Streaming Algorithms for Cosmological Simulations and Beyond

Vladimir Braverman
Johns Hopkins University

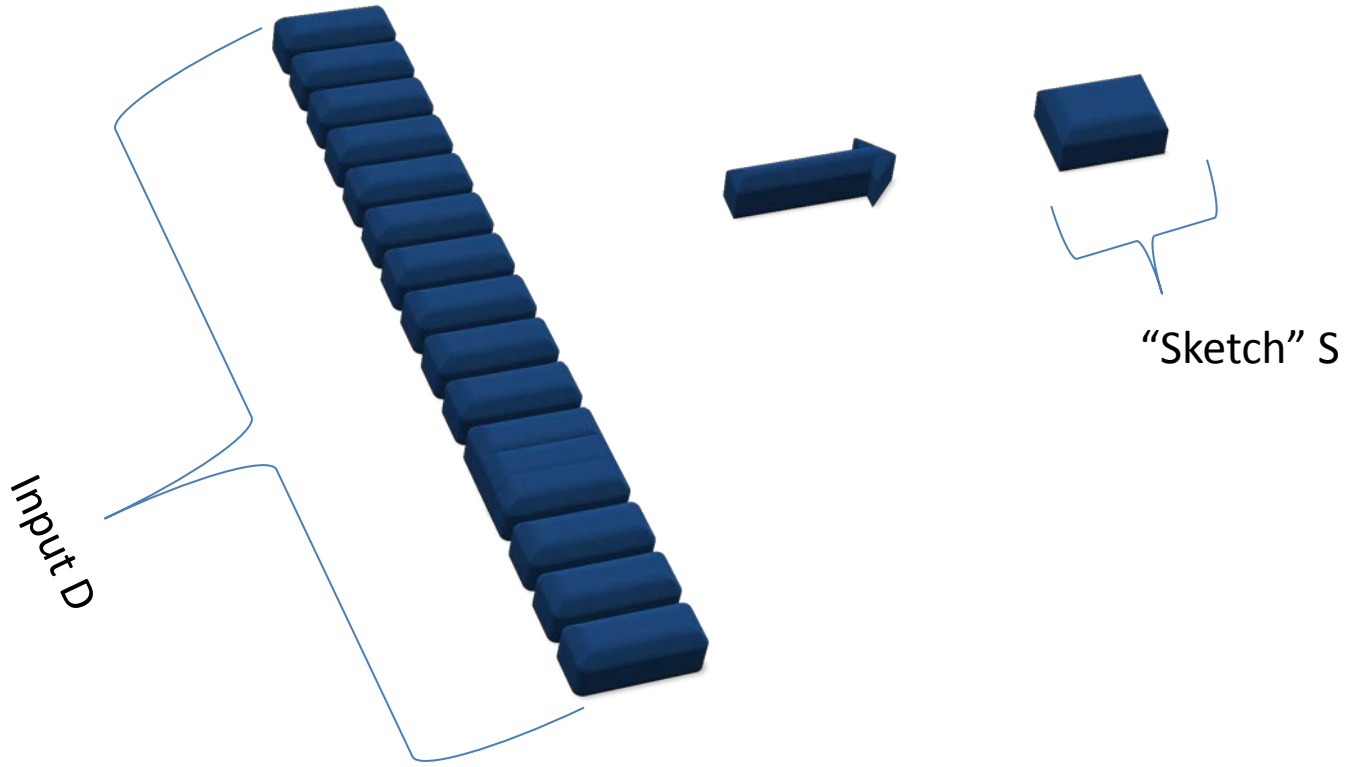
supported in part by the
National Science Foundation under Grant No. 1447639, by the Google Faculty Award and by DARPA grant N660001-1-2-4014.

Joint works with

Alexander S. Szalay, Zaoxing Liu, Greg Vorsanger, Vyas Sekar, Nikita Ivkin , Lin F. Yang , Mark Neyrinck , Gerard Lemson, Tamas Budavari, Randal Burns, Xin Wang, Stephen Chestnut, Harry Lang, Keith Levin…

Google

What are Streaming Algorithms?

