POLITECNICO
MILANO 1863

Processes and quality of data

Barbara Pernici

CAiSE, June 17, 2016

Outline

Why

- Implications of poor quality

What

- Processes and data
- Data quality dimensions
- Evaluations

How

- A control flow perspective
- Assessment
- Improvement – data quality blocks
- Repair

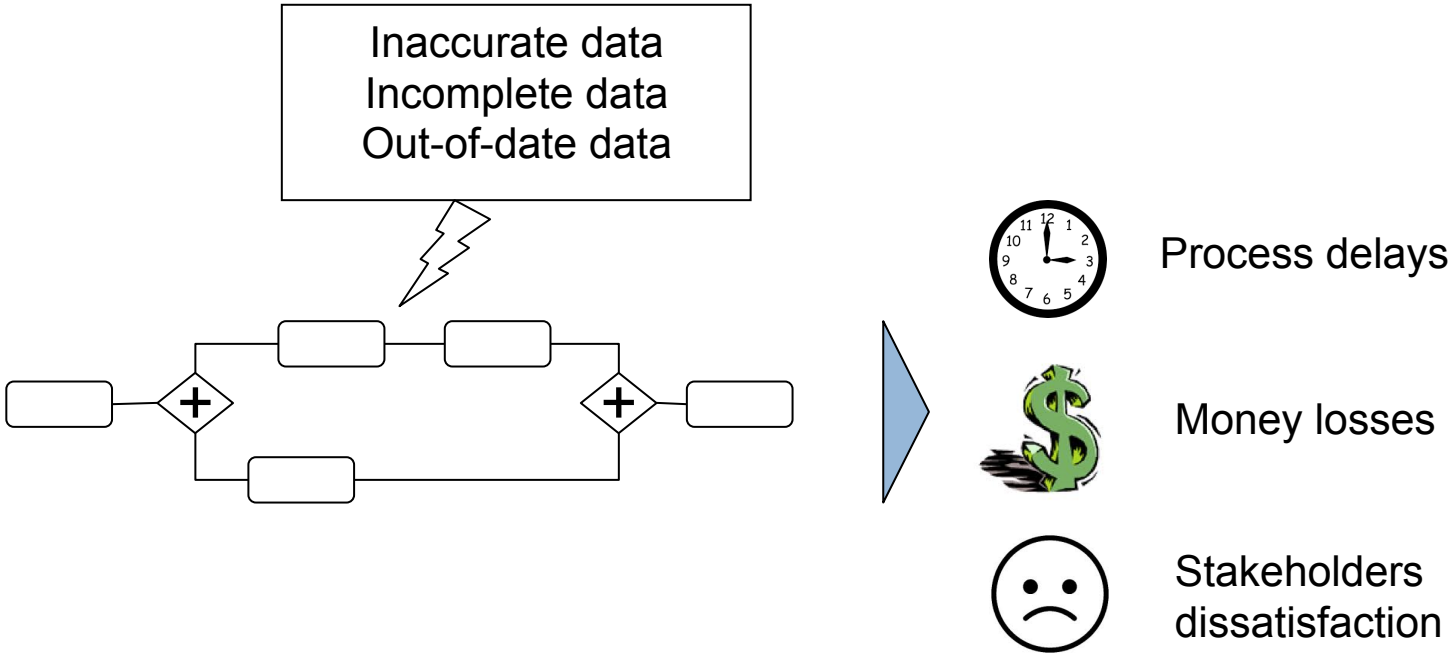
Other perspectives

Open issues

WHY

General statement

Poor quality negatively affects the efficiency and effectiveness of business processes



Where are we coming from?

BPM

DQ

ISE

WS

Implications of poor data quality in processes

- Wrong outputs
- Different courses of action
- Wrong analyses
- Failures
- Delays
- No effect …

logs

Typical causes

Input data

- Wrong / missing
- Access to external sources
- Received messages

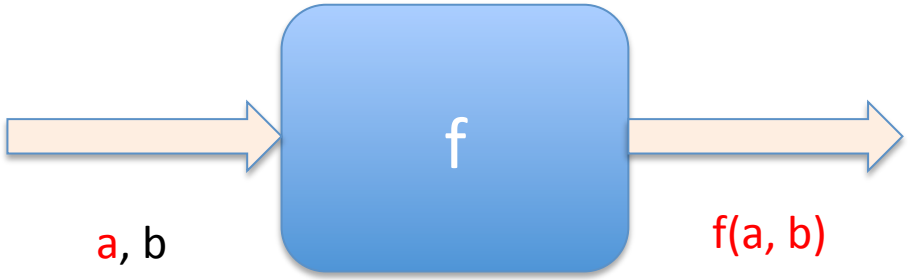
Work-arounds

e.g., post-factum information changes, fictitious entity instances (Soffer 2016)

