Snorkel: Easier-to-use Machine Learning Systems

Chris Ré

STANFORD
INFOLAB

Machine learning is **harder** than traditional programming, but it should be **easier**.

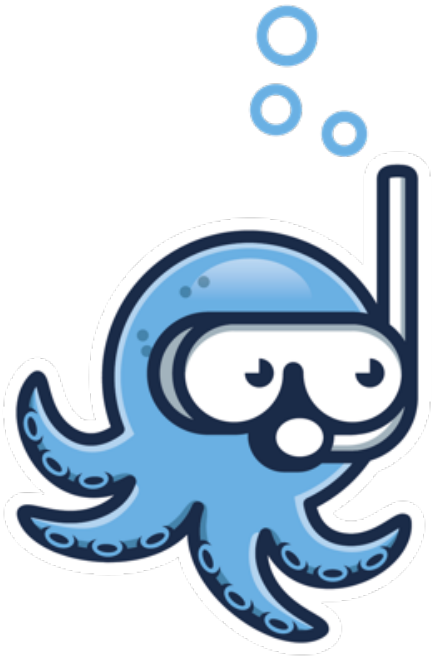
Radically easier to use ML systems

Make routine-ML, easy-ML

- Classification tasks
- Data cleaning & integration
- Entity & relationship extraction

Stretch goal: world-class quality in hours.

Snorkel @ Snorkel.Stanford.edu – over Spark!

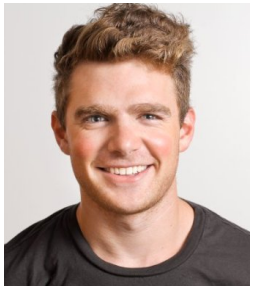


snorkel

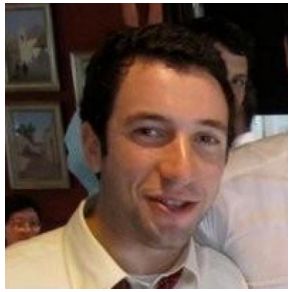
The Real Work

Stephen Bach

Chris De Sa



Henry Ehrenberg



Alex Ratner

Paroma Varma

The Rise of Automatic Feature Libraries

Pain: Users struggle to **write *good features.***

Deep learning (and others) removes feature engineering and is a **commodity** for many tasks

ML's Dirty Little Secret

Deep learning and much of ML needs **large training** sets.

"Bigger N allows more noise"

IMAGENET

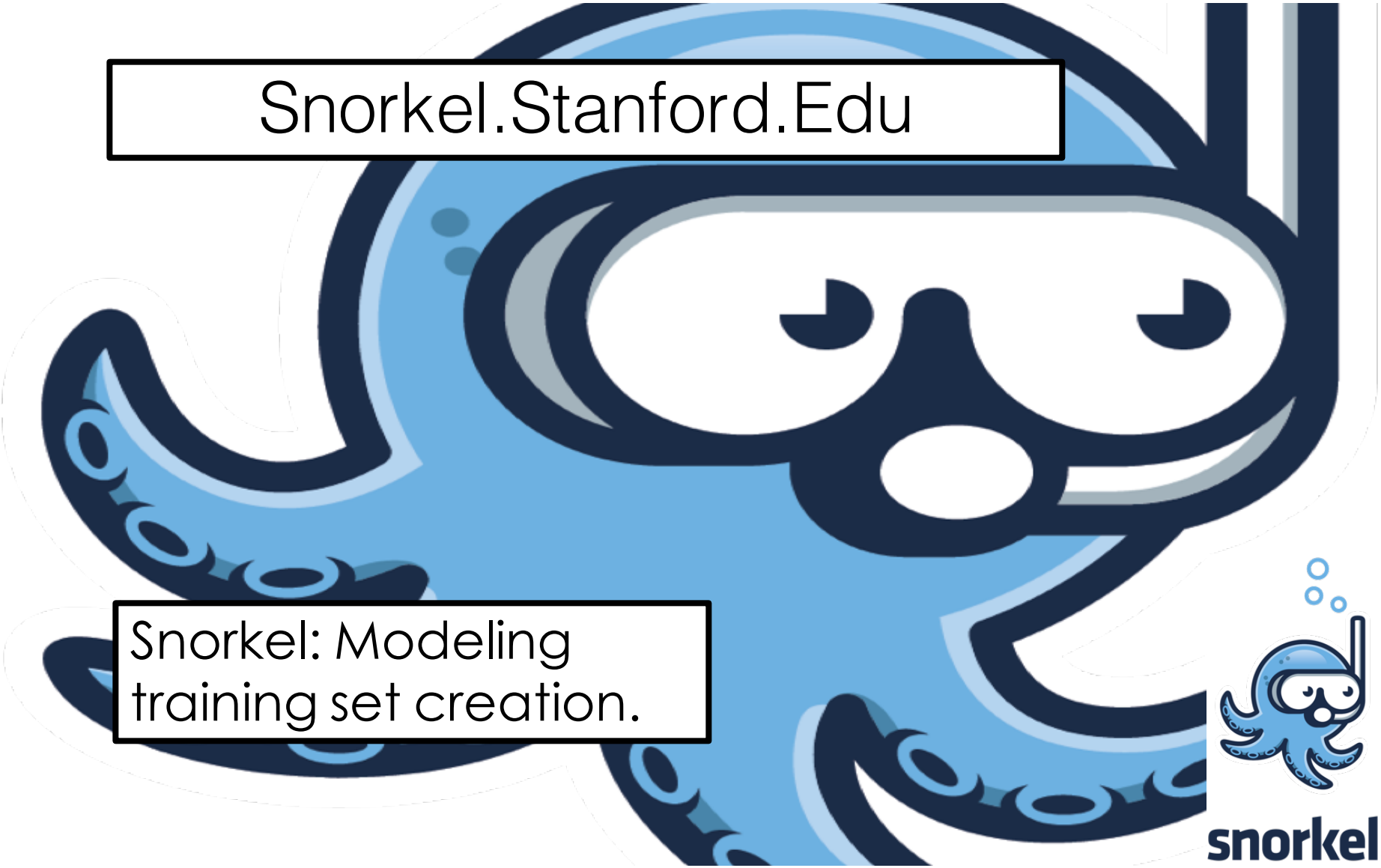
Creating hand-labeled training data the bottleneck.

