Data intensive flows

Knowledge objectives

1. Recognize the importance of usability and taking a user-centered approach
2. Remember BPMN elements regarding flow objects, swimlanes, connections, and data artifacts

Understanding Objectives

1. Assign ETL uses to BPMN elements

Application Objectives

1. Given a description of an ETL process, model it using BPMN

User centered design

"It is users and not data that are important."

- Focus on the users
- Needed activities
  - Specify the context of use
  - Specify the user and business requirements
  - Design the product
  - Evaluate the design

Usable systems

- Effectiveness
  - Does it do the job?
- Efficiency
  - How easily does it do the job?
- Satisfaction
  - How enjoyable is it to do the job?

Consistent representation

- Complete
  - The user can get all information
- Correct
  - The user cannot derive any other information

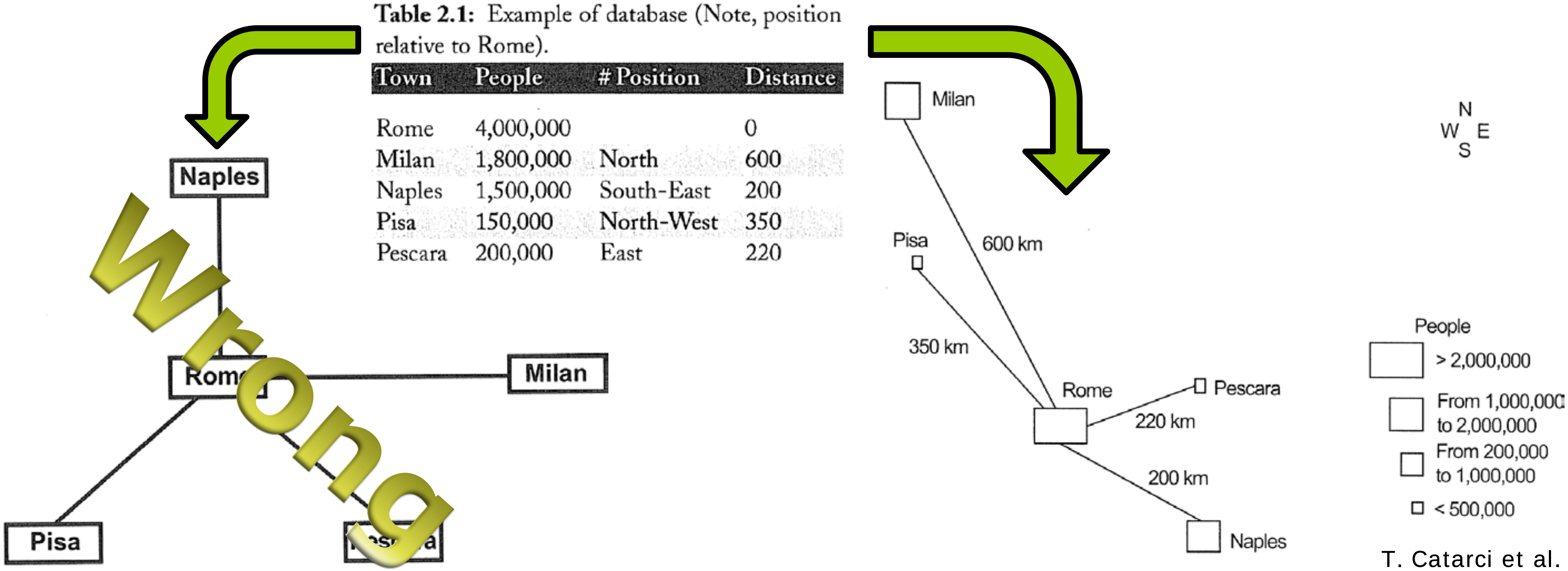
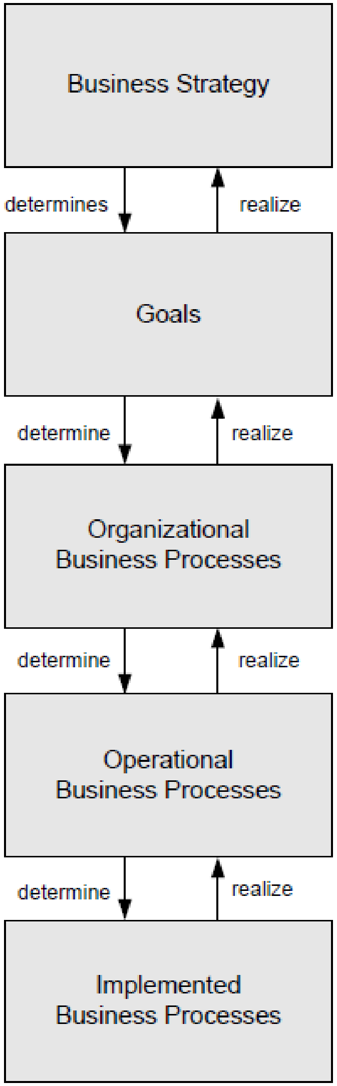


Table 2.1: Example of database (Note, position relative to Rome).

Town	People	# Position	Distance
Rome	4,000,000		0
Milan	1,800,000	North	600
Naples	1,500,000	South-East	200
Pisa	150,000	North-West	350
Pescara	200,000	East	220

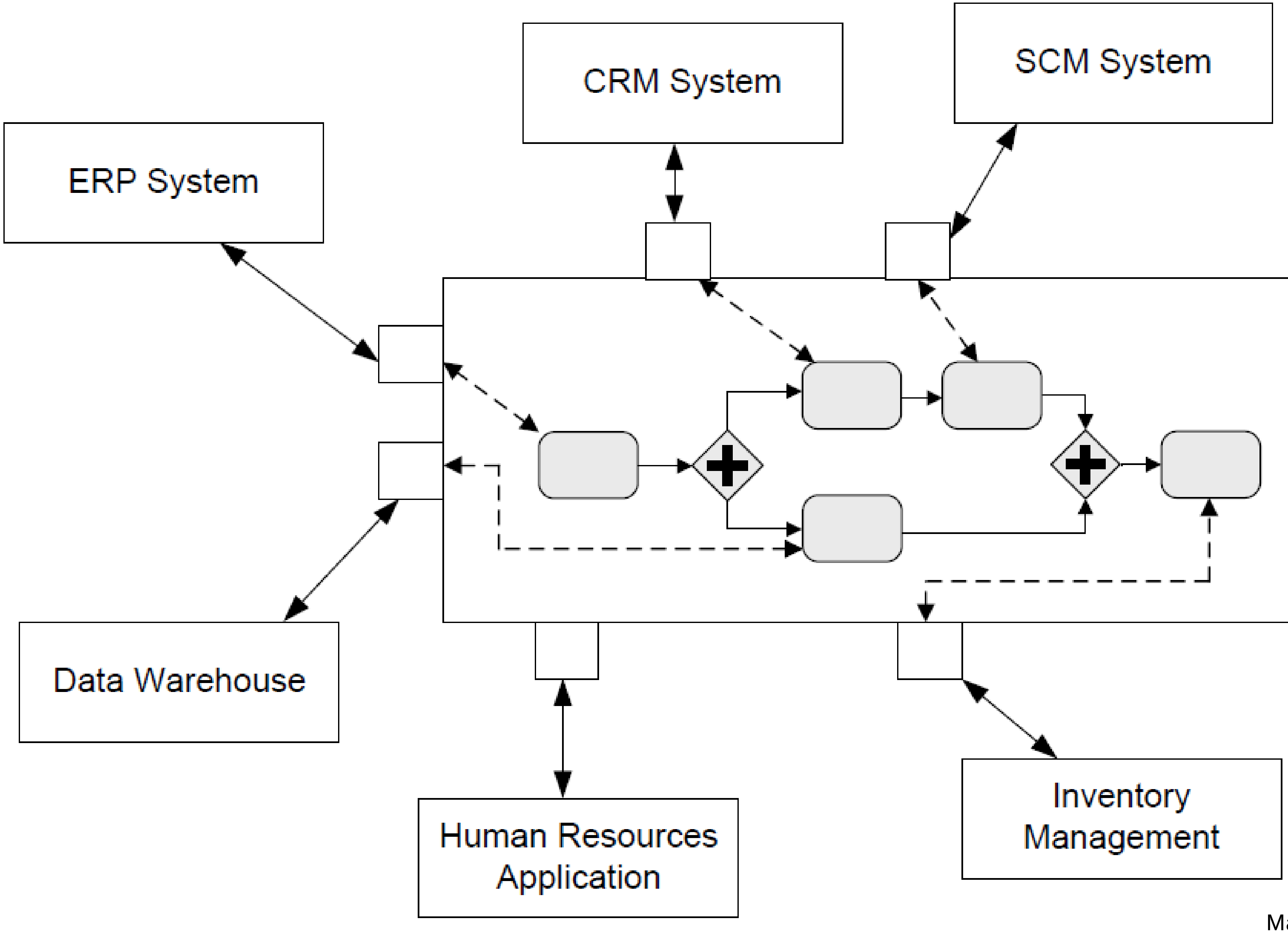
T. Catarci et al.

