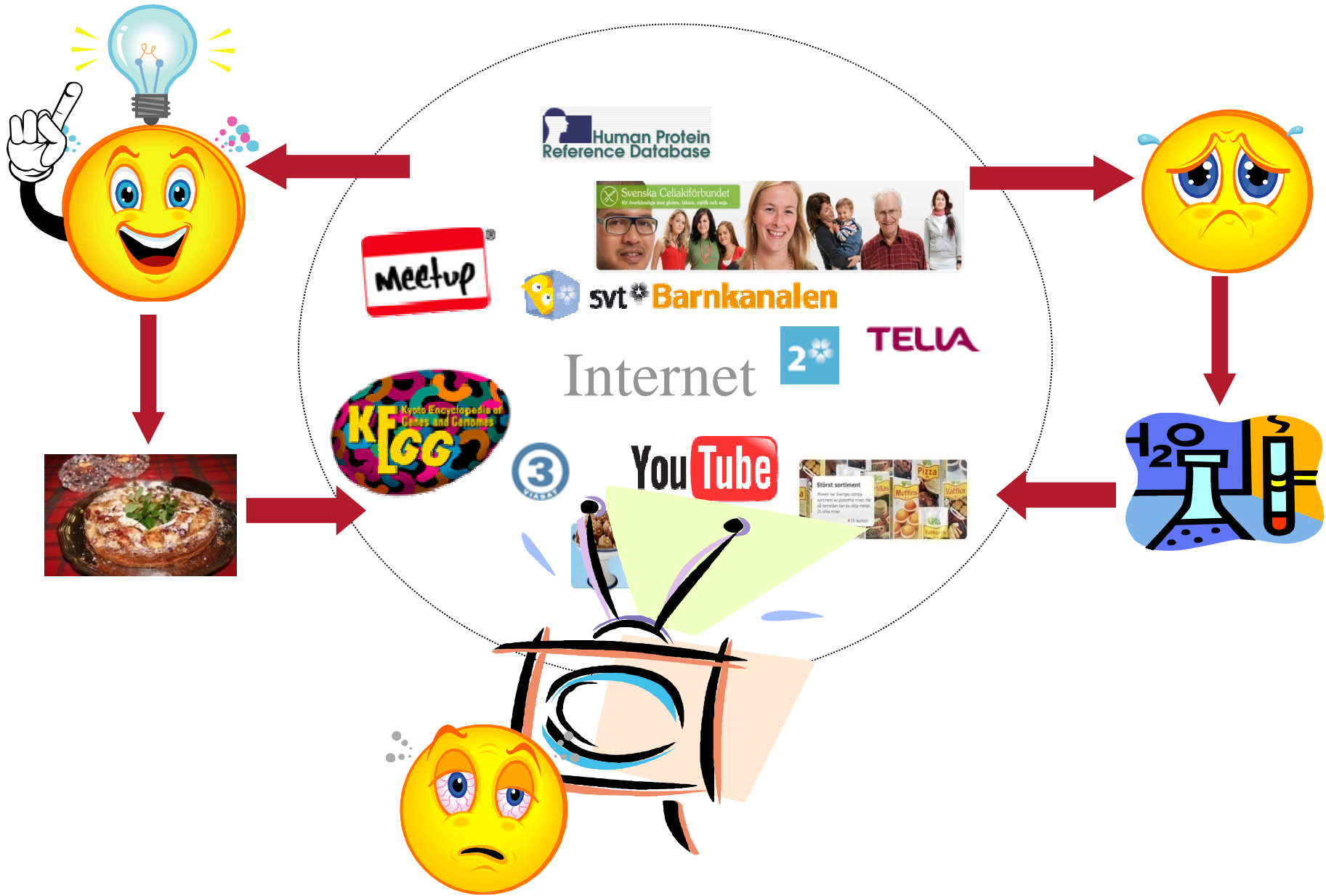


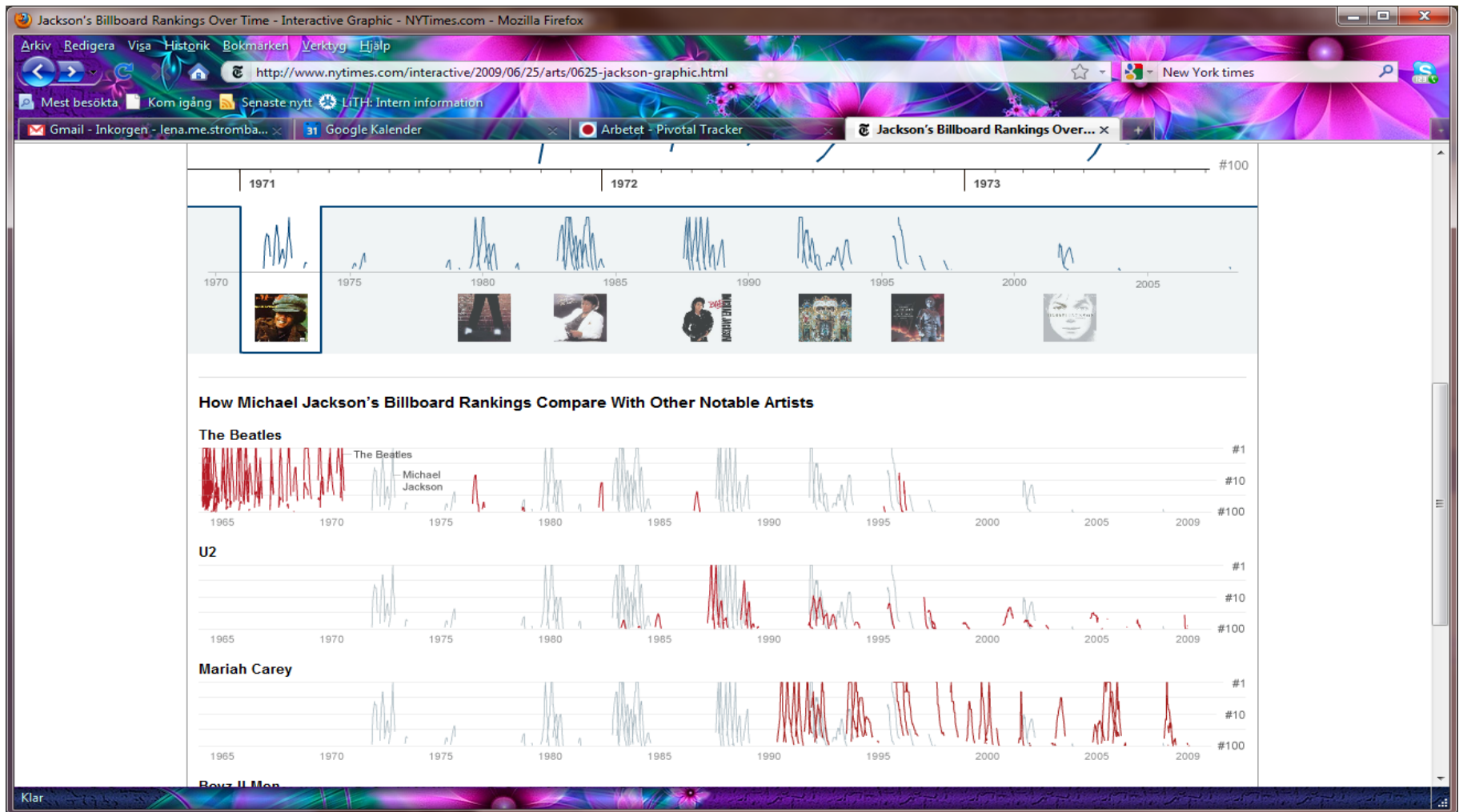
Effective Management and Exploration of Scientific Data on the Web.

