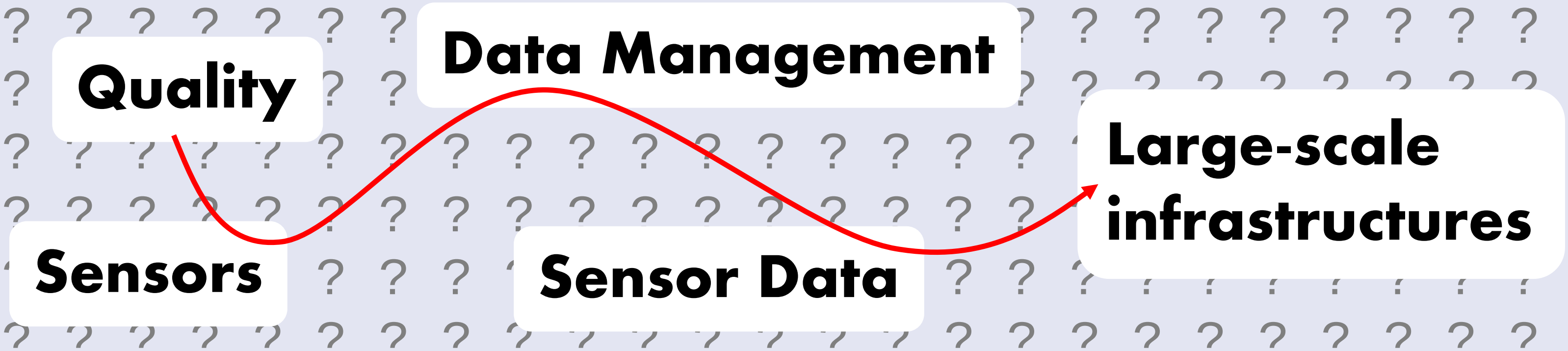
Universität Bamberg

Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing,
June 25 – June 30, 2017 in Crete, Greece

Prof. Dr. Daniela Nicklas
Chair of Computer Science, Mobile Software Systems / Mobility
daniela.nicklas@uni-bamberg.de

Overview

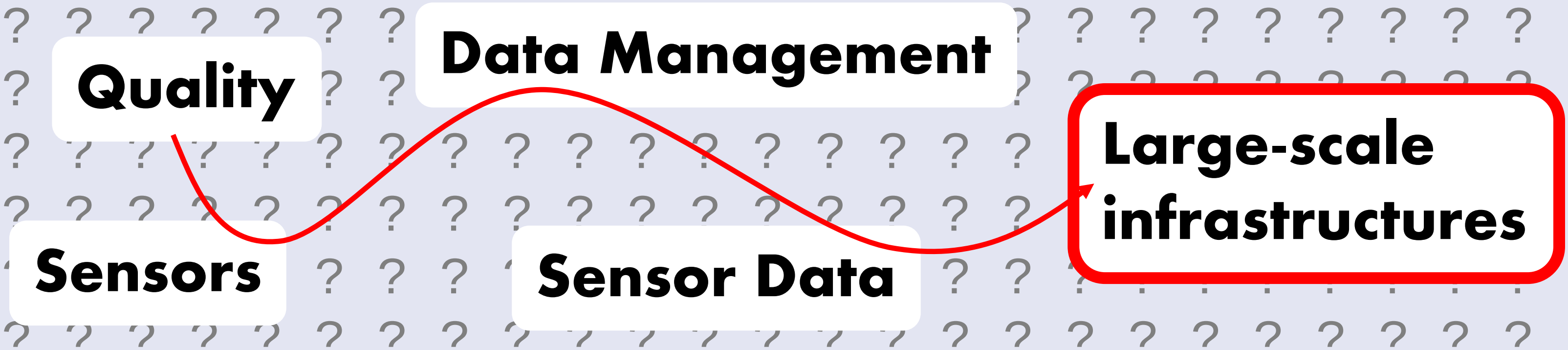


Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing, June 25 – June 30, 2017 in Crete, Greece

- Quality of what?
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management?
- Large-scale = Big = Very Large? How big is large?
- Infrastructures for what? For data or for real things?

Overview



Quality matters: Sensor Data Management for Large-scale Infrastructures

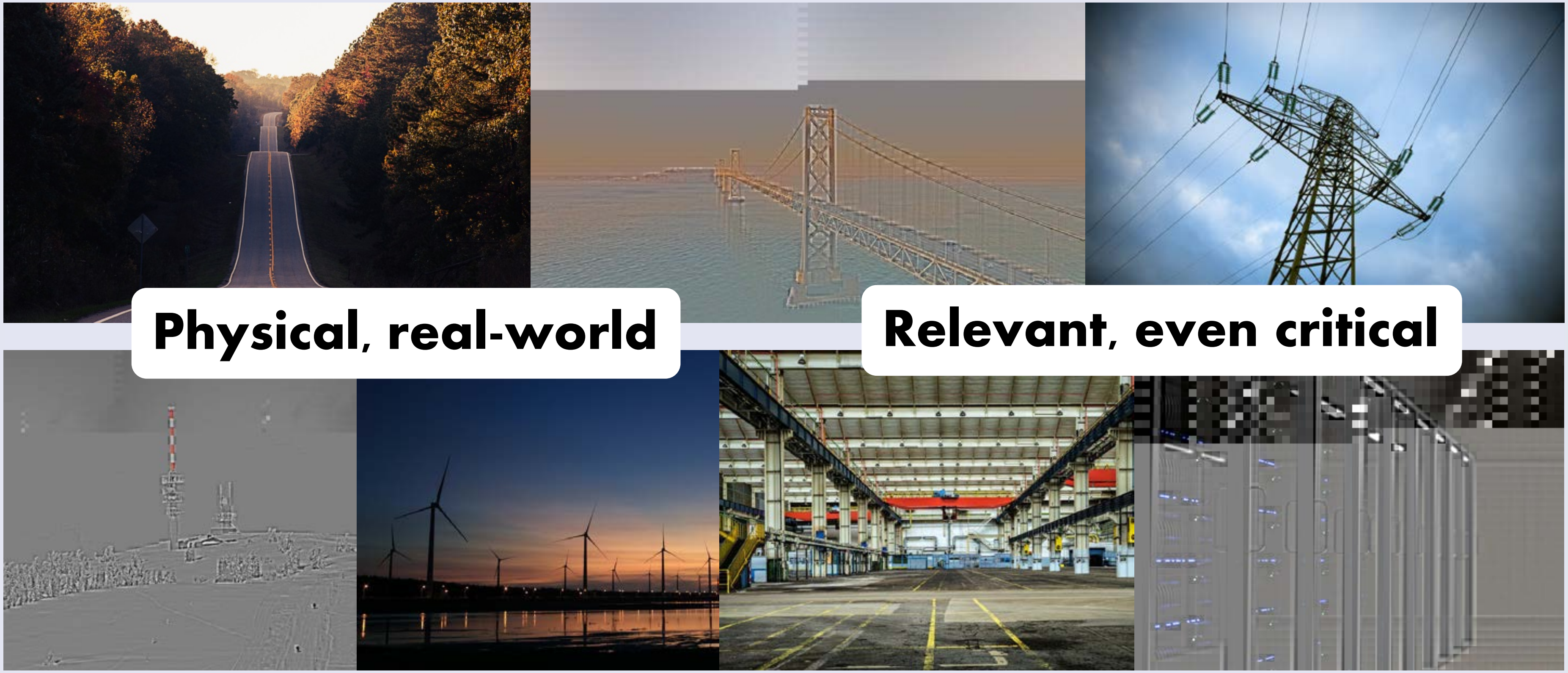
11th Symposium and Summer School On Service-Oriented Computing, June 25 – June 30, 2017 in Crete, Greece

- Quality of what?
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management?
- Large-scale = Big = Very Large? How big is large?
- Infrastructures for what? For data or for real things?

Infrastructure

Infrastructure refers to the fundamental facilities and systems serving a country, city, or area, including the services and facilities necessary for its economy to function.

https://en.wikipedia.org/wiki/Infrastructure



Physical, real-world

Relevant, even critical

Large-scale infrastructures

- Large-scale = Big = Very Large? How big is large?

VLDB 2017 Munich · Germany

43rd International Conference on Very Large Data Bases

How many software engineers does it need to change a light bulb?

A1: That's a hardware problem.

A2: One, but if he changes it, the whole building will probably fall down.

http://www.lightbulbjokes.com/

Large-scale / Very Large / Big Infrastructures:

- Many people involved
- Complex structure
- So large that conventional techniques have problems handling it or can‘t handle it at all

new techniques become conventional over time:
→ „big“ is a moving target

Why Sensors for Large-Scale Infrastructures?

Universität Bamberg

Physical, real-world

Relevant, even critical

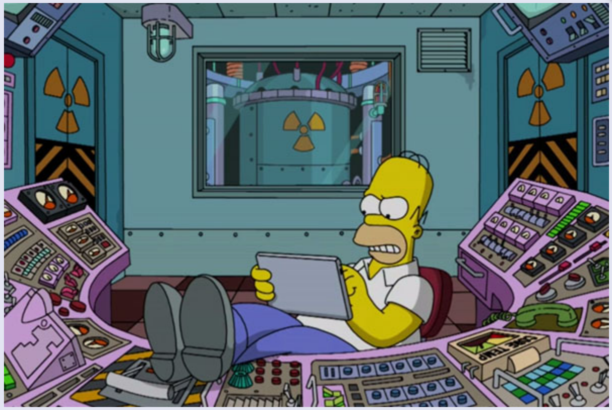
Should be monitored:

- **Normal operation**
- **Effects of changes**

Sensors!

Human monitoring?

- **Does not scale well**
- **Boring → Errors**



Applications

for X in …

Meter

Grid

Factory

Home

City

Phone

Transportation

X

+ sensors

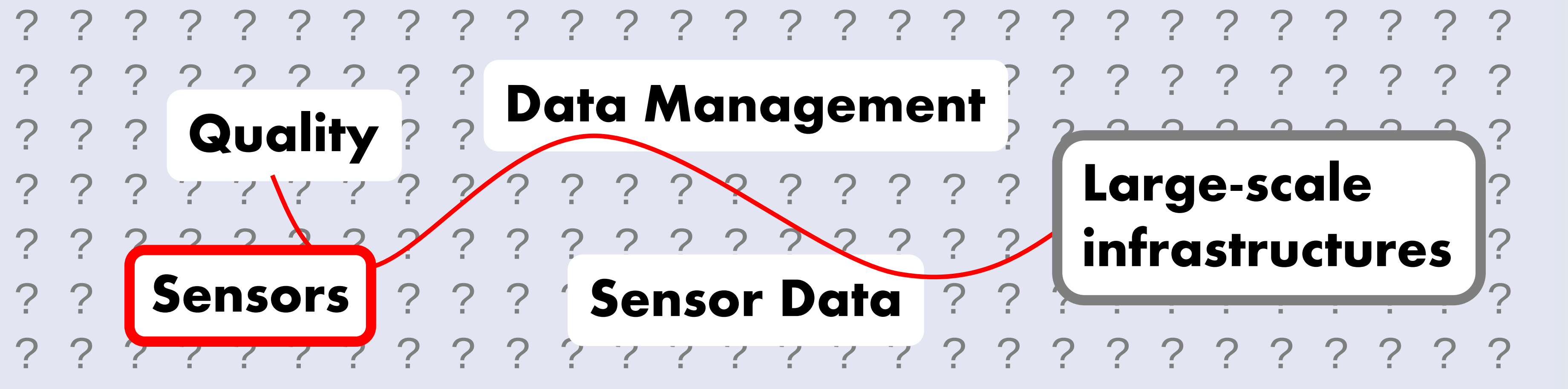
+ magic

= SmartX

Realized by: IoT (Internet of Things)

And similar challenges in pervasive computing, ambient intelligence, physical computing, …

