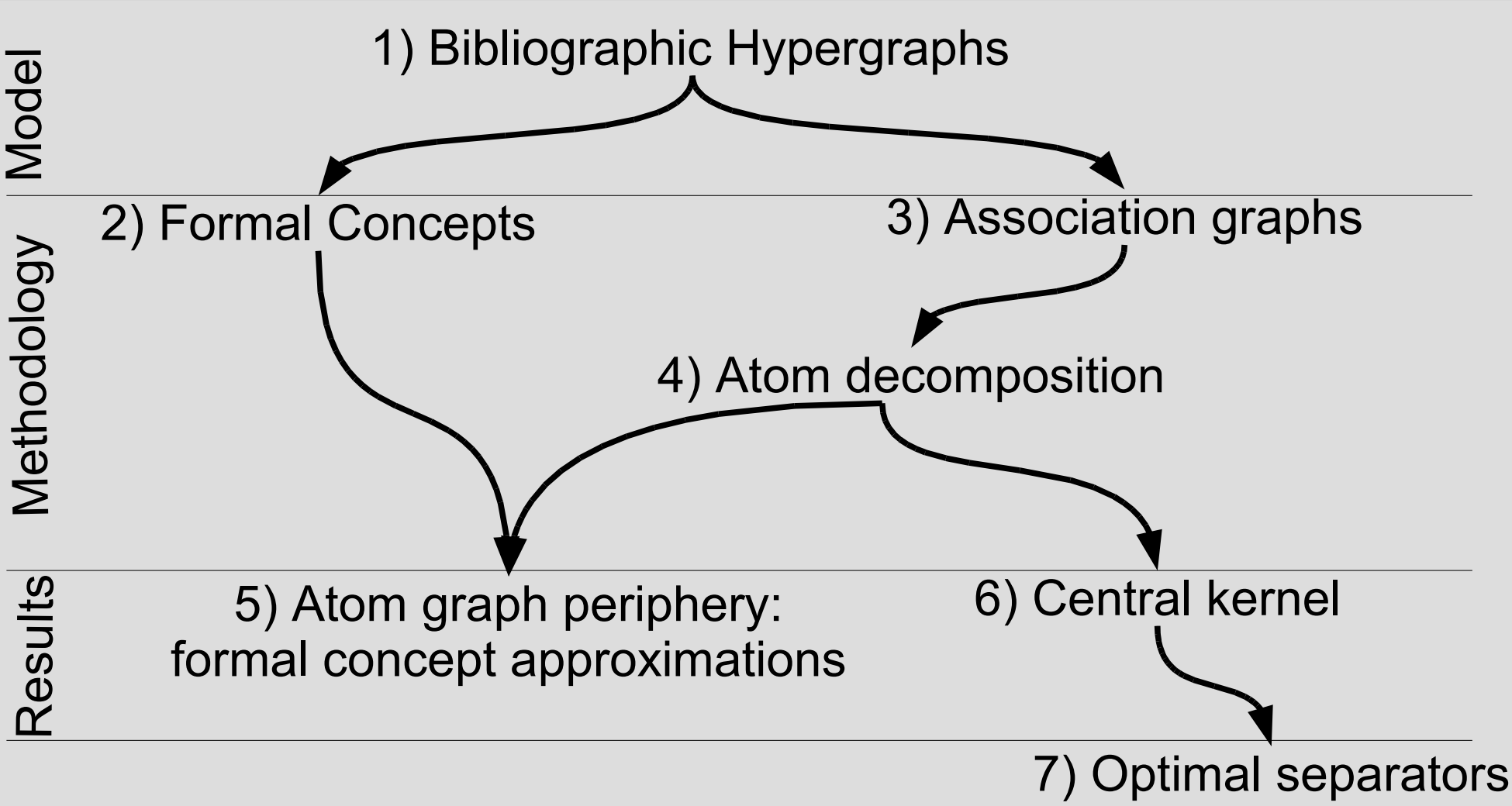
Hypergraph Modeling and Graph Clustering Process Applied to Co-word Analysis

Bangaly Kaba
LIMOS, Université Blaise Pascal Clermont 2, France
kaba@isima.fr

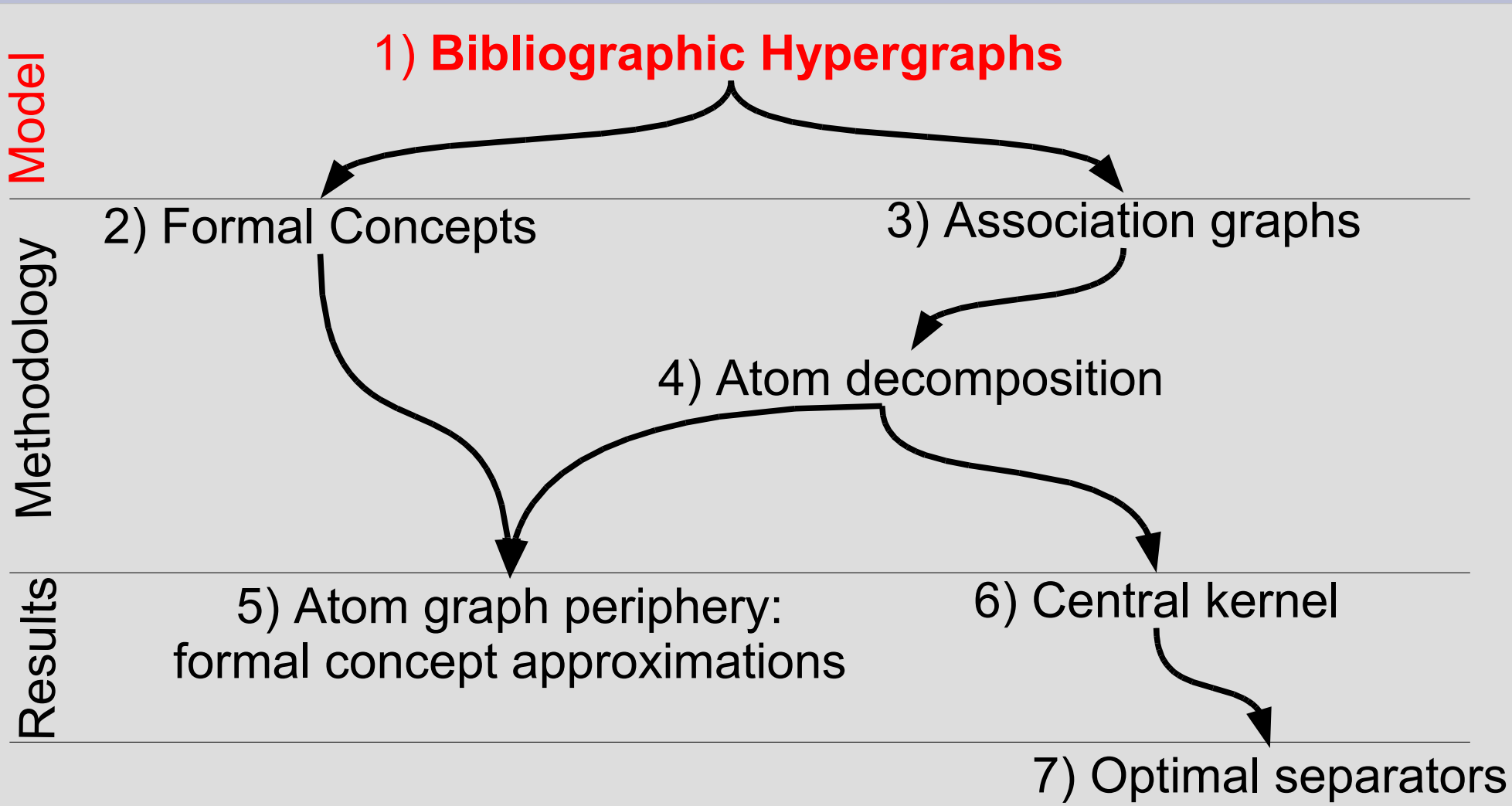
Xavier Polanco
LIP6, Université Pierre et Marie Curie, France
xavier.polanco@lip6.fr

