BigDataBench Tutorial

Jianfeng Zhan (1), Gang Lu (2), and Xinhui Tian(1)

http://prof.ict.ac.cn/BigDataBench

(1)ICT, Chinese Academy of Sciences
(2)Beijing Academy of Frontier Science and Technology

ASPLOS 2016 ATLANTA, GEOGRIA, USA

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

Acknowledgements

- BigDataBench contributors


Lei Wang, Chunjie Luo, Zhen Jia, Dr. Rui Han, Wanling Gao, Xinhui Tian, Gang Lu


Gang Lu, Jingwei Li


Shujie Zhang, Dr. Chuliang Weng


Xiaona Li


Bizhu Qiu


Kent Zhan, Zijian Ming

BigDataBench Tutorial Program (1)

- 8:30-9:00 Jianfeng Zhan
  - What is BigDataBench?
- 9:00-9:30 Gang Lu
  - BigDataBench data sets and workloads
  - How to use BigDataBench
- 9:30-10:00 Gang Lu
  - How to generate Large-scale data sets?

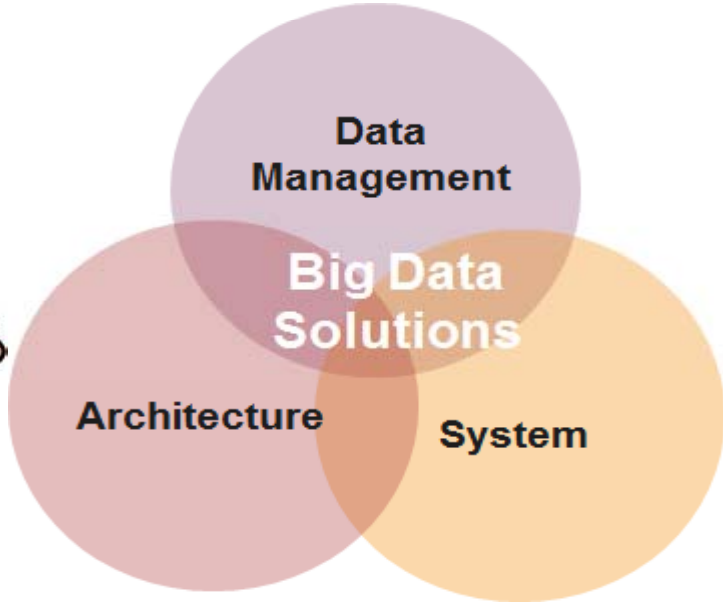
Program (2)

- 10:00-10:20 Coffee break
- 10:20-10:50 Gang Lu
  - Multi-tenancy version of BigDataBench
- 10:50-11:20 Xinhui Tian
  - Subsetting and characterizing big data workloads
- 11:20-11:50 Xinhui Tian
  - BigDataBench Dwarfs

First part

- *What is BigDataBench?*
- BigDataBench benchmarking methodology

Why Big Data Benchmarking?



Measuring big data systems and architectures quantitatively

What is *BigDataBench*?

- An open source big data benchmarking project
  - http://prof.ict.ac.cn/BigDataBench
  - Search Google using "**BigDataBench**"

BigDataBench 3.2 Overview

BDGS(Big Data Generator Suite) for scalable data

Wikipedia Entries	Amazon Movie Reviews	Google Web Graph
Facebook Social Network	E-commerce Transaction	ProfSearch Resumes
ImageNet	English broadcasting audio	DVD Input Streams
Image scene	Genome sequence data	Assembly of the human genome
SoGou Data	MNIST	MovieLens Dataset

15 Real-world Data Sets

Search Engine	Social Network	E-commerce
Multimedia	Bioinformatics	

37 Workloads

Impala Flink JStorm Alibaba NoSql

GraphX Shark GraphLab Carnegie Mellon HIVE ORACLE MySQL

Hadoop RDMA MVAPICH MPI

Spark Lightning-Fast Cluster Comp hadoop DataMPI

Software Stacks

What's New in BigDataBench 3.2

BigDataBench support for Flink

- WordCount, Grep, Naïve Bayes, PageRank, K-means

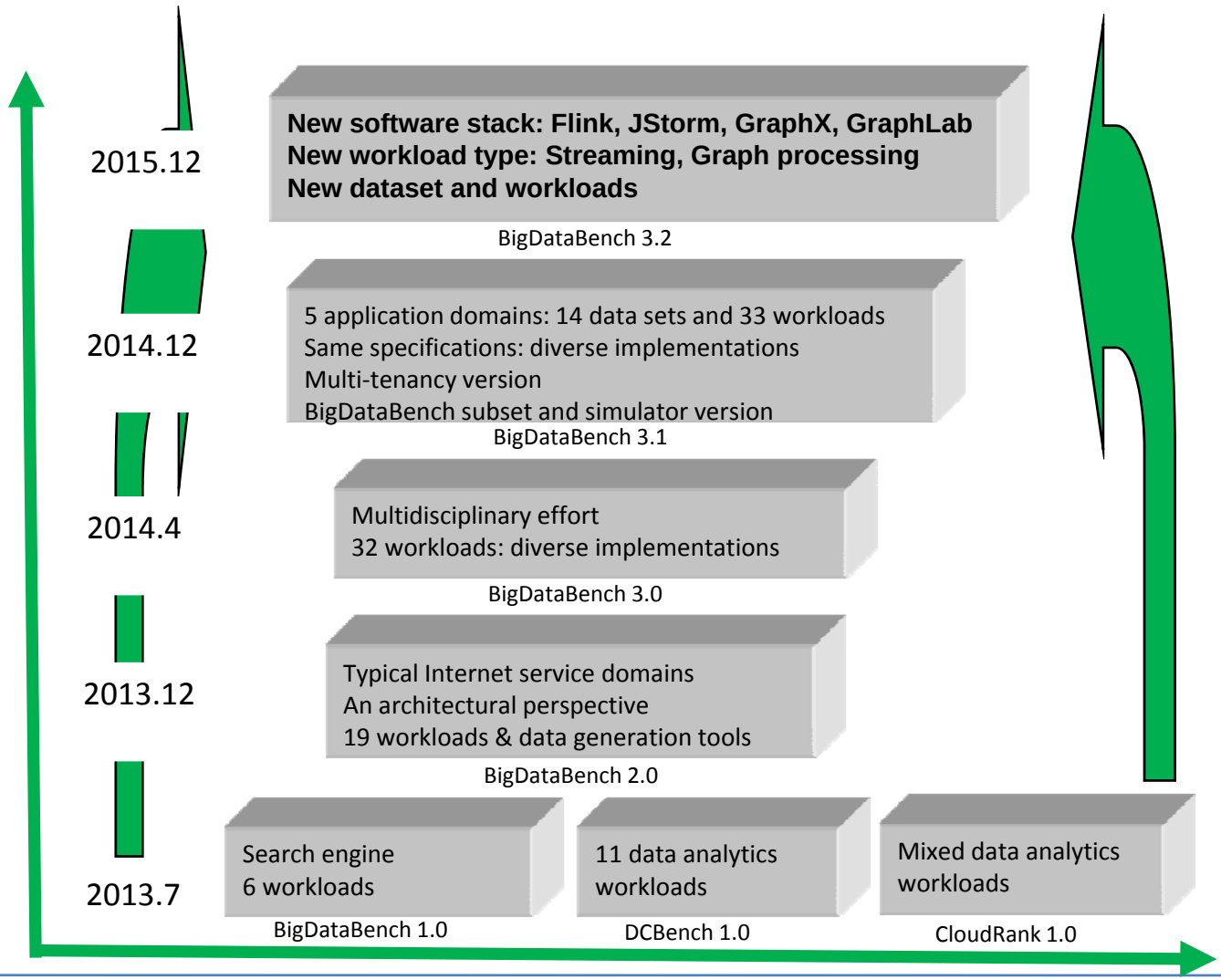
Streaming

- JStorm, Spark Streaming

Graph processing

- GraphX, GraphLab, Flink Gelly

BigDataBench evolution



BigDataBench Users

- http://prof.ict.ac.cn/BigDataBench/users/
- Industry users
  - Accenture, BROADCOM, SAMSUMG, Huawei, IBM
- About 100 academia groups published papers using or citing BigDataBench
  - VLDB/SIGMOD, SC, FAST, ASPLOS, ISCA/Micro/ HPCA, ICPP and etc.

