BIG DATA AGE VS. MOBILE AGE – A CHALLENGE FOR BOTH AREAS

PROFESSOR JOAN LU, UNIVERSITY OF HUDDERSFIELD, UK, 20 JULY 2014, PARIS, FRANCE

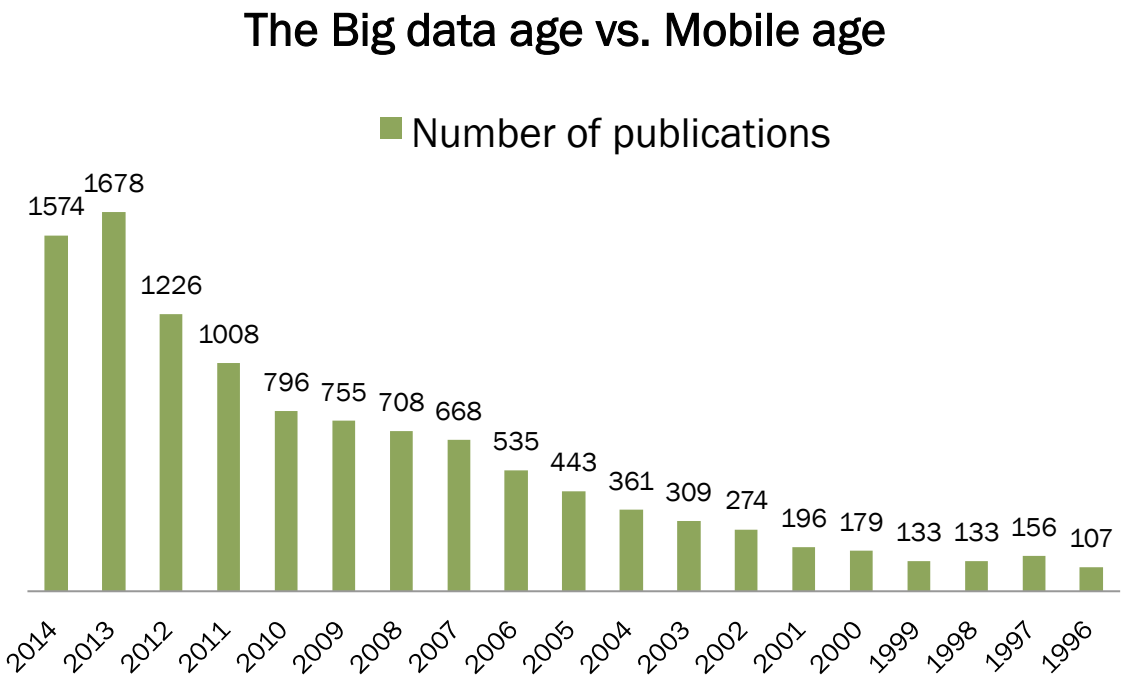
Welcome

Professor Joan Lu
Informatics,
University of Huddersfield
United Kingdom
Email: j.lu@hud.ac.uk

TODAY'S OVERVIEW

1. Trend of research on the big data age vs. mobile age
2. Mobile context
3. Challenges for the big data research in mobile age
4. Where we are and what is the next

RESEARCH BACKGROUND

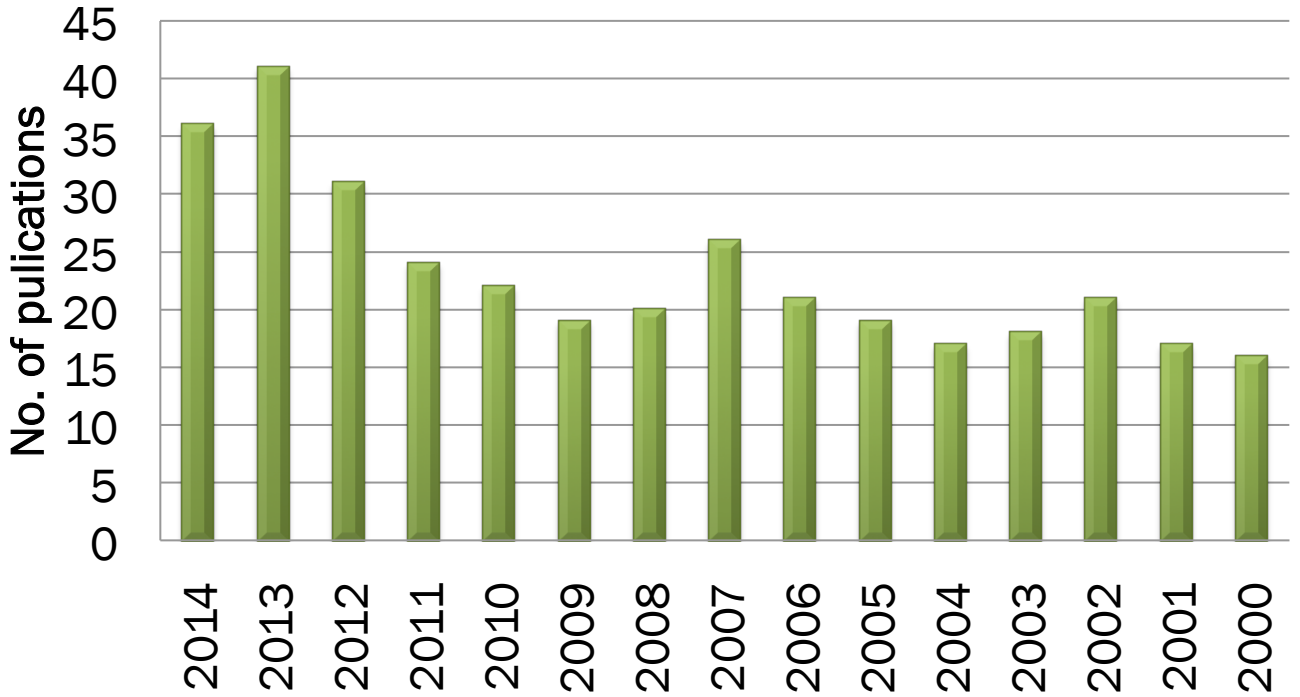


A new challenge for both areas

For 19 years the publication number is increased from 107 to 1574, more than 10 times.

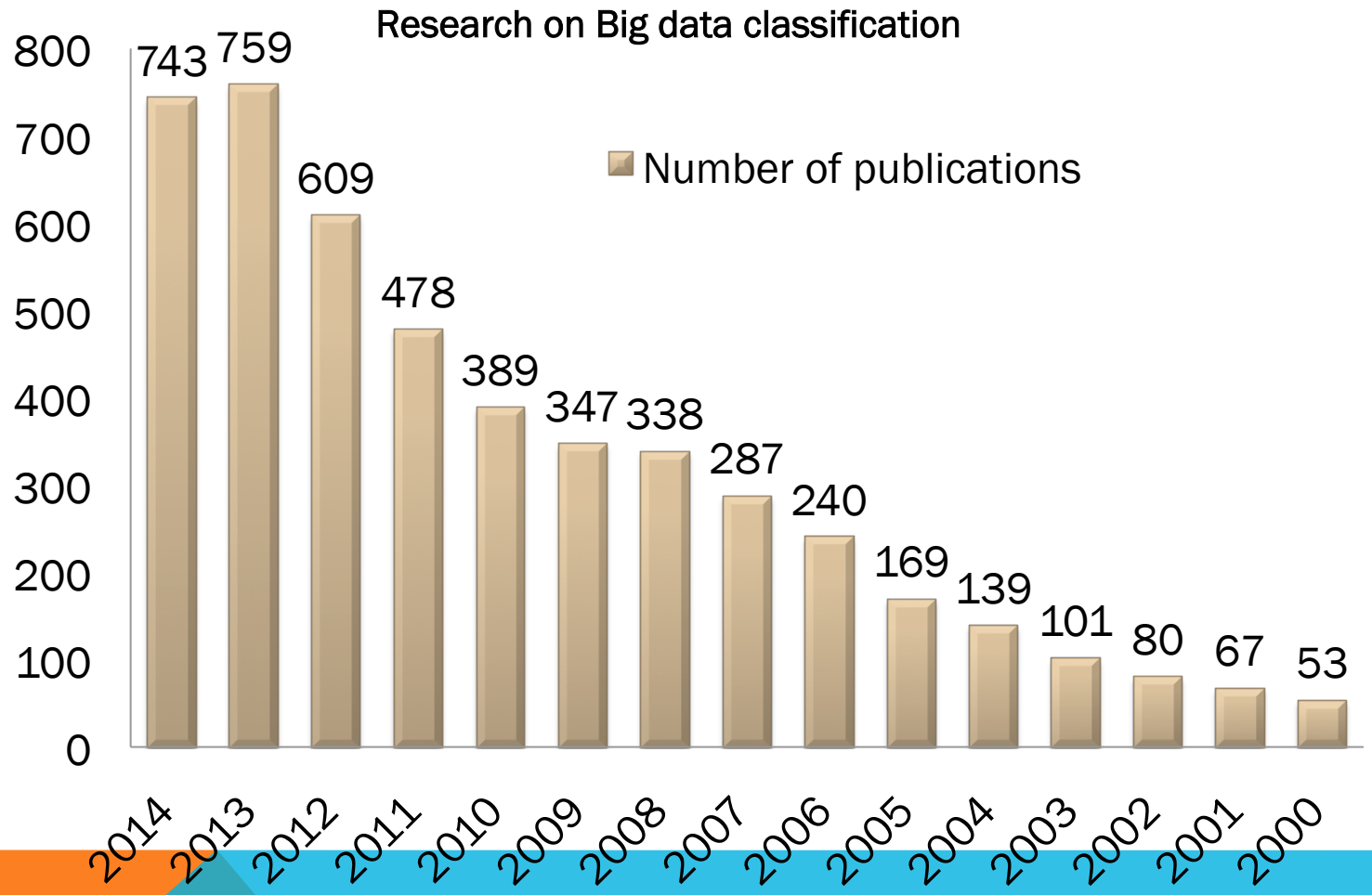
http://www.sciencedirect.com accessed by 16 July 2014

Research on Big data analytic tools



For 15 years, the number of publications is not increased significantly.

http://www.sciencedirect.com accessed by 16 July 2014



For 15 years, the publication number is increased about 14 times

http://www.sciencedirect.com accessed by 16 July 2014

Mobile context

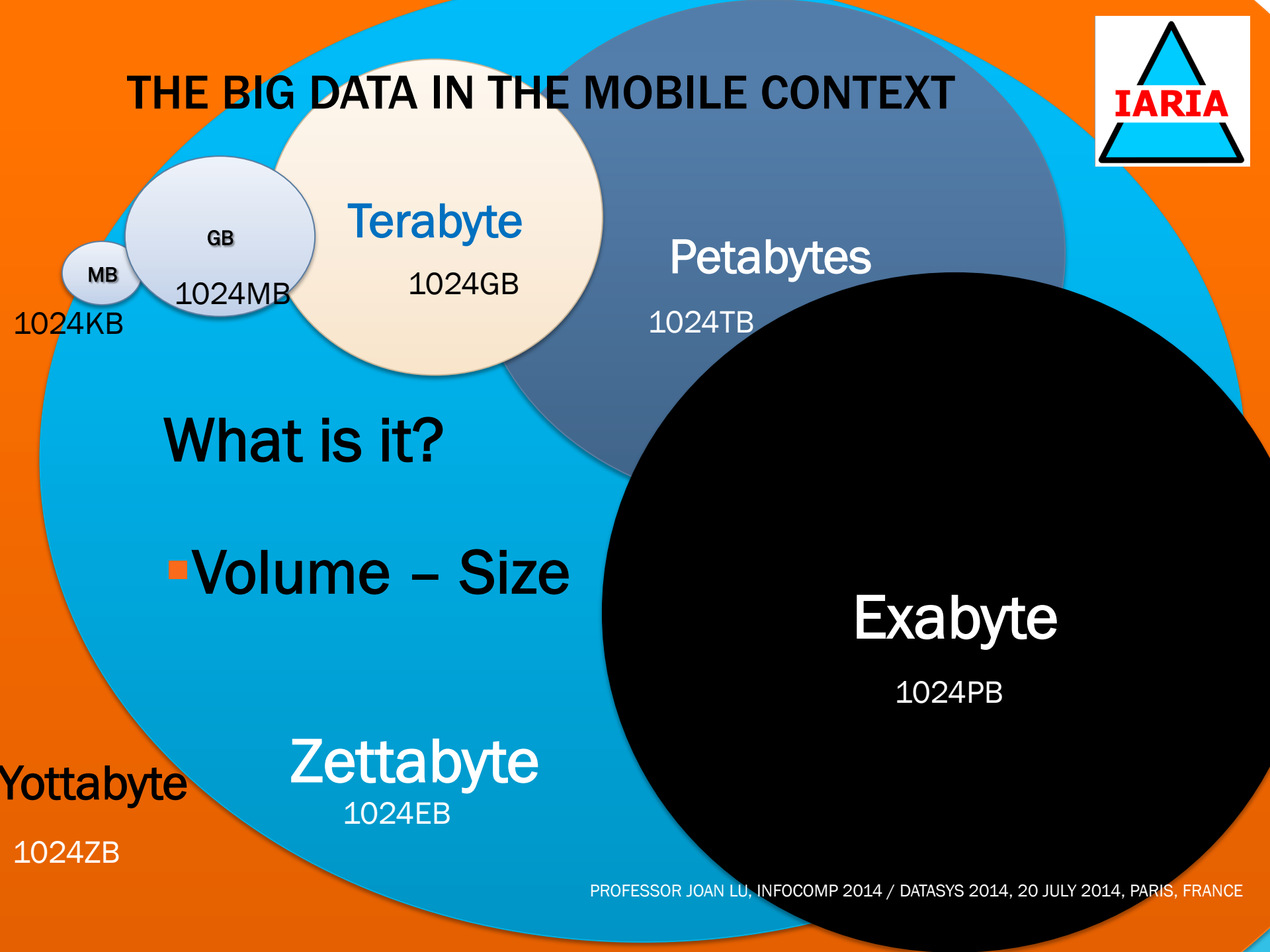
THE BIG DATA IN THE MOBILE CONTEXT

What is it?