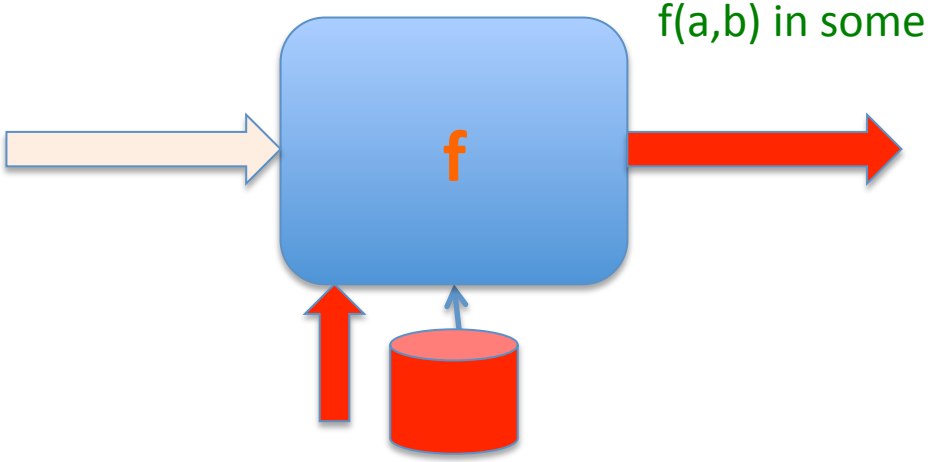
Temporal aspects

- Untimely information
- Delayed recording of information

Propagation of effects in processes



f(a,b) in some cases…



Propagation of effects

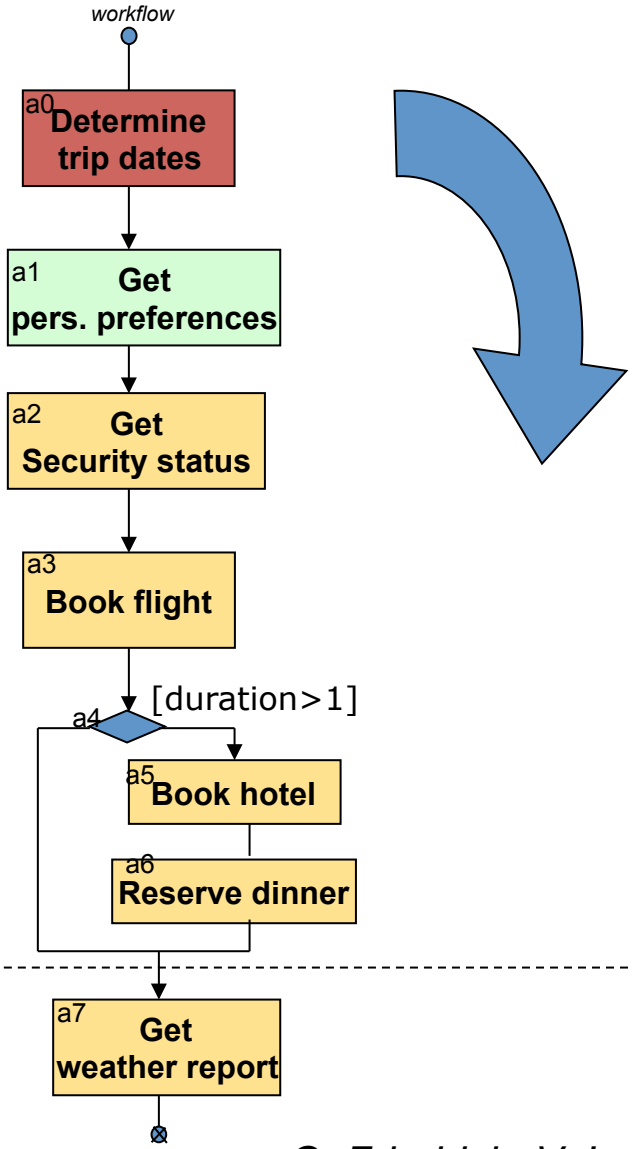
abnormal

e.g. some inputs provided by a Human are faulty

ok

possibly infected

failure detected



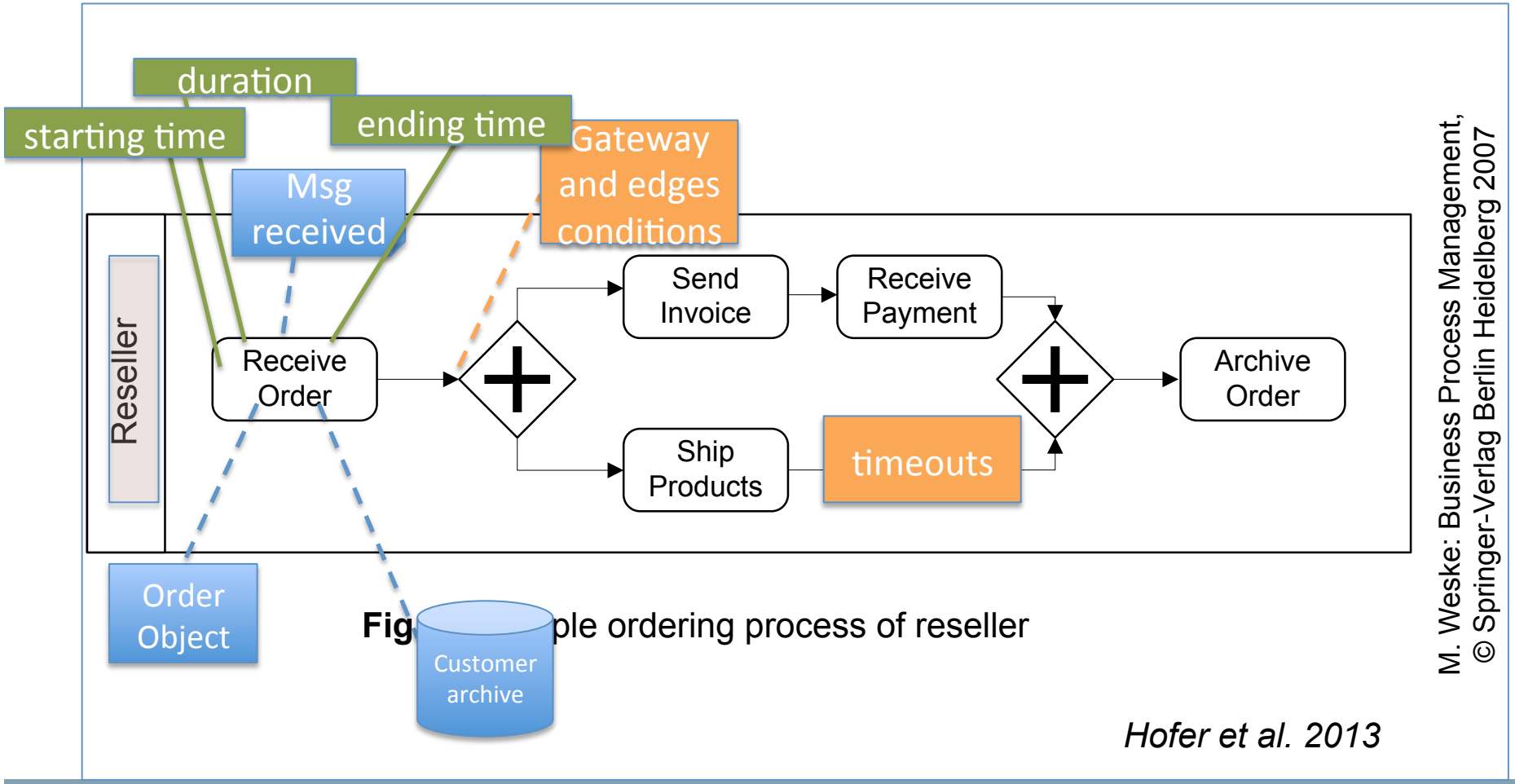
executed

not executed

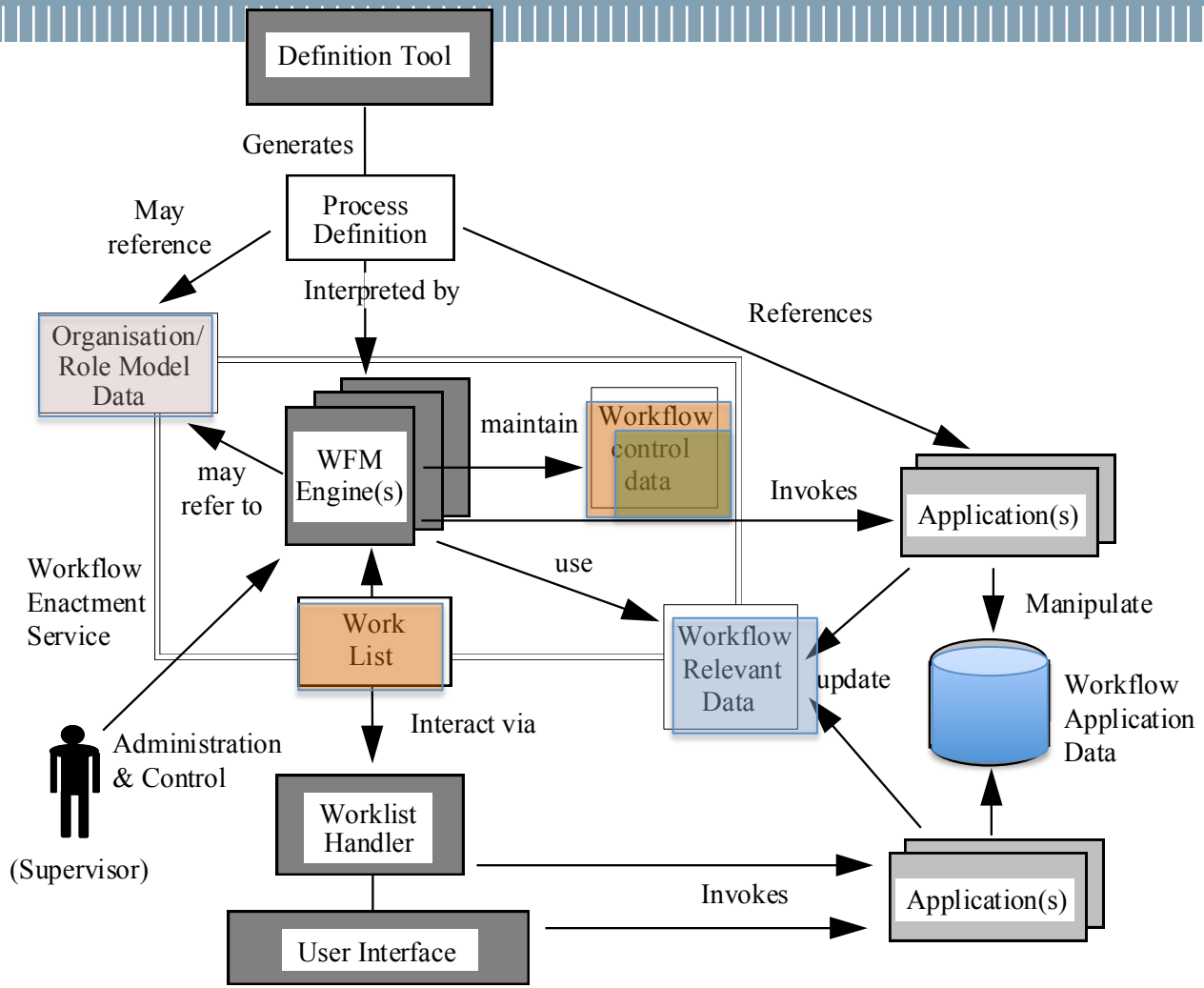
G. Friedrich, V. Ivanchenko, Univ. Klagenfurt

WHAT

A control flow perspective



WF Product Implementation Model (WfMC)



Software component
System control data
External product/data

Data Quality

"fitness for use"

Studying:

- Producing information products
- Cleaning and merging data (datawarehouses)

Main techniques:

- Data quality dimensions and techniques for improving the quality
- Record linkage: recognizing entities/objects
- Structured/semistructured data analysis (e.g. address formats)

Data Quality dimensions

Fitness for use

Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Added Value, Timeliness, Completeness, Amount of data, Interpretability, Easy of understanding, Consistency, Concise representation

179 dimensions

Batini Scannapieco 2016

Data Quality Attributes identified in BP modelling

DQ ► Dimension / Work ▼	Integrity	Accuracy	Uniqueness	Completeness	Non-Obsolescence	Consistency	Timeliness	Objectivity	Believability	Reputation	Accessibility	Security	Relevancy	Value-added	Amount of Data	Interpretability	Understandability	Concise Rep.	Consistent Rep.	Easy of Manipulation
Lu et al. (2000)	x																			
Soffer (2010)		x																		
Bringel et al. (2004)		x		x			x	x	x	x	x	x	x	x	x	x	x	x	x	x
el Abed (2011)		x	x	x	x	x	x													
Heravizadeh et al (2008)		x		x			x	x	x	x	x	x	x	x	x					

Cappiello et al 2012

Data quality dimensions

The ones mainly considered in process modeling and management

- Accuracy
- Completeness
- Timeliness

(sometimes merged into incorrectness)

- Linkability to source (provenance)

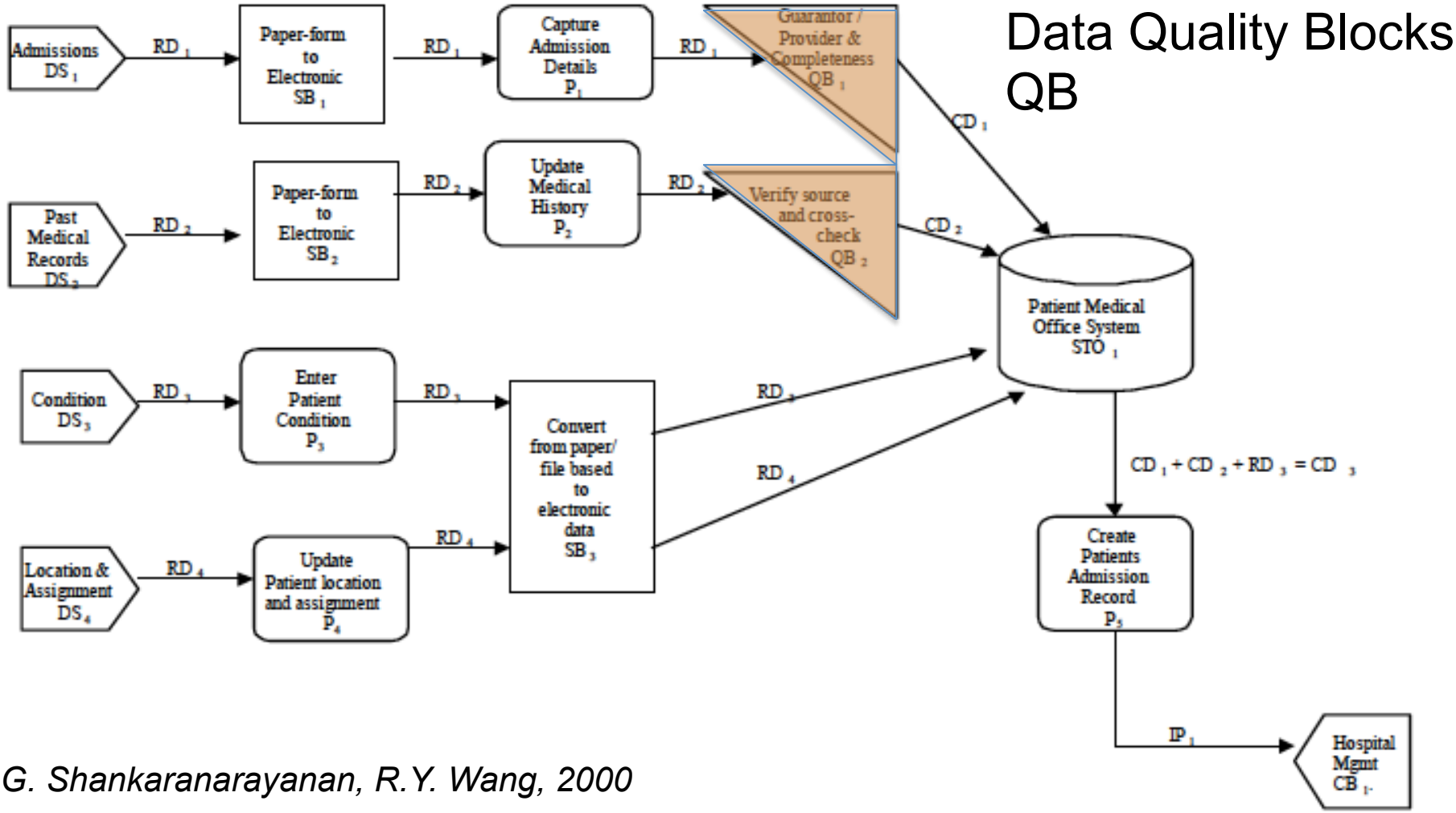
HOW

HOW

Modeling data quality in processes

IP-MAPS – Information as product

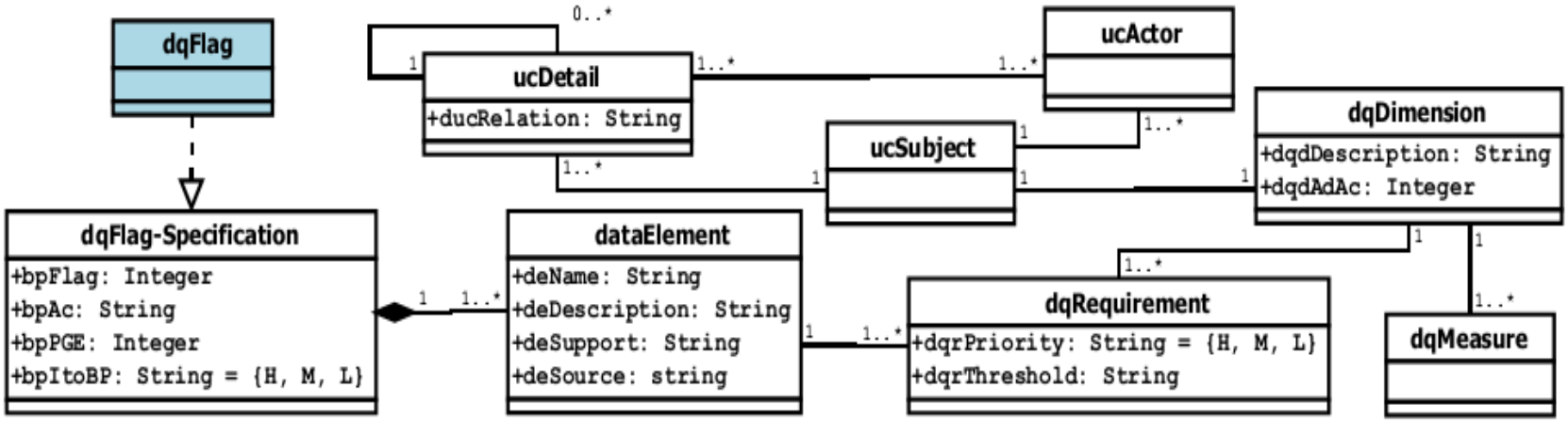
Figure 1: IP-MAP for Patients Admissions Record



Data Quality Blocks QB

G. Shankaranarayanan, R.Y. Wang, 2000

A conceptual model - dqBP

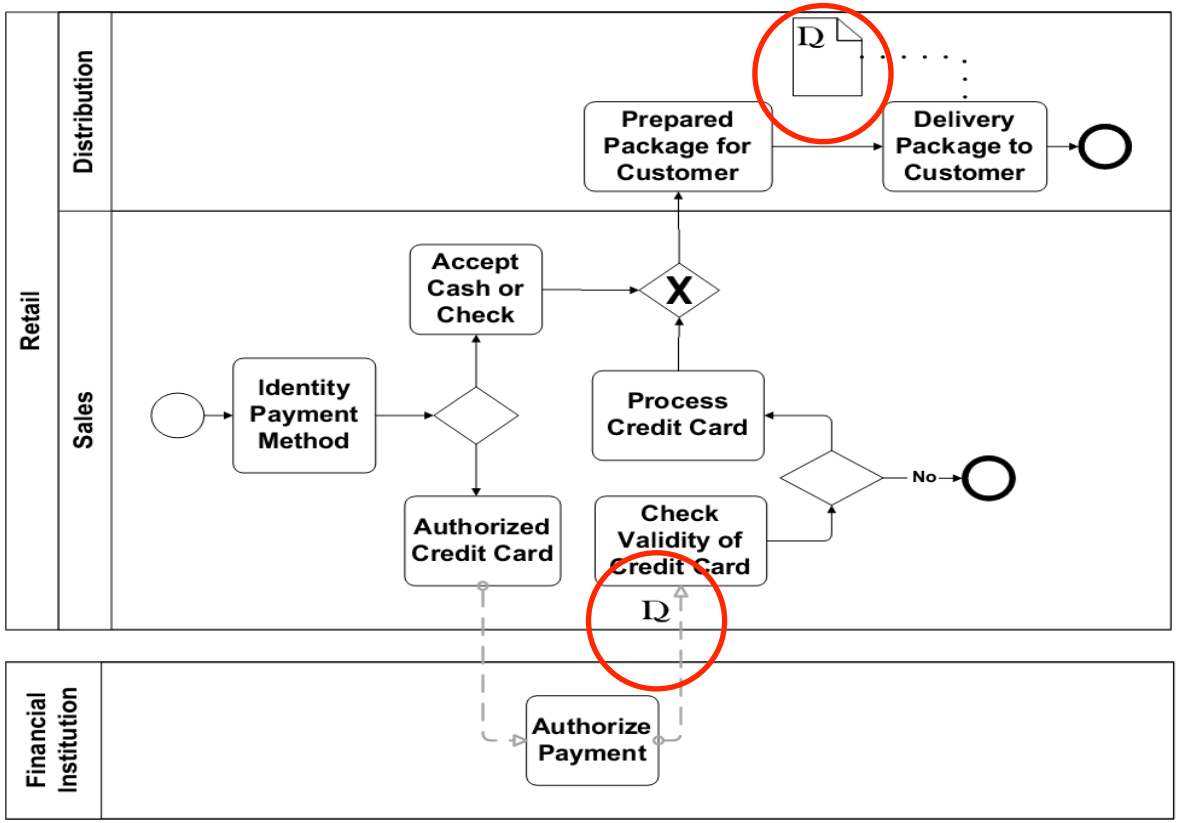


dqFlag: Abstract class containing data quality flag specifications associated with a BP element in the BPMN model. Each data quality flag must be indicated in detail in dqFlag-Specification realization.

