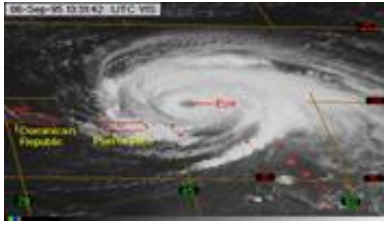
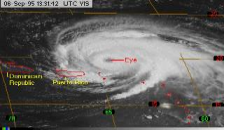
Incremental and Interactive Visualizations

Marco Angelini, Gabriele Curzi, **Giuseppe Santucci**

Università di Roma “La Sapienza”



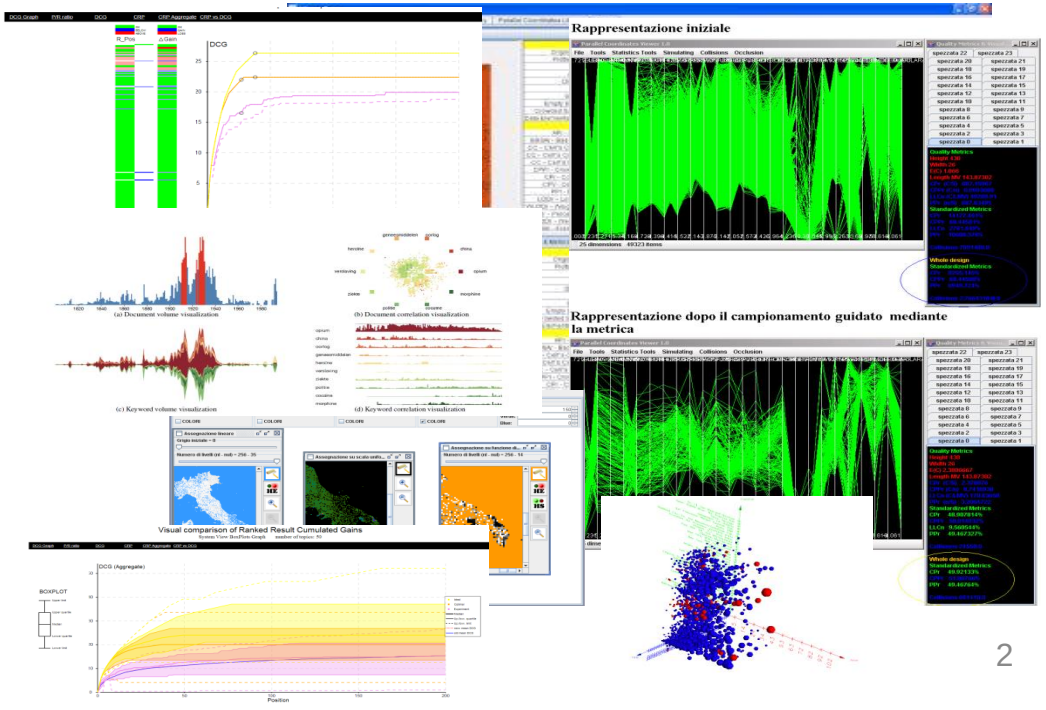
SAPIENZA
UNIVERSITÀ DI ROMA



Who am I?

(University of Rome is so big…)

- **VisDis** and the **Database & User Interface** groups are two tightly connected research groups at the **Department of Computer, Control, and Management Engineering** (32 full professors, 19 associate ,and 13 assistant professors) of **Rome Faculty of Computer Science, Automatic, and Statistics**
- The VisDis and the Database/Interface group background is about:
  - Visual Information Access
  - Data quality
  - Data integration
  - User Centered Design
  - Usability and Accessibility
  - Infovis evaluation
  - Visual quality metrics
  - Visual Analytics
    - Data sampling
    - Density map optimization
    - Interactive Visualization
    - VA for Information Retrieval



Outline

- Introduction
- Applications' classification
- A model for interactive visualizations
- Examples
- Work in progress
- Conclusions

Introduction

- A number of applications call for the incremental/iterative drawing of a visualization (+ interaction)
  - Data streams
  - Computational Intensive Visual Analytics
  - Slow Cloud streaming
- ...either to follow the natural evolution of the data or to speed up the **interaction** and the **visualization**

Data Streaming

- **Continuously changing data**
- **Alerting changes is often the main activity[1]**
- **In some cases it is not possible to store the whole stream (Efficient statistical indicators are mandatory)**

[1] e.g. XIE Z., WARD M. O., RUNDENSTEINER E. A.: Visual exploration of stream pattern changes using a data-driven framework

Cloud Streaming

- **Visualization of big amount of data**
- Slow connection rate
- Large amount of time can elapse between the data request and their visualization
- “Smart” incremental visualization could mitigate the problem

We assume that statistical information about the dataset is available and that the streaming is performed in a random fashion

Computational Intensive V.A. applications

- Intensive computations slow down interaction speed, that is a mandatory feature of V.A. applications
- Incrementally visualizing **partial results** provides means for quicker interaction

- **Partial results ? They come from:**
  - Inherently iterative algorithms
  - Forcing non-iterative algorithms to be iterative, e.g. applying a non iterative algorithm to random chunks of data (errors will likely rise…)

