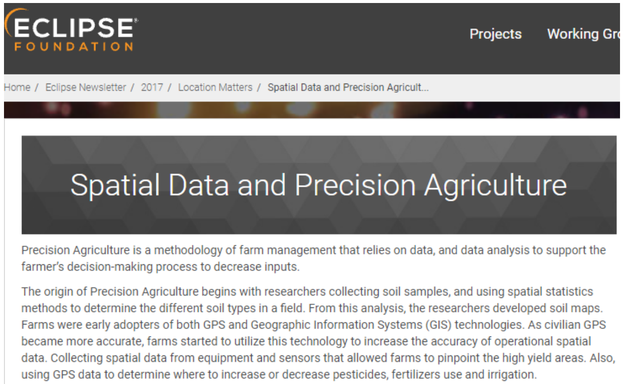
Machine Learning for Big Spatial Data and Applications

Mohamed F. Mokbel
University of Minnesota

The Ubiquity of Big Spatial Data and Applications

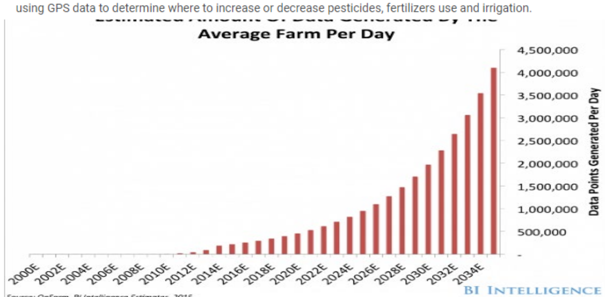
Big Spatial Data in Agriculture


ECLIPSE FOUNDATION
Projects Working Gr
Home / Eclipse Newsletter / 2017 / Location Matters / Spatial Data and Precision Agricult...

Spatial Data and Precision Agriculture

Precision Agriculture is a methodology of farm management that relies on data, and data analysis to support the farmer's decision-making process to decrease inputs.

The origin of Precision Agriculture begins with researchers collecting soil samples, and using spatial statistics methods to determine the different soil types in a field. From this analysis, the researchers developed soil maps. Farms were early adopters of both GPS and Geographic Information Systems (GIS) technologies. As civilian GPS became more accurate, farms started to utilize this technology to increase the accuracy of operational spatial data. Collecting spatial data from equipment and sensors that allowed farms to pinpoint the high yield areas. Also, using GPS data to determine where to increase or decrease pesticides, fertilizers use and irrigation.



Agriculture Technology: How GIS Can Help You Win the Farm

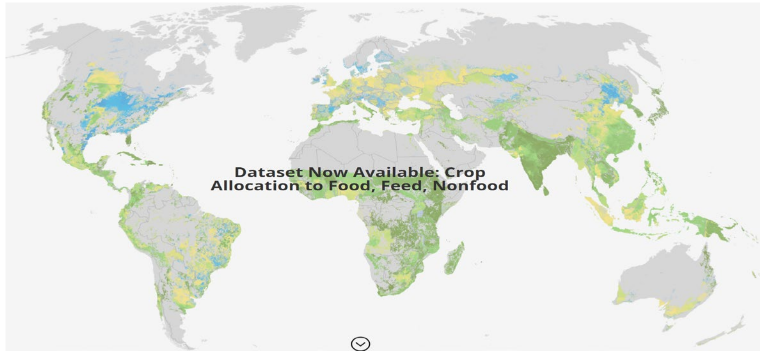
By: GISGeography • Last Updated: August 4, 2021

Agriculture Technology from Location

Today's farmers use sophisticated **agriculture technology** because they can save time and money.

Because crops are location-based, this makes Geographic Information Systems (GIS) an EXTREMELY relevant tool for farmers.

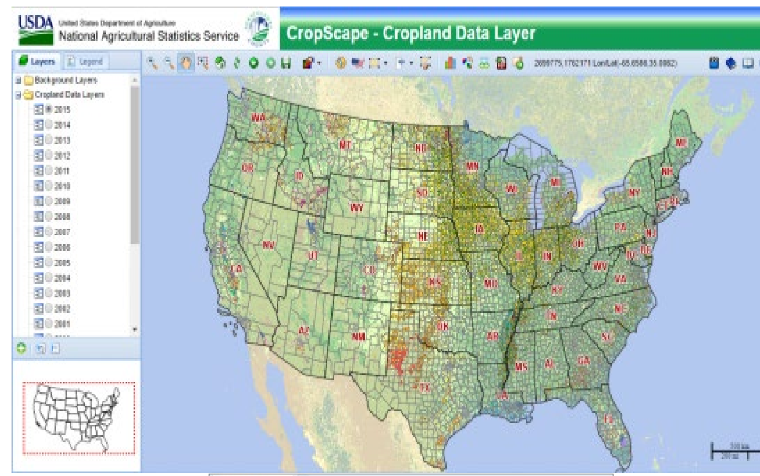
For example, farmers use precision GPS on the field to save fertilizer. Also, satellites and drones collect vegetation, topography, and weather information from the sky.



EARTHSTAT

EarthStat serves geographic data sets that help solve the grand challenge of feeding a growing global population while reducing agriculture's impact on the environment.

CropScape is developed by USDA-NASS where farmers can see *what* crops are growing *where* and *how much*. CropScape is also used for food security, land-cover change and pesticide control: https://nassgeodata.gmu.edu/CropScape/



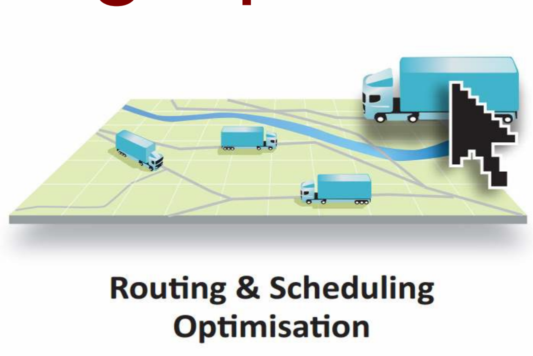
Open Access Feature Paper Article

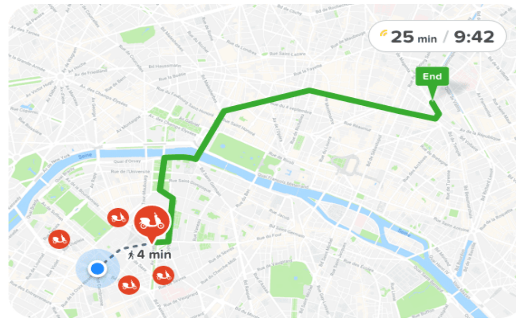
LandQ^{V2}: A MapReduce-Based System for Processing Arable Land Quality Big Data

by Xiaochuang Yao [1,*], Mohamed F. Mokbel [2], Sijing Ye [3], Guoqing Li [1], Louai Alarabi [2], Ahmed Eldawy [4], Zuliang Zhao [5], Long Zhao [5] and Dehai Zhu [5,*]

1 Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
2 Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA
3 State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China
4 Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA
5 College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

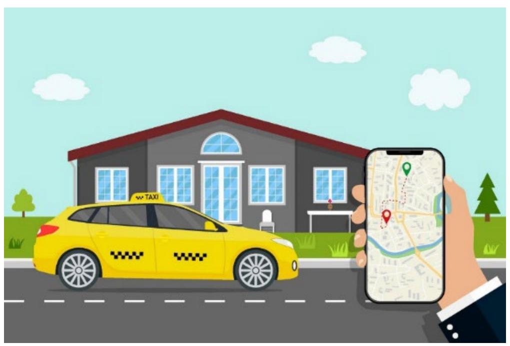
Big Spatial Data in Transportation

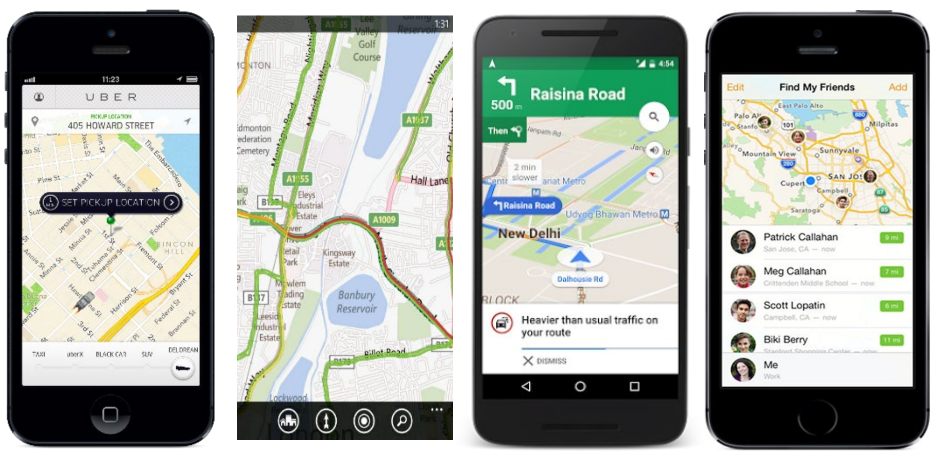




OpenStreetMap

Intelligent Transportation Systems





COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE

Home / Magazine Archive / April 2021 (Vol. 64, No. 4) / Traffic Routing in the Ever-Changing City of Doha / Full Text

ARAB WORLD SPECIAL SECTION: HOT TOPICS

Traffic Routing in the Ever-Changing City of Doha

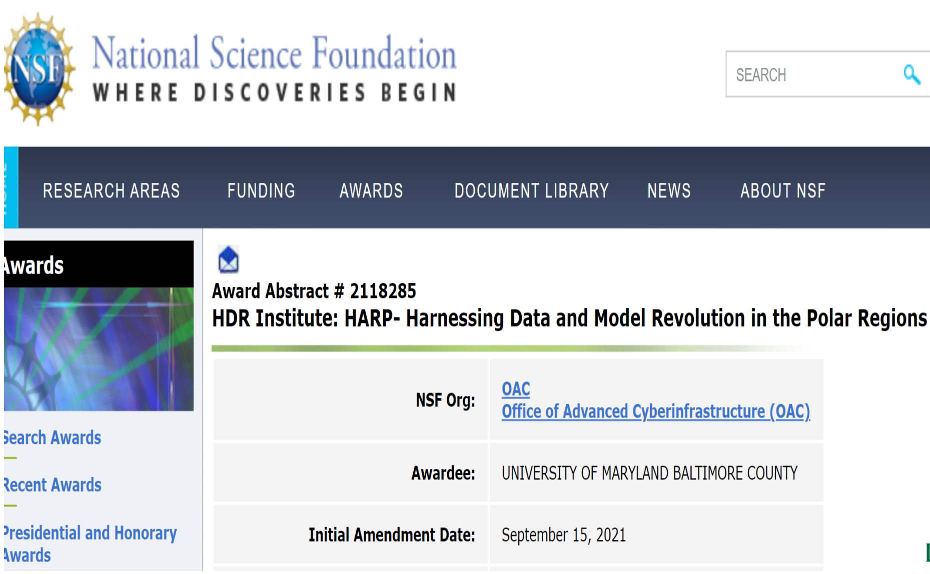
By Sofiane Abbar, Rade Stanojevic, Shadab Mustafa, Mohamed Mokbel
Communications of the ACM, April 2021, Vol. 64 No. 4, Pages 67-68
10.1145/3447731
Comments

VIEW AS: SHARE:

On December 2, 2010, Qatar was announced to host 2022 FIFA World Cup. That was time for celebrating the first-ever Middle Eastern country to organize the tournament. The 1.8M population of Qatar then (2.8M today) never imagined the journey their country was about to embarked. Indeed, in less than 10 years, the population grew by more than a half, pushing the available urban resources and services to their limit. At the same time, the country undertook an ambitious investment plan of \$200B on various infra-structural projects including a brand new three-line metro network, six new stadiums, several new satellite cities, and an astonishing 4,300km of new roads, which tripled the size of the road network in only five years.[3]

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة

Big Spatial Data in Polar Regions



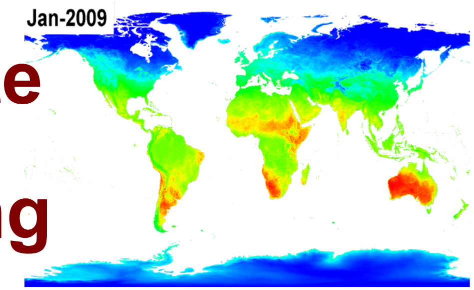
UMBC
TEXAS The University of Texas at Austin
UAF UNIVERSITY OF ALASKA FAIRBANKS
AMHERST COLLEGE
Boulder
BOWIE STATE UNIVERSITY 1865
DARTMOUTH
NASA

The HDR Institute aims to harness massive heterogeneous, noisy, and discontinuous data in space and time and integrate data with numerical and physical models

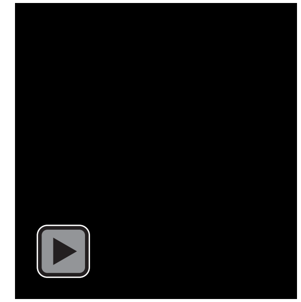
Researchers at i-HARP are investigating novel data science techniques including deep generative adversarial networks, graph neural networks, meta learning, hybrid networks, physics-informed machine learning, causal artificial intelligence, data assimilation, spatio-temporal deep learning, and scalable algorithms.

