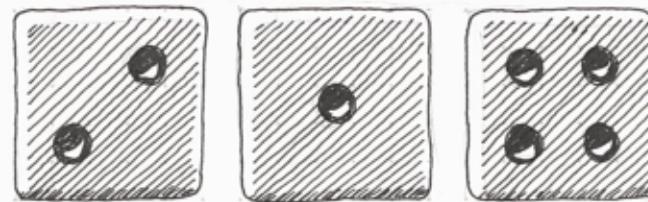


# Understanding RANDOMNESS

Without

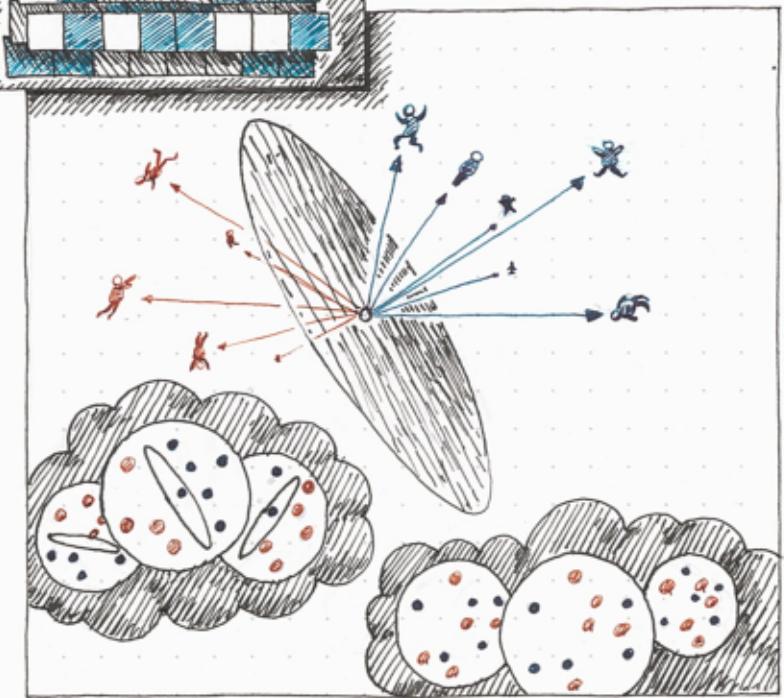
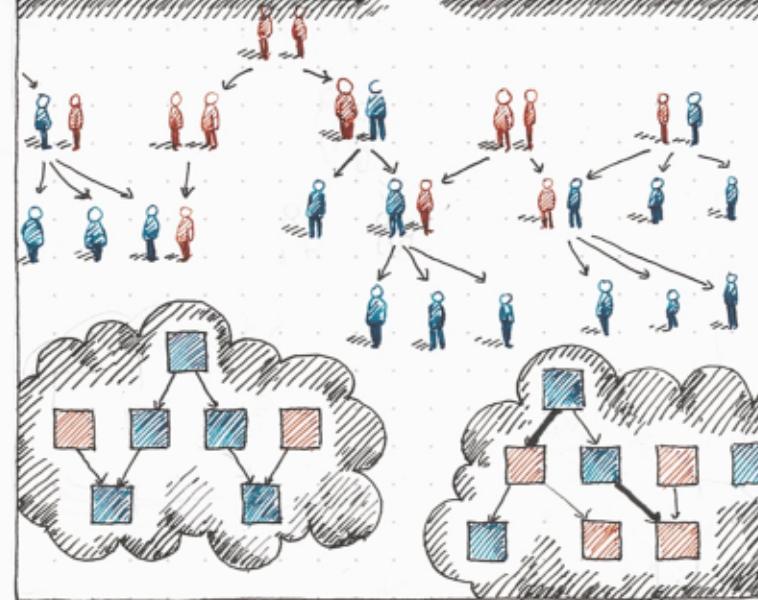
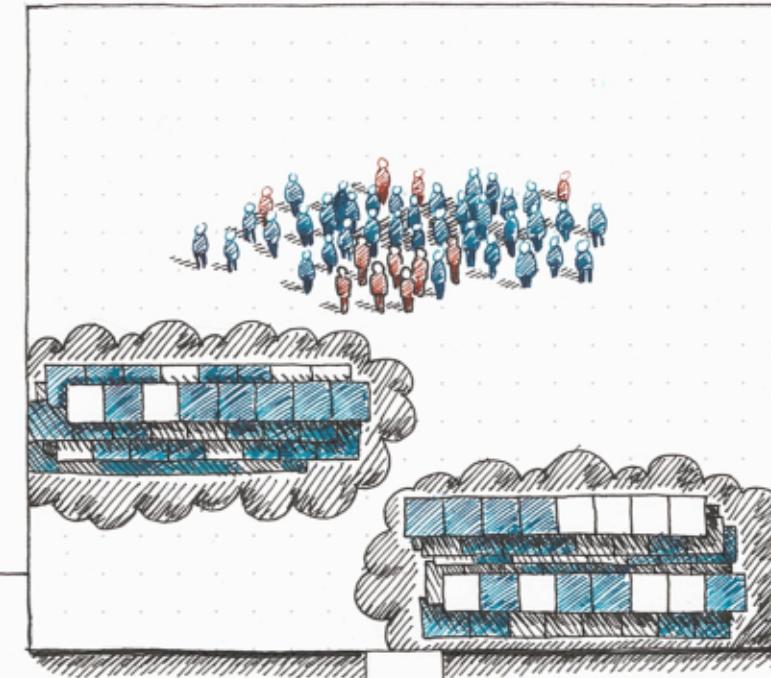
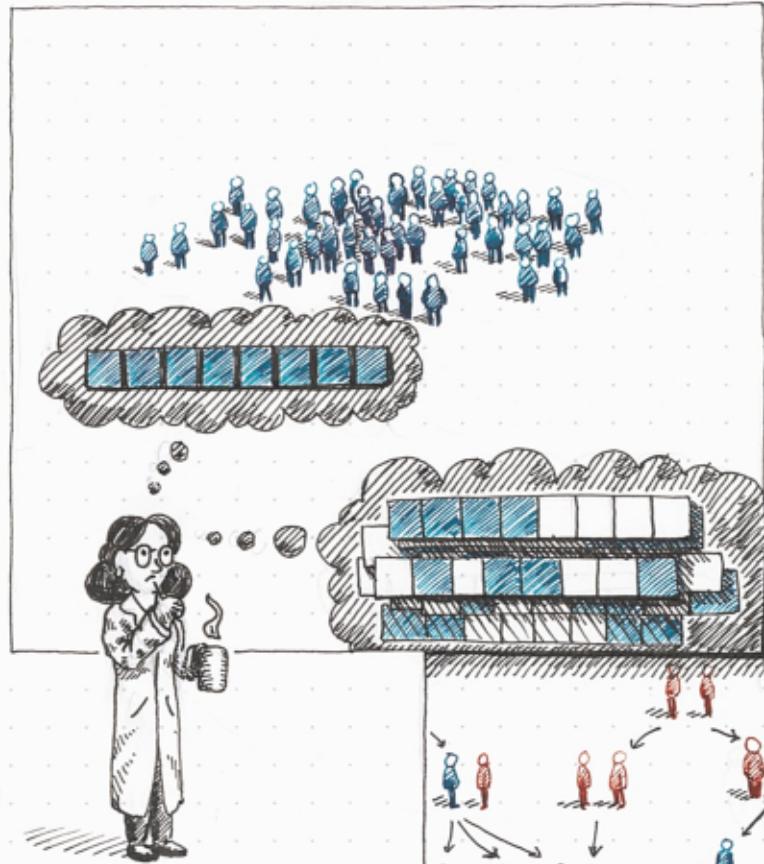


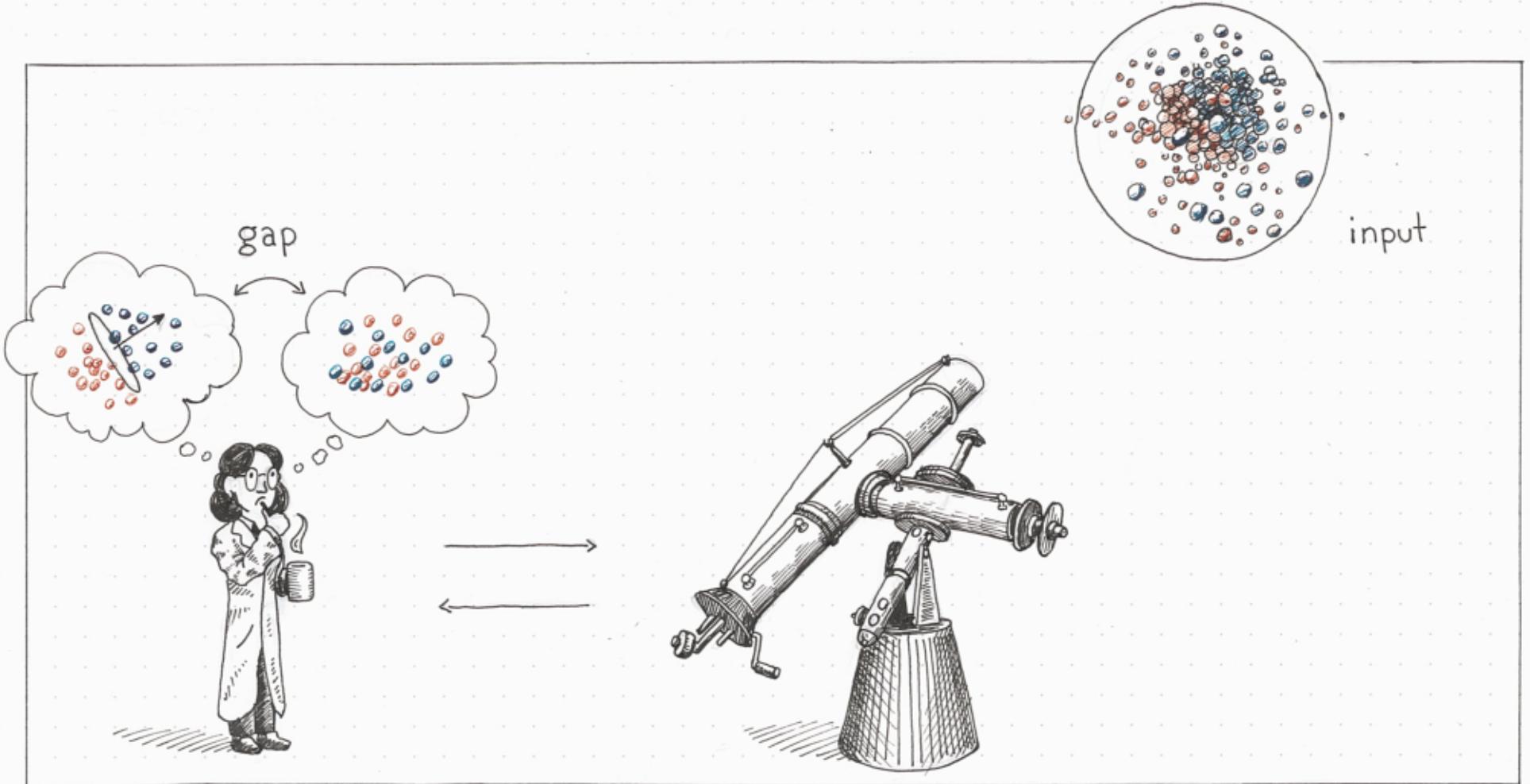
# RANDOMNESS



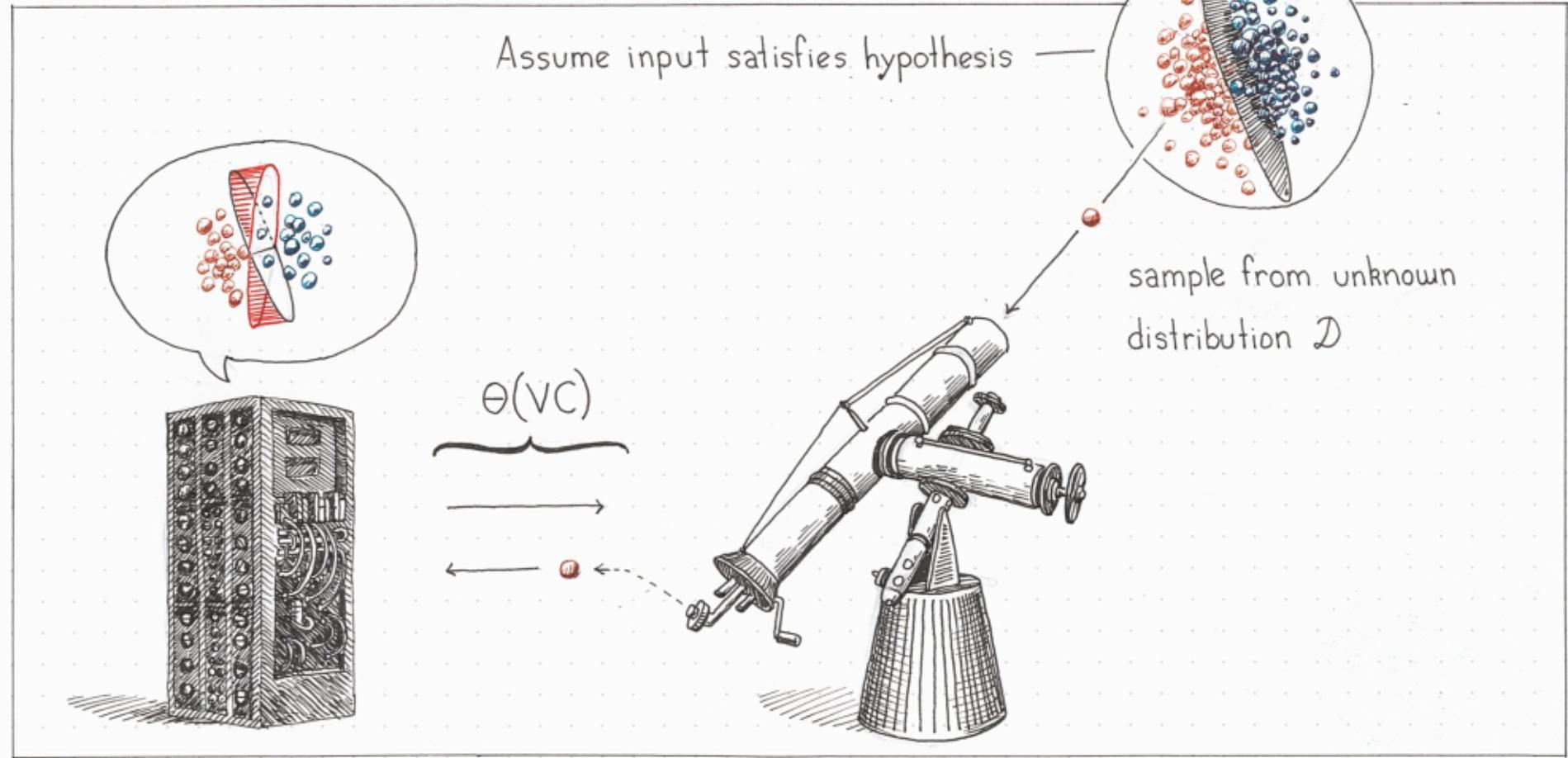
Structural explanations of the power of randomized  
algorithms

Nathan Harms (EPFL)





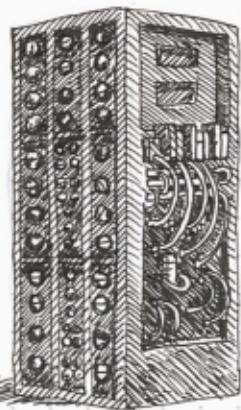
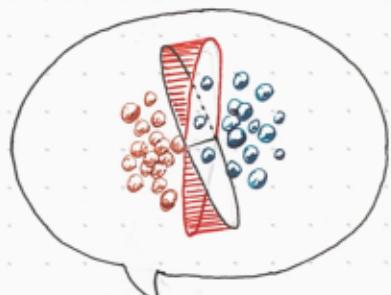
Property Testing



Learning

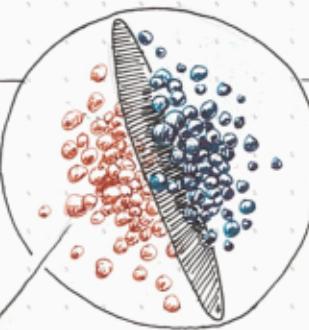
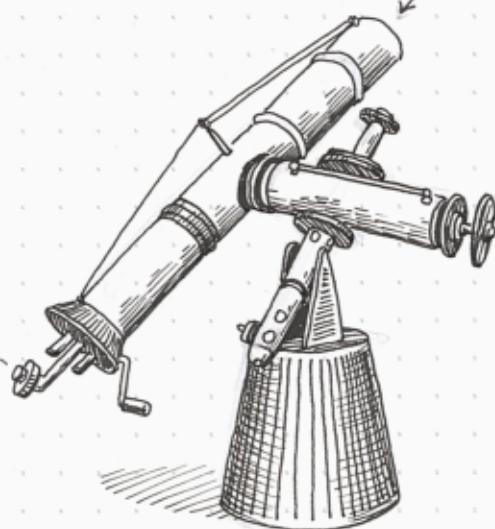
Testing vs Learning [GGR'98]: Testing requires ??? samples

