

MassDAL

Massive Data Analysis Lab

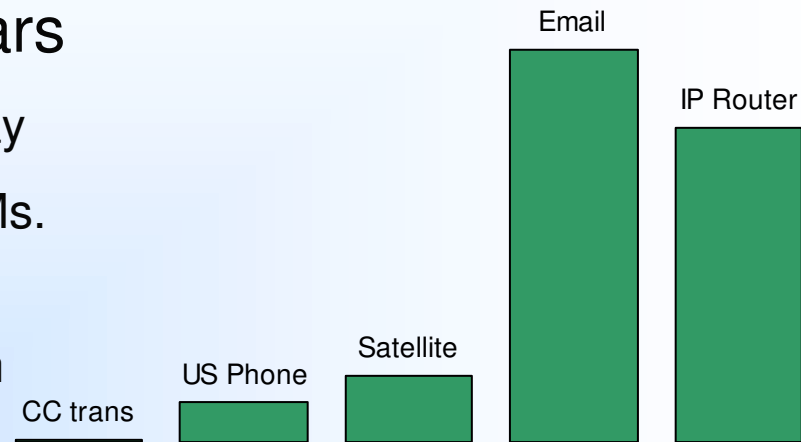
S. Muthu Muthukrishnan
Dept of Computer Science

<http://www.cs.rutgers.edu/~muthu/massdal.html>

Massive Scale of Data

Explosion of Data In Recent Years

- 3 Billion Telephone Calls in US each day
- 30 Billion emails daily, 1 Billion SMS, IMs.
- **Scientific data:** NASA's observation satellites generate billions of readings each per day.
- **IP Network Traffic:** up to 1 Billion packets per hour per router. Each ISP has many (hundreds) of routers!
- **Compare to "human scale" data:** "only" 1 billion worldwide credit card transactions per month.



New data scales demand new approaches from databases, algorithms, networks, systems and engineering.

Muthu Muthukrishnan, Rutgers Univ.

Agenda

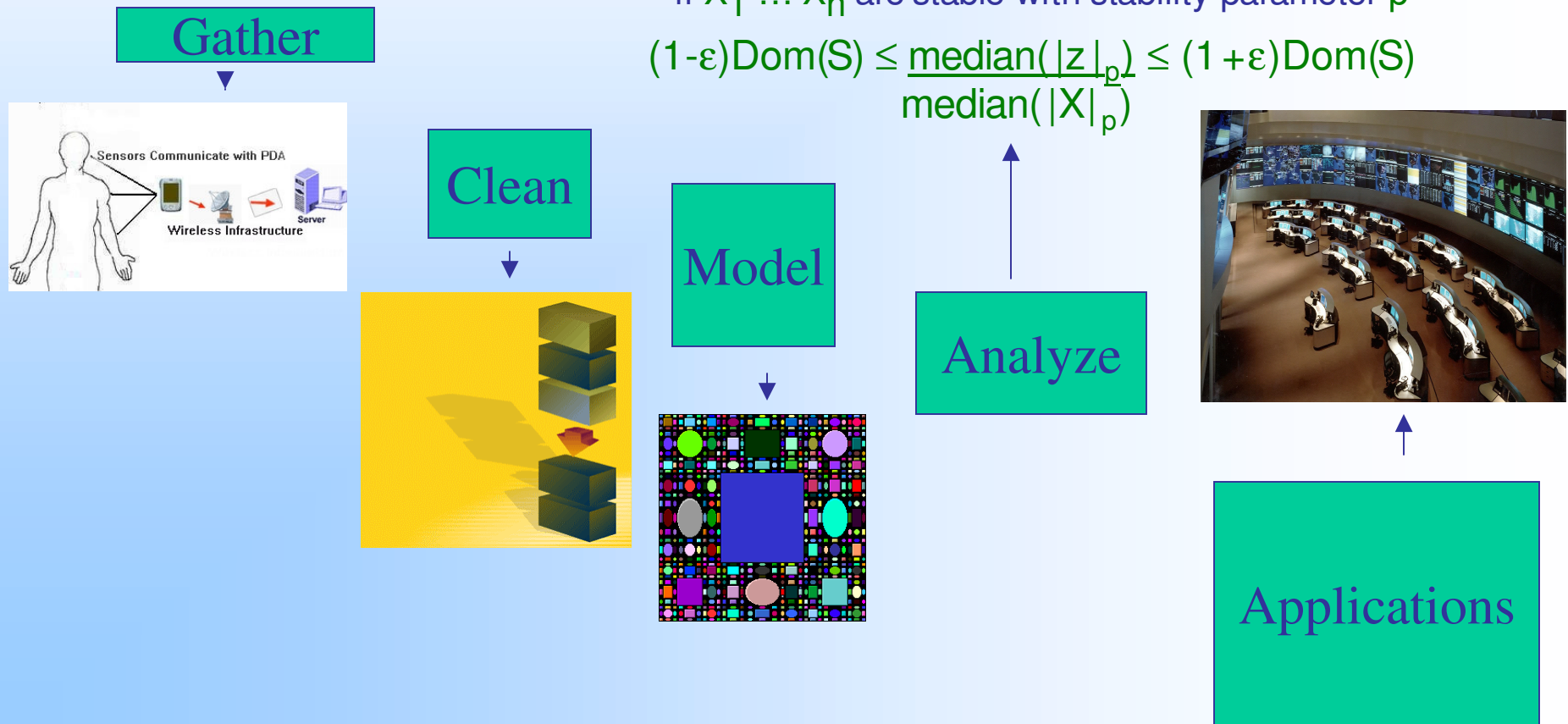
- **MassDAL** researches the entire lifecycle of “massive data”.

