Ediflow: data-intensive interactive workflows for visual analytics

Véronique Benzaken[2], Jean-Daniel Fekete[1], Pierre-Luc Hémery[1], **Wael Khemiri**[1,2] and Ioana Manolescu[1,2]

[1]INRIA Saclay – île de France
[2]LRI, Université Paris Sud

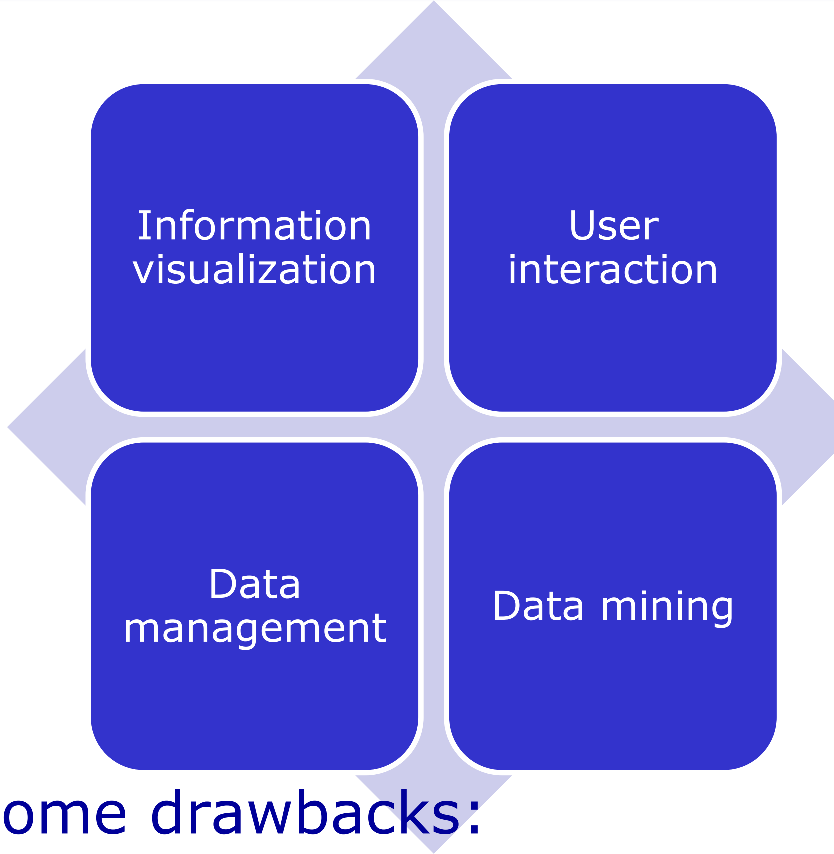
ICDE 2011

Outline

- Motivation
- Ediflow architecture
- Isolation management
- Use cases
- Robustness evaluation
- Conclusion and perspectives

Motivation – Visual analytics field

"Visual analytics is especially focused on situations where *the huge amount of data and the complexity of the problem make automatic reasoning impossible without human interaction"*



Current visual analytics tools have some drawbacks:

- Scalability issues
- No multi-user environment
- Data cannot be shared and reused

Scientific workflows vs. visual analytics

Scientific workflow systems share many characteristics with visual analytics

- Complex analysis tasks backed by persistent storage

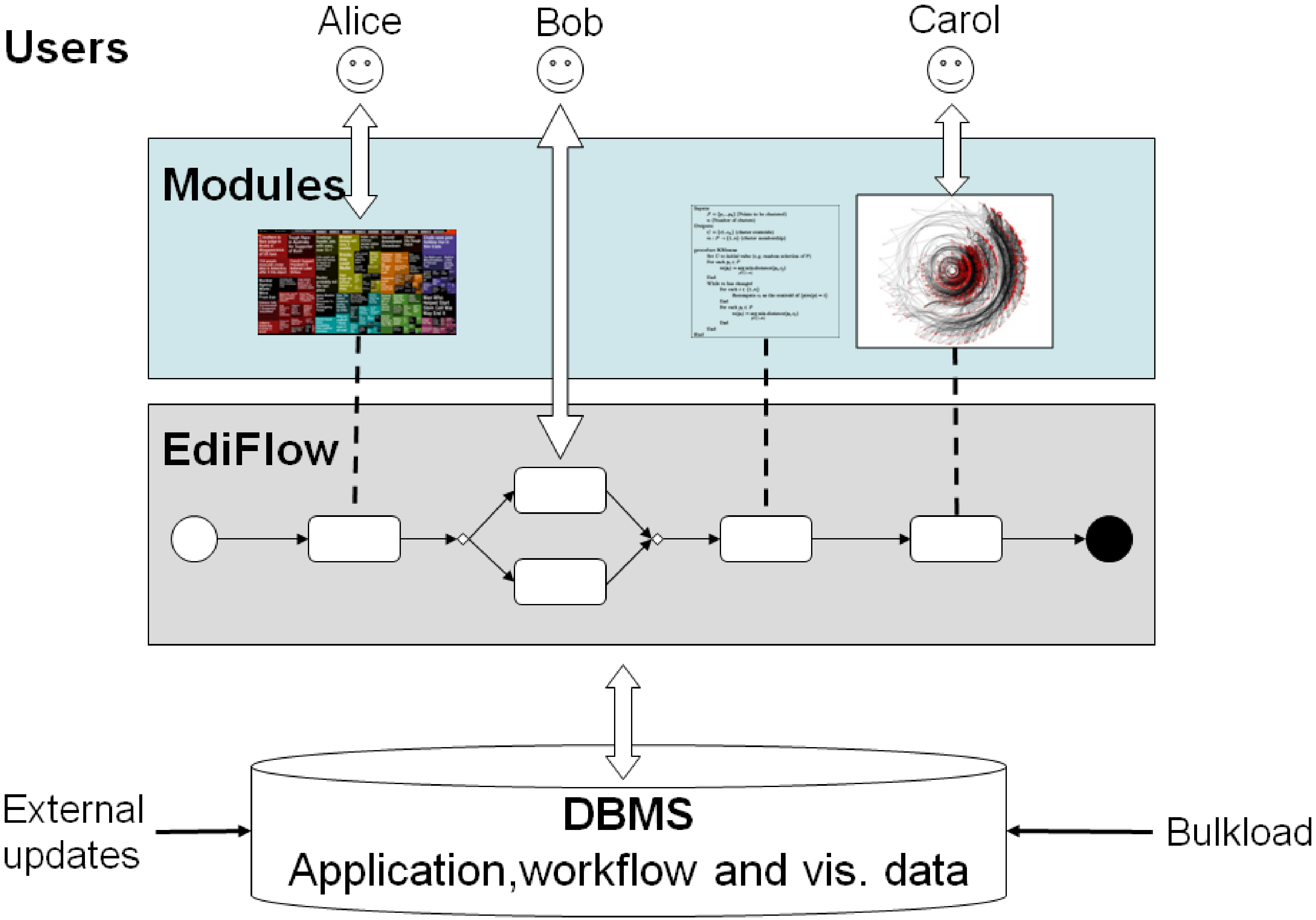
Differently from visual analytics platforms, scientific workflows:

- are designed to carry automated analytical processes to completion
- do not manage dynamic data
- offer little or no visualization

Our goals:

1) Integrating scientific workflows with visual analytics
2) Managing dynamic data

