Method of knowledge extraction applied to post-graduate computer science studies

Overview of CREA method

LSE – Lightning Talks

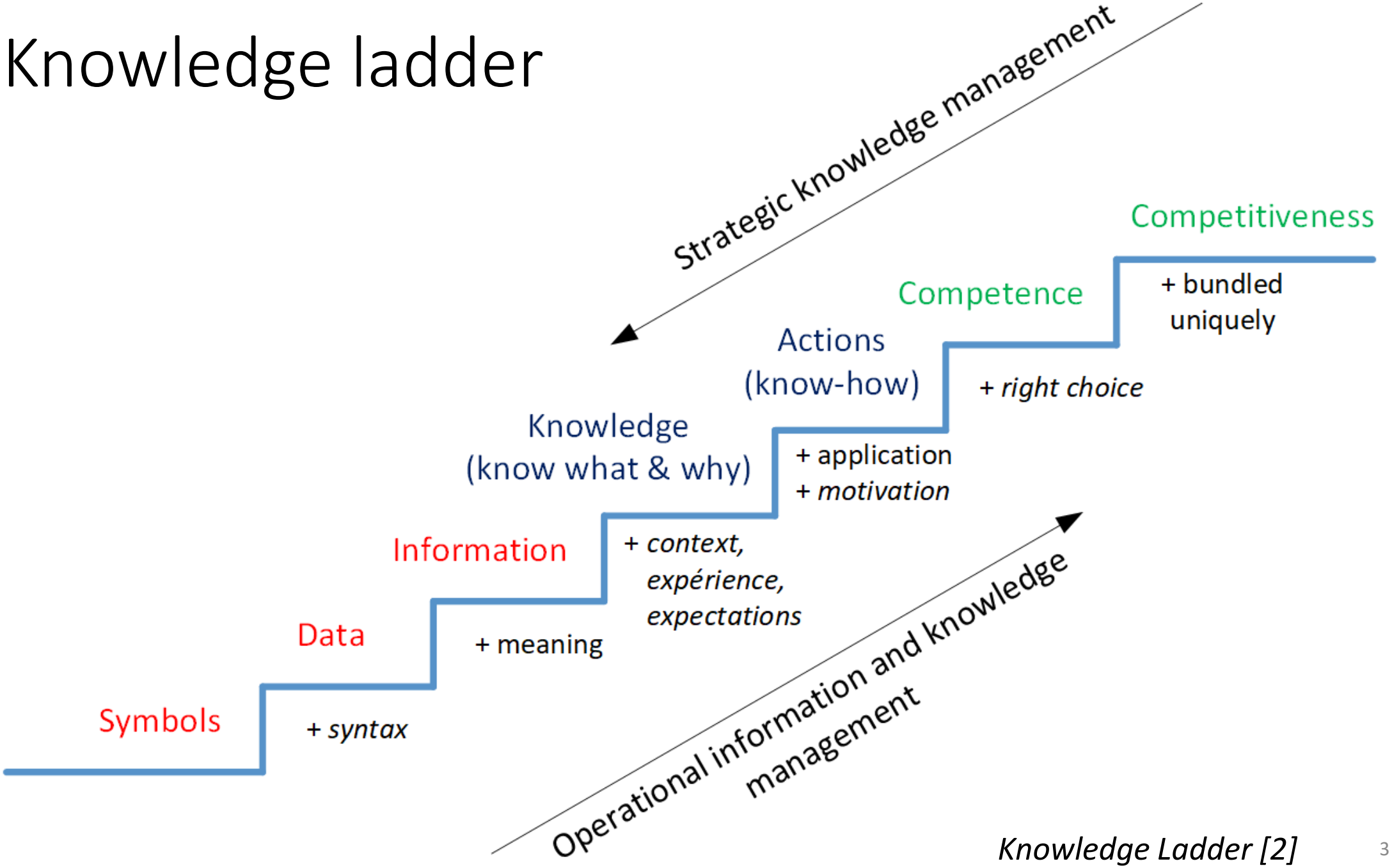
Fabrice BOISSIER

04 avril 2022

Knowledge and teaching

- How to help a teacher build a new course?
  - Ask for colleagues who have taught the same topic
  - Gather previous course materials and/or from other teachers
  - Find books, articles, … talking about the same topic
- *Knowledge Intensive Process*
  (not just a regular Business Process)

