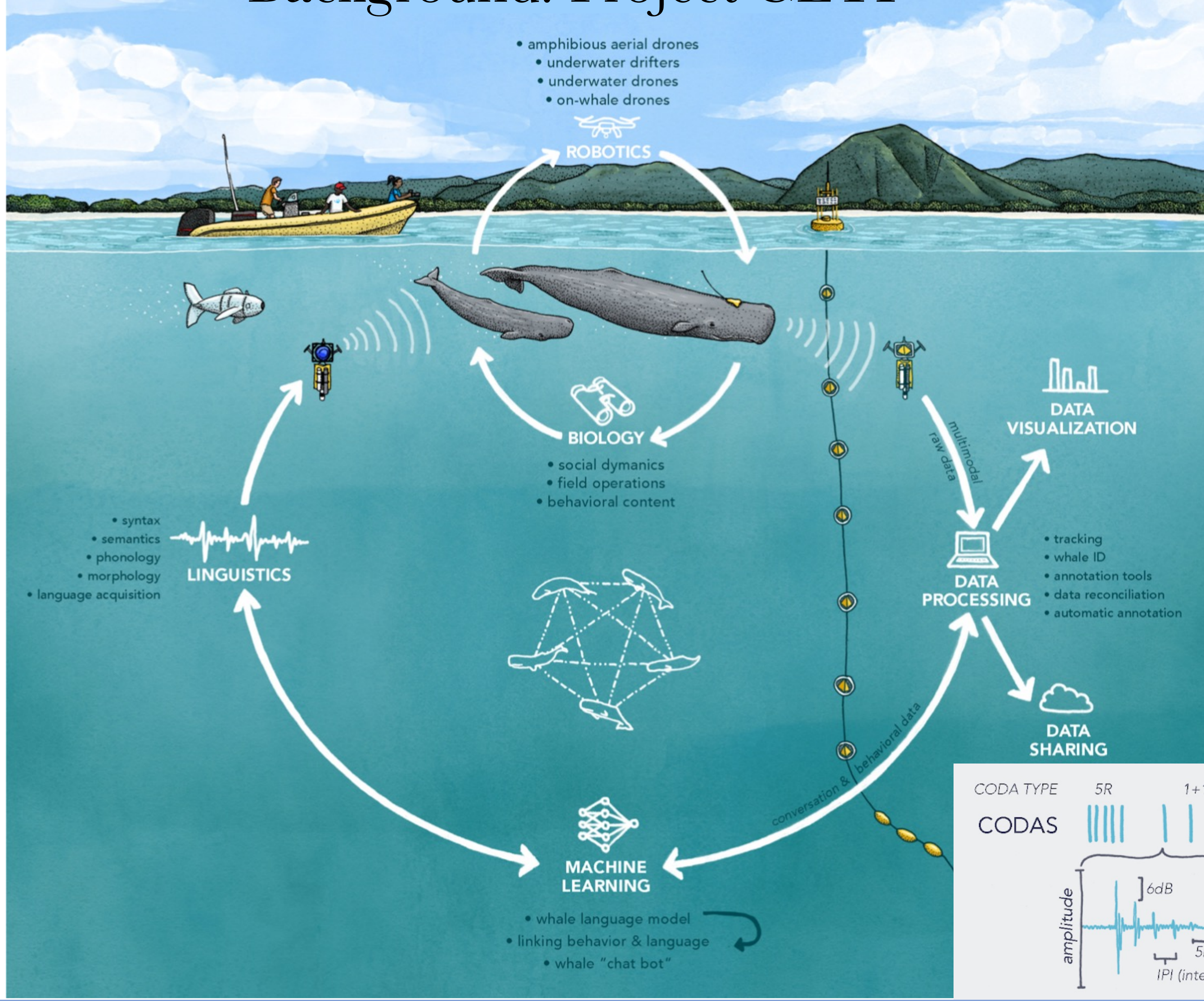
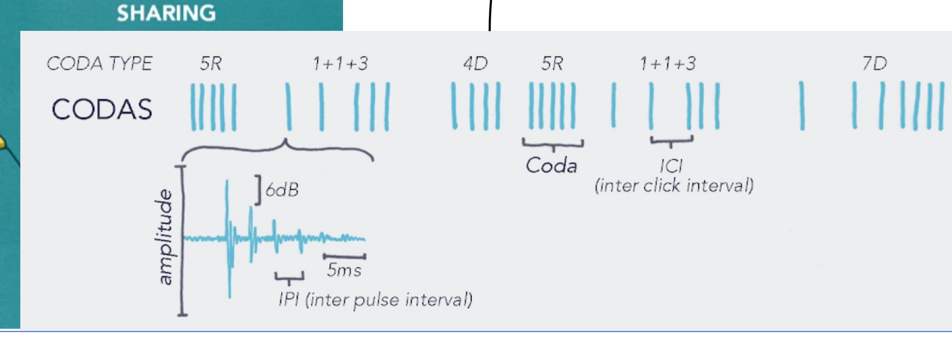
A Theory of Unsupervised Translation Motivated by Understanding Animal Communication