Industry Standard: BigDataBench-DCA

- China's first industry-standard big data benchmark suite
  - http://prof.ict.ac.cn/BigDataBench/industry-standard-benchmarks/
  - Telecom Research Institute of Ministry of Industry and Information Technology, ICT, CAS, Huawei, China Mobile, Sina, ZTE, Intel (China), Microsoft (China), IBM CDL, Baidu, INSPUR , ZTE, 21viane and UCloud

Why BigDataBench?

	Specification	Application domains	Workload Types	Work loads	Scalable data sets (from real data)	Multiple implementations	Multitenancy	Subsets	Simulator version
BigDataBench	Y	Five	Six	37	8	Y	Y	Y	Y
BigBench	Y	One	Three	10	3	N	N	N	N
Cloud-Suite	N	N/A	Two	8	3	N	N	N	N
HiBench	N	N/A	Two	10	3	N	N	N	N
CALDA	Y	N/A	One	5	N/A	Y	N	N	N
YCSB	Y	N/A	One	6	N/A	Y	N	N	N
LinkBench	Y	N/A	One	10	1	Y	N	N	N
AMP Benchmarks	N	N/A	One	4	N/A	Y	N	N	N

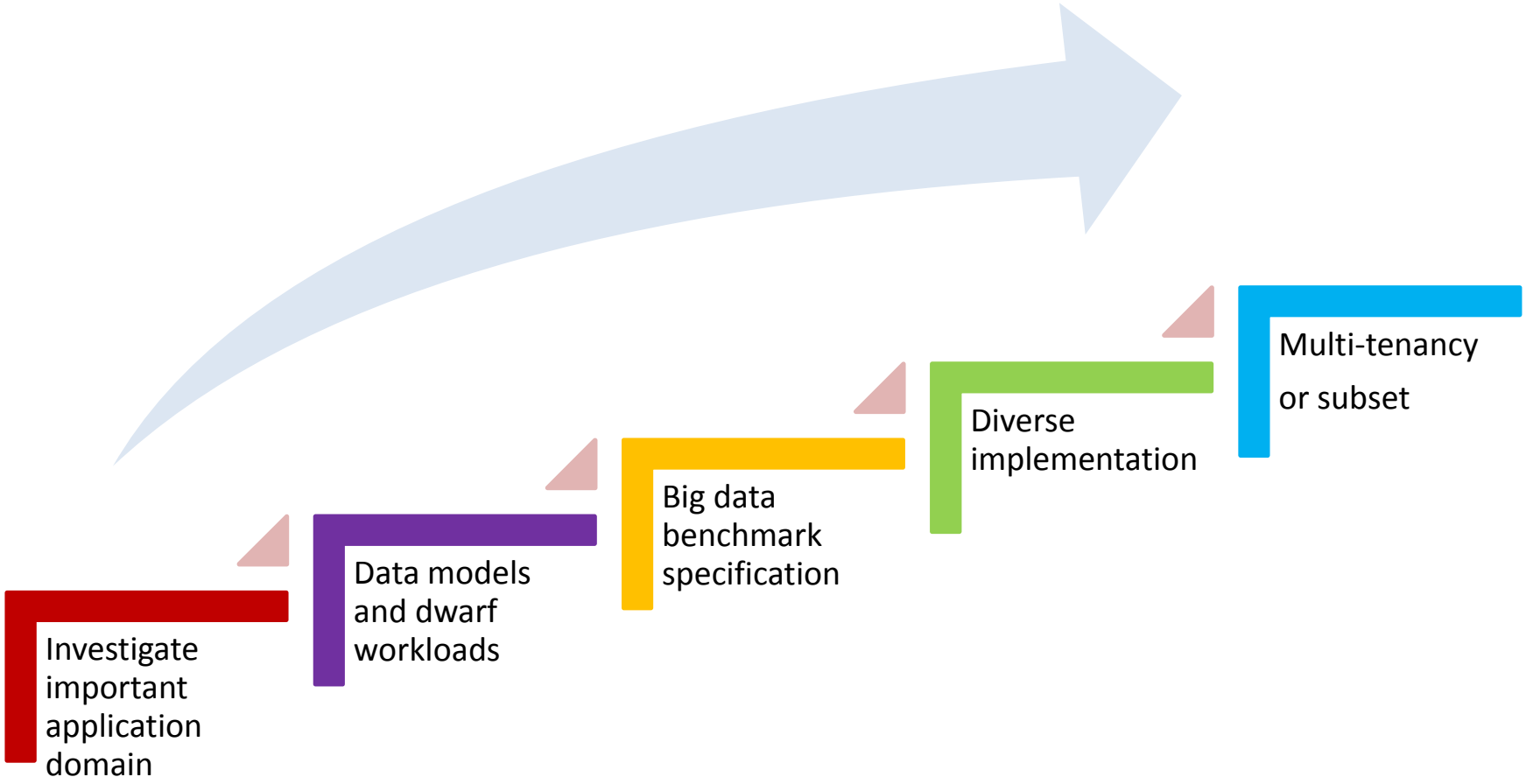
BigDataBench Publications

- BigDataBench: a Big Data Benchmark Suite from Internet Services. 20th IEEE International Symposium On High Performance Computer Architecture (**HPCA-2014**).
- Characterizing data analysis workloads in data centers. 2013 IEEE International Symposium on Workload Characterization (**IISWC 2013**)（Best paper award）
- BigOP: generating comprehensive big data workloads as a benchmarking framework. 19th International Conference on Database Systems for Advanced Applications (**DASFAA 2014**)
- BDGS: A Scalable Big Data Generator Suite in Big Data Benchmarking. The Fourth workshop on big data benchmarking (WBDB 2014)
- Identifying Dwarfs Workloads in Big Data Analytics arXiv preprint arXiv:1505.06872
- BigDataBench-MT: A Benchmark Tool for Generating Realistic Mixed Data Center Workloads arXiv preprint arXiv:1504.02205