Knowledge ladder

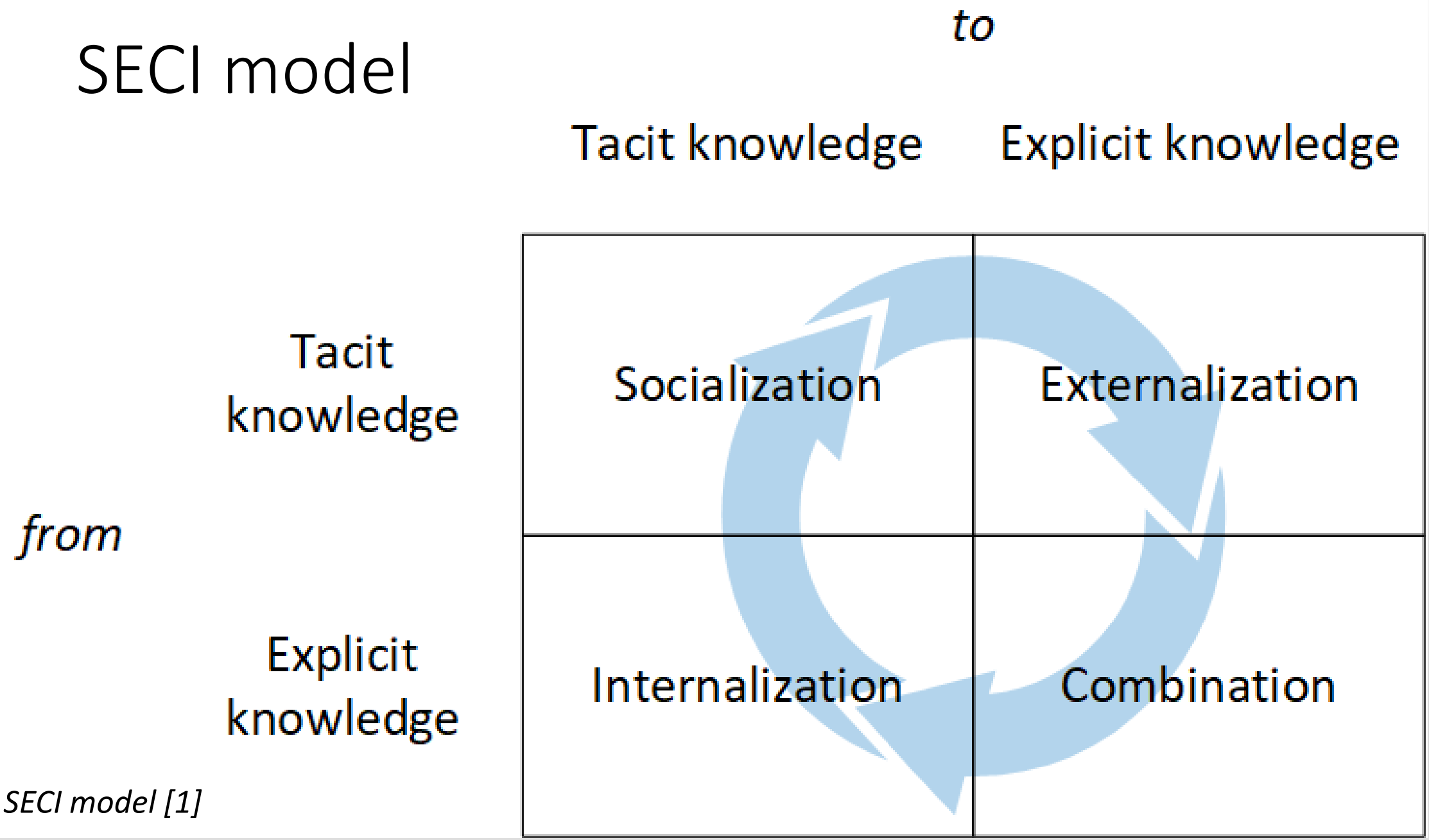


Knowledge Ladder [2]

Knowledge

- Knowledge: explicit and tacit [1][2][3]
  - Explicit knowledge:
    structured, codified, formalised with schemas, formulas, texts, …
  - Tacit knowledge:
    intern to each person, depends on the 5 senses and past experiences
  - « *Knowledge is not an object, [but it] exists in interaction, is linked and created through actions, requires an interpretative framework* » [3]

SECI model



SECI model [1]

SECI model

- Combination : (explicit → explicit)
  - Teacher is combining multiple books, articles, existing course materials, …
  - Student searches for different information sources than the course (wikipedia, books, …)
- Internalization : (explicit → tacit)
  - Teacher gives lecture to students
  - Student uses notions and appropriates them during exercises, labs, projects
- Socialization : (tacit → tacit)
  - Students works together on a topic
  - Teacher discusses with other teachers
  - Teacher helps a particular student on an exercise
- Externalization : (tacit → explicit)
  - Teacher adapts its course material for the next session (or next year)
  - Student prepares a presentation or a report on its project

Knowledge and teaching

- How to help a teacher build a new course?

- Help him by automatizing *combination* of documents
  - Show the keywords of the documents
  - Show graphically which document(s) are irrelevant
  - Build clusters of related notions for each session of the course

CREA method

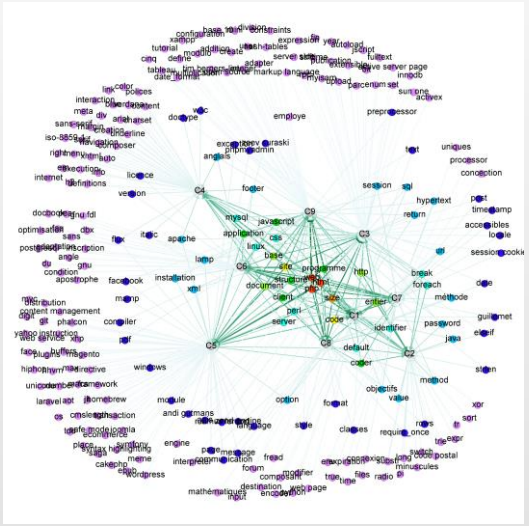
Semantic Pre-Processing

- Extraction of the texts from chosen documents
- Disambiguation with NLP techniques
- Preparation of the texts for structural analysis

