Machine Learning Meets Big Spatial Data

Ibrahim Sabek

Mohamed Mokbel

University of Minnesota
Driven to Discover

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

The Rise of Machine Learning

SmartDataCollective

The Rise of Machine Learning and AI is Improving Lives in 2018

Take a dive into how Machine Learning and AI have impacted the way we live our daily lives.

Bhupinder Kour

January 5, 2018

Forbes

Rise Of The Machines: The Future Of Data Science And Machine Learning

Meghann Chilcott Forbes Councils

Forbes Technology Council CommunityVoice

POST WRITTEN BY

Meghann Chilcott Jul 30, 2018, 07:00am

Senior Vice President of OrderInsite, delivering executive leadership in innovative pharmacy technology solutions. Connect with me.

Broadcast

The rise of machine learning

By Adrian Pennington | 25 September 2017

AI is an increasingly important tool for media companies, helping to automate repetitive tasks and free up staff to focus on delivering quality content.

Much of what is now referred to as Artificial Intelligence (AI) and Machine Learning (ML) is, in reality, just advanced image or metadata analysis. Rather than ‘learning’ by themselves, machines need to be trained in detail to get good results and will only get better through additional training.

DZone AI Zone

The Rise of Machine Learning

Let's take a look at a brief article that explores machine learning and how the recent surge of data has empowered a field of computer science.

by olu campbell · Aug. 25, 18 · AI Zone · Opinion

IDC forecasts that spending on Machine Learning will grow from $12 billion in 2017 to $57.6 billion by 2021. What's more, Machine Learning patents grew at a 34 percent CAGR between 2013 and 2017, making it the third-fastest growing category of all patents granted.

ORACLE ORACLE MAGAZINE Topics Roles Issues

FROM THE EDITOR

The Rise of Machine Learning

When smartphones, cars, and other devices learn, businesses and people win.

By Tom Haunert

July/August 2016

Futurists and science fiction writers have created some high expectations over the years. “Where's my flying car?” has become a classic rhetorical question as people look

TechRepublic. SEARCH

Why machine learning will see explosive growth over the next 2 years

By **Macy Bayern** in **Artificial Intelligence**

on September 18, 2018, 7:21 AM PST

While current production of machine learning projects are low, 96% of companies expect them to increase in the next couple years.

PHYS.ORG Nanotechnology Physics

The rise of machine learning in astronomy

eptember 4, 2018, Particle

he SKA will have over 2000 radio dishes and 2 million low-frequency antennas once finished.

When mapping the universe, it pays to have some smart programming. Experts share how machine learning is changing the future of astronomy.

Packt Search... Web Development Data Mobile Programming

The rise of machine learning in the investment industry

By Natasha Mathur - February 15, 2019 - 4:00 am 894 0

The investment industry has evolved dramatically over the last several decades and continues to do so amid increased competition, technological advances, and a challenging economic environment. In this article, we will review several key trends that have shaped the investment environment in general, and the context for algorithmic trading more

BANK INFO SECURITY

The Rise of Machine Learning in Cybersecurity

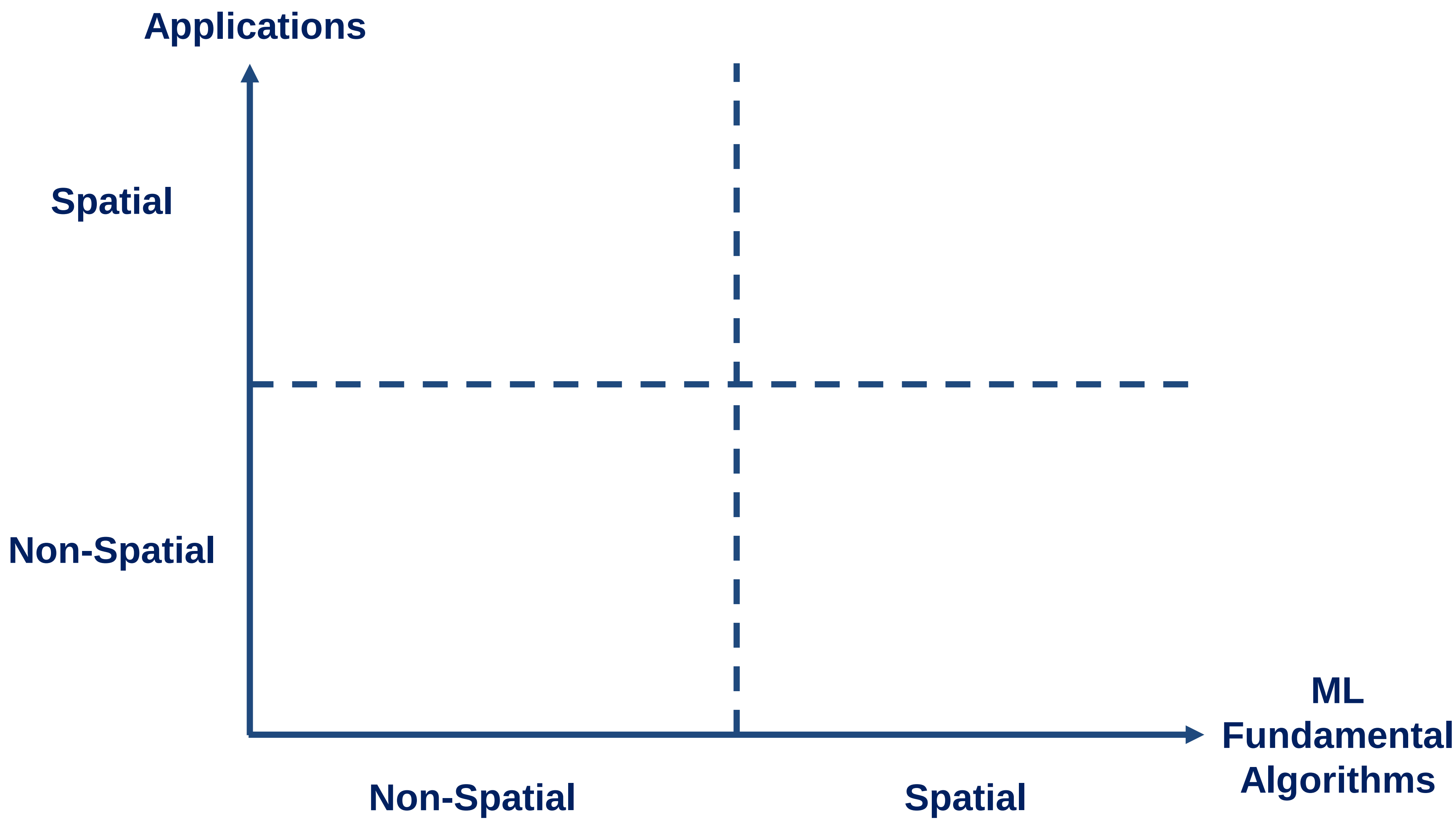
CrowdStrike · August 28, 2018

How the critical capability of machine learning can help prevent today's most sophisticated attacks

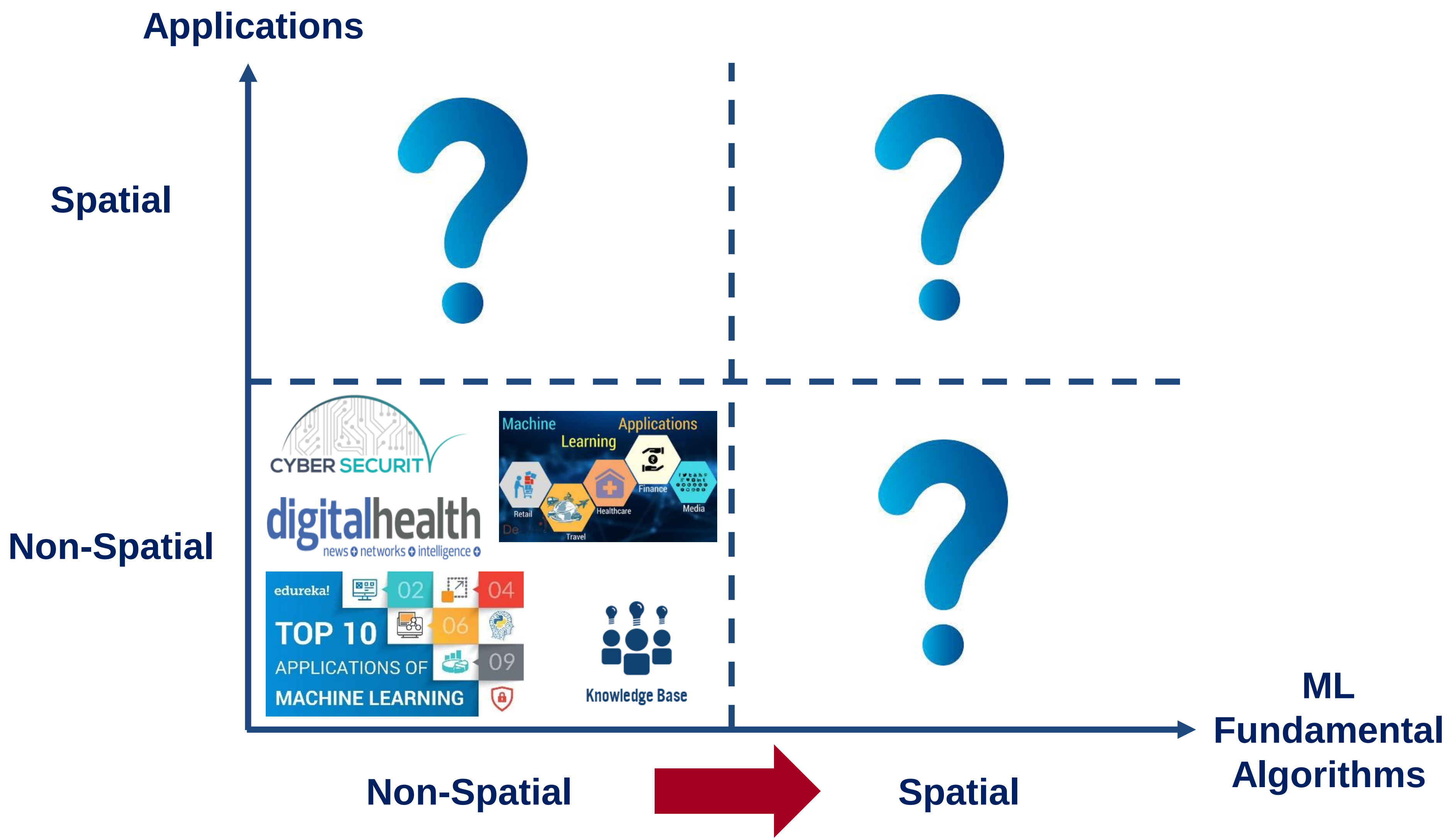
“Machine learning is a core, transformative way by which we’re rethinking everything we’re doing.” -Google CEO Sundar Pichai

The Ubiquity of Big Spatial Data

Machine Learning meets Big Spatial Data



Machine Learning meets Big Spatial Data



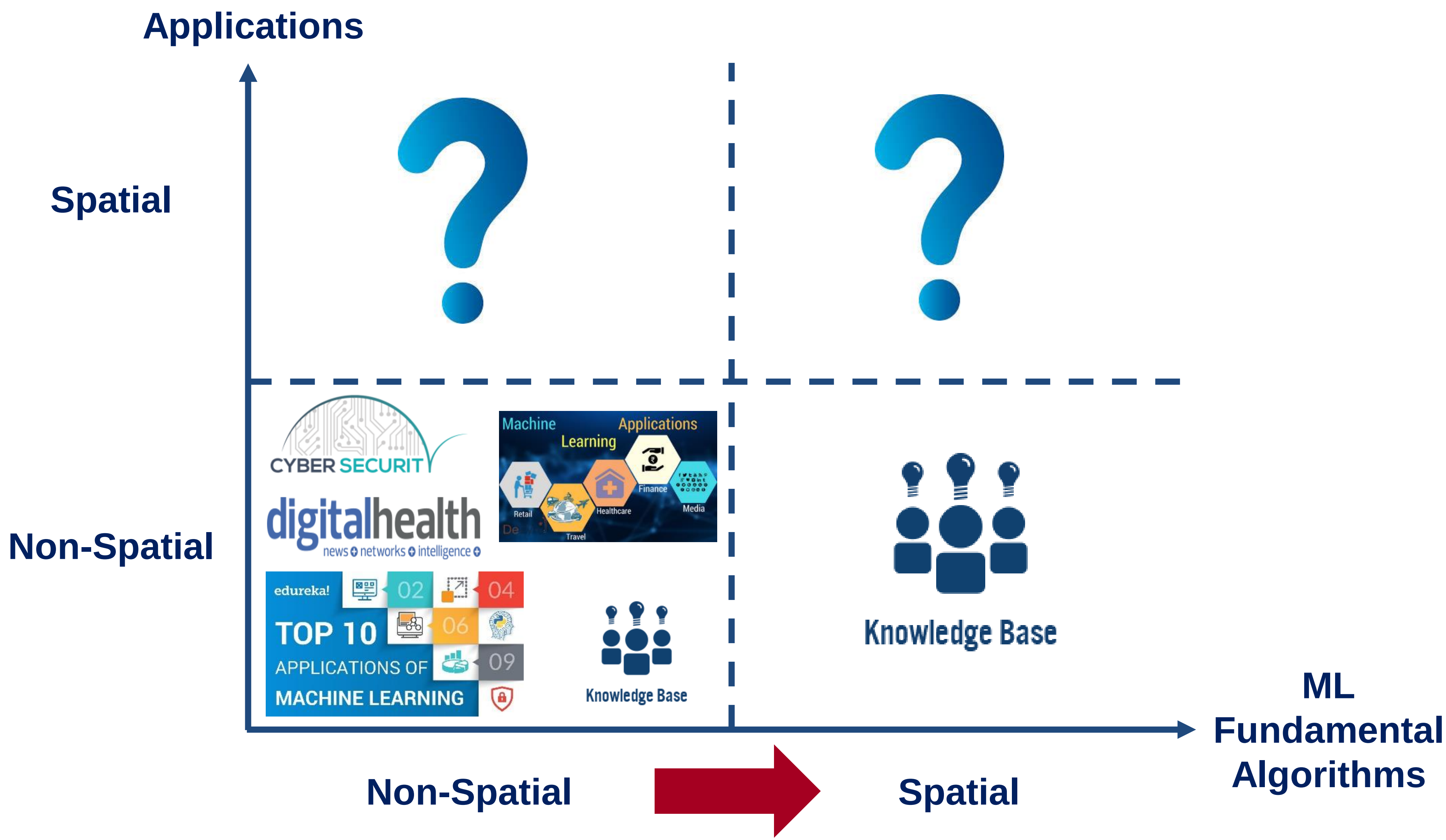
University of Minnesota
Driven to Discover

QCRI
Qatar Computing Research Institute
HAMAD BIN KHALIFA UNIVERSITY

Outline

- Introduction
- Motivation
- Detailed Techniques
- End-to-End Systems

Machine Learning meets Big Spatial Data



Knowledge Base Construction

Knowledge Base Rules

Knowledge Base Construction System

Relations Extraction

tables, relations of facts

Spouses KB

Person 1	Person 2
Barack	Michelle

Probabilistic 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

Probabilistic 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 SystemT StatSnowBall

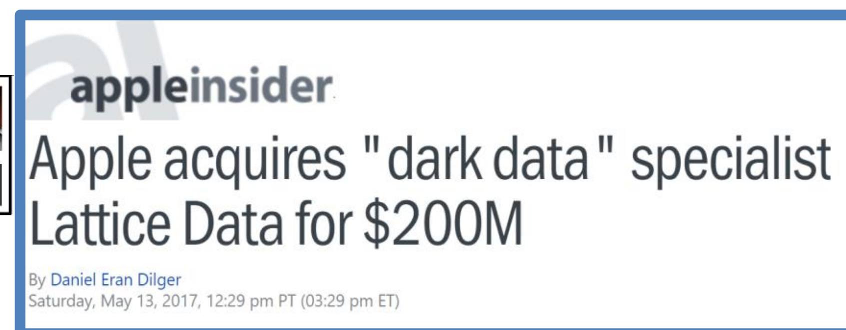
DARPA MEMEX

Fight Human Trafficking
Crime Investigation

DeepDive ProbKB

SCIENTIFIC AMERICAN WSJ 60 MINUTES CNN Forbes BBC

lixto DELIVERING COMPETITIVE ADVANTAGE



appleinsider

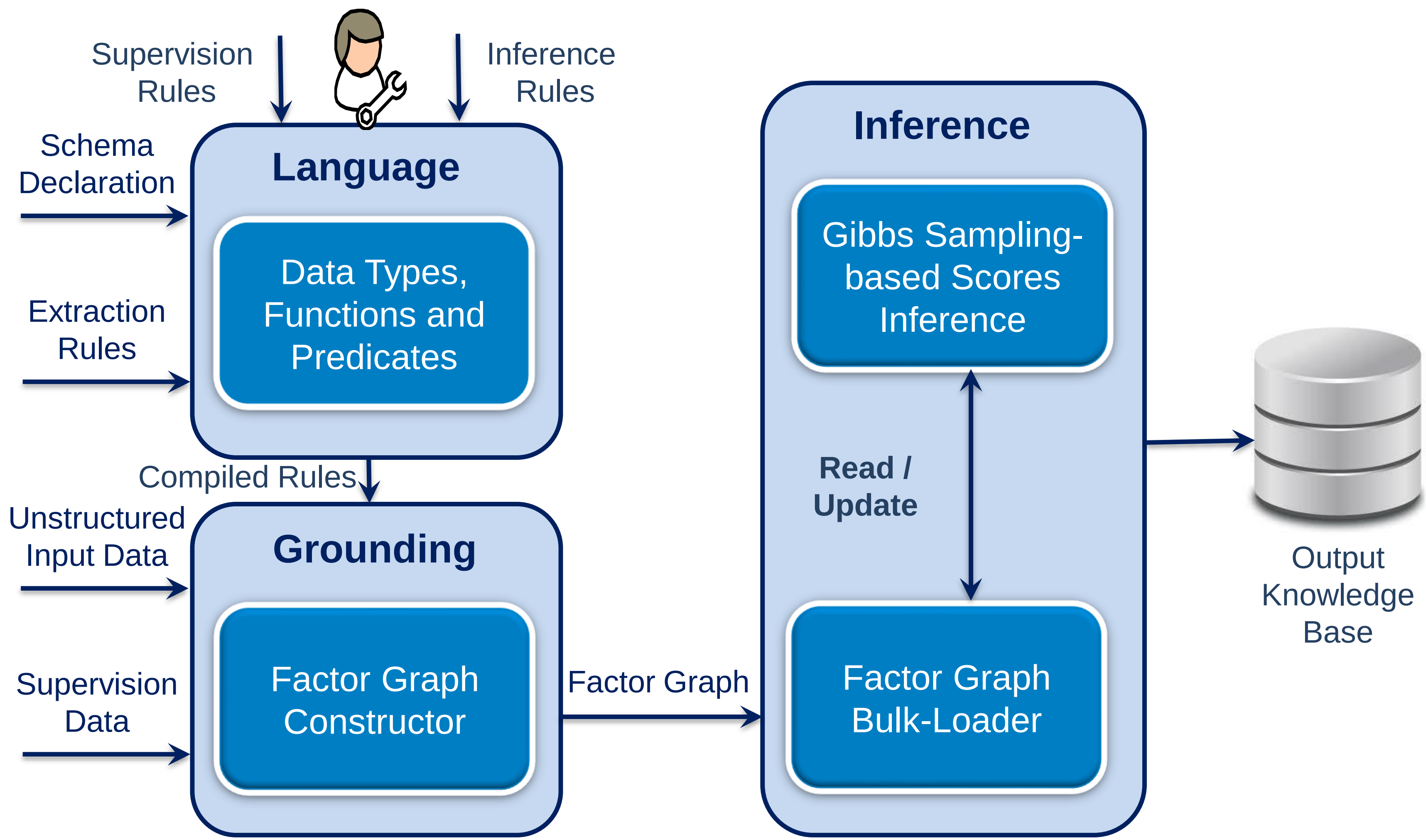
Apple acquires "dark data" specialist Lattice Data for $200M

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

DBLife Google Vault

DeepDive: ML-based Knowledge-Based Construction

Built on scalable implementation of Markov Logic Networks



C. Zhang, C. Ré, M. Cafarella, J. Shin, F. Wang, S. Wu. "*DeepDive: Declarative Knowledge Base Construction*" In **Communications of ACM** 60(5): 93-102 (2017)

F. Niu, C. Ré, A. Doan, J. Shavlik. "*Tuffy: Scaling up Statistical Inference in Markov Logic Networks using an RDBMS*" In **PVLDB** 4(6): 373-384 (2011)

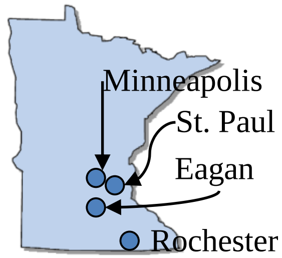
DeepDive with Spatial Data …

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

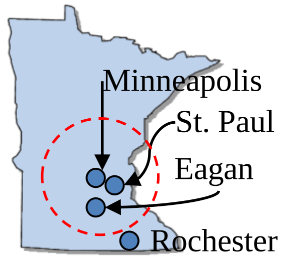
DeepDive with Spatial Data …

Crime rates in Minnesota

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
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

