Data Mining of Chemical Compounds Using Functional Groups

Ali Rathore

Chabot College

Electrical Engineering & Computer Science

Mentor: Sayan Ranu

Advisor: Dr. Ambuj Singh

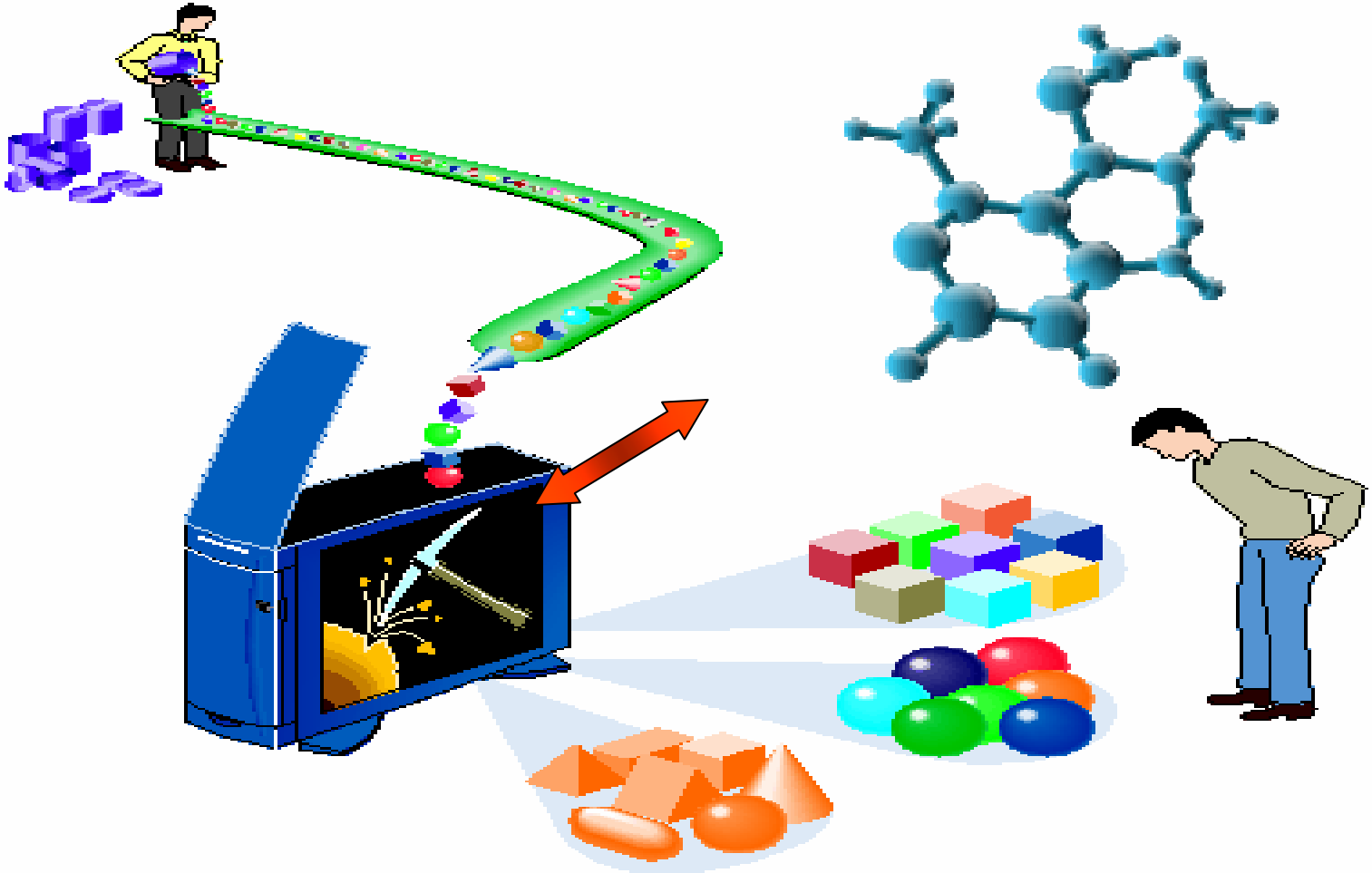
Department: Computer Science

Database & Bioinformatics Lab (DBL)