Structural Analysis

- Calculation of metrics with FCA
- Relevancy of input documents
- Building of clusters of terms

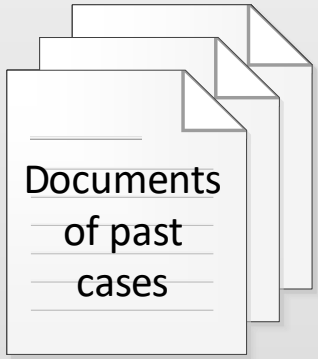
Mutual impact graph



C1	Web	Site		
C2	HTML	HTTP	PHP	JS
C3	Code	Foreach	If	
C4	Cookie	Session	Time	

Clusters of terms

CREA method: Semantic Pre-Processing



Selection of documents by the user

Extraction of texts

Cleaning of extracted texts

Disambiguation

Filtering of terms

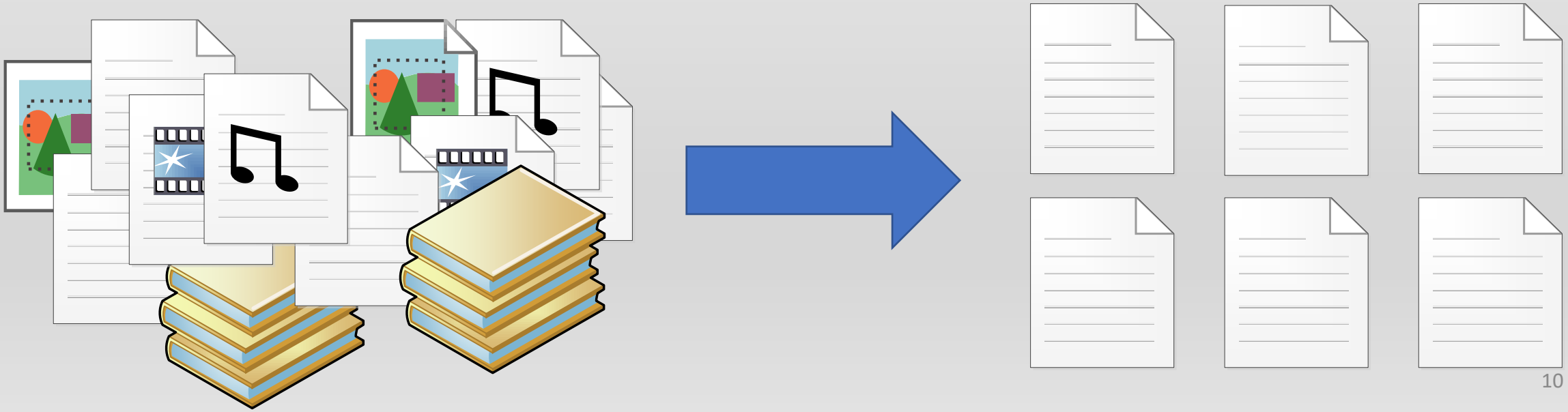
	T1	T2	T3	T4
D1	0	36	0	2
D2	7	14	5	0
D3	1	8	12	1

Occurrence matrix

1 - CREA method: Semantic Pre-Processing

Selection of documents by the user

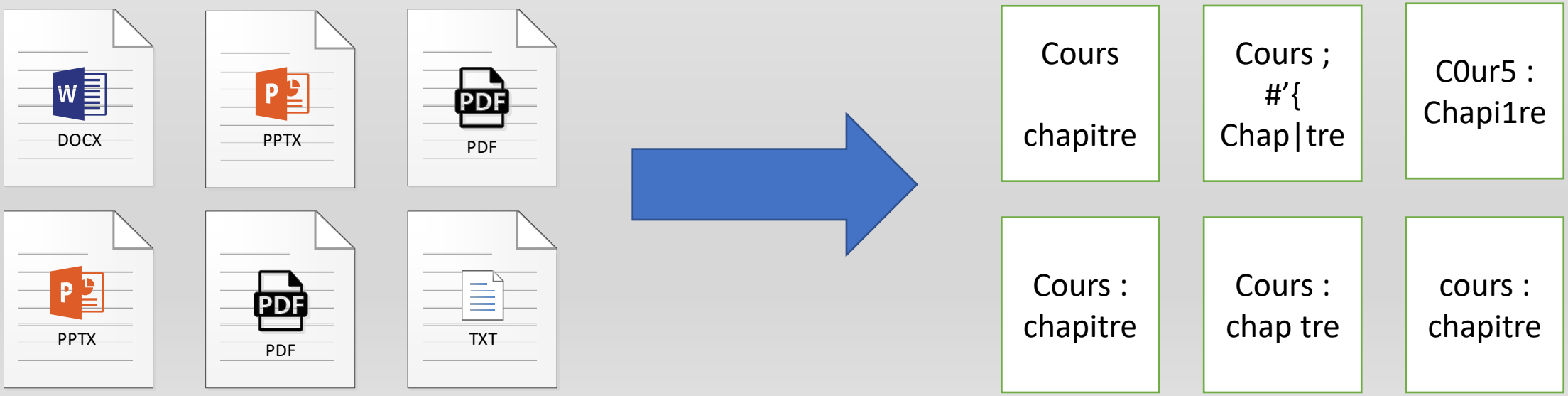
- Select documents based on their title, abstract, or syllabus
- Documents must be in a text format with enough content to analyze *(pictures and arrays are not managed)*



