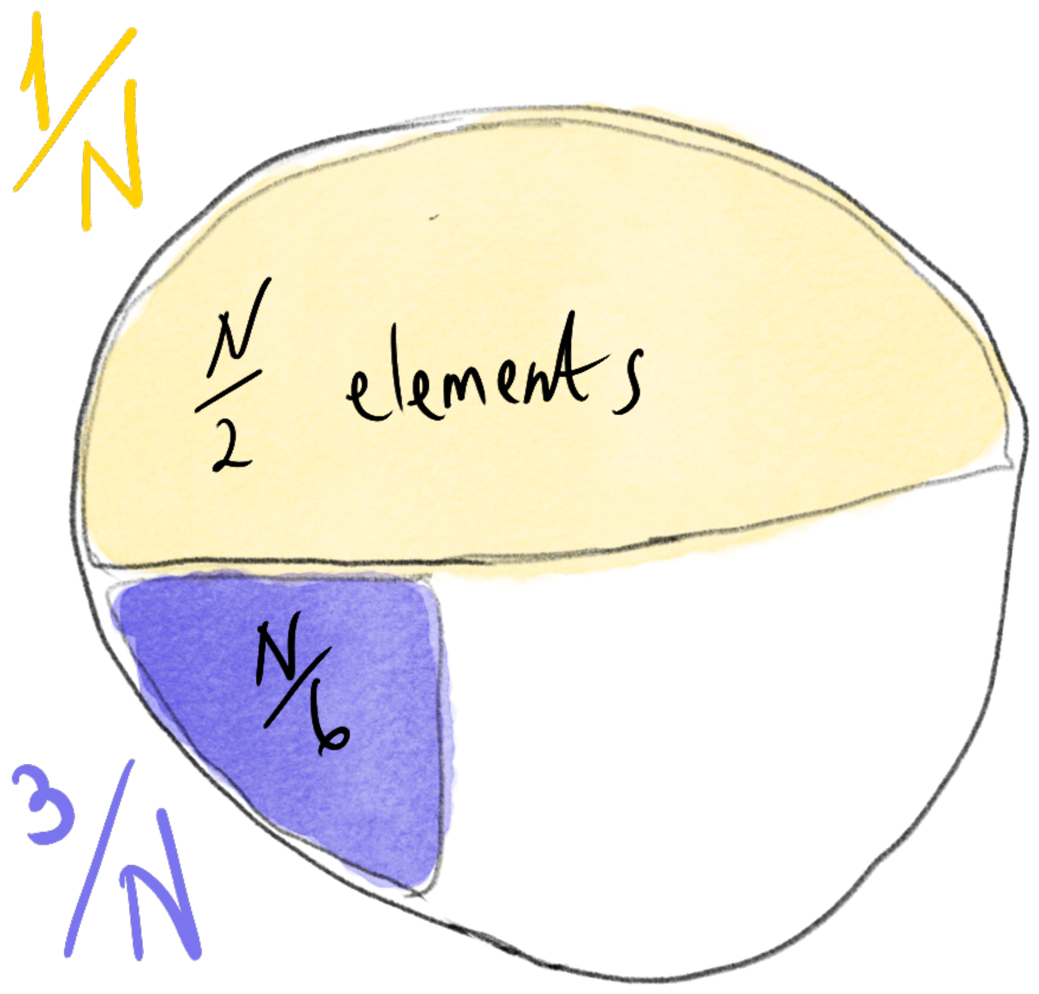
- *(Tolerantly)* $\Theta(N/\log N)$ *samples.* *[VV'11]*

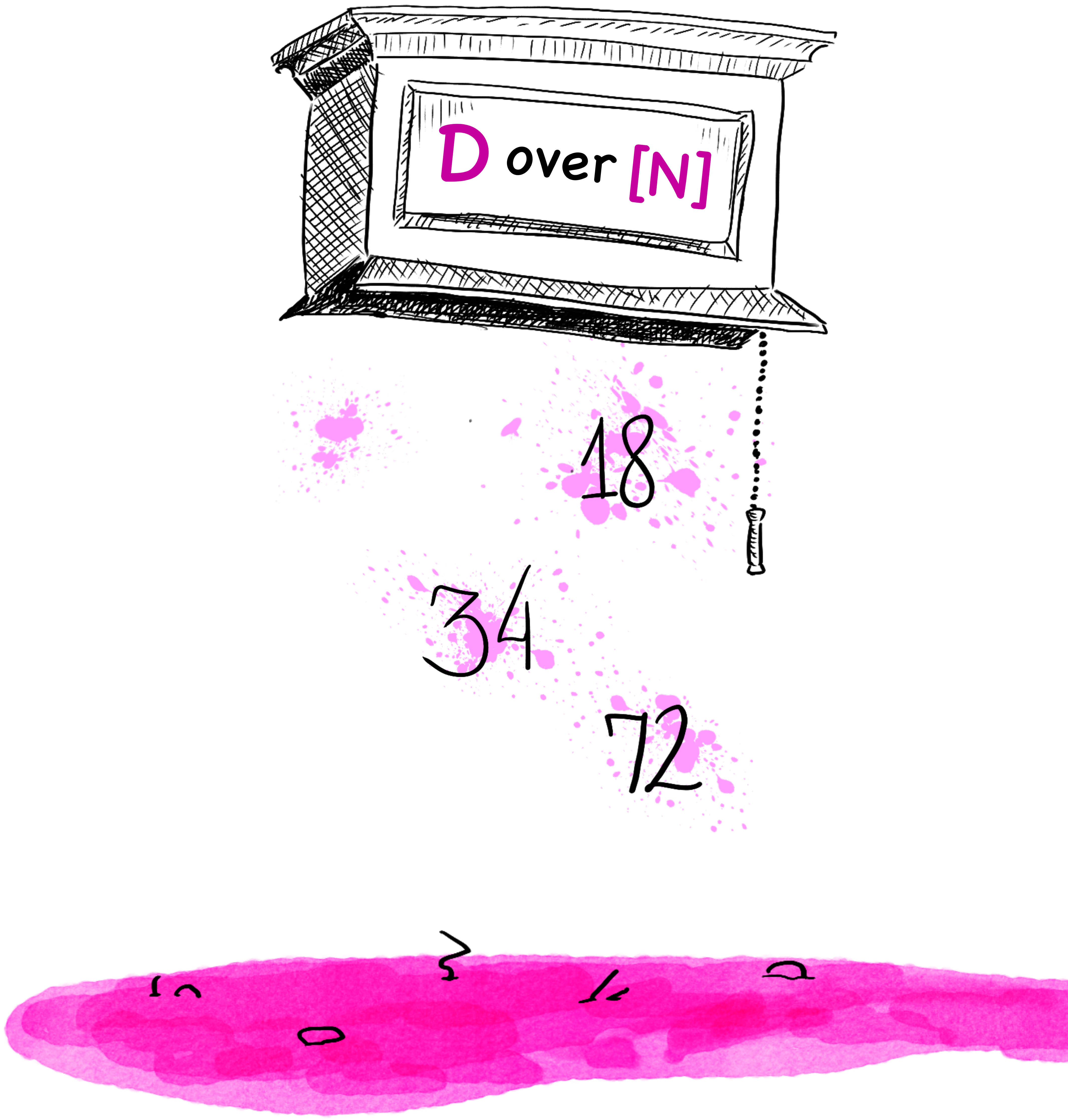


Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over* ***some half*** *of their domain.*

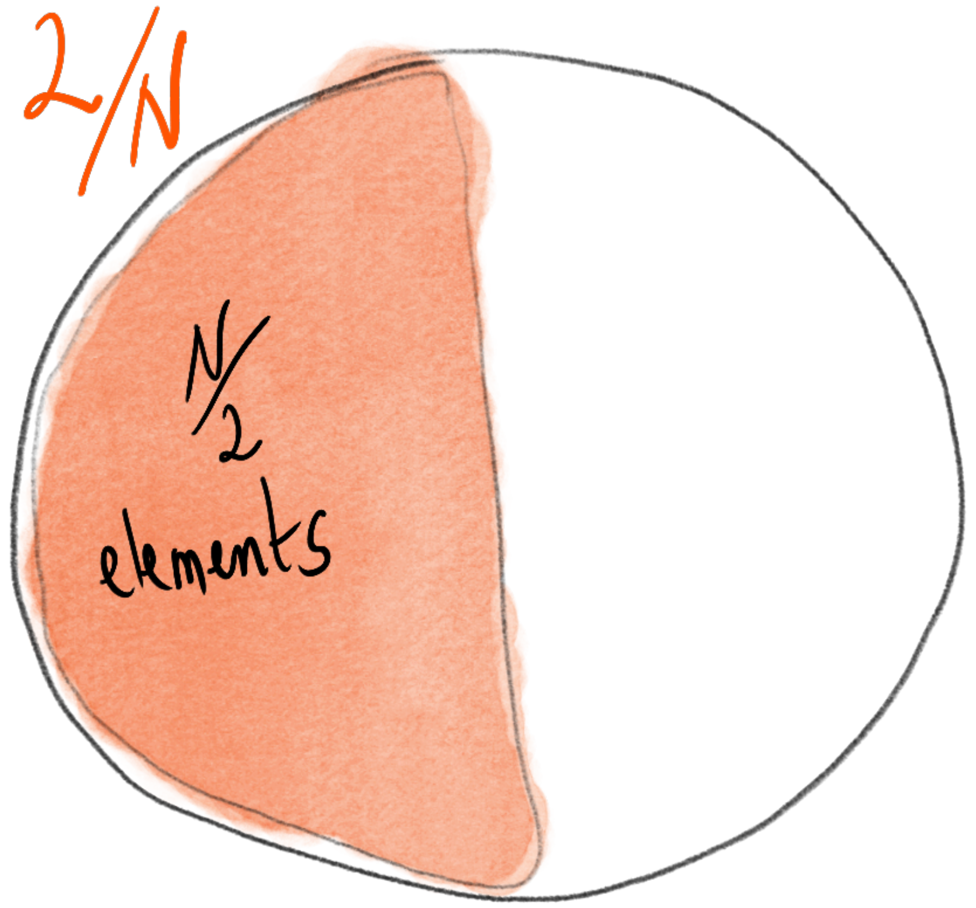


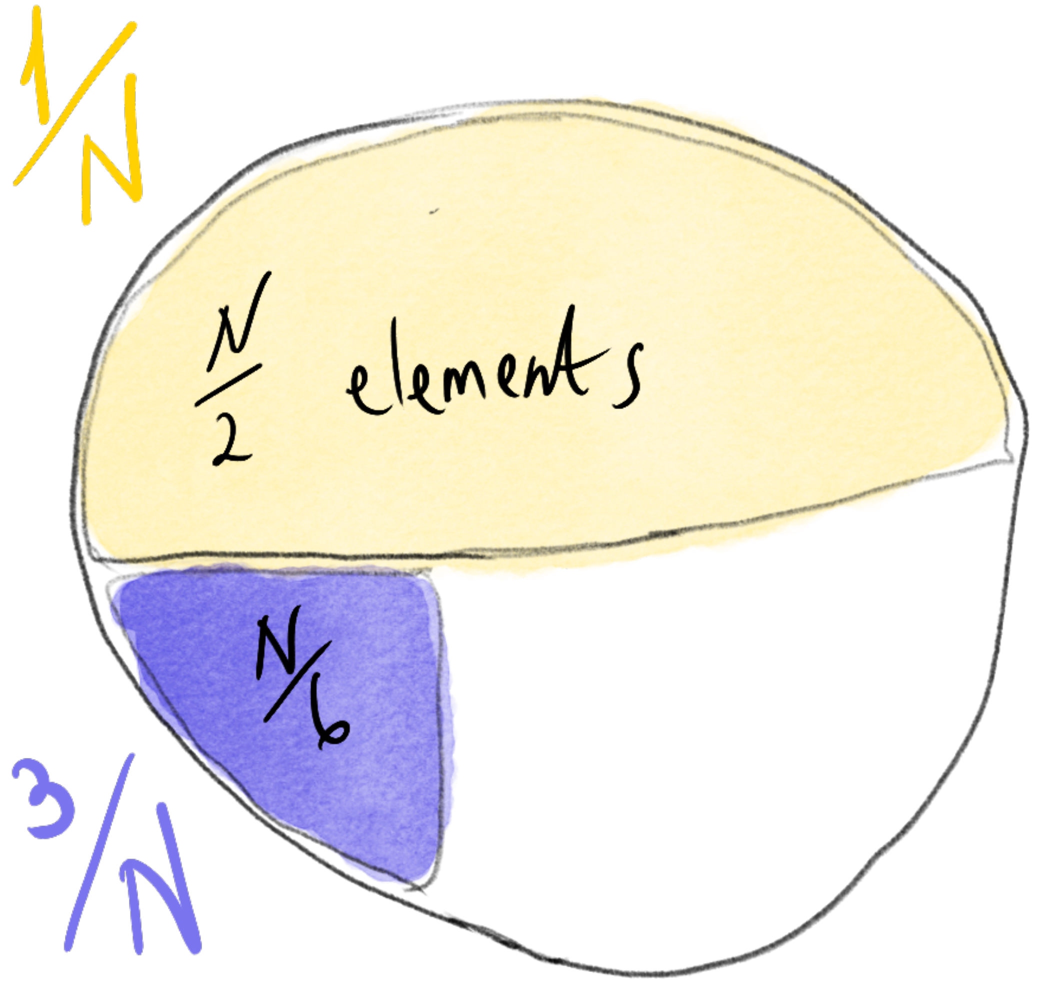




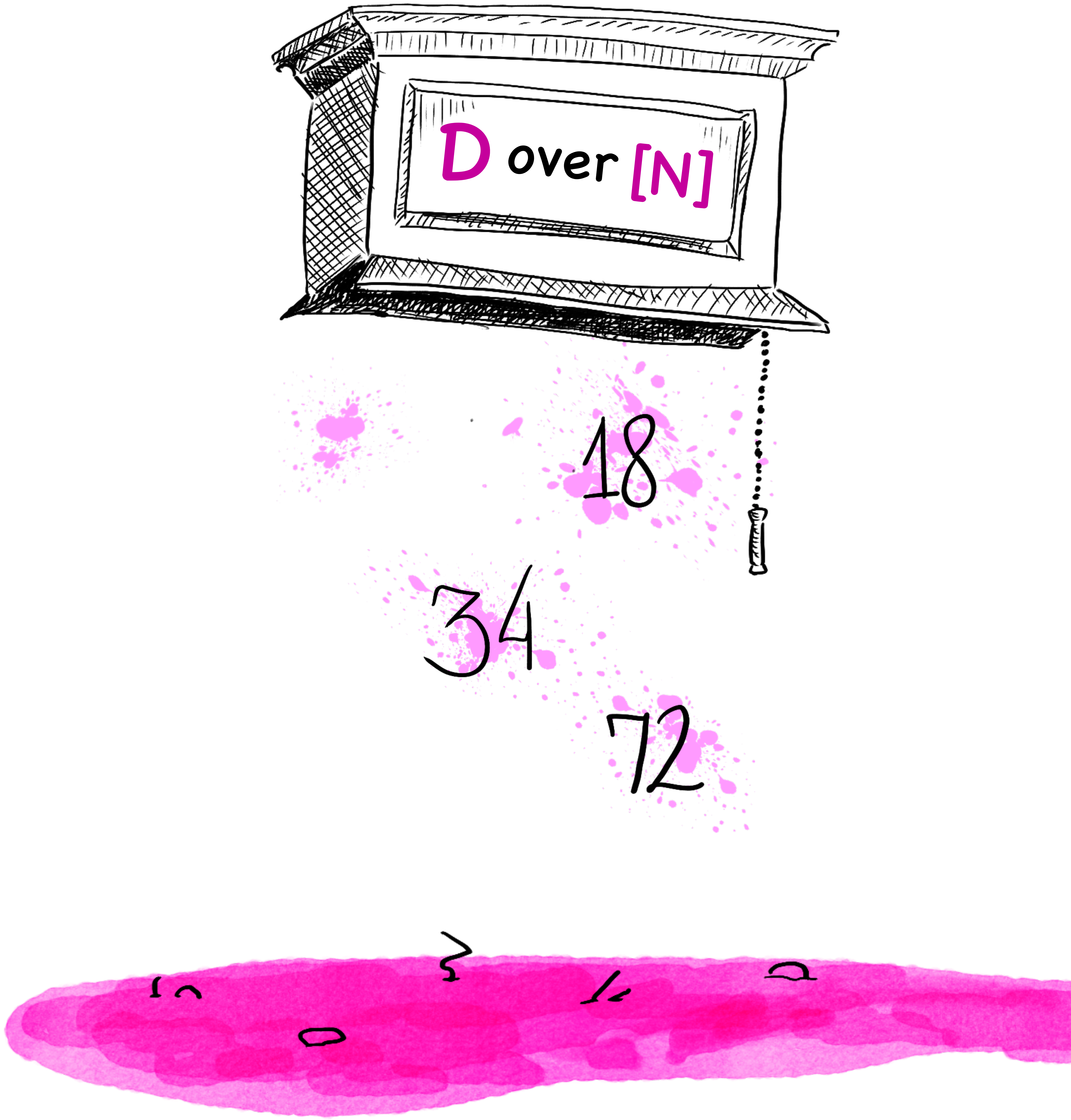
Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over* ***some half*** *of their domain.*



