INSTITUTE OF
COMPUTER SCIENCE
Masaryk University

(Big) Data Quality

Mouzhi Ge, Faculty of Informatics, Masaryk University

Agenda

- Research motivation
- Overview of data quality research
- Data governance and DQM
- Data quality research community
- Future research agenda

Data Quality is pervasive

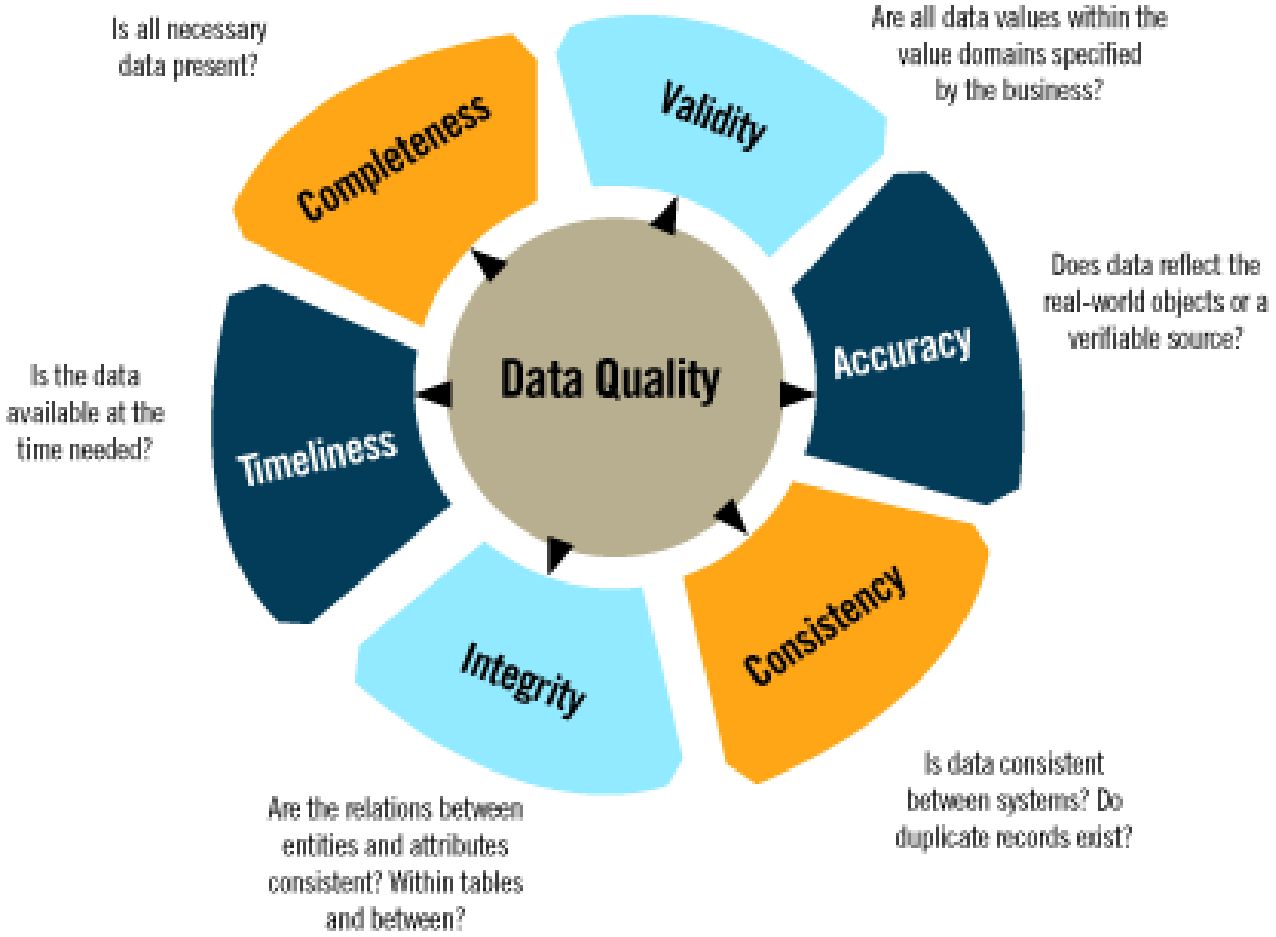
Costs of Poor Data Quality

- TDWI: Data quality problems costs US business $600 billion a year (5% of US GDP) (Eckerson 2002)
- PWC Report: 75% of those surveyed reported major problems resulting from faulty data, (Informatica. 2005:).
- “Every business function will have direct costs associated with poor data quality” (Dun and Bradstreet (2006).
- A Survey of 193 organisations sponsored, 39% of which had revenues in excess of US $1 billion: 33% rated their data quality as poor at best, whilst only 4% reported it as excellent (Information Difference 2009).
- 33% of Fortune 100 organisations will experience a data crisis arising from their inability to value, govern or trust their enterprise data (Gartner 2014)
- On average global companies feel that 26% of their data is inaccurate (up 25% on last year) and 80% do not have a sophisticated approach to data quality. (Experian 2015)
- Nearly 60% of organisations do not measure the cost of poor data quality (Gartner 2017)

Data Quality – more than Accuracy

- Syntactic Quality: degree to which data conforms to the metadata.
- Semantic Quality: degree to which data corresponds to represented external phenomena (or trusted surrogate).
- Pragmatic Quality: degree to which data is suitable and worthwhile for a given use (as decided by the actual users).