Ediflow architecture



Users

Alice

Bob

Carol

Modules

EdiFlow

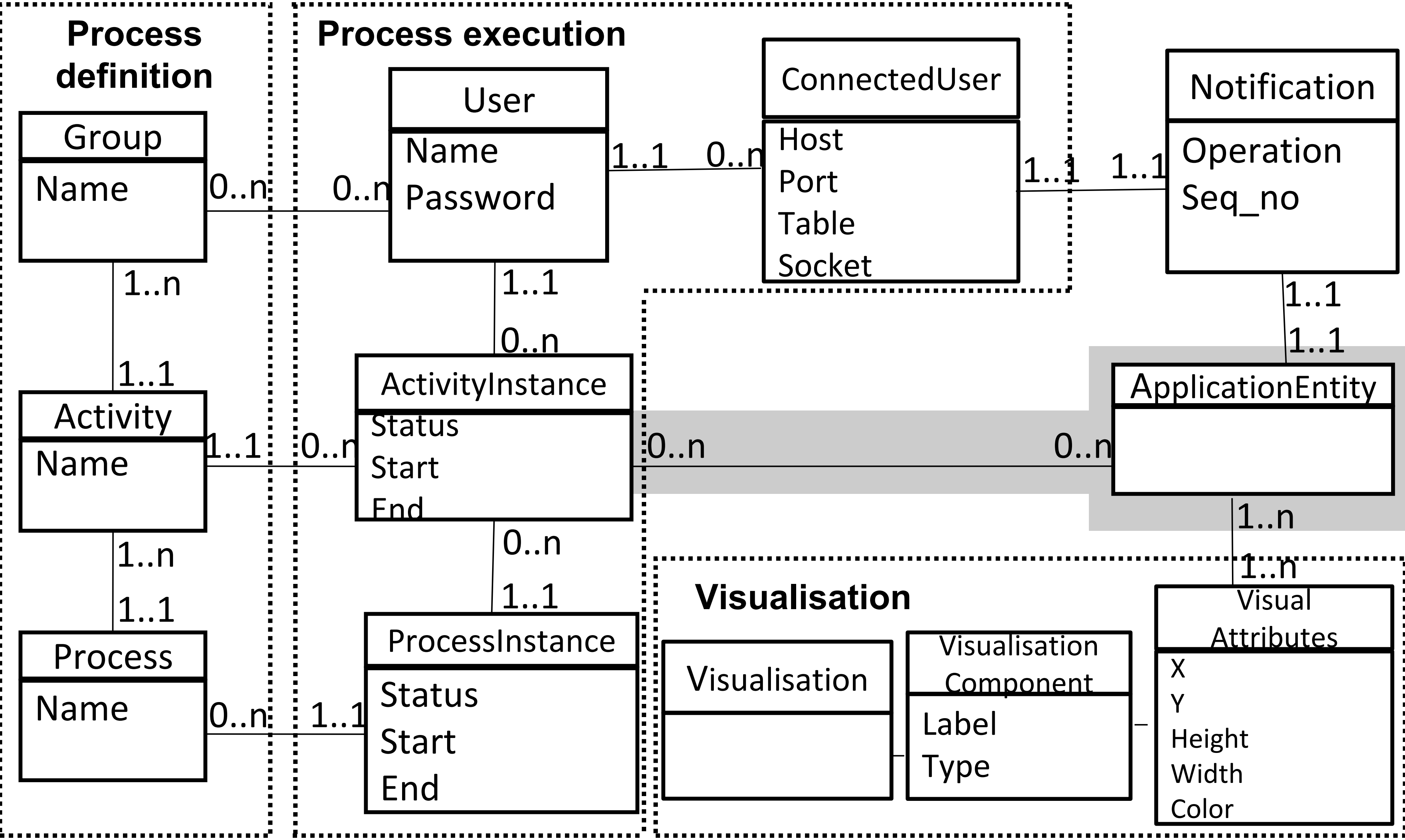
External updates

DBMS

Application,workflow and vis. data

Bulkload

Data model


Process definition
Process execution
Group
Name
0..n
0..n
User
Name
Password
1..1
0..n
ConnectedUser
Host
Port
Table
Socket
1..1
1..1
Notification
Operation
Seq_no
1..n
1..1
1..1
0..n
1..1
1..1
Activity
Name
1..1
0..n
ActivityInstance
Status
Start
End
0..n
0..n
ApplicationEntity
1..n
1..n
0..n
1..n
1..1
1..1
Process
Name
0..n
1..1
ProcessInstance
Status
Start
End
Visualisation
Visualisation
Visualisation Component
Label
Type
Visual Attributes
X
Y
Height
Width
Color

Process model

Core process model:

- Structured processes
- Workflow management coalition model
  - Sequence
  - OR split, OR join
  - AND split, AND join
  - IF-THEN
  - Procedure

Extension: reactive processes

- Reactive: propagate changes between the database and the running workflow instances through the process

Process model zooms

Procedure

- Computation unit
- Black box developed outside the DB engine (Java, C++, Matlab etc.)
- E.g. clustering algorithms, statistical analysis tools

Delta handler

- Helper procedures used to reflect the impact of data changes on process execution
- Ediflow recuperates the result of handler invocation and injects it into the process
- The implementation of handlers is opaque to the process execution framework

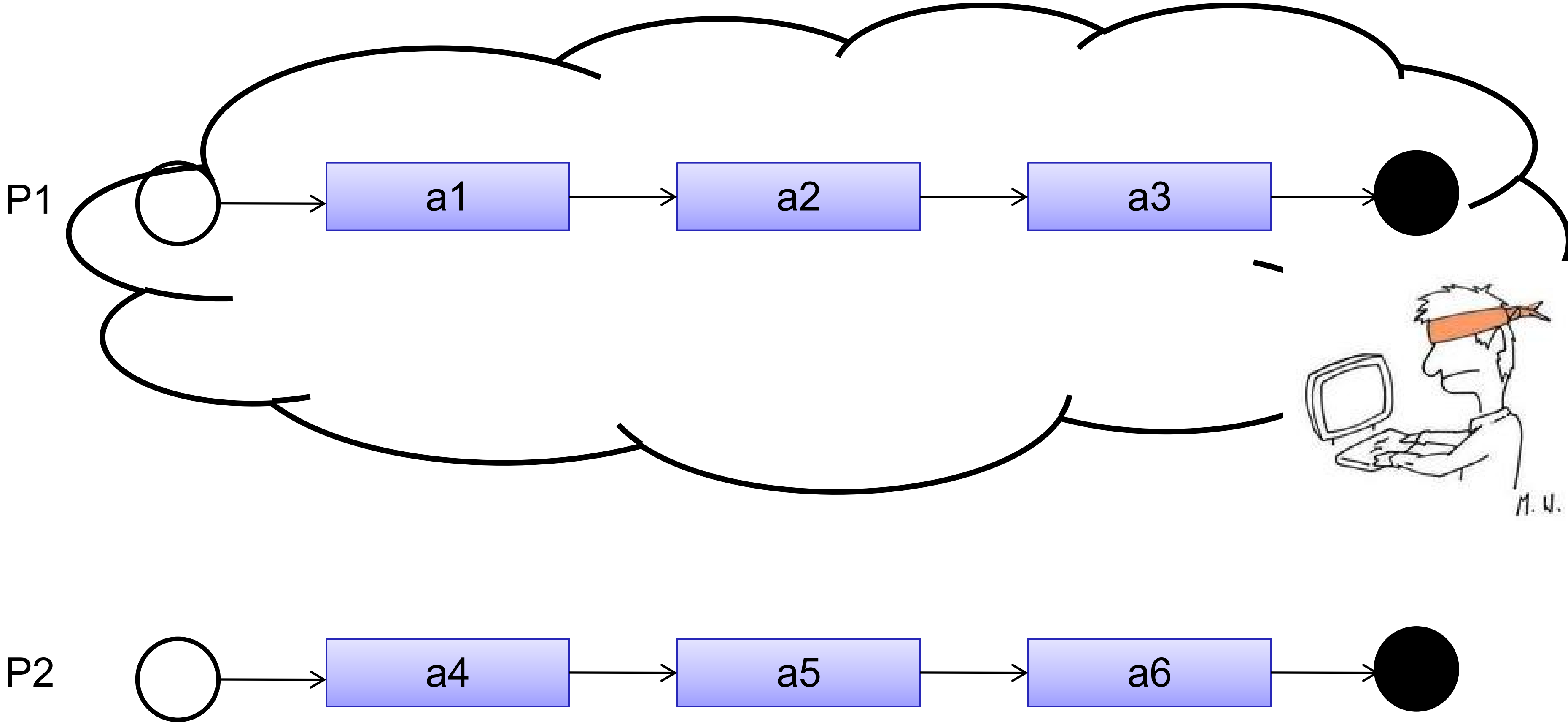
Reactive process model

Update propagation in reactive processes:

- Ignore ΔR for the execution of all processes which had started executing before $t_{\Delta R}$
- Ignore ΔR for the execution of all activities which had started executing before $t_{\Delta R}$
- ΔR are propagated to instances of all activities that are yet to be started in a running process
- Propagate the update ΔR to all the terminated instances of a given activity
- Propagate the update ΔR to all the running instances of a given activity, whether they had started before $t_{\Delta R}$ or not

Isolation management

- Process- and activity-based isolation



Isolation management

Time-based isolation

- Data visible to a given activity or process instance may depend on the starting time of that instance

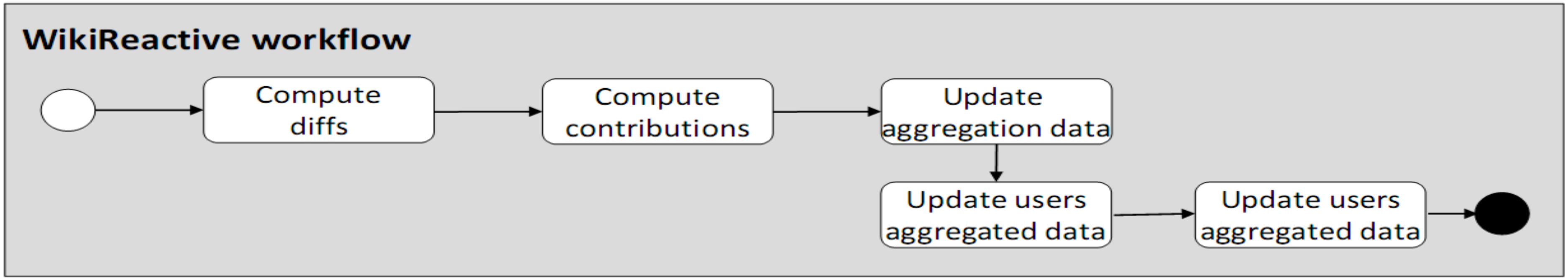
Associating to each application table R a creation timestamp

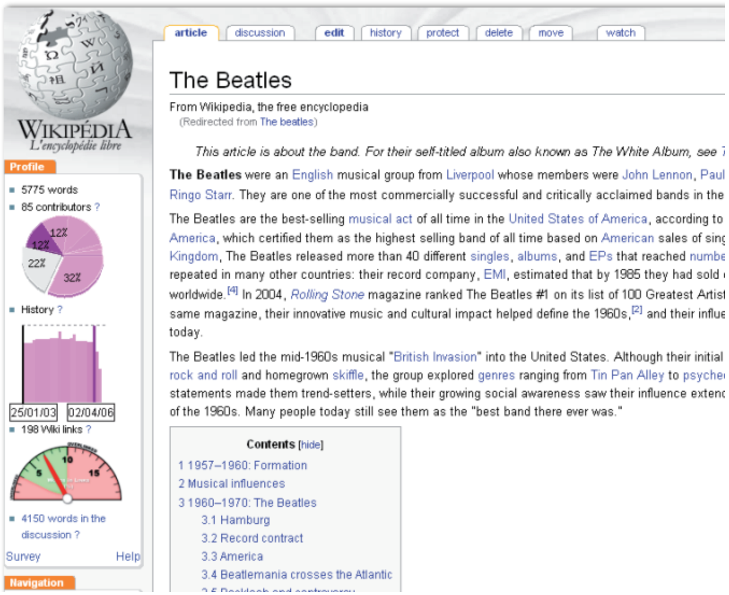
Problem with tuple deletion:

- Tuples are not actually deleted from R until the end of the process execution
- Tuples are added to a deletion table R (tid, tdel, pid, ___)
- Rewriting queries

Use case 1: WikiReactive scenario

Goal: Proposing to Wikipedia readers and contributors some measures of the history of an article


WikiReactive workflow



The Beatles

From Wikipedia, the free encyclopedia
(Redirected from The beatles)

This article is about the band. For their self-titled album also known as The White Album, see

The Beatles were an English musical group from Liverpool whose members were John Lennon, Paul
Ringo Starr. They are one of the most commercially successful and critically acclaimed bands in the

The Beatles are the best-selling musical act of all time in the United States of America, according to
America, which certified them as the highest selling band of all time based on American sales of sing
Kingdom, The Beatles released more than 40 different singles, albums, and EPs that reached numbe
repeated in many other countries: their record company, EMI, estimated that by 1985 they had sold
worldwide.[4] In 2004, *Rolling Stone* magazine ranked The Beatles #1 on its list of 100 Greatest Artist
same magazine, their innovative music and cultural impact helped define the 1960s,[2] and their influe
today.

The Beatles led the mid-1960s musical "British Invasion" into the United States. Although their initial
rock and roll and homegrown skiffle, the group explored genres ranging from Tin Pan Alley to psyche
statements made them trend-setters, while their growing social awareness saw their influence extend
of the 1960s. Many people today still see them as the "best band there ever was."

Contents [hide]

Compute the differences between successive versions of each article

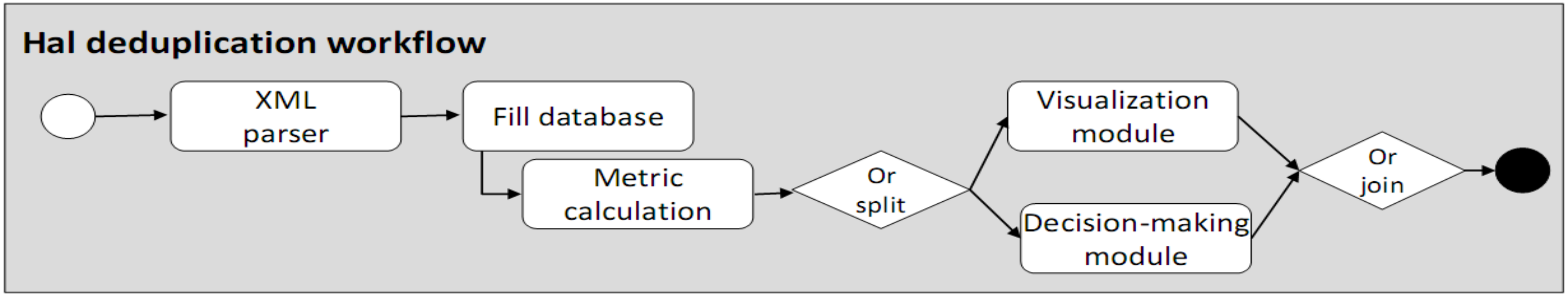
For each user, maintain the total number of characters added, deleted and moved

