Big Data

overview, issues, challenges and opportunities

C. Onime

(onime@ictp.it)

Outline

- Interactive session
  - Introduction to Big-Data
  - Issues/challenges
  - Taxonomy classifications
- Conclusion
  - Opportunities and future

Pre-exercise

- Before providing a formal definition, let's try answer the questions:
  - What exactly is Big-Data?
  - Can you identify it?

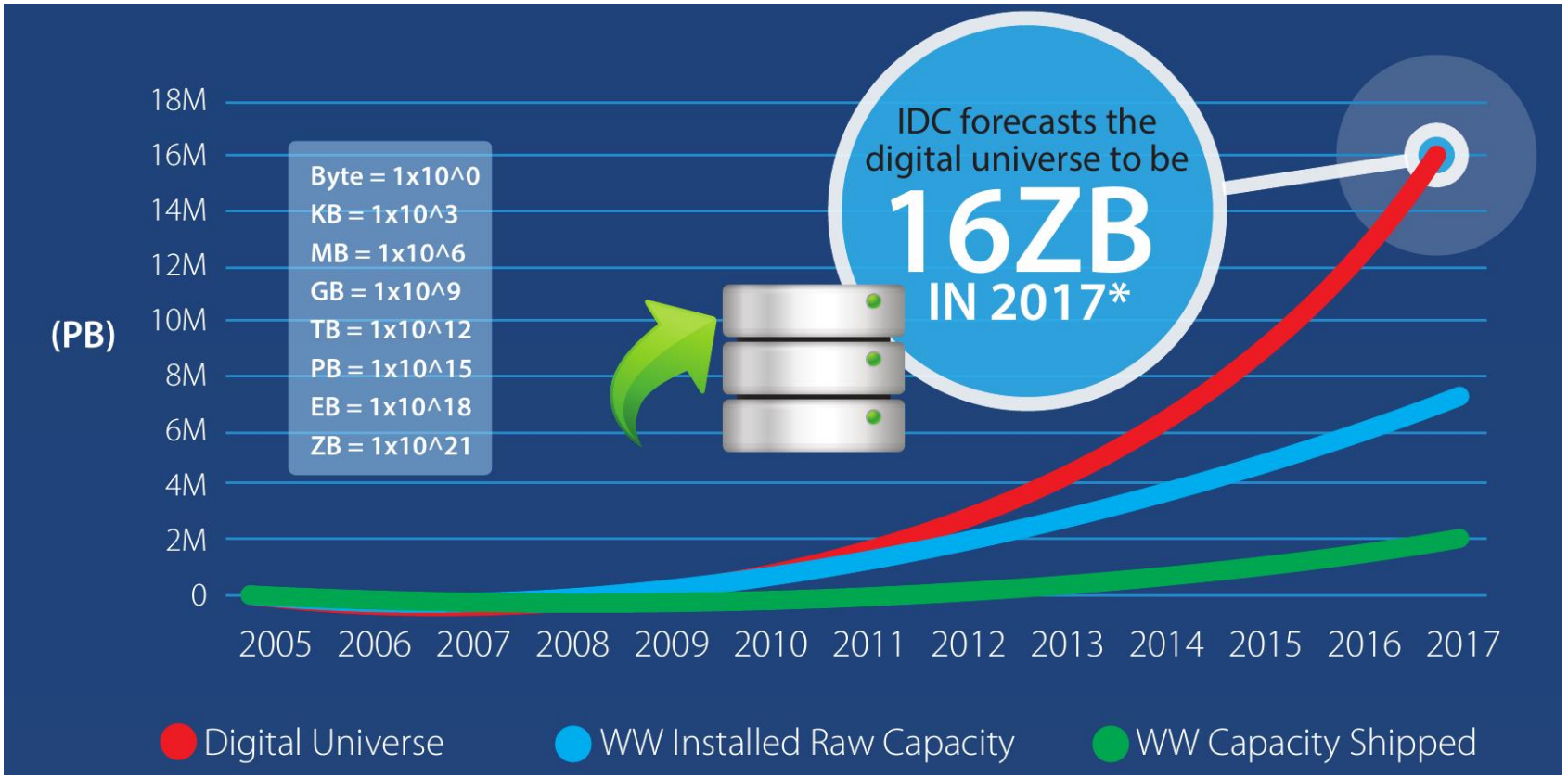
Definition(s)

- The term Big-Data by definition is used for data that is "massive" in one of the following areas:
  - Volume: quantity
  - Velocity: generated at high speed
  - Variety: wide spread from diverse sources and types.
  - Variability: constantly changing meaning
  - Veracity: making data accurate (removing bad data)
  - Visualization: presenting and conveying meaning
  - Value: applying findings and taking action

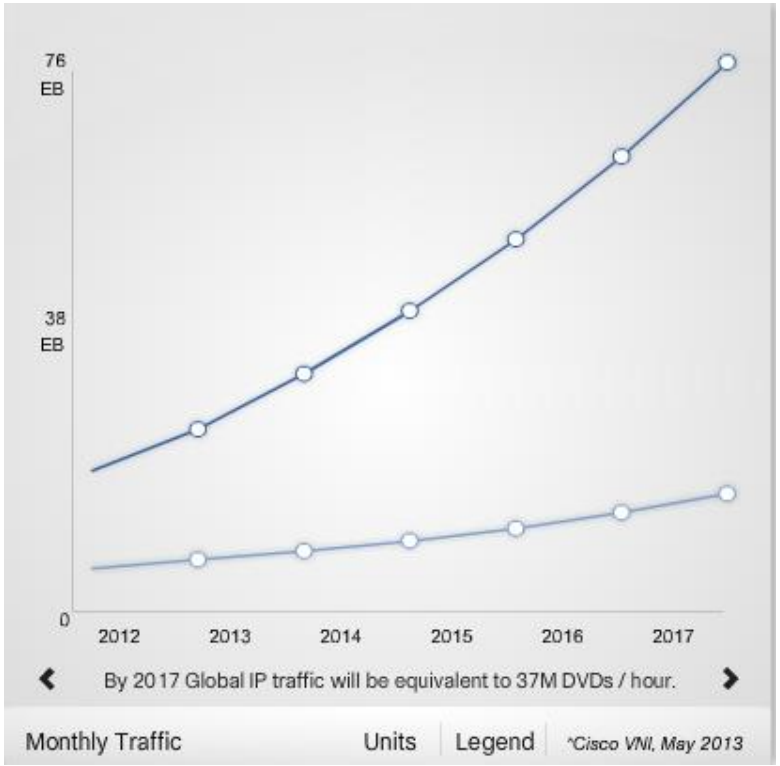
Big-Data examples

- Astronomical Image data from a telescope exceeds 1TB/day
- Environmetal monitoring
- Government: Census, National Health Records/Systems, etc.
- Industry: Amazon, Google, Ebay...

World wide storage



Another forecast



- 0.076 ZB = 76 EB
- 76 EB = 76M PB
- Current estimate is that 82% of global IP traffic will be video by 2020

Preamble

- So what is driving Big Data?
  - Mainly industry related paradigms & applications
    - Data mining, Business Intelligence, Knowledge Management and now Big Data Management

Data Mining

- A process of analyzing data from different perspectives and summarizing it into useful information, [...] which allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

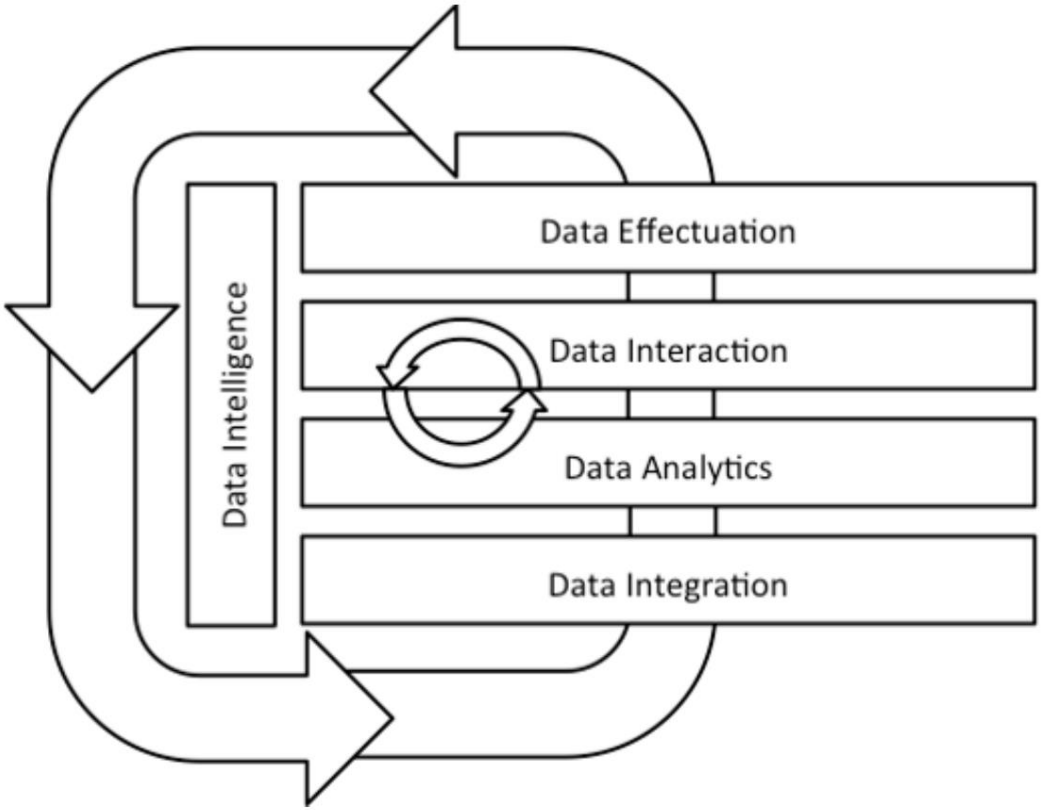
Business Intelligence

- A process of finding, gathering, aggregating and analyzing information for decision-making. It makes use of a set of technologies that allow the acquisition and analysis of data to improve company decision making and work flows.

