UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Statistical Modeling of Large-Scale Sensor Count Data

Padhraic Smyth
Department of Computer Science
University of California, Irvine

UCIrvine
University of California, Irvine

Acknowledgements

- Collaborators:

Jon Hutchins

Alex Ihler

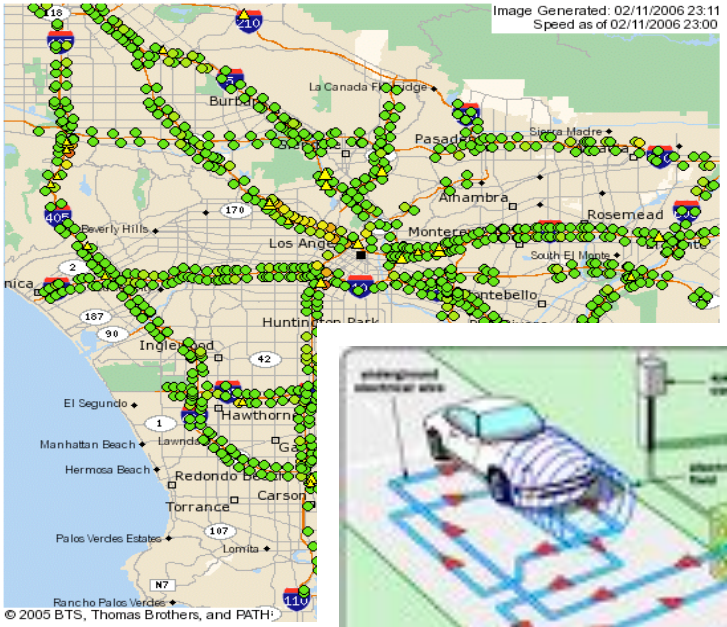
- Funding:
  - US National Science Foundation
  - California Department of Transportation

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Sensors Measuring Human Activity

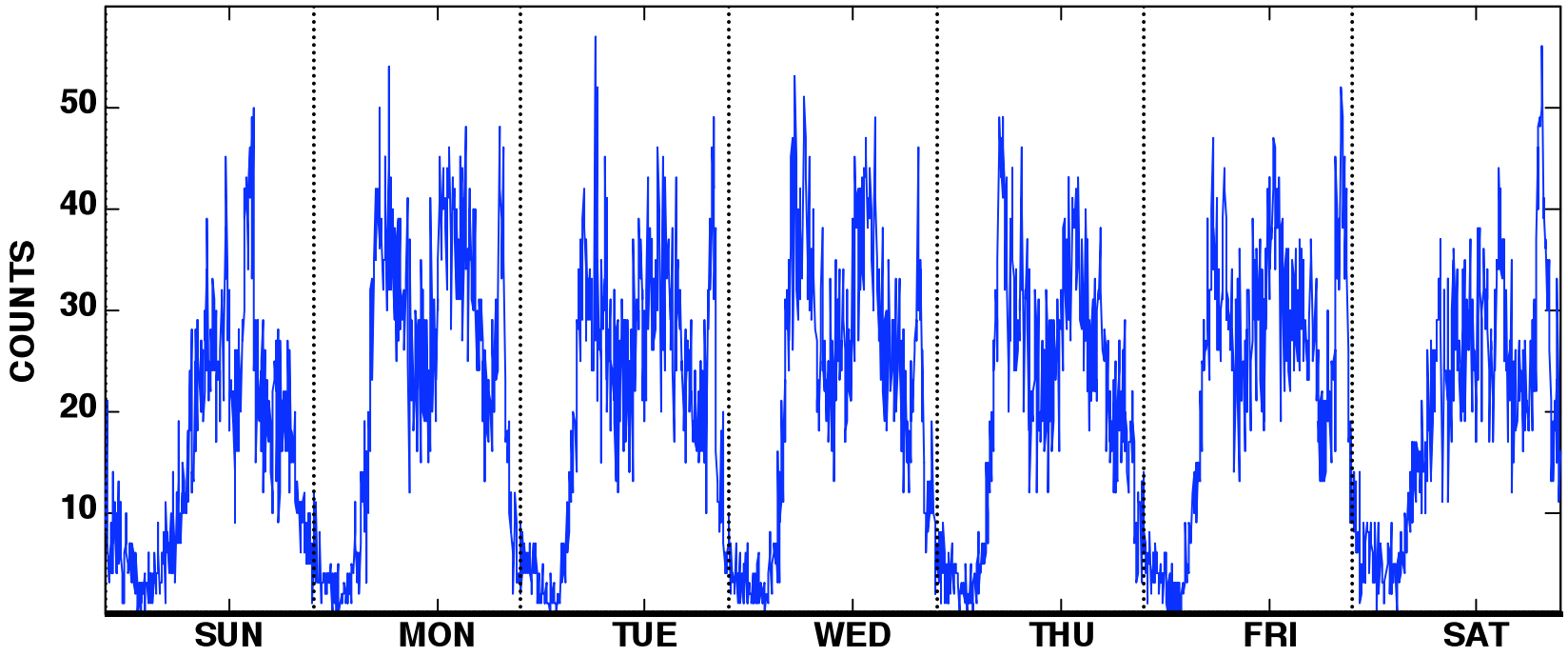
Optical people counter at a building entrance on campus

Loop sensors on Southern California freeways

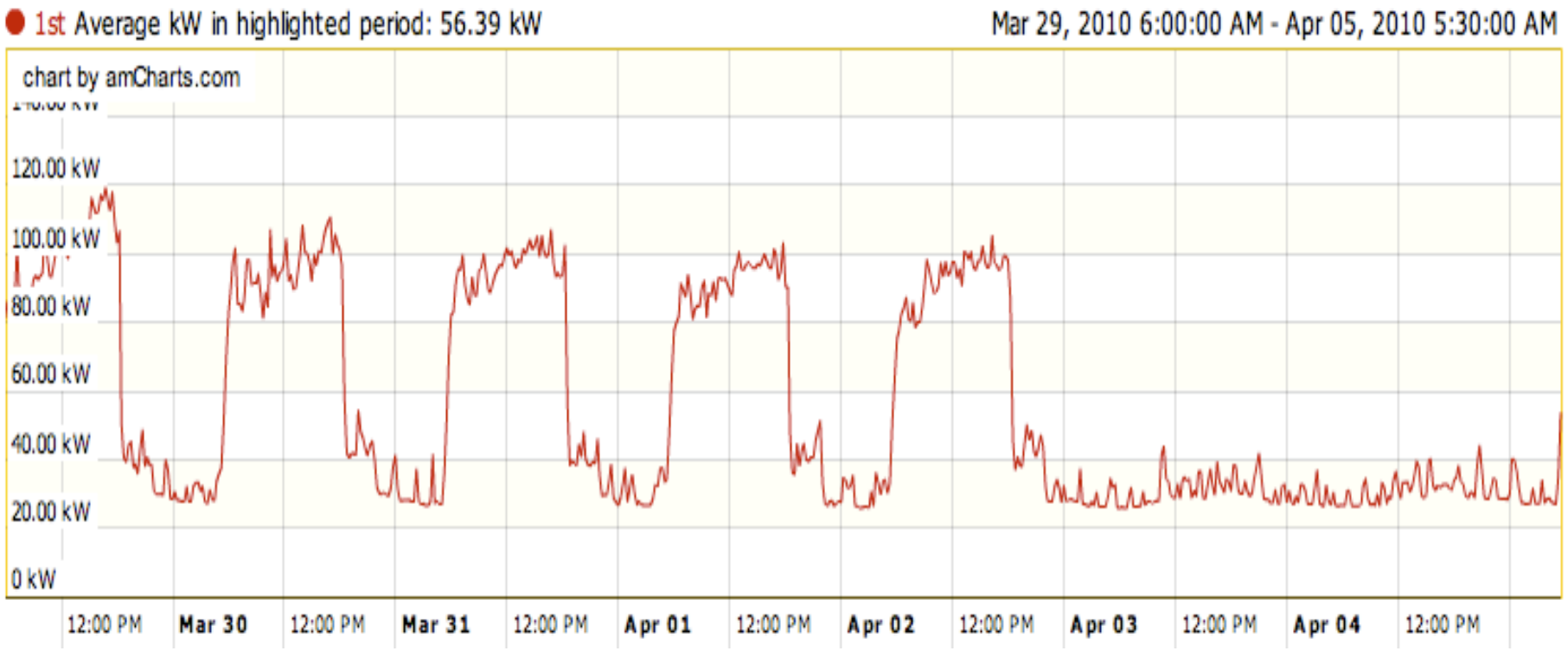


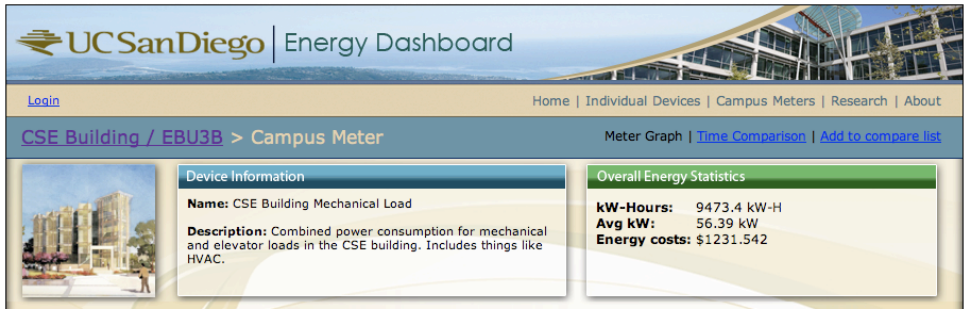
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

One Week of Traffic Data



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

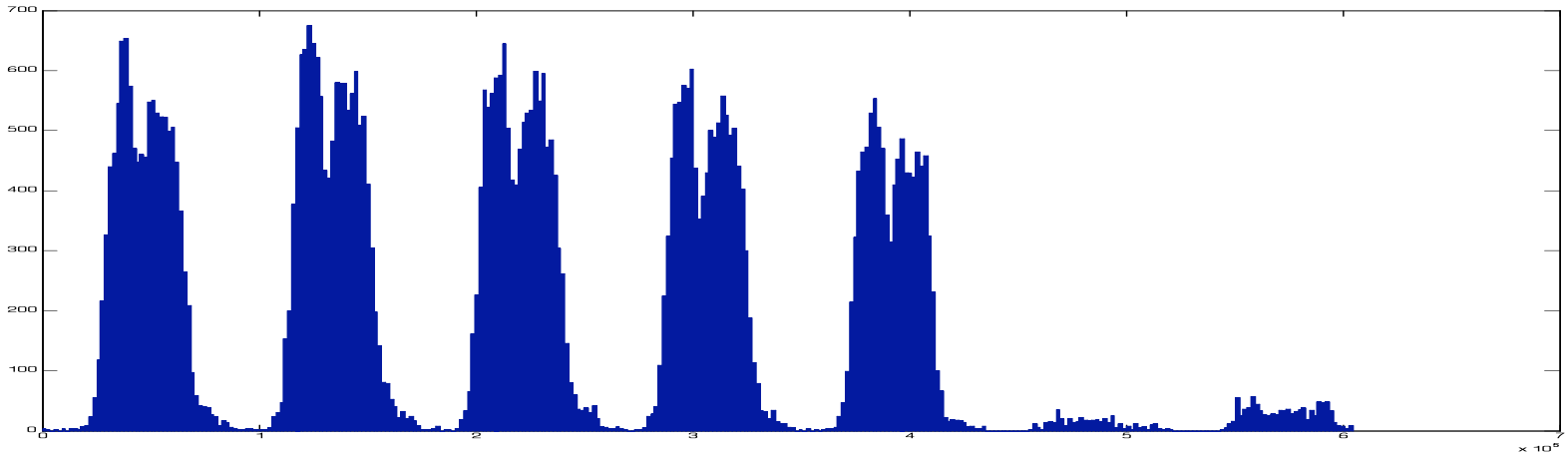




UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Email Activity over 1 Week

number of emails sent



Time

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

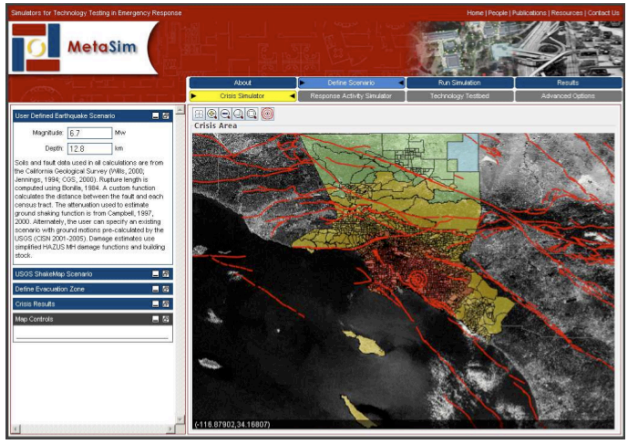
Research Problems

- Detection of anomalous events
- Inference of “typical behavior”
- Occupancy and flow estimation
- Prediction
- Missing data inference
- ….and more

UCIrvine
University of California, Irvine

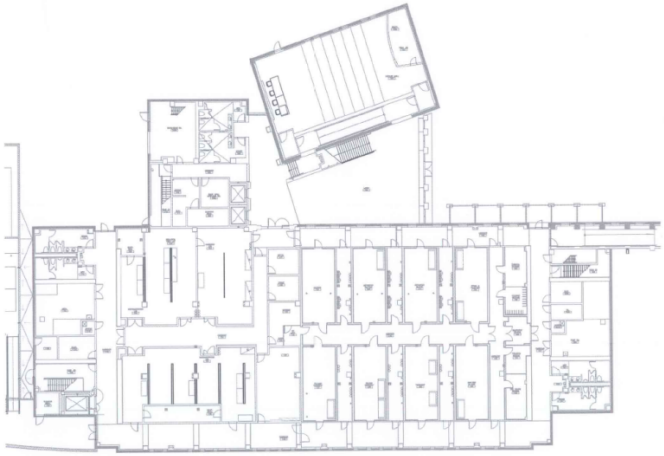
Applications

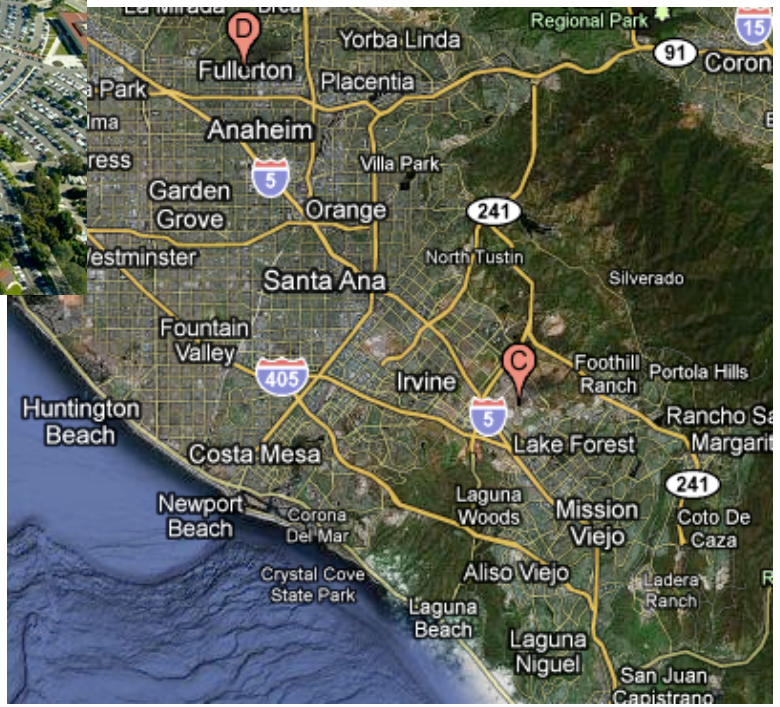
- Traffic engineering
- Urban planning
- Energy monitoring
- Disaster response



UCIrvine
University of California, Irvine

Spatial Scale





Urban Sensing

- Combining multiple data sources
  - Satellite and aerial images (e.g., MODIS)
  - Energy usage (e.g., electricity)
  - Road sensors
  - Network traffic
  - Search engine data
  - Cell-phone data
  - Census data
  - ……

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Outline

- Graphical models for count time-series
- Experimental results
  - Detecting events in traffic data and building sensor data
  - Estimating building occupancy
- Ongoing work and conclusions

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

General Context

- Aggregated counts at discrete times (people, vehicles, emails, etc)
- Data has underlying "signal"
  - e.g., daily/weekly rhythms of human behavior
  - occasional "bursts" of counts related to events
- Large amounts of data (e.g., years)
- …but little prior knowledge
  - detailed time patterns of human behavior are unknown
  - algorithm has no knowledge of scheduled events

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Related Work

- Markov-modulated Poisson processes
  - Heffes and Lucantoni (1994)
  - Scott (1998)
  - Scott and Smyth (2003)

- Segmentation with Markov/Poisson models
  - Kleinberg (2002)
  - Salmenkivi and Mannila (2005)

- Statistical modeling of traffic
  - Bickel et al, *Statistical Science*, 2007 + related work

Points to make re: Vehicle Data

- Volume of data (# sensors x # measurements)
- Basic challenges and difficulties (missing, bad data)
- Why is this interesting to machine learning?

UCIrvine
University of California, Irvine

More points to make

- Why traffic data is important
  - Air quality and the environment
  - Transportation and planning
  - Real-time prediction
  - Census correlations
  - Economics

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

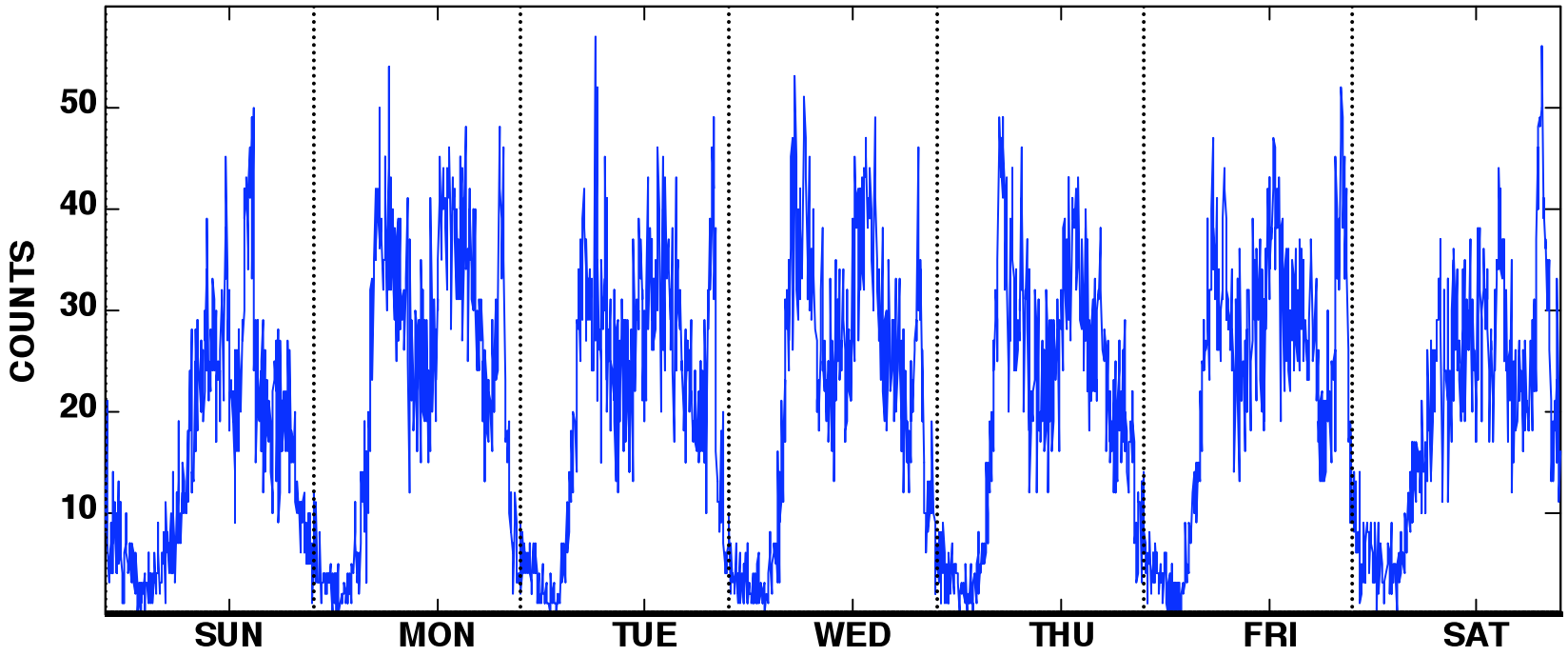
Freeway Traffic Data

- Onramp sensor near Dodger baseball stadium

- Aggregated vehicle counts every 5 minutes
- 2 years of data
- Partial ground truth for event data = baseball games

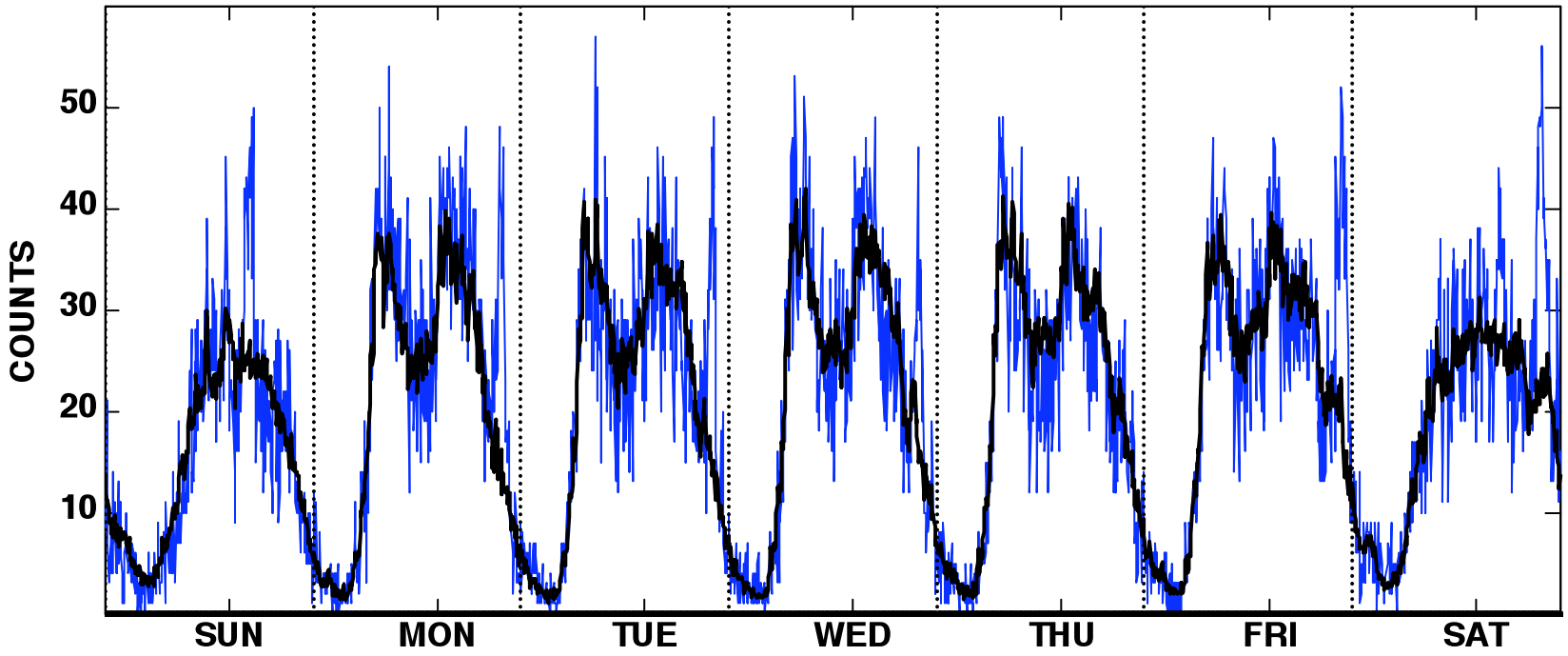
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