Data Quality Dimensions

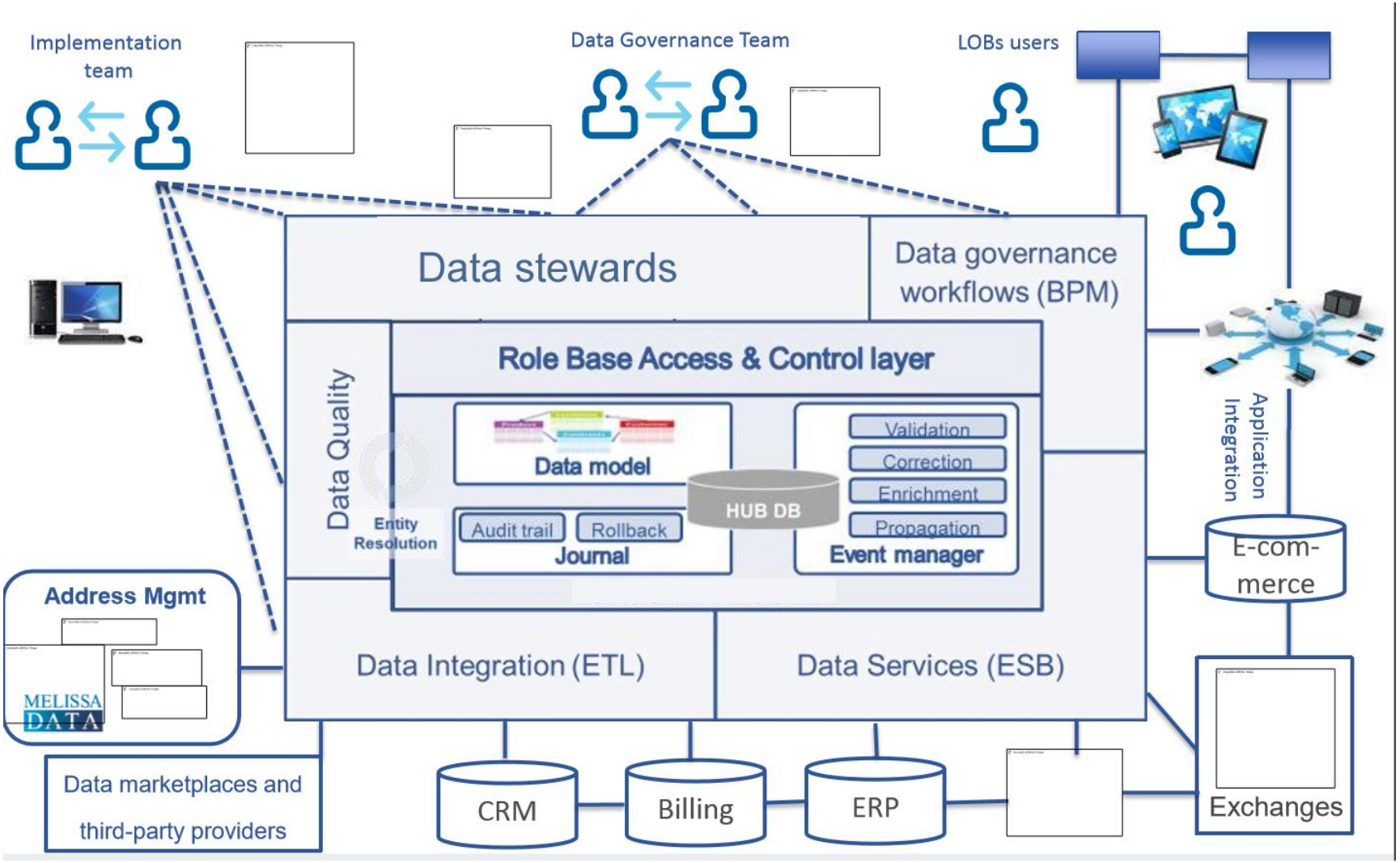


Data Quality Research Domains

Enterprise Data Governance	Data Quality Management	Master Data Management	Metadata Management	Data Protection	Data Strategy & Diagnostics
Enterprise data governance framework including processes, organization, and Infosys Data Governance technology solution. Setting data life cycle management policies.	Data quality assessment, profiling, monitoring, transformation framework, and toolkit backed up by pre-built solutions and accelerators.	Leverages Infosys deep experience on master data management implementations, and unique methodology.	Frameworks, accelerators backed up by deep competency in industry-leading tools and technologies.	Deploy data security across the enterprise with cost-effective, scalable solution, ensuring data protection and privacy, as a service.	Data Assessment Toolkit to identify optimization areas in data landscape, redundancy, simplification, and data life cycle management

Data Governance - Roles and Operations

	Data Operations	Data Quality Monitoring	Data Quality Improvement
Data Manager	Data Architecture Managment	Data Quality Planning	Data Stewardship/ Flow Management
Data Administrator	Data Design	Data Quality Criteria Setup	Data Error Causes Analysis
Data Technician	Data Processing	Data Quality Measurement	Data Error Correction



adapted from Talend Inc. Data Management

Data Quality Research Community

- *Database Community*: VLDB, SIGMOD, ICDE, you always can find the data quality track or topics of interests in top DB conferences, this community focuses more on data cleasing, record linkage etc. hard-core DQ.

- *Information System Community*: ICIS, ECIS, AMCIS, ICEIS etc, you also can find data quality track in top IS conferences, it is sometimes called information quality, they focus on user and social perspectives of data quality management.

Central Conference and Networking

- MIT International Conference on Information Quality (ICIQ)
- MIT http://mitiq.mit.edu/