Stable distributions have property that

$$a_1X_1 + a_2X_2 + \dots + a_nX_n = \|(a_1, a_2, \dots, a_n)\|_p X$$

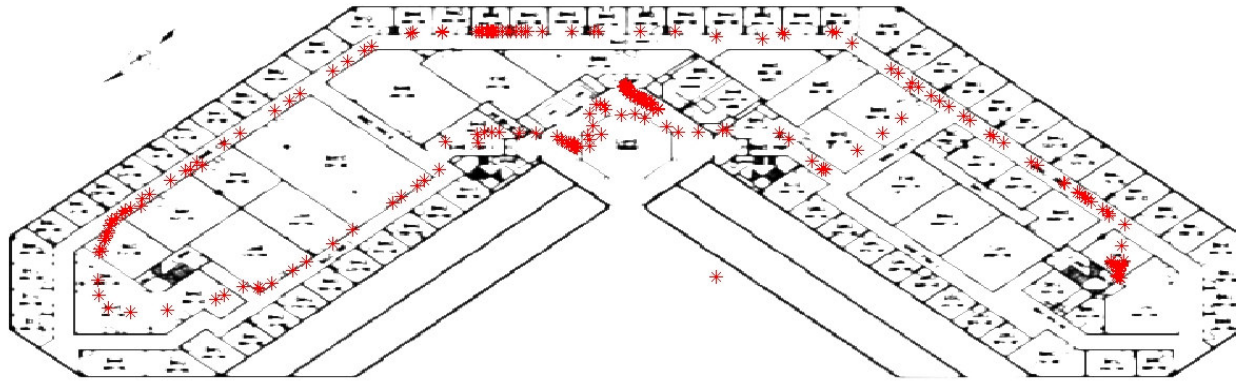
if $X_1 \dots X_n$ are stable with stability parameter p

$$(1-\varepsilon)\text{Dom}(S) \leq \frac{\text{median}(|z|_p)}{\text{median}(|X|_p)} \leq (1+\varepsilon)\text{Dom}(S)$$



