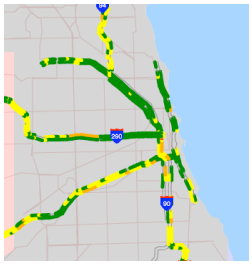
Using Baselines and Change Detection to Improve Information Quality

Joseph Bugajski, Robert Grossman,
Eric Sumner, & Zhao Tang
& Pantheon Gateway Team

Part 1:
Introduction & Background

Problem

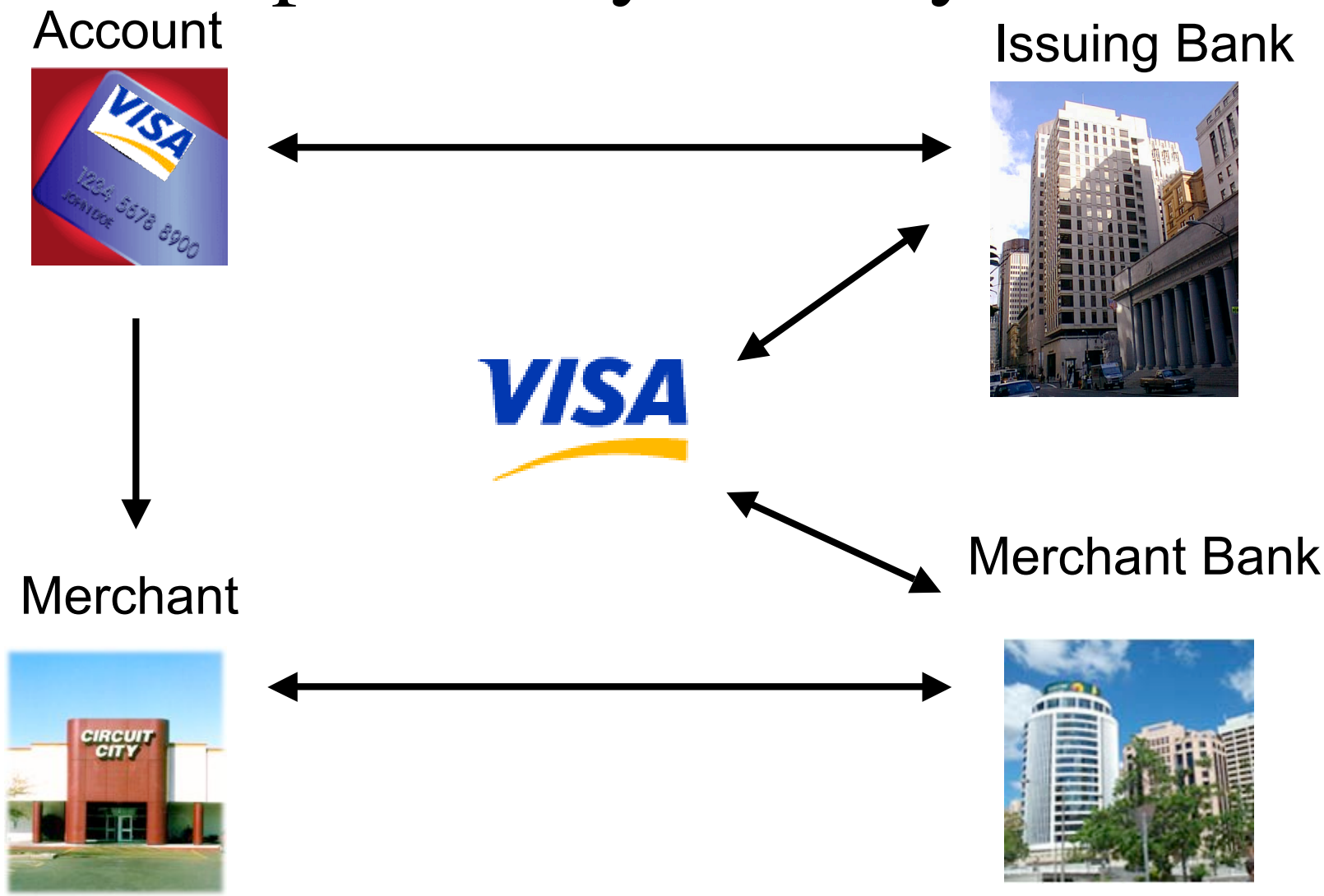
- How can we monitor, alert, and ameliorate information quality for complex distributed systems?
- Examples:
  - Payment systems
  - Distributed sensor systems
  - Homeland defense systems



