Introducing Relational^{AI}

Probabilistic Graphical Models –and-Statistical Relational Learning

Molham Aref -- Relational^{Al}

OVERVIEW

Age: 9 months old AI/ML startup

Team:

- 16 people, 10 PhD's, 4 former university professors
- Faculty Network: 9 @ ~20% + summers, additional ~12+ in extended network

Location: distributed team (SV, Atlanta, Seattle, Toronto, NYC, Utrecht, London) Offices in Berkeley and Utrecht

Financials: Funded through 2019 + revenue from consulting

Industries served: TBD (looking at Financial Services, Business Intelligence)

 If all else fails we're going to pivot to Deep Quantum Crypto Blockchain of Things and hope for the best

HOW WE SEE OURSELVES

Expertise in systems (8 people)

- We built sophisticated compilers & interpreters (OOPSLA, ECOOP, ICFP)
- We built databases that advanced the state of the art (SIGMOD, VLDB)

Expertise in theory (7 people)

- We prove things (PODS, ICDT, TODS, JACM, POPL)
- We designed first-of-a-kind asymptotically efficient algorithms
 - Worst case optimal join algorithms
 - Asymptotically superior query plans

Expertise in ML & AI (7 people)

- We implemented scalable ML & probabilistic systems (NIPS, ICML, AAAI)
- We developed sophisticated statistical and mathematical models

INDUSTRIAL EXPERIENCE

HNC

- Financial services: Fraud detection -- credit card, insurance (23 of top 25 credit card issuing banks)
- Neural networks on proprietary HW accelerators (also computational intelligence, <u>Database</u> mining)
- IPO 1995, acquired by FICO in 2002

Retek

- Retail: Demand forecasting, supply chain optimization, pricing (majority of Retail Global 250)
- Time series with approximate optimization first hierarchical forecasting solution to scale to Retail volumes
- IPO 1999, acquired by Oracle in 2005

Optimi

- Telecom: Wireless network optimization (AT&T, Cingular, Next, America Movil, Telefonica)
- Monte Carlo simulation, heuristic search (simulated annealing)
- Acquired by Ericsson in 2010

Brickstream

- Retail: in-store video analytics (>30% market share globally)
- Old-school computer vision (pre-deep learning) first industrial use of stereo cameras
- Acquired by Point-Grey which was acquired by FLIR in 2015

LogicBlox/Predictix

- Retail: Demand forecasting, supply chain optimization, pricing (3 of top 6 US retailers + dozen large global retailers)
- Factorization machines, linear programming, integer programming first on cloud
- Acquired by Infor in 2016







7







OUR SECRET SAUCE

- We know how to exploit problem structure to make optimization asymptotically faster
 - We know how to perform stochastic gradient descent and batch gradient descent directly on normalized relational data
 - Most ML Methods can be solved well with SGD, some with BGD
- Asymptotically faster optimization means anything that depends on it can go faster
 - Learning/Parameter optimization
 - Hyperparameter optimization
 - Feature engineering
 - Structure learning
 - Inference
 -

STRUCTURE FOR TRADITIONAL/DISCRIMINATIVE ML

5 COMPONENTS OF MACHINE LEARNING

- Method (or model class)
 - e.g. FM, decision tree, neural network, ...
- Loss (error) function
 - e.g. Absolute error (L1 norm), Square error (L2 norm), ...
- Generalization mechanism
 - e.g. Regularization (norm * penalty), cross validation
- Evaluation function
 - Takes the model parameters and the input and produces a prediction
- Optimizer
 - E.g. Gradient descent, EM algorithm

5 TYPES OF METHODS or MODEL CLASSES

- Regression: predict a number
 - Linear regression with VIF
 - LASSO regression
 - Multi-time series prediction
 - Nonparametric regression
 - Mixture of experts
 - Factorization machines
 - Polynomial regression
- Classification: predict a category
 - Naïve Bayes classifier
 - Non-parametric Bayes classifier
 - K-nearest neighbor classifier
 - Support vector machine
 - Decision tree
 - Hidden Markov model

- Density Estimation: find likelihood of objects
 - Histograms
 - Kernel density estimation
- Clustering: find natural groups
 - K-means
 - Spectral clustering
 - Mean shift clustering
- Dimension Reduction: combine features
 - Singular value decomposition
 - Maximum variance unfolding
 - Non-negative matrix factorization
 - Kernel principal component analysis
 - (Ensemble) Singular value decomposition
 - GROUSE
 - Random projections
 - Tensor factorization PARFAC/CANDECOMP