Gathering Data: Persistent Health Monitoring

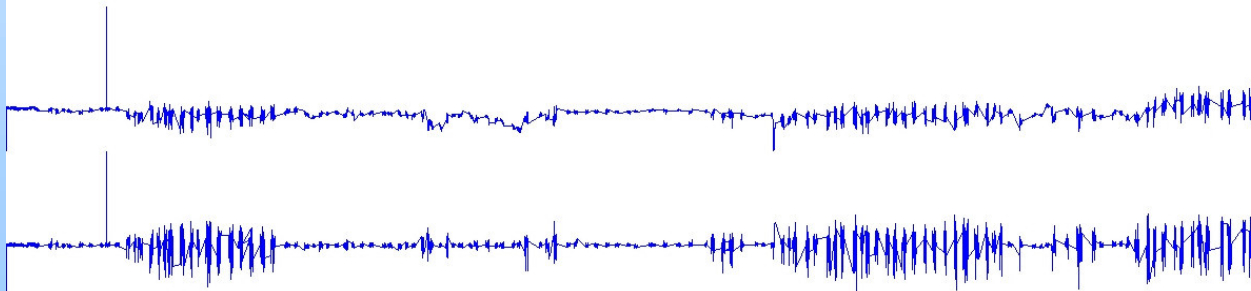
MassDAL
Massive Data Analysis Lab



Project:
Continuously
gather Location,
Audio, EMG, ECG
signals from each
person

Sample appl here:
How to use these
data to better
localize users?

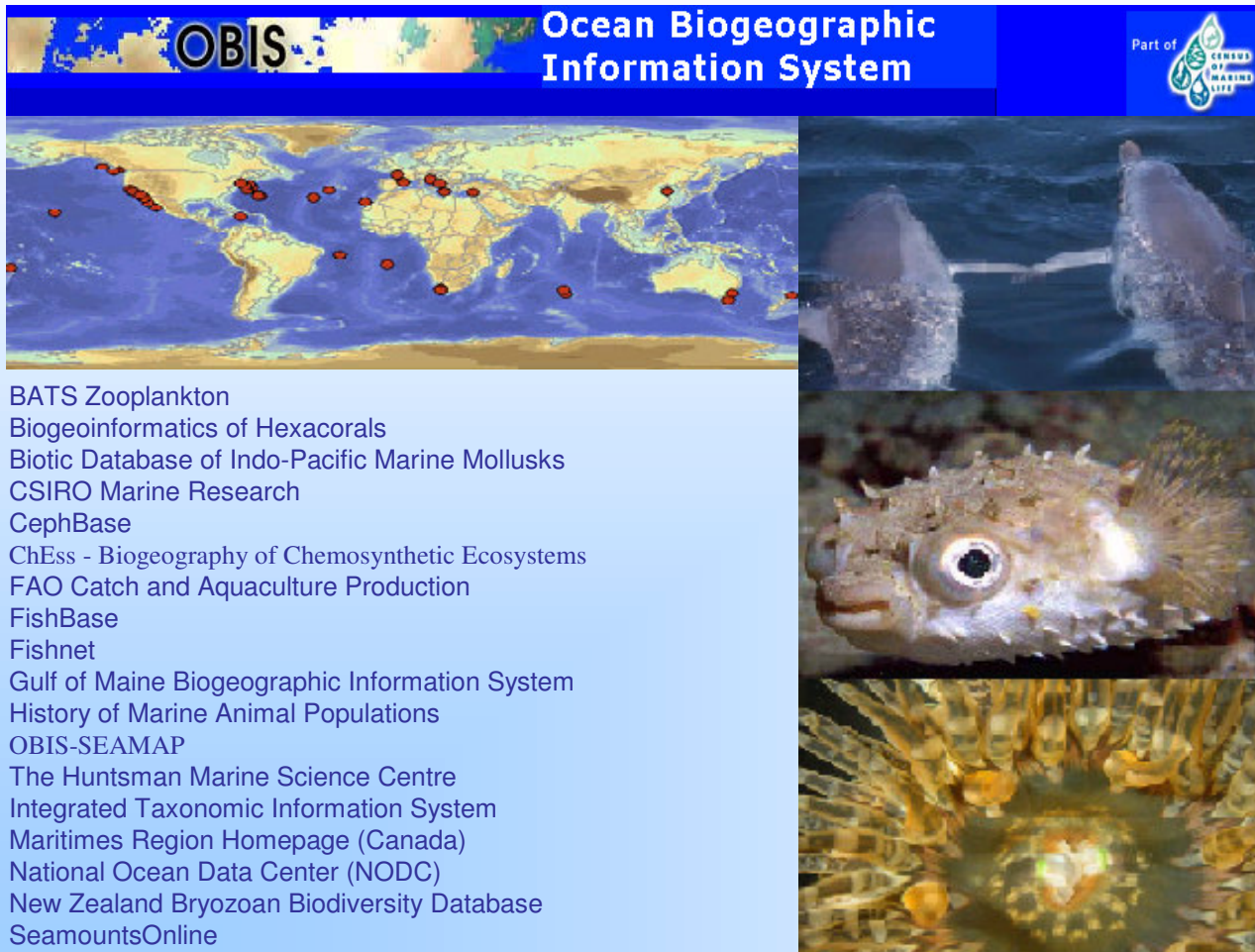
Other appl:
persistent health
monitoring.



