Valverde, F.J., González, J.M., Peñas, A., & Peláez, Carmen. (2016). Supporting scientific knowledge discovery with extended, generalized Formal Concept Analysis. *Expert Systems with Applications*, 44, pp. 198-216.

DOI:10.1016/j.eswa.2015.09.022

Supporting scientific knowledge discovery with extended, generalized Formal Concept Analysis

Francisco J. Valverde-Albacete[a], José María González-Calabozo[b], Anselmo Peñas[a], Carmen Peláez-Moreno[b,*]

[a] Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid 28040, Spain
[b] Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Leganés 28911, Spain

ABSTRACT

Keywords:
Scientific knowledge discovery
Exploratory Data Analysis
Landscapes of Knowledge
Metaphor theory
Formal Concept Analysis
$\mathcal{K}$-Formal Concept Analysis
Extended Formal Concept Analysis
Semiring theory
Confusion matrix
Relation extraction
Gene expression data

In this paper we fuse together the Landscapes of Knowledge of Wille's and Exploratory Data Analysis by leveraging Formal Concept Analysis (FCA) to support data-induced scientific enquiry and discovery.

We use extended FCA first by allowing $\mathcal{K}$-valued entries in the incidence to accommodate other, non-binary types of data, and second with different modes of creating formal concepts to accommodate diverse conceptualizing phenomena.

With these extensions we demonstrate the versatility of the Landscapes of Knowledge metaphor to help in creating new scientific and engineering knowledge by providing several successful use cases of our techniques that support scientific hypothesis-making and discovery in a range of domains: semiring theory, perceptual studies, natural language semantics, and gene expression data analysis.

While doing so, we also capture the affordances that justify the use of FCA and its extensions in scientific discovery.

1. Introduction

Formal Concept Analysis (FCA) can be understood as an unsupervised analysis technique for matrices with boolean entries (Ganter & Wille, 1999). The underlying theoretical construct is a Galois Connection induced by a relation $I \subseteq 2^{G \times M}$ between a set of objects G and a set of attributes M[1]. In this case, the Galois connections adopts the guise of a pair of functions, the *polars*, induced by the relation transforming sets of objects into sets of attributes and vice versa: pairs of sets of objects and attributes that are mutually transformed are called *formal concepts* and the inclusion order of object sets (dually, that of attribute sets) turns out to be a complete lattice, providing an excellent canvas against which to cast many of the properties hidden in the data. A lightweight introduction to FCA can be found in Section 2.1.

In spite of the width of the work trying to make FCA a sound technique for the exploratory analysis of knowledge, as reviewed, for instance in Poelmans, Kuznetsov, Ignatov, and Dedene (2013b), Poelmans, Ignatov, Kuznetsov, and Dedene (2013a), it has found but a feeble echo in the Statistical and Machine Learning communities.

Tukey was the figure who crystallized the cry for exploratory, discovery-driven methods in the Statistics community. He called his proposal *Exploratory Data Analysis* (EDA) and advocated a complementary curriculum for Statisticians balancing EDA against Statistical Hypothesis Testing (that he called Confirmatory Data Analysis) (Tukey, 1980). At present, EDA is standard practice of good statistics, it is taught at basic statistics courses in academia (Tukey, 1977), and supported by widely-used data processing software (Matlab, 2012; R Core Team, 2014).

FCA falls under the definition of an EDA technique for boolean matrices, yet very few work in the FCA community or otherwise, highlights this fact. Furthermore, extensions to FCA exist that allow it to work with more quantitative data like FCA in a fuzzy setting (Bělohlávek, 2002) or $\mathcal{K}$-FCA (Valverde-Albacete & Peláez-Moreno, 2006) where the numeric data take values in different kinds of semirings.

1.1. *(Formal Conceptual) Landscapes of Knowledge*

One of the founders of FCA, Wille, a mathematician and philosopher, developed in the late 1990s a program for knowledge

[*] Corresponding author. Tel.: +34 916248771; fax: +34 916248749.
E-mail addresses: fva@lsi.uned.es (F.J. Valverde-Albacete), melopsitaco@gmail.com (J.M. González-Calabozo), anselmo@lsi.uned.es (A. Peñas), carmen@tsc.uc3m.es (C. Peláez-Moreno).

[1] *G* for German *Gegenstunde* "object" and *M* for German *Merkmale* "attribute" are customary in FCA theory.

Table 1
The activities supporting the Landscape of Knowledge metaphor, their descriptions (from (Wille, 1999)) and some aliases that we use for them. The last two rows are our proposal for inclusion in the list.

Activity	Description	Aliases
Focusing	Selecting what one wants to look into.	*contextualizing*
Exploring	"...looking for something of which one only has a vague idea..."	*browsing*
Searching	"...looking for something which one can more or less specify but not localize..."	
Recognizing	"...perceiving clearly circumstances and relationships..."	*Clarifying*
Identifying	"...determining the taxonomic position of an object within a given classification..."	*categorizing, placing in a taxonomy*
Analyzing	"...examining data in their relationships while guided by the theoretical views and declared purposes..."	
Investigating	"...study by close examination and systematic enquiry..."	
Deciding	"...resolving a situation of uncertainty by an order..."[a]	
Improving	"...enhancement in quality and value. ..."	*distilling, valorizing*
Restructuring	"...to reshape a given structure, which, within the scope of our discussion, is conceptual in its nature..."	
Memorizing	"...a process of committing and reproducing what has been learned and retained. ..."	*committing to memory*
Hypothesizing	Abducing facts and relationships based on data.	*forming an opinion*
Indexing	Indexing other knowledge resources using concepts, objects or attributes	*interfacing resources*

[a] There is a possible ambiguity here in the meaning of *order*, in spite of Wille's being an *order mathematician* we have decided to understand here *command*.

discovery based in FCA. He baptized it the "Conceptual Landscapes of Knowledge Paradigm" (LofK) but we believe it can be better understood as a *metaphor* in the spirit of Lakoff and Johnson (1996).

Metaphorical interpretation eases the comprehension of an unfamiliar domain of knowledge in terms of an already-known domain: the entities and issues in the new domain are mapped onto the well-known domain to make them more apprehensible. For a review of tenets and notation in metaphor theory see Lakoff (1987), Lakoff and Johnson (1996).

The basic metaphor in LofK seems to be

Metaphor 1. UNDISCOVERED KNOWLEDGE IS UNCHARTED TERRITORY.

and is founded in the analogy between the investigation of hitherto unknown knowledge and the physical exploration of non-mapped terrain. The analogy is reinforced by a series of subsidiary metaphors like

Submetaphor 2. THE INTRICACIES OF KNOWLEDGE ARE FEATURES OF THE LANDSCAPE.

This emphasizes that navigating the nitty-gritty details of knowledge is like exploring difficult to access places.

Once the initial identification of knowledge and terrain is accepted, Wille proposes that FCA can help in fleshing out this metaphor by providing a physical embodiment for knowledge that can be acted upon as if it were a *map of the terrain*. The first step of this exploration is to explicitly highlight the data being considered by forming a *formal context* which is a two-mode relation between objects and attributes. This action is called *contextualizing*:

Metaphor 3. BUILDING A FORMAL CONTEXT IS SETTING YOURSELF ON A VANTAGE POINT ON THE KNOWLEDGE LANDSCAPE.

At the same time, contextualizing sets the limits of the knowledge you can see, since you choose to concentrate on some information to the detriment of some other. This is further emphasized by the context-lattice duality of FCA and the following

Metaphor 4. A CONCEPT LATTICE IS A MAP OF THE KNOWLEDGE LANDSCAPE.

The different operations on the formal context would subsequently represent acting on the territory being charted, with the idea of using the concept lattice to explore the unknown territory until it becomes tame.

Table 1 lists the knowledge-exploring activities supported by these metaphors. Most of these activities are expressed in either the main or the secondary domain of the metaphor, like "browsing" in the physical domain, and "analyzing" in the cognitive domain. Note that these metaphors have already been put to the test in the building of semantic file systems (Martin, 2004; Martin & Eklund, 2005) and Logical Information Systems (Ferré, 2007; Ferré & Ridoux, 2001) as Information Retrieval systems. A thorough review of the use of FCA to model knowledge is Poelmans et al. (2013b), while applications are detailed in Poelmans et al. (2013a).

In later work, Wille has concentrated in developing these metaphors with a view towards making them *actionable* by providing epistemic interpretation for the constructs of FCA (Wille, 2006). But even his early papers show usage of the actions enunciated in Table 1 or precursors of them.

1.2. FCA as an instance of Exploratory Data Analysis

We believe that the following metaphor captures one of the tenets of exploratory analysis:

Metaphor 5. KNOWLEDGE IS AN EXHIBIT

We refer here to "exhibit" much as in a work of art in a museum or an item in a shop. The idea is that knowledge can be taken "out in the open" to be shared, subjected to public scrutiny and commentary.

In our experience, FCA is a method of EDA (Tukey, 1977) that supports the knowledge-as-exhibit metaphor through two submetaphors:

Submetaphor 6. KNOWLEDGE CAN BE VISUALIZED.

This is achieved through the Hasse diagram of the concept lattice, while

Submetaphor 7. KNOWLEDGE CAN BE (BODILY) MANIPULATED.

is achieved through (a) the polars that enable the transformation of objects and attributes into formal concepts, and (b) the operations between formal concepts allowing us to obtain more general or more specific ones.

A summary of metaphors to be fleshed out can be consulted in Table 2.

The first endeavor of this paper is to show how FCA as interpreted in the LofK metaphors can be leveraged as an EDA technique for scientific and engineering discovery by helping with the elicitation, structuring and manipulation of knowledge. The first result of this endeavor in this paper is *a methodology to carry out exploratory analysis of two-mode data with (extended, generalized) FCA*, which we put to the test and exemplify in Section 3.

1.3. Extensions to Formal Concept Analysis

Unfortunately, the fact that *standard* FCA can only deal with boolean incidences often prevents scientists with other type of data

Table 2
Summary of the metaphors in LofK considered in this paper. Sub-metaphors are further indented.

Metaphor number	Statement
1	Undiscovered knowledge is uncharted territory.
2	The intricacies of knowledge are features of the landscape.
3	Building a formal context is setting yourself on a vantage point on the knowledge landscape.
4	A concept lattice is a map of the knowledge landscape.
5	Knowledge is an exhibit.
6	Knowledge can be visualized.
7	Knowledge can be (bodily) manipulated.

from resorting to these techniques. First and foremost it must be stated that many-valued extensions to FCA have existed for some 20+ years. A review of extensions can be found in Poelmans et al. (2013a,b).

In this paper we are using both standard FCA and $\mathcal{K}$-Formal Concept Analysis, an analysis technique for matrices with $\mathcal{K}$-valued entries where $\mathcal{K}$ is an *idempotent semifield* (Valverde-Albacete & Peláez-Moreno, 2011), a type of structure that is well-known from Morphological Processing (morphological calculus), Artificial intelligence (path algebras) and Graph Theory (the Viterbi algorithm). $\mathcal{K}$-FCA has already been successfully used for the analysis of confusion matrices (Peláez-Moreno, García-Moral, & Valverde-Albacete, 2010), Gene Expression Data (GED) (González-Calabozo, Peláez-Moreno, & Valverde-Albacete, 2011) and to model the batch retrieval task in Information Retrieval (Pedraza-Jiménez, Valverde-Albacete, & Navia-Vázquez, 2006).

Some concern may be raised as to how FCA provides a *biased* manner of eliciting and manipulating knowledge: the formal concepts spoken of so far aggregate objects and attributes "in the positive", meaning they capture incidence of objects onto attributes an vice-versa. But there are other modes of incidence, for instance objects that lack some attributes or attributes that lack some objects. In Valverde-Albacete and Peláez-Moreno (2011) the standard modes of conceptualization of FCA in the multi-valued setting have been formalized and examined to show that they share many of the desirable properties of standard FCA.

A second endeavor of this paper is to show *how multi-valued and non-standard extensions of FCA dovetail into the standard theory to complement and eradicate its shortcomings on multi-valued or special dataset instances.* To this purpose we provide extensive evidence in Section 3.

1.4. Reading guide

In Section 2.1 we first revisit the basics of FCA, and then its multi-valued and generalized conceptualization mode-enabled extensions in Sections 2.2 and 2.3, respectively.

In Section 3 we present our main contributions:

- An analysis and listing of the main affordances of extended, generalized, multi-valued FCA as an embodiment of LofK for EDA support in scientific discovery.
- A methodology for using FCA in many numeric or boolean knowledge EDA, instantiated and proven over a number of tasks.

We end the paper with a Discussion and some Conclusions.

2. Materials and methods

2.1. A crash course on Formal Concept Analysis

FCA is a procedure to render lattice theory more concrete and manipulative (Ganter & Wille, 1999). It stems from the realization that a binary relation between two sets $I \in 2^{G \times M}$—where G and M are conventionally called the set of objects and attributes, respectively—defines a Galois connection between the powersets $X \equiv 2^G$ and $Y \equiv 2^M$ endowed with the inclusion order (Erné, Koslowski, Melton, & Strecker, 1993). The triple $\mathbb{K} = (G, M, I)$ is called a *formal context* and the pair of maps that build the connection are called the *polars (of the context)*:

$$\forall A \in 2^G, A^\uparrow = \{m \in M \mid \forall g \in A, gIm\} \qquad (1)$$
$$\forall B \in 2^M, B^\downarrow = \{g \in G \mid \forall m \in B, gIm\}.$$

Pairs of sets of objects and attributes that map to each other are called *formal concepts* and the set of formal concepts is denoted by

$$\mathfrak{B}(G, M, I) = \{(A, B) \in 2^G \times 2^M \mid A^\uparrow = B \wedge A = B^\downarrow\}.$$

The set of objects of a concept is called its *extent* while the set of attributes is called its *intent*. Formal concepts are partially ordered by the inclusion (resp. reverse inclusion) of extents (resp. intents)

$$c_1 = (x_1, y_1), c_2 = (x_2, y_2) \in \mathfrak{B}(G, M, I) \quad c_1 \leq c_2 \Leftrightarrow x_1 \subseteq x_2$$
$$\Leftrightarrow y_1 \supseteq y_2 \qquad (2)$$

With the concept order, the set of formal concepts $\langle \mathfrak{B}(G, M, I), \leq \rangle$ is actually a complete lattice called the *concept lattice* $\underline{\mathfrak{B}}(G, M, I)$ *of the formal context* (G, M, I).

Most concept lattice-building algorithms available output *Hasse diagrams* developed to easily describe partial orders. Concept lattices can profitably be represented and grasped in such form: nodes in the diagram represent concepts, and the links between them the hierarchical partial order between immediate neighbors. For instance, Fig. 1 is an example of concept lattice where the objects are algebraic structures and the attributes the properties they fulfill.

