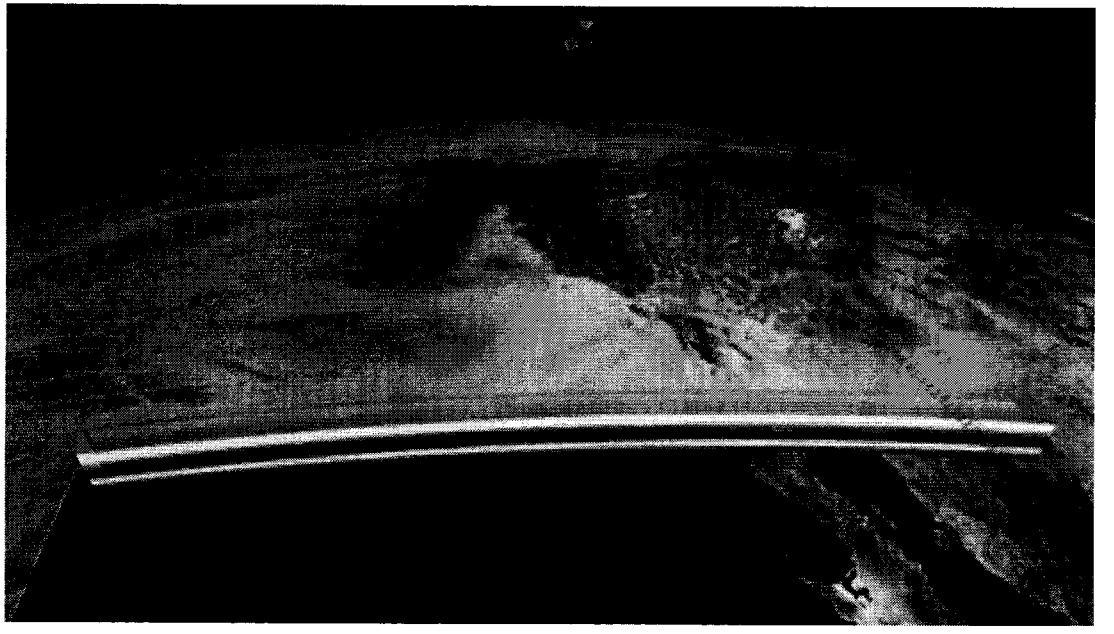
Visual Data Mining for Quantized, Spatial Data



Amy Braverman
Jet Propulsion Laboratory
California Institute of Technology
Mail Stop 169-237
4800 Oak Grove Drive
Pasadena, CA 91109-8099
Amy.Braverman@jpl.nasa.gov

Outline

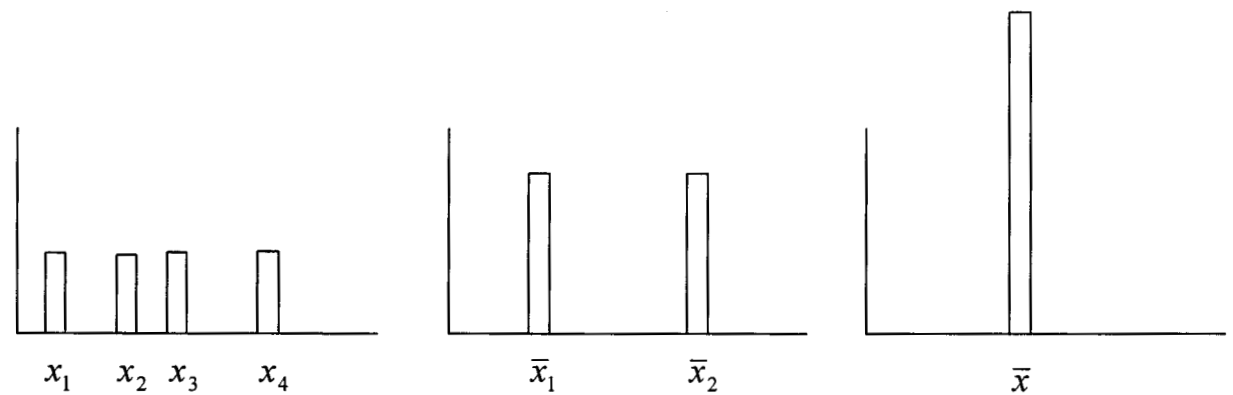
1. Motivation.
2. Approach.
3. AIRS data collection.
4. Quantization..
5. Visual data mining (I).
6. Visual data mining (II).
7. Hierarchical Quantization.
8. Visual data mining (III).
9. Summary.