Su Chen & Amit Gaur (Rutgers);
David Rosenbluth (Telcordia)

Data Cleaning: Federated Marine Data

MassDAL
Massive Data Analysis Lab



BATS Zooplankton
Biogeoinformatics of Hexacorals
Biotic Database of Indo-Pacific Marine Mollusks
CSIRO Marine Research
CephBase
ChEss - Biogeography of Chemosynthetic Ecosystems
FAO Catch and Aquaculture Production
FishBase
Fishnet
Gulf of Maine Biogeographic Information System
History of Marine Animal Populations
OBIS-SEAMAP
The Huntsman Marine Science Centre
Integrated Taxonomic Information System
Maritimes Region Homepage (Canada)
National Ocean Data Center (NODC)
New Zealand Bryozoan Biodiversity Database
SeamountsOnline
Species 2000
ZooGene

Phoebe Y. Zhang (Rutgers, IMCS)
Wei Zhuang (Rutgers, Comp. Sci.)

Question: How to clean the data in federated stores such as the OBIS, and ensure its data integrity?

Solution:
Apply our Probabilistic Approximate Constraints approach.

Data Modeling: Multi-Fractal Nature of IP Traffic

Observation:
IP traffic →
has multi-fractal
distribution.

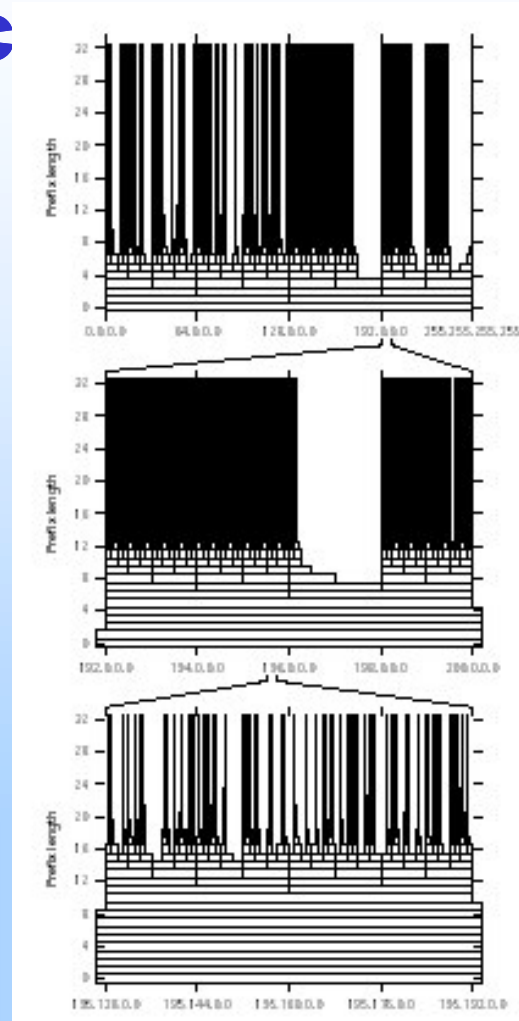
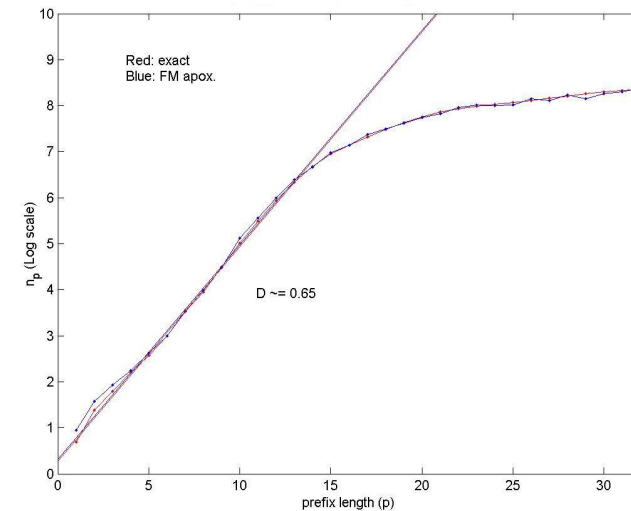


