ICCUB School
Machine Learning
and Data Mining in Physics

October 20

MACHINE LEARNING: A VIEW FROM THE TRENCHES

Dr. Vicens Gaitan

Grupo AIA

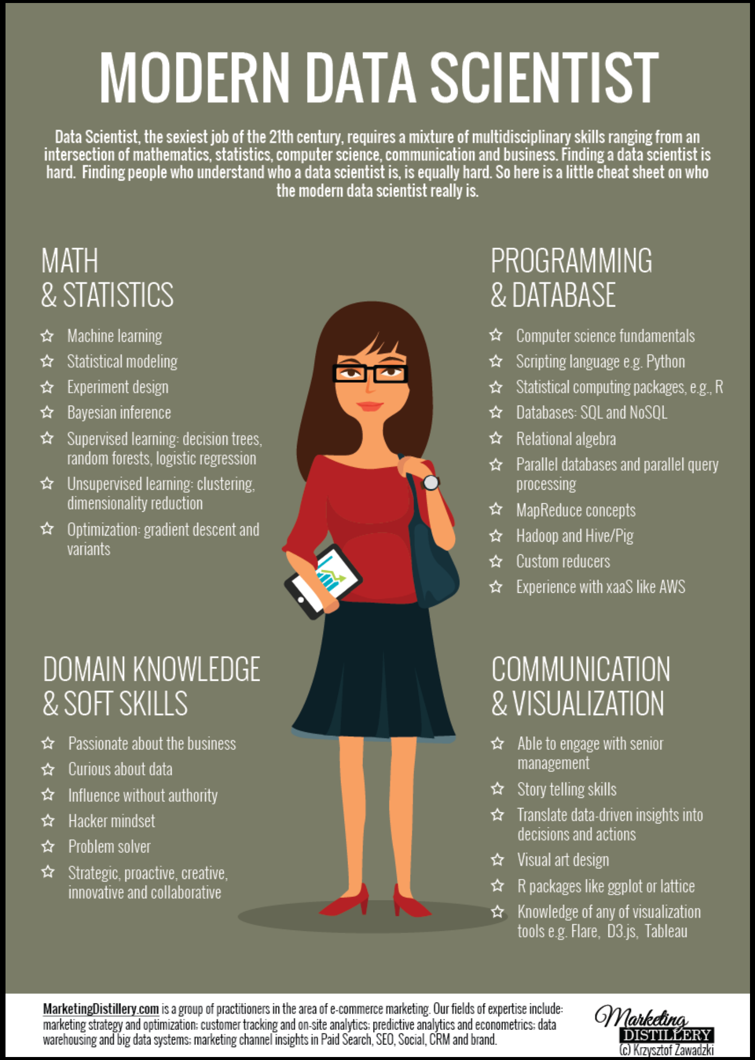
AGENDA

- Data Science
- Model driven or Data driven
- Some examples on Basic Science Industrialization
- ML + Big Data : The perfect couple
- Closing the loop: Machine Learning in HEP
- Conclusions

DATA SCIENCE

- There is *Science* without *Data* ?...
- Science: Model from reality, predictive and falsifiable
- Why now? Huge amounts of data + computing power

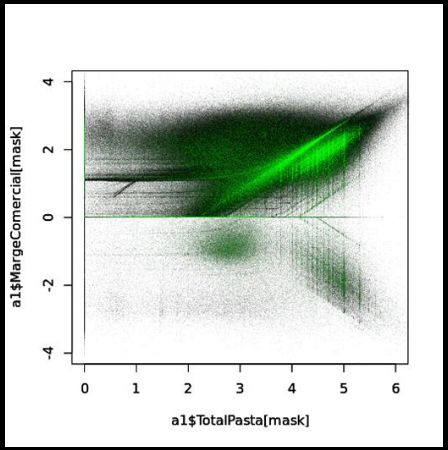
DATA SCIENTIST

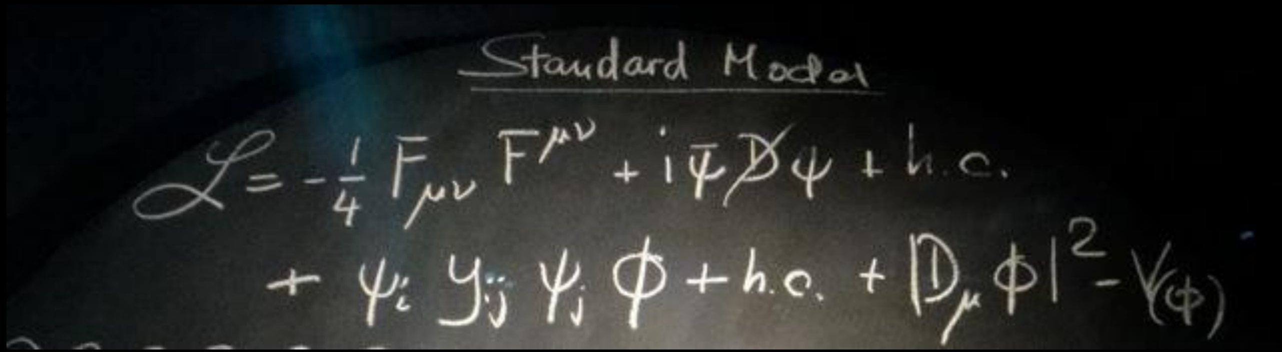


- “Leonardo” in the current days:
  - Machine learning
  - Data munging (Tb of them)
  - Fast programmer
  - Hacker
  - Modeler
  - Good communicator
  - Artist ;)

MODEL DRIVEN O DATA DRIVEN

- Here is where physicist can make a difference…
- The best model is a huge amount of data,… but correlation dose'ny imply causality.
- …a realistic analytical modes is better than thousand deep neural networks





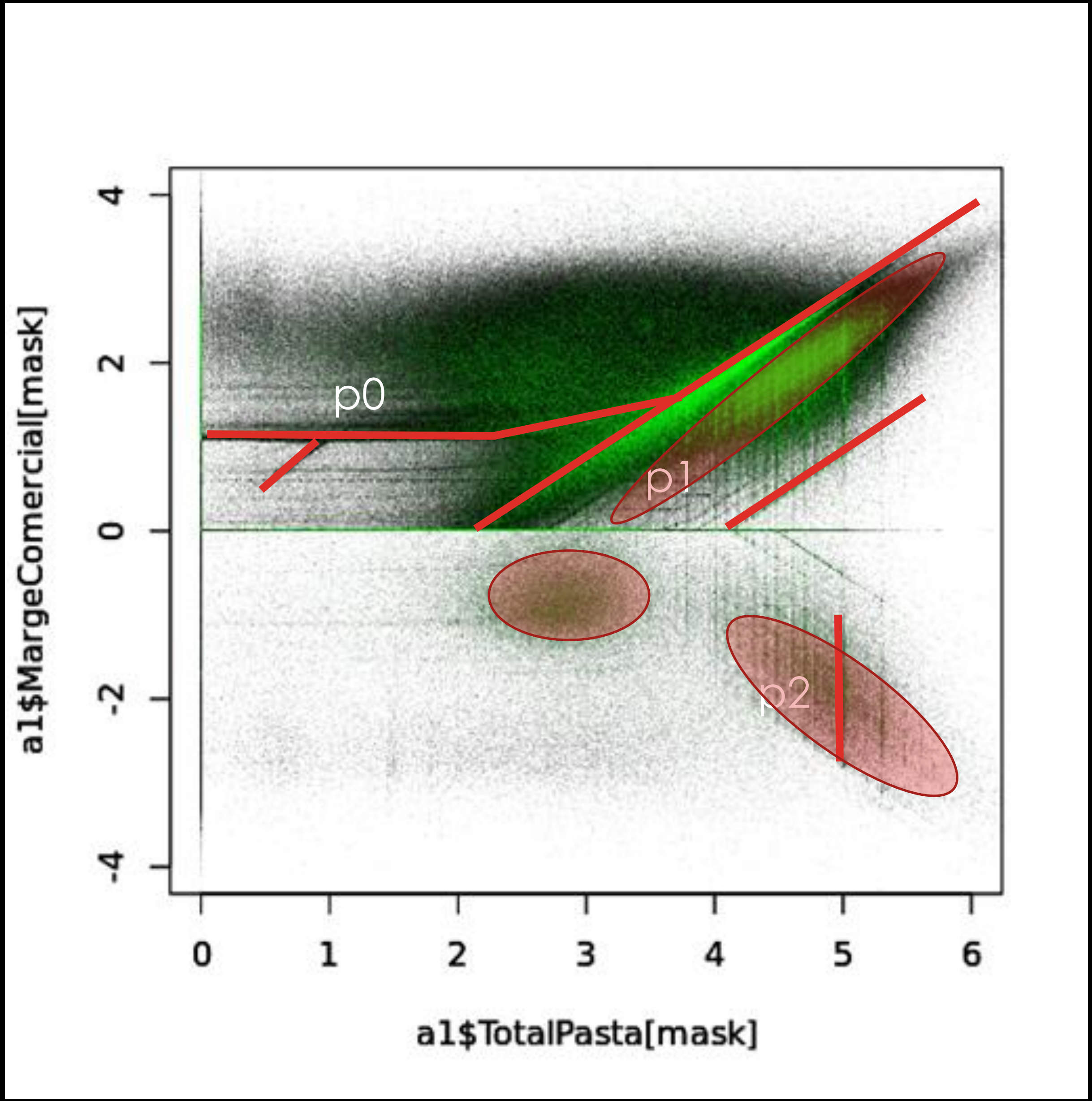
Data+Models: explosive mixture

Data+Machine Learning= Quantitative

But..

Data+Model= Quantitative+Causal+Predictive

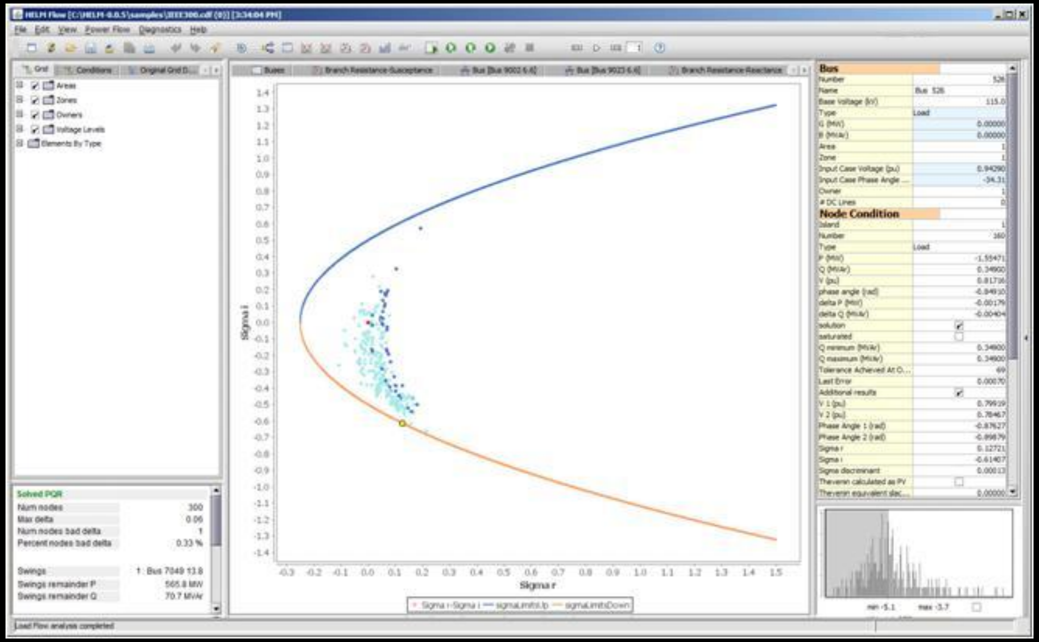
Be aware: use machine learning when analytical models are not enough



