Indexing and Mining Topological Patterns for Drug Discovery

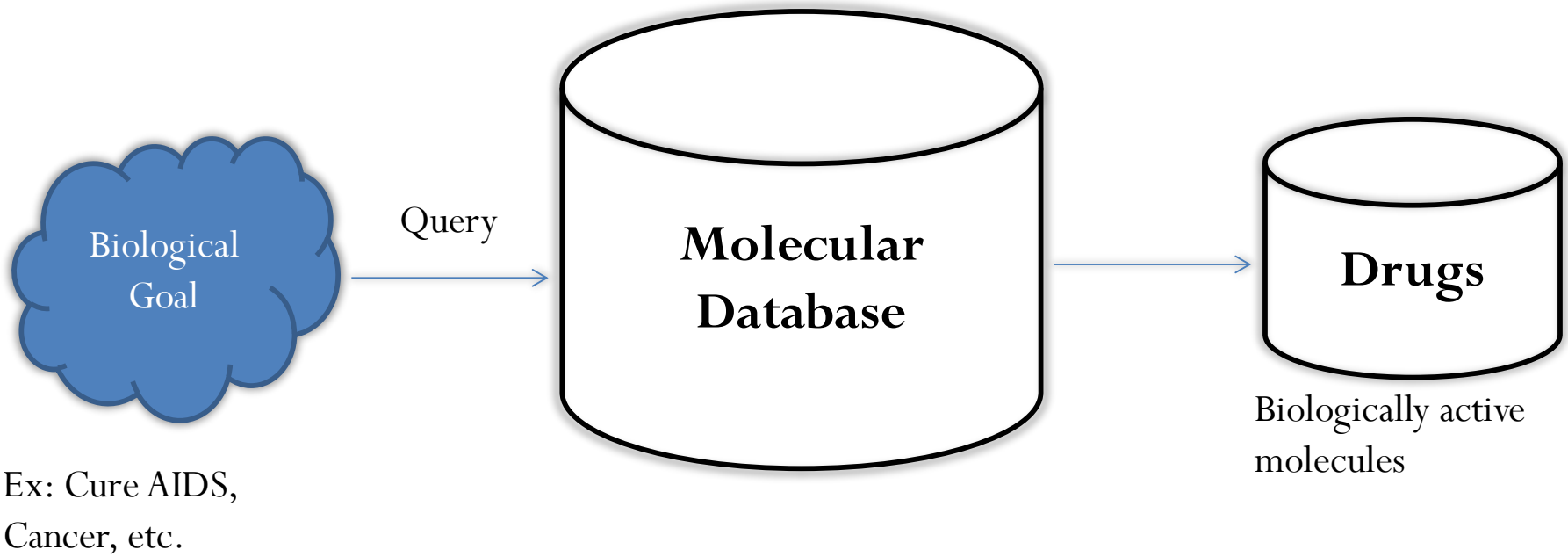
Sayan Ranu and Ambuj K. Singh

The Data Mining and Bioinformatics Lab

UC SANTA BARBARA

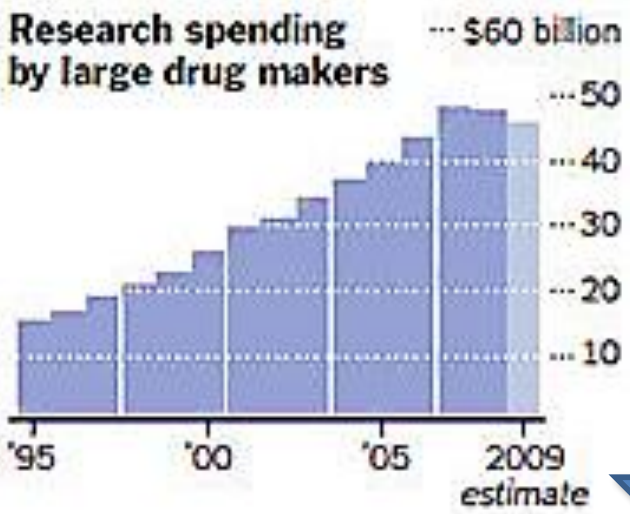
UNIVERSITY OF CALIFORNIA

What is Drug Discovery?



The Economics of Drug Discovery

Company	Total Revenues (USD billions)
Johnson & Johnson[2]	61.90[3]
Pfizer[4]	50.01[3]
Roche[5]	47.35[3]
GlaxoSmithKline[6]	45.83[3]
Novartis[7]	44.27[3]
Sanofi[8]	41.99[3]
AstraZeneca[9]	32.81[10]
Abbott Laboratories[11]	30.76[10]
Merck & Co.[12]	27.43[10]
Bayer HealthCare[13]	22.30



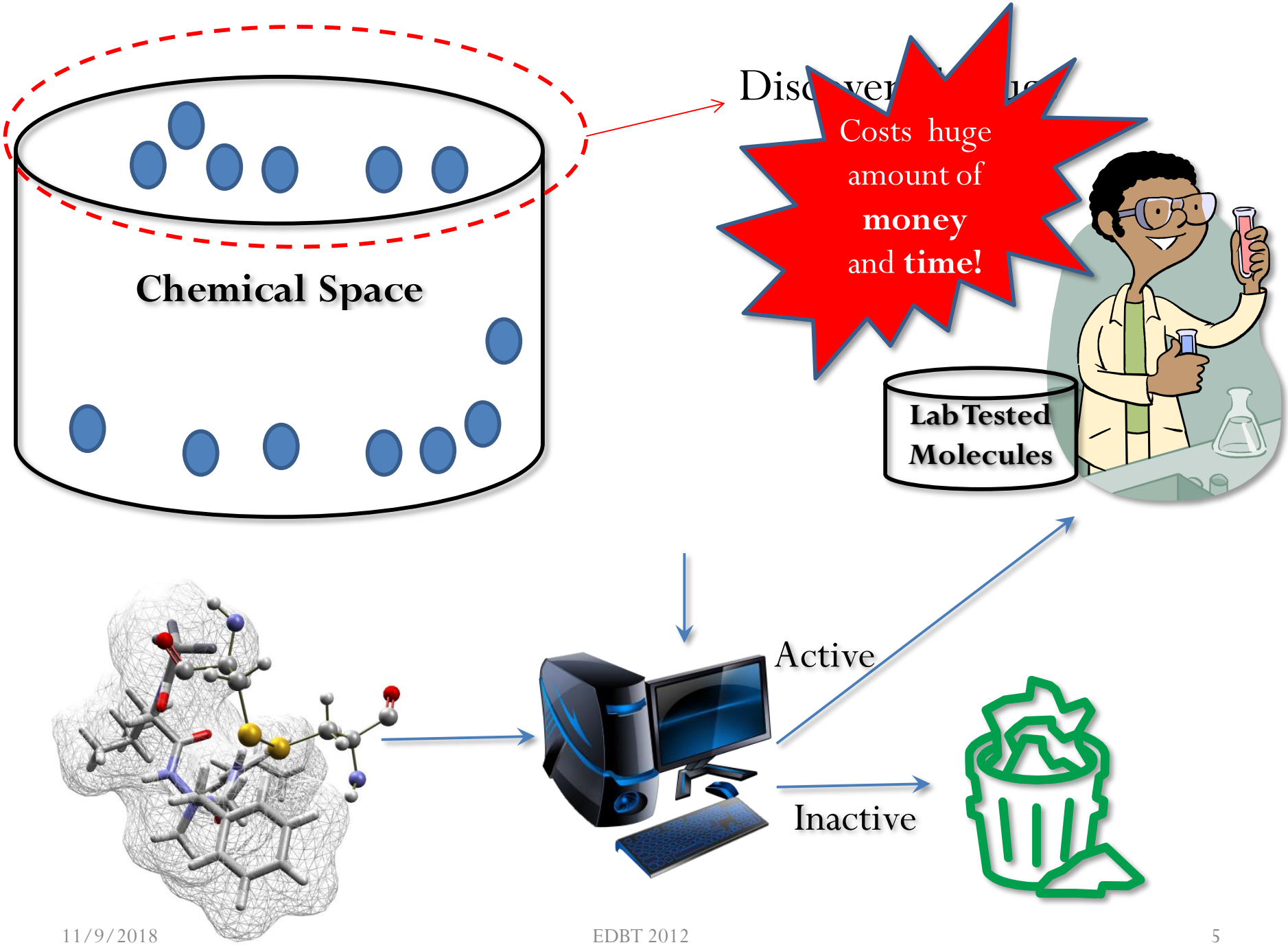
New Drug:
Cost: $880 M- $1.3 B
Time: 15-16 years

Why?

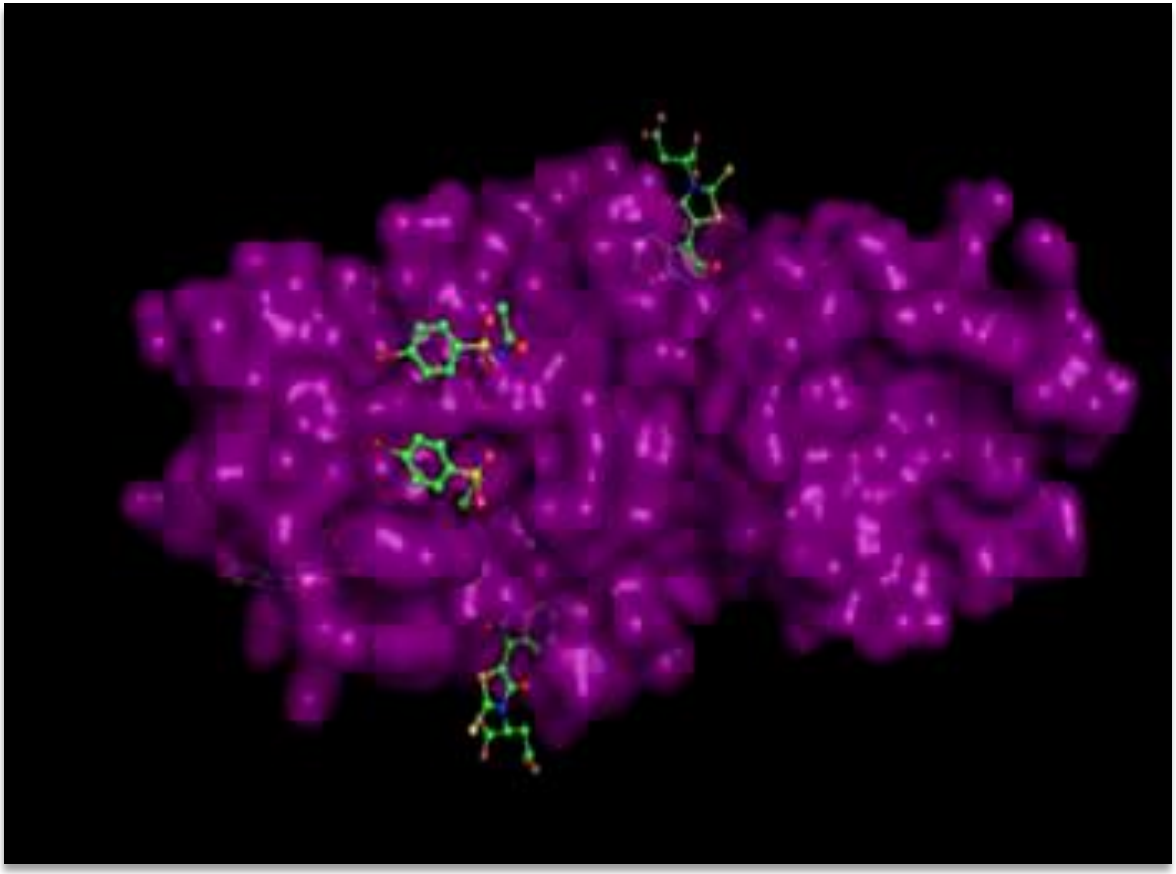
Then

Now

"Low hanging fruits" have already been picked!


Chemical Space
Costs huge amount of money and time!
Lab Tested Molecules
Active
Inactive

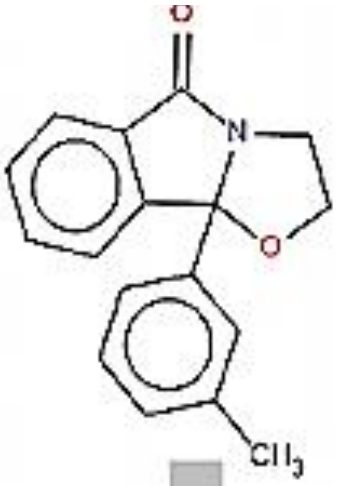
Drug-Protein Binding



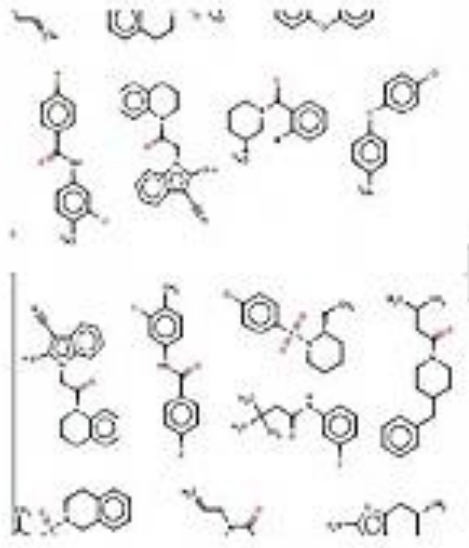
Common Prediction Approaches

How can we *index* molecular databases?

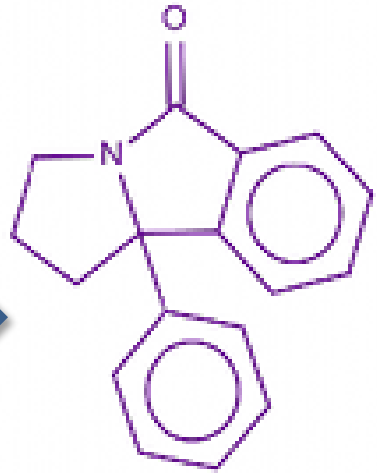
Known Active



Repository of Compounds

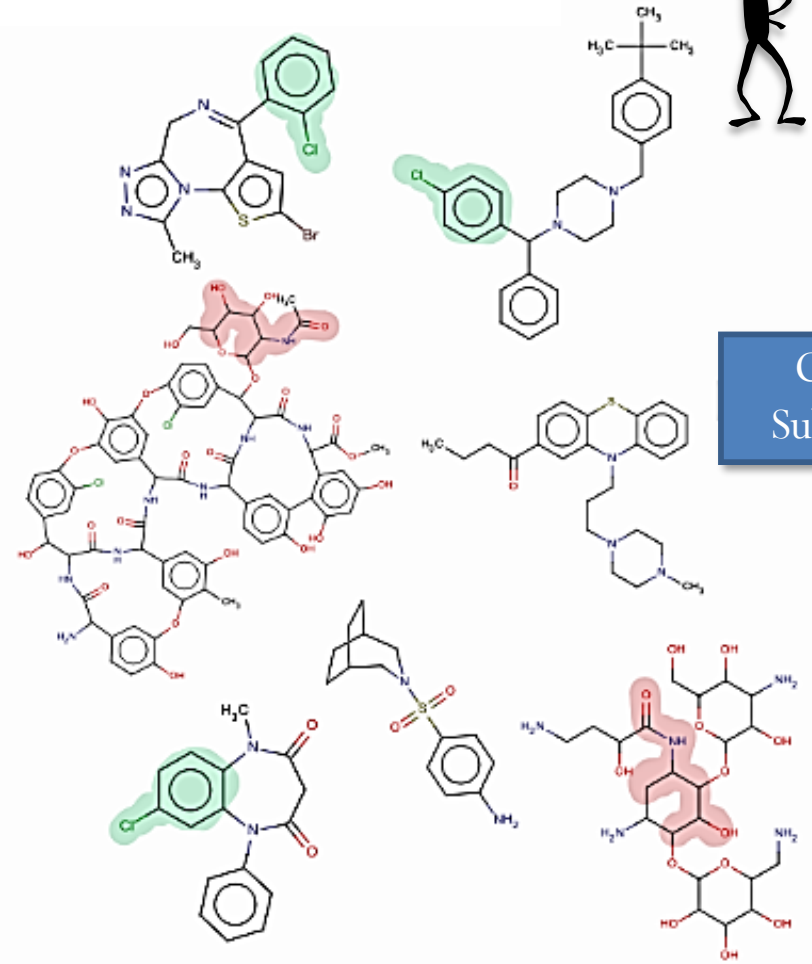


Similarity Search



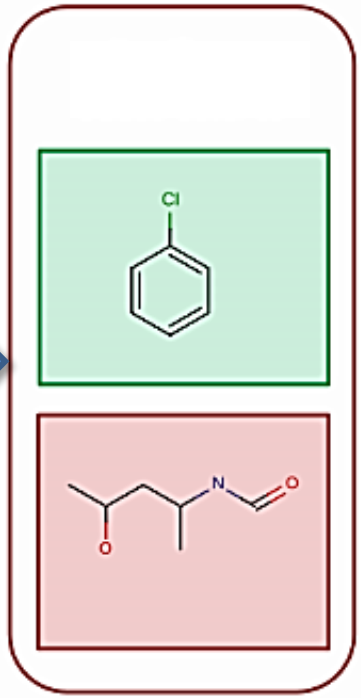
Common Prediction Approaches

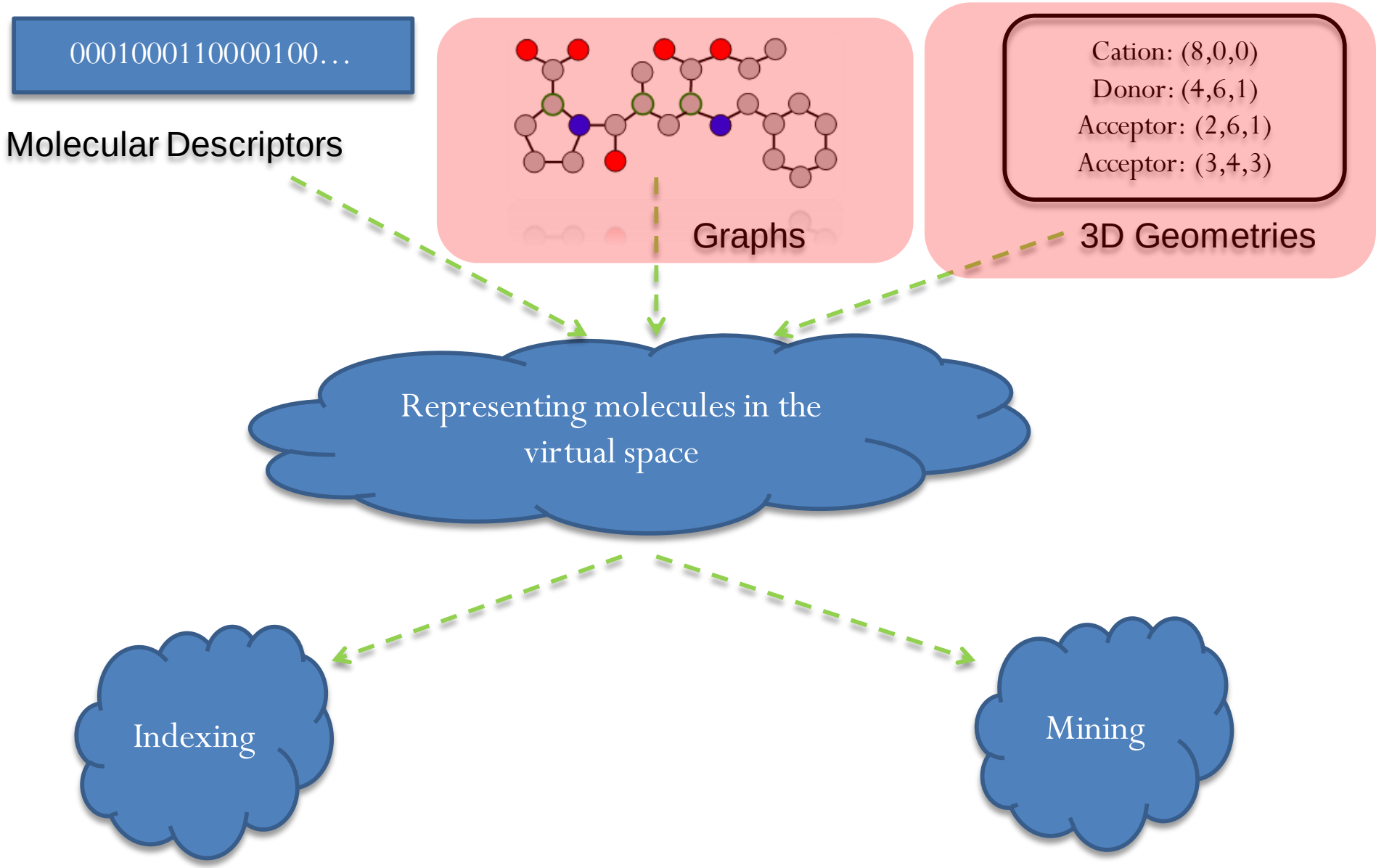
Database of Active Molecules



How can we *mine* molecular databases?

Correlated Substructures

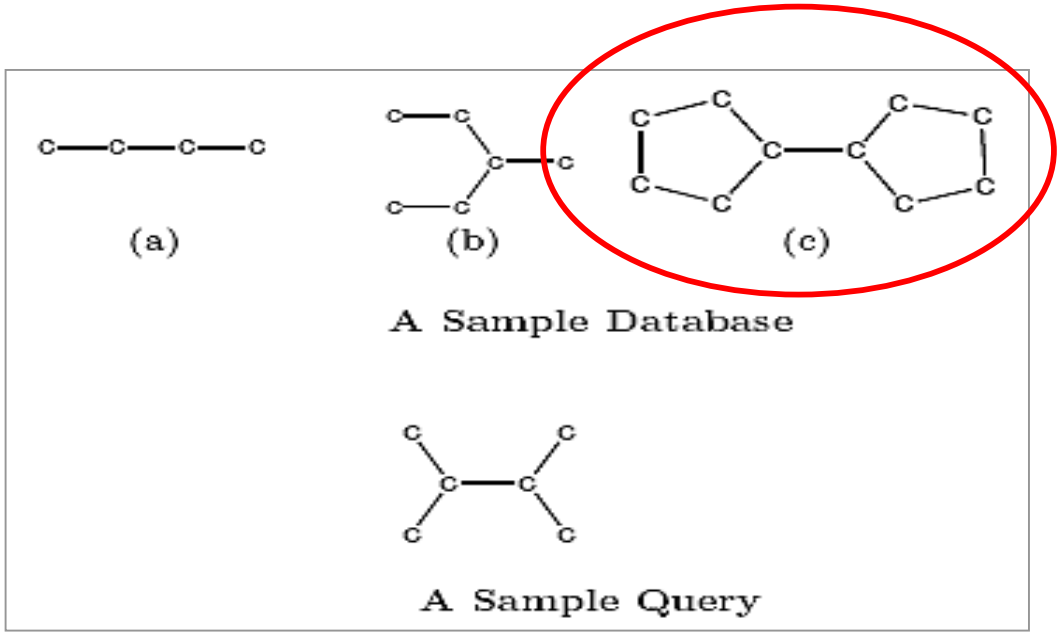



0001000110000100…
Molecular Descriptors
Graphs
Cation: (8,0,0)
Donor: (4,6,1)
Acceptor: (2,6,1)
Acceptor: (3,4,3)
3D Geometries
Representing molecules in the virtual space
Indexing
Mining

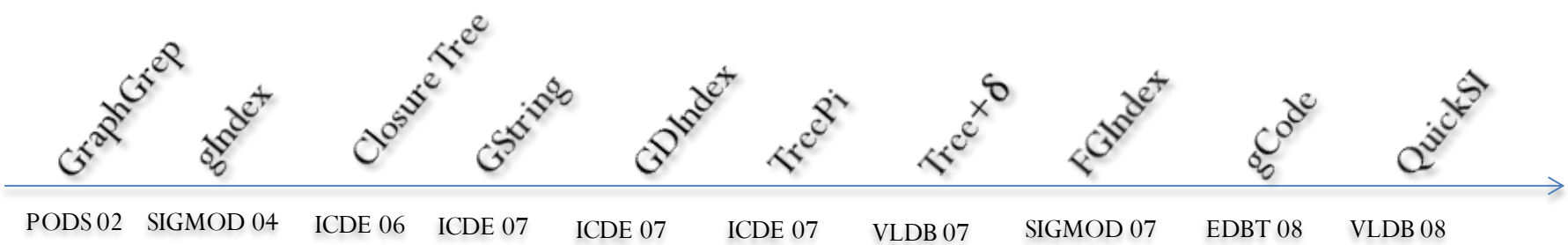
Queries

- *Subgraph* Searches
  - Find molecules *containing* a specific functional group
  - Find molecules *containing* a substructure with a known desired activity
  - Computer Science problem: subgraph isomorphism
    - NP-complete
- *Similarity* Searches
  - Find molecules *structurally similar* to a known active
  - Computer Science problem: top-*k*/range search
  - Graph based distance measures

