IQ&AI

Supervised Elicitation Model for Causal Analysis of Companies' Performance

9th Annual Bayesialab Conference , 2021

October, 15th

Joël Pain
Emmanuel KEITA
Christophe Thovex

IQ&AI

The team

Joël Pain
CEO Up&Up

Strategy consulting and restructuring

Emmanuel KEITA
Data boundary spanner

Human+Machine continuum : Bayesian networks, Information system
AI Associate Senior Consultant (Aveyo Consulting)
Blockchain project (Avkee.com)

Christophe THOVEX
Network scientist (PhD) – Associated scientist (LAFMIA)

Applied research in network science
R&D in network science/data science and big data
First author, editorial board member, reviewer of scientific publications

Summary

Joël Pain
Christophe Thovex
Emmanuel KEITA

IQ&AI

Context & purposes

- We led an assignment for a large multinational company (MNC) willing to optimize its +1000 agencies network. The purpose is :
  1) To spot key success/failure factors related to branches' performance
  2) Eventually to build a decision-making model aiming at providing branch footprint optimization guidance.
- Context : the MNC has been working on it for a while, but the recent health crisis has made it more urgent to identify the key strategic and operational issues to be dealt with. It has then become urgent to treat both :
  - Structural issues that existed prior to the crisis (growth, profitability…)
  - New issues resulting from the crisis (shift in customer behaviour, digital competition…)
- The objective is to assist the MNC in making the appropriate strategic and organizational decisions. To do this, we must provide the relevant "decision making" facts and figures, i.e. those that will allow to :
  - Identify possible sources of underperformance and avenues for optimization,
  - Identify the relevant levers of action,
  - Prioritize them so as to maximize the chances of success of an optimization and adjustment plan
- The purpose it then to better anticipate : we developed a methodology intended to shed light on possible upcoming events
- The methodology we elaborated is based on data processing and involves :
  - Bayesian networks (BayesiaLab)
  - Neural networks
  - Humain expertise (domain experts).

IQ&AI

Initial approach

Source	Target	Correl. [Pearson]
Variable 1	Variable 28	0,520830077
Variable 1	Variable 43	0,518258786
Variable 1	Variable 30	0,516407663
Variable 1	Variable 39	0,515761265
Variable 1	Variable 86	0,515564785
Variable 1	Variable 17	0,515179052
Variable 1	Variable 34	0,5149961
Variable 1	Variable 91	0,507724536
Variable 2	Variable 37	0,50682865
Variable 2	Variable 72	0,506706964
Variable 2	Variable 4	0,505722602
Variable 2	Variable 40	0,504405526

- The first stage consisted in identifying the variables that could make sense as regards to the questions asked:
  - What defines an agency that "performs“ or one that "underperforms"?
  - What are the indicators that can help anticipate developments and adapt agencies and the agencies network accordingly?
- To achieve this, we needed to move away from usual assumptions, and to adopt a non-siloed, "holistic" approach.
- Therefore, we collected and gather data from various origins and sources : the expectation was that this "diversity" would lead us to spot unexpected correlations between variables
- If we knew the reason why were doing this, ~~but~~ we didn’t know what we were looking for: the variables that would prove important could come either from the financial sphere, or be an indicator that characterizes HR issues, or others such as commercial activity, market shares, business fields, customer size, geographical location... no hypothesis should be ruled out, except those that were obviously far-fetched (e.g. little chance that the color of the walls of the agency could have an impact on turnover.... although...)
- Some data supposed to exist proved to be missing or unavailable… in conclusion, this “collect and centralization” phase has been a tedious and long one, but we knew that the rigor with which this was carried would condition the relevance of the observations and conclusions that will result…
- It resulted in 1,116 columns and 1,032 rows centralization document (i.e. more than one million data), ready to be refined for the purpose of the data processing …

Summary

Joël Pain

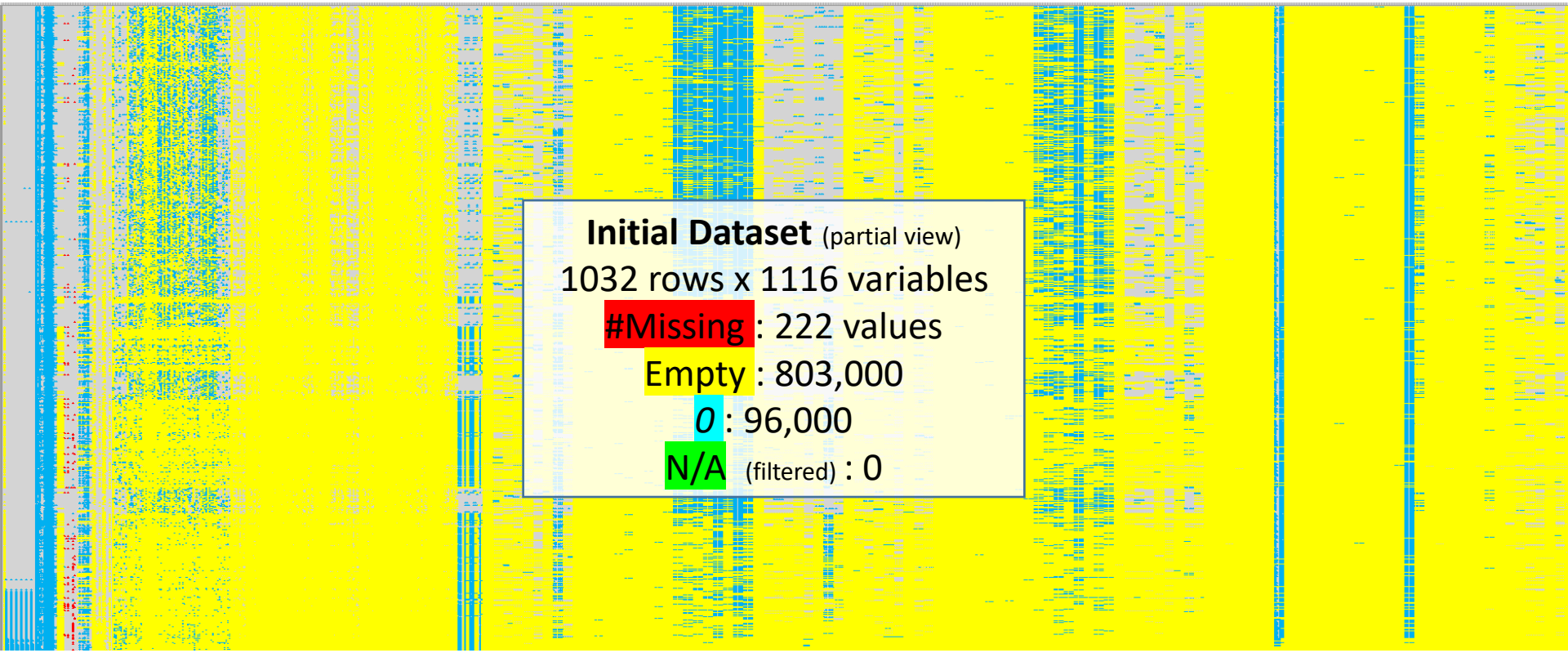
Christophe Thovex

Emmanuel KEITA

IQ&AI

'Noisy' signal in data ?

Empty, #Missing, 0 : What does that ***really*** mean ?



https://money.howstuffworks.com/business-communications/host-conference-call-using-your-mobile-device.htm
http://www.cdrp74.fr/voitures-robustes-route-de-montagne

'Noisy' signal in data ?

Dilemma : Missing vs Filtered alternative [bias]

https://money.howstuffworks.com/business-communications/host-conference-call-using-your-mobile-device.htm
http://www.cdrp74.fr/voitures-robustes-route-de-montagne

IQ&AI

'*Noisy*' signal in data ?