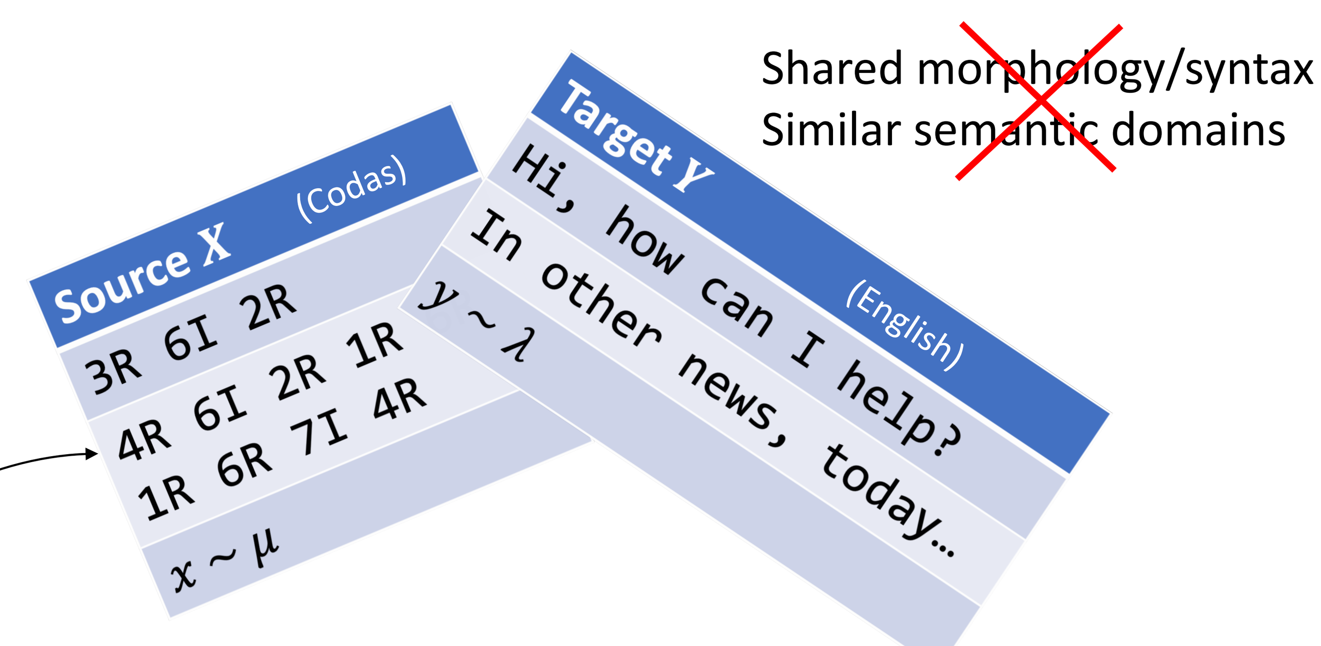
Shafi Goldwasser David F. Gruber Adam Tauman Kalai Orr Paradise