Subgraph Searches

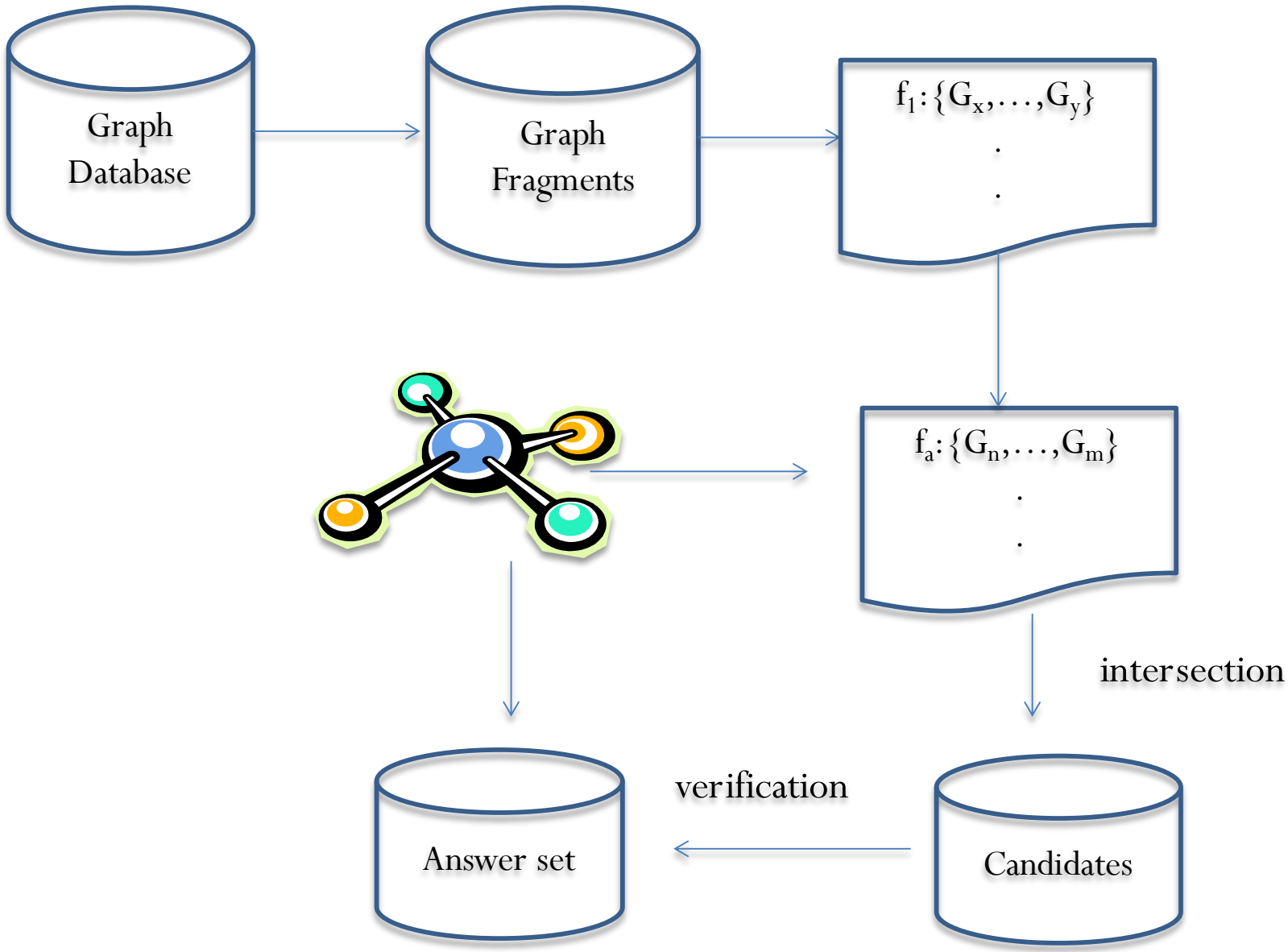


Fragment Based Indexing



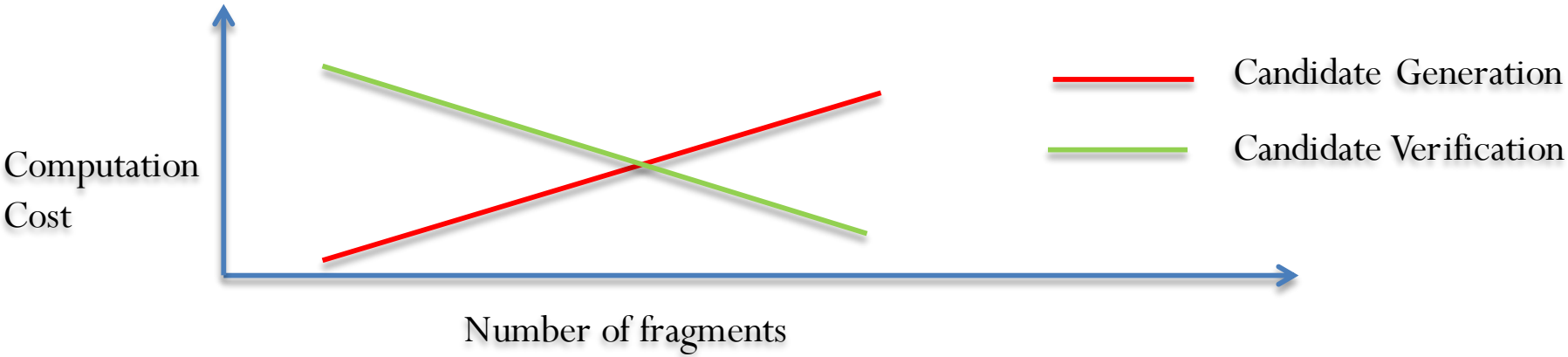
gIndex: Pruning Idea

- *Cheaper* to perform subgraph isomorphism on *small graphs*
- If graph fragment $x \in q$ and $x \notin g$, then $q \notin g$
  - *q: query graph*
  - *g: a database graph*

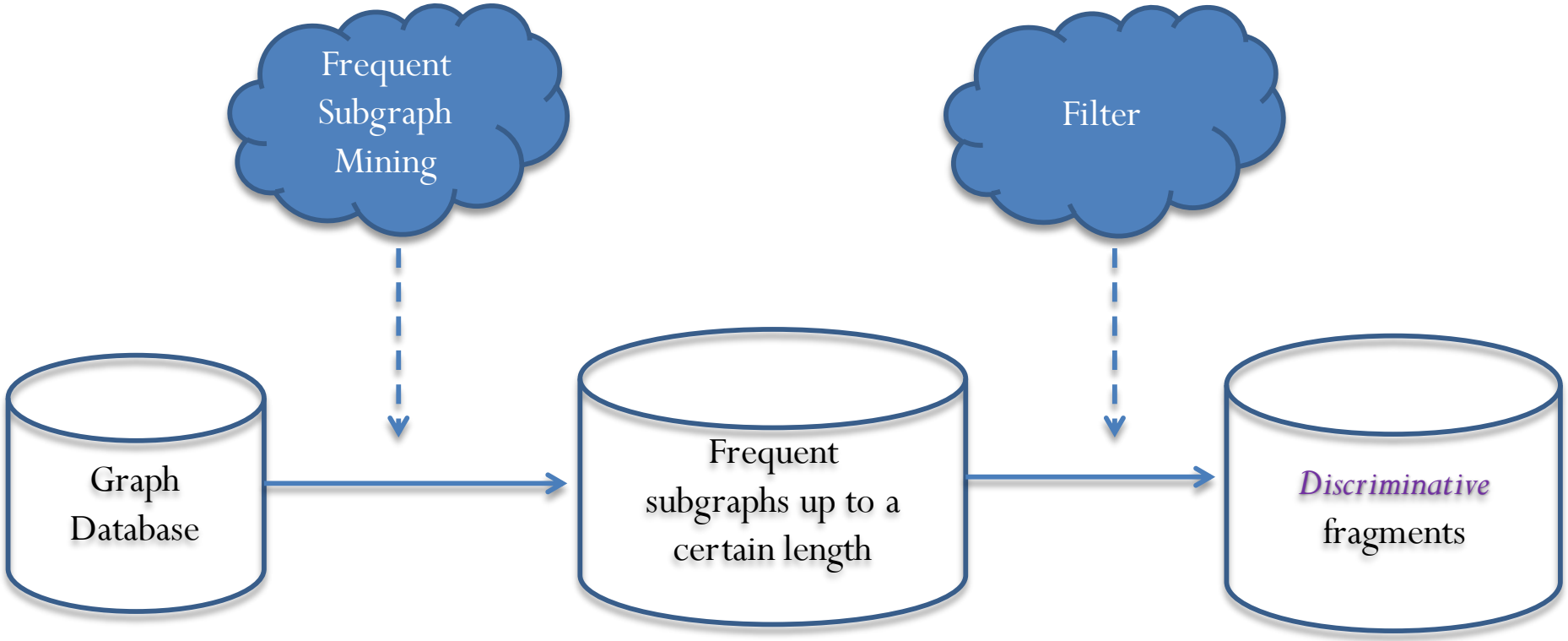

Graph Database
Graph Fragments
$f_1:\{G_x,\ldots,G_y\}$
$f_a:\{G_n,\ldots,G_m\}$
intersection
verification
Answer set
Candidates

Which fragments to Index?

- Cost= $(|F| * c_f)+(|C| * c_q)$
  - F = indexed fragments
  - c_f = average cost of subgraph isomorphism for fragments
  - C= candidate set
  - c_q = average cost of subgraph isomorphism on candidates



gIndex: Discriminative Fragments



Discriminative Fragments

- Indexed fragment set $\mathbb{F}$
- Should we index fragment X?
  - Discriminative ratio:
    - $r = \frac{|\cap_{f \in \mathbb{F}, f \subseteq X} D_f|}{|D_x|}$

Candidates *without indexing* X

Candidates if X is *indexed*

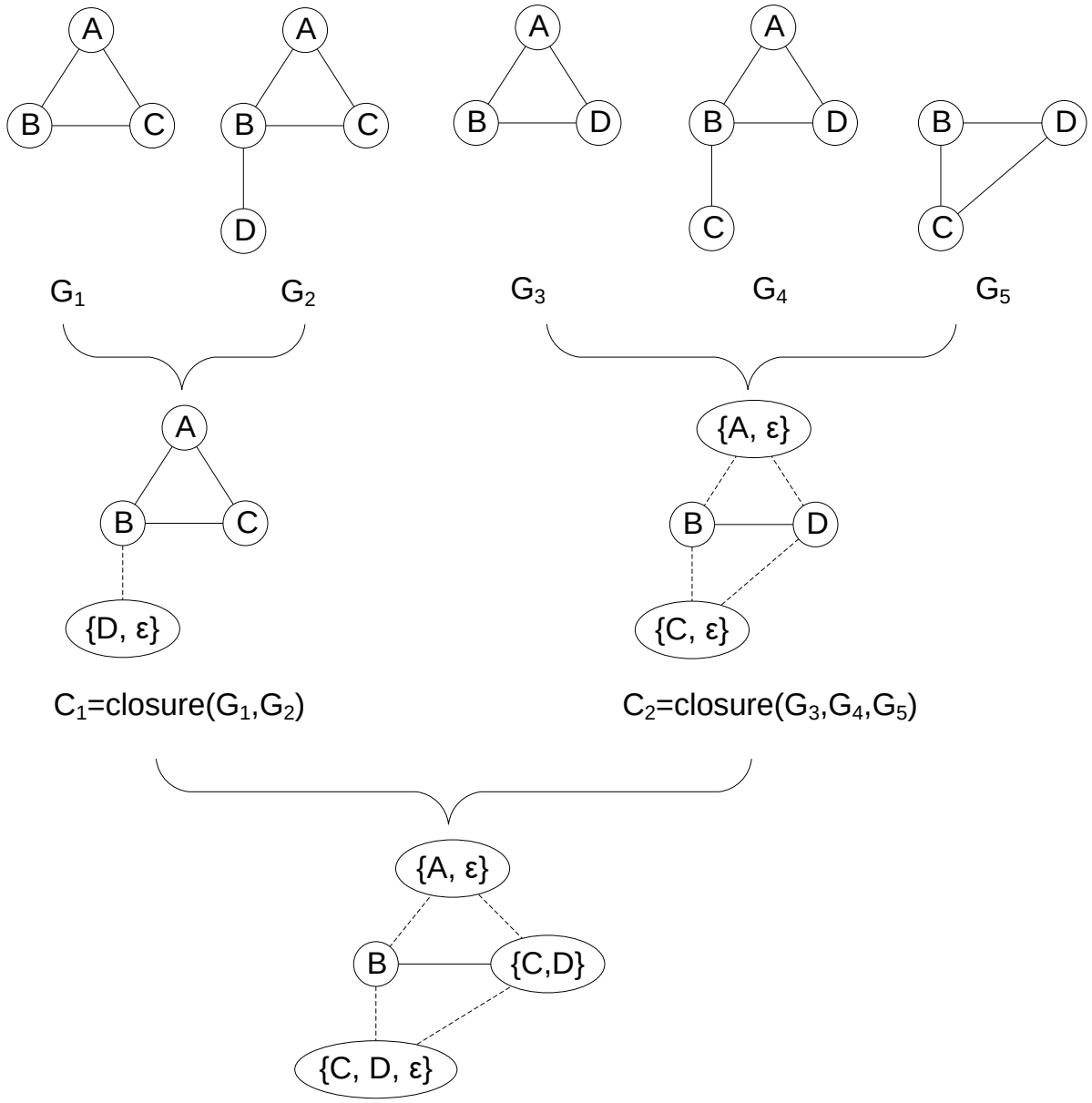
  - D_x: number of database graphs containing fragment X

- Select X if $r \geq \theta$

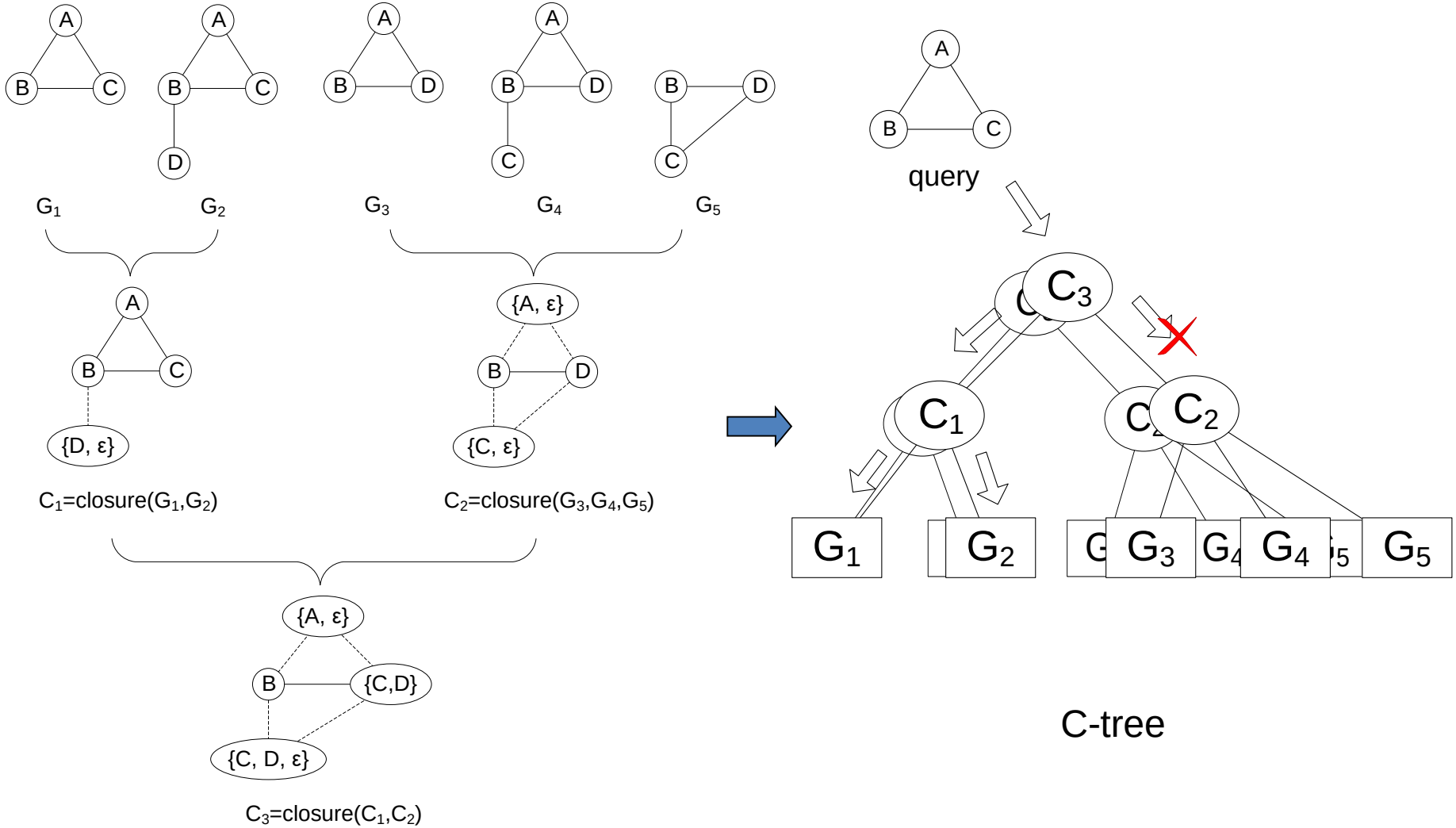
Closure tree: Basic idea

- A closure is a *summary* of multiple graphs
- Let C be closure of graphs $g_1, \ldots g_n$
  - if query $q \not\subseteq C$, then $q \not\subseteq g_i \ \forall_i$

Graph Closures



Closure Tree



Advantage of Closure tree

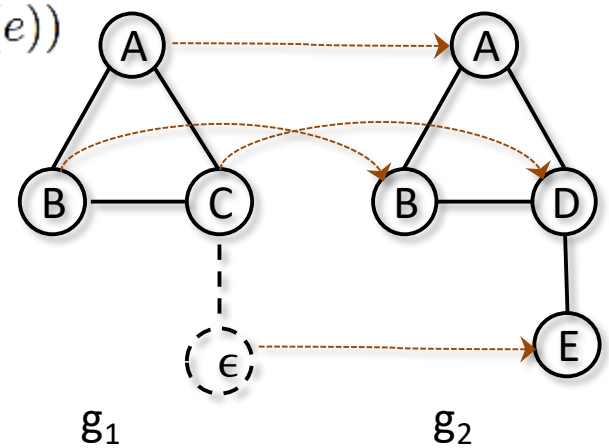
- Easily adaptable to similarity queries
  - if query $d(q, C) \geq \theta$, then $d(q, g_i) \geq \theta$, $\forall g_i \in C$
    - C is a closure
    - $d(g_1, g_2)$ is *Edit Distance* between graphs

Graph Edit Distance

- Graph mapping ϕ

- $$dist_\phi(g_1, g_2) = \sum_{v \in V_1^*} dist(v, \phi(v)) + \sum_{e \in E_1^*} dist(e, \phi(e))$$