- This is a small community, we have a SIG IQ
- But the topic has broad publication venues

Selected data quality publications

- Mouzhi Ge, Theodoros Chondrogiannis. Assessing the Quality of Spatio-textual Datasets in the Absence of Ground Truth. In Proceedings of the 21st European Conference on Advances in Databases and Information Systems, 2017. (**ADBIS**)
- Qishan Yang, Mouzhi Ge, Markus Helfert, Data Quality Guidelines for Data Integration in a Data Warehouse - Designed using the TPC-DI Benchmark, 19th International Conference on Enterprise Information Systems, Porto, Portugal 2017 (**ICEIS**)
- Mouzhi Ge, Tony O'Brien, Markus Helfert: Predicting Data Quality Success - The Bullwhip Effect in Data Quality. Perspectives in Business Informatics Research - 16th International Conference, Copenhagen, Denmark, August 28-30, 2017 (**BIR**)
- Mouzhi Ge, Markus Helfert, Big Data Quality - Towards an Explanation Model, 21st International Conference on Information Quality, Ciudad Real, Spain, 2016. (**ICIQ**)
- Mouzhi Ge, Markus Helfert, Impact of Information Quality on Supply Chain Decisions, Journal of Computer Information Systems, Vol. 53, No. 4, 2013. (**JCIS**)
- Mouzhi Ge, Markus Helfert, Dietmar Jannach, Information Quality Assessment: Validating Measurement Dimensions and Process, 19th European Conference on Information Systems, Helsinki, Finland, 2011. (**ECIS**)
- Mouzhi Ge and Markus Helfert, Challenges of Teaching Information Quality: Demonstrating an Adaptation of a Popular Management Game in Teaching Information Quality, 16th Americas Conference on Information Systems, Lima, Peru, 2010. (**AMCIS**)
- Mouzhi Ge and Markus Helfert, Effects of Information Quality on Inventory Management. International Journal of Information Quality, Vol. 2, No. 2, pp 176-191, 2008. (**IJIQ**)
- Mouzhi Ge and Markus Helfert, Develop a Research Agenda: A Review of Information Quality Research, 12th International Conference on Information Quality, MIT USA. November 9-11, 2007. (**ICIQ**)

Where are we heading?