One Week of Traffic Data



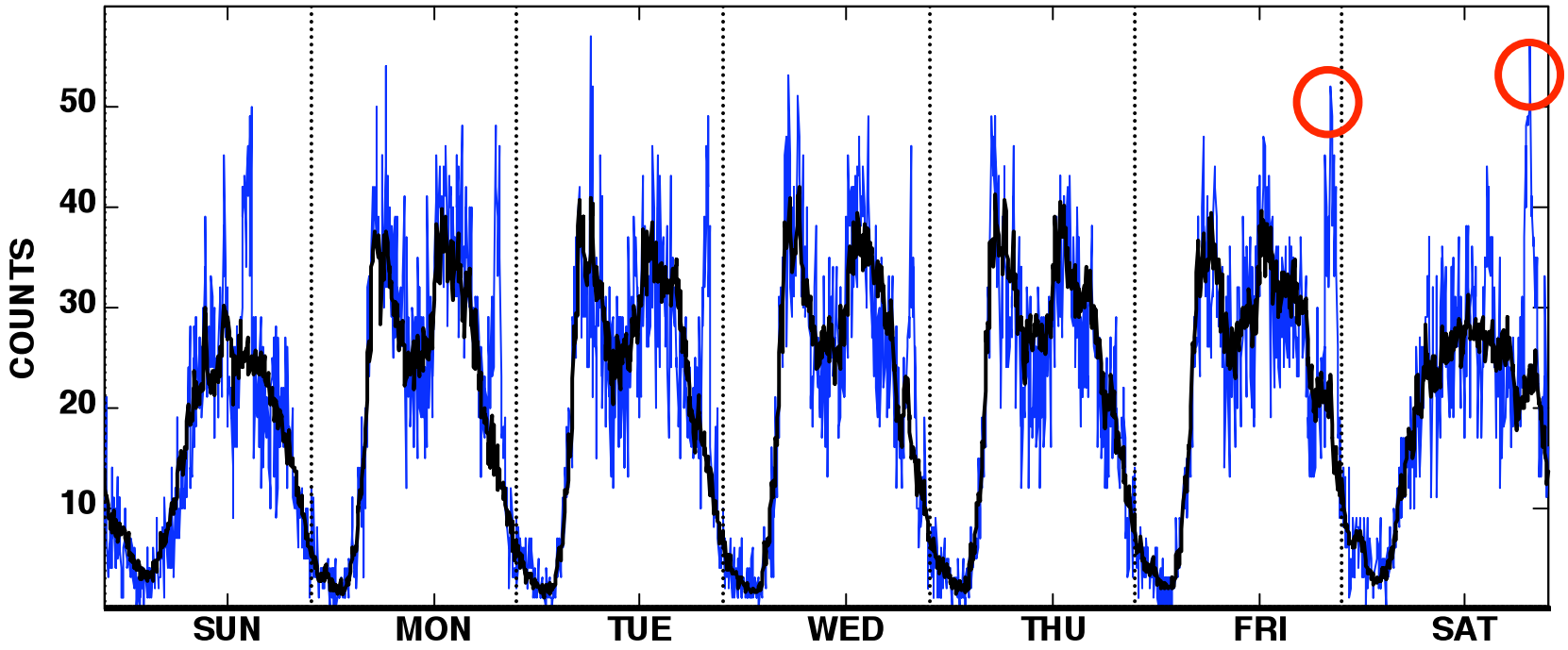
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

One Week of Traffic Data



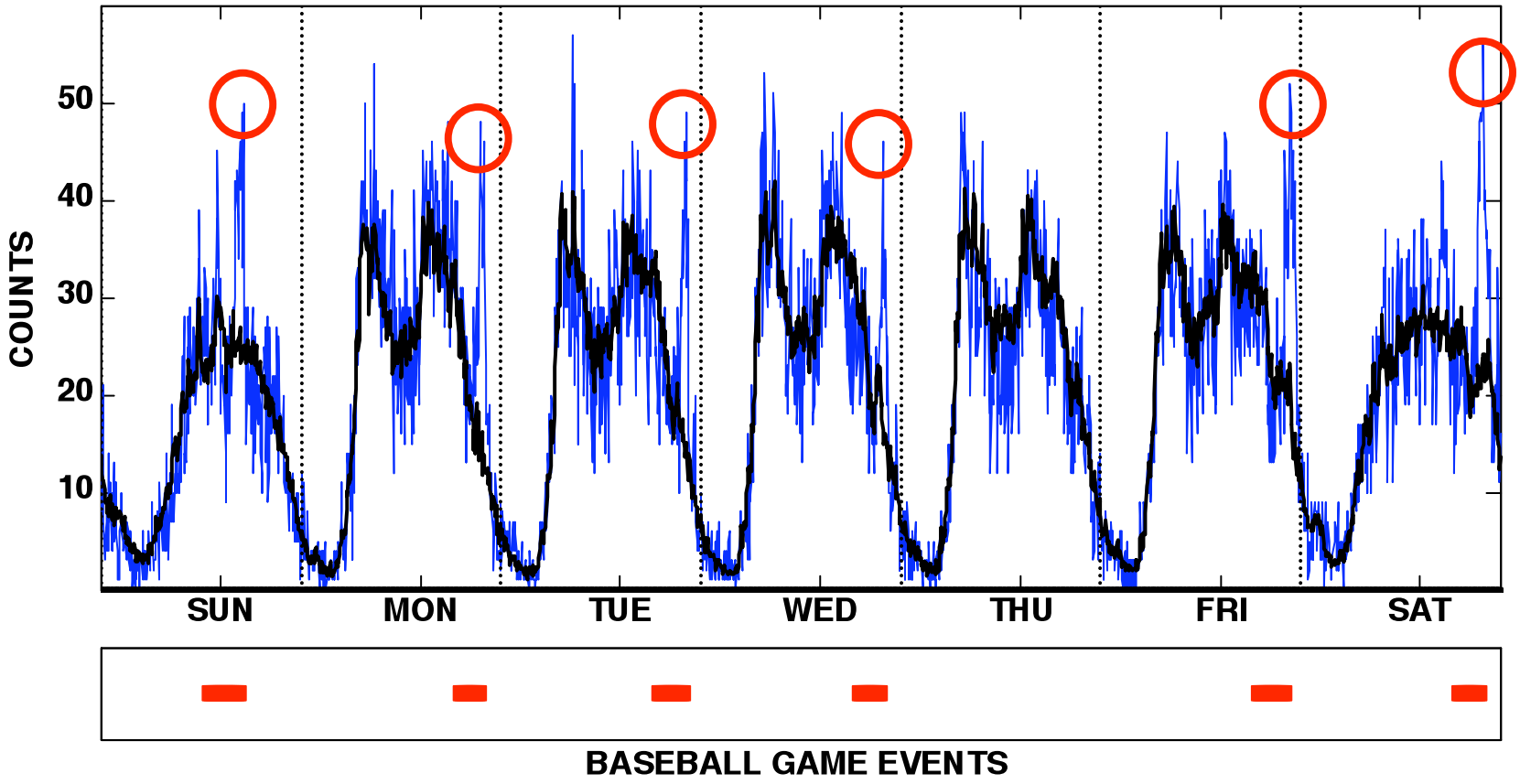
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

One Week of Traffic Data



UCIrvine
University of California, Irvine

One Week of Traffic Data



"Event" = large scale activity,
unusual relative to normal patterns

UCIrvine
University of California, Irvine

Building Entry Data

Optical sensors at door entrances of Calit2 building on UC Irvine campus

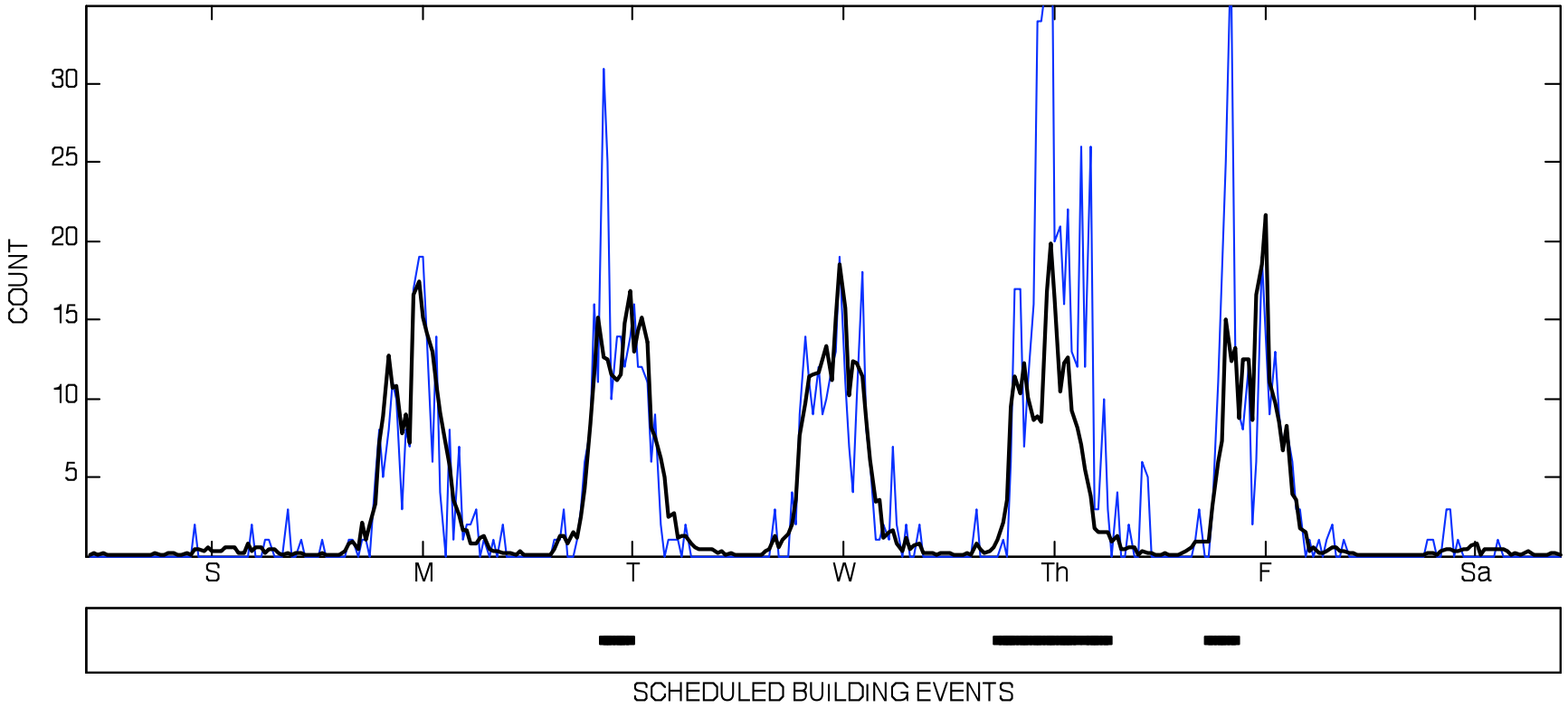
Aggregated counts every 30 minutes

6 months of data

Partial ground truth for events = building calendar

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

1 Week of Building Data



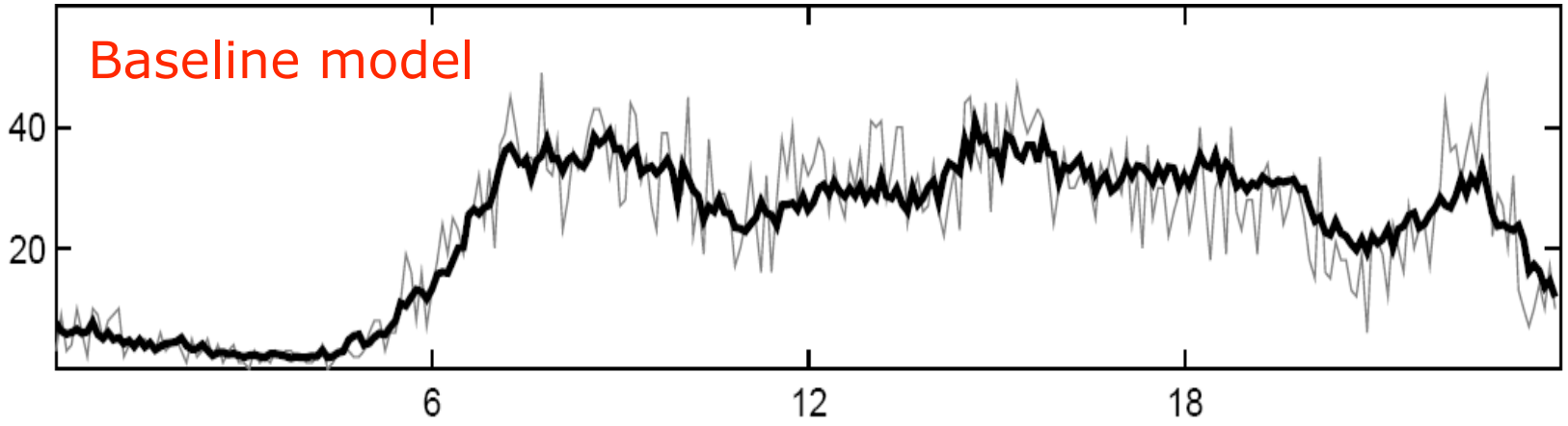
UCIrvine
University of California, Irvine

Problem Statement

- Given:
  - Data in the form of count time-series (of people, vehicles, etc)
  - Regular patterns of behavior + bursty events
  - Unlabeled
- Learn:
  - How to model normal recurrent behavior
  - How to detect abnormal events

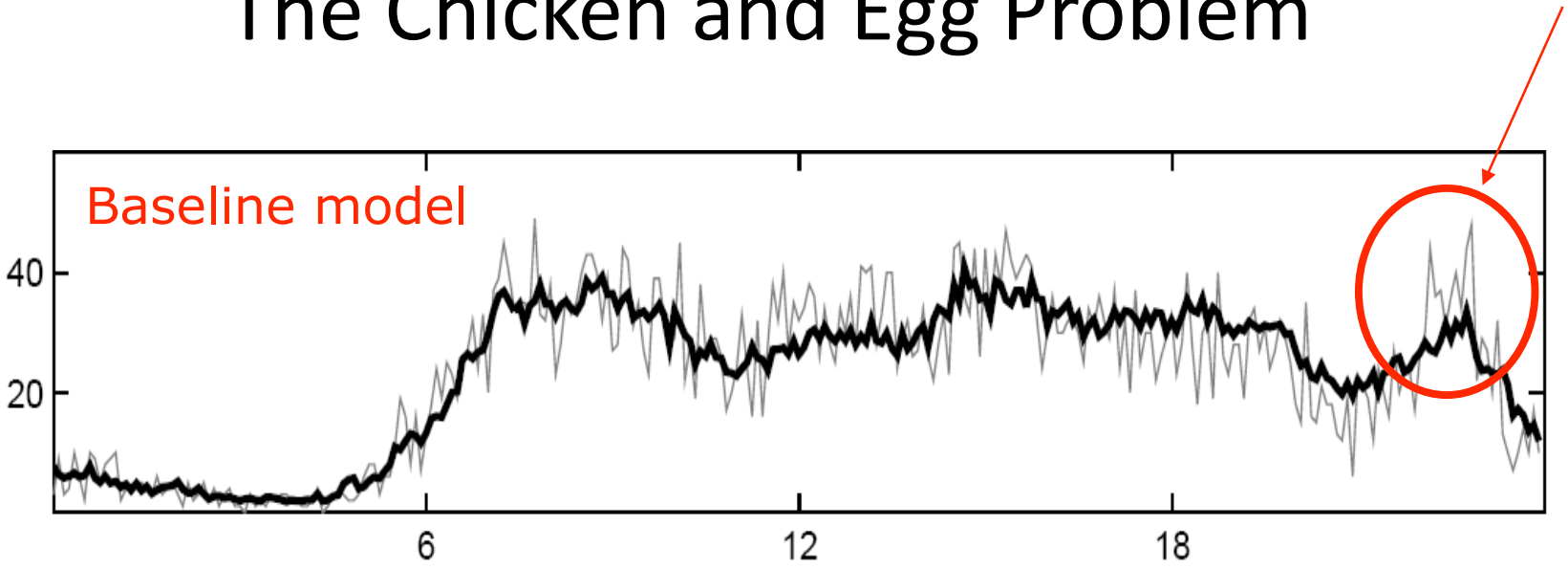
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

The Chicken and Egg Problem



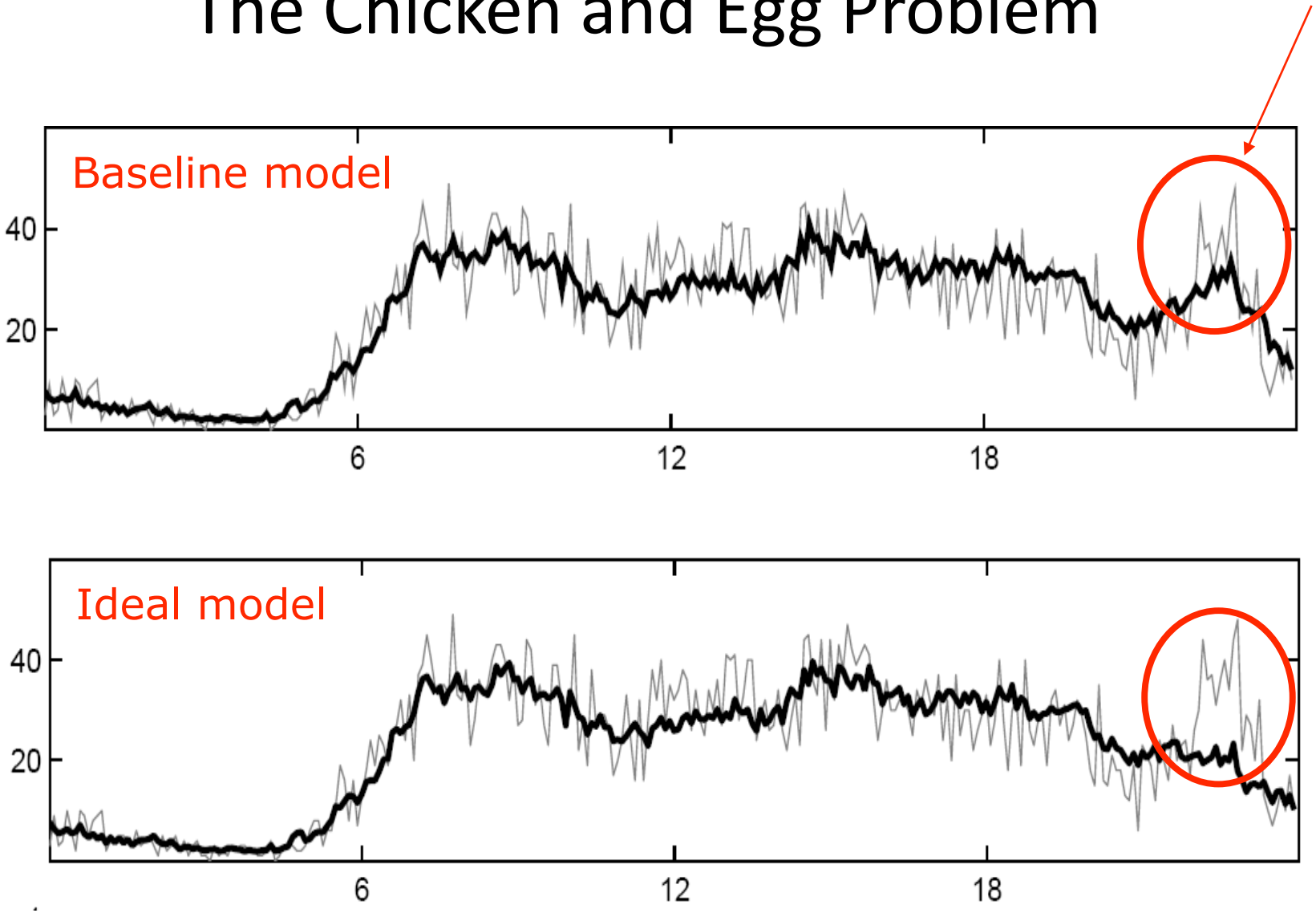
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

The Chicken and Egg Problem



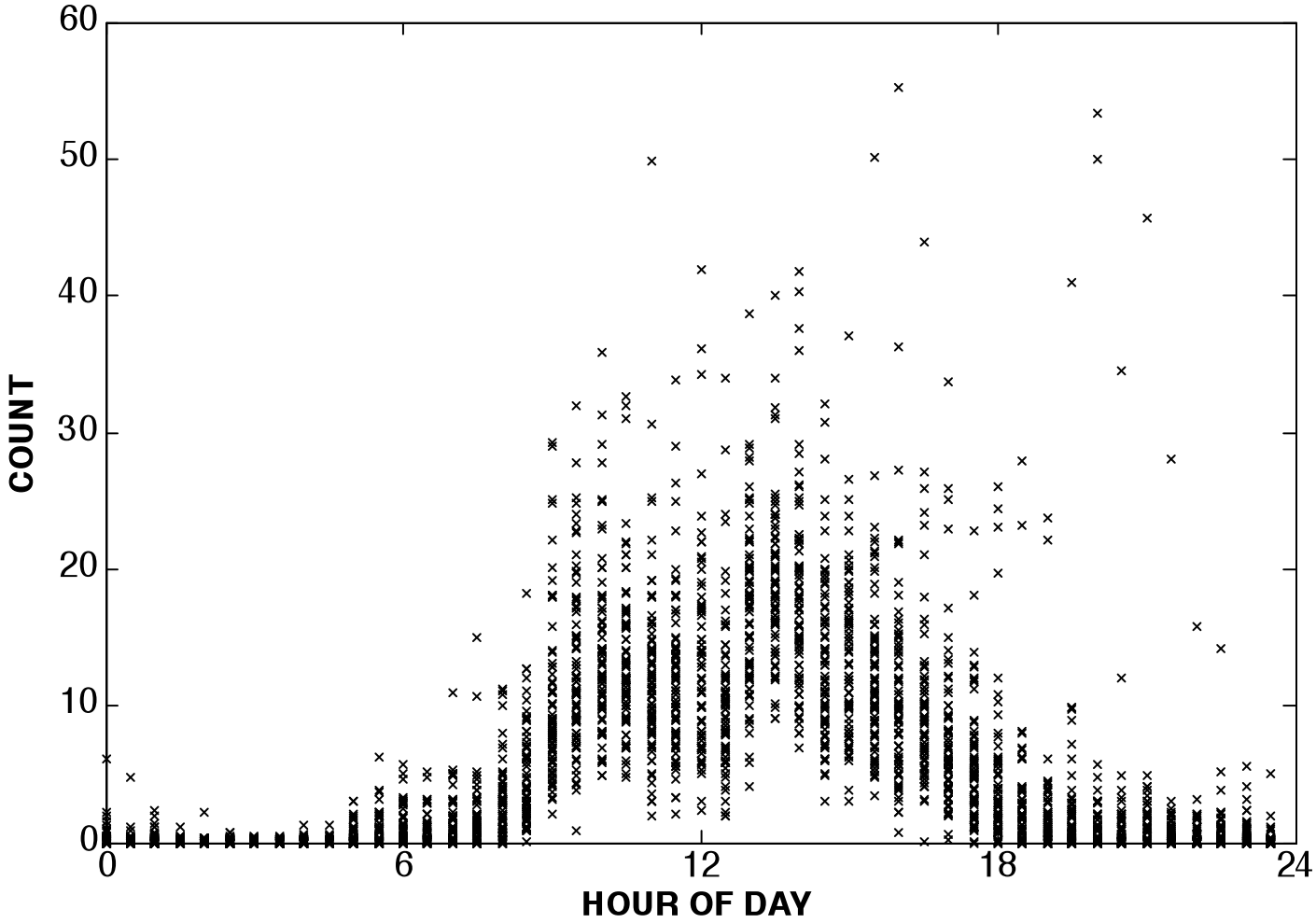
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

The Chicken and Egg Problem



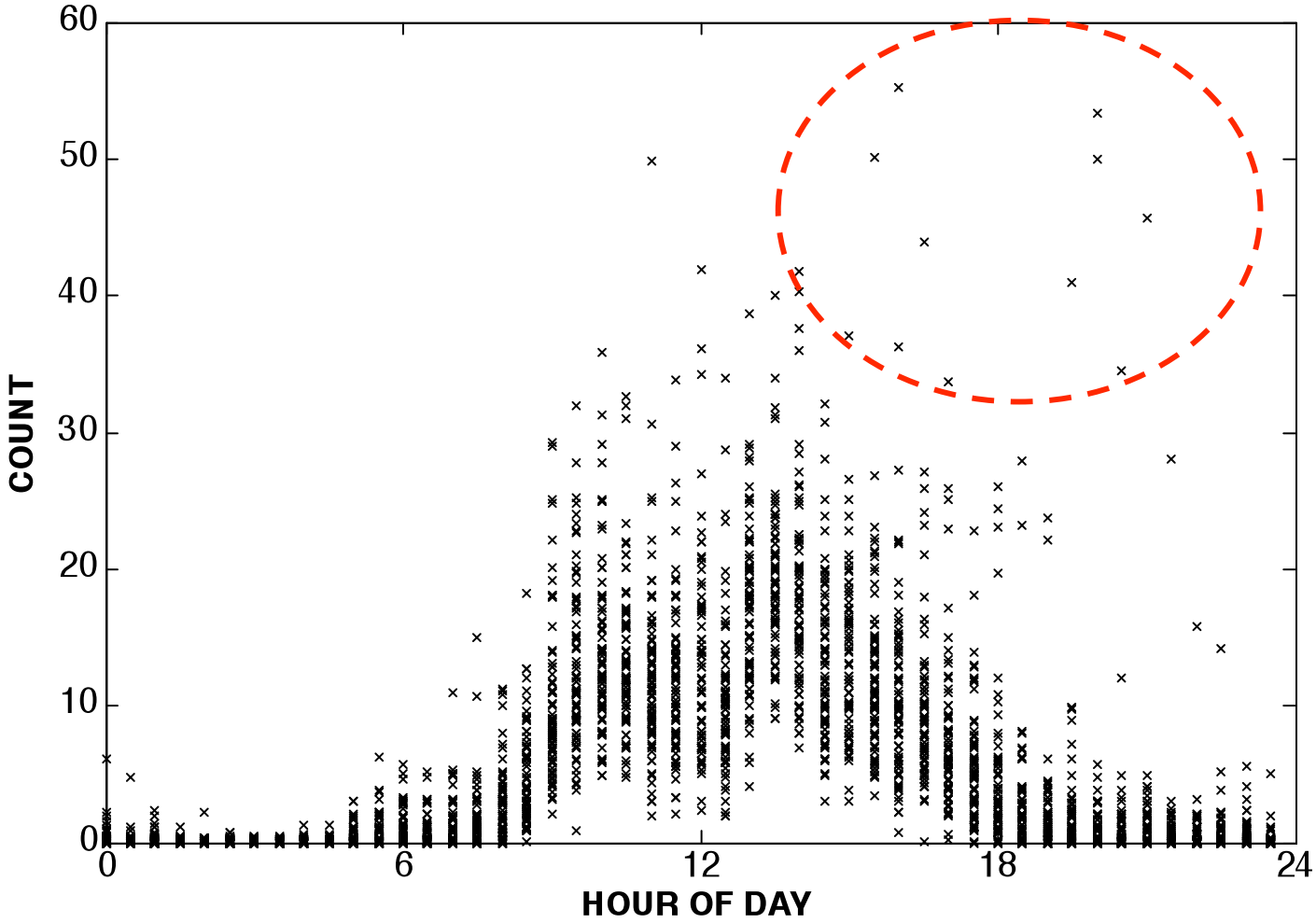
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Why this problem is hard….