DeepDive

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

Result

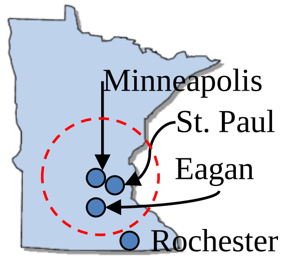
DeepDive with Spatial Data …

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

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   (struck through)
Rule: P1&P2&P3 → Y has high crime rate
```

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

```
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		

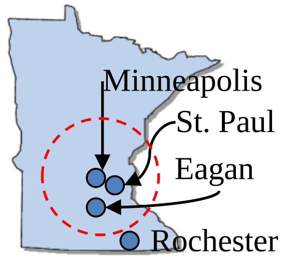
DeepDive with Spatial Data …

Crime rates in Minnesota

Crimes

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

Education



Data

Ebola infection rates in Liberia

Infections

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

Sanitation

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
```

```
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	

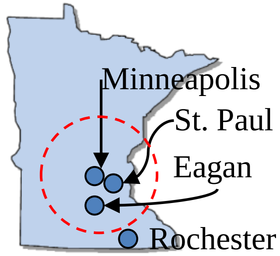
DeepDive with Spatial Data …

Crime rates in Minnesota

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

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
```

DeepDive

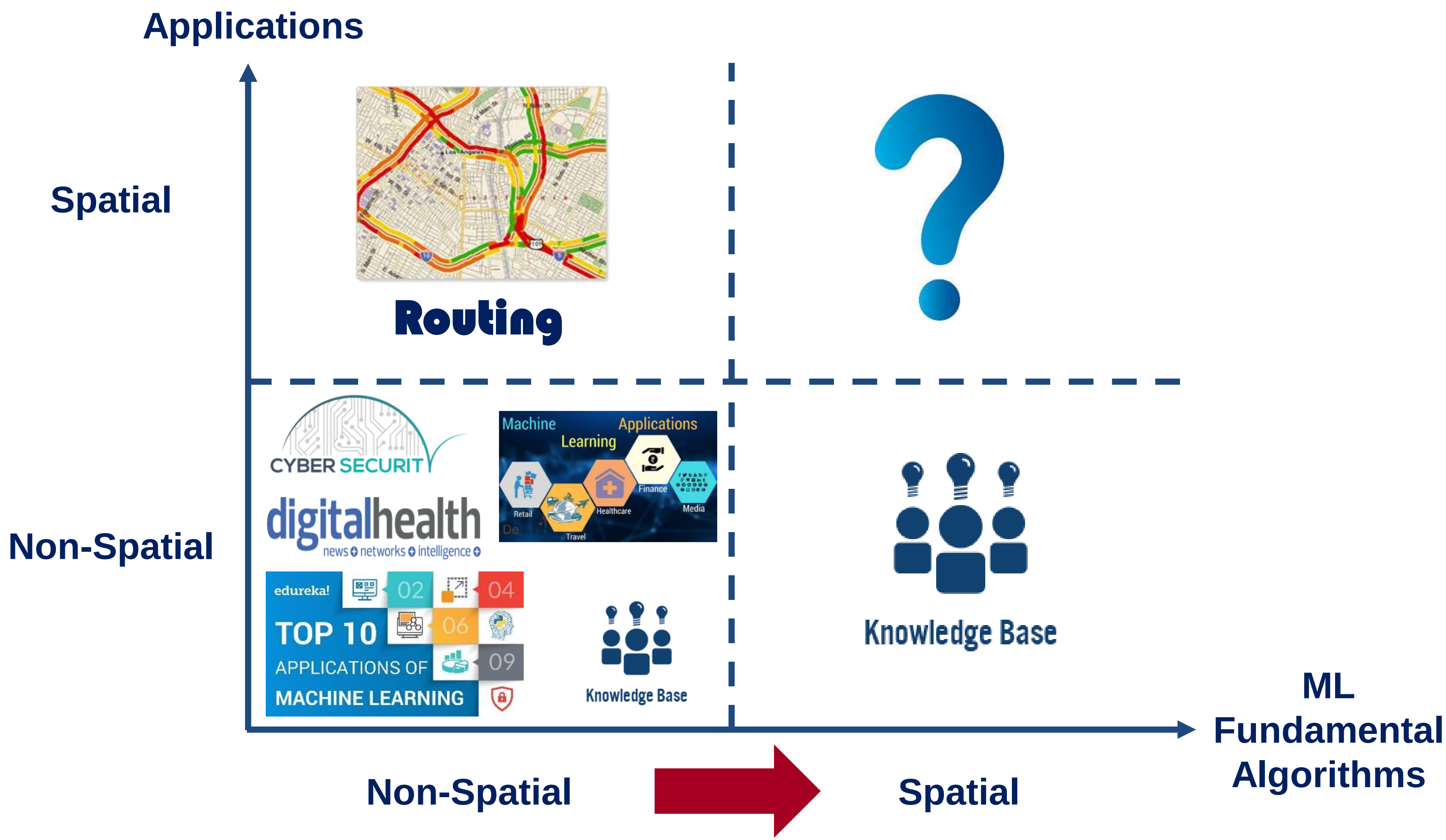
DeepDive

Result

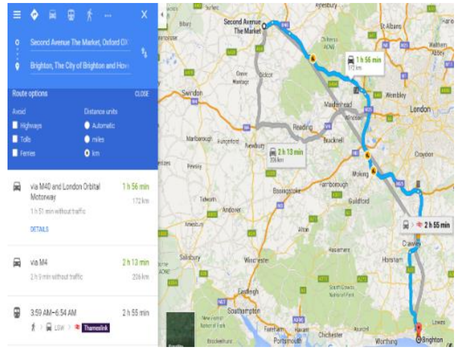
City	Confidence		
St. Paul	~~0.5~~	~~0.7~~	
Eagan	~~0.5~~	~~0.7~~	
Rochester	~~0.5~~	~~0~~	

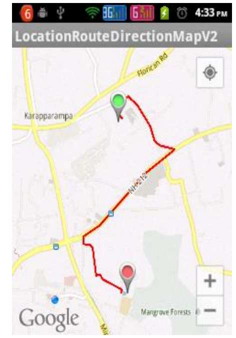
City	Confidence		
Margibi	~~0.54~~	~~0.51~~	
Bong	~~0.52~~	~~0.45~~	
Gbarpolu	~~0.63~~	~~0.06~~	

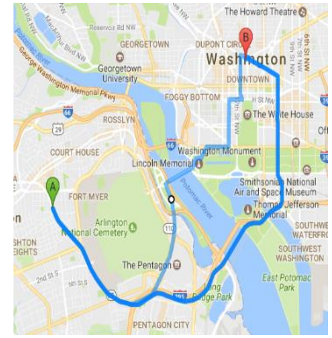
Machine Learning meets Big Spatial Data

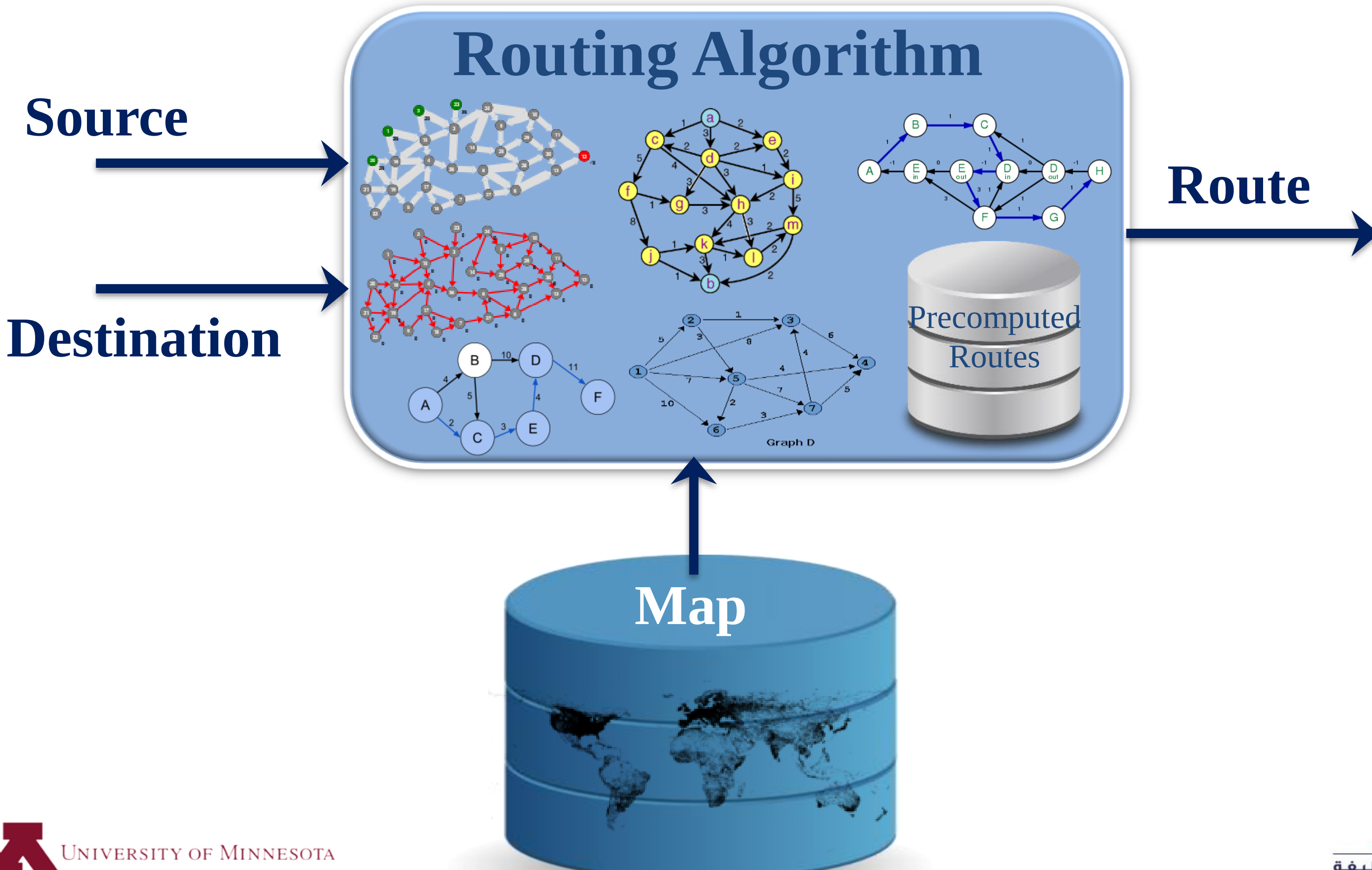


Routing..

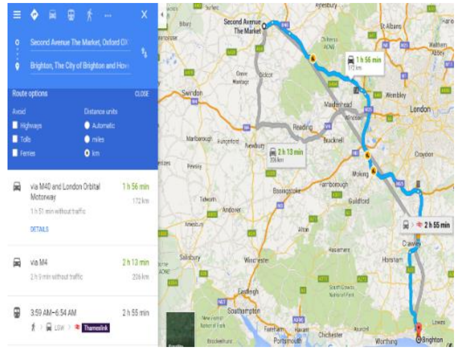


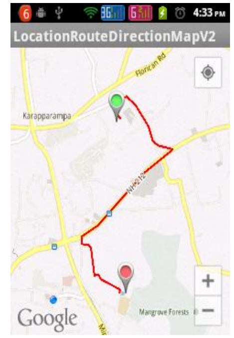

LocationRouteDirectionMapV2
Google

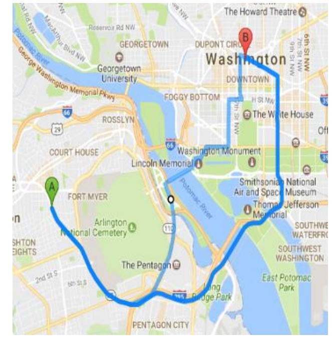



Routing Algorithm
Source
Destination
Route
Precomputed Routes
Map
Graph D

Routing..

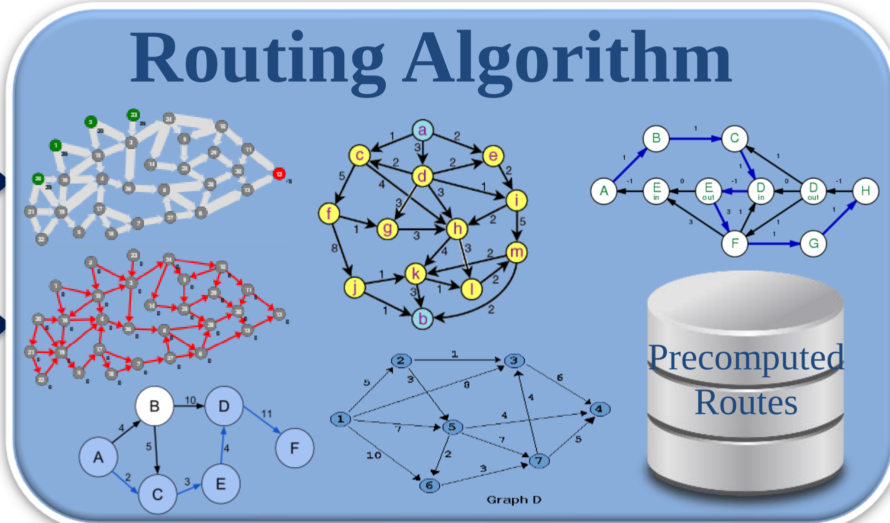






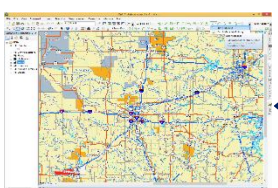
Source

Destination



Route

Topology



Metadata

University of Minnesota
Driven to Discover

QCRI
Qatar Computing Research Institute
HAMAD BIN KHALIFA UNIVERSITY

npr

NEWS | ARTS & LIFE | MUSIC | SHOWS & PODCASTS | SEARCH

TECHNOLOGY

Google Maps Leads About 100 Drivers Into A 'Muddy Mess' In Colorado

June 27, 2019 · 11:35 AM ET

MERRIT KENNEDY

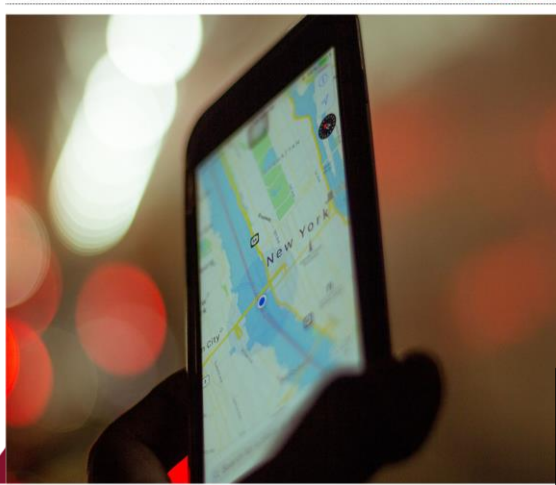
Nearly 100 drivers were recently led astray on a Google Maps-suggested detour near Denver's airport.

POPULAR SCIENCE

Google and Apple Maps have plenty of errors. Here's how to fix them.

Flag missing roads, update restaurants' opening hours, and more.

By David Nield | November 1, 2018



GADGET HACKS

BEST NAVIGATION APPS

Google Maps vs. Apple Maps vs. Waze vs. MapQuest

Best Maps & Navigation Apps for Mobile

GADGET HACKS	Google Maps	Apple Maps	Waze	MapQuest
General				
Platform	Android, iOS, macOS, Windows	iOS, macOS	Android, iOS, Windows	Android, iOS, macOS, Windows
Map Features				
Countries & Terriories Mapped	266	181	72	252
Countries & Territories with Driving Directions	256	101	72	252
Street View	Yes	No	No	No
Overlays	Satellite, Terrain, Transit, Traffic, Bicycling	Satellite, Transit	None	Satellite
3D View	3D Structures	3D Renderings	No	No
Live Location Sharing	Yes	Yes	No	No
Location History	Yes	Yes	Yes	Yes
Cultural Hotspot Indicators	Yes	No	No	No
Weather Data	None	Weather, Temperature, Air Quality	None	Weather, Temperature
Indoor Maps	Airports, Malls, Museums	Airports, Malls	No	No
Offline Maps	Yes	Yes	No	No
Navigation Features				
Traffic Data	Alternate Routes, Accidents, Road Work, Speed Traps	Alternate Routes, Accidents, Road Work	Accidents, Alt. Routes, Road Work, Potholes, Police, Speed Traps	Accidents, Road Work, Traffic Cameras
Traffic Data Source	In-House, User Curated	In-House, Third-Party	In-House, User Curated	Third-Party, User Curated
High Traffic Warnings	Yes	Yes	No	No
Speed Limits	Yes	Yes	Yes	Yes
Lane Guidance	Yes	Yes	Yes	Yes
Add Toll & HOV Passes	No	No	Yes	No
Avoid Tolls & Highways	Yes	Yes	Yes	Yes
Choose Different Routes	Yes	Yes	Yes	Yes
Add Pit Stops	Unlimited	1	1	Unlimited
Show Gas Prices	Yes	Yes	Yes	Yes
Hands-Free Control in App	Yes	Yes	Yes	No
Directions Using Other Modes of Transport	Transit, Biking, Walking, Ride Share	Transit, Walking, Ride Share	Motorcycles, Taxis	Biking, Walking
Re-Center	Yes	Yes	Yes	Yes
Accessible Navigation	Yes	Yes	No	No
Save Parking Spot	Yes	Yes	No	No
Offline Navigation	Yes	Yes	No	No
Works With Screen Off	Yes	Yes	Yes	Yes
App Features				
Dark Mode	Yes	Yes	Yes	Yes
Ride Share Integration	Uber, Lyft, Lime	Uber, Lyft	None	None