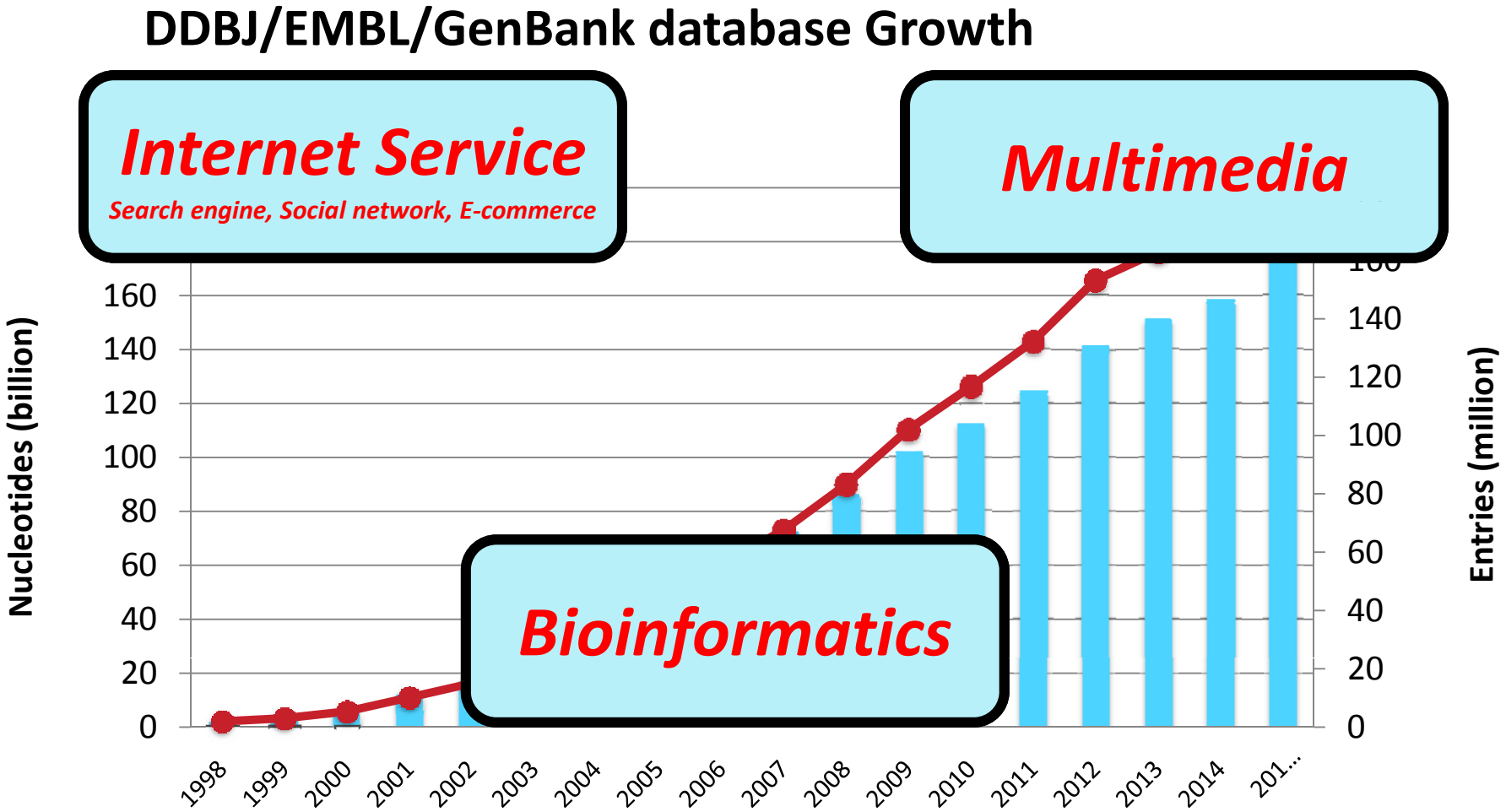
First part

- What is BigDataBench?
- *BigDataBench benchmarking methodology*

Five Steps

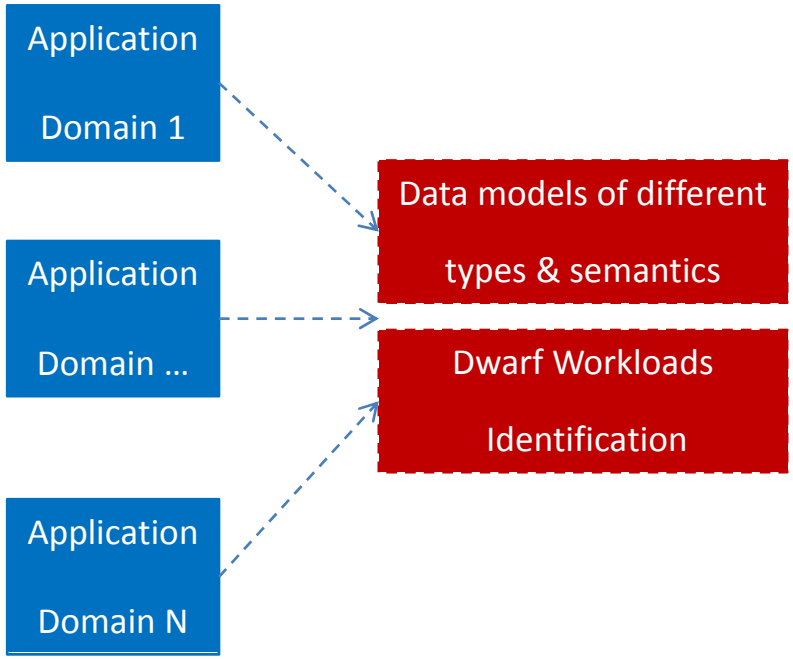

Investigate important application domain
Data models and dwarf workloads
Big data benchmark specification
Diverse implementation
Multi-tenancy or subset

Five Application Domains



http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html#dbgrowth-graph

BigDataBench Methodology



Identify Dwarf Workloads in Big Data Analytics

How to define a representative big data benchmark ?

- One att… …mpt
  - Using … …ains of
  - Subje…

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

Dwarf workloads !