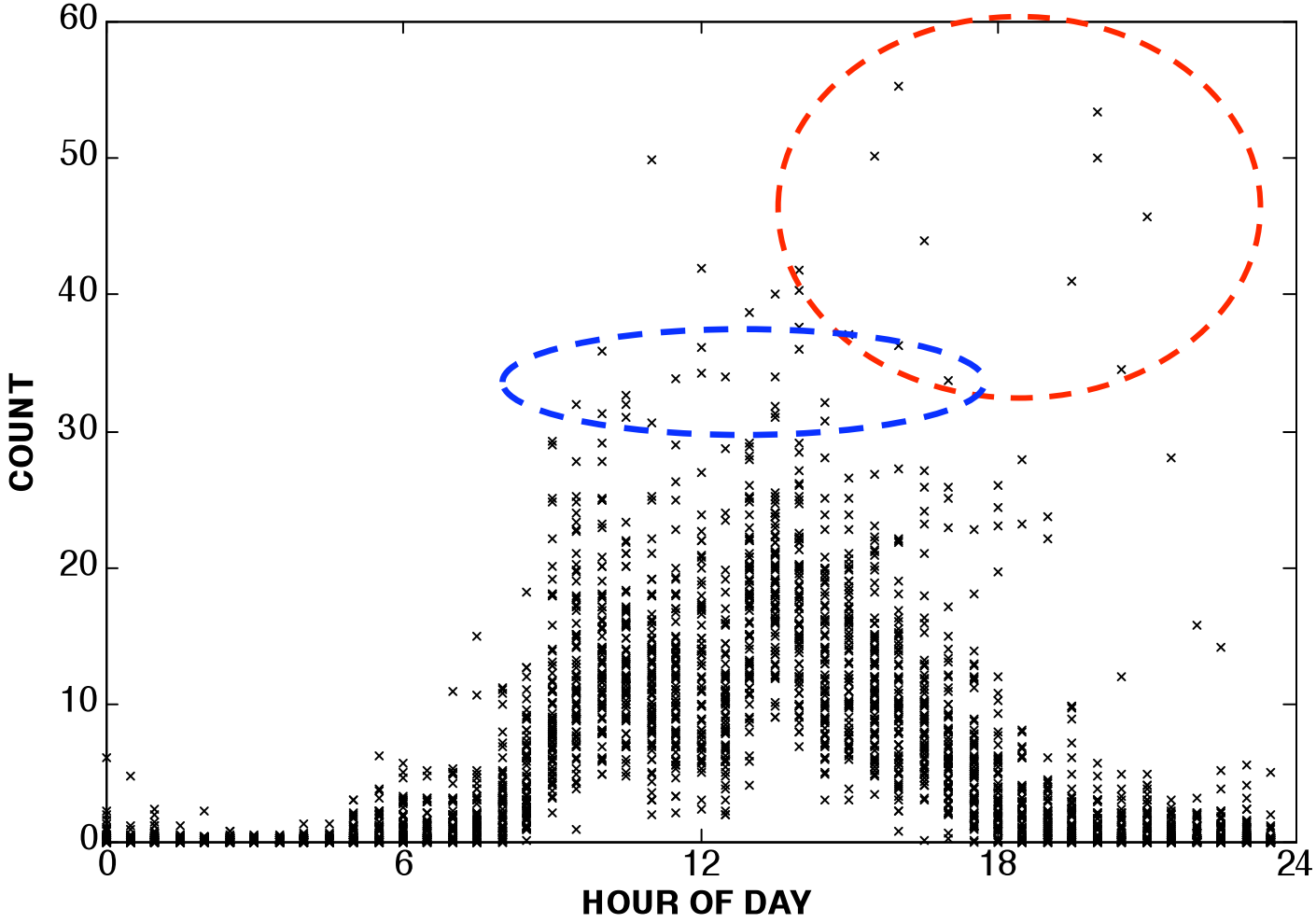
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Why this problem is hard….



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Why this problem is hard….



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Proposed Model

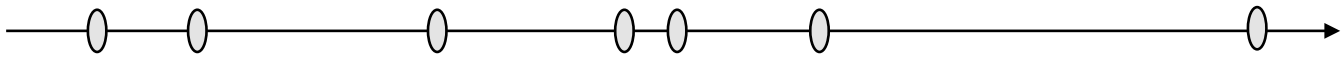
$$N(t) = N_0(t) + N_E(t)$$

OBSERVED COUNT

NORMAL COUNT (UNOBSERVED)

EVENT COUNT (UNOBSERVED)

Poisson Process



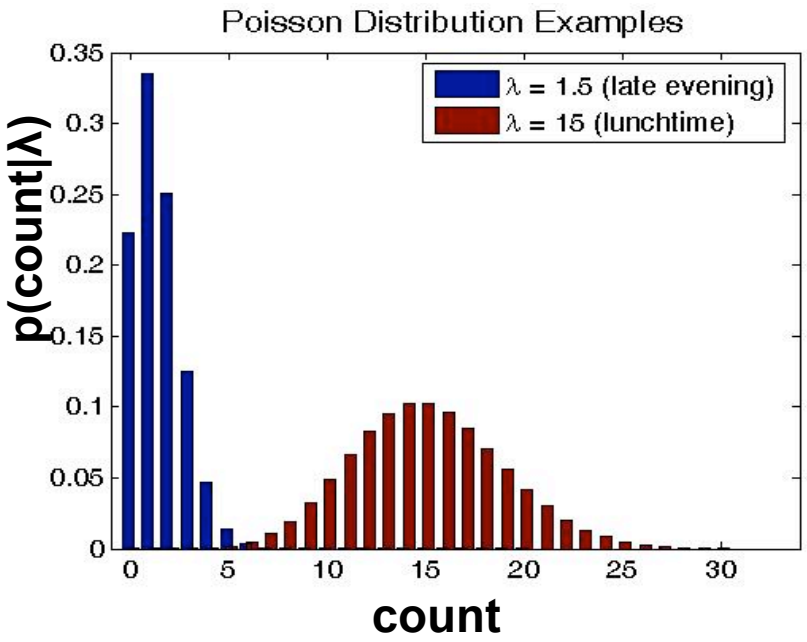
- Simple stochastic model for count data
  - Single parameter λ = "rate"

λ = expected number of events per unit time

Time-Varying Poisson Process

- The rate $\lambda(t)$ is a function of time ("inhomogeneous")

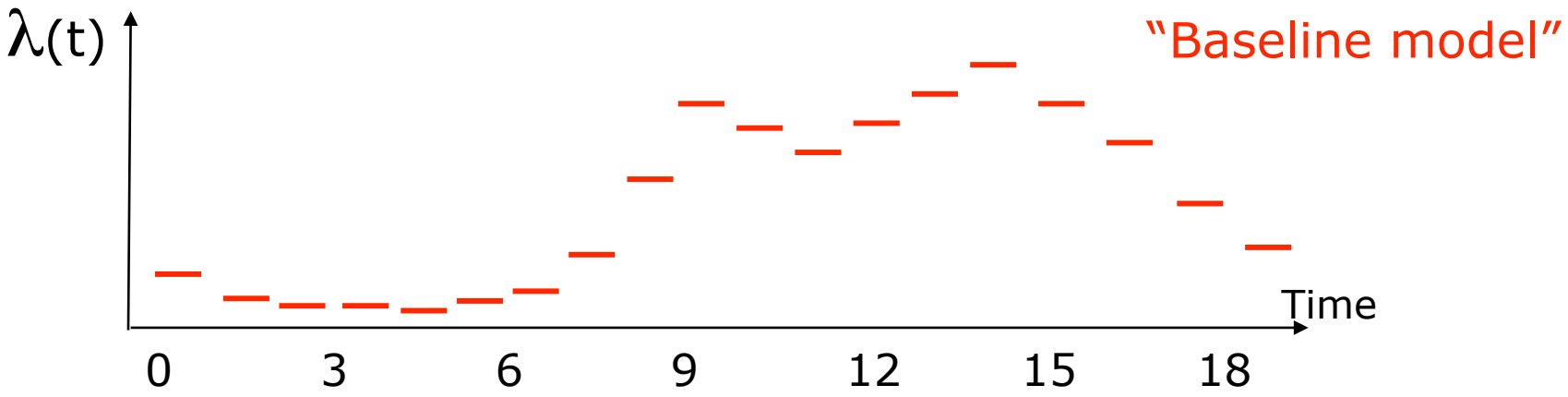
- Simple parametrization
  - Different λ's for different discrete times of day, day of week



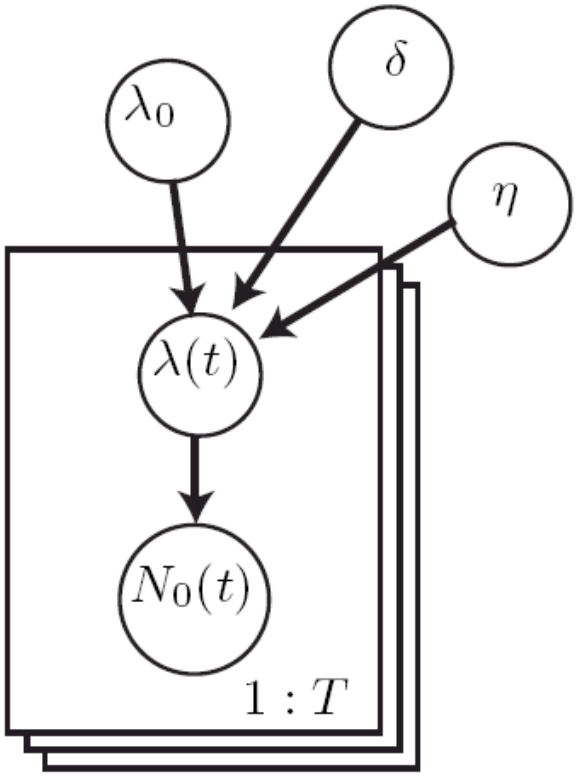
Inhomogeneous Poisson Process

- The rate $\lambda(t)$ is a function of time ("inhomogeneous")

- Piecewise constant parametrization
  - Different λ's for different discrete times of day, day of week

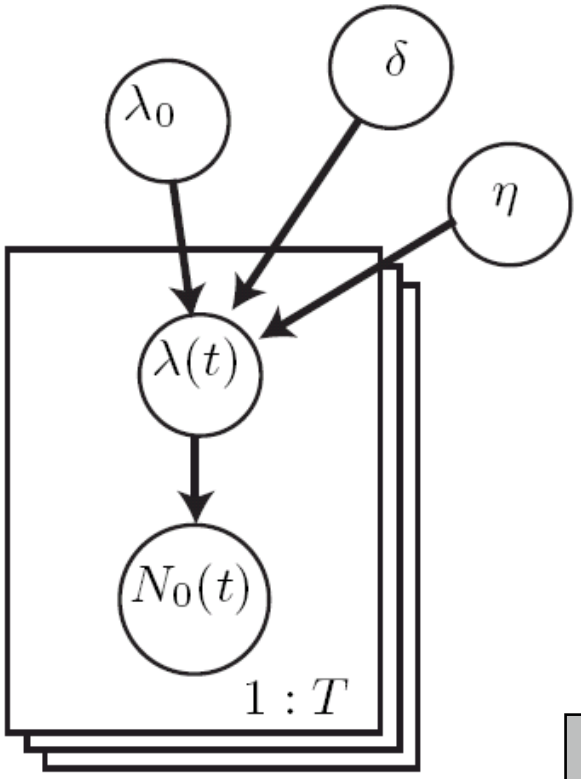


Model for Normal Counts



$$\lambda(t) = \lambda_0 \ \delta_{d(t)} \ \eta_{d(t),h(t)}$$

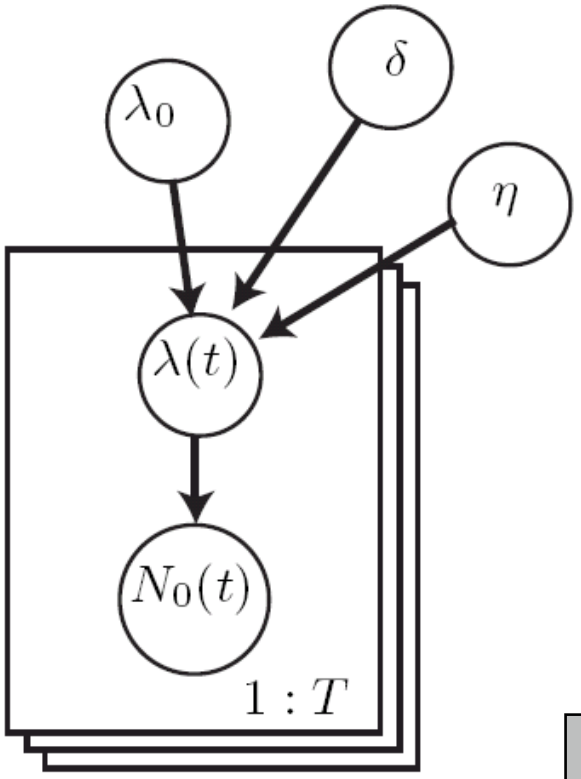
Model for Normal Counts



$$\lambda(t) = \lambda_0 \ \delta_{d(t)} \ \eta_{d(t),h(t)}$$

Overall mean rate

Model for Normal Counts



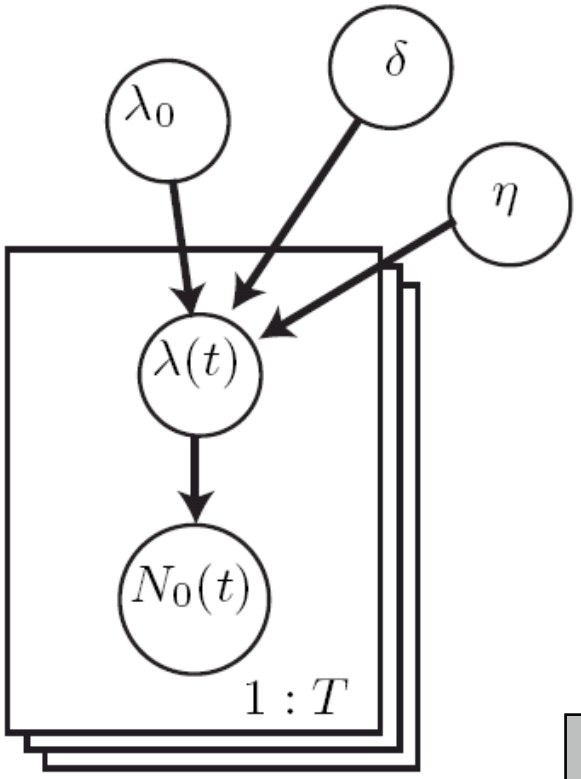
$$\lambda(t) = \lambda_0 \ \delta_{d(t)} \ \eta_{d(t),h(t)}$$

Daily rate component

Overall mean rate

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Model for Normal Counts



$$\lambda(t) = \lambda_0 \ \delta_{d(t)} \ \eta_{d(t),h(t)}$$

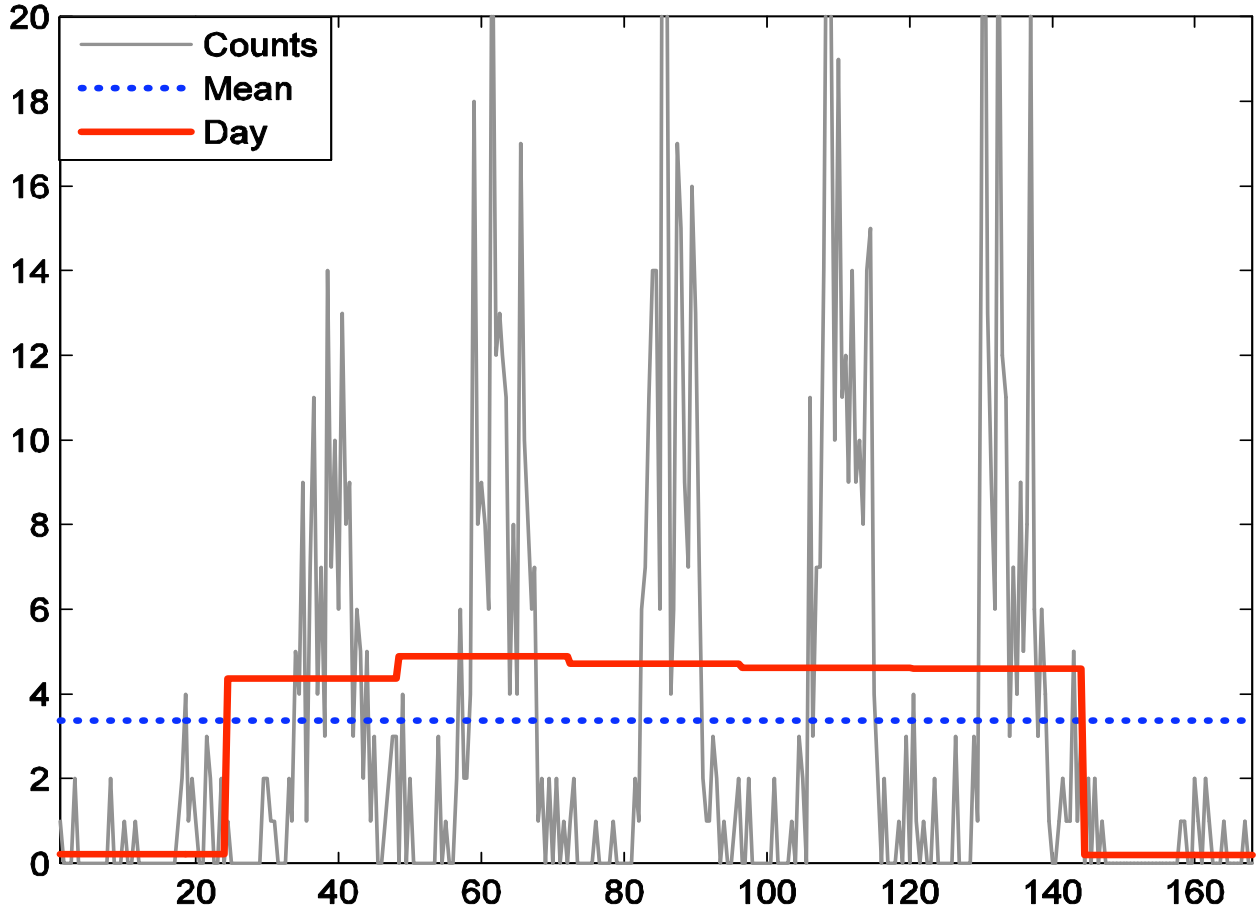
Daily rate component

Overall mean rate

Local time bin component, given day and time

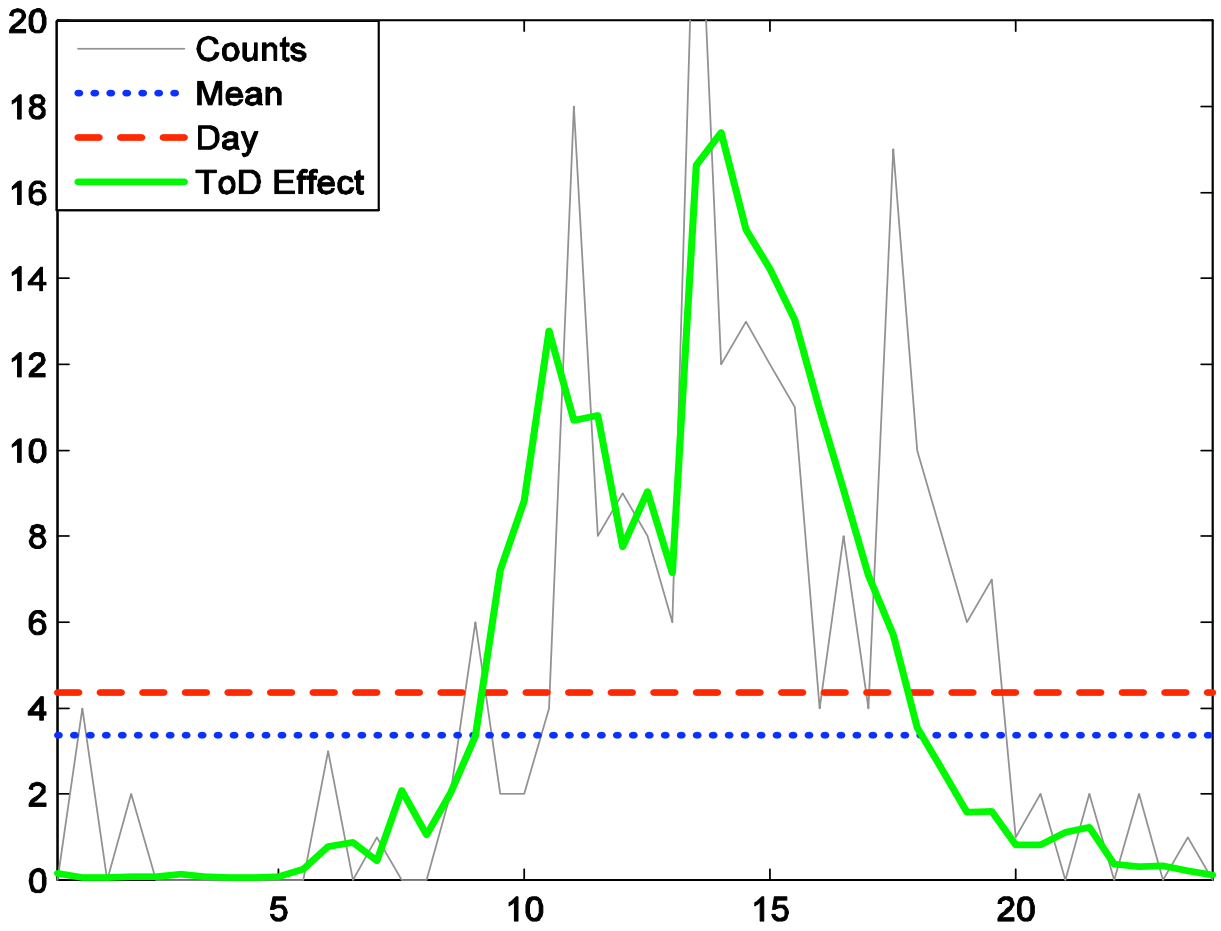
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Daily Rate Component



UCIrvine
University of California, Irvine

Time of Day Rate Component



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Proposed Model

$$N(t) = N_0(t) + N_E(t)$$

OBSERVED COUNT

NORMAL COUNT (UNOBSERVED)

EVENT COUNT (UNOBSERVED)

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Proposed Model

$$N(t) = N_0(t) + N_E(t)$$

OBSERVED COUNT

NORMAL COUNT (UNOBSERVED)

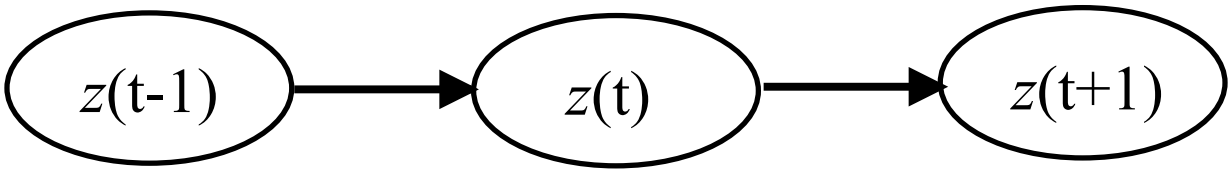
EVENT COUNT (UNOBSERVED)