Lena Strömbäck
lena.stromback@liu.se

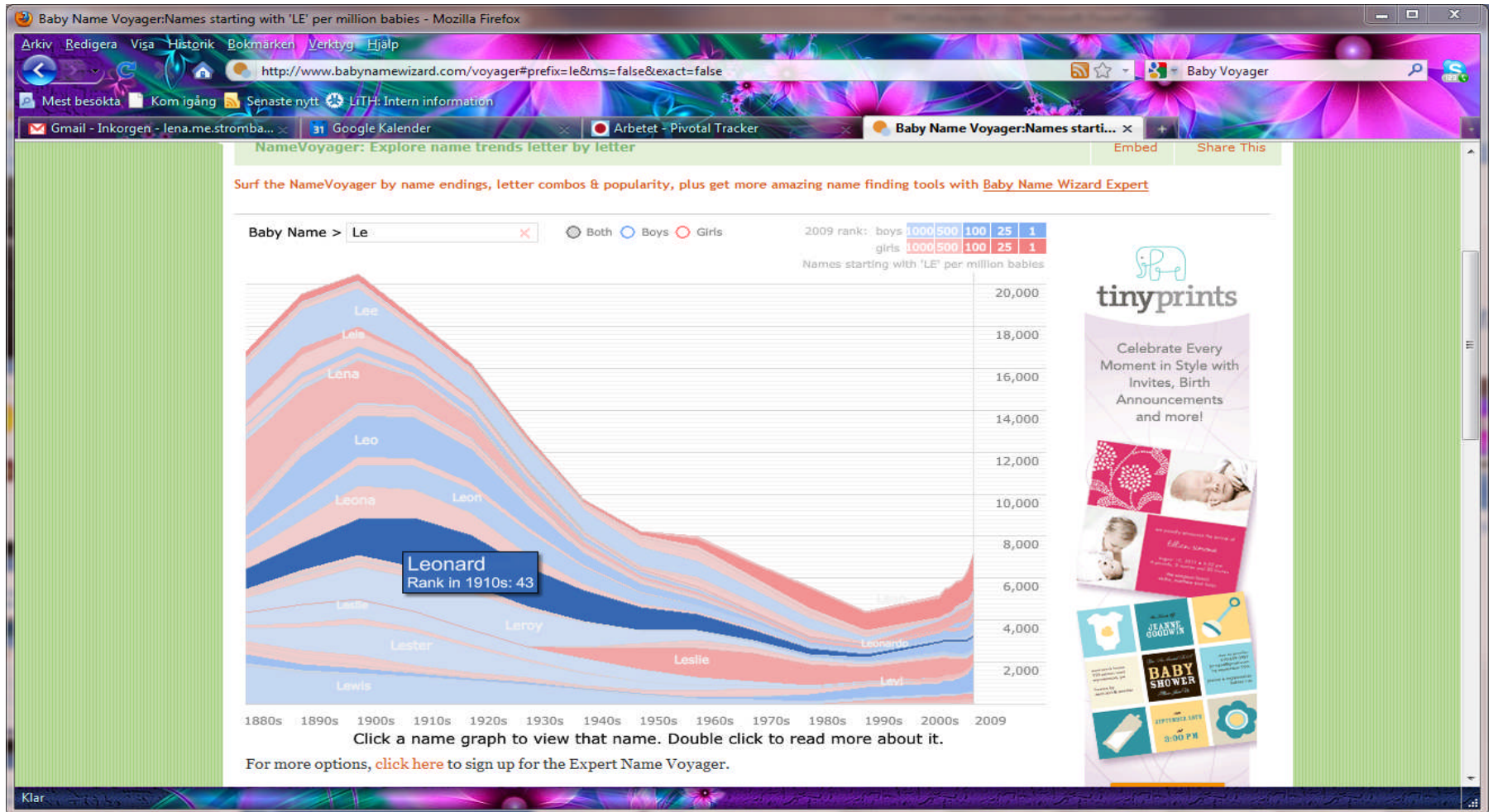
Linköping University



Example: New York Times



Example: Baby Name Vizard Laura Wattenberg – Generation Grownup



Example: Many Eyes

IBM Research and the IBM Cognos software group

Many Eyes : Browsing visualizations - Mozilla Firefox

Arkiv Redigera Visa Historik Bokmärken Verktyg Hjälp

http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations?page=2

My Eyes

Mest besökta Kom igång Senaste nytt LITH: Intern information

Gmail - Inkorgen - lena.me.stromba... 31 Google Kalender Arbetet - Pivotal Tracker Many Eyes : Browsing visualizations

Many Eyes Log in IBM

An experiment brought to you by IBM Research and the IBM Cognos software group

Visualizations Search

Listing Visualizations

Subscribe

Showing 31-60 of 79077

Previous 1 2 3 4 5 6 7 8 9 ... 2635 2636 Next

Sorted by date Sort by rating

Thumbnail	Title	Time
[Blank]	MA PSP website text	Yesterday at 02:54 PM
[Colorful dots]	Name Data Chart	Yesterday at 02:43 PM
[Bar chart]	Egienvector Centrality in Upstream Hydrocarbon Investments	Yesterday at 02:37 PM
[Heatmap]	Student Ages	Yesterday at 02:29 PM
[Colorful dots]	NAMES	Yesterday at 02:25 PM
[Colorful dots]	Student Names	Yesterday at 02:23 PM
[Text]	terra 11 vezes	Yesterday at 01:47 PM
[Colorful dots]	2010 Major League Baseball Salaries	Yesterday at 01:36 PM
[Word cloud]	11 anos de bamba que valem	Yesterday at 01:11 PM
[Word cloud]	11 anos de bambas	Yesterday at 12:52 PM
[Colorful dots]	Uncommitted pledges of Aid to Haiti by country 2010	Yesterday at 12:45 PM
[Treemap]	SHU Computing modules treemap	Yesterday at 12:14 PM

E-Science data

- Complex data
- Not easily human interpretable
- Need for integration and comparison
- Powerful computation needed

To further complicate the task

- Standardization and agreement of common formats is a prerequisite for efficient data management
- The Web is an ad-hoc platform where new data formats and actors occurs all the time

Content of this presentation:

- Two scientific application areas
 - Provenance/Scientific workflows
 - Bioinformatics
- Three different aspects
 - Interfaces for exploration
 - Seamless data integration
 - Effective data exploration

Content of this presentation:

- **Two scientific application areas**
 - **Provenance/Scientific workflows**
 - **Bioinformatics**
- Three different aspects
 - Interfaces for exploration
 - Seamless data integration
 - Effective data exploration

Biological data

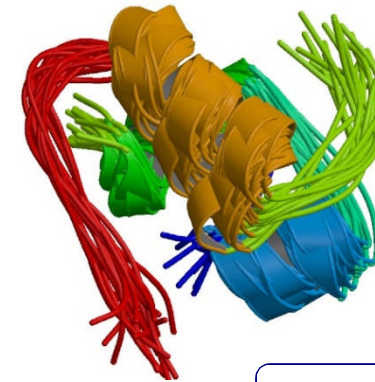
LOCUS HUMINS01 4044 bp DNA
 DEFINITION Human insulin gene, complete cds.
 ACCESSION J00265
 VERSION J00265.1 GI:186429

```

3841 aggggccagg gatggtgggg ccaactgagaa gtgacttctt gttcagtagc tctggactct
3901 tggagtcgcc agagaccttg ttcaggaaaag ggaatgagaa cattccagca attttccccc
3961 cacctagccc tcccaggttc tatttttaga gttatttctg atggagtcce tgtggaggga
4021 ggaggctggg ctgagggagg ggggt
  
```

DNA seq.

GenBank



Tertiary str.
PDB

Entry information

Entry name **INS_HUMAN**
 Primary accession number **P01308**

Comments

- FUNCTION:** Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.
- SUBUNIT:** Heterodimer of a B chain and an A chain linked by two disulfide bonds.
- SUBCELLULAR LOCATION:** Secreted.
- DISEASE:** Defects in INS are the cause of familial hyperproinsulinemia [MIM:176730]

```

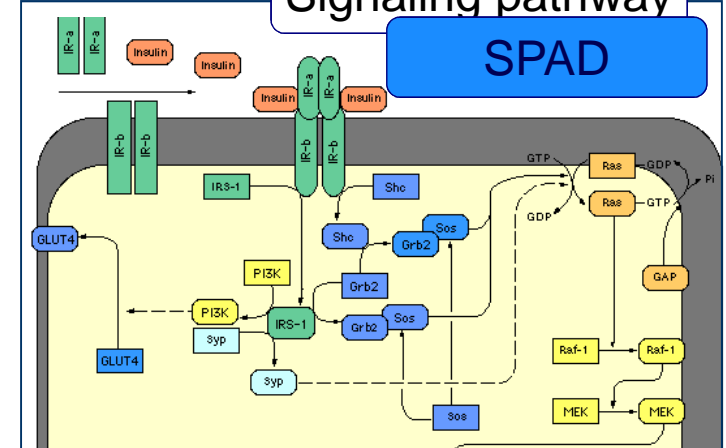
10 20 30 40 50 60
MALWMRLRLPL LALLALWGPD PAAAFVNHQL CGSHLVEALY LVCGERGFY TPKTREAED
70 80 90 100 110
LQVQVELGG GPGAGSLQPL ALEGLQKRG IVEQCCTSLC SLYQLENYCN
  
```

Protein seq.

SWISS-PROT

Signaling pathway

SPAD



General information about the entry

Entry name **INSULIN**
 Accession number **PS00262**
 Entry type **PATTERN**

Name and characterization of the entry

Description Insulin family signature.
 Pattern C-C- {P} - {P} -x-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C.

Secondary str.

PROSITE

INSULIN

Taxonomy

AmiGO

all: all (184297)
 GO:0003674 : molecular function (184297)
 GO:0005488 : binding (33171)
 GO:0005102 : receptor binding (1864)
 GO:0005179 : hormone activity (325)
 GO:0004871 : signal transducer activity (8874)
 GO:0005102 : receptor binding (1864)
 GO:0005179 : hormone activity (325)

Capturing provenance

- Provenance of scientific artifacts is necessary to reproduce, validate and share scientific results
- Provenance can be as important as the results!

▼ ————— *Dictionary* —————

prov•e•nance |ˈprāvənəns|

noun

the place of origin or earliest known history of something : *an orange rug of Iranian provenance.*

- the beginning of something's existence; something's origin : *they try to understand the whole universe, its provenance and fate.*