Assume input satisfies hypothesis —



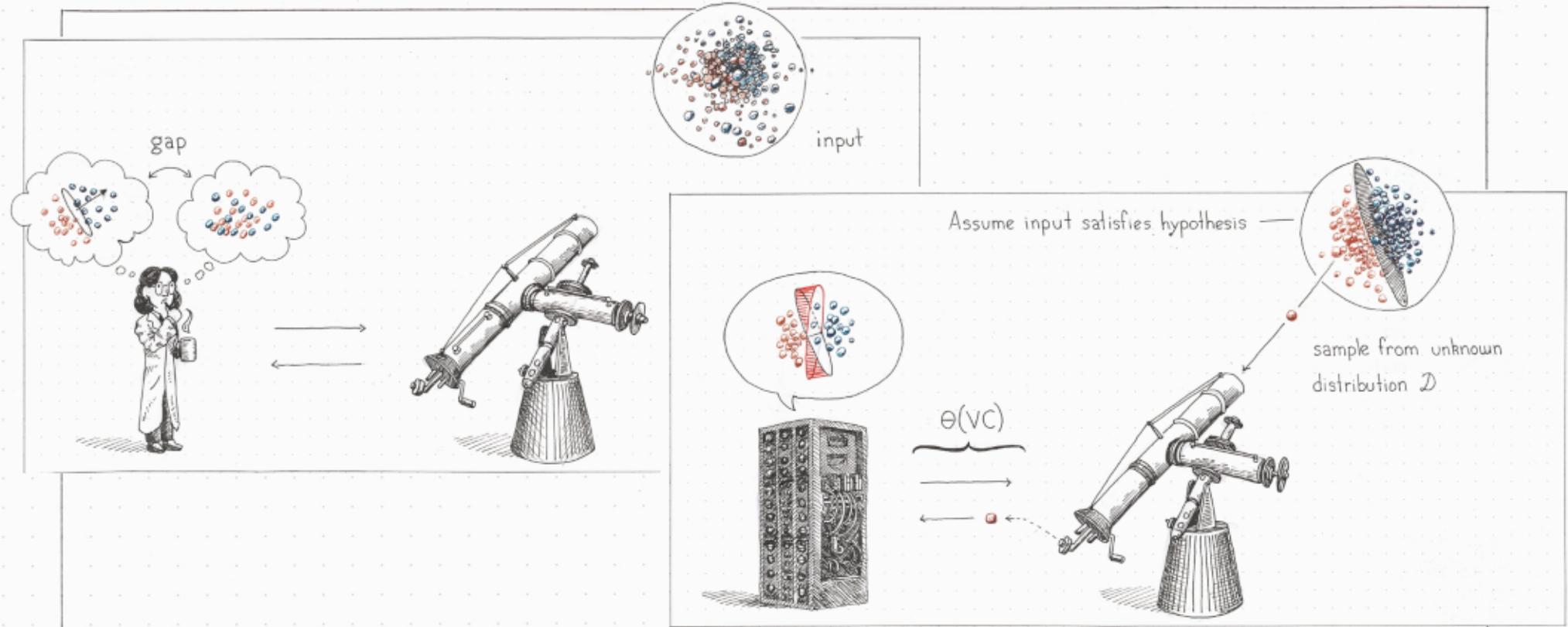
$$\theta(\text{vc})$$

A mathematical expression where the Greek letter theta is followed by the text "vc" in parentheses, with a brace underneath indicating it spans both the theta and the vc.

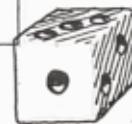


sample from unknown  
distribution  $\mathcal{D}$

Learning

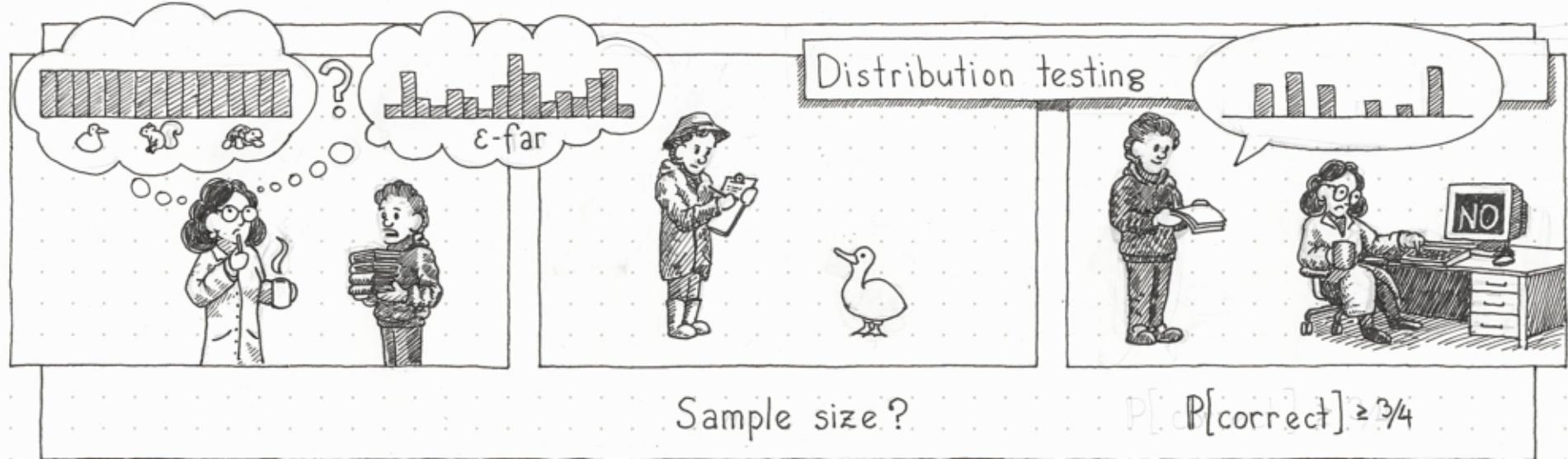


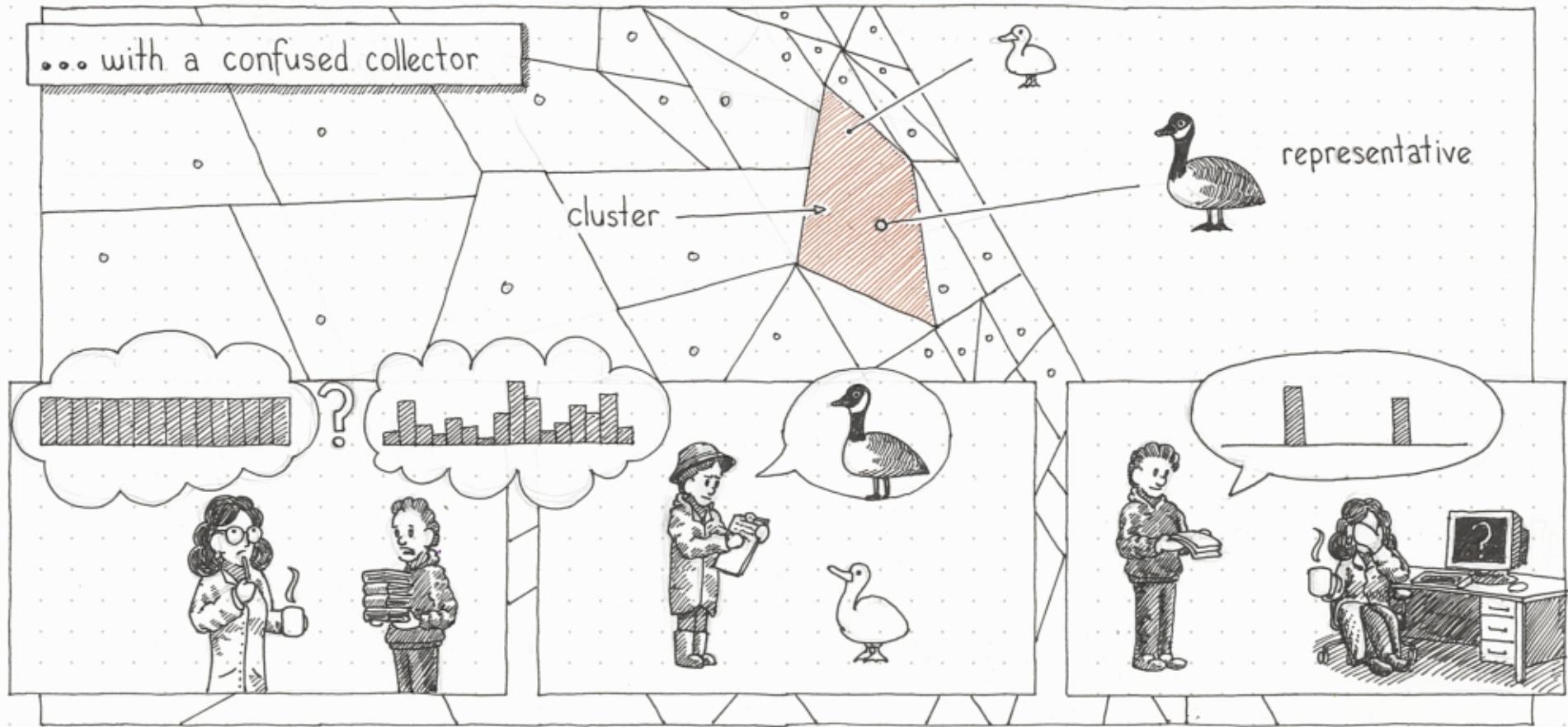
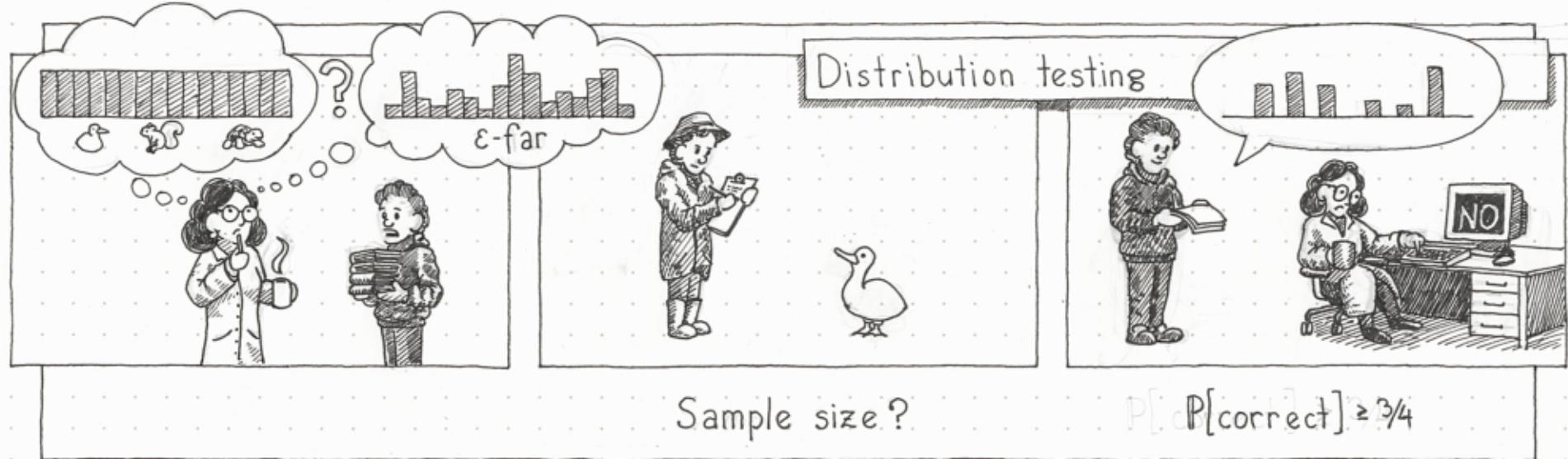
Theorem: Hypotheses with "large certificates"  $\Rightarrow \approx \text{VC}$   
 require  $\Omega(\text{VC}/\log \text{VC})$  samples to test. [BFH'21]



Sometimes tight!

Deeper connection to distribution testing?





Results: [FH'24]

Algorithms for

random clustering

adversarial clustering!

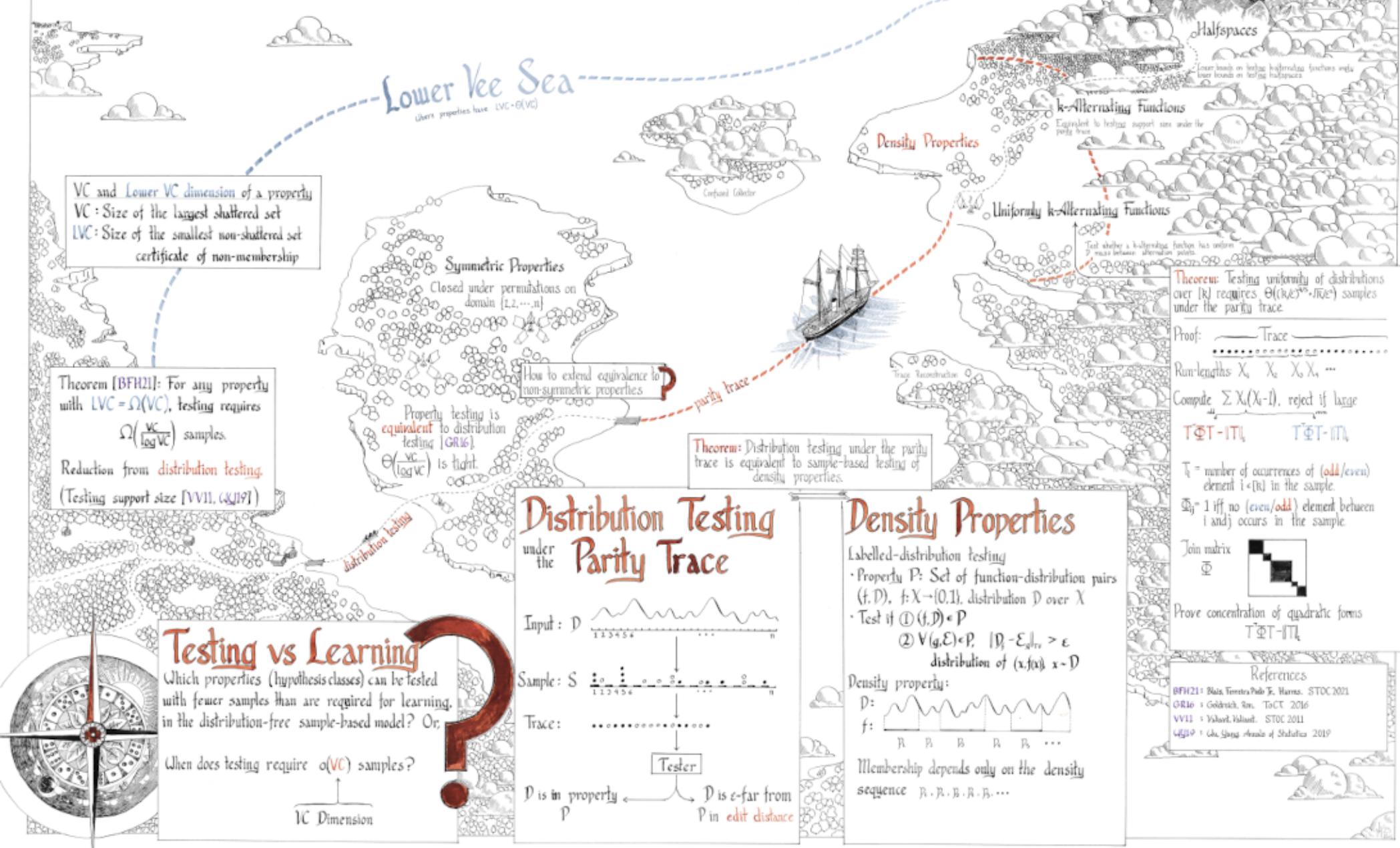
... with a confused collector

cluster

representative



# Distribution Testing & Testing vs Learning



Left out... [H, SODA'19], [HY, ICALP'22], [BBH, ITCS'24]

