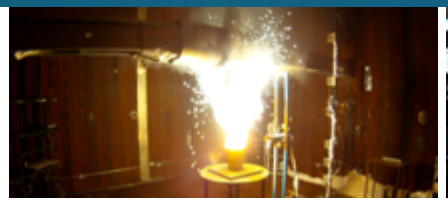
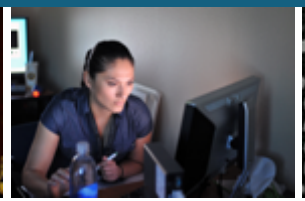
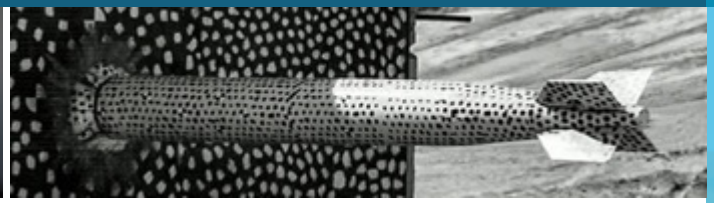
SAND2021-14134PE

Sandia National Laboratories

Big Data Actionable Intelligence Architecture







Presented by: Tian J. Ma

Distinguished Member of R&D, S&E Staff

Sandia National Laboratories

505-284-1238

tma@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

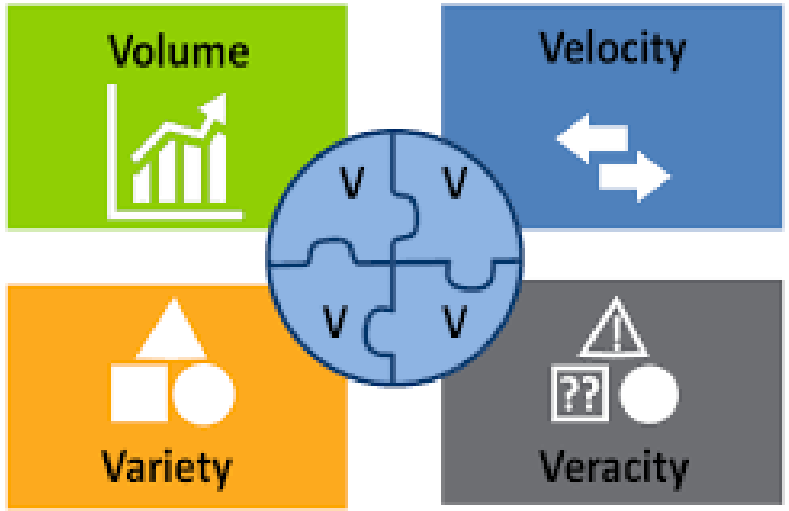
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

OUTLINE

- Problem Overview
- Big Data Actionable Intelligence Architecture
- Results
- Conclusion

What is Big Data?

- Big data – a collection of large or unstructured data sets that cannot be processed by traditional processing applications
- Major characteristics
  - Volume (Terabyte to Exabyte)
  - Velocity (e.g. streaming)
  - Variety (data in many forms)
    - big structured data (i.e. relational table) that cannot be processed by traditional applications
    - semi structured (e.g. xml, json)
    - unstructured data (e.g. text, signal, audio, video, graph)
  - Veracity (i.e. data in doubt due to inconsistency and incompleteness of data)



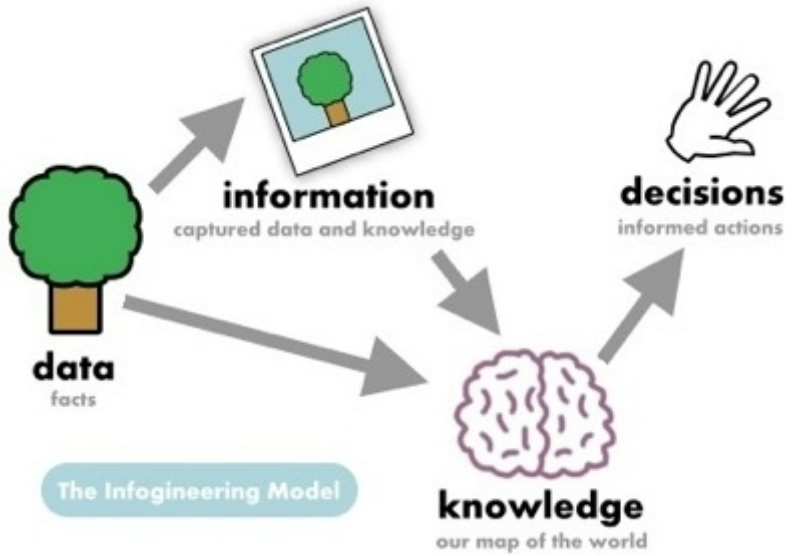
Problems with Big Data

- Significant increase in amount of digital information
  - “90% of the data in the world today has been created in the last two years alone, at 2.5 quintillion bytes of data a day! And, says the report, with new devices, sensors, and technologies emerging, the data growth rate will likely accelerate even more.” - IBM Marketing Cloud (December 2016)
- Terabytes of digital data collected by intelligence agencies each day (phys.org 2017)
- “Intelligence agencies are swamped” (phys.org 2017)
- The need to quickly and automatically turn data into useful insights

