Identify Dwarfs Workloads in Big Data Analytics

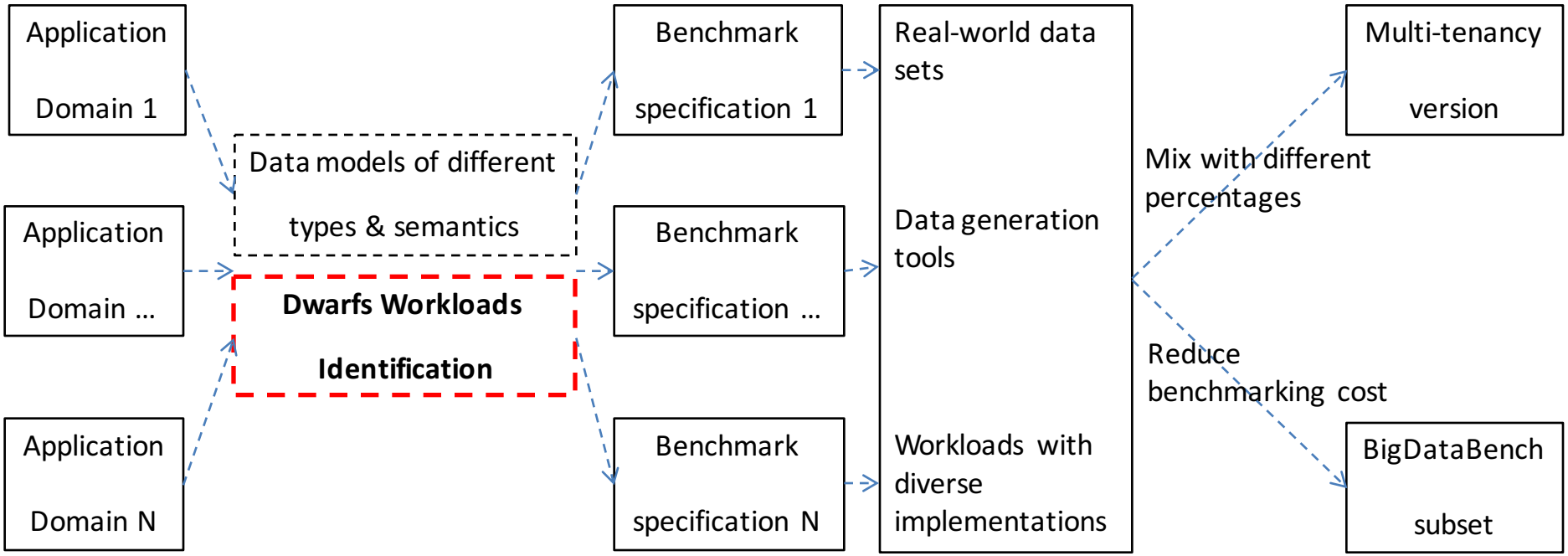
Xinhui Tian

Institute of Computing Technology, Chinese Academy of Sciences

BigDataBench Tutorial
ASPLOS 2016 Atlanta, GA

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

BigDataBench Methodology



How to define a representative big data benchmark ?

- One atte
  - Using
  - Subje
- Another attempt
  - ains of

How can we construct a benchmark suite using ***a minimum workload set*** to represent ***maximum patterns*** of big data analytics?

Dwarfs workloads !

Inspiration

Successful Compute Abstractions

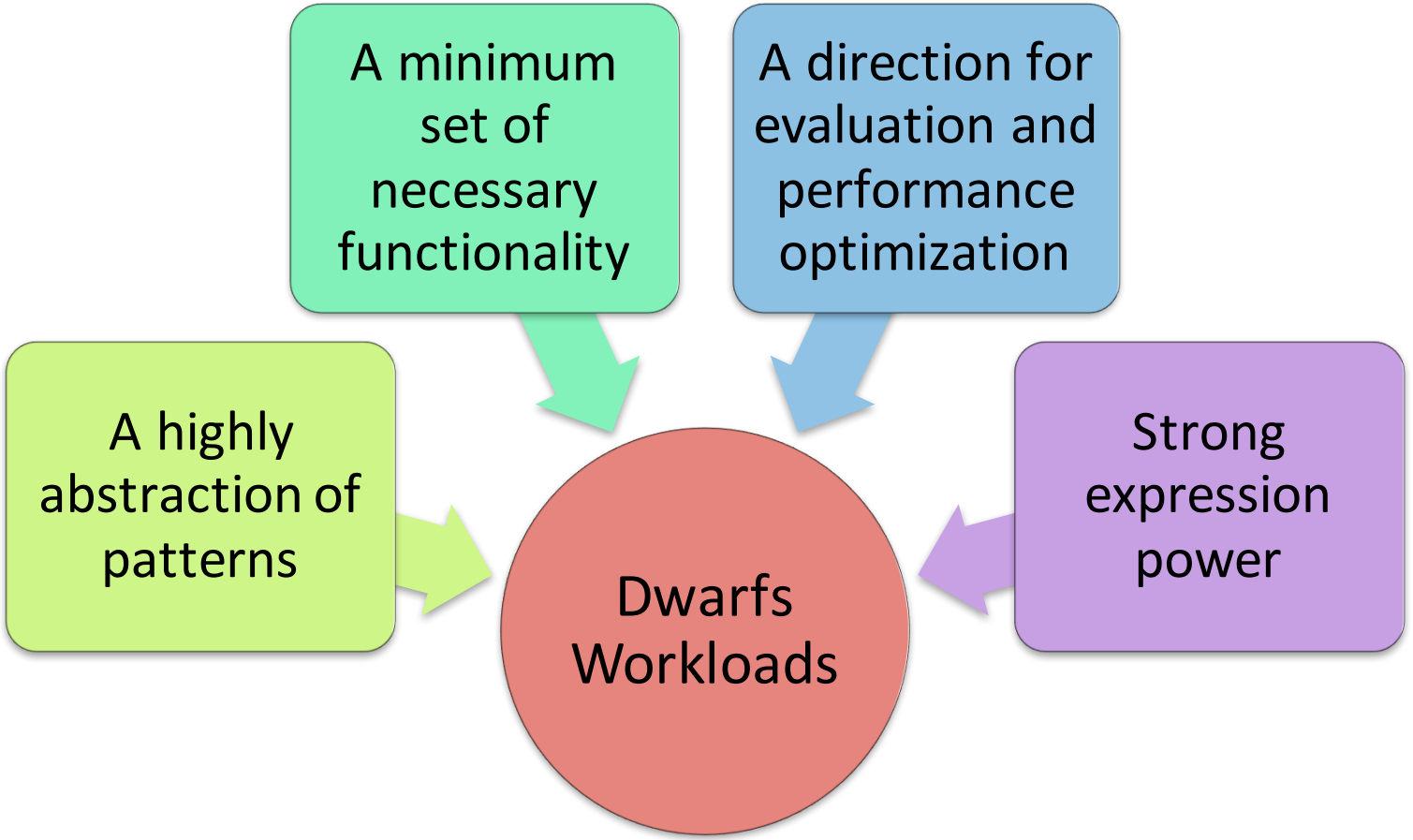
- Relational algebra
  - 5 primitive operations
  - Select, Project, Product, Union, Difference
- Parallel computing
  - Computational & communication patterns
  - 13 dwarfs