Big Spatial Data for .....

Remote Sensing



https://lpdaac.usgs.gov
LP DAAC archive exceeds 1PB
5 Trillion points Temperature data
Vegetation data at 250m2 resolution (16 times larger)



GIS INNOVATION CENTER

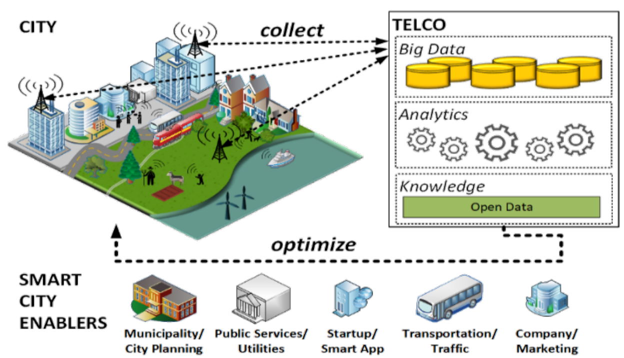
72 months × 14 Billion points/month = **1 Trillion points**

A. Eldawy, M. Mokbel, S. Alharthi, A. Alzaidy, K. Tarek, S. Ghani. "*SHAHED: A MapReduce-based System for Querying and Visualizing Spatio-temporal Satellite Data*". **ICDE 2015**

Telco Data

Telco Big Data Awareness

The aim of this project is to develop next generation telco big data management architectures that can help in understanding urban phenomena (e.g., traffic in a city, mobility patterns for emergency response or city planning, improve the Quality of Service) at a very high spatio-temporal resolution. The project deals with algorithms and structures to ingest in the most compact manner huge amounts of network logs perform big data exploration and analytics within a tolerable elapsed time.

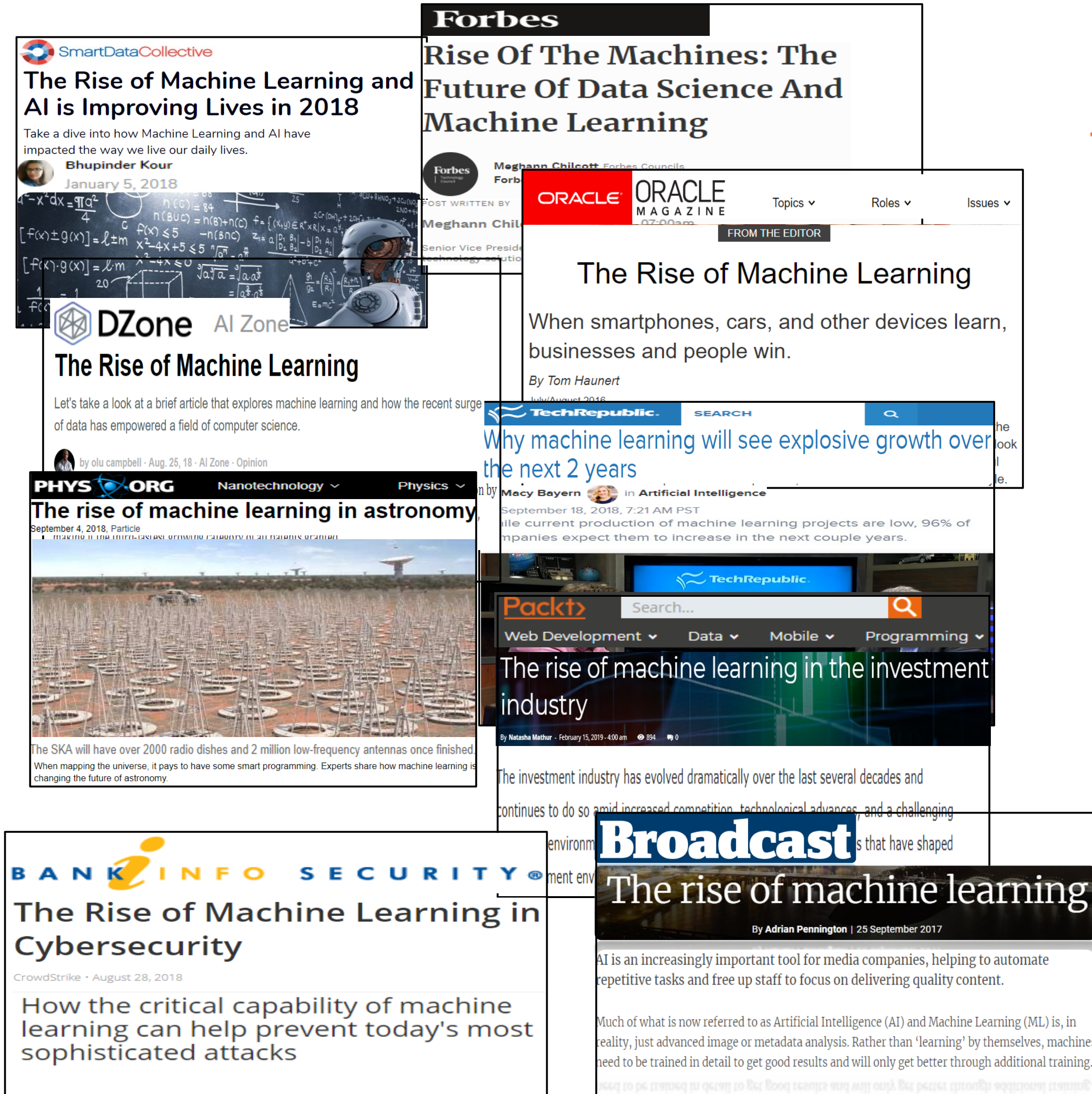


C. Costa, G. Chatzimilioudis, D. Zeinalipour-Yazti, M. Mokbel: Efficient Exploration of Telco Big Data with Compression and Decaying. ICDE 2017: 1332-1343

University of Cyprus

. . . . . .

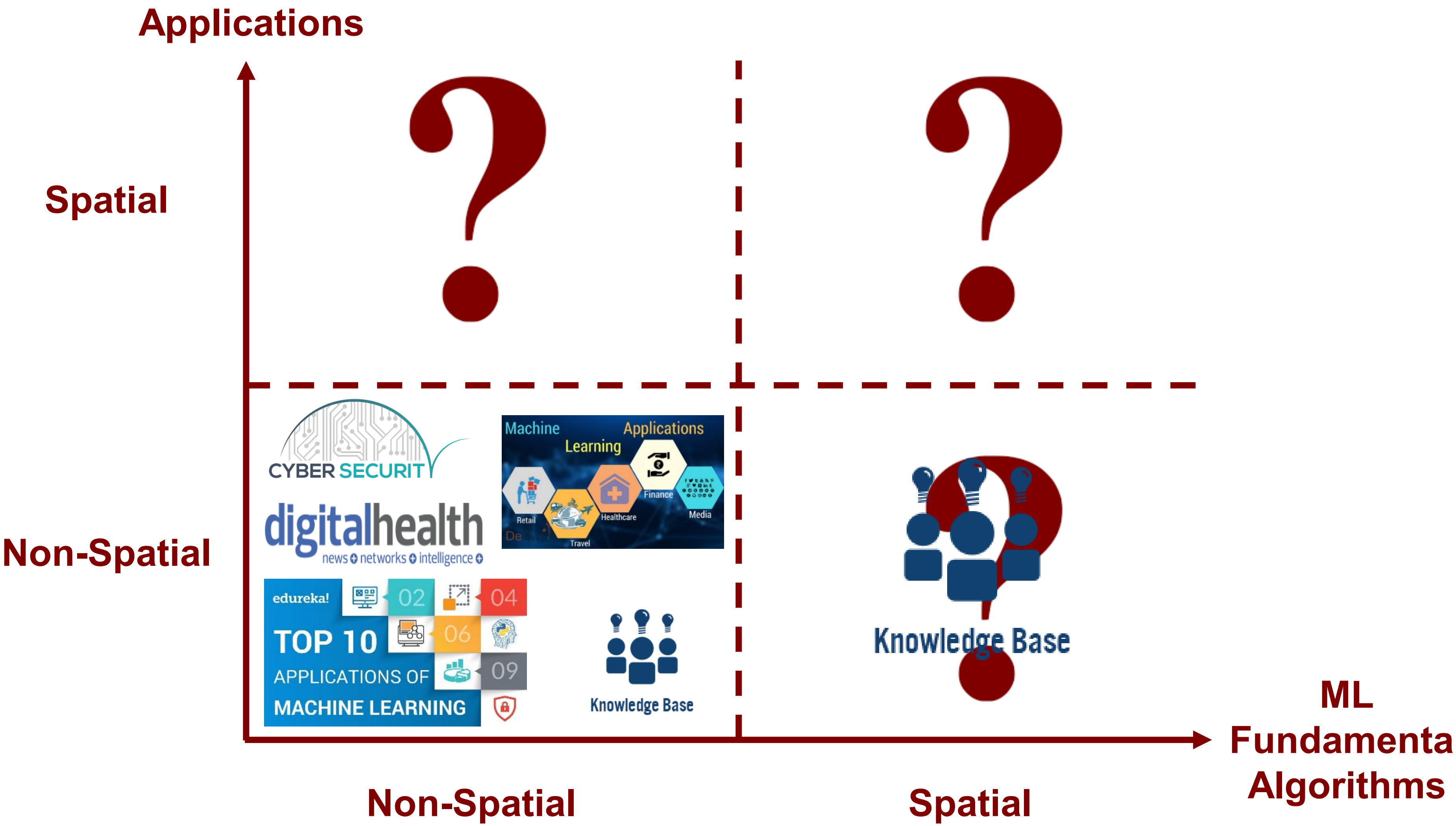
Meanwhile, the Rise of Machine Learning



TensorFlow PYTORCH mxnet Chainer Spark Caffe MLlib theano Microsoft Cognitive Toolkit Caffe2

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

Machine Learning meets Big 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
Barack	Michelle

NELL

yago select knowledge

SystemT

StatSnowBall

DeepDive

lixto DELIVERING COMPETITIVE ADVANTAGE

DBLife

Google Vault

ProbKB

DARPA MEMEX

Fight Human Trafficking

Crime Investigation

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)

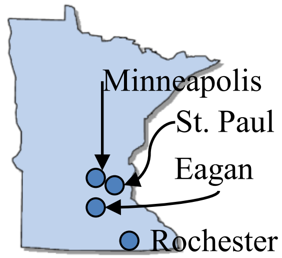
DeepDive with Spatial Data …

Crime rates in Minnesota

City	C	E
Minneapolis	1	0.7
St. Paul	?	0.7
Eagan	?	0.7
Rochester	?	0.7

Crimes

Education



Data

```
P1: City X has high crime rate
P2: Cities X&Y have same education level

Rule: P1&P2 → Y has high crime rate
```

Inference Rules

DeepDive

City	Confidence		
St. Paul	0.5		
Eagan	0.5		
Rochester	0.5		

Result

DeepDive with Spatial Data …

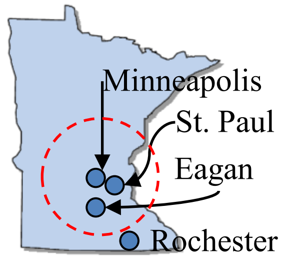
Crime rates in Minnesota

Data

City	C	E
Minneapolis	1	0.7
St. Paul	?	0.7
Eagan	?	0.7
Rochester	?	0.7

Crimes

Education



Inference Rules

```
P1: City X has high crime rate
P2: Cities X&Y have same education level
P3: Cities X&Y are within 80 miles
```

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

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

DeepDive

Result

City	Confidence		
St. Paul	~~0.5~~	0.7	
Eagan	~~0.5~~	0.7	
Rochester	~~0.5~~	0	

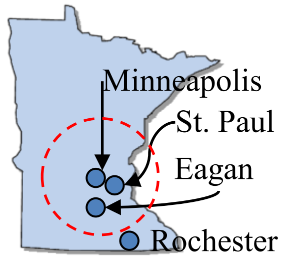
DeepDive with Spatial Data …

Crime rates in Minnesota

City	C	E
Minneapolis	1	0.7
St. Paul	?	0.7
Eagan	?	0.7
Rochester	?	0.7

Crimes

Education



Ebola infection rates in Liberia

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

Infections

Sanitation

Data

```
P1: City X has high crime rate
P2: Cities X&Y have same education level
P3: Cities X&Y are within 80 miles
```