Our approach

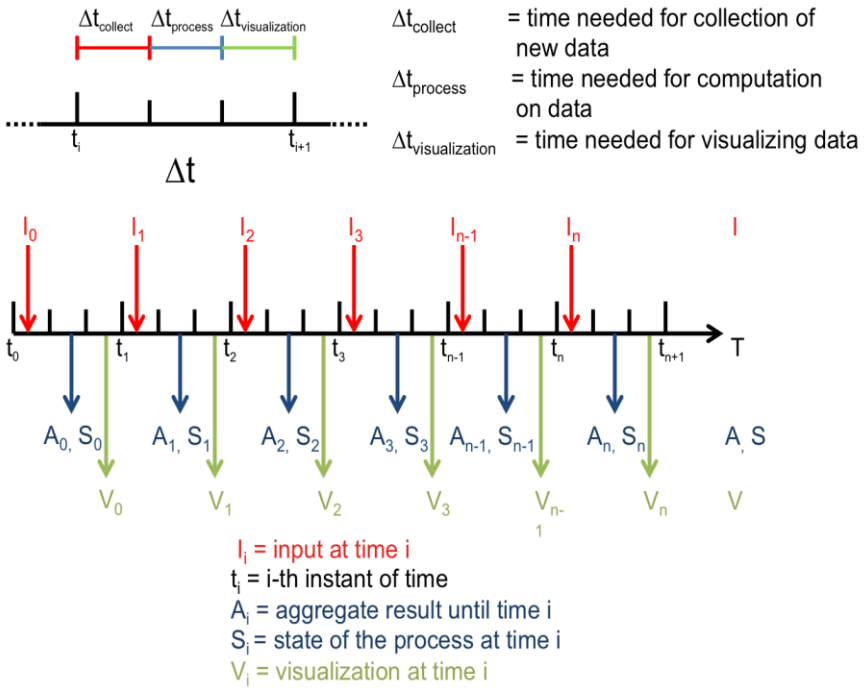
We are working on a formal model for characterizing the drawing of interactive visualizations:

- …grouping different kinds of applications under the unique umbrella of their need of dealing with interactive visualizations
- ...describing practical issues
- ...outlining the main evaluation parameters
- ...instantiating the model on 3 proof-of-concept scenarios

Quick overview of the model

Modeling time-oriented issues

- The incremental nature of the visualizations calls for the choice of a quantum of time Δt composed by several sub-intervals:
  - One for collecting data
  - One for processing data
  - One for visualizing data
- …plus several functions for modeling the actual state



$$\begin{cases} A_{i+1} = F(A_i, S_i, I_{i+1}) \\ S_{i+1} = G(S_i, I_{i+1}) \end{cases}$$

Quality indicators

- We use some metrics for modeling several relevant aspects:
  - the difference between two internal state, e.g.:

$$\Delta data^{actual}(R_i, R_j) = \sum_{k=1}^{N} \frac{\left| d_{j,k} - d_{i,k} \right|}{N}$$

  - the difference between two visualizations, e.g.:

$$\Delta vis^{actual}(R_i, R_j) = \sum_{k=1}^{N} \frac{\left| color(d_{j,k}) - color(d_{i,k}) \right|}{N}$$

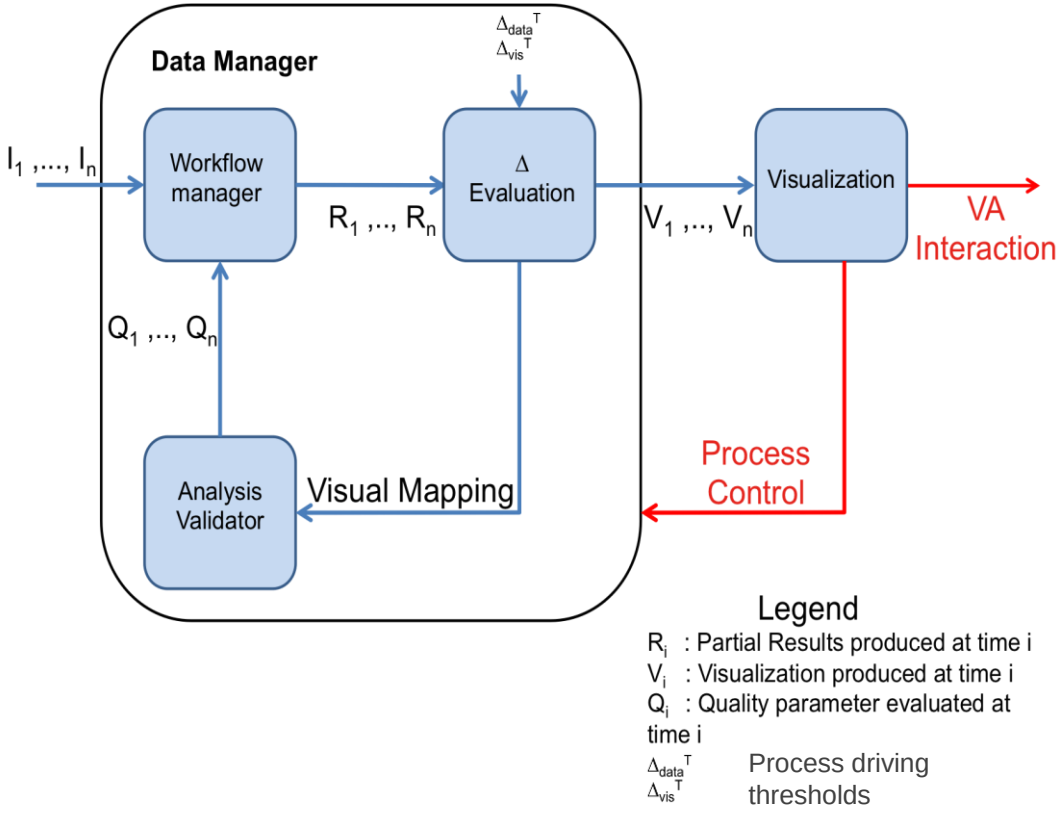
  - the error introduced in the computation of intermediate results, e.g.:

$$Q\varepsilon_{current} = Q\Delta_{current} + \varepsilon_{stat}(i)$$

in order to:

- assess the actual image quality
- detect image changes
- drive the overall process workflow (e.g., skipping useless operation and/or stopping the process)

