# Few-shot learning of weak supervision sources in Snorkel

(or, learning weakly supervised weak supervisors)

Jesse Mu

#### **Project Outline**

- Replicate Snorkel causal relation extraction system
- Learn weak supervision sources from tiny sets of annotated examples, and compare performance to (1)





### The *New* New Oil

## snorkel

A training data creation and management system focused on information extraction

### Stanford **DAWN**

"We don't have **better algorithms** than anyone else; we just have **more data**"

Peter Norvig Chief Scientist, Google

#### How to get more labeled training data?

Traditional Supervision: Have subject matter experts (SMEs) hand-label more training data

Too expensive!

Active Learning: – Estimate which points are most valuable to solicit labels for Semi-supervised Learning: Use structural assumptions to automatically leverage unlabeled data Weak Supervision: Get lower-quality labels more efficiently and/or at a higher abstraction level Transfer Learning: Use models already trained on a different task

Get cheaper, lower-quality labels from non-experts Get higher-level supervision over unlabeled data from SMEs

Use one or more (noisy / biased) pre-trained models to provide supervision





#### Data Programming: Creating Large Training Sets, Quickly

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré Stanford University {ajratner,cdesa,senwu,dselsam,chrismre}@stanford.edu

Large labeled training sets are the critical building blocks of supervised learning methods and are key enablers of deep learning techniques. For some applications, creating labeled training sets is the most time-consuming and expensive part of applying machine learning. We therefore propose a paradigm for the programmatic creation of training sets called *data programming* in which users express weak supervision strategies or domain heuristics as *labeling functions*, which are programs that label subsets of the data, but that are noisy and may conflict. We show that by explicitly representing this training set labeling process as a generative model, we can "denoise" the generated training set, and establish theoretically that we can recover the parameters of these generative models in a handful of settings. We then show how to modify a discriminative loss function to make it noise-aware, and demonstrate

