Searching for Cancer Biomarkers with Apache OODT

Rishi Verma
NASA Jet Propulsion Laboratory,
California Institute of Technology
February 27th, 2013

APACHECON
NORTH AMERICA 2013
Open Source Community Leadership Drives
Enterprise-Grade Innovation

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Prezi

Searching for Cancer Biomarkers with Apache OODT

Rishi Verma
NASA Jet Propulsion Laboratory,
California Institute of Technology
February 27th, 2013

APACHECON
NORTH AMERICA 2013
Open Source Community Leadership Drives
Enterprise-Grade Innovation

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Prezi

Agenda

* Introduction
* Science of cancer biomarker research
* Apache OODT and data processing
* Example: EDRN LabCAS Proteomics Pipeline
* Tutorial: Build your own processing system

Prezi

Introduction

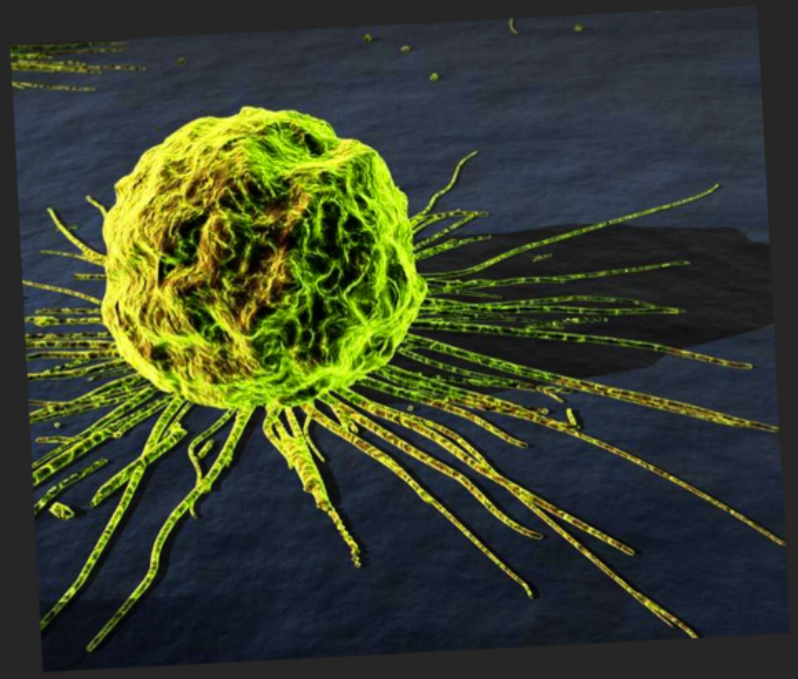
Rishi Verma

NASA Jet Propulsion Laboratory

- Software engineer specializing in data services for scientific discovery
- Involved in:
  - Cancer research
  - Mars Science Laboratory
  - Deep Space Network
  - CO2 data management
  - Earthquake data management
- Program management committee member and committer for Apache OODT

Prezi

Science of cancer biomarker research



Prezi

Cancer biomarkers

- biological indicators
- used in many fields: medicine, biology, geology... etc.

"A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease."

- National Cancer Institute

Prezi

biomarkers

- biological indicators
- used in many fields: medicine, biology, geology... etc.

Prezi

Cancer biomarkers

- biological indicators
- used in many fields: medicine, biology, geology... etc.

"A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease."

- National Cancer Institute

Prezi

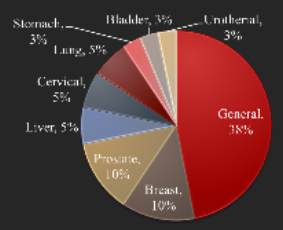
Cancer biomarkers: facts

Applications:

- Screening
- Early detection
- Treatment

Types

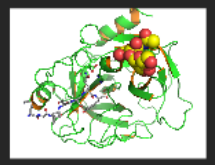
- Proteins
- Gene expression and DNA



Examples

Prostate specific antigen (PSA)

- Cancer type: prostate
- Substence type: glycoprotein
- Use: like other markers, currently screening is not recommended



Prezi

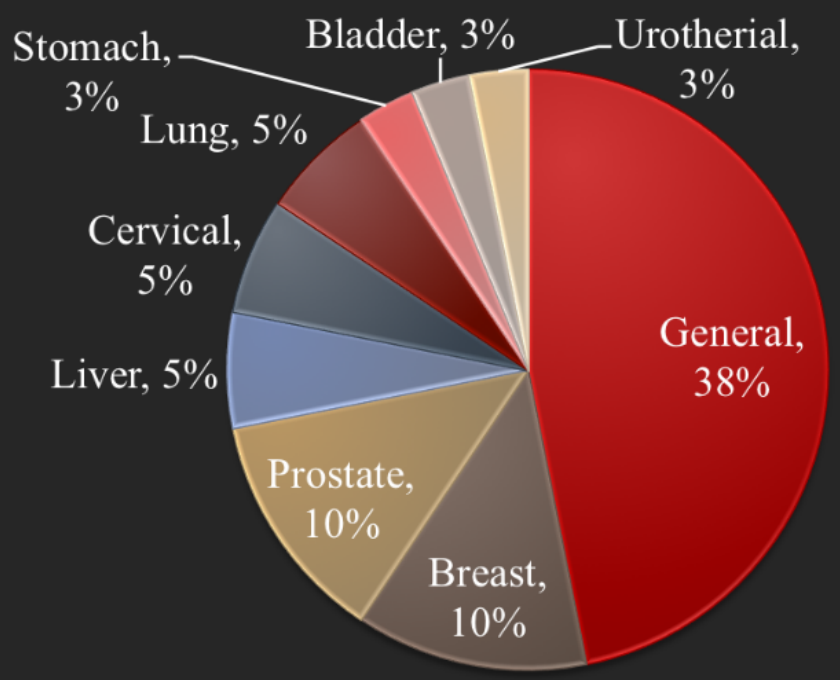
Applications:

- Screening
- Early detection
- Treatment

Prezi

Types

- Proteins
- Gene expression and DNA

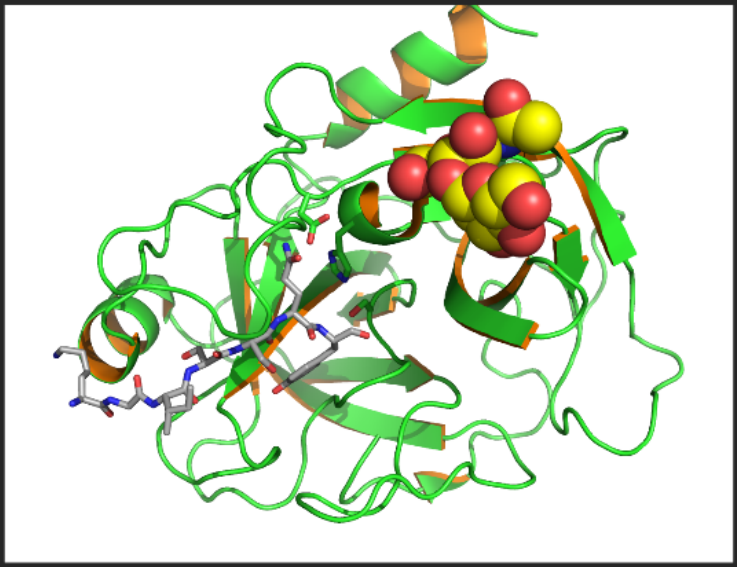


Prezi

Examples

Prostate specific antigen (PSA)

