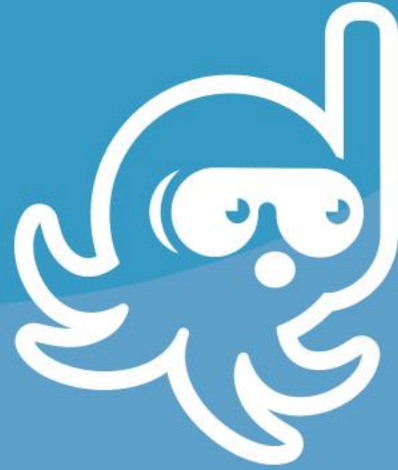


# **Snorkel: Rapid Training Data Creation with Weak Supervision**



**Hira Cansu Değirmenbaşı**

**191101053**

**Hüseyin Emre Aksoy**

**191101036**

# Snorkel Nedir ve Neden kullanılır?

- Snorkel, hızlı bir şekilde eğitim verileri oluşturmak için Weak Supervision kaynaklarını birleştiren ilk uçtan uca sistem.
- Eğitim verilerini etiketlemek, makine öğrenimi sistemlerini devreye almada giderek artan en büyük darboğaz haline geliyor.
- Snorkel önceki buluşsal yaklaşımlara göre ortalama %132 tahmin performansında iyileştirme sağlar.

# Snorkel'in dizayn prensipleri

- Bring All Sources to Bear: sistem bütün weak supervision kaynaklarından gelen label ları kullanmaya hazır olmalıdır.
- Training Data as the Interface to ML: Sistem, her veri noktası için tek bir olasılık etiketi üretmek üzere etiket kaynaklarını modellemeli ve bu kaynakların ötesinde genelleme yapmak için çok çeşitli sınıflandırıcılardan herhangi birini eğitmelidir.
- Supervision as Interactive Programming: Sistem, kullanıcı denetimine yanıt olarak hızlı sonuçlar sağlamalıdır. Makine öğrenimi için weak supervision ı REPL benzeri olarak tasavvur ediyoruz.

# SNORKELE Mimari

- Writing Labeling Functions: Snorkel kullanıcıları, eğitim verilerini elle etiketlemek yerine, çeşitli zayıf denetim kaynaklarını ifade etmelerine olanak tanıyan etiketleme işlevleri yazar.
- Modeling Accuracies and Correlations: Snorkel, etiketleme işlevleri üzerinden otomatik olarak üretken bir model öğrenir ve bu, bunların doğruluklarını ve korelasyonlarını tahmin etmesine olanak tanır.
- Training a Discriminative Model: Snorkelin çıktısı, çok çeşitli son teknoloji makine öğrenimi modellerini eğitmek için kullanılabilen bir dizi olasılık etiketidir.

# Snorkelin yapısı

- Hand-Defined Labeling Functions: Bizim el ile tanımladığımız arbitrary python kodlarıyla oluşturulur.
- Declarative Labeling Functions: Snorkel, en yaygın zayıf denetim işlevi türlerini kodlayan bir declarative operasyonlar içeren kitaplığı içerir.

# Modelleme Yapısı

## Structure Learning:

- Birbirinin varyasyonu olan fonksiyonlar arasında,
- Benzer girdiler üzerinde çalışan fonksiyonlar,
- İlişkili bilgi kaynaklarını kullanan işlevler

# Hand Model Neden Pratik Değildir ?

- 1) Uzman olmayan birinin bağımlılıkları belirlemesi zordur.
- 2) Etiketleme işlevleri tekrar ettikçe bağımlılıklar değişkenlik gösterebilir..
- 3) Bağımlılık yapısı veri setine özgün ise başka bir veri setinde değişiklikler olabilir..