- Volume – Size
- Velocity – speed (in-out)
- Variety – types/resources
- Vitality – life cycle

Current popular concepts are discussed by a number of publications, e.g. Zikopoulos, et al.[31]

THE BIG DATA IN THE MOBILE CONTEXT



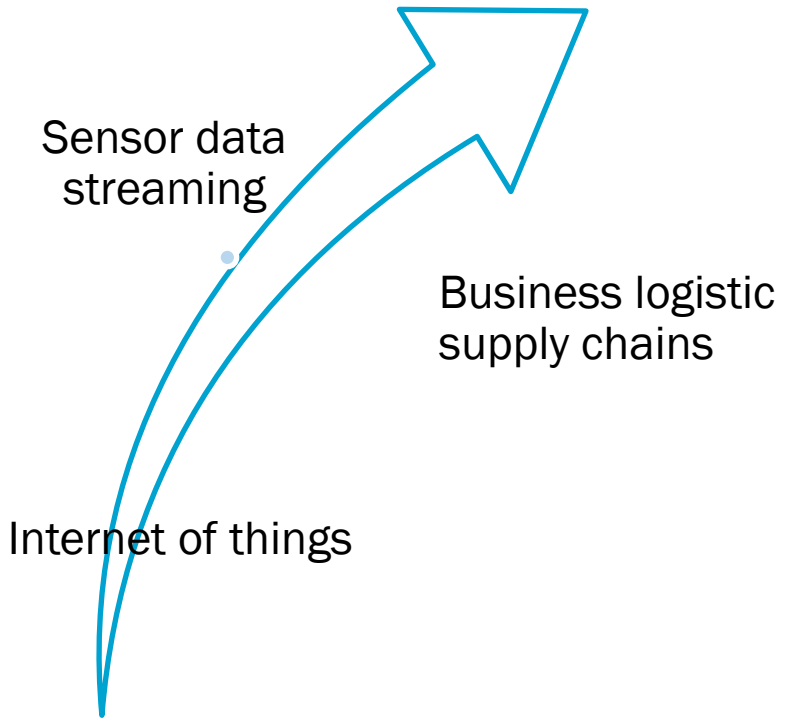
What is it?

- Volume – Size

THE BIG DATA IN THE MOBILE CONTEXT

Velocity – speed (in-out)

- Sensor data streaming:
  - Transportation
  - Smart city
  - Police communications
  - Manufacturing production lines
- Business:
  - logistic supply chain
  - e-commerce
  - e-banking
  - e-finance
- Internet of things
  - Tracking sale records, product flow paths, etc.
  - Multiple sectors involved



THE BIG DATA IN THE MOBILE CONTEXT

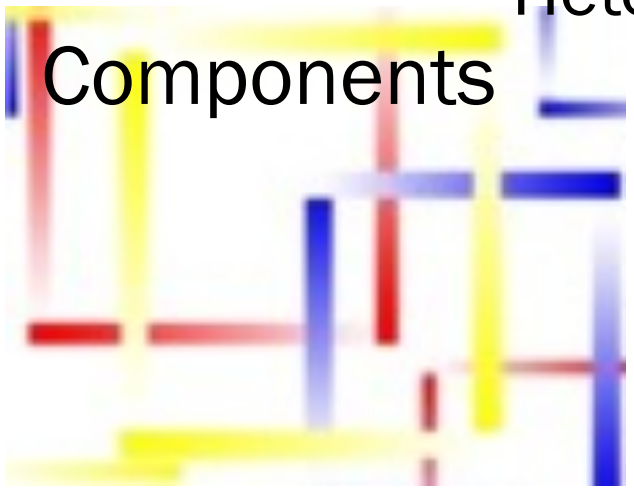
Variety – types/resources

A big challenge to the data workers is:

- Unstructured data or semi- structured data
- e.g. audio, video, other data streaming, etc.

THE BIG DATA IN THE MOBILE CONTEXT

Heterogeneity



Components

- Tables
- Text documents
- Math notations
- Static and dynamic Graphics
- Etc.

Formats

- MSWord/Latex
  - Reports
  - Papers
- Spreadsheets
  - Numeric data, including some visual presentations
- Etc.

Datatypes

- character
- string
- integer
- float
- double
- Etc.

THE BIG DATA IN THE MOBILE CONTEXT

Variety – types/resources

- Text,
- Images
- Video clips
- RFID tags
- Documents
- Integrated web pages
- Instant snap shots, e.g. cameras, detective instrumer
- Etc.,

http://www.123rf.com/ accessed 11, May 2014
http://www.bbc.co.uk/ accessed 11, May 2014

THE BIG DATA IN THE MOBILE CONTEXT

Technology assigns new meaning for the mobile at the mobile age,

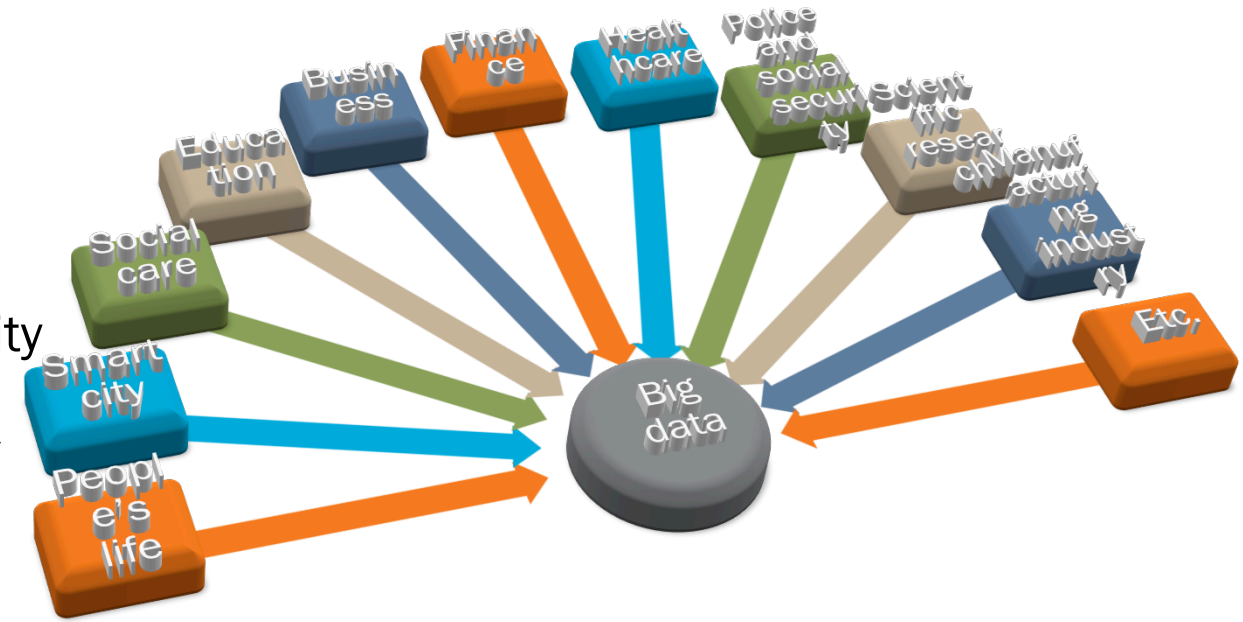
- Smartphone, e.g. iPhone, Microsoft windows phone, Android phone, Nokia phone, blackberry phone, etc.
- Handheld devices
  - Tablets, portable devices, laptop, iPad/iPod touch, etc.
- Industrial sensors
  - Embedded system, actuator, manufacturing workshops, etc.,





MOBILE AGE GENERATING BIG DATA

- Mobile is everywhere, the big data is pervasive
  - People's life
  - Smart city
  - Social care
  - Education
  - Business
  - Finance
  - Healthcare
  - Police and social security
  - Scientific research
  - Manufacturing industry
  - Etc.



MOBILE AGE GENERATING BIG DATA

- Manufacturing industry
  - for one sensor with sampling rate 5MHz, one hour data logging file will be: 3600x5000000x8byte= 144GB.
  - For an 8 hours' shift per day, one sensor will produce more than one TB data.
  - For a production line, it could have hundreds in one location or millions of sensors at multiple locations in use, multiple pet-bytes are usually achievable.

Massive datasets are generated by the sensor data streaming
The data are constantly communicated between sensors and machines

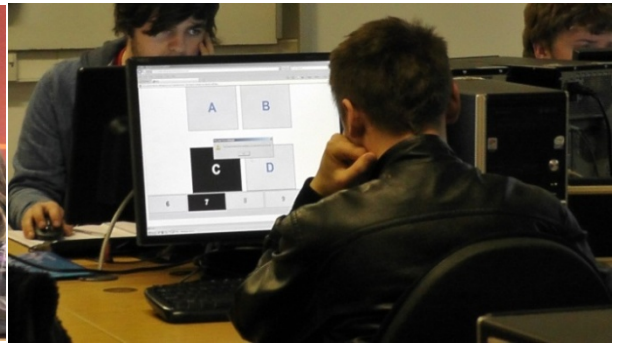
MOBILE AGE GENERATING BIG DATA

- Education
  - Student Response System as a case study in the mobile learning sessions, using the following pedagogical model.
  - In one class, if 20 to 100 students click responses from iPod/ iPad touch, what if for multi-classes, subjects, for multi-departments of over 1000 students, etc.

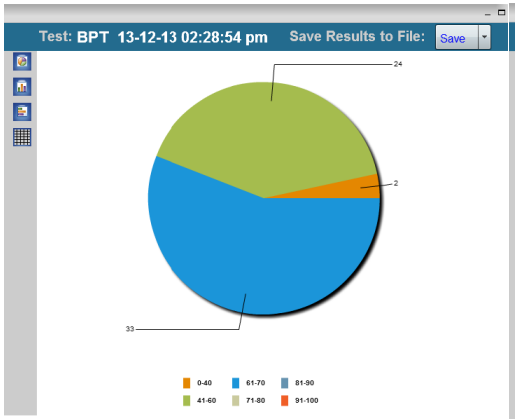
Massive datasets are generated by the responding – understanding process

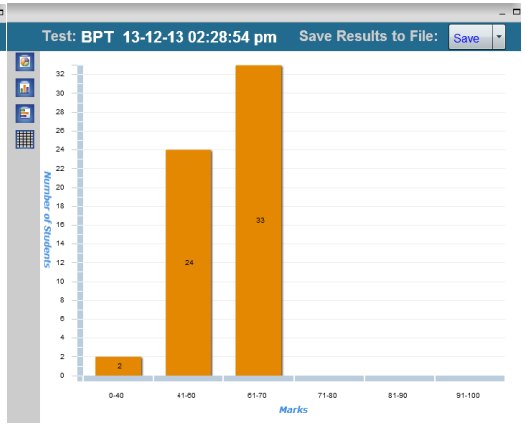
MOBILE AGE GENERATING BIG DATA

Understanding – use student response system



Knowledge harvesting – assessment



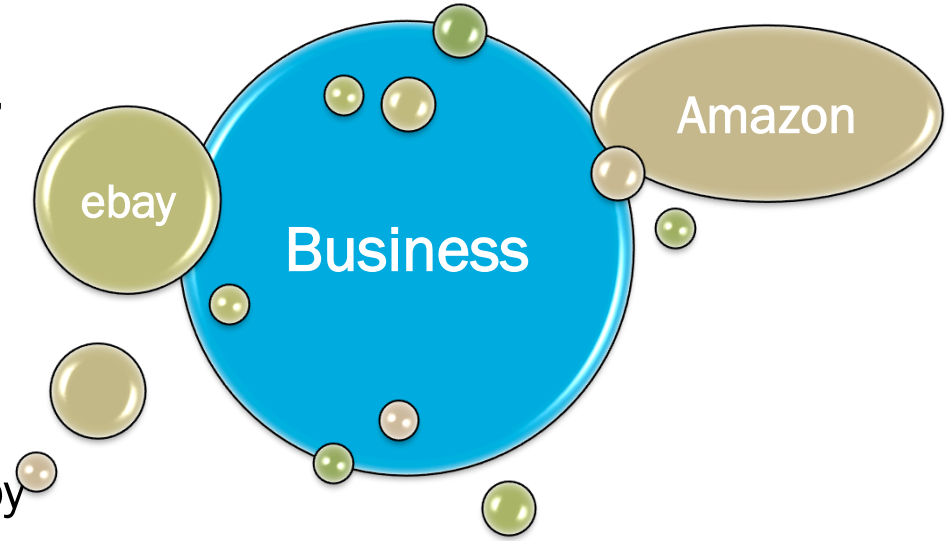


- Massive datasets are generated by assessment process.

- The data will be constantly communicated between human actions and machine access retrieval interfaces.

MOBILE AGE GENERATING BIG DATA

- Business,
  - One customer presses an order per day, what if for multi-customers press the multi-orders per minute, per day, or per week, per month, per year, from single chain of product supply, or from multi-chains of product supply, etc.