Overview



Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing,
June 25 – June 30, 2017 in Crete, Greece

- Quality of what?
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management?
- Large-scale = Big = Very Large? How big is large? → Bigger than usual
- Infrastructures for what? For data or for real things? → Real things

Sensors

A **sensor** is an electronic component, module, or subsystem whose purpose is to detect events or changes in its environment and send the information to other electronics, frequently a computer processor.

https://en.wikipedia.org/wiki/Sensor

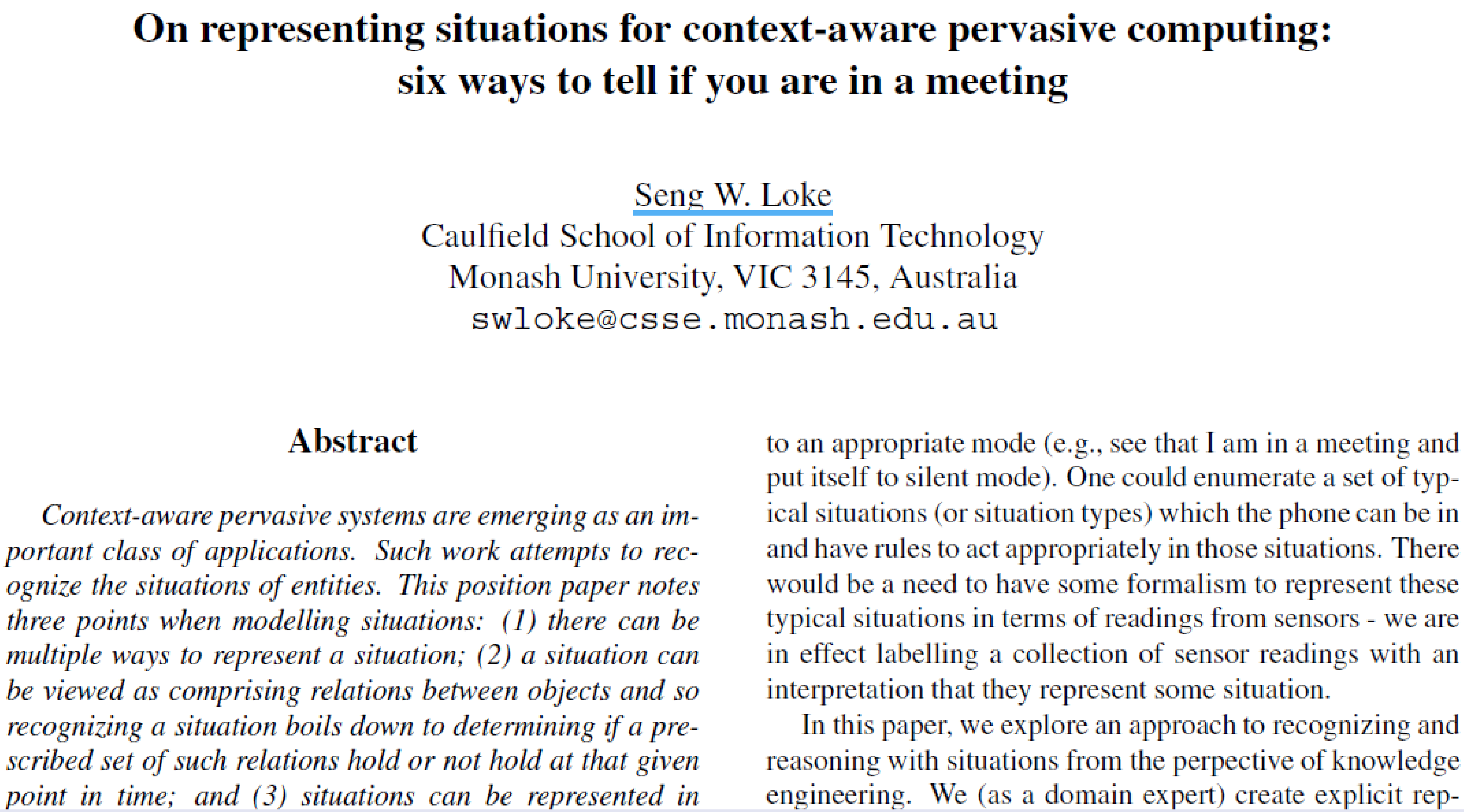
- Technical systems can achieve situational awareness by using data from sensors
- However, sensor data is often …
  - incomplete (not everything can be sensed)
  - late (latency is not always as good as it should be)
  - inaccurate (values are not exact)
  - mobile (sensed by moving systems)
- To make things worse, sensor data needs to be interpreted

  … and interpretations can cause further errors

How to choose a sensor (system)?

- Phenomenon:
  - Physical? (Light, noise, acceleration, radio signals, ..)
  - Chemical? (Substances in gas or fluids)
  - Social? (Behaviour, communication, …)
  - Technical? (Proper operation, …)
- Measurement:
  - Direct or derived?
  - Latency?
- Redundancy:
  - One sensor or many?
  - Same sensors or different?
- Installation:
  - Static or mobile?
  - Wired or wireless?
  - One-hop or multi-hop?
  - Calibration?
- Aging:
  - Battery?
  - Saturation?
  - Re-calibration?
- Cost <-> Quality – Tradeoff!

For a given use case, many solutions …



On representing situations for context-aware pervasive computing: six ways to tell if you are in a meeting

Seng W. Loke
Caulfield School of Information Technology
Monash University, VIC 3145, Australia
swloke@csse.monash.edu.au

Abstract

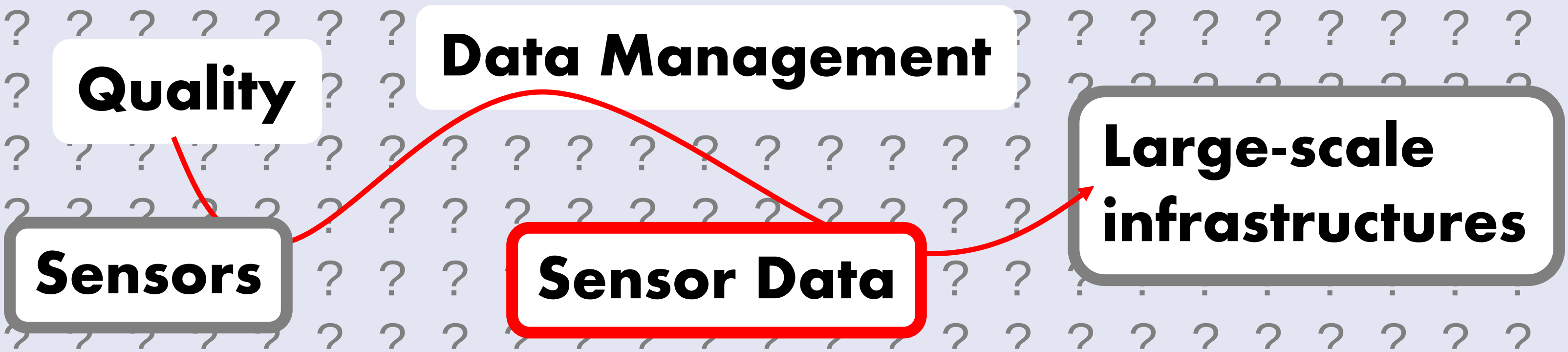
Context-aware pervasive systems are emerging as an important class of applications. Such work attempts to recognize the situations of entities. This position paper notes three points when modelling situations: (1) there can be multiple ways to represent a situation; (2) a situation can be viewed as comprising relations between objects and so recognizing a situation boils down to determining if a prescribed set of such relations hold or not hold at that given point in time; and (3) situations can be represented in

to an appropriate mode (e.g., see that I am in a meeting and put itself to silent mode). One could enumerate a set of typical situations (or situation types) which the phone can be in and have rules to act appropriately in those situations. There would be a need to have some formalism to represent these typical situations in terms of readings from sensors - we are in effect labelling a collection of sensor readings with an interpretation that they represent some situation.

In this paper, we explore an approach to recognizing and reasoning with situations from the perpective of knowledge engineering. We (as a domain expert) create explicit rep-

S. W. Loke, “On representing situations for context-aware pervasive computing: six ways to tell if you are in a meeting,” 2006, pp. 35–39.

Overview



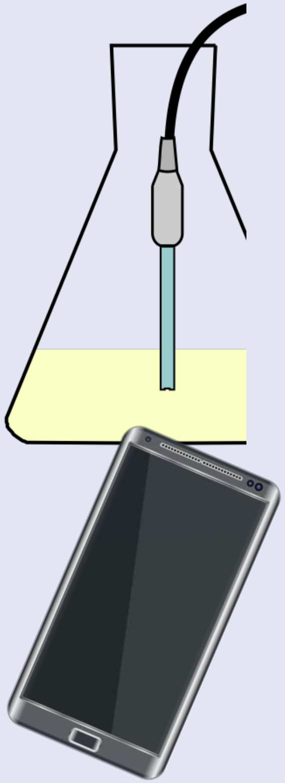
Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing, June 25 – June 30, 2017 in Crete, Greece

- Quality of what?
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management?
- Large-scale = Big = Very Large? How big is large? → Bigger than usual
- Infrastructures for what? For data or for real things? → Real things

Sensor Data

- Sensors implement transfer function:
  - Input: A state of the observed phenomenon
  - Output: A signal (analog or digital) → sensor data
  - Most sensors have a linear transfer function
- Sensitivity of a sensor:
  - How much does the sensor output change when the input changes?
    → Slope of the transfer function
- A sensor system contains:
  - One or many sensors
  - A processing unit (fixed or configurable) to derive data from the sensor signal
  - A communication unit (wired or wireless) to transfer the data to an other system



Types of sensor (system) data

- Format:
  - Structured
    - E.g. (Timestamp, Value), or (Value, Value, Value)
  - Unstructured
    - E.g. image stream (video) or audio stream
  - Semi-structured
    - E.g. photo + DXF meta data (timestamp, location, resolution, …)
- Semantic levels:
  - Raw: just the signal
  - Feature: a typed attribute of a real-world entity, e.g. the location
  - Object: multiple attributes grouped together for an object
  - Situation: a complex sitation was detected
  - → Higher levels are often results of sensor data fusion
- Validity: How long is the sensor value valid?
  1. Only at timestamp
     - if sensor sends with fixed frequency
  2. Fixed until next data comes in
     - if sensor sends when value deviates from last value by threshold
  3. Changing according to model
     - if sensor sends when value deviates from a function of time
     - „dead reckoning“ → often used for moving objects (but can be applied to other phenomena, too)

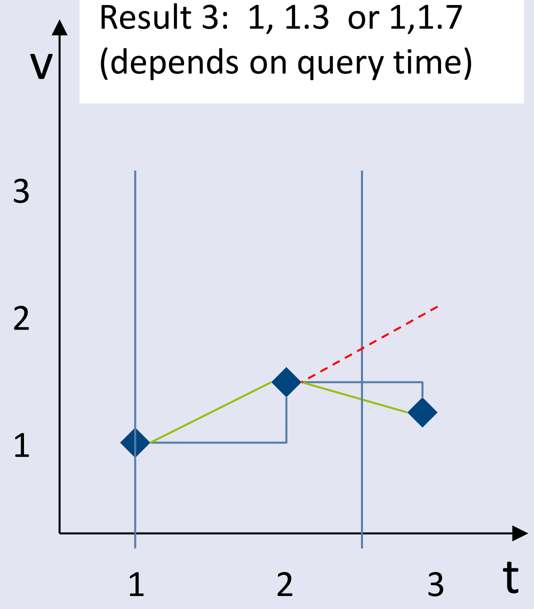
select t, v from sensordata
where t = 1 or t = 2.5