~~**Rule:** P1&P2 → Y has high crime rate~~

Rule: P1&P2&P3 → Y has high crime rate

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

DeepDive

DeepDive

Result

City	Confidence		
St. Paul	~~0.5~~	0.7	
Eagan	~~0.5~~	0.7	
Rochester	~~0.5~~	0	

City	Confidence		
Margibi	0.54		
Bong	0.52		
Gbarpolu	0.63		

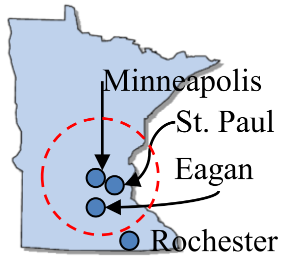
DeepDive with Spatial Data …

Crime rates in Minnesota

City	C	E
Minneapolis	1	0.7
St. Paul	?	0.7
Eagan	?	0.7
Rochester	?	0.7

Crimes

Education



Ebola infection rates in Liberia

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

Infections

Sanitation

Data

```
P1: City X has high crime rate
P2: Cities X&Y have same education level
P3: Cities X&Y are within 80 miles
```

~~**Rule:** P1&P2 → Y has high crime rate~~

Rule: P1&P2&P3 → Y has high crime rate

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

DeepDive

Result

City	Confidence		
St. Paul	~~0.5~~	0.7	
Eagan	~~0.5~~	0.7	
Rochester	~~0.5~~	0	

City	Confidence		
Margibi	~~0.54~~	0.51	
Bong	~~0.52~~	0.45	
Gbarpolu	~~0.63~~	0.06	

DeepDive with Spatial Data …

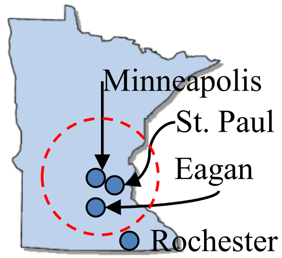
Crime rates in Minnesota

Data

City	C	E
Minneapolis	1	0.7
St. Paul	?	0.7
Eagan	?	0.7
Rochester	?	0.7

Crimes

Education



Inference Rules

```
P1: City X has high crime rate
P2: Cities X&Y have same education level
~~P3: Cities X&Y are within 80 miles~~
P3: The closer Y&X the higher Y crime rate

~~Rule: P1&P2 → Y has high crime rate~~
Rule: P1&P2&P3 → Y has high crime rate
```

DeepDive

Result

City	Confidence		
St. Paul	~~0.5~~	~~0.7~~	
Eagan	~~0.5~~	~~0.7~~	
Rochester	~~0.5~~	~~0~~	

Ebola infection rates in Liberia

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

Infections

Sanitation


```
P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
~~P3: Counties X&Y are within 150 miles~~
P3: The closer Y&X the higher Y infect rate

~~Rule: P1&P2 → Y has high infection rate~~
Rule: P1&P2&P3 → Y has high infection rate
```

DeepDive

City	Confidence		
Margibi	~~0.54~~	~~0.51~~	
Bong	~~0.52~~	~~0.45~~	
Gbarpolu	~~0.63~~	~~0.06~~	

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`

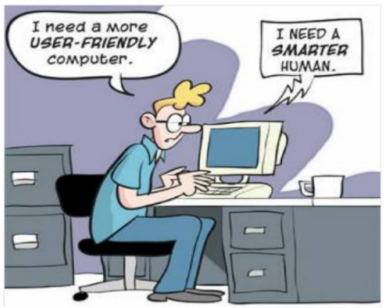
Need to build ***Spatial Markov Logic Networks (SMLN)***, a full-fledged MLN framework with a native support for spatial data and applications

Markov Logic Networks (MLN)

Towards Data Science | HOME | DATA SCIENCE | MACHINE LEARNING | PROGRAMMING | VISUALIZATION | AI | PICKS | CONTRIBUTE

Making Deep Learning User-Friendly, Possible?

Eric Feuilleaubois · Apr 4 · 12 min read



CONTEXT | Business R... 2018): Desp... machine le... discoveries... academics... are struggl... machine learning to solve real business problems. In short, the gap for most companies isn't that machine learning doesn't work, but that they struggle to actually use it.

Machine learning — a form of artificial intelligence that uses algorithms and large data sets to derive insights in real time — is way more than hype.

Gartner predicts that by 2018, 45 percent of the fastest-growing companies will have fewer employees than instances of smart machines.

It's clear that machine learning offers companies a competitive advantage, but is it something that small- and medium-sized business can adopt? The algorithms churning the data are often opaque, and things can go wrong, from the humorous (automated email replies that write "I love you" to a

Forbes | Billionaires | Innovation | Leadership | Money

53,950 views | Jan 1, 2018, 08:33pm

Why Do Developers Find It Hard ...ine Learning?

...earning, and Internet of Things

... the most critical skill of current times. ...cation of ML, is becoming pervasive. From autonomous vehicles to self-tuned databases, AI and ML are found everywhere. Industry analysts often refer to AI-driven automation as the job killer. Almost every domain and industry vertical are getting impacted by AI and ML. Platform companies with massive investments in AI research are shipping new tools and frameworks at a rapid pace.

Need experts and highly-trained scientists, specially for deep learning

- **MLN is an end-to-end ML solution**
  - Covers wide range of ML problems
  - Thousands of lines of ML code can be done in few MLN formulas

Rules as MLN formulas → Markov Logic Network (MLN) → Rule weights

ACM SIGMOD Houston, Texas, USA 2018

Alchemy - Open Source AI

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

Tuffy

Scalable RDBMS-based MLN System

DARPA MEMEX

DeepDive

datanami — DATA SCIENCE • AI • ADVANCED ANALYTICS

July 3, 2018

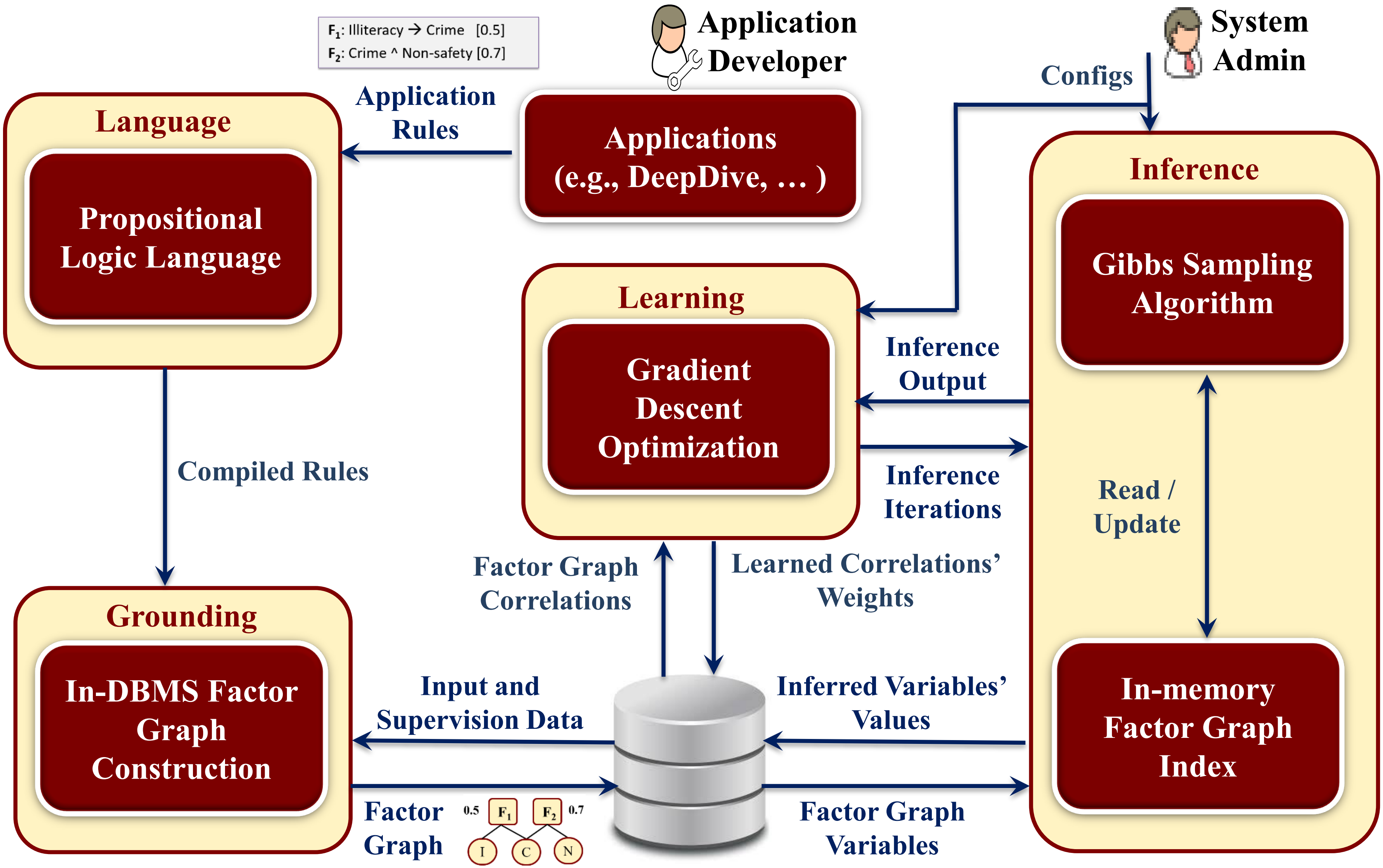
Can Markov Logic Take Machine Learning to the Next Level?

Alex Woodie

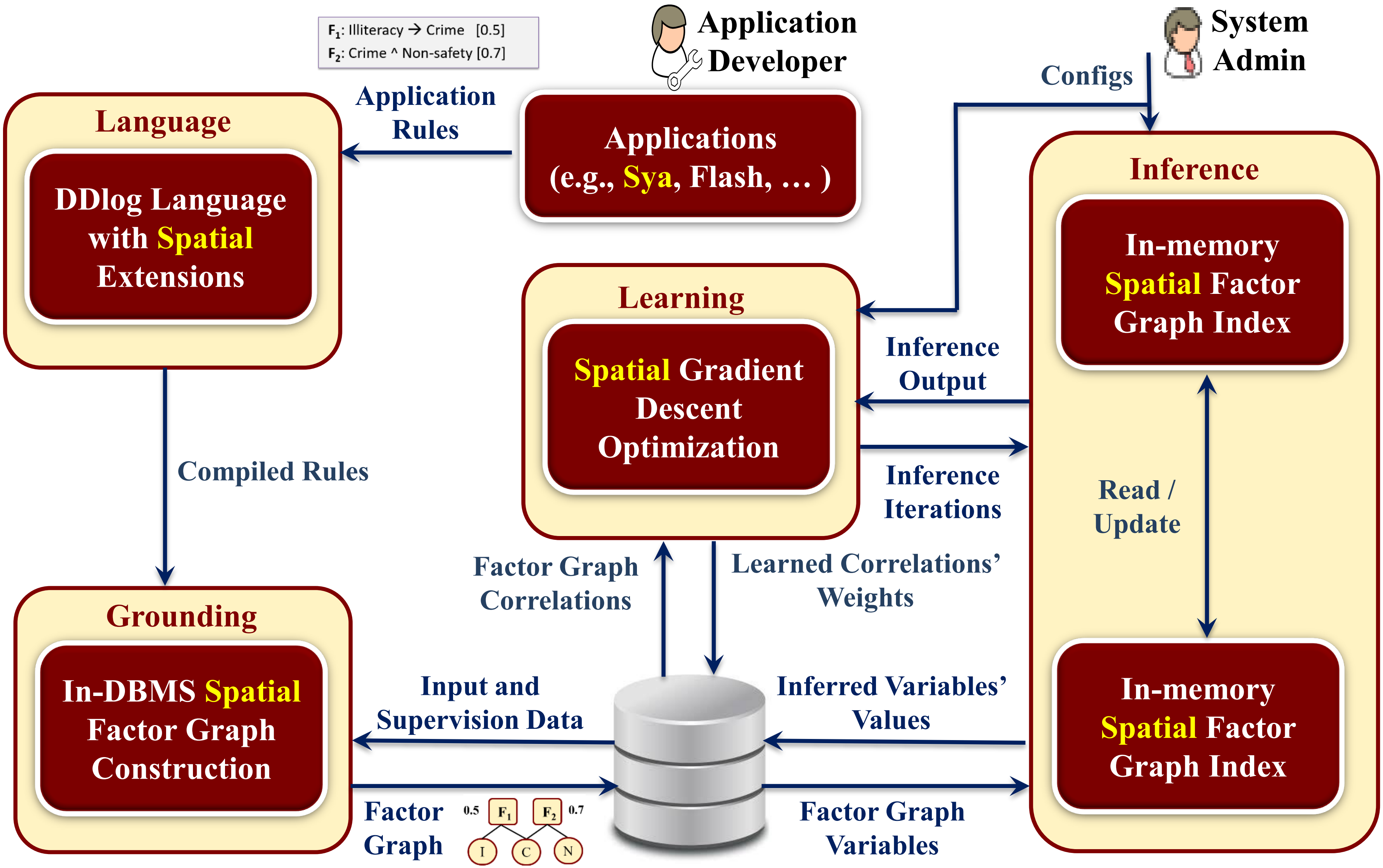
Advances in machine learning, including deep learning, have propelled artificial intelligence (AI) into the public conscience and forced executives to create new business plans based on data. However, the