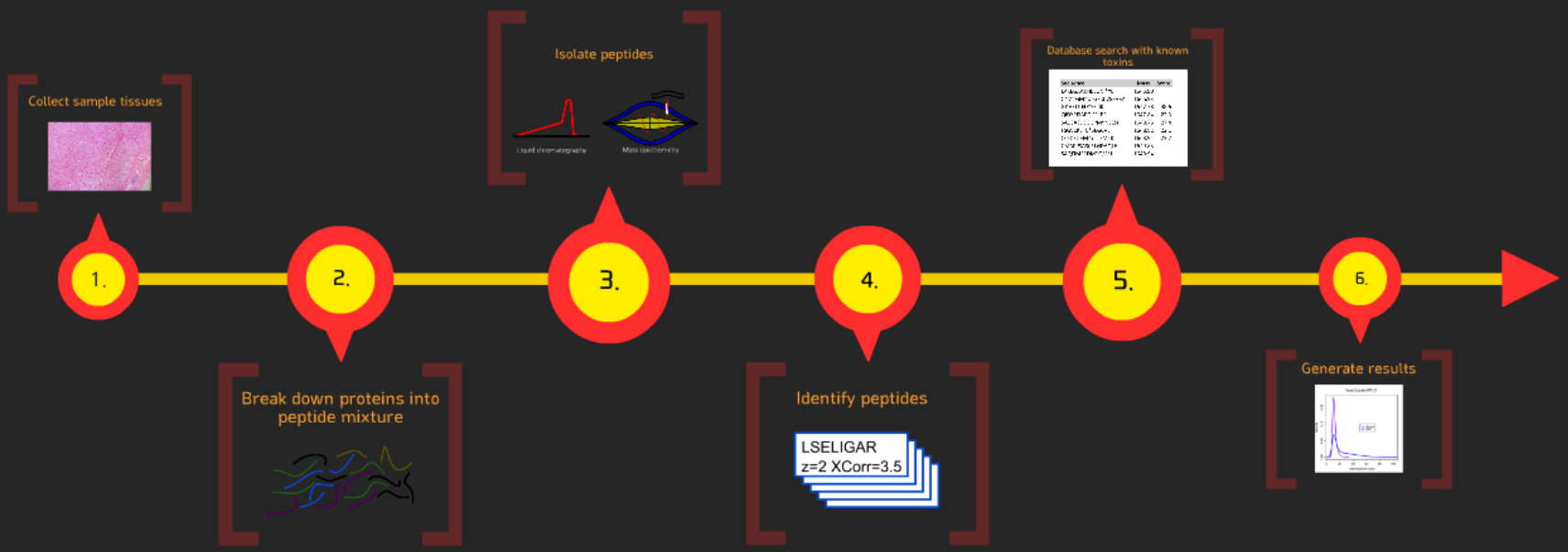
- Cancer type: prostate
- Substence type: glycoprotein
- Use: like other markers, currently screening is not recommended



Prezi

Cancer biomarker: discovery

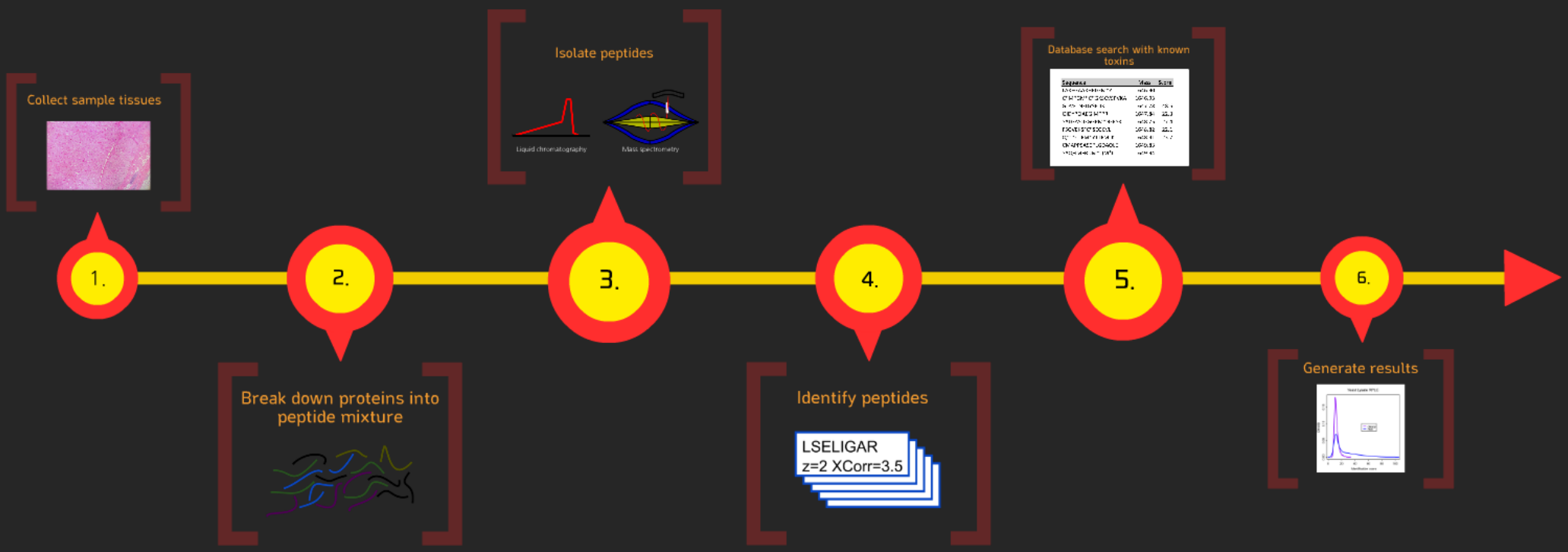
Example: protein identification



Prezi

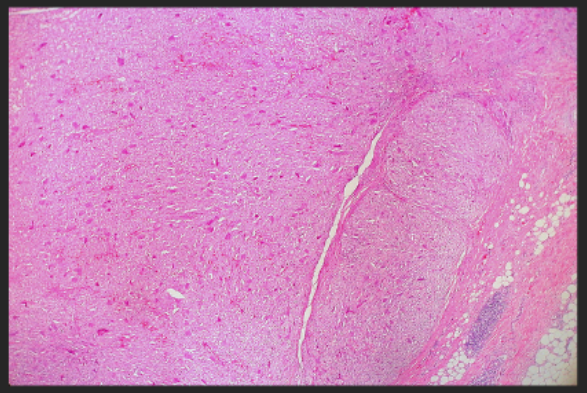
Cancer biomarker: discovery

Example: protein identification



Prezi

Collect sample tissues



1.

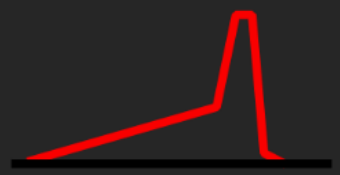
Prezi

2.

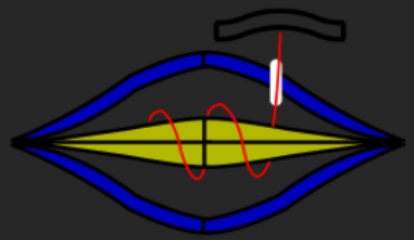
Break down proteins into peptide mixture

Prezi

Isolate peptides



Liquid chromatography



Mass spectrometry

3.

Prezi

4.

Identify peptides

LSELIGAR
z=2 XCorr=3.5

Prezi

Database search with known toxins

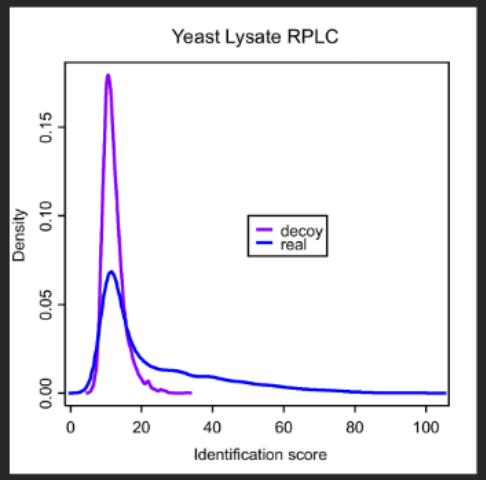
Sequence	Mass	Score
LAKEGAAKHELGM*YL	1646.90	
C*M*GM*C*GKSCVSPVKA	1646.93	
GDVFTNFDYSELIK	1647.78	48.6
QIDYPEIAEGIM*PR	1647.84	22.3
SALGASLGGGPM*NEESR	1648.75	27.4
FSGVEKSFC*SEGCVL	1648.82	22.1
Q*C*CLEM*YTTEMLK	1648.91	23.7
CMAPPSASC*LGDAQLE	1649.83	
SAQFIMFRDM*QM*L	1649.94	

5.

