Data as a Service – Models and Engineering

Hong-Linh Truong
Distributed Systems Group,
Vienna University of Technology

truong@dsg.tuwien.ac.at
http://dsg.tuwien.ac.at/staff/truong

Outline

- Data provisioning and data service units
- Data-as-a-Service concepts
- Data concerns for DaaS
- Evaluating data concerns

What is the common point here?

„Use of several health, food and recipe services, in order to collect general food information”

“Measure and report water quality metrics”

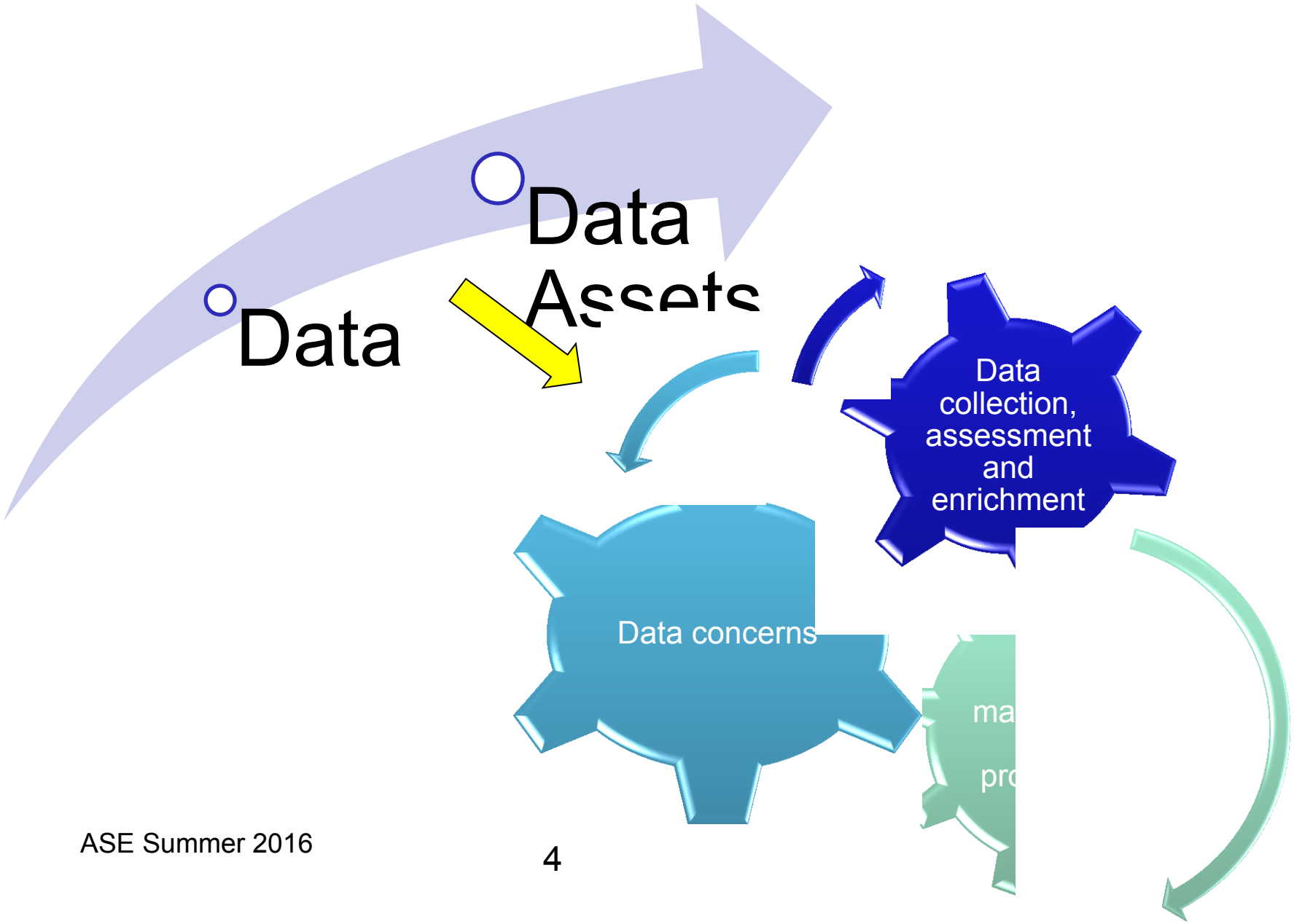
“Latest data on air quality is fetched from London Air API”

“give data about crimes in an area …. ranking of data quality ”

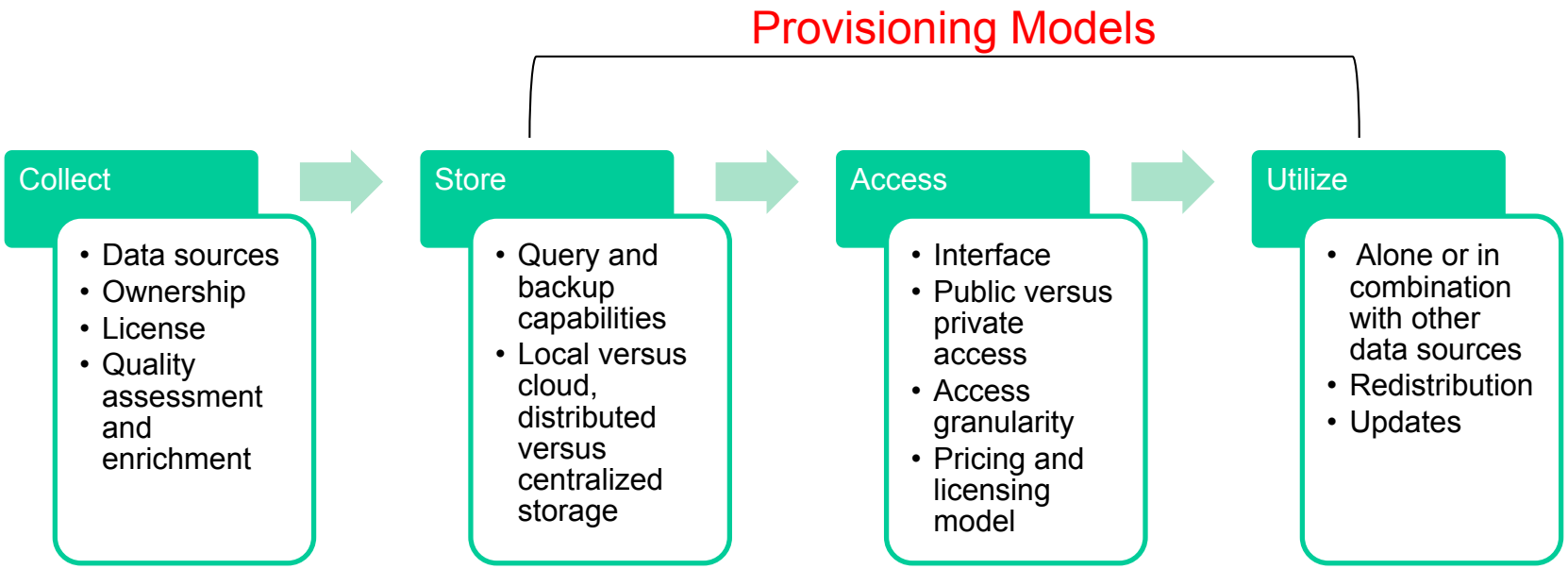
„collect location-data from multiple Sources …. combine location- with social-data“

„real time production information from photovoltaic panels”

Data versus data assets

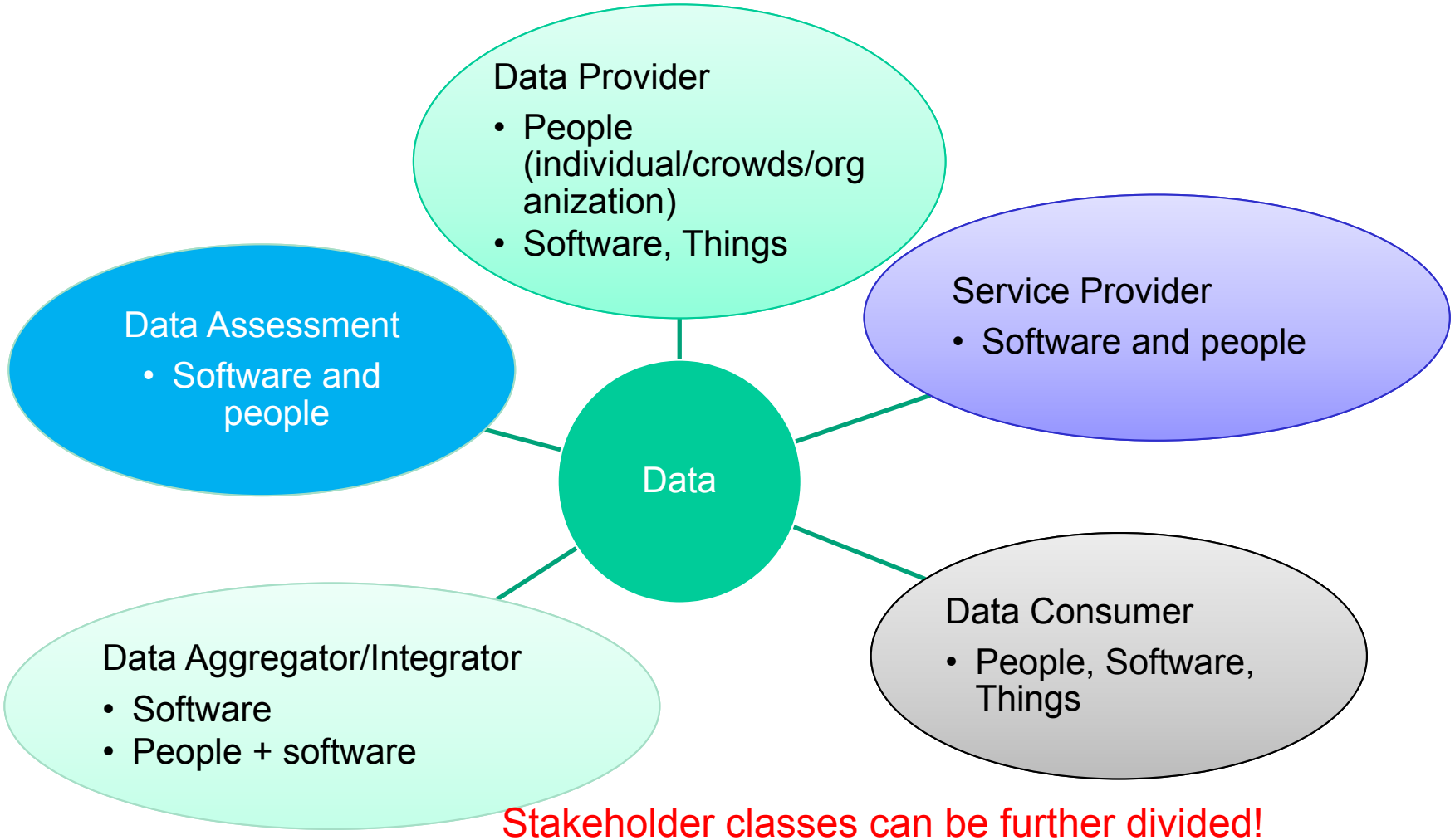


Data provisioning activities and issues



Non-exhausive list! Add your own issues!

Stakeholders in data provisioning



Stakeholder classes can be further divided!
Domain-specific versus domain-independent functions

Data service unit

Consumption, ownership, provisioning, price, etc.

Service model

Data

Unit Concept

„basic component“/“basic function“ modeling and description

Data service unit

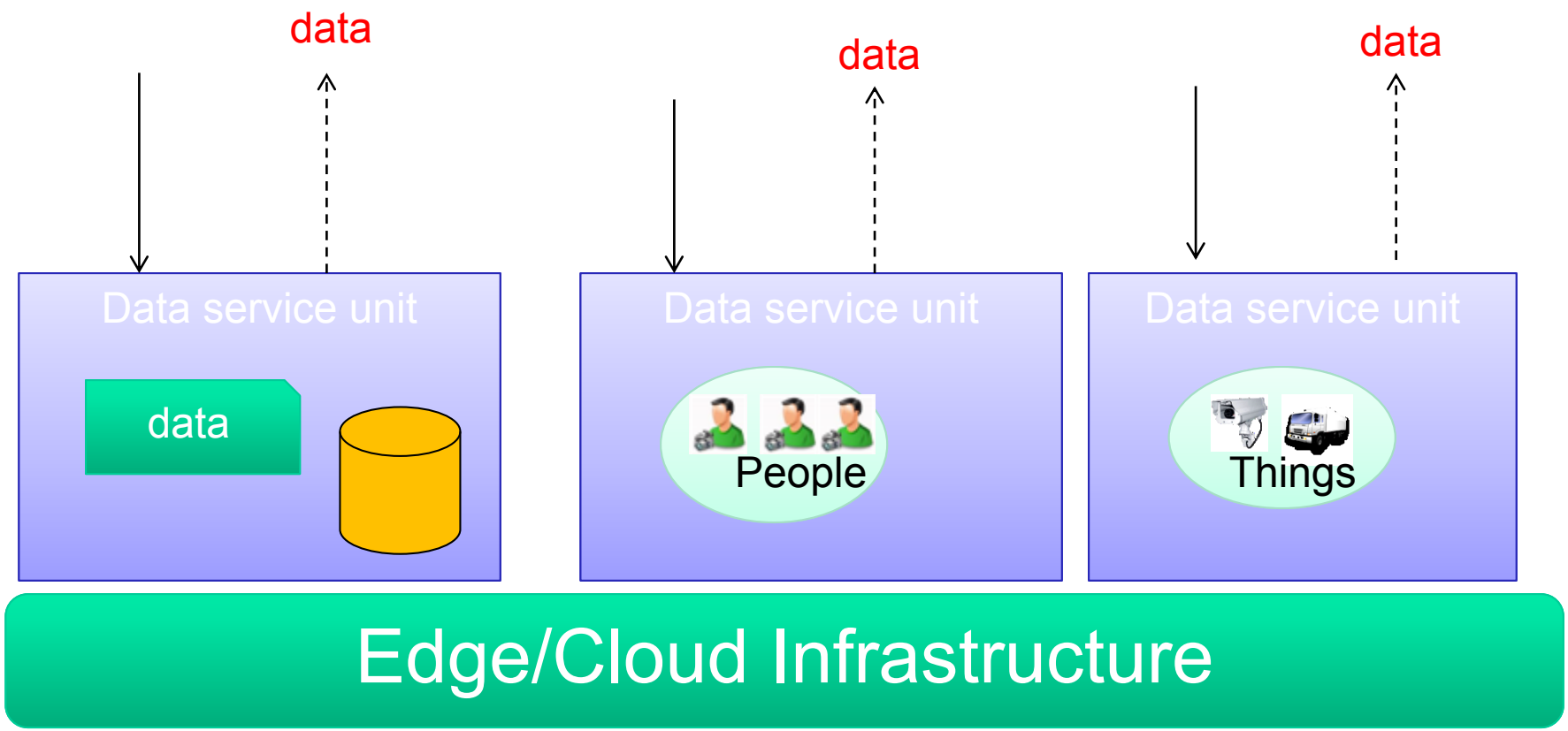
What about the granularity of the unit?