Initial dataset

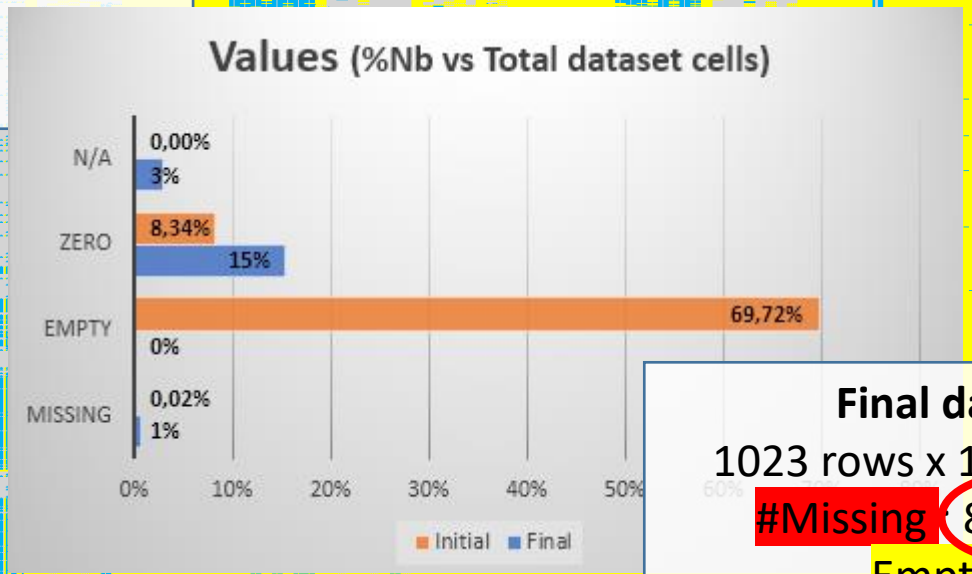
1032 rows x 1116 variables

#Missing : 222 values

Empty : 803,000

0 : 96,000

N/A : 0



Final dataset

1023 rows x 106 variables

#Missing : 832 values

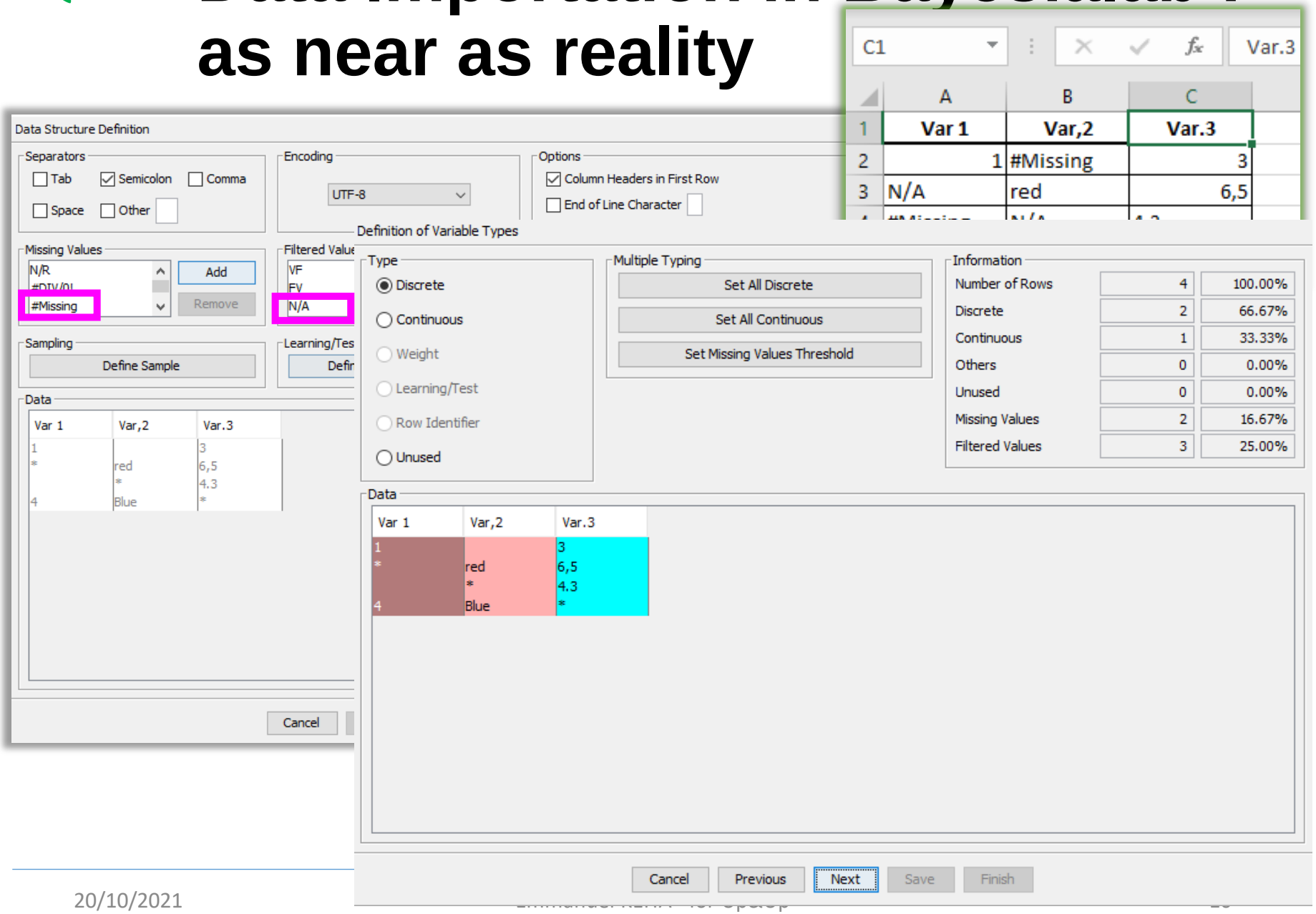
Empty : 0

0 : 16,800

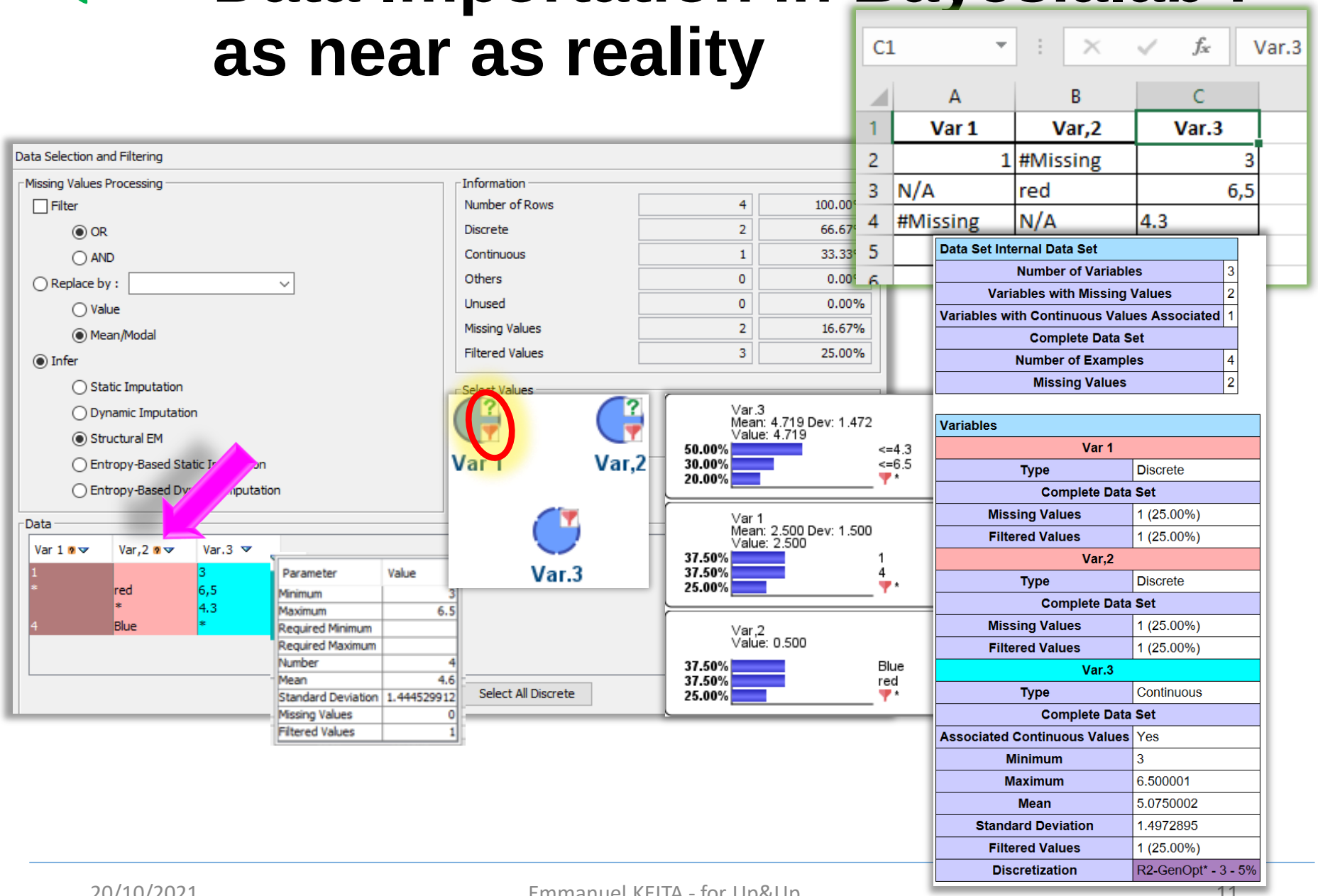
N/A : 3,138

IQ&AI

Data Importation in Bayesialab : as near as reality



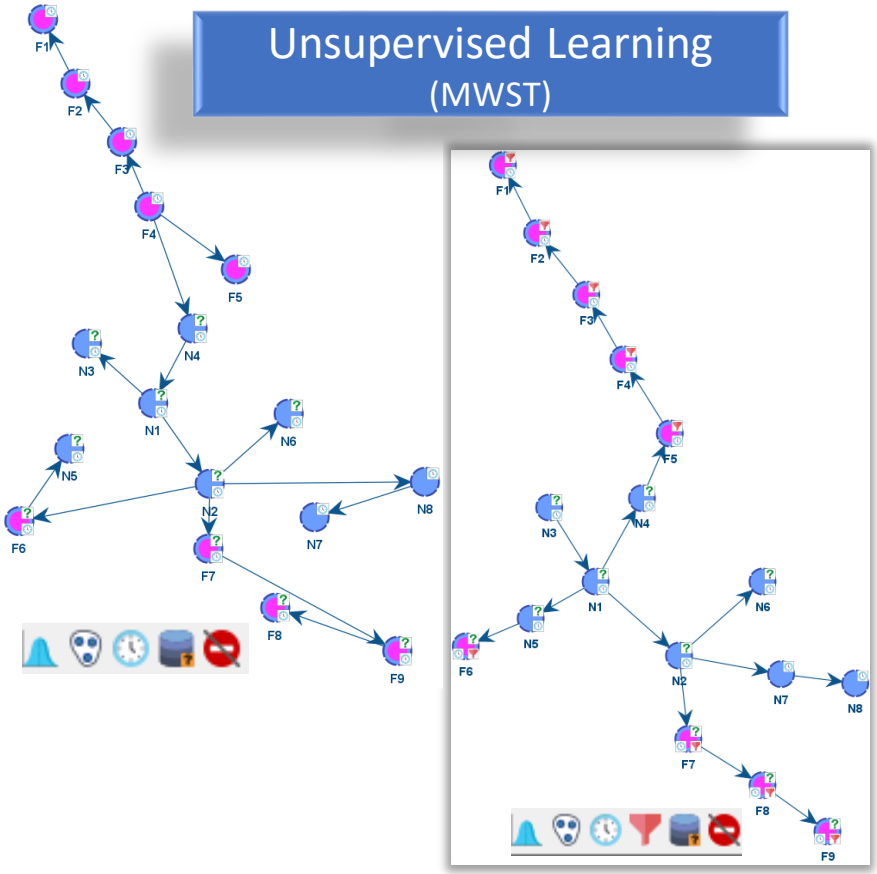
Data Importation in Bayesialab : as near as reality