Funding: National Science Foundation

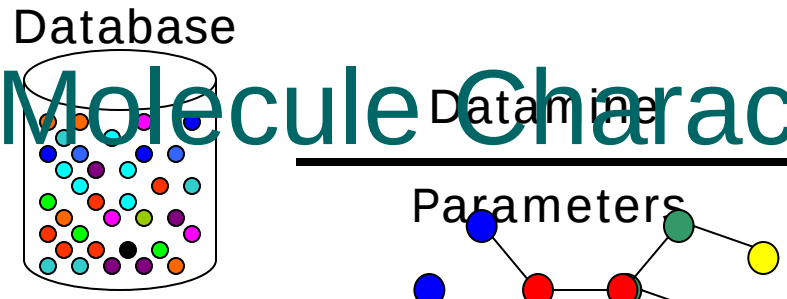
Division of Information & Intelligent Systems

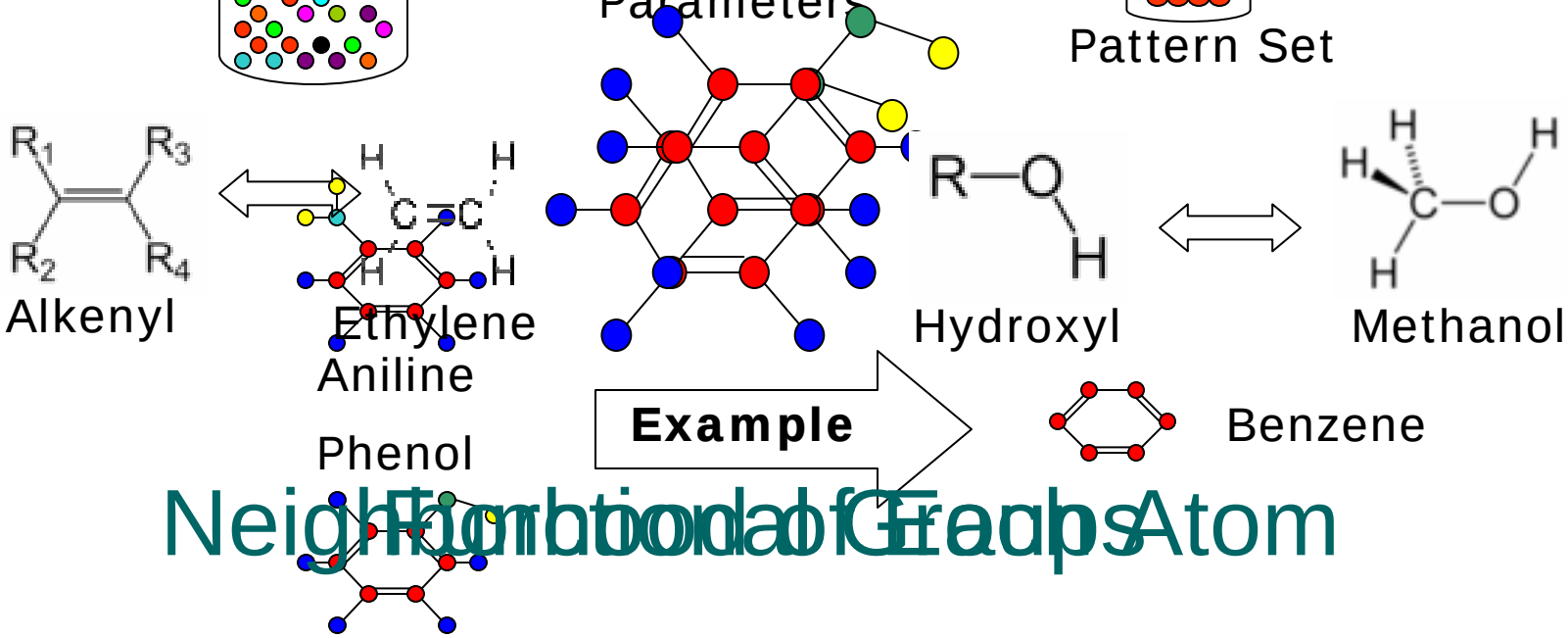
# of Chemical Compounds



Research Goals

Molecule Characterization

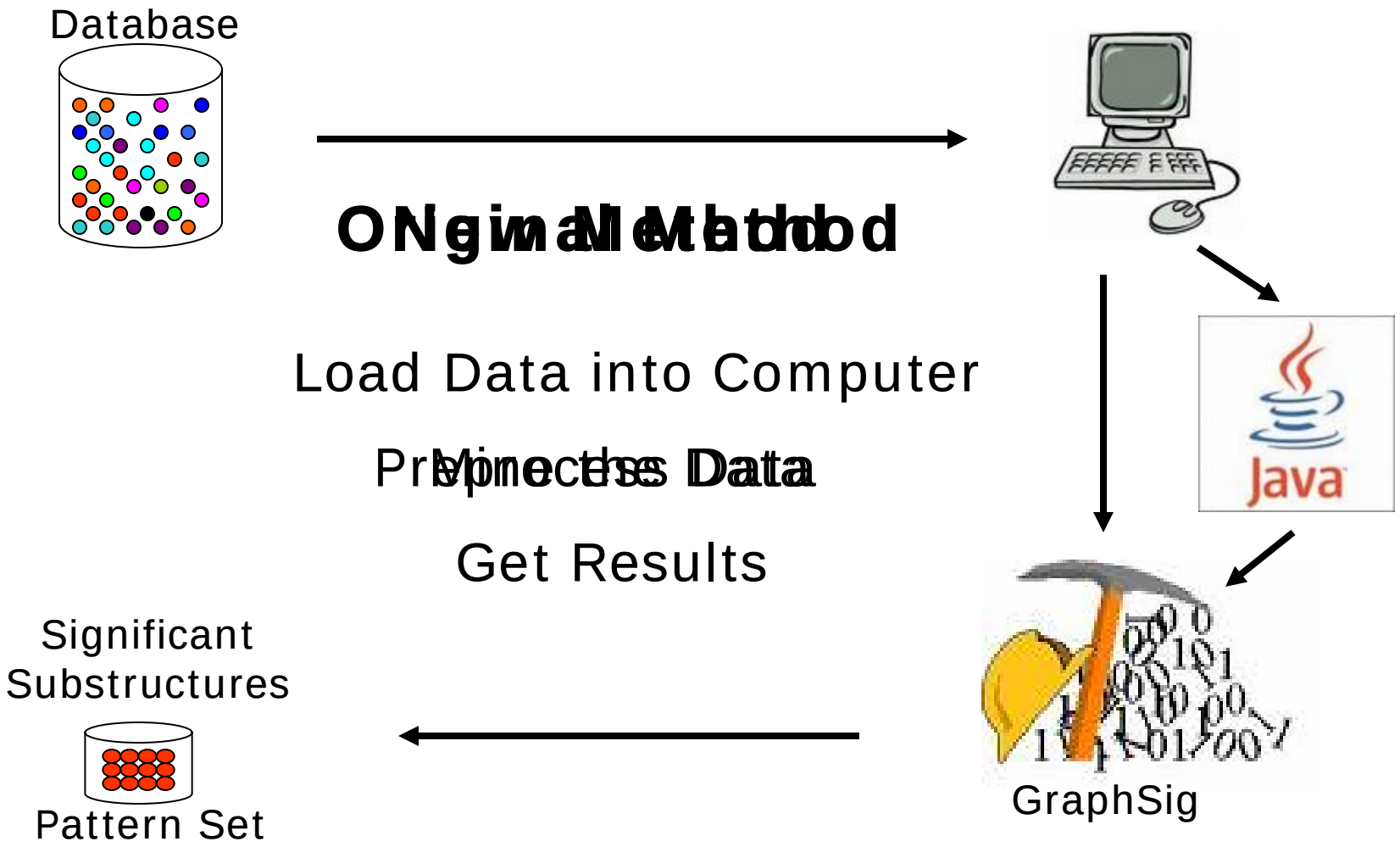




Neighborhood of Each Atom

Formation of Groups

Research Method



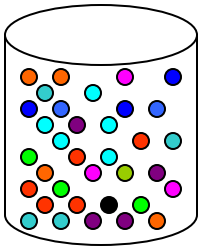
GraphSig Results

Comparison of "Accuracy" (Score out of 100)			
Database	Optimal Assignment Kernel	Scalable Leap Search	**GraphSig**
MCF-7	68	76	**77**
MOLT-4	65	72	**74**
NCI-H23	79	79	**80**
OVCAR-8	67	78	**79**
P388	79	84	**84**
PC-3	66	76	**76**
SF-295	75	77	**80**
SN12C	75	80	**80**
SW-620	70	76	**77**
UACC-257	65	75	**81**
Yeast	64	71	**73**
Average	**70.2**	**76.7**	**78.2**