- Can be used for private or public
- Can be elastic or not

Data service units in clouds

- *Provide data capabilities* rather than provide computation or software capabilities
- Providing data in clouds/internet is an increasing trend
  - In both business and e-science environments
- Now often in a combination of data + analytics of the data → to provide data assets

Data service units in distributed edge and cloud systems



Data as a Service -- characteristics

Let us use NIST's definition

- *On-demand self-service*
  - Capabilities to provision data at different granularities
- *Resource pooling*
  - Multiple types of data, big, static or near-realtime,raw data and high-level information
- *Broad network access*
  - Can be access from anywhere
- *Rapid elasticity*
  - Easy to add/remove data sources
- *Measured service*
  - Measuring, monitoring and publishing data concerns and usage

Data as a Service – service models and deployment models

Data-as-a-Service – service models

- Data publish/subcription middleware as a service
- Database-as-a-Service (Structured/non-structured querying systems)
- Sensor-as-a-Service
- Storage-as-a-Service (Basic storage functions)

deploy

Edge and/or Cloud Systems

Examples of DaaS

Windows Azure Marketplace

Region: United States | Support | Sign In

Learn | Applications | Data | My Account | Publish | Search the Marketplace

HOME > DATA

category
- BANKING (2)
- CAPITAL MARKETS (6)
- INSURANCE (1)

publisher
- BOUNDARY SOLUTIONS, INC. (3)
- CDYNE CORPORATION (1)
- CUSTOM WEB APPS, LTD (1)
- DEUTSCHE BOERSE AG (1)
- DIGITAL TROWEL INC. (2)
- DNB (2)
- DUN & BRADSTREET (12)
- > SEE ALL PUBLISHERS...

41 Results in: DATA | PAID | BUSINESS AND FINANCE

Sort By: Date Added | Name | Publisher

1 2 3 4 5 ▶

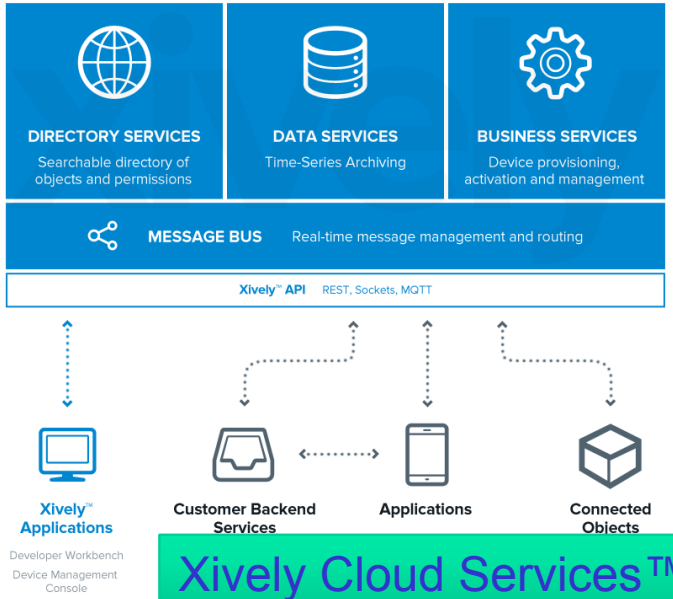
Bustling Manufacturers & Business Services List — data
published by: DNB

Bustling Manufacturers & Business Services list is a market segmentation that includes over 30,000 large manufacturers and businesses with an average annual sales volume of $40 million. The companies in this list also have high trade activity, maintained steady size in last 4 years and have been in business for an average of 20 years.

Crime Statistics for England & Wales — data
published by: Custom Web Apps, Ltd

The crime data is released by the National Policing Improvement Agency (NPIA) at the end of every month and contains all recorded crime and anti-social behaviour for England & Wales. Data is available from Dec 2010 to present to a level of full UK postcode as well as postcode sector, postcode district, and postcode area.

data



Xively Cloud Services™
https://xively.com/

GNIP — The Social Media API™ — Product

Gnip is the Largest Provider of Social Media Data to the Enterprise - Never Miss a Tweet, Post, Comment or Like

Try Gnip! CONTACT US TODAY

Twitter Feeds GET STARTED!

DATA.GOV.UK Beta
Opening up Government

Home | Data | Participate | Apps | Location | Linked Data | Library | Lab | About

Search | Map Search | Publishers | Tags | Public Roles & Salaries | Spend Browser | Spend Reports

Search Datasets

8729 Datasets

Search... Search

Tags View all tags »

national-indicators Health health Spending Data care spend-transactions communities school NERC_DDC local-government transparency nhs children health-well-being-and-care population finance child health-and-social-care education disclosure

Publishers View all publishers »

- Office for National Statistics (847)
- Department for Communities and Local Government (739)
- NHS Information Centre for Health and Social Care (514)
- British Geological Survey (364)
- Centre for Ecology & Hydrology (326)
- Department for Environment, Food and Rural Affairs (322)
- Welsh Government (241)
- Department of Health (239)
- Department for Children, Schools and Families (227)
- Home Office (221)

UK Location Conduct Map Based Search »

The UK Location Programme has introduced over 1000 location data records into data.gov.uk and tools to support their use. To find which of these datasets cover a particular location, you can use Map Based Search.

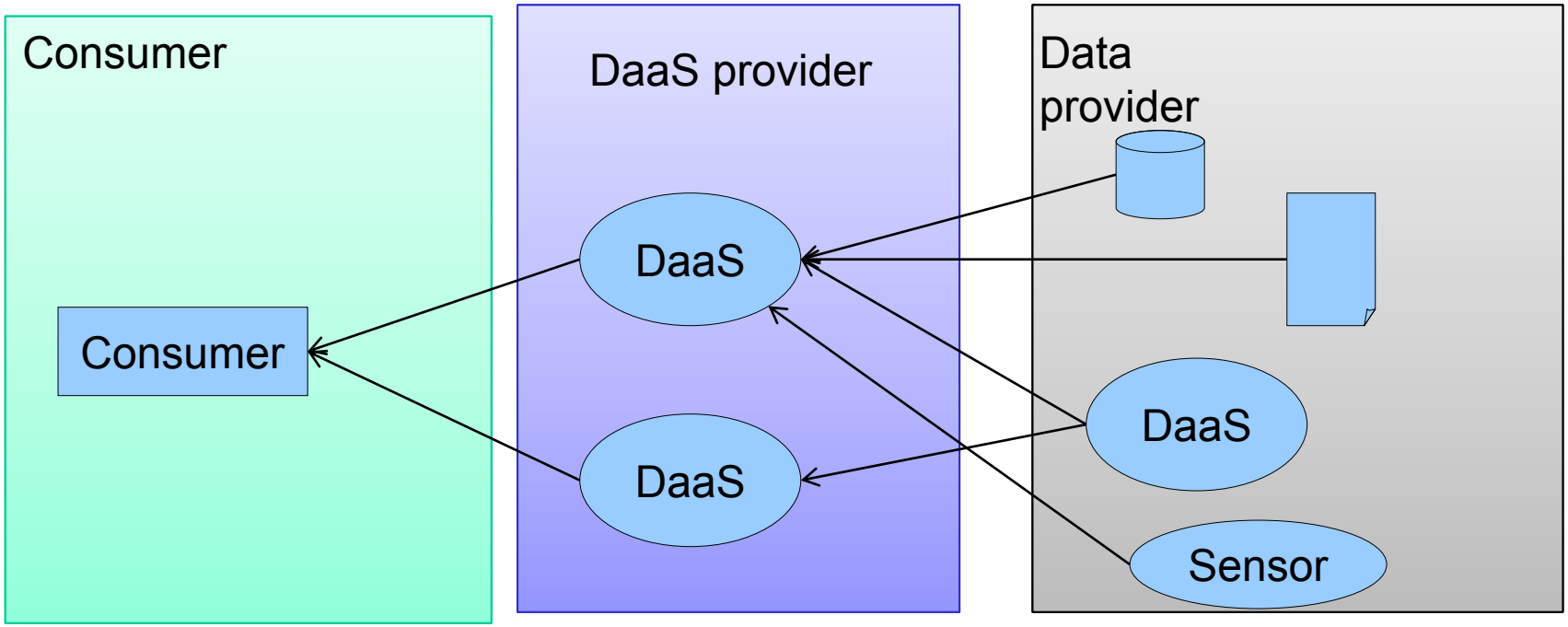
Many of these datasets provide a Web Map Service too, and for some a preview of the data is available. Click to find out more about Map Based Search and about Preview on Map.

DaaS design & implementation – APIs

- Read-only DaaS versus CRUD DaaS APIs
- Service APIs versus Data APIs
  - They are not the same wrt data/service concerns
- SOAP versus REST
- Streaming data API

DaaS design & implementation – service provider vs data provider

- The DaaS provider is separated from the data provider



Example: DaaS provider =! data provider

Windows Azure Marketplace

Region: United States ▼ | Support | Sign In

Learn | Applications | Data | My Account | Publish | Search the Marketplace

HOME > DATA

category
- BANKING (2)
- CAPITAL MARKETS (6)
- INSURANCE (1)

publisher
- BOUNDARY SOLUTIONS, INC. (3)
- CDYNE CORPORATION (1)
- CUSTOM WEB APPS, LTD (1)
- DEUTSCHE BOERSE AG (1)
- DIGITAL TROWEL INC. (2)
- DNB (2)
- DUN & BRADSTREET (12)
- > SEE ALL PUBLISHERS...

41 Results in: DATA | PAID | BUSINESS AND FINANCE

Sort By: Date Added | Name | Publisher

1 2 3 4 5 ▶

Bustling Manufacturers & Business Services List — data

published by: DNB

Bustling Manufacturers & Business Services list is a market segmentation that includes over 30,000 large manufacturers and businesses with an average annual sales volume of $40 million. The companies in this list also have high trade activity, maintained steady size in last 4 years and have been in business for an average of 20 years.

Crime Statistics for England & Wales — data

published by: Custom Web Apps, Ltd

The crime data is released by the National Policing Improvement Agency (NPIA) at the end of every month and contains all recorded crime and anti-social behaviour for England & Wales. Data is available from Dec 2010 to present to a level of full UK postcode as well as postcode sector, postcode district, and postcode area.