Knowledge Management

- A business process that formalizes the management and use of an enterprise's intellectual assets." KM promotes a collaborative and integrative approach to the creation, capture, organization, access and use of information assets, including the tacit, un-captured knowledge of people.
- *A systematic process of finding, selecting, organizing, distilling and presenting information in a way that improves an employee's comprehension in a specific area of interest which supports an organization to gain insight and understanding from its own experience.*

Big Data Management



Other drivers

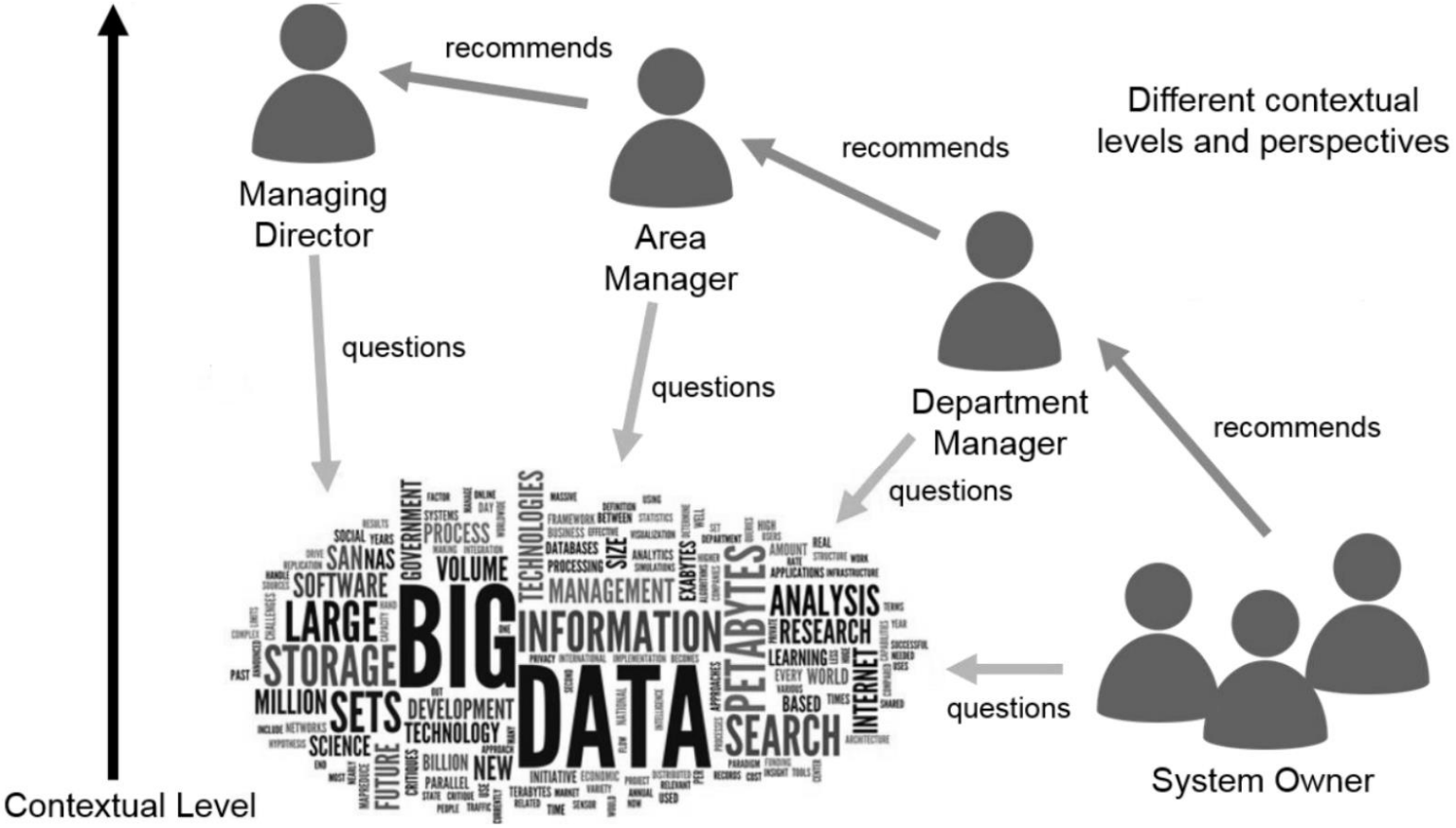
- Scientific Research
  - High Performance Computing (LHC, SKA, Genomics)
- Improvements in hardware technology
  - Heading towards Nano-circuits, clocking resolutions, etc
- Improvements in computing platforms
  - Networks: always connected devices, capacity; Clouds: anytime, anywhere on-demand metered access to resources
- Every user is a now a provider/consumer
  - Social networking

Issues and challenges

- Perspectives – backgrounds, use cases
- Taxonomies, ontologies, schemas, workflow
- Bits – raw data formats and storage methods
- Cycles – algorithms and analysis
- Infrastructure (screws) to support Big Data

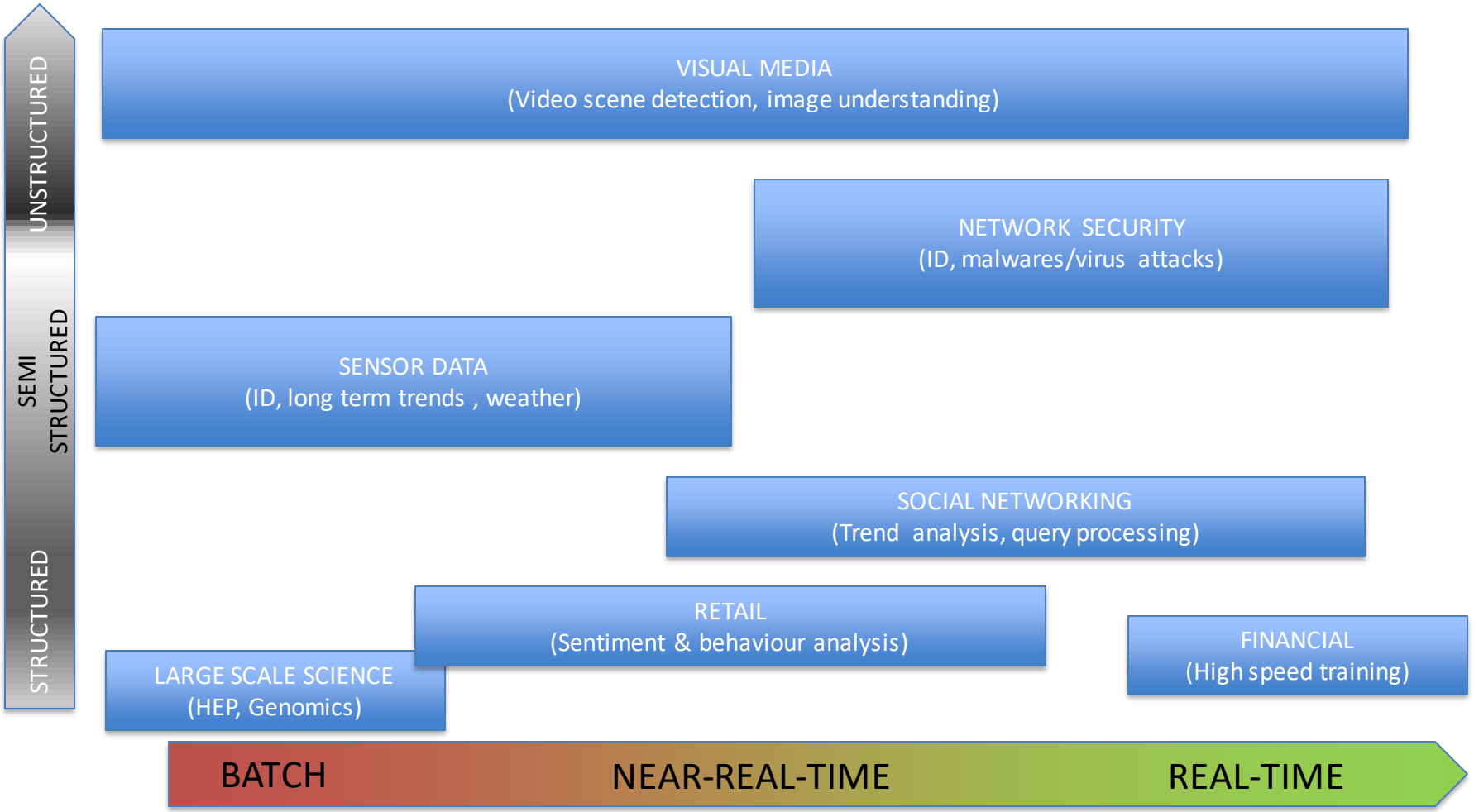
  – *From presentation by Michael Cooper & Peter Mell of NIST*