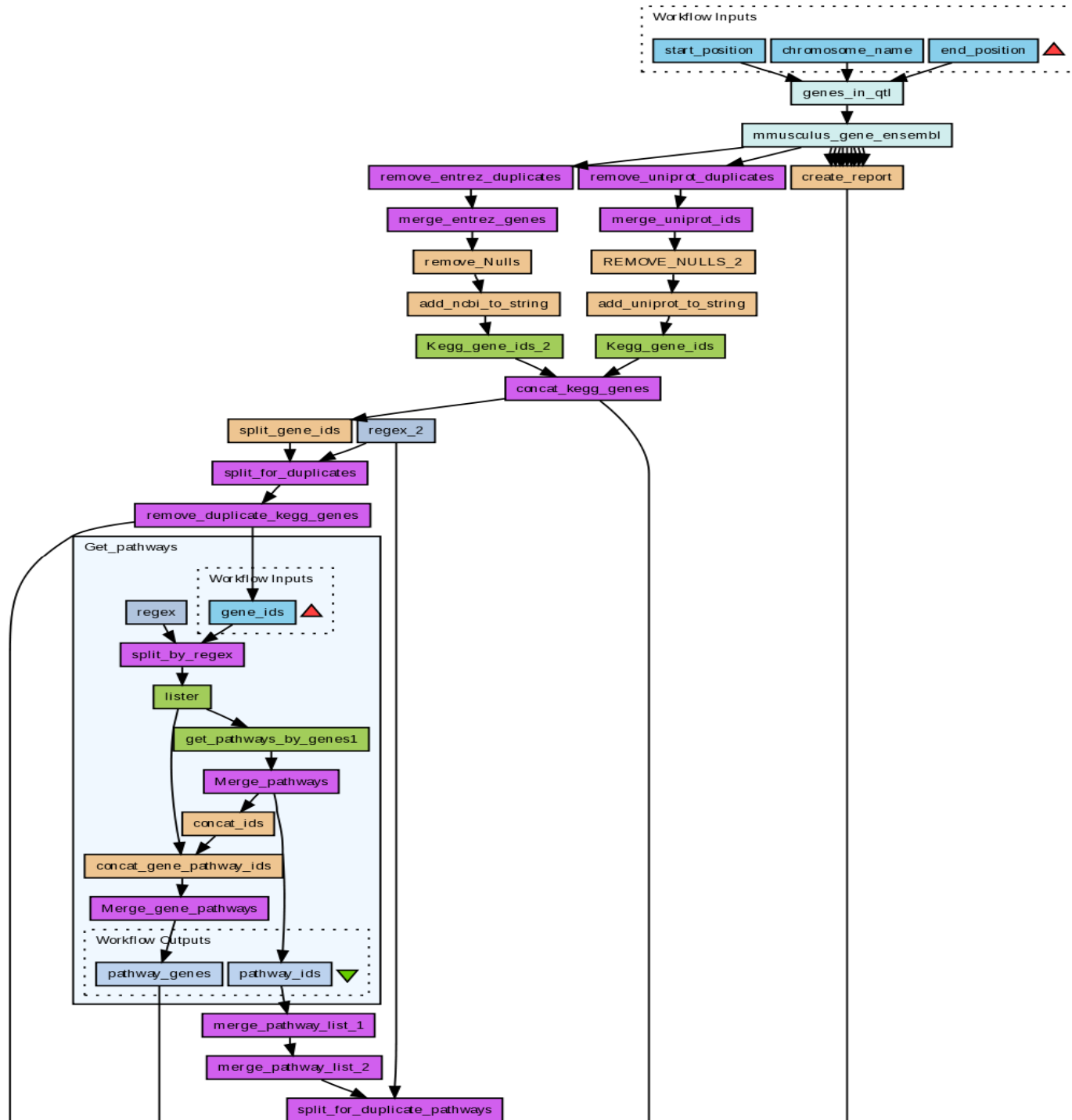
See note at **ORIGIN** .

- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality : *the manuscript has a distinguished provenance.*

ORIGIN late 18th cent.: from French, from the verb *provenir* 'come or stem from,' from Latin *provenire*, from *pro-* 'forth' + *venire* 'come.'

Scientific workflows and provenance – capturing biological data integration

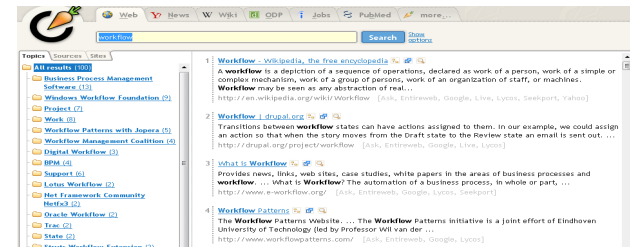
The screenshot displays the VisTrails Builder interface for a workflow named 'libSBMLdemo.vt*'. The main workspace contains a directed acyclic graph (DAG) of modules. The workflow starts with 'sbmlReadFromFile', which feeds into 'sbmlGetModelFromDocument'. This module then branches into two parallel paths: 'sbmlGetBQModelsResource' and 'sbmlGetBQModelsDescribedByResource'. Each of these paths leads to a 'getLocations' module, which in turn feeds into a 'displayWebpages' module. A 'Methods' panel on the right side of the interface is currently empty, and a 'Set Methods' panel is also visible below it. The left sidebar shows a list of available modules, including 'Basic Modules', 'Dialogs', 'HTTP', 'PythonCalc', and 'VTK'. The bottom of the image shows the Windows taskbar with several open applications and the system clock at 16:21.



Scientific workflows

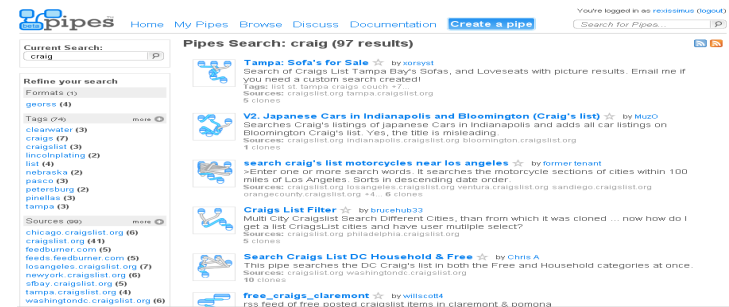
Advantage of workflows

- Easy to edit
- Reusable
- Sharable



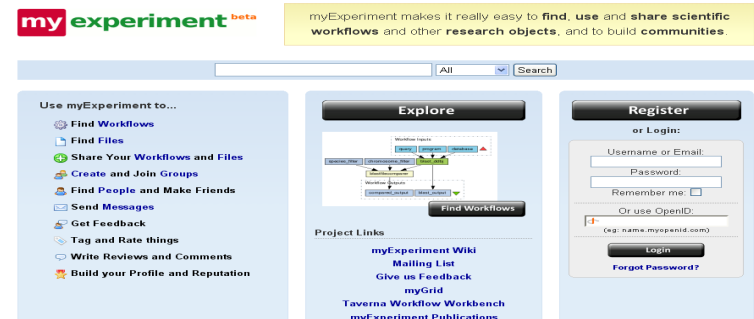
Reusing workflows

- Large collections have become available
- How to take advantage of this information?



Finding specific workflows

- Workflow Search Engines
- Workflow Query Languages

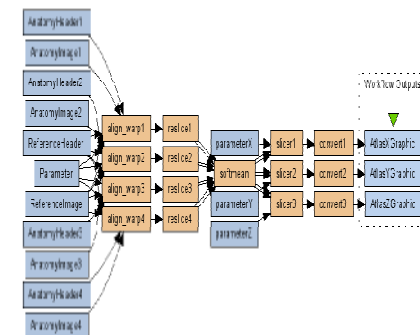
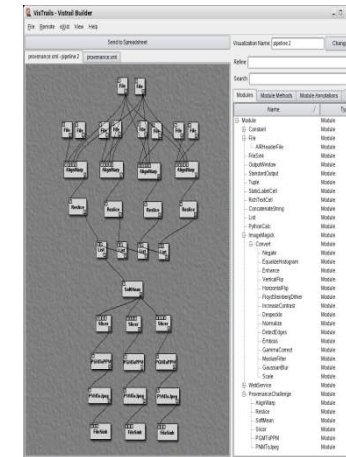


Content of this presentation:

- Two scientific application areas
 - Provenance/Scientific workflows
 - Bioinformatics
- Three different aspects
 - **Interfaces for exploration**
 - Seamless data integration
 - Effective data exploration

Issues in workflow search

- Different types of search methods
 - Keywords
 - Structured queries – workflow query language
 - Workflow similarity clustering
- Capturing the user intent
- How to rank results
 - Calculate most relevant workflow from a user query
- How to display result
 - Workflow snippets, descriptions, thumbnails




Workflow snippets – state of the art


 **Recumbent Bike Finder (Mountain US)** ☆ by jillian [View Results](#) [View Source](#)

This pipe searches every Craigslist classified in the following states (CO, WY, UT) and outputs all results with the word "recumbent" into an RSS feed, sorted by date. It allows text entry to refine the search further (example: Rans, Bacchetta, Gold Rush, etc)



Sources: [craigslist.org](#) [denver.craigslist.org](#) [sattlakecity.craigslist.org](#) [boulder.craigslist.org](#) [fortcollins.craigslist.org](#) +10... 1 clones

 **Fetch** ☆ by AndresVia [View Results](#) [View Source](#)


Fetch any URL, that has feeds.
3 clones

 **Discover_proteins_from_text** (v2)

Created: 15/11/07 @ 08:58:00 | **Last updated:** 15/11/07 @ 09:12:34

Credits:  Marco Roos  AID

License: Creative Commons Attribution-Share Alike 3.0 License

 This workflow discovers proteins from plain text. It is built around the AIDA 'Named Entity Recognize' web service by Sophia Katrenko (service based on LingPipe), from which output it filters out proteins. The Named Recognizer services uses the pre-learned genomics model, named 'MedLine', to find genomics concepts in plain text.

Rating: 0.0 / 5 (0 ratings) | **Versions:** 2 | **Reviews:** 0 | **Comments:** 0 | **Citations:** 0

Viewed internally: 64 times | **Downloaded internally:** 24 times

Tags (7):
[AIDA](#) | [BioAID](#) | [biorange_nl](#) | [protein](#) | [text_mining](#) | [text_mining_network](#) | [VL-e](#)

BioMart_hsapiens_gene_ensembl_variation_Noom_Edit_09_12_2551 (v1)

