Text mining in KNIME

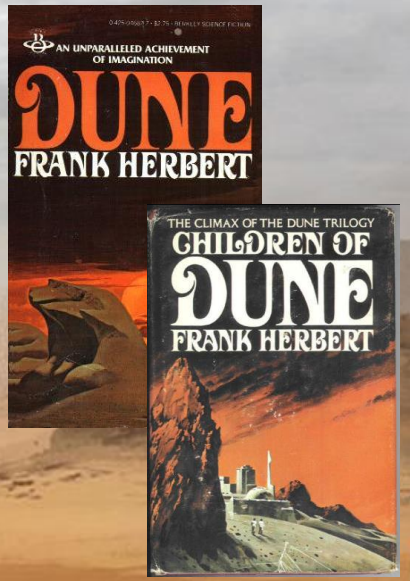
Peter Kuna | Business Analyst

© 2015 Infomatix B.V. All Rights Reserved

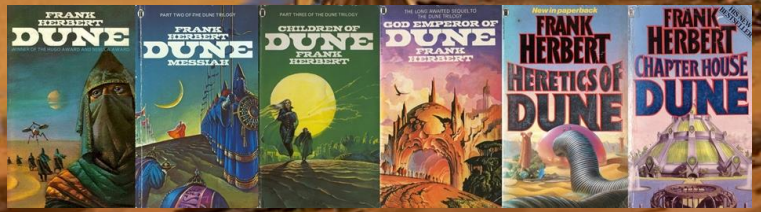
Dune series by Frank Herbert

Sci-fi series, Dates back to 1965

Over 20 books and other short stories



Lots of text to analyze.

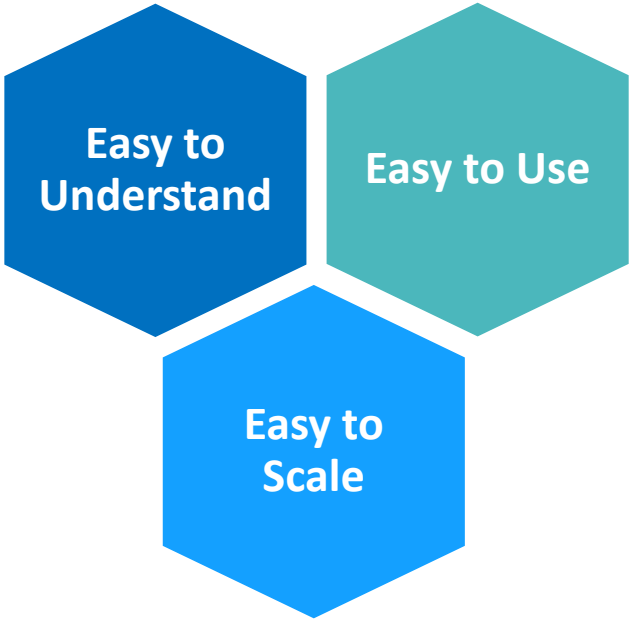


© 2015 Infomatix B.V. All Rights Reserved

…but what is KNIME?

