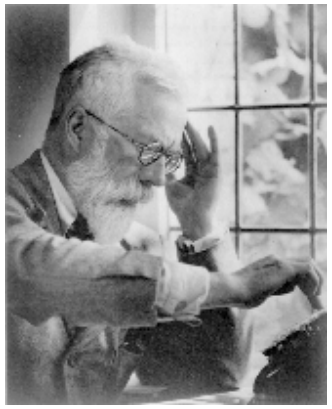
Modeling Highly Heterogeneous (Large) Data Sets:

Towards a Billion Models

Robert Grossman
University of Illinois at Chicago &
Open Data Group

Traditional Statistics

- One small data set
- A few attributes
- Vector-valued data



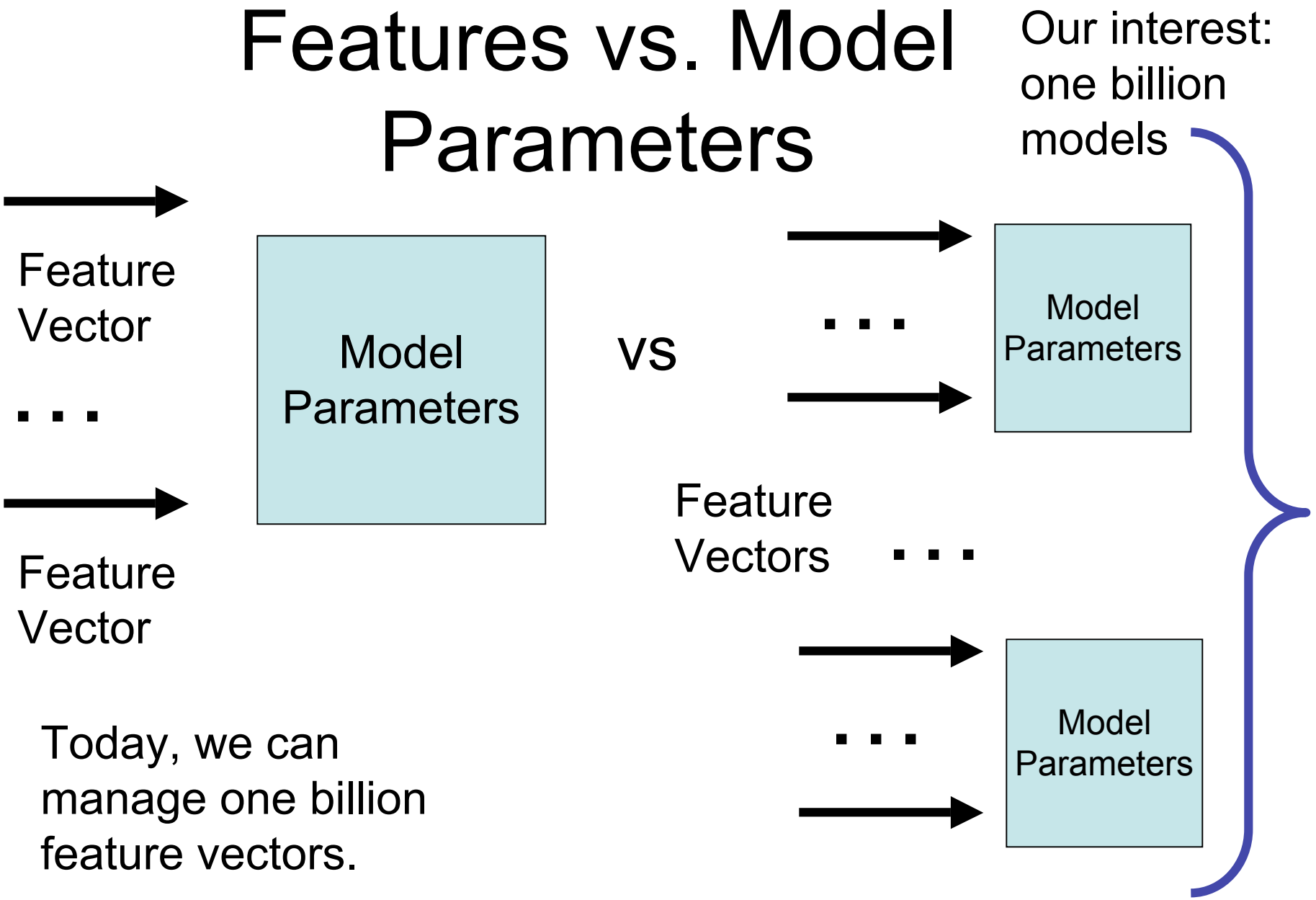
Data Mining

