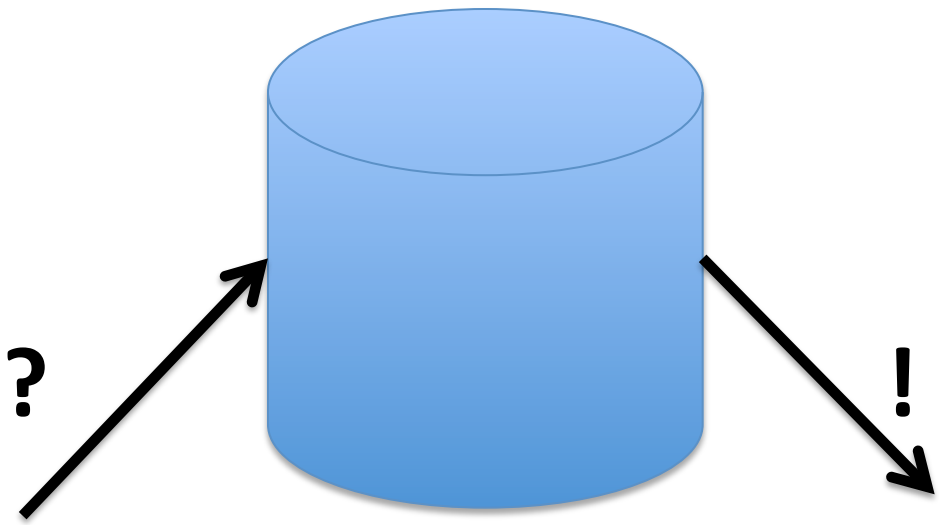
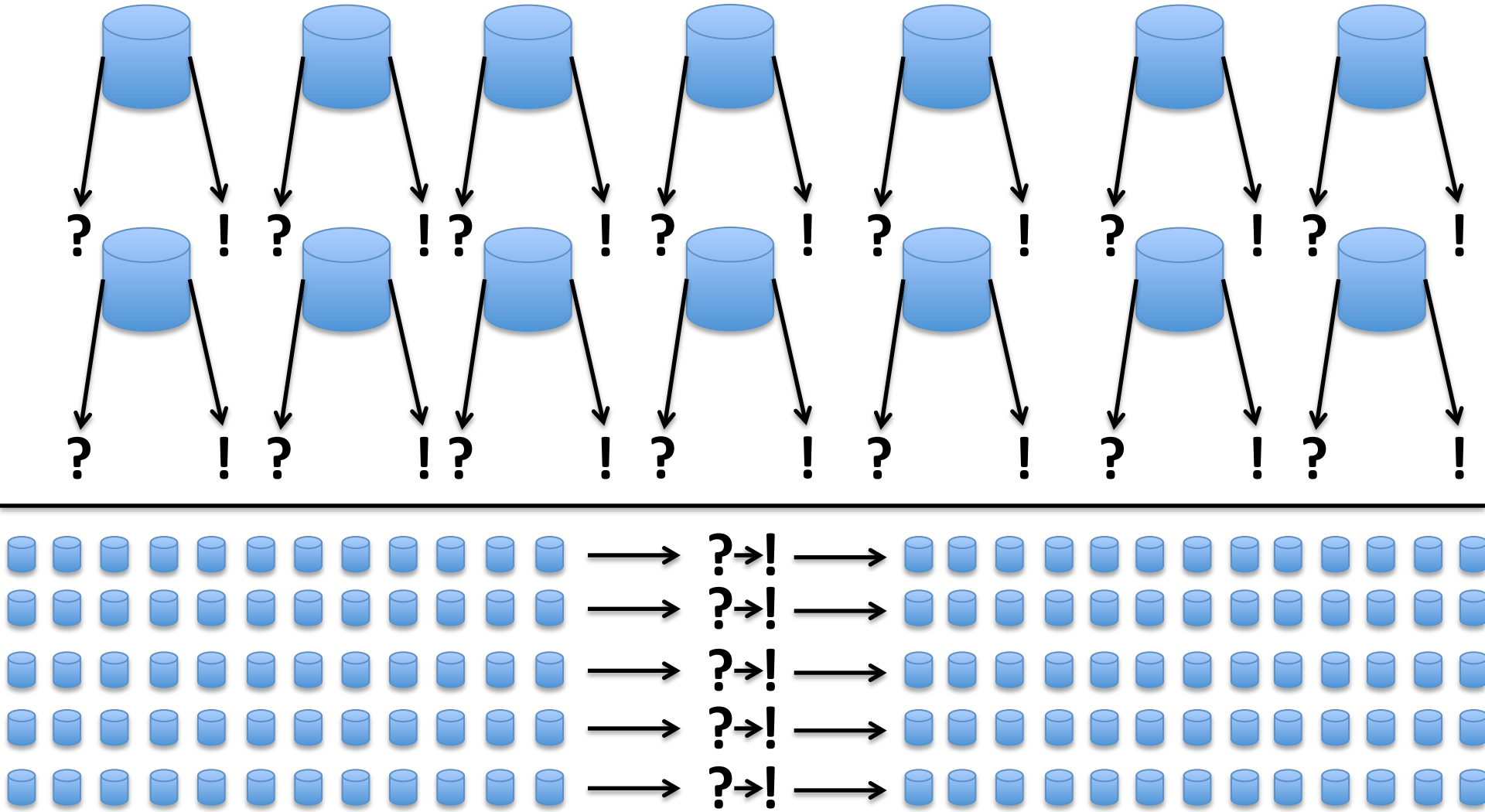
Survey of Streaming Applications using Streams

Scott Schneider
scott.a.s@us.ibm.com
IBM Research

The Old Way

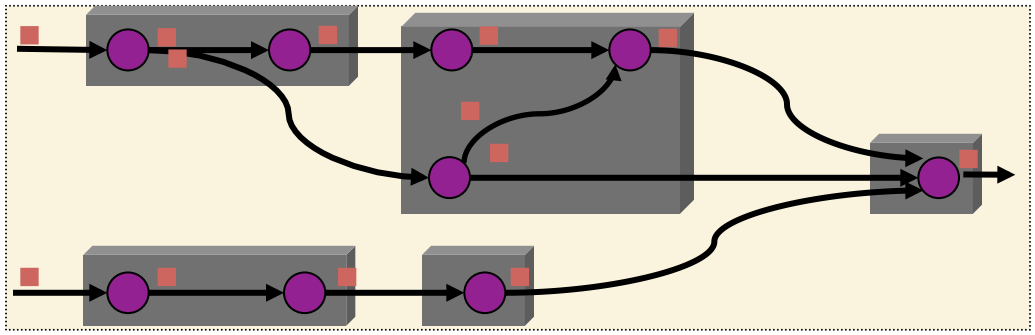


Big Data



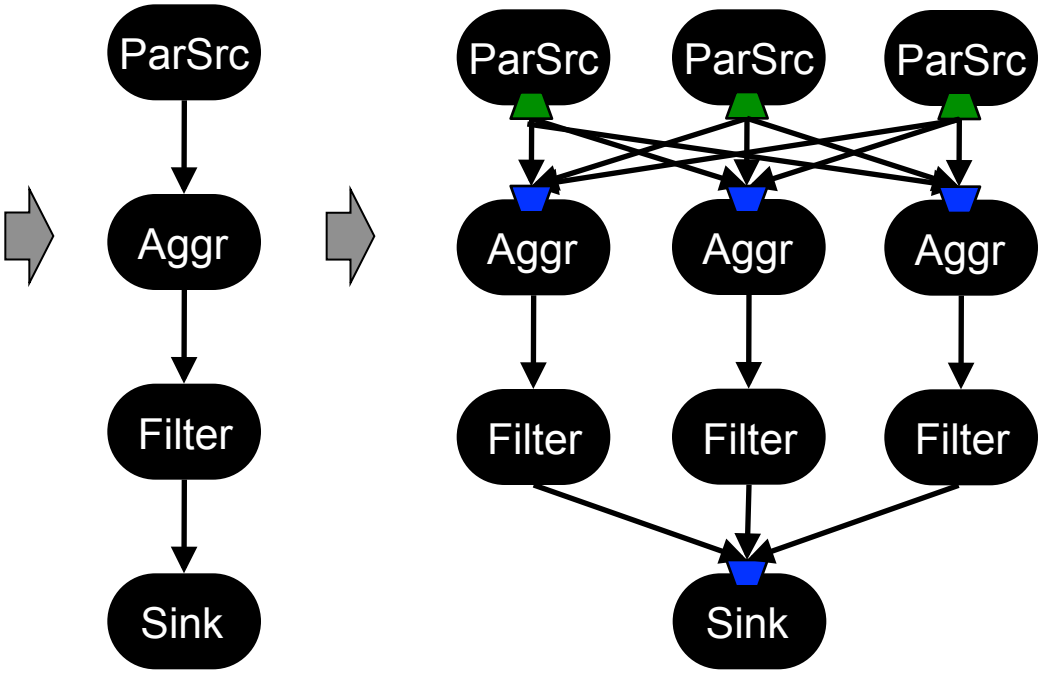
Streams Programming Model

- Streams applications are data flow graphs that consist of:
  - **Tuples**: structured data item
  - **Operators**: reusable stream analytics
  - **Streams**: series of tuples with a fixed type
  - **Processing Elements**: operator groups in execution