## TESTING CONVEX SETS

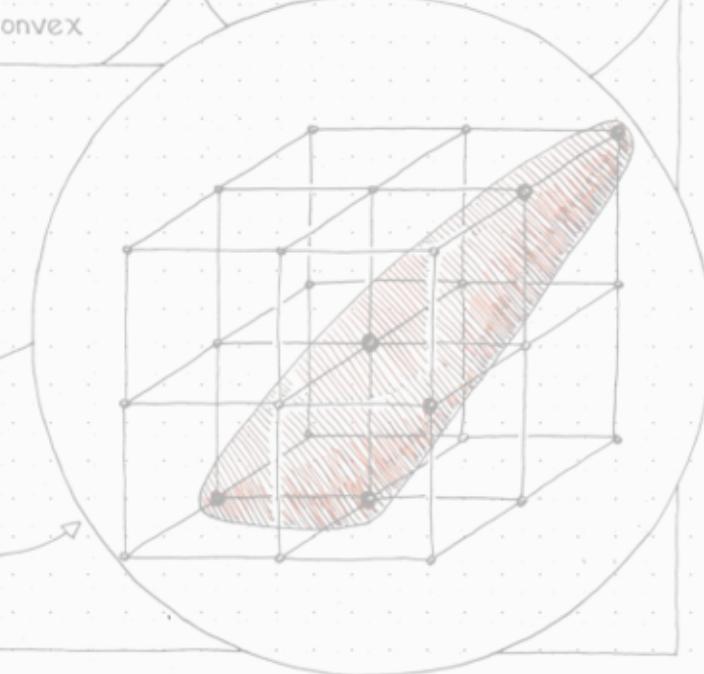
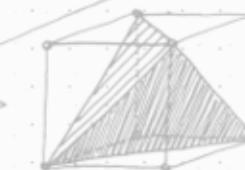


Tight bounds:

- Low dimension: [BMR'16 x 3, R'03]
- Gaussian over  $\mathbb{R}^n$ : [CFSS'17, KOS'08]  
↳ samples only

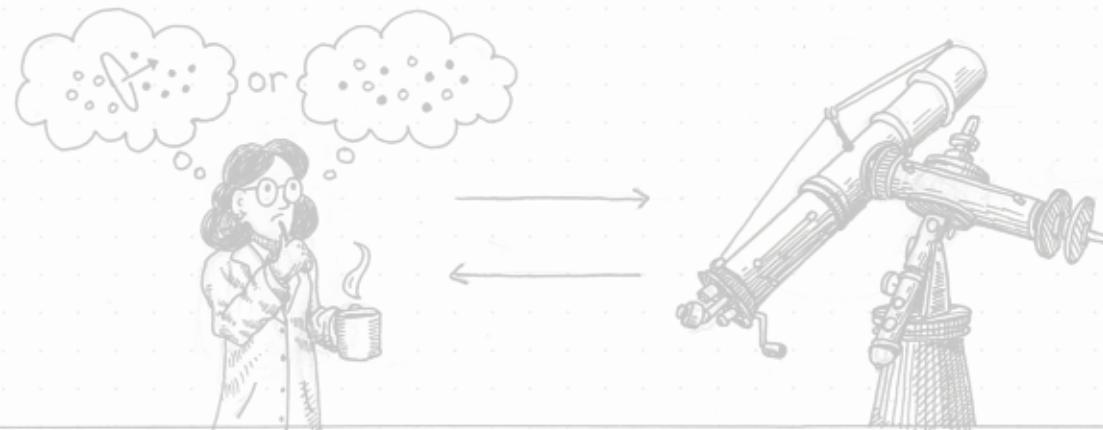
Unclear how to use queries [BB'20, RV'04]

Discrete convex sets?



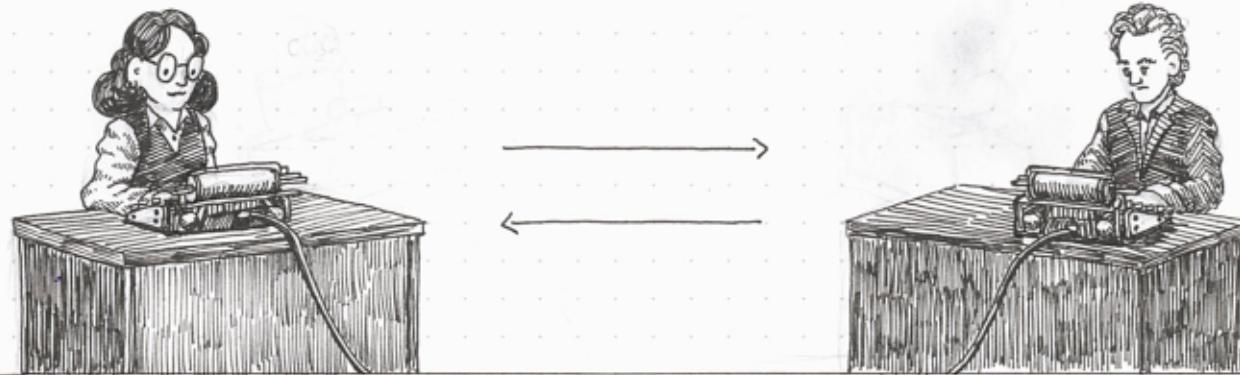
1

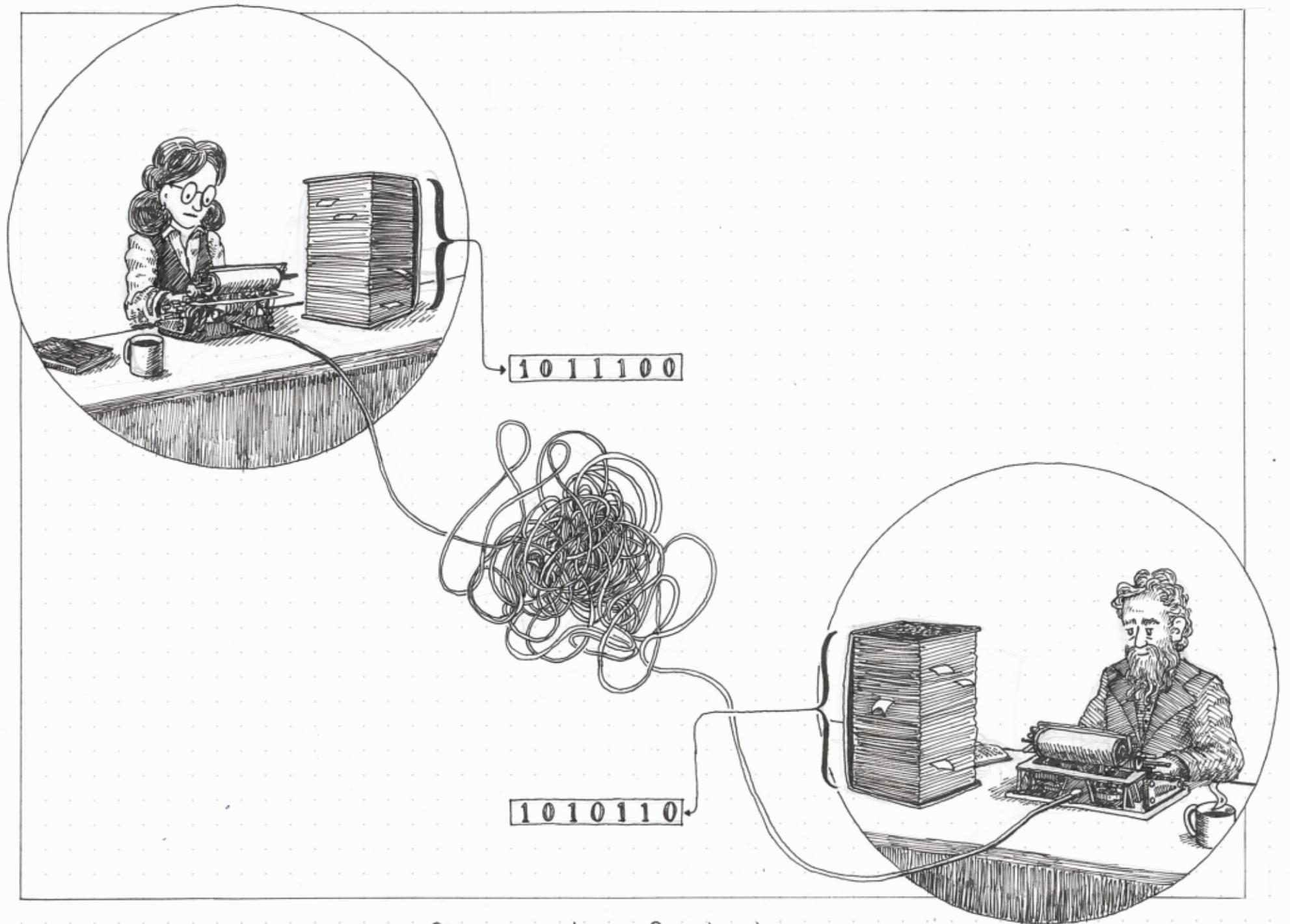
## Property Testing



2

## Communication Complexity





Communication Complexity

## Why?

### ① Power of randomness

- Most extreme case
- Most basic lower bound
- “Fine-grained” understanding
- Standard techniques fail
- Complete characterization?
- Most evident structure

### ② Connections

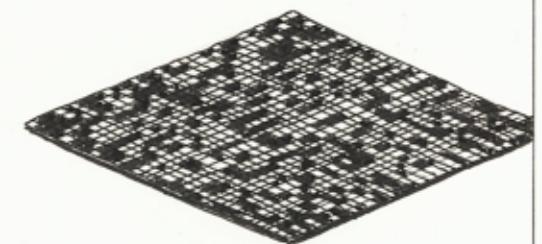
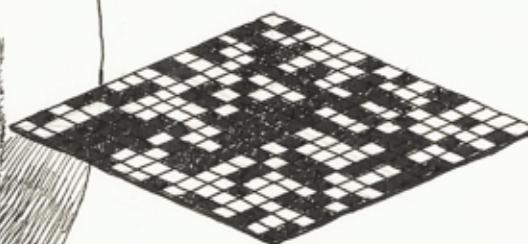
- Matrix representations
- Structural graph theory
  - New concepts
  - New techniques
- Algebra
- Learning theory