Notation: 

Cappiello et al. 2012

Illustrative example: BPMN model with DQ Flags



Cappiello et al. 2012

Representation of the combination of Data-related BPMN elements and DQ flags

Graphical Representation	Intended use of the Graphical Representation
DQ Message	It represents that data contained in the message might satisfy some DQ requirements for the sake of the business success, e.g. Completeness and Consistency in a drug prescription from the doctor to a patient.
DQ Message flow	It represents that data implicitly contained in the message (the message does not appear in the flow) might satisfy some DQ requirements to develop successfully the business, e.g. Currentness for a credit card authorization from the bank.
DQ Conversation	It represents that data in some messages contained in the conversation might satisfy some DQ requirements for the sake of the success the business process, e.g., Security and Accuracy of the data interchanged between a customer and an airline Web application during the flight booking process.
DQ Data Object	It represents that data in the data object might satisfy some DQ requirements to successfully achieve the goals of the business process, e.g. Completeness, Consistency and/or Accuracy of the data required to successfully deliver and ordered package to a customer.
DQ Data Store	It represents that data contained in a data store might satisfy some DQ requirements for the sake of the success of the business process, e.g. Checking the completeness of the data updated about product sale.
DQ Activity	It represents that used/produced data in the activity might satisfy some DQ requirements to the business success, e.g. Checking the Precision and Accuracy of the budget generated as the output of one activity.

Cappiello et al. 2012

DQ Flags specifications

DQFlag1 → DQFlagSpecification1	
BPMN element: Data Object	**P. exec.**: 75%
Influence: High	**Overhead**: 25%
Name: Delivery Order	**Support**: Electronic
Description: Delivery order (customer information)	**Source** Internal
DQ Requirements Accuracy (High) and Completeness (Medium)	

DQFlag2 → DQFlagSpecification2	
BPMN element: Message Flow	**P. exec.**: 50%
Influence: Medium	**Overhead**: 12,5%
Name: Financial institution response	**Support** Electronic
Description: Delivery order (customer information)	**Source** Internal
DQ Requirements Currency (High)	

Cappiello et al. 2012

HOW

Assessment

Measuring impact of non-quality

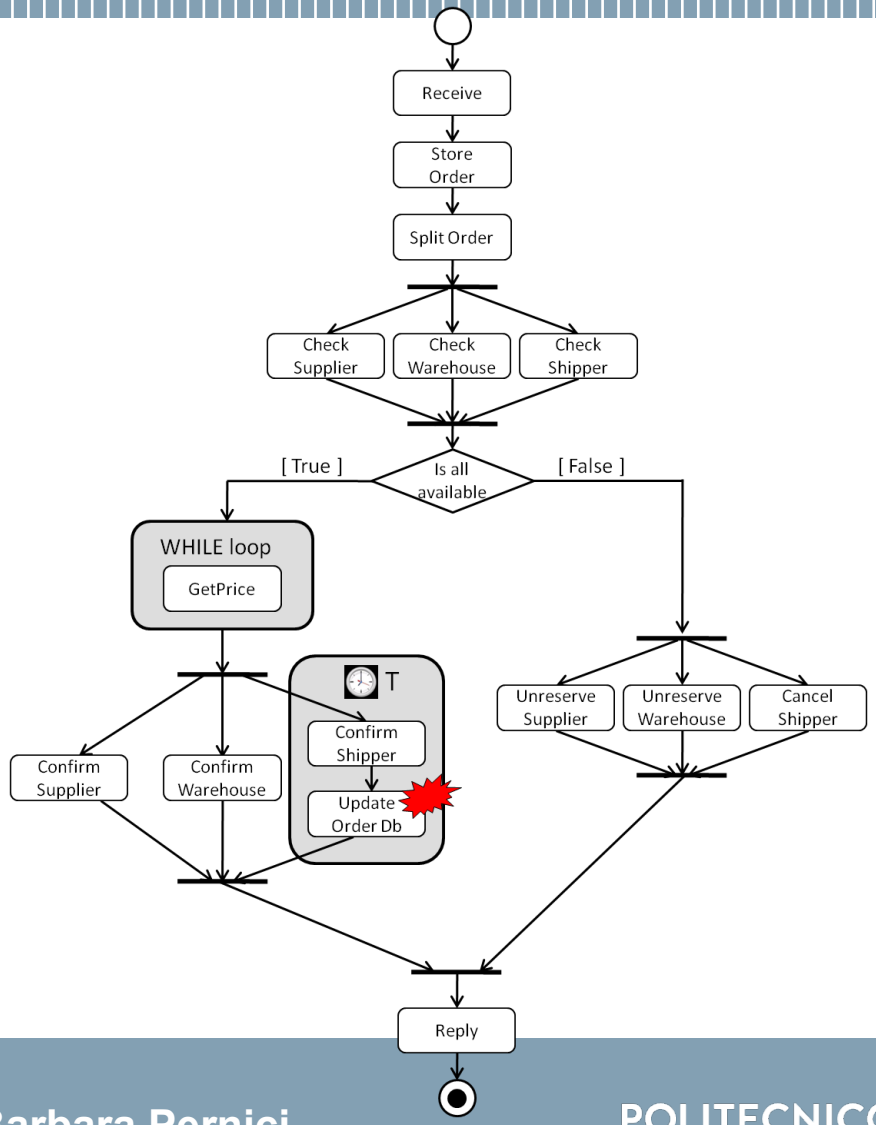
Economical evaluations

- Direct costs and cost of non-quality (IP-MAP)

Evaluation of the impact of data errors

Multiple dimensions evaluation

Testing data faults



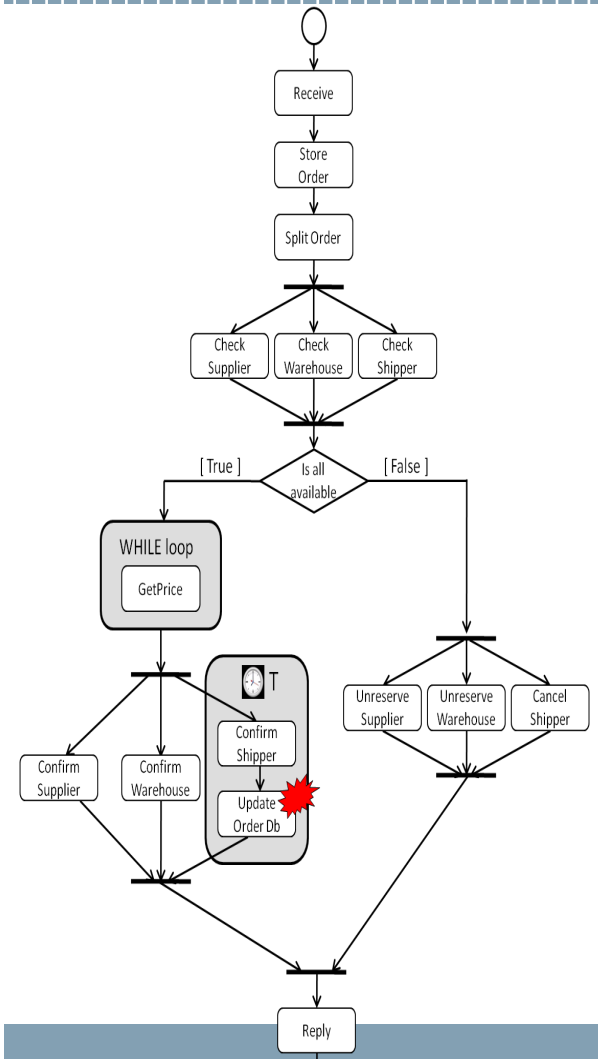
Test example: a simple foodshop

A BPEL process (Shop) and four simple services: LocalShop, Supplier, Warehouse, Shipper.

Inject data faults in exchanged messages and in local databases.

Fugini, Pernici, Ramoni, ISF, 2009

Test case: Data Faults



Operation	Data	Data fault		
		Typo	Null	Misalignment
Shop - Receive	CustomerInfo	ND	D(0)	ND
Shop - Receive	ItemList	D(1)	D(0)	ND
LocalShop - StoreOrder	ItemList	D(1)	F	D(10)
LocalShop - StoreOrder	CustomerInfo	ND	ND	ND
LocalShop - SplitOrder	ItemList	D(1)	F	D(11)
Supplier - Check	ItemList	F	F	D(13)
Warehouse - Check	ItemList	D(2)	F	D(14)
Shipper - Check	ItemList	ND	F	ND
LocalShop - GetPrice	ItemList	D(0)	F	ND
Supplier - Confirm	ItemList	D(1)	F	ND
Warehouse - Confirm	ItemList	D(1)	F	ND
Shipper - Confirm	ItemList	D(1)	F	ND
LocalShop - Update	ItemList	D(1)		

Fugini et al. 2009

Impact analysis – redesign and value changes scenarios

Primitives to define impact

- Primitive 1 (P_1): An **activity** affects a **data item**

- A **data item** affects
  - Primitive 2 (P_2): **another data item**
  - Primitive 3 (P_3): A data item affects **another data item** through an activity
  - Primitive 4 (P_4): A data item affects an **activity**
  - Primitive 7 (P_7): A data item directly affects **a routing constraint**

- **A routing constraint** affects
  - Primitive 5 (P_5): affects an **activity**
  - Primitive 6 (P_6): a **gateway**

Analysis

- *redesign* and *value changes* scenarios

Techniques

- Indirect impacts
- Trace analysis and reachability

Results derived as **queries**

Tsouri et al 2016

Impact analysis – redesign and value changes scenarios

Q1 for P_1 – returns data items affected by an activity	Q2 for P_2 - returns data items affected by a data item	Q3.1 for P_3 - returns data items affected by a data item through an activity (outputs)
Select name, 'd', 'P_1' From output_of Where activity_Id=@key	Select affected_name, 'd', 'P_2' From related_to Where effecting_name= @key	Select affected_name, 'd', 'P_3' From affected_through Where effecting_name= @key

Q3.2 for P_3 - returns activities affected by a data item to create an output	Q4 for P_4 - returns activities affected by a data item	Q5 for P_5 - returns activities affected by a routing constraint
Select activity_Id, 'a', 'P_3' From affected_through Where effecting_name=@key	Select activity_Id, 'a', 'P_4' From input_for Where name=@key	Select activity_Id, 'a', P_5' From Flow Where routingC_Id=@key

Q6 for P_6 – returns gateways affected by a routing constraint	Q7 for P_7 - returns routing constraints affected by a data item
Select gateway_Id, 'g', 'P_6' From Flow Where routingC_Id=@key	Select routingC_Id, 'r', P_7 from used_in where name=@key

Indirect impacts
Trace analysis and reachability

Fig. 3. Generic SQL queries for extracting impacts of process elements

Tsouri et al 2016

Elements for assessments – running processes

Identification of:

Data quality metrics
- Which data
- Which measuring devices
- Measurement scale
- Measurement procedures

Evaluation of processes
- KPIs

Otto et al., ICIQ 2009

Assessment of global quality

Multiple quality dimensions

- Weighted sum

$$q = \Sigma\, w_i dq_i$$

- How to determine weights?

- A global measure for the process (for each execution path)
  - minimum dq values

HOW

Improvements – design and run time

HOW

Improvements – design time

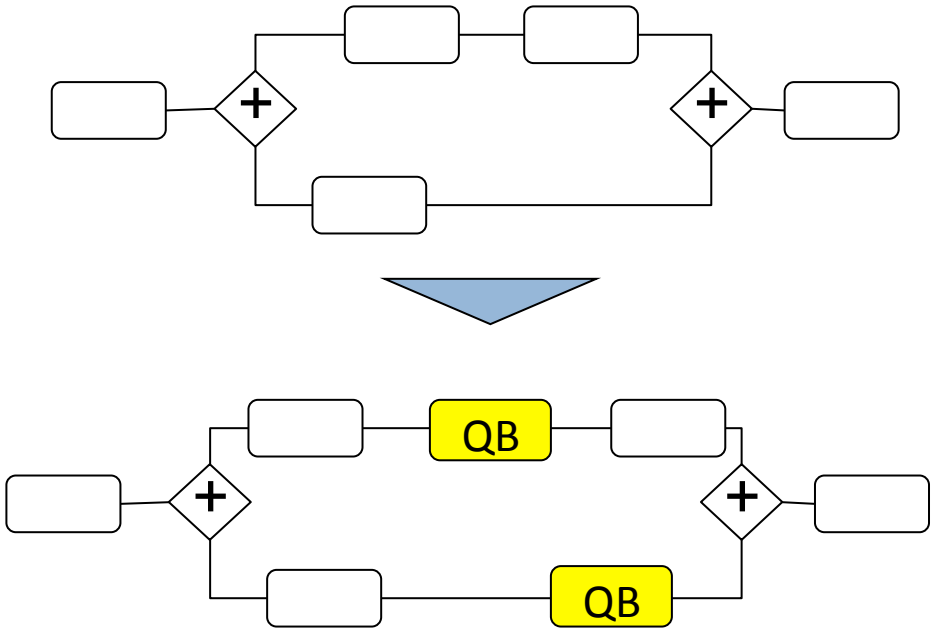
Data Quality Activities Repository: an example

DQ Dimension	Definition	Improvement Activities	Examples
Accuracy	The extent to which data reflects a real-world view within a context and a specific process [1, 18, 20].	- Determine the data set, which requires accuracy. - Verify data provided against the right domain. - Verify data coming from alternatives sources. - Clean database to achieve the required level of accuracy.	- The price received by the client for a booking hotel must be accurate. - In a medical prescription, the name of the medicines can be confronted with the Vademecum. - The weight of a package to be delivered must be contained within a specific range of values.
Timeliness	The extent to which data are sufficiently updated for the context and a specific process [1, 19, 20].	- Verify if data have the required age for the task. - From different sources, select the one providing data with the age required for the process. - Check if data are delivered within the required time.	- Check if the same data are in different company's source and if it is closer to the right age required, and then take values from this source. - Bank's response to check a credit card must be lower than 5 seconds.
Completeness	The extent to which data have all values necessary for a successful execution of a process in a specific domain and context [1, 19, 20].	- Specify which data are mandatory - Verify/Ensure whether all mandatory items of data have values. - Complete data provided with other sources of data. - Use a procedure to force the delivery of all mandatory data.	- Check if the same data are in different company's and then complete the golden register - To deliver a package, all data about the address and customer identification must be complete.

