Adopting Markov Logic Networks for Big Spatial Data and Applications

Ibrahim Sabek

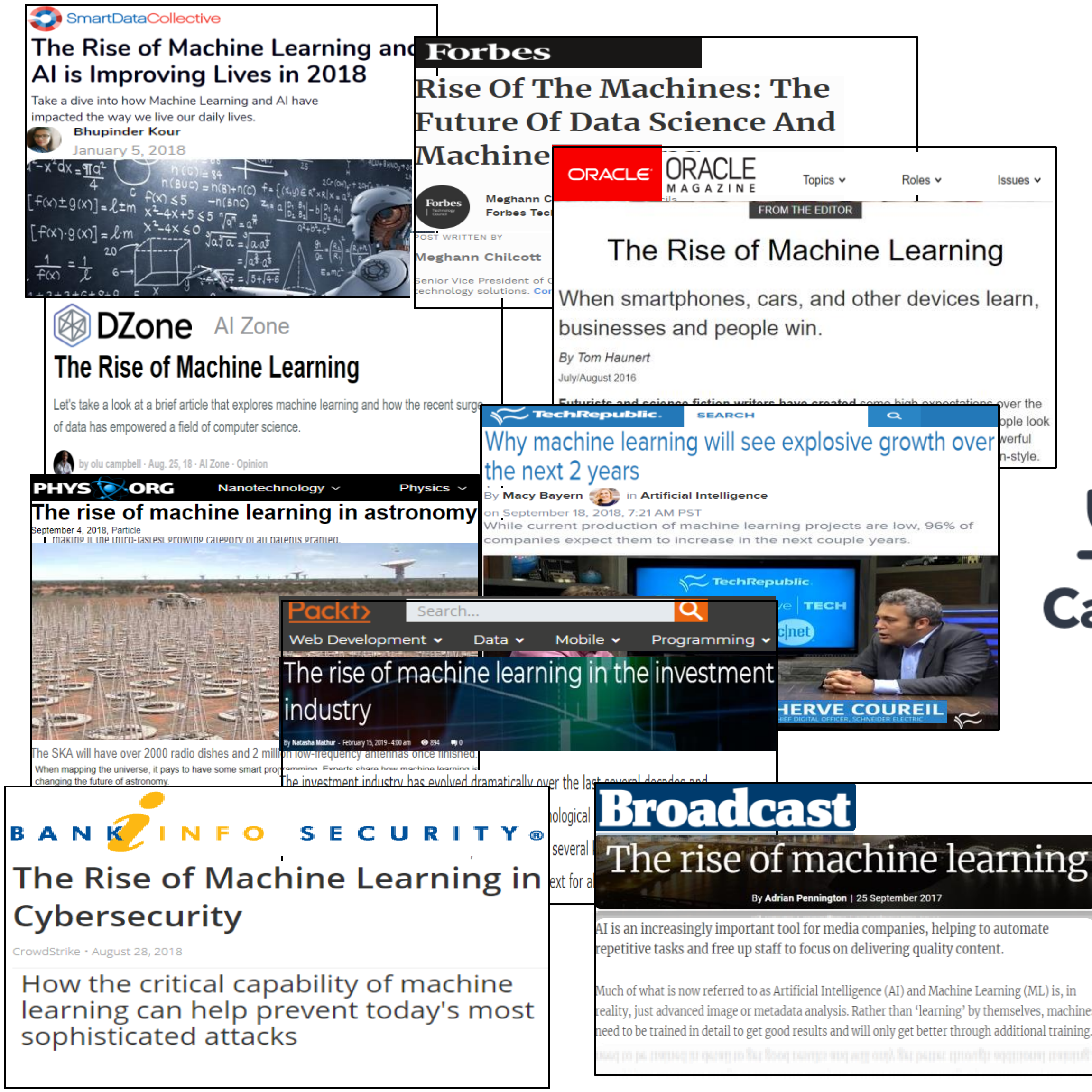
Thomas Lord Department of Computer Science

USC Viterbi School of Engineering

Spatial Sciences Institute (SSI)

USC Dornsife Dana and David Dornsife College of Letters, Arts and Sciences

The Rise of Machine Learning (ML)



"Machine learning is a core, transformative way by which we're rethinking everything we're doing." -Google CEO Sundar Pichai

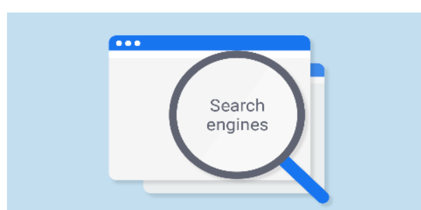
ML and Big Data

Knowledge Base

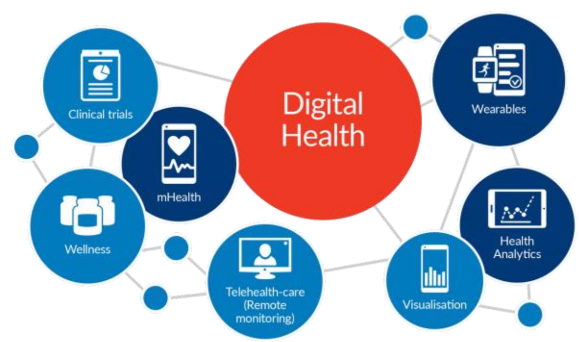
Event Detection

Image Analysis

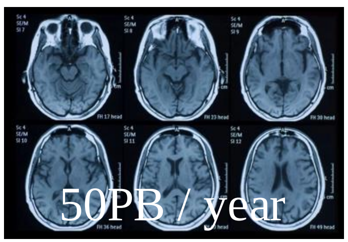
Crowdsourcing

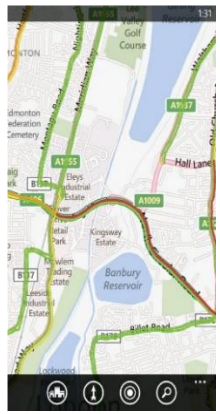


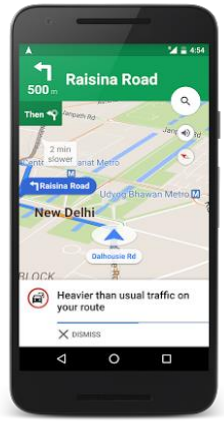
Search Engines

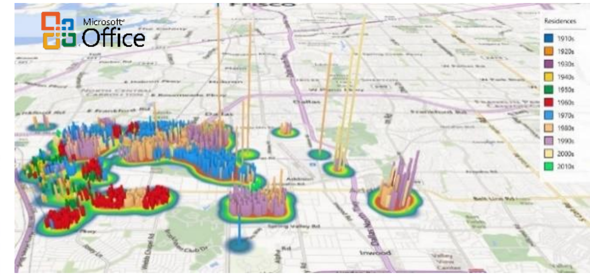


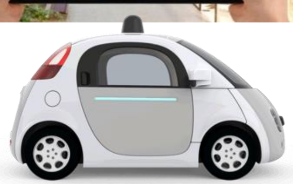
Meanwhile, … Big Spatial Data

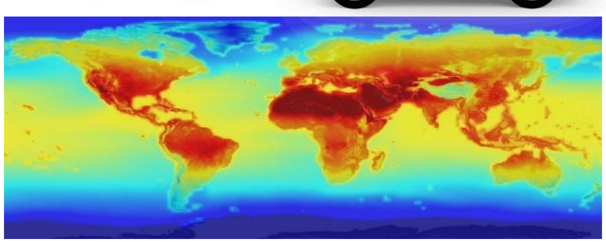


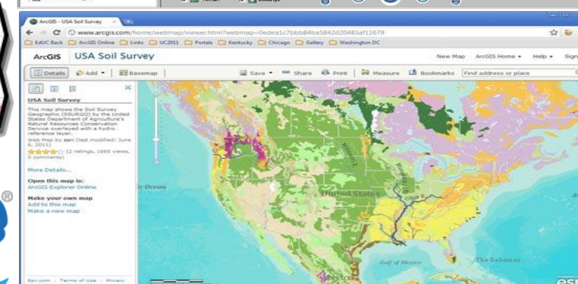


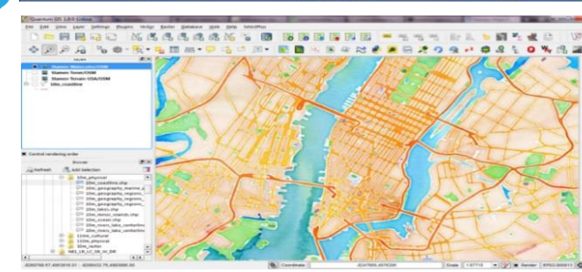












ML and Big Spatial Data

Event Detection

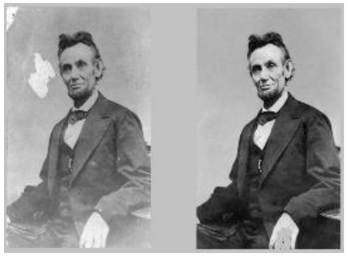
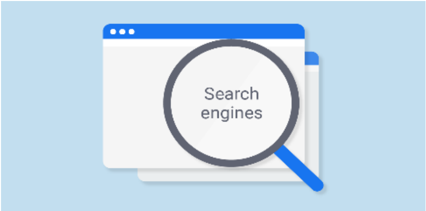
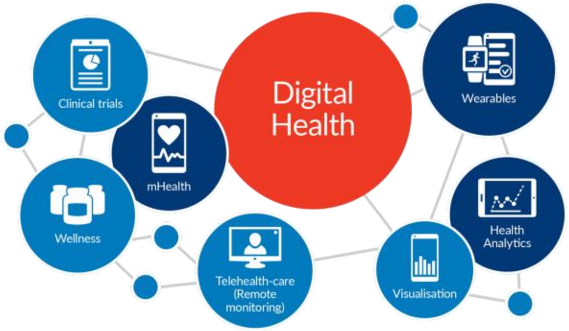


Image Analysis

Crowdsourcing



Search Engines



- **ML internals ignore the properties of spatial data and relationships**

"Thinking Spatial" … Can we adapt ML internals to properly use spatial data?

Knowledge Base Construction

Knowledge Base Rules

Probabilistic
Knowledge Base
Construction System

Relations Extraction

Factual Scores Inference

tables, relations of facts

Spouses KB

Person 1	Person 2	Confidence
Barack	Michelle	0.85
Joe	Katy	0.52
Joe	Lily	0.46

Knowledge Base Construction

Knowledge Base Rules

Probabilistic
Knowledge Base
Construction System

Relations Extraction

Factual Scores Inference

tables, relations of facts

Spouses KB

Person 1	Person 2	Confidence
Barack	Michelle	0.85
Joe	Katy	0.52
Joe	Lily	0.46

NELL

yago select knowledge

SystemT

StatSnowBall

DARPA MEMEX

Fight Human Trafficking
Crime Investigation

DeepDive

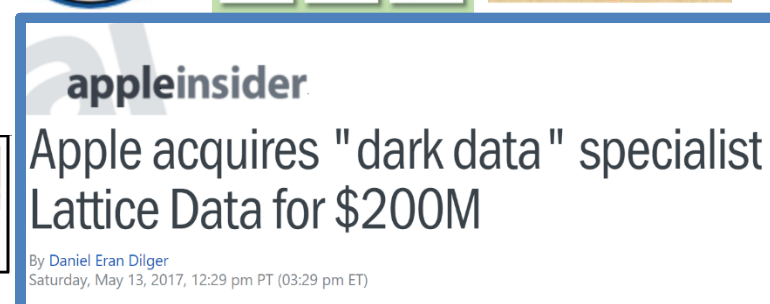
lixto DELIVERING COMPETITIVE ADVANTAGE

DBLife

Google Vault

ProbKB

SCIENTIFIC AMERICAN WSJ 60 MINUTES CNN Forbes BBC


appleinsider

Apple acquires "dark data" specialist Lattice Data for $200M

By Daniel Eran Dilger
Saturday, May 13, 2017, 12:29 pm PT (03:29 pm ET)

DeepDive: Introduction

- **Extracting structured data from unstructured data.**
  - Structured data: SQL tables, Knowledgebases, association rules …
  - Unstructured data: text, image, PDFs, tables, ….
- **Infrastructure for probabilistic machine learning and data mining algorithms.**
- **Think of features not algorithms.**
- **Declarative inference rules:**

```
person_smokes(p) =>
    person_has_cancer(p) :-
    person(p, _).
```

http://deepdive.stanford.edu/

DeepDive: Smoke Example

```
person (
  person_id bigint,
  name text
).
person_has_cancer? (
  person_id bigint
).
person_smokes? (
  person_id bigint
).
friends (
    person_id bigint,
    friend_id bigint
).
@weight(0.5)
person_smokes(p) =>
 person_has_cancer(p) :- person(p, _).

@weight(0.4)
person_smokes(p1) =>
 person_smokes(p2) :-
  person(p1, _), person(p2, _),
  friends(p1, p2).
```

- **`person_has_cancer` and `person_smokes` need to be inferred.**
- **Implication relation depends on Boolean logic (AND, OR)**
  - What if the implication relation has spatial semantics, e.g. meet, neighbor, north of?
- **Variables are linked to each other through ID matching (Hash join).**
  - What if the variables should be matched based on their overlap areas (Spatial Join)?

DeepDive with Spatial Data …

Ebola infection rates in Liberia

Inference Rules

```
P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level

Rule: P1&P2 ➔ Y has high infection rate
```

Data

County	I	S
Montserrado	1	0.6
Margibi	?	0.6
Bong	?	0.6
Gbarpolu	?	0.6

Infections

Sanitation

DeepDive

Result

County	Confidence				Ground Truth
Margibi	0.54				[0.6, 1]
Bong	0.52				[0.4, 0.6[
Gbarpolu	0.63				[0.2, 0.4[

DeepDive with Spatial Data …

Ebola infection rates in Liberia

Inference Rules

```
P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
P3: Counties X&Y are within 150 miles
```

~~**Rule:** P1&P2 ➔ Y has high infection rate~~

Rule: P1&P2&P3 ➔ Y has high infection rate

DeepDive

Data

County	I	S
Montserrado	1	0.6
Margibi	?	0.6
Bong	?	0.6
Gbarpolu	?	0.6

Infections

Sanitation

Result

County	Confidence				Ground Truth
Margibi	~~0.54~~	0.51			[0.6, 1]
Bong	~~0.52~~	0.45			[0.4, 0.6[
Gbarpolu	~~0.63~~	0.06			[0.2, 0.4[

DeepDive with Spatial Data …

Ebola infection rates in Liberia

Inference Rules

```
P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
P3: Counties X&Y are within 150 miles   (0.01)
P4: Counties X&Y are within 148.5 miles (0.02)
                     ..
P102: Counties X&Y are within 1.5 miles (1)
```

~~**Rule:** P1&P2 ➔ Y has high infection rate~~

Rule: P1&..&P102 ➔ Y has high infection rate

DeepDive

Data

County	I	S
Montserrado	1	0.6
Margibi	?	0.6
Bong	?	0.6
Gbarpolu	?	0.6

Infections

Sanitation

Result

County	Confidence				Ground Truth
Margibi	~~0.54~~	~~0.51~~	0.63		[0.6, 1]
Bong	~~0.52~~	~~0.45~~	0.48		[0.4, 0.6[
Gbarpolu	~~0.63~~	~~0.06~~	0.14		[0.2, 0.4[

Execution time of these rules in the **grounding** phase explodes !!

DeepDive with Spatial Data …

Ebola infection rates in Liberia

Inference Rules

P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
~~P3: Counties X&Y are within 150 miles (0.01)~~
~~P4: Counties X&Y are within 148.5 miles (0.02)~~
..
~~P102: Counties X&Y are within 1.5 miles (1)~~
P3: The closer Y&X the higher Y infection rate

~~**Rule:** P1&P2 ➔ Y has high infection rate~~
Rule: P1&P2&P3 ➔ Y has high infection rate

DeepDive

Data

County	I	S
Montserrado	1	0.6
Margibi	?	0.6
Bong	?	0.6
Gbarpolu	?	0.6

Infections

Sanitation

Result

County	Confidence				Ground Truth
Margibi	~~0.54~~	~~0.51~~	~~0.63~~		[0.6, 1]
Bong	~~0.52~~	~~0.45~~	~~0.48~~		[0.4, 0.6[
Gbarpolu	~~0.63~~	~~0.06~~	~~0.14~~		[0.2, 0.4[

Where Is the Problem?

- **DeepDive is built on top of Markov Logic Networks (MLN)**
  - MLN is designed for *binary logic* only
    - E.g., bitwise-AND, bitwise-OR, and imply

- **MLN is not spatially- aware**
  - It can not interpret the *gradual semantics* of spatial predicates
    - E.g., `P3: The closer Y&X the higher Y infect rate`

We propose *Spatial Markov Logic Networks (SMLN)*, a full-fledged MLN framework with a native support for spatial data and applications

Outline

- Motivation
- Introduction to Spatial Markov Logic Networks (SMLN)
  - MLN in a Nutshell
  - SMLN Architecture
- SMLN for Knowledge Base Construction
- SMLN for Spatial Analysis
- Summary

Markov Logic Networks (MLN)

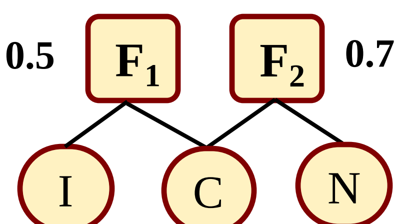
- **MLN is an end-to-end ML solution**
  - Covers wide range of ML problems
  - User-friendly and efficient
    - No need for ML expert to use it
    - Thousands of lines of ML code can be done in very few MLN formulas

Weighted rules as MLN formulas → **Markov Logic Networks (MLN)** → Learning and inference outputs

First-order logic rules

F_1: Illiteracy → Crime [0.5]

F_2: Crime ^ Non-safety [0.7]



Factor graph

Alchemy - Open Source AI

ACM SIGMOD HOUSTON, TEXAS, USA 2018

ACM SIGMOD/PODS International Conference on Management of Data

June 10 – June 15, 2018 Houston, TX, USA

SIGMOD 2018: Keynote Talks

Machine Learning for Data Management: Problems and Solutions

datanami

DATA SCIENCE • AI • ADVANCED ANALYTICS

July 3, 2018

Can Markov Logic Take Machine Learning to the Next Level?