Classification of business processes


Business Strategy
determines
realize
Goals
determine
realize
Organizational Business Processes
determine
realize
Operational Business Processes
determine
realize
Implemented Business Processes

Mathias Weske

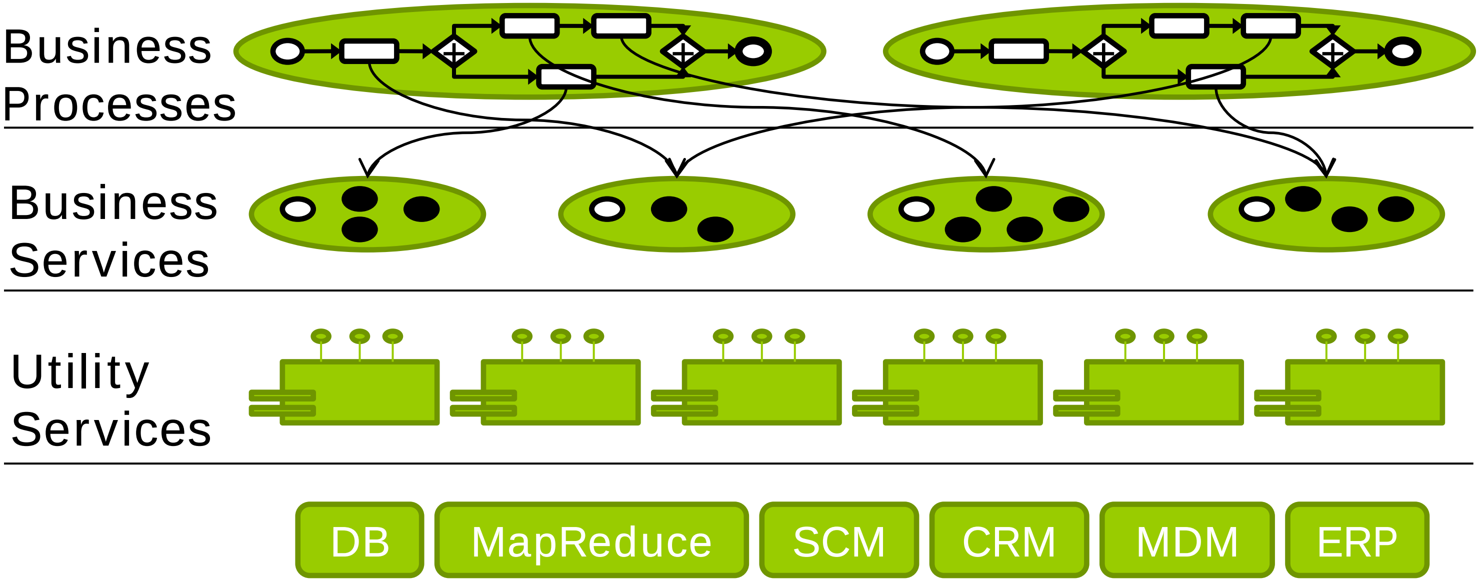
Workflow management system



Mathias Weske

Technological Challenges

- Business Process Management
- Service Composition
- Service Infrastructure and Management



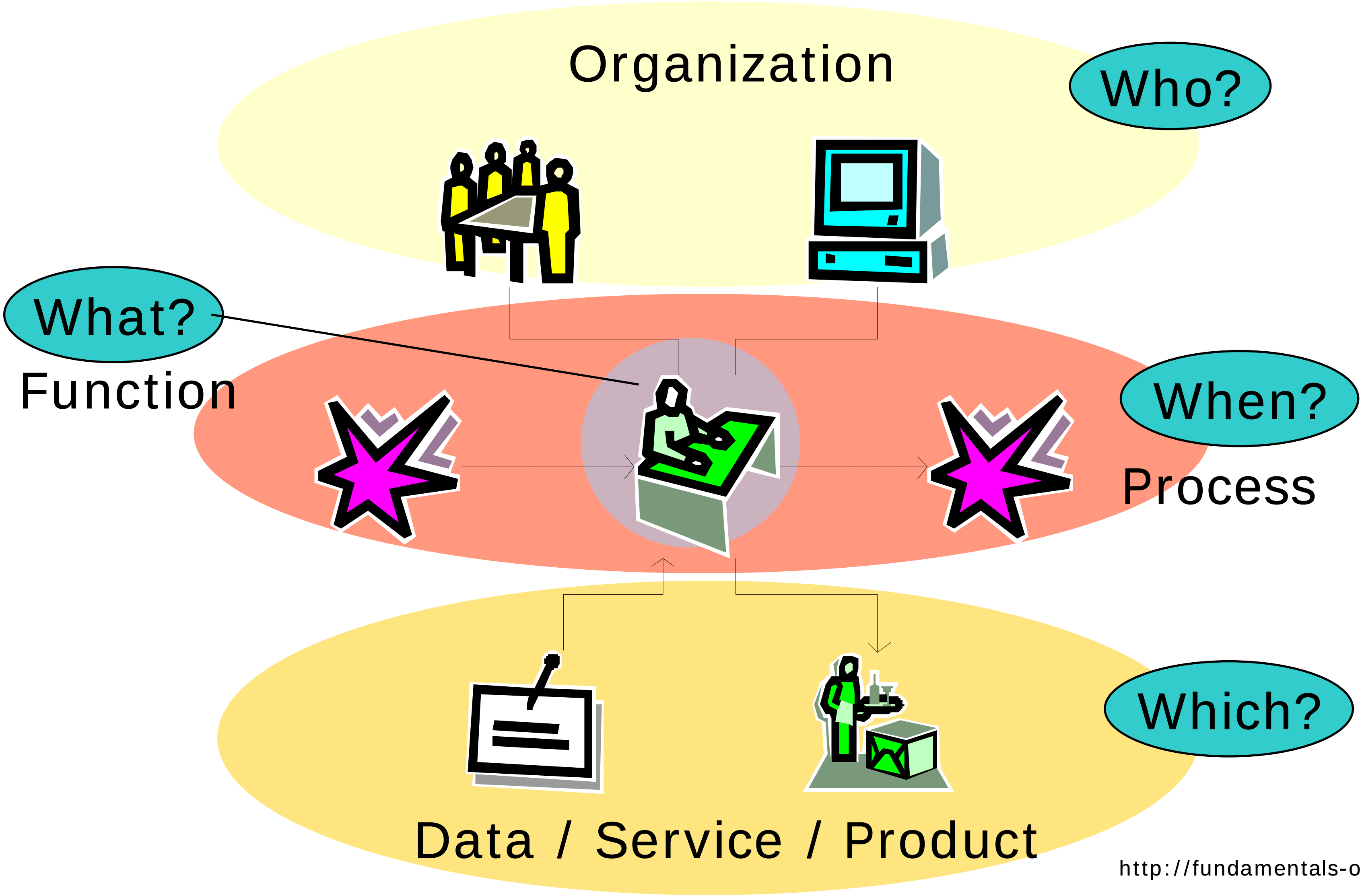
Comparison between ETL and BPM

- Benefits of treating ETL as a type of process
  - Provide an abstract view (implementation independent)
  - Monitor and report in terms of the abstract view
- ETL is batch oriented, while BPM is event oriented
  - We can also consider pipelining ETL
    - This is more appropriate for streaming

ETL operations

- Extraction
- Schema modification
  - Projection
  - Field splitters
  - Attribute addition
- Aggregation
- Value derivation
  - Value mapper
  - Lookups
  - String processing
  - Scripting
  - Cryptography
- Dataset alteration
  - Filtering
  - Duplicate removal
  - Sampling
- External calls
  - Check for existence
  - Send e-mail
  - Write to log
- Others
  - Delay row
  - Blocking step
  - Abort

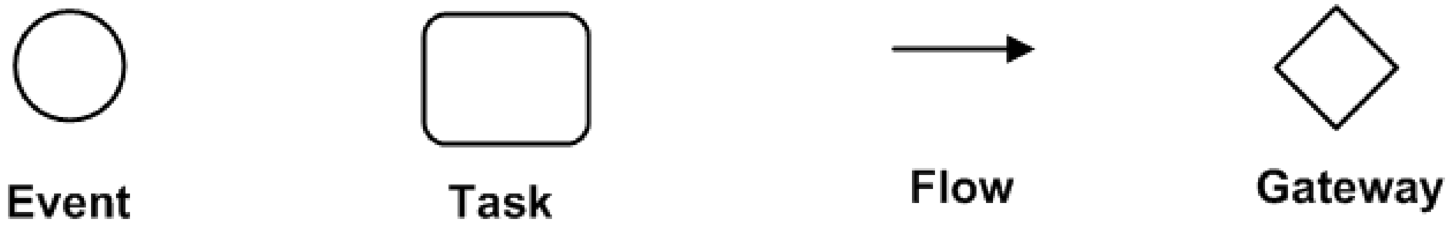
Process Modelling Viewpoints



http://fundamentals-of-bpm.org

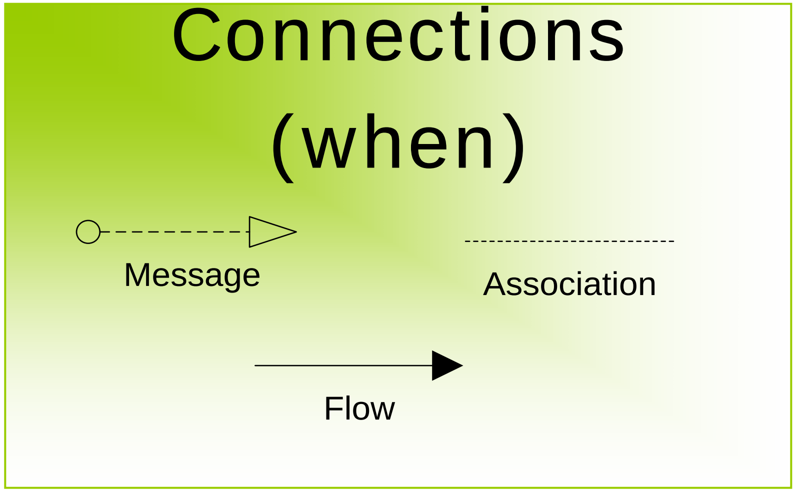
BPMN idea

A BPMN process model is a graph consisting of four types of elements (among others):

