7th Workshop on Visual Analytics, Information Visualization and Scientific Visualization

Honorable Mention

@

Niterói - Brazil
SIBGRAPI'17
30th Conference on Graphics, Patterns and Images
17 - 20 OCTOBER

The Dark Side of Progressive Visual Analytics

Marco Angelini and **Giuseppe Santucci**
[angelini][santucci]@dis.uniroma1.it

A.WA.RE: **A**dvanced **V**isualization & **V**isual **A**nalytics **RE**search Group

@ La Sapienza, University of Rome

SAPIENZA
UNIVERSITÀ DI ROMA

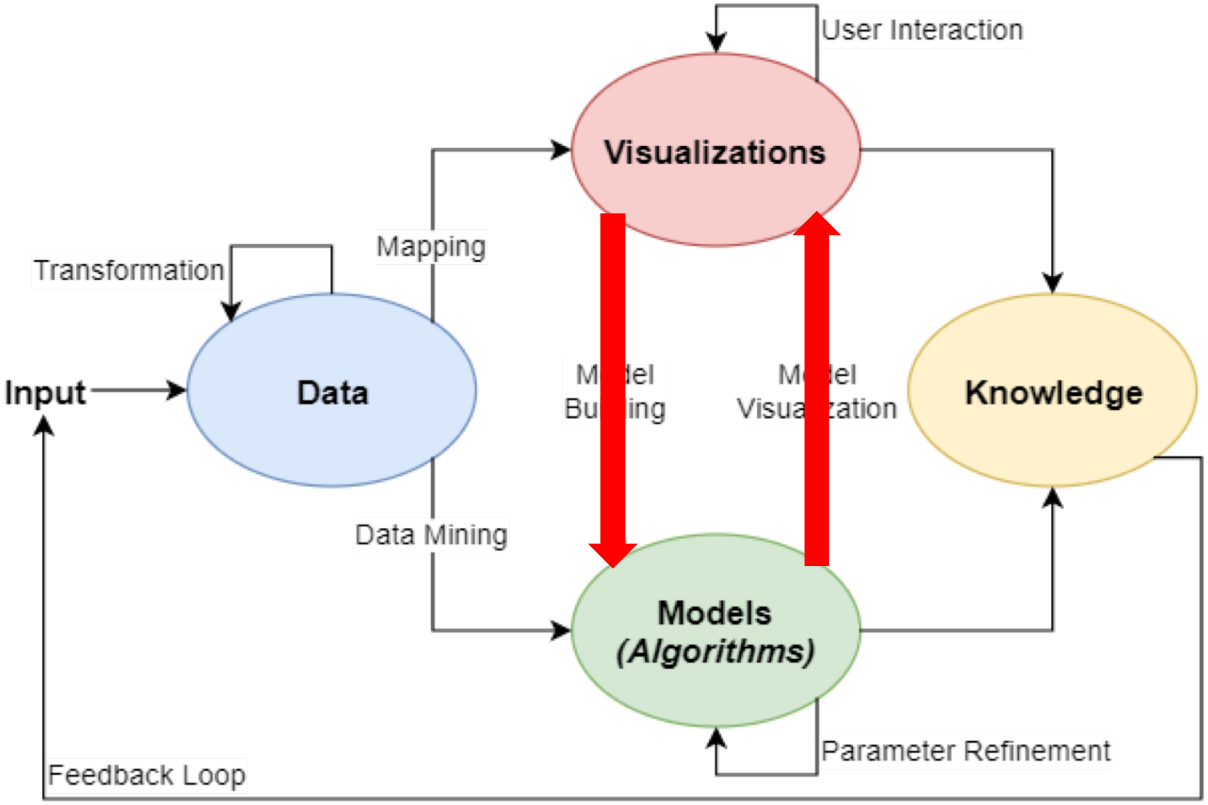
CIS SAPIENZA
CYBER INTELLIGENCE AND INFORMATION SECURITY

OUTLINE

- Progressive Visual Analytics ?
- Problems and issues (i.e., the dark side!)
- A clarifying example
- Open issues

VISUAL ANALYTICS

Combination of automated analysis techniques with interactive visualizations to process and analyze the information spaces