Open for Innovation
KNIME

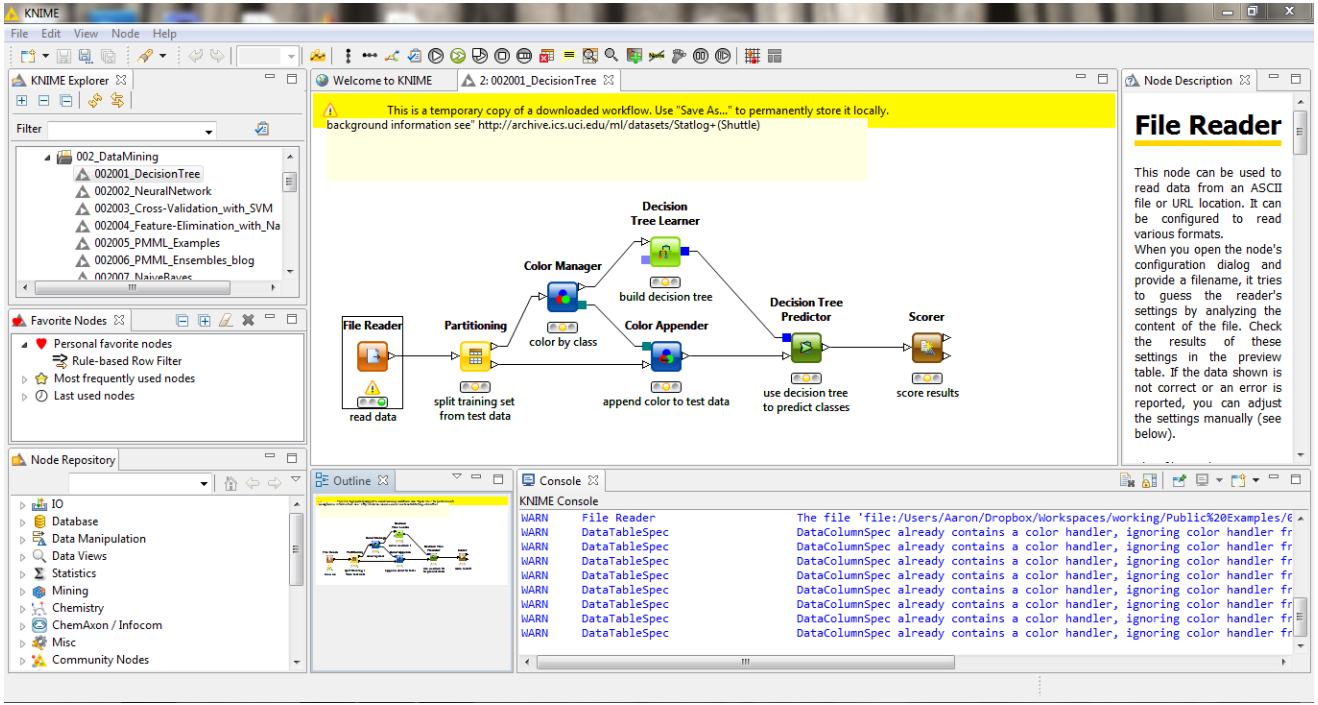
An open source workflow based data mining and analysis platform.



Data Access
Data Transformation
Data Mining
Advanced Analytics
Data Visualization
Data Exploitation