DESIGN MATRIX



Entities

DESIGN MATRIX IS A VIEW ON STRUCTURED (RELATIONAL) DATA





Entities

ID	x1	x2	x3	•••	у

Features

WE CAN EXPLOIT THE RELATIONAL STRUCTURE

We use algebraic structure (e.g. semi-rings) to "push the aggregations through the joins" to implement lifted stochastic and batch gradient descent for efficient learning of a variety of model classes

- Linear regression
- Polynomial regression
- Factorization machines
- Decision trees
- Neural nets
- ••••

(many more on the way)

		v ₁	V2	V3	V4
Join Representation	Listing	774M	3.614G	3.614G	3.614G
(#values)	Factorized	37M	169M	169M	169M
Compression	Fact/List	20.9×	21.4×	21.4×	21.4×
Join Computation (PSQL) for 1	R, TensorFlow, libFM	50.63	216.56	216.56	216.56
Factorized Computation of 43	Counts over Join	8.02	34.15	34.15	34.15
	I	inear regression	ı		
Features	without FDs	33 + 55	33+55	33+1340	33+2702
(continuous+categorical)	with FDs	same a	s above, there are	no FDs	33+2653
Aggregates	without FDs	595+2,418	595+2,421	595+111,549	595+157,735
(scalar+group-by)	with FDs	same a	s above, there are	no FDs	595+144,589
MADLib (ols)	Learn	1,898.35	8,855.11	> 79,200.00	-
R (QR)	Export/Import	308.83	-	-	-
	Learn	490.13	-	-	-
TensorFlow (FTLR)	Export/Import	74.72	372.70	372.70	372.70
(1 epoch, batch size 1K)	Learn	2762.50	12710.53	12724.94	12708.11
F	Aggregate	93.31	424.81	OOM	OOM
	Converge (runs)	0.01 (359)	0.01 (359)		
AC/DC	Aggregate	25.51	116.64	117.94	895.22
	Converge (runs)	0.02 (343)	0.02 (367)	0.42 (337)	0.66 (365)
AC/DC+FD	Aggregate		same as AC		380.31
	Converge (runs)		there are no FDs		8.82 (366)
Speedup of AC/DC+FD over	MADlib	74.36×	75.91×	> 669.14×	00
	R	33.28×	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	00
	TensorFlow	113.12×	$114.01 \times$	$112.49 \times$	$34.17 \times$
	F	3.65×	3.64×	00	00
	AC/DC	same as	AC/DC, there are	no FDs	2.30×

BENCHMA	RK (2/3)
---------	----------

		v ₁	V2	V3	V4
Join Representation	Listing	774M	3.614G	3.614G	3.614G
(#values)	Factorized	37M	169M	169M	169M
Compression	Fact/List	20.9×	21.4×	21.4×	21.4×
Join Computation (PSQL) for I	R, TensorFlow, libFM	50.63	216.56	216.56	216.56
Factorized Computation of 43	Counts over Join	8.02	34.15	34.15	34.15
	Polynor	nial regression d	legree 2		
Features	without FDs	562+2,363	562+2,366	562+110,209	562+154,033
(continuous+categorical)	with FDs	same a	562+140,936		
Aggregates	without FDs	158k+742k	158k+746k	158k+65,875k	158k+46,113k
(scalar+group-by)	with FDs	same a	158k+36,712k		
MADlib (ols)	Learn	> 79, 200.00	> 79,200.00	> 79,200.00	-
AC/DC	Aggregate	132.43	517.40	820.57	7,012.84
	Converge (runs)	3.27 (321)	3.62 (365)	349.15 (400)	115.65 (200)
AC/DC+FD	Aggregate		same as AC/DC		1,819.80
	Converge (runs)		219.51 (180)		
Speedup of AC/DC+FD over	MADlib	> 583.64×	> 152.01×	> 67.71×	00
	AC/DC	same as	AC/DC, there are	no FDs	$3.50 \times$