Prezi

6.

Generate results

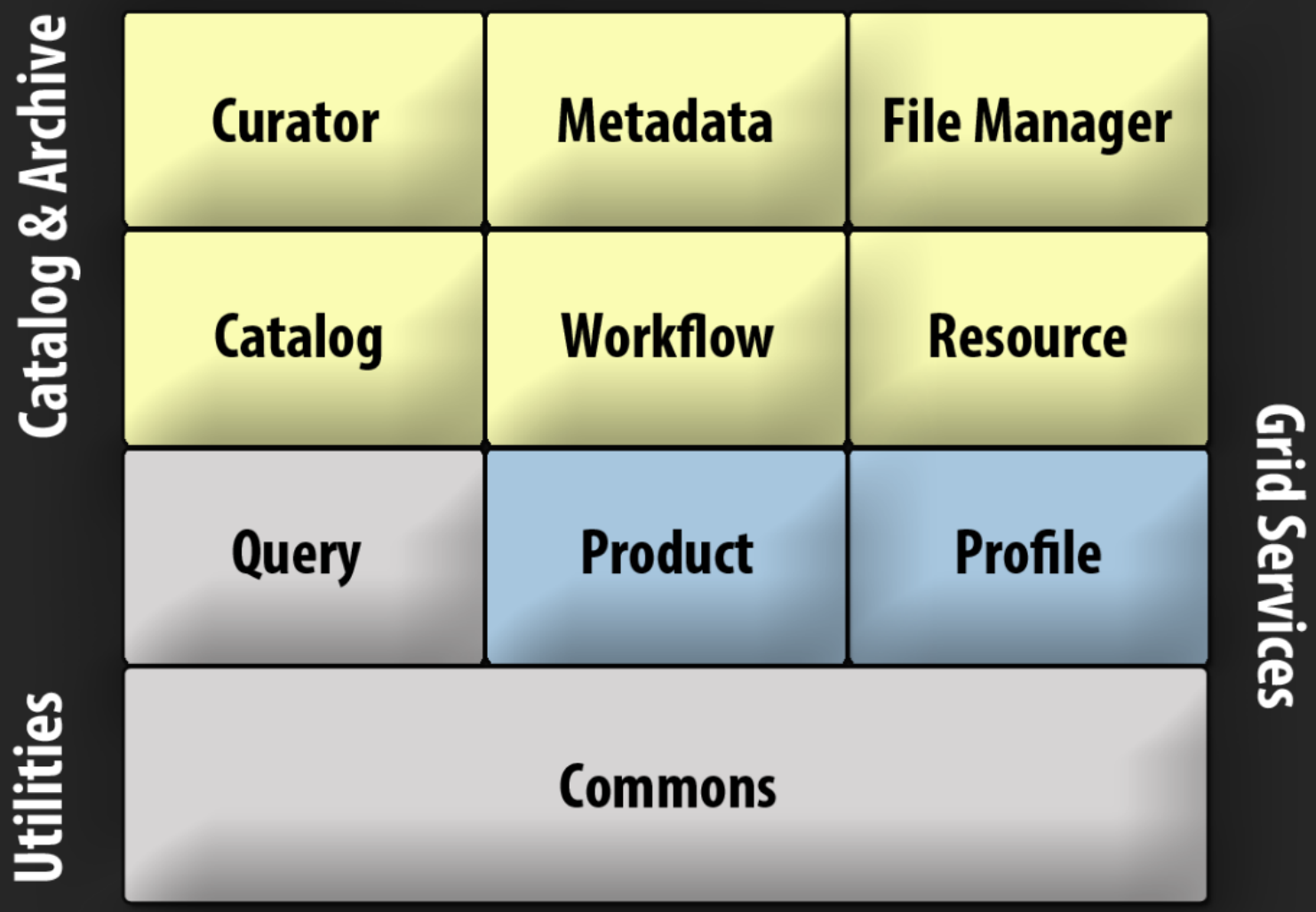


Prezi

Apache OODT and data processing

Prezi

What is Apache OODT?