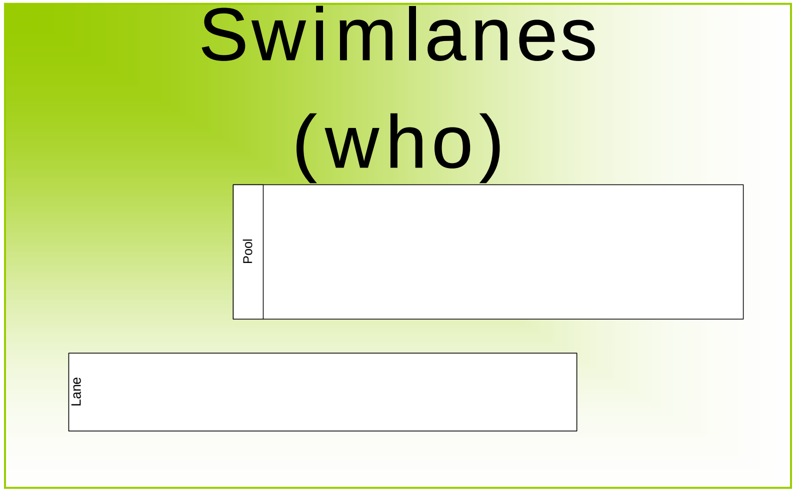


http://fundamentals-of-bpm.org

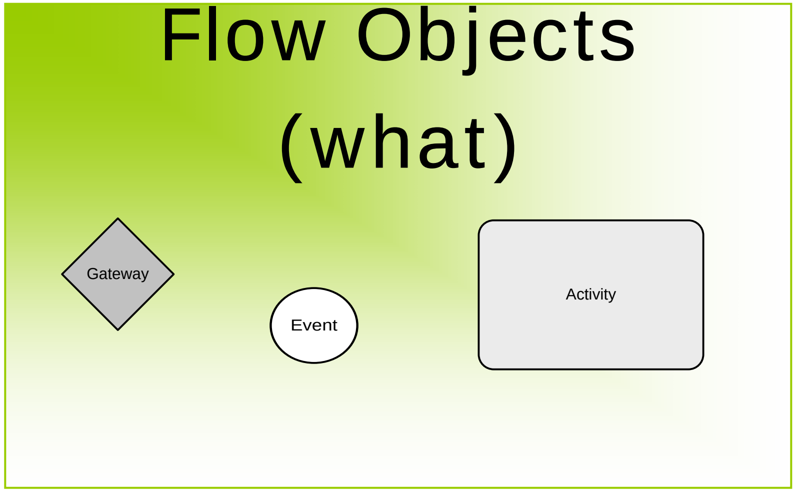
BPMN main elements


Connections

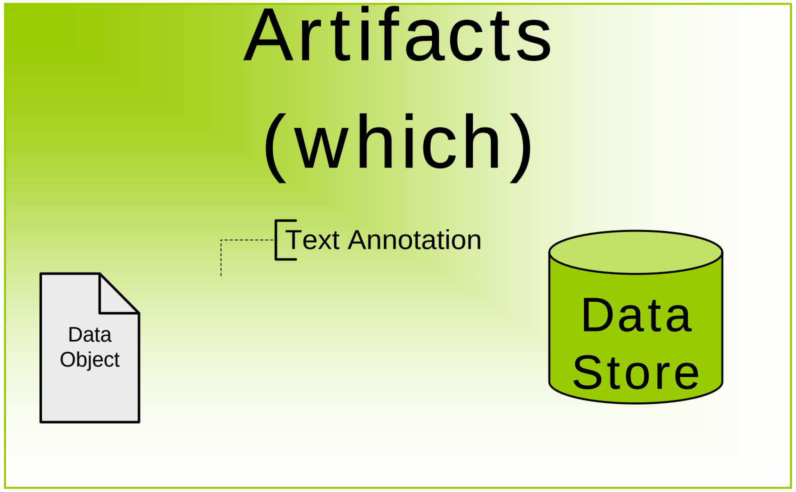
(when)


Swimlanes

(who)


Flow Objects

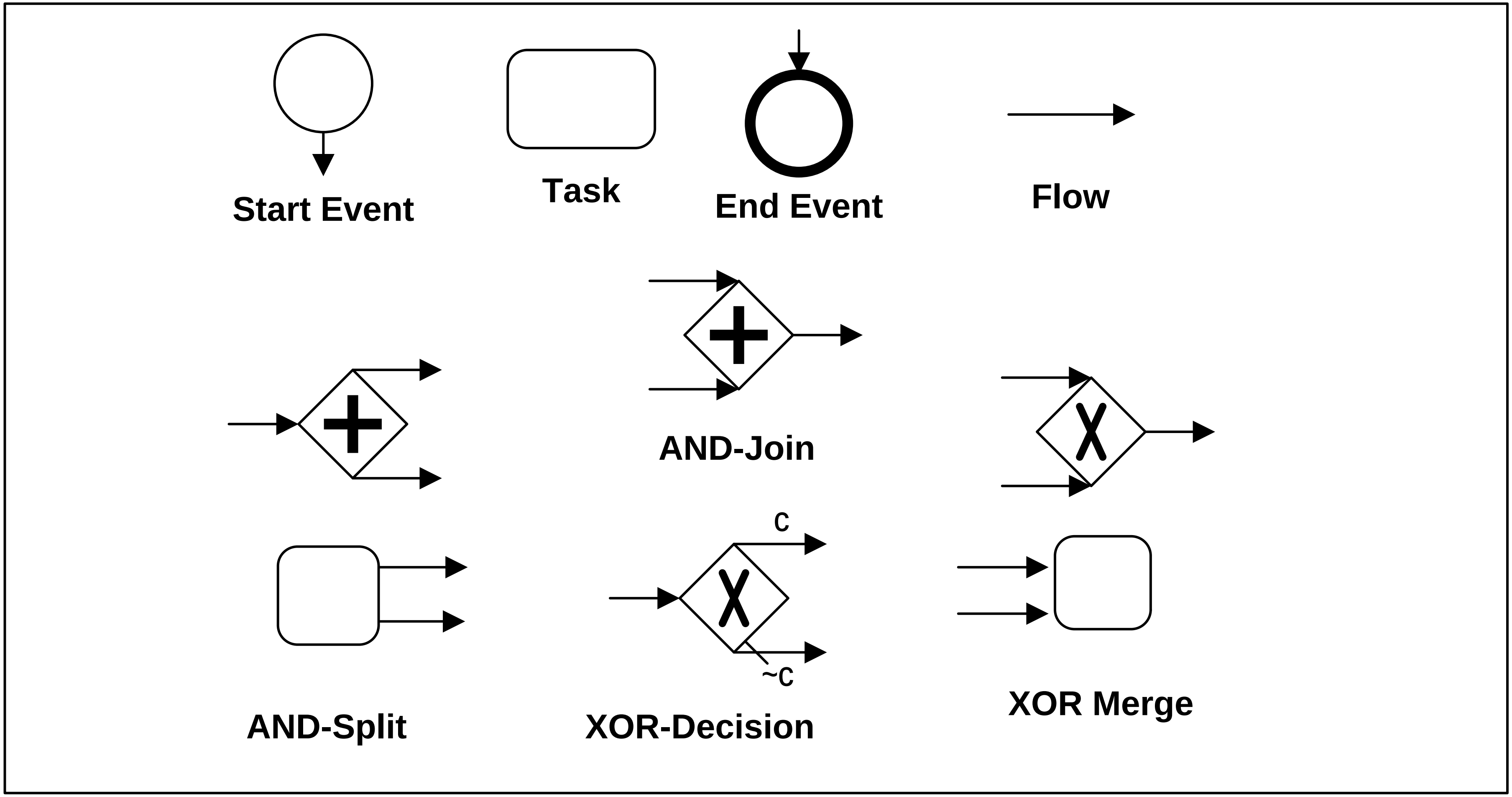
(what)


Artifacts

(which)

http://fundamentals-of-bpm.org

Flow elements

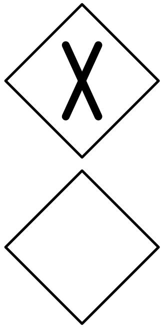


http://fundamentals-of-bpm.org/

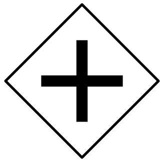
Data intensive flows

Gateways

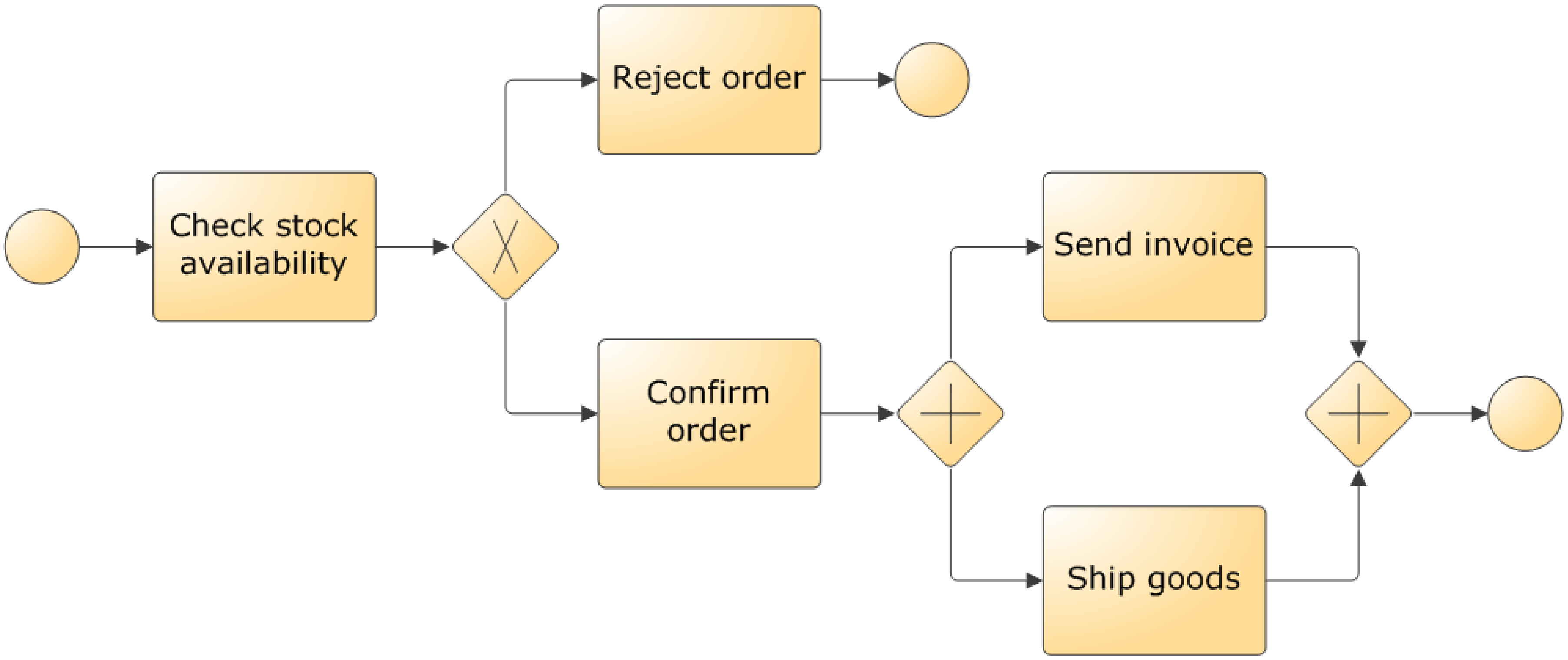
- Exclusive Decision / Merge
  - Indicates locations within a business process where the sequence flow can take two or more alternative paths
  - **Only one** of the paths can be taken



- Parallel Fork / Join
  - Provide a mechanism to synchronize parallel flow and to create parallel flow
  - Depicted by a diamond shape that *must* contain a marker that is shaped like a plus sign