Data is “Gold” only if you can uncover an insight

Decision Makers

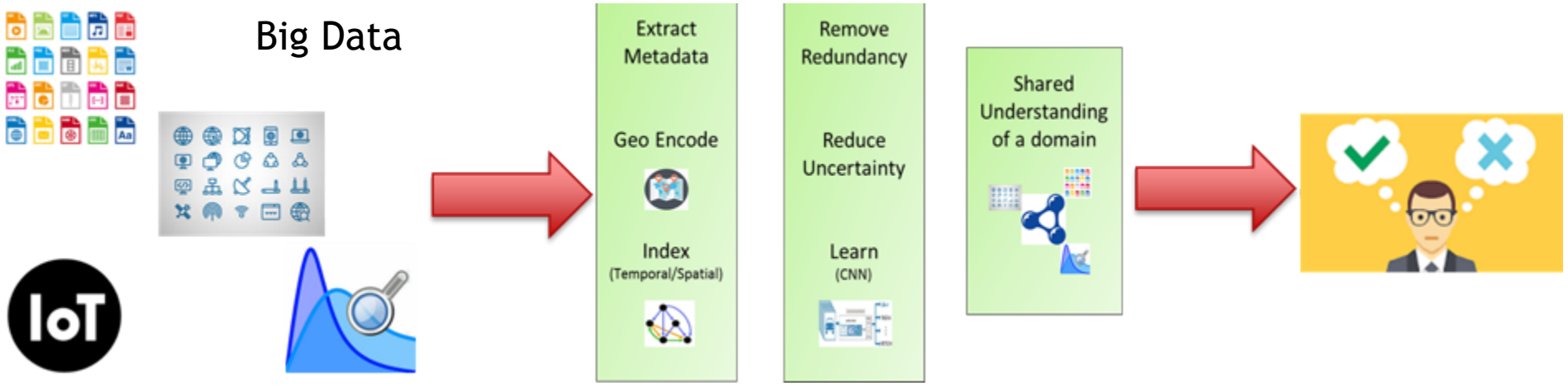
- Relies on multiple relevant data sources to support decision making
- Problem – Time is critical!
  - Many data sources are available
  - Difficult to sort through all relevant information
  - Often takes a lot of time to extract information from data sources
  - Some data sources may provide conflicting reports
    - Requires additional data sources or further analysis to resolve possible conflicts
  - Some data sources could be redundant
    - Requires removing redundancy to save time and leverage redundancy to improve reliability
  - Information from a data source could be incomplete
  - Requires connecting dots between data sources to extract key pieces of information



How can I quickly sort through relevant information in a short time and generate actionable insights to make decisions?

What is Actionable Intelligence?

- Actionable Intelligence - the next level of Big Data Analysis
- Quickly turns real-time streaming data from a variety of sources into actionable insights that enables decision makers (political leaders or field commanders) to take appropriate actions when faced with a security threat



Key Research Question

- Can we create a (near) real-time, (semi) data-agnostic, software architecture that is capable of fusing many disparate sources where it automatically generates Actionable Intelligence?
- Exemplar to prove the architecture:
  - Requires significant data diversity and high data value
- Exemplar: generate Actionable Intelligence regarding traffic congestion in Chicago

Traffic Raw Data Details

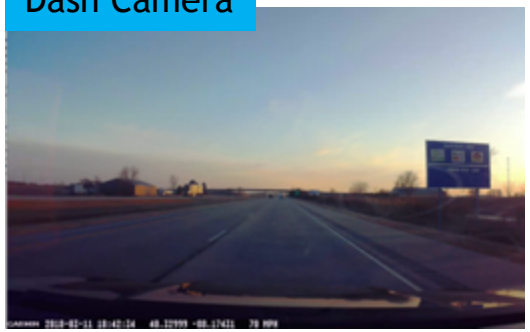
Web Camera Locations

Tweet

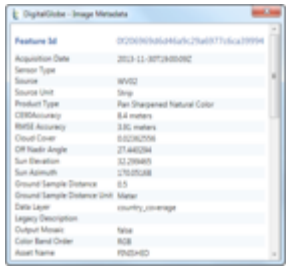

```
{
    "createdAt": "Mar 24, 2018 9:38:28 AM",
    "id": 977570348328775700,
    "text": "traffic is terrible omg",
    "source": "<a href=\"http://twitter.com/download/iphone\"
    rel=\"nofollow\">Twitter for iPhone</a>",
    "isTruncated": false,
    "inReplyToStatusId": -1,    "inReplyToUserId": -1,
    "isFavorited": false,
    …
```