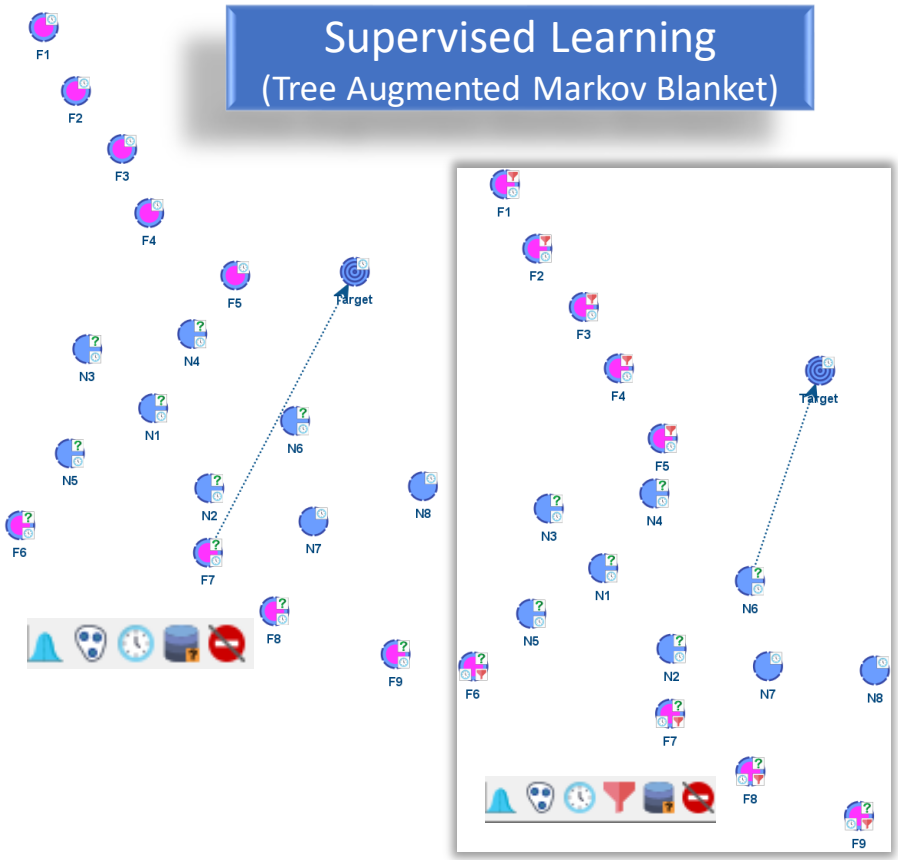


IQ&AI

Filtered values option

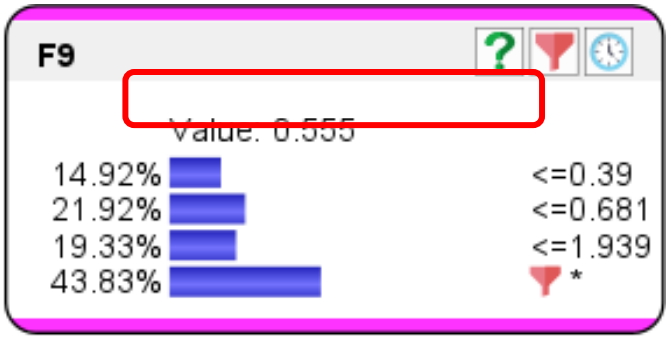
Differences with machine learning…



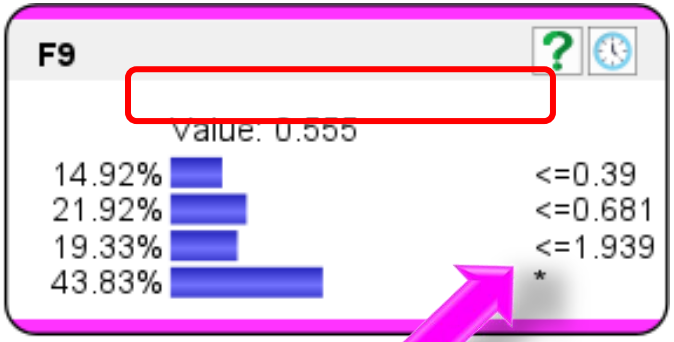


Filtered option in Bayesialab

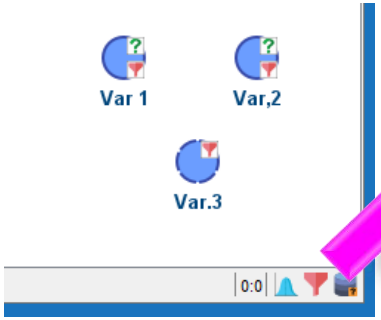
… Because of differences of perception !