The *New* New Oil

A Fundamental Problem in Machine Learning

Key idea: Model **process** or provenance of training set creation.

Snorkel.Stanford.Edu



Snorkel: Modeling training set creation.

snorkel

Case Study: **Lightweight** Extraction

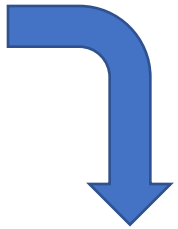
- Better than human extraction systems still take **months or years** to build using state-of-the-art ML systems

- Build systems that answer questions in ***hours to days***

What is holding us back?

Example: Chemical-Disease Relation Extraction from Text

TITLE:
Myasthenia gravis presenting as weakness after magnesium administration.
ABSTRACT:
We studied a patient with no prior history of neuromuscular disease who became virtually quadriplegic after parenteral magnesium administration for preeclampsia. The serum magnesium concentration was 3.0 mEq/L, which is usually well tolerated. The magnesium was stopped and she recovered over a few days. While she was weak, 2-Hz repetitive stimulation revealed a decrement without significant facilitation at rapid rates or after exercise, suggesting postsynaptic neuromuscular blockade. After her strength returned, repetitive stimulation was normal, but single fiber EMG revealed increased jitter and blocking. Her acetylcholine receptor antibody level was markedly elevated. Although paralysis after magnesium administration has been described in patients with known myasthenia gravis, it has not previously been reported to be the initial or only manifestation of the disease. Patients who are unusually sensitive to the neuromuscular effects of magnesium should be suspected of having an underlying disorder of neuromuscular transmission.

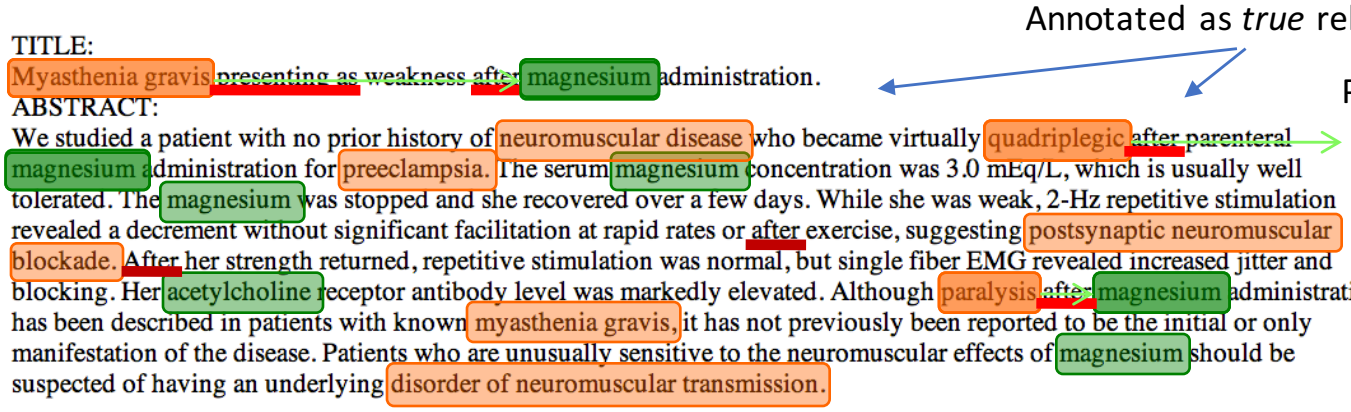


ID	Chemical	Disease
00	magnesium	Myasthenia gravis
01	magnesium	quadriplegic
02	magnesium	paralysis

Input: A corpus of text.

