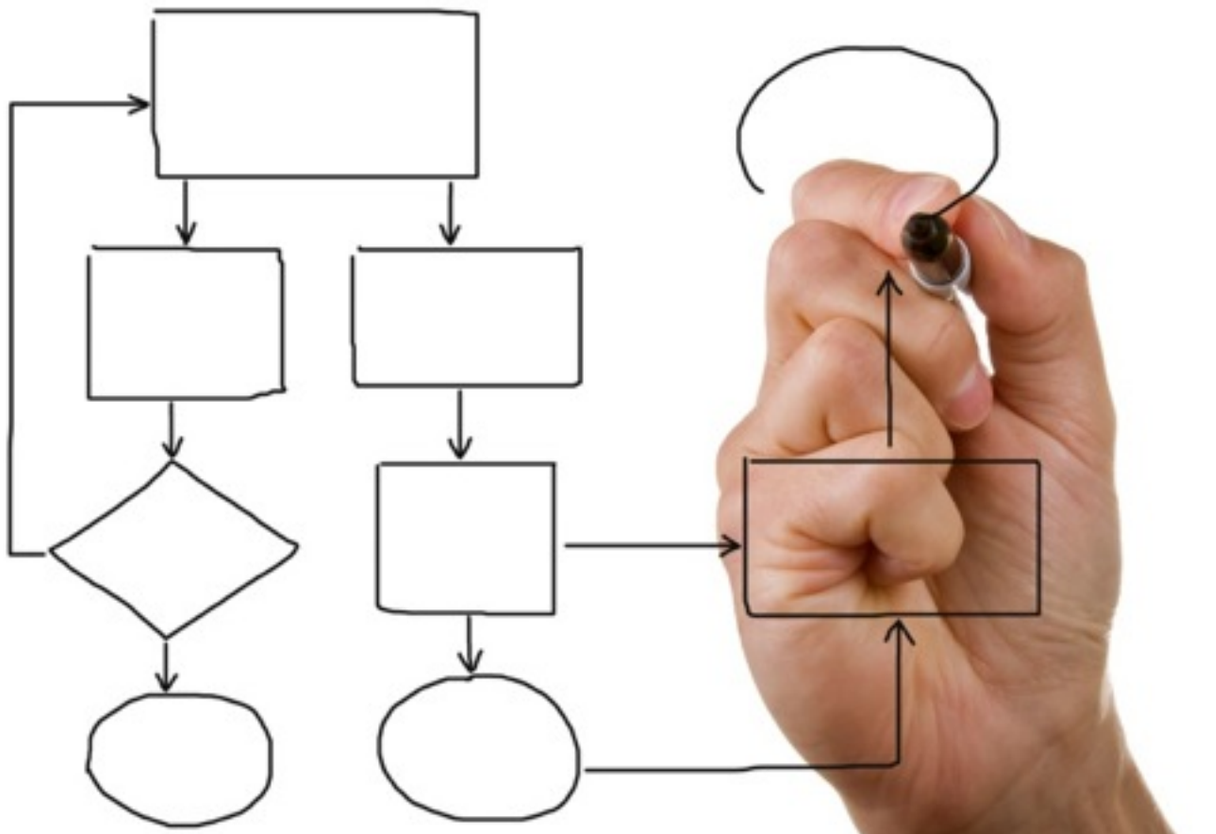
CAPS

Capturing and Managing Provenance Information in Scientific Workflows



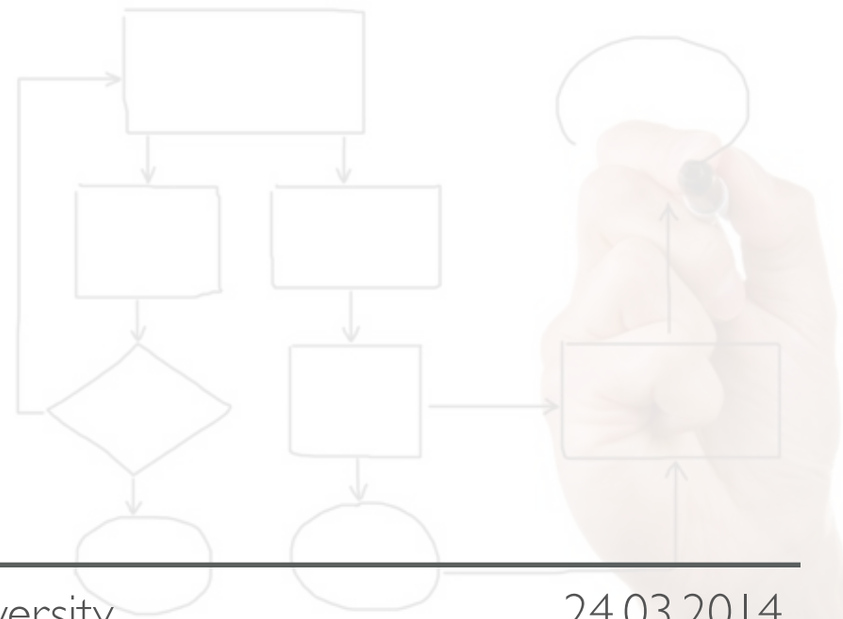
Peer Brauer

pcb@informatik.uni-kiel.de
Software Engineering Group
Kiel University

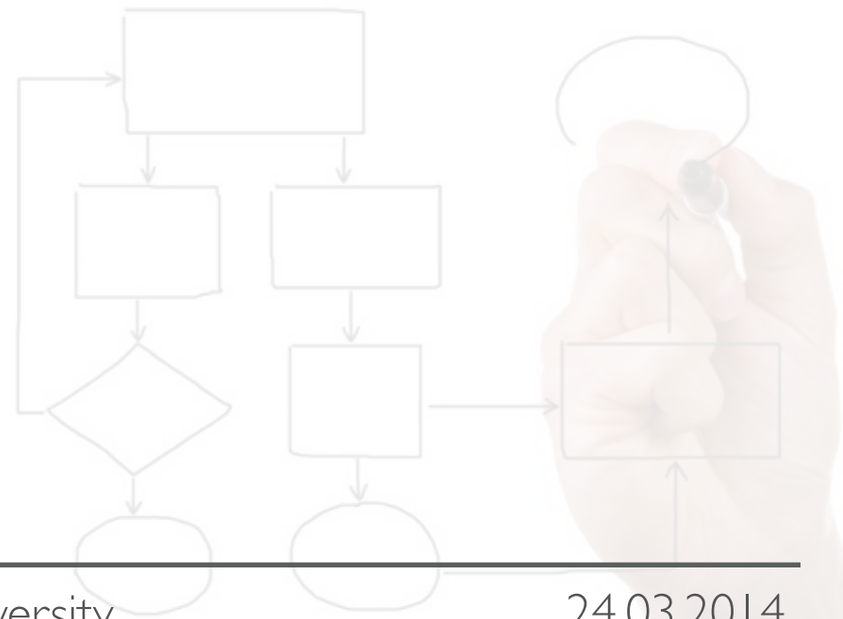
Hamburg, 24.03.2014

AGENDA

- Motivation
- Research Questions
- CAPS
- Evaluation



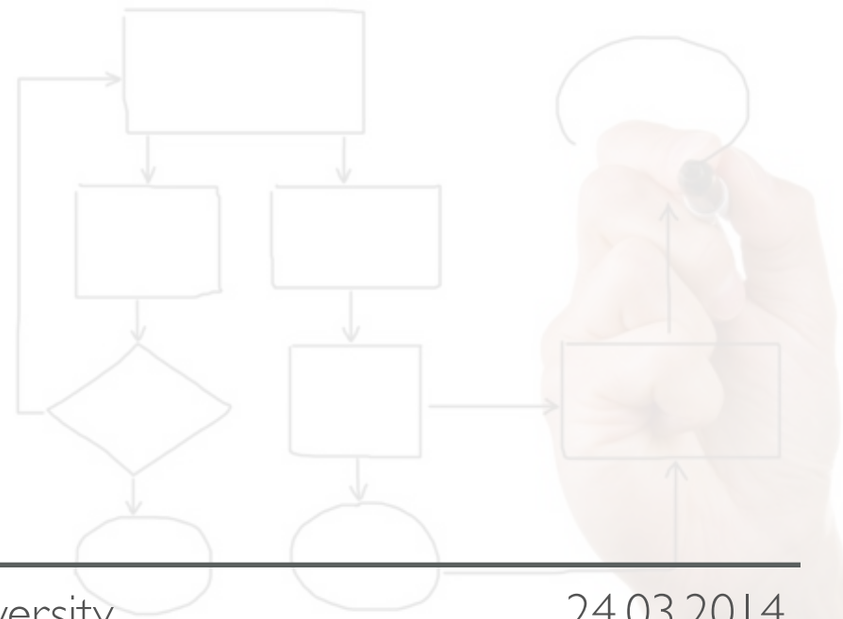
Motivation



MOTIVATION

Provenance

How was the scientific data I use
for my research created?



Coordinates

Position (RA):	9 0 12.02
Position (Dec):	-45° 57' 0.24"
Field of view:	33.16 x 33.16 arcminutes
Orientation:	Die Nordrichtung liegt 90.1° links zur Vertikale

Description

The oddly shaped Pencil Nebula (NGC 2736) is pictured in this image from ESO's La Silla Observatory in Chile. This nebula is a small part of a huge remnant left over after a supernova explosion that took place about 11 000 years ago. The image was produced by the Wide Field Imager on the MPG/ESO 2.2-metre telescope at ESO's La Silla Observatory in Chile

About

Name:	NGC 2736
Typ:	• Milky Way : Nebula : Type : Supernova Remnant • X - Nebulae
Distance:	800 Lightyears
Constellation:	Vela

Coordinates

Position (RA):	9 0 12.02
Position (Dec):	-45° 57' 0.24"
Field of view:	33.16 x 33.16 arc minutes
Orientation:	North is 90.1° left of vertical

Metadata

Id:	eso1236a
Type:	Observation
Release date:	12 September 2012, 12:00
Related releases:	eso1236
Size:	8357 x 8357 px

Description

The oddly shaped Pencil Nebula (NGC 2736) is pictured in this image from ESO's La Silla Observatory in Chile. This nebula is a small part of a huge remnant left over after a supernova explosion that took place about 11 000 years ago. The image was produced by the Wide Field Imager on the MPG/ESO 2.2-metre telescope at ESO's La Silla Observatory in Chile

About

Name:	NGC 2736
Typ:	• Milky Way : Nebula : Type : Supernova Remnant • X - Nebulae
Distance:	800 Lightyears
Constellation:	Vela

Coordinates

Position (RA):	9 0 12.02
Position (Dec):	-45° 57' 0.24"
Field of view:	33.16 x 33.16 arc minutes
Orientation:	North is 90.1° left of vertical

Metadata

Id:	eso1236a
Type:	Observation
Release date:	12 September 2012, 12:00
Related releases:	eso1236
Size:	8357 x 8357 px

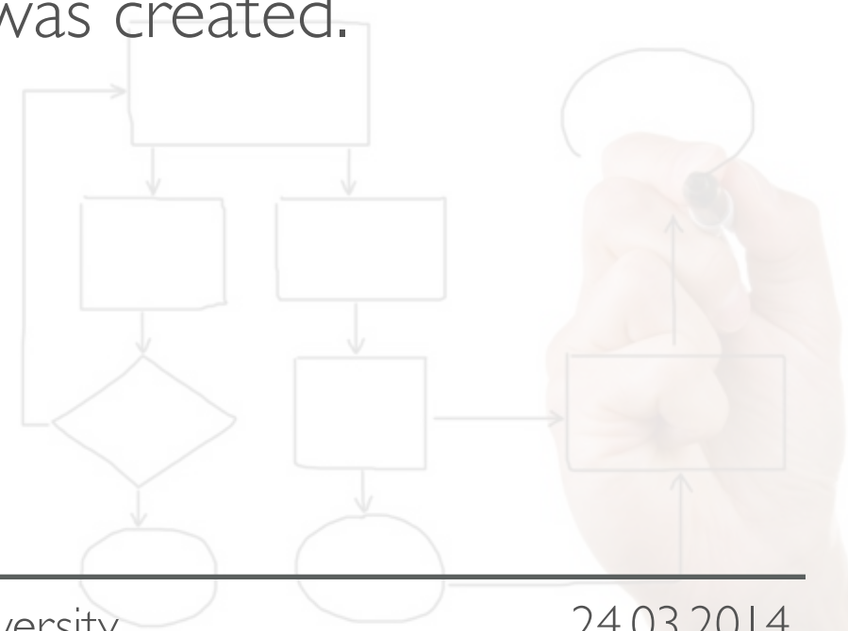
Provenance Information helps to assess (scientific) Data !

MOTIVATION

Definition Provenance

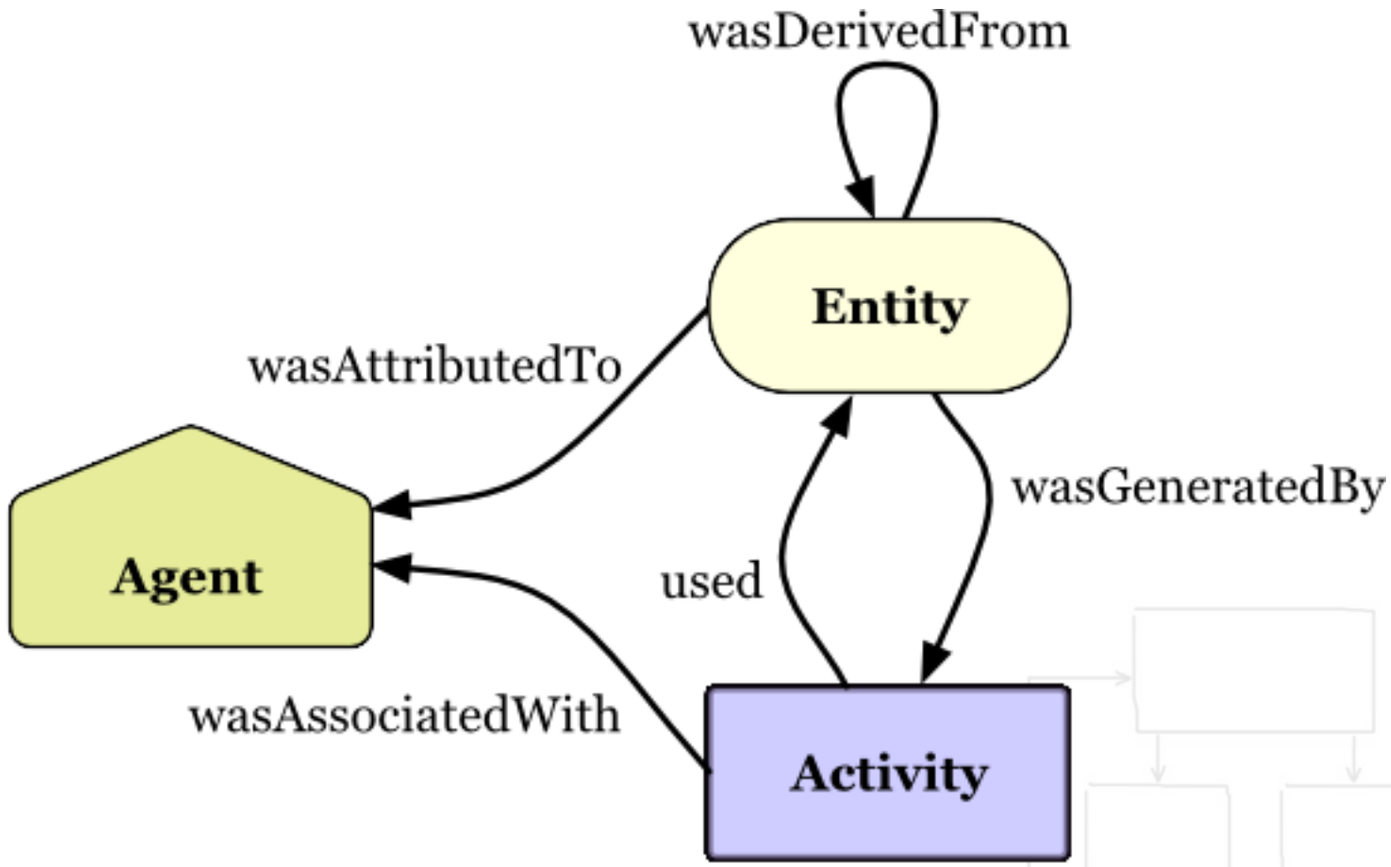
Provenance information describes how,

when, where and by whom data was created.