Chaotic
- No standrads
- Reactive approach
- No master data plan
- No strategy

Reactive
- Standards established
- Basic DQM process established
- Master data plan identified
- Strategy defined and communicated

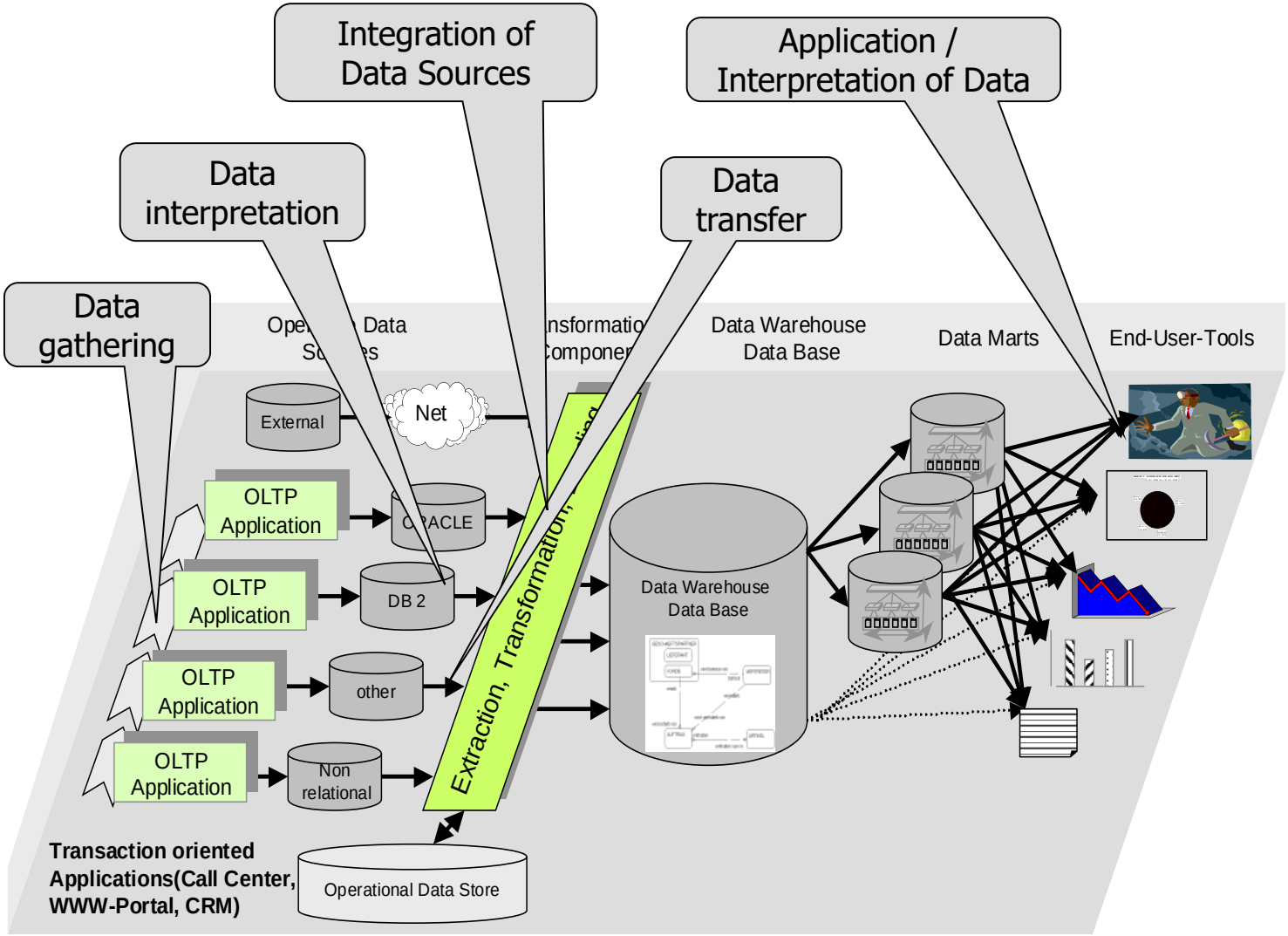
Defined
- KPIS identified & measured
- Data dictionary and rule dictionary documented and maintained
- N-Tiered stewardship established
- Master data plan executed
- Supporting technology framework deployed
- Root cause for issues being tracked and measured