Cappiello et al, ECIS 2013

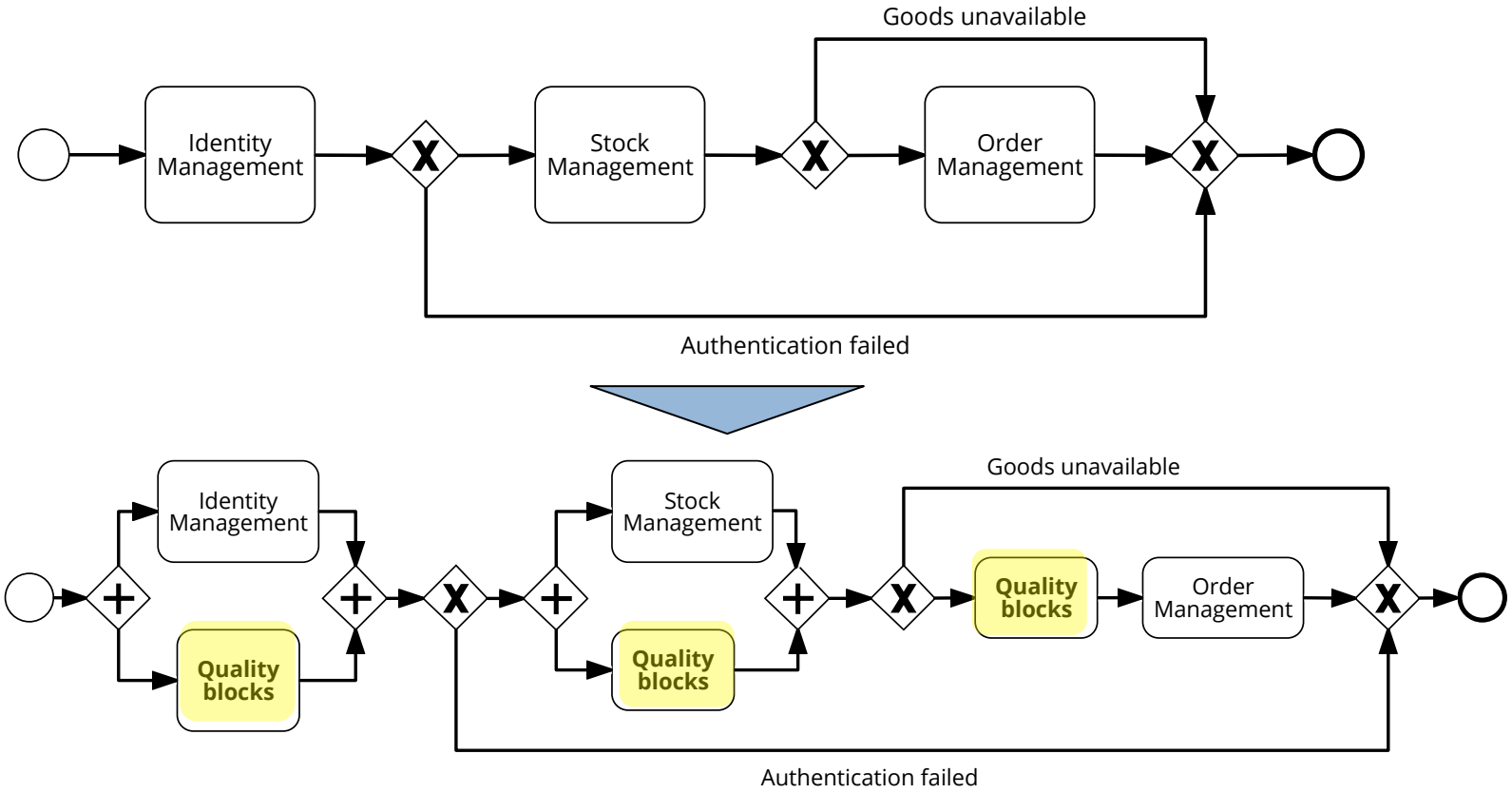
Quality-aware redesign of the business process

Insert Data Quality Blocks inside the process to improve its data quality level



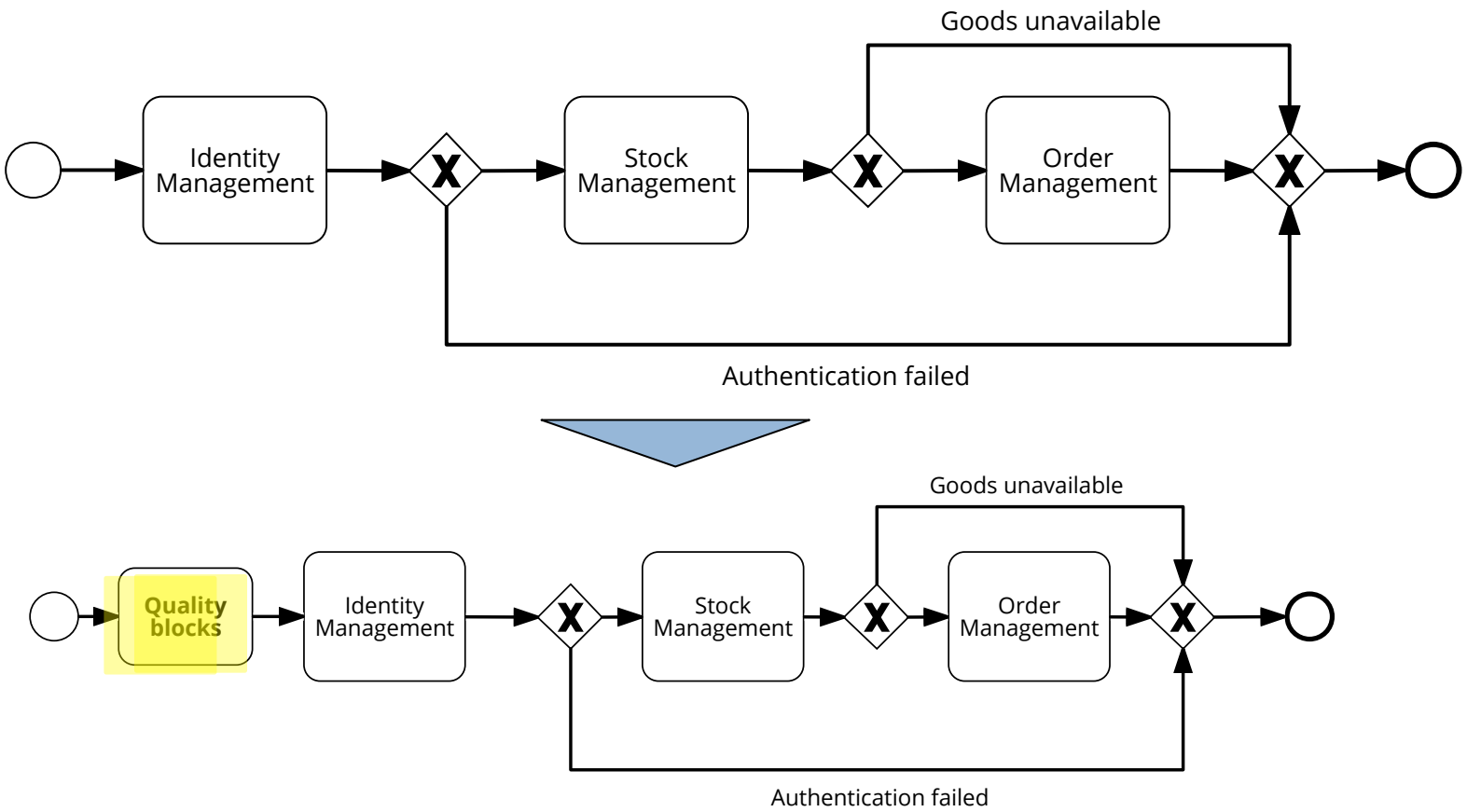
Cappiello, Pernici, Villani, 2014

Local check



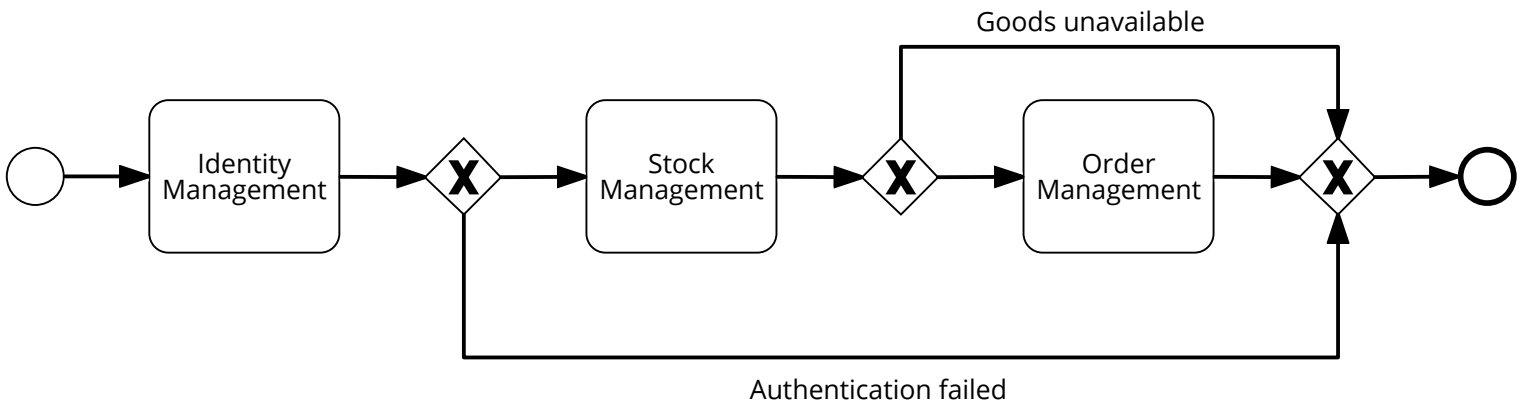
Cappiello, Pernici, Villani, 2014

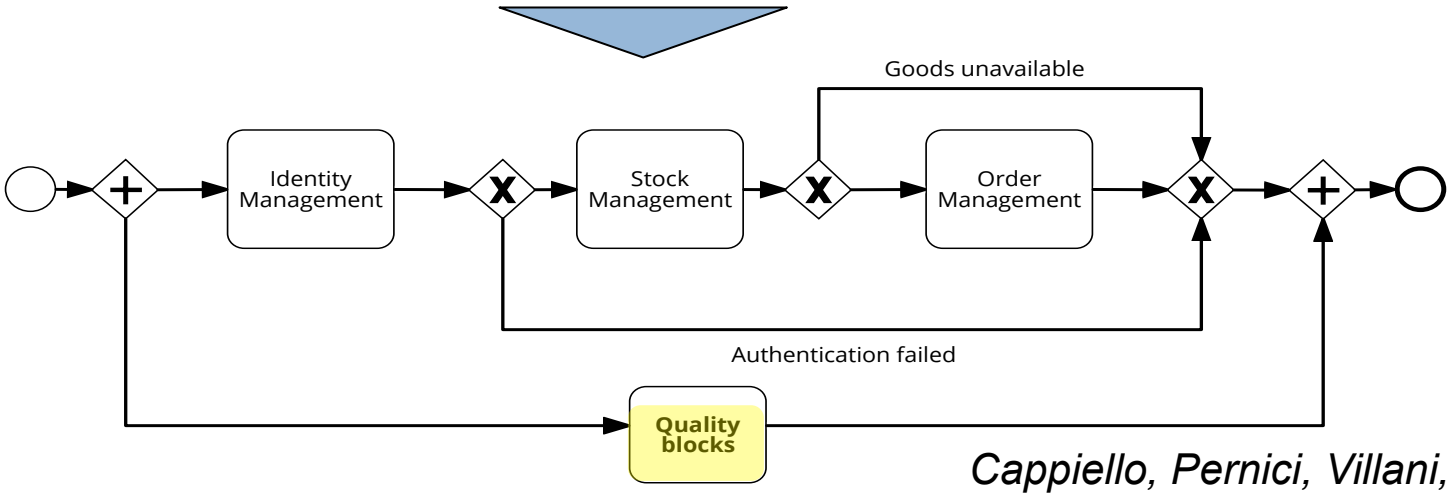
Preliminary Check



Cappiello, Pernici, Villani, 2014

Parallel Check





Cappiello, Pernici, Villani, 2014

How to assess quality checks

Are the controls effective?

do they capture the errors?

What is the residual impact of errors?

outcome of the process

do the controls really improve the outcome?

What are the weights for different dimensions?

Fault injection again!

An example – large financial institution
Probability of Default (PD) assessment
Multinational Group Wide Counterparts (Multinational, Banks, Funds, ...)