1 - CREA method: Semantic Pre-Processing

Extraction of texts

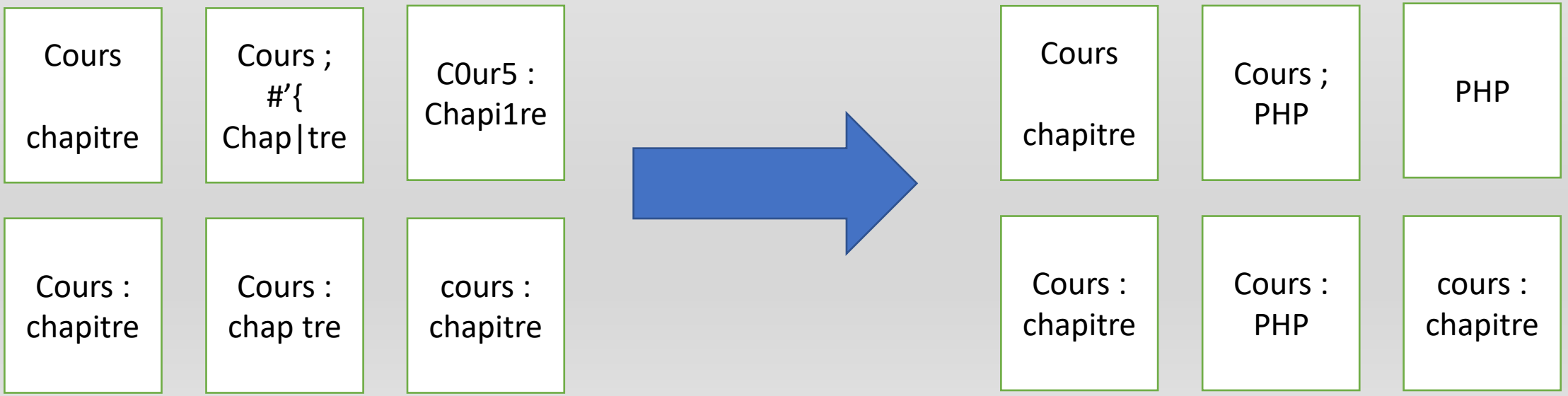
- Standardization from various input format into flat text
- *[Usage of an OCR and/or PDF extractor, but it's not the objective to dig these techniques]*



1 - CREA method: Semantic Pre-Processing

Cleaning of extracted texts

- Removes non printable characters, symbols, useless grammar classes
- TreeTagger [5] for POS tagging and a list of grammar classes to keep



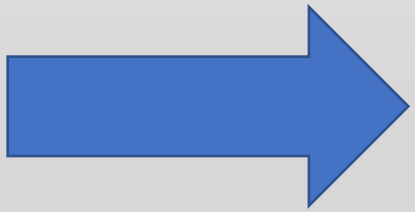
1 - CREA method: Semantic Pre-Processing

Disambiguation

- Find concepts and named entities from the text of each document
- Word Sense Disambiguation and Entity Linking with BabelFy [6]
  (so, with BabelNet [7]… and therefore with Wikipedia)

Cours : chapitre PHP

Langage de programmation initialement dédié aux pages webs personnelles.
Se lie facilement à une base de données utilisant le SGBD MySQL.
…



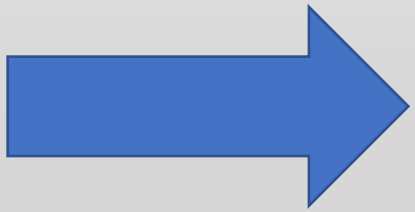
Cours ; 0,1 ; bn:XXX ; (1,5)
chapitre ; 0,04 ; bn:XXX ; (9,16)
PHP ; 0,98 ; bn:XXX ; (18,20)
…
Programmation ; 0,96 ; bn:XXX ; (,)
…
Base de données ; 0,94 ; bn:XXX ; (,)
…

1 - CREA method: Semantic Pre-Processing

Filtering of terms

- Removes irrelevant terms
- Keeps terms that have a coherence score above 0,05 (empirical value)

Cours ; 0,1 ; bn:XXX ; (1,5)
chapitre ; 0,04 ; bn:XXX ; (9,16)
PHP ; 0,98 ; bn:XXX ; (18,20)
…
Programmation ; 0,96 ; bn:XXX ; (,)
…
Base de données ; 0,94 ; bn:XXX ; (,)
…



Cours ; 0,1 ; bn:XXX ; (1,5)
PHP ; 0,98 ; bn:XXX ; (18,20)
Programmation ; 0,96 ; bn:XXX ; (,)
Base de données ; 0,94 ; bn:XXX ; (,)
…