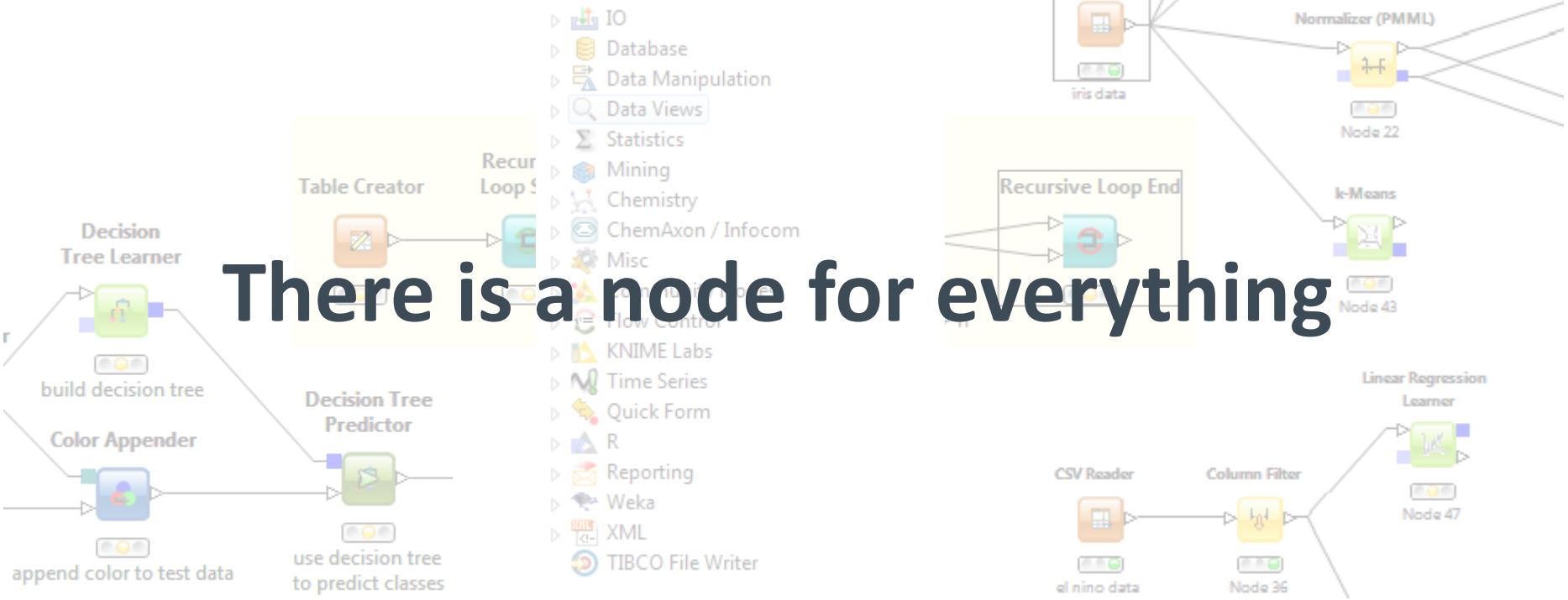
© 2015 Infomatix B.V. All Rights Reserved

THE INTERFACE AND THE LOGIC



© 2015 Infomatix B.V. All Rights Reserved

Talking about nodes…

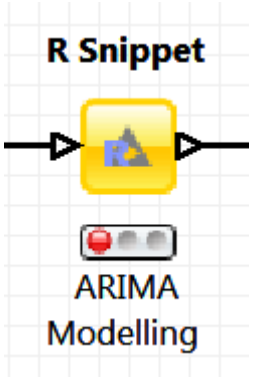


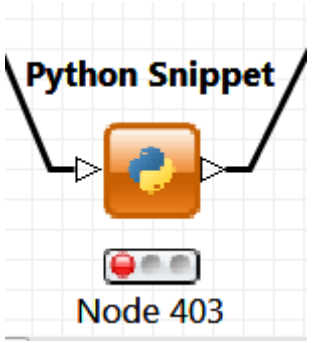
There is a node for everything

© 2015 Infomatix B.V. All Rights Reserved

Most importantly though

R, Python, Java integration!





All in all – lots of opportunities for advanced analysis.