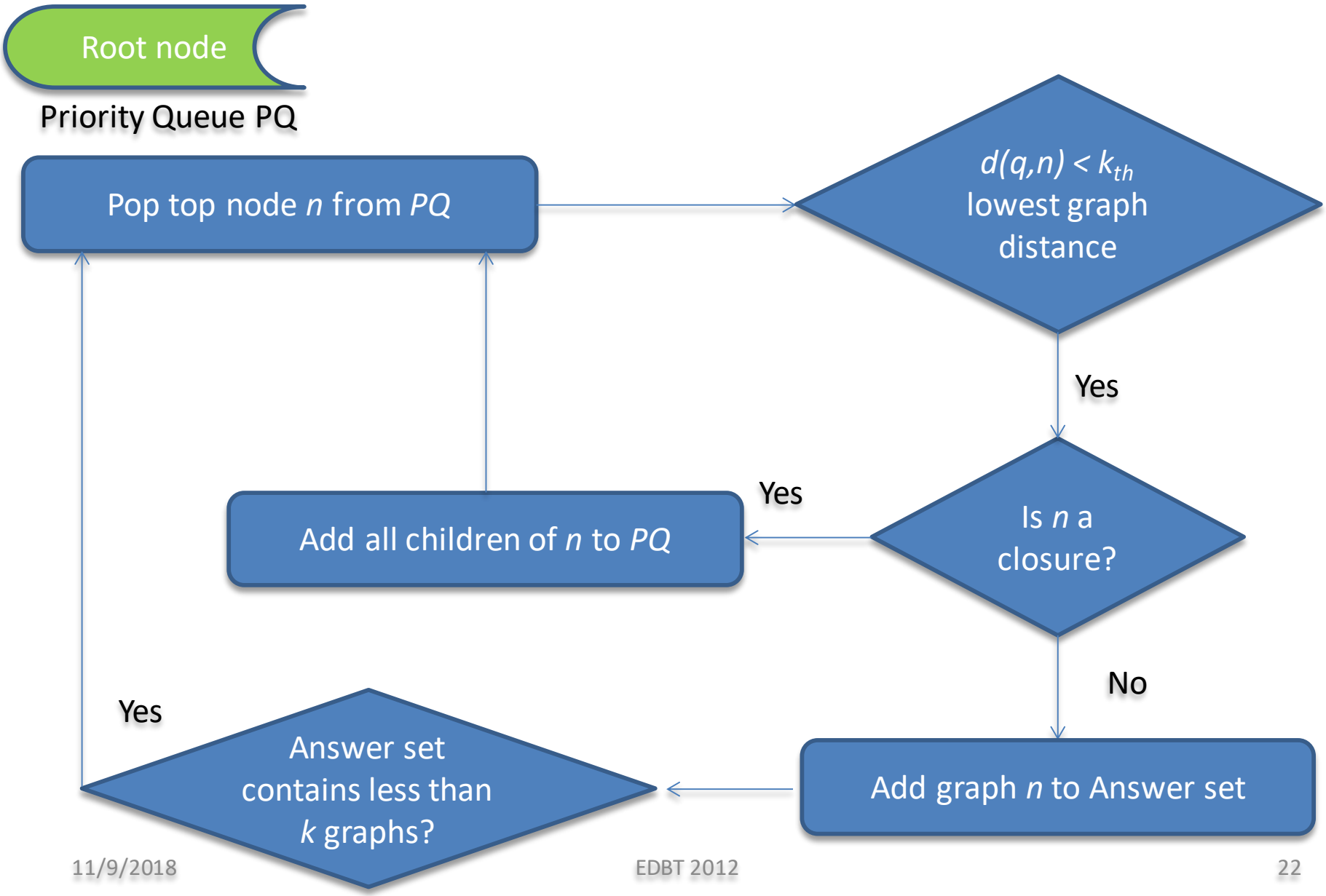
  – $d(g_1, g_2)=3$

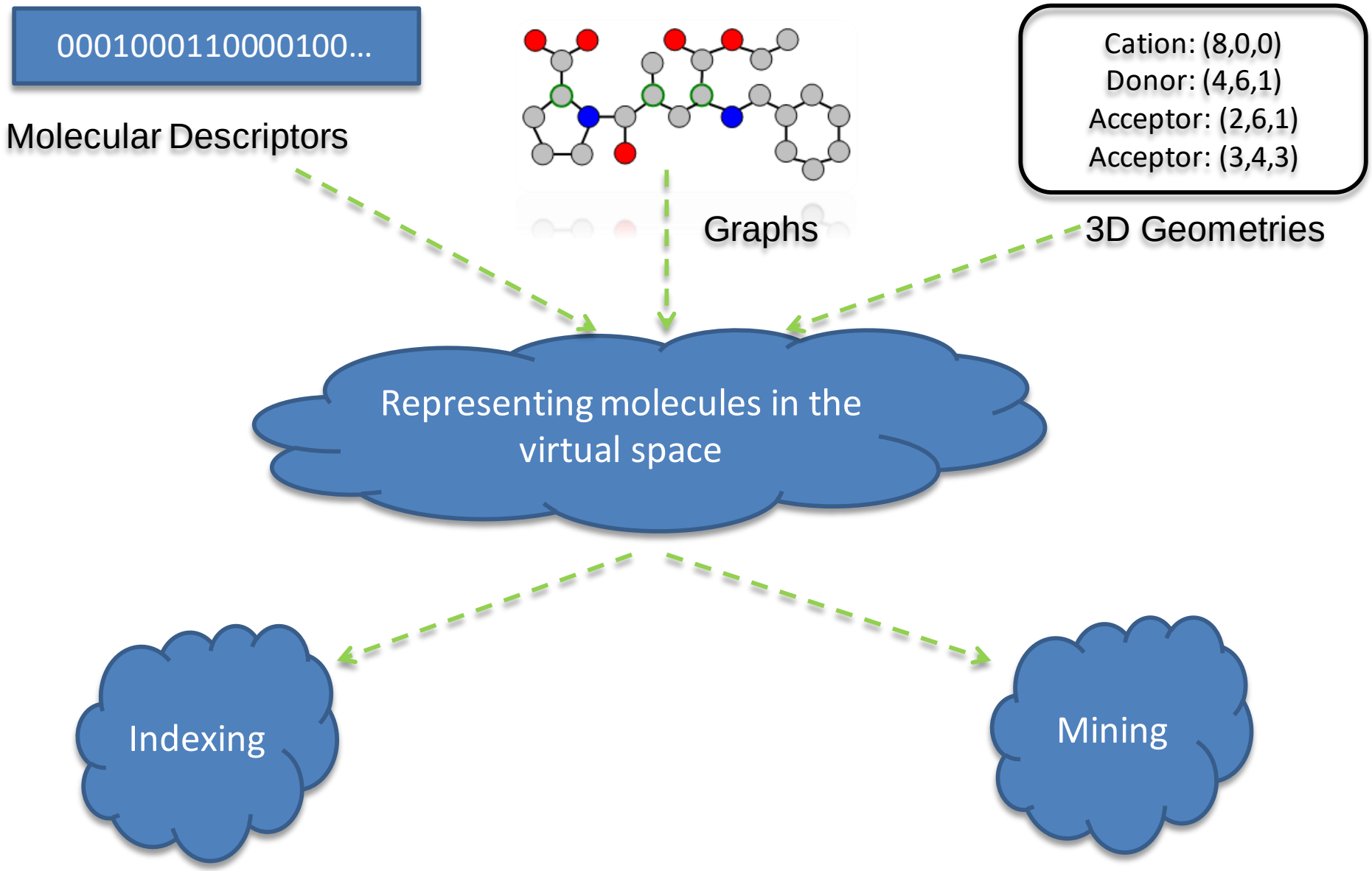


- Edit distance

  – $dist(g_1, g_2) = \min_\phi \{dist_\phi(g_1, g_2)\}$

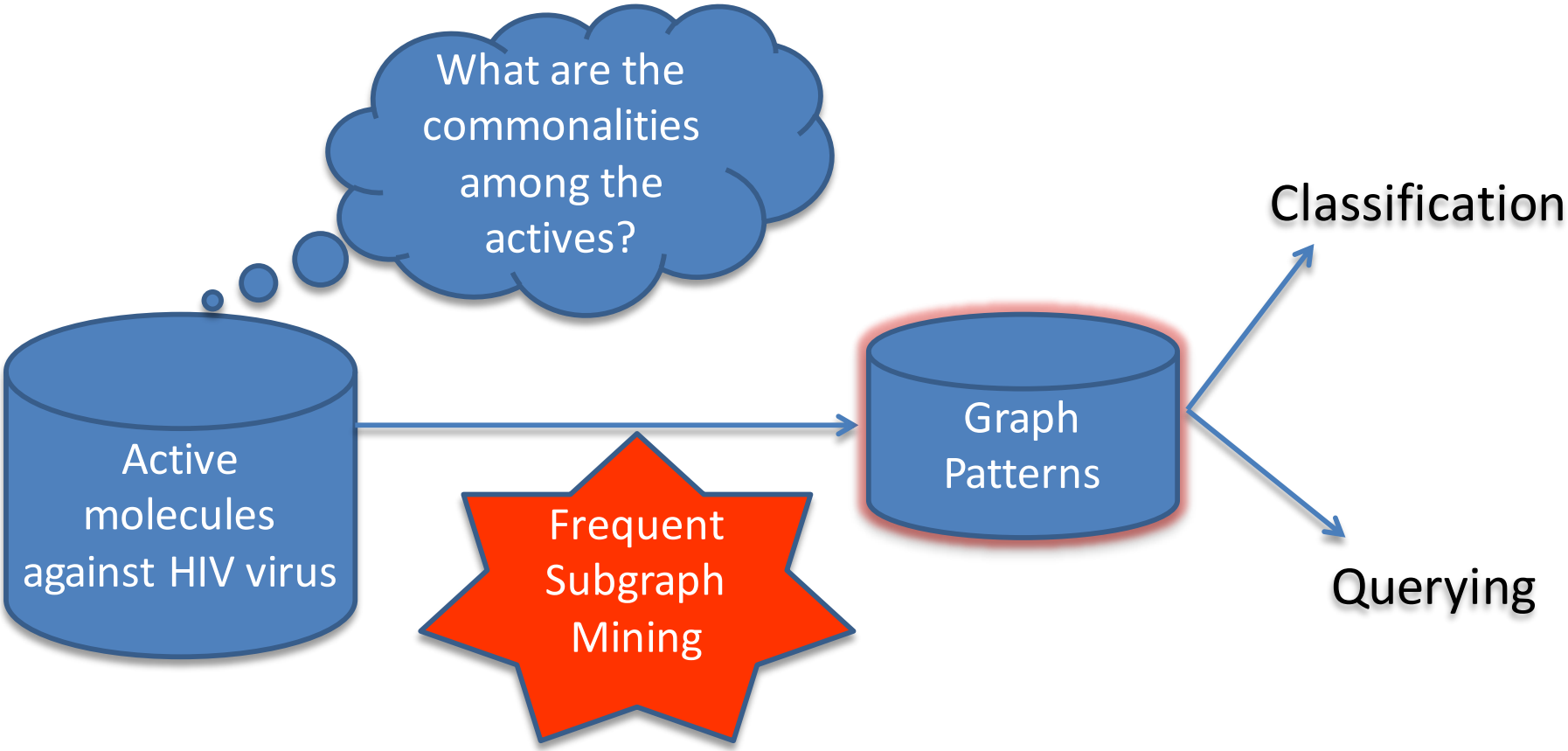
Top-*k* similarity search




0001000110000100...
Molecular Descriptors
Graphs
Cation: (8,0,0)
Donor: (4,6,1)
Acceptor: (2,6,1)
Acceptor: (3,4,3)
3D Geometries
Representing molecules in the virtual space
Indexing
Mining

Graph Pattern Mining

- Identify *hidden* characteristics of a dataset



What are graph patterns?

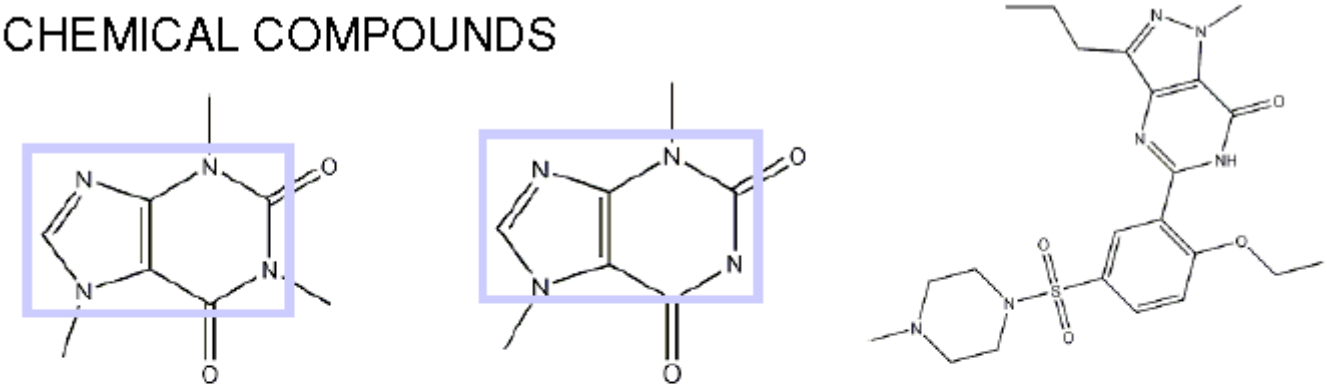
- Given a function *f(g)* and a threshold θ, find all subgraphs *g*, such that *f(g)* ≥ θ.
- Example: frequent subgraph mining.

Given a graph dataset D, find subgraph g, s.t.

$$freq(g) \geq \theta$$

where $freq(g)$ is the percentage of graphs in D that contain g.

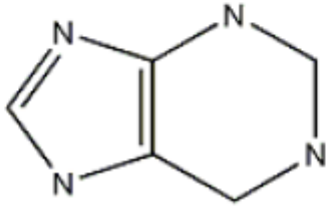
CHEMICAL COMPOUNDS



(a) caffeine (b) diurobromine (c) viagra

FREQUENT SUBGRAPH

Θ=50%



Is this the only frequent subgraph?

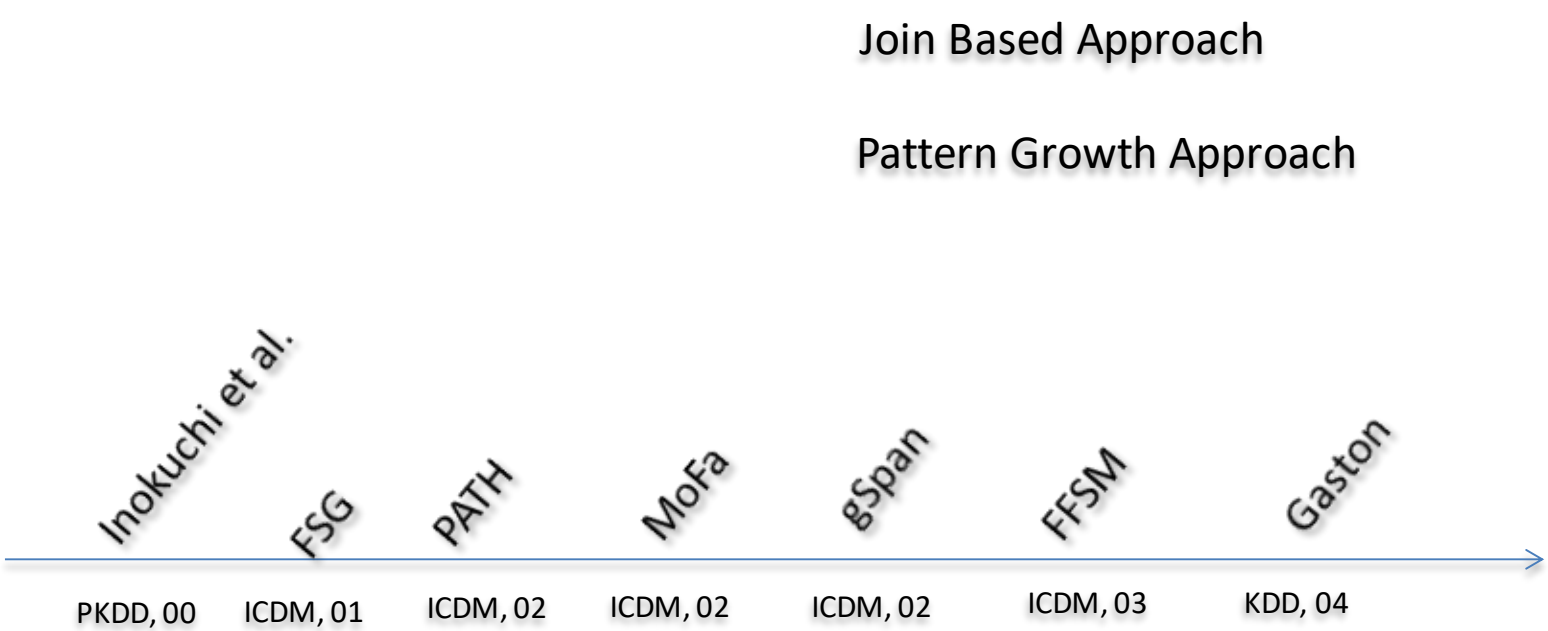
NO!

Apriori Property

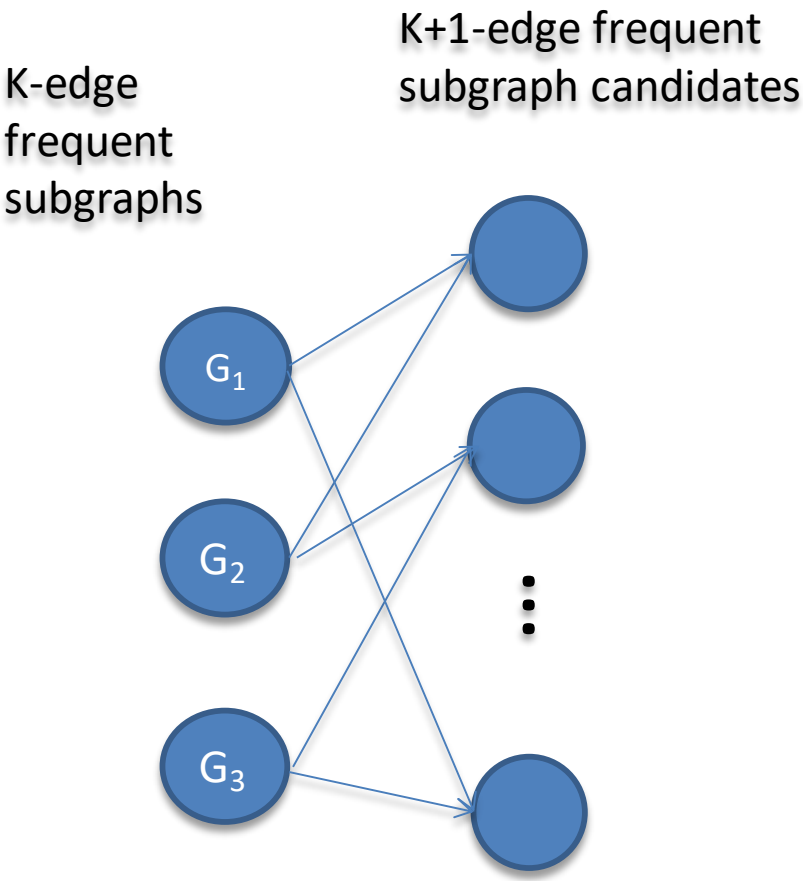
If a graph is frequent, all of its subgraphs are frequent.

Why is graph mining hard?

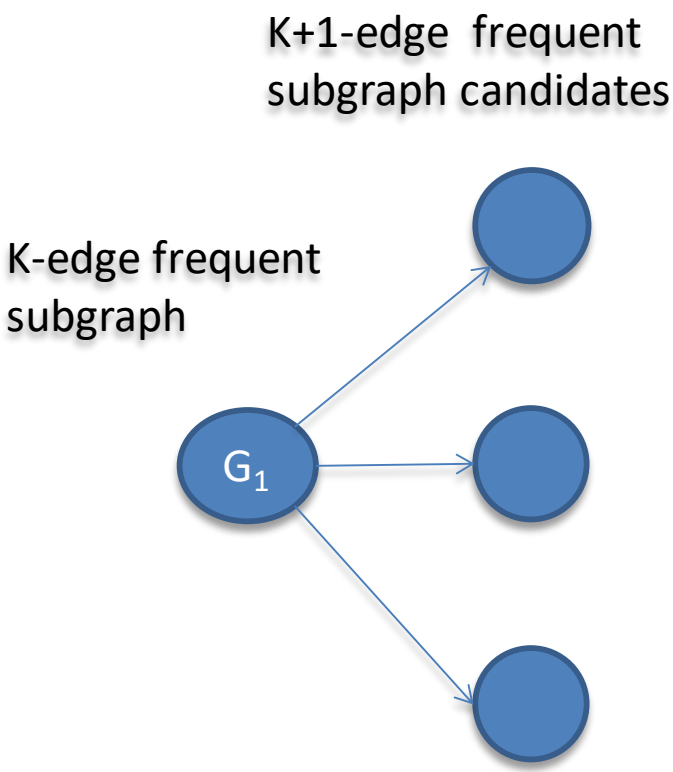
- Worst Case Scenario: A graph with n edges has 2^n subgraphs
- *Exponential* search space!



Frequent Pattern Mining Approaches

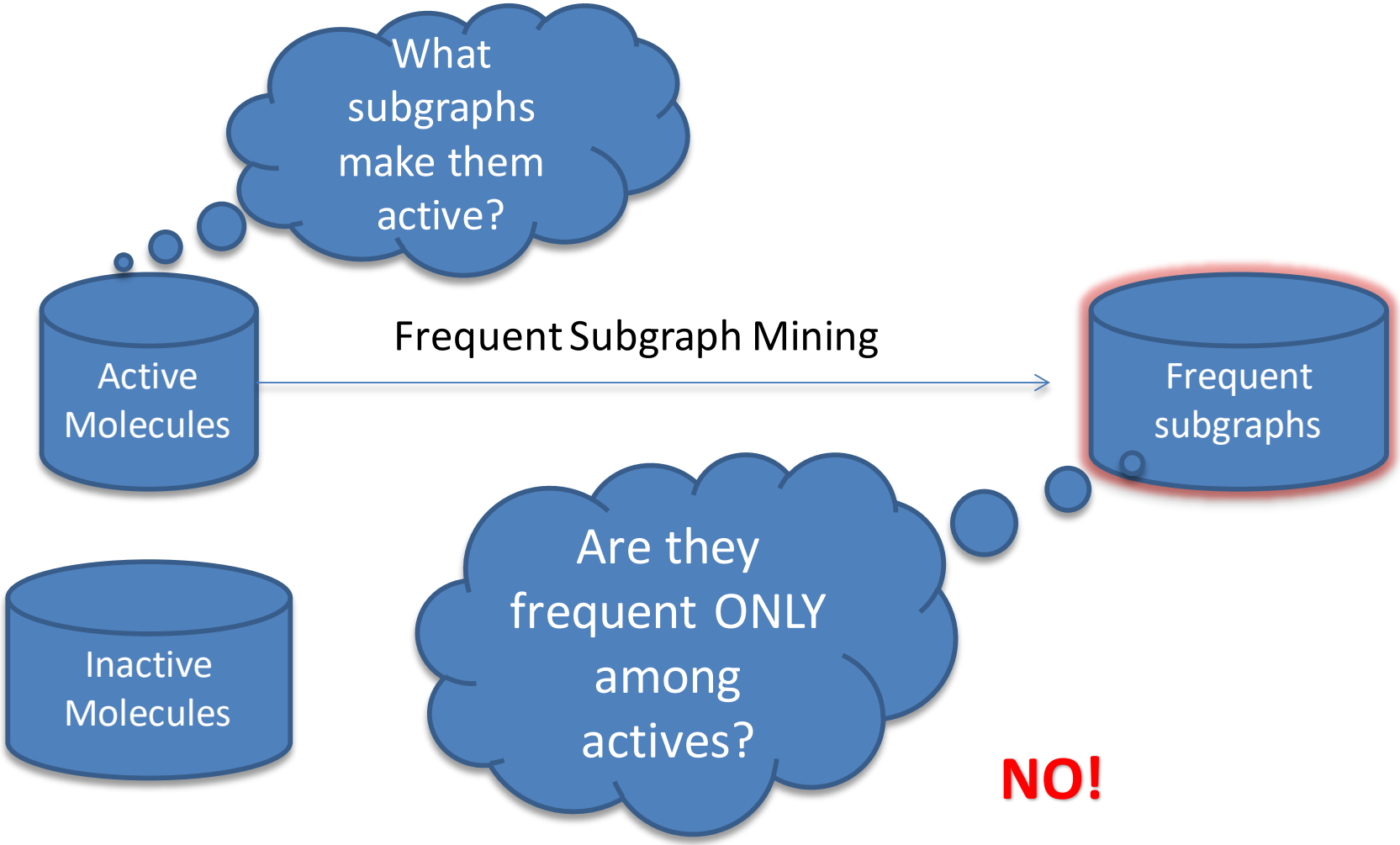


Join based approach



Pattern growth approach

Are frequent patterns enough?

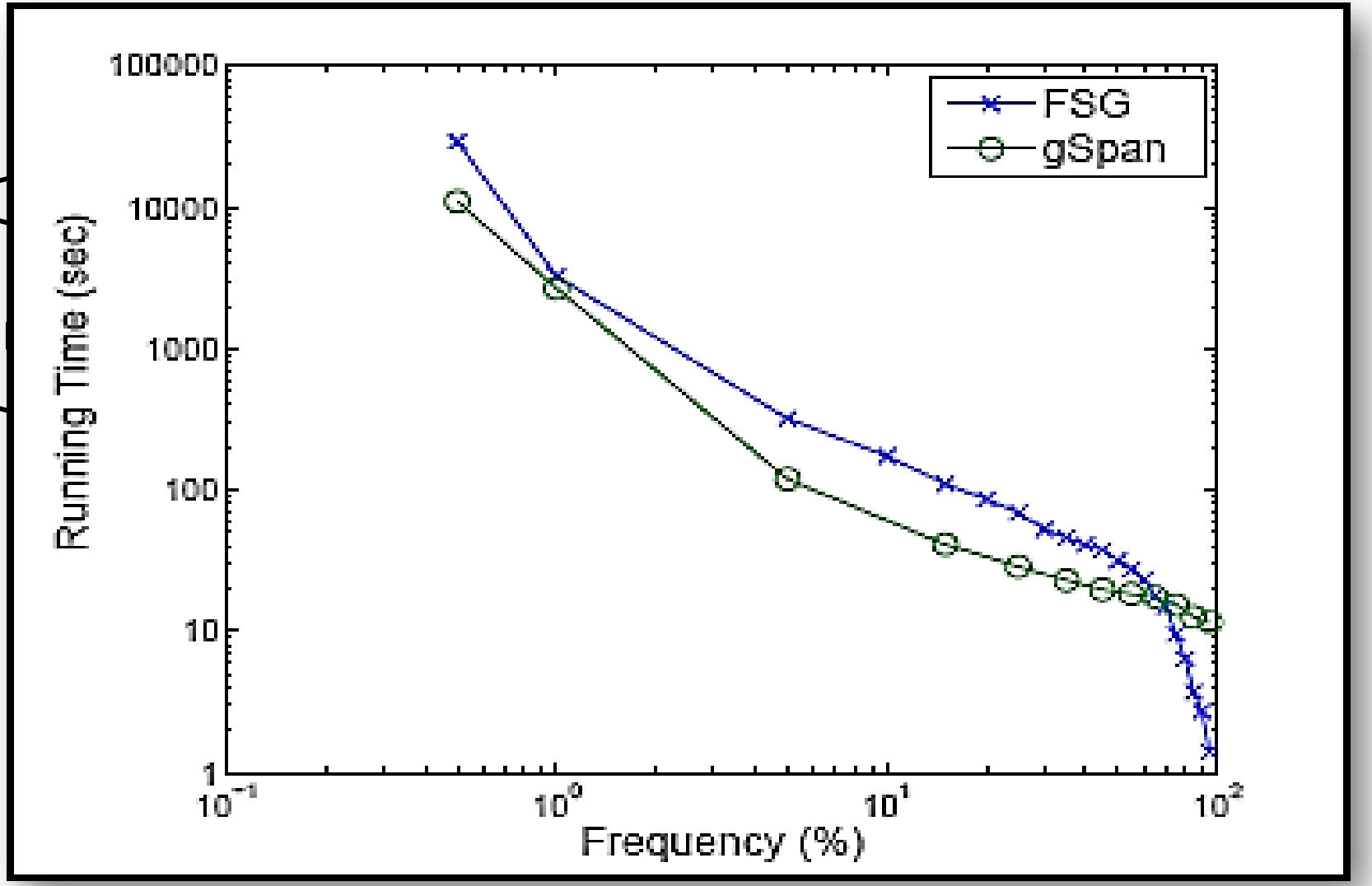


What are the *statistically significant* subgraphs?