Successful Benchmarks

- TPC-C
  - OLTP domain
  - Functions of abstraction
- HPCC
  - High performance computing
  - Seven basically tests

K. Asanovic,, etc. The landscape of parallel computing research: A view from berkeley.

Why Dwarfs are Important



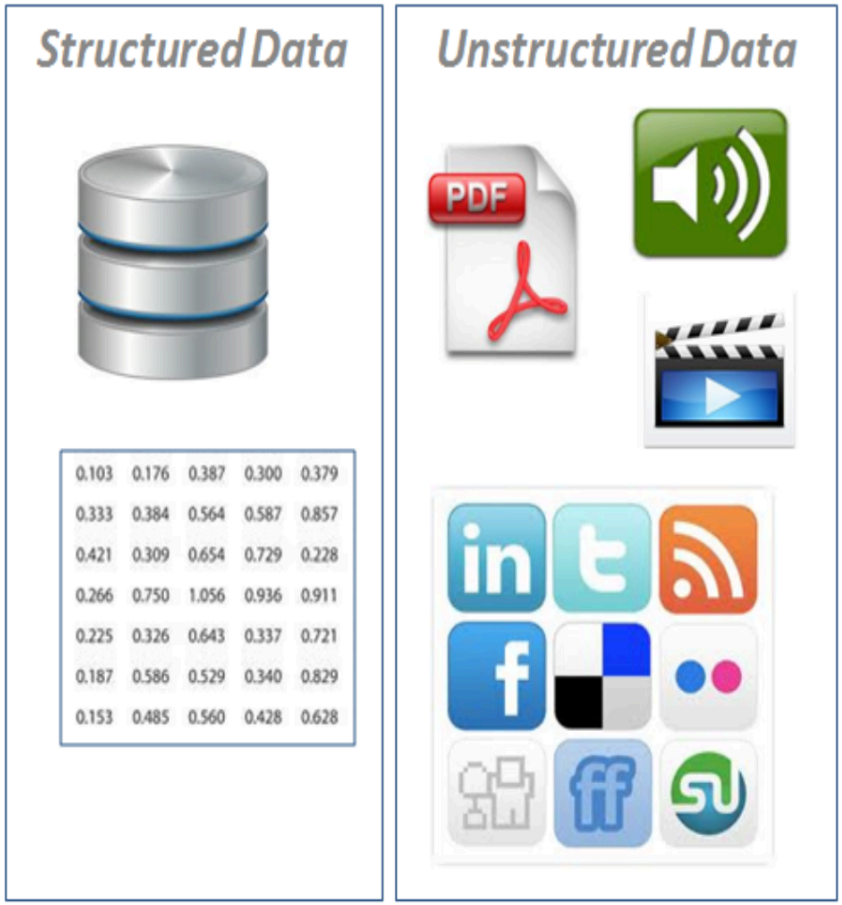
http://www.krellinst.org/doecsgf/conf/2014/pres/jhill.pdf
http://cacs.usc.edu/education/cs596/DavidPatterson.pdf

Fundamental Issues

What are the challenges for big data dwarfs?

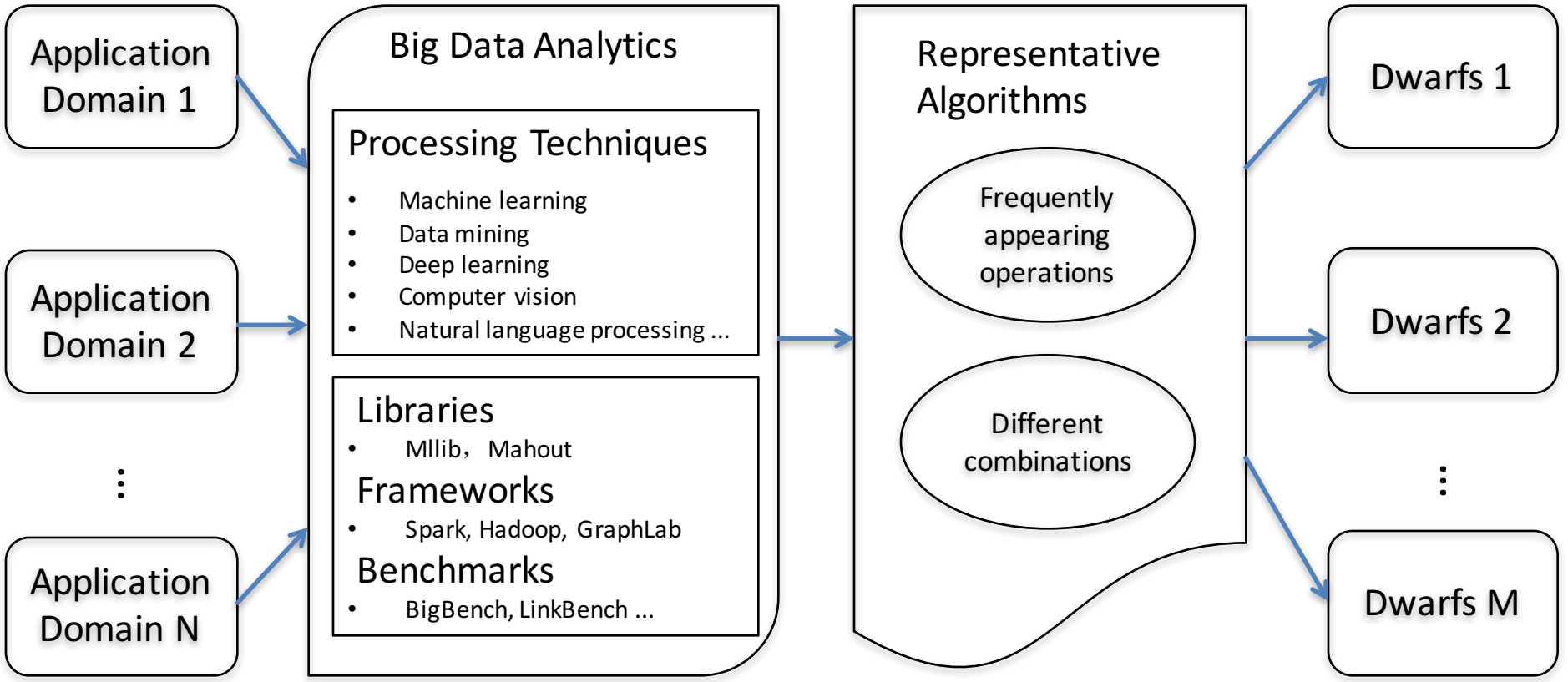
How to find the dwarfs workloads in big data analytics?

