12th eScience IEEE INTERNATIONAL CONFERENCE 2016

2016 IEEE 12th International Conference on eScience, Baltimore

A Comprehensive Scenario Agnostic Data LifeCycle Model for an Efficient Data Complexity Management

Amir Sinaeepourfard, **Jordi Garcia**,
Xavier Masip-Bruin, Eva Marín-Tordera

Universitat Politècnica de Catalunya – UPC Barcelona**Tech**

CRAAX Lab
UPC-BARCELONATECH
Advanced Network Architectures Lab

Data is important …

- … in **eScience**, but also in many other disciplines!
- There is **a lot of data** in the world …
- … and **data generation rate** is growing exponentially
  - The particle accelerator (LHC) at CERN (European Organization for Nuclear Research, Switzerland) generates **40 TB per second** during experimentation

- So **data management and organization** is a huge concern

Purpose of our research

Understanding how **data**
is **organized and managed** in complex systems …

… and, contribute with a **data management model**
appropriate for our scenario

Data LifeCycle (DLC) model

- Integral data management framework
  - From collection, to processing, and preservation, till removal
  - Specify policies for each phase, and define relationship among phases
- Main goals of data lifecycle models
  - Operate efficiently
  - Eliminate waste
  - Provide quality and security
  - Prepare data for efficient use
- Benefits of designing a good DLC
  - Facilitate the planning and complexity design
  - Create sustainable software, …

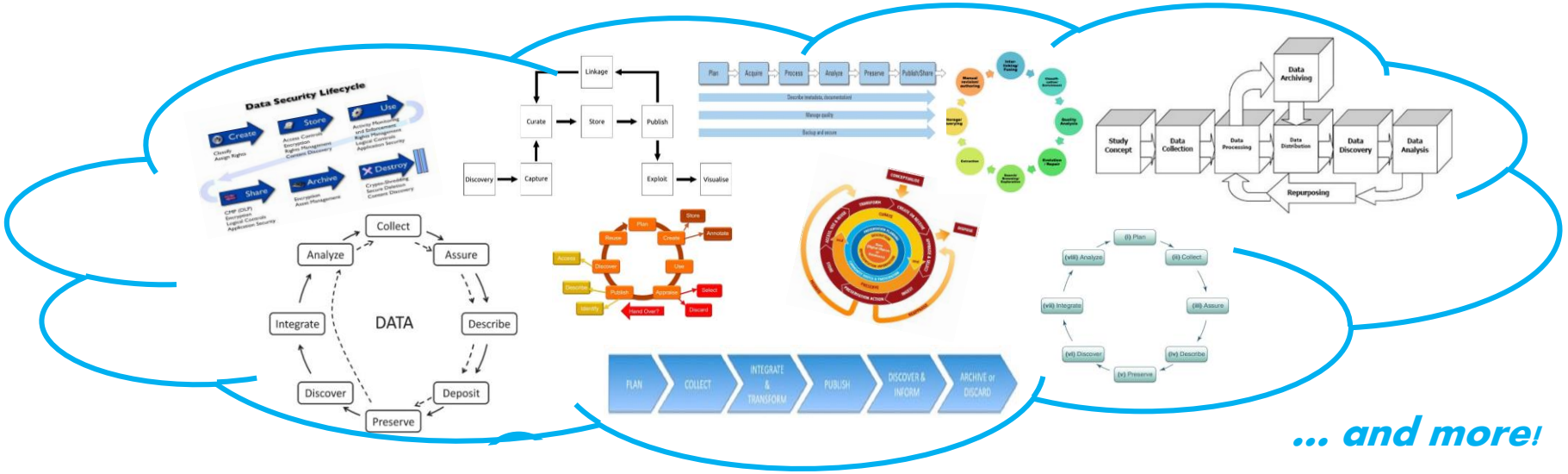
In this presentation …

- Contextualization of this research
- Preliminary work: Survey of current DLC models
  - → Limitations on current proposals
- Our proposal: The comprehensive scenario agnostic DLC model
  - Model description
  - Model assessment
- Use cases: Adaption to different scenarios
- Conclusions & following research directions

Preliminary work (1)

A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges
Technical Report (UPC-DACRR-2015-18), 2015

- Survey of all DLC models found in the literature (*and more!*)