For the purpose of reading extents and intents off the lattice diagram, concepts could be annotated graphically with a *complete labeling*, by listing for each concept the set of object labels in the concept extent and the set of attribute labels in the concept intent. But since this implies repeating many times each object and attribute throughout the lattice the following, *reduced labeling* is preferred: we put the label of each attribute only in the highest (most abstract) concept it appears, and the label of each object only in the lowest (most specific) concept it appears. Attribute labels usually appear *just above* the corresponding concept and object labels appear *just below* and each only once for the whole lattice, diminishing the visual clutter. This is quite straightforward to accomplish if we consider mappings $\gamma : G \to \underline{\mathfrak{B}}(\mathbb{K})$ and $\mu : M \to \underline{\mathfrak{B}}(\mathbb{K})$ obtaining *object and attribute concepts*, respectively: we put each object and attribute label in the concept obtained by means of these maps. This is the type of labeling shown throughout the paper and the most usual, though different lattice-building tools use variations of it.

The following notions can also be read straightforwardly from the Hasse diagram of the concept lattice:

- The order between concepts described in Eq. (2) is consistent with the inclusion of extents and the reverse inclusion of intents.
  The intuition behind this order is that a concept (A, B) is "less defined" than another concept (C, D), equivalently higher in the diagram, if it has more objects in its extent or less attributes (read *properties*) in its intent, in complete agreement with traditional extensional-intensional distinctions in logic.

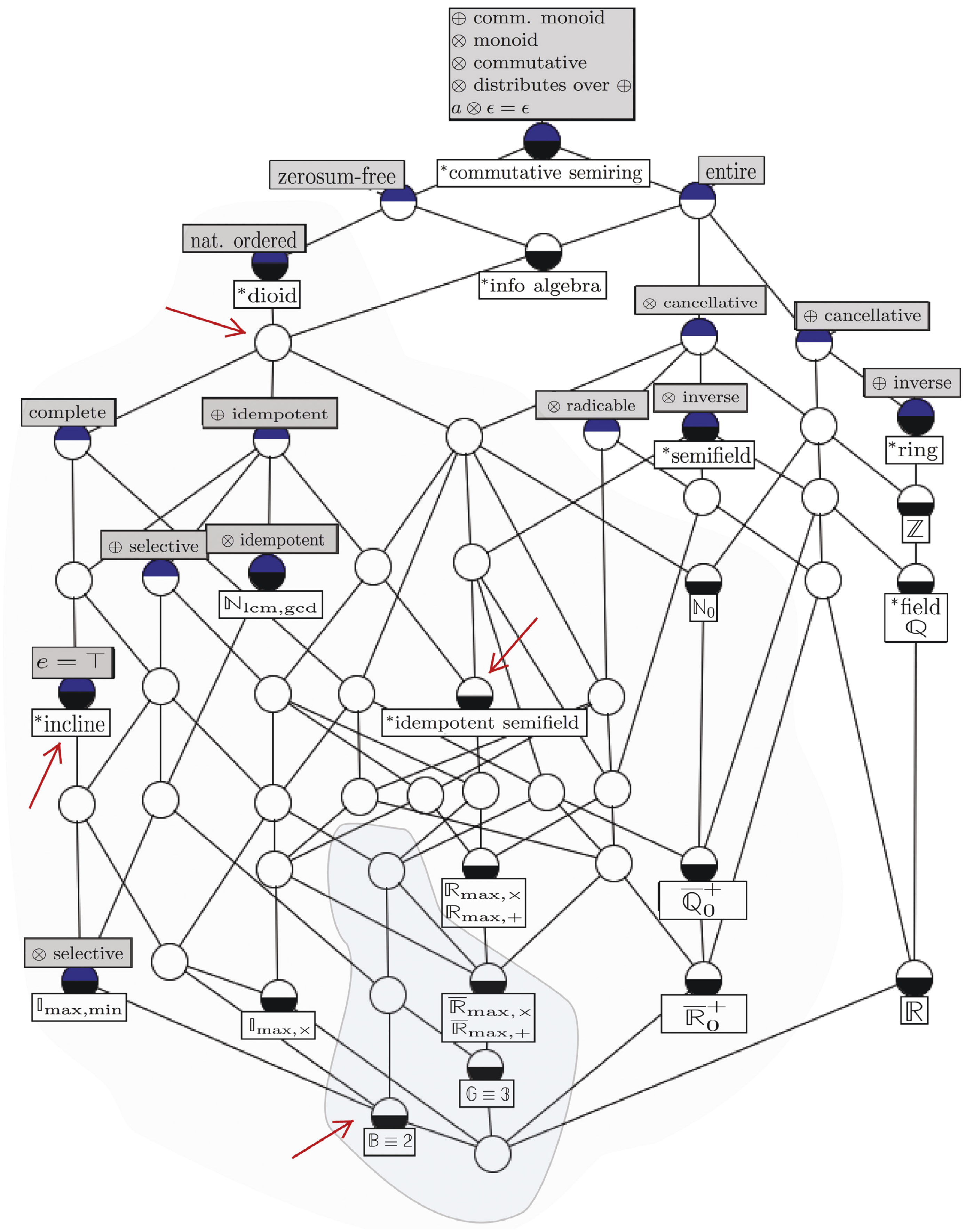


Fig. 1. (Color online) Concept lattice of dioids in the context of commutative semirings. Each node is a concept of abstract algebra: its properties are obtained from the gray labels in nodes upwards, and its structures from the white labels in nodes downwards. The picture is related to the *chosen* sets of properties and algebras and does not fully reflect the structure of the class of semirings .

- A *top concept*, $\top = (G, G^{\uparrow})$, whose extent includes all objects and whose intent includes all attributes predicable of all objects (the most general concept).
- A *bottom concept* $\bot = (M^{\downarrow}, M)$, whose intent includes all attributes and whose extent includes all objects which enjoy all attributes (the least general concept).
- The two total binary operations making the set a lattice can be calculated from the diagram:
  - The *join* of two concepts, $c_1 \vee c_2$, by following the links *upwards* from them to find the lowest concept which is more abstract than either, their *lowest upper bound*.
  - The *meet* of two concepts, $c_1 \wedge c_2$, by following the links *downwards* starting from them to find the highest concept which is less abstract than either of them, that is their *highest lower bound*.
  - The set of *join-irreducibles*, that is the set of concepts which can be *join*-ed to obtain all other concepts. These are normally depicted as nodes bottom-half filled in black.
  - The set of *meet-irreducibles*, that is, the set of concepts which can be *meet*-ed to obtain all other concepts. These are normally depicted as nodes top-half filled in blue.

  These two procedures can be generalized to any number of concepts, given that the lattice is complete.

Besides the information given by its Hasse diagram as a generic ordered set, the diagram of a concept lattice offers the following items of information:

- It equates the order expressed by the lattice with the *order of generalization* between concepts. Hence the top is the *most general concept*, while the bottom is the *least general concept*. Essentially, by *navigating* the diagram links we are generalizing formal concepts when going upwards and specializing them when going downwards.
- It equates every element in the lattice (node in the Hasse diagram) with a formal concept, (A, B) whose intent and extent are easily accessible: the extent is found accumulating all the object labels in all walks from the concept node to the bottom, while the intent is found by accumulating all attribute labels from the concept node to the top.
- Finally, for a concept lattice, the *distribution of crosses* in I essentially imposes the *shape* of the lattice $\mathfrak{B}(G, M, I)$ while the sets of objects and attributes (and the intuitive meaning of the incidence, of course) allow us to *interpret* the lattice as referring to properties (attributes) of some entities (objects)[2].

2.2. $\mathcal{K}$-Formal Concept Analysis

$\mathcal{K}$-FCA (Valverde-Albacete & Peláez-Moreno, 2006; 2007a; 2011) is an extension of FCA for relations or matrices with non-boolean entries. Consider the $\overline{\mathcal{K}}$-formal context $(G, M, R\,)_{\overline{\mathcal{K}}}$ where the incidence matrix relating objects and attributes $R \in K^{G\times M}$ has entries in a special kind of algebra $\overline{\mathcal{K}}$, an *idempotent semifield*[3]. While in FCA an object is either incident or not with a given attribute, in $\mathcal{K}$-FCA an object i has an attribute *j to a certain degree R_{ij}* . Examples of such algebras include the (completed) max-plus and min-plus semirings—denoted $\overline{\mathbb{R}}_{\max,+}$ and $\overline{\mathbb{R}}_{\max,+}$ in Fig. 1—as discussed in Section 3.2.1.

In this setting, it is better to think of sets of objects and attributes as *vectors in object- and attribute-vector spaces* over the idempotent semifield $\mathcal{X} \equiv \overline{\mathcal{K}}^{G}$ and $\mathcal{Y} \equiv \overline{\mathcal{K}}^{M}$. The crucial technical condition for the existence of a Galois connection (Erné et al., 1993) that induces the concept lattices in this setting is the existence of a dot product $\langle x \mid R \mid y\rangle = x^{\mathrm{T}} \otimes R \otimes y$ between left and right vector spaces over $\overline{\mathcal{K}}$ (Cohen, Gaubert, & Quadrat, 2004). Then Valverde-Albacete and Peláez-Moreno (2011) describes how the *polar (functions)* induce a Galois connection between the vector spaces: $[(\,\cdot\,)^{\uparrow}_{R,\varphi}, (\,\cdot\,)^{\downarrow}_{R,\varphi}] : \overline{\mathcal{K}}^{G} \leftharpoondown \overline{\mathcal{K}}^{M}$:

$$(x)^{\uparrow}_{R,\varphi} = \bigvee\{\, y \in Y \mid \langle x \mid R \mid y\rangle \le \varphi\,\}$$

$$(y)^{\downarrow}_{R,\varphi} = \bigvee\{\, x \in X \mid \langle x \mid R \mid y\rangle \le \varphi\,\}$$

Analogously to FCA, for object- and attribute-vectors a and b, the *φ-formal concept* $(a,\ b)_{\varphi}$ is a pair such that $(a)^{\uparrow}_{R,\varphi} = b$ and $(b)^{\downarrow}_{R,\varphi} = a$with a the *φ-extent* and b the *φ-intent*. As usual, φ-concepts can be ordered by extents or dually by intents

$$(a_1, b_1) \le (a_2, b_2) \Leftrightarrow a_1 \le a_2 \Leftrightarrow b_1 \le^{\mathrm{d}} b_2 \tag{3}$$

and the set of φ-concepts with this order is actually a lattice, the φ-concept lattice. It is easy to see the analogues of the top and bottom concepts and the join and meet operations in these lattices.

The parameter $\varphi \in \overline{\mathcal{K}}$ is called the *threshold of existence* and it can be proven to describe a *maximum degree value* allowed for pairs $(a, b) \in \overline{\mathcal{K}}^{G} \times \overline{\mathcal{K}}^{M}$ to be considered as members of the φ-formal concept lattice $\mathfrak{B}^{\varphi}(G, M, R\,)_{\overline{\mathcal{K}}}$. It can also be interpreted as a *focusing level*: the higher, the coarsest the φ-concepts appear, the lower, the finest. Therefore, to take an overall view of the information in the formal context, it seems necessary to calculate the concept lattices for the different φ threshold values. We call this process *lattice exploration*, an instance of which is described in Section 3.2.5. For an in-depth explanation of the issues involved see Valverde-Albacete and Peláez-Moreno (2007b; 2011).

We can also change the type of FCA carried out by changing the underlying semifield in which the formal context is interpreted: for instance, by inverting all the elements and their order in the semifield, instead of exploring in the concepts with maximum degree of existence, we would be exploring those with *minimum degree of existence*. This is taken up in Sections 3.2.2 and 3.2.5.

2.3. Extended Formal Concept Analysis

In Valverde-Albacete and Peláez-Moreno (2011) the FCA formalism was systematically augmented to consider different modes of conceptualization by changing the basic dual isomorphism in a modal-logic motivated way. This was actually the systematization of similar work laid out in Düntsch and Gediga (2002; 2003) and creates three new types of concepts and lattices of extended FCA, viz. the *lattice of neighborhood of objects*, the *lattice of neighborhood of attributes* and the *lattice of unrelatedness*. Note that extensions for fuzzy FCA also exist for some of these (Konecny, 2011), and that these types of Galois connections appear in a much wider setting in Erné et al. (1993).

The basic technique is to compose the basic (antitone) Galois connections of Sections 2.1 and 2.2 with order- and dual order-isomorphisms, or *anti-isomorphism*:

- We take the TYPE OO Galois connection to be a basic adjunction composed with an *even* number of anti-isomorphisms *on the domain and range orders*. These generate the lattices of neighborhood of objects.
- To obtain a TYPE OI Galois connection, we compose a basic adjunction with an *odd* number of anti-isomorphisms *on the range*. These are the standard concept lattices of FCA.
- To get a TYPE IO Galois connection, we compose a basic adjunction with an *odd* number of anti-isomorphisms *on the domain*. These connections generate the lattices of neighborhood of attributes.
- Finally, a TYPE II Galois connection is a basic adjunction with an *odd* number of anti-isomorphisms composed *on both the domain*

[2] There is also a nice duality which allows to reify attributes and convert objects to properties.

[3] The main peculiarity in idempotent semifields (in idempotent semirings, in general), is that addition is idempotent, thereby ruling out additive inverses.

and range. These are the lattices of unrelatedness between objects and attributes, and they will not be used in this paper.

Such alternate modes of conceptualization can be implemented by negation operations on contexts and sets in standard FCA, and by the inversion of vectors and relations in $\mathcal{K}$-FCA, whereby the constructions of the lattices of these new connections can be reduced to the original TYPE OI Galois connection. The rather mathematical details can be found in Valverde-Albacete and Peláez-Moreno (2011).

3. Results

Since the explanation of the metaphor of LofK is summarized in Section 1 and detailed elsewhere (Wille, 1999; 2006), what we contribute here is, first, our analysis of the *affordances* of such a metaphor as embodied in knowledge-intensive tasks (Section 3.1), and second, a methodology that leverages FCA as a framework capable of supporting the enterprise of scientific and engineering enquiry (Section 3.2).

3.1. Generic affordances of Formal Concept Analysis as Landscapes of Knowledge for Exploratory Data Analysis

First we will introduce some generic affordances available across all applications and then specific cases to exemplify particular instances.

3.1.1. Formal, domain-independent and ubiquitous modeling

The first idea that we want to put forward is that

Affordance 1. The formal quality of FCA makes it an optimal technique for domain-independent data analysis.

This is a mixture of the abstract or *formal* quality of FCA already mentioned, that is *interpretation-independent*, and the ubiquity of two-mode binary relations as data for all scientific domains. Of course, this is the result of a dedicated and painstaking effort on the part of the developers of FCA to present a well-founded theory and practice of formal contexts and concept lattices.

The second affordance stems from the duality of contexts and lattices and empowers FCA as an EDA technique,

Affordance 2. Visualization and manipulation of data in table format and hierarchical format are mutually warranted.