Alex Woodie

Advances in machine learning, including deep learning, have propelled artificial intelligence (AI) into the public conscience and forced executives to create new business plans based on data. However, the

Tuffy

Scalable RDBMS-based MLN System

DeepDive

HoloClean

Markov Logic Networks (MLN)

- **Combination of two things**
  - First-order Logic rules
    - Handles reasoning
  - Mar
    -
- **Examp**

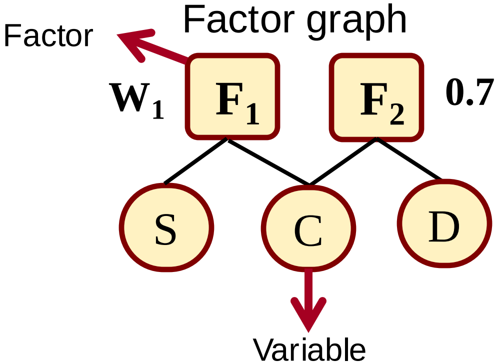
To solve a problem with MLN, find equivalent variables and rules representation. That is it!

1. Express rules with weights
2. Ground to a factor graph
3. ...ts from the factor ... **Gibbs sampling** and ... **scent optimization**
4. ...wn variables based on weights using **Gibbs sampling**

Rules

F$_1$: Smoke → Cancer [W$_1$]
F$_2$: Cancer ^ Die [0.7]

Factor graph



Smoke

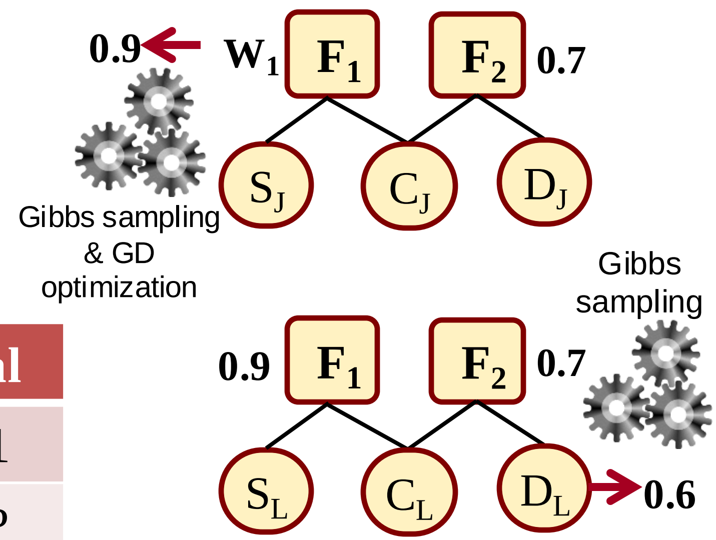
Var	Val
Joe	0.9
Lily	0.7

Cancer

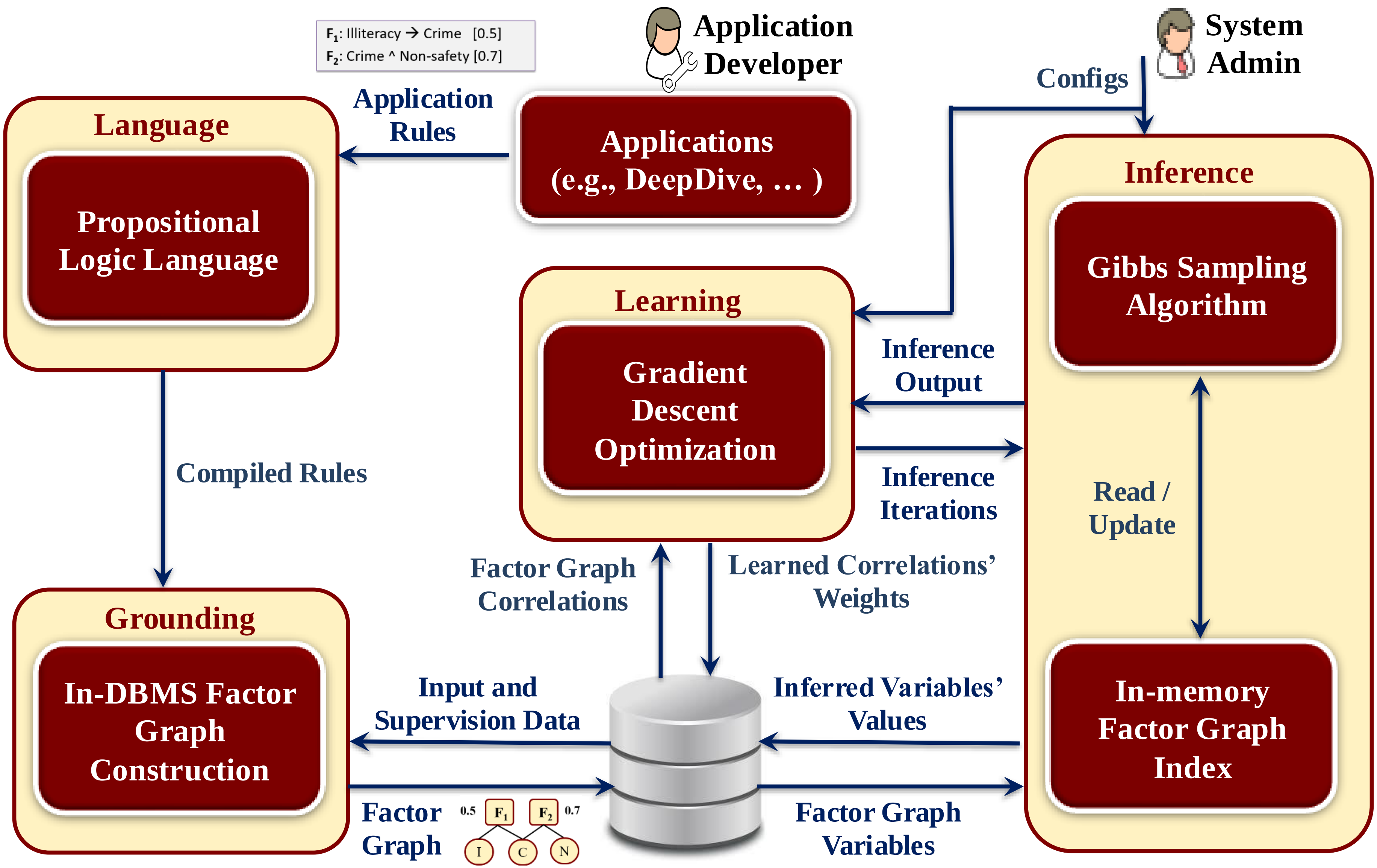
Var	Val
Joe	0.8
Lily	0.5

Die

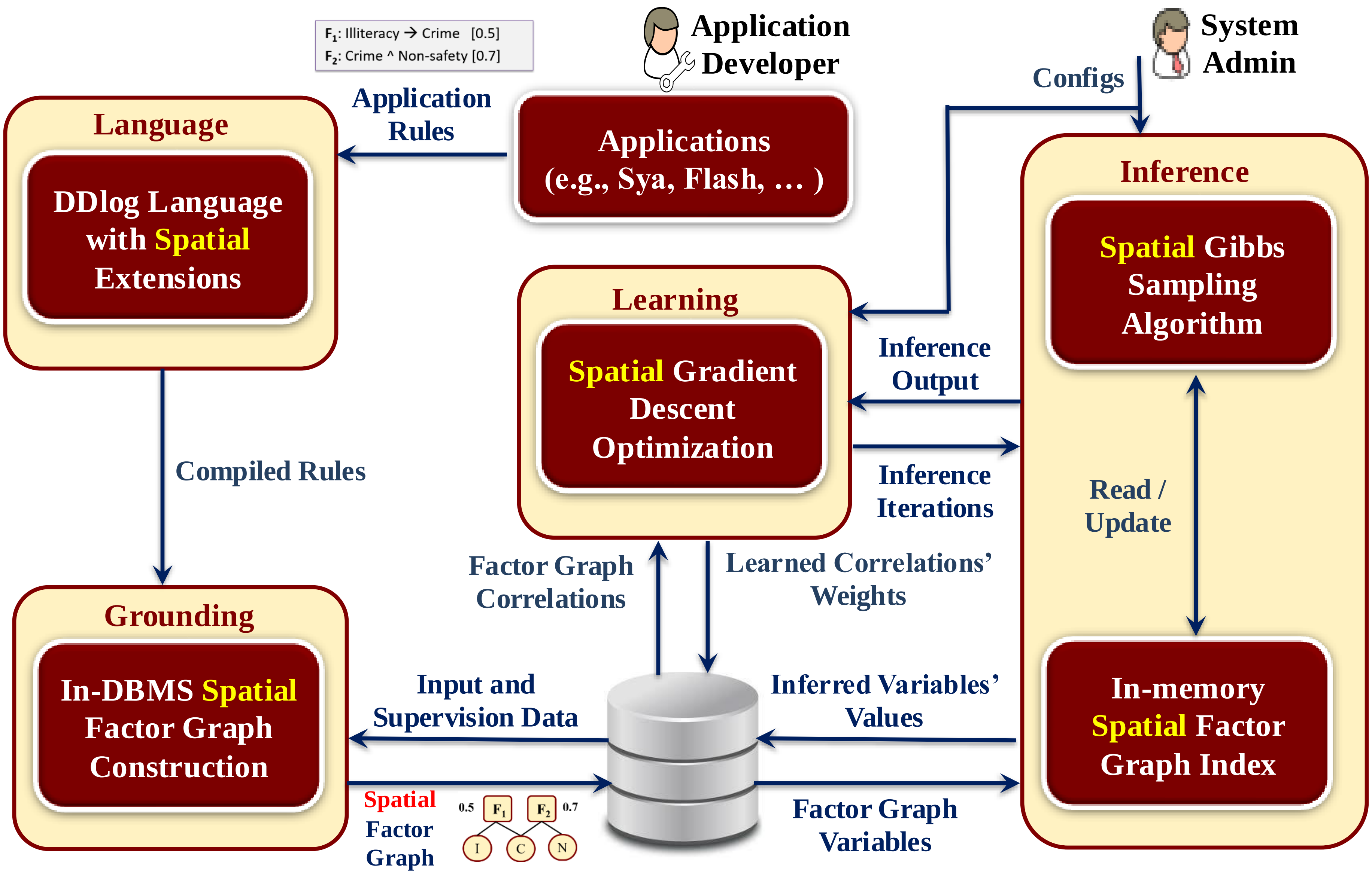
Var	Val
Joe	1
Lily	?



MLN Architecture



SMLN Architecture



Outline

- Motivation
- Introduction to Spatial Markov Logic Networks (SMLN)
- **SMLN for Knowledge Base Construction**
  - Sya: A Spatial Probabilistic Knowledge Base Construction System [ICDE’2020, SIGMOD’18]
- **SMLN for Spatial Analysis**
- **Summary**

Going Back to the Ebola Example ...

Ebola infection rates in Liberia

Inference Rules

P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
~~P3: Counties X&Y are within 150 miles (0.01)~~
~~P4: Counties X&Y are within 148.5 miles (0.02)~~
..
~~P102: Counties X&Y are within 1.5 miles (1)~~
P3: The closer Y&X the higher Y infection rate

~~**Rule:** P1&P2 ➔ Y has high infection rate~~
Rule: P1&P2&P3 ➔ Y has high infection rate

Data

County	I	S
Montserrado	1	0.6
Margibi	?	0.6
Bong	?	0.6
Gbarpolu	?	0.6

Infections

Sanitation

DeepDive

Result

County	Confidence				Ground Truth
Margibi	~~0.54~~	~~0.51~~	~~0.63~~		[0.6, 1]
Bong	~~0.52~~	~~0.45~~	~~0.48~~		[0.4, 0.6[
Gbarpolu	~~0.63~~	~~0.06~~	~~0.14~~		[0.2, 0.4[

Objective: Achieving more **accurate** confidence scores than DeepDive, while keeping the execution time **efficient**

Going Back to the Ebola Example ...

Ebola infection rates in Liberia

Inference Rules

P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
~~P3: Counties X&Y are within 150 miles (0.01)~~
~~P4: Counties X&Y are within 148.5 miles (0.02)~~
..
~~P102: Counties X&Y are within 1.5 miles (1)~~
P3: The closer Y&X the higher Y infection rate

~~**Rule:** P1&P2 ➔ Y has high infection rate~~
Rule: P1&P2&P3 ➔ Y has high infection rate

Sya

Data

County	I	S
Montserrado	1	0.6
Margibi	?	0.6
Bong	?	0.6
Gbarpolu	?	0.6

Infections

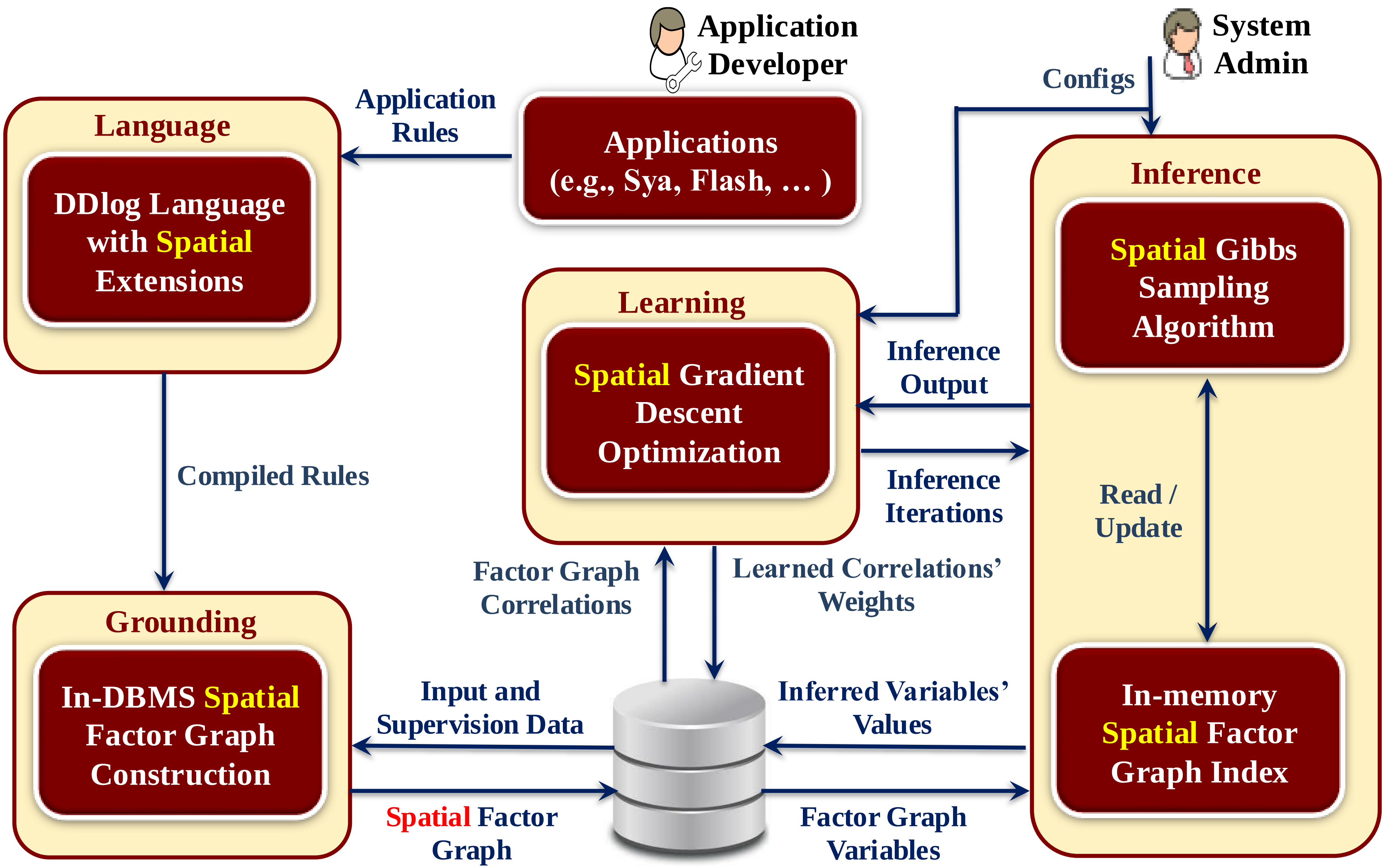
Sanitation

Result

County	Confidence				Ground Truth
Margibi	~~0.54~~	~~0.51~~	~~0.63~~	0.76	[0.6, 1]
Bong	~~0.52~~	~~0.45~~	~~0.48~~	0.53	[0.4, 0.6[
Gbarpolu	~~0.63~~	~~0.06~~	~~0.14~~	0.22	[0.2, 0.4[

Objective: Achieving more **accurate** confidence scores than DeepDive, while keeping the execution time **efficient**

SMLN Architecture



Language Module

- **Extending the DDlog language**
  - Easy to express spatial functionalities

- **Example: some rules from the Ebola KB example**

Spatial data types

#Schema Declaration
S1: County (id bigint, location **point**, hasLowSanitation bool).

@spatial(exp)
S2: HasEbola? (id bigint, location **point**).

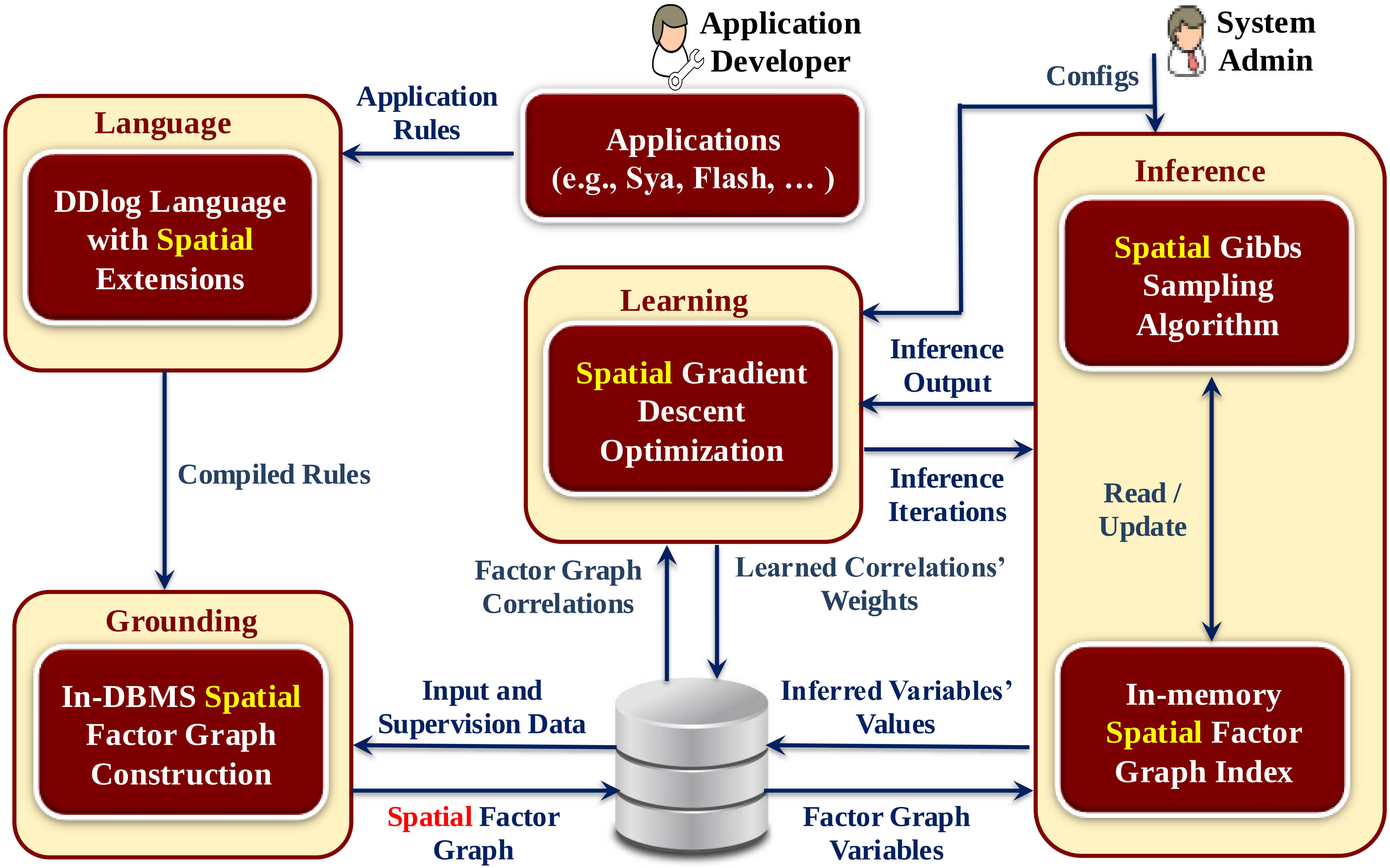
Spatial parameters

#Derivation Rule
D1: HasEbola(C1, L1) = NULL :- County(C1, L1, -).