BENCHMARK	(3/3)
-----------	-------

		v ₁	v ₂	V3	V4		
Join Representation	Listing	774M	3.614G	3.614G	3.614G		
(#values)	Factorized	37M	169M	169M	169M		
Compression	Fact/List	20.9×	21.4×	21.4×	21.4×		
Join Computation (PSQL) for I	R, TensorFlow, libFM	50.63	216.56	216.56	216.56		
Factorized Computation of 43	Counts over Join	8.02	34.15	34.15	34.15		
	Factorizatio	on machine degr	ee 2 rank 8				
Features	without FDs	530+2,363	530+2,366	530+110,209	530+154,033		
(continuous+categorical)	with FDs	same a	562+140,936				
Aggregates	without FDs	140k+740k	140k+744k	140k+65,832k	140k+45,995k		
(scalar+group-by)	with FDs	same a	same as above, there are no FDs				
libFM (MCMC)	Export/Import	412.84	1462.54	3,096.90	3,368.06		
	Learn (runs)	19,692.90 (300)	103,225.50 (300)	79,839.13 (300)	87,873.75 (300)		
AC/DC	Aggregate	128.97	498.79	772.42	6,869.47		
	Converge (runs)	3.03 (300)	3.05 (300)	262.54 (300)	166.60 (300)		
AC/DC+FD	Aggregate	same as AC/DC		1,672.83			
	Converge (runs)		there are no FDs		144.07 (300)		
Speedup of AC/DC+FD over	libFM	152.70×	209.03×	80.34×	50.33×		
	AC/DC	same as	AC/DC, there are	e no FDs	3.87×		

STATISTICAL RELATIONAL LEARNING

SRL and StarAl

MOTIVATION

- Graphical models are considered by some to be "one of the most exciting advances in machine learning (AI, signal processing, coding, control, ...) in the last decades"
- Graphical models allow us to gain global insight based on local observations
- There are different types of graphical models
 - Directed: eg. Bayesian Networks (aka belief networks)
 - Undirected: e.g. Mark Networks (aka Markov Random Fields), Factor Graphs
 - Mixed: e.g. Chain Graphs both directed acyclic graphs and undirected graphs are special cases of chain graphs, which can therefore provide a way of unifying and generalizing Bayesian and Markov networks
- Statistical Relational models generalize PGM's in the same way that first order logic generalizes propositional logic – they allow us to quantify over individuals/entities

STATISTICAL RELATIONAL LEARNING (SRL)

- Knowledge is represented as a distribution over possible worlds
 - Finite and infinite sets of possible worlds are supported
 - Undirected models: via integrity constraints
 - We specify the constraints that determine the legal set of possible worlds & a function to score each of them.
 - We don't have to provide a program to generate each possible world.
 - Directed models: via probabilistic programs
 - We have to provide a program to generate each possible world. Score by normalizing frequency of a given world relative to all others.
 - Normalize the score of each world by the sum of scores of all the worlds
- Inference
 - Unlike "traditional" methods where prediction is the input applied to the parameters of the model class, inference in SRL requires expensive optimization or (approximate) integration over possible worlds
- Learning
 - Unlike traditional learning algorithms, just one instance to learn from (the relational DB)
 - Structure learning uses inference during each step

SEMANTICS

Ordinary minimal model semantics:



Slide thanks to Benny Kimelfeld

UNDIRECTED MODELS

Use Integrity Constraints to specify a set of possible worlds & define a scoring function for each

SMOKERS AND FRIENDS

Smoking and Quitting in Groups

Researchers studying a network of 12,067 people found that smokers and nonsmokers tended to cluster in groups of close friends and family members. As more people quit over the decades, remaining groups of smokers were increasingly pushed to the periphery of the social network.



CERTAIN KNOWLEDGE WITH INTEGRITY CONSTRAINTS

A logical Knowledge Base is a set of Integrity Constraints that define a set of possible worlds:

```
person(x)
smokes(x) -> person(x)
cancer(x) -> person(x)
friends(x, y) -> person(x), person(y)
```

Assuming persons Alice (A) and Bob (B), then there are 4 possible relations for each of: smokes, cancer:





А	В	А
 		В

There are 16 possible relations for friends



So how many possible worlds are there?

• 2 bits for for each of smokes cancer and 4 bits for friends: And how many if we add a 3rd person Carla (C)?

В

• 3 bits * 2 unary relations + 9 bits for binary relation:

2⁸ or 256 possible worlds



EACH POSSIBLE TUPLE IS A NODE IN A GRAPHICAL MODEL



PROBABILISTIC KNOWLEDGE WITH WEIGHTED INTEGRITY CONSTRAINTS

Smoking causes cancer Friends have similar smoking habits

Represent probabilistic knowledge with soft (weighted) Integrity Constraints

```
person(x)
smokes(x) -> person(x)
cancer(x) -> person(x)
friends(x, y) -> person(x), person(y)
w1 smokes(x) -> cancer(x)
w2 smokes(x), friends(x, y) -> smokes(y)
```

When a world violates a formula, it becomes less probable, not impossible Weights give indication of certainty in domain knowledge or expertise

- 0 -> no confidence -- might as well not have the IC
- Infinity -> absolute certainty hard constraint
- Infinity -> absolute certainty in the converse hard constraint on negated IC

GRAPHICAL MODEL



ALL WORLDS



31

ALL WORLDS where B smokes



32





QUANTIFYING OVER POSSIBLE WORLDS

Observations/measurements eliminate some of the possible worlds

Finding the mostly likely world can be computed with optimization

Computing the probability of any world requires us to aggregate/integrate over all possible worlds.