Created: 17/12/08 @ 09:08:50 | **Last updated:** 26/12/08 @ 07:51:03

Credits:  Kasikrit

License: Creative Commons Attribution-Share Alike 3.0 License

 No description

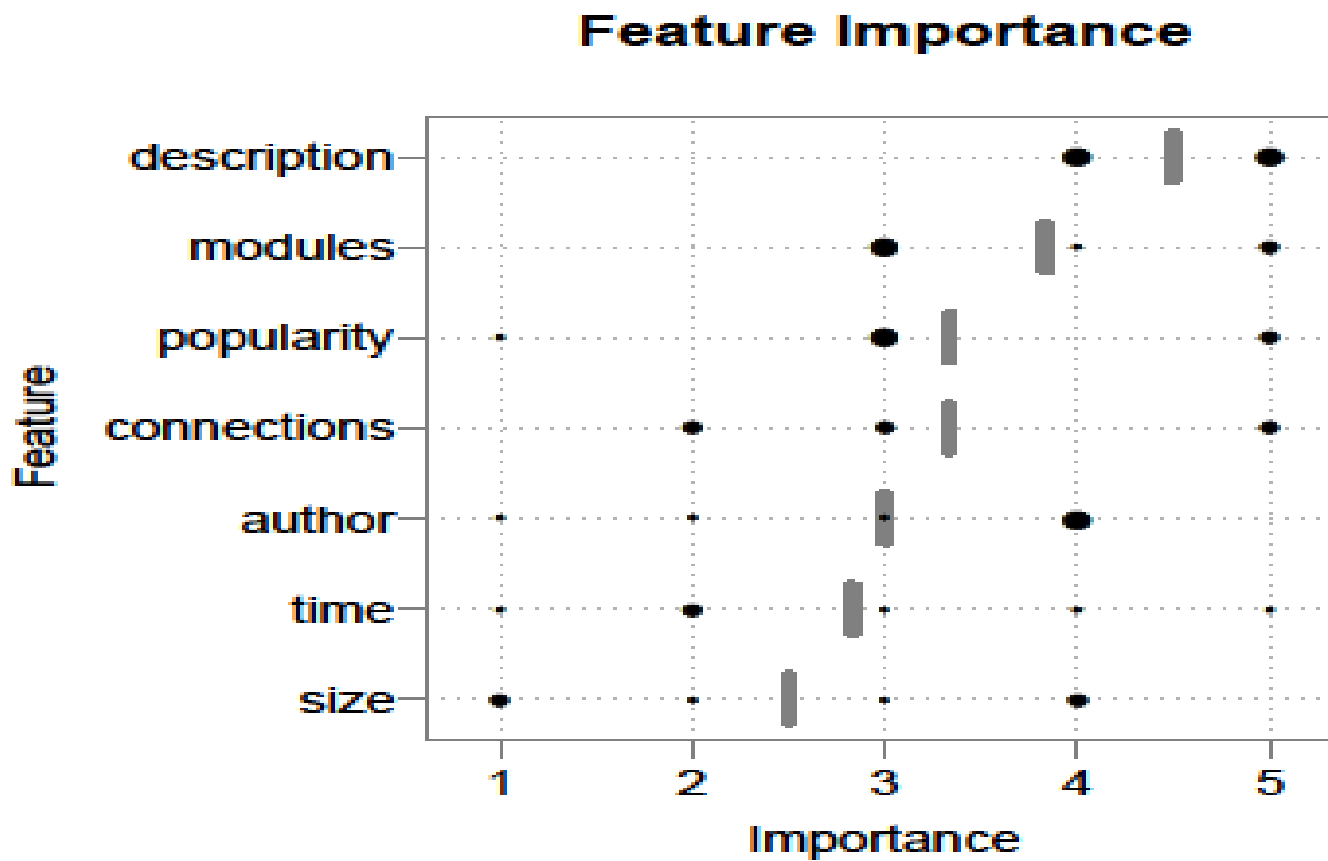
Rating: 0.0 / 5 (0 ratings) | **Versions:** 1 | **Reviews:** 0 | **Comments:** 0 | **Citations:** 0

Viewed internally: 49 times | **Downloaded internally:** 6 times

This Workflow has no tags!

- Emphasis on meta-data
- Low quality when information is insufficient or absent

Important features



Requirements for snippets

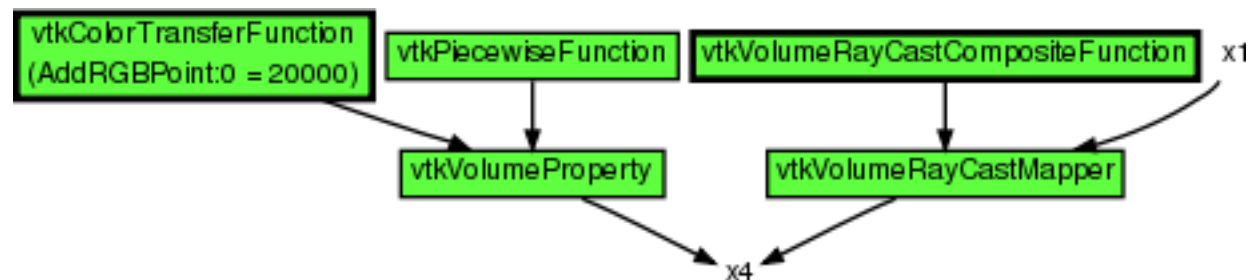
- Self-contained
 - A snippet should contain the **context of a keyword**
- Representative
 - The user should be able to **grasp the essence of the result** from its snippet.
- Distinguishable
 - The snippet should make the corresponding query result **distinguishable from other results**
- Small
 - A snippet should be **small** so that it is easy to browse several results
- **Huang, Liu and Chen (2008) Query biased snippet generation in XML search. SIGMOD 2008.**

Requirements for workflow snippets

- Self-contained
 - If a **keyword matches a module, its parameters or annotation** then that module should **be included in the snippets**.
- Representative
 - Include the modules **representing the most prominent features** of a workflow and include them in the snippet.
- Distinguishable
 - Find and display the **structural differences** among the workflows
- Small
 - We show **maximum g modules**

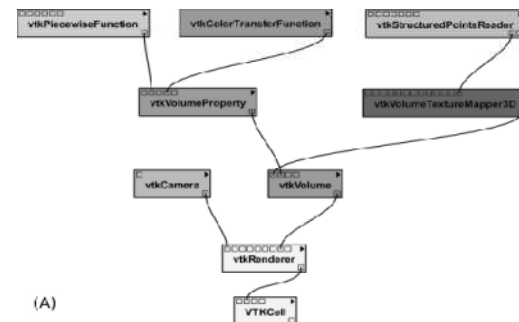
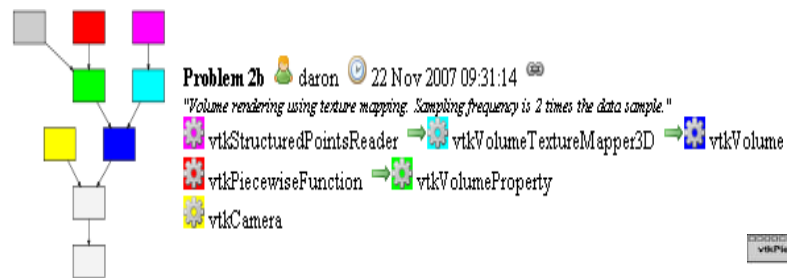
Selection strategy 1: Query neighborhood

- Identify the most important modules in the neighborhood of modules matching the keywords.
- Algorithm:
 1. Choose the modules matching the keywords
 2. Traverse the neighborhood to find closest modules with the highest IDF-values



Selection strategy 2: IDF

- Find a set of representative by choosing the modules with the highest IDF values.



Selection strategy 3: Grouping

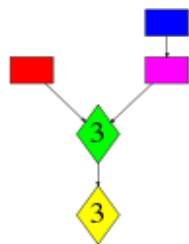
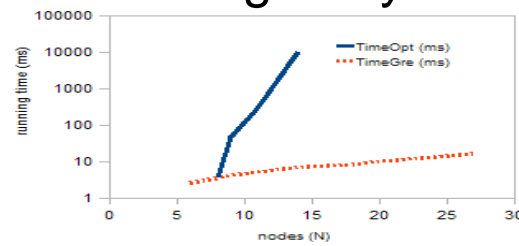
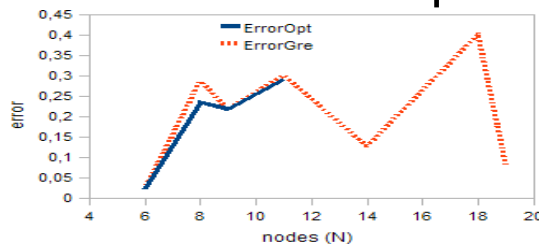
- Find co-occurring modules as they correspond to a specific functionality or semantic entity.

- Jaccard distance:

$$MScore (M_n) = \frac{\sum_{m_i, m_j \in M_n} dist(m_i, m_j)}{|M_n|}$$

$$GScore (G) = \sum_{M_i \in G} MScore (M_i)$$

- Problem: NP-complete, we use a greedy version:



Problem 2b daron 22 Nov 2007 09:31:14

"Volume rendering using texture mapping. Sampling frequency is 2 times the data sample."

vtkColorTransferFunction (AddRGBPoint:0 = 20000) vtkVolumeProperty vtkVolume

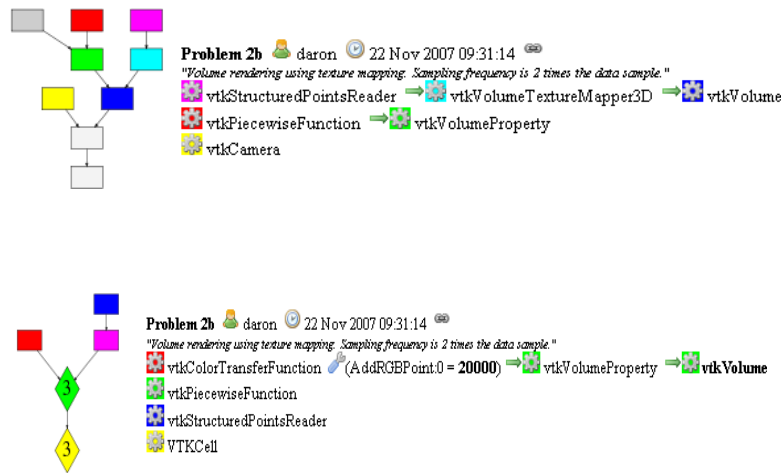
vtkPiecewiseFunction

vtkStructuredPointsReader

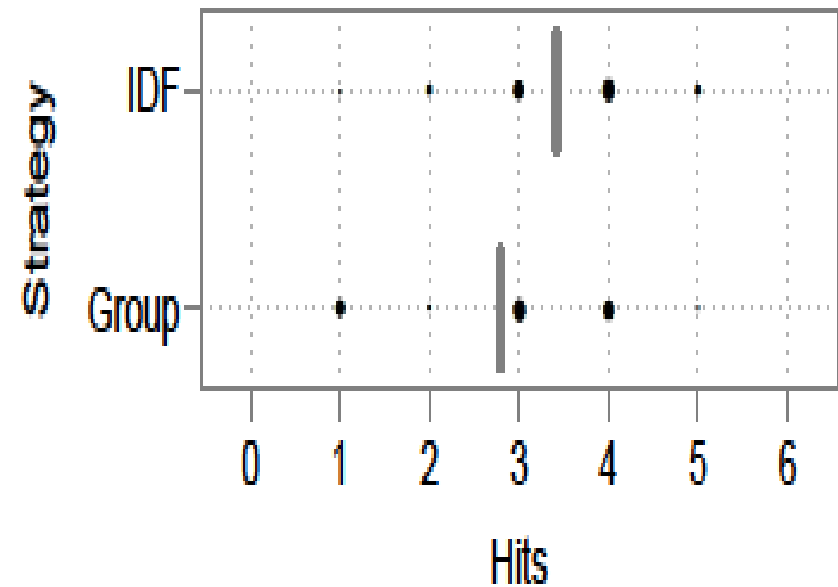
WTKCell

Evaluation: Important modules – compared to strategies

- Choose the six most important modules in the workflow.