That this is one of is its greatest strengths can be seen from the main theorem of FCA—stating that up to isomorphism, *every* complete lattice emerges as the concept lattice of some formal context and vice-versa—the many different types of information that emerge from either representation of the data—hierarchical order, implications, attribute and object reduction,—, and the many manipulative techniques to extend, restrict or combine formal contexts and concept lattices: essentially, there is a whole *algebra of contexts and lattices* that allows partitioning and aggregating interesting data in different ways under either guise (see Deiters and Erné, 2009; Ganter and Wille, 1999, Chaps. 3–7).

3.1.2. Setting the board

FCA is limited in the data it can analyze by its own stage-setting. The next affordance describes exactly what it can do,

Affordance 3. Standard Formal Concept Analysis provides a generic framework to deal with two-mode data in a homogeneously meaningful, contextualized way.

By "homogeneously meaningful" we mean that it provides the same kind of manipulations and interpretations over the whole set of application domains. By "contextualized" we mean that building a formal context exactly sets the board of the investigation to be addressed.

To wit, the standard version of FCA described in Section 2.1 rests heavily on the binary nature of the incidence relation. In fact, a formal context is another name for a binary relation where the elements of the carrier sets have been "contextualized" by giving them names and interpretations as objects and attributes. Building the context itself supports the scientific activity of "setting the board" before a problem is attacked. The context frames the relevant entities and their properties and grants them names to support some sort of semantic grounding for discursive rationality. Nothing more, and nothing less than the objects, their attributes and the incidence relationship can appear in questions about the data. Note that if the previous affordances were missing, this would be a very stringent limitation.

3.1.3. Analyzing

Co-clustering of objects and attributes and their hierarchic organizations is a straightforward affordance of FCA.

Affordance 4. A concept lattice describes a hierarchical, non-partitional co-clustering between sets of objects and attributes.

Non-partitional means that although there is a covering for either set of objects and attributes, the sets in the covering are not incompatible. Note that there are no widely accepted measures—and, indeed, very few used measures at all—for the assessment of non-partitional clustering at present (Mirkin, 1996; 2005)[4].

Sometimes, there are readily observable subsystems in the data. The identification of adjoint sub-lattices as independent sources of structuring and the analysis of their corresponding sub-contexts is a valuable contribution to help formulating hypothesis. This affordance is particularly made evident in the applications of Sections 3.2.2 and 3.2.3.

Affordance 5. Adjoint sublattices provide qualitatively and quantitatively different behaviors for associated sets of objects and attributes.

$\mathcal{K}$-FCA's free parameter φ is a useful tool for exploring degrees of incidence at various levels of detail, from the big picture where only the most remarkable structures appear to the most detailed where even the tiniest confusions appear. Observing several concept lattices at different φ is a good starting point when one has only a vague idea of the phenomenon under study. This affordance will be specially developed in Sections 3.2.2 and 3.2.5.

Affordance 6. $\mathcal{K}$-Formal Concept Analysis allows the exploration of multi-valued data at different levels of detail modulating the φ parameter.

3.2. A methodology for LofK based in FCA based on use cases

The following analysis cases have a three-fold purpose: they focus on a basic affordance, they demonstrate the affordance, and they support the abduction of scientific or engineering hypotheses.

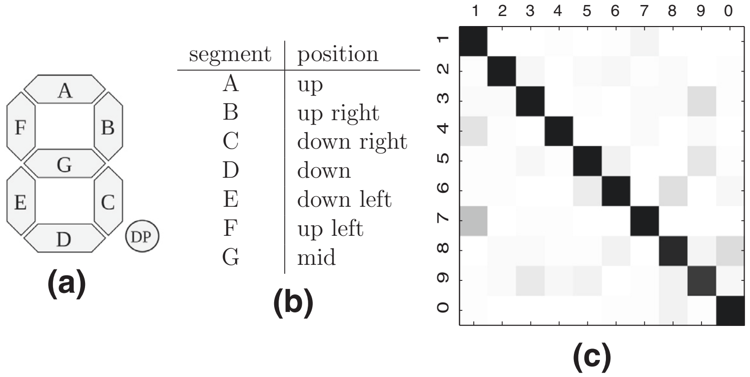
3.2.1. Use case: extending FCA theory itself

In this use case, we show how to organize formal knowledge using FCA so that new research hypotheses suggest themselves upon lattice browsing.

Affordance 7. Standard Formal Concept Analysis supports postulating new research hypotheses.

Example. For instance, consider the theory of semirings as compiled in Golan (1999), a rather specialized branch of abstract algebra. To describe what a semiring is, the author starts with an *abstract description* of what constitutes a semiring, then proceeds to define *properties* through which semirings can be studied, and to induce better

[4] This adds insult to injury in the case of the evaluation of non-supervised learning tasks, like concept-lattice inferencing.



segment	position
A	up
B	up right
C	down right
D	down
E	down left
F	up left
G	mid

Fig. 2. 7-segment numeral schematics (a) image from Wikipedia Commons, (b) mapping to the mnemonics used in this paper, and (c) heatmap of human confusions in recognizing the 10 numerals.

grasping of the material intersperses *concrete examples* that show the properties.

It so happens that the type of "algebras" that allow a matrix with entries in those algebras to display the behavior of FCA seems to be intimately tied up to semirings:

- The Boolean algebra that allows us to define the polars of FCA is itself a semiring.
- The most widespread multi-valued extension to FCA, FCA in a fuzzy setting (Bělohlávek, 2002; Burusco & Fuentes-González, 1994) is based on fuzzy-algebras, a kind of *incline*, also a semiring.
- Another extension is $\mathcal{K}$-FCA (Valverde-Albacete & Peláez-Moreno, 2006; 2011), based on the concept of a *complete idempotent semifield*, also a semiring.

Given these data, the research question an FCA practitioner should ponder is whether other kinds of multi-valued extensions are possible.

One method to explore this question with FCA suggests itself readily:

1. **Contextualizing:** Gather all abstract and concrete algebras as objects and the properties of semirings as attributes. Build the formal context where an algebra is incident to a property if that algebra has the property .
2. **Data preparation:** In this case, use data as is, even keeping duplicates and null columns and rows.
3. **Visualizing & Identifying:** Obtain the concept lattice of this context and represent it, as in Fig. 1.
4. **Exploring the data:** Locate the Boolean algebra (a concrete structure), inclines and idempotent semifields–two abstract structures—in this lattice: they are marked with a bottom, left and right red arrows in the Figure, respectively. Check that the Boolean algebra is an instance of both abstract algebras by finding their meet and the Boolean algebra in its extent. This is easy to do in the lattice and indeed the Boolean algebra is an instance of both these concepts.
5. **Abstracting the question:** Locate the join of the formal concepts to which inclines and idempotent semifields belong in Fig. 1, marked with red arrow near the top. We see that this node (unnamed) is a meet of other concepts namely the object concept for "dioid" and the object concept for "information algebra".
   We look into dioids to find that they are semirings that have a "natural order" within them and that they are as "far away" as can be from "usual" rings. To this, information algebras add the property of being *entire*, e.g. having no zero divisors, and being *complete in the order*.
   Browsing downwards from the unnamed node we find that some peculiar concrete structures are actually complete entire dioids, like the positive reals $\mathbb{R}_0^+$ and rationals $\mathbb{Q}_0^+$.
6. **Abducting hypotheses:** We posit that complete dioids are actually those semirings capable of inducing the Galois connection structure that we know as FCA.

Result: We should find a construction for such Galois connections that puts into play the specific features of these two (non-idempotent) semifields $\mathbb{R}_0^+$ and $\mathbb{Q}_0^+$ and possibly others. □

3.2.2. Use case: the analysis of contingency or confusion matrices

Confusion matrices are a long-standing technique to quantify the performance of multiclass classifiers. They are devices to measure the performance of any type of classifier—whether embodied or artificial—by means of tallying the errors and successes of iterated acts of classification.

A detailed confusion matrix is often presented in the form of *heatmaps*—see Fig. 2(c)—or numerically (Congalton & Green, 1999). Also, behavioral sciences frequently choose this type of representation to illustrate the results of perceptual experiments (Miller & Nicely, 1955). The purpose of displaying the confusions either way is to provide a more *friendly* representation of the phenomenon under study (be it biologically or machine-originated) that allows the elicitation of knowledge.

We contend that the inclusion order of confusions between input and output responses in a CM is more efficiently represented by concept lattices and it is more convenient for knowledge discovery.

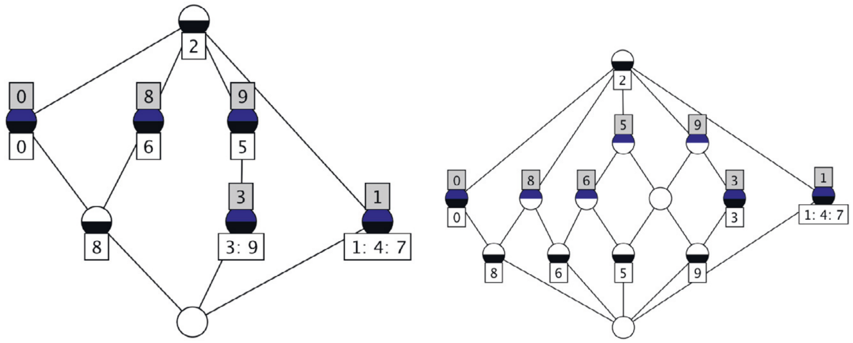
Affordance 8. $\mathcal{K}$-Formal Concept Analysis supports the detection of regularities in the structure of the data that have theoretical significance.

Example. For this example and that in Section 3.2.3 we use the data on human confusions of segmented numerals collected in Keren and Baggen (1981). Note that this task is also used in Mirkin (2005) to exemplify a number of unsupervised data analysis methods facilitating a comparison with our results.

In this visual acuity task $k = 10$ seven-segment numerals of digital displays were presented to human testers under different conditions designed to guarantee that confusions would appear. Fig. 2 shows the schematics of these numerals in terms of their segments, some mnemonics for later reference, and a heatmap representation of human confusions in the task.

Next we apply our methodology and discuss its associated affordances:

1. **Contextualizing:** Let $V_X = \{x_i\}_{i=1}^k$ be an *input alphabet* and $V_Y = \{y_j\}_{j=1}^{k'}$ an *output alphabet*. When we consider experiments in human performance in perceptual tasks, we model stimuli as input



(a) At $\varphi = 0.05$.

(b) At $\varphi = -0.43$.

Fig. 3. Concept lattices of the segmented digits confusion matrix at different φ . *Stimuli* are labeled in white and *responses* in gray.

symbols and responses as output symbols. For observational purposes, the basic experiment is: "presenting a stimulus x_i to obtain a response y_j from the subject," $(X = x_i, Y = y_j)$. The process of classification is enacted by a sequence of m independent experiments $(X_m = x_i, Y_m = y_j)$, whose results are tallied and aggregated into a *(count) confusion matrix* $N_{XY} \in \mathbb{N}^{V_X \times V_Y}$ where

$$N_{XY}(x_i, Y_j) = \sum_m \delta_{X_m Y_m}(x_i, y_j) \tag{4}$$

gives the count that y_j was returned as a response to x_i. A confusion matrix N_{XY}, then, is a particular kind of *contingency table* for a classification process (Mirkin, 2005, p.51), the "contingency" being that input stimuli get confused with output responses. Indeed all contingency tables with comparable rows and columns are susceptible of the same treatment.

2. **Data preparation:** To cast absolute frequencies onto an idempotent-semiring we obtained the *pointwise mutual information matrix* (Fano, 1961): Call $n_{ij} = N_{X,Y}(x_i, y_j)$, $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$. Note that in our setting[5], $m = \sum_{ij} n_{ij}$. From the count matrix of (4), we can estimate the empirical estimate or maximum-likelihood estimator of the joint probability distribution between inputs and outputs as $\hat{P}_{X,Y}(x_i, y_j) = \frac{n_{ij}}{m}$. Whence the (empirical pointwise) *mutual information*[6]$\hat{I}_{XY}(x_i, y_j)$ between X and Y (Fano, 1961, Section 2.3) is

$$\hat{I}_{XY}(x_i, y_j) = \log \frac{\hat{P}_{XY}(x_i, y_j)}{\hat{P}_X(x_i) \cdot \hat{P}_Y(y_j)} = \log \frac{n_{ij} \cdot m}{n_{i\cdot} \cdot n_{\cdot j}}. \tag{5}$$

3. **Visualizing & identifying:** We explored the multi-valued context with the technique of Section 2.2. Fig. 3(a) and (b) are examples of two such structural lattices for the data.
Note that, in this case, both inputs and outputs have the same denominations. To distinguish them in the lattice we will use white boxes for the former and gray for the latter. In the text, the inputs will be denoted with boldface characters. For instance, in Fig. 3(a) the object concept for object **5** is $\gamma(\mathbf{5}) = (\{\mathbf{3}, \mathbf{5}, \mathbf{9}\}, \{9\}) = \mu(9)$ which is also the attribute concept $\mu(9)$, therefore the concept lattice is labeled on the node directly below the top and to the right of it with a "**5**" on white denoting an object and a "9" on gray, denoting an attribute.
4. **Restructuring:** In the example of Fig. 3(a) we can clearly observe three differentiated groups as adjunct sub-lattices, namely:
   - **Right sub-lattice**: *stimuli* "**1**", "**4**" and "**7**" (white label) are perceived as "1" (gray label),
   - **Center sub-lattice**: *stimuli* "**3**" and "**9**" (white label) are perceived as "3" and "9" (reading the gray labels from the node upwards) and also *stimuli* "**5**" (white label) is perceived as "9" (gray label) and
   - **Left sub-lattice**: *stimulus* "**8**" (white label) is (wrongly) perceived as "0" and (correctly) as "8" (gray label) whilst for *stimulus* "**0**" (white label) we get the correct *response* "0" (gray label) and for *stimulus* "**6**" (white label) we get the wrong *response* "8" again. In this case, we can assert that *stimulus* "**8**" is more *generic* than "**0**" and "**6**" since, the former is confused with both the latter whilst the latter are only taken as a single *response*. These are therefore more *specific* and we can see them appearing in upper nodes of the lattice.
5. **Analyzing:** The digit proximities denoted by these three sublattices also agree to a certain extent with another representation, the *directed threshold graph*, proposed in (Mirkin, 2005, p.52), that provides the following interpretation: in the case of the right sublattice, its stimuli share the fact that the *lowest* segment of the representation is missing. For the center sub-lattice, a distinctive characteristic is the absence of the lower-left segment. Finally, for the left sublattice, the number of common present segments is very high (both "0" and "8", on the one hand and "6" and "8" share 6 out of the 7 possible segments).
Fig. 3(b) allows us to zoom in some of the confusions. For example, "**5**" and "**6**" are both mutually confused. Also, "**9**" is more *specific* than "**3**" since the former is taken for "5", the latter and itself while "3" is only taken for "9" and itself. This is the reason for the node with the white label "**9**" to appear lower in the lattice than "**3**".
6. **Abducting hypotheses:** In view of the previous analysis we can abduct the following interpretation: in the case of the right sublattice, its stimuli share the fact that the *lowest* segment of the representation is missing. For the center sub-lattice, a distinctive characteristic is the absence of the lower-left segment[7]. Finally, for the left sublattice, the number of common present segments is very high.