EXEMPLES ON BASIC SCIENCE INDUSTRIALIZATION

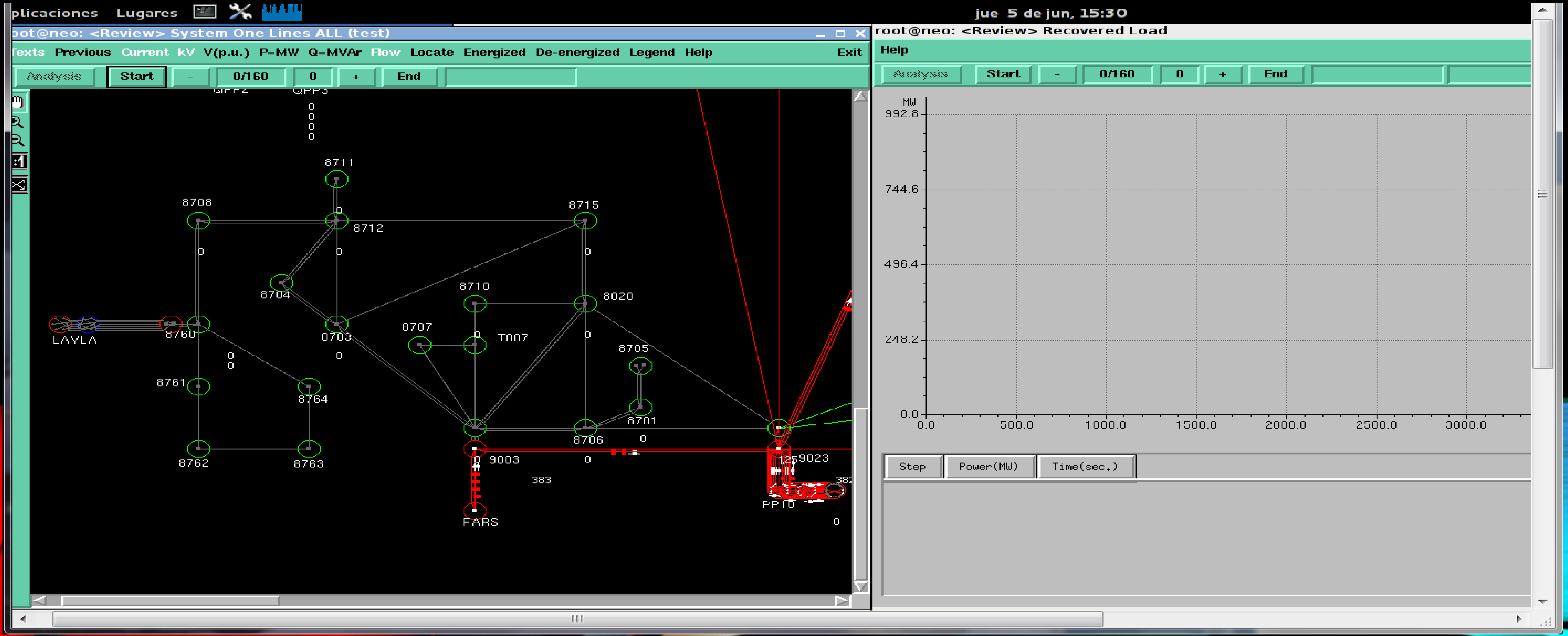
- **Holomorphic embedding**: using some properties of algebraic corves on the complex plane, allow to solve nonlilear multivalued equations non iteratively

- Power systems (Load Flow equations).
- Si*/Vi*= Sum Yij Vj
- Automatic action generation
- Working with NASA in autonomous systems



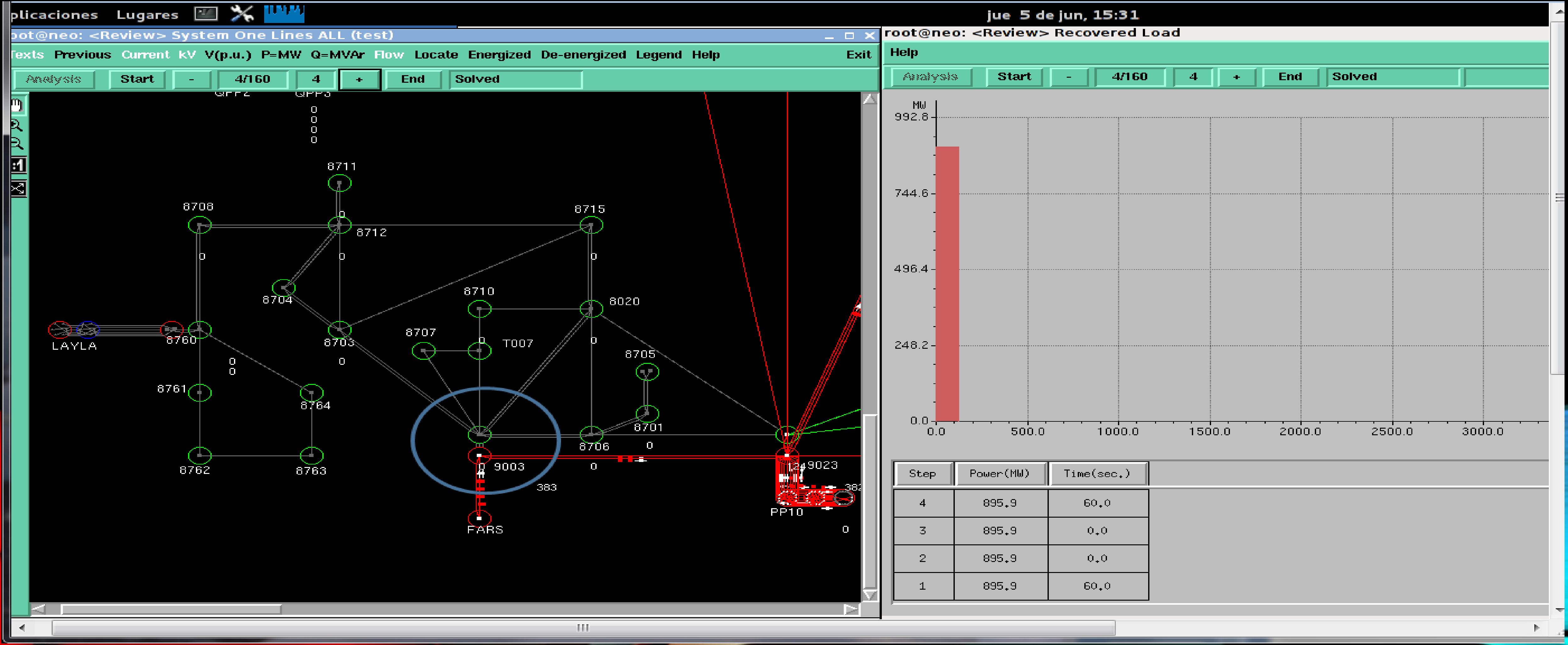
AGORA Restoration

AGORA can monitor in the simulation environment of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



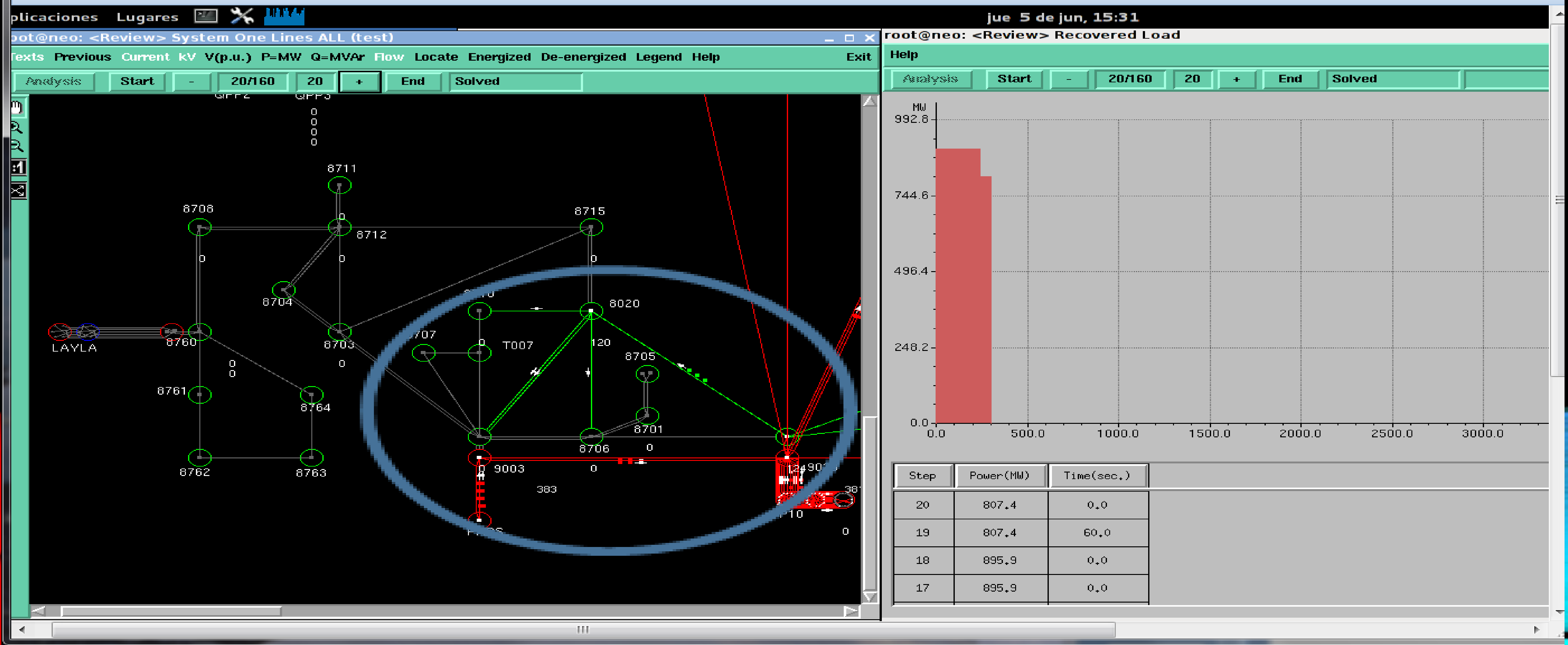
AGORA Restoration. Monitoring as Exp.Recovered Load.

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



AGORA Restoration. Propagate from strong Bus

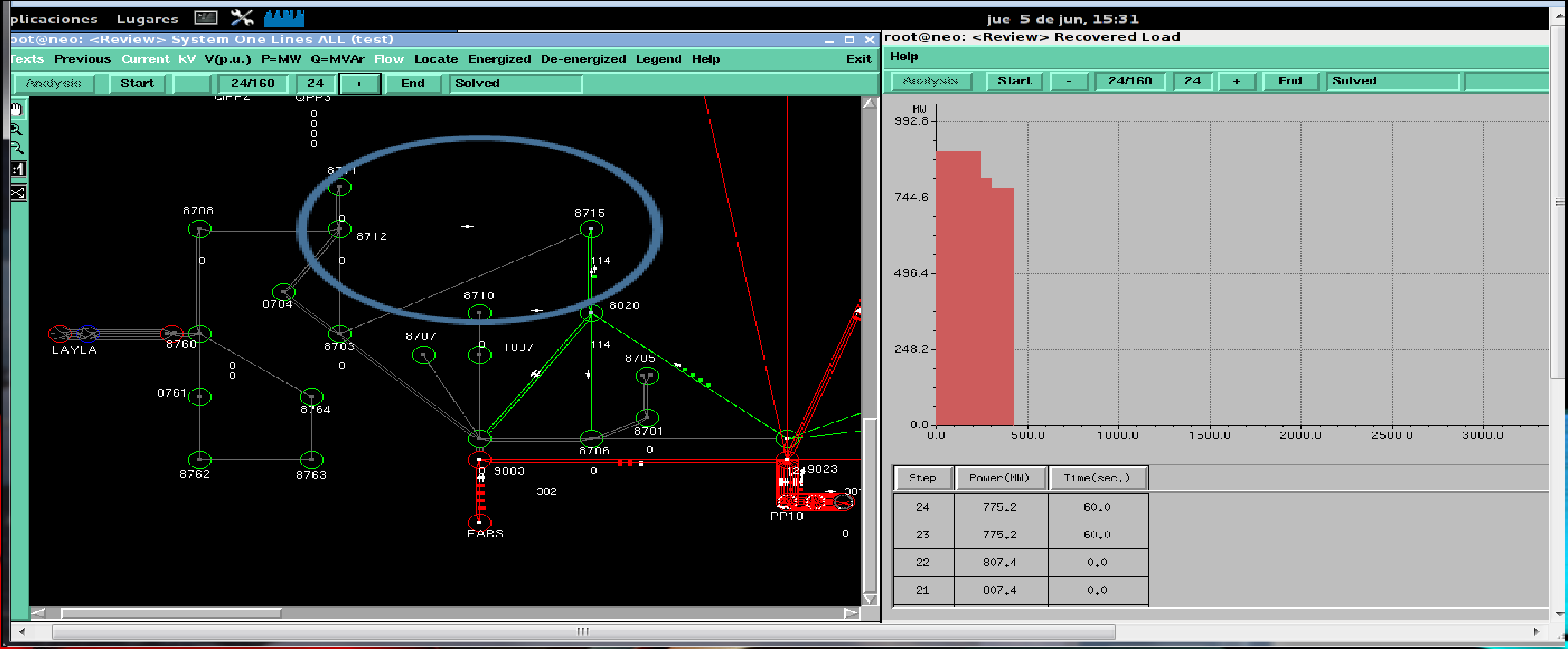
AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