Proactive
- Continuous improvement feedback loops operating
- Root cause analysis feeding into feedback process
- Proactive approach to management of data and rules dictionary
- DQM process automating measurement of function performance
- All information silos fully integrated with master data systems

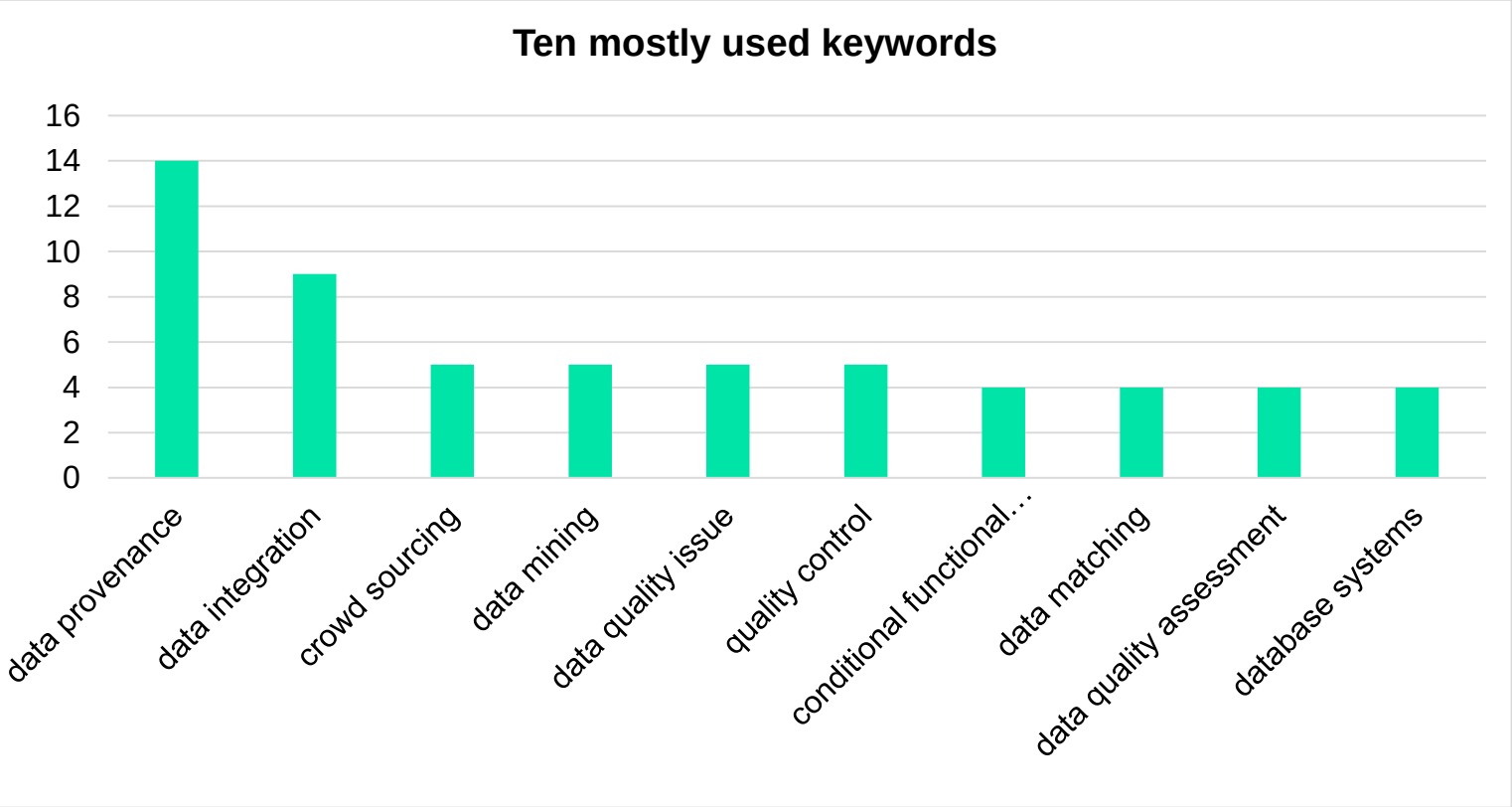
Predictive
- Process feedback loops are tuning as opposed to fixing
- DQM processes fully automated with complete audit trail
- Top-down strategy fully in tune with the bottom up application of stewardship=> complete cultural alignment across the enterprise
- People, Process, and Technology operating in harmony

