Fourth Brazil-France Workshop

On High Performance Computing and Scientific Data Management Driven by Highly Demanding Applications

Unveiling objects in Big Data

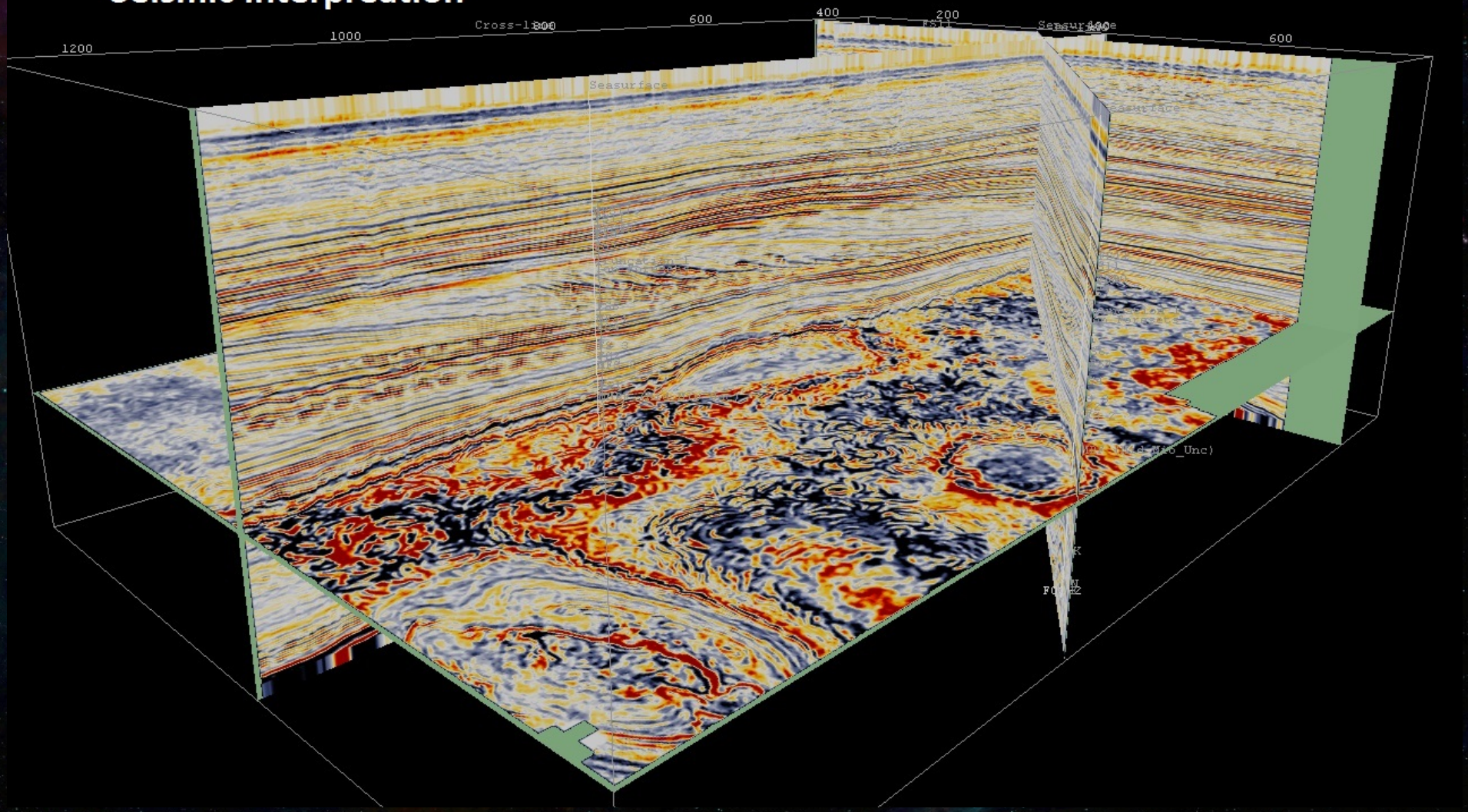
supervisor: Fabio Porto

By Amir Khatibi

Sep 2014

Laboratório Nacional de Computação Científica

Seismic Interpreation



Big-Data (in science) Data Challenges

- Data Representation
  - Different Data Models:
    - Data structure and query languages
    - Graphs, Matrixes, Key-Value,…
- Data Uncertainty
  - Data is uncertain
  - uncertainty quantification on data
- Data Partitioning
  - in sync with data processing
- Data Heterogeneity
  - Data Granularity