© 2015 Infomatix B.V. All Rights Reserved

Typical challenges with advanced functionalities

Lack of knowledge

- I don't know what this can mean to me

Lack of confidence

- I am not sure I can manage this

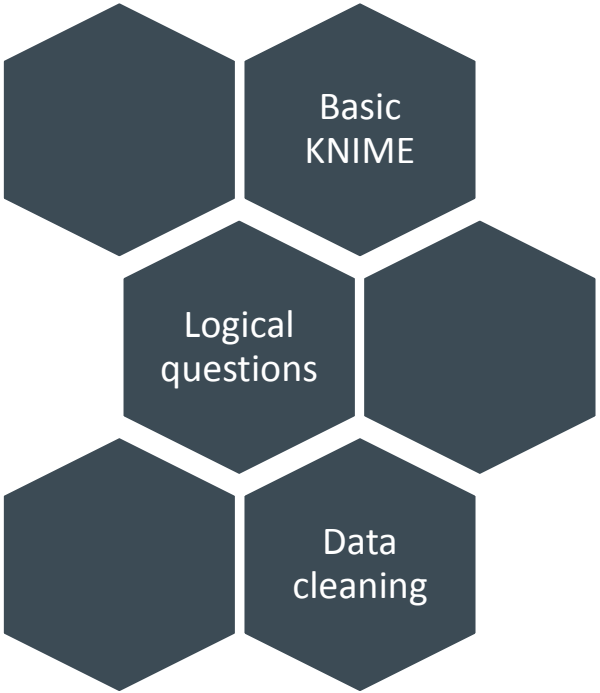
Lack of initiative

- I'm good with the current solution

In KNIME
most of the methods can be made simple and wanted.
Correlation, K-means, *seasonality* and..
TEXT MINING

© 2015 Infomatix B.V. All Rights Reserved

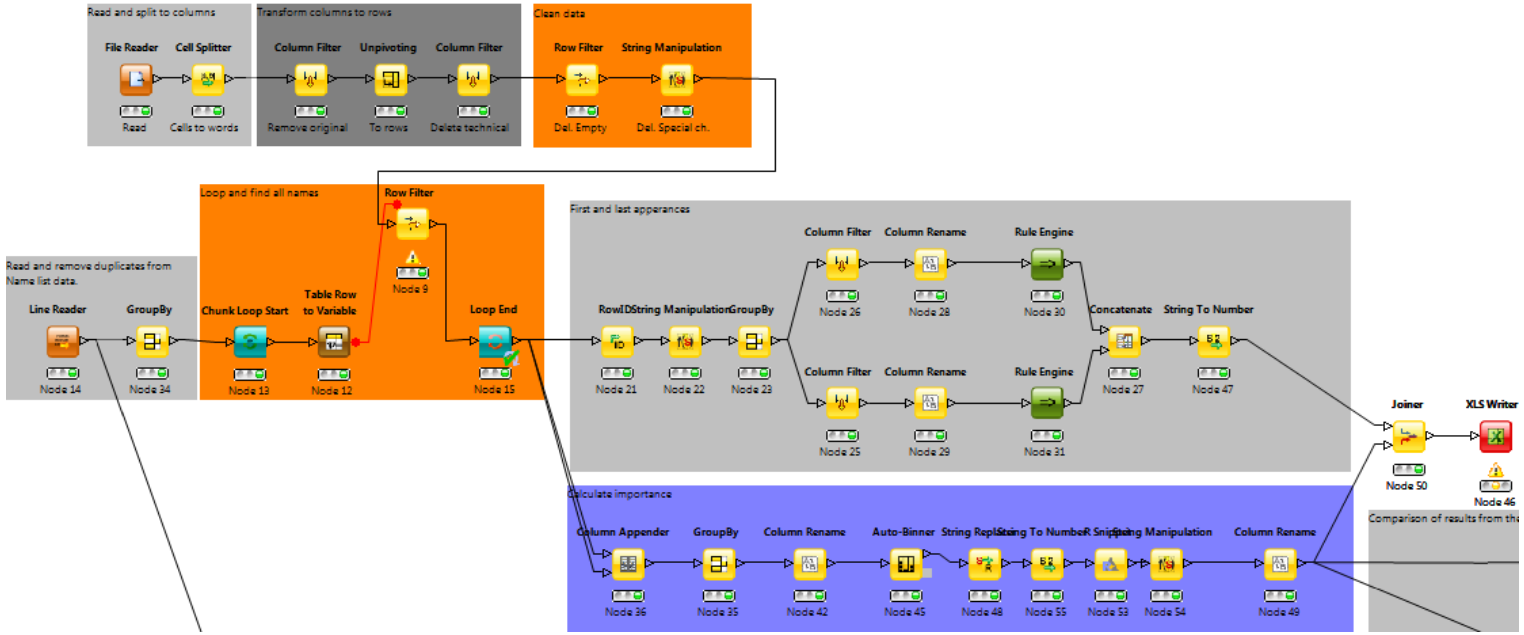
The KNIME "STARTER" KIT



- Nodes knowledge, what do connections mean..
- How important a fact is? How often does someone appear…
- Do I need punctuation? Capital letters…

