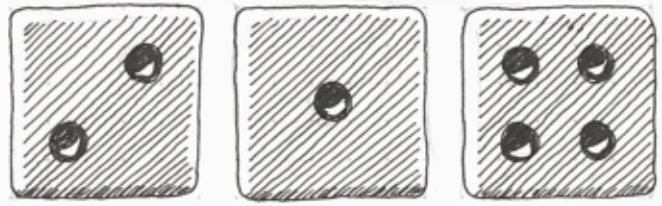
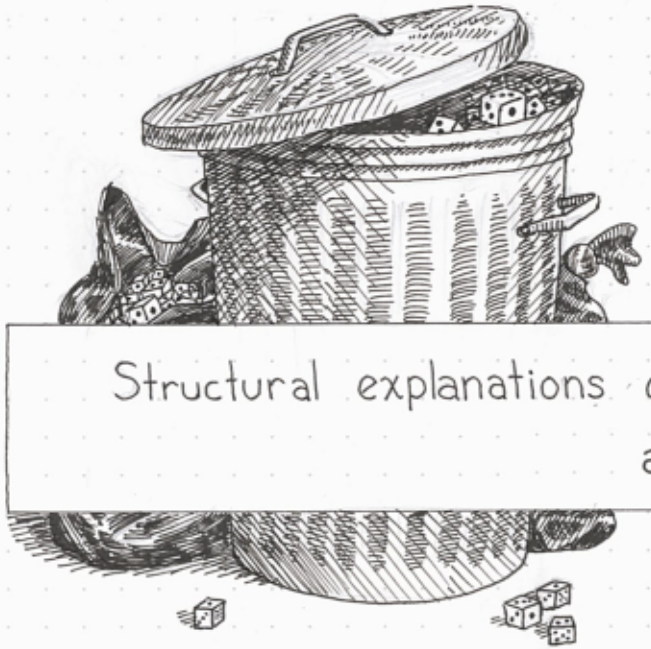
Understanding RANDOMNESS



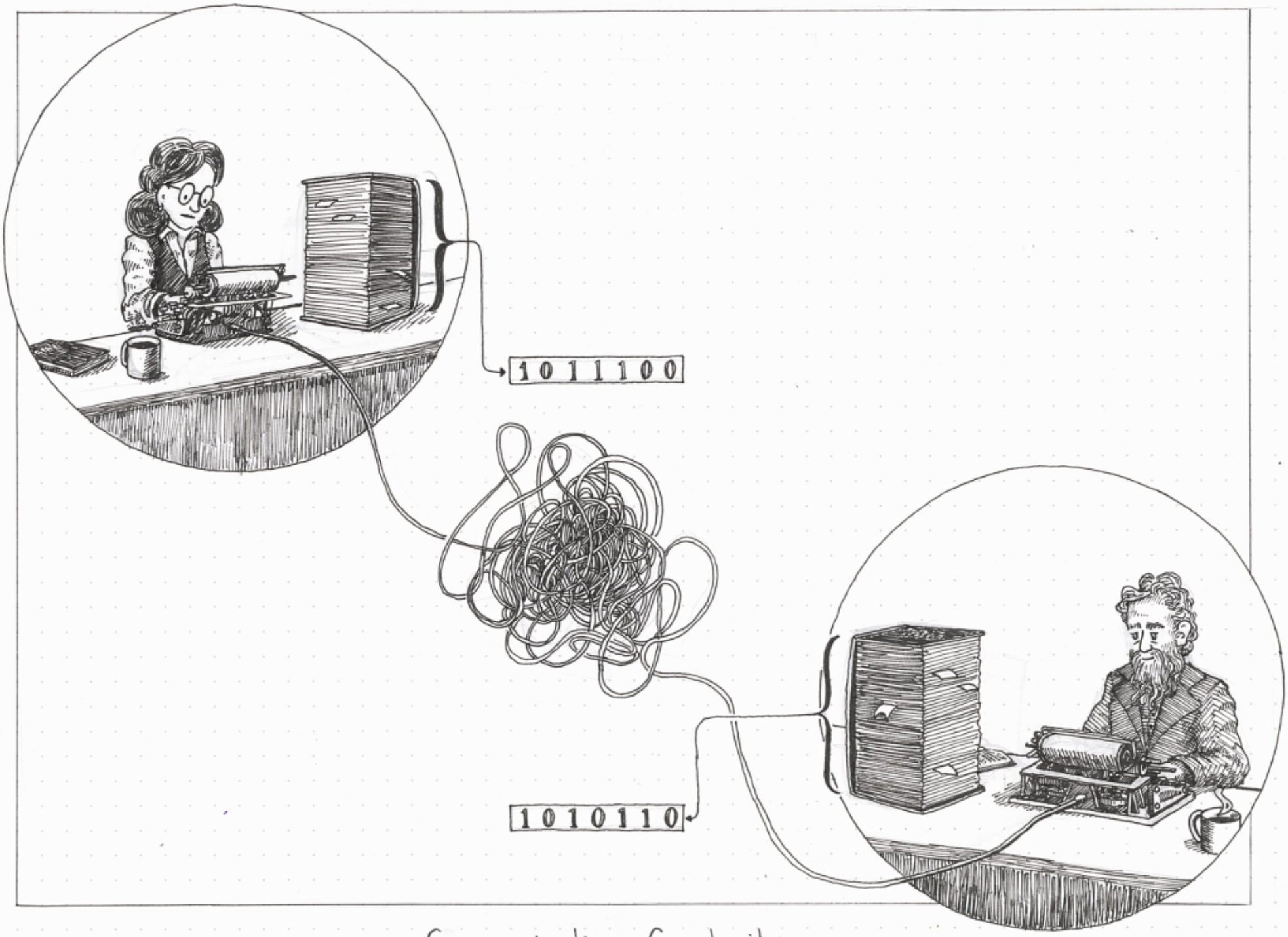
Without

RANDOMNESS

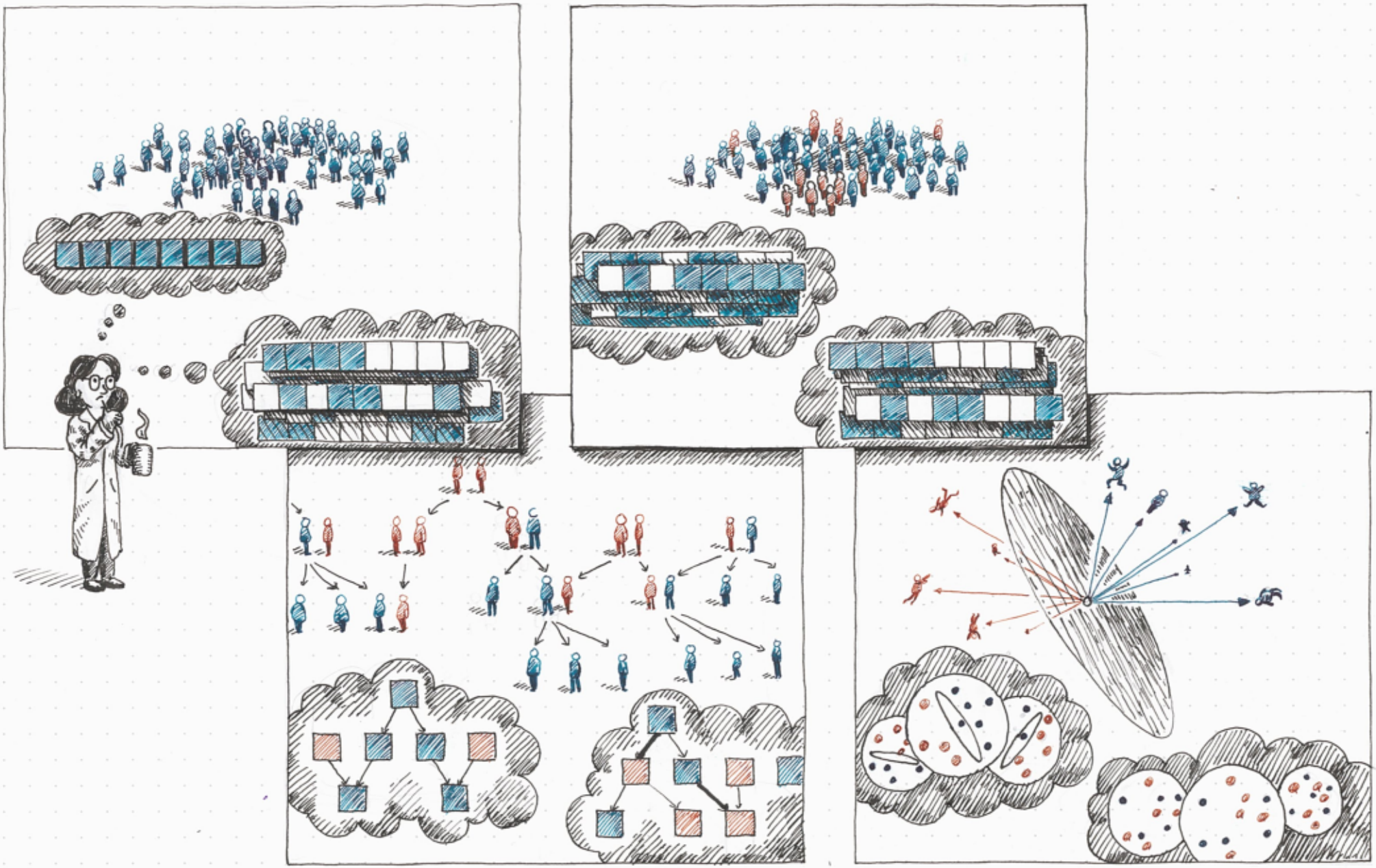


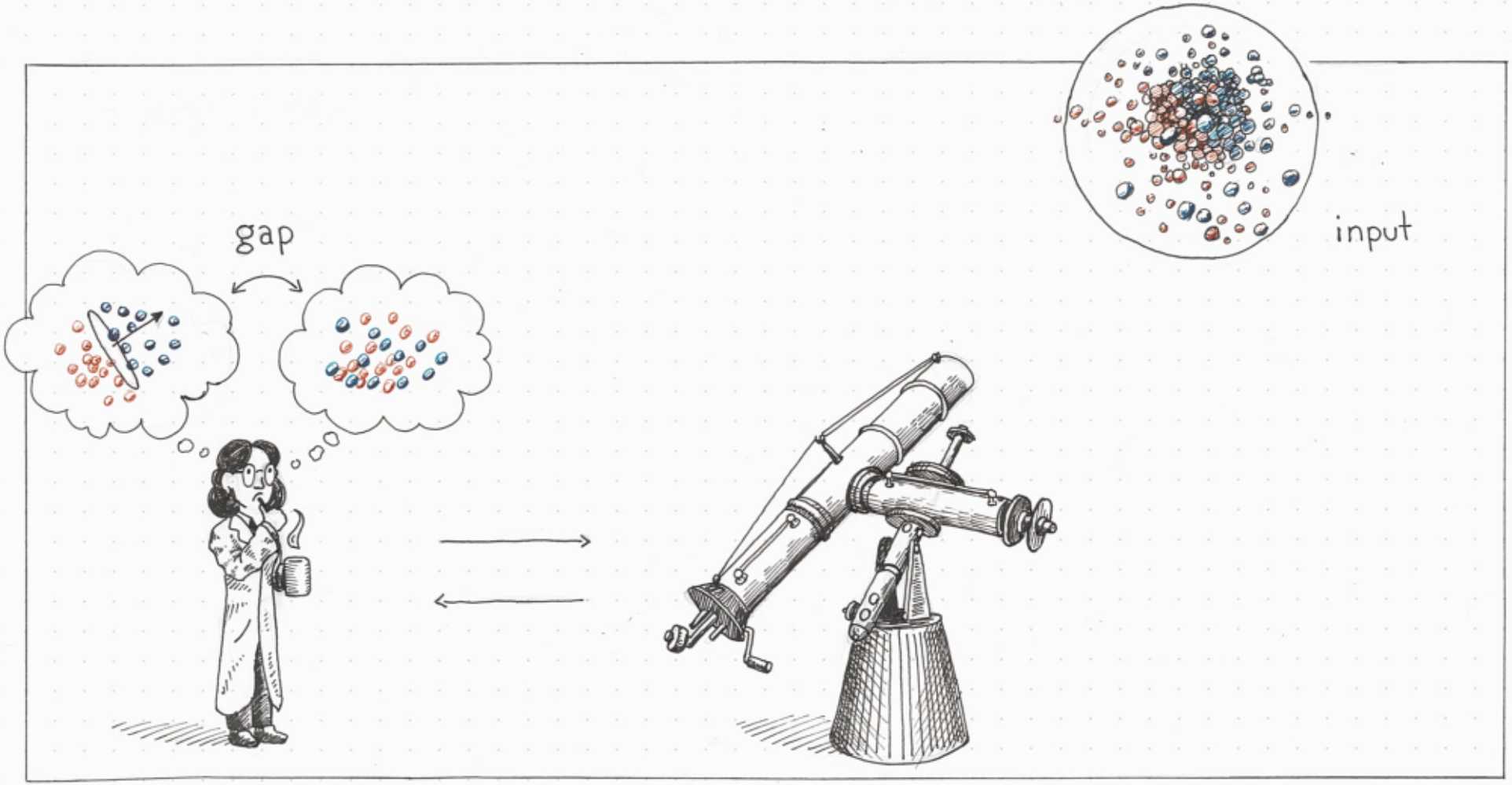
Structural explanations of the power of randomized algorithms

Nathan Harms (EPFL)