Perspectives

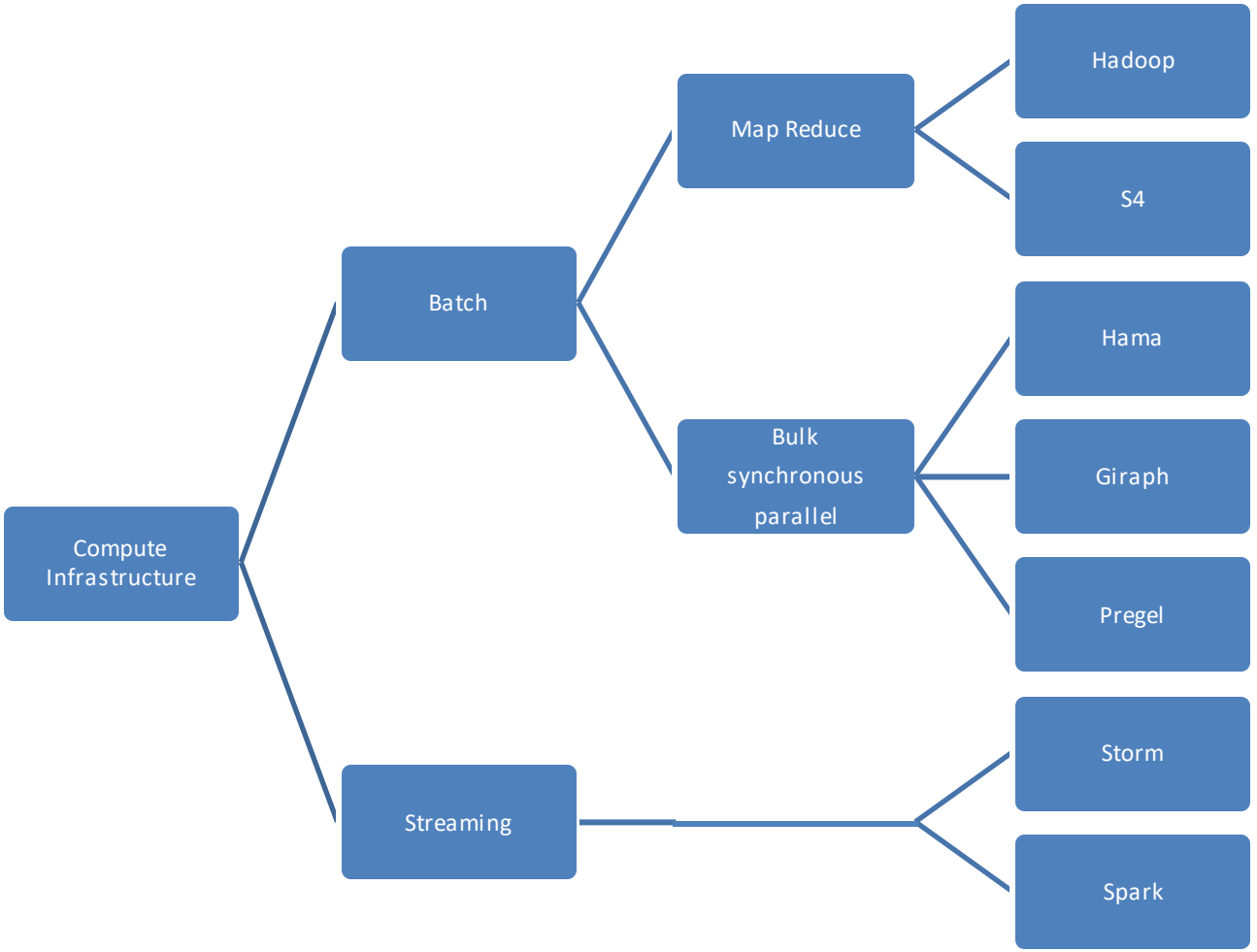


Six dimensional Taxonomy

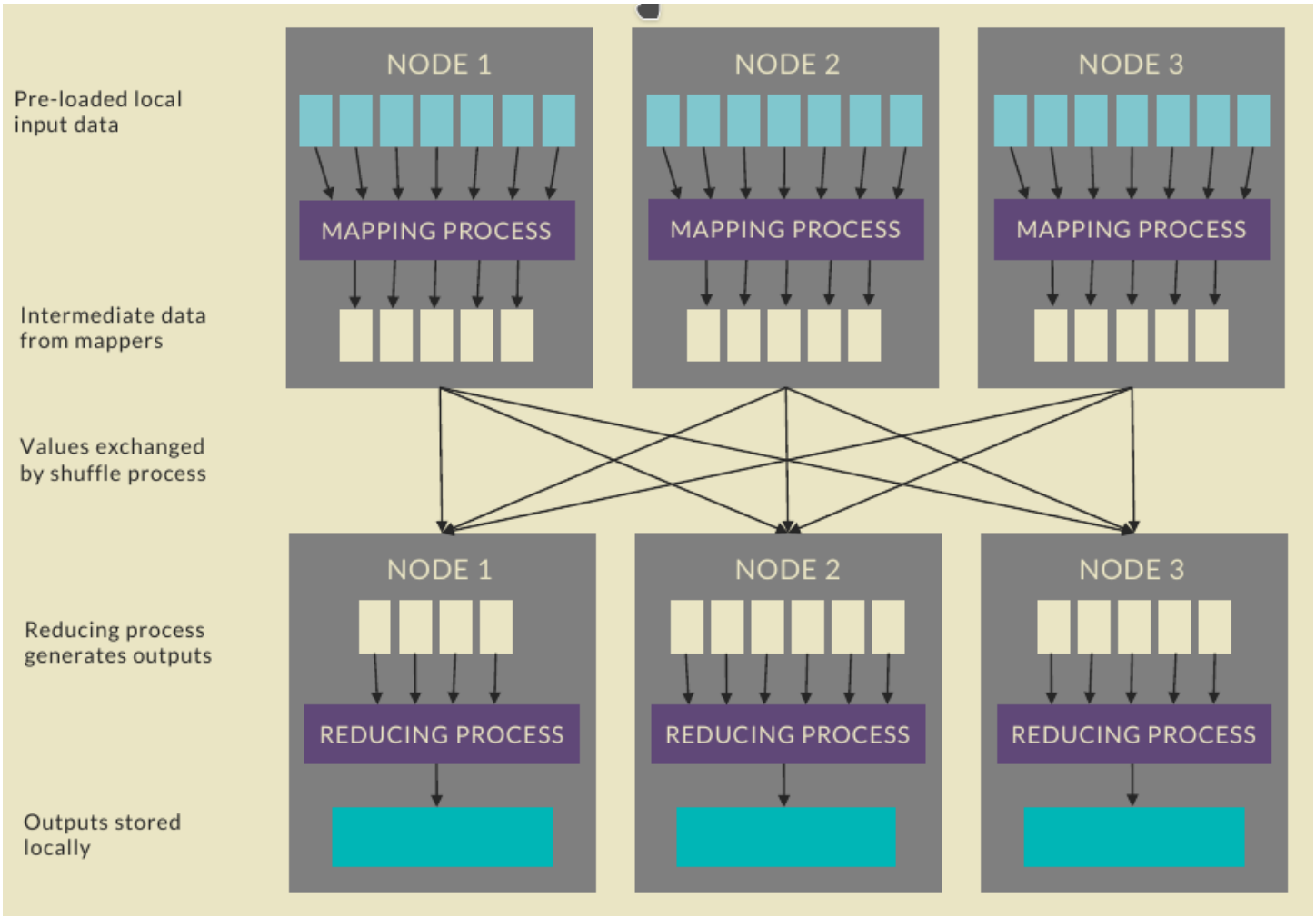
Data Mapping examples



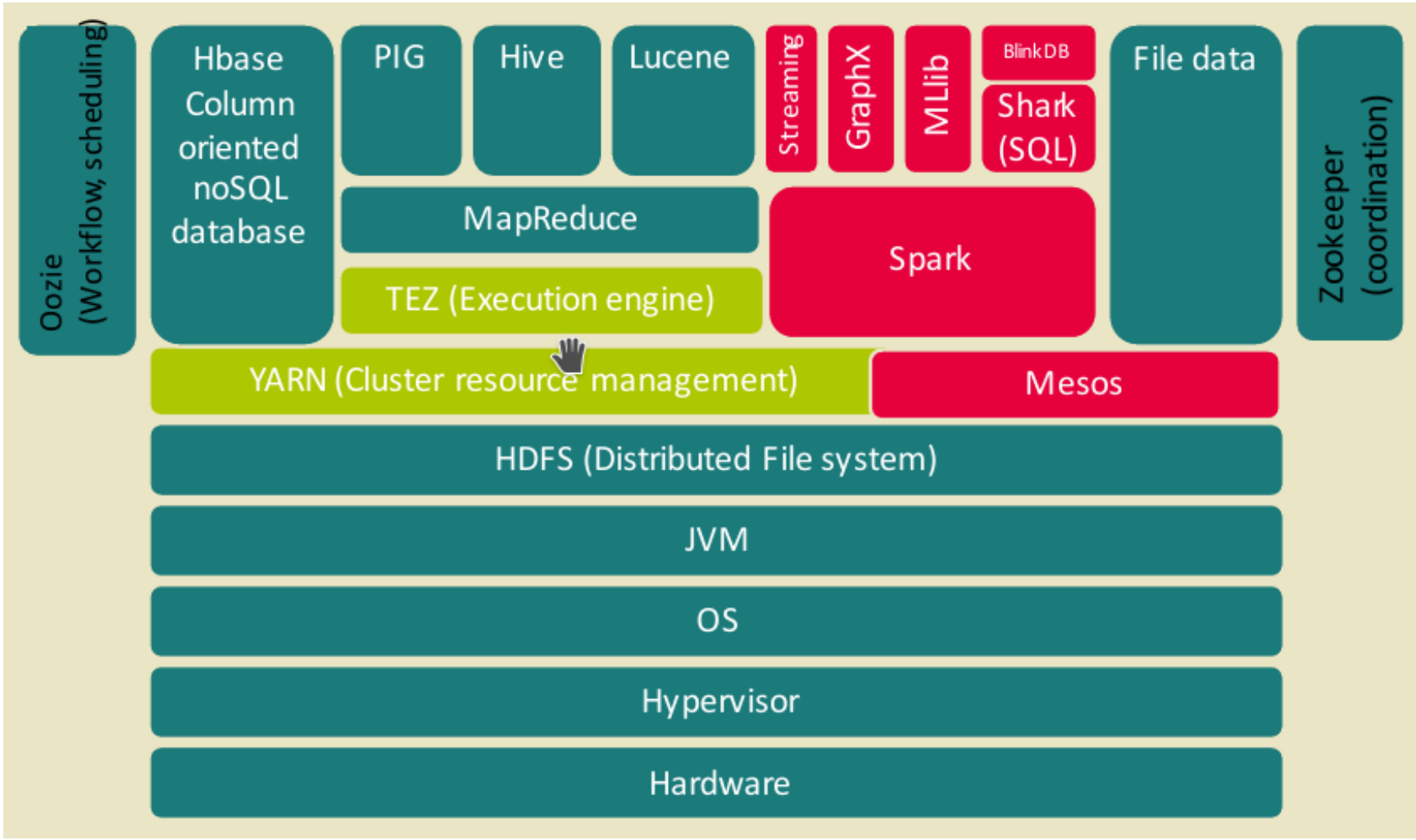
Compute infrastructure



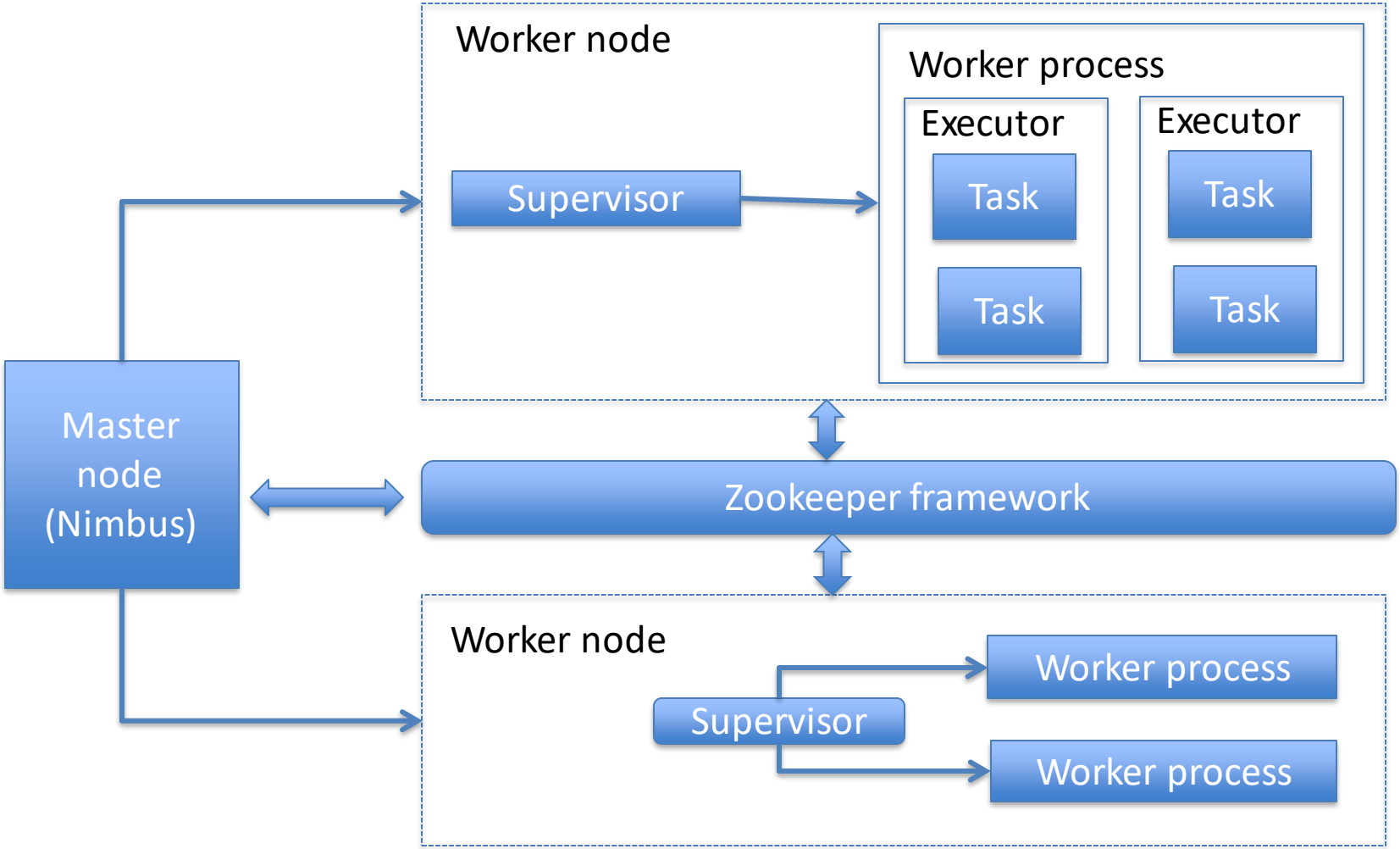
Overview of Hadoop MapReduce



Hadoop 2.0 Ecosystem