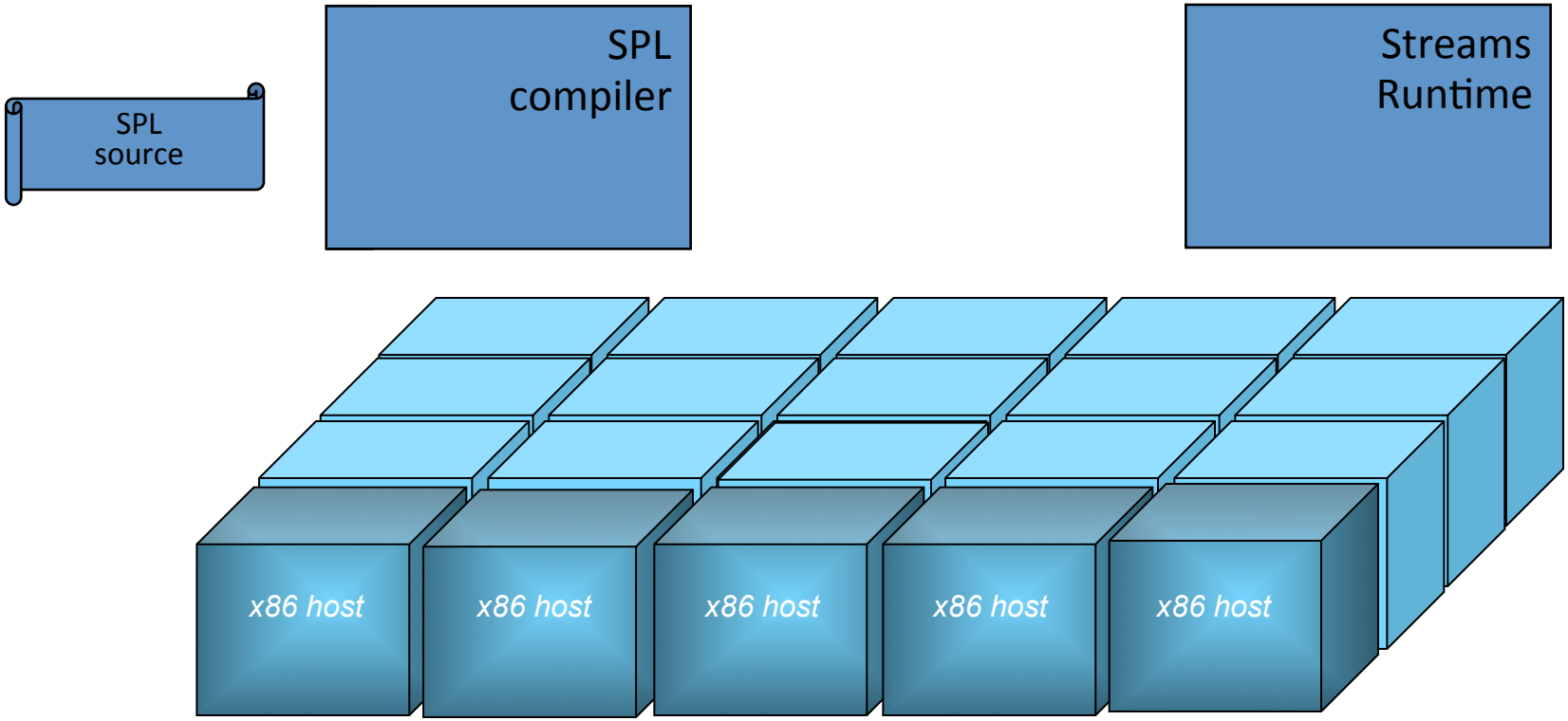


Streams Programming Language

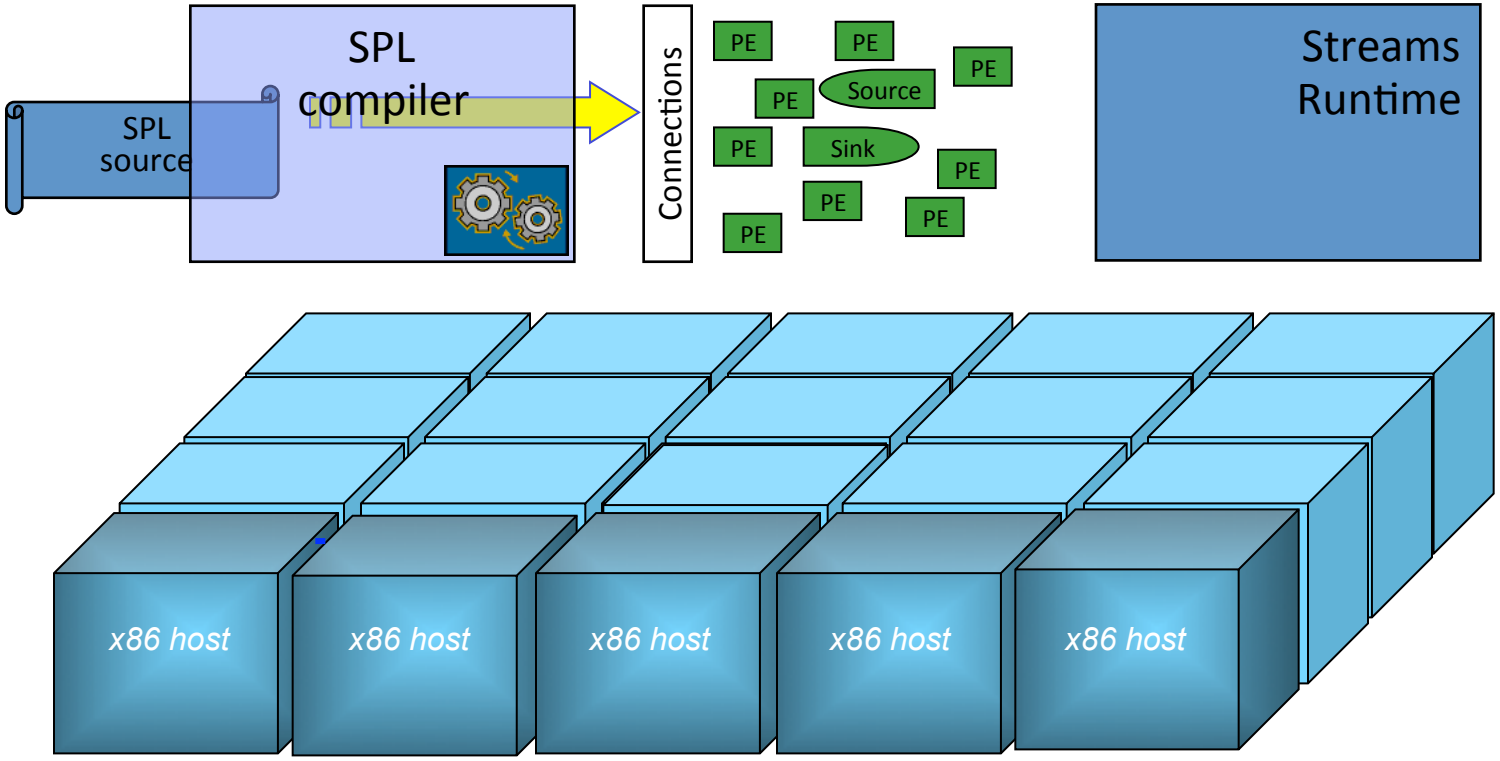
```
composite Main {
type
  Entry = int32 uid, rstring server,
          rstring msg;
  Sum = uint32 uid, int32 total;
graph
  stream<Entry> Msgs = ParSource() {
    param servers: "logs.*.com";
          partitionBy: server;
  }

  stream<Sum> Sums = Aggregate(Msgs) {
    window Msgs: tumbling, time(5),
                 partitioned;
    param partitionBy: uid;
  }

  stream<Sum> Suspects = Filter(Sums) {
    param filter: total > 100;
  }

  () as Sink = FileSink(Suspects) {
    param file: "suspects.csv";
  }
}
```