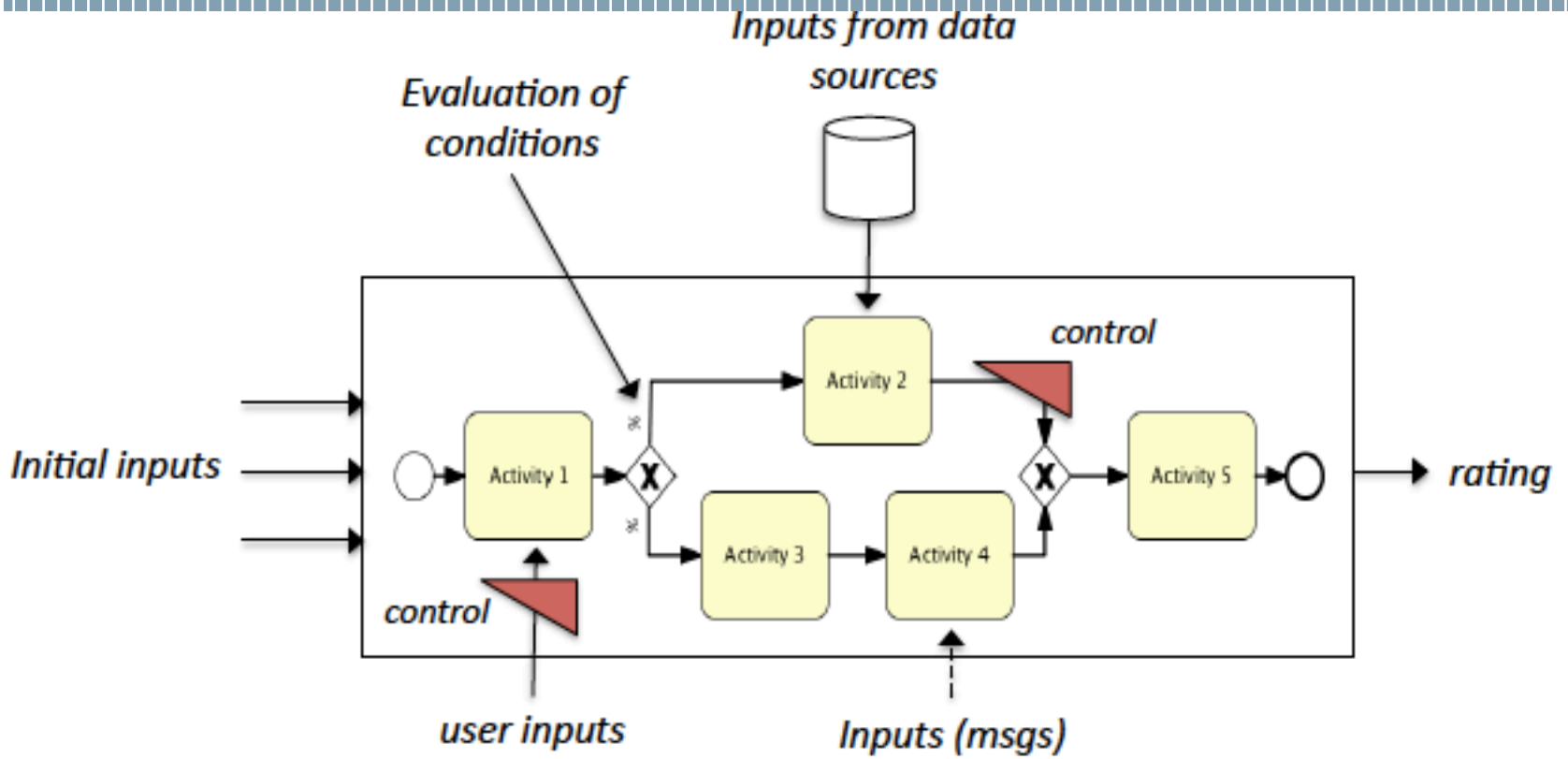


Fig. 1. Rating processes

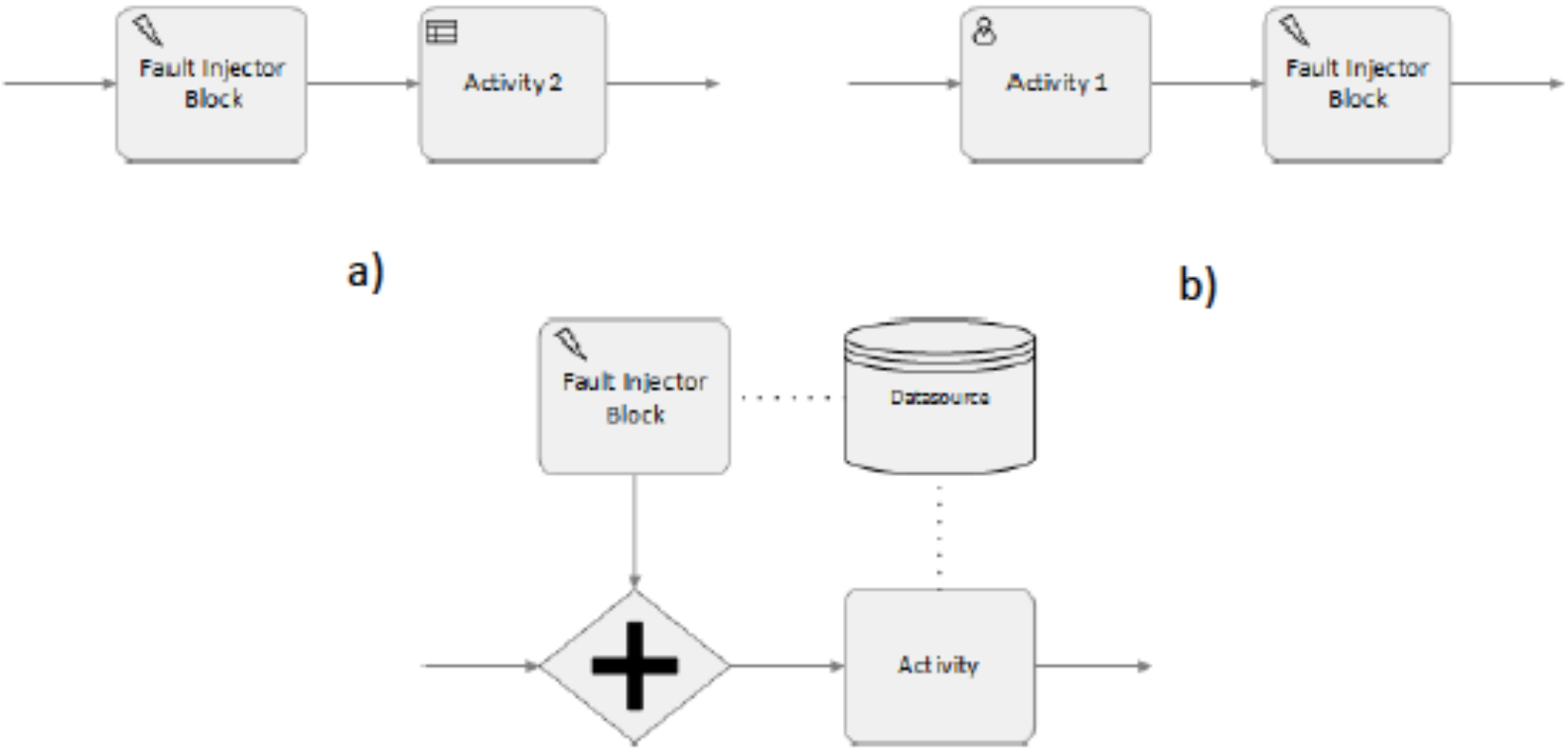
Standard & Poor's
AAA
AA+
AA
AA-
A+
A
A-
BBB+
BBB
BBB-
BB+
BB
BB-
B+
B
B-
CCC+
CCC
CCC-
CC
C
D

more than 2,500 counterparts

Goal: determine the overall Data Quality level of the evaluations performed by the financial institute

Cerletti Fratto Cappiello Pernici, in preparation

Injection patterns



Cerletti Fratto Cappiello Pernici, in preparation

Fault injection for validation

Fault types

- Missing values/documents
- Altered values
    - Within ranges
    - Outside ranges

Assessments

- Data Quality **controls validation**
- Process **behaviors validation**
- Data Quality dimension **weights validation**

Cerletti Fratto Cappiello Pernici, in preparation

Controls validation

150 faulty executions

$$v = \begin{bmatrix} control_1 \\ control_2 \\ \dots \\ control_j \\ \dots \\ control_m \end{bmatrix}$$

Compare fault-free control vector v with v' executing the process with one fault

Possible results

- One control affected: it captures the fault
- No control
- Several controls: duplication of efforts

Cerletti Fratto Cappiello Pernici, in preparation

Behaviour validation

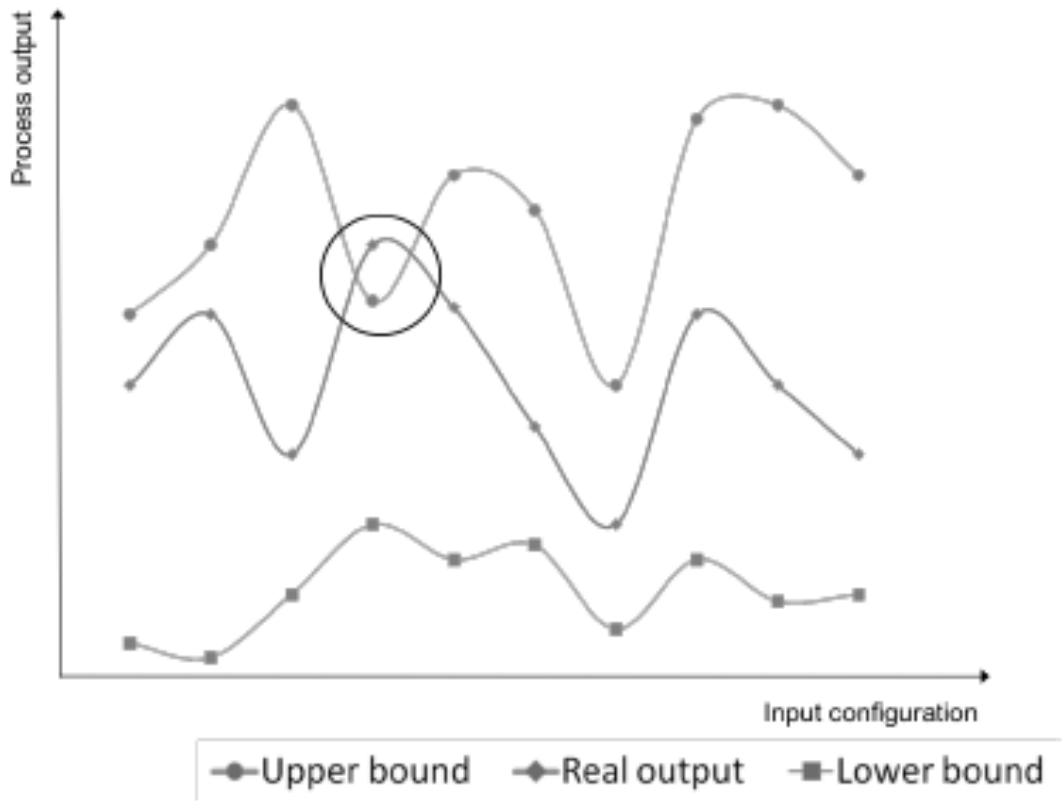
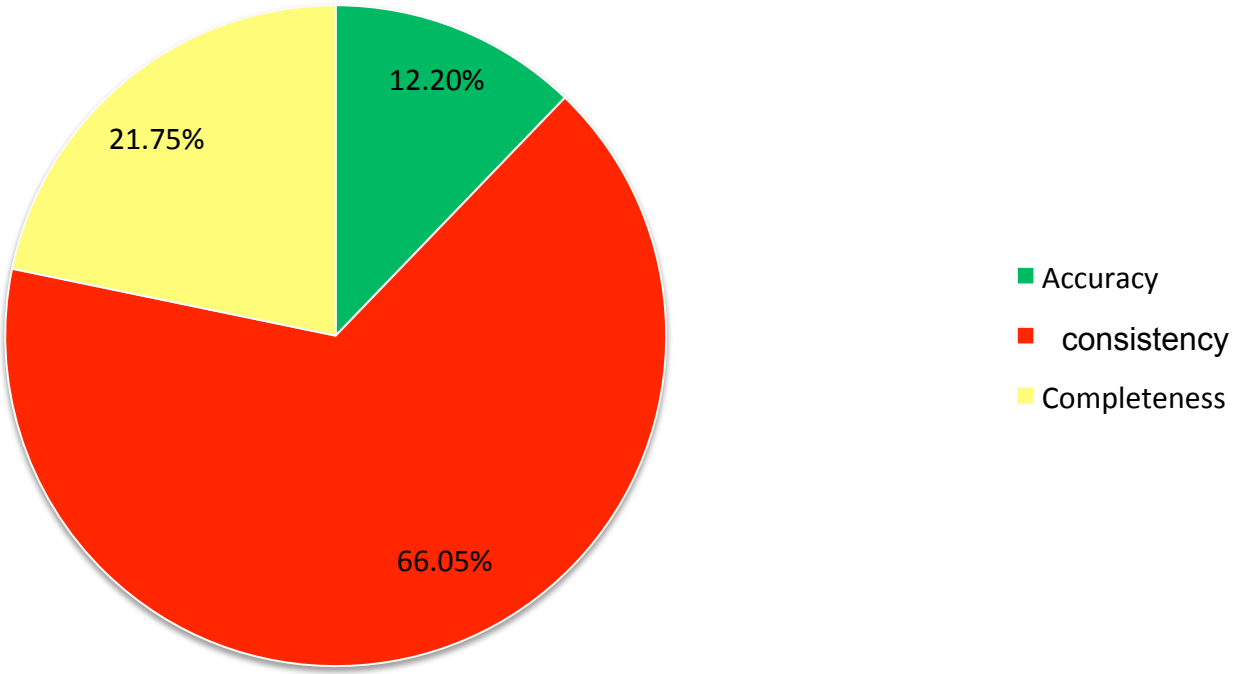


