# PCAP TO HDFS

## OARC25

### DNS-OARC WORKSHOP

DALLAS - TX

cira

BUILDING A BETTER ONLINE CANADA


**Presented by:**

**Elson Oliveira**


October – 16th – 2016

# CANADIAN INTERNET REGISTRATION AUTHORITY

- Aka CIRA
  - .CA registry with over 2.5 million domains
    - CiraLabs
    - D-Zone
    - Fury
    - IPT

.ca | D-Zone Anycast DNS

cira LABS

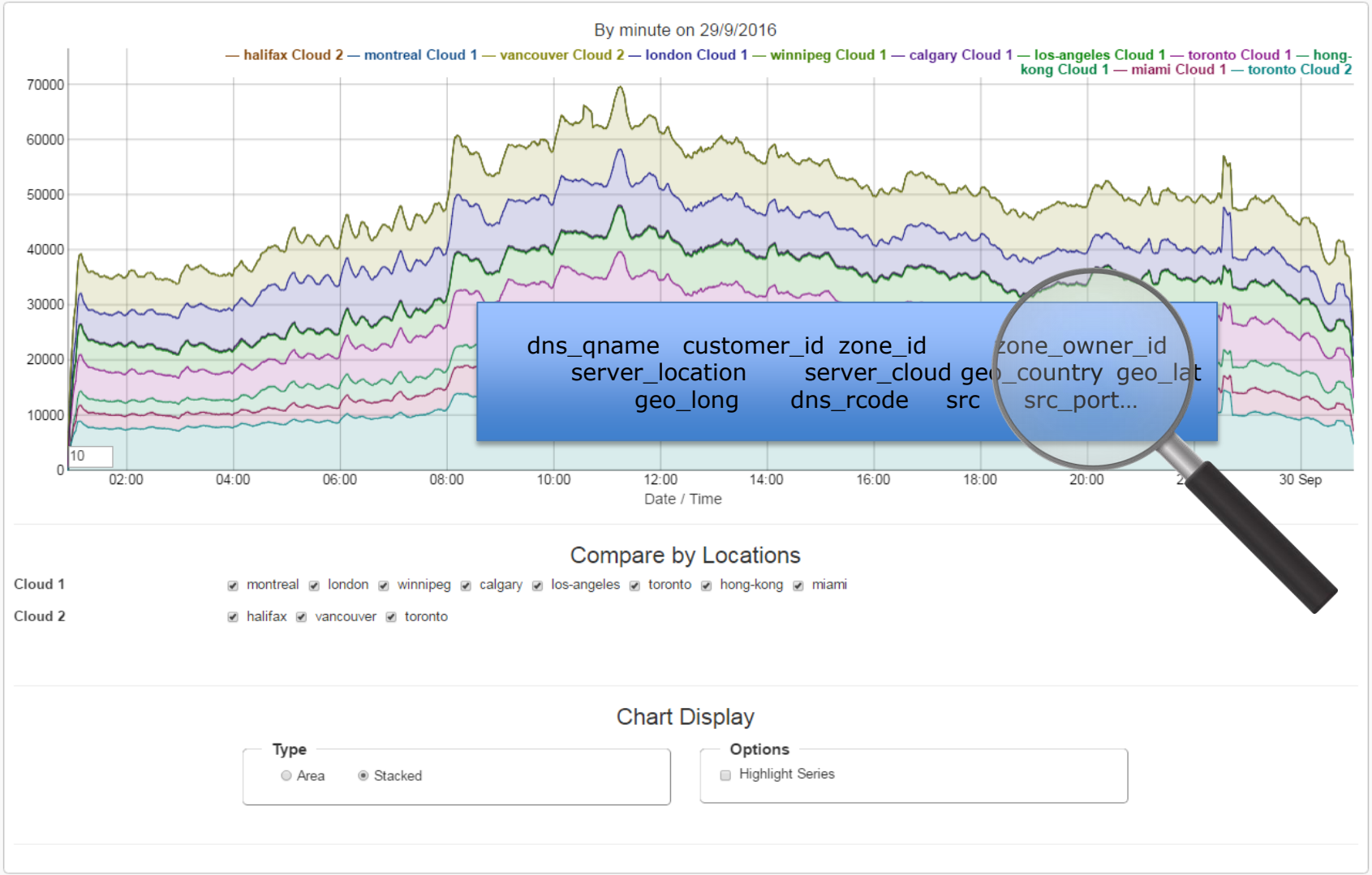
.ca | Internet Performance Test

.FURY

# MOTIVATION – ENHANCED D-ZONE REPORTING

- Previous Stats
  - Total queries
  - Half hour interval
- Lost data
- Bind specific endpoints
  - Parsing
  - Delta calculation
- Scalability (RDBM?)
- Enhance customer metrics and troubleshooting capabilities

# D-ZONE ANALYTICS



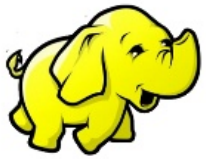
# OPTIONS – MORE PERFORMANCE?

- passiveDNS
- DNSTap
- Bro  - Network Security Monitor
- Vertica
- ENTRADA – ( PCAP + HDFS + Parquet + Impala )
  - Starting point

# INGESTION ENGINE

- Challenges
  - Read n files
  - Join packets
  - Overlay customer information
    - Relational database
  - Geolocate IPs
  - Expensive operations
- Solution
  - Async and parallel execution
  - Actor model
  - Chain architecture
  - Task specialized actors

# OUR TOOL KIT – LEVERAGING ENTRADA



- PCAP decoder
- Hadoop
  - Cloudera distro
- AVRO staging
  - Data model
  - Flume Data Stream
- Parquet file conversion
  - File size optimization
  - 256MB
  - Oozie