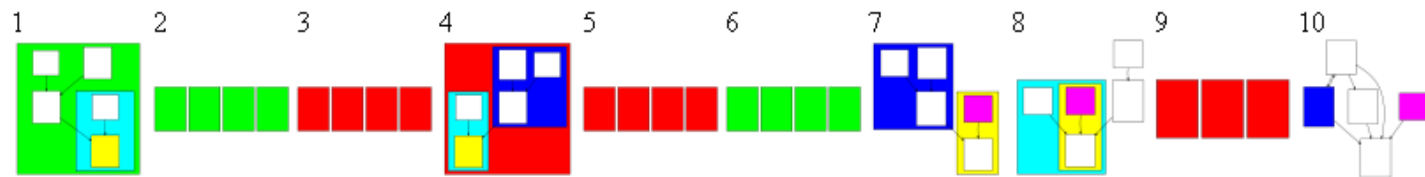
Strategy Hits



Selection strategy 4: Difference highlighting

- Display differences and similarities among workflows in a result set
- Identify the most prominent differences

12 x  vtkVolumeTextureMapper3D vtkPiecewiseFunction vtkVolumeProperty **vtkVolume** ... + 6
9 x  vtkVolumeRayCastMapper vtkVolumeRayCastCompositeFunction vtkPiecewiseFunction vtkVolumeProperty ... + 7
3 x  vtkVolumeTextureMapper3D vtkPiecewiseFunction vtkVolumeProperty **vtkVolume** ... + 2
4 x  vtkCamera VTKCell vtkRenderer
3 x  vtkCamera
3 x  vtkCamera CellLocation VTKCell vtkRenderer