Example of gateways



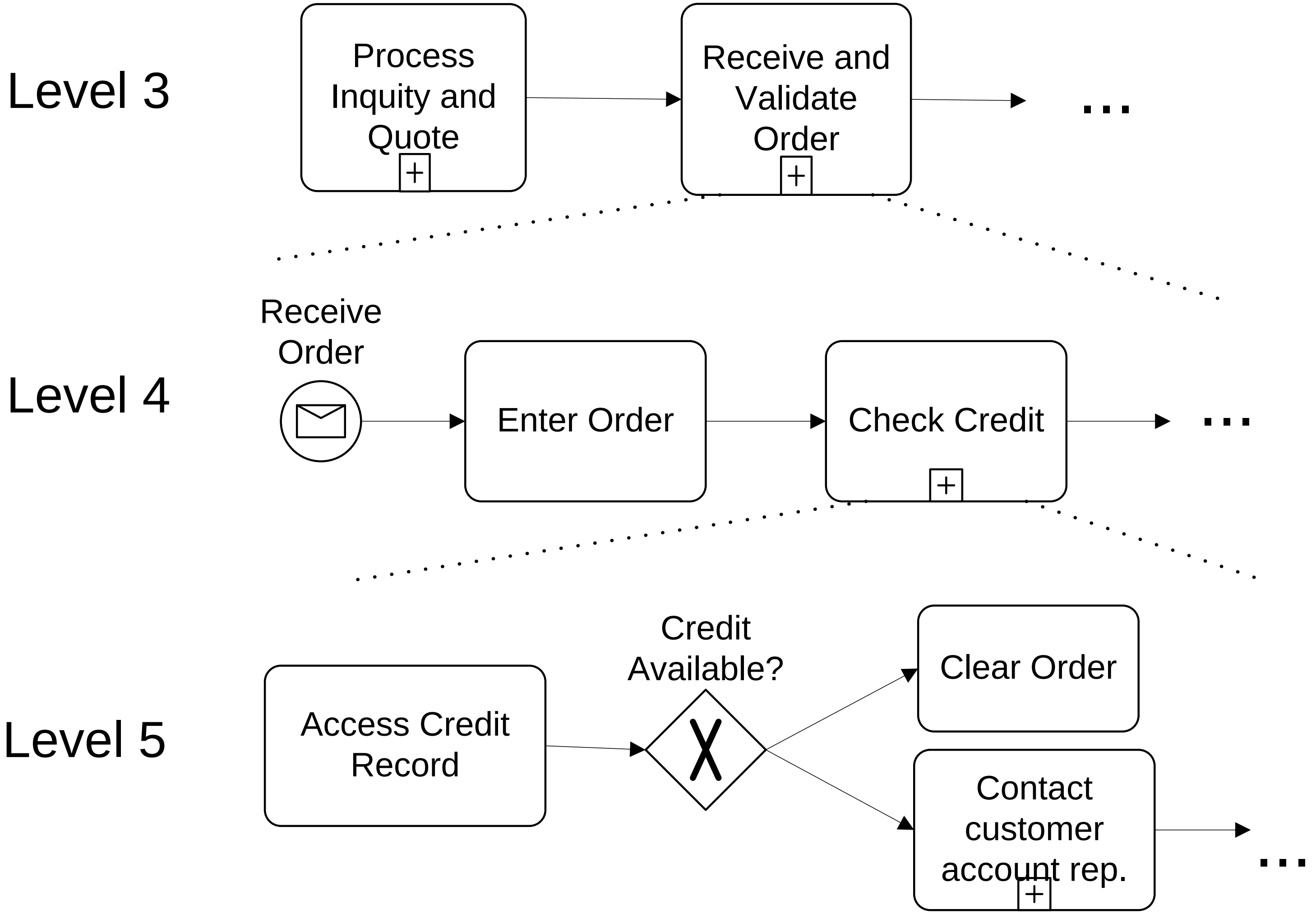
http://fundamentals-of-bpm.org

Sub-processes

- An activity in a process can "invoke" a separate (sub-)process
- Use this feature to:
  1. Break down large models into smaller ones, making them easier to understand and maintain
     → process hierarchies
  2. Share common fragments across multiple processes
     → shared subprocesses
  3. Identify parts of a process that should be:
     - Repeated
     - Executed multiple times in parallel
     - Cancelled
- Good practice is that the top-level process should be simple (no gateways) and should show the main phases of the process
  - This is sometimes called a "value chain"
  - Each phase then becomes a sub-process

http://fundamentals-of-bpm.org

Example of process hierarchies



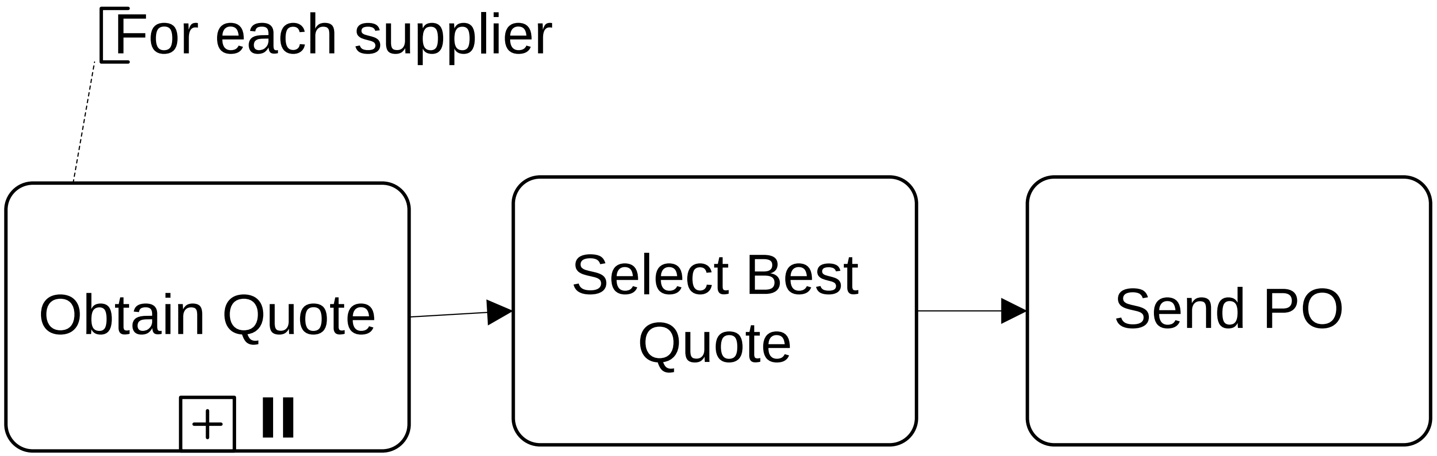
http://fundamentals-of-bpm.org

Multiple instance marker

- ≡: Sequential repetition of an activity/sub-process
- |||: Parallel repetition of an activity/sub-process
- Useful when the same activity should be executed for multiple entities or data items,
  - Examples:
    - Request quotes from multiple suppliers
    - Check the availability for each line item in an order separately
    - Send and gather questionnaires for multiple witnesses in the context of an insurance claim