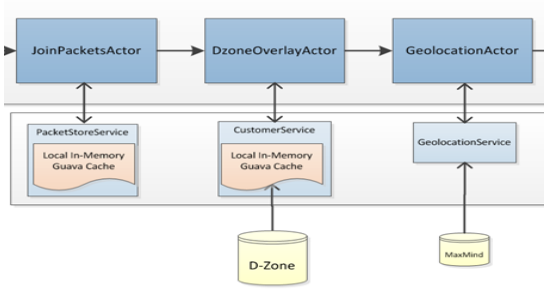

- Impala query engine
  - JDBC
  - Impyla

Aggregation

- Oozie / sqoop
  - Daily
  - Monthly
- Postgres

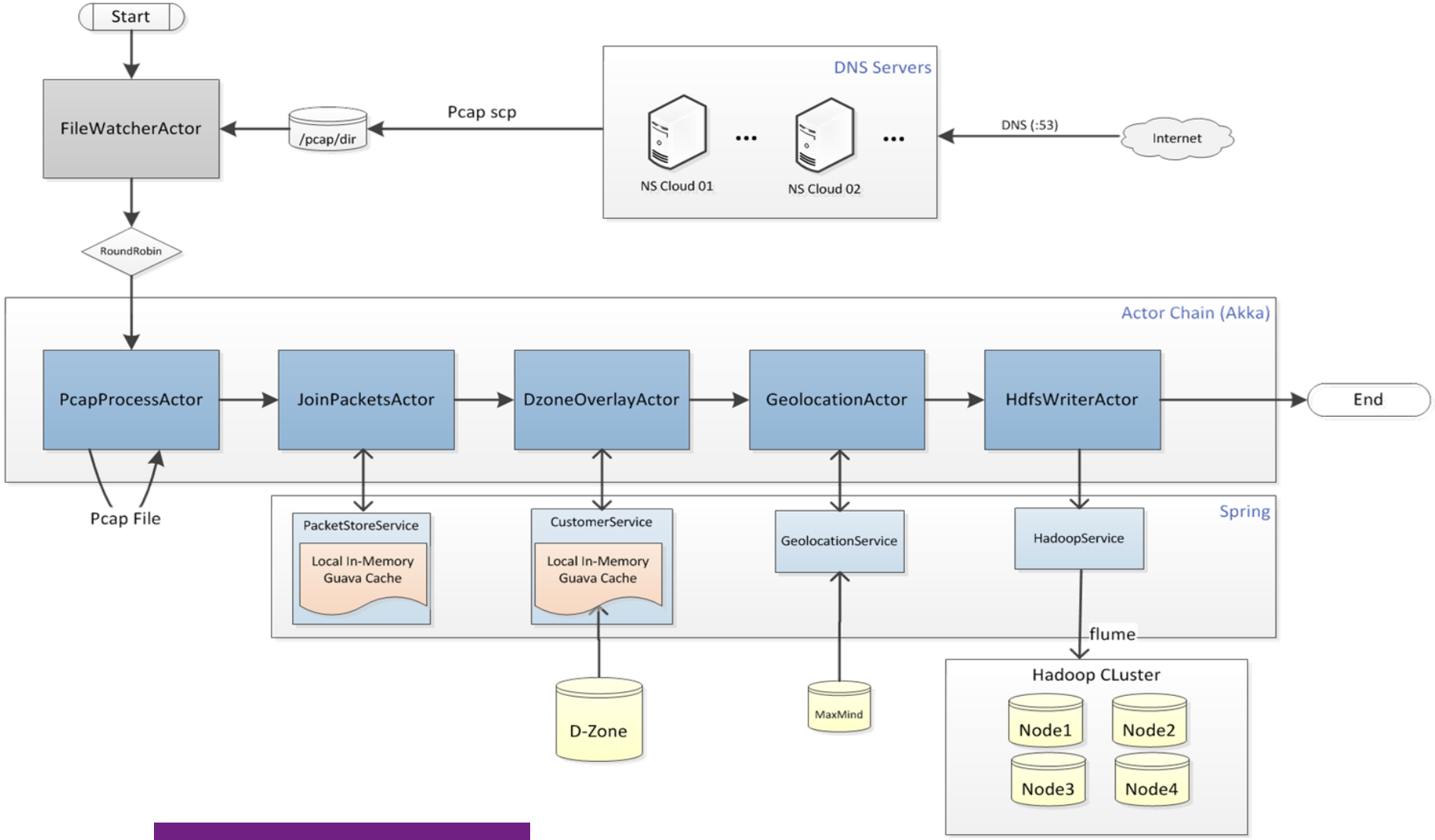
# THE ACTOR CHAIN

- Concurrent processing - Message driven actors
- Configurable number of instances
- Easily scalable
- +performance ☺




- The AKKA toolkit
  - Open source toolkit for concurrent and distributed applications on the JVM.
  - Supports multiple models
  - Actor-based concurrency