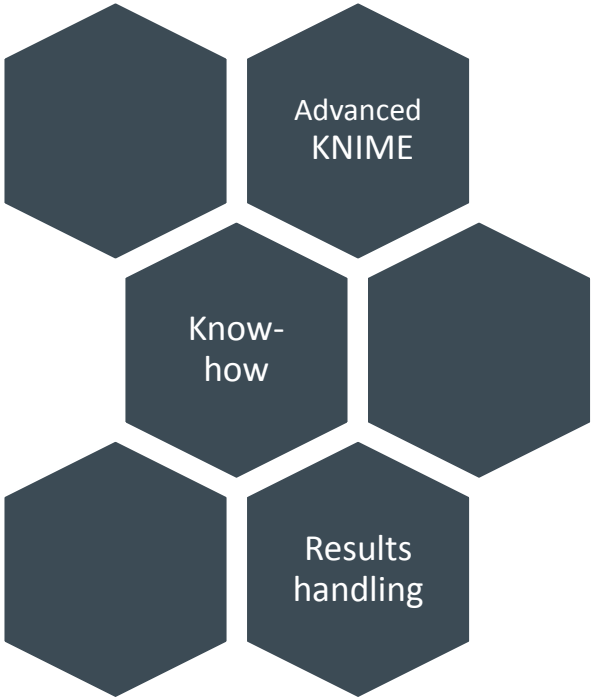
© 2015 Infomatix B.V. All Rights Reserved

HOW MY WORKFLOW LOOKS



© 2015 Infomatix B.V. All Rights Reserved

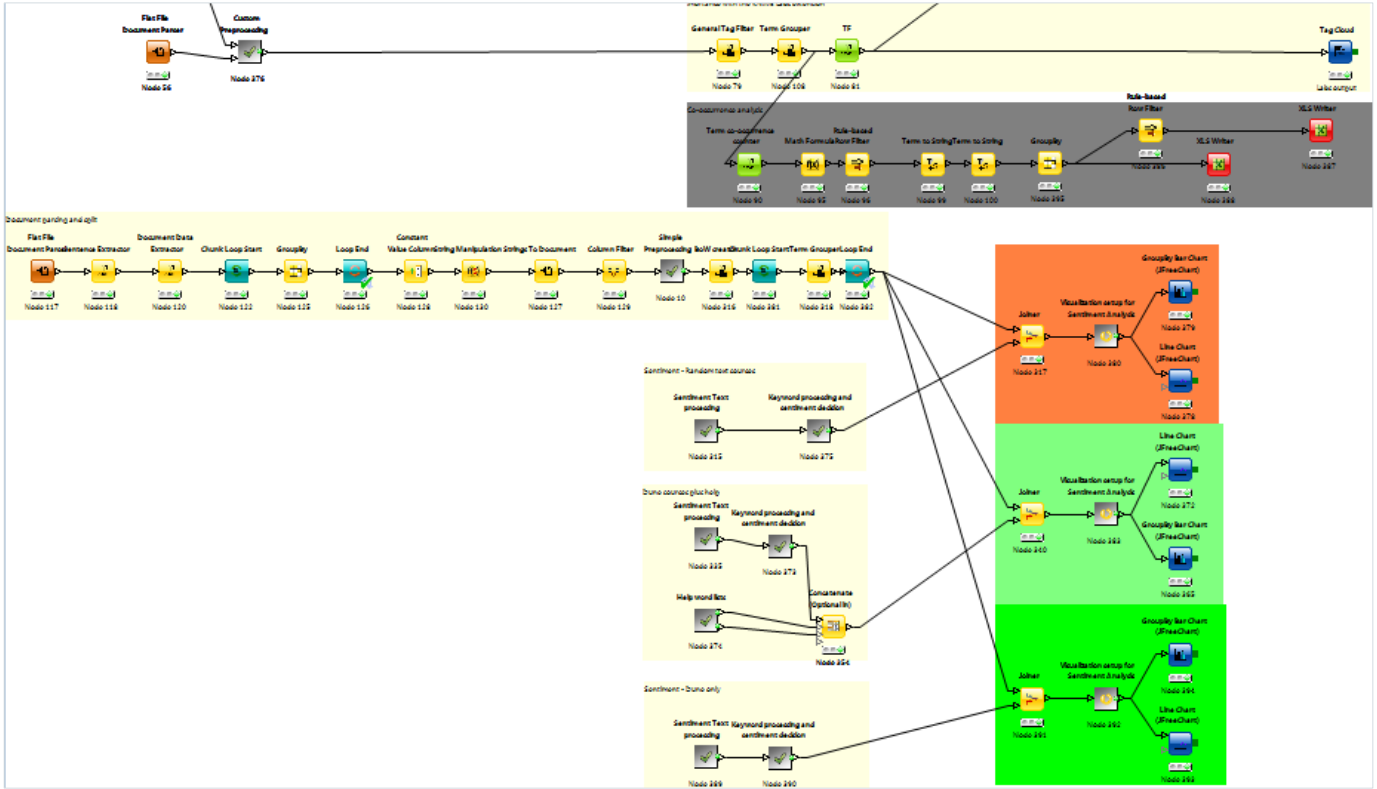
KNIME “INTERMEDIATE”



