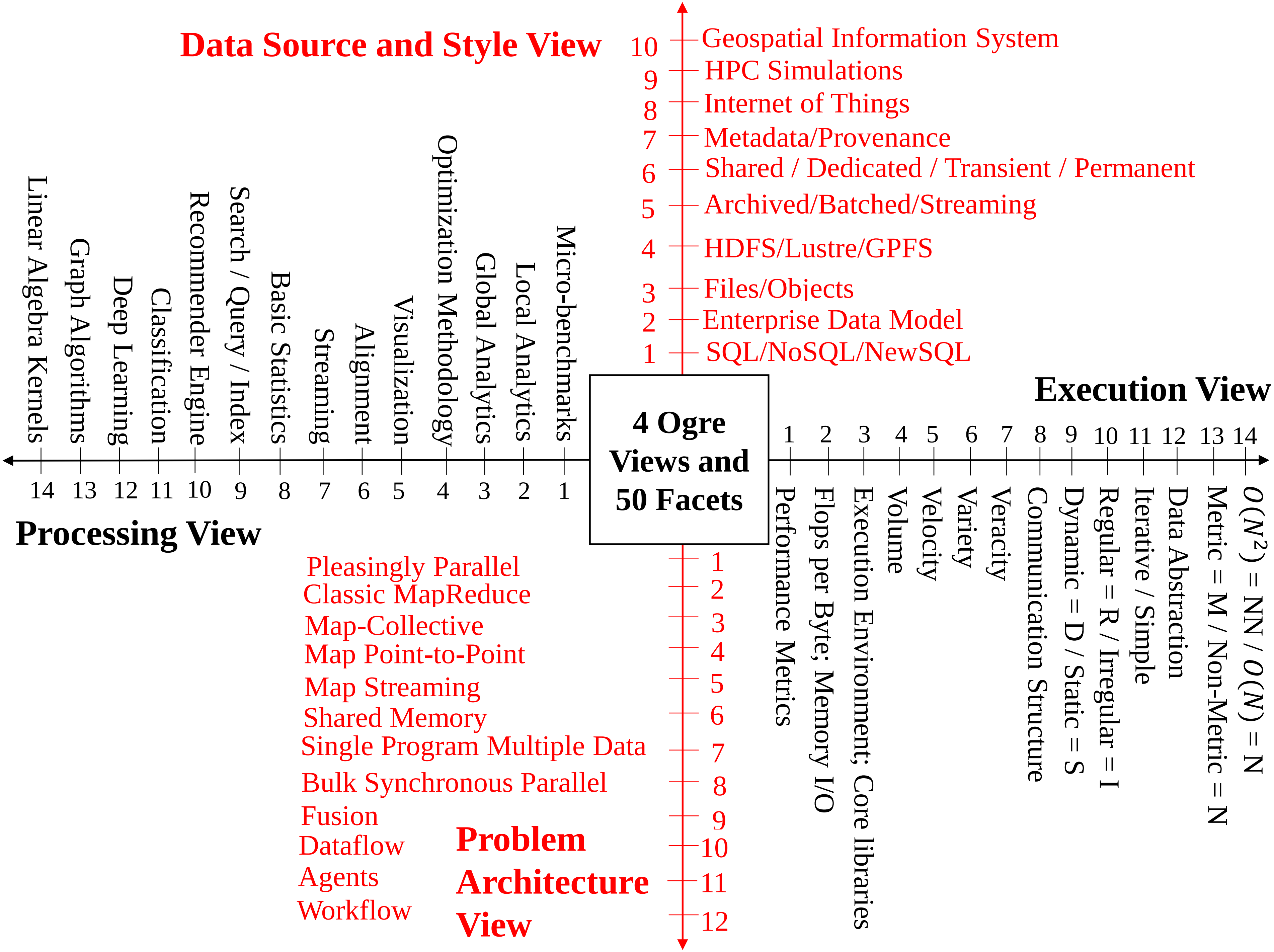
Big Data Ogres and their Facets

Geoffrey Fox, Judy Qiu, Shantenu Jha, Saliya Ekanayake

- **Big Data Ogres** are an attempt to characterize applications and algorithms with a set of general common features that are called **Facets**
- Originally derived from NIST collection of 51 use cases but refined with experience
- The 50 facets capture common characteristics (shared by several problems)which are inevitably multi-dimensional and often overlapping. **Divided into 4 views**
- One view of an Ogre is the overall **problem architecture** which is naturally related to the machine architecture needed to support data intensive application.
- The **execution (computational) features** view, describes issues such as I/O versus compute rates, iterative nature and regularity of computation and the classic V's of Big Data: defining problem size, rate of change, etc.
- The **data source & style** view includes facets specifying how the data is collected, stored and accessed. Has classic database characteristics
- **Processing** view has facets which describe types of processing steps including nature of algorithms and kernels e.g. Linear Programming, Learning, Maximum Likelihood
- **Instances of Ogres** are particular big data problems and a set of Ogre instances that cover enough of the facets could form a comprehensive **benchmark/mini-app** set
- Ogres and their instances can be **atomic** or **composite**