- … and assessment with respect to the main data challenges: the **6Vs**
  - **V**alue, **V**olume, **V**ariety, **V**elocity, **V**ariability, and **V**eracity

Preliminary work (2)

A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges
Technical Report (UPC-DACRR-2015-18), 2015

- Conclusions (of survey):

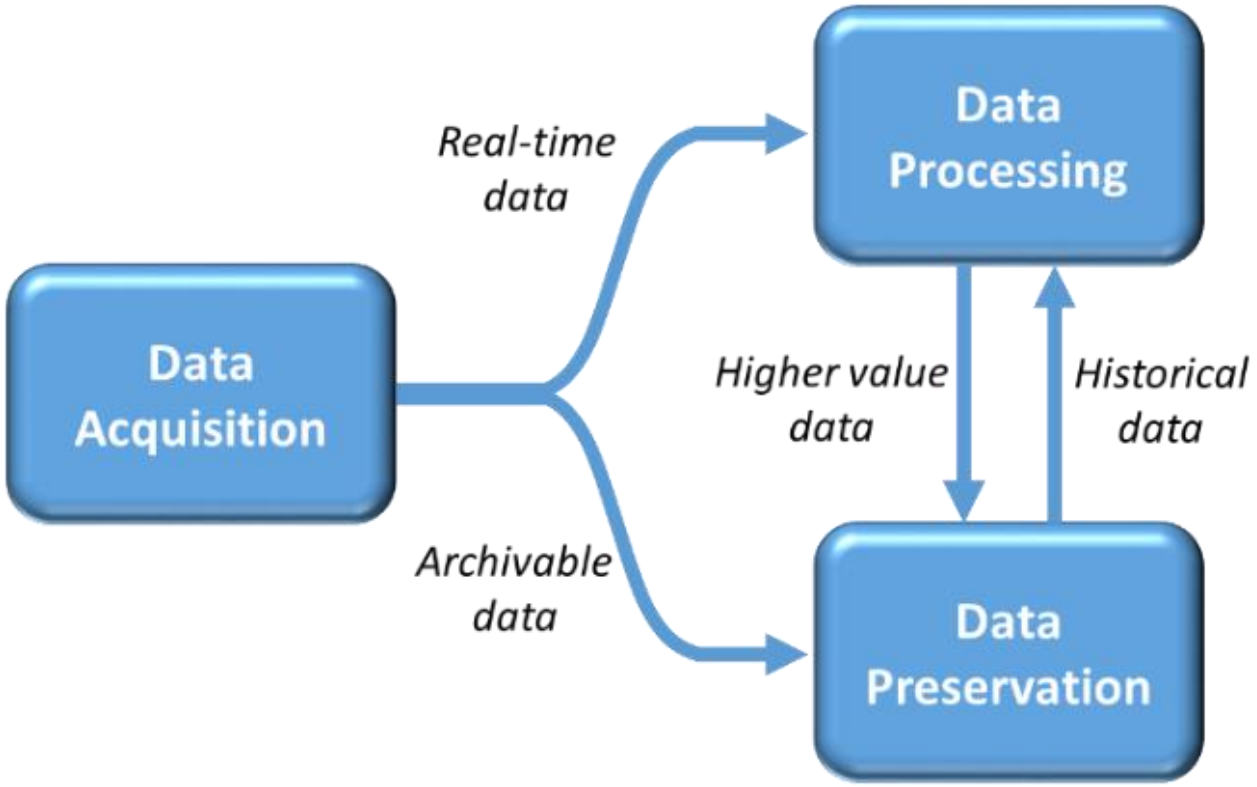
1. *Each model addresses a particular scenario, so it is not general*
2. *No model addresses successfully all 6Vs challenges*

- Proposal for this current research stage: Design a **comprehensive scenario agnostic** Data LifeCycle (COSA-DLC) model

→**Comprehensive** = Addresses all 6Vs challenges
→**Scenario agnostic** = Can be adapted to any specific scenario