- Document handling, and specific types
- What nodes can be used – what do they mean
- Is it what I expected?

© 2015 Infomatix B.V. All Rights Reserved

HOW MY WORKFLOW LOOKS



© 2015 Infomatix B.V. All Rights Reserved

RESULT OF THE “BASICS”

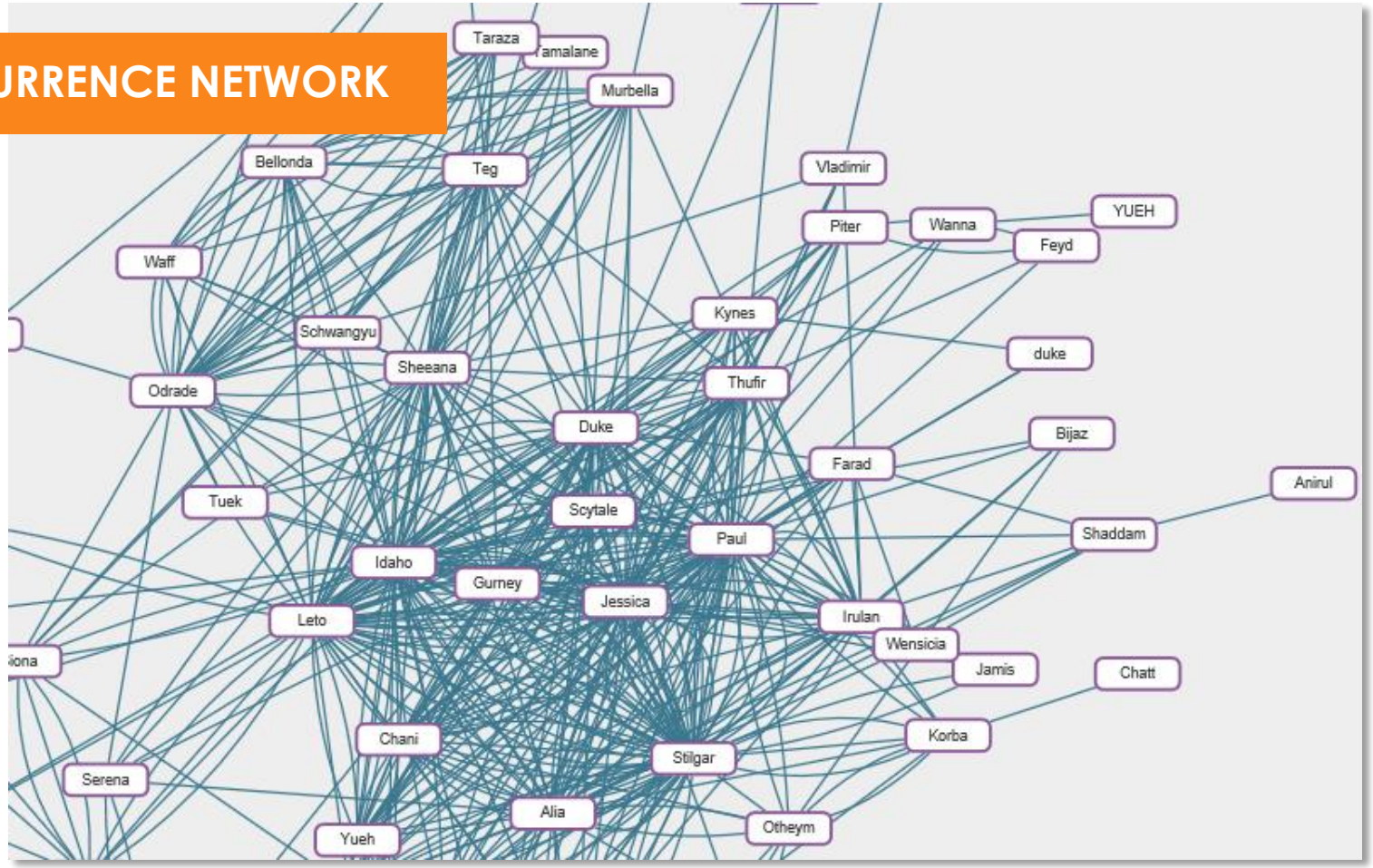
© 2015 Infomatix B.V. All Rights Reserved