### ③ It is cool

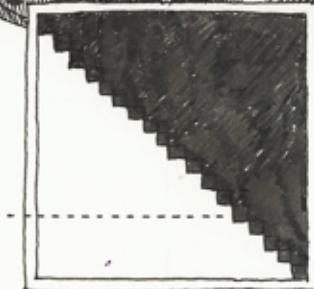
# Randomized Communication



BPP<sup>0</sup>: Constant Cost

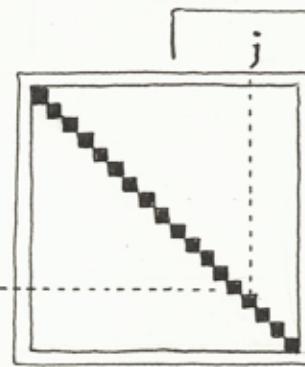


$i < j ?$

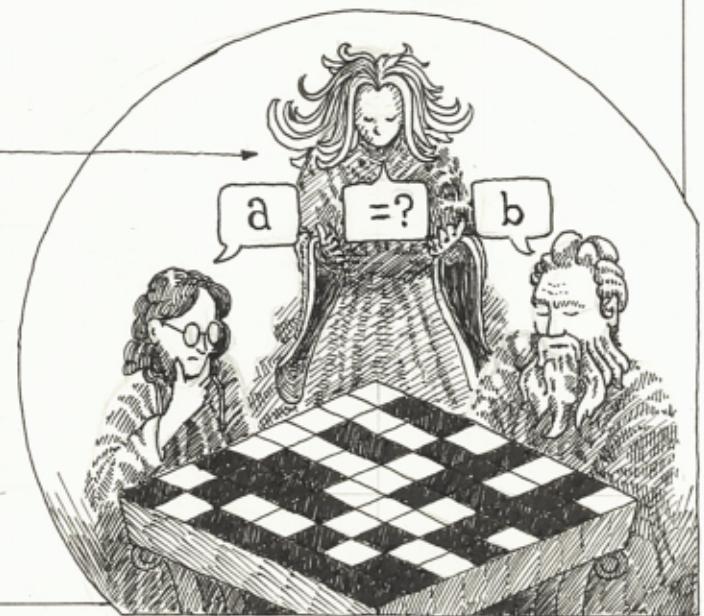


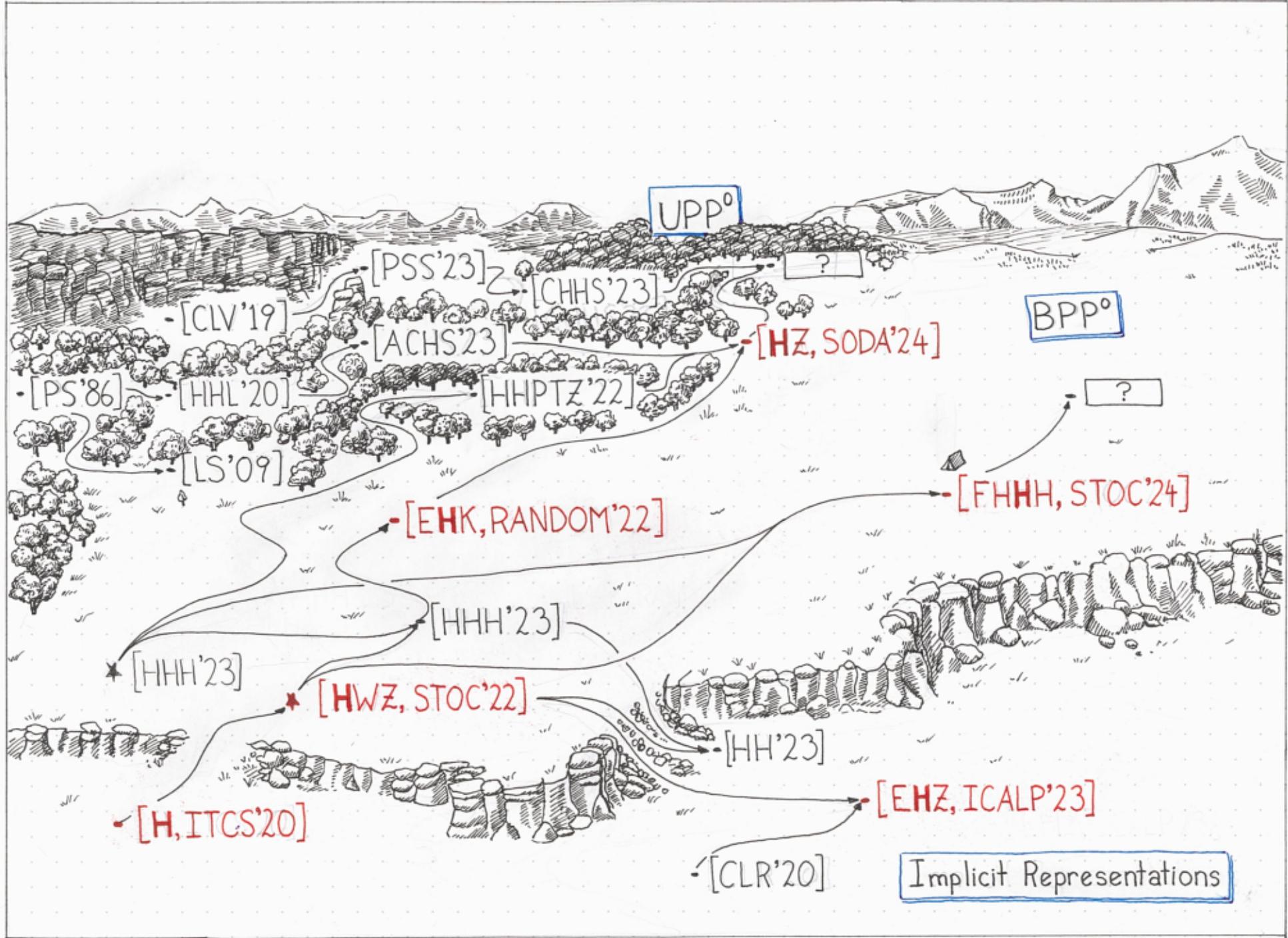
GREATER-THAN

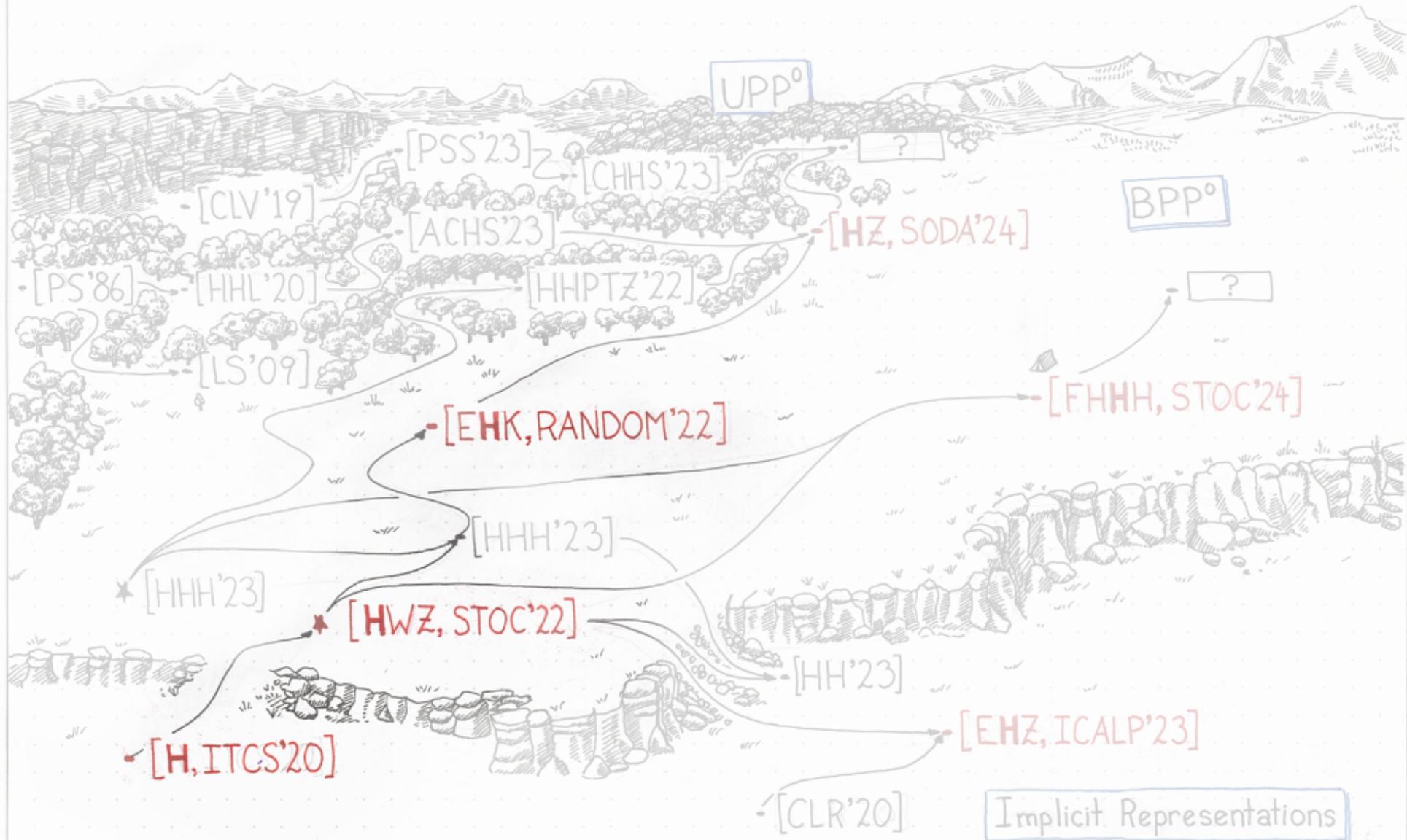
$i = j ?$

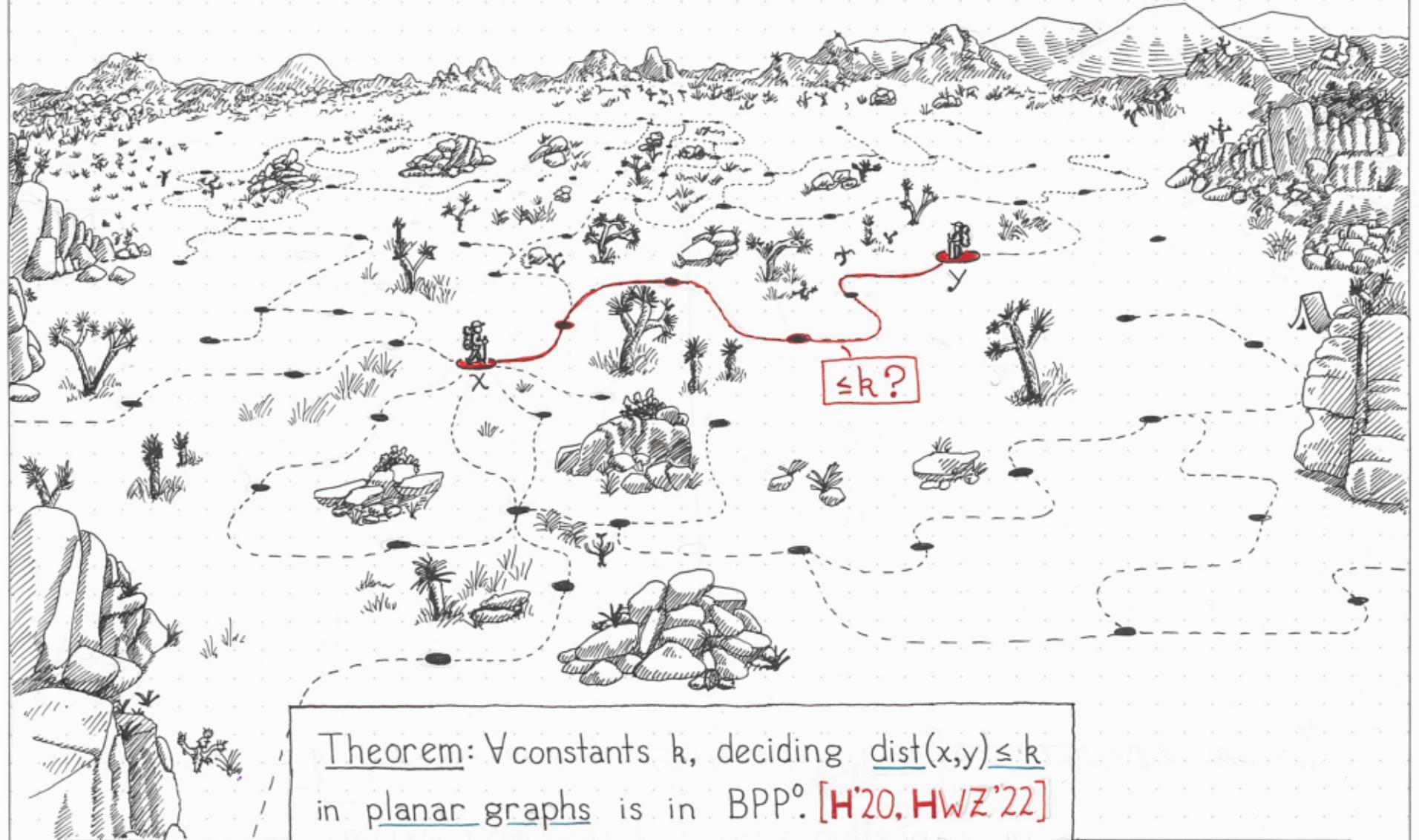


EQUALITY







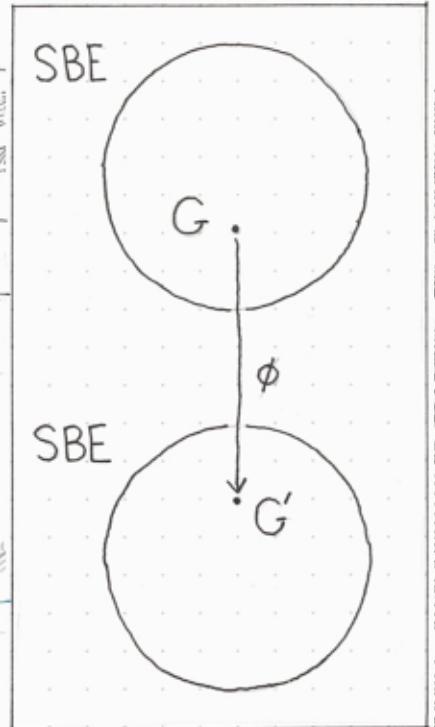
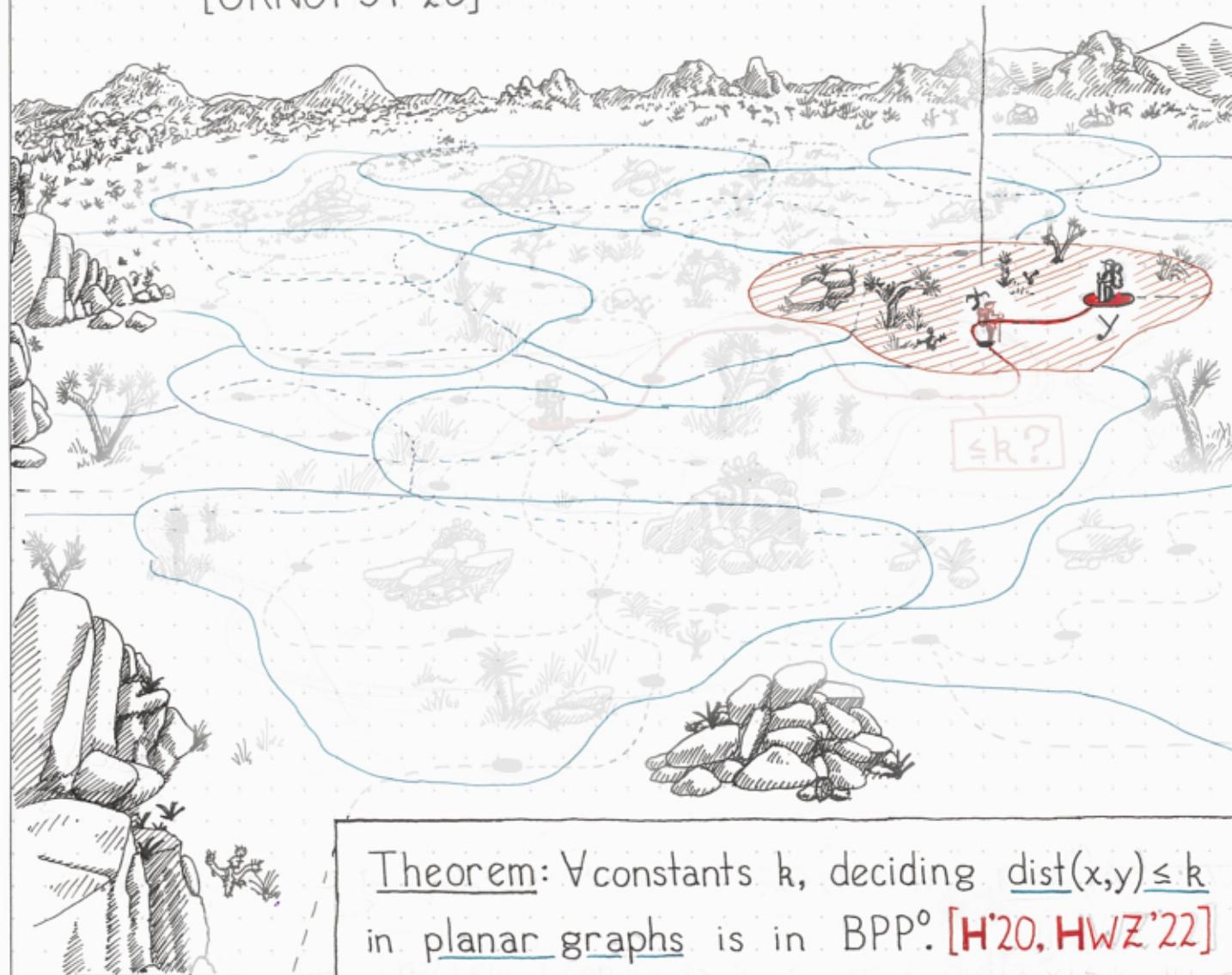


$$\phi(x,y) = \exists p_0, \dots, p_k \in V : x = p_0 \wedge y = p_k \wedge (p_0 = p_1 \vee E(p_0, p_1)) \wedge \dots \wedge (p_{k-1} = p_k \vee E(p_{k-1}, p_k))$$

"Structurally bounded expansion"

[GKNOPST'20]

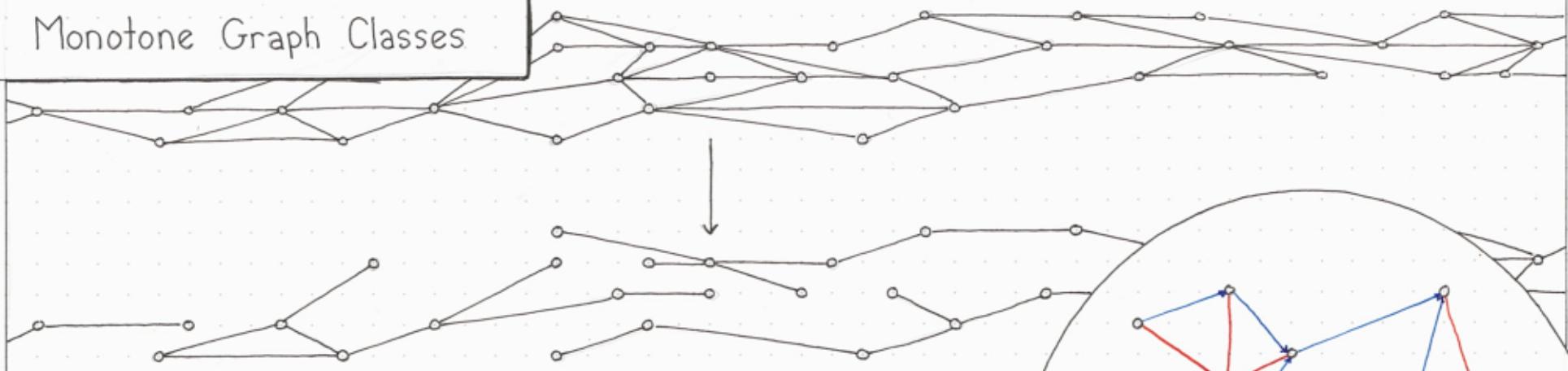
bounded shrubdepth



Theorem:  $\forall$  constants  $k$ , deciding  $\text{dist}(x,y) \leq k$  in planar graphs is in  $\text{BPP}^{\circ}$ . [H'20, HWZ'22]

$$\phi(x,y) = \exists p_0, \dots, p_k \in V : x = p_0 \wedge y = p_k \wedge (p_0 = p_1 \vee E(p_0, p_1)) \wedge \dots \wedge (p_{k-1} = p_k \vee E(p_{k-1}, p_k))$$

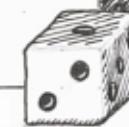
## Monotone Graph Classes



Theorem: If  $\mathcal{G}$  is a monotone graph class:

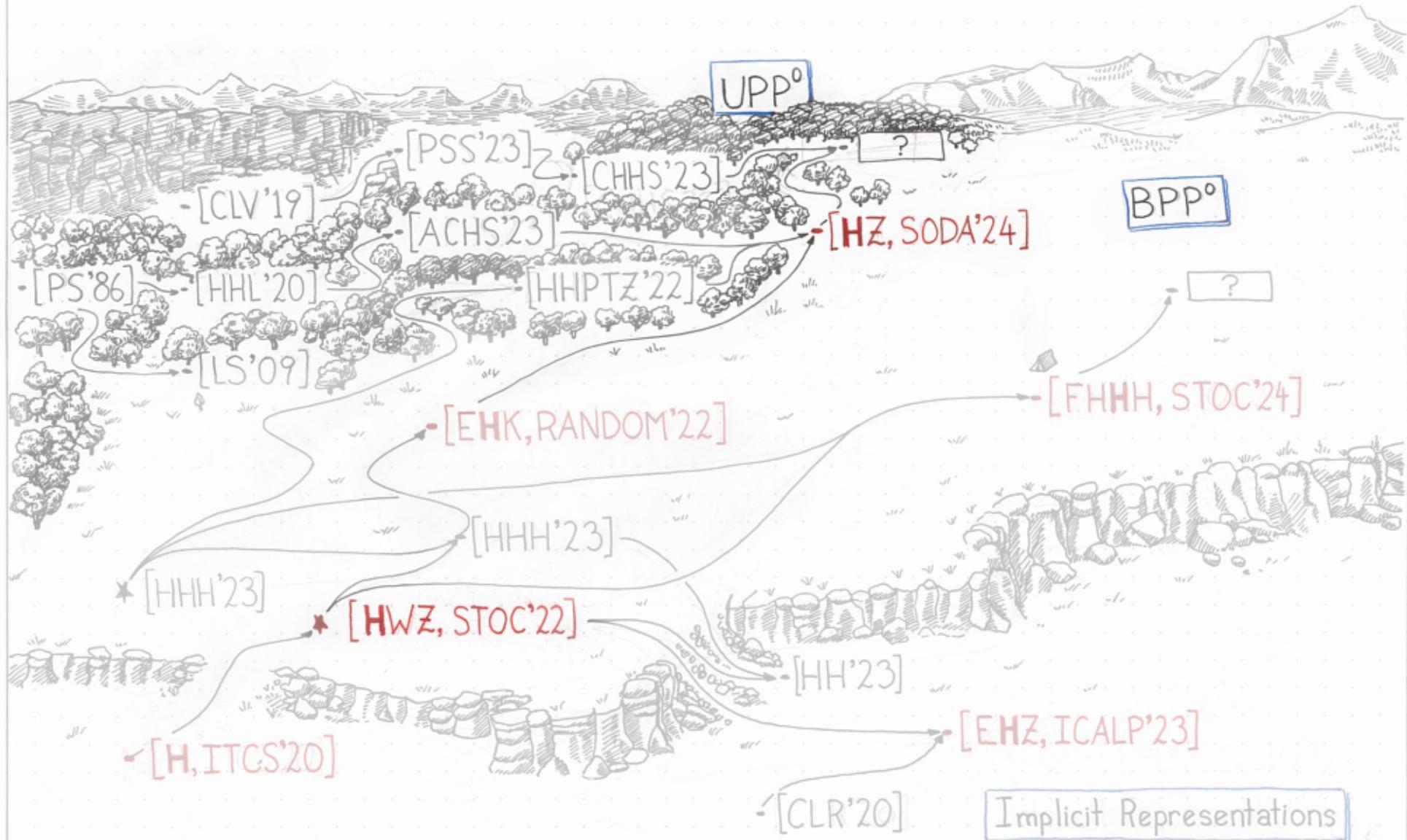
- ① Adjacency is in  $BPP^0$   
 $\Leftrightarrow \mathcal{G}$  has bounded arboricity.
- ② Distance  $k$  is in  $BPP^0$ .  $\forall$  constants  $k$   
 $\Leftrightarrow \mathcal{G}$  has bounded expansion.