Challenges

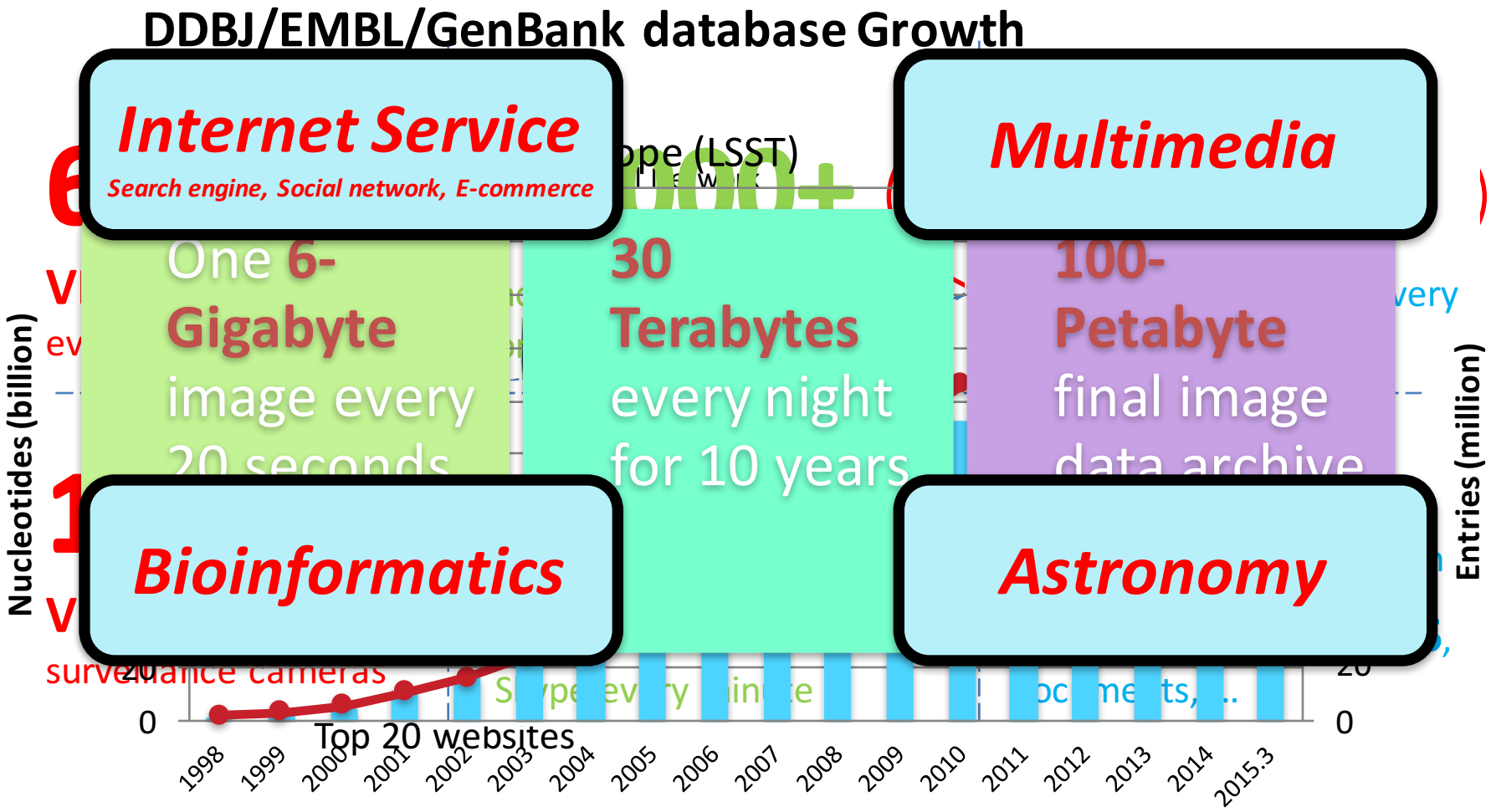


- *80%* data growth are unstructured data
- Operations on big data are *complicated*
  - Pipeline? Parallel?