HOW DO WE MAKE SRL EFFICIENT?

There are 2 important dimensions to consider

- Brawn (i.e. the constant factors)
 - Latency hiding: memory hierarchy and network latencies (e.g. in memory)
 - Specialization: specialize for workload (e.g. JIT compilation), specialize for data
 - Parallelization: SIMD, multi-core, accelerators (e.g. GPU, TPU), in memory computing
- Brain (i.e. the asymptotics)
 - Lifting and Structure exploitation: algebraic (e.g. semi rings, groups), combinatorial, statistical, geometric
 - Approximation (with error bars): e.g. variational methods

both ---> approximate lifted inference

SUMMARY: ADVANTAGES OF (STATISTICAL) RELATIONAL AI

- Performance
 - Exploits the relational structure for asymptoticly better performance
- Understandability
 - Declarative relational language can be used to codify knowledge/expertise (human to machine) and to return insight (machine to human)
- Quality
 - Fewer assumptions regarding independence, identical distributions, # of observations per example, etc. can produce more accurate models
- Versatility
 - Generalized inference: from observations to unknowns in any direction

WORST-CASE OPTIMAL MULTI-WAY JOIN

- Worst-Case Optimal Join Algorithms: Techniques, Results, and Open Problems. Ngo. (Gems of PODS 2018)
- Worst-Case Optimal Join Algorithms: Techniques, Results, and Open Problems. Ngo, Porat, Re, Rudra. (Journal of the ACM 2018)
- What do Shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another? Abo Khamis, Ngo, Suciu, (PODS 2017 - Invited to Journal of ACM)
- Computing Join Queries with Functional Dependencies. Abo Khamis, Ngo, Suciu. (PODS 2017)
- Joins via Geometric Resolutions: Worst-case and Beyond. Abo Khamis, Ngo, Re, Rudra. (PODS 2015, **Invited to TODS 2015**)
- Beyond Worst-Case Analysis for Joins with Minesweeper. Abo Khamis, Ngo, Re, Rudra. (PODS 2014)
- Leapfrog Triejoin: A Simple Worst-Case Optimal Join Algorithm.
 Veldhuizen (ICDT 2014 Best Newcomer)
- Skew Strikes Back: New Developments in the Theory of Join Algorithms. Ngo, Re, Rudra. (Invited to SIGMOD Record 2013)
- Worst Case Optimal Join Algorithms. Ngo, Porat, Re, Rudra. (PODS 2012 – Best Paper)