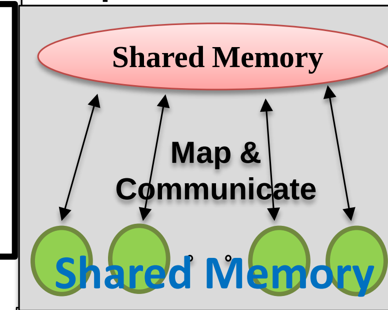

Data Source and Style View
10 Geospatial Information System
9 HPC Simulations
8 Internet of Things
7 Metadata/Provenance
6 Shared / Dedicated / Transient / Permanent
5 Archived/Batched/Streaming
4 HDFS/Lustre/GPFS
3 Files/Objects
2 Enterprise Data Model
1 SQL/NoSQL/NewSQL
4 Ogre Views and 50 Facets
Processing View
14 Linear Algebra Kernels
13 Graph Algorithms
12 Deep Learning
11 Classification
10 Recommender Engine
9 Search / Query / Index
8 Basic Statistics
7 Streaming
6 Alignment
5 Visualization
4 Optimization Methodology
3 Global Analytics
2 Local Analytics
1 Micro-benchmarks
Execution View
1 Performance Metrics
2 Flops per Byte; Memory I/O
3 Execution Environment; Core libraries
4 Volume
5 Velocity
6 Variety
7 Veracity
8 Communication Structure
9 Dynamic = D / Static = S
10 Regular = R / Irregular = I
11 Iterative / Simple
12 Data Abstraction
13 Metric = M / Non-Metric = N
14 $O(N^2)$ = NN / $O(N)$ = N
Problem Architecture View
1 Pleasingly Parallel
2 Classic MapReduce
3 Map-Collective
4 Map Point-to-Point
5 Map Streaming
6 Shared Memory
7 Single Program Multiple Data
8 Bulk Synchronous Parallel
9 Fusion
10 Dataflow
11 Agents
12 Workflow

Benchmarks/Mini-apps spanning Facets

- **Look at** NSF Dibbs Project, NIST 51 use cases, Baru-Rabl review
- **Catalog facets** of benchmarks and choose entries to cover “all facets”
- **Micro Benchmarks:** SPEC, EnhancedDFSIO (HDFS), Terasort, Wordcount, Grep, MPI, Basic Pub-Sub ….
- **SQL and NoSQL Data systems, Search, Recommenders:** TPC (-C to x–HS for Hadoop), BigBench, Yahoo Cloud Serving, Berkeley Big Data, HiBench, BigDataBench, Cloudsuite, Linkbench
    - includes MapReduce cases Search, Bayes, Random Forests, Collaborative Filtering
- **Spatial Query:** select from image or earth data
- **Alignment:** Biology as in BLAST
- **Streaming:** Online classifiers, Cluster tweets, Robotics, Industrial Internet of Things, Astronomy; BGBenchmark; choose to cover all 5 subclasses
- **Pleasingly parallel (Local Analytics):** as in initial steps of LHC, Pathology, Bioimaging (differ in type of data analysis)
- **Global Analytics:** Outlier, Clustering, LDA, SVM, Deep Learning, MDS, PageRank, Levenberg-Marquardt, Graph 500 entries
- **Workflow** and **Composite** (analytics on xSQL) linking above

6 Data Analysis Architectures

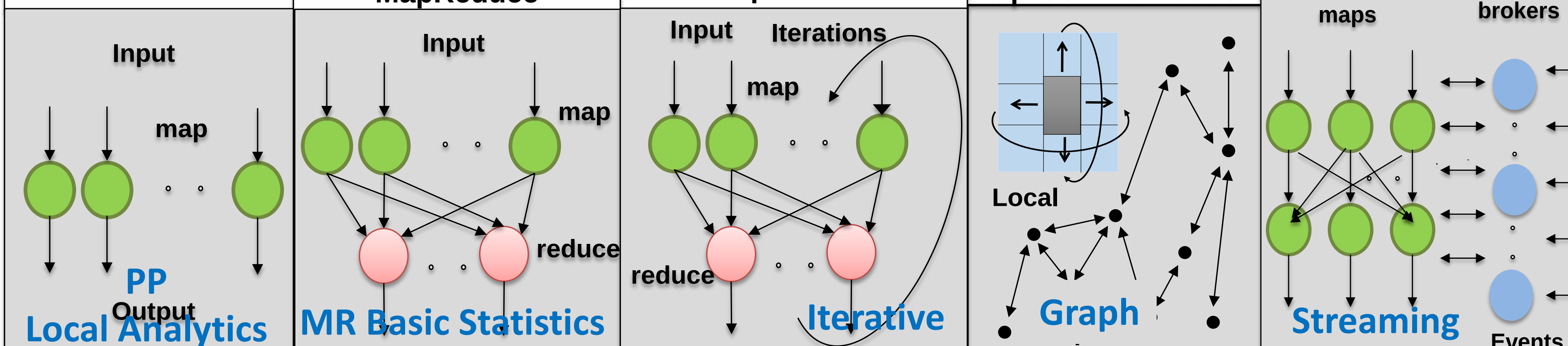
(6) Shared memory Map Communicates



Difficult to parallelize asynchronous parallel Graph Algorithms

Classic Hadoop in classes 1) 2) but not clearly best in class 1)

Many Task)



(1) Map Only	(2) Classic MapReduce	(3) Iterative Map Reduce or Map-Collective	(4) Point to Point or Map-Communication	(5) Map Streaming
BLAST Analysis Local Machine Learning Pleasingly Parallel	High Energy Physics (HEP) Histograms Web search Recommender Engines	Expectation maximization Clustering Linear Algebra, PageRank	Classic MPI PDE Solvers and Particle Dynamics Graph	Streaming images from Synchrotron sources, Telescopes, IoT
MapReduce and Iterative Extensions (Spark, Twister)			MPI, Giraph	Apache Storm
Harp – Enhanced Hadoop				Maps are Bolts