### NIPS 2016



BMC Systems Biology	Automatica prod.	<b>BMC Neuroscience</b>	Automatica State S	BMC Systems Biology	-	BMC Neuroscience	-37-m	<b>BMC Neuroscience</b>	America
fair postation for the post flatter manufactures and yets factor first. The latter and states Onlys	100000	Non-productor Hop location effects conter-surround modulation model of V1 Manufactor functions and Manufactoreses	CONTRACTOR OF A	A place evolution A place evolution and for improved departure of Data (1), Natural <sup>114</sup> and improved VO		Non-production Inhibitory Southark in a small CA3-conterest Free Solicement' and Wyse ( Vallman	1000000	Or provide A probabilitie framework to the com- study of charge detection and adaptation field framework and balance frame	activity from function a
Man fortugeness tell fort off mens tell fortugeness contractions in any second path		<ul> <li>Bernstein Marchael and State Annual Annua Annual Annual Ann</li></ul>		eller Verservieller Verster Verser Versit von die eller versitet im Verservieller eller Verservieller versitet im Verservieller Mathematiken versiteten Transpolatie		Address of the second constants. All seconds are a final second s		where is a prince has been used using the line.	
where a sub-		Company of the Compan		in state of the sign in the second state of the second state		PLANT DOLARS AND ADDRESS AND ADDRESS			
Statement and an and an and an and		No. of Concession, Name				Andre rapidli Millionen William and auf approximation		In Advanced Income Concession of Stationary Concession Names of Stationary On Stationary	
		Annual States of Conception of Conception		Bright Bright Bright Barrier Barrier				Annual Statements and statements	
Figure and the second second second		hand a second second second		The rest of the re		The advance optimizer of the standard standard state of the state of t		Name official interference of the States	
		Participant of the second state of the second		P Browner Berner Ber		And in the second of long we have the		1 Wheel a factor and the same	
	Sec.141		hap-al		Sec.11		Page Lat		10000
							and the second second		
		BMC Pharmacology	<u></u>	BMC Neuroscience		BMC Pharmacology	<u></u>	BMC Genomics	<u></u>
IMC Represeitence Interpreteiter Interfectuation Interfectuation (Interfectuation) Interfectuation (Interfectuation)		Finally detect PIGES a dealth adged second in inclusion formation Realthris Mandal (One Science Contact Name Realthris Mandal (Dears Science) October	A De Skillert	Even presentes Episamical insights on the bistory dependence do presentation of disaling stimut insight which if strange from the strange baseback to the	-12-	BMC Pharmacology Drawning Ricogae data therein is patients with choice or patients y backmans, (*10%) or patients y typestical (#40)		Technical area Bight PCB provides sampling and also the provides and also Refuel A White TP That C. Makey' Ji Claim	inter maliferation for high inte l'un and l'explore I Qualer
IMC Neuroscience India in committee raises and analog company and Comm <sup>21</sup> and included in Charlo <sup>124</sup>		rissingulared PIREs a standar science and an inclusion format Research March 2010 Strategy (Marchael Sciences	All and a second	Announcements Bywanical begins on the basetoy dependence do proceedings of reality altered balls (2019), Channe Broy, Straphile Basetowi ( The Calments)	and and a second	EMC Pharmacology Departure Research and a lapation with dense to Research approximate (CERR) or phoneory hypothesis (CERR) based to Charlow, and a lapation of induction from research and induction of the only phone from research and the dense of the dense of the only phone from research and the dense of the dense of the dense of the dense from research and the dense of the dense of the dense of the dense from research and the dense of the dense from research and the dense of t		Digital PCR provides sanative and shad	Later calibration for high into fue-and frequencial Quality
INC Neuroscience to provint to the constant sease of andia conputed to the constant sease of andia conputed to the constant sease of the constant to the constant sease of the constant sease the constant sease of the constant sease of the constant sease the constant sease of the constant sease of the constant sease the constant sease of the constant sease of the constant sease the constant sease of the constant sease of the constant sease the constant sease of the constant sease of the constant sease the constant sease of the constant		"Here data" PIAS a dealle ciget avant in kolonish here heredes a dealle ciget instant (Christen Mare Vriget Brach), fager brach, fager Mare, Mal Brach Brach, Hale Mare, Brach, fager Mare, Mal Brach Brach, Hale Mare, Brach, and Arthur Mare, Brach, Hale Mare, Status, and Status, and the Brach Brach, and Arthur Mar, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Mare, and Arthur Mare, and	A Statistics	From promotion Byranetical insights are the Natury-dependence di promotection of challing science finite Carlor", Control Insis, Stransfor functions, I Tone Carlor",	and and a second	BMC Pharmacology Desense: Respective des strates is private with device of private and particular is private with device of represented (FAR), una st hereof (or fundam) can't thereas . Include (organization)		To booking units Biglind FCR profiles searching and shad throughted a segmenting. Robust A Mitter IP That C Manage' JI Carlo and Strange and A Mitter IP That C Manage' JI Carlo Mitter Parameters and A Mitter A Mitter A Mitter A Mitter Mitter Parameters and A Mitter A Mitter A Mitter A Mitter Mitter A Mitter A Mitter A Mitter A Mitter A Mitter A Mitter Mitter A Mitter A Mitter A Mitter A Mitter A Mitter A Mitter Mitter A Mitter A Mitter A Mitter A Mitter A Mitter A Mitter Mitter A Mitter A Mitter A Mitter A Mitter A Mitter A Mitter A Mitter Mitter A Mitter A Mitt	Late raliferation for high into Fact and Peoples II Could of Statistic Peoples in a County People of Statistics
MC Hearseclence asymptotic house the values of aning comparison the Case of animatics of the Case of a comparison of the Case of a state of the Case of a comparison of the Case of th		"Here data" PIAS a dealle ciget avant in kolonish here heredes a dealle ciget instant (Christen Mare Vriget Brach), fager brach, fager Mare, Mal Brach Brach, Hale Mare, Brach, fager Mare, Mal Brach Brach, Hale Mare, Brach, and Arthur Mare, Brach, Hale Mare, Status, and Status, and the Brach Brach, and Arthur Mar, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Arthur Mare, and Mare, and Arthur Mare, and	A Statistics	Anna presentation Byparamitad lengths on the bistory dependence du presentation of Analysis physical and the Baseline's Inter California (California Barriel California) Inter California Inter California Inter California Inter California Inter California Inter California	and and a second	EMC Pharmacology Constraints C		Persentage units Explaint PCD provides sometime and allow through a second age And a System IP - Syndre Manay - H - Ope And a System IP - Syndre Manay - H - Ope And A - Ope - System Age And A - Ope - System Age A	tate rationate for high the rationate for high the factor of higher I Code of the state of the state of the state of the state of the state of the state of the official state of the state of the official state of the state of
MC Hearseclence asymptotic house the values of aning comparison the Case of animatics of the Case of a comparison of the Case of a state of the Case of a comparison of the Case of th		Transport	A Statistics	Anter provinces Beneficial integration of the Network dependence in Product Carry To, Carrier Barry, Tangata Theorem 2019 The Carrier Carry To, Standard Theorem 2019 Theorem 2019	and and a second	BMC Pharmacology Creation and an end of the second		Parameters and the parameters an	tate rationate for high the rationate for high the factor of higher I Code of the state of the state of the state of the state of the state of the state of the official state of the state of the official state of the state of
		Taking object PGREs & decide and provide its inclusion is brown for status of additional transmission of the status of the baseline inclusion of status of the status of the status baseline inclusion of status of the status of	A Statistics	Anny spectrates: By exercise it suggests are the basis of dependence of the suggest of the suggest of the suggest of the suggest basis of exercises: The suggest of the suggest of the suggest of the suggest the suggest of the suggest of the suggest of the suggest of the suggest of the	and and a second	BMC Pharmacology Conversion Conve		Persentage units Explaint PCD provides sometime and allow through a second age And a System IP - Syndre Manay - H - Ope And a System IP - Syndre Manay - H - Ope And A - Ope - System Age And A - Ope - System Age A	tate rationate for high the rationate for high the factor of higher I Code of the state of the state of the state of the state of the state of the state of the official state of the state of the official state of the state of
INC. Neuroscience Internetica		Hang start Hang start Hannes Andels - Une and A being the start Hannes Andels - Une and A being the Hannes Andels - Une and A being the Hannes - International Andels - International Hannes - International Andels - International Andels - International Hannes - International Andels - International Andels - International Hannes - International Andels - Internati	A Statistics	Anter provinces Beneficial integration of the Network dependence in Product Carry To, Carrier Barry, Tangata Theorem 2019 The Carrier Carry To, Standard Theorem 2019 Theorem 2019	and and a second	BMC Pharmacology Draws Pharmacology Pharmaco	and a second sec	Researching and the second sec	tare rationales de las
INC Neuroscience Incomentations of adaption Mark Characteristics of adaption and Characteristics of adaption and Characteristics of adaption description de		Personal State	A Statistics			<section-header><section-header><section-header><section-header><section-header><text><text><text></text></text></text></section-header></section-header></section-header></section-header></section-header>	La constanti constant	The second secon	
MC Neuroscience Instruction Instruction and and and and and and Instruction and and and and and and Instruction and and and and and and and Instruction and And And And And And And And Instruction and A		Transmitter     Transmiter     Transmitter     Transmiter     Transmitter     Transmitter	A Statistics		and and a second	BMC Pharmacology Draws Pharmacology Pharmaco	La constanti constant	Report of the second se	An additional and tables     An addition