HoloClean

MLN Architecture



Spatial MLN Architecture



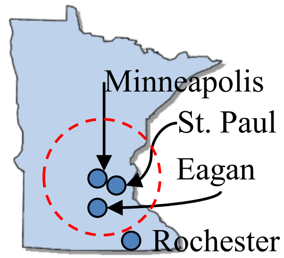
Sya

Crime rates in Minnesota

City	C	R
Minneapolis	1	0.7
St. Paul	?	0.7
Eagan	?	0.7
Rochester	?	0.7

Crimes

Education



Ebola infection rates in Liberia

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

Infections

Sanitation

Data

Inference Rules

```
P1: City X has high crime rate
P2: Cities X&Y have same education level
~~P3: Cities X&Y are within 80 miles~~
P3: The closer Y&X the higher Y crime rate

~~Rule: P1&P2 → Y has high crime rate~~
Rule: P1&P2&P3 → Y has high crime rate
```

```
P1: County X has high Ebola infection rate
P2: Counties X&Y have same sanitation level
~~P3: Counties X&Y are within 150 miles~~
P3: The closer Y&X the higher Y infect rate

~~Rule: P1&P2 → Y has high infection rate~~
Rule: P1&P2&P3 → Y has high infection rate
```

Sya

Sya

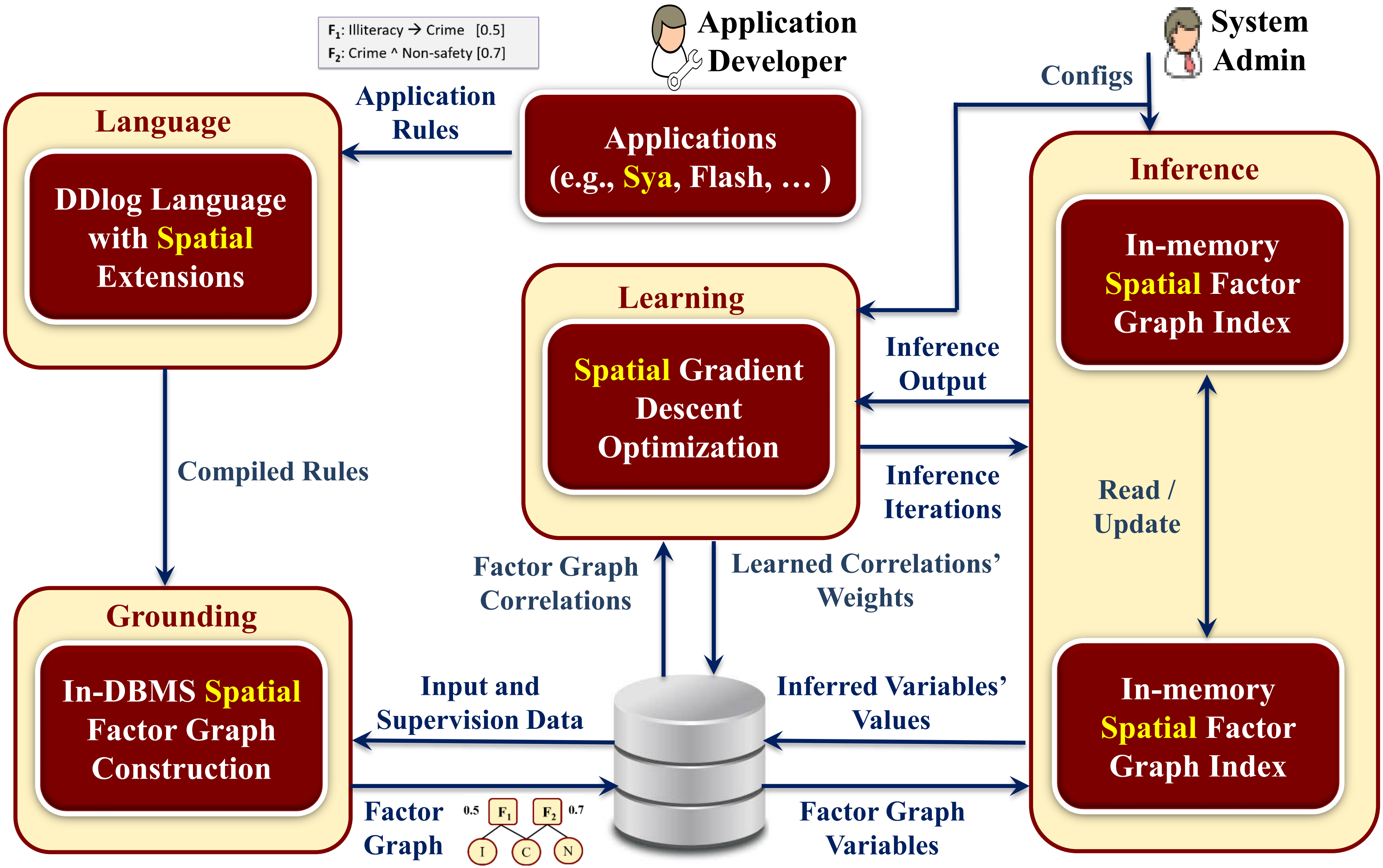
Result

City	Confidence		
St. Paul	~~0.5~~	~~0.7~~	0.9
Eagan	~~0.5~~	~~0.7~~	0.7
Rochester	~~0.5~~	~~0~~	0.3

City	Confidence		
Margibi	~~0.54~~	~~0.51~~	0.76
Bong	~~0.52~~	~~0.45~~	0.53
Gbarpolu	~~0.63~~	~~0.06~~	0.22

Machine Learning meets Big Spatial Data

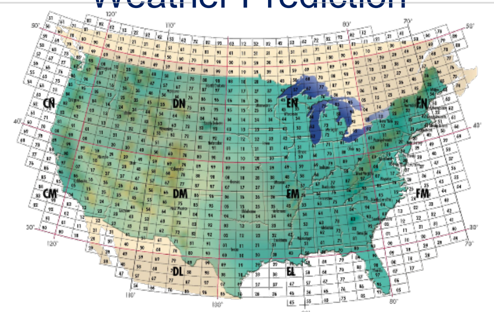
Spatial MLN Architecture



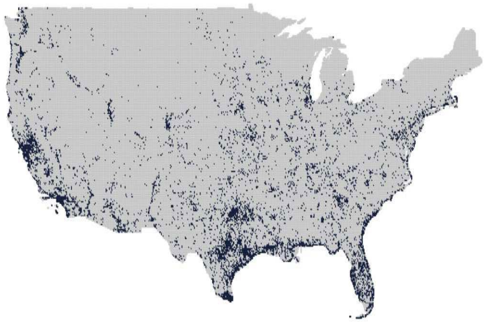
Spatial (Autologisitc) Regression

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

Weather Prediction



Birds Migration



Features

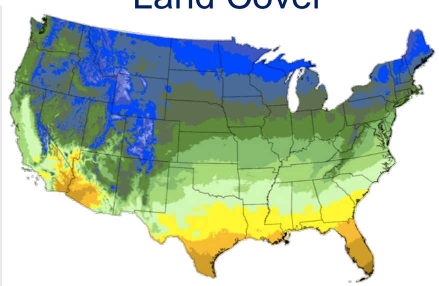
$(x_1 = 0, x_2 = 1)$ $(x_1 = 1, x_2 = 1)$

l_1 0	l_2 1	l_3 1	l_4 1
l_5 1	l_6 1	l_7 0	l_8 1
l_9 1	l_{10} 1	l_{11} 0	l_{12} 0
l_{13} 1	l_{14} ?	l_{15} 0	l_{16} 0

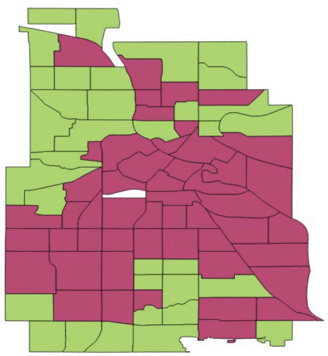
$(x_1 = 1, x_2 = 0)$

Phenomenon value

Land Cover



Crimes Distribution



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

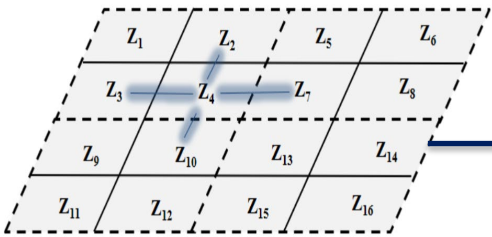
Learning regression parameters for 80K cells takes more than one day ☹

Spatial Regression as SMLN Problem

Rules as MLN formulas → Markov Logic Network (MLN) → Rule weights

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

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



→ SMLN Transformation → SMLN Rules → SMLN Engine → ß1 , η

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

$$\left\{ \begin{array}{l} \boxed{\log \frac{Pr(z_1=1)}{Pr(z_1=0)} = \beta_1 x_1 + \eta z_2 + \eta z_3} \\ \vdots \\ \log \frac{Pr(z_4=1)}{Pr(z_4=0)} = \beta_1 x_1 + \eta z_2 + \eta z_3 + \eta z_7 + \eta z_{10} \\ \vdots \\ \log \frac{Pr(z_{16}=1)}{Pr(z_{16}=0)} = \beta_1 x_1 + \eta z_{14} + \eta z_{15} \end{array} \right.$$

SMLN Rules
[Z_1 ^ X_1, ß1]
[Z_1 ^ Z_2, η]
[Z_1 ^ Z_3, η]
[Z_2 ^ X_1, ß1]
[Z_2 ^ Z_4, η]
[Z_2 ^ Z_5, η]
[Z_3 ^ X_1, ß1]
[Z_3 ^ Z_4, η]
....

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

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

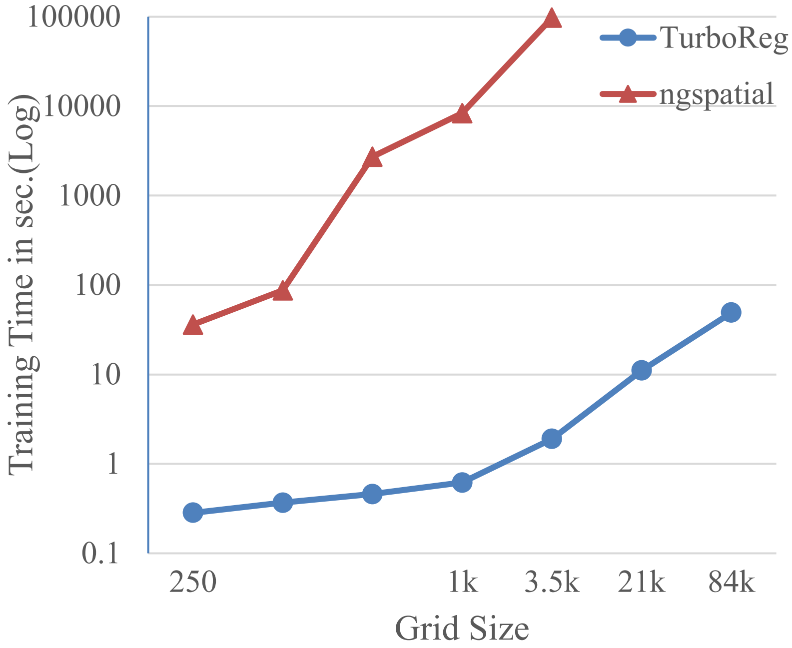
$$\text{Prediction} \quad \{0,1,2\} \; z_i \equiv \begin{matrix} \boxed{z_i(0)} \\ z_i(1) \\ z_i(2) \end{matrix} \; \{0,1\}$$

$$\text{Feature} \quad \{0,1,2\} \; x_j \equiv \begin{matrix} x_j^{0,0} & x_j^{0,1} & x_j^{0,2} \\ x_j^{1,0} & x_j^{1,1} & x_j^{1,2} \\ x_j^{2,0} & x_j^{2,1} & x_j^{2,2} \end{matrix} \; \{0,1\}$$