MOTIVATION

W3C Prov-DM



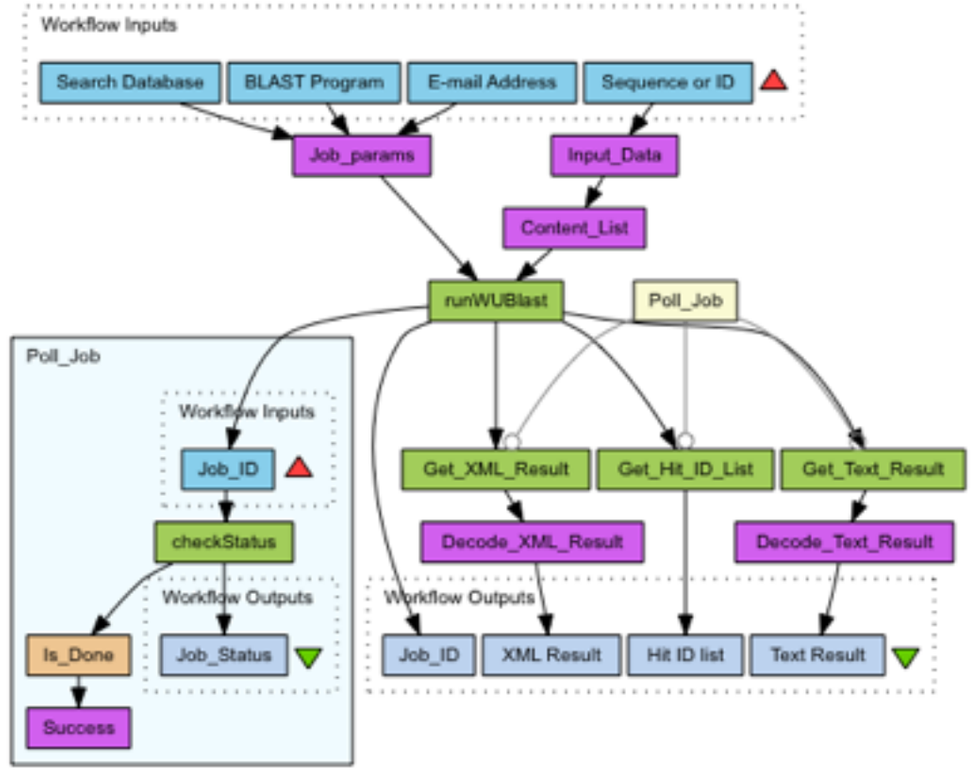
Source: http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

MOTIVATION

Scientific Workflows

Scientific Workflows …

- … automate recurrent tasks
- … are widely used in data intensive domains
- … orchestrate complex analysis or computations



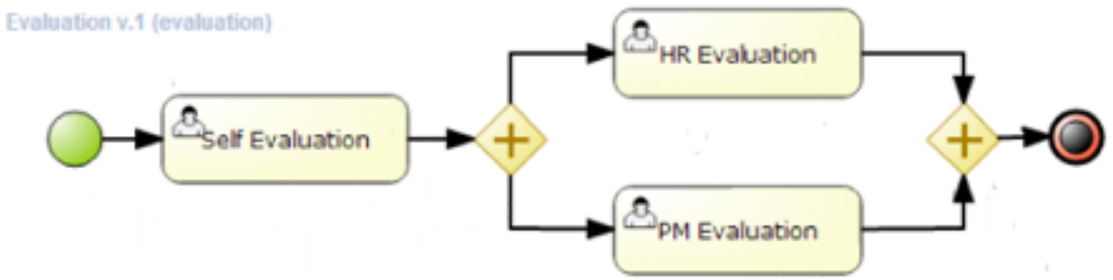
Source: www.myexperiment.org/workflows/197.html

MOTIVATION

Scientific Workflows

Scientific Workflows …

- … can have many different forms
- … consist mainly of Activities, Gateways and Events



Quelle: http://docs.jboss.org/jbpm/v6.0.1/userguide/images/Chapter-1-Overview/Overview.png

MOTIVATION

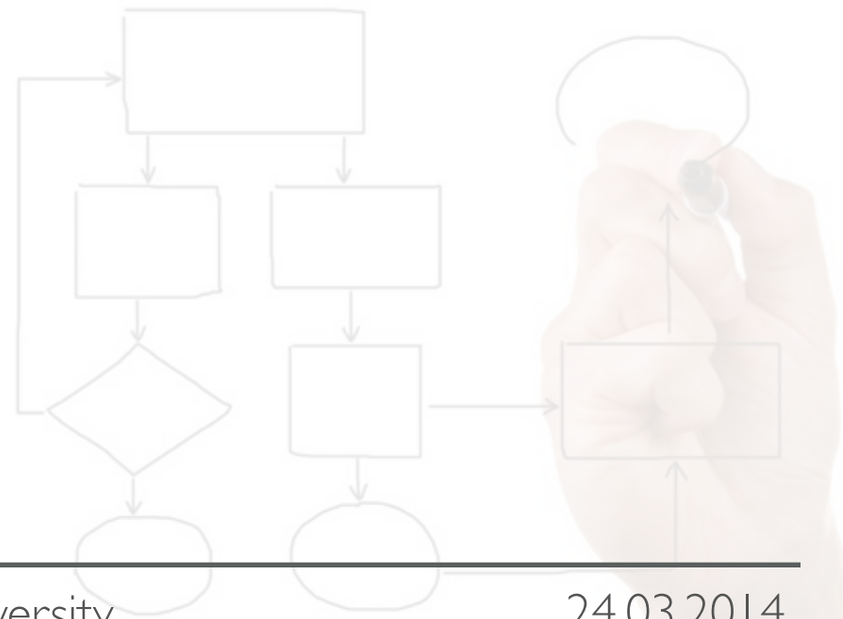
Problemstatement

- Scientific workflow engines are often domain specific and developed by small teams with less resources > only really needed functionality is implemented

- also business products are widely used

How to assure the capturing of provenance information in these systems ?

Research Question



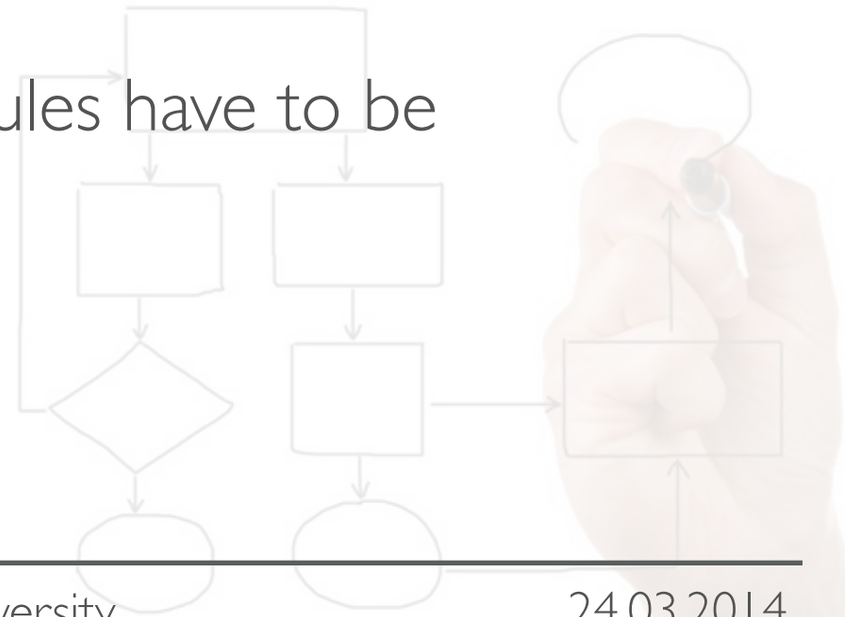
RESEARCH QUESTIONS

Aspect-Oriented Provenance Capturing

Data provenance as a cross-cutting concern

- Is it possible to use aspect-oriented programming techniques to acquire provenance information from a scientific software system?

- How to determine, which methods and modules have to be monitored?

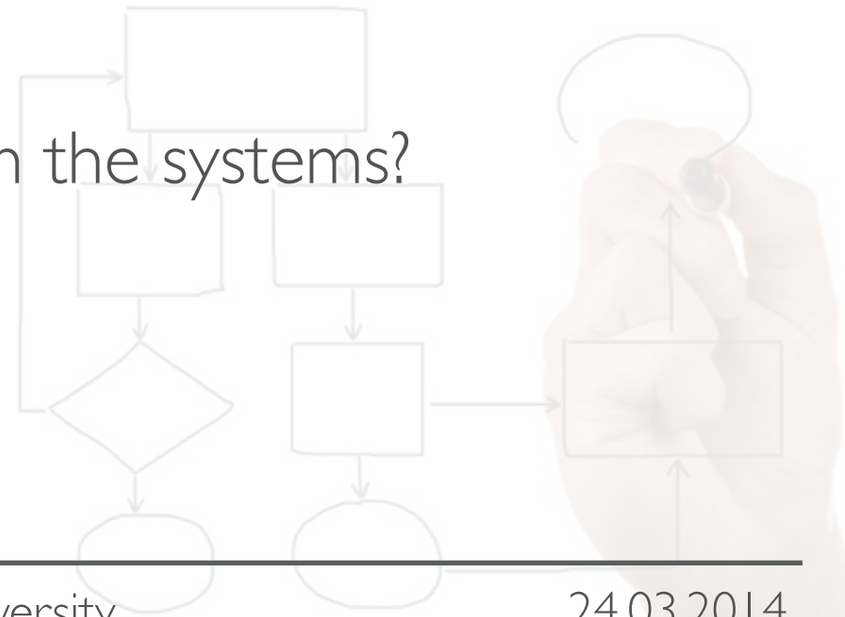


RESEARCH QUESTIONS

Scientific Workflows

Capturing Provenance Information in scientific workflows

- What are the demands for capturing provenance information in scientific workflows?
- Which informations have to be gathered from the systems?

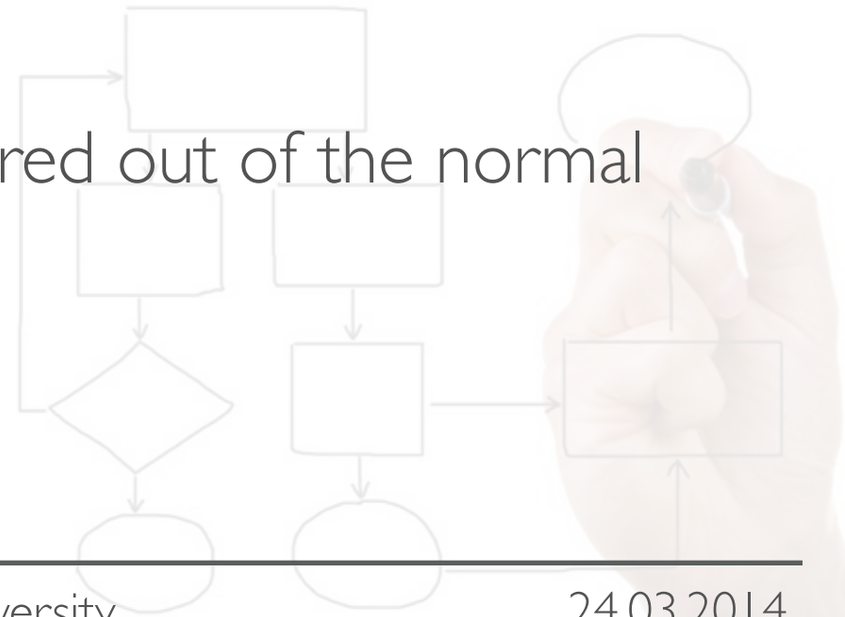


RESEARCH QUESTIONS

Provenance Information Reconstruction

Integrating provenance information from different systems and reconstruction the provenance graph

- What are the demands for constructing a provenance graph out of monitoring informations received from different software systems?
- How can the provenance information be filtered out of the normal monitoring records?