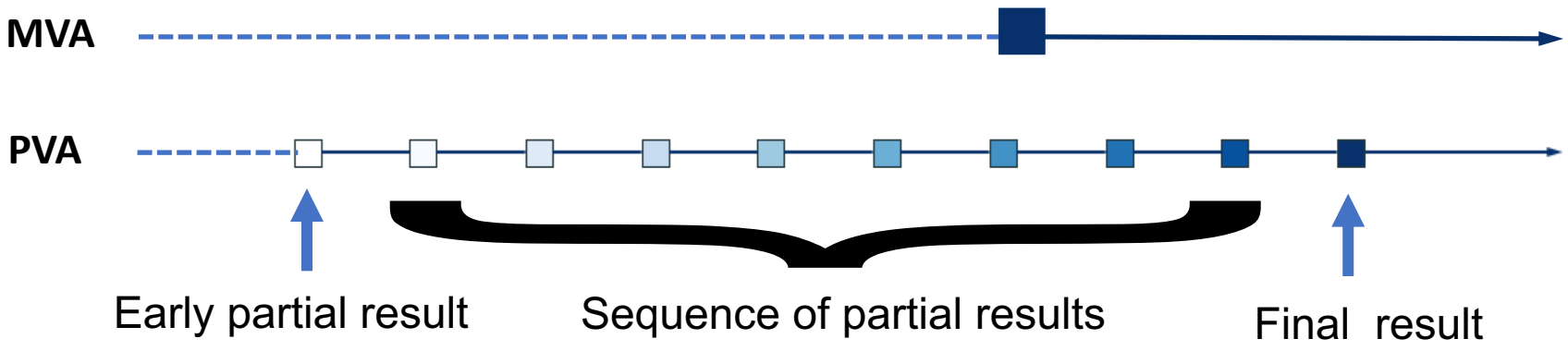


Long-running cycles making the approach unfeasible !!!

PROGRESSIVE VISUAL ANALYTICS (PVA)

When parallel computation and faster CPU are not enough, PVA addresses the problem producing

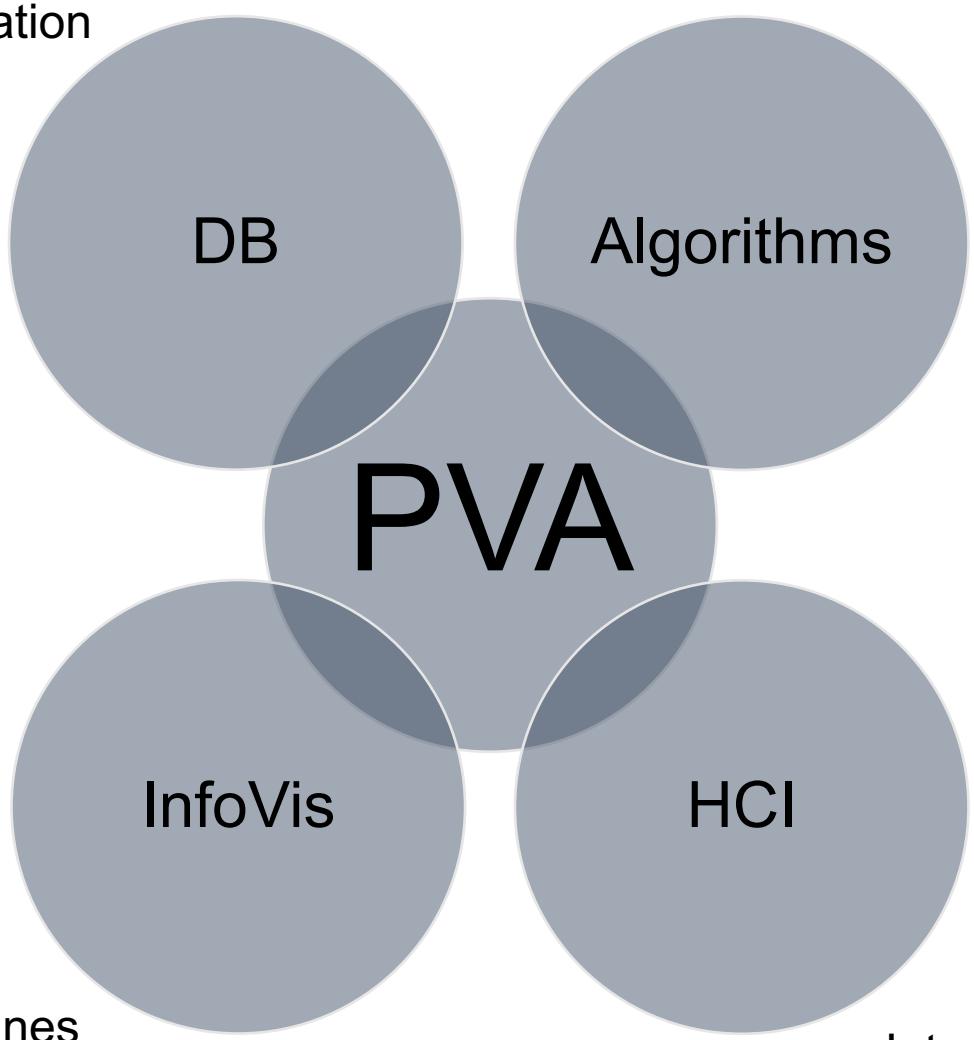
- an **early partial result,** followed by
- a **sequence of partial results,** till
- a **final result**



FIELDS (HIDDENLY) SHARING THE SAME PROBLEM

On line DB aggregation
...