Visualization control-flow

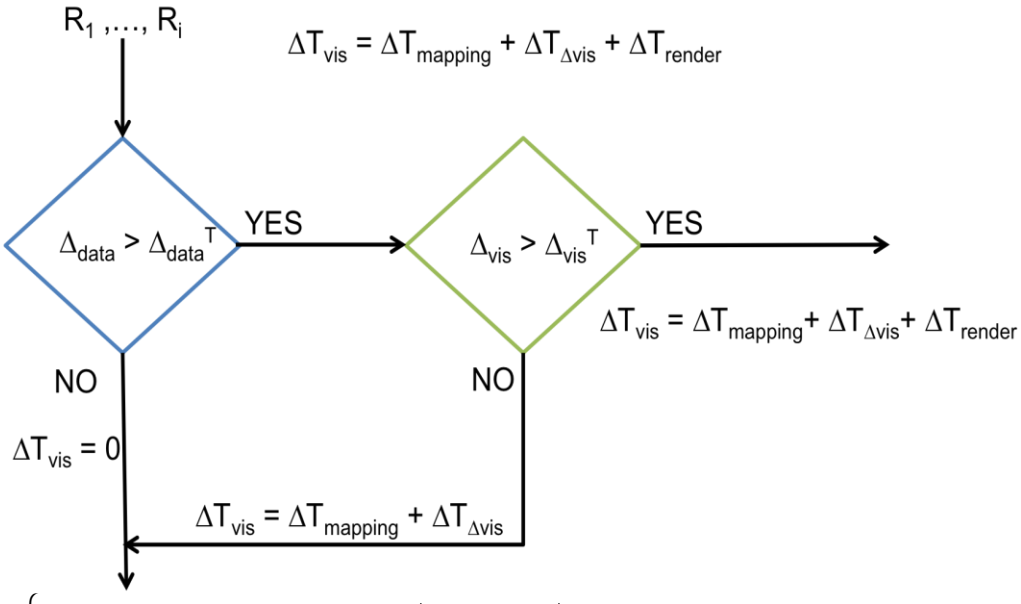


- Time modeling and quality indicators allow for driving the workflow of an incremental visualization process
- Why produce a new visualization when the underlying data has not changed?

Some details about the workflow

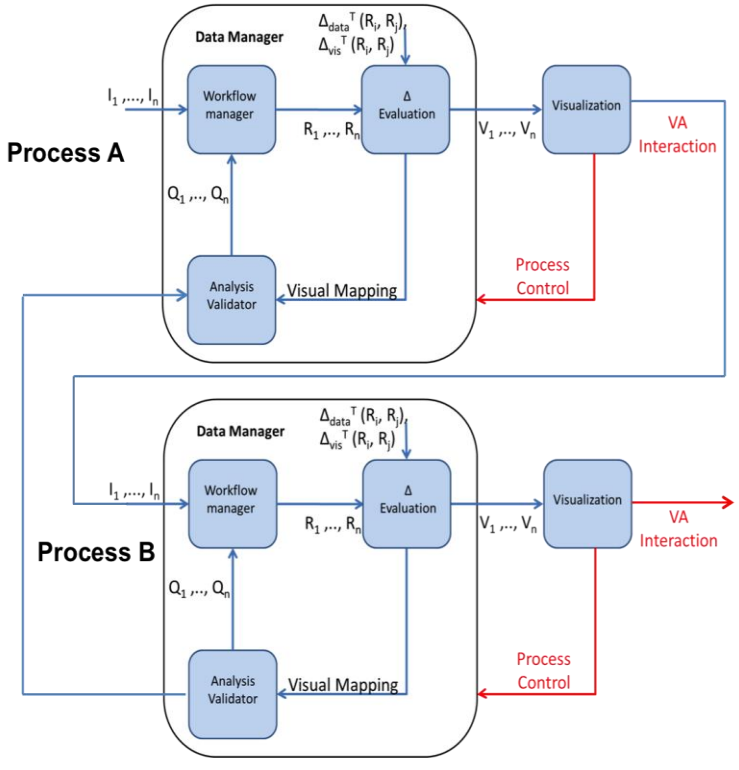
-The model allows to save computation cycles by comparing actual Δ_{data} and/or Δ_{vis} against threshold values, skipping useless activities

-This behavior helps in achieving a quicker interaction with the visualization



$$\begin{cases} Ddata^{actual}(R_i, R_j) = \sum_{k=1}^{N} \frac{\left| d_{j,k} - d_{i,k} \right|}{N} \\ Dvis^{actual}(R_i, R_j) = \sum_{k=1}^{N} \frac{\left| color(d_{j,k}) - color(d_{i,k}) \right|}{N} \end{cases}$$

Interaction with partial visualizations

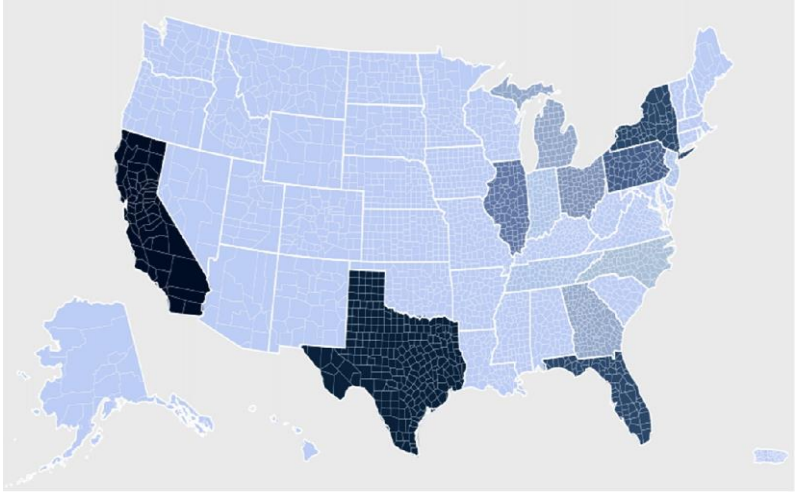


- A typical VA interactive exploration is to use the actual visualization to start a new analysis (automated or visual) on a subset of the data associated with the current visualization
- This subset is the input of a new process, totally distinct from the original one
- The model provides formal means for synchronizing such activities