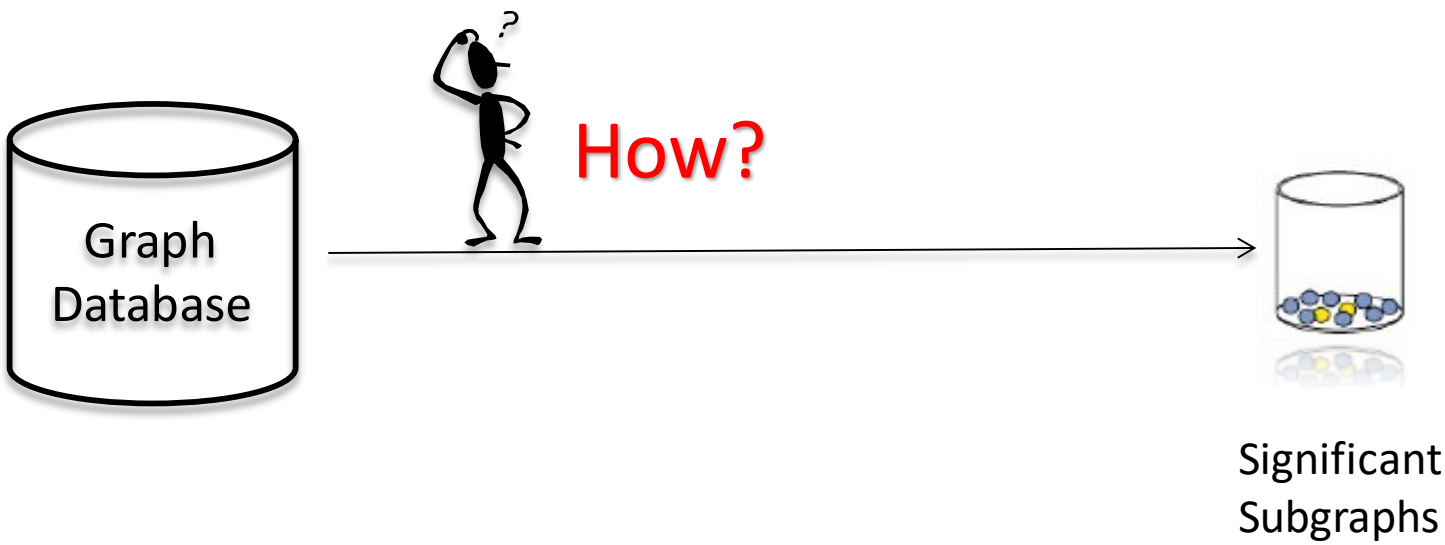
Limitation of Frequent Subgraphs

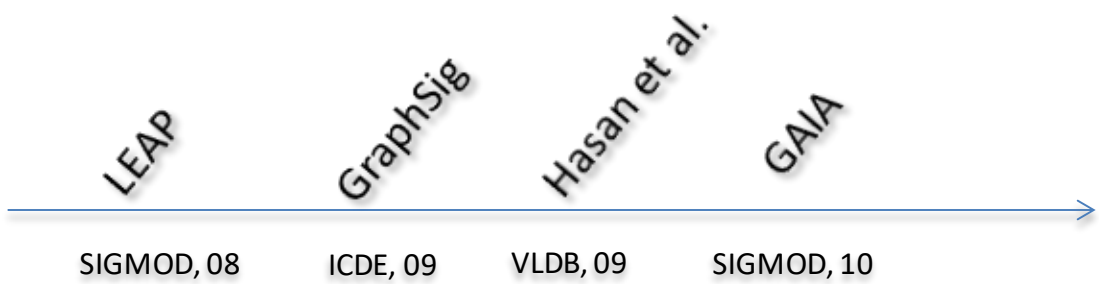
- High frequency does not imply high significance and vice versa
- A subgraph with frequency 1% can be statistically significant if the *expected frequency* is 0.1%

Naïve Approach



Direct Mining of Significant Subgraphs



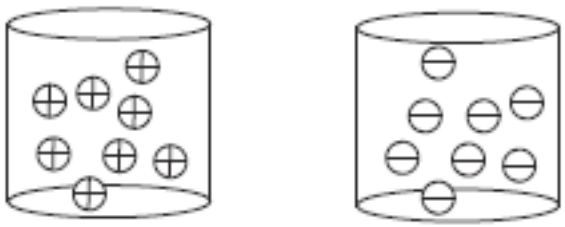


LEAP

- Uses g-test score to quantify significance

$$G_t = 2m(p \cdot ln\frac{p}{q} + (1-p) \cdot ln\frac{1-p}{1-q})$$

  - m: number of active molecules
  - p, q: frequencies in active and inactive datasets
- Find subgraphs with g-test score > Θ



Actives Inactives

g_1	p_1	q_1
g_2	p_2	q_2

If,

$p_1 > p_2$

$q_1 < q_2$

Then,

$g\text{-}test(g_1) > g\text{-}test(g_2)$

Leap: Pruning Heuristics

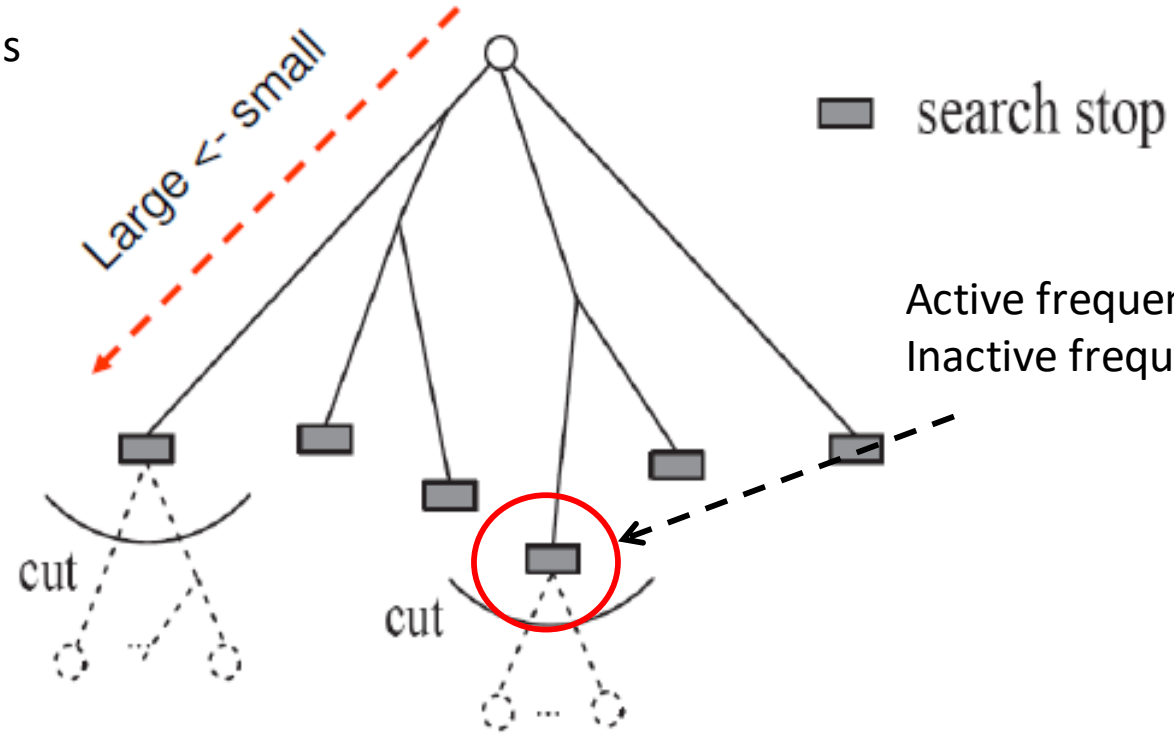
- Vertical Pruning ← *Optimal*
- Horizontal Pruning ← *Non-optimal*

Vertical Pruning

0-edge subgraphs

1-edge subgraphs

k-edge subgraphs



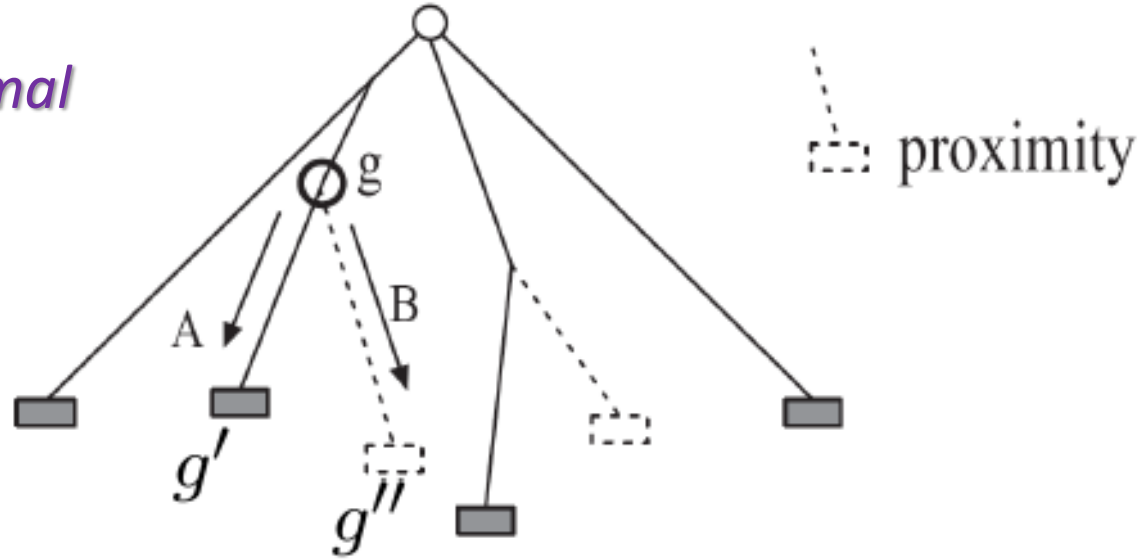
Active frequency: p
Inactive frequency: q

$$\max(F(p,\epsilon), F(\epsilon,q)) < \Theta$$

$$where, \epsilon \sim 0$$

Horizontal Pruning

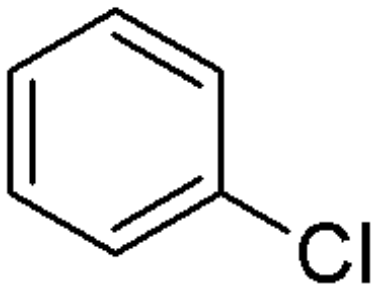
Can be *non-optimal*



$$g' \sim g'' \Rightarrow F(g') \sim F(g'').$$

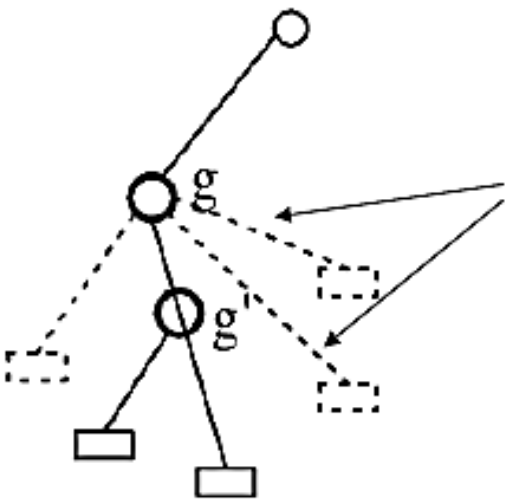
$$F(g') \ll \Theta \qquad \Rightarrow F(g'') \ll \Theta$$

Horizontal Pruning: Example



Similar Frequencies

Horizontal Pruning: Pruning Heuristic



if g' and g are close enough,
cut branches except g'.

If *freq(g)-freq(g')*$<\epsilon$ then,

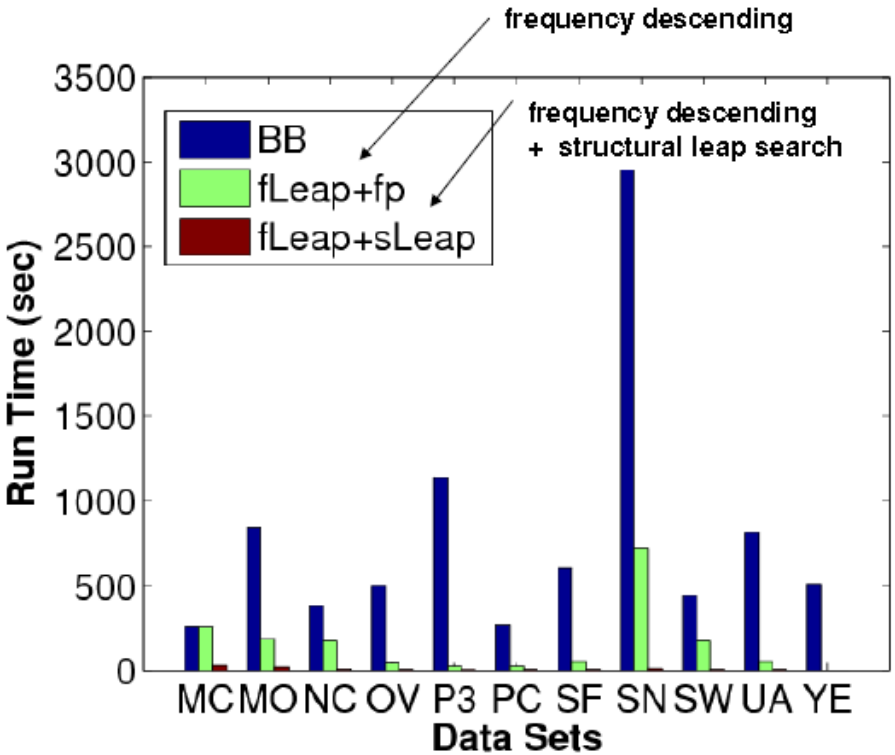
skip all *sibling branches* of g'

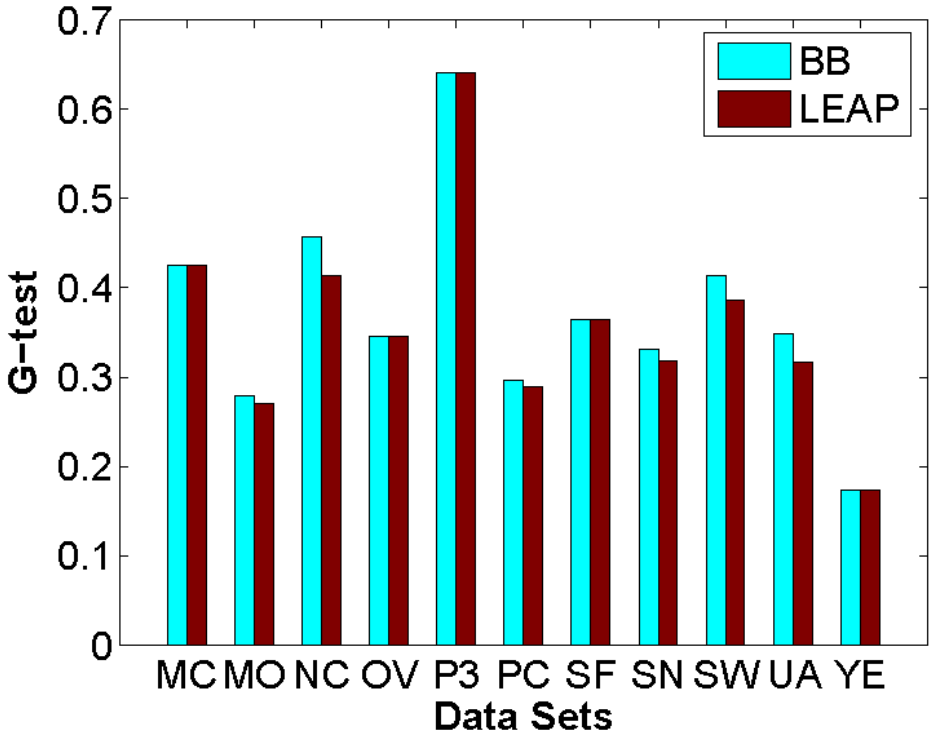
Experimental Results: Datasets

- 11 Cancer Datasets

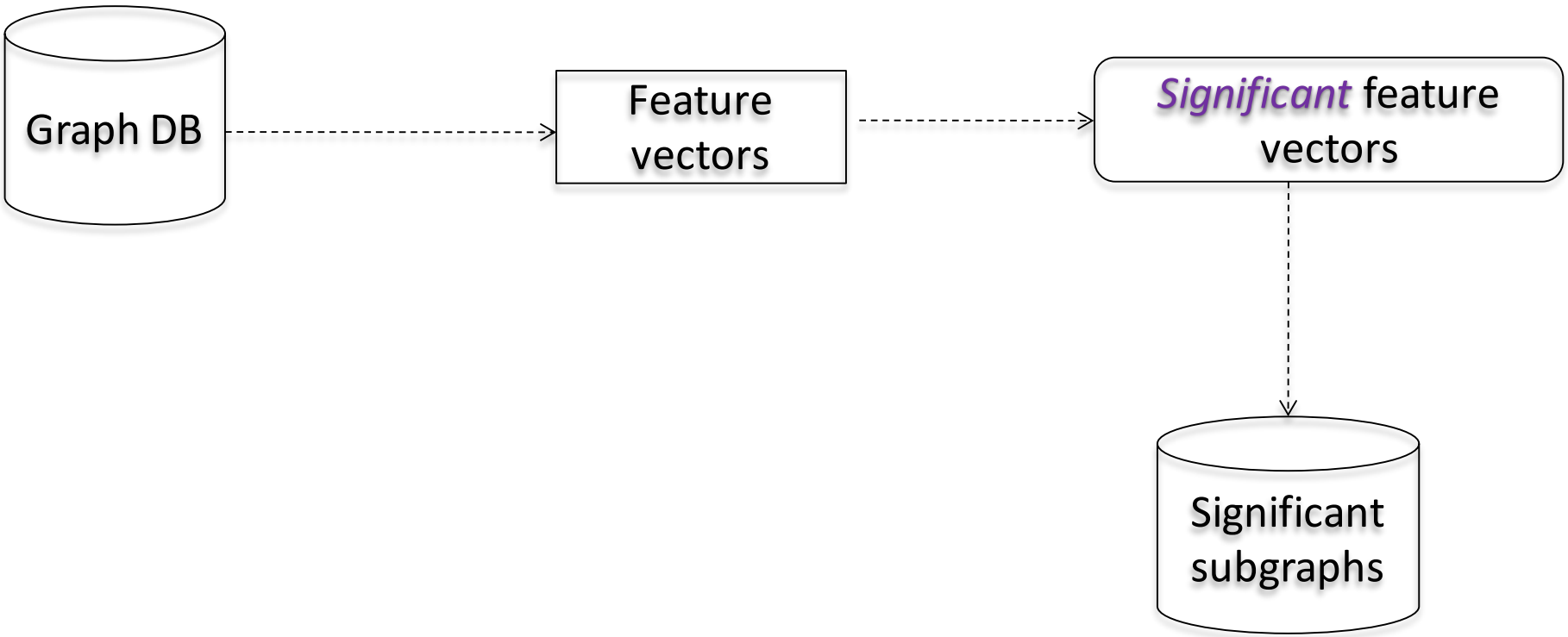
Dataset	size	# of actives	description
MCF-7	28972	1989	breast
MOLT-4	41810	3391	leukemia
NCI-H23	42164	2235	lung
OVCAR-8	42386	2255	ovarian
P388	46440	2549	leukemia
PC-3	28679	1692	prostate
SF-295	40350	1936	central nervous system
SN12C	41855	2123	renal
SW-620	42405	2623	colon
UACC-257	41864	1807	melanoma
yeast	83933	10257	yeast anticancer

Leap: Empirical Evaluation





Approach of GraphSig



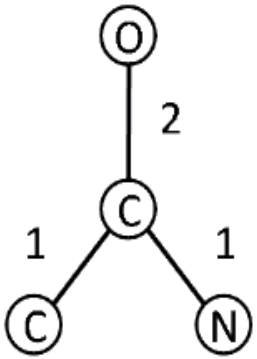
GraphSig: Problem Formulation

- Find answer set:
  - $A = \{g \mid p\text{-value}(g) < \Theta, g \subseteq G, G \in D\}$
  - D : Graph Database
  - Θ : Significance Threshold
  - $g \subseteq G$: g is a subgraph of G
- *Lower* the p-value, *higher* is the significance

Random Walk with Restarts (RWR)

- RWR on *each node* of a graph
  - Captures distribution of edge-types around each node
  - Discretized into 10 bins

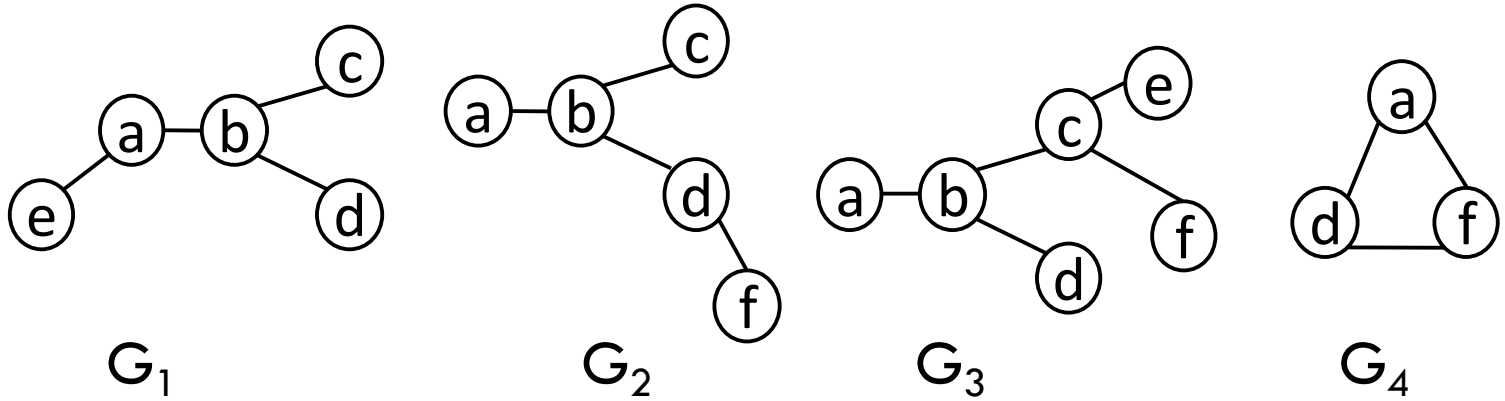
Sample Graph



Random Walk Results

ID	Starting Node	O-2-C	C-1-C	C-1-N
h_1	O	4	2	2
h2	C	2	3	3
h3	C	2	4	2
h4	N	2	2	4

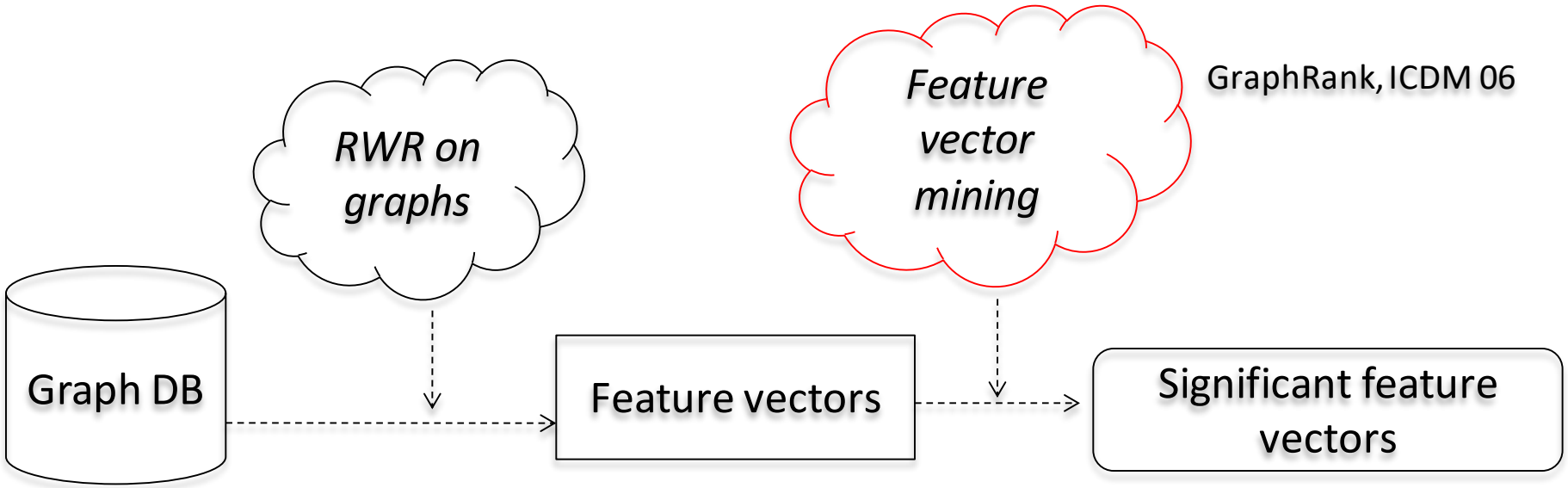
Can we capture the presence of a common subgraph?



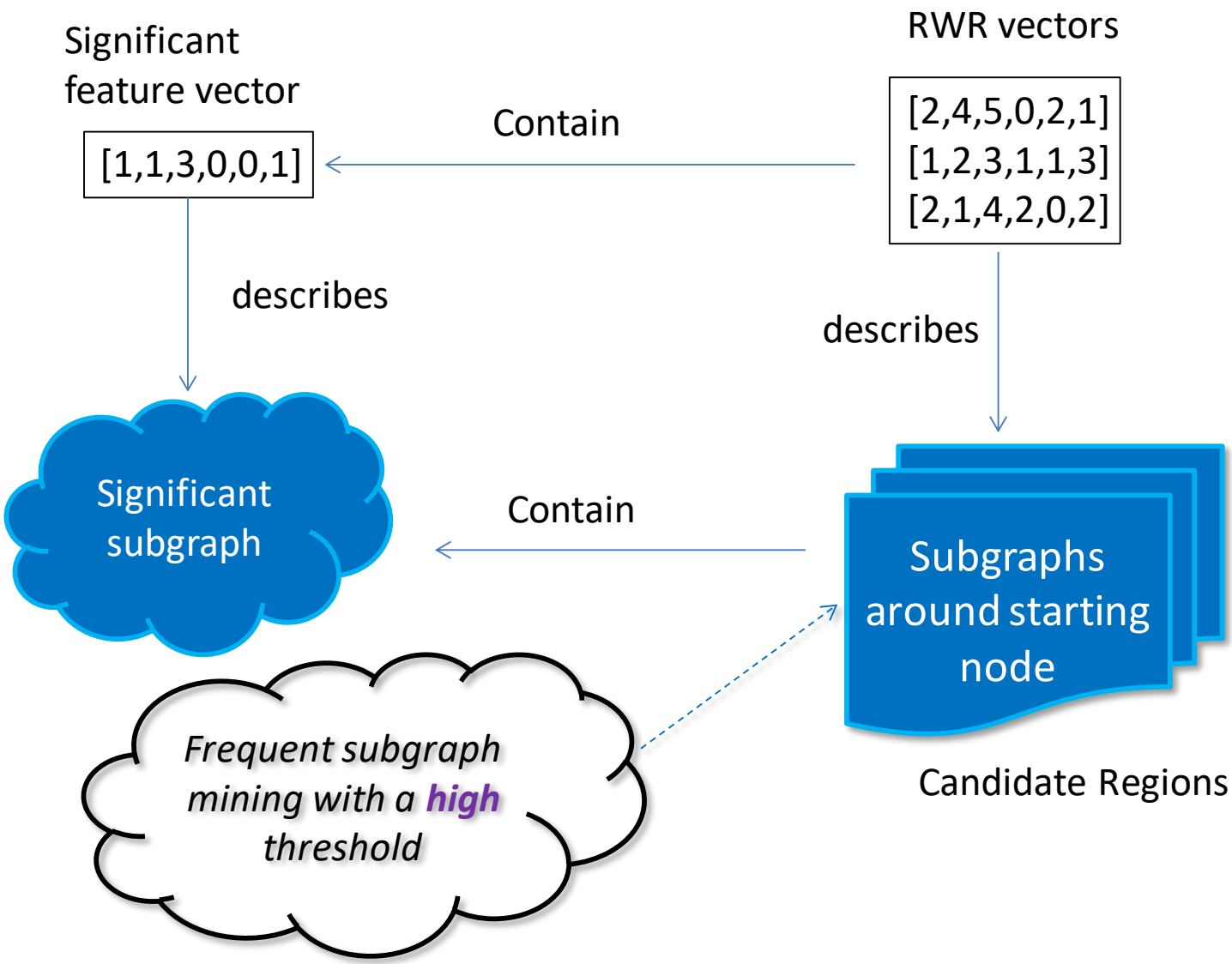
Vector	a-b	a-d	a-e	a-f	b-c	b-d	c-e	c-f	d-f
G_1	2	0	3	0	1	1	0	0	0
G_2	4	0	0	0	2	1	0	0	1
G_3	3	0	0	0	1	2	1	1	0
G_4	0	3	0	3	0	0	0	0	2

- *Floor* of G_1,G_2,G_3: [2,0,0,0,1,1,0,0,0]
- *Floor* of G_1,G_2,G_3,G_4 : [0,0,0,0,0,0,0,0,0]
- Can we measure the significance of the floors?