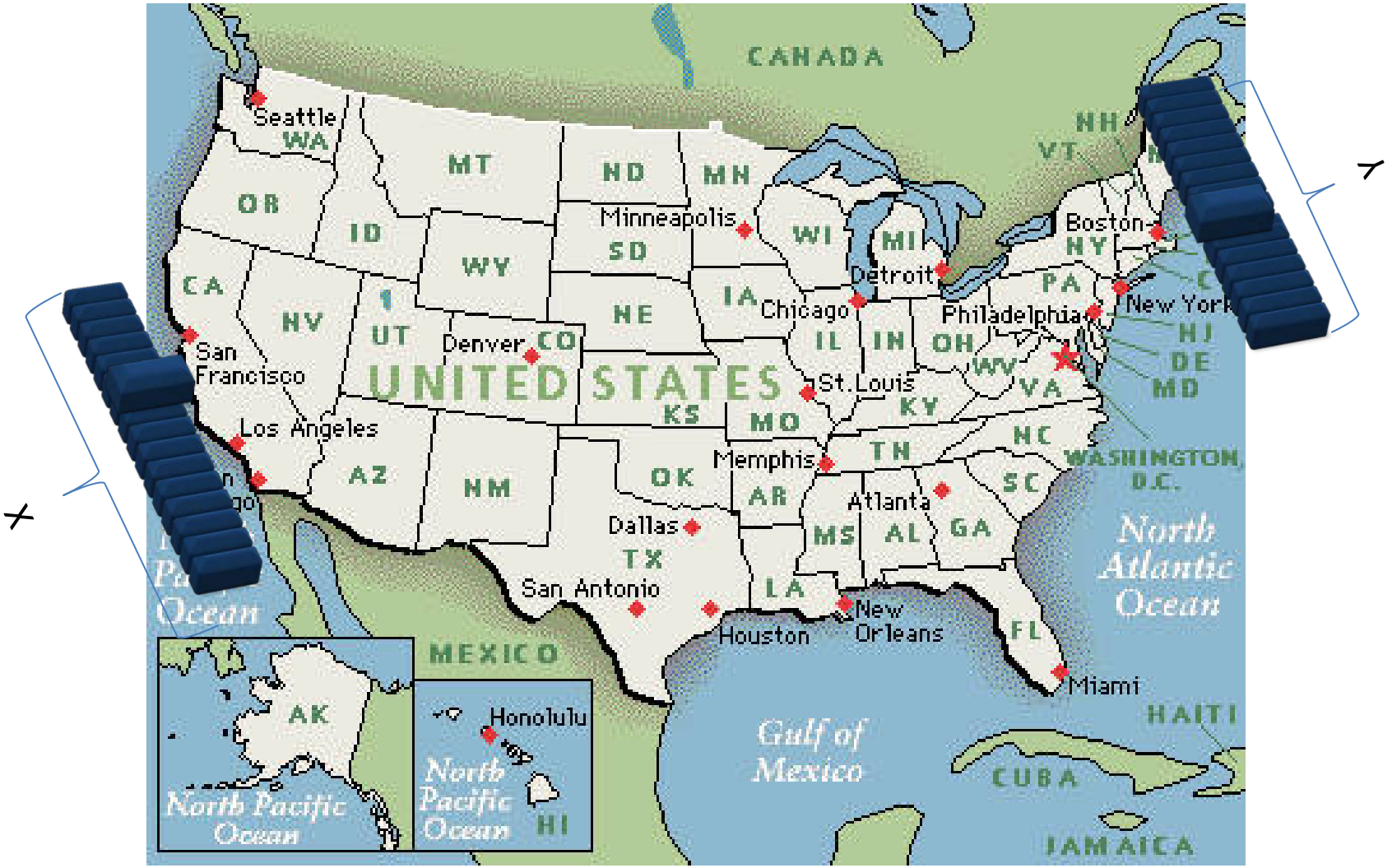


Goal: Compute F(D)

F(D) ≈ F'(S)

How does it work?