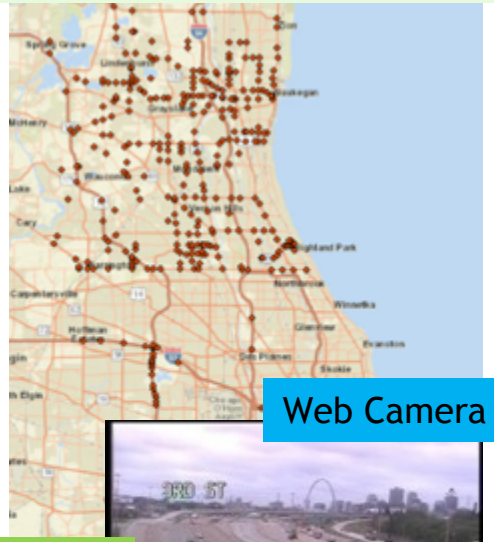

http://content.dot.wi.gov/travel/milwaukee/cameras/cam110.jpg|WI-100 at North Ave|43.06005|-88.04787
http://content.dot.wi.gov/travel/milwaukee/cameras/cam111.jpg|WI-100 at Watertown Plank Rd|43.0458|-88.04701
http://content.dot.wi.gov/travel/madison/cameras/cam265.jpg|I-39/90 at Milwaukee St|42.69652|-88.98436

Dash Camera



Digital Globe





Web Camera

Electronic Traffic Sign

Object Detections

car: 50%
car: 36%
car: 31%
car: 51%
bus: 45%
car: 53%
car: 79%

MapQuest


```
"incidents": [
  {
    "id": "68786261",
    "type": 3,
    "severity": 2,
    "eventCode": 73,
    "lat": 41.71352,
    "lng": -87.600822,
    "startTime": "2018-03-08T11:37:00",
    "endTime": "2018-03-08T12:35:33",
    "impacting": true,
    "shortDesc": "Bishop Ford Fwy N/B: delays increasing in Chicago",
    "fullDesc": "Delays increasing and delays of six minutes on Bishop Ford Fwy Northbound in Chicago. Average speed 15 mph.",
    "delayFromFreeFlow": 6.880000114440918,
    …
```