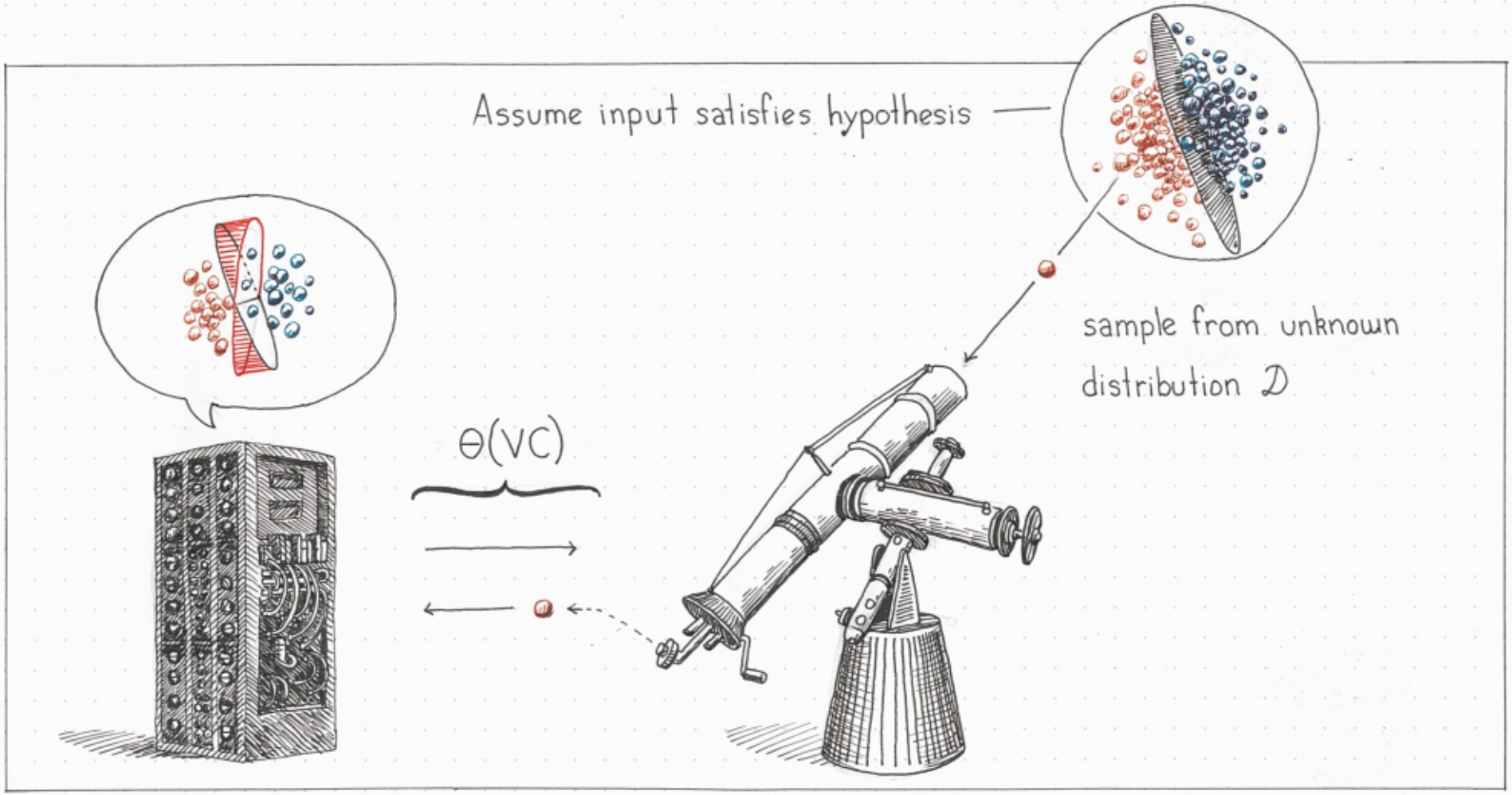


Communication Complexity



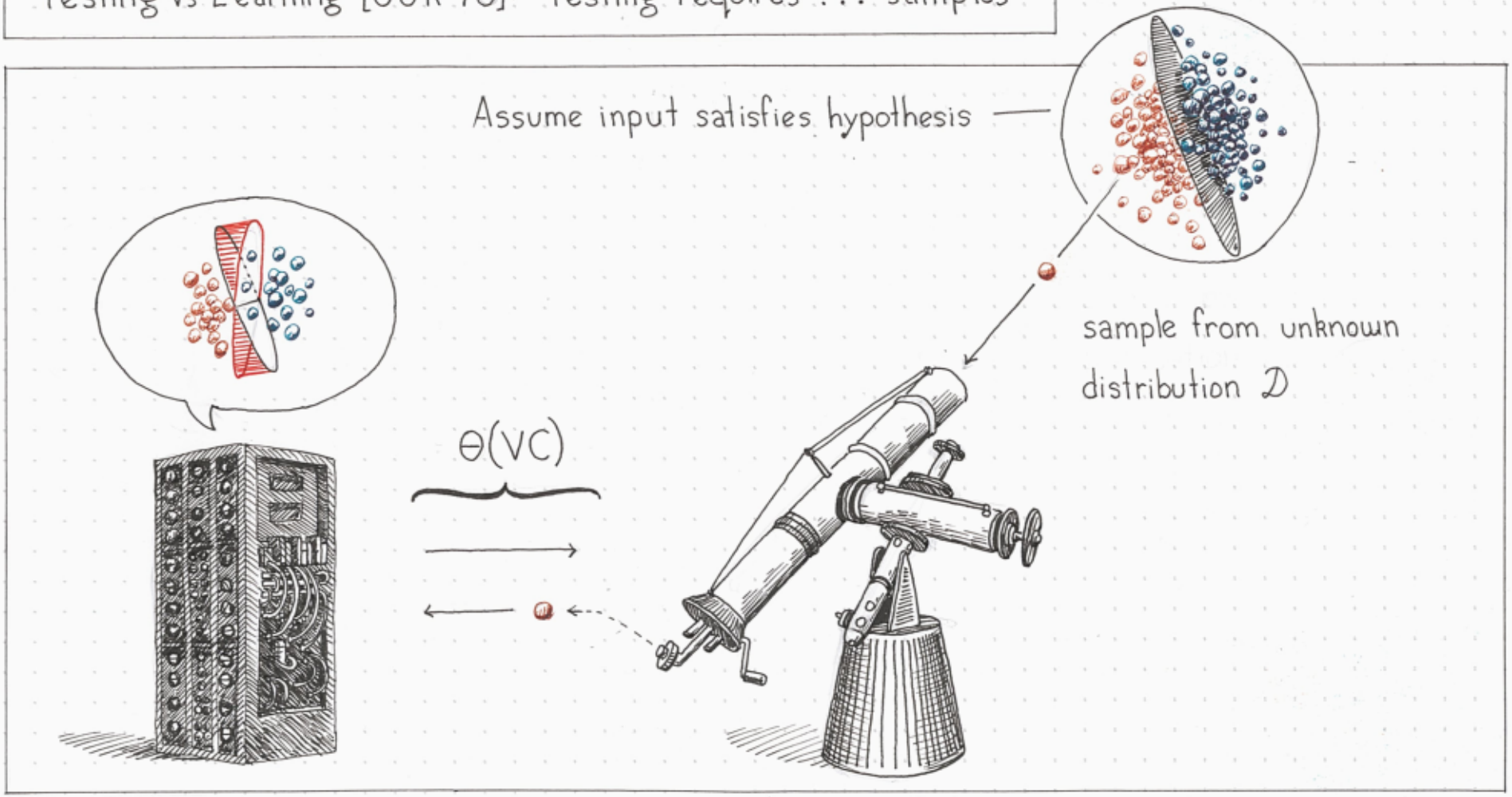


Property Testing

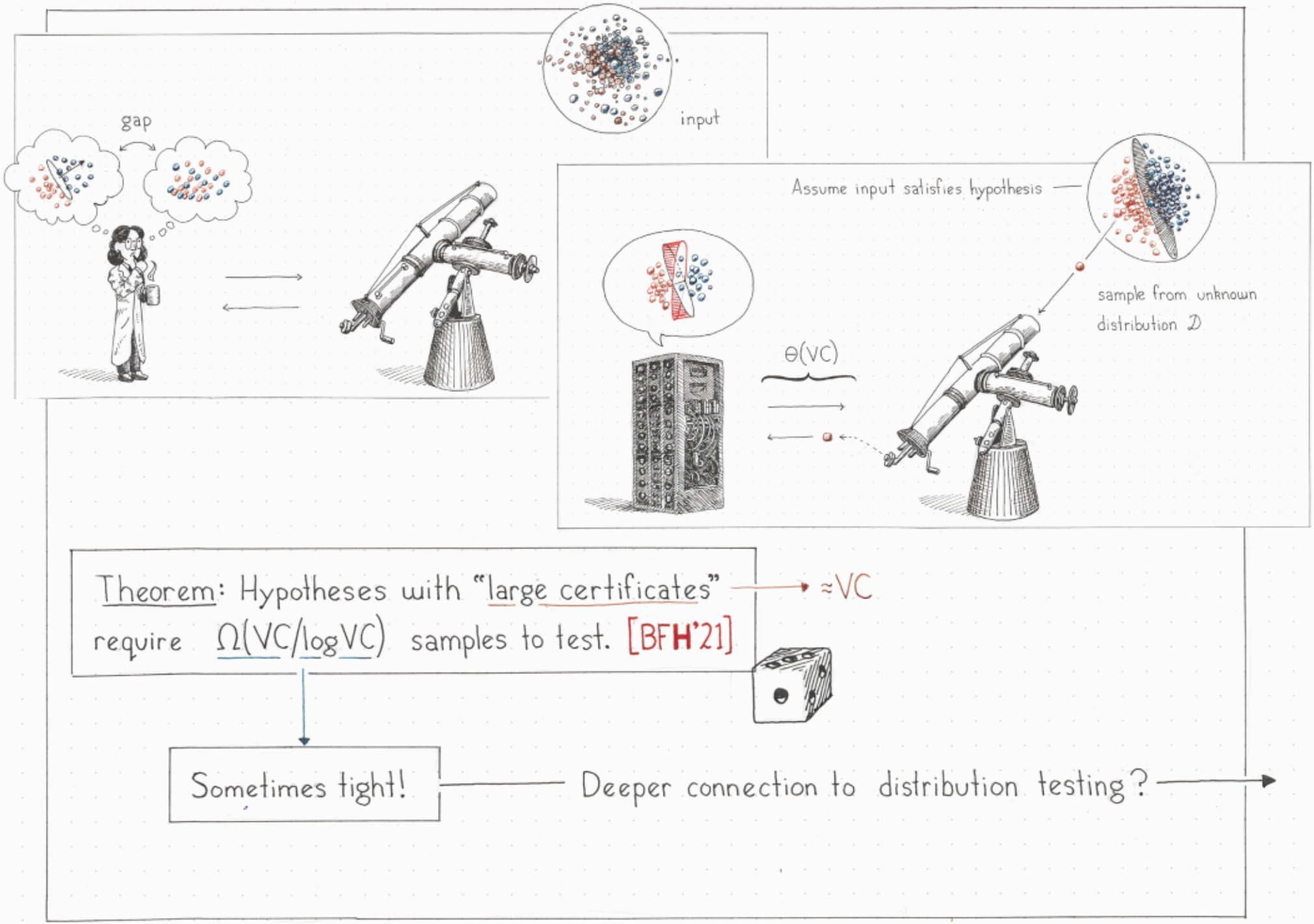

Assume input satisfies hypothesis
sample from unknown distribution $\mathcal{D}$
$\Theta(VC)$

Learning

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

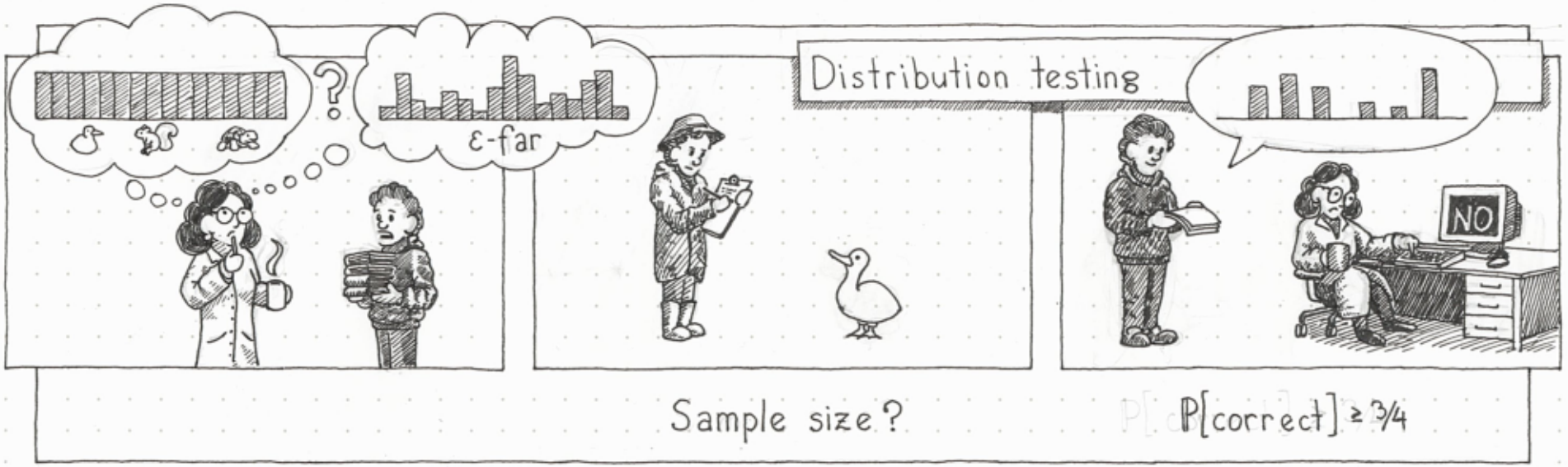


Learning

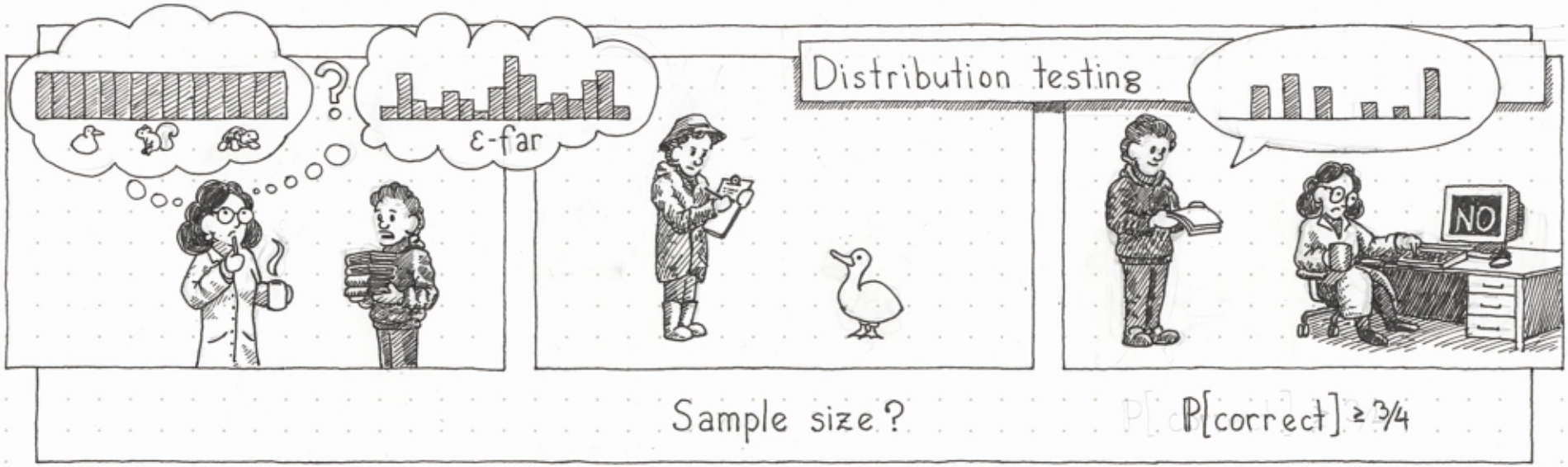


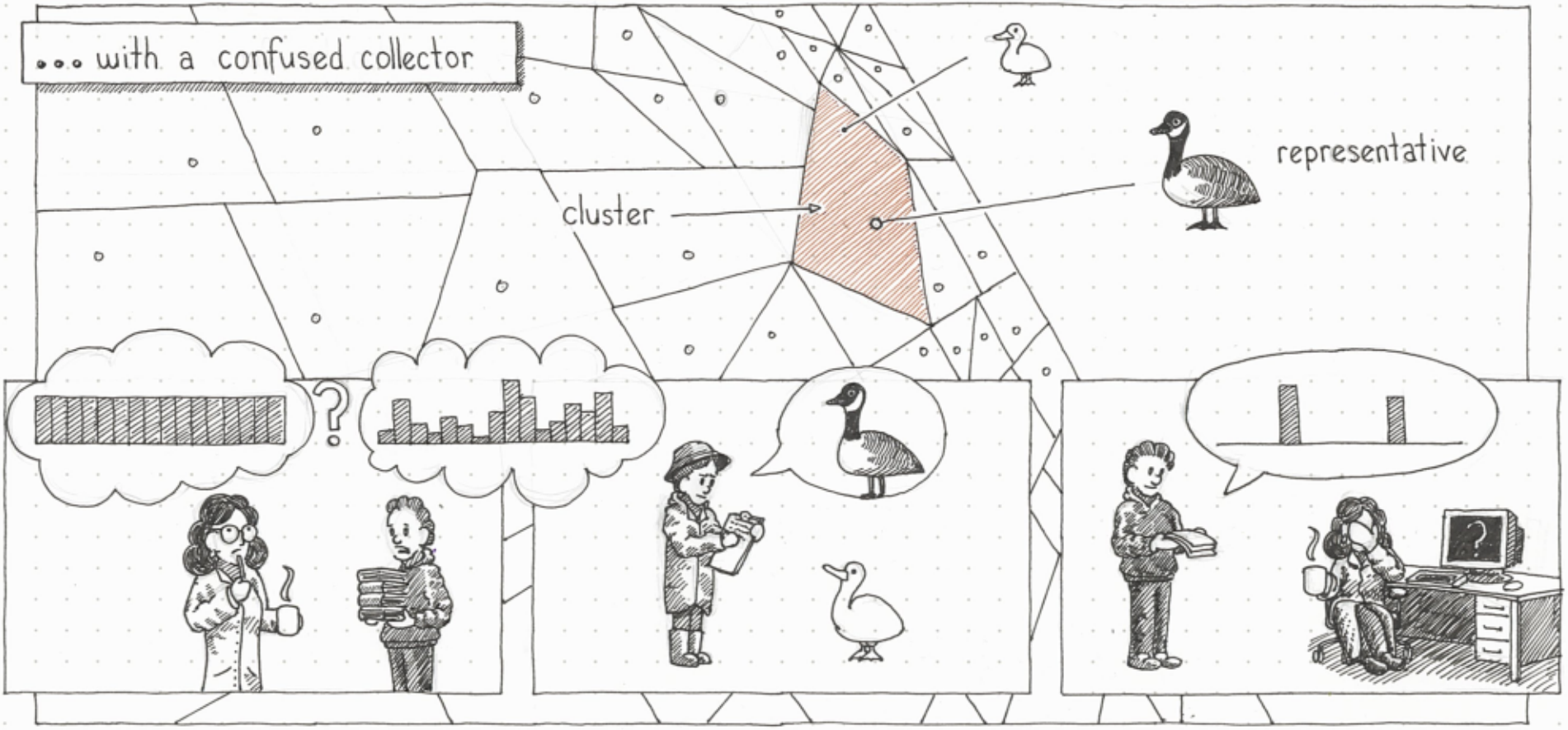
Theorem: Hypotheses with "large certificates" require $\Omega(VC/\log VC)$ samples to test. [BFH'21] → $\approx VC$

Sometimes tight! — Deeper connection to distribution testing? →


Distribution testing
ε-far
NO
Sample size?
P[correct] ≥ 3/4

... with a confused collector
cluster
representative


Distribution testing
ε-far
NO
Sample size?
P[correct] ≥ 3/4

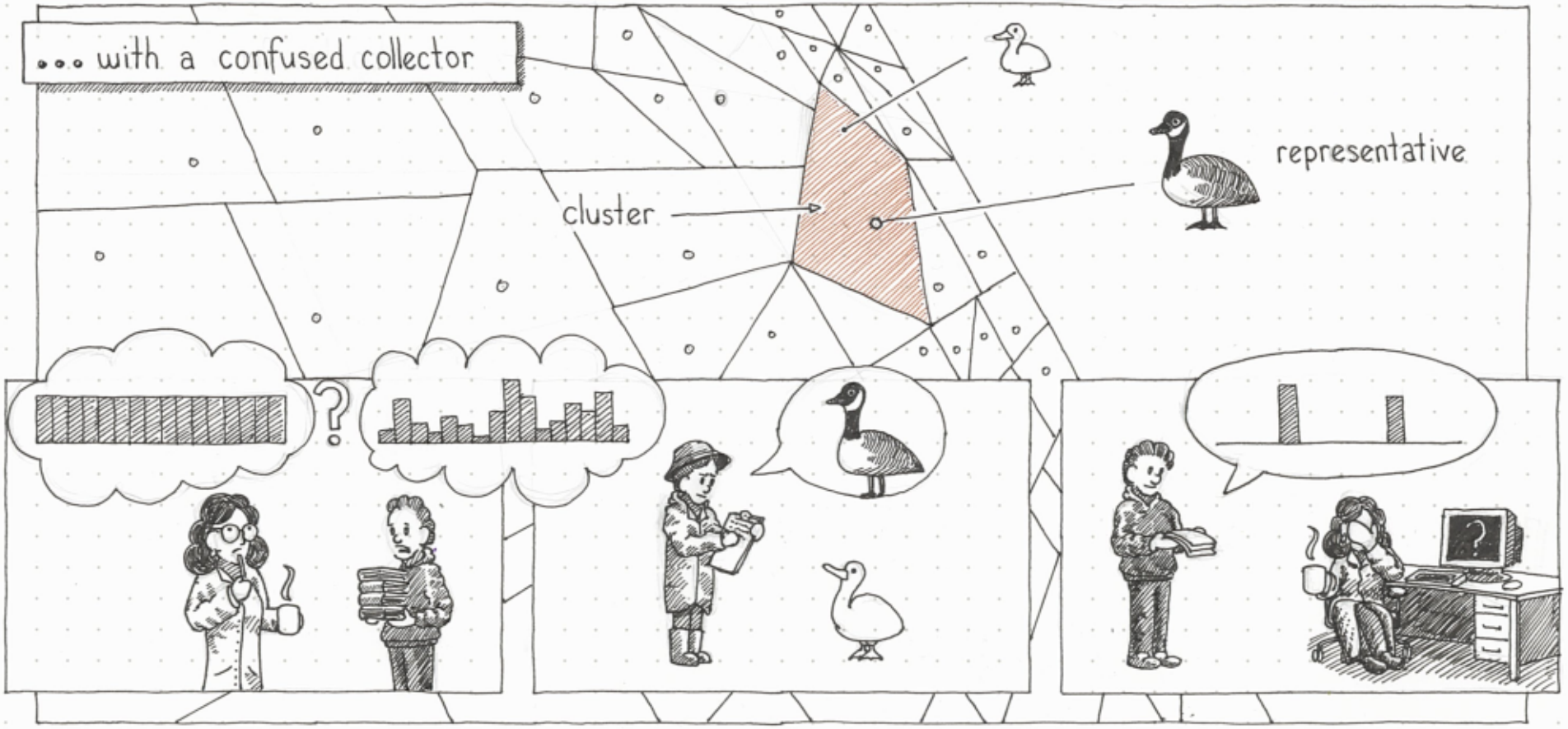

... with a confused collector
cluster
representative
?