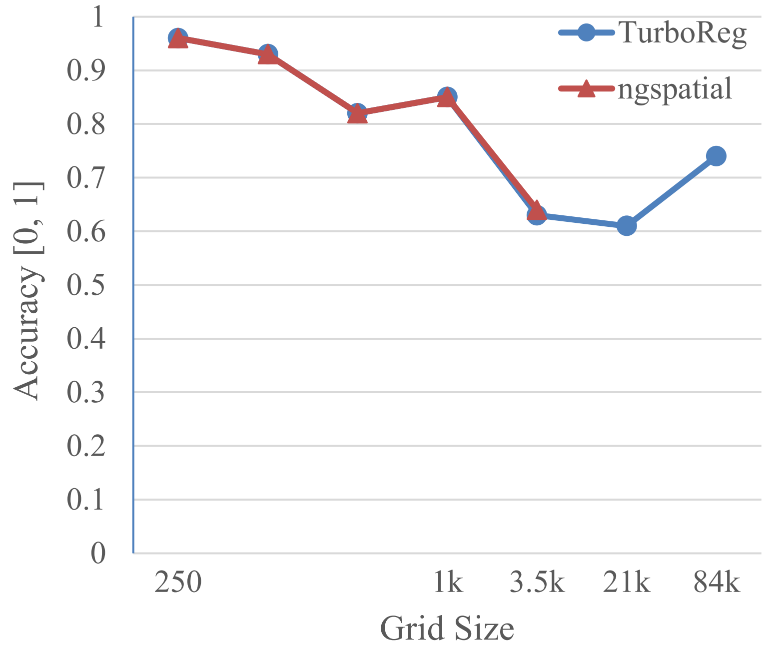
$$\begin{cases} \log \frac{Pr(z_i(1)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i(0)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})} = \sum_{j=1}^{m} \sum_{t \in \mathcal{D}_{x_j}} \beta_j^{1,t} x_j^{1,t} + \sum_{k \in \mathcal{N}_i} \sum_{s \in \mathcal{D}_{z_k}} \eta_{1,s} z_k(s) \\ \log \frac{Pr(z_i(2)=1|\mathcal{X}(i),\mathcal{Z}_{\mathcal{N}_i})}{Pr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

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

Scalability



Accuracy

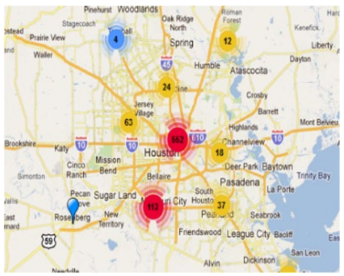


At **least three orders of magnitude** performance gain, while accuracy is **almost the same**.

Spatial Probabilistic Graphical Modeling (SPGM)

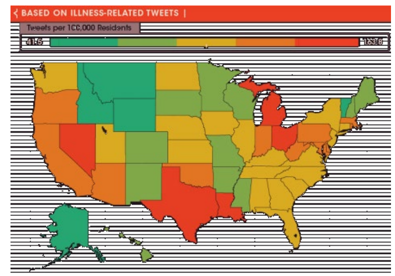
- Performing *uncertain* (i.e., prob.) predictions over spatial data
  - Classical ML approaches (e.g., regression) ignore the probabilistic relationships

Disaster Analysis



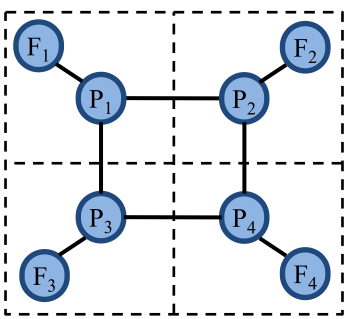
Crime Analysis

Public Health Monitoring

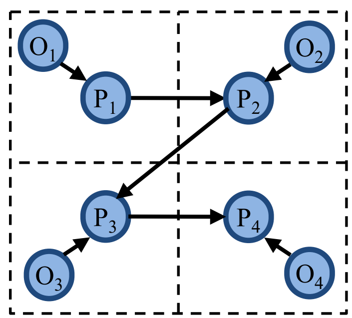


Geo-tagged Ads

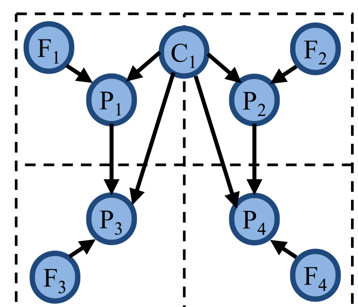
- Representing the world as a collection of *random variables* with joint probabilistic distribution
  - Tasks: learning the distribution, and inferring unknown variables via the distribution



Spatial Markov Random Field (SMRF)



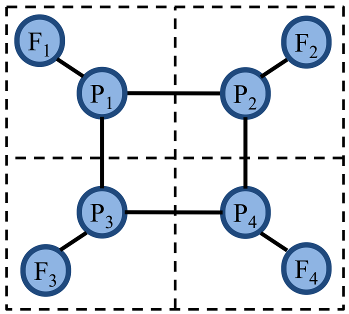
Spatial Hidden Markov Model (SHMM)



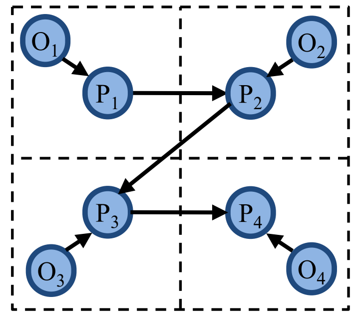
Spatial Bayesian Network (SBN)

SMLN for SPGM

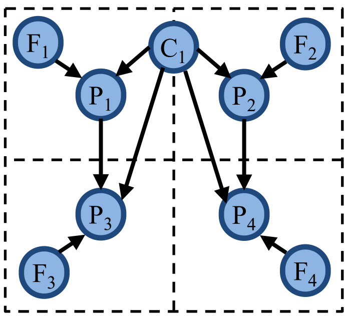
- Generates an equivalent set of weighted rules containing logical predicates for any SPGM input
  - Weights represent the original SPGM parameters
  - Rules follow the syntax of the DDlog logic programming framework



Spatial Markov Random Field (SMRF)



Spatial Hidden Markov Model (SHMM)



Spatial Bayesian Network (SBN)

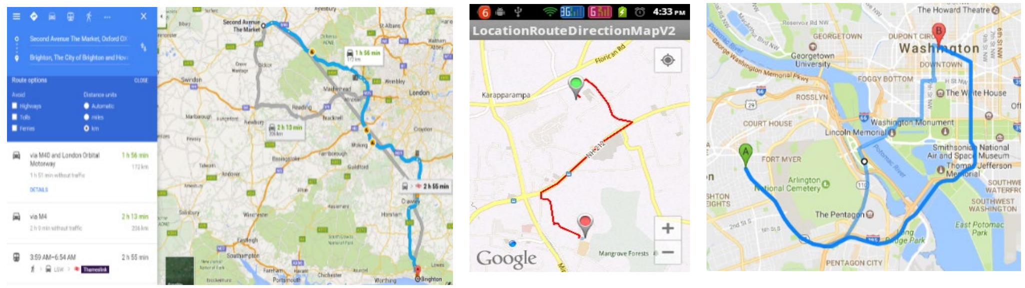
MLN Rules
$[P_1 \wedge F_1, \beta_1]$
$[P_1 \wedge P_2, \eta]$
$[P_1 \wedge P_3, \eta]$
$[P_2 \wedge F_2, \beta_1]$
$[P_2 \wedge P_4, \eta]$
…….

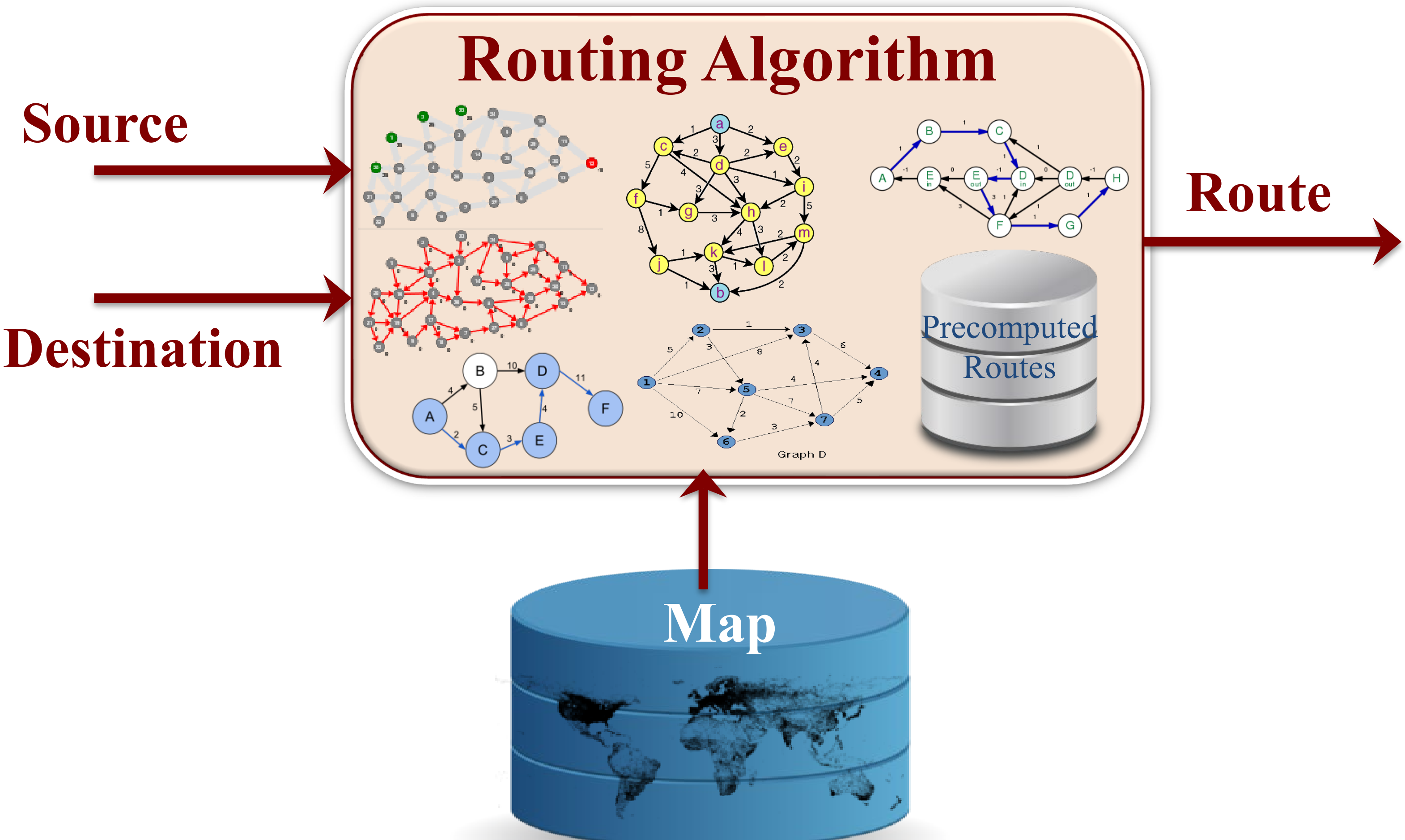
MLN Rules
$[O_1 \rightarrow P_1, b]$
$[P_1 \rightarrow P_2, a]$
$[O_2 \rightarrow P_2, b]$
$[P_2 \rightarrow P_3, a]$
$[O_3 \rightarrow P_3, b]$
…….