Picture In Picture	Yes (Android Only)	No	No	No
Lock Screen Navigation	Yes	Yes	Yes	Yes
Show Festivals & Protests	No	No	Yes	No
Personalized Recommendations	Yes	No	No	No
Book Dinner Reservations	Via OpenTable	Via OpenTable	No	No
Report Traffic Issues	No	No	Yes	No
Post Reviews	Yes	No	No	No
Car Support	Android Auto, CarPlay	CarPlay	Android Auto, CarPlay	No
AR Features	Interactive Street View	Flyover	None	None
Widgets	Yes	Yes	Yes	No
Music Integration	Yes	Yes	Yes	No

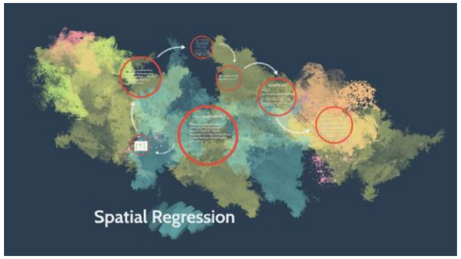
https://smartphones.gadgethacks.com/how-to/best-navigation-apps-google-maps-vs-apple-maps-vs-waze-vs-mapquest-0194591/

Machine Learning meets Big Spatial Data

Applications

Spatial

Routing



Spatial Regression

Non-Spatial

CYBER SECURIT

digitalhealth
news networks intelligence

Knowledge Base

Knowledge Base

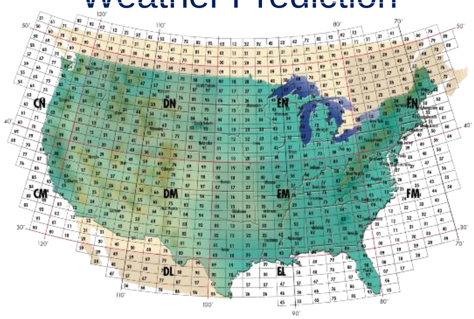
ML Fundamental Algorithms

Non-Spatial → **Spatial**

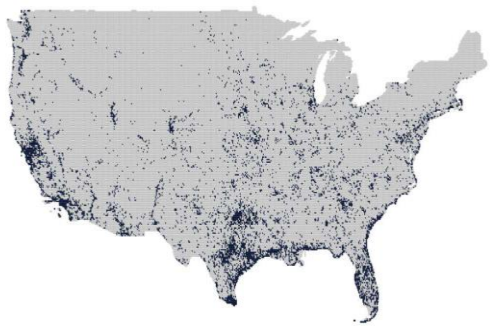
Spatial (Autologisitc) Regression

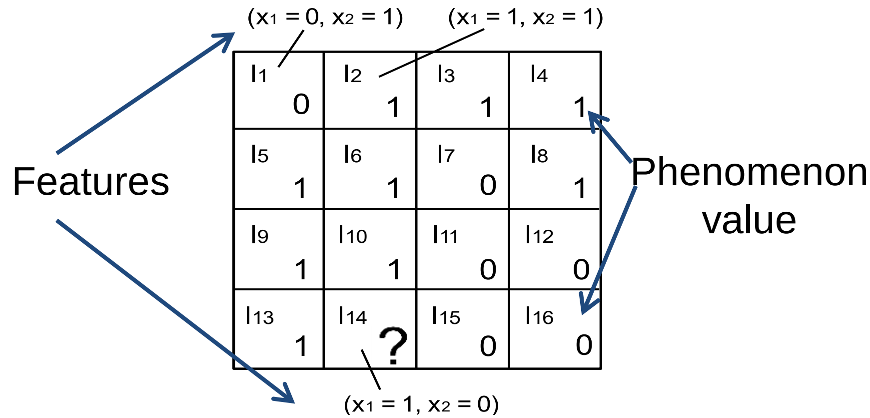
- Find whether a spatial phenomenon exists or not, based on neighbor values and features

Weather Prediction

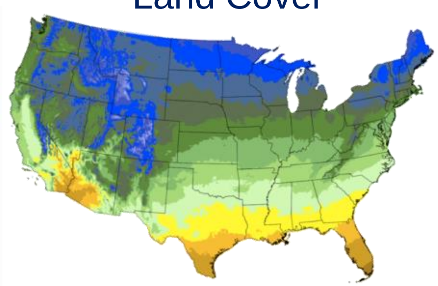


Birds Migration

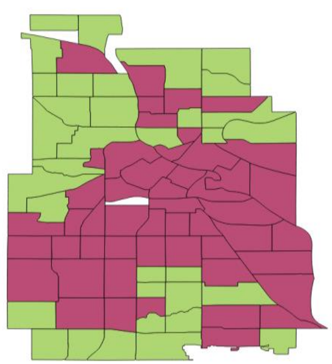




Land Cover



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 ☹

Outline

- Introduction
- Motivation
- Detailed Techniques
- End-to-End Systems

Machine Learning meets Big Spatial Data

Applications

Spatial

Non-Spatial

CYBER SECURIT

digitalhealth
news networks intelligence

Knowledge Base

Knowledge Base

Event Detection

Image Analysis

ML Fundamental Algorithms

Non-Spatial → **Spatial**

University of Minnesota
Driven to Discover

QCRI
Qatar Computing Research Institute
معهد قطر لبحوث الحوسبة
HAMAD BIN KHALIFA UNIVERSITY
جامعة حمد بن خليفة

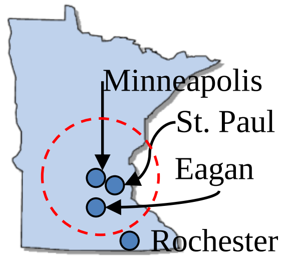
DeepDive with Spatial Data …

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

DeepDive

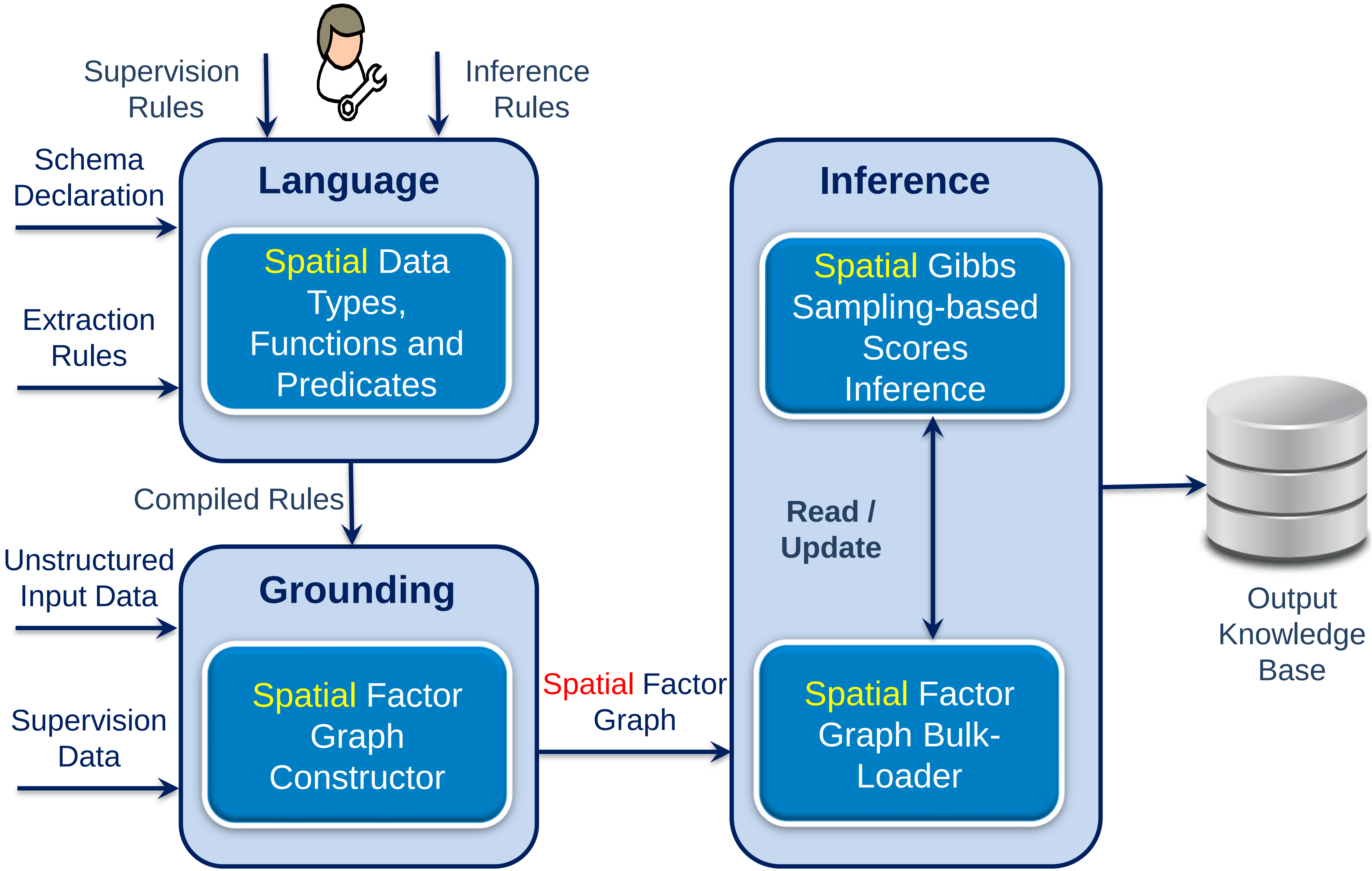
DeepDive

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~~	

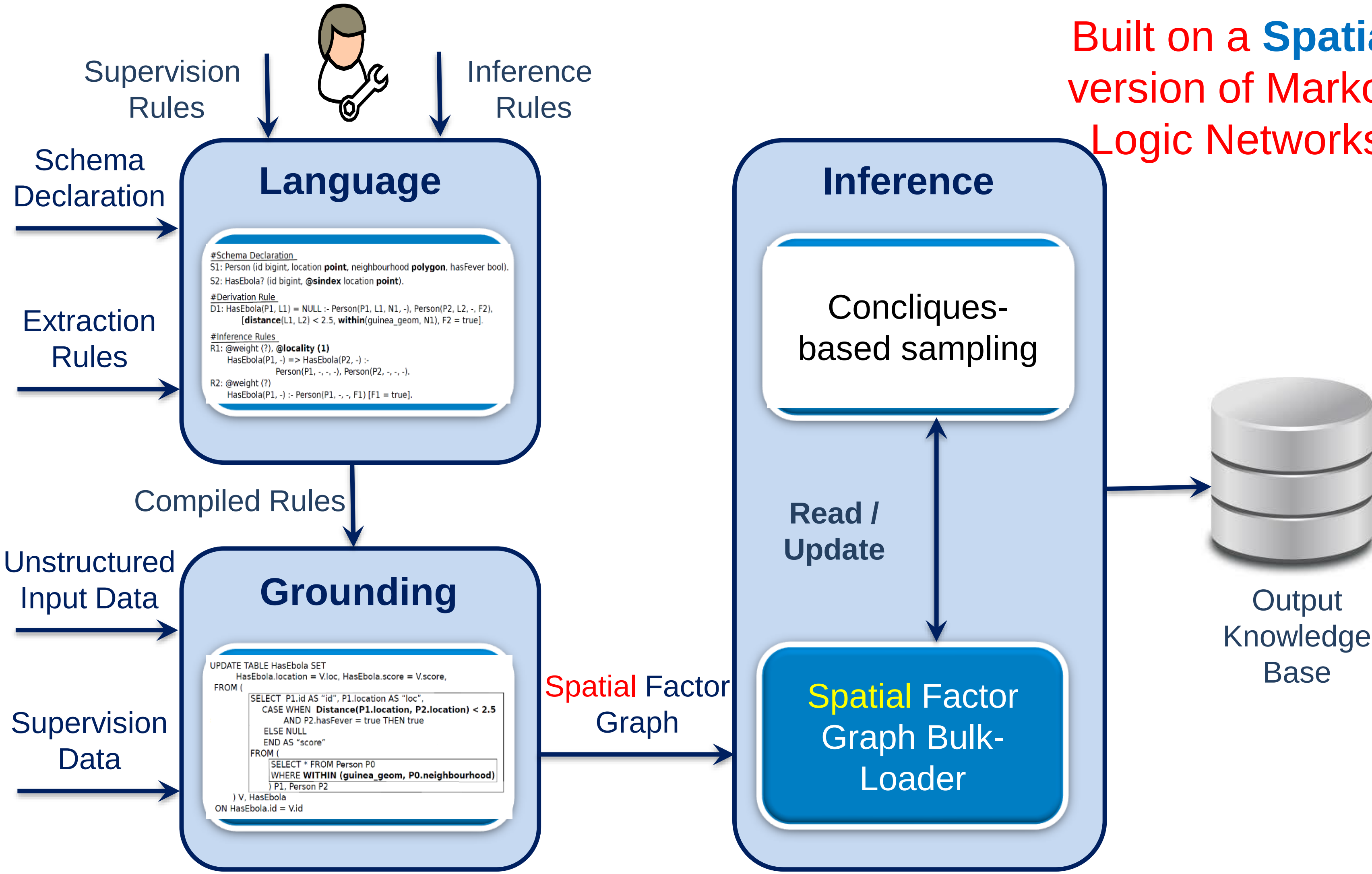
From DeepDive to Sya: A Spatially-Aware Knowledge-Based Construction System



I. Sabek, M. Musleh, and M. F. Mokbel. "*A Demonstration of Sya: A Spatial Probabilistic Knowledge Base Construction System*". In **SIGMOD 2017**

From DeepDive to Sya: A Spatially-Aware Knowledge-Based Construction System

Built on a **Spatial** version of Markov Logic Networks



I. Sabek, M. Musleh, and M. F. Mokbel. “*A Demonstration of Sya: A Spatial Probabilistic Knowledge Base Construction System*”. In **SIGMOD 2017**

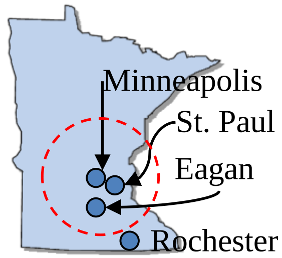
Knowledge-Base Construction with Sya

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

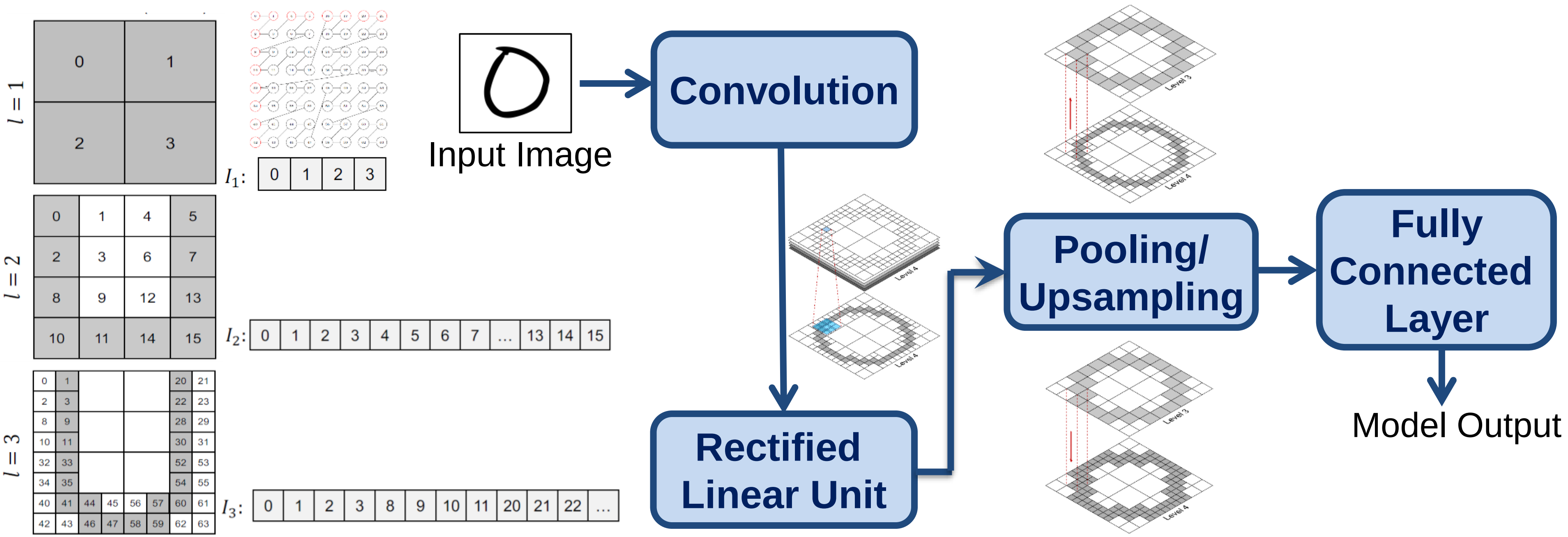
Result

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

Spatially-Aware ML-based Image Analysis

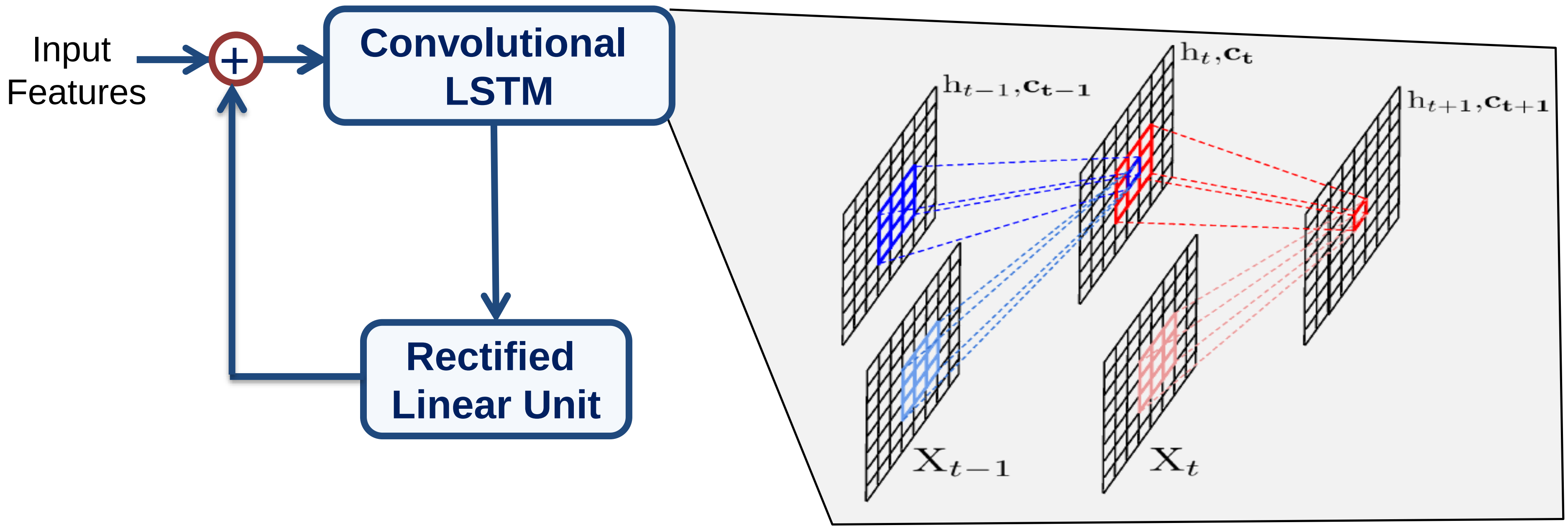
- Sparse object detection in images (e.g., OCR)
- Using Quadtree to improve the performance of Convolutional Neural Networks (CNN) for sparse datasets (e.g., handwriting)
  - Traditional CNNs are optimized for dense datasets



P. Jayaraman, J. Mei, J. Cai *et al*. "*Quadtree Convolutional Neural Networks*". In **ECCV 2018**

Spatially-Aware ML-based Event Detection

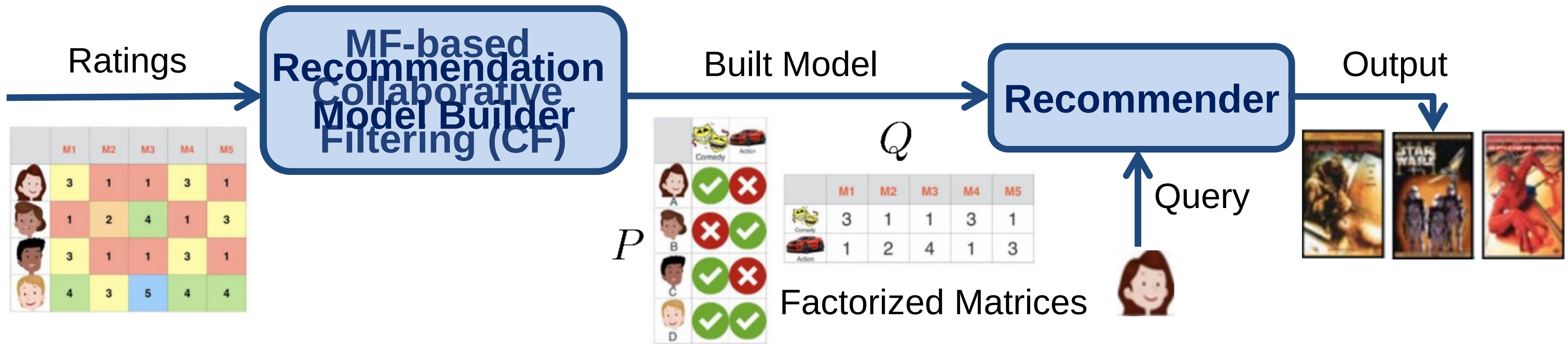
- Predicting a sequence of spatiotemporal tweet counts
  - Traditional modeling uses Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) → focuses only on temporal aspect
- Combining the spatial convolution with LSTM networks



H. Wei, H. Zhou, J. Sankaranarayanan, S. Sengupta, and H. Samet. "*Residual Convolutional LSTM for Tweet Count Prediction*". In **WWW 2018**

Spatially-Aware ML-based Recommender System

- Analyze user behavior to recommend interesting items



- Spatio-temporal Collaborative Filtering
  - Exploiting spatial and temporal correlations across users/items

Spatial Regularization for Users

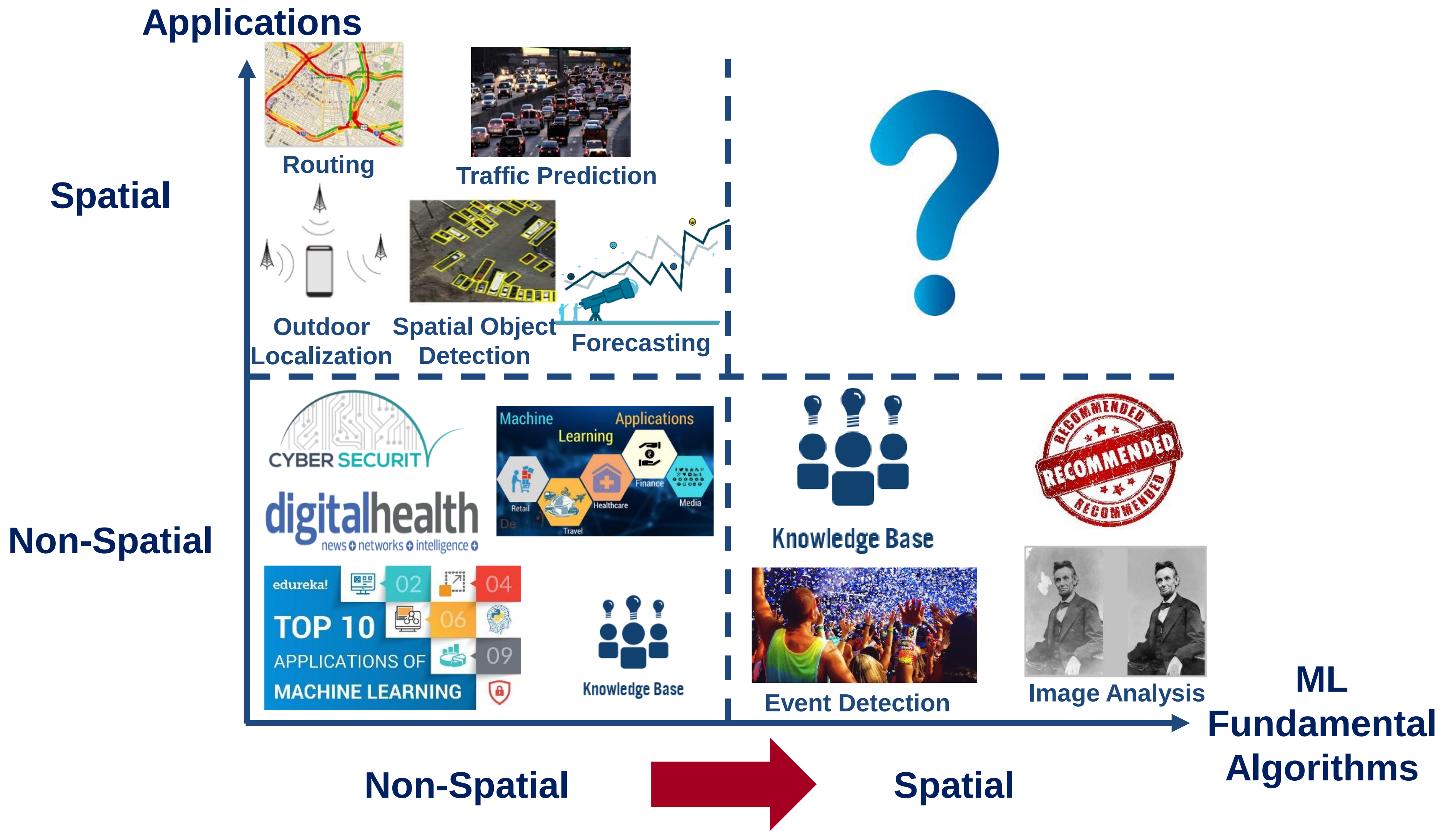
$$\| B \odot (R - PQ^T) \|_F^2 + \lambda(\| P \|_F^2 + \| Q \|_F^2) + \boxed{\sum_{u,v} W_p^{(u,v)} \| p^{(u)} - p^{(v)} \|^2}$$

$$+ \boxed{\sum_{i,j} W_q^{(i,j)} \| q^{(i)} - q^{(j)} \|^2}$$

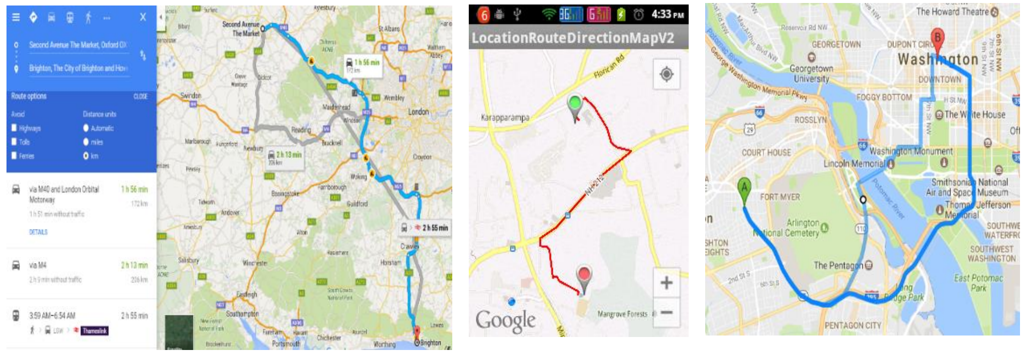
Spatial Regularization for Items

Z. Lu, D. Agarwal, and I. Dhilllon. "*A Spatiotemporal Approach to Collaborative Filtering*". In **RecSys 2009**

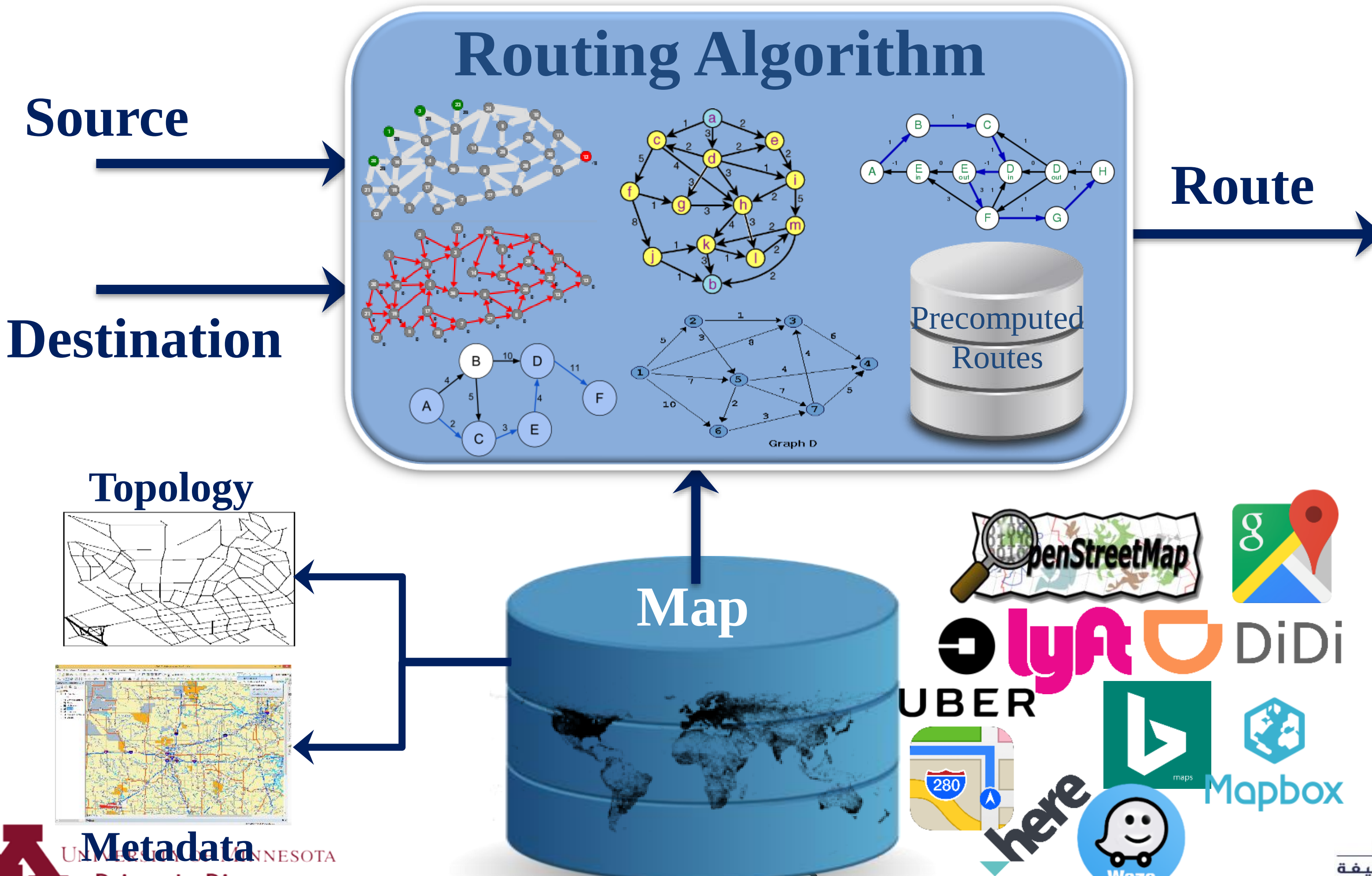
Machine Learning meets Big Spatial Data



Routing..