Online algorithms
...



Modeling Viz pipelines
...

Interaction techniques
...

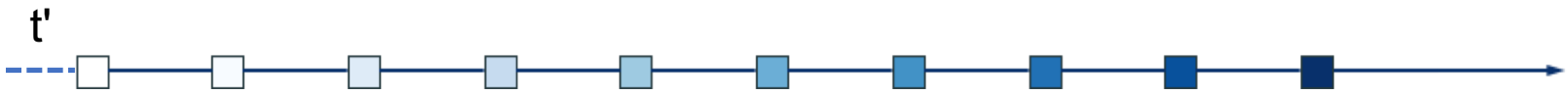
WHAT DO THEY HAVE ALL IN COMMON?

- An unbearable delay t

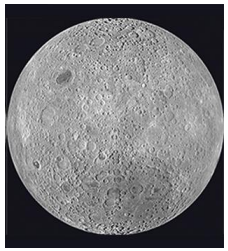
unbearable depends on the the task

Design goal	Unbearable delay
Perceptual update	t > 0.1s
Immediate response	t > 1s
Unit task completion	t > 20-30 s
Tractable analysis	t > days
Within analyst life	t > hundreds of years

- A **problem** originating the delay
- A **technical solution** tricking the problem and producing a **sequence of results** with a bearable t' t'<<t



- Some **issues**



CAUSES & TECHNICAL SOLUTIONS

Main Causes

- Large data sets
- Computational complexity
- Slow connections
- A malign  combination of two or more of them!☹

Main Technical solutions

- Split the **data** in **chunks** and process each chunk individually
  ⇨ Partial results of increasing *completeness*

- Split the **analytical algorithm** into computational **steps** that iteratively refine previous results
  ⇨ Partial results of increasing *precision*

- Or both of them!

Main issues

MVA MVA produces just ONE and precise result

PVA

How much can I trust this result ?

How far am I from the final result?

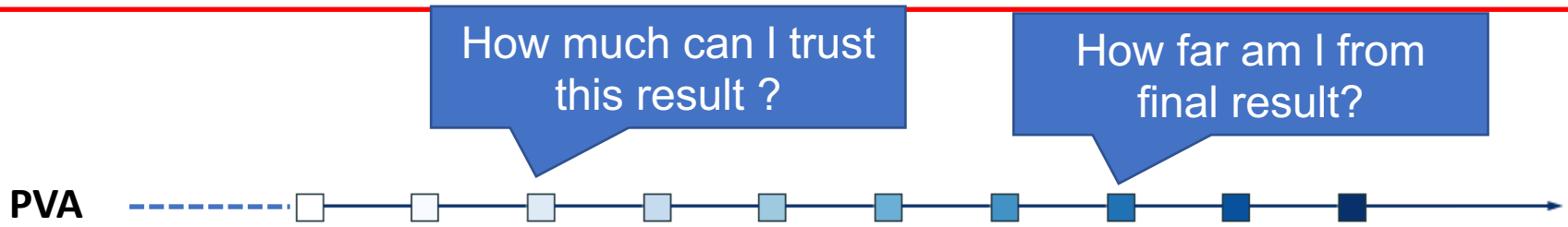
Approximation and errors introduced by either data chunking or process chunking

Early, but useful?