# PCAP TO HDFS WORKFLOW



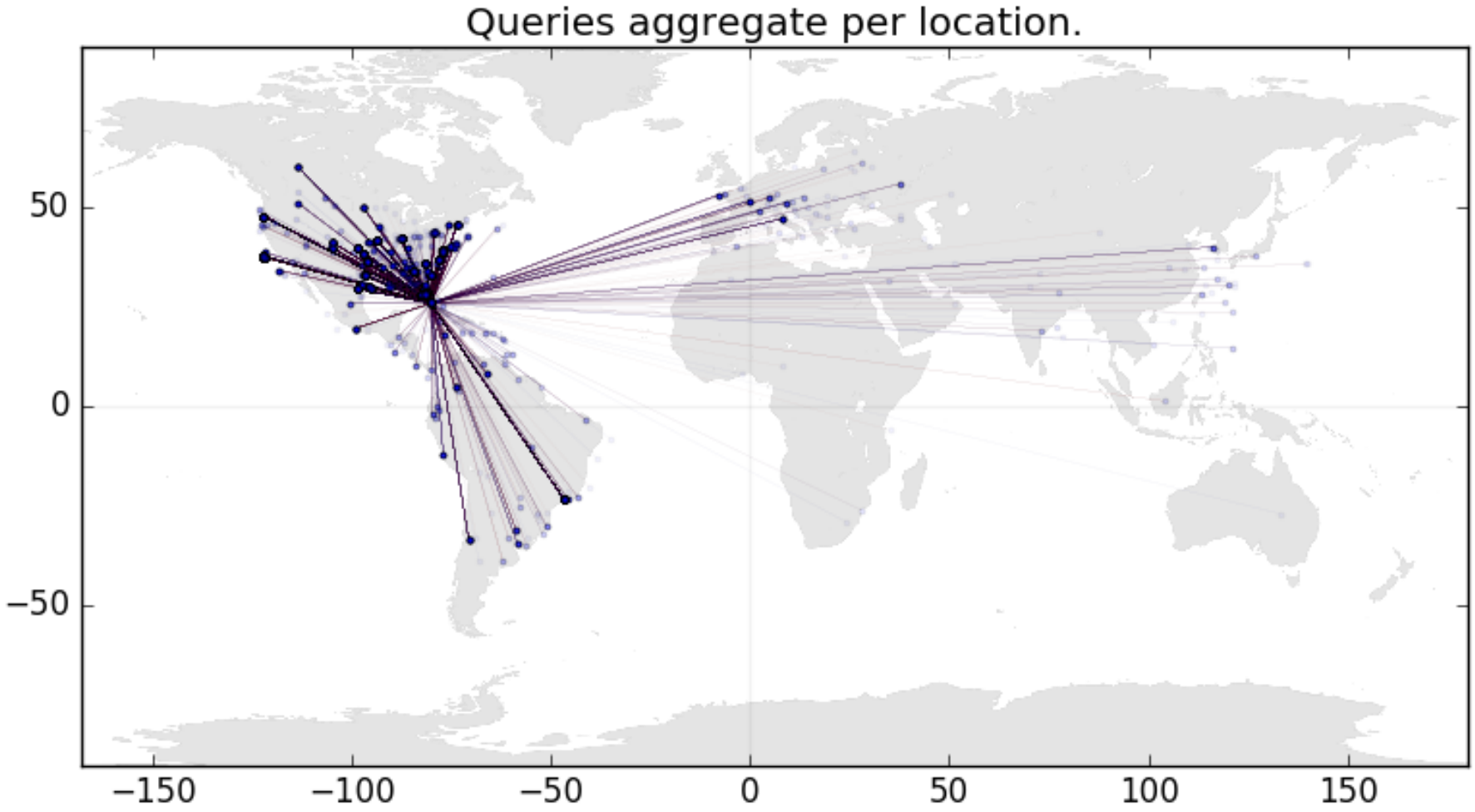
# IMPLEMENTATION SCOPE

- D-Zone
  - Secondary DNS
    - Anycast cloud
  - 1k queries/sec
  - 2 data nodes
  - 1GB / 5 min PCAPS