Source ➔ Compilation ➔ Execution

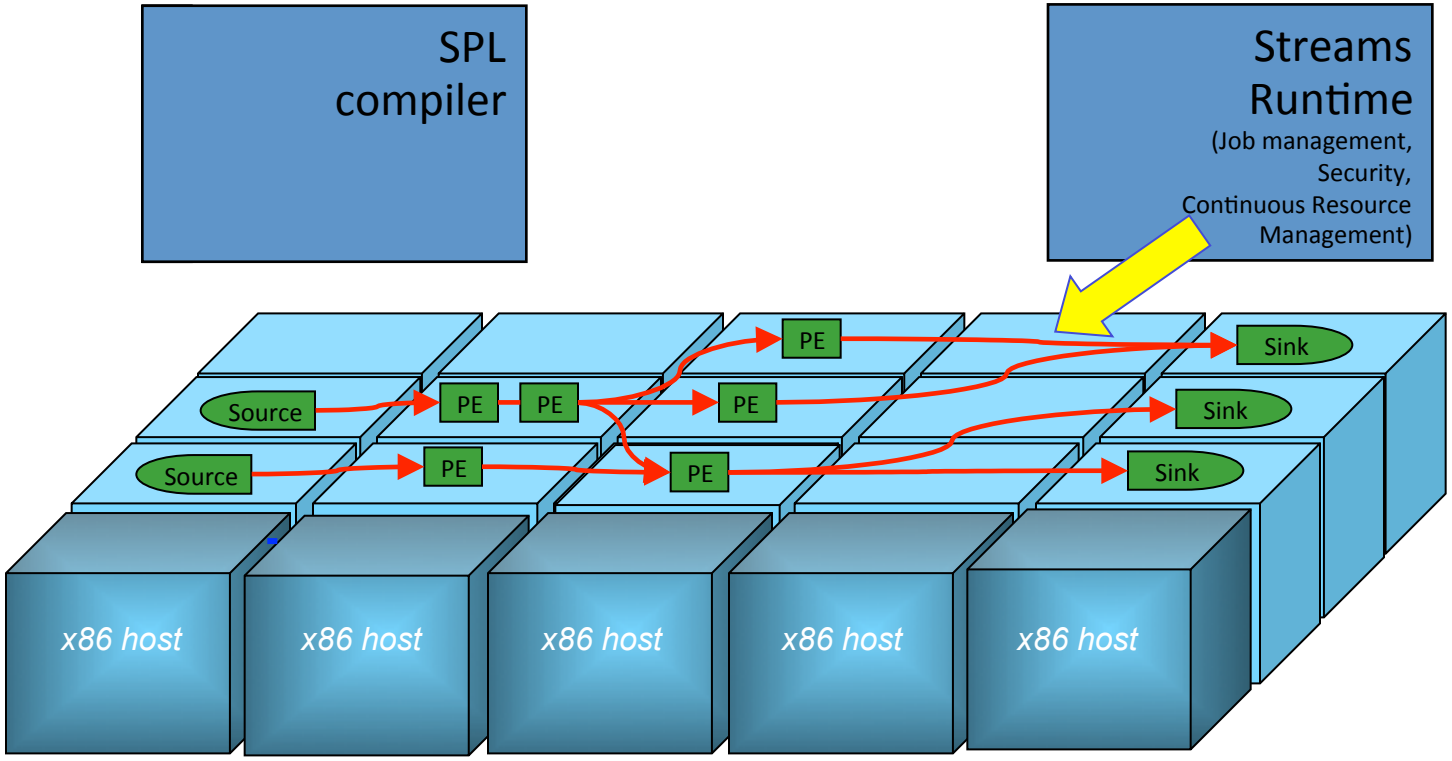


Source ➔ Compilation ➔ Execution



7

Source ➔ Compilation ➔ Execution



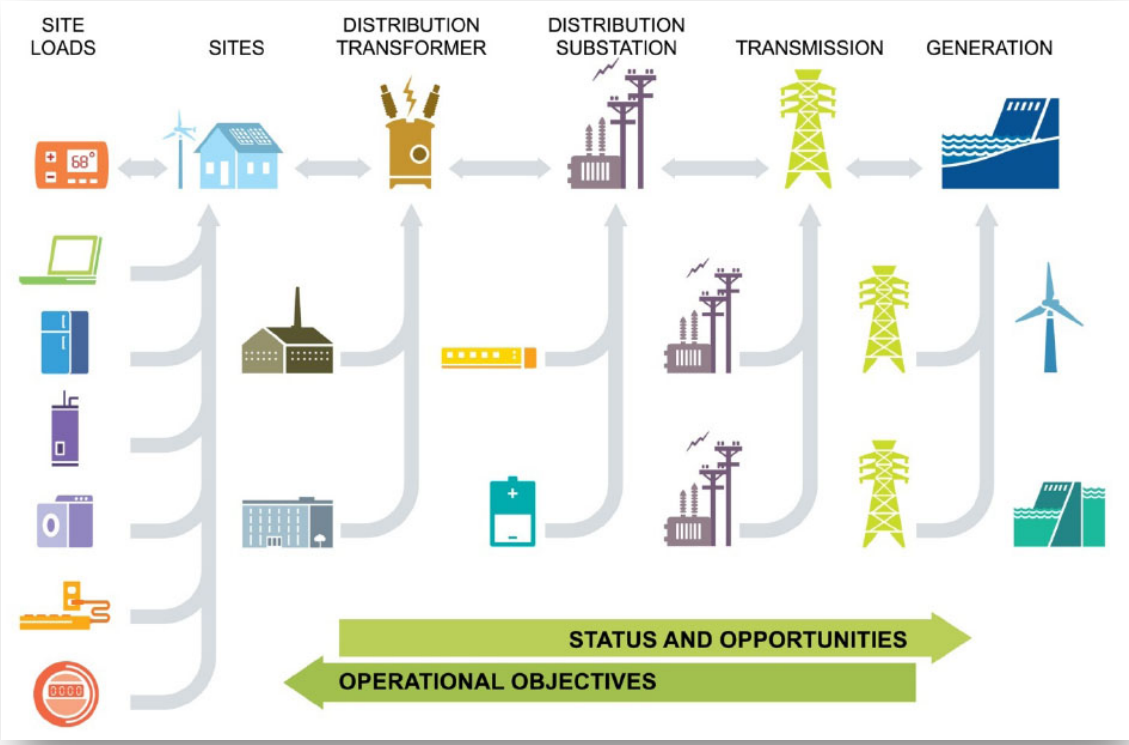
Applications

- Smart Grid
  - Connecting generators, distributors and transmitters
- Social Media Analytics
  - Sentiment analysis of movie marketing
  - Disease tracking
- Medical
  - SickKids: neonatal monitoring
  - ANS modeling