Prezi

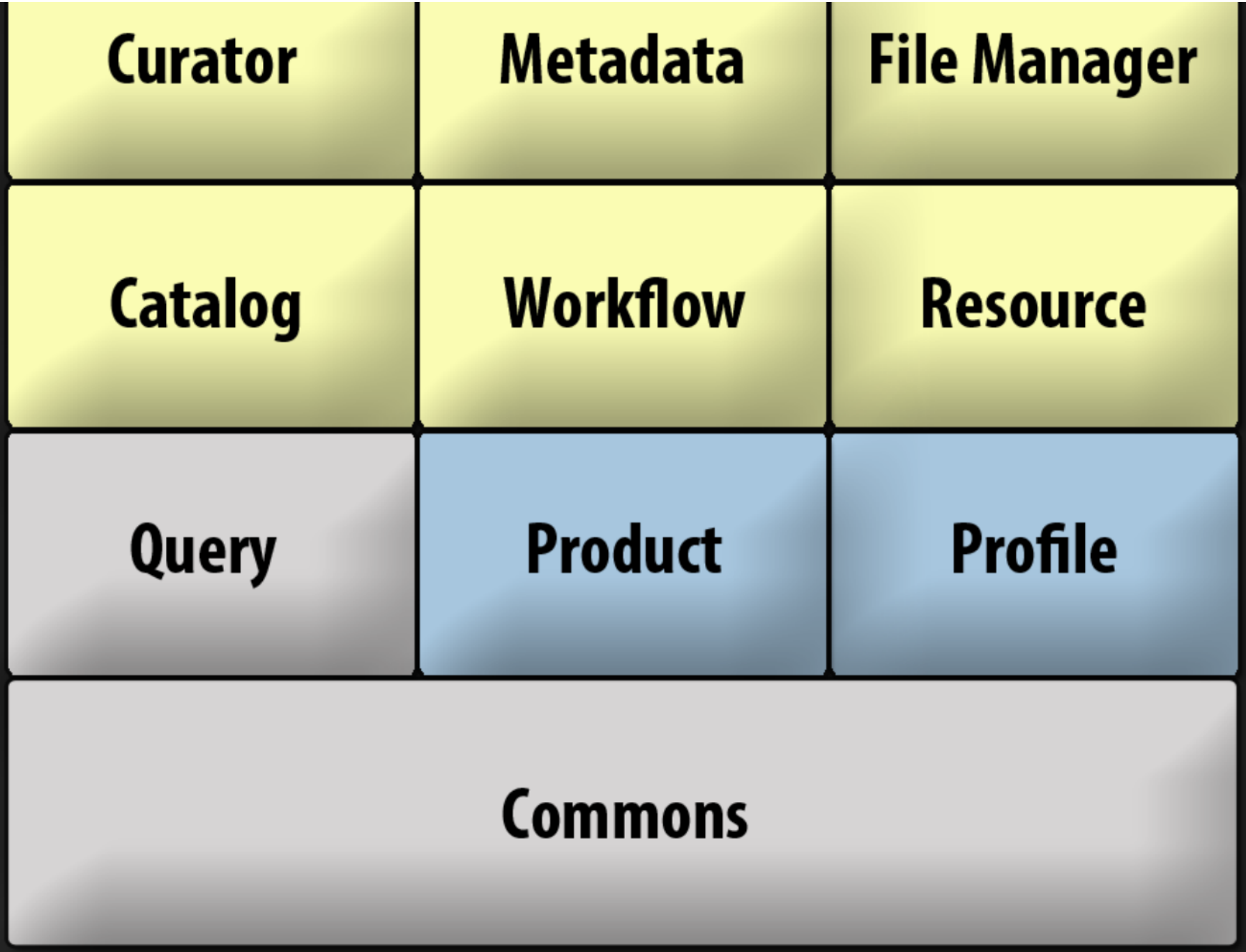
What is Apache OODT

Catalog & Archive

Curator	Metadata	File Manager
Catalog	Workflow	Resource
Query	Product	Profile

Grid Services

Prezi



Grid Services

Prezi

Ca

Query

Product

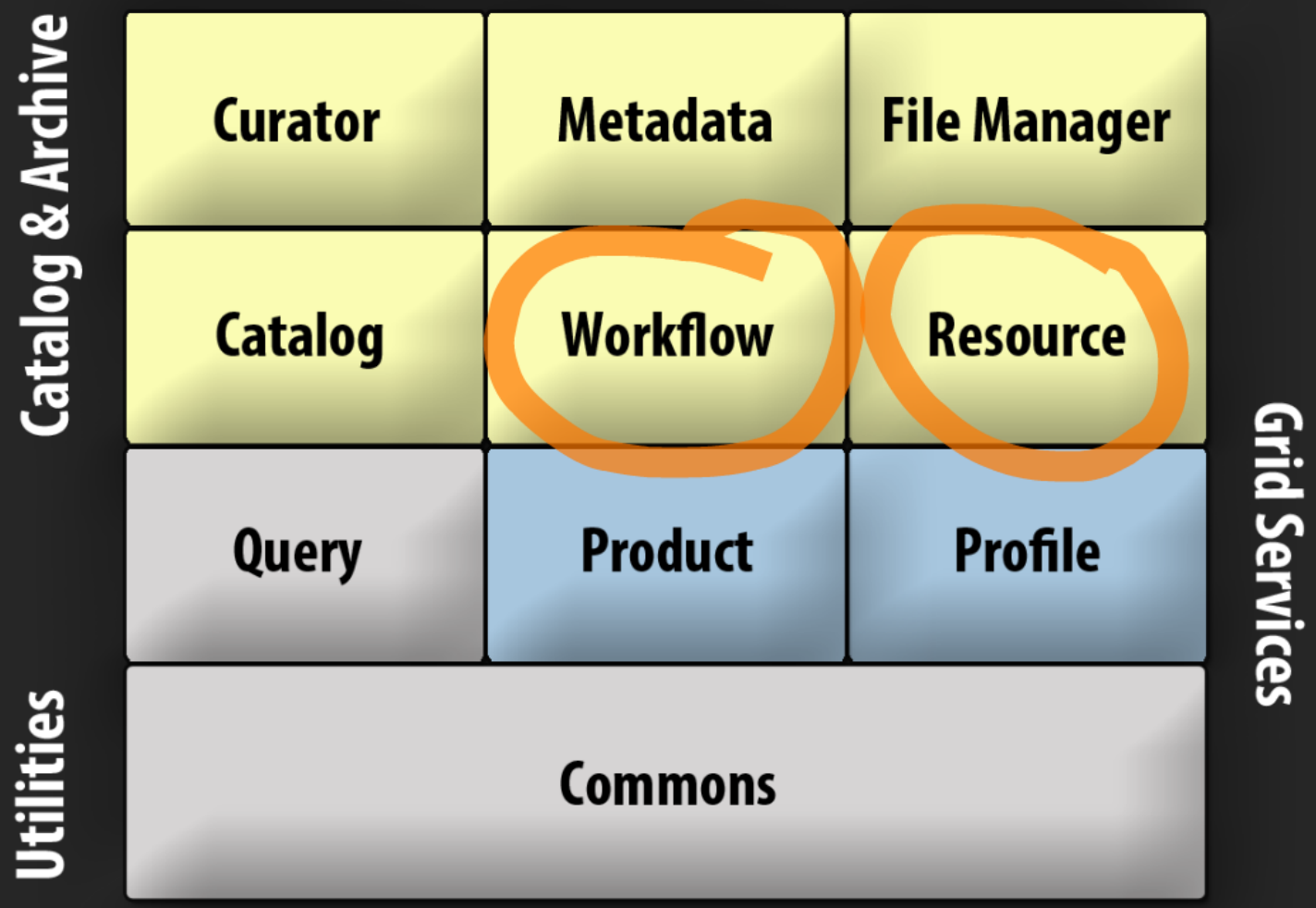
Pr

Utilities

Commons

Prezi

OODT Data Processing Components



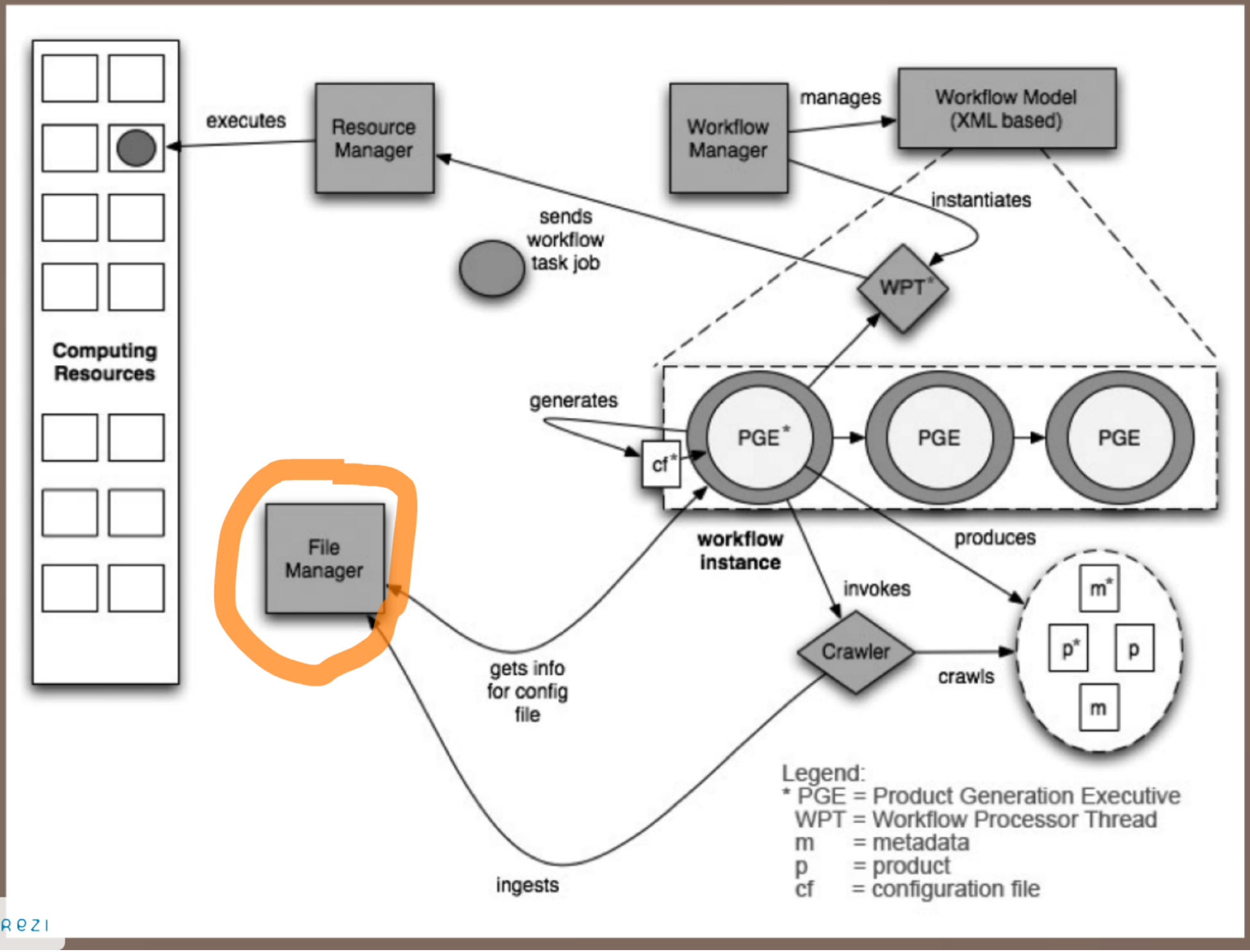
Prezi

Process Control System (PCS)