Initial MDL score: 27,623.982

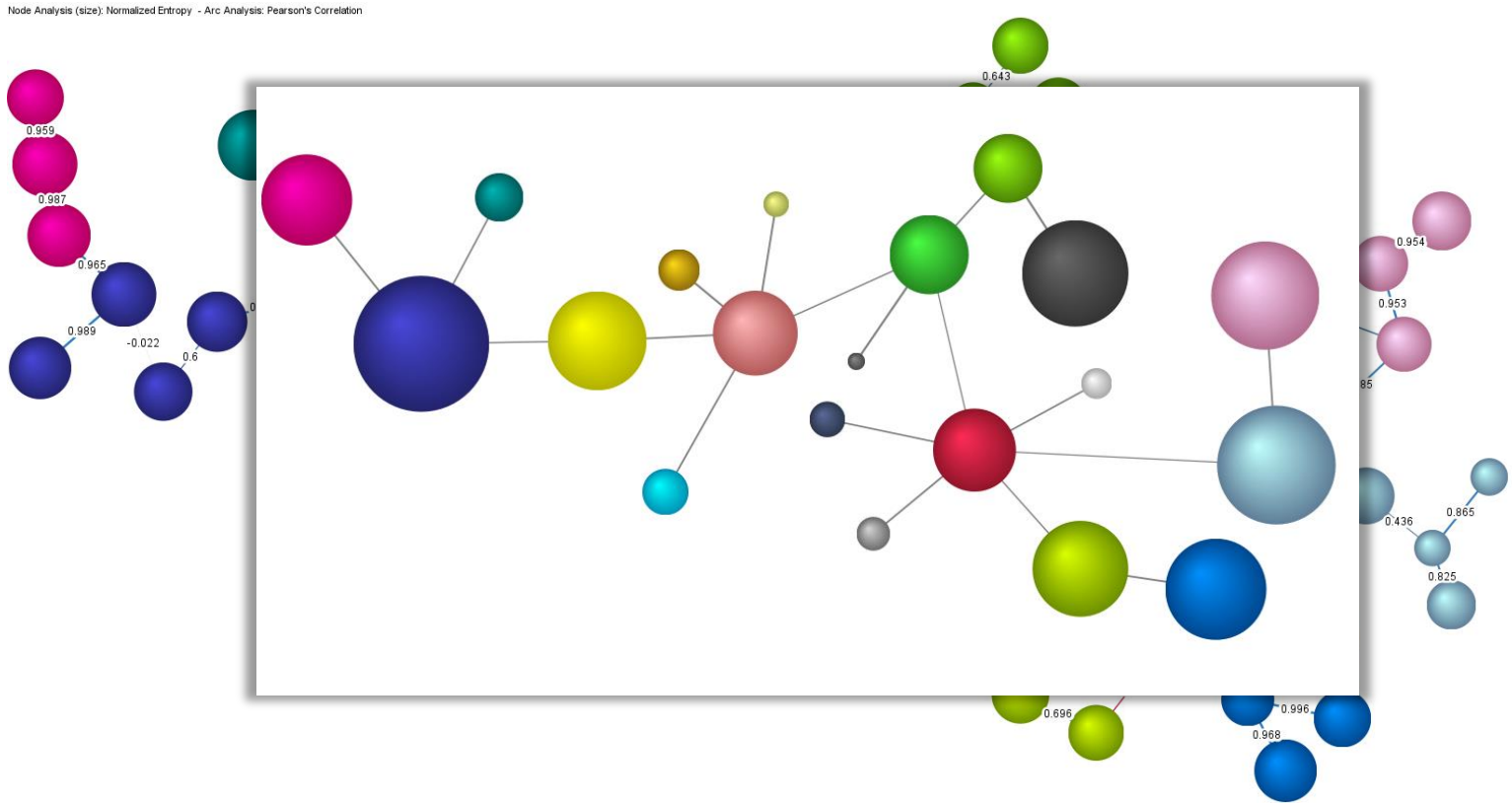


Initial MDL score: 34,380.139

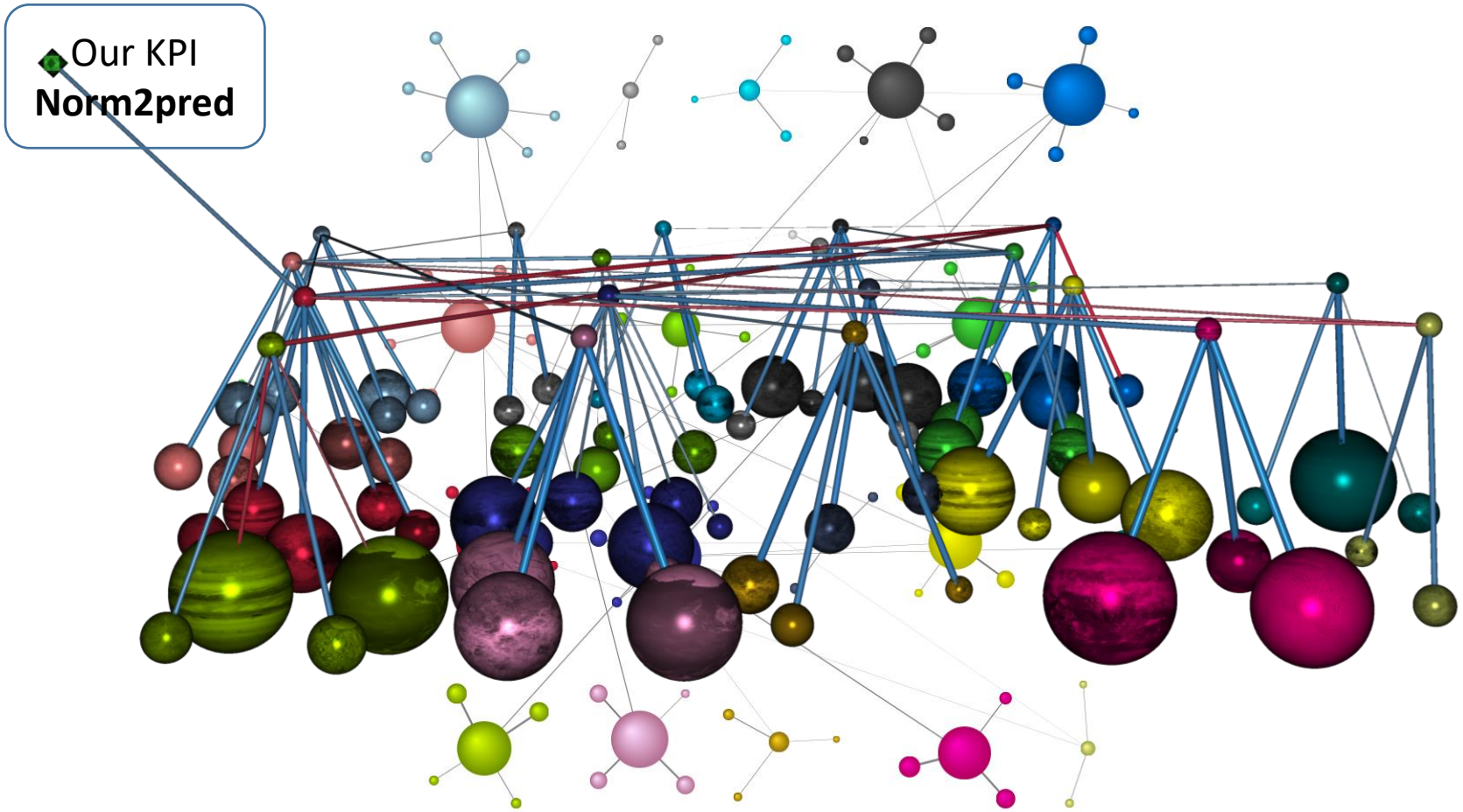


IQ&AI

PSEM 1 : Learning + Var. Clustering

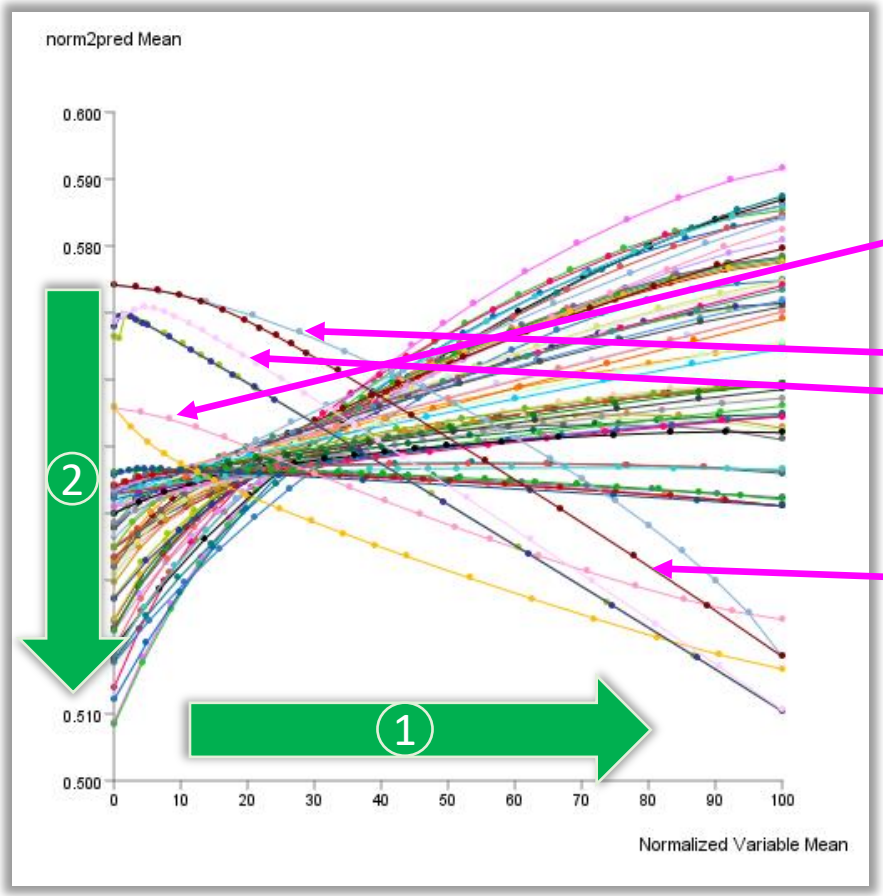


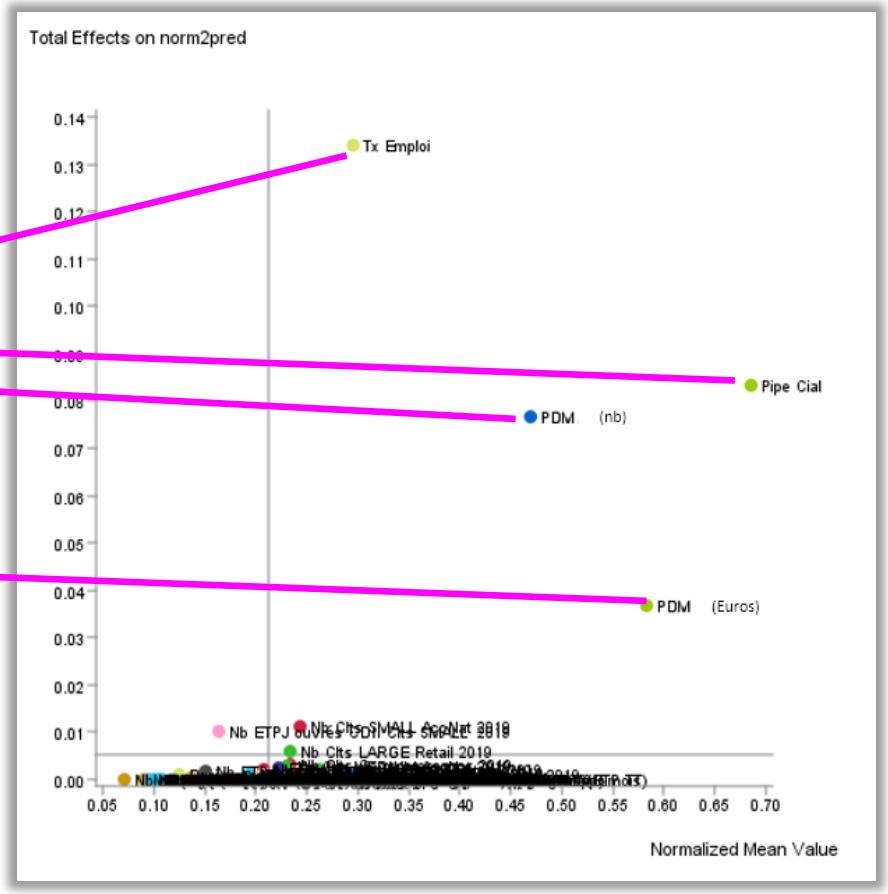
IQ&AI PSEM 2 : Data Clustering + Learning