#Inference Rule
R1: @weight (0.35)
HasEbola(C1, L1) => HasEbola(C2, L2) :- County(C1, L1, -), County(C2, L2, S2)
[**distance**(L1, L2) < 150, **within**(liberia_geom, L1), S2 = true].

Spatial functions

SMLN Architecture



Grounding Module: Spatial Factors

- **Introducing a new spatial factor type**
  - Considers the spatial correlation over variables based on their distance

Spatial variables

$$\rho_{j,k} = \begin{cases} e^{w_{d(v_j,v_k)}} & v_j = v_k \\ e^{-w_{d(v_j,v_k)}} & otherwise \end{cases}$$

Distance-based weight

- **Extended to support the categorical case**
  - Favors similar domain values from close variables

$$\rho_{j,k}(t_j,t_k) = \begin{cases} e^{w_{d(v_j,v_k)}} & v_j(t_j) = v_k(t_k) = 1, t_j = t_k \\ e^{-w_{d(v_j,v_k)}} & v_j(t_j) = v_k(t_k) = 1, t_j \neq t_k \\ 1 & otherwise \end{cases}$$

Domain values

- **Spatial factor graph**
  - Combines spatial and logical (i.e., non-spatial) factors in an efficient manner

Grounding Module: Spatial Factor Graph Construction

- **Generating the spatial factor graph using SDBMS (e.g., PostGIS)**
  - Rules are translated into spatial SQL queries
  - e.g., Rule R1 from the Ebola example

```
#Inference Rule
R1: @weight (0.35)
HasEbola(C1, L1) => HasEbola(C2, L2) :- County(C1, L1, -), County(C2, L2, S2)
        [distance(L1, L2) < 150, within(liberia_geom, L1), S2 = true].
```

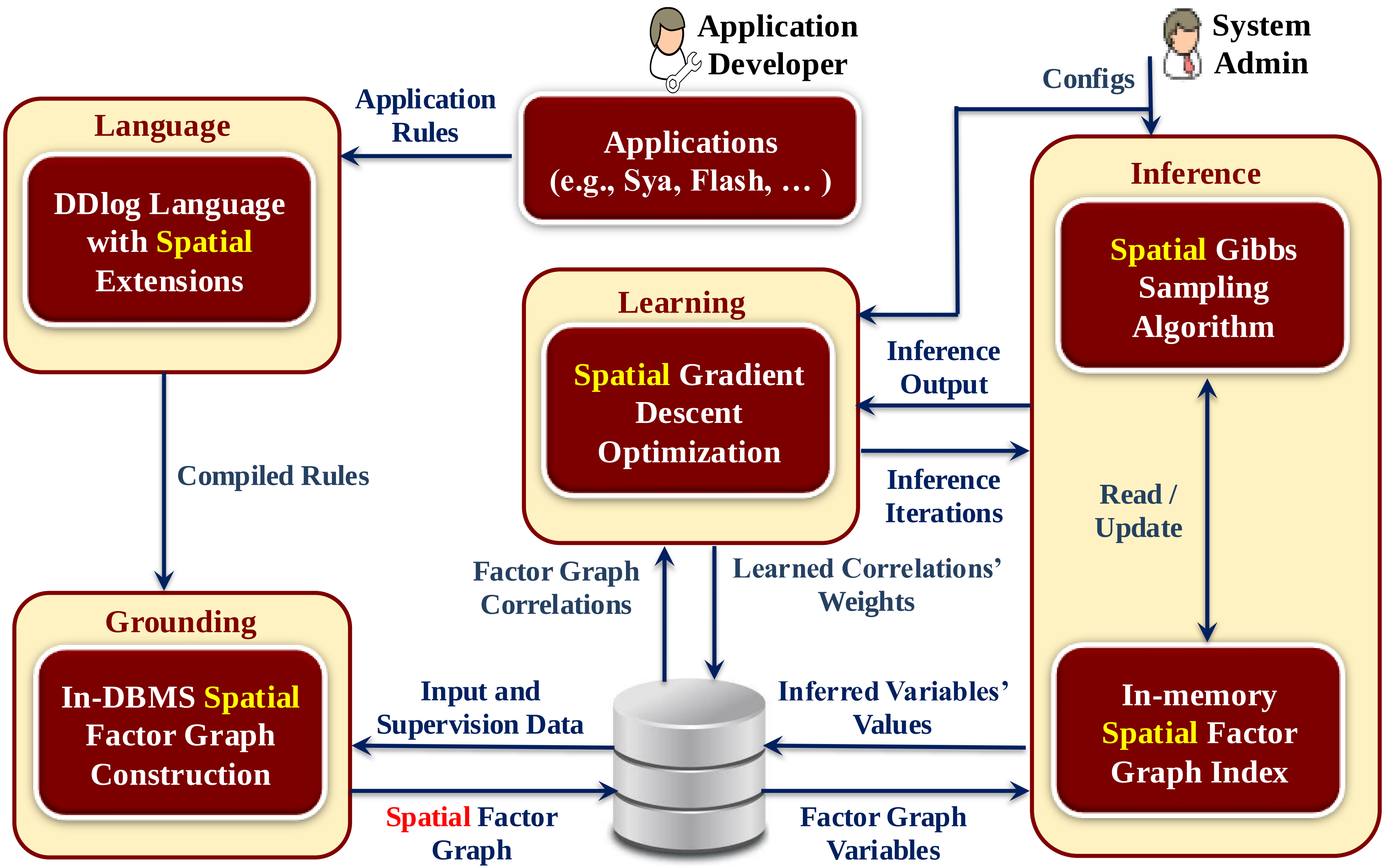
```
INSERT INTO R1_Factors (var1, var2, type, weight)
(
  SELECT C1.id AS "var1", C2.id AS "var2", "imply", 0.35
  FROM (
        SELECT * FROM County C0
        WHERE WITHIN (liberia_geom, C0.location)
        ) C1, County C2
  WHERE DISTANCE (C1.location, C2.location) < 150
        AND C2.hasLowSanitation = true
)
```

Range Query

Spatial join

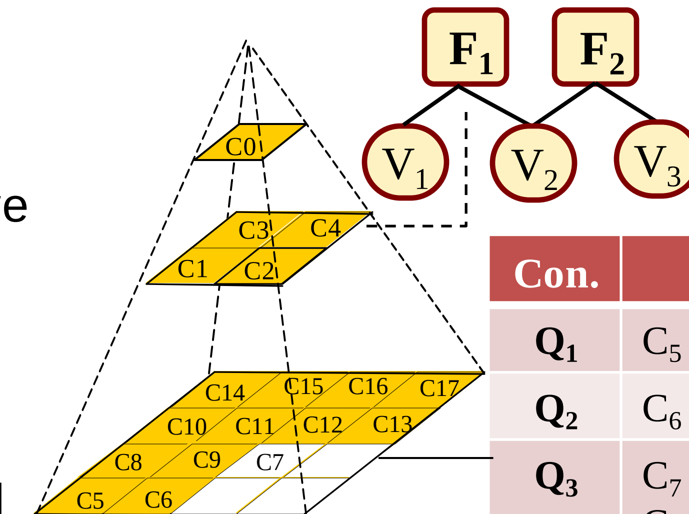
- **Two effective optimizations**
  - Providing a heuristic query optimizer (e.g., spatial queries reordering)
  - Using co-occurrence statistics to predict and remove *inactive* spatial factors based on training data

SMLN Architecture



Inference Module: Spatial Gibbs Sampling

- **Existing Gibbs sampling algorithms are inefficient**
  - Sequential or single-site sampling updates within the same epoch
  - Slow convergence when having spatial correlations
- **Spatial variation of Gibbs sampling**
  - Instead of sequential sampling, we use *concliques-based sampling*
    - A conclique is a set of locations such that no two locations are neighbors
    - Designed for sampling over[1] spatially-dependent variables
  - Guarantees both <u>efficiency</u> and <u>accuracy</u> in our case



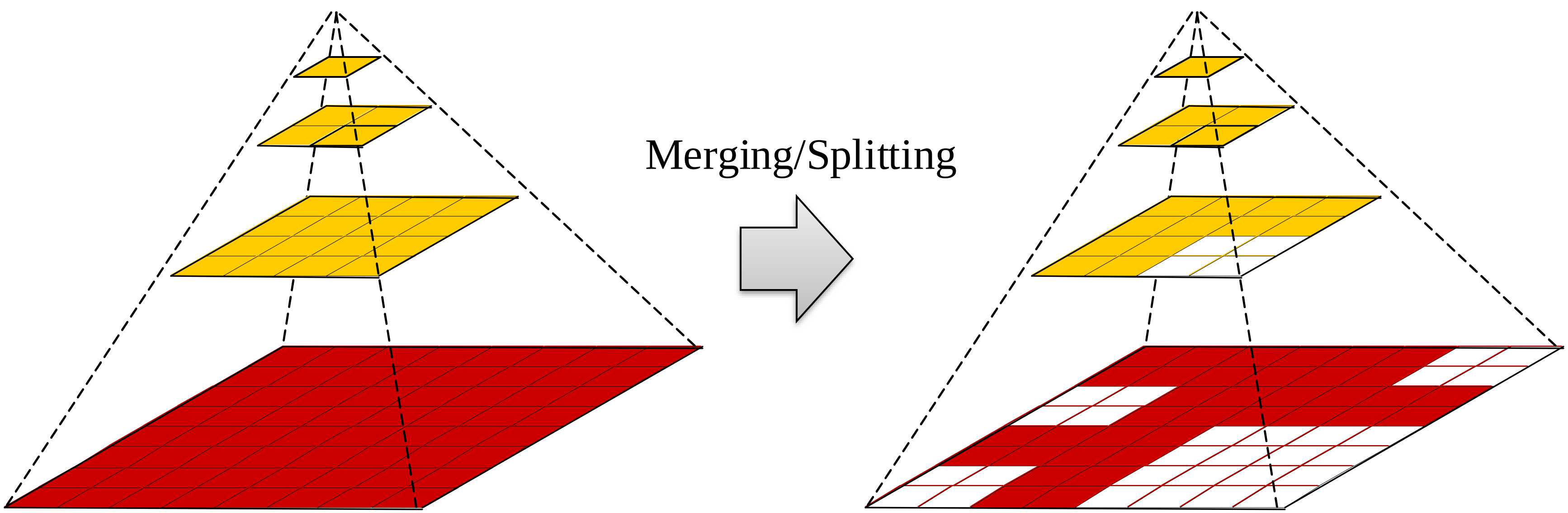
Con.	Cells
Q_1	$C_5\, C_{10}\, C_{12}$
Q_2	$C_6\, C_{11}\, C_{13}$
Q_3	$C_7\, C_8\, C_{14}\, C_{16}$
Q_4	$C_9\, C_{15}\, C_{17}$

In-memory Spatial Factor Graph Index

[1] M. Kaiser, S. Lahiri, and D. Nordman. Goodness of Fit Tests for a Class of Markov Random Field Models. The Annals of Statistics, 2012

Inference Module: Spatial Factor Graph Index Maintenance

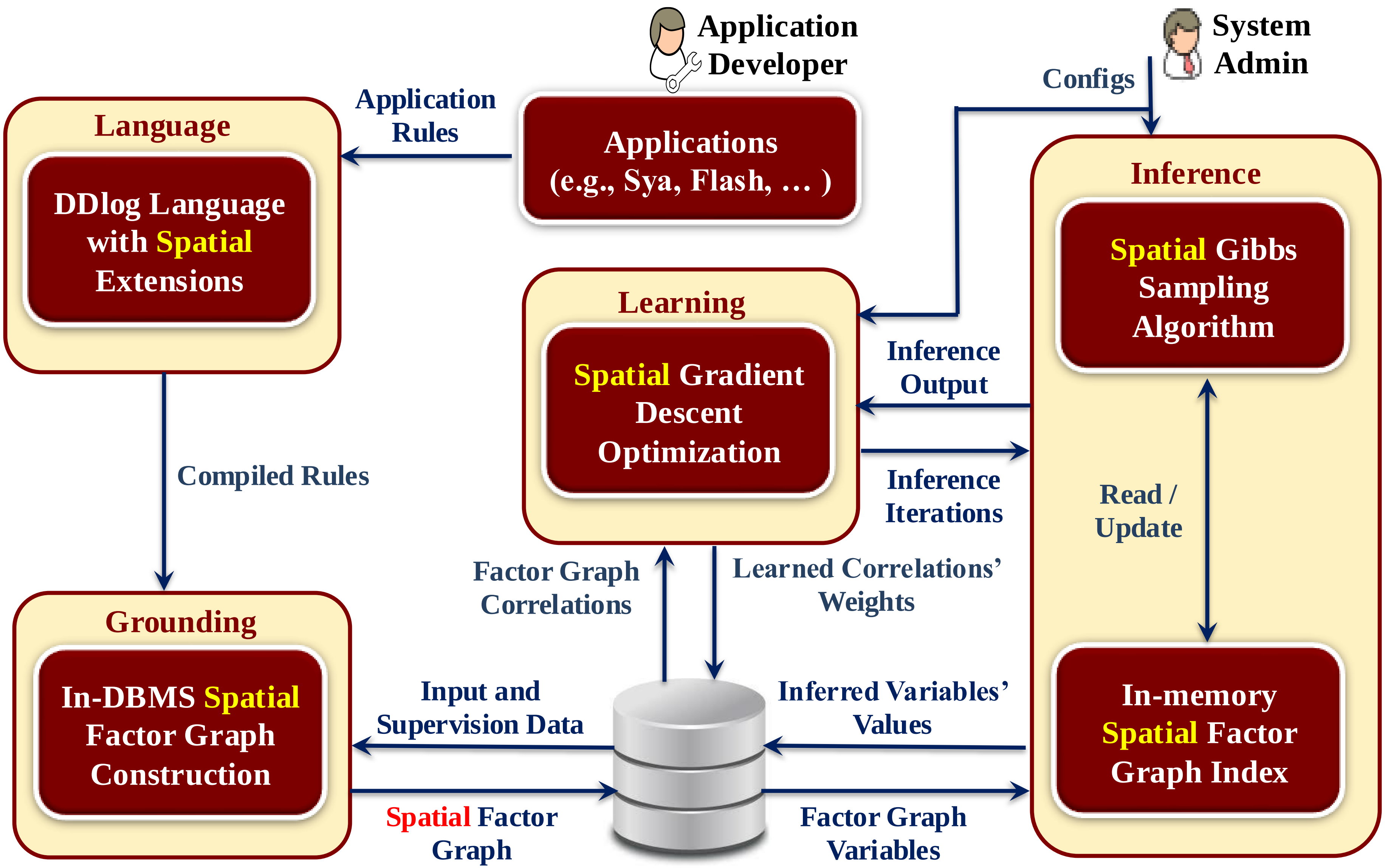
- **Merging**
  - 4-cell quadrant at level (h+1) "merged" into parent at level h
  - Merging decision made on trade-off between *locality loss* and *scalability gain*
- **Splitting**
  - Opposite operation as merging
  - Splitting decision made on trade-off between *locality gain* and *scalability loss*



Initial Complete Pyramid Index

Final Partial Pyramid Index

SMLN Architecture



Learning Module

- **Introducing the concept of *Correlation Locality***
  - Correlations between spatially close variables should have higher effect on learned weights than correlations between distant variables
  - Very important for spatial analysis applications

- **Spatial variation of gradient descent optimization**
  - We employ the inverse-weight method to weigh gradient updates [1]

$$w_s = w_s + \boxed{\frac{m(m-1)}{2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d(v_i, v_j)}}\alpha g$$

Current weight

Previous weight

Inverse-weight

Step size

Gradient sign (i.e., -1 or 1)

- **Employing parallel technique for high throughput** [2]

[1] G. Lu and D. Wong. An Adaptive Inverse-distance Weighting Spatial Interpolation Technique. In Computers and Geosciences, 2008

[2] M. Zinkevich, M. Weimer, L. Li, and A. Smola. Parallelized Stochastic Gradient Descent. In NIPS, 2010

Experimental Setup

- **Building two knowledge bases, each from different dataset**
  - KB about the water quality in Texas
    - Texas Ground Water Database (GWDB) about 9831 wells
    - 11 inference rules with spatial relationships
  - KB about the air pollution concentrations in New York
    - New York Heals and Mental Hygiene dataset (NYCCAS)
    - 5 inference rules with spatial relationships