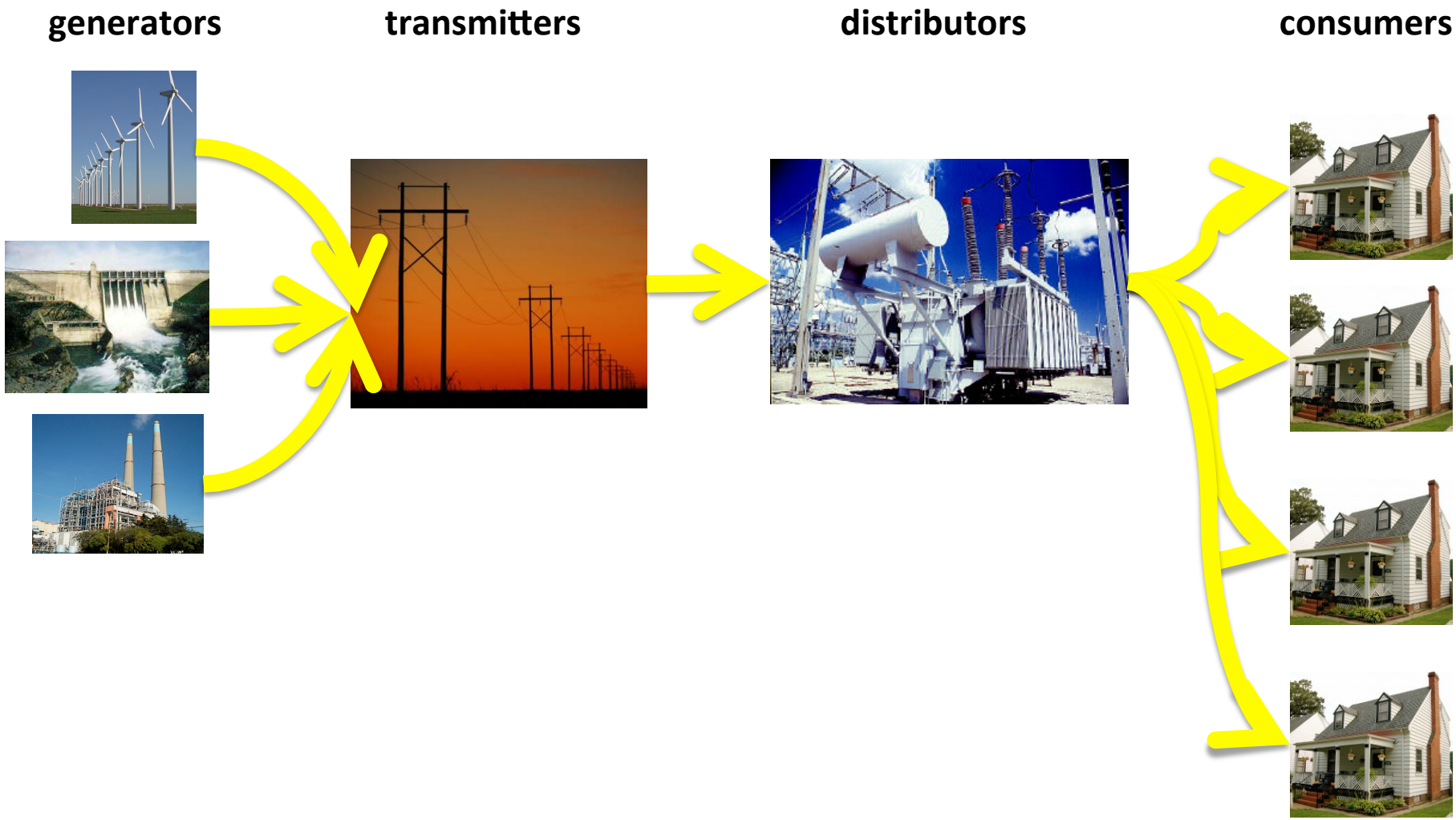
Smart Grid

Pacific Northwest Smart Grid

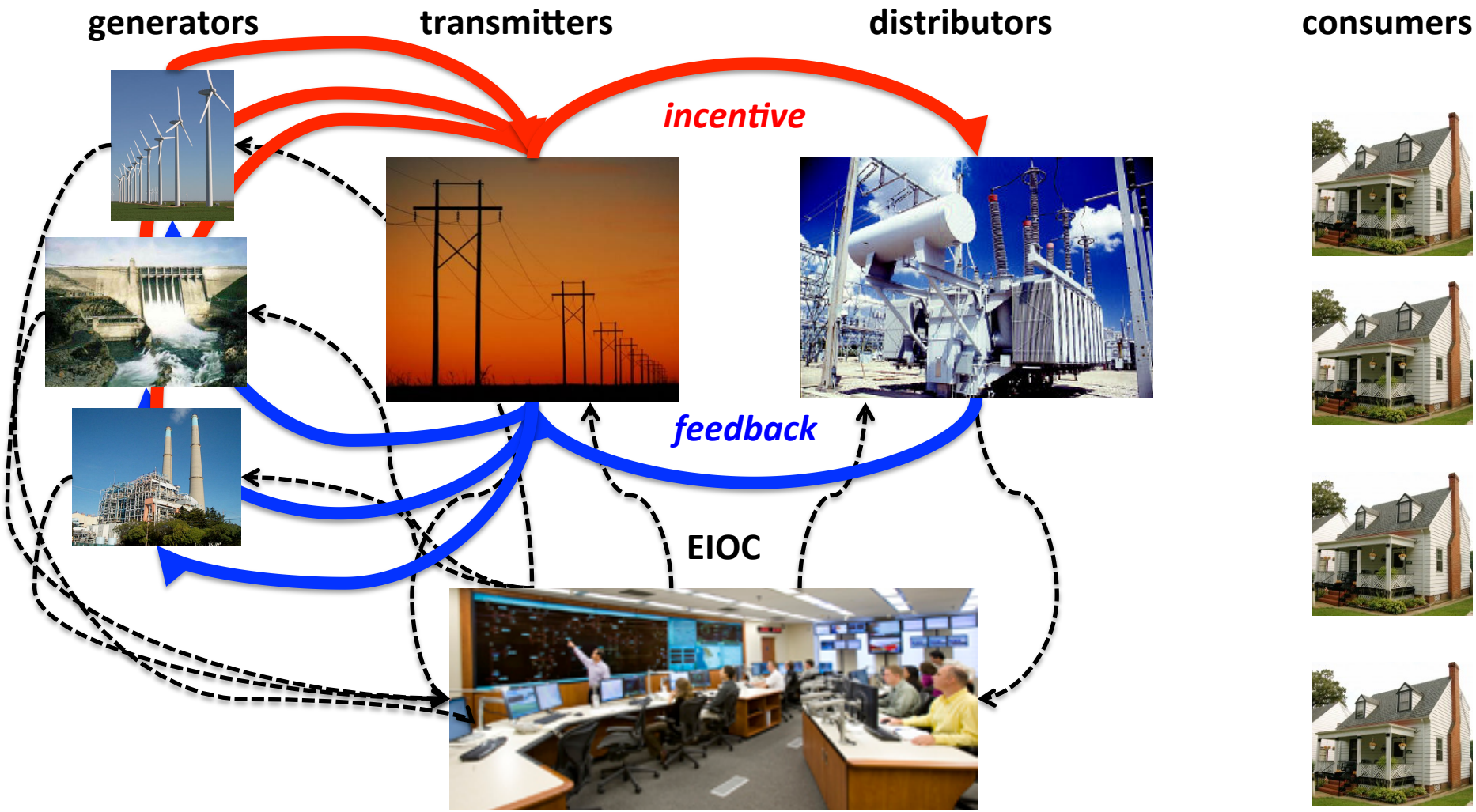
- Goals
  - Manage peak demand
  - Integrate renewable energy sources
  - Address constrained resources
  - Reliablity, effeciency
  - Dynamically choose most econmical resource