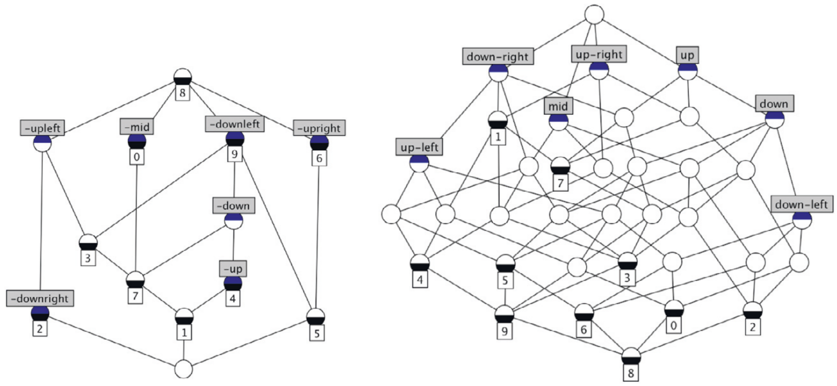
Result: We found three perceptual sublattices that embodied in different ways similarities and differences between stimuli and responses. □

This procedure was also employed for the analysis of human and machine phonetic confusions in Peláez-Moreno, García-Moral, and Valverde-Albacete (2009) Peláez-Moreno et al. (2010) Valverde-Albacete (2010). In Peláez-Moreno et al. (2010) we analyzed the adjoint sublattices appearing at different φ in the human phonetic

[5] In the context of machine learning it is customary to have m as the number of feature vectors to be classified and n as their dimension. In a statistical context n is often taken to mean a count frequency.

[6] Also called the *transinformation* in the context of communication channels.

[7] A similar interpretation, from a different point of view will also be presented in Section 3.2.3.



(a) Type OO (left adjunction). (b) Type OI (Galois connection).

Fig. 4. (Color online) Concept lattices of the encoding of segmented digits in terms of their present or absent segments. *Digits* are labeled in white and the segments present (b) or absent (a) in gray.

confusions of (Miller & Nicely, 1955) corroborating some of the conclusions of the most influential papers on Human Speech Recognition (HSR) to date. The original authors had manually clustered the sounds under analysis determining that phone recognition is grounded on *hierarchical categorical discrimination*, that is, English consonant sounds form groups identified in terms of hierarchical clusters of *articulatory features*. Furthermore, they introduced the notion of *virtual articulatory communication channels*, identifying five of them: voicing, nasality, affrication, duration and place.

In our analysis with concept lattices we were capable of clearly observing the appearance of *nasality* as the most outstanding channel, closely followed by *voicing* but then, the hierarchy of the channels differs for unvoiced consonants, where the most distinctive feature is *friction* followed by *manner* (and a glimpse of *duration*) and voiced ones, where the *place of articulation* is more relevant.

In this case the analysis with $\mathcal{K}$-FCA corroborated analyses done decades ago that were being questioned at present, whence we conclude,

Affordance 9. $\mathcal{K}$-FCA analyses can lend support to hypotheses independently obtained.

The same analysis conducted on HSR to identify Miller & Nicely's articulatory channels described in Section 3.1.3 was carried out on an Automatic Speech Recognizer (ASR) (Peláez-Moreno et al., 2010), the most striking conclusion being that the former's most salient feature, *nasality*, was not clearly observed. This leads us to hypothesize that, either the feature extraction mechanisms or the acoustic models of the machine recognizer are not adequate for representing this characteristic of speech, apparently the easiest for humans. Further work to improve those mechanisms should help reduce the gap between HSR and ASR.

Affordance 10. By rendering data to a common, comparable form, FCA is capable of comparing performances across different data domains, when the data setting are compatible.

This is the case with human- and machine-induced acoustical confusions.

3.2.3. *Use case: extended FCA analysis of confusion matrices*

By using the techniques in Section 2.3 we can sometimes enrich certain kinds of analyses by catering to more phenomena than those observable through standard FCA.

Affordance 11. Extended, generalized FCA can analyze data to find a wider array of phenomena than FCA.

Example. In this Section we enrich the analysis of confusion matrices sketched above by providing evidence for perceptions that take into consideration the absence of certain parts of stimuli. We use the same data setting and processing.

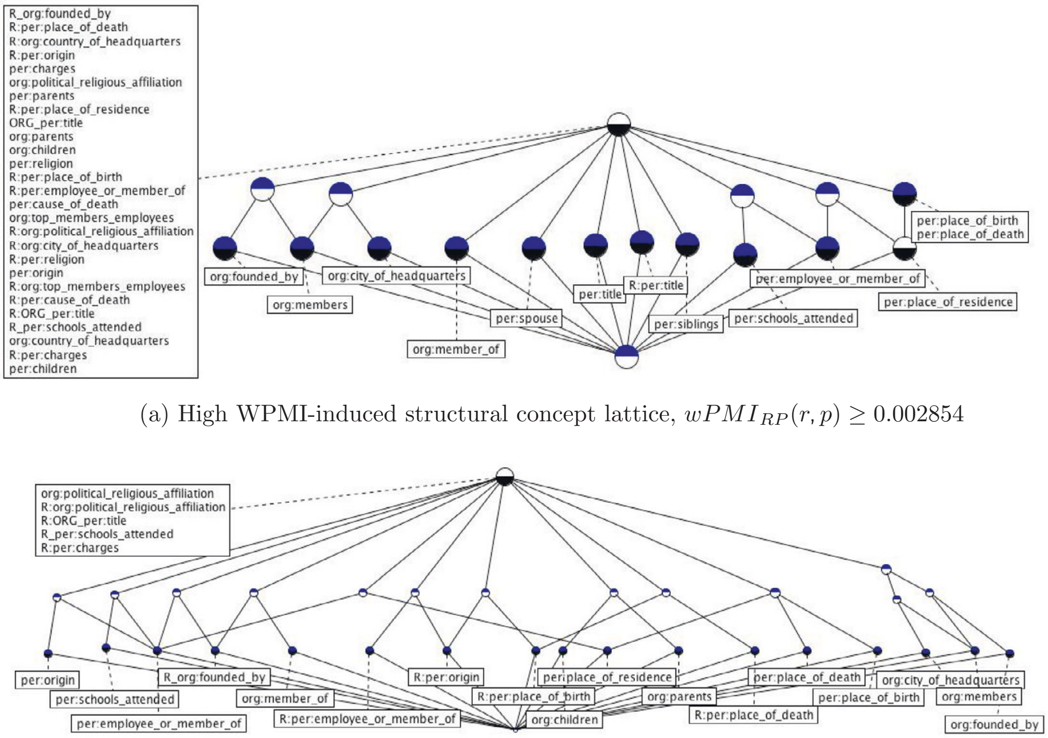
1. **Contextualizing:** Fig. 4 shows two instances of the first and second types of extended conceptual lattices for the incidence matrix of segmented numerals (Keren & Baggen, 1981) specifying for each digit—the *objects*—every segment activated to delineate it—the *attributes* (see Table 2(b)). Note that this is a mere boolean description of the composition of the numerals in terms of their segments and it does not include any information about human confusions as those of Fig. 2(c) in Section 3.2.2.
   In the lattice to the left, the object concepts are read with negative attributes, that is "number 2 does not have the up-left or down-right segments", while in that to the right they are read in the positive, that is "number 2 has segments up, down, mid, down-left, and up-right".
2. **Visualizing & identifying and analyzing:** The comparison of the two concept lattices evidences the usefulness of TYPE OO extensions: while Fig. 3(b) is a highly complex lattice where the *recognition* or *analysis* of any relationship, clustering or organization in the geometry of the task is hardest, Fig. 3(a) is a less complex description.
3. **Abducting hypotheses:** Following Occam's razor, we choose the left lattice as a more accurate representation of the underlying phenomena (confusions) and we posit the hypothesis is that the absence of segments explains better some of the errors in segment perception.
   Note how observing this lattice we can formulate a hypothesis to explain the right and center adjoint sublattices of Fig. 3(a) but it cannot provide a clear account for the left sublattice. □

Result: In concrete dataset analyses it pays to consider non-usual versions of concepts and standard concept lattices side-by-side. □

3.2.4. *Use case: structuring open-domain relation extraction*

Relation Extraction is the task of obtaining *relation instances*, triples of two entities and a (reified) pre-specified relation between them. *Open Relation Extraction* (Fader, Soderland, & Etzioni, 2011; Yates et al., 2007) aims at automatically extracting relation instances, where the relation is directly mined from texts instead of using a pre-specified set. Given a set of extracted instances the question arises whether they show any relationship between them and what such relationship might be:

Affordance 12. FCA allows us to elicit hierarchical structure from apparently unstructured domains.



(a) High WPMI-induced structural concept lattice, $wPMI_{RP}(r, p) \geq 0.002854$

(b) Low WPMI-induced structural concept lattice, $wPMI_{RP}(r, p) \geq 0.000598$

Fig. 5. Illustration of the structure of high *wPMI*$_{RP}$-induced concept lattices. The attributes (patterns) are not shown to lessen the visual clutter. The little structure that there is seems to be semantically-motivated, but the reciprocal relations (prefixed with "R:") are never clustered close to their direct counterparts.

Example. Next we show how the use of FCA helps to bring hierarchical structure to a domain as long as it can be captured by a two-mode data matrix.

1. **Contextualizing:** In spite of the data for this task being three-moded, and since the theoretical development and software for *triadic concept analysis* is so scarce (Wille, 1995), we can use FCA by encoding the pairs of entities as if they were attributes predicated of relations. Here relations are objects and pairs of entities are attributes[8].
2. **Data preparation:** To build the dataset of relation candidate patterns over which we will perform our analyses, we aligned relation tuples extracted from a Linked Data base and very general lexical-syntactic patterns mined from a large document collection. Then we transformed the data to a format amenable to FCA processing and carried out the exploration. The data sources used in these experiments were:
   - A large open-domain document collection: the source data from the TAC 2010 Knowledge Base Population Evaluation. This data release consists of 1 777 888 files, from which over 1.2M are newswire and around 490K are Web documents (Ji, Grishman, & Dang, 2011).
   - A Linked Data base: we downloaded a full data *dump* of Freebase, dated May 26th, 2013 (Google, 2013). This dataset contains 585 million RDF-serialized assertions.

   After the linguistic processing, we transformed the clustered pattern sequences into (complex) pattern expressions thus obtaining a count-valued context (R, P, N_{RP}), where R is the set of relations, P is the set of such (complex) patterns and $N_{RP} \subseteq \mathbb{N}^{R \times P}$ is the $\mathbb{N}$-valued relation of occurrence counts.

   Using these counts of relation-patterns we estimated the empirical joint probability $P_{RP}(r, p)$ between relation names and patterns and calculated the *weighted point-wise mutual information (wPMI)*, using the following formula where $I_{RP}(r, p)$ is the PMI of (5),

   $$wPMI_{RP}(r, p) = P_{RP}(r, p) \cdot I_{RP}(r, p)$$

   to create an exploration formal context (R, P, wPMI_{RP}).
3. **Exploring:** Next we carried out the exploration of this context as described in Section 2.2. Since the cardinal was high, we sampled a 20% of the total number of values in the range of φ to reduce the computation time. Later, when we chose some of these measures to carry out the structuring task, we noticed that interesting lattices were already in the sample.
4. **Analyzing:** In Fig. 5 we present two concept lattices obtained by min-plus analysis exploration of the context (cfr. Section 3.2.5). Note that the complex patterns are not represented to prevent visual clutter. The first outstanding result is that even at high positive thresholds as in Fig. 5(a) there are some semantically-motivated hierarchical relations, for instance `per:place_of_birth, per:place_of_death` and `per:place_of_residence`, although at that point most of the relations are either not captured—the big extent of the top concept—or completely disjoint from each other—like the many nodes only related through top and bottom.

   At lower thresholds, as in Fig. 5(b), there is an increase in hierarchy height. But relations and their reciprocals are still not placed in the same cluster—see, for instance the two clusters on right and left around `org:founded_by` and its reciprocal `R:org:founded_by`, respectively. We believe this to be produced by the different valencies of relations and their reciprocals: the complex patterns are *ordered* pairs of entities and the reciprocity inverts it.

[8] In fact, they are *tags for classes* of entities as detected by a Named-Entity Recognition algorithm, but we want to concentrate on relations for this application.

Results: This exploration reveals that there is very little structure in the set of relations extracted with these particular techniques from the Freebase data and documents we used[9].

When there is, the proposed scheme is capable of detecting structure induced by apparent semantic proximity of the relations, barring reciprocity. Since from the names of the relations it is evident that they are semantically related, the initial data extraction is a good candidate for improvement.

Note also that when the data is originally very much related, the extraction of structure is much more apparent as made evident for FrameNet data in Valverde-Albacete (2008). □

3.2.5. *Use case: the analysis of overexpression and underexpression in Gene Expression Data*

Grouping genes by means of gene expression similarity data has been the goal of many researchers for well over the past 20 years as a way to induce the proteins being synthesized in a sample of cells for each precise condition under study. Under- or over-expression of genes offers information about the co-regulation underlying expression mechanisms. This co-regulation process is not exempt of the problems of high variability due to the unreliability of the measurements. Currently the most popular technique to obtain gene expression profiles is still using gene micro-arrays, but in recent years there have been some advances in next generation sequencing techniques making it possible to obtain gene expression profiles with a higher quality (Metzker, 2010; Xie, Wu, Tang, Luo, Patterson, Liu, Huang, He, Gu, Li, Zhou, Li, Xu, Wong, & Wang, 2014).

In both cases, Gene Expression Data (GED) must be *explored* to various ends: to *recognize* relationships between the probesets under different empirical conditions, to *identify* the biological functions of genes yet unknown, to *investigate* the role of measurement errors in the diversity found in different instances of the same experiment and finally to provide friendlier visual representations to human cognition and *memorization* that make this tool interactive and helpful for *decision making*.

FCA multivalued extensions are essential for this task: since a gene is considered to be under-expressed (resp. over-expressed) under some *particular* conditions with respect to some *reference* condition if its m-RNA concentration level is lower (resp. higher) for those *particular* conditions than for the reference case, we can employ two opposite semifields, for instance max-plus and min-plus, to explore both phenomena under the $\mathcal{K}$-FCA framework (González-Calabozo et al., 2011; 2012).

Affordance 13. Changing the underlying semifield in $\mathcal{K}$-Formal Concept Analysis allows us to measure processes of opposite polarities.

Example. To illustrate how to realize this affordance with WebGeneKFCA (González-Calabozo, Peláez-Moreno, & Valverde-Albacete, 2012), a tool for GED analysis[10], we present here an example of analysis intended as a guide through the *exploratory process* designed to discover the relationships among genes belonging to different human tissues[11].

