Research Data Administration in CITIES Project

Xiufeng Liu, *Postdoc.*

xiuli@dtu.dk

Department of Engineering Management, DTU

26 October 2015

Table of Contents

Introduction – *smart city data for research*

Data types:

- Building, environment, climate, traffic, GPS, LBS, energy, and customer info., etc.

Characteristics:

- From heterogeneous sources, and diverse
- Possibly have bad qualities
- Contain sensitive data, e.g., personal info.

The requirements of the data used for research:

- Have good quality
- Ensure the safety of the data
- Protect privacy
- Be shared/published & be cited like academic papers
- ...

CITIES

DTU

Research Data Sharing/Publishing

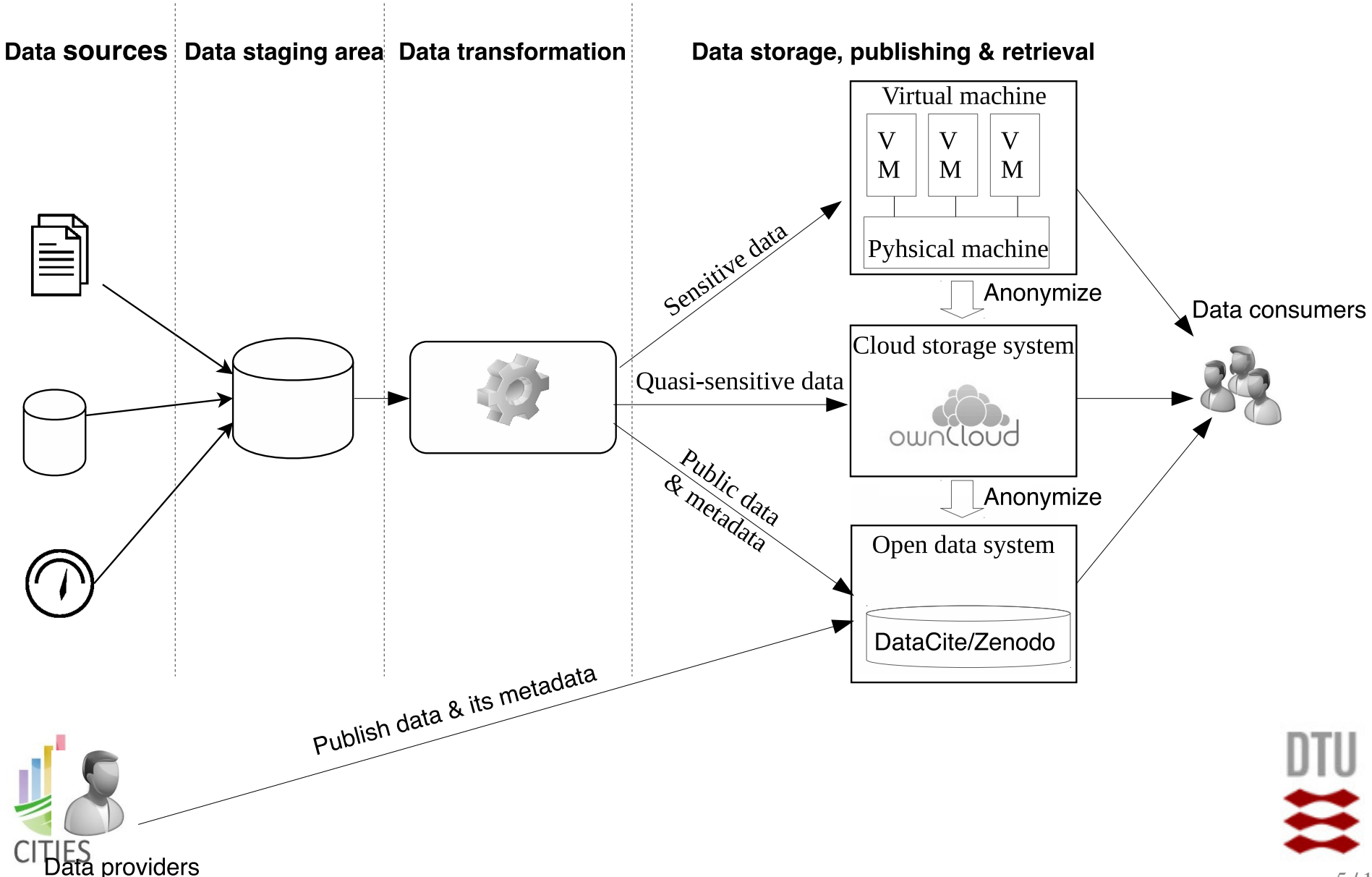
Therefore, there are two main issues:

1. Data quality issue
   - Cleanse data before publishing
   - Instruction of data quality, e.g., comments, quality score.
2. Security & privacy issue:
   - Classify data attribute values according to the risk level:
     - High sensitive data (e.g., CPR, name)
     - Sensitive data (e.g., age, gender, address, families., etc)
     - Quasi-sensitive data (e.g., salary, energy consumption)
     - Public data (or open data)
   - Use a different sharing/publishing strategy for the data of each risk level

CITIES

DTU

The Proposed Data Sharing/publishing Framework



Sensitive Data Sharing

Use virtual machine (VM) based secured environment:

- Create VM from the image with customized software installation
- Can access the data of all risk levels
- Run the applications, models, algorithms within VM

CITIES

DTU

Quasi-sensitive Data Sharing

Use the cloud-based storage system – OwnCloud:

- User authorization – university account
- Fine-granular accession permission control – Users/Groups
- Users can use the data on their own computers.

CITIES

DTU

Open Data Publishing

Use the open data platform – Zenodo:

- Publish open data sets, anonymized data, and metadata of (quasi-)sensitive data
- Use a unique digital data object identifier (DOI) to identify each published data set.
- A published data set can be cited by DOI
- Publish the metadata of local data store via OAI-PMH.