Data Deluge and importance of Data Analytics

- The rapid increase in the amount of published information and data hide interesting objects from users.
- Extract meaning of large volumes of data

Problem Formulation

- Lack of kno
- The si
- T
- We
- Pattern Query

similarity match

acteristics

or by g (query by example)

Pattern Query:
List of object's features
Q = {q1, q2, …, qn}

FRCS: Find and Rank the Candidate Solutions

first step) Define a strategy to find candidate solutions to Q

- $$Result = F_{shape}(F_{element}(Query))$$

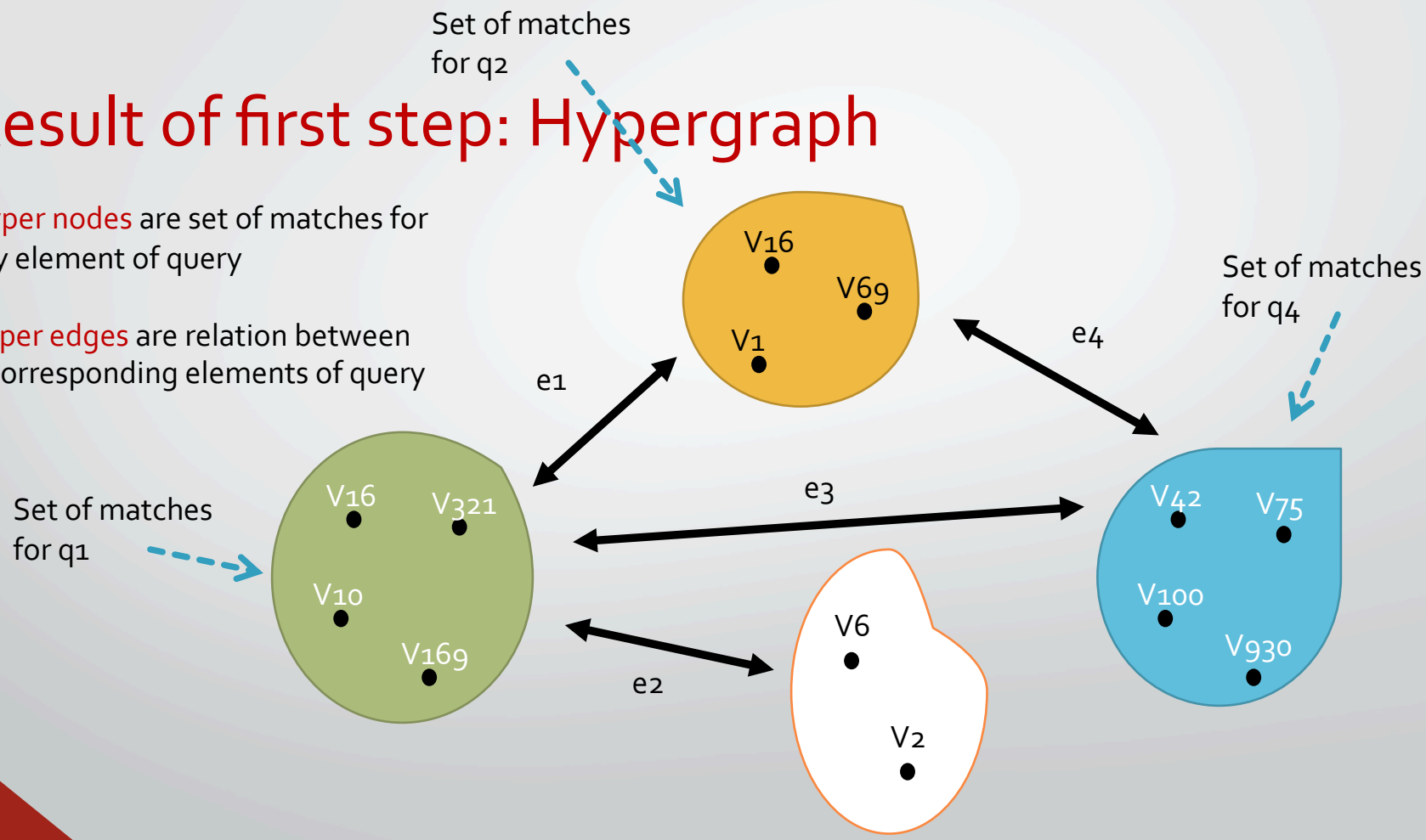
We propose a function which provides transformation like Shape Context (Point Matching Algorithm)