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

Machine Learning meets Big Spatial Data

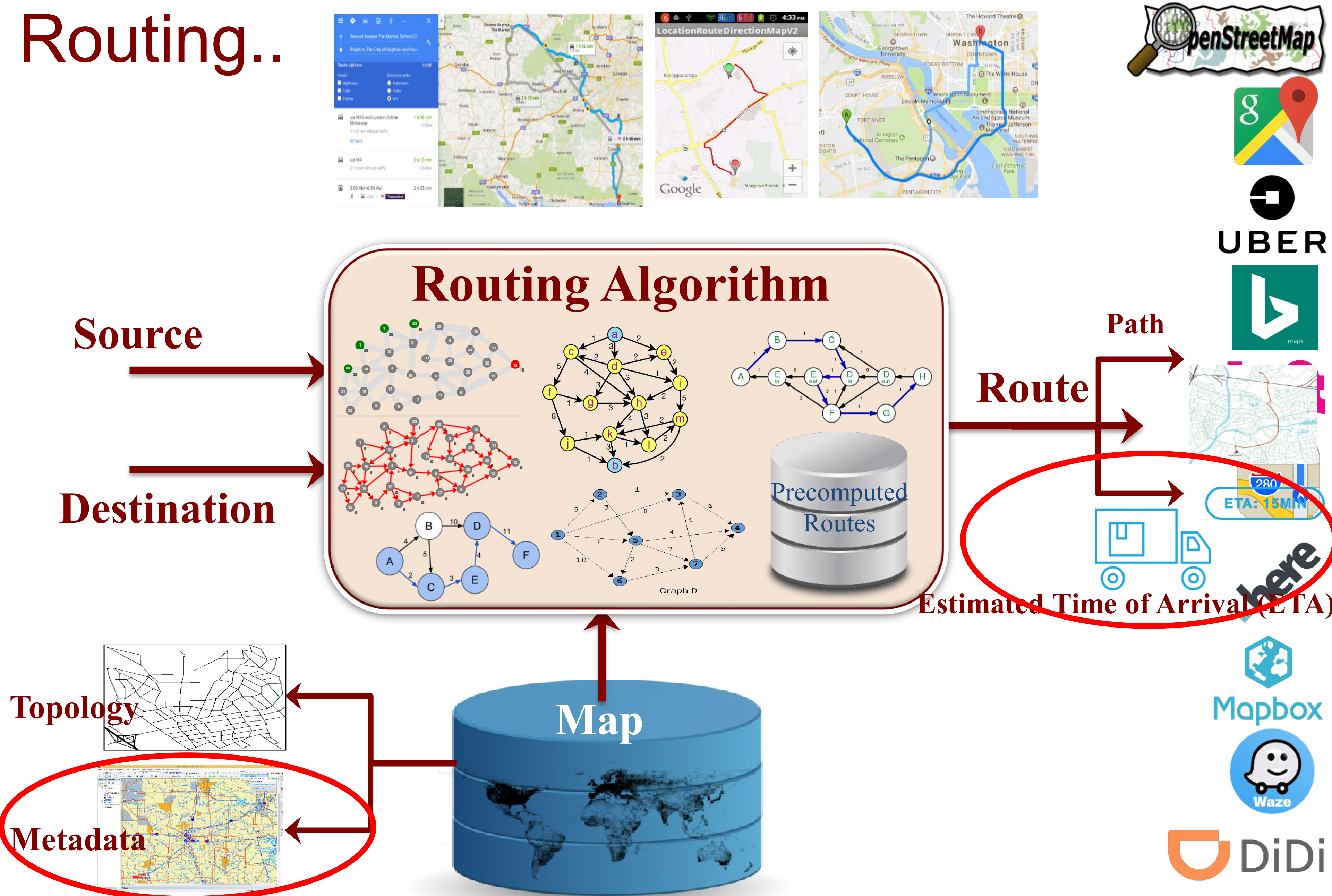
Routing..





Source
Destination
Routing Algorithm
Precomputed Routes
Route
Map

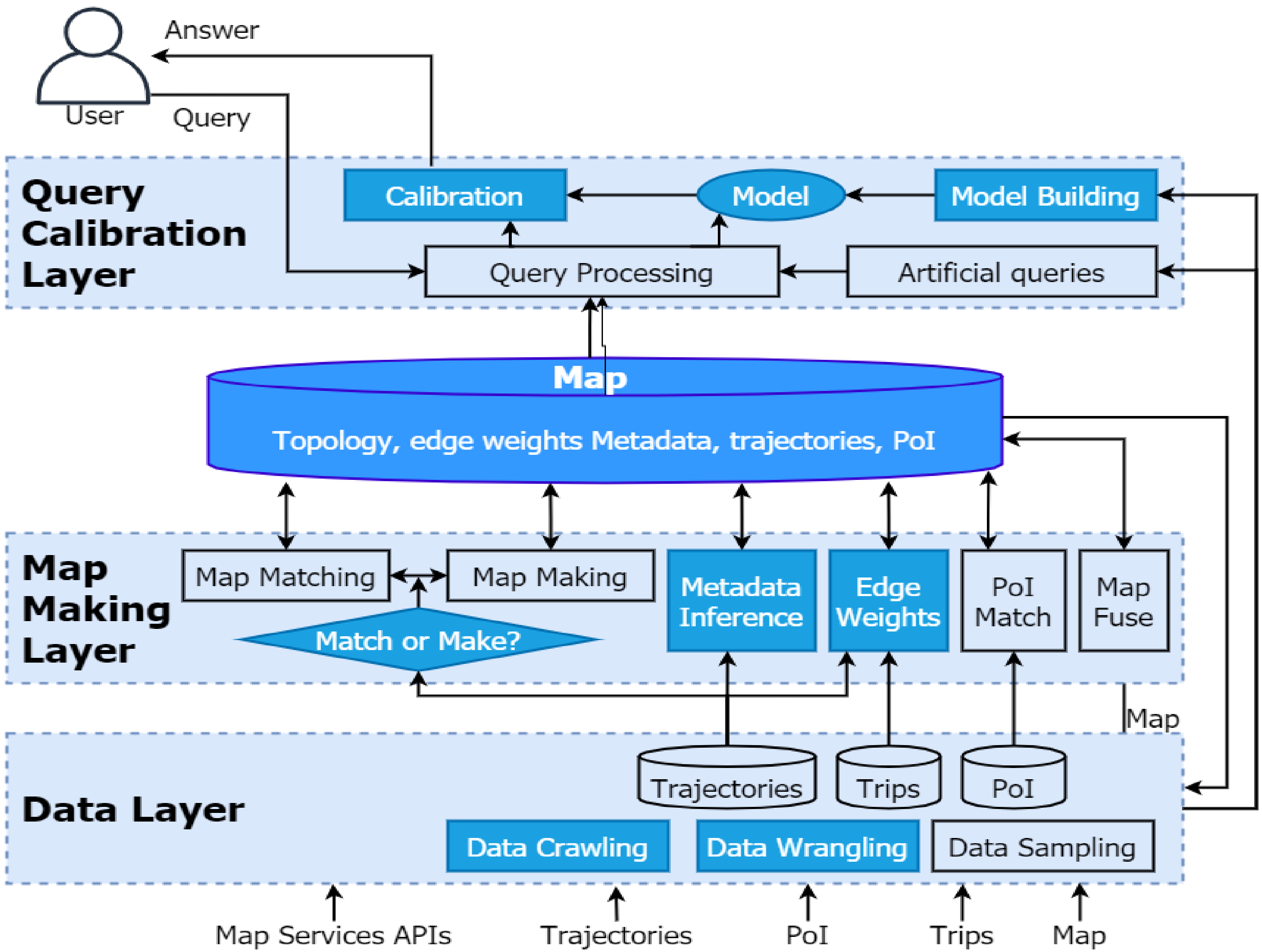
Routing..



Routing Algorithm
Source
Destination
Route
Path
Estimated Time of Arrival (ETA)
Precomputed Routes
Topology
Metadata
Map
ETA: 15MIN
UBER
Mapbox
Waze
DiDi

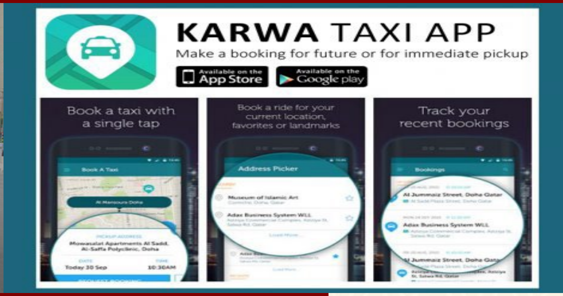
QARTA: An ML-based System for Accurate Map Services

- Map-Centric: QARTA *learns* its own map in terms of topology and metadata
- Query Calibration: QARTA *learns* the error margins of various algorithms and use it to calibrate its answer



QARTA: Why..??

- Problem came up from the Taxi company working in Qatar



Too much construction and road changes in town (in preparation to FIFA 2022)

Commercial maps cannot cope with such changes in road networks, and are not cheap

The Peninsula
Local focus, Global vision
QATAR'S DAILY NEWSPAPER
3rd Best News Website in the Middle East in 2017

Qatar road network increased three times between 2013-18: Ashghal

24 Apr 2018 - 11:58

Al Muhannadi said that the length of the road network increased by about three times between 2013 and 2018 compared to before 2013. He said that the volume of roadworks carried out over the past five years also increased from 1,700 km to 6,000 kilometers, while sanitation capacity doubled, rainwater drainage grew 7 times, and pedestrian trails increased 12 times during the same period.

Raya Daily (Sept. 8, 2020), 20

CACM, April 2021

hot topics arab world

Digital Mapping DOI:10.1145/3447731

Traffic Routing in the Ever-Changing City of Doha

BY SOFIANE ABBAR, RADE STANOJEVIC, SHADAB MUSTAFA, AND MOHAMED MOKBEL

ON DECEMBER 2, 2010, Qatar was announced to host 2022 FIFA World Cup. That was time for celebrating the first-ever Middle Eastern country to organize the tournament. The 1.8M population of Qatar then (2.8M today) never imagined the journey their country was about to embarked. Indeed, in less than 10 years, the population grew by more than a half, pushing the available urban resources and services to their limit. At the same time, the country undertook an ambitious investment plan of $200B on various infrastructural projects including a brand new three-line metro network, six new stadiums, several new satellite cities, and an astonishing 4,300km of new roads, which tripled the size of the road network in only five years.[3]

While this enterprise boosted the socio-economical life of people in Qatar, it did disrupt the way they navigate the urban space and their mobility patterns in general. Simple commutes to work, drops and pickups of kids to and from schools, became challenging and impossible to plan with daily changes in the road layout, including temporary and permanent closures, deviations, new connections, conversions of roundabouts into signaled intersections, turn restrictions, to name but a few. A commute to school that lasted 10 minutes yesterday, could last 25 minutes today. Cab drivers in the city of Doha (Qatar's capital), who are mostly foreigners, also wish they could rely on popular navigation services such as Google Maps, Here, or Tomtom.

Yet, all such systems fall short in coping up with the rapid urbanization and the ever-changing roads in Doha. This was actually depicted in a very popular caricature in one of the most widely distributed daily local newspapers showing Google maps as a limping turtle that is helplessly trying to catch a bunny representing the road changes in the city of Doha.[4]

Besides the general public who is not happy with the routes offered by navigation systems, other stakeholders from public and private sectors were struggling with the poor quality of existing digital maps. For example, the Ministry of Transport and Communication was facing issues getting access to the most accurate map of the road network, needed for their traffic modeling. Also, transportation, delivery, and logistics companies that heavily rely on accurate maps, routes, and travel time estimates were tired of the many lost drivers and missed rendezvous.

Early work: Silent maps are not enough. The issue of inaccurate local maps has triggered an early work at Qatar Computing Research Institute (QCRI) in collaboration with Qatar Mobility Innovation Center (QMIC) to come up with an accurate map for the city of Doha, Qatar.[7] The idea was to use data collected from a fleet of vehicles that are continuously tracked, for accurate and timely detection of road changes, such as new roads, road closures, and detours. Though that early work was successful in coming up with a more accurate map than what navigation systems have, it was not enough to address the main problem of routing. Accurate topological maps do not say much about the time needed to go through each road segment—a main functionality needed for any routing application.

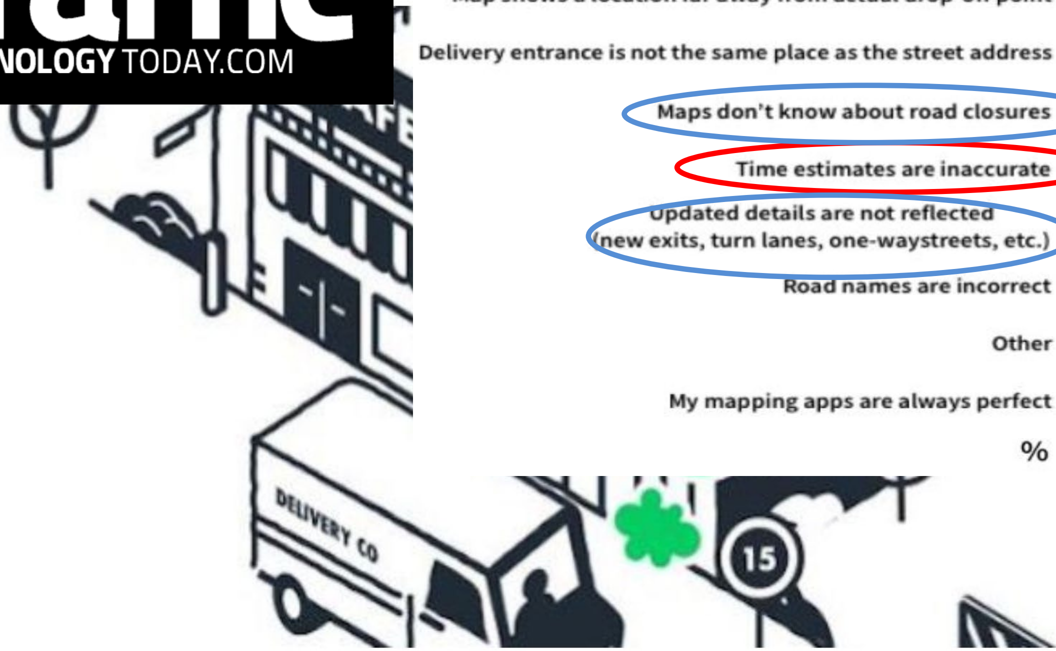
Data access and collaboration. To address the routing problem in the ever-changing roads of Doha, we partnered with the national taxi company Karwa. The collaboration gives us access to all taxi data (both historic and live) that took place in the country, including pick-up and drop-off locations, time, duration, speed, fare, route, as well as sampled GPS points for each trip—a gold mine for our research agenda. But most importantly, we also learned from our partners about the real challenges they face, which helped us prioritize our projects.