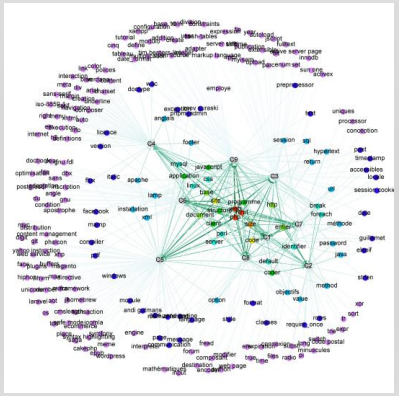
CREA method: Structural Analysis

	T1	T2	T3	T4
D1	0	36	0	2
D2	7	14	5	0
D3	1	8	12	1

Occurrence matrix

Formal Concept Analysis

Construction of the mutual impact graph



Mutual impact graph

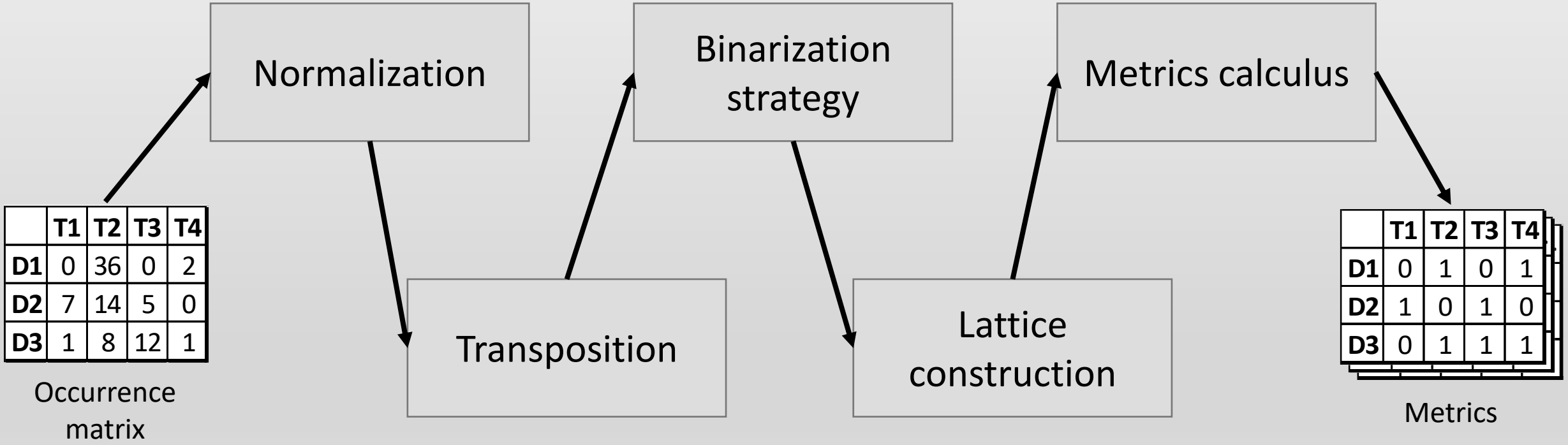
Construction of the clusters

C1	Web	Site		
C2	HTML	HTTP	PHP	JS
C3	Code	Foreach	If	
C4	Cookie	Session	Time	

Clusters of terms

2 - CREA method: Structural Analysis

Formal Concept Analysis [8]



	T1	T2	T3	T4
D1	0	36	0	2
D2	7	14	5	0
D3	1	8	12	1

	T1	T2	T3	T4
D1	0	1	0	1
D2	1	0	1	0
D3	0	1	1	1

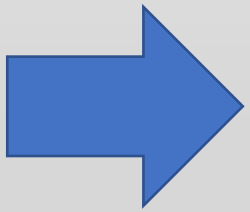
2 - CREA method: Structural Analysis

A - Formal Concept Analysis: Normalization

- Makes the values independent of the length of documents
- Calculate proportions of occurrences per documents (%)

	web	php	sql	mysql
Cours 1	10	10	10	10
Cours 2	1	2	2	0
Cours 3	0	0	1	0

Occurrence matrix



	web	php	sql	mysql
Cours 1	25	25	25	25
Cours 2	20	40	40	0
Cours 3	0	0	100	0

Normalized matrix

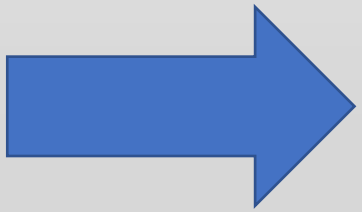
2 - CREA method: Structural Analysis

B - Formal Concept Analysis: Transposition

- Change the point of view
  - *From :* each document containing terms
  - *To :* each term present in documents

	web	php	sql	mysql
Cours 1	25	25	25	25
Cours 2	20	40	40	0
Cours 3	0	0	100	0

Normalized matrix



	Cours 1	Cours 2	Cours 3
web	25	20	0
php	25	40	0
sql	25	40	100
mysql	25	0	0

Normalized transposed matrix

2 - CREA method: Structural Analysis

C - Formal Concept Analysis: Binarization strategy

- Build a « Formal Context » with binarization strategies [4]

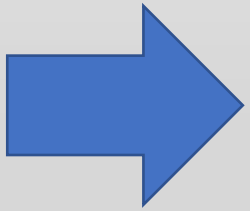
- Transforms a multivaluated matrix into binary matrices
- Two interesting matrices:
  - Matrix of presence/lack of terms
  - Matrix of higher frequencies of presence of terms

	Cours 1	Cours 2	Cours 3
web	25	20	0
php	25	40	0
sql	25	40	100
mysql	25	0	0

Normalized transposed matrix

Frequencies matrix

	Cours 1	Cours 2	Cours 3
web	56%	44%	0%
php	38%	62%	0%
sql	15%	24%	61%
mysql	100%	0%	0%