- Few large data sets
- Many attributes
- Complex data

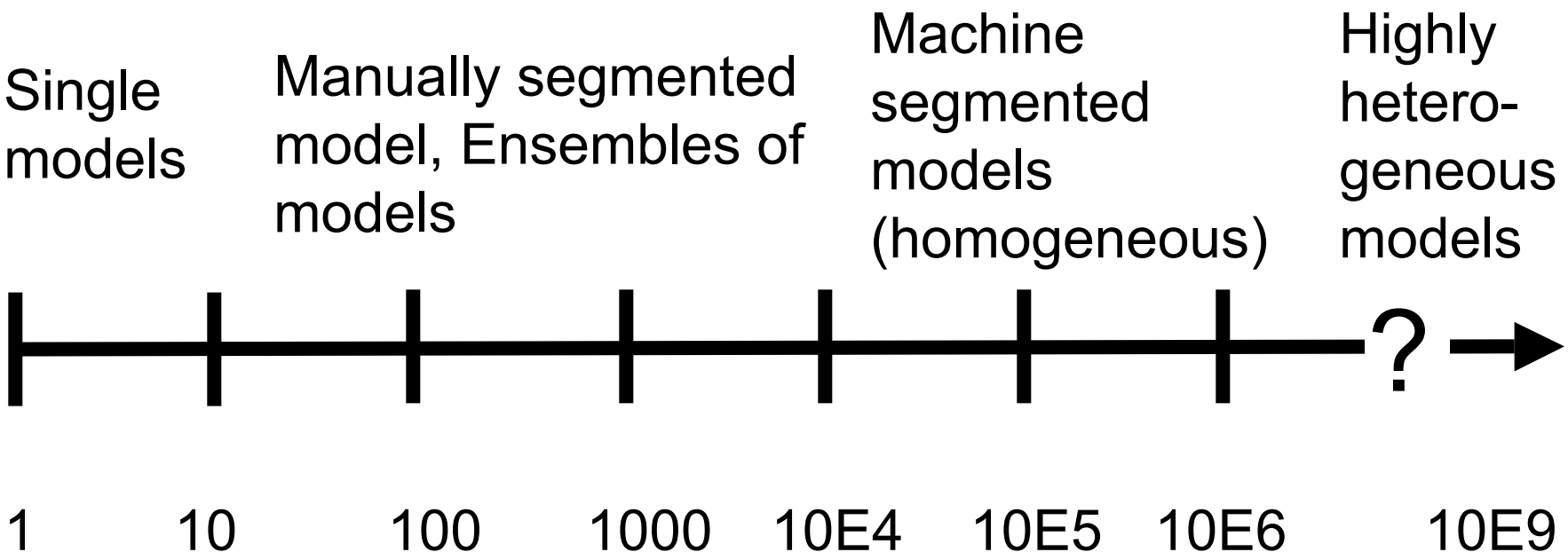
But Large Data Is Not Homogeneous

	Statistics	Large Data Today	Large, Highly Heterogeneous Data (Tomorrow)
Data	Small	Large	Large
Attributes	Few	Many	Many
Structure	Vector	Complex	Complex
Populations	One	Several	Many