1. **Contextualizing:** Five different human muscles were considered: *vastus lateralis, quadriceps, deltoid* and *vascular smooth muscle*. Samples were downloaded from the NCBI gene expression database (Edgar, Domrachev, & Lash, 2002), from different experiments and analyzed with the HG-U133-Plus Affymetrix micro-array chip, whose heterogeneity intends to make evident clear-cut distinctions to demonstrate the applicability of the tools.
   - **Vastus lateralis**: Two samples from the *quadriceps* extracted from healthy women used as a control group for a polycystic ovary syndrome study.
   - **Quadriceps**: Two samples from an unspecified muscle that prior to sample extraction was deprived of blood with a tourniquet applied to the lower limb after the administration of spinal anesthesia, as per the normal protocol for knee arthroplasty surgery.
   - **Deltoid**: Two samples from a much bigger study with 622 samples collected for the characterization of several human tissues.
   - **Vascular smooth muscle**: Two samples extracted from a healthy control group to analyse muscle cells infected with *Trypanosoma cruzi*.

   Note that we have intentionally selected quite disparate samples so that the outcomes of the analysis are well-known and falsifiable.

   Robust Multichip Average (RMA) summarization was performed using the Affymetrix Power Tools (Affymetrix, 2013a) resulting on a $G = 54\,675$ by $M = 8$ matrix containing an expression of a single probeset per row. The correspondence between genes and probesets is given by the Affymetrix Annotation file, but most of the probesets identify a single gene unequivocally (Affymetrix, 2013b).
2. **Data preparation:** Prior to starting the $\mathcal{K}$-FCA exploration we need to perform some preprocessing to allow the comparison of the expression of different probesets. Since these have different ranges of expression, a normalization is necessary. Several alternatives are implemented in WebGeneKFCA (González-Calabozo et al., 2011) and here we have chosen to apply the natural logarithm of each probeset i for the condition j divided by the arithmetic mean of the gene expression over all considered conditions as it better agrees to the nature of the quantities (m-RNA logarithmic concentration) being considered[12].
3. **Exploring: the evolution of the number of concepts** An important feature of the sequence of concept lattices produced in the exploration is the *evolution of the number of concepts along the* φ—the threshold of existence for max-plus exploration capturing under-expression—, *and* ϕ *axes*—for min-plus exploration to capture over-expression. It provides an indication of the size and complexity of the concept lattice: if the absolute value of the threshold of existence is large, then large absolute values of the gene concentrations will be required for the concepts to exist, meaning that the lattice will only show the most salient relationships or co-clusters. On the contrary, if the absolute values required for existence are low, many nodes of the lattice appear—sometimes showing spurious relations due to the uncertainty in data collection or measurement noise. The graphical interface of WebGeneKFCA, as an interactive tool for analysis and decision, allows the user to navigate along this parameter, zooming in and out to observe rougher or finer groupings.

   Fig. 6 is a depiction of this evolution for the data described previously. Since $\varphi \in (-\infty, 0]$ and $\phi \in [0, \infty)$ the left part of the curve corresponds to the analysis of co-regulated gene under-expression while the right one represents the analysis of over-expression. The maximum number of concepts $2^M = 256$ is attained at $\varphi = \phi = 0.0$. We can observe that the overall negative slope of the $\overline{\mathbb{R}}_{\min,+}$ part is higher and smoother than that of $\overline{\mathbb{R}}_{\max,+}$ which lead us to think that the over-expressed genes are fewer than under-expressed ones. Sudden changes in the slope reveal values of the threshold of existence where the lattice changes substantially and therefore are interesting to look into.

[9] We also checked these results with plain joint probability P_{RP} and PMI I_{RP}. wPMI and joint probability provided the best results.

[10] Available at http://webgenekfca.com

[11] This particular instance of analysis is available at https://webgenekfca.com/webgenekfca/kfcaresultses/4. Other results on *Arabidopsis Thaliana* are also available in (González-Calabozo et al., 2011).

[12] These data are available from https://webgenekfca.com/webgenekfca/datamatrices/4 including heatmaps of gene expression levels.

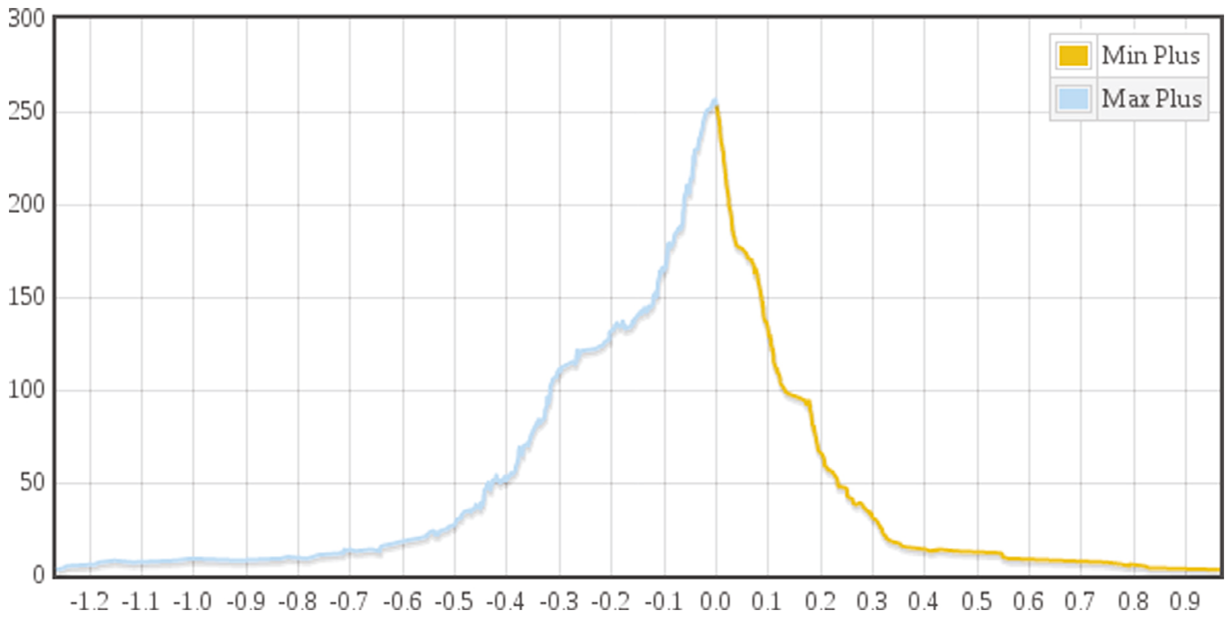


Fig. 6. (Color online) Number of concepts vs. φ (light blue, to the left of 0.0) and number of concepts vs. ϕ (drab yellow, to the right of 0.0) for the context being explored. The positions of this curve can be explored in an interactive way at https://webgenekfca.com/webgenekfca/kfcaresultses/4.

4. **Exploring: biclustering of genes.** The way to present the concept lattice in a tool designed as an aid for knowledge exploration is crucial. Even more so if we take into account the size of the contexts analyzed. There are many different available algorithms to visualize a concept lattice, each with its advantages and disadvantages (Eklund & Villerd, 2010).
   Since the use of $\mathcal{K}$-FCA requires the visualization of a *sequence* of concept lattices as a function of φ or ϕ, in González-Calabozo et al. (2012) we proposed a scheme having the distinctive feature that the formal concepts with the same intent belonging to different concept lattices ($CL(\varphi)$ or $CL(\phi)$) are always plotted in the same spatial position. This means that as the user explores the values of φ or ϕ, she will easily see how the extent of each concept evolves, increasing or decreasing in size. They might even disappear.
   In our particular example, with $M = 8$ samples, this silhouette will exhibit 9 levels of rows including the top and bottom (see Figs. 7–9). The intents of these 8 concepts of the second row will have a single element: each of the samples. The second row will be composed of up to $\binom{8}{2}$ 2-combinations of the previous row concepts' intents, the third, up to $\binom{8}{3}$ 3-combinations, and so on.
   Thus, we say that the *higher* in the lattice a concept appears, the more *specific* the co-cluster so profiled is, since the set of genes that comprise the extent apply to fewer samples and therefore are the ones that make it *distinct*. Conversely, the *lower* the concept, the more *generic* the cluster is, since the genes enumerated in its extent apply to a larger number of samples and hence are more *common*.
   Next, a few examples of the analysis afforded at particular points of the graph in Fig. 6 will be presented.
5. **Analyzing gene under-expression:** According to the observations made in point 3 we choose the concept lattice at $\varphi = -0.4$ since it seems to be an inflection point in Fig. 6. The concept lattice for this φ is shown in 7. In this position a large co-cluster can be found (highlighted in blue), whose genes are under-expressed or down-regulated in both *vascular smooth* muscles' probes. Since all the concept lattices are plotted over the silhouette of a boolean concept lattice, the concept with this extent always appears in the second row to the right, connecting the last two concepts of the previous row that correspond to the last (*vascular2*) and last but one (*vascular2*) probes[13]. The 721 genes of its extent profile these two probes (at this particular $\varphi = -0.4$).
   This information will allow us to navigate the gene ontologies indexed by WebGeneKFCA in search of an explanation for this grouping as explained in Section 3.2.6.
   The researcher can now identify the most under-expressed genes in this cluster, and she will find that the analysis seems reasonable since in this experiment the only smooth muscles are the two vascular ones: the rest are striated. These under-expressed genes in smooth muscles are over-expressed in *striated* muscles.
   Moving on to a lower value, $\varphi = -0.22$, we find another change of slope in the number of concepts of Fig. 8. The *vascular* co-cluster described above persists (with 1 748 new genes) and three new salient co-clusters are added, related to the *quadriceps* samples. The two co-clusters in the second row specify the genes privative to each of the two samples whilst the one in the third, lists the genes in common.
   Another very populated cluster appears in the third row formed by the probesets which belong to under-expressed *quadriceps* genes (displayed as a blue circle in the concept lattice of Fig. 8). A priori, we should expect little difference between the smooth *vastus lateralis* samples and those of the *quadriceps*, but in this case the differences are due to the conditions in which these probesets have been collected: a tourniquet applied to the quadriceps deprives it from some oxygen and nutrients, so we hypothesize that this is the reason why this co-cluster becomes salient. On a more detailed analysis we find under-expressed genes (like EPOR) with functions related to red blood cell survival that have a direct relation with the tourniquet applied and the lack of enough blood. Note that to draw these conclusions we used the aid of the interface to gene ontologies the tool provides, explained in Section 3.2.6.
6. **Analyzing gene over-expression:** The same exploration can now be applied using min-plus analysis to discover information about over-expression. For example the concept lattice appearing around $\phi = 0.15$ is shown in Fig. 9. It shows clearly the *vascular*

[13] The reader is encouraged to explore the lattices using the on-line application WebGeneKFCA as the labels of the nodes appear when selected by the click of the mouse. Figs. 7–9 can hardly approximate the wealth of information in the interactive tool.

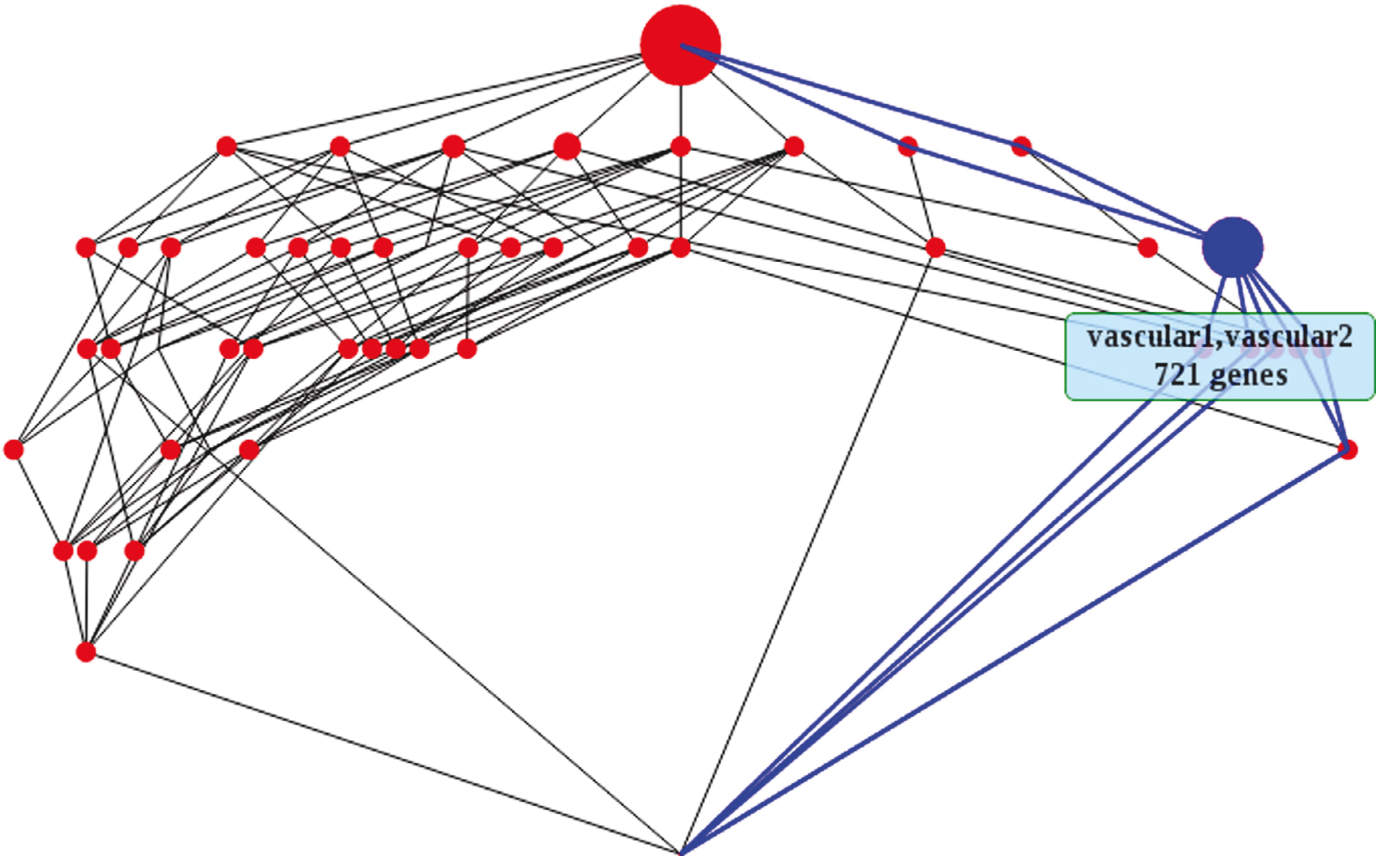


Fig. 7. (Color online) Gene under-expression lattice for $\varphi = -0.4$. This lattice can be explored in an interactive way at https://webgenekfca.com/webgenekfca/kfcaresultses/4.