Inspiration

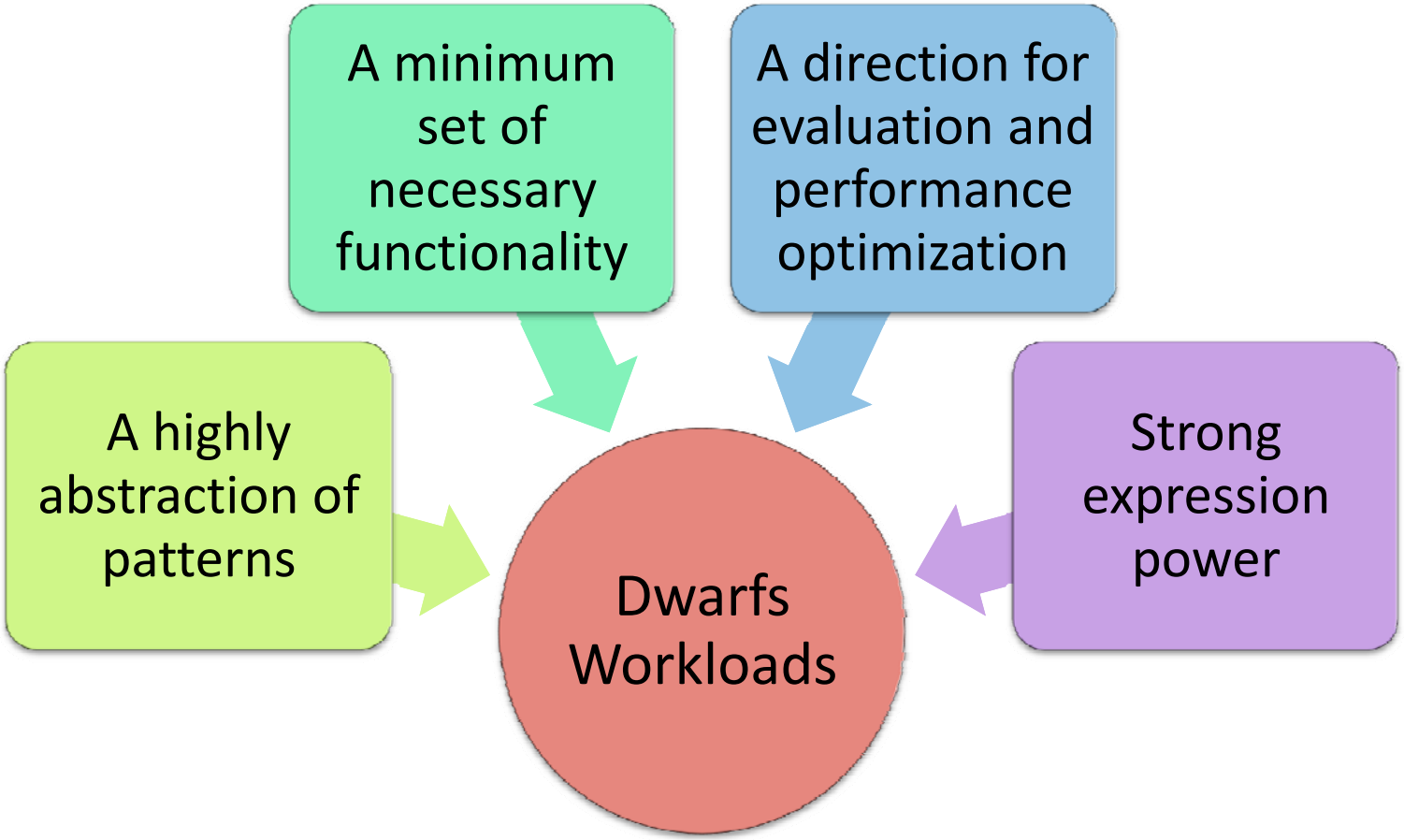
Successful Compute Abstractions

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

Successful Benchmarks

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

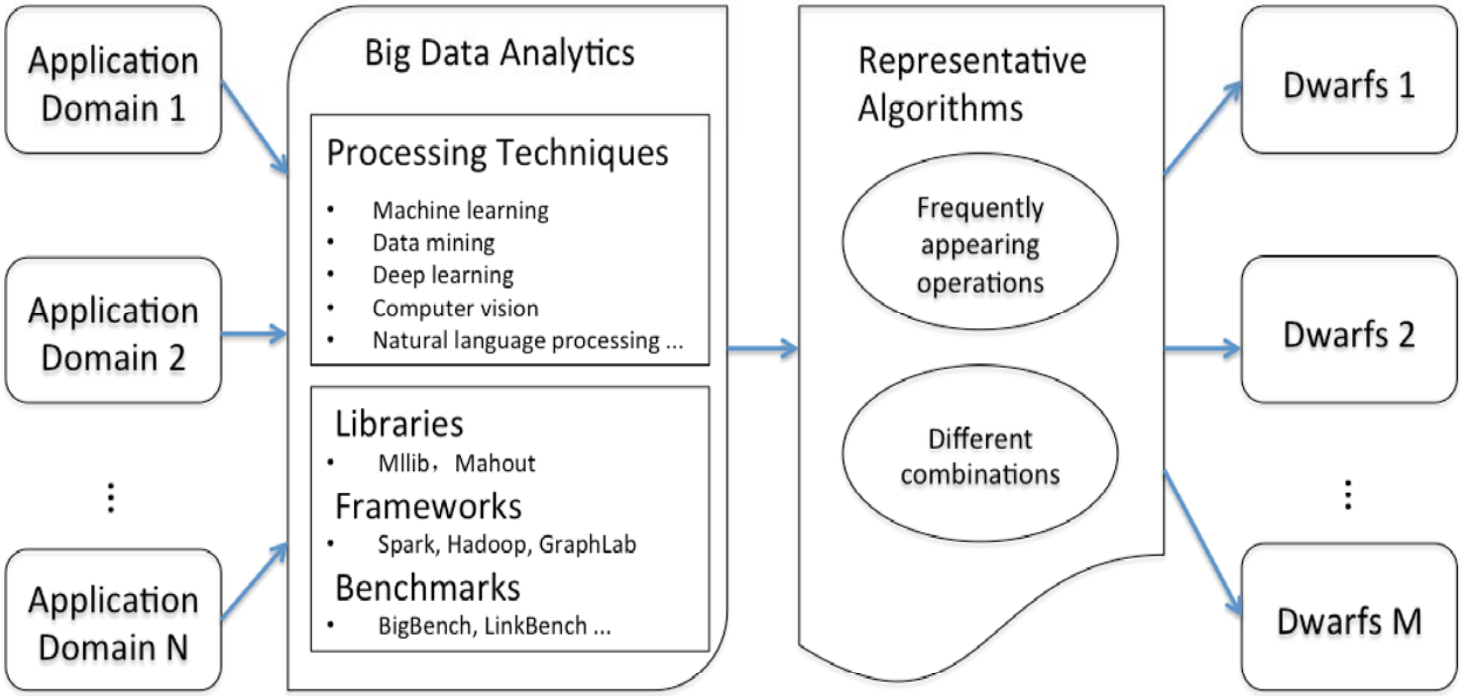
Why Dwarfs are Important



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

Dwarfs in Big data analytics

- ***A minimum set*** to represent ***maximum patterns*** of big data analytics



Our identified Big data Dwarfs

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

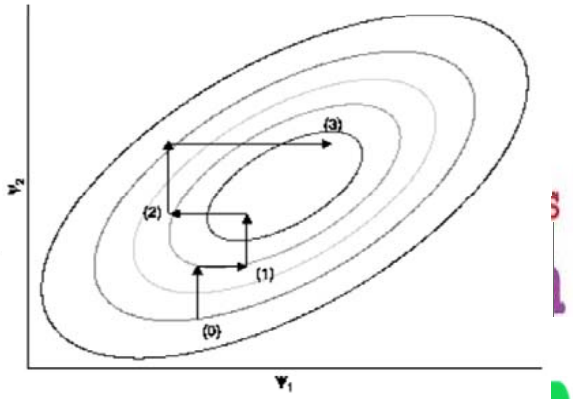


Figure 3.4: Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations

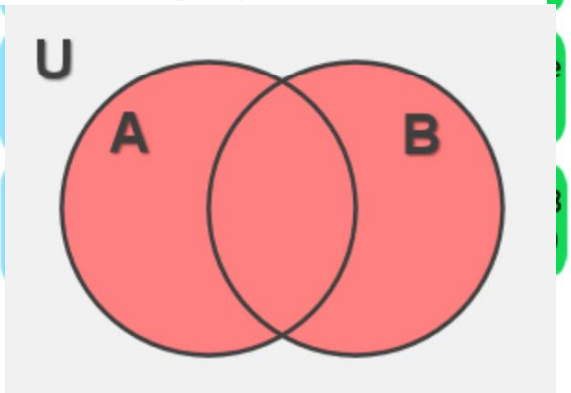
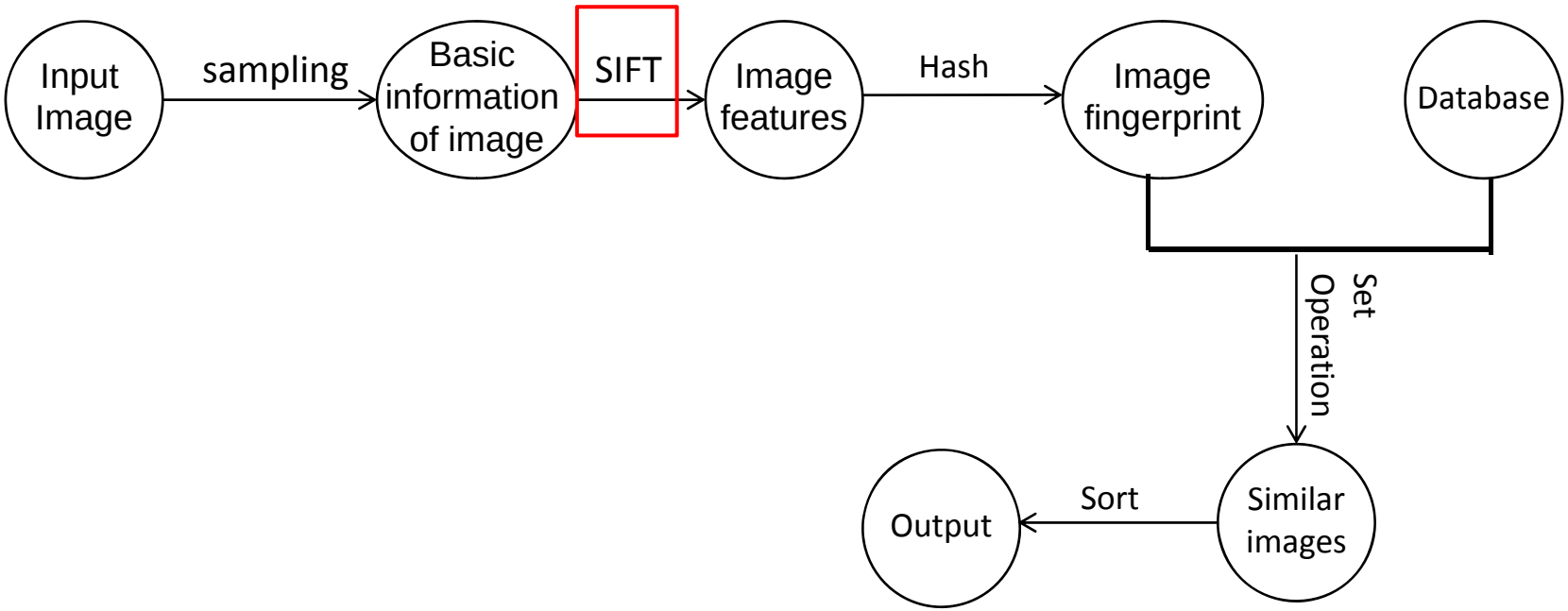
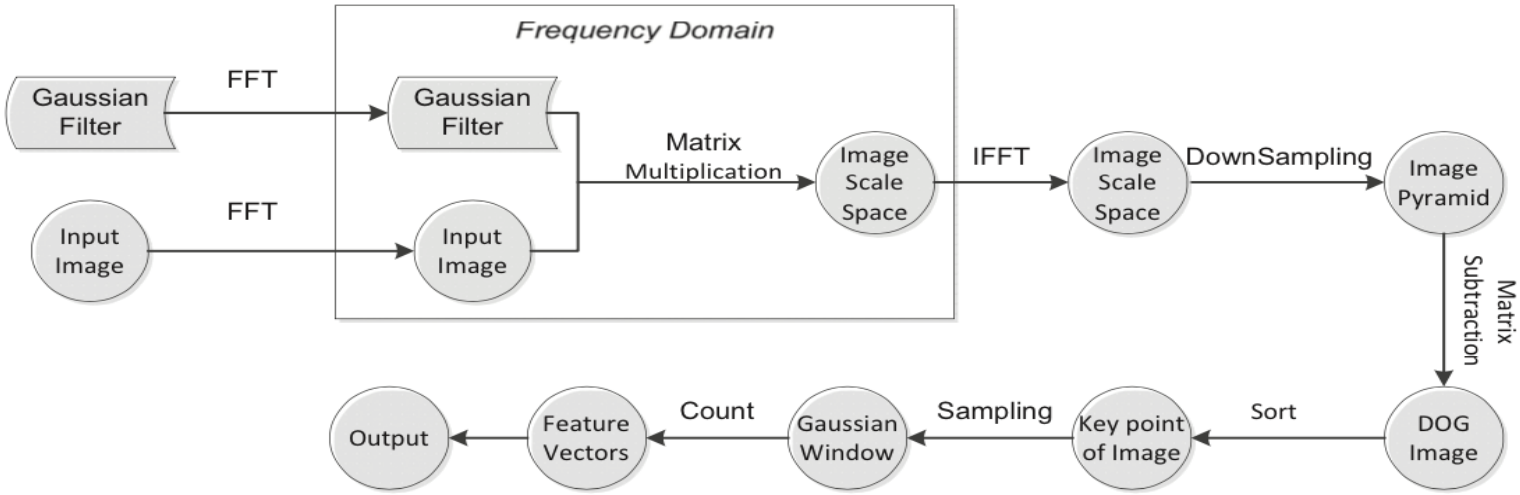