Data Sources Description

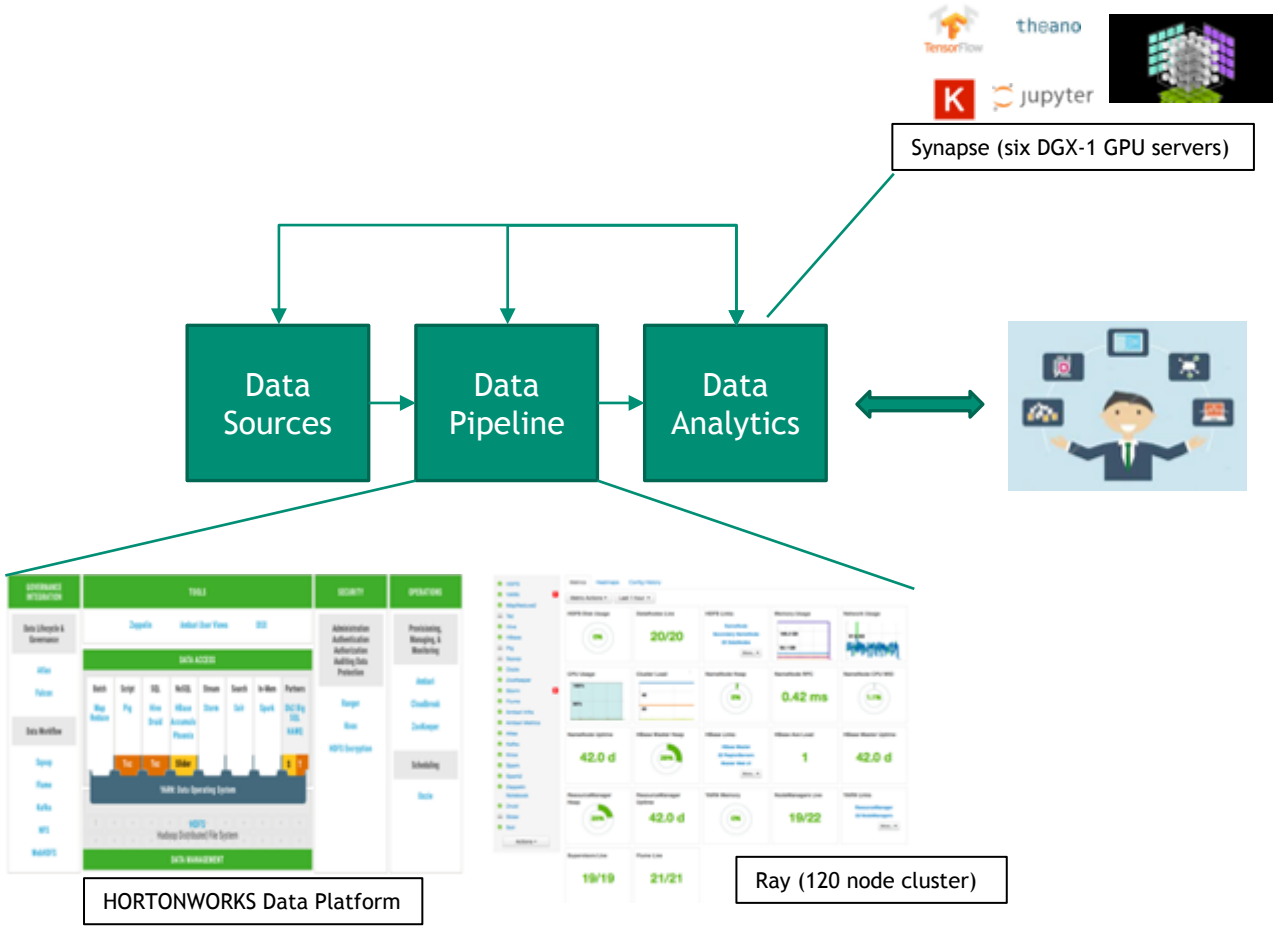
Data Sources	Source Type	Frequency	Description
Twitter [10]	Live text	• Live. Query every 5 mins	• Decahose – Geo-tagged tweets within Chicago city limits
Travel Mid-west [11]	Various	• Traffic camera images every 15 min • Vehicle Detection System (VDS) every 10 min • Dynamic Message Sign (DMS) every 10 min • Thousands of camera locations	• Traffic Cameras • VDS - Vehicle Speeds, Vehicle Occupancy • DMS – Traffic times, Lane Closures, Accidents
City of Chicago [12]	Various	• Traffic Segments every 10-15 mins • Traffic Region every 10-20 mins • Construction Moratorium - Infrequent	• Traffic Segments – Vehicle Speeds, Vehicle Occupancy • Traffic Region - Vehicle Speeds, Vehicle Occupancy • Construction Moratorium – Road closures
GDELT [13]	Various	Every 15 minutes	Global Knowledge Graph – provides context and feeling between people, organizations, and locations Event Mentions, Events
MapQuest [14]	Various	Every 5 minutes	Reported Incidents
Digital Globe [15]	Satellite Imagery	1-3 images a day	Satellite Imagery (limited number of images)
Dash Camera	3-hour Video	Field experiment	Dash Camera Video (Live Experiment and Validation)