IQ&AI

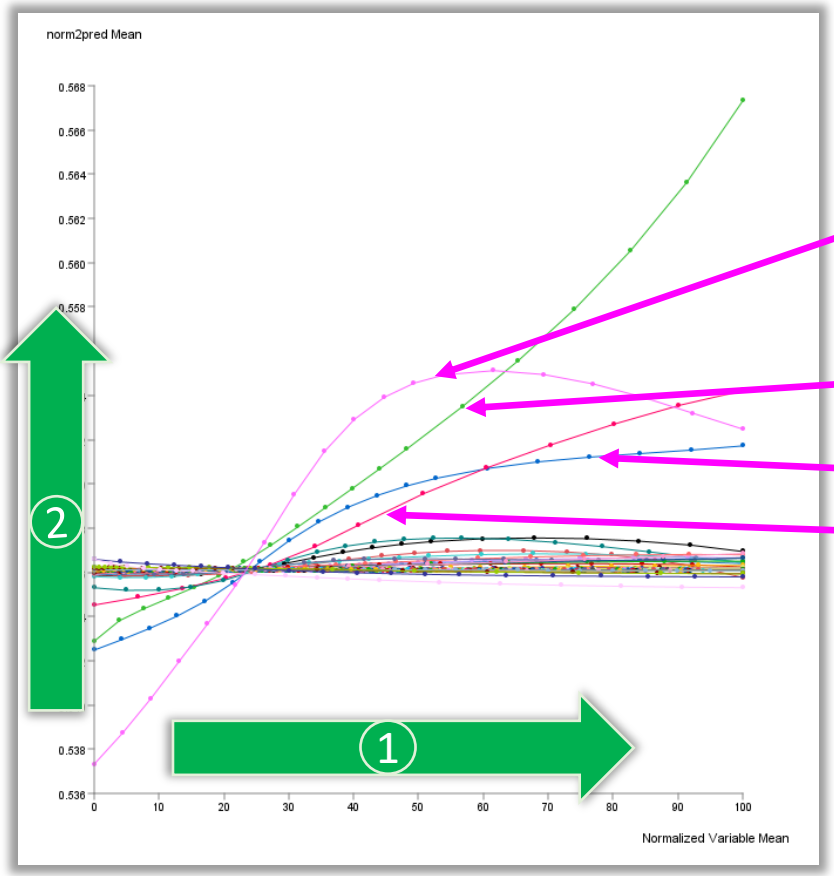
Total Effects on Target

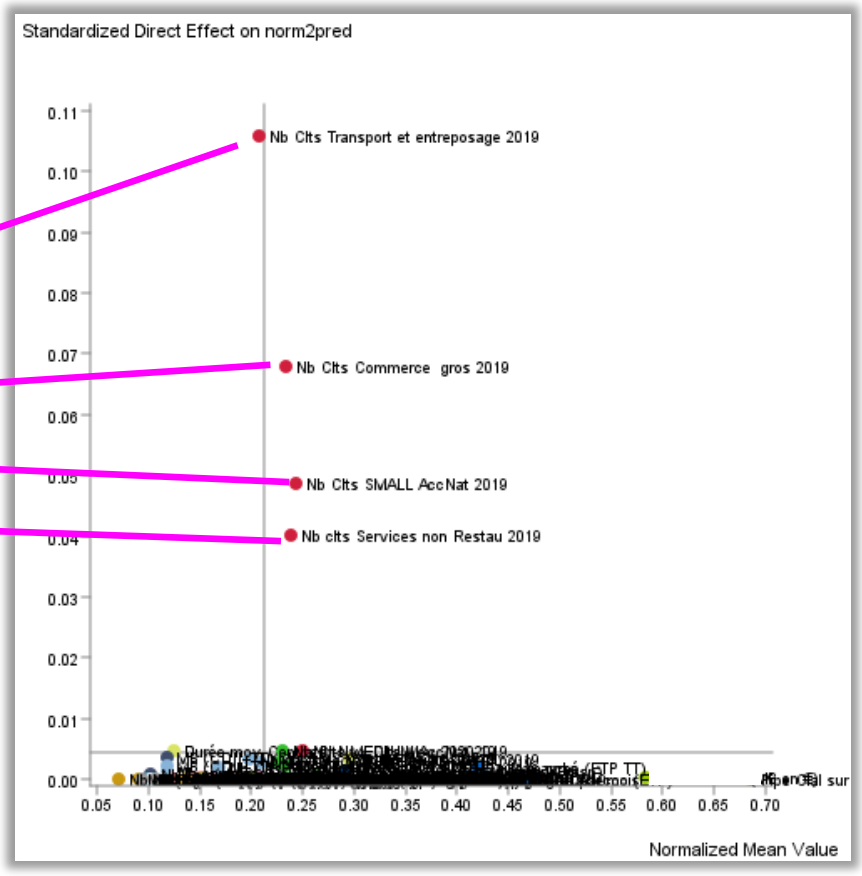




IQ&AI

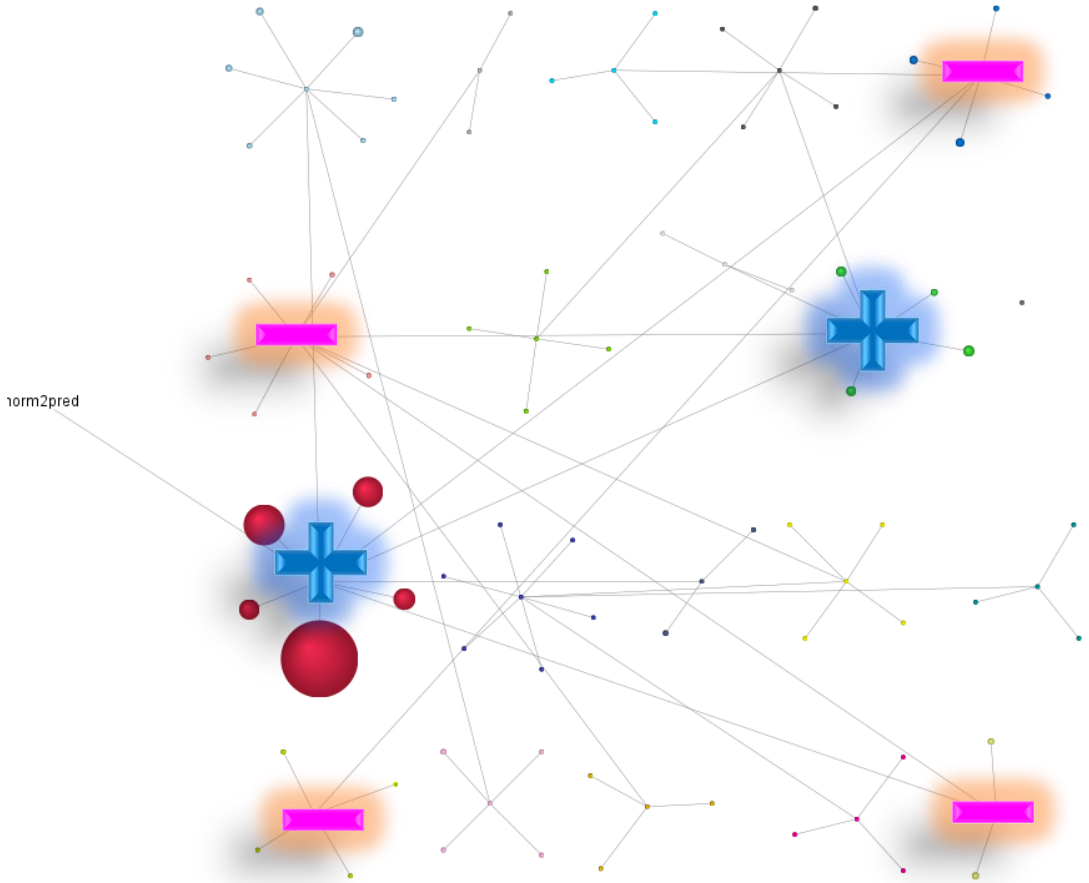
Direct Effects on Target





IQ&AI

Top 10 causal effects

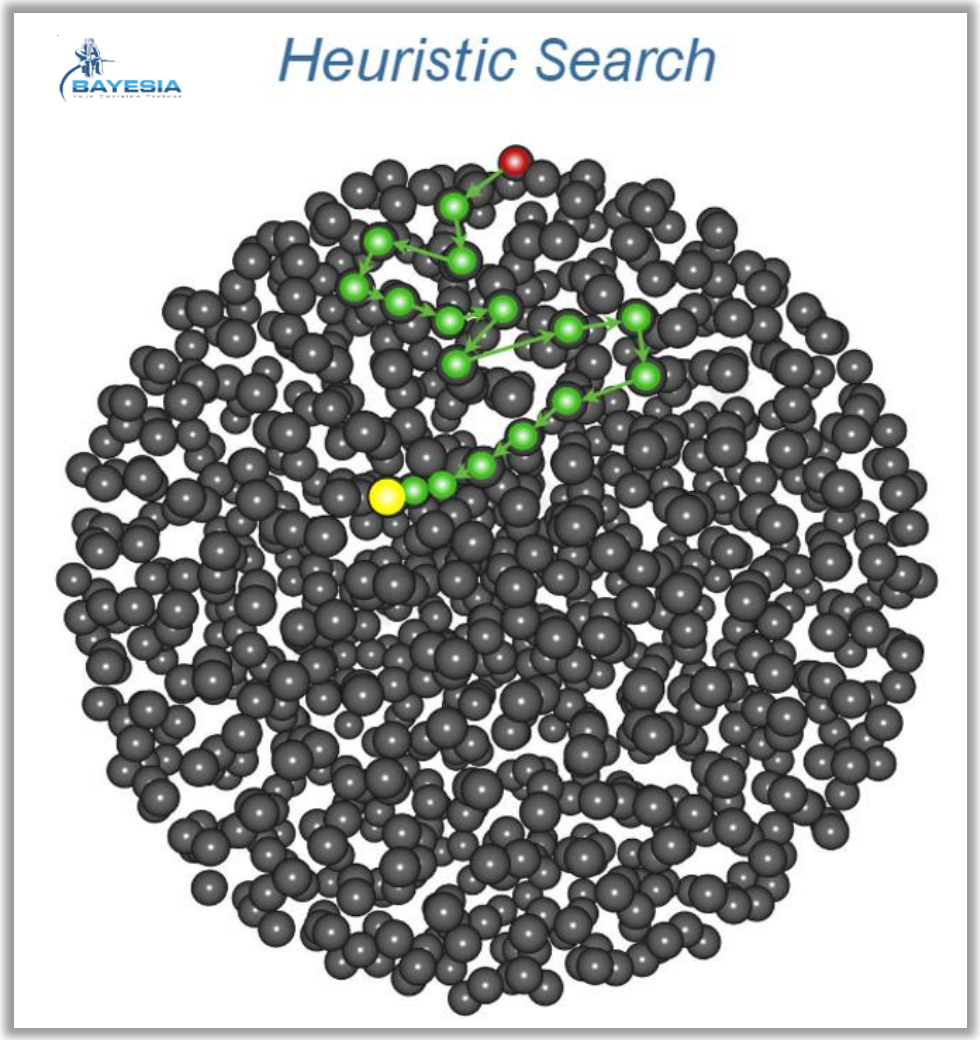


Positive

Negative

IQ&AI

That’s a great job !



Summary

Joël Pain

Christophe Thovex

Emmanuel KEITA

IQ&AI

Network Science and data elicitation

2006:
The United States National Research Council defines **network science** as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena."[1]

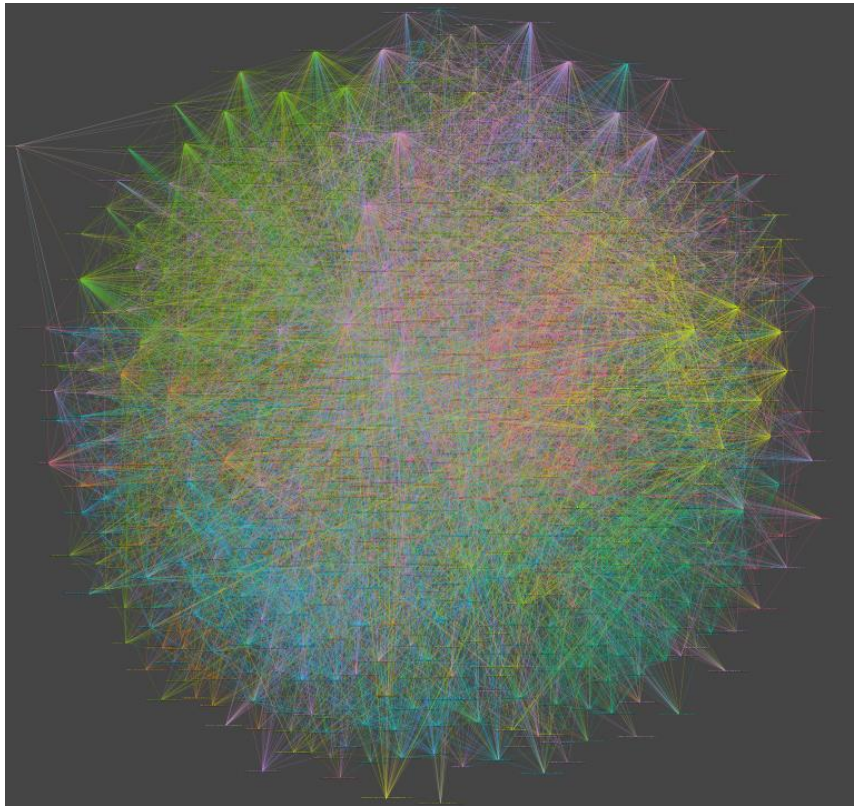
1. Committee on Network Science for Future Army Applications (2006). Network Science. National Research Council. doi:10.17226/11516. ISBN 978-0309653886.

IQ&AI

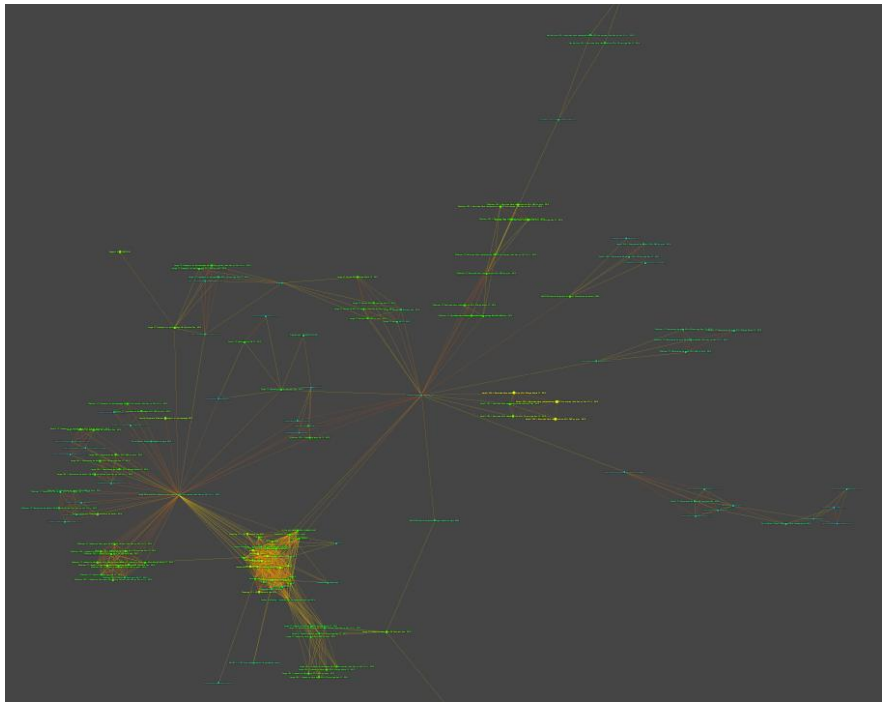
Network Science and data elicitation

High dimension and sparse data sets:
supporting data preparation with correlations network

First:
670 nodes, 8 530 correlated pairs, 10 clusters [Blondel2008]



Pearson > 0,5: **sparse** correlated structure
225 nodes, 1 100 pairs

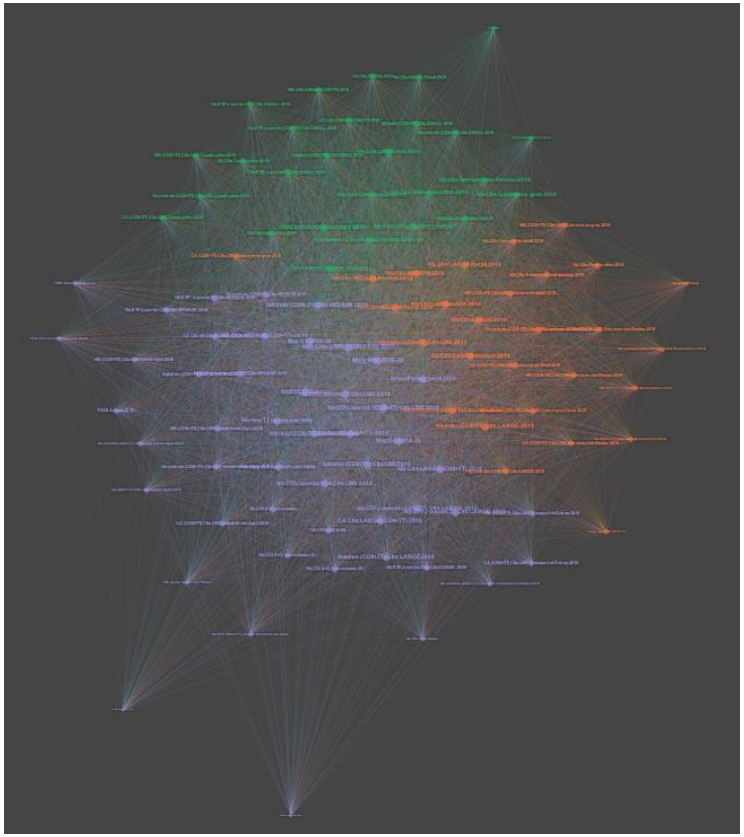


IQ&AI

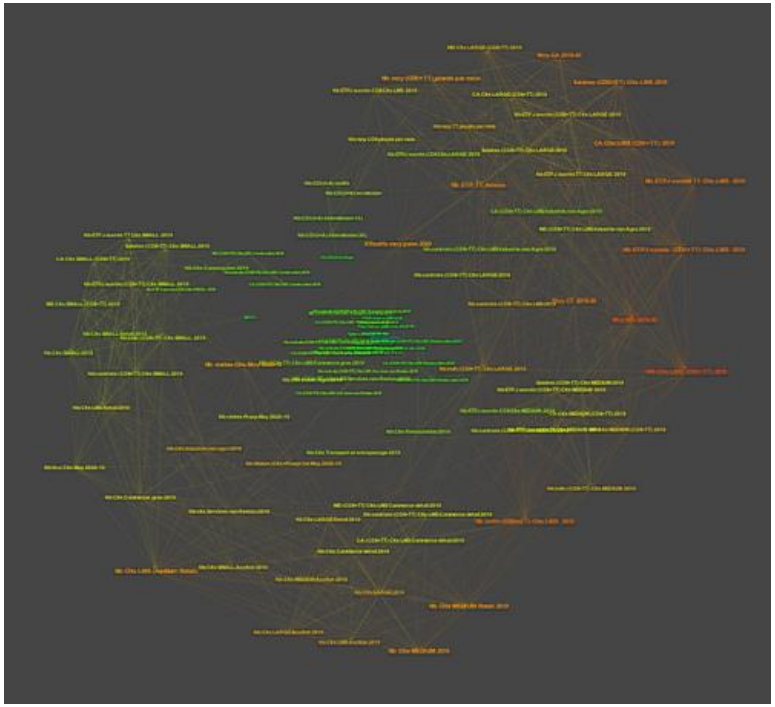
Network Science and data elicitation

Supporting data preparation with correlations network:
reduced dimension and sparsity