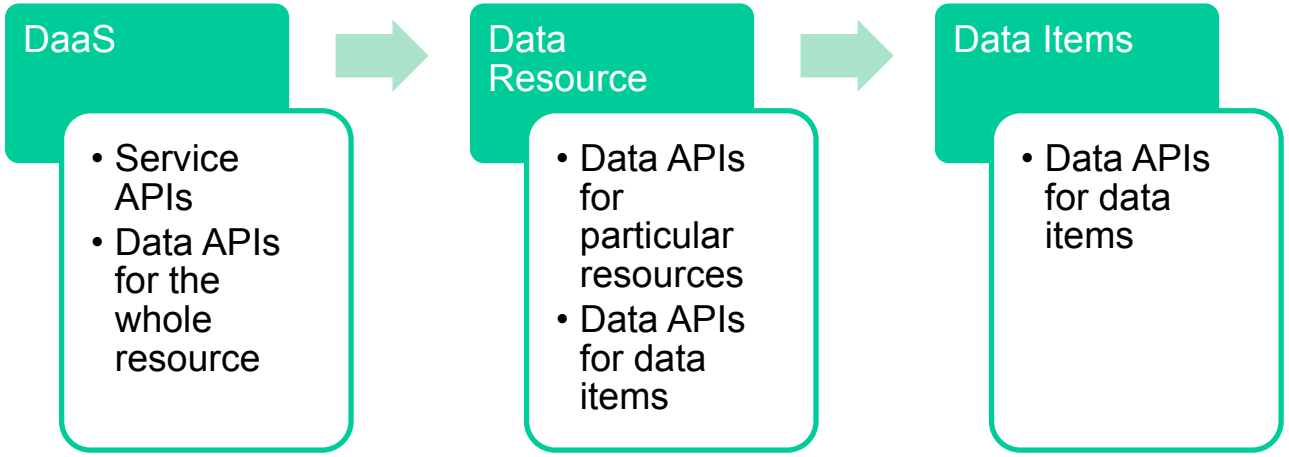
D&B Business Insight — data

published by: DNB

With D&B Business Insight, professionals can generate a list of

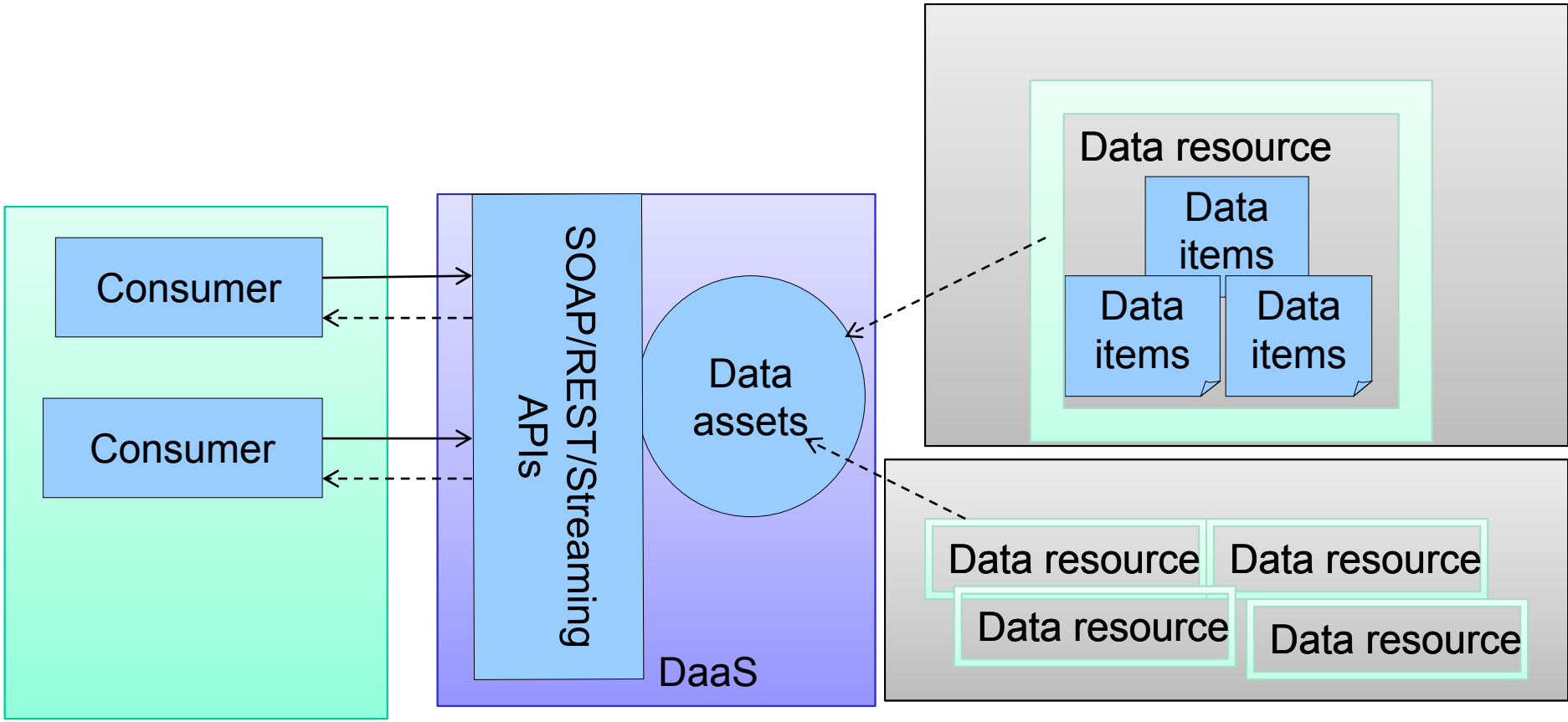
DaaS design & implementation – structures

Three levels

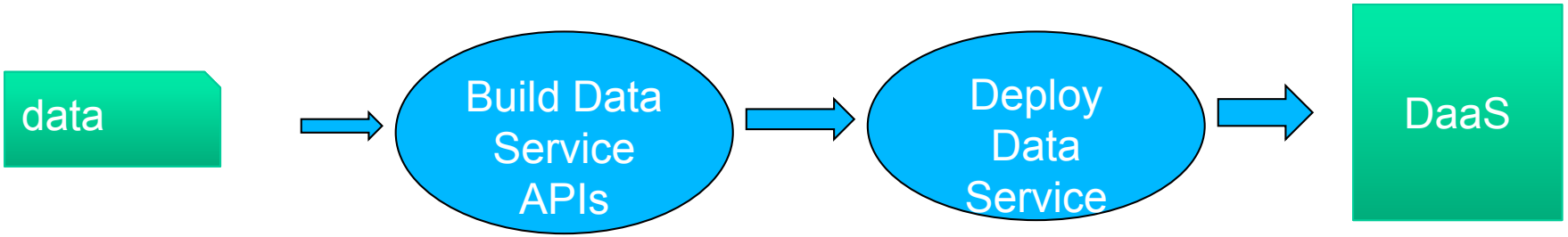


- DaaS and data providers have the right to publish the data

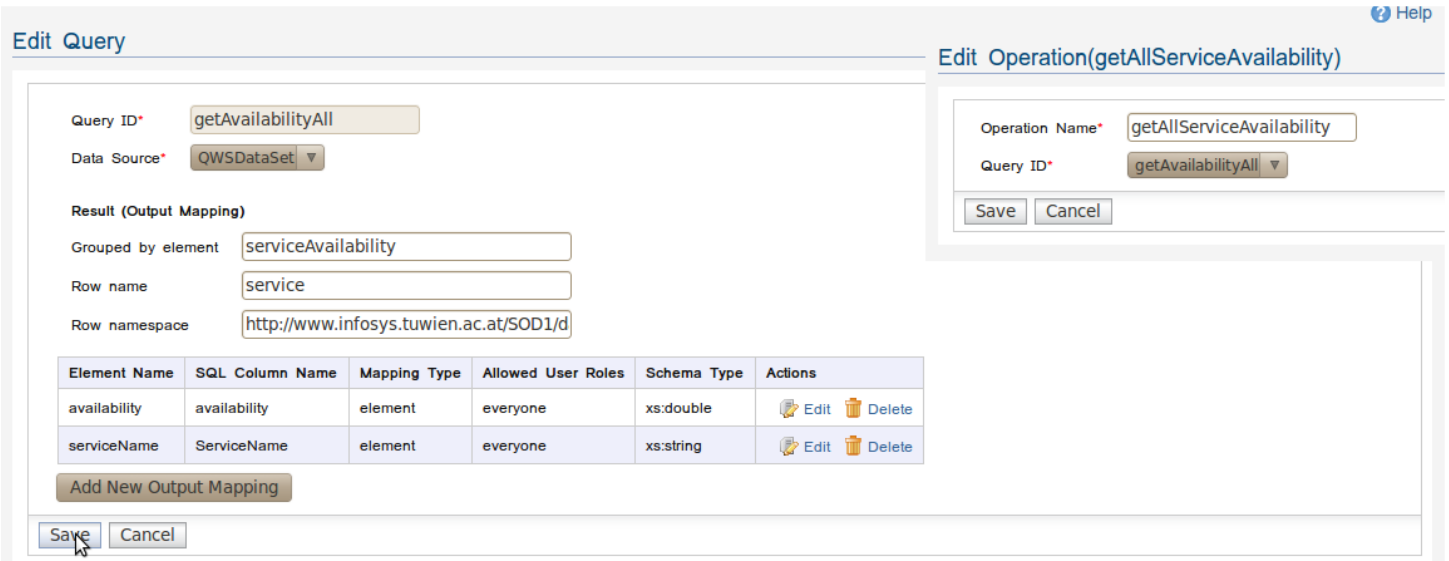
DaaS design & implementation – structures (2)



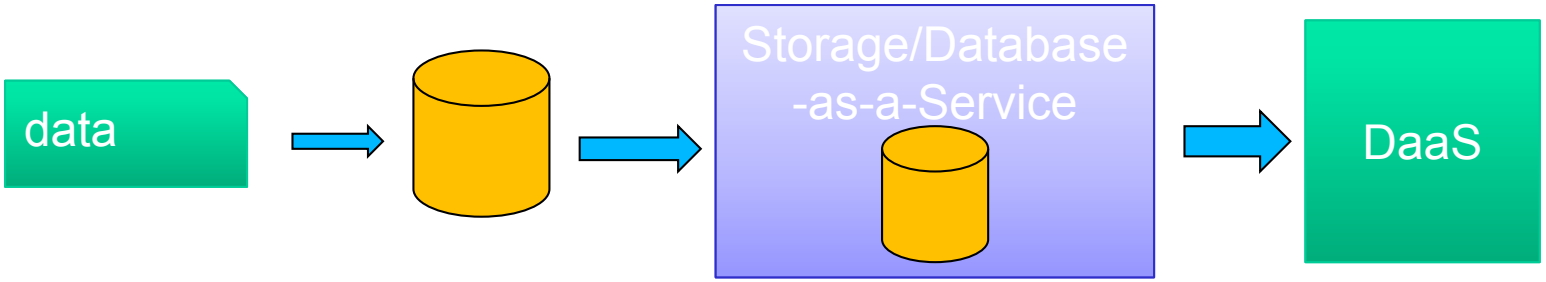
DaaS design & implementation – patterns for „turning data to DaaS“ (1)



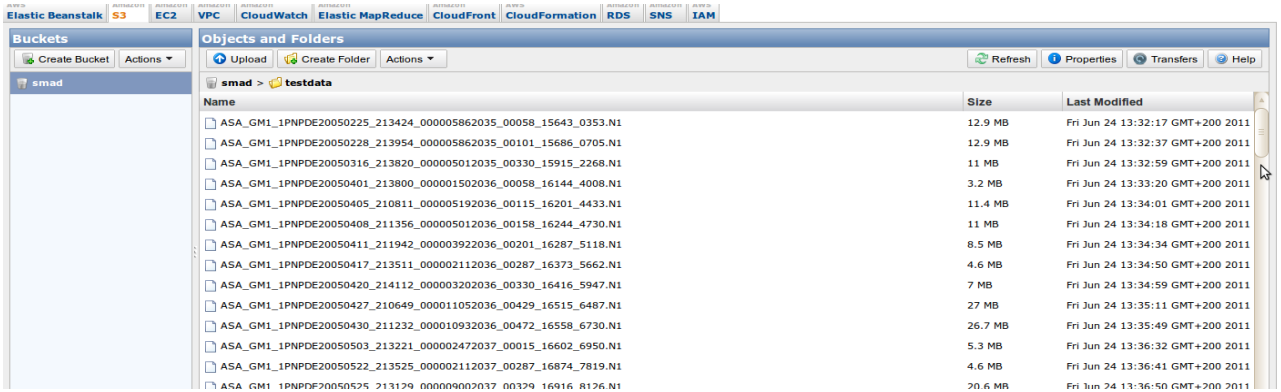
Examples: using WSO2 data service



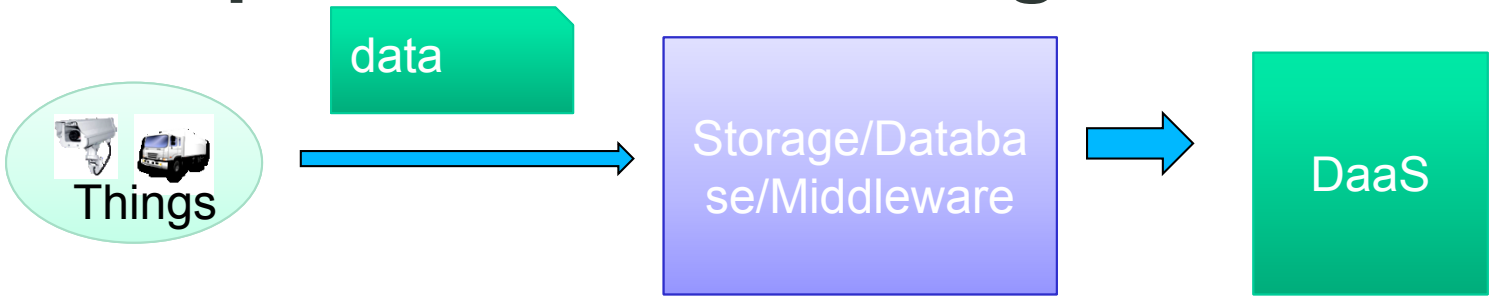
DaaS design & implementation – patterns for „turning data to DaaS“ (2)



Examples: using Amazon S3

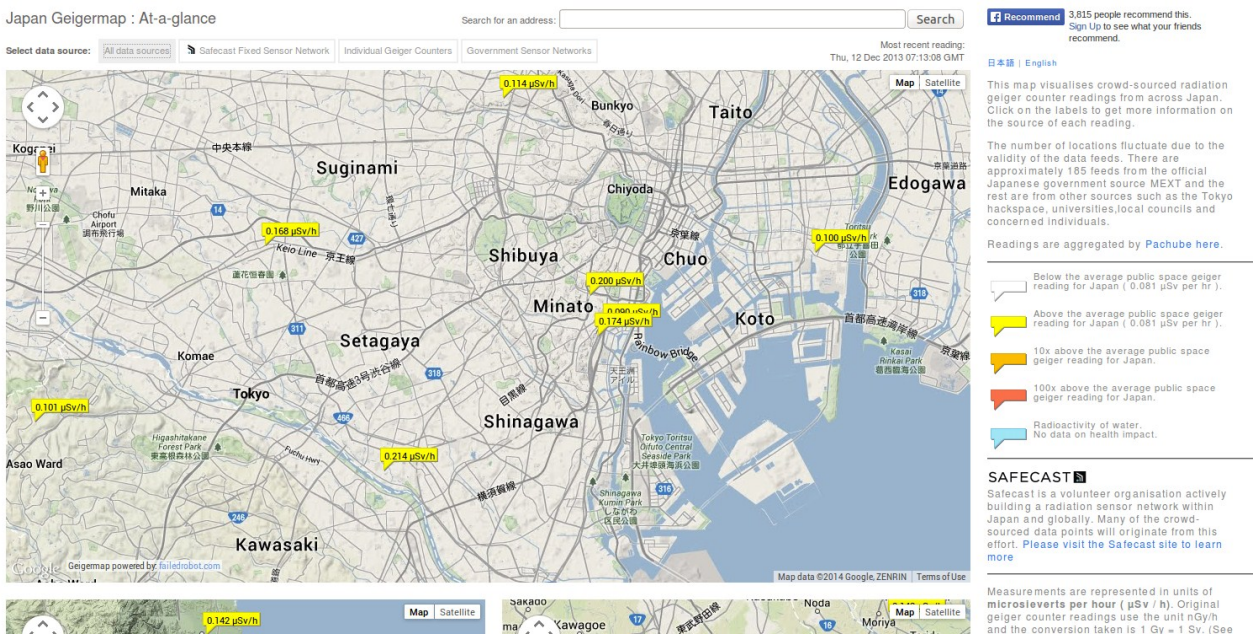


DaaS design & implementation – patterns for „turning data to DaaS“ (3)

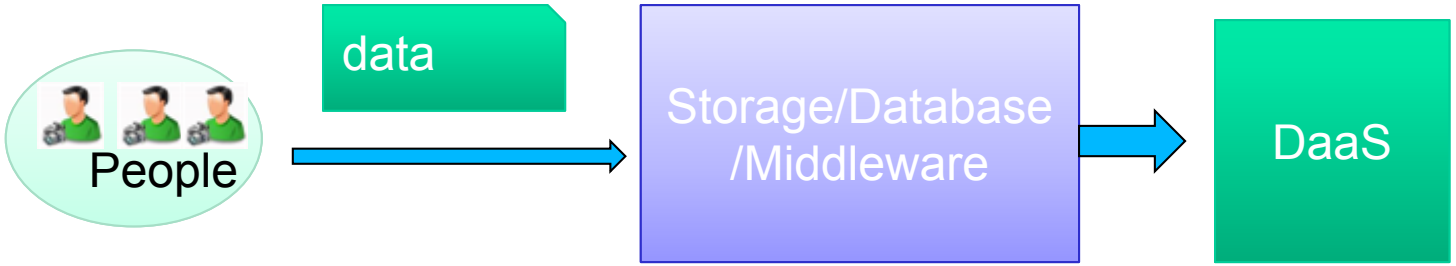


One Thing → 10000... Things

Examples: using Crowd-sourcing with Pachube (the predecessor of Xively)



DaaS design & implementation – patterns for „turning data to DaaS“ (4)



Examples: using Twitter

Developers API Health Blog Discussions Documentation Search Sign in

Home

REST API v1.1 Resources

Jump to

Timelines

Timelines are collections of Tweets, ordered with the most recent first.