Time-varying Poisson

Markov with Poisson counts

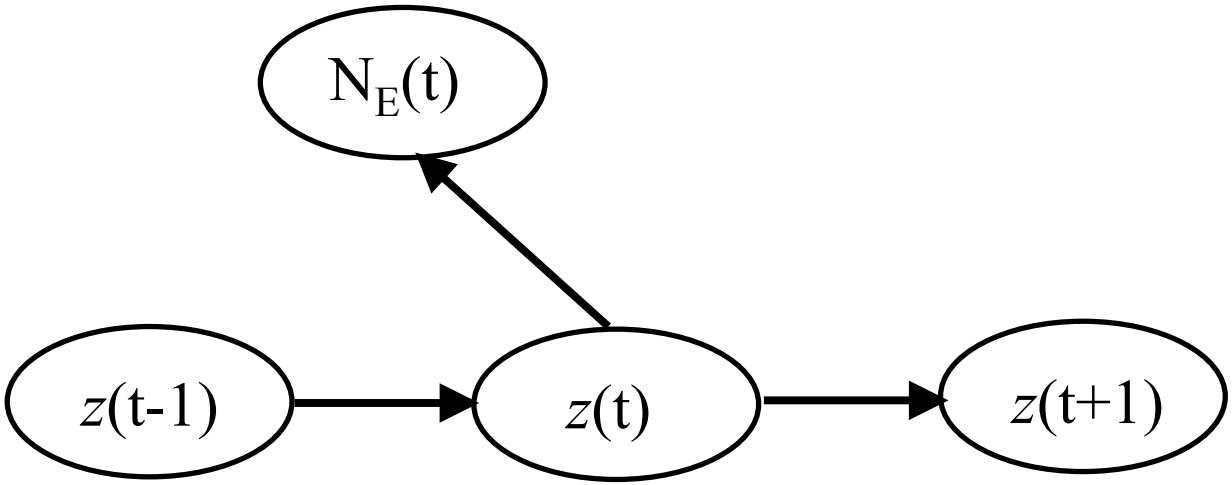
Adding a Hidden Event Process

$$z(t) = \begin{cases} 0 & \text{if there is no event at time } t \\ +1 & \text{if there is a positive event} \\ -1 & \text{if there is a negative event} \end{cases}$$

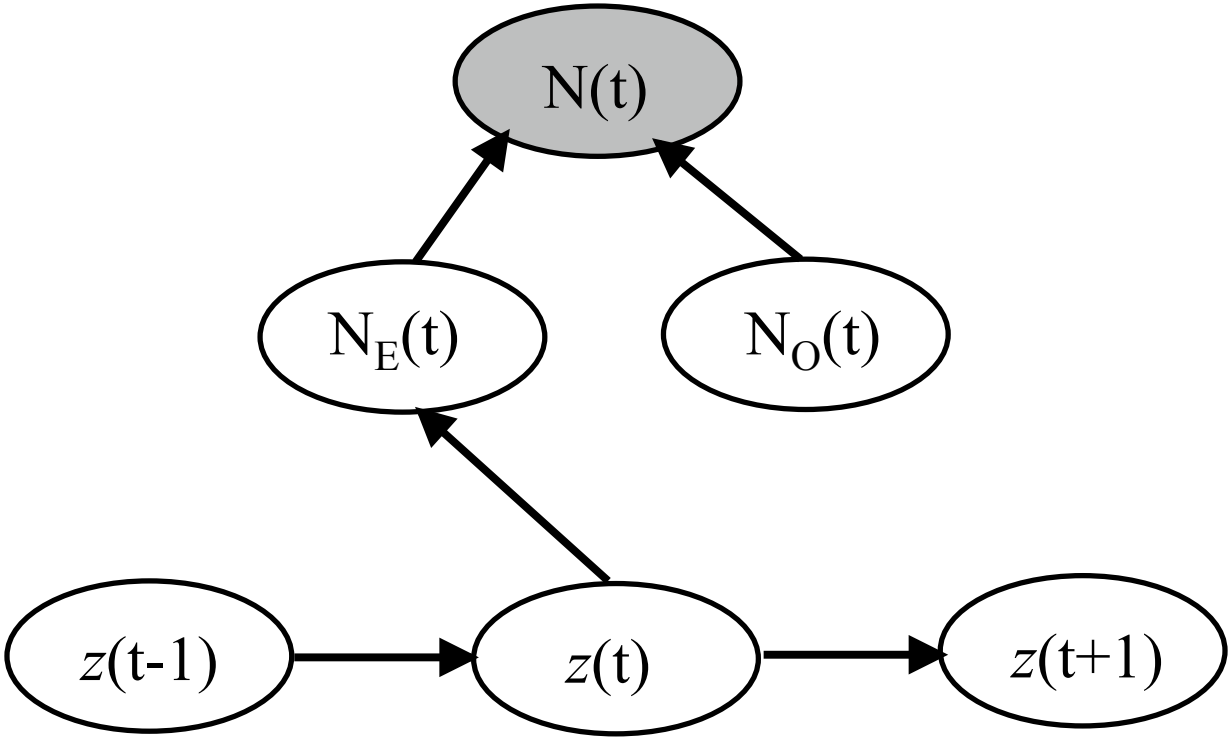


Adding a Hidden Event Process

$$N_E(t) \sim \begin{cases} 0 & z(t)=0 \\ \text{Poisson}(N\,;\,\gamma) & z(t)=1 \text{ or } -1 \end{cases}$$

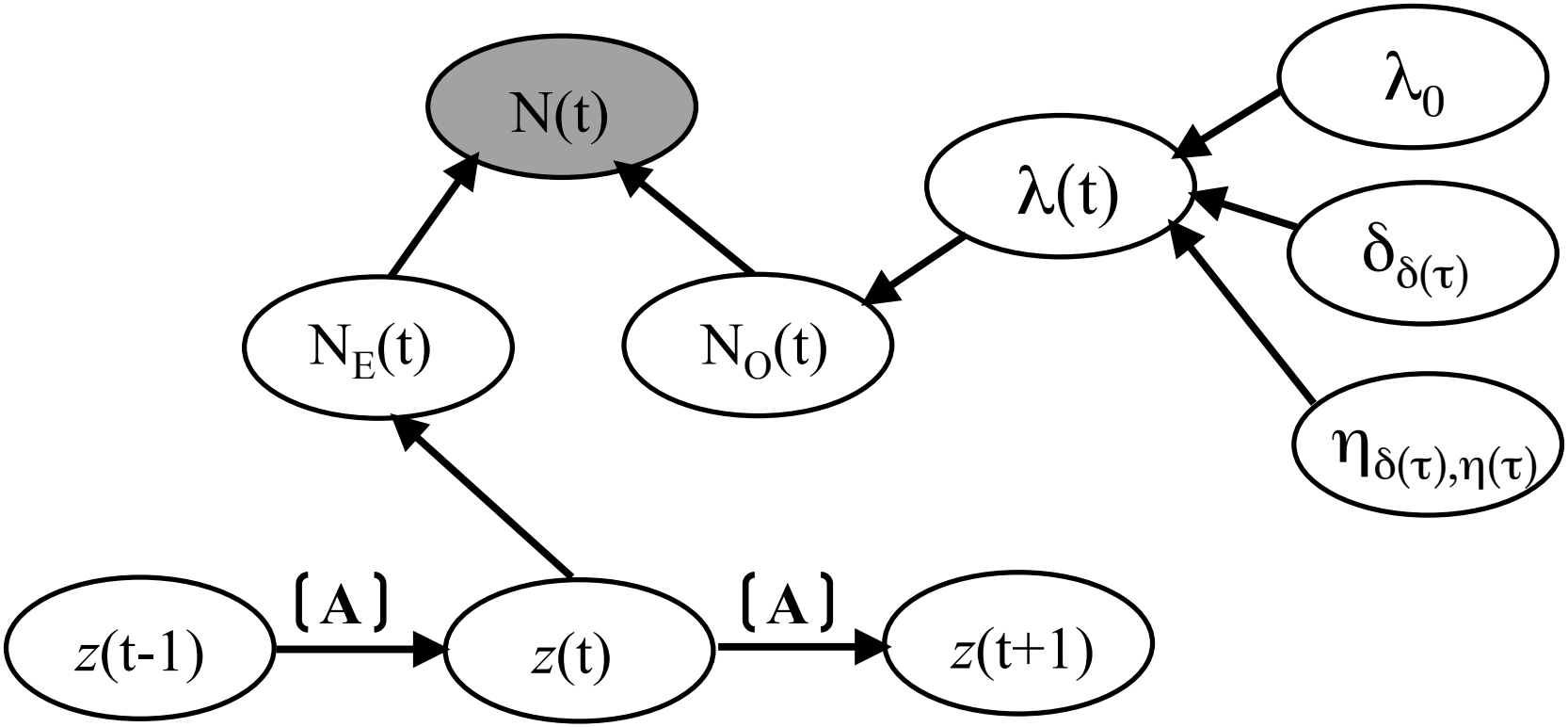


Adding a Hidden Event Process



Markov Modulated Poisson Process (MMPP) model
e.g., see Heffes and Lucantoni (1994), Scott (1998)

Graphical Model



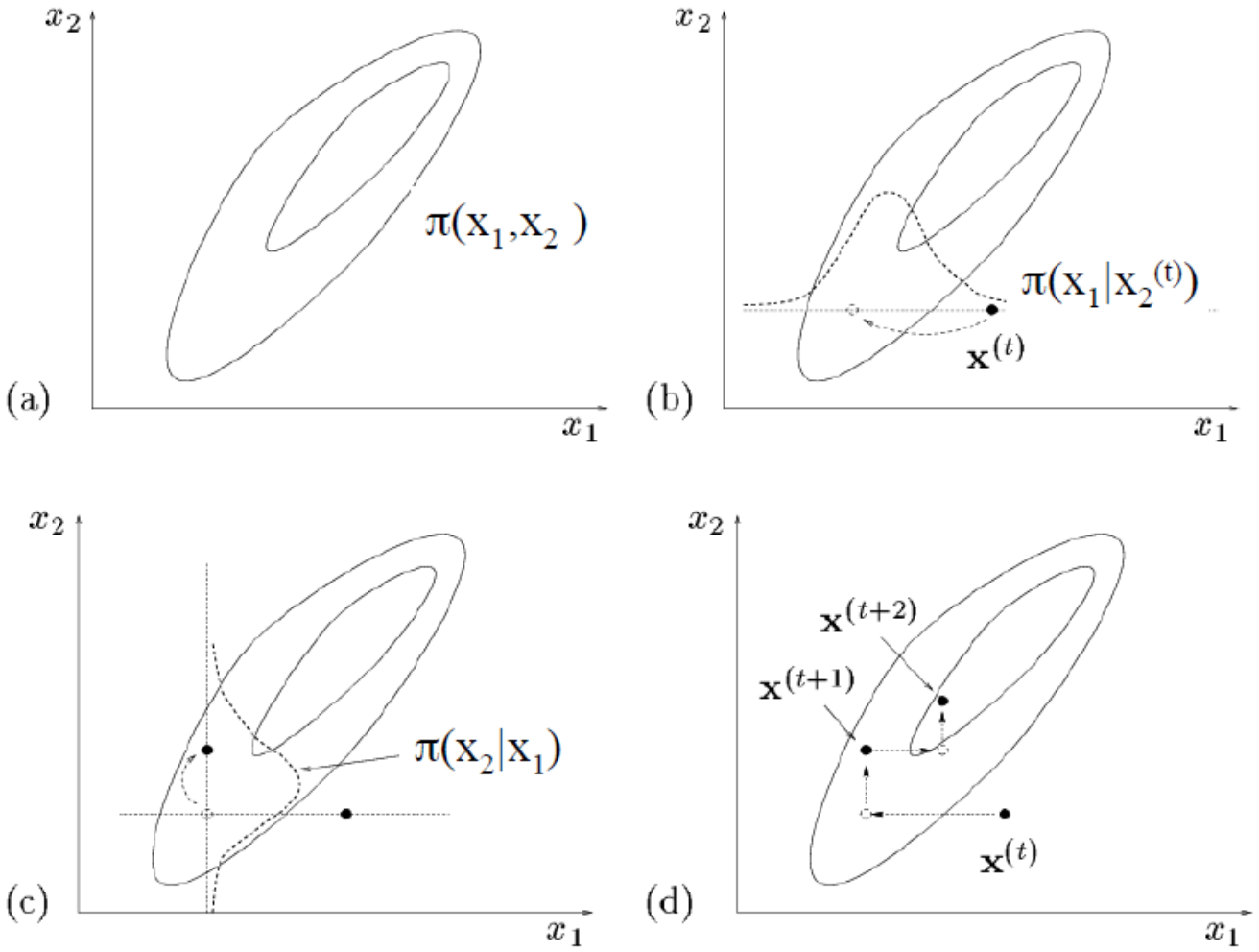
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Bayesian Learning

- Unknowns:
  - Poisson parameters: λ, $\delta(t)$, $\eta(t)$
  - Markov parameters: A, $N_E(t)$
  - Hidden state sequence: $z(t)$
- Use uninformative priors in general
  - E.g., Gamma priors on Poisson rates
  - Informative priors on event transition probabilities
- Use Gibbs sampling to get a sample-based estimate of the posterior distribution given the counts
- Time complexity is linear in T, number of time points
  - Typically 10 iterations for burn-in, and then 50-100 samples after that

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

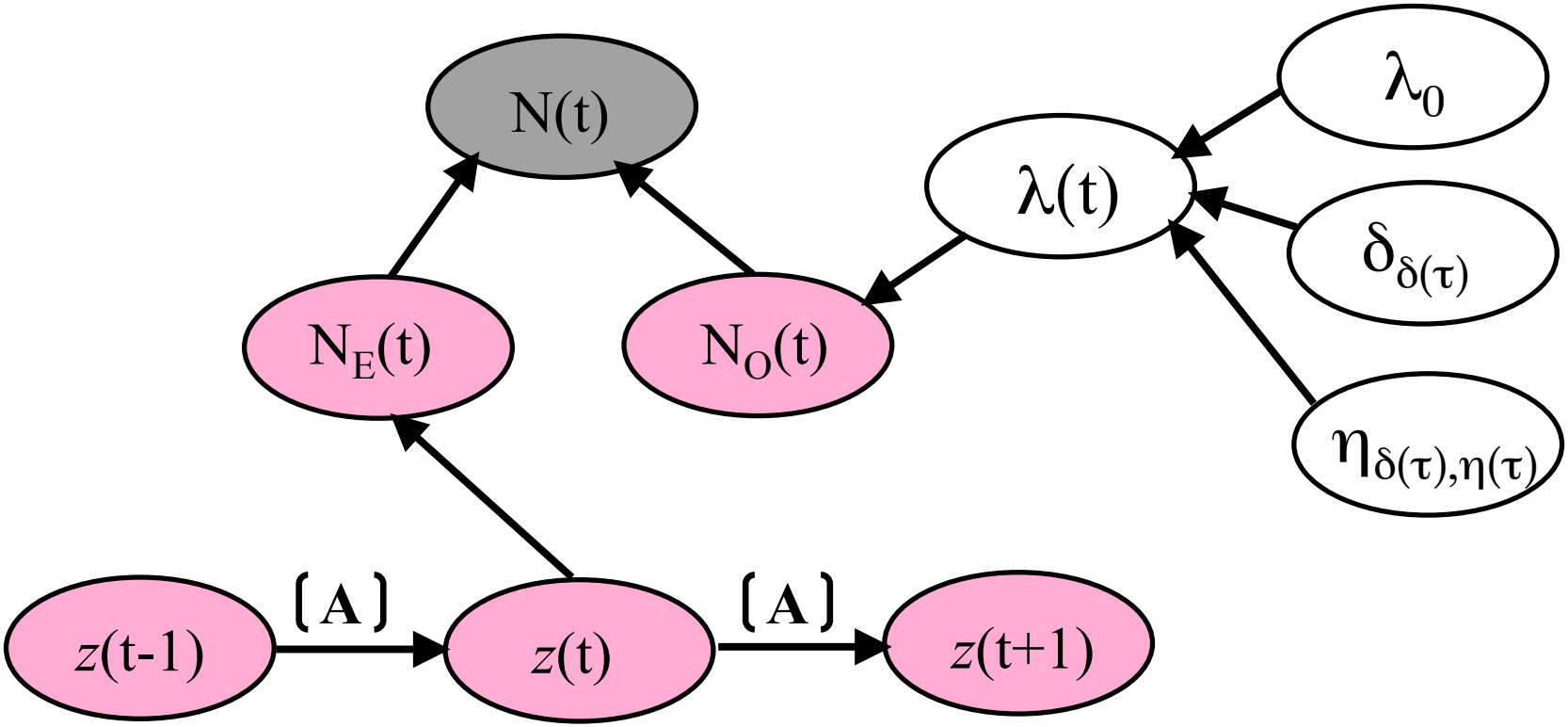
Gibbs Sampling



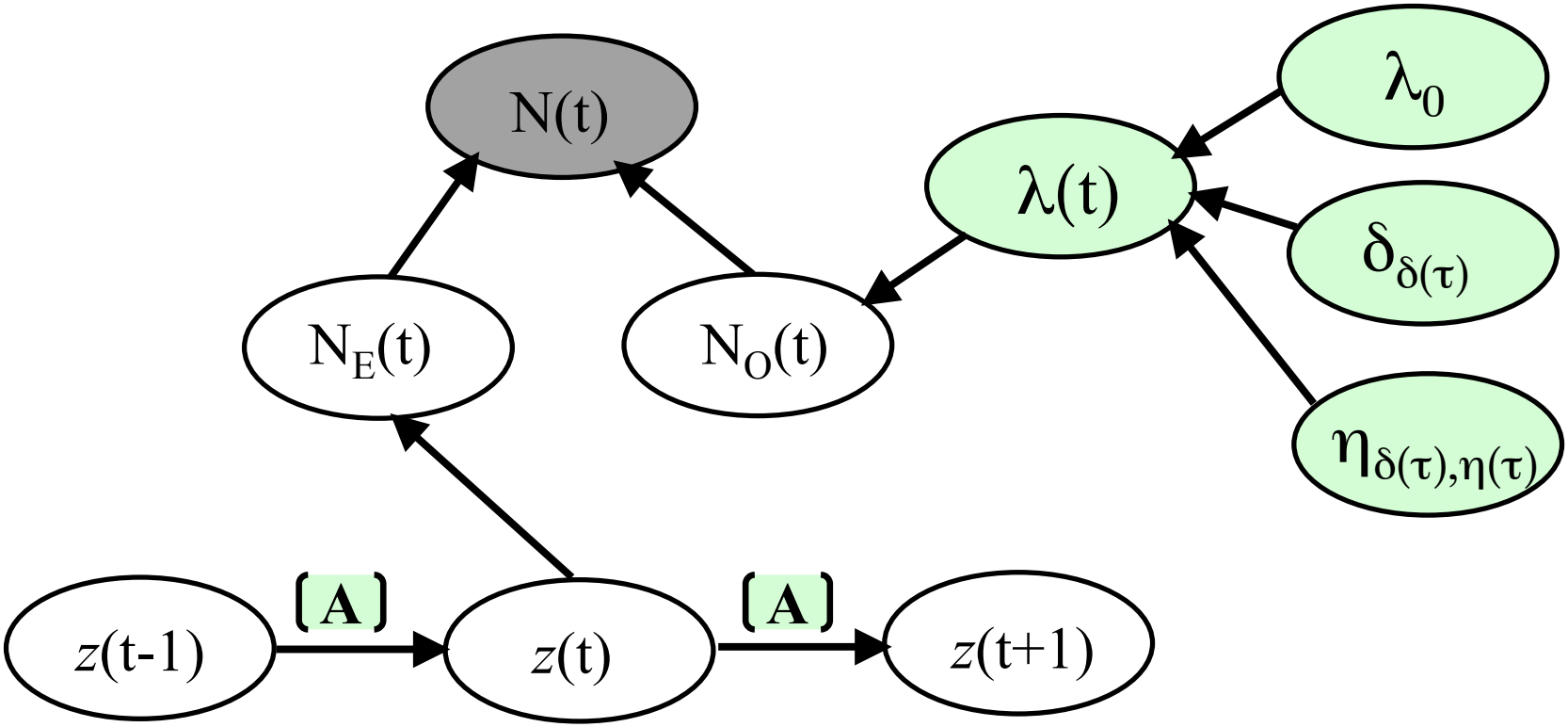
From online MCMC tutorial notes by Frank Dellaert, Georgia Tech

Sampling the Hidden Variables

S. Scott, JASA, 2002



Sampling the Parameters



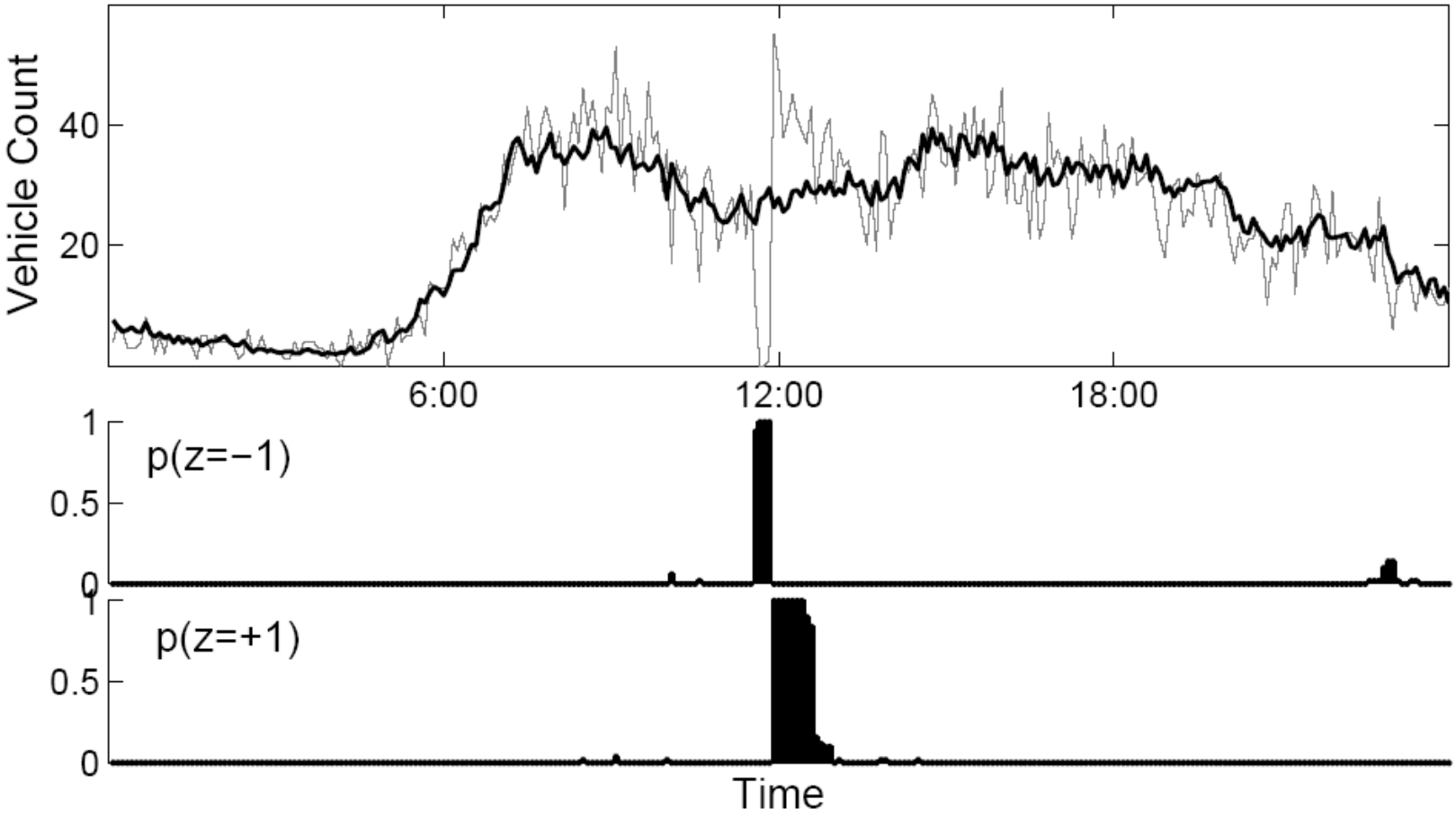
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Experimental Results

- Fit our probabilistic model to sensor streams:
  - UC Irvine building people counts
  - Freeway on-ramp near Dodger Stadium
- Perform inference on several months of data
  - Order of several minutes using Gibbs sampling
- Evaluate models by comparing with (hidden) ground truth
  - Model detects events when $p(z = 0) < 0.5$

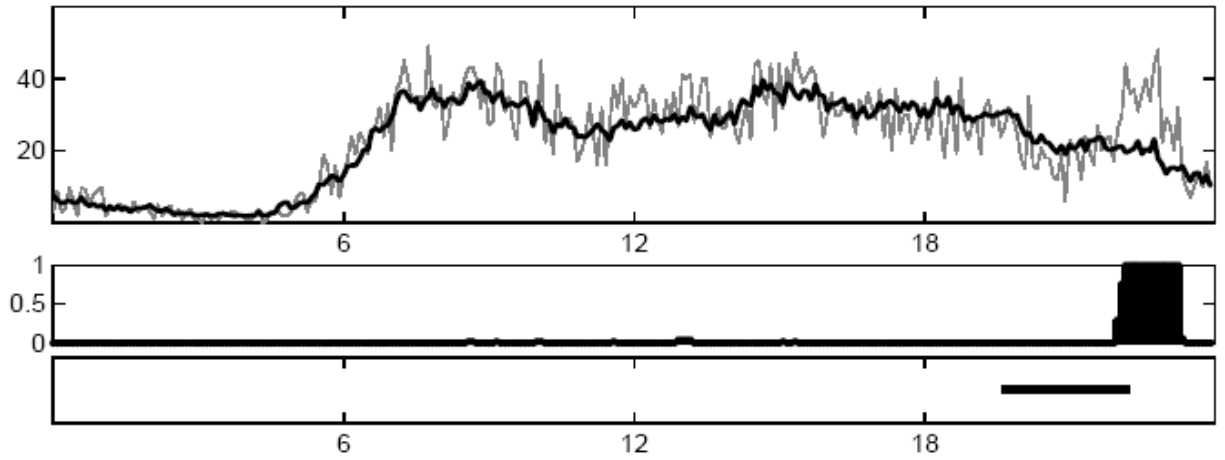
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Example: Traffic Data