Prezi

Archive Management

Prezi


executes
Resource Manager
Workflow Manager
manages
Workflow Model (XML based)
instantiates
sends workflow task job
WPT*
Computing Resources
generates
cf*
PGE*
PGE
PGE
workflow instance
produces
File Manager
invokes
Crawler
crawls
m*
p*
p
m
gets info for config file
ingests
Legend:
* PGE = Product Generation Executive
WPT = Workflow Processor Thread
m = metadata
p = product
cf = configuration file

Process Control System (PCS)

Prezi

products

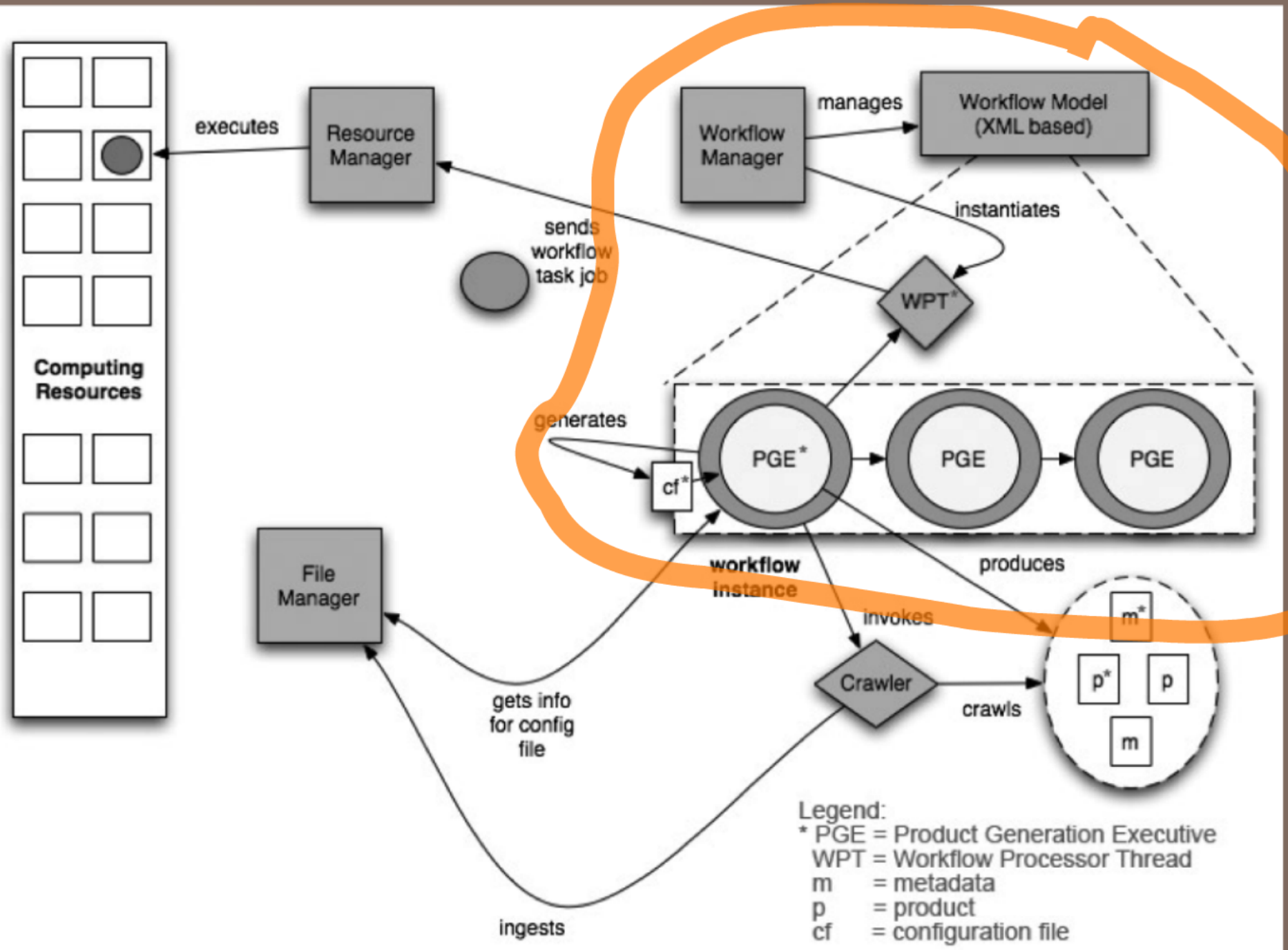
Workflow Manager

Compu

products

Science Algorithm Management

Prezi


Computing Resources
executes
Resource Manager
sends workflow task job
Workflow Manager
manages
Workflow Model (XML based)
instantiates
WPT*
generates
cf*
PGE*
PGE
PGE
workflow instance
produces
File Manager
invokes
Crawler
crawls
m*
p*
p
m
gets info for config file
ingests
Legend:
* PGE = Product Generation Executive
WPT = Workflow Processor Thread
m = metadata
p = product
cf = configuration file

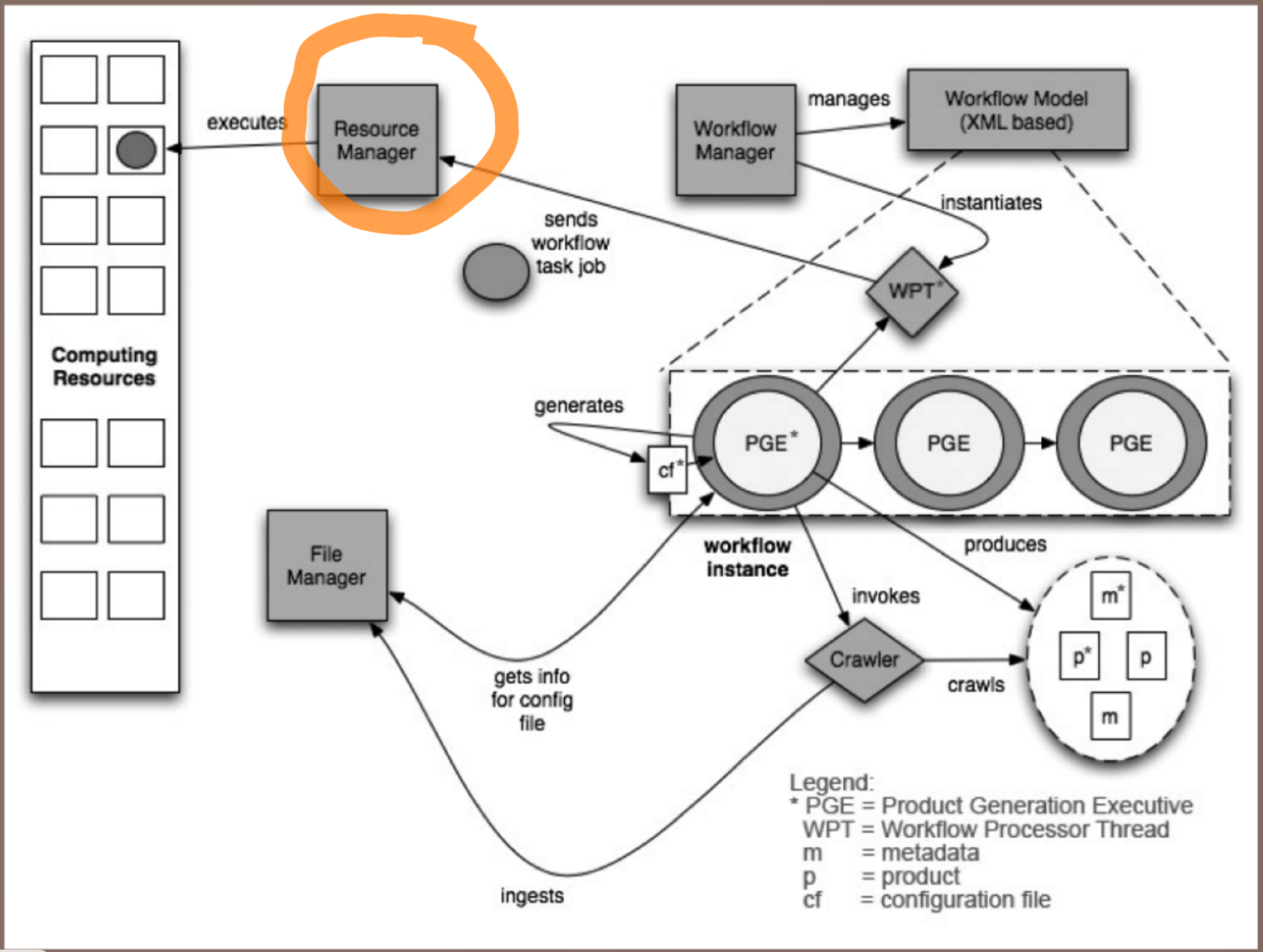
Process Control System (PCS)