THE TAG CLOUD

Aramsham
Chatt Lichna Anirul
Faroula Mohiam Bijaz Kobat Agamemnon
Geoff Ramallo Dama Thufir Wensicia Chenoeh Shoab
Kudu Shaddam Edric Tuek Chani Tamalane Serena Burah Orlop
Mattai Edrik Bellonda Duke Jessica Alia Schwangyu Otheym Xavier
Ghanima Moneo Leto Sheeana Idaho Yueh Piter Nefud
Hellica Scytale Murbella TegPaul WaffUxtal Omnius Margot
Paulus Jamis Gurney Siona Odrade Taraza Kynes Harah Wose Kaleff
Pook Tharthar Korba Hwi Stilgar Khrone Erasmus Mapes Syaksa
Aurelius Turok Wanna Malky Irulan Anteac Luyseyal Shishakli Ikanicre
Duro Ajax Farok Vladimir Logno Norma Helena
Elrood Ajidica Feyd Hasimir Gilbertus

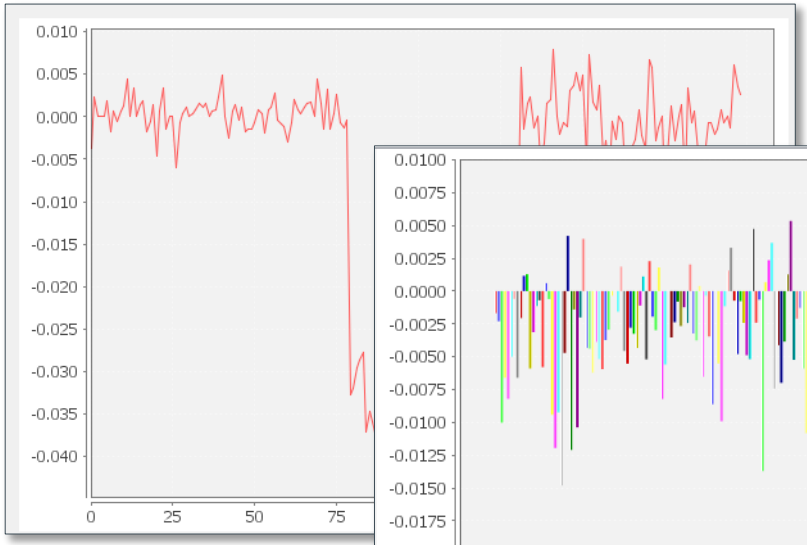
© 2015 Infomatix B.V. All Rights Reserved