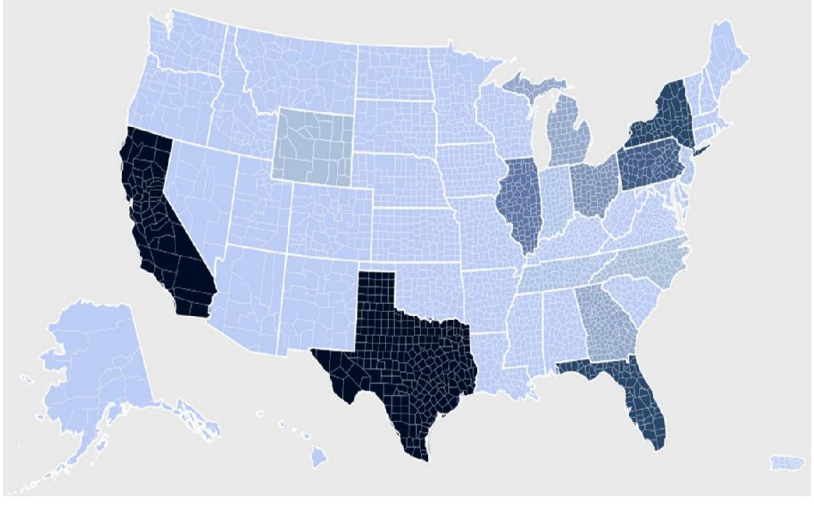
Examples

Data streaming: detecting changes

We simulate a data stream using the NHTSA Fatality Analysis Reporting System data about 20 years of fatal car accidents



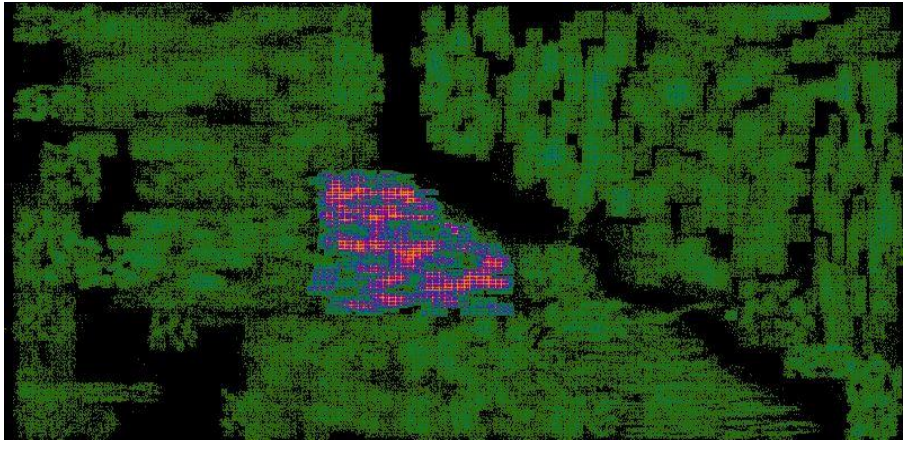
distribution of fatal car accidents at time t_i



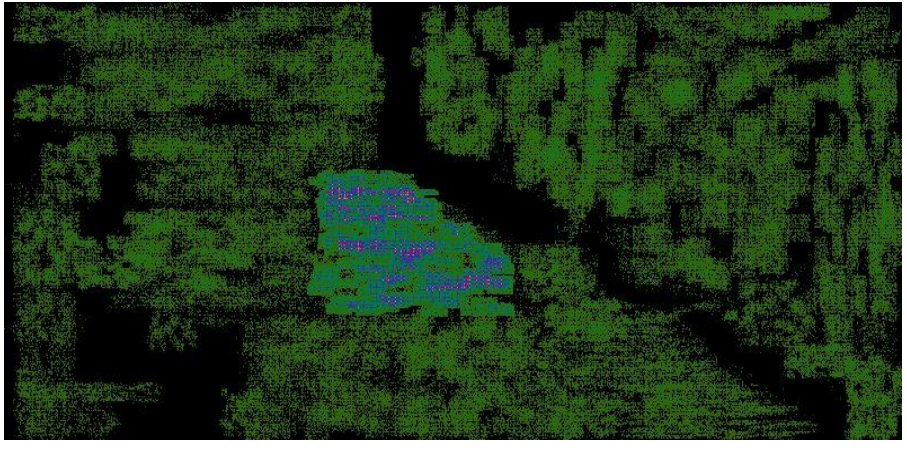
detecting a change at time t_j(changes in both actual Δ_{data} and Δ_{vis} are above the thresholds)

Cloud streaming: detecting convergence

We incrementally visualize random chunks of the VAST 2011 mini-challenge 1 using Δ_{vis} to estimate the convergence of the process



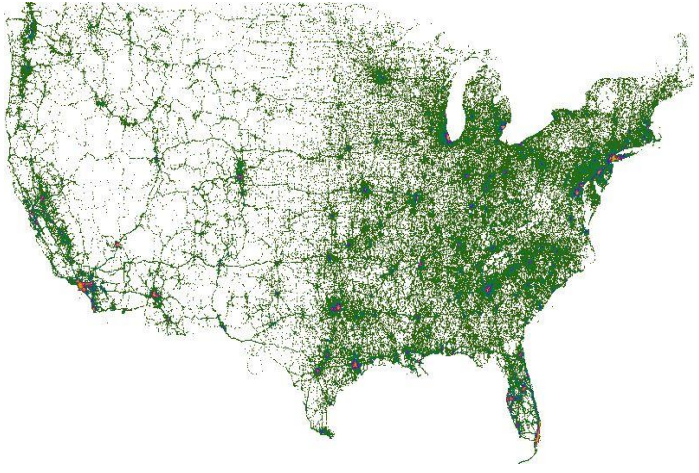
Whole dataset visualization



First intermediate valid state (40% of the dataset): quality indicators are below the threshold

Computational intensive VA algorithm

NHTSA Fatality Analysis Reporting System data: 310,254 fatal accidents, from 2001 to 2009, plotted by their longitude and latitude