Step	Power(MW)	Time(sec.)
20	807.4	0.0
19	807.4	60.0
18	895.9	0.0
17	895.9	0.0

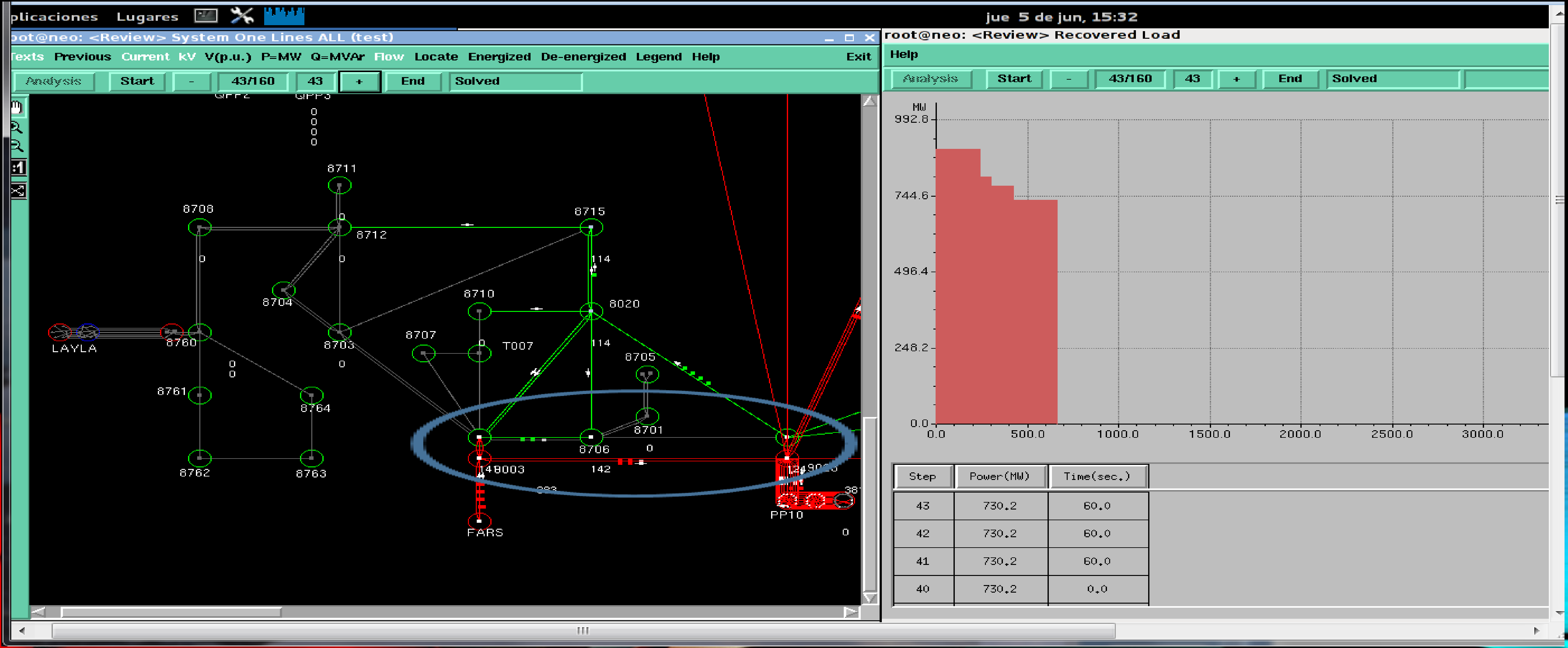
AGORA Restoration

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



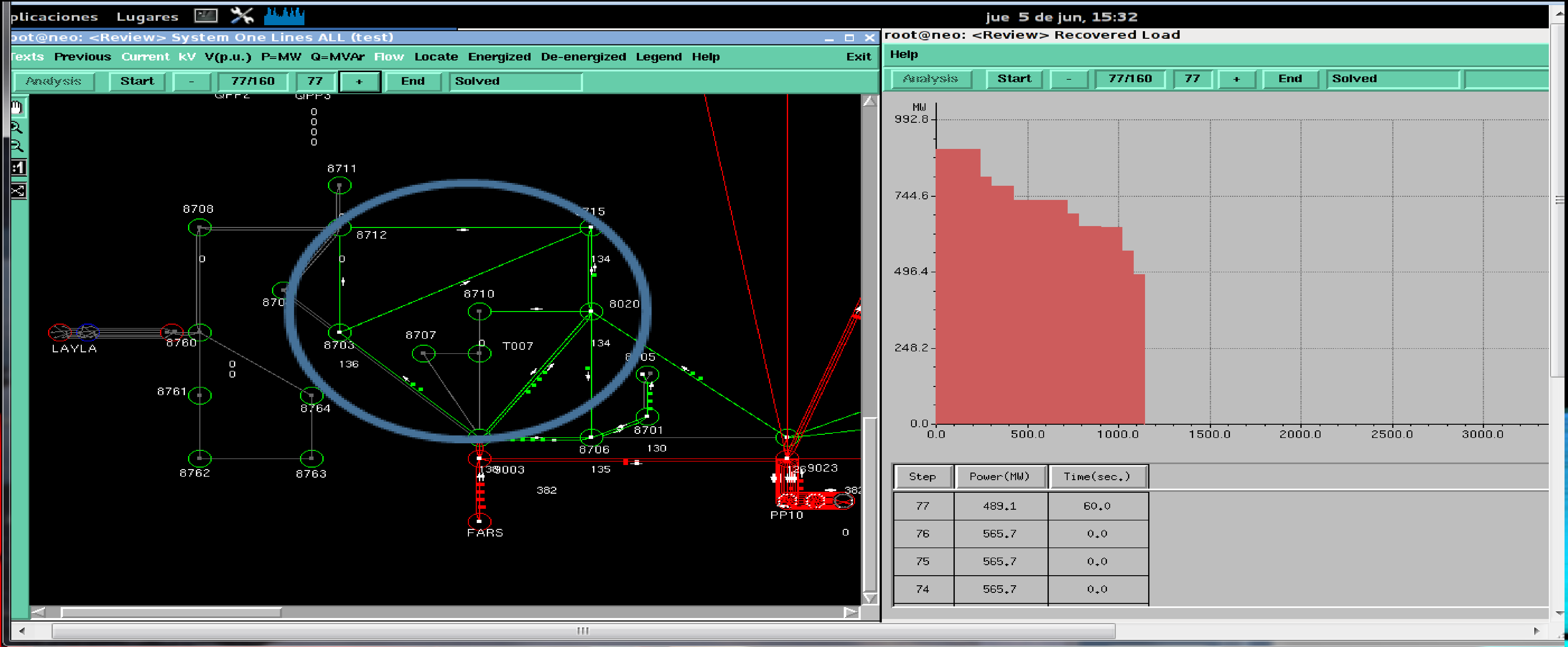
AGORA Restoration. Mesh the network

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



AGORA Restoration

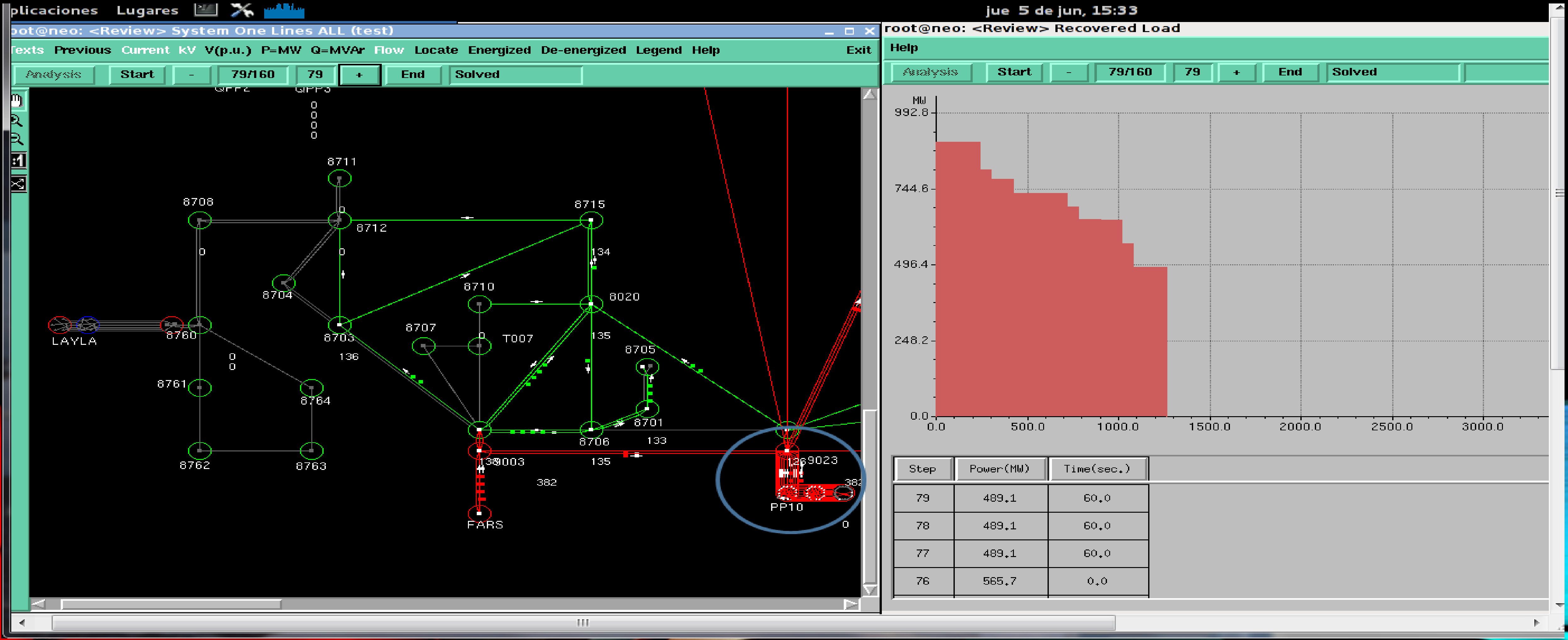
AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



Step	Power(MW)	Time(sec.)
77	489.1	60.0
76	565.7	0.0
75	565.7	0.0
74	565.7	0.0

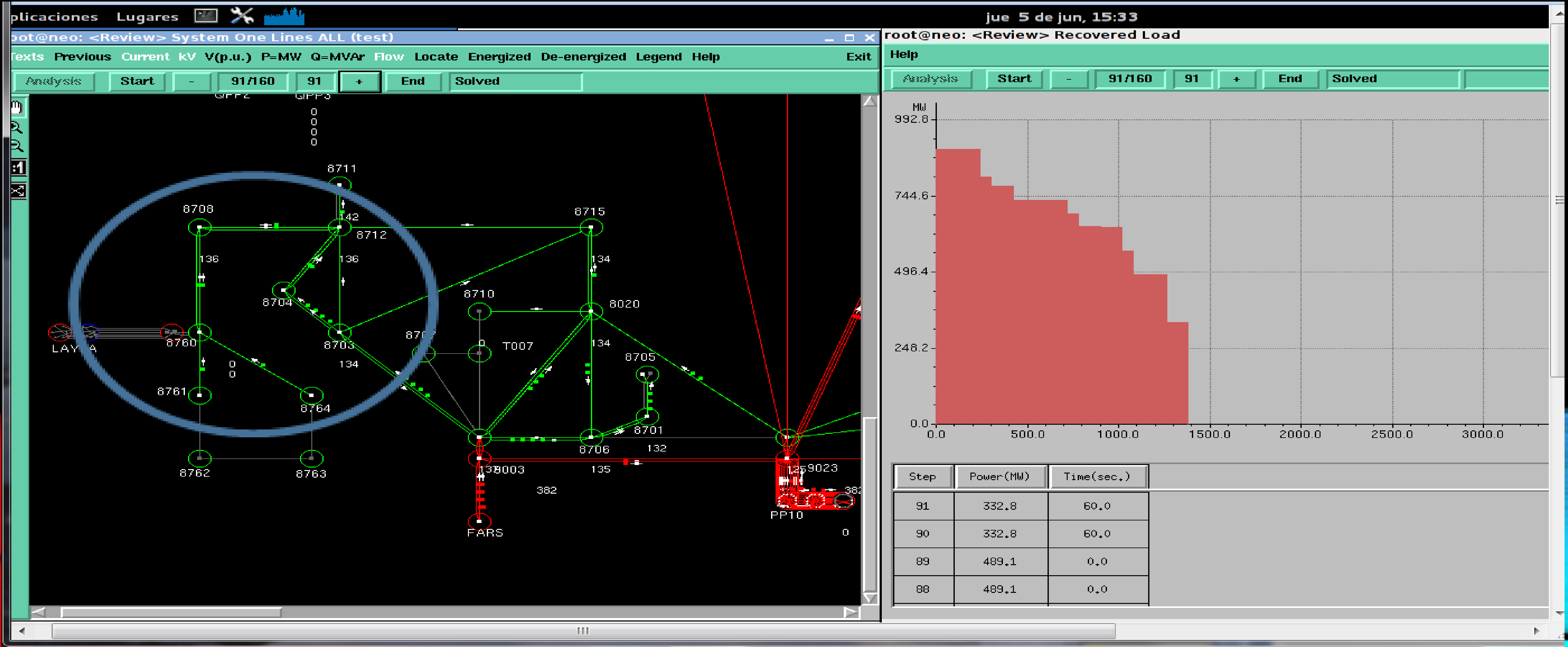
AGORA Restoration. Synchronizes generation