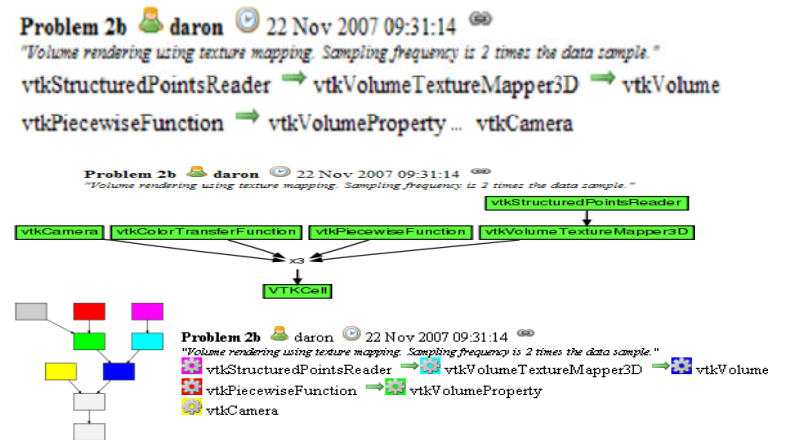
Snippet presentation

Independent of selection strategy there are several options for presentation

–Text-based

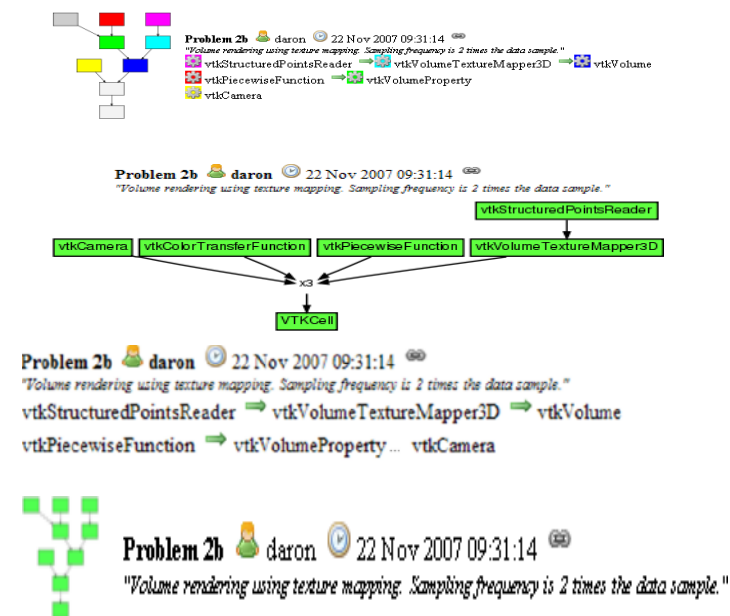
–Dynamic image

–Legend

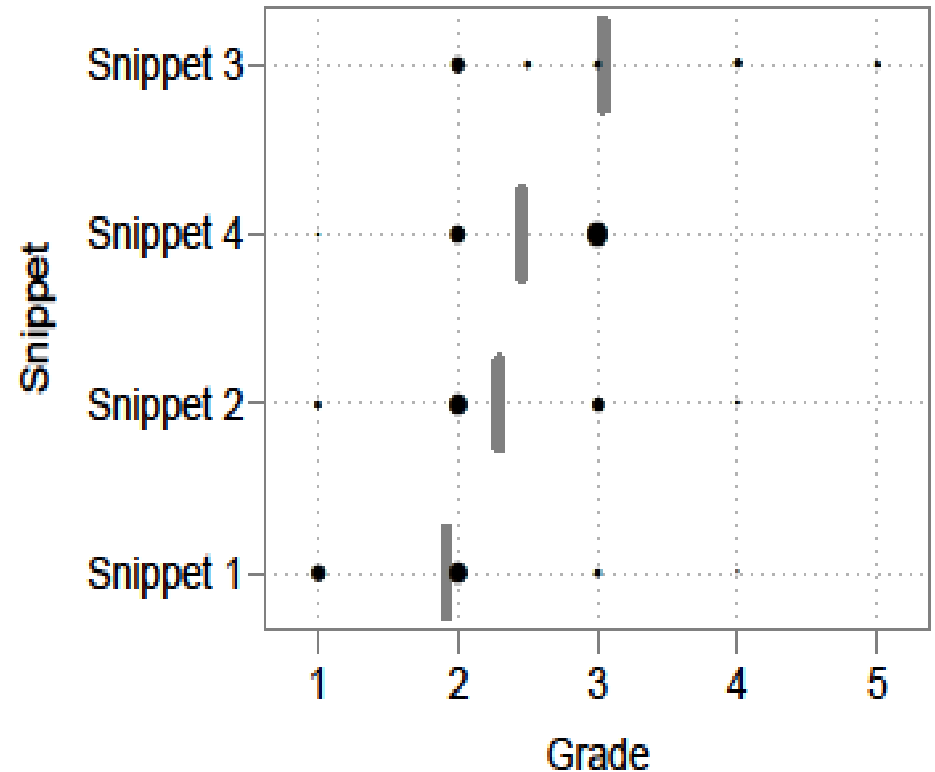


Evaluation: Important features

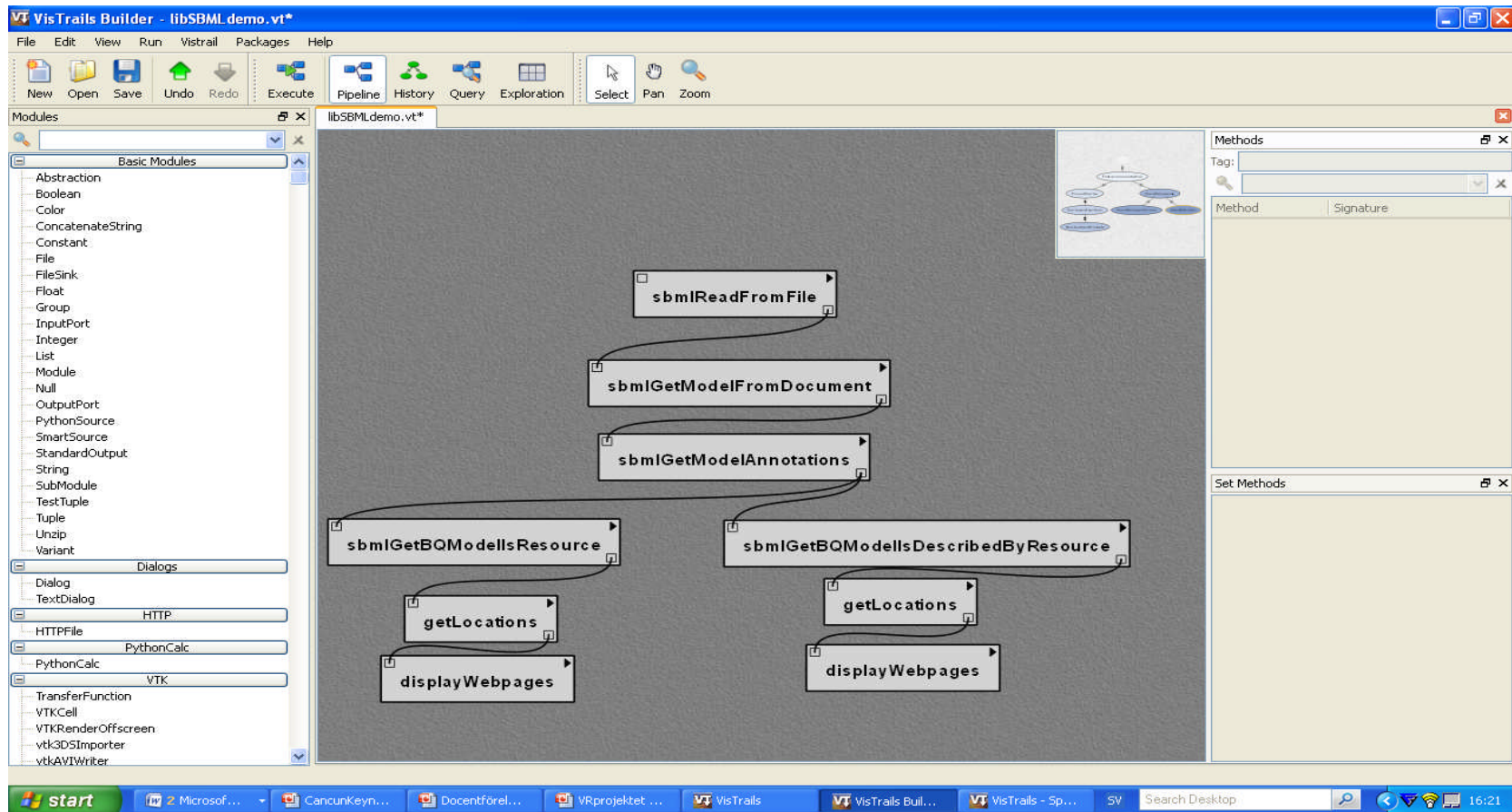
- Part3: Score workflow snippets



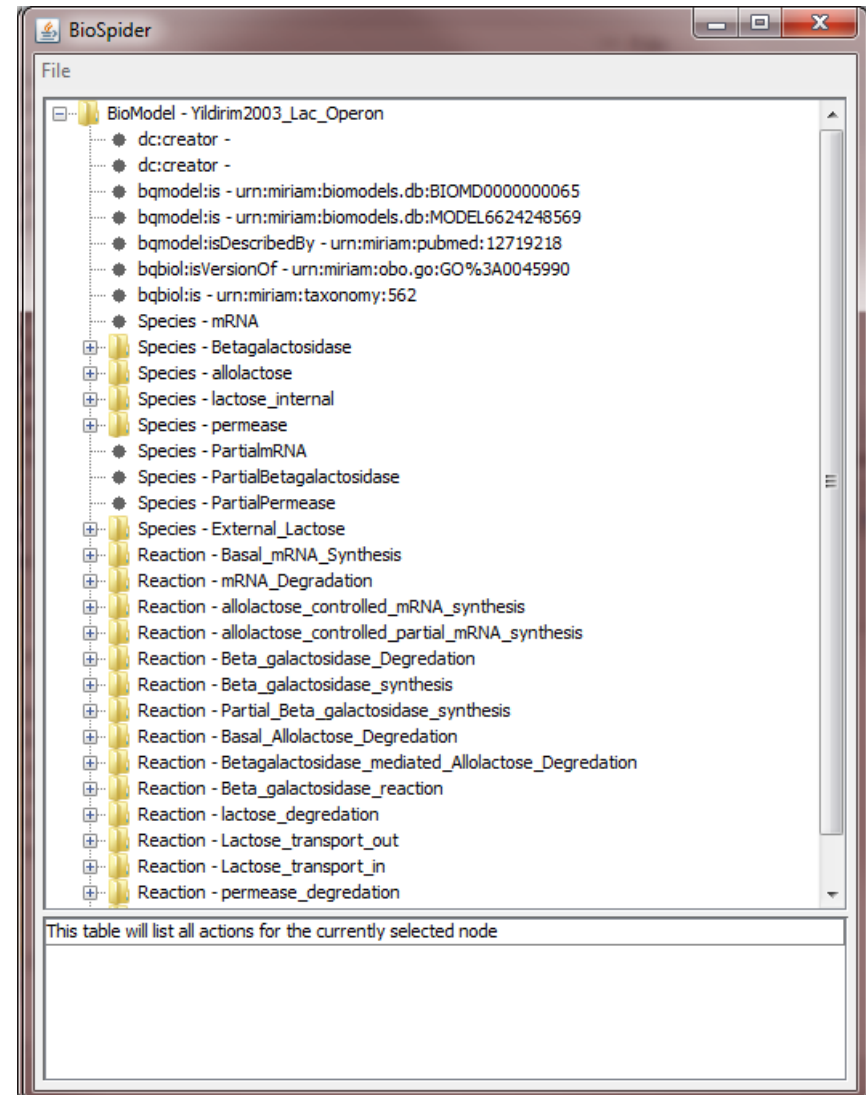
Snippet Grades



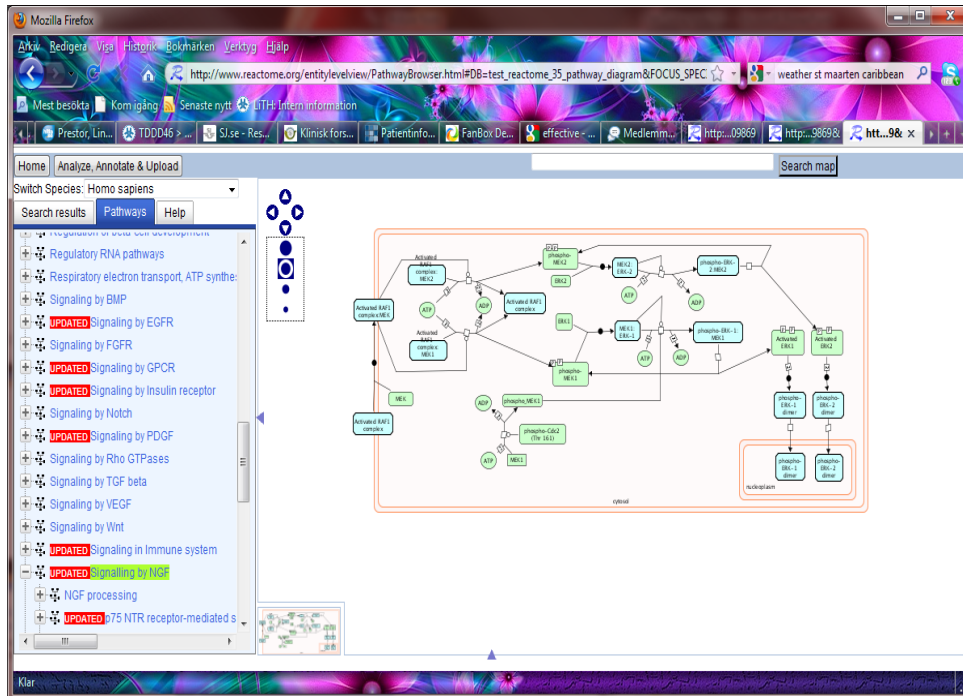
Scientific workflows for exploring Bioinformatics Web sources



BioSpider



BioSpider



BioSpider
File
BioModel - Levchenko2000_MAPK_noScaffold
● dc:creator -
● bqmodel:is - urn:miriam:biomodels.db:BIOMD000000011
● bqmodel:is - urn:miriam:biomodels.db:MODEL6615234250
● bqmodel:isDescribedBy - urn:miriam:pubmed:10823939
● **bqbiol:isHomologTo - urn:miriam:reactome:REACT_634**
● bqbiol:isVersionOf - urn:miriam:obo.go:GO%3A0000165
● bqbiol:is - urn:miriam:taxonomy:8355
● bqmodel:isDerivedFrom - urn:miriam:biomodels.db:BIOMD0000000009
+ Species - MAPK
+ Species - MAPK_MEK-PP
+ Species - MAPK-P
+ Species - MAPK phosphatase
+ Species - MAPK-P_MAPKase
+ Species - MAPK-P_MEK-PP
+ Species - MAPK-PP
+ Species - MAPK-PP_MAPKase
+ Species - MEK
+ Species - MEK-P
● Species - MEK phosphatase
+ Species - MEK-P_MEKase
+ Species - MEK-PP
+ Species - MEK-PP_MEKase
+ Species - MEK-PP_RAF-P
+ Species - MEK_RAF-P
+ Species - RAFK
+ Species - RAF-P
● Species - RAF phosphatase
+ Species - RAF-P_RAFase
+ Species - RAF_RAFK
bqbiol:isHomologTo - urn:miriam:reactome:REACT_634
Follow link
Go to webpage http://www.reactome.org/cgi-bin/eventbrowser_st_id?FROM_REACTOME=1&ST_ID...

Content of this presentation:

- Two scientific application areas
 - Provenance/Scientific workflows
 - Bioinformatics
- Three different aspects
 - Interfaces for exploration
 - **Seamless data integration**
 - Effective data exploration

Seamless data integration

- The BioSpider allows:
 - Easy integration of data from various web sources
 - Tracking of data provenance
 - Little programming knowledge of the end user
- However,
 - Each new object type (database) must be added as a new module
 - Requires large programming skills
- How can we improve?

Seamless data integration

The screenshot shows a Mozilla Firefox browser window displaying the MIRIAM Resources website. The address bar shows the URL <http://www.ebi.ac.uk/miriam/main/datatypes/MIR-00000018>. The page title is "MIRIAM Resources - Mozilla Firefox". The browser's address bar and tabs are visible at the top. The website's navigation bar includes "EMBL-EBI", "EB-eye Search", and "All Databases". The main content area displays the "Data type: Reactome" page. The page has a sidebar with navigation links and a main content area with a table of information.