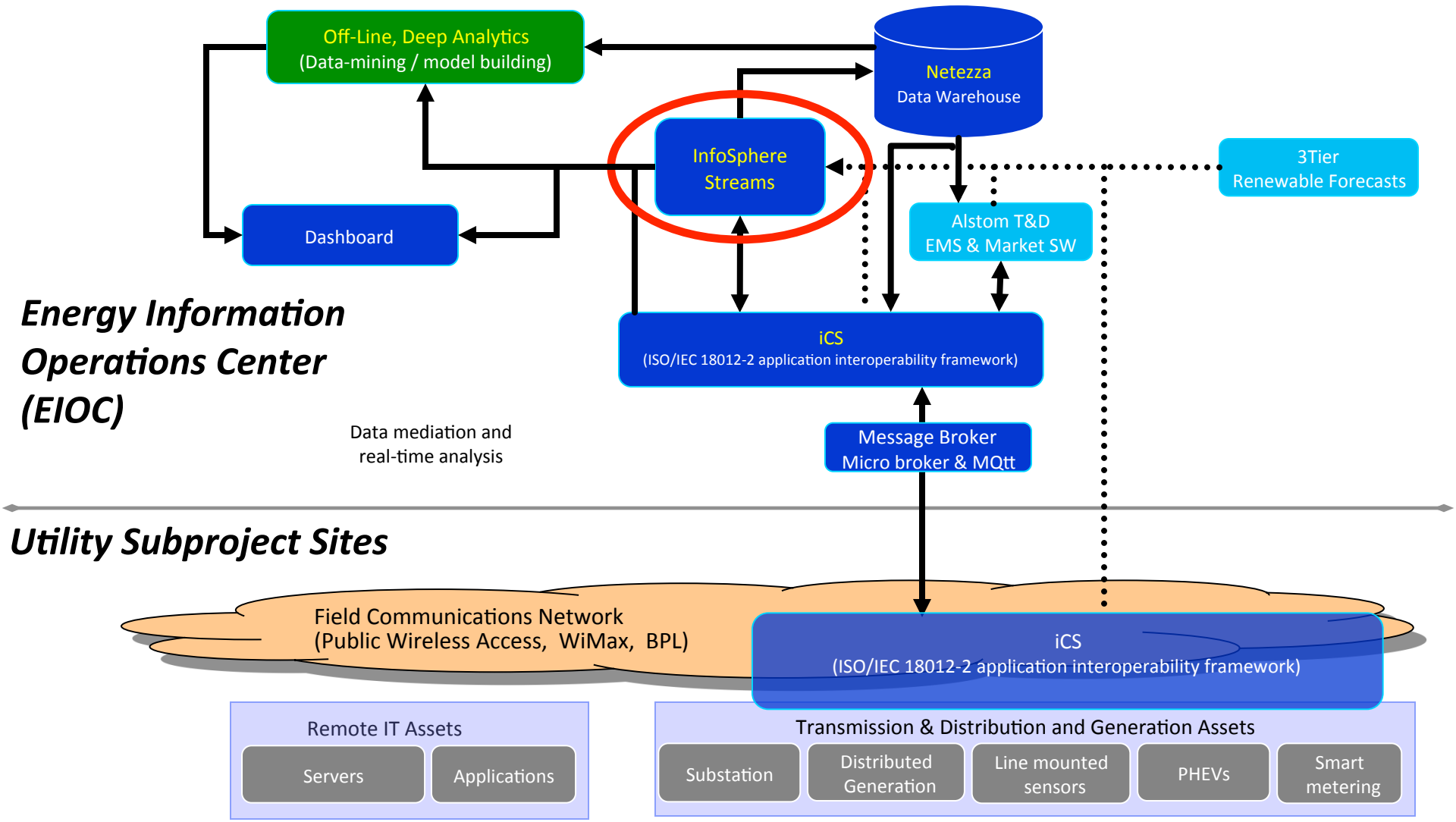
Power and Demand Flow



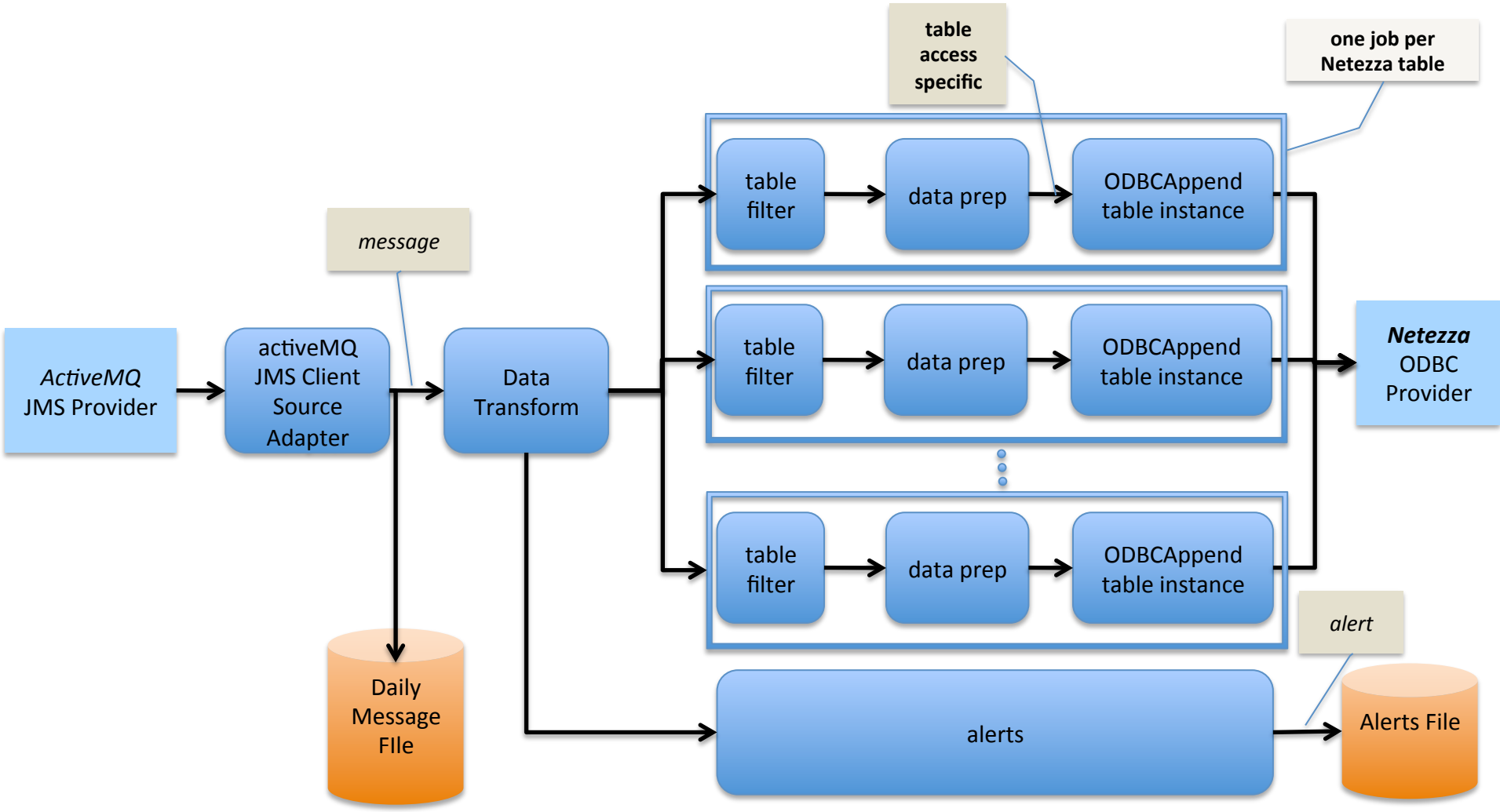
Transactive Control Flow



Transactive Architecture



High Speed Data Mediation

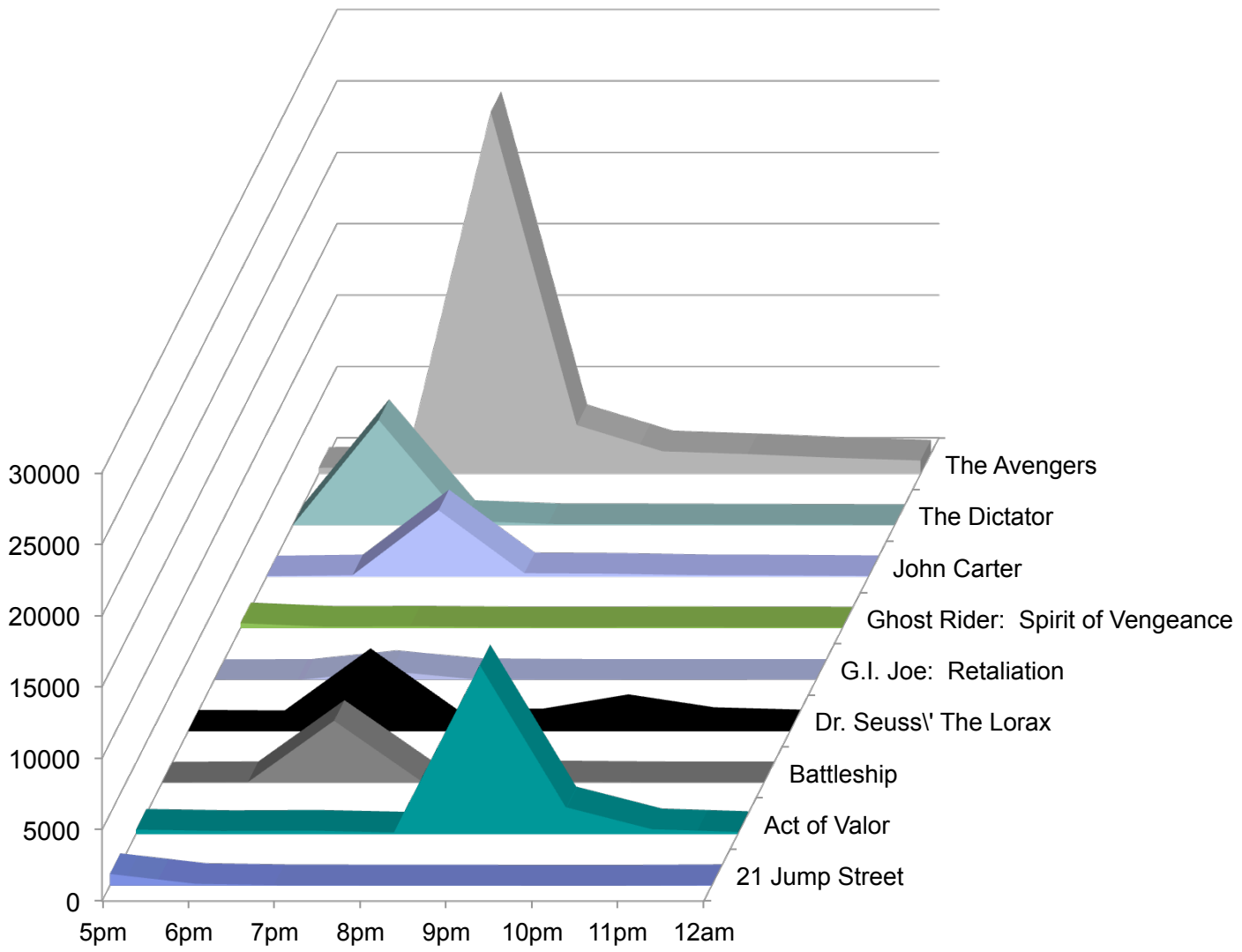


Social Media

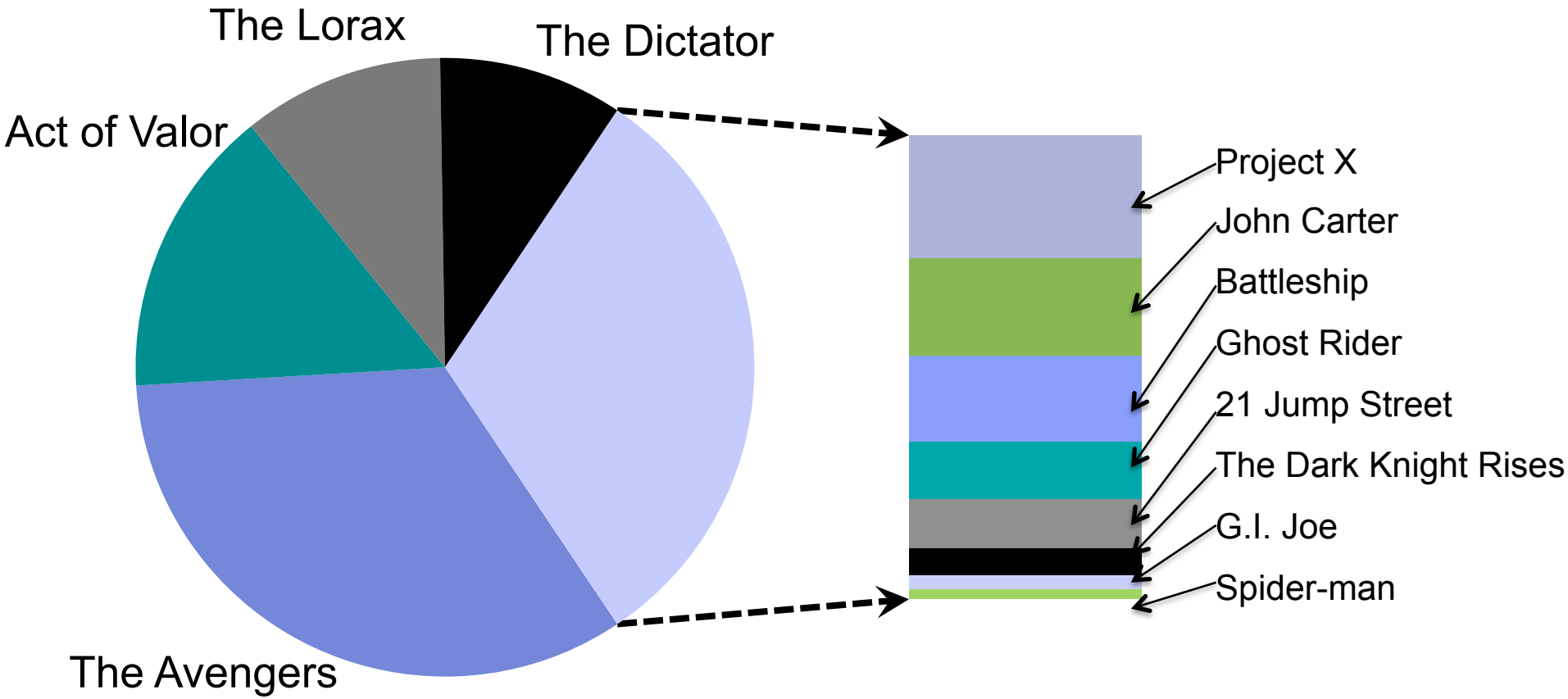
Super Bowl Movie Trailers

- Effectiveness
  - How many people are talking about the movies?
  - Who are they?
  - Do they intend to see it?
  - How does this compare to other movies?
- Big Data
  - Over 1 billion messages
  - Over 30 million profiles extracted