# Otomatik Olarak Model Seçme

## TRADEOFF

PREDICTIVE PERFORMANCE X COMPUTATION COST

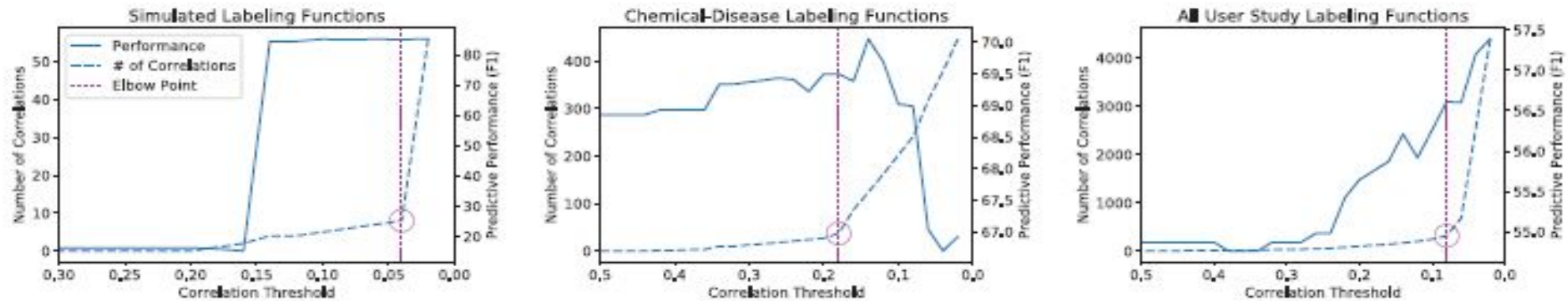


Figure 5: Predictive performance of the generative model and number of learned correlations versus the correlation threshold  $\epsilon$ . The selected elbow point achieves a good tradeoff between predictive performance and computational cost (linear in the number of correlations). Left: simulation of structure learning correcting the generative model. Middle: the CDR task. Right: all user study labeling functions for the Spouses task.



Snorkel

Generative Model

Discriminative Model

# Bazı Araştırmalar

Scientific Articles (Chem)

Electronic Health Records (EHR)

Chemical-Disease Relations (CDR)

Abnormality Detection in Lung Radiographs (Rad)

Crowdsourcing (Crowd)

# Relation Extraction from Text

Table 3: Evaluation of Snorkel on relation extraction tasks from text. Snorkel’s generative and discriminative models consistently improve over distant supervision, measured in F1, the harmonic mean of precision (P) and recall (R). We compare with hand-labeled data when available, coming within an average of 1 F1 point.

Task	Distant Supervision			Snorkel (Gen.)				Snorkel (Disc.)				Hand Supervision		
	P	R	F1	P	R	F1	Lift	P	R	F1	Lift	P	R	F1
Chem	11.2	41.2	17.6	78.6	21.6	33.8	+16.2	87.0	39.2	54.1	+36.5	-	-	-
EHR	81.4	64.8	72.2	77.1	72.9	74.9	+2.7	80.2	82.6	81.4	+9.2	-	-	-
CDR	25.5	34.8	29.4	52.3	30.4	38.5	+9.1	38.8	54.3	45.3	+15.9	39.9	58.1	47.3
Spouses	9.9	34.8	15.4	53.5	62.1	57.4	+42.0	48.4	61.6	54.2	+38.8	47.8	62.5	54.2

# Cross-Modal: Images & Crowdsourcing

Table 4: Evaluation on cross-modal experiments. Labeling functions that operate on or represent one modality (text, crowd workers) produce training labels for models that operate on another modality (images, text), and approach the predictive performance of large hand-labeled training datasets.

Task	Snorkel (Disc.)	Hand Supervision
Radiology (AUC)	72.0	76.2
Crowd (Acc)	65.6	68.8

# Araştırma Sonuçları

	SNORKEL	HAND
Ortalama performanslı Snorkel Kullanıcısı :	30.4 F1	20.9 F1
İyi performanslı Snorkel Kullanıcısı :	48.7 F1	29.5 F1
Kötü performanslı Snorkel Kullanıcısı :	12.0 F1	11.7 F1 ( en düşük)

# Sonuç..

## Hand-Labeled x Snorkel

Araştırmalar, Snorkel güçlü makine öğrenimi modellerini eğitmenin maliyetini ve zorluğunu önemli ölçüde azalttığını, önceki weak supervision yöntemlerini aştığını ve büyük, Hand-Labeled eğitim setlerinin kalitesine yaklaştığını gösteriyor.

**Dinlediğiniz için teşekkür ederiz..**

Snorkel