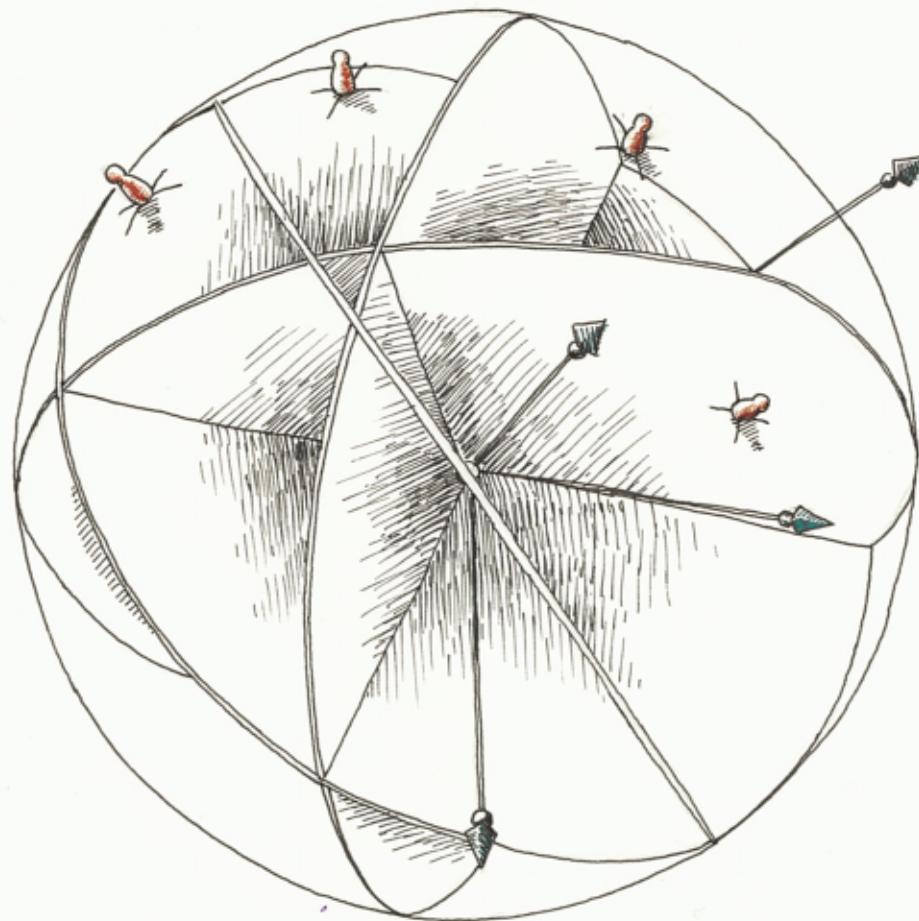
And the EQUALITY oracle suffices. [EHK'22]



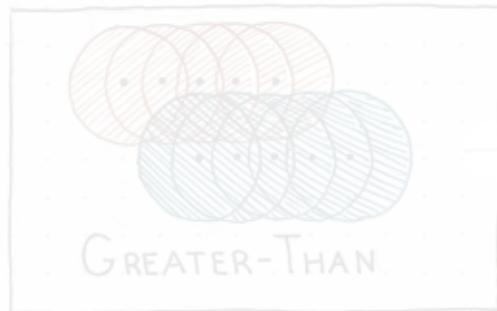
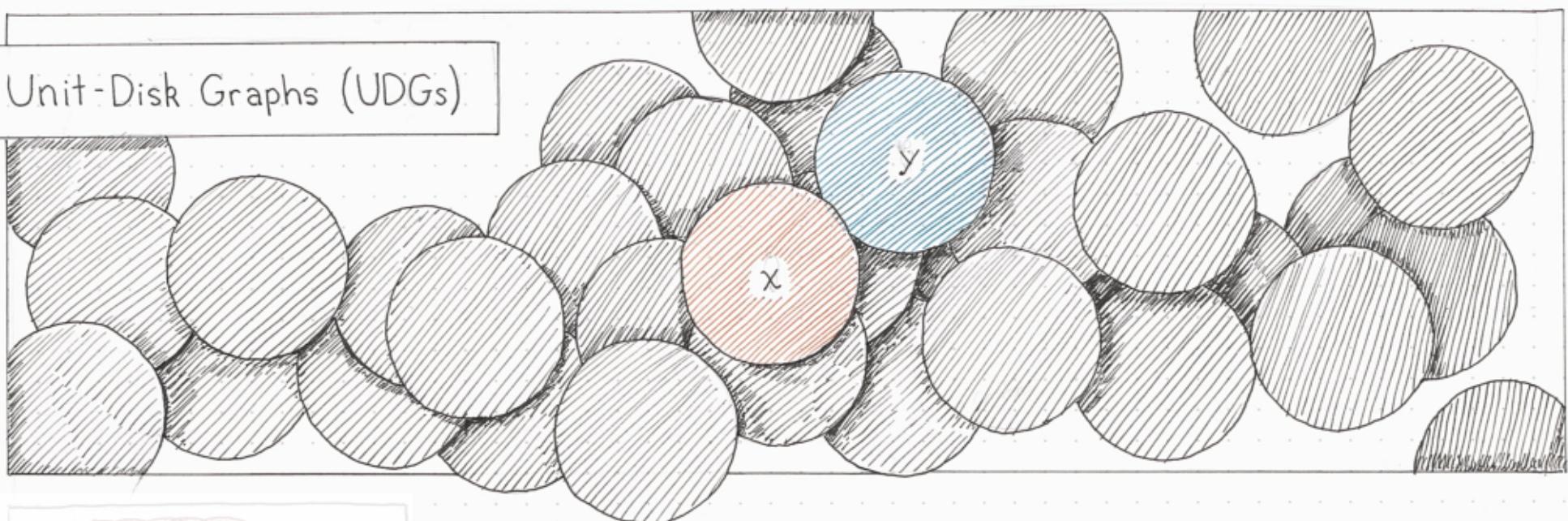
contraction has average degree  $f(r)$ . [e.g. NO'12]





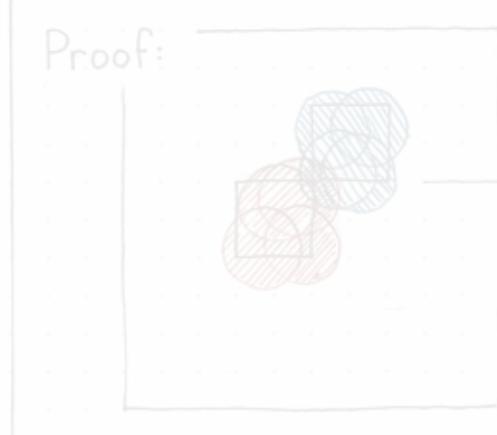


## Unit-Disk Graphs (UDGs)

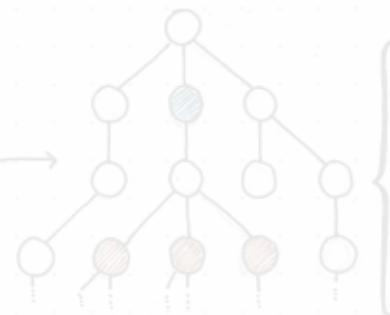


remove

Theorem: UDGs have constant-cost protocols for adjacency iff. they are stable. [HZ'24]



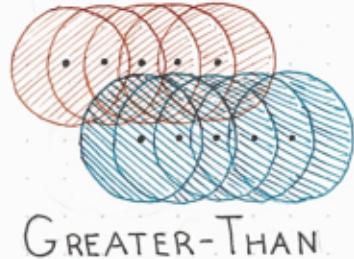
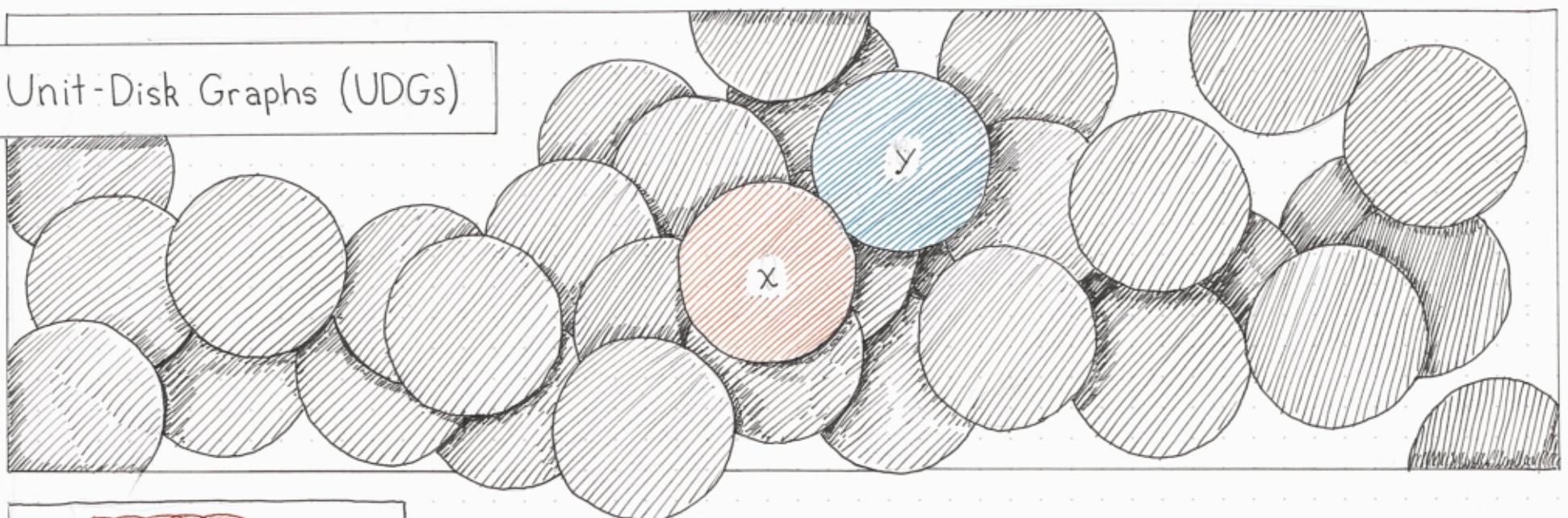
[OPS'22]: "Gyárfás decomposition"



Ramsey's theorem  
↓  
Recurse on subgraph,  
shrink GREATER-THAN



## Unit-Disk Graphs (UDGs)



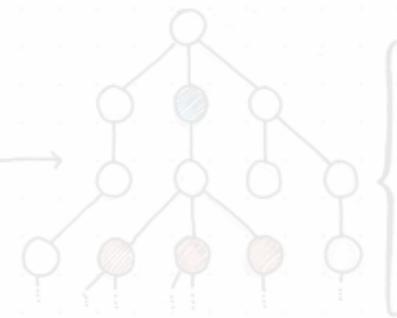
remove

Theorem: UDGs have constant-cost protocols  
for adjacency iff. they are stable. [HZ'24]

Proof:



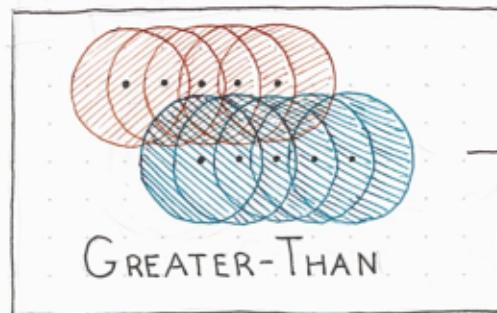
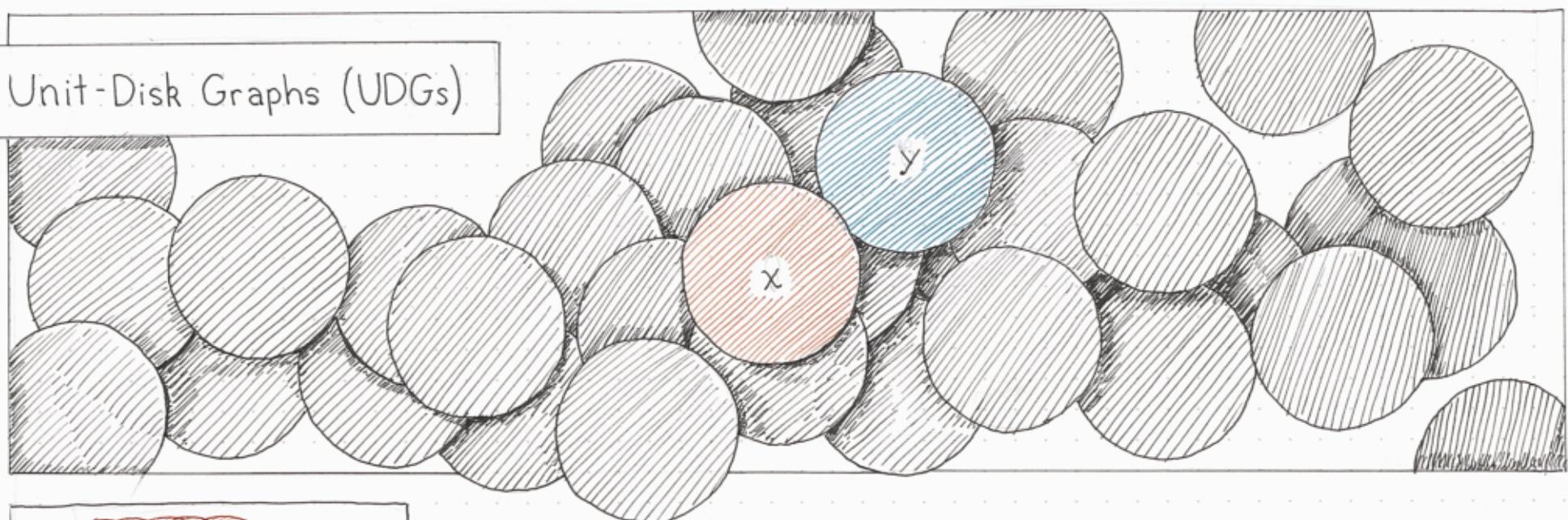
- free



Ramsey's theorem  
↓  
Recurse on subgraph,  
shrink GREATER-THAN

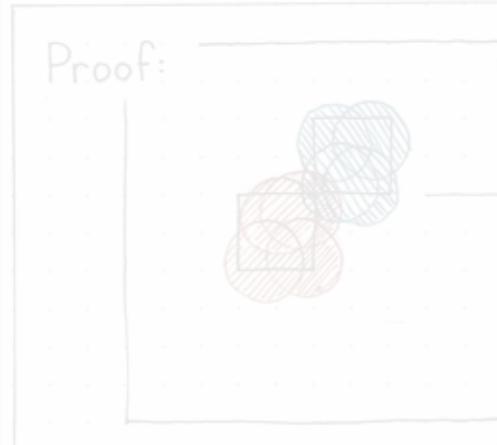
[OPS'22]: "Gyárfás decomposition"