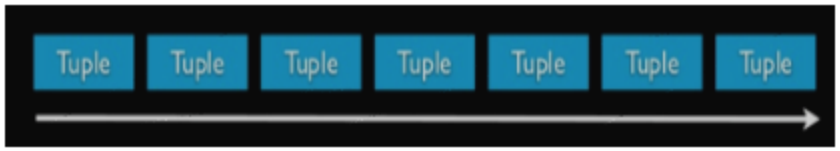


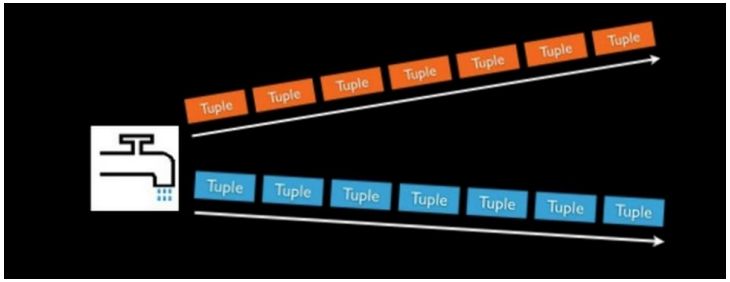
Storm Cluster

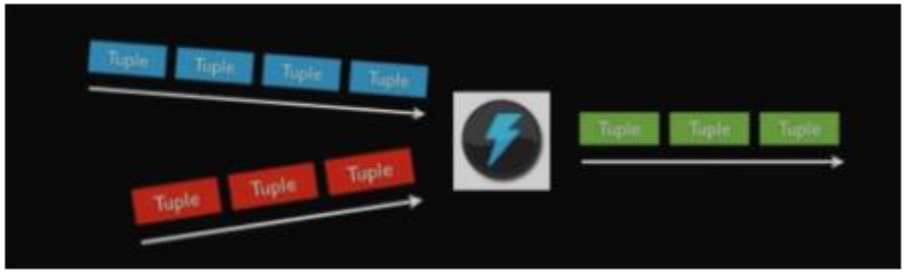


Storm Basics

- Tuple
  - Key-value pairs
- Streams
  - Sequence of tuples pairs
- Spout
  - Source of streams
- Bolt
  - Processing element
  - (filers, join, transform, e.t.c)