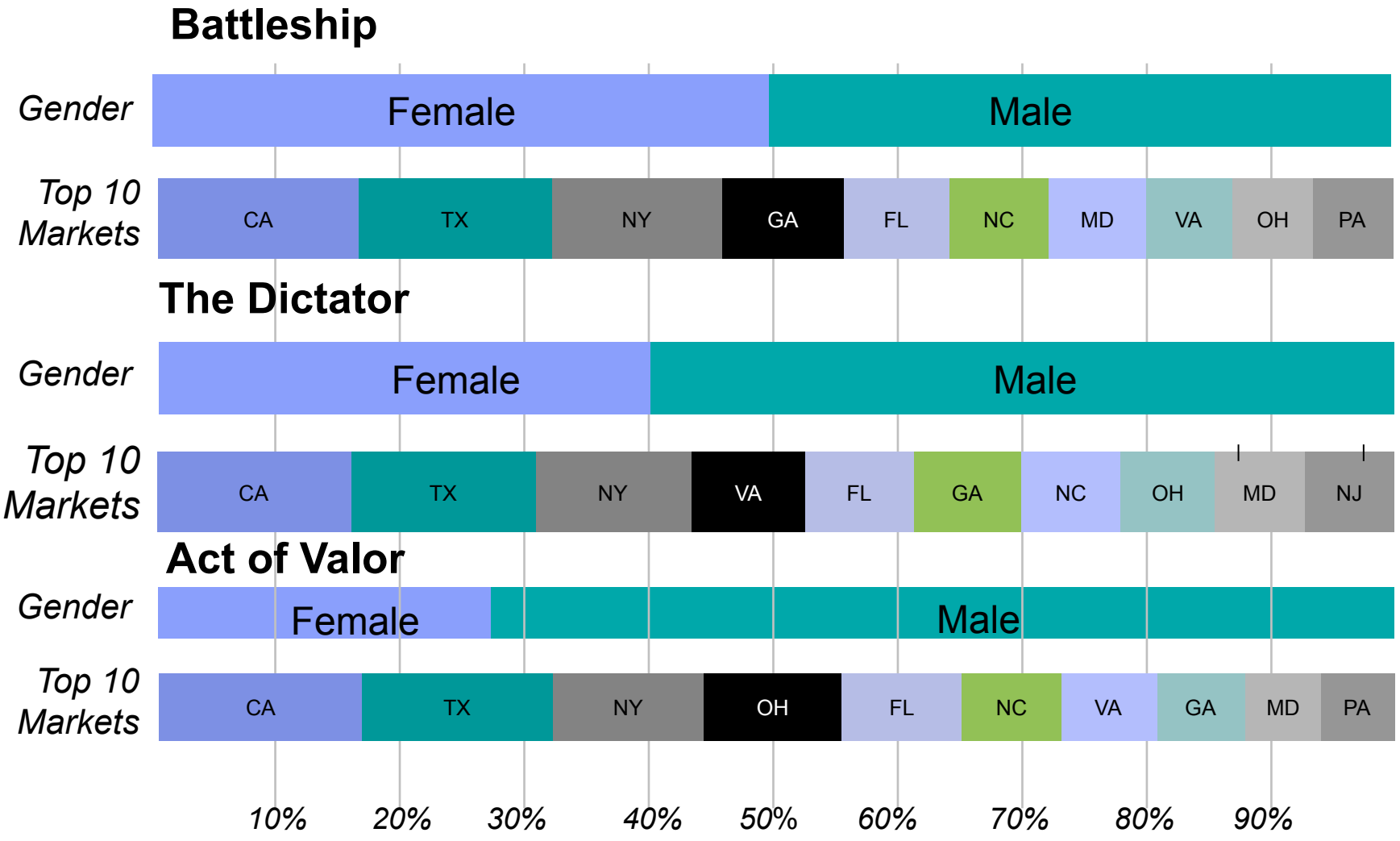
Buzz During Super Bowl



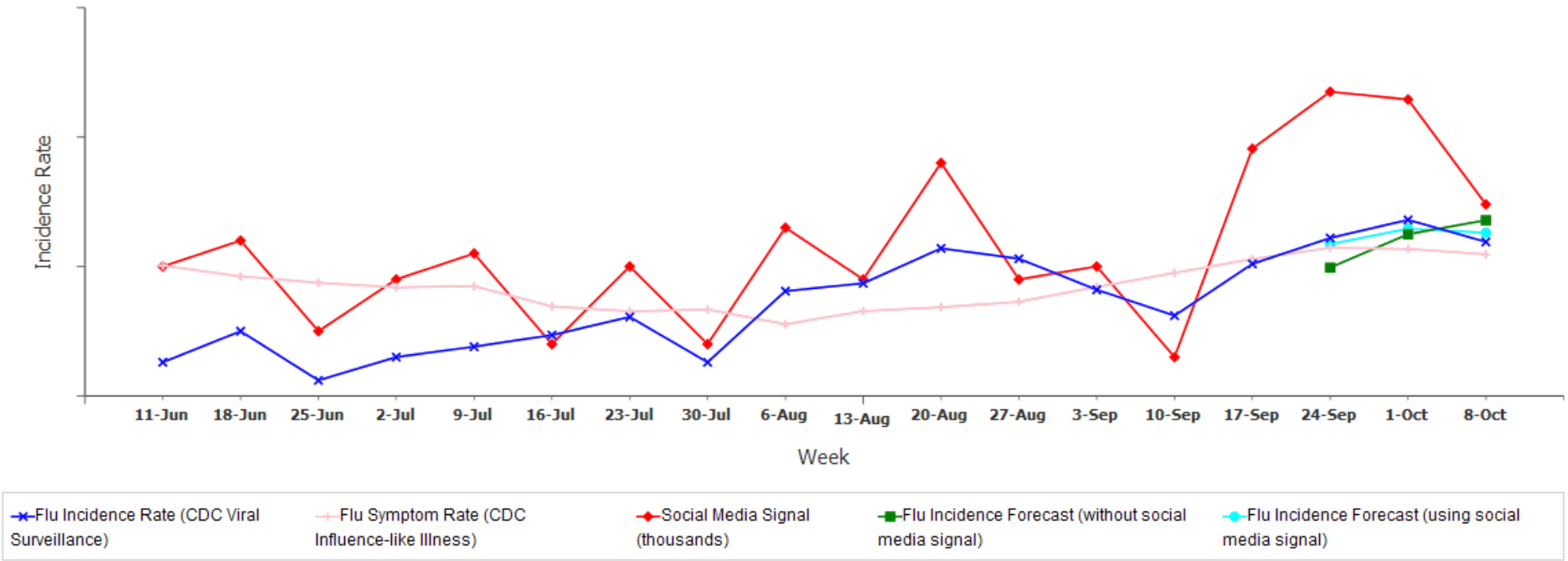
Avid Movie Goer Buzz During Super Bowl



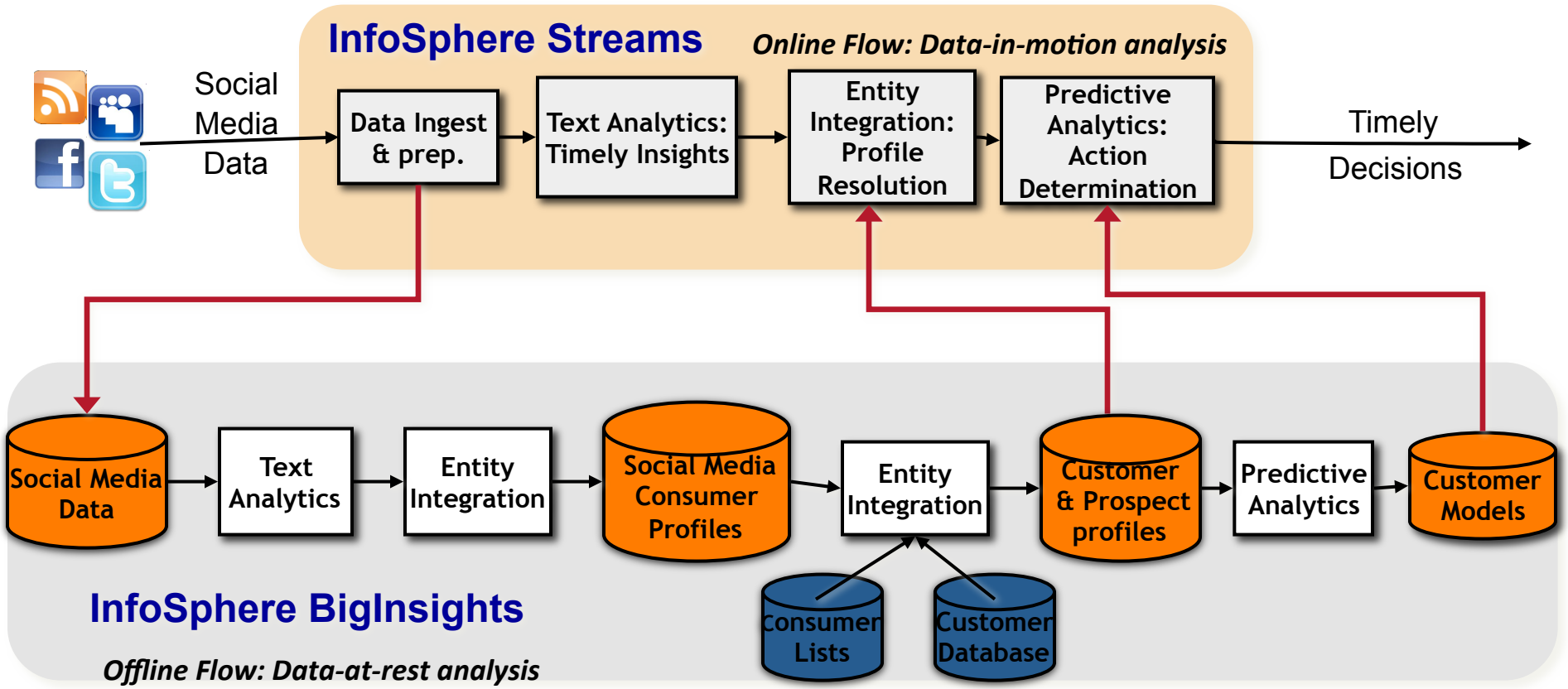