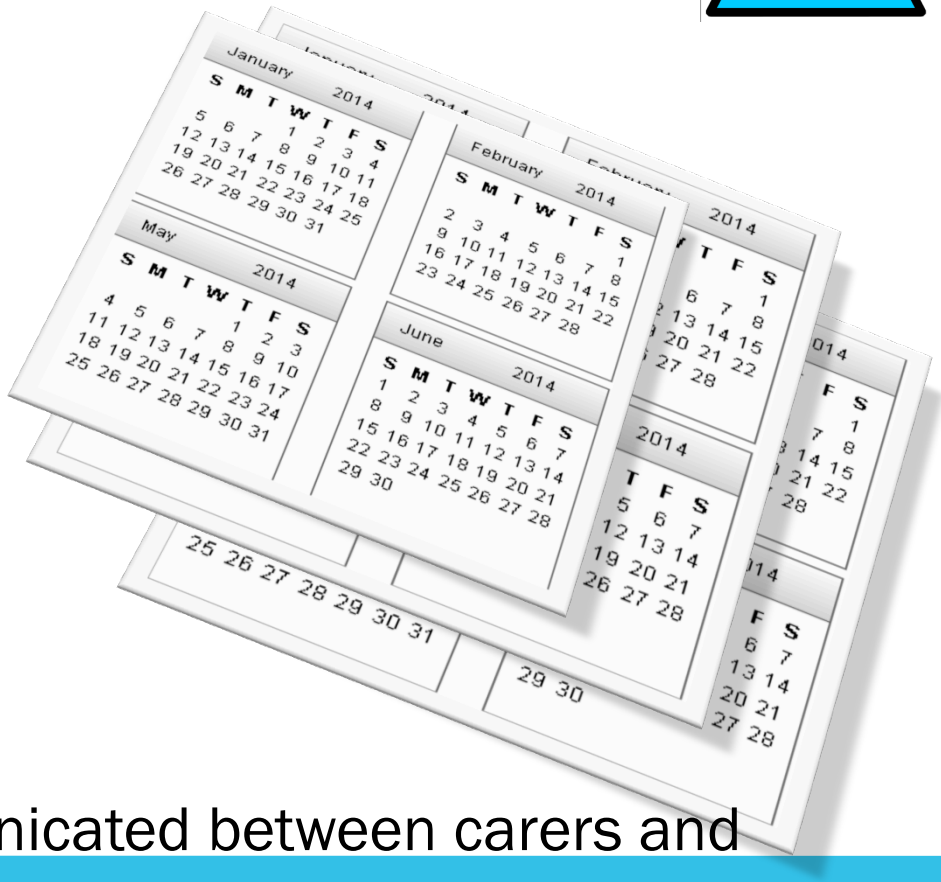


  - Massive datasets are generated by ordering process
- The data are constantly communicated between customers and suppliers.

Mobile age generating the big data

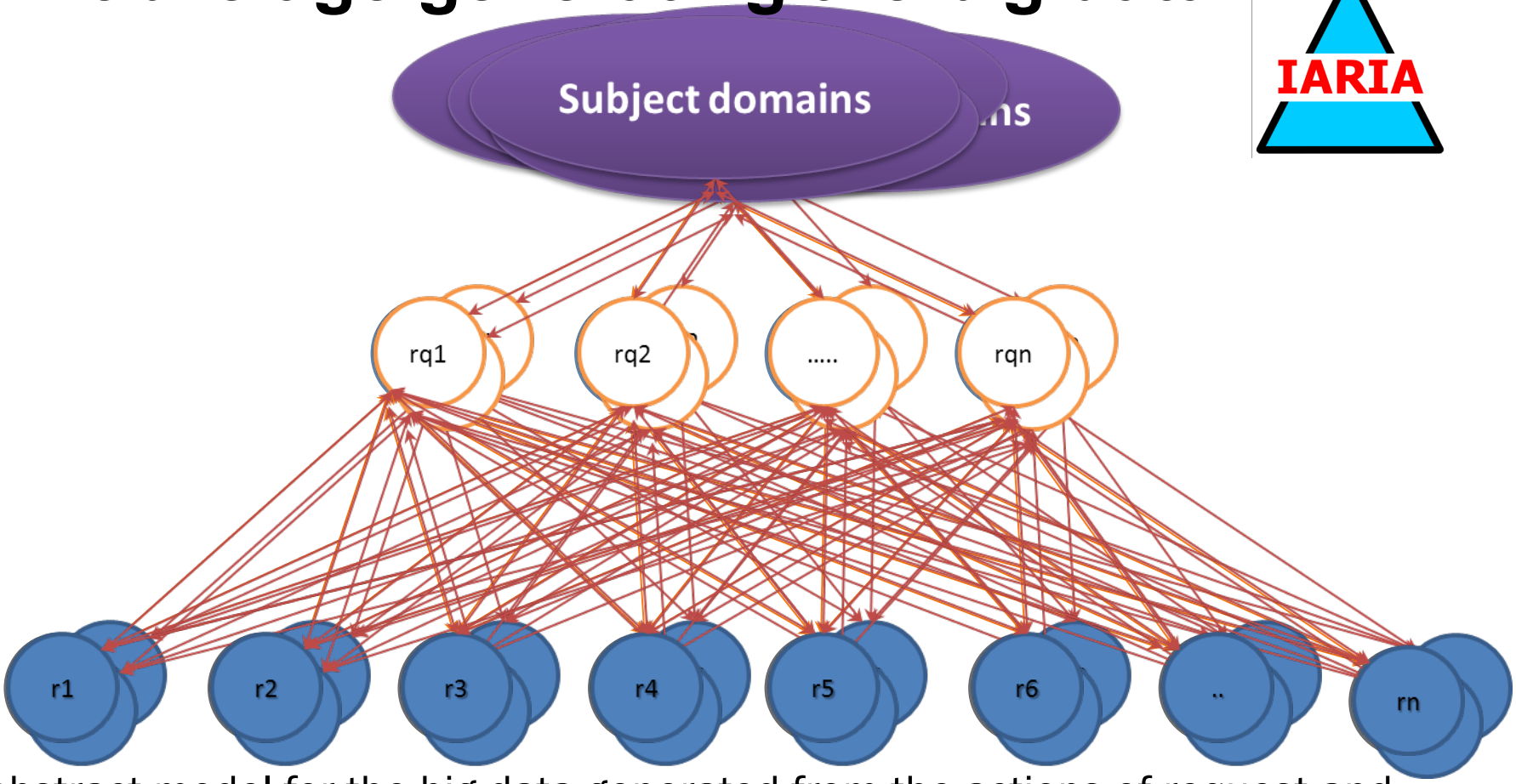
Healthcare or social care,

- For one person to make an appointment per day at one location, what if multi-people for multi-appointments, per week, per month, per year, at multi-locations,
- Massive datasets are generated by single or multi-care notes, text messages, etc.,



The data are constantly communicated between carers and careers.

Mobile age generating the big data



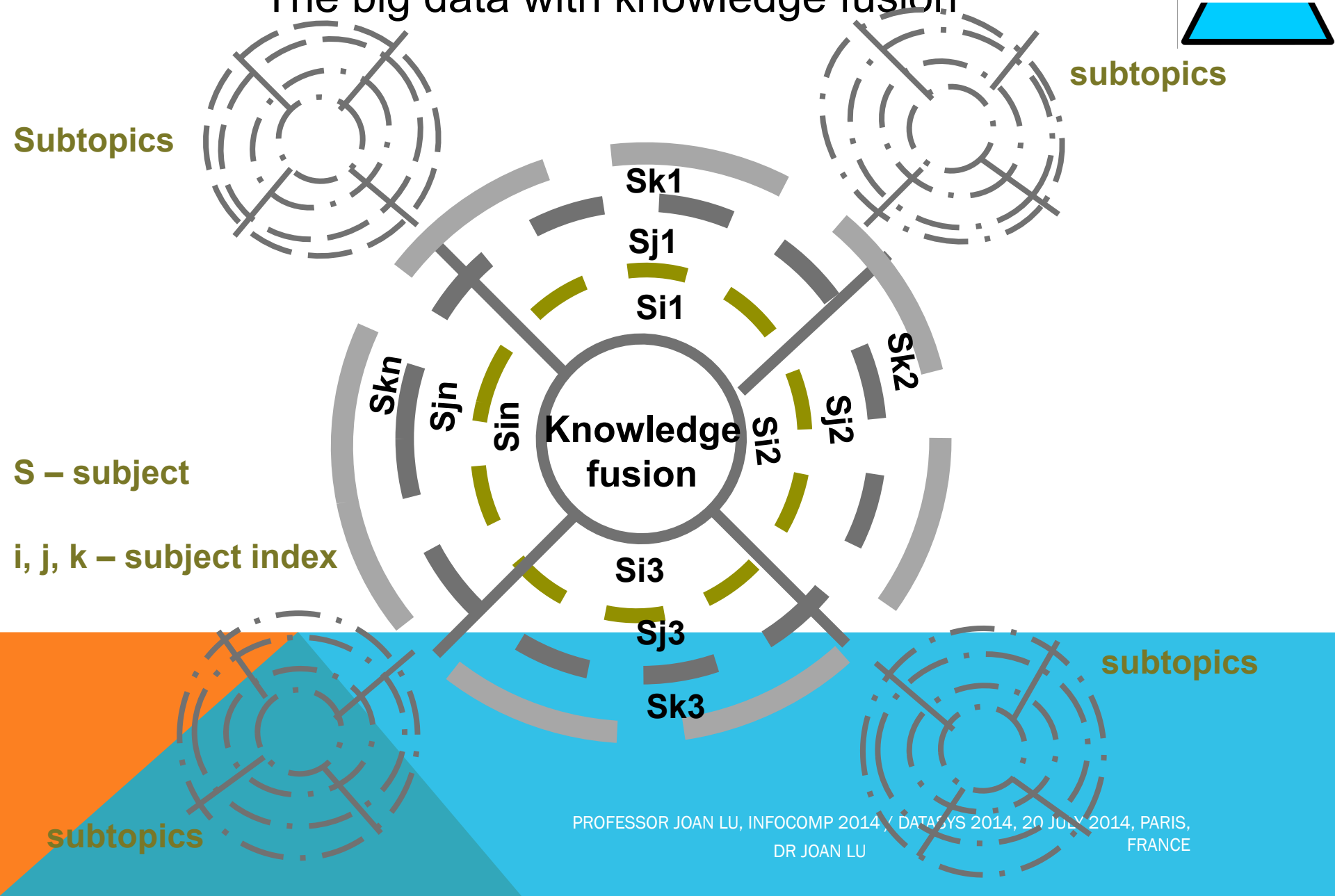
An abstract model for the big data generated from the actions of request and response system at the mobile age

$$SD = \sum rqi \longleftrightarrow \sum rj$$

Where SD means a subject domain, rqi stands for request sent from a device, rj stands for a specific response sent from a device.