Compute the contribution table storing the identifier of the user who entered it

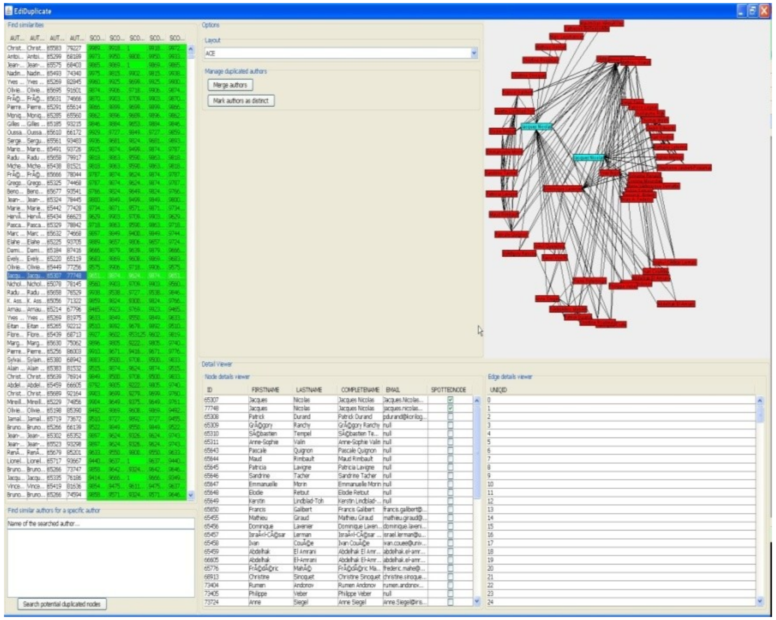
Compute the number of distinct contributors
Maintain the total number of characters

Use case 2: publication cleaning scenario

Goal: Detecting and helping remove duplicated author entries in a large database of publications.



Hal deduplication workflow



EdiDuplicate
Find similarities
Options
Manage duplicated authors
Merge authors
Mark authors as distinct
Detail Viewer
Find similar authors for a specific author
Name of the searched author...
Search potential duplicated nodes

Layout: ACE

AUT...	AUT...	AUT...	AUT...	SCO...	SCO...	SCO...	SCO...	SCO...
Christ...	Christ...	65583	79227	9989...	9918...	1	9918...	9972...
Antoi...	Antoi...	65299	68189	9973...	9950...	9800...	9950...	9933...
Jean-...	Jean-...	65575	68403	9865...	9869...	1	9869...	9865...
Nadin...	Nadin...	65493	74340	9975...	9815...	9902...	9815...	9938...
Yves ...	Yves ...	65269	82945	9960...	9925...	9699...	9925...	9900...
Olivie...	Olivie...	65695	91601	9874...	9906...	9718...	9906...	9874...
FrÃ©...	FrÃ©...	65631	74666	9870...	9903...	9708...	9903...	9870...
Pierre...	Pierre...	65291	65614	9866...	9899...	9699...	9899...	9866...
Moniq...	Moniq...	65285	65560	9862...	9896...	9689...	9896...	9862...
Gilles ...	Gilles ...	65185	93215	9846...	9884...	9653...	9884...	9846...
Oussa...	Oussa...	65610	66172	9929...	9727...	9849...	9727...	9859...
Serge...	Sergu...	65561	93483	9936...	9681...	9824...	9681...	9893...
Mario...	Mario...	65491	93726	9915...	9874...	9498...	9874...	9787...
Radu ...	Radu ...	65658	79917	9818...	9863...	9590...	9863...	9818...
Miche...	Miche...	65438	81521	9818...	9863...	9590...	9863...	9818...
FrÃ©...	FrÃ©...	65666	78044	9787...	9874...	9624...	9874...	9787...
Grego...	Grego...	65325	74468	9787...	9874...	9624...	9874...	9787...
Beno...	Beno...	65677	93541	9766...	9824...	9649...	9824...	9766...
Jean-...	Jean-...	65324	78445	9800...	9849...	9498...	9849...	9800...
Marie...	Marie...	65442	77428	9734...	9871...	9571...	9871...	9734...
HervÃ...	HervÃ...	65434	66623	9629...	9903...	9709...	9903...	9629...
Pasca...	Pasca...	65329	78842	9718...	9863...	9590...	9863...	9718...
Marc ...	Marc ...	65632	74668	9897...	9849...	9400...	9849...	9744...
Elahe ...	Elahe ...	65225	93705	9889...	9657...	9806...	9657...	9724...
Dami...	Dami...	65184	87416	9666...	9879...	9639...	9879...	9666...
Evely...	Evely...	65220	65119	9683...	9869...	9608...	9869...	9683...
Olivie...	Olivie...	65449	77256	9575...	9906...	9718...	9906...	9575...
Jacqu...	Jacqu...	65307	77748	9651...	9874...	9624...	9874...	9651...
Nichol...	Nichol...	65078	78145	9560...	9903...	9708...	9903...	9560...
Radu ...	Radu ...	65658	76529	9938...	9538...	9727...	9538...	9846...
K. Ass...	K. Ass...	65056	71322	9859...	9824...	9300...	9824...	9766...
Arnau...	Arnau...	65214	67796	9465...	9923...	9768...	9923...	9465...
Yves ...	Yves ...	65269	81975	9633...	9849...	9550...	9849...	9633...
Eitan ...	Eitan ...	65265	92212	9510...	9892...	9678...	9892...	9510...
Flore...	Flore...	65439	68713	9927...	9602...	953125	9602...	9819...
Marg...	Marg...	65630	75062	9896...	9805...	9222...	9805...	9740...
Pierre...	Pierre...	65256	86003	9910...	9671...	9416...	9671...	9776...
Sylvai...	Sylain...	65380	68942	9883...	9500...	9708...	9500...	9833...
Alain ...	Alain ...	65383	81532	9515...	9874...	9624...	9874...	9515...
Christ...	Christ...	65639	76914	9849...	9500...	9708...	9500...	9833...
Abdel...	Abdel...	65459	66605	9792...	9805...	9222...	9805...	9740...
Christ...	Christ...	65689	92164	9903...	9699...	9279...	9699...	9760...
Mireil...	Mireil...	65229	74856	9904...	9649...	9375...	9649...	9751...
Olivie...	Olivie...	65198	85390	9492...	9869...	9608...	9869...	9492...
Jamal...	Jamal...	65719	73672	9510...	9727...	9892...	9727...	9455...
Bruno...	Bruno...	65266	66139	9522...	9849...	9550...	9849...	9522...
Jean-...	Jean-...	65302	65352	9897...	9624...	9326...	9624...	9743...
Jean-...	Jean-...	65523	93298	9897...	9624...	9326...	9624...	9743...
RenÃ...	RenÃ...	65679	85201	9633...	9550...	9800...	9550...	9633...
Lionel...	Lionel...	65717	93667	9440...	9637...	1	9637...	9440...
Bruno...	Bruno...	65266	73747	9858...	9642...	9324...	9642...	9646...
Jacqu...	Jacqu...	65335	76186	9414...	9666...	1	9666...	9349...
Vince...	Vince...	65419	81636	9854...	9475...	9611...	9475...	9637...
Bruno...	Bruno...	65266	74594	9858...	9571...	9324...	9571...	9646...