Goal: Populate a a table with pairs of chemicals reported to cause a disease.

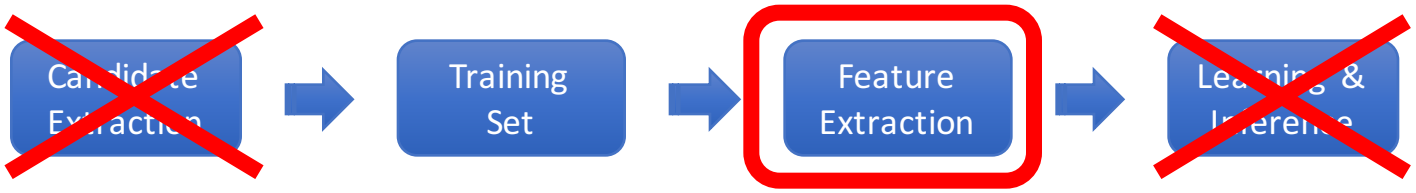
Relation Extraction with Machine Learning



Example binary features:

- `PHRASE_BTWN["presenting as"]`
- `WORD_BTWN["after"]`

TODAY:

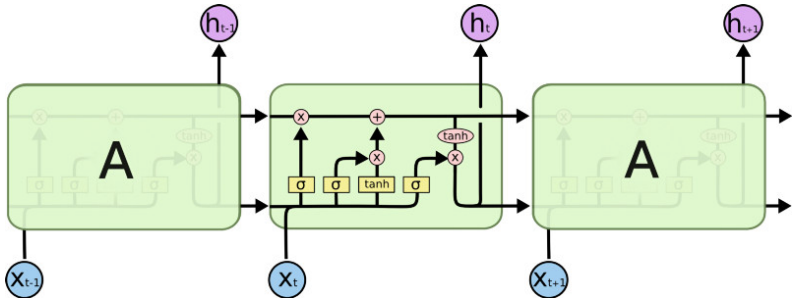


Used to be: *Feature engineering* is the bottleneck

Rise of Deep Learning

3. The BiLSTM Hegemony

**To a first approximation,
the de facto consensus in NLP in 2017 is
that no matter what the task,
you throw a BiLSTM at it, with
attention if you need information flow**



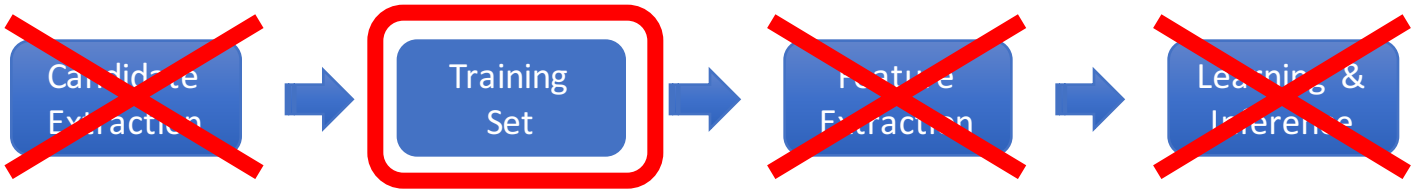
Feature engineering is dying!

Relation Extraction with Machine Learning

Annotated as *true* relations

TITLE:
Myasthenia gravis presenting as weakness after magnesium administration.
ABSTRACT:
We studied a patient with no prior history of neuromuscular disease who became virtually quadriplegic after parenteral magnesium administration for preeclampsia. The serum magnesium concentration was 3.0 mEq/L, which is usually well tolerated. The magnesium was stopped and she recovered over a few days. While she was weak, 2-Hz repetitive stimulation revealed a decrement without significant facilitation at rapid rates or after exercise, suggesting postsynaptic neuromuscular blockade. After her strength returned, repetitive stimulation was normal, but single fiber EMG revealed increased jitter and blocking. Her acetylcholine receptor antibody level was markedly elevated. Although paralysis after magnesium administration has been described in patients with known myasthenia gravis, it has not previously been reported to be the initial or only manifestation of the disease. Patients who are unusually sensitive to the neuromuscular effects of magnesium should be suspected of having an underlying disorder of neuromuscular transmission.

For a basic real-world use case:



…If we have ***massive*** training sets.

CRAZY IDEA: Noise-aware learning

By modeling noise in training set creation **process**,

we can use **low-quality** sources to train **high-quality** models.