Image Search

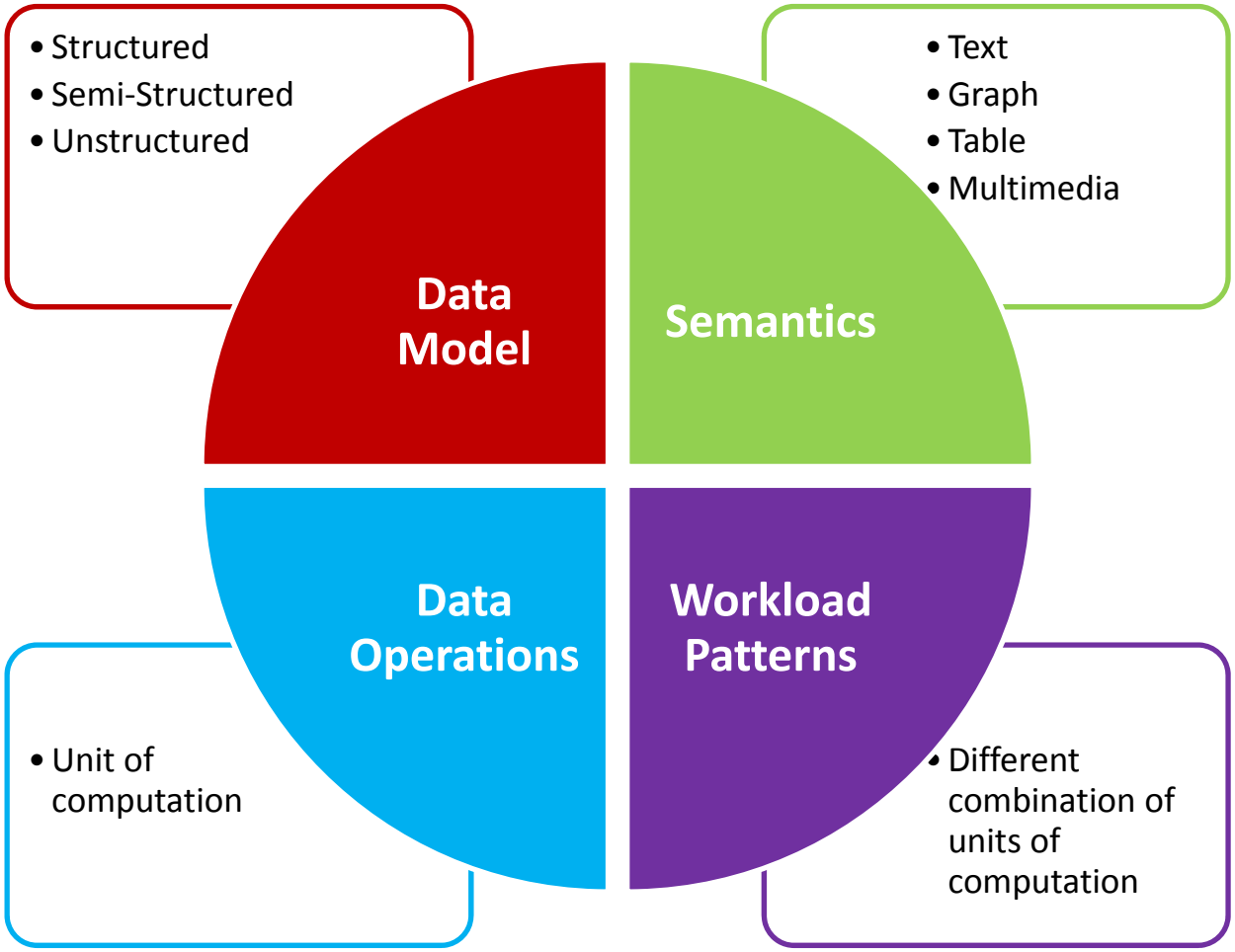
- Perceptual Hash Algorithm



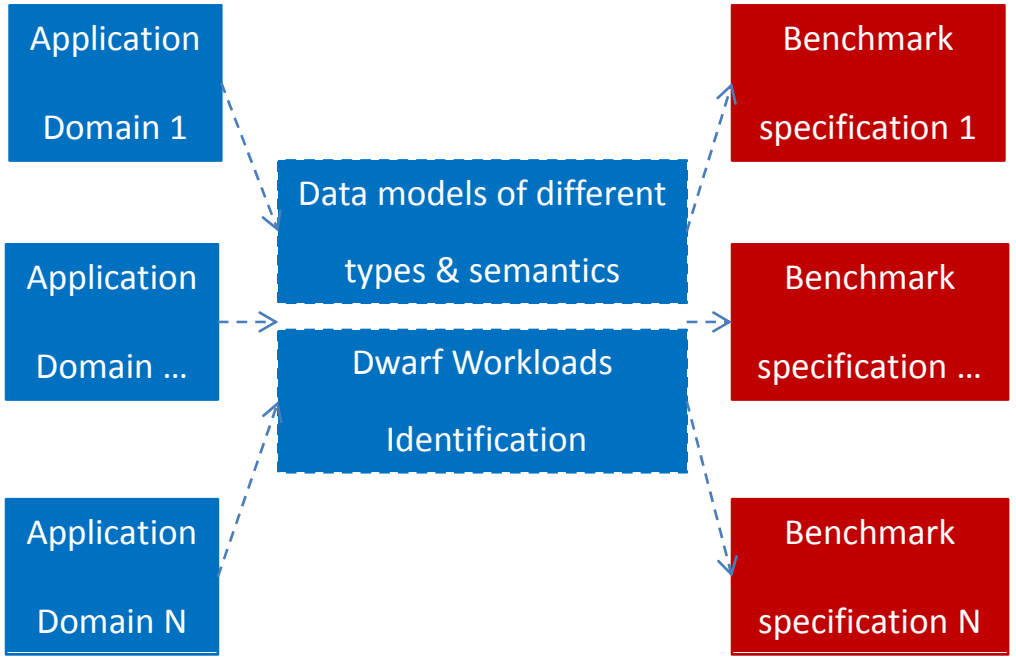
Feature Extraction--SIFT



Workload & Data Set



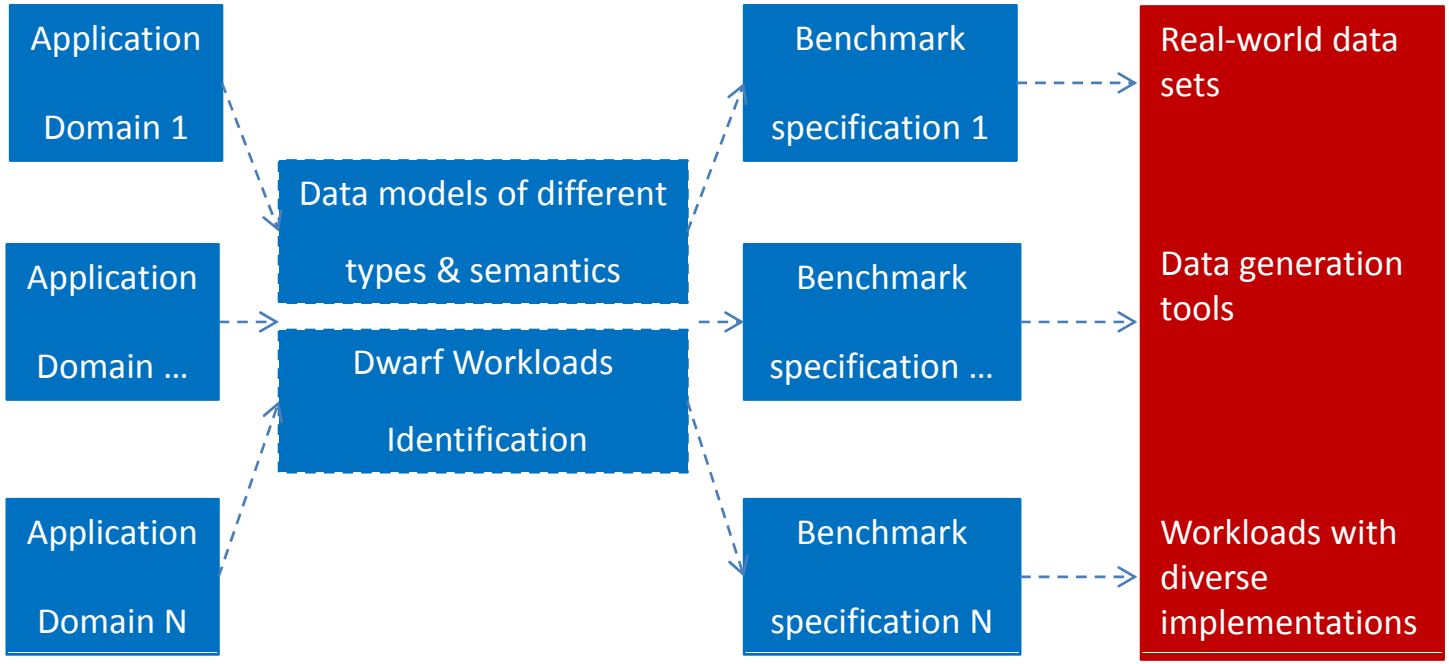
BigDataBench Methodology



Data management's tradition

- Specification First.
- Functions of abstraction are units of computation that appear frequently in the application domain being benchmarked.
- They are expressed in a generic form that is independent of the underlying system implementation.

BigDataBench Methodology



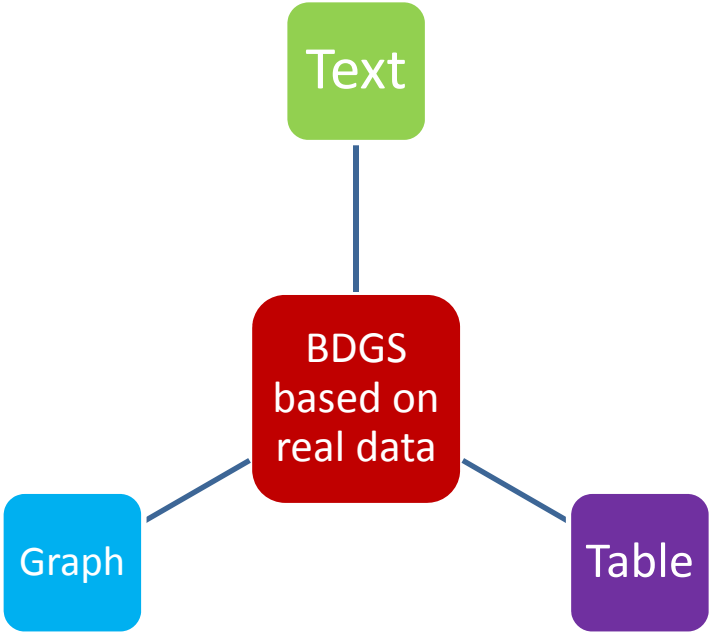
Real-World Data sets

- 15 real-world data sets

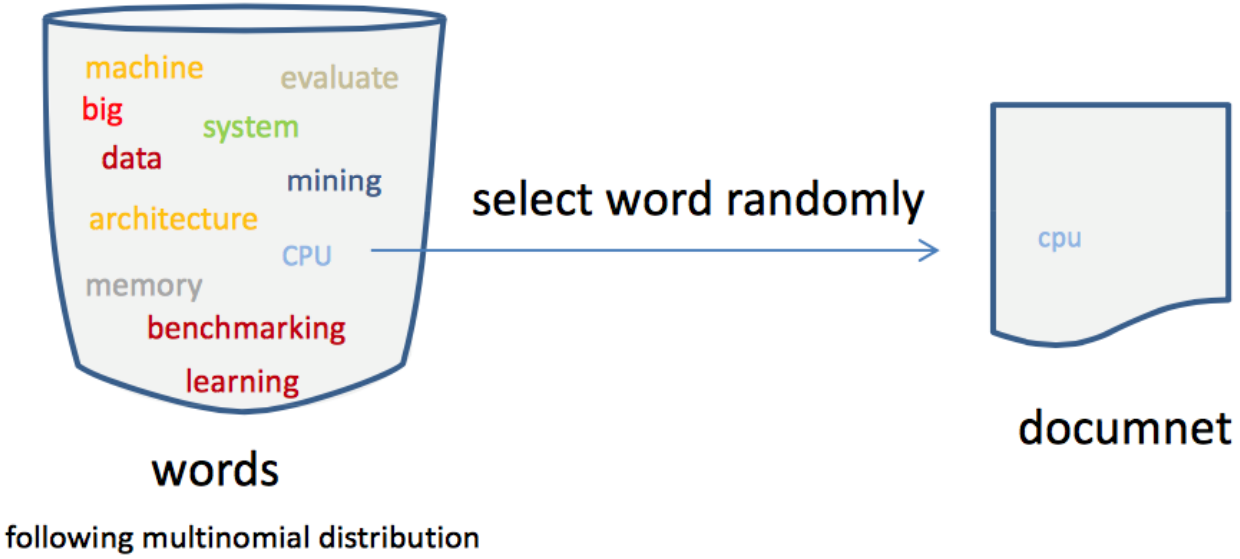
No.	data sets	data set description[1]	scalable data set
1	Wikipedia Entries	4,300,000 English articles (unstructured text)	Text Generator of BDGS
2	Amazon Movie Reviews	7,911,684 reviews (semi-structured text)	Text Generator of BDGS
3	Google Web Graph	875713 nodes, 5105039 edges (unstructured graph)	Graph Generator of BDGS
4	Facebook Social Network	4039 nodes, 88234 edges (unstructured graph)	Graph Generator of BDGS
5	E-commerce Transaction Data	Table 1: 4 columns, 38658 rows. Table 2: 6 columns, 242735 rows (structured table)	Table Generator of BDGS
6	ProfSearch Person Resumés	278956 resumés (semi-structured table)	Table Generator of BDGS
7	ImageNet	ILSVRC2014 DET image dataset (unstructured image)	ongoing development
8	English broadcasting audio files	Sampled at 16 kHz, 16-bit linear sampling (unstructured audio)	ongoing development
9	DVD Input Streams	110 input streams, resolution:704*480 (unstructured video)	ongoing development
10	Image scene	39 image scene description files (unstructured text)	ongoing development
11	Genome sequence data	cfa data format (unstructured text)	4 volumes of data sets
12	Assembly of the human genome	fa data format (unstructured text)	4 volumes of data sets
13	SoGou Data	the corpus and search query data from SoGou Labs (unstructured text)	ongoing development
14	MNIST	handwritten digits database which has 60,000 training examples and 10,000 test examples (unstructured image)	ongoing development
15	MovieLens Dataset	User's score data for movies, which has 9,518,231 training examples and 386,835 test examples(semi-structured text)	ongoing development