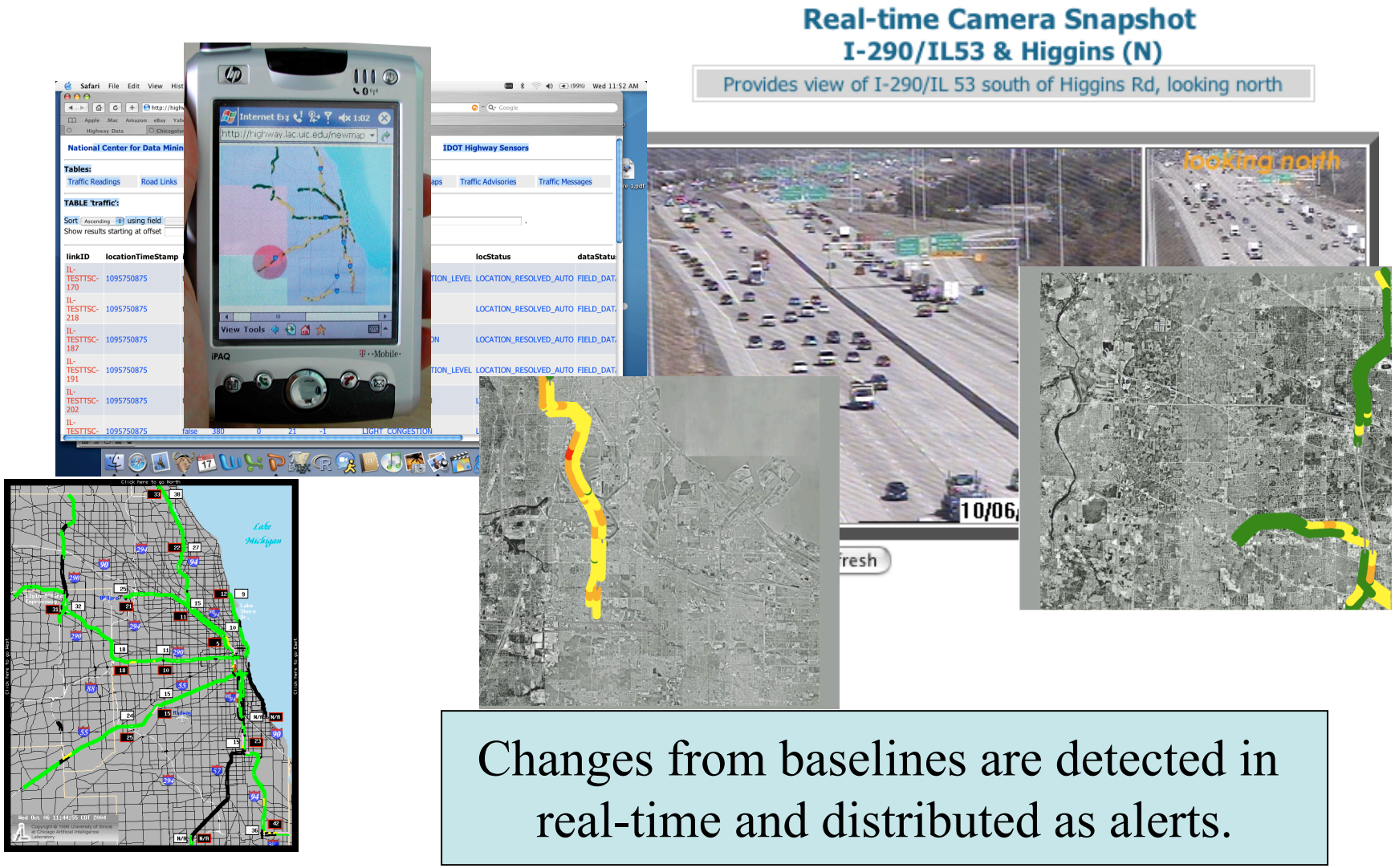
Challenges

1. Large, high volume, complex, distributed streaming data
2. Multiple parties involved, each of which can modify the data in subtle ways
3. System is sufficiently complex that establishing accuracy is a challenge

Example 1: Payment Systems

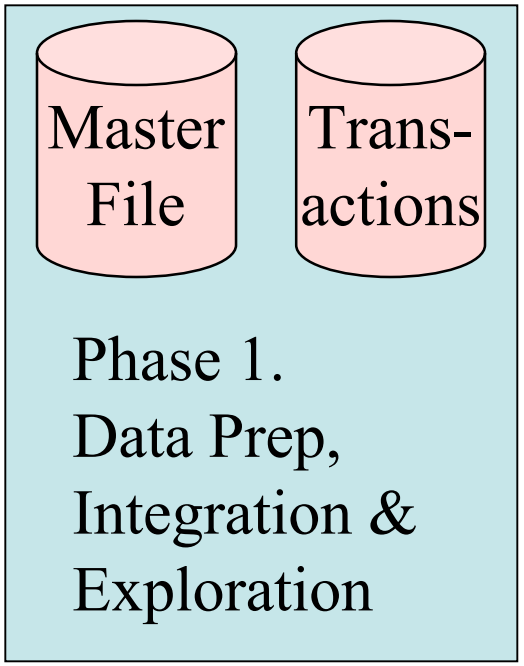


Example 2: Highway Traffic Data



Changes from baselines are detected in real-time and distributed as alerts.

Why Do I Care?
Data Mining: Process View



- If there are data quality problems with the data preparation, then the statistical and data mining problems are generally useless.
- This is a very common in practice.

Some Questions

- Is the number of payment exceptions from this merchant (frequent traveler, Sheraton, Friday, 9am, June, no holiday, etc.) **different** than the baseline?
- Is this traffic today leaving this workshop (Friday, June, 6pm, no convention events, no rain) **different** than the baseline?

Part 2:
Overview

Key Ideas

1. Don't try to measure accuracy (for example) - focus instead on change from established baseline distributions for accuracy (for example)
   - Typically create 10^3 to 10^6 different baseline models.
   - Use a variety of different change detection models.
2. Think of data as real time stream of events. Keep persistent state information.
   - Update scores for each event
   - Use multiple models change detection models
3. Investigate alerts generated from baseline models by hand in order to improve accuracy of baseline models.

What I'm Not Discussing

1. Create baselines (off line analysis)
   - Exploratory and statistical analysis of data
2. Monitor event stream (on line deployment)
   - Automatically generated by alert management systems
3. Root cause analysis
   - Domain experts investigate alerts
4. Ameliorate
   - Formal models used to give system designers and developers better tools to understand data and metadata