Data Programming in Snorkel

- The user
  - ***Loads in*** unlabeled data
  - ***Writes*** labeling functions (LFs)
  - ***Chooses*** a discriminative model, e.g., LSTMs

- Snorkel
  - ***Creates*** a noisy training set- *by applying the LFs to the data*
  - ***Learns*** a model of this noise- *i.e. learns the LFs' accuracies*
  - ***Trains*** a *noise-aware* discriminative model

Importantly, **no** hand-labeled training sets.

Data Programming in Snorkel

- The user
  - ***Loads in*** unlabeled data
  - ***Writes*** labeling functions (LFs)
  - ***Chooses*** a discriminative model, e.g., LSTMs

Main user input!

- Snorkel
  - ***Creates*** a noisy training set- *by applying the LFs to the data*
  - ***Learns*** a model of this noise- *i.e. learns the LFs' accuracies*
  - ***Trains*** a *noise-aware* discriminative model

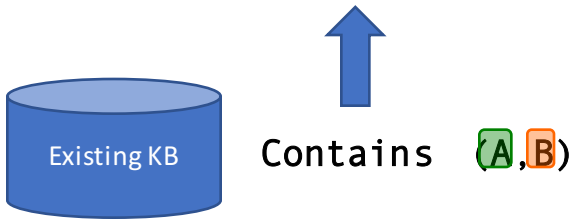
Labeling Functions

snorkel

- Traditional "distant supervision" rule relying on external KB

```
def lf1(x):
  cid =(x.chemical_id,x.disease_id)
  return 1 if cid in KB else 0
```

"Chemical A is found to cause disease B under certain conditions…"



→ **Label = TRUE**

This is likely to be true… *but*

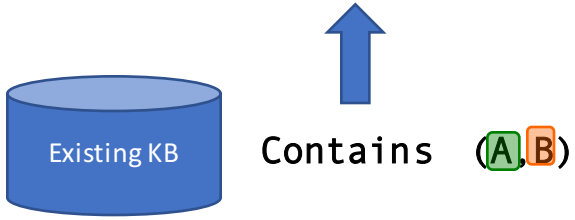
Labeling Functions

snorkel

- Traditional "distant supervision" rule relying on external KB

```
def lf1(x):
  cid =(x.chemical_id,x.disease_id)
  return 1 if cid in KB else 0
```

"Chemical A was found on the floor near a person with disease B..."



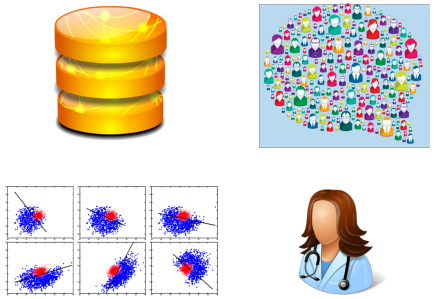
→ Label = TRUE

...can be false!

We learn **accuracy** and **correlations** for a handful of rules.
Experts: Generative model—without hand-labeled training data.

A Unifying Method for Weak Supervision

- Distant supervision
- Crowdsourcing
- Weak classifiers
- Domain heuristics / rules



$$\lambda : X \mapsto Y \cup \{\emptyset\}$$

You don't have to choose just one source! *Use them all!*

Hackathons with bio-* experts

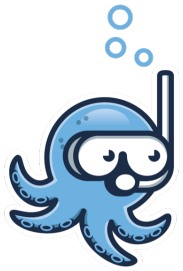
These systems can match or beat benchmark results ***without any labeled training data***!

- Ex: Three chemical / disease tagging tasks

System	NCBI Disease (F1)	CDR Disease (F1)	CDR Chem. (F1)
TaggerOne (Dogan, 2012)*	**81.5**	79.6	**88.4**
Snorkel: Logistic Regression	79.1	79.6	**88.4**
Snorkel: LSTM + Embeddings	79.2	**80.4**	88.2

A handful of labeling functions is competitive with the state-of-the-art supervised approach

Snorkel new, but in use!



Stanford MEDICINE

Shah Lab

U.S. Department of Veterans Affairs

Helix Group

FDA

ESOC Empirical Studies of Conflict

DARPA MEMEX

PharmGKB The Pharmacogenomics Knowledgebase

CISAC STANFORD

ExxonMobil Energy lives here

Greene Lab

accenture High performance. Delivered.

mallicklab

OIL

mobilize

District Attorney New York County

SING Stanford Information Networks Group

TOSHIBA

Canary Center at Stanford

It's raw. We welcome contributors – especially for Spark!

We can **fundamentally change** programming for ML by enabling **less precise** programming.

snorkel

STANFORD INFOLAB

DAWN

Peter Bailis
Kunle Olukotun
Matei Zaharia

http://dawn.cs.stanford.edu/

Conclusion

Machine learning can make programming **radically easier**.

Tutorials for extraction, data cleaning, and crowd sourcing—all on Spark!

Snorkel.Stanford.edu