Prezi

Resource Manager

Computation Management

Prezi


executes
Resource Manager
Computing Resources
Workflow Manager
manages
Workflow Model (XML based)
instantiates
sends workflow task job
WPT*
generates
cf*
PGE*
PGE
PGE
workflow instance
produces
File Manager
invokes
Crawler
crawls
m*
p*
p
m
gets info for config file
ingests
Legend:
* PGE = Product Generation Executive
WPT = Workflow Processor Thread
m = metadata
p = product
cf = configuration file

Example: EDRN LabCAS Proteomics Pipeline

Prezi

National Cancer Institute

Early Detection Research Network

Biomarkers: the key to early detection

- Founded in 2000
- Objectives:
  - Identify clinically applicable cancer biomarkers
  - Connect disparately located researchers
  - Establish a national biomarker research infrastructure

Prezi

EDRN LabCAS and the Proteomics Pipeline

Laboratory Catalog and Archive Service (LabCAS)

- Data management infrsastructure for EDRN researchers
- Targeted at pre-publication, experimental research data

Proteomics Pipeline

- Centralized data processing system for LabCAS
- Configurable workflows for analyzing tissue sample data using diverse algorithms
- Generation of higher-level products

Prezi

EDRN LabCAS and the Proteomics Pipeline

Laboratory Catalog and Archive Service (LabCAS)

- Data management infrsastructure for EDRN researchers
- Targeted at pre-publication, experimental research data

Proteomics Pipeline

- Centralized data processing system for LabCAS
- Configurable workflows for analyzing tissue sample data using diverse algorithms
- Generation of higher-level products

Prezi

Laboratory Catalog and Archive Service (LabCAS)

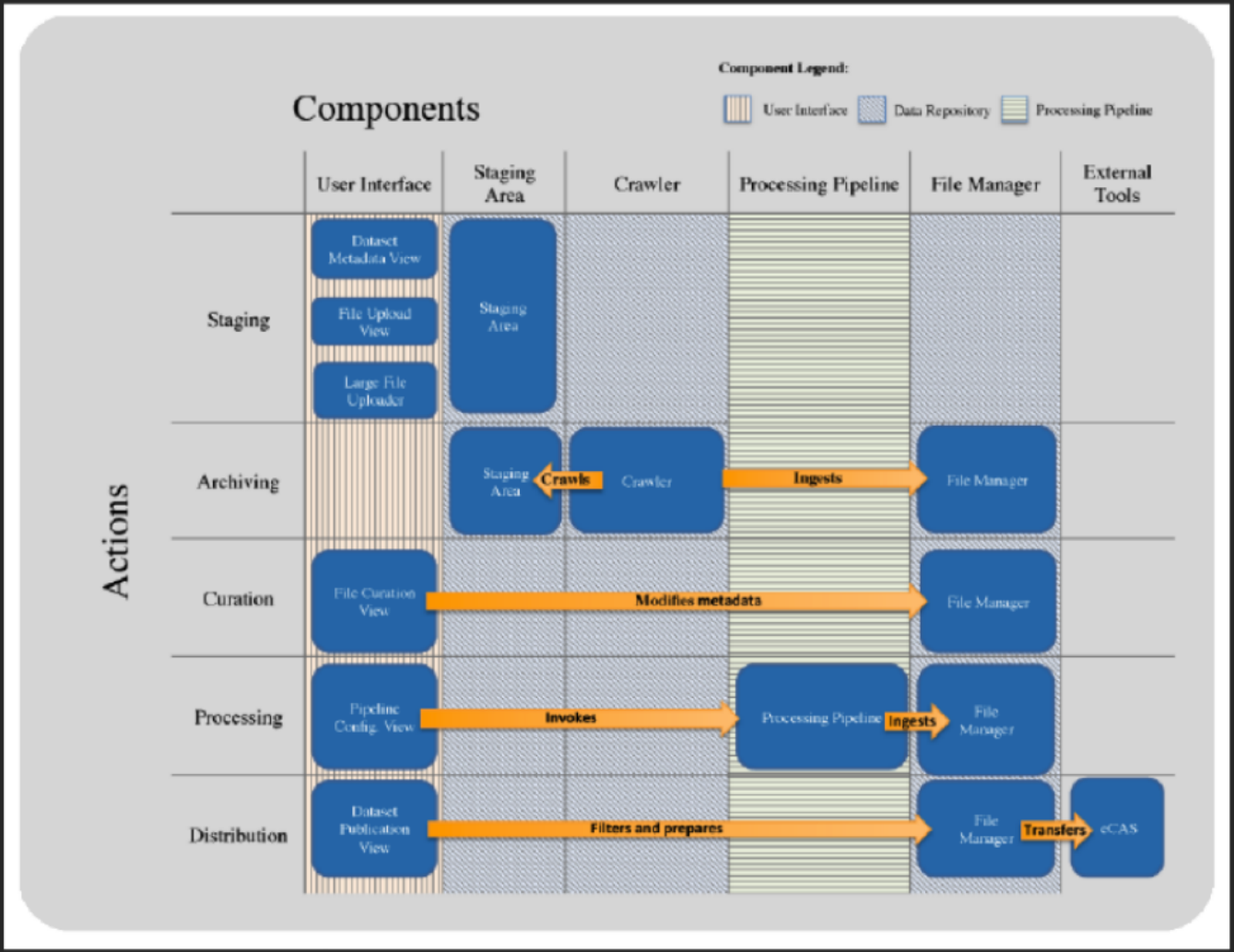
- Data management infrsastructure for EDRN researchers
- Targeted at pre-publication, experimental research data

