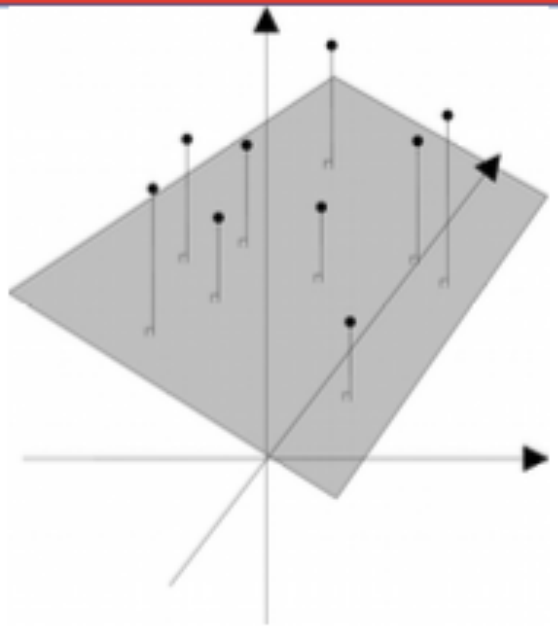
Scalable Big Data Clustering by Random Projection Hashing

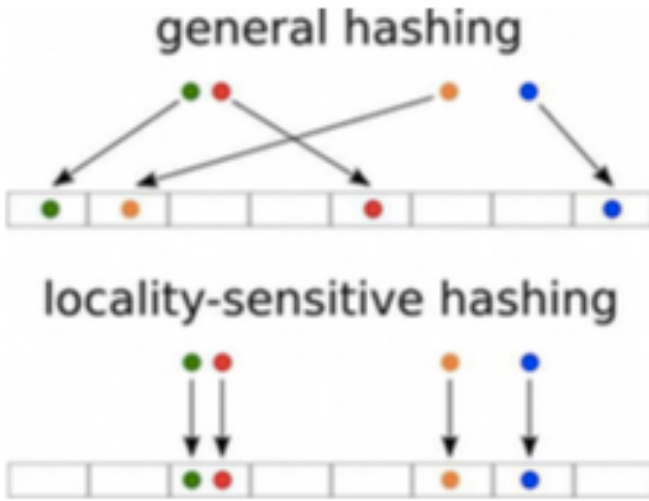
PHILIP WILSEY, UNIVERSITY OF CINCINNATI

Cybersecurity Research Acceleration Workshop and Showcase, Indianapolis, IN
October 11, 2017

Scalable Big Data Clustering by Random Projection Hashing

- High-performance data clustering on high-dimensional data
- Random projection to reduce dimensionality
- Locality Sensitive Hashing
- Distributed, map-reduce capability that operates without data exchange
- Very fast





Algorithm	ARI	Purity	WCSSE	Time
Kmeans	0.461	0.600	182,168.70	24.746
SL	0.000	0.189	556,519.10	413.880
CL	0.327	0.377	222,044.40	414.330
AL	0.332	0.359	236,142.80	414.100
Ward's	0.491	0.660	191,441.10	414.450
SOTA	0.314	0.397	210,490.10	14.244
RPHash	0.363	0.508	210,628.60	0.484
RPHash_Tree	0.449	0.609	194,688.70	0.262