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



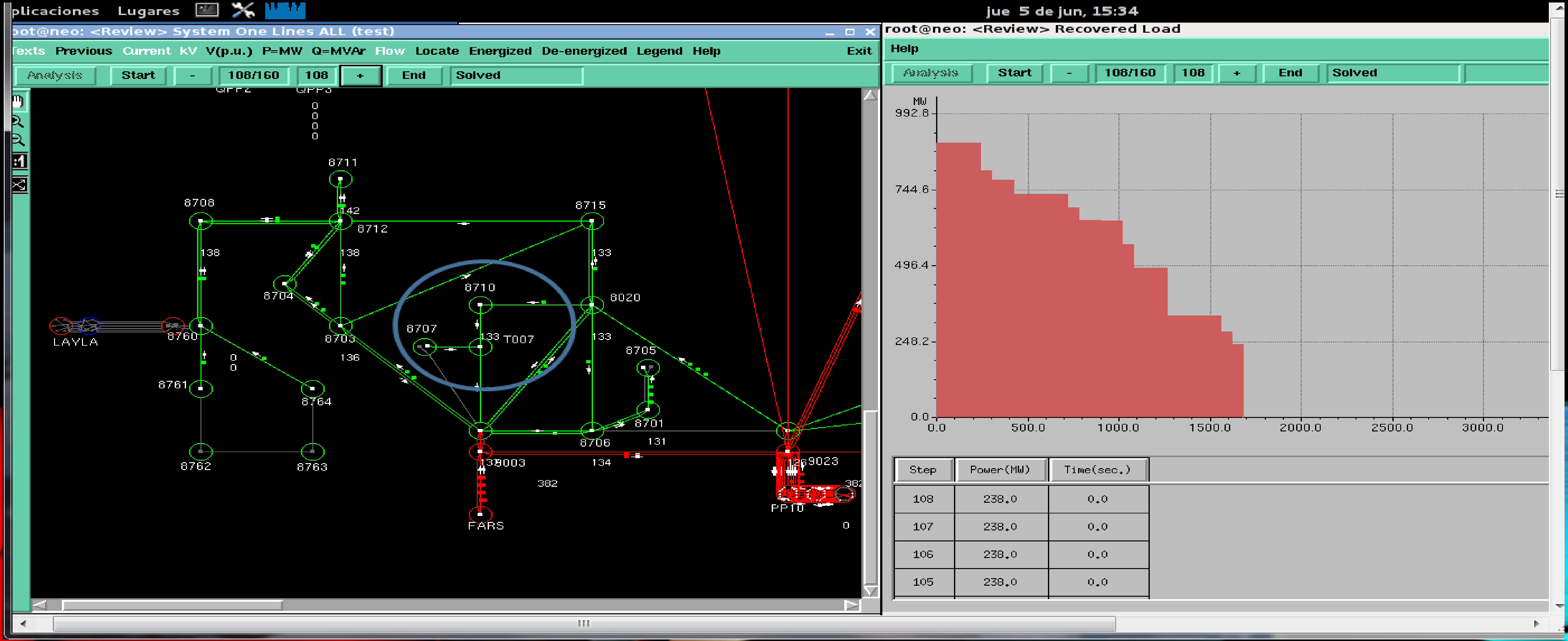
AGORA Restoration. Recovering Larger Loads

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



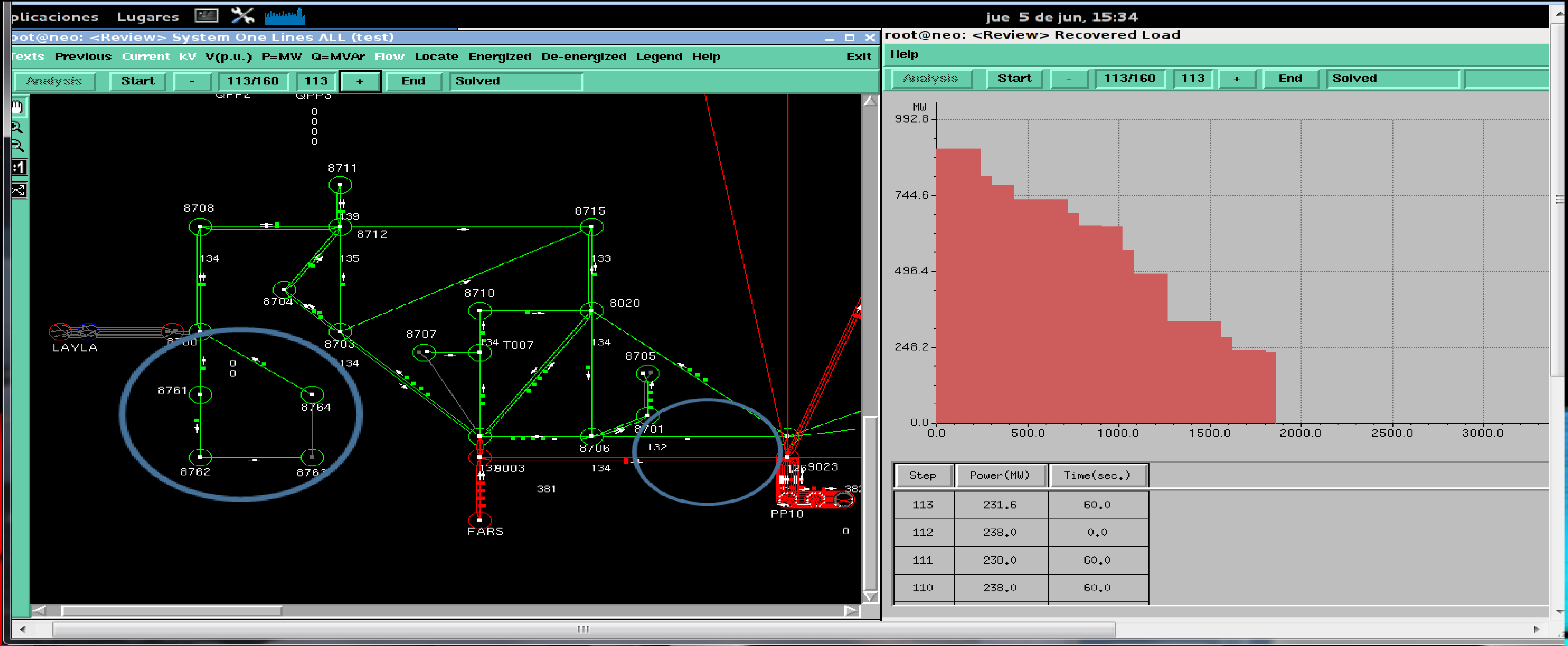
AGORA Restoration. Stabilize Flows & Avoid Overloads

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



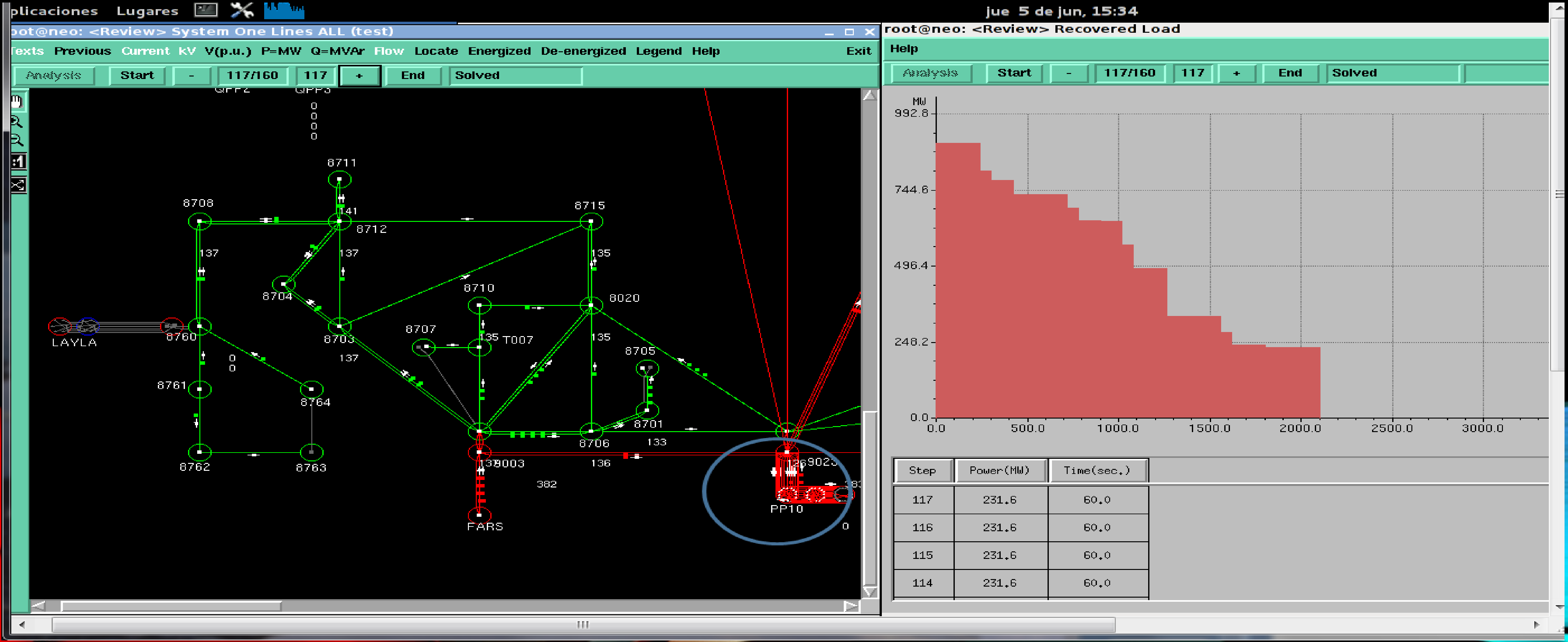
AGORA Restoration

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



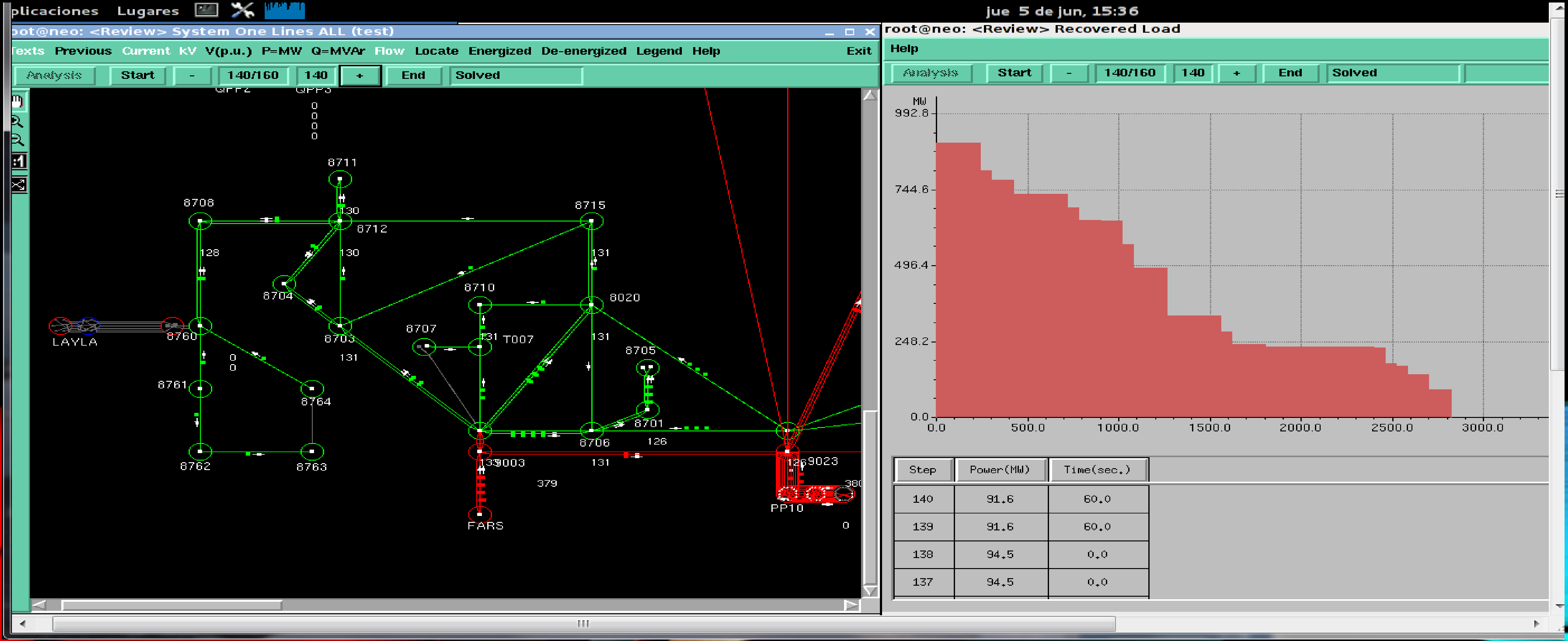
AGORA Restoration

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



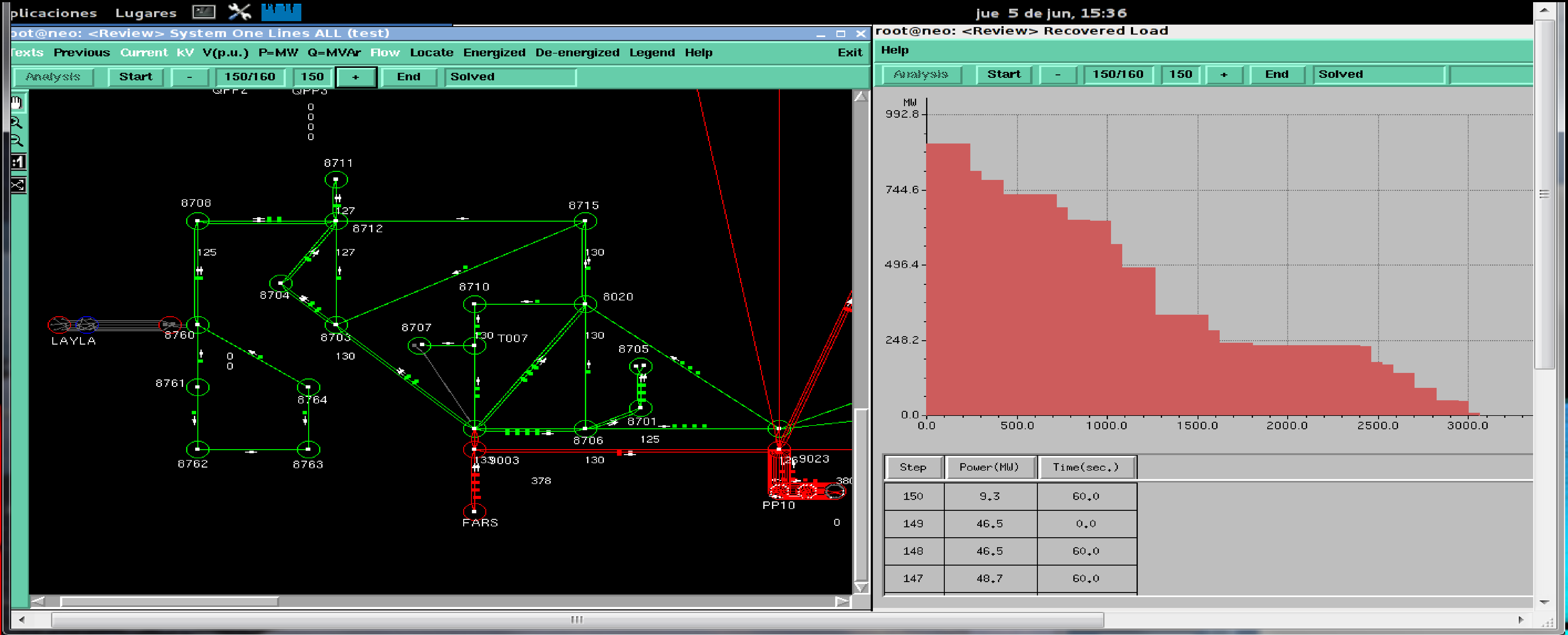
AGORA Restoration. Recovering Smaller loads

AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



AGORA Restoration

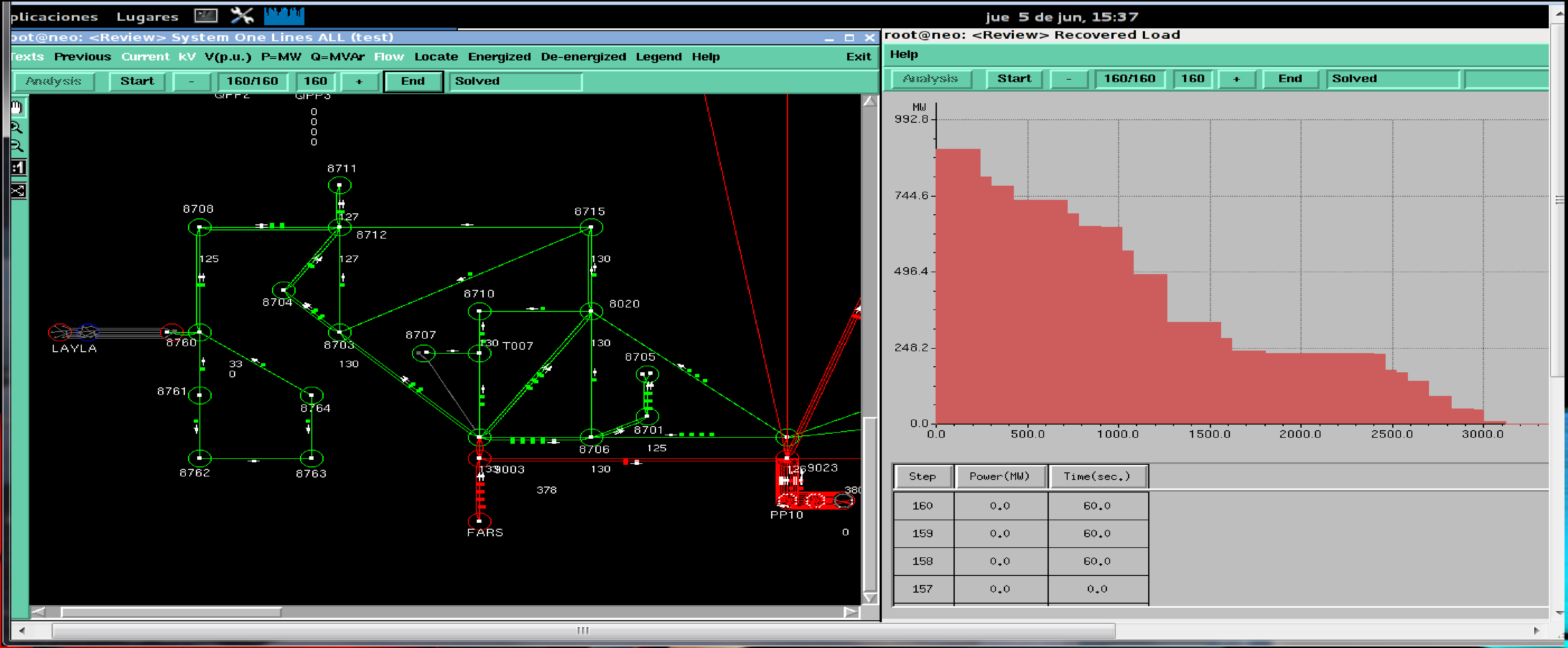
AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



Step	Power(MW)	Time(sec.)
150	9.3	60.0
149	46.5	0.0
148	46.5	60.0
147	48.7	60.0

AGORA Restoration

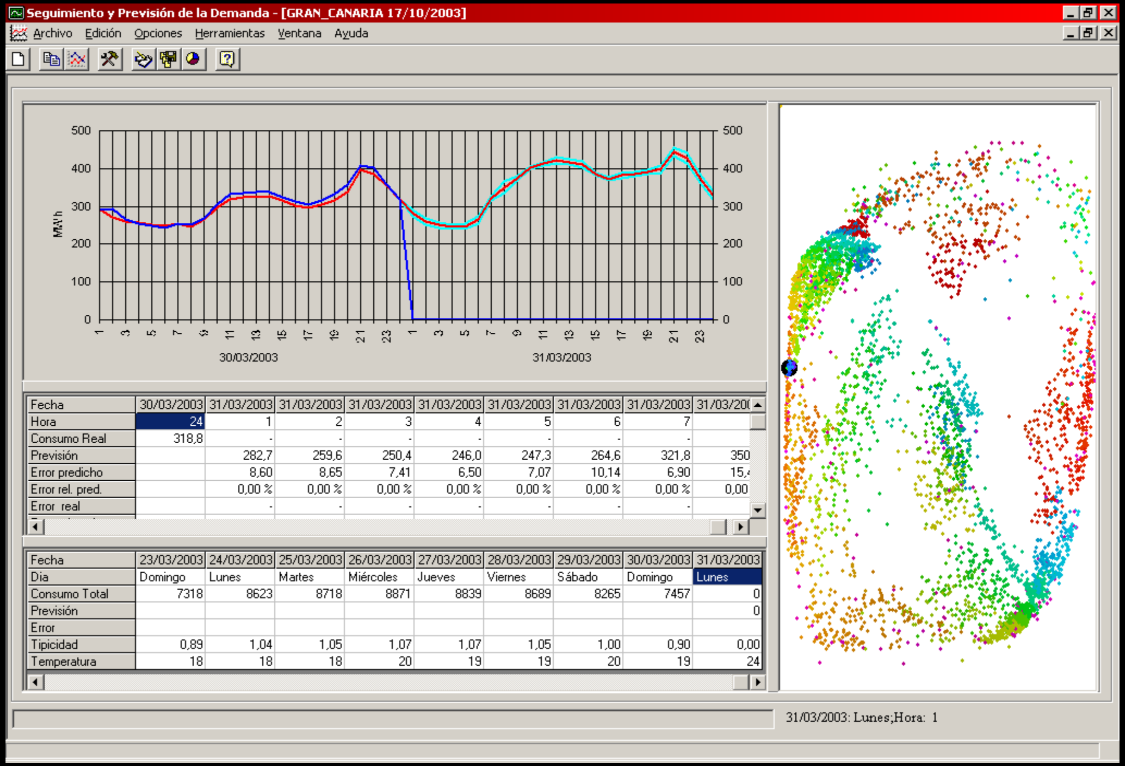
AGORA can monitor the simulation of the restoration plan, step by step or blocks of local objectives. We will show evolution of the plan for the disturbance.



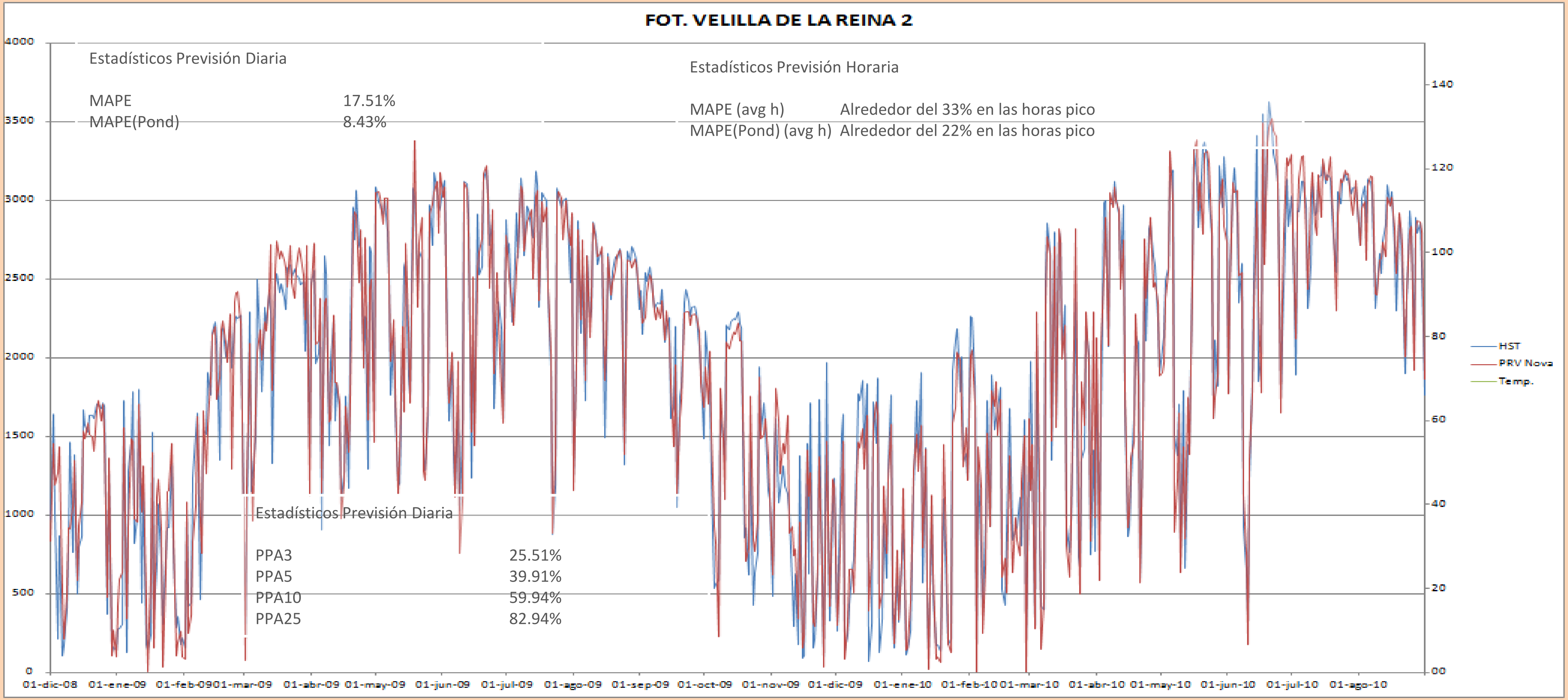
EXEMPLES ON BASIC SCIENCE INDUSTRIALIZATION

- Phase Space: load forecasting
- Gas, Electricity, ATM cash, phone calls…

- Embedding of the time dynamics on a low dimensional space



PHOTOVOLTAIC: HOURLY FORECASTING



2009

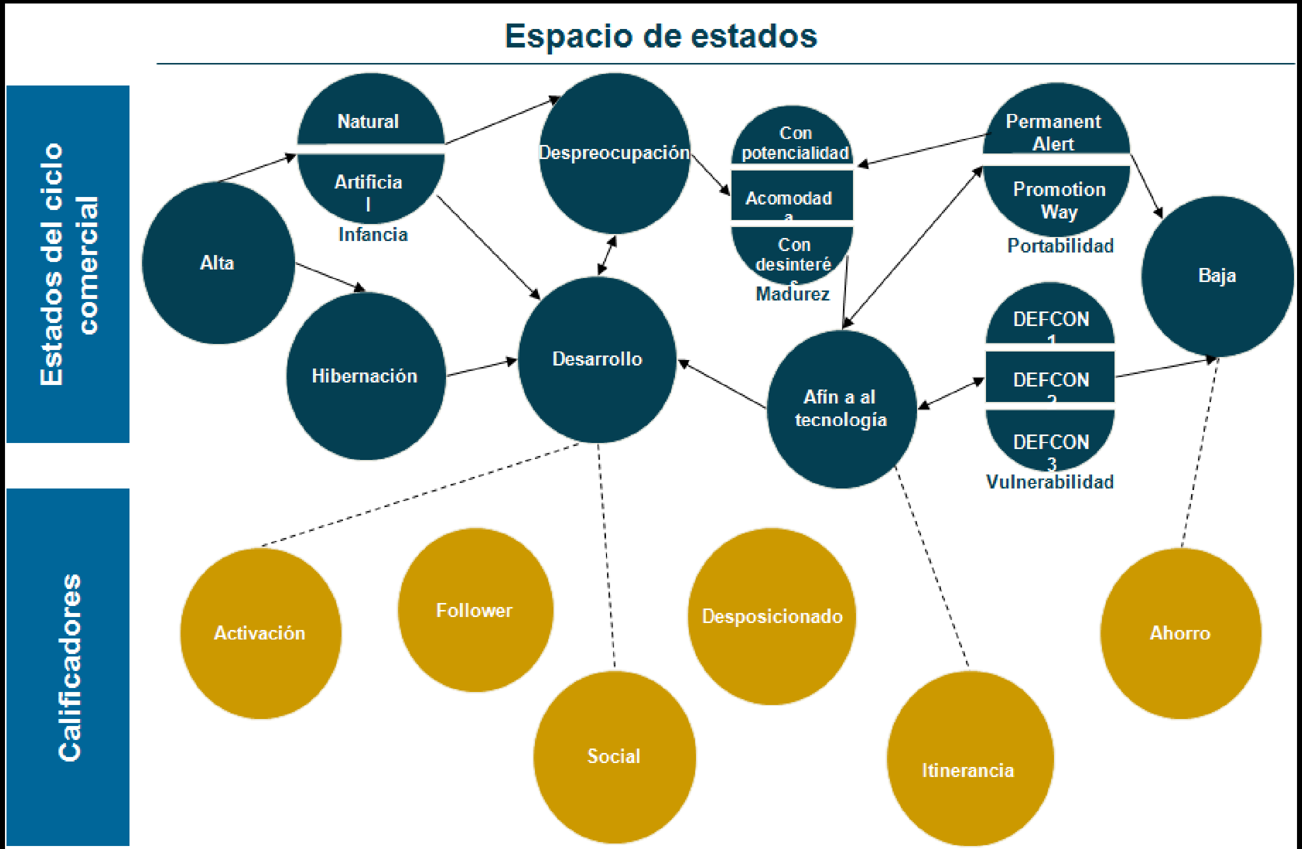
EXEMPLES ON BASIC SCIENCE INDUSTRIALIZATION