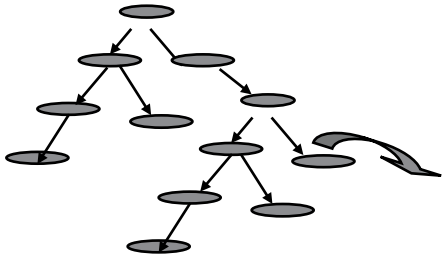
MOBILE AGE GENERATING THE BIG DATA

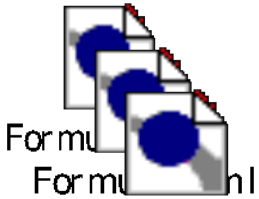
The big data with knowledge fusion



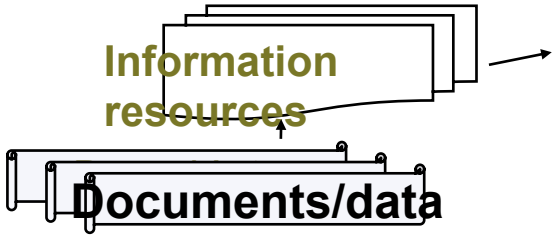
MOBILE AGE GENERATING THE BIG DATA

Subject Domain

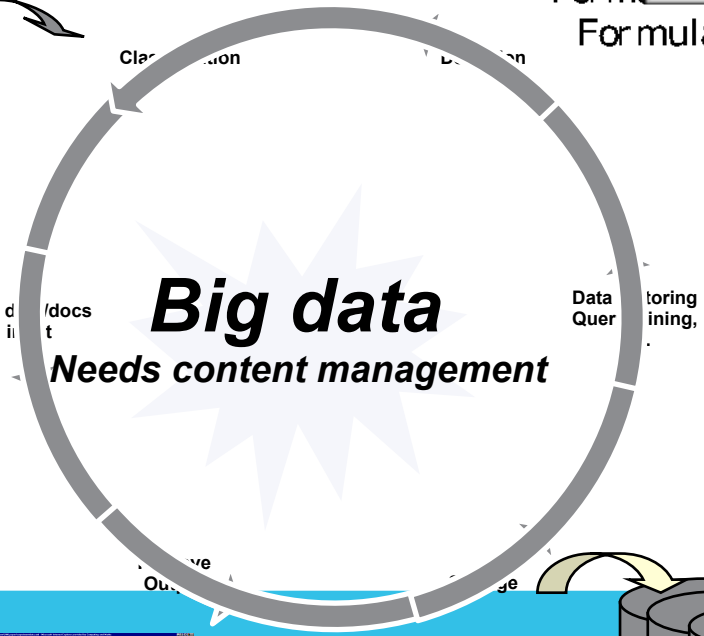


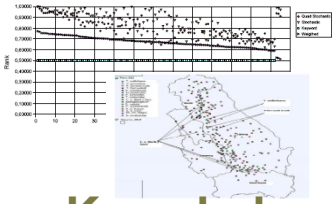


Data definition and representation

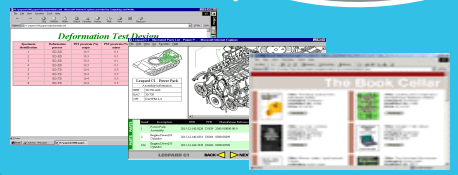


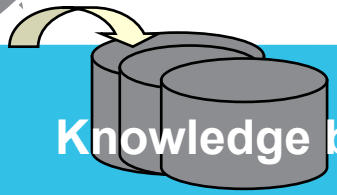
Information resources





Knowledge discovery



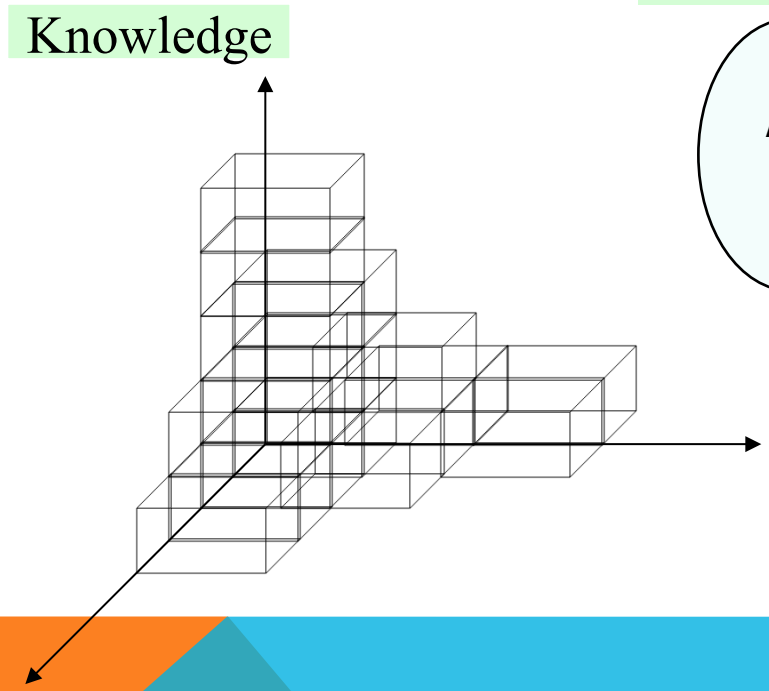


Knowledge base

MOBILE AGE GENERATING THE BIG DATA

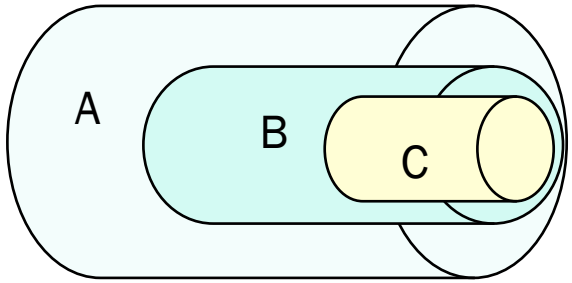
The Big data needs content management

Dimensionality

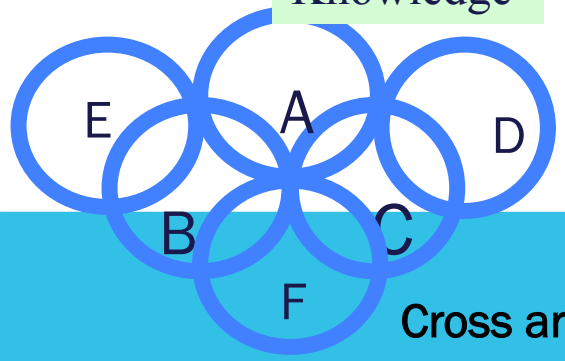


Knowledge

Inclusion



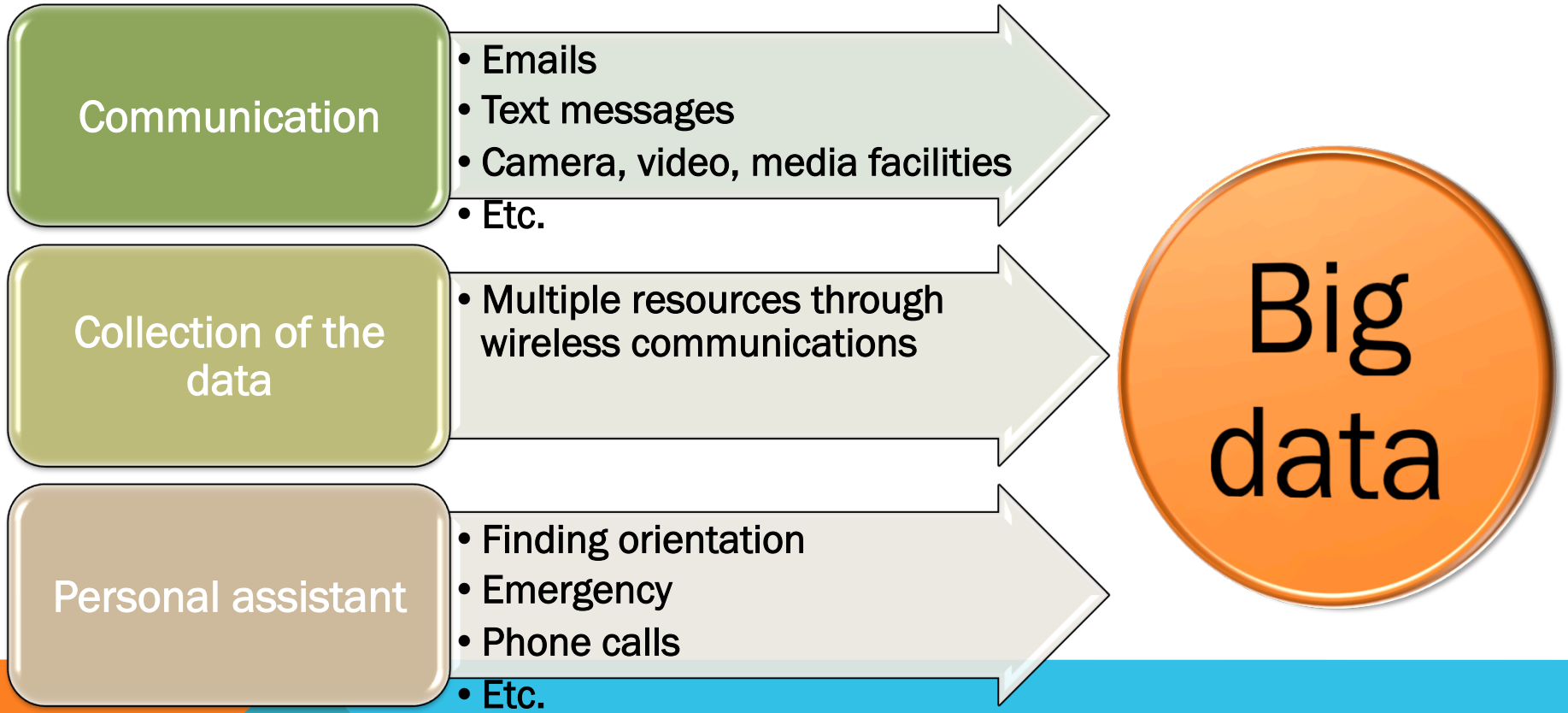
Knowledge



Cross areas

IMPACT ON THE REAL WORLD

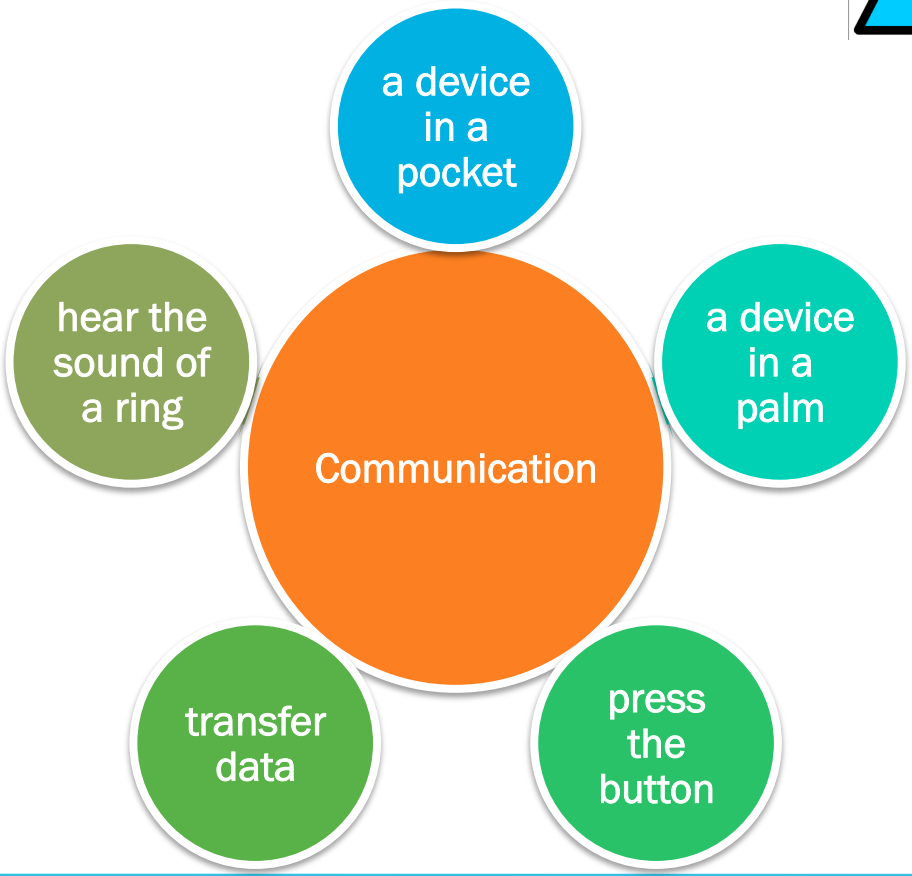
Mobile in people's life – what mobile can do:



IMPACT ON THE REAL WORLD

Communication via

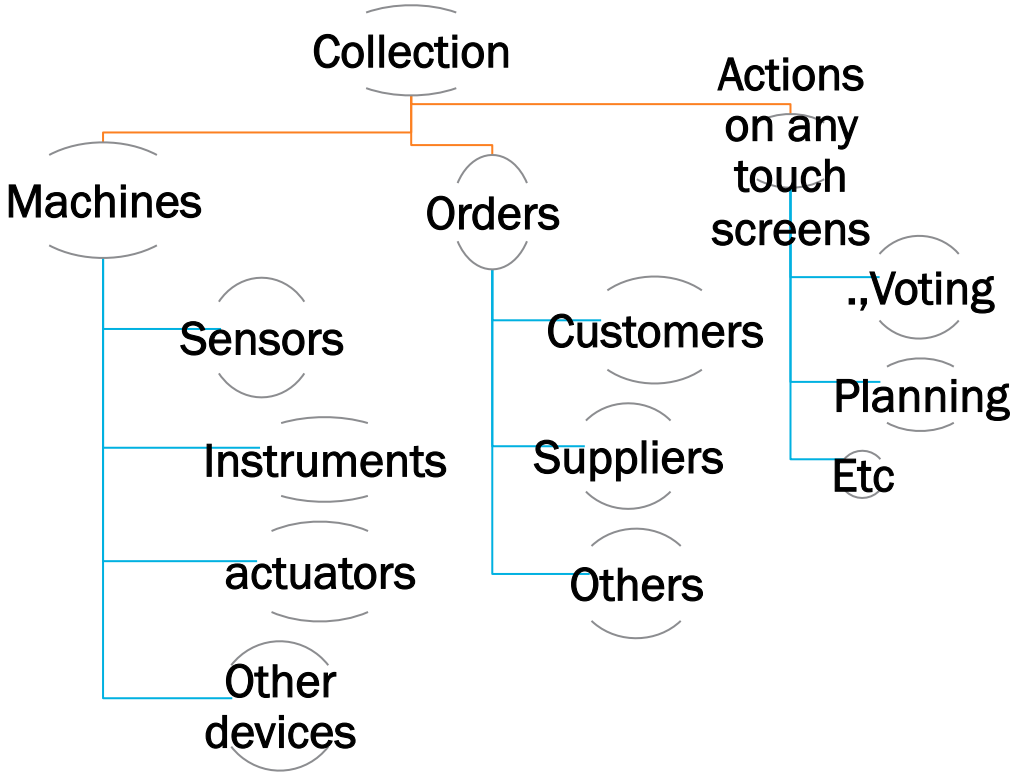
- Devices in a pocket, or a palm, e.g. Phones, PDAs, walking talking, etc.
- Signals, e.g. voice, sound of rings, images, video clips, etc.
- Data transfer, e.g. sending messages, making appointments,
- Actions, e.g. press buttons



IMPACT ON THE REAL WORLD

Data collections

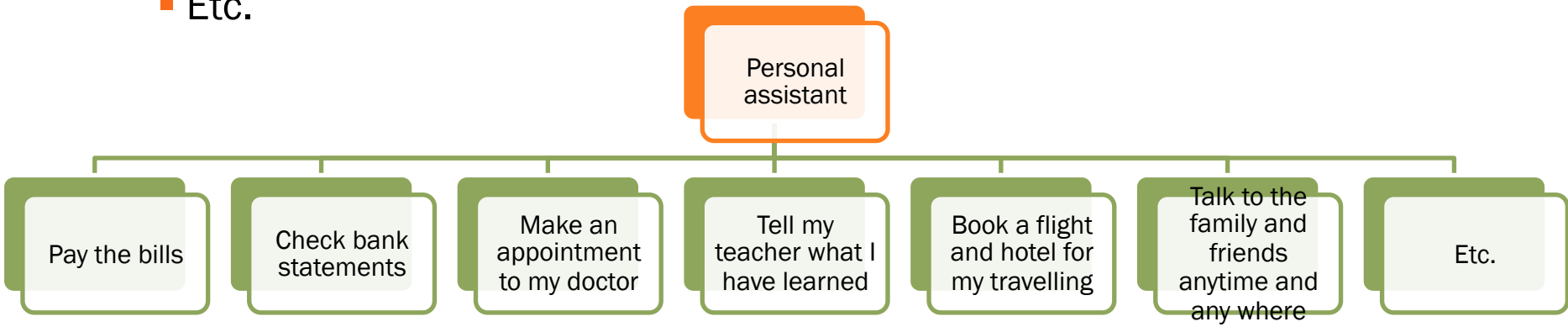
- Industrial production lines
  - Machines
    - Sensors
    - Instruments
    - Other mobile devices
- Business processes
  - Orders
    - Customers
    - Suppliers
    - Other third parties
- Other social events
  - Voting
  - Planning
  - Etc.



IMPACT ON THE REAL WORLD

Personal assistant

- Pay the bills
- Check bank statements
- Make an appointment to my doctor
- Tell my teacher what I have learned
- Book a flight and hotel for my travelling
- Talk to the family and friends anytime and any where
- Etc.