Sponsor

- NSF: Grant ACI-1440420

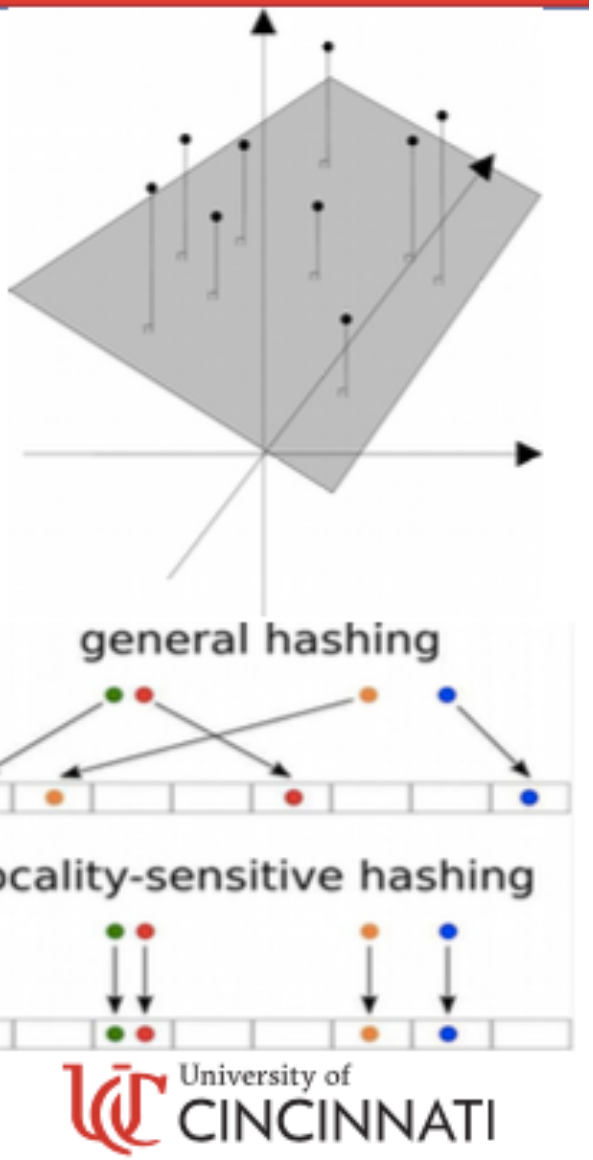
Contact Us

- wilseypa@gmail.com

University of CINCINNATI

Scalable Big Data Clustering by Random Projection Hashing

- High-performance data clustering on high-dimensional data
- Random projection to reduce dimensionality
- Locality Sensitive Hashing/Tree Walk
- Distributed, map-reduce capability that operates without data exchange
- Very fast, scalable performance