Eric SanJuan
LIA, Université d'Avignon, France
eric.sanjuan@univ-avignon.fr

Summary

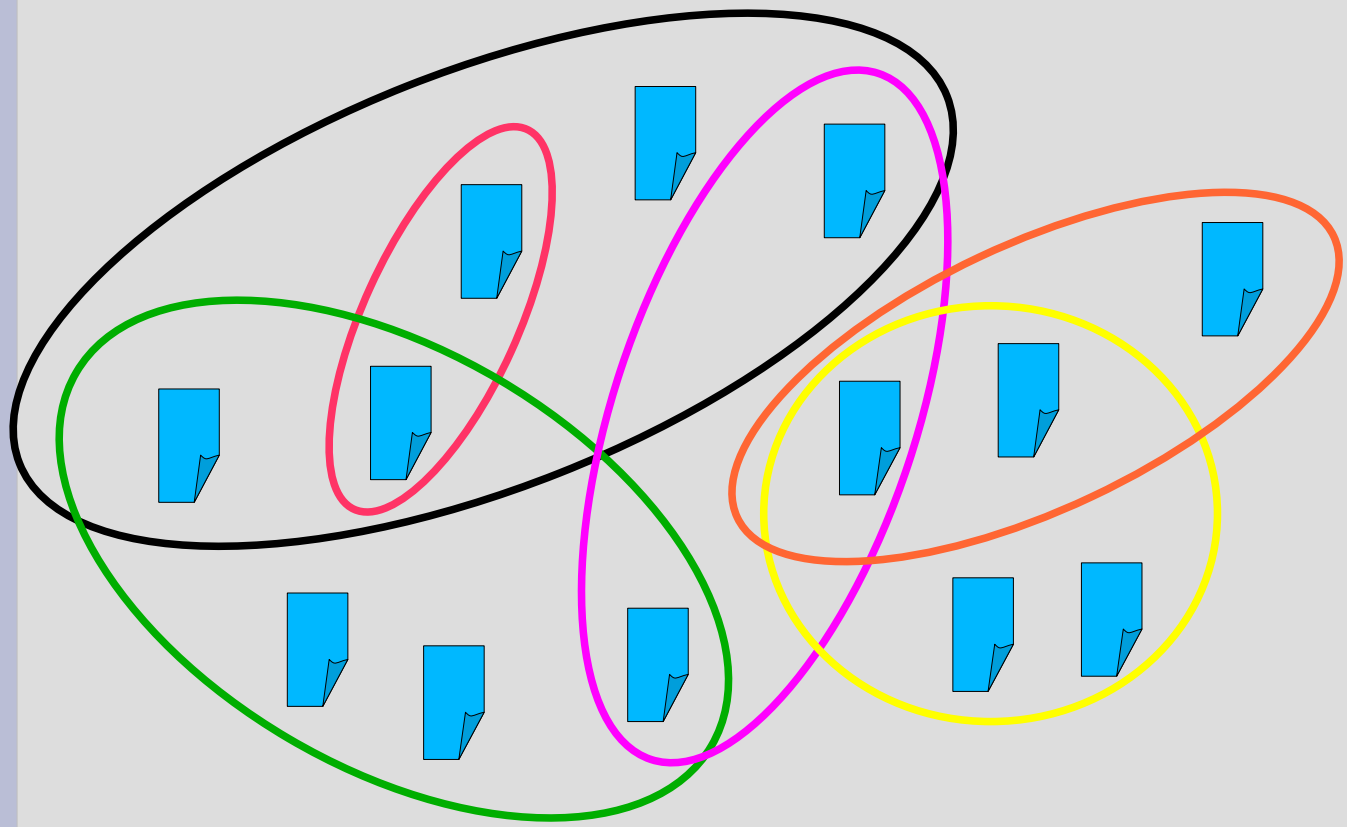

Model
Methodology
Results
1) Bibliographic Hypergraphs
2) Formal Concepts
3) Association graphs
4) Atom decomposition
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

Summary


Model
Methodology
Results
1) Bibliographic Hypergraphs
2) Formal Concepts
3) Association graphs
4) Atom decomposition
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

Bibliographic Hypergraph

- Family of sets (hyper-edges) of documents (hyper-vertex)
- Each document item defines an hyper-edge