http://fundamentals-of-bpm.org

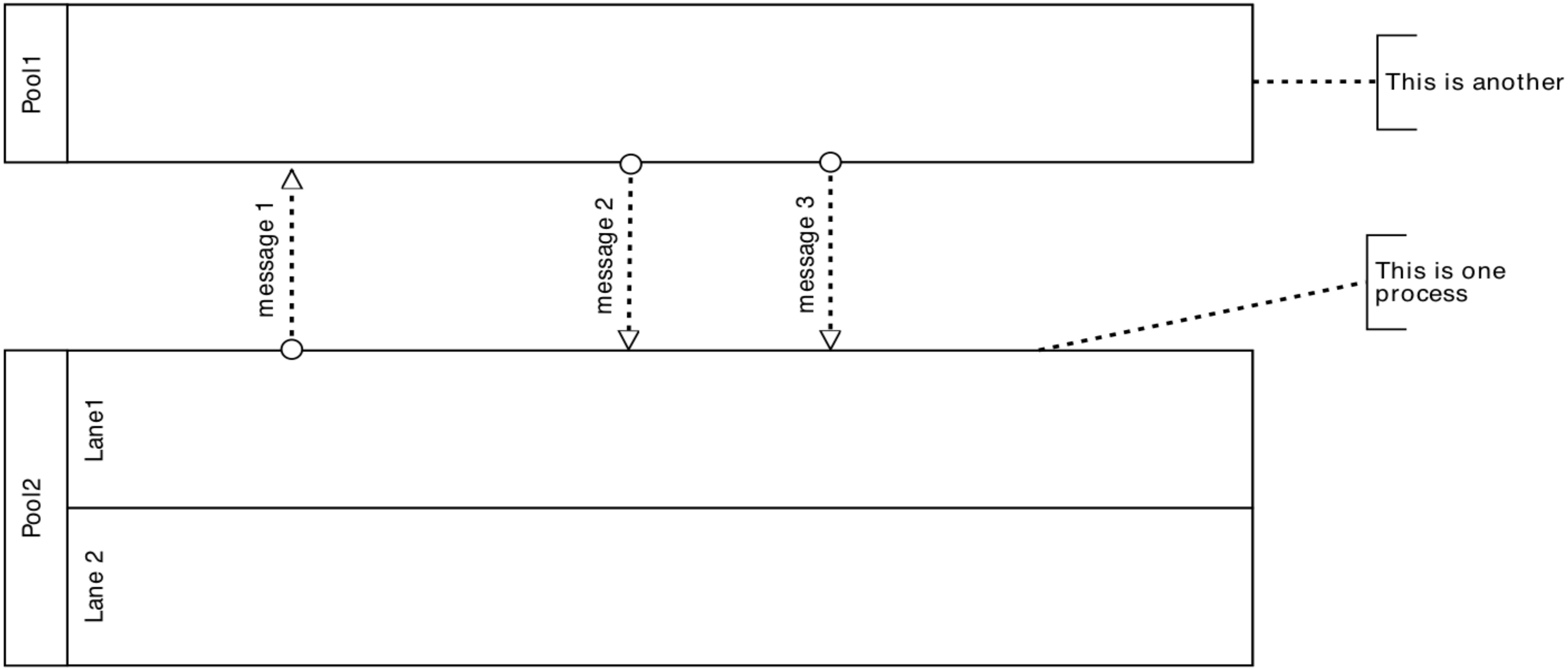
Example of Multiple instance activity



Resource elements

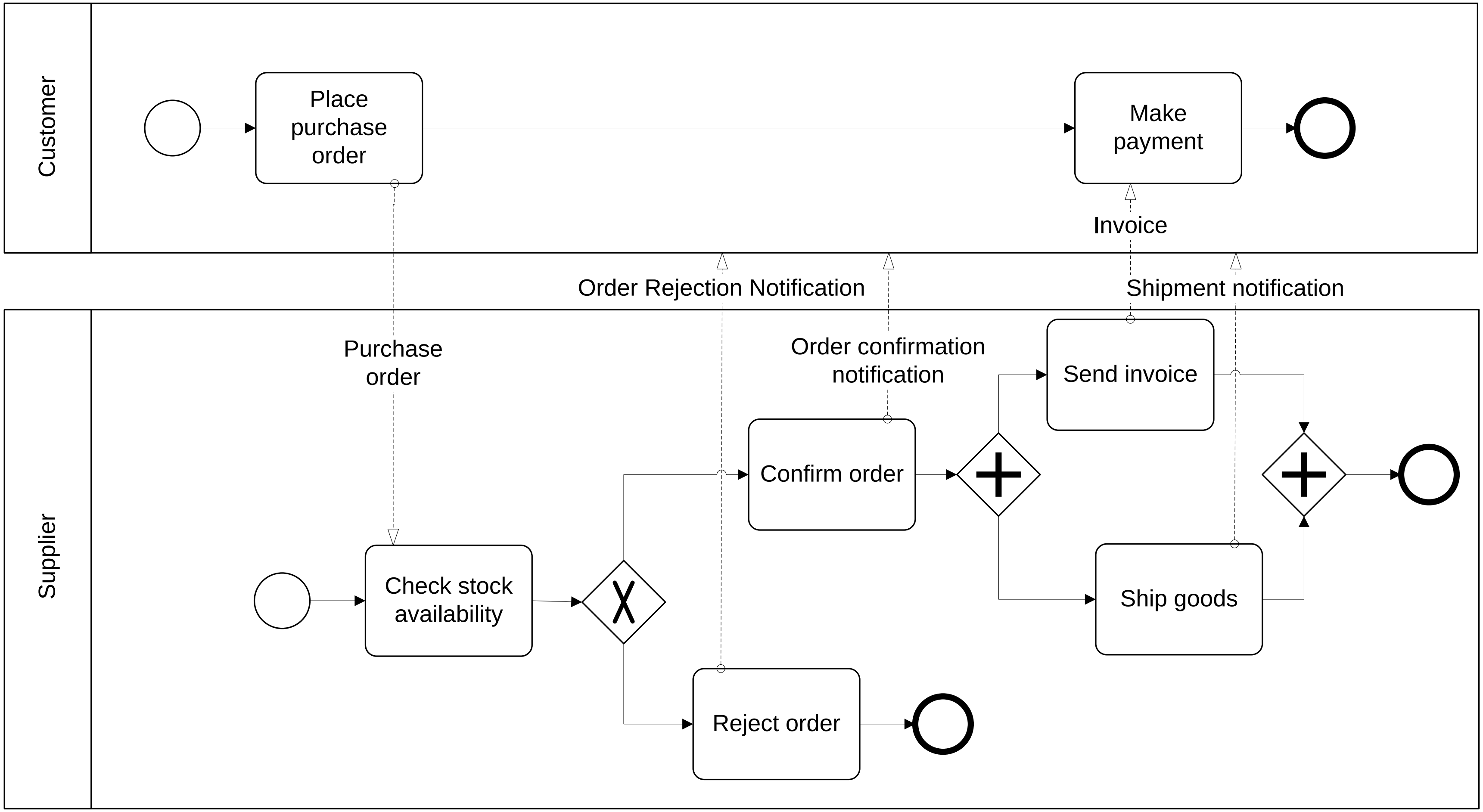
- Resource classes are captured using:
  - Pools – independent organizational entities
    - E.g., Customer, Supplier, East-Tallinn Hospital, Tartu Clinic
  - Lanes – resource classes in the same organizational space and sharing common systems
    - Sales Department, Marketing Department
    - Clerk, Manager, Engineer
- Resource class is a set of resources with shared characteristics
  - E.g., Clerk, Manager, Insurance Officer
- A *resource class* may be a
  - Role (skill, competence, qualification)
    - Classification based on what a resource can do or is expected to do
  - Group (department, team, office, organizational unit)
    - Classification based on the organization's structure

Pools and Swimlanes



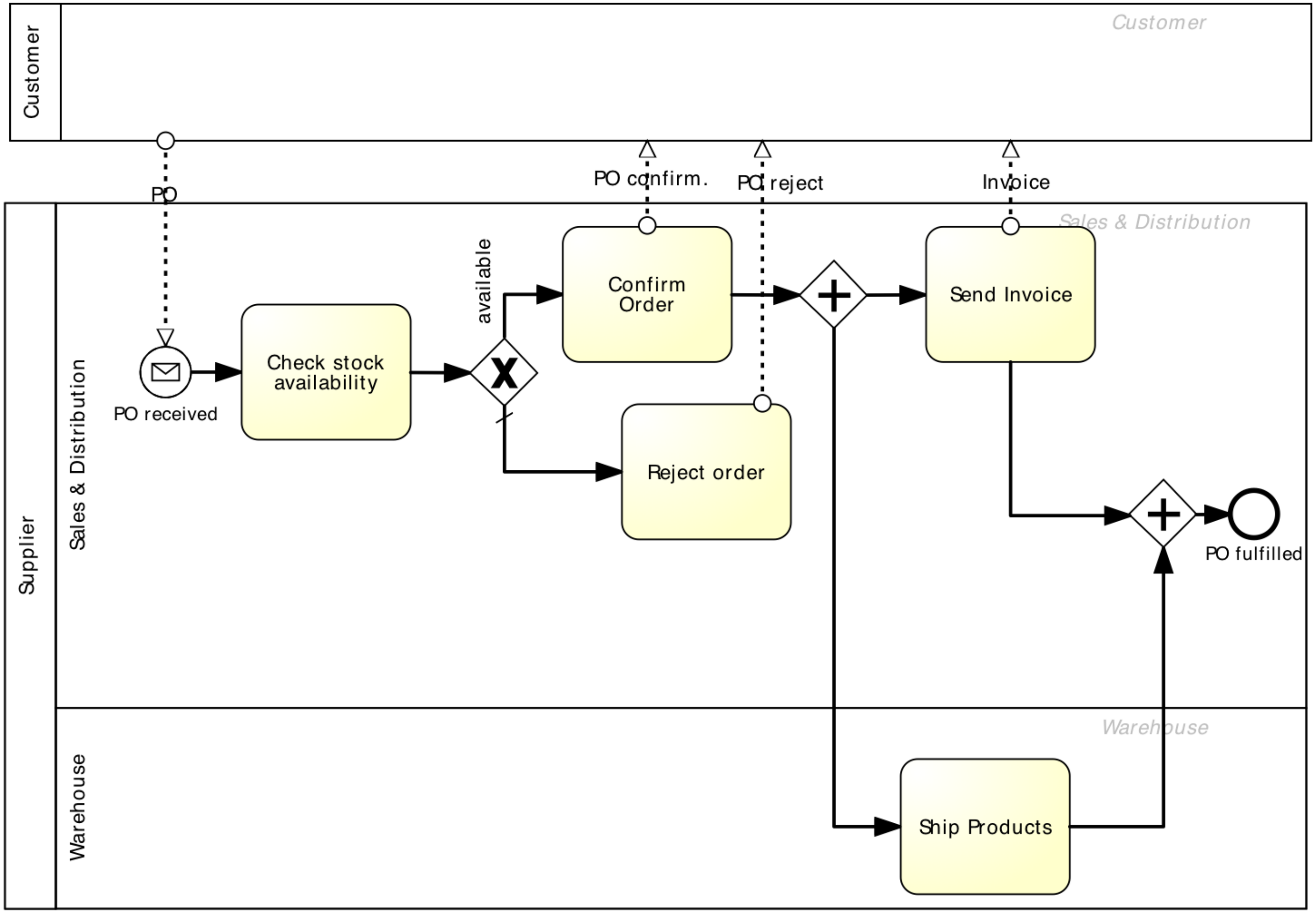
http://fundamentals-of-bpm.org

Example of Pools



http://fundamentals-of-bpm.org

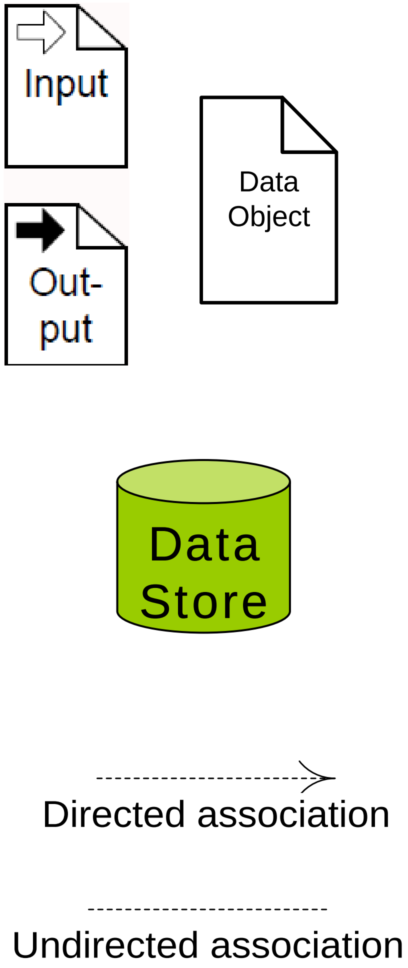
Example of Lanes



http://fundamentals-of-bpm.org

Data intensive flows

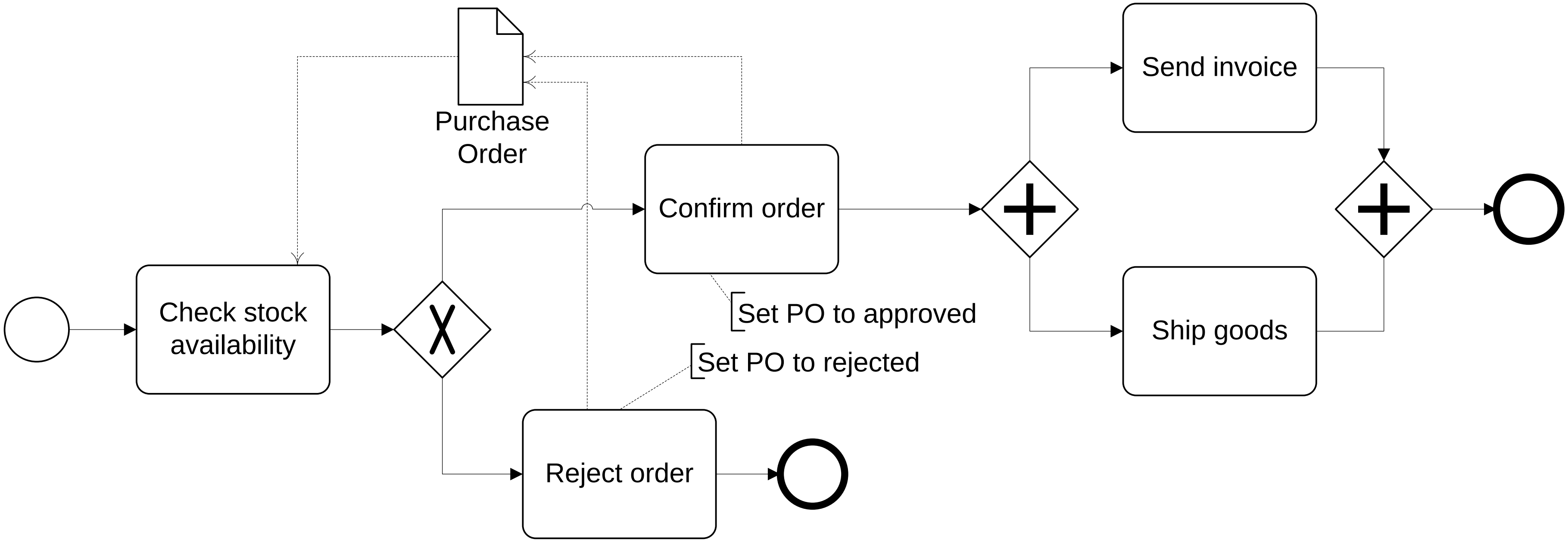
Artifacts



- Data Objects are a mechanism to show how data is required or produced by activities
  - Are depicted by a rectangle that has its upper-right corner folded over
  - Represent input and output of a process activity
- Data stores are containers of data objects that need be persisted beyond the duration of a process instance
- Associations are used to link artifacts such as data objects and data stores with flow objects (e.g., activities)