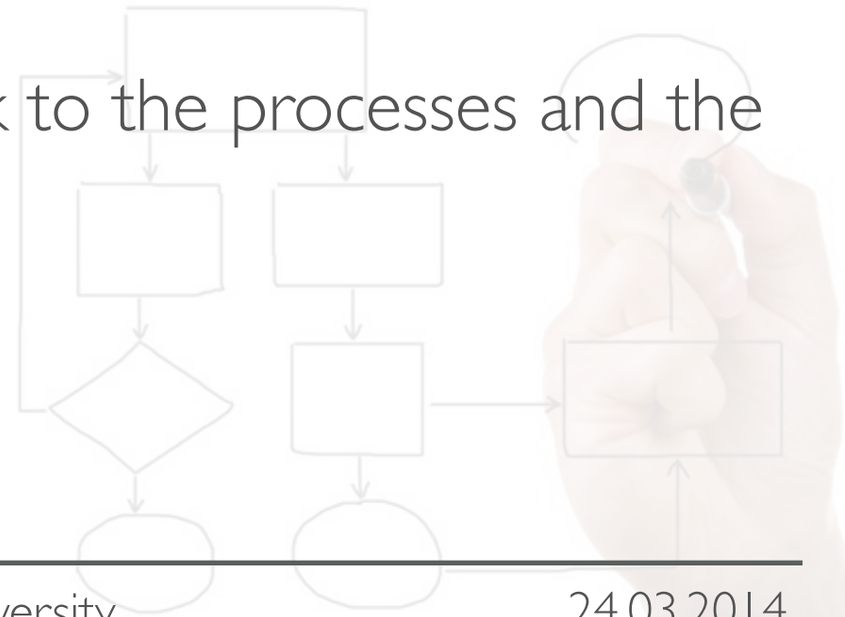
RESEARCH QUESTIONS

Provenance Information Archival

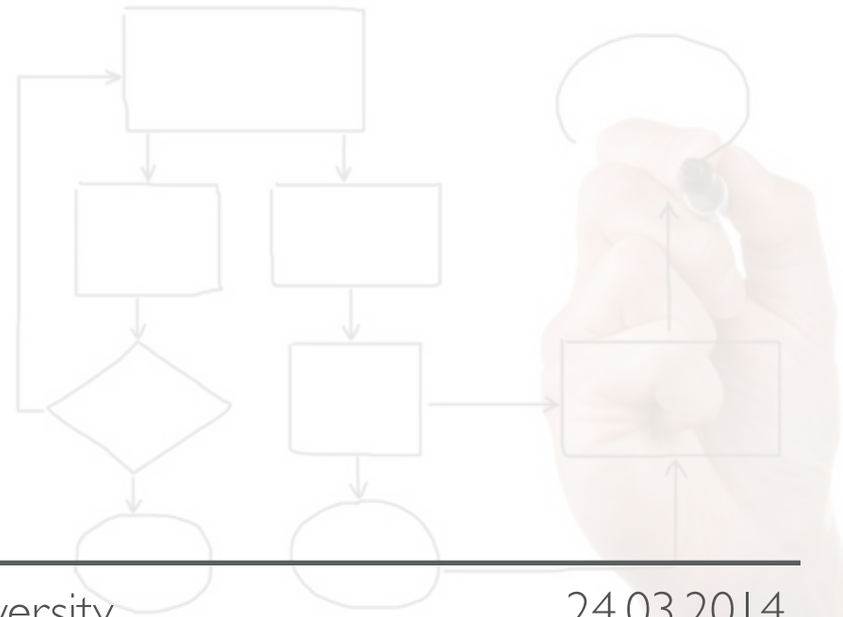
Archiving the provenance information

- How to store the provenance information, so it can be accessed by scientist and data managers to validate their data?

- How to link the provenance information back to the processes and the data it evolved of?