Result 1: 1, NULL

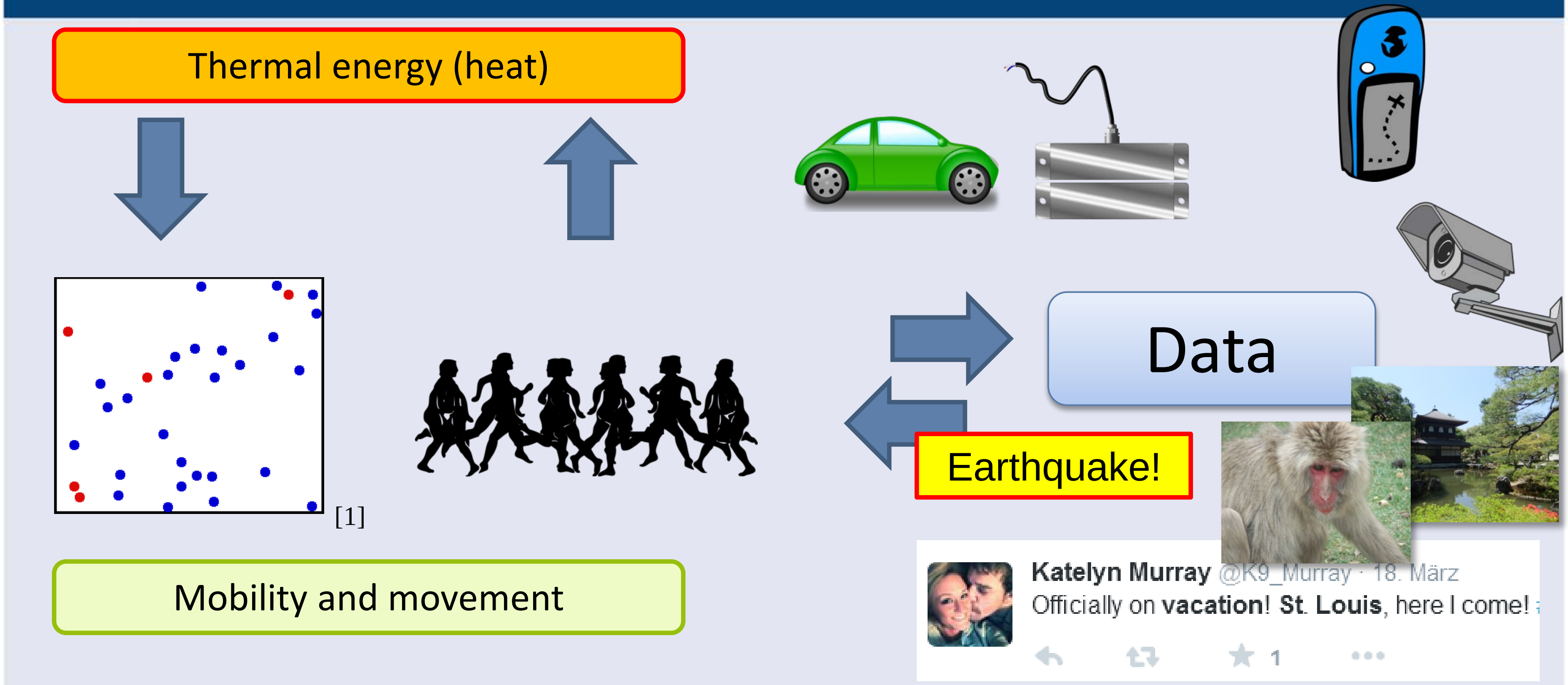
Result 2: 1, 1.5

Result 3: 1, 1.3 or 1,1.7

(depends on query time)



Data and mobility



Thermal energy (heat)

[1]

Mobility and movement

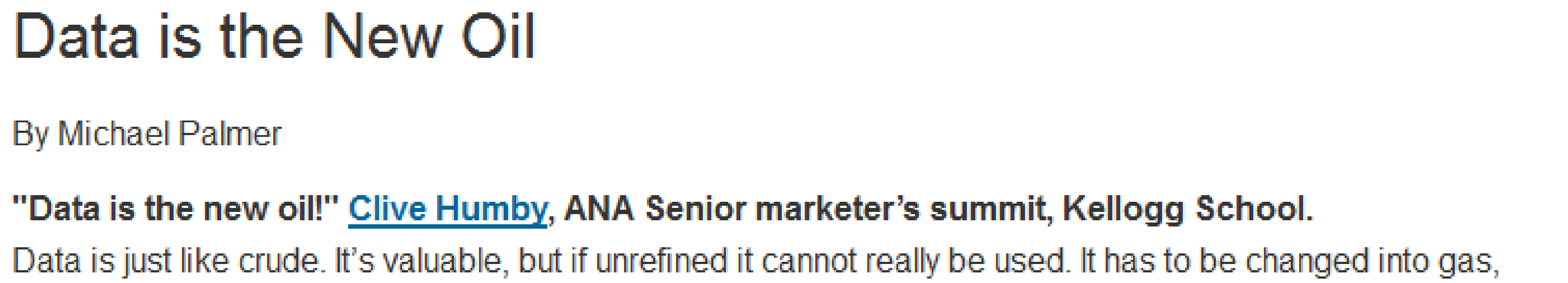
Data

Earthquake!

Katelyn Murray @K9_Murray · 18. März

Officially on vacation! **St. Louis**, here I come!

2006, http://ana.blogs.com/maestros/2006/11/data_is_the_new.html, retrieved 21.3.2015



Data is the New Oil

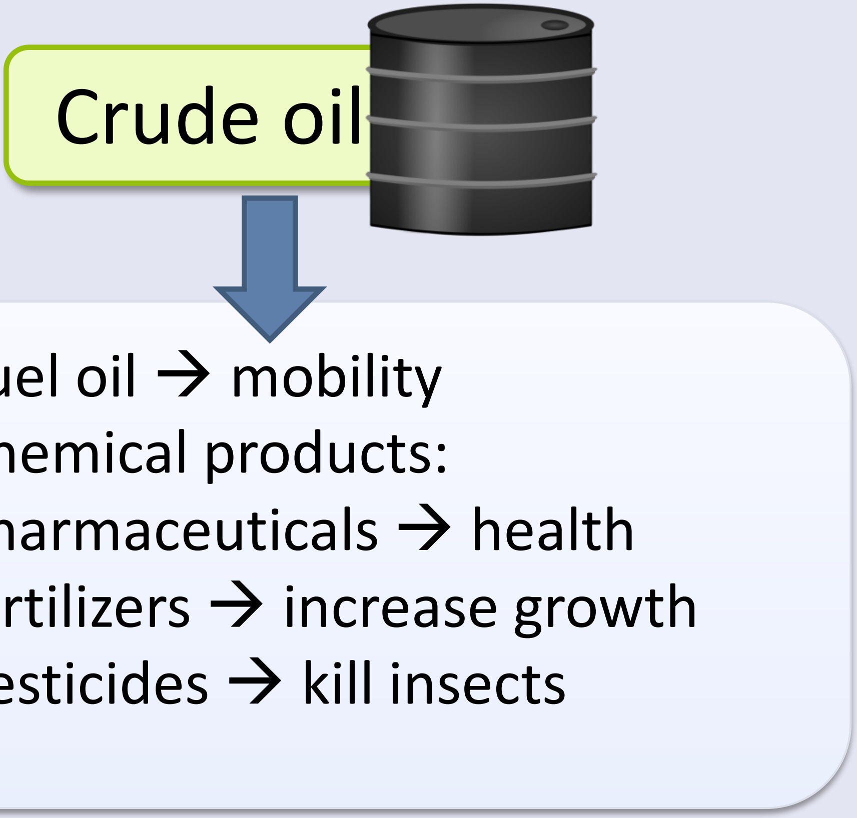
By Michael Palmer

"Data is the new oil!" Clive Humby, ANA Senior marketer's summit, Kellogg School.

Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas,

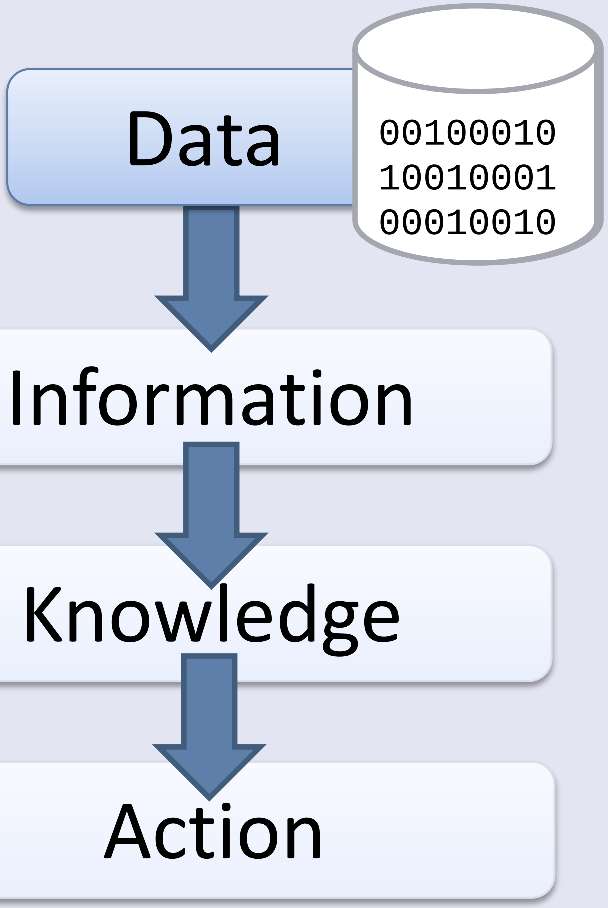
Data is the new crude oil …

… it needs to be refined to be valuable.



Crude oil

- Fuel oil → mobility
- Chemical products:
  pharmaceuticals → health
  fertilizers → increase growth
  pesticides → kill insects
- …



Data

```
00100010
10010001
00010010
```

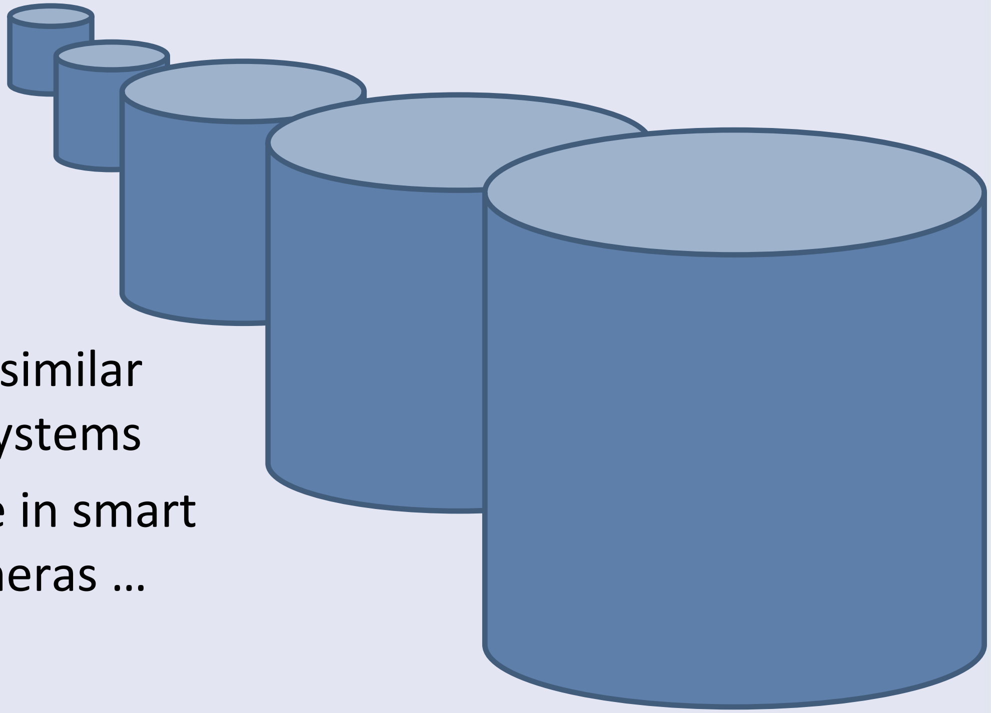
Information

Knowledge

Action

DATA and mobility

- Googles self-driving car: nearly 1MB data per second[1]
  - Per day: 85 GB
  - Per year: ~30 TB
  - If 10% of the cars would be like this, or 50%, or …
    (> 1 Billion cars on the world)
  - … not only by self-driving cars, similar for advanced driver assistant systems
  - … plus data from infrastructure in smart cities, like induction loops, cameras …



→Big Data!