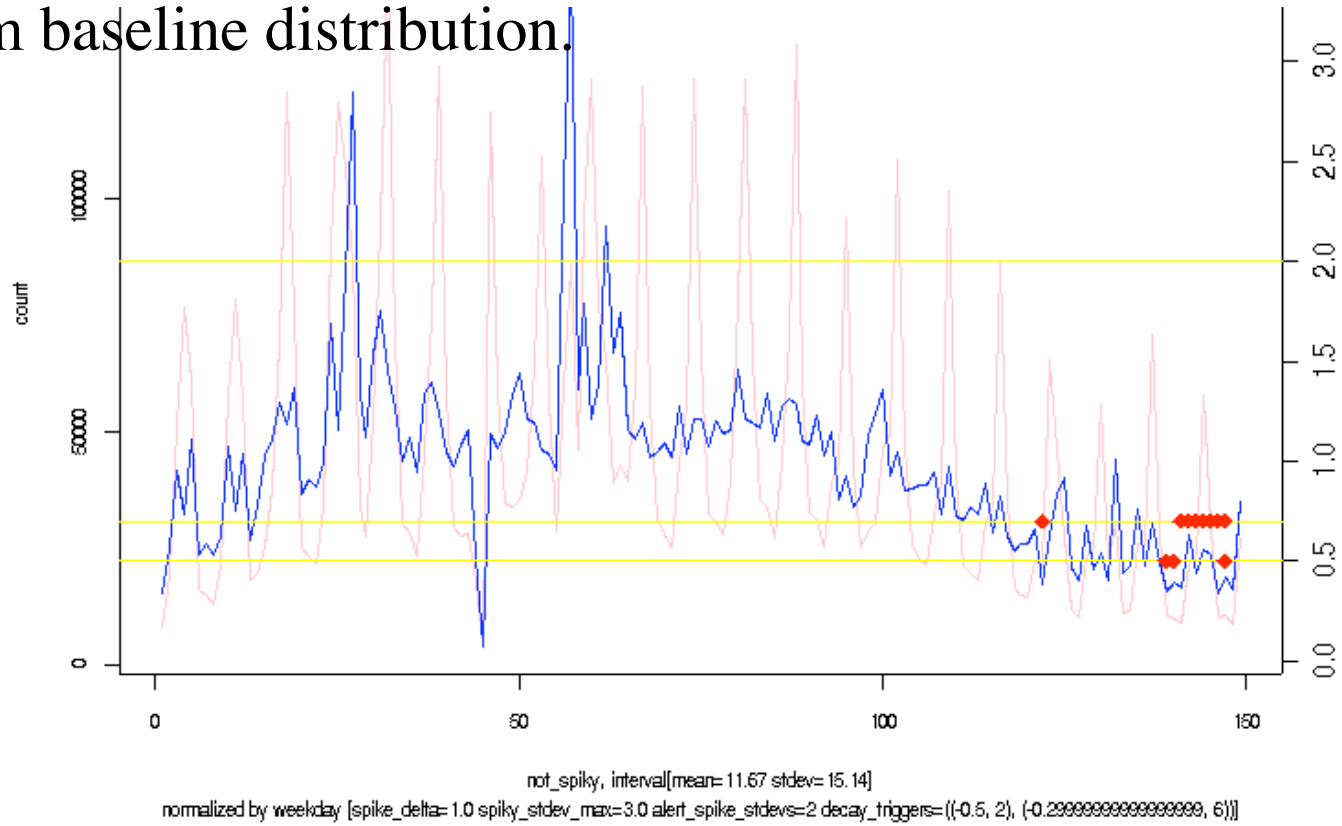
Part 3:
Technical Approach

Idea 1: Focus on Changes

- Traditional dimensions of data quality
  - Accuracy, completeness, consistency, timeliness, uniqueness & validity
- Point of view here:
  - One of more measures from the dimensions above
  - Build (many) baseline distributions for (some) of these dimensions and measure deviations from baseline distributions