After:
104 nodes, 5 350 correlated pairs, 3 clusters



Pearson > 0,5 : **core-clustered** structure
104 nodes, 784 pairs



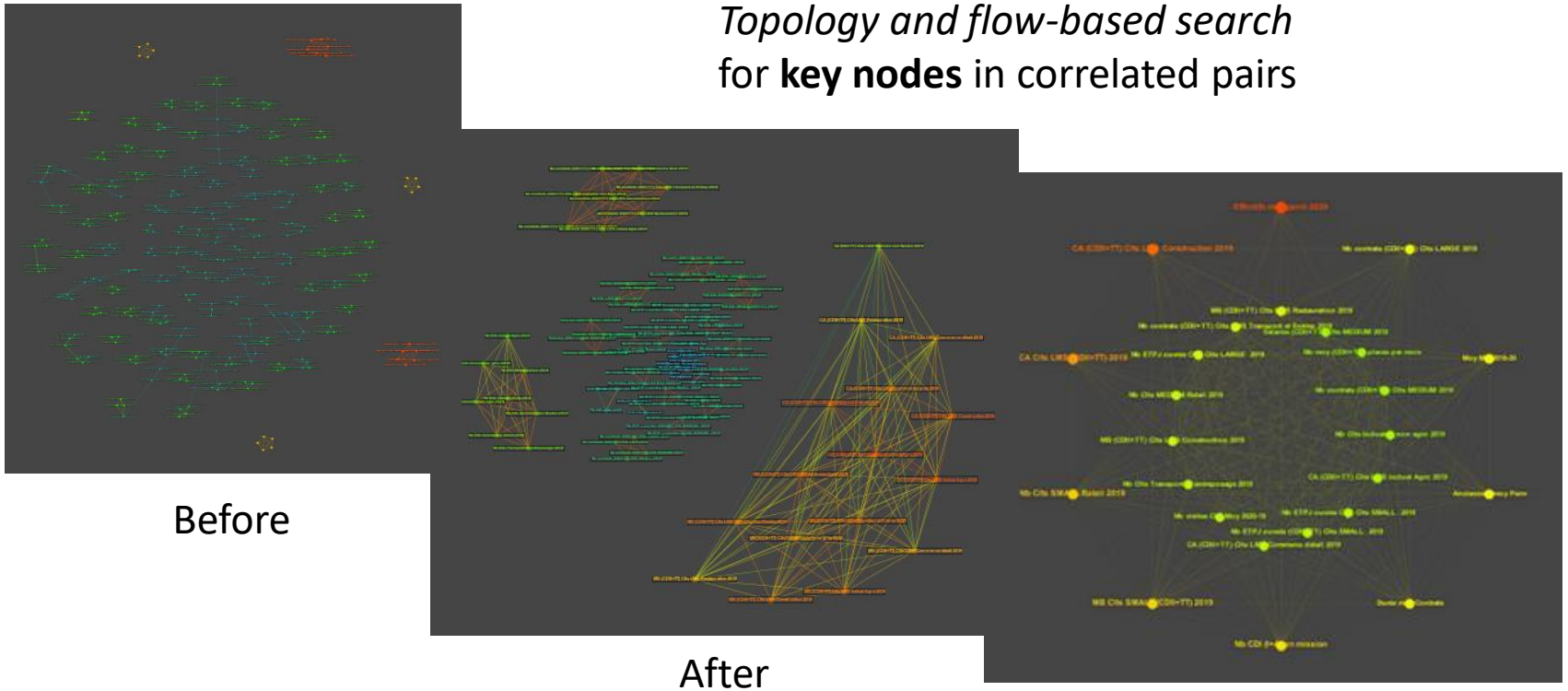
Network Science and data elicitation

Supporting data elicitation with network science

How ?

Exper. graphs using Hilbert-Schmidt Independence Criterion (NHSIC), Automatic Relevance Determination (ARD) and Bridging Centrality [Hwang2006].

Topology and flow-based search
for **key nodes** in correlated pairs



Before

After

Key-nodes structure

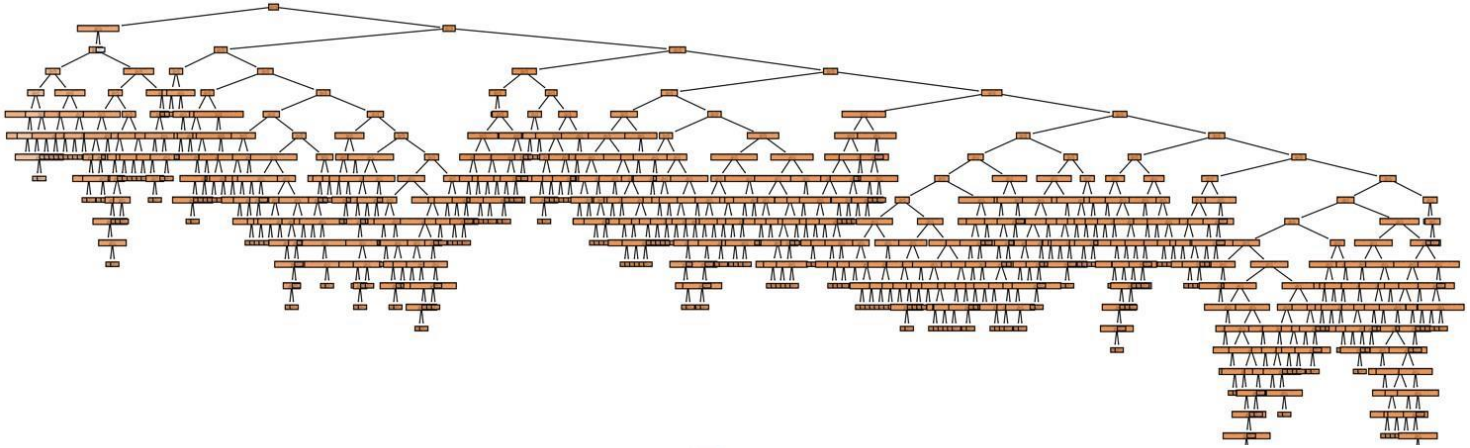
IQ&AI

Network Science and data elicitation

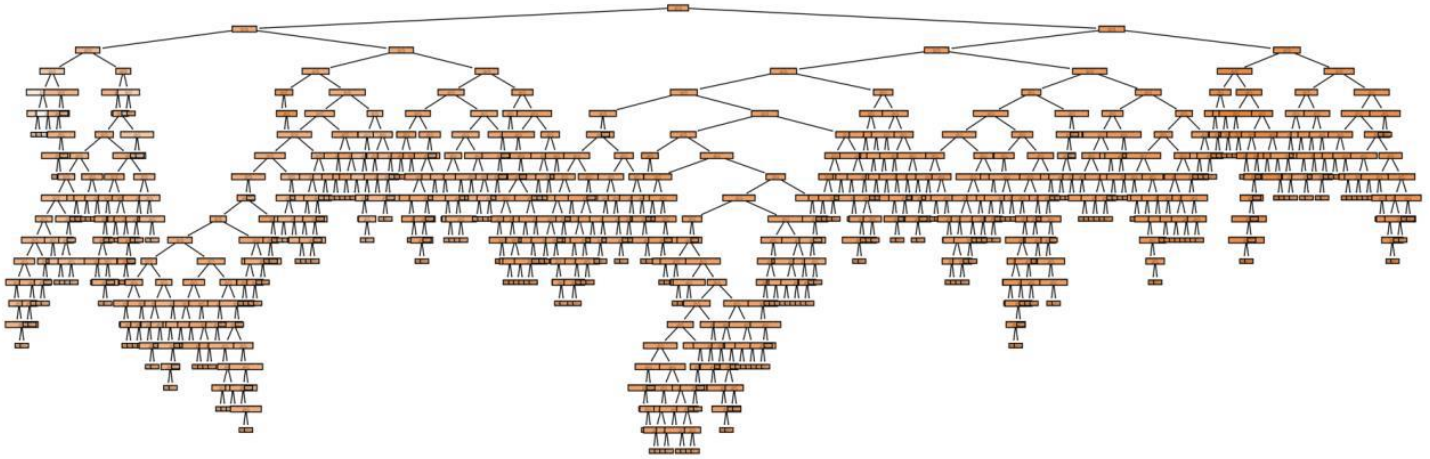
Generic effects on regression trees for BN in Bayesia Lab - decision support

Random tree forest regressor – representative examples (scikit-learn 0.24.1):

Before



After

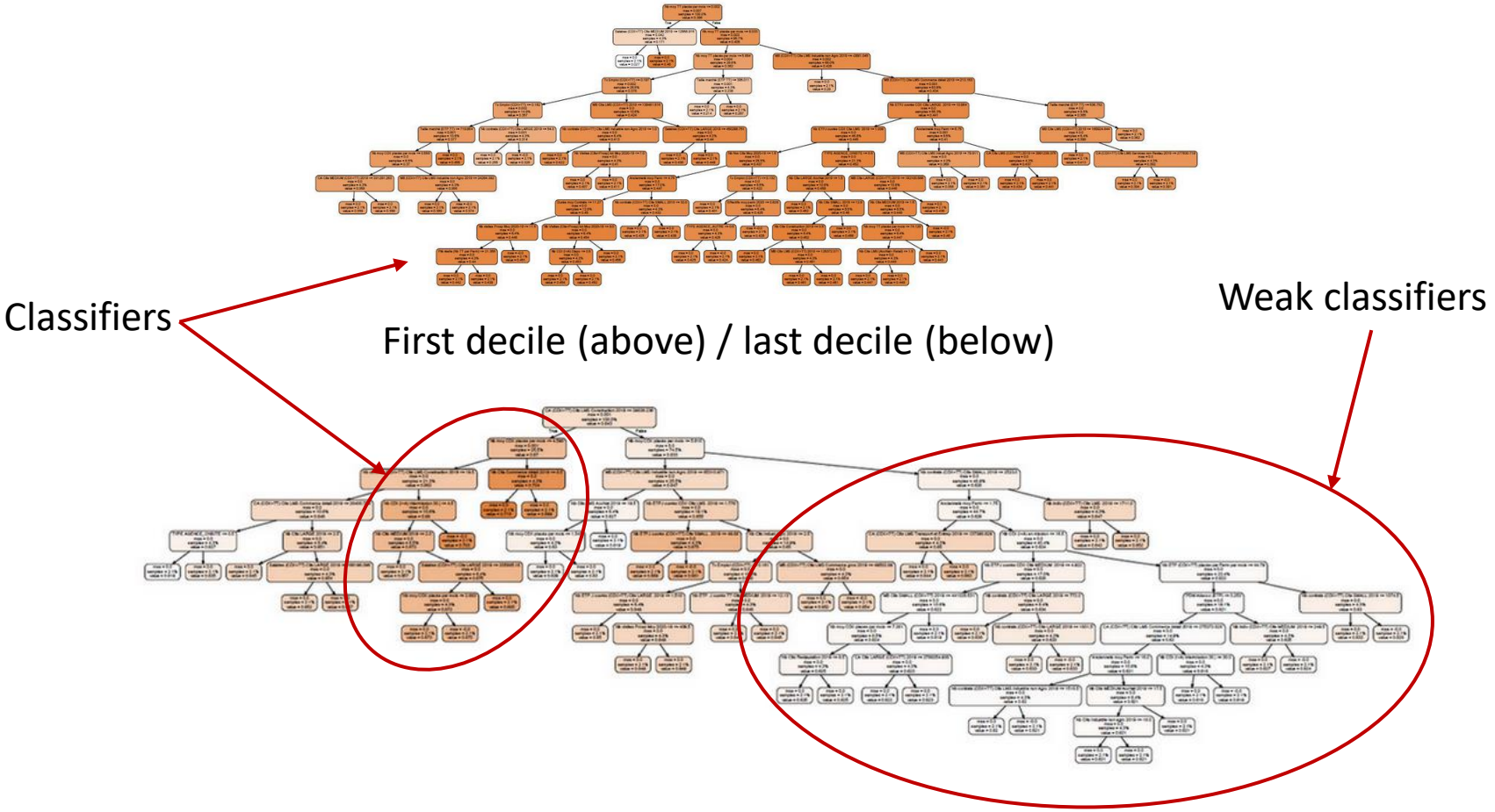


IQ&AI

Network Science and data elicitation

Generic effects on regression trees for BN in Bayesia Lab - decision support

Random tree forest regression on deciles – representative examples :