Node details viewer

ID	FIRSTNAME	LASTNAME	COMPLETENAME	EMAIL	SPOTTEDNODE
65307	Jacques	Nicolas	Jacques Nicolas	Jacques.Nicolas...	☑
77748	Jacques	Nicolas	Jacques Nicolas	jacques.nicolas...	☑
65308	Patrick	Durand	Patrick Durand	pdurand@korilog...	□
65309	GrÃ©gory	Randy	GrÃ©gory Randy	null	□
65310	SÃ©bastien	Tempel	SÃ©bastien Te...	null	□
65311	Anne-Sophie	Valin	Anne-Sophie Valin	null	□
65643	Pascale	Quignon	Pascale Quignon	null	□
65644	Maud	Rimbault	Maud Rimbault	null	□
65645	Patricia	Lavigne	Patricia Lavigne	null	□
65646	Sandrine	Tacher	Sandrine Tacher	null	□
65647	Emmanuelle	Morin	Emmanuelle Morin	null	□
65648	Elodie	Retout	Elodie Retout	null	□
65649	Kerstin	Lindblad-Toh	Kerstin Lindblad-...	null	□
65650	Francis	Galibert	Francis Galibert	francis.galibert@...	□
65455	Mathieu	Giraud	Mathieu Giraud	mathieu.giraud@...	□
65456	Dominique	Lavenier	Dominique Laven...	dominique.laveni...	□
65457	IsraÃ«l-CÃ©sar	Lerman	IsraÃ«l-CÃ©sar ...	israel.lerman@u...	□
65458	Ivan	CouÃ©e	Ivan CouÃ©e	ivan.couee@univ...	□
65459	Abdelhak	El Amrani	Abdelhak El Amr...	abdelhak.el-amr...	□
66605	Abdelhak	El-Amrani	Abdelhak El-Amr...	abdelhak.el-amr...	□
65776	FrÃ©dÃ©ric	MahÃ©	FrÃ©dÃ©ric Ma...	frederic.mahe@...	□
68913	Christine	Sinoquet	Christine Sinoquet	christine.sinoque...	□
73404	Rumen	Andonov	Rumen Andonov	rumen.andonov...	□
73405	Philippe	Veber	Philippe Veber	null	□
73724	Anne	Siegel	Anne Siegel	Anne.Siegel@iris...	□

Edge details viewer

UNIQID
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

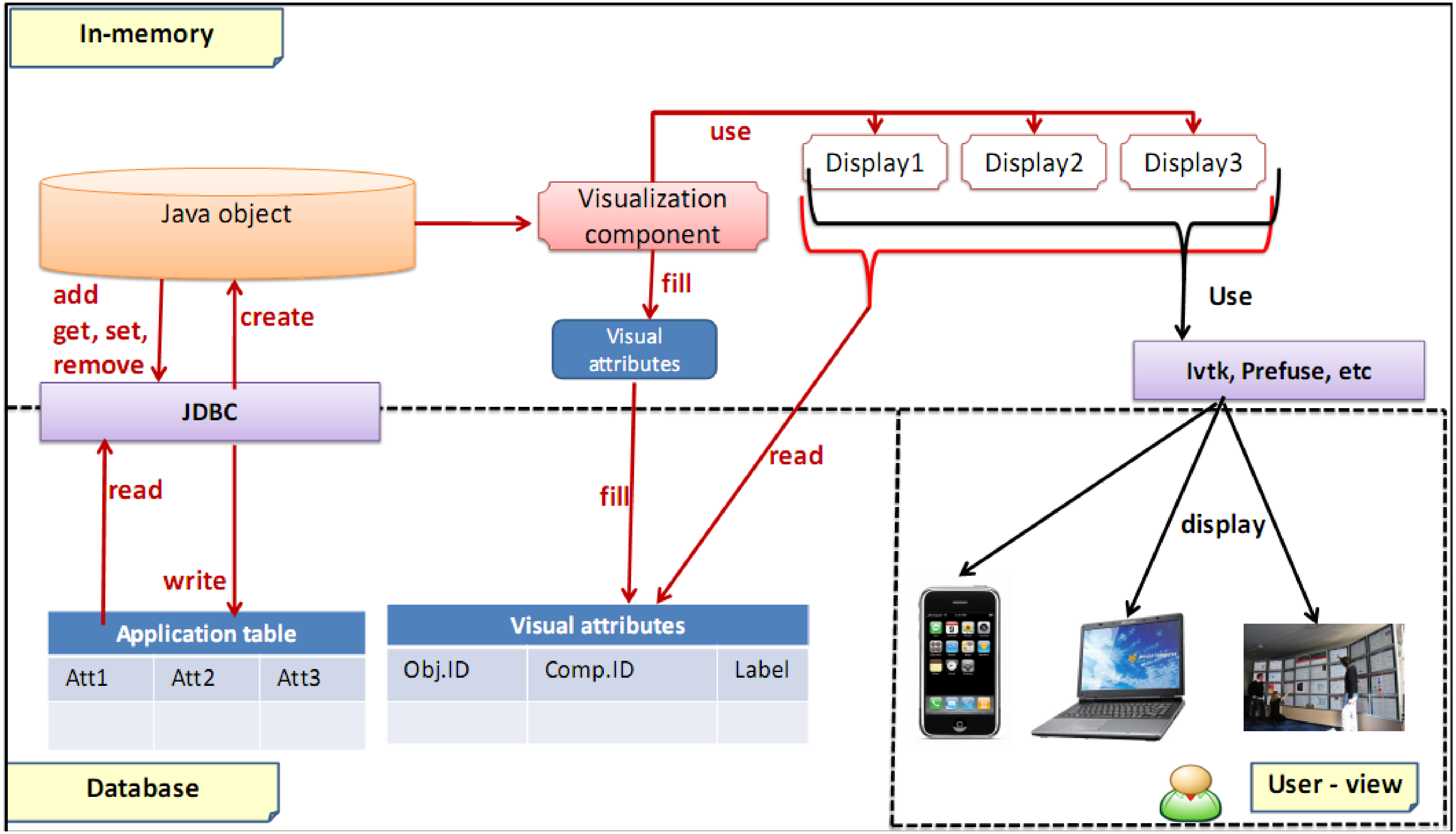
Compute the similarities between the inserted author and all the other already in the table

Show the results of similarities and co-publications graph of an author

Allow the user to decide whether two authors are identical or not

Visualization views management

- Ediflow can maintain several visualization views for one visualization



Visualization views management

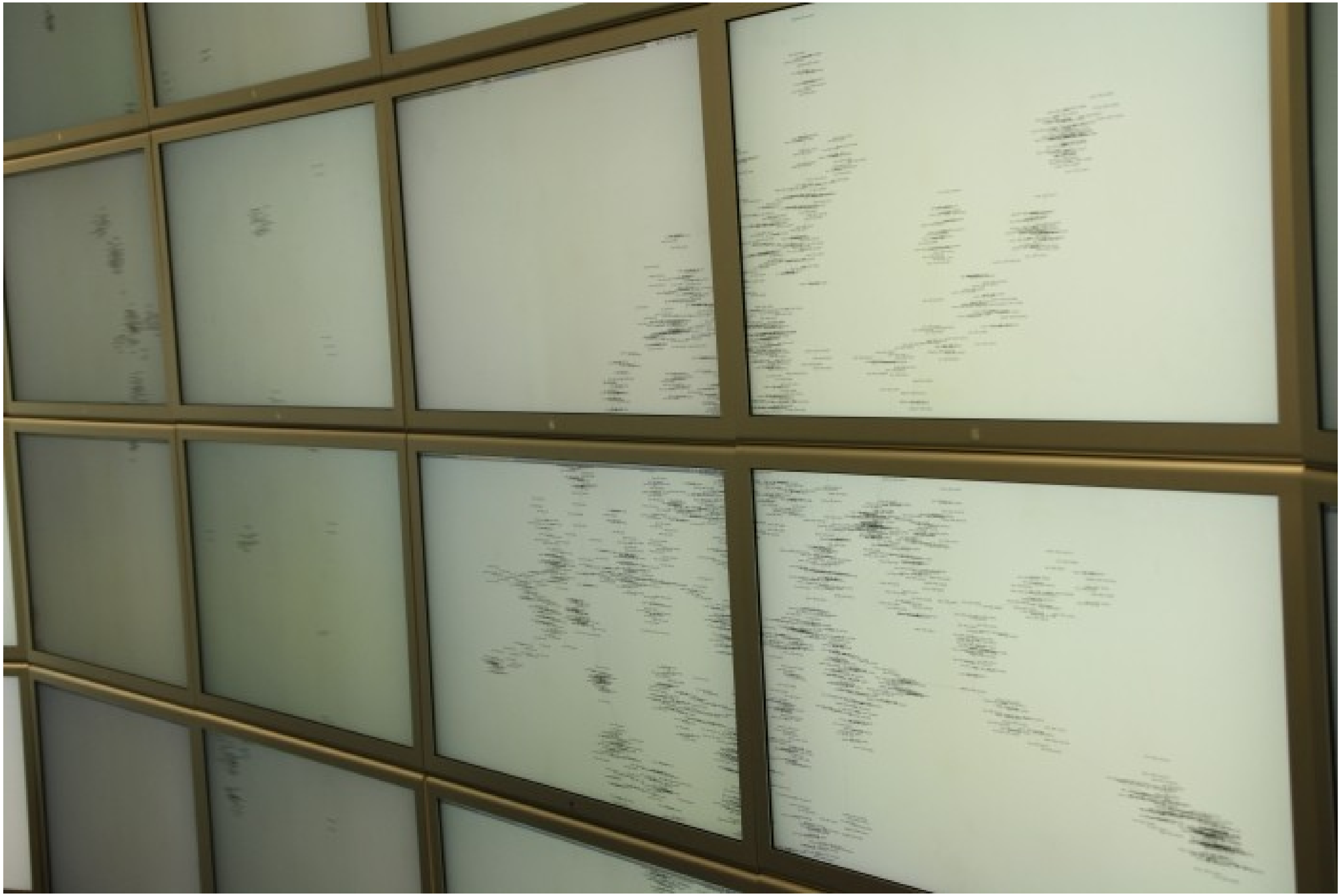
Benefits of this architecture:

- It allows sharing visual attributes by several views
- The computation of visual attributes is done only once
- In line with visual analytics recommended software architecture

Example of co-publications graph in the WILD:

- A cluster of 16 machines to display the graph over 32 screens
- Each machine controls two screens
- Each machine runs an Ediflow instance

Visualization views management



Ediflow tool implementation

- Implemented in Java
- On top of Oracle 11g DBMS
- Procedures: Java modules in OSGi Service Platform
- A procedure instance is a concrete class implementing the Ediflow Process interface
- Ediflow process requires four methods:
  - initialize()
  - run(ProcessEnv env)
  - update(ProcessEnv env)
  - string getName()

Robustness evaluation

Goal: Study how the Ediflow event processing chain scales when confronted with changes in the data

The DBMS is connected via 100 MHz Ethernet connection to two Ediflow instances running on two machines

- The first Ediflow machine computes visual attributes (runs the layout procedures)

- The second machine extracts nodes from VisualAttributes table and displays the graph

⇨ Adding increasing numbers of tuples to the database

Robustness evaluation

Inserting tuples requires performing a sequence of steps:

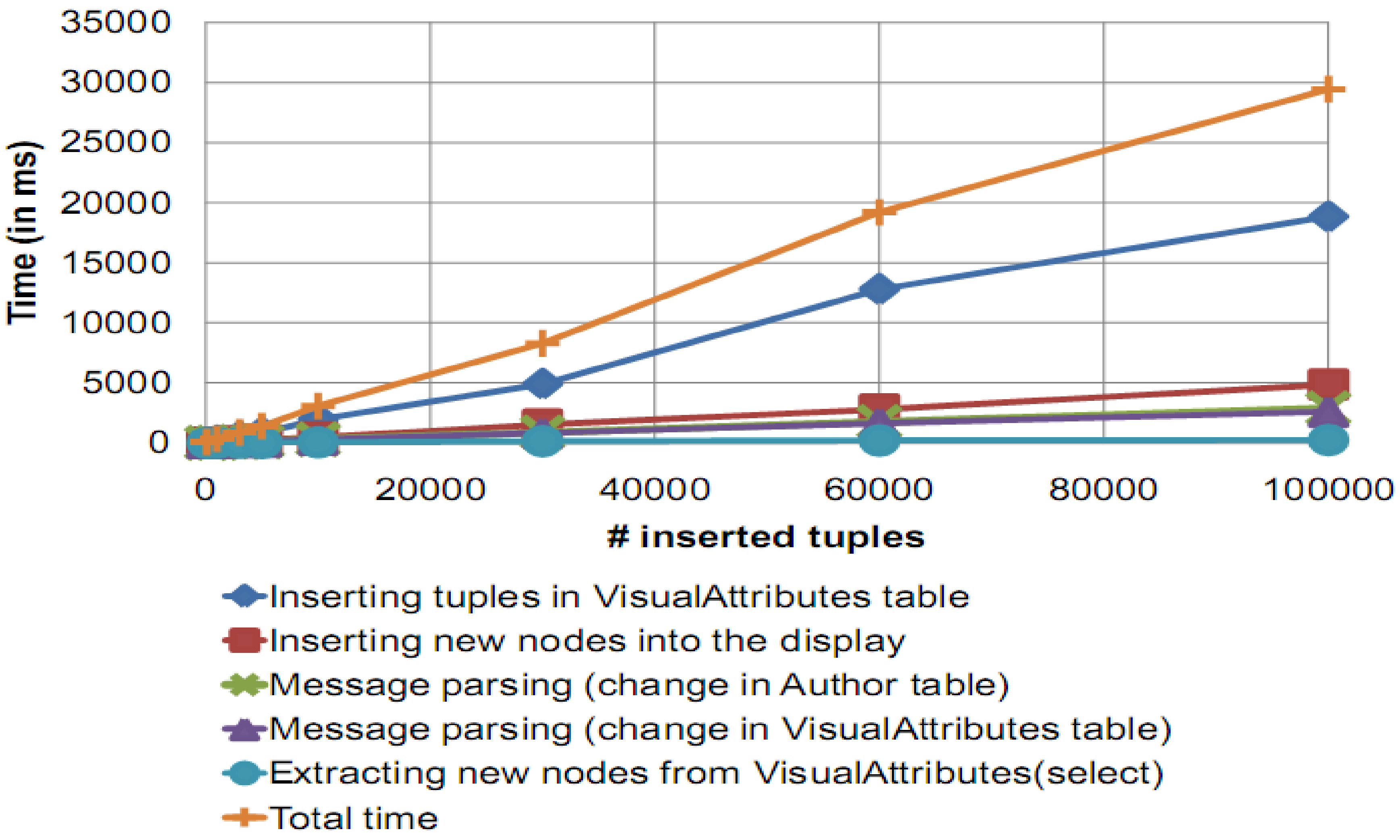
First machine

- Parse the message involved after insertion in nodes table
- Insert the resulting tuples in the VisualAttributes table

- Parse the message involved after the insertion in the VisualAttributes table
- Extract the visual attributes of the new nodes
- Insert new nodes into the display screen of the second machine

Second machine

Robustness evaluation



Robustness evaluation

Results:

- The times are compatible with the requirements of interaction
- They grow linearly with the size of inserted tuples
- The dominating time is required to write in the VisualAttributes table

⇨ The price to pay for having these attributes stored in a persistent database

Summary

- Design and implementation of Ediflow
- Workflow platform supporting visual analytics
- Ediflow unifies the data model used by all its components
- Supports standard data manipulation through procedures
- Reflects changes in the data through update propagations
- Several options are offered to react to such changes

Perspectives

- Improve the visual table schema
- Specify a collaboration management mechanisms
- Integration with current scientific workflow systems (Vistrail, Kepler, etc)

Thank you.

References

C.Scheidegger and T.VoHuy and D.Koop and J. Reire and C.T.Silva. **Querying and re-using workflows with VsTrails.**
SIGMOD 2008

I. Altintas and C. Berkley and E.Jaeger and M. Jones and B. Ludascher and S. Mock. **Kepler : An Extensible System for Design and Execution of Scientic Workows**.
SSDBM 2004.