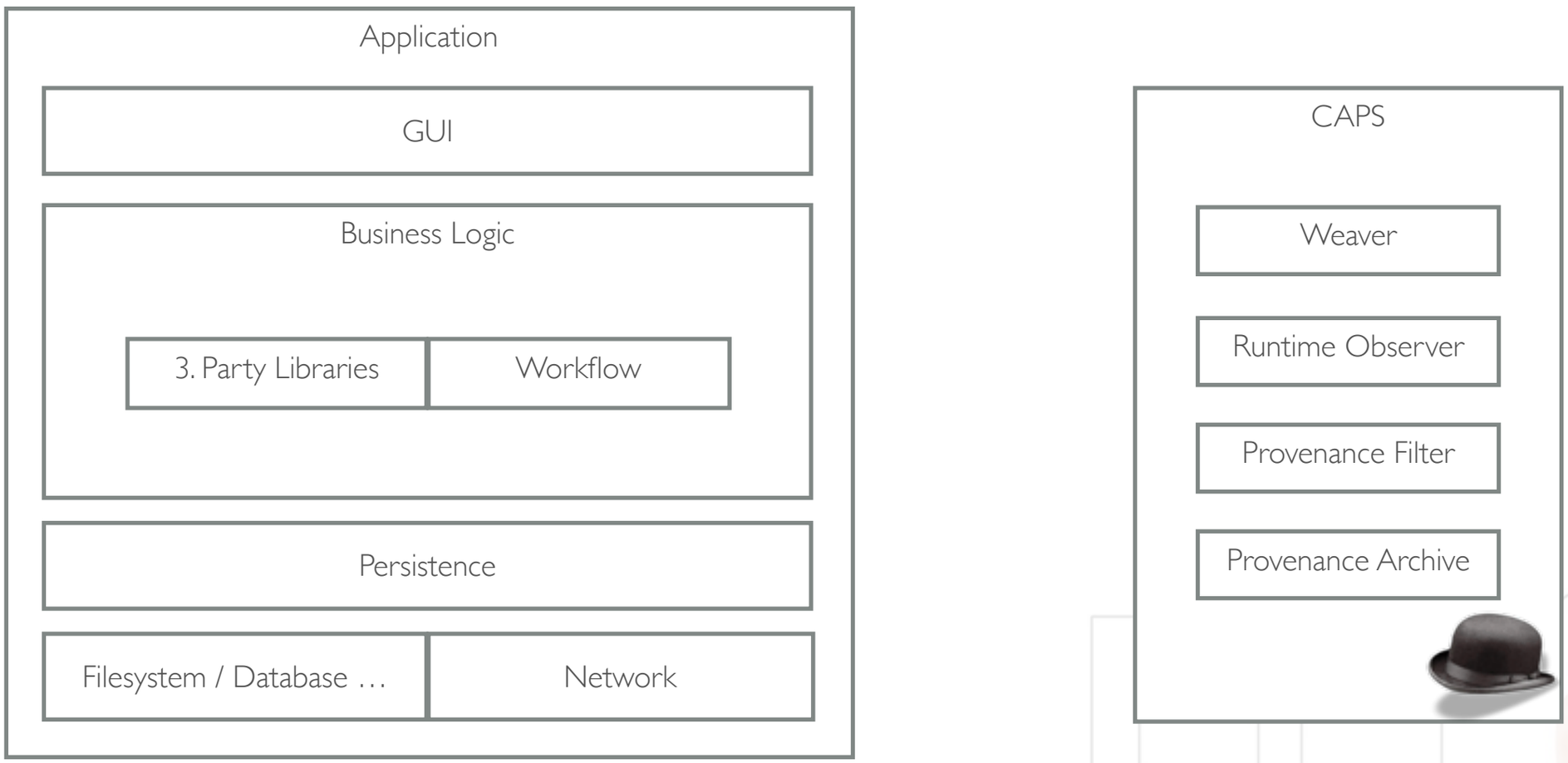


CAPS



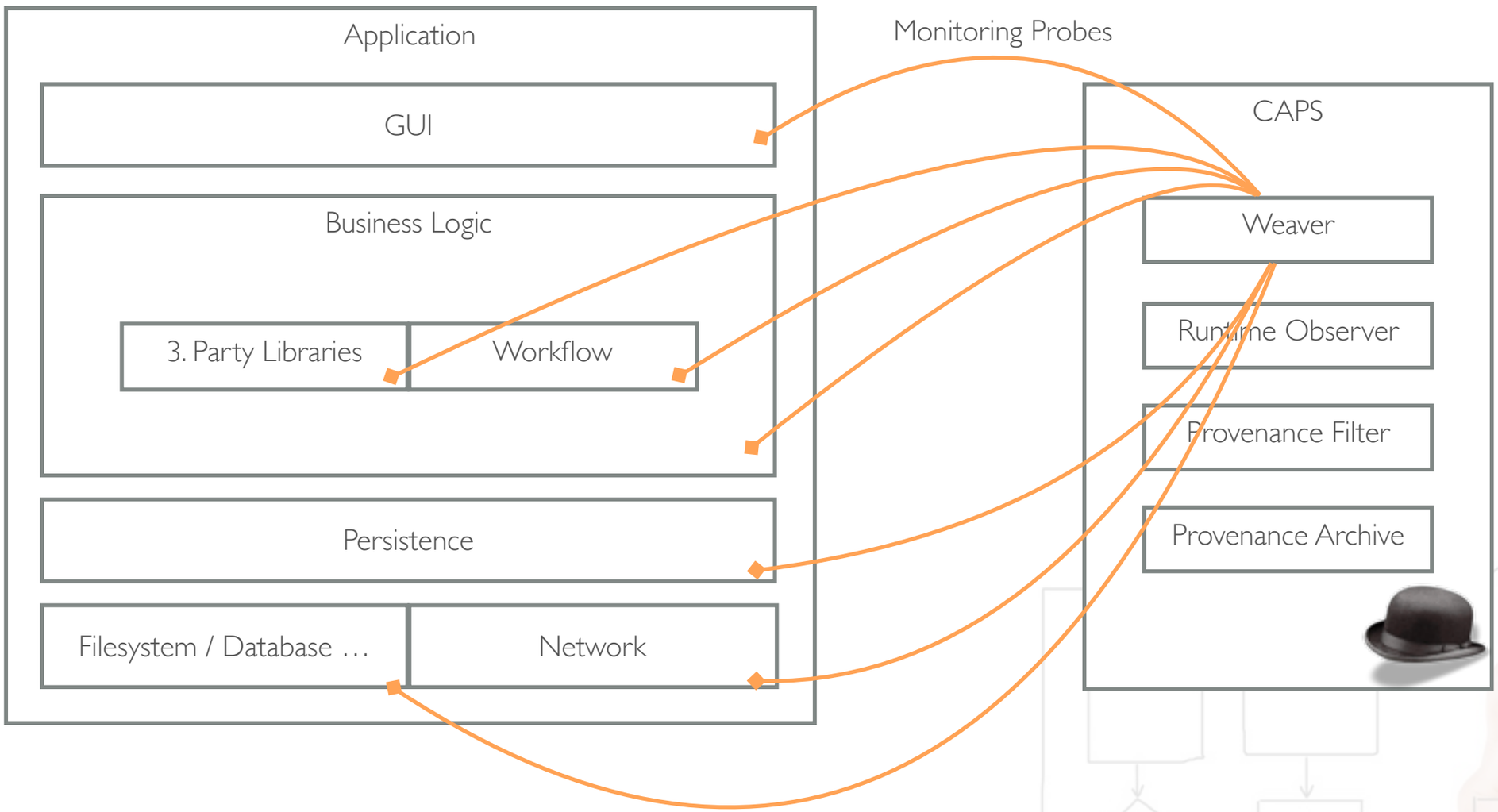
CAPS

Integration in the Guest System



CAPS

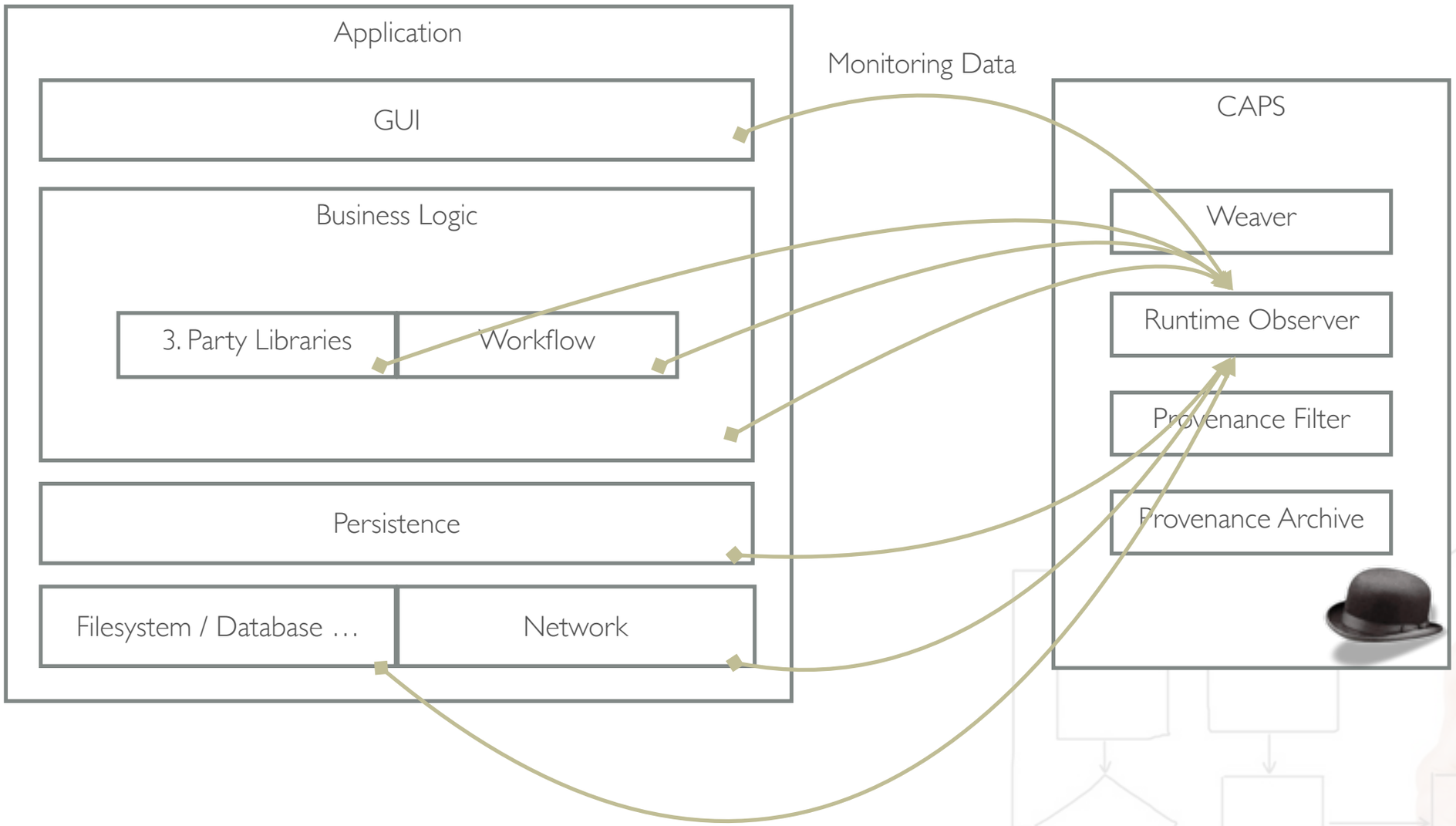
Integration in the Guest System



Analyze Application & Integrate Provenance Awareness

CAPS

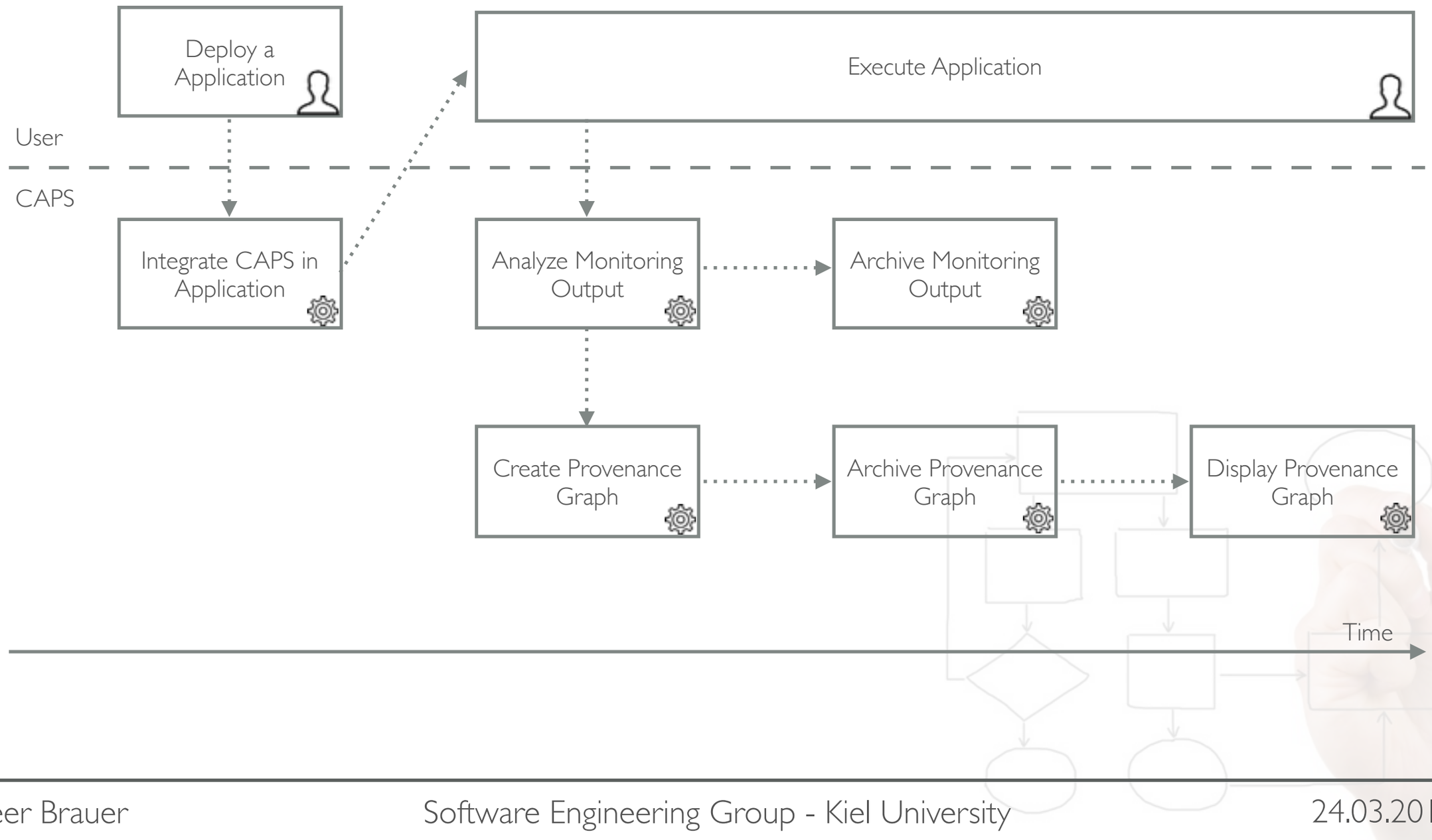
Integration in the Guest System