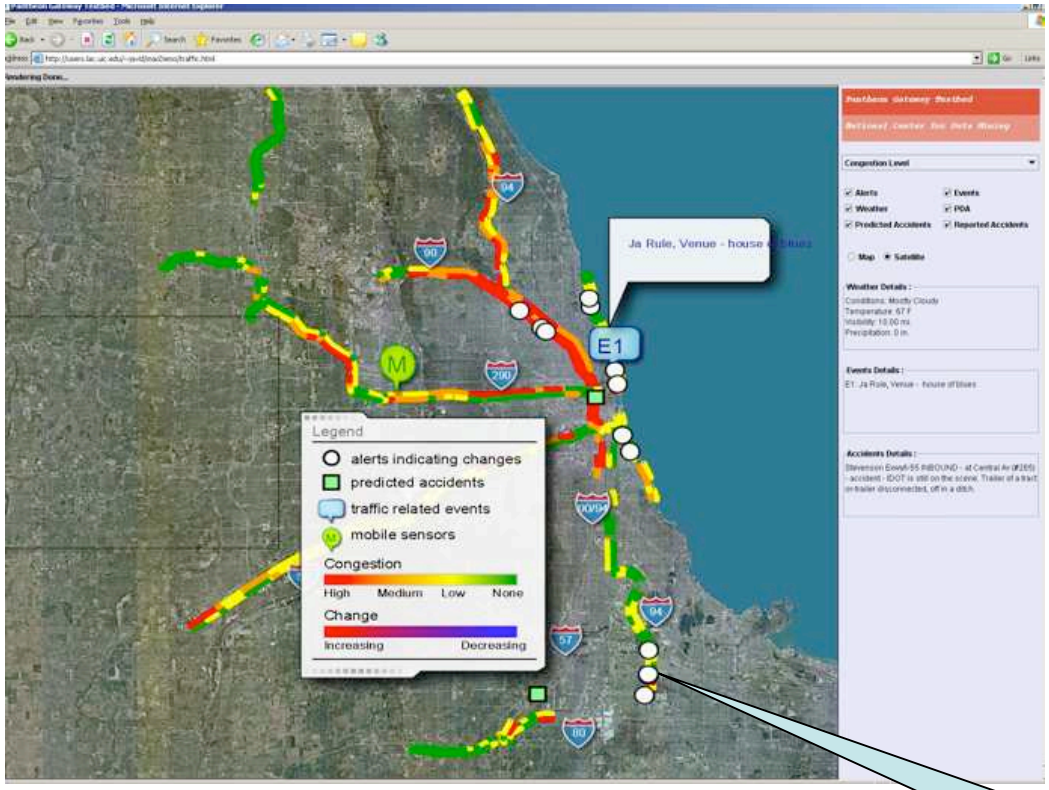
Features vs. Model Parameters



Progress to Date



Example 1 - 42,000 Models

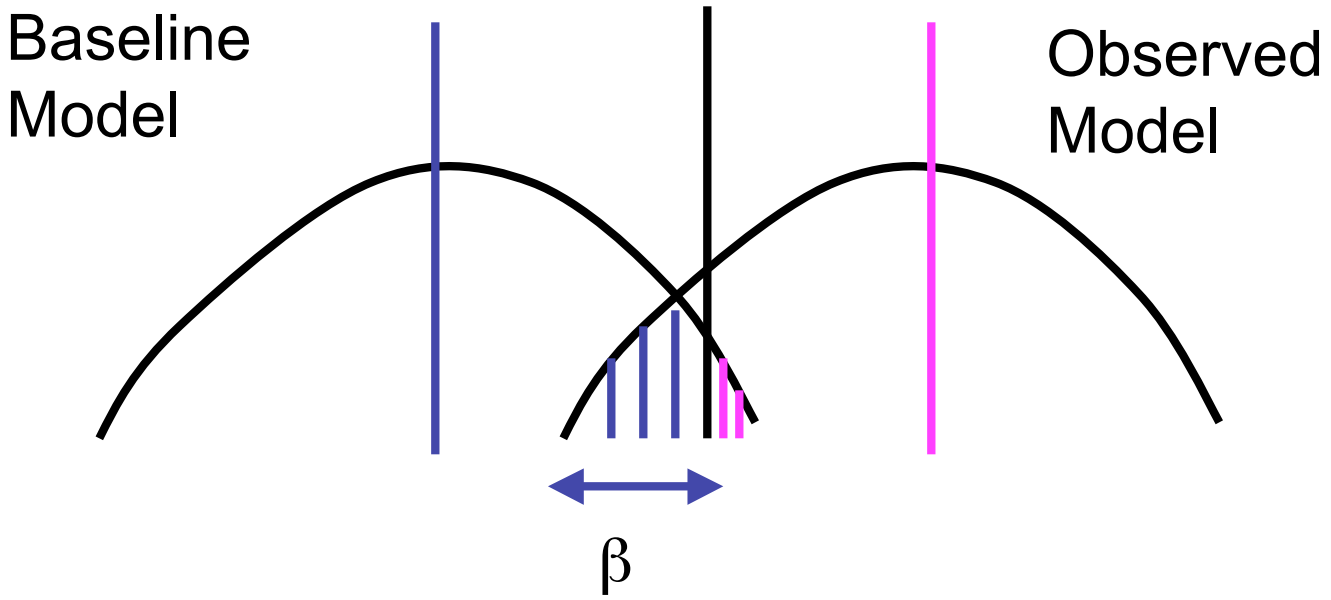


- Is the traffic speed and volume today (Tuesday, May 15, 4:30 pm,, no rain) **different** than the baseline model?