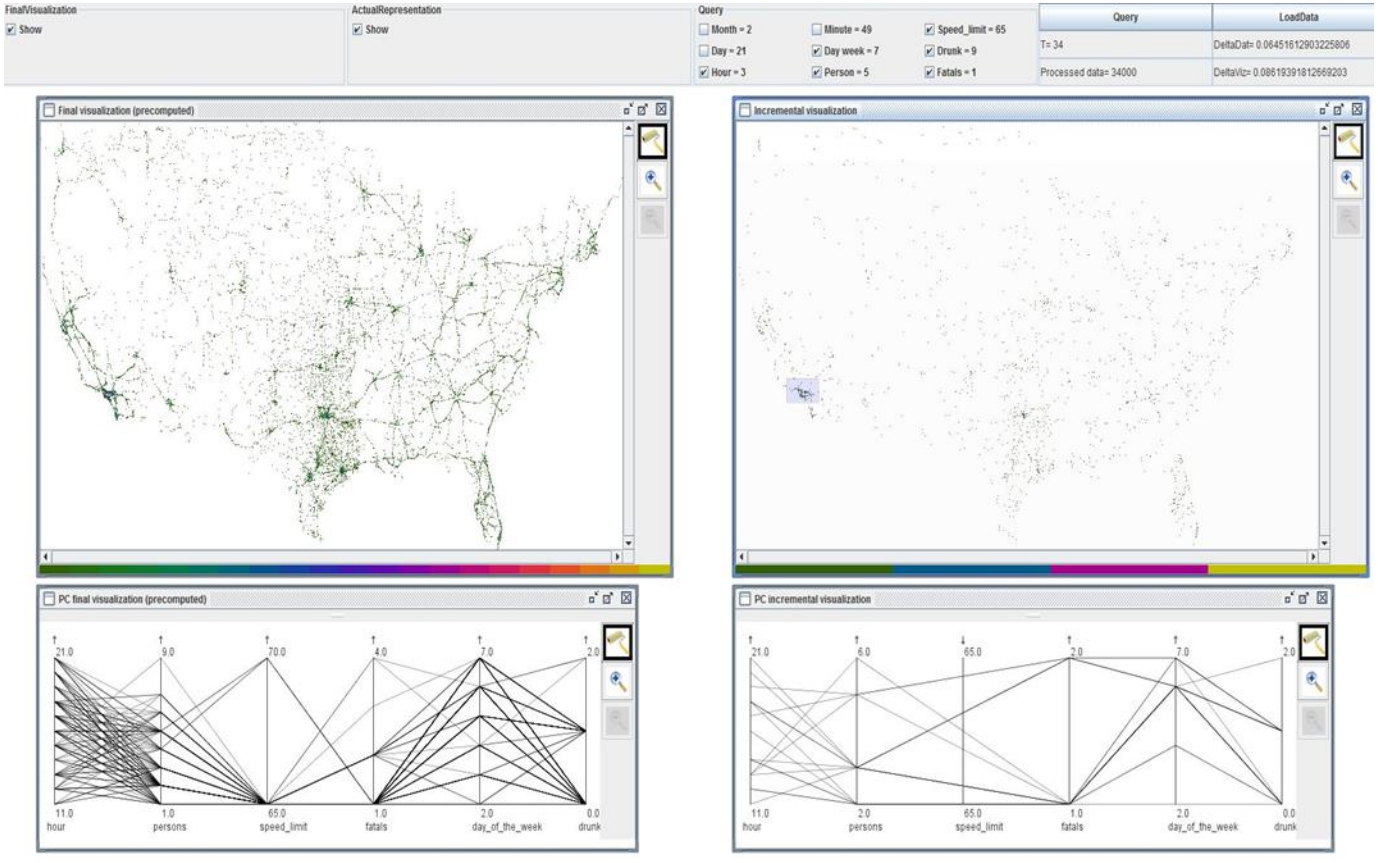


Goal: the user selects a representative accident and asks the system to compute the subset of similar (Euclidian distance) accidents and to plot them on a density map

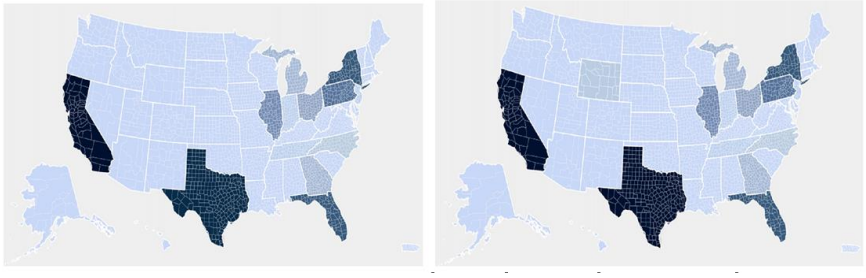
- **Representative accident:**

1) Time of the accident
2) Number of involved people
3) Speed limit
4) Dead people
5) Day of the week
6) Driver drunkenness

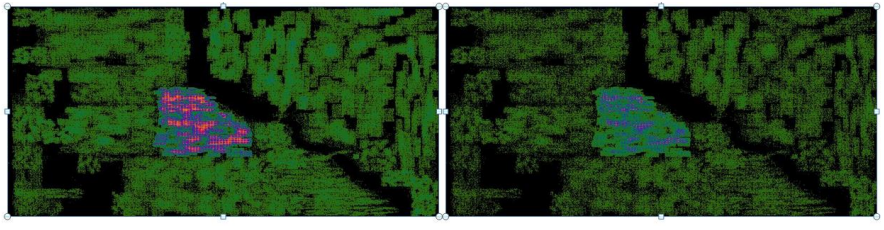
Video


FinalVisualization
Show
ActualRepresentation
Show
Query
Month = 2
Minute = 49
Speed_limit = 65
Day = 21
Day week = 7
Drunk = 9
Hour = 3
Person = 5
Fatals = 1
Query
LoadData
T= 34
DeltaDat= 0.0645161290322580 6
Processed data= 34000
DeltaViz= 0.0861939181266920 3
Final visualization (precomputed)
Incremental visualization
PC final visualization (precomputed)
21.0
9.0
70.0
4.0
7.0
2.0
11.0
1.0
65.0
1.0
2.0
0.0
hour
persons
speed_limit
fatals
day_of_the_week
drunk
PC incremental visualization
21.0
6.0
65.0
2.0
7.0
2.0
11.0
2.0
65.0
1.0
2.0
0.0
hour
persons
speed_limit
fatals
day_of_the_week
drunk

Δ_{vis}?