[1] Bill Gross, Founder and CEO of Idealab
https://www.linkedin.com/today/post/article/20130502024505-9947747-google-s-self-driving-car-gathers-nearly-1-gb-per-second

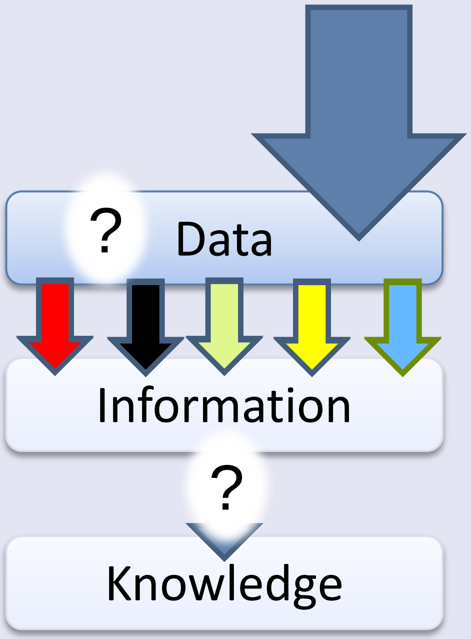
Mobility of the future?

Rush Hour by Fernando Livschitz, Black Sheep films
http://vimeo.com/106226560

Big Data Challenges

- Many definitions, often by a numer of 3-5 "V" challenges:

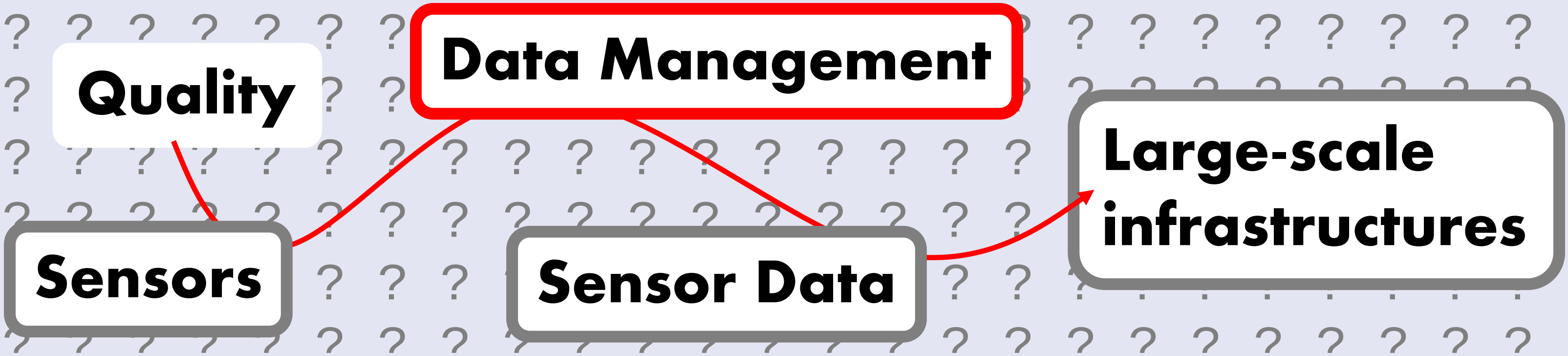
Volume	A lot of data (amount varies)
Variety	Data differs in structure
Variability	Structure changes
Velocity	Many updates
Veracity	Unclear source or quality



Not in list of challenges: Pricacy

(analysis of sensible data, how to adhere to legal / societal constraints?)

Overview



Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing, June 25 – June 30, 2017 in Crete, Greece

- Quality of what?
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management?
- Large-scale = Big = Very Large? How big is large? → Bigger than usual
- Infrastructures for what? For data or for real things? → Real things

Data management

- CRUD: create, read, update, delete
  - by ID: “get measurement value with ID 47111981231“

    → one value
  - by query:
    SELECT location, avg(temperature) FROM values WHERE sensor_location=‘Crete’ group by location

    → a set of average temperatures at locations
  - by search: „Knossos“

    → all data sets that contain that term
- Transactions: uninterupted sequence of operations
  - „All or nothing“
  - e.g.: financial transaction
    - money should be here or there
- In addition: Pub/sub or Continuous Queries

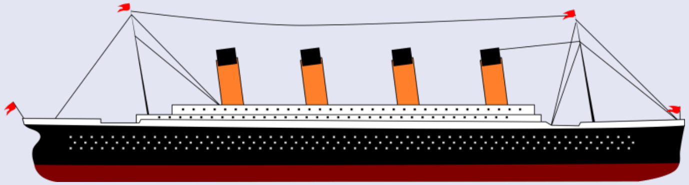
Requirements for data management

SQL systems

1. Data management
2. Scalability
3. Heterogeneity
4. Efficiency
5. Persistence
6. Reliability
7. Consistency
8. Non-redundancy
9. Multi-User support

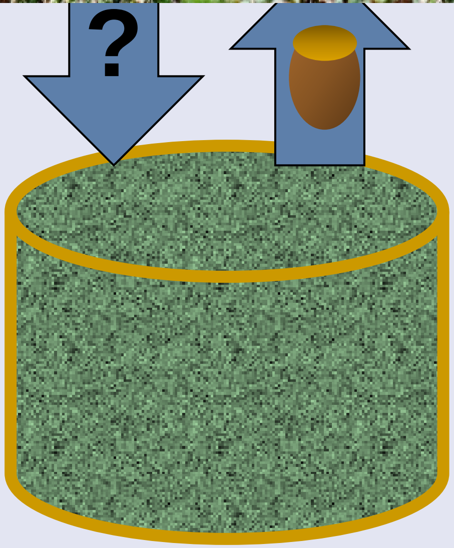
NOSQL systems

1. Data structure complexity
2. Schema independence
3. Sparseness
4. Self-descriptiveness
5. Variability
6. Scalability
7. Volume



… which properties are you willing to relax / neglect?

Bild: elli60 / pixelio.de



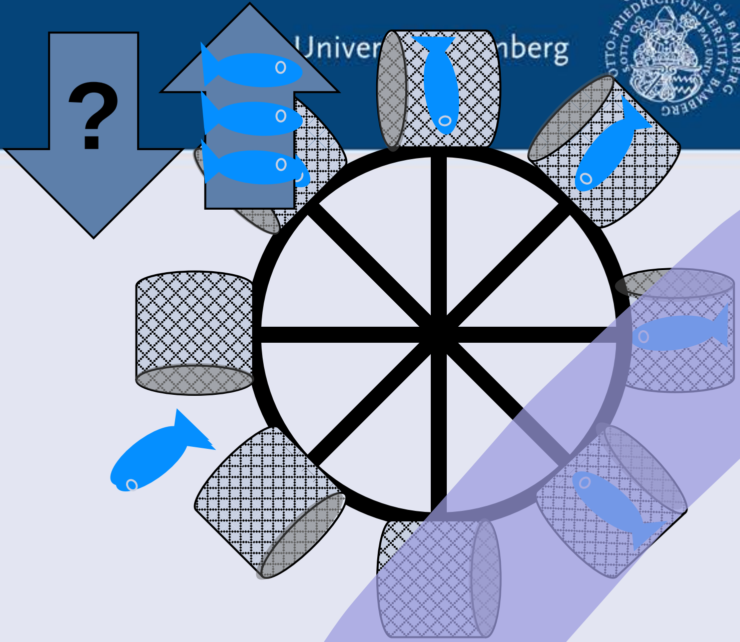


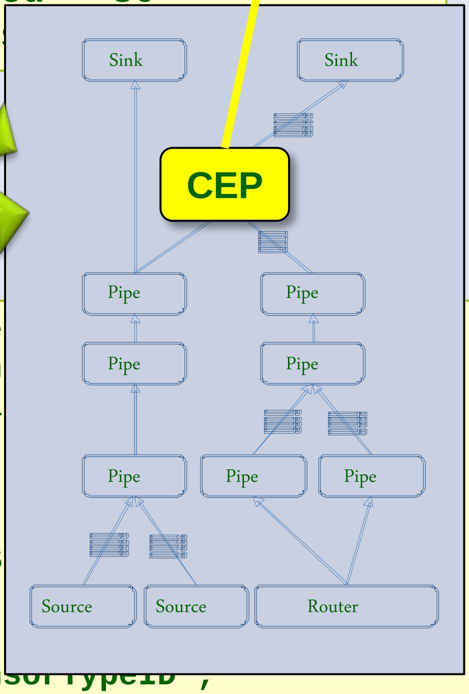
Bild: Ronny Senst / pixelio.de

Features of Data Stream Management Systems

- Programming Abstraction
  - declarative: query
  - functional: flow graph
  - → enables optimizations
  - → better maintanance of systems
  - → using a DSMS on data streams is like using a DBMS instead of files
- Easy to combine with complex event processing (CEP)
- Parallel execution of operators in graph → no shared memory
- Data streams can be unbounded:
  - issues with sorting, joins, aggregation
  - → approximate answers
  - → window semantics

```
SELECT ego.pos
RANGE 10 secon
radar RANGE 15
WHERE ego.speed > 30 AND
radar.speed > 30
AND s2.pos
```

Some DSMS provide CEP operators



```
stream<uint64 curre
sensorId, uint8 sen
measureValue1, floa
float64 distanceV,
uint8 speed, uint8
direction, uint64 t
Aggregate(HigherGPS

{window HigherGPS :
param groupBy : sensorTypeID ;
output MapGPS : avgSpeed = Average(speed)
;}
```

Data stream management and Big Data

- More "velocity", less "volume"
- Direct processing
    - Online, (hard/soft) real time, "right time"
- More information, less data
    - Enrichment of data streams
        - E.g., product information for an RFID tag
    - Interpretation and reasoning
        - E.g., classification ("this is a car")
    - Data cleansing
        - Remove redundancy, anomaly detection
- Online quality assessment
- Enables built-in privacy methods
    - Online pseudonmization and anonymization
    - Data economy
    - Certify and/or publish your query plans

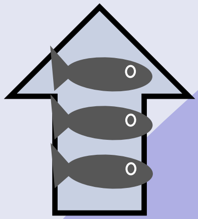
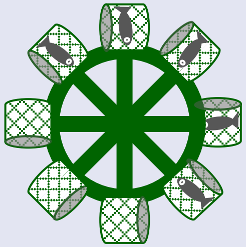
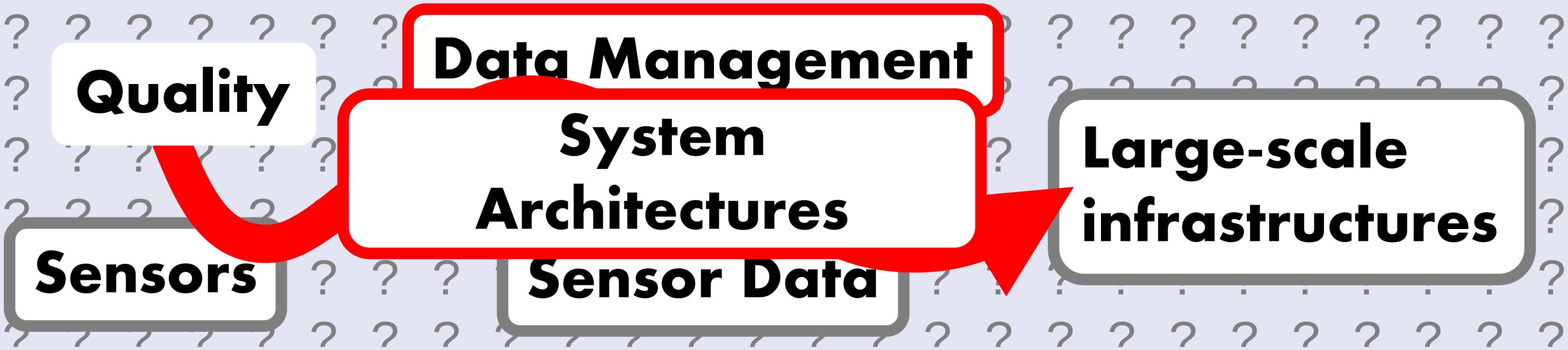