## Unit-Disk Graphs (UDGs)

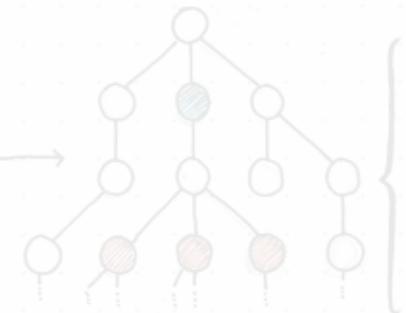


remove

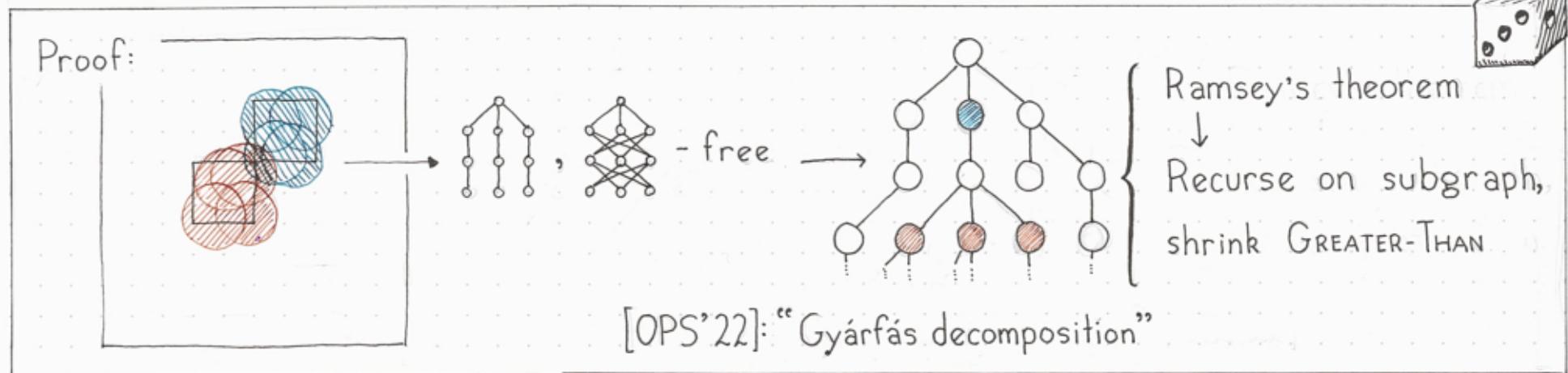
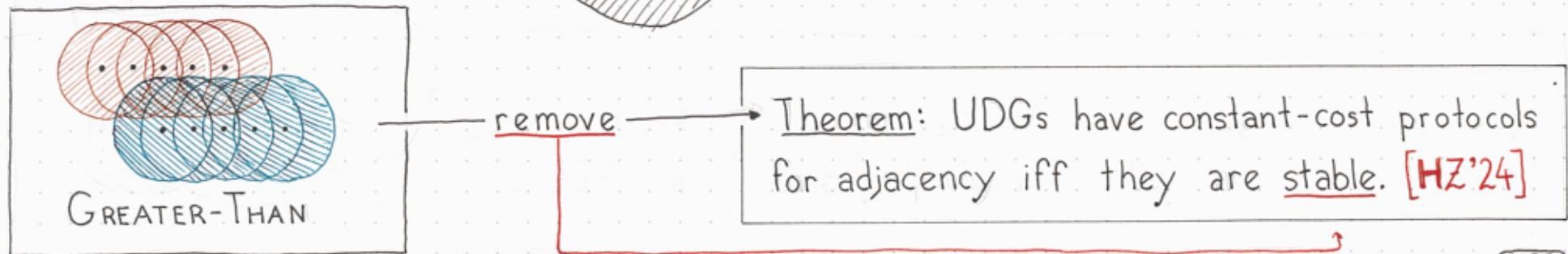
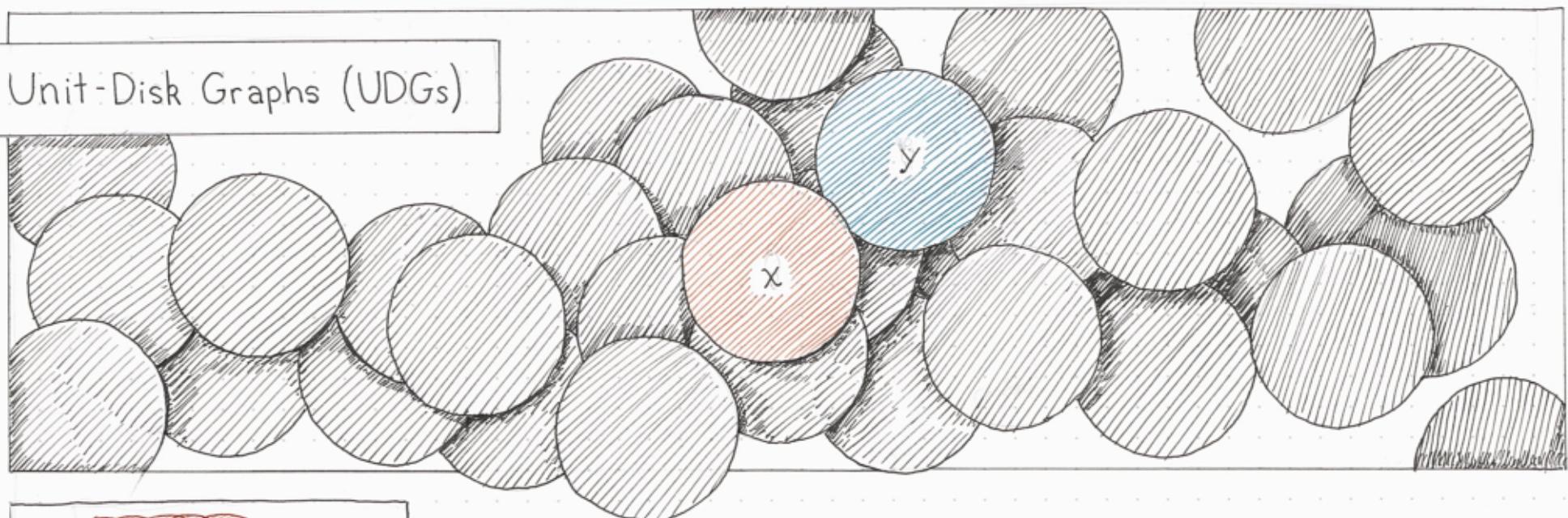
Theorem: UDGs have constant-cost protocols for adjacency iff they are stable. [HZ'24]



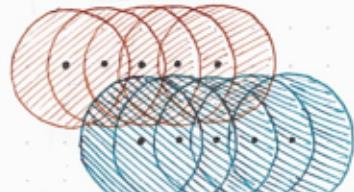
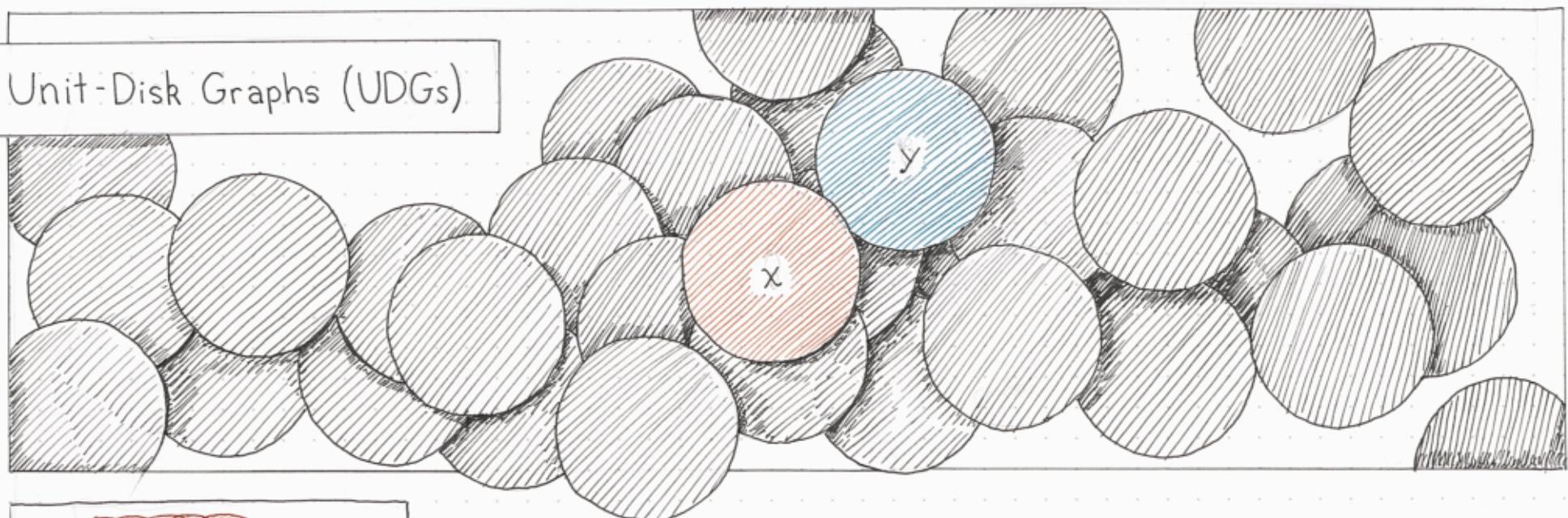
[OPS'22]: "Gyárfás decomposition"



## Unit-Disk Graphs (UDGs)



## Unit-Disk Graphs (UDGs)



GREATER-THAN

remove

Theorem: UDGs have constant-cost protocols for adjacency iff they are stable. [HZ'24]

Works also for:

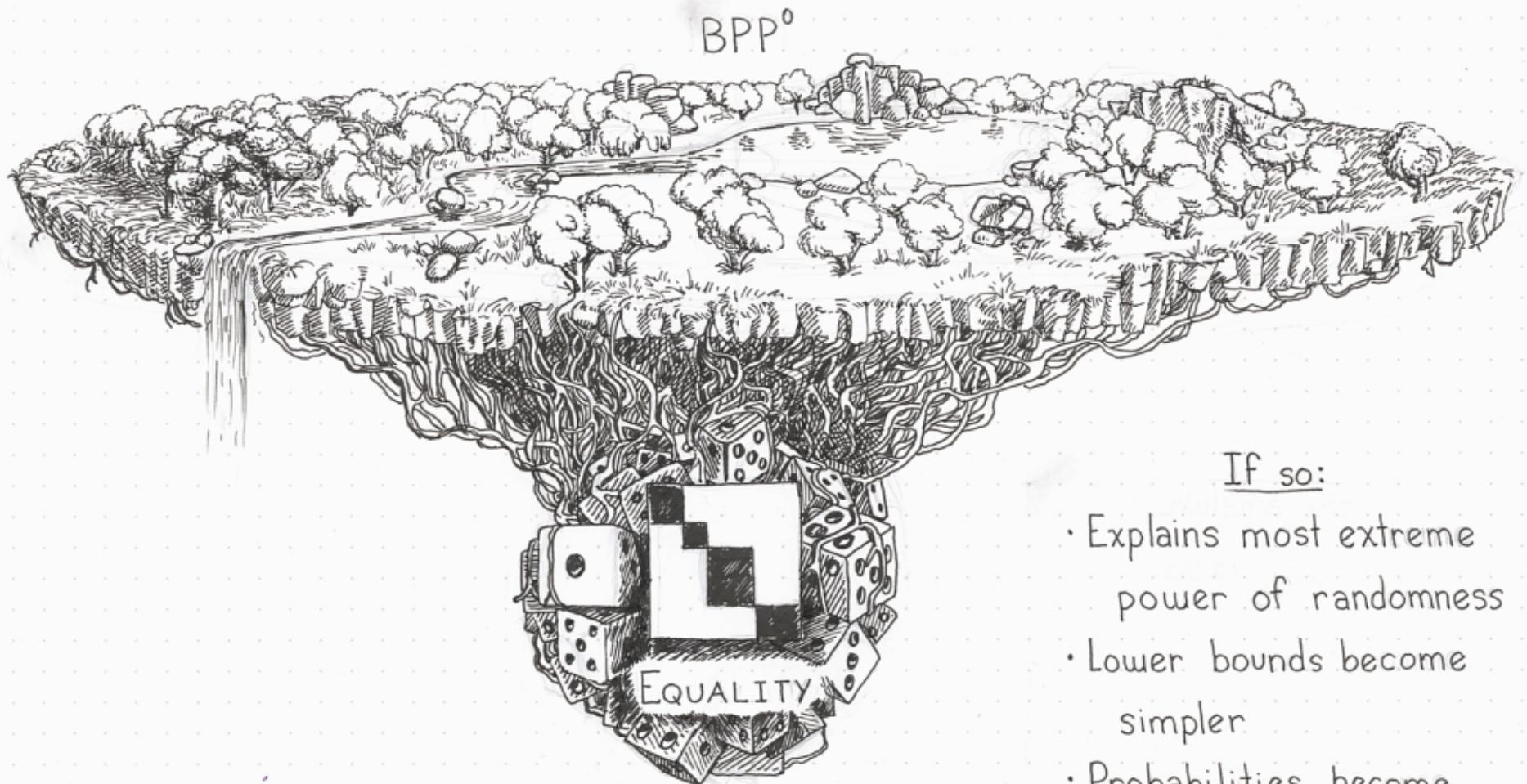
- Dimension 3 halfspaces
- Interval graphs
- Permutation graphs
- etc...

}  $\in \text{UPP}^0$



Only uses EQUALITY!

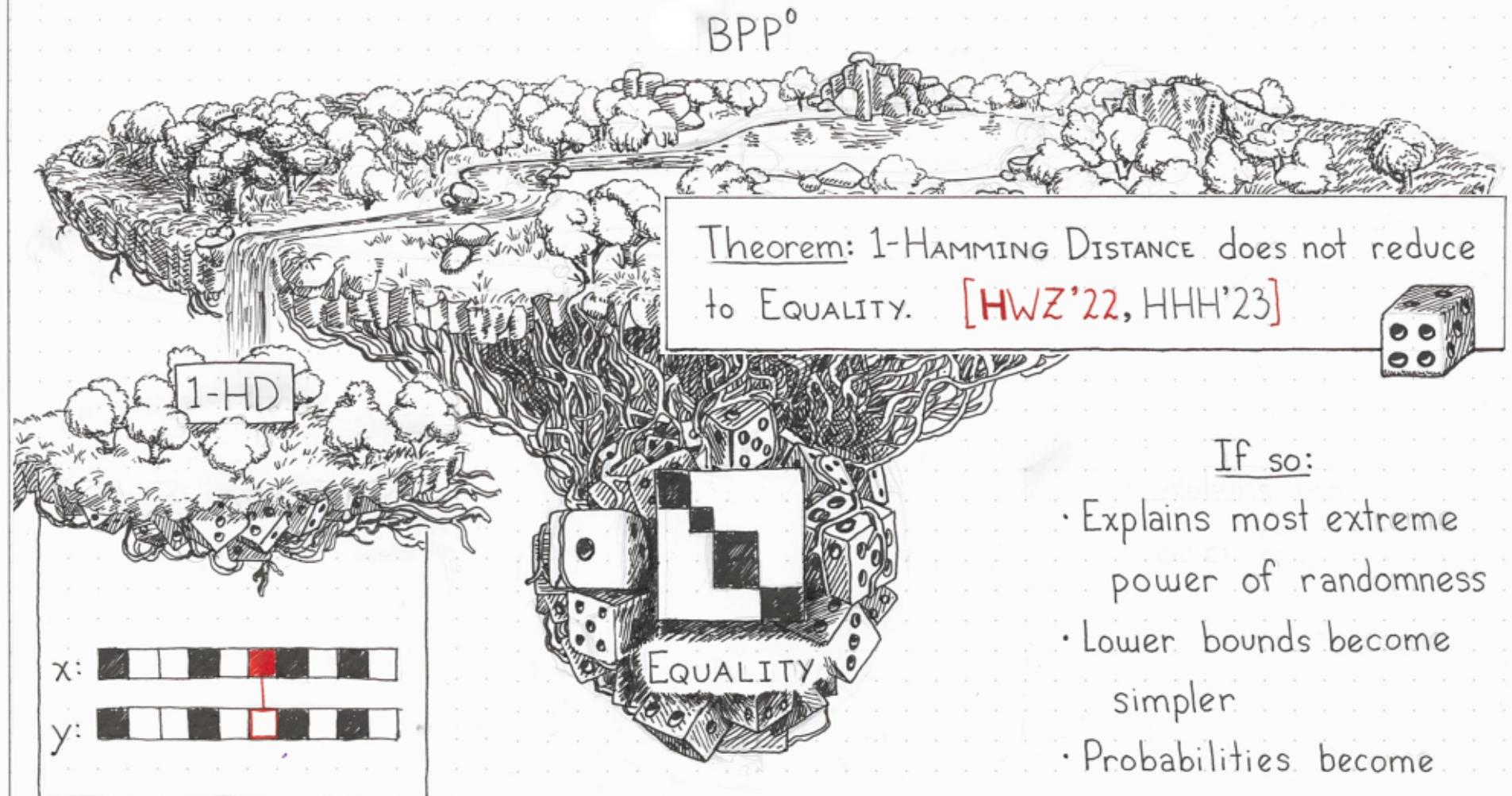
Is EQUALITY “complete” for constant-cost communication?