Is X equal to Y ?



Streaming Sketch

- If $h_1, \ldots, h_n$ are *i.i.d.,* $\boldsymbol{h_i} \sim U(\{-1,1\})$

Compare the inner products:

$\sum_{i=1}^{n} x_i h_i$ $\sum_{i=1}^{n} y_i h_i$

New Theory

- The Johnson-Lindenstrauss Lemma and metric embedding
- Stable Distributions and Pseudorandom generators
- Dvoretzky Theorem (local theory of Banach spaces)

Algorithms for :

- Clustering
- Sliding Windows
- Correlations
- Trends
- Frequent Events
- …

Streaming Algorithms for Halo Finders

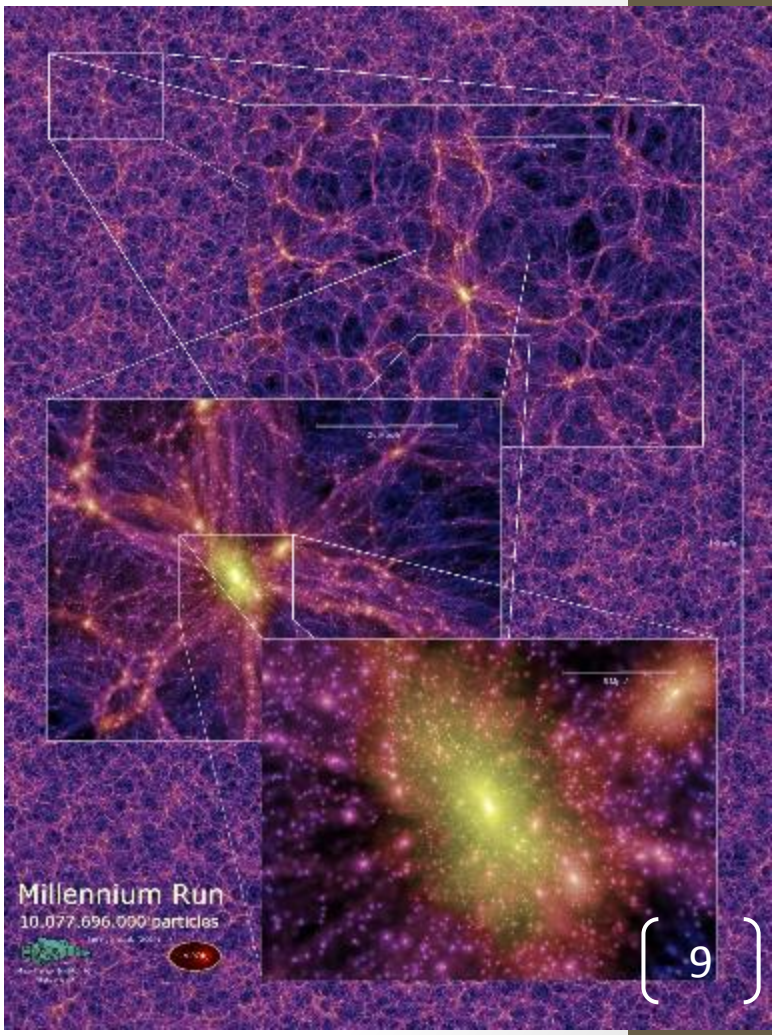
Zaoxing Liu , Nikita Ivkin , Lin F. Yang , Mark Neyrinck ,
Gerard Lemson, Alexander S. Szalay, Vladimir Braverman,
Tamas Budavari, Randal Burns, Xin Wang

Johns Hopkins University, Baltimore, MD, USA

Cosmological Simulations

Simulation:

- is a gravitational evolution of the system of particles
- provides distribution of particles in space and time
- helps to understand the processes of forming galaxies