GraphSig Flowchart

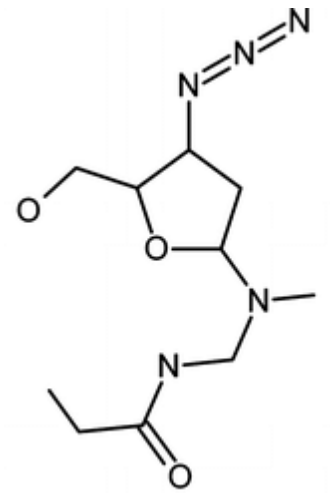


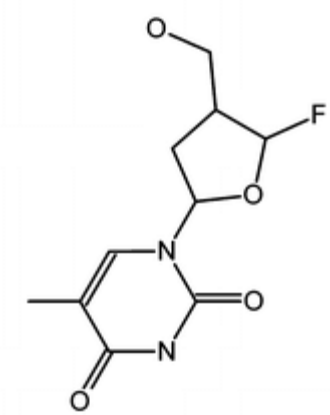
Mapping Significant Vector to Significant Subgraph



Quality of Patterns: AIDS dataset

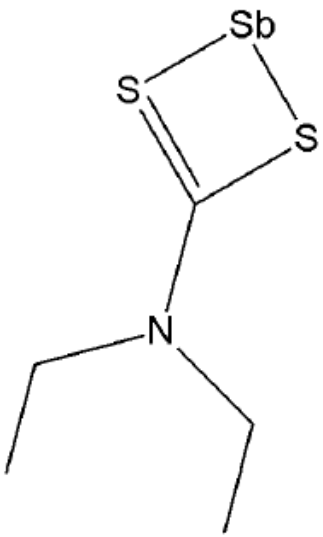
- Substructure of AZT
  - *most widely adopted medicine* to control the HIV virus
- Substructure of FDT
  - fluorinated analog of AZT. It is more active against AIDS than AZT
  - also displays a higher level of toxicity

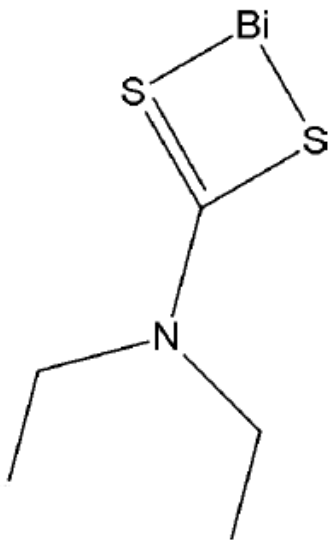




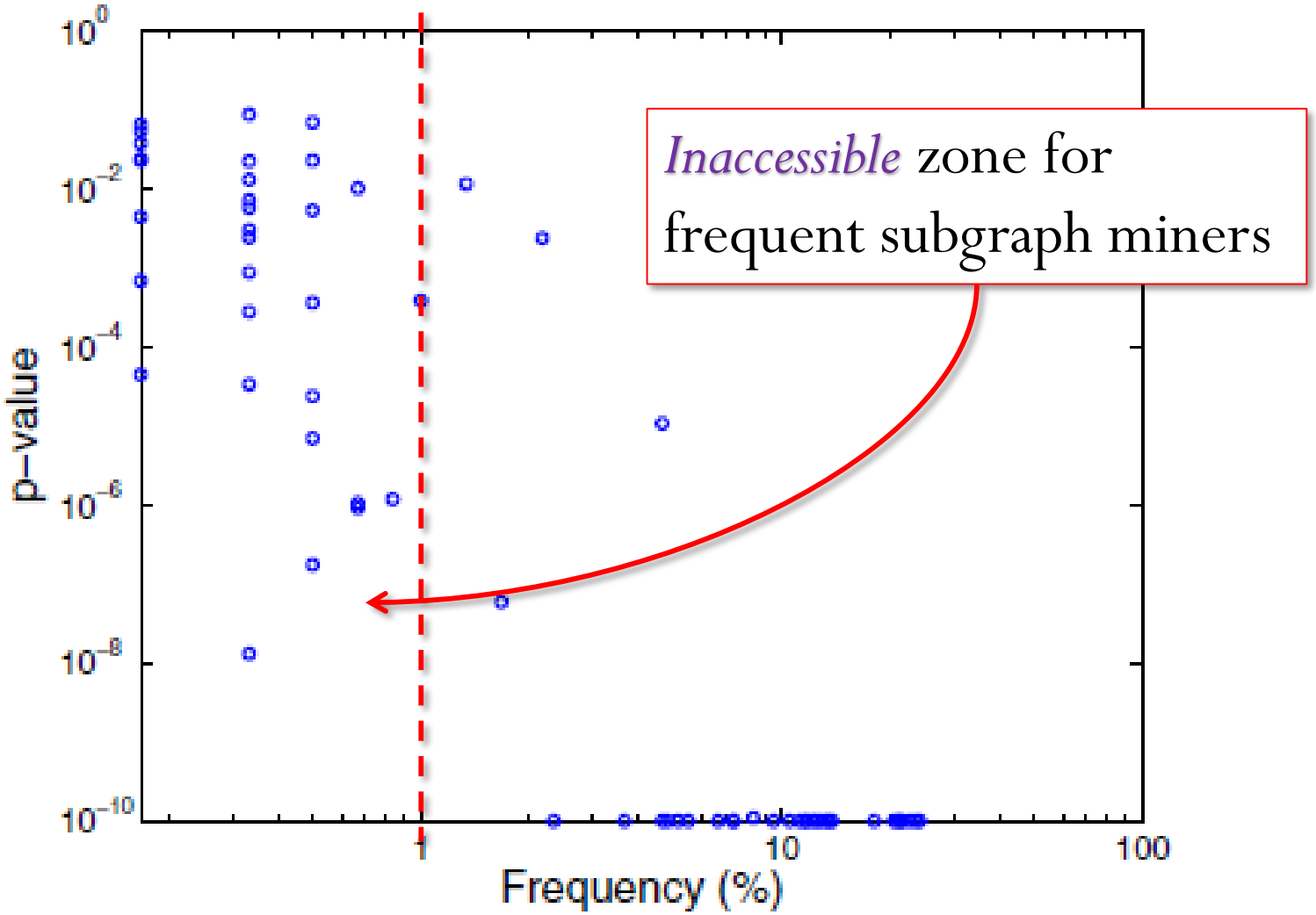
Quality of Patterns

- *Benzene* was *not* reported as *significant*
- Subgraphs mined from molecules active against Leukemia
  - Sb and Bi are found at a frequency *below 1%*
  - Frequent subgraph miners are unable to scale to such low frequencies



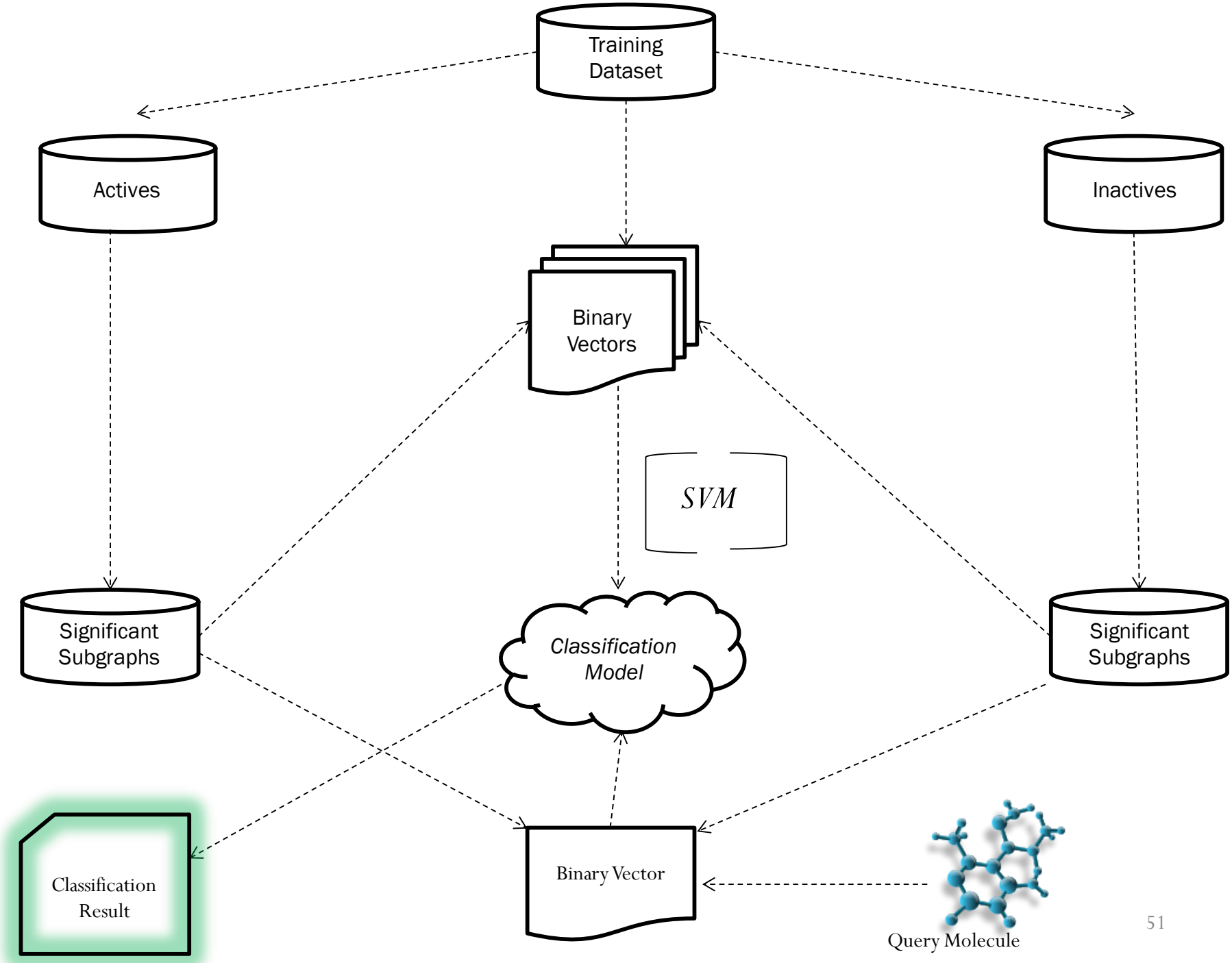


Distribution of Significant Subgraphs

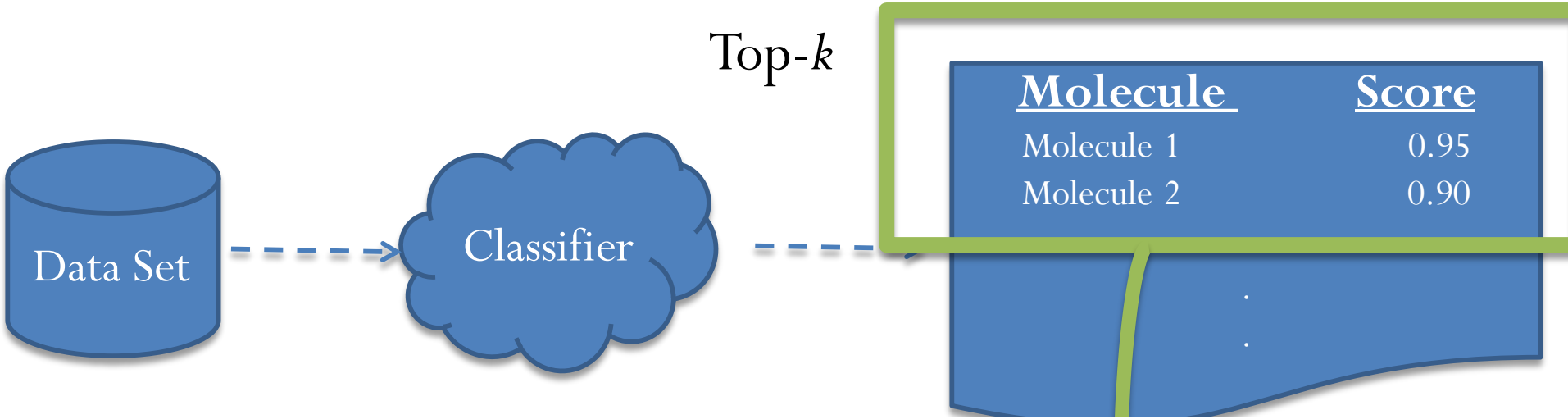


Significant Subgraph Mining: Summary

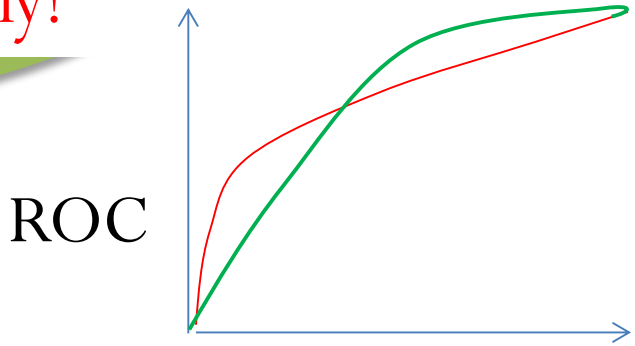
- Leap and GraphSig *overcome* the *scalability bottleneck* of frequent subgraph mining techniques
- Significant subgraphs *correlate* with *biological activity*
  - Provides excellent platform for *molecular classification*


Training Dataset
Actives
Inactives
Binary Vectors
SVM
Significant Subgraphs
Classification Model
Significant Subgraphs
Classification Result
Binary Vector
Query Molecule

BEDROC Metric [JCIM, 2007]



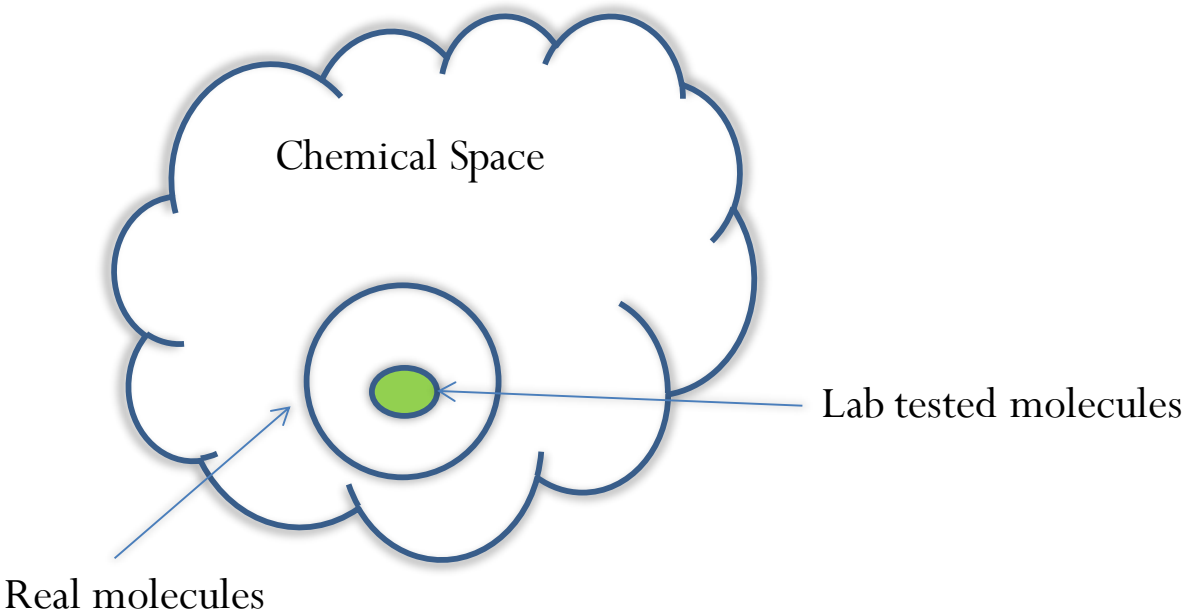
Early part of the ranked list is more important. Weigh the early part exponentially!



Performance Comparison on BEDROC

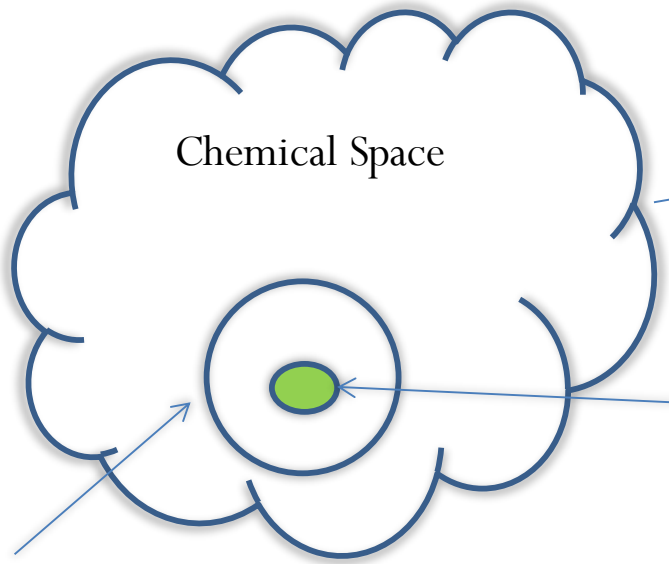
Data set	Daylight	GraphSig
MCF-7	0.41	**0.61**
MOLT-4	0.42	**0.45**
NCI-H23	0.44	**0.63**
OVCAR-8	0.40	**0.65**
P388	0.50	**0.55**
PC-3	0.33	**0.62**
SF-295	0.32	**0.63**
SN12C	0.40	**0.62**
SW-620	0.36	**0.60**
UACC-257	0.34	**0.65**
Yeast	**0.38**	**0.38**
average	0.39	**0.57**

How applicable are the techniques?



Can the rest of the molecules be used to improve our knowledge of significant subgraphs?

How to estimate the activity of the unlabeled molecules?



Molecule	P(active)
Molecule 1	0.05
Molecule 2	0.90
.	
.	
.	

pGraphSig: GraphSig on probabilistically labeled data [Molecular Informatics, 2011]

Molecule	P(active)
Molecule 1	0.05
Molecule 2	0.90
.	
.	
.	

Significant substructure mining →

Answer Set

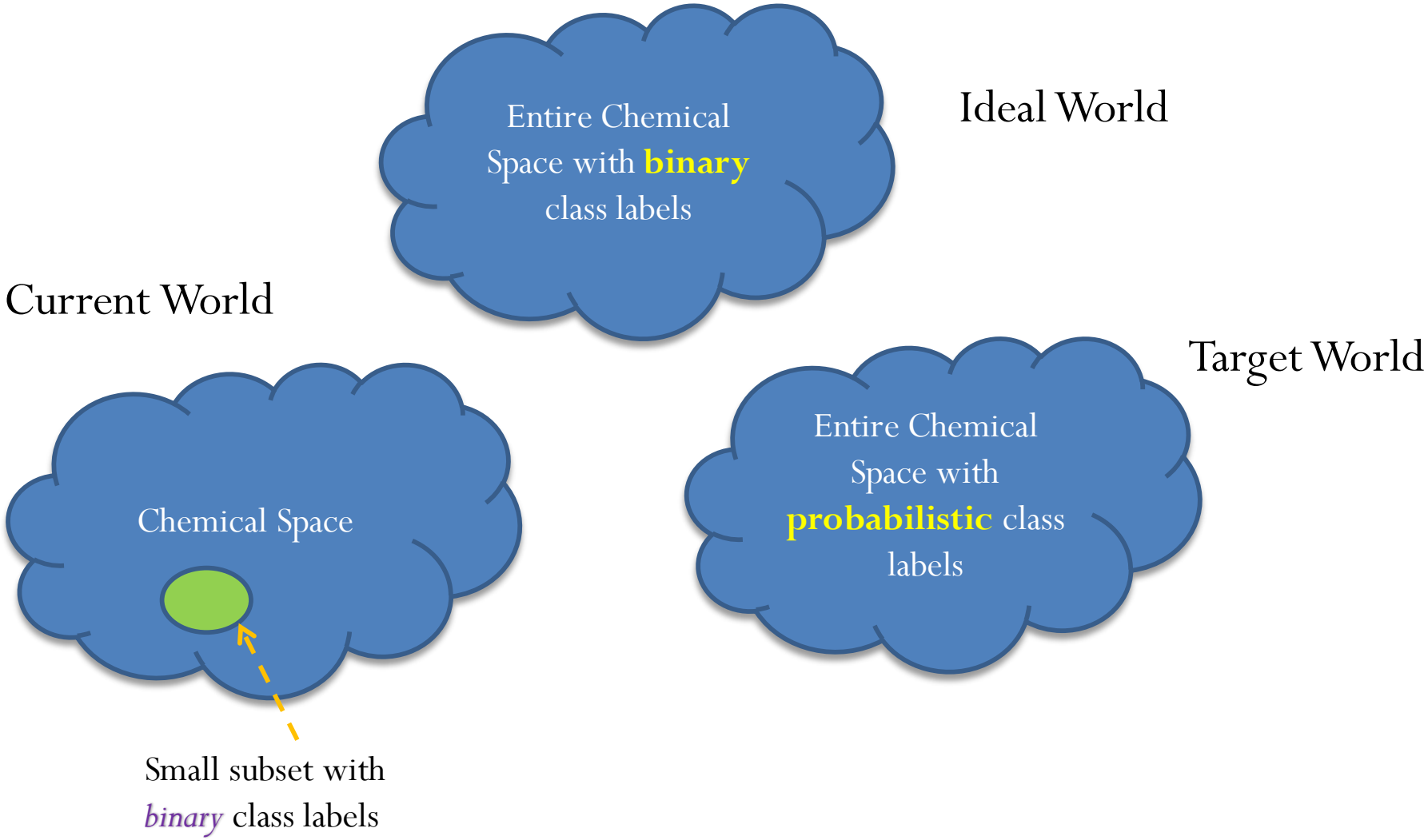
- How to compute the frequency of a subgraph under probabilistic class labels ?