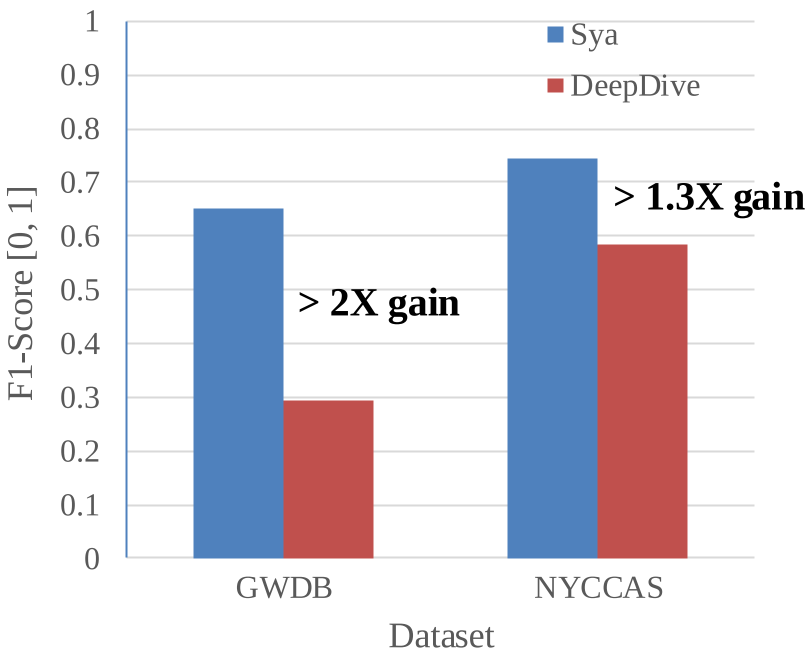
- **Evaluation metrics**
  - F1-score for quality
  - Total Inference time for scalability

- **State-of-the-art system to compare with: DeepDive**[1]

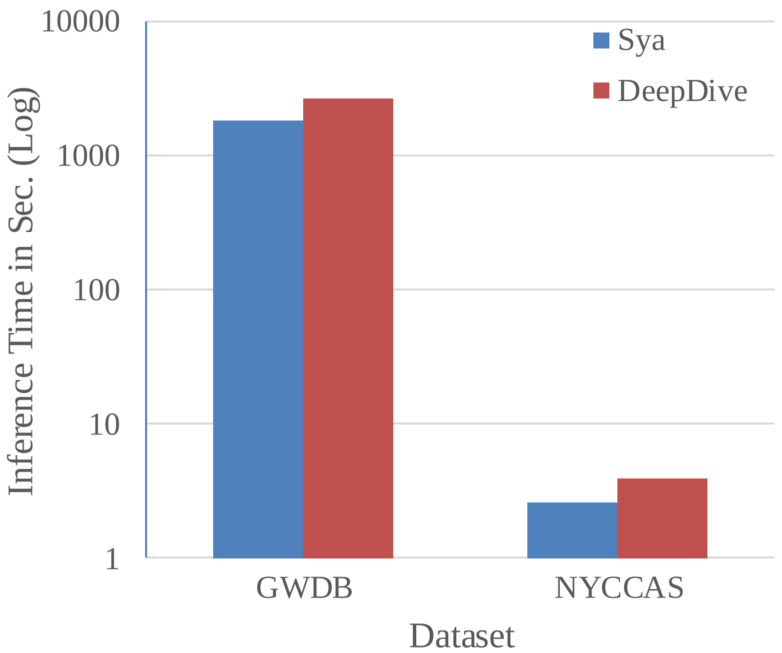
[1] J. Shin, S. Wu, F. Wang, C. D. Sa, C. Zhang, and C. Re. Incremental Knowledge Base Construction Using DeepDive. VLDB, 2015

Sya Results

Quality



Scalability



Sya can achieve <u>two times</u> accuracy gain over DeepDive, while scalability is <u>a little bit better</u>

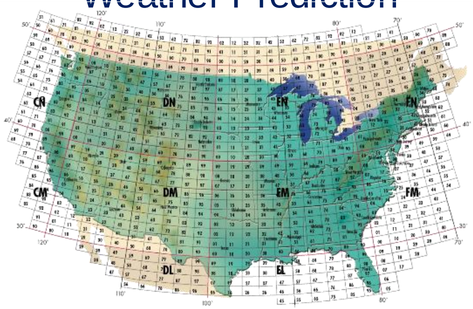
Outline

- **Motivation**
- **Introduction to Spatial Markov Logic Networks (SMLN)**
- **SMLN for Knowledge Base Construction**
- **SMLN for Spatial Analysis**
  - TurboReg: A Framework for Scaling Up Autologistic Regression Models [ACM TSAS'19, SIGSPATIAL'18]
  - Flash: Scalable Spatial Probabilistic Graphical Modeling [SIGSPATIAL Special'20, VLDB'19, SIGSPATIAL'19]
- **Summary**

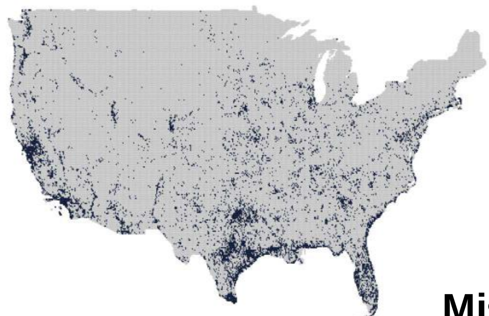
Autologistic Regression

- **Predict whether a spatial phenomenon exists or not, based on neighbor values and features**

Weather Prediction

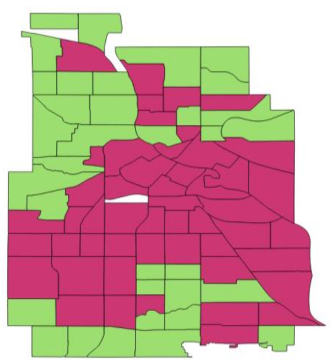


Birds Migration



Land Cover

Crimes Distribution



Features

$(x_1 = 0, x_2 = 1)$ $(x_1 = 1, x_2 = 1)$

l1 0	l2 1	l3 1	l4 1
l5 1	l6 1	l7 0	l8 1
l9 1	l10 1	l11 0	l12 0
l13 1	l14 ?	l15 0	l16 0

$(x_1 = 1, x_2 = 0)$

Phenomenon value

Missing value

Features

Neighbor values

$$\log \frac{Pr(z_i=1|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i=0|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})} = \sum_{j=1}^{m} \beta_j x_j + \eta \sum_{k\in\mathcal{N}_i} z_k$$

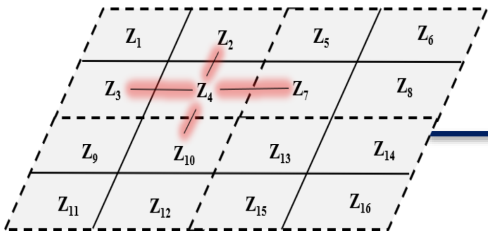
Regression Parameters

Learning regression parameters for 80K cells takes more than one day ☹

TurboReg Using SMLN

$$\log \frac{Pr(z_i=1|\mathcal{X},\mathcal{Z}_{N_i})}{Pr(z_i=0|\mathcal{X},\mathcal{Z}_{N_i})} = \sum_{j=1}^{m} \beta_j x_j + \eta \sum_{k\in\mathcal{N}_i} z_k$$

Regression Equation → SMLN Transformation → SMLN Rules → SMLN Engine → Rule weights = Regression Parameters



→ SMLN Transformation → SMLN Rules → SMLN Engine → β_1 , η

$$\log \frac{Pr(z_i=1|\mathcal{X},\mathcal{Z}_{N_i})}{Pr(z_i=0|\mathcal{X},\mathcal{Z}_{N_i})} = \beta_1 x_1 + \eta \sum_{k\in\mathcal{N}_i} z_k$$

$$\boxed{\log \frac{Pr(z_1=1)}{Pr(z_1=0)} = \beta_1 x_1 + \eta z_2 + \eta z_3}$$

⋮

$$\log \frac{Pr(z_4=1)}{Pr(z_4=0)} = \beta_1 x_1 + \eta z_2 + \eta z_3 + \eta z_7 + \eta z_{10}$$

⋮

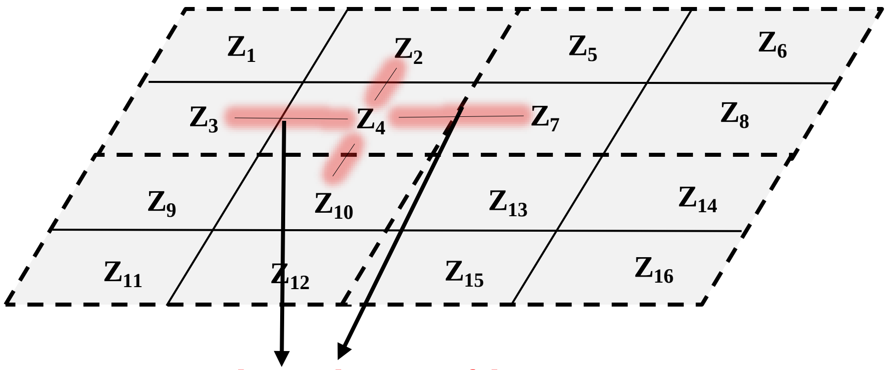
$$\log \frac{Pr(z_{16}=1)}{Pr(z_{16}=0)} = \beta_1 x_1 + \eta z_{14} + \eta z_{15}$$

SMLN Rules
[Z_1 ^ X_1, β_1]
[Z_1 ^ Z_2, η]
[Z_1 ^ Z_3, η]
[Z_2 ^ X_1, β_1]
[Z_2 ^ Z_4, η]
[Z_2 ^ Z_5, η]
[Z_3 ^ X_1, β_1]
[Z_3 ^ Z_4, η]
…….

Theoretical proof of the Autologistic Regression-SMLN equivalence is in the paper

Two More Benefits

Generalized Models



- **Conditional dependency**
  - Traditional model:

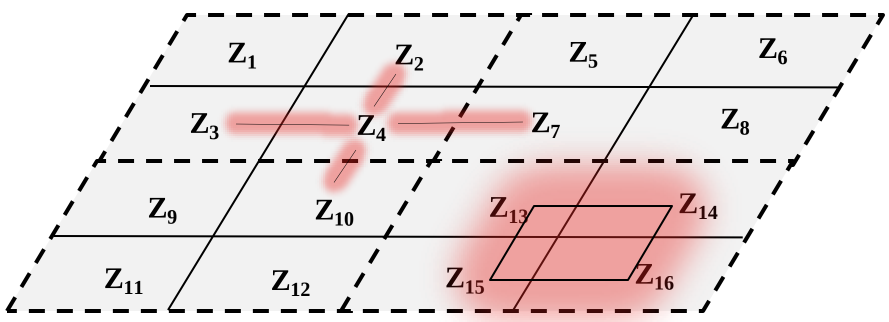
    $Z_2 + Z_3 + Z_7 + Z_{10}$ ✗
  - TurboReg model:

    $\boxed{(Z_3 \vee Z_7) \wedge Z_4}$ ✓

    $Z_2 \wedge Z_4$

    $Z_{10} \wedge Z_4$

Higher-degree Interactions



- **Complex dependency**
  - Traditional model:
    - Expensive matrix computations ✗
  - TurboReg model:
    - Same computation, yet, with longer factors

      $Z_{13} \wedge Z_{14} \wedge Z_{15} \wedge Z_{16}$ ✓

Multinomial Autologistic Regression

- **Prediction and feature variables are multinomial (i.e., categorical)**
  - Domain values are predefined values (e.g., {0, 1, 2})
  - Represent <u>each multinomial</u> variable with <u>a set of binary</u> variables

Pivot

Prediction

$$\{0,1,2\} \quad z_i \equiv \begin{array}{l} \boxed{z_i(0)} \\ z_i(1) \\ z_i(2) \end{array} \quad \{0,1\}$$

Feature

$$\{0,1,2\} \quad x_j \equiv \begin{array}{ccc} x_j^{0,0} & x_j^{0,1} & x_j^{0,2} \\ x_j^{1,0} & x_j^{1,1} & x_j^{1,2} \\ x_j^{2,0} & x_j^{2,1} & x_j^{2,2} \end{array} \quad \{0,1\}$$

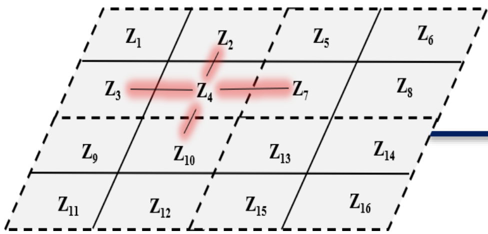
$$\begin{cases} \log \frac{Pr(z_i(1)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i(0)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})} = \sum_{j=1}^{m} \sum_{t \in \mathcal{D}_{x_j}} \beta_j^{1,t} x_j^{1,t} + \sum_{k \in \mathcal{N}_i} \sum_{s \in \mathcal{D}_{z_k}} \eta_{1,s} z_k(s) \\ \log \frac{Pr(z_i(2)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i(0)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})} = \sum_{j=1}^{m} \sum_{t \in \mathcal{D}_{x_j}} \beta_j^{2,t} x_j^{2,t} + \sum_{k \in \mathcal{N}_i} \sum_{s \in \mathcal{D}_{z_k}} \eta_{2,s} z_k(s) \end{cases}$$

$$Pr(z_i(0)=1) = 1 - Pr(z_i(1)=1) - Pr(z_i(2)=1)$$

RegRocket: Multinomial Case Using SMLN

$$\log \frac{Pr(z_i=1|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i=0|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})} = \sum_{j=1}^{m} \beta_j x_j + \eta \sum_{k \in \mathcal{N}_i} z_k$$

Regression Equation → SMLN Transformation → SMLN Rules → SMLN Engine → Rule weights = Regression Parameters



→ SMLN Transformation → SMLN Rules → SMLN Engine → β_1 , η

$$\log \frac{Pr(z_i=1|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i=0|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})} = \beta_1 x_1 + \eta \sum_{k \in \mathcal{N}_i} z_k$$

$$\boxed{\begin{cases} \log \frac{Pr(z_1(1)=1)}{Pr(z_1(0)=1)} = \sum\limits_{t \in \mathcal{D}_{x_1}} \beta_1^{1,t} x_1^{1,t} + \sum\limits_{s \in \mathcal{D}_{z_2}} \eta_{1,s}[z_2(s) + z_3(s)] \\ \log \frac{Pr(z_1(2)=1)}{Pr(z_1(0)=1)} = \sum\limits_{t \in \mathcal{D}_{x_1}} \beta_1^{2,t} x_1^{2,t} + \sum\limits_{s \in \mathcal{D}_{z_2}} \eta_{2,s}[z_2(s) + z_3(s)] \end{cases}}$$

:

$$\begin{cases} \log \frac{Pr(z_{16}(1)=1)}{Pr(z_{16}(0)=1)} = \sum\limits_{t \in \mathcal{D}_{x_1}} \beta_1^{1,t} x_1^{1,t} + \sum\limits_{s \in \mathcal{D}_{z_{14}}} \eta_{1,s}[z_{14}(s) + z_{15}(s)] \\ \log \frac{Pr(z_{16}(2)=1)}{Pr(z_{16}(0)=1)} = \sum\limits_{t \in \mathcal{D}_{x_1}} \beta_1^{2,t} x_1^{2,t} + \sum\limits_{s \in \mathcal{D}_{z_{14}}} \eta_{2,s}[z_{14}(s) + z_{15}(s)] \end{cases}$$

SMLN Rules
[Z_1 (1) ^ $X_1^{1,0}$, $\beta_1^{1,0}$]
[Z_1 (1) ^ $X_1^{1,1}$, $\beta_1^{1,1}$]
[Z_1 (1) ^ $X_1^{1,2}$, $\beta_1^{1,2}$]
…….
[Z_1 (1) ^ Z_2 (0), $\eta_{1,0}$]
[Z_1 (1) ^ Z_2 (1), $\eta_{1,1}$]
[Z_1 (1) ^ Z_2 (2), $\eta_{1,2}$]
[Z_1 (1) ^ Z_3 (0), $\eta_{1,0}$]
[Z_1 (1) ^ Z_3 (1), $\eta_{1,1}$]
[Z_1 (1) ^ Z_3 (2), $\eta_{1,2}$]
…….

An extended theorem is provided for the equivalence of multinomial case as well

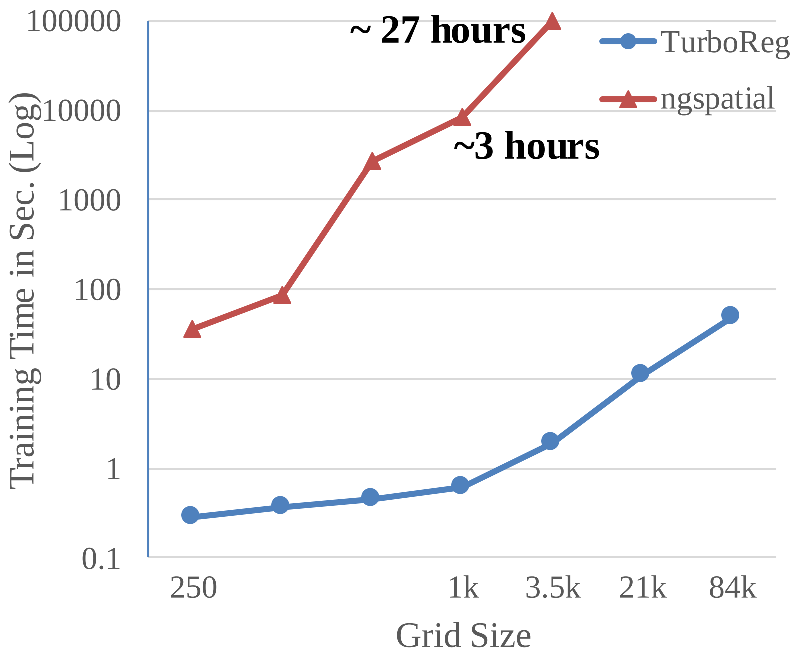
Experimental Setup

- **Three datasets, different variations, different data sizes**
  - Ebird dataset, with 3 predictors, ranging from 250 to 84K cells
  - MNCrime dataset, covering 87 neighborhoods, with 11 binary predictors
  - MNLandCover dataset, with 3 predictors, ranging from 1K to 1M cells
- **Parameters and configurations**
  - 85% training and 15% testing
  - 7 threads, 200 factor graph grid partitions
- **Evaluation metrics**
  - Total training time
  - Ratio of correctly predicted cells
  - F1-score
- **State-of-the-art system to compare with: ngspatial**[1]