Motivation

- Earth Observing System satellites return "massive" data volume.
- Data to be studied by wide range of researchers at universities and elsewhere. Large segment of intended users can't handle the volume.
- Need to produce low-volume summary data sets to be used in place of original data for analysis.
- Traditional approach: produce maps of one degree averages and standard deviations for each parameter of interest.
- Good news: this is easy, practical, and everybody understands it.
- Bad news: the method throws away almost all of the distributional information in the data including covariance and higher-order statistics.
- **Need: to "mine" the data, i.e. how do characteristics of joint distributions change in time and space, and across resolutions? Characterize forcings and feedbacks.**

Approach

- New approach: produce an estimate of the joint (empirical) probability distribution of variables of interest within each one degree grid cell.
- Make it as "small" as possible without sacrificing "too much" accuracy.



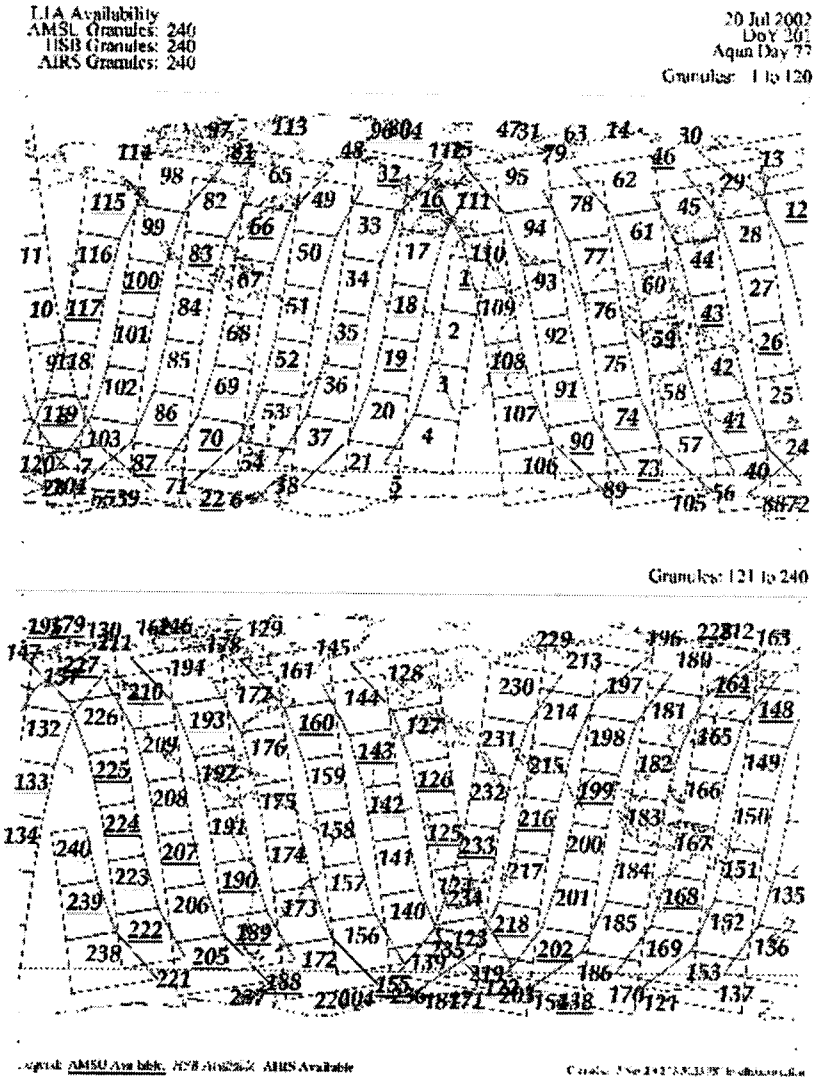
No compression, perfect fidelity.

Intermediate compression intermediate fidelity.

Full compression, minimum fidelity

- Collection of all 180 x 360 = 64,800 grid cell distribution estimates is a proxy for the original data.

AIRS Data Collection



- AIRS on EOS-Aqua in polar orbit.
- Measures radiance in 2378 spectral channels at 15 km resolution.
- Derived geophysical products include column profiles of water vapor and temperature, total column ozone, methane and carbon dioxide, and cloud properties.
- Complete global coverage every 3 days.