1-A: Think Distributions, Not Values

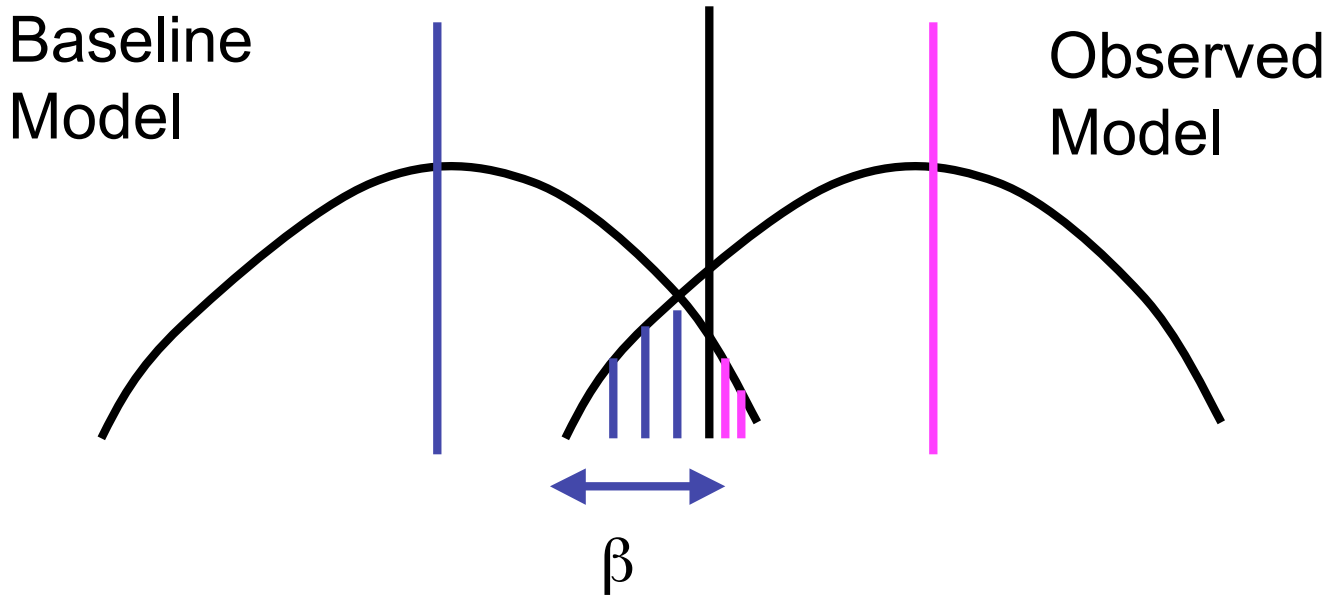
- Look at distributions visually
- Focus on distributions of values, pairs of values, triples, etc.
- Trigger (red) alerts when observed distribution deviations from baseline distribution.



1-B: Use Lots of Baselines

- Payment Systems
  - each field (hundreds) x each acquirer (thousands) x each merchants (millions)
- Highway Traffic Data
  - each day (7) x each hour (24) x each sensor (hundreds) x each weather condition (5) x each special event (dozens)

1-C: Exploit Changes



- Sequence of events x[1], x[2], x[3], …
- Two different distributions
- Question: is the observed distribution different than the baseline distribution?

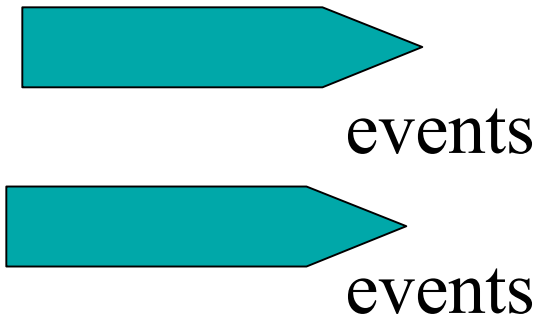
1-D: Use Standards to Lower Cost of Deployment (PMML)