Challenges for the big data research in mobile age

PROFESSOR JOAN LU, INFOCOMP 2014 / DATASYS 2014, 20 JULY 2014, PARIS, FRANCE

ESSENTIAL FACING CHALLENGES IN THE BIG DATA ISSUES

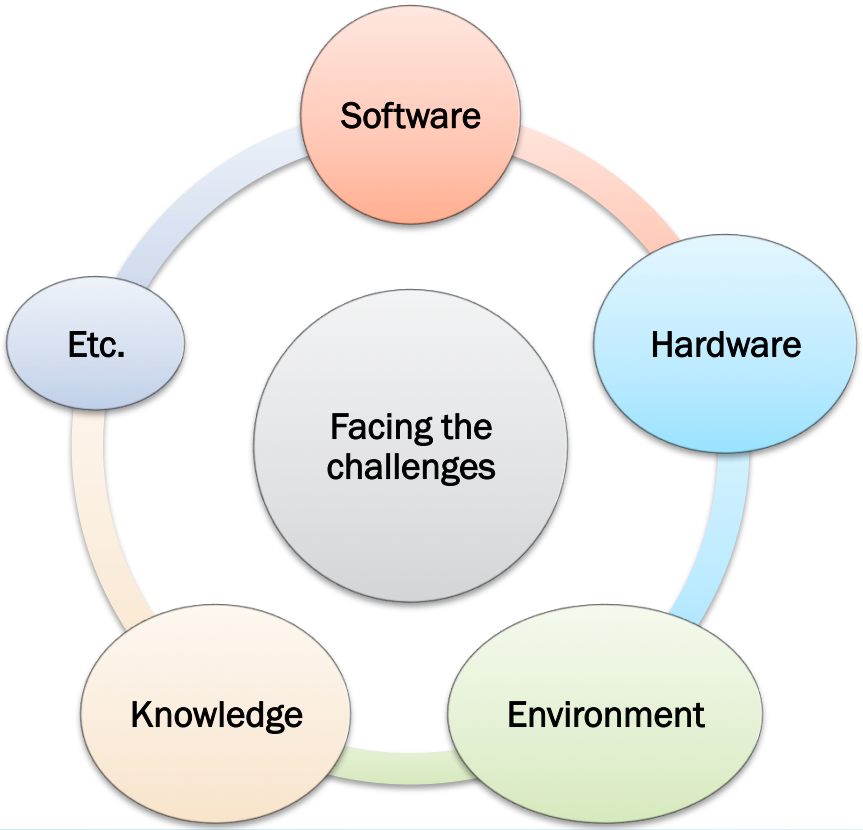
Software – open source, stand alone, etc.

Hardware – internal, external physic facilities, etc.

Environments – private cloud, public cloud, hybrid cloud, etc.

Knowledge – interdisciplinary, multidisciplinary expertise, etc.

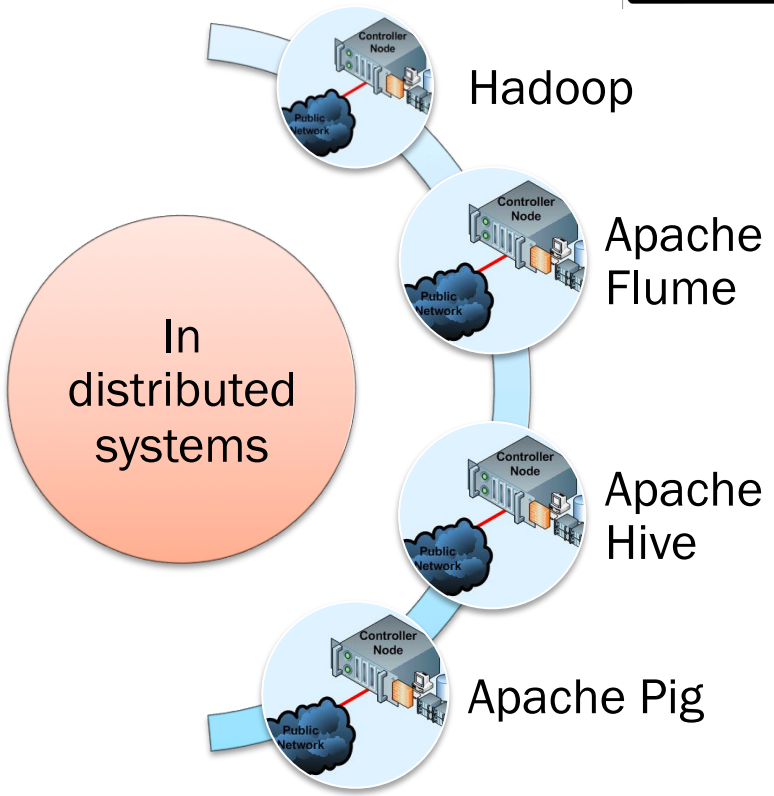
Etc.



CHALLENGES IN THE SOFTWARE

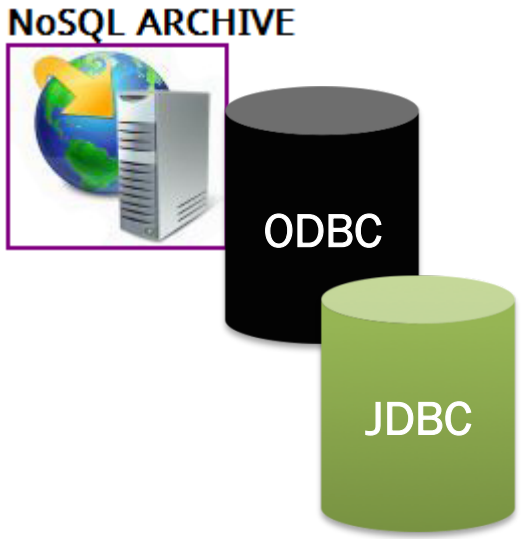
- **In distributed system**

- Hadoop[2-5] is open source software that is freely available from the apache.org source code repository.
- Apache Flume is a distributed system for collecting, aggregating, and moving large amounts of data from multiple sources into HDFS or another central data store[1].
- Apache Hive and Apache Pig are programming languages that simplify development of applications employing the MapReduce framework[1].



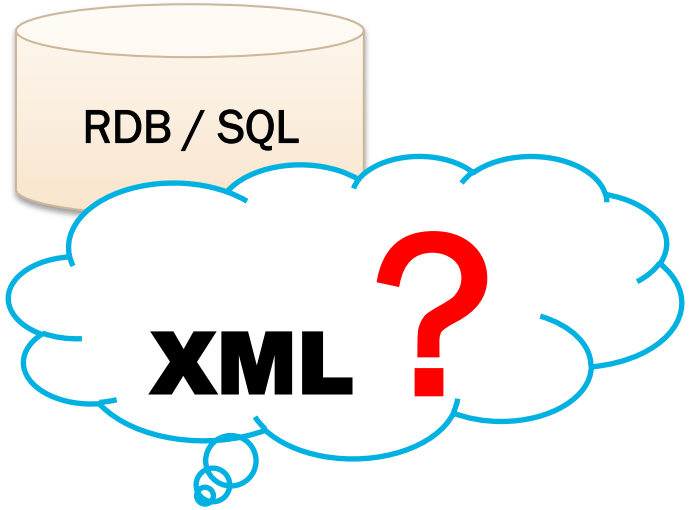
CHALLENGES IN THE SOFTWARE

- Database related
- Hbase is a database that can run on top of the Hadoop cluster and provides random, real time read/write access to data. But it is a NoSQL database [30].
- ODBC/JDBC Connectors for HBase and Hive are often proprietary components included in distributions for Apache Hadoop software[1]



CHALLENGES IN THE SOFTWARE

- Database related
- Apache Sqoop is a tool for transferring data between Hadoop and relational databases[1].
- Hive provides warehousing capabilities on top of the existing Hadoop cluster. It also provides an SQL like interface to the warehouse. Hive is also a batch processing system and is not a good fit if something real time is needed.

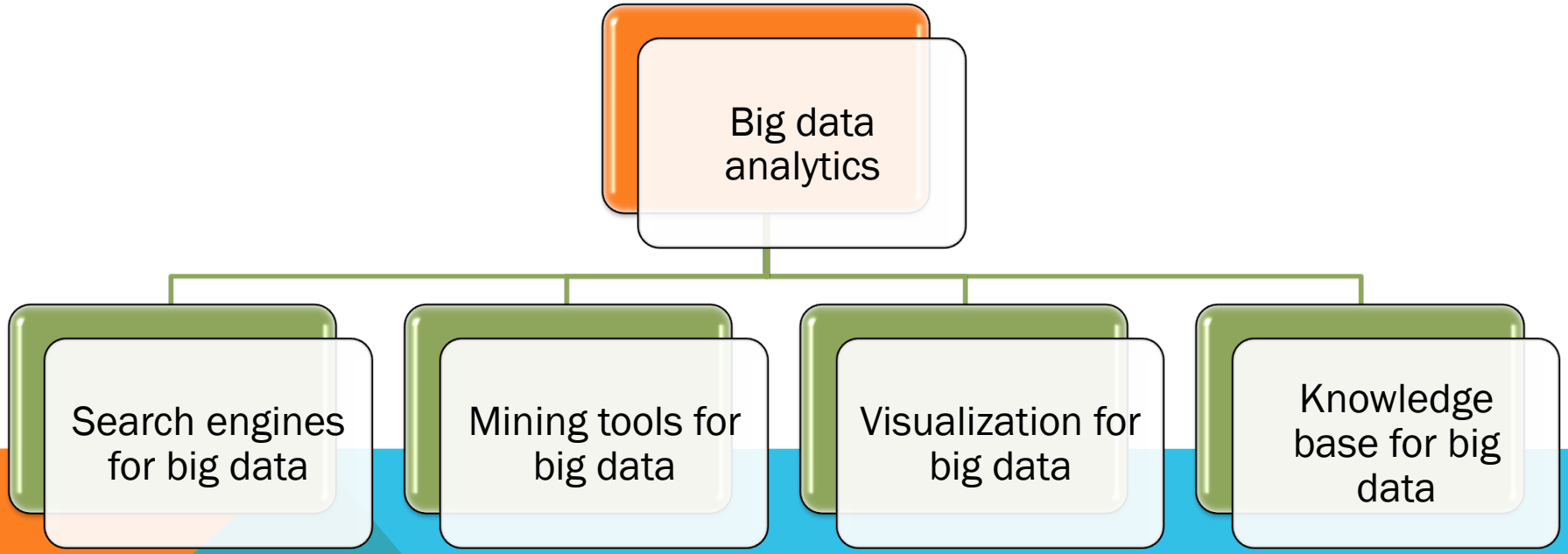


CHALLENGES IN THE SOFTWARE

Analytic tools for big data

From the UK's largest and most prestigious cross-industry event – BIG DATA ANALYTICS 2014: the big data is exponentially growing, every day about 2.5 quintillion bytes are created [28].

The big data needs analytic tools to process.



CHALLENGES IN THE SOFTWARE

Analytic tools for big data

Challenges to data workers:

Not many tools available on the market.

Most tools are at the experimental stage in the research labs. But IBM system Z and DB2 claimed to have the solution for big data analysis for business sector [29].

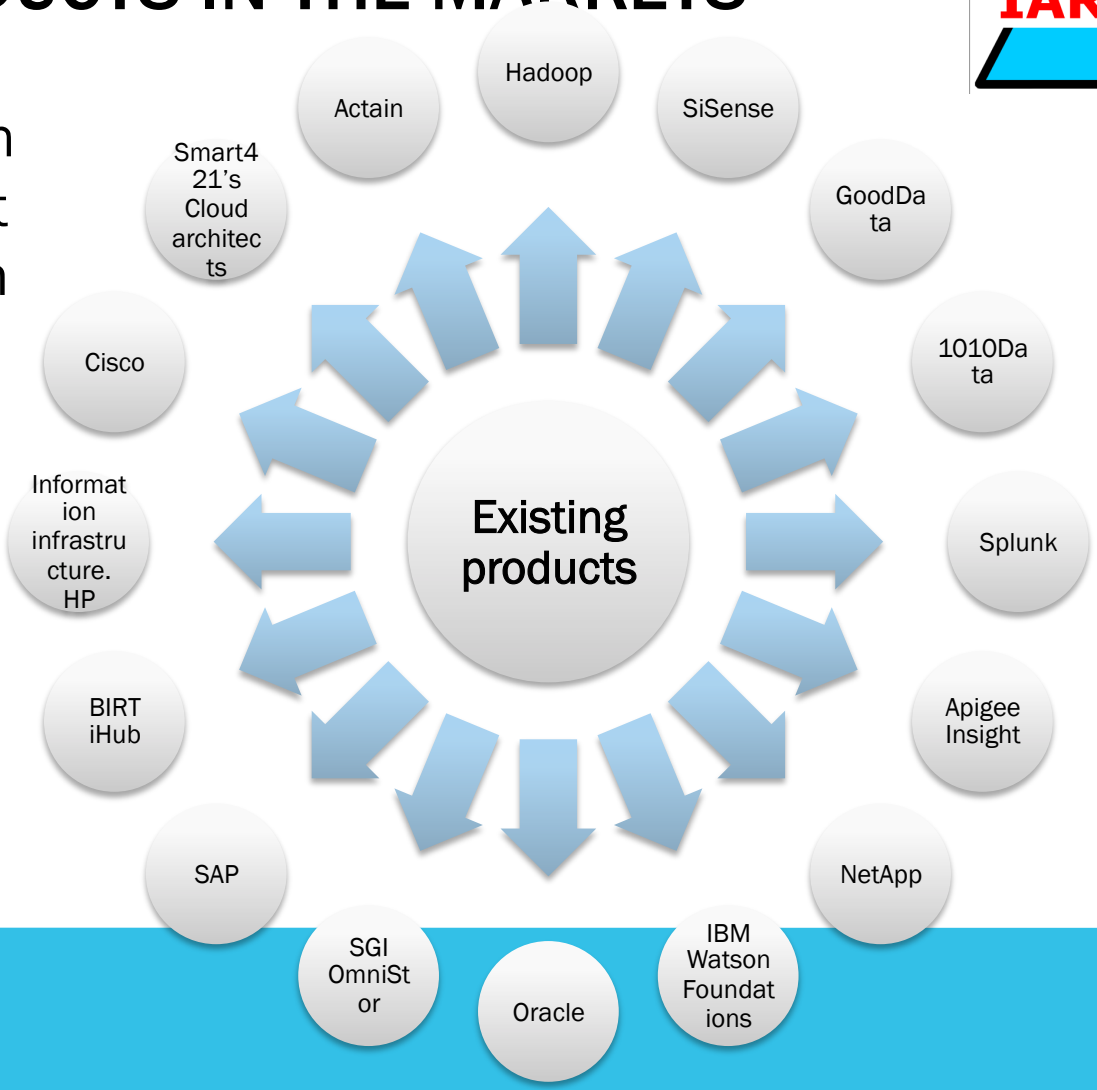
http://www.sciencedirect.com accessed by 16 July 2014

EXISTING PRODUCTS IN THE MARKETS

More than 16 products in the markets to claim that they are able to deal with the big data.