Customer life cycle representes in a State Space

Markov Chains

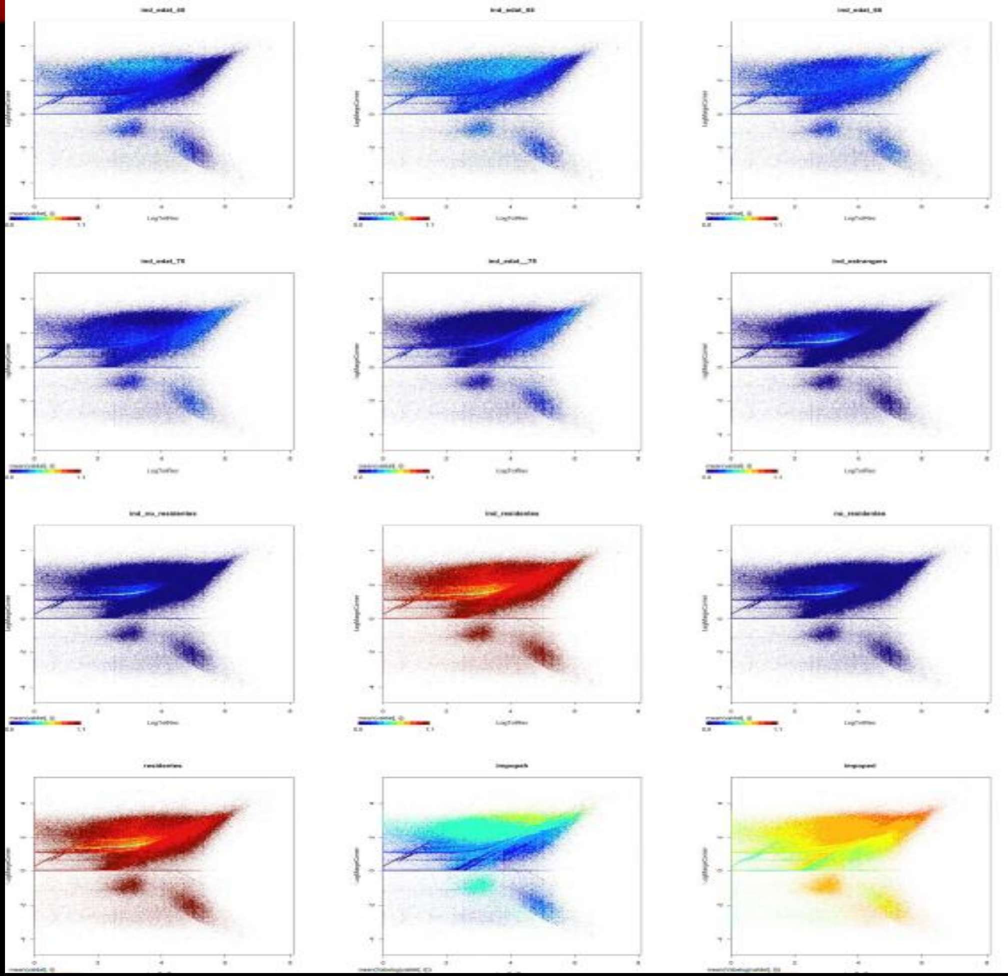
-> Stochastic
-> Quantum

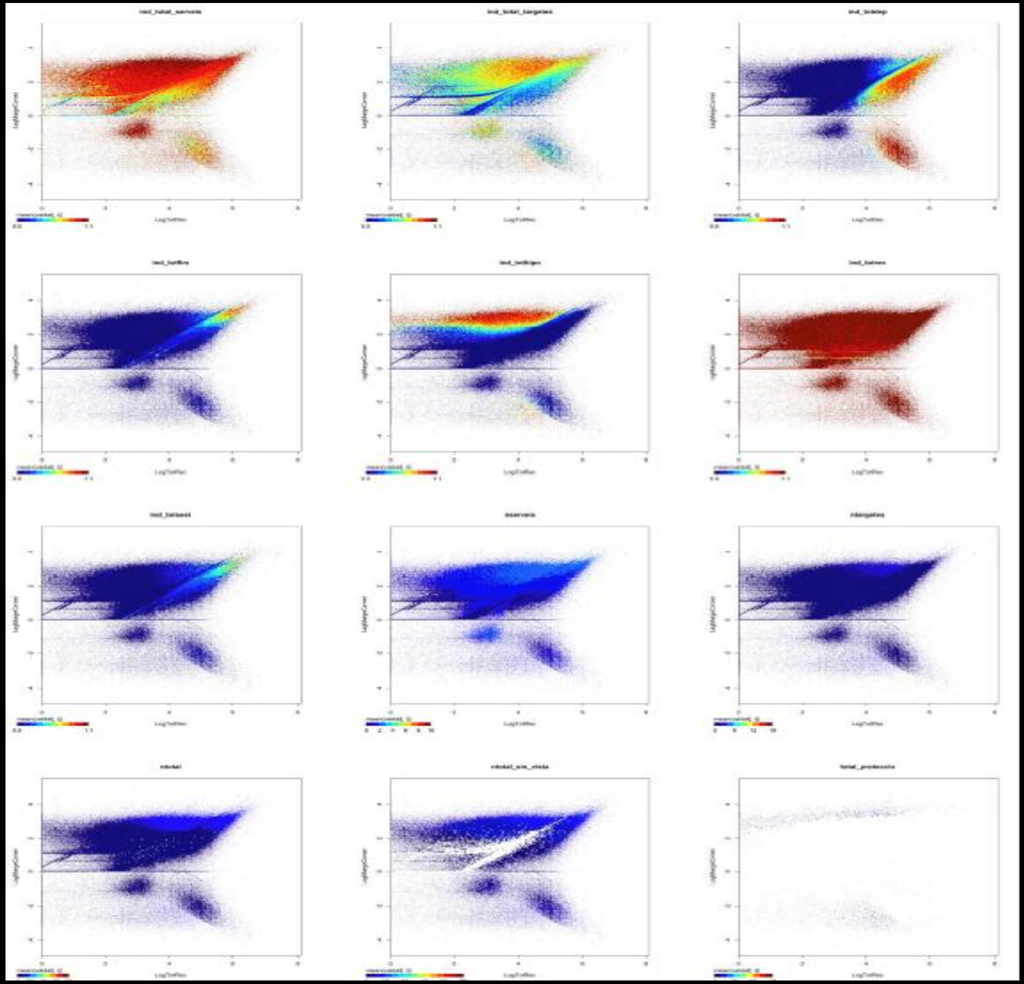
Customer evolution forecasting



ML +BIG DATA: The perfect couple

State SPECTROSCOPY





5 - Examples. Propensity Models for Marketing

Goals

- Developing Purchasing Propensity Models of products.
  - Calculate purchasing propensity of specific financial products for each customer.
  - Quick design and development of models, adapted to the campaign's context.

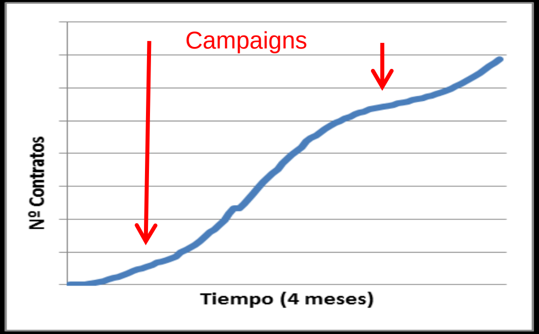
Methods

- **Conceptualization**: Customer characterization according to its commercial relationship with the Financial Institution.
- **Prediction**: Purchasing Propensity Modeling, based on customers previous purchasing identity information and knowledge..

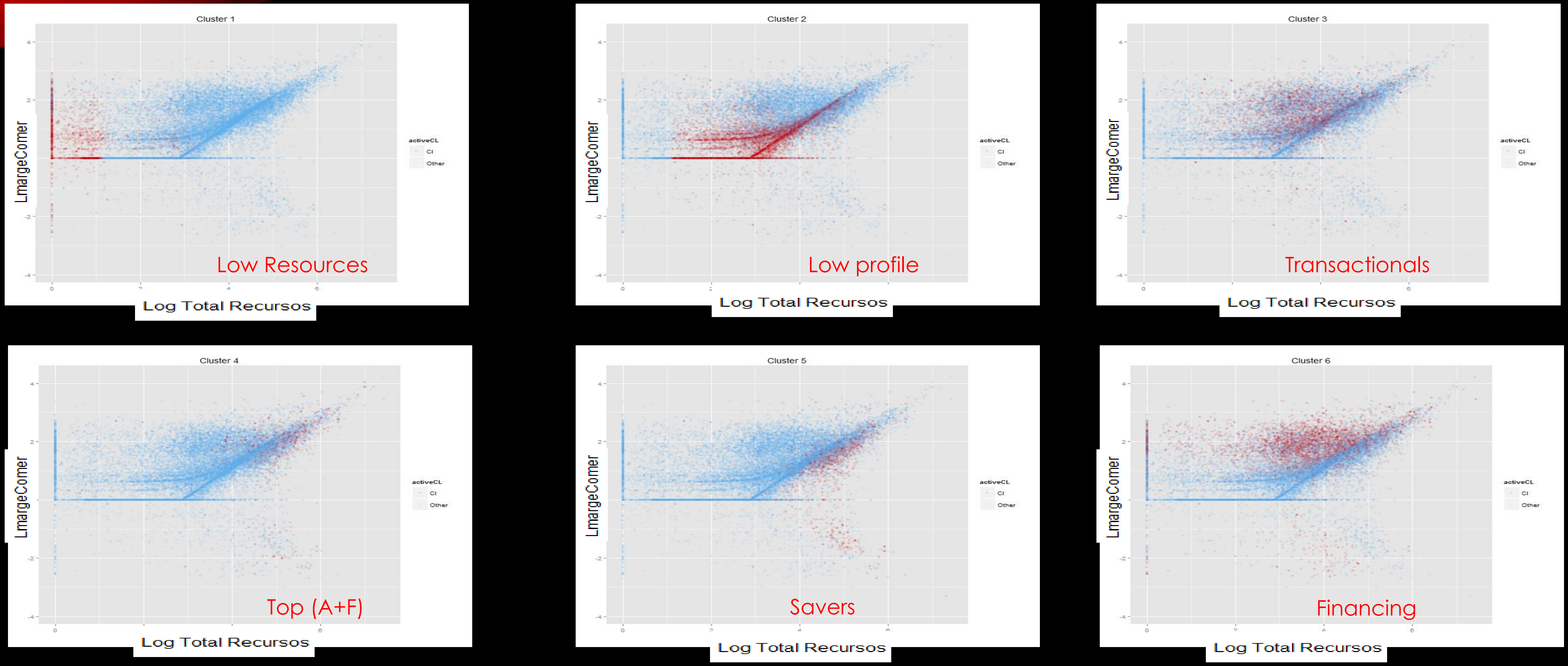
Results

- Customer behavior Knowledge.
- Determining the best campaigns targets, considering all the customers.
- Complete Customers life cycle modeling, including aspects such as Churn.

5 - Examples. Propensity Models for Marketing

	Goals	Results
Pension Plan	• Improvement in the purchase of financial products with information from previous purchases. • Customers Target focused on non- traditional company products.	• Doubling the number of products purchased. • New profile of the contractor: regular contributions (youngest, less income, less contribution than traditional contractor) • Non-stationary sales.
CIALP	• Improvement in purchasing new financial products (without previous purchases information) • First targets extrapolated from the Pension Plan model (regular contributions) • New model and frequent updating.	• Accelerating effect of purchasing. 
Consumer Finance	• Estimate the need for funding. • Discriminate basic needs from variable needs. • Select the action levers adapting the value proposition (Product, Price, Channel).	• Use of *datalake* for achieving models in line with business objectives. • Continuous improvement by acquired knowledge. Ex: Increased card use: High probability purchasing target identification based on the volatility of forecasting needs.