Estimated Support

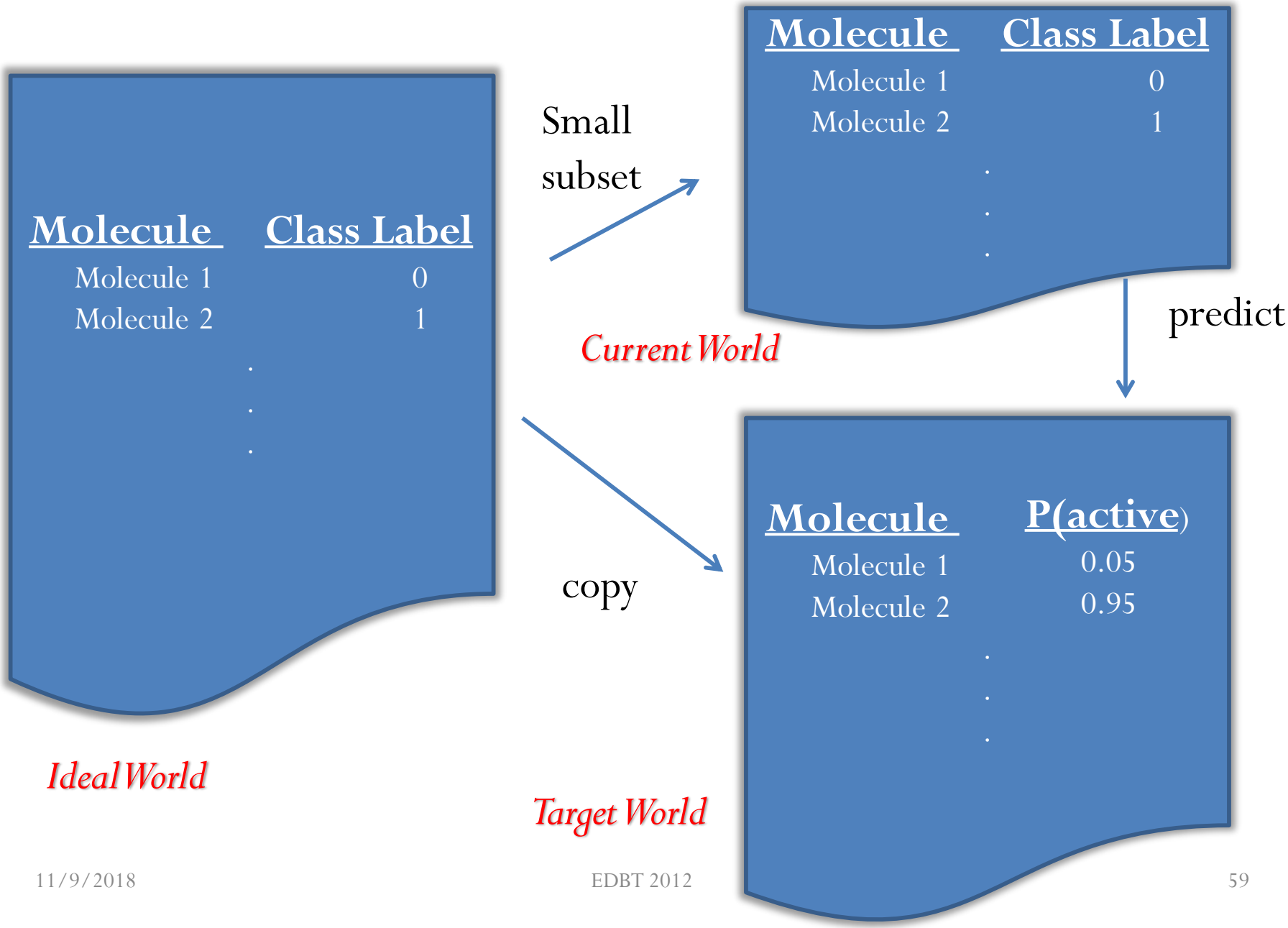
- Due to probabilistic class labels, we can only *estimate* the support

- Support(x)$=\sum P(g)$, $\forall_g \in \mathbb{S}$
  - x is a subgraph
  - $\mathbb{S} = \{g \mid x \in g,\ g \in \mathbb{D}\}$
  - $\mathbb{D}$: graph database

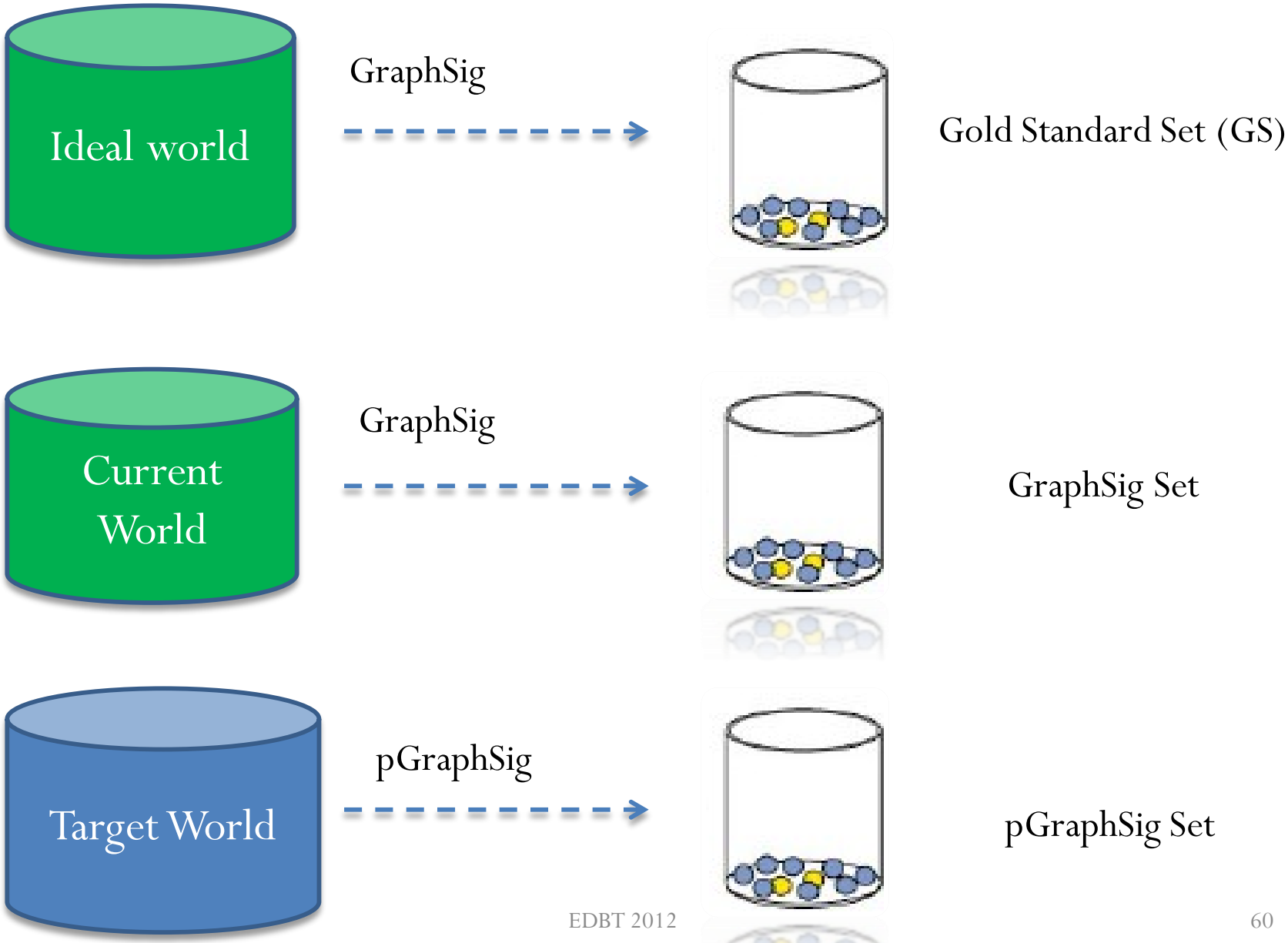
Evaluation Framework



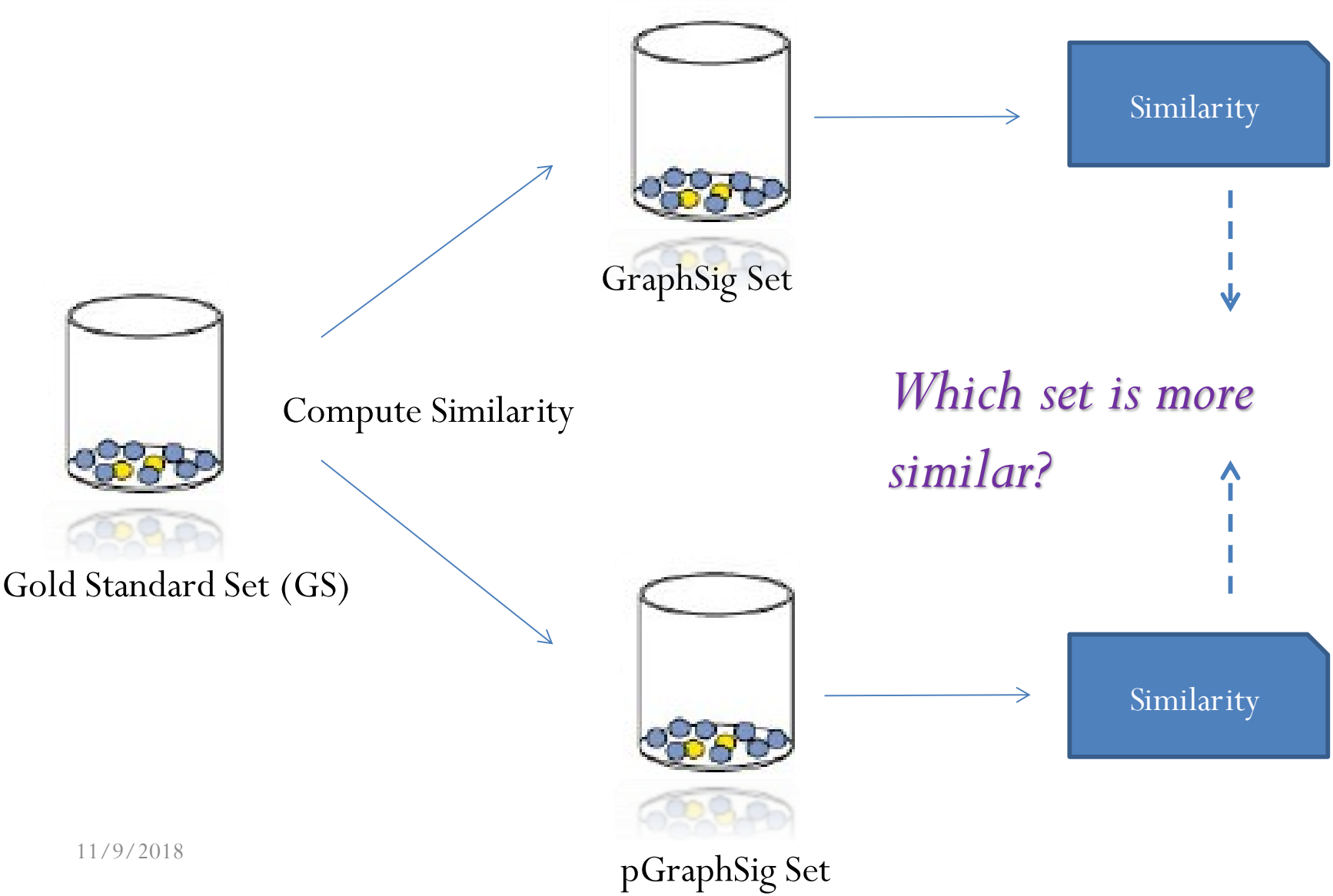
Simulation of the Three Worlds



Evaluation Sets



Comparing the Answer Sets

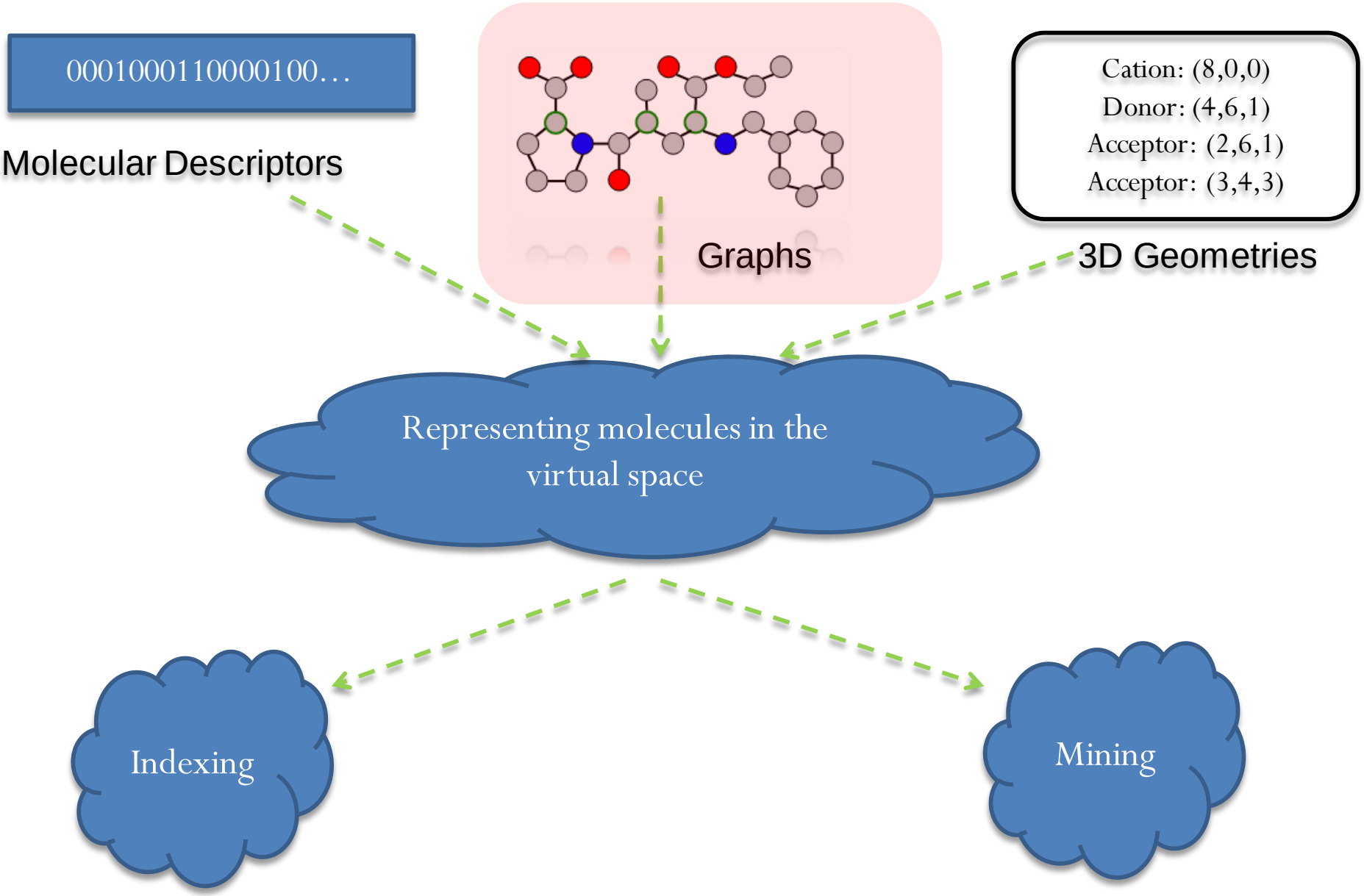


Results: Similarity with Gold Standard

Dataset	Tanimoto		Edit Distance	
	GraphSig	pGraphSig	GraphSig	pGraphSig
BAZ	0.47	**0.79**	0.27	**0.85**
HLM	0.80	**0.83**	**0.70**	0.65
JMJ	0.61	**0.72**	0.32	**0.54**
TDP	0.69	**0.77**	0.67	**0.87**
Average	0.64	**0.77**	0.49	**0.72**

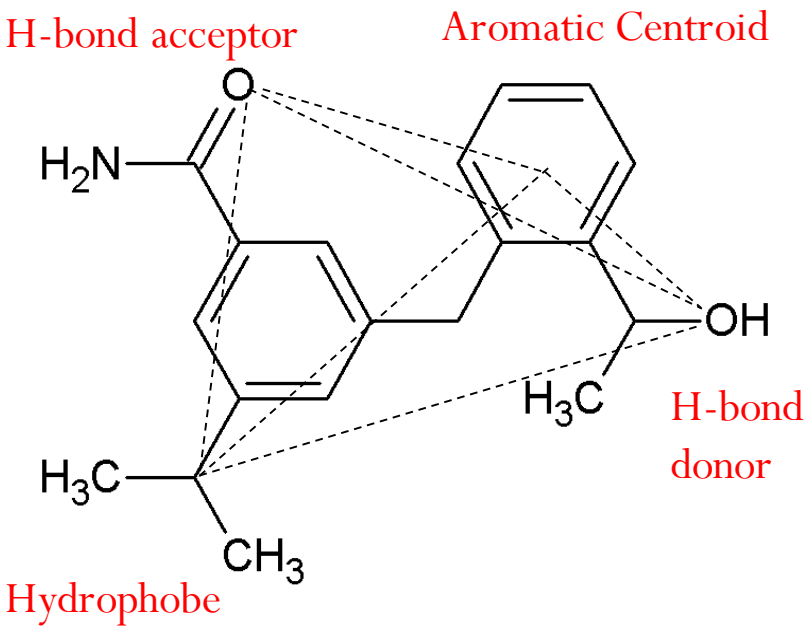
pGraphSig: Summary

- Addition of probabilistic information *expands* our *knowledge base*
- Ability to handle probabilistically labeled data significantly *increases* the *applicability* of pGraphSig


0001000110000100…
Molecular Descriptors
Graphs
Cation: (8,0,0)
Donor: (4,6,1)
Acceptor: (2,6,1)
Acceptor: (3,4,3)
3D Geometries
Representing molecules in the virtual space
Indexing
Mining

Geometry Based Representation

- Pharmacophore: based on modeling the *interactions* between a small molecule and protein target
- Higher level labeling of atoms

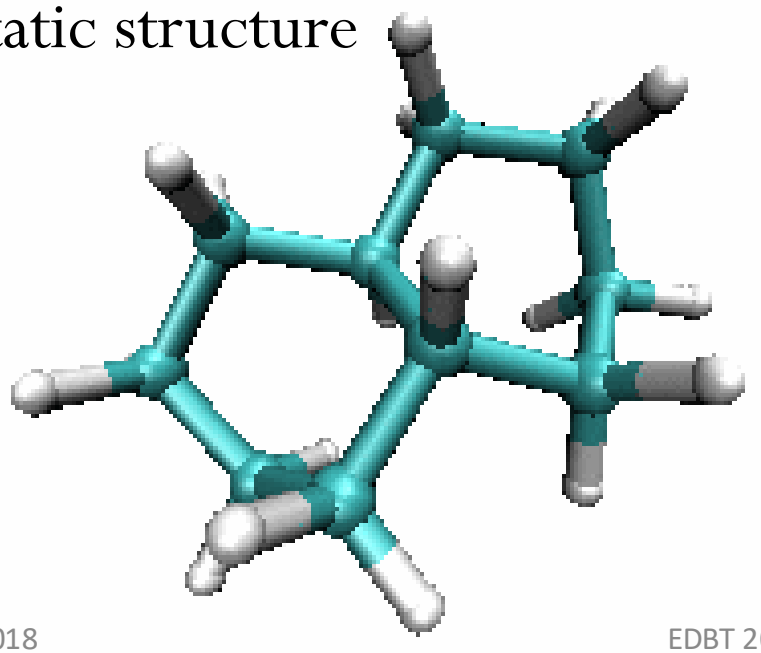


acceptor: $[x_1, y_1, z_1]$
aromatic: $[x_2, y_2, z_2]$
donor: $[x_3, y_3, z_3]$
hydrophobe: $[x_4, y_4, z_4]$

Graph Vs. 3D Geometry

Graph

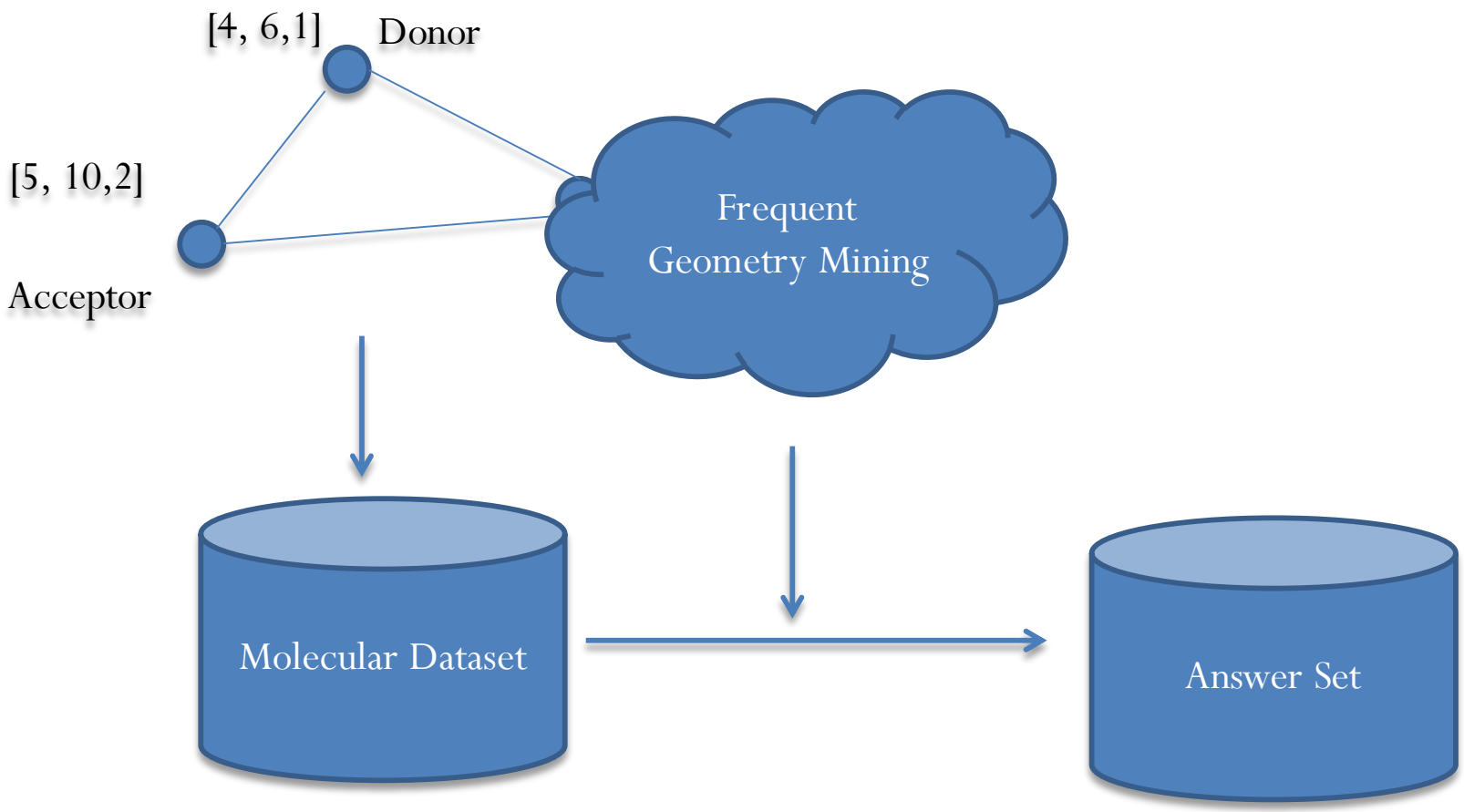
1. Atoms
2. Well defined edges
3. Static structure



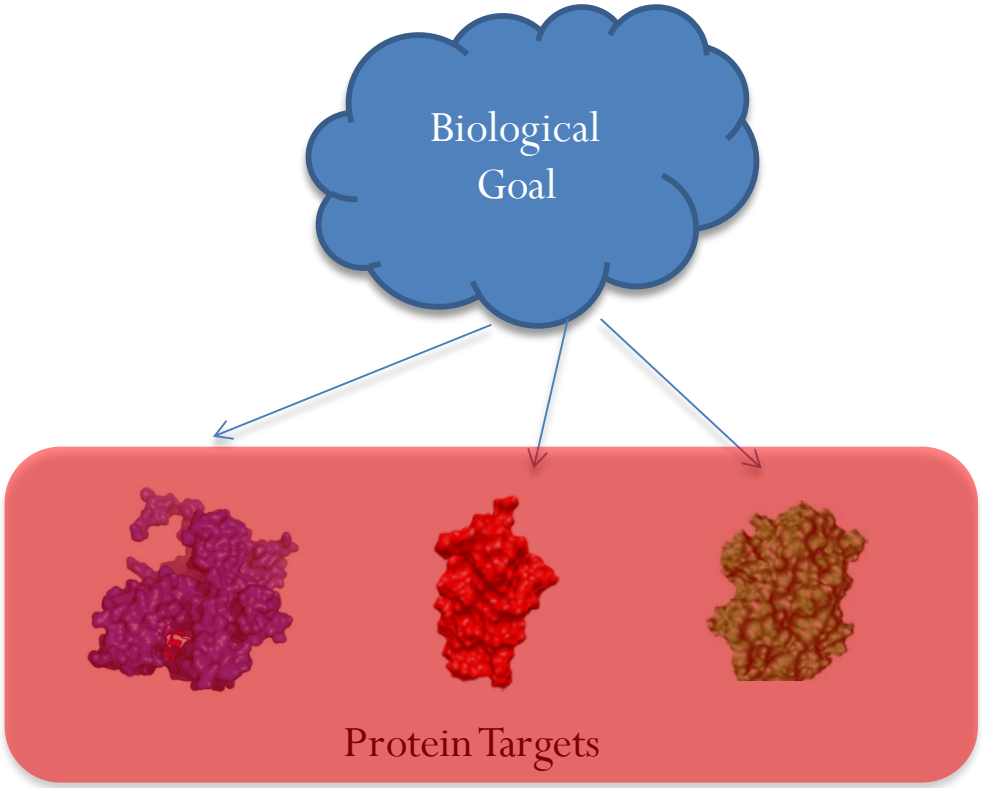
3D Geometry

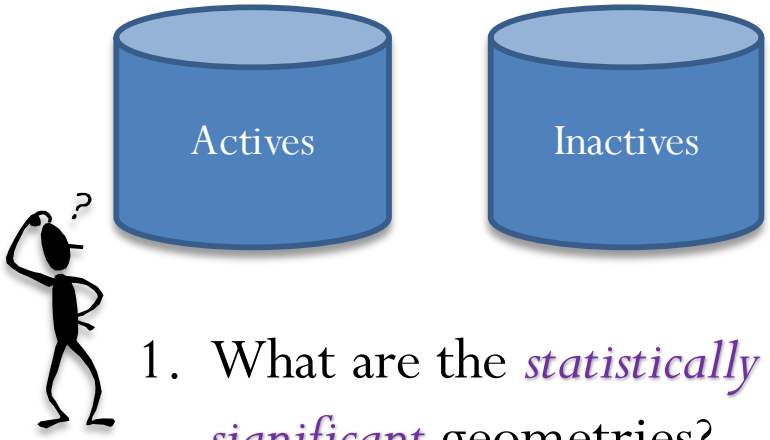
1. Pharmacophores
2. No edges
3. Dynamic structure
   - Multiple *conformations* per molecule