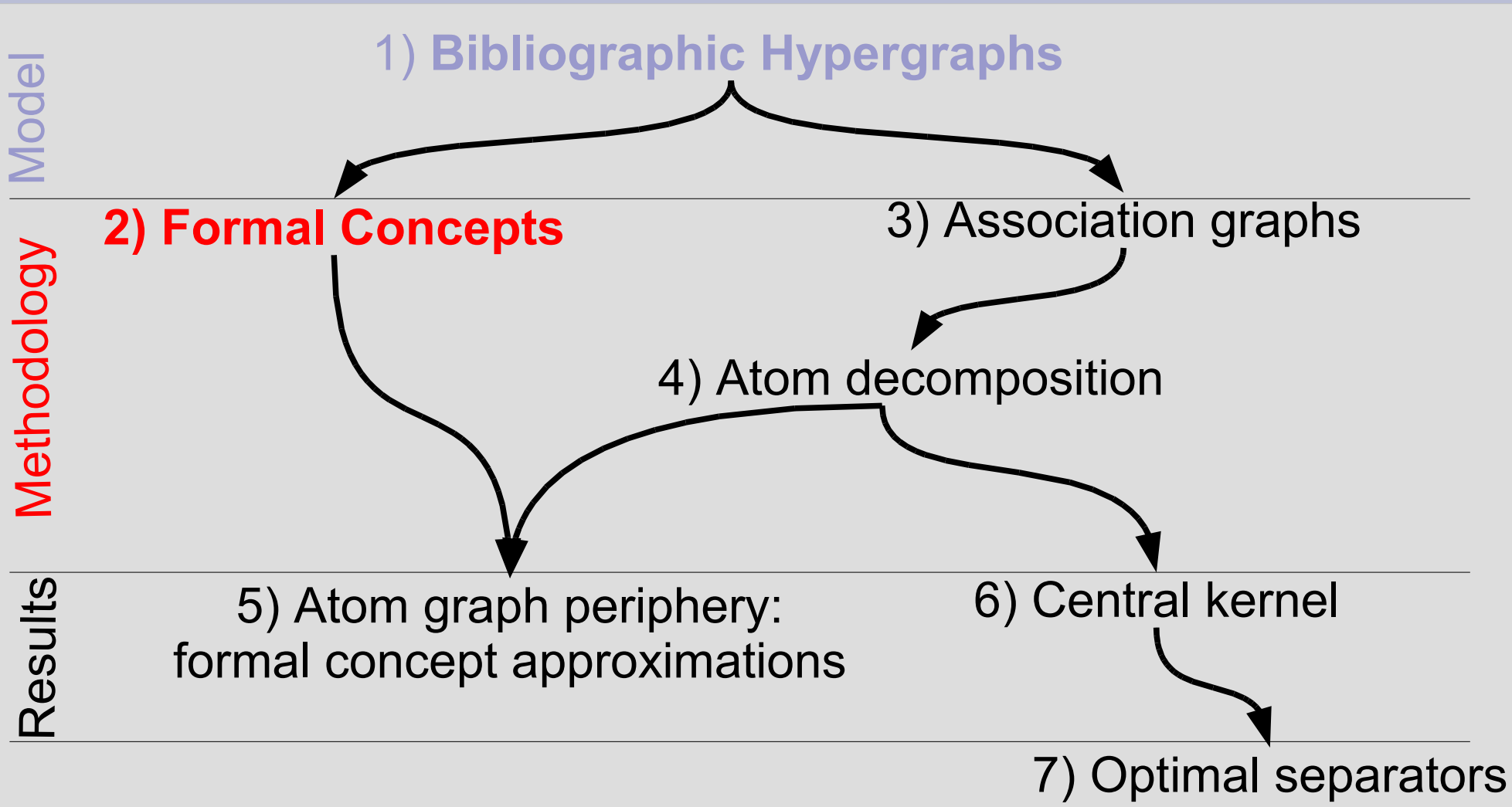


Items:
AU = Smith
KW = IR
KW = Text mining
CI = John
YR = 2005
KW = Graph

Bibliographic Hypergraph

- 3,671 records extracted from the SCI database containing the keywords data mining and text mining over the period 2000-2006.
- Indexed by a set of 8,040 keywords.
- The average number of keywords per record is 5.
- **1,524** keywords (hyper-edges) of frequency > 1 indexing **2,615** records (hyper-vertex).

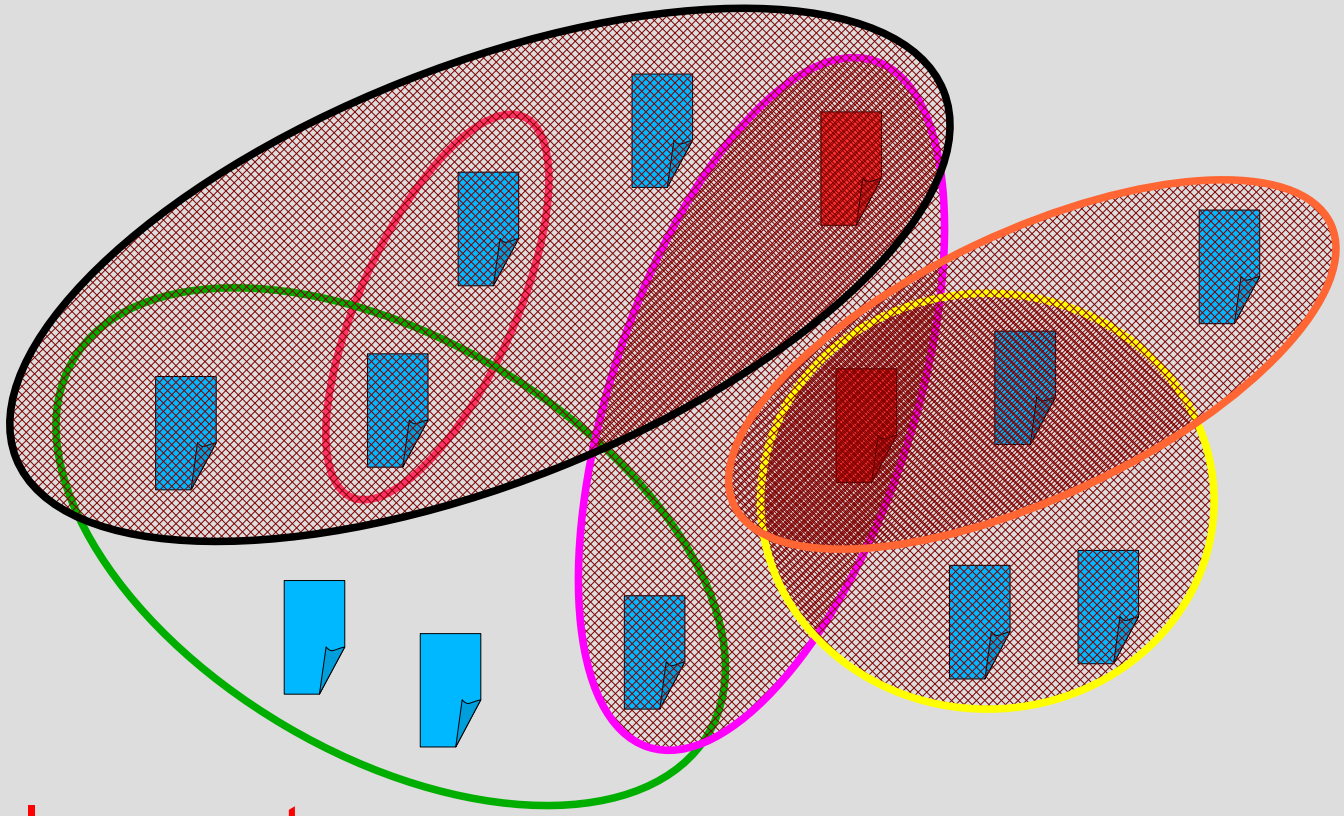
Summary


Model
Methodology
Results
1) Bibliographic Hypergraphs
2) Formal Concepts
3) Association graphs
4) Atom decomposition
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

Formal concepts

- Association rules (7,082, s > 10%, c > 80%):
  - gene expression, genetic algorithms => rough sets
  - database, information retrieval => visualization
  - ...
- Closed set of items (2,526):
  1. bioinformatics, cancer, data mining, genomics, proteomics
  2. data analysis, data mining, dimensionality reduction, feature extraction, pattern recognition
  3. clustering, machine learning, microarray, proteomics, text mining
- Closed set of documents = Hypergraph minimal tranversals