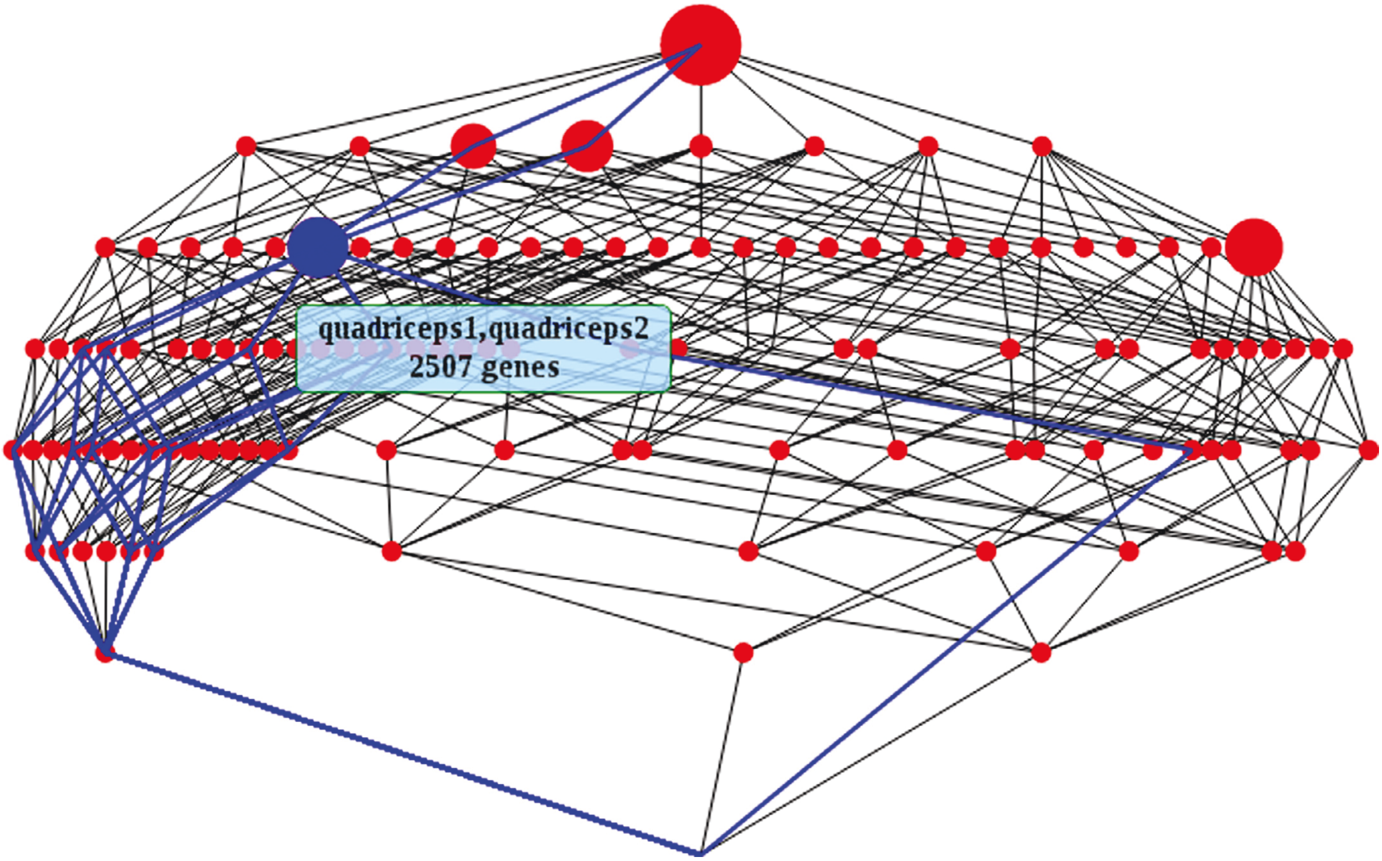


Fig. 8. (Color online) Gene under-expression lattice for $\varphi = -0.22$. This lattice can be explored in an interactive way at https://webgenekfca.com/webgenekfca/kfcaresultses/4.

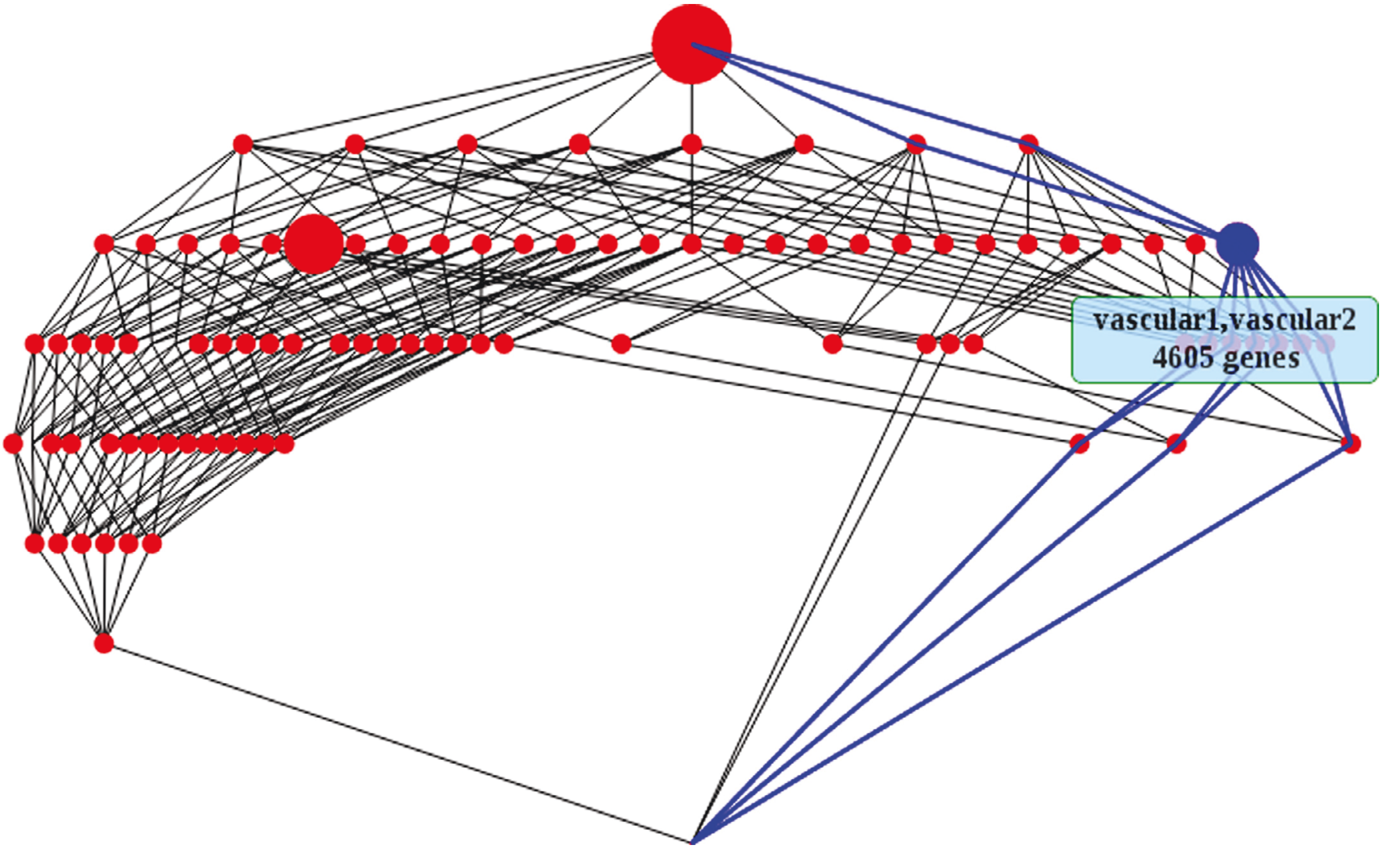


Fig. 9. (Color online) Gene over-expression lattice for $\phi = 0.15$. This lattice can be explored in an interactive way at https://webgenekfca.com/webgenekfca/kfcaresultses/4.

smooth muscle co-cluster (big right blue circle) and the *quadriceps* co-cluster (big left red circle) both in the second row. The cluster of over-expressed *striated* muscles is identified as the small red dot at the bottom on the left part of the lattice (eighth row, left). There are more than 4 000 over-expressed probesets in the *vascular smooth* muscle and again an analysis of the most over-expressed genes can be carried out.

Result: The use of max-plus analysis and min-plus analysis is an effective mechanism for exploring the phenomena of under-expression and over-expression in GED data, respectively. □

3.2.6. *Use case: interfacing other resources*

In this section we analyze an affordance that does not issue directly from the LofK metaphor by Wille. It comes from the work of Godin and collaborators (Godin, Gecsel, & Pichet, 1989; Godin, Saunders, & Gecsei, 1986) who very early realized that a concept lattice can be turned into an indexing device for an Information Retrieval system,

Affordance 14. A concept lattice is a content-based indexing device onto other knowledge resources.

Example. Recall the WebgeneKFCA tool introduced in Section 3.2.5.

1. **Indexing:** A very abstract scheme of interaction of the tool is drawn in Fig. 10. In there we see how WebGeneKFCA takes its inputs from a database manager of expression profiles for particular Genomes. By reading the meta-data from this DataBase Manager System (DBMS), WebGeneKFCA provides a brief description for each gene selected from a cluster. This description has references to the NCBI gene database[14] to ease the access to the latest online description of that gene.

   There are also references to each of the Gene Ontologies[15] the gene belongs to, with one link to each ontology description.

   Finally WebGeneKFCA shows if the gene is known to belong to a pathway and provides a link to KEGG pathway database [16].

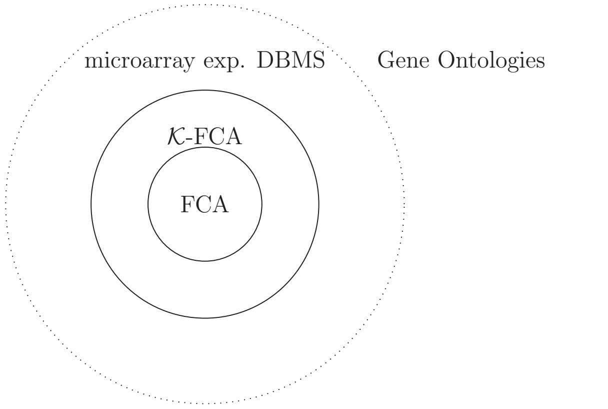


Fig. 10. An interpretation of the interaction between FCA, $\mathcal{K}$-FCA and external resources. In the example, gene ontologies are indexed by FCA.

Result: Concept lattices can be used to support interaction with other knowledge management tools like ontologies by indexing these based in extents and intents.

[14] http://www.ncbi.nlm.nih.gov/gene

[15] http://amigo.geneontology.org/amigo/

[16] http://www.genome.jp/kegg/pathway.html

Although using different extensions that $\mathcal{K}$-FCA, the semantic file systems of Ferré (2007) Martin (2004) Martin and Eklund (2005), and the post-retrieval clustering results of Carpineto and Romano (2004) 2005) are also proof of this. Examples of the use of FCA for the bibliographic analysis of different domains can be found in Poelmans et al. (2013a) Poelmans et al. (2013b). A more in-depth analysis of affordances of FCA for the analysis and synthesis of Information Retrieval systems is Valverde-Albacete and Peláez-Moreno (2013). □

4. Discussion

4.1. Other frameworks for EDA within Data Analysis

To analyze the suitability of LofK for EDA we will compare to other methods of EDA previously proposed. For this purpose, we believe that an explanation of conceptual frameworks for overall Data Analysis (DA), including both EDA and Confirmatory Data Analysis (CDA) is necessary. Note that the issue of what questions can DA answer is still being debated (Leek & Peng, 2015).

To make this comparison manageable we will use very basic Metaphor Theory (Lakoff, 1987) to make evident the metaphors underlying these frameworks. For other approaches to characterize DA, see, e.g. Džeroski (2007) Mannila (2000).

4.1.1. The Classical Approach of Data Analysis

The *classical framework* of EDA was originally proposed as a hypothesis-inducing stage of the whole data analysis process achieved through exploring the data with as wide an array of tools as possible (Tukey, 1977). We believe it is a blend of three metaphors (Behrens, 1997; Tukey, 1977)[17]:

- Data Analysis is Trial-by-jury
- Data Analysis is an Interaction
- EDA Tools are Back-of-the-Envelope Calculations

We briefly sketch these metaphors.

Data Analysis is Trial-by-jury. The “trial-by-jury” is a construct of the Anglo-Saxon culture that seems to have heavily influenced Statistics from its inception. This metaphor assigns different “legal roles” to the participants in the DA process: the suspects are the hypotheses issued from the data, the practitioner is a detective or prosecutor depending on the phase of the analysis, exploratory analyses and plots are exhibits to be presented as evidence in the trial, where statistical orthodoxy and practice is the judge, test statistics are the jury, and significance is the verdict.

In this metaphor, EDA is about indicting, CDA is about convicting. It is made explicit in different submetaphors: EDA is detective work (Tukey, 1977, p. 1), CDA is Prosecutor Work (directing detectives to collect evidence) (Behrens, 1997, p.132), and Final CDA is the trial Jury and Judge Work (Behrens, 1997). Interestingly, this metaphor is still productive in the sense that modern authors still talk of building tools as if they were a lineup (Buja et al., 2009)[18].

Standard statistical practice has learned to break away from this metaphor, by insisting that the data for EDA, the *training data*, be different to those for final CDA, the *test data*. It is also interesting to see how difficult this code of conduct is to grasp for new practitioners.

A very interesting variation of this framework is explained at length in Buja et al. (2009). They see two problems with classical EDA: the confirmation of purported “discoveries” and the control of over-interpretation of data. Since these seem to align with the traditional purposes of CDA they suggest to use *visual analogues* of the tools and procedures of CDA in the metaphor Visual EDA is Inferential Analysis.

This framework could be stated as follows: EDA is Inferential Analysis, Plots are Test Statistics, Human Cognition is Statistical Testing, Hypothesis Discovery is Contrasting Collected Data Plots vis-a-vis Random Data Plots, Discoveries are Rejections of Null Hypotheses. To avoid the over-interpretation of data, they develop the data analyst’s ability to detect novelty with methods designed to offset cognitive biases in the analyses, Calibration of the Discovery Process is Control of False Positives.

This comes from statisticians whose education has endowed them with the Inferential Analysis framework for DA, hence grounding EDA into CDA seems natural to them. But in fact the methods and tools of CDA are alien to the “default” human interpretation of data, the reason they were developed in the first place. And the question remains whether this is necessary for the EDA, born precisely as a reaction to the strictures of CDA.