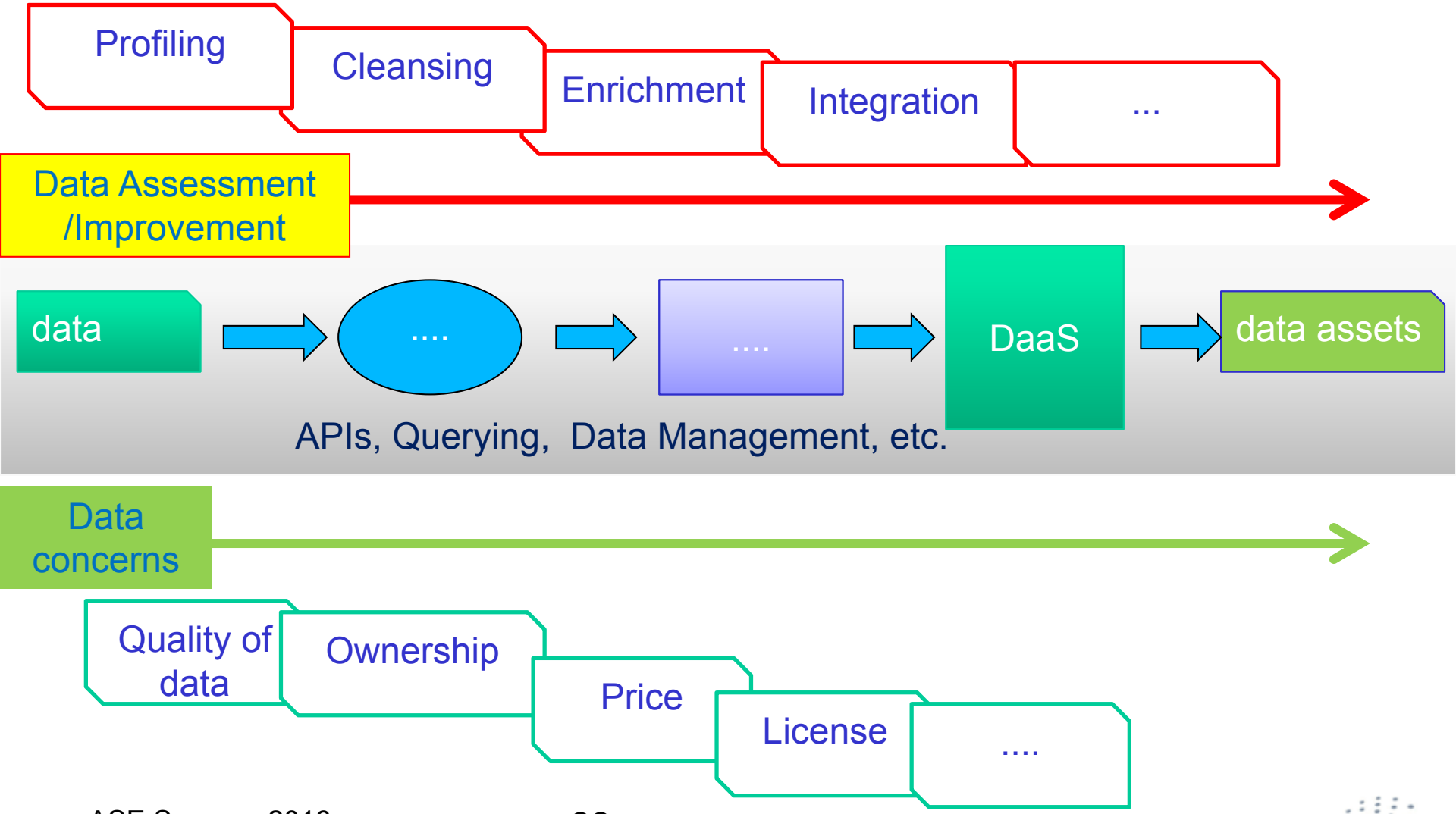
Resource	Description
GET statuses/mentions_timeline	Returns the 20 most recent mentions (tweets containing a users's @screen_name) for the authenticating user. The timeline returned is the equivalent of the one seen when you view your mentions on twitter.com. This method can only return up to 800 tweets. See Working with Timelines for...
GET statuses/user_timeline	Returns a collection of the most recent Tweets posted by the user indicated by the screen_name or user_id parameters. User timelines belonging to protected users may only be requested when the authenticated user either "owns" the timeline or is an approved follower of the owner. The timeline...
GET statuses/home_timeline	Returns a collection of the most recent Tweets and retweets posted by the authenticating user and the users they follow. The home timeline is central to how most users interact with the Twitter service. Up to 800 Tweets are obtainable on the home timeline. It is more volatile for users that follow...
GET statuses/retweets_of_me	Returns the most recent tweets authored by the authenticating user that have been retweeted by others. This timeline is a subset of the user's GET statuses/user_timeline. See Working with Timelines for instructions on traversing timelines.

Tweets

Tweets are the atomic building blocks of Twitter, 140-character status updates with additional associated metadata. People tweet for a variety of reasons about a multitude of topics.

Resource	Description
GET statuses/retweets/:id	Returns a collection of the 100 most recent retweets of the tweet specified by the id parameter.

DaaS design & implementation – not just „functional“ aspects (1)

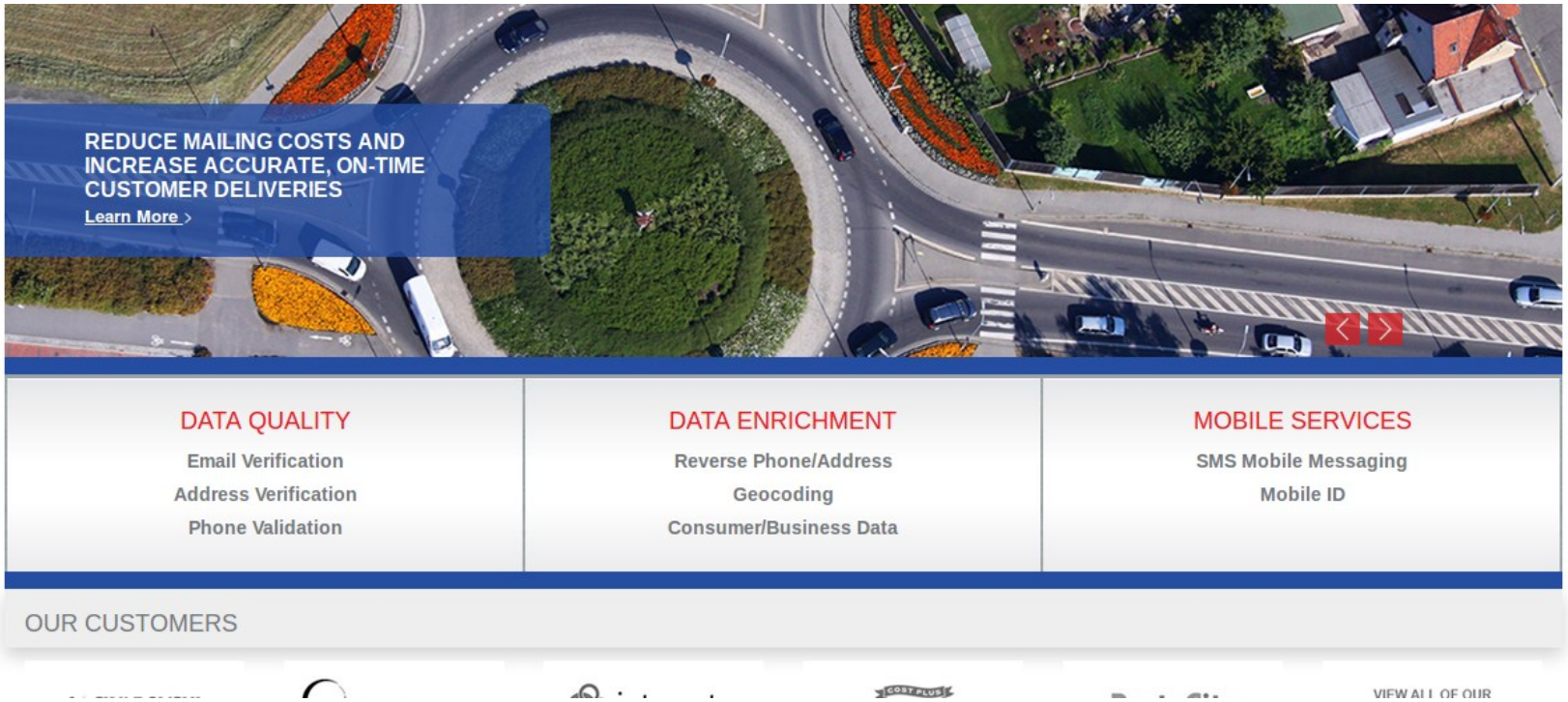


DaaS design & implementation – not just „functional“ aspects (2)

Understand the DaaS ecosystem

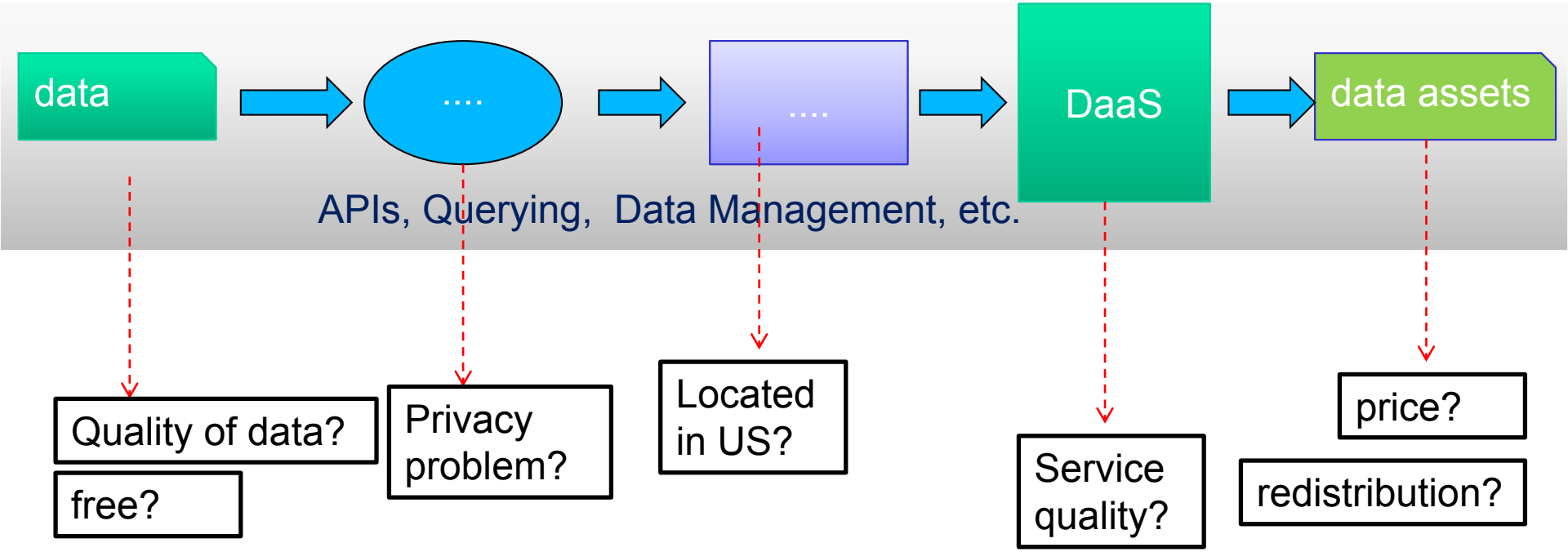
Specifying, Evaluating and Provisioning *Data concerns and Data Contract*

Example - http://www.strikeiron.com/



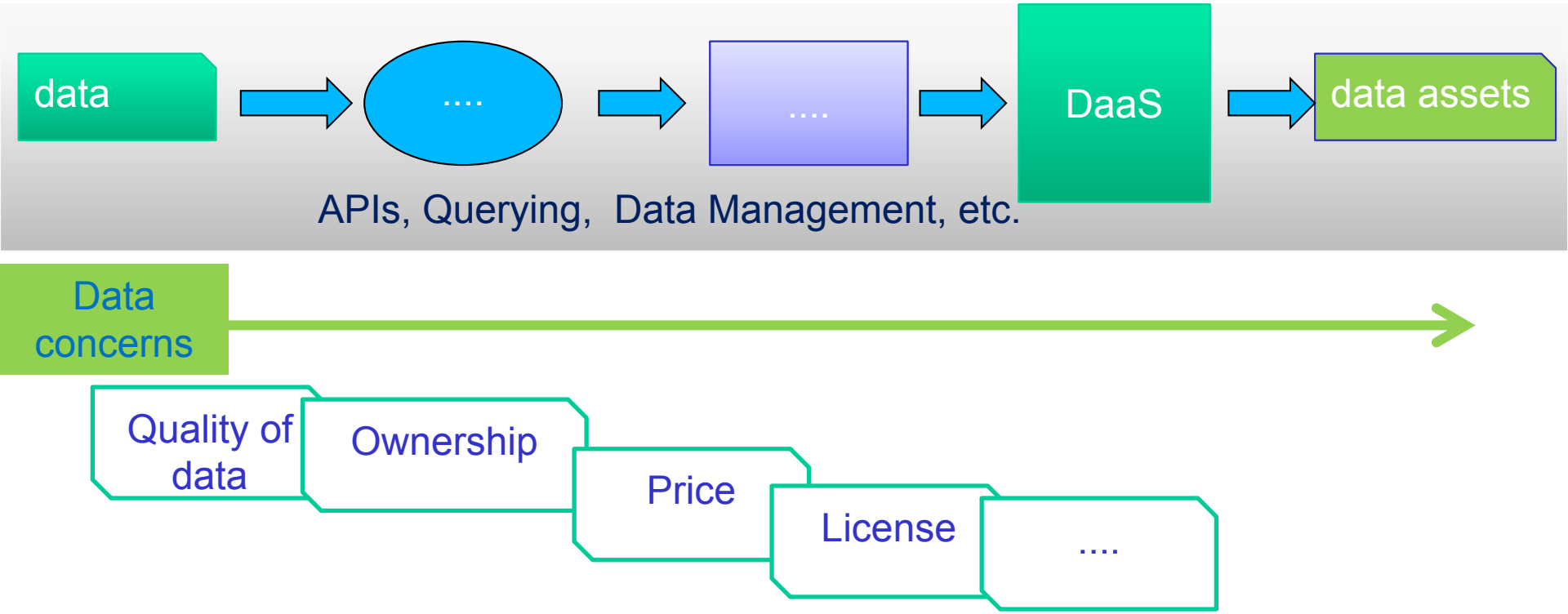
DATA CONCERNS

What are data concerns?