Big Data Generation Tool--BDGS

- Provide scalable data set extracted from real-world data sets

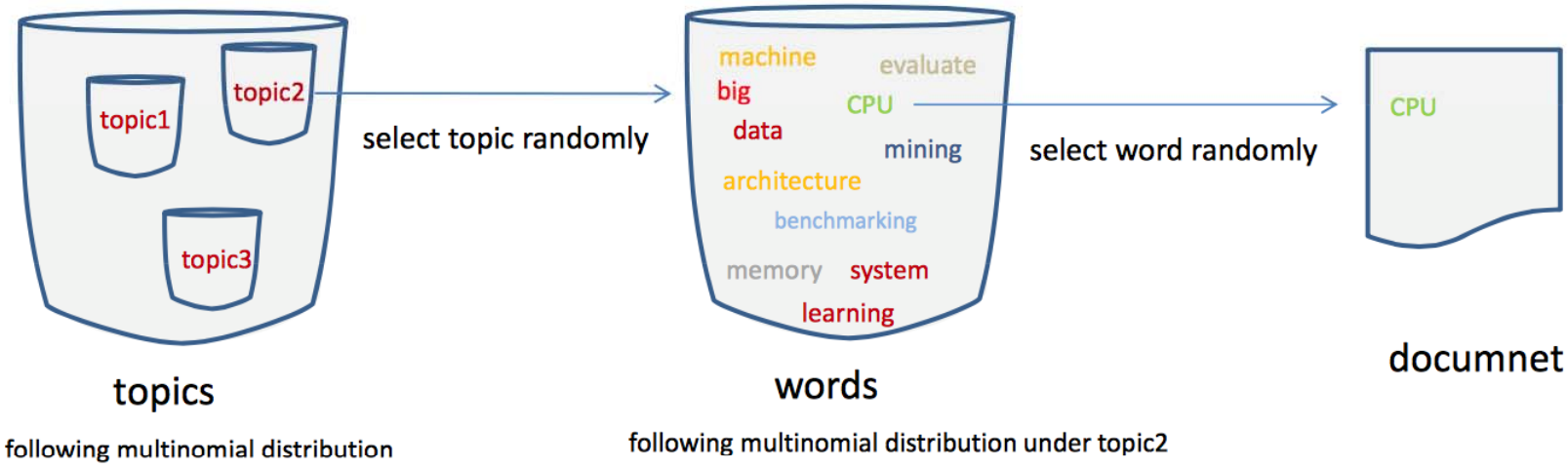


Naïve Text Generator



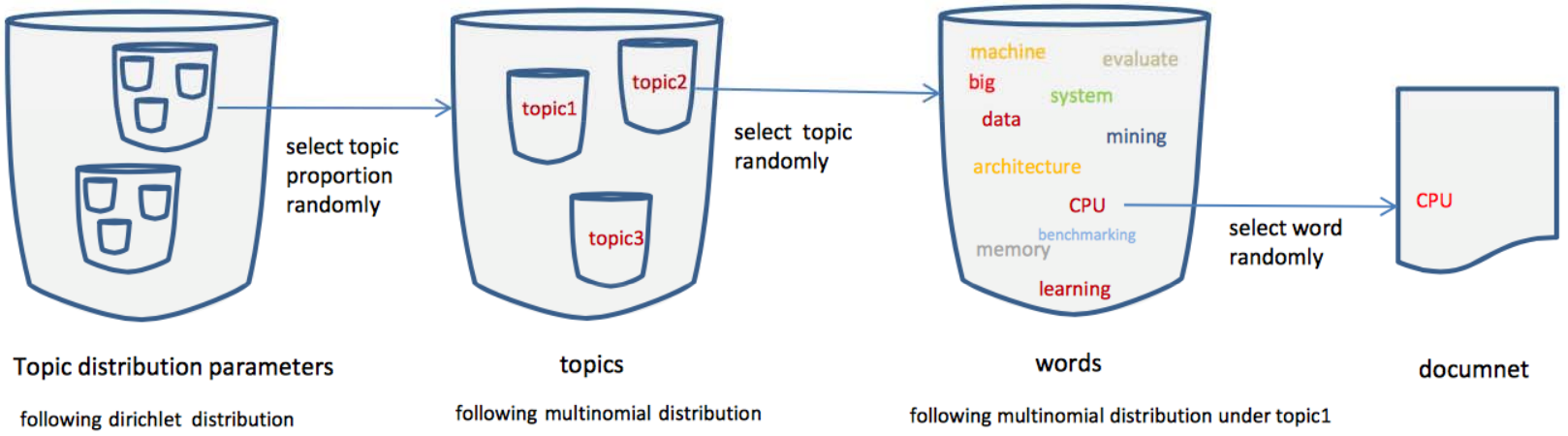
- Only modeling on word level;
- Words are selected according to the same distribution

Improved Text Generator



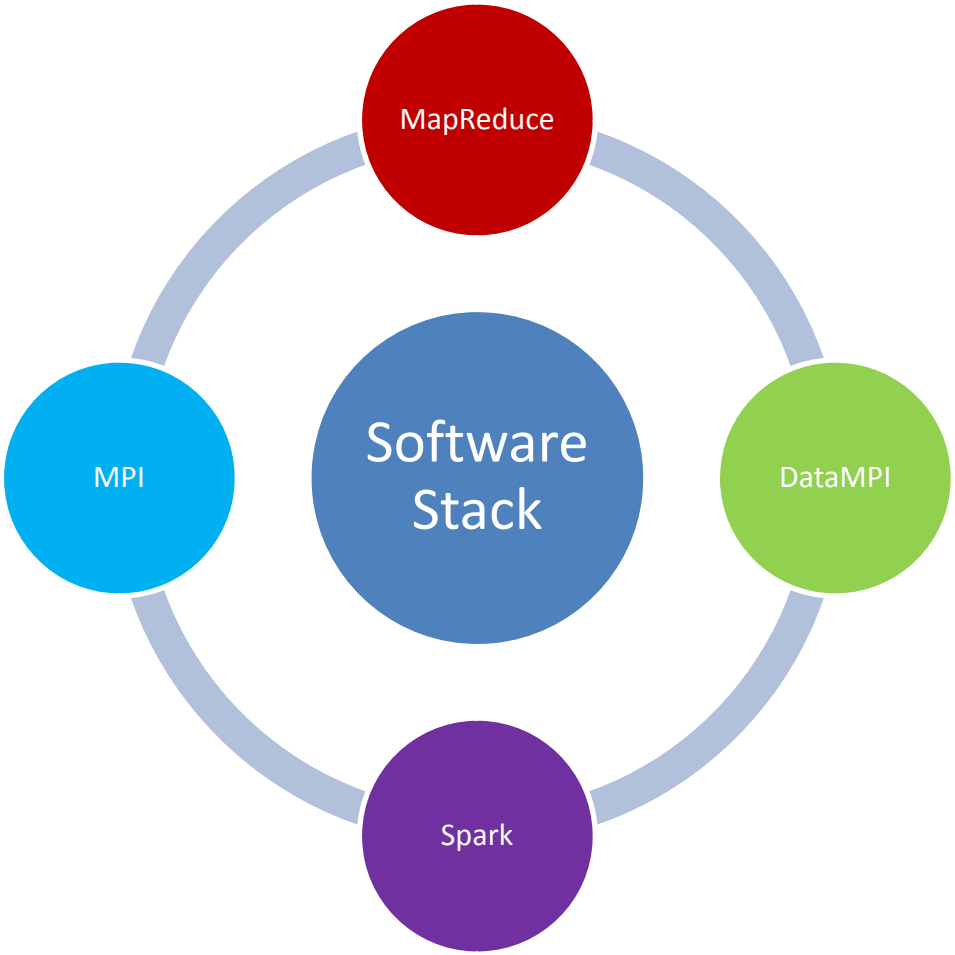
- Modeling on topic and word level
- Words are drew from distribution under particular topic
- Topics are drew from same distribution, as a result, each document has same topic proportion