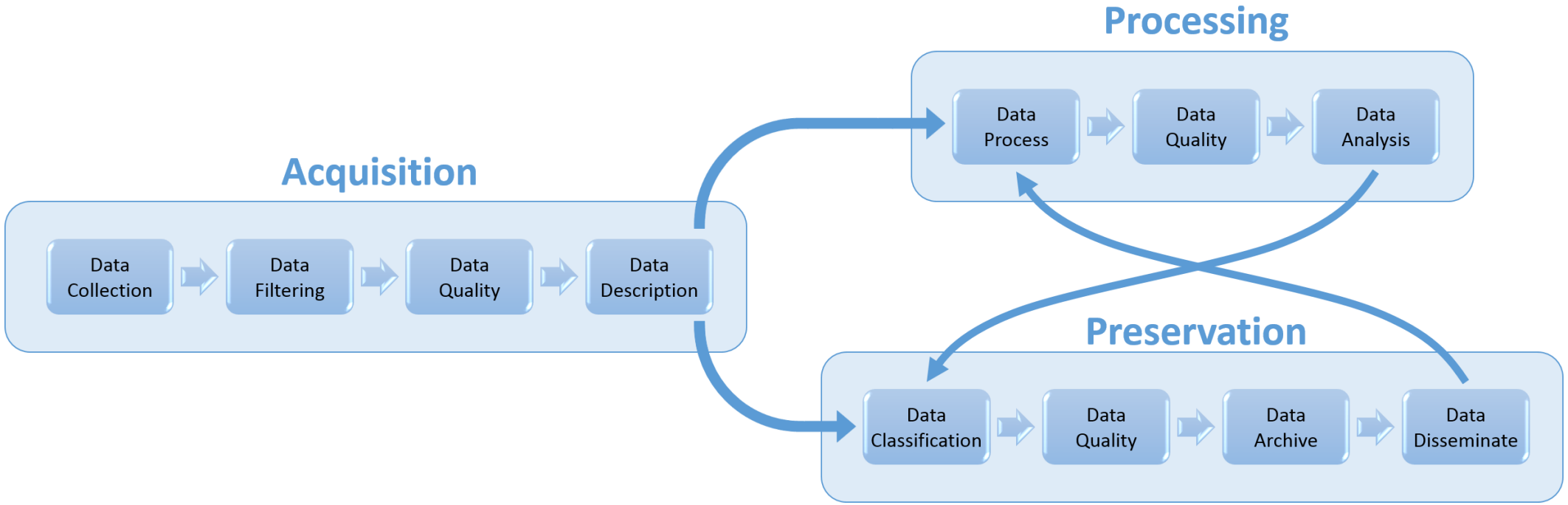
The COSA-DLC model (1)

- At block level

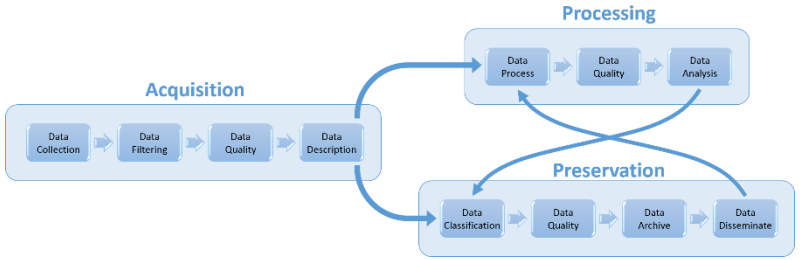


The COSA-DLC model (2)