Background: Project CETI





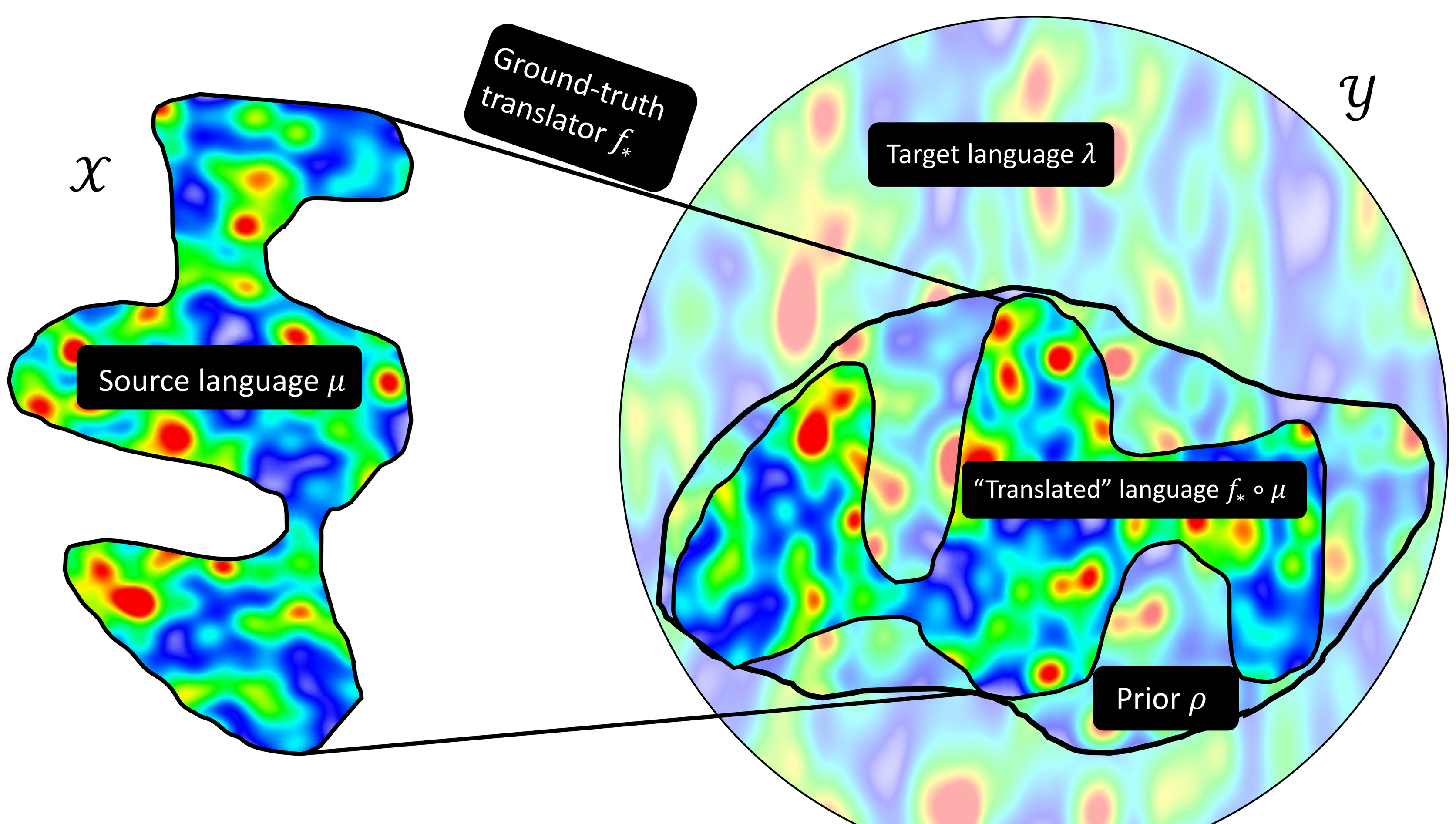
Unsupervised Machine Translation (UMT)



Given **unlabeled** samples, $x_1, x_2, \dots, y_1, y_2, \dots$
learn $f_\theta : X \to Y$ with small error $\mathcal{L}(\theta) \coloneqq \mathbb{E}_{x\sim\mu}[\ell(f_\star(x), f_\theta(x))]$

Hypothetical ground-truth translator $f_\star : X \to Y$

Main Tool: Prior over Translations

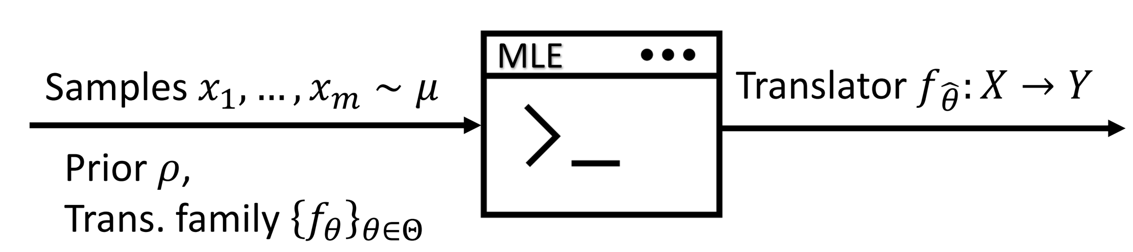


Setup. Assume access to prior over translations $\rho \approx f_\star \circ \mu$

Definition (γ-ambiguities). $\Theta_\gamma(\rho) \coloneqq \left\{\theta \in \Theta : \mathbb{P}_{x\sim\mu}[\rho(f_\theta(x)) = 0] \le \gamma\right\}$, $\varepsilon_\gamma(\rho) \coloneqq \max_{\theta\in\Theta_\gamma(\rho)}(\mathcal{L}(\theta))$

Assumption (Realizability). $\star \in \Theta_0(\rho)$

Sample-complexity bounds for UMT with a prior



Theorem: For any $\mu, f_\star, \rho$ such that $\star \in \Theta_0(\rho)$, and $\delta, \gamma \in (0,1)$.