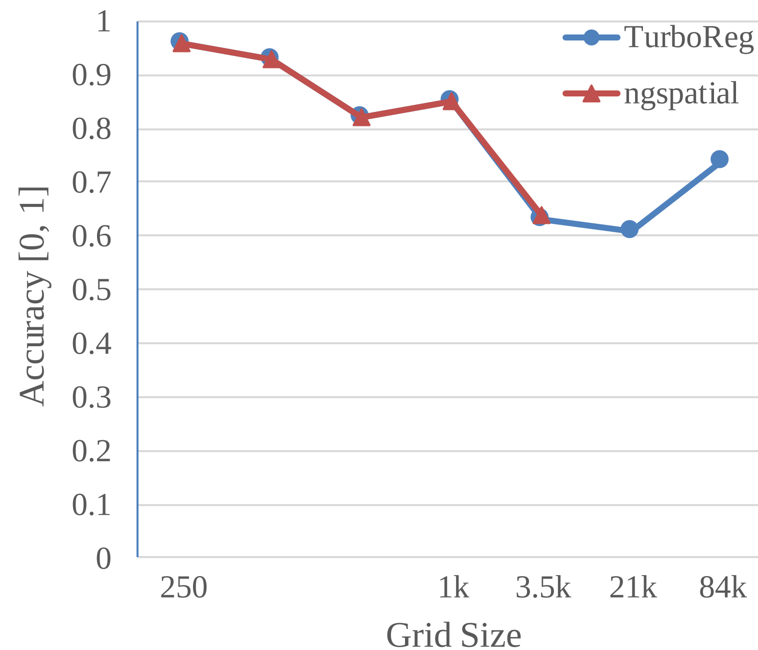
[1] John Hughes. ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data. The R Journal, 2014

TurboReg Results

Scalability



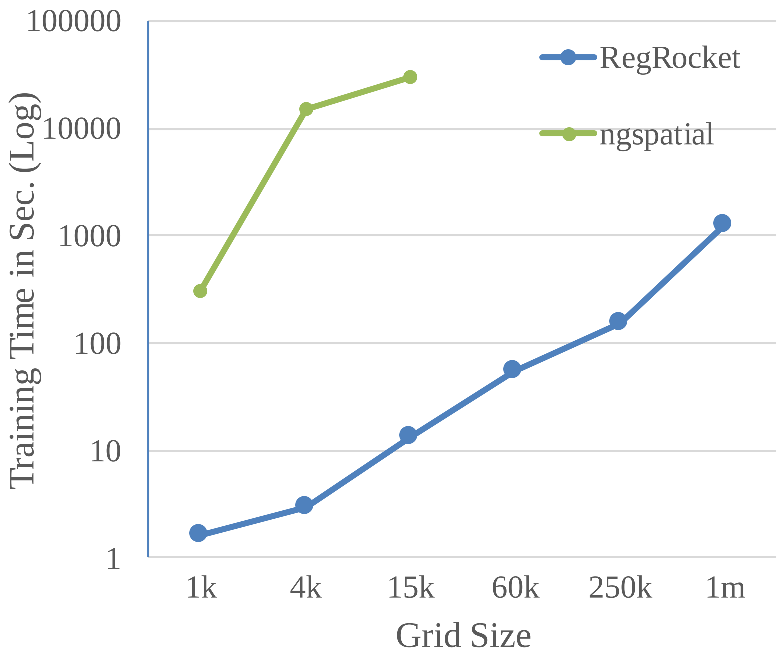
Accuracy



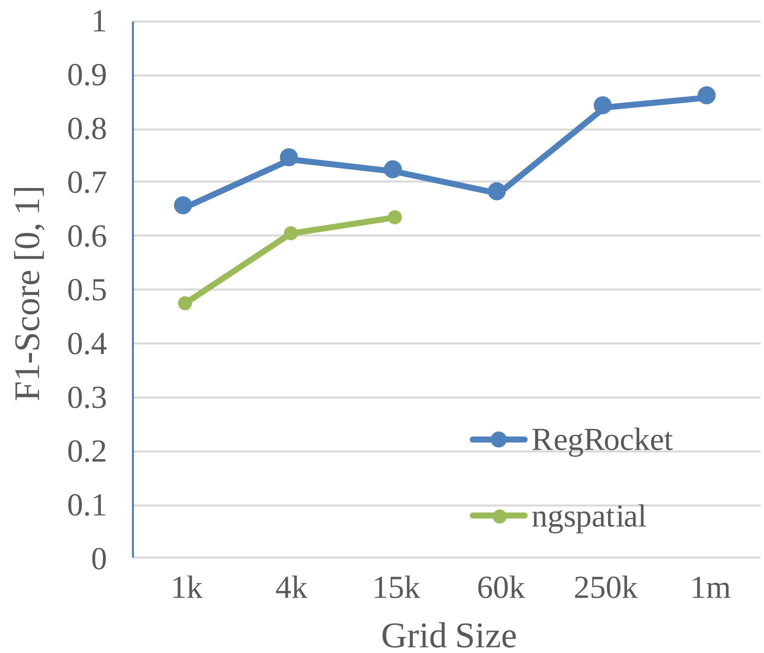
TurboReg achieves at least three orders of magnitude performance gain, while accuracy is almost the same

RegRocket Results

Scalability



F1-Score



RegRocket can handle 1 million grid cells in <u>few minutes</u> only and with <u>30% average F1-score improvement</u>

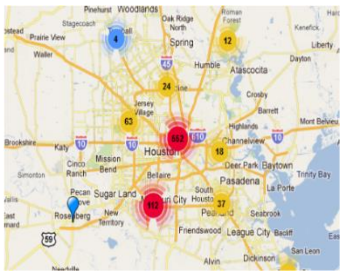
Outline

- Motivation
- Introduction to Spatial Markov Logic Networks (SMLN)
- SMLN for Knowledge Base Construction
- **SMLN for Spatial Analysis**
  - TurboReg: A Framework for Scaling Up Autologistic Regression Models [ACM TSAS’19, SIGSPATIAL’18]
  - Flash: Scalable Spatial Probabilistic Graphical Modeling [SIGSPATIAL Special’20, VLDB’19, SIGSPATIAL’19]
- **Summary**

Spatial Probabilistic Graphical Modeling (SPGM)

- **Performing *uncertain* (i.e., prob.) predictions over spatial data**
  - Classical ML approaches (e.g., regression) ignore the probabilistic relationships

Disaster Analysis

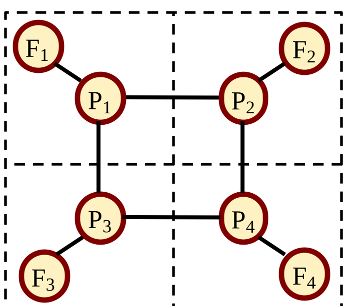


Crime Analysis

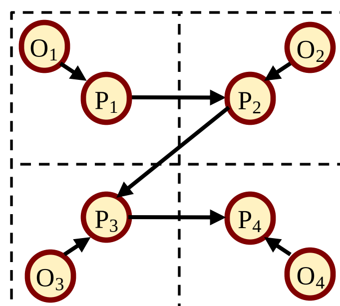
Public Health Monitoring

Geo-tagged Ads

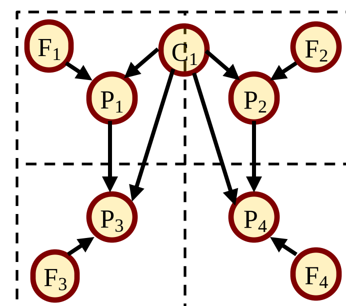
- **Representing the world as a collection of *random variables* with joint probabilistic distribution**
  - Tasks: learning the distribution, and inferring unknown variables via the distribution



Spatial Markov Random Field (SMRF)



Spatial Hidden Markov Model (SHMM)



Spatial Bayesian Network (SBN)

SPGM Challenges

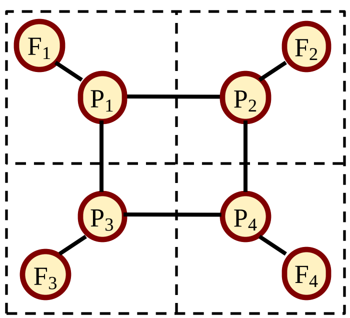
- **Scalability Issue**
  - Existing SPGM solutions can not scale beyond prototypes over small spatial datasets
    - E.g., existing SMRF solutions take more than 24 hours to perform learning and inference over 80k grid cells

- **Reusability Issue**
  - Existing SPGM solutions are tailored for domain-specific applications
    - A developer would need to re-implement and optimize the same solution for different applications

We need to employ scalable ML frameworks (e.g., SMLN) to build SPGM models with efficient learning and inference operations

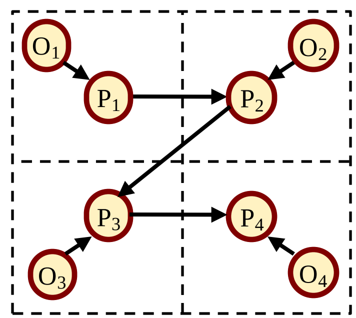
Flash using SMLN

- **Generates an equivalent set of weighted rules containing logical predicates for any SPGM input**
  - Weights represent the original SPGM parameters
  - Rules follow the syntax of the DDlog logic programming framework



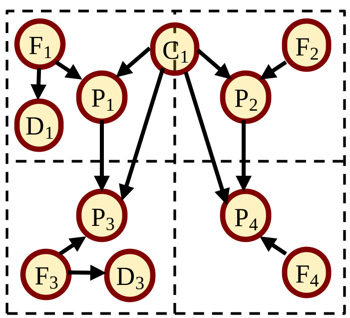
Spatial Markov Random Field (SMRF)

MLN Rules
$[P_1 \wedge F_1, ß_1]$
$[P_1 \wedge P_2, \eta]$
$[P_1 \wedge P_3, \eta]$
$[P_2 \wedge F_2, ß_1]$
$[P_2 \wedge P_4, \eta]$
…….



Spatial Hidden Markov Model (SHMM)

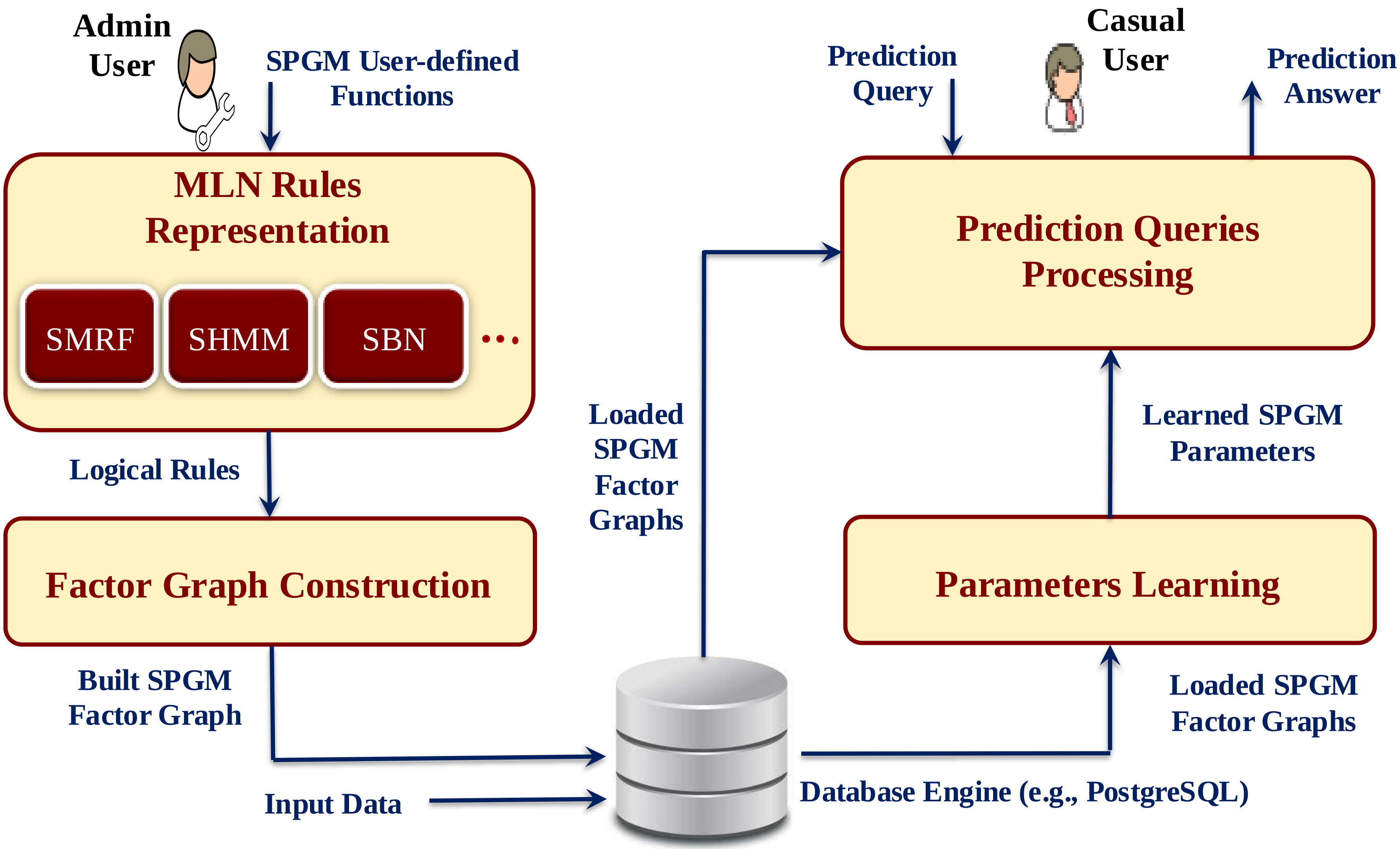
MLN Rules
$[O_1 \rightarrow P_1, b]$
$[P_1 \rightarrow P_2, a]$
$[O_2 \rightarrow P_2, b]$
$[P_2 \rightarrow P_3, a]$
$[O_3 \rightarrow P_3, b]$
…….



Spatial Bayesian Network (SBN)

MLN Rules
$[!P_1 \vee !F_1 \vee !C_1]$
$[!P_3 \vee !P_1 \vee !F_3 \vee !C_1]$
$[!P_2 \vee !F_2 \vee !C_1]$
$[!P_4 \vee !P_2 \vee !F_4 \vee !C_1]$
$[!D_1 \vee !F_1]$
…….

Flash Architecture



Experimental Setup

- **Three SPGM applications, with three different datasets**
  - Bird monitoring: SMRF model + Ebird dataset
    - Competitor: ngspatial [1]
  - Safety analysis: SHMM model + Chicago crime dataset
    - Competitor: shmm[2]
  - Land use change tracking: SBN model + Minnesota land cover dataset
    - Competitor: bnspatial[3]
- **Training and testing configurations**
  - 85% training and 15% testing
- **Evaluation metrics**
  - Learning time (Scalability), and ratio of correctly predicted cells (Accuracy)

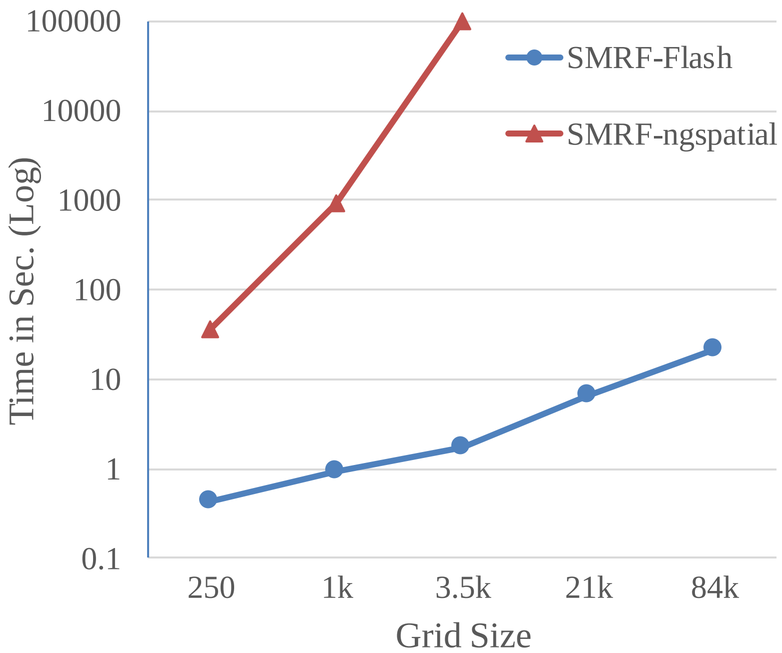
[1] John Hughes. ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data. The R Journal, 2014

[2] shmm: An R Implementation of Spatial Hidden Markov Models. github.com/mawp/shmm, 2019

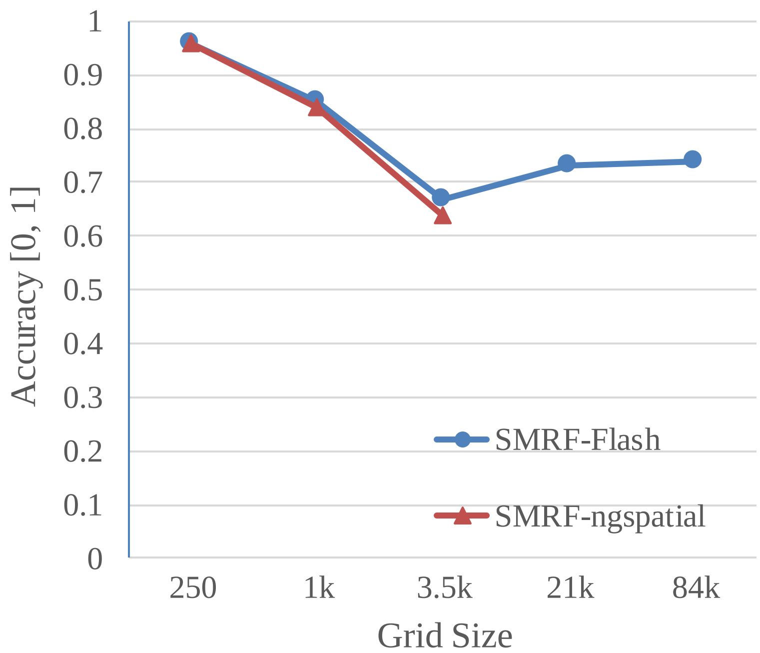
[3] bnspatial: Spatial Implementation of Bayesian Networks. cran.r-project.org/web/packages/bnspatial, 2019

Flash Results

Scalability



Accuracy



Flash is at least <u>two orders of magnitude faster</u> than state-of-the-art computational methods in learning SPGM parameters