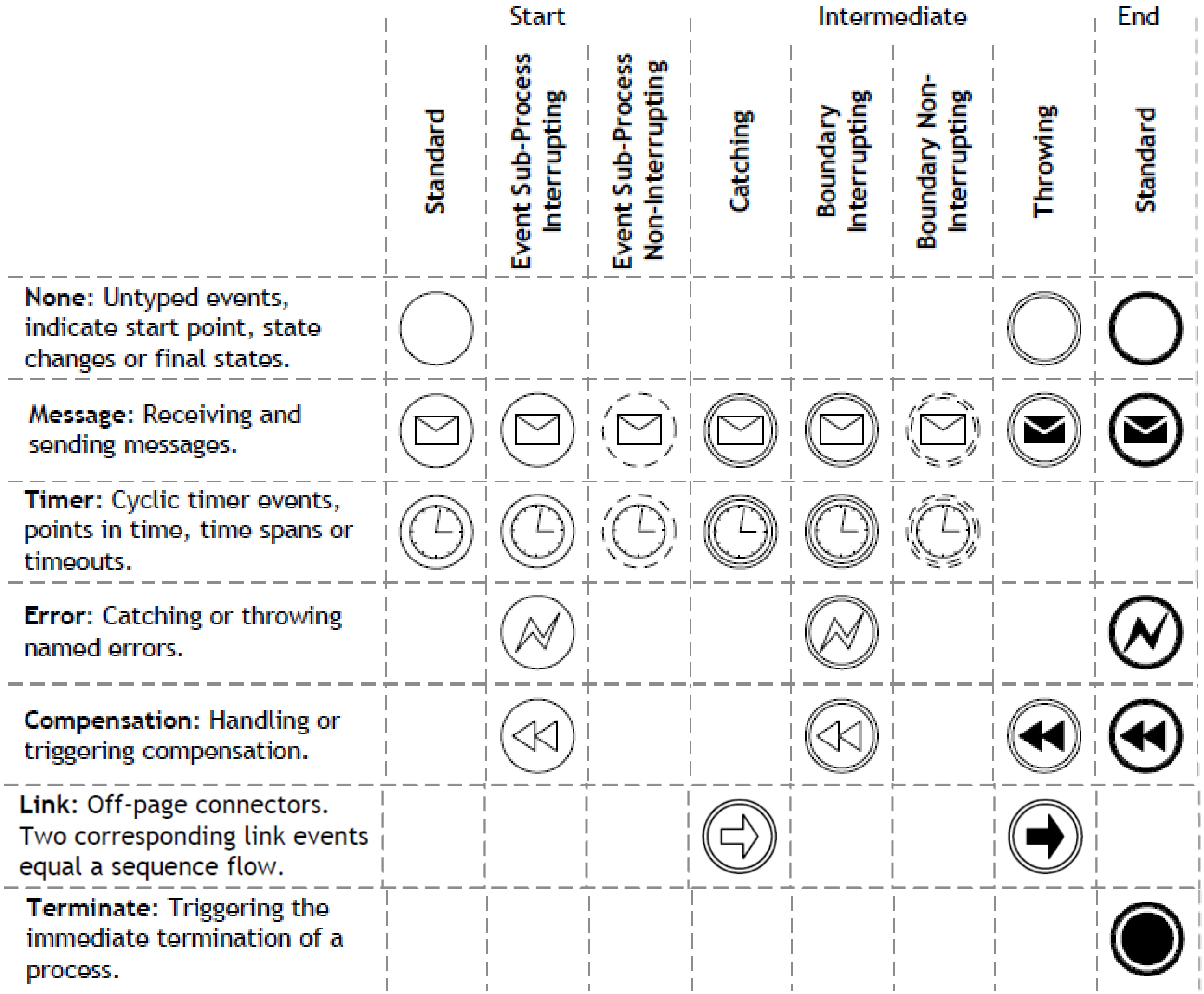
http://fundamentals-of-bpm.org/

Example of Artifacts



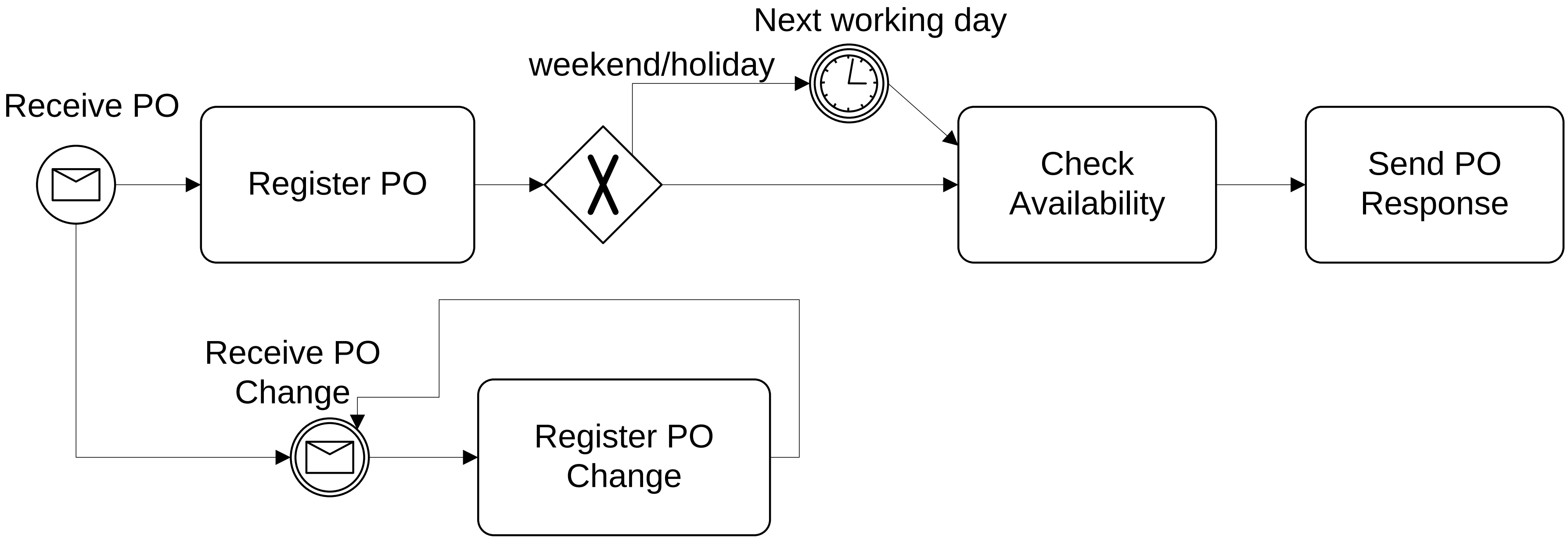
http://fundamentals-of-bpm.org/

Events



	Start			Intermediate				End
	Standard	Event Sub-Process Interrupting	Event Sub-Process Non-Interrupting	Catching	Boundary Interrupting	Boundary Non-Interrupting	Throwing	Standard
None: Untyped events, indicate start point, state changes or final states.	○						◎	○
Message: Receiving and sending messages.	✉	✉	✉	✉	✉	✉	✉	✉
Timer: Cyclic timer events, points in time, time spans or timeouts.	⏲	⏲	⏲	⏲	⏲	⏲		
Error: Catching or throwing named errors.		⚡			⚡			⚡
Compensation: Handling or triggering compensation.		⏪			⏪		⏪	⏪
Link: Off-page connectors. Two corresponding link events equal a sequence flow.				⇨			➡	
Terminate: Triggering the immediate termination of a process.								●