	Cours 1	Cours 2	Cours 3
web	1	1	0
php	1	1	0
sql	1	1	1
mysql	1	0	0

Direct strategy

High strategy (β = 0,50)

	Cours 1	Cours 2	Cours 3
web	1	0	0
php	0	1	0
sql	0	0	1
mysql	0	0	0

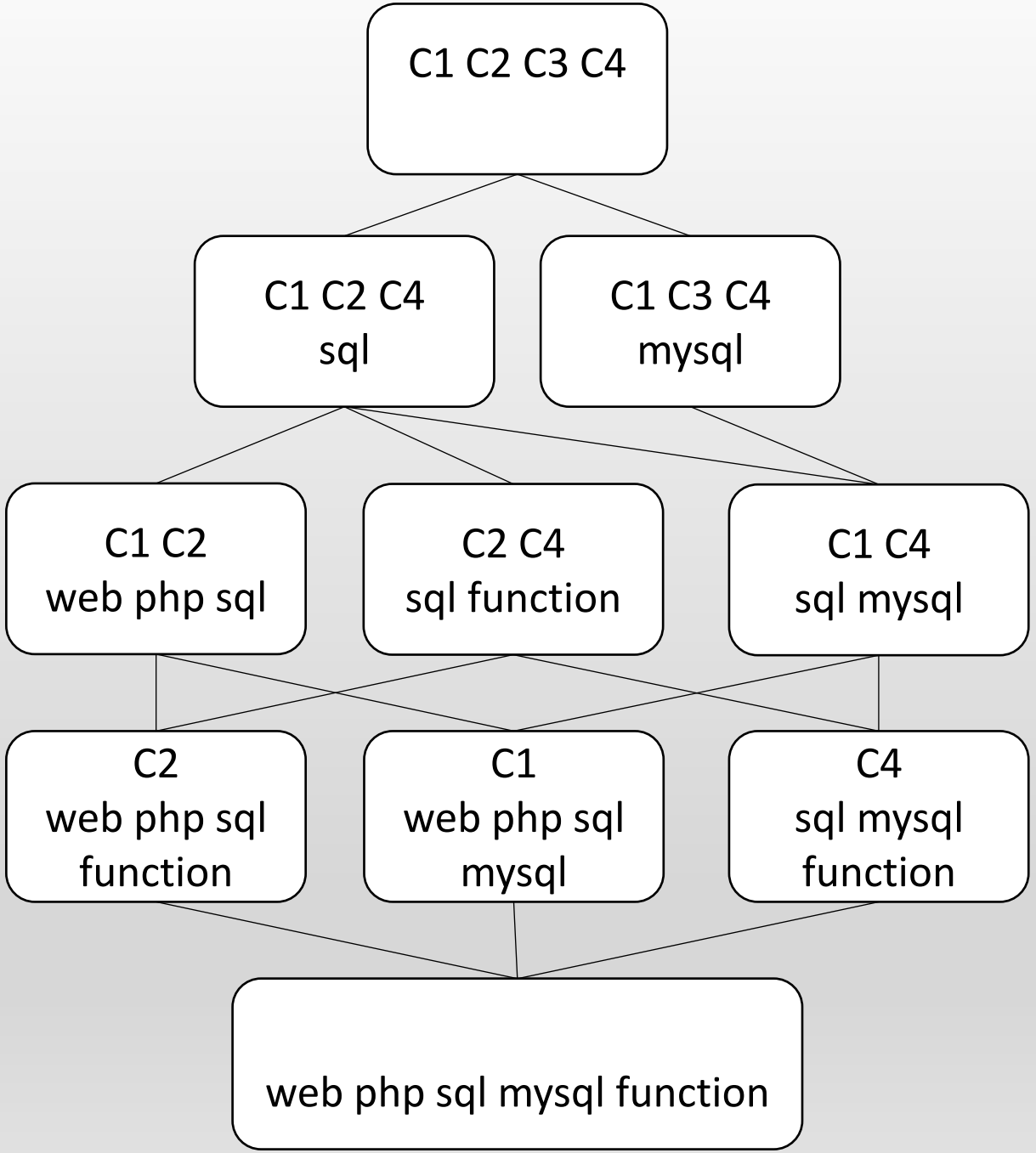
2 - CREA method: Structural Analysis

D - Formal Concept Analysis: Lattice construction

- Prepare data for the metrics to calculate
- Build a lattice from a Formal Context
  - Objects: terms
  - Attributes: documents
  - « Terms » are described by their presence within « Documents »

	C1	C2	C3	C4
web	1	1	0	0
php	1	1	0	0
sql	1	1	0	1
mysql	1	0	1	1
function	0	1	0	1

Formal Context



2 - CREA method: Structural Analysis

E - Formal Concept Analysis: Metrics calculus

- Calculates mutual impact and conceptual similarity metrics in lattice

- Mutual Impact [4]:

$$\mathrm{MI}(O_i, A_j) = \frac{Nb\ of\ Concepts\ containing\ O_i\ \boldsymbol{AND}\ A_j}{Nb\ of\ Concepts\ containing\ O_i\ \boldsymbol{OR}\ A_j}$$

- Conceptual Similarity [4]:

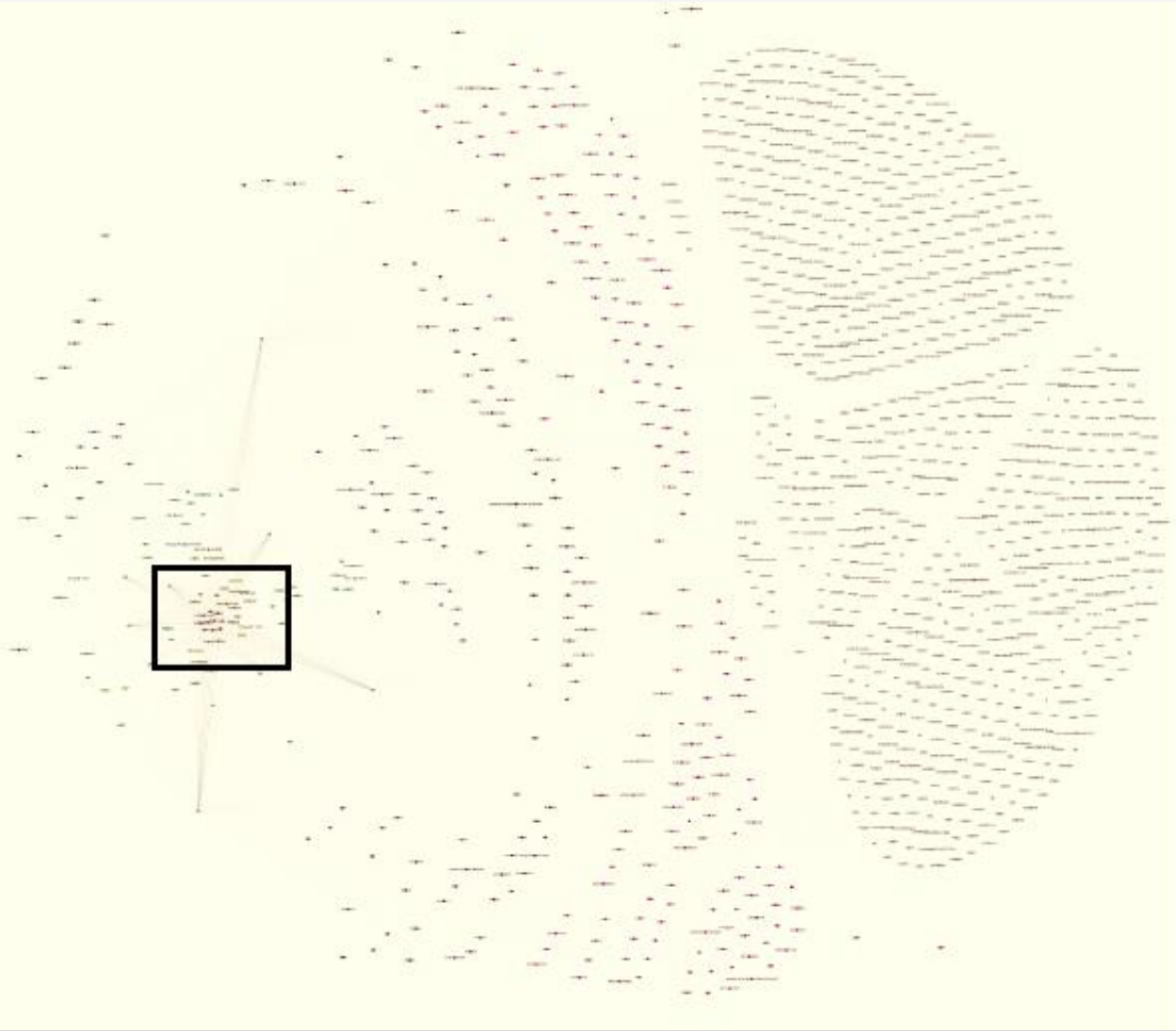
$$\mathrm{CS}(O_i, O_j) = \frac{Nb\ of\ Concepts\ containing\ O_i\ \boldsymbol{AND}\ O_j}{Nb\ of\ Concepts\ containing\ O_i\ \boldsymbol{OR}\ O_j}$$