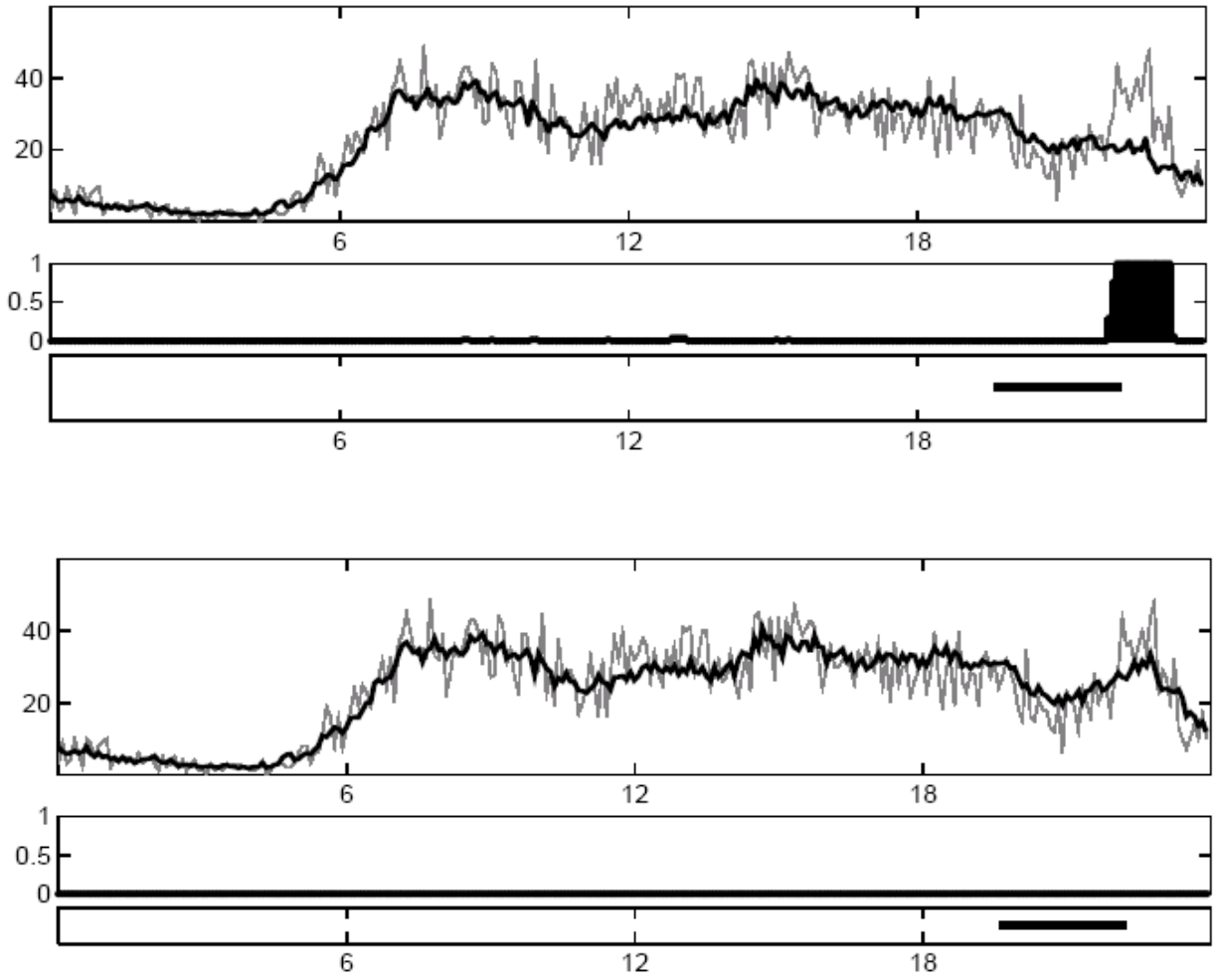
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Detection of Friday Night Baseball Game

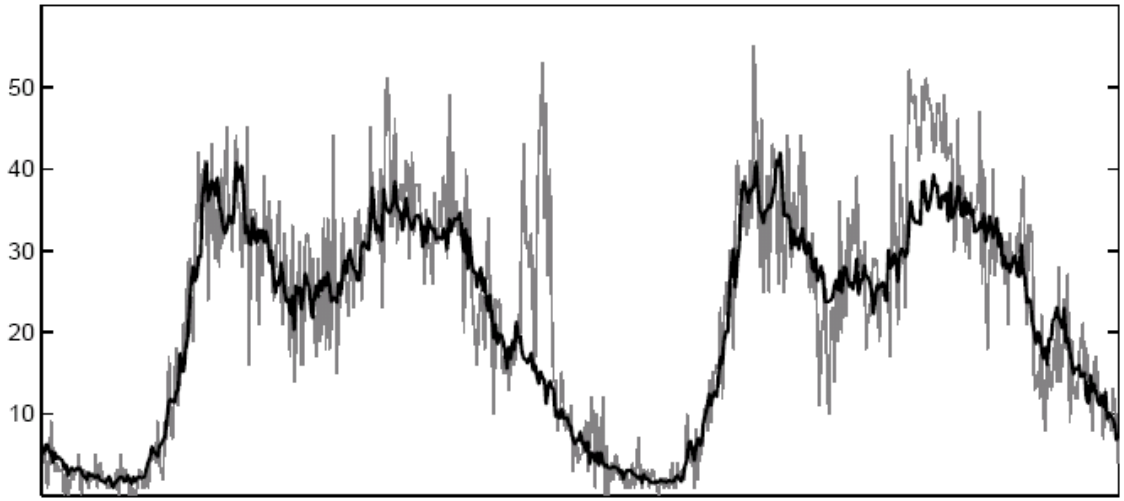


UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Detection of Friday Night Baseball Game

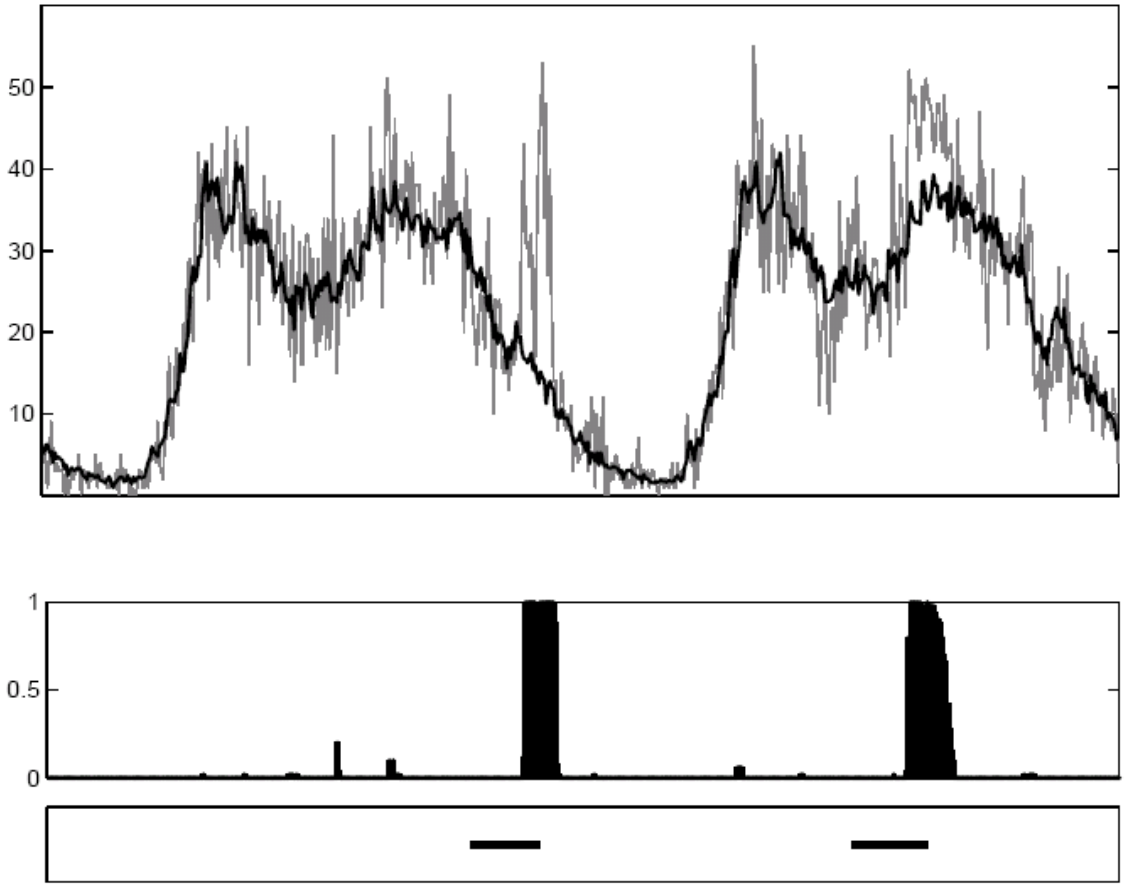


Detection of Baseball Games



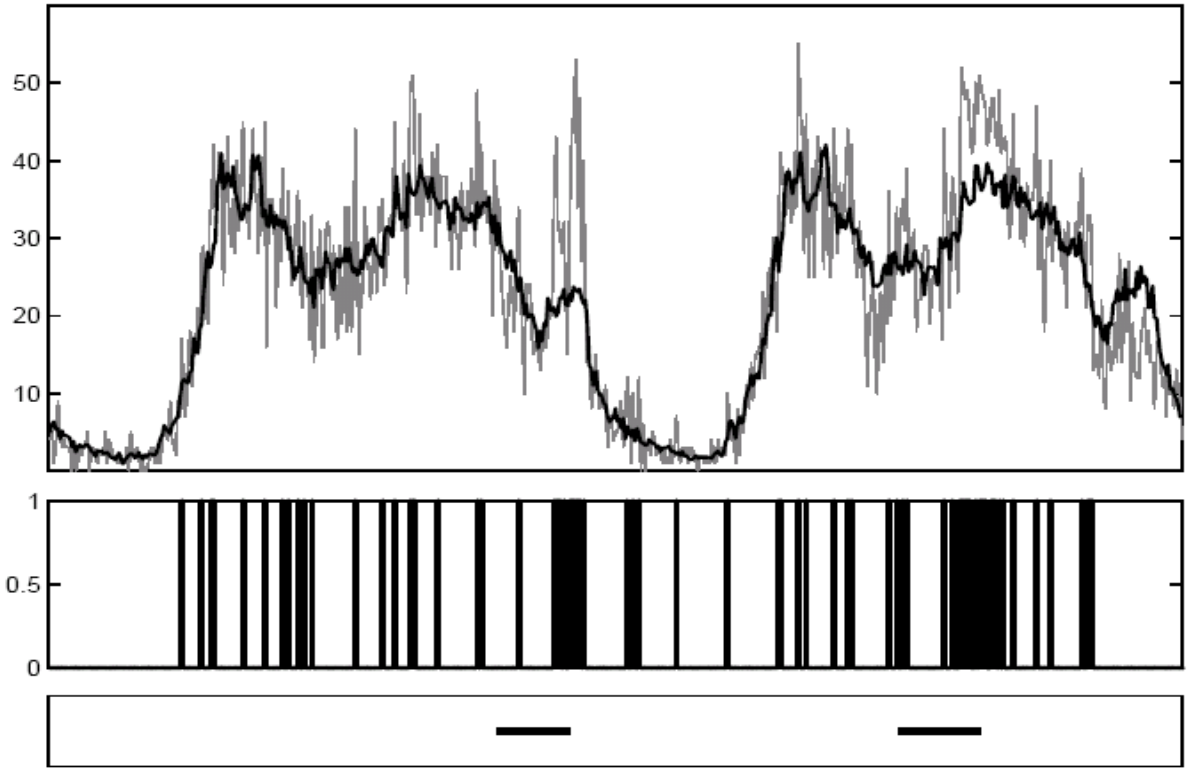
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Detection of Baseball Games

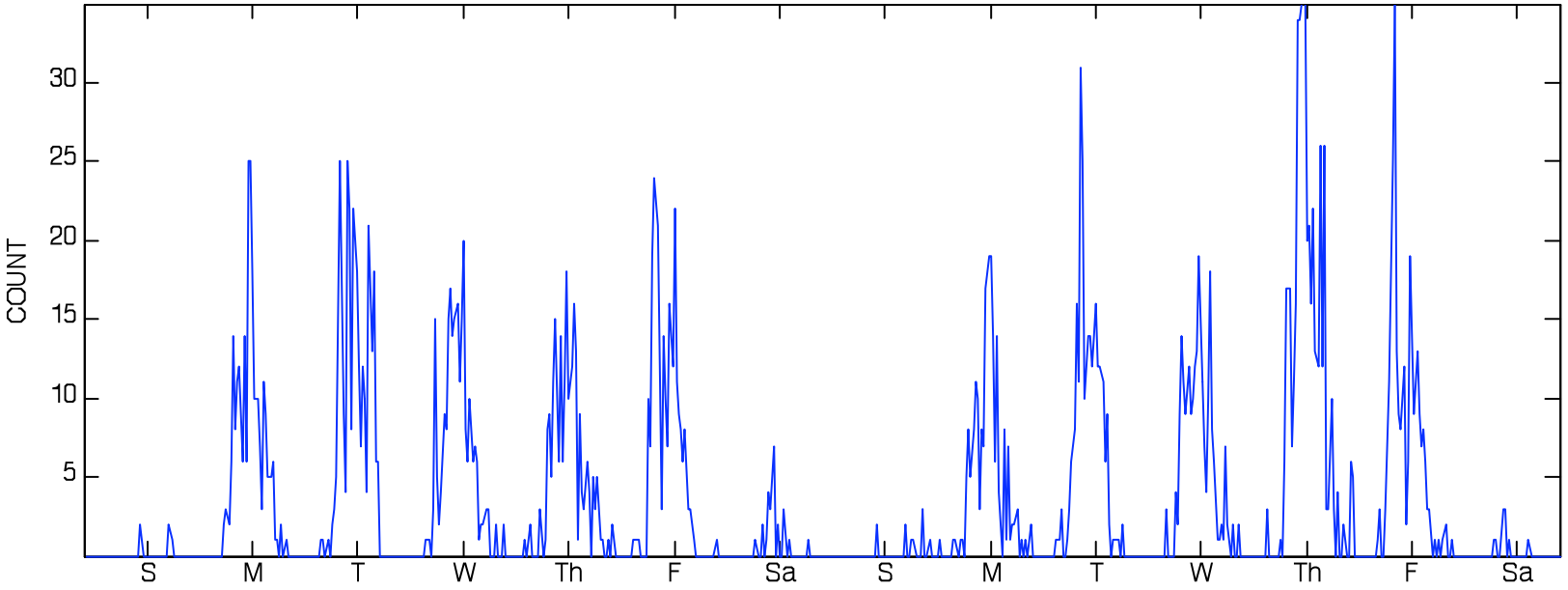


UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

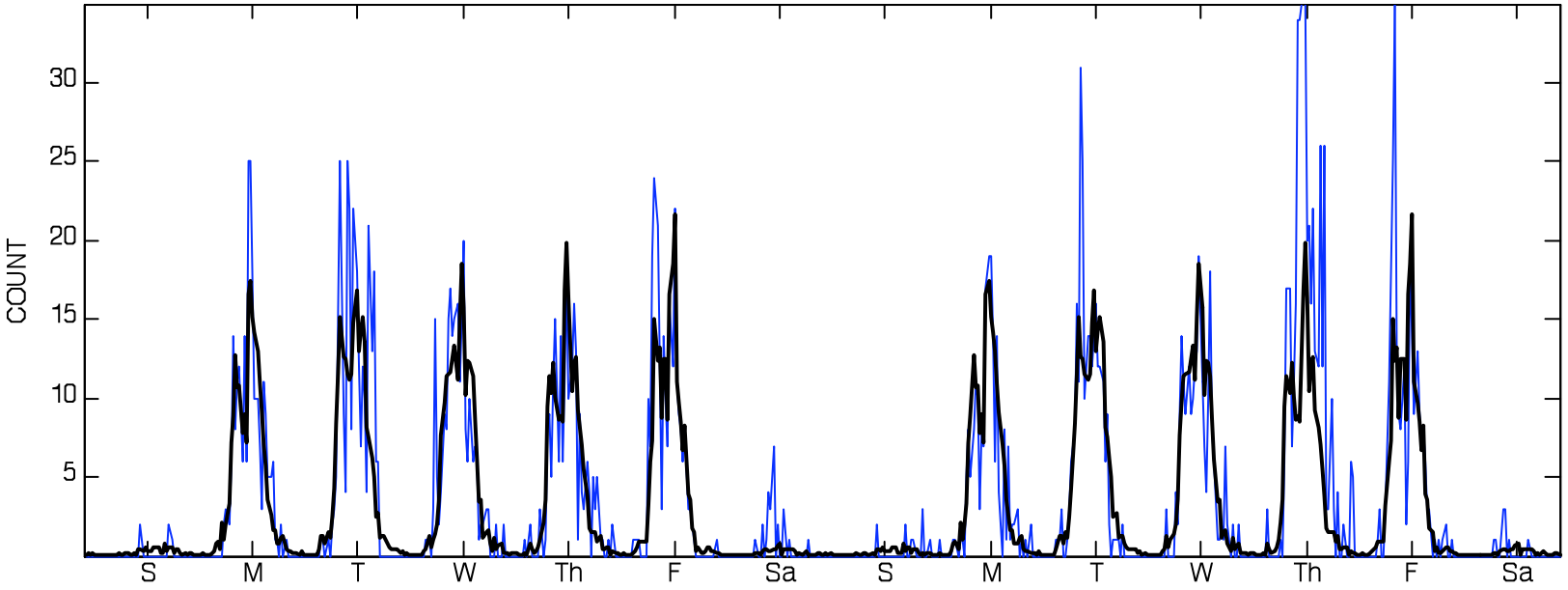
Baseline Poisson Model without Events



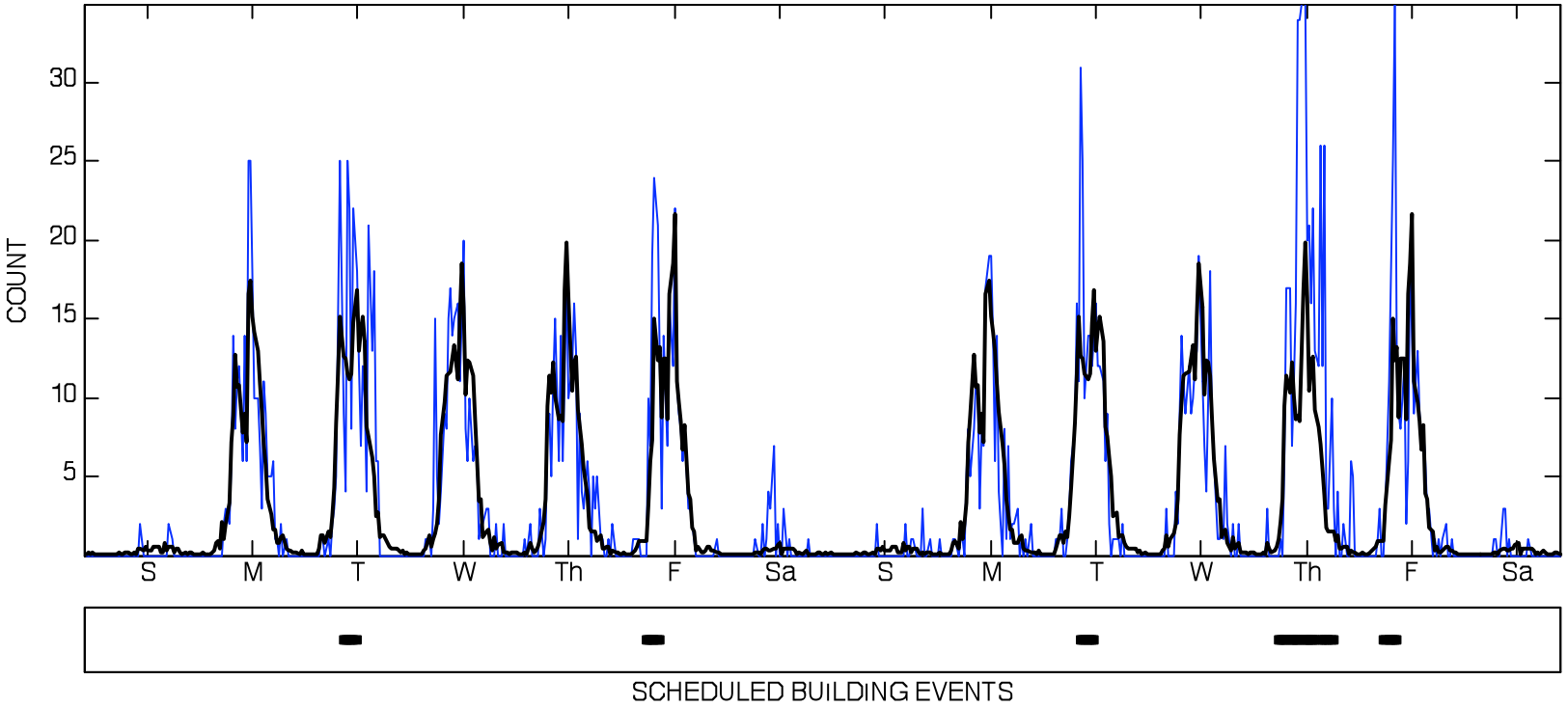
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