	r	Leap	frog Triej	oin: A Simple, Worst-Case Optimal Join Algorithm
ABS: Efficier		Wo	orst-case	Optimal Algorithms for Conjunctive Queries with Functional Dependencies
studied gorithm a novel	ABSTR		What and	do Shannon-type Inequalities, Submodular Width, Disiunctive Datalog have to do with one another?
worst-c Atseria	Recent y			
conjun	forth AG			Worst-case Optimal Join Algorithms
the bod they re	on the in	ABCT	м	worst-case Optimal John Algorithmis
theoret	and Rud	ABSI		HUNC O NGO U-in-in-it p-fi-1- SIDIV
whethe	ing join a tional to	been de		FLV POPAT Bas-Ilan University
to data	log syste	mal tim		CHRISTOPHER RÉ. Stanford University
join ph	leapfrog	gorithm	ABSTRA	ATRI RUDRA, University at Buffalo, SUNY
case or	spurred	dencies	Recent work	
indeper	per we es	and sho	query with fi	Efficient join processing is one of the most fundamental and well-studied tasks in database research. In this
proof e and Ma	over, lear	subsets	information	work, we examine algorithms for natural join queries over many relations and describe a new algorithm to
that thi	finer-gra	lattice of	gous output l	process these queries optimally in terms of worst-case data complexity. Our result builds on recent work
and Th	straints o Wo show	deep, no	of these bou	by Atserias, Grohe, and Marx, who gave bounds on the size of a natural join query in terms of the sizes of the individual relations in the body of the cuery. These bounds however, are not constructions they rely
inequal	than NP	and wo timal al	related to SI	of the individual relations in the body of the query. These bounds, nowever, are not constructive: they rely on Shearer's entropy inequality, which is information-theoretic. Thus, the previous results leave open the
used to	leapfrog	present	Shannon-typ	question of whether there exist algorithms whose runtimes achieve these optimal bounds. An answer to this
Cata	to \exists_1 que	dependi	inequalities".	question may be interesting to database practice, as we show in this article that any project-join style plans,
Care	addition	Keywe	PANDA, whi	such as ones typically employed in a relational database management system, are asymptotically slower than the optimal for some queries. We present an algorithm whose runtime is worst-case optimal for all natural
n.2.9 (ing easy concise o	Worst-c	size bound p	join queries. Our result may be of independent interest, as our algorithm also yields a constructive proof of
Gene	concise o	cies, ent	as a black-be fractional hy	the general fractional cover bound by Atserias, Grohe, and Marx without using Shearer's inequality. This
Algorit	Genera	1 1	times for age	bound implies two famous inequalities in geometry: the Loomis-Whitney inequality and its generalization,
	Algorith	Sever	dependencies Our result	we discuss how our algorithm can be used to evaluate full conjunctive gueries optimally, to compute a relaxed
Keyv		lead to	First, our bo	notion of joins and to optimally (in the worst-case) enumerate all induced copies of a fixed subgraph inside
Join Al	1. INI	algorith	eral class of oueries are a	of a given large graph.
1, 100	Join pi studied r	timal al	matches pree	$\texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \textbf{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \texttt{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \texttt{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \texttt{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \texttt{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \texttt{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet \textbf{Information systems} \rightarrow \texttt{Relational database model}; \bullet \textbf{Theory of computation} \rightarrow \texttt{CCS Concepts:} \bullet CCS Concep$
	can be fo	plans, i wise ioi	previous algo mial in this b	Database query processing and optimization (theory):
		The rur	for queries w	Additional Key Words and Phrases: Join Algorithms, fractional cover bound, Loomis-Whitney inequality,
Permiss		the data	dencies, and Overall, or	Bollobás-Thomason inequality
personal not mad	Permission	triangle	three seemin	
bear this republie	personal or not made or	case tin	on proot seq of independe	
permiss	bear this no	While		A preliminary version of this article was presented at PODS'12 as Reference [62]. We thank Georg Gettlob for sending
Copyrig	permission :		Keywords	us a full version of his work [30]. We thank XuanLong Nguyen for introducing us to the Loomis-Whitney inequality.
	Copyright 2		Submodular	We thank Dung Nguyen for catching some errors in the earlier statement of our algorithm. We thank the anonymous PODS'12 and IACM reference for many helpful comments that have areally improved the preparentation of the activity IAN's
		Permissio	equalities; Er	work is partly supported by NSF Grants No. CCF-1161196 and No. CCF-1319402. C.R. acknowledges the National Science
		personal o	cess patterns	Foundation (NSF) under CAREER Awards No. IIS-1353606 and No. CCF-1356918, the Office of Naval Research (ONR) under Awards No. NOROLLELIONLI and No. NOROLLELION for Research Fellowship the Marco Post Action Data
		bear this i	Permission to mal classroom use ic o	revards No. N000141210041 and No. N000141310129, the Stoan Research Fellowship, the Moore Foundation Data Driven Investigator award, and gifts from American Family Insurance, Google, Lightspeed Ventures. and Toshiba. A.R.'s work on
		republish, permissio	for profit or comm on the first page	this project is supported by NSF Grants No. CCF-0844796 and No. CCF-1319402.
		Copyright	author(s) must be l	Authors' addresses: H. Q. Ngo and A. Rudra, 338 Davis Hall, University at Buffalo, Buffalo, NY, 14214. USA; emails: Bummun, strikibuffalo adu E. Borst Ban-Ban University Pamat-Can 5980002 Izrad, email: noratebolier bia acil: C. Ré
			and/or a fee. Requ	Gates Computer Science Building, 353 Serra Mall, Stanford, CA 94305. USA; email: chrismre@cs.stanford.edu.
	l		PODS'17, May © 2017 Copyright	Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee
			ISBN 978-1-4503-	provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored.
			DOI: http://dx	Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires
				prior specific permission and/or a fee. Request permissions from permissions@acm.org. 0.2018.ACM.0004-5411/2018/03-ART16.\$15.00
		L		https://doi.org/10.1145/3180143
				Journal of the ACM Val 65 No. 3. Article 16. Publication date: March 2018