Results: [FH'24]

Algorithms for

random clustering

adversarial clustering!



Renato Ferreira Pinto Jr., University of Waterloo
Nathan Harms, EPFL

Distribution Testing & Testing vs Learning

Lower Vee Sea

where properties have LVC = Θ(VC)

VC and Lower VC dimension of a property
VC: Size of the largest shattered set
LVC: Size of the smallest non-shattered set certificate of non-membership

Theorem [BFH21]: For any property with LVC = Ω(VC), testing requires $\Omega\left(\frac{VC}{\log VC}\right)$ samples.

Reduction from distribution testing.
(Testing support size [VV11, CJL19])

Confused Collector

Symmetric Properties

Closed under permutations on domain {1,2,⋯,n}

Property testing is equivalent to distribution testing [GR16].
$\Theta\left(\frac{VC}{\log VC}\right)$ is tight.

How to extend equivalence to non-symmetric properties?

distribution testing

parity trace

Density Properties

Halfspaces

Lower bounds on testing halfspaces implies lower bounds on testing halfspaces.

k-Alternating Functions

Equivalent to testing support size under the parity trace.

Uniformly k-Alternating Functions

Test whether a k-alternating function has uniform $\frac{n}{k+1}$ sign changes between alternation points.

Trace Reconstruction

Theorem: Testing uniformity of distributions over [k] requires $\Theta((k/\varepsilon)^{4/3}\cdot\sqrt{k}/\varepsilon^2)$ samples under the parity trace.

Proof: Trace

Run-lengths X_0 X_1 X_2 X_3 ⋯

Compute $\sum X_i(X_i-1)$, reject if large

$T\Phi T - \|T\|_4$ $T\Phi T - \|T\|_4$

T_i = number of occurrences of (odd/even) element $i \in [k]$ in the sample.

Φ_{ij} = 1 iff no (even/odd) element between i and j occurs in the sample.

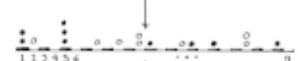
Join matrix Φ

Prove concentration of quadratic forms $T\Phi T - \|T\|_4$

Distribution Testing under the Parity Trace

Input: D

Sample: S



Trace:

Tester

D is in property P

D is ε-far from P in edit distance

Density Properties

Labelled-distribution testing

- Property P: Set of function-distribution pairs (f, D), $f: X \to \{0,1\}$, distribution D over X
- Test if ① $(f, \hat{D}) \in P$
  ② $\forall (g, \mathcal{E}) \in P$, $|\hat{D} - \mathcal{E}_g|_{TV} > \varepsilon$
  distribution of $(x, f(x))$, $x \sim D$

Density property:

D:

f:

Membership depends only on the density sequence R, R, B, R, R, ⋯

Testing vs Learning

Which properties (hypothesis classes) can be tested with fewer samples than are required for learning, in the distribution-free sample-based model? Or,

When does testing require o(VC) samples?

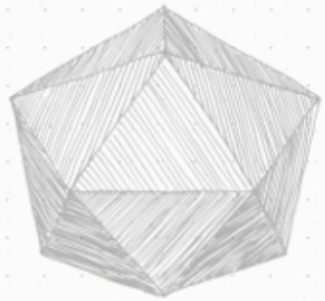
VC Dimension

References

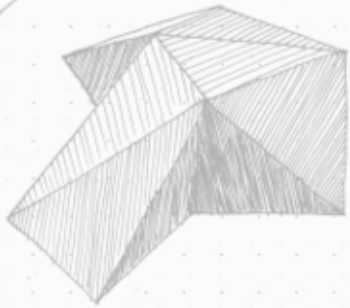
BFH21: Blais, Ferreira Pinto Jr., Harms. STOC 2021
GR16: Goldreich, Ron. ToCT 2016
VV11: Valiant, Valiant. STOC 2011
CJL19: Chu, Jiang, Acharya, Indyk. 2019

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

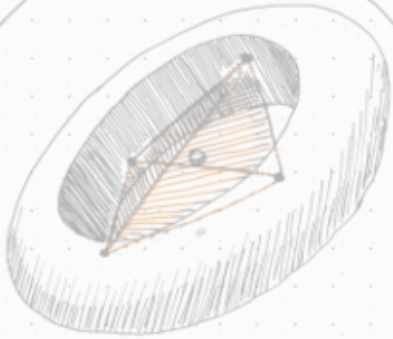
Testing Convex Sets



Convex



ε-Far from Convex

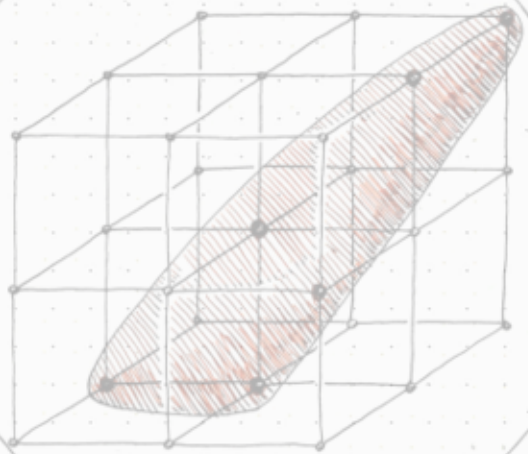


Tight bounds:

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

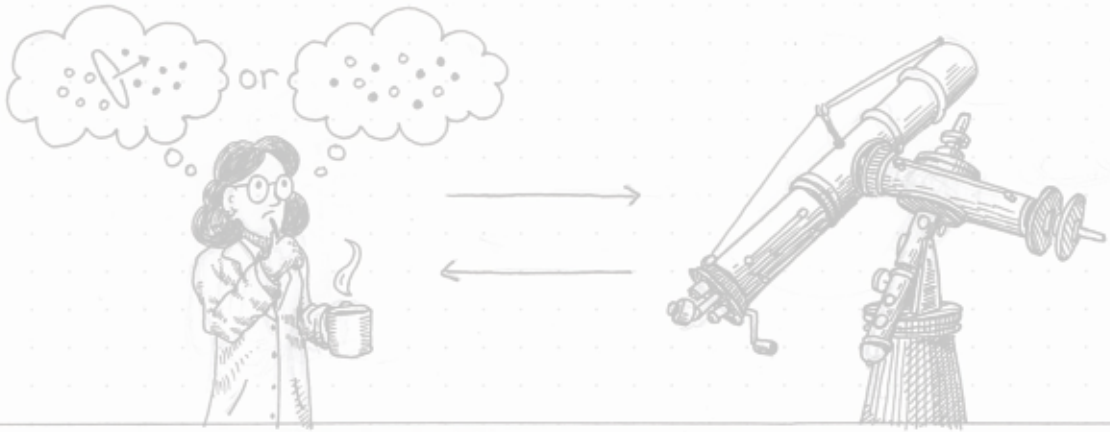
Unclear how to use queries [BB'20, RV'04]

Discrete convex sets?



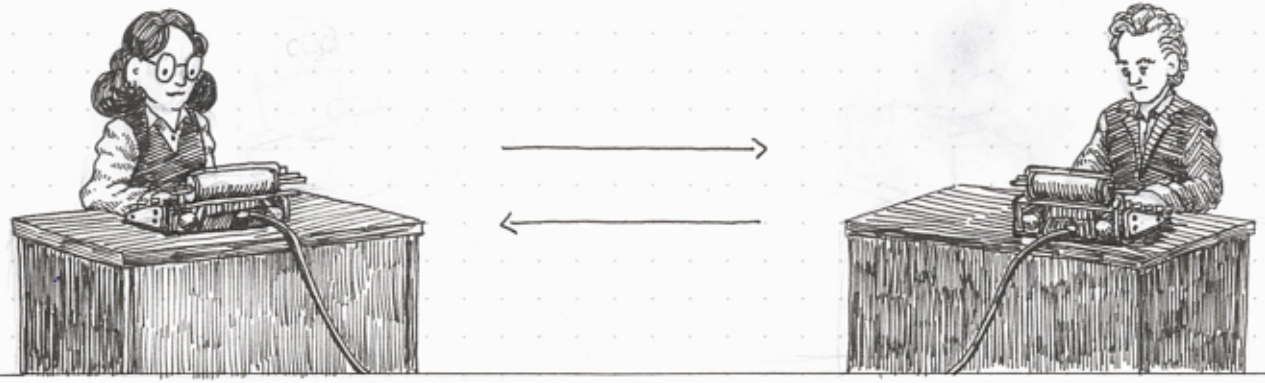
1

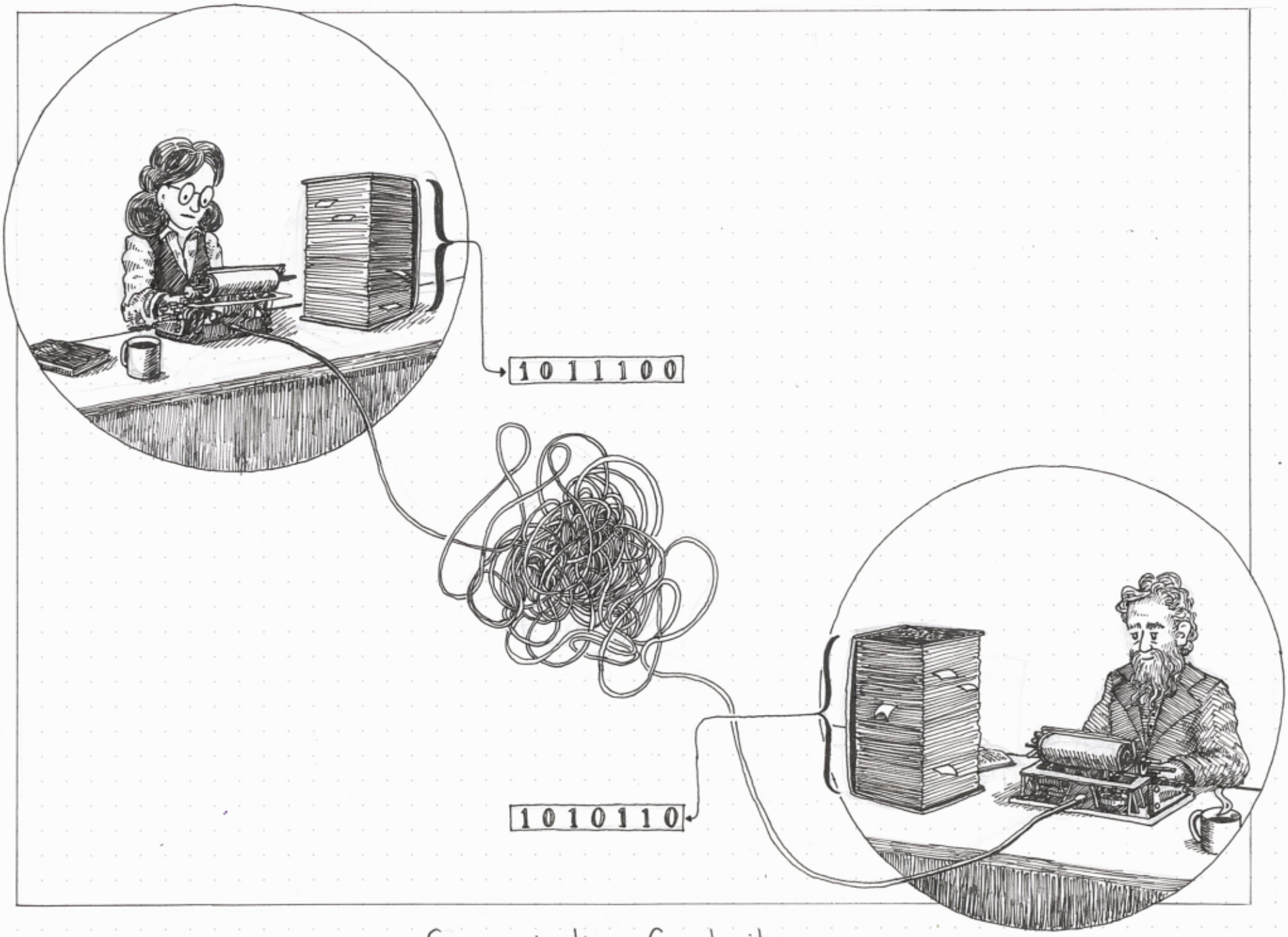
Property Testing



2

Communication Complexity





Communication Complexity

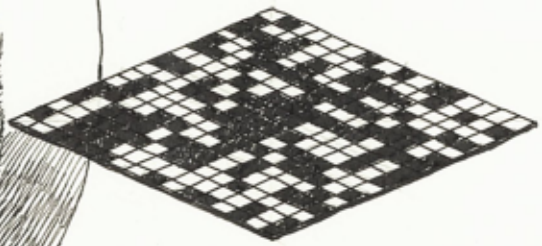
Why?

① Power of randomness
- Most extreme case
- Most basic lower bound
- "Fine-grained" understanding
- Standard techniques fail
- Complete characterization?
- Most evident structure

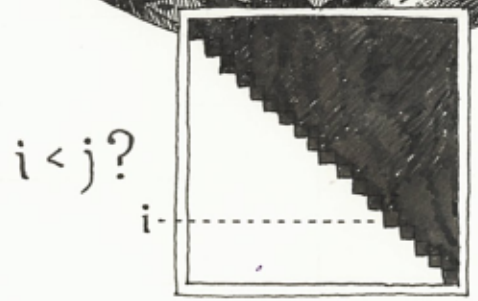
② Connections
- Matrix representations
- Structural graph theory
  - → New concepts
  - → New techniques
- Algebra
- Learning theory