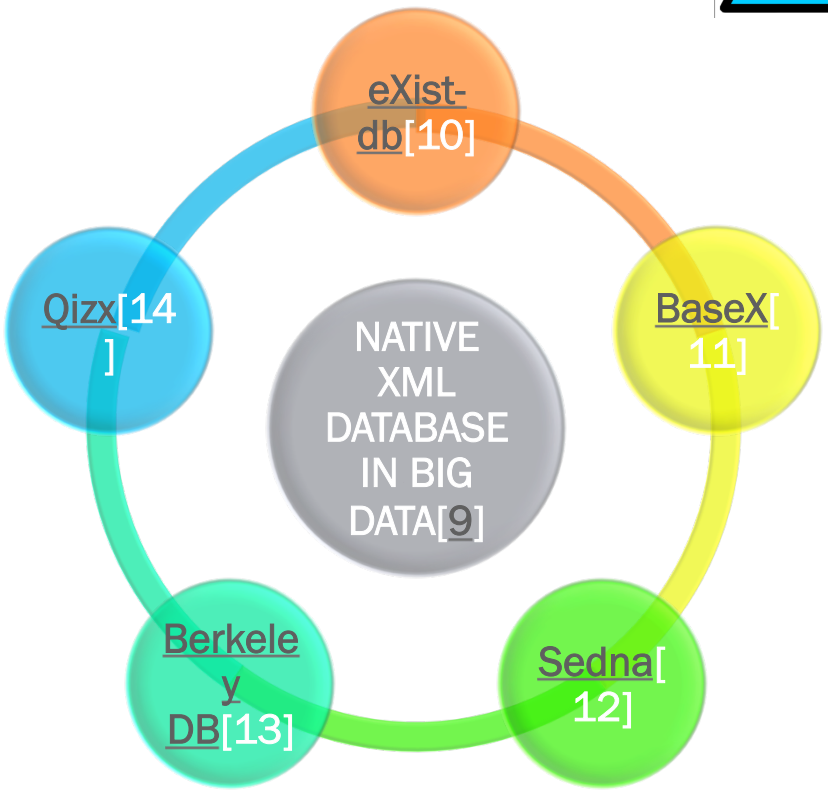
Challenges:

- Standalone and not open source for most products
- A huge financial pressure on SME



EXISTING PRODUCTS IN THE MARKET

- Dealing with big data, XML has its advantages, e.g.
  - Understandable by both machine and human beings.
  - Unstructured or semi-structured data
- It has been reported that about 5 NATIVE XML DATABASE used by BIG DATA [9]



CHALLENGES IN THE HARDWARE

The Intel Xeon processor E5 family provides a strong foundation for many Hadoop workloads[1].

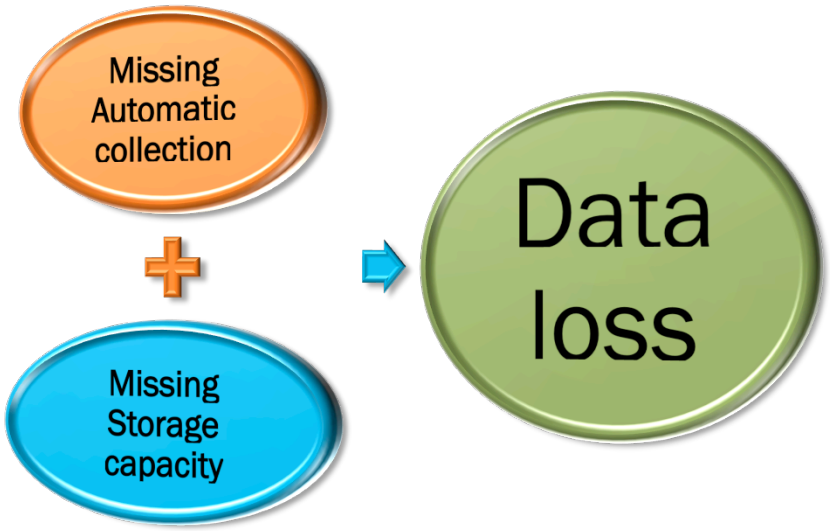
Hadoop typically requires 48 GB to 96 GB of RAM per server, and 64 GB is optimal in most cases[1].

Intel SSD 710 Series SATA SSDs provide significantly faster read/write performance that mechanical hard drives[1].

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The data loss:

- as there is missing facility for automatic data collection or data capture from variety of resources including devices, formats and volume, etc., and the lack of capacity to store generated data



CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The poor interoperability:

- as most products in market are standalone with limited accessibility not suitable for dynamic business patterns.

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

- The poor understanding by both human and machines:
  - as most products in the markets are still using traditional binary code packages, such as relational database.
  - They are neither semantically applicable for efficient information retrieval and knowledge discovery, nor suitable for the massive data streaming or sensor streaming in big data age.

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The high cost:

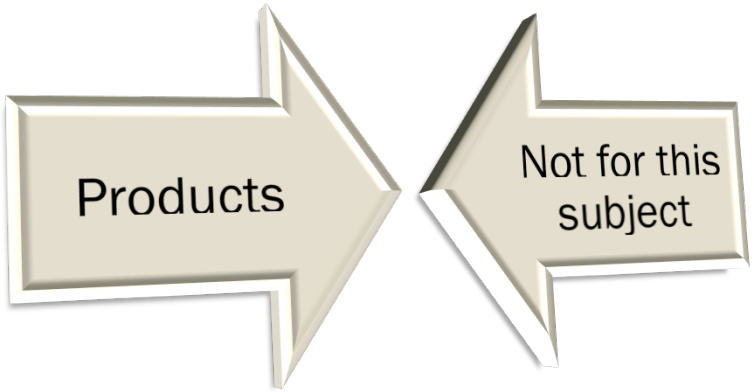
- A huge financial pressure for building up the infrastructures of big data supply chain, particularly for middle or SMEs;
- Additional cost on training for using products that need high level IT background or share the IT expertise.

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The poor adaptation for:

- Multi-disciplinary applications:

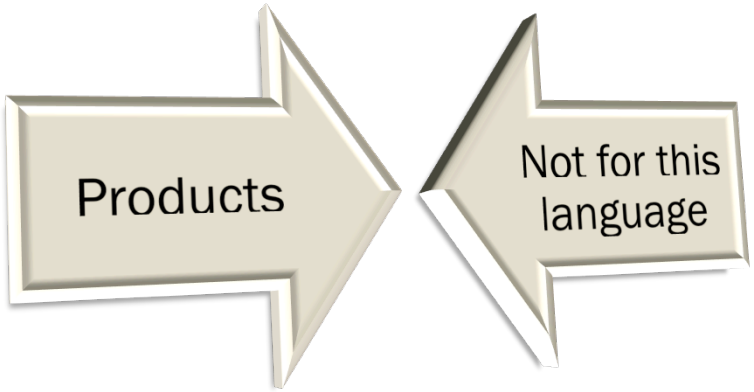
- Although a large numbers of products are available in the markets, they are rarely extended to be subject independent system, particularly for science and technologies



CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The poor adaptation for Multi-multilingual users:

- A large number of products are available for single lingual based application. The users will cost more to employ or train the workers to use the products produced by a specific language.



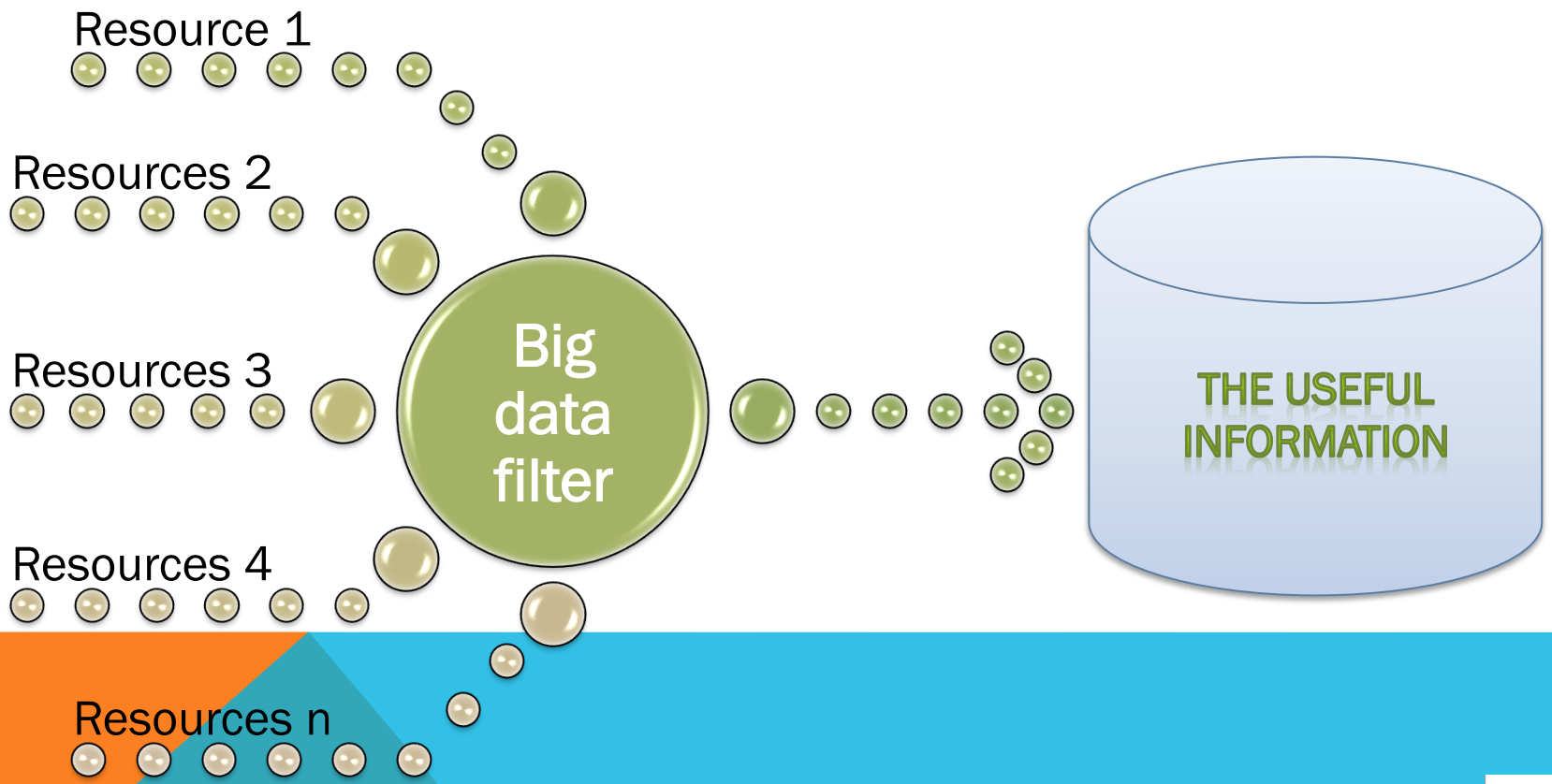
CHALLENGES FOR THE BIG DATA IN MOBILE AGE

Difficult to find the values from big data

- Traditional data mining techniques in:
  - **Theoretic models**
  - **Algorisms**
  - **Tools**
  - **Platforms**
- not suitable for the dimensions of 4Vs in the big data age
- Knowledge base built upon from the big data is needed.

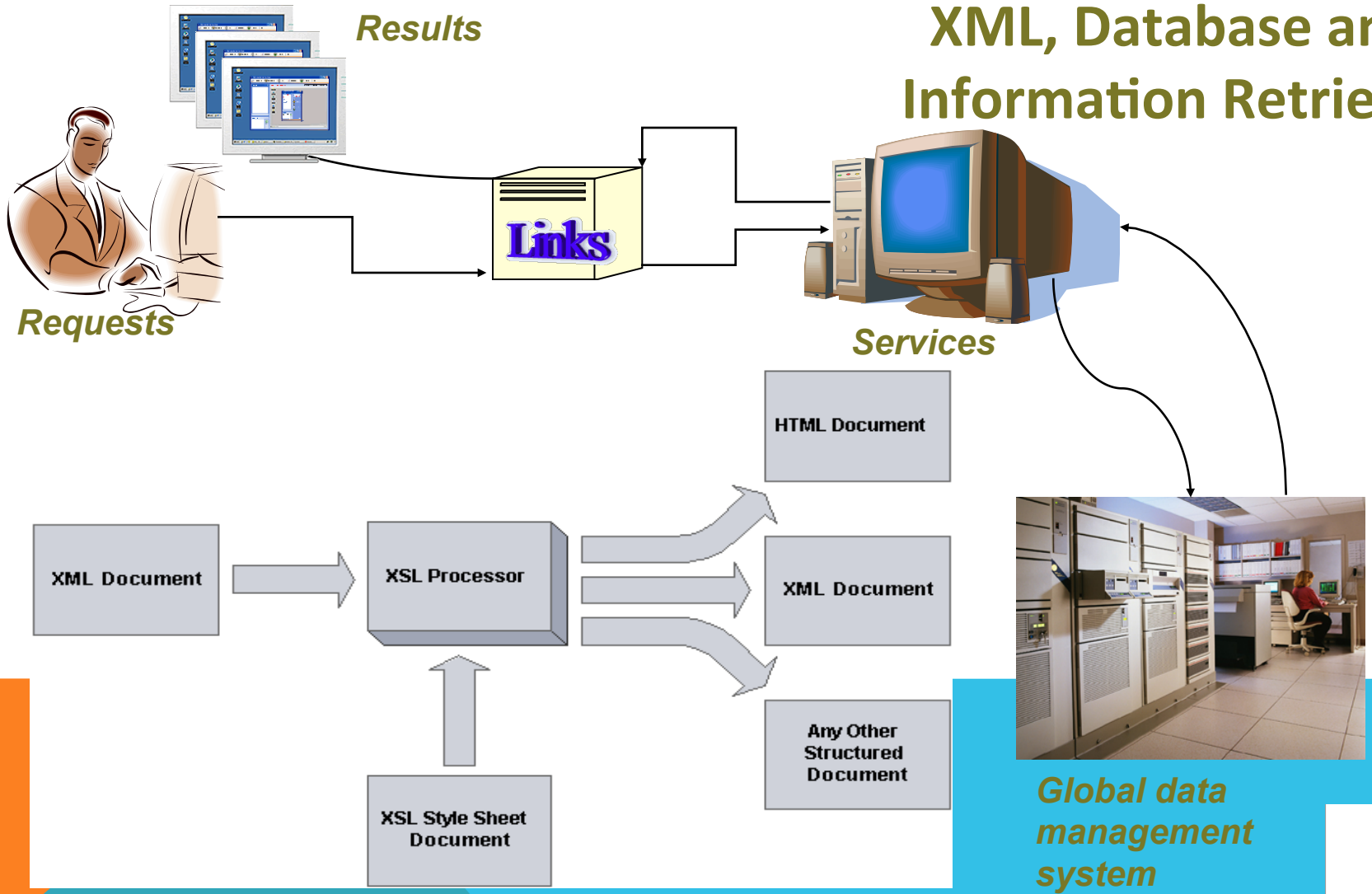
CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The Big data needs filter(s) to keep the useful information



CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The big data needs service oriented open framework



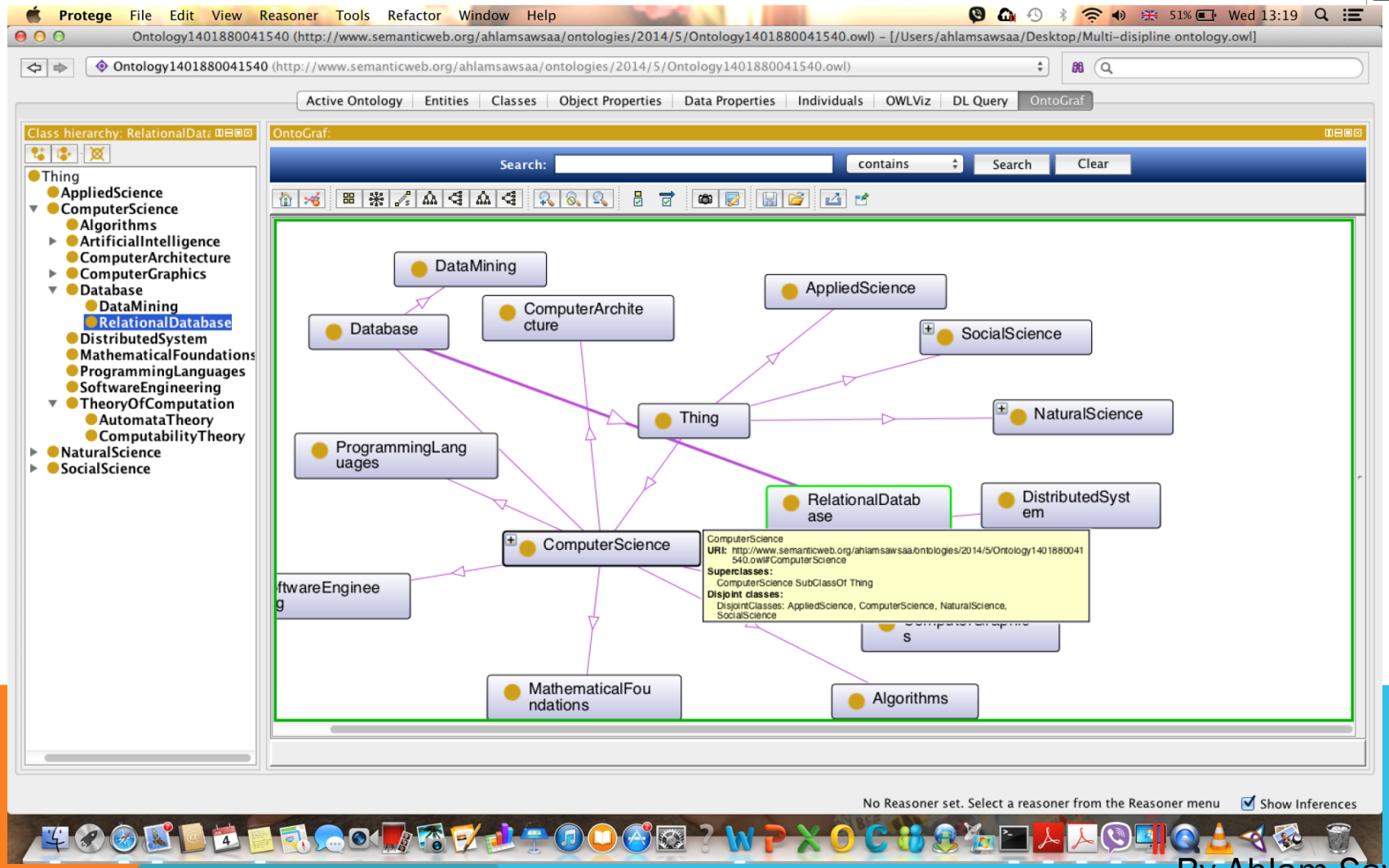
CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The big data needs classification for the useful information

- Rules, tools for the big data in information extractions and visualization
  - Taxonomic relationships or non-taxonomic relationships
  - Ontology
  - XML
  - GATE [32]
  - Protégé- OWL [32]
  - JAPE [32]
  - Etc.

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

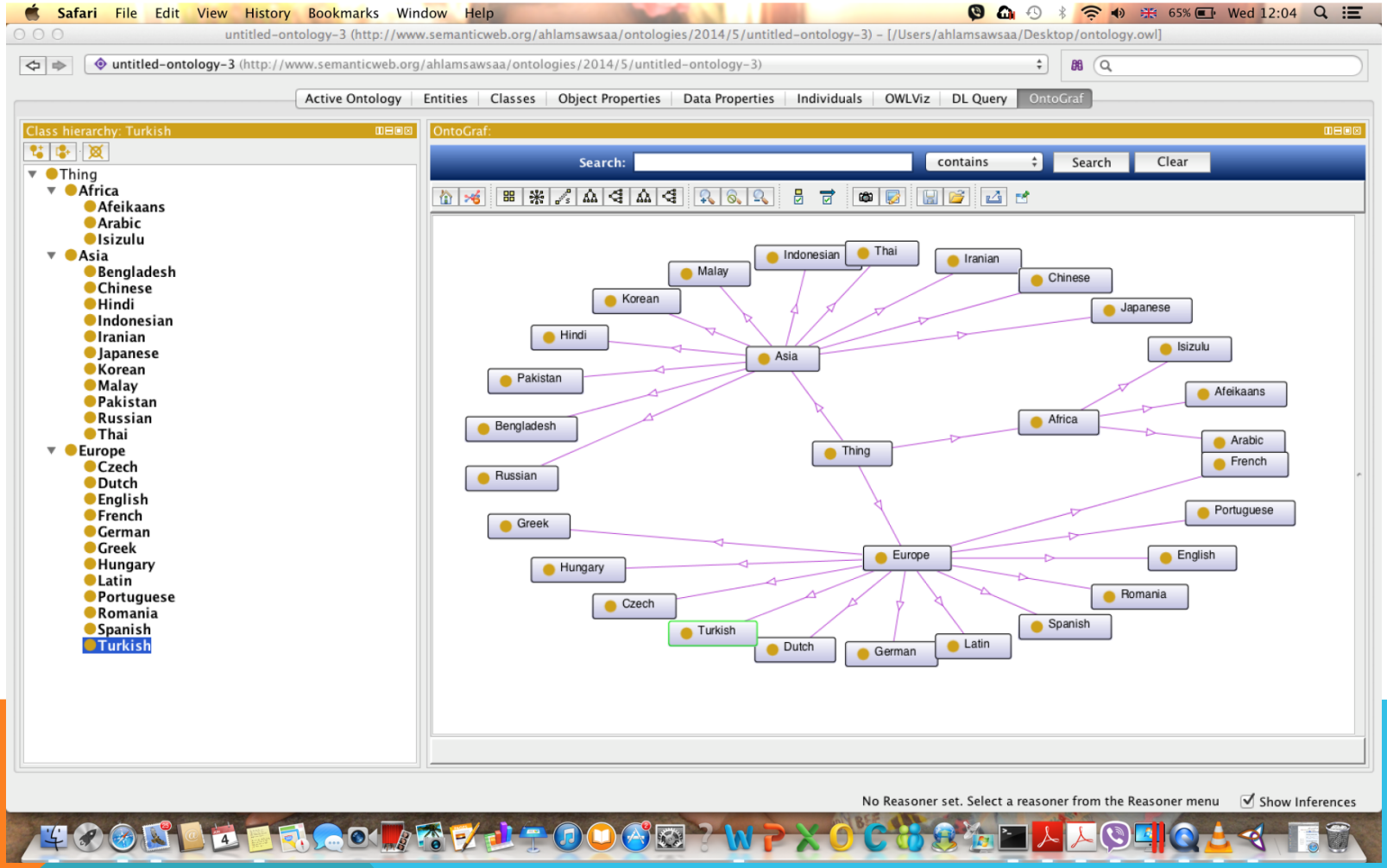
The big data tools need to consider the applications for multidisciplinary users.



By Ahlam Sawsaa

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The big data tools need to consider cross board users.



By Ahlam Sawsaa

PROFESSOR JOAN LU, INFOCOMP 2014 / DATASYS 2014, 20 JULY 2014, PARIS, FRANCE

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

The big data needs new rules to discover the true knowledge

- Models for the big data
- Algorisms for the big data
- Tools for the big data
- Platforms for the big data

Decision making based on the big data

CHALLENGES FOR THE BIG DATA IN MOBILE AGE

CHANGES ARE NEEDED

Why?

- Cost effective, especially for EU users.

Who will be the driving force?

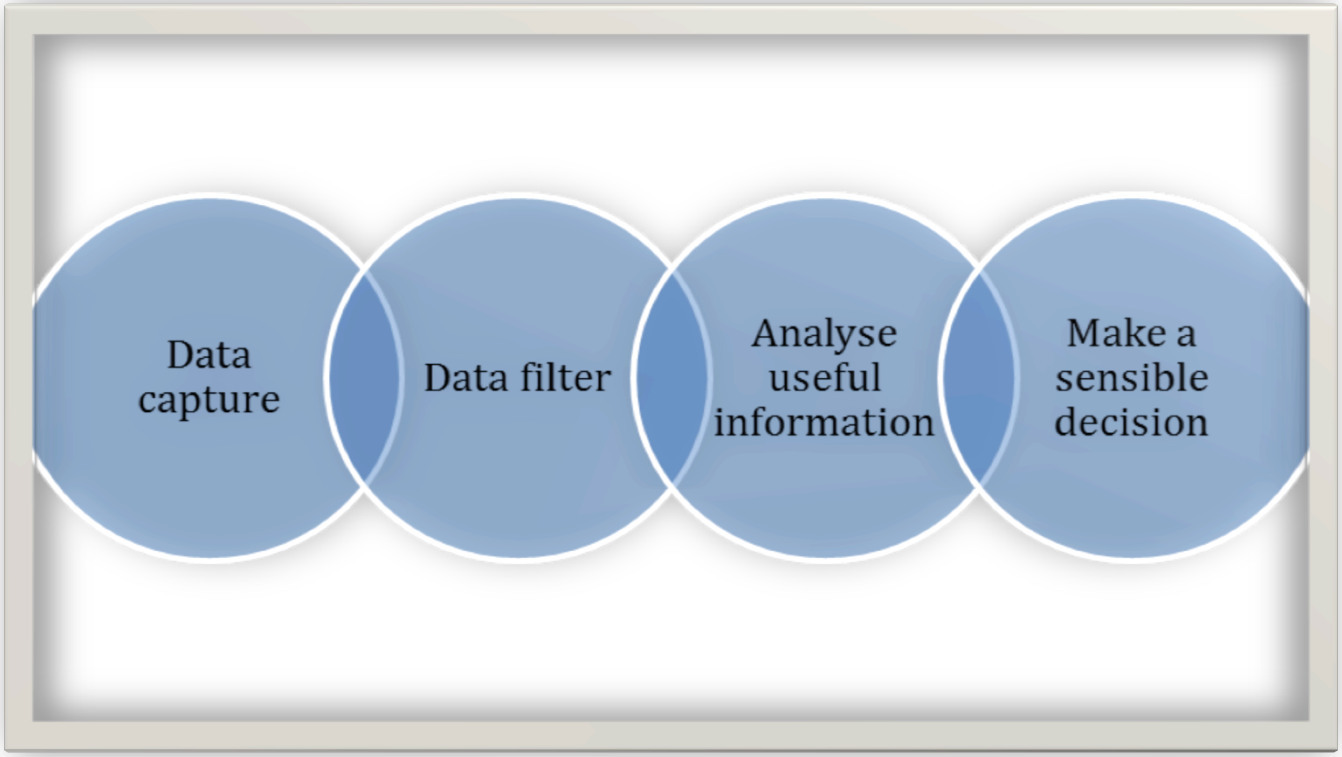
- Data users from multidisciplinary sectors
- Data users from cross countries
- Data workers in SMEs

What?

- Methodology vs. technology
  - Traditional mining and statistic models are significantly challenged by the requirements of big data solutions.