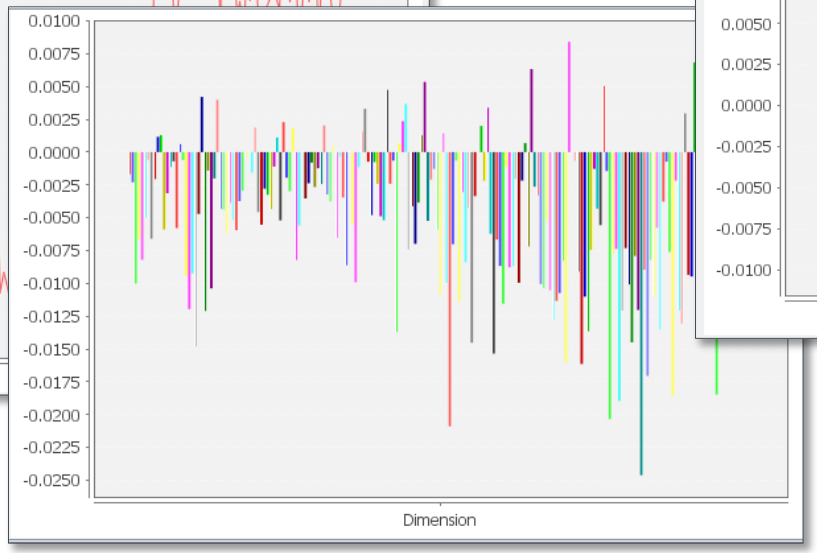
THE CO-OCCURRENCE NETWORK

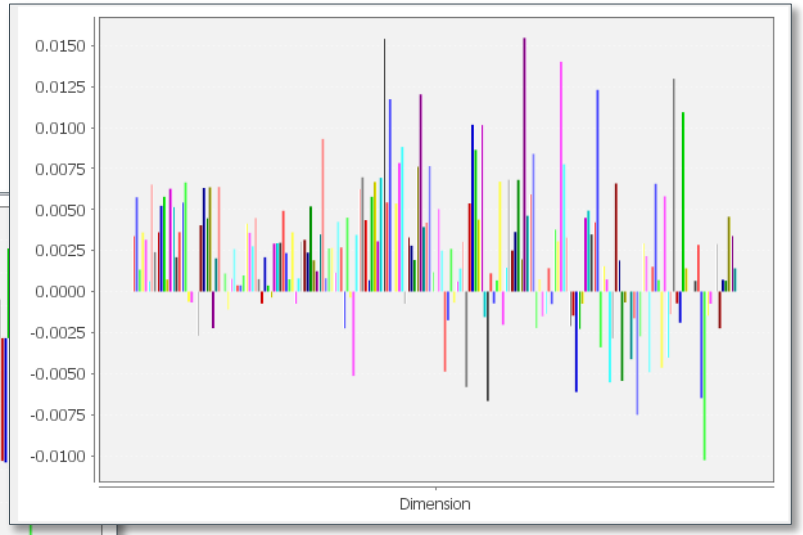


© 2015 Infomatix B.V. All Rights Reserved

THE SENTIMENT







© 2015 Infomatix B.V. All Rights Reserved

Peter Kuna | Business Analyst | pkuna@infomatix.net | www.infomatix.net

InfomatiX
Clarity from Complexity

© 2015 Infomatix B.V. All Rights Reserved

Let’s start the HARVEST