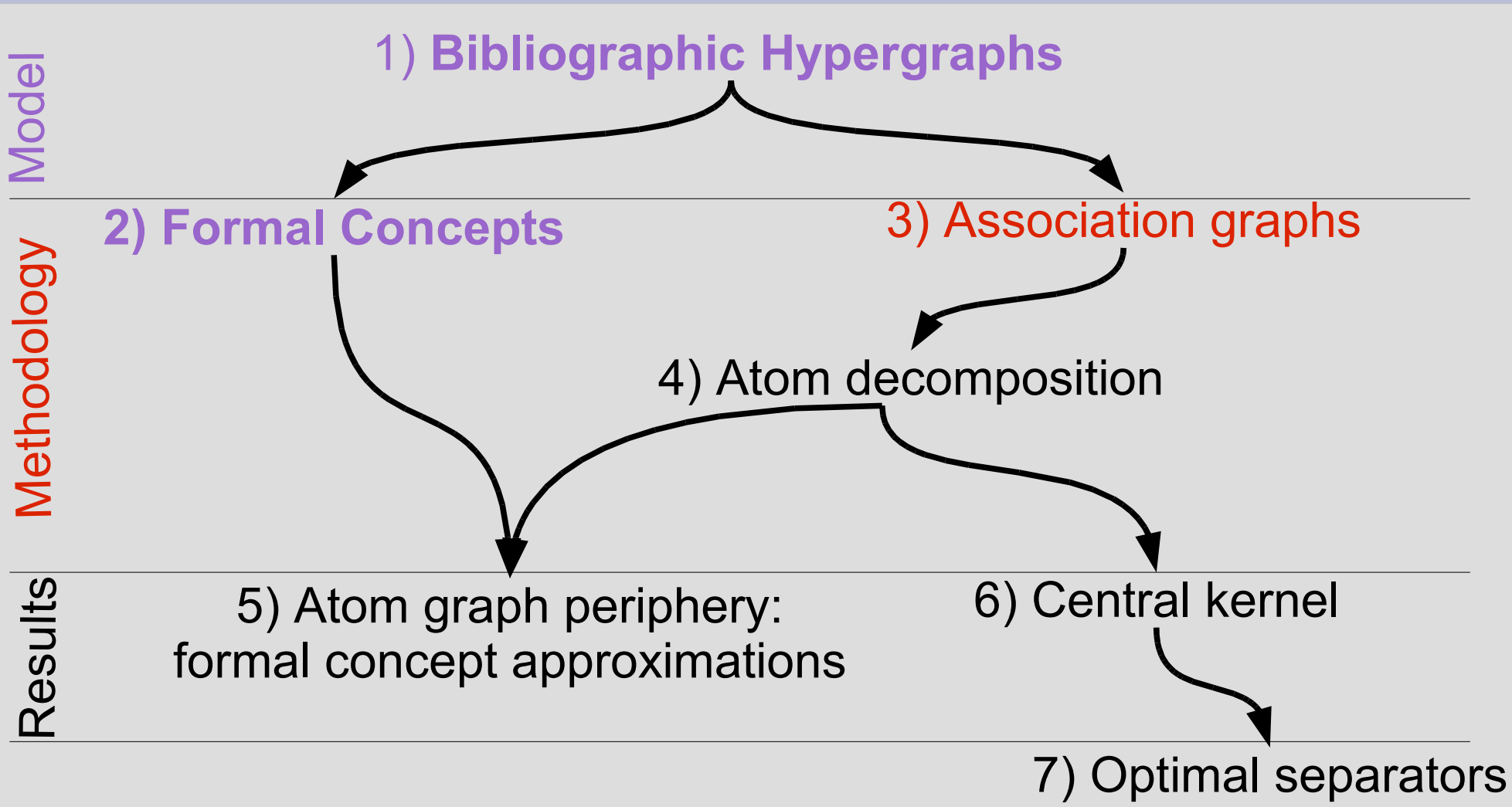
Formal concepts (illustration)



Formal concept :

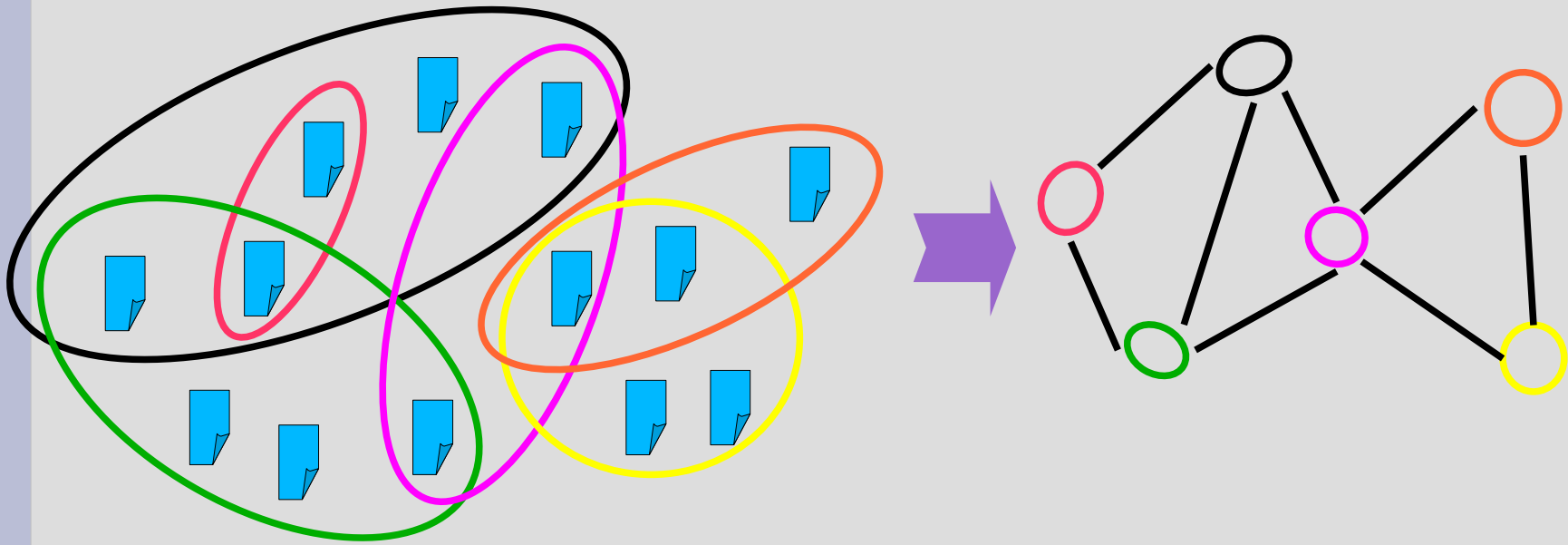
- a closed item set (*intension*)
- with its correspondent set of documents (*extension*)

Summary


Model
1) Bibliographic Hypergraphs
Methodology
2) Formal Concepts
3) Association graphs
4) Atom decomposition
Results
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