Summary

SMLN

Sya | TurboReg | Flash

Language

DDlog Language with Spatial Extensions

Grounding

Spatial Factor Graph

Learning

Spatial Gradient Descent Optimization

Inference

Spatial Gibbs Sampling Algorithm

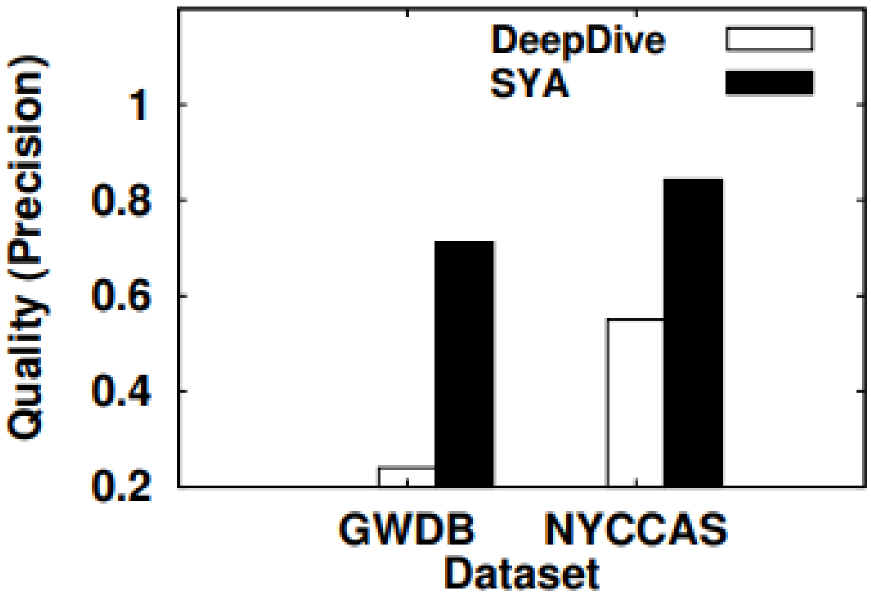
In-memory Spatial Factor Graph Index

Thank You

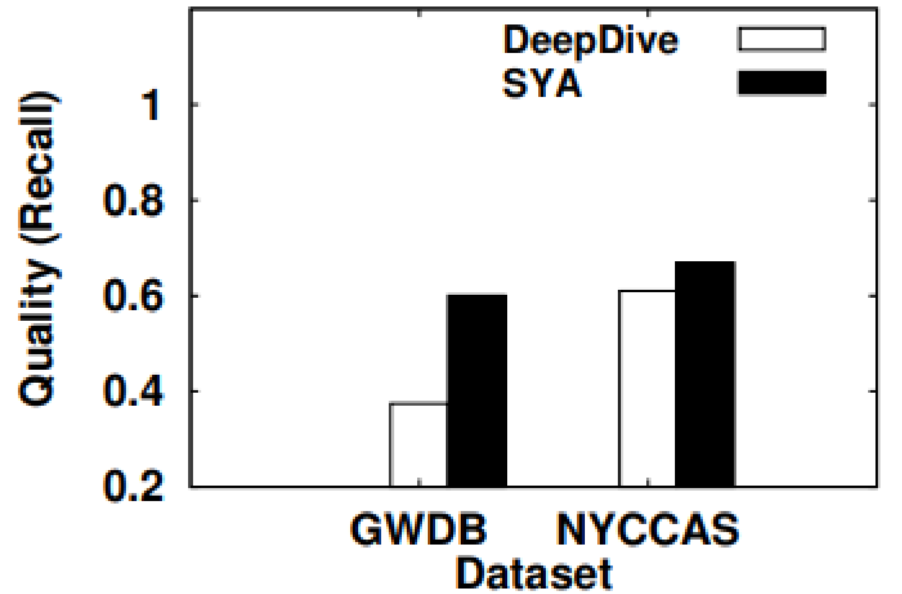
Questions

Sya Results – Extension (1/7)

- **Precision and Recall**



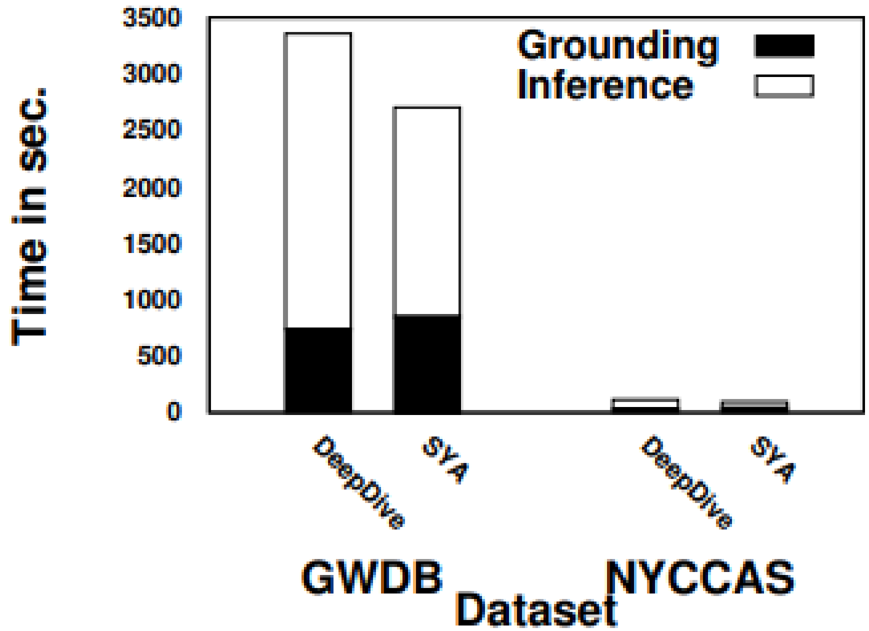
Dataset vs. Precision



Dataset vs. Recall

Sya Results – Extension (2/7)

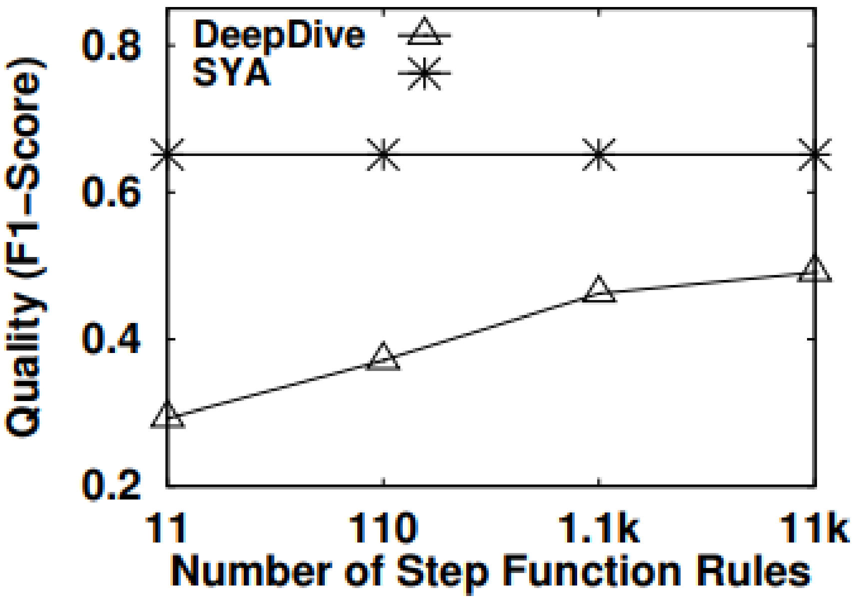
- Execution time breakdown



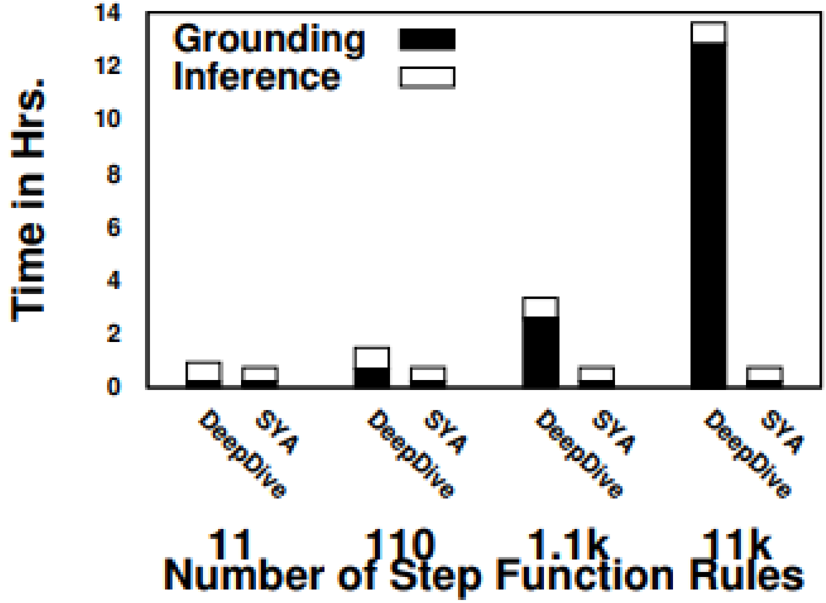
Dataset vs. Execution Time

Sya Results – Extension (3/7)

- Effect of number of step function rules in DeepDive



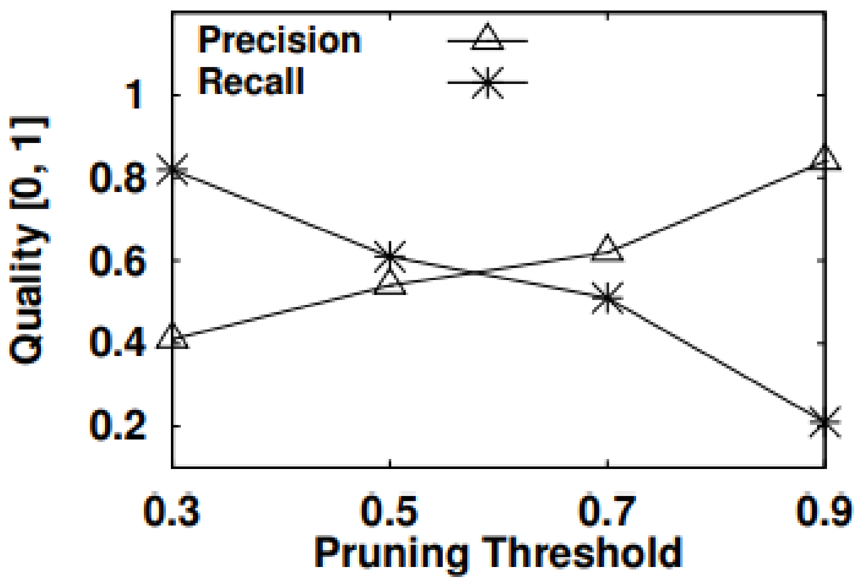
Accuracy



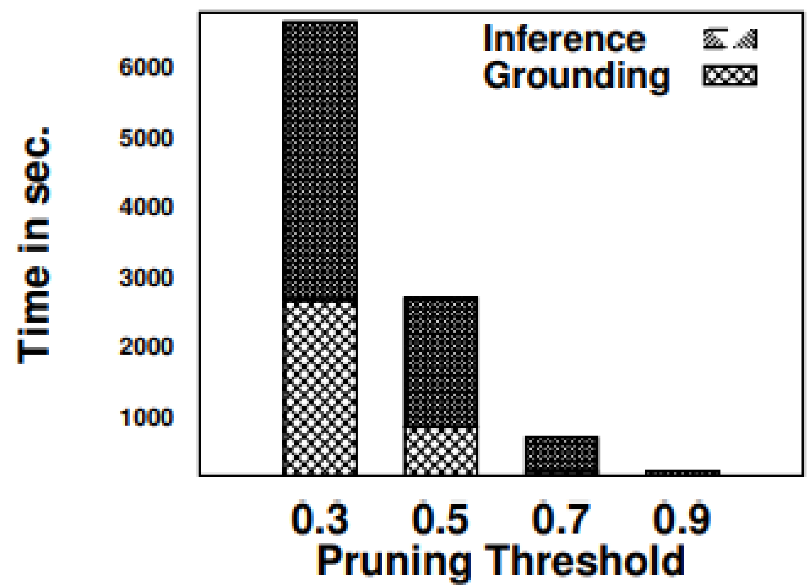
Scalability

Sya Results – Extension (4/7)

- Effect of pruning threshold



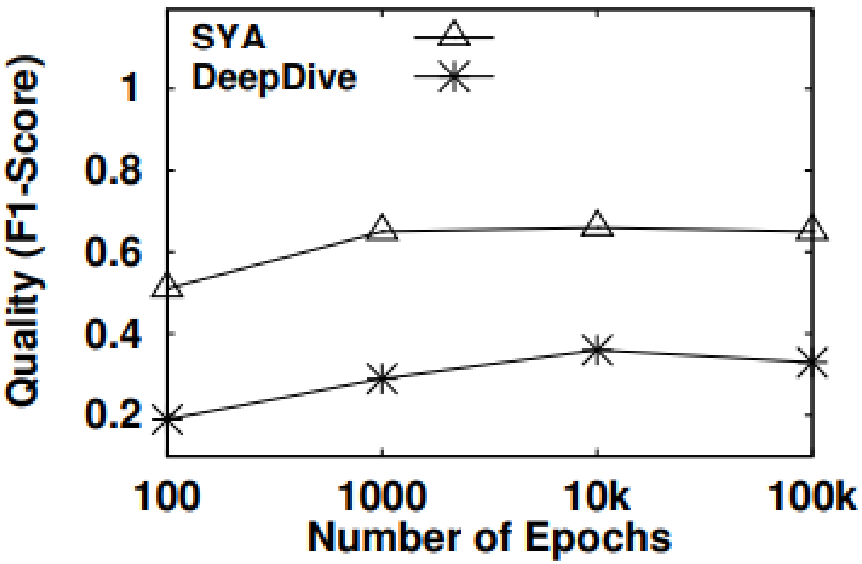
Accuracy



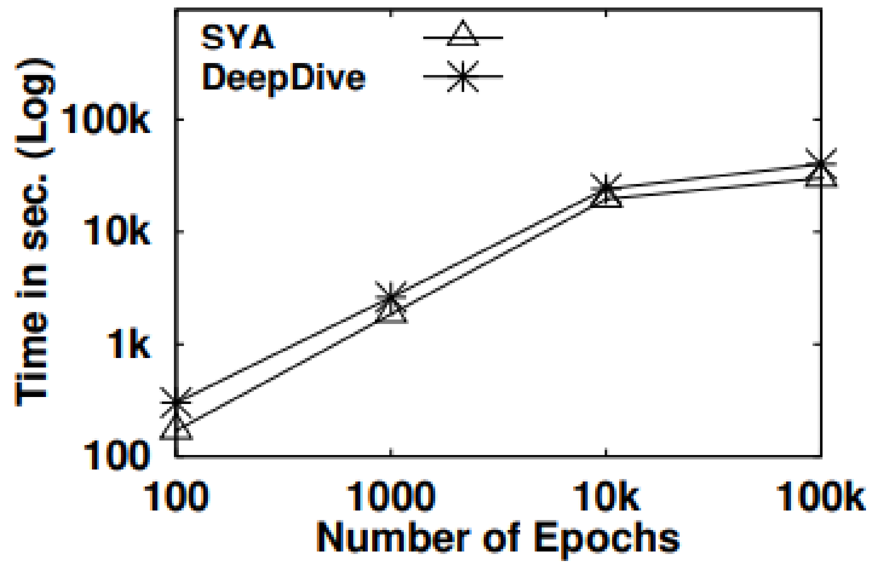
Scalability

Sya Results – Extension (5/7)

- Effect of inference epochs



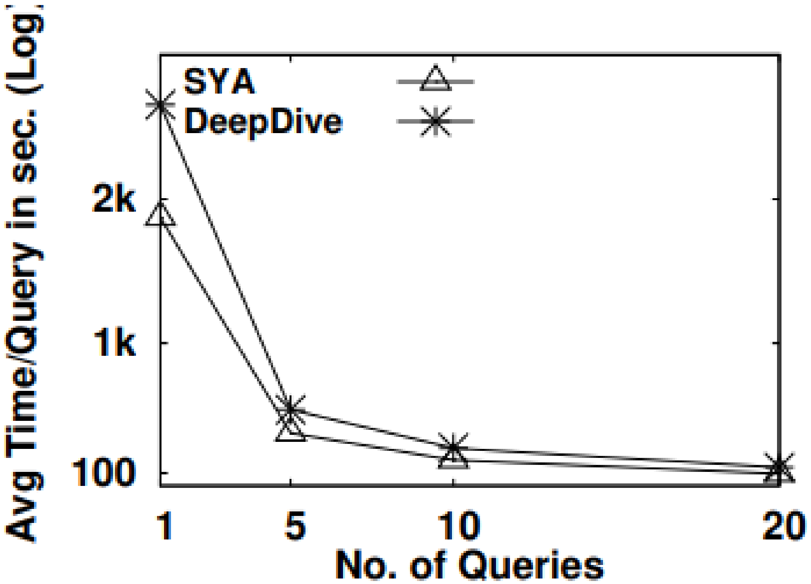
Accuracy



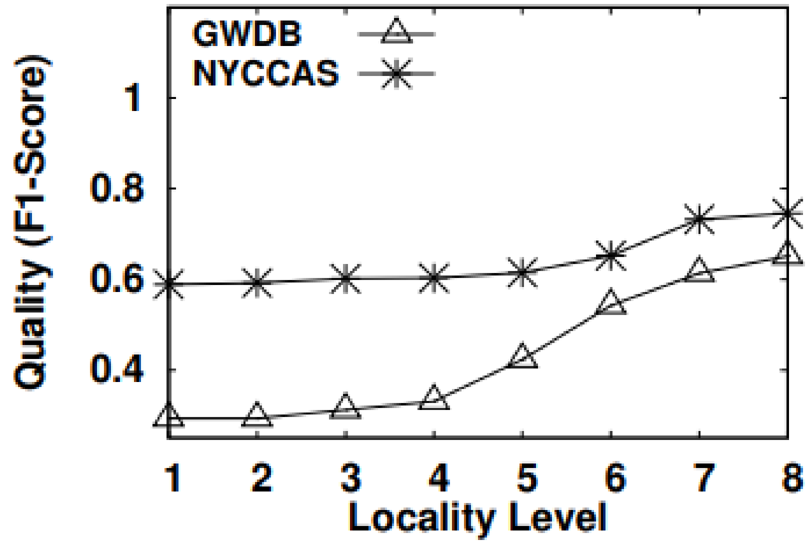
Inference Time

Sya Results – Extension (6/7)