Image from Kohler et al. 2002

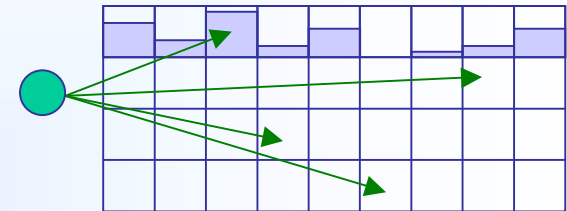
Result: We can estimate and learn the parameters of the multi-fractal very accurately at the line speed in IP networks, with very small space!



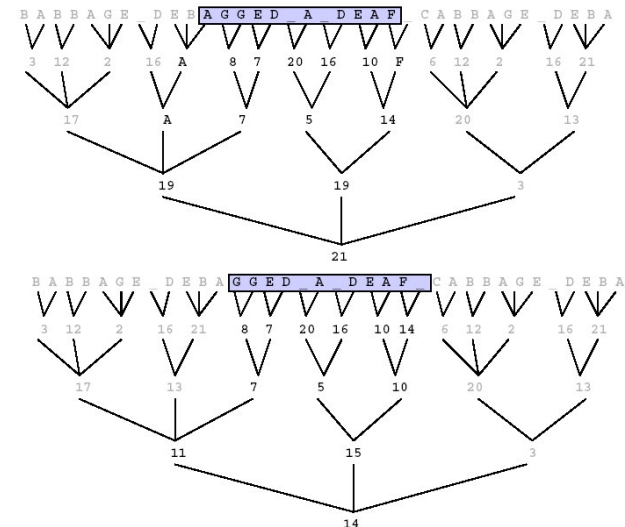
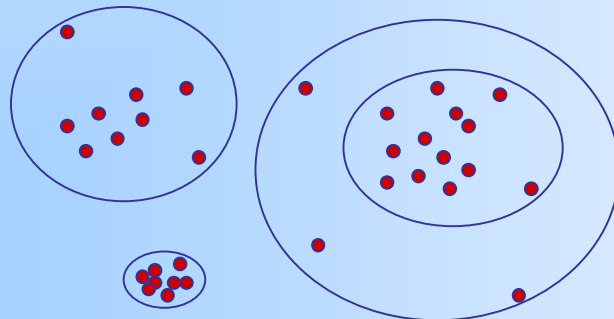
Suhrid Balakrishnan (Rutgers); Flip Korn (AT&T Research)

Data Analysis: Streaming Algorithms

- What's new, what's hot, what's next?
 - Detect trends in massive streams of data
 - Frequently occurring items, and rare ones
 - What is different between yesterday and today?
What is expected for tomorrow?