OPTIMAL QUERY PLANS FOR MULTI-WAY JOINS

- What do Shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another? Abo Khamis, Ngo, Suciu, (PODS 2017 - Invited to Journal of ACM)
- FAQ: Questions Asked Frequently, Abu Khamis, Ngo, Rudra, (PODS 2016 – Best Paper, Invited to Journal of ACM).
- Juggling functions inside a database, Abo Khamis, Ngo, Suciu (Invited to SIGMOD Record)

		FAQ: Questions Asked	Frequ	ently		
		Juggling Functions	Inside	a Datab	base	
ABSTR	ma	What do Shannon-typ and Disjunctive Data	pe Ineq alog ha	ualities, ve to do	Submodular Wig	dth, er?
We define problem, v tions in e tions, prob main cone	ABSTRA We define as	Mahmoud Abo Khamis LogicBlox Inc.	Hung (LogicBl	Q. Ngo lox Inc.	Dan Suciu LogicBlox Inc. and University of Washington	
We then solve this traditional programm adds a cou	problem, wł a very wide logic, matri ical models, Simply put,	ABSTRACT		1. INTRO	DDUCTION	
order to de tage of Gr and of the join algori	to evaluate : into a set of query is the	Recent works on bounding the output size of a query with functional dependencies and degree b shown a deep connection between fundamental information theory and database theory. We p gous output bounds for disjunctive database rules	conjunctive ounds have questions in rove analo- and answer	from four difference of the four four difference of the four differe	erent research threads, and establis even those threads. Bound for Full Conjunctive	h new con-
As is the model infe to solve an <i>variable o</i> work is a p is "semant the input	join algorith mally evalu exciting dev work tighth with a vast ner, showin bluwing the	gous output bounds for any several open questions regarding the tightness ar of these bounds along the way. The bounds ar related to Shannon-type information inequaliti vise the notion of a "proof sequence" of a spec Shannon-type information inequalities called "SI inequalities". We then show how a proof seque	Grohe and and Gottlob, connection be tive query wi and informat this bound to	Marx [30], Atserias, Grohe, and Lee, Valiant and Valiant [27] develo etween the output size bound of th (or without) functional depende ion theory. Our first problem is degree constraints, and to study w	Marx [13], oped a deep a conjunc- encies (FD) to extend whether the	
tion algori has the be host of kn ference, m	good datab solver, relat gine, and m The Insid	used as symbolic instructions to guide an algor PANDA, which answers disjunctive datalog rules size bound predicted. We show that PANDA c as a black-box to devise algorithms matching p fractional hypertree width and the submodular	ithm called within the an be used recisely the width run-	We associat $\mathcal{H} = ([n], \mathcal{E}), a$ Its atoms are	te a full conjunctive query Q to a l $\mathcal{E} \subseteq 2^{[n]}$. The query's variables are $R_F, F \in \mathcal{E}$. The query is:	hypergraph $A_i, i \in [n].$
We also worst-case	scribed in t general prob upon the be	times for aggregate and conjunctive queries with dependencies and degree bounds.	functional	A	$Q(\mathbf{A}_{[n]}) := \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{A}_F),$	(1)
over comp	specializes t in graphical lational join	First, our bounds and algorithms are for the muc- eral class of disjunctive datalog rules, of which	h more gen- conjunctive	goal is to com the input dat	pute an upper bound on the output abase satisfies a set of degree const	≤ [n]. Our t size, when traints. De-