Bild: Ronny Senst / pixelio.de



Overview



Quality

Data Management

System Architectures

Sensors

Sensor Data

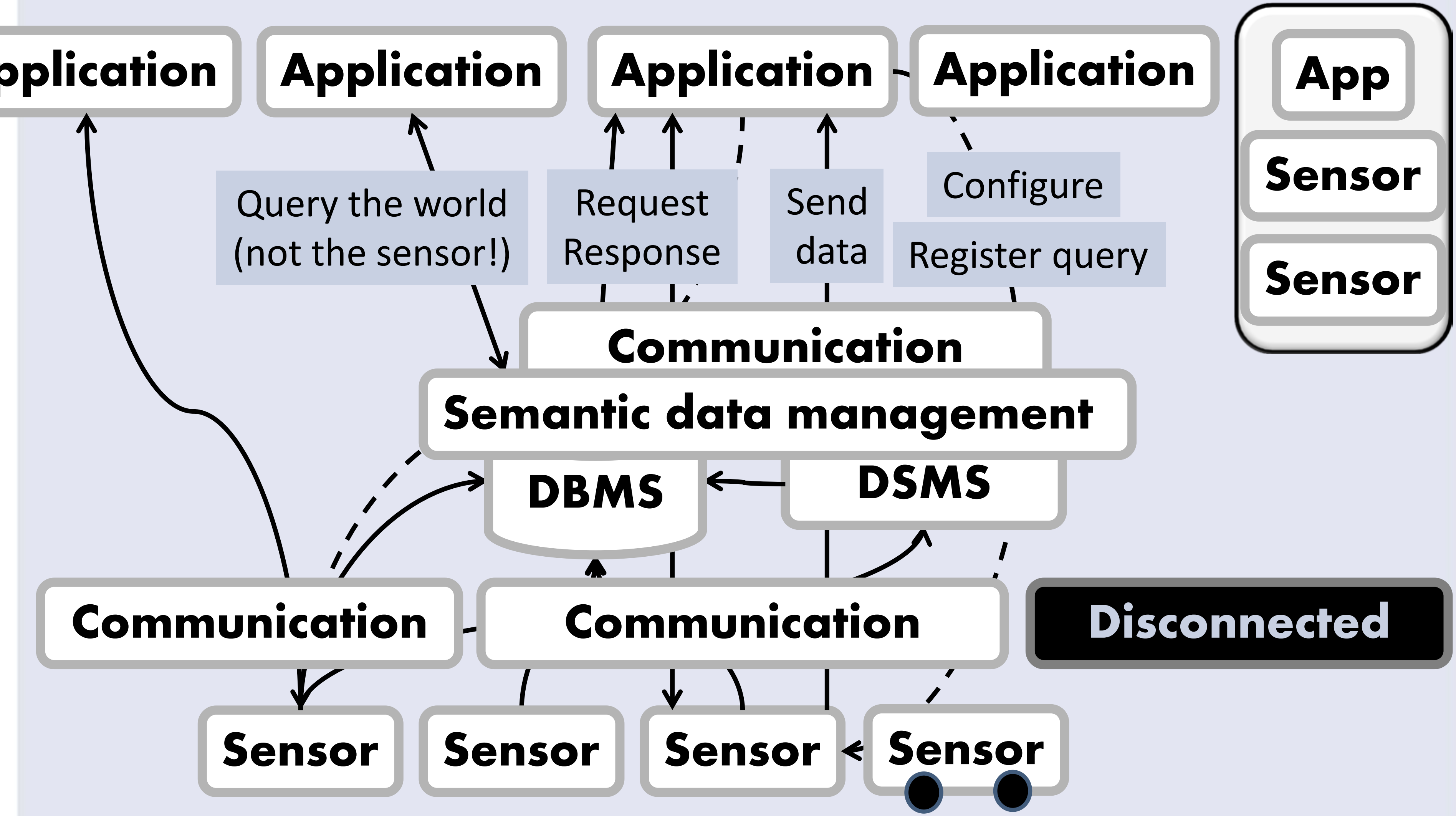
Large-scale infrastructures

Quality matters: Sensor Data Management for Large-scale Infrastructures

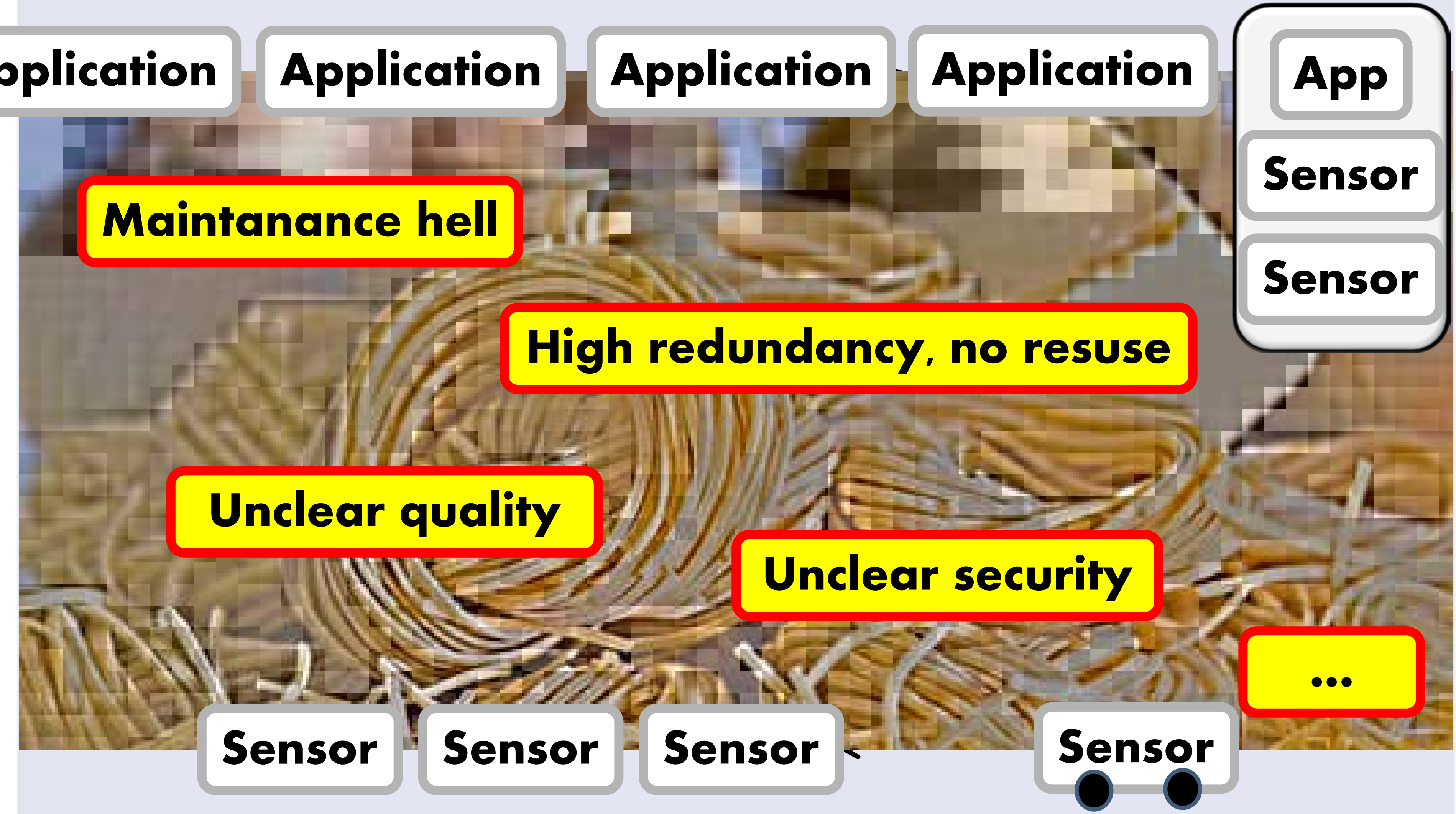
11th Symposium and Summer School On Service-Oriented Computing, June 25 – June 30, 2017 in Crete, Greece

- Quality of what? → The sensor data
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management?
- Large-scale = Big = Very Large? How big is large? → Bigger than usual
- Infrastructures for what? For data or for real things? → Real things

Evolution of a sensor-based system



Evolution of a sensor-based system



pplication

Application

Application

Application

App

Sensor

Sensor

Maintanance hell

High redundancy, no resuse

Unclear quality

Unclear security

…

Sensor

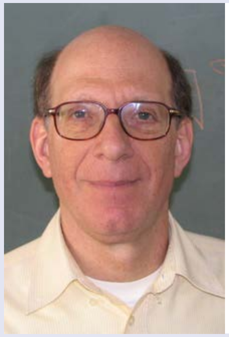
Sensor

Sensor

Sensor

Better architectures? *AAS?

- Shouldn‘t we help with a standard?
- From a systematic mapping study [1] on 35 studies on IoT and Cloud:
  - 15 provide Software as a Service (SaaS)
  - 13 provide Platform as a Service (PaaS)
  - 10 provide Infrastructure as a Service (IaaS)
  - 1 provides Network as a Service (NaaS)
  - 2 provide Sensing as a Service (SaaS)
  - 2 provide Sensing and Actuation as a Service (SAaaS)
  - 1 provides Smart Object as a Service (SOaaS)



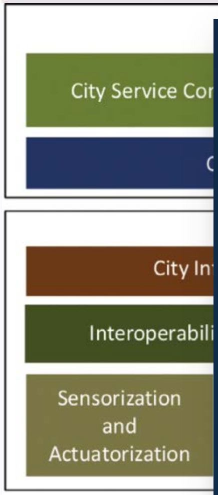
> „The nice thing about standards is that you have so many to choose from.“
>
> A. Tanenbaum, *Computer Networks*, 2nd ed., p. 254

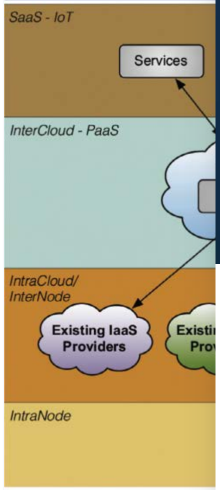
Application Domains:

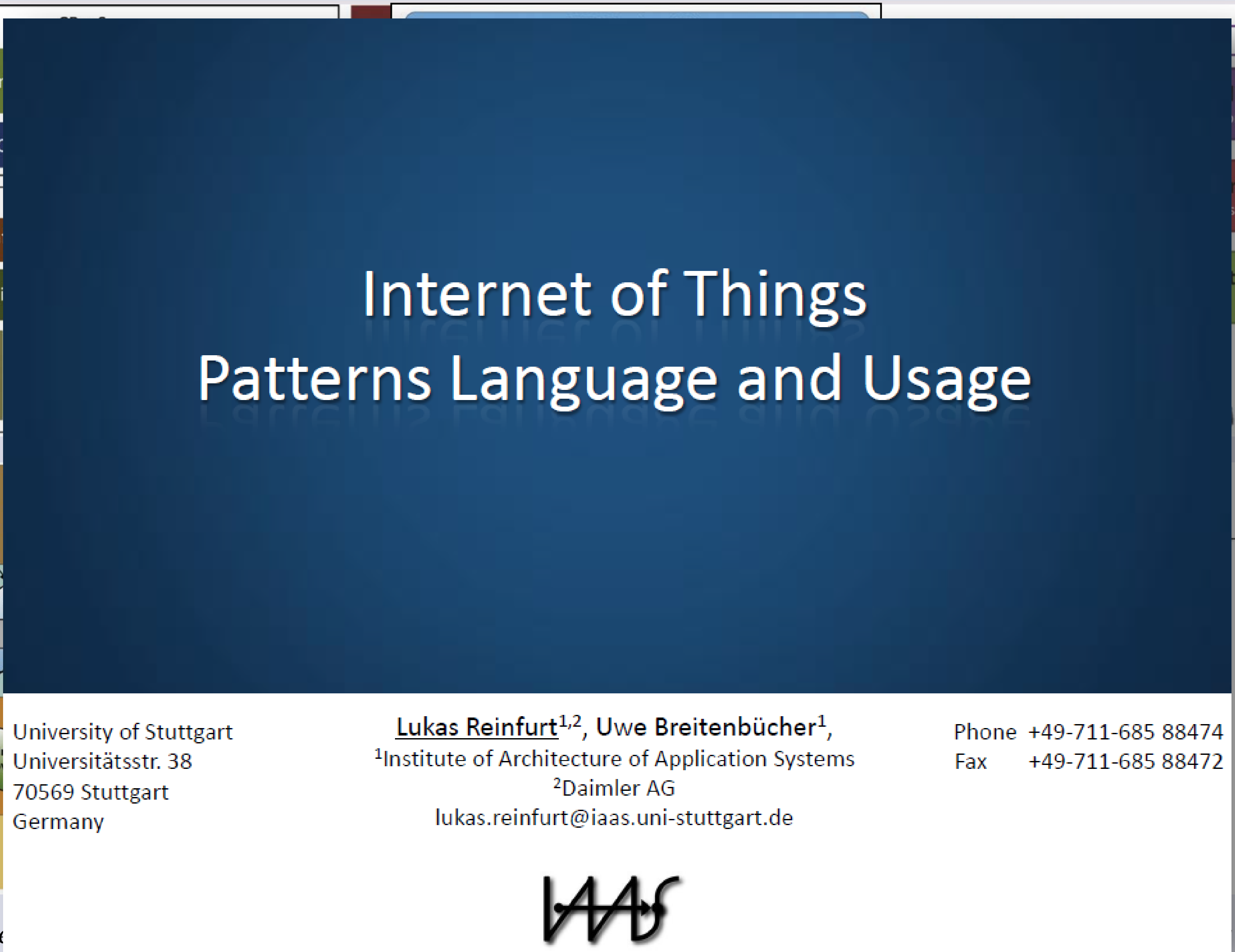
- Healthcare (4)
- Smart cities (2)
- Ambient Assisted Living (1)
- Smart homes (1)
- Mobile applications (2)
- Intelligent business services (1)
- Supply chain management (1)

[1] E. Cavalcante *et al.*, “On the interplay of Internet of Things and Cloud Computing: A systematic mapping study,” *Computer Communications*, vol. 89–90, pp. 17–33, Sep. 2016.

To understand and assess this zoo, Pattern to the Help!







Internet of Things Patterns Language and Usage

University of Stuttgart
Universitätsstr. 38
70569 Stuttgart
Germany

Lukas Reinfurt[1,2], Uwe Breitenbücher[1],
[1]Institute of Architecture of Application Systems
[2]Daimler AG
lukas.reinfurt@iaas.uni-stuttgart.de

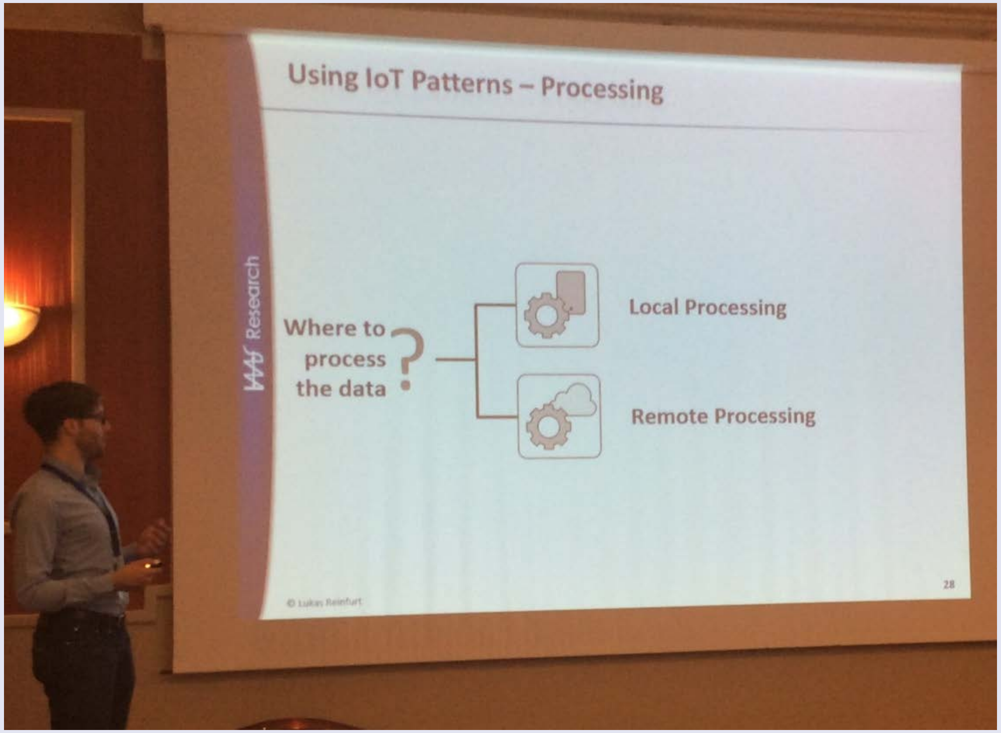
Phone +49-711-685 88474
Fax +49-711-685 88472

E. Cavalcante … Communications, vol. 89–90, pp. 17–33, Sep. 2016.

Sensor to Cloud?

Internet of Things
Patterns Language and Usage

Lukas Reinfurt, SummerSoc 2017



Using IoT Patterns – Processing

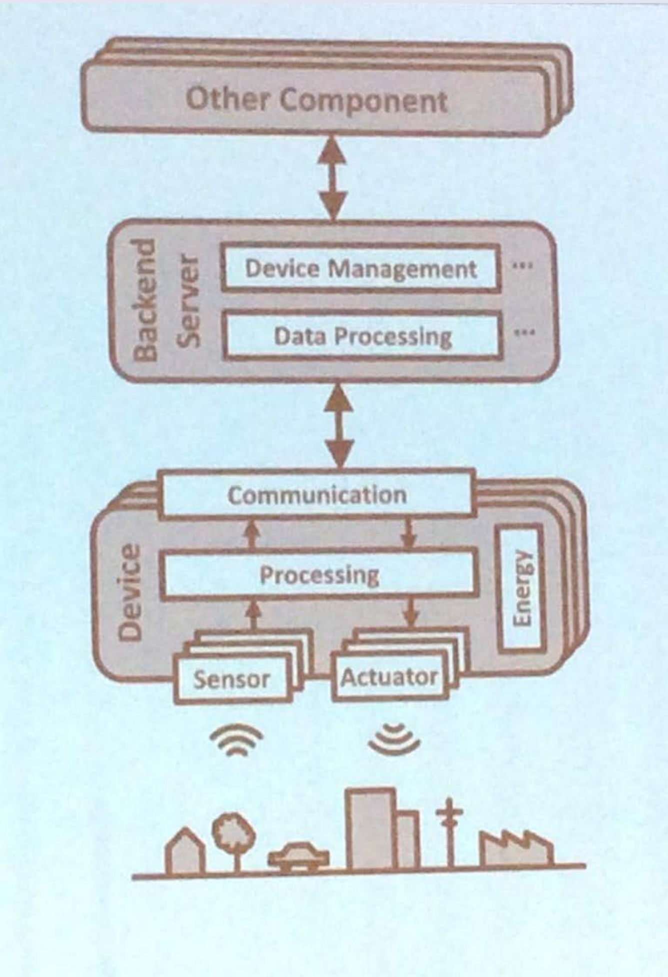
Where to process the data?

Local Processing

Remote Processing

© Lukas Reinfurt

28



Next step in architectures: Fog Computing

- Sending all raw sensor data to the cloud cannot be the final solution:
  - Bandwith
  - Energy comsumption
    - (computing needs less than communication)
  - Application needs, e.g., privacy
- Edge computing:
  - Move the processing to the edge of the network
- Fog computing:
  - Utilize further processing nodes on the way

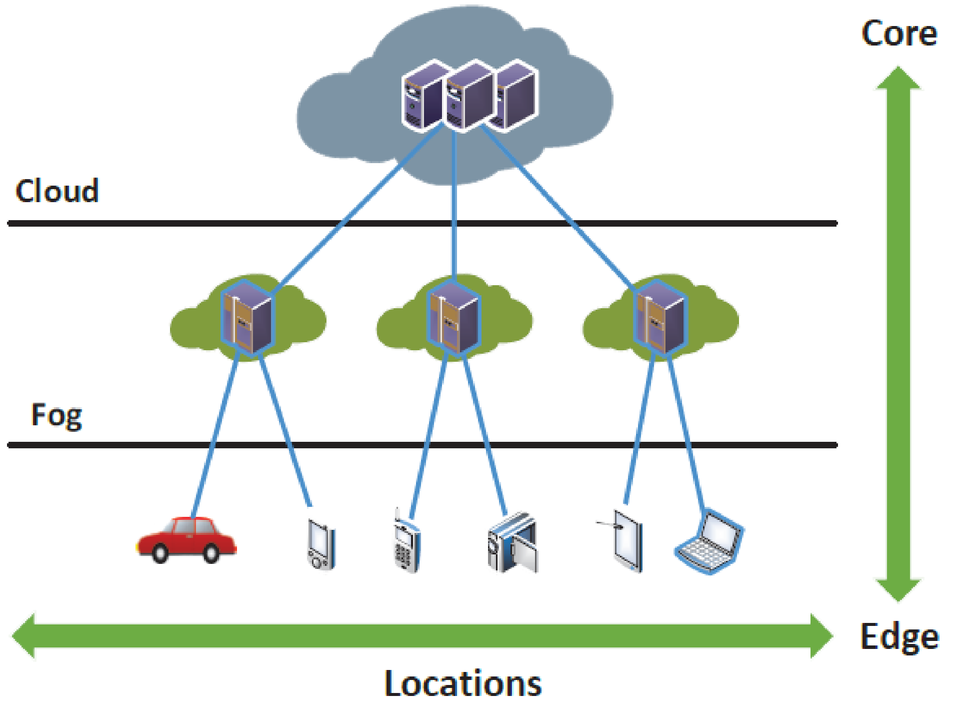


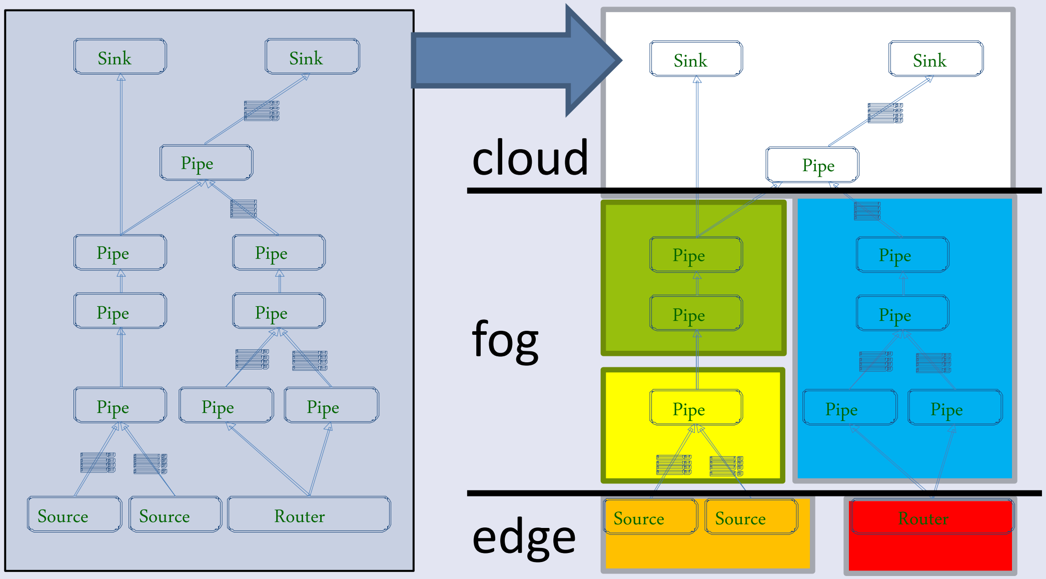
Fig. 1. Fog between edge and cloud.