- At phases level



The 6Vs challenges



1. Value: *of information*

Processing and analysis provide added value to data

2. Volume: *(huge) of data*

Addressed in collection and archive phases

3. Variety: *from different sources*

Mainly collection, but also filtering, description and classification phase

4. Velocity: *data rate & efficiency*

Designed for high performance: data collection & data processing

5. Variability: *meaning over time*

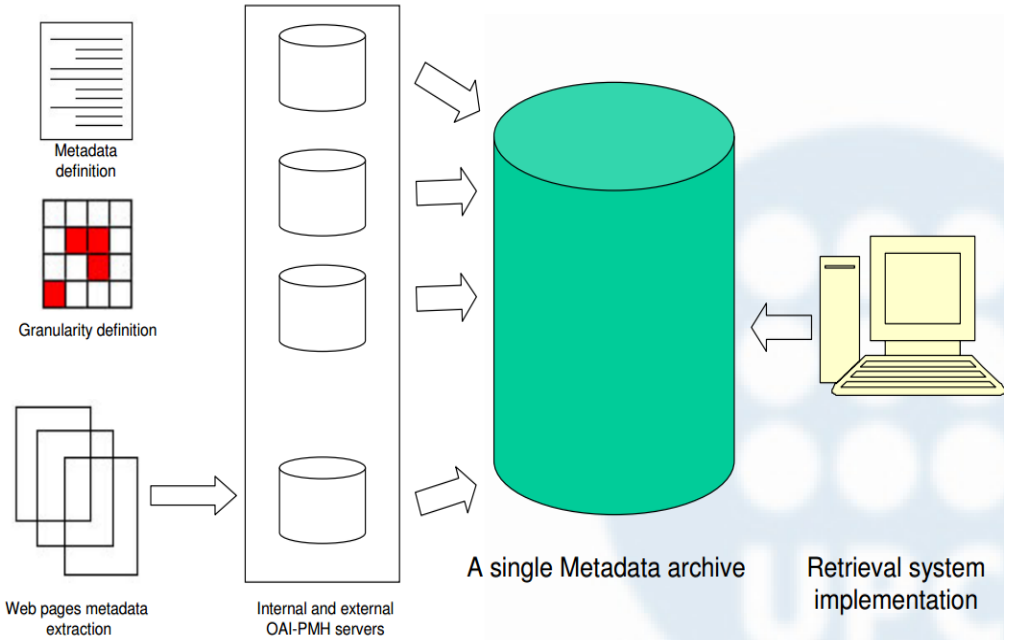
Addressed in description and classification phases. Data analysis?

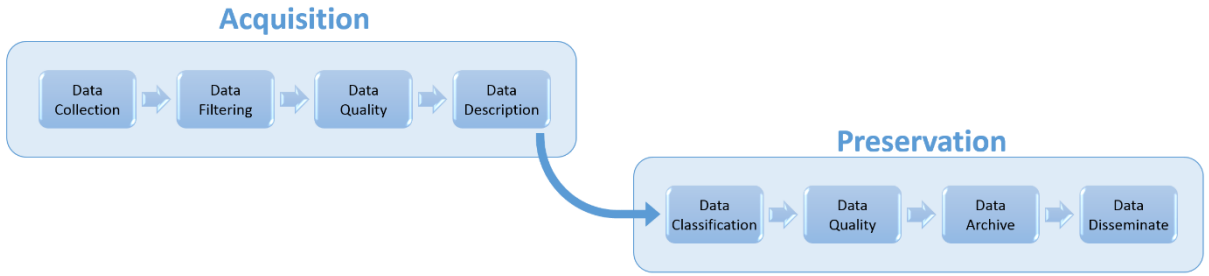
6. Veracity: *quality and/or security*

One data quality phase in each block (depends on business model)

Ease of adaptation (1)