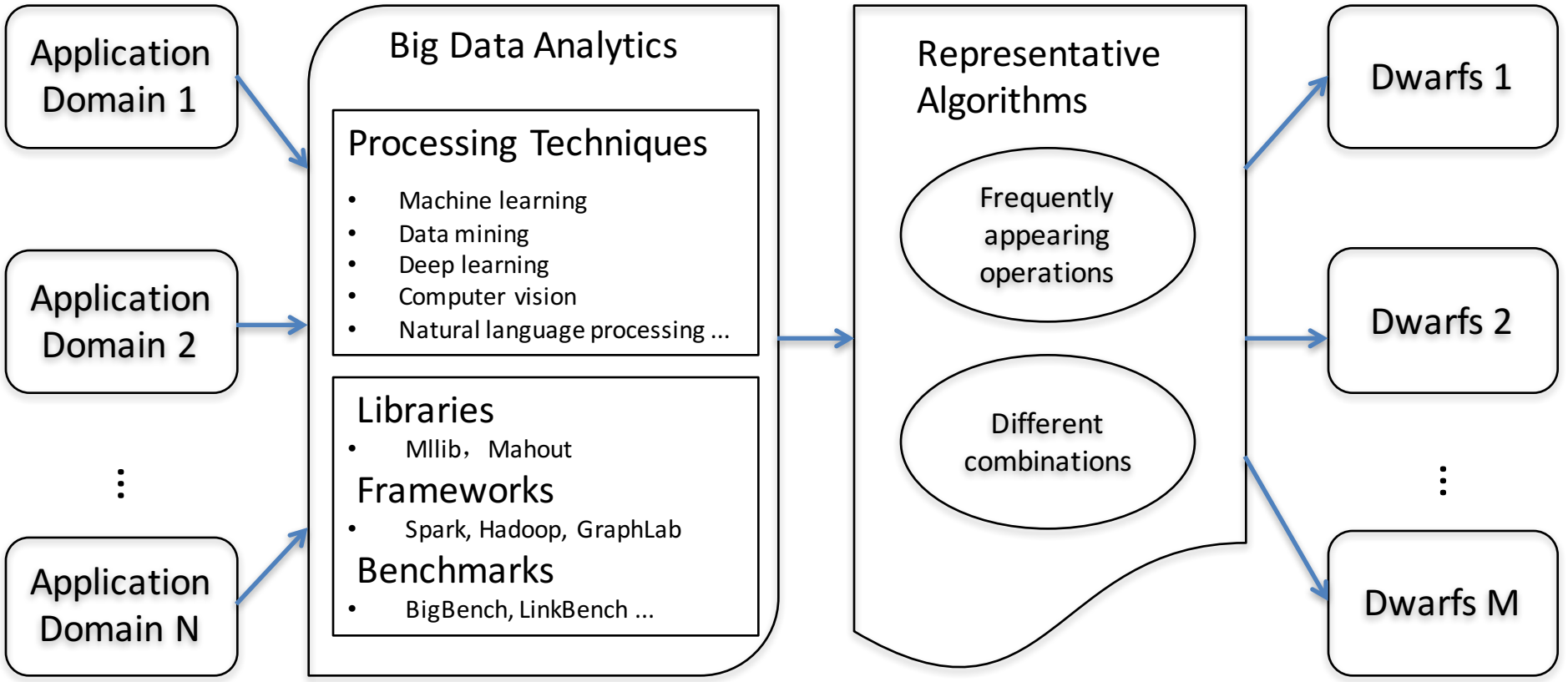
Methodology of Dwarfs



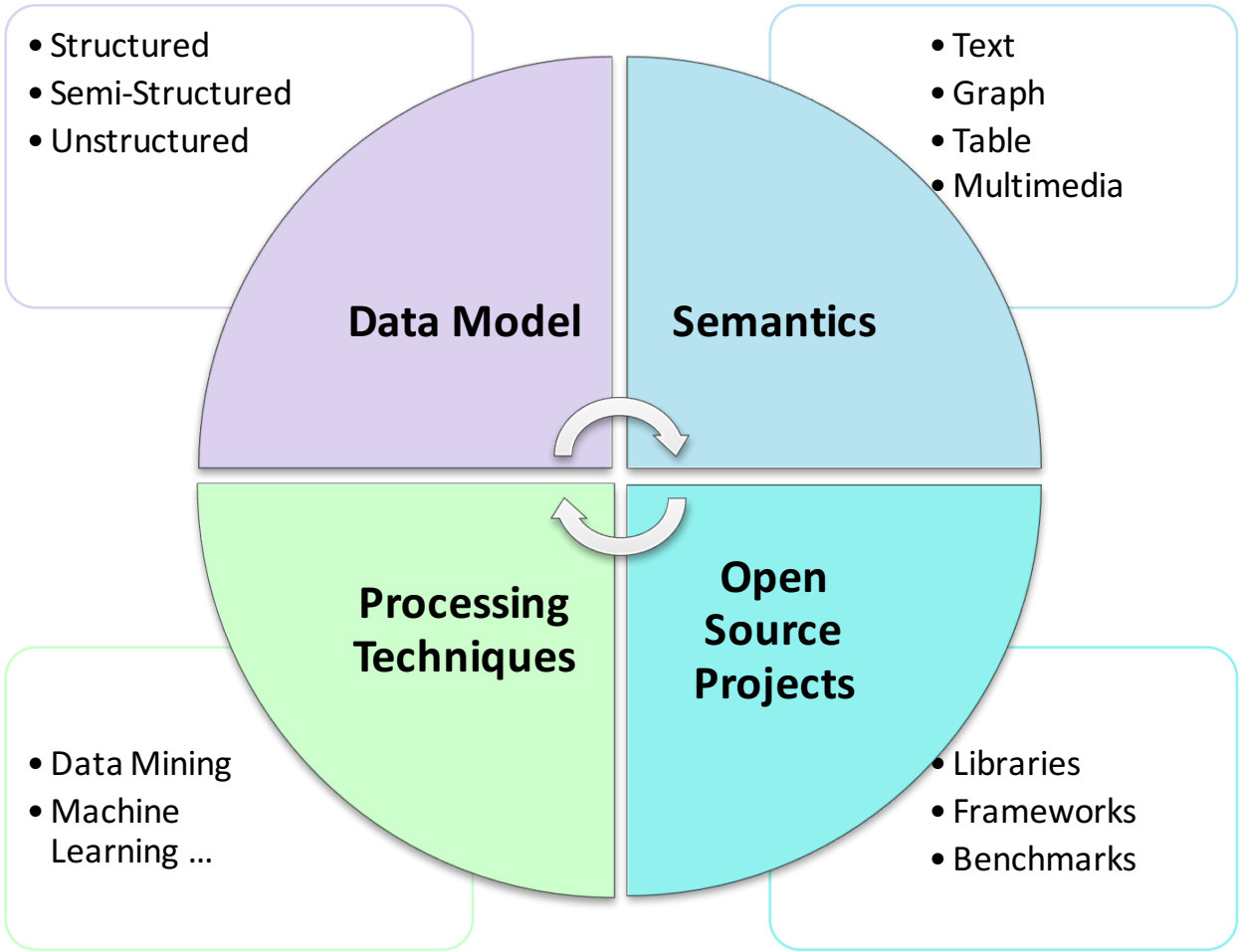
Choosing Application Domain



Methodology of Dwarfs



Big Data Analytics



Libraries & Frameworks & Benchmarks

Libraries

- Opencv, Mlib, Weka, AstroML …

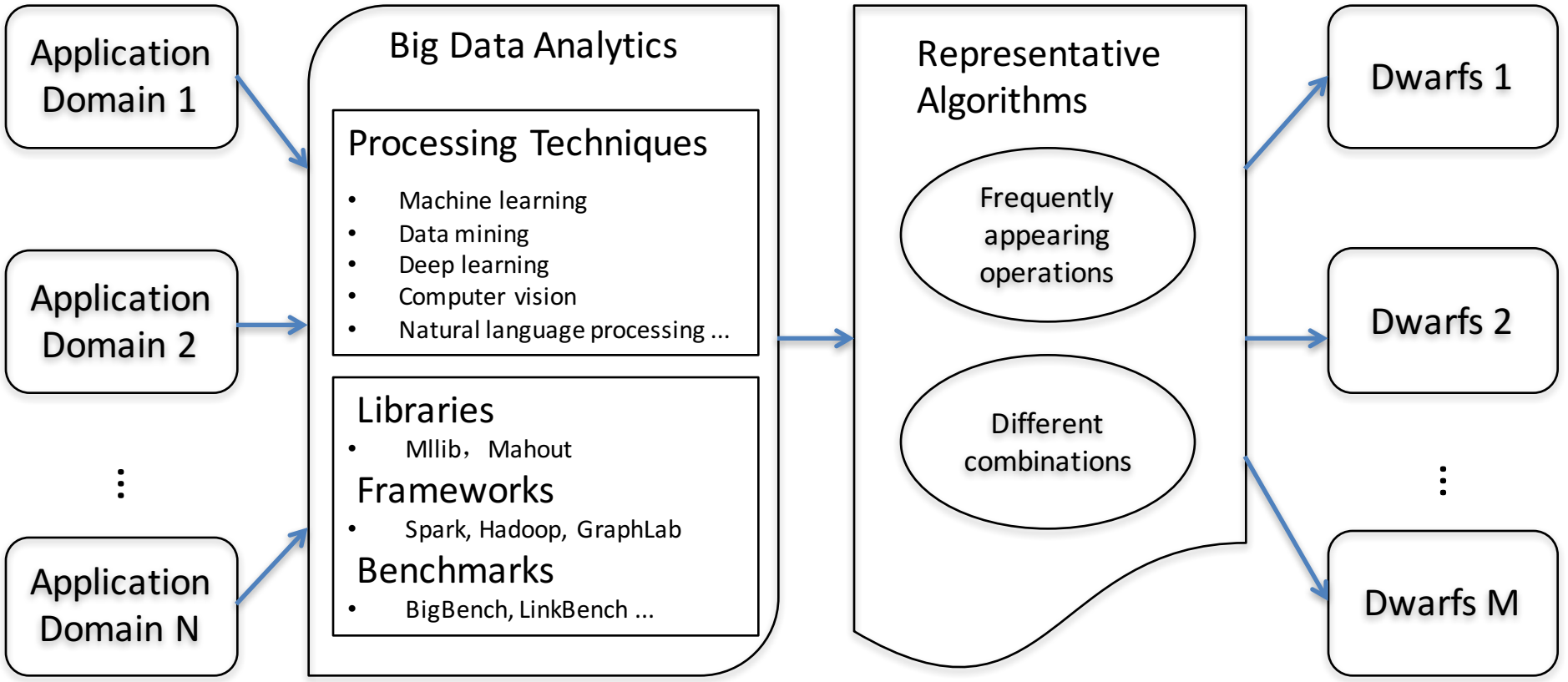
Frameworks