Fluctuation could confuse the user

Variability of the convergence and usefulness of the first result

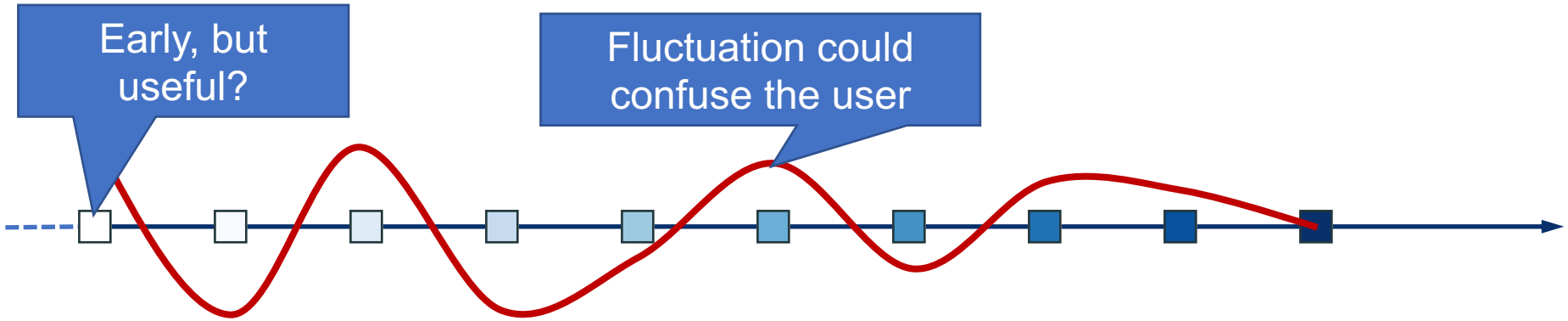
MITIGATING ISSUE 1



Approximation and errors introduced by either data chunking or process chunking

- Use metrics to measure/estimate approximation and errors
- Make explicit them to the user, helping him to better judge results usefulness and convergence

MITIGATING ISSUE 2



Variability of the convergence and usefulness of the first result

- Design the production of the first result in a careful way
  - Respecting the t' constraints
  - Working on highly relevant data
  - Preserving the structure of the viz
  - Allowing full interaction on it
- Minimize fluctuations impact
  - Use animation
  - Maintain the context
  - Slow down and avoid non relevant changes
  - Block user interaction while updating the result

An example: The combinatorial Telecom case

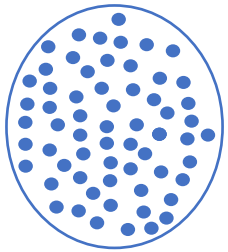
Mobile consumers switch from one operator to another based on the best fares: **advertising campaigns** are very important

The TIM goal is to locate a campaign target of the 110 Italian **provinces,** e.g., the top10 that maximize an objective function

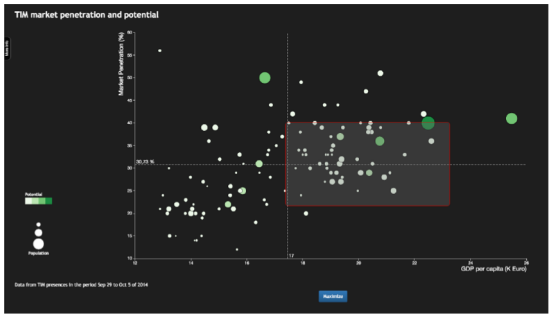
The objective function uses stored data (e.g., traffic between provinces): it requires the exploration of all combinations $\binom{n}{k}$

$\binom{110}{10}$ exploration requires hundreds of years, and the designed solution allows the analyst for interactively investigate on smaller province subsets

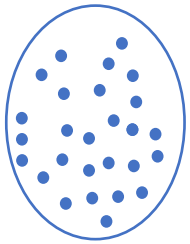
AN EXPLORATIVE APPROACH



110 provinces

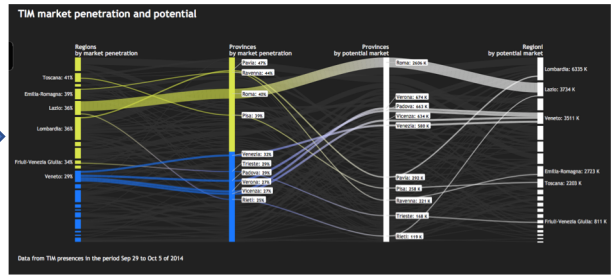


interactive scatterplot



user selection (e.g., 40 provinces)

top10



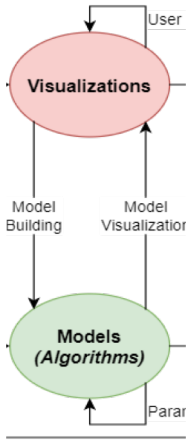
Sankey plot

like it!

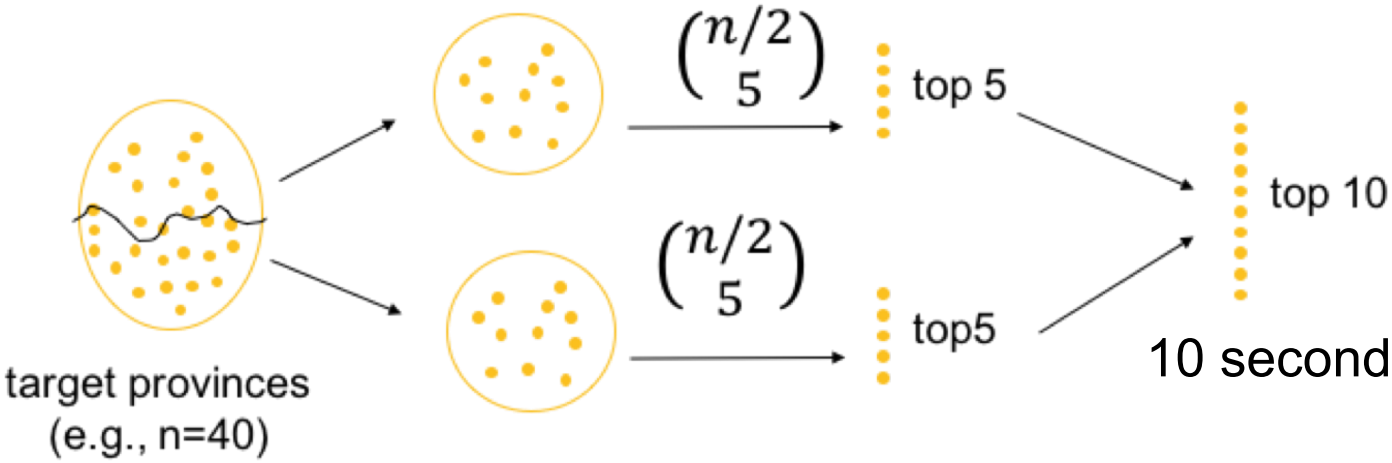
do not like it!