③ It is cool

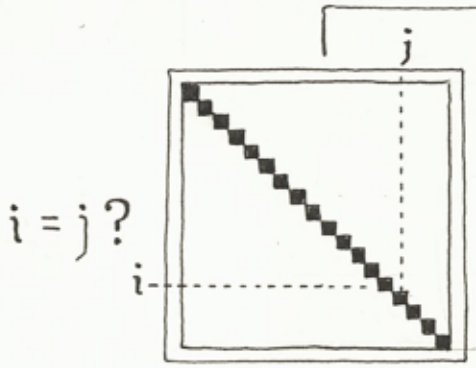
Randomized Communication



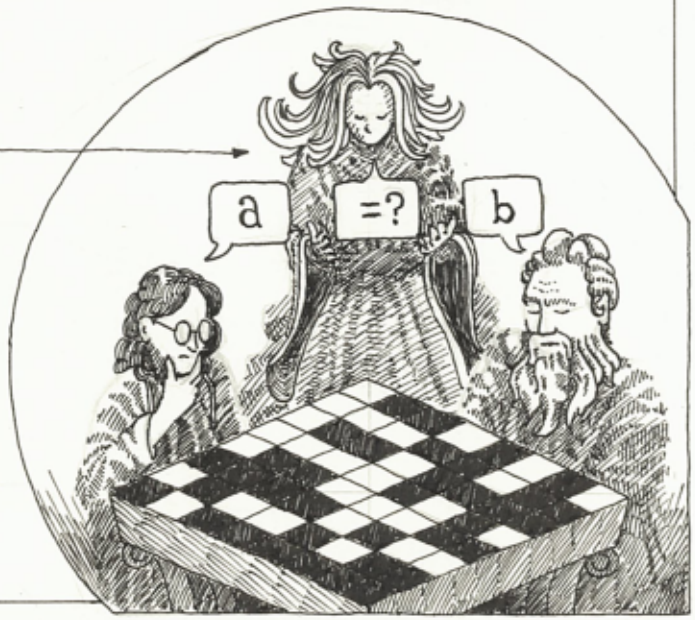
BPP^0: Constant Cost

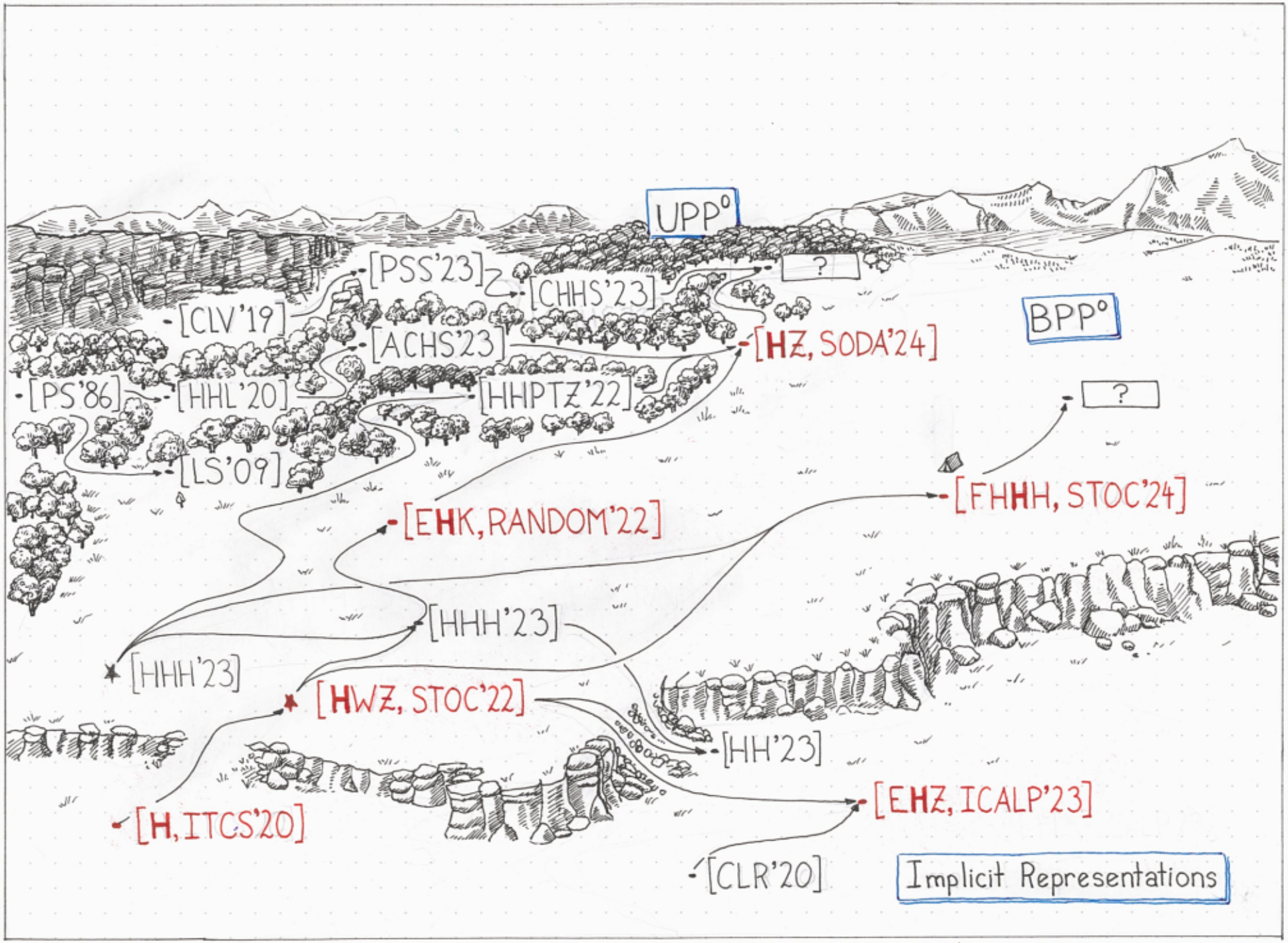


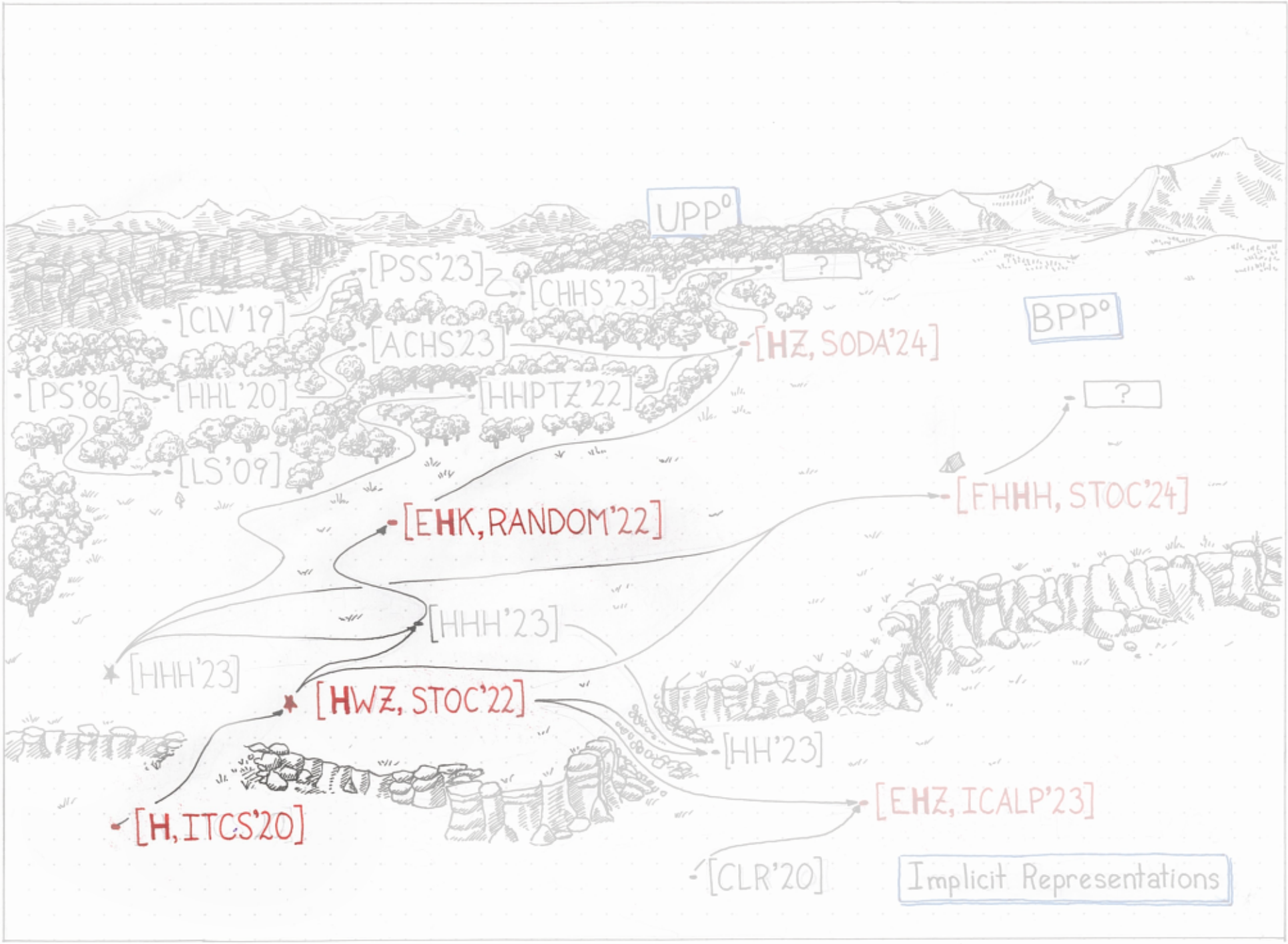
GREATER-THAN

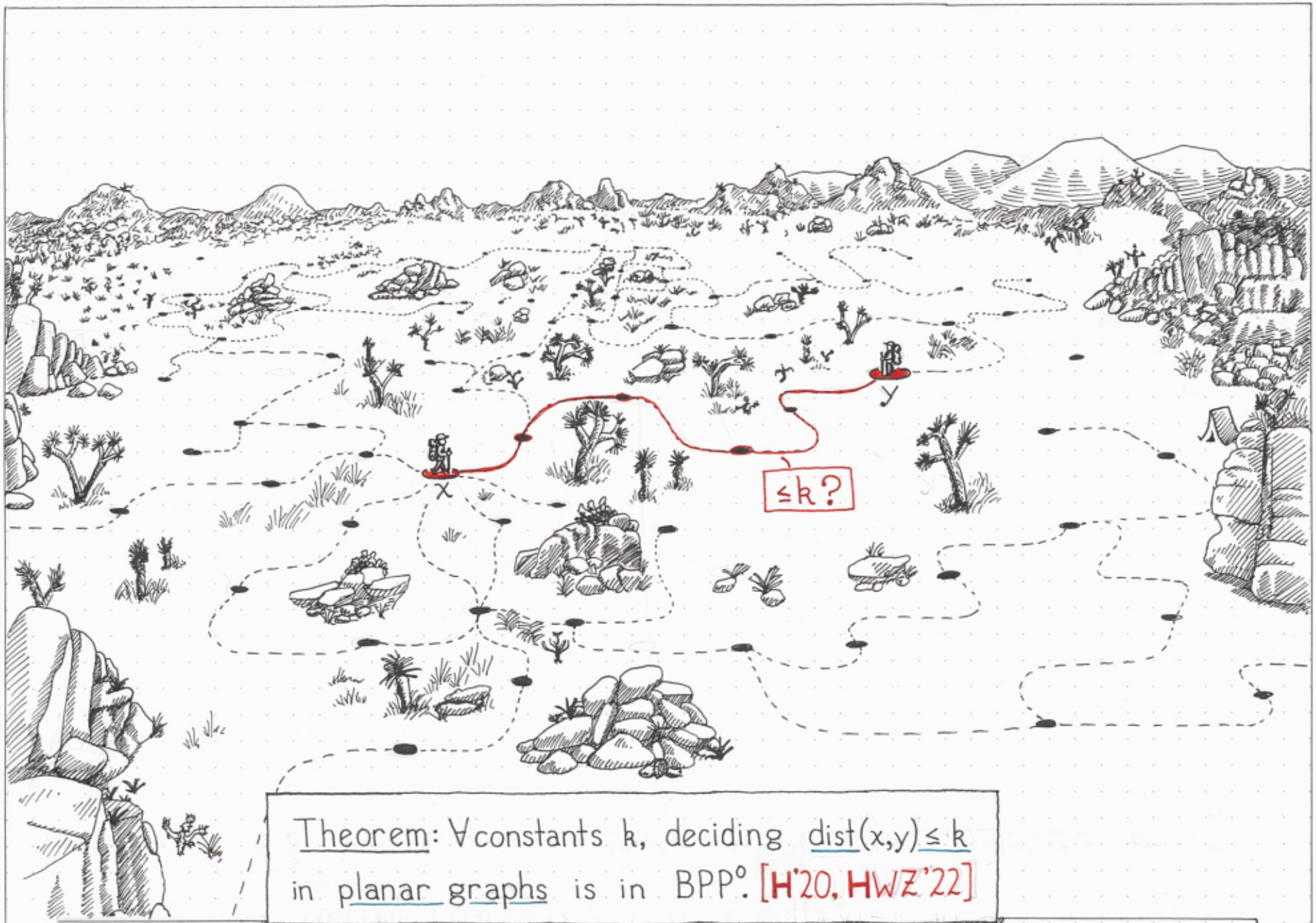


EQUALITY




UPP°
[PSS'23]
[CHHS'23]
?
[CLV'19]
BPP°
[ACHS'23]
[HZ, SODA'24]
[PS'86]
[HHL'20]
[HHPTZ'22]
?
[LS'09]
[FHHH, STOC'24]
[EHK, RANDOM'22]
[HHH'23]
[HHH'23]
[HWZ, STOC'22]
[HH'23]
[EHZ, ICALP'23]
[H, ITCS'20]
[CLR'20]
Implicit Representations


UPP°
[PSS'23]
[CHHS'23]
?
[CLV'19]
BPP°
[ACHS'23]
[HZ, SODA'24]
[PS'86]
[HHL'20]
[HHPTZ'22]
?
[LS'09]
[FHHH, STOC'24]
[EHK, RANDOM'22]
[HHH'23]
[HHH'23]
[HWZ, STOC'22]
[HH'23]
[EHZ, ICALP'23]
[H, ITCS'20]
[CLR'20]
Implicit Representations

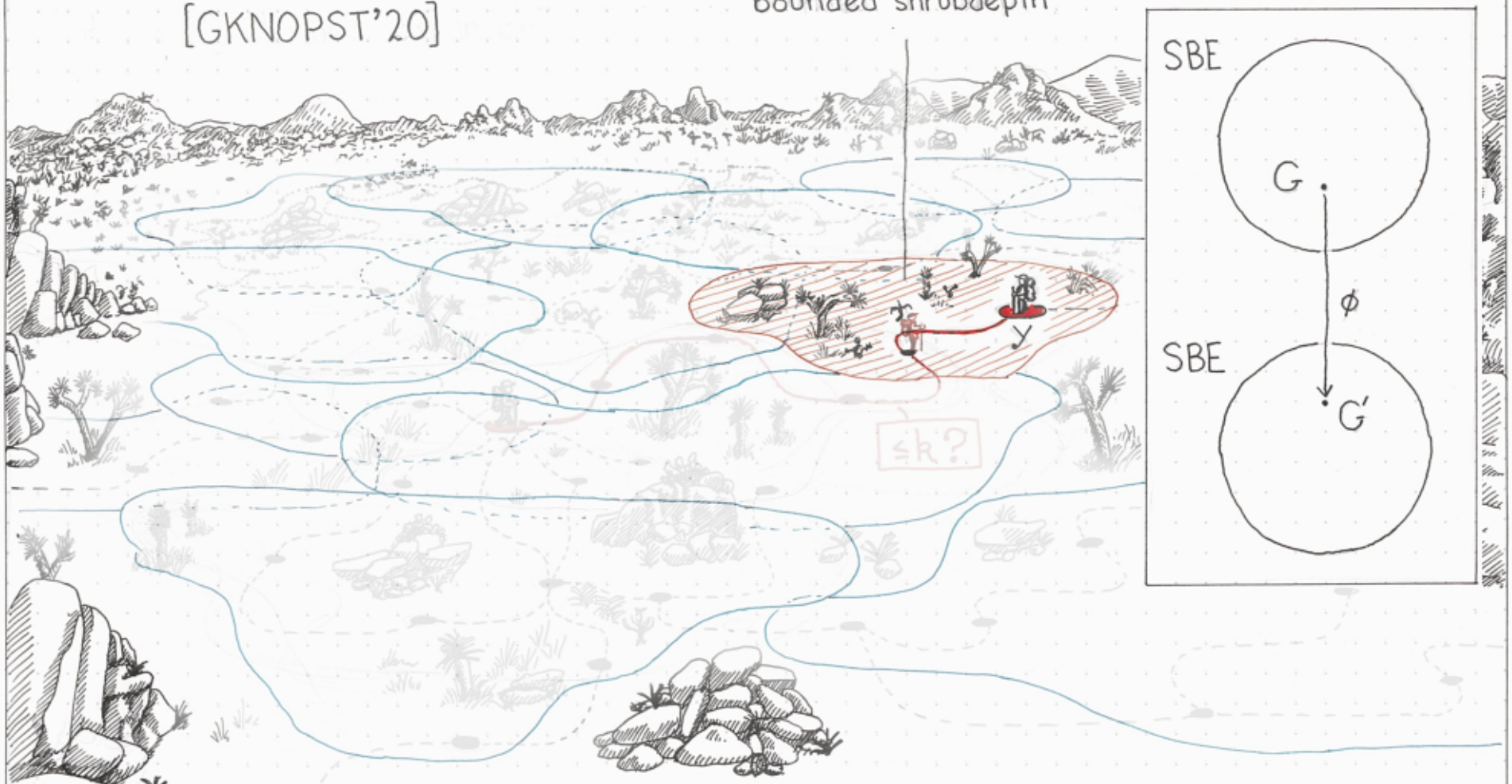


Theorem: ∀ constants k, deciding $\mathrm{dist}(x,y) \le k$ in planar graphs is in BPP°. [H'20, HWZ'22]

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

"Structurally bounded expansion"
[GKNOPST'20]

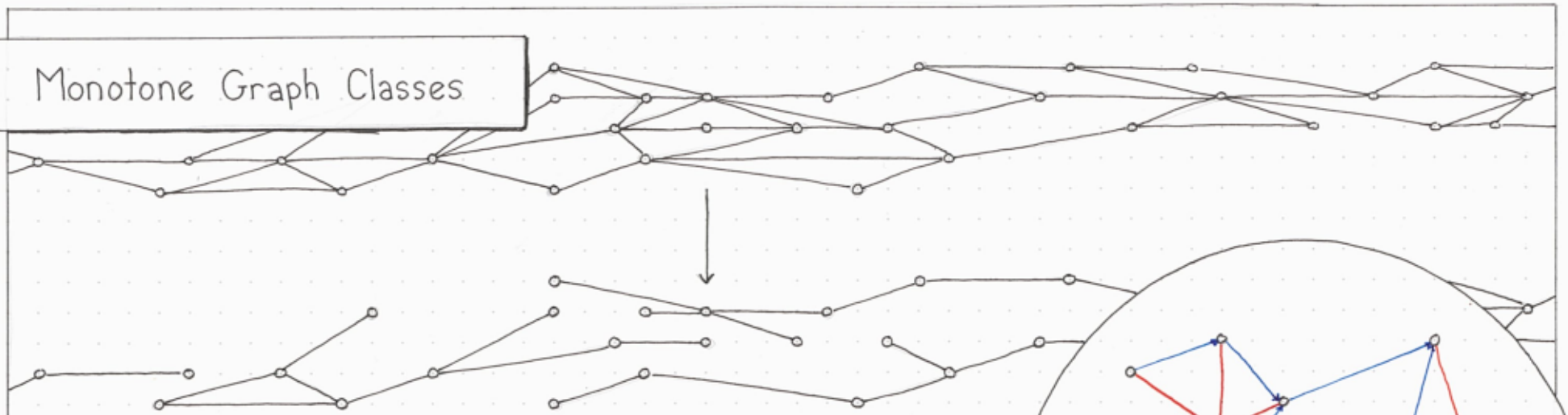
bounded shrubdepth

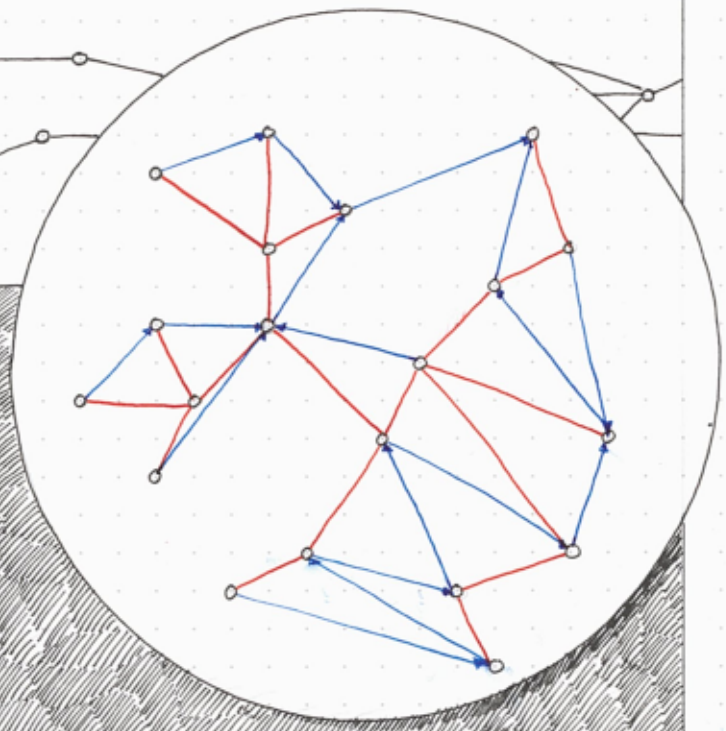


Theorem: ∀ constants k, deciding $\mathrm{dist}(x,y) \le k$ in planar graphs is in BPP°. [H'20, HWZ'22]