- Spark, Hadoop, Graphlab …

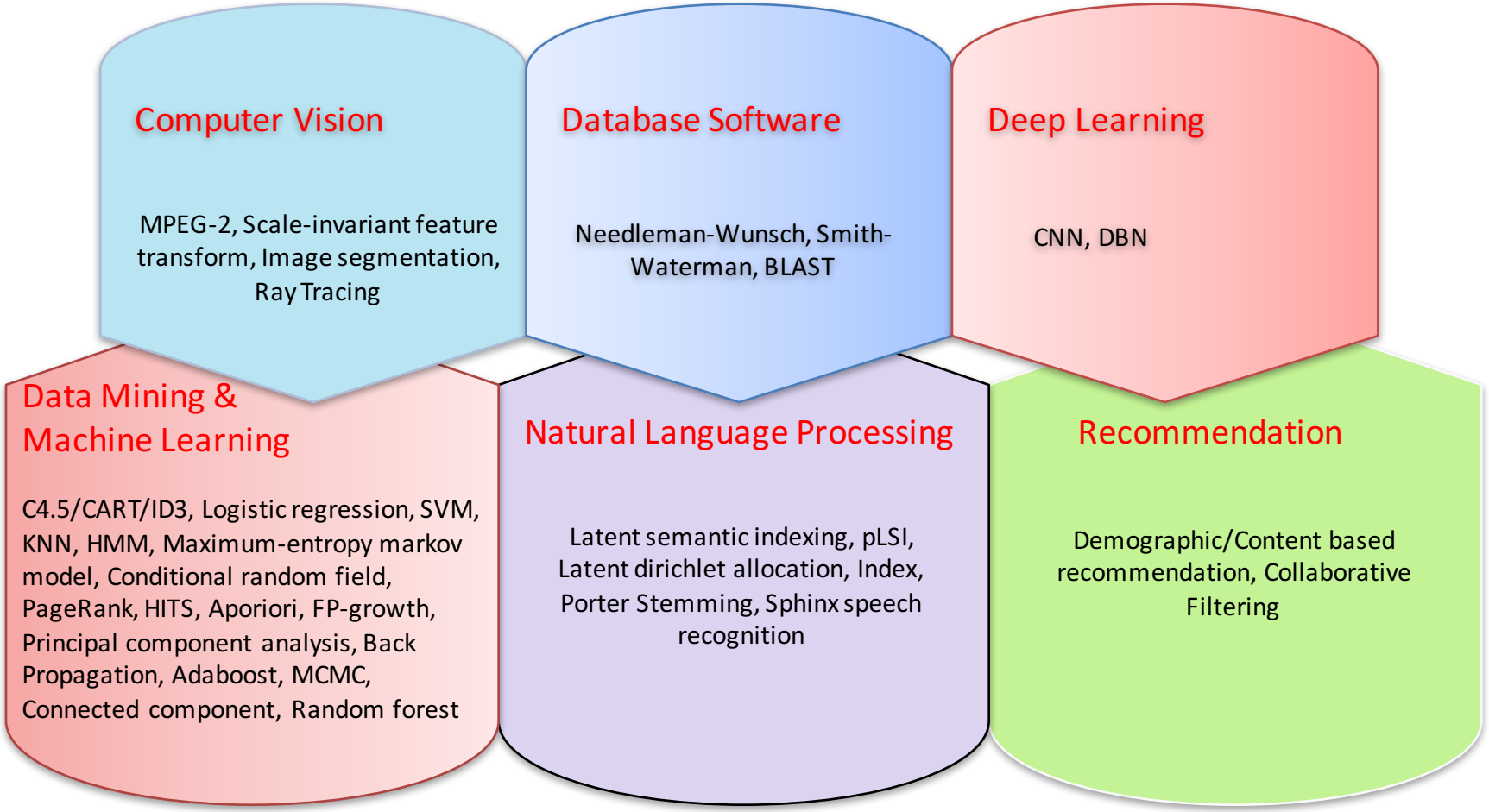
Benchmarks

- BigDataBench, Linkbench …

Methodology of Dwarfs



Algorithms Chosen to Investigate



Frequently-appearing Operations

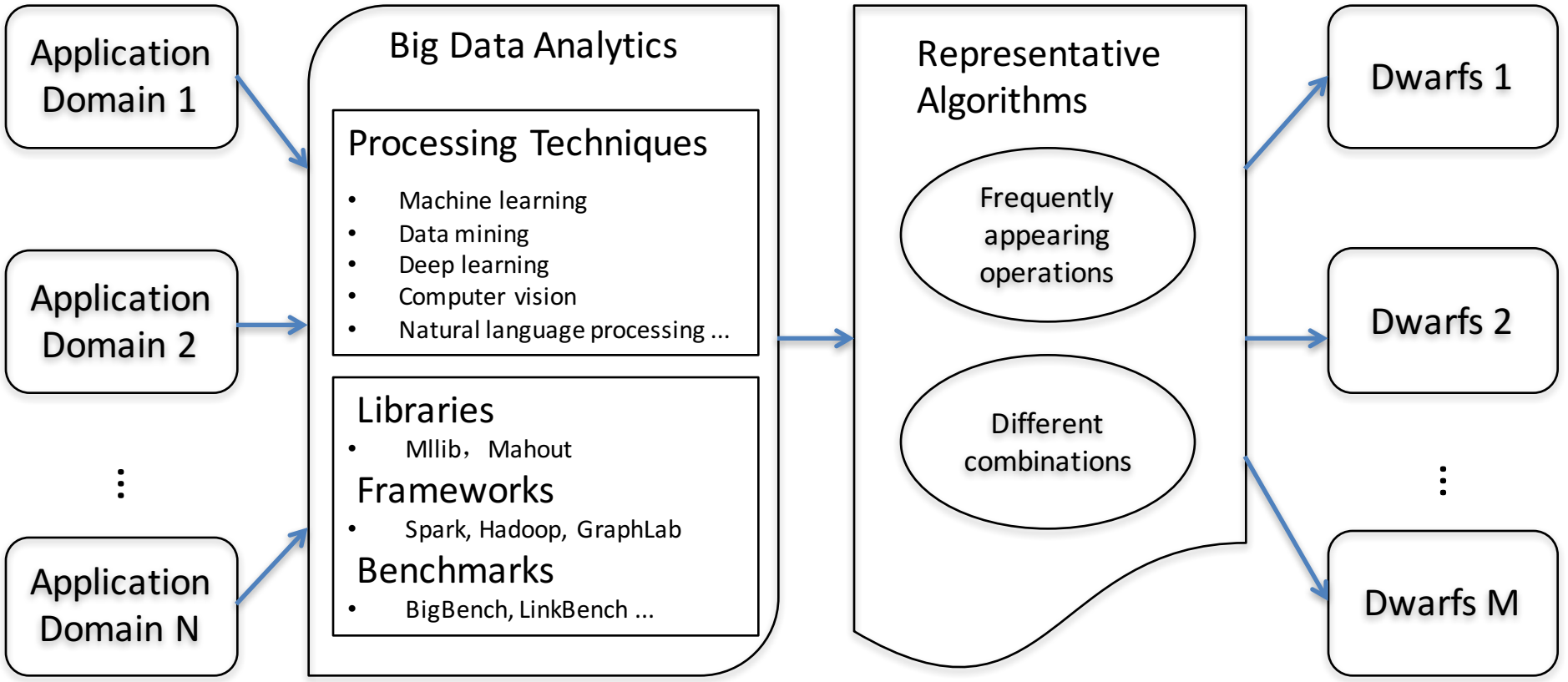
Statistic Operation

- Probability calculation
- LSI, pLSI, Latent dirichlet allocation …