1 2 3 4 5

Source: InfoSys 2015

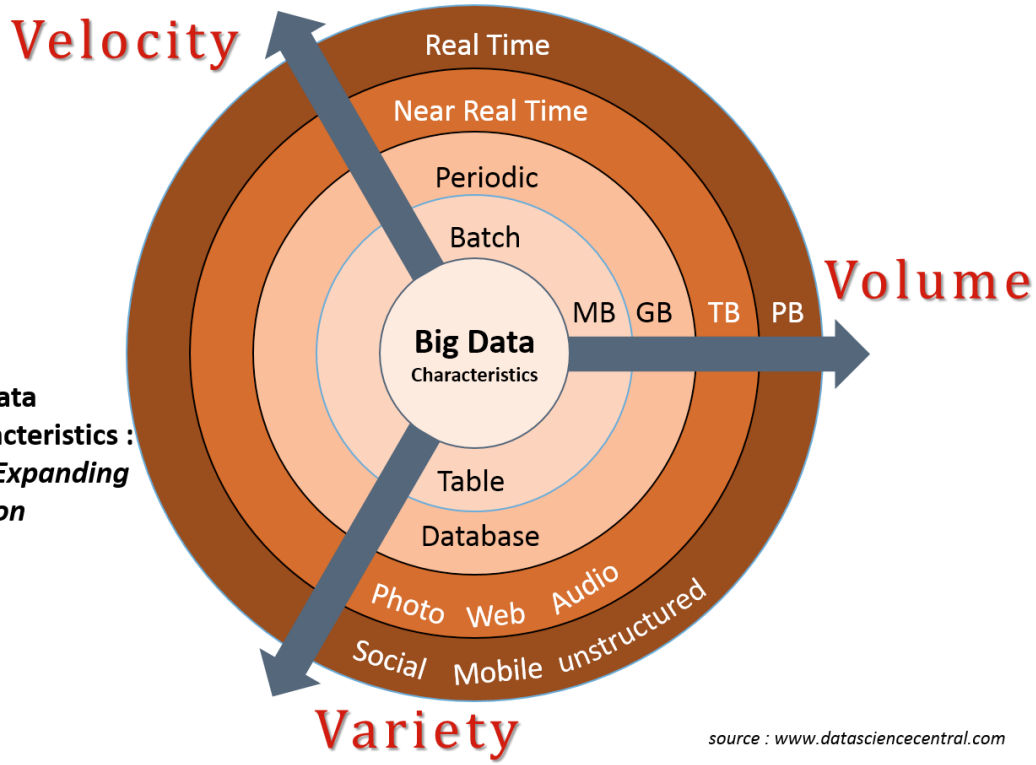

Integration of Data Sources
Application / Interpretation of Data
Data interpretation
Data transfer
Data gathering
Data Warehouse Data Base
Data Marts
End-User-Tools
External
Net
OLTP Application
OLTP Application
OLTP Application
OLTP Application
DB 2
other
Non relational
Extraction, Transformation,
Data Warehouse Data Base
Transaction oriented Applications(Call Center, WWW-Portal, CRM)
Operational Data Store

Results of a Literature Review



Ge at al. 2016

Recap the Big Data



Big Data Characteristics : ***Ever Expanding horizon***

source : www.datasciencecentral.com

“The Value of (BIG) DATA lies in the ability to derive meaningful insights from otherwise ‘noisy’ data, to make impactful decisions based on those insights and to execute decisive actions that result in increased revenue & profitability, reduction of risk & liability, and/or improved operational efficiency.“