5 - Examples. Propensity Models for Marketing



Client segments (red) identified as emerging form, represented by the rest (in blue) in a Commercial margin vs Resources diagram.

5 - Examples. Pricing Models in Financial Products

Goals

- Models development that optimize the price of liability products considering:
  - The minimum price that the client will accept for a specific product.
  - The Global Strategies of the Financial Institution.

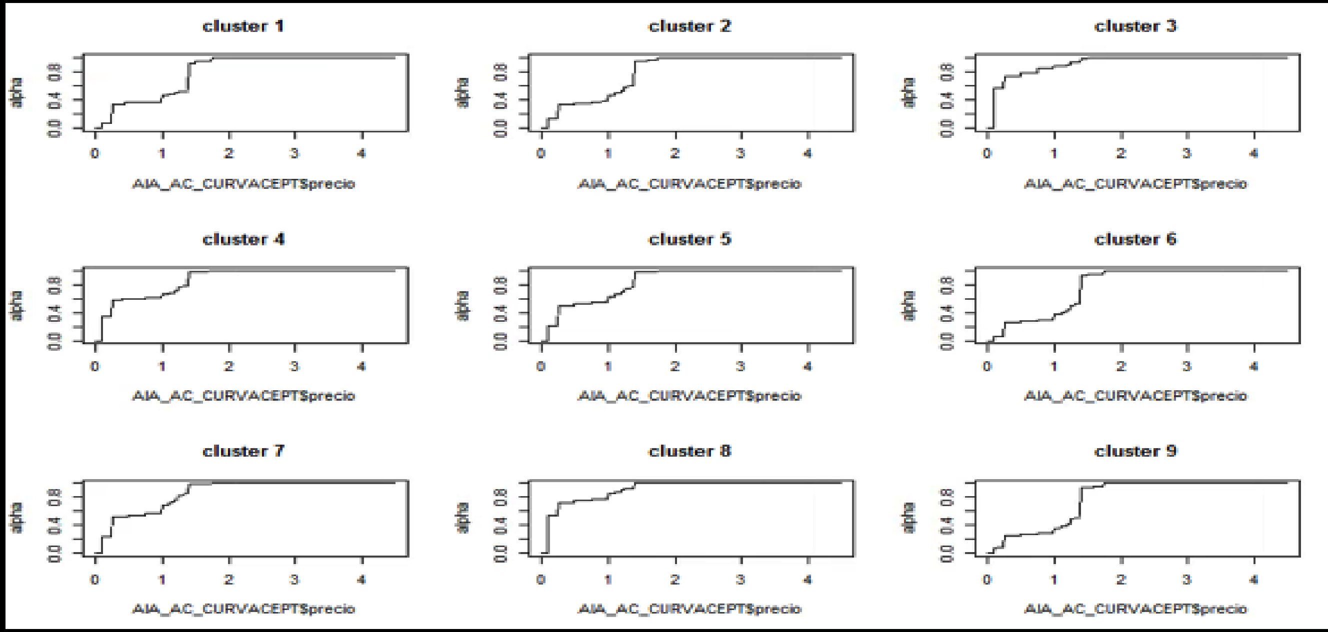
Methods

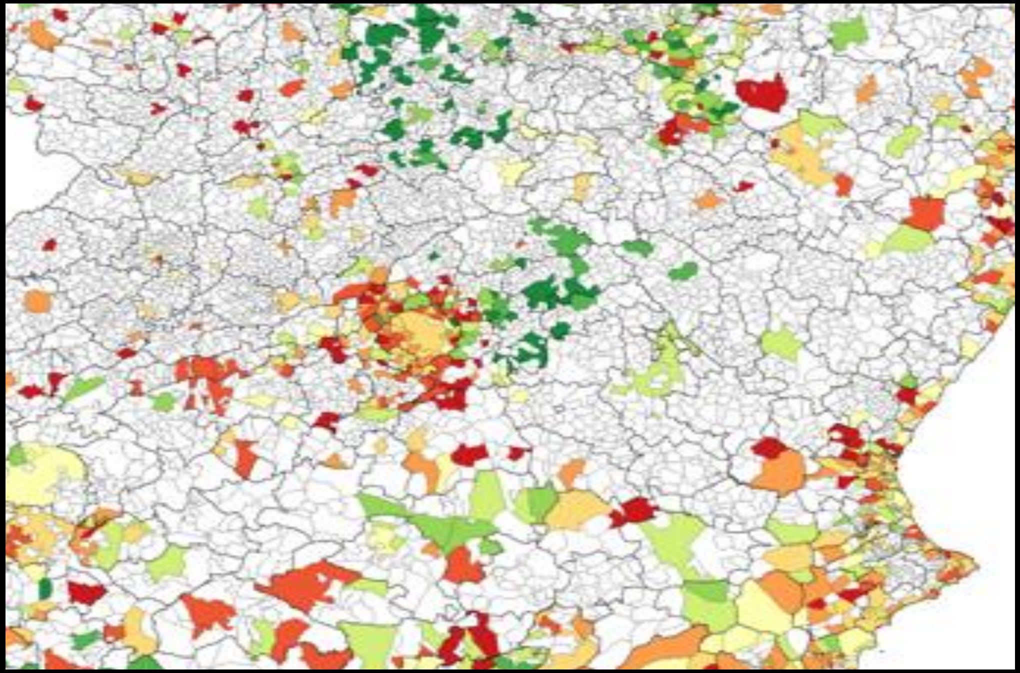
- **Conceptualization:** Client’s characteristics, including its negotiation skills.
- **Prediction**: Models based on GBM (Gradient Boosting Machine) to obtain the minimum price that the client will accept for a specific product.
- **Optimization:** Global optimization considering the Financial Institution’s Strategies.

Results

- Optimum prices for the different products:
  - Individualized per client.
  - Alternatively by client’s clusters.
- Software for monitoring and pricing simulation.

5 - Examples. Pricing Models for Financial Products





Probability curves of Price Acceptance vs. Price calculated to different clients segments (left.). The actual hiring is monitorable for better follow-up of the activity (r.).

5 – Examples. Monitoring Large Infrastructures

Goals

- Monitor and detect anomalies in large information structures:
  - Using both structured and non-structured information.
  - With real-time processing capacity.

Methods

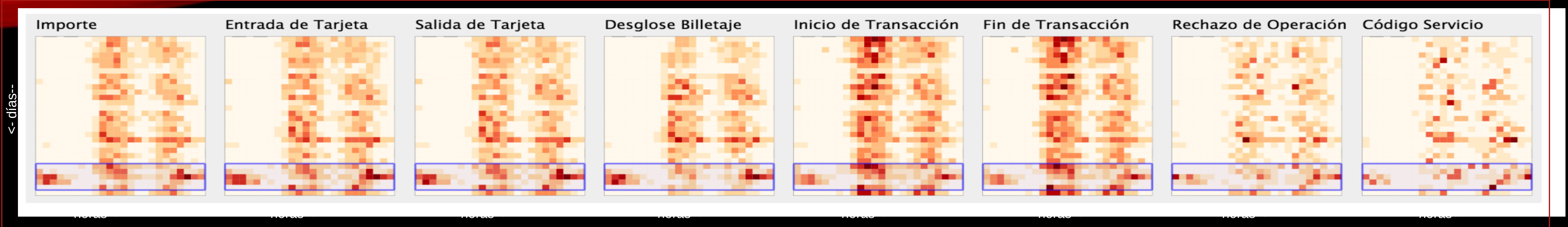
- **Conceptualization**:
  - Unsupervised methodology based on auto-encoders to detect anomalies.
  - Unsupervised methodology to identify the different aspects that could be considered anomalies(SIO), to offer results based on business vocabulary.

Results

- Anomalies detection system based on multiple information sources both structured and non-structured.
- Visualization tools and analysis of detected anomalies, oriented to search explanatory factors.

BIG DATA PROJECT DEVELOPMENT

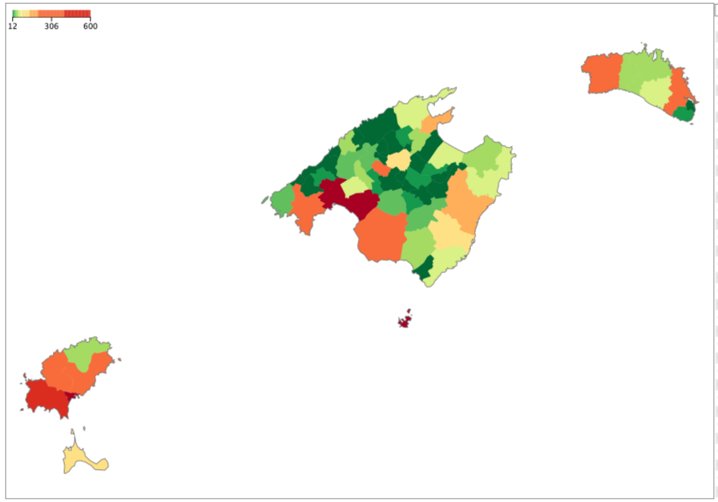
5 – Examples. Monitoring Large Infrastructures



Visualizaciones en Cassiopeia 3.0

DEC201409

Baleares coloreada por actividad en la noche de Sábado 2014-09-27 20:00-05:00)



Código	Municipio	Actividad
07040	Palma de Mallorca	2675
07026	Eivissa	945
07048	San José	436
07011	Calvià	307
07046	Sant Antoni de Portmany	262
07031	Llucmajor	248
07032	Mahón	236
07054	Santa Eulalia del Río	232
07015	Ciutadella de Menorca	231
07008	Binissalem	203
07033	Manacor	166
07003	Alcúdia	164
07024	Formentera	137
07027	Inca	119
07022	Felanitx	83
07057	Santanyí	72
07055	Santa Margalida	66
07062	Son Servera	65
07014	Capdepera	62
07036	Marratxí	59
07051	Sant Llorenç des Cardassar	55
07002	Alaior	53
07042	Pollença	51
07061	Sóller	44
07044	Pobla (Sa)	41
07006	Artà	39
07037	Mercadal (Es)	38
07902	Migjorn Gran (Es)	37
07013	Campos	37
07023	Ferreries	34
07056	Santa María del Camí	30
07050	Sant Joan de Labritja	30
07043	Porreres	29
07010	Bunyola	28
07004	Algaida	27
07005	Andratx	26
07065	Vilafranca de Bonany	26
07001	Alaró	24
07052	Sant Lluís	19
07012	Campanet	18
07035	Maria de la Salut	18
07038	Montuïri	18
07020	Esporles	17
07060	Sineu	12
07016	Consell	12

ATMs in Palma and Ibiza with unusual night activity. Visualization tools helps understand the nature of this anomaly.

5 - Examples. Text analysis applied to campaigns

Goals

- Use of textual descriptive information from customers relationship to improve propensity models to buy financial products.
- Unsupervised use of the large volume of written text in natural language.

Methods

- **Conceptualization:** unsupervised methodologies such as Word2Vec, gather words referring to a common matter.
- **Prediction**: Propensity models based on GBM (Gradient Boosting Machine), to incorporate the information coming from texts to behavioral models.

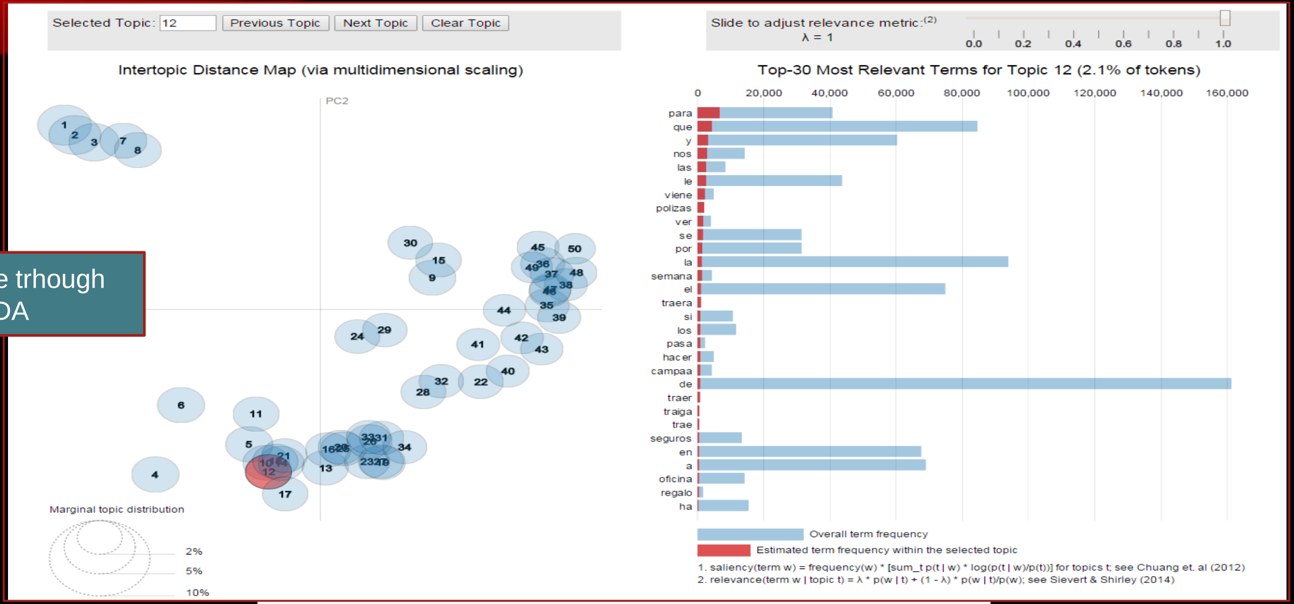
Results

- Purchase Propensity models which improve the traditional models based only on customers behavior.
- Can be applied to multiple campaigns of financial products.