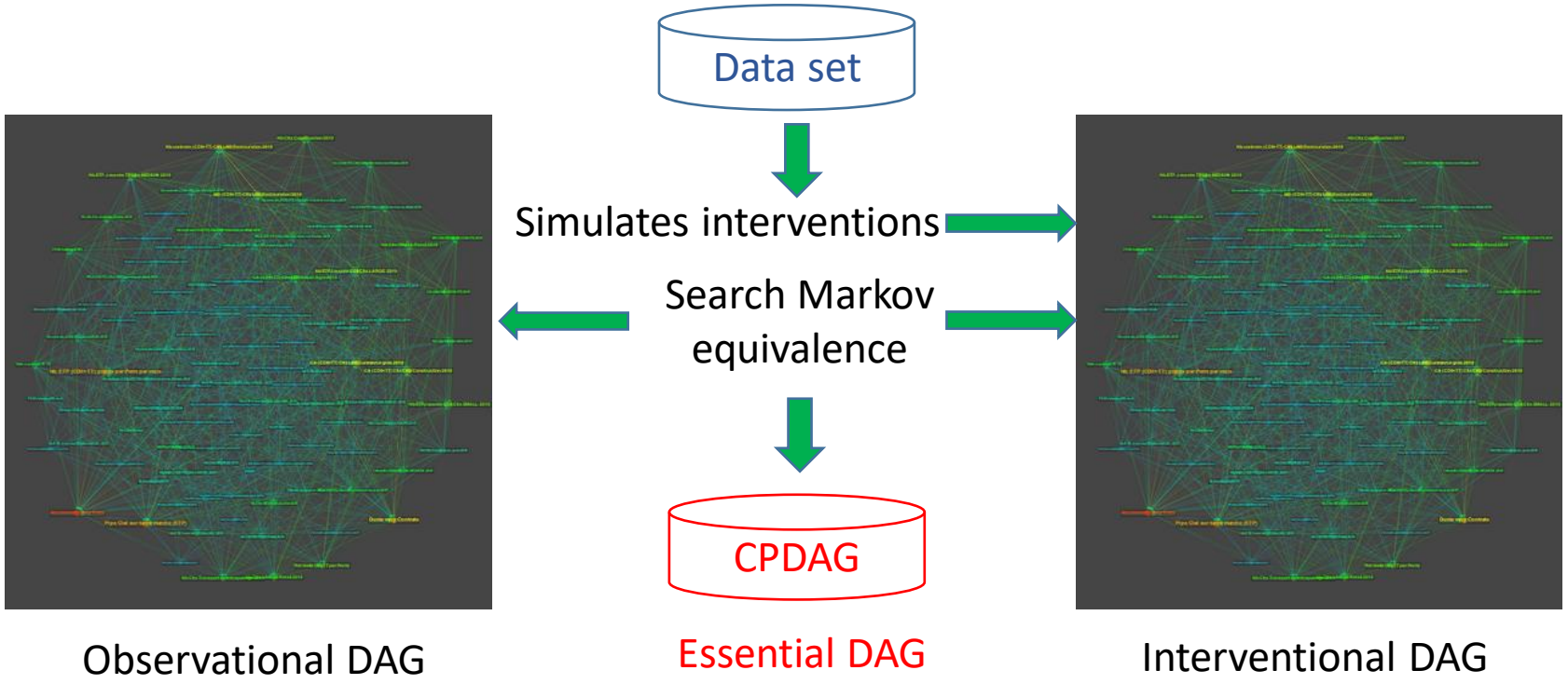


IQ&AI

Causal discovery in high dimensions

Following flows convergences through large graphs for pulling causal signals from noise

Step #1 : GIES - Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs [Hauser2012]

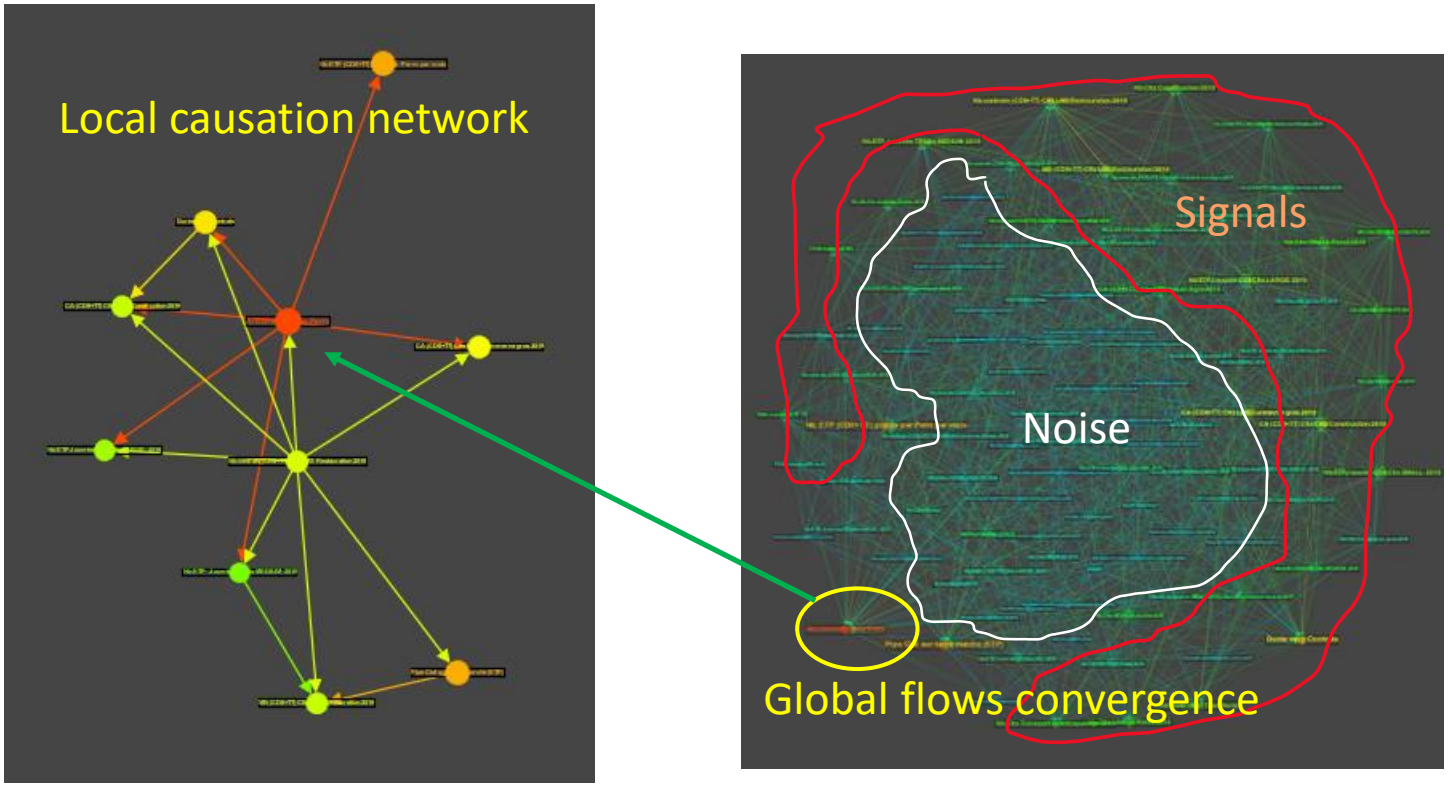


IQ&AI

Causal discovery in high dimensions

Following flows convergences through large graphs for
pulling causal networks from noise

Step #2 : using *P-rank* – [Yan2011] for
page-ranking essential DAGs with nodes in/out degree as a prominent value.



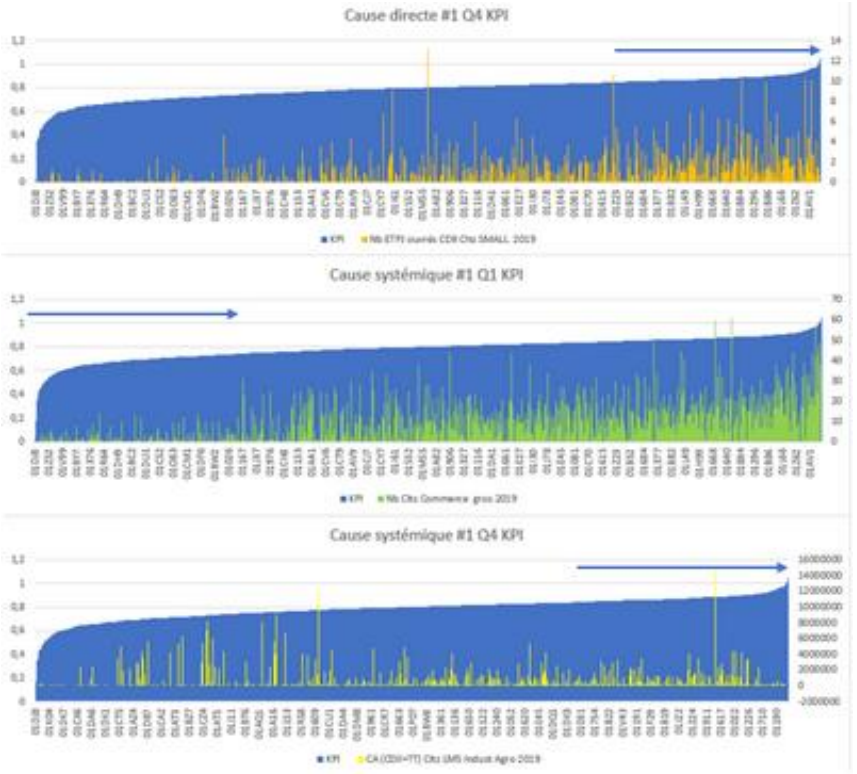
IQ&AI

Causal discovery in high dimensions

Experimentation: Top 1 **causes vs. KPI** on quartiles

Top ranked causes for quartiles Q1 and Q4 look coherent with our key performance indicator (KPI) for these quartiles, although they may be uncoherent for Q3.