- Low value = nothing interesting has happened
- High value = wake up the user !!



- High value = we are fare from converging, keep drawing
- Low value = two consecutive images are very close each other: stop drawing

Work in progress

We are going to analyze the most widely used algorithm in Visual Analytics, and for each of those:

1. Is the algorithm inherently iterative?

2. If it is not inherently iterative, there exists an iterative version of the algorithm?

3. The error introduced in the partial results can vary in a not-monotonic way?(can the algorithm change it's mind)?

4. If the algorithm was not inherently iterative, but has an iterative implementation, are the two results equals?

Paper analysis

Number of **Visual Analytics papers** analyzed:
173

Number of **papers containing algorithms**:
102 (59% of analyzed papers)

Number of papers containing **"Well Known" algorithms**:
56 (55% of papers with algorithm)

On the table are reported the occurrences of the "Well Known" algorithms.

The "Well Known" algorithms used only once have been cut from the table (19 algorithms).

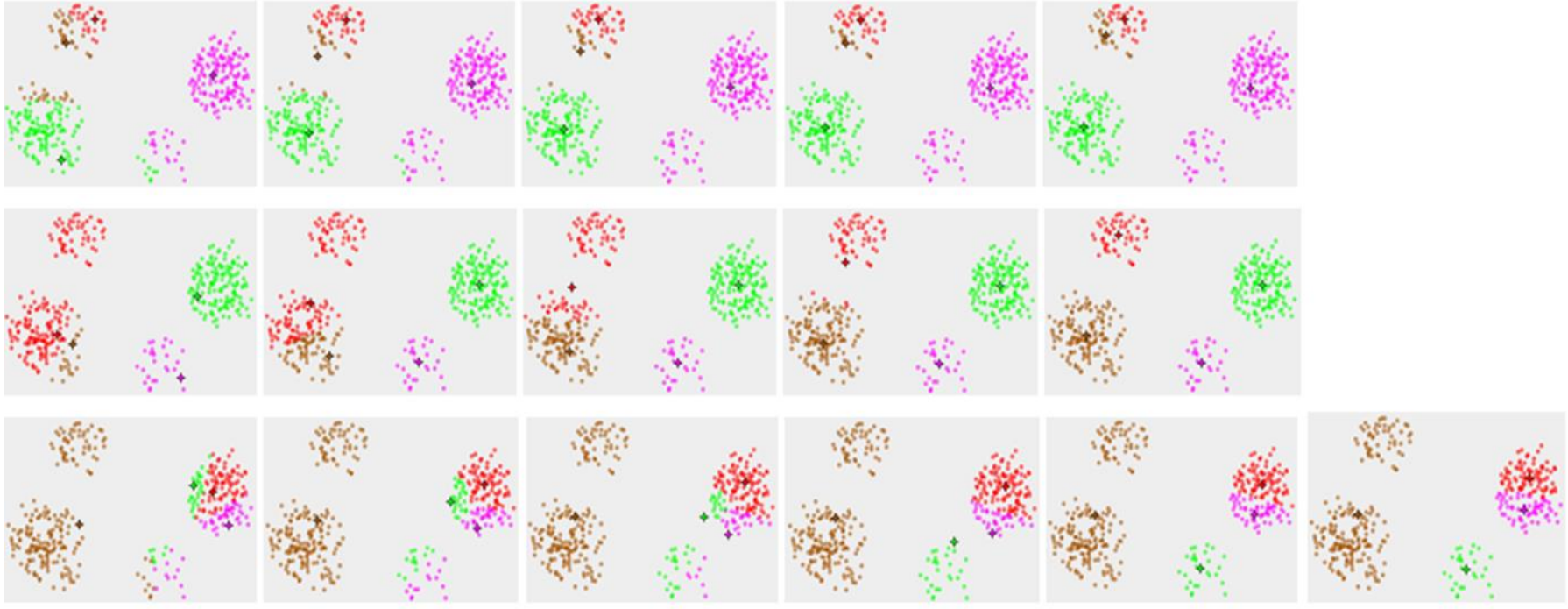
Note: inside a single paper there could be used more than a single algorithm.

Algorithm	% use on Visual Analytics papers	Algorithm family			
		Clustering	Dimension Reduction	Classification	Other
K-Means	13% (13/102)	X	-	-	-
Hierarchical clustering	8% (8/102)	X	-	-	-
DBSCAN	6% (6/102)	X	-	-	-
SOM	6% (6/102)	X	-	-	-
PCA	11% (11/102)	-	X	-	-
LDA	6% (6/102)	-	X	-	-
k-nearest neighbor	7% (7/102)	-	-	X	-
TF-IDF	3% (3/102)	-	-	X	-
SVM	2% (2/102)	-	-	X	-
CUSUM	3% (3/102)	-	-	-	X
Kruskal MST	2% (2/102)	-	-	-	X
E-M	2% (2/102)	-	-	-	X