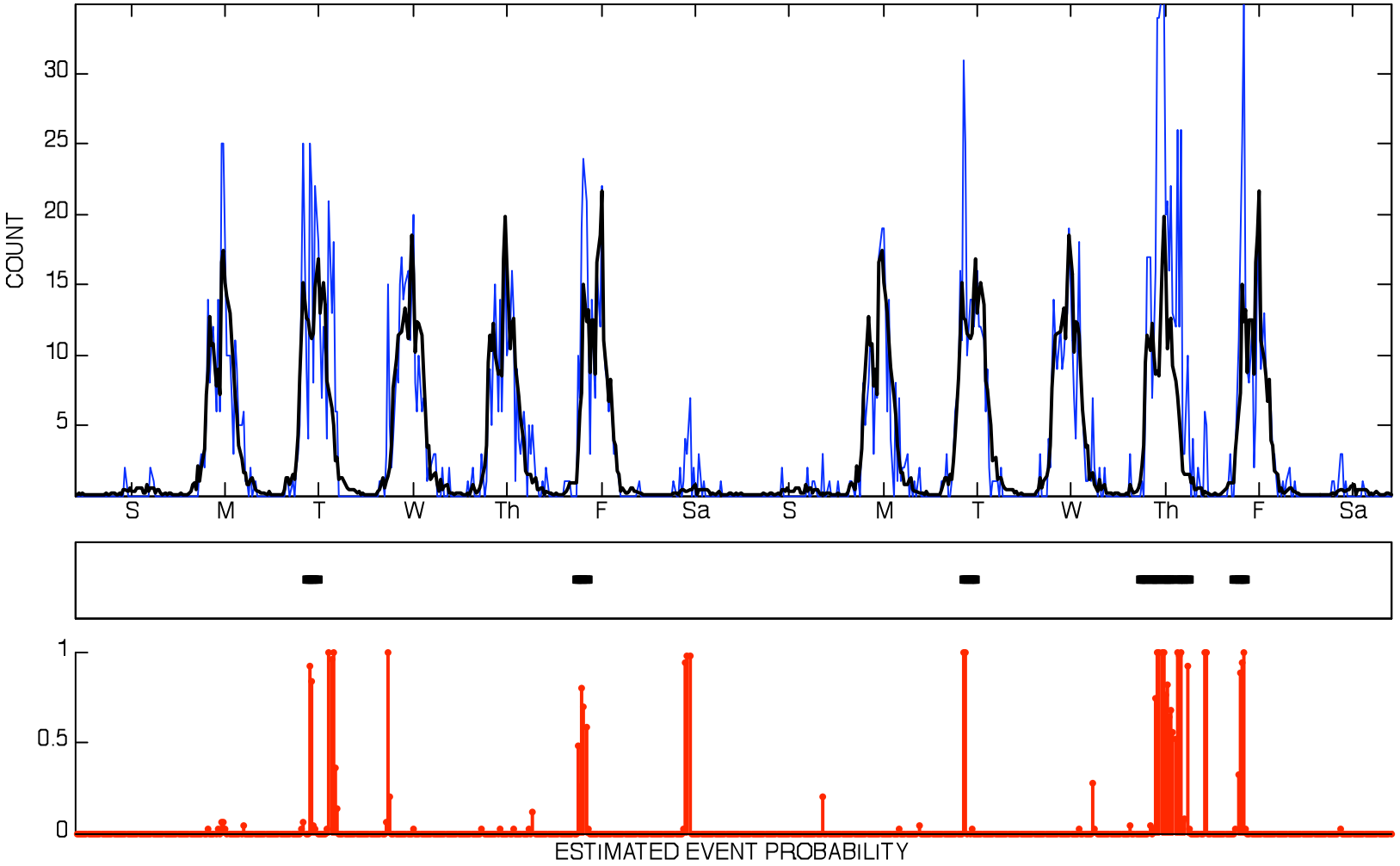


UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

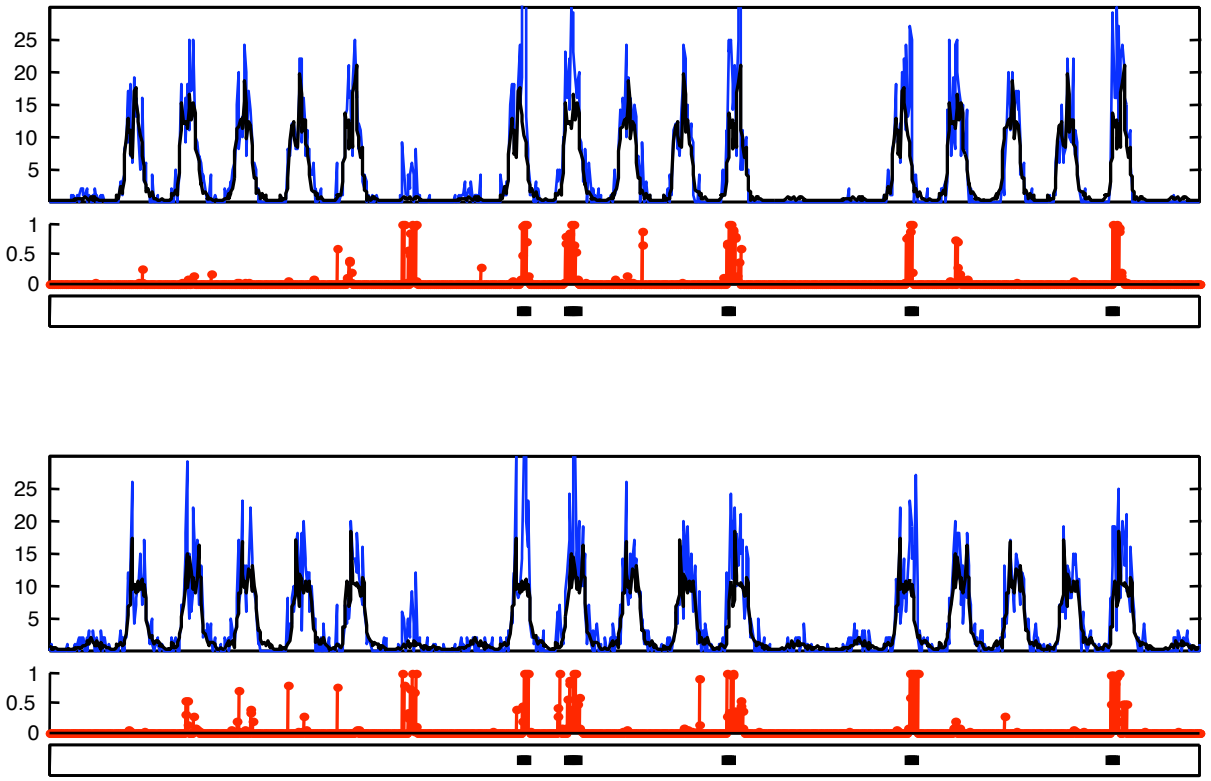



COUNT
30
25
20
15
10
5
S
M
T
W
Th
F
Sa
S
M
T
W
Th
F
Sa
1
0.5
0
ESTIMATED EVENT PROBABILITY

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

6 Weeks of Building Data

- **Actual data**
- **Mean profile λ(t)**
- **Event probability p(z)**
- **Actual events**



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Evaluation

- Compare predictions with ground truth
  - Building data: official schedule of events
  - Freeway: times/days of Dodger baseball games
- Ground truth is partial
  - Model can detect additional “unknown” events
- Important
  - Known events were hidden from model during training, training is completely unsupervised

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Fraction of known events that are detected:

Building Data

MMPP Model	Threshold Model
100.0%	89.7%
86.2%	82.8%
79.3%	65.5%

Traffic Data

MMPP Model	Threshold Model
100.0%	85.9%
100.0%	82.1%
100.0%	66.7%
97.4%	55.1%

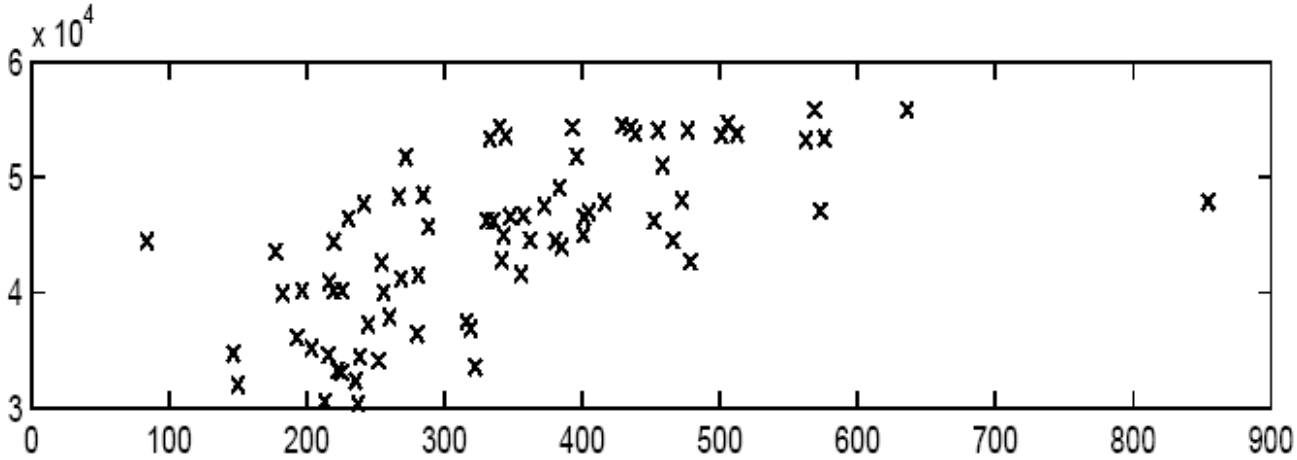
Adjust priors on transition probabilities and
on thresholds to vary number of events predicted

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Estimation of Event Size

$N_E(t)$ estimates number of additional counts per event…

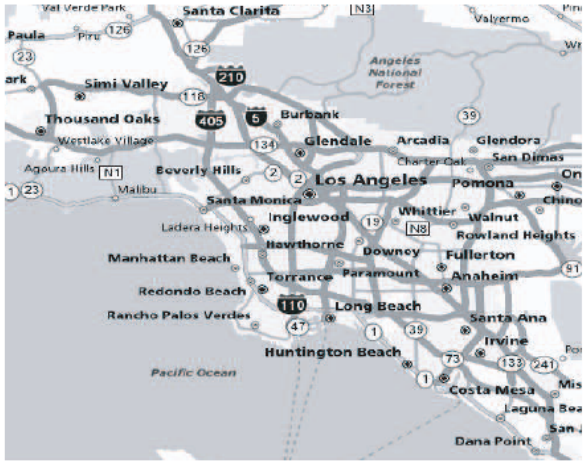
Baseball Game Attendance

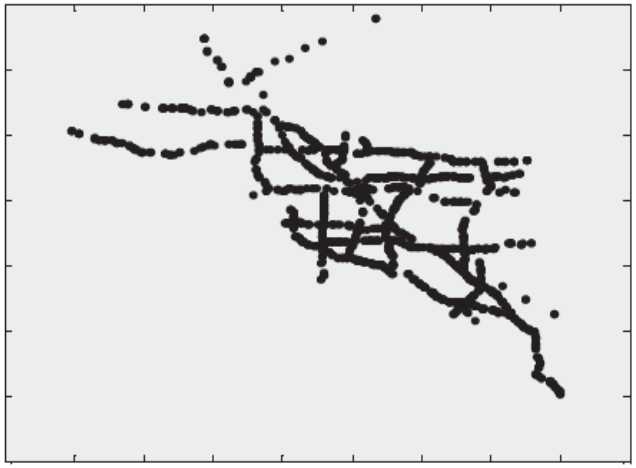


$N_E(t)$

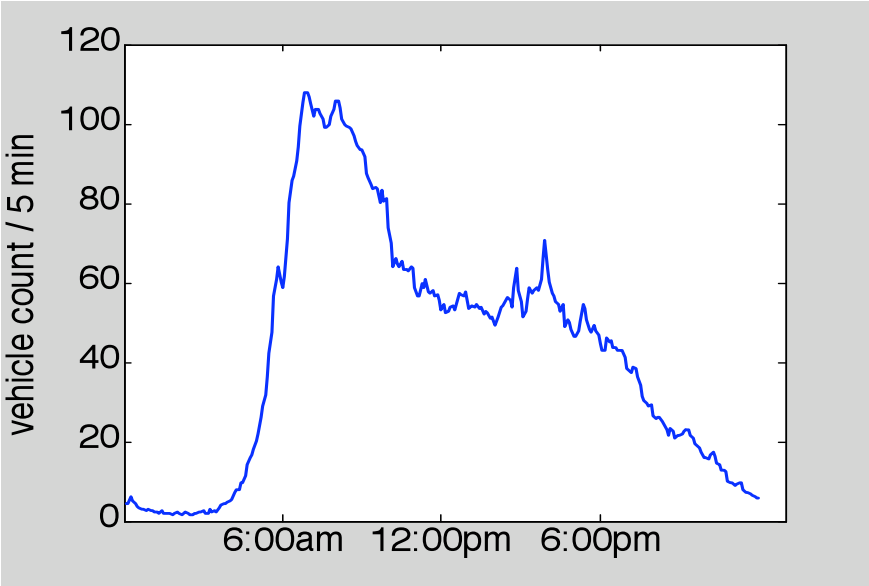
UCIrvine
University of California, Irvine

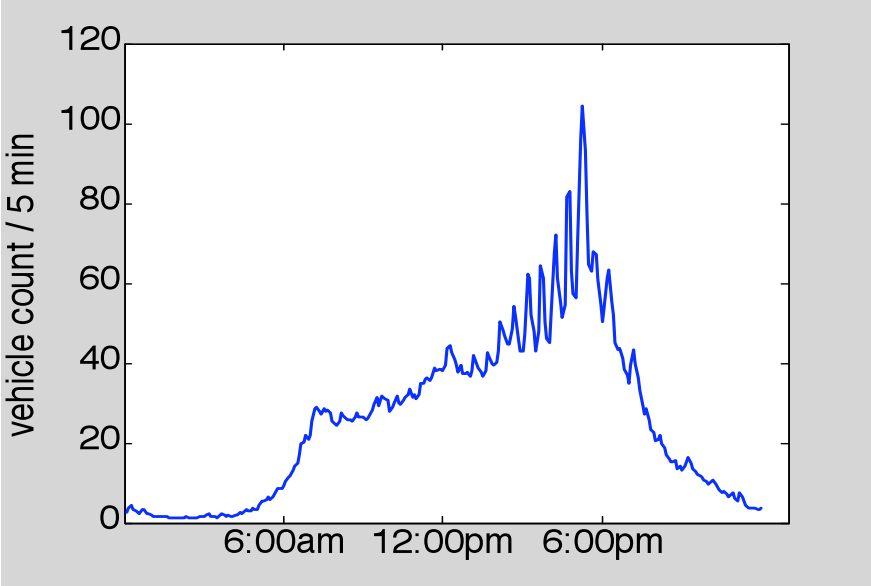
A Large-Scale Study





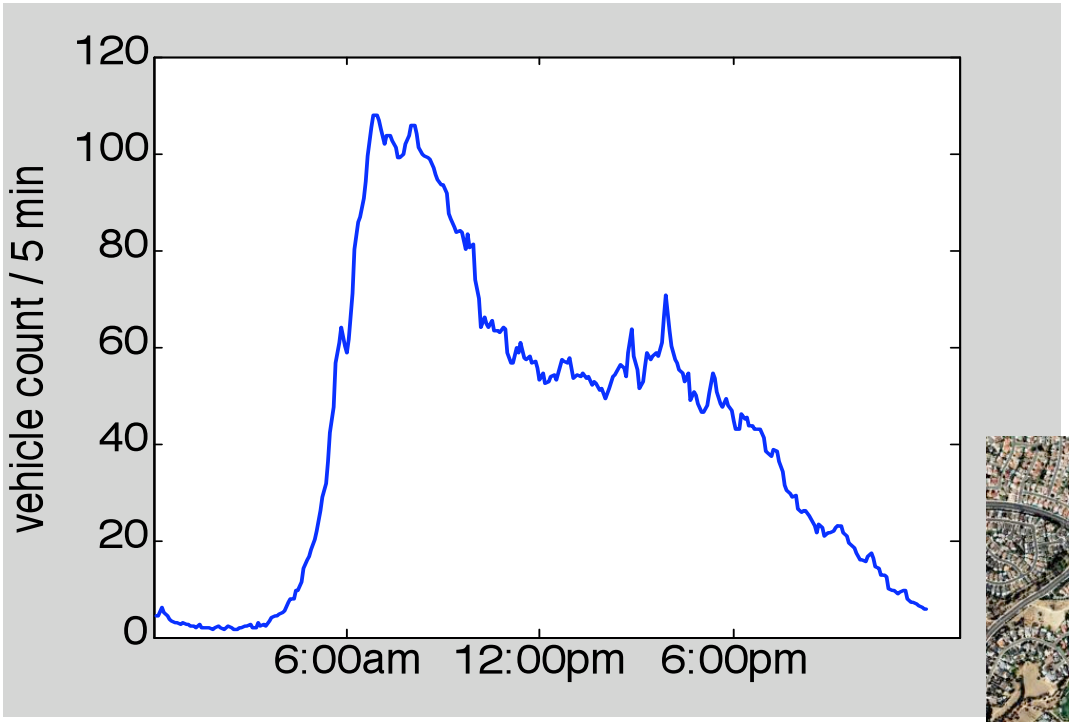
- All entrance and exit ramps in LA and OC for 7 months
- 1716 traffic sensors with data
  - 300 million measurements
  - 29% missing measurements per sensor on average





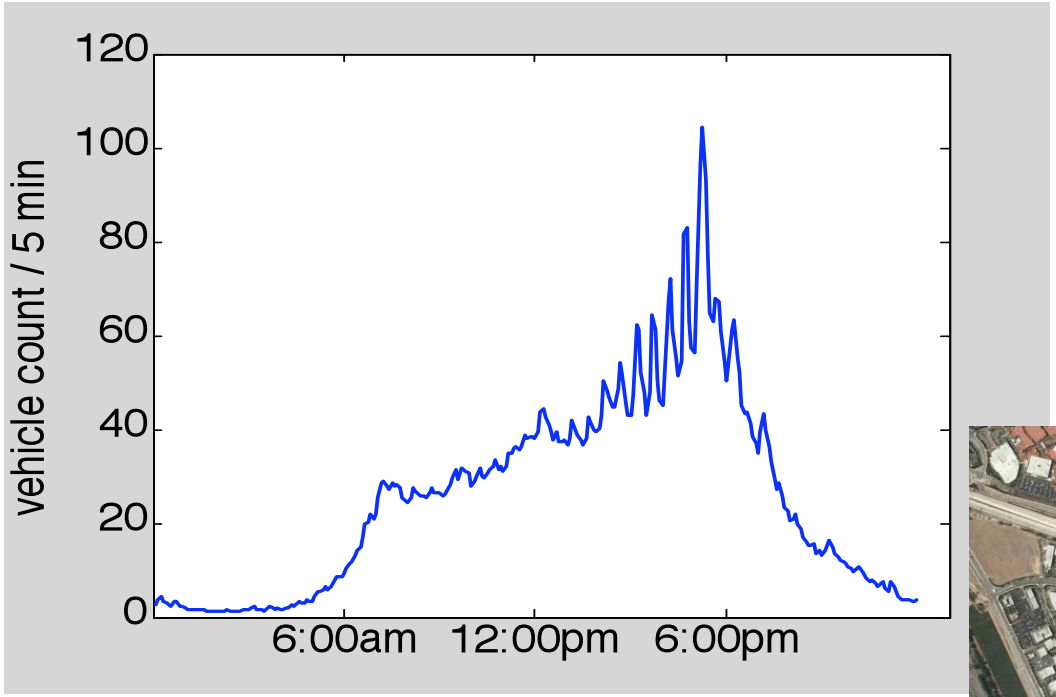
Profile Shapes (on-ramps)

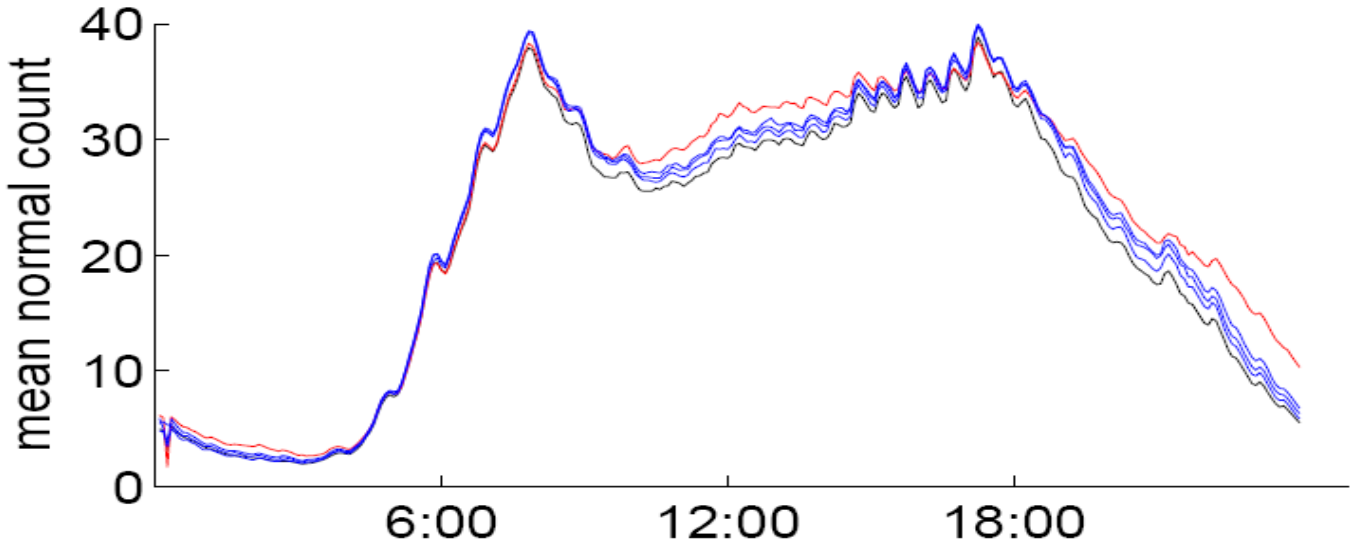
Suburban Area

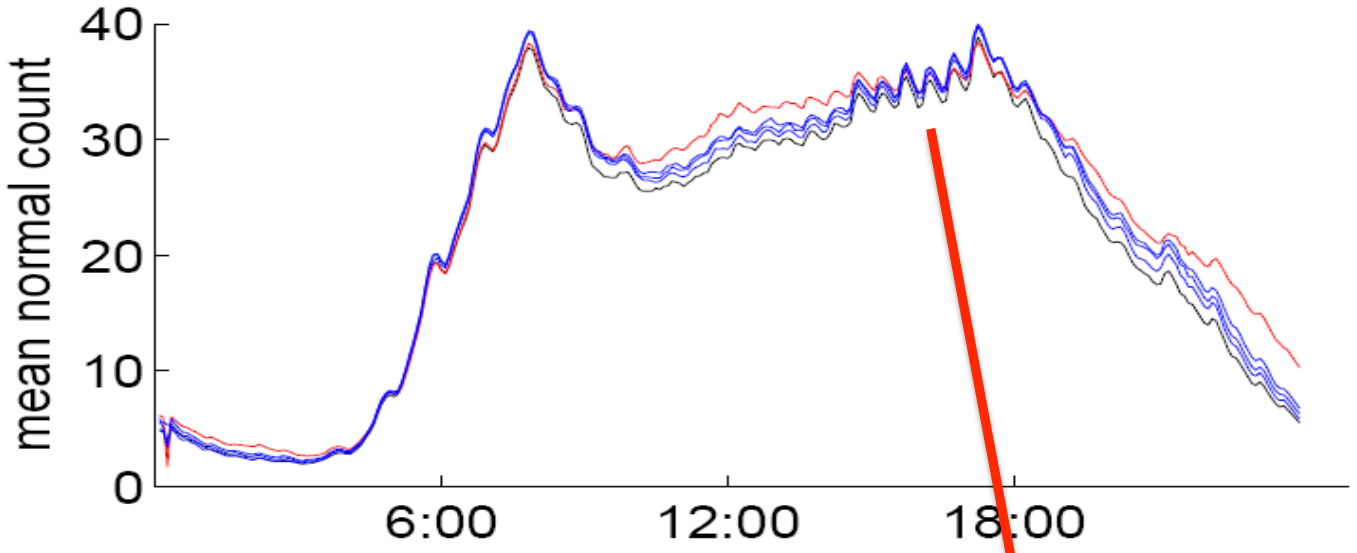


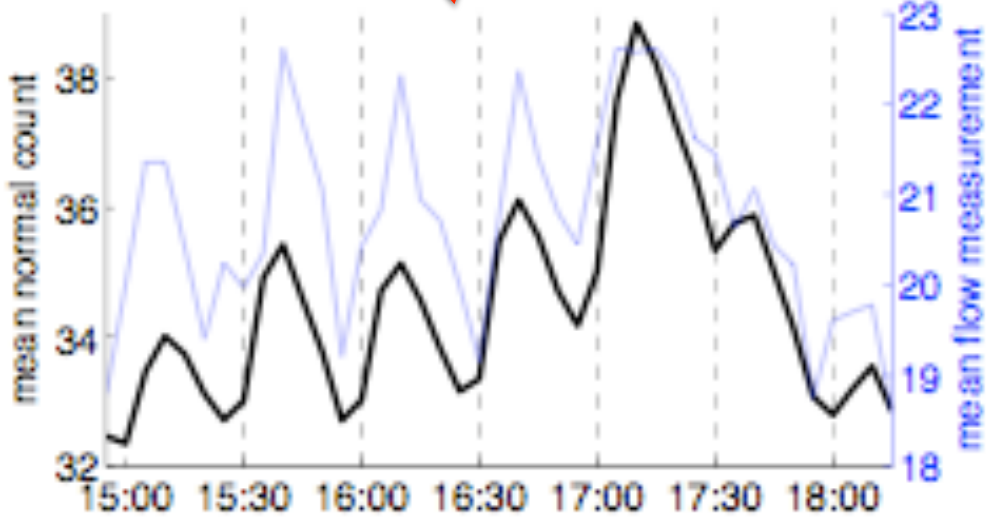
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