S. Ranu, A. Singh. "GraphSig: A Scalable Approach to Mining Significant Subgraphs in Large Graph Databases"

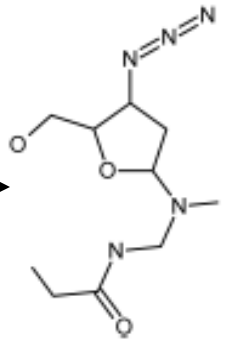
GraphSig Results

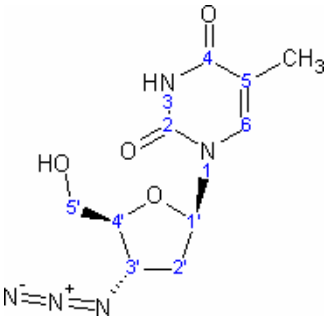
AIDS Database



GraphSig

Parameters

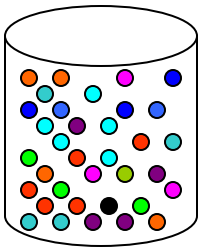




3-azido-thymidine (AZT)

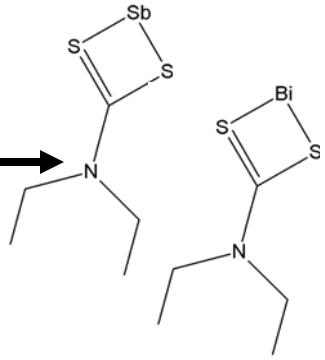
Most used medicine for controlling HIV virus.

Leukemia Database



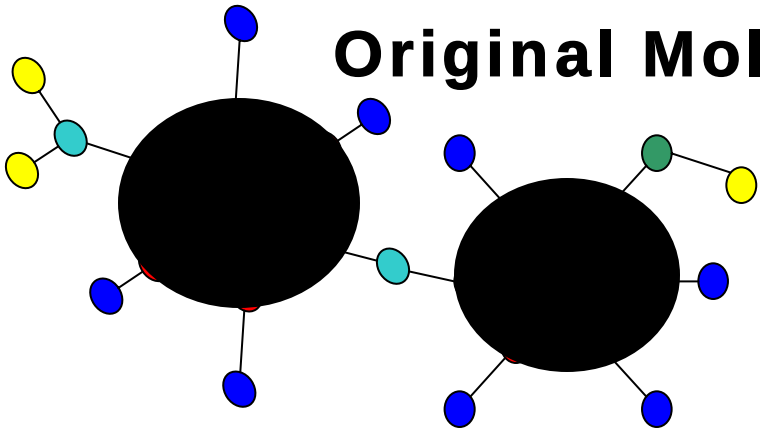
GraphSig

Parameters



- Only difference is presence of Antimony (Sb) and Bismuth (Bi)
- May lead chemists to try other metals from same group
- Sb & Bi cannot be mined using other techniques.

Preprocessing Method

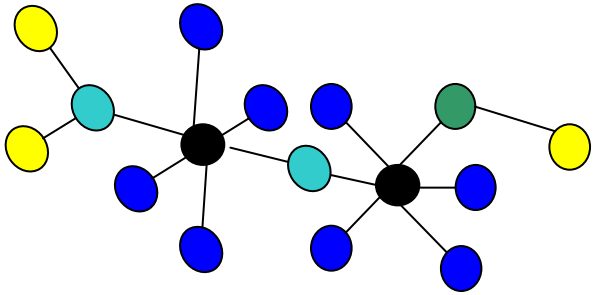


Original Molecules

Find Functional Groups

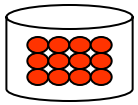
Replace with “Atoms”

New Molecules



GraphSig

Significant Substructures



“Better” Pattern Set

Data Mining Of Chemical Compounds

- Automated extraction of implicit information.
- Discovery of previously unknown patterns.
- Analysis of databases of chemical compounds.
- Allows chemists to:
  - Predict behavior of new compounds.
  - Identify compounds with wanted properties.
- Allows pharmacists to:
  - Create drugs using significant substructures.
  - Classify compounds as active or inactive.