With probability $\ge 1-\delta$, given $m \ge \frac{1}{\gamma}\ln\frac{|\Theta|}{\delta}$ samples from μ, and prior ρ,

MLE outputs $\hat\theta$ with $\mathcal{L}(\hat\theta) \le \varepsilon_\gamma(\rho)$.

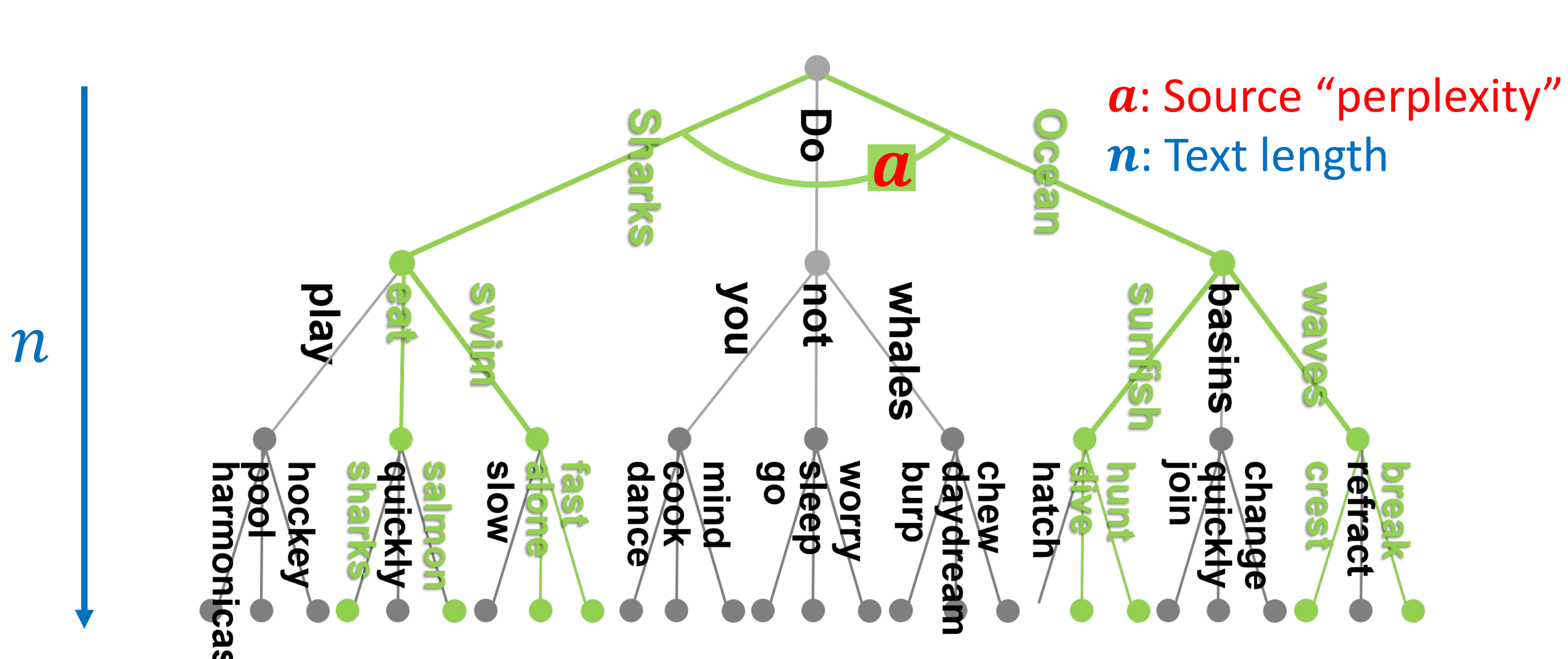
Comparable to Supervised MT: $m \ge \frac{1}{\varepsilon}\ln\frac{|\Theta|}{\delta}$ for error ε (Occam's Razor)

Information-theoretic bound! Optimizing $\hat\theta$ may be hard…

$\mathcal{L}(\hat\theta) \xrightarrow{m\to\infty} \varepsilon_0(\rho)$
"Best translator (up to ambiguities)"