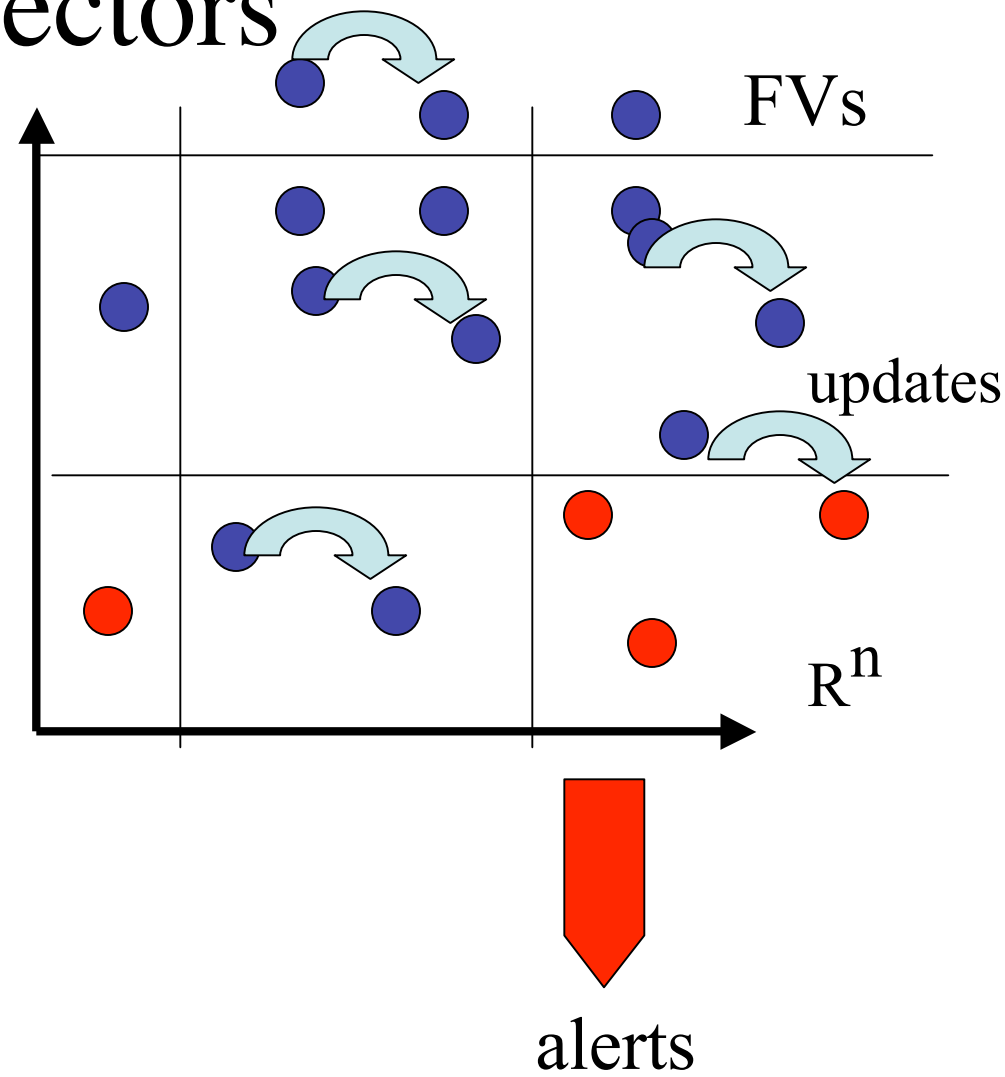
```
<SegmentAssignment
  <SegmentAssignmentField name="latitude" />
  <SegmentAssignmentType name ="regular-partition" />
  <ParameterList>
        <Parameter name = "left-endpoint" value = "0.0" />
        <Parameter name = "right-endpoint" value = "90.0" />
        <Parameter name = "number-partitions" value = "10" />
  </ParameterList>
</SegmentAssignment>
```

Idea 2: What is Event Based Data Mining?

- Events – transactional data about entities of interest
- Features or State Vectors – statistical summaries incorporating derived and aggregated attributes of events
- Updates - each new event updates one or more feature vectors
- Alerts – the results of scoring events and features using statistical and data mining models