Fig. 4. Example of abnormal behavior

Cerletti Fratto Cappiello Pernici, in preparation

Weights assessment

Normalized variability



Cerletti Fratto Cappiello Pernici, in preparation

HOW

Improvements – run time

Use of variables

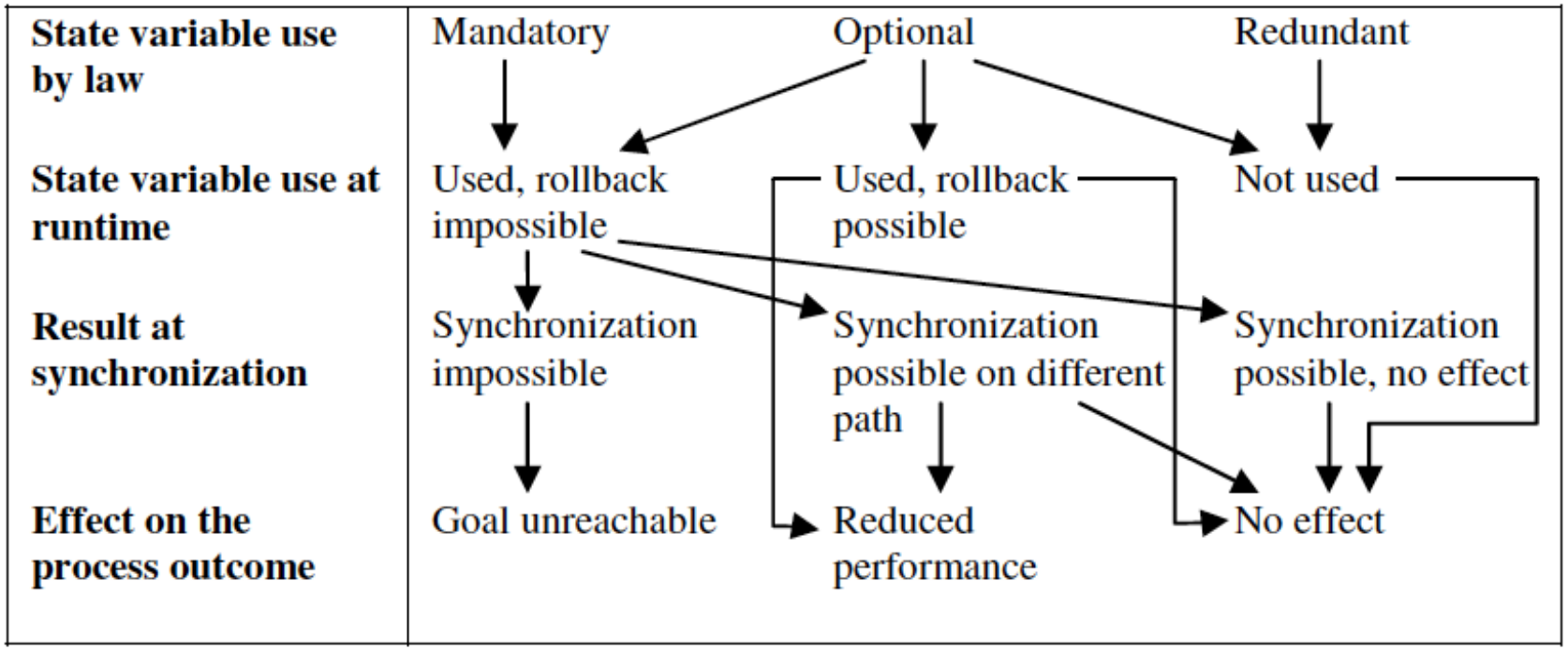
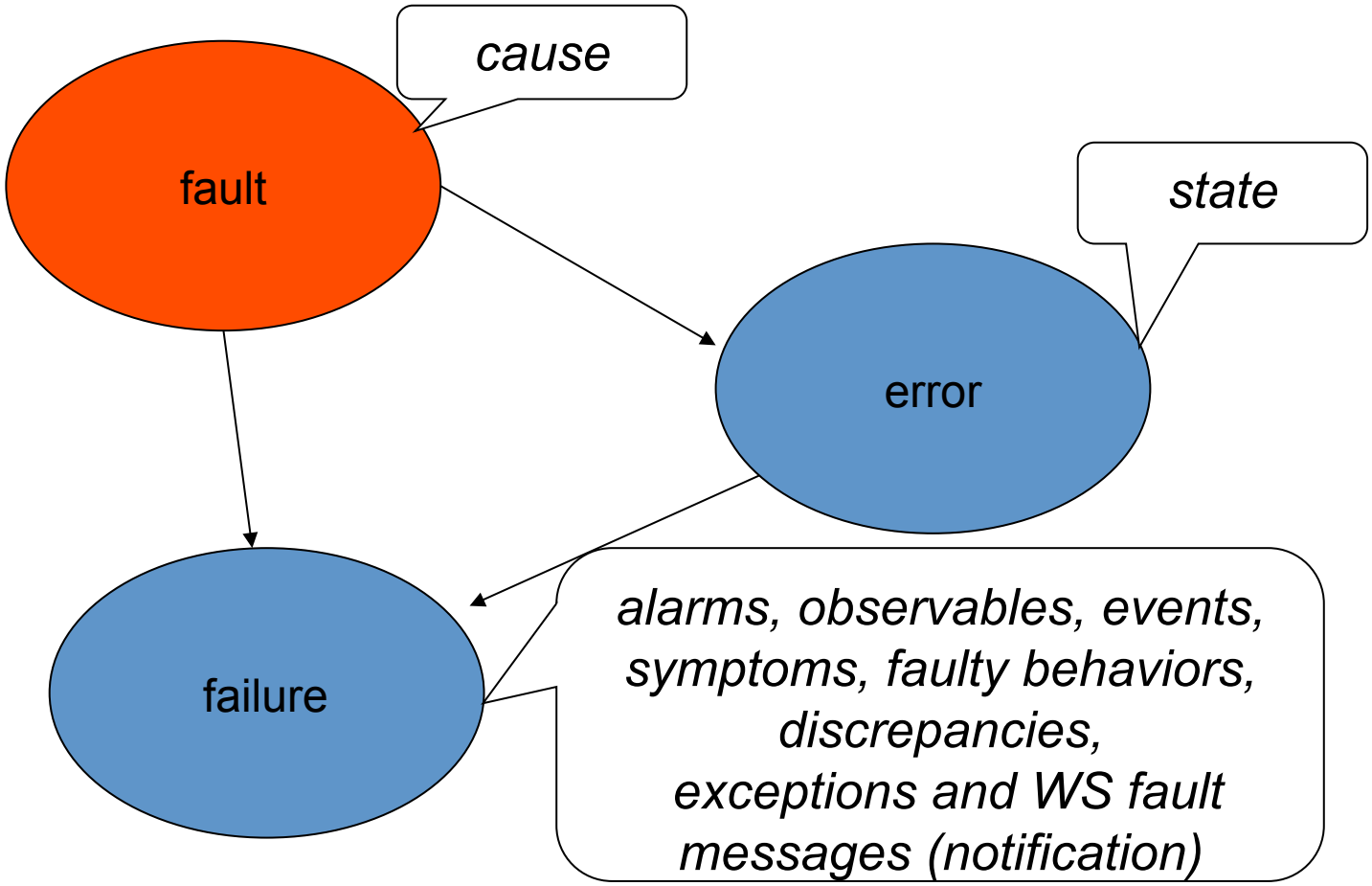


Fig. 2. Possible results of data inaccuracy

Soffer, 2010

Fault-Error-Failure cycle



Ws-Diamond

Service compositions

Focus on repairing failed processes

Techniques:

- Analysis of propagation of data errors
- Diagnosis of causes of failures
- Repair plans (minimal)

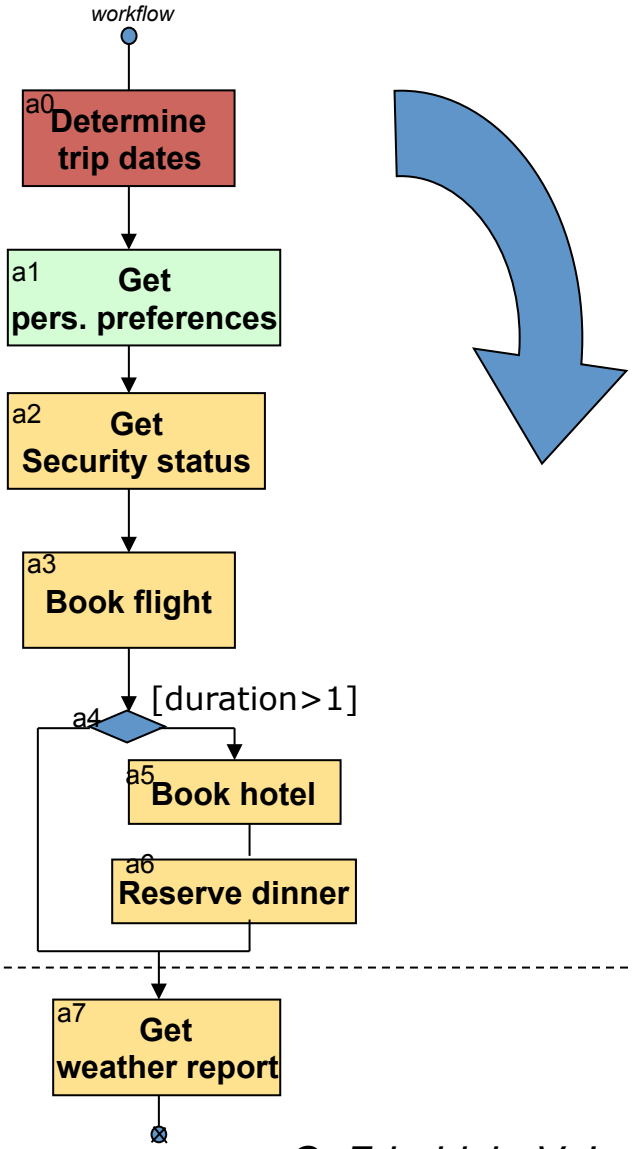
abnormal

e.g. some inputs provided by a Human are faulty

ok

possibly infected

failure detected

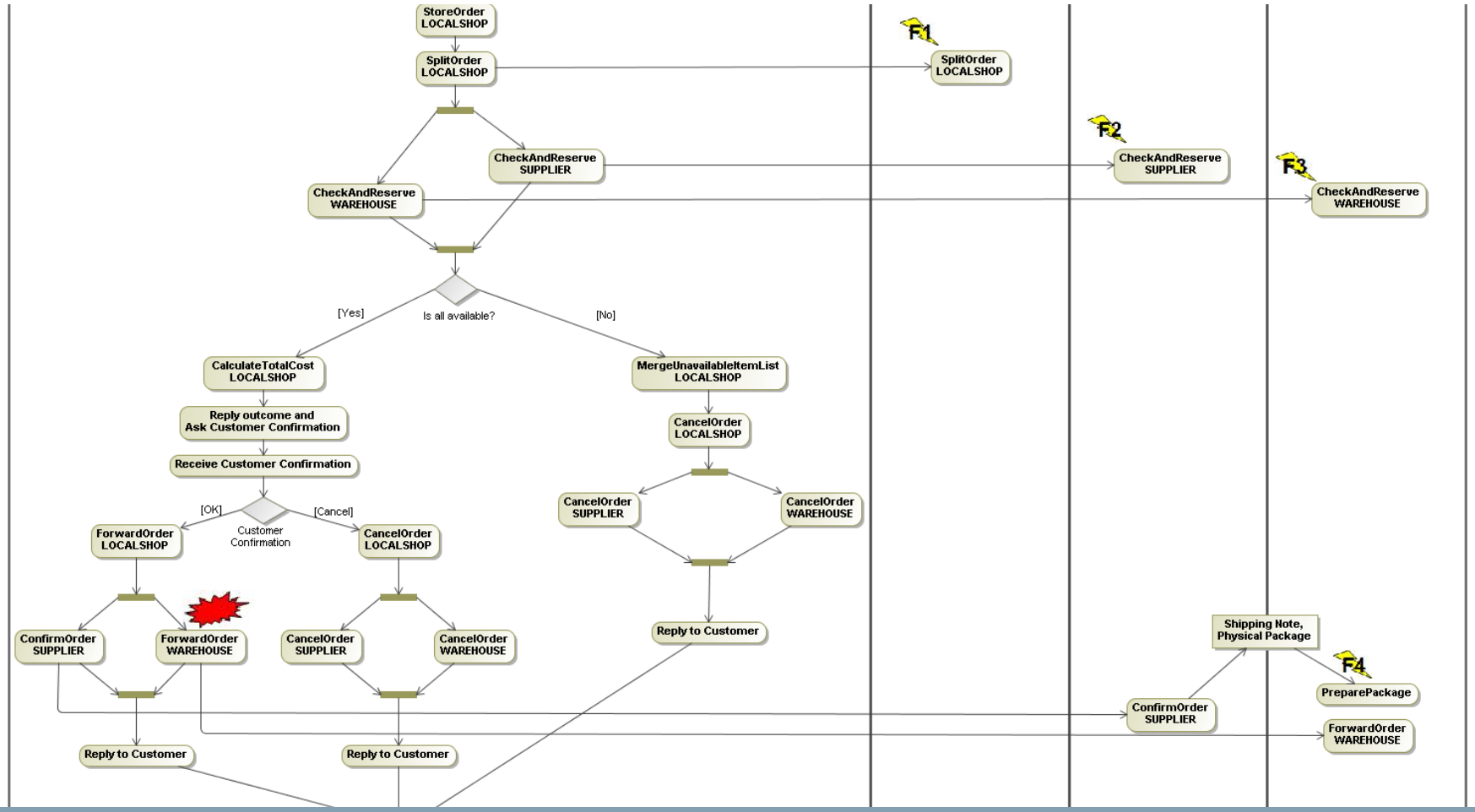


executed

not executed

G. Friedrich, V. Ivanchenko, Univ. Klagenfurt

WS-Diamond – an approach based on diagnosis and repair plans



Friedrich et al, 2010

Examples of Faults in the FoodShop Process

A misalignment inside the LocalShop database that, given a purchased item name, makes the service returning a wrong item code

F1 = <LocalShop Web service, SplitOrder>

The Supplier reserves a different item instead of the purchased one

F2 = <Supplier Web service, Check&Reserve>

The Warehouse reserves a different item instead of the purchased one

F3 = <Warehouse Web service, Check&Reserve>

The Warehouse creates a package containing a different item with respect to the reserved one