Buzz by Demographics



Disease Tracking



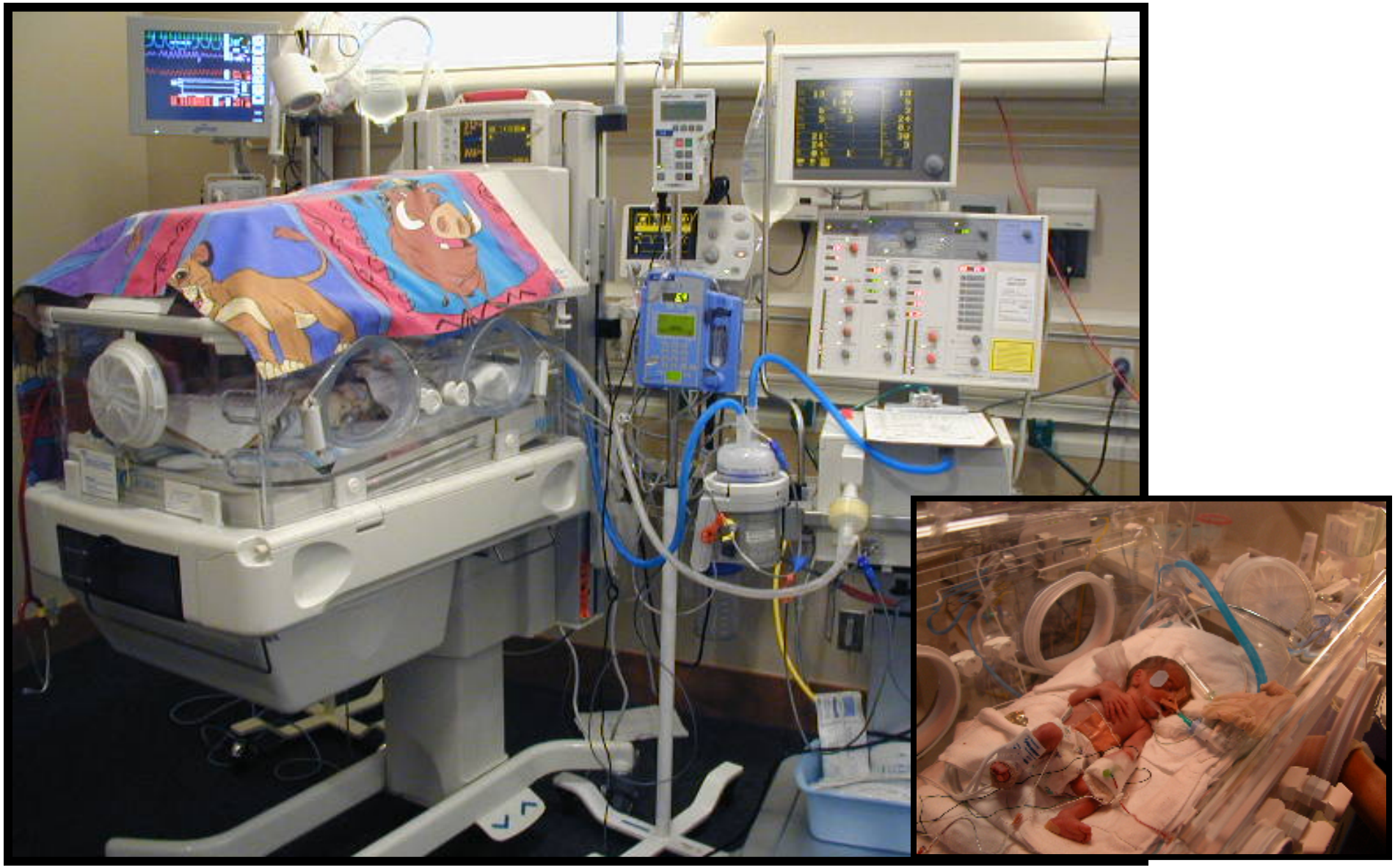
Social Media Analytics Architecture



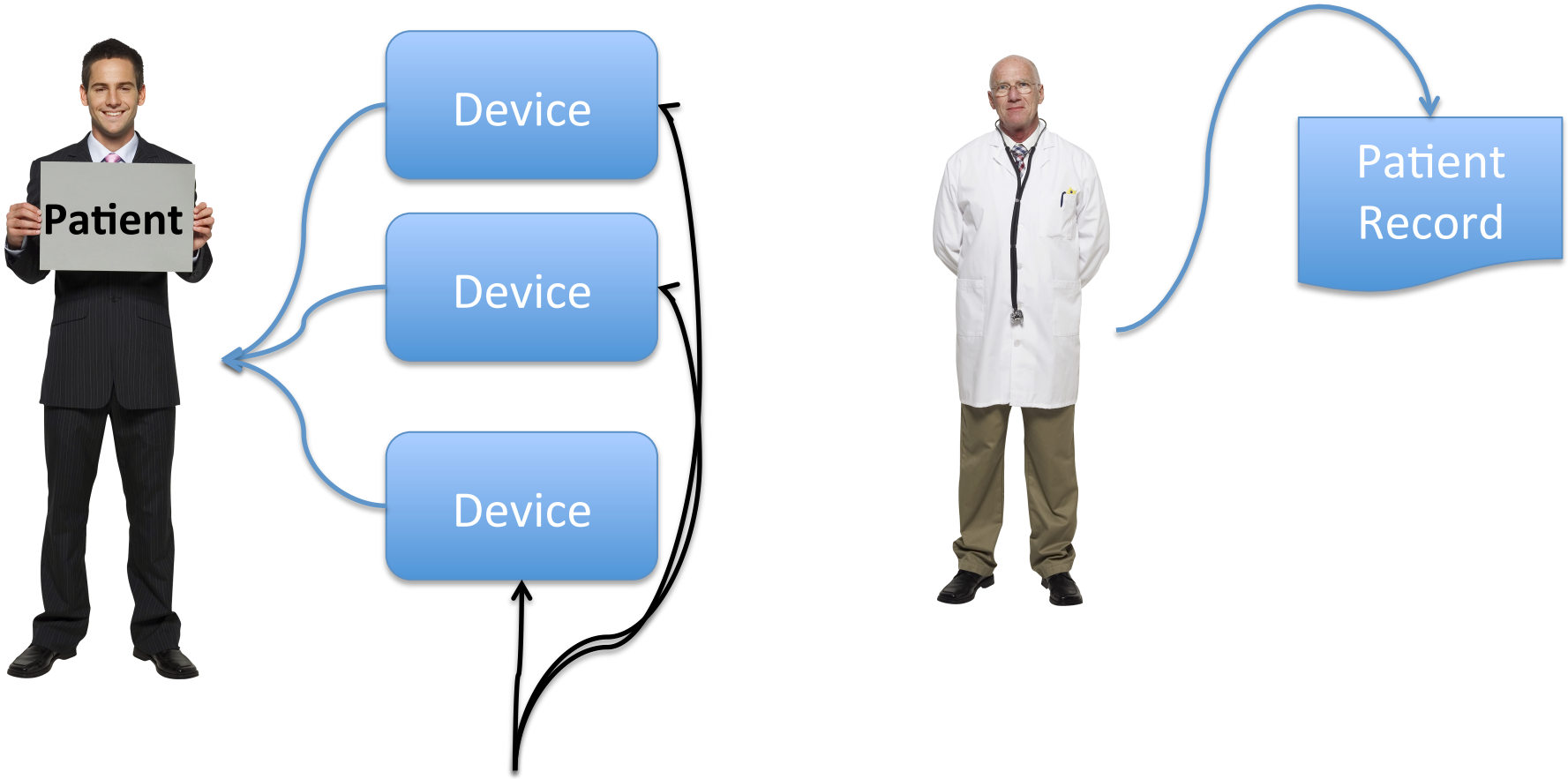
Medical

"Data Baby"