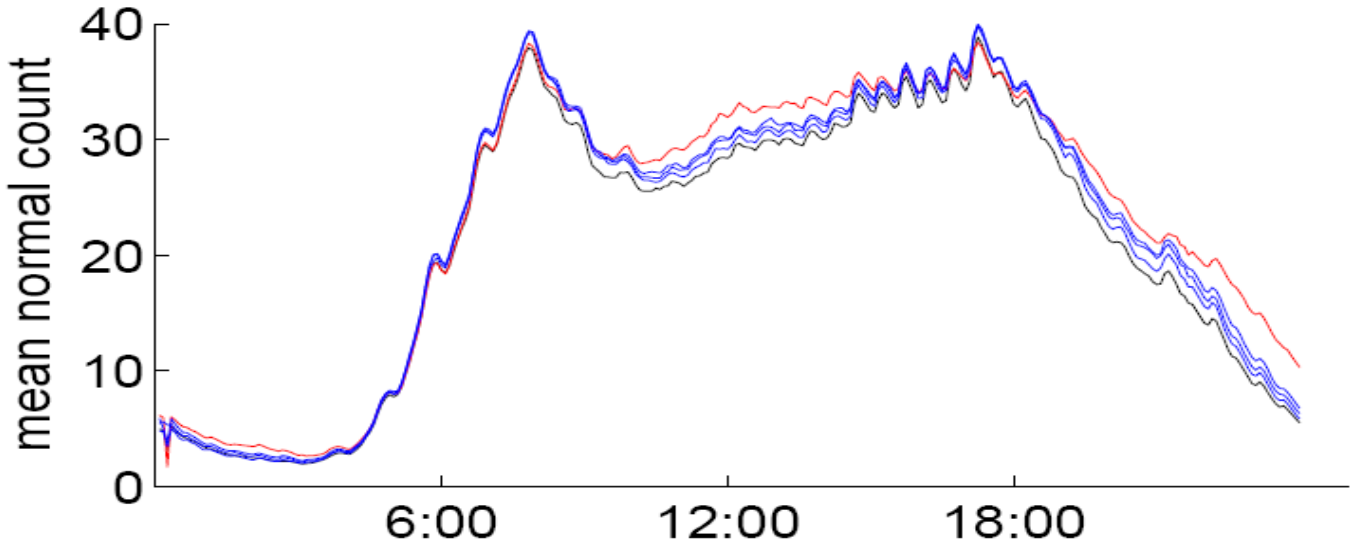
Industrial Area

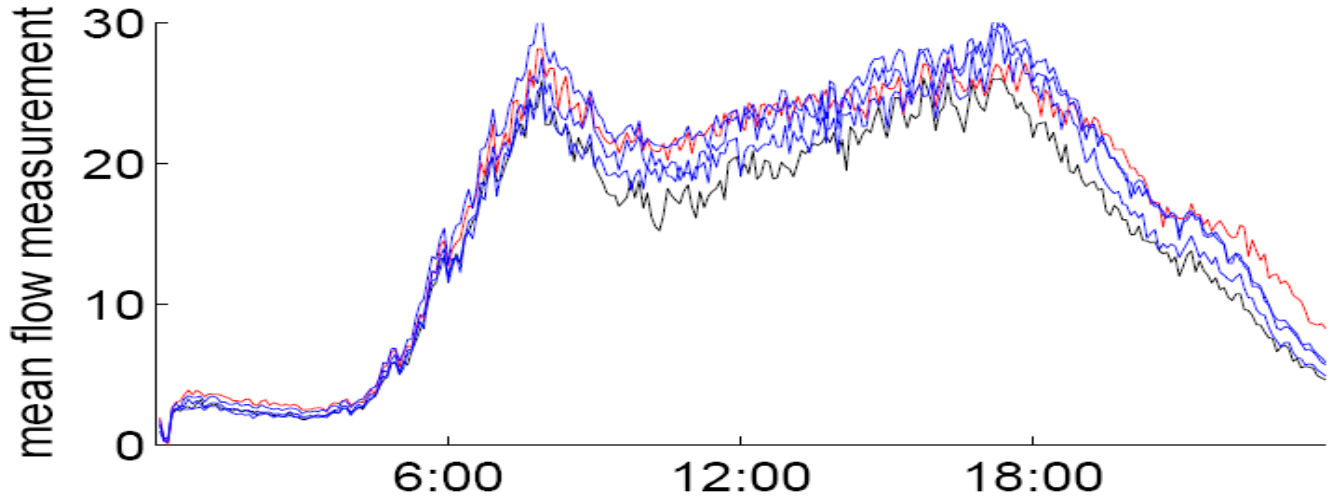


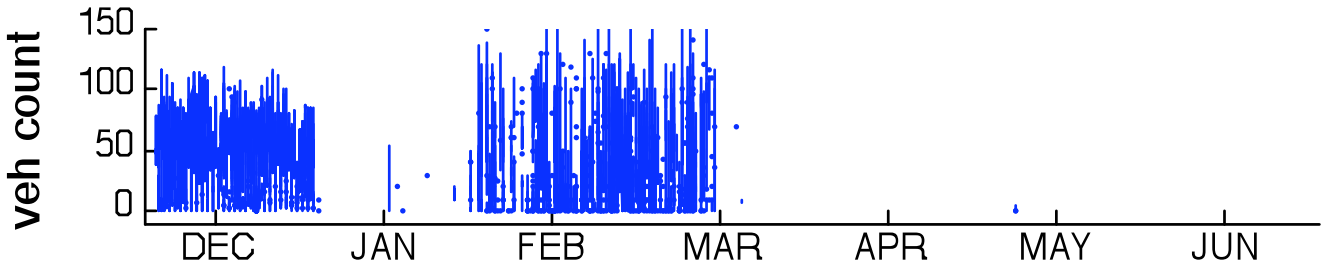

mean normal count
40
30
20
10
0
6:00
12:00
18:00

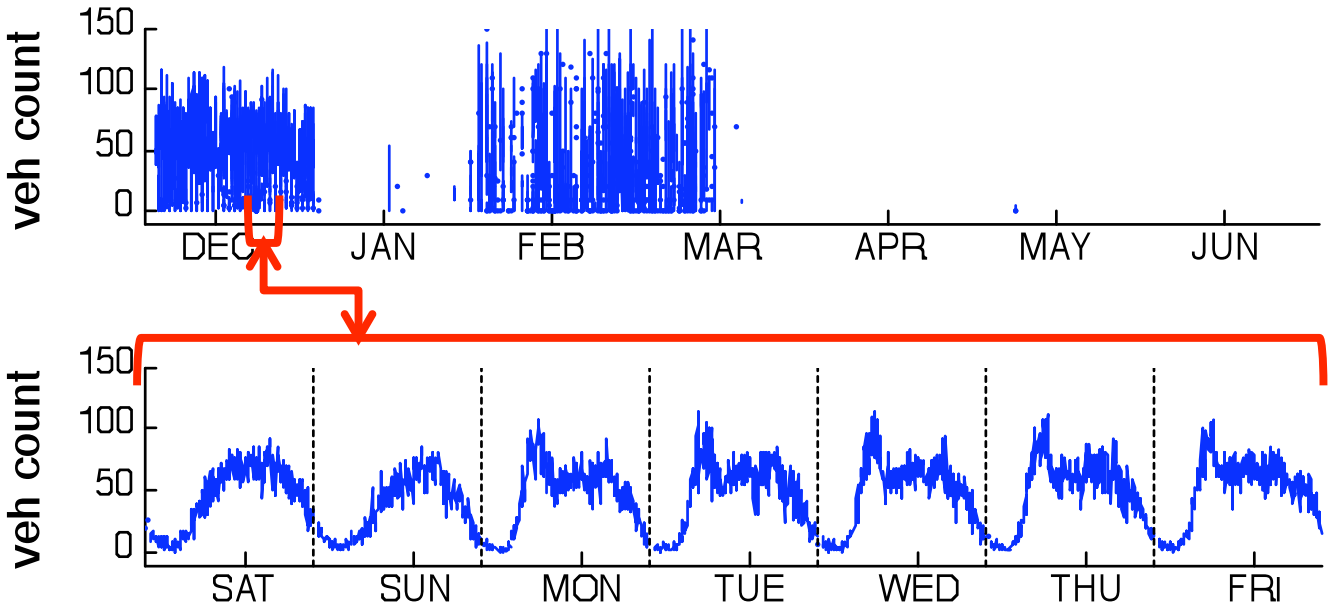

mean normal count
40
30
20
10
0
6:00
12:00
18:00

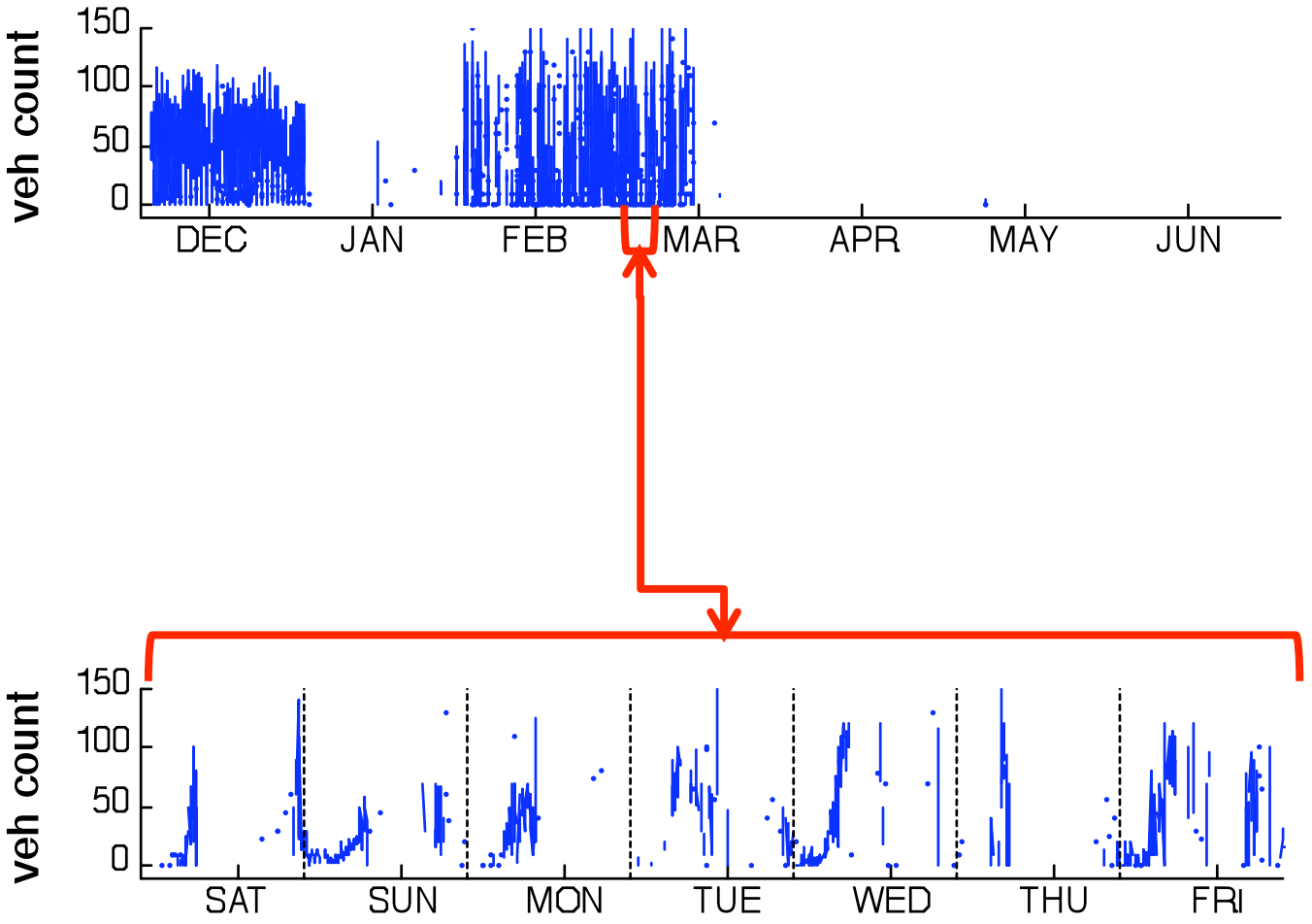

mean normal count
38
36
34
32
mean flow measurement
23
22
21
20
19
18
15:00
15:30
16:00
16:30
17:00
17:30
18:00

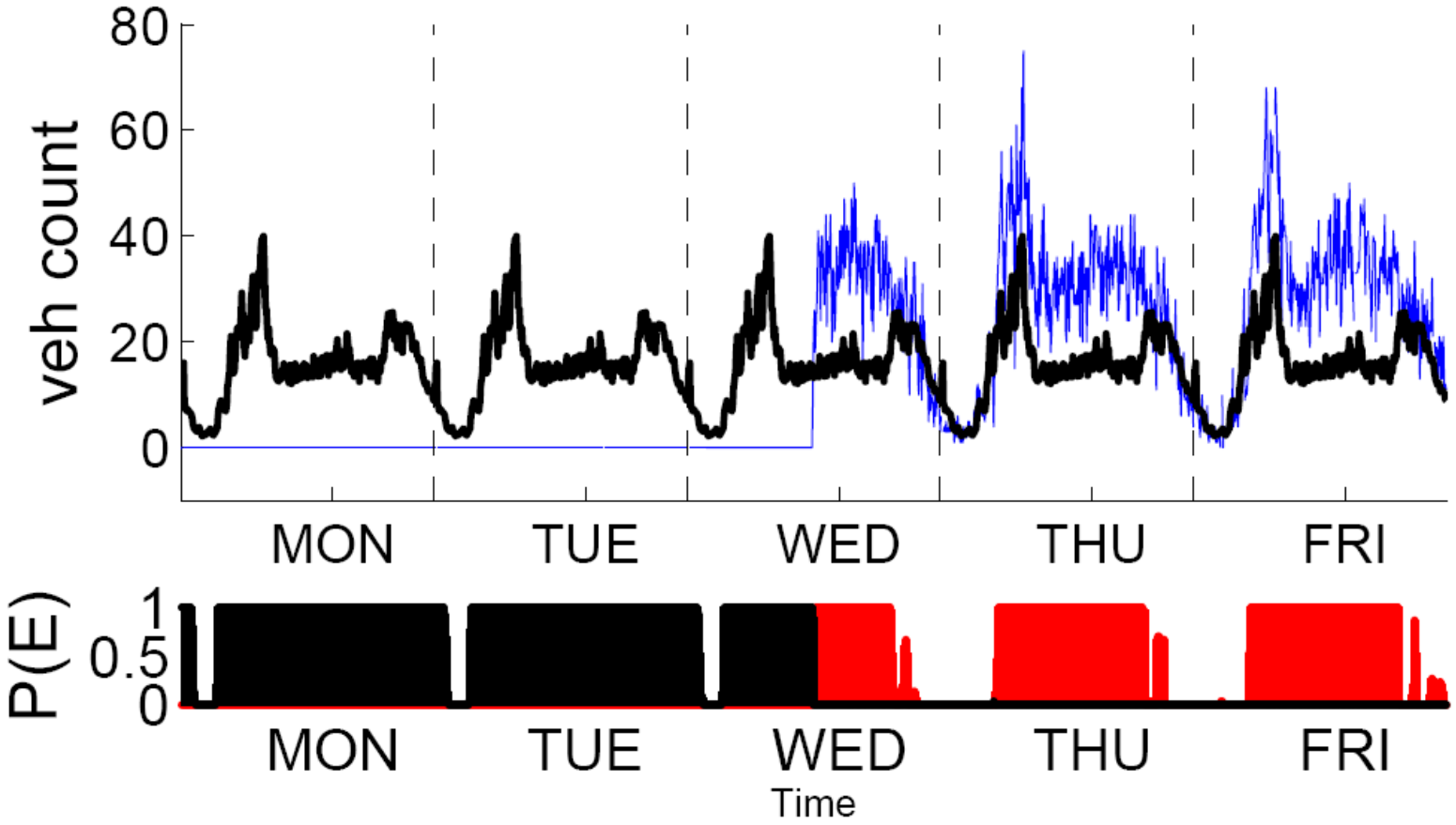

mean normal count
40
30
20
10
0
6:00
12:00
18:00

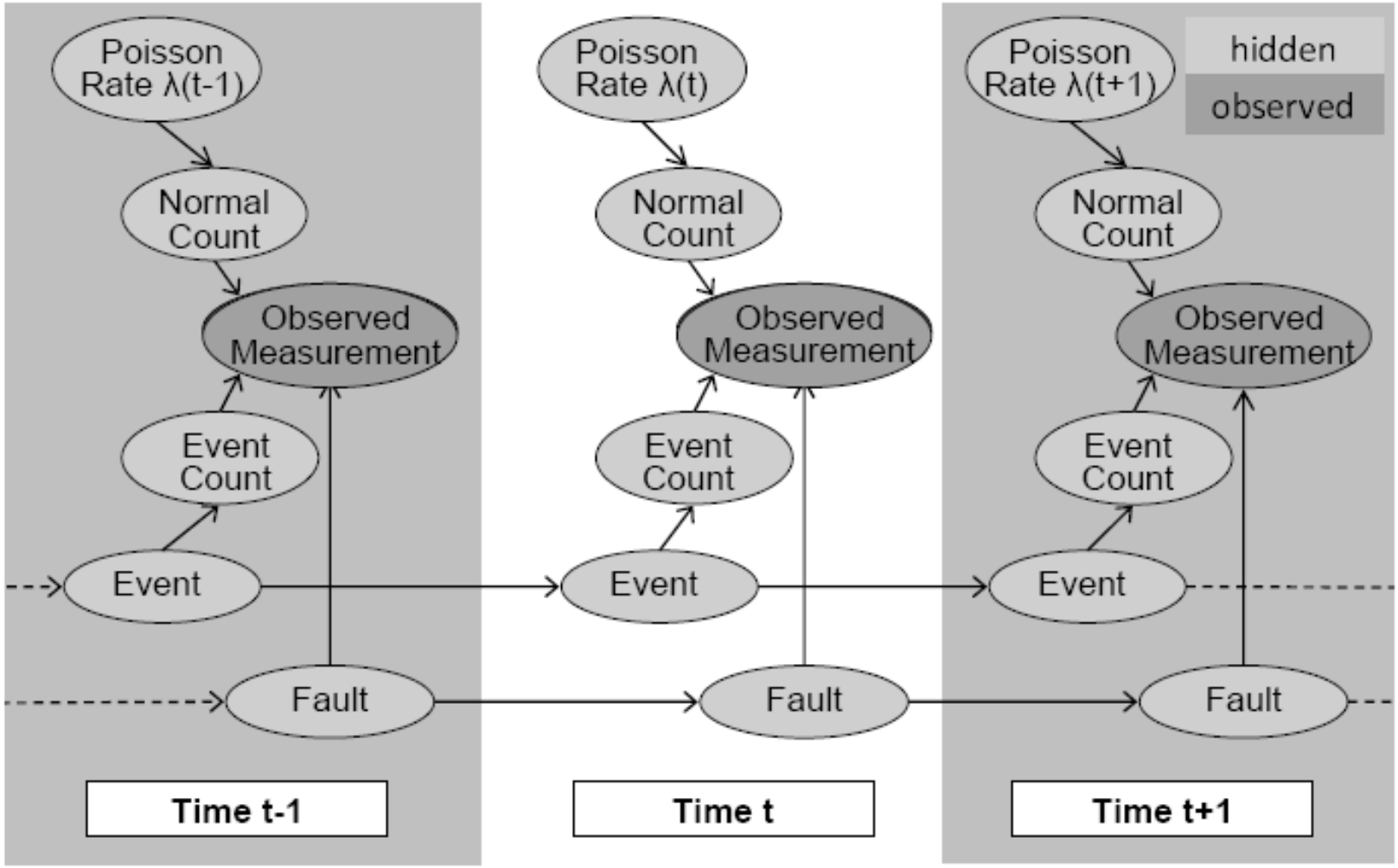

mean flow measurement
30
20
10
0
6:00
12:00
18:00


veh count
150
100
50
0
DEC
JAN
FEB
MAR
APR
MAY
JUN

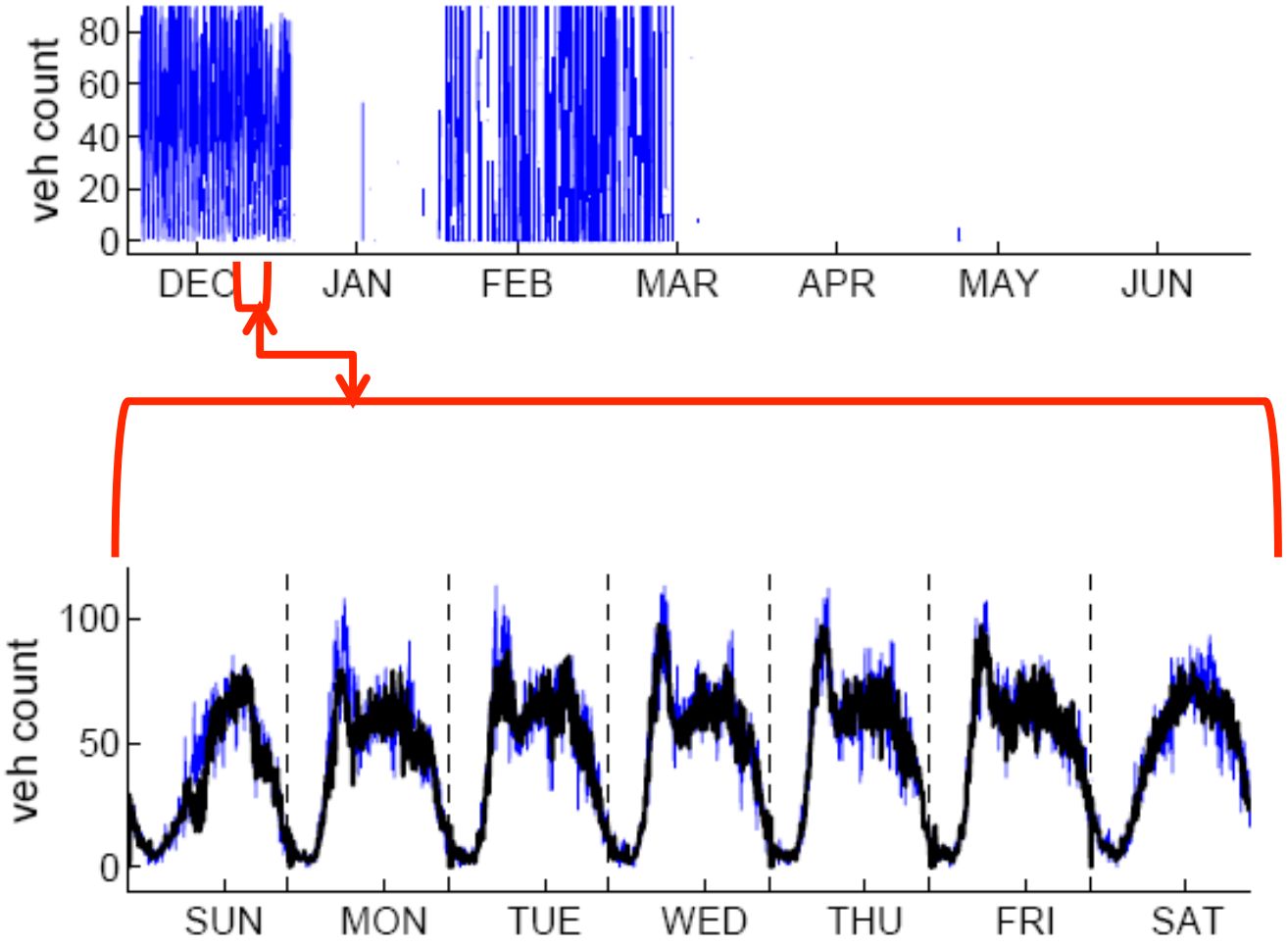

veh count
150
100
50
0
DEC
JAN
FEB
MAR
APR
MAY
JUN
veh count
150
100
50
0
SAT
SUN
MON
TUE
WED
THU
FRI


veh count
150
100
50
0
DEC
JAN
FEB
MAR
APR
MAY
JUN
veh count
150
100
50
0
SAT
SUN
MON
TUE
WED
THU
FRI


80
60
40
20
0
veh count
MON
TUE
WED
THU
FRI
P(E)
1
0.5
0
MON
TUE
WED
THU
FRI
Time



Hutchins, Ihler, Smyth, 2008


veh count
80
60
40
20
0
DEC
JAN
FEB
MAR
APR
MAY
JUN
veh count
100
50
0
SUN
MON
TUE
WED
THU
FRI
SAT

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

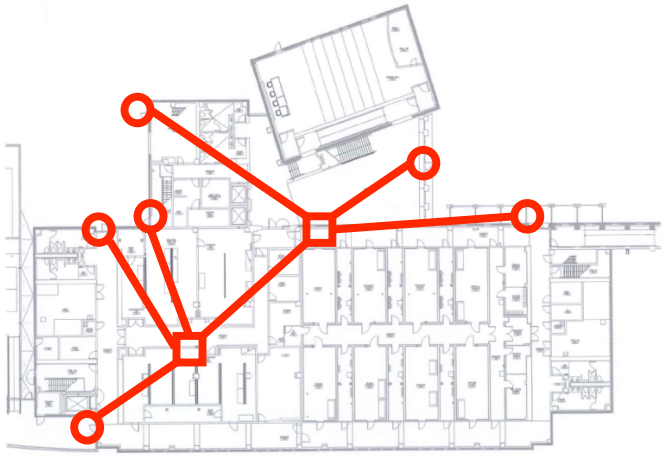
Animation showing spatio-temporal signature of traffic event

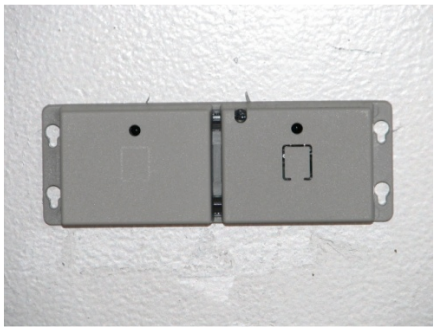
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Estimating building occupancy….