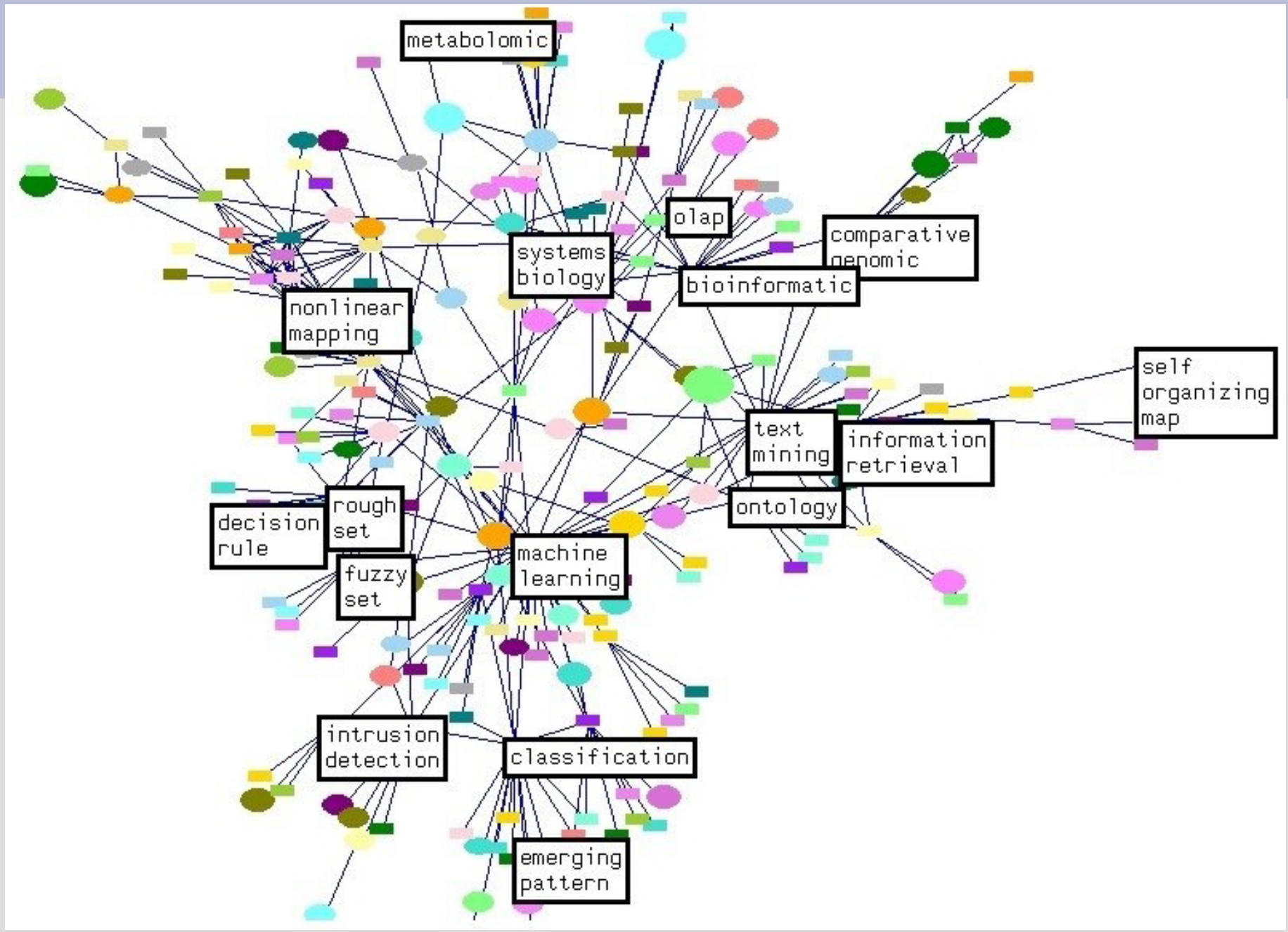
Association graphs

- Intersection graphs derived from the biblio-graphic hypergraph are Small Worlds.

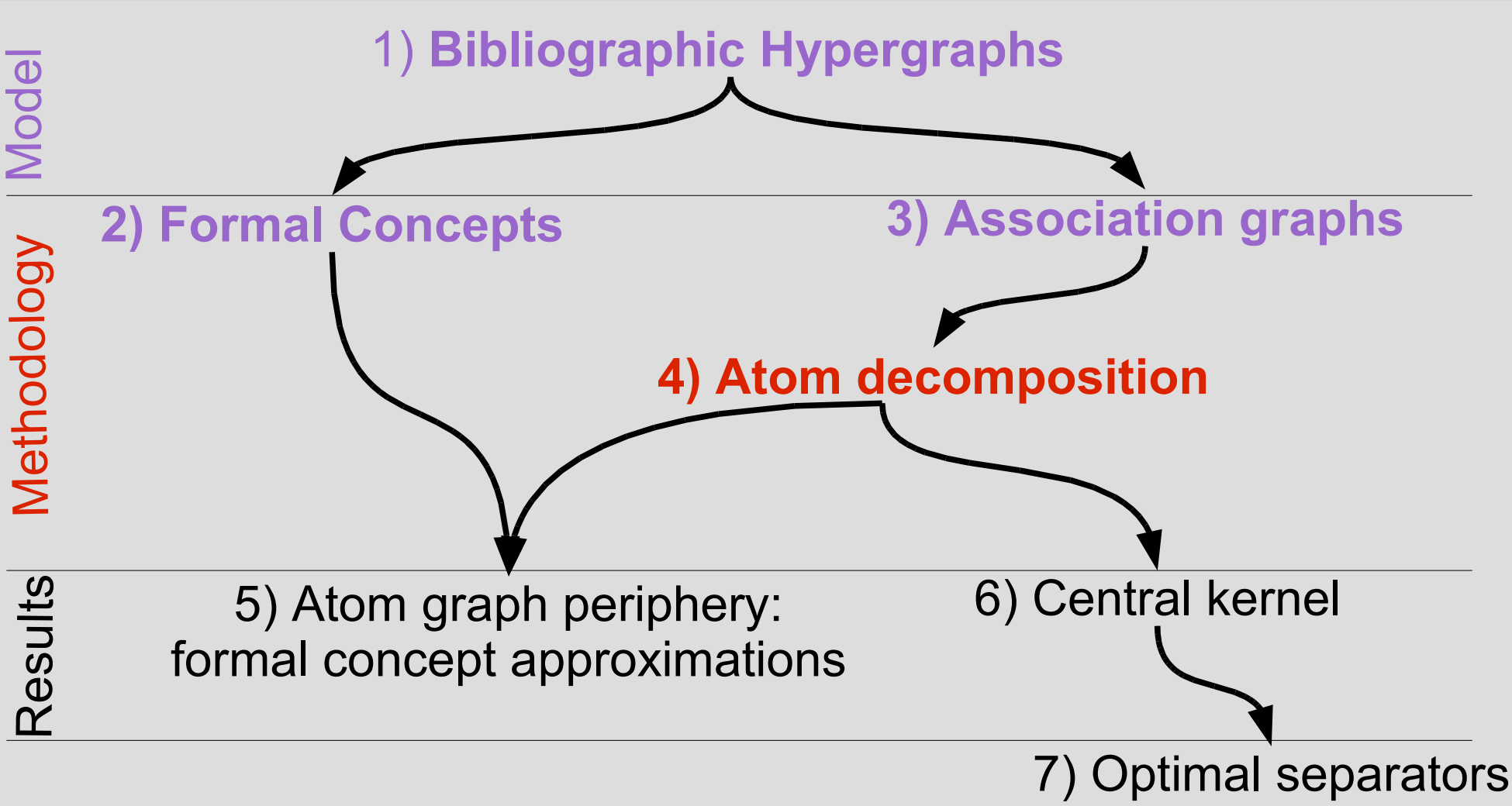


- A threshold can be set on intersection car-dinalities and a coefficient can be defined on edges (mutual information).