Data Analysis is an Interaction. This legal metaphor may be clashing against another: it is a tenet of this type of trial-by-jury that no man can be judged twice for the same crime[19], but standard EDA practice seems to imply that there is a form of interaction between different stages of EDA and CDA. In this metaphor, an initial, rough EDA will be followed by a quick CDA, followed by a finer EDA and so on until the final CDA concludes the process. This is the kernel of the DA is an Interaction metaphor where EDA supplies the hypotheses and CDA accepts them or suggests they be revised (Tukey, 1980). In this metaphor there seems to be an implicit quantity that is accruing or refining, perhaps the validity of the analysis. Note that this interaction could also be conceived as a conversation.

Explicit consideration of this quantity as accruing, leads to an enriched metaphor DA is an Opening spiral where the interaction is actually used to accrue be it evidence, knowledge of the data, information, etc. We find no trace of the enriched metaphor in the first Tukey (Tukey, 1977), but the Analysis of Variance (ANOVA) has this flavor and the later Tukey considers it as crucial for EDA (Tukey, 1993). In there, the quantity accruing is the explained variance, but in modern incarnations of this metaphor the quantity may also be decreasing, like uncertainty in De Bie (2011).

EDA tools are back-of-the-envelope calculations. This metaphor is explicitly invoked by (Tukey, 1977, p. 1) but later modulated to include the possibility of (at the time scarce) computer aids (Tukey, 1977, ch. 20). From an initial emphasis on “pen-and-paper” it soon migrated to “procedure-based” view (Tukey, 1993) which fostered the advent of computer-supported toolsets and environments.

The embodiment of this classical framework seems to be that of a *statistical operating system*, that is a console onto a Read-Eval-Print-Loop (REPL)—in the spirit of LISP—and quick prototyping—perhaps with a customized user interface—and a plethora of different exploration tools available. The user environment of languages like S (Becker, Chambers, & Wilks, 1988), R (Ihaka & Gentleman, 1996) or Octave all agree with this description, and embrace open source and social software development practices—extensibility by means of packages, code repositories, community blogs, etc., although other software adopts the guise of standalone libraries or rapid prototyping environments. The analysis of the environments, user interfaces, and workflow types allowed by such tools would warrant a paper of its own.

Note that with complex graphic toolsets a price is paid: every plot needs a learning exercise on the part of the user, so they cannot be used profusely in communicating data analysis to a lay audience.

[17] But see also Data is a Data table below for a fourth metaphor that we believe is acting in this model.

[18] A legal procedure where a witness is requested to pick a suspect from a set of otherwise random people

[19] Without new evidence accruing, the finer point asserts.

4.1.2. *The approach of Data Mining*

The more modern field of Data Mining (DM) has also recognized the dichotomy between Exploratory Data Mining and Confirmatory Data Mining[20], but is also leaning towards Predictive Data Mining. This has to be understood in the context of the different questions that can be posed with a data analysis procedure (Leek & Peng, 2015).

Data Mining seems to be supported, at least, in the following metaphors:

- Knowledge extracted from data is gold
- Data is a Data table
- Data Analysis is an Interaction

The third metaphor is similar to that explained above, so we will concentrate in the other two.

Knowledge extracted from data is Gold. This metaphor inspires the name of "Data Mining" as well as frames the activity: Data Analyzing is Mining for gold, so A (data) Pattern is a Gold nugget, and the Interestingness of a pattern is the Value of the gold in the nugget.

Data is a Data table. This is a metaphor that is induced by the standard representation of data in the Database community—where the Knowledge Discovery in Databases community originated—who coined the expression "Data Mining". DM post-dates DA and many of its practitioners are trained in Statistics and DA, as well as Computer Science. To them, embodying data in a table that follows Codd's Relational Algebra (RA) (Codd, 1970) is as natural as considering Data analysis an Interaction.

We will consider that a metaphor backed up by a Mathematical formalism becomes a *model* that can be implemented, e.g. as RA is embodied into Relational Database Management Systems (RDBMS). Once the practitioner is knowledgeable with the implementation, this model makes operating with data extremely easy. So adopting the Data is a Data table model provides the ability to implement DM systems within RDBMS, which might explain the success of the DM community. Notice that this is an advantage that modern Statistical environments tend to adopt, e.g. the *dplyr* package within R (Wickham & Francois, 2015).

4.1.3. *Towards a mixing of metaphors*

Blending, mixing and actually identifying terms in a metaphor is natural for humans and a major mode of conceptualization and learning (Lakoff & Johnson, 1996). It is no wonder that there is a permeation of the metaphors of the most successful data analysis framework, DM towards classical (statistical) DA, acknowledging that statistical DA first informed DM.

For instance, modern embodiments of the statistical environments, embracing the back-of-envelope metaphor, have adopted technologies from Computer Science—like RA for data tables or incremental and open source toolset construction as in the CRAN mechanism for the R language—, Information Visualization—like the *grammar of graphics* inspired in the visual discipline promoted by Tufte (1992)—, and Machine Learning, like interactive plots with regression lines.

In another instance, the Data as a Data table model is to be contrasted to Data is an iid[21] Sample from a joint probability distribution of Classical Statistics and Machine Learning (Murphy, 2012): the mathematical apparatus backing up the latter is well-founded, but requires a longer training than RA to become proficient at, and can be very easily misused, which is why CDA was developed. Indeed we have waited until the DM metaphors were explained to introduce this model/metaphor, because Data is a Data table seems to be the more successful metaphor (in the number of practitioners) so far. On the contrary, using Data is an iid Sample from a joint probability distribution provides a unique advantage in solving questions of Predictive Data Analysis, and this is why Machine Learning, that embraces this model is flourishing.

That these models can be blended to advantage can be seen in the work by De Bie and co-authors (De Bie, 2011; De Bie & Spyropoulou, 2008). They have developed information-theoretic EDA in a number of papers tackling association rules, clustering, bi-clustering and other unsupervised machine learning tasks for which EDA is specially suitable. Their fundamental tenet is that the interestingness of patterns comes from the subjective knowledge state of the user, clearly an instance of the Data is Gold metaphor. The different types of data that they address suggest that they are using the Data is a Data table model. But they resort to Information-Theory and the Maximum Entropy Principle to iteratively induce joint distributions for the users preconceptions about the data, and this suggest that Data is an iid Sample of a distribution model. They concentrate in patterns that provide high compression rates vs. low descriptive complexity, to guide an iterative, and possibly interactive process of uncertainty reduction, like "sifting" through the information in the data. This manner of proceeding hints also that established metaphors can be improved upon and transformed into models very easily by means of different ways of understanding DA. And this is a meta-process of DA on the metaphors—exploratory and rough— and the models—confirmatory and final.

4.2. *A comparison of Landscapes of Knowledge to other EDA frameworks*

In this section we include both a comparison of LofK to other EDA frameworks and a critical analysis of the affordances of $\mathcal{K}$-FCA as a tool for EDA.

4.2.1. *Challenges for $\mathcal{K}$-FCA as a tool for EDA*

There are many techniques of EDA in the mainstream of DA, like (unsupervised) clustering (Mirkin, 1996), latent semantic analysis (Landauer, McNamara, Dennis, & Kintsch, 2007), multidimensional scaling, principal component analysis (Murphy, 2012), etc. Despite its privileged relation to association rule extraction (Ganter & Wille, 1999), FCA is not counted amongst them at present and in this section we provide our take in this matter.

First, note that standard Machine Learning libraries and environments, like Weka (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009), or languages, like R (Ihaka & Gentleman, 1996), do not have FCA facilities. Nor do FCA environments, like ConExp (Yevtushenko, 2000), include other standard unsupervised machine learning techniques, apart from association rule extraction. This lack of an open-source, standard toolkit is acknowledged and often regretted in the FCA community, and further efforts are needed to address it.

In our opinion, one feature that often misleads novel practitioners of FCA is its "formal" qualification. Although results obtaining from such lattices are often called "conceptual"—such as "conceptual clustering" or "conceptual search"—they should rather be read as "formal conceptual", that is, as pairings of sets of *abstract* objects and *abstract* attributes. "Abstract" in the previous sentence should be read in the same sense as in "abstract algebra": indeed, any pair of sets can be used instead of objects and attributes, even sets that are formally the same.

For instance, when documents are taken as objects and words or lemmas are taken as attributes one is tempted to declare the intents of formal concepts—that in this instance take the form of sets of such terms related to sets of documents—as a conceptual description in

[20] We will still refer to them as EDA and CDA, respectively for the reasons explained in the text.

[21] "Independent and Identically Distributed".

the usual epistemological sense[22]. To the best of our knowledge, this has not been granted by any research.

As soon as one becomes used to different instances of object-attribute pairs—and Section 3 presents many—this illusion of capturing "epistemological concepts" is dispelled. But it must account for a lot of people being attracted to FCA and getting equally disillusioned with it shortly afterwards. Since Science is essentially a sociological phenomenon, this might help explain why FCA in spite of having been available for the past thirty years has not been adopted in mainstream DA.

On the positive side, FCA is quite remarkable in the depth and number of results that it lays at the disposal of the practitioner when the data are binary and of high quality, as shown in Section 2.1. We have tried to make this evident in the use case of Section 3.2.1. On the negative, to the best of our knowledge, FCA has been so far unable to incorporate effectively a notion of *noise* in incidences. Our hope is that multi-valued extensions will eventually be able to bring noise into the picture to widen the range of applicability of the extended techniques. The difficulty here is rethinking the concept of noise in complete idempotent semifields, or semirings.

Clearly, FCA adheres to the DATA IS A DATA TABLE model, and a formal context, as a two-mode incidence relation with names on rows and columns is but another name for a matrix. As a boon, this is well-known to be a crypto-morphism of bipartite graphs and related to bi-clustering (co-clustering) (Hartigan, 1972; Mirkin, 1996). For DA, a matrix like a contingency table is a type of *wide data table*, to be contrasted to the *long data table* that normal forms of RA require (Codd, 1970), but modern data languages are endowed with primitives to transfer between the two, e.g `reshape` in R.

Furthermore, FCA adheres to the VISUAL INFERENTIAL ANALYSIS metaphor: concept lattices provide EXHIBITS for the analysis to proceed and the duality of contexts and lattices allows to transform the visual hypotheses into data hypotheses.

What seems to be missing from FCA practice is a serious attempt at formalizing the iterative process in the DA IS AN INTERACTION metaphor. This is a consequence of the FCA formalism not having any free parameters susceptible of adjustment. The only actual freedom of analysis available in FCA is which of the four modes in Section 2.3 to consider for eliciting the lattice, and this changes the interpretation of the data. Note that the alternate modes of analysis are not frequent even in FCA practice.

However, $\mathcal{K}$-FCA has the φ free parameter which we have used to define a flavor of exploratory analysis: changing this value the level of detail at which the context is being explored varies, entailing differences in the concept lattice. Indeed, Fig. 6 is an exploratory plot of the consequences of varying such parameter in the particular case of analyzing gene expression.

4.2.2. Challenges for LofK as a DA framework

Clearly, LofK is a metaphor comparable to the DA IS A TRIAL-BY-JURY and DA IS MINING FOR GOLD metaphors. We believe that it is more natural and suggestive, as a metaphor, for the average practitioner than DA IS A TRIAL-BY-JURY, if it is made to work as effectively. Also, metaphors carry with them not only good associations, but also the bad. We believe that for some time the DATA IS GOLD impinged subconsciously in many data providers, and this may have delayed the arrival of the Open Data and Open Science movements.

The adoption of FCA and its variants clearly indicates that the DATA IS A DATA TABLE (of the wide variety) is blended to it, and the fact that FCA is well-formalized seems to bid well for the uptake of LofK. But the lack of support for FCA as an EDA tool is a burden: analysts must rely on discipline and ingenuity to carry it out, and this entails the same sort of difficulty as adhering to the DA IS A TRIAL-BY-JURY metaphor.

We believe that a considerable amount of use cases have still to crop up before an attempt at giving support to LofK is successful. A gamut of such use cases are selected in this paper aimed at covering several issues ranging from the purely abstract and theoretical example of Section 3.2.1, to the simple and well-know perceptual task of segmented numerals—chosen for the availability of other published solutions to validate against—, to the as yet insufficiently explored problem of relation extraction, to the complexity of GED analysis and its wealth of $\mathcal{K}$-FCA exploratory affordances as demonstrated in a Web application.

Perhaps the most daunting task for FCA and its generalizations is to provide effective and efficient ways to use it as a tool in Confirmatory and Predictive Data Analysis. This would allow, among others, blending the DA AS INTERACTION metaphor into LofK, that is, we believe, the common divisor of all DA techniques. Barring this we believe FCA and LofK will be doomed to a niche tool in binary relation visualization.

4.3. Future research directions

The analyses in the previous sections suggest the following research directions:

- Developing FCA-implementing tool sets for mainstream libraries and languages, perhaps using the paradigm of Open Science, to increase the chance of mainstream acceptance for FCA.
- Strengthening the exploratory capabilities of FCA by developing a hierarchy of visualization alternatives for concept lattices that address the specific requirements of the meso- and macro-scale.
- Developing the geometry of semimodules over semirings, leading to new techniques of unsupervised and supervised analysis, for instance, those stemming from spectral and singular value decompositions that mimic the same techniques in conventional vector spaces. This and the previous research direction would enhance the uses of FCA for EDA.
- Developing Confirmatory and Predictive Analysis tools based in FCA that enable, for instance the blending of the DA AS AN INTERACTION with LANDSCAPES OF KNOWLEDGE. This will probably entail heavy research in the line below.
- Providing a satisfactory treatment of noise in semifields or semirings, that leads to models of contexts or lattice capable of accepting effect + noise decompositions. Combined with the previous capabilities this would blend all the affordances of DA IS AN OPENING SPIRAL with LANDSCAPES OF KNOWLEDGE, and would make it an alternative to DA IS TRIAL-BY-JURY and DA IS MINING GOLD.

5. Conclusions

Our analysis shows that Wille's metaphor of the Landscapes of Knowledge, as embodied in Formal Concept Analysis also seems to embody the metaphors implied by Tukey's proposals for Exploratory Data Analysis. The formalization of data inherent in formal context formation and its transformation into the concept lattice provide the basis to apply the Landscapes of Knowledge metaphor on scientific data as a kind of exploratory analysis.

This metaphor is well-aligned with Exploratory Data Analysis in the sense that concept lattice visualization and manipulation support a range of the activities related to scientific enquiry associated with knowledge elicitation and manipulation, and such procedures can be carried out on two-mode (matrix) data in whatever data-driven domain of knowledge. In this form, standard Formal Concept Analysis is capable of structuring and visualizing scientific knowledge to the point where it allows us to carry out even hypothesis formation (cfr. Section 3.2.1).

[22] In this case, related to the choice of documents and terms, the data instances as evinced in the incidence, etc.