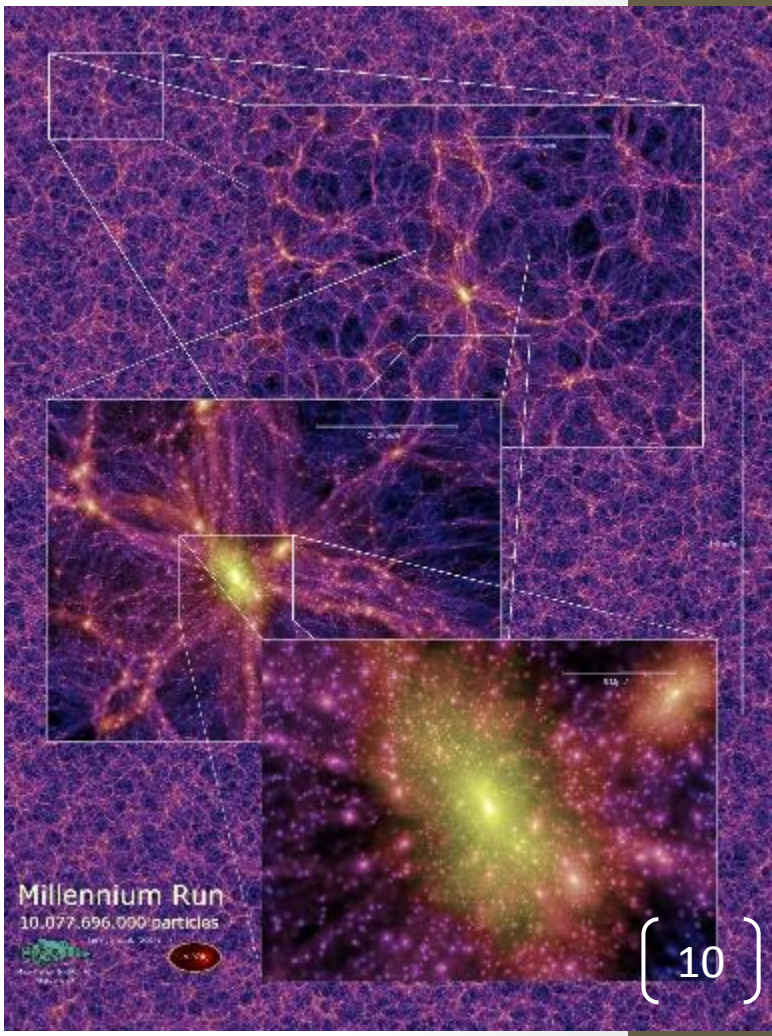
Halo

In terms of Physics:

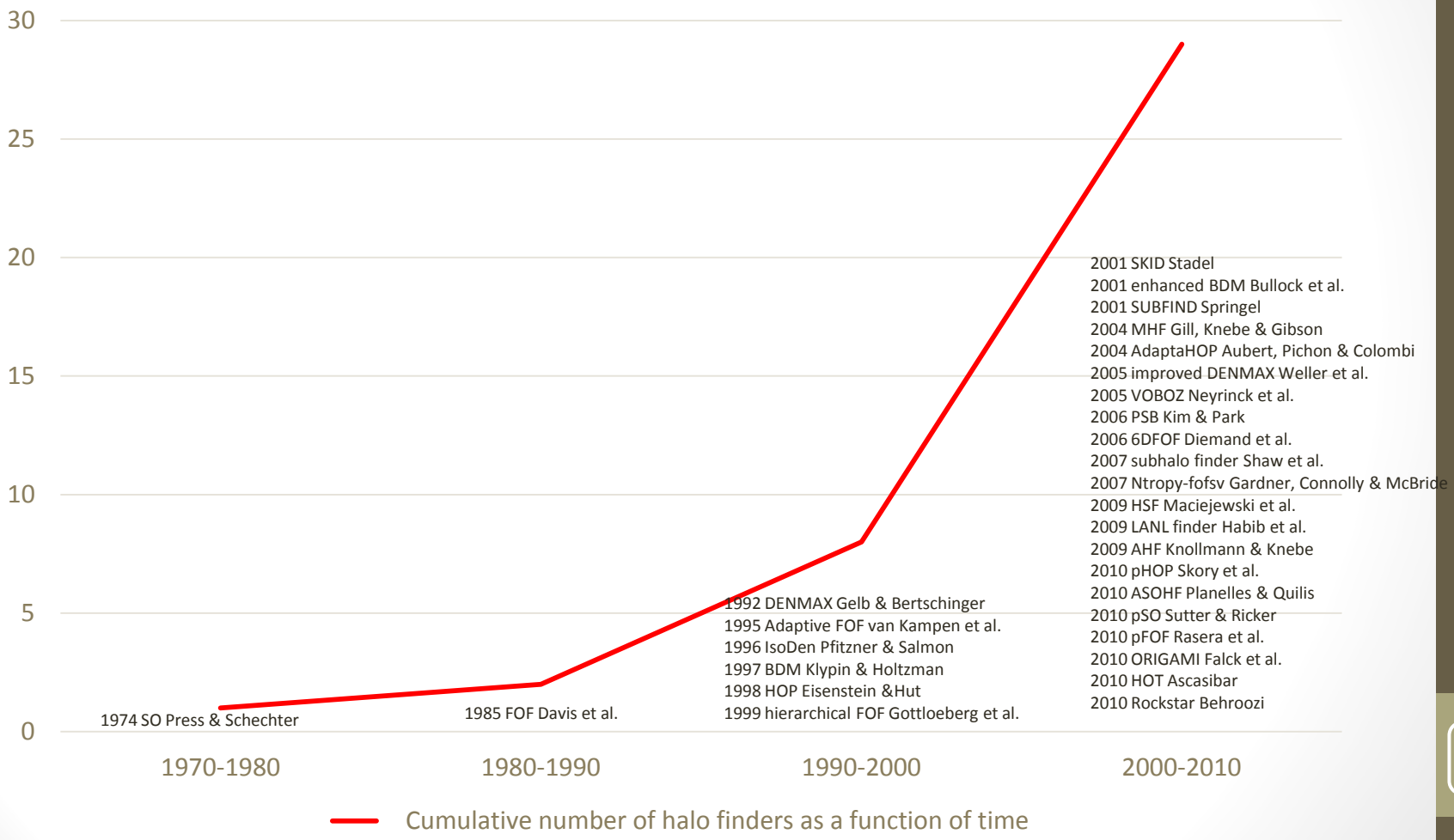
- Galaxies are thought to form in halos

Defining property:

- Macro structure with high mass concentration



Halo finding algorithms



The Halo-Finder Comparison Project
[Knebe et al, 2011]

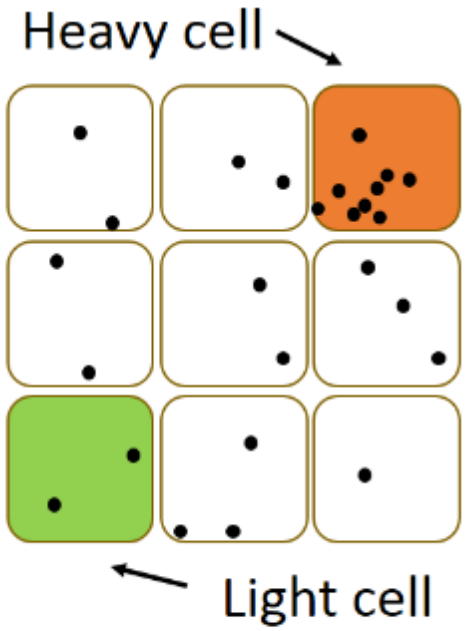
Streaming Solution:

Our goal:

- Reduce halos finding problem to one of the existing problems in streaming setting
- Apply ready-to-use algorithms

haloes ≈ heavy hitters?