Summary

Acknowledgements

Liu-Yen Kramer, CNSI Education Programs Development Analyst

Dr. Evelyn Hu, CNSI Scientific Director

Jens-Uwe Kuhn, INSET Program Coordinator

Dr. Nick Arnold, INSET Faculty Coordinator

Sayan Ranu, Graduate Student Mentor

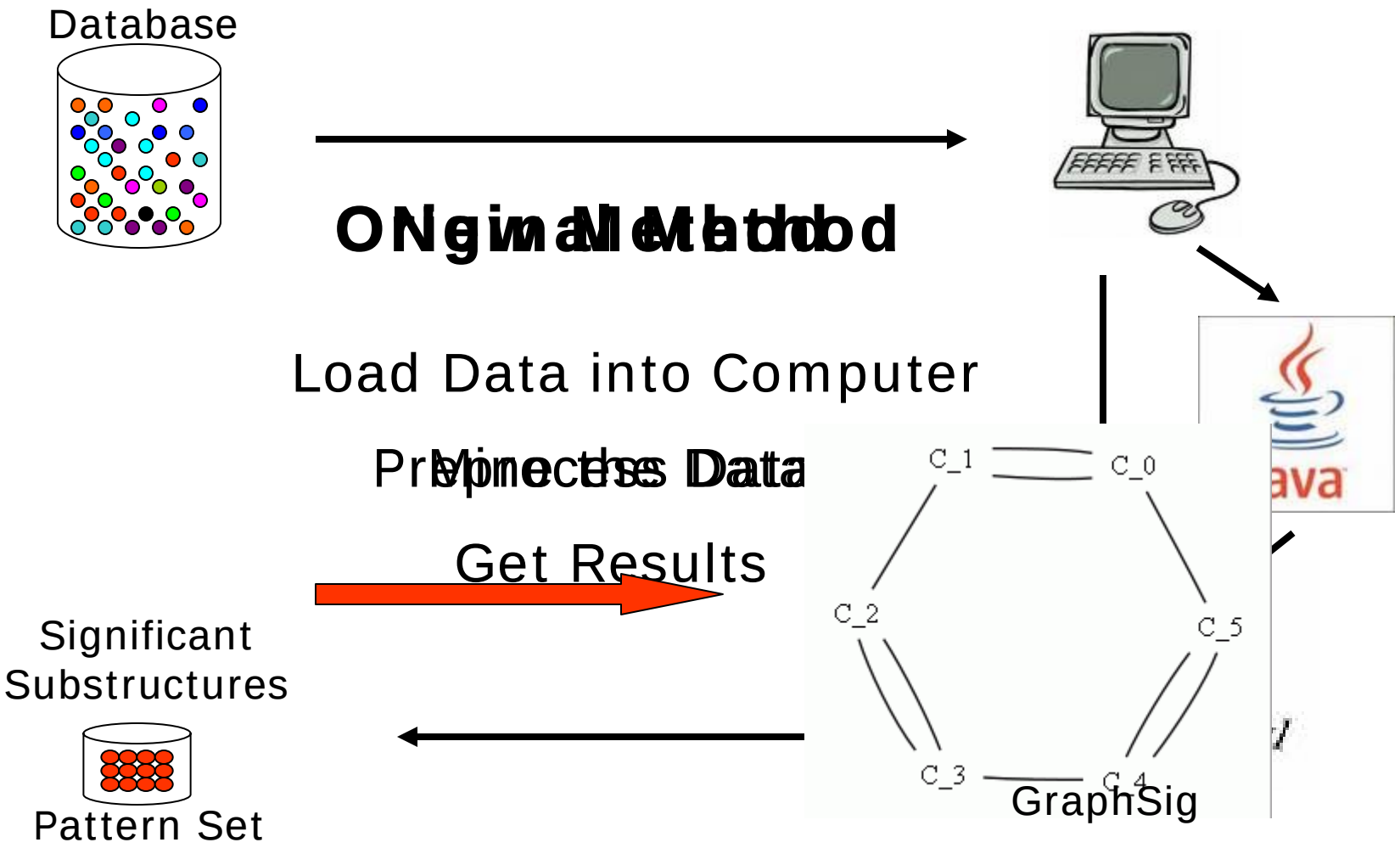
Dr. Ambuj Singh, Computer Science Faculty Advisor