Read: Carlo Batini, Monica Scannapieco: Data and Information Quality - Dimensions, Principles and Techniques. Data-Centric Systems and Applications, Springer 2016, ISBN 978-3-319-24104-3, pp. 1-449

DaaS concerns



DaaS concerns include QoS, quality of data (QoD), service licensing, data licensing, data governance, etc.

Why DaaS/data concerns are important?

- Too much data returned to the consumer/integrator are not good
- Results are returned without a clear usage and ownership causing data compliance problems
- Consumers want to deal with dynamic changes

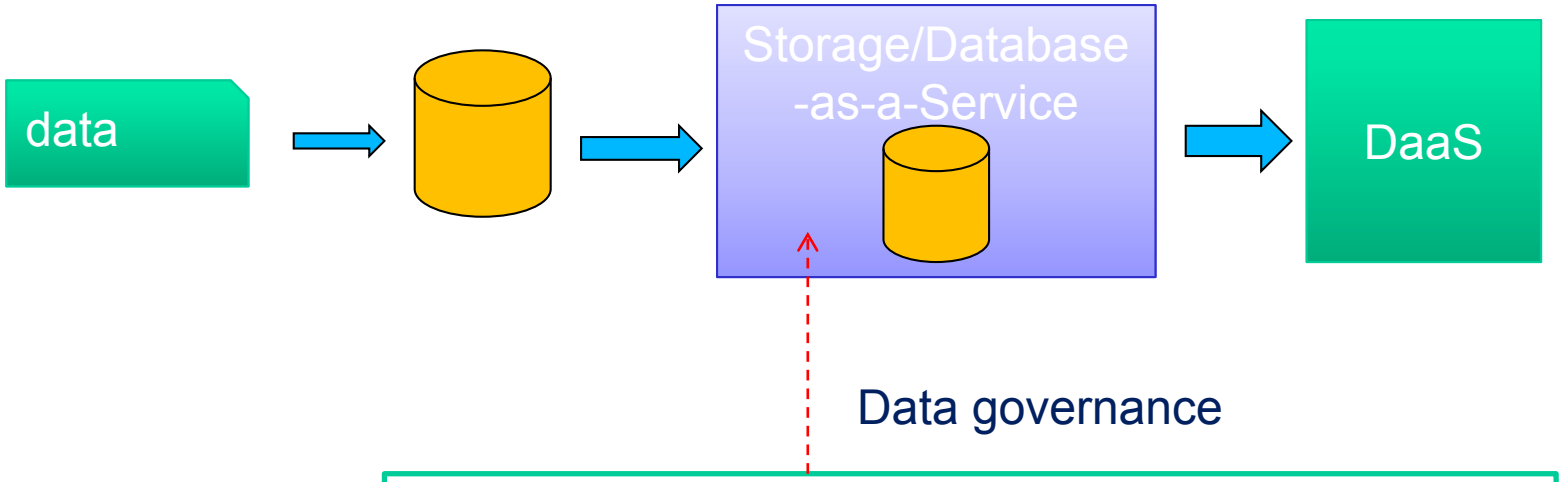
Ultimate goal: to provide *relevant* data with *acceptable constraints on data concerns in different provisioning models*

DaaS concerns analysis and specification

- Which concerns are important in which situations?
- How to specify concerns?

Hong Linh Truong, Schahram Dustdar On analyzing and specifying concerns for data as a service. APSCC 2009: 87-94

Data governance



Important factor, for example, the security and privacy compliance, data distribution, and auditing

Quality of data

Read-only DaaS

- Important factor for the selection of DaaS.
- For example, the accurary and compleness of the data, whether the data is up-to-date

CRUD DaaS

- Expected some support to control the quality of the data in case the data is offered to other consumers

Data and service usage

Read-only DaaS

- Important factor, in particular, price, data and service APIs licensing, law enforcement, and Intellectual Property rights

CRUD DaaS

- Important factor, in paricular, price, service APIs licensing, and law enforcement

Quality of service

Read-only DaaS

- Important factor, in particular availability and response time

CRUD Daas

- Important factor, in particular, availability, response time, dependability, and security

Contextual information

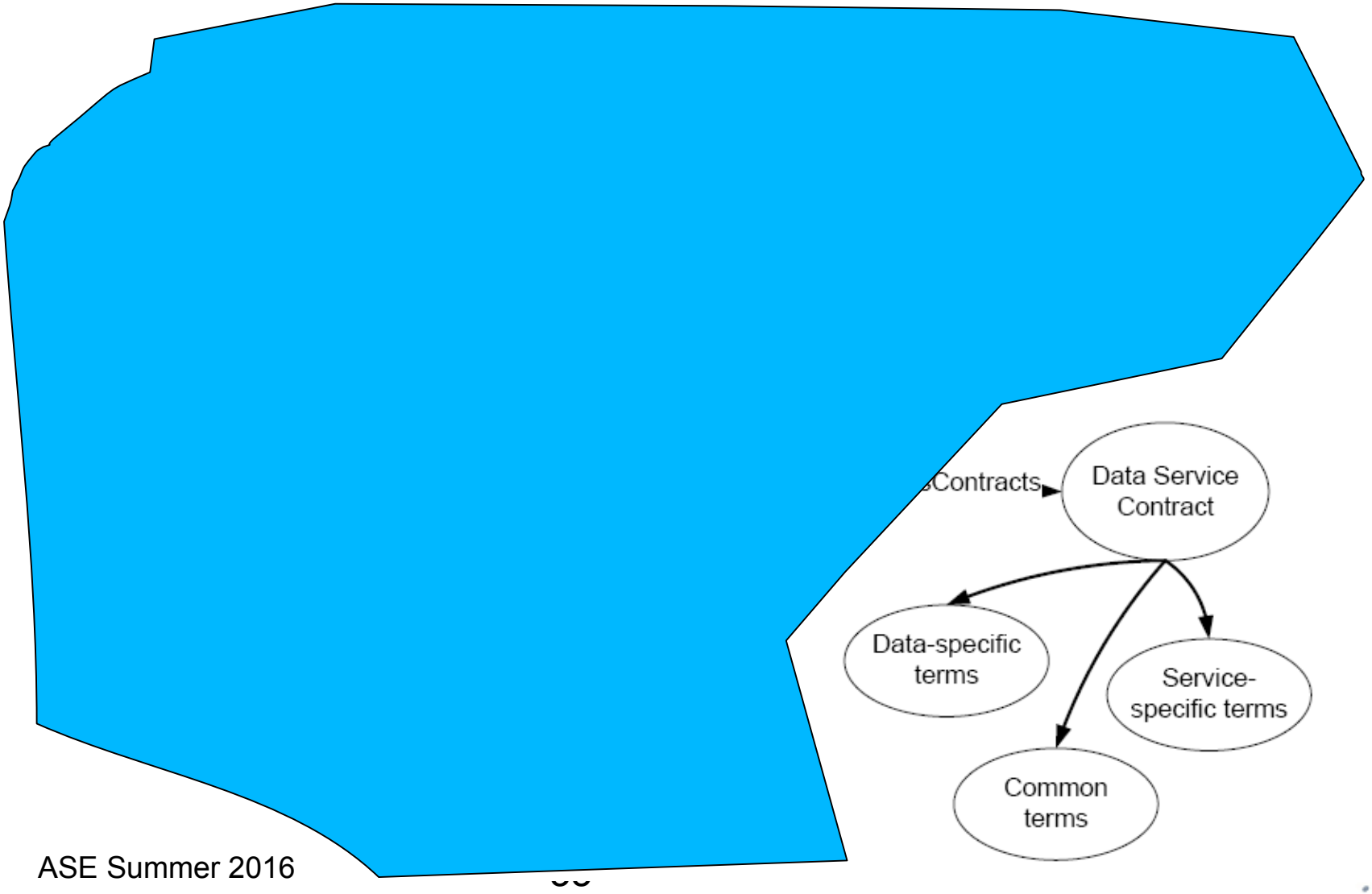
Read-only DaaS

- Useful factor, such as classification and service type (REST, SOAP), location

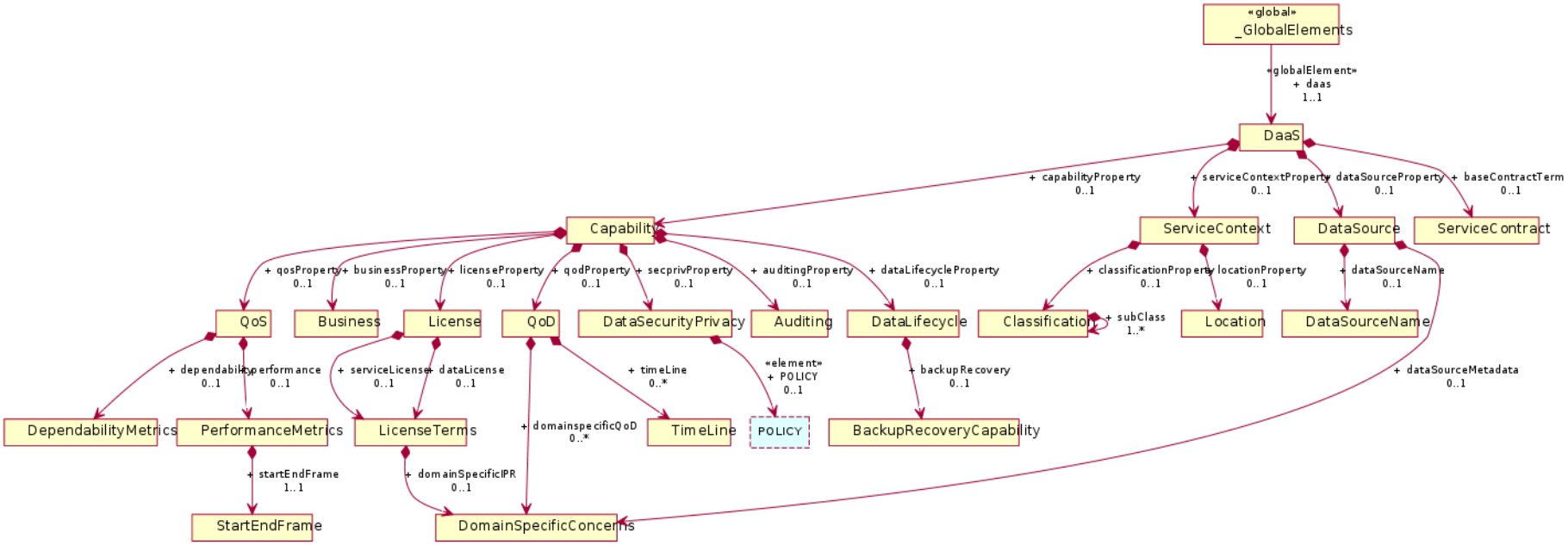
CRUD DaaS

- Important factor, e.g. location (for regulation compliance) and versioning

Conceptual model for DaaS concerns and contracts



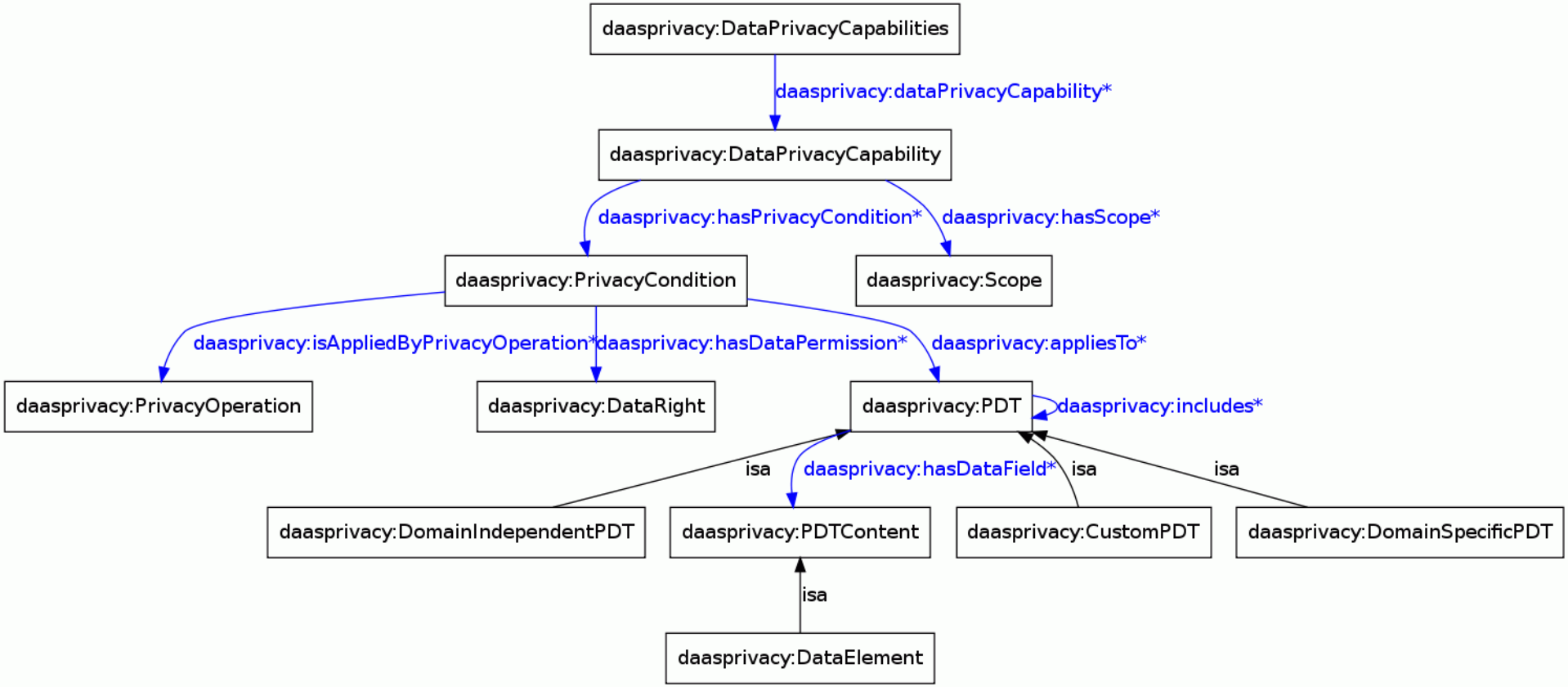
Implementation (1)