- Separate model for 7 days x 24 hours x 250 locations = 42,000 models

- 833 road sensors
- weather data (images, xml)
- text data about special events

Anomalies

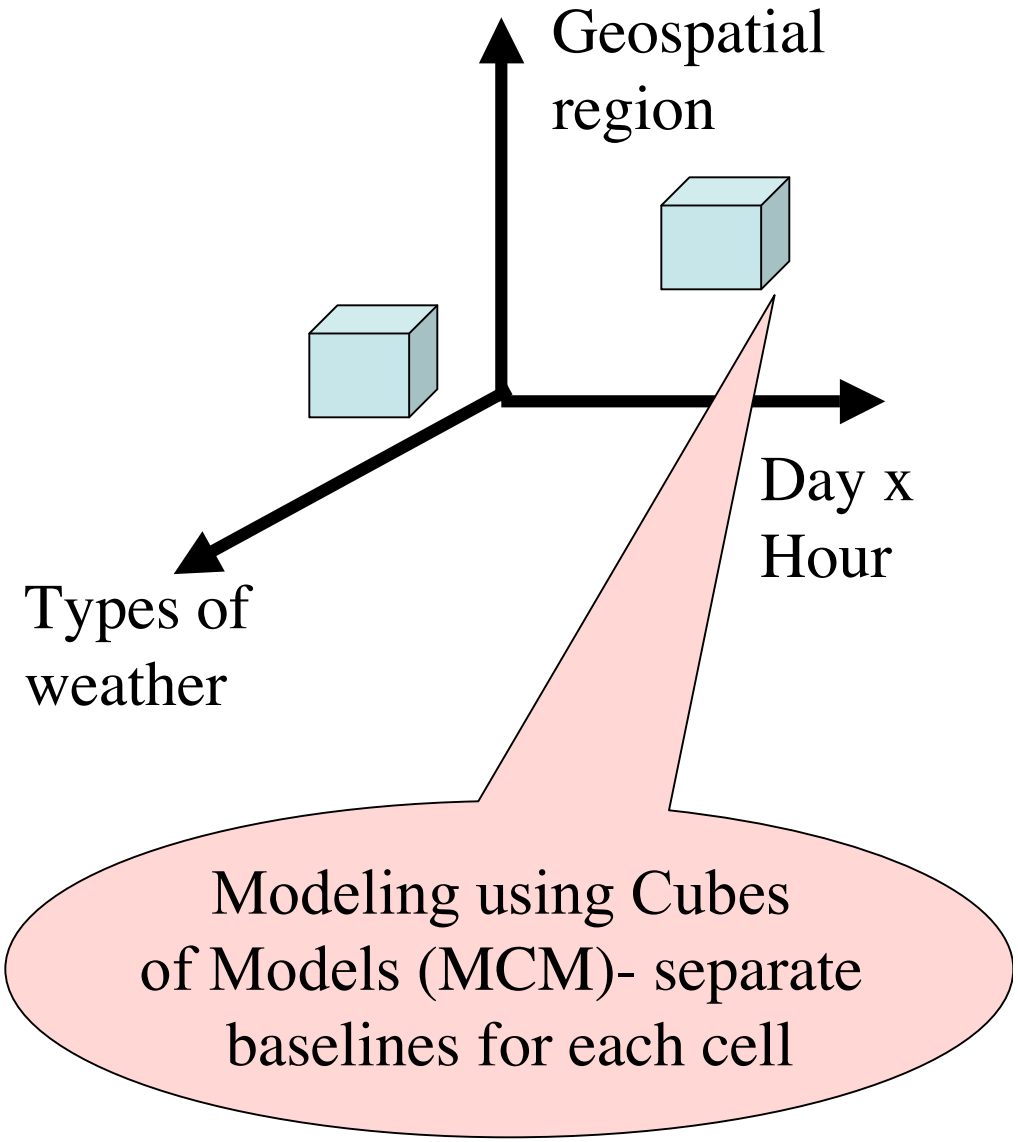
GLR Change Detection Algorithms (Single Model)