- To make a reduction to heavy hitters we need to discretize the space.
- Naïve solution is to use 3D mesh:
  - Each particle now replaced by cell id
  - Heavy cells represent mass concentration
  - Grid size is chosen according to typical halo size



Memory

- Dataset size: $\sim 10^9$ particles
  - Any in-memory algorithm: 12 GB
  - Pick-and-Drop: 30 MB

- GPU acceleration
  - One instance of Pick-and-Drop algorithm can be fully implemented by separate thread of GPU
  - Count Sketch algorithm have two time-consuming procedures: evaluating the hash functions and updating the queue. The first one can be naively ported to GPU

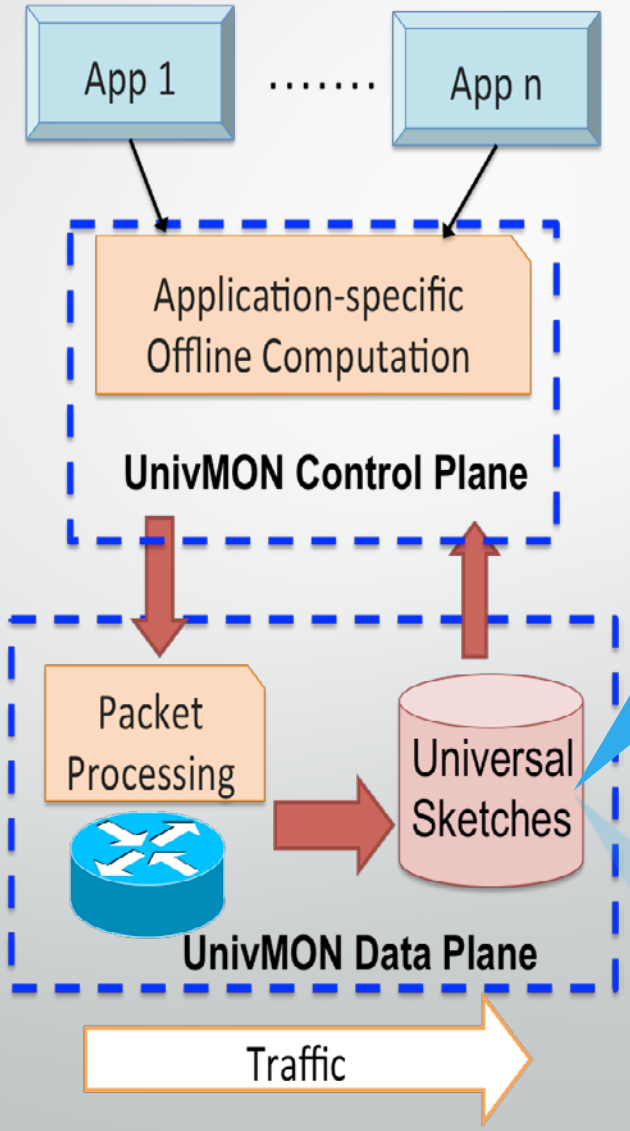
HotNets 2015

Enabling a "RISC" Approach for Software-Defined Monitoring using Universal Streaming

Zaoxing Liu*, Greg Vorsanger*, Vladimir Braverman*, Vyas Sekar†

*Johns Hopkins University, † Carnegie Mellon University

A "RISC" method called Universal Monitoring (UNIVMON)