Dynamic Updating of Feature Vectors

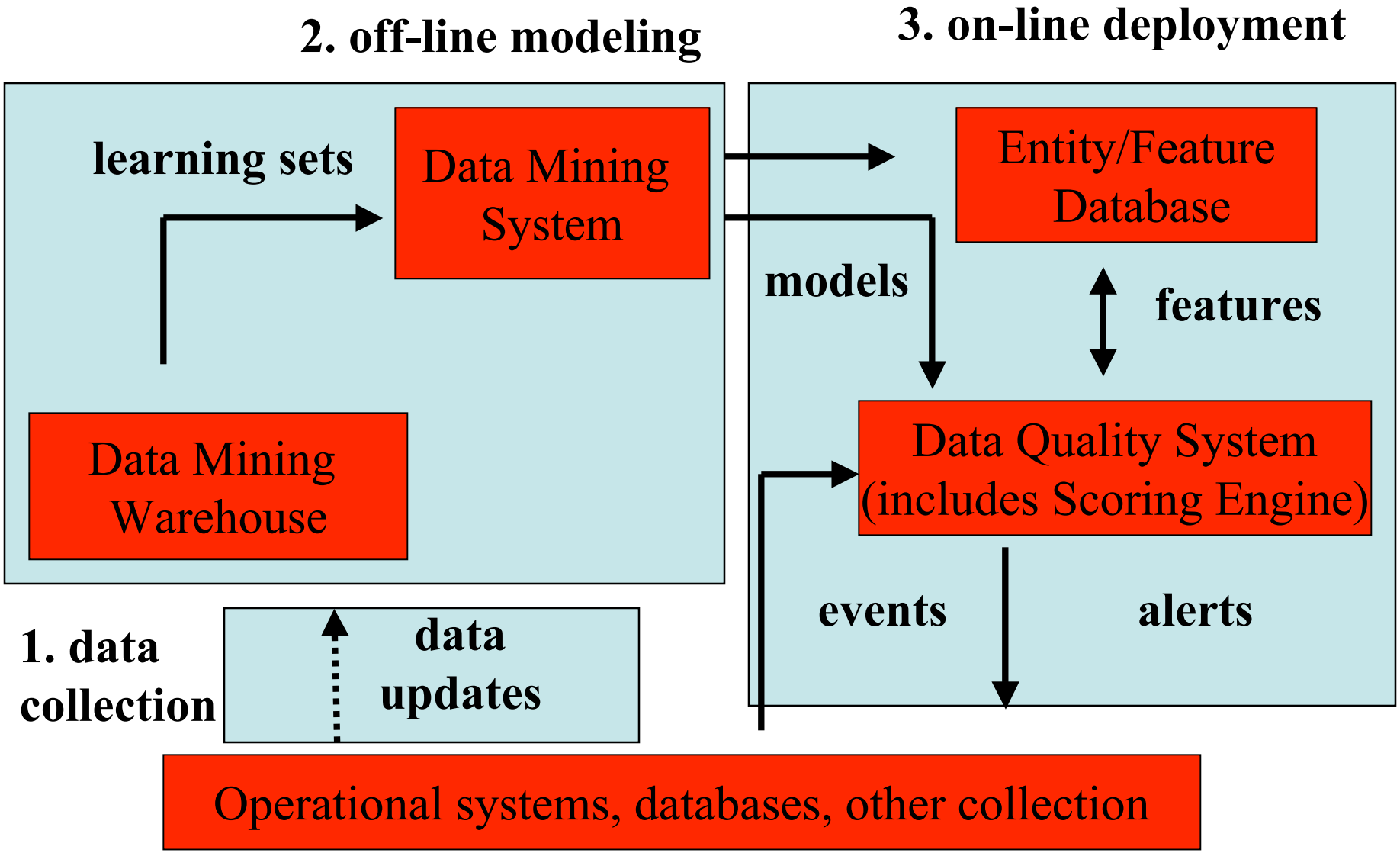




Event loop:

1. take next event
2. update one or more associated FVs
3. rescore updated FVs to compute alerts

Event Based Architecture for DQ



Part 4:
Summary & Conclusion

Summary

- Instead of focusing on accuracy, completeness, consistency, etc. focus instead on **deviations from baselines** for measures defined from these…
- Build **many** baseline models, one for each cell in a data cube…
- We have applied to this methodology to 5 different application areas, all of which are currently in pre-production or production
- Methodology and approach appear promising. Not a lot of alternatives that we are aware of.
- Working with PMML standards group to create standards for baseline models.

For more information:
http://www.opendatagroup.com