Quantization

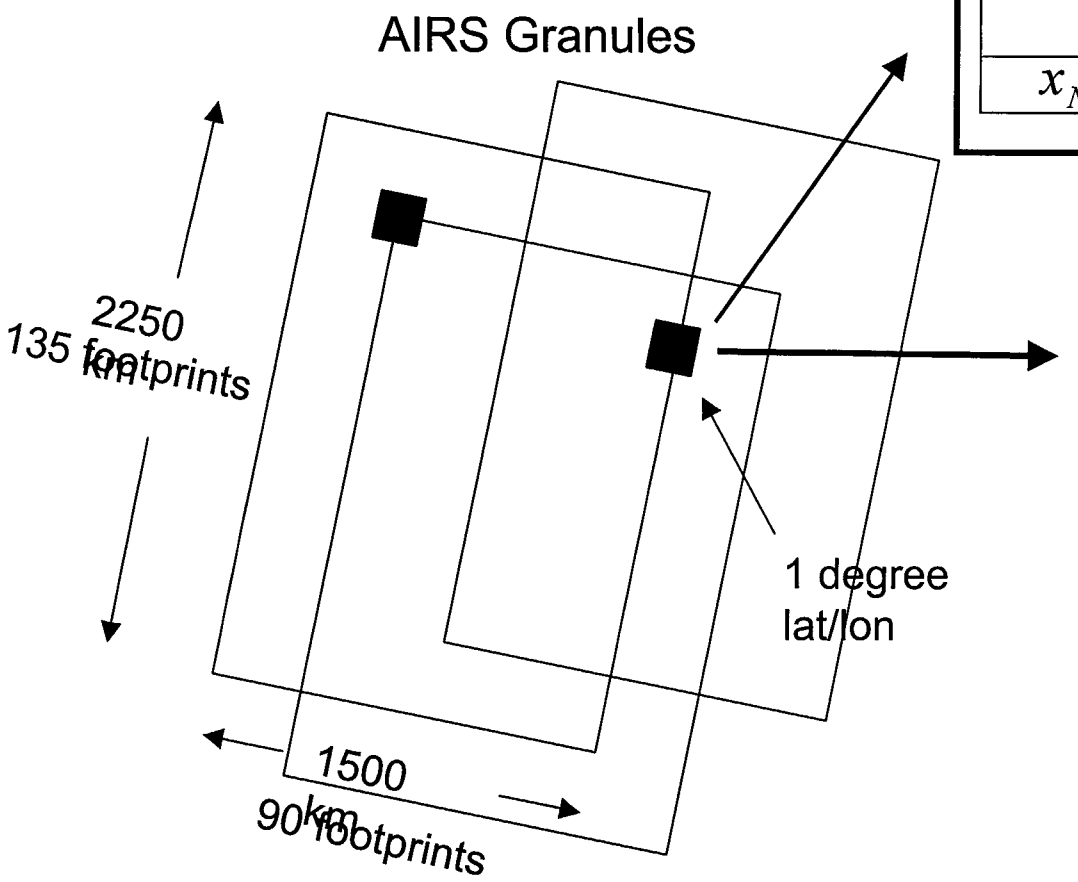
x_1	1
x_2	1
x_N	1

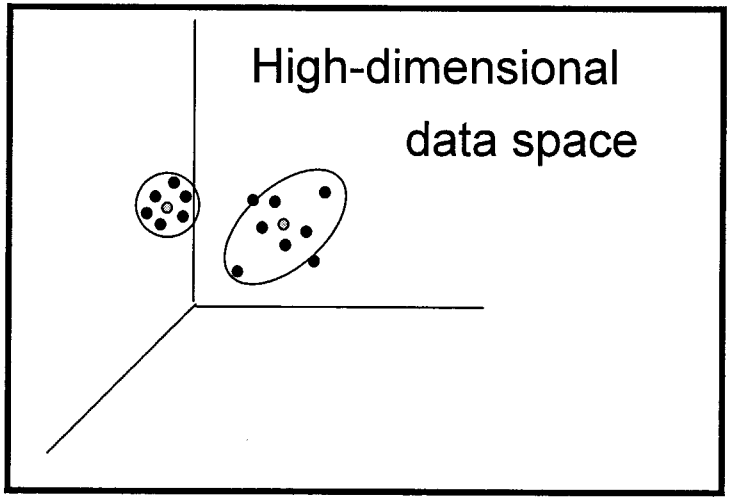
y_1	N_1
y_2	N_2
y_K	N_K

$$y_k = \frac{1}{N_k}\sum_{n=1}^{N} x_n \mathbb{1}[x \in k]$$

$$N_k = \sum_{n=1}^{N} \mathbb{1}[x \in k]$$

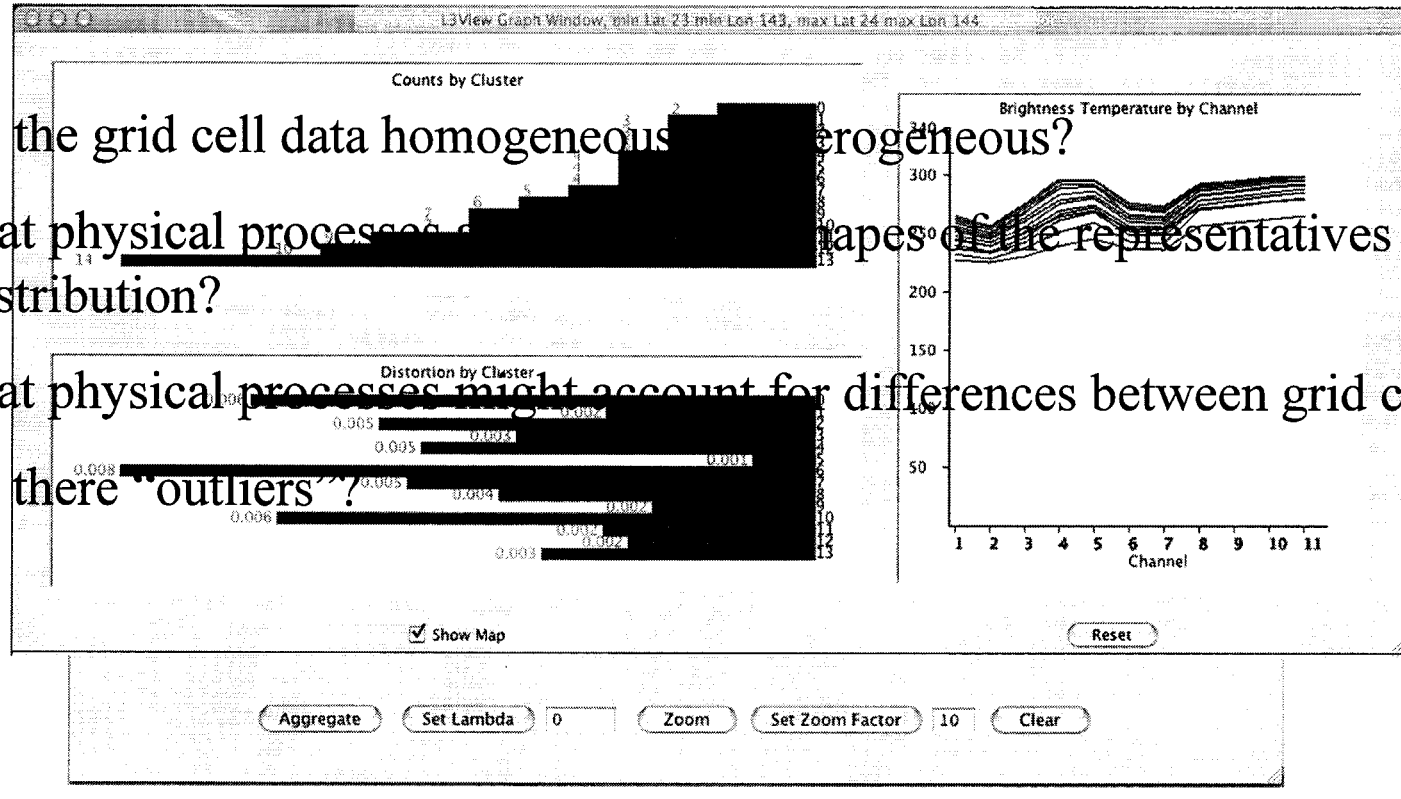
$$X \longrightarrow Y = E(X \mid Y) \quad \textbf{(!)}$$