LocationRouteDirectionMapV2
Google
Washington



Routing Algorithm
Source
Destination
Route
Precomputed Routes
Topology
Metadata
Map
Graph D
OpenStreetMap
UBER
DiDi
Mapbox
here
Waze

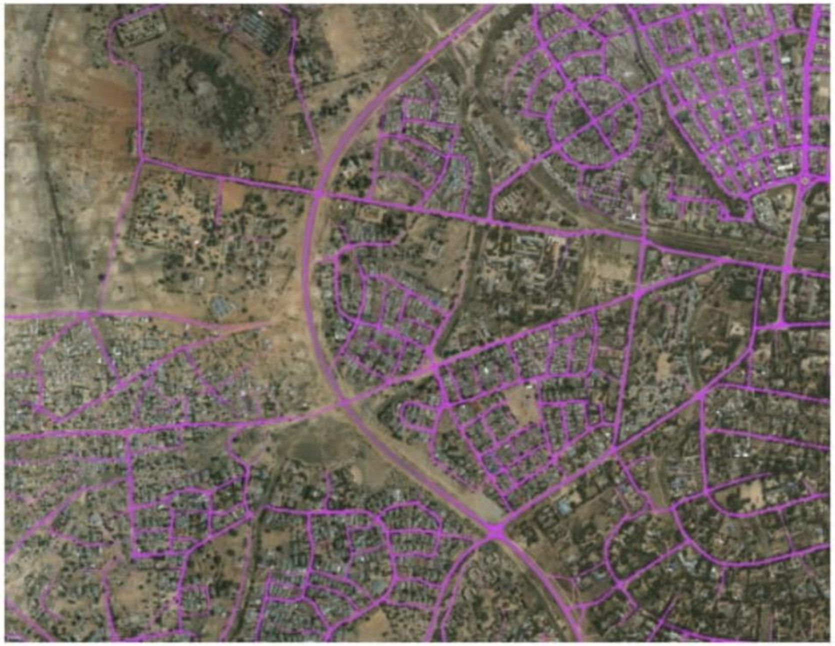
ML for Map Making (Topology)

- Automatic construction of road maps from images
  - Incremental route building (point by point)
  - Using Convolution Neural Networks (CNN) to search next point

F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden and D. DeWitt. "*RoadTracer: Automatic Extraction of Road Networks from Aerial Images*". In **CVPR 2018**

ML for Map Making (Topology)

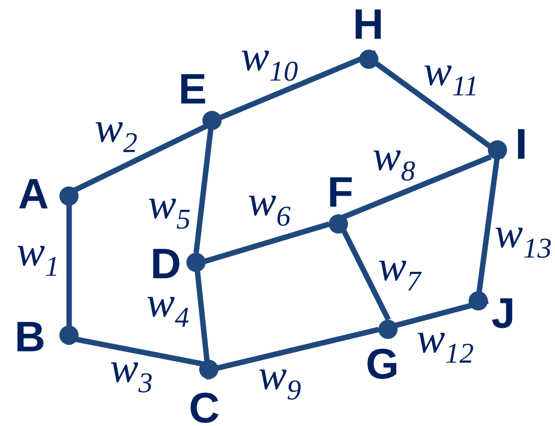
- Facebook AI provides “MapWithAI” to improve open-source mapping (e.g., OpenStreetMaps)
  - Weakly supervised learning from satellite Images using CNN
  - Apps: FB Marketplace, FB Local, and disaster response service



https://mapwith.ai/

ML for Map Making (Metadata)

- Learning Edge weights per time granularity (e.g., hour)
- **Input:** Trips (Pickup time/location, Drop off time/location, [Optional] Path)



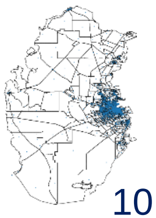
(A, F, 15) ➔ $w_2 + w_5 + w_6 = 15$

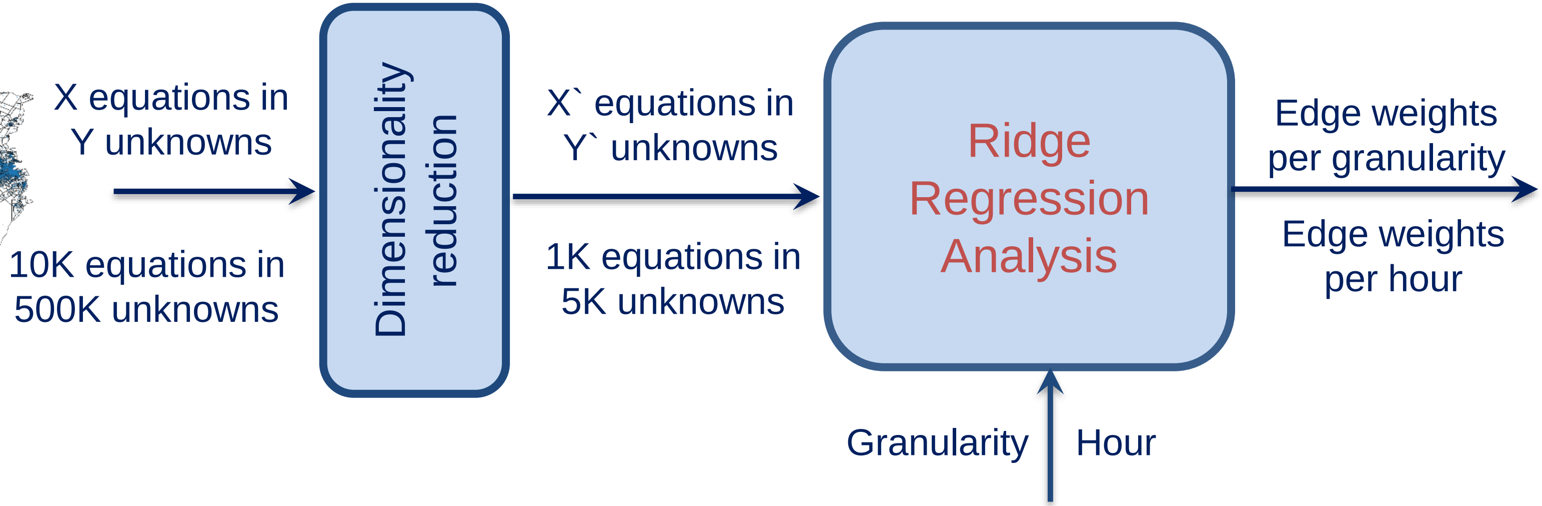
(B, H, 28) ➔ $w_3 + w_7 + w_8 + w_9 + w_{11} = 28$

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

…

…

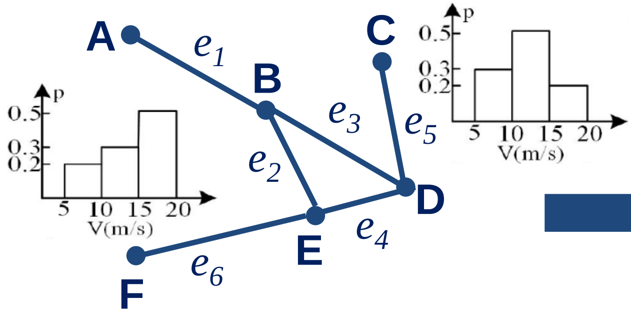




R. Stanojevic, S. Abbar, M. Mokbel. "W-edge: Weighing the Edges of the Road Network". In **ACM SIGSPATIAL 2018**

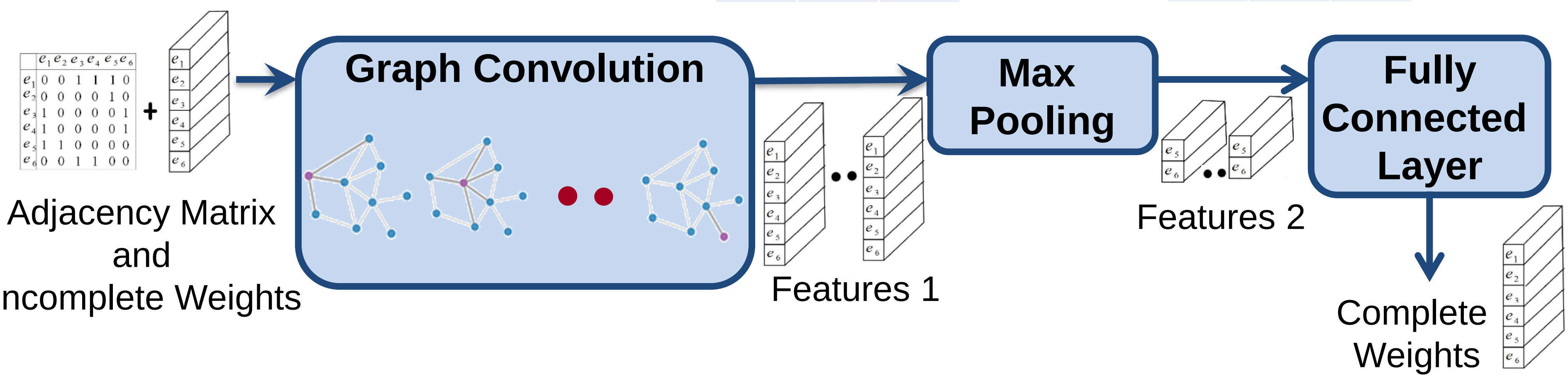
ML for Map Making (Metadata)

- Input: Speed distribution for certain time granularity



	[5, 10)	[10, 15)	[15, 20)
e_1	?	?	?
e_2	?	?	?
e_3	?	?	?
e_4	?	?	?
e_5	0.3	0.5	0.2
e_6	0.2	0.3	0.5

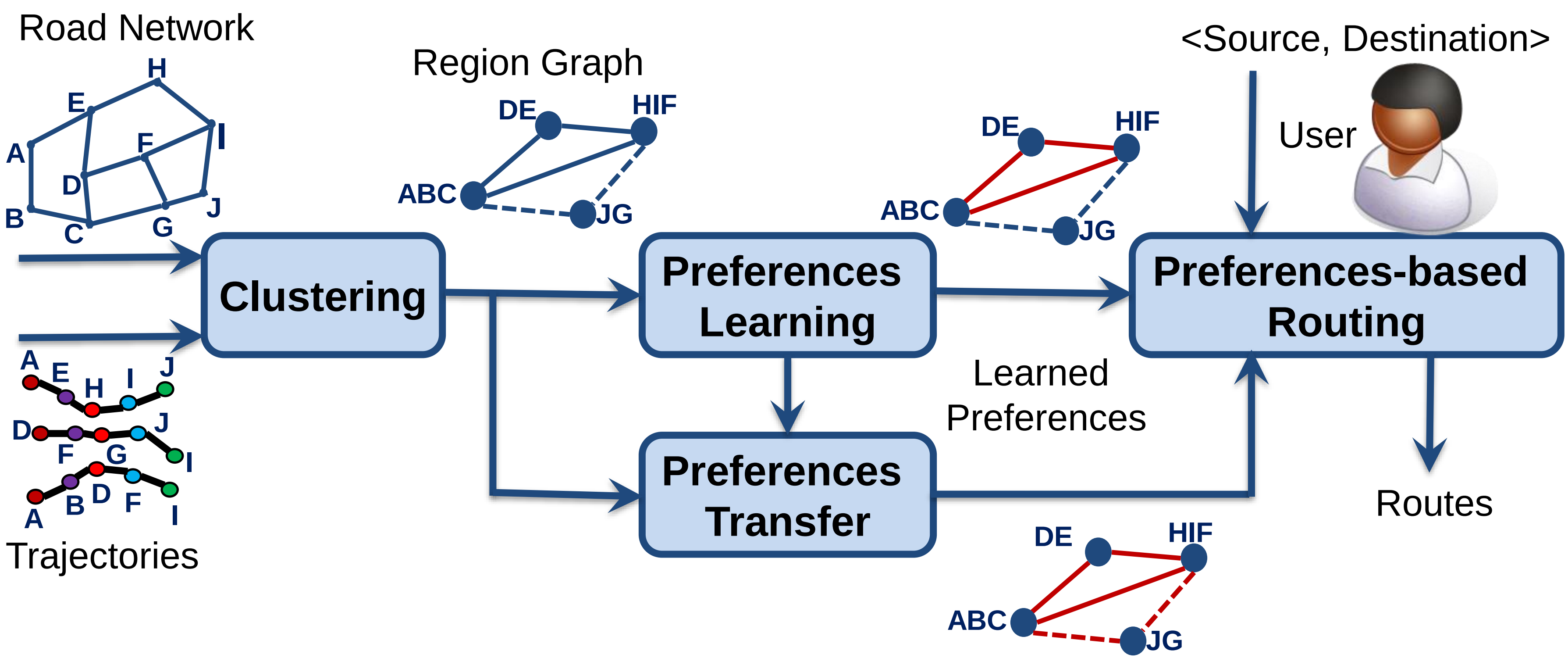
	[5, 10)	[10, 15)	[15, 20)
e_1	0.4	0.5	0.1
e_2	0.1	0.7	0.2
e_3	0.4	0.2	0.4
e_4	0.2	0.8	0
e_5	0.3	0.5	0.2
e_6	0.2	0.3	0.5



J. Hu, C. Guo, B. Yang, and C. S. Jensen. "*Stochastic Weight Completion for Road Networks using Graph Convolutional Networks*". In **ICDE 2019**

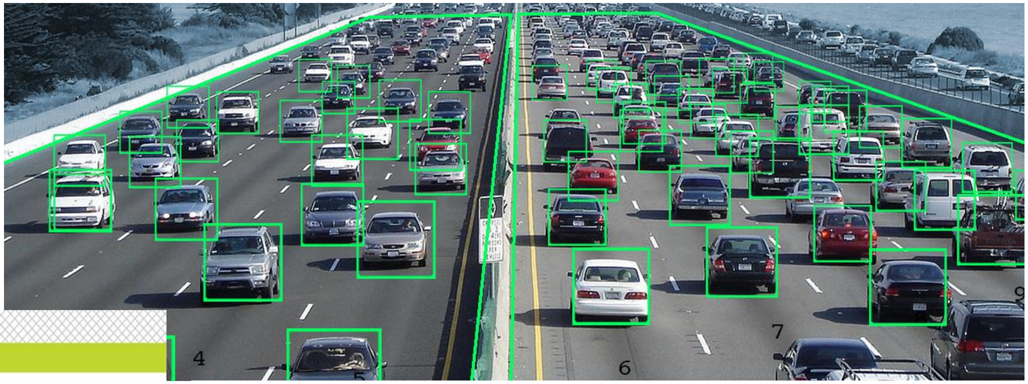
ML for Routing

- Using available trajectories to learn a better routing
  - A good route is determined by different preferences other than distance (e.g., road condition)
  - Similar trajectories can have similar preferences



C. Guo, B. Yang, J. Hu, C. S. Jensen. "*Learning to Route with Sparse Trajectory Sets*". In **ICDE 2018**

Traffic Monitoring & Prediction



Technology Profile

Real-time traffic management or short-term prediction?

traffic TECHNOLOGY INTERNATIONAL

Short-term prediction, decision support systems and predictive modeling: all familiar concepts to experts in transportation, but given the many real-time adaptive traffic management systems, do we really need them?

When it comes to adaptive traffic management systems, the UK market is the world's most saturated. Most cities and towns face high levels of congestion on a daily basis. A well-configured SCOOT system can handle many of the day-to-day challenges facing such networks.

The best investment an authority can make to manage its network and maximize its physical assets through technology is an adaptive system, be it a highway ATMS (advanced traffic management system) or a city-wide adaptive solution. All carry a high cost/benefit ratio, though they are generally restricted to dealing with immediate traffic situations.

In the UK, the Transport Technology Forum suggests that transport congestion, safety and emissions add up to a £100bn a year (US$130bn) problem.

We are all familiar with the frustrations of the daily commute. The problem is magnified when there is an unplanned incident in the network, whether a lane closure, emergency road works, an accident, or even an ITS failure (it does happen). With air-quality policies starting to influence cities' traffic management strategies, congestion rising, and capacity in the network remaining largely the same, timely intervention could well be the measure of success by which road users judge operators.

Predicting the future

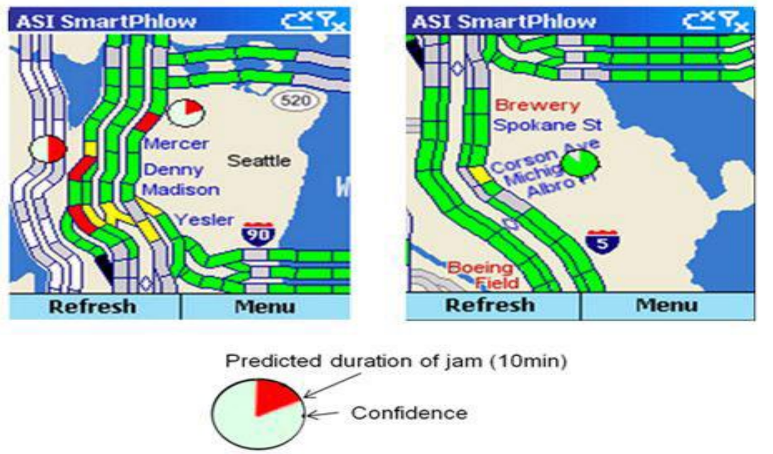
This is where predictive decision support systems (DSS) come in, working together and enhancing these real-time systems by looking past the current situation and assessing, analyzing and predicting the

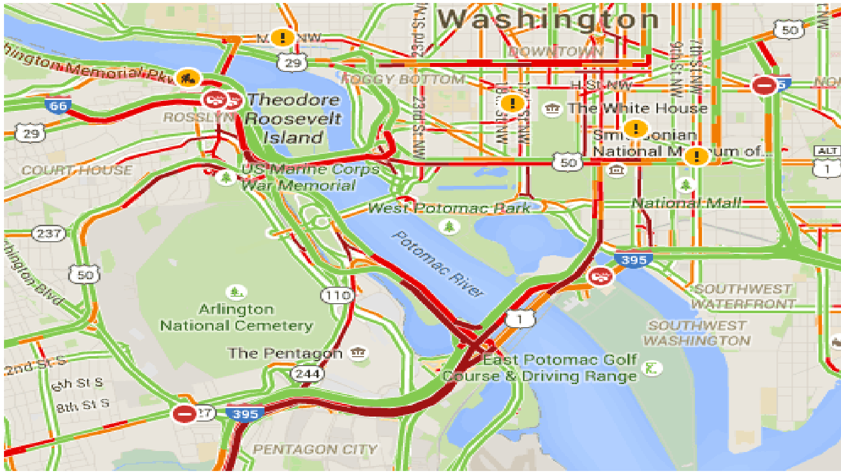
Above: **Aimsun Live helps with earlier communication of incident information**

Right: **The Aimsun Live user control panel features live feeds and simulations**