- Sequence of events x[1], x[2], x[3], …
- Question: is the observed distribution different than the baseline distribution?
- Use simple CUSUM & Generalized Likelihood Ratio (GLR) tests
- ... but use thousands of them

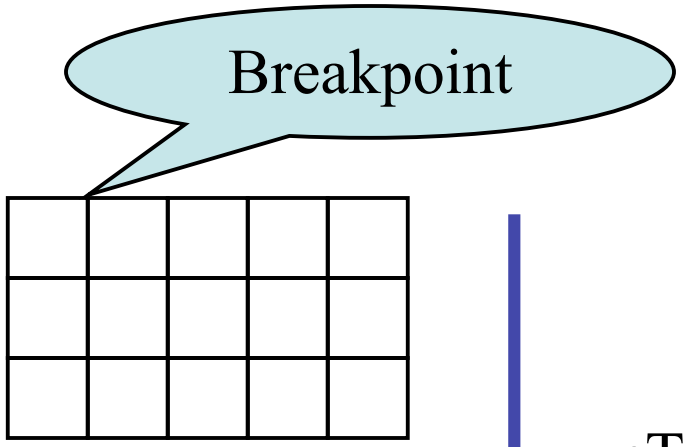
Build 10^4+ Models

1. Build segmented models using multidimensional data cubes
2. For each distinct cube, estimate parameters for separate statistical model
3. Detect changes from baselines and send alerts in real time

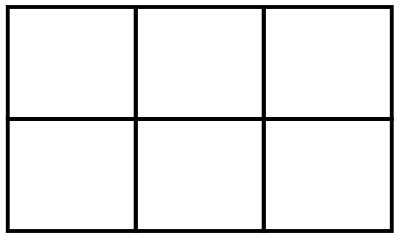


Greedy Meaningful/Manageable Balancing (GMMB) Algorithm

- More alerts
- Alerts more ***meaningful***
- To increase alerts, add breakpoint to split cubes, order by number of new alerts, & select one or more new breakpoints