I. Zachary and G.Todd and K. Grigoris and T. Nicholas and T. Val and T. Partha Pratim and J. Marie and P. Fernando
The ORCHESTRA Collaborative Data Sharing System.
SIGMOD 2008

The Kepler scientific workflow platform

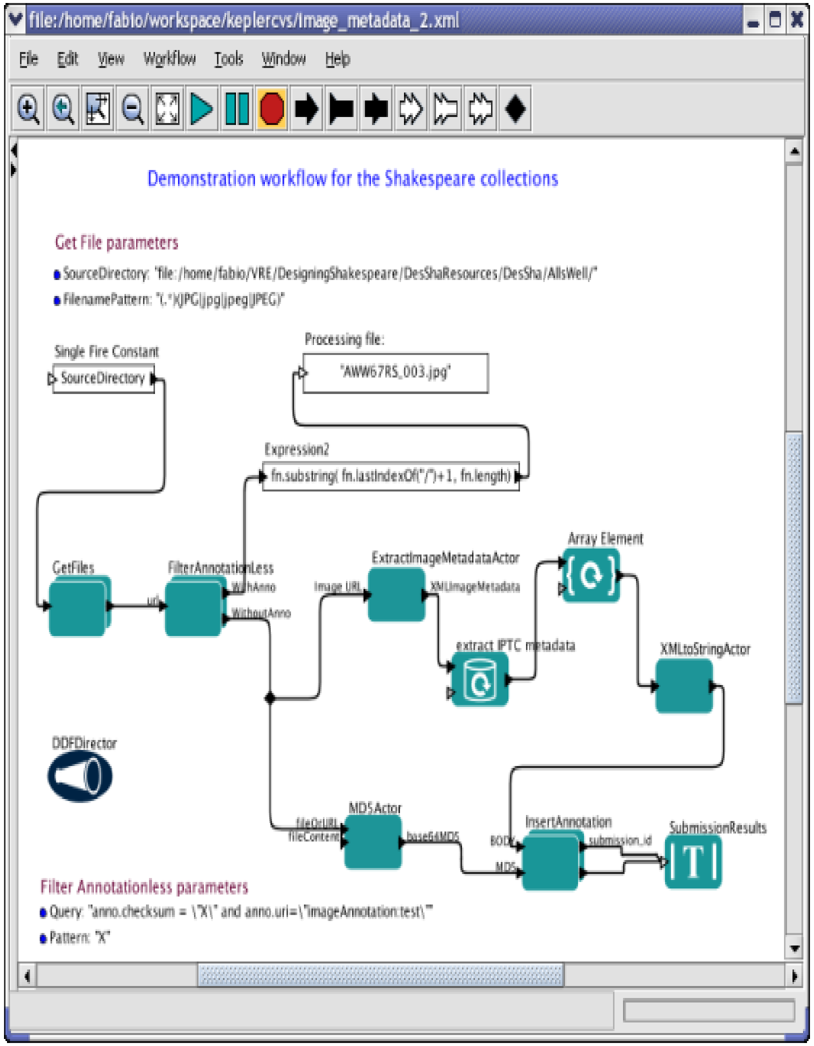
Recent and well-developed scientific workflow project

Helps scientists and analysts to create, execute and share models

Provides a GUI to create scientific workflow

No mechanism to handle dynamic data

Visualization remains external



Demonstration workflow for the Shakespeare collections

Get File parameters

- SourceDirectory: "file:/home/fabio/VRE/DesigningShakespeare/DesShaResources/DesSha/AllsWell/"
- FilenamePattern: "(.*)(JPG|jpg|jpeg|JPEG)"

Filter Annotationless parameters

- Query: "anno.checksum = \"X\" and anno.uri=\"imageAnnotation:test\""
- Pattern: "X"

The Orchestra platform

Data-centric P2P platform for scientific applications

Dedicated to bioinformatics

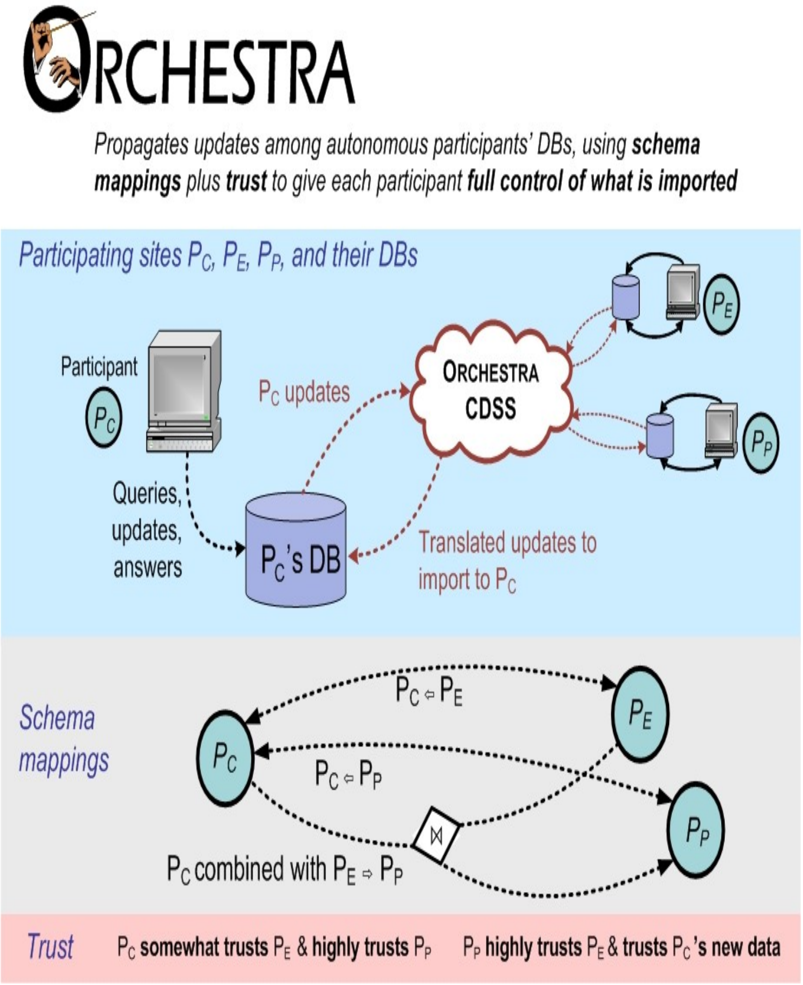
Focuses on data exchange mapping across different sources

Each peer's DB is updated to reflect updates in the other peers

No visualization

No interactivity

Leaves out external computations


ORCHESTRA
Participating sites P_C, P_E, P_P, and their DBs
Participant
P_C
P_C updates
ORCHESTRA CDSS
P_E
P_P
Queries, updates, answers
P_C's DB
Translated updates to import to P_C
Schema mappings
$P_C \Leftarrow P_E$
$P_C \Leftarrow P_P$
P_C combined with $P_E \Rightarrow P_P$
Trust
P_C somewhat trusts P_E & highly trusts P_P
P_P highly trusts P_E & trusts P_C's new data

Propagates updates among autonomous participants' DBs, using ***schema mappings*** *plus* ***trust*** *to give each participant* ***full control of what is imported***