Sponsor: NSF: Grant ACI-1440420

Contact Us: Philip A. Wilsey wilseypa@gmail.com

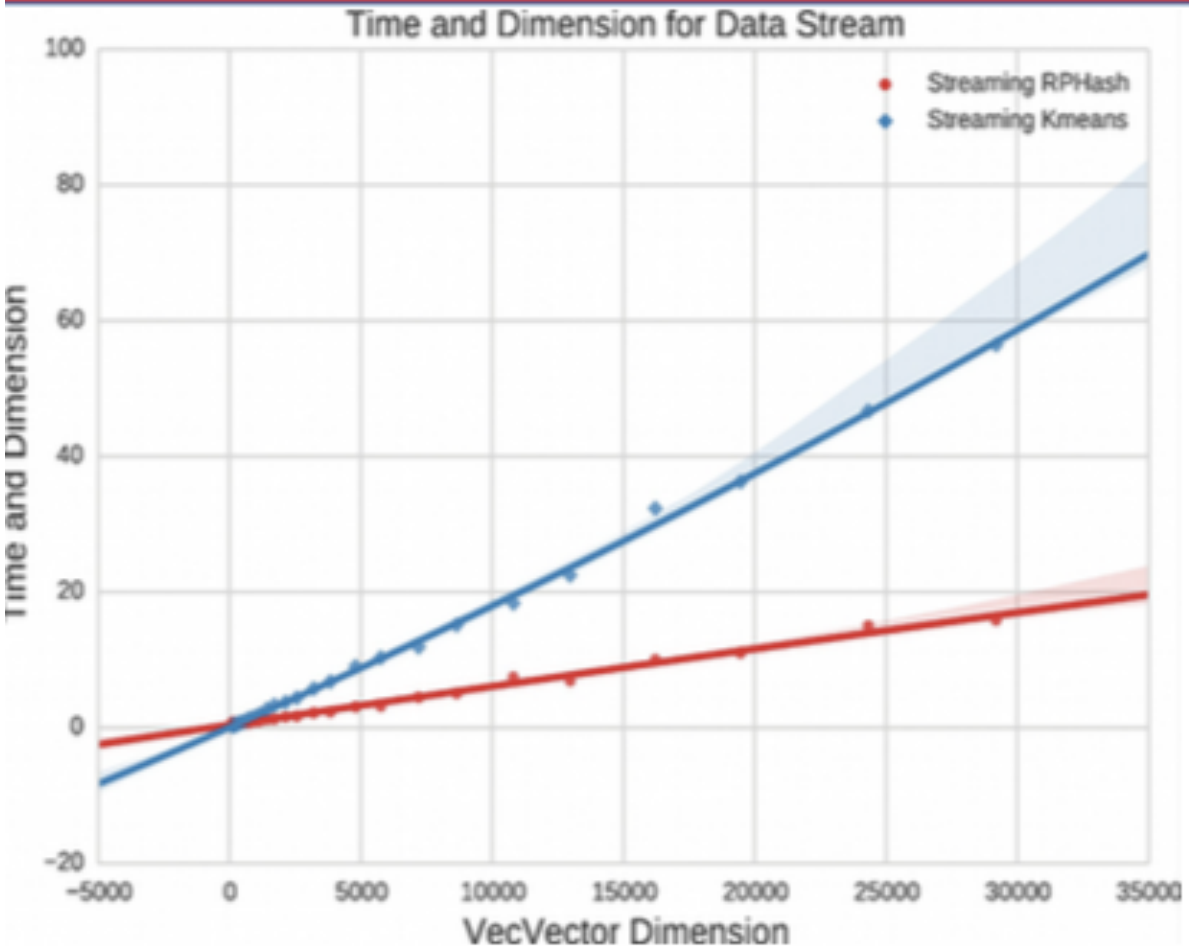
University of CINCINNATI

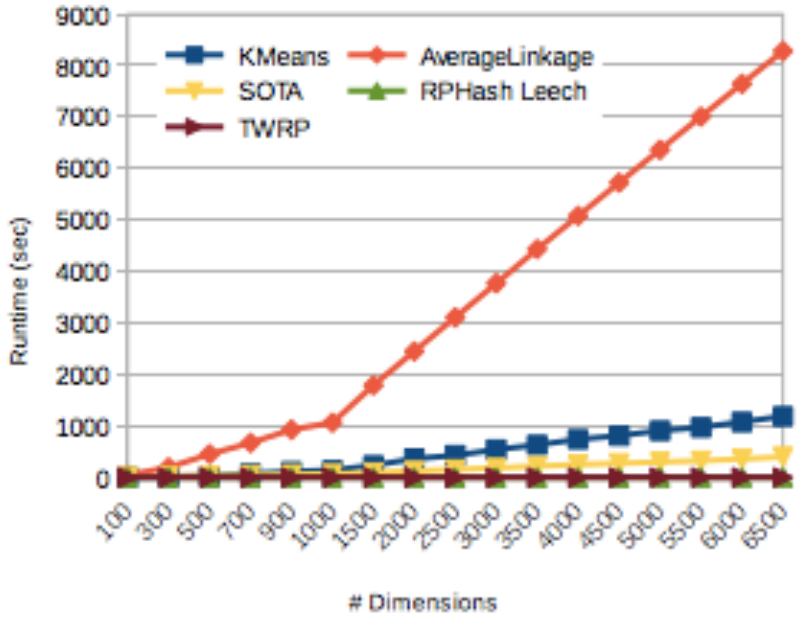
Performance Comparison

Algorithm	ARI	Purity	WCSSE	Time (sec)
Kmeans	0.461	0.600	182,168.70	24.746
SL	0.000	0.189	556,519.10	413.880
CL	0.327	0.377	222,044.40	414.330
AL	0.332	0.359	236,142.80	414.100
Ward's	0.491	0.660	191,441.10	414.450
SOTA	0.314	0.397	210,490.10	14.244
RPHash	0.363	0.508	210,628.60	0.484
RPHash_Tree	0.449	0.609	194,688.70	0.262

University of CINCINNATI

Scalability





University of CINCINNATI