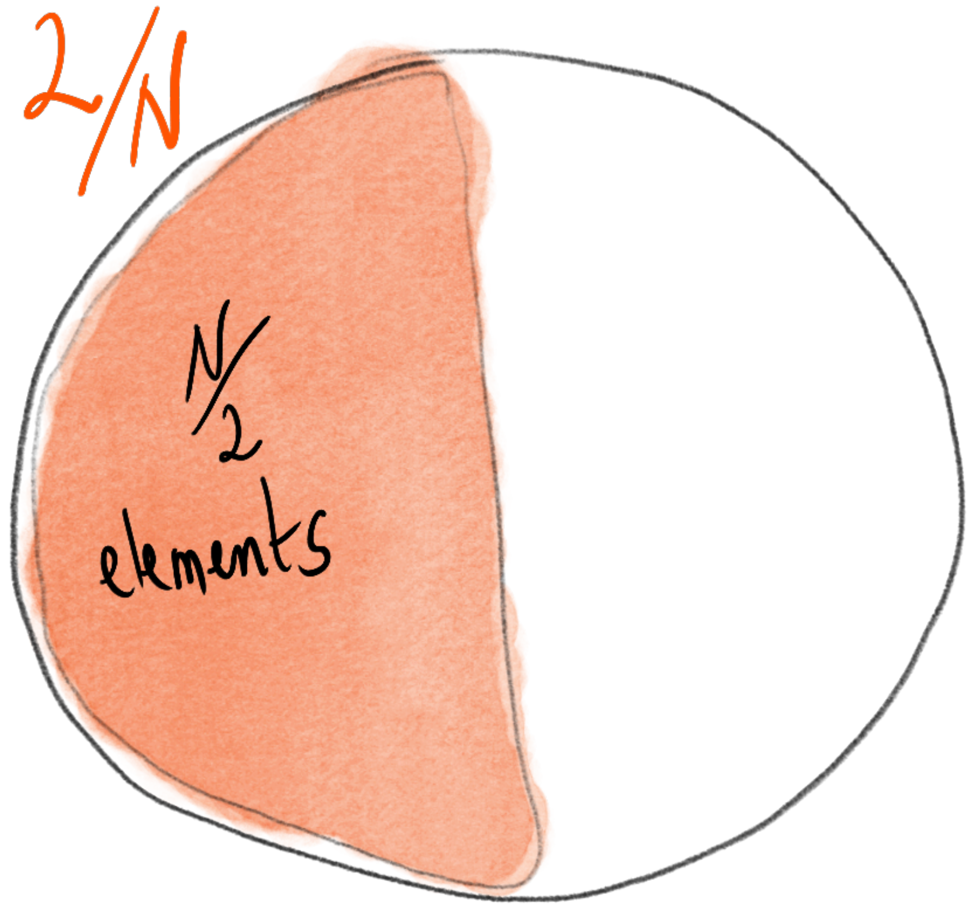


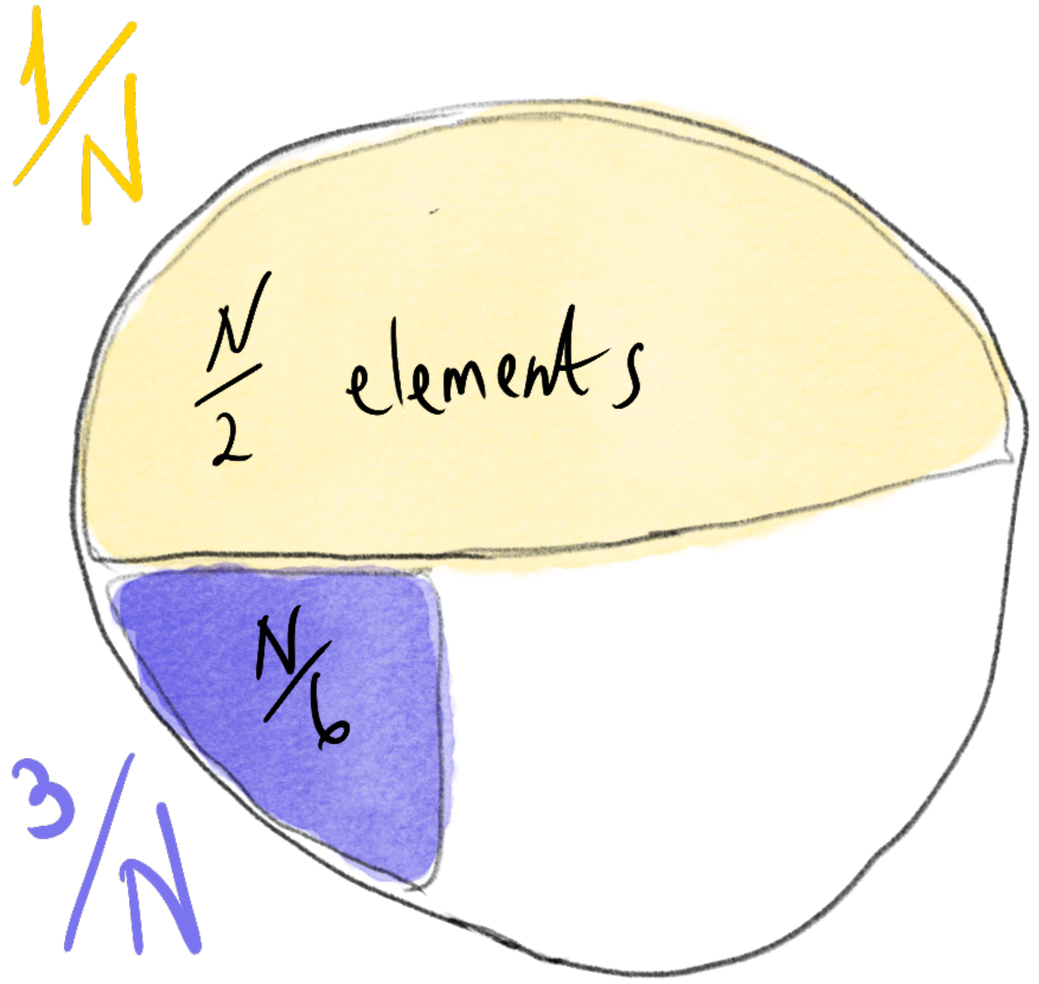
Probability of seeing a 2-collision : $\sum_x D(x) \cdot D(x)$

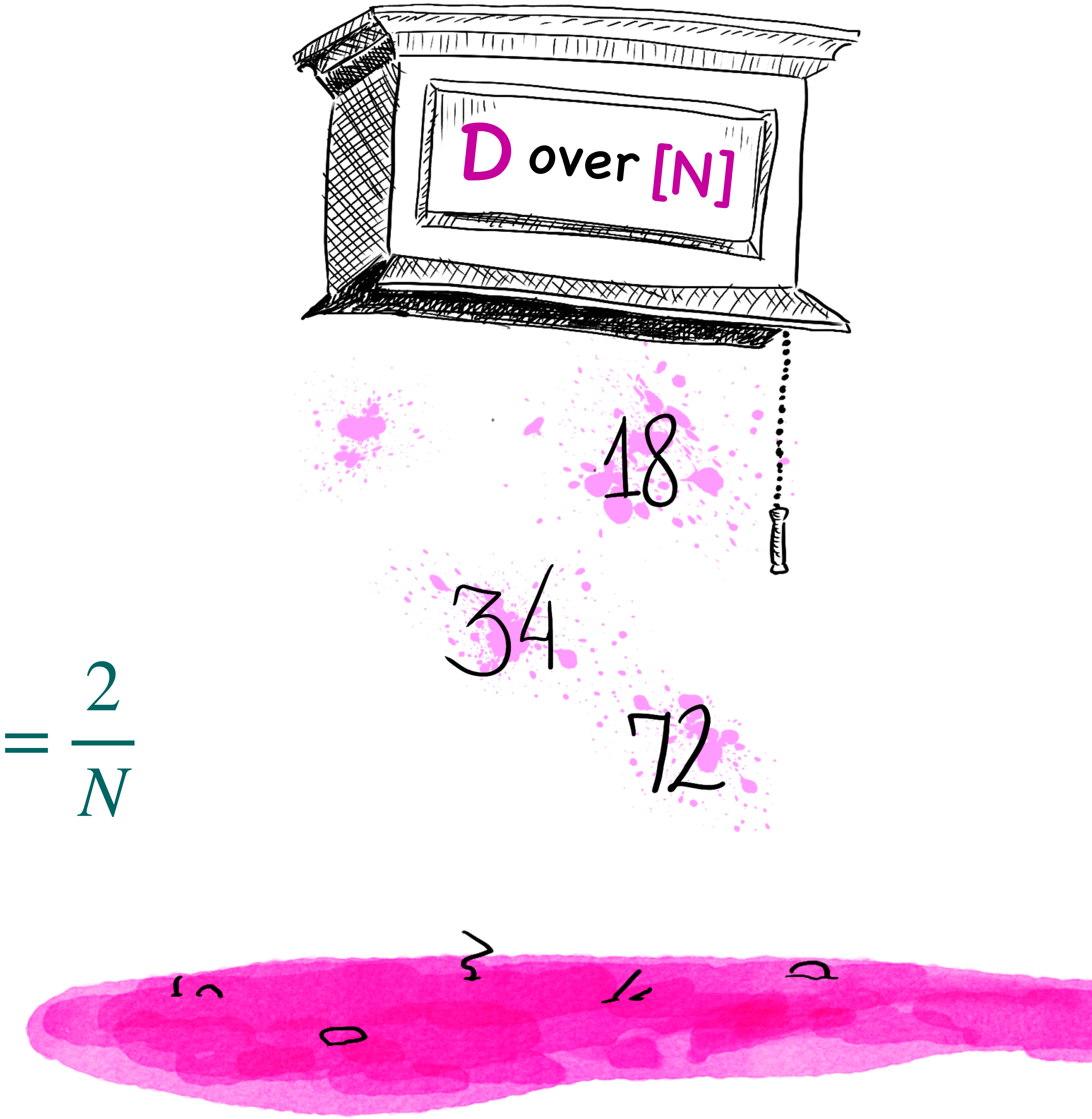


Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over **some half** of their domain.*



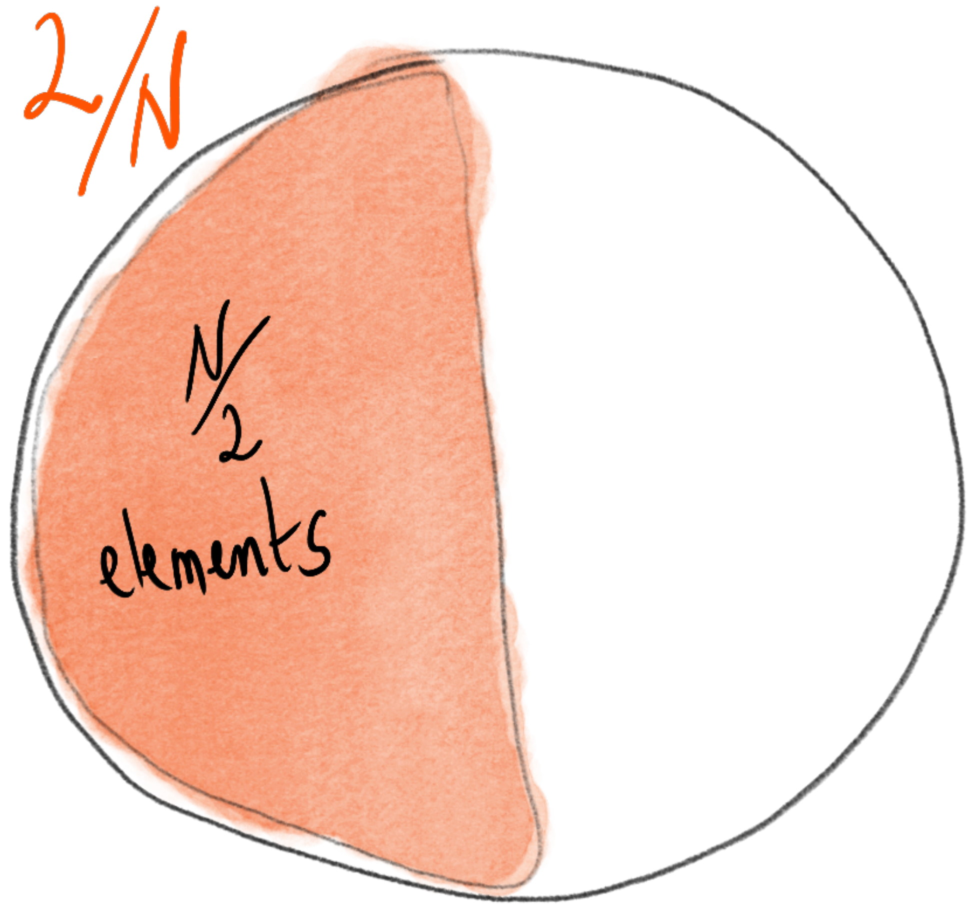


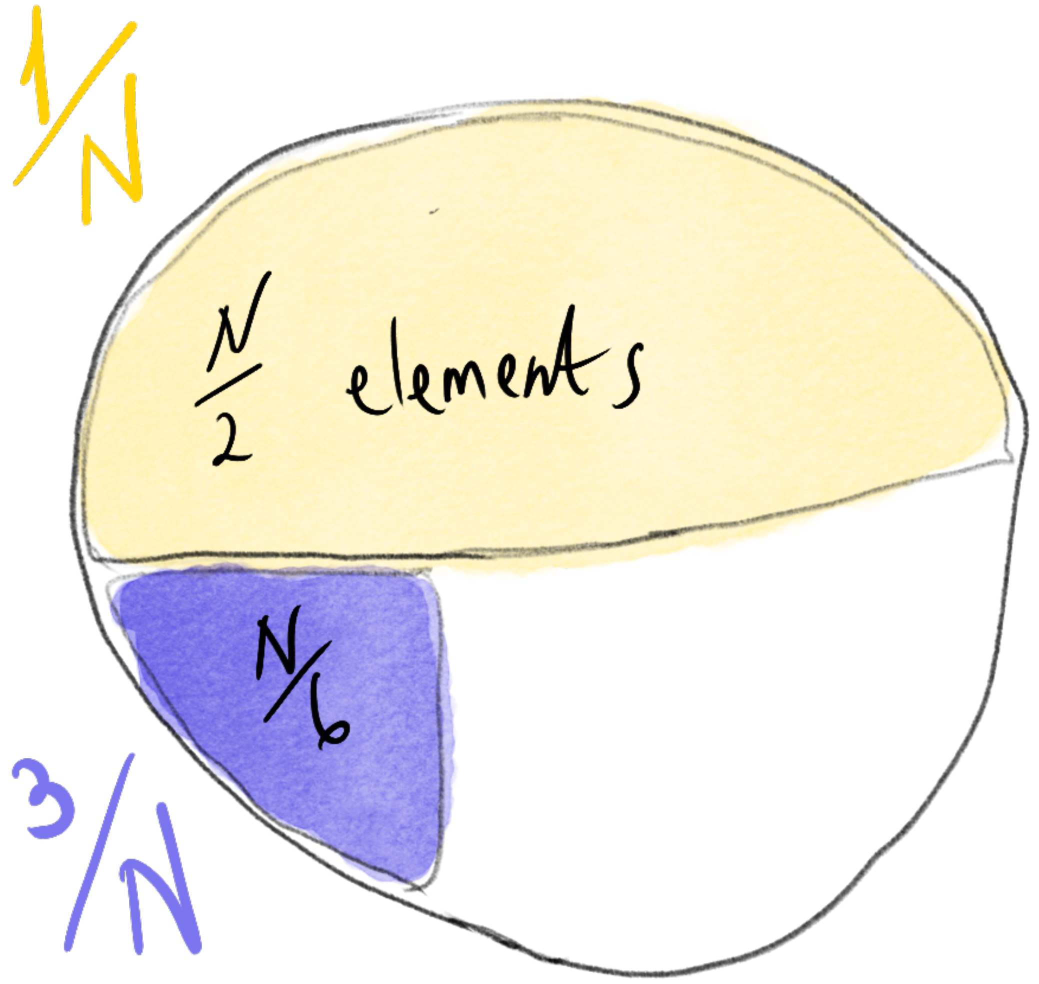


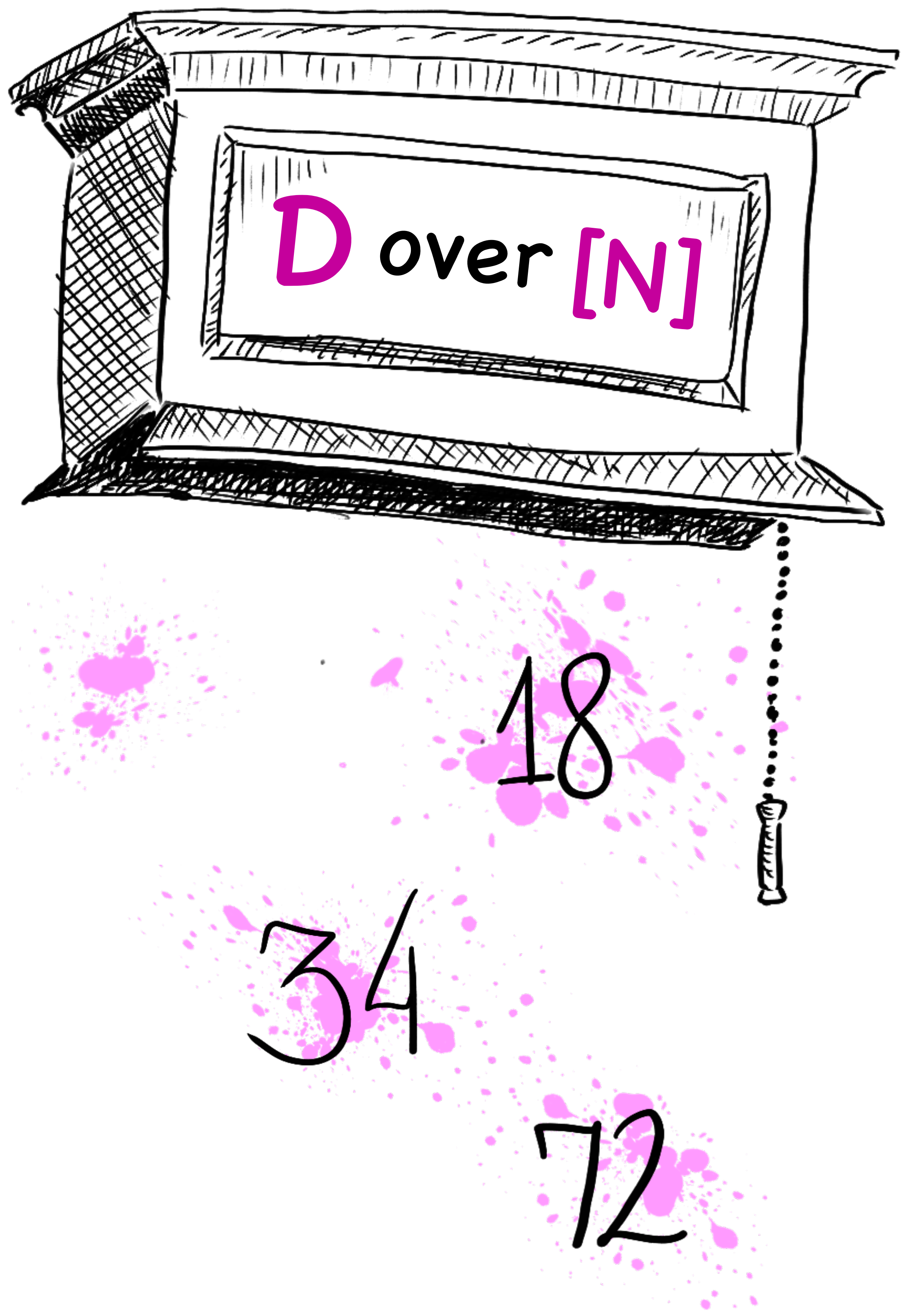
Probability of seeing a 2-collision : $\sum_x D(x) \cdot D(x) = \frac{2}{N}$

Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over **some half** of their domain.*