System Requirement

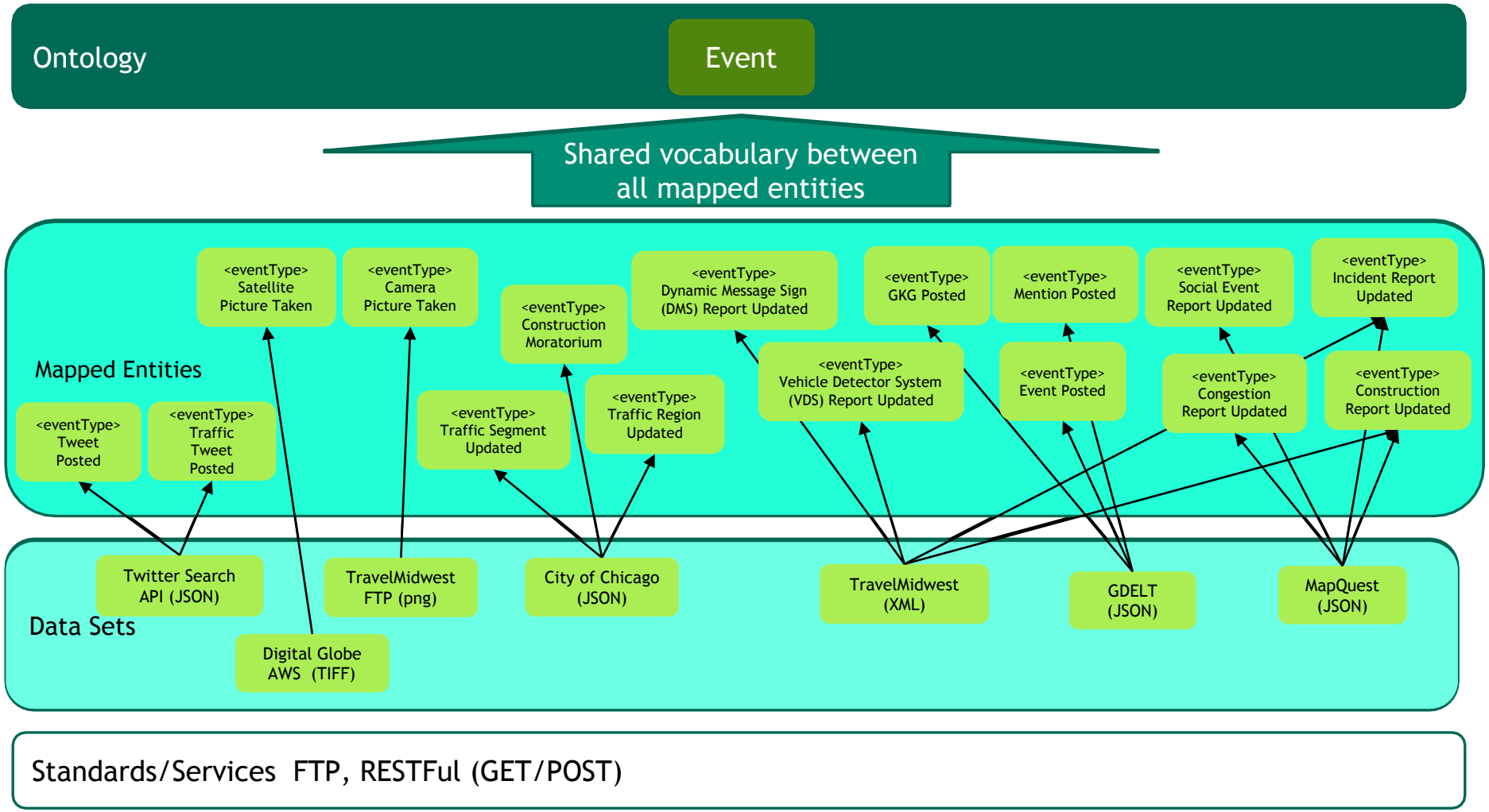
Requirements	Descriptions	Goal	Threshold
Scalability	Number of streaming location supported	150 streaming location	100 streaming location
Data Variety	Structured, Unstructured, Semi-structured	Structured, Unstructured, Semi-structured	Structured, Semi-structured
Average Throughput Per Location	Average data transfer rate per location	1 Mbps per source Location	0.50 Mbps per source location
Average Data Latency	Time measured from data creation to the time the data has arrived and indexed into our system	Less than or equal to the polling frequency	max (polling frequency, data update frequency) + 2 minutes