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

Monotone Graph Classes



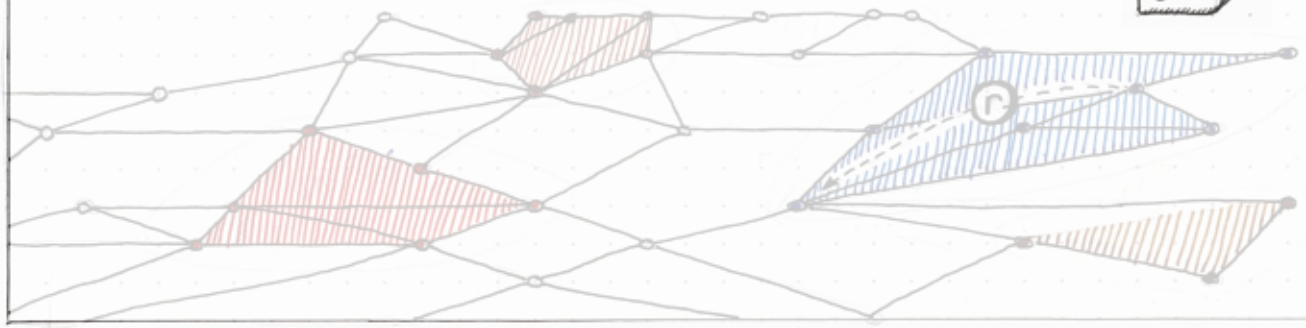


Theorem: If $\mathcal{G}$ is a monotone graph class:

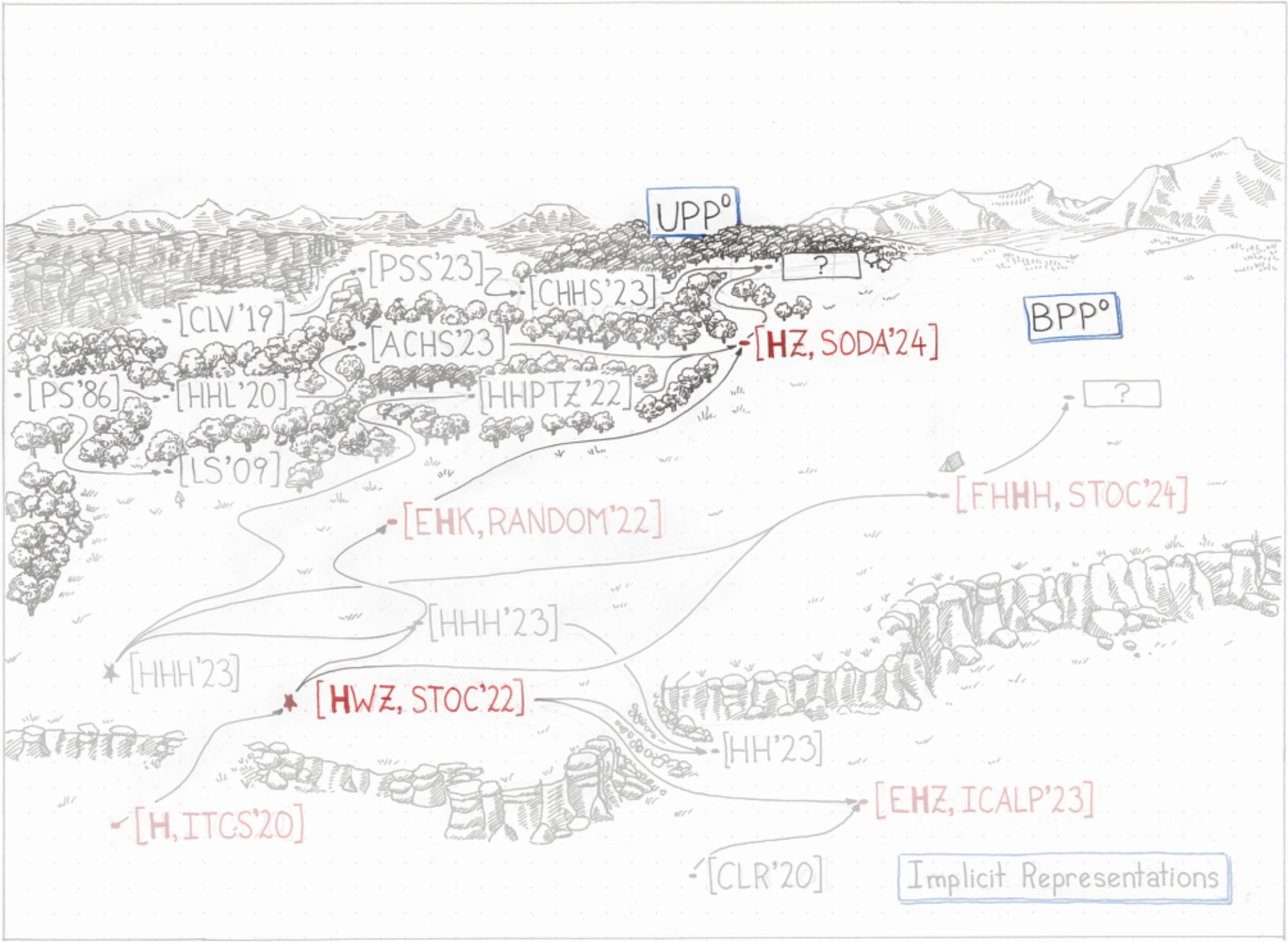
① Adjacency is in BPP^0
$\Leftrightarrow$ $\mathcal{G}$ has bounded arboricity.

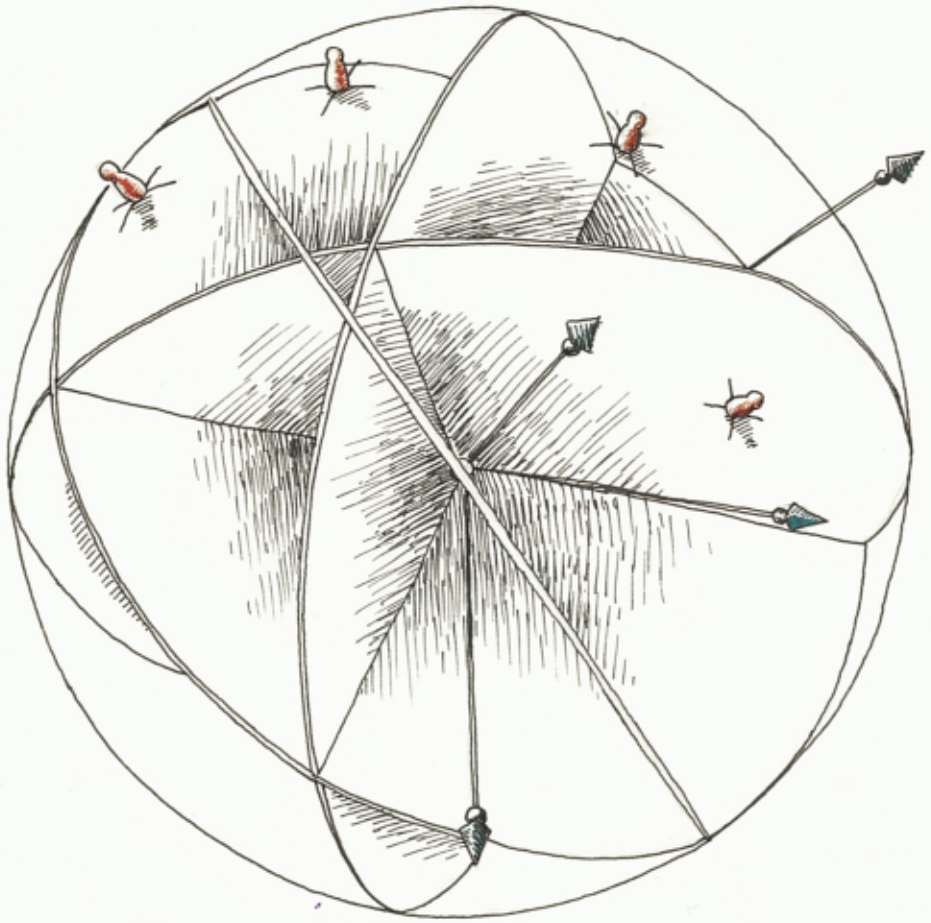
② Distance k is in BPP^0 $\forall$constants k
$\Leftrightarrow$ $\mathcal{G}$ has bounded expansion.

And the EQUALITY oracle suffices. [EHK'22]

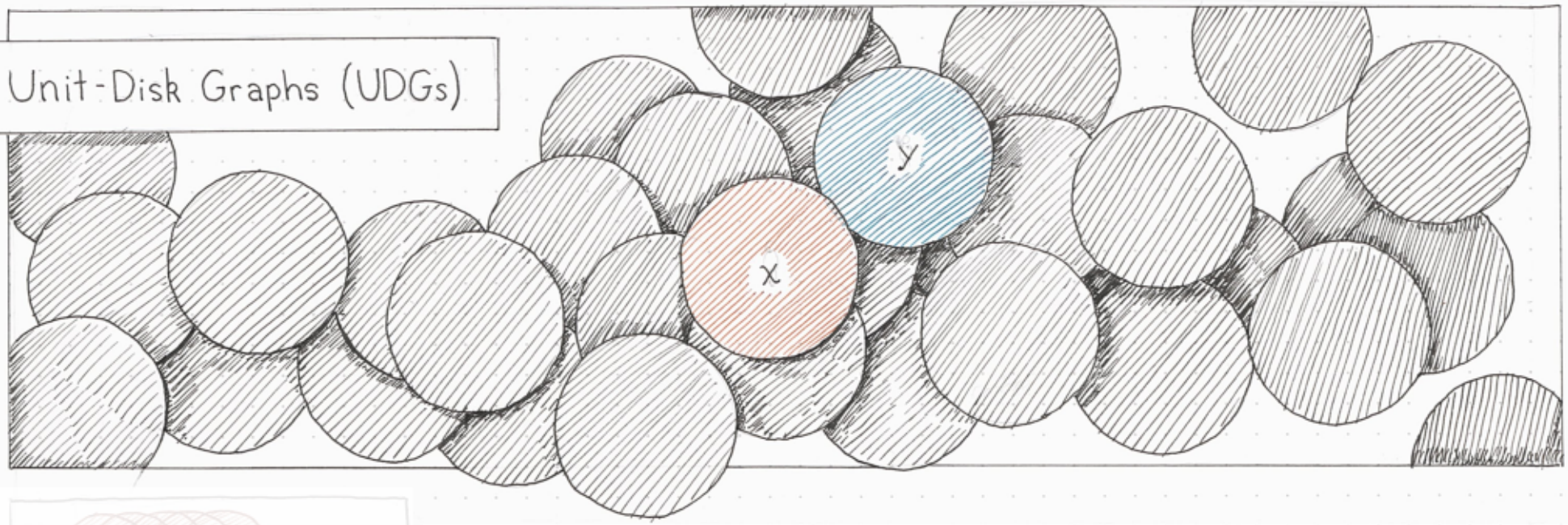


contraction has average degree f(r). [e.g. NO'12]


UPP°
?
BPP°
[PSS'23]
[CHHS'23]
[CLV'19]
[ACHS'23]
[HZ, SODA'24]
[PS'86]
[HHL'20]
[HHPTZ'22]
?
[LS'09]
[FHHH, STOC'24]
[EHK, RANDOM'22]
[HHH'23]
[HHH'23]
[HWZ, STOC'22]
[HH'23]
[EHZ, ICALP'23]
[H, ITCS'20]
[CLR'20]
Implicit Representations



Unit-Disk Graphs (UDGs)



GREATER-THAN

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

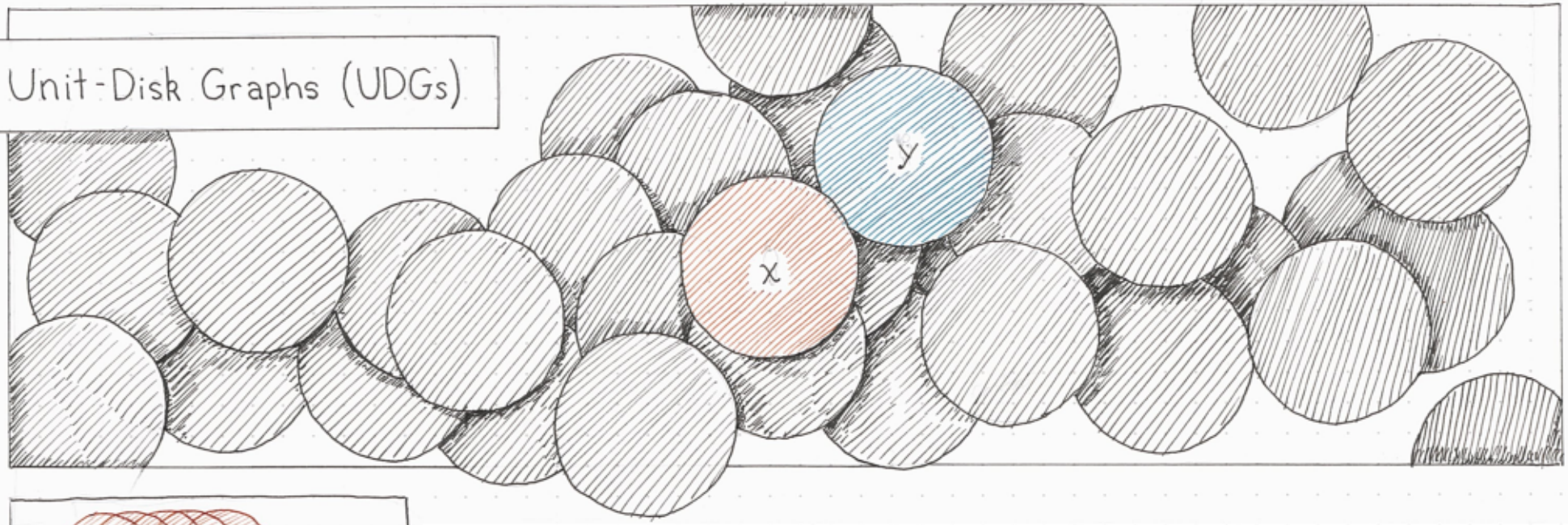
Proof:

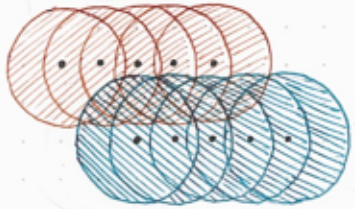
→ , , -free →

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

[OPS'22]: "Gyárfás decomposition"

Unit-Disk Graphs (UDGs)





GREATER-THAN

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

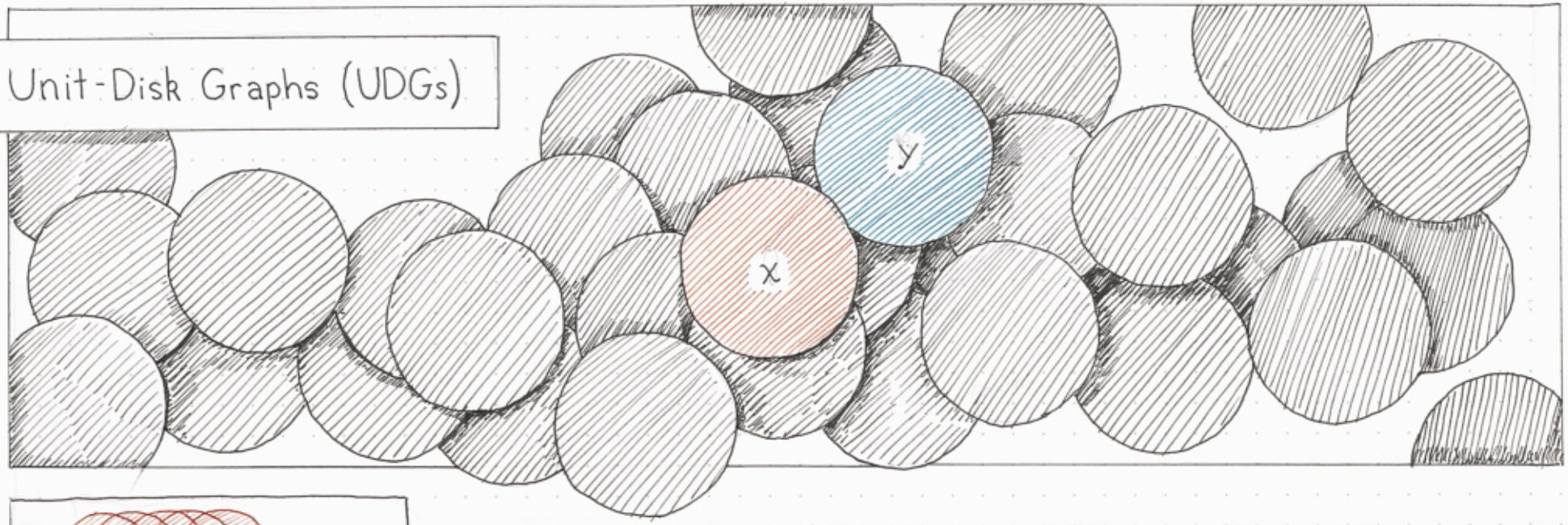
Proof:

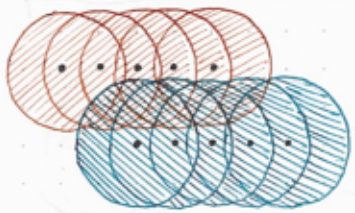
→ , , -free →

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

[OPS'22]: "Gyárfás decomposition"

Unit-Disk Graphs (UDGs)