Analysis of Geometric Patterns



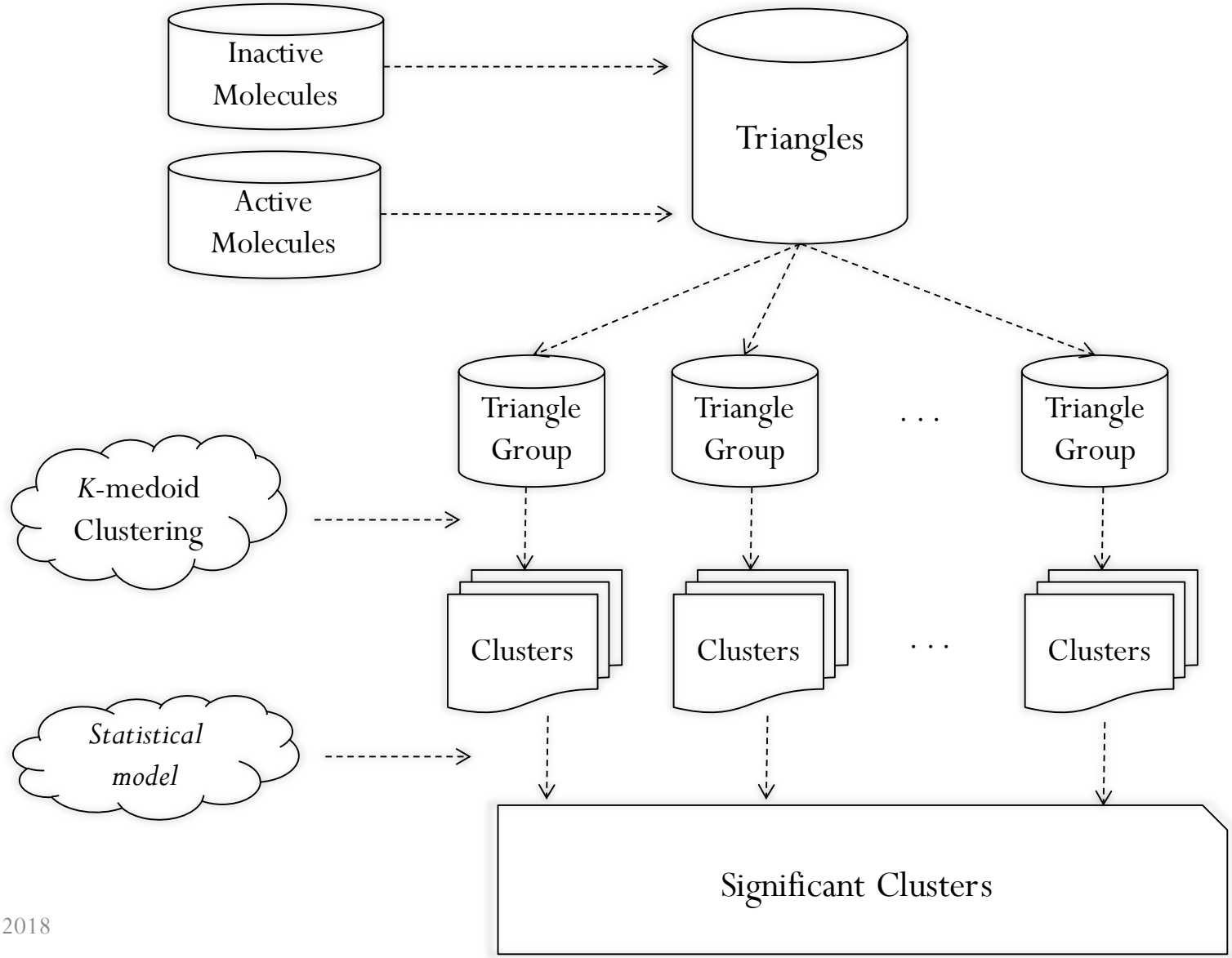
Mining Statistically Significant Geometries[JCIM 2011]



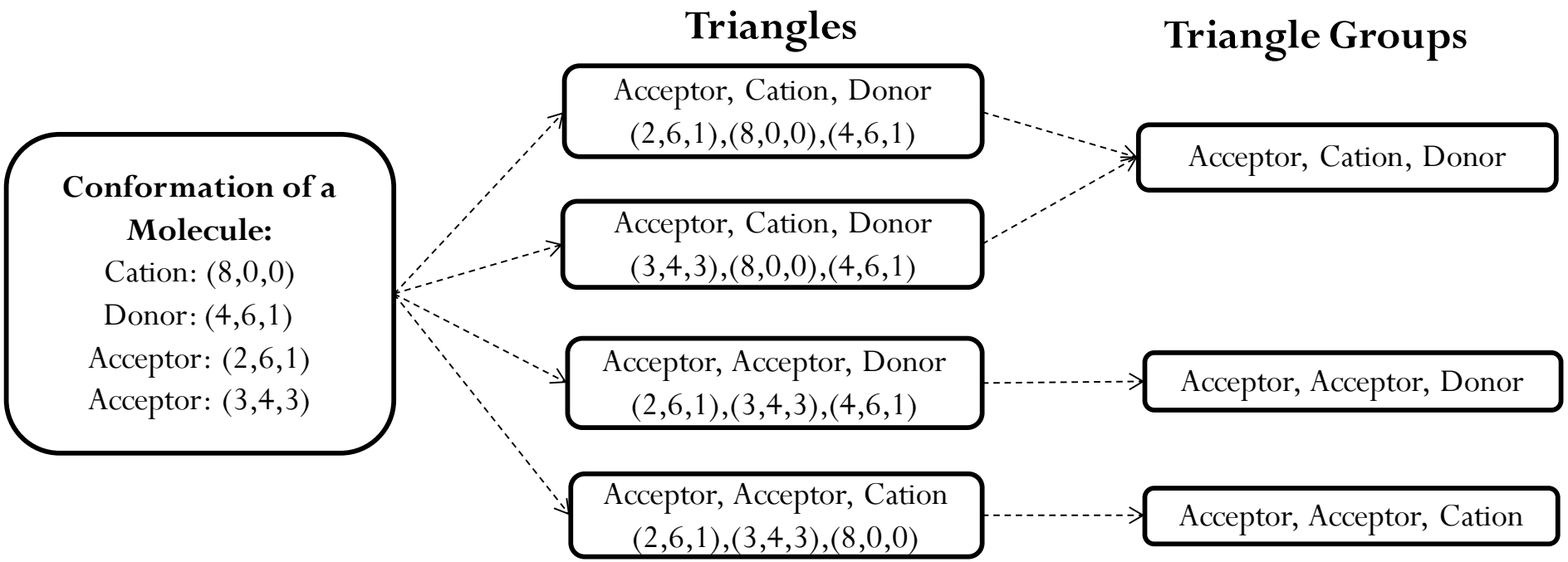


1. What are the *statistically significant* geometries?
2. Can we *divide* the actives into groups based on their *binding mechanisms*?

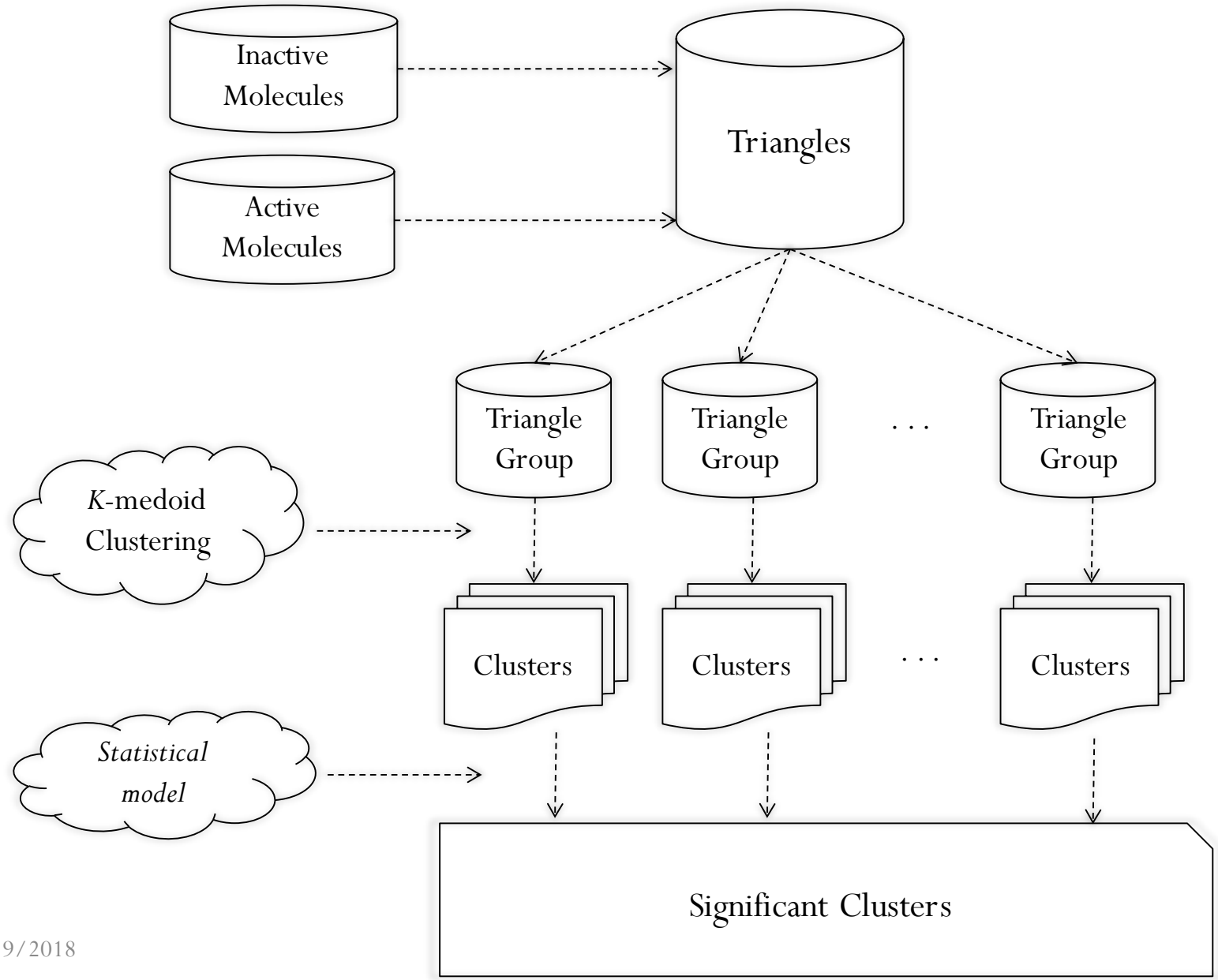
Mining Workflow



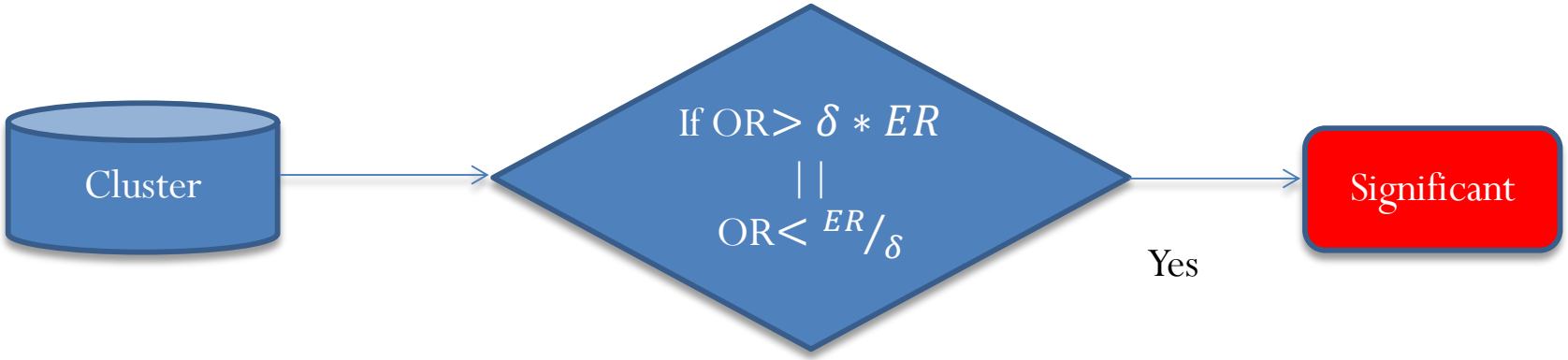
Method: Triangle Extraction



Mining Workflow



Identifying *significant* clusters



Expected Ratio (ER): $\frac{\#\ of\ triangles\ from\ actives\ in\ Database}{\#\ of\ triangles\ from\ inactives\ in\ Database}$

Observed Ratio (OR): $\frac{\#\ of\ triangles\ from\ actives\ in\ Cluster}{\#\ of\ triangles\ from\ inactives\ in\ Cluster}$

Evaluation: Datasets

- CDK5 inhibitors (single target setting)
  - 102 actives, 10,000 inactives
- DUD Datasets (multi-target setting)
  - 20 different targets
  - Actives and inactives corresponding to each target

Results: CDK-5 dataset

- Each triangle-group is divided into 50 clusters

Expected Ratio: 0.01

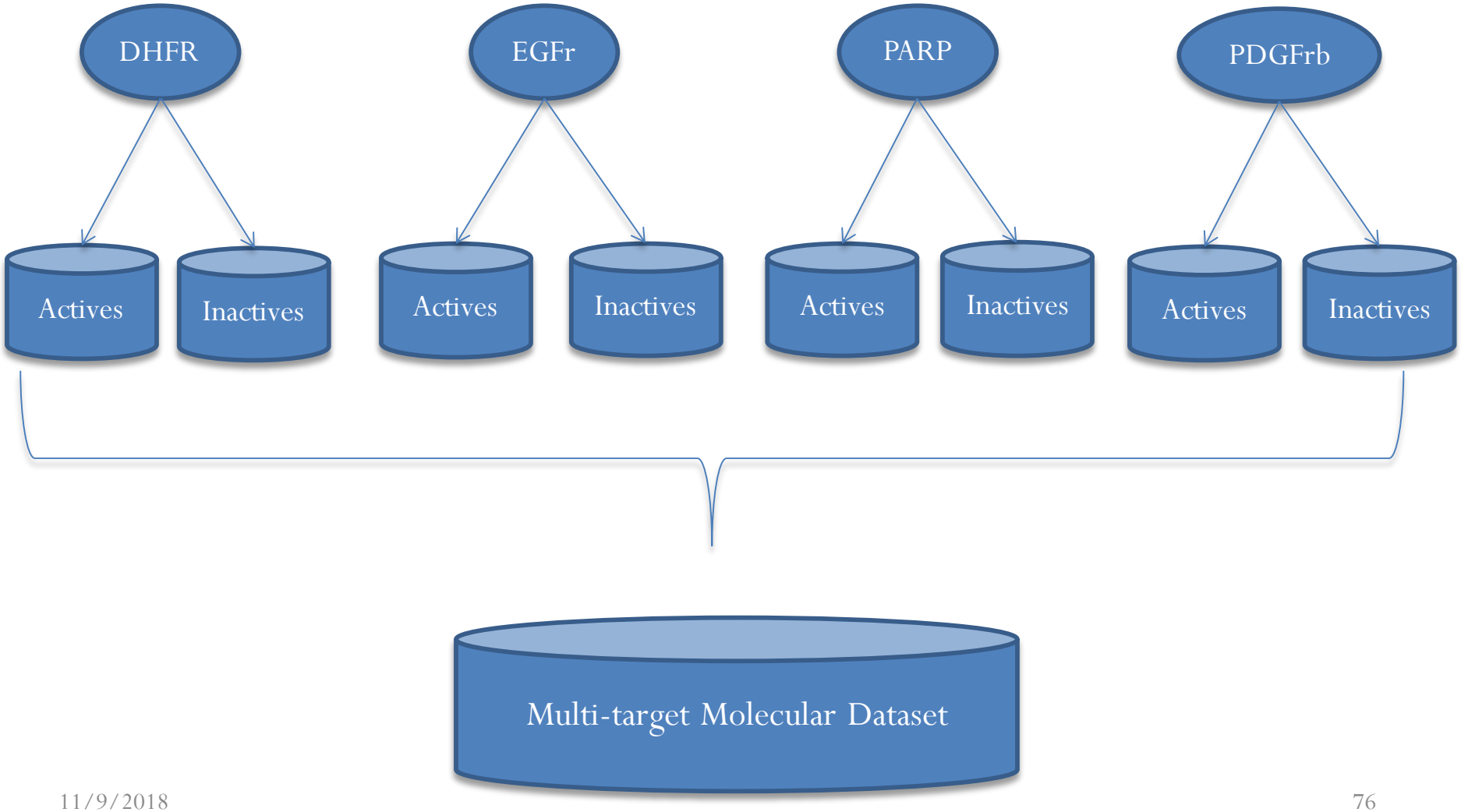
cluster ID	triangle type	cluster size	OR	p-value
1	aromatic−aromatic−aromatic	268	0.24	1.75×10^{-68}
2	aromatic−aromatic−donor	455	0.38	1.71×10^{-223}
3	aromatic−aromatic−donor	545	0.19	5.54×10^{-94}
4	aromatic−aromatic−donor	625	0.08	1.31×10^{-31}
5	aromatic−donor−donor	436	0.25	4.47×10^{-123}
6	aromatic−donor−donor	353	0.3	4.56×10^{-118}
7	aromatic−donor−acceptor	461	0.22	6.38×10^{-102}
8	donor−donor−acceptor	469	0.22	6.92×10^{-104}

All active triangles in a *single* cluster!

Results: Multi-target setting

- Key questions:
  - Do significant clusters exist in the multi-target setting?
  - Can the actives be grouped based on their binding target?

Results: Multi-target setting



Results: Multi-target setting

cluster ID	triangle type	cluster size	ER	OR	p-value
1	aromatic−aromatic−aromatic	468	0.012	0.35	8.39×10^{-184}
2	aromatic−donor−donor	508	0.012	0.69	0
3	acceptor−donor−donor	884	0.012	0.26	1.28×10^{-191}
4	acceptor−donor−donor	1840	0.012	0.33	0
5	aromatic−aromatic−aromatic	1089	0.014	0.38	0
6	aromatic−aromatic−aromatic	132	0.014	0.26	3.88×10^{-34}
7	aromatic−aromatic−donor	1049	0.014	0.24	9.49×10^{-218}
8	aromatic−aromatic−acceptor	2051	0.001	0.04	2.69×10^{-83}
9	aromatic−aromatic−aromatic	311	0.005	0.35	7.77×10^{-144}
10	aromatic−aromatic−donor	445	0.005	0.23	1.35×10^{-92}
11	aromatic−donor−donor	266	0.005	0.28	2.26×10^{-91}

- ✓ Do significant clusters exist in the multi-target setting?
- Does each significant cluster correspond to a single target?

Results: Multi-target setting

- Observed Ratio for each target

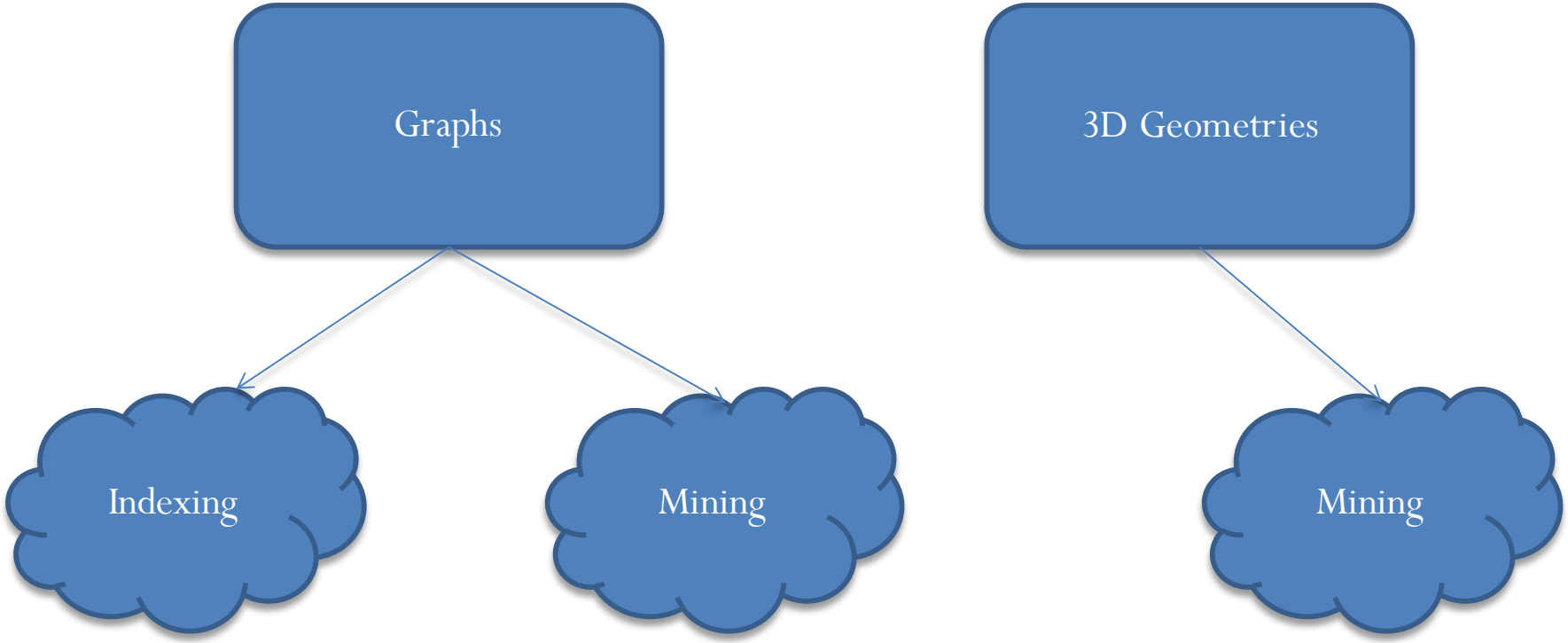
cluster ID	DHFR	EGFr	PARP	PDGFrb
1	**0.34**	0.01	0	0
2	**0.69**	0	0	0
3	**0.23**	0.03	0	0
4	**0.27**	0.01	0	0.05
5	0.05	**0.33**	0	0
6	0	**0.26**	0	0
7	0	**0.24**	0	0
8	0	0	**0.04**	0
9	0	0.04	0	**0.31**
10	0.03	0.03	0	**0.17**
11	0	0.03	0	**0.25**

Each significant cluster corresponds to a single target!

Summary

- Significant clusters *exist* in the *multi-target setting*.
- Significant clusters can be used to *group* actives *based* on their *binding mechanisms*.

Wrap UP



- Fragment based indexing
  - gIndex
- Closure tree

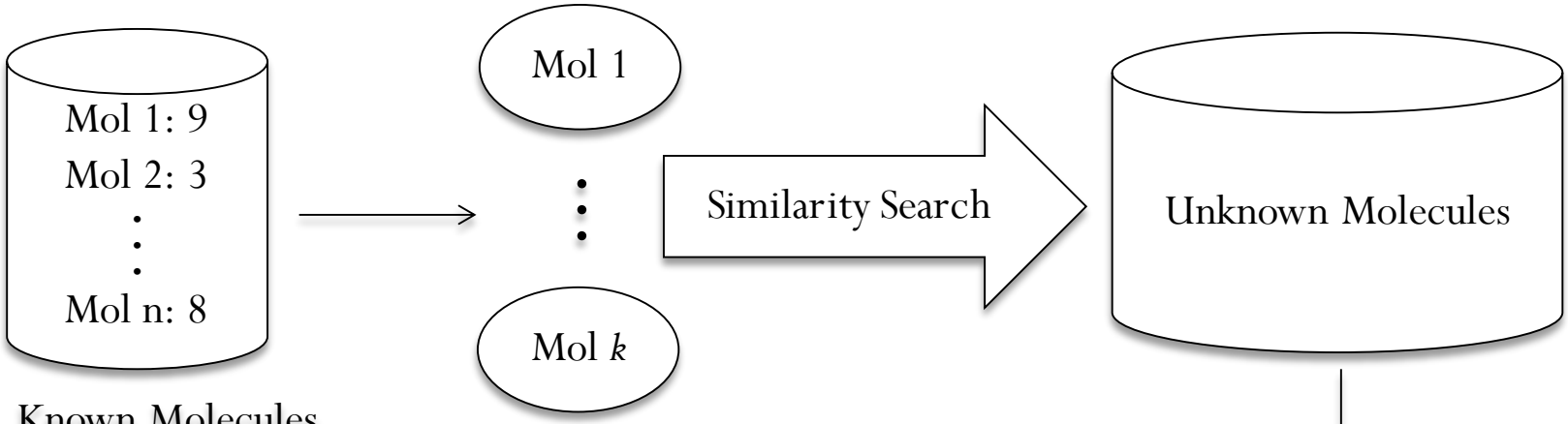
- Frequent Subgraph Mining
  - Join Based Approach
  - Pattern Growth Approach
- Significant Subgraph Mining
  - Leap
  - GraphSig and pGraphSig
  - Molecular Classification

- Mining Significant Geometric Patterns

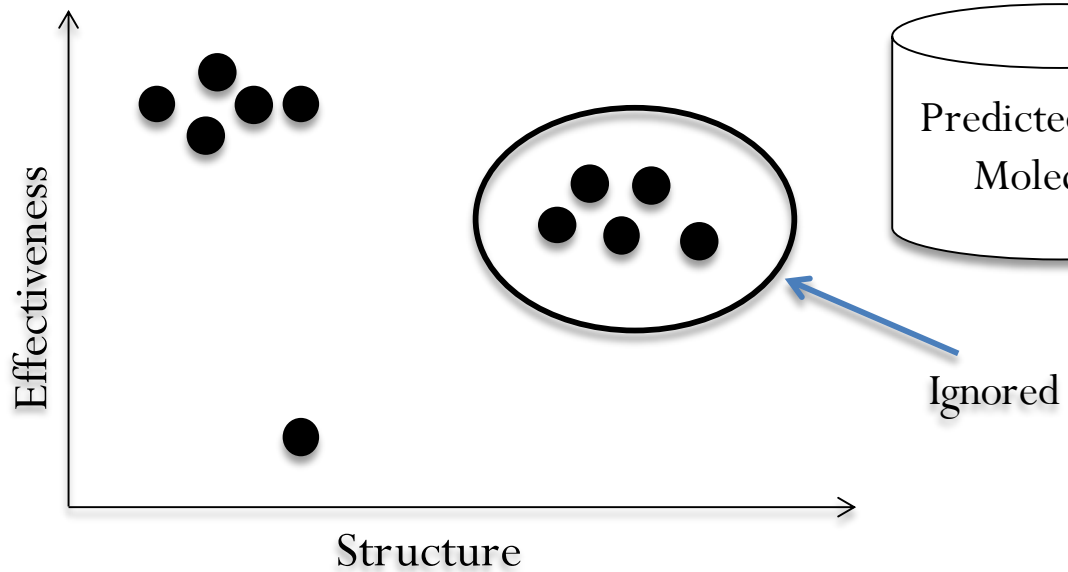
Future Research Directions

- *Budget-aware* querying and mining
  - Traditional top-*k* is not enough
- Drug Repurposing

Similarity search based prediction



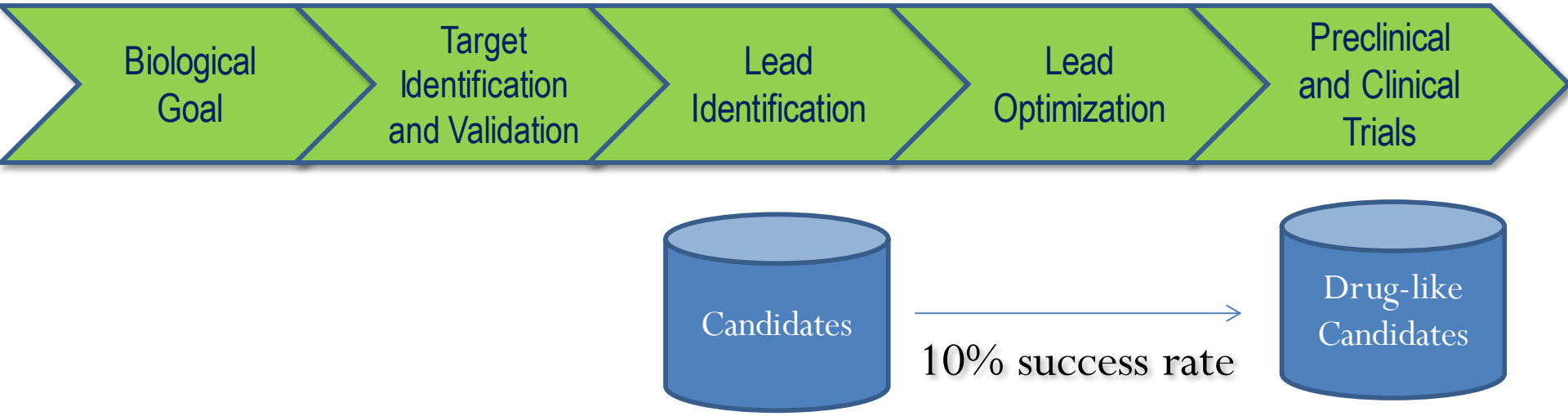
Select effective molecules from *multiple clusters!*



Budget-aware subgraph mining

- Budget
  - *k*
- Mine *k* *best* patterns
- How to *quantify* "best"?
  - Dimensionality reduction in vector space
  - How to *model orthogonality* in structural space?