If so:

- Explains most extreme power of randomness
- Lower bounds become simpler
- Probabilities become unnecessary

Is EQUALITY “complete” for constant-cost communication?



If so:

- Explains most extreme power of randomness
- Lower bounds become simpler
- Probabilities become unnecessary

... Then, is 1-HAMMING DISTANCE complete for  $\text{BPP}^0$ ?

All previous techniques fail ...

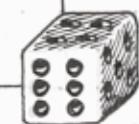
[HWZ'22, HZ'24, HHH'23, CLV'19, CHHS'23, PSS'23]

But... Theorem: 2-HAMMING DISTANCE does not  
reduce to 1-HAMMING DISTANCE [FHHH'24]

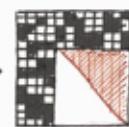
Theorem: There is no complete problem for  $\text{BPP}^0$ .



Theorem: The  $k$ -HAMMING DISTANCE problems form an infinite hierarchy within  $\text{BPP}^0$ . [FHHH'24]

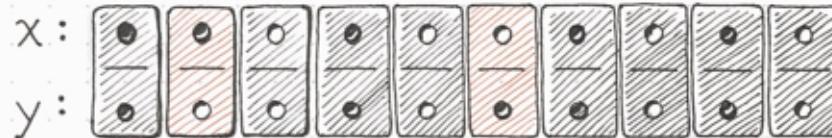


① Every oracle in  $\text{BPP}^0$  is stable.  $\rightarrow \{ \text{O}(1)$



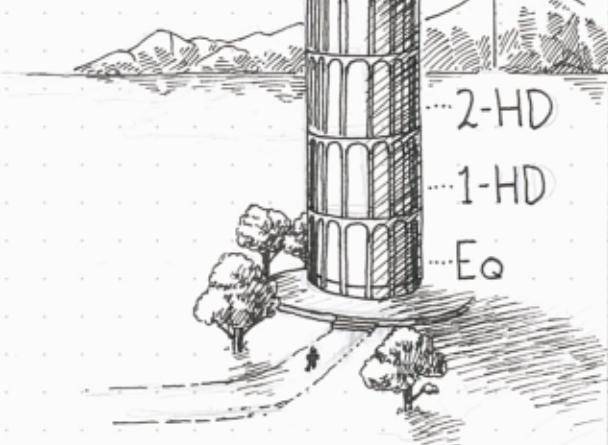
$k\text{-HD}$

②  $k$ -HAMMING DISTANCE is permutation-invariant:



If the oracle queries are not permutation-invariant:

$\rightarrow$  Hypergraph Ramsey theorem  $\Rightarrow$  stability is violated.



③ One query computes distance  $k$  vs.  $k+2$  for all weight  $n/2$  strings. Stability is violated.

## Future work

### Testing vs. learning

distribution testing

trace reconstruction\*

confused collector

parity trace

k-alternating — ? → halfspaces

multiple collectors

random clusters (trees?)

small certificates?

VC-like theory

### Constant-cost communication

$BPP^0 \neq UPP^0$

model theory?

dimension 4

$UPP^0 \cap BPP^0 = EQ$

hereditary problems

other types

stability

complete characterization?

finitely-defined problems

rectangle size

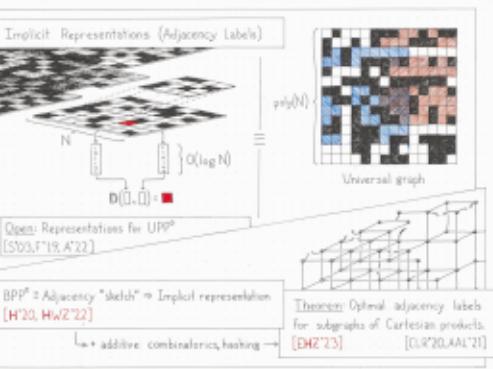
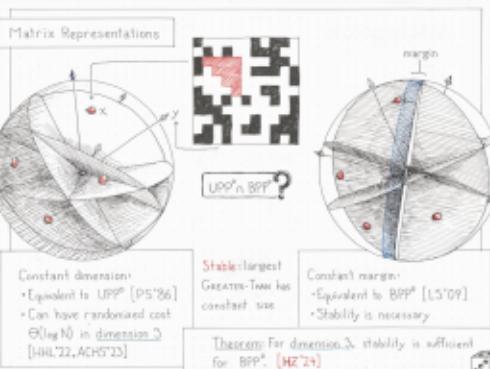
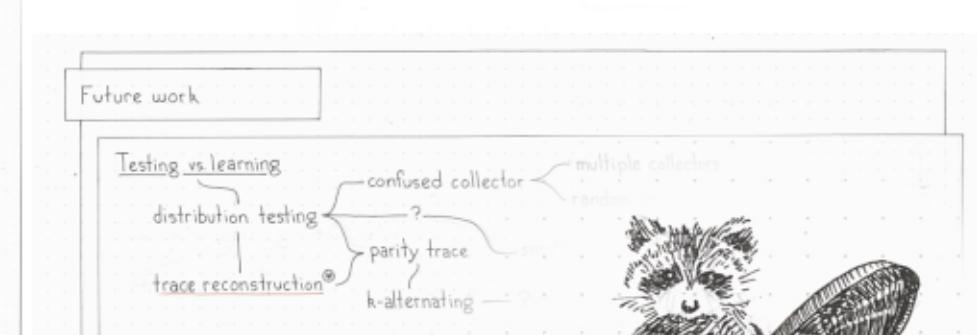
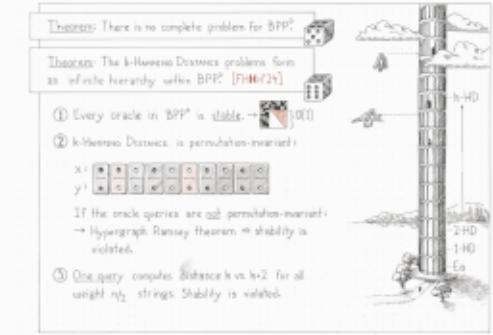
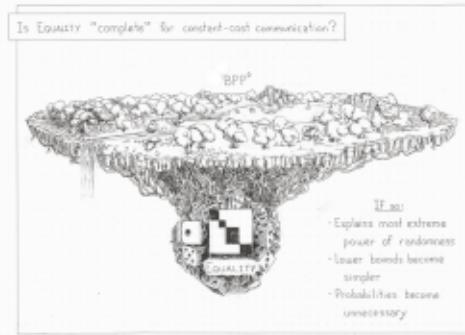
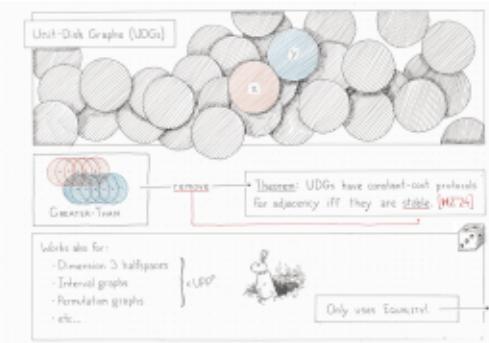
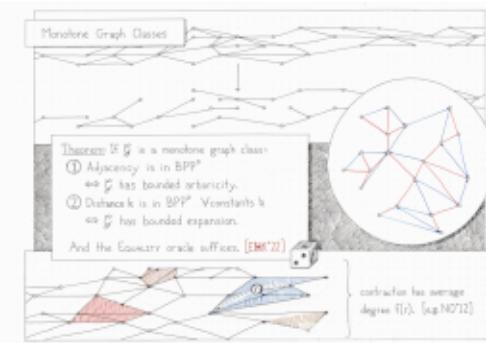
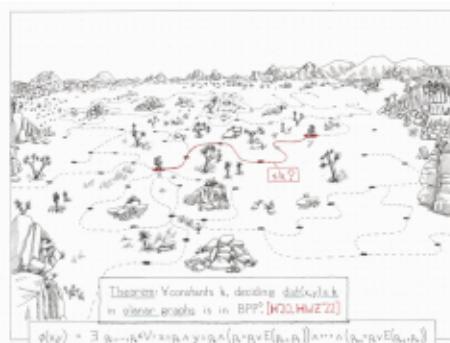
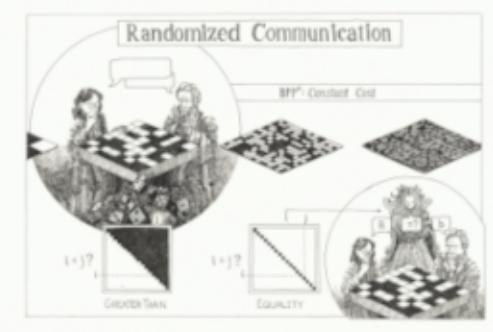
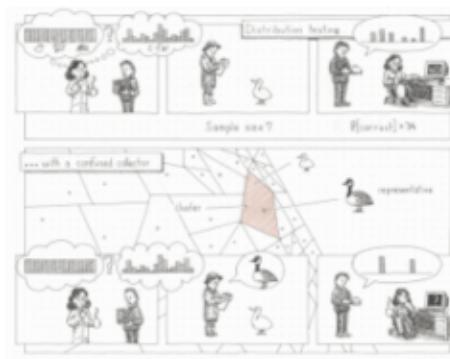
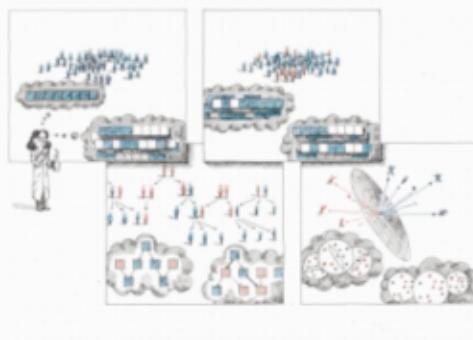
insight into higher complexity?

1-vs.-2-sided error

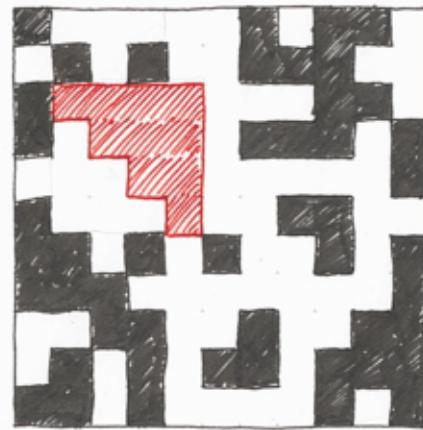
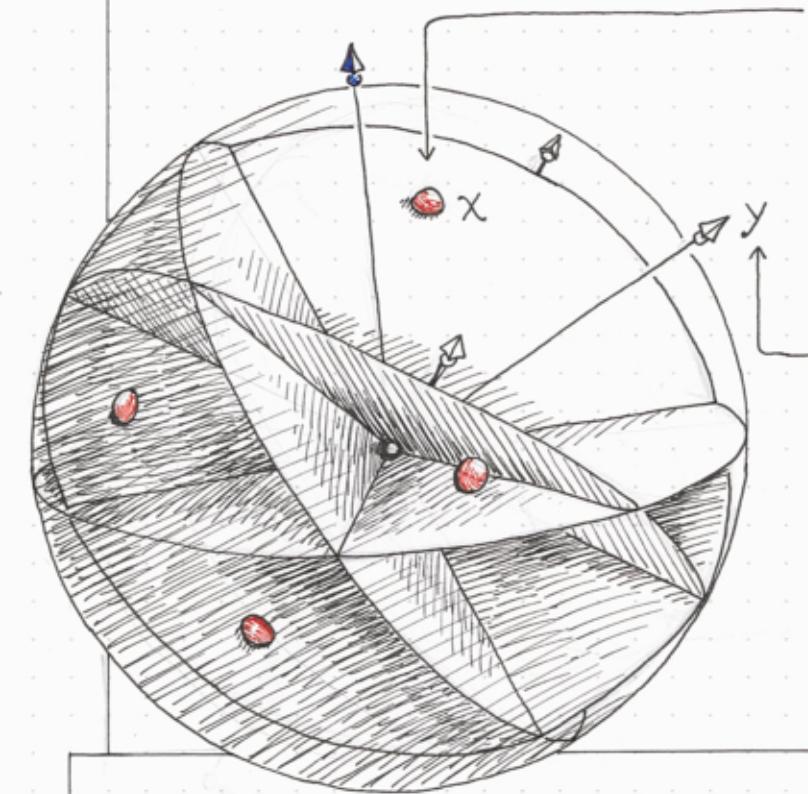
gap-hamming

randomized "log-rank"?

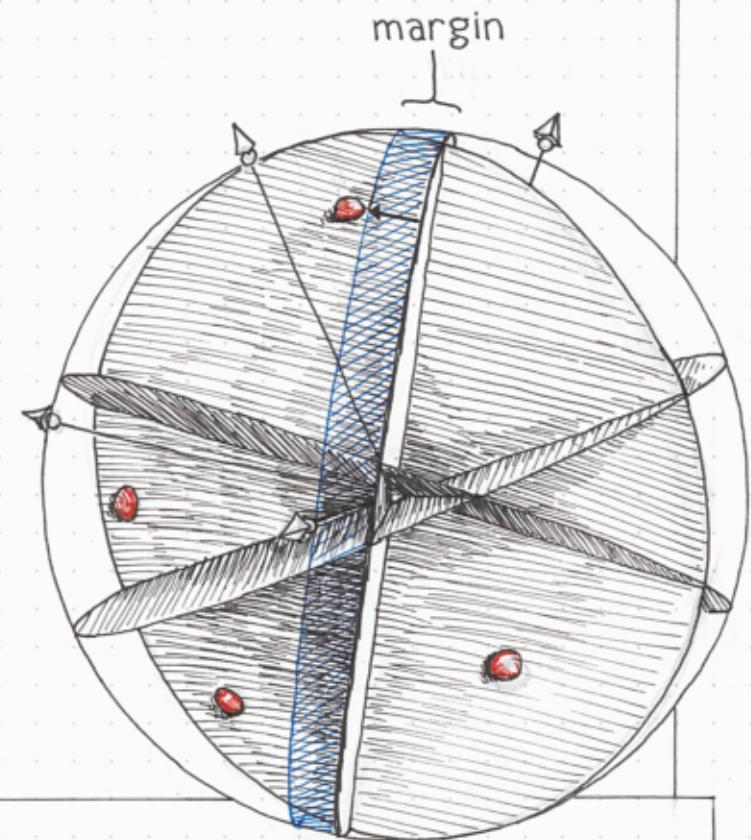
lower bounds for implicit representations



## Matrix Representations



UPP<sup>0</sup> n BPP<sup>0</sup> ?