For systemic causes - i.e. a specific variation not presented here -, when they influence or depend on a large *spectrum* of other nodes (causes and/or consequences in causal chains), direct correlation to KPI may become unsignificant and the determinating aspect – combinatorial – of these nodes may not be expressible from such a type of visual representation (cf. last graphic here).



Summary

I. Initial Approach

II. Bayesialab in the process : from data to causality

III. Network science's contributions

IV. Conclusions

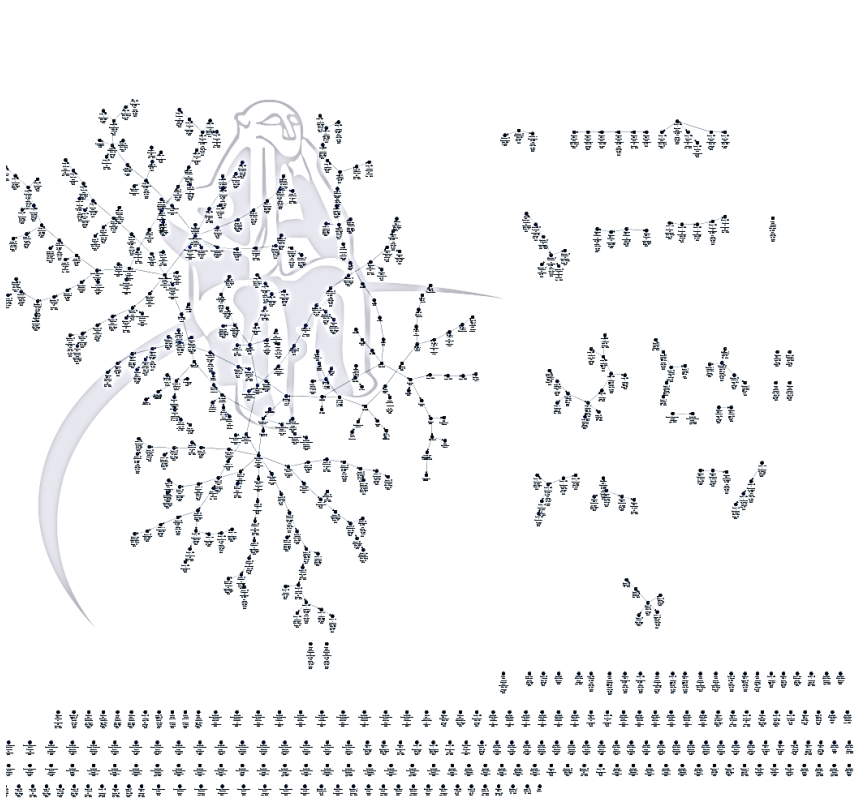
1. Dimensional view of the work & Questions to Blabers
2. Perspectives

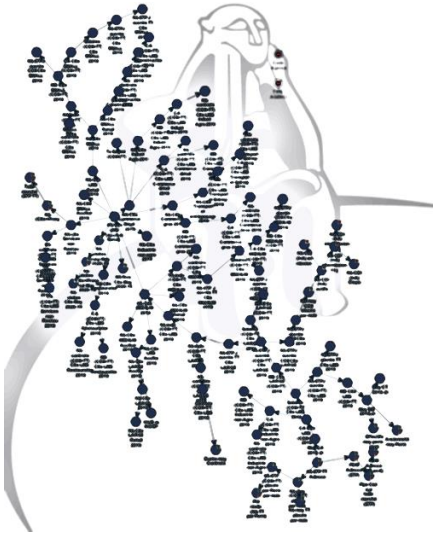
Joël Pain

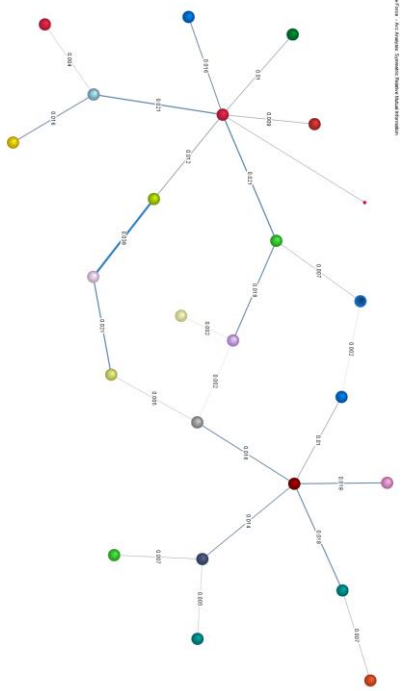
Christophe Thovex

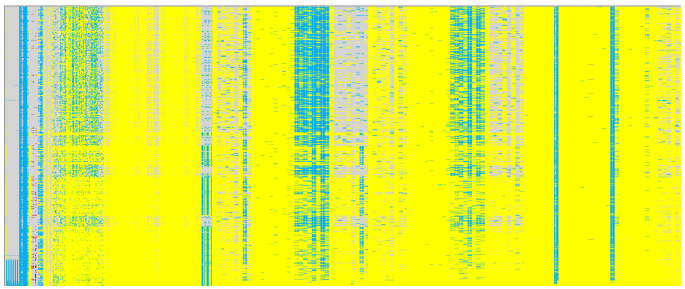
Emmanuel KEITA

Another view of the work

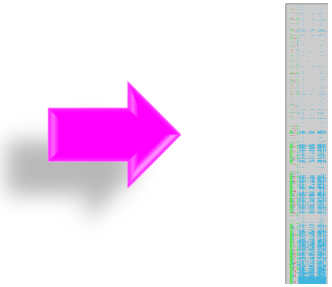




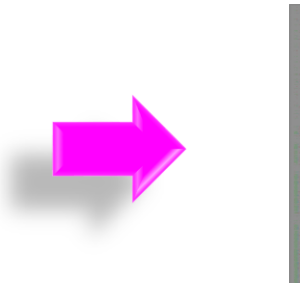




1032 rows x 1116 variables



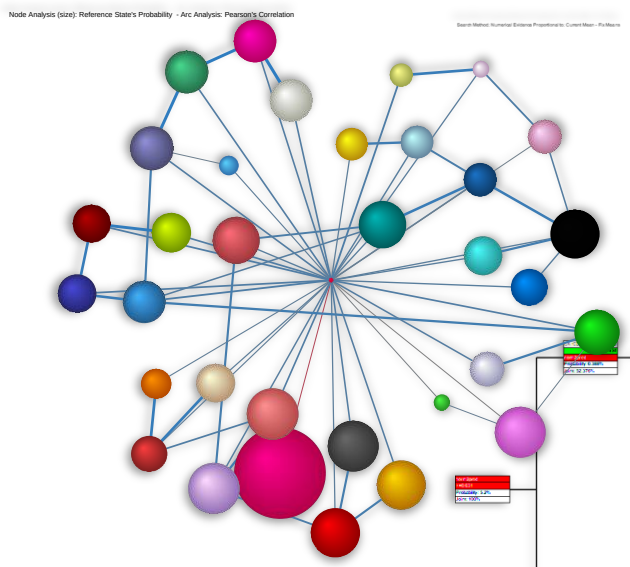
1023 rows x 106 variables



1023 rows x 23 Factors

IQ&AI

From observation to action : Enligthen the decision makers



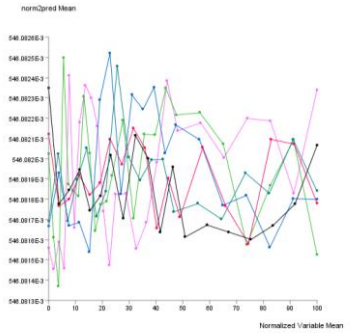
Best Solutions

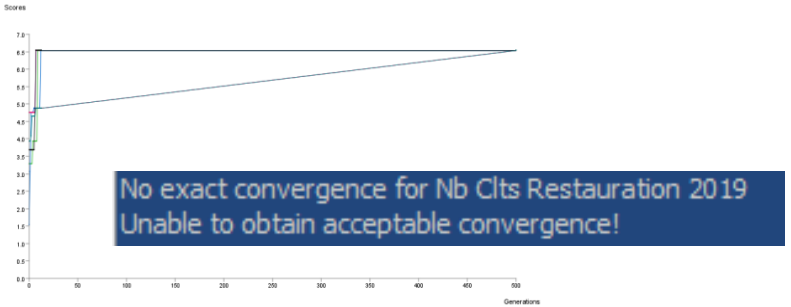
V1	V2	V3	V4	Nb of actions	Weight of Evidence	GBF
	<=26(1/4)			1	3.871	14.632
<=11(1/4)				1	3.516	11.442
<=11(1/4)	<=26(1/4)			2	3.416	10.672
			<=16.5(1/4)	1	3.369	10.33
	<=26(1/4)		<=16.5(1/4)	2	3.285	9.747
<=11(1/4)			<=16.5(1/4)	2	3.191	9.131
		<=0.856(1/4)		1	3.182	9.074
	<=26(1/4)	<=0.856(1/4)		2	3.147	8.855
<=11(1/4)	<=26(1/4)		<=16.5(1/4)	3	3.145	8.844
		<=0.856(1/4)	<=16.5(1/4)	2	3.099	8.568

IQ&AI

4 questions to Blabers

1) Switch to the « remove all Filtered States » option ?
2) Variable clustering : modifying clusters composition?
3) Convergence problems (direct effects, MRE)?





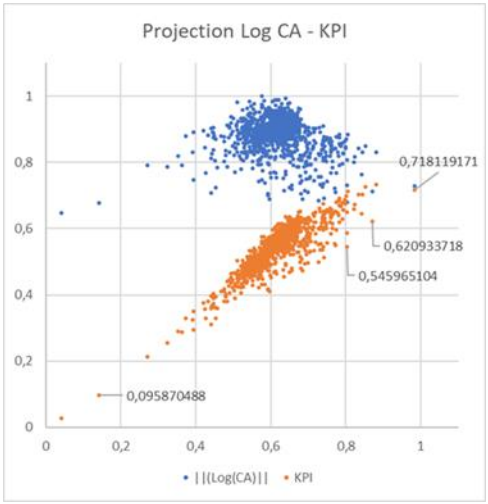
4) Explain in one sentence a GBF ?
5) Found a MDL score's local minimum for SPL, using different steps in the settings ?

[MDL_100steps > …> MDL_39steps > *MDL_40steps* < MDL_39 steps <….]

IQ&AI

Conclusion

- As showed, all these processes provided us with a lot of information, even if at the end we got more questions than answers…
- But it enabled us to look in the right direction, to study deeper some specific issues
- It raised many interrogations, it brought up many hypothesis, and it allowed us to ask the client “did you already consider this ?”, “have you already noticed that such variable could have an influence on the agencies performance”…
- Depending on the client feedback, we were then encouraged to deepen our analysis so as to confirm (or not) that a correlation we identified was a causality (or not)



- As a result, our approach proved efficient and brought some results the MNC found interesting and promising : we identified
  - a few “performance patterns”
  - some keys reasons that explain underperformance,
  - we draw their attention to some characteristics they hadn’t noticed previously,
  - We spotted some inconsistencies in their data which put some major weakness related to their data in evidence…
  - … and they seem happy with all that !
- We are currently still in the process of working with them, and we hope they will soon be willing to broaden the scope of analysis to the rest of the World and to external variables…
- …which is precisely our initial purpose : we intend to build a model that integrates all relevant parameters and that can be played with thanks to BayesiaLab ;-)

Thank you for your attention !

Joël Pain

Emmanuel Kéita

Christophe Thovex

Thank you for your attention !

Joël Pain
Christophe Thovex
Emmanuel KEITA