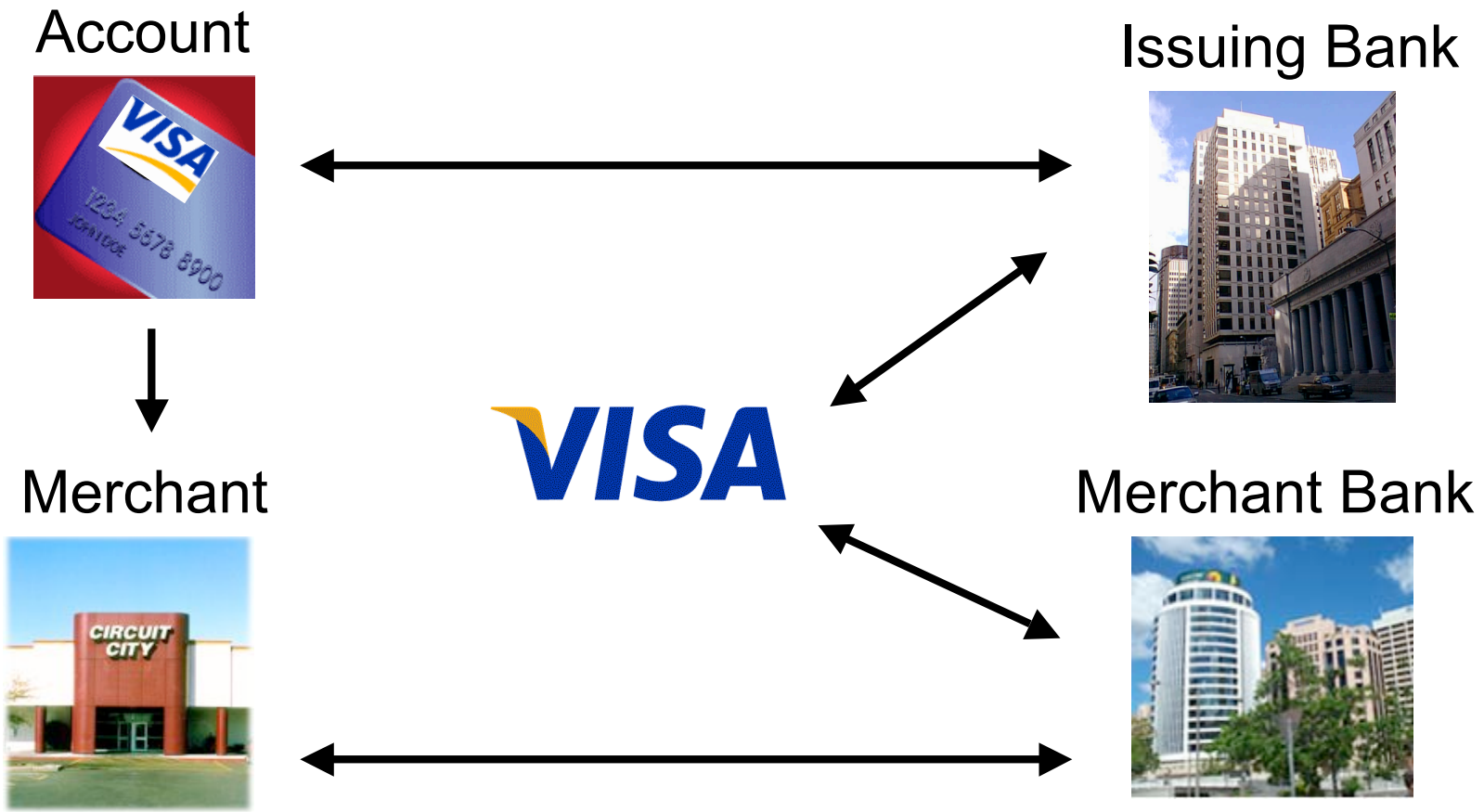


One model for each cell in data cube



- Fewer alerts
- Alerts more ***manageable***
- To decrease alerts, remove breakpoint, order by number of decreased alerts, & select one or more breakpoints to remove

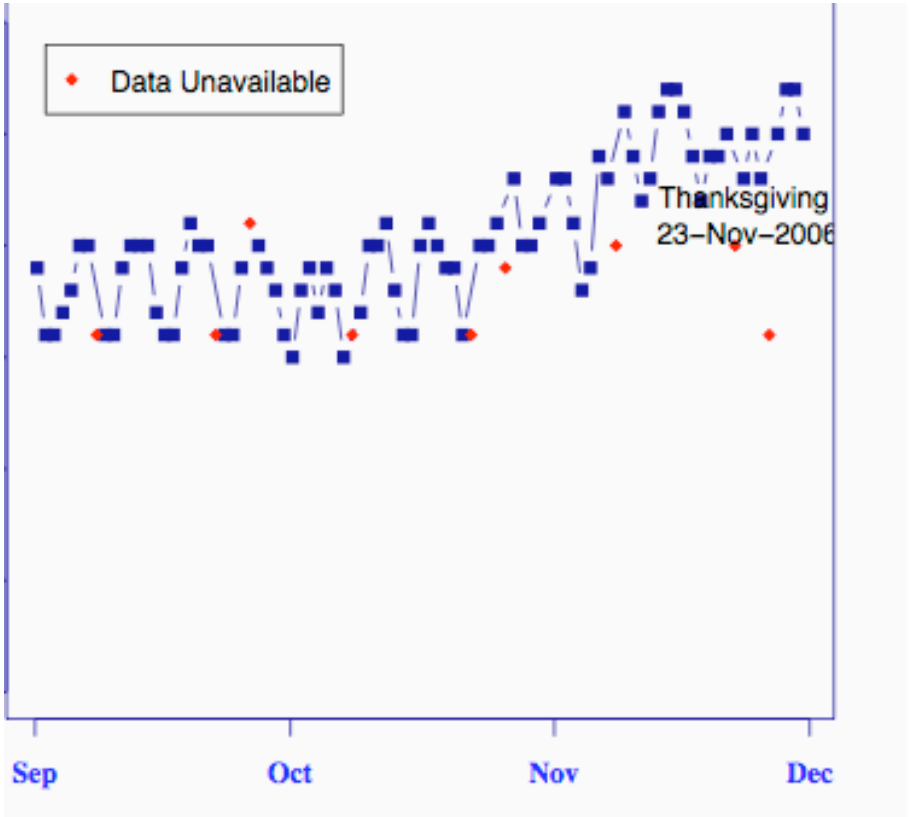
Example 2: Data Quality for Payment Systems