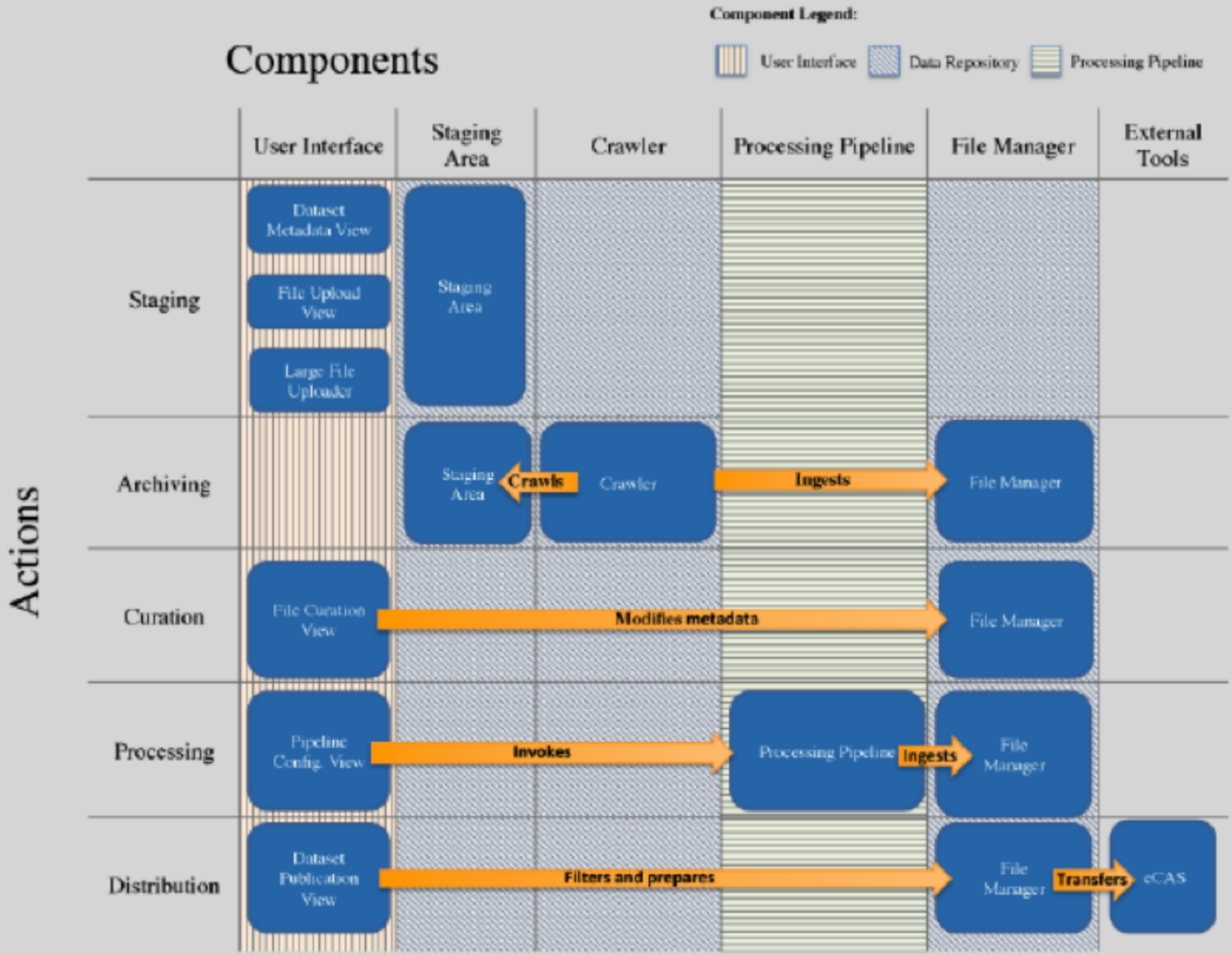
Proteomics Pipeline

- Centralized data processing system for LabCAS
- Configurable workflows for analyzing tissue sample data using diverse algorithms
- Generation of higher-level products

Prezi

Leveraging the OODT framework




Component Legend:
User Interface
Data Repository
Processing Pipeline
Components
User Interface
Staging Area
Crawler
Processing Pipeline
File Manager
External Tools
Actions
Staging
Archiving
Curation
Processing
Distribution
Dataset Metadata View
File Upload View
Large File Uploader
Staging Area
Staging Area
Crawls
Crawler
Ingests
File Manager
File Curation View
Modifies metadata
File Manager
Pipeline Config. View
Invokes
Processing Pipeline
Ingests
File Manager
Dataset Publication View
Filters and prepares
File Manager
Transfers
eCAS

LabCAS Demonstration

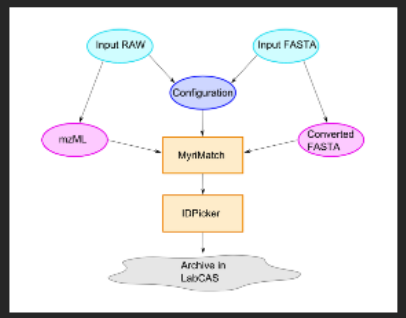
http://goo.gl/24M4U

Prezi

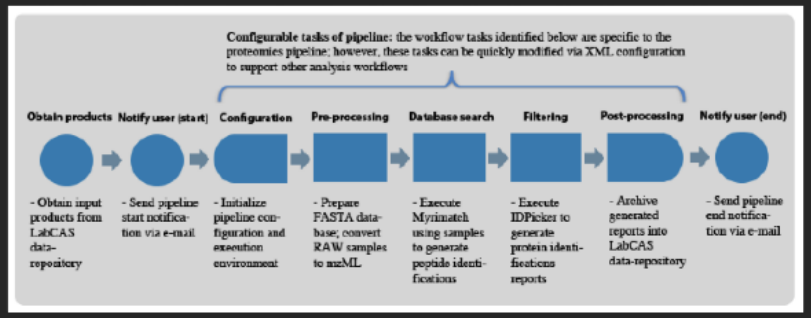
Proteomics Pipeline

- Centralize the the analysis of LabCAS data sets
- Make available a variety of analysis algorithms

