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.
Mass DAL researches the entire lifecycle of “massive data”.

Stable distributions have property that

\[ a_1X_1 + a_2X_2 + \ldots + a_nX_n = \| (a_1, a_2, \ldots, a_n) \|_p X \]

if \( X_1, \ldots, X_n \) are stable with stability parameter \( p \)

\[(1-\epsilon)\text{Dom}(S) \leq \text{median}(|z|_p) \leq (1+\epsilon)\text{Dom}(S) \]

\[
\text{median}(|X|_p)
\]
Gathering Data: Persistent Health Monitoring

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)
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.

Phoebe Y. Zhang (Rutgers, IMCS)
Wei Zhuang (Rutgers, Comp. Sci.)
Data Modeling:
Multi-Fractal Nature of IP Traffic

Observation: IP traffic has multi-fractal distribution.

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!

Image from Kohler et al. 2002

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