Drug Repurposing



Inefficient

- Cost
- Time

Alternative Route

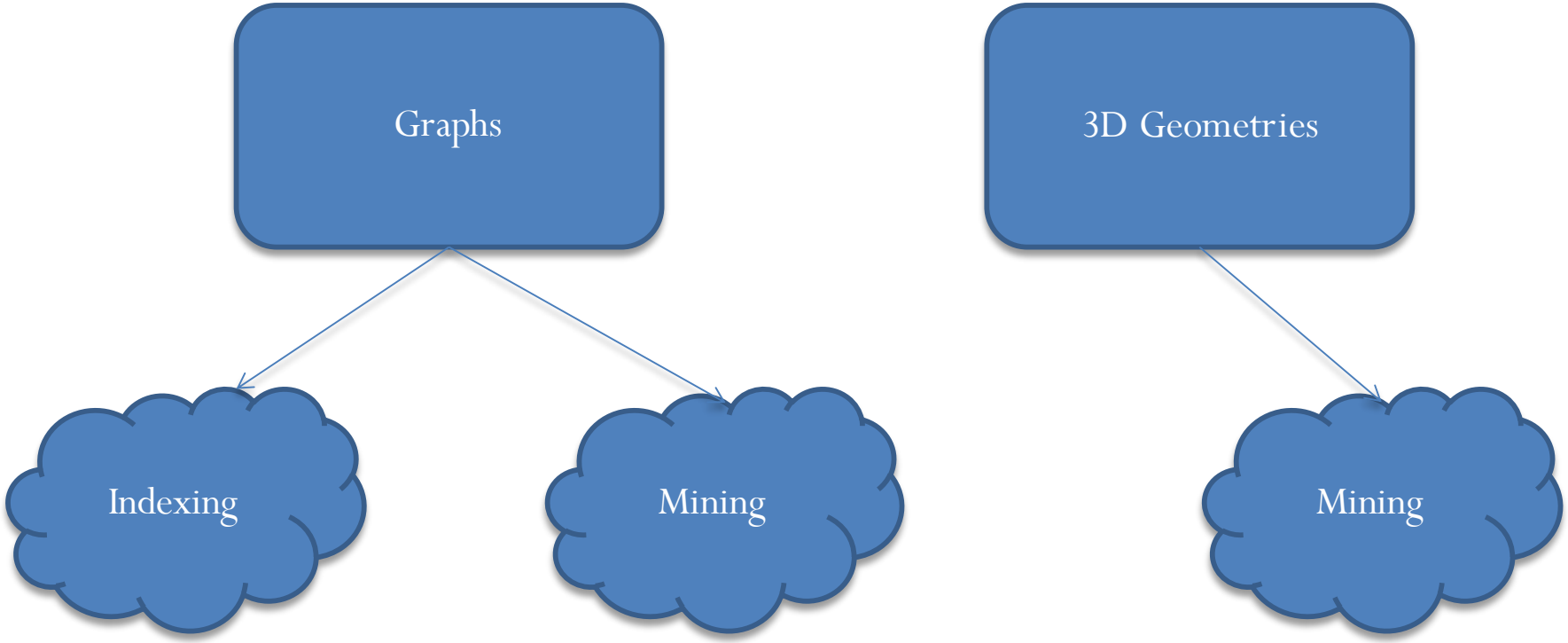
- Find *new uses* of *existing drugs*

Conclusion

- Computer Science *plays a key role* in drug discovery
- Modeling *molecules as graphs* allows us to apply powerful graph analysis tools
- Future Directions: Range $\longrightarrow$ top-k
  - *Maximize information content* in top-k answer set

Thank You!

Conclusion

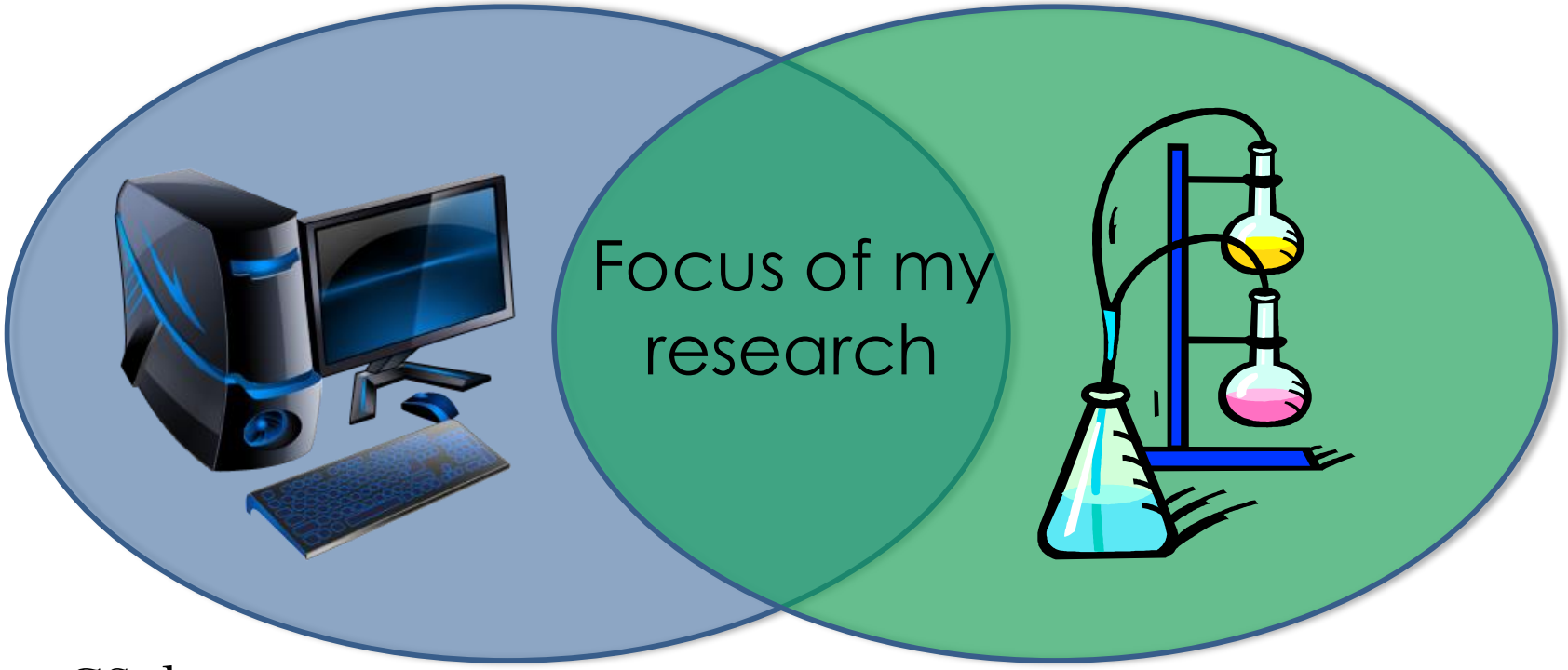


- gIndex
- Closure tree

- Frequent Subgraph Mining
  - Join Based Approach
  - Pattern Growth Approach
- Significant Subgraph Mining
  - Leap
  - GraphSig and pGraphSig

- Mining Significant Geometric Patterns

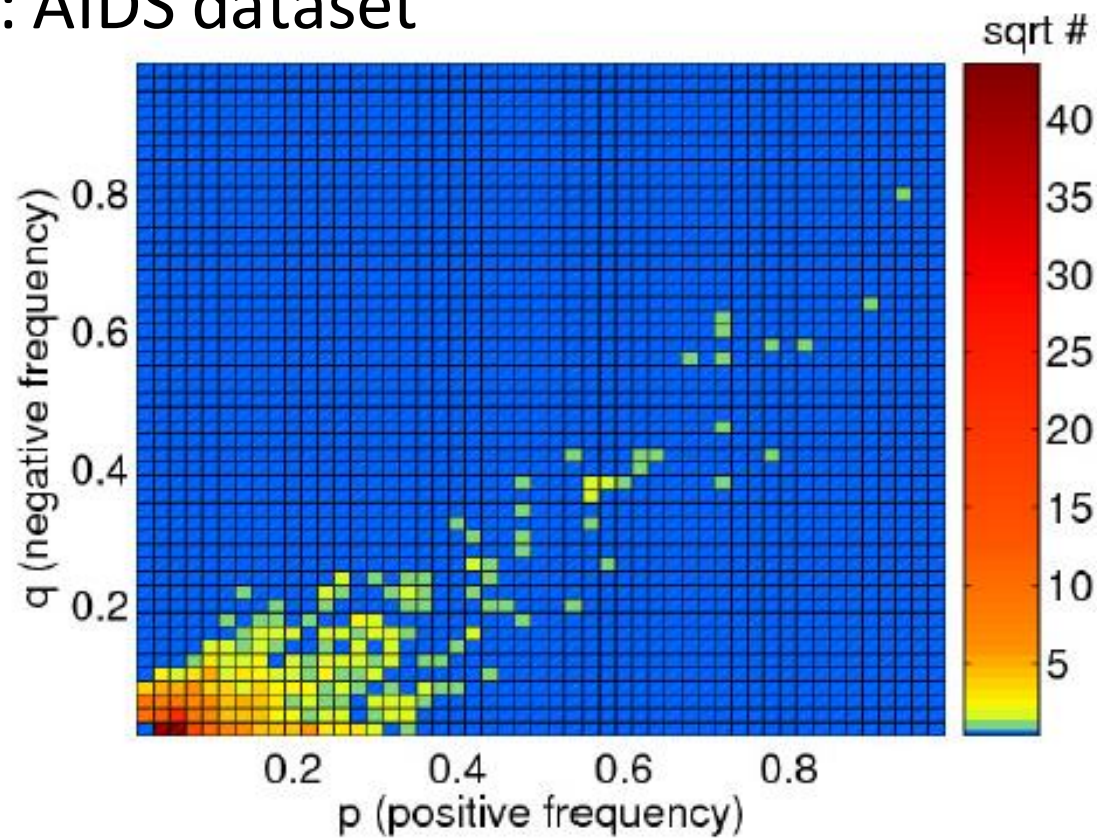
Research Summary



- CS driven
  - A difficult problem in CS with applications in chemistry
  - Publish in CS conferences (SIGMOD, VLDB, ICDE, KDD etc.)
- Chemistry driven
  - Use CS techniques to solve a problem in chemistry
  - Publish in Chemistry Journals (JCIM, Bioinformatics etc.)

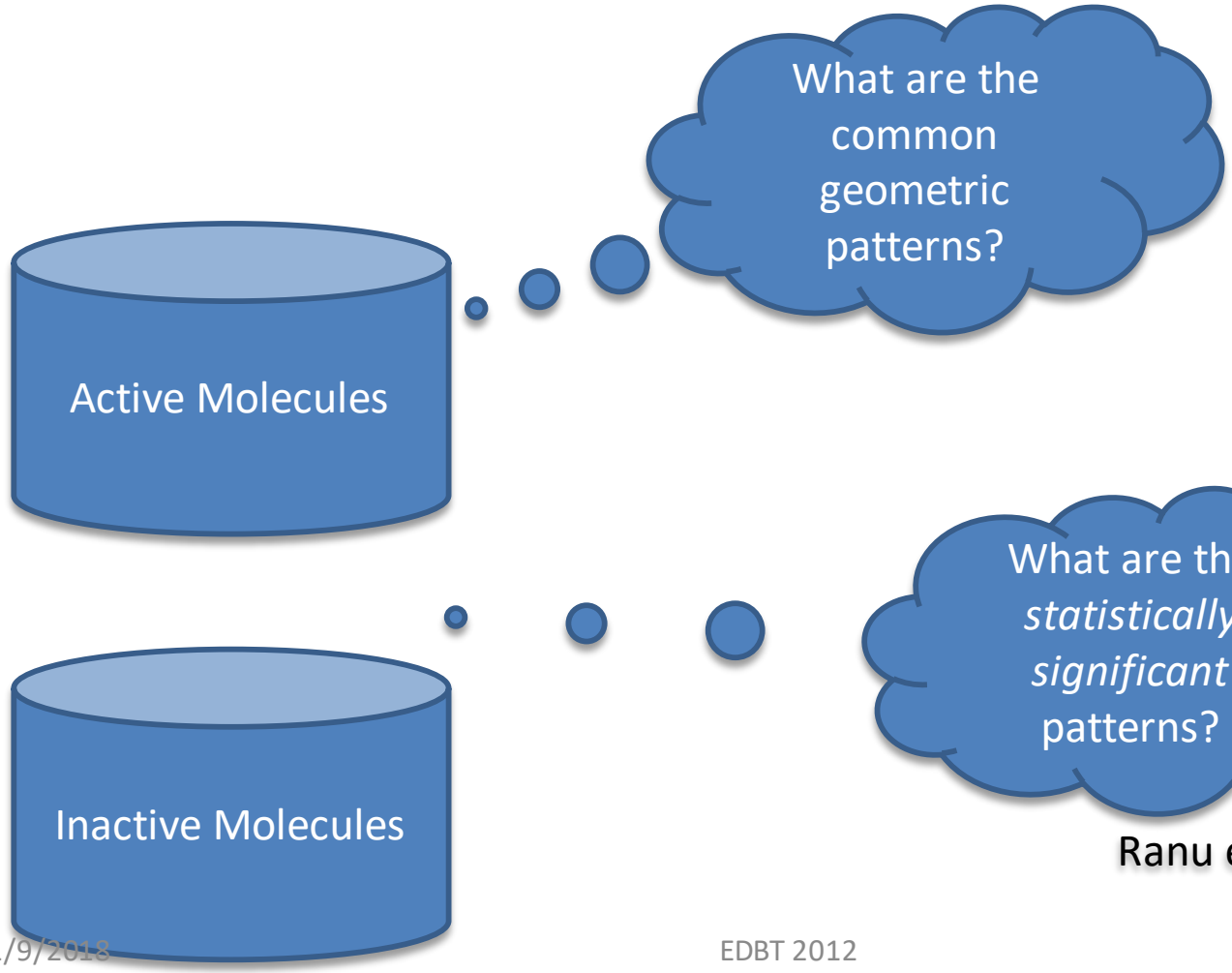
Horizontal Pruning: Verification

Dataset: AIDS dataset



Many subgraphs share the same frequencies!

Mining Geometric Patterns

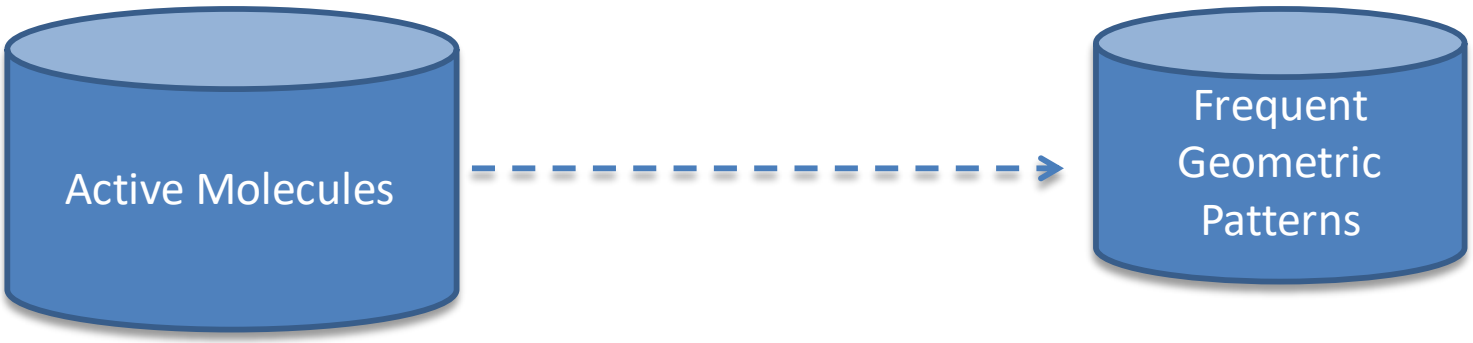


Podolyan et al. JCIM 2009.

Ranu et al. JCIM 2011.

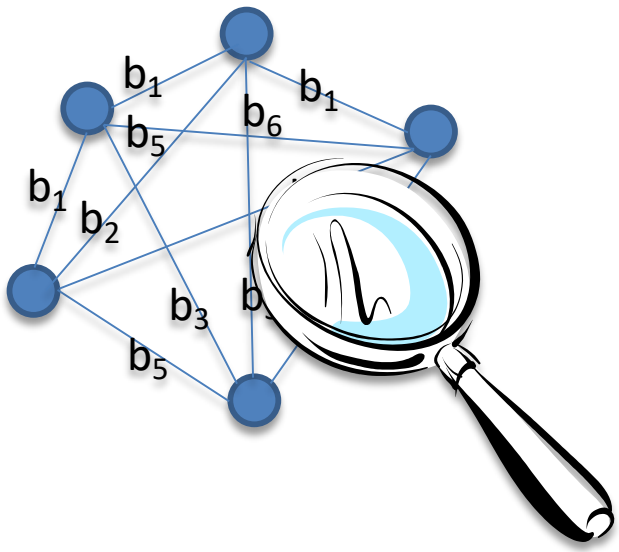
Mining Frequent Geometric Patterns [JCIM, 2009]

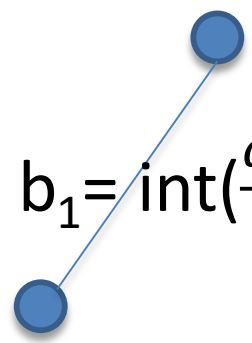
Geometries required to bind to a protein target



Can we represent geometries as graphs?

Representing Geometries as Cliques



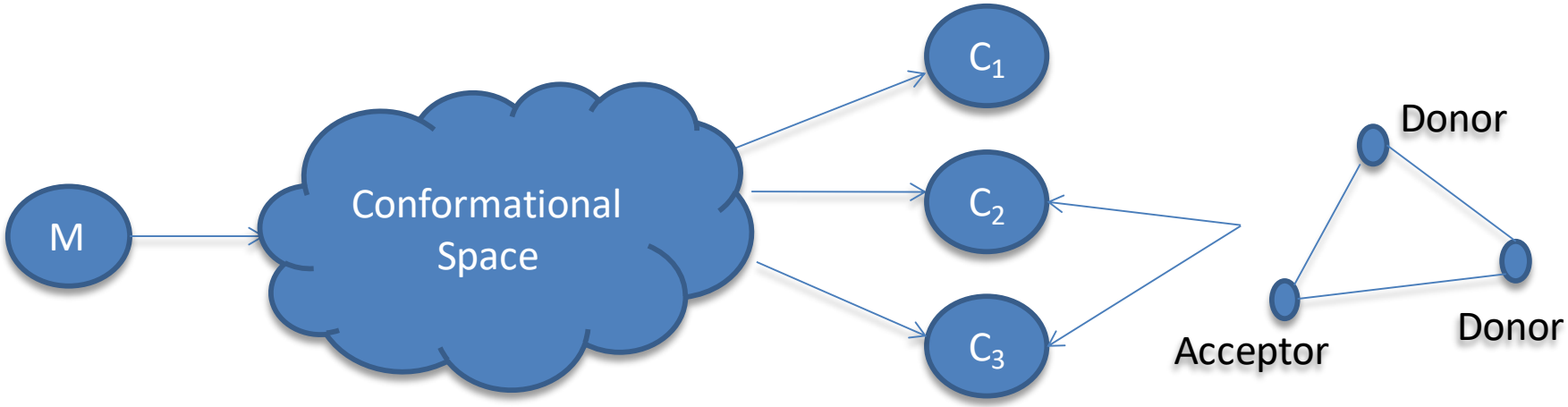


$$b_1 = \text{int}\left(\frac{distance}{\delta}\right)$$

3D Geometry → Clique

Mining frequent geometries → Mining frequent cliques

Managing the conformational space

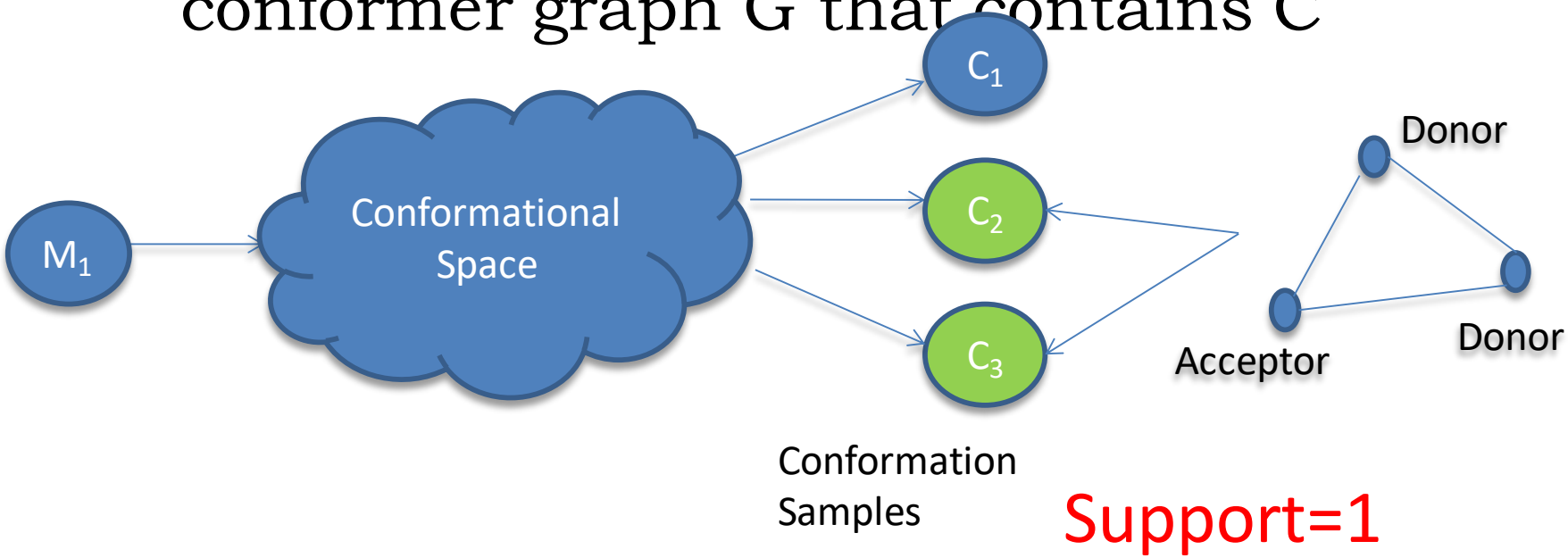


Conformation Samples

What is the support? 2 or 1?

Computing Support

- sup(C) = |M|,
  - where $M \subseteq \{M_1, \ldots, M_n\}$
  - each molecules in M has at least one conformer graph G that contains C



Mining frequent geometries

Graph Vs. 3D Geometry

- ✓ No edges
- ✓ Dynamic structure
  - ✓ Multiple *conformations* per molecule

Re-use frequent subgraph mining techniques!

Significant Pattern Mining [JCIM, 2011]