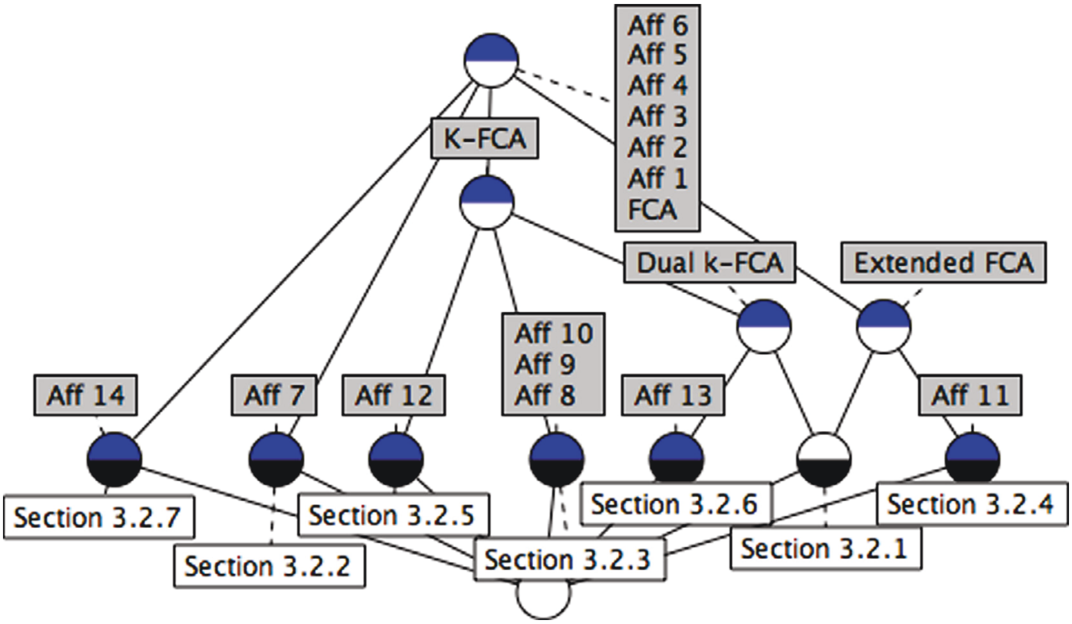


Fig. 11. Concept lattice describing the sections and the techniques and affordances of FCA and some of its extensions.

As a matter of fact, Fig. 11 summarizes the content of this paper in a very schematic way: by listing, for each section, what affordances of FCA and the extensions mentioned in the paper are treated therein.

The use of $\mathcal{K}$-Formal Concept Analysis allows the practitioner to encompass a wider range of datasets, including, for instance, probabilities or pointwise mutual information and variants thereof (cfr. Section 3.2.2) as well as (gene expression product) concentrations (cfr. Section 3.2.5). The idiosyncrasies of idempotent semifields allows them to model the increase or decrease of cost or concentration data, making them well-matched for exploring over- and under-expression in GED (cfr. Section 3.2.5). Furthermore, the use of extended FCA allows us to investigate different types of phenomena conceptualization: not only association between objects and attributes, but also rejection or negative evidence (cfr. Section 3.2.3).

In providing such an overarching grasp on data, we believe that formal concept analysis is on a par with vector spaces or coordinate spaces as a metaphor for extracting, visualizing and manipulation the scientific knowledge embodied in data. Parallel research to that presented here also points in this direction (Valverde-Albacete & Peláez-Moreno, 2008).

Acknowledgments

FJVA and AP were partially supported by EU FP7 project LiMoSINe (contract 288024) for this research. CPM was partially supported by the Spanish Ministry of Economics and Competitiveness projects TEC2014-61729-EXP and TEC2014-53390-P.

References

Affymetrix (2013a). Affymetrix power tools. http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx. Accessed 10.02.15. Affymetrix (2013b). What is a probeset? http://www.affymetrix.com/support/help/faqs/mouse_430/faq_8.jsp. Accessed 10.02.15.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). The new S language: a programming environment for data analysis and graphics. *Chapman & Hall/CRC*.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160.

B˘elohlávek, R. (2002). Fuzzy relational systems. Foundations and principles. *IFSR Int. Series on Systems Science and Engineering*: 20. Kluwer Academic.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., et al. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 367*(1906), 4361–4383.

Burusco, A., & Fuentes-González, R. (1994). The study of the L-fuzzy concept lattice. *Mathware and Soft Computing, 1*(3), 209–218.

Carpineto, C., & Romano, G. (2004). Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science, 10*, 8 . Carpineto, C., & Romano, G. (2005). *Concept data analysis: Theory and Applications*. Chichester, UK: John Wiley & Sons, Ltd.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377–387.

Cohen, G., Gaubert, S., & Quadrat, J.-P. (2004). Duality and separation theorems in idempotent semimodules. *Linear Algebra and Its Applications, 379*, 395–422. Congalton, R. G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC Press.

De Bie, T. (2011). An information theoretic framework for data mining. In *Proceedings of the 17th international conference on knowledge discovery in databases (KDD)* (pp. 564–572).

De Bie, T., & Spyropoulou, E. (2008). A theoretical framework for exploratory data mining: recent insights and challenges ahead. *Machine learning and knowledge discov-ery in databases* (pp. 612–616). Berlin, Heidelberg: Springer.

Deiters, K., & Erné, M. (2009). Sums, products and negations of contexts and complete lattices. *Algebra Universalis, 60*(4), 469–496.

Düntsch, I., & Gediga, G. (2002). Modal-style operators in qualitative data analysis. *Proceedings of the 2002 IEEE international conference on data mining, ICDM 2002* (pp. 155–162).

Düntsch, I., & Gediga, G. (2003). Approximation operators in qualitative data analysis. In H. de Swart, et al. (Eds.), *Theory and applications of relational structures as knowledge instruments*. In *Lecture Notes in Computer Science: 2929* (pp. 214–230). Springer.

Džeroski, S. (2007). Towards a general framework for data mining. *Analysis of time series data with predictive clustering trees* (pp. 259–300). Berlin, Heidelberg: Springer. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research, 30*(1), 207–210.

Eklund, P., & Villerd, J. (2010). A survey of hybrid representations of concept lattices in conceptual knowledge processing. *Formal Concept Analysis*, 296–311.

Erné, M., Koslowski, J., Melton, A., & Strecker, G. (1993). A primer on Galois connections. In A. Todd (Ed.), *Proceedings of the 1991 Summer Conference on General Topology and Applications in Honor of Mary Ellen Rudin and Her Work*. In *Annals of the New York Academy of Sciences: 704* (pp. 103–125). Madison, WI, USA: New York Academy of Science.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the conference of empirical methods in natural language processing. Scotland, UK*.EMNLP '11 Edinburgh.

Fano, R. M. (1961). *Transmission of information: a statistical theory of communication*. The MIT Press.

Ferré, S. (2007). Camelis: Organizing and browsing a personal photo collection with a logical information system. In *Proceedings of the 5th international conference on concept lattices and their applications, (CLA07)* (pp. 112–123).

Ferré, S., & Ridoux, O. (2001). Searching for objects and properties with logical concept analysis. In H. Delugach, & G. Stumme (Eds.), *Proceedings of the ICCS 2001*. In *LNAI: 2120* (pp. 187–201). Berlin Heidelberg: Springer-Verlag.

Ganter, B., & Wille, R. (1999). *Formal concept analysis: mathematical foundations*. Berlin, Heidelberg: Springer.

Godin, R., Gecsel, J., & Pichet, C. (1989). Design of a browsing interface for information retrieval. *Proceedings of the 12th international conference on research and development in information retrieval (ACM SIGIR '89)* (pp. 32–39). Cambridge, MA: ACM.

Godin, R., Saunders, E., & Gecsei, J. (1986). Lattice model of browsable data spaces. *Information Sciences, 40*, 89–116.

Golan, J. S. (1999). *Semirings and their applications*. Kluwer Academic.

González-Calabozo, J., Peláez-Moreno, C., & Valverde-Albacete, F. J. (2011). Gene expression array exploration using $\mathcal{K}$-formal concept analysis. In P. Valtchev, & R. Jäschke (Eds.), *Proceedings of the ICFCA11*. In *LNAI: 6628* (pp. 119–134). Springer.

González-Calabozo, J., Peláez-Moreno, C., & Valverde-Albacete, F. J. (2012). WebGeneKFCA: an on-line conceptual analysis tool for genomic expression data. In L. Szathmary, & U. Priss (Eds.), *Proceedings of the ninth conference on concept lattices and applications (CLA 07)* (pp. 345–346). Fuengirola, Málaga: Universidad de Málaga.

Google (2013). Freebase data dumps. https://developers.google.com/freebase/data. Accessed 26.05.13.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations Newsletter, 11*(1), 10–18.

Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association, 67*(337), 123–129.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299–314.

Ji, H., Grishman, R. Dang, H. T. (2011). Overview of the TAC2011 knowledge base population track.

Keren, G., & Baggen, S. (1981). Recognition models of alphanumeric characters. *Perception & Psychophysics, 29*(3), 234–246.

Konecny, J. (2011). Isotone fuzzy galois connections with hedges. *Information Sciences, 181*(10), 1804–1817.

Lakoff, G. (1987). *Women Fire and dangerous things*. University Of Chicago Press.

Lakoff, G., & Johnson, M. (1996). *Metaphors we live by*. University Of Chicago Press.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Lawrence Erlbaum Associates.

Leek, J. T., & Peng, R. D. (2015). Statistics. what is the question? *Science, 347*(6228), 1314–1315.

Mannila, H. (2000). Theoretical frameworks for data mining. *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations Newsletter, 1*(2), 30–32.

Martin, B. (2004). Formal concept analysis and semantic file systems. In P. Eklund (Ed.), *Concept lattices* (pp. 88–95). Springer Berling Heidelberg.

Martin, B., & Eklund, P. (2005). Applying formal concept analysis to semantic file systems leveraging wordnet. In *Proceedings of the 10th australasian document computing symposium* (pp. 1–8).

Matlab (2012). Matlab and statistics toolbox release 2012b. Natick, Massachusetts, USA, The MathWorks Inc.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics, 11*(1), 31–46.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *Journal of the Acoustical Society of America, 27*(2), 338–352.

Mirkin, B. (1996). Mathematical classification and clustering. *Nonconvex Optimization and Its Applications*: 11. Kluwer Academic Publishers.

Mirkin, B. (2005). Clustering for data mining. a data recovery approach. *Computer Science and Data Analysis Series*. Boca Raton (FL), US: Chapman & Hall.

Murphy, K. P. (2012). *Machine learning. a probabilistic perspective*. MIT Press.

Pedraza-Jiménez, R., Valverde-Albacete, F. J., & Navia-Vázquez, A. (2006). A generalisation of fuzzy concept lattices for the analysis of web retrieval tasks. In *Proceedings of the international conference on information processing and management of uncertainty in knowledge-based systems*.(IPMU'06)

Peláez-Moreno, C., García-Moral, A. I., & Valverde-Albacete, F. J. (2009). Eliciting a hierarchical structure of human consonant perception task errors using formal concept analysis. In *Proceedings of the interspeech* (pp. 828–831).

Peláez-Moreno, C., García-Moral, A. I., & Valverde-Albacete, F. J. (2010). Analyzing phonetic confusions using formal concept analysis. *Journal of the Acoustical Society of America, 128*(3), 1377–1390.

Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., & Dedene, G. (2013a). Formal concept analysis in knowledge processing: a survey on applications. *Expert Systems with Applications, 40*(16), 6538–6560.

Poelmans, J., Kuznetsov, S. O., Ignatov, D. I., & Dedene, G. (2013b). Formal concept analysis in knowledge processing: a survey on models and techniques. *Expert Systems with Applications, 40*(16), 6601–6623.

R Core Team, (2014). R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Tufte, E. R. (1992). *The visual display of quantitative information*. Graphics Press.

Tukey, J. W. (1977). Exploratory data analysis. *Behavioral Science series*. Addison Wesley.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician, 34*(1), 23–25.

Tukey, J. W. (1993). Exploratory data analysis: past, present and future. *Technical report 302*. Princeton University: Department of Statistics.

Valverde-Albacete, F. J. (2008). Extracting frame-semantics knowledge using lattice theory. *Journal of Logic and Computation, 18*(3), 361–384.

Valverde-Albacete, F. J. (2010). Detecting features from confusion matrices using generalized formal concept analysis. *Hybrid artificial intelligence systems* (pp. 375–382). Peláez-Moreno, C.: Springer Berlin Heidelberg.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2006). Towards a generalisation of formal concept analysis for data mining purposes. In R. Missaoui, & J. Schmid (Eds.), *Proceedings of the international conference on formal concept analysis, ICFCA06, (Dresden, Germany)*. In *Lecture Notes on Artificial Intelligence: 3874* (pp. 161–176). Berlin, Heidelberg: Springer.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2007a). Further Galois connections between semimodules over idempotent semirings. In J. Diatta, & P. Eklund (Eds.), *Proceedings of the 4th conference on concept lattices and applications (cla 07), Montpellier* (pp. 199–212).

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2007b). Galois connections between semimodules and applications in data mining. In S. Kusnetzov, & S. Schmidt (Eds.), *Formal Concept Analysis. Proceedings of the 5th international conference on formal concept analysis, ICFCA 2007*. In *LNAI: 4390* (pp. 181–196). Clermont-Ferrand, France: Springer.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2008). Spectral lattices of $(R_{max,+})$-formal contexts. formal concept analysis. In *Proceedings of the 6th international conference, ICFCA 2008, Montreal, Canada*. In *Lecture Notes in Computer Science (LCNS 4933)* (pp. 124–139).February 25–28, 2008.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2011). Extending conceptualisation modes for generalised formal concept analysis. *Information Sciences, 181*, 1888–1909.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2013). Systems vs. methods: an analysis of the affordances of formal concept analysis for information retrieval. In *Proceedings of formal concept analysis meets information retrieval (fcair)*.Workshop co-located with ECIR-2013 (pp. Best workshop paper award).

Wickham, H. Francois, R. (2015). dplyr: a grammar of data manipulation. R package version 0.4.3.

Wille, R. (1995). The basic theorem of triadic concept analysis. *Order, 12*(2), 149–158.

Wille, R. (1999). Conceptual landscapes of knowledge: a pragmatic paradigm for knowledge processing. In W. Gaul, & H. Locarek-Junge (Eds.), *Classification in the information age, studies in classification, data analysis, and knowledge organization* (pp. 344–356). Springer Berlin Heidelberg.

Wille, R. (2006). Methods of conceptual knowledge processing. *Formal Concept Analysis*, 1–29.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). *SOAPdenovo-trans: de novo transcriptome assembly with short rna-seq reads.* (pp. 1–7). Oxford, England: Bioinformatics.

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., & Soderland, S. (2007). Textrunner: open information extraction on the web. *Proceedings of the human language technologies. the annual conference of the north american chapter of the association for computational linguistics: demostrations* (pp. 25–26). ACL.

Yevtushenko, S. A. (2000). 127–134, System of data analysis “Concept Explorer”. In Proceedings of the 7th national conference on Artificial Intelligence KII-2000, 127–134. (In Russian) http://sourceforge.net/projects/conexp. Accessed 10.02.15.