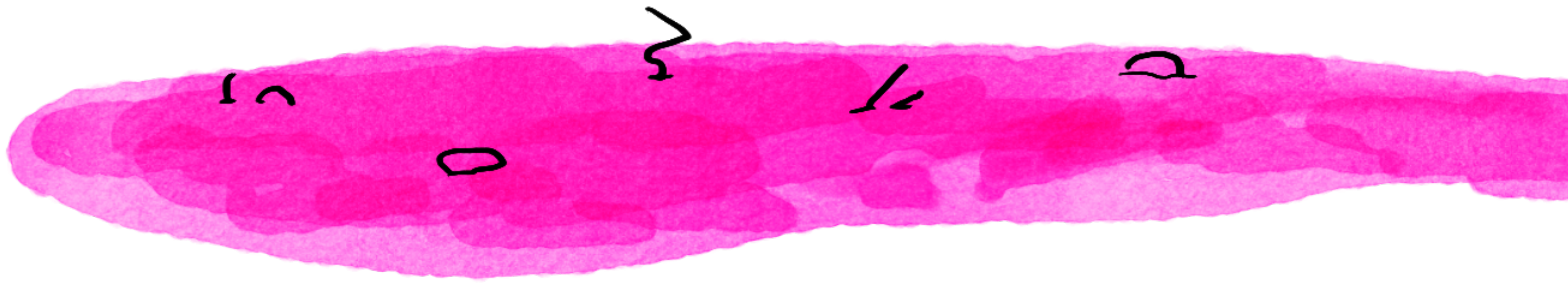






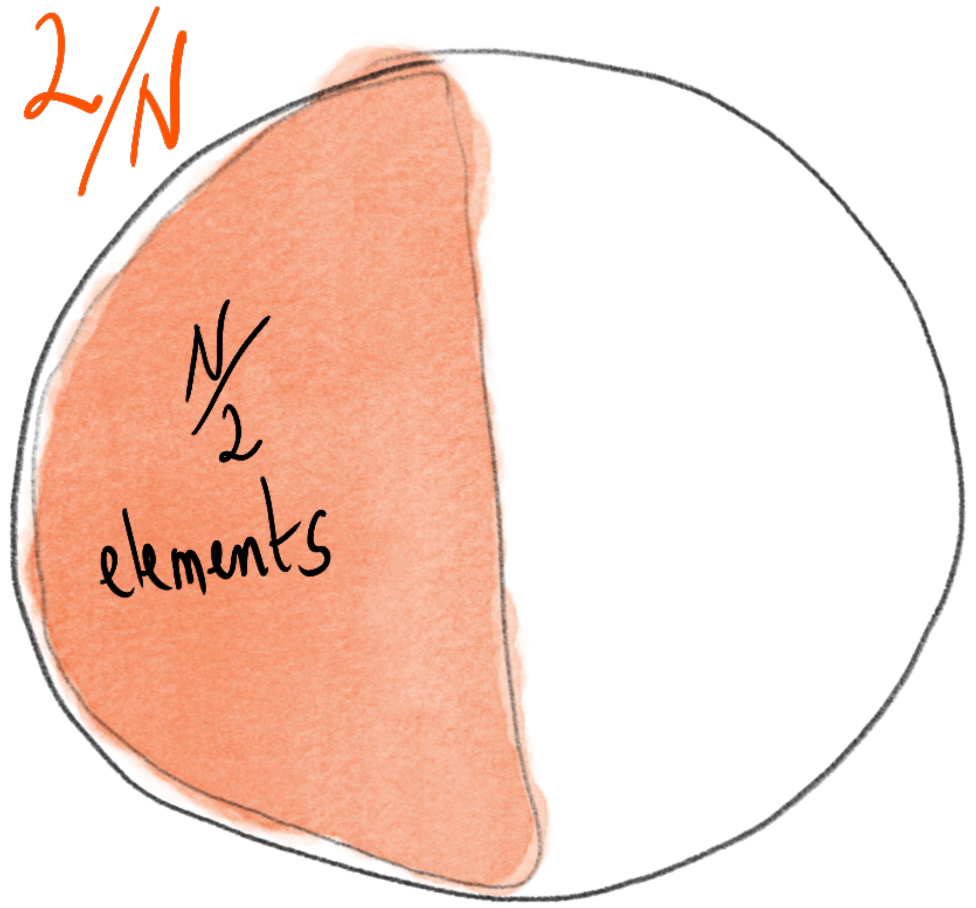
Probability of seeing a 2-collision : $\sum_x D(x) \cdot D(x) = \frac{2}{N}$

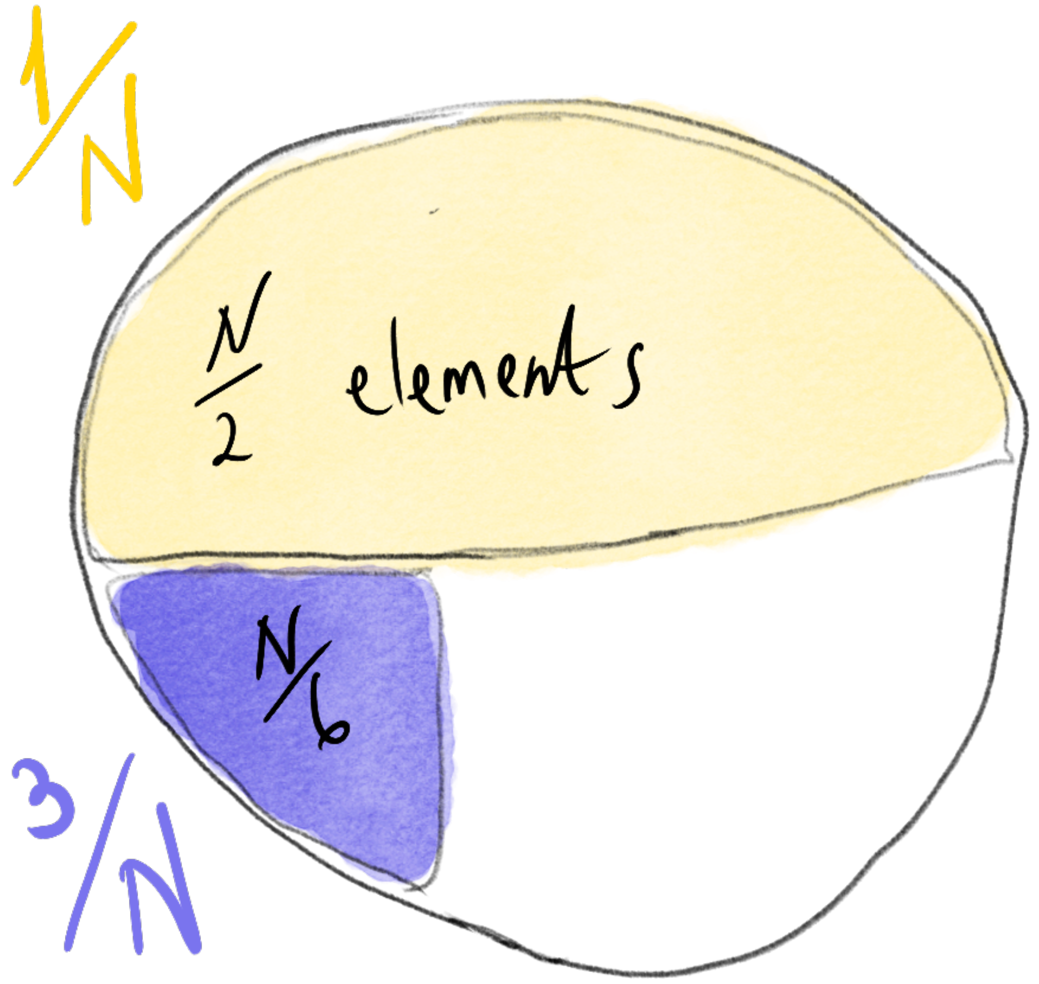
Can't test with $o\left(N^{2/3}\right)$

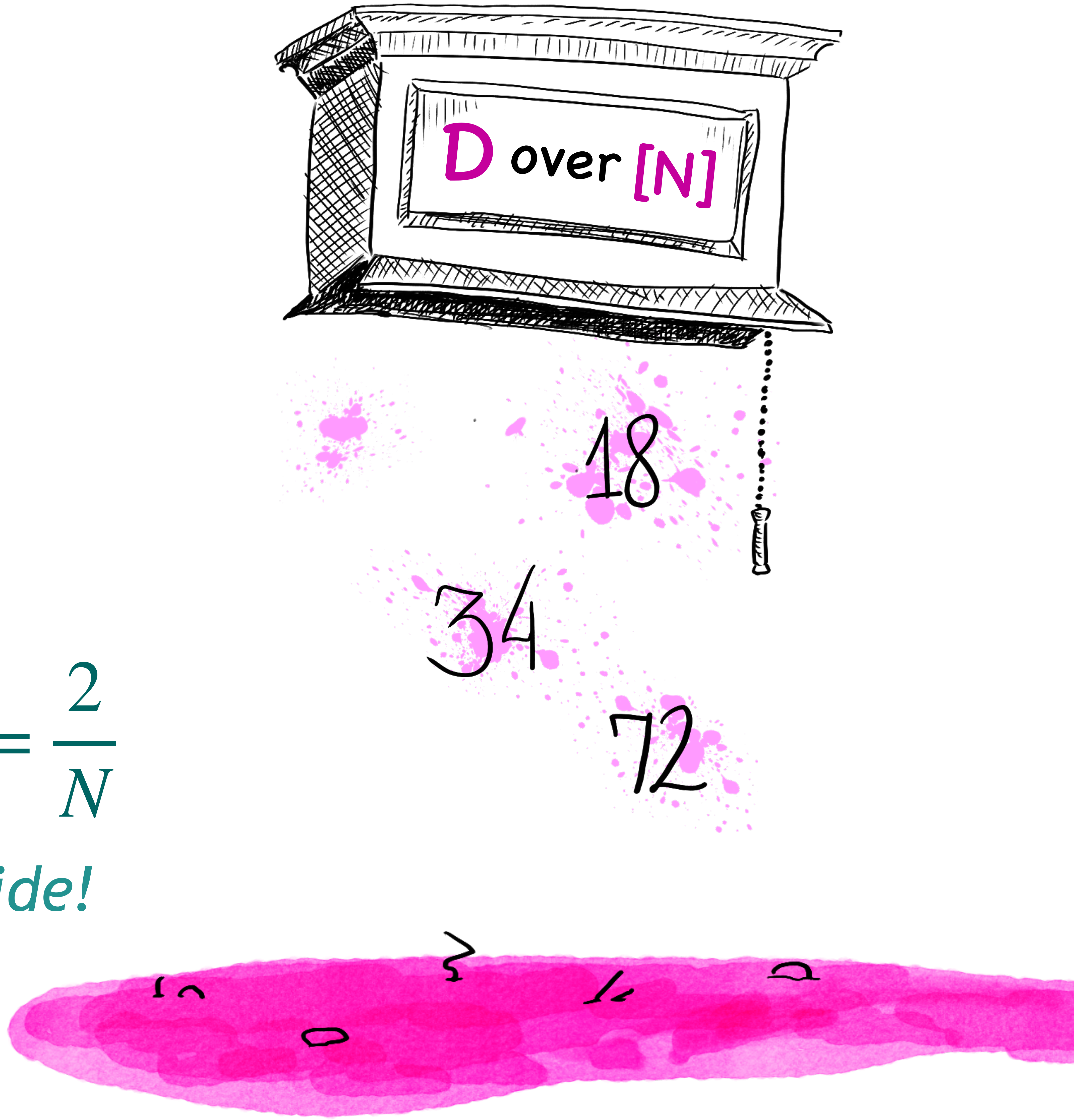


Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over* ***some half*** *of their domain.*





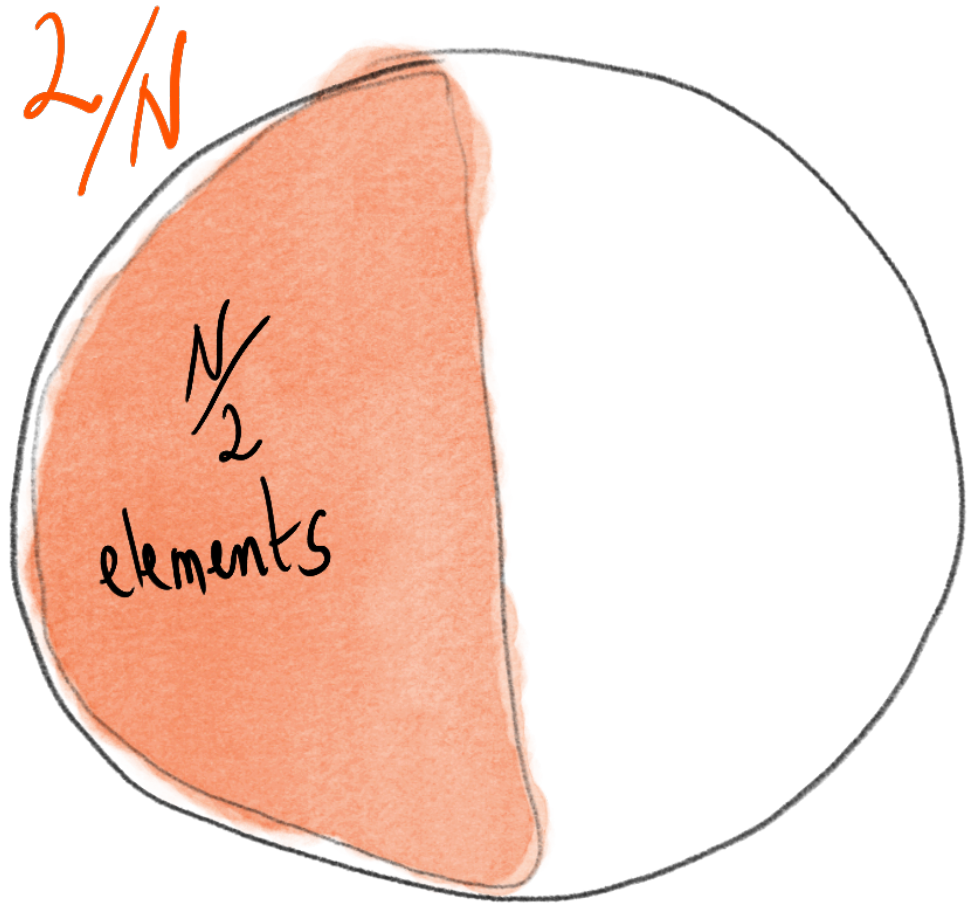


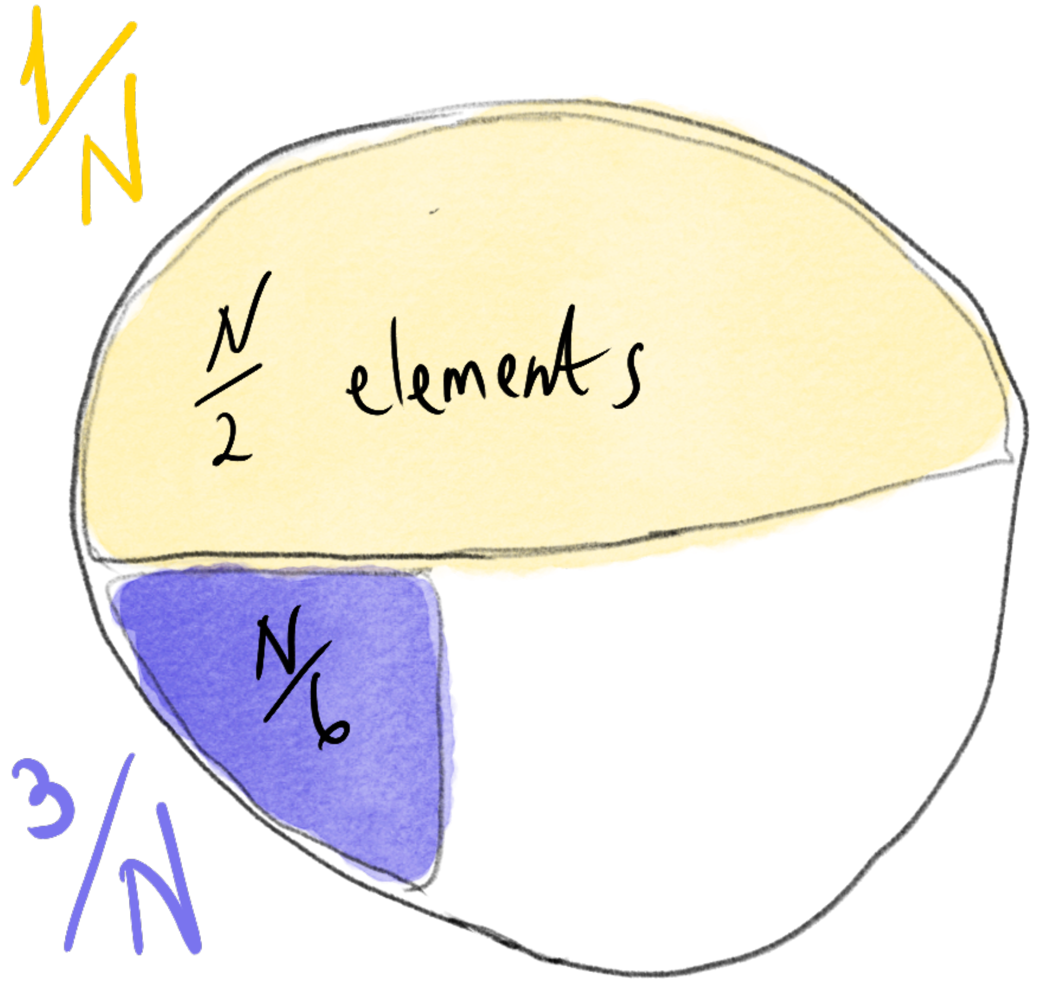
Probability of seeing a *2-collision* : $\sum_x D(x) \cdot D(x) = \frac{2}{N}$

With $O\left(N^{2/3}\right)$ samples exhibit 3-collisions & decide!

Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over **some half** of their domain.*





Probability of seeing a *2-collision* : $$\sum_x D(x) \cdot D(x) = \frac{2}{N}$$

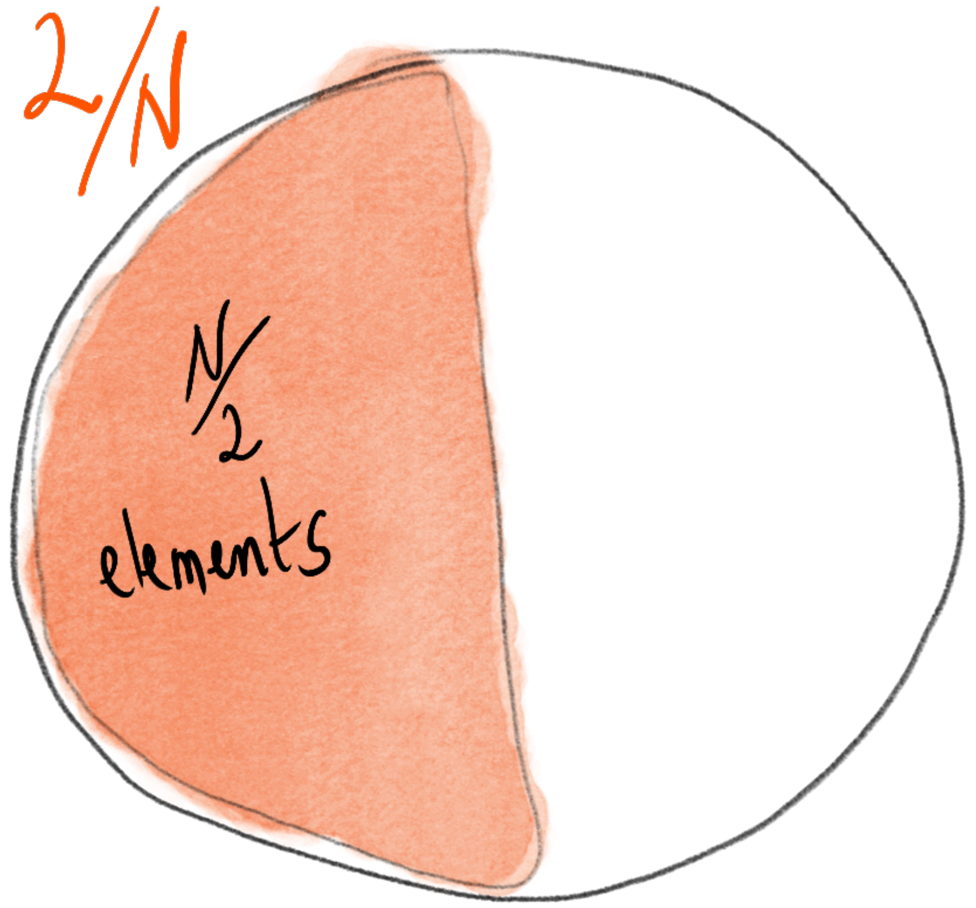
With $O\left(N^{2/3}\right)$ samples exhibit 3-collisions & decide!

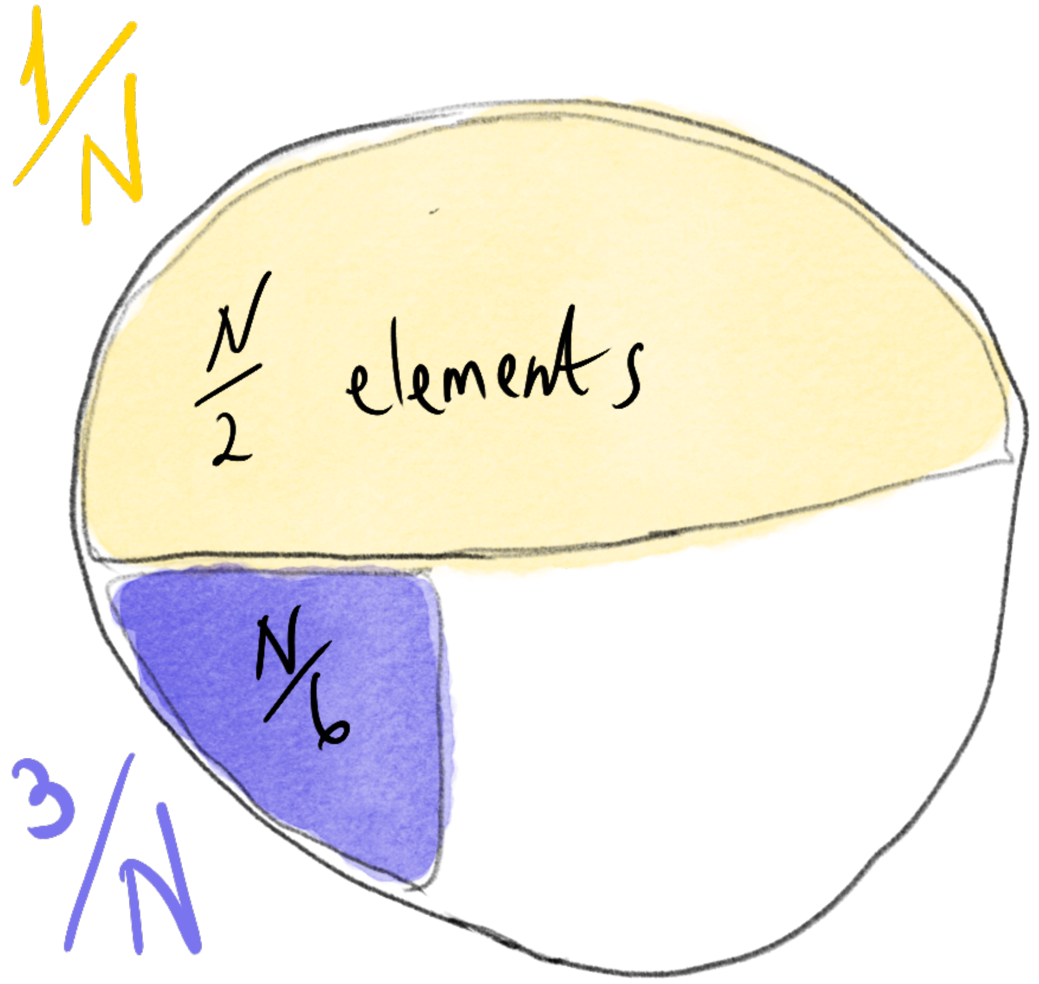
Observation:

Support sizes are different.

Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over* ***some half*** *of their domain.*





Probability of seeing a *2-collision* : $\sum_x D(x) \cdot D(x) = \frac{2}{N}$

With $O\left(N^{2/3}\right)$ samples exhibit 3-collisions & decide!

Observation:

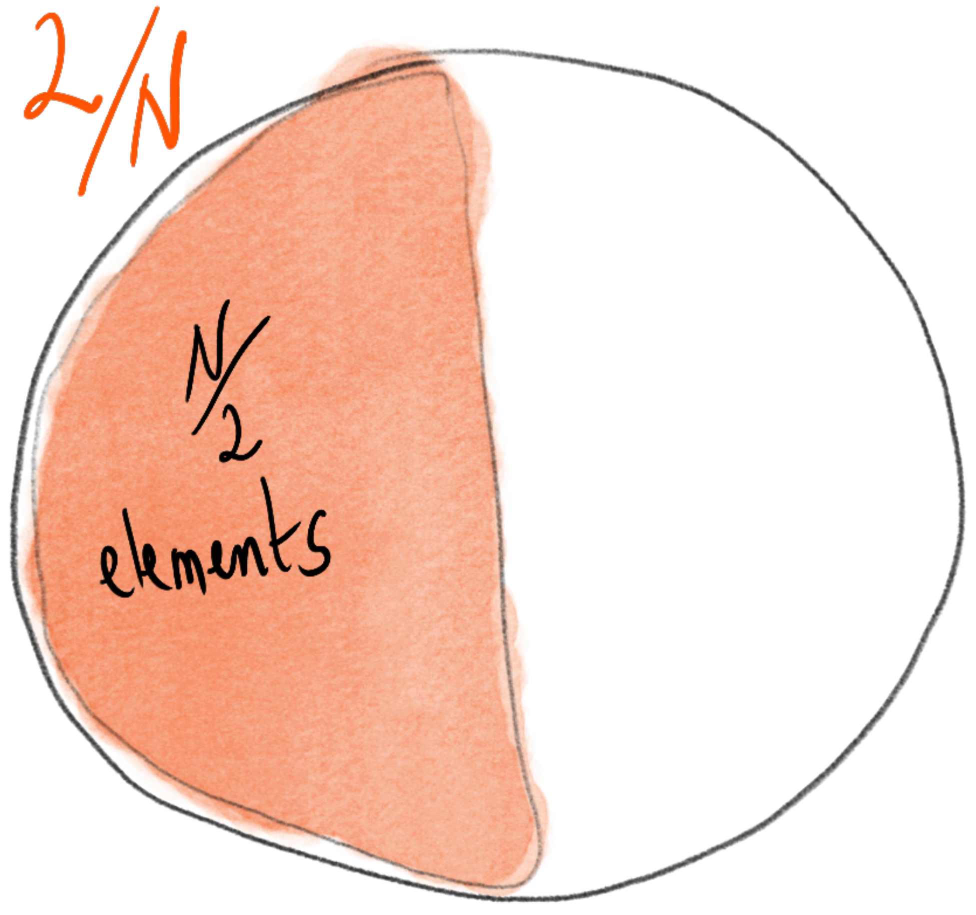
Support sizes are different.

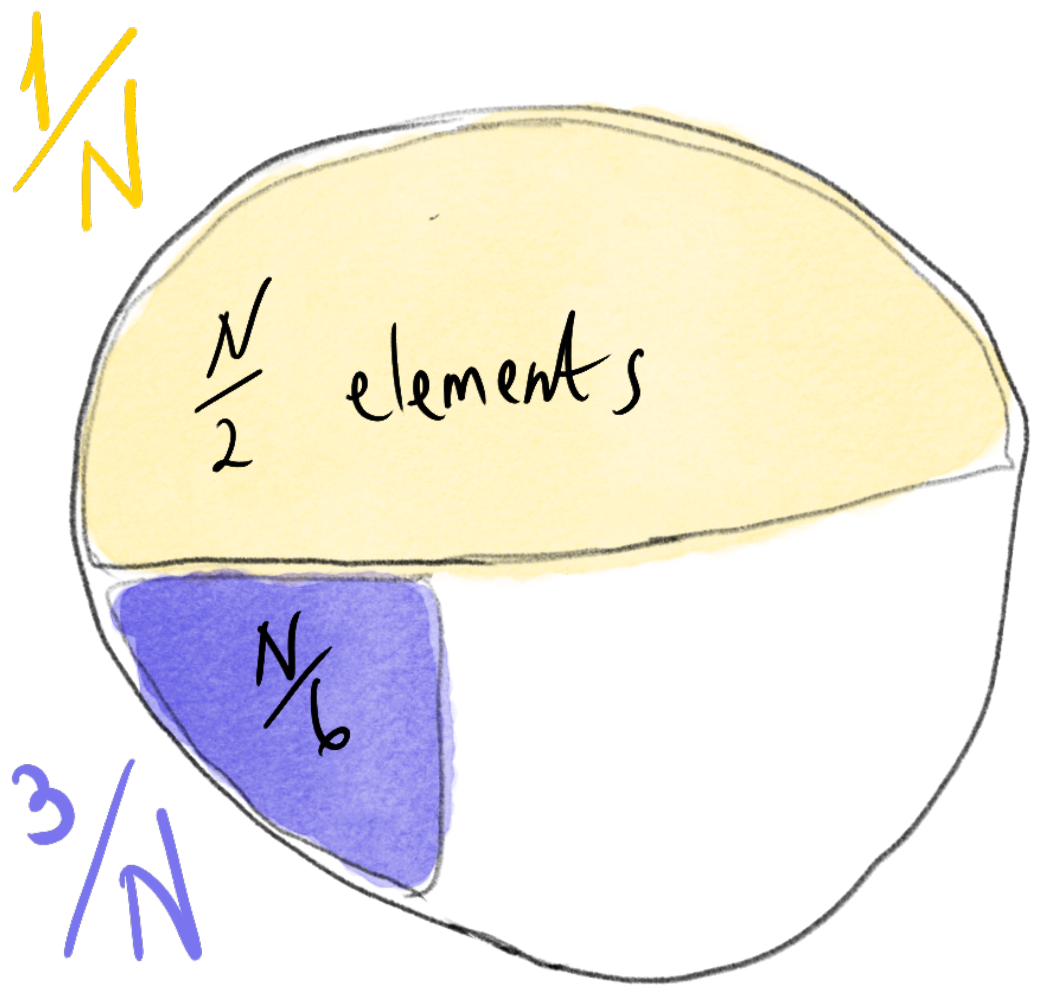
$\sum_x D^2(x) = \frac{2}{N}$ AND $\delta_{TV}\left(D, U_{\text{half}}\right) \geq \varepsilon$

$|\text{Supp}(D)| \geq \frac{N}{2}\left(1 + \Theta\left(\varepsilon^2\right)\right)$

Example: Verifying Uniformity Over Half Domain

U_{half} *All distributions uniform over **some half** of their domain.*





Probability of seeing a *2-collision* : $\sum_x D(x) \cdot D(x) = \frac{2}{N}$

With $O\left(N^{2/3}\right)$ samples exhibit 3-collisions & decide!

Observation:

Support sizes are different.

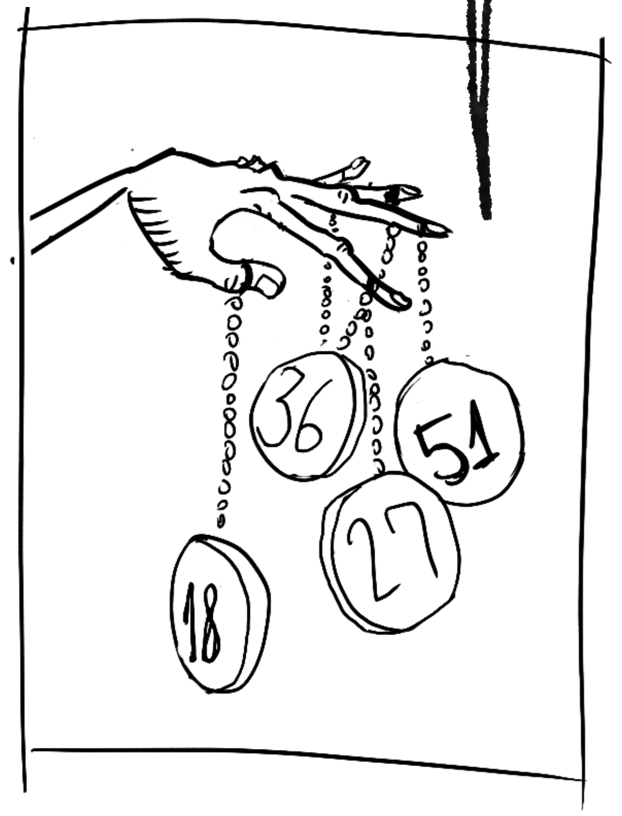
Every NO-case either has:

- 2-collision prob. $\neq 2/N$;
- or, support larger than.

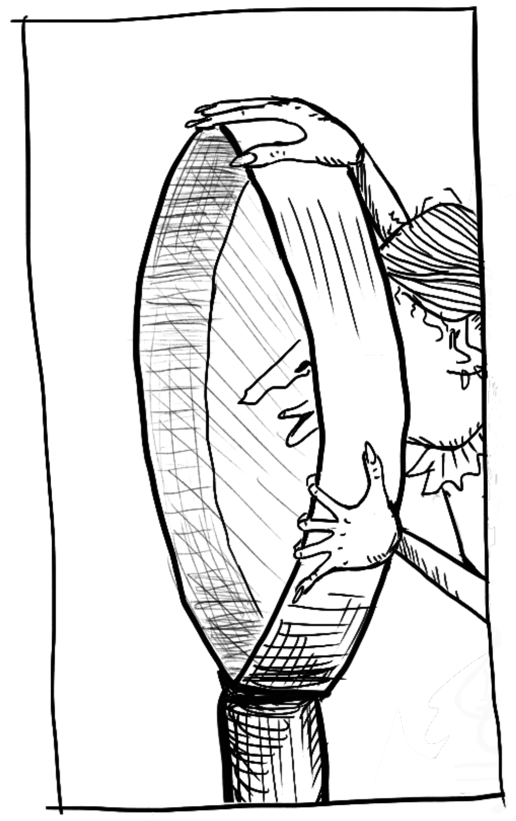
$$\frac{N}{2}\left(1+\Theta\left(\varepsilon^2\right)\right)$$