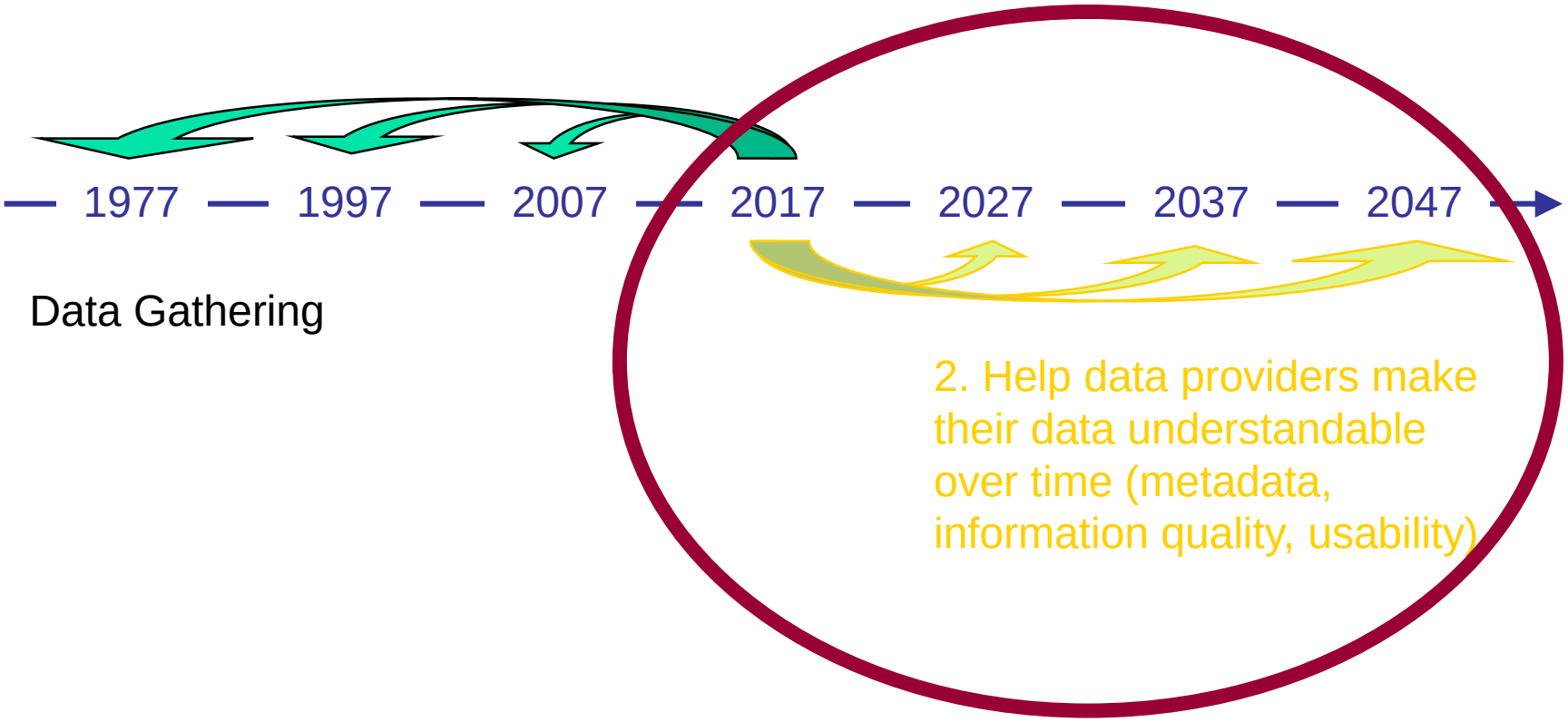
siliconANGLE

Research Agenda

- Data quality issues in data integration/ETL
- Big Data Quality
- Data quality model
- Master data management
- Data quality assessment methodology
- Data quality improvement/cleaning
- Big Data analytics and Big Data value

Information Usage

1. Help information users understand historic data



INSTITUTE OF
COMPUTER SCIENCE
Masaryk University

Thank you and question?

Mouzhi.Ge@muni.cz