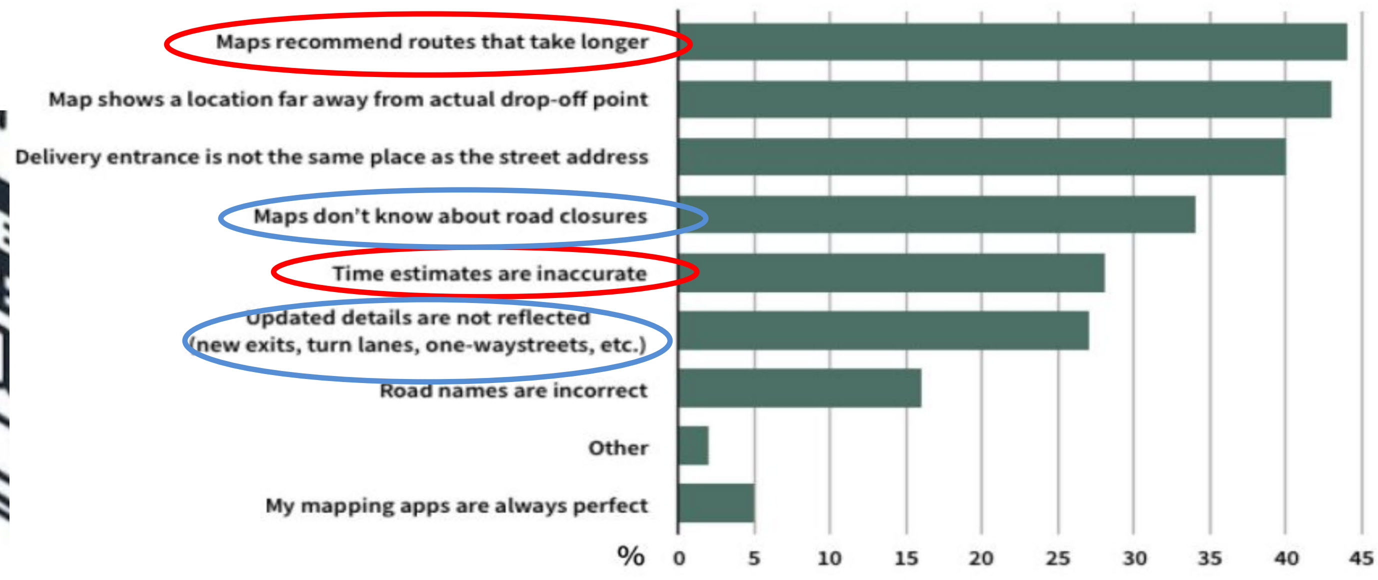
Map enrichments for traffic-aware routing. Our first project with Karwa was to enrich the topological maps with traffic information, that is, accurate edge weights for each road segment for each hour of the day. Inferring traffic information from a large number of vehicles can be relatively straightforward. However, the problem is much more challenging when the data is sparse and does not cover many roads with large frequency. We tackle these problems in Stanojevic et al.[5,6] and derive a traffic layer with an accuracy comparable to the commercial maps using only sparse data available to us either from Karwa Taxi data as in Stanojevic et al.[5] or from using commercial map APIs as in Stanojevic et

PHOTO BY PHILIP LANGE/SHUTTERSTOCK.COM

APRIL 2021 | VOL. 64 | NO. 4 | COMMUNICATIONS OF THE ACM 67

TRAFFIC TECHNOLOGY TODAY.COM





Poor maps costing delivery companies US$6bn annually

0

BY ADAM FROST ON FEBRUARY 19, 2020 MAPPING

Based on a survey of delivery drivers in the USA and conducted by an independent research firm, the first 'Mapping in Logistics Report' has revealed that 'broken maps' are costing the logistics sector an estimated US$6bn annually.

https://www.traffictechnologytoday.com/news/mapping/poor-maps-costing-delivery-companies-us6bn-annually.html

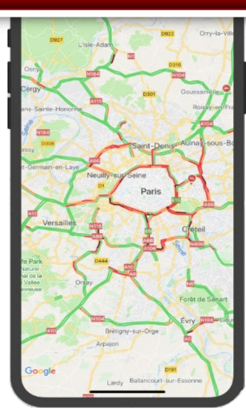
Edge Weight Inference: Who is doing it?

- Traffic departments: Loop detectors or plate recognition



Edge Weights are considered as proprietary information, not to be shared

- Co

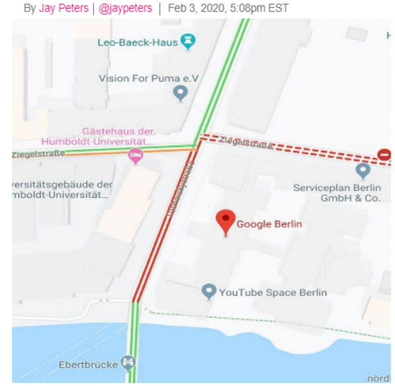



ORE
with

99 phones and a little red wagon

The streets were mostly empty, but the map showed a traffic jam

By Jay Peters | @jaypeters | Feb 3, 2020, 5:08pm EST



Edge Weight Inference in QARTA:

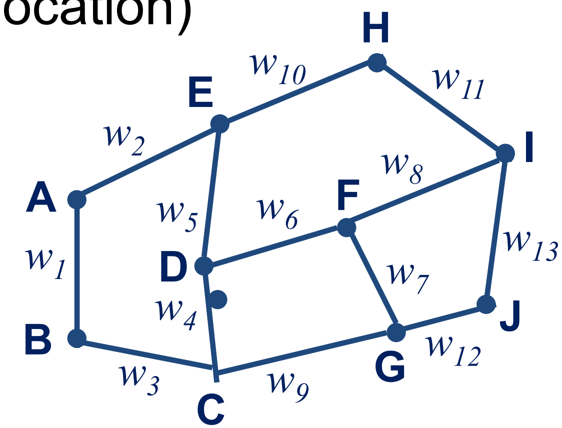
- **Input:** Trips (Pickup time/location, Drop off time/location)

(A, F, 15) ➔ $w_2 + w_5 + w_6 = 15$

(B, H, 28) ➔ $w_3 + w_7 + w_8 + w_9 + w_{11} = 28$

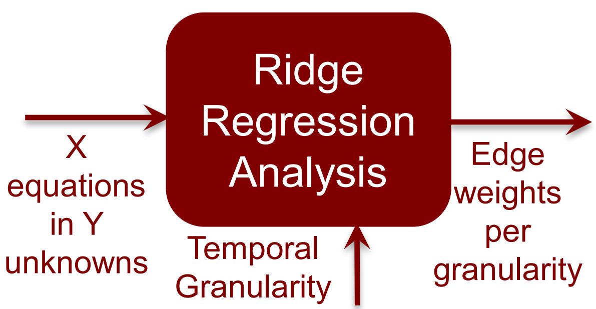
(A, I, 19) ➔ $w_1 + w_3 + w_7 + w_8 + w_9 = 19$

…



- **Objective:** Given a set of edges, each with length l_e and unit *length weight* W_e, a set of trips T, each with a path P_t, find W_e that minimize:

$$\sum_{t\epsilon T}\left(\sum_{e\epsilon Pt} W_e l_e - \delta_t\right)^2$$



- **Challenges:**
  - A direct solution may result in zero or negative weights
  - Scalability is a major issue: Hundreds of thousands of edges with millions of trajectories
  - Over-fitting for unreliable edges
  - Need to accommodate for a fine granularity (e.g., 168 hours per week)

Edge Weight Inference in QARTA

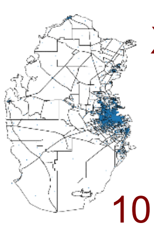
- After several tuning steps: (Details in the paper)
- **Objective:**

Regularization strength

Average speed

$$\sum_{t\epsilon T}\left(\sum_{g:P_t\cap Hg\neq\emptyset} W_g L_g + W_0 \sum_{e\,\epsilon\,(Pt\setminus H)} l_e - \delta_t\right)^2 + \alpha\sum_{g}(W_g - \sigma)^2$$

$$L_g = \sum_{e\,\epsilon\,Hg} l_e$$



X equations in Y unknowns

10K equations in 500K unknowns

Tuning Steps

X\` equations in Y\` unknowns

1K equations in 5K unknowns

Ridge Regression Analysis

Granularity

Hour

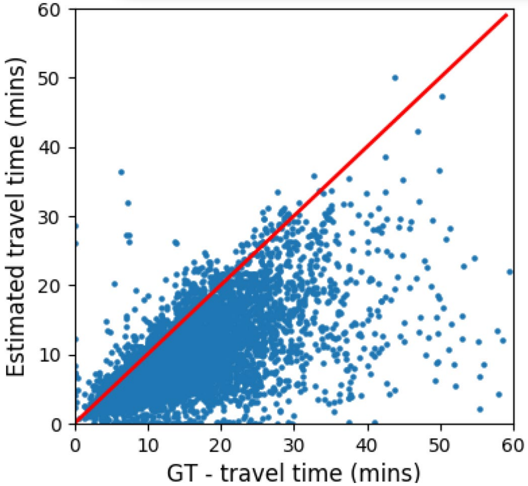
Edge weights per granularity

Edge weights per hour

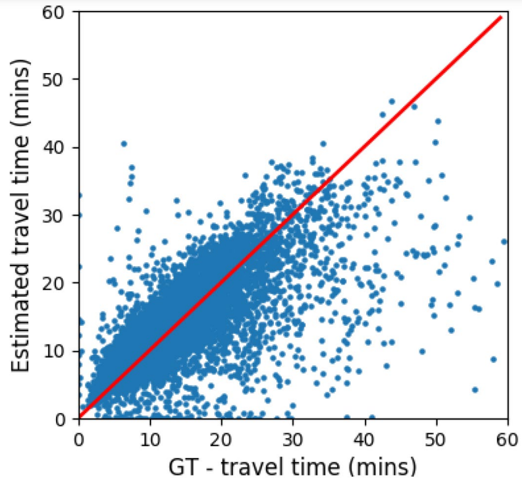
Estimated Time of Arrival (ETA)

- The accuracy of query answers heavily rely on Estimated time of Arrival

Idea: Can we study the error patterns of each algorithm under various context, and use to adjust the query answer.



OSRM



Google Maps

VB

Uber taps ClimaCell to improve ETA estimates with hyper-local weather data

PAUL SAWERS @PSAWERS FEBRUARY 6, 2020 8:00 AM

Uber is partnering with weather technology company ClimaCell to enable more accurate estimated time of arrival (ETA) predictions for drivers and riders.