[1]

[1] I. Stojmenovic and S. Wen, "The Fog Computing Paradigm: Scenarios and Security Issues," 2014, pp. 1–8.

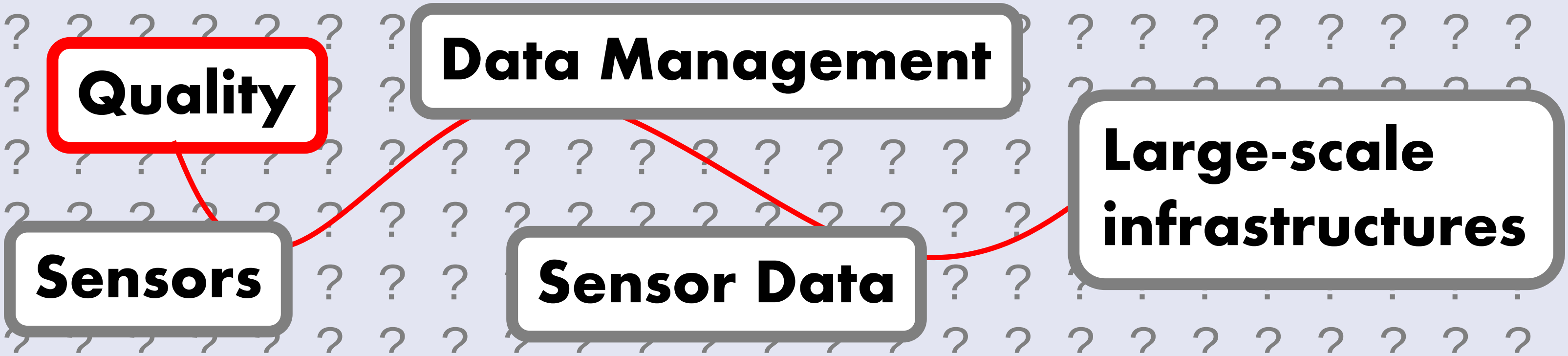
Fog computing and distributed data stream management

- Data stream management:
  - Provides a higher-level abstraction to stream-based data processing
- Distributed stream management:
  - Distributes the execution of the data stream processing over nodes
  - Finds an optimized query execution plan
  - Can adapt to changing situations and migrate the execution



We can use distributed DSMS to implement sensor data management in a fog-computing architecture

Overview



Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing, June 25 – June 30, 2017 in Crete, Greece

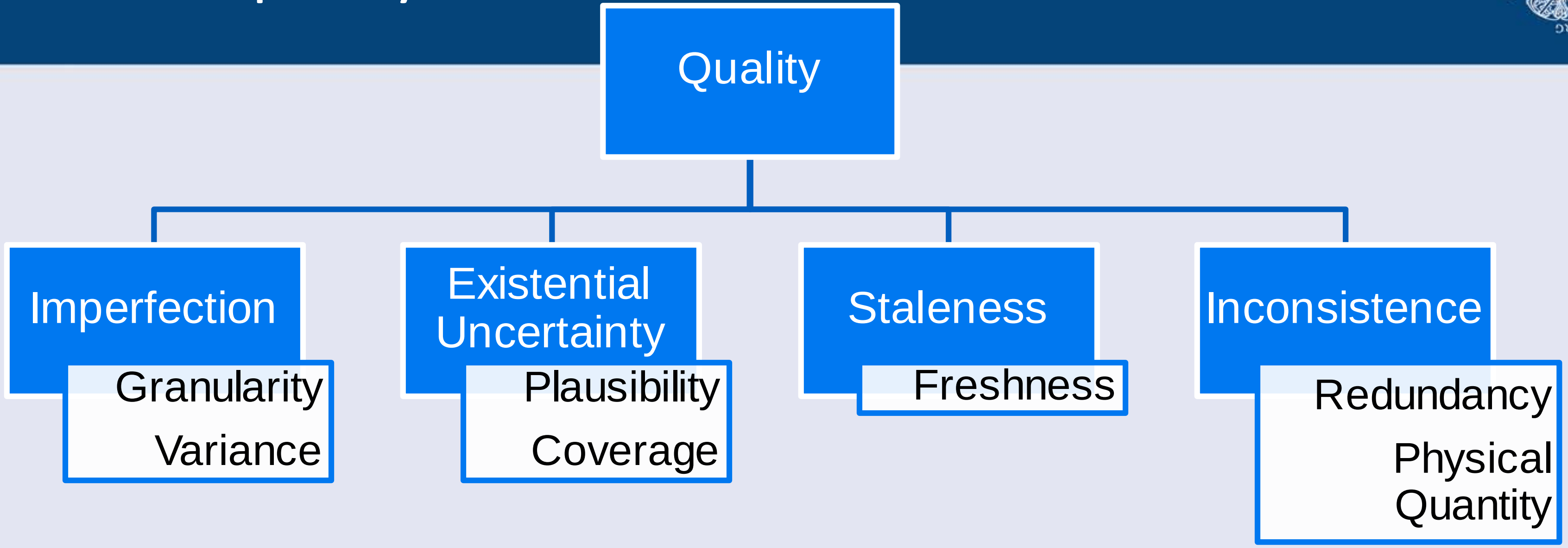
- Quality of what? → Of sensor data, if relevant to application
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management? → All of it
- Large-scale = Big = Very Large? How big is large? → Bigger than usual
- Infrastructures for what? For data or for real things? → Real things

Some common quality issues

- Data source
  - Measurement method, e.g. low frequency of sensor for fast moving objects
  - Environment, e.g., temperature too high for good measurements
  - …
- Data processing
  - Wrong training data for classifier
  - Over-simplified models or missing concepts
  - Not enough input data for algorithm
  - Stale models (due to concept drift)
  - …
- Some can be detected after installation of system, some occur later

→ Decisions based on inpresise data

Relevant quality dimensions



- Imperfection: some information is there, but it‘s inaccurate or not detailed enough
- Existential uncertainty: you have information about something (e.g., an object), but you are not sure whether it exists
- Staleness: your information might be outdated
- Inconsistency: you have redundant information (overlapping in time/space/content), and it contradicts each other

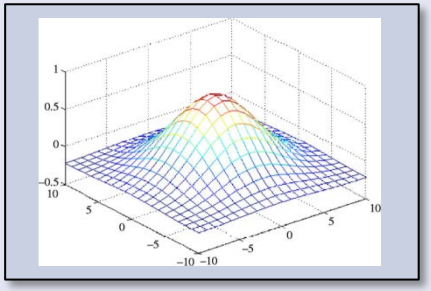
Features for a quality-aware DSMS

Goal: programming abstractions for dealing with non-perfect data

Approach:

1. develop unified data model to represent data quality
2. consider data quality in operators

→ data management can attach combined quality metadata to result



type (probability)
bicycle (0.8)
pedestrian (0.1)
other (0.1)

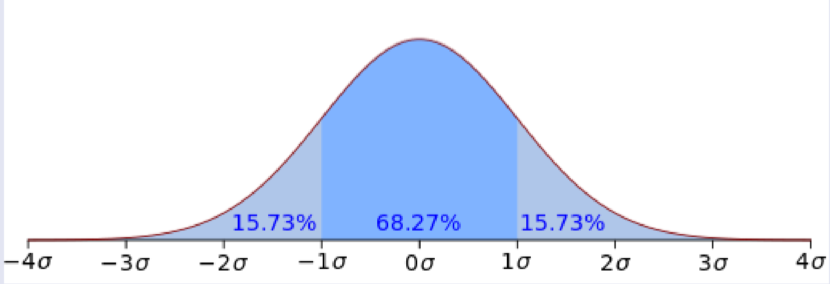
- How to determine data quality and correlations?
  - given by data source / sensor (e.g., accuracy)
  - given by algorithm (e.g., confidence)
  - learned by observation (requires redundancy)

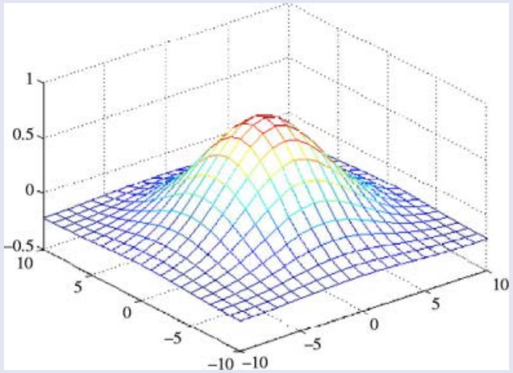
→ store in sensor relationship model

Quality Matters: Supporting Quality-aware Pervasive Applications by Probabilistic Data Stream Management, DEBS2014

Representation of accuracy

- Discrete, multiple values with existence probability:
  [car, 0.3 ; bike, 0.5 ; other, 0.2]
  - → Leads to multiple possible worlds
- Continuous values
  - Probability density functions, may be correlated (→ covariance)





- And what to do with it?

Multiple possible worlds

object A

type (probability)
bicycle (0.8)
pedestrian (0.1)
other (0.1)

object B

type (probability)
bicycle (0.3)
pedestrian (0.7)
other (0.0)

world 1 **0,24**
object A: bicycle
object B: bicycle

world 2 **0,56**
object A: bicycle
object B: pedestrian

world 3
object A: bicycle
object B: other

worl
ob
obj

0,07

world 6
object A: pedestrian
object B: other

world 7
object A: other
object B: bicycle

0,07
object A: other
object B: pedestrian

world 9
object A: other
object B: other

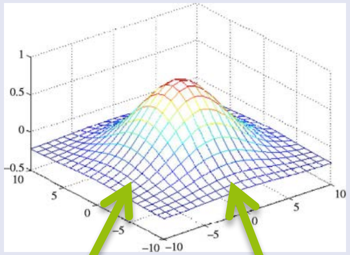
but only if probabilities are independent!