- .CA
  - .CA Name Servers
  - 10k queries/sec
  - 4 data nodes

cira
LABS

# QUERIES SNAPSHOT

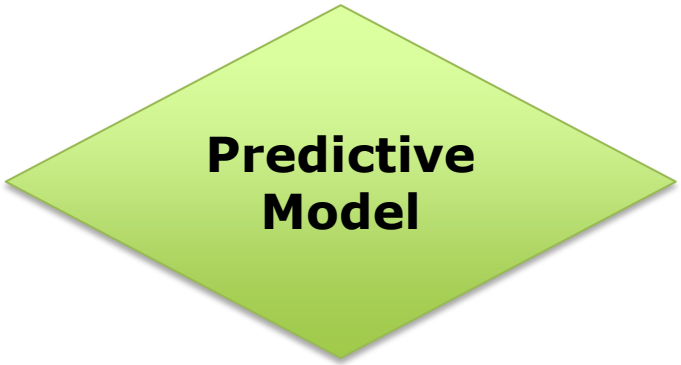


# CIRA LABS
# MACHINE LEARNING EXPERIMENTS

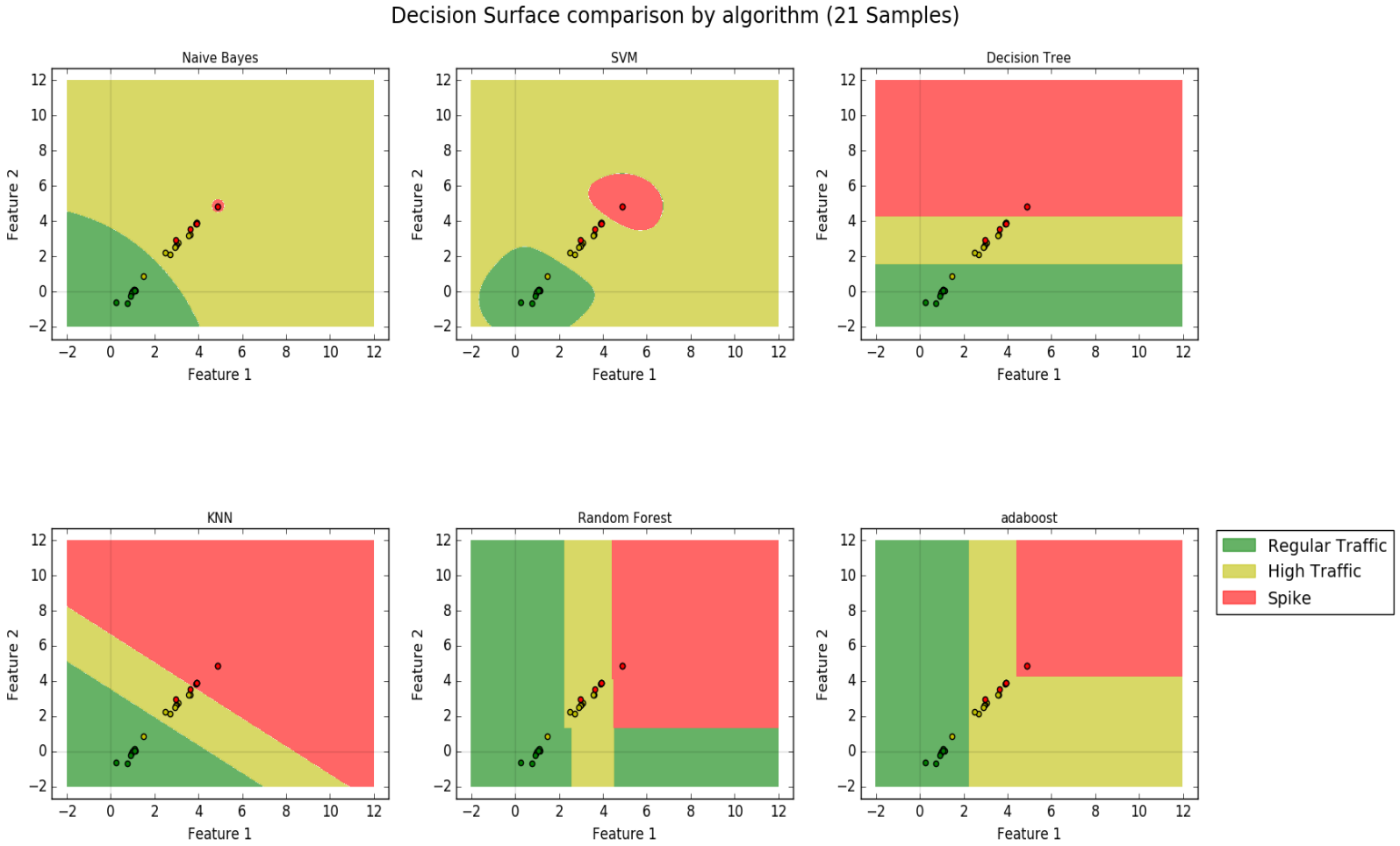
- Python / Sklearn
- Supervised Learning
  - Spike detection
    - Total queries
      - Location
      - Type (A, MX, NS…)
      - Source IP
    - Non responded queries
    - Non asked queries
  - Unsupervised
    - Queries Clustering
      - K-Means
    - Dimension reduction
      - PCA