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Multiple Door Sensors





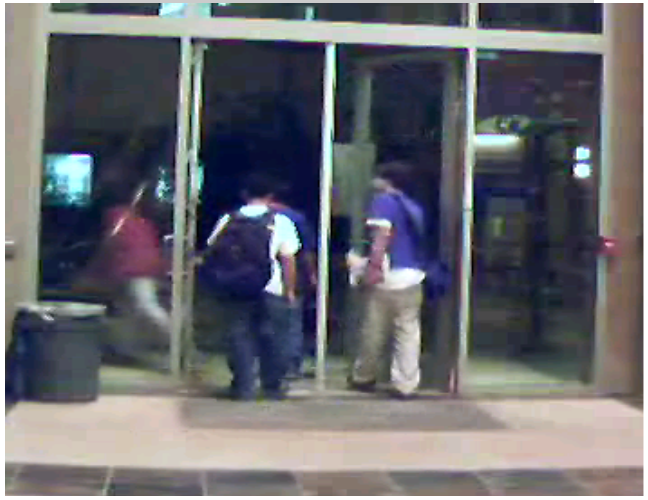
CalIt2 Building, UC Irvine campus

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Sensor Noise

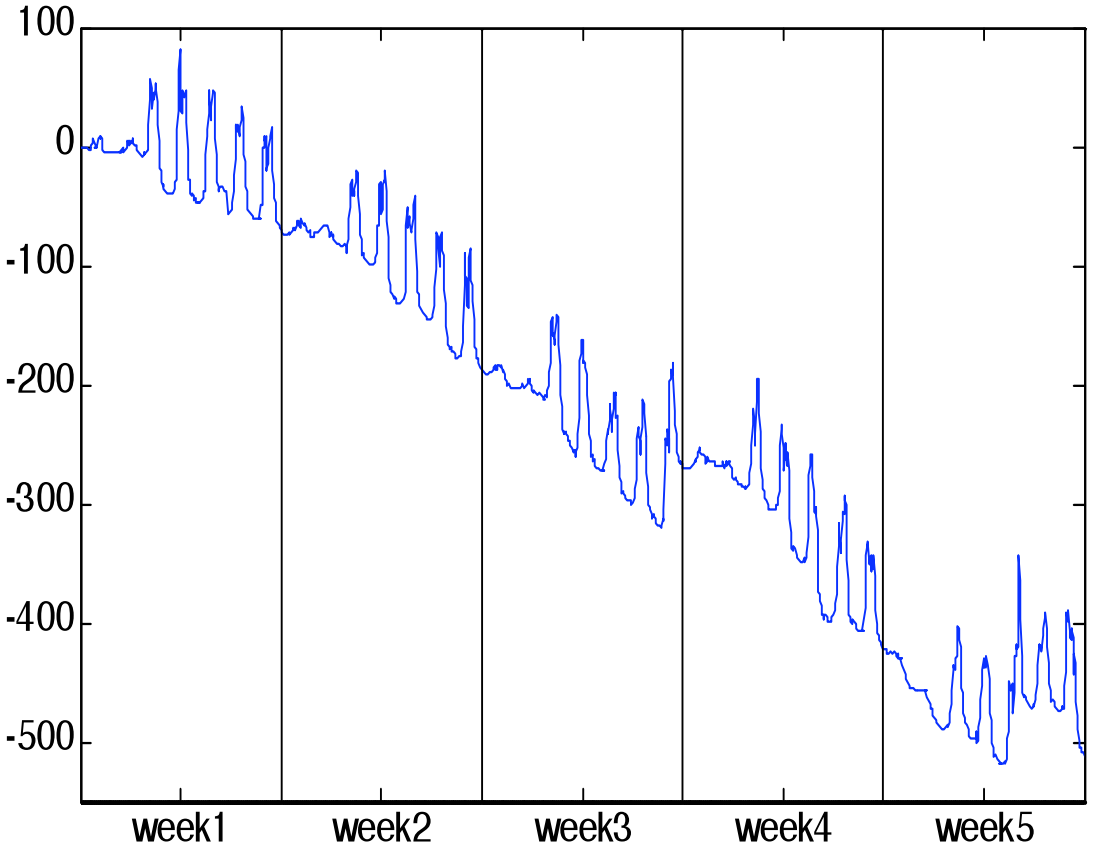
Over-counting

Under-counting



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

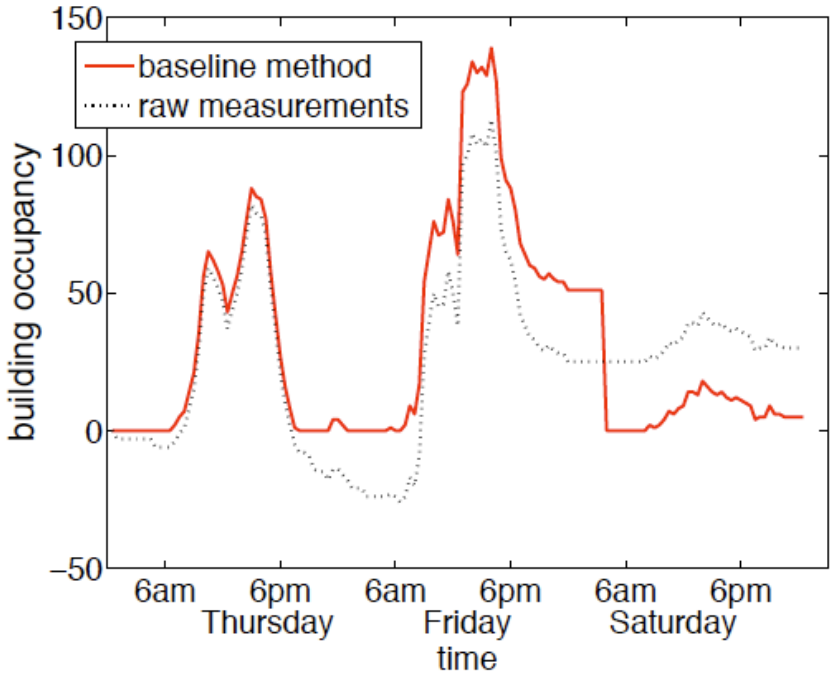
Difference of In and Out Counts



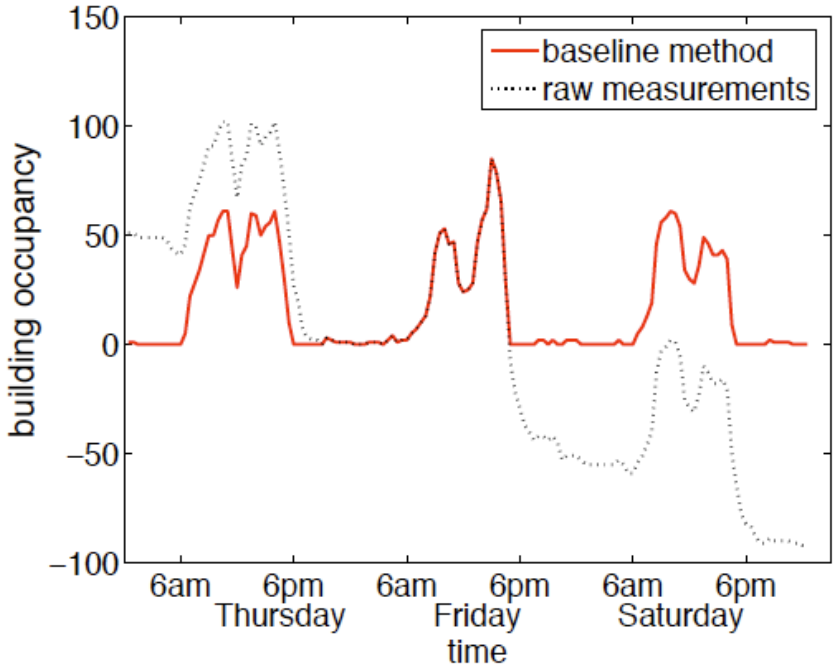
Systematic bias in raw counts

UCIrvine
University of California, Irvine

Extra In Counts



Extra Out Counts



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Probabilistic Model

- For each sensor:
  - Same MPMM model as before, learn normal Poisson behavior, subtracting out events
  - Add a sensor noise model:

$$o_t = a_t + \Upsilon_t^O - \Upsilon_t^U$$

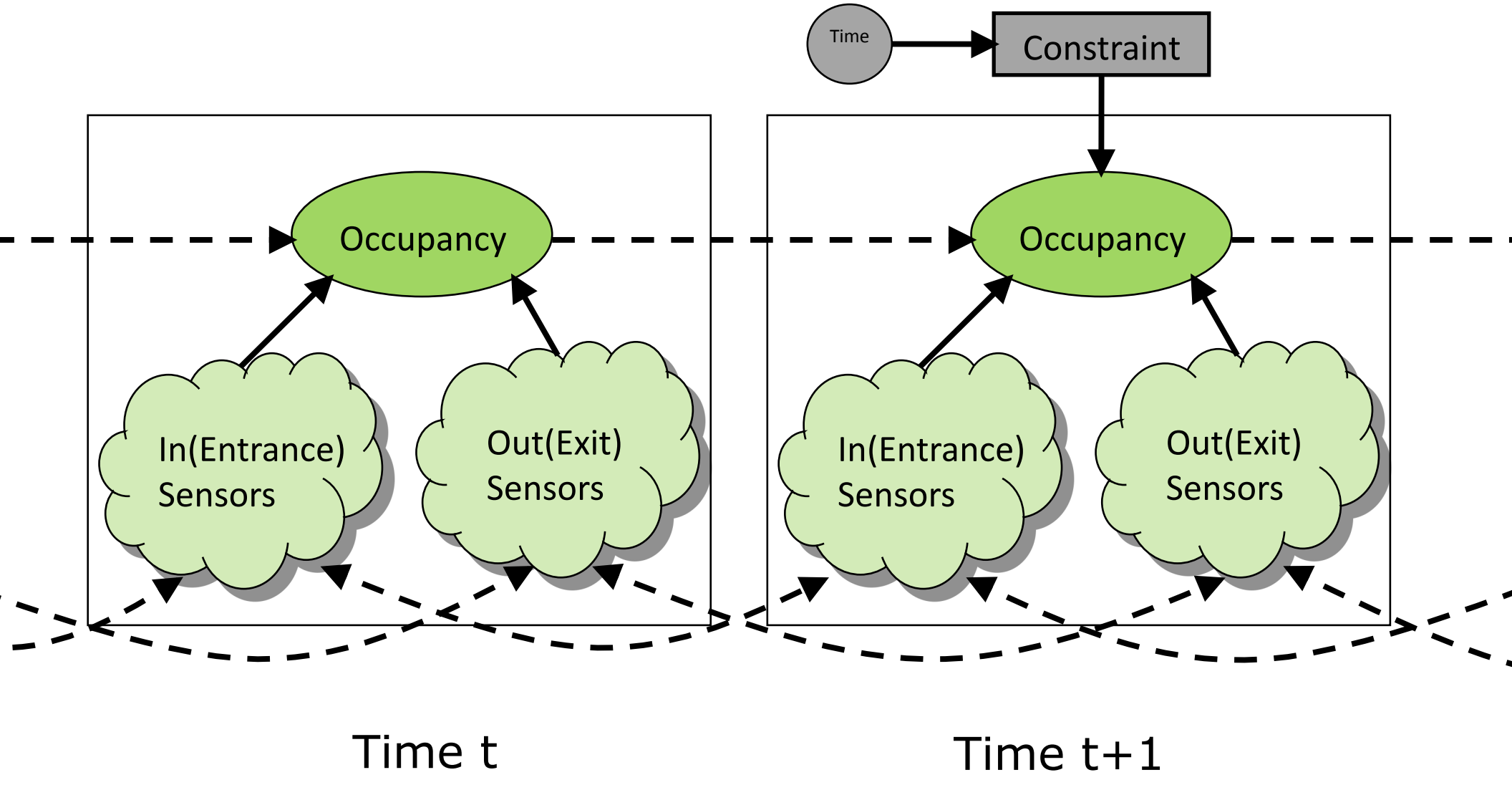
- Occupancy(t)

  = sums/differences of unobserved sensor counts + Occupancy(t-1)

- Add a prior on late night occupancy
  - Geometric(0.9) prior on occupancy at 3am - mean ~ 1.1 person

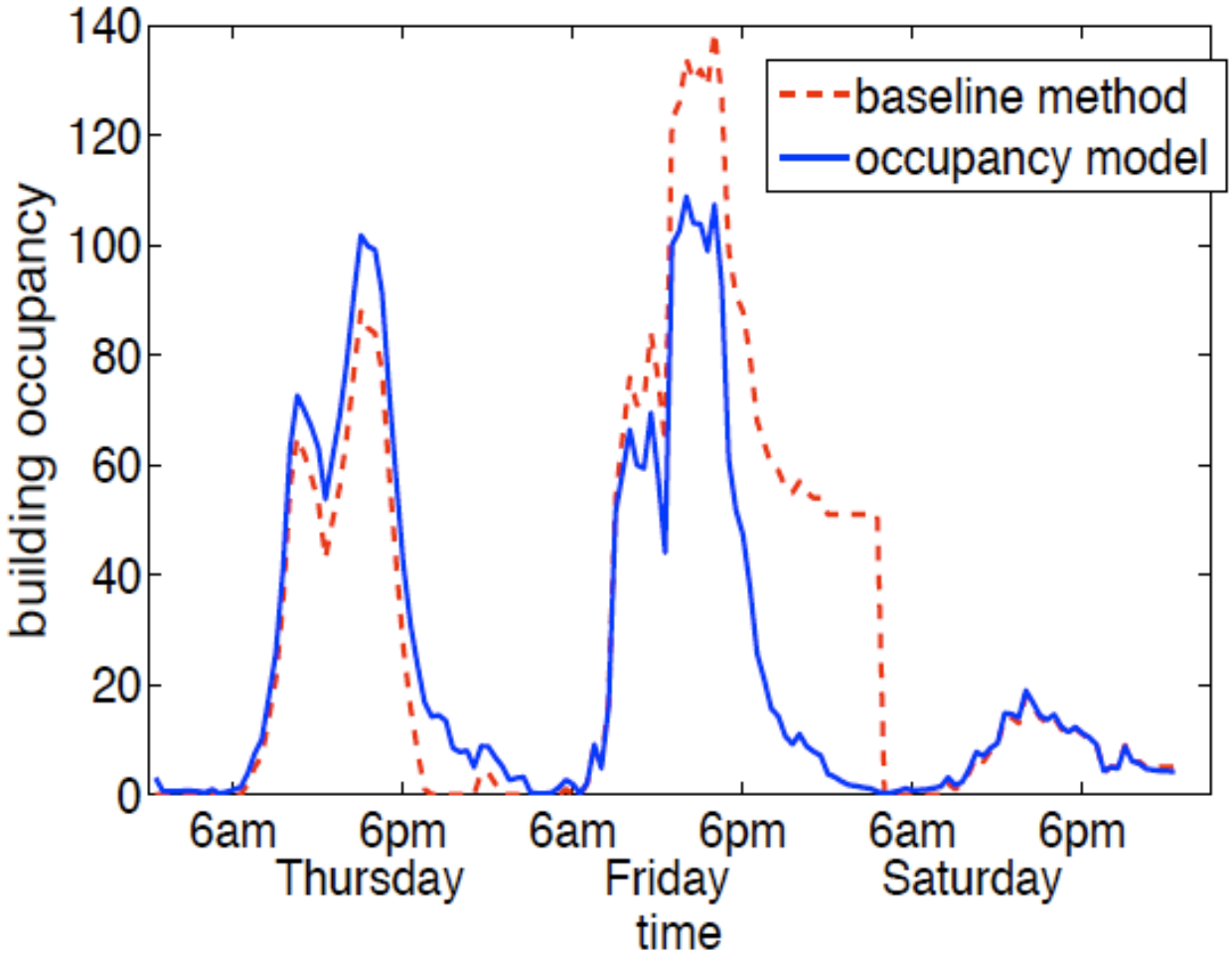
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Sketch of Graphical Model



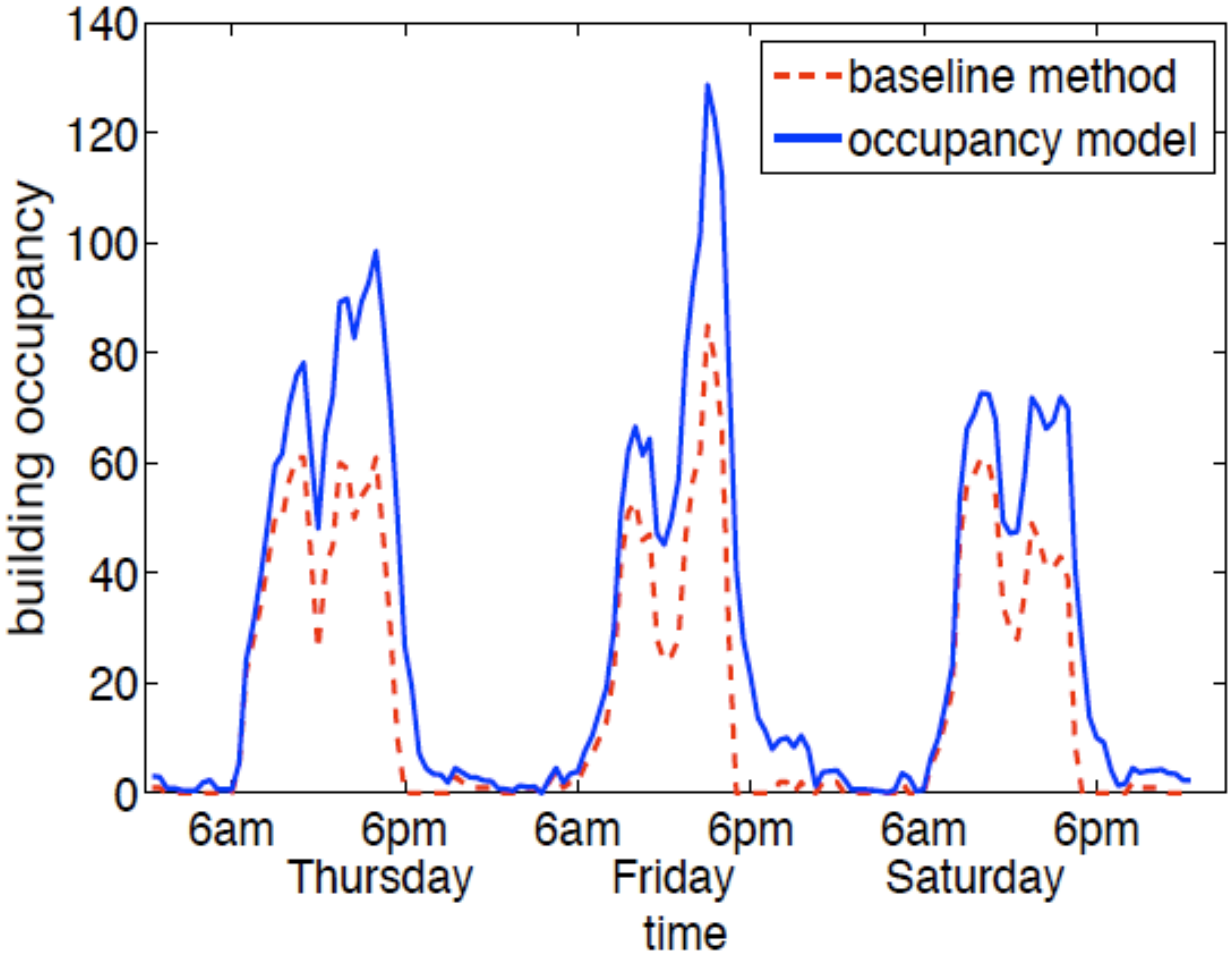
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Estimation with Extra In Counts



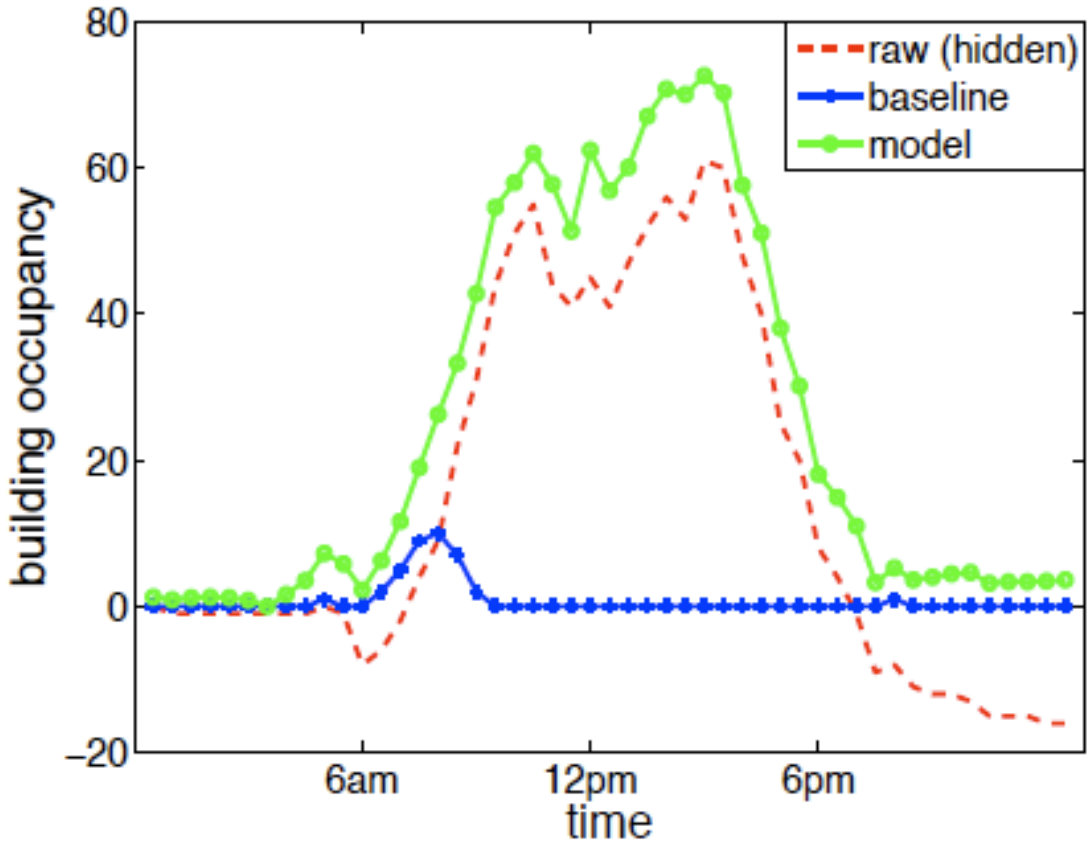
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Estimation with Extra Out Counts



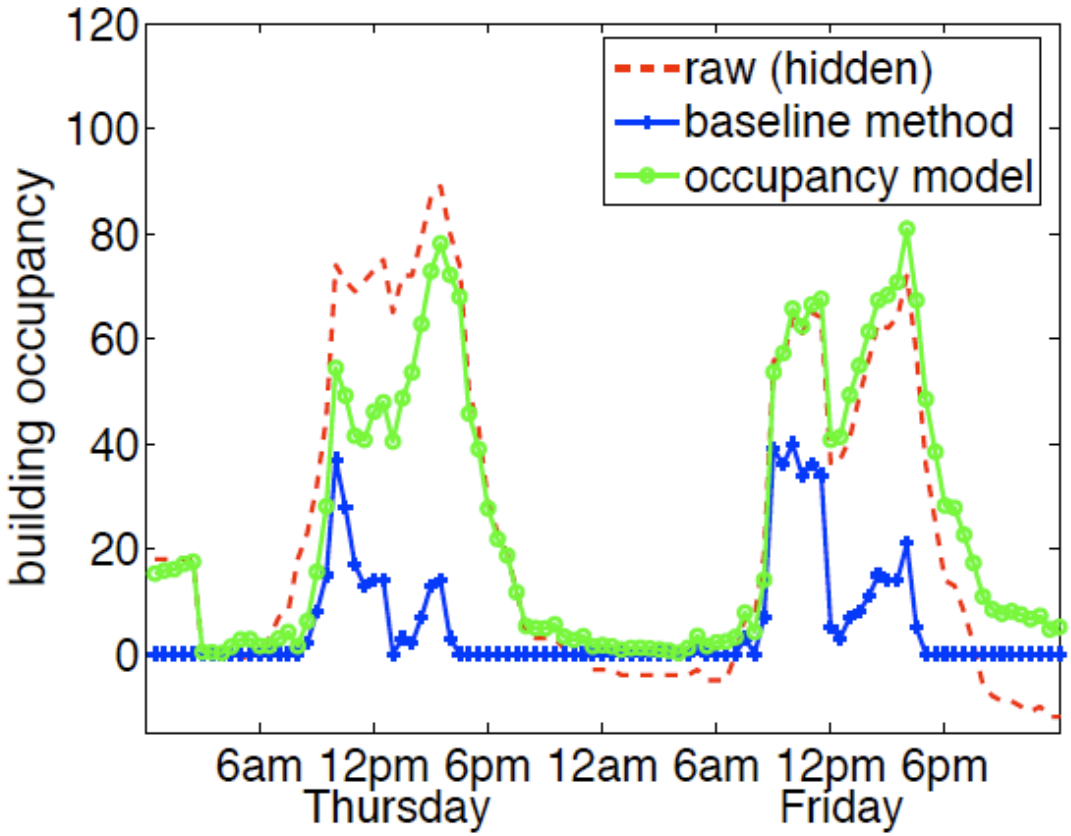
UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Missing Data



UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

Corrupted Data Experiment



Summary

- Time-series data of human behavior
  - Increasingly available
  - Interesting research problems and applications
  - Strong temporal regularities + bursts of anomalous events
  - Hidden Markov Poisson models are useful and scalable
  - Same models can be applied to multiple data types and applications

- Ongoing and future directions
  - Accident detection and characterization
  - Spatial models and missing data, e.g., sparse Markov networks
  - Linking traffic sensors to census and image data
  - Modeling trends, changes, random effects…