F4 = <Warehouse Web service, PreparePackage>

Case 1: Temporary Fault

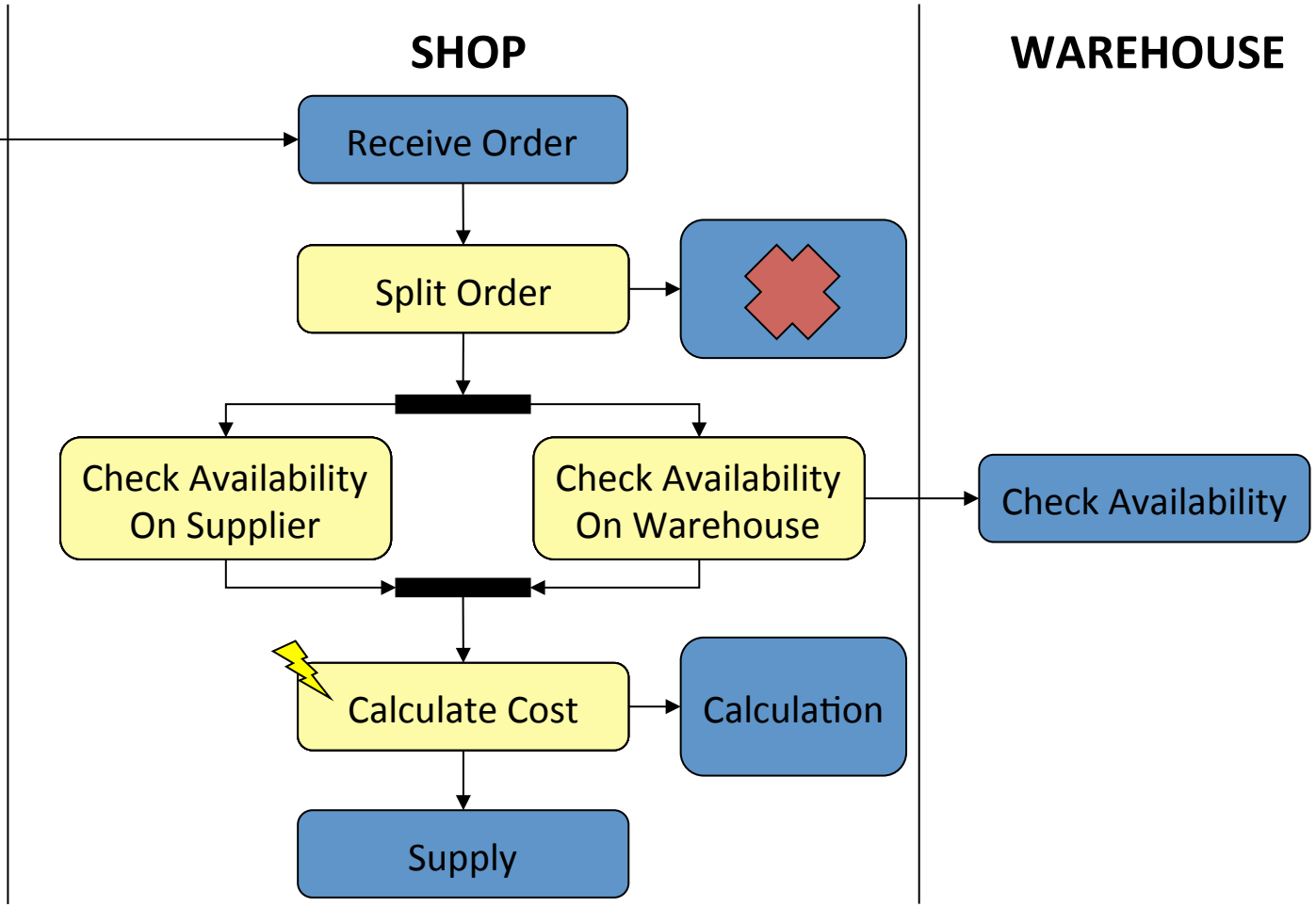
The SHOP is declared faulty

The Recovery Selector **stops** the SHOP and the WAREHOUSE

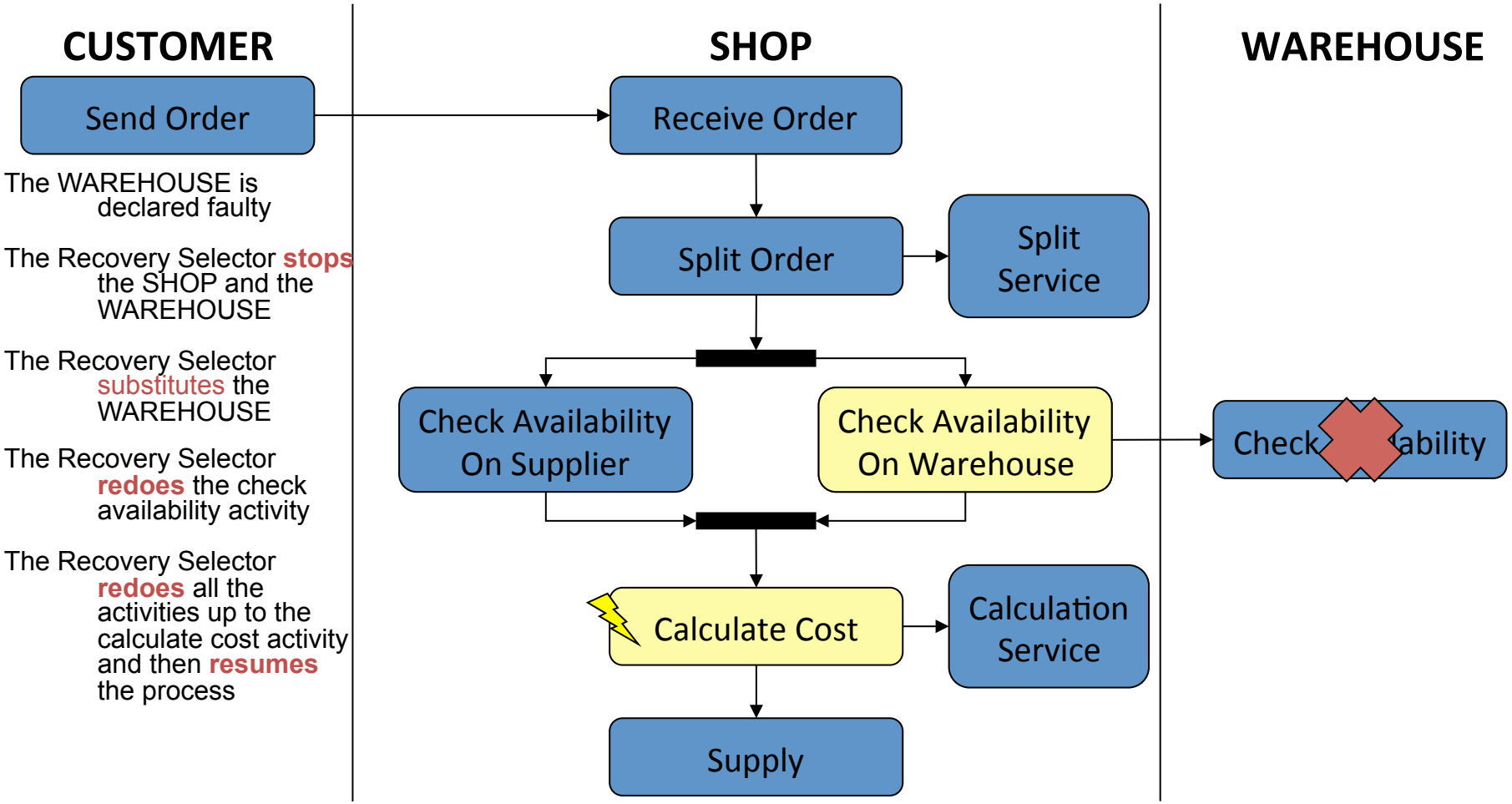
The split activity is repaired

The Recovery Selector **retries** the split activity

The Recovery Selector **retries** all the activities up to the calculate cost activity and then **resumes** the process



Case 2: Permanent Fault



The WAREHOUSE is declared faulty

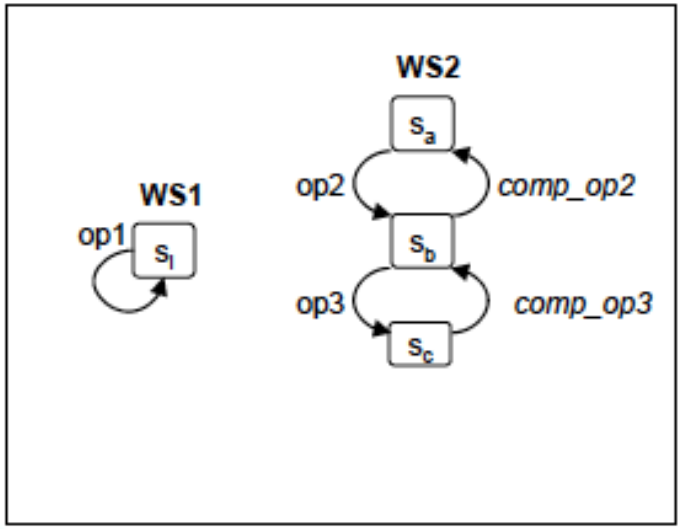
The Recovery Selector **stops** the SHOP and the WAREHOUSE

The Recovery Selector substitutes the WAREHOUSE

The Recovery Selector **redoes** the check availability activity

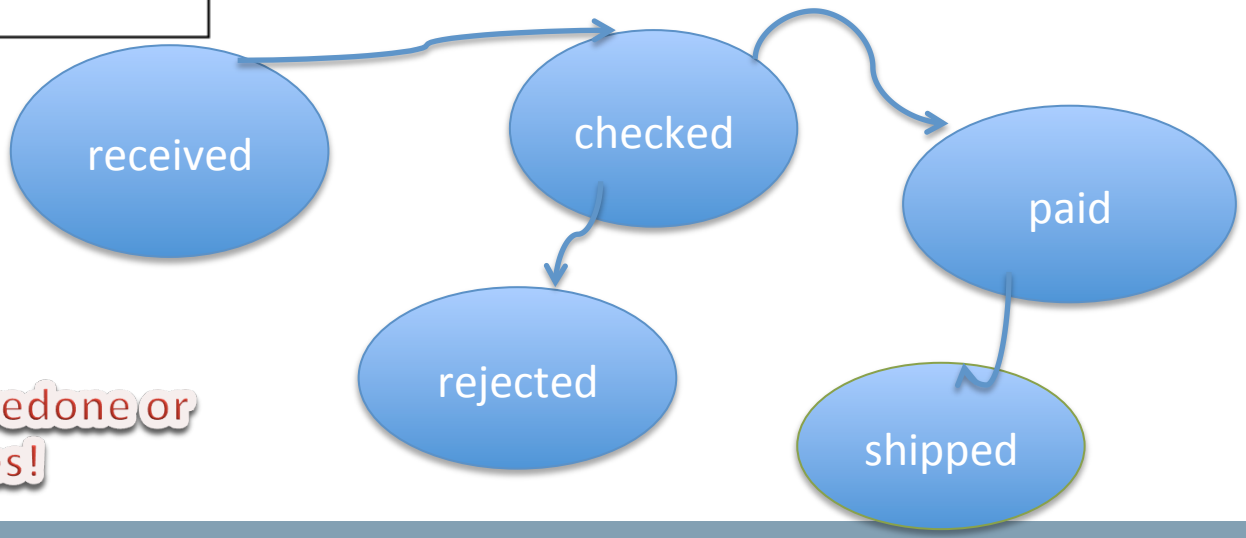
The Recovery Selector **redoes** all the activities up to the calculate cost activity and then **resumes** the process

Activities and objects have states



COMPENSATION ACTIONS

ORDER



Not all activites can be redone or Compensated at all times!

Instance Repair Actions

Retry/Execute an activity

Compensate an activity, that is invoking an operation which is defined as a compensation for a given one in a given state