EBI > Groups > Computational Neurobiology > Research > MIRIAM Resources

Data type: Reactome

General | Tags | Example Usage | Web Services

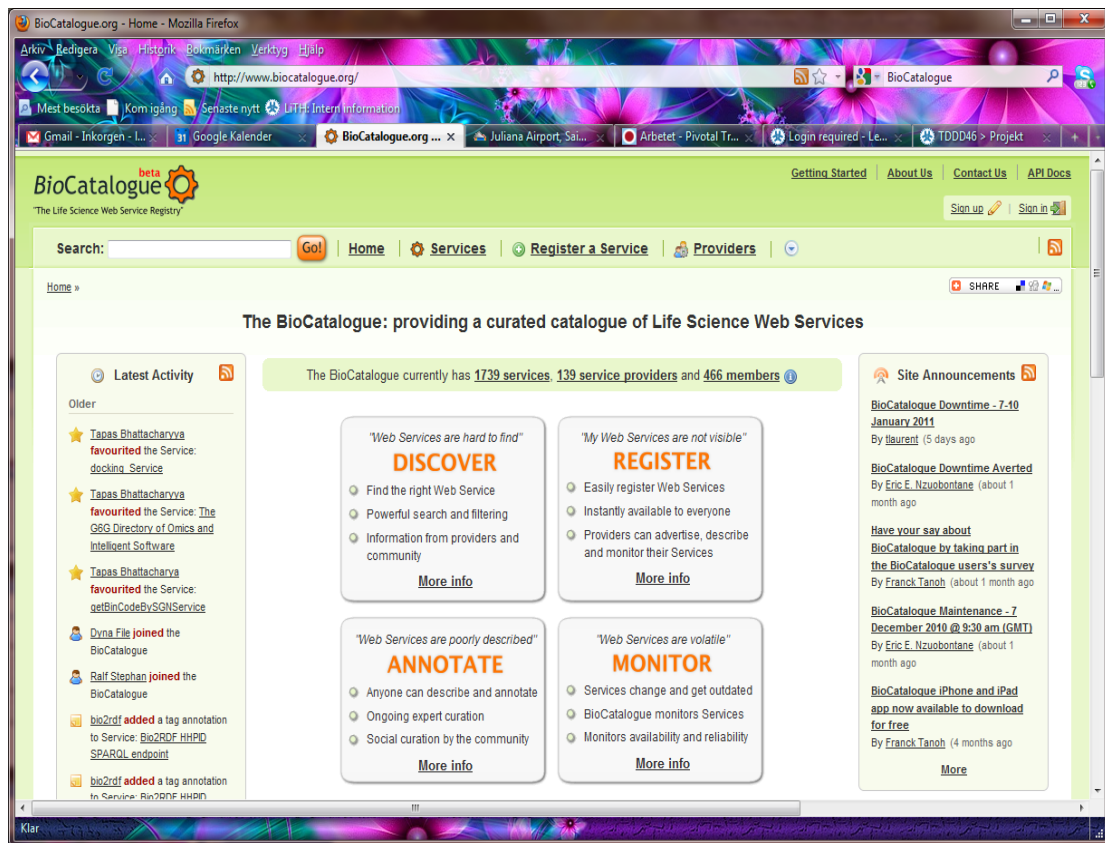
General information about the data type

Name	
Identifier	MIR:00000018
Name	Reactome
URIs	
MIRIAM URN	urn:miriam:reactome
Deprecated	http://www.reactome.org/
Information	
Definition	The Reactome project is a collaboration to develop a curated resource of core pathways and reactions in human biology.
Identifier Pattern	<code>^REACT_(id+(Lid+)?\$</code>
Physical Locations	
Resource	Access URL
MIR:00100026	http://www.reactome.org/cgi-bin/eventbrowser_st_id?FROM_REACTOME=1&ST_ID=\$id [Example: REACT_1590]
	Website
	http://www.reactome.org/
	Description
	Reactome, a curated knowledgebase of biological pathways
	Institution
	Cold Spring Harbor Laboratory and European Bioinformatics Institute, USA / United Kingdom
References	
URL(s)	http://srs.ebi.ac.uk/srsbin/cgi-bin/wqetb?2-view+MedlineFull+medline-PMID:15608231
Miscellaneous	
Date of creation	2006-08-14 19:38:06 GMT
Date of last modification	2009-04-21 15:49:13 GMT

[Go back to the list of data types](#) [Suggest modifications to this data type](#)

Seamless data integration

- Using available resources
 - MIRIAM
 - BioCatalogue



Seamless data integration

- Using available resources
 - MIRIAM
 - BioCatalogue
- Allowing users to add new methods and knowledge

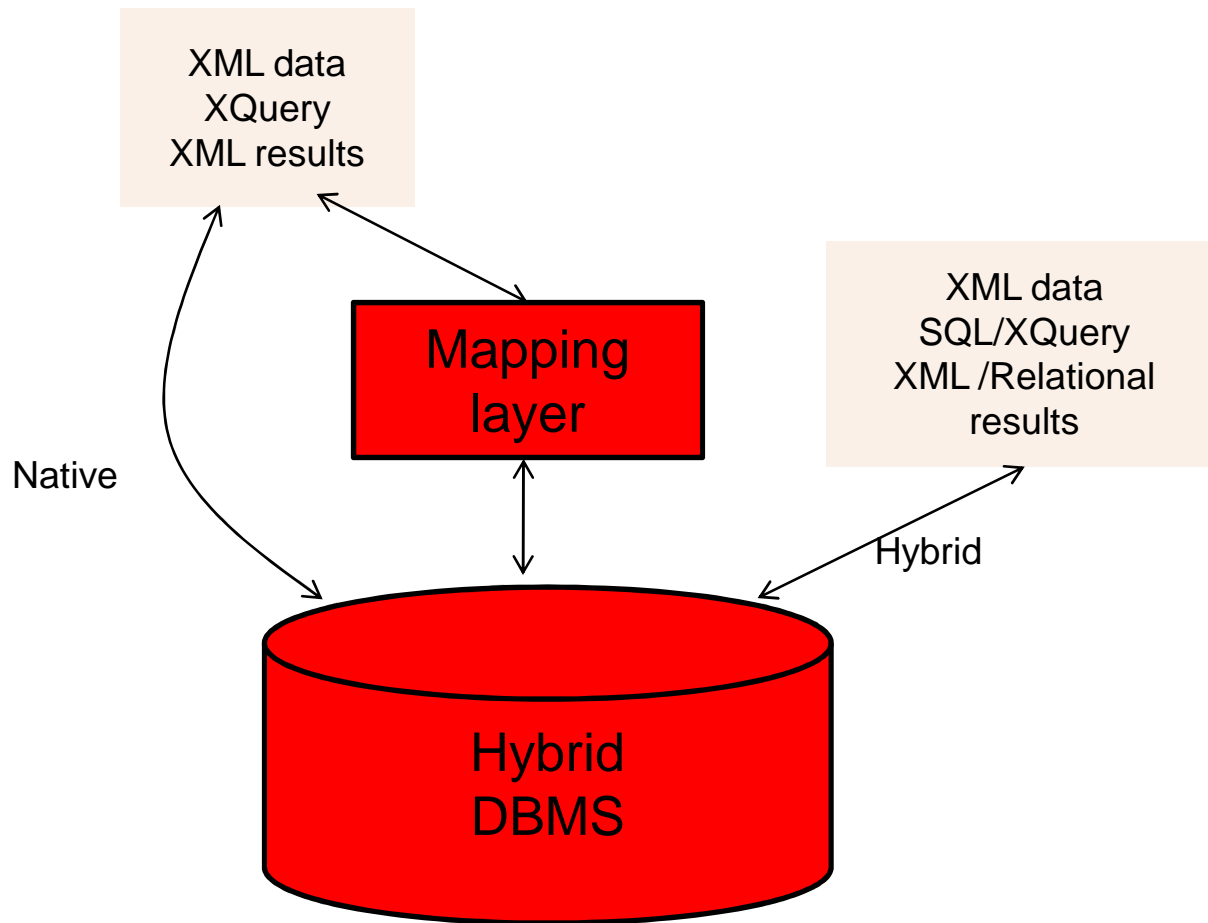
Content of this presentation:

- Two scientific application areas
 - Provenance/Scientific workflows
 - Bioinformatics
- Three different aspects
 - Interfaces for exploration
 - Seamless data integration
 - **Effective data exploration**