Example of Events

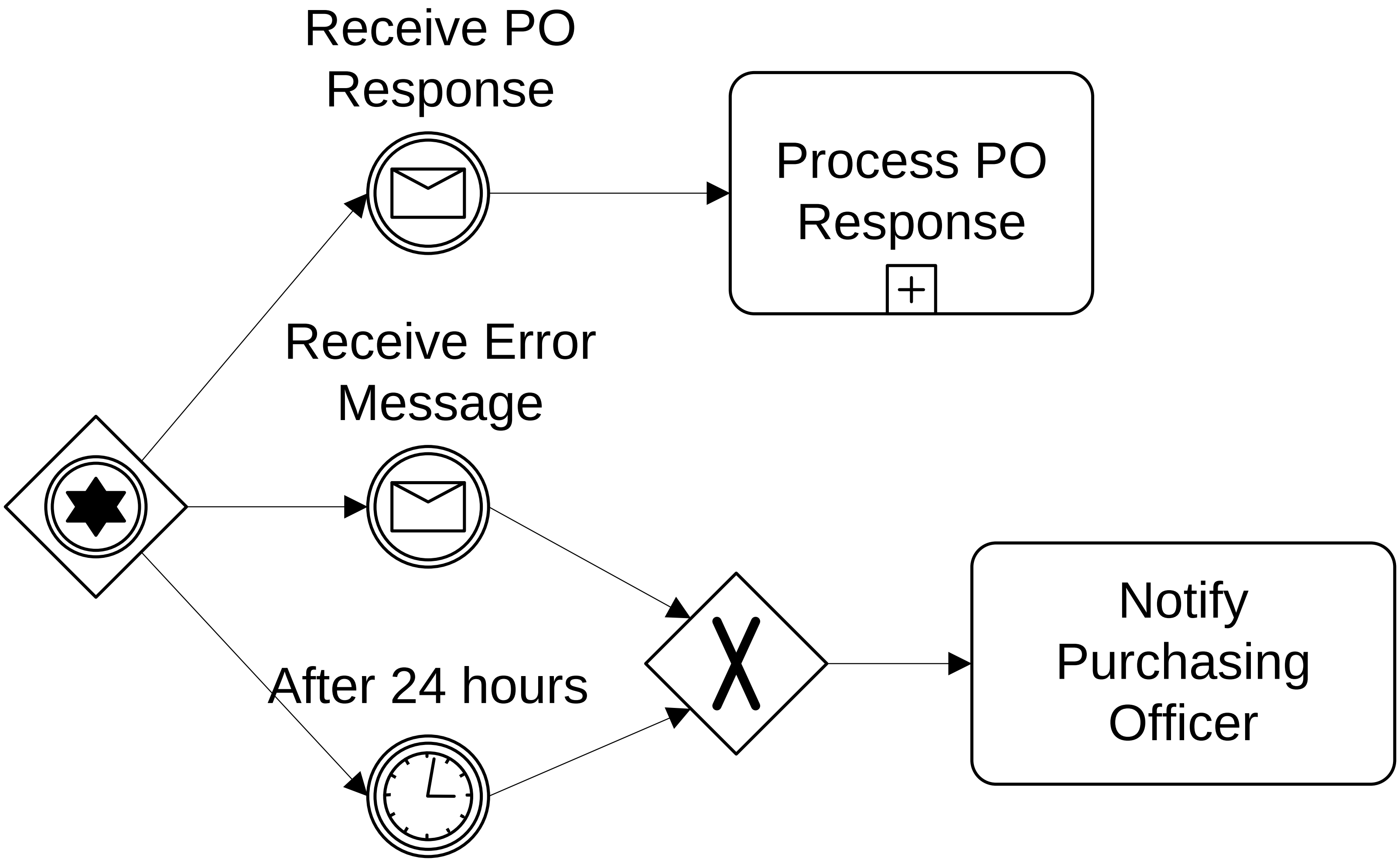


http://fundamentals-of-bpm.org

Data-based vs. event-based decision

- In an XOR-split gateway, one branch is chosen based on expressions evaluated over available data
  - → Choice is made immediately when the gateway is reached
- Sometimes, the choice must be delayed until something happens
  - → Choice is based on a “race between events”
- BPMN distinguishes between:
  - Exclusive decision gateway (XOR-split)
  - Event-based decision gateway

Example of Event-based Decision

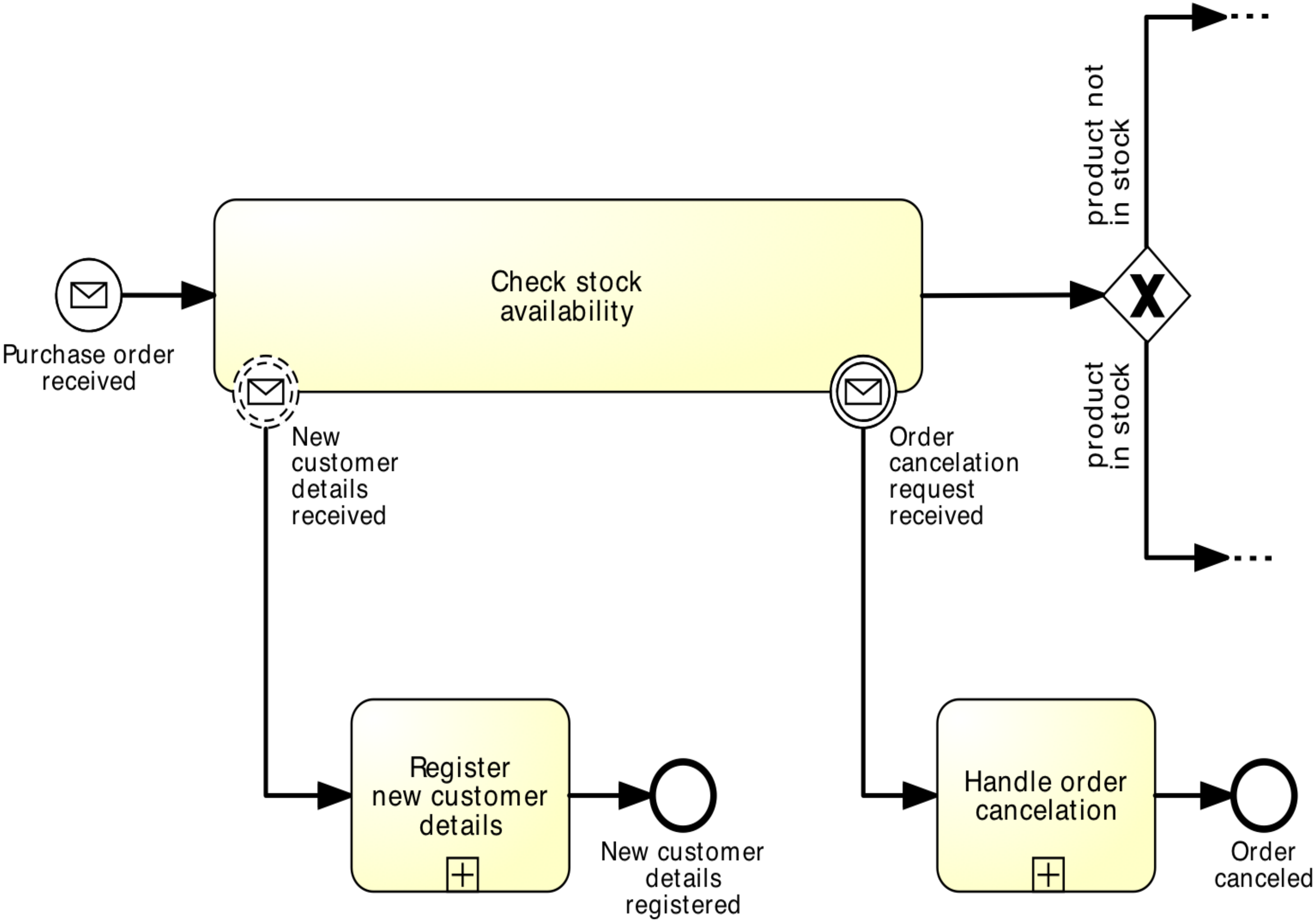


http://fundamentals-of-bpm.org

Boundary events

- Sometimes during a sub-process execution, some event may occur that needs some action...
- Such events are placed at the boundaries of the sub-process (boundary events)
- Two flavors:
  - Interrupting boundary events
  - Non-interrupting boundary events

Boundary Events – Example



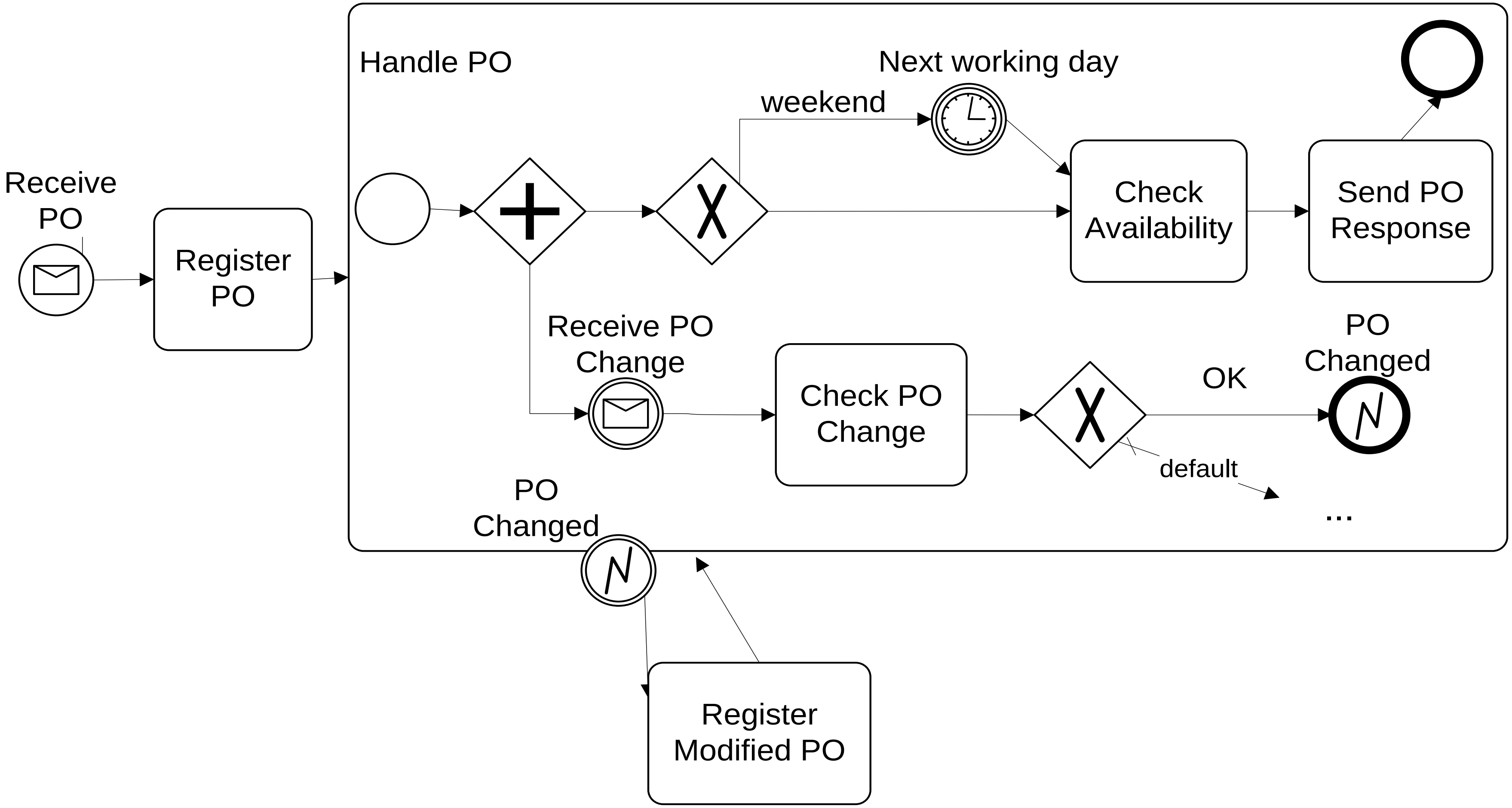
http://fundamentals-of-bpm.org

Exception handling (error events)