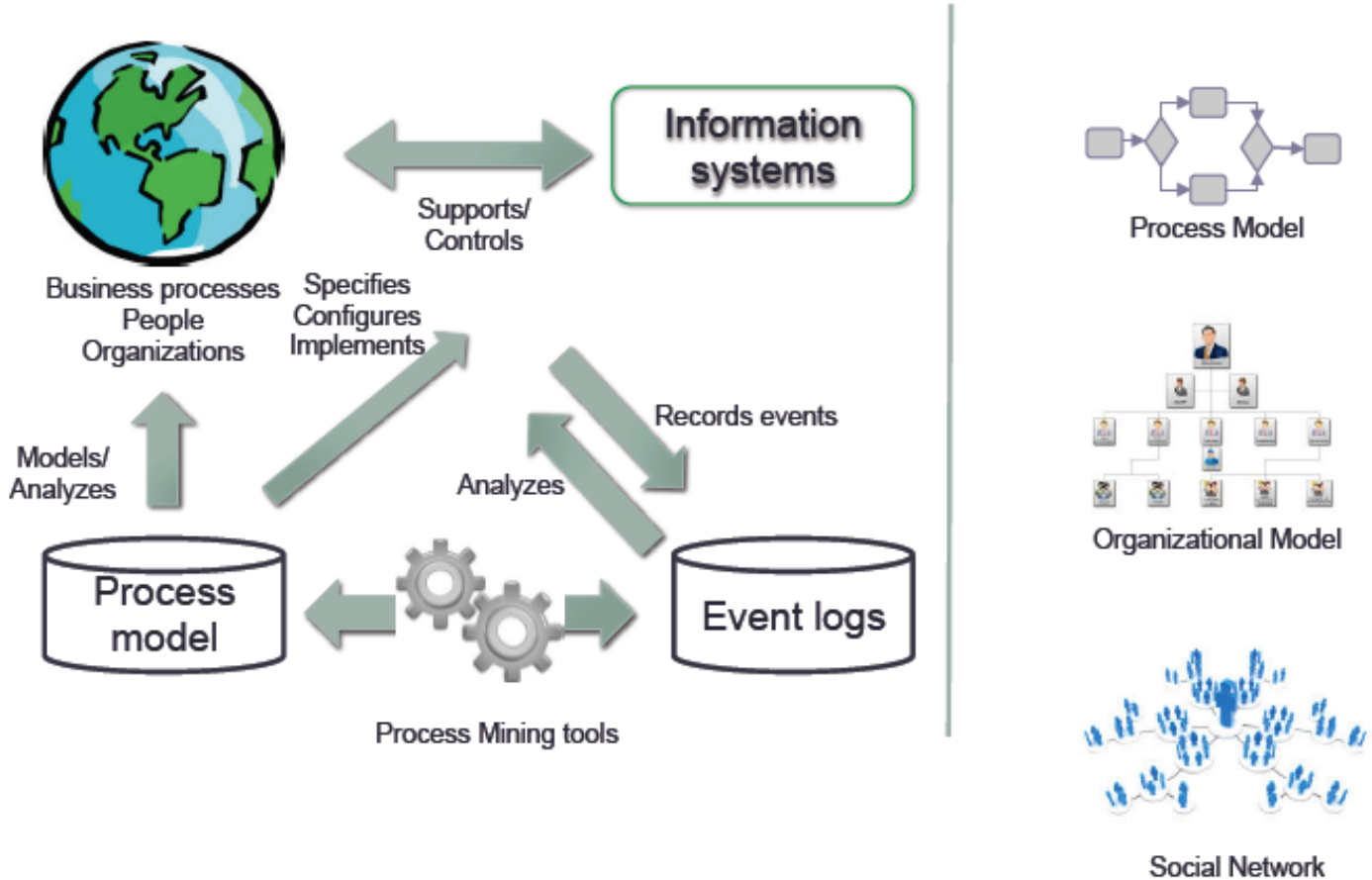
Substitute a Web service

Problems: session and state management, choreography

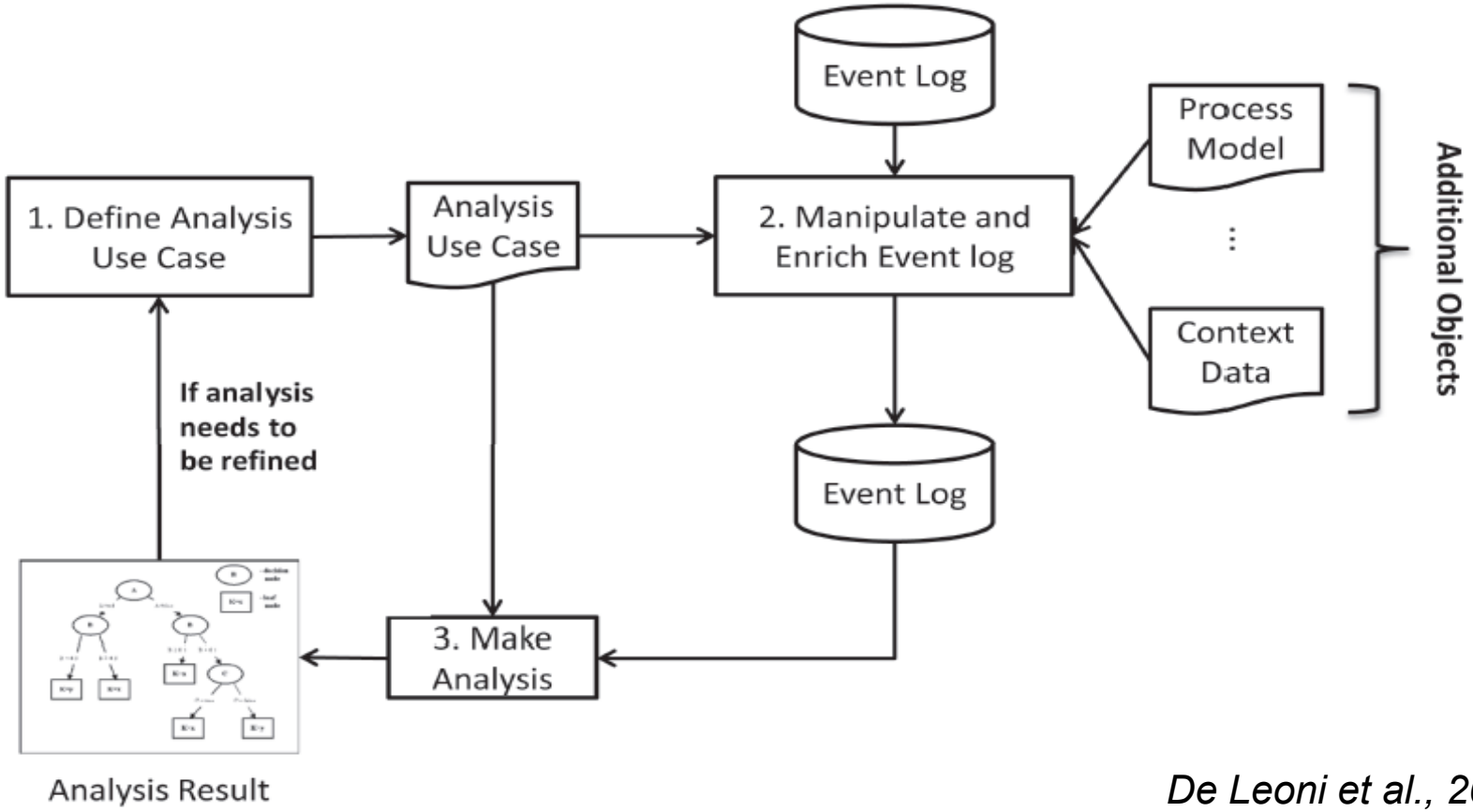
Other issues

Mining

Process mining and information systems

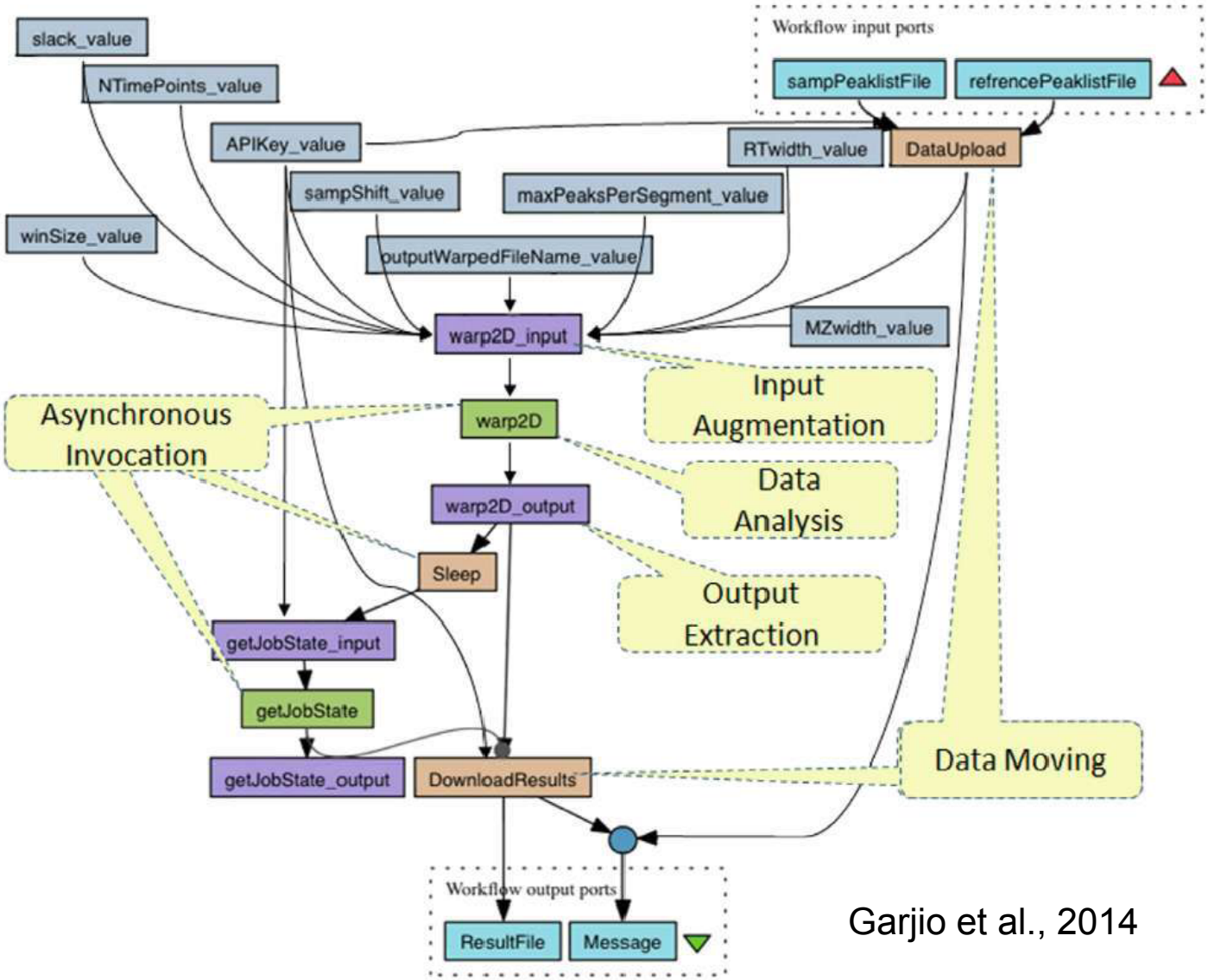


Van der Aaalst



De Leoni et al., 2014

Scientific workflows



Garjio et al., 2014

Barbar

ig. 2. Sample motifs in a Taverna workflow for functional genomics. The workflow transfers data files containing proteomics data to a remote server and augments several arameters for the invocation request. Then the workflow waits for job completion and inquires about the state of the submitted warping job. Once the inquiry call is eturned the results are downloaded from the remote server.

Scientific workflow motifs

Data operation motifs

- Data preparation
  - Combine
  - Filter
  - Format transformation
  - Input augmentation
  - Output extraction
  - Group
  - Sort
  - Split
- Data analysis
- Data cleaning
- Data movement
- Data retrieval
- Data visualization

Garjio et al., 2014

Workflow oriented motifs

Additional issues in scientific workflows

Data provenance:

- Why a given decision was taken?
- How results derived (support)

Definition: “description of the origins of a piece of data and the process by which it arrived in the database” (Batini et al. 2016)

Many open research problems

Measuring quality

- a costly activity
- efficient and effective measures are needed
- remember fitness for use

Evaluate the impact of errors

- study and model different types of data quality problems
- propagation and impact evaluation
- minimal repair

How to design stronger processes

- how to balance the different quality requirements
- where to insert data quality blocks
- consider awareness and un-awareness of data incorrectness

Related issues

Role of monitoring

Role of infrastructure

And more….

Processes in many areas

- scientific workflows
- data analysis
- production processes

Are conceptual modeling and associated tools and techniques going to support them in managing data quality aspects?

QUESTIONS?

References

- M.G. Fugini, B. Pernici, F. Ramoni: Quality analysis of composed services through fault injection. Information Systems Frontiers 11(3): 227-239 (2009)
- G. Friedrich, M.G. Fugini, E. Mussi, B. Pernici, G. Tagni: Exception Handling for Repair in Service-Based Processes. IEEE Trans. Software Eng. 36(2): 198-215 (2010)
- M. Vitali, B. Pernici: PiE - Processes in Events: Interconnections in Ambient Assisted Living. OTM Workshops 2015: 157-166
- C. Cappiello, B. Pernici, L. Villani: Strategies for Data Quality Monitoring in Business Processes. WISE Workshops 2014: 226-238
- A. Caro, A. Rodriguez, C.L. Yonke, C. Cappiello, Designing Business Processes Able to Satisfy Data Quality Requirements, ICIQ, 2012
- A. Rodríguez, A. Caro, C. Cappiello, I. Caballero, A BPMN Extension for Including Data Quality Requirements, 4th International Workshop, BPMN 2012, Vienna, Austria, September 12-13, 2012. Springer
- G. Shankaranarayanan, Richard Y. Wang, IP-MAP: representing the manufacture of an information product, ICIQ, 2000
- M. Scannapieco, B. Pernici, E. M. Pierce: IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework. IQ 2002: 279-291
- WS-Diamond, Web Service Diagnosability, Monitoring and Diagnosis, FP6 project, http://wsdiamond.di.unito.it/

C. Batini, M. Scannapieco, Data and Information Quality, Springer, 2016

P. Soffer: Mirror, Mirror on the Wall, Can I Count on You at All? Exploring Data Inaccuracy in Business Processes. BMMDS/EMMSAD 2010: 14-25

A. Tsoury, P. Soffer, I. Reinhartz-Berger, Towards Impact Analysis of Data in Business Processes, BPMDS, 2016

M. Ofner, B. Otto, H. Österle: A Maturity Model for Enterprise Data Quality Management. Enterprise Modelling and Information Systems Architectures 8(2): 4-24 (2013)

Y. Evron, Data inaccuracy-aware design of business processes, CAiSE DC, 2016

W. Van Der Aalst, Data scientist: The Engineer of the Future, Proceedings of the I-ESA Conferences 7, 2014

M. de Leoni, W.M.P. van der Aalst, M. Dees: A General Framework for Correlating Business Process Characteristics. BPM 2014: 250-266

A. Meyer, S. Smirnov, M. Weske, Data in Business Processes, Hasso-Plattner-Institutes, Universität Potsdam, 2011

D. Garijo, P. Alper, K. Belhajjame, O. Corcho, Y. Gil, C. Goble, Common motifs in scientific workflows: An empirical analysis, Future Generation Computer Systems 36, 338-351, 2014

N. Outmazgin, P. Soffer, A process mining-based analysis of business process work-arounds, Softw. Syst. Model 15, pp. 309-323, 2016

B. Otto, K.M. Hüner, H. Ősterle, Identification of business oriented data quality metrics, ICIQ, 2009

P. Glowalla, A. Sunyaev, Process-driven data quality management. A critical review on the application of process modeling languages, ACM Journal of Data and Information Quality, August 2014