Data Management Guarantees	Level of guarantee on which message to be processed	Fully process each message	Drop message on failure
Traffic Classification Accuracy	Traffic Classification Accuracy	95% accuracy on trained location	90% accuracy on trained location

Big Data Actionable Intelligence (BDAI) Architecture

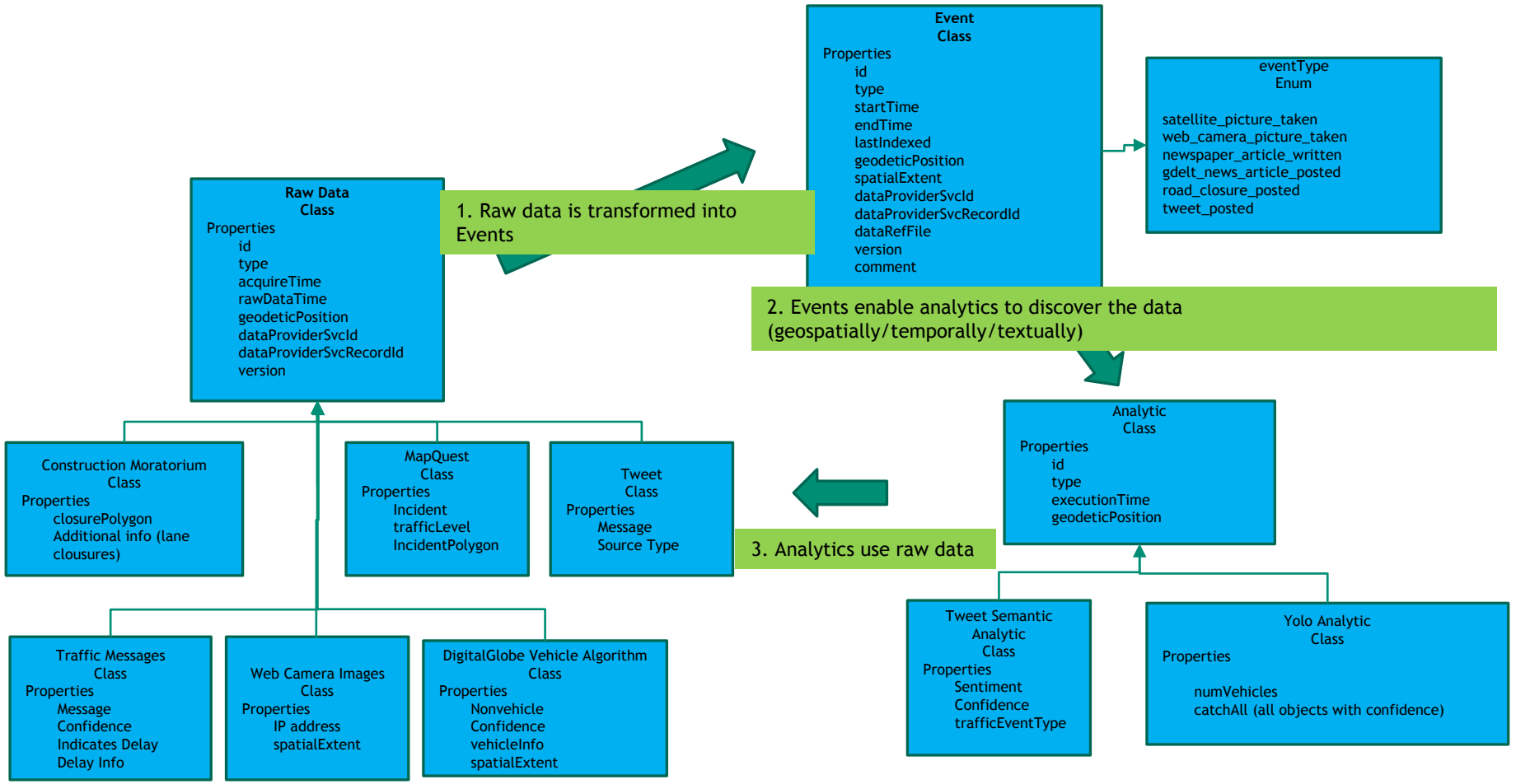
The data architecture provides a shared understanding of all data sources collected. We normalize the data into a common frame of reference so that meaningful comparisons of events in space, time and context are possible.