- 6000+ peak transactions per second.

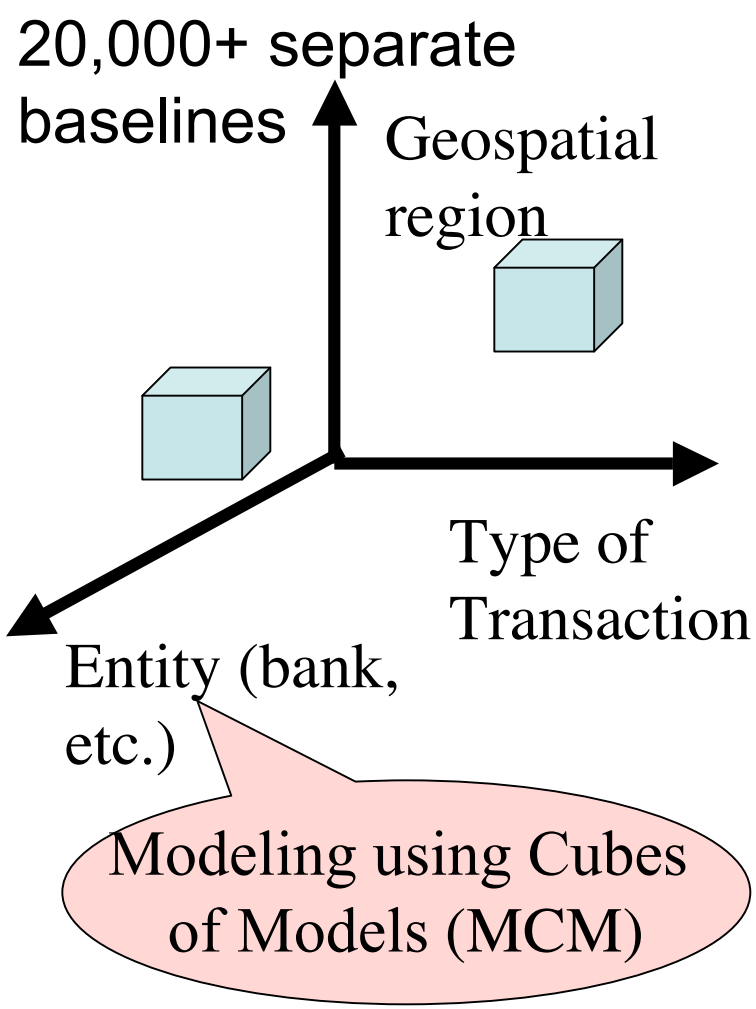
Payments Data is Highly Heterogeneous

- Variation merchant to merchant
- Variation bank to bank
- Daily variation
- Variation season to season