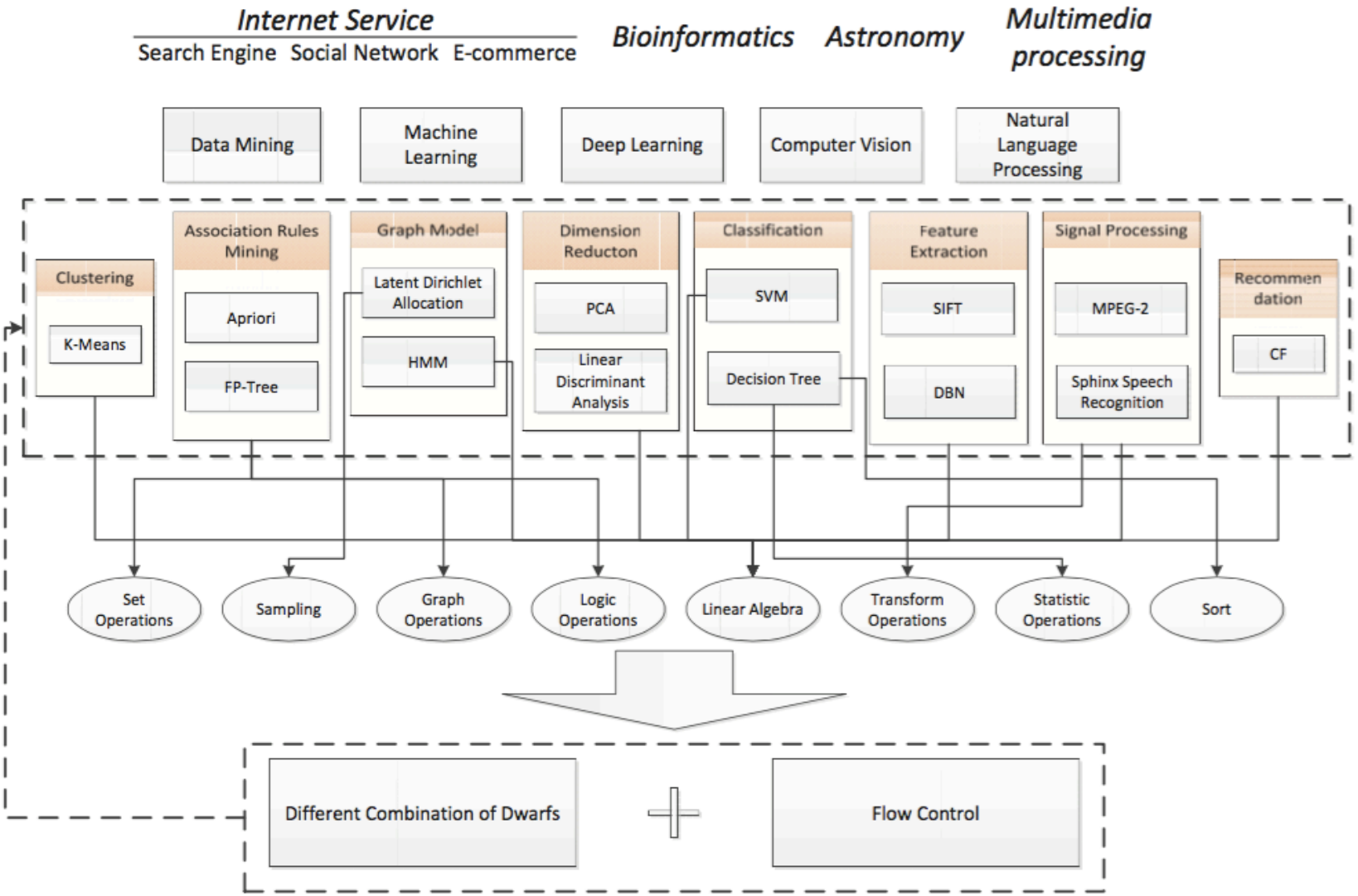
Sort

- Partial sort, quick sort, Top k sort…
- K-means, Decision tree …

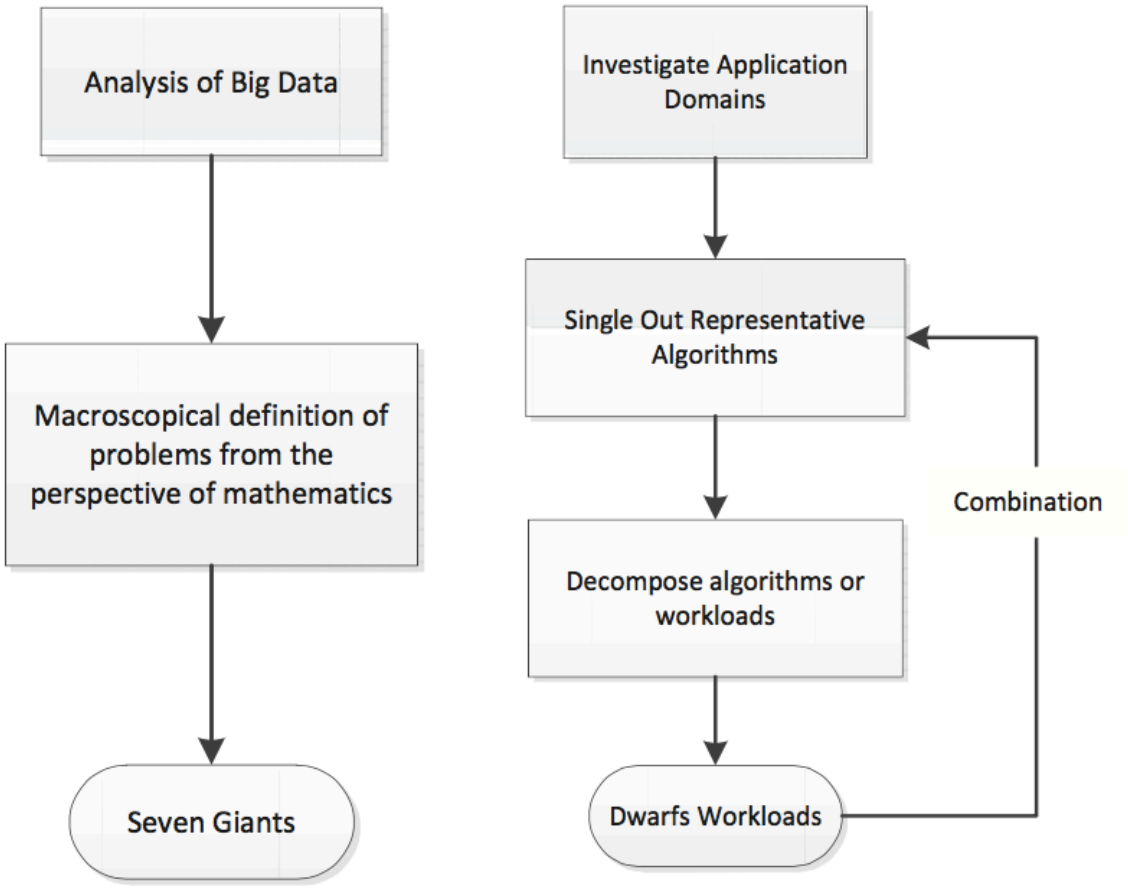
Methodology of Dwarfs



Methodology Summary



Comparison With NRC's Giants



(a) NRC Seven Giants (b) Our Eight Dwarfs