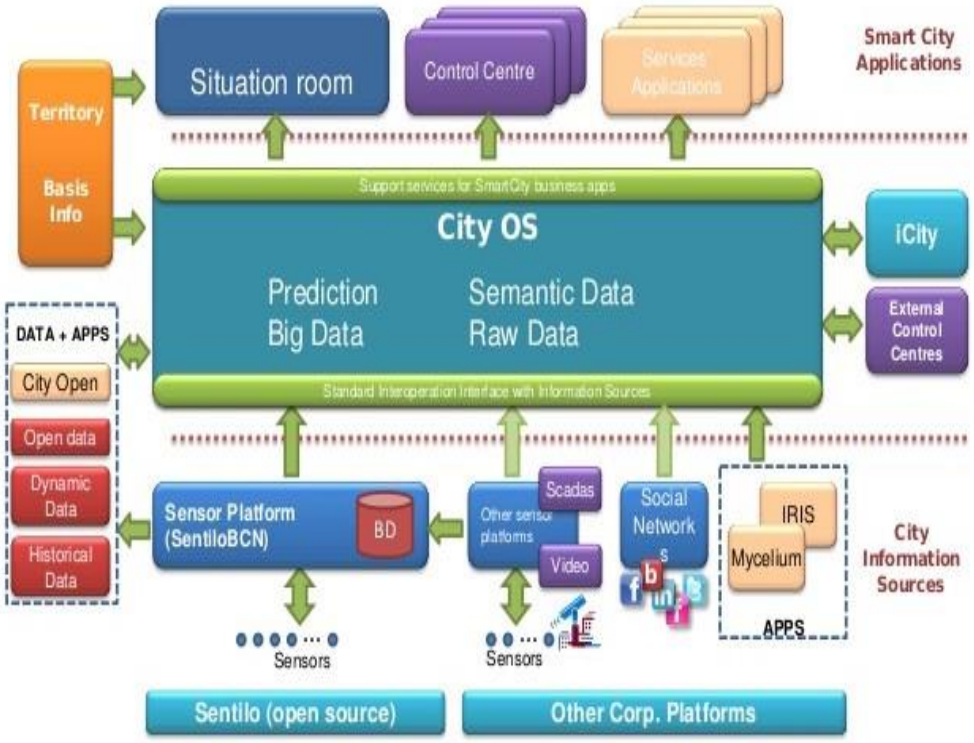
- The UPC Barcelona**Tech** library

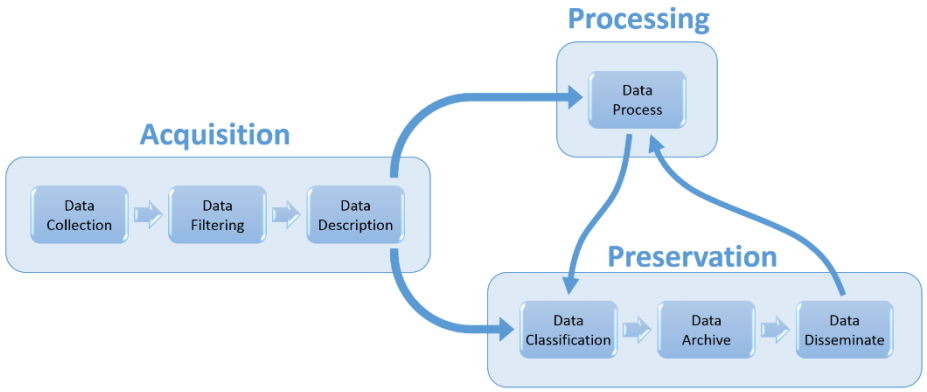




Ease of adaptation (2)

- The Barcelona Smart City architecture





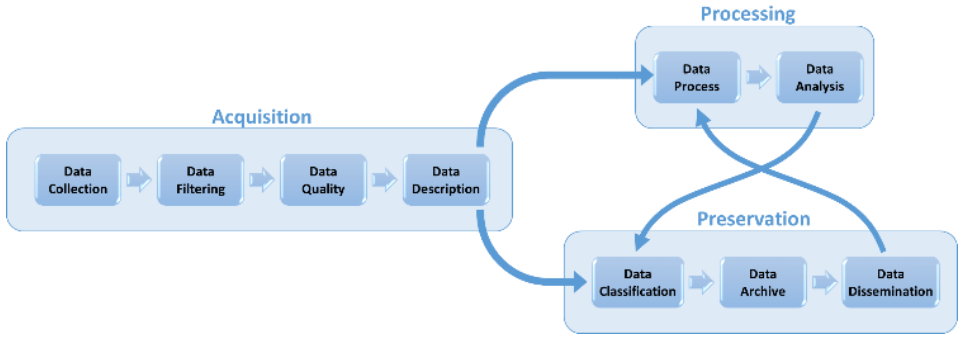
Conclusions

- COSA-DLC model
  - → **CO**mprehensive, with respect to the 6Vs challenges
  - → **S**cenario **A**gnostic, easily adaptable to any scenario

- Advantages / applicability
  - Use this model in new data management design
  - Modifications and / or extensions easily done
  - Guarantee that all 6Vs challenges are addressed …
  - … or facility to detect an eventual lack of data quality

Future (current) directions

- Adaptation of the COSA-DLC model to our specific Smart City scenario



  - Global / heterogeneous resources management strategy
  - From fog to cloud (F2C) architecture, providing the best of
    - High computing capabilities at **cloud level**
    - Low latencies at **fog level** (real-time)
      - … but also network traffic reduction
      - … plus increased security through closest data accesses

Thanks for your attention

jordig@ac.upc.edu