Founded in 2016, Boston-based ClimaCell specializes in real-time weather forecasts. Rather than relying on government data typically garnered from

Model Building

- Trip: (Pickup time/location, Drop off time/location, δ)
  - δ is the difference between actual and estimated time of the trip

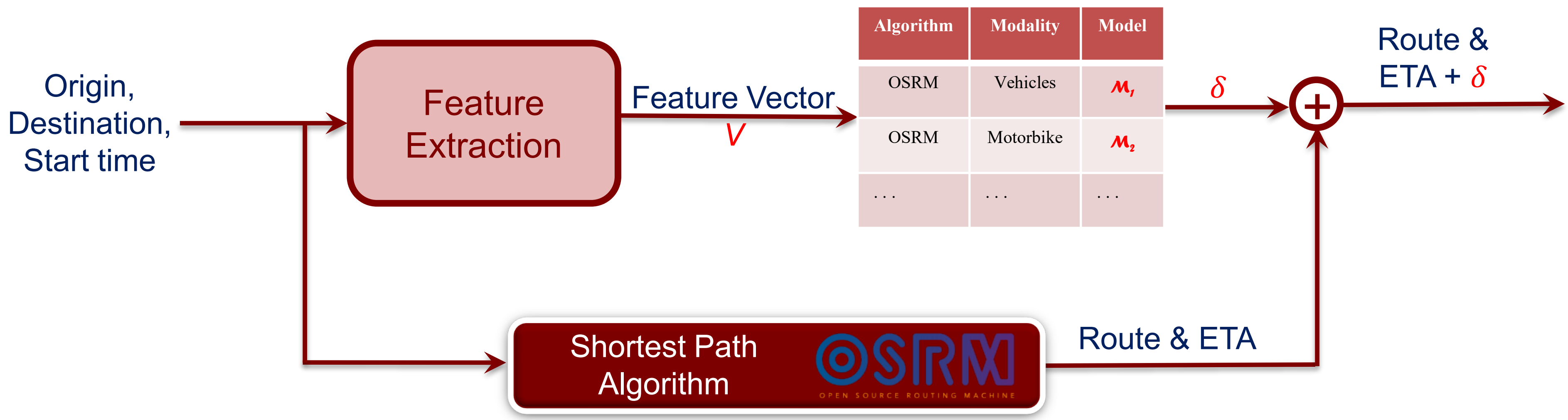
- Features in V that impact δ
  - Spatial Zoning
    - Origin
    - Destination
  - Temporal Zoning
    - Pickup time
    - Drop off time
  - Trip Characteristics
    - Trip distance
    - Trip duration

- A model $\mathcal{M}$ will be built for each ETA algorithm and driving modality

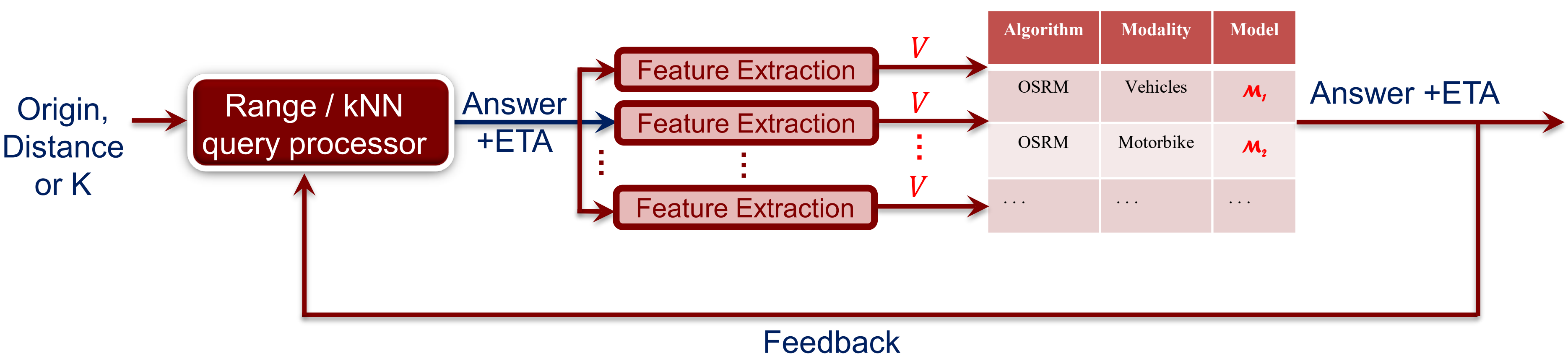
Algorithm	Modality	Model
OSRM	Vehicles	$\mathcal{M}_1$
OSRM	Motorbikes	$\mathcal{M}_2$
. . .	. . .	. . .

Query Calibration in QARTA

- Shortest Path queries



- Range and kNN queries



QARTA in Deployment

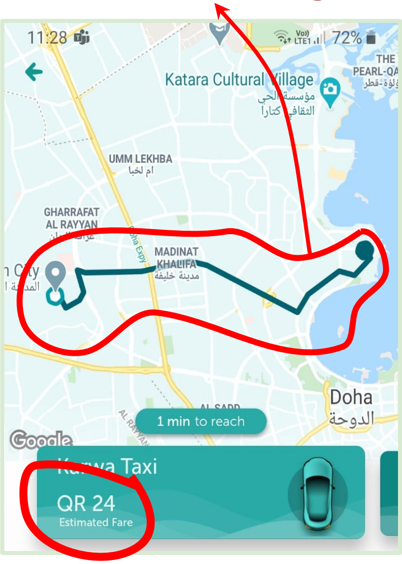
QARTA is deployed in *all* Taxis in Qatar $\sim 4K$ vehicles

A local food delivery company $\sim 3K$ motorbiks

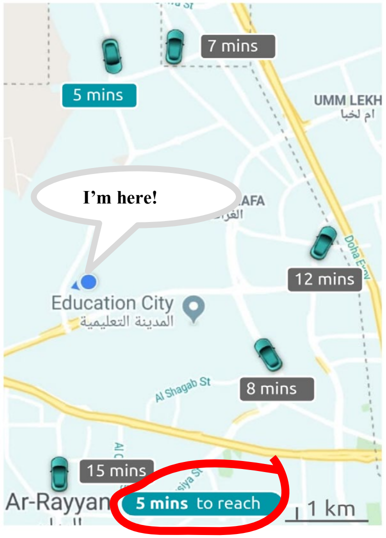
- **QARTA receives:**
  - $\sim 235K$ daily API calls
  - $\sim 1\ Million$ daily GPS tracks

- **APIs & Services:**
  - In-traffic routes
  - Travel time estimation
  - Complex route planning
  - OD matrices
  - Search & addresses

Routing



Fare estimation

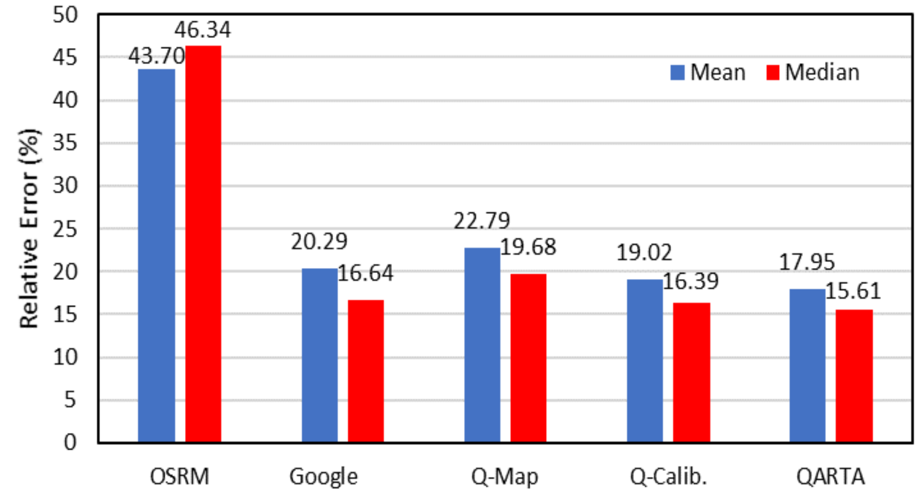


Taxi Dispatching

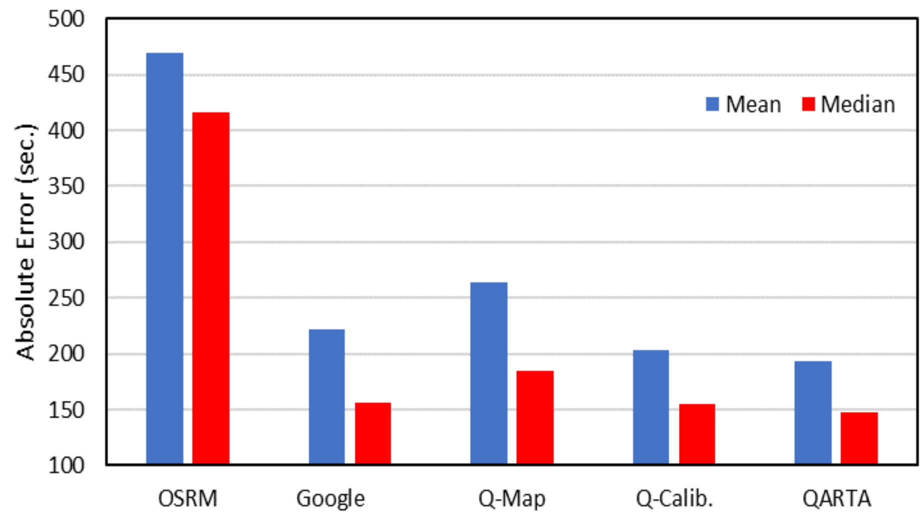
Link: https://qarta.io

QARTA vs Other Map Services: Shortest Path Query

- **Q-Map**: Runs QARTA Map Making layer without any calibration
  - OSRM on QARTA map



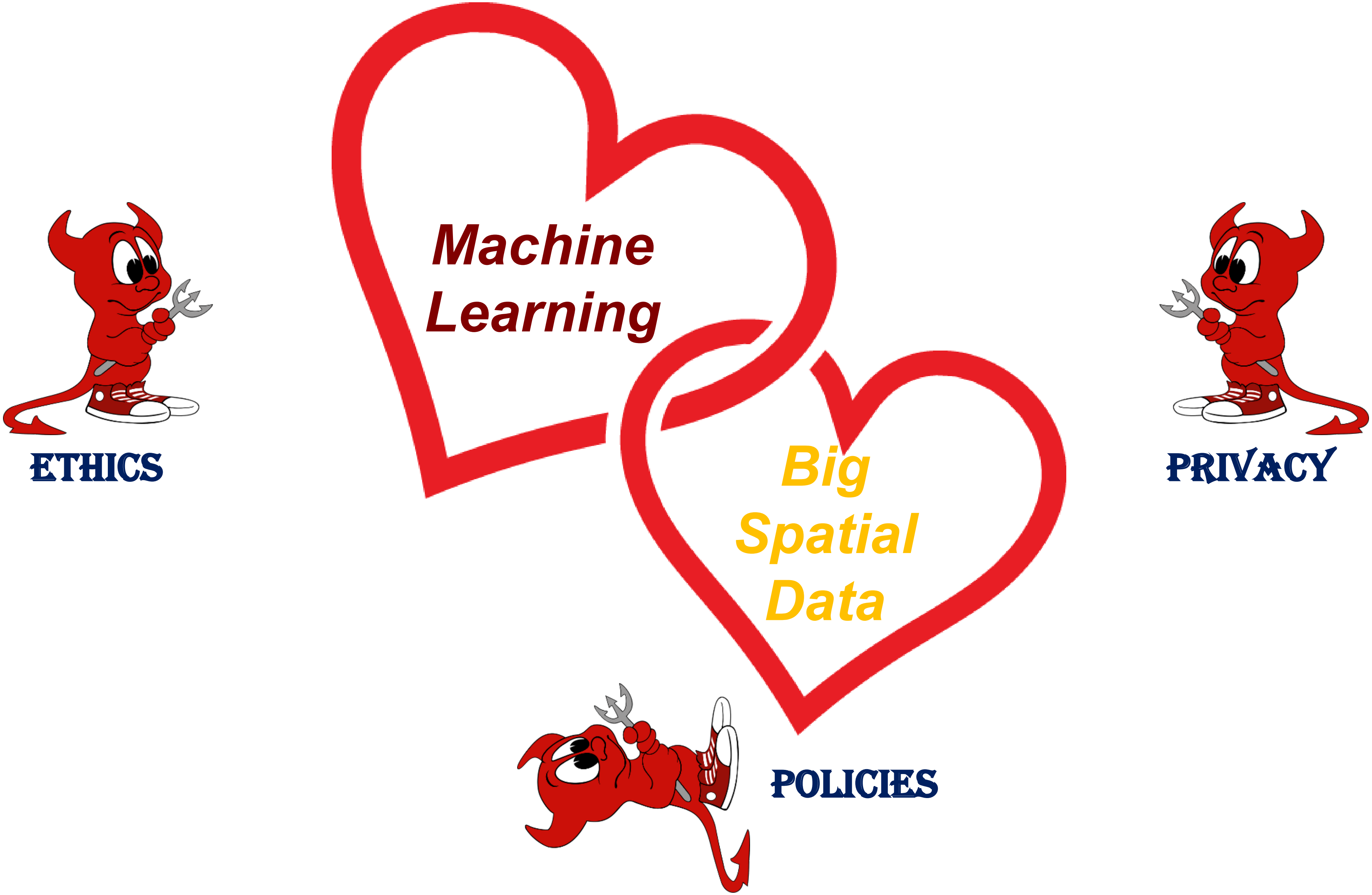
- **Q-Calib**: Runs QARTA calibration without Map Making layer
  - Calibrating OSRM engine



Summary:

Applications
Spatial
Non-Spatial
Routing
SPATIAL REGRESSION MODELS FOR THE SOCIAL SCIENCES
Quantitative Applications in the Social Sciences
155
Spatial Regression Models
SECOND EDITION
Spatial Regression
CYBER SECURIT
digitalhealth
news networks intelligence
Machine Learning Applications
Retail
Travel
Healthcare
Finance
Media
edureka!
02
04
06
09
TOP 10
APPLICATIONS OF
MACHINE LEARNING
Knowledge Base
Knowledge Base
Non-Spatial
Spatial
ML Fundamental Algorithms

Machine Learning meets Big Spatial Data




Thank you