General Event Schema



BDAI Data Management



Data Architecture – Data Pipeline

Data Sources

Twitter:
- Tweets around Chicago: ~125k tweets / day
- Tweets filter on traffic keywords and emojis

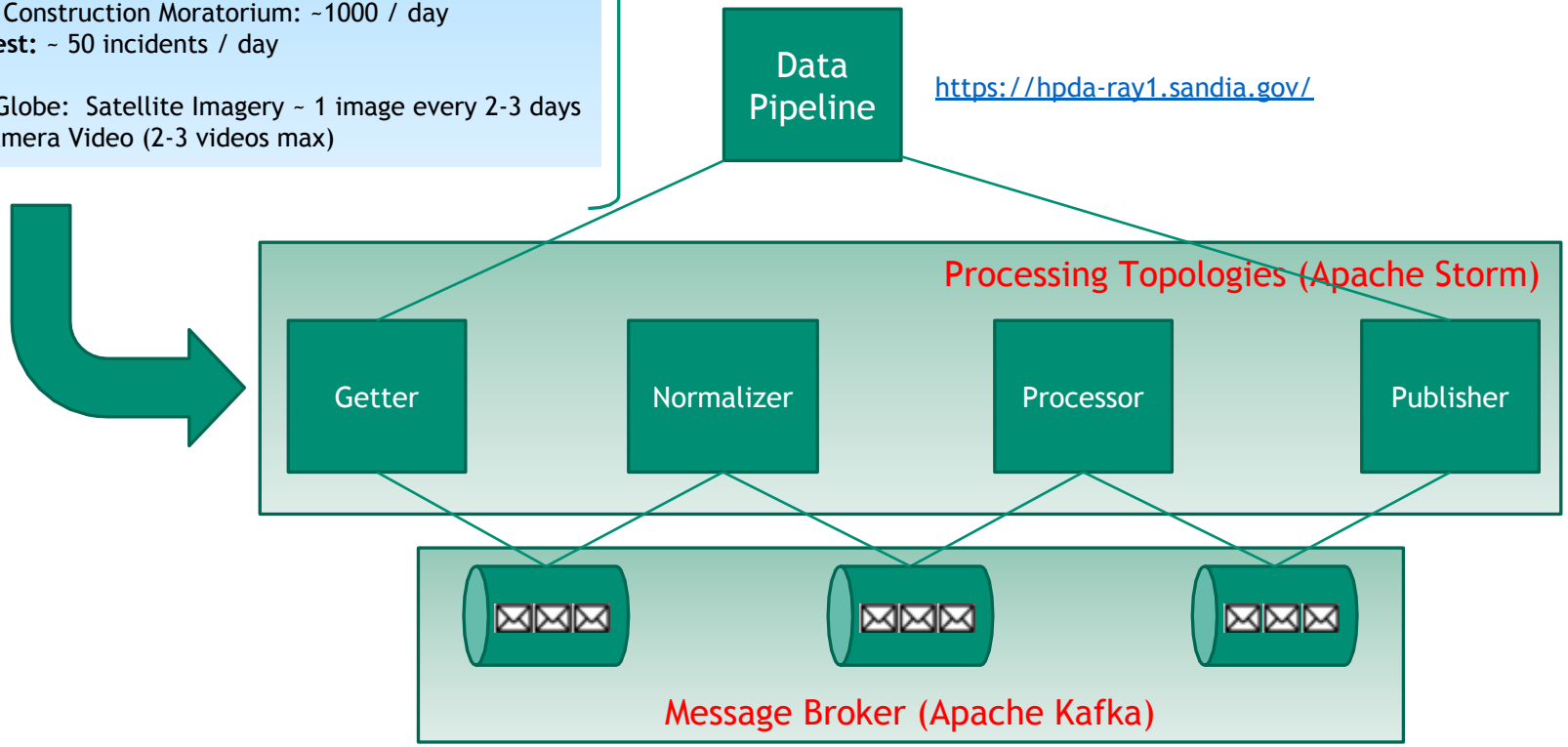
TravelMidwest:
- Web Camera Images: ~20k / day
- Vehicle Detection System readings: ~30k / day
- Dynamic Message Sign records: ~20k / day