WHAT IS THE NEXT

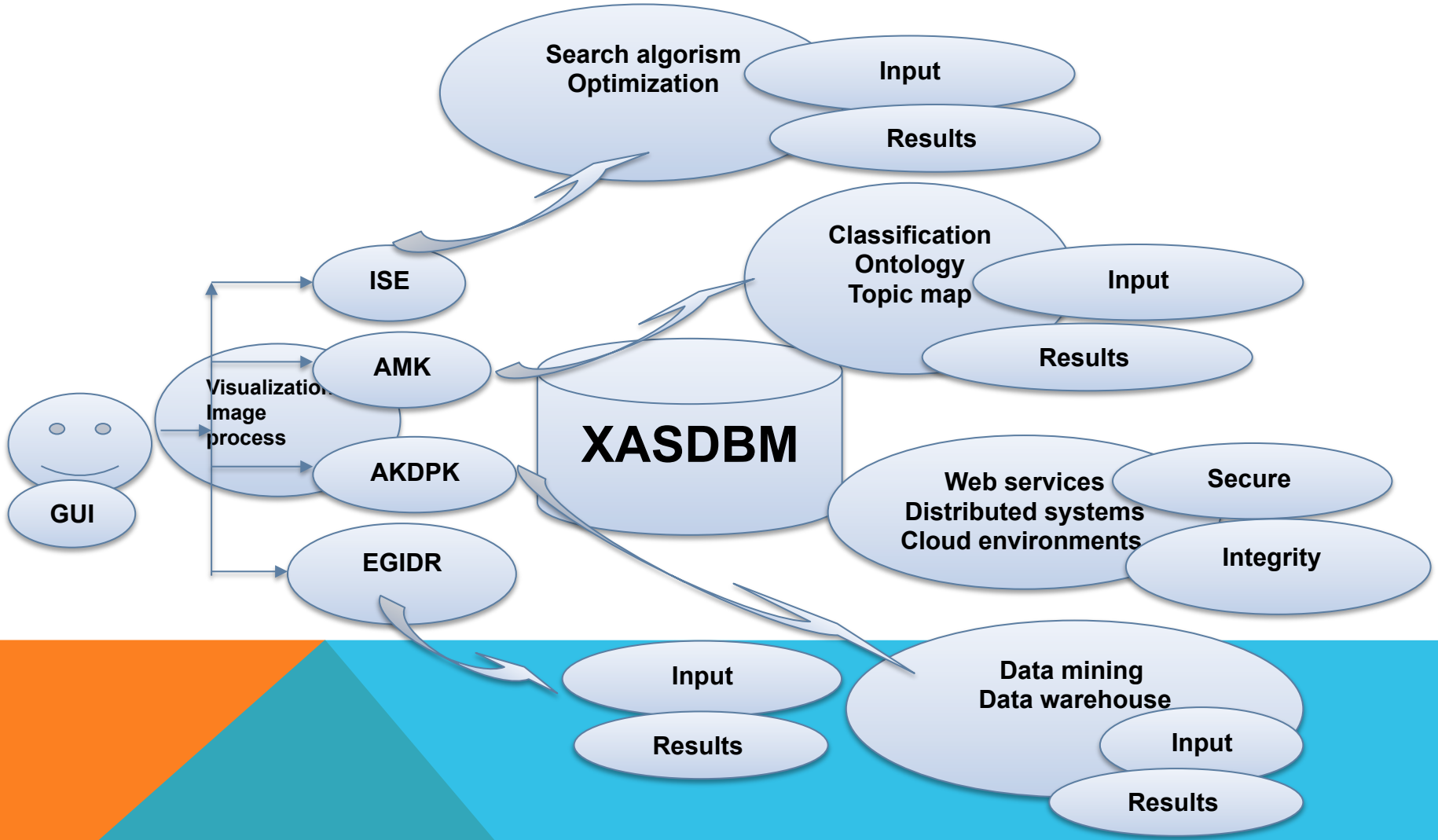
An action is in a nut shell



IARIA

WHAT IS THE NEXT

INTELLIGENT BIG DATA MANIPULATION SYSTEM

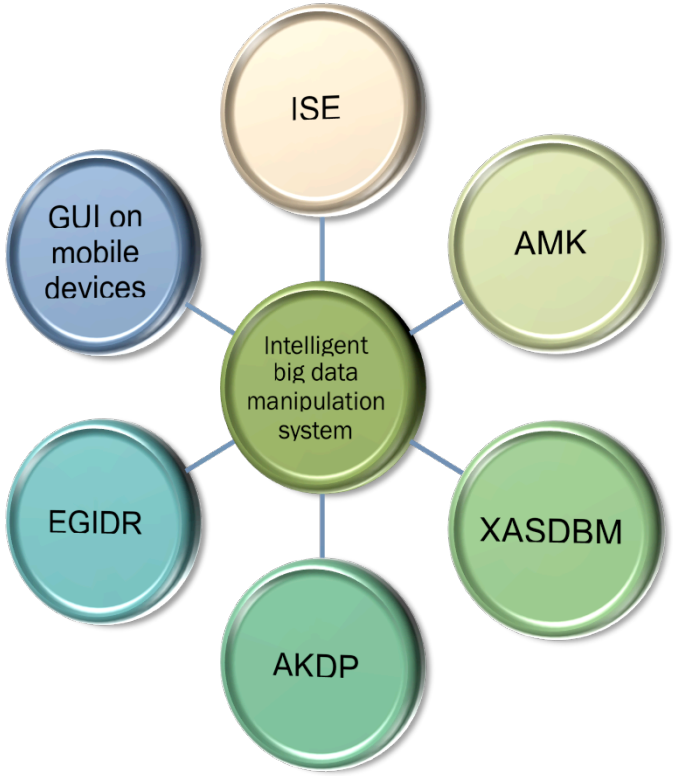


WHAT IS THE NEXT

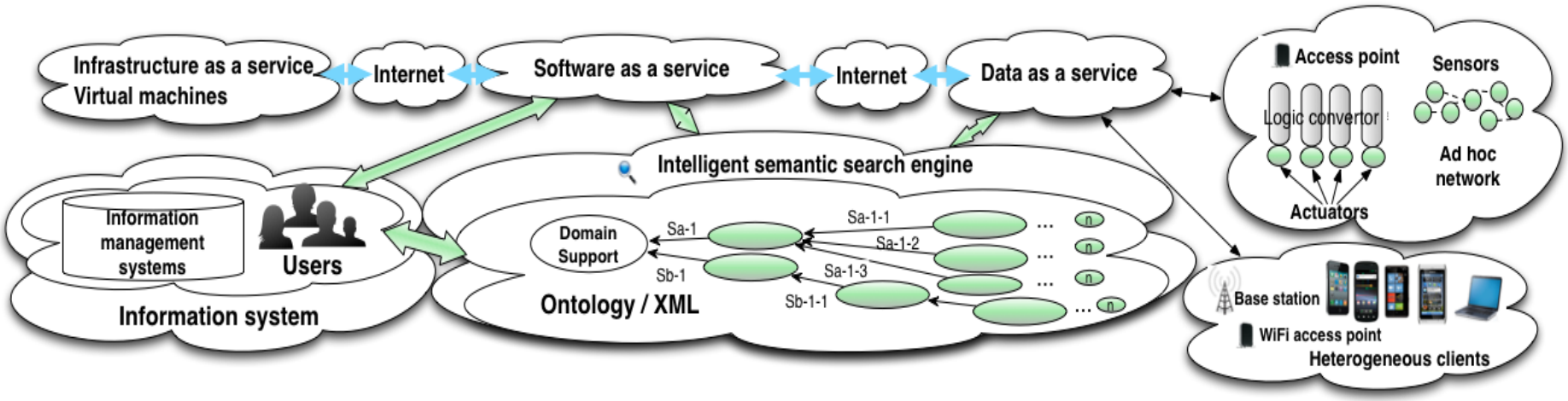
INTELLIGENT BIG DATA MANIPULATION SYSTEM

Systems designed to achieve the following facilities in the prototype:

- ISE – Intelligent search engine for data/ documents;
- AMK – Automatic mapping Knowledge base for the information repository;
- XASDBM – XML based advanced secure database management system;
- AKDP – Agent related knowledge discovery processor;
- EGIDR – Efficient graphic/image display/retrieval;
- GUI on mobile devices – User friendly graphic interface for multidisciplinary and multilingual users.



WHAT IS THE NEXT

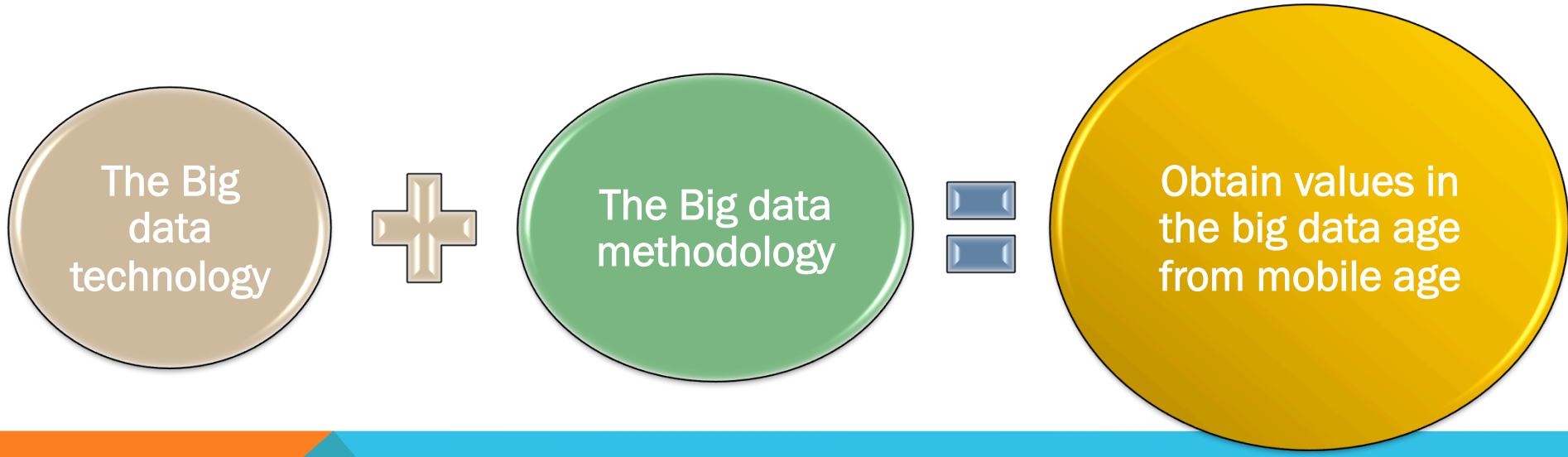


Storage, analysis, retrieval, resource sharing , information mining, finally, finding the true knowledge from the big data.

IARIA

CONCLUSION

- At the mobile age, the big data is still at large.
- A large number of trained data professional workers are needed [27].



REFERENCES

1. Intel. *Extract, transform, and load big data with Apache Hadoop*. 2014; Available from: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf.
2. Correia, M., et al. *On the Feasibility of Byzantine Fault-Tolerant MapReduce in Clouds-of-Clouds* in *Reliable Distributed Systems (SRDS), 2012 IEEE 31st Symposium on* 2012. Irvine, CA
3. Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters*, in *Communications of the ACM - 50th anniversary issue: 1958 - 20082008*: New York, NY, USA p. 107-113.
4. Costa, P., et al., *On the Performance of Byzantine Fault-Tolerant MapReduce*. Dependable and Secure Computing, IEEE Transactions on, 2013. 10(5): p. 301 - 313
5. Li, F., et al., *Distributed data management using MapReduce*. ACM Computing Surveys (CSUR), 2014. 46(3): p. 1.
6. exchange, T.k. *What do customer centricity and big data have in common?* 2014; Available from: http://www.sas.com/knowledge-exchange/business-analytics/featured/what-do-customer-centricity-and-big-data-have-in-common.
7. Xiao, Z. and Y. Xiao, *Achieving Accountable MapReduce in cloud computing*. Future Generation Computer Systems, 2014. 30: p. 1-13.
9. LINUXLINKS.com. *Native XML databases for big data*. 2014; Available from: http://www.linuxlinks.com/article/20130419184303676/NativeXMLDatabases.html.
10. eXistdb. *eXistdb*. 2014; Available from: http://exist-db.org/exist/apps/homepage/index.html.
11. BaseX. *BaseX*. 2014; Available from: http://basex.org/.
12. Sedna. *Sedna*. 2014; Available from: http://www.sedna.org/.
13. Oracle. *Berkeley DB XML*. 2014; Available from: http://www.oracle.com/technetwork/database/database-technologies/berkeleydb/overview/index-083851.html.
14. AXYANA. *Qizx*. 2014; Available from: http://www.axyana.com/qizx/.
15. LINUXLINKS.com. *Data analysis tools for big data*. 2014; Available from: http://www.linuxlinks.com/article/20130406024357813/DataAnalysisTools.html.
16. Hadoop, A. *HDFS*. 2014; Available from: http://hadoop.apache.org/.
17. systems, H. *HPCC systems*. 2014; Available from: http://hpccsystems.com/.
18. Apache. *Storm*. 2014; Available from: http://hortonworks.com/hadoop/storm/.
19. Drill, A. *Apache Drill*. 2014; Available from: http://incubator.apache.org/drill/.
20. Rapoidminer. *RapidMiner*. 2014; Available from: http://rapidminer.com/.
21. Pentaho. *Pentaho*. 2014; Available from: http://events.pentaho.com/trial-Pentaho.html?leadsource=GoogleAds&utm_campaign=EMEA-Pentaho-Brand-Name&ad_group=Pentaho&keyword=pentaho&campaign_id=70150000000fU21&gclid=CJOZquHOnr4CFbDJtAodVQsAGQ.
22. LINUXLINKS.com. *File systems for big data*. 2014; Available from: http://www.linuxlinks.com/article/20130411155608341/FileSystems.html.
23. Quantcast. *Quantcast file system*. 2014; Available from: https://www.quantcast.com/engineering/qfs/.
24. Ceph. *Ceph*. 2014; Available from: http://ceph.com/.
25. OpenSFS. *Lustre*. 2014; Available from: http://lustre.opensfs.org/.
26. Gluster. *GlusterFS*. 2014; Available from: http://www.gluster.org/.
27. EU Horizon2020 work programme, h2020-wp1415-leit-ict_en[1].pdf, published 10 December 2013.
28. IBM DB2, http://searchbusinessanalytics.techtarget.com/
29. IBM system z project, http://searchbusinessanalytics.techtarget.com/
30. NoSQL, http://2014.nosql-matters.org/cgn/
31. Zikopoulos Chris Eaton, Paul, Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGaw Hill Professional, 2011
32. Sawsaa, A. and Lu, J. (2014) ' Building an Advance Domain Ontology Model of Information Science (OIS) *International Journal of Digital Information and Wireless Communications (IJDIWC)* . 4 (2). pp. 90-98. ISSN 2225-658X