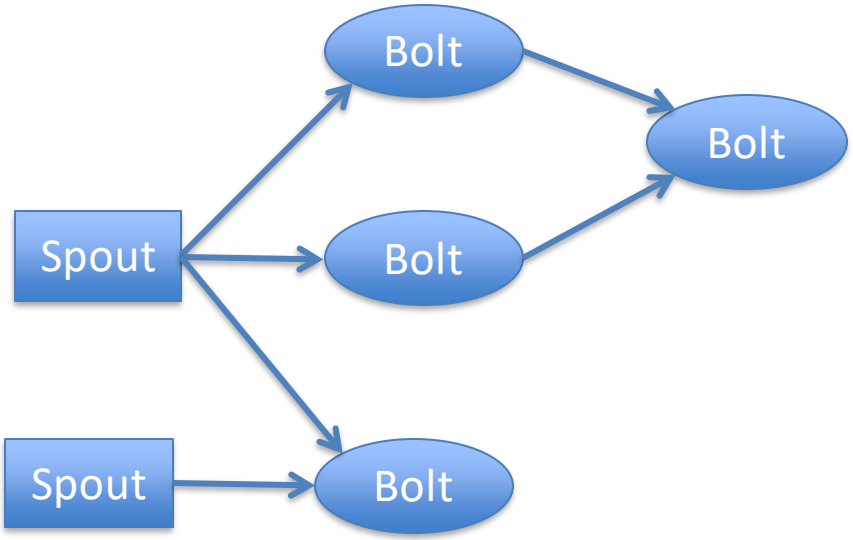




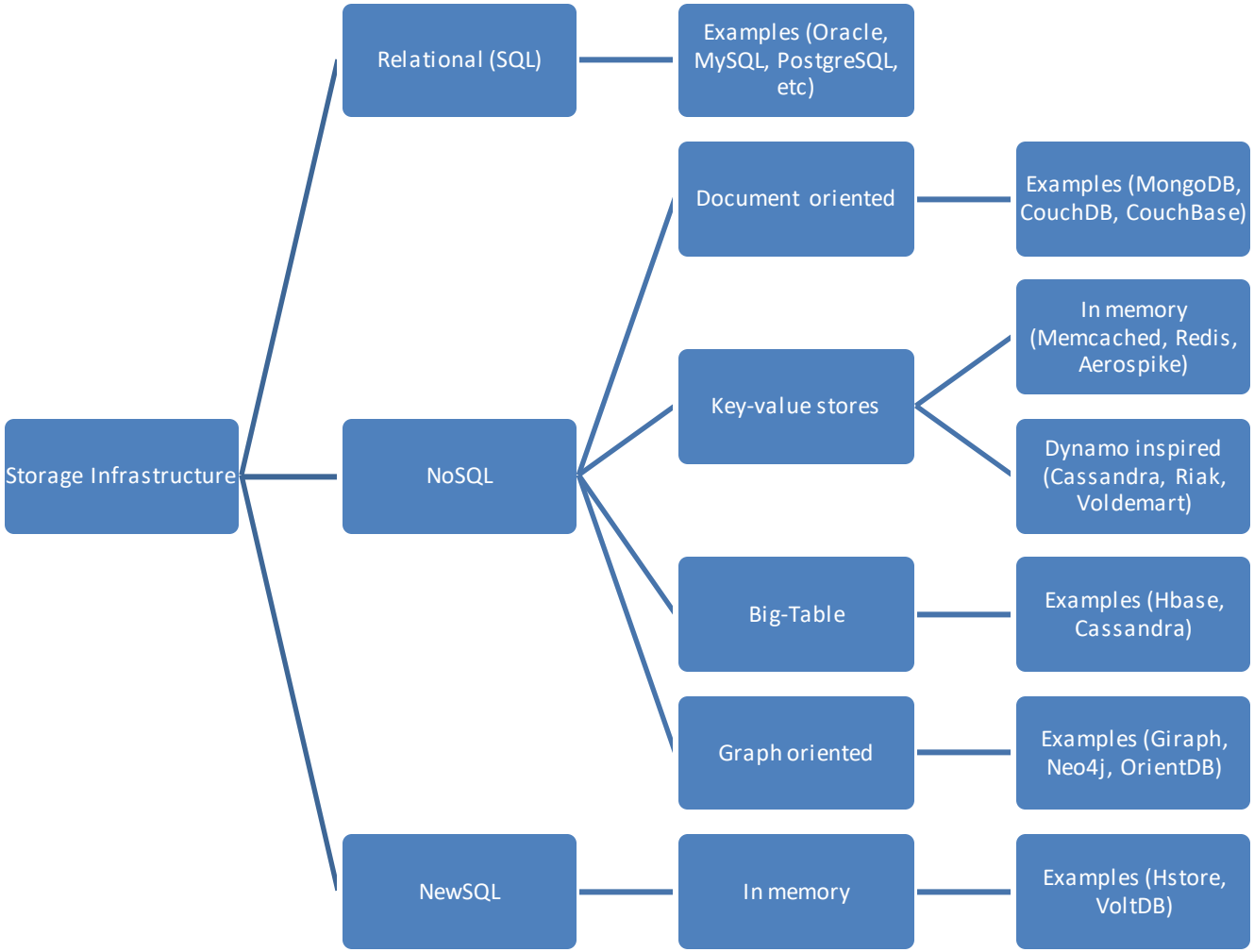


Storm topology

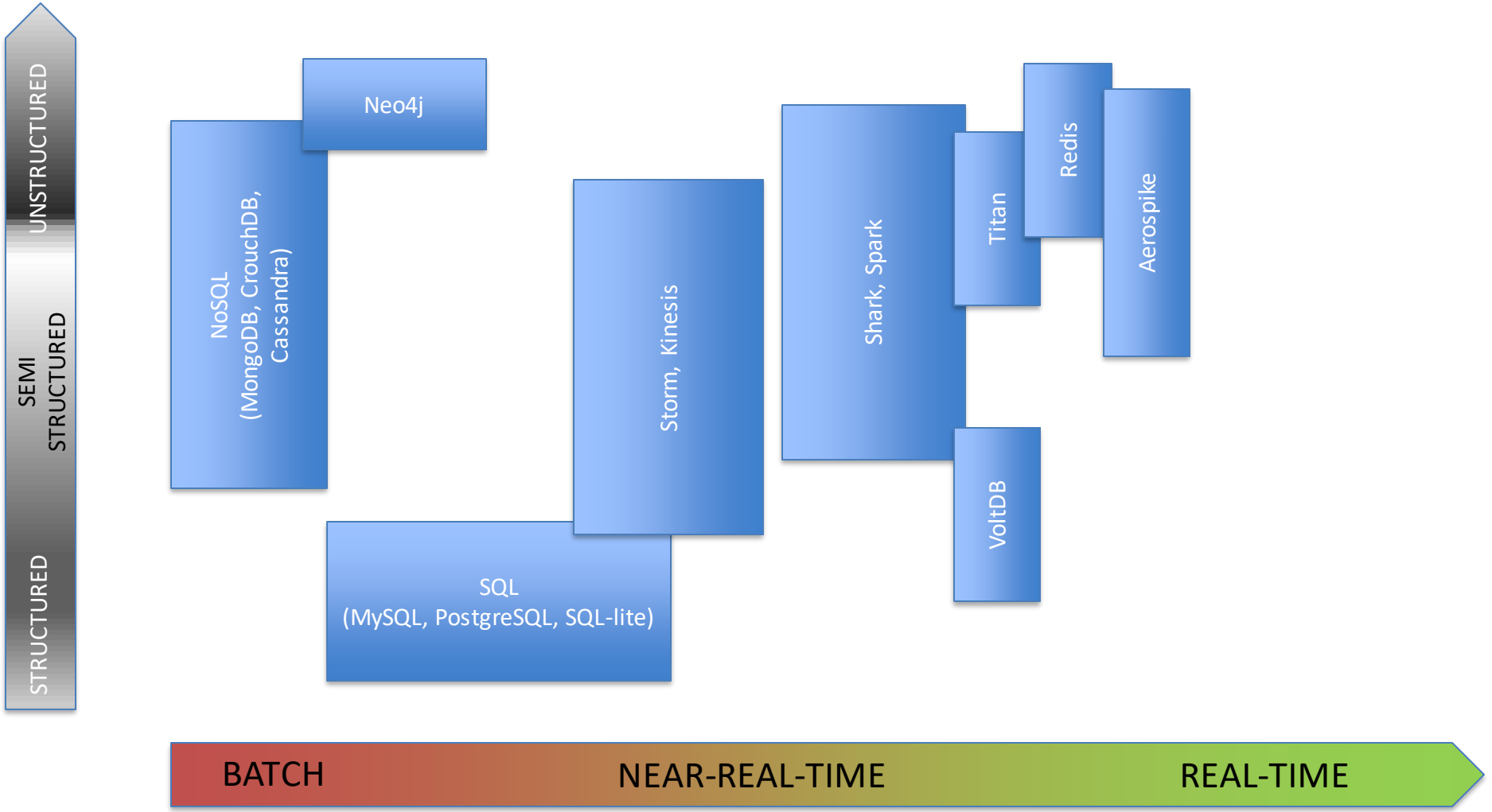
- Graph of Computation
  - Network of spouts and bolt
  - Parallel & cyclic execution
- Groupings
  - Shuffle, all, Global, fields
- Example:
  - Twitter analytics: spout, bolts: parse, count, ranks, report