GREATER-THAN

— remove →

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

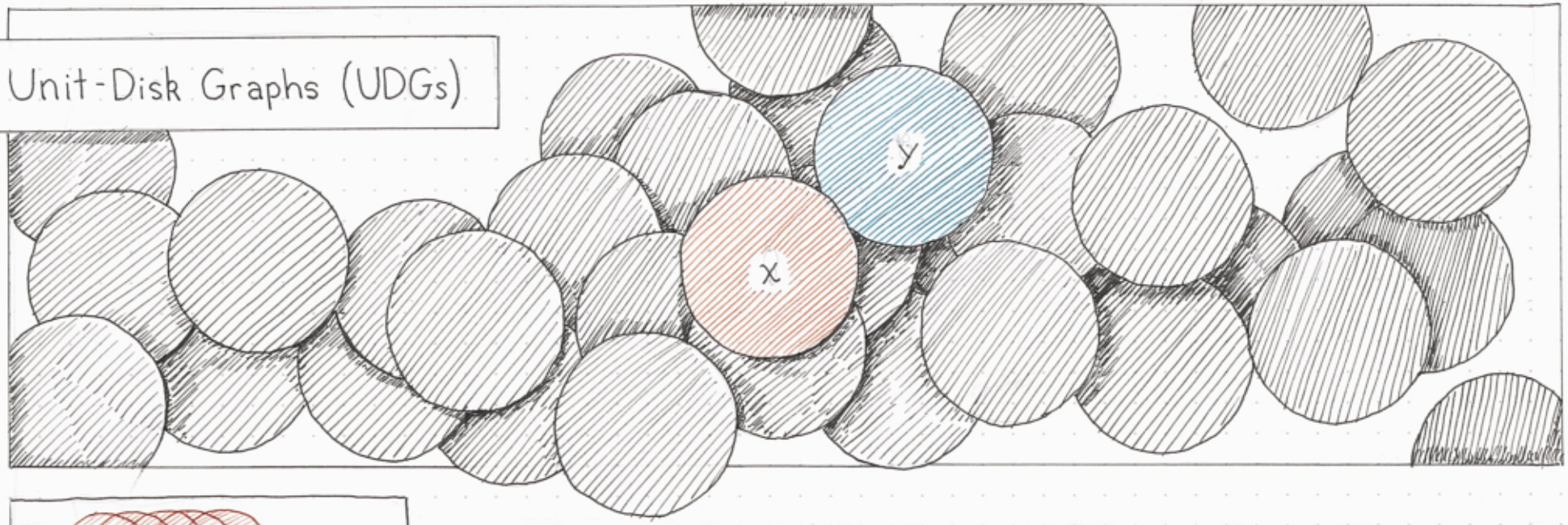
Proof:

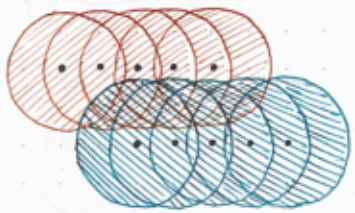
- free →

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

[OPS'22]: "Gyárfás decomposition"

Unit-Disk Graphs (UDGs)





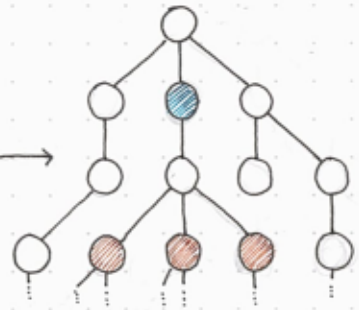
GREATER-THAN

remove

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

Proof:

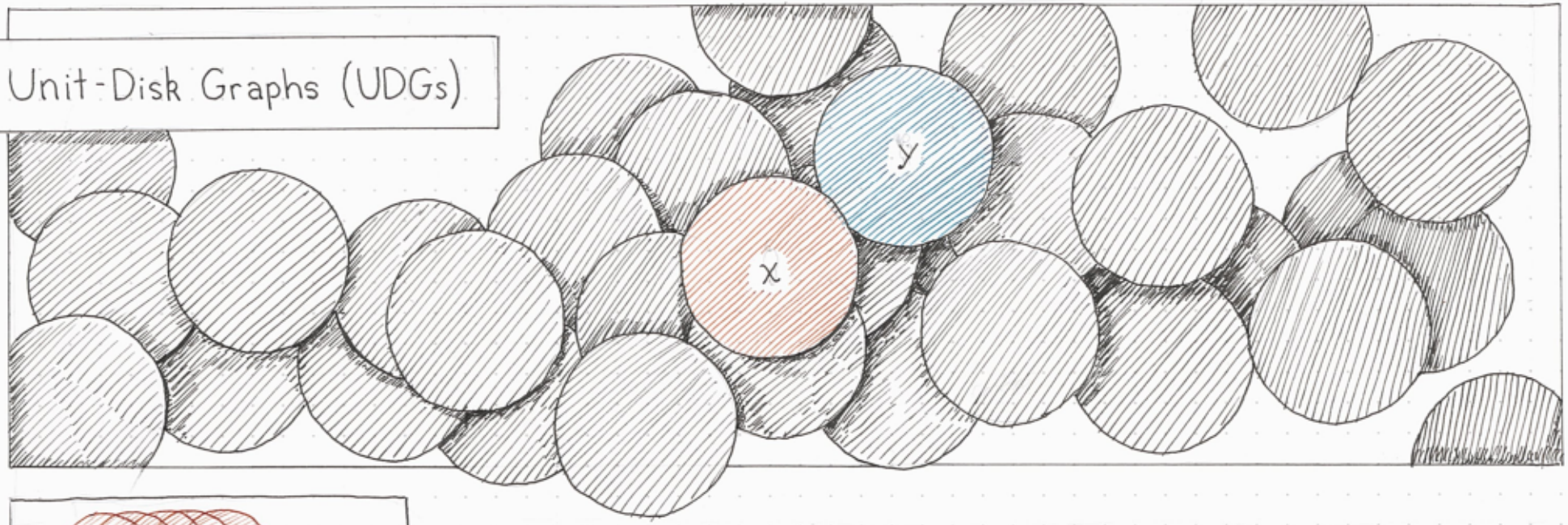
- free

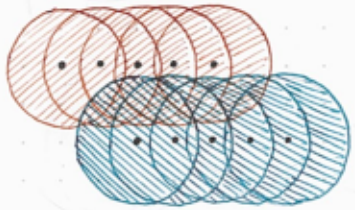


- Ramsey's theorem
- ↓
- Recurse on subgraph,
- shrink GREATER-THAN

[OPS'22]: "Gyárfás decomposition"

Unit-Disk Graphs (UDGs)





GREATER-THAN

remove

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

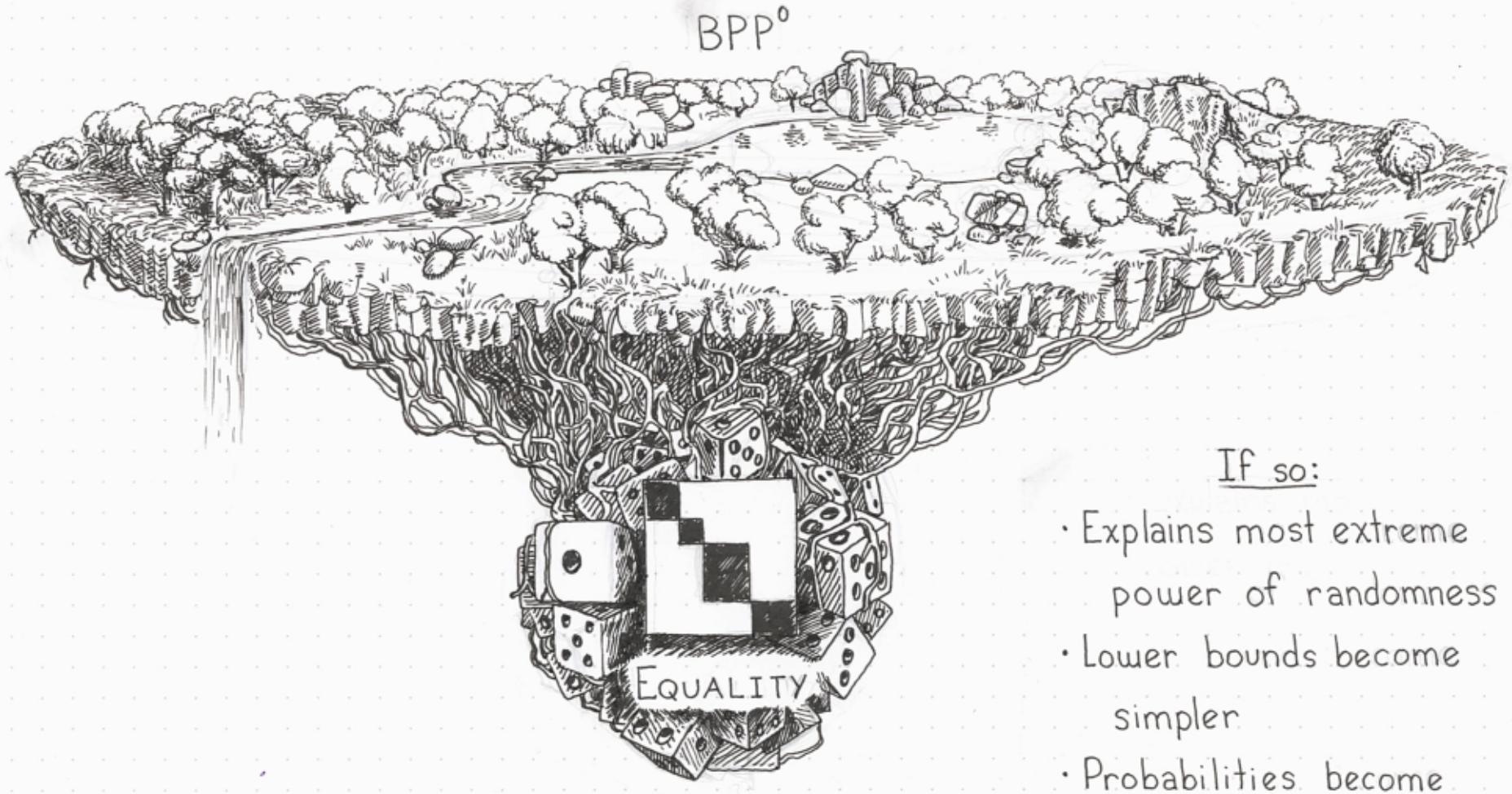
Works also for:

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

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

Only uses EQUALITY!

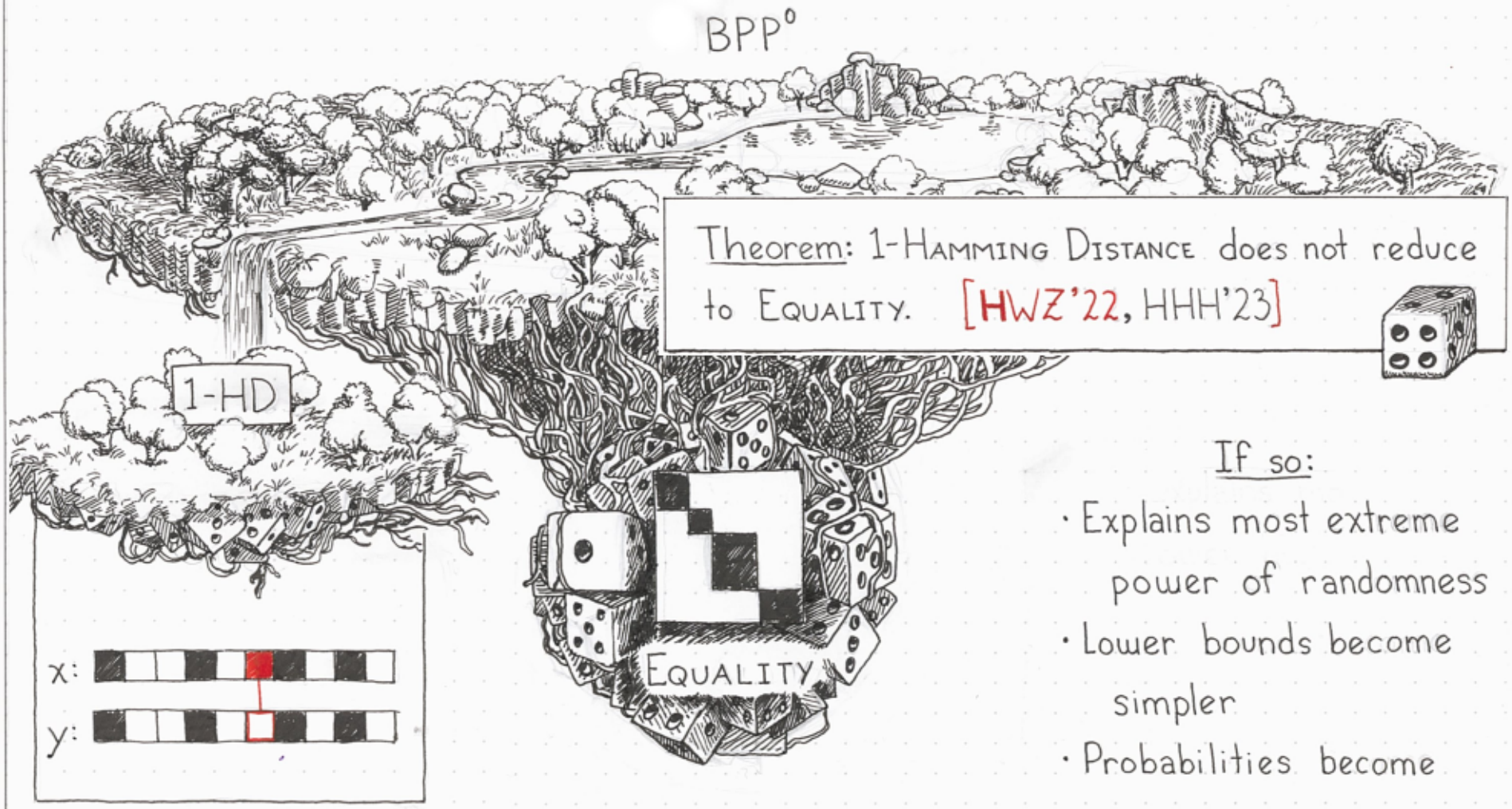
Is EQUALITY "complete" for constant-cost communication?



If so:

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

Is EQUALITY "complete" for constant-cost communication?



Theorem: 1-HAMMING DISTANCE does not reduce to EQUALITY. [HWZ'22, HHH'23]

If so:

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

... Then, is 1-Hamming Distance complete for 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 BPP°.

Theorem: The k-Hamming Distance problems form an infinite hierarchy within BPP°. [FHHH'24]

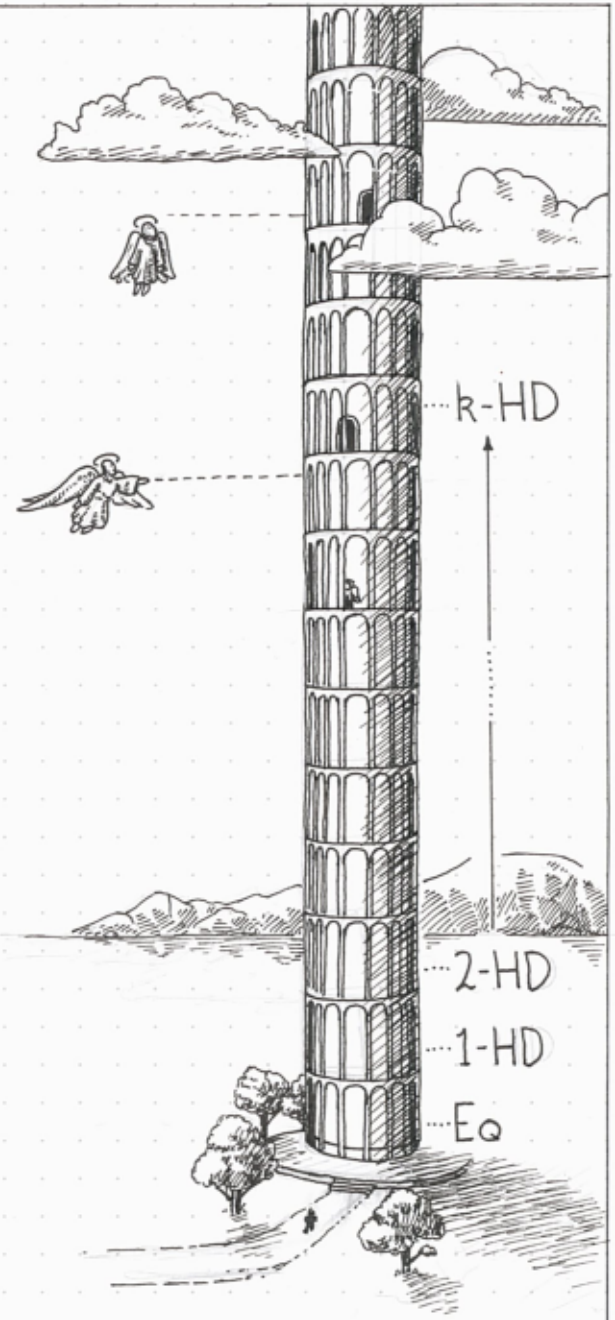
① Every oracle in BPP° is stable. → } O(1)

② k-Hamming Distance is permutation-invariant:

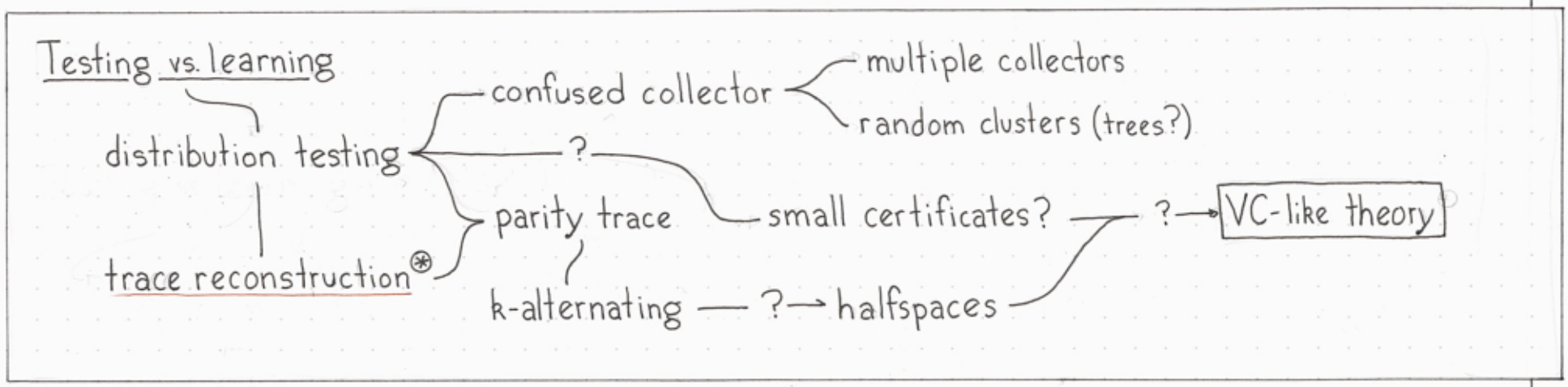
x :
y :

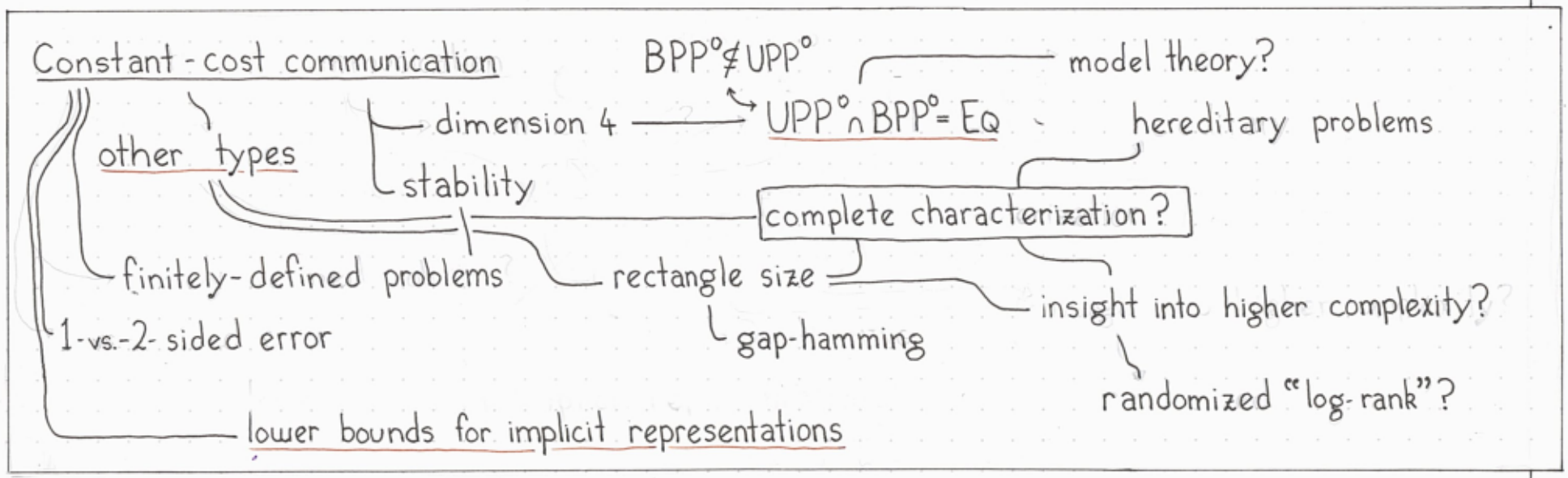
If the oracle queries are not permutation-invariant:
→ Hypergraph Ramsey theorem ⇒ stability is violated.

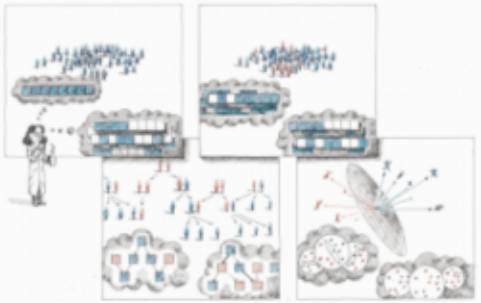
③ One query computes distance k vs. k+2 for all weight $n/2$ strings. Stability is violated.



Future work

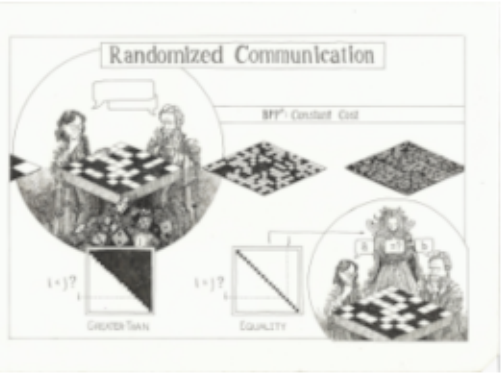

Testing vs. learning
distribution testing
trace reconstruction*
confused collector
multiple collectors
random clusters (trees?)
?
parity trace
small certificates?
?
VC-like theory
k-alternating
?
halfspaces

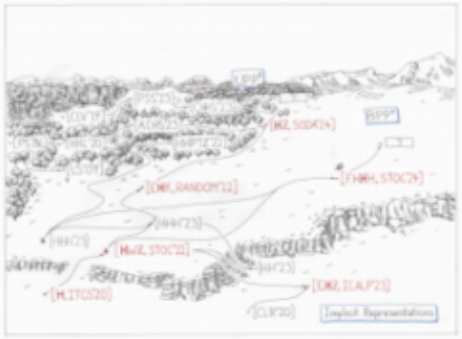

Constant-cost communication
other types
dimension 4
stability
BPP° ⊄ UPP°
UPP° ∩ BPP° = EQ
model theory?
hereditary problems
complete characterization?
finitely-defined problems
rectangle size
gap-hamming
insight into higher complexity?
randomized "log-rank"?
1-vs.-2-sided error
lower bounds for implicit representations

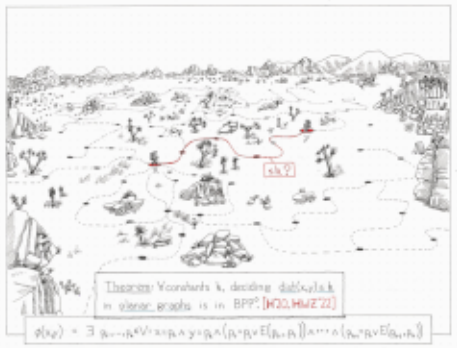


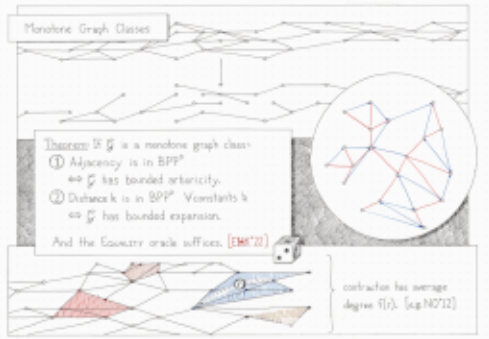

gap
input
Property Testing

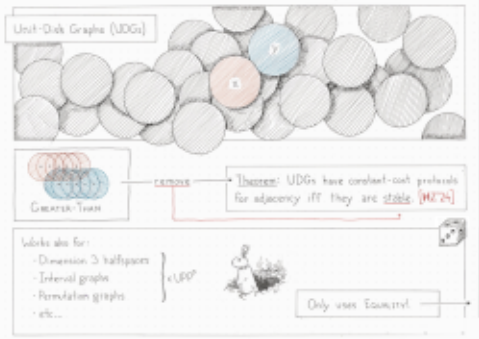

Distribution Testing
Sample size?
...with a confused collector
cluster
representative

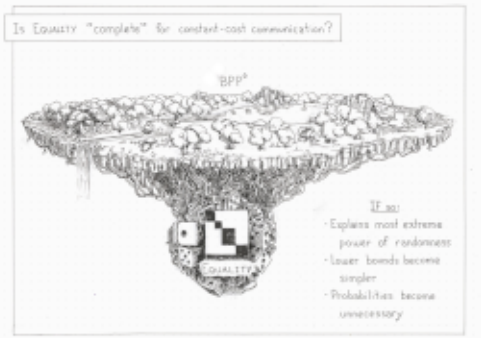

Randomized Communication
BPP*: Constant Cost
GREATER-THAN
EQUALITY

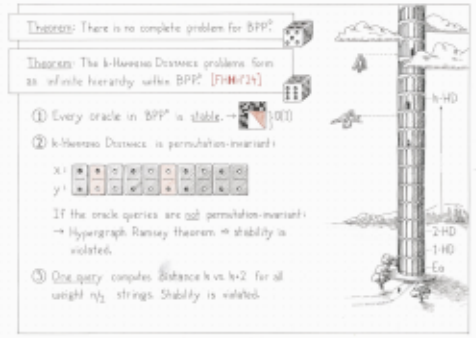

BPP*
[MZ, SODA'24]
[FHHH, STOC'24]
[EMK, RANDOM'22]
[MWZ, STOC'22]
[EMZ, ICALP'23]
Implicit Representations

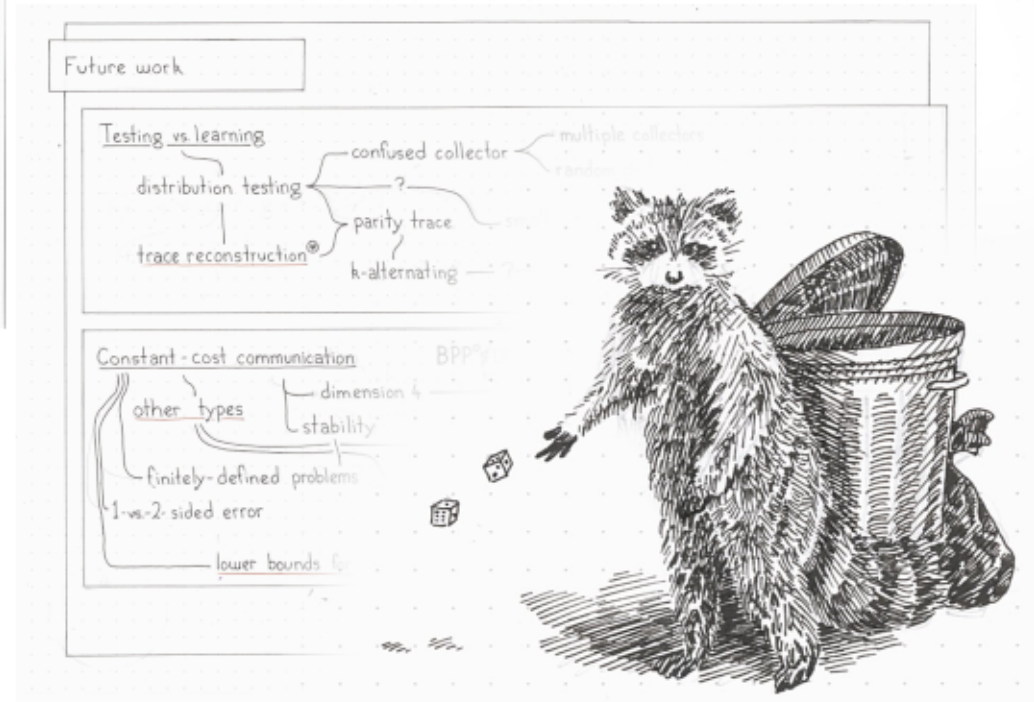

Theorem: ∀ constants k, deciding dist(x,y)≤k
in planar graphs is in BPP*. [HIZ0, HMZ'22]

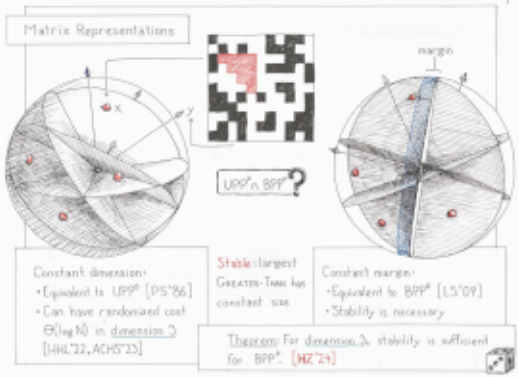

Monotone Graph Classes
Theorem: If 𝒢 is a monotone graph class:
① Adjacency is in BPP*
⇔ 𝒢 has bounded arboricity.
② Distance k is in BPP*. ∀ constants k
⇔ 𝒢 has bounded expansion.
And the Equality oracle suffices. [EMK'22]
contraction has average degree f(r). [Leg NO'12]

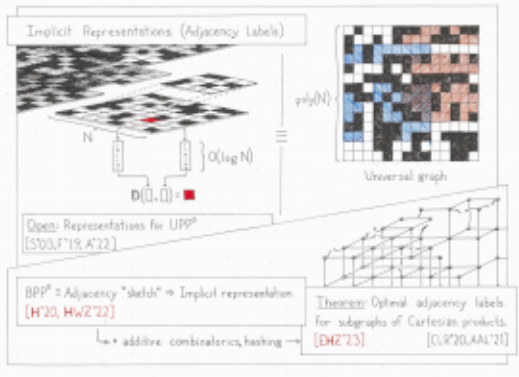

Unit-Disk Graphs (UDGs)
GREATER-THAN
reduces
Theorem: UDGs have constant-cost protocols
for adjacency iff they are stable. [MZ'24]
Works also for:
· Dimension 3 halfspaces
· Interval graphs
· Permutation graphs
· etc...
<UPP*
Only uses Equality!


Is Equality "complete" for constant-cost communication?
BPP*
EQUALITY
If so:
· Explains most extreme power of randomness
· Lower bounds become simpler
· Probabilities become unnecessary

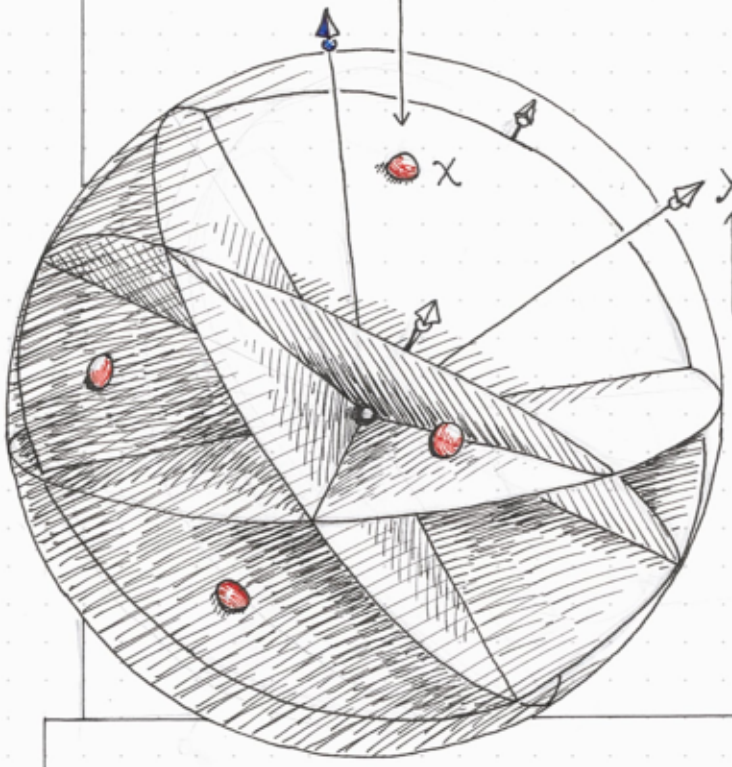

Theorem: There is no complete problem for BPP*.
Theorem: The k-Hamming Distance problems form
an infinite hierarchy within BPP*. [FHH'24]
① Every oracle in BPP* is stable.
② k-Hamming Distance is permutation-invariant.
If the oracle queries are not permutation-invariant:
→ Hypergraph Ramsey theorem ⇒ stability is violated.
③ One query computes distance k vs k+2 for all
weight n/2 strings. Stability is violated.
k-HD
2-HD
1-HD
Eq

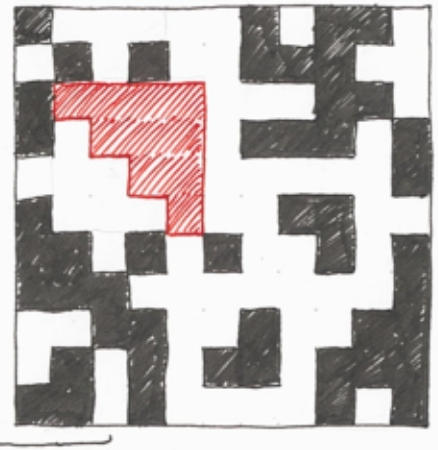

Future work
Testing vs. learning
distribution testing
trace reconstruction®
confused collector
multiple collectors
?
parity trace
k-alternating
Constant-cost communication
other types
dimension 4
stability
finitely-defined problems
1-vs.-2-sided error
lower bounds


Matrix Representations
margin
UPP* ∩ BPP*?
Constant dimension:
· Equivalent to UPP* [PS'86]
· Can have randomized cost Θ(log N) in dimension 3
[HHL'22, ACHS'23]
Stable: largest Greater-Than has constant size.
Constant margin:
· Equivalent to BPP* [LS'09]
· Stability is necessary
Theorem: For dimension 3, stability is sufficient for BPP*. [MZ'24]

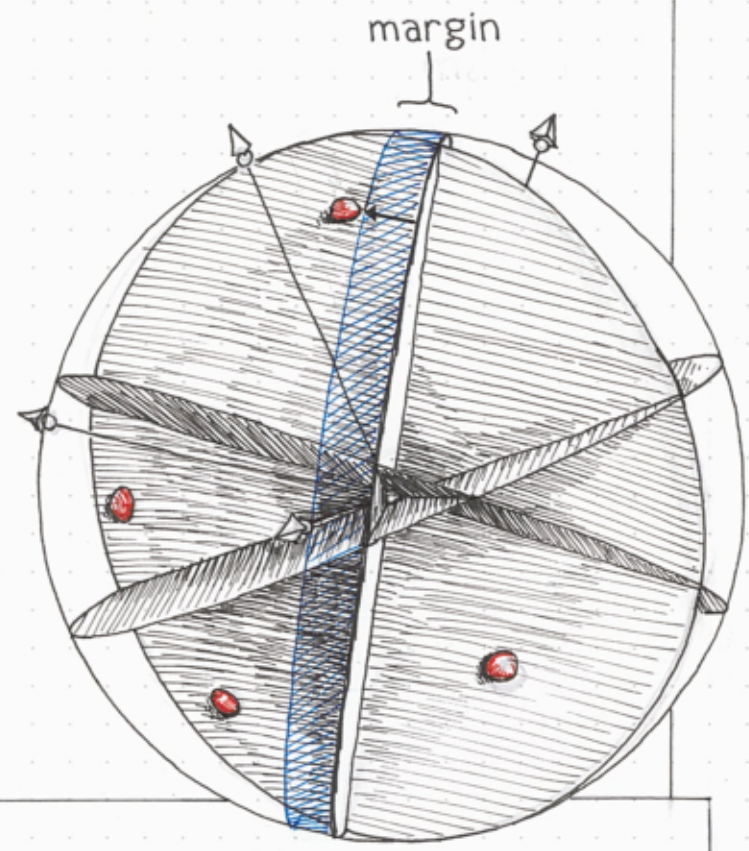

Implicit Representations (Adjacency Labels)
N
O(log N)
poly(N)
Universal graph
Open: Representations for UPP*
[SO03, FT19, A'22]
BPP* = Adjacency "sketch" ⇒ Implicit representation
[M'20, HMZ'22]
+ additive combinatorics, hashing
Theorem: Optimal adjacency labels
for subgraphs of Cartesian products.
[EMZ'23]
[CLW'20, KAL'21]

Matrix Representations





UPP^0 ∩ BPP^0 ?