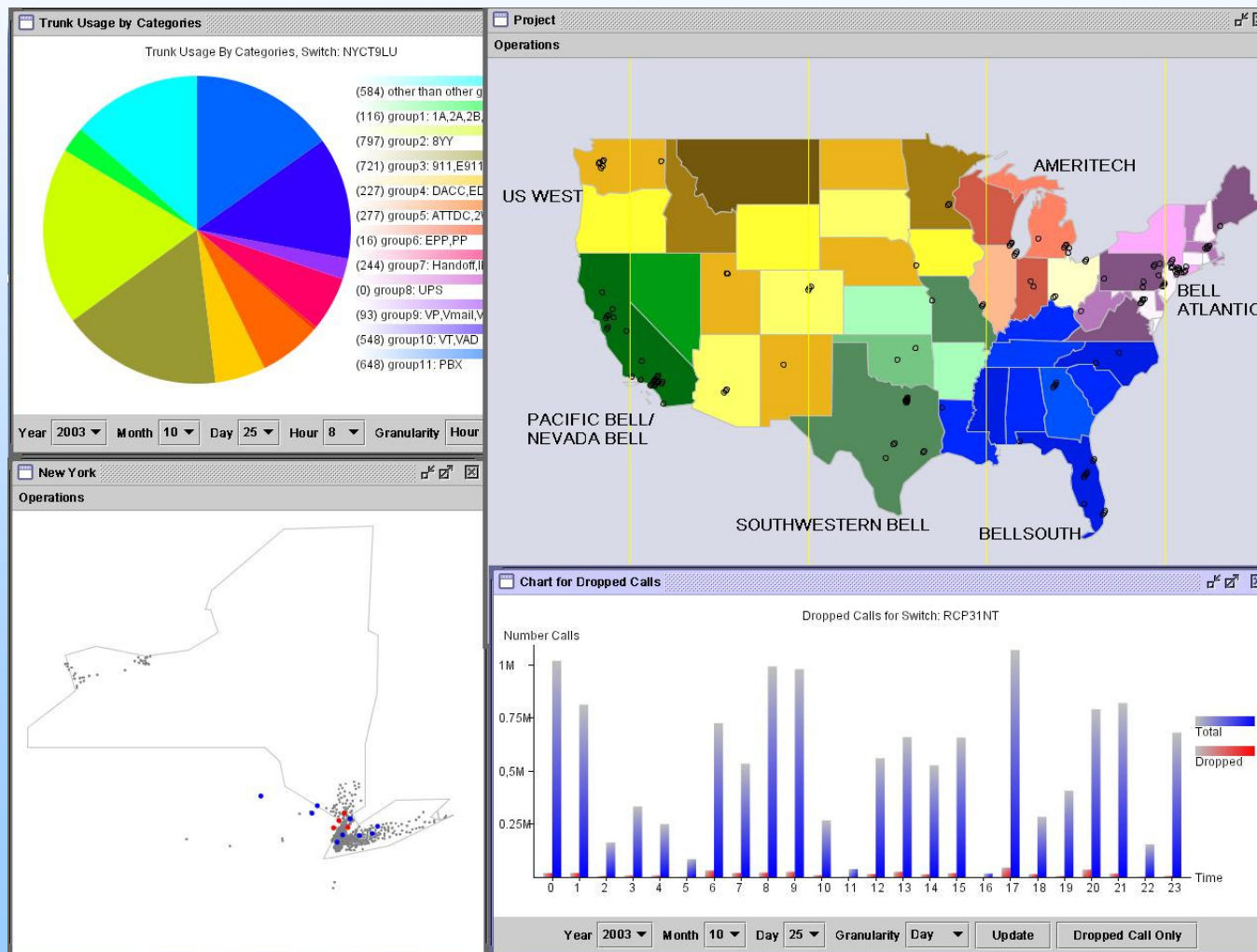


- Application: Burst analysis on text streams.
 - Find terms that are suddenly important
 - Keyword tracking, topic clustering
 - Finding groups and multiple identities



Graham Cormode, DIMACS

Applications: Visualize Geospatial Patterns



Visualize
telecom
traffic data
as it
evolves,
finding
geographic
patterns.

Qi Yan,
Rutgers.

Lifecycle of Data



- **MassDAL** has methods to manage massive streams during the **entire lifecycle of data**: collect, clean, analyze and integrate into applications.
- **Applications**: Homeland security, Persistent health monitoring, Telecom traffic monitoring, Federated scientific databases, Social networking and Epidemiology.
- Working with academic and industrial partners