Optimized Text Generator

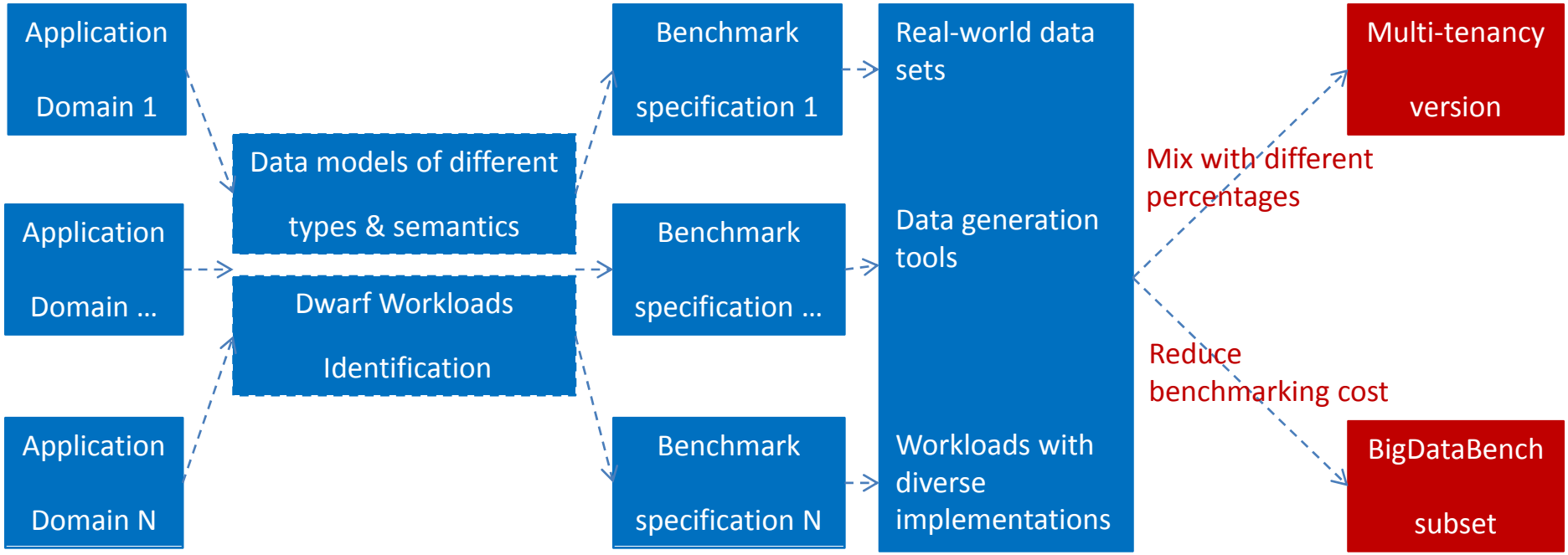


- Modeling on topic and word level
- Words are drew from distribution under particular topic
- Topics are selected from different distribution with parameters following a dirichlet distribution

Workloads With Diverse Implementations

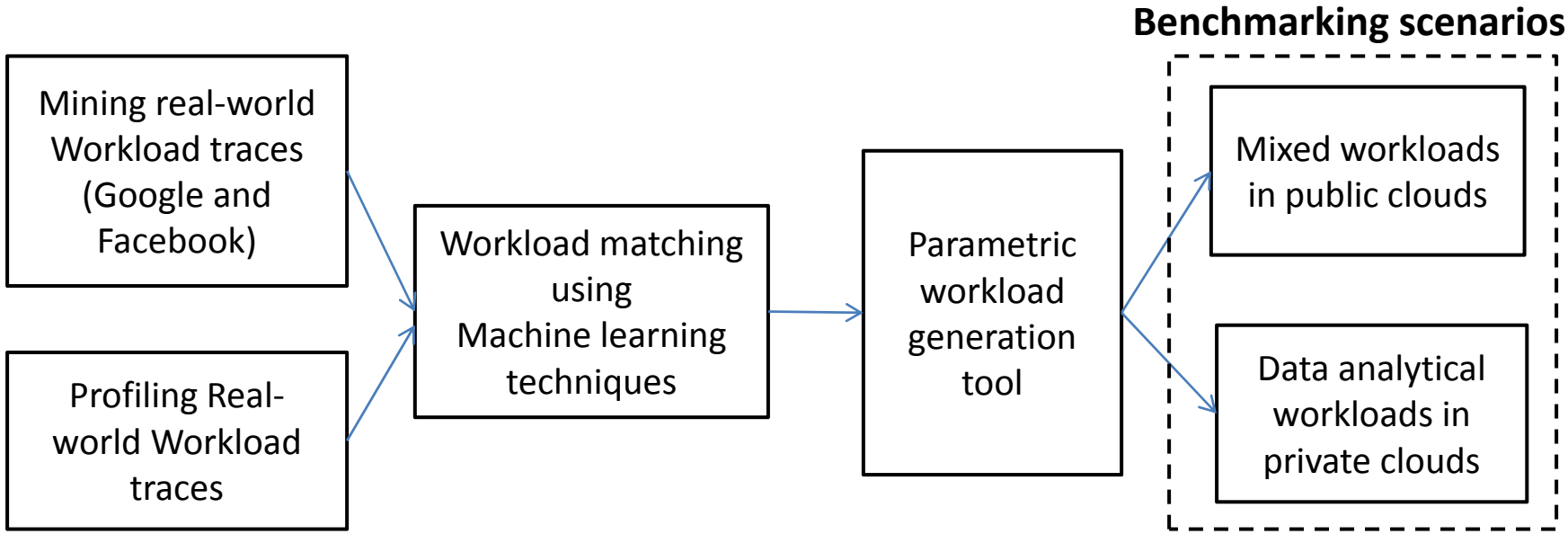


BigDataBench Methodology



Multi-tenancy version of BigDataBench

- Scenarios of multiple tenants running heterogeneous applications in cloud datacenters
  - Latency-critical online services
  - Latency-insensitive offline batch applications



BigDataBench Subset

- Motivation
  - Expensive to run all the benchmarks for system and architecture researches
    - multiplied by different implementations
    - BigDataBench 3.0 provides about 77 workloads

Methodology of Subsetting

- Identify a comprehensive set of workload characteristics from a specific perspective

- Eliminate the correlation data in those metrics
- Map the high dimension metrics to a low dimension

- Use the clustering method to classify
- Choose representative workloads from each category

Simulator version of BigDataBench

- Please refer to the user manual.

- http://prof.ict.ac.cn/BigDataBench/simulatorversion/#Simulator_version


QUESTIONS
And
Answers