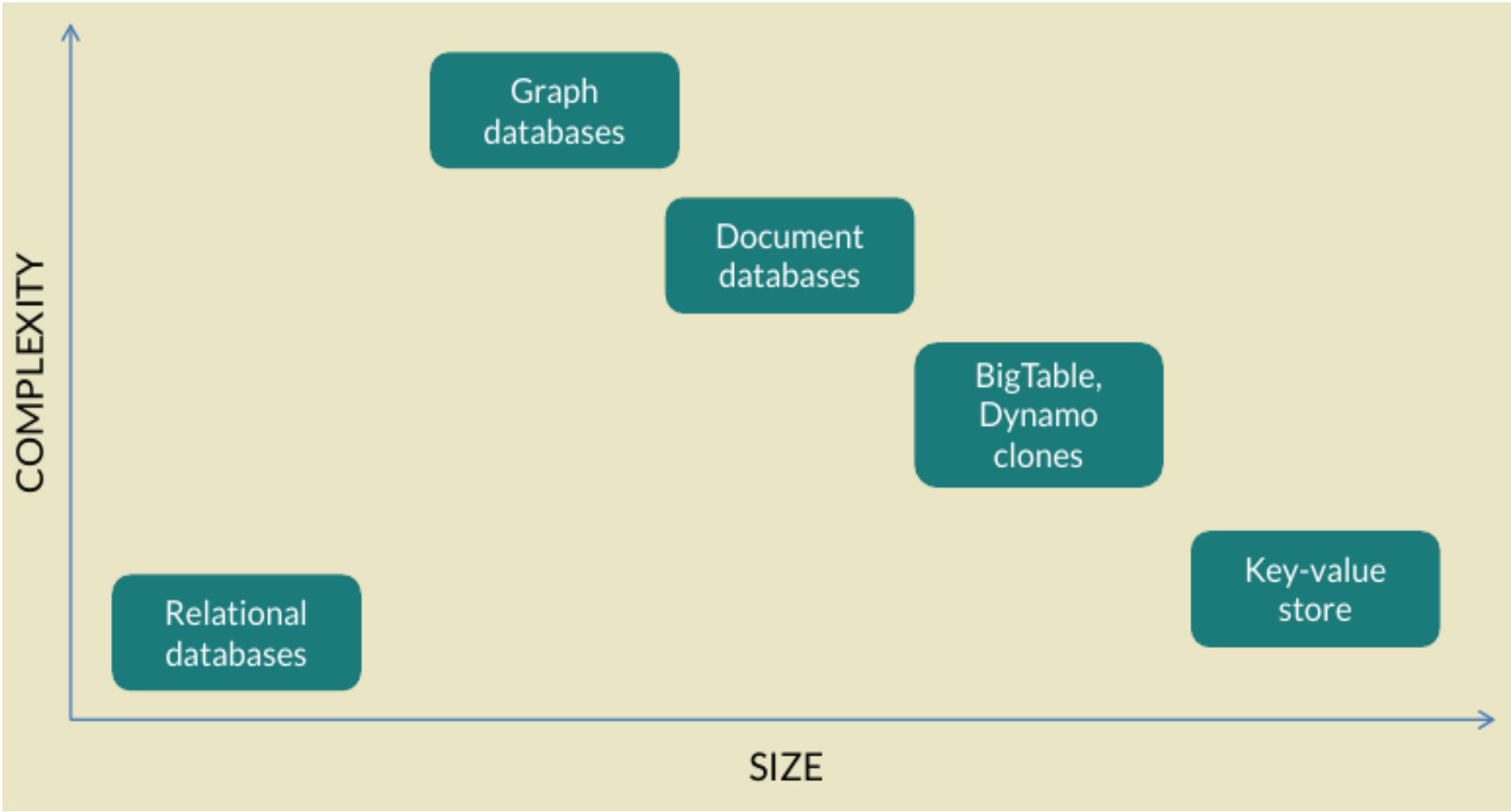
Storage infrastructure



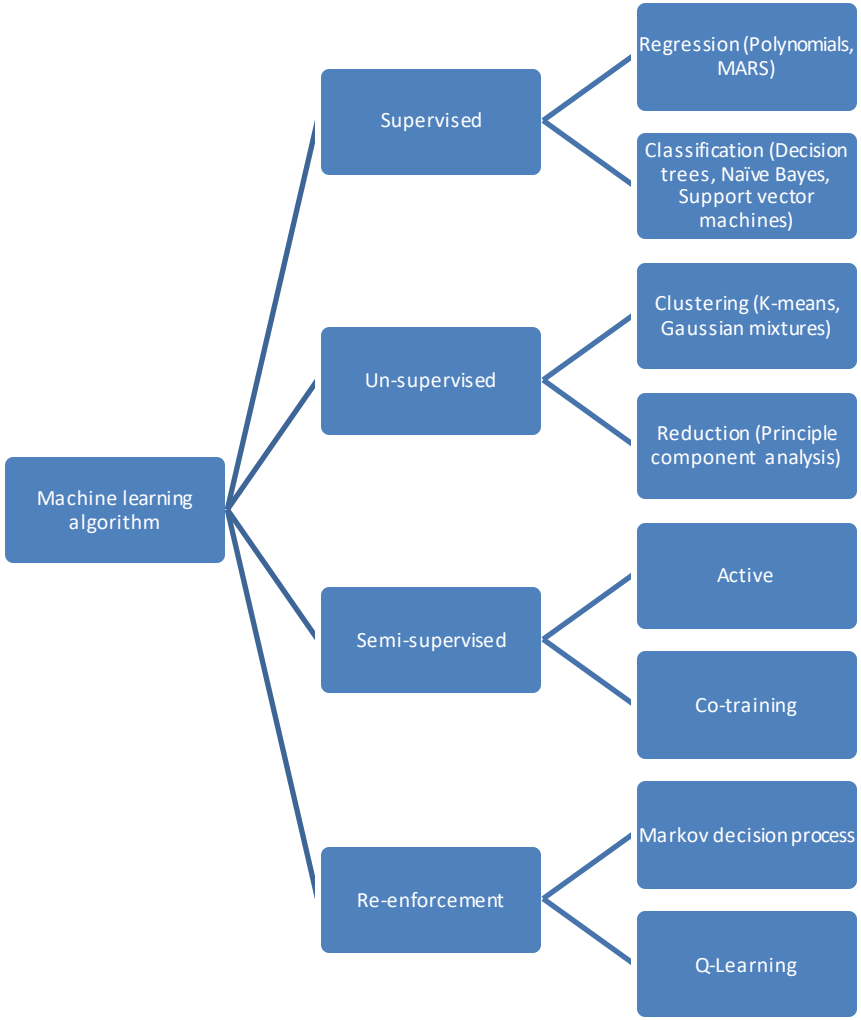
Infrastructure mapping



Storage complexity/size



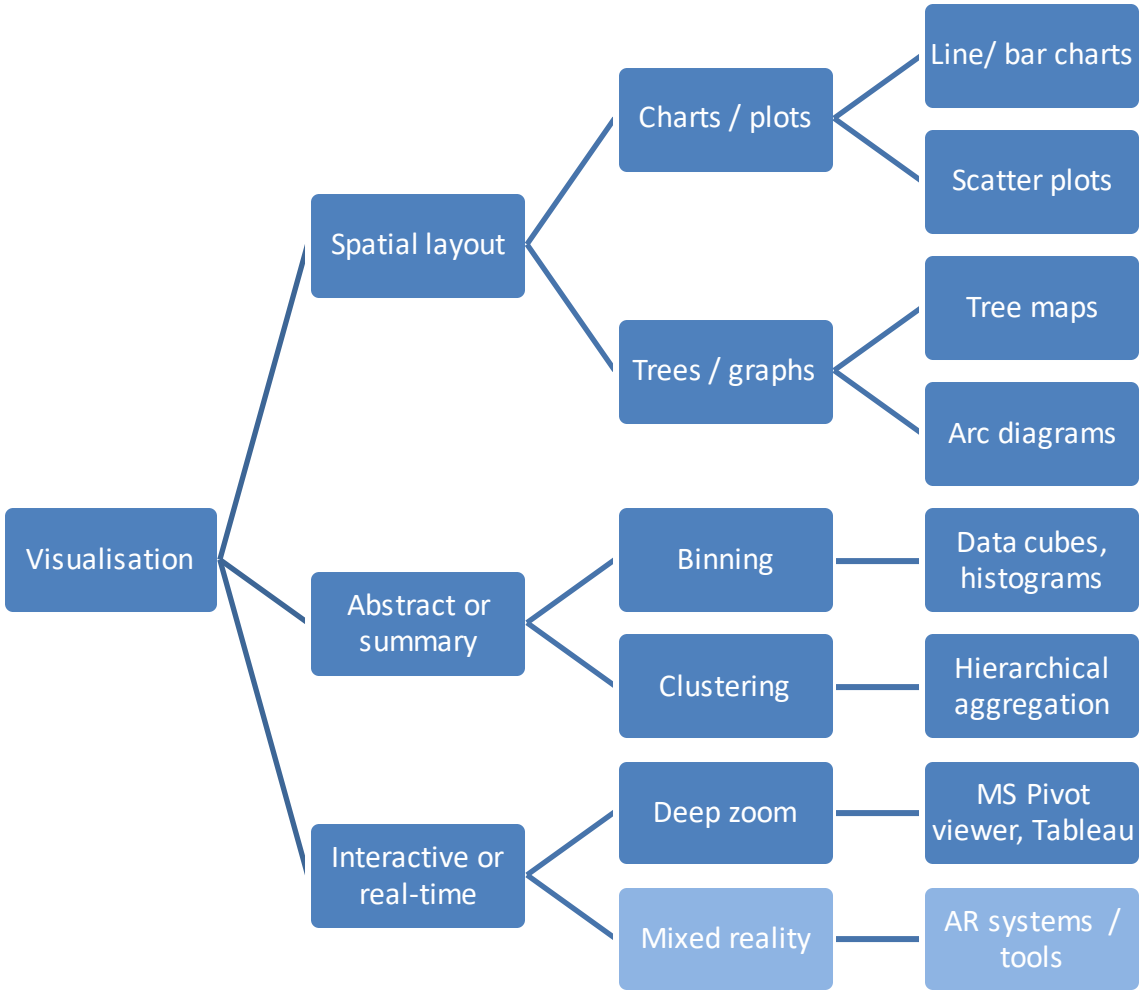
Analytics


Machine learning algorithm
Supervised
Regression (Polynomials, MARS)
Classification (Decision trees, Naïve Bayes, Support vector machines)
Un-supervised
Clustering (K-means, Gaussian mixtures)
Reduction (Principle component analysis)
Semi-supervised
Active
Co-training
Re-enforcement
Markov decision process
Q-Learning