The Vistrail platform

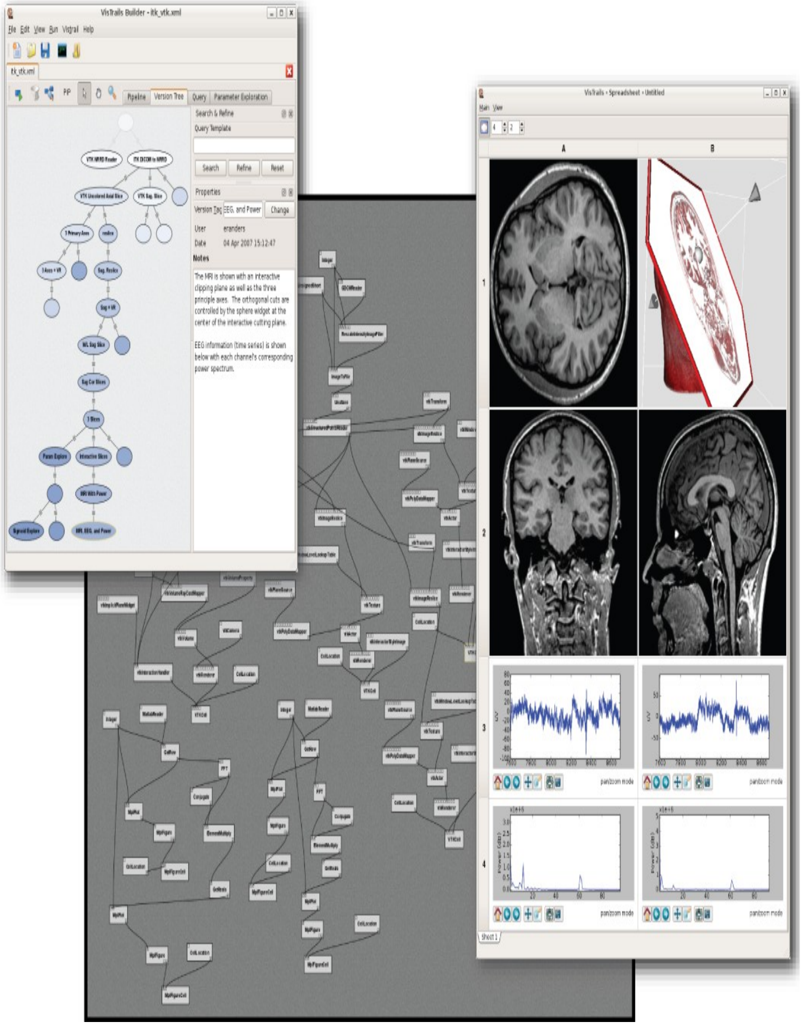
Combines features of workflow systems and visual analytics

Manages exploratory activities

Iteratively refines computational tasks

Maintains detailed provenance of the exploration process

No support for data dynamicity



Version Tag EEG, and Power

User eranders

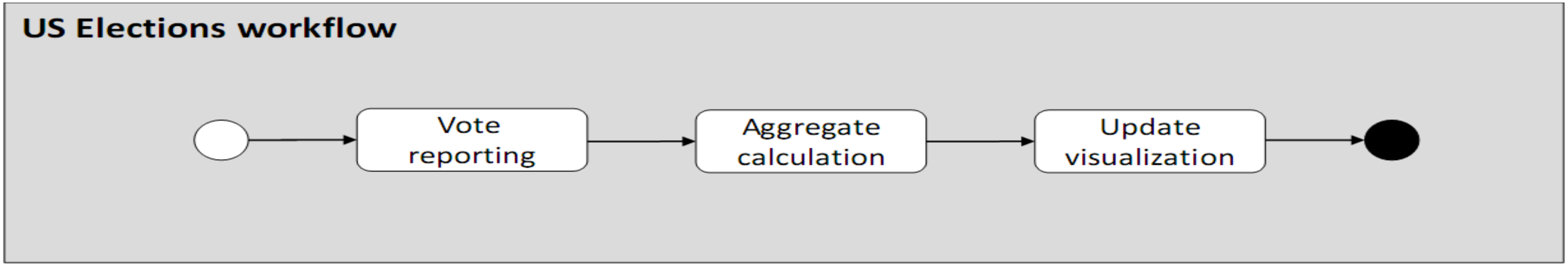
Date 04 Apr 2007 15:12:47

The MRI is shown with an interactive clipping plane as well as the three principle axes. The orthogonal cuts are controlled by the sphere widget at the center of the interactive cutting plane.

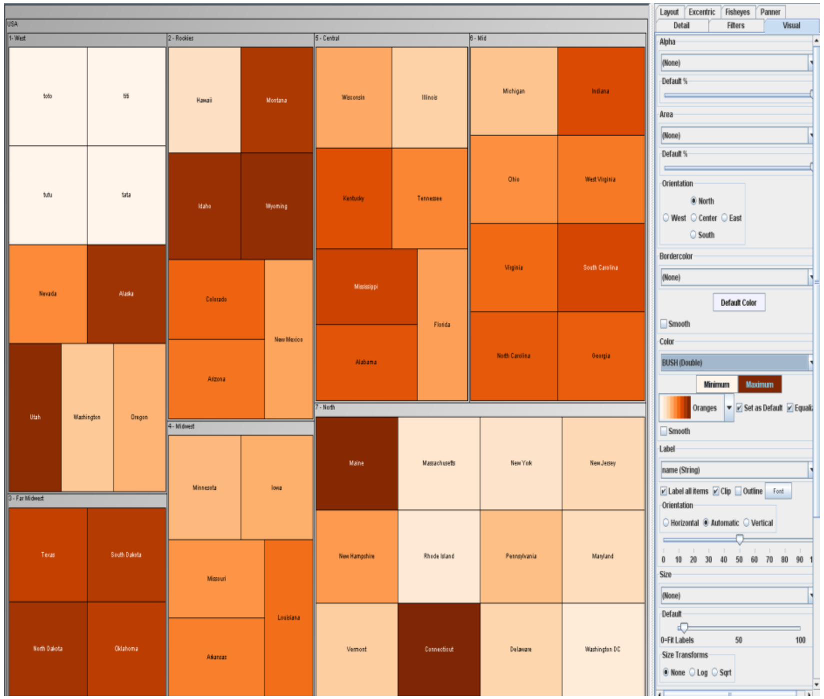
EEG information (time series) is shown below with each channel's corresponding power spectrum.

Use case 1: election scenario

Goal: Monitoring the results of the American presidential election


US Elections workflow

Vote reporting → Aggregate calculation → Update visualization



Alpha: (None)

Default %

Area: (None)

Default %

Orientation: ● North ○ West ○ Center ○ East ○ South

Bordercolor: (None)

□ Smooth

Color: BUSH (Double)

Oranges ☑ Set as Default ☑ Equali

□ Smooth

Label: name (String)

☑ Label all items ☑ Clip □ Outline

Orientation: ○ Horizontal ● Automatic ○ Vertical

Size: (None)

Default

0=Fit Labels 50 100

Size Transforms: ● None ○ Log ○ Sqrt

Retrieve the results of votes and update the database

Compute and store the number of votes of each party in each state in an aggregated table

Update the visualization and the view to reflect the new votes

Process model – Structured process

Process	: :=	Configuration Constant* Variable+ Relation+ Procedure* StructProcess
Configuration	: :=	DBdriver DBuri DBuser
Constant	: :=	name value name $\in N$, value $\in V$
Variable	: :=	name type name $\in N$, type $\in T$
Relation	: :=	name primaryKey RelType
RelationType	: :=	(attName attType)* attName $\in N$, attType $\in T$
Procedure	: :=	name classPath
StructuredProcess	: :=	Activity | Sequence | AndSplitJoin | OrSplitJoin | ConditionalProcess
Sequence	: :=	Activity **,** StructuredProcess
AndSplitJoin	: :=	**AND-split** (StructuredProcess)+ **AND-join**
OrSplitJoin	: :=	**OR-split** (StructuredProcess)+ **OR-join**
ConditionalProcess	: :=	**IF** Condition StructuredProcess
Activity	: :=	activityName Expression
Expression	: :=	askUser | callProcedure | runQuery