- Exceptions are events that deviate a process from its "normal" course
- Handling exceptions often involves stopping a sub-process and performing a special activity
- Achieved using two event nodes:
  - An "end error event" that stops the enclosing subprocess execution
  - An "intermediate error event" attached to the enclosing subprocess – this is where the process execution will continue after the error

http://fundamentals-of-bpm.org

Example of Error events

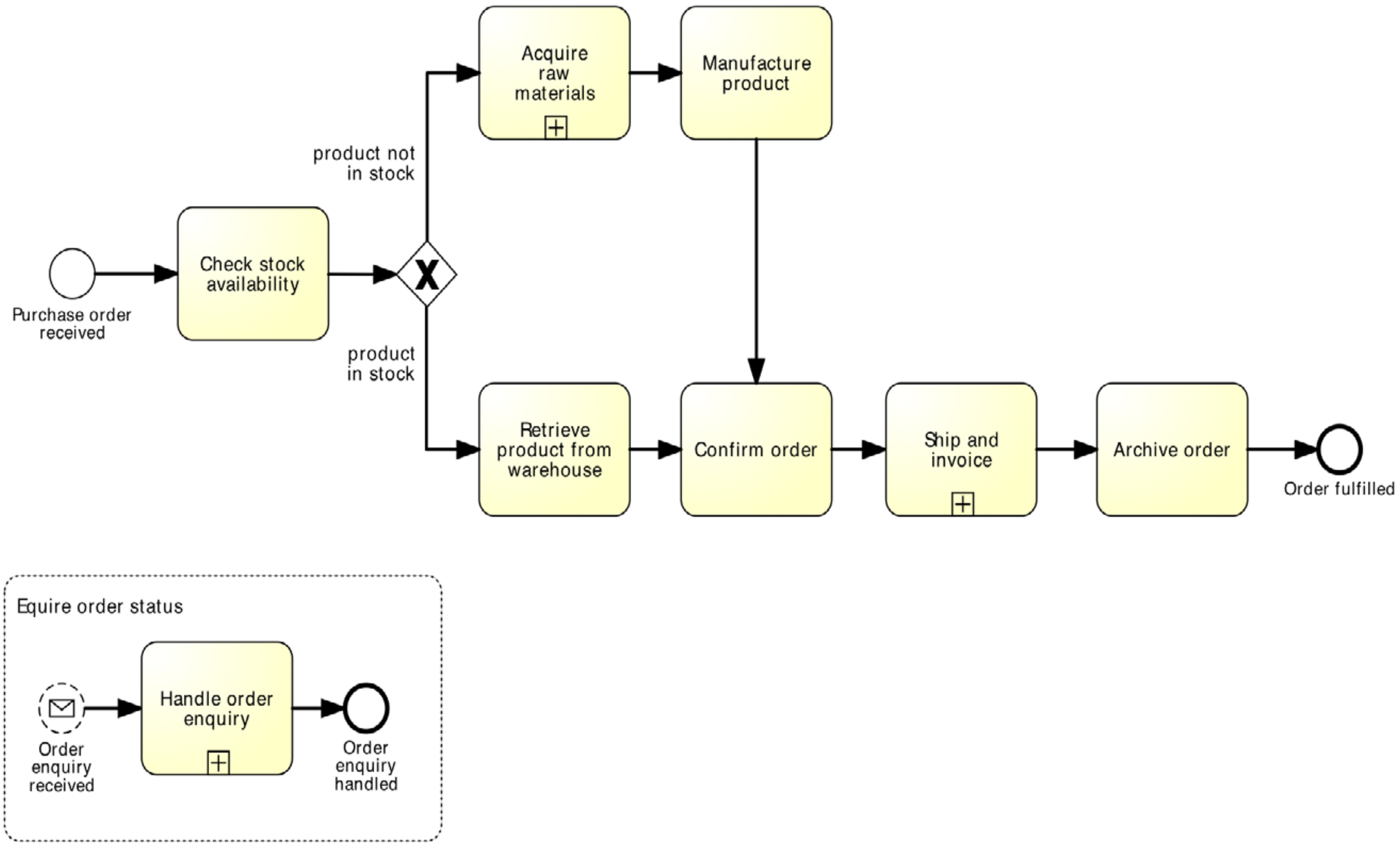


http://fundamentals-of-bpm.org

Event sub-processes

- An event sub-process are processes attached to a parent process, that are triggered when an event happens
- Alternative to putting a boundary non-interrupting event around the parent process

Example of Event sub-processes

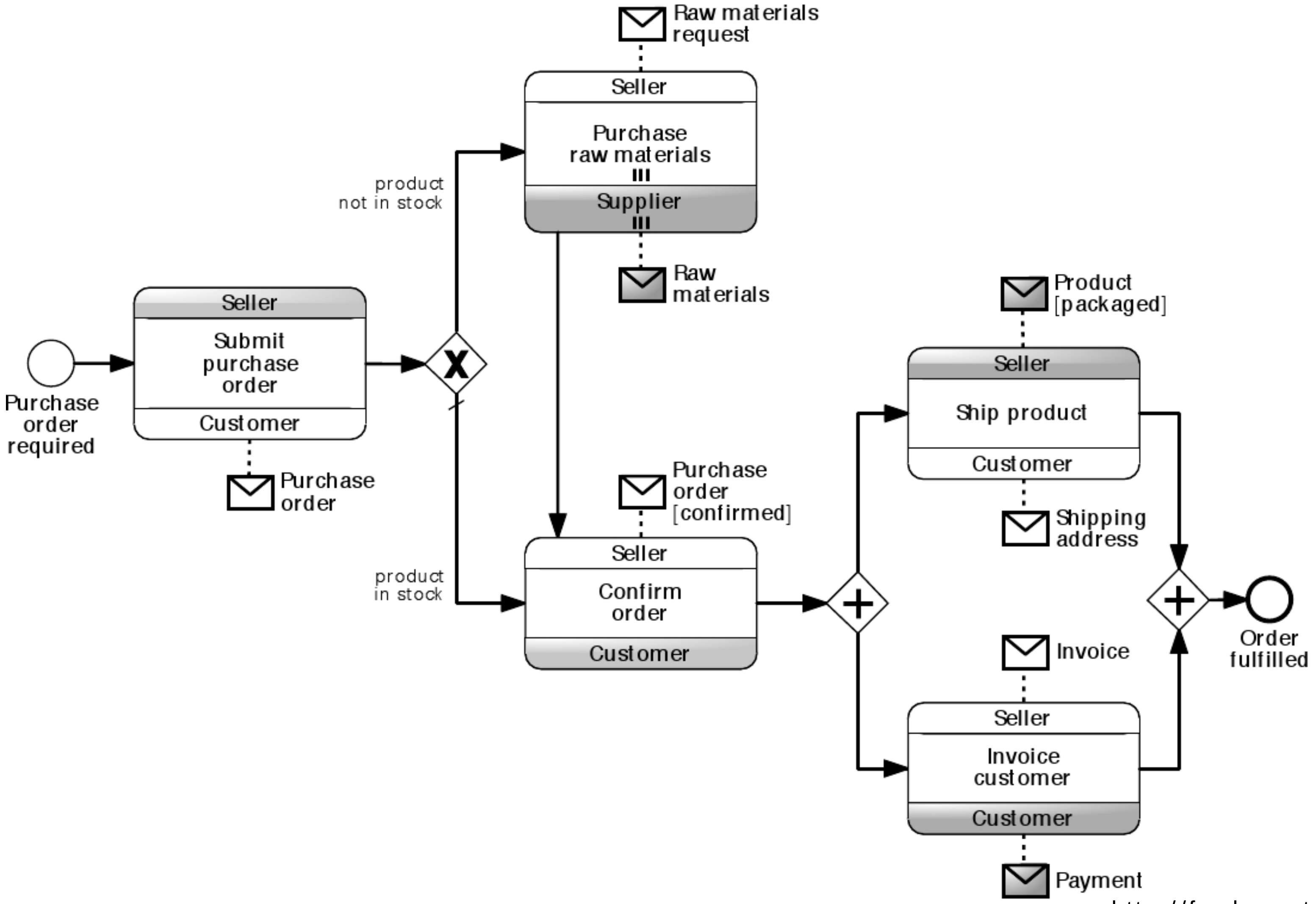


http://fundamentals-of-bpm.org

Choreographies

- Focus on interactions occurring between two or more parties
  - Each interaction involves an exchange of messages (one or more)
- Each activity element contains the information of the participants
  - Light band for the initiator
  - Dark band for the recipient

Example of choreographies



http://fundamentals-of-bpm.org

Good practices

- Hierarchical design
  a) By using BPMN levels (1&2) notation
     i. Main flow
     ii. Exception handling
  b) By drilling down activities into subprocesses
- Completeness
- Clarity (unambiguous)
- Shareability between business and IT
- Structural consistency (use standards)

Activity

- *Objective: Use BPMN to model an ETL process*
- *Tasks:*
  1. *(15') Individually draw a proposal of the corresponding ETL part*
  2. *(15') Match all three proposals in to one*
  3. *Hand in the merged proposal*
- *Roles for the team-mates during task 2:*
  - a) *Explains his/her material*
  - b) *Asks for clarification of blur concepts*
  - c) *Mediates and **controls time***

Summary

ETL	BPMN
Extraction/Load	Data store
Input/Output	Data objects
Parallelism	AND-gateway
Load balance	XOR-gateway
Complex task	Subprocess
Pipelining	Multiple instance marker
Multiple components	Swimlanes
Multiple resources	Pools
Exception handling	Error events
Compensation actions	Compensation events
Control flow	Even based decisions & Boundary events & Event subprocess

Bibliography

- T. Catarci et al. *User-Centered Data Management*. Morgan & Claypool, 2010
- R. T. Ng et al. *Perspectives on Business Intelligence*. Synthesis Lectures on Data Management. Morgan- & Claypool, 2012
- M. Weske. *Business Process Management – Concepts, Languages, Architectures*. Springer, 2007
- B. Silver. *BPMN Method & Style*. Cody-Cassidy Press, 2011 (2nd edition)
- M. Dumas et al. *Fundamentals of Business Process Management*. Springer, 2012
- A. Vaisman and E. Zimanyi. *Data Warehouse systems*. Springer, 2014
- K. Wilkinson et al. *Leveraging Business Process Models for ETL design*. ER'2010

Data intensive flows

Resources

- http://www.signavio.com
- http://www.bpmb.de/images/BPMN2_0_Poster_EN.pdf
- http://oozie.apache.org
- https://sqoop.apache.org
- http://flume.apache.org