Studying translatability in two stylized models of language

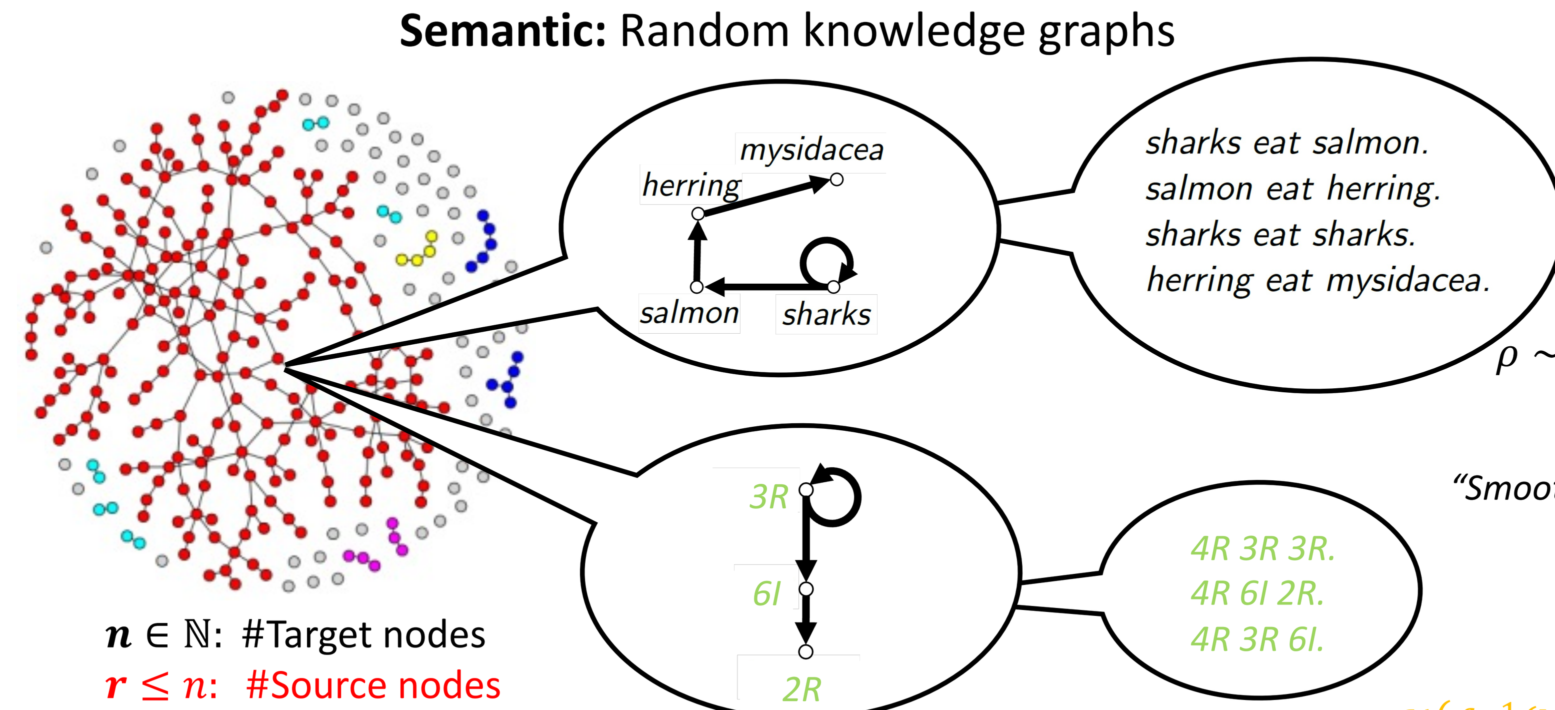
Syntactic: Random sentence trees



$\rho \sim \mathcal{U}(\text{Tree})$ $\mu \sim \mathcal{U}(f_\star^{-1}(\text{Subtree}))$

0-1 loss for m samples: $\mathcal{L}_{01}(\hat\theta) = O\left(\max\left(\frac{1}{m}, \frac{1}{a^n}\right)\cdot \log|\Theta|\right)$

Semantic: Random knowledge graphs



$\rho \sim \frac{1}{2}\mathcal{U}(\text{Graph}) + \frac{1}{2}\mathcal{U}(\text{All})$
"Smoothed uniform"

$\mu \sim \mathcal{U}(f_\star^{-1}(\text{Noisy subgraph}))$

$n \in \mathbb{N}$: #Target nodes
$r \le n$: #Source nodes
$d \le r$: Average degree.
$\alpha \in (0,1)$: Agreement (α-correlated Erdős–Rényi graphs)

$$\mathcal{L}_{01}(\hat\theta) = O\left(\frac{\log n}{\alpha^2 d} + \frac{1}{\alpha}\sqrt{\frac{r \log n}{m}}\right)$$

Full version