Verifying Uniformity Over Half Domain

$\mathsf{D} \in \mathbb{U}_{N/2}$



Prover



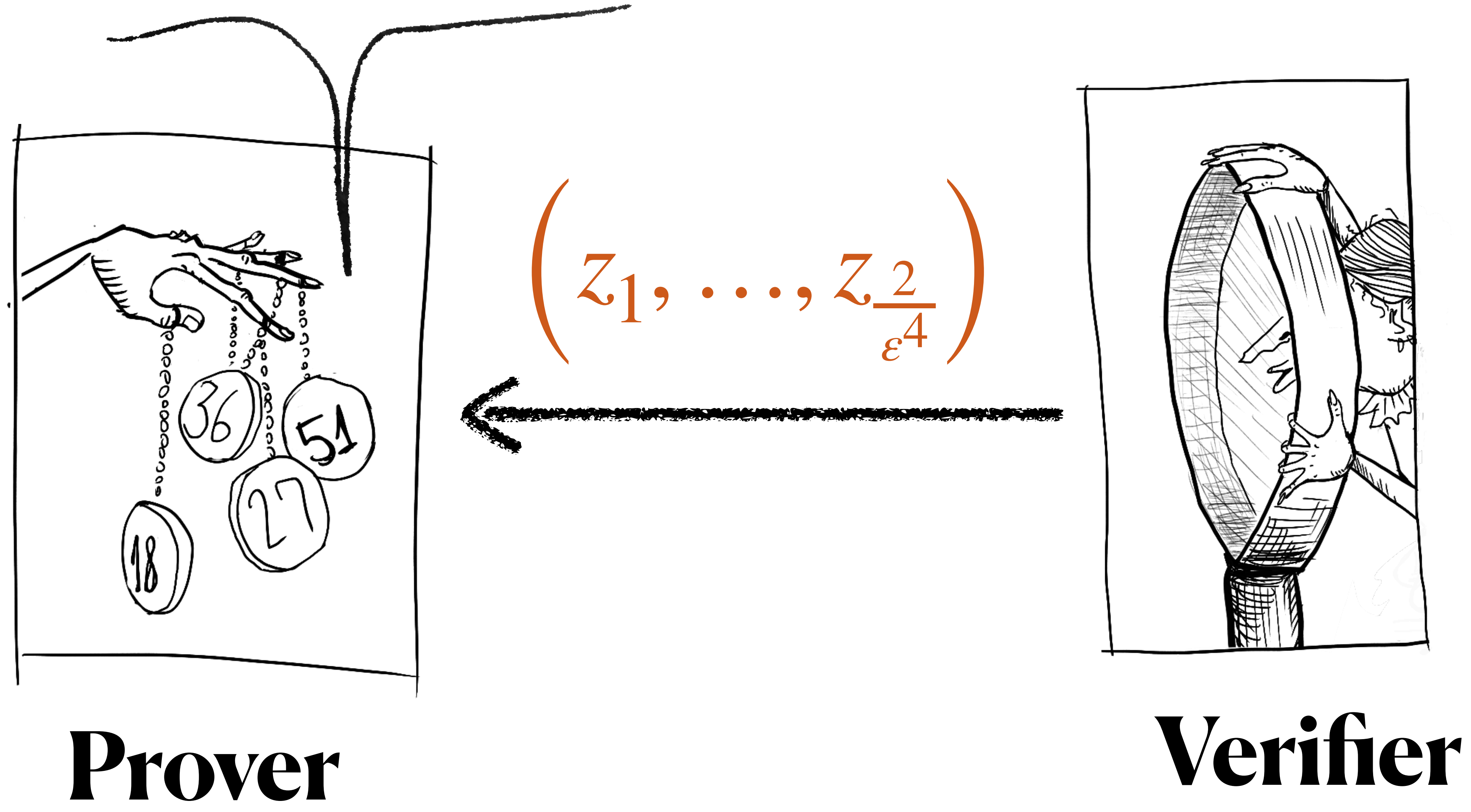
Verifier

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

Verifying Uniformity Over Half Domain

$D \in \mathbb{U}_{N/2}$



1. Draw $s = O\left(\sqrt{N}\right)$ samples.

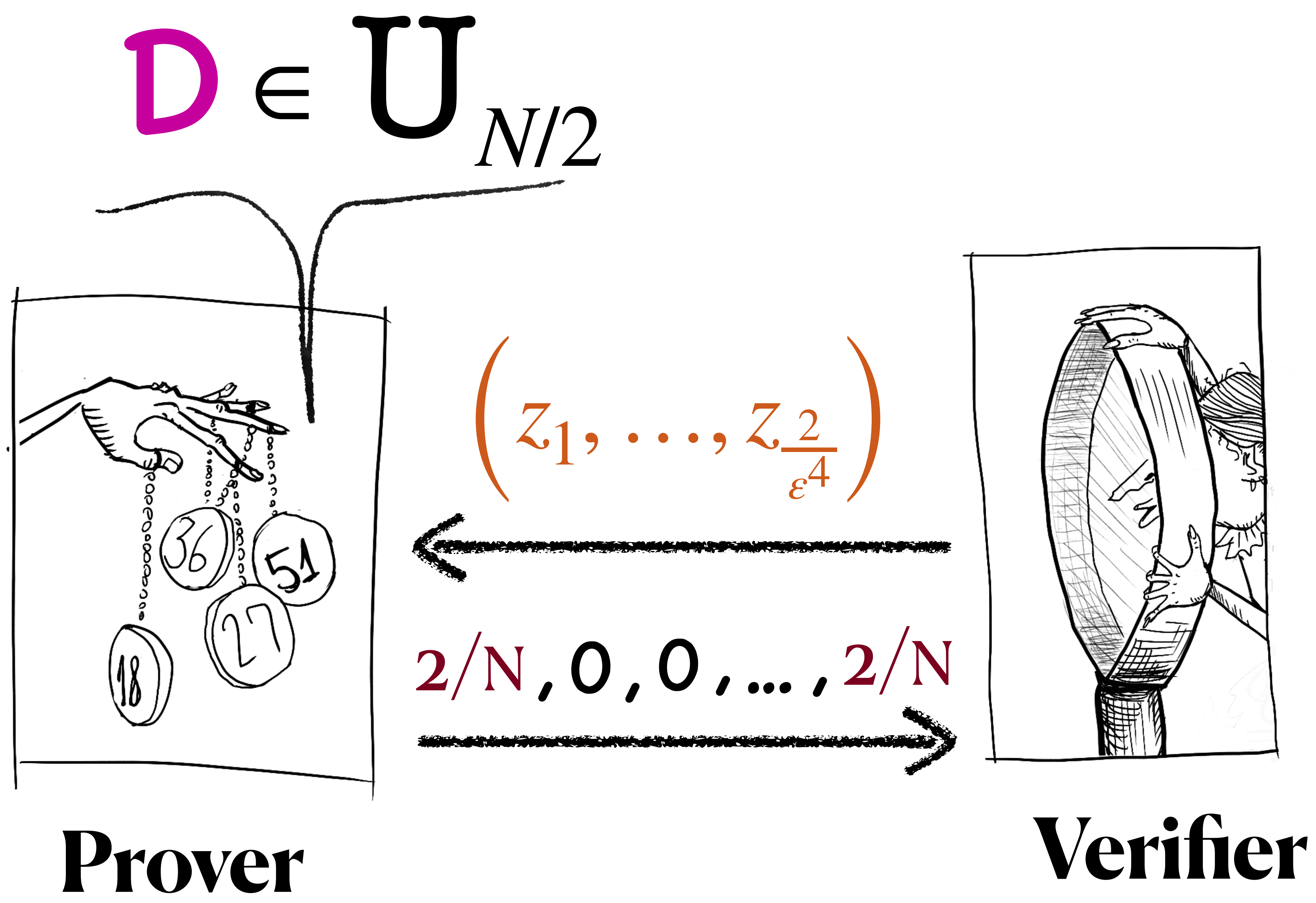
   Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

   Draw ε^{-4} samples from D.

   Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

Verifying Uniformity Over Half Domain



1. Draw $s = O\left(\sqrt{N}\right)$ samples.

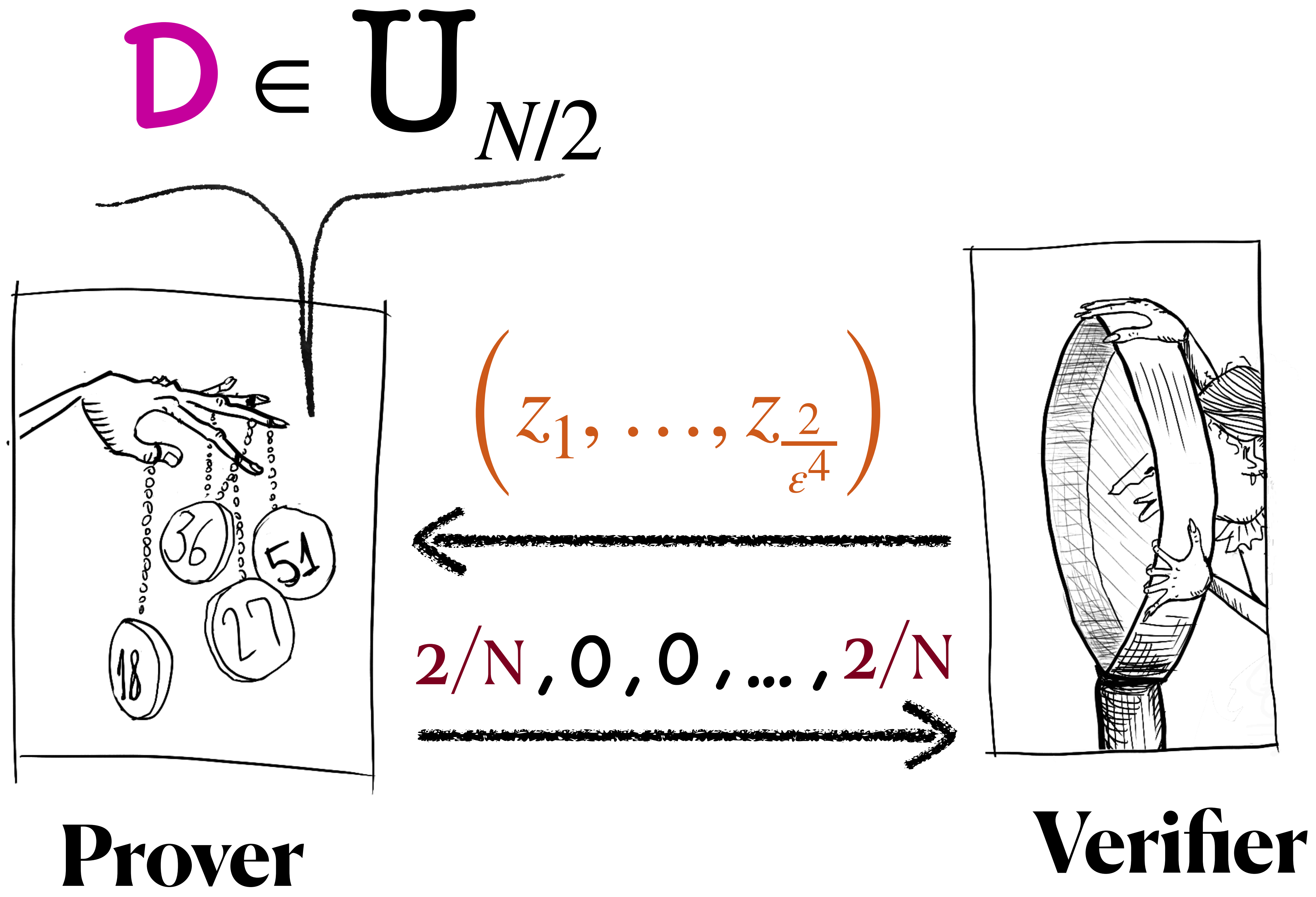
Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

Draw ε^{-4} samples from D.

Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

Verifying Uniformity Over Half Domain



1. Draw $s = O\left(\sqrt{N}\right)$ samples.

   Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

   Draw ε^{-4} samples from D.

   Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

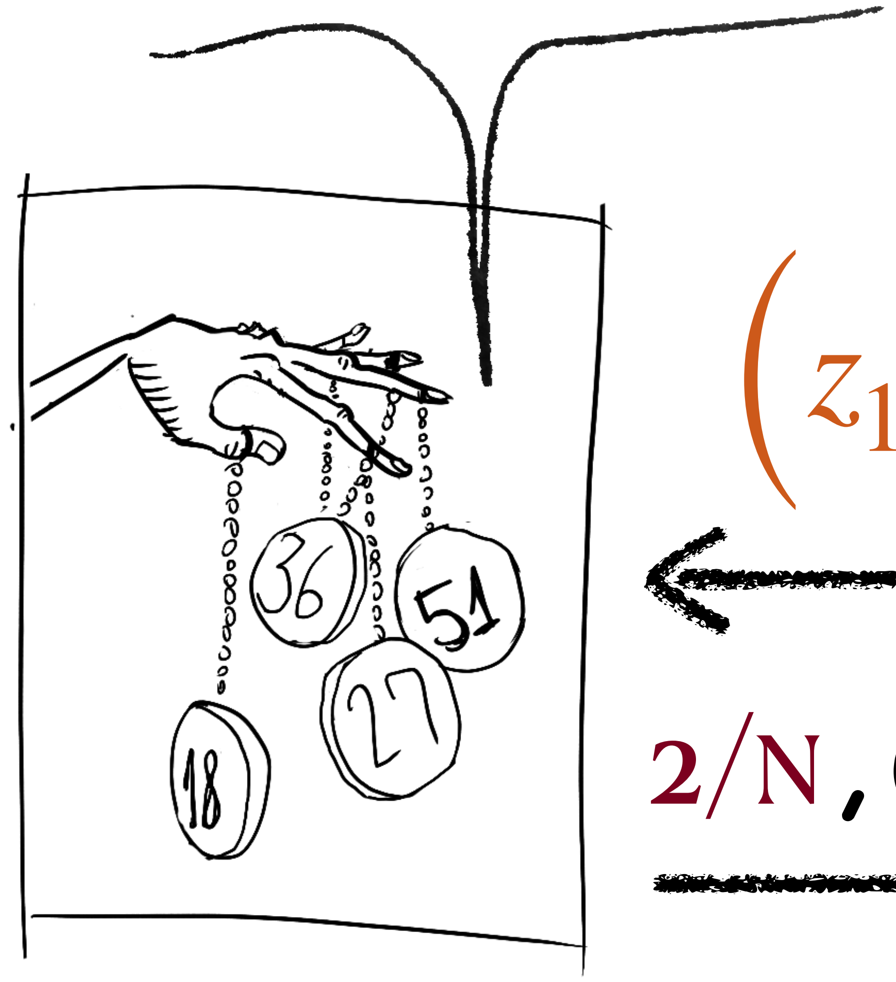
   Test 2:

   $\approx$ ***half*** the $U_{[N]}$-samples have tag $2/N$

   ALL D-samples have tag $2/N$

Verifying Uniformity Over Half Domain

$\mathbf{D} \in \mathbf{U}_{N/2}$

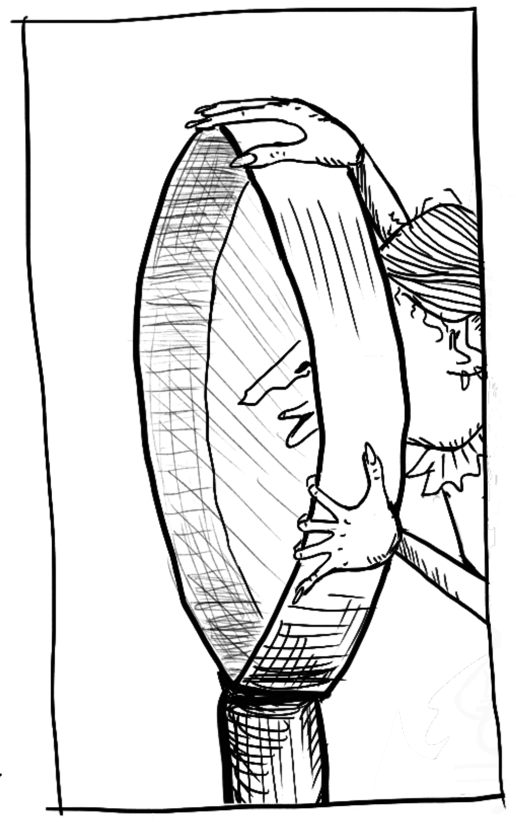


$\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

$\longleftarrow$

2/N , 0 , 0 , ... , 2/N

$\longrightarrow$



Prover

Verifier

Completeness:

Soundness:

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

Draw ε^{-4} samples from D.

Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

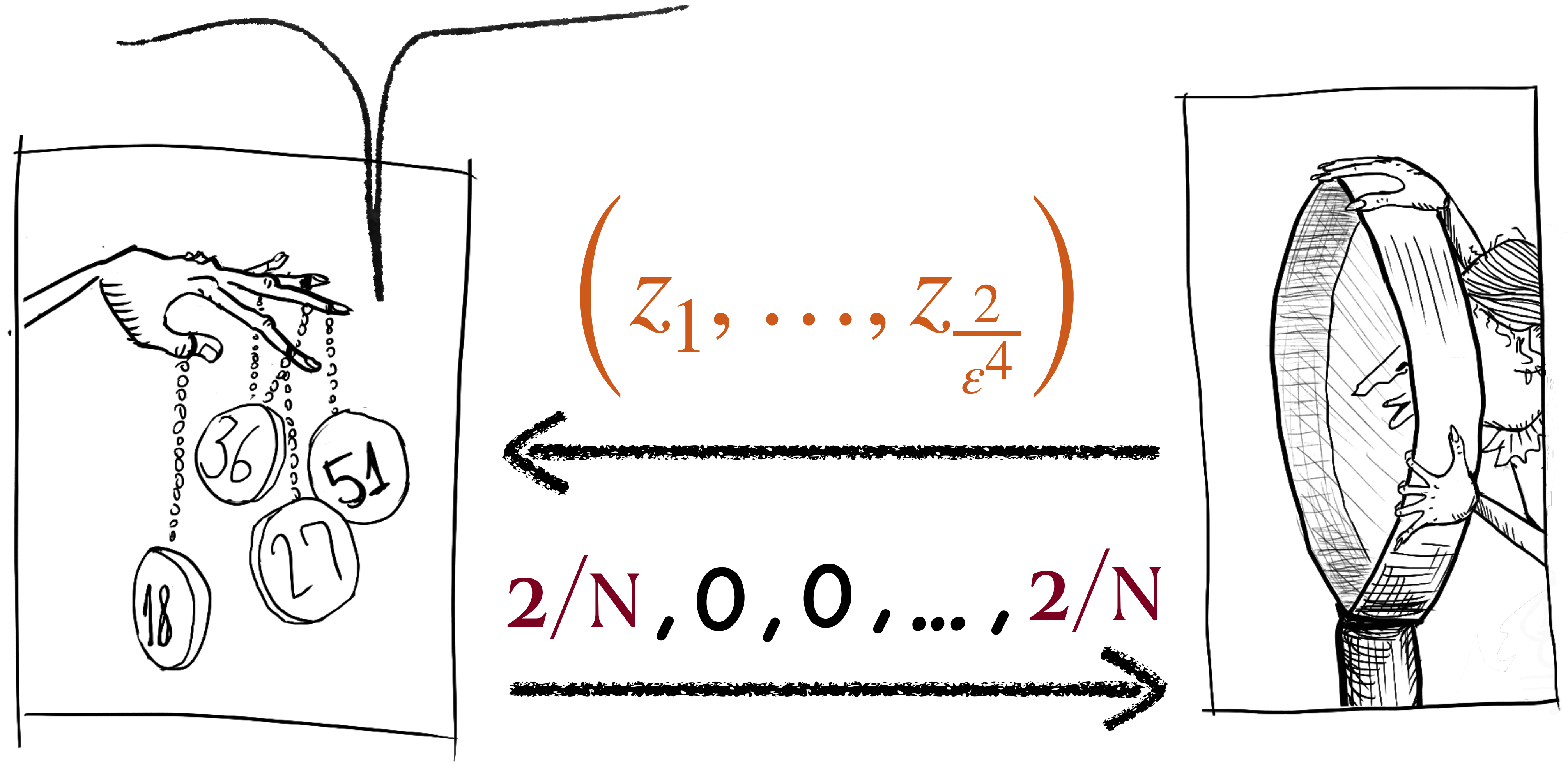
Test 2:

$\approx$ ***half*** the $U_{[N]}$-samples have tag 2/N

ALL D-samples have tag 2/N

Verifying Uniformity Over Half Domain

$D \in U_{N/2}$



Prover **Verifier**

Completeness:

Soundness:

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

Draw ε^{-4} samples from D.

Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

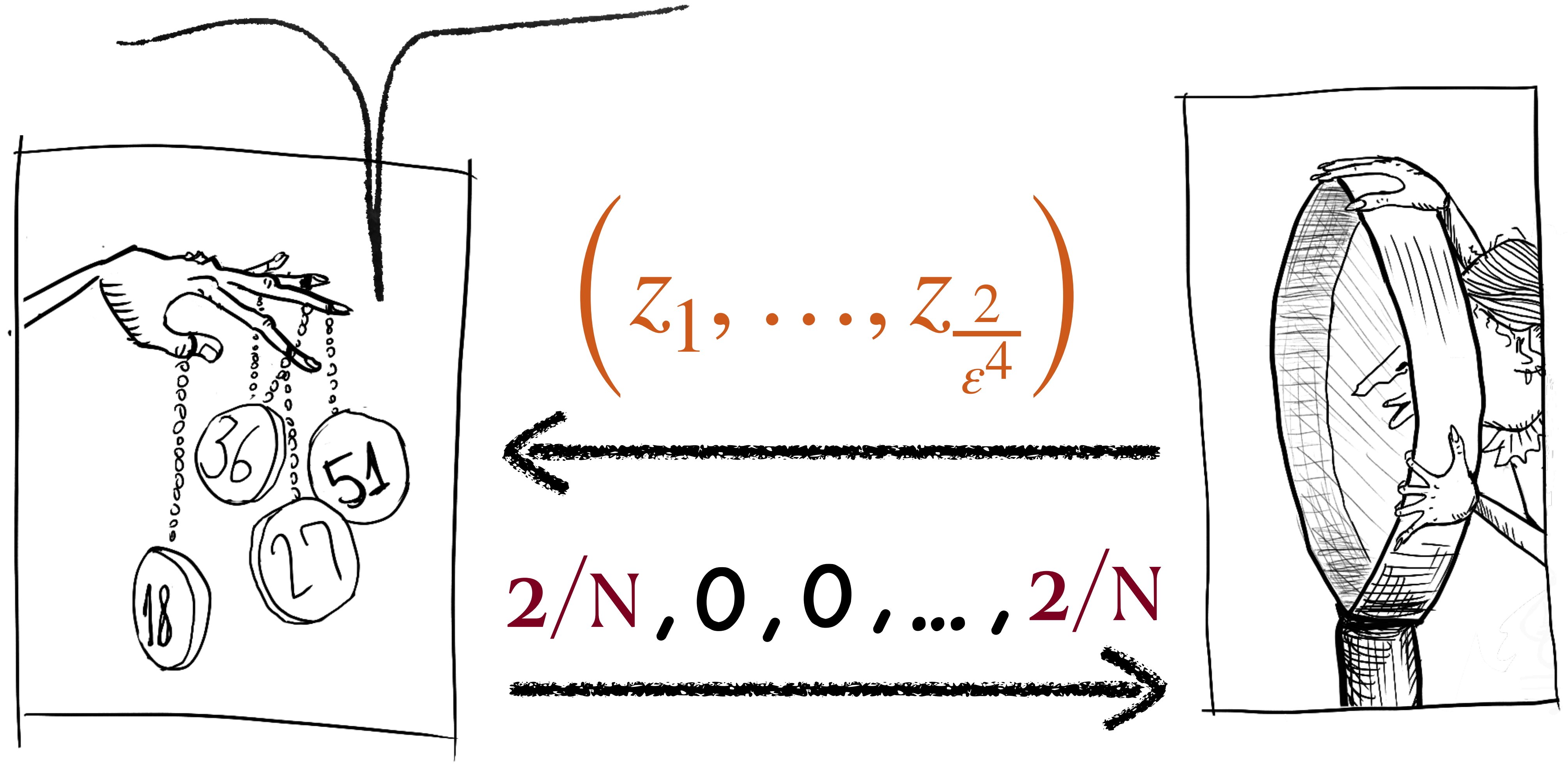
Test 2:

$\approx$ ***half*** the $U_{[N]}$-samples have tag $2/N$

ALL D-samples have tag $2/N$

Verifying Uniformity Over Half Domain

$\mathsf{D} \in \mathsf{U}_{N/2}$



Prover **Verifier**

Completeness: ✓

Soundness: ✓

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

Draw ε^{-4} samples from D.

Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

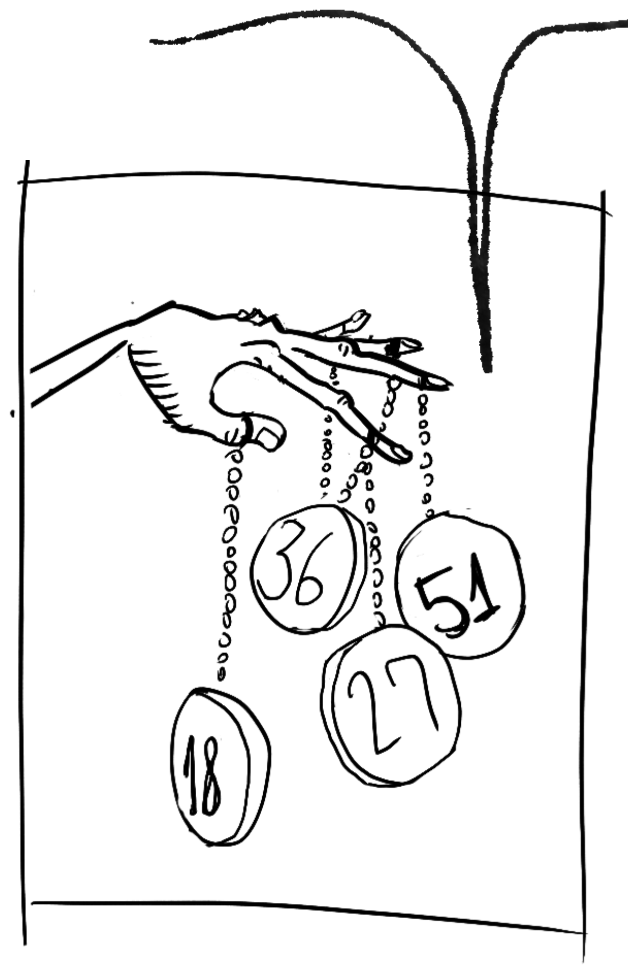
Test 2:

$\approx$ ***half*** the $U_{[N]}$-samples have tag $2/N$

ALL D-samples have tag $2/N$

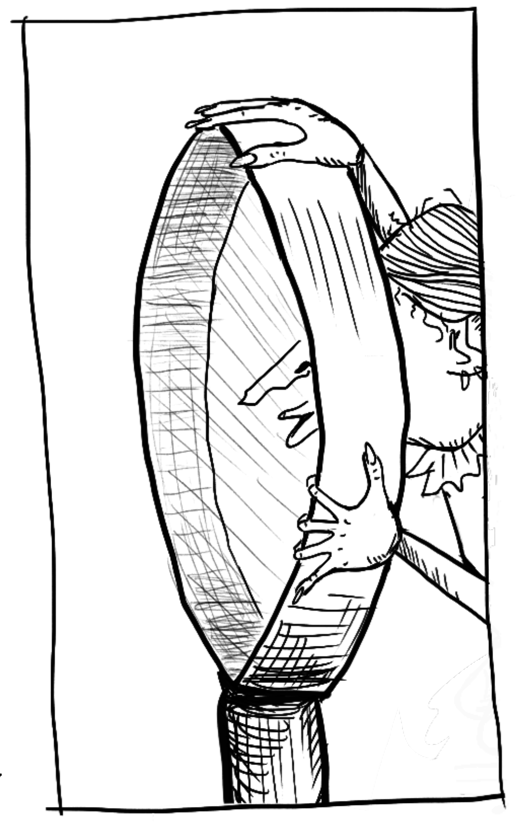
Verifying Uniformity Over Half Domain

$D \in \mathrm{U}_{N/2}$



$\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

2/N , 0 , 0 , ... , 2/N



Prover

Verifier

P samples: $O(N)$

V samples: $O\left(\sqrt{N}\right)$

Comm.: $O\left(\varepsilon^{-4}\right)$

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

   Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{2}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

   Draw ε^{-4} samples from D.

   Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

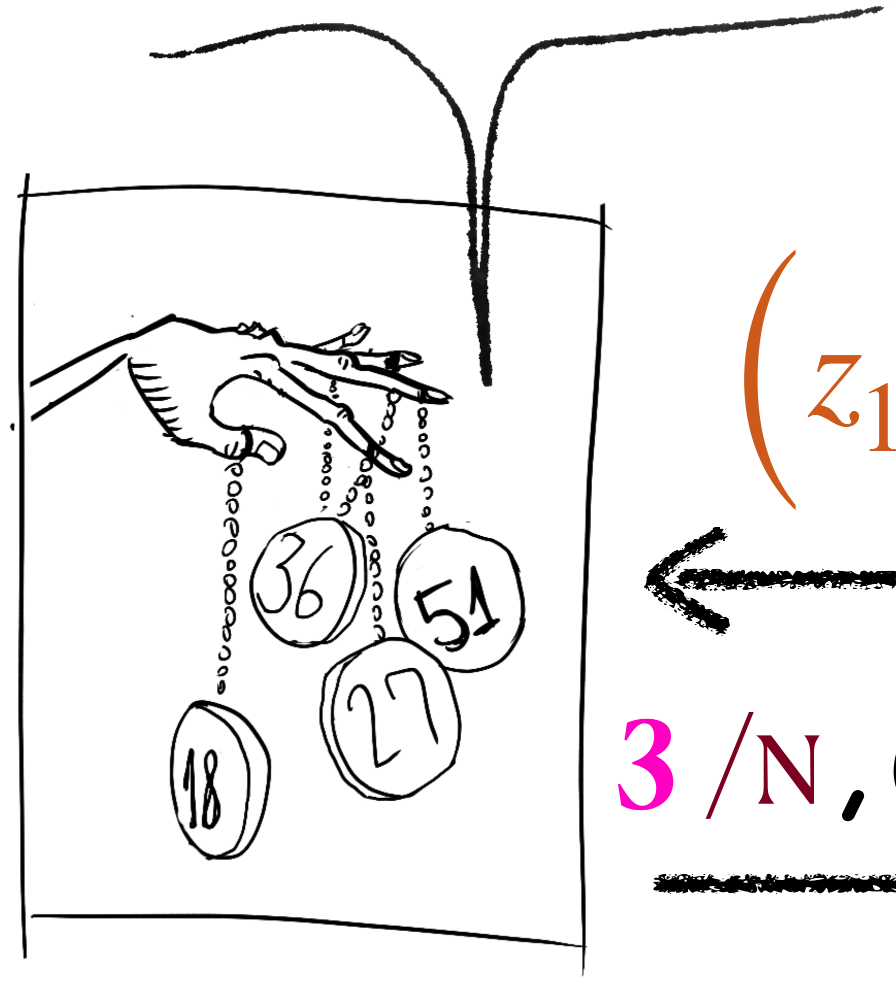
   Test 2:

   $\approx$ ***half*** the $U_{[N]}$-samples have tag 2/N

   ALL D-samples have tag 2/N

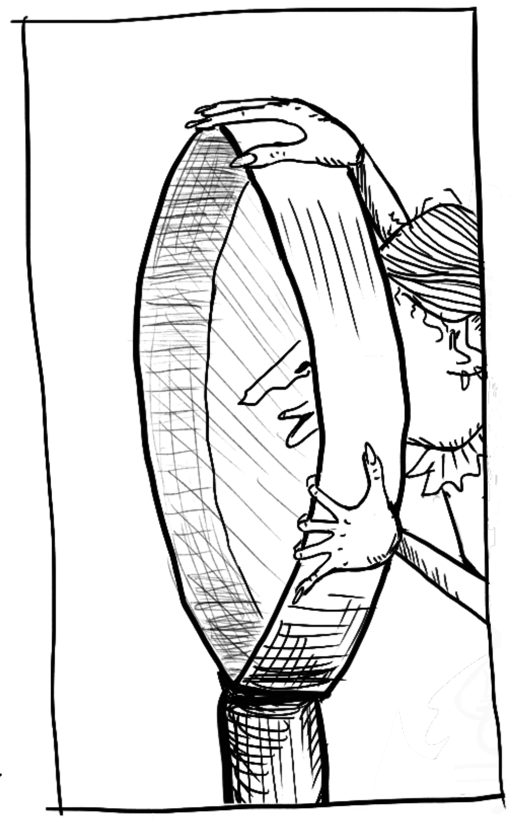
Verifying Uniformity Over Third Domain

$\mathsf{D} \in \mathsf{U}_{N/3}$



$\xleftarrow{\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)}$

$\xrightarrow{3/\mathrm{N},\ 0,\ 0,\ \ldots,\ 3/\mathrm{N}}$



Prover **Verifier**

P samples: $O(N)$

V samples: $O\left(\sqrt{N}\right)$

Comm.: $O\left(\varepsilon^{-4}\right)$

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

   Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{3}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

   Draw ε^{-4} samples from D.

   Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

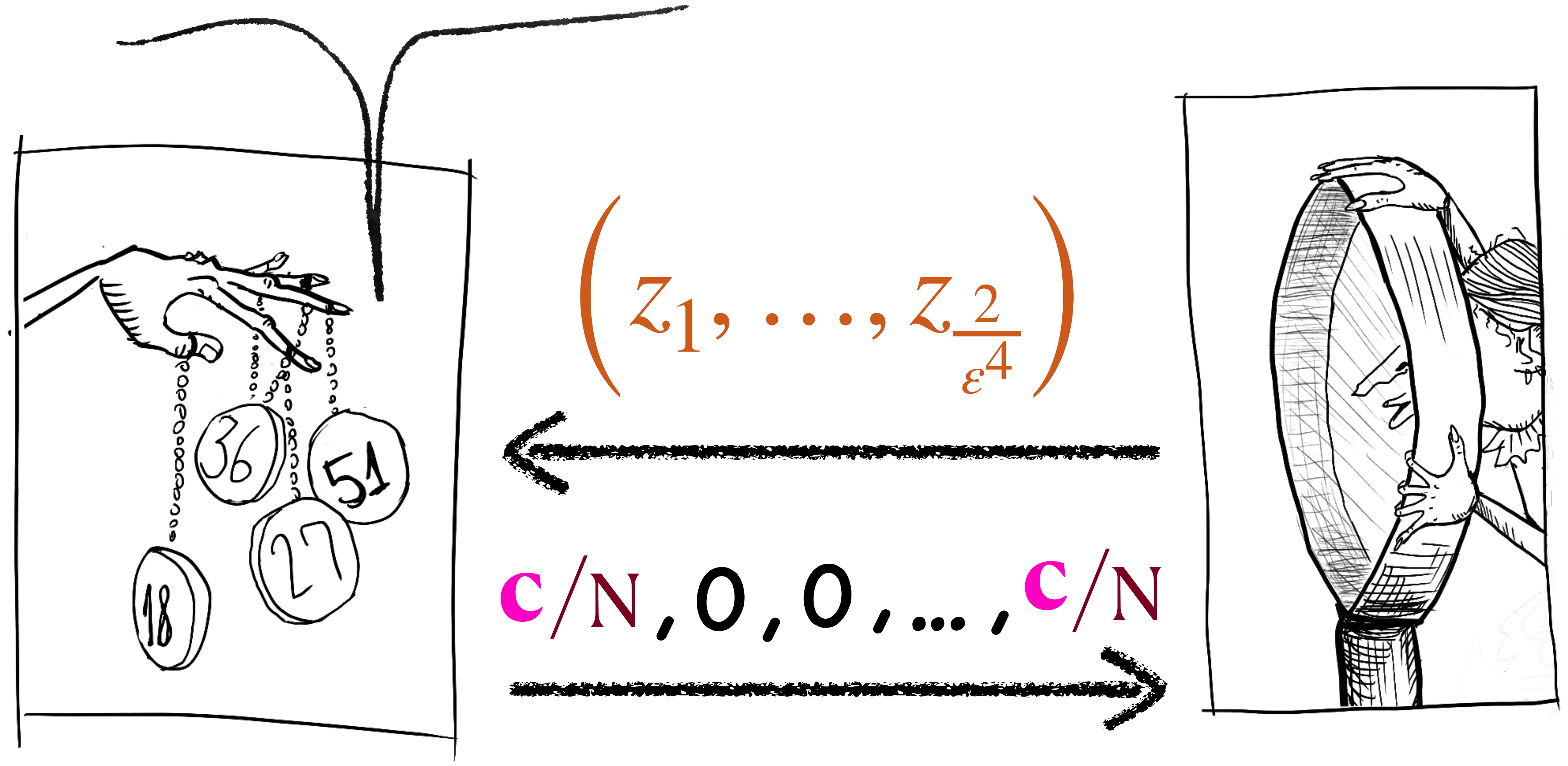
   Test 2:

   $\approx$ ***third*** the $U_{[N]}$-samples have tag $3/\mathrm{N}$

   ALL D-samples have tag $3/\mathrm{N}$

Verifying Uniformity Over Fraction of Domain

$\mathsf{D} \in \mathsf{U}_{N/\mathbf{c}}$



Prover **Verifier**

P samples: $O(N)$

V samples: $O\left(\sqrt{N}\right)$

Comm.: $O\left(\varepsilon^{-4}\right)$

1. Draw $s = O\left(\sqrt{N}\right)$ samples.

   Test 1: #of 2-collisions is $\approx \binom{s}{2} \cdot \frac{\mathbf{c}}{N}$

2. Draw ε^{-4} samples from $U_{[N]}$.

   Draw ε^{-4} samples from D.