Association graphs (clustered)

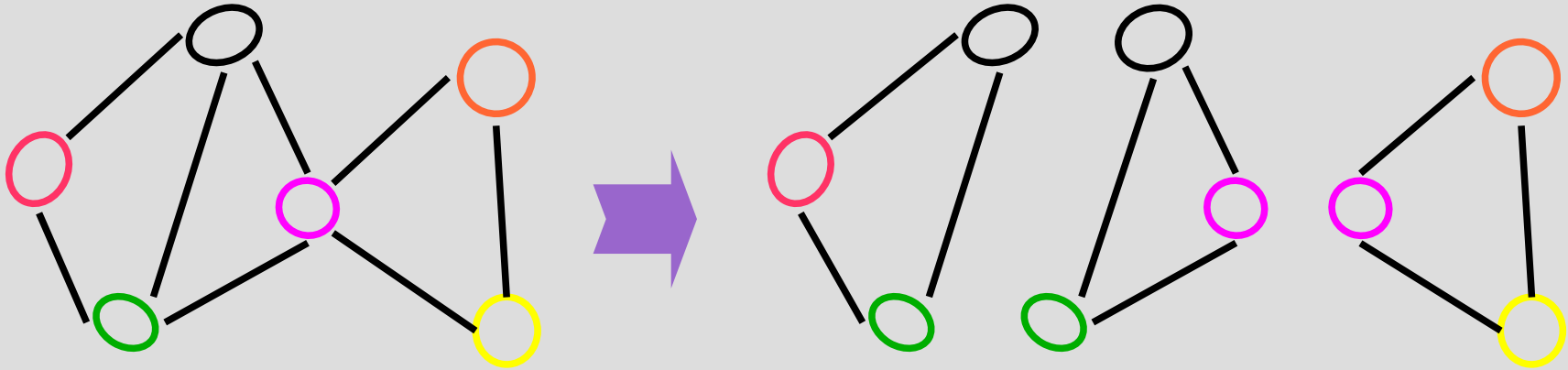


Summary


Model
Methodology
Results
1) Bibliographic Hypergraphs
2) Formal Concepts
3) Association graphs
4) Atom decomposition
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

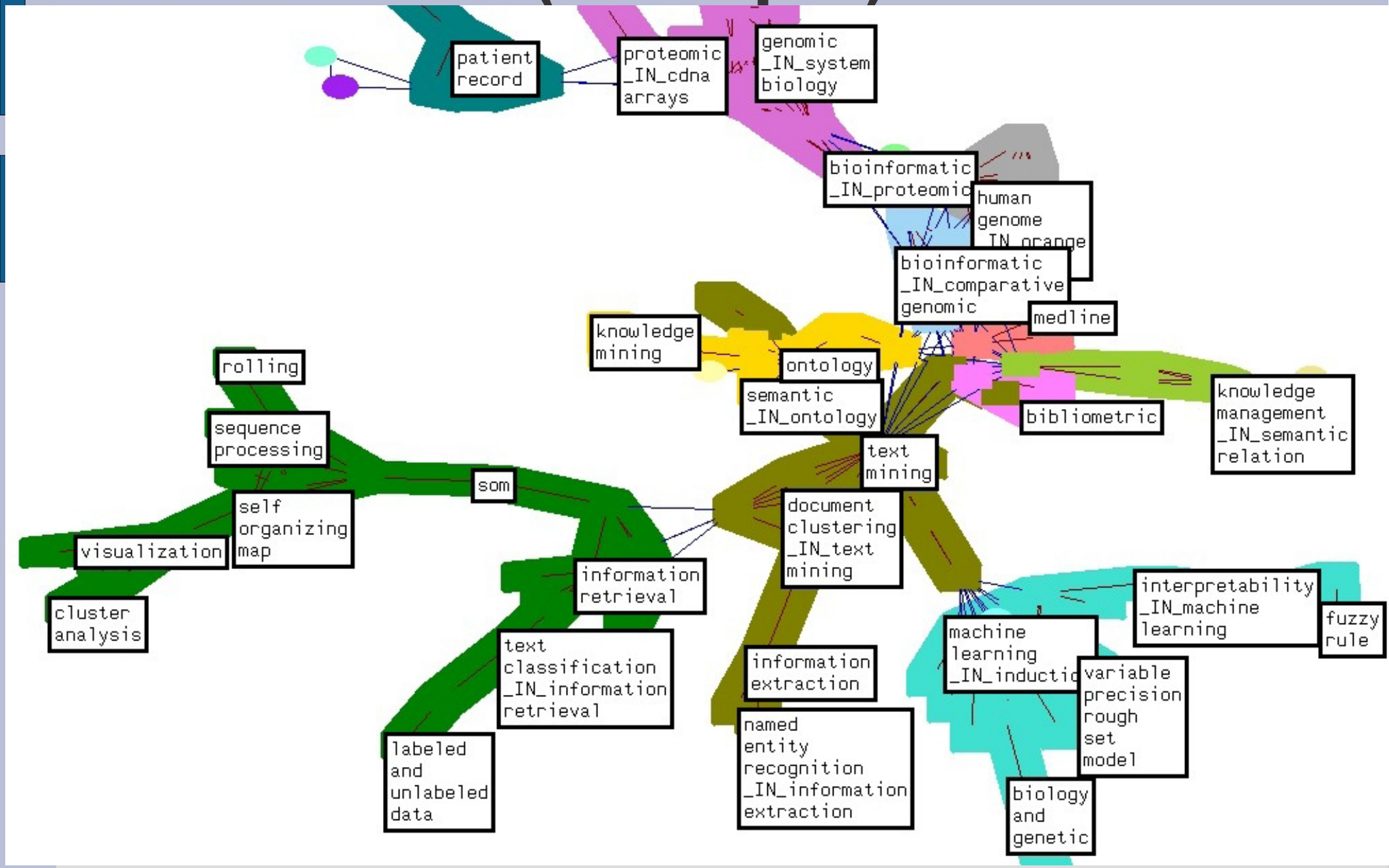
Atom Graph decomposition

- Connected Subgraphs without complete separators:

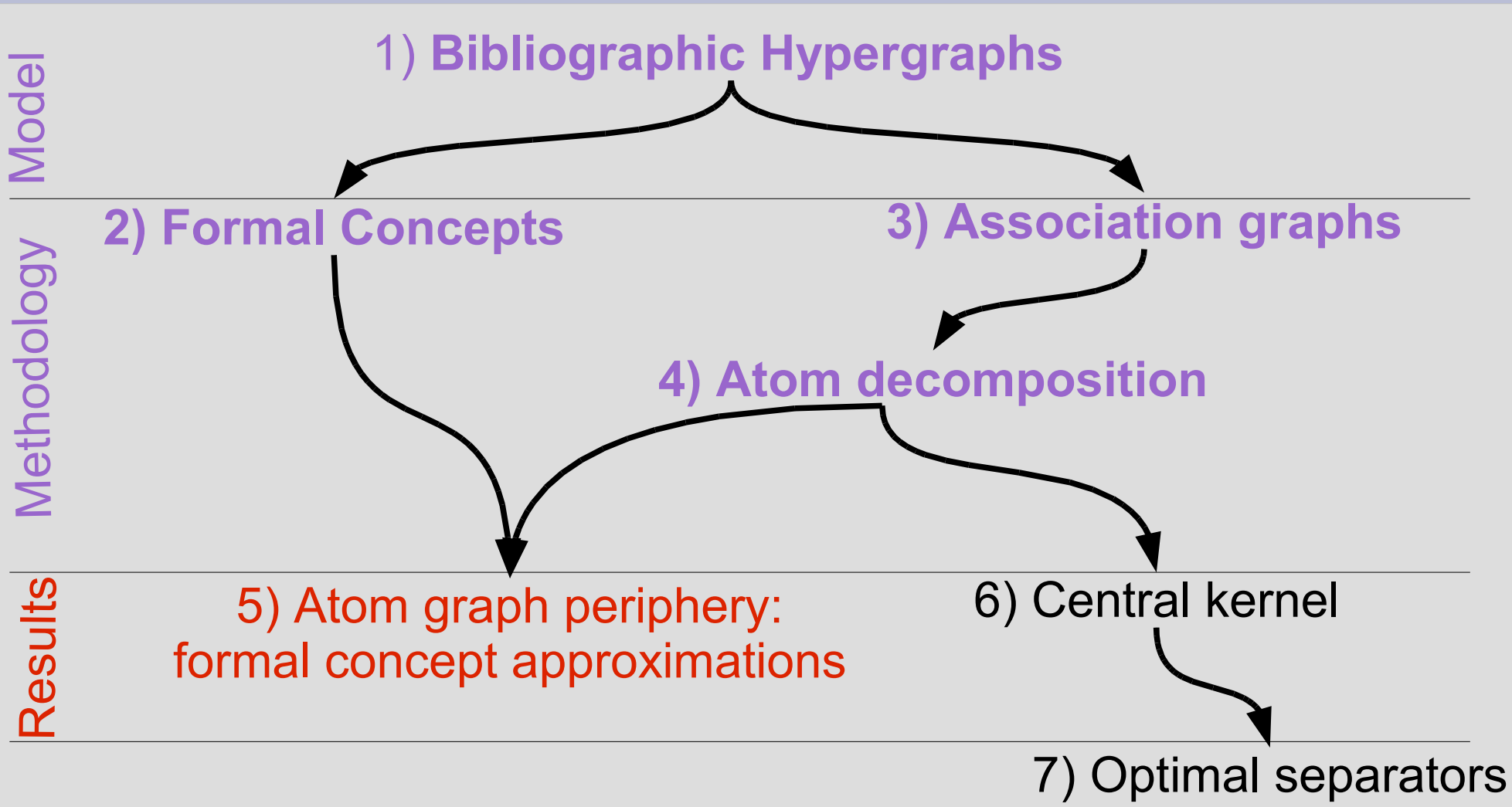


- Unique solution
- Complexity = O(#Edges.#vertex)
- Atoms labeled by their center

Atom Graph decomposition (example)



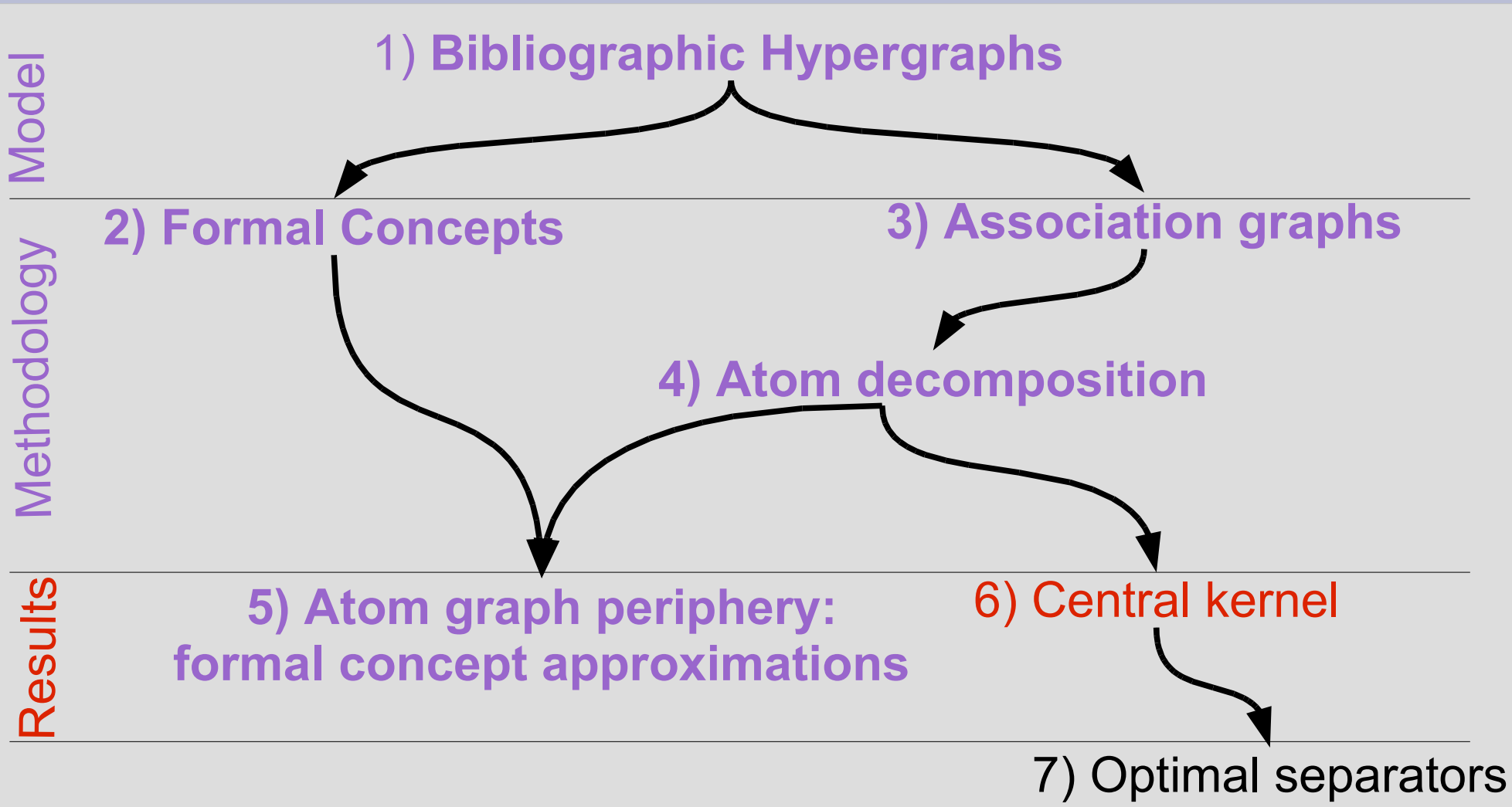
Summary


Model
Methodology
Results
1) Bibliographic Hypergraphs
2) Formal Concepts
3) Association graphs
4) Atom decomposition
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

Atom graph periphery: formal concept approximations

- Association graphs are divided into a central core and a periphery:
  - The main component of the association graph has 645 vertices, 1, 057 edges and 404 atoms.
  - A central atom involves 298 vertices. The remaining 403 atoms have less than13 vertices.
- 96% of the 403 small atoms are closed itemsets and thus formal concept intensions.
- graph of peripheral atoms has
  - 598 vertices that represent pairs of atoms and keywords.
  - It involves 201 different keywords.
  - Like for concept intensions, the overlap between atoms is important.

Summary


Model
1) Bibliographic Hypergraphs
Methodology
2) Formal Concepts
3) Association graphs
4) Atom decomposition
Results
5) Atom graph periphery:
formal concept approximations
6) Central kernel
7) Optimal separators