- *N. Council. Frontiers in massive data analysis. The National Academies Press Washington, DC, 2013.*

Dwarfs in Big Data Offline Analytics

- Linear Algebra
- Sampling
- Transform operation
- Graph operation
- Logic operation
- Set operation
- Statistic operation
- Sort

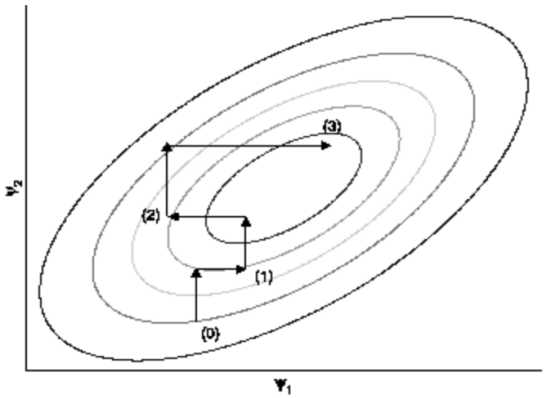
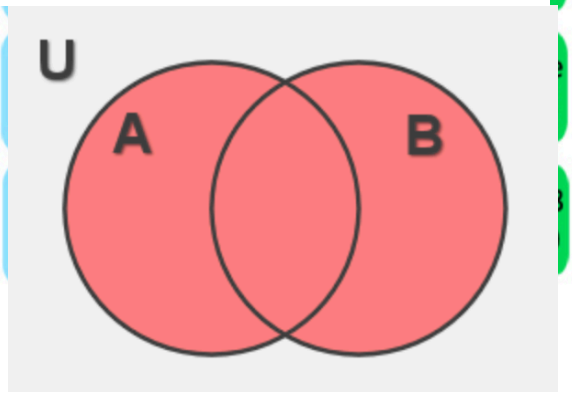