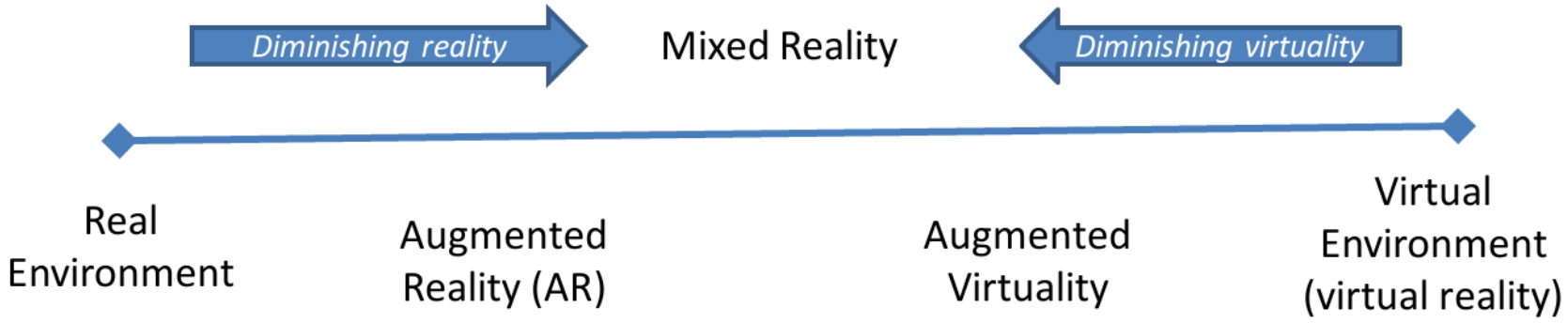
Comparison of Data analysis paradigms

Statistics	Machine learning
Model	Network, Graphs
Data point	Examples/instances
Response	Label
Parameters	Weights
Covariate	Feature
Fitting/Estimation	Learning
Test set performance	Generalization
Regression/Classification	Supervised Learning
Density estimation, Clustering	Unsupervised Learning

Visualisation



Mixed Reality Environments



$$E_{MR} = \int (R + V)$$

where

$$E_{MR} = \begin{cases} E_R, & \text{if } V = 0 \\ E_{AR}, & \text{if } R > V \\ E_{AV}, & \text{if } R < V \\ E_{VR}, & \text{if } R = 0 \end{cases}$$

VR and AR

Virtual Reality (VR) CAVE

- Computer generated virtual environment
- Creates a completely virtual environment that is without real objects
- Portable
  - Headsets, wearable devices
  - Custom and typically not cost effective

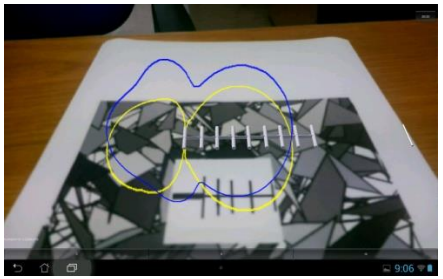
Augmented Reality(AR)

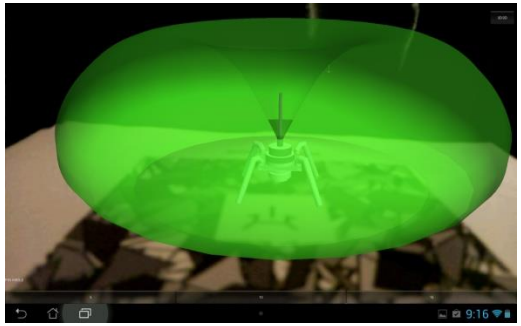
- Real-time integration of computer generated information into a 3D world.
- Blends into real world and supports real objects
- Mobile
  - Commodity devices: smart-phones and tablets
  - Cost effective