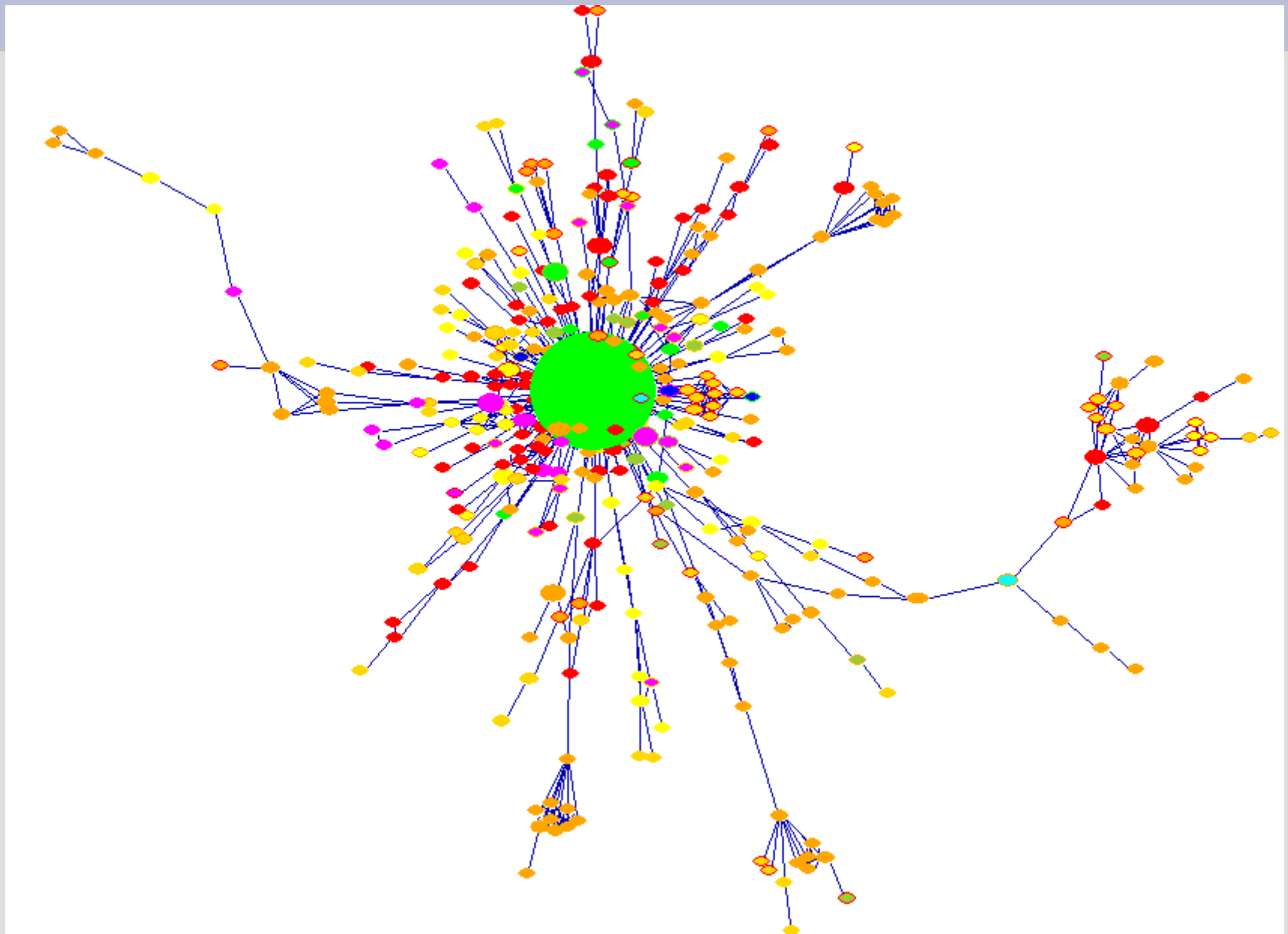
Central Kernel

- Clustered association graph based on a Single Link variant (CPCL) allows to visualize the central atom
  - It has 84 vertices that represent clusters of keywords in the central atom.
  - The biggest cluster has only 10 vertices
  - 86% of these clusters are closed itemsets.

- Thus, CPCL clustering on association graphs seems to be coherent with formal concepts. However it reveals less formal concepts than peripheral atoms.

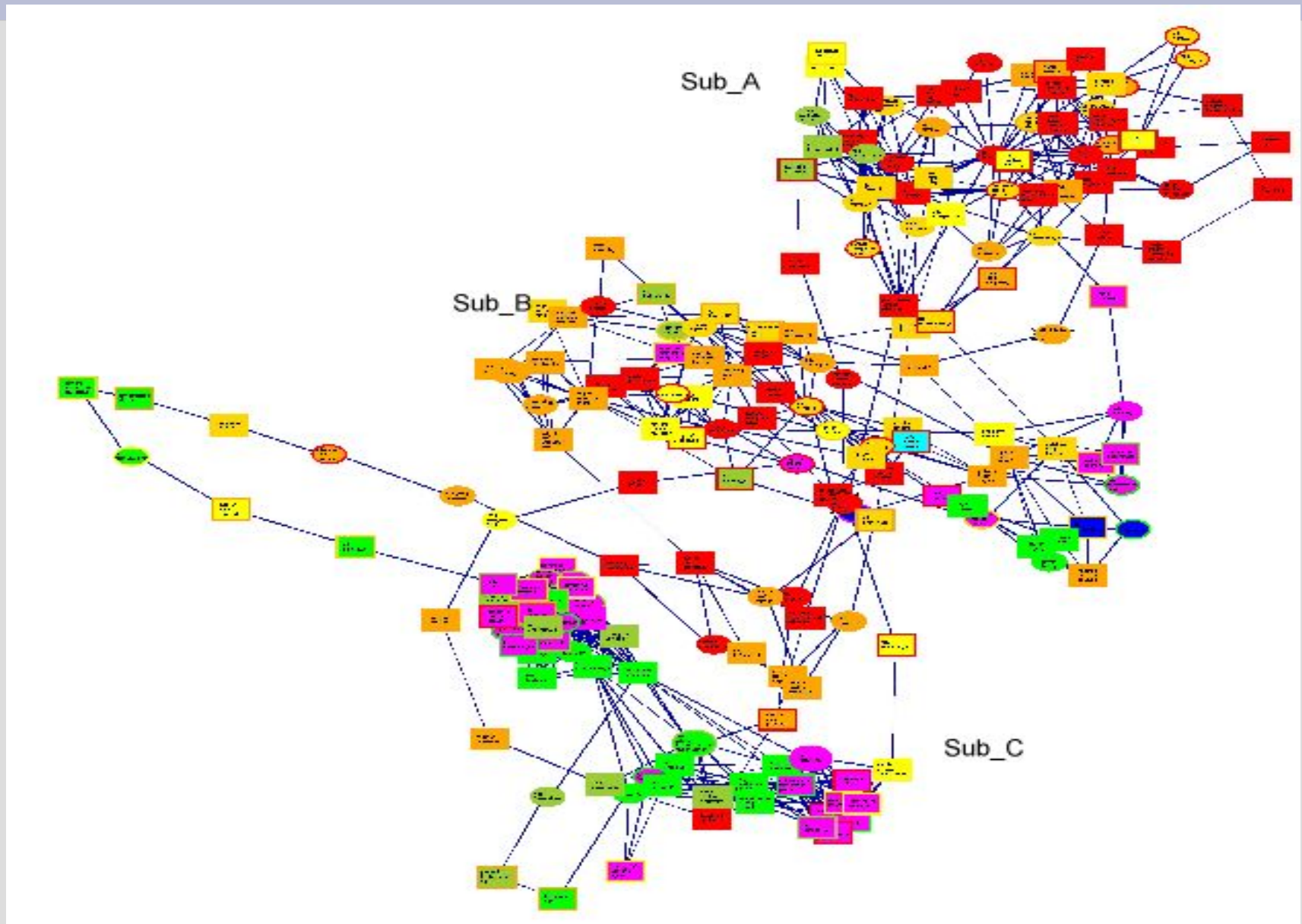
Central Kernel

(application by Fidelia Ibekwe – Lyon 3 - France)



Optimal separators

(Work in progress with Marie Jean Meurs - LIA)



Thank you for your attention