done!

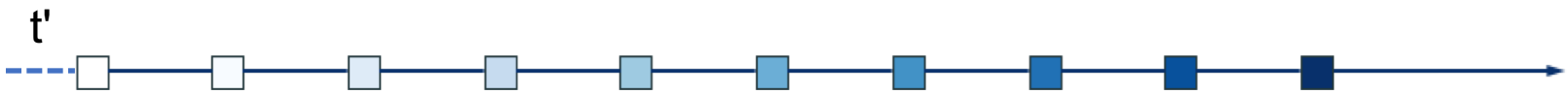
THE PROBLEM & THE SOLUTION



The problem: $\binom{40}{10}$ enumeration requires about **75 hours**, making the explorative analysis not possible: t_{max}=20-30 s



10 seconds vs 75 hours



The solution: partitioning the user selection and computing the top ten as the union of local optima allow for fine tuning t'; further partial results are obtained covering the user selection with increasing size subsets

ISSUE 1 AND MITIGATION

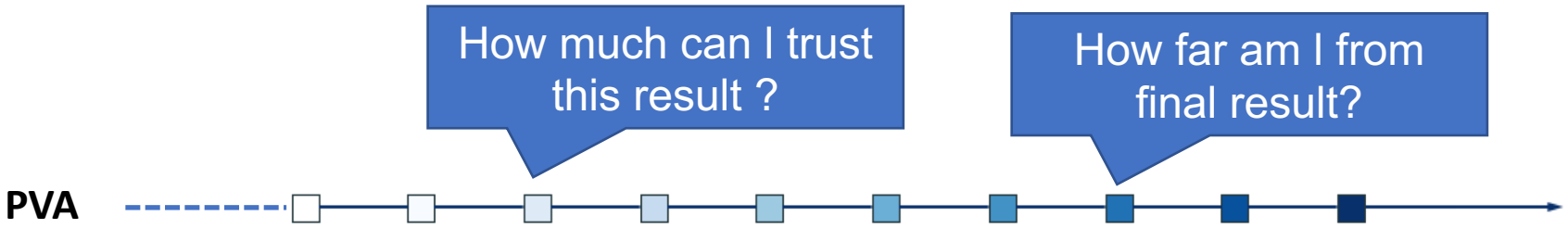
The solution introduces errors and approximation

We defined 2 metrics

$$FunctionRatio(FR) = \frac{Estimated\ function}{optimum\ value}$$

$$Top10Proportion(TP) = \frac{Estimated\ top10 \cap real\ top10}{10}$$

- The first one allows for making decision on the value of a partial result
- The second one allows for estimating the stability of the partial result
- We show them to the user, together with the Sankey plot



ISSUE 2 AND MITIGATION

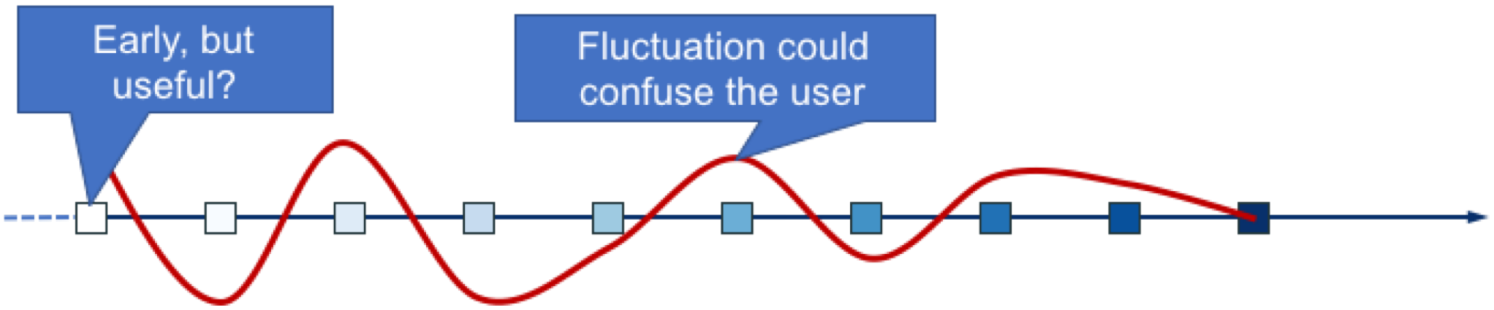
Partial results oscillate (e.g., a province disappear from the top10 and after some iterations it is back)