1. INT 1.1 Me	within any of pled way of the LogicBlo This, work	queries are a special case. Second, the runtime matches precisely the submodular width bound previous algorithm by Marx has a runtime that mial in this bound. Third, our bounds and algor	of PANDA l, while the t is polyno- ithms work	fine $\deg_F(\mathbf{A}_Y)$ gree constrain $N_{Y X}$, for X tion of the fo	$ \mathbf{A}_X\rangle \stackrel{\text{def}}{=} \max_{\mathbf{t}} \Pi_{\mathbf{A}_Y}(\sigma_{\mathbf{A}_X=\mathbf{t}}(R_F)) ;$ at is an assertion of the form deg _F (. $\subset Y \subseteq F$, A cardinality constraint $\operatorname{rm} R_F \leq N_F$, for some $F \in \mathcal{E}$; is	then, a de- $\mathbf{A}_Y \mathbf{A}_X) \leq$ is an asser- t is exactly
domains sl Example a series of	1319402 and 0009. The U distribute re ing any cop	for queries with input cardinality bounds, functi dencies, and degree bounds. Overall, our results showed a deep connecti- three seemingly unrelated lines of research; and, on proof sequences for Shanpon flow inequality	onal depen- on between our results se might be	the degree co- tional depend $N_{X \cup Y X} = 1$ both cardinal	nstraint $\deg_F(\mathbf{A}_F \emptyset) \leq N_{F \emptyset} \stackrel{\text{def}}{=} N$ lency $\mathbf{A}_X \rightarrow \mathbf{A}_Y$ is a degree const . Thus, degree constraints strictly ity constraints and FDs.	F. A func- traint with generalize
	(c) ACM 201 Questions Ask 4191-2/16/06, DOI: http:// to make digit	of independent interest.	5 mgn oc	The first ou who establish only, known t and degree o	tput size upper bound was pioneered ed a tight bound, for cardinality coday as the AGM bound. Extensionstraints were discussed in [27] a	d in [13,30], constraints ons to FDs and [3], re-
Permission to personal or o not made or bear this noti republish, to reemission a	sonal or class are not mad and that cop page. Copyr than ACM n ted. To cop	Keywords Submodular width; Disjunctive datalog; Shann equalities; Entropy; Functional dependencies; Re cess patterns; Degree bounds; Join algorithms	on-type in- stricted ac-	spectively, wh are tight. Ha strong practic scribed a new independent q	to left open the question whether the andling queries with degree constr- cal motivation. Armbrust et al. [w approach to query evaluation, c puery processing, which guarantees a	ese bounds aints has a 10–12], de- alled <i>scale</i> - a fixed run-
Copyright 20	redistribute fee. Request	Permission to make digital or hard copies of all or part of this word classroom use is granted without fee provided that copies are not ma for profit or commercial advantage and that copies bear this notice and on the first page. Copyrights for components of this work owned by author(s) must be honored. Abstracting with credit is permitted. To cc republish, to post on servers or to redistribute to lists, requires prior sp	t for personal or de or distributed d the full citation others than the opy otherwise, or scific permission	time even wh bound; this g write explicit derive upper upper bounds	en the size of the database increas guarantee is provided by asking de degree constraints, then using h bounds on the query output. Thus on the size of the query arguments	ses without velopers to euristics to s, improved we immedia
l	6	anaror a rec. Request permissions from permissions #km.org. PODS'17, May 14 - 19, 2017, Chircago, IL, USA © 3017 Copyright held by the owner/author(s). Publication rights li ISBN 978-1-4303-4198-11705513.00 DOI: http://dx.chi.org/10.1145/3034786.3056105	censed to ACM.	ate applicatio eral complexi ("is the outpu in [15–17].	ns to scale-independent query proce ty results on the associated decisic t size of the query bounded?") were	essing. Sev- on problem considered