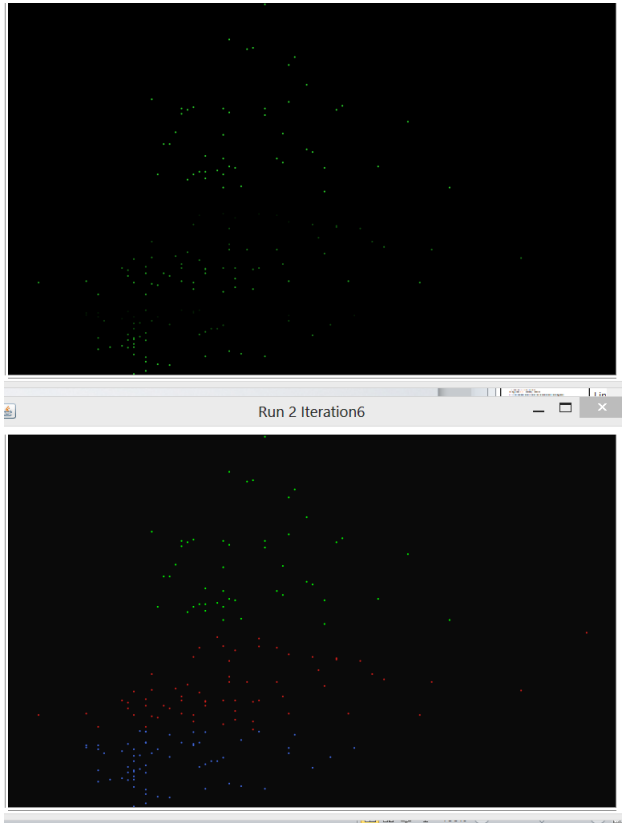
Algorithm analysis

Algorithm	% use on Visual Analytics papers	Algorithm family			Complexity	Inherently Iterative	Can be made an iterative version	Monotonicity	Introduced error on iterative version	Multi-run version
		Clustering	Dimension Reduction	Classification						
K-Means	13% (13/102)	X	-	-	$O(n^{kd})$	Yes	-	No	-	Yes
Hierarchical clustering	8% (8/102)	X	-	-	$O(n^3)$	Yes	-	Yes	-	No
DBSCAN	6% (6/102)	X	-	-	$O(n^2)$	Yes	-	Yes	-	No
SOM	6% (6/102)	X	-	-	$O(nmd)$	Yes	-	Partial	-	Yes
PCA	11% (11/102)	-	X	-	$O(nd^2)$	No	Yes	Yes	Partial	?
LDA	6% (6/102)	-	X	-	$O(mnt+t^3)$	No	?	?	?	?
k-nearest neighbor	7% (7/102)	-	-	X	$O(knlog_n)$	Yes	-	Yes	-	Yes

e.g., K-means (multi-run version)



e.g.,K-means (multi-run version)



Usage scenario: The Panoptesec Project

The total budget of the project is approximately 7,5 million € and the European Commission is funding 70%
START September 2013
STOP September 2016
The project full title is “Dynamic Risk Approaches for Automated Cyber Defence” and will be completed in 2016

Panoptesec's objectives

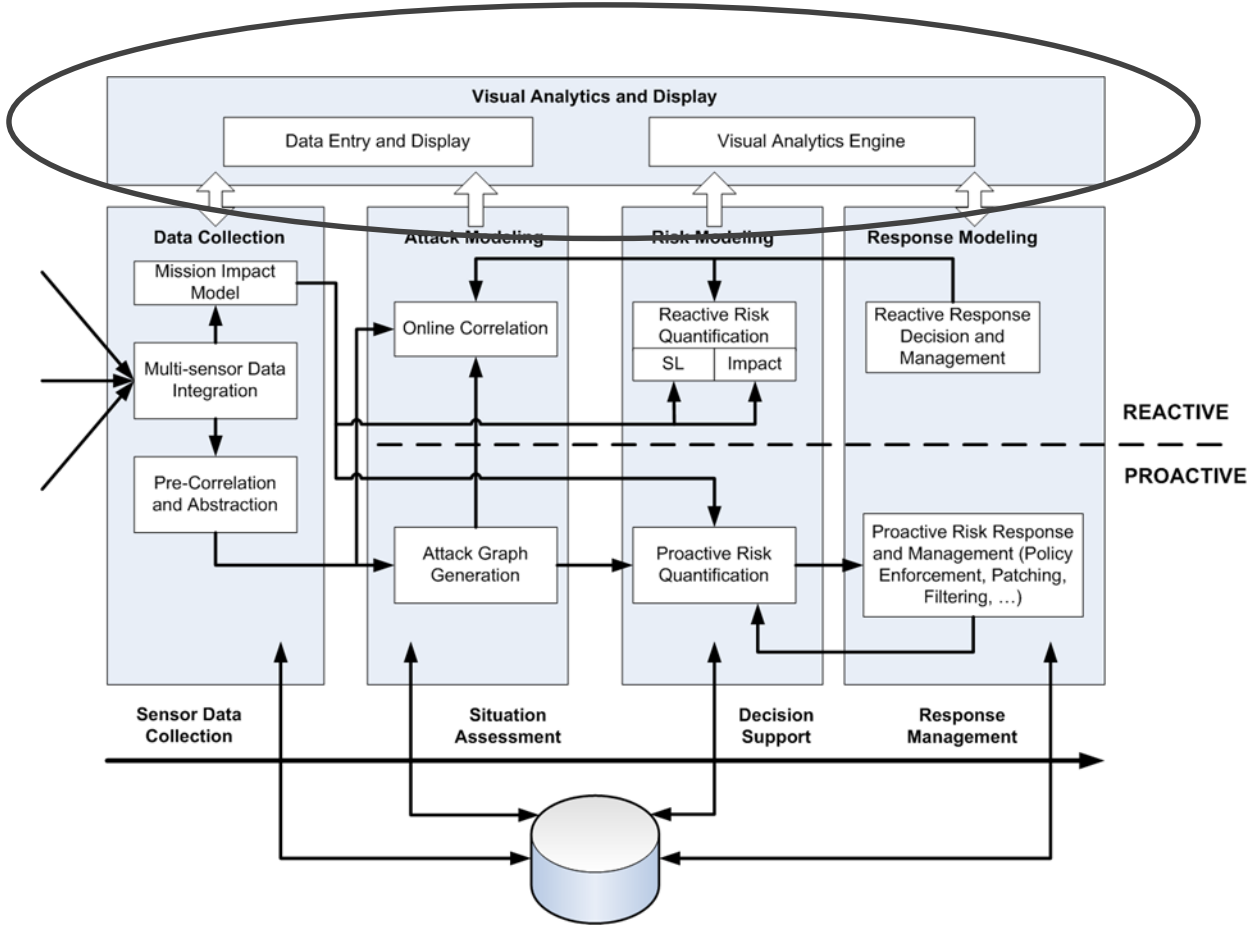
- Objective ICT-2013.1.5 Trustworthy ICT