Algorithm

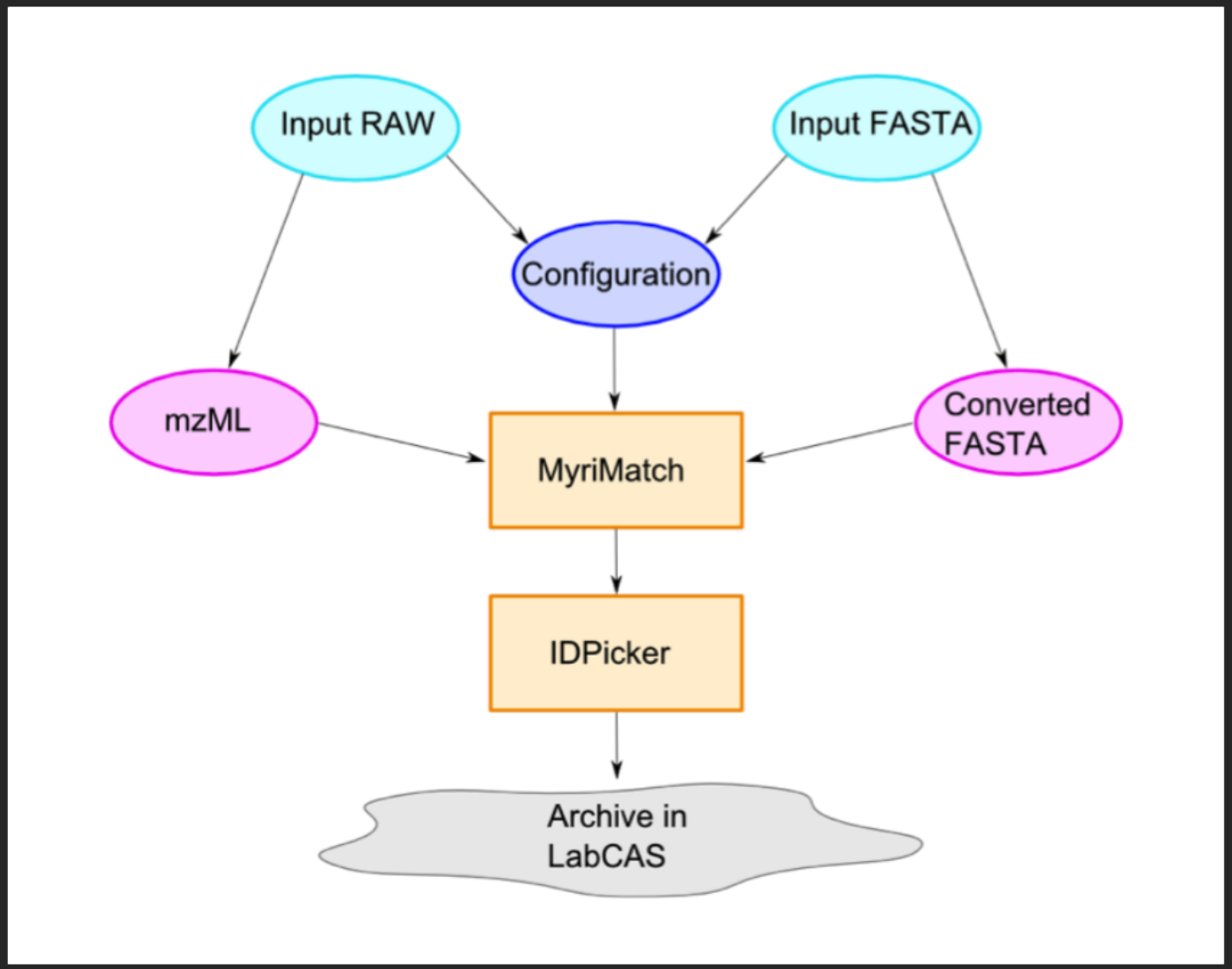


Workflow Implementation



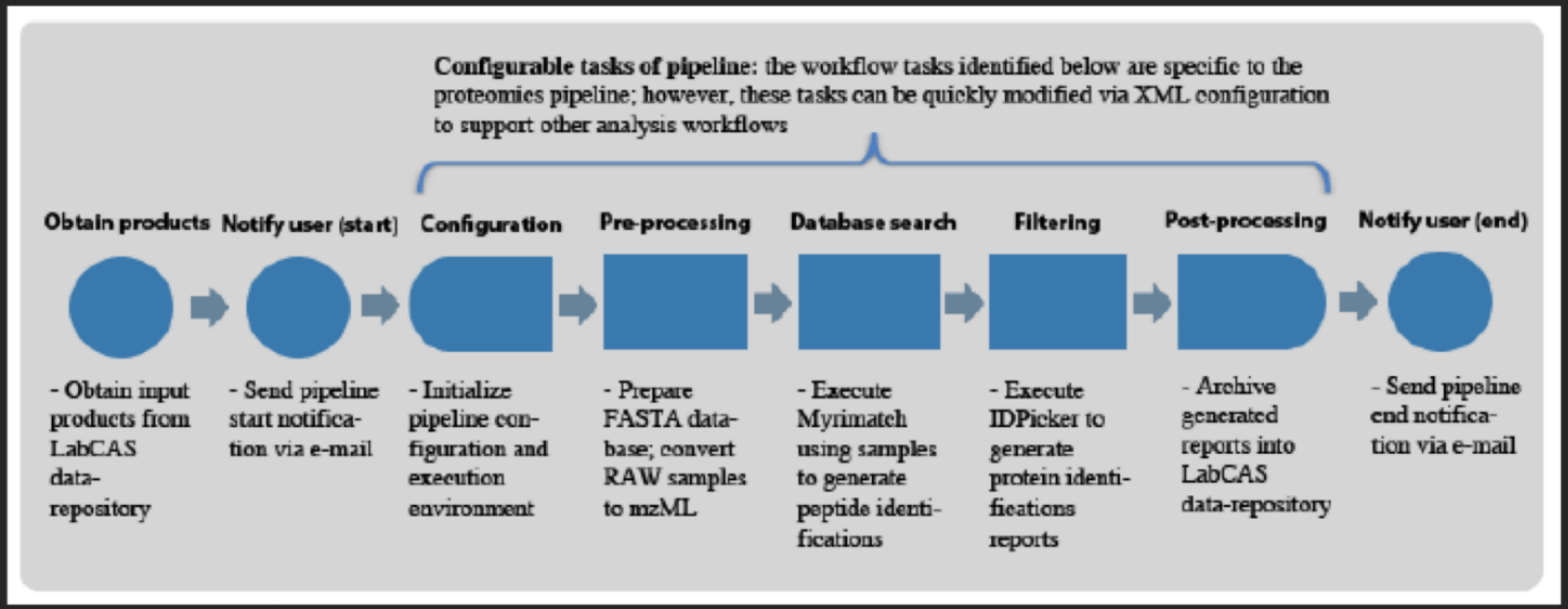
Prezi

Algorithm



Prezi

Workflow Implementation



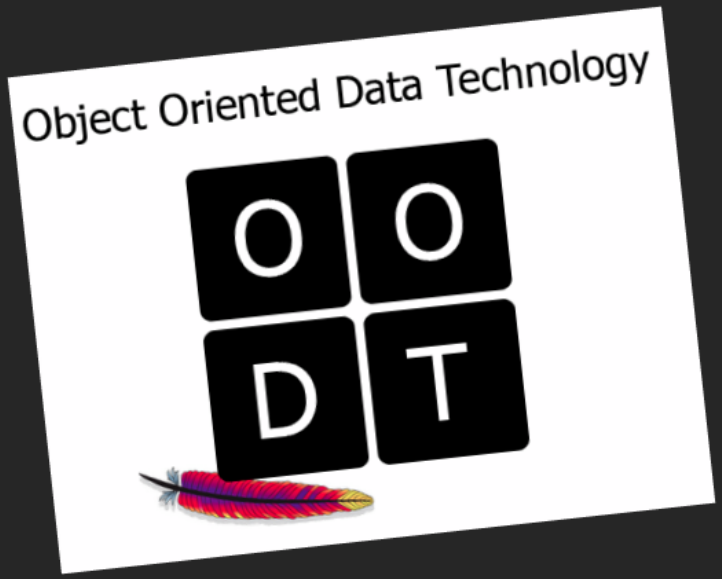
Prezi

Proteomics Pipeline Demonstration

http://goo.gl/7sZjl

Prezi

Tutorial: Build your own processing system



Prezi

Review of steps:

1. Download OODT PCS framework:
   - File Manager
   - Workflow Manager
   - Resource Manager
   - Crawler ...
2. Configure components
3. Wrap science algorithms using PGEs
4. Configure user interface

Prezi

Easier method?

Try Apache OODT's RADiX tool!

- Contains distribution of Apache OODT
- Convention over configuration
- Reduce ramp up time
- Eliminate some configuration

Details:

- Uses Apache Maven Archetype construct
- Auto-generation of OODT base project

Prezi

Using OODT RADiX

```
% mvn archetype:generate -DarchetypeGroupId=org.apache.oodt -DarchetypeArtifactId=radix-archetype -DarchetypeVersion=0.6-SNAPSHOT -Doodt=0.6-SNAPSHOT -DgroupId=com.mycompany -DartifactId=dms -Dversion=0.1-SNAPSHOT
% cd dms; mvn package
% tar -xvf distribution/target/dms-distribution-0.1-SNAPSHOT-bin.tar.gz -C ../oodt-deploy
% cd ../oodt-deploy
% ./bin/oodt start
```

Try out the base deployment, then:

- Customize your workflows by wrapping science algorithms
- Customize your resource management strategy
- Customize your archival strategy

Prezi

Using OODT RADiX

```
% mvn archetype:generate -DarchetypeGroupId=org.apache.oodt -DarchetypeArtifactId=radix-archetype -DarchetypeVersion=0.6-SNAPSHOT -Doodt=0.6-SNAPSHOT -DgroupId=com.mycompany -DartifactId=dms -Dversion=0.1-SNAPSHOT
% cd dms; mvn package
% tar -xvf distribution/target/dms-distribution-0.1-SNAPSHOT-bin.tar.gz -C ../oodt-deploy
% cd ../oodt-deploy
% ./bin/oodt start
```

Try out the base deployment, then:

- Customize your workflows by wrapping science algorithms
- Customize your resource management

Prezi

```
deploy
% cd ../oodt-deploy
% ./bin/oodt start
```

Try out the base deployment, then:

- Customize your workflows by wrapping science algorithms
- Customize your resource management strategy
- Customize your archival strategy

Prezi

Thanks! Questions?

Contact: Rishi Verma
- E-mail: Rishi.Verma@jpl.nasa.gov
- Twitter: @reesh24

Presentation contributions:
- Cameron Goodale, NASA Jet Propulsion Laboratory
- David Tabb, Vanderbilt University Medical Center
- Chris Mattmann, NASA Jet Propulsion Laboratory
- Paul Ramirez, NASA Jet Propulsion Laboratory

Bibliography:
- http://www.cancer.gov/cancertopics/factsheet/Detection/tumor-markers
- http://en.wikipedia.org/wiki/Cancer_biomarkers
- http://sunset.usc.edu/~mattmann/pubs/SMCIT09.pdf

Prezi