Check http://www.infosys.tuwien.ac.at/prototyp/SOD1/dataconcerns

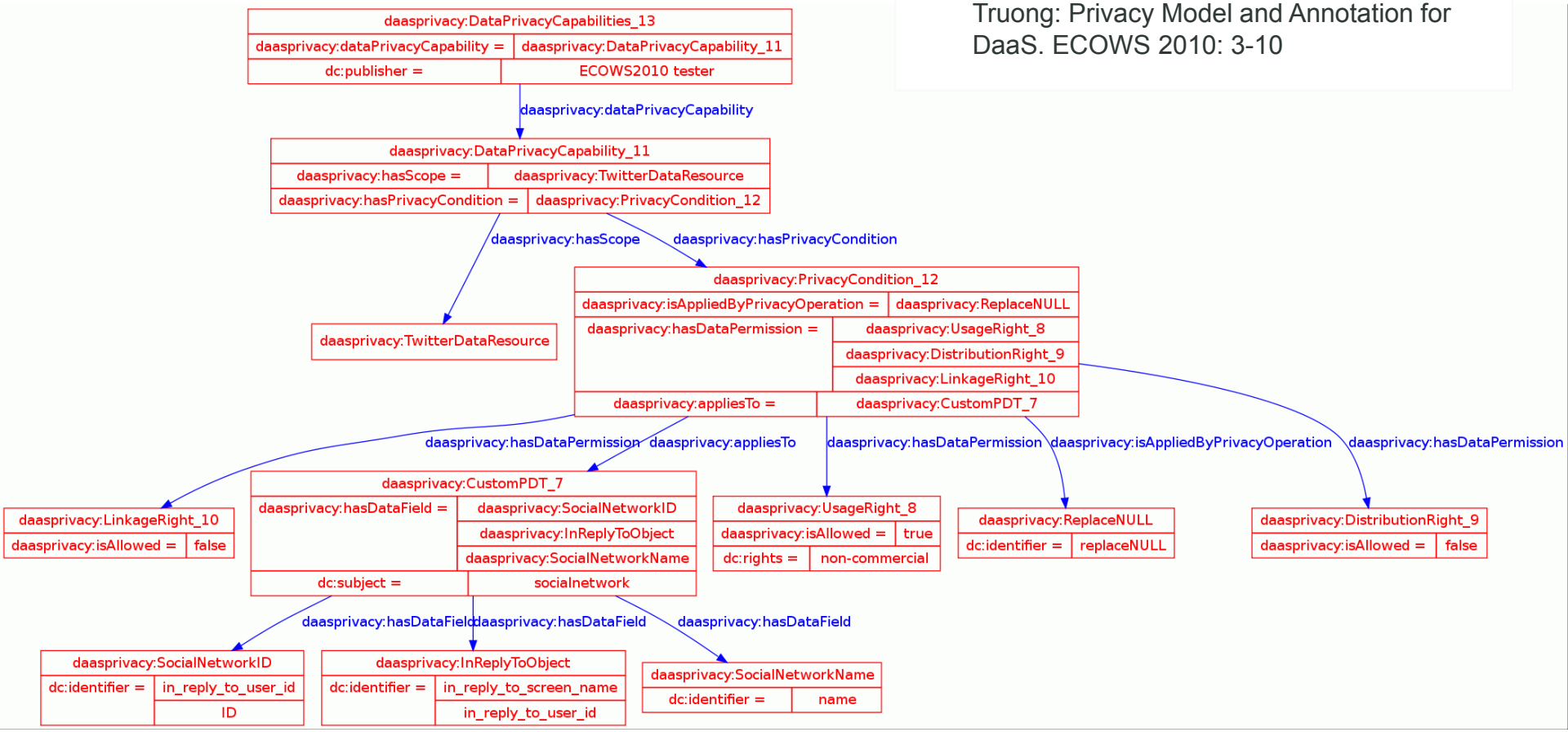
Implementation (2)

- Data privacy concerns are annotated with WSDL and MicroWSMO



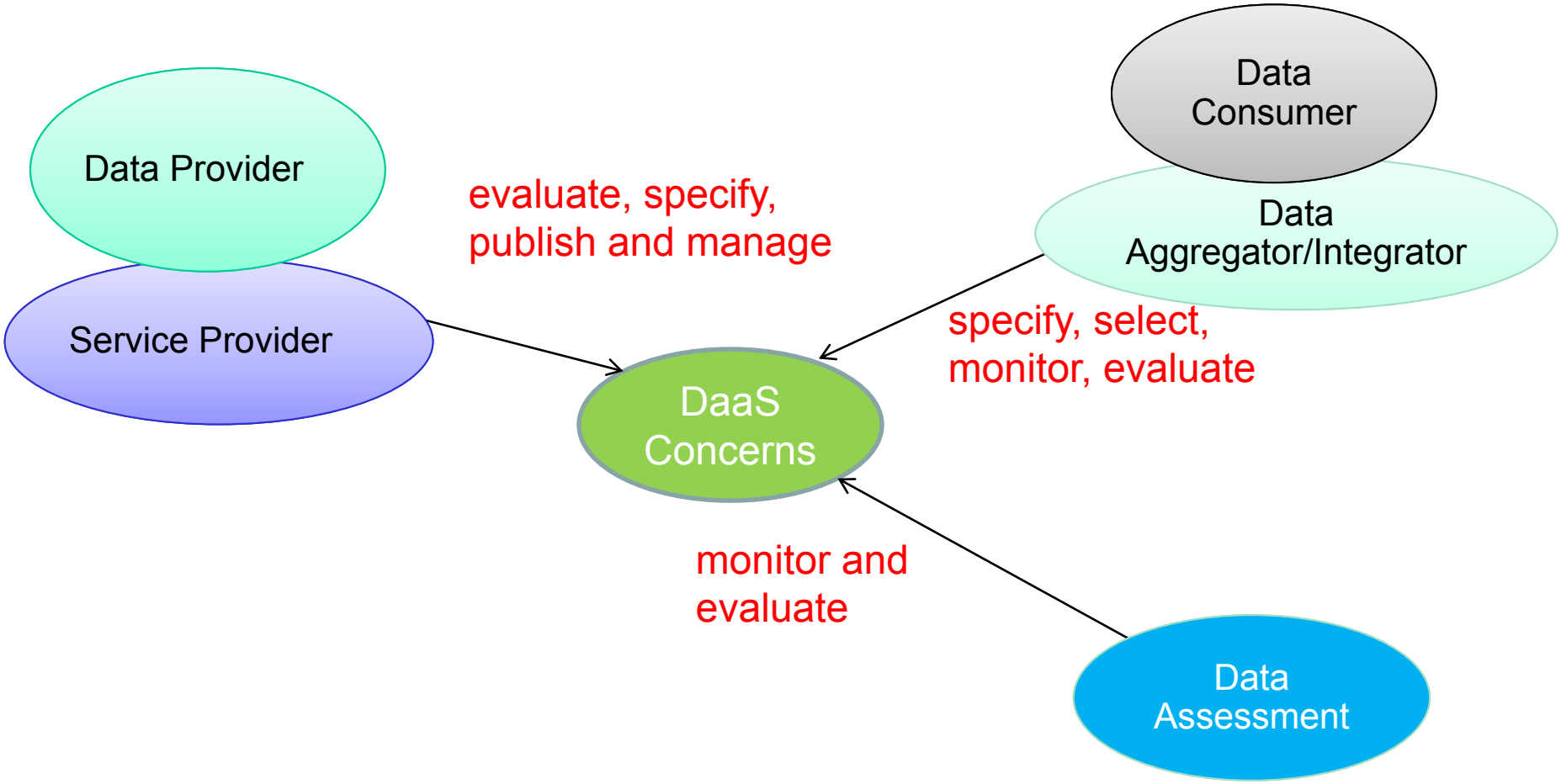
Implementation (3)

Michael Mrissa, Salah-Eddine Tbahriti, Hong Linh Truong: Privacy Model and Annotation for DaaS. ECOWS 2010: 3-10