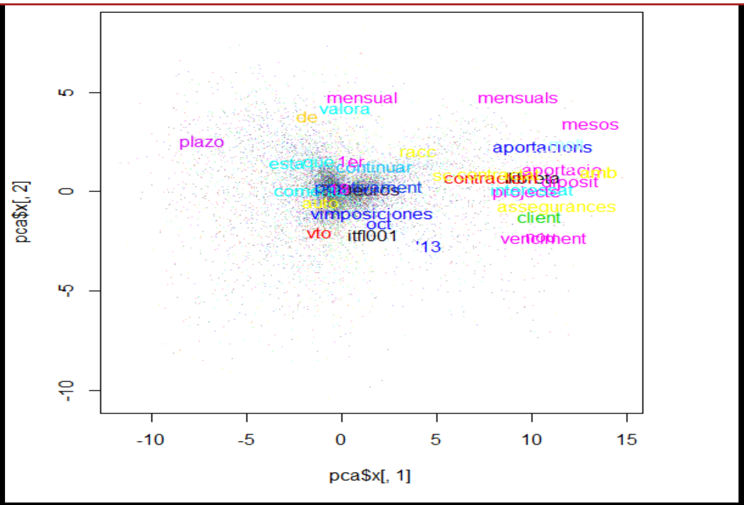
BIG DATA PROJECT DEVELOPMENT

5 - Examples. Text analysis applied to campaigns

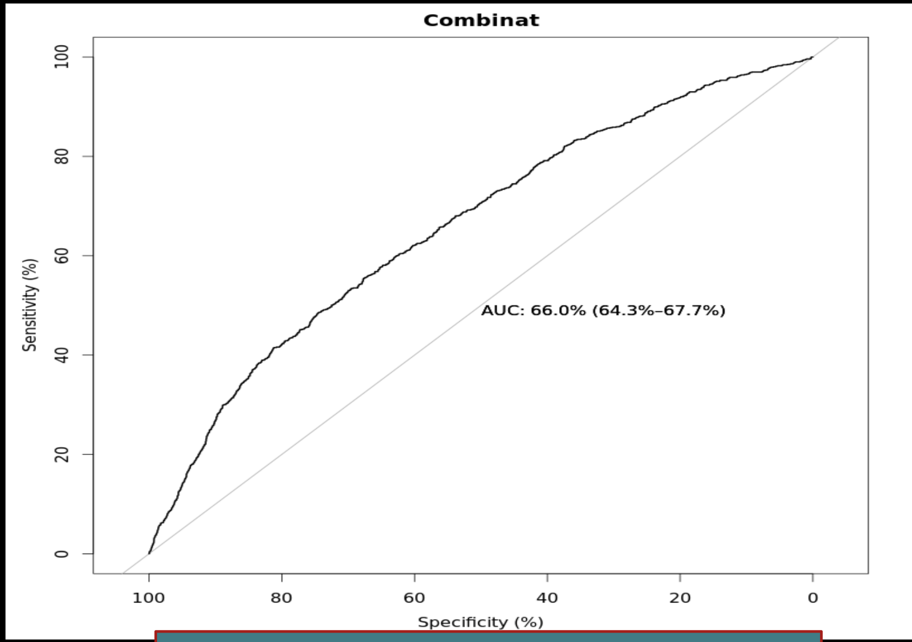
Structure trhough LDA



Structure through word2vec



Combined Word2Vec + Behavioral

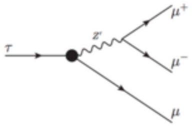


ROC Curve of Combination Model

The choice of the most suitable structuring technology of unsupervised text allows the improvement of target selection in commercial campaigns.

COMPETITIVE DATA SCIENCE

"Flavour of Physics"
Kaggle Challenge



Completed • $15,000 • 673 teams

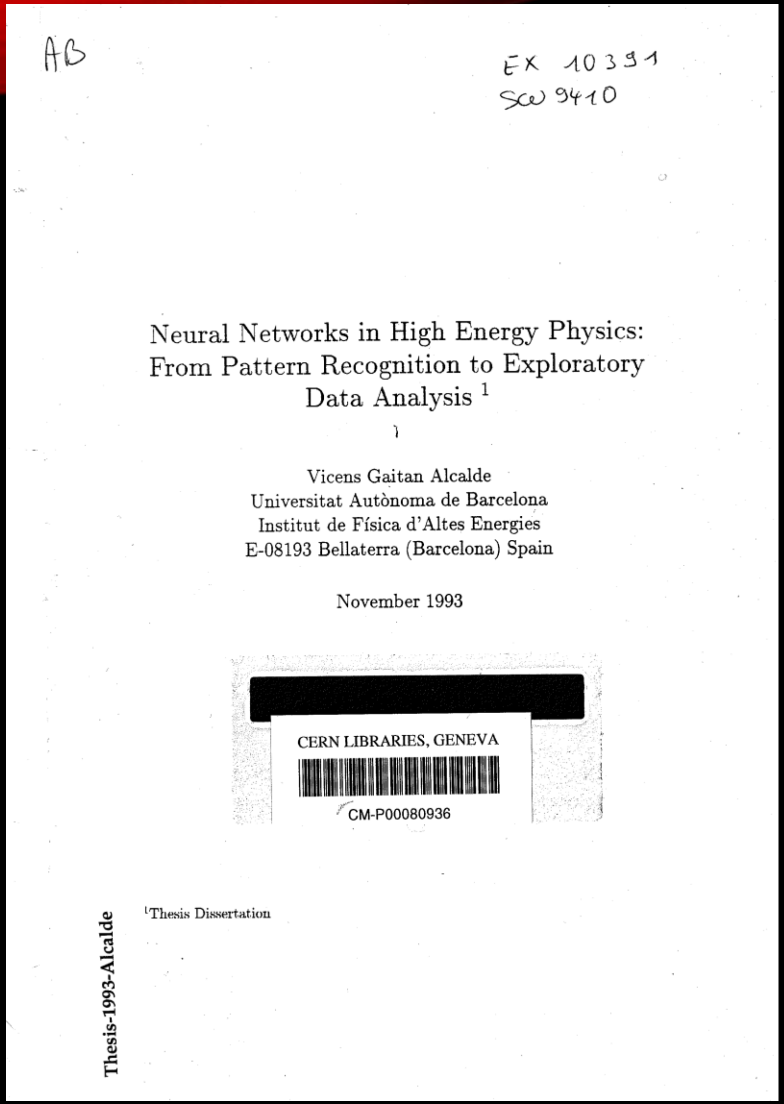
Flavours of Physics: Finding τ → μμμ

Mon 20 Jul 2015 – Mon 12 Oct 2015 (4 months ago)

CERN · LHCb · SCHOOL OF DATA ANALYSIS · Yandex Data Factory · intel · University of Zurich UZH · THE UNIVERSITY OF WARWICK · Higher School of Economics

#	Δrank	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Go Polar Bears ‡ *	1.000000	49	Mon, 12 Oct 2015 22:57:38
2	↑1	Alexander Gramolin ‡ *	0.999998	12	Mon, 12 Oct 2015 18:38:07
3	↓1	Josef Slavicek ‡ *	0.999897	25	Mon, 12 Oct 2015 21:49:53
4	—	Michal Wojcik	0.999225	35	Mon, 12 Oct 2015 23:57:46 (-3h)
5	—	rakhlin	0.998338	31	Mon, 12 Oct 2015 23:32:18 (-5.8h)
6	—	Archy ‡	0.997784	47	Mon, 12 Oct 2015 20:31:53 (-7.8h)
7	—	Faron	0.995918	66	Mon, 12 Oct 2015 18:15:46
8	—	Alejandro Mosquera	0.994946	28	Mon, 12 Oct 2015 15:23:51 (-19.7h)
9	—	Anton Laptiev	0.994894	61	Mon, 12 Oct 2015 23:56:37
10	—	Andrzej Prałat	0.993957	14	Mon, 12 Oct 2015 18:25:39 (-0.3h)
11	—	Ivanhoe	0.993692	35	Mon, 12 Oct 2015 23:17:39
12	—	George Solymosi	0.993646	95	Mon, 12 Oct 2015 23:58:45 (-0.6h)
13	—	PhysicsTau	0.993099	90	Mon, 12 Oct 2015 22:30:42
14	↑1	Grzegorz Sionkowski	0.992031	49	Mon, 12 Oct 2015 23:50:56 (-27.2h)
15	**↓1**	**Vicens Gaitan [0.989012 physically sound]**	**0.991860**	**85**	**Mon, 12 Oct 2015 20:56:04 (-5.9h)**
16	—	achm	0.991841	105	Mon, 12 Oct 2015 13:06:31 (-44.1h)
17	—	bgeol	0.991709	14	Tue, 06 Oct 2015 03:56:14 (-5.3d)

CLOSING THE LOOP


AB

EX 10391
Scu 9410

Neural Networks in High Energy Physics:
From Pattern Recognition to Exploratory
Data Analysis [1]

Vicens Gaitan Alcalde
Universitat Autònoma de Barcelona
Institut de Física d'Altes Energies
E-08193 Bellaterra (Barcelona) Spain

November 1993

[1]Thesis Dissertation

Machine Learning in HEP

Todat we have:

- Good tools
- Open Data
- CPU power

MACHINE LEARNING IN HEP

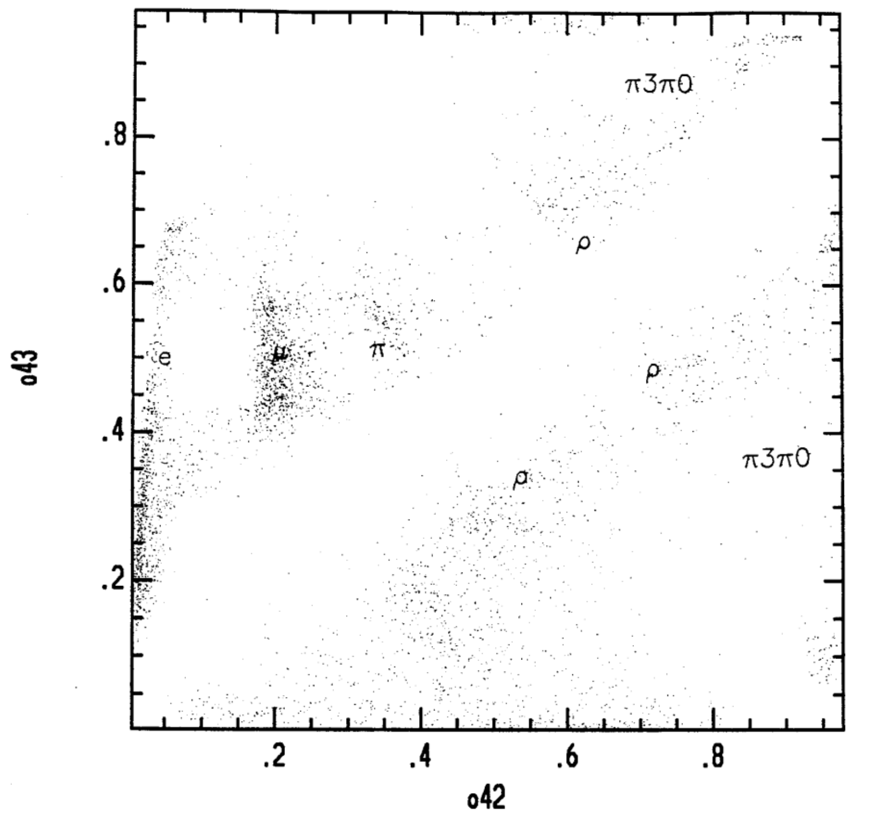
Example: exploring tau decay at LEP (ALEPH 1993)
(yes, e+ e- physics is cleaner…)

Feeding an autoencoder with "elaborated" detector data we are able to "discover" different decay modes looking at the compressed representation without a physics model (MC)

Today is possible to do "end to end" autoencoding from raw detector data

Input neuron	Variable Description	Kolmogorov C.L.
1	Number of charged tracks in hemisphere +	0.947
2	Number of charged tracks in hemisphere -	0.339
3	Number of neutral tracks in hemisphere +	0.010
4	Number of neutral tracks in hemisphere -	0.047
5	Total charged energy in hemisphere +	0.131
6	Total charged energy in hemisphere -	0.078
7	Total neutral energy in hemisphere +	0.874
8	Total neutral energy in hemisphere -	0.995
9	Number of identified μ in hemisphere +	1.000
10	Number of identified μ in hemisphere -	1.000
11	Number of identified electrons in hemisphere +	0.367
12	Number of identified electrons in hemisphere -	0.921
13	Number of identified γ in hemisphere +	0.258
14	Number of identified γ in hemisphere -	0.746
15	Planarity	0.489
16	Total momentum in hemisphere +	0.523
17	Total momentum in hemisphere -	0.534
18	Invariant mass	0.90621
-	Output neuron	0.457

5.4 Unsupervised τ classification 75



CONCLUSIONS

- The combination of Data + Models + Machine learning is changing the way we see the world
  - We can forecast, optimize, action and learn from the experience
- Physicist had good skills and knowledge for the modeling aspects
  - But is also necessary programming, hacking, data munging...
- Problems beyond basic science are also challenging ;).
  - Thera are interesting problems and solutions outside the academia. Cross breedin can be fruitful