Collect Provenance Information

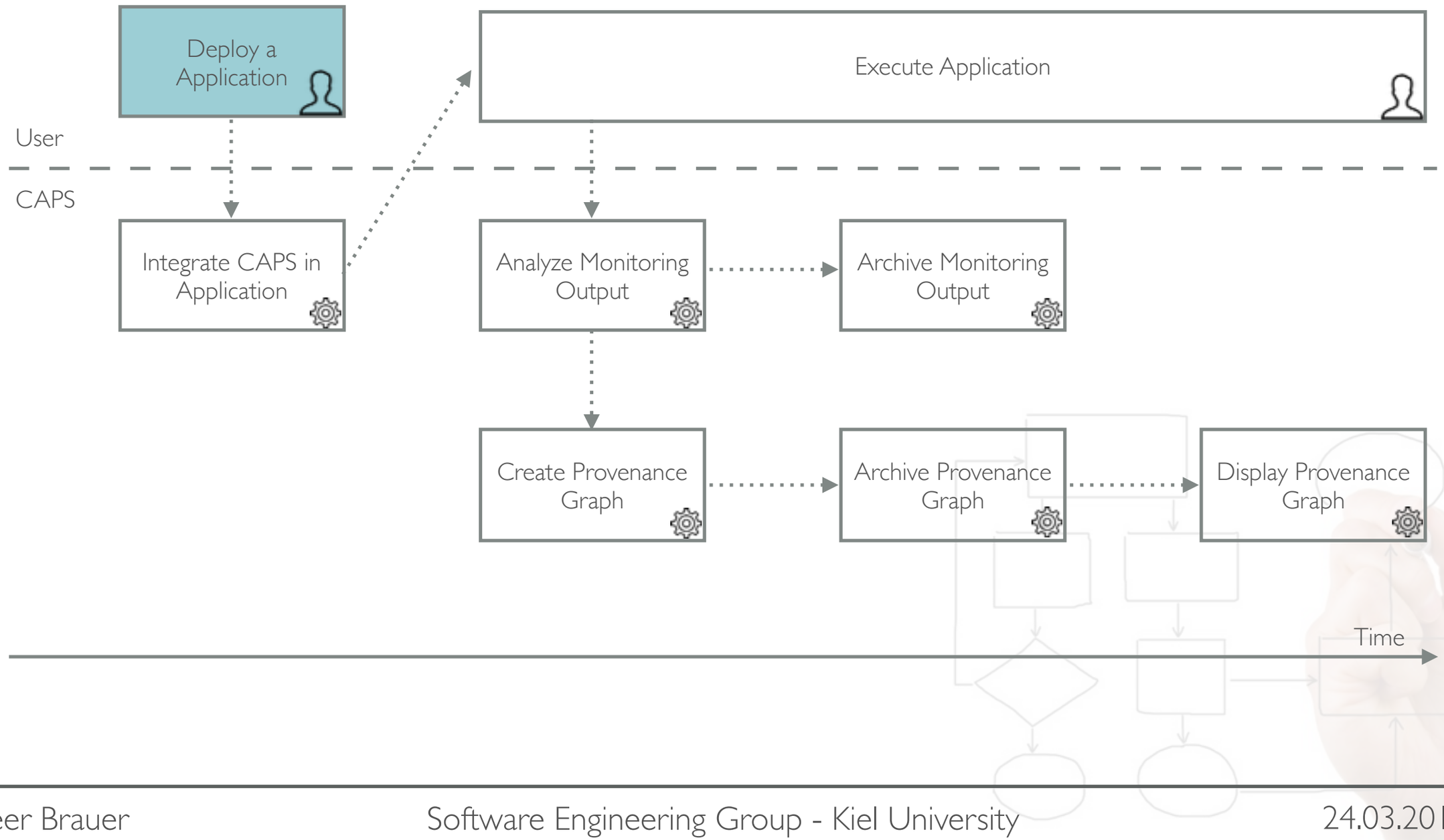
CAPS

Basic Workflow



CAPS

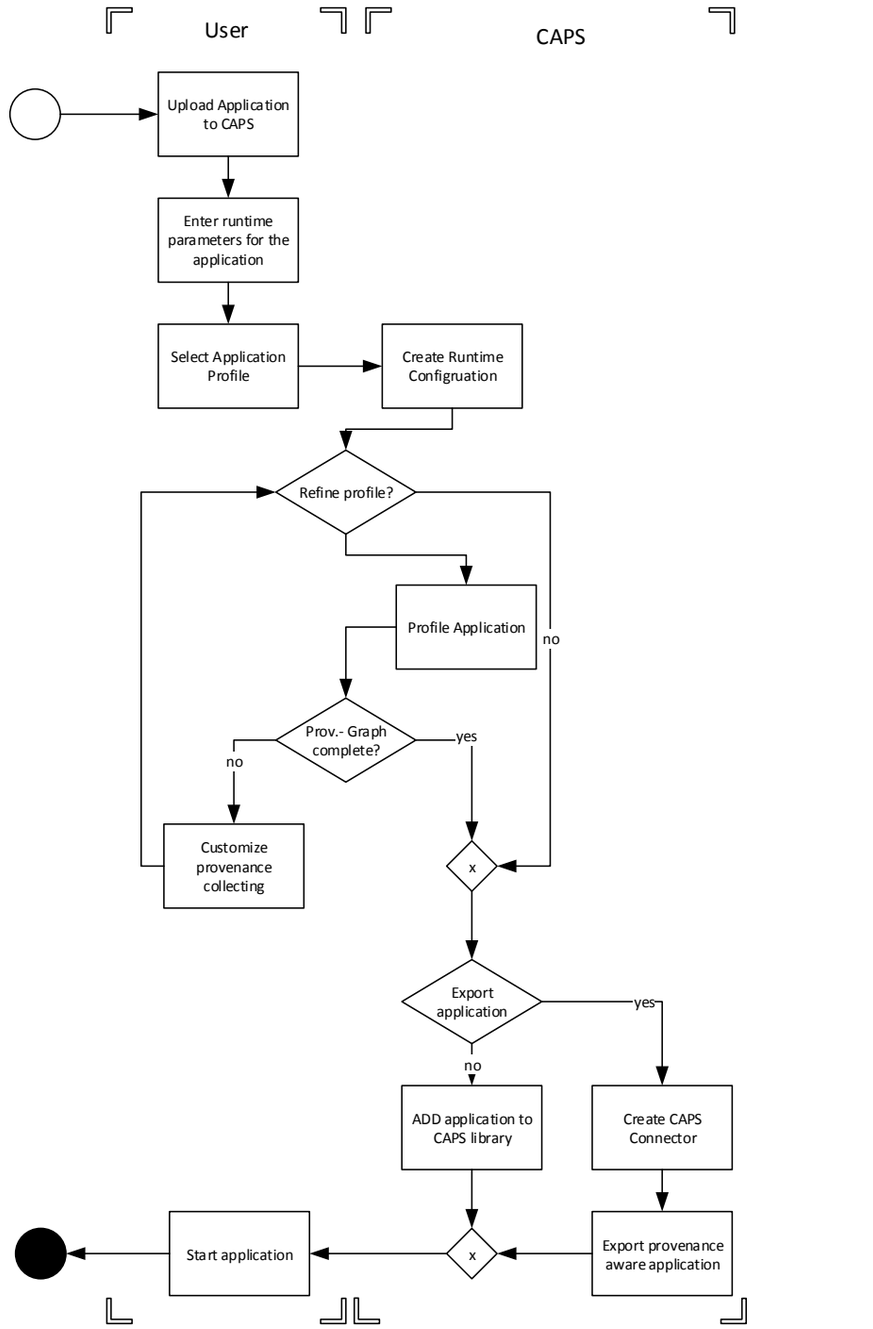
Basic Workflow



CAPS

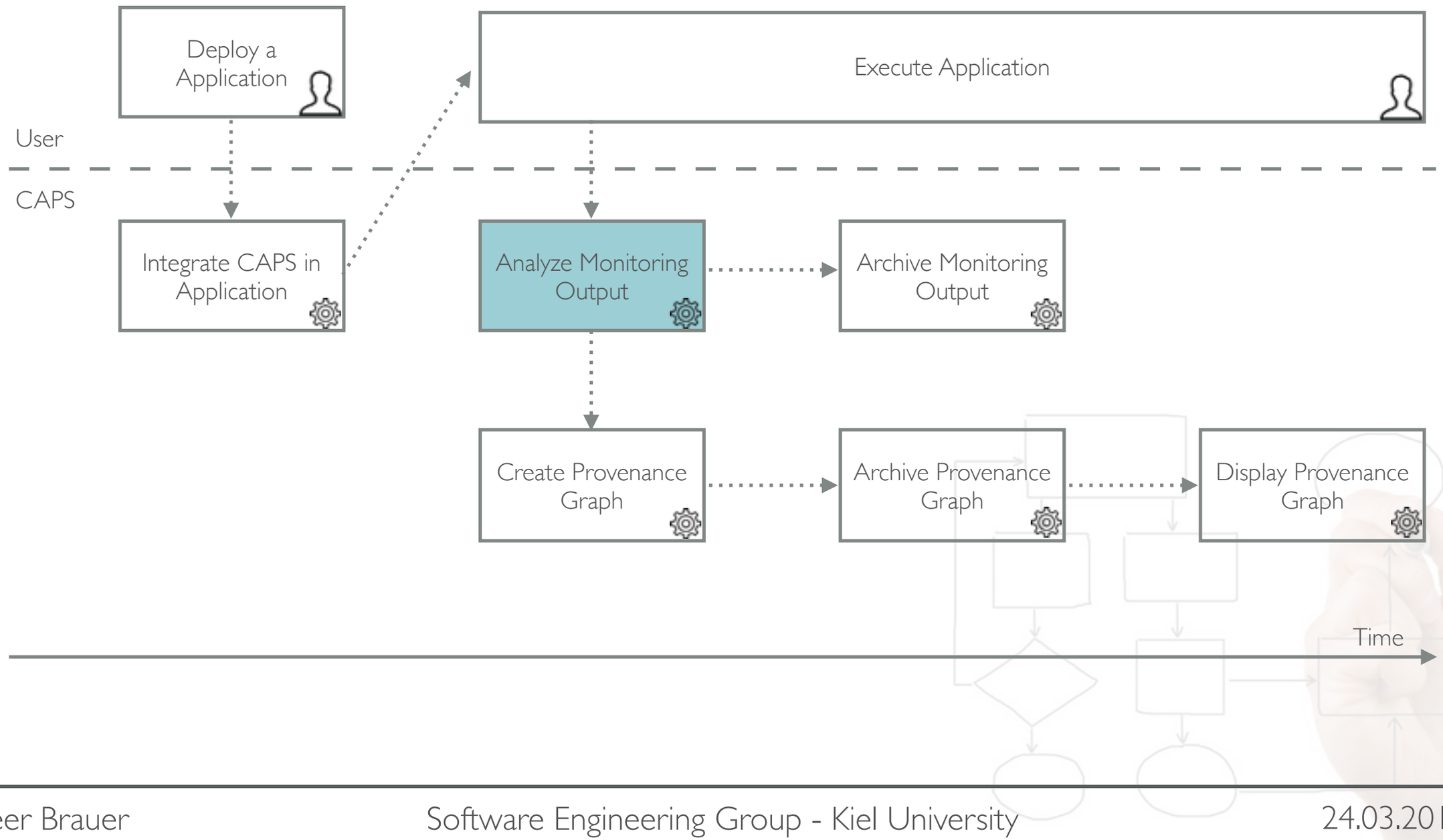
Deploying an Application

- Upload an application to CAPS
- Choose an application profile
- Create a runtime configuration
- Create an application wrapper