Constant dimension:

- Equivalent to UPP<sup>0</sup> [PS'86]
- Can have randomized cost  $\Theta(\log N)$  in dimension 3 [HHL'22, ACHS'23]

**Stable**: largest  
GREATER-THAN has  
constant size

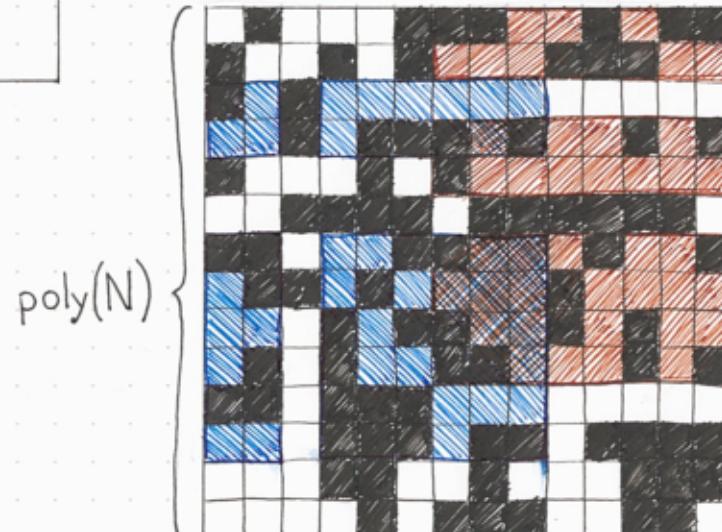
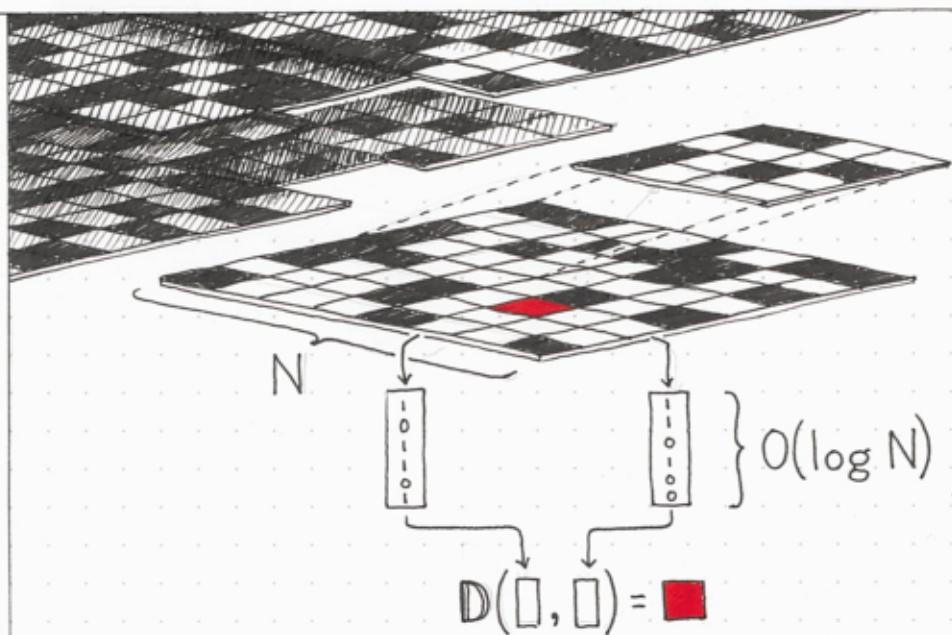
Constant margin:

- Equivalent to BPP<sup>0</sup> [LS'09]
- Stability is necessary

Theorem: For dimension 3, stability is sufficient  
for BPP<sup>0</sup>. [HZ'24]



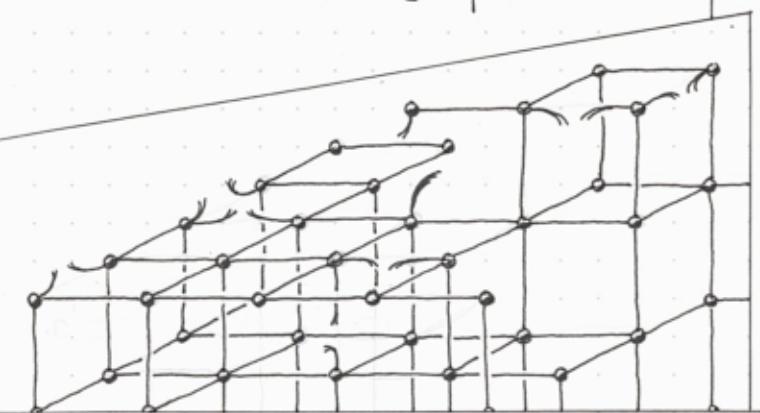
## Implicit Representations (Adjacency Labels)



Open: Representations for  $\text{UPP}^0$   
[S'03, F'19, A'22]

$\text{BPP}^0 \equiv \text{Adjacency "sketch"} \Rightarrow \text{Implicit representation}$   
[H'20, HWZ'22]

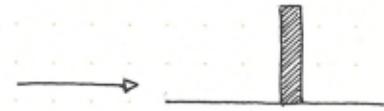
↳ + additive combinatorics, hashing →



Theorem: Optimal adjacency labels  
for subgraphs of Cartesian products.  
[EHZ'23] → [CLR'20, AAL'21]

# PART I: ADVERSARIAL CLUSTERS

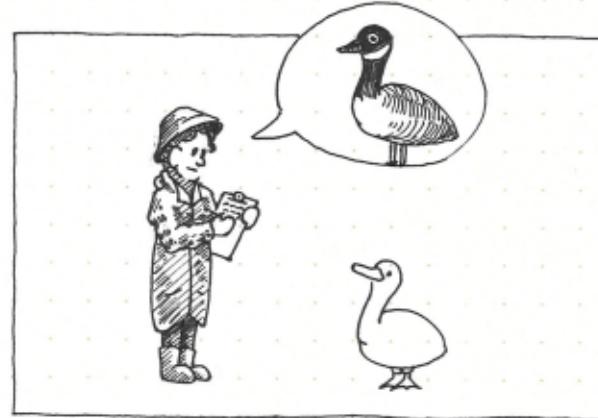
Impossible!



1.

Restrict clustering to  $\mathcal{U}$

Replace TV with  
earth-mover (EMD)  
Domain  $\rightarrow$  metric space



3.

Queries

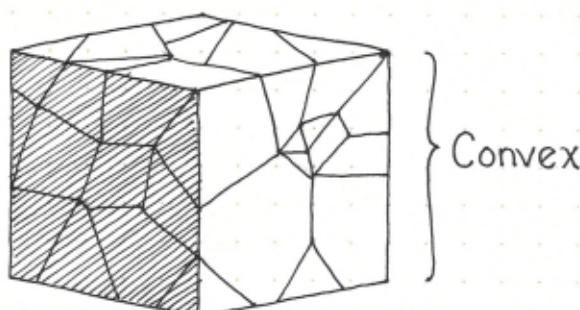


4.

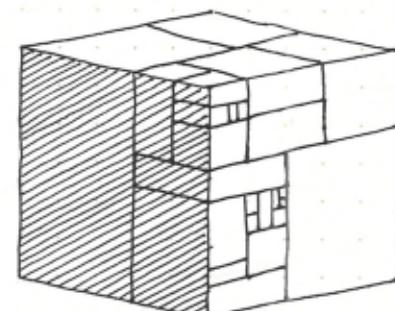
CLUSTER-REJECT

$\mathcal{U}$ : Universe

E.g.



$G \subseteq \mathcal{U}$ : "Good" clusterings



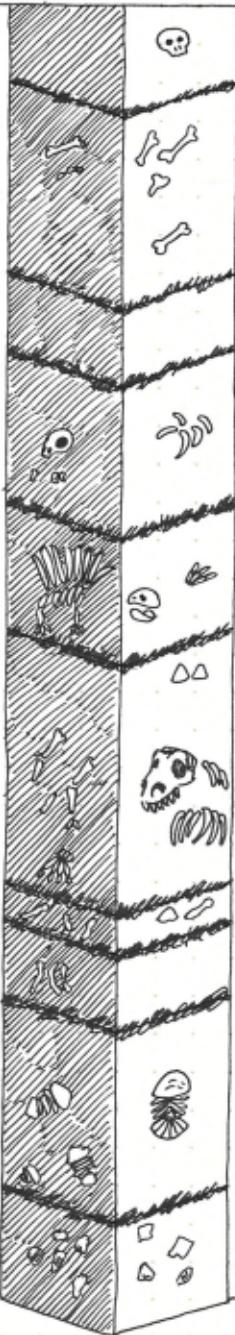
not allowed!

depends on distribution

High-probability of low diameter  
boxes, decision trees

RESULT: learning cells is not  
necessary

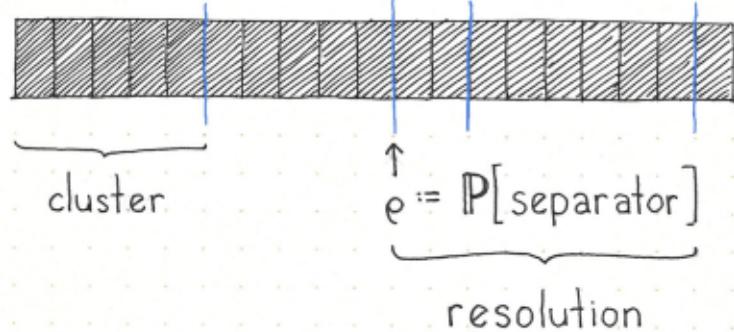
# PART II: RANDOM CLUSTERS



Motivations:

- Environmental randomness
- Randomized classifier training

We study: Testing uniformity,



Standard:

X: histogram



Analyze  $X^T I X - \|X\|_1$



Naïve (known clusters):

$$O\left(\frac{\sqrt{n}}{\rho^{3/2} \varepsilon^2}\right)$$

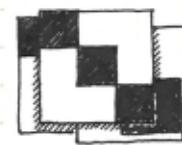
Without queries,  $(\rho \geq \tilde{\Omega}(n^{-1/5} \varepsilon^{-4/5}))$

$$\tilde{O}\left(\frac{\sqrt{n}}{\rho^{3/2} \varepsilon^2}\right)$$

With queries

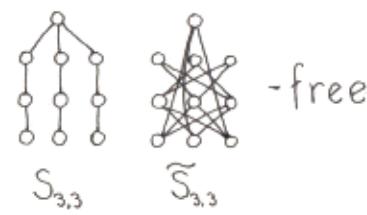
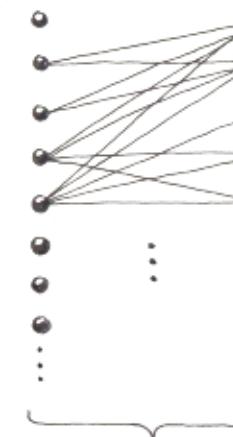
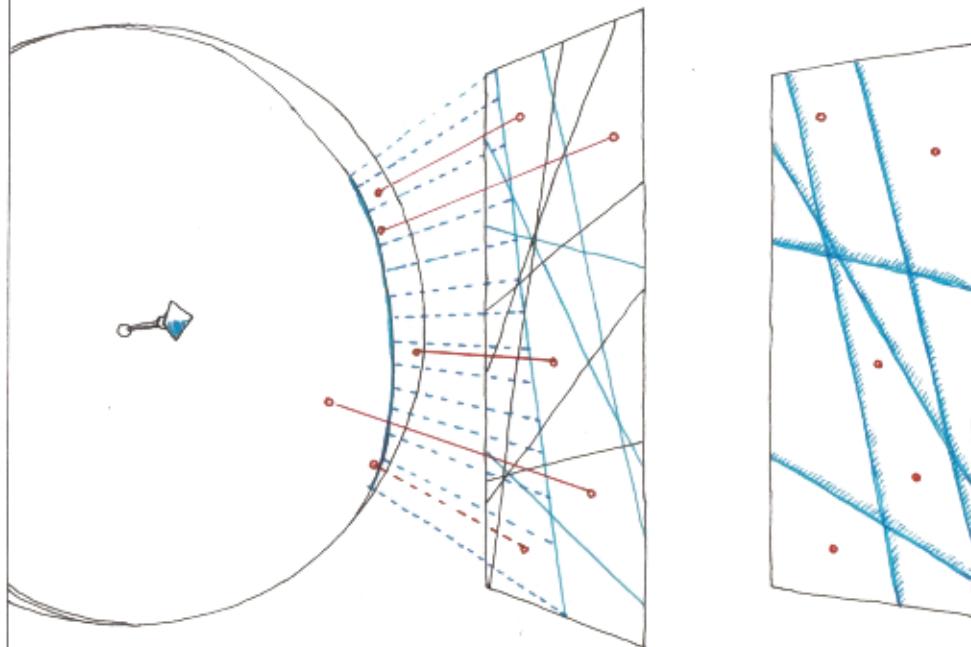
$$O\left(\frac{\sqrt{n}}{\rho \varepsilon^2}\right) \text{ Use [W'17]}$$

Now:



Analyze  $X^T \Phi X - \|X\|_1$

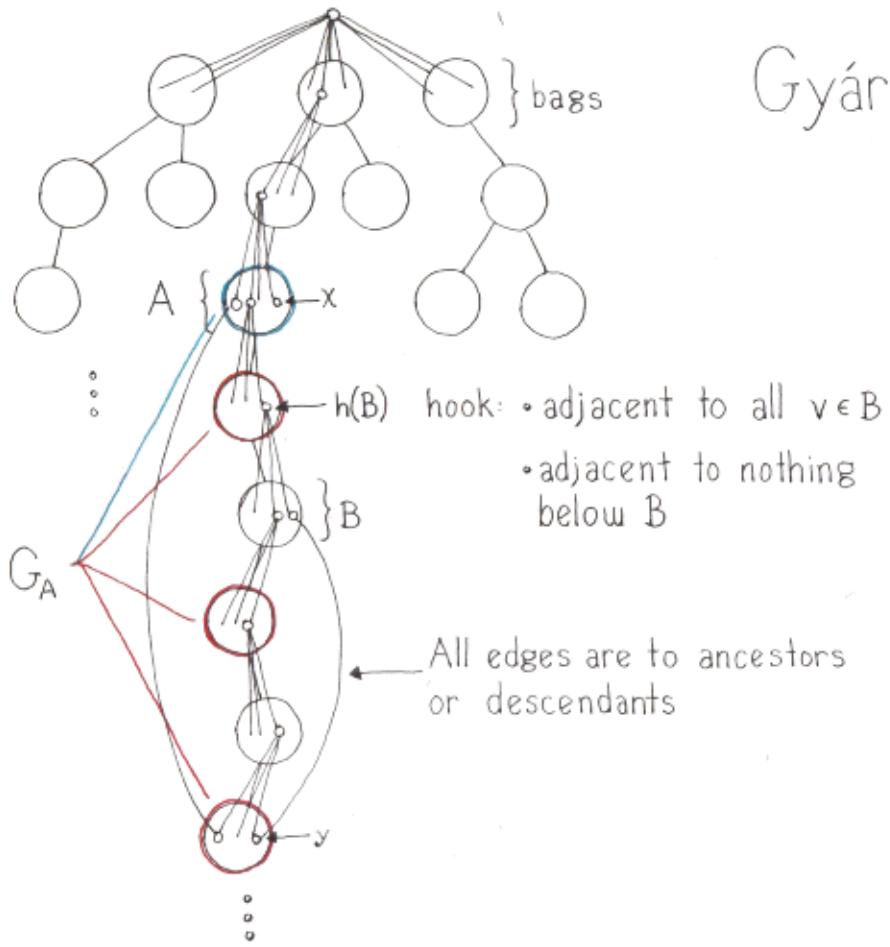
# Step 1



Lemma:

Constant-cost protocol for adjacency in stable  $S_{s,t}, \tilde{S}_{s,t}$ -free bipartite graphs.

## Step 2

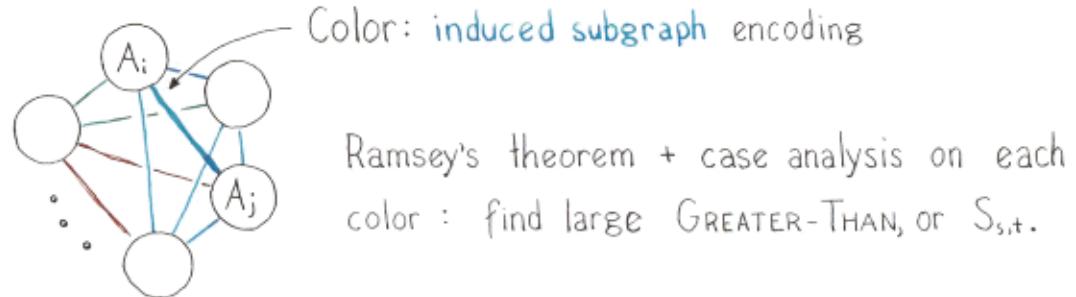
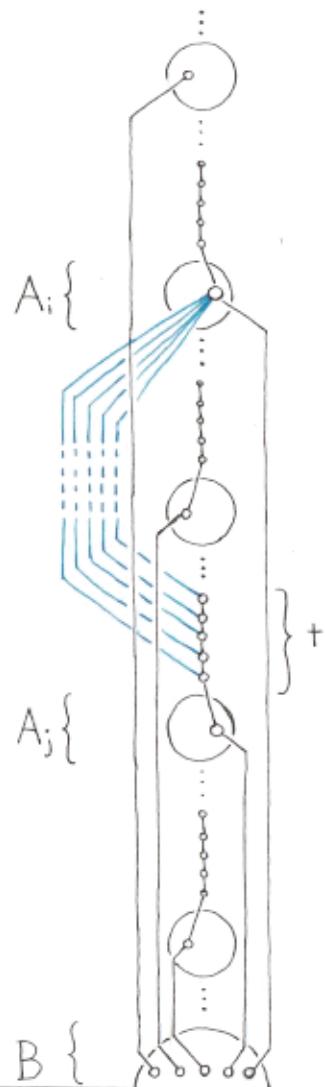


Gyárfás decomposition [POS'22]

- Players agree on ancestor bag  $A$   
 $\Rightarrow$  "Recurse on  $G_A$  or  $\tilde{G}_A$  with smaller GREATER-THAN."
- Stable  $\Rightarrow O(1)$  recursions
- $S_{s,t}, \tilde{S}_{s,t}$ -free  $\Rightarrow$  always  $S_{s,t}$ -free

## Step 3

Lemma: Stable,  $S_{s,t}$ -free  $\Rightarrow$  each bag has edges to  $O(1)$  ancestor bags.



Players need  $O(1)$  EQUALITY queries to agree on the ancestor, or output "non-adjacent."