def lf1(x): cid = (x.chemical\_id, x.disease\_id) return 1 if cid in KB else 0

Contains  $(\mathbf{A}, \mathbf{B})$ 

def lf2(x): m = re.search(r'.\*cause.\*', x.between) return 1 if m else 0

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

Existing

KB







Example Weak Supervision Sources Technical Challenge: Integrating & Modeling Diverse Sources Use Weak Supervision to Train End Model

#### Extension: from *examples* to *labeling functions*

- Labeling functions (LFs) require programming experience and abstraction.
- Can we learn noisy labelers from few examples, without a single line of code?
- Given sentences and relations, generate many candidate LFs that distinguish LF from surrounding sentences

Several diseases that appear to be heritable, but not genetically defined, have been observed at low frequency in the breed.<u>11</u>, <u>12</u>, <u>13</u> Many of these disorders have evolved with the domestic dog over time and inherited by descent as breeds have been created [<u>3</u>]. Except for hip dysplasia, which is considered one of the more serious disorders of Samoyed, most heritable and potentially heritable disease traits of the breed have been of minor importance.<u>11</u> There are only three simple deleterious genetic disorders in Samoyed with defined causes, X-linked glomerulopathy [<u>4</u>], X-linked progressive retinal atrophy [<u>5</u>], and an incomplete dominant short-limbed defect with ocular abnormalities [<u>6</u>, <u>7</u>].



#### Extension: from *examples* to *labeling functions*

- Labeling functions (LFs) require programming experience and abstraction.
- Can we learn noisy labelers from few examples, without a single line of code?
- Given sentences and relations, generate many candidate LFs that distinguish LF from surrounding sentences

2 questions:

- How dumb are LFs generated in this way?
- How dumb can LFs be before Snorkel begins to break down?

# Few-shot learning of weak supervision sources in Snorkel

(or, learning weakly supervised weak supervisors)

Jesse Mu

Application	# of LFs	Coverage	$ S_{\lambda \neq 0} $	Overlan	Conflict	F1 Score Improvement	
Application				Overlap		HT	LSTM
KBP (News)	40	29.39	2.03M	1.38	0.15	1.92	3.12
Genomics	146	53.61	256K	26.71	2.05	1.59	0.47
Pharmacogenomics	7	7.70	129K	0.35	0.32	3.60	4.94
Diseases	12	53.32	418K	31.81	0.98	N/A	N/A

Table 2: Labeling function (LF) summary statistics, sizes of generated training sets  $S_{\lambda\neq0}$  (only counting non-zero labels), and relative F1 score improvement over baseline IRT methods for hand-tuned (HT) and LSTM-generated (LSTM) feature sets.