CAPS

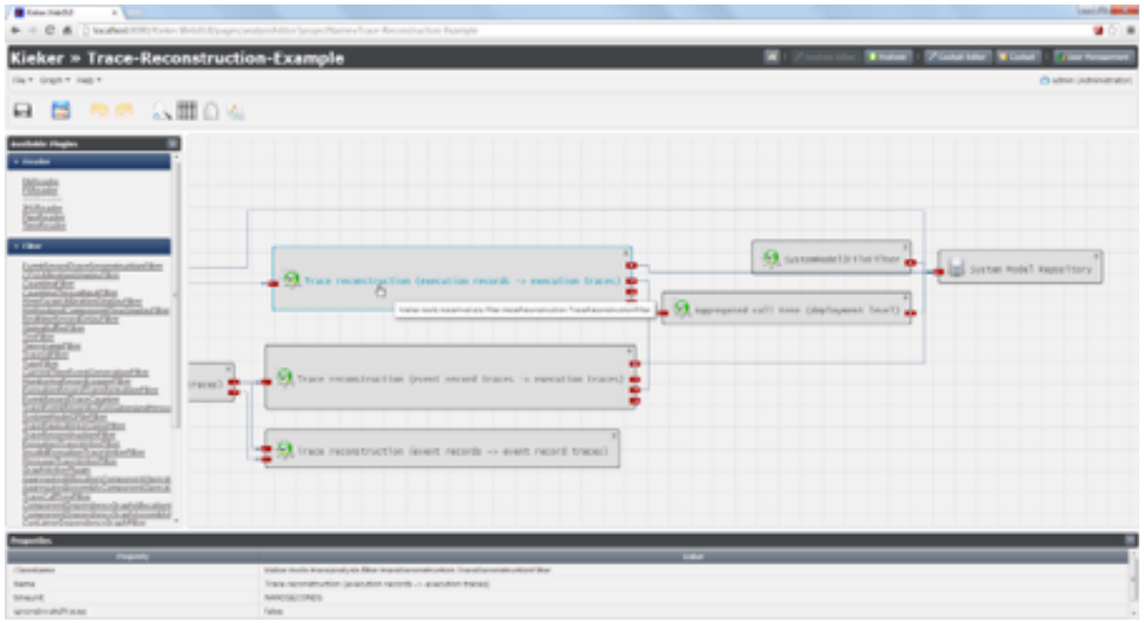
Basic Workflow



CAPS

Analyzing and Filtering the Results

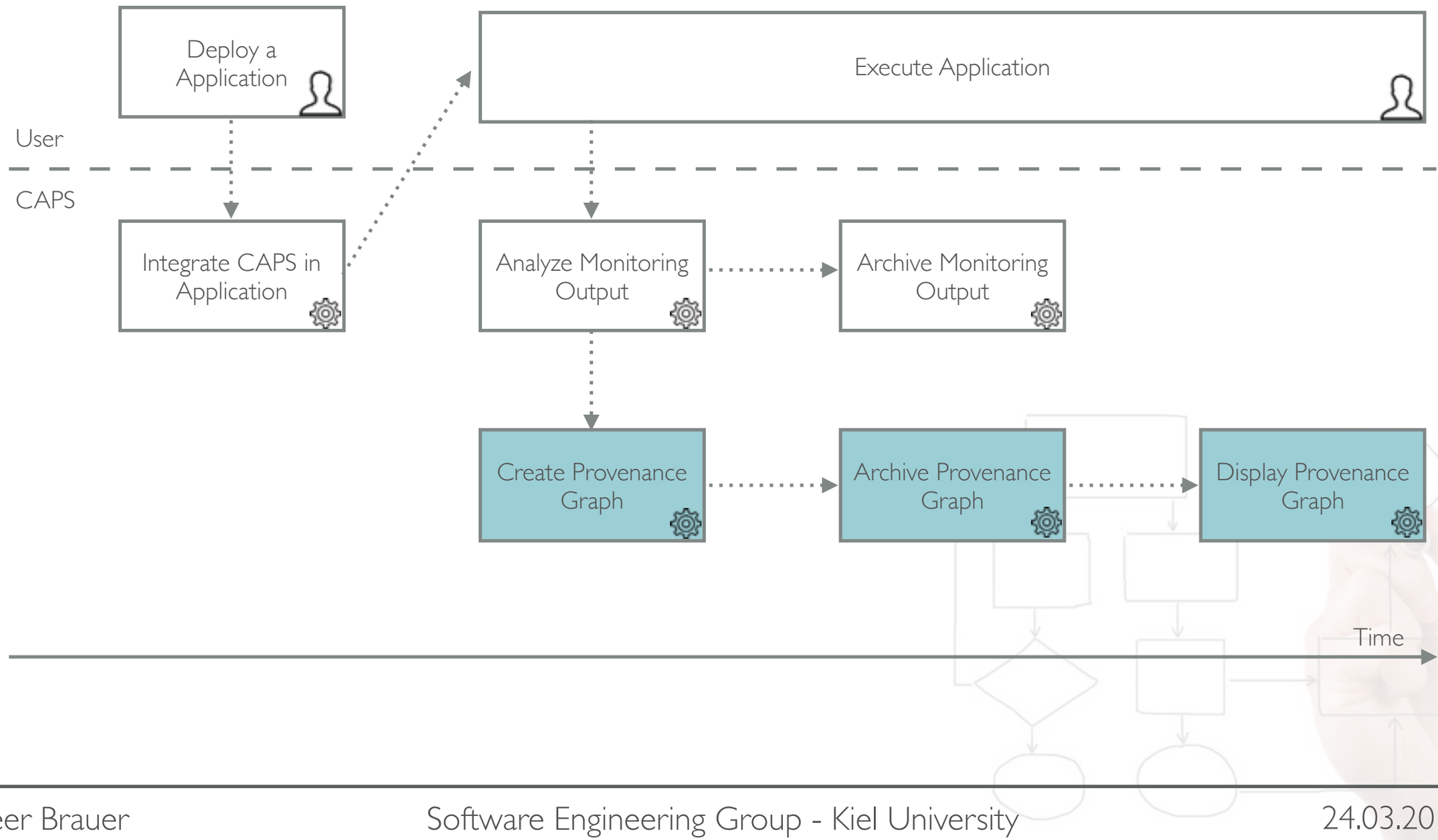
- Observation data analysis is build upon Kieker Webgui
- Data is processed by an integrated pipe and filter analysis framework
- Dedicated provenance filter reconstruct the provenance graph, workflow
- Additional information can be processed (Memory usage, CPU time, …)
- Own filters and analysis plugins can be added



Source: http://kieker-monitoring.net/features/webgui/attachment/kieker-webgui-02/

CAPS

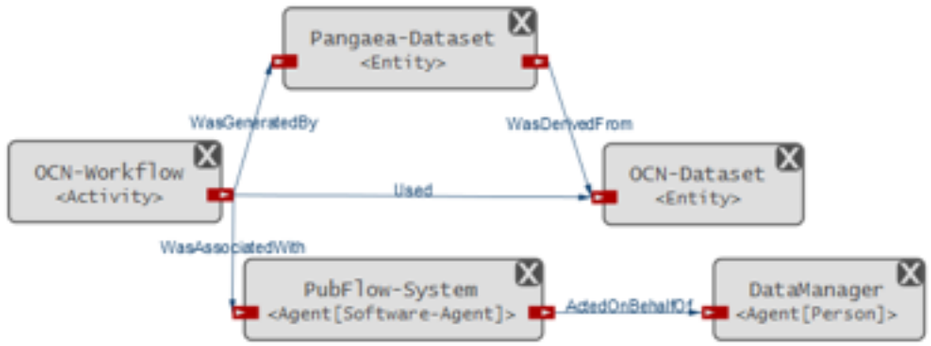
Basic Workflow



CAPS

Archiving the Provenance Data

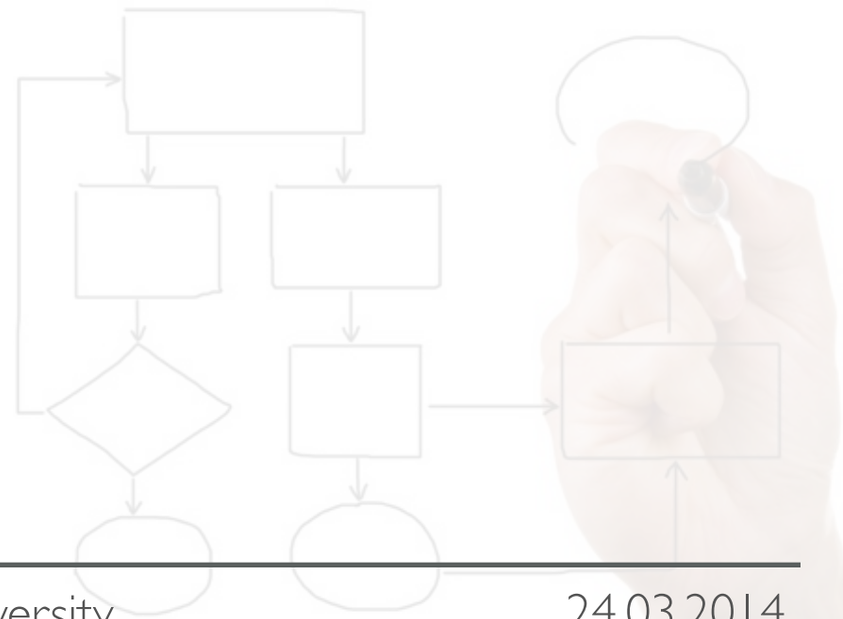
- Provenance archive build upon Neo4J, EMF and Prov-DM



- API, Webservice and WebGUI for accessing the provenance information

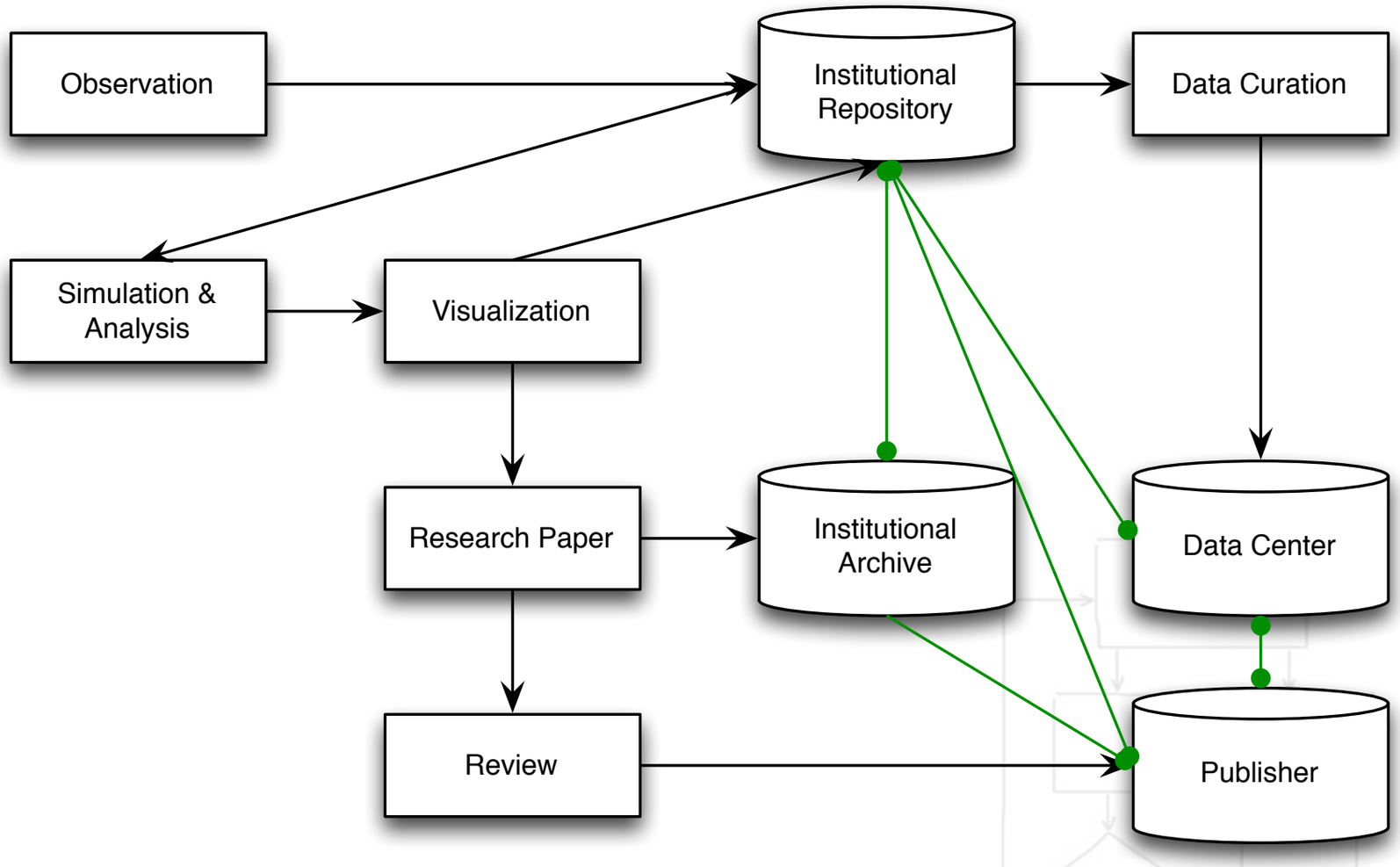
- Export of the provenance data to data centers (i.e. Zenodo)