Everyone at Database & Bioinformatics Lab

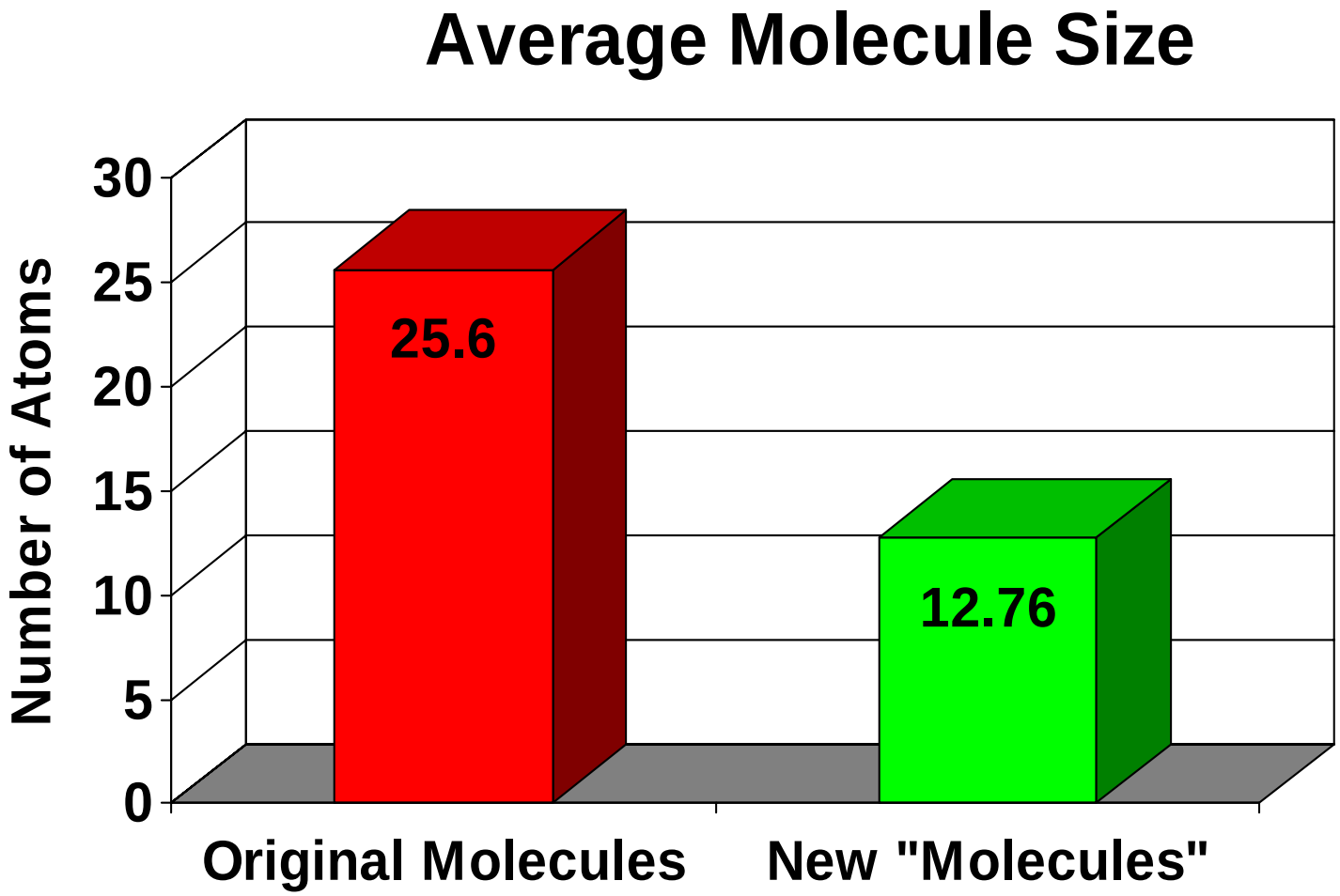
Thank You

Questions?