as traffic signals, ramp meters and message signs.

At the heart of the DSS is the Aimsun Live modeling package, configured and integrated into the system by the SANDAG

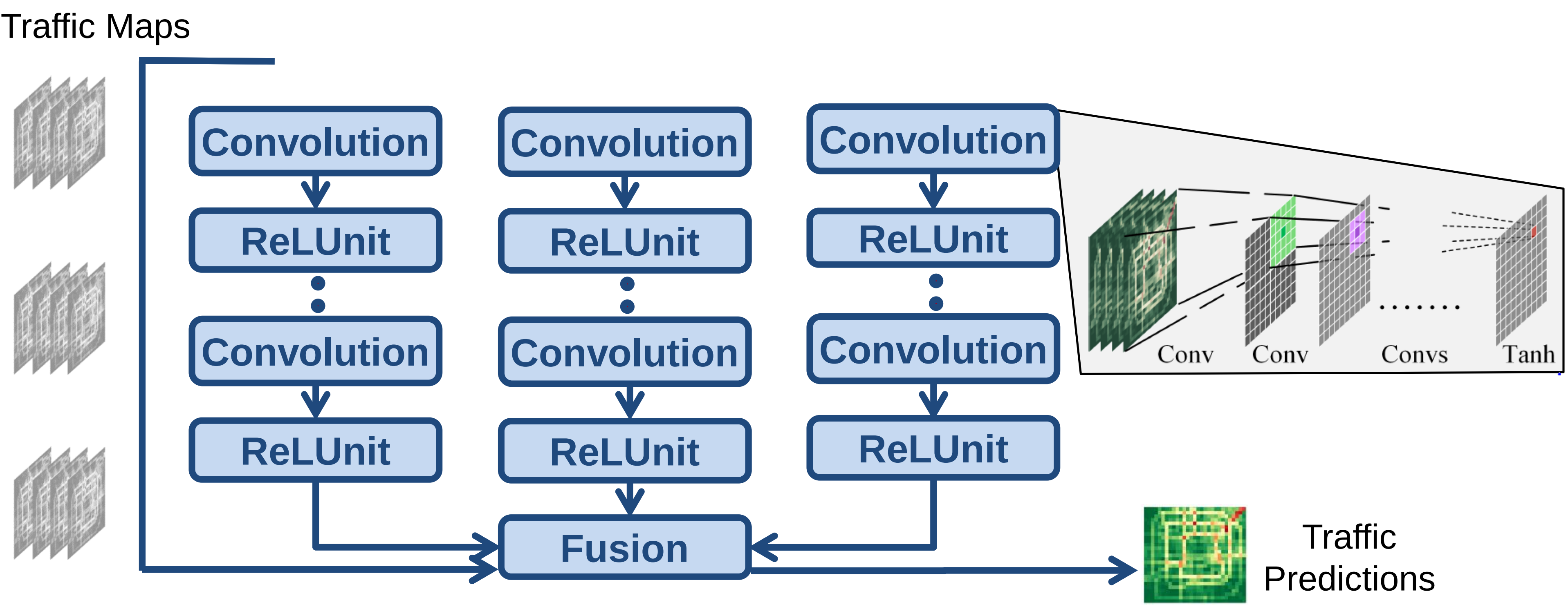




Need Long-term Traffic Prediction.!!?

ML for Traffic Prediction: Residual Networks

- Using convolution-based residual networks to handle both *spatial* and *temporal* dependencies
  - Inputs are divided into time spans, then each span is processed with a residual network, and finally all outputs are fused together



J. Zhang, Y. Zheng, and D. Qi. "*Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction*". In **AAAI 2017**

ML for Traffic Prediction: DCRNN

- Modeling the traffic prediction as a graph problem
  - Traffic sensors are nodes, and edge weights denote spatial proximity among these nodes → capturing spatial correlation
- Employing Diffusion Convolutional RNN (DCRNN)
  - Diffusion processes for *inflow* and *outflow* traffic flows

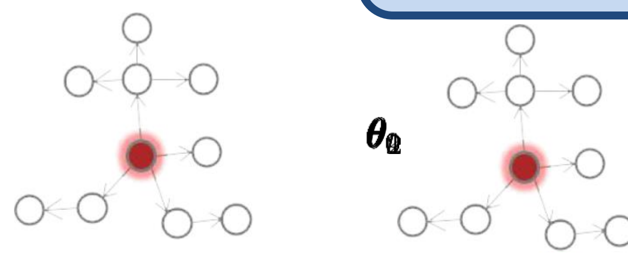
Traffic Sensors Data

Diffusion Graph Constructor

Traffic Sensors Graph

Diffusion CRNN Model

Prediction Outputs



Inflow Diffusion Process

Outflow Diffusion Process

$$X_{:,p} \star_{\mathcal{G}} f_{\boldsymbol{\theta}} = \sum_{k=0}^{K-1} \left(\theta_{k,1} \boxed{\left(D_O^{-1}W\right)^k} + \theta_{k,2} \boxed{\left(D_I^{-1}W^{\intercal}\right)^k}\right) X_{:,p}$$

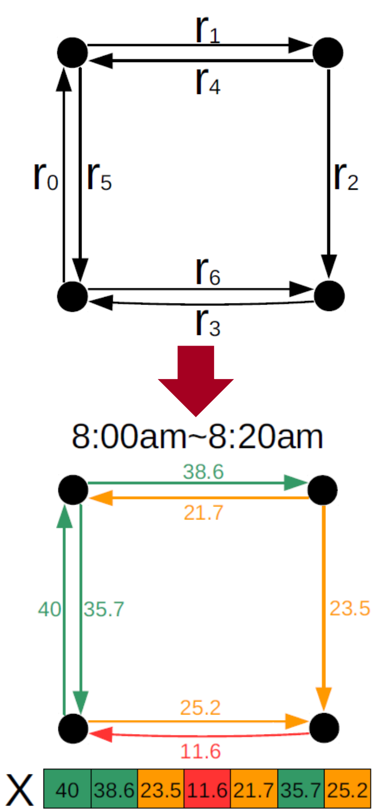
Graph Signal

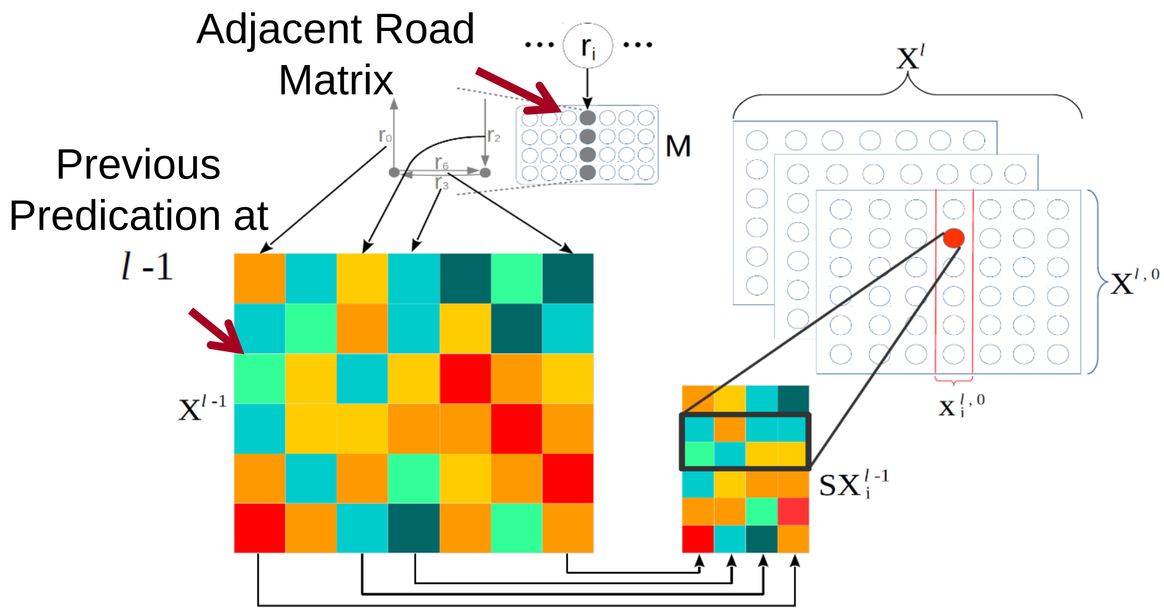
Filter

Y. Li, R. Yu, C. Shahabi, and Y. Liu. "*Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting*". In **ICLR 2018**

ML for Traffic Prediction: GCNN

- Using Graph Convolutional Neural Network (GCNN)
  - A novel road network embedded convolution method to learn meaningful spatial and speed features



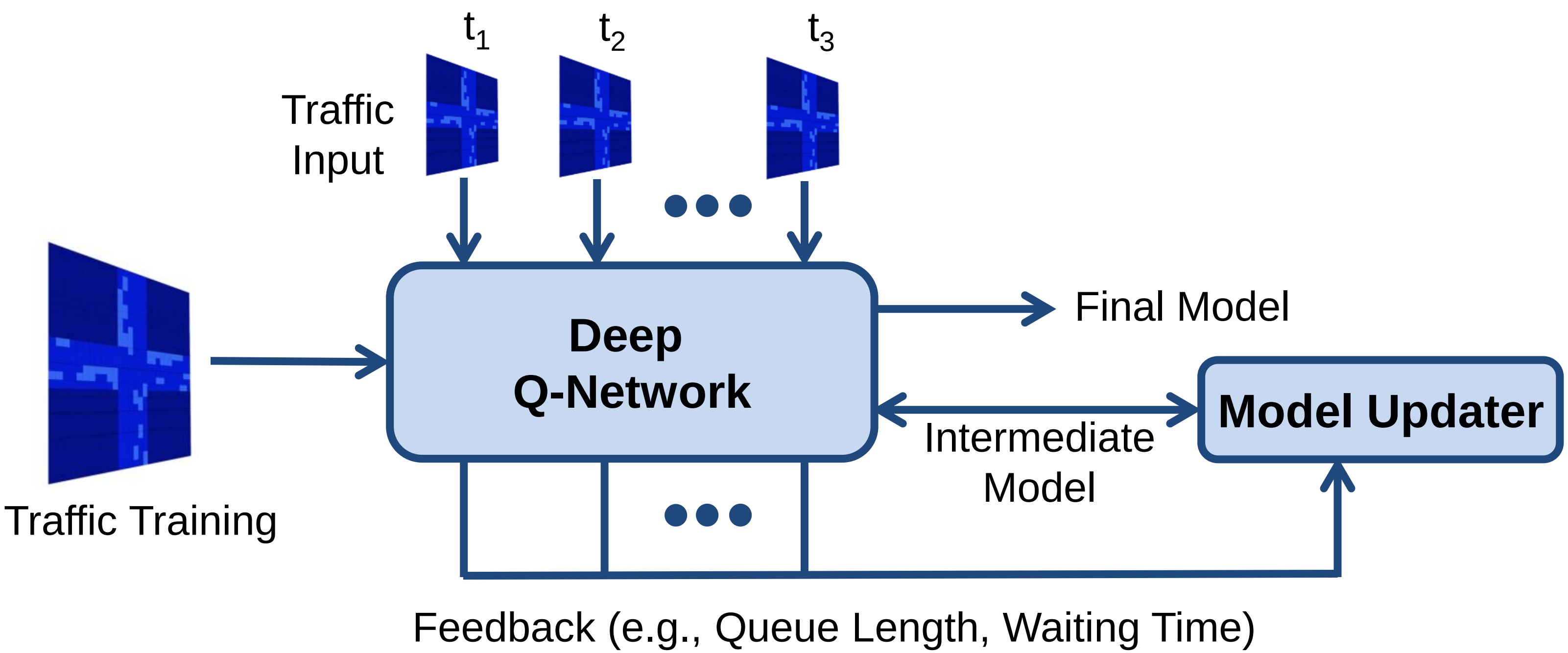


$$x_{j,i}^{l,k} = relu\left(\sum_{m=0}^{h}\sum_{n=0}^{A} w_{m,n}^{l,k}\, sx_{j+m,n}^{l-1} + b^{l,k}\right)$$

Z. Lv, J. Xu, K. Zheng, H. Yin, P. Zhao, and X. Zhao. "*LC-RNN: A Deep Learning Model for Traffic Speed Prediction".* In **IJCAI 2018**

ML for Traffic Prediction: Reinforcement Learning

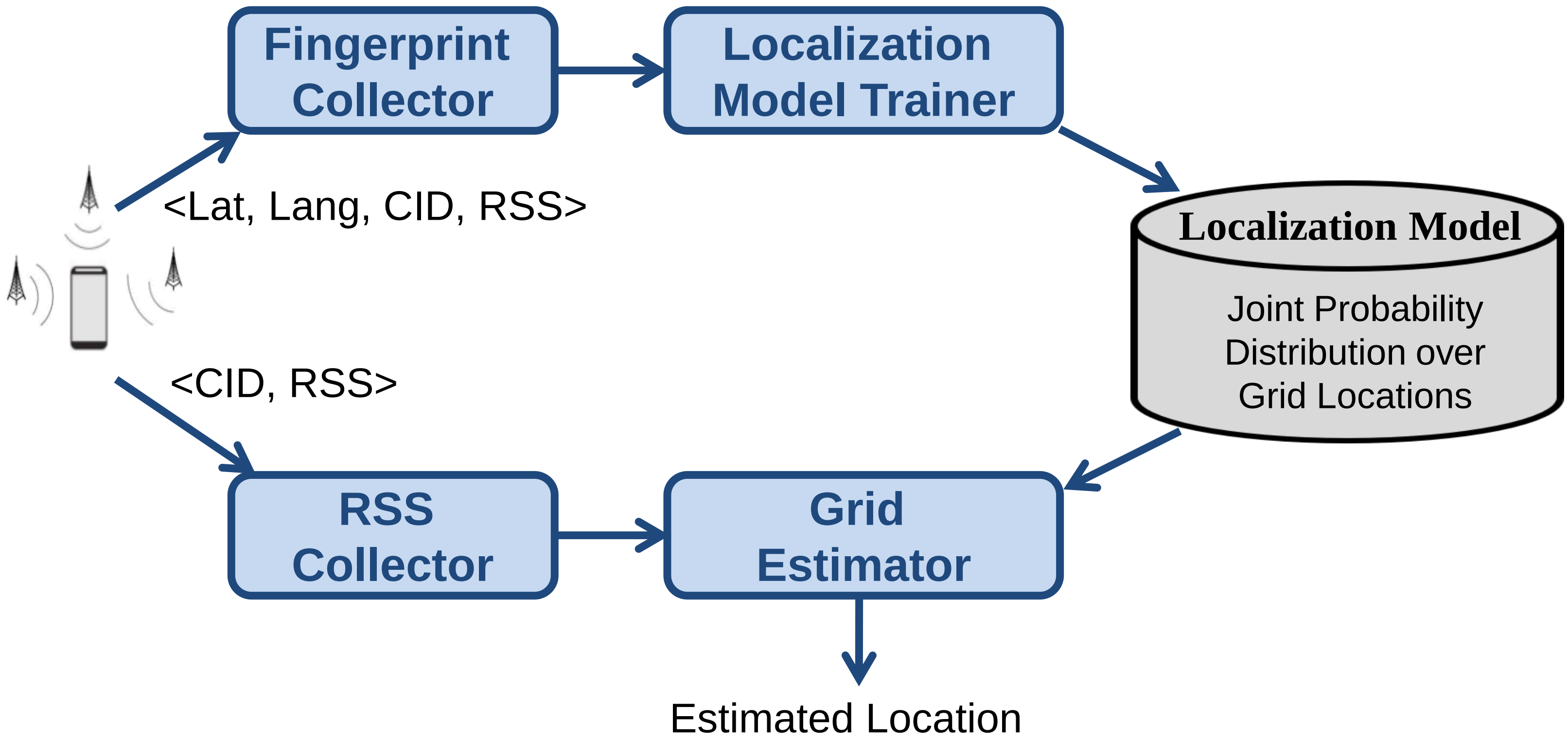
- Real-time Traffic Lights Control via Reinforcement Learning
  - Using non-spatial signals (e.g., waiting time) to update the model



H. Wei, G. Zheng, H. Yao, and Z. Li. "*IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control*". In **ACM SIGKDD 2018**

ML for Outdoor Localization

- Localizing people using their phones and without GPS
  - Fingerprinting using Received Signal Strength from cell towers
  - Having offline (training) and online (tracking) phases



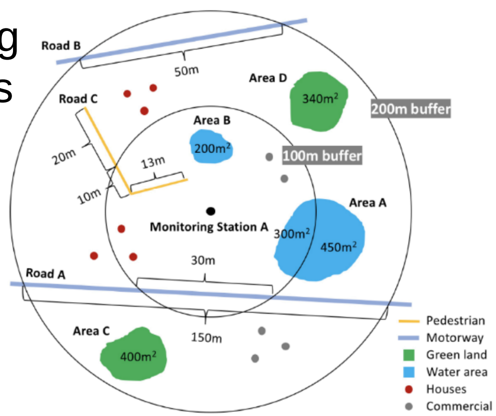
A. Shokry, M. Torki, M. Youssef. "*DeepLoc: A Ubiquitous Accurate and Low-overhead Outdoor Cellular Localization System*". In **ACM SIGSPATIAL 2018**

ML for Air Quality Forecasting

- Prediction at a certain location and within a time period
  - Traditional approach uses a hybrid regression model to *separately* deal with spatial and temporal correlations
  - Using joint spatiotemporal-aware CNN is more efficient
    - Defining more “important” geo-context features than distance

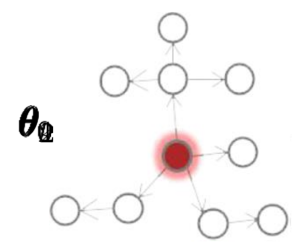
Air Monitoring Observations

→ **Geo-Context Features Extractor** → Feature Vectors → **Geo-Context Graph Constructor** → Graph → **Diffusion CRNN Model** → Prediction Output



	l_1	l_2	l_3	l_4
l_2	0.4			
l_3	0.1	0.4		
l_4	0.3	0.6	0.2	
l_5	0.5	0.3	0.2	0.1

Similarity-based Weights



$$X_{:,p} \star_G f_\theta = \sum_{k=0}^{K-1} \left(\theta_k (D^{-1}A)^k\right) X_{:,p}$$

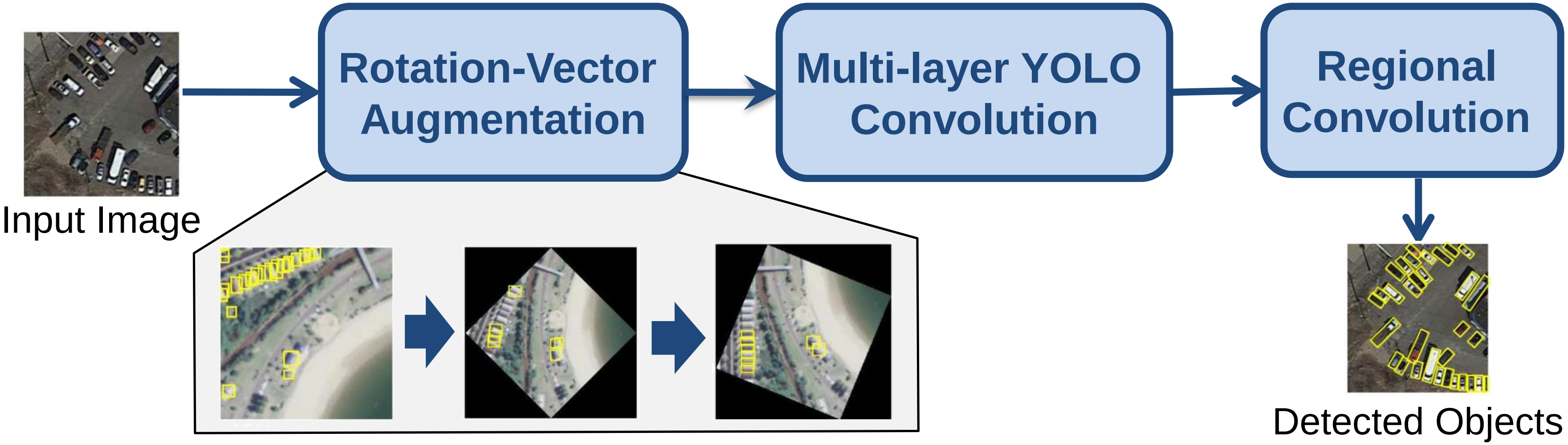
Graph Signal

Filter

Y. Lin, N. Mago, Y. Gao, Y. Li, Y. Chiang, C. Shahabi, and J. L. Ambite. “*Exploiting Spatiotemporal Patterns for Accurate Air Quality Forecasting Using Deep Learning*”. In **ACM SIGSPATIAL 2018**

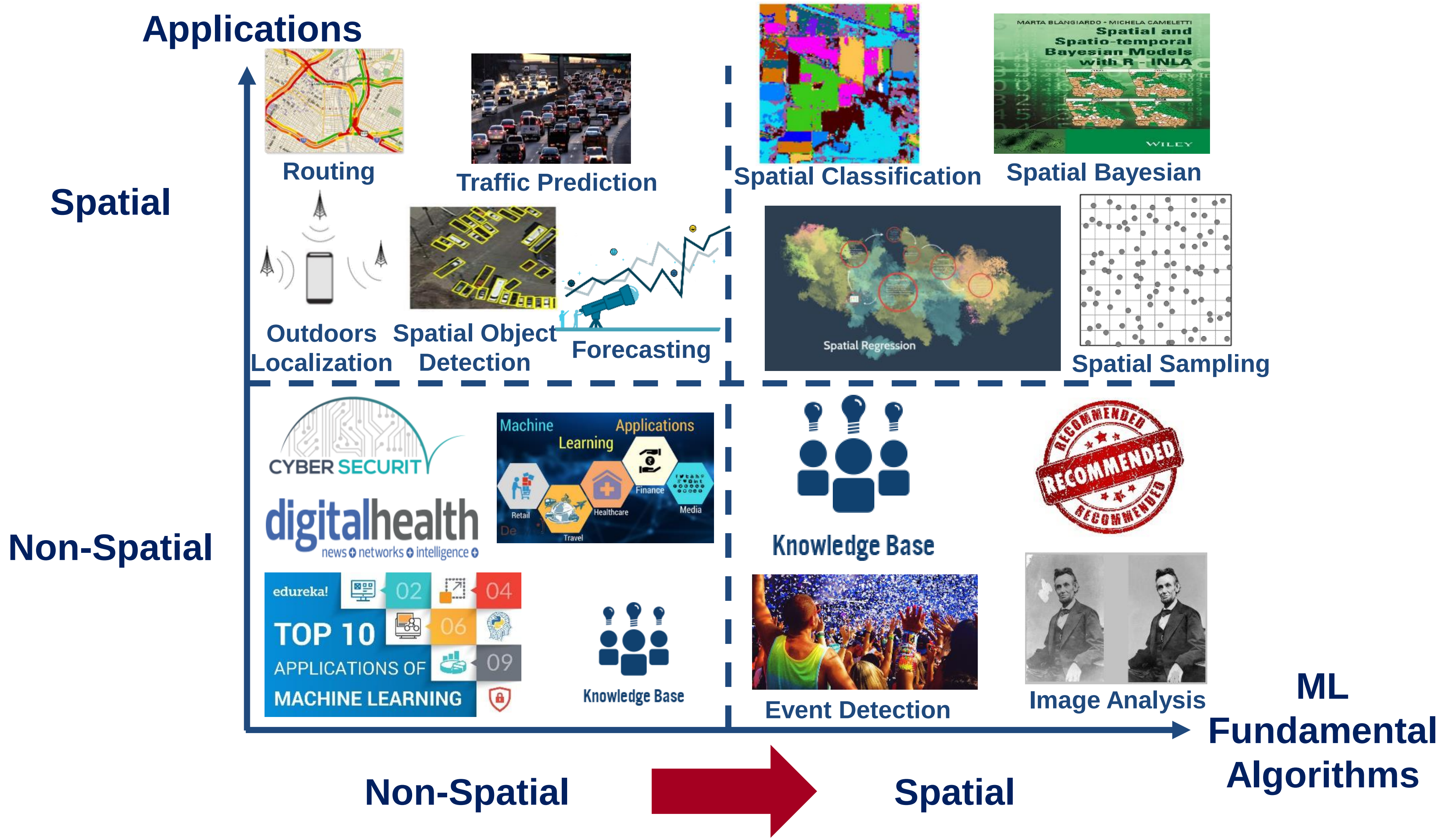
ML for Geospatial Object Detection

- Detecting geospatial objects (e.g., buildings) from images
  - Challenging as directions are not parallel to the orthogonal axes
  - Existing techniques detect the Minimum Orthogonal Bounding Rectangles (MOBR) of objects only (e.g., YOLO Framework)
- Main idea is to extract features from rotated images
  - No need for new training data with different rotations



Y. Xie, R. Bhojwani, S. Shekhar, and J. Knight. "*An Unsupervised Augmentation Framework for Deep Learning Based Geospatial Object Detection: A Summary of Results*". In **ACM SIGSPATIAL 2018**