Real Neonatal ICU



Current ICU Monitoring

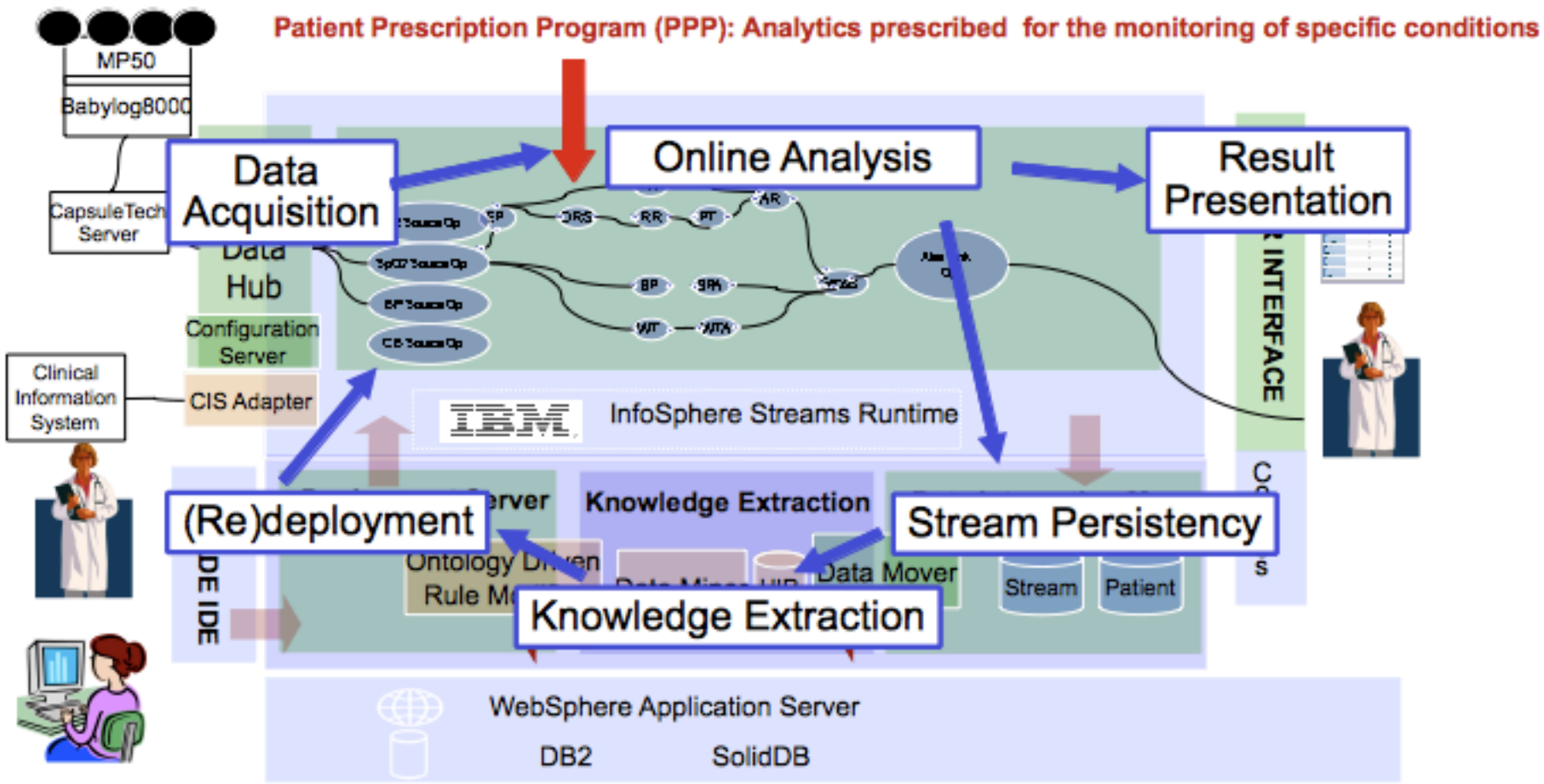


Fixed, threshold based alerts
(e.g.: heart rate below 45 bpm)

What's Missing

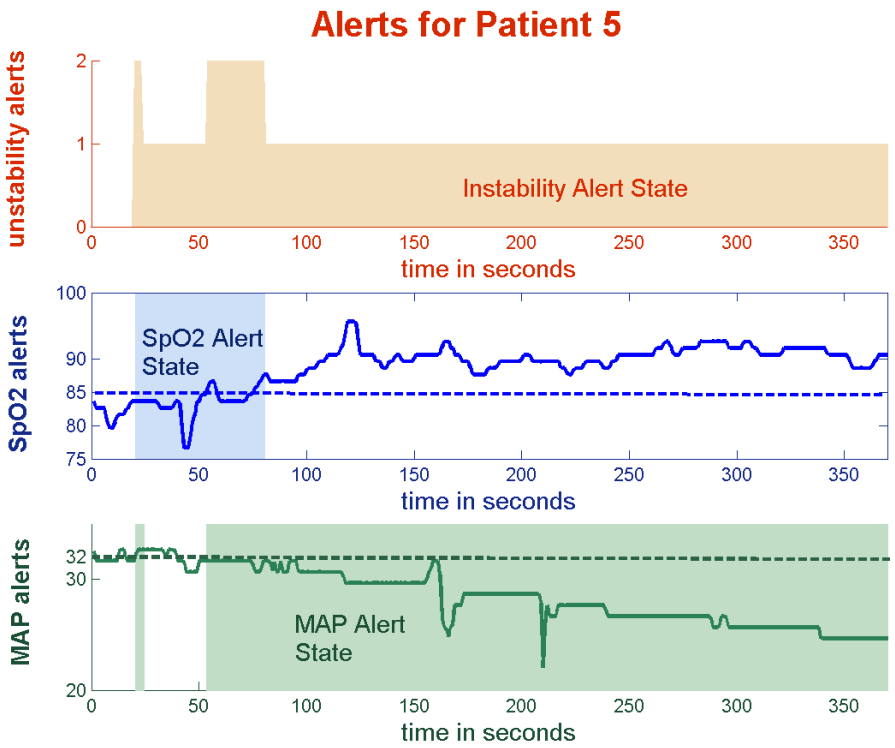
- Automation
- Data retention
- Complex temporal correlations
  - e.g., Alert when both SpO2 and blood pressure drop below a threshold for 45 seconds
  - e.g., Alert when variability of the HR is below a threshold for a 30 minute window
  - e.g., Alert when probability of observing bad event in next 6 hours is high
- Further exploration of physiological data

SickKids Deployment

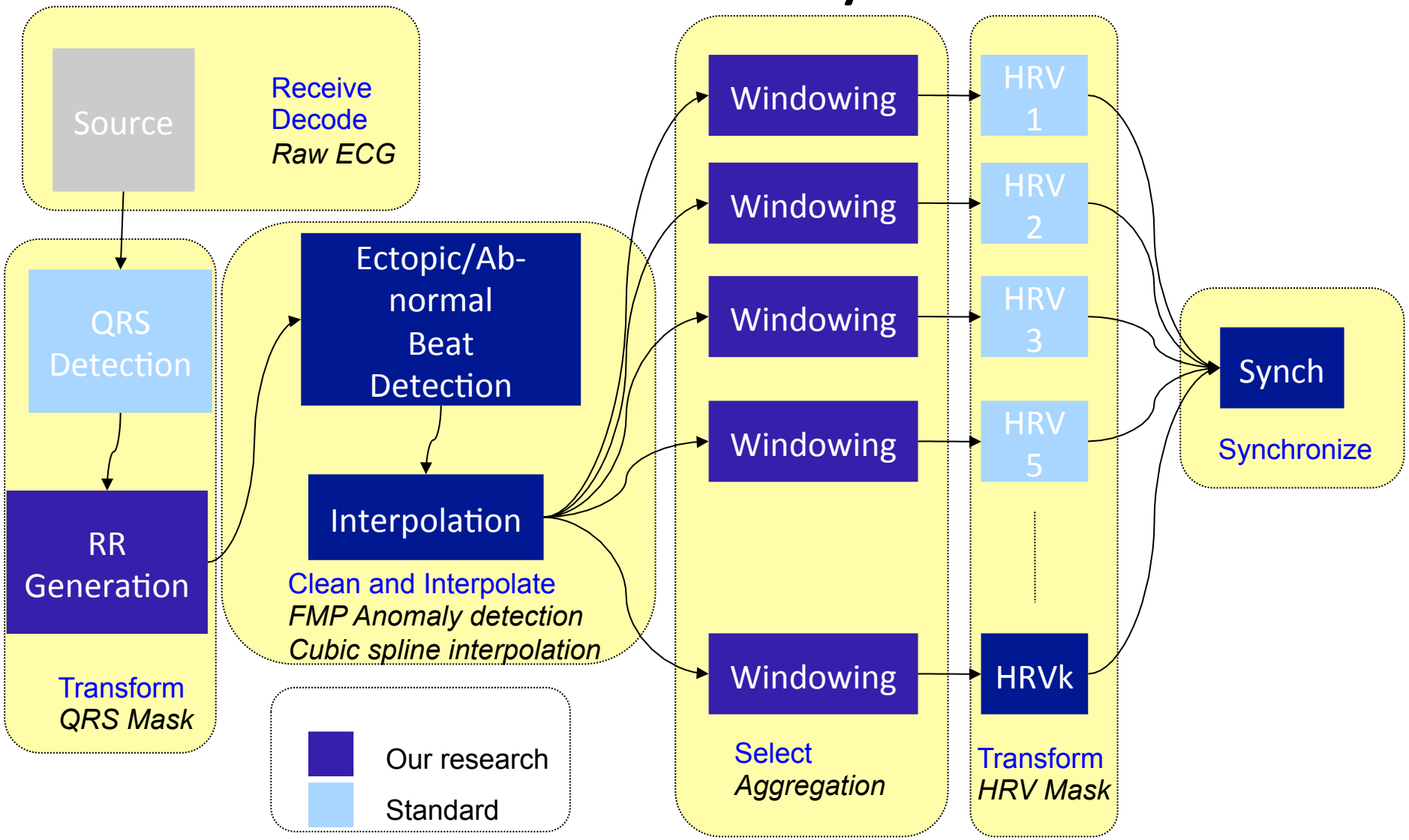


Simple Example

- "Baby crashing"
  - Correlate SpO2 with mean arterial pressure
    - SpO2 < 85
    - Mean BP < gestational age
  - For 20 seconds



Not So Simple Example

- Detecting onset of sepsis in premature infants
- Data sources
  - Systolic, diastolic and mean arterial blood pressure
  - SpO2
  - Respiration rates
  - Electrocardiograms
  - Static clinical data
- Fusion and scoring
  - Based models specified by physicians
  - Data driven approach uses machine learning to discover new models
- Deployed at SickKids hospital in Ontaria, Canada monitoring NICU patients

Modeling the ANS with Heart Rate Variability