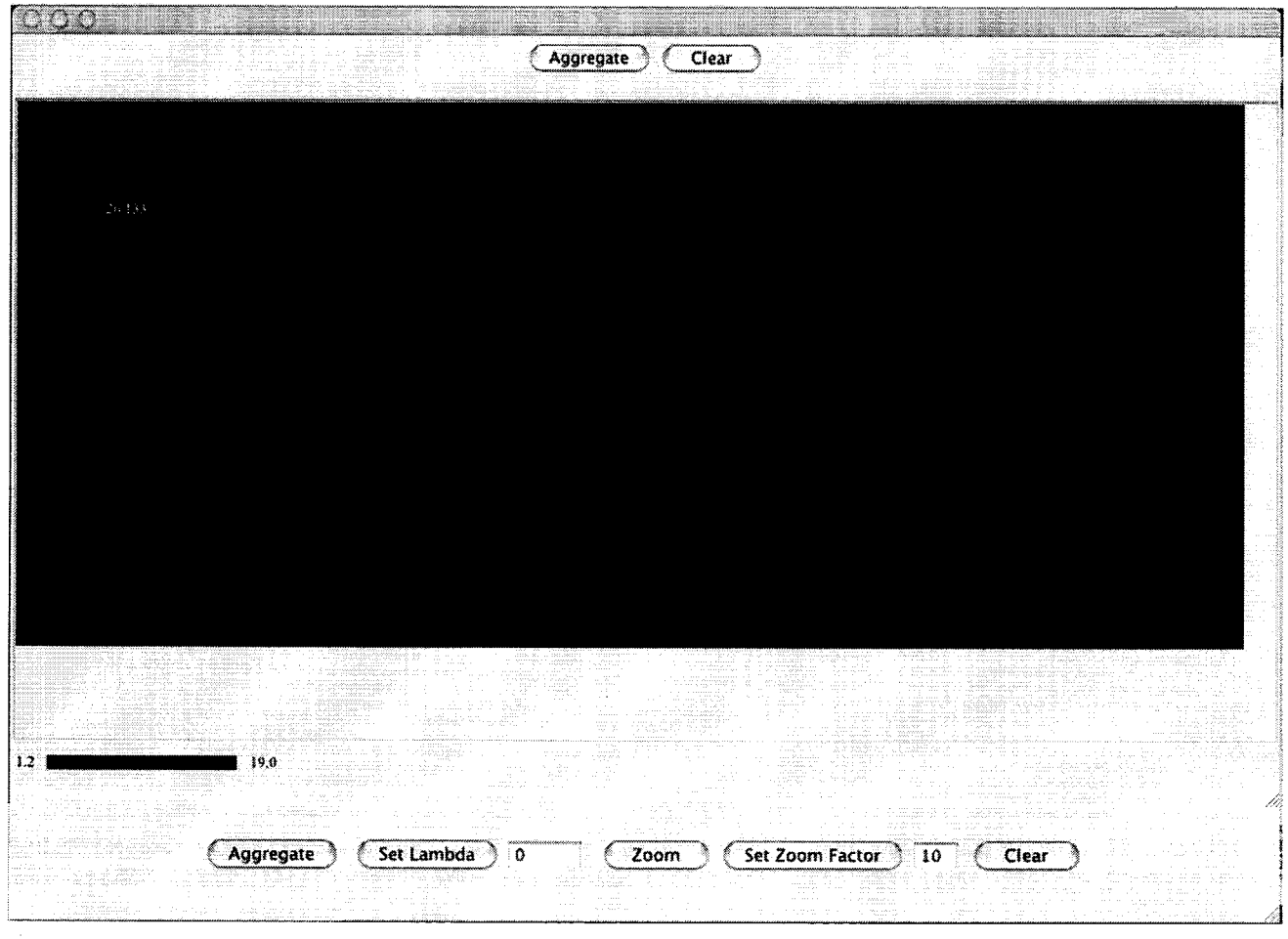
Visual Data Mining (I)

- Data: 11 AIRS channels observed over 3 days (July 20-22, 2002).
- Compare joint distributions among grid cells:



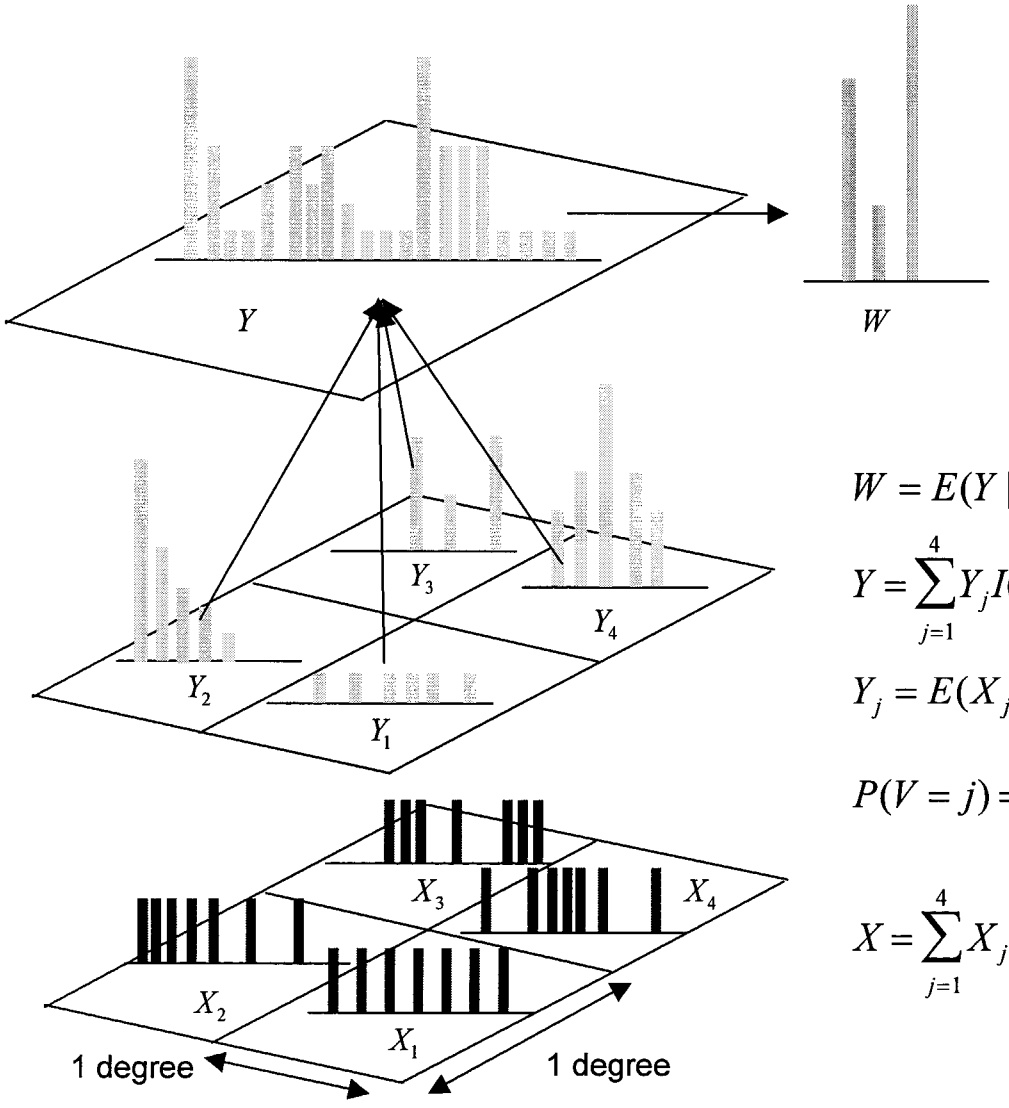
- Are the grid cell data homogeneous [redacted] erogeneous?
- What physical proces[redacted]apes of the representatives and the distribution?
- What physical [redacted] differences between grid cells?
- Are there "outliers"?

Visual Data Mining (II)



- Data in this region: 10,498 clusters representing 60,681 observations.
- Can we summarize the whole region as one?

Hierarchical Quantization



$$\delta(X_j, Y_j) = E\left\|X_j - Y_j\right\|^2$$

$$\delta(X, Y) = E\left\|X - Y\right\|^2$$

$$\delta(Y, W) = E\left\|Y - W\right\|^2$$

$$\boxed{\delta(X, W) = \delta(Y, W) + \sum_{j=1}^{4} \delta(X_j, Y_j) P(V = j)}$$