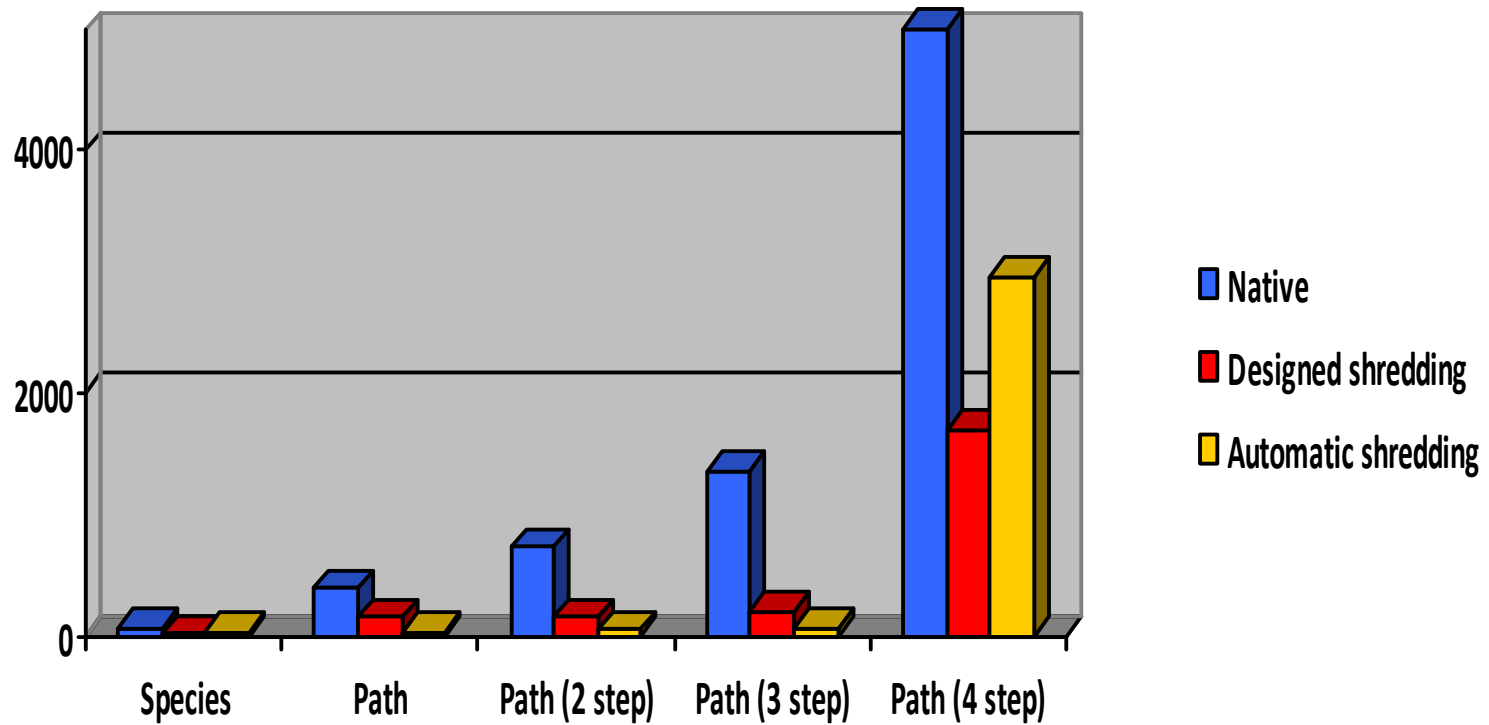
Effective data exploration

- Complex data structure – often graph structure
- Need for effective exploration methods
- Data often represented as XML or RDF

Hybrid XML Storage



Efficiency: Increasing query complexity



Tool development: HShreX

The screenshot shows the HShreX application window with the following components:

- Title Bar:** HShreX version 1.0 - August 26th, 2010 - optional_string_attribute-1.xsd
- Menu Bar:** File, Connections, View Scripts, Settings
- Schema Tree (Left Panel):**
 - movies (Folder)
 - movie (Folder)
 - title (Field)
 - year (Field)
 - goodMovie (Field)
- Navigation Tabs:** XML Schema, About Mappings, Relational Schema, Mapping Editor, Query
- Table View (Right Panel):** Shows the 'movies' schema selected. Below it is a table with the following data:

Field Name	SQL Type	SQL Type L...	isNullable	isPrimaryKey	isForeignKey	refTableName
shrex_id	SQL_INT	default value	false	true	false	
shrex_pid	SQL_INT	default value	false	false	true	movies
goodMovie	SQL_STRING	default value	true	false	false	
title	SQL_STRING	default value	false	false	false	
year	SQL_STRING	default value	false	false	false	

Starting parse of schema "optional_string_attribute-1.xsd". This can take a while for large and/or complicated schemas...
Parsing of schema "optional_string_attribute-1.xsd" completed.

Active connection: SimulatedConnection

Working with HShreX:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:shrex="http://www.cse.ogi.edu/shrex">

  <xs:element name="families">
    <xs:complexType>
      <xs:sequence maxOccurs="unbounded">
        <xs:element name="family" type="familyType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:complexType name="familyType">
    <xs:sequence>
      <xs:element name="parent" type="parentType" >
        <xs:element name="child" type="childType" >
        </xs:sequence>
      </xs:complexType>

  <xs:complexType name="parentType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="job" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="childType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="school" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
```

Families

Id	Pid
0	-

Families_family

Id	Pid
1	0

Families_family_parent

Id	Pid	Name	Job
2	1	Lena	Lektor

Families_family_child

Id	Pid	Name	School
3	1	Ludvig	Skolan

Working with HShreX:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:shrex="http://www.cse.ogi.edu/shrex">

  <xs:element name="families">
    <xs:complexType>
      <xs:sequence maxOccurs="unbounded">
        <xs:element name="family" type="familyType"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:complexType name="familyType">
    <xs:sequence>
      <xs:element name="parent" type="parentType" >
        <xs:element name="child" type="childType"
          shrex:maptoxml="true">

        </xs:sequence>
      </xs:complexType>

  <xs:complexType name="parentType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="job" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="childType">
    <xs:sequence>
      <xs:element name="name" type="xs:string"/>
      <xs:element name="school" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

Families

Id	Pid
0	-

Families_family

Id	Pid	Child
1	0	<child> <name>Ludvig</name> <school>Skolan/school</school> </child>

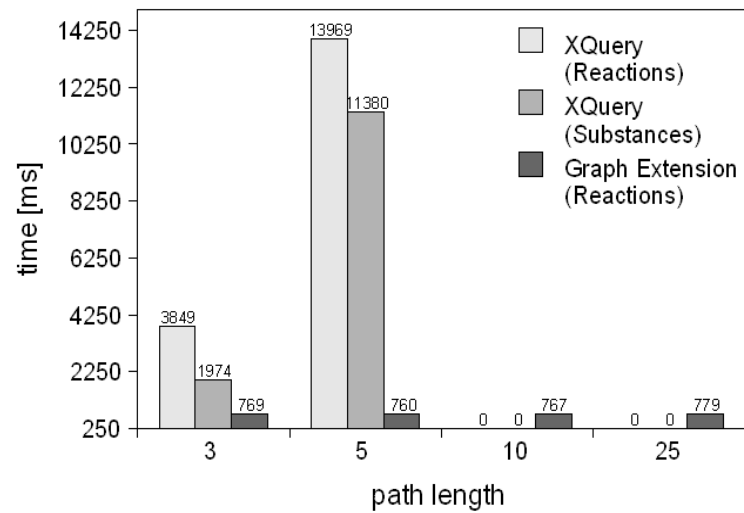
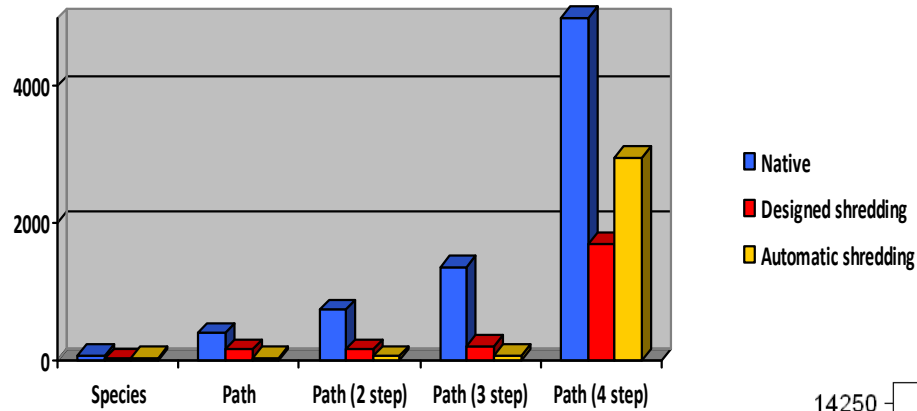
Families_family_parent

Id	Pid	Name	Job
2	1	Lena	Lektor

Guidelines for Shredding XML:

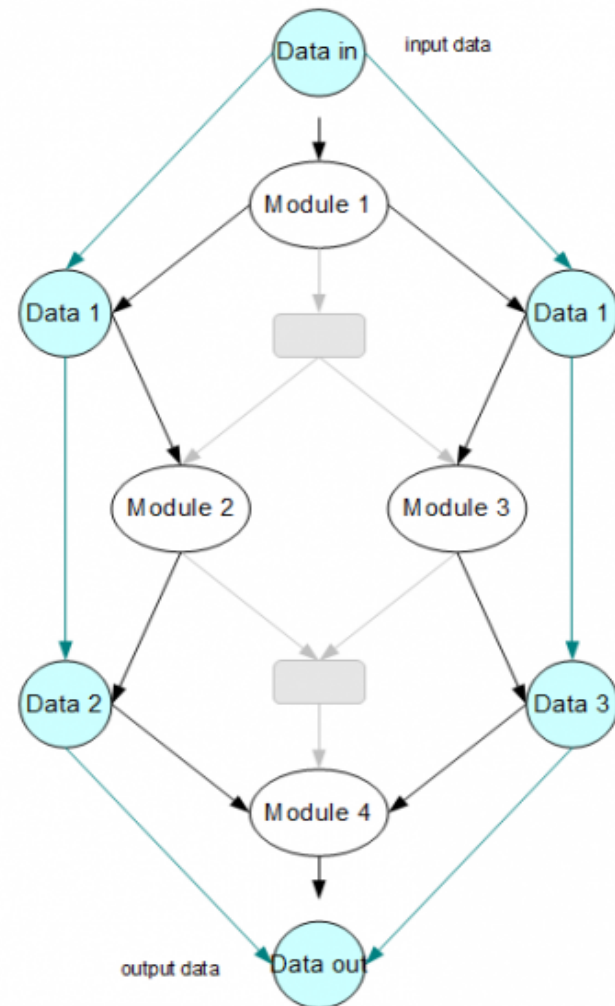
- Keep together what naturally belong together
- Do not shred parts of the XML where the schema allows large variation
- Take variations of the actual data into account
- Shred elements that are critical for performance
- Prefer the representation that is required for query results

Efficiency for graph queries



Effective querying for workflows

- Tool independent
 - capture all features of OPM
- Complex queries on
 - structure,
 - versions,
 - subworkflow
 - similarity
- Infrastructure for evaluation



Collaborators

Bioinformatics standards: Patrick Lambrix, He Tan

Workflow snippets: Tommy Ellkvist, Juliana Freire,
Lauro Didier Linz

BioSpider: Mikael Åsberg, Rickard Pettersson

HShreX and hybrid storage: Mikael Åsberg, David Hall,
Valentina Ivanova, Juliana Freire

Efficient storage for workflows: Valentina Ivanova,
Juliana Freire

Thanks!





Linköping University

expanding reality

www.liu.se