We propose a function which provides scaling like DTW (Sequence Matching Algorithm)

Result of first step: Hypergraph

V: hyper nodes are set of matches for every element of query

e: hyper edges are relation between the corresponding elements of query



FRCS: Find and Rank the Candidate Solutions

second step) Define the cost function to rank the candidate solutions.

Cost function: descriptor that weights the presence of the characteristics of query or/and distance between elements of candidate solutions.

Result of second step: Cost Function

- Total_Cost = $\sum$ wi*(Match_Cost of element i) + $\sum$ (1-wi)*(Distance_Cost between elements i, i+1)

wi = criteria weights

Application

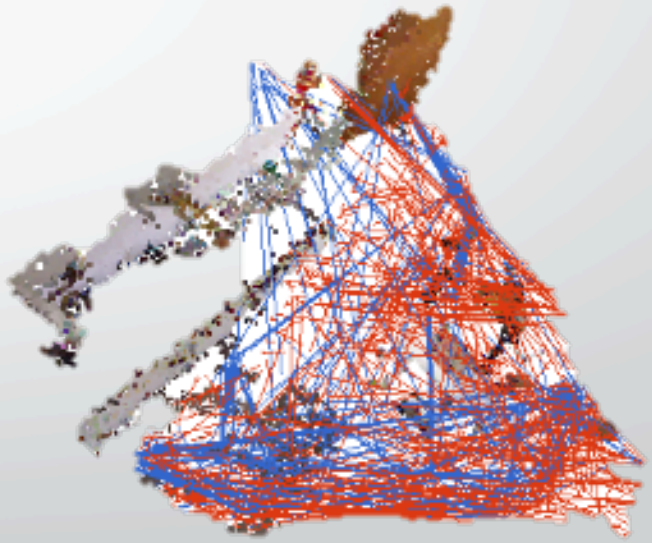
a sample company wants to advertise a new waterpark to specific people

- Simple Pattern Query:

  Finding people that during four weeks repeat these actions:

  - go to a pool
  - go to any fast food restaurant
  - go to a beach

- Constraint: the same action is performed at different times and places by different people, possibly at different speeds.



Trajectory of people's daily life
Colors show the people's motion level

Solution: FRCS technique

First step: Find the candidate solutions

Second step: Rank the candidate solutions

- First factor: Big Data: everyday trajectories
- $F_{distance}$: everyday distance between the elements; candidates with smaller distance between their elements have lower costs
- $F_{element}$: find people(go to a pool, go to any fast food, go to beach)
- F_{shape}: combination of elements(people who went to all these places at least four times for each)

Total_Cost = w1*(Match_Cost) + w2*(Distance_Cost)

wi = weight the criteria

Thank you

for your attention