Some Examples

VR Environments

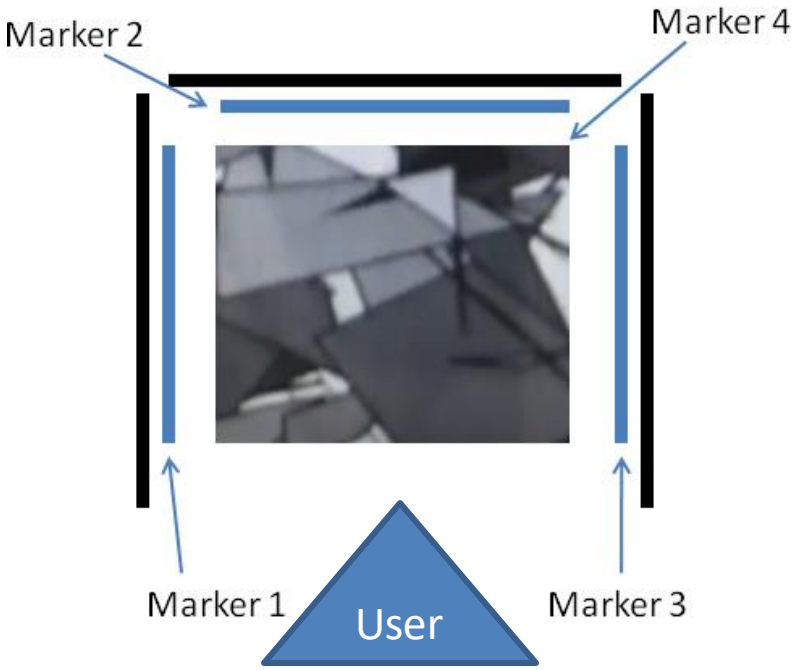
AR Environments

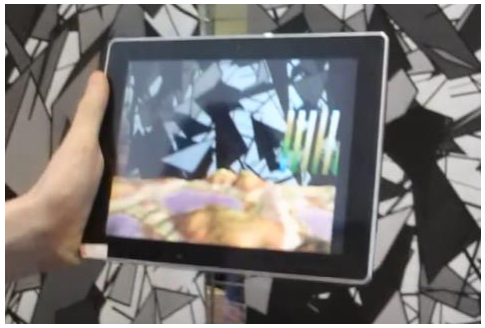




AR Cubicle

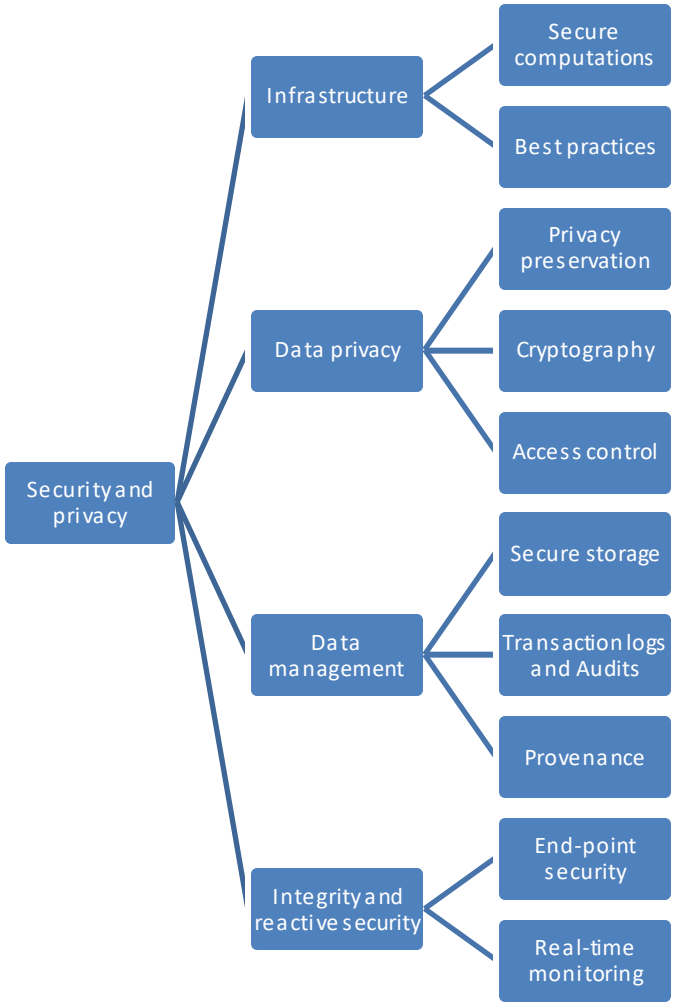
AR immersive cubicle





180° horizontal by 3 markers on walls and 90° vertical by marker on floor

Security and privacy



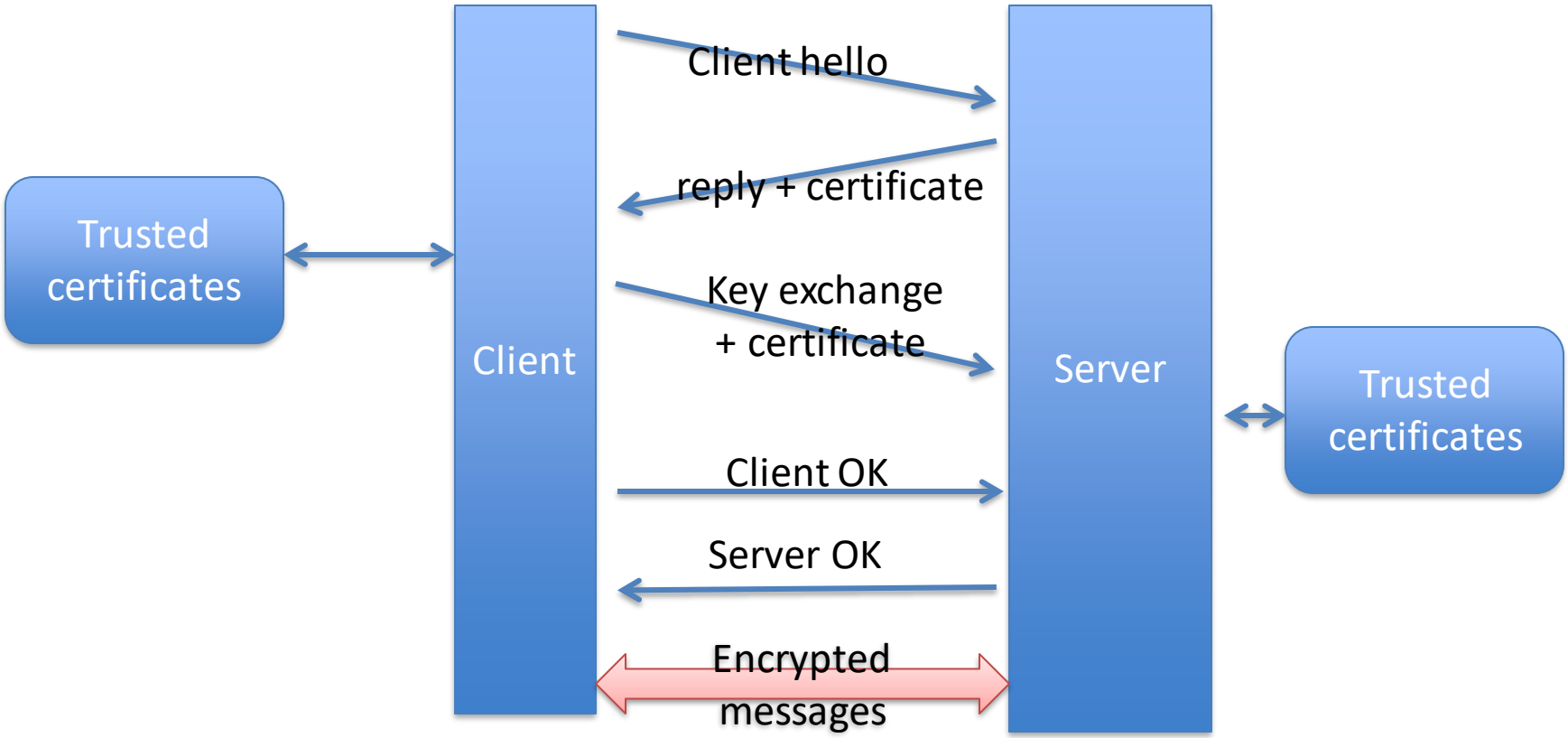
Public Key Cryptography

- Asymmetric cryptography
  - A pair of keys: one public and the other private
  - Useful for authentication and encryption
  - Depends mainly on the impracticability of computing the equivalent private key from its public component.
  - Public key may be freely exchanged without secure channels such as public key servers, etc..
  - Computationally intensive mathematical algorithms

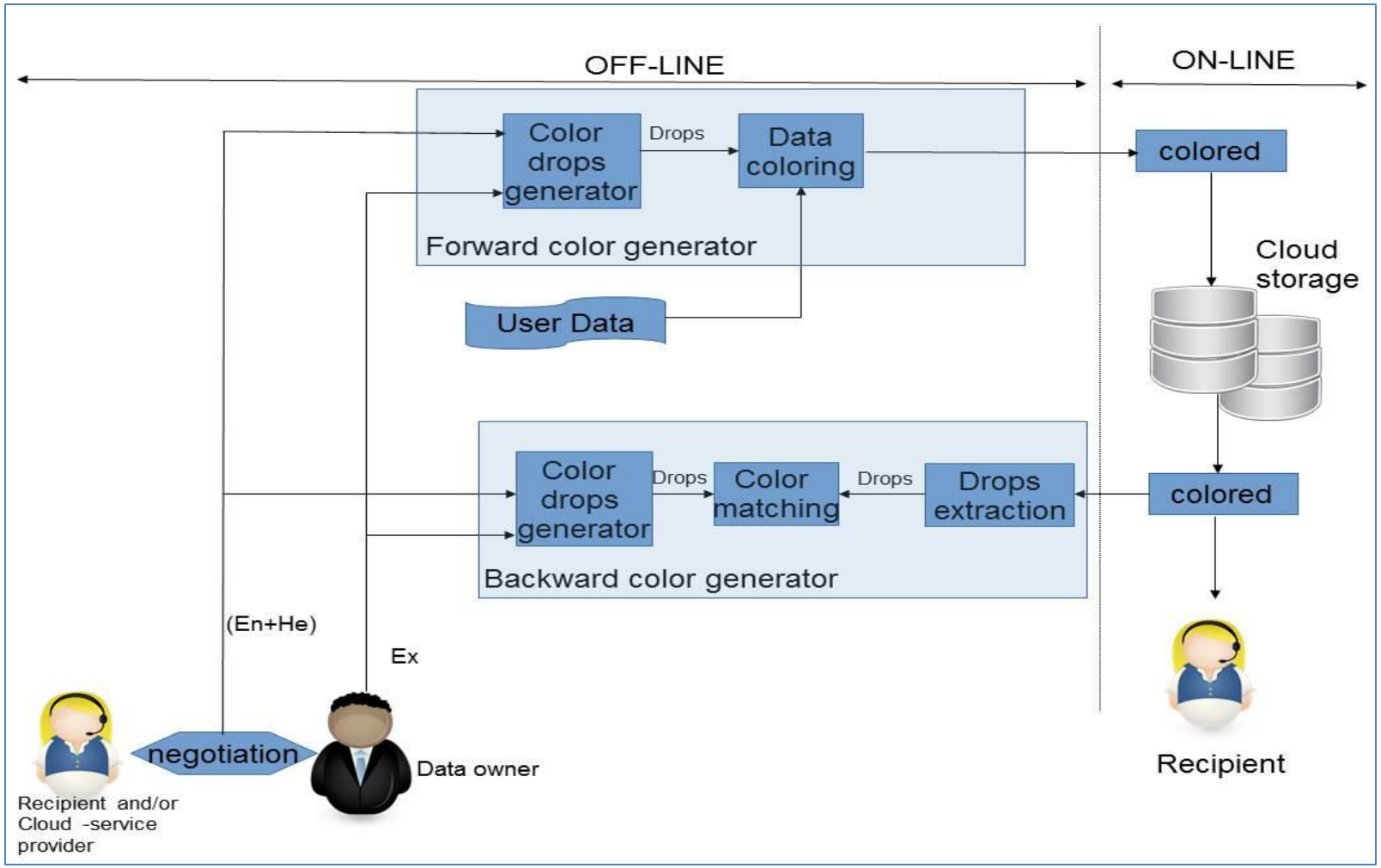
Digital Certificates

- Similar to travel passport
  - Provides forgery resistant identifying information
    - Name of holder
    - Serial number
    - Expiration date
    - Copy of holder's public key (used for encryption)
    - Digital signature of issuing authority (CA)

SSL Transport



Data colouring



Conclusion

- The potentiality of Big Data is now all around us in our everyday lives. Every device will be connected and constantly generating data.
- Good mapping of big-data is fundamental to understanding/selecting infrastructure (compute & storage) , analytics, visualization and protection (security and privacy).
- New frontiers such as “Data Science” is bringing many of the ideas/techniques from Big-Data Analytics to almost any field or discipline.

2017 opportunities @ ICTP

- Workshop on Open Source Solutions for the Internet of Things – June 28 – July 7th, 2017
- The CODATA RDA Advanced School of Research Data Science for Extreme sources of Data, Bioinformatics and IoT/Big-Data Analytics – July 3rd -28th, 2017
- Two other CODATA/RDA schools on Data Science
  - South Africa and Brazil , *maybe a HPC school in Mexico*
- Masters degree in HPC, Trieste, Italy
- Graduate studies @ East African Institute for Fundamental Research (EAIFR), Kigali, Rwanda

References

- Michael Cooper & Peter Mell, “Tackling Big Data”, NIST Information Technology Laboratory, 2010
- Big Data Working Group, “Big Data Taxonomy”, Cloud Security Alliance, 2014
- M. Bornschlegl et al, “IVIS4BigData: A Reference Model for Advanced Visual Interfaces Supporting Big Data Analysis in Virtual Research Environments”, 2016
- S. Rajendran, Apache Storm: A scalable distributed & fault tolerant real time computation system, 2015

That’s all folks!!

questions