# MACHINE LEARNING

- dns_qname
- server_location
- geo_country
- geo_lat
- dns_rcode
- src
- src_port
- dst
- dst_port
- ts_delta
- ts_query
- ts_resp
- dns_qtype
- dns_queryid
- ip_header_len
- ...

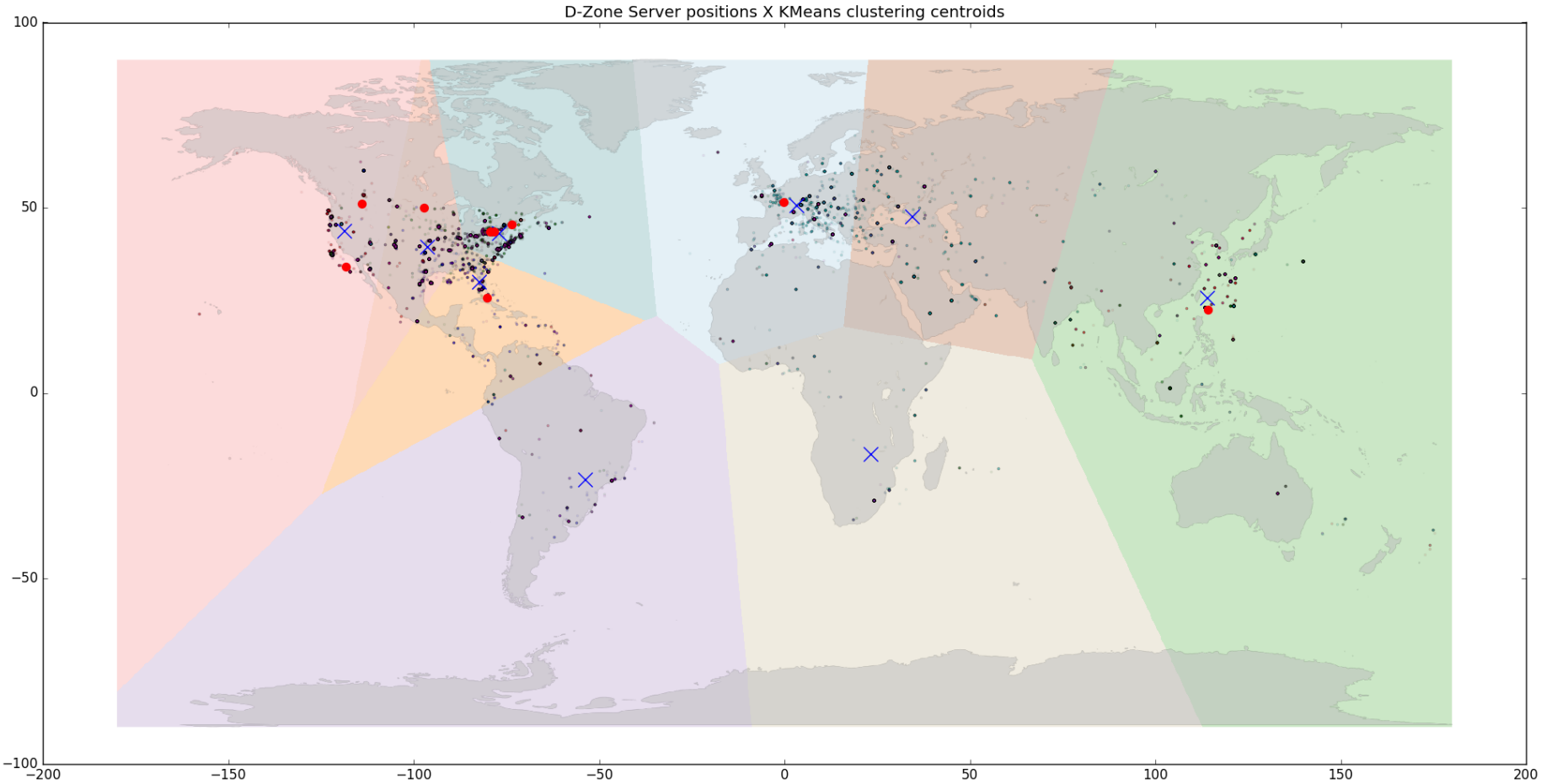


# DECISION SURFACE

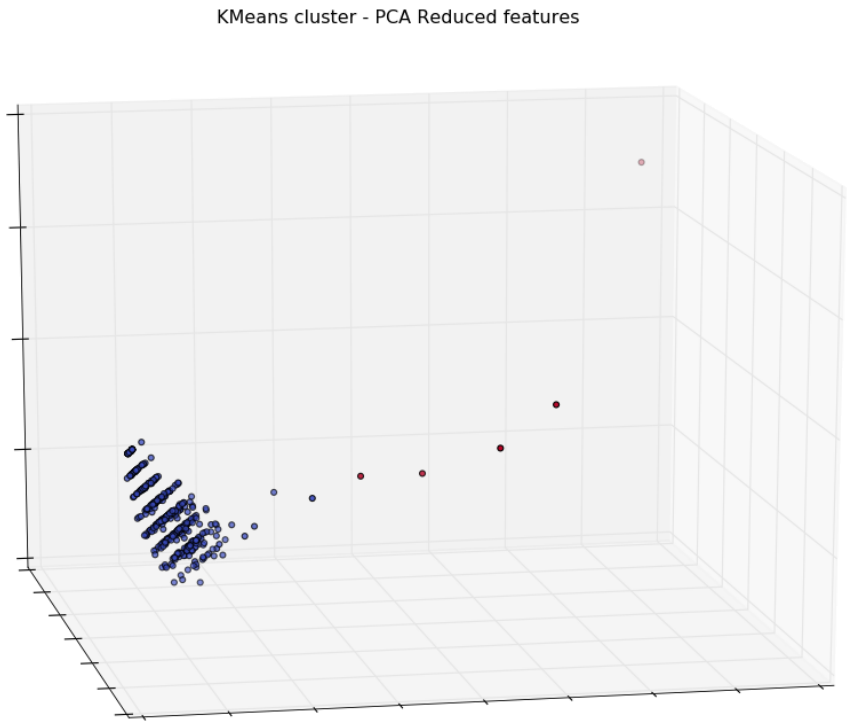


# SPIKE DETECTION

# SERVER DISTRIBUTION ANALYSIS



# CLUSTERING




Features:

- Total queries
- Distinct servers
- Distinct sources
- Distinct rcode

| dns_qname | rcode_count | src_count | server_count | tot_queries |
|---|---|---|---|---|
| • domain1.ca. | 3.0 | 104.0 | 9.0 | 332.0 |
| • domain2.ca. | 3.0 | 177.0 | 8.0 | 455.0 |
| • domain3.ca. | 3.0 | 18.0 | 8.0 | 723.0 |

# WHAT'S NEXT?

- Analytics
  - Improve use of stored data
  - Lower granularity of search, enhance filtering capabilities
- R&D
  - Improve machine learning experiments in order to enable more DNS behaviour patterns detection
  - Apply existing ones
- Architecture
  - Make use of a distributed streaming platform for real-time traffic capture (pcap scps?)
    - Apache Kafka, Amazon Kinesis Streams
  - Distributed Engine

# QUESTIONS?

elson.oliveira@cira.ca

cira

LABS

cira.ca/blogs/cira-labs