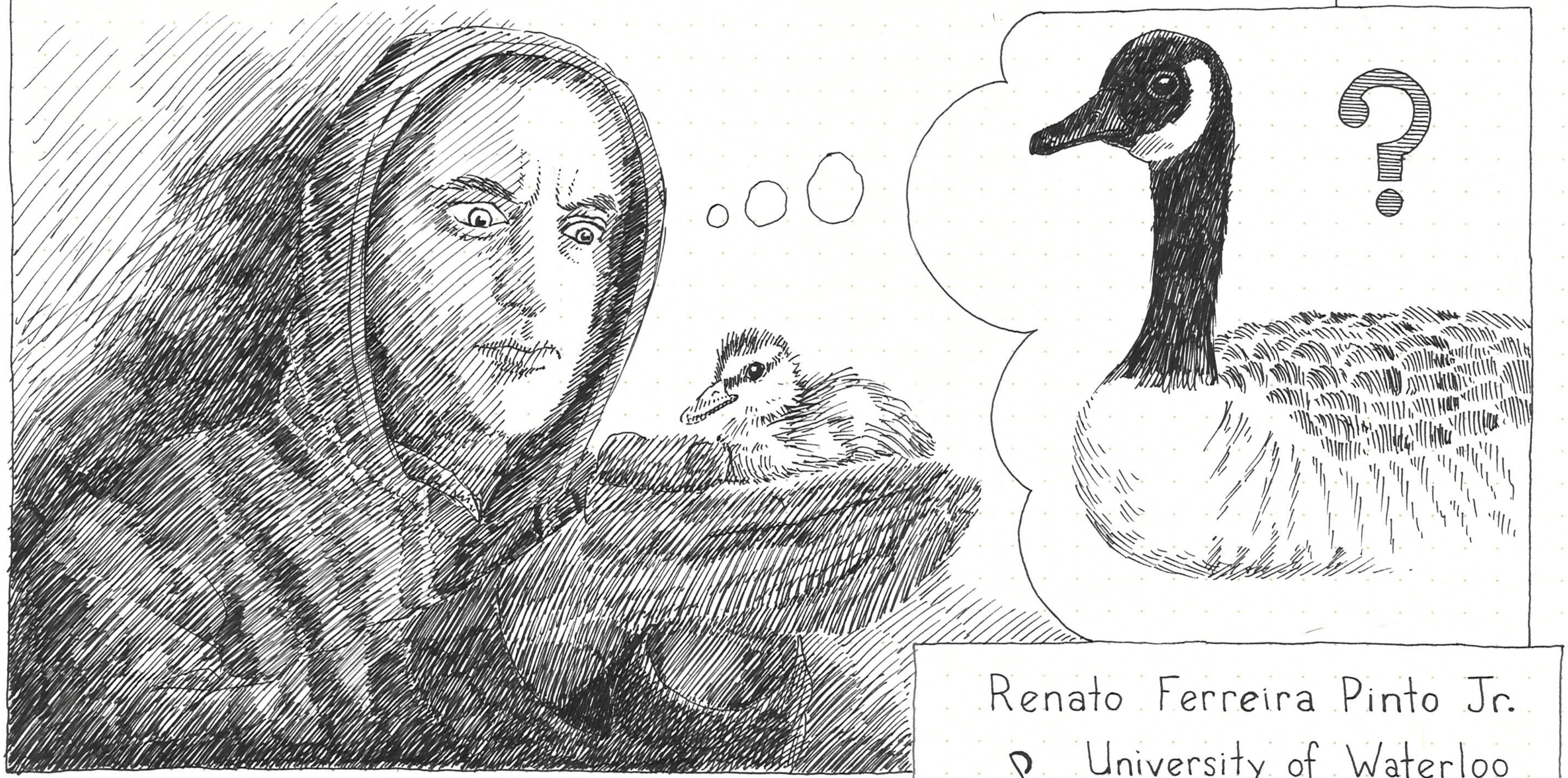
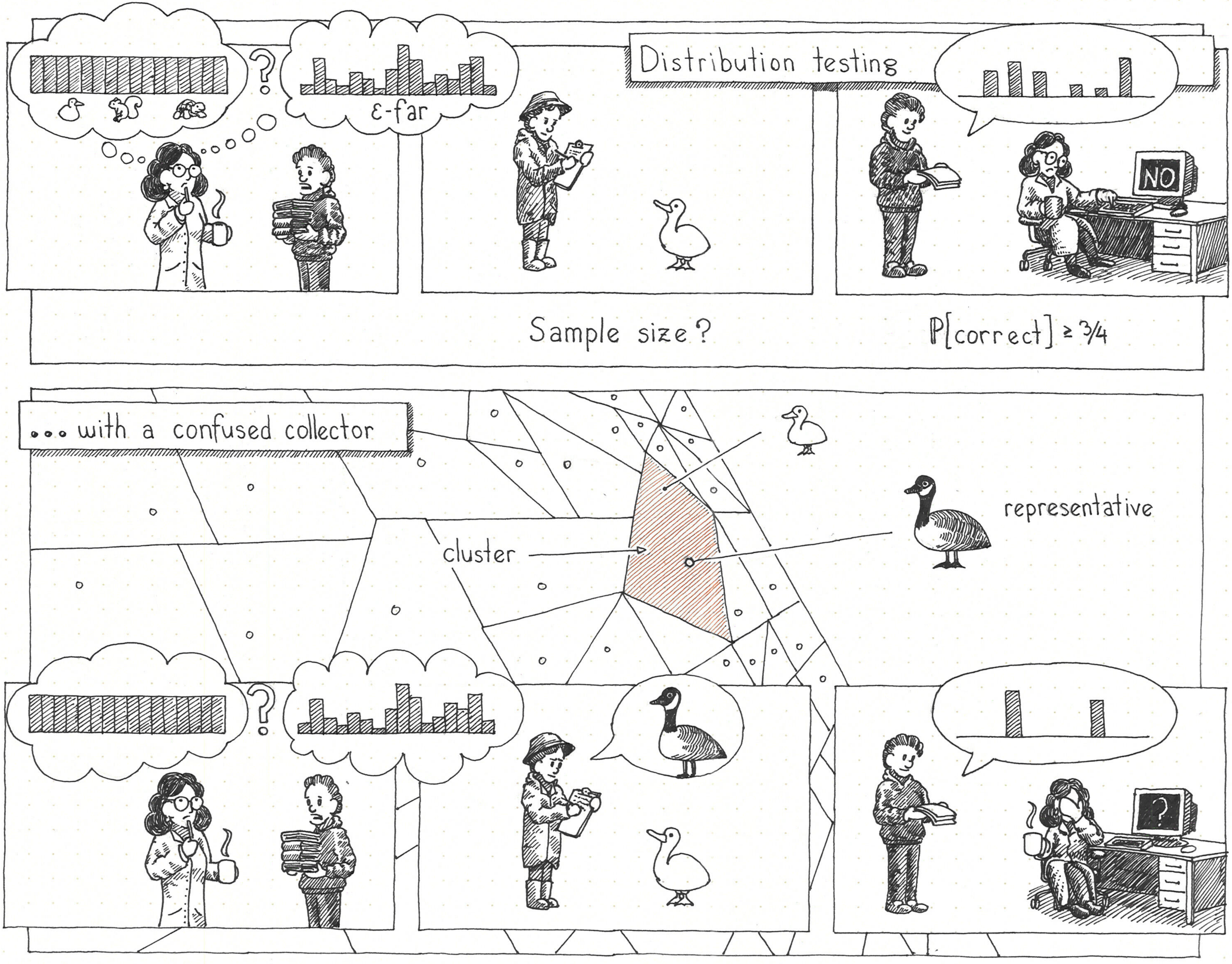
DISTRIBUTION TESTING with a CONFUSED COLLECTOR