   Randomly mix: $\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

   Test 2:

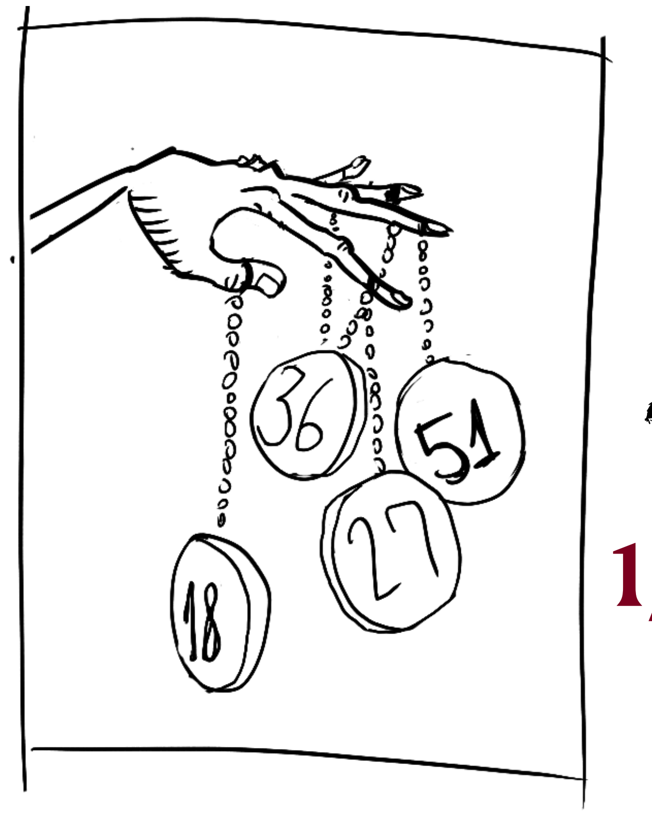
   $\approx$ *1/c* the $U_{[N]}$-samples have tag $\mathbf{c}/\mathrm{N}$

   ALL D-samples have tag $\mathbf{c}/\mathrm{N}$

D is supported over:

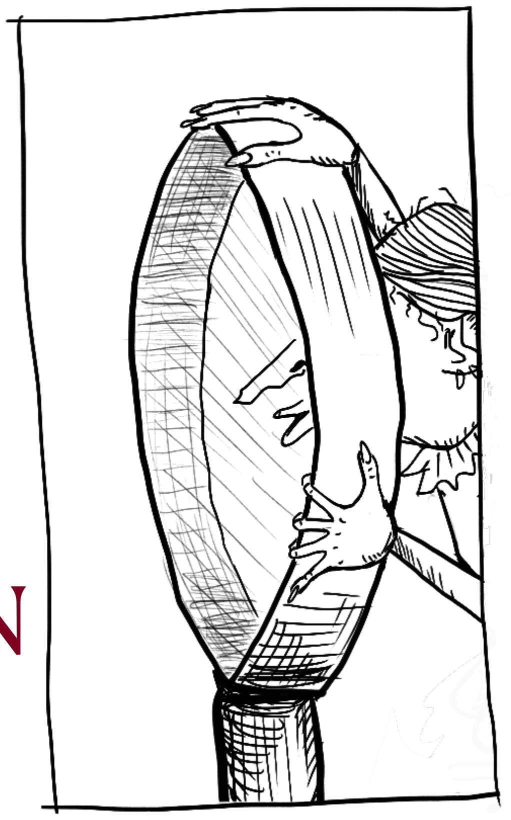
S_1 elements of probability $1/N$

S_2 elements of probability $2/N$



$\left(z_1, \ldots, z_{\frac{2}{\varepsilon^4}}\right)$

$1/N, 2/N, 0, \ldots, 1/N$



Prover

Verifier

P samples: $O(N)$

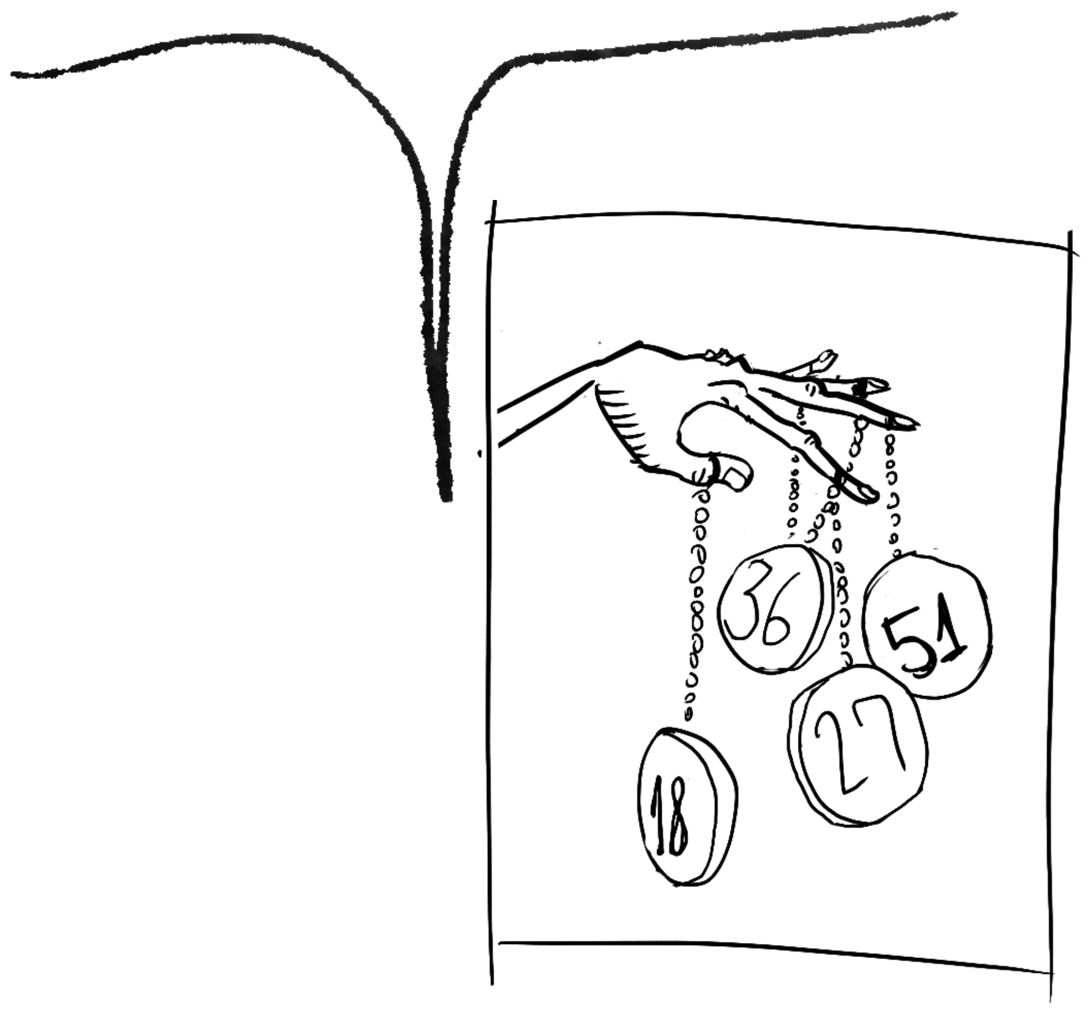
Comm.: $O\left(\varepsilon^{-4}\right)$

V samples: $O\left(\sqrt{N}\right)$

$\forall j \in [\log N \,/\, \varepsilon]$

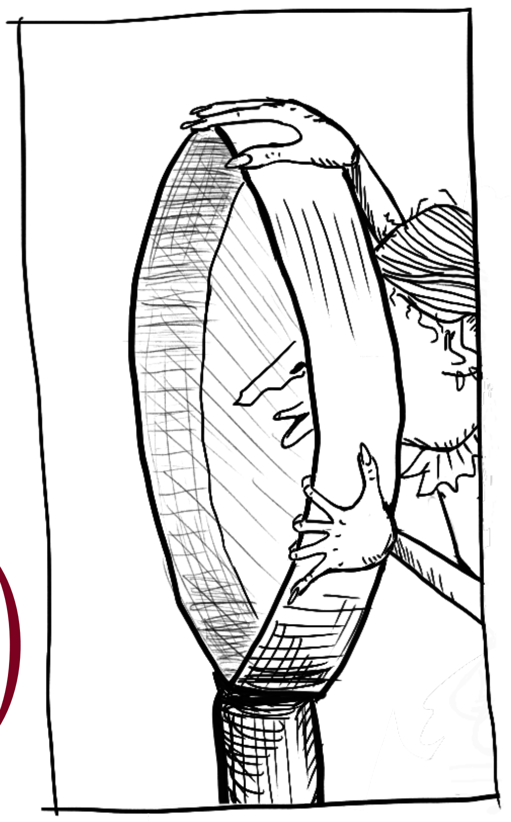
D is supported over:

S_j elements w.p. $\approx \dfrac{\varepsilon \cdot (1+\varepsilon)^j}{N}$



$\left(z_1, \ldots, z_s\right)$

$\left(\widetilde{D}_{z_1}, \widetilde{D}_{z_2}, \ldots, \widetilde{D}_{z_s}\right)$



Prover

Verifier

Partition of $[0,1]$:

$\left(\dfrac{\varepsilon}{N}, \dfrac{\varepsilon}{N}\cdot(1+\varepsilon)\right]$

$\left(\dfrac{\varepsilon}{N}\cdot(1+\varepsilon), \dfrac{\varepsilon}{N}\cdot(1+\varepsilon)^2\right]$

$\vdots$

$\left(\dfrac{\varepsilon}{N}\cdot(1+\varepsilon)^j, \dfrac{\varepsilon}{N}\cdot(1+\varepsilon)^{j+1}\right]$

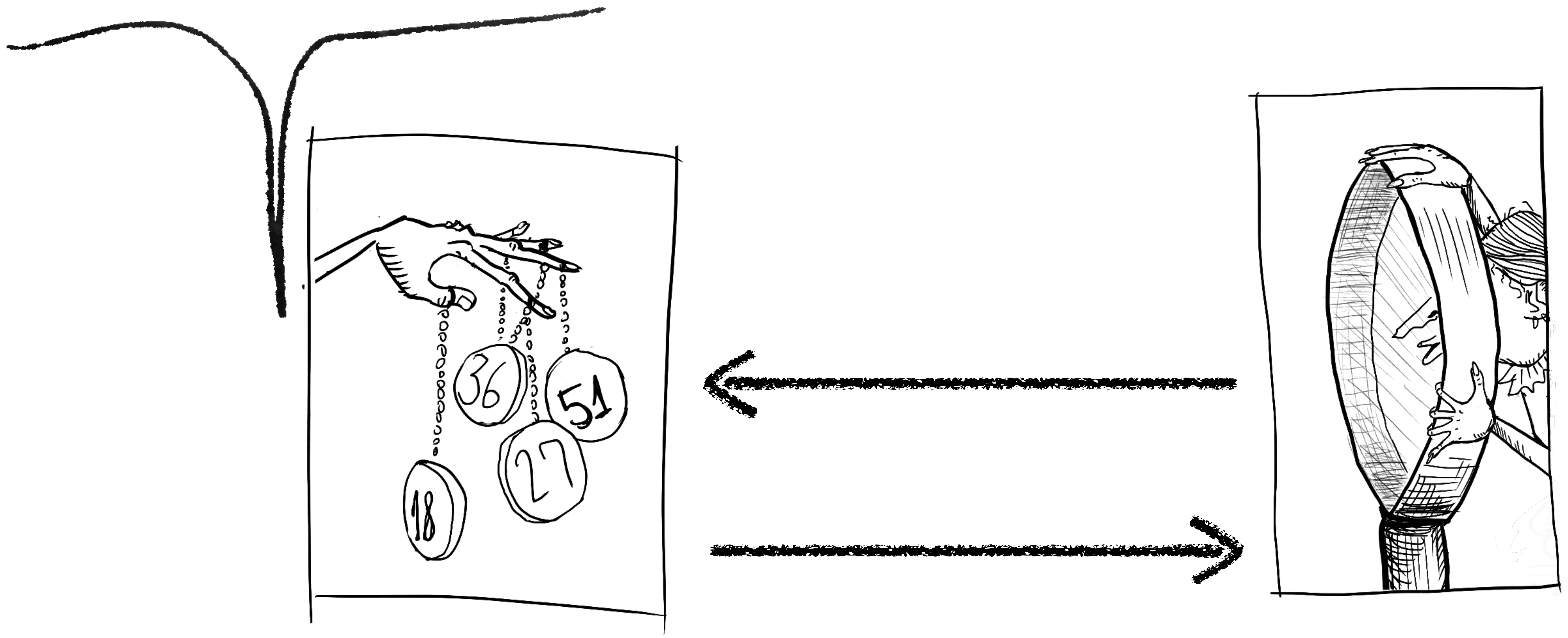
$\vdots$

P samples: $O(N)$

Comm.: $O\left(\sqrt{N}\right)$

V samples: $O\left(\sqrt{N}\right)$

$\mathcal{D}$ has probability histogram $hist_0$



Prover

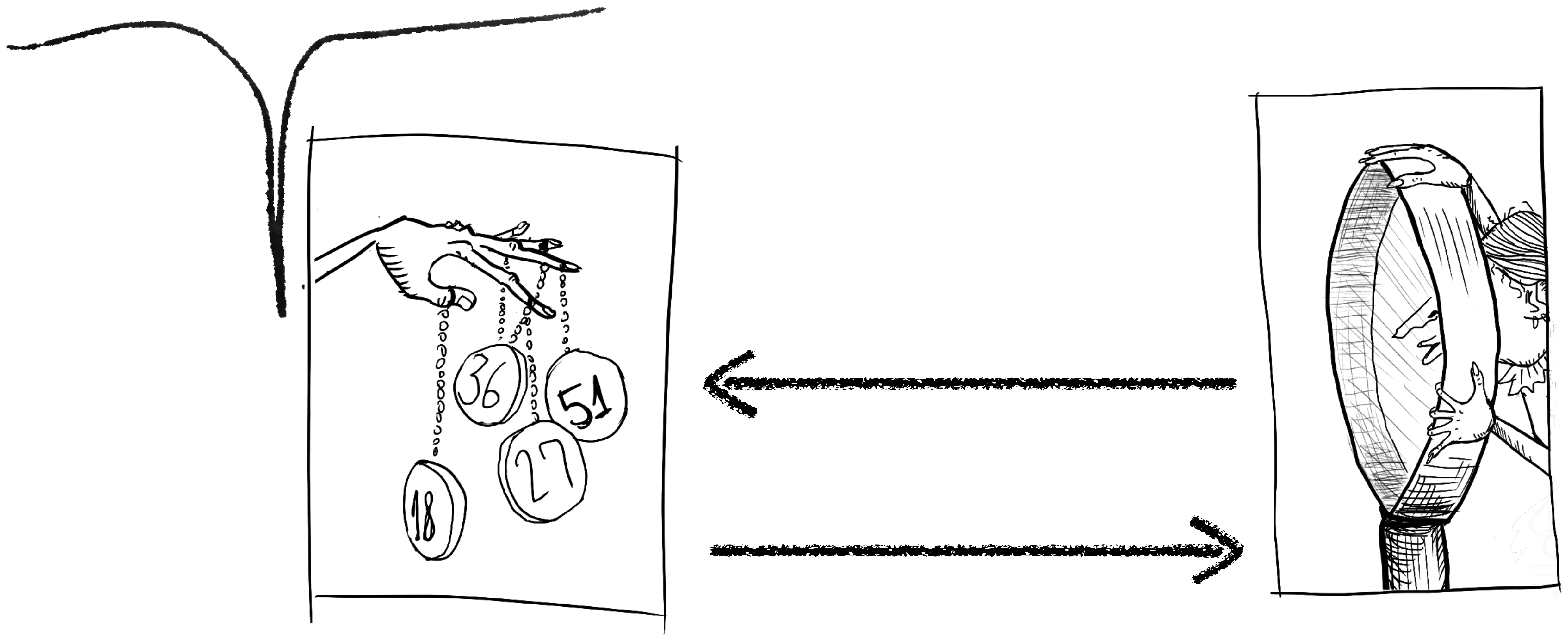
Verifier

P samples: $O(N)$

Comm.: $O\left(\sqrt{N}\right)$

V samples: $O\left(\sqrt{N}\right)$

$\mathcal{D}$ has probability histogram $hist_0$



Prover

Verifier

Verify histogram, then compute:

- Shannon entropy!
- Distance from uniformity!
- Support size!

P samples: $O(N)$

Comm.: $O\left(\sqrt{N}\right)$

V samples: $O\left(\sqrt{N}\right)$

Theorem [H-Rothblum'22,H-Rothblum'23]

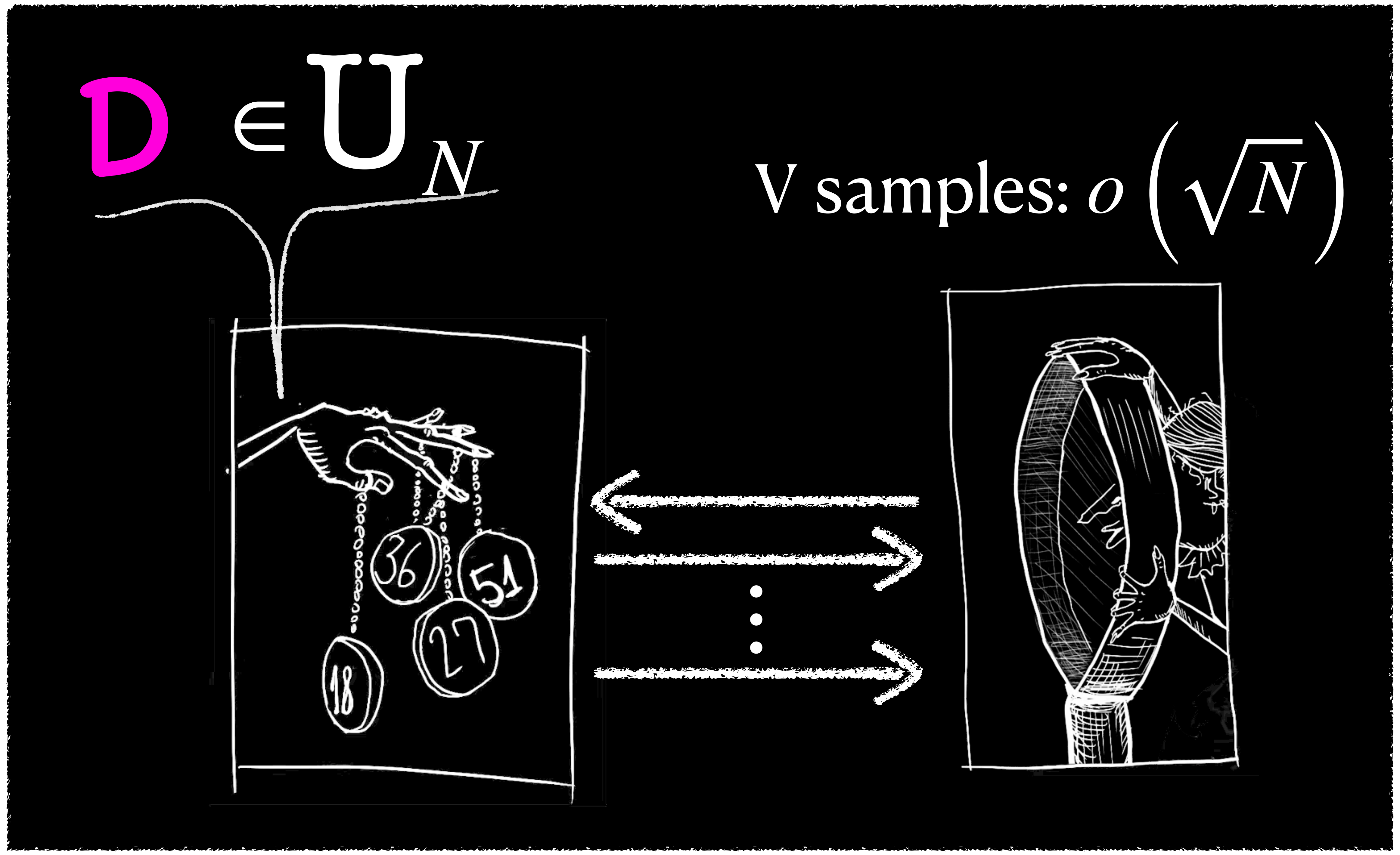
Every **label-invariant** distribution property admits a 2-message tolerant proof system with:*

P samples: $O(N)$ **Comm.**: $O\left(\sqrt{N}\right)$ **V samples**: $O\left(\sqrt{N}\right)$

Can we verify uniformity with $o\left(\sqrt{N}\right)$*?*

Can we verify uniformity with $o\left(\sqrt{N}\right)$?

Assume...



Can we verify uniformity with $o\left(\sqrt{N}\right)$?