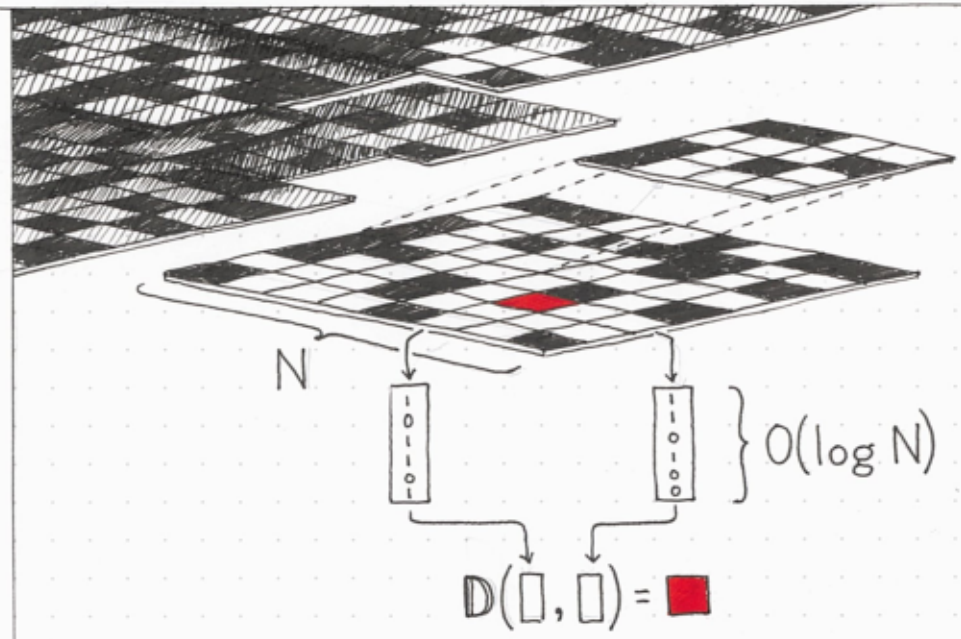
Constant dimension:
- Equivalent to UPP^0 [PS'86]
- Can have randomized cost $\Theta(\log N)$ in dimension 3 [HHL'22, ACHS'23]

Stable: largest GREATER-THAN has constant size

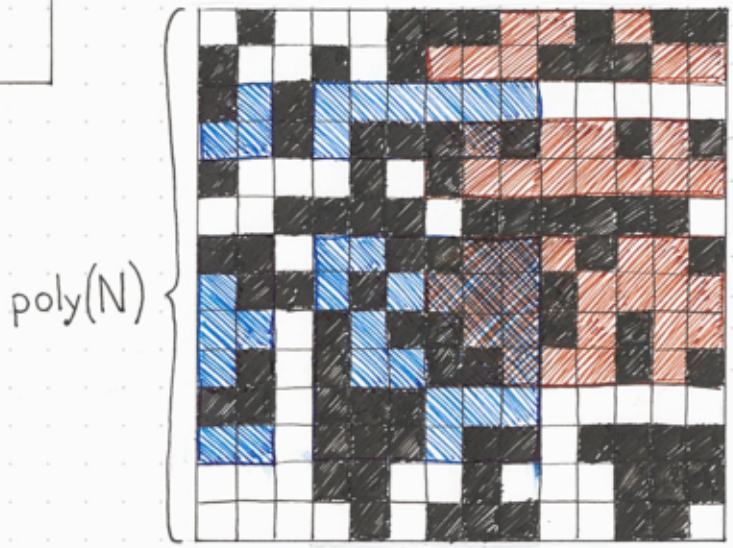
Constant margin:
- Equivalent to BPP^0 [LS'09]
- Stability is necessary

Theorem: For dimension 3, stability is sufficient for BPP^0. [HZ'24]

Implicit Representations (Adjacency Labels)

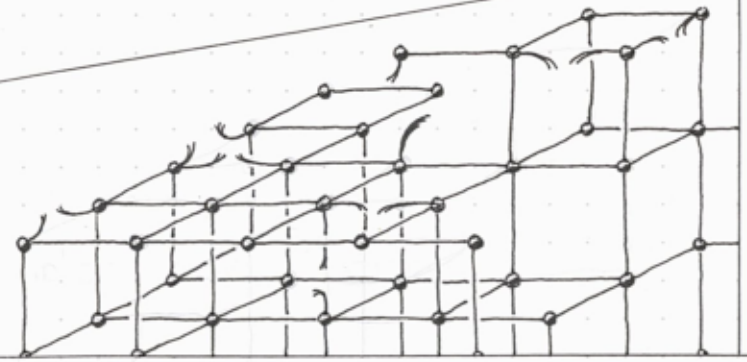


≡



Universal graph

Open: Representations for UPP^0
[S'03, F'19, A'22]



BPP^0 ≡ Adjacency "sketch" ⇒ Implicit representation
[H'20, HWZ'22]

↳ + additive combinatorics, hashing →

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

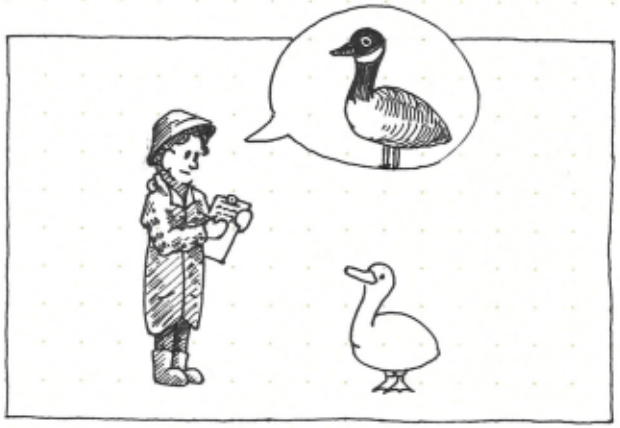
Part I: Adversarial Clusters

Impossible!

1. Restrict clustering to $\mathcal{U}$

2. Replace TV with earth-mover (EMD)

Domain → metric space

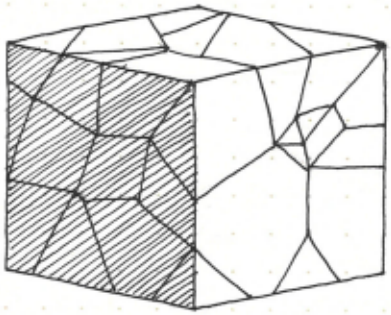


3. Queries

4. Cluster-Reject

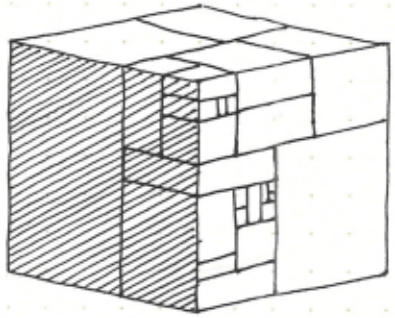
$\mathcal{U}$: Universe

E.g.



} Convex

$G \subseteq \mathcal{U}$: "Good" clusterings — not allowed!

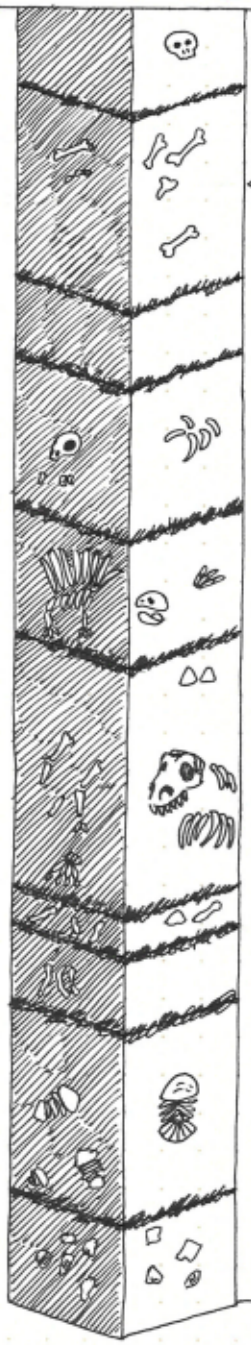


} depends on distribution

High-probability of low diameter boxes, decision trees

Result: learning cells is not necessary

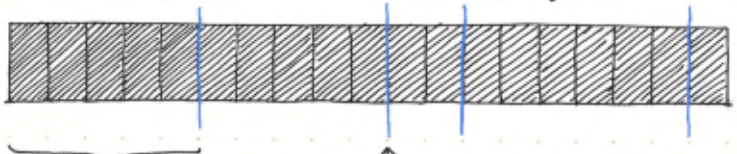
Part II: Random Clusters



Motivations:

- Environmental randomness
- Randomized classifier training

We study: Testing uniformity,



Standard:

X: histogram 

Analyze $X^T \boxed{I} X - \|X\|_1$

Naïve (known clusters):

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

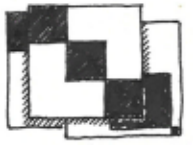
Without queries, $\left(\rho \geq \tilde{\Omega}\left(n^{-1/5}\varepsilon^{-4/5}\right)\right)$

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

With queries

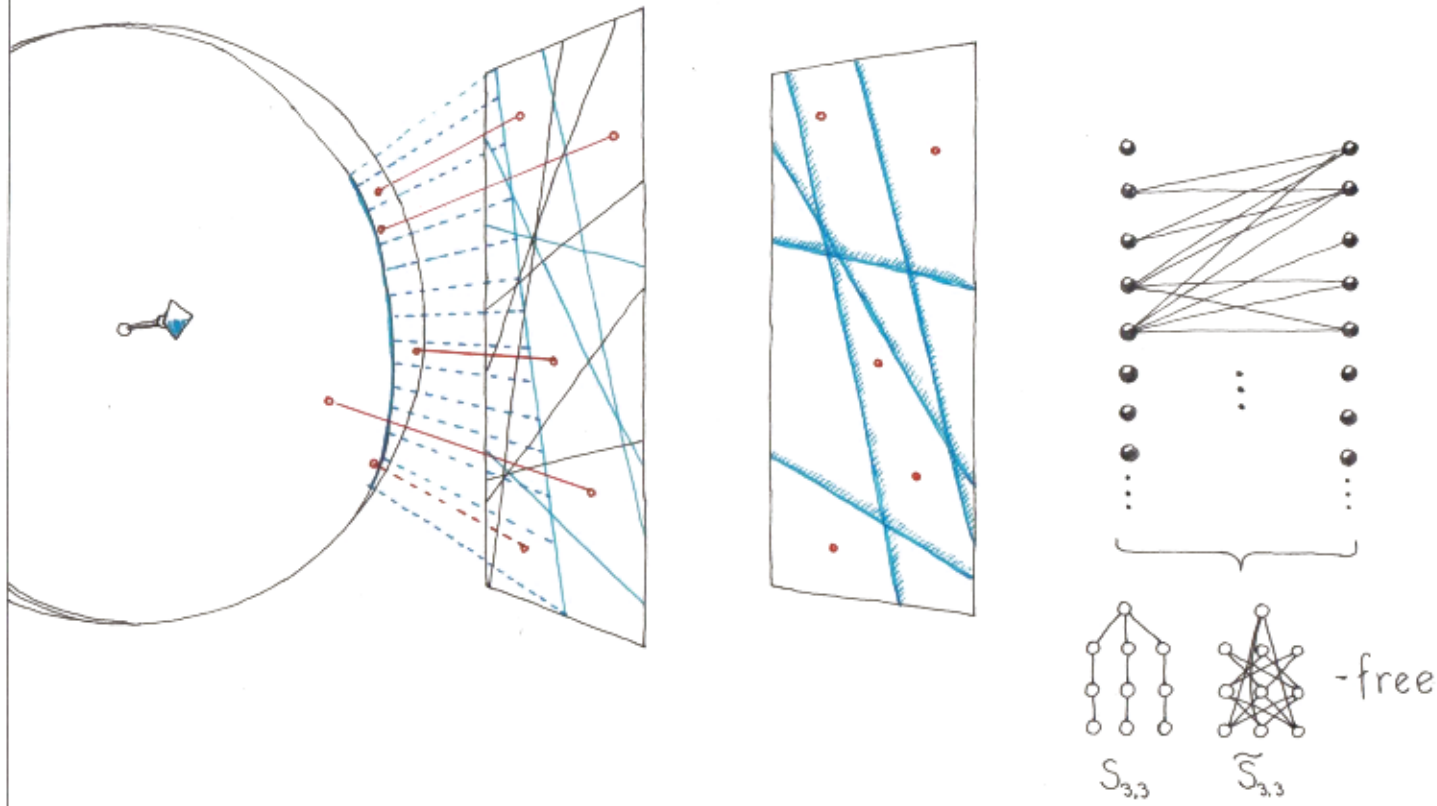
$$O\left(\frac{\sqrt{n}}{\rho\,\varepsilon^2}\right)$$ Use [VV'17]

Now:



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

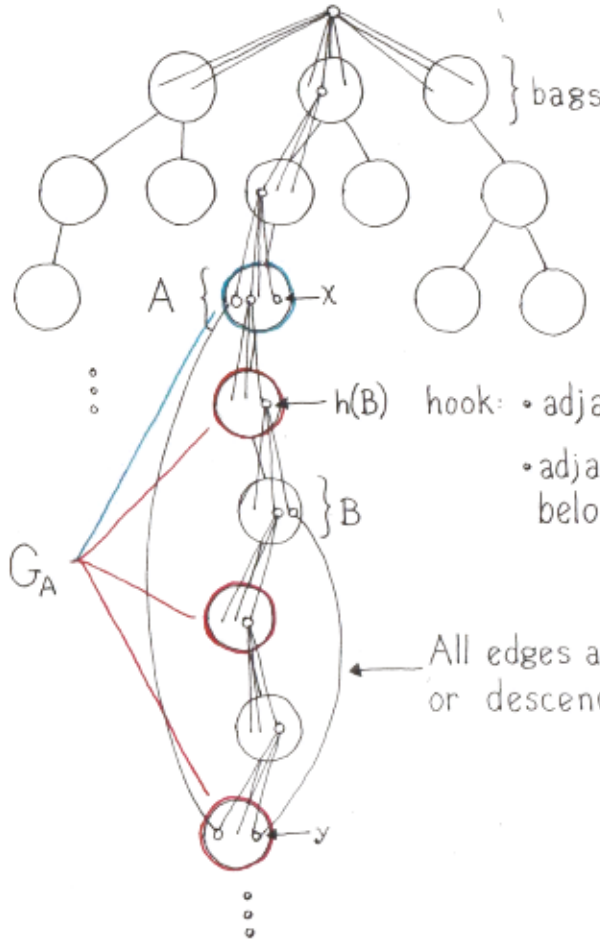
Step 1



Lemma:

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

Step 2

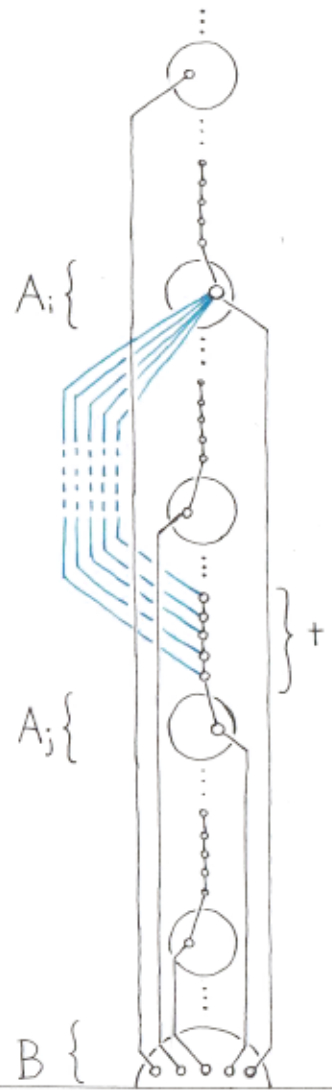


hook: • adjacent to all $v \in B$
• adjacent to nothing below B

All edges are to ancestors or descendants

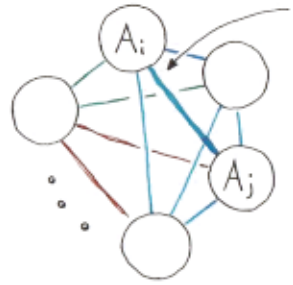
Gyárfás decomposition [POS'22]

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



Step 3

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



Color: induced subgraph encoding

Ramsey's theorem + case analysis on each color : find large GREATER-THAN, or $S_{s,t}$.

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