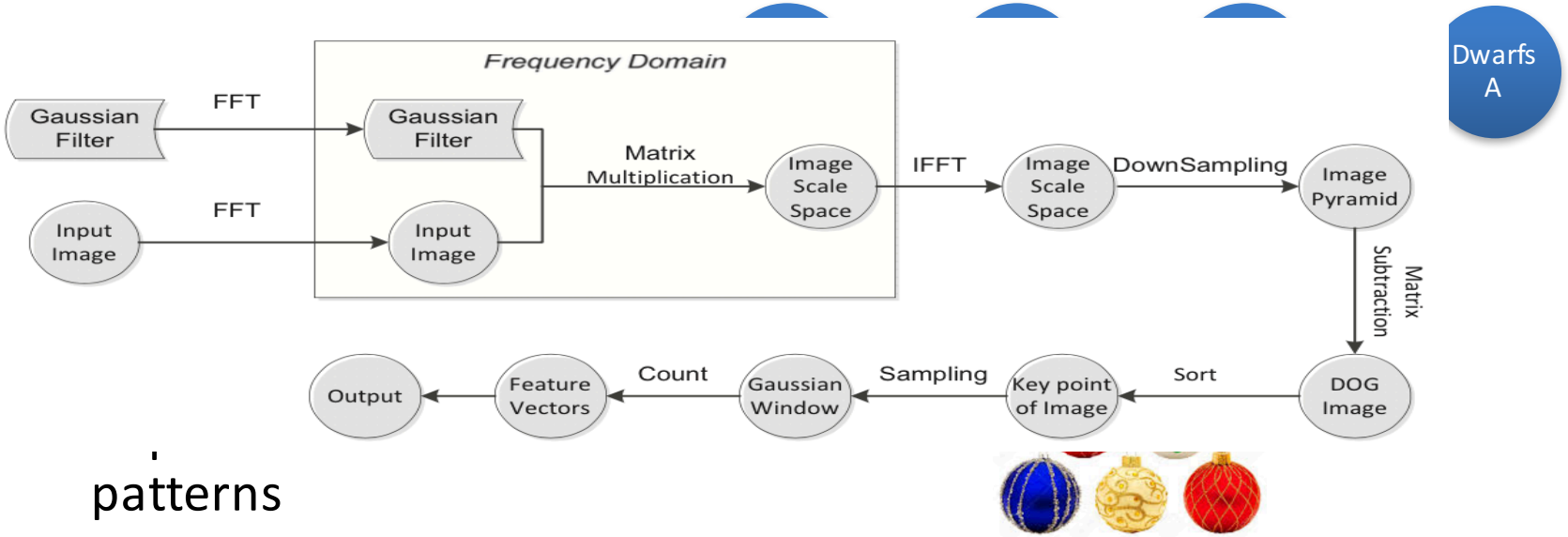
Figure 3.4: Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations



Properties

- Composability
  - One Algorithm can be composed of one or more dwarfs

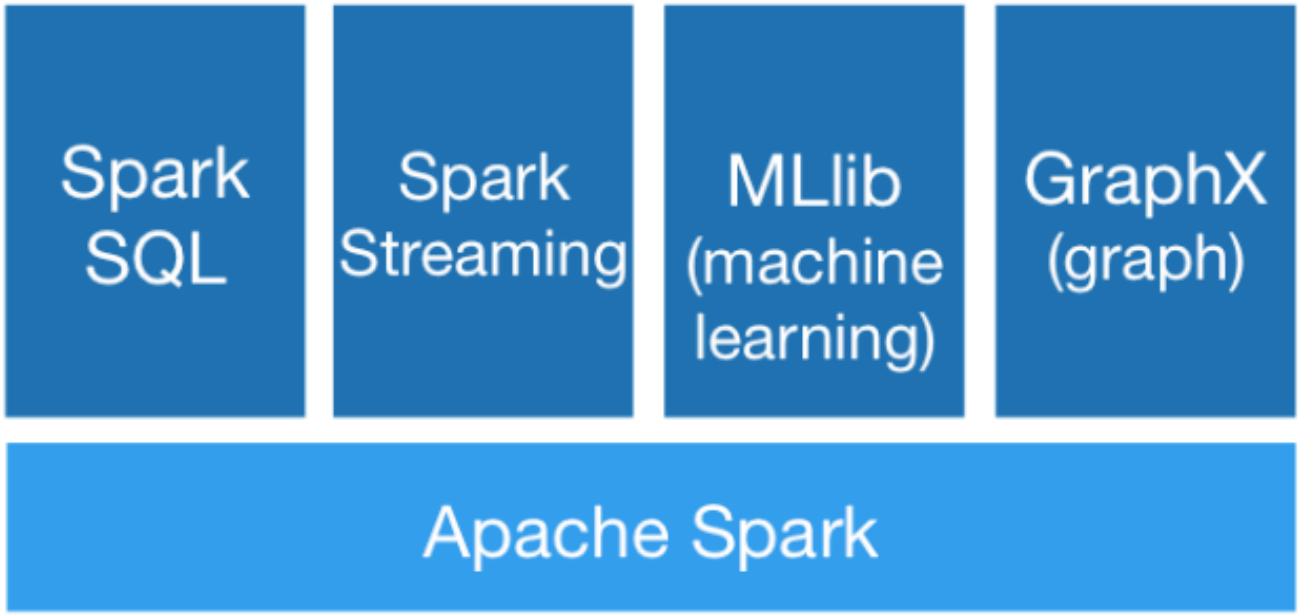
Feature extraction – SIFT Algorithm



- patterns

Overlap? Yes

- View of applications
  - Same application can be implemented with different abstractions
- Provide a method for application optimization and finding bottleneck in different levels



Thank You!

Eight Basic Dwarfs Implementation

No.	Dwarfs	Implementation
Dwarf-1	Linear Algebra	Matrix Multiplication
Dwarf-2	Sampling	MCMC for GMM
Dwarf-3	Transform Operation	FFT
Dwarf-4	Graph Operation	BFS
Dwarf-5	Logic Operation	MD5
Dwarf-6	Set Operation	Union
Dwarf-7	Sort	QuickSort
Dwarf-8	Statistic Operation	Count