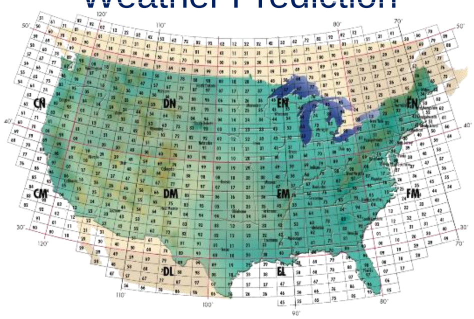
Machine Learning meets Big Spatial Data



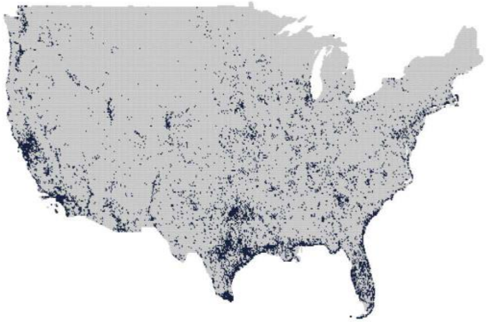
Spatial (Autologisitc) Regression

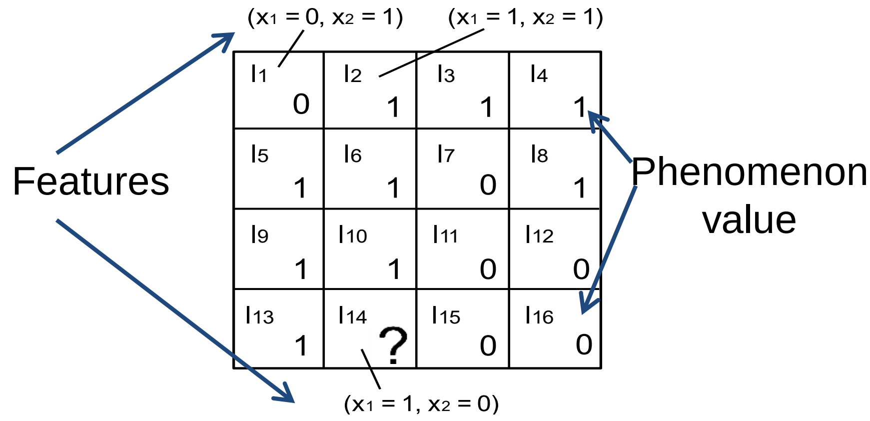
- Find whether a spatial phenomenon exists or not, based on neighbor values and features

Weather Prediction



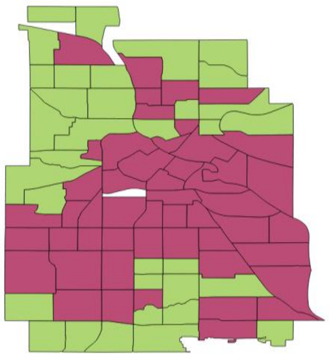
Birds Migration





Land Cover

Crimes Distribution



Missing value

Features

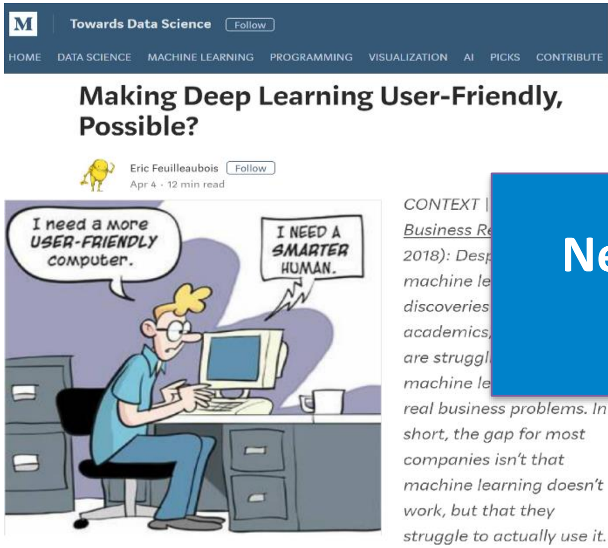
Neighbor values

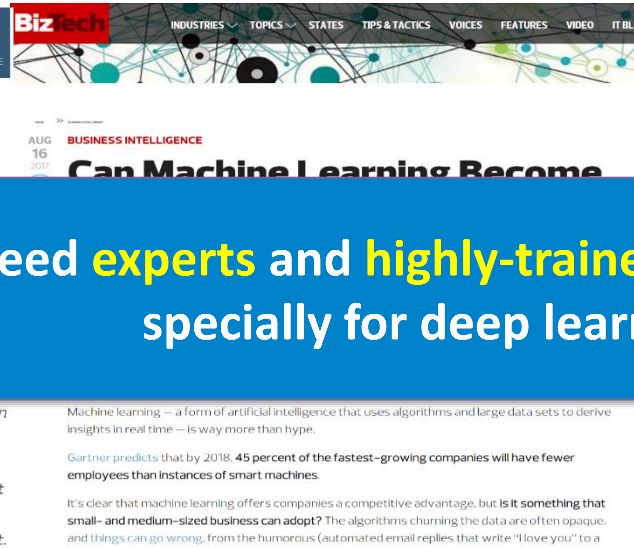
$$\log \frac{Pr(z_i=1|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})}{Pr(z_i=0|\mathcal{X},\mathcal{Z}_{\mathcal{N}_i})} = \sum_{j=1}^{m} \beta_j x_j + \eta \sum_{k\in\mathcal{N}_i} z_k$$

Regression Parameters

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

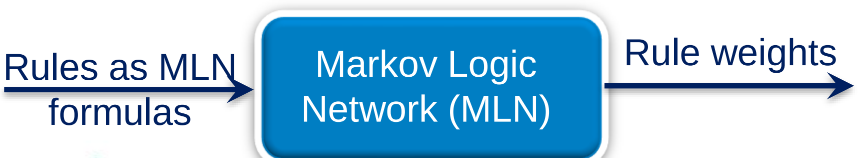
Markov Logic Networks (MLN)





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



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

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

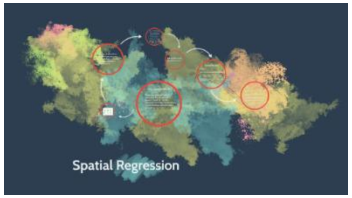
DeepDive

HoloClean

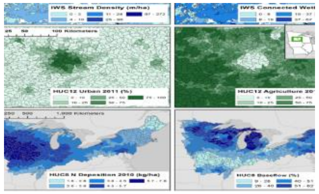
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

From MLN to Spatial MLN

Spatial Regression



Spatial Classification



Spatial Bayesian Networks

. . . . .

Language — **Spatial DDLog**

SMLN Rules

Grounding — **Spatial Factor Graph**

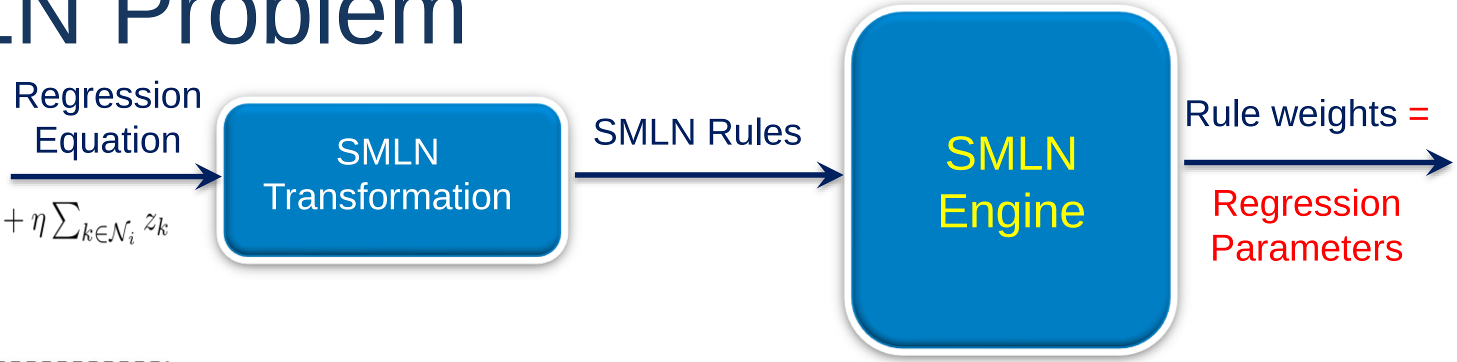
Constructed Graph

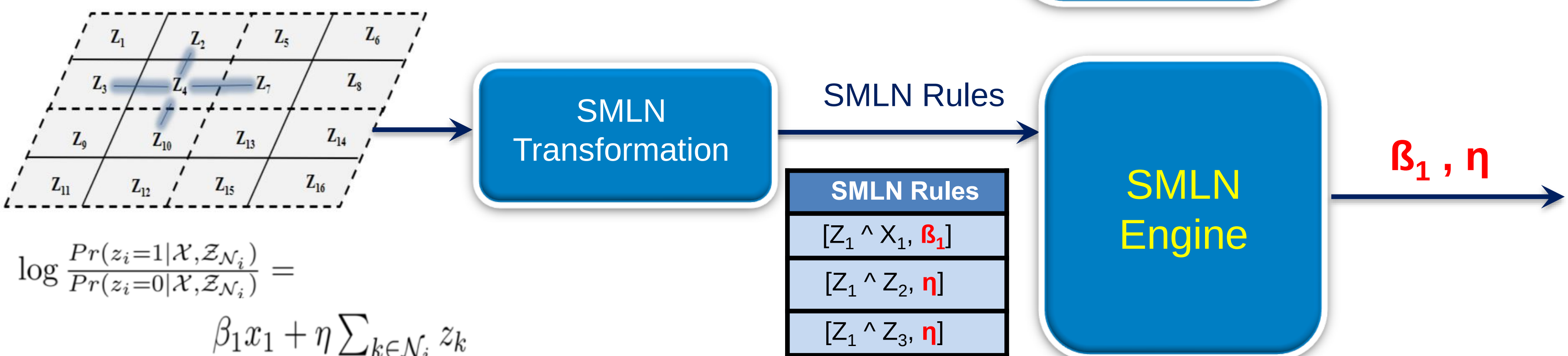
Inference — **Spatial Gibbs Sampling**

Spatial MLN

Spatial Regression as SMLN Problem

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





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

SMLN Rules
[Z_1 ^ X_1, **$ß_1$**]
[Z_1 ^ Z_2, **η**]
[Z_1 ^ Z_3, **η**]
[Z_2 ^ X_1, **$ß_1$**]
[Z_2 ^ Z_4, **η**]
[Z_2 ^ Z_5, **η**]
[Z_3 ^ X_1, **$ß_1$**]
[Z_3 ^ Z_4, **η**]
…….

I. Sabek, M. Musleh, and M. F. Mokbel. "*TurboReg: A Framework for Scaling Up Spatial Logistic Regression Models*". In **SIGSPATIAL 2018**

Bayesian Modeling

- Analyzing spatial data for prediction, estimating parameters, and capturing correlations
  - Traditional assumption is Gaussian processes
  - Estimating parameters is a bottleneck in case of big data

$$y = \boxed{\beta} X + \epsilon \longrightarrow \sim Normal(0, \boxed{\sigma^2} \rho)$$

Nx1 Vector of Outcomes

px1 Vector of Slopes

Nxp Matrix of Features

NxN Covariance Matrix

- Using Bayesian inference, the joint posterior distribution can be estimated in a closed form

$$p(\beta, \sigma^2 | y) = p(\sigma^2 | y) p(\beta | \sigma^2, y)$$

$$\sim InverseGamma(\sigma^2, c) \qquad \sim Normal(\beta | Mm, \sigma^2 M)$$

Need to calculate c, M, m efficiently on a large scale

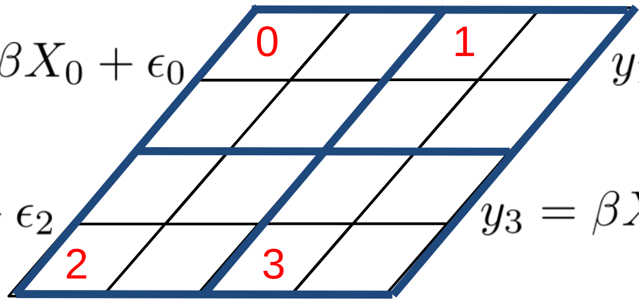
Quadtree-based Bayesian Modeling

- Main Idea: exploiting likelihood decomposition
  - Replacing the joint posterior distribution with a composite one that assumes independence across partitions

$$y = \beta X + \epsilon$$

$m_0, M_0 \leftarrow y_0 = \beta X_0 + \epsilon_0$

$y_1 = \beta X_1 + \epsilon_1 \rightarrow m_1, M_1$



$m_2, M_2 \leftarrow y_2 = \beta X_2 + \epsilon_2$

$y_3 = \beta X_3 + \epsilon_3 \rightarrow m_3, M_3$

$$m = \sum_{k=0}^{K-1} (m_k - (1 - \tfrac{1}{K}) C_\beta)$$

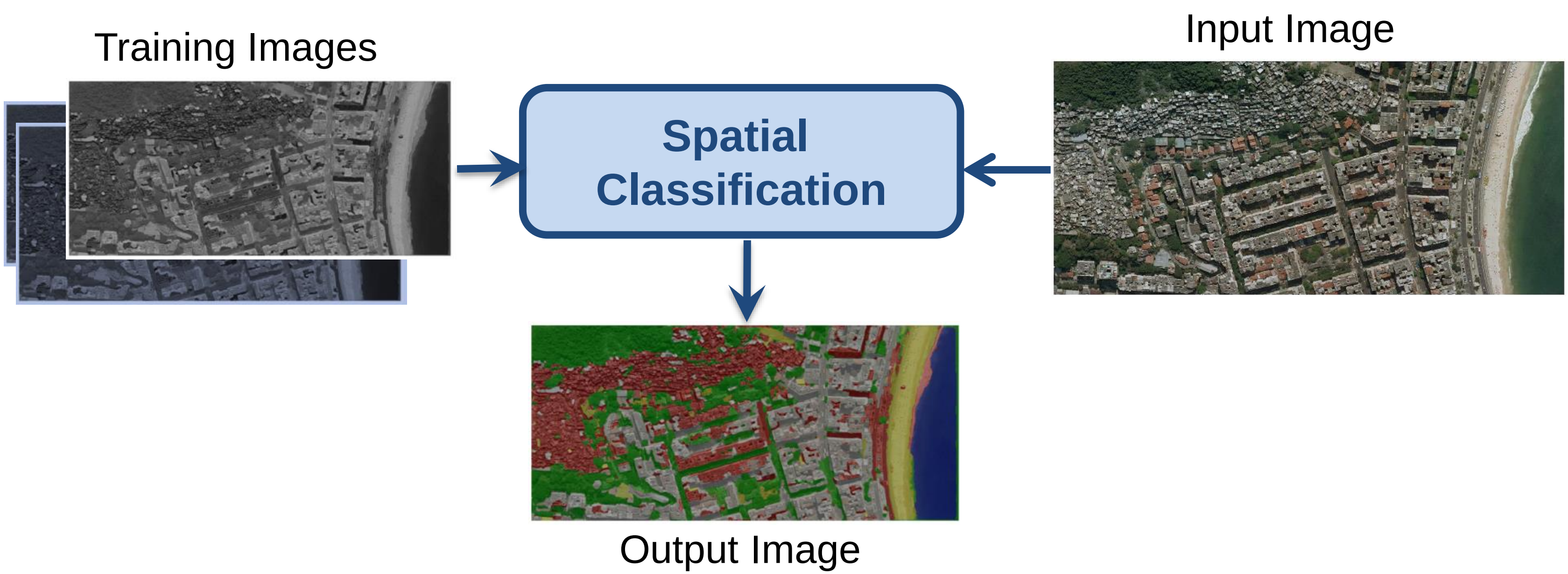
$$M^{-1} = \sum_{k=0}^{K-1} (M_k^{-1} - (1 - \tfrac{1}{K}) H_\beta)$$

Calculate C based on *m* and *M* using a closed form

R. Guhaniyogi, and S. Banerjee. "*Meta-Kriging: Scalable Bayesian Modeling and Inference for Massive Spatial Datasets".* In **Journal of Technometric 2018**

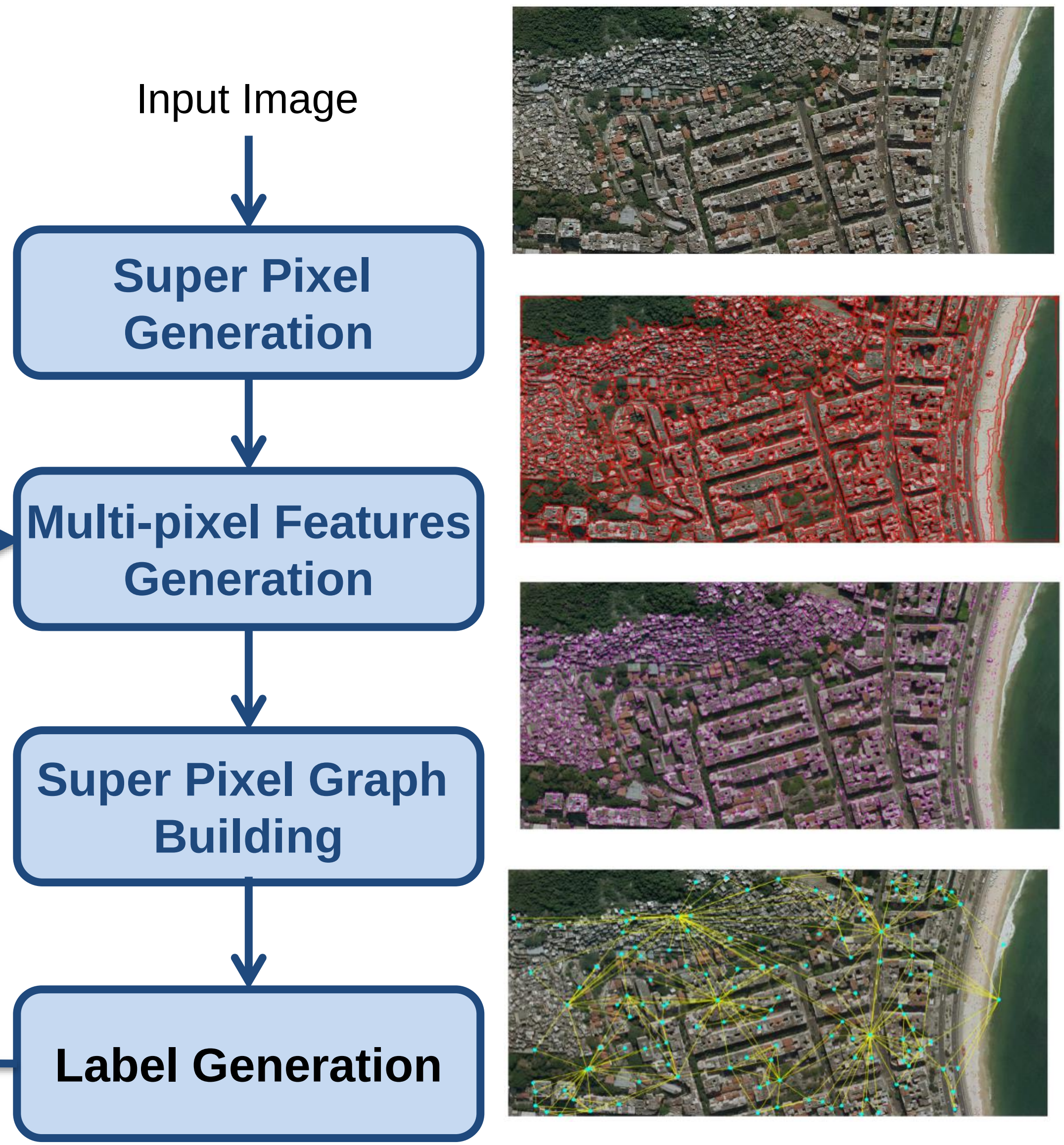
Spatial Classification

- Input
  - Training images labeled with pre-defined spatial classes
  - Unknown image
- Output
  - The same input image, yet, labeled with one or more of the spatial classes

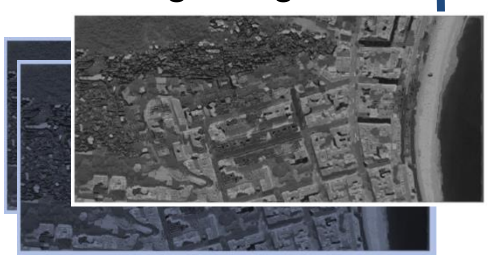


Scalable ML-based Spatial Classification

- Existing ML approaches are not scalable for high resolution images
- Employing the concept of *super pixel* to simplify computation



Training Images

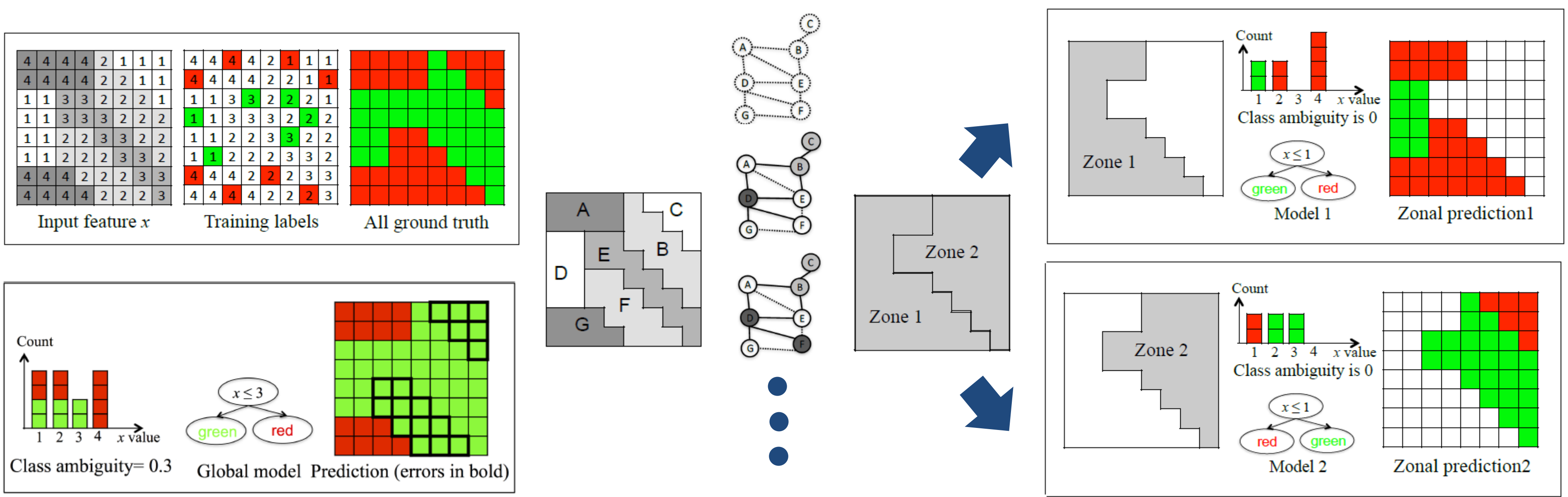


Output Image

M. Sethi, Y. Yan, A. Rangarajan, R. R. Vatsavai, and S. Ranka. "*Scalable Machine Learning Approaches for Neighborhood Classification Using Very High Resolution Remote Sensing Imagery*". In **SIGKDD 2015**

Ensemble Learning Spatial Classification

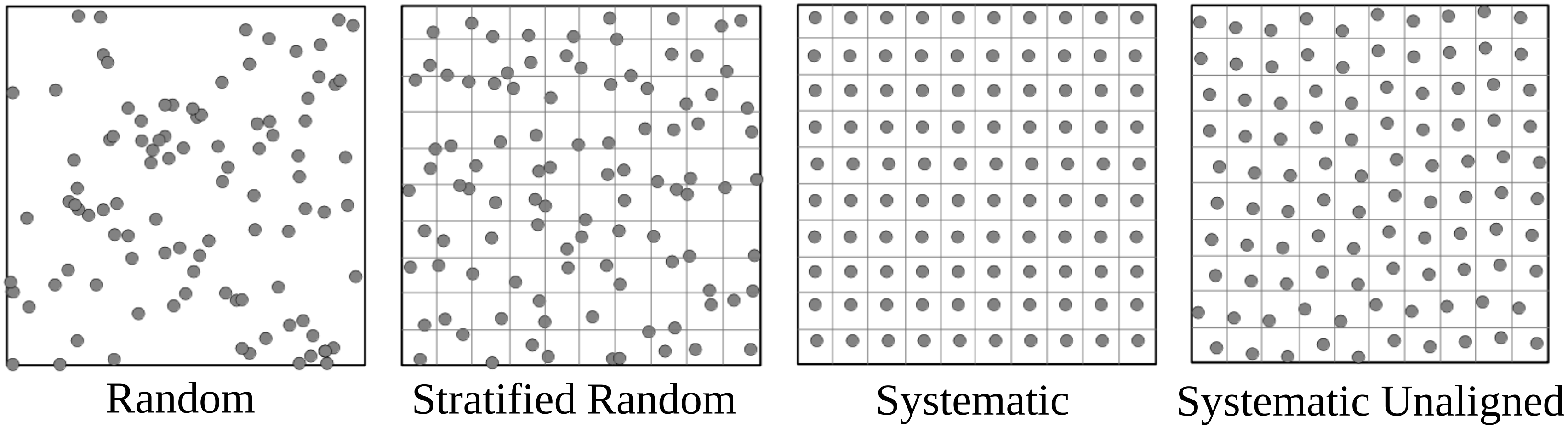