IN-DB LEARNING

- In-Database Learning with Sparse Tensors, Abo Khamis, Ngo, Nguyen, Olteanu, Schleich (PODS 2018)
- AC/DC: In-Database Learning Thunderstruck, Abo Khamis, Ngo, Nguyen, Olteanu, Schleich (DEEM 2018)
- Modelling Machine Learning Algorithms on Relational Data with Datalog. Makrynioti, Vasiloglou, Pasalic, Vassalos. (DEEM 2018)
- In-Database Factorized Learning, Ngo, Nguyen, Olteanu, Schleich (AMW 2017)
- Data Science with Linear Programming. Makrynioti, Vasiloglou, Pasalic, Vassalos. (DeLBP 2017)

Hung Q.	In-Dat Ngo ¹ , XuanL	tabase Factorized Learn	ing ximilian Schleid	ch ³		
1 Intr		In-Database Learning Mahmoud Abo Khamis ¹ Dan Olteanu ³ M	with Spa Q. Ngo ¹ Maximilian Schle	rse Tensor: KuanLong Nguyen	S 2	
In this pa class of o and forec	А ВСТВА	AC/DC: In-D	atabase Le	earning Thu	nderstruck	
polynomi	ABSI KA In-database	Relational AI, Inc	Relation	alAI, Inc	University of Michigan	
classificat fined by a	avoids the c with on a d vert data fo	Dan Ol University of	teanu of Oxford	Maximilian S University of	Schleich Oxford	
to large of analytics Pushi time usu systems a challenge technique computat factorized We furth optimizat work attr 2 Pro We use a In the normal fa inner pro a databas Suppose t y and n in Q(D). Fo \mathbb{R}^m and multivari $(h_j)_{j \in [m]}$	port the pa cally fundar. This pape evaluating a topolynomial component from relation outproposed and problems and problems and problems and gine and use with signifi- solutions the solutions the Although the solutions the analytic second of data set tary. Mean discovery Mean and Second generation and solutions the solutions the solutions the solutions the solutions the solutions the solutions the solution the solutio	<text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text>	e AC/DC gradient s over normalized robben into a set ions. It then uses prove the solution dutabase size, the nd the large num- times the solution into the solution into the solution into the solution of aggregates. factorization ma- u-world dataset of factorization ma- u-world factorization factorization factorization factorization factorization factorization main factorization of the solution of the solution of the solution solution prior solution main dispersion to make dataset and the solution solution of the solution of the solution of the solution solution of the solution of the solution of the solution solution of the solution of the	currently under develop- recent effort to bring at the interface between ACDCC ¹ solves equi- relations. It is a unified by fauture extrac- relations. It is a unified plan with asymptotical them in isolation. This First, ACDC decor- solver that iteratively until i reaches conver- solver that iteratively until in reaches conver- solver that iteratively until in reaches conver- sation of the solver that interatively and the solver scalars for combination su- calars for combination su- ered combinations that a encoding. These are com- bined by the solver and the expresentations that a encoding. These are com- plexity. We categorical features by a Second. ACDC for Second. ACDC for the only other in-data form the underlying the in- data for the underlying the acid or in-between the solves of the only other in-data form the underlying the con- spondent over key-foreig ACDCC granularity. The (ACCC means)	pment at RelationalAI, Inc. Is nat malytics inside the database [7, 9, na database spress and databatis initiation or database of po- long querics over database of po- lagoreash for computing both It you are a spressively the two tas- space, they are intertwined in on glabases queries, the two tas- space and the spressively than that ol- goreas a given optimization proj- goreas a given optimization proj- ogenes a given optimization proj- ogenes and the solutions to it geness. The aggregates capture to functional the group-base ag- tith at least one categorical factures only. It database queries are called a strategories of the in-database methics are categorical factures only. It database queries on the result of the factures only avoid the expension wire one-hot en- oded the resultance, introduced and particle the compared of the factures only of the relations entropy introduced magnated on the result of the facture tabase learning system (13) that at tabase learning system (13) that at tabase learning system (13) that at tabase learning system (13) that tabase learning system (13) that at tabase learning system) (13) that at tabase learning system) (13) that at tabase learning system) (13) that tabase learning system) (13) that at tabase learning system) (13) that tabase learning system) (13) that at tabase learning system) (13) that tabase learning system) (13) that tabase learning system) (13) that ta- tabase learning system) (13) that the same the same the decompose and cannot recover the good core spression and cannot recover the good core s	scribe to a $(12, 13, 21)$ (12, 13, 21) (12, 13, 21) (13,
		must ue nonorea, sussancially With refail is permitted, 10 copy to post on servers or to redistribute to lists, requires prior speci- fee. Bequest permissions from permissions@ucm.org. DEEM, June 2018, Hoacton, Texas © 2018 Association for Computing Machinery.	sunerwise, or republish, fic permission and/or a	and their aggregates h ¹ AC/DC supports both type processing. Its name allures a and at the fast-paced sound of	ave similar structures. s of features, i.e., categorical and contin t another duality; that of alternating and di of a homonymous rock band.	uous, and fai screte current

IN-DB & LIFTED LP/QP/IP SOLVERS

- SolverBlox: Algebraic Modeling in Datalog. Borraz-Sanchez, Klabjan, Pasalic, Aref. (Declarative Logic Progamming – Morgan & Claypool 2018)
- The Symbolic Interior Point Method. Mladenov, Belle, Kersting. (AAAI 2017)
- Lifted Inference for Convex Quadratic Programs.
 Mladenov, Kleinhans, Kersting. (AAAI 2017)
- RELOOP/ A Python-Embedded Declarative Language for Relational Optimization. Mladenov, Heinrich, Kleinhans, Gonsior, Kersting. (AAAI 2016)
- Relational Linear Programs. Kersting, Mladenov, Tokmanov. (2015)
- Lifted Linear Programming. Mladenov, Ahmadi, Kersting. (AISTATS 2012)



FOR A SMALL GROUP OF REBELS TO BEAT THE EMPIRE, WE HAVE TO ...



FIN