To **minimize fluctuation** effect:

- We show the top0 highlighting it on the 110 provinces (maintaining the context)
- We use animation along updates
- We block updates while the user interacts with the plot
- We update the viz only if there exist a significant improvement (e.g., 5%)

The **first result usefulness** is improved by:

- Using all the selected provinces for computing the first top10
- Allocating a slot time of maximum size (30 sec) and compute as many as possible combinations before presenting the user with the first result



An example of a partial result

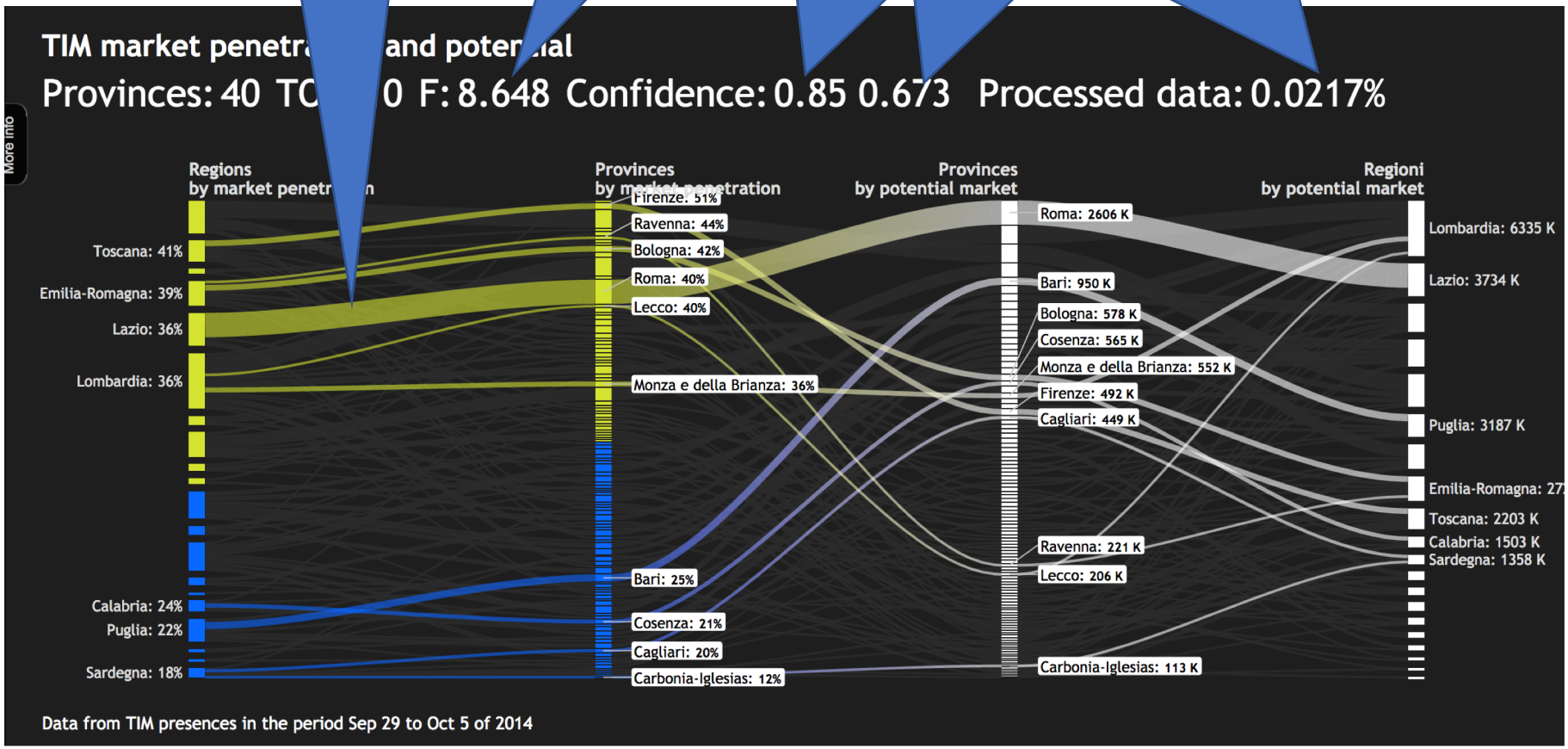
Top10 higlighted on the 110 provinces

Current c… functior…

Stal… in th… lik…