- Efficient classification over heterogonous spatial data
  - Class ambiguity: same feature values belong to different classes in different locations
- Learn ensembles on spatial neighborhoods in parallel
  - Global models have higher error rates and are much slower



Z. Jiang, Y. Li, S. Shekhar, L. Rampi, and J. Knight. "*Spatial Ensemble Learning for Heterogeneous Geographic Data with Class Ambiguity: A Summary of Results*". In **SIGSPATIAL 2017**

Spatial Sampling

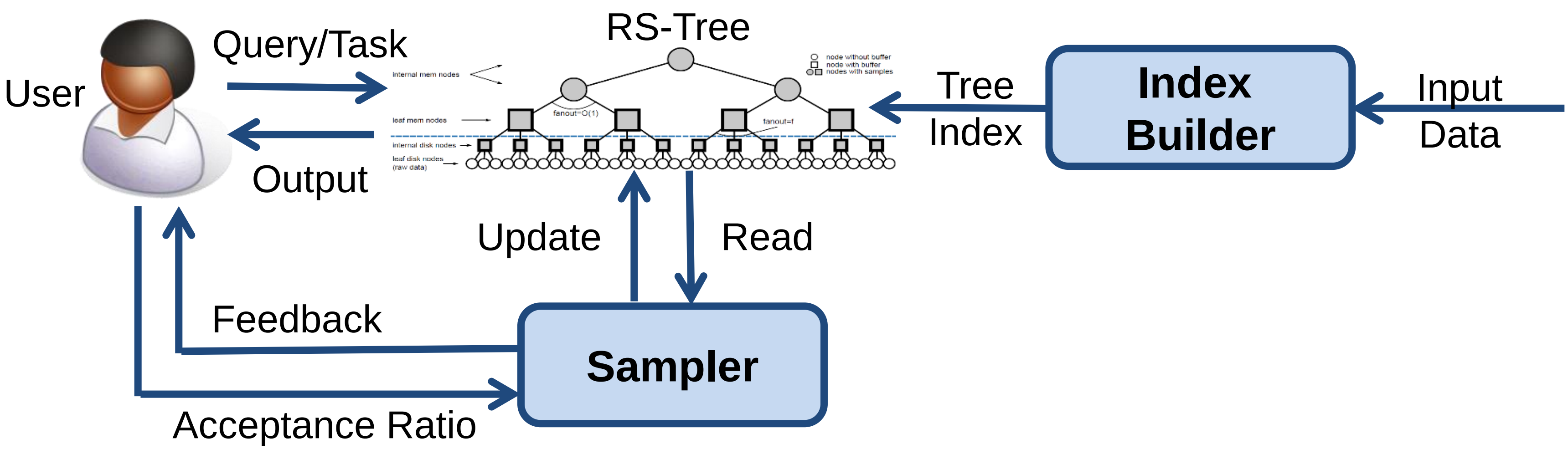
- Collecting representative samples in a 2-D framework
  - Could have a second-phase to reduce errors in initial samples



- Spatial sampling is more challenging with “big” data
  - Can be easily dragged to “biased” sampling
  - Exploiting ML to learn more “accurate” spatial samples

ML-based Incremental Spatial Sampling

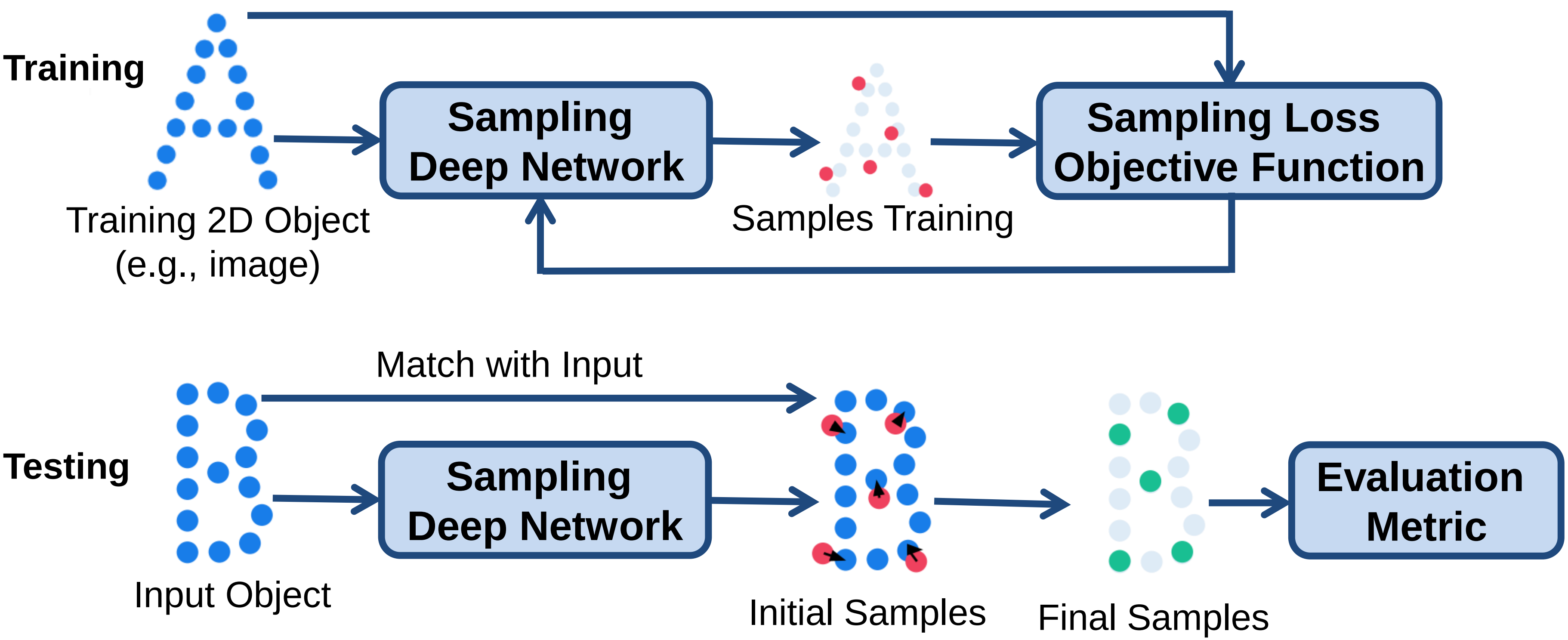
- No need to wait for the whole sampling to be done
  - Iterative sampling iterations with feedback from users
- The main idea is "level sampling"
  - Embedding samples into indexing structures (e.g., R-tree)
  - Lazy exploration for efficient processing
    - Visiting the children of any cell only after its sample buffer is exhausted (either consumed or rejected)



L. Wang, R. Christensen, F. Li, and K. Yi. "*Spatial Online Sampling and Aggregation".* In **VLDB 2015**

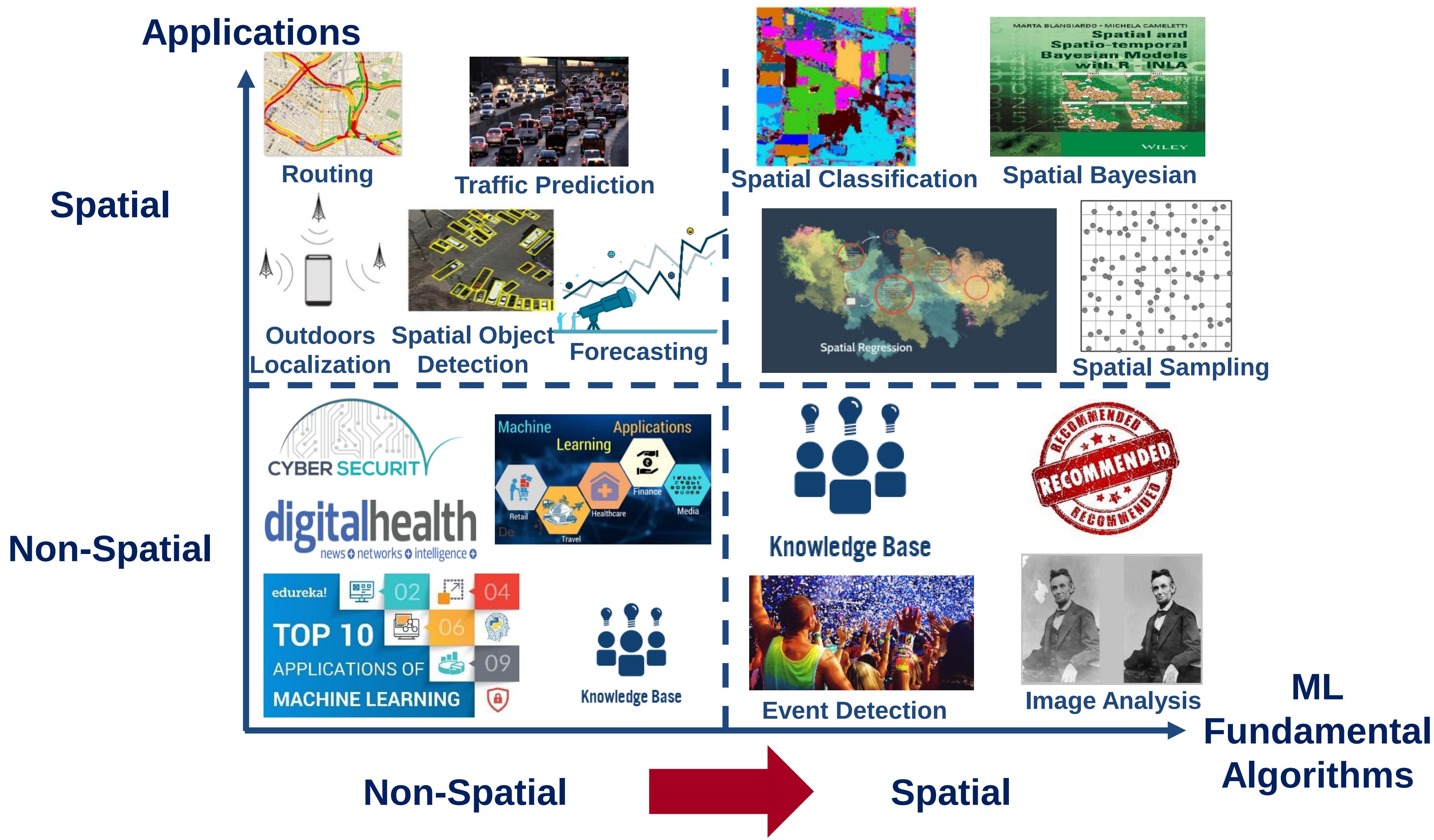
Deep Learning-based Spatial Sampling

- Exploiting deep learning to learn spatial samples
  - Training: a deep network is trained to preserve the original shape
  - Testing: generated samples are matched with the input to estimate the error for feedback



O. Dovrat, I. Lang, and S. Avidan. "*Learning to Sample*". In **CVPR 2019**

Machine Learning meets Big Spatial Data



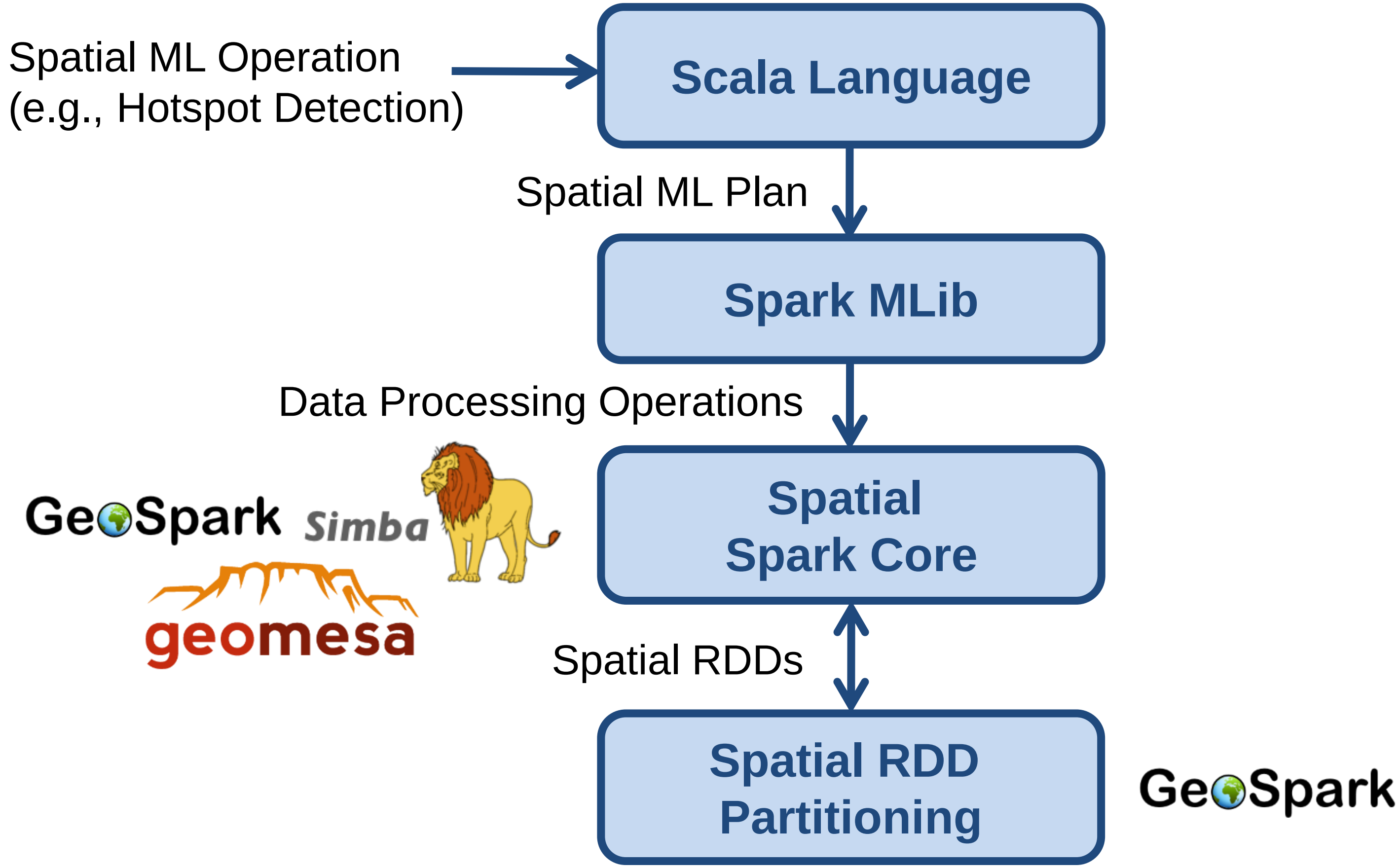
Outline

- Introduction
- Motivation
- Detailed Techniques
- End-to-End Systems

Spark-based Spatial ML Systems

APACHE Spark

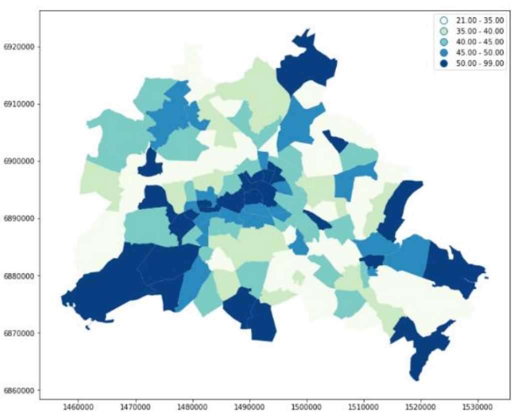
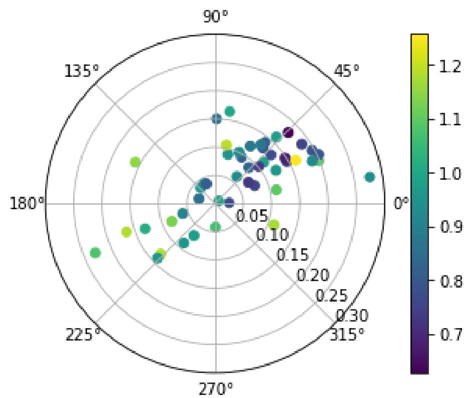
- Integrating Spark MLib with spatially-equipped spark core and RDD operations

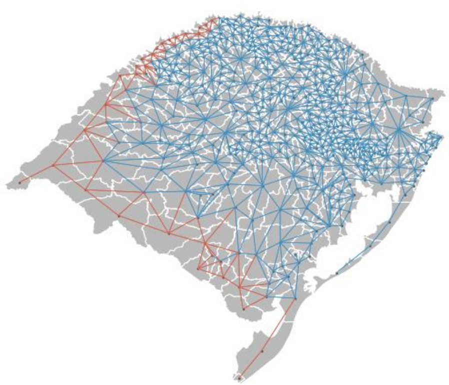


PySAL: Python Spatial Analysis Library

- Open source cross-platform library for geospatial data science on vector data
  - Spatial clusters, hot-spots, and outliers
  - Spatial regression and statistical modeling
  - Spatial econometrics
  - Exploratory spatio-temporal data analysis

Visualization
Spatial Analytic Functions
Spatial Modeling Functions
Numpy, Scipy
Python

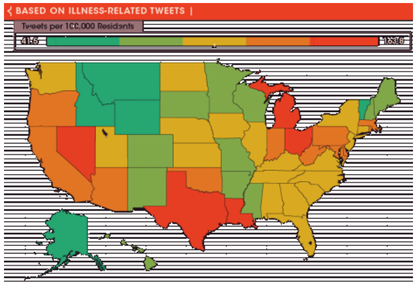


S. J. Rey, and L. Anselin. "*PySAL: A Python Library of Spatial Analytical Methods*". In **Review of Regional Studies** 37, 5-27 2007 https://pysal.org/

Flash: Scalable Spatial Data Analysis Using Markov Logic Networks

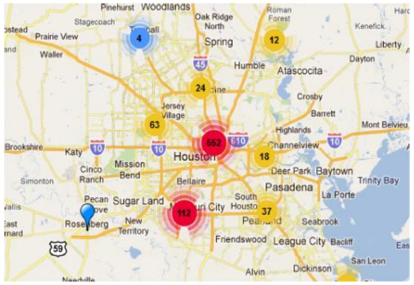
- Based on the design and deployment of Spatial Awareness in Probabilistic Graphical Models
  - Spatial Markov Random Fields (SMRF)
  - Spatial Hidden Markov Models (SHMM)
  - Spatial Bayesian Networks (SBN)

Public Health Monitoring



Geo-tagged Ads

Disaster Analysis



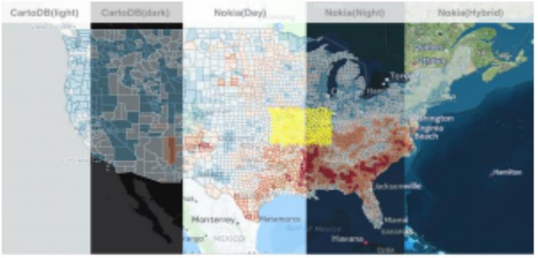
Crime Analysis

I. Sabek, M. Musleh, and M. F. Mokbel. "*Flash in Action: Scalable Spatial Data Analysis Using Markov Logic Networks*". In **VLDB 2019**

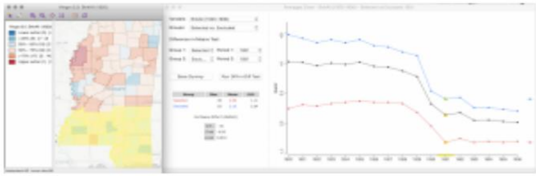
GeoDA: An Introduction to Spatial Data Analysis

- Exploratory spatial data analysis tool
  - Enriched visualization tools
  - Including spatial clustering, outliers detection, spatial regression

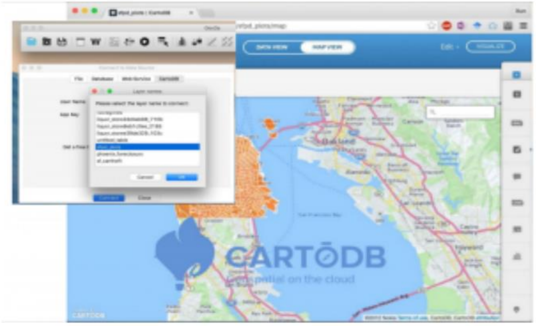
- Latest versions are open source (OpenGeoDa)
  - Cross-platform
  - Support cloud-based computation
    - Designed to support datasets with more than 170000 observations efficiently



Basemaps help contextualize the main map layer.



The Averages Chart aggregates trends across time and space.



The latest version of GeoDa integrates with CartoDB.

L. Anselin, I. Syabri, and Y. Kho. "*GeoDa: An Introduction to Spatial Data Analysis*". In **Journal of Geographical Analysis 2006** http://geodacenter.github.io/

ESRI ArcGIS GeoAI & GeoAnalytics

Sample Training Data

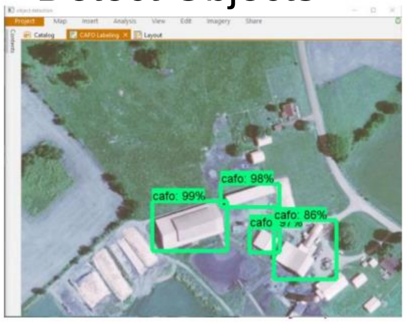
Add Imagery Source

Export Training Data

Train CNN

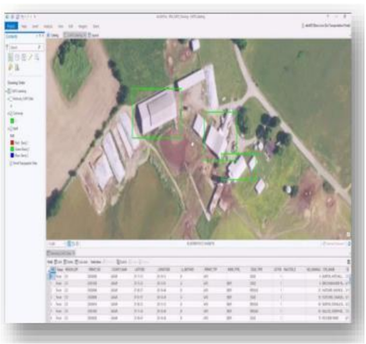
GeoAI tools integrated with Tensor flow for deep learning, classification, clustering, regression, etc.

Detect Objects



Call the model from

ArcGIS Pro



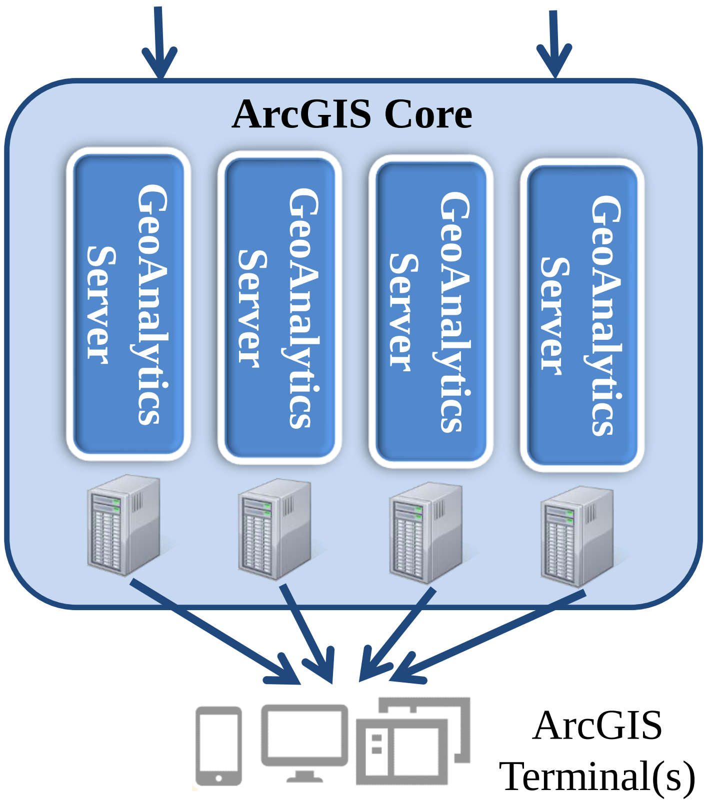
Python Function