City of Chicago
- Traffic Segments : ~120k records / day
- Traffic Region: ~4k records / day
- Construction Moratorium: ~1000 / day

MapQuest: ~ 50 incidents / day

Digital Globe: Satellite Imagery ~ 1 image every 2-3 days
Dash Camera Video (2-3 videos max)

~300k records / day



Events to Decision

Traffic Sensors (All Normalized)

YOLO

- Vehicle Count, Y_VC

car: 82% car: 62% car: 55% car: 50%

Traffic Volume Report

- Volume, TVR_V – Vehicle Count
- Occupancy, TVR_O
- Speed, TVR_S
- Impact, TVR_I

I-74 NB from 7th Ave to 4th Ave RAMP
occupancy: 0.26666668
speed: 17.314816

Traffic Message Signs

- Message Sentiment, TMS_S

TIME TO IL-59 ORCHARD RD MIN 11 19 / / SEE FLASHING LIGHTS MOVE OVER GIVE THEM DISTANCE

Traffic Incident Reports

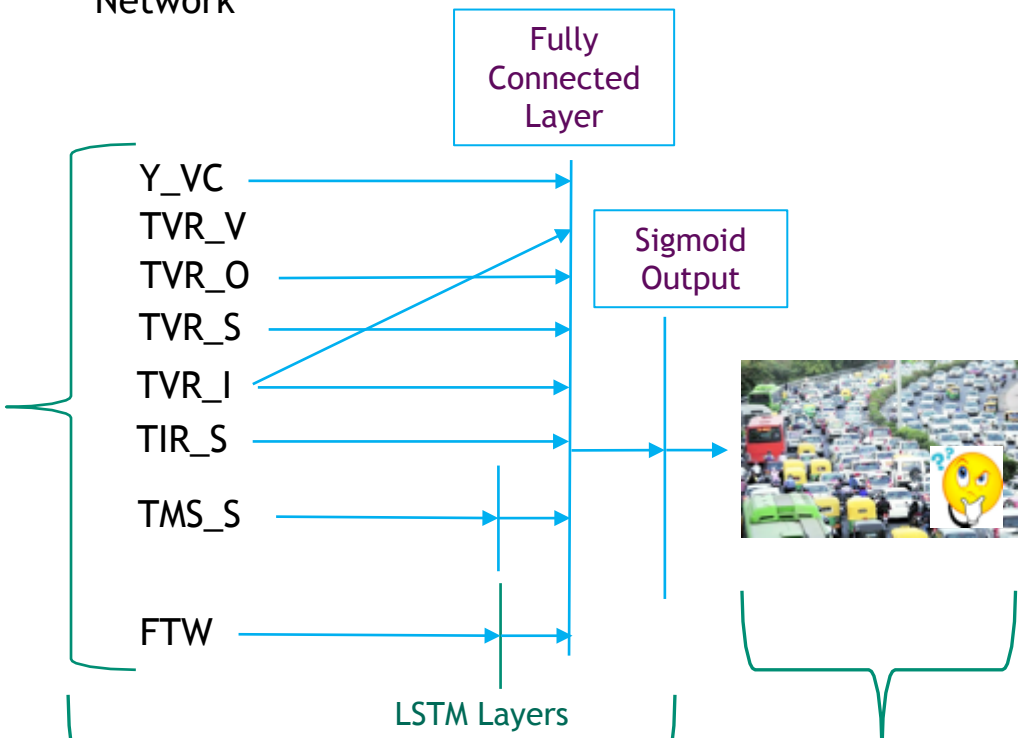
- Severity, TIR_S

Filtered Tweets, FTW

OMG somebody call me i'm falling asleep in this traffic...😩

Nice, clean data

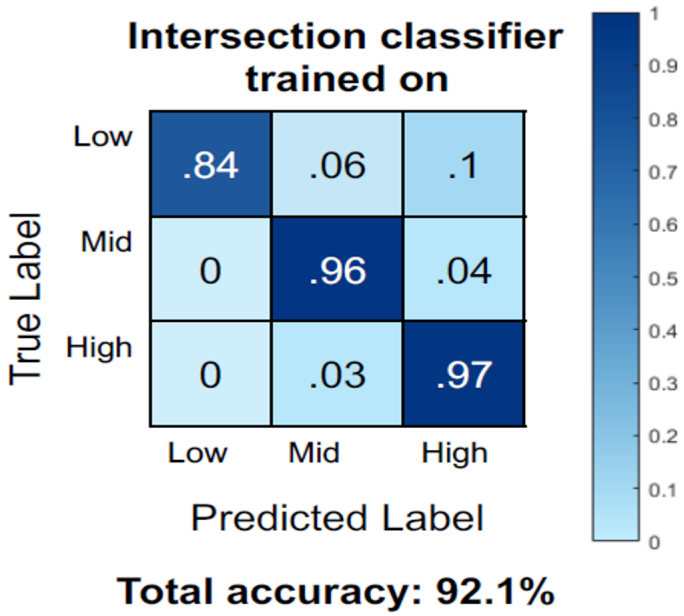
Traffic Jam Merged Neural Network



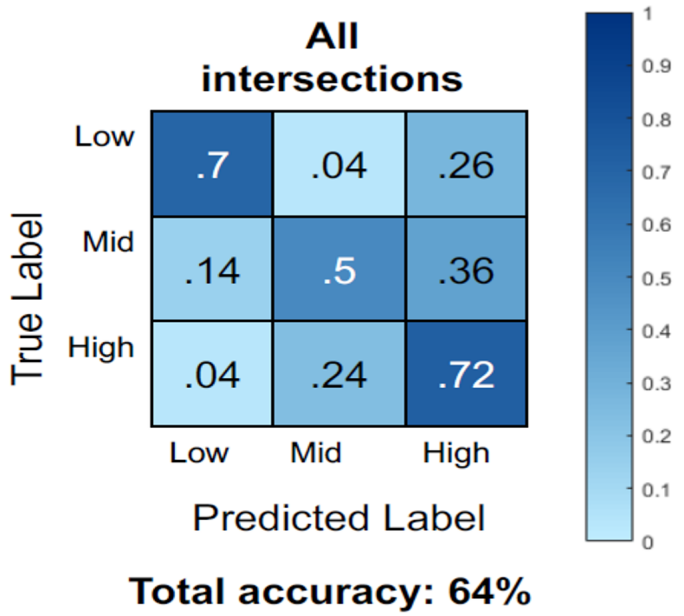
Run it through an algorithm

Provide a Traffic Jam Probability

Classification Accuracy



Total accuracy: 92.1%



Total accuracy: 64%

Low = Minimal slow down

Mid = Some traffic impact

High = Significant impact

Latency Performance vs Requirement

Event Type	Avg Record Per Day	Num of Source Station	Query Freq (min)	Polling Freq (min)	Average throughput (bytes/sec)	Average Latency (Measured)	Status
Tweets	132K	1	5 or 15 (subject to rate limit)	5 or 15 minutes subject to rate limit)	5873	5.5 min	Met Goal
Tweet Traffic Posts	0.5K	1	5 or 15 (subject to rate limit)	5 or 15 (subject to rate limit)	25	5.8 min	Met Goal
Camera Images	14K	150	15	15	56879	3 min	Met Goal
Gdelt Global Knowledge Graphs	4.2K	1	15	15	5386	3 min	Met Goal
Gdelt Mention Posts	4.5K	1	15	15	391	1 min	Met Goal
Gdelt Event Posts	1.7K	1	15	15	25	1.7 min	Met Goal
Dynamic Message Sign Report	20.9K	150	15	10-12	532	34.9 min	Failed

The larger latency in Dynamic Message Sign Report was associated with an inconsistent update interval provided in the server rather than the actual latency in our system.

Conclusion

- BDAI architecture provides a framework for machine-learning algorithms
- Data-agnostic solution
- Fully realized and deployed in a bare metal system
- Proven with near-real-time data from hundreds of sources
- Future work: enhance decision science
- Research has been featured News Media
  - Sandia Labs News Service (2019). “Wrangling Big Data”, Albuquerque Journal, November 4, 2019. https://www.abqjournal.com/1386752/wrangling-big-data-to-locate-actionable-info-a-lot-faster.html
- Research has been published in the Springer Journal of Big Data (peer-reviewed and listed in Scopus)
  - Tian J. Ma, Rudy J. Garcia, Forest Danford, Laura Patrizi, Jennifer Galasso, and Jason Loyd; “Big Data Actionable Intelligence Architecture”, Journal of Big Data, September 2020, DOI: 10.1186/s40537-020-00378-7 https://rdcu.be/cbdiv