$$W = E(Y \mid W)$$

$$Y = \sum_{j=1}^{4} Y_j I(V = j)$$

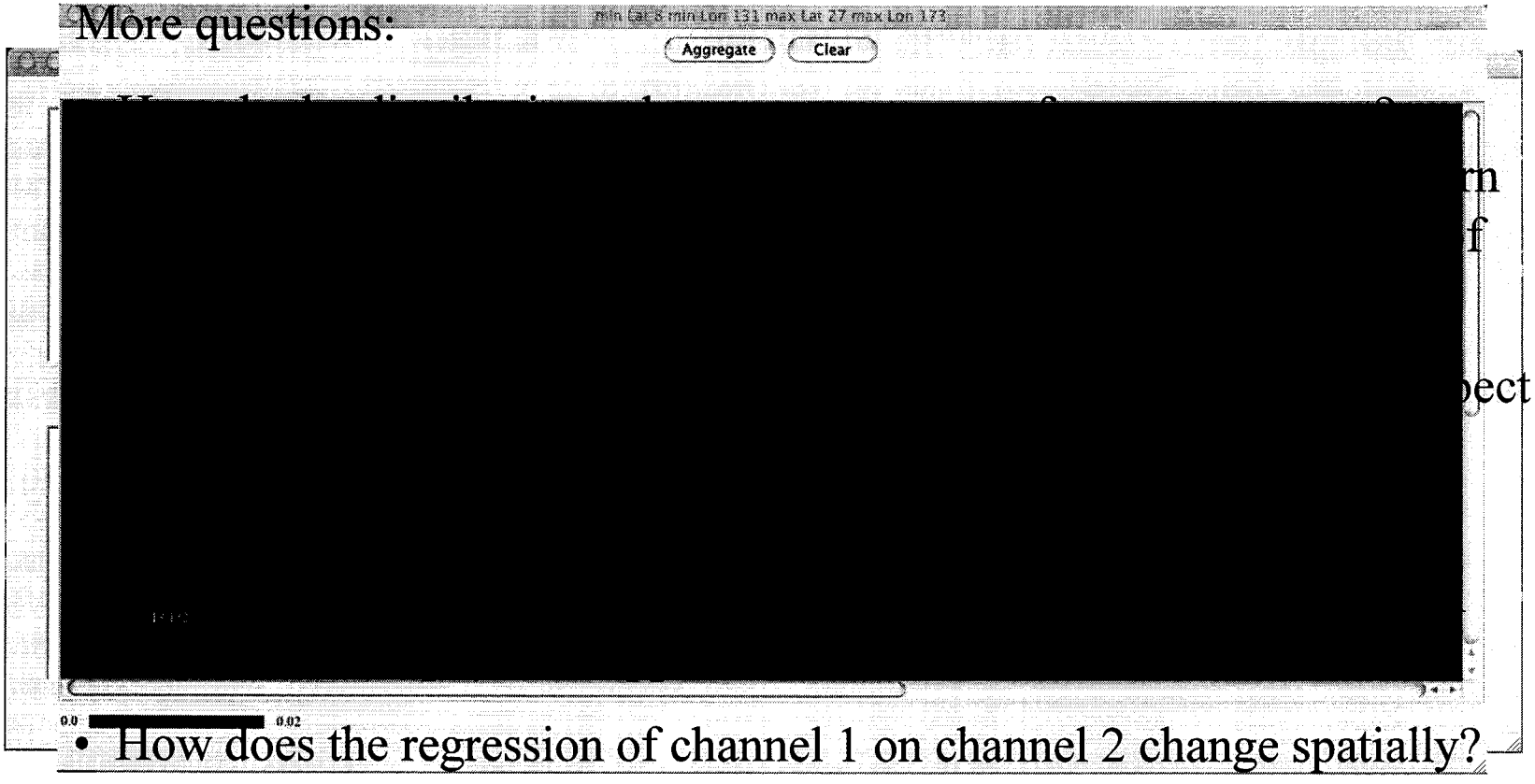
$$Y_j = E(X_j \mid Y_j)$$

$$P(V = j) = \frac{N_j}{\sum N_j}$$

$$X = \sum_{j=1}^{4} X_j I(V = j)$$

Visual Data Mining (III)

More questions:



- How does the regression of channel 1 on channel 2 change spatially?

- From where do the clusters come?

Summary

- Accept coarser spatial resolution (one degree) to achieve replication and estimate distributions.
- Explore quantized data interactively by comparing distributions at different levels of aggregation and in different places.
- We are mining the data, not making inferences. No spatial statistical models.
- Live demo on request.
- AIRS data are free and available at http://daac.gsfc.nasa.gov/atmodyn/airs/index.html.
- More information about AIRS: http://www-airs.jpl.nasa.gov.