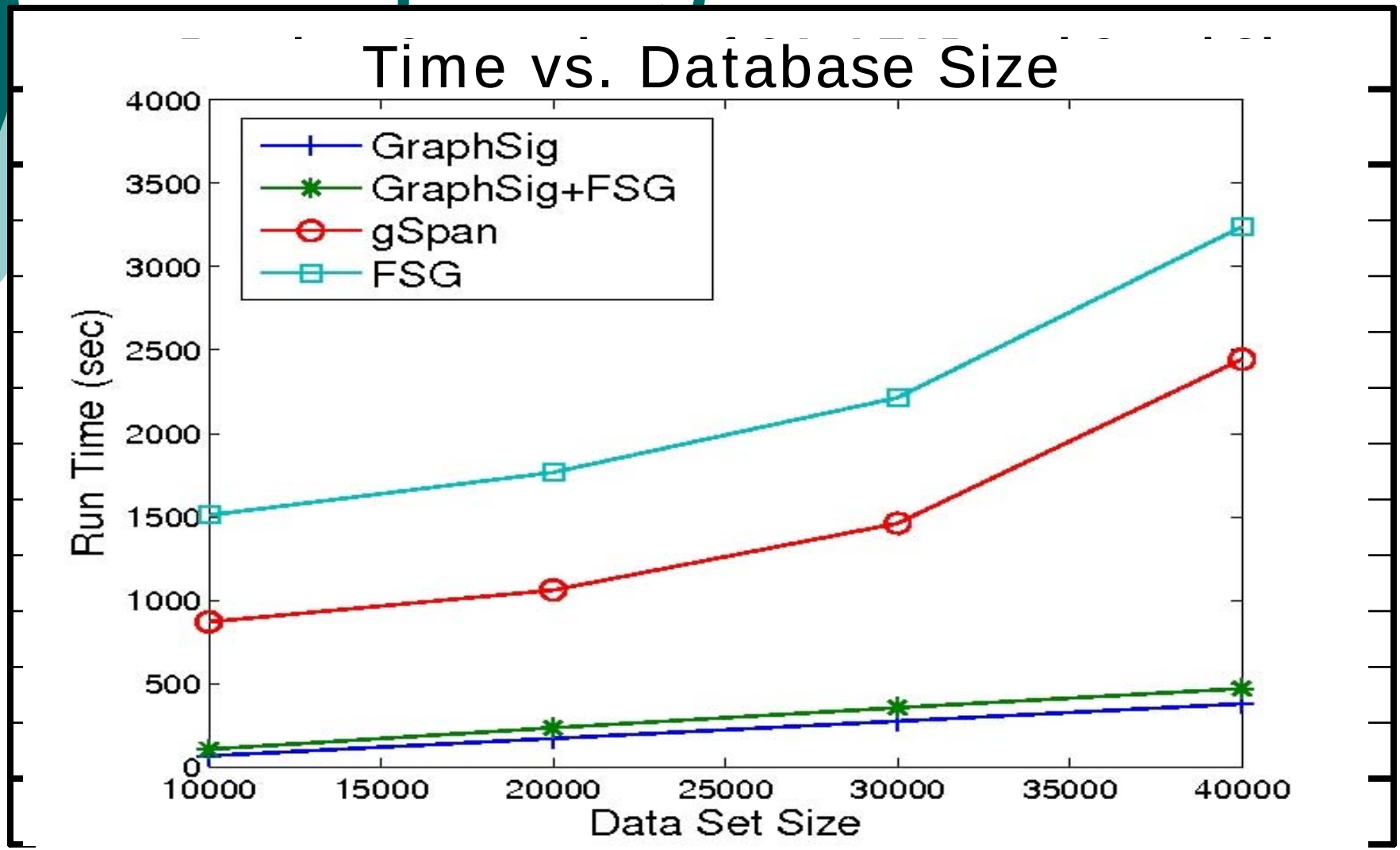
Research Method



Preprocessing Results



GraphSig Results



S. Ranu, A. Singh. “GraphSig: A Scalable Approach to Mining Significant Subgraphs in Large Graph Databases”