- **Effect of incremental inference and locality level**



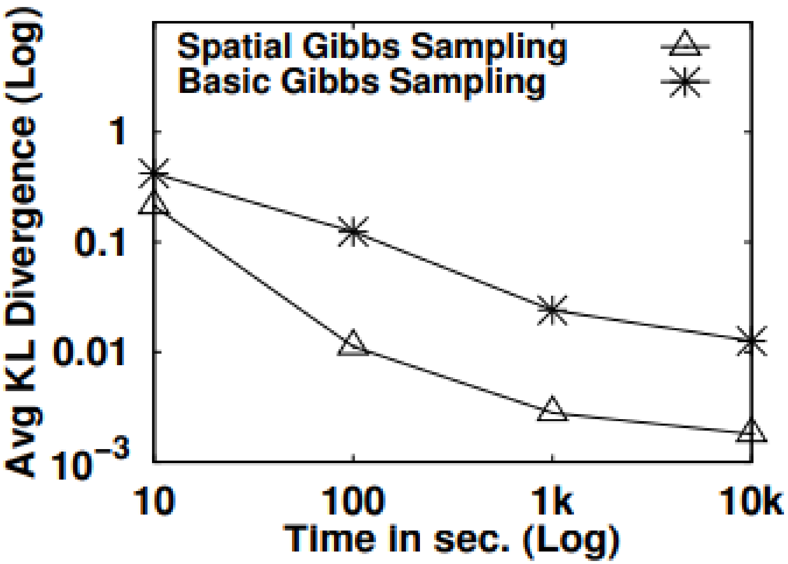
Incremental Inference



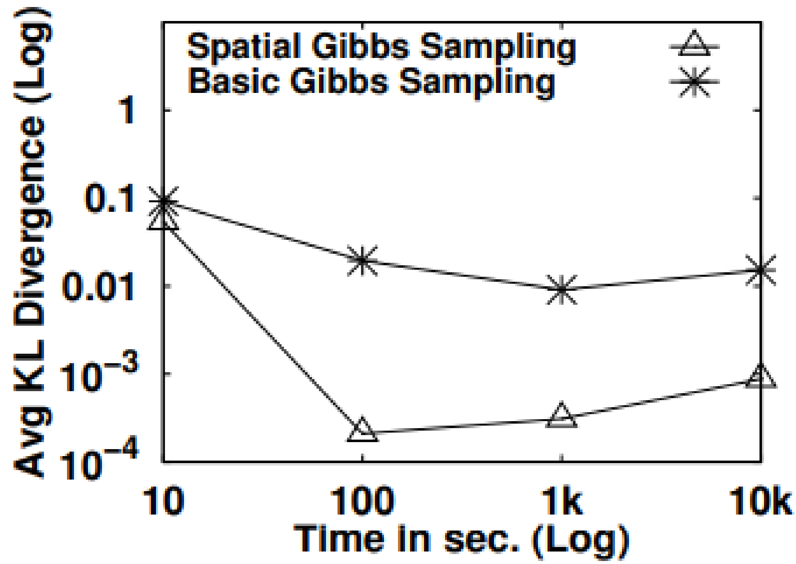
Locality Level

Sya Results – Extension (7/7)

- Spatial Gibbs sampling



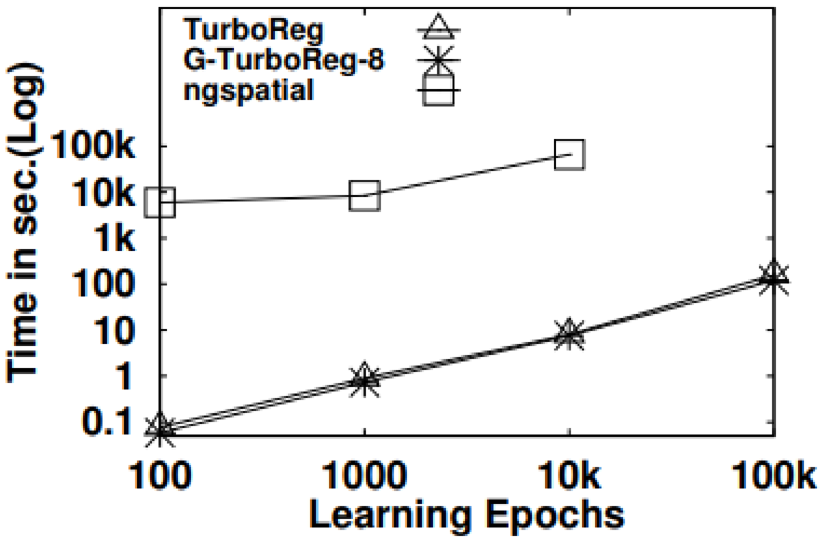
GWDB Dataset

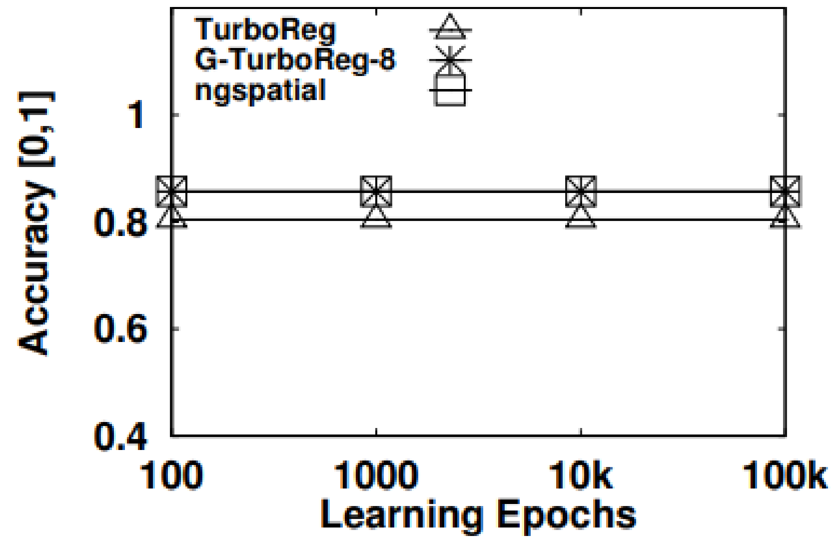


NYCCAS Dataset

TurboReg Results – Extension (1/3)

- Effect of learning epochs



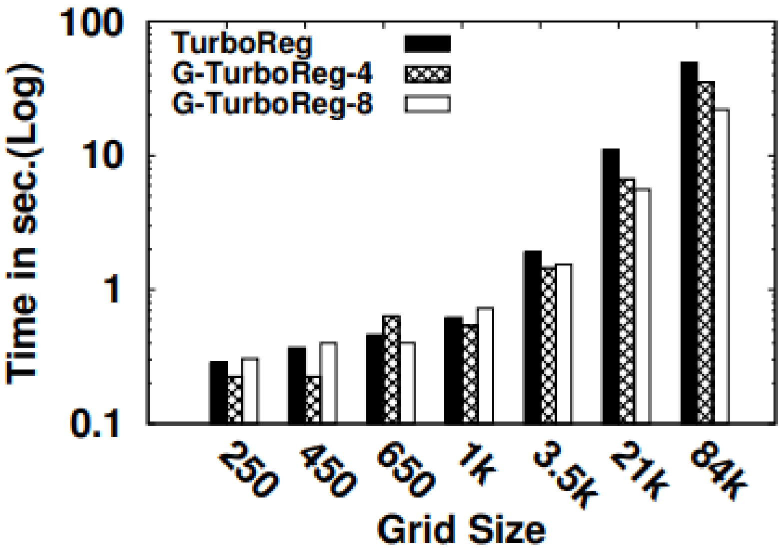


Scalability

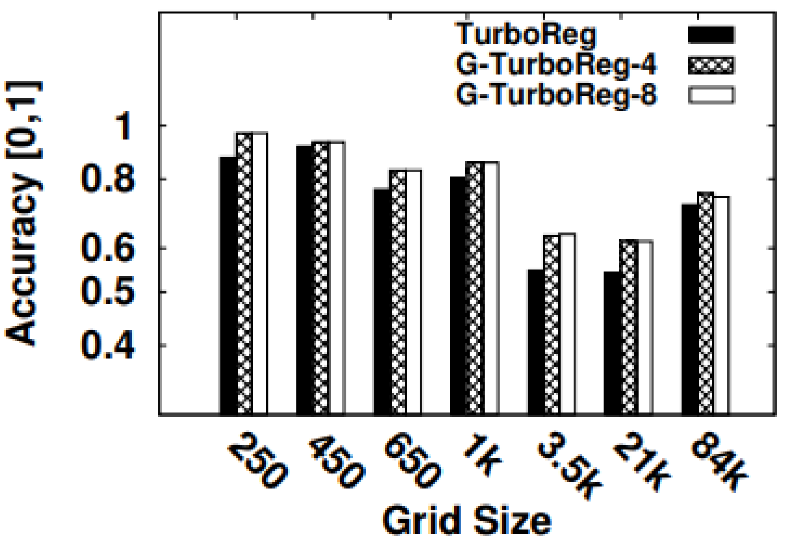
Accuracy

TurboReg Results – Extension (2/3)

- Effect of neighborhood degree



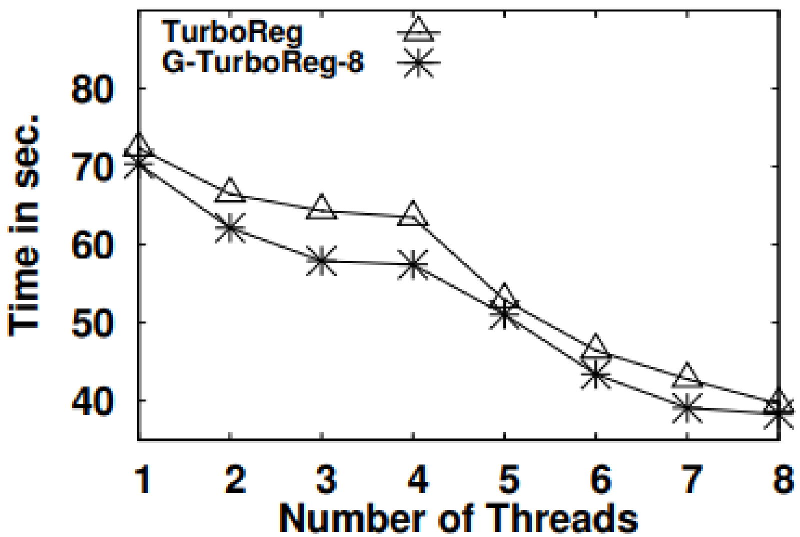
Scalability



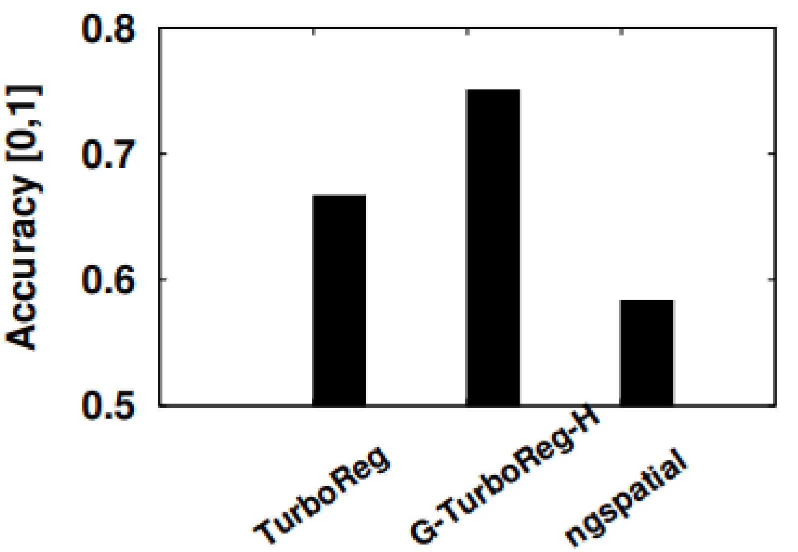
Accuracy

TurboReg Results – Extension (3/3)

- **Effect of number of threads and hybrid neighborhood degree**



Threads Number vs. Scalability



Hybrid Neighborhood Degree vs. Accuracy

RegRocket Results – Extension (1/9)

- **Effect of grid size on accuracy (Table results)**

Grid Size	Metric	ngspatial	*RegRocket*	*RegRocket-4*	*RegRocket-8*
$1k$	Prec.	0.498	0.746	0.872	**0.731**
	Rec.	0.491	0.757	0.837	**0.763**
	F1	0.476	0.653	0.708	**0.683**
$4k$	Prec.	0.667	0.803	0.808	**0.933**
	Rec.	0.601	0.834	0.856	**0.871**
	F1	0.606	0.742	0.704	**0.782**
$15k$	Prec.	0.671	0.804	0.906	**0.962**
	Rec.	0.741	0.832	0.898	**0.903**
	F1	0.635	0.721	0.841	**0.834**
$60k$	Prec.	N/A	0.822	0.913	**0.976**
	Rec.	N/A	0.821	0.919	**0.919**
	F1	N/A	0.678	0.736	**0.798**
$250k$	Prec.	N/A	0.864	0.932	**0.967**
	Rec.	N/A	0.893	0.912	**0.915**
	F1	N/A	0.839	0.781	**0.806**
$1m$	Prec.	N/A	0.878	0.929	**0.961**
	Rec.	N/A	0.908	0.931	**0.895**
	F1	N/A	0.859	0.868	**0.873**

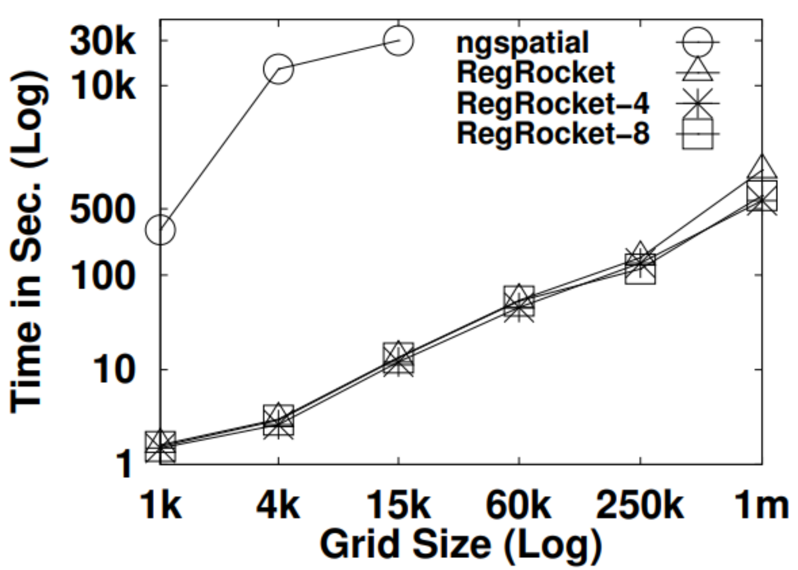
MNLandCover Dataset

Grid Size	Metric	ngspatial	*RegRocket*	*RegRocket-4*	*RegRocket-8*
250	Prec.	0.551	0.846	0.847	**0.858**
	Rec.	0.951	0.966	0.976	**0.985**
	F1	0.698	0.902	0.907	**0.917**
$1k$	Prec.	0.503	0.801	0.876	**0.883**
	Rec.	0.981	0.986	0.965	**0.961**
	F1	0.665	0.884	0.918	**0.921**
$3.5k$	Prec.	0.477	0.865	0.916	**0.901**
	Rec.	0.977	0.991	0.992	**0.985**
	F1	0.641	0.924	0.952	**0.941**
$5k$	Prec.	N/A	0.885	0.875	**0.912**
	Rec.	N/A	0.984	0.986	**0.984**
	F1	N/A	0.932	0.927	**0.947**
$21k$	Prec.	N/A	0.864	0.866	**0.895**
	Rec.	N/A	0.984	0.991	**0.991**
	F1	N/A	0.921	0.924	**0.941**
$84k$	Prec.	N/A	0.889	0.929	**0.919**
	Rec.	N/A	0.991	0.993	**0.991**
	F1	N/A	0.937	0.956	**0.954**

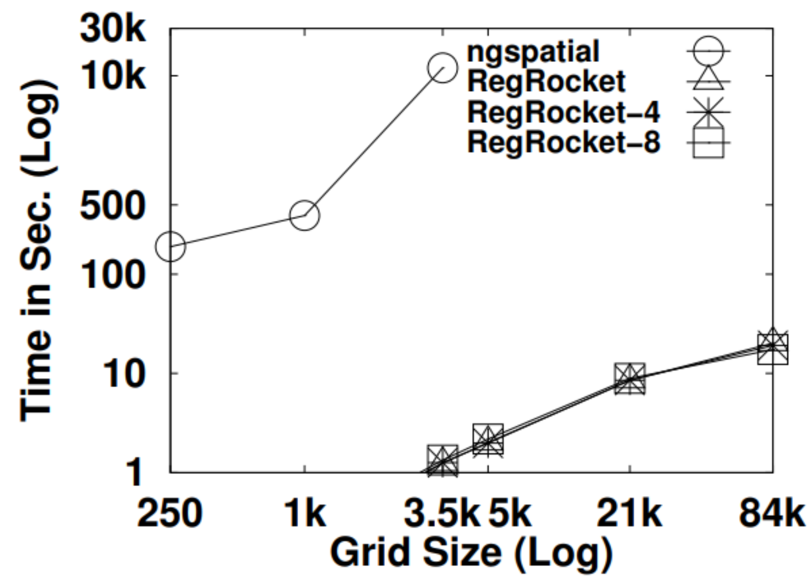
Ebird Dataset

RegRocket Results – Extension (2/9)

- Effect of grid size on scalability



MNLandCover Dataset



Ebird Dataset

RegRocket Results – Extension (3/9)

- **Effect of learning epochs on accuracy**

Num. of Epochs	Metric	*RegRocket*	*RegRocket-4*	*RegRocket-8*
	Prec.	0.815	0.883	0.906
100	Rec.	0.845	0.864	0.854
	F1	0.772	0.732	0.715
	Prec.	**0.864**	**0.932**	**0.967**
1000	Rec.	**0.893**	**0.912**	**0.915**
	F1	**0.839**	**0.781**	**0.806**
	Prec.	0.881	0.931	0.966
10k	Rec.	0.866	0.909	0.915
	F1	0.826	0.785	0.795

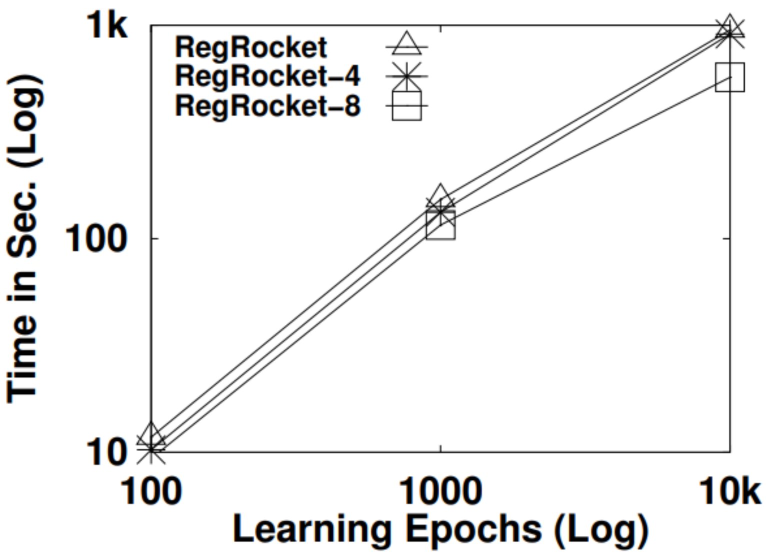
MNLandCover Dataset

Num. of Epochs	Metric	*RegRocket*	*RegRocket-4*	*RegRocket-8*
	Prec.	0.849	0.899	0.909
100	Rec.	0.845	0.835	0.825
	F1	0.847	0.866	0.865
	Prec.	**0.889**	**0.929**	**0.919**
1000	Rec.	**0.991**	**0.993**	**0.991**
	F1	**0.937**	**0.961**	**0.954**
	Prec.	0.909	0.919	0.919
10k	Rec.	0.925	0.935	0.995
	F1	0.917	0.927	0.955