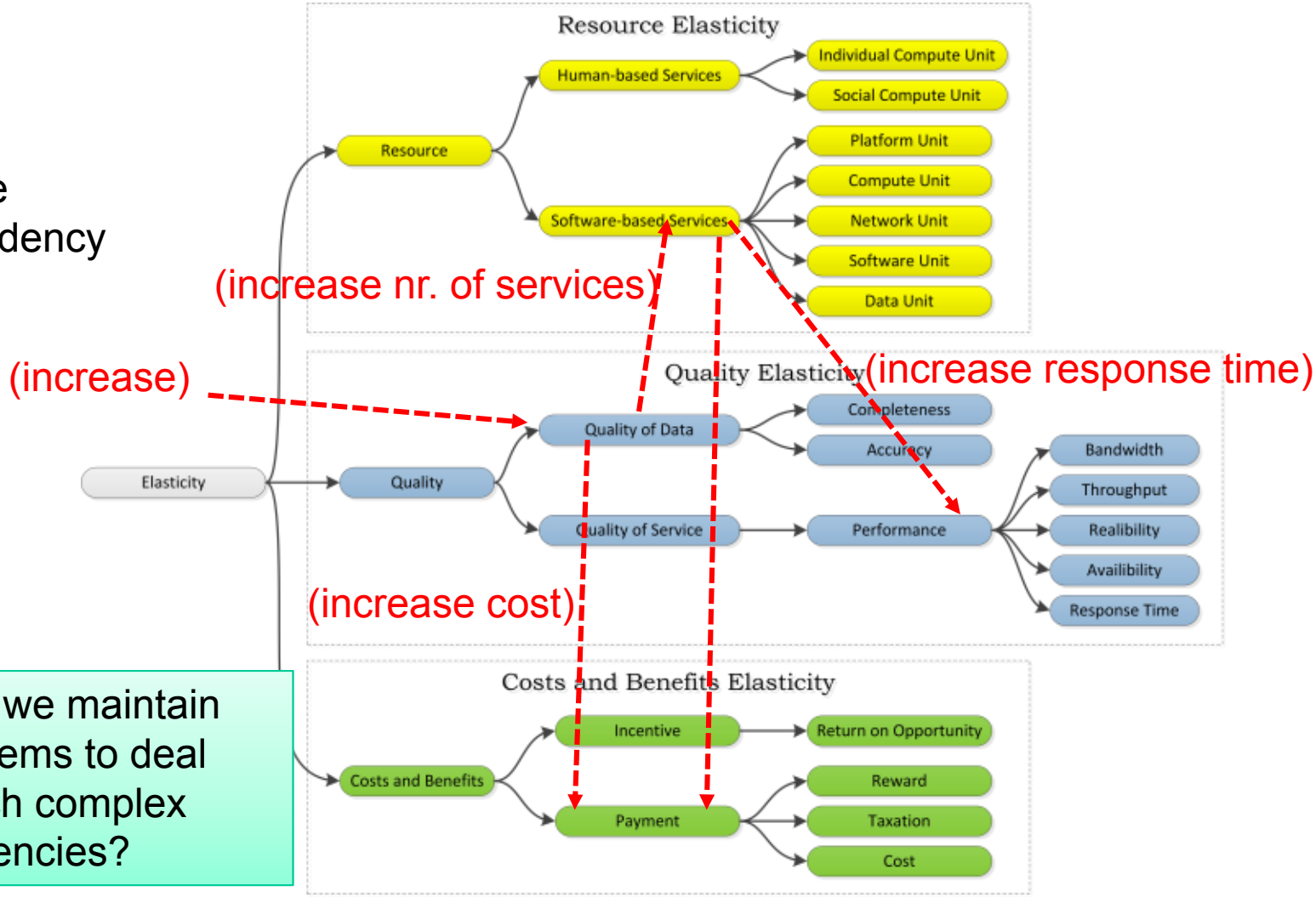
Populating DaaS concerns

The role of stakeholders in the most trivial view



Data concerns in multi-dimensional elasticity

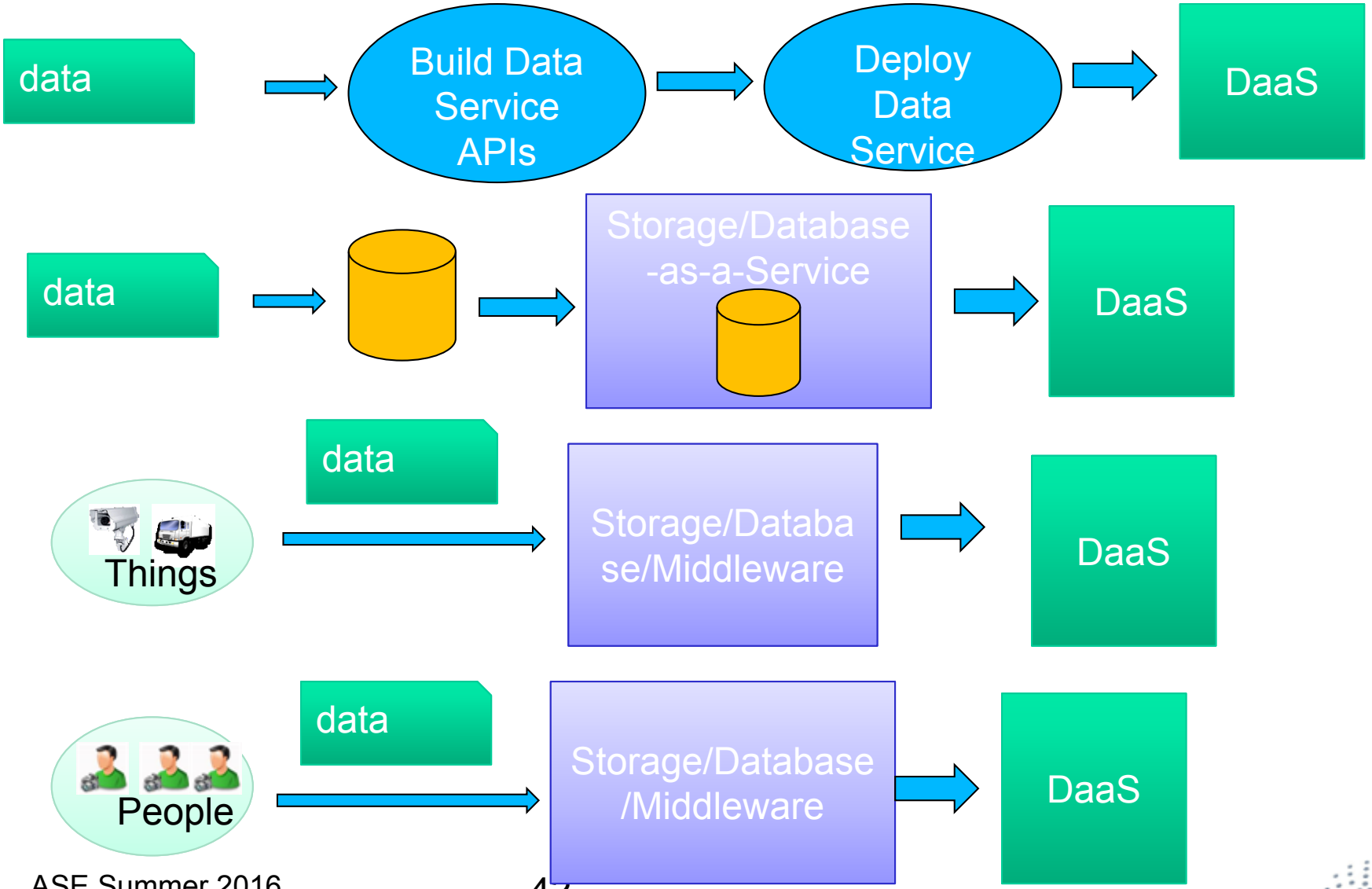
Simple dependency flows



How do we maintain our systems to deal with such complex dependencies?

HOW TO EVALUATE DATA CONCENRS FOR DATA ASSETS IN DAAS?

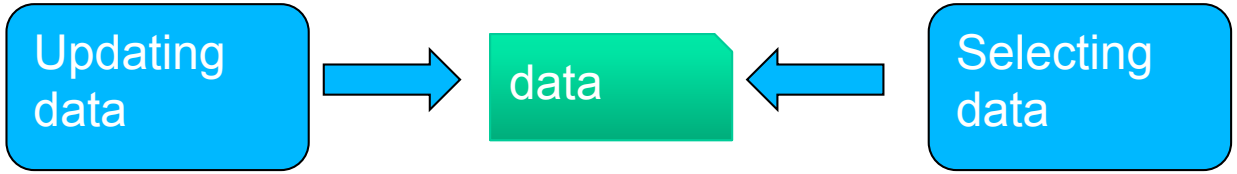
Patterns for „turning data to DaaS“



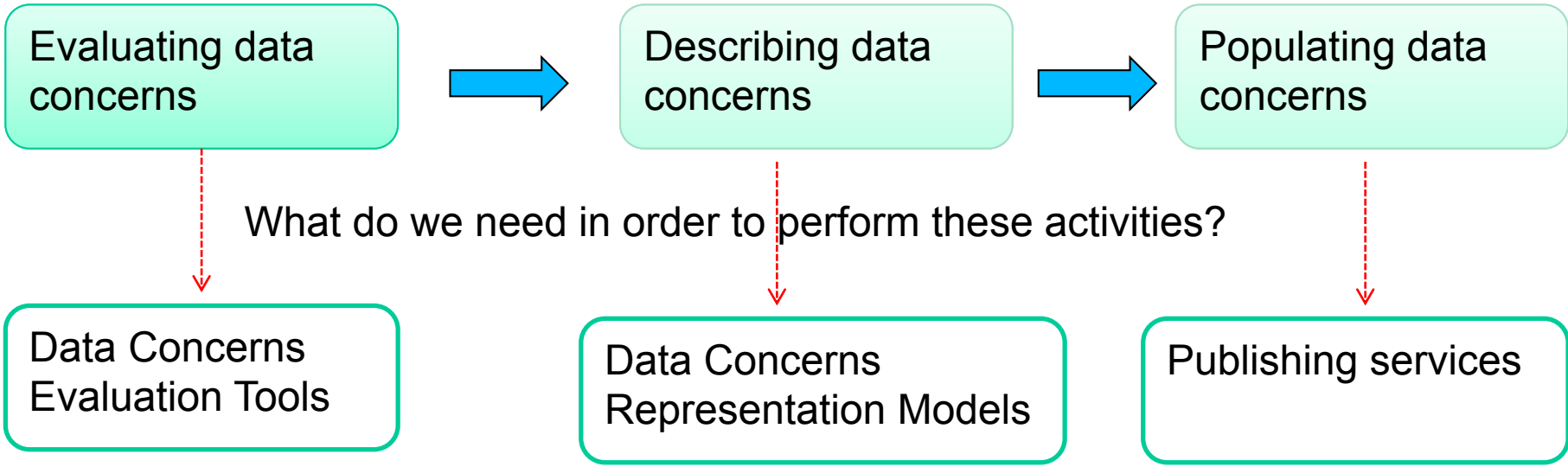
Data-related activities

Typical activities for data wrapping and publishing

Typical activities for data updating & retrieval

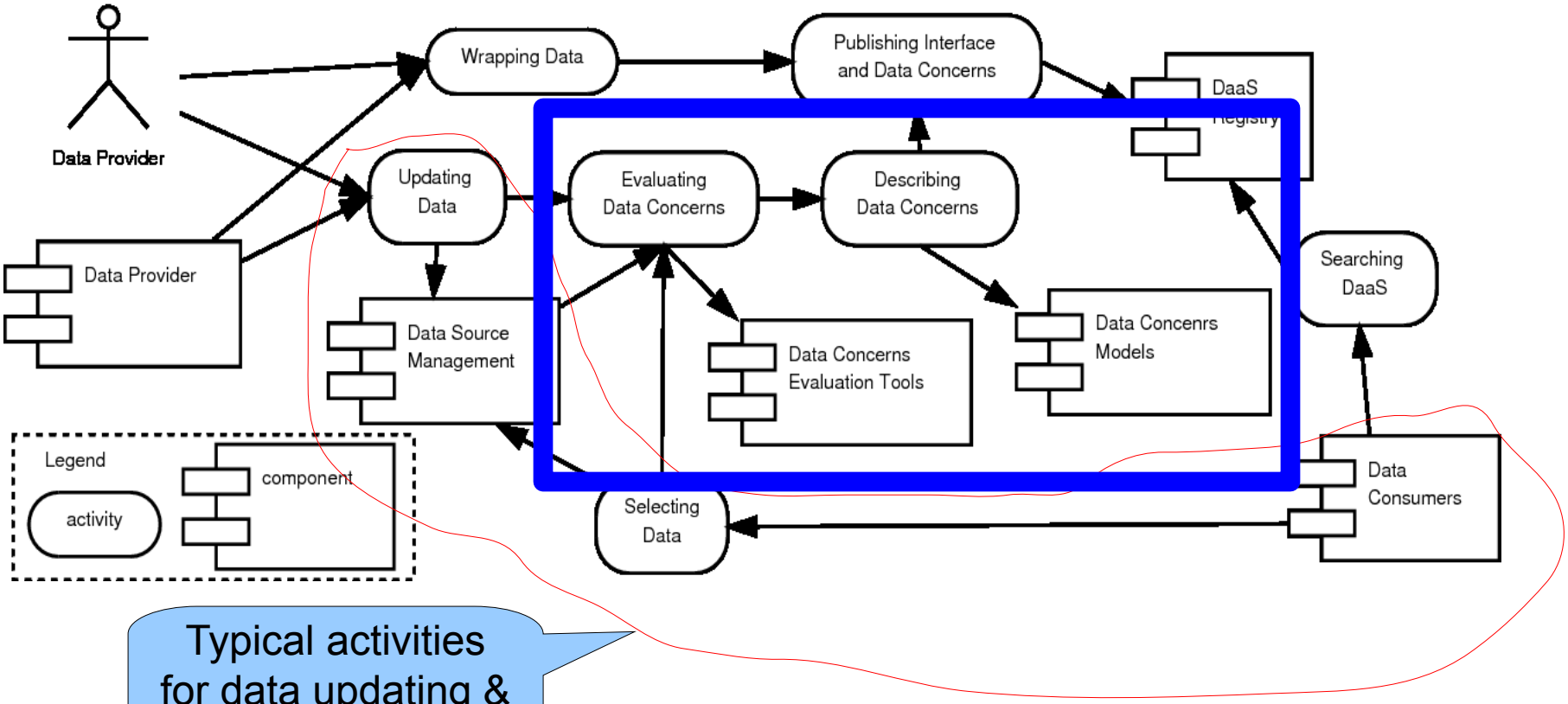


Typical data concern evaluation



Data concern-aware DaaS engineering process

Typical activities for data wrapping and publishing



Typical activities for data updating & retrieval

Hong Linh Truong, Schahram Dustdar: On Evaluating and Publishing Data Concerns for Data as a Service. APSCC 2010: 363-370

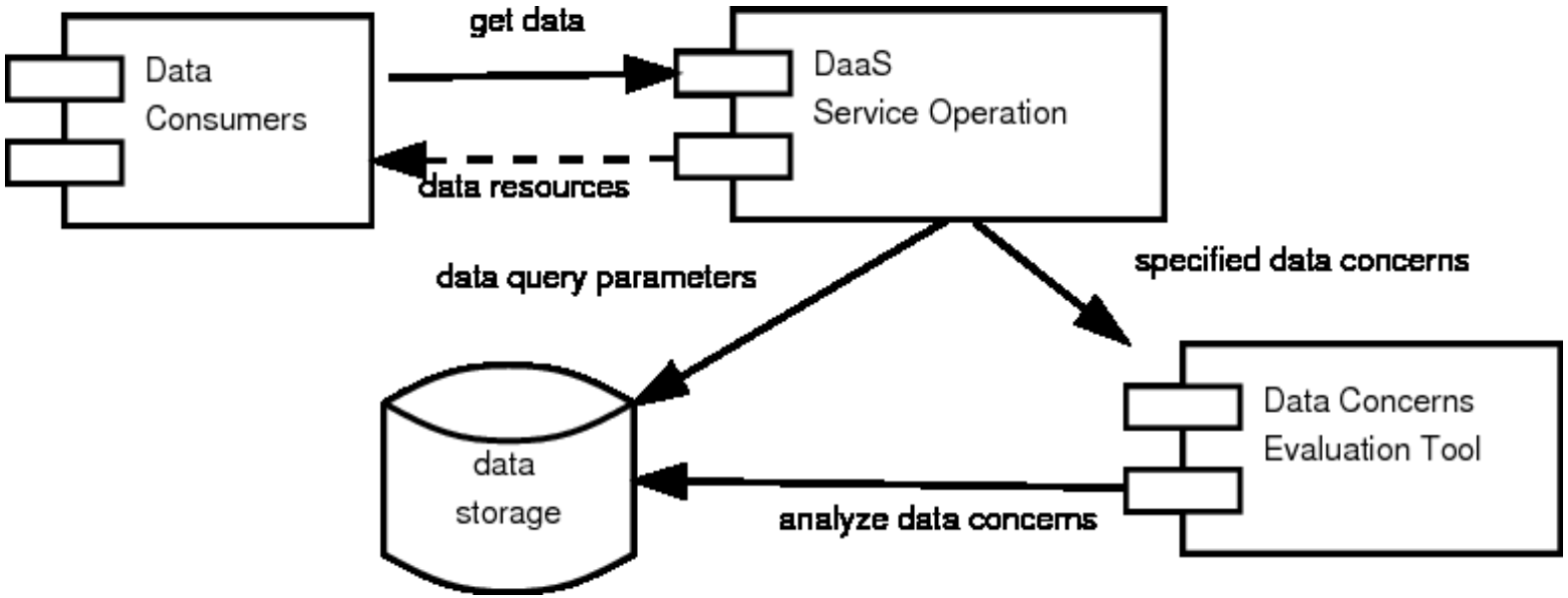
Evaluating data concerns – the three important points

evaluation scope	• At which level the evaluation is performed?
evaluation modes	• When the evaluation is done?
integration model	• How the evaluation tool is invoked?