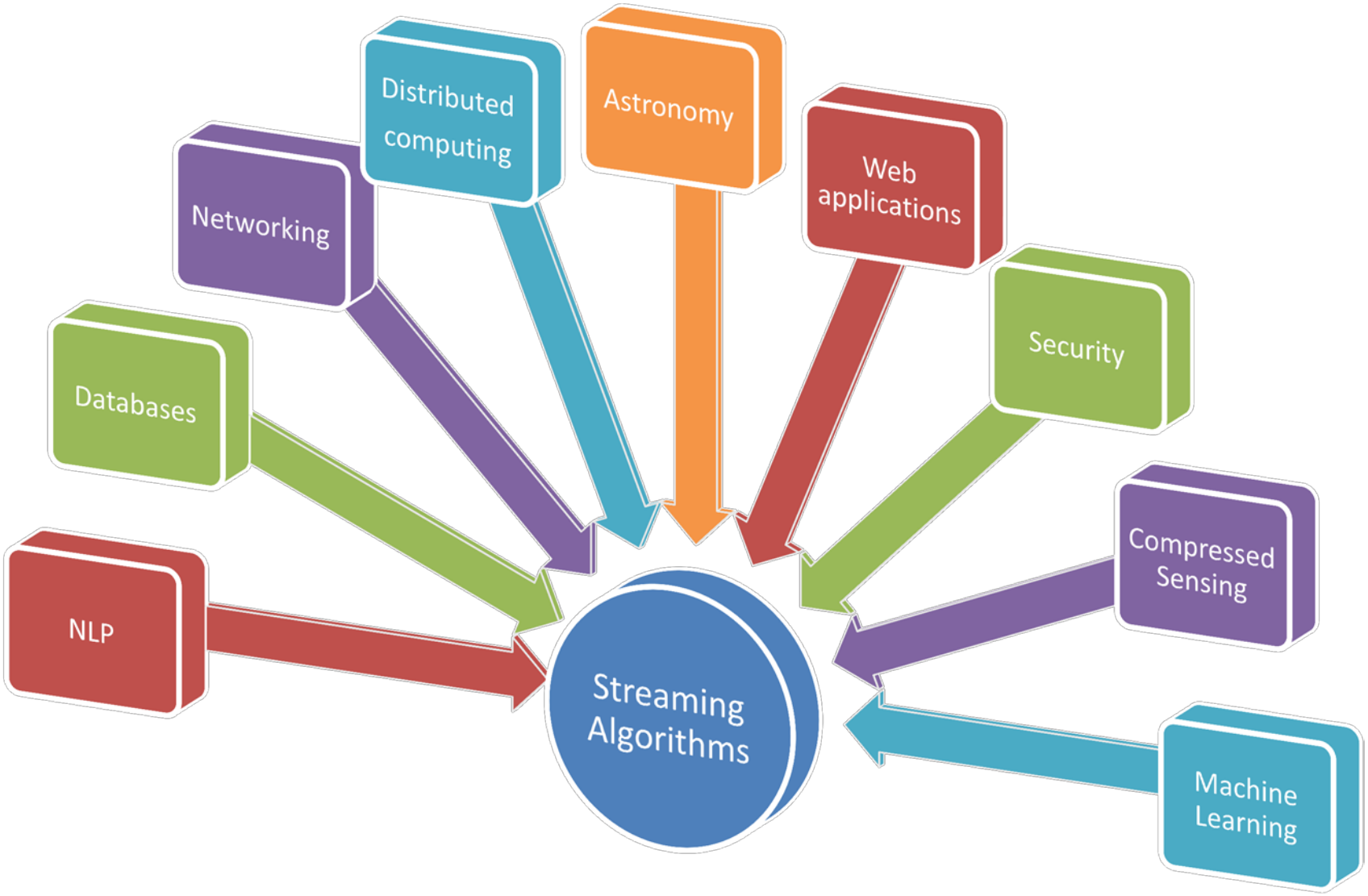
Calculate different metrics without processing the traffic again

Aggregative queries for OLAP

- HyperRoll: Start-up company, Israel
- Acquired by Oracle in 2009
- From **InformationWeek , September 2009:**

  "…HyperRoll has acquired customers in the retail, financial services, and consumer goods sectors... Its products can shorten data warehouse loading times and speed up query executions by a factor of 10…"

Applications



Thank you

- vova@cs.jhu.edu

- http://www.cs.jhu.edu/~vova