Usefulness: The actual…

Progress: The PVA application has processed 0.02% of the possible combinations

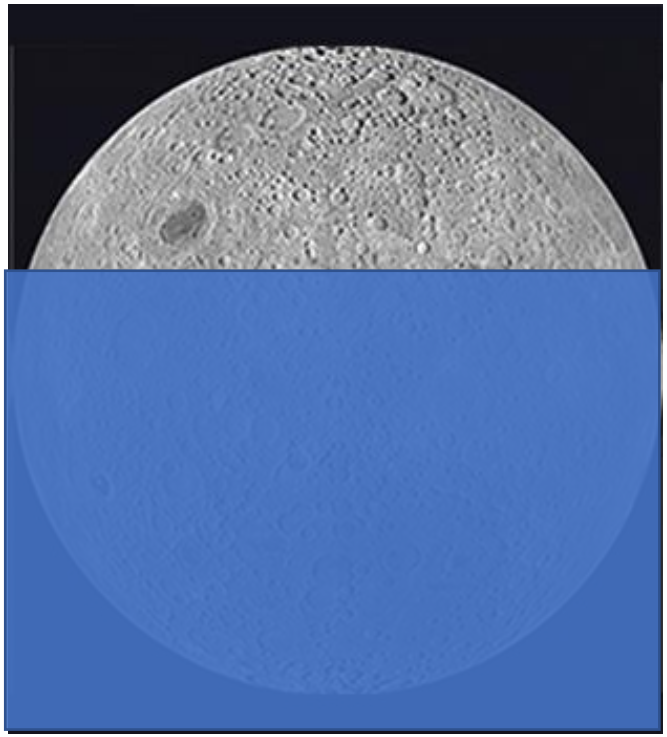


Conclusion & Open issues

- This talk just tickled the dark side of PVA, pointing out issues associated with solutions and providing an initial set of mitigations
- We foresee a lot of open issues that deserve to be dealt with:

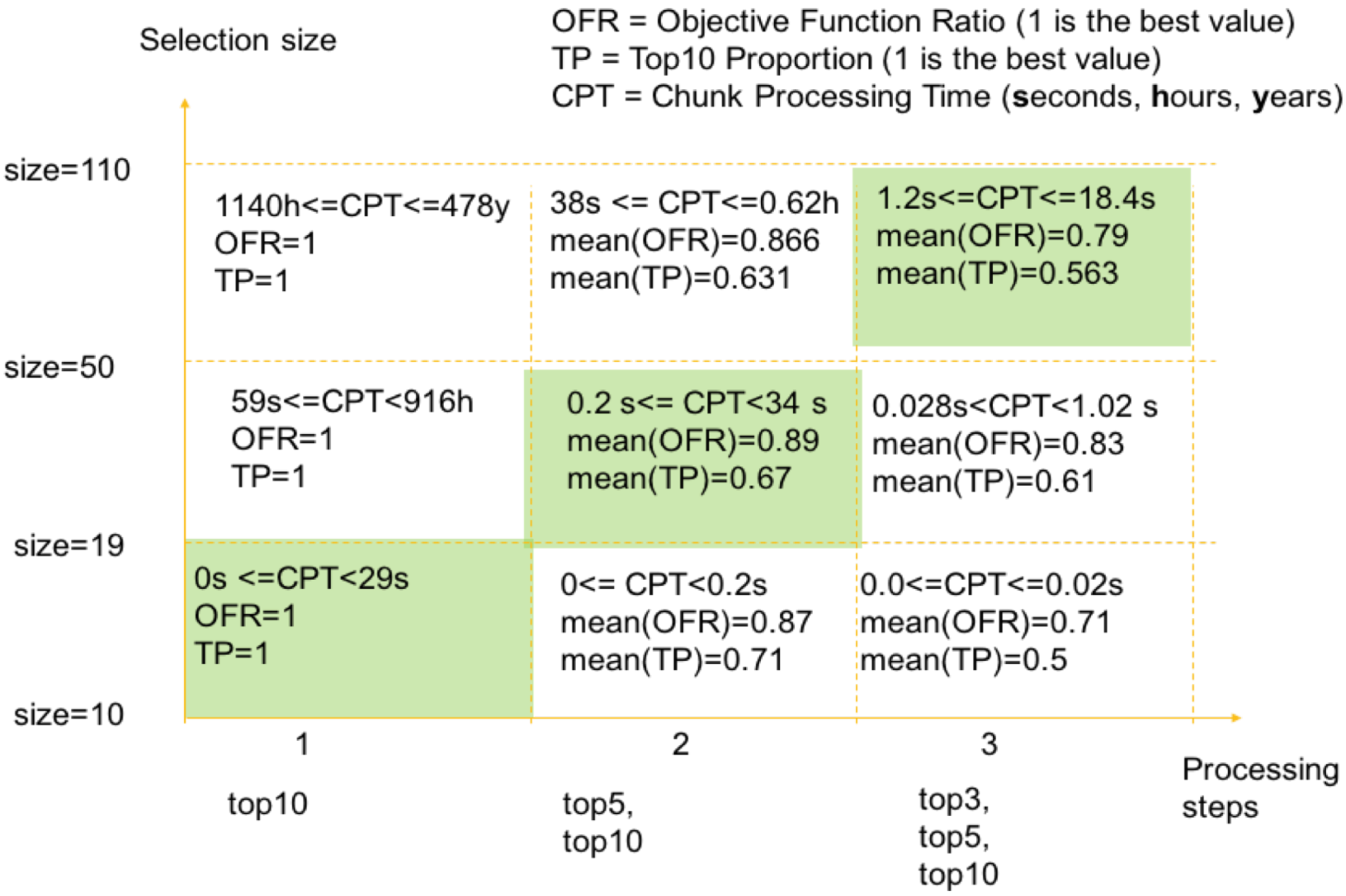
1. Provide a deeper characterization of the evolution of the partial results
2. Better characterize the issues (they are more than 2!)
3. Better characterize mitigations
4. Involve the user in the loop: support the inspection and parametrization of non-transparent computations
5. ...