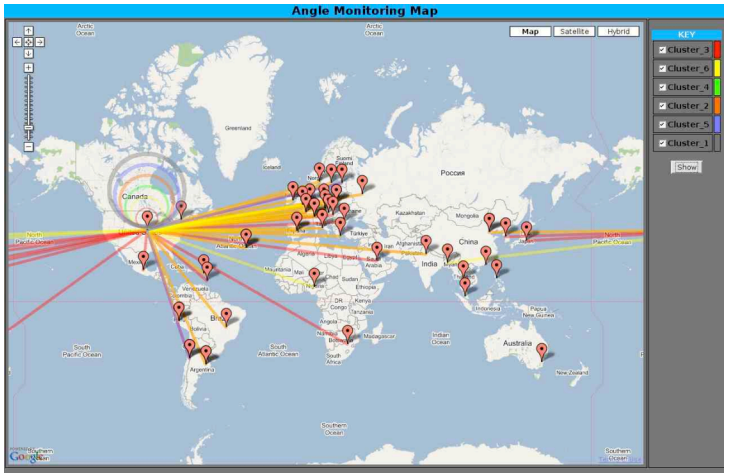


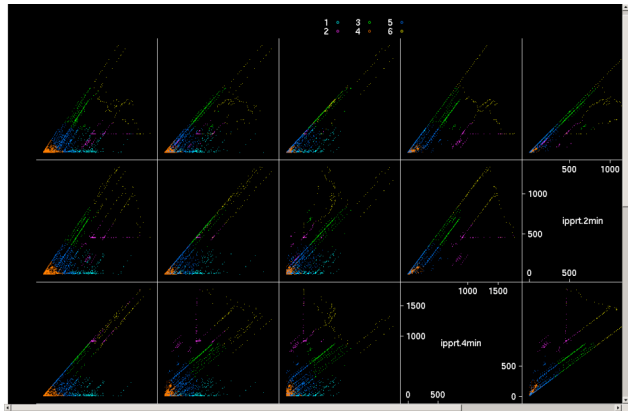
Data Cubes of Models - Payments Systems

- Build separate model for each bank (c. 1000)
- Build separate model for each geographical region (6 regions)
- Build separate model for each different type of merchant (c. 800 types of merchants)
- For each distinct cube, establish separate baselines for each metric of interest (declines, etc.)
- Detect changes from baselines



Example 3 - Emergent Behavior Network Packet Data





- Data collected in real time from several different distributed sensors (Angle)
- Still investigating best dimensions for cube
- Build separate *cluster* model for each cell in cube

Angle Scoring Functions for Each Cube in Data Cube of Models

- Update features using new packets and evolve features
- Divide clusters into good (B or Blue), neutral, and bad (R or Red)
- Blue - score using good clusters
- Red - score using bad clusters
- Purple - score using both good and bad clusters

- Hard scoring - use max / min

$$s(x) = \max_{k \in B} s_k(x)$$

- Soft scoring use sum

$$s(x) = \sum_{k \in B \cup R} s_k(x)$$

- Scoring function for single cluster

$$s_k(x) = \theta_k \frac{1}{\sigma_k} \exp\left(\frac{-\lambda \left\| x - \mu_k \right\|^2}{2\sigma_k^2} \right)$$

$$\sum_k \theta_k = 1$$

The Challenge

- This methodology can work quite well in practice.
- Develop some of the theory to guide this methodology and improve the methodology.

Other Applications

- George Church's challenge individual predictive models for each human genome 6.5 Billion humans x 6 Billion Base Pairs
- Consumer Marketing - large advertisers will see 1-3 Billion different consumers
- Network defense / cyberdefense - 4 billion IPv4 addresses; billions of users; billions+ of IPv6 addresses

What About the Data?

- Highway change detection data is available highway.ncdm.uic.edu
- Angle network anomalies will be available

What About the Software?

- Augustus - Will be available from Source Forge

References

- Robert L. Grossman, Michal Sabala, Javid Alimohideen, Anushka Aanand, John Chaves, John Dillenburg, Steve Eick, Jason Leigh, Peter Nelson, Mike Papka, Doug Rorem, Rick Stevens, Steve Vejcik, Leland Wilkinson, and Pei Zhang, Real Time Change Detection and Alerts from Highway Traffic Data, ACM/IEEE International Conference for High Performance Computing and Communications (SC '05).
- Joseph Bugajski, Robert L. Grossman, Eric Sumner and Steve Vejcik, Monitoring Data Quality for Very High Volume Transaction Systems, Proceedings of the 11th International Conference on Information Quality, 2006.
- Joseph Bugajski, Chris Curry, Robert L. Grossman, David Locke, Steve Vejcik, Detecting Changes in Large Data Sets of Payment Card Data: A Case Study, Proceedings of The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2007.