c) Development, demonstration and innovation in cyber security

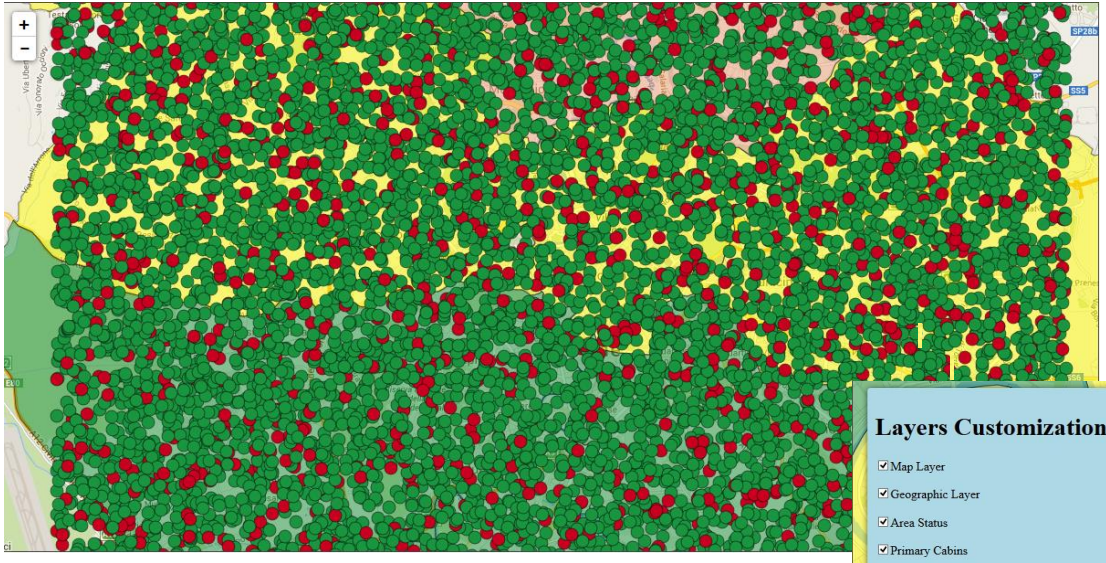
"This activity addresses the application of technologies to ***increase the level of cyber security*** in Internet. This includes the development and demonstration of technologies, methodologies and processes to ***prevent, detect, manage and react*** to cyber incidents in real-time, and to support the breach notifications, improving the ***situational awareness*** and supporting the ***decision making*** process. It will also develop and demonstrate advanced technologies and tools that will empower users, notably individuals and SMEs, in ***handling security incidents*** and ***protecting their privacy***."

Real Time geographical visualization : incremental clustering!

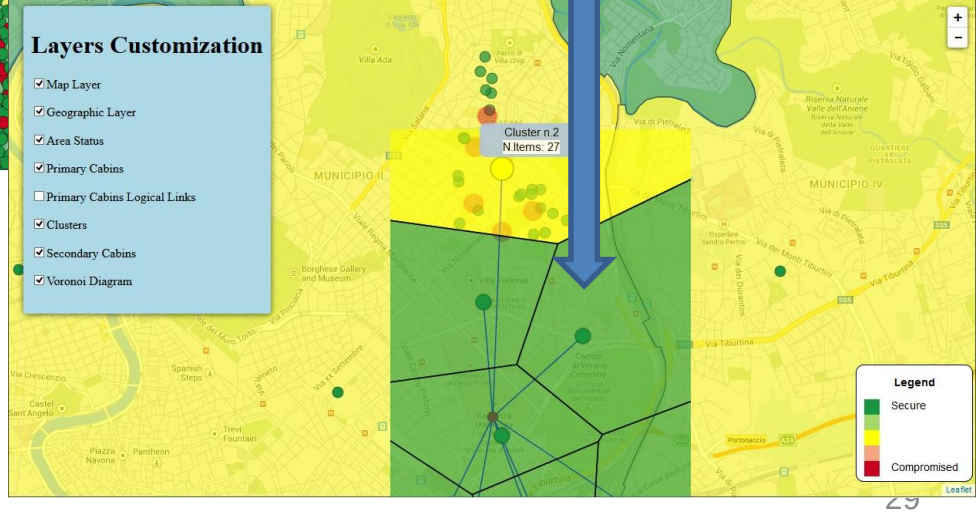
Architetcture

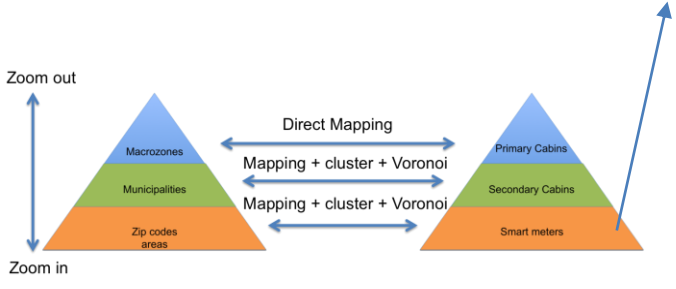


Incremental K-means clustering



Voronoi maps?





Conclusions

Our work addresses the problem of formal modeling a generic incremental drawing of a visualization, practical issues and parameters that can be used to drive and evaluate the whole process

We are currently:

- investigating the most used VA algorithms to adapt them to our model
- investigating how to provide the user a feedback on the errors and stability of displayed images
- modeling the error that rise from the visual interaction with partial results that is used to start new analytical activities