So, most of the dark side is still hidden...



Q U E S T I O N S ?

ANSWER 1


Selection size
OFR = Objective Function Ratio (1 is the best value)
TP = Top10 Proportion (1 is the best value)
CPT = Chunk Processing Time (seconds, hours, years)
size=110
size=50
size=19
size=10
1140h<=CPT<=478y
OFR=1
TP=1
38s <= CPT<=0.62h
mean(OFR)=0.866
mean(TP)=0.631
1.2s<=CPT<=18.4s
mean(OFR)=0.79
mean(TP)=0.563
59s<=CPT<916h
OFR=1
TP=1
0.2 s<= CPT<34 s
mean(OFR)=0.89
mean(TP)=0.67
0.028s<CPT<1.02 s
mean(OFR)=0.83
mean(TP)=0.61
0s <=CPT<29s
OFR=1
TP=1
0<= CPT<0.2s
mean(OFR)=0.87
mean(TP)=0.71
0.0<=CPT<=0.02s
mean(OFR)=0.71
mean(TP)=0.5
1
2
3
Processing steps
top10
top5, top10
top3, top5, top10

References

[2] C. D. Stolper, A. Perer, and D. Gotz, “Progressive visual analytics: User-driven visual exploration of in-progress analytics,” IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 1653-1662, 2014.

[3] T. Muhlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit, “Opening the black box: Strategies for increased user involvement in existing algorithm implementations,” IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1643-1652, 2014.

[5] J. M. Hellerstein, P. J. Haas, and H. J. Wang, “Online aggregation,” SIGMOD Rec., vol. 26, pp. 171{182, June 1997.

[7] J. Fekete and R. Primet, “Progressive analytics: A computation paradigm for exploratory data analysis,” CoRR, vol. abs/1607.05162, 2016.

[8] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, “Interactions with big data analytics,” ACM, May 2012.

[9] R. Rosenbaum, J. Zhi, and B. Hamann, “Progressive parallel coordinates,” in 2012 IEEE Pacic Visualization Symposium, pp. 25-32, Feb 2012.

[11] S. K. Badam, N. Elmqvist, and J.-D. Fekete, “Steering the craft: Ui elements and visualizations for supporting progressive visual analytics,” Computer Graphics Forum, vol. 36, no. 3, pp. 491-502, 2017.

[13] H.-J. Schulz, M. Angelini, G. Santucci, and H. Schumann, “An enhanced visualization process model for incremental visualization,” IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 7, pp. 1830-1842, 2016.

[14] N. Pezzotti, B. P. F. Lelieveldt, L. v. d. Maaten, T. Hllt, E. Eisemann, and A. Vilanova, “Approximated and user steerable tsne for progressive visual analytics,” IEEE Transactions on Visualization and Computer Graphics, vol. 23, pp. 1739-1752, July 2017.

[15] C. Turkay, E. Kaya, S. Balcisoy, and H. Hauser, “Designing progressive and interactive analytics processes for high-dimensional data analysis,” IEEE Transactions on Visualization and Computer Graphics, vol. 23, pp. 131-140, Jan 2017.