Evaluation



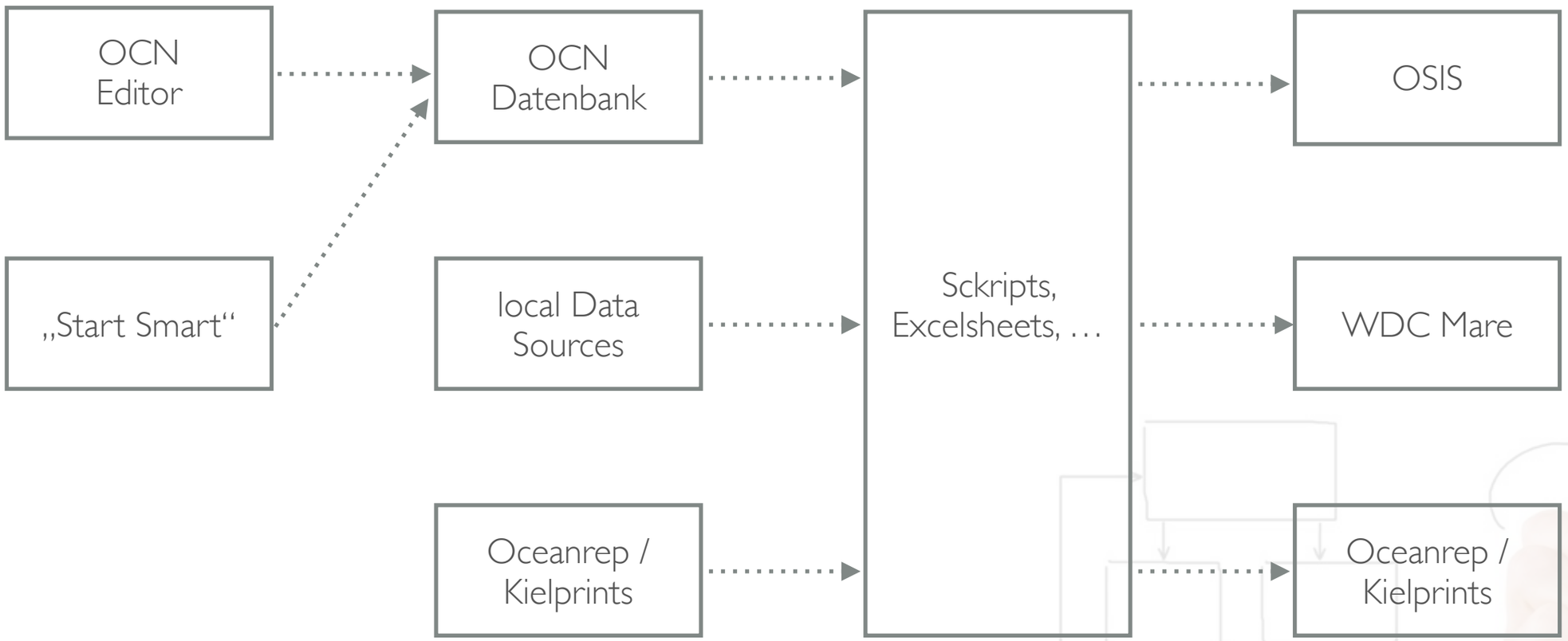
EVALUATION

PubFlow



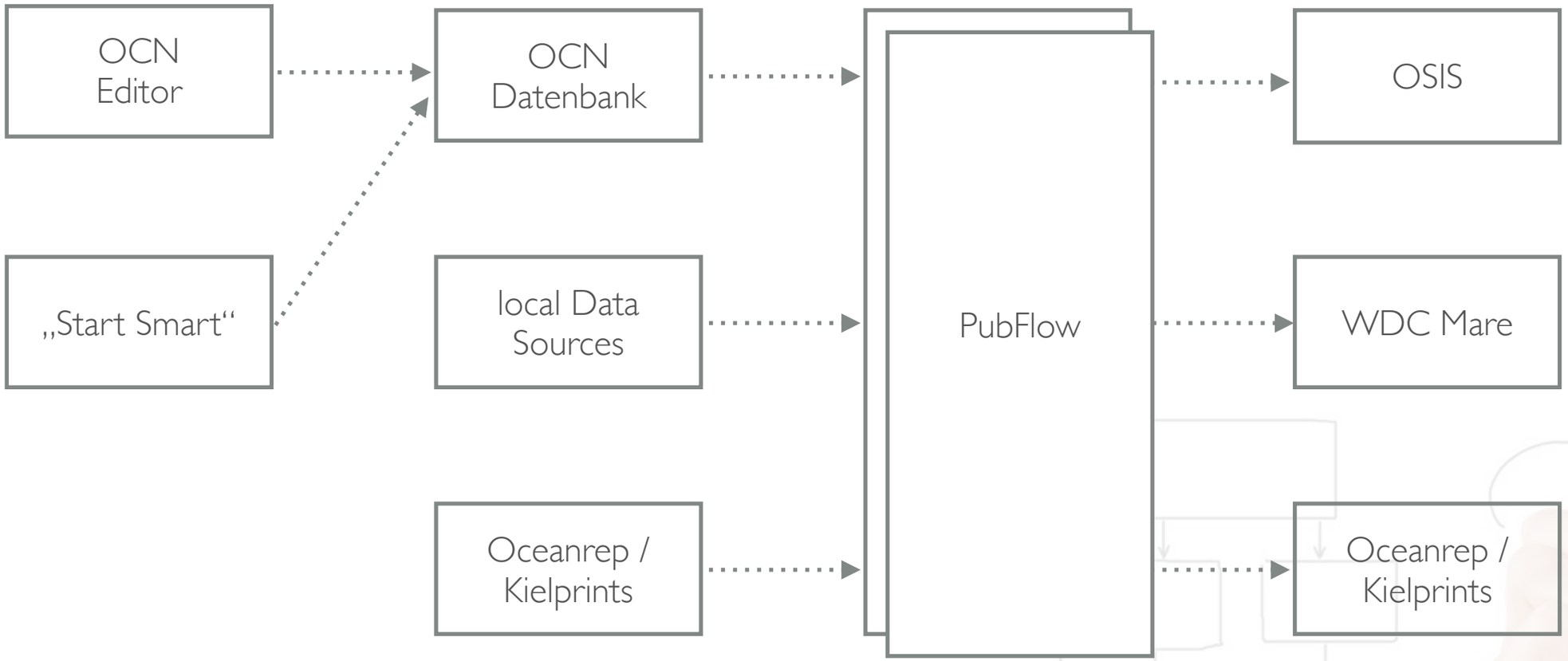
EVALUATION

System Landscape



EVALUATION

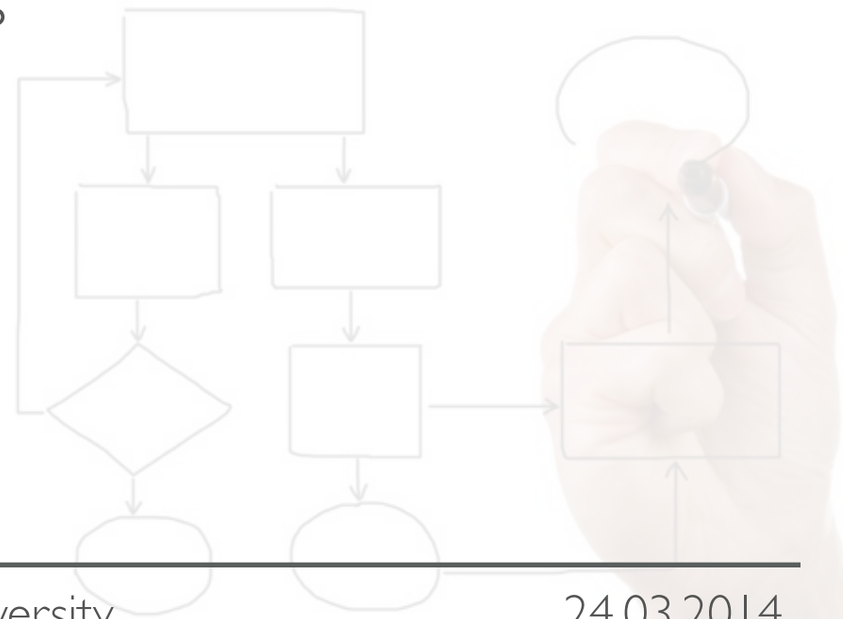
System Landscape



EVALUATION

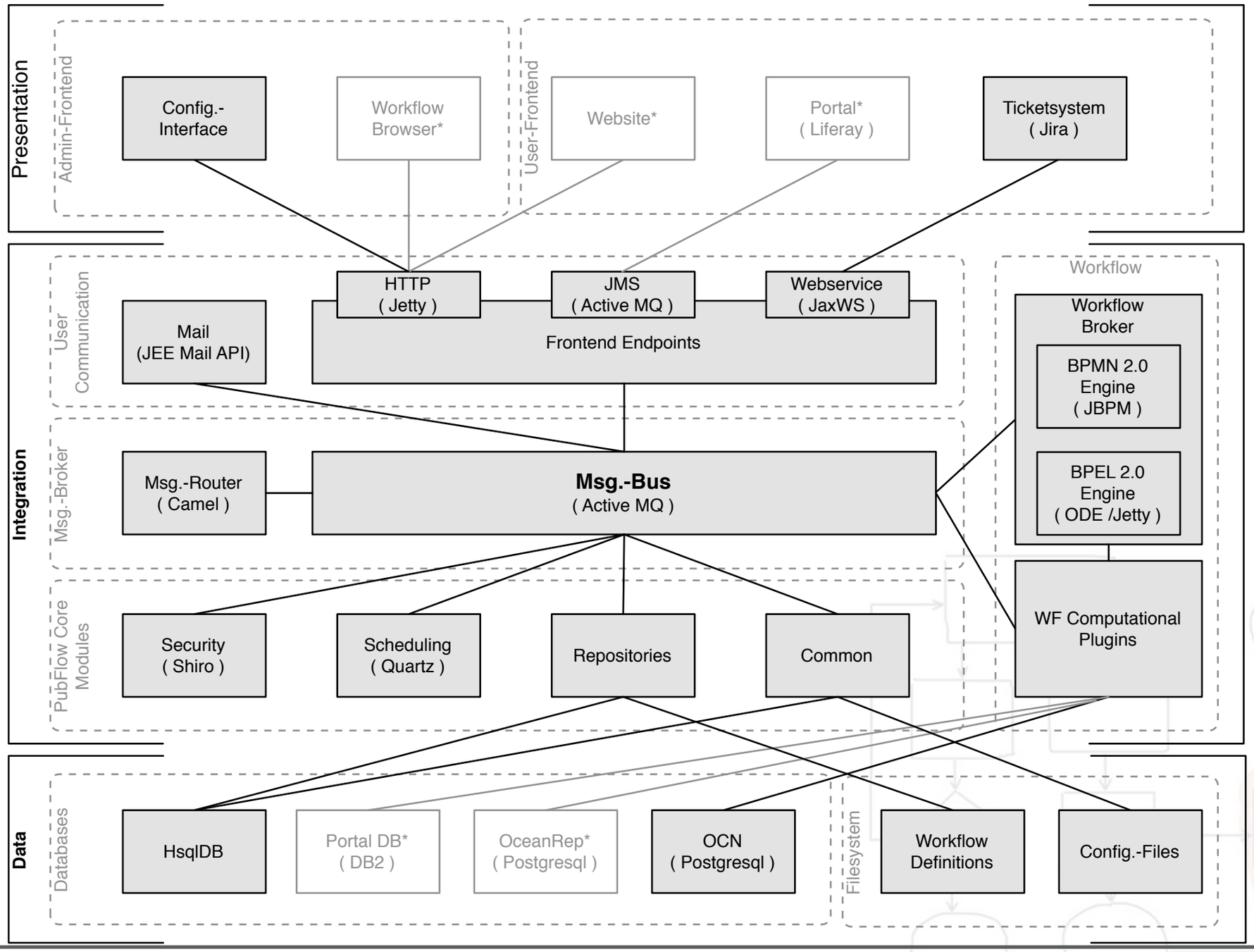
PubFlow : Goals

- Development of a workflow based tool for data publication
- Integration of PubFlow in the existing systems



EVALUATION

PubFlow



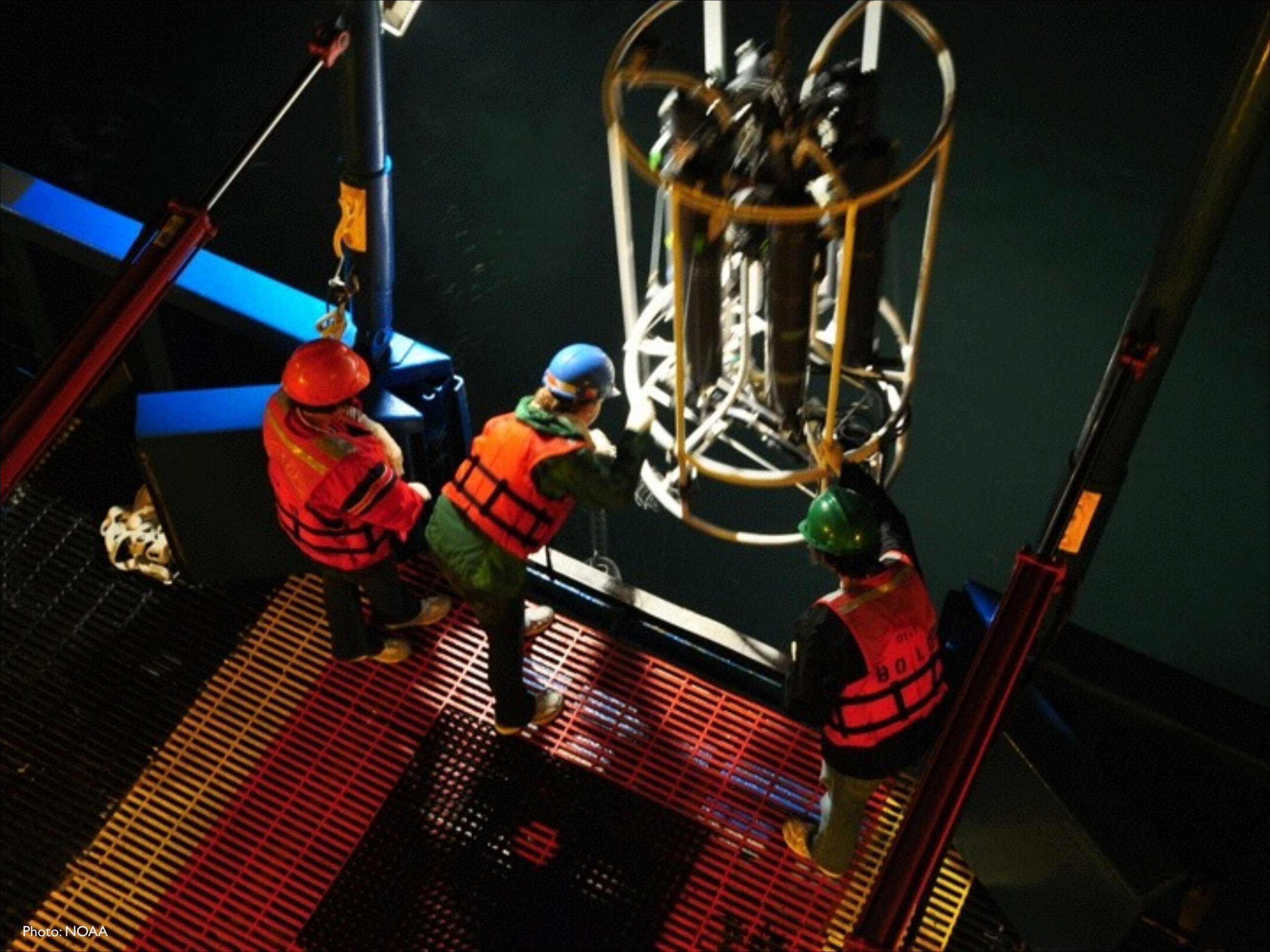
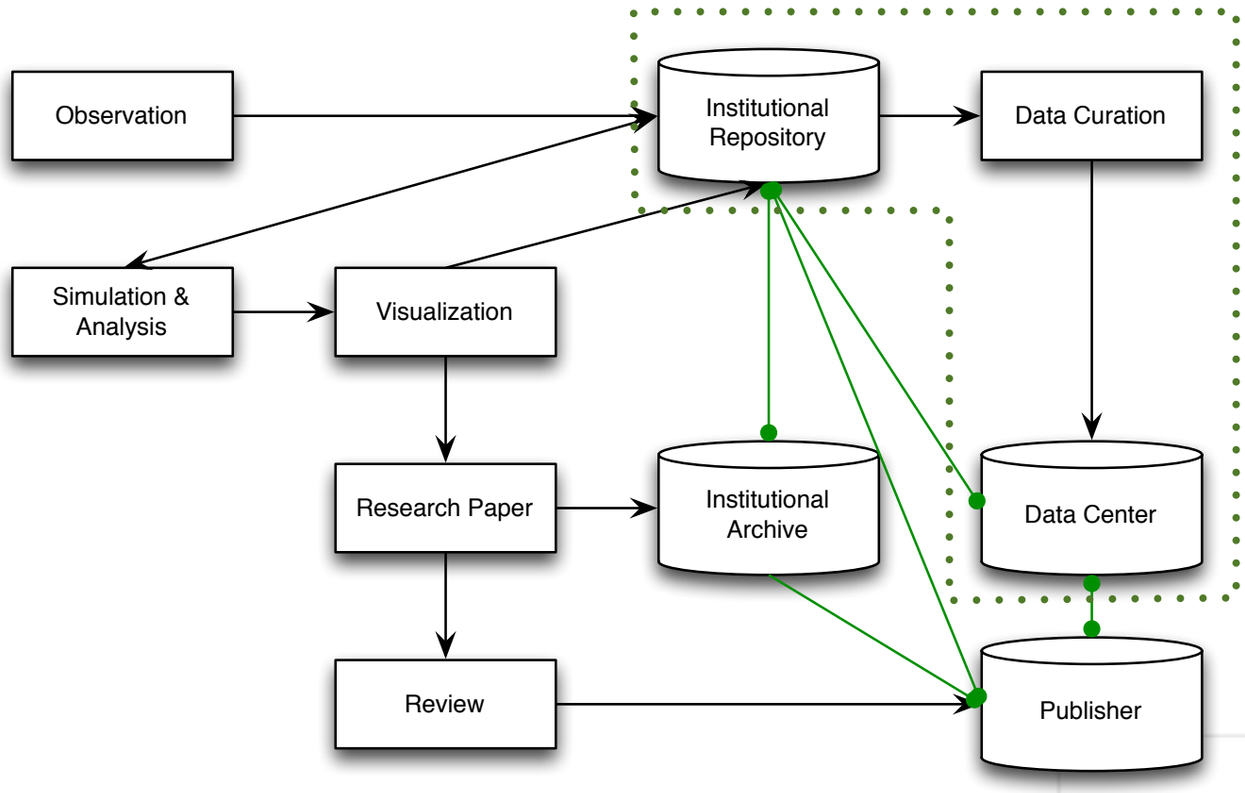