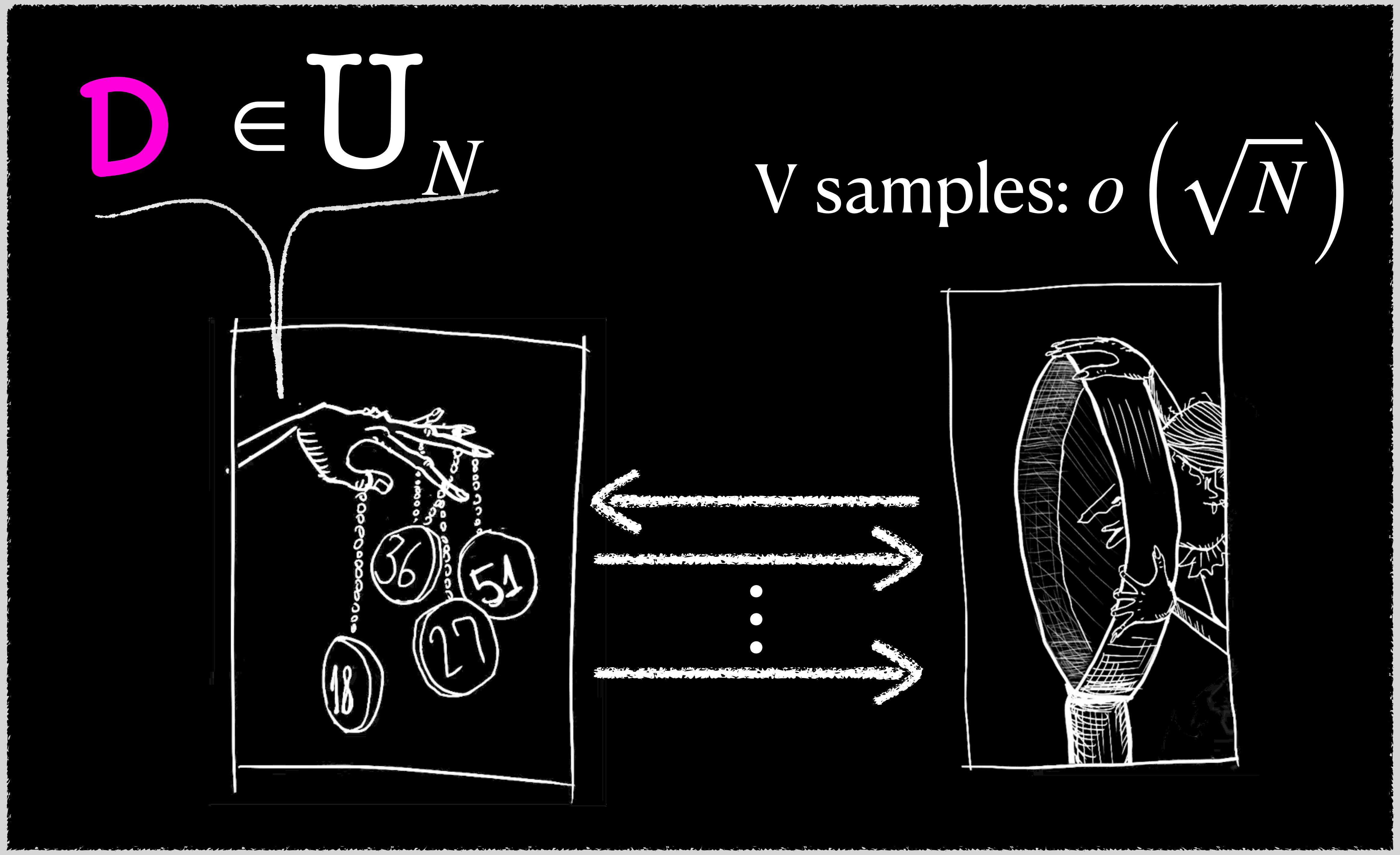
Tester [CG'17]

Total samples: $o\left(\sqrt{N}\right)$

Assume...

Simulate by answering according to $U_{[N]}$

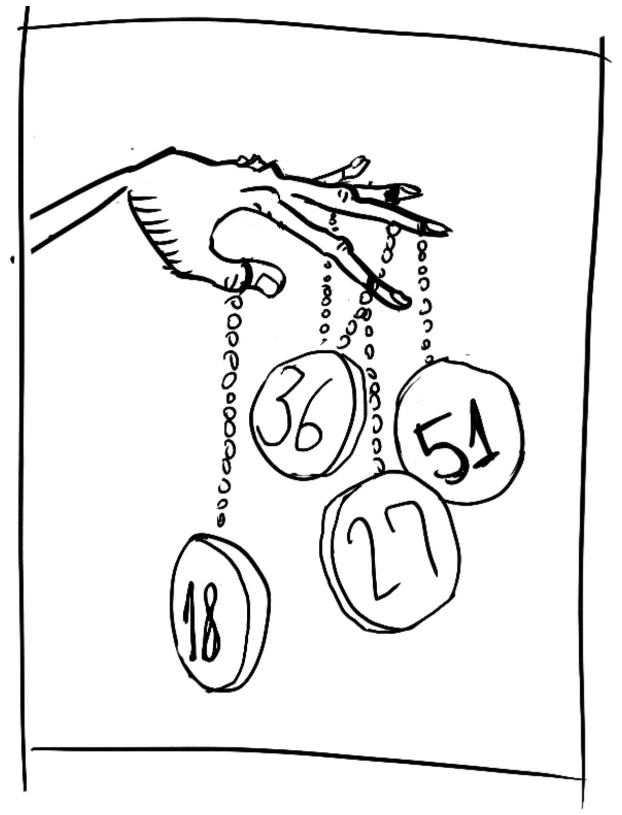
No samples needed!



Simulate V

$o\left(\sqrt{N}\right)$ ***samples***

Main tool:



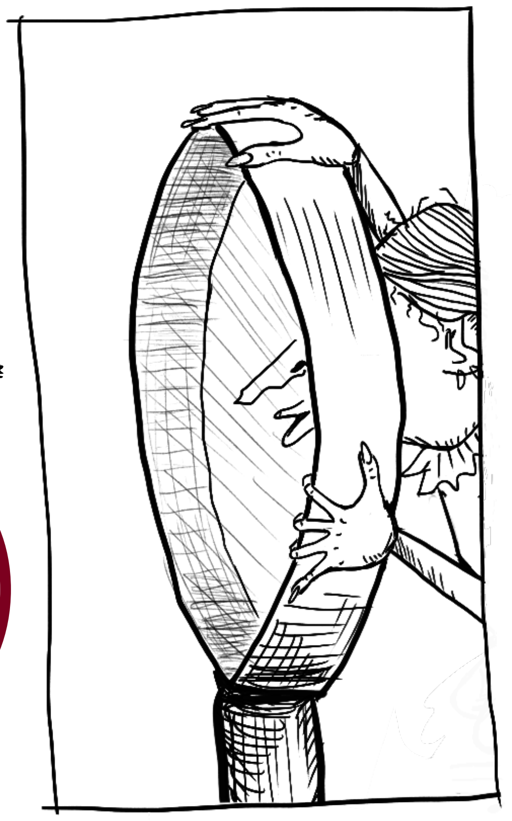
Prover

$\left(z_1, \ldots, z_s\right)$

$\longleftarrow$

$\left(\widetilde{D}_{z_1}, \widetilde{D}_{z_2}, \ldots, \widetilde{D}_{z_s}\right)$

$\longrightarrow$



Verifier

$$\left(z_i, \widetilde{D}_{z_i}\right)_{i\in[s]} \text{ s.t. } \widetilde{D}_{z_i} \approx D(z_i)$$

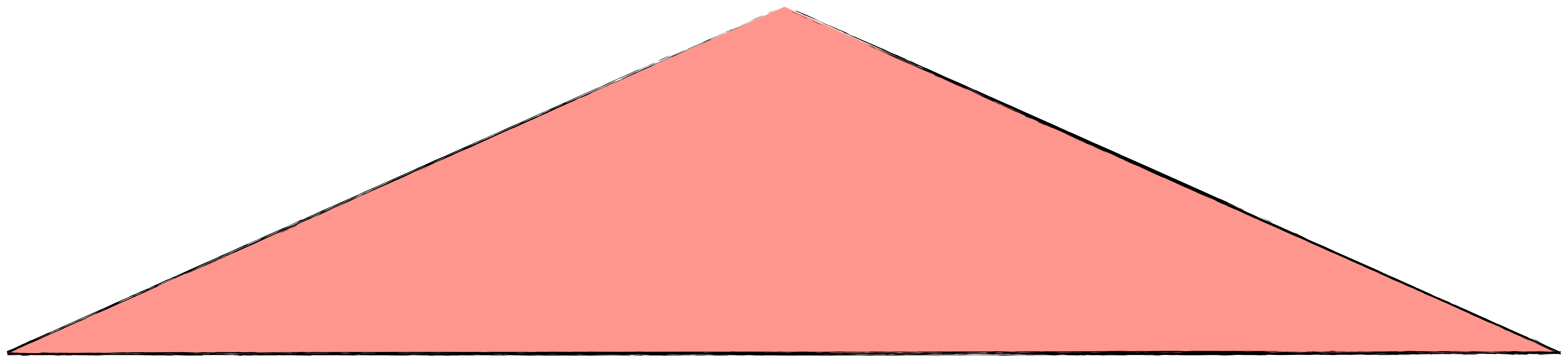
P samples: $O(N)$

Comm.: $O\left(\sqrt{N}\right)$

V samples: $O\left(\sqrt{N}\right)$

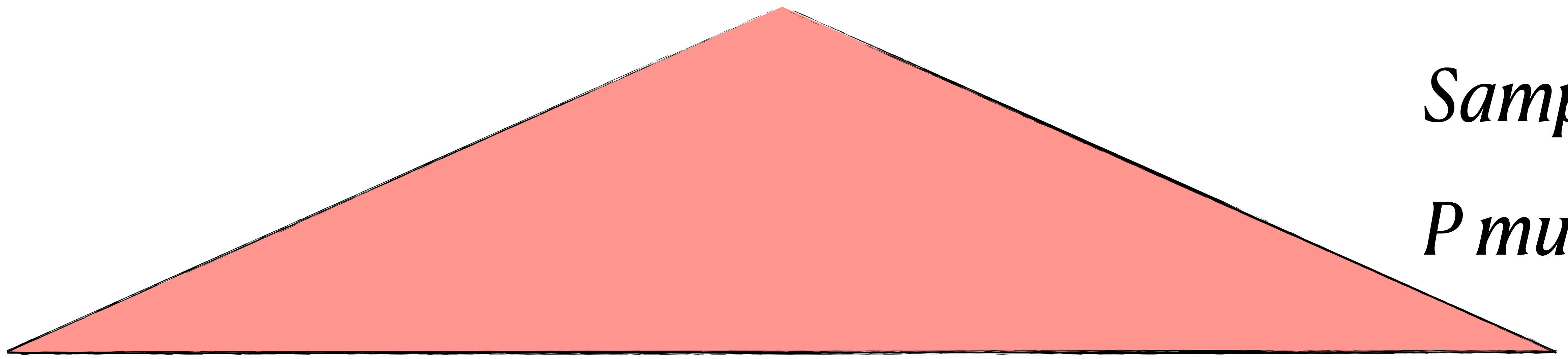
$$\mathbf{D} \quad \left((1, D(1)), (2, D(2)), \ldots, (N, D(N)) \right)$$

$b \in \{0,1\}$



D $\quad \big((1, D(1)), (2, D(2)), \ldots, (N, D(N)) \big)$

$b \in \{0,1\}$



Almost ***IPP...***

Sample and not query access

P must know the accessed locations.

$\mathbf{D} \quad \big((1, D(1)), (2, D(2)), \ldots, (N, D(N)) \big)$

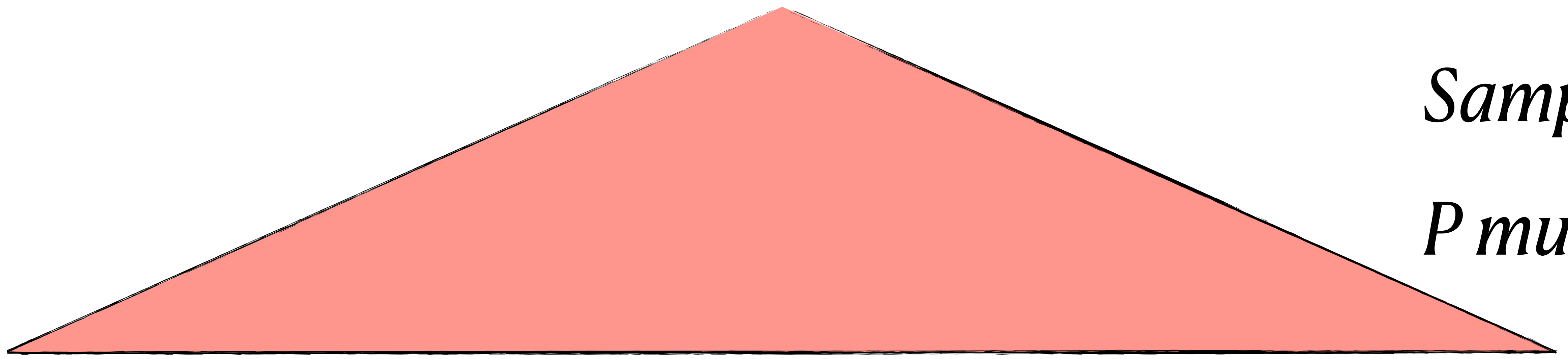
Theorem [H-Rothblum'24]

Every distribution property of ***reasonable complexity*** *admits a*

$(\log N)$*-messages tolerant proof system with:*

P samples: $O\left(N^{1.1}\right)$ **Comm.**: $O\left(N^{0.95}\right)$ **V samples**: $O\left(N^{0.9}\right)$

$b \in \{0,1\}$



D $\left((1,D(1)), (2,D(2)), \ldots, (N,D(N)) \right)$

Almost ***IPP...***

Sample and not query access

P must know the accessed locations.

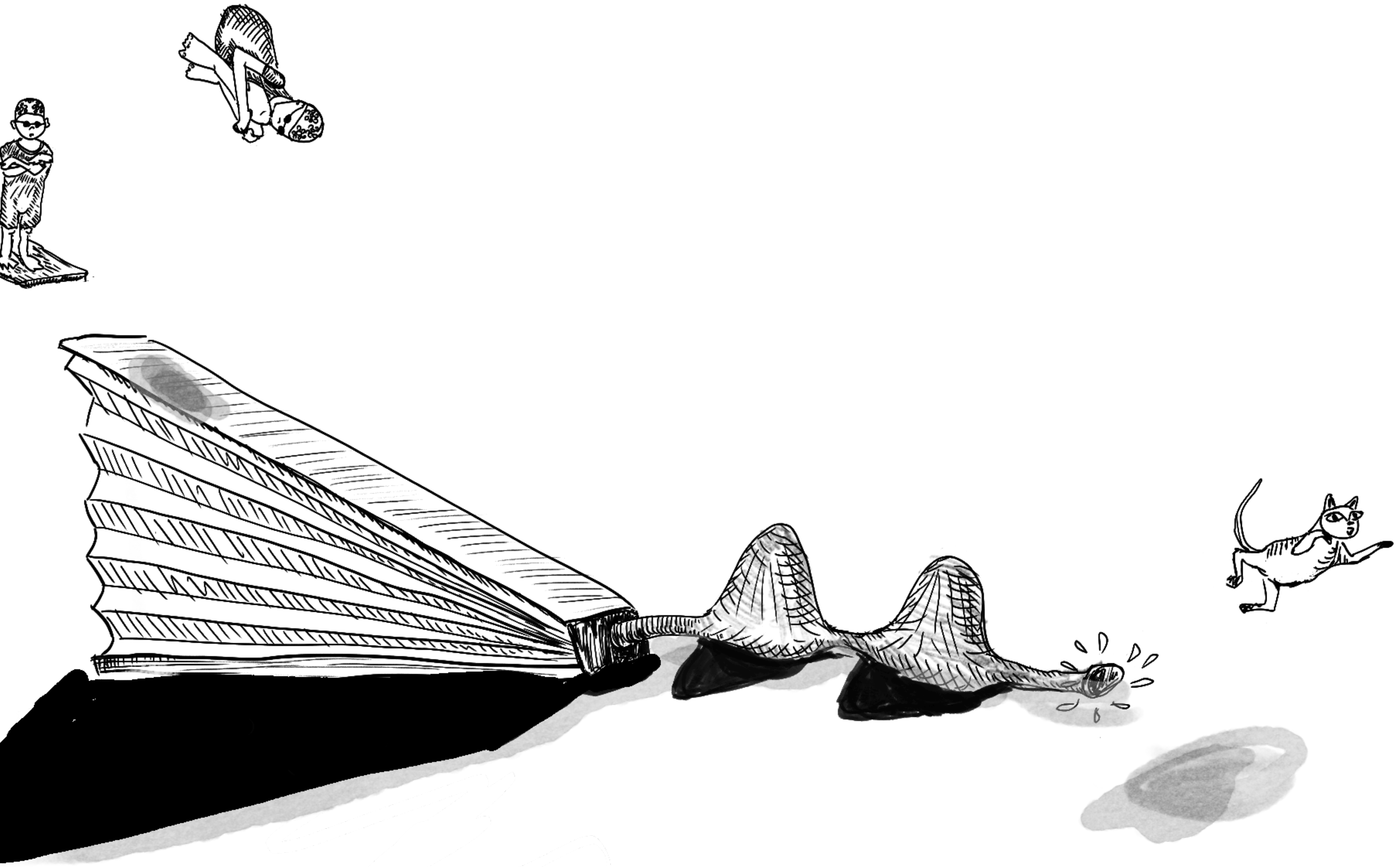
More Results...

- Exploring ***features of protocols***: public coin [**H**'24], differential privacy [DDTZ'26], zero knowledge [CPSTY'26], doubly sublinear proof system [G**H**R'25])
- ***Argument systems*** for every* property with $O\left(N^{1/2}\right)$ V samples & comm. [**H**-Rothblum'25].
- Exploring ***different access models***: PCond verification [BBCS'25]
- **Lower bounds:**
    - ***Proving*** membership in distribution properties can be ***strictly harder than testing*** [**H**-Rothblum'25].
    - ***The comm. complexity*** of verifying label-invariant properties is $\widetilde{\Theta}\left(N \cdot \sum_x D^2(x)\right)$ [**H**'26].

Summary and Future work

Verifying distribution properties can be done more efficiently than testing.

New lens for insights into distribution properties, and a rich area for future work.



Summary and Future work

Verifying distribution properties can be done more efficiently than testing.

New lens for insights into distribution properties, and a rich area for future work.

Open Questions:

Improving parameters, exploring different features (ZK, DP...).

Proofs for different models (continuous distributions, structured domains...)

Applications:

- *Verifiable Data Science, verifiable learning*
- Applications for AI Verification:
  - Unique properties, domains, and access models.