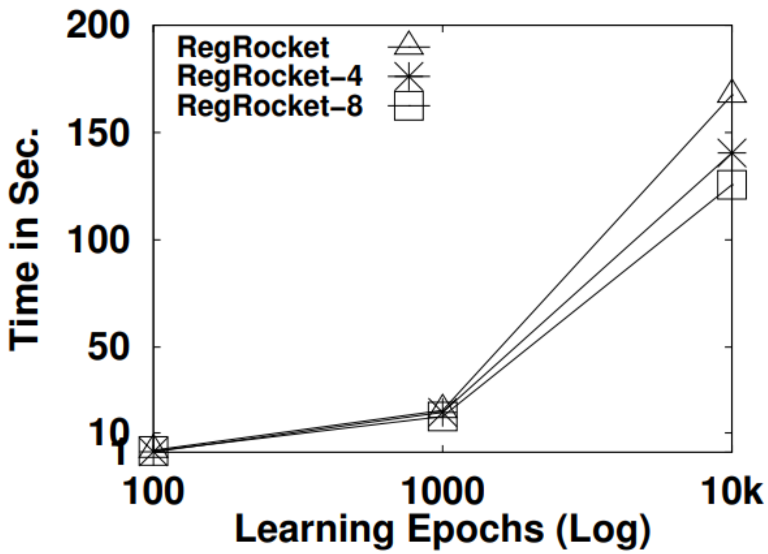
Ebird Dataset

RegRocket Results – Extension (4/9)

- Effect of learning epochs on scalability



MNLandCover Dataset



Ebird Dataset

RegRocket Results – Extension (5/9)

- **Effect of optimization step size on accuracy**

Step Size	Metric	*RegRocket*	*RegRocket-4*	*RegRocket-8*
	Prec.	0.829	0.921	0.966
0.0001	Rec.	0.816	0.789	0.915
	F1	0.782	0.825	0.875
	Prec.	**0.864**	**0.932**	**0.967**
0.001	Rec.	**0.893**	**0.912**	**0.915**
	F1	**0.839**	**0.781**	**0.806**
	Prec.	0.819	0.871	0.926
0.01	Rec.	0.806	0.838	0.875
	F1	0.756	0.745	0.795
	Prec.	0.779	0.861	0.916
0.1	Rec.	0.766	0.828	0.865
	F1	0.676	0.745	0.785

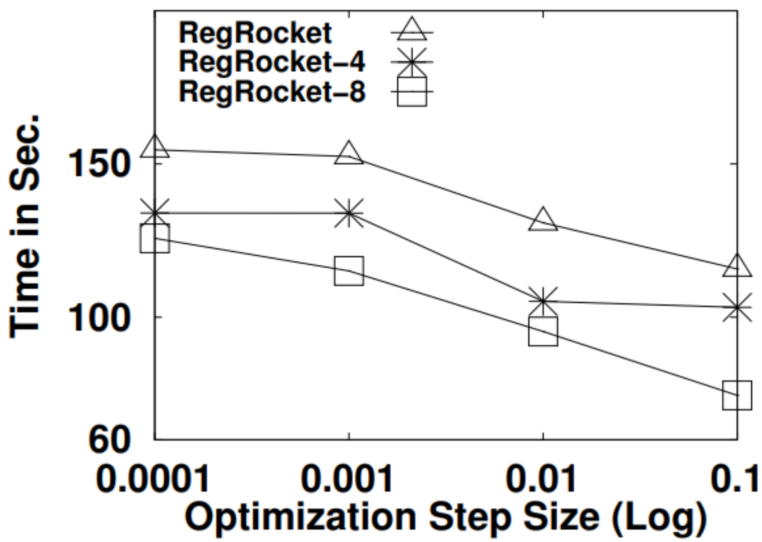
MNLandCover Dataset

Step Size	Metric	*RegRocket*	*RegRocket-4*	*RegRocket-8*
	Prec.	0.914	0.909	0.929
0.0001	Rec.	0.993	0.998	0.995
	F1	0.952	0.951	0.961
	Prec.	**0.889**	**0.929**	**0.919**
0.001	Rec.	**0.991**	**0.993**	**0.991**
	F1	**0.937**	**0.956**	**0.954**
	Prec.	0.879	0.909	0.899
0.01	Rec.	0.985	0.985	0.985
	F1	0.929	0.945	0.941
	Prec.	0.779	0.884	0.879
0.1	Rec.	0.985	0.895	0.895
	F1	0.871	0.889	0.887

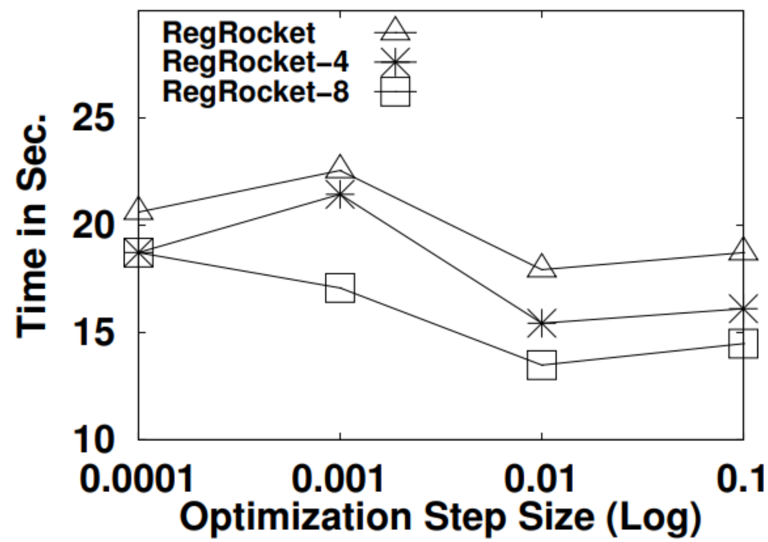
Ebird Dataset

RegRocket Results – Extension (6/9)

- **Effect of optimization step size on scalability**



MNLandCover Dataset



Ebird Dataset

RegRocket Results – Extension (7/9)

- **Effect of factor graph partitions on accuracy**

Num. of Partitions	Metric	*RegRocket*	*RegRocket-4*	*RegRocket-8*
50	Prec.	0.945	0.962	0.971
	Rec.	0.894	0.931	0.912
	F1	0.852	0.877	0.914
100	Prec.	0.913	0.954	0.961
	Rec.	0.891	0.923	0.931
	F1	0.843	0.812	0.861
200	Prec.	**0.864**	**0.932**	**0.967**
	Rec.	**0.893**	**0.912**	**0.915**
	F1	**0.839**	**0.781**	**0.806**
300	Prec.	0.782	0.812	0.815
	Rec.	0.734	0.831	0.821
	F1	0.689	0.701	0.712

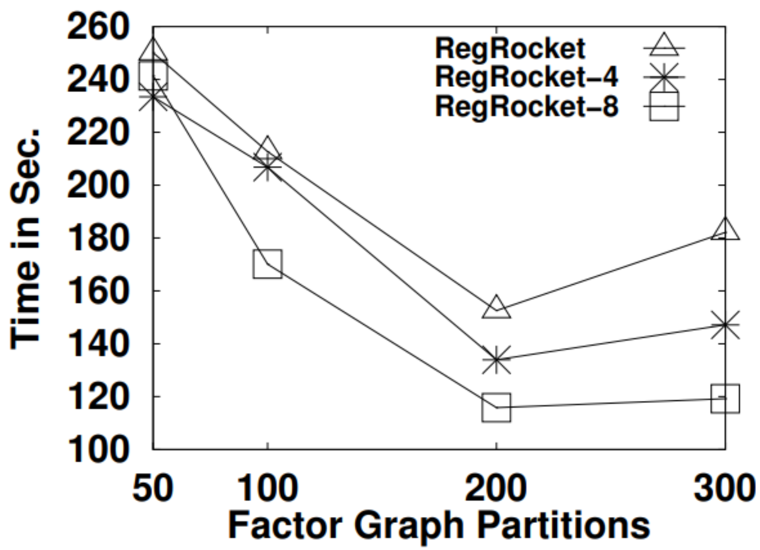
MNLandCover Dataset

Num. of Partitions	Metric	*RegRocket*	*RegRocket-4*	*RegRocket-8*
50	Prec.	0.967	0.944	0.968
	Rec.	0.992	0.991	0.982
	F1	0.979	0.967	0.975
100	Prec.	0.923	0.941	0.937
	Rec.	0.971	0.981	0.983
	F1	0.946	0.961	0.959
200	Prec.	**0.889**	**0.929**	**0.919**
	Rec.	**0.991**	**0.993**	**0.991**
	F1	**0.937**	**0.959**	**0.953**
300	Prec.	0.674	0.789	0.792
	Rec.	0.782	0.712	0.812
	F1	0.724	0.748	0.802

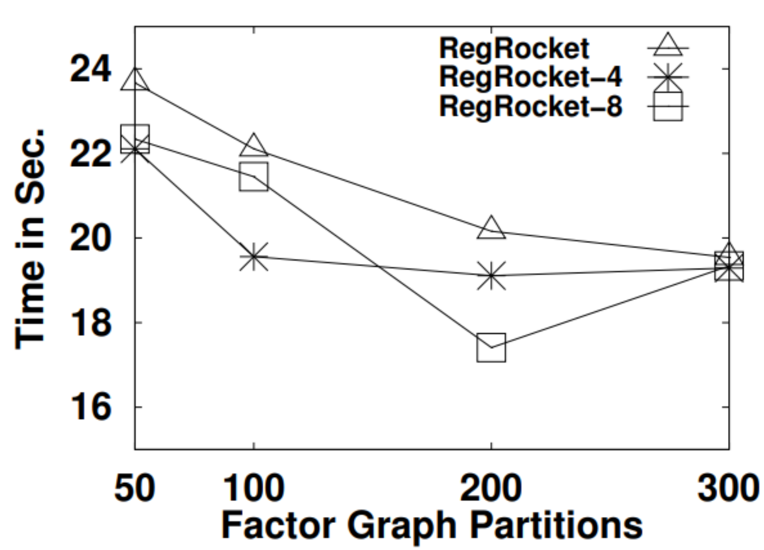
Ebird Dataset

RegRocket Results – Extension (8/9)

- Effect of factor graph partitions on scalability



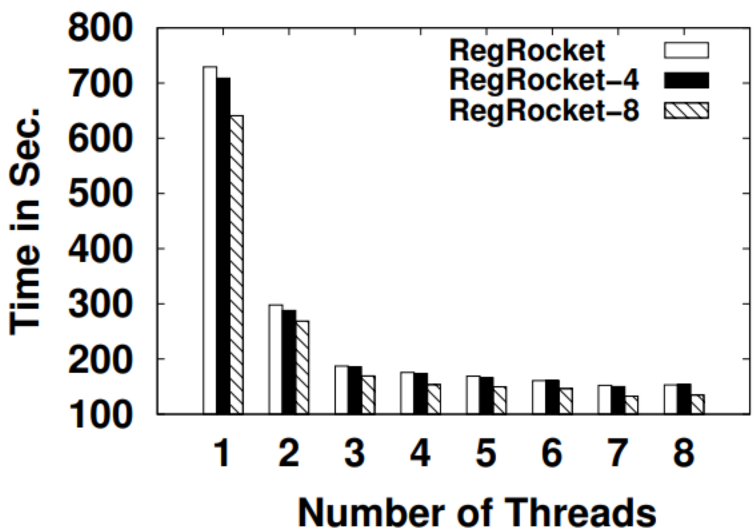
MNLandCover Dataset



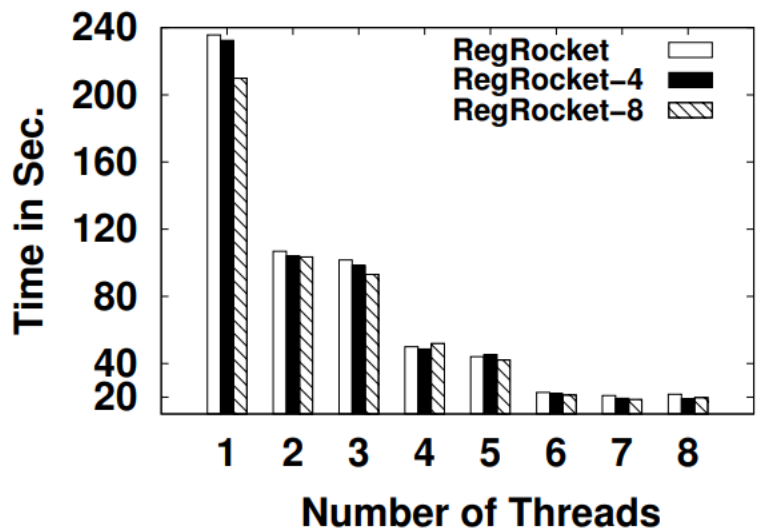
Ebird Dataset

RegRocket Results – Extension (9/9)

- Effect of number of threads on scalability



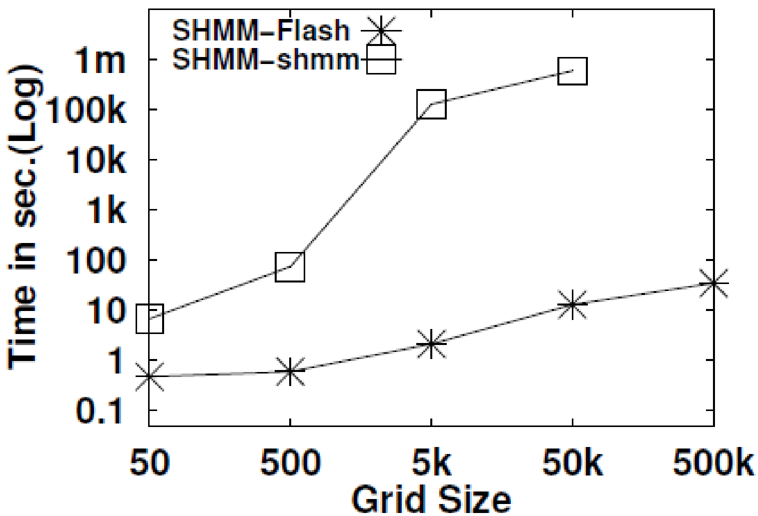
MNLandCover Dataset



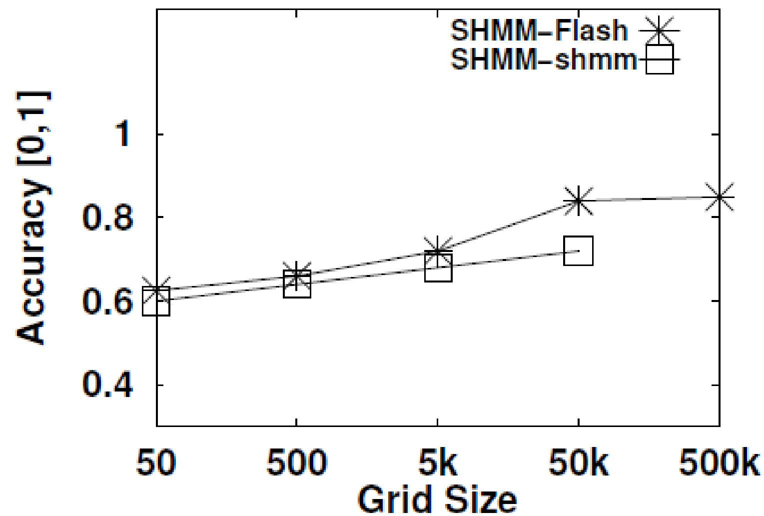
Ebird Dataset

Flash Results – Extension (1/2)

- **SHMM accuracy and scalability**



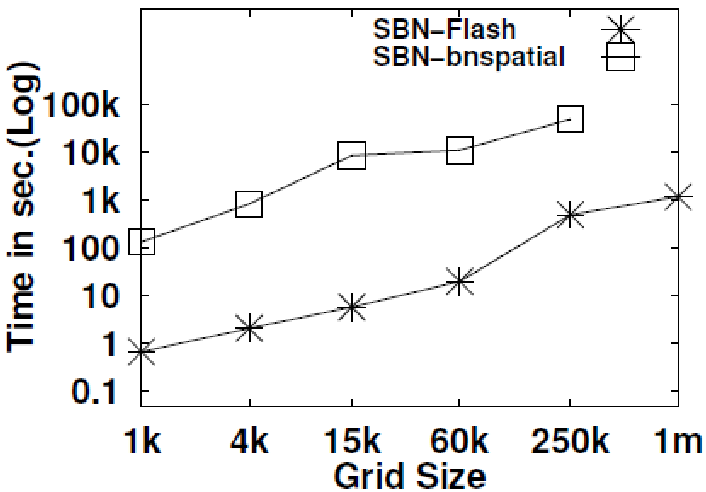
Scalability



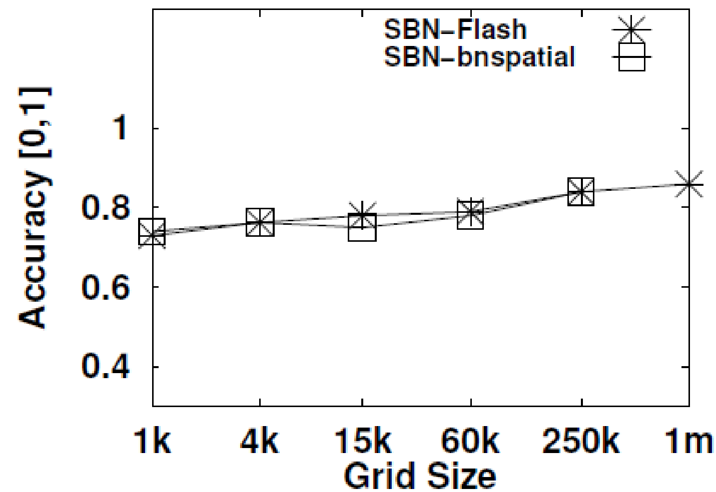
Accuracy

Flash Results – Extension (2/2)

- SBN accuracy and scalability



Scalability



Accuracy