Photo: NOAA

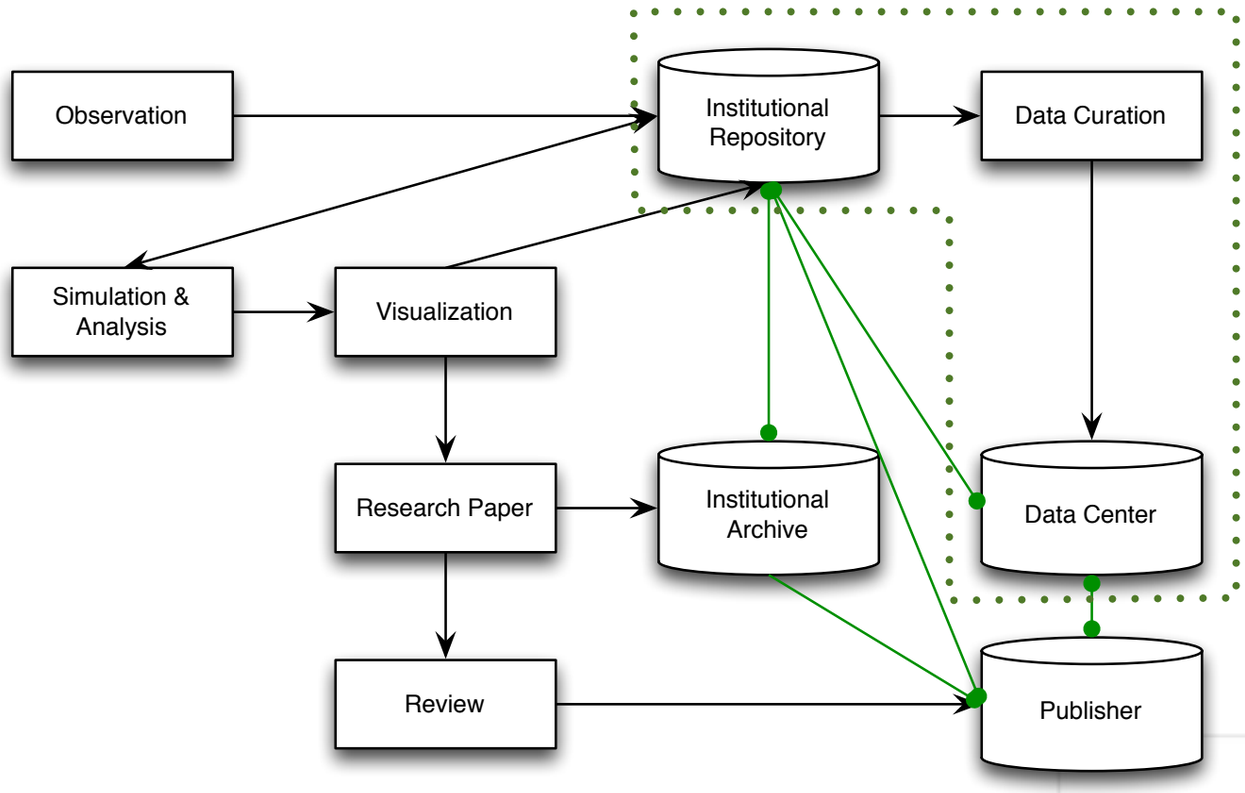
EVALUATION

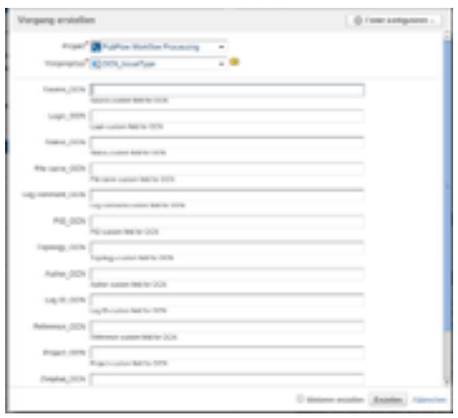
PubFlow : The Workflow



EVALUATION

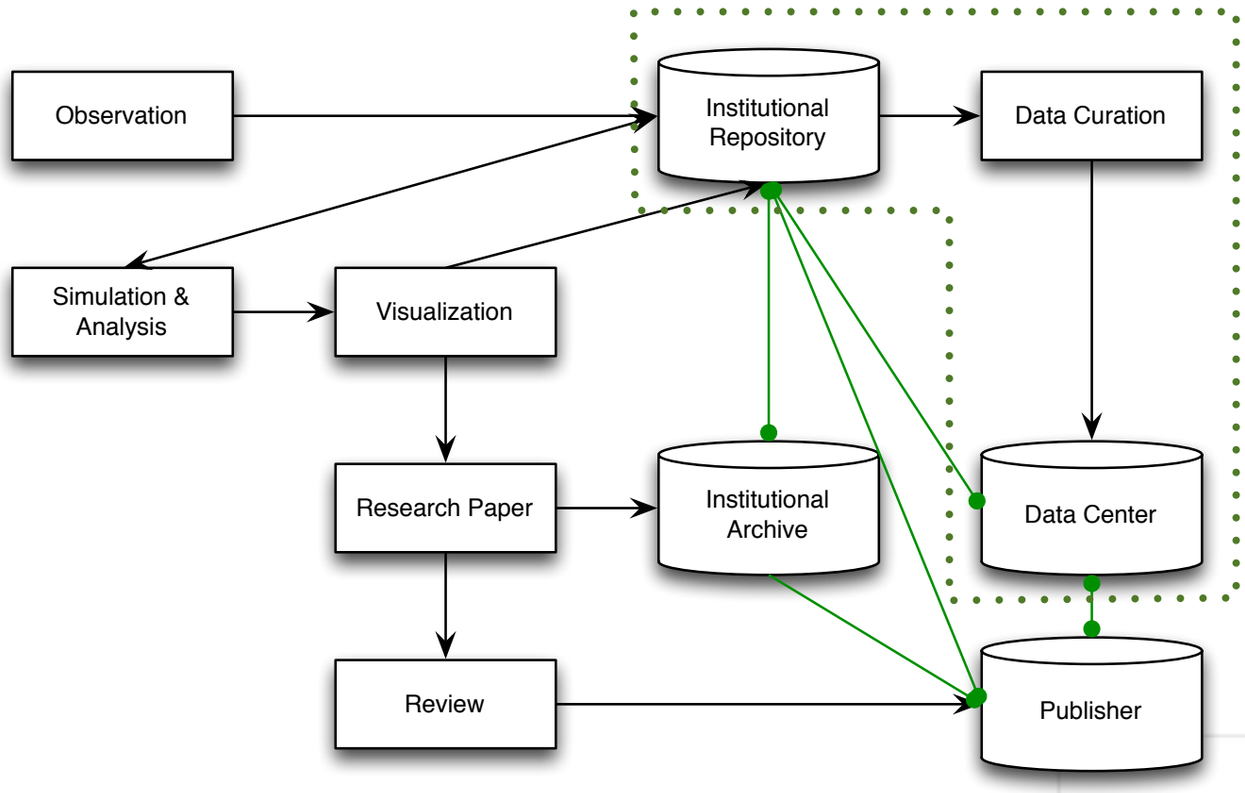
PubFlow : The Workflow

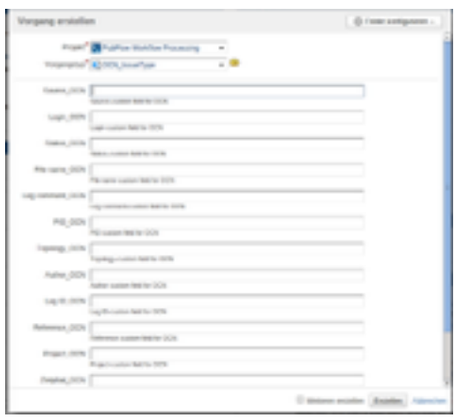




EVALUATION

PubFlow : The Workflow

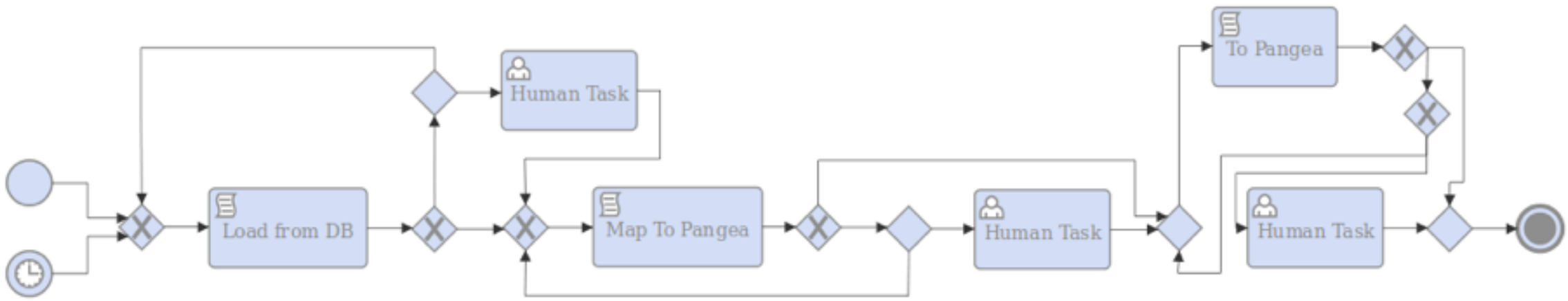




EVALUATION

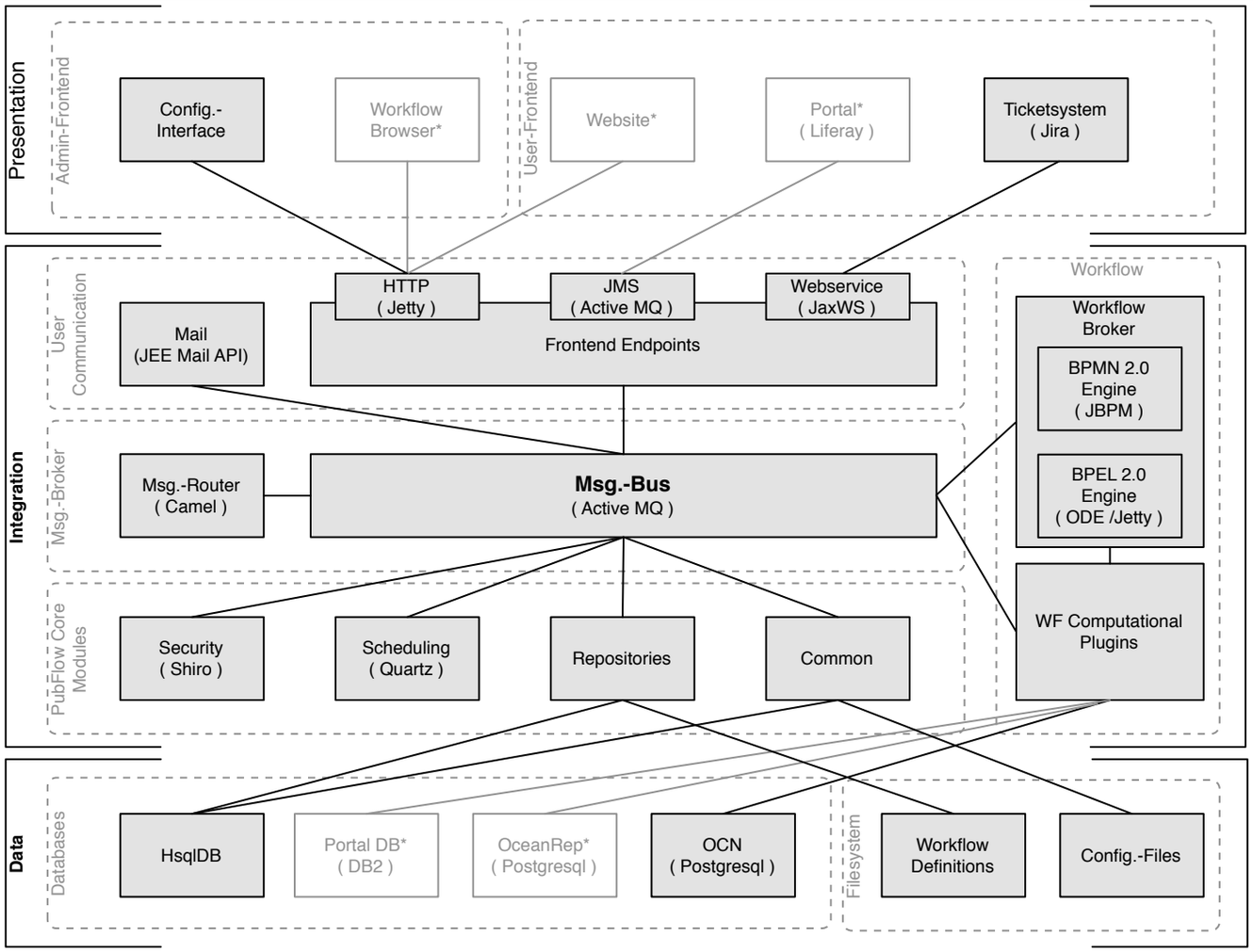
PubFlow : The Workflow

- Loading the Data
- Transformation
- Export the Data to the WDC



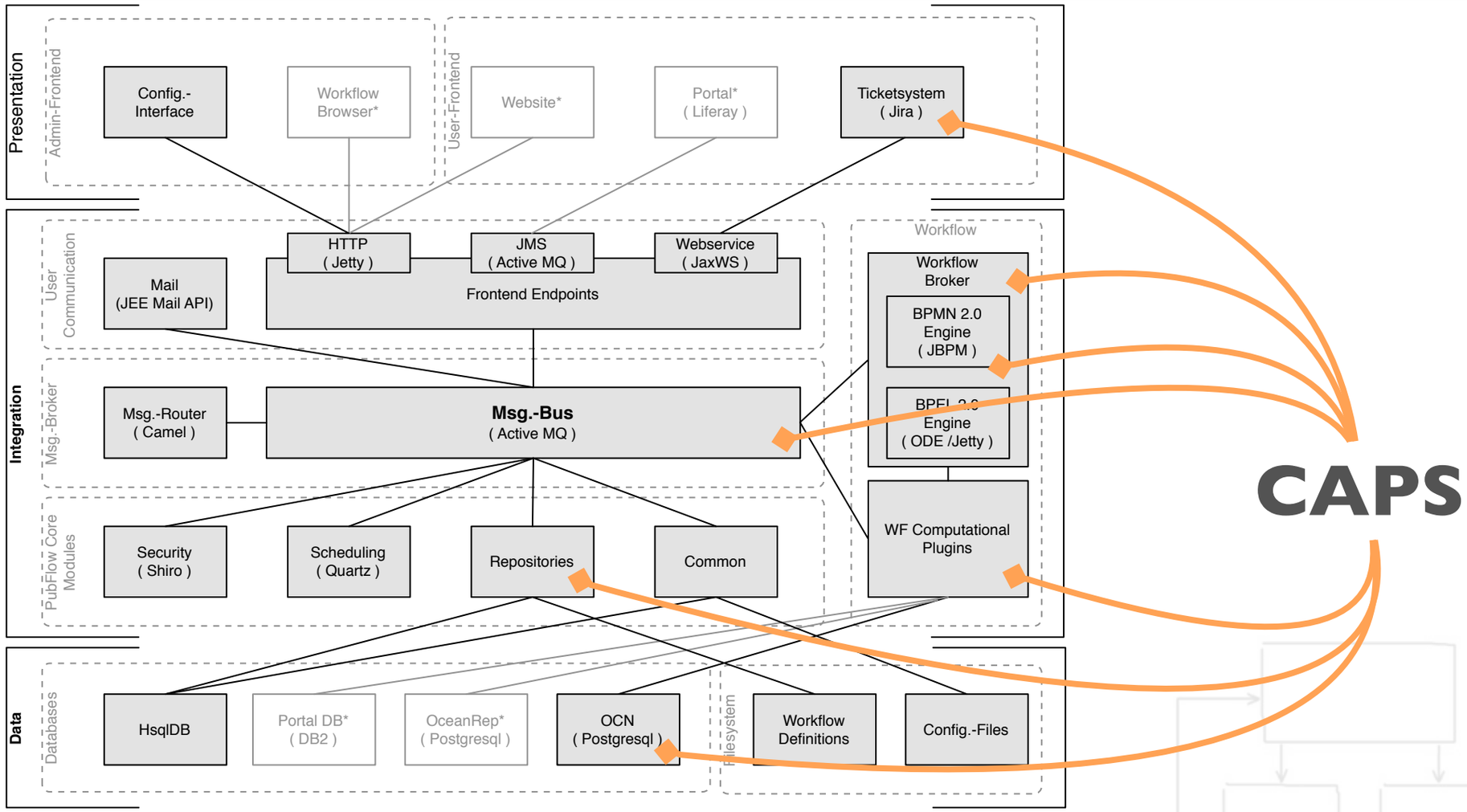
EVALUATION

PubFlow



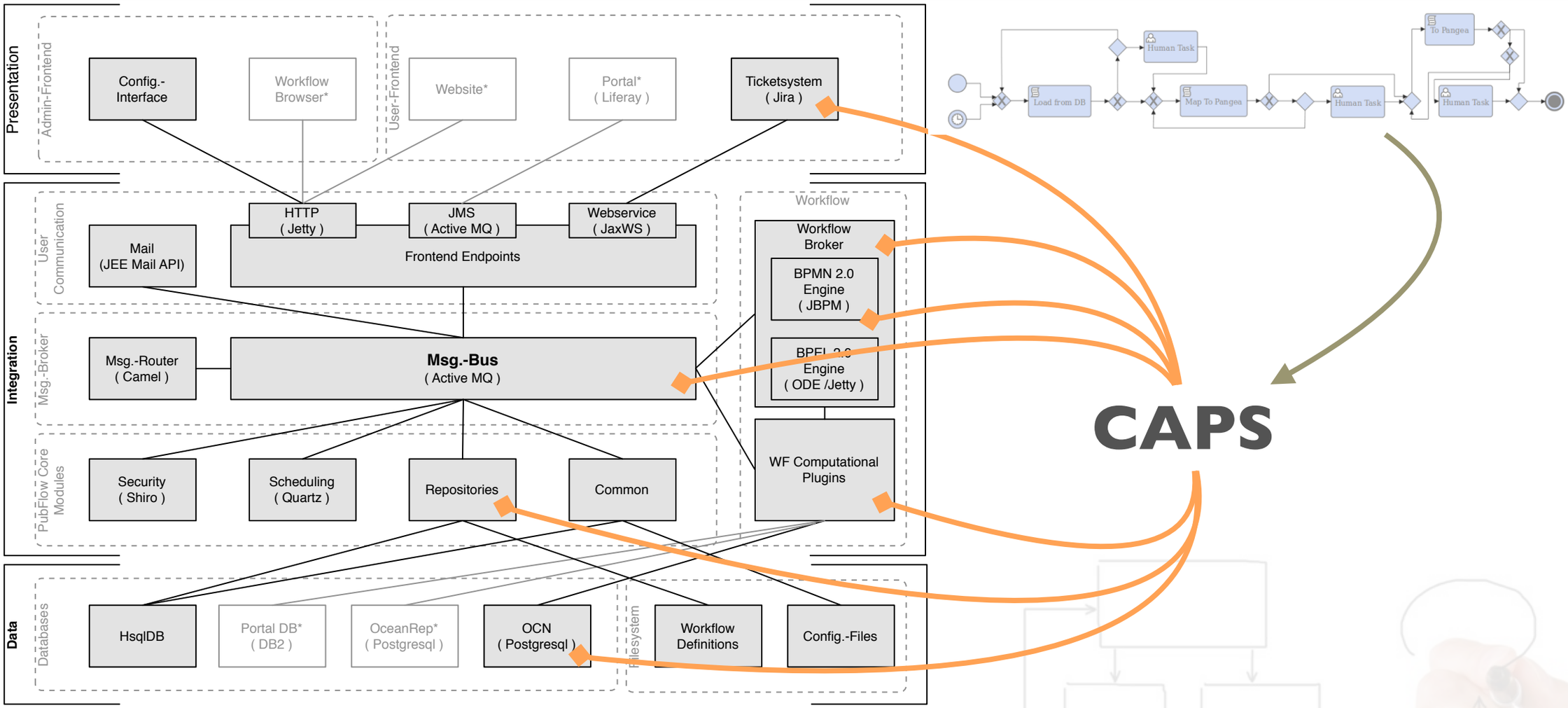
EVALUATION

PubFlow



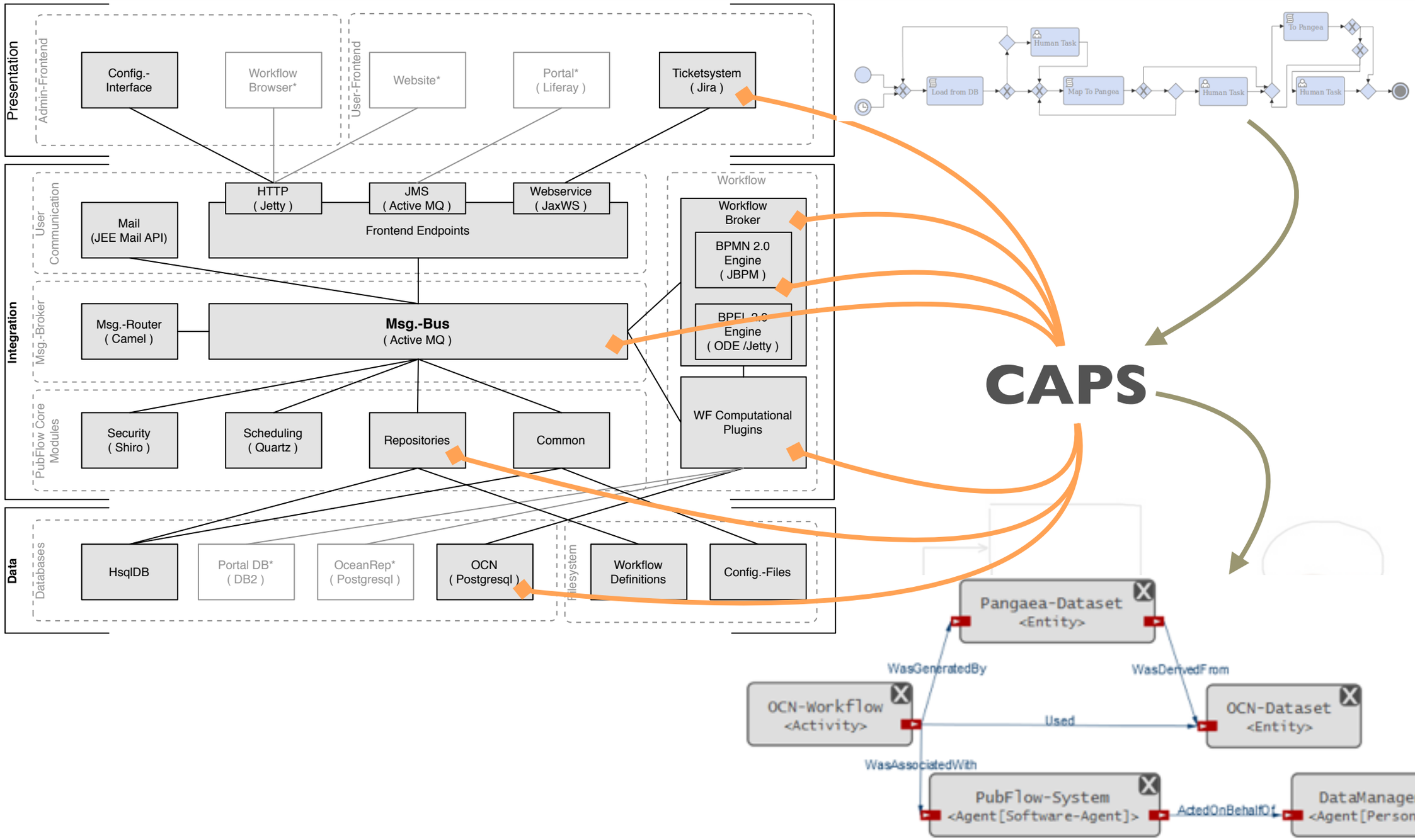
EVALUATION

PubFlow



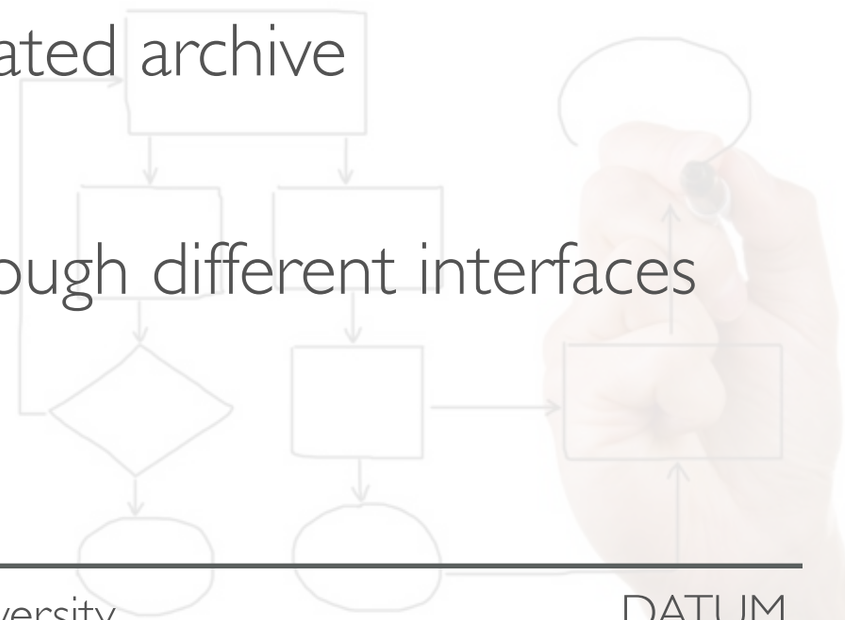
EVALUATION

PubFlow



CONCLUSION

- CAPS captures provenance information using AspectJ and workflow monitoring
- Uses proven technologie like Kieker, Neo4J, Emf, GWT, …
- Archives provenance information in an integrated archive
- Makes provenance information accessible through different interfaces



CONCLUSION

Thanks for your attention !