```
134 lines (112 sloc) | 5.15 KB

import numpy as np
import tensorflow as tf

class Detect():

    def __init__(self):
        self.name = "Detect Object Function"
        self.description = ""

        self.modelPath = None
        self.clsLblPath = None
        self.threshold = 0.8
        # self.num_classes = 1
        #self.batch = 128

    def getParameterInfo(self):
        return [
```

GeoAnalytics Distributed Server for scalability

ML Operation

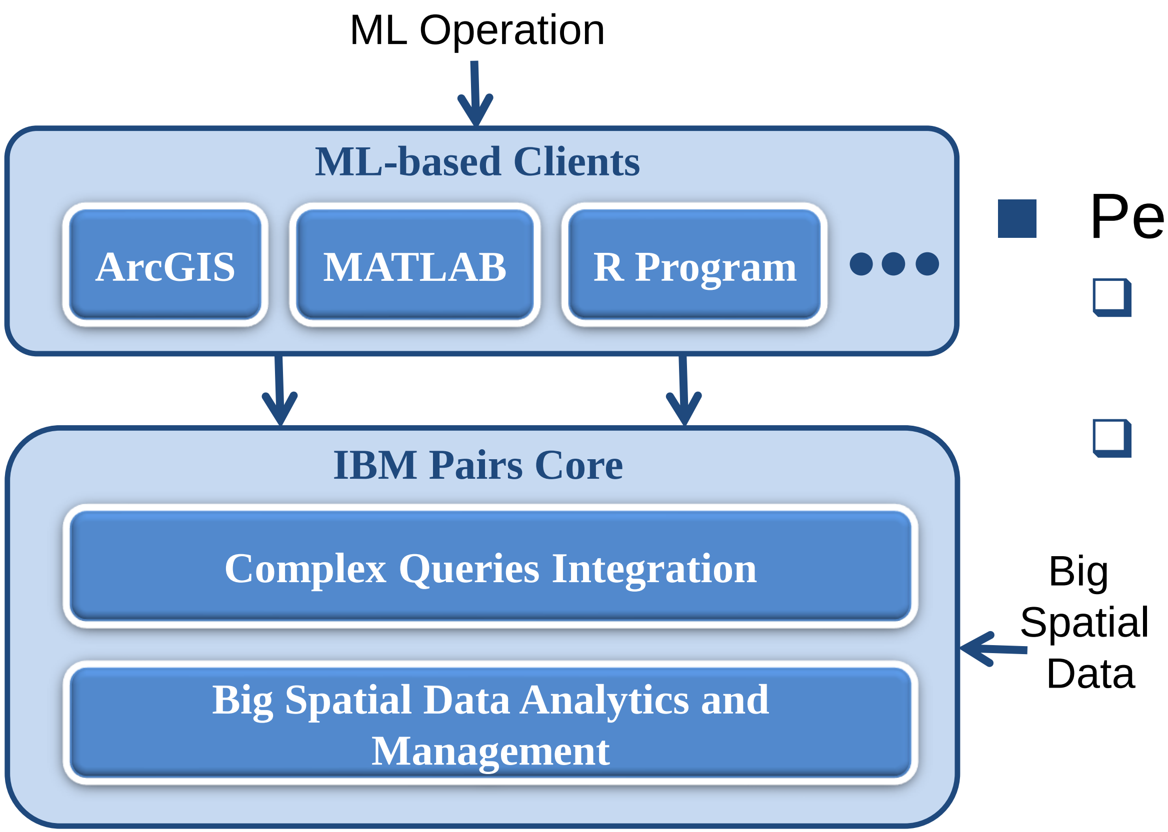
Big Spatial Data



esri.com/en-us/arcgis/products/arcgis-geoanalytics-server

IBM PAIRS GeoScope

- A cloud-based service for geospatial analytics and machine learning modeling

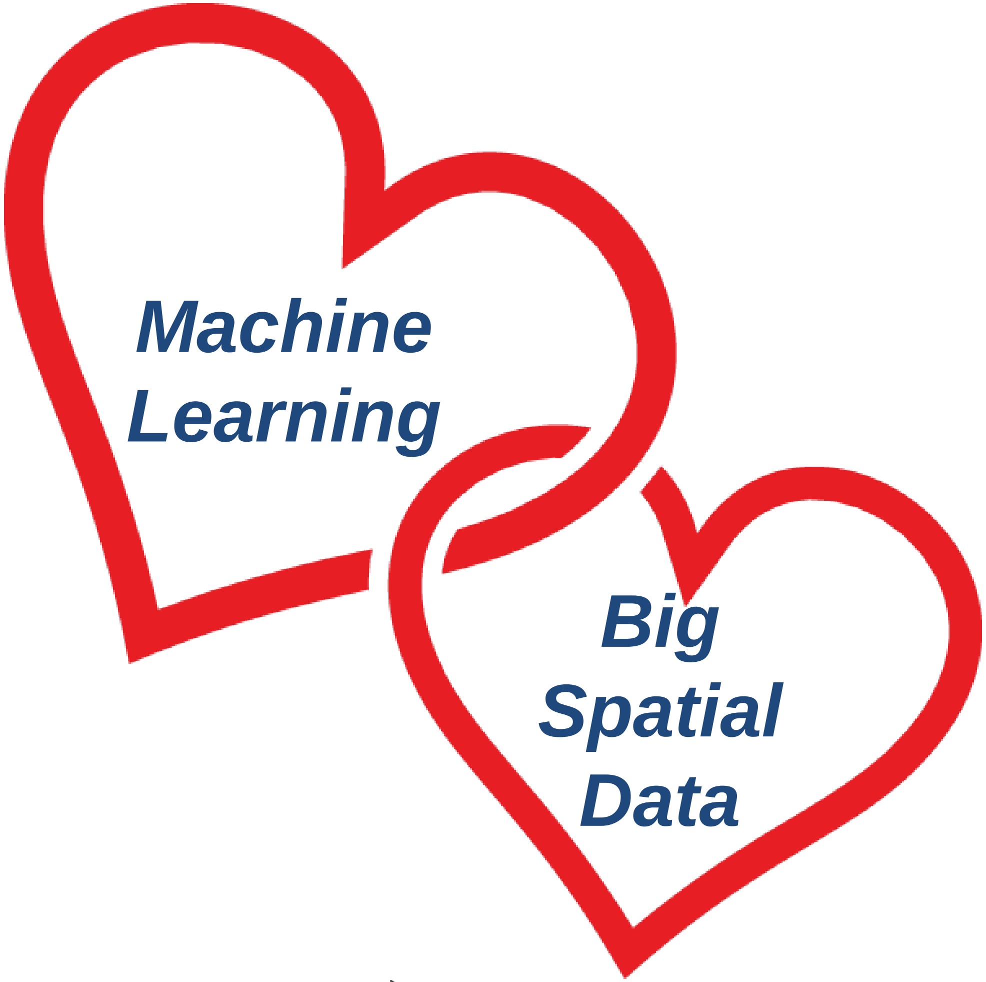


- Performing data curation
  - Enabling complex queries to be performed in real time
  - Success stories in weather prediction

L. J. Klein, F. J. Marianno, C. Albrecht, M. Freitag, S. Lu, N. Hinds, X. Shao, S. Rodriguez, and H. F. Hamann. “*PAIRS: A Scalable Geo-spatial Data Analytics Platform”.* In **IEEE Big Data 2015**

Machine Learning meets Big Spatial Data

ETHICS



PRIVACY

POLICIES


Thank you