Hong Linh Truong, Schahram Dustdar: On Evaluating and Publishing Data Concerns for Data as a Service. APSCC 2010: 363-370

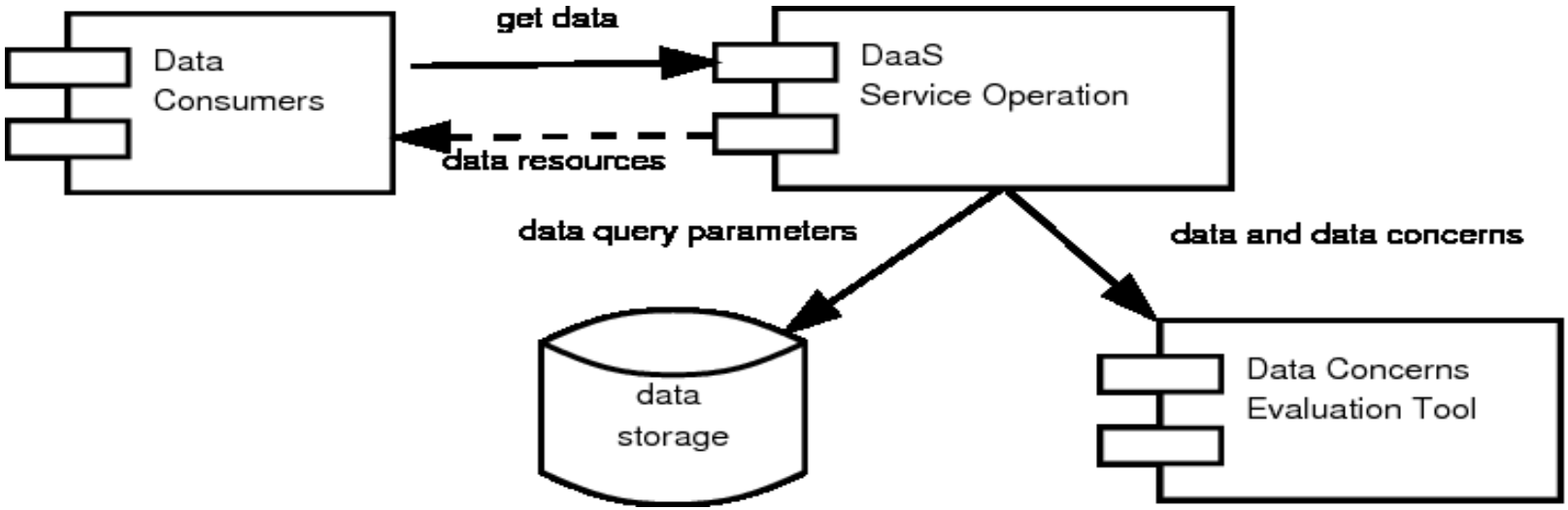
Evaluating data concerns – some patterns (1)

Pull, pass-by-references



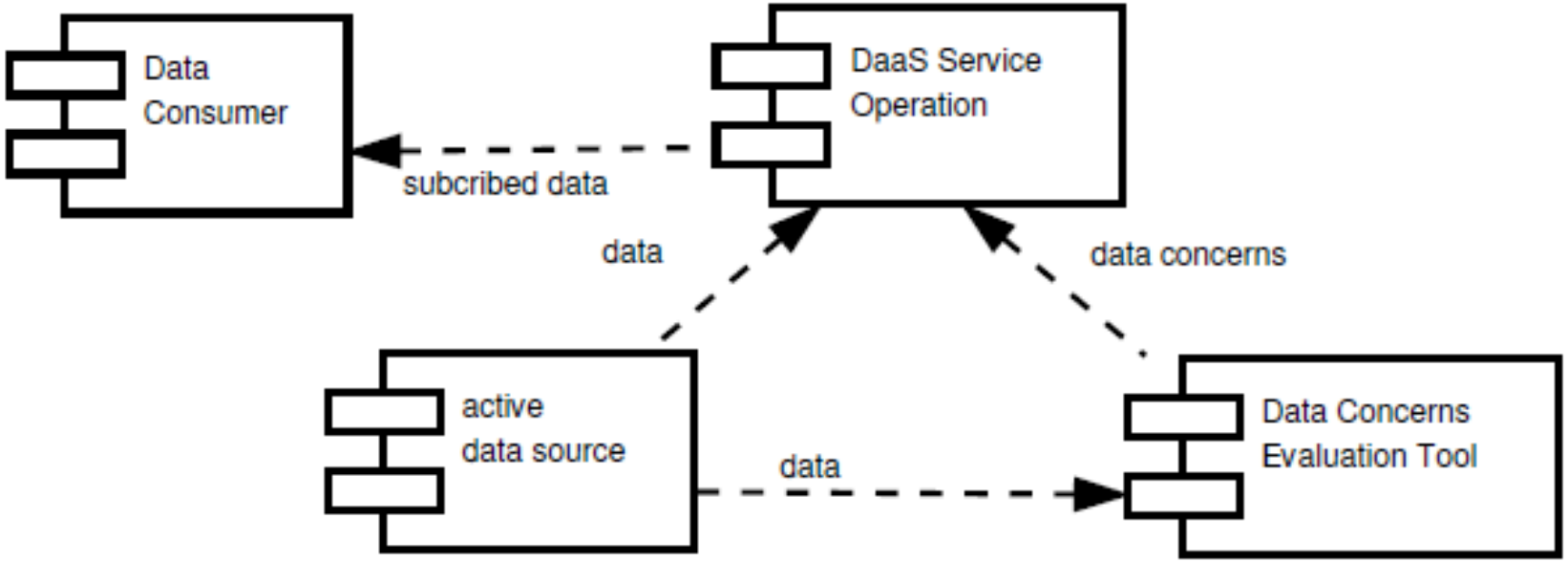
Evaluating data concerns – some patterns (2)

Pull, pass-by-values



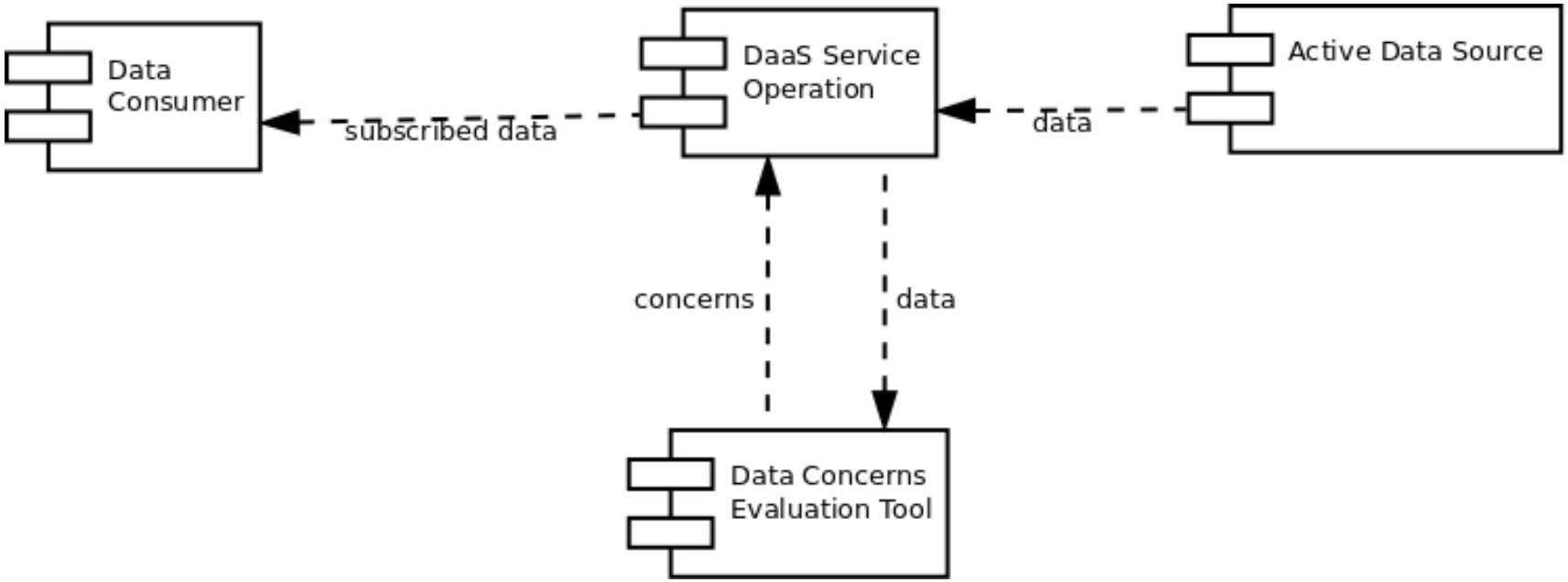
Evaluating data concerns – some patterns (3)

Push, pass-by-values (1)



Evaluating data concerns – some patterns (4)

Push, pass-by-values (2)



Evaluation Tool – Internal Software components

- Self-developed or third-party software components for evaluation tool
- Advantages
  - Tightly couple integration → performance, security, data compliance
  - Customization
- Disadvantages
  - Usually cannot be integrated with other features (e.g., data enrichment)
  - Costly (e.g., what if we do not need them)

Evaluation tool – using cloud services

- Evaluation features are provided by cloud services
- Several implementations
  - Informatica Cloud Data Quality Web Services, StrikeIron,
- Advantages
  - Pay-per-use, combined features
- Disadvantages
  - Features are limited (with certain types of data)
  - Performance issues with large-scale data
  - Data compliance and security assurance

Evaluation Tool -- using human computation capabilities

- Professionals and Crowds can act as data concerns evaluators
  - For complex quality assessment that cannot be done by software
- Issues
  - Subjective evaluation
  - Performance
  - Limited type of data (e.g., images, documents, etc.)

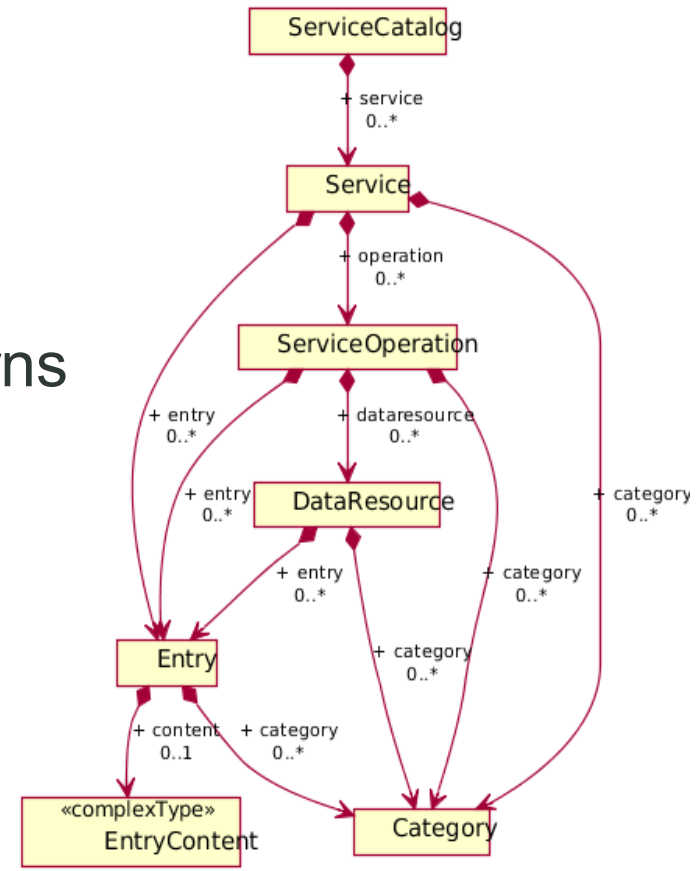
Michael Reiter, Uwe Breitenbücher, Schahram Dustdar, Dimka Karastoyanova, Frank Leymann, Hong Linh Truong: A Novel Framework for Monitoring and Analyzing Quality of Data in Simulation Workflows. eScience 2011: 105-112

Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, Jens Lehmann: Crowdsourcing Linked Data Quality Assessment. International Semantic Web Conference (2) 2013: 260-276

Óscar Figuerola Salas, Velibor Adzic, Akash Shah, and Hari Kalva. 2013. Assessing internet video quality using crowdsourcing. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (CrowdMM '13). ACM, New York, NY, USA, 23-28. DOI=10.1145/2506364.2506366 http://doi.acm.org/10.1145/2506364.2506366

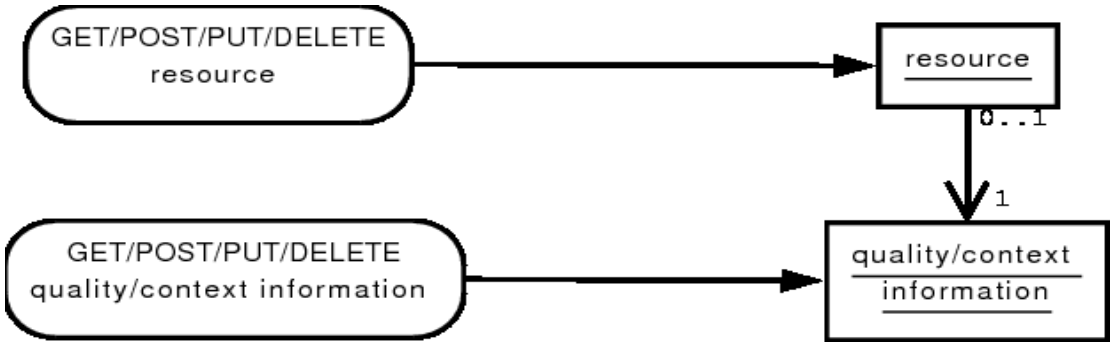
QoD framework: publishing concerns (1)

- Off-line data concern publishing, e.g.
  - a common data concern publication specification
  - a tool for providing data concerns according to the specification
  - supported by external service information systems



QoD framework: publishing concerns (2)

- On-the-fly querying data concerns associated with data resources, e.g.,
  - Using REST parameter convention
  - Based on metric names in the data concern specification



Hong Linh Truong, Schahram Dustdar, Andrea Maurino, Marco Comerio: Context, Quality and Relevance: Dependencies and Impacts on RESTful Web Services Design. ICWE Workshops 2010: 347-359

QoD framework: publishing concerns (3)

- Specifying requests by using utilizing query parameters the form of **metricName=value**

```
GET/resource?crq.accuracy="0.5"&crq.location=’’Europe”
```

- Obtaining contex and quality by using context and quality parameters without specifying value conditions

```
curl http://localhost:8080/UNDataService/data/query/Population annual growth rate
(percent)?crq.qod
{”crq.qod” : {
”crq.dataelementcompleteness ”: 0.8654708520179372,
”crq.datasetcompleteness”: 0.7356502242152466,
...
}}
```

Exercises

- Read mentioned papers
- Check characteristics, service models and deployment models of mentioned DaaS (and find out more)
- Identify services in the ecosystem of some DaaS
- Write small programs to test public DaaS, such as Xively, Microsoft Azure and Infochimps
- Turn some data to DaaS using existing tools

Exercises (2)

- Identify and analyze the relationships between data concerns evaluation tools and types of data
- Analyze trade-offs between on-line and off-line evaluation and when we can combine them
- Analyze how to utilize evaluated data concerns for optimizing data compositions
- Analyze situations when software cannot be used to evaluate data concerns

Thanks for your attention

Hong-Linh Truong
Distributed Systems Group, TU Wien
truong@dsg.tuwien.ac.at
http://dsg.tuwien.ac.at/staff/truong
@linhsolar