On Discovering Data Preparation Modules Using Examples

Khalid Belhajjame, PSL, Paris-Dauphine University
kbelhajj@gmail.com

PROVENANCE WEEK 2020

Data Preparation

Despite the impressive body of work in the field of data preparation, there is no single generic one-shop-stop solution that can be utilized by the scientists to prepare their data prior their analysis.

Data preparation tasks are numerous, can be difficult to generalize.

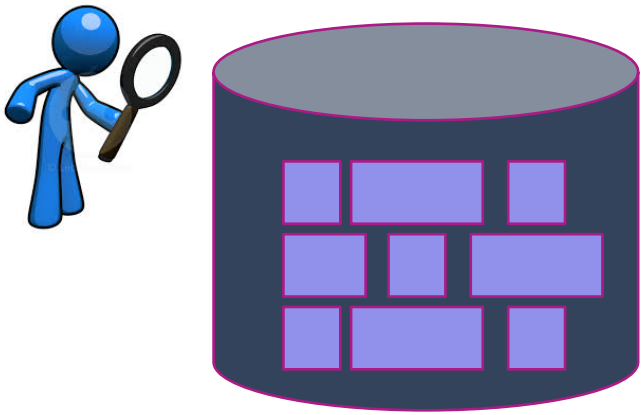
Scientists tend to develop their own program/script using their favorite language, e.g., Python, R or Perl, to prepare their data.

To overcome the above problem, several researchers have been calling for the creation of repositories dedicated to scientific modules in general, such as Bio.Tools and Galaxy Tools, and data preparation tasks in particular, e.g., BigGorilla

Galaxy Tool Shed

elixir bio.tools

BigGorilla



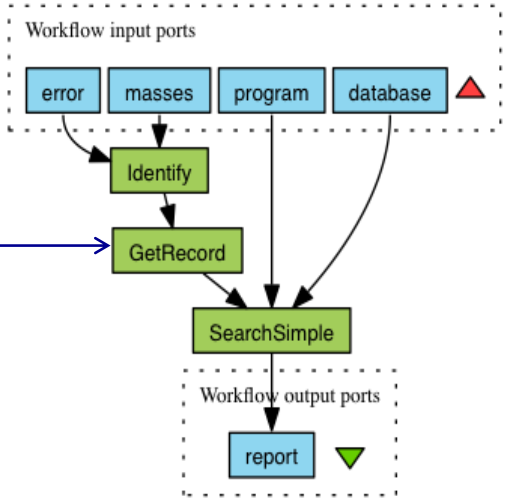
Repository of Modules

Problem Investigated

Given a repository of data preparation modules, how can we assist scientists in their exploration and discovery ?

Data Examples

Data Example



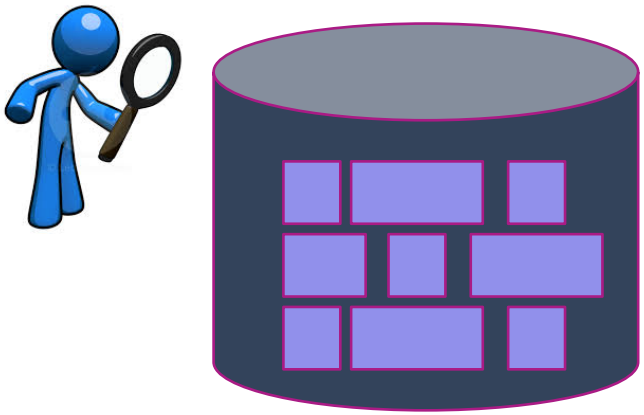
Describes >

GetRecord.i

P17110

GetRecord.o

>sp|P17110|CH36_CERCA Chorion protein S36 OS=Ceratitis capitata
MNCFLFTLFFVAAPLATASYGSSSGGGGGGSSYLSSASSNGLDELVQAAAGGAQQAGGTI
TPANAEIPVSPAEVARLNQVQAQLQALNSNPVYRNLKNSDAIAESLAESSLASKIRQGNI
NIVAPNVIDQGVYRSLLVPSGQNNHQVIATQPLPPIIVNQPALPPTQIGGGPAAVVKAAP
VIYKIKPSVIYQQEVINKVPTPLSLNPVYVKVYKPGKKIDAPLVPGVQQNYQAPSYGGSS
YSAPAASYEPAPAPSYSAAPAQSYNAAPAPSYSAAPAASYGAAPSASYDAAPAASYGAES
SYGSPQSSSSYGSAPPASGY



Repository of Modules

Research Problem 1

How to select/generate the data examples that characterize modules ?

Using modules' metadata and retrospective provenance of the modules' executions [1]

Research Problem 2

How to explore/discover modules using data examples?

[1] K. Belhajjame, On characterizing scientific modules using data examples, in the proceeding of EDBT, 2014.

Module Discovery

- To discover a module, a user can provide data examples that characterize the module
  - specifying data examples that characterize the desired module can be time-consuming
- Instead, we make use of metadata (semantic annotations) to help the user narrow down the candidate module that need to be explored.
  - In particular, the user specify the domain of the input and output of the desired modules.

	Data examples		User feedback	
	protein name	**accession**	**expected**	**unexpected**
δ_1	Chorion protein S36	CH36_CERCA	X	
δ_2	Zinc metalloproteinase	VMDM_VIRST		X

- The user then examines the data examples of the candidate modules and specifies the ones that meet the expectations and the ones that do not by labelling the data examples.

Issues that needs to be addressed

- How to reduce the number of data examples that need to be displayed to the user?
- A module that meets the behavior expected by the user may not exist in the repository.
  - What is the module that meets best user needs?
    - Module Ranking

On Discovering Data Preparation Modules Using Examples

Khalid Belhajjame, PSL, Paris-Dauphine University
kbelhajj@gmail.com

PROVENANCE WEEK 2020