Unified data model

- Extend stream data types to allow for:
  - tuples with existential probabilities (e.g., events)
  - discrete attributes with multiple values (e.g., classification results)
  - continuous attributes with probabilistic distributions (e.g., temperature)
  - continuous attributes with conditional distributions (e.g., location)
- Logical view (stream schema):

ID	**Type**	**Location.X**	**Location.Y**
discrete value	discrete prob. value	correlated continuous probabilistic values	

- Physical view (stream data):
  - tuple meta data
  - payload (data) with multi-values
  - distributions and correlations

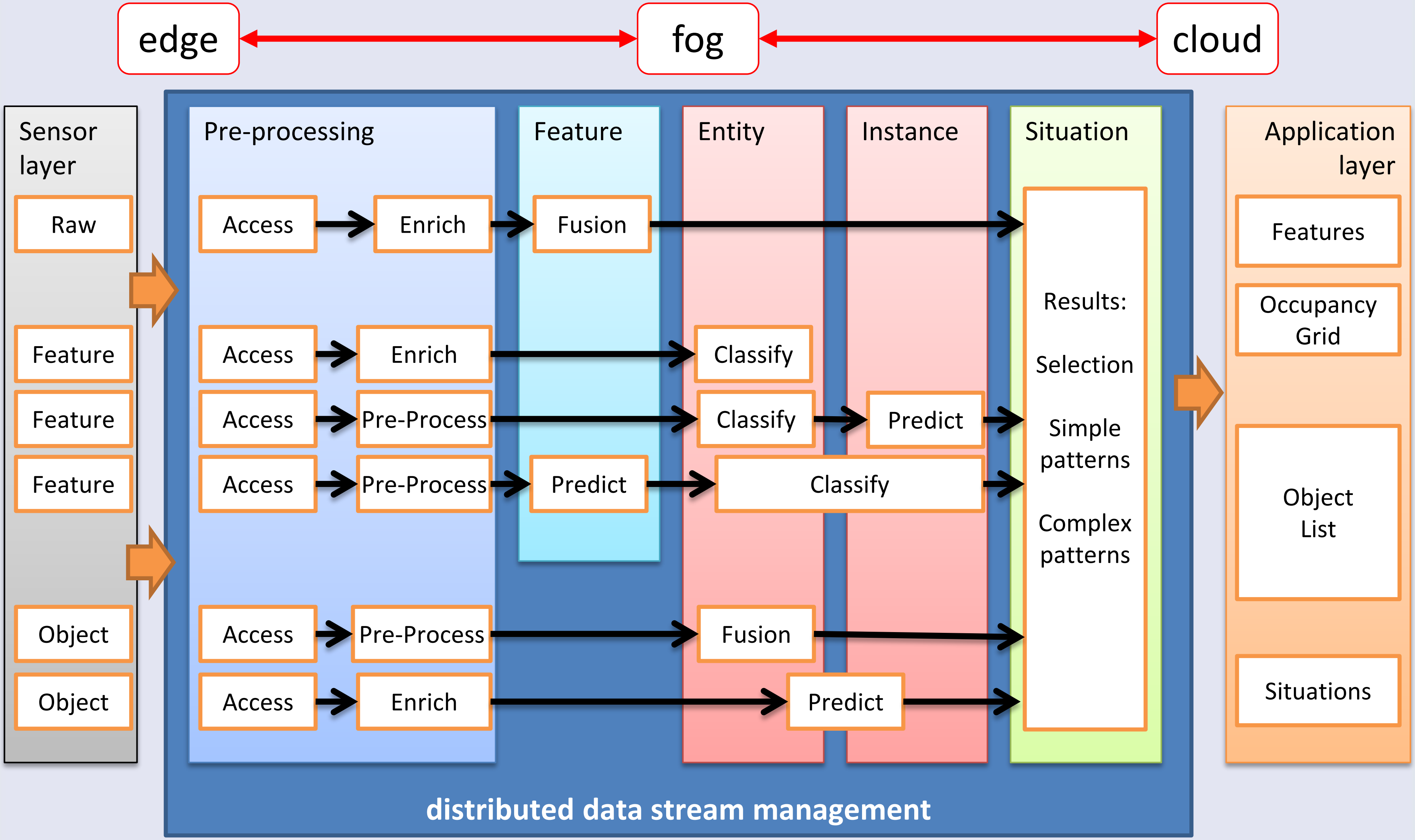
Timestamp: 15:14		**Existence: 0.9**	
Distribution[0]:			
4815162342	**<car, 30%> <bike, 50%> <other, 20%>**	**(0 : 0)**	**(0 : 1)**

Consider data quality in processing

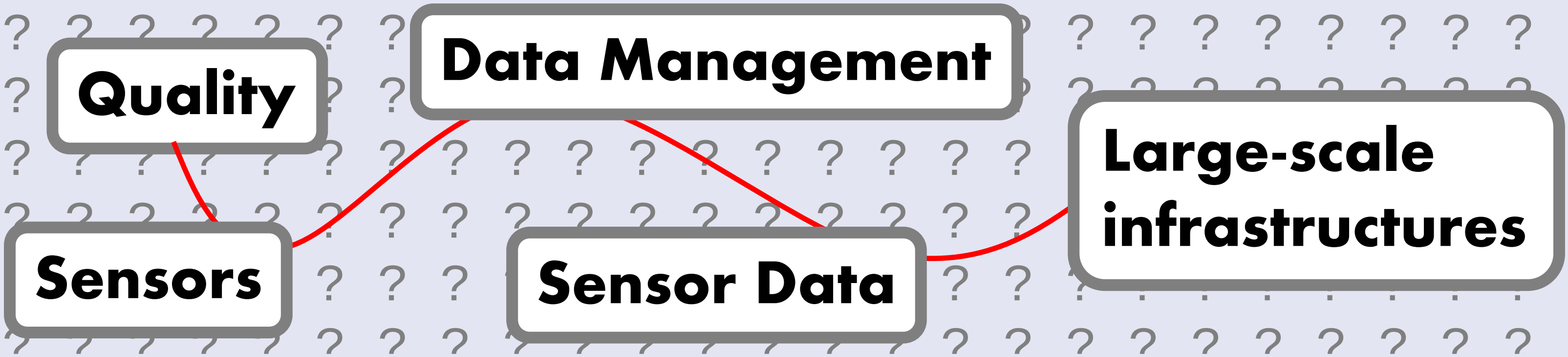
- System resources are limited
- Probabilistic data requires additional processing
  - Integration of multivariate probability distributions
  - Processing of multiple possible worlds
- Ideas:
  - Change result accuracy depending on available resources
    - Filtering of data with low probability ("quality shedding")
    - Decrease number of samples for result estimation
  - Quality-aware query rewriting
    - Quality-based optimizations
  - Include a priori knowledge (sensor relationship model) in data stream queries
    - e.g., sensor observing sensors [KN12]
  - Quality monitoring
    - for service level agreements, minimum quality requirements
- Integrate with multi-layer processing

Multi-Layer Semantic Proceessing
based on JDL fusion levels

Universität Bamberg



Overview



Quality matters: Sensor Data Management for Large-scale Infrastructures

11th Symposium and Summer School On Service-Oriented Computing,
June 25 – June 30, 2017 in Crete, Greece

- Quality of what? → Of sensor data, if relevant to application
- (Sensor Data) Management or Sensor (Data Management)? Sensor Management? → All of it
- Large-scale = Big = Very Large? How big is large? → Bigger than usual
- Infrastructures for what? For data or for real things? → Real things

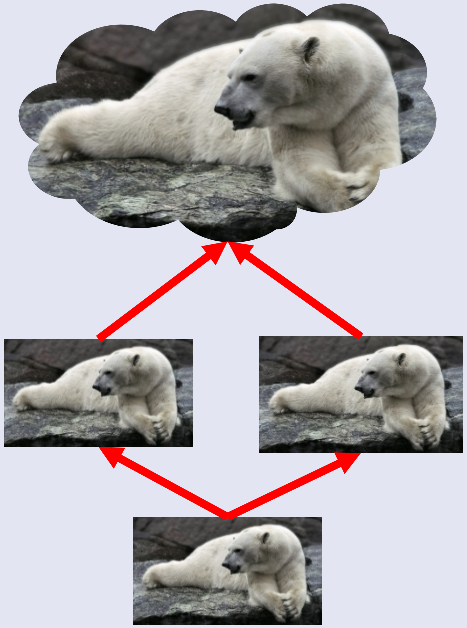
Summary and outlook

- Monitoring large-scale infrastructures with sensors can lead to large-scale sensor data management systems

- Issues to solve:
  - The „V“ challenges → maybe you do not need to store everything in the cloud
  - The „P“ challenge → maybe you can anonymize or aggregate at the edge or in the fog
  - The „Q“ challenge → know thy quality, before and during operation

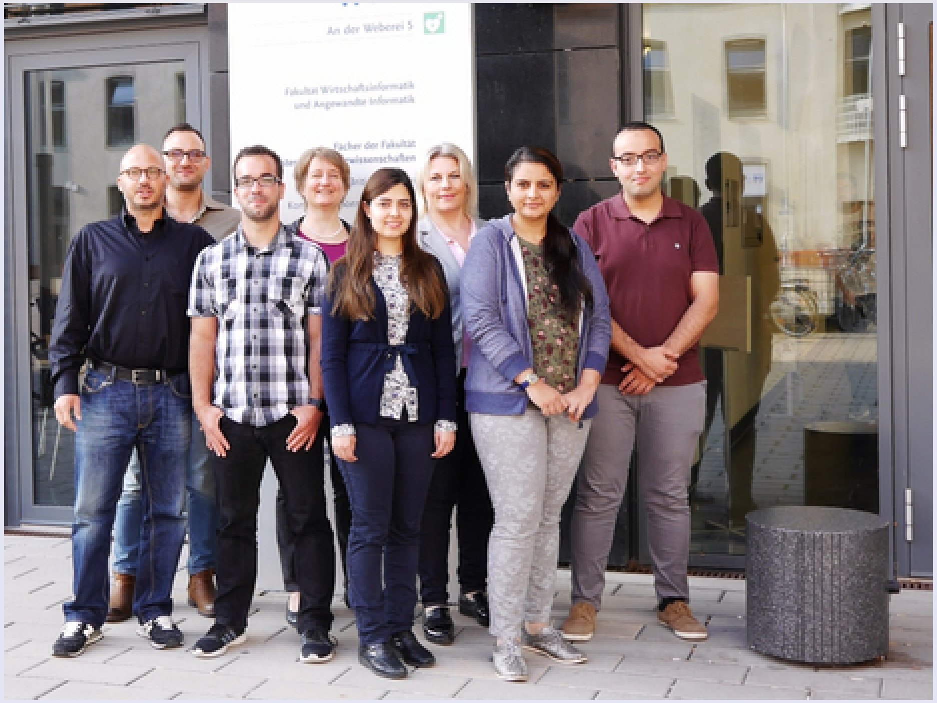
- IoT platforms can help, but are only slowly moving towards fog architectures („who owns the data?“)

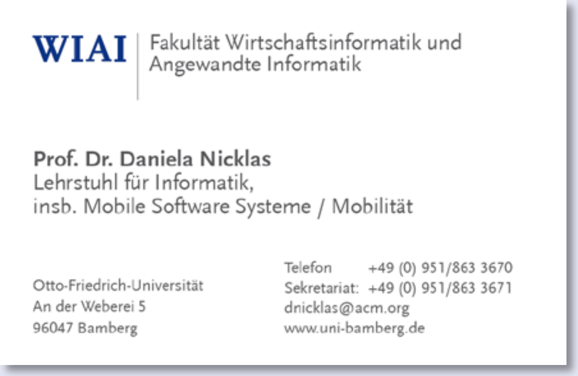
  → Distributed data stream processing revisited?



Ronny Senst / pixelio.de

Thank‘s for all the fish!





WIAI | Fakultät Wirtschaftsinformatik und Angewandte Informatik

Prof. Dr. Daniela Nicklas
Lehrstuhl für Informatik,
insb. Mobile Software Systeme / Mobilität

Otto-Friedrich-Universität
An der Weberei 5
96047 Bamberg

Telefon +49 (0) 951/863 3670
Sekretariat: +49 (0) 951/863 3671
dnicklas@acm.org
www.uni-bamberg.de

Bild: Ronny Senst / pixelio.de