Data Pre-processing – *for data quality*

- Data quality checking
  - Use a linear regression model (different data properties as the variables) for scoring the data quality
  - Automate quality checking
  - Quality scores and comments as the metadata for publishing
- Rule-based data cleansing
- Analytics-based for complex data cleansing (mostly semantic problems)

CITIES

DTU

Data Pre-processing – *for sensitive data*

- Data anonymization
  - Anonymization methods: *suppression & generation*
  - Anonymity metrics: *k-anonymity & l-diversity*

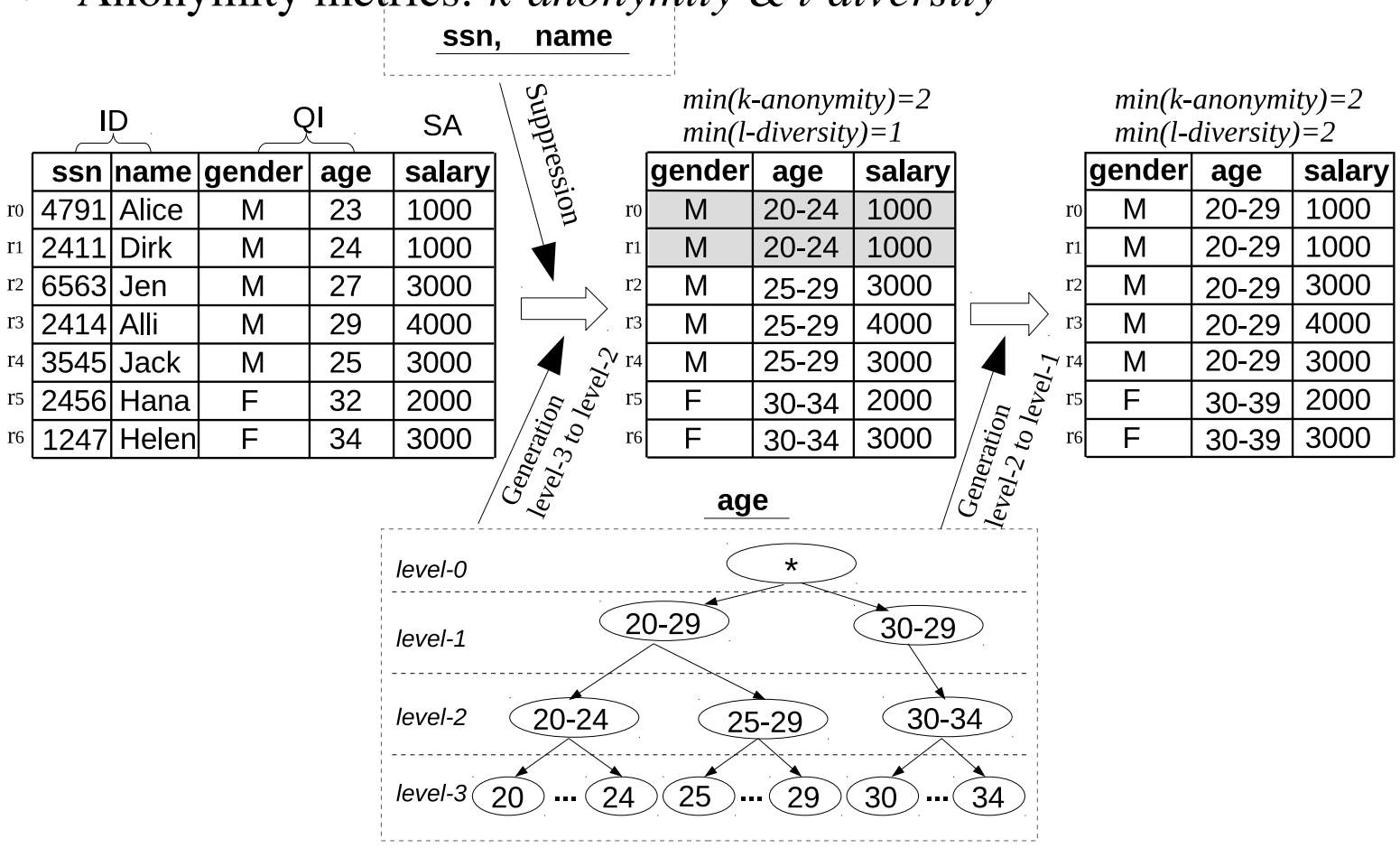


Figure : An example of data anonymization implementation with the requirements, $k\text{-}anonymity \geq 2$ & $l\text{-}diversity \geq 2$

Finished and Future Work

The work finished:

- BigETL: A scalable data processing system
- Create an ETL job to automate transferring district heating data from ftp server to OwnCloud
- Data anonymization module

Future work:

- Deploy BigETL into a planning sever
- Deploy Zenodo
- Create ETLs for different types of the data (cleansing, scoring, anonymization)
- VM for sensitive data

CITIES

DTU

CITIES Data – We have

- Sønderborg, hourly district heating usage data from 140 buildings delivered every day since 2014.11.20 (53 meters).
- National energy demand data from the ministry (in cooperation with Project Zero and Ålborg university)
- BBR register data (organized by the PhD student, *Panagiota*)

CITIES

DTU

CITIES Data – We will have

We can get:

- Sønderborg, data from Project Zero (to work with Nicolas)
- Sønderborg, electrical demand data from some other buildings - Dong Energy
- Sønderborg, gas data

We want to get:

- Århus data from Affaldvarme Århus (AVA) - building data will come in a few month
- Odense was discussed
- Albertslund data

CITIES

DTU

Questions?