Renato Ferreira Pinto Jr.
University of Waterloo

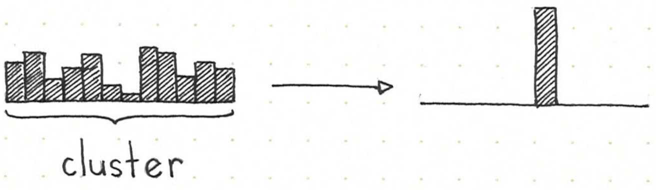
&

Nathan Harms
EPFL


?
ε-far
Distribution testing
NO
Sample size?
ℙ[correct] ≥ 3/4
... with a confused collector
cluster
representative
?
?

Part I: Adversarial Clusters

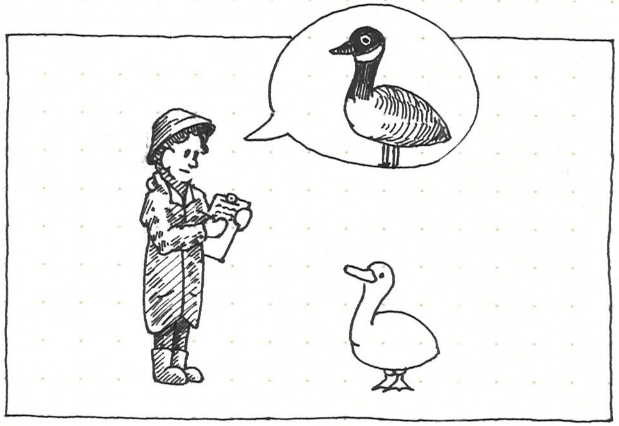
Impossible!



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

2. Replace TV with earth-mover (EMD)

Domain → metric space

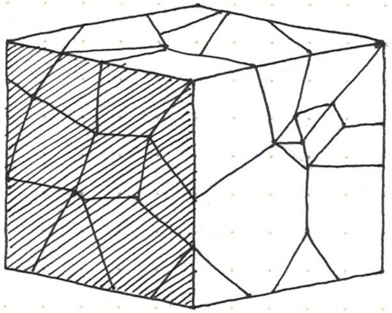


3. Queries

4. Cluster-Reject

$\mathcal{U}$: Universe

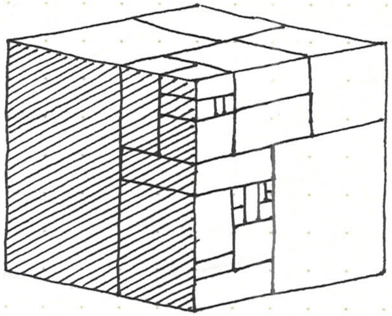
E.g.



Convex

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

not allowed!

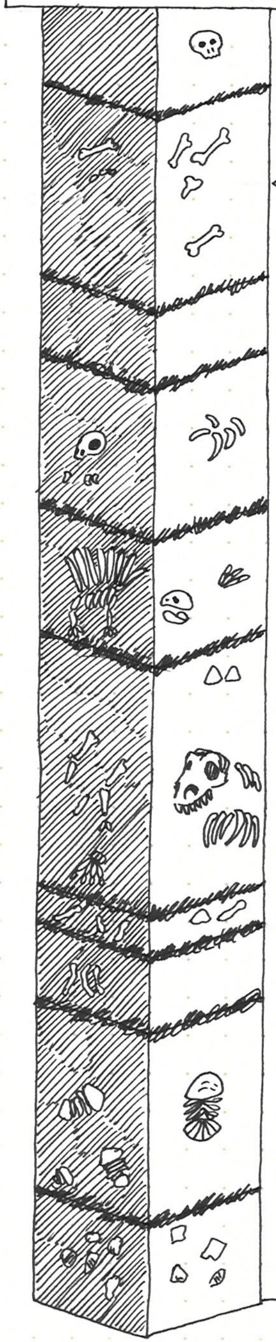


depends on distribution

High-probability of low diameter boxes, decision trees

Result: learning cells is not necessary

Part II: Random Clusters



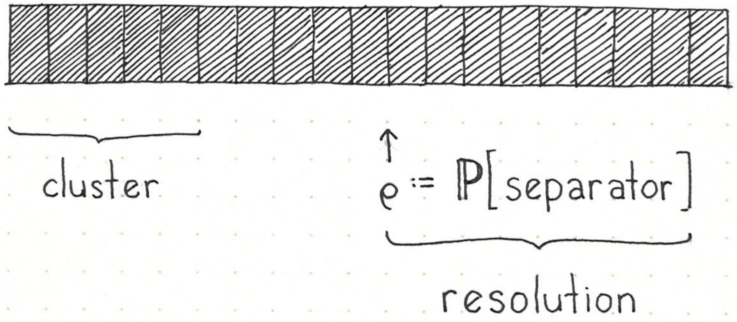
Motivations:

- Environmental randomness
- Randomized classifier training

Naïve (known clusters):

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

We study: Testing uniformity,



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

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

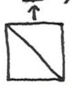
With queries

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

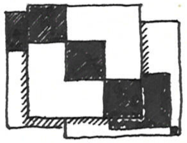
Standard:

X: histogram

Analyze $X^{T} I X - \|X\|_{1}$



Now:



Analyze $X^{T} \Phi X - \|X\|_{1}$