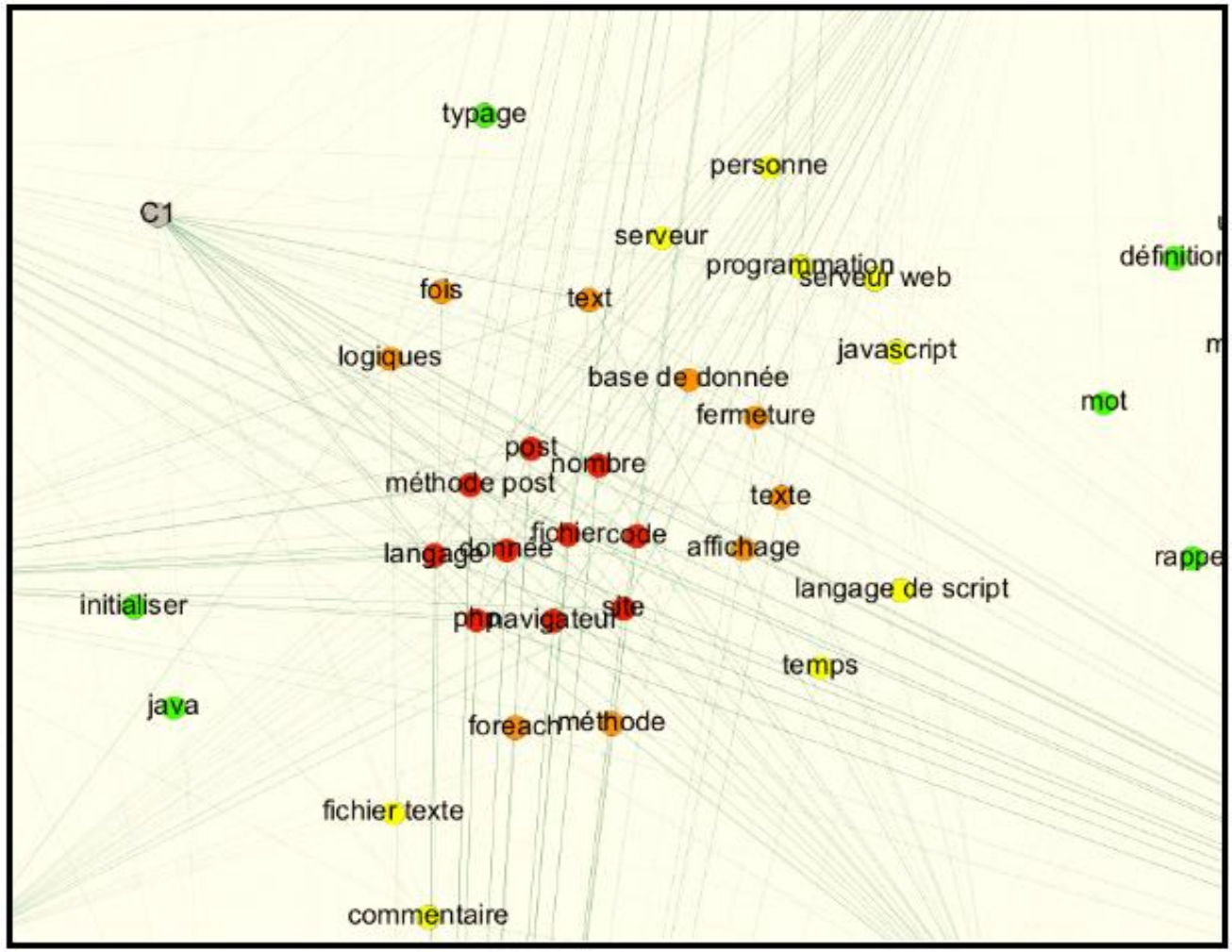
2 - CREA method: Structural Analysis

Construction of the mutual impact graph

- Show graphically:
  - Relevancy of each document
  - Keywords of the documents

- Gephi on Mutual Impact matrix (Terms X Documents)
  - Force Atlas spatialization (force-based/force-directed)
  - Nodes produces repulsive force
  - Edges produces attractive force




typage
personne
C1
serveur
programmation
définition
serveur web
fois
text
logiques
javascript
base de donnée
mot
fermeture
post
nombre
méthode post
texte
fichiercode
langage
donnée
affichage
rappe
langage de script
initialiser
site
php
navigateur
temps
java
foreach
méthode
fichier texte
commentaire


C6
C9
C8
C1
C2
C7
C4
C5
C3

2 - CREA method: Structural Analysis

Construction of the clusters

- Proposes group of notions to explain/present together
- Group terms based on their similarity in documents
  - Regular HCA [9] for non-overlapping clusters

Stratégie Haute (β = 1,00)

1	php	code	fois	post	jour	foreach	cle	classe	class	mysqli	
2	page web	navigateur	serveur web	texte	concerner	délimiter	utilisateur	associer	personne	machine	mysql
3	url	langage	case	fermeture	session	chaîne	entête	avoir accès			
4	fichier	commentaire	case à cocher	interpréter	côté serveur	serveur	côté client				
5	typage	mot	moteur	affiche	transaction	visiteur					
6	base de donnée	insert	varchar	null							
7	xml	configuration	composer	doctype							
8	donnée	text	méthode post	programmation	site	langage de script	list	méthode	timestamp	files	

Experiments

- 5 scenarios

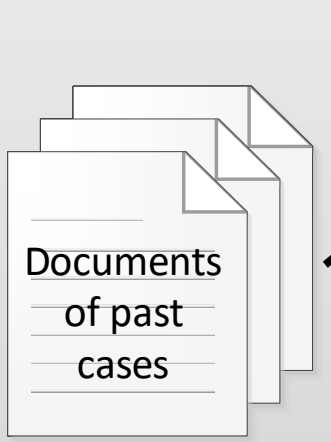
- PHP courses [french]
  - 9 docs: 6 Slides, 3 Texts
  - 10 docs: [9 PHP] + 1 Java text
  - 18 docs: 11 Slides, 7 Texts
  - 7 docs: 7 Texts [Correction of a document]
- Statecharts
  - 13 documents of various nature in english (webpage, article, slides, …)

Limitations

- Unable (for now) to dissociate text from meta-data
- Filtering with TreeTagger is probably useless as BabelFy might does it
- The generation of the clusters is currently difficult to manage
  - Too many terms are inside clusters in some cases

- Format of documents has an impact on the graph
- Only regular text is managed…
  - What about code? Arrays? Pictures?
- Experiments done only on computer science courses
- Clusters are difficult to read for a beginner in the domain
  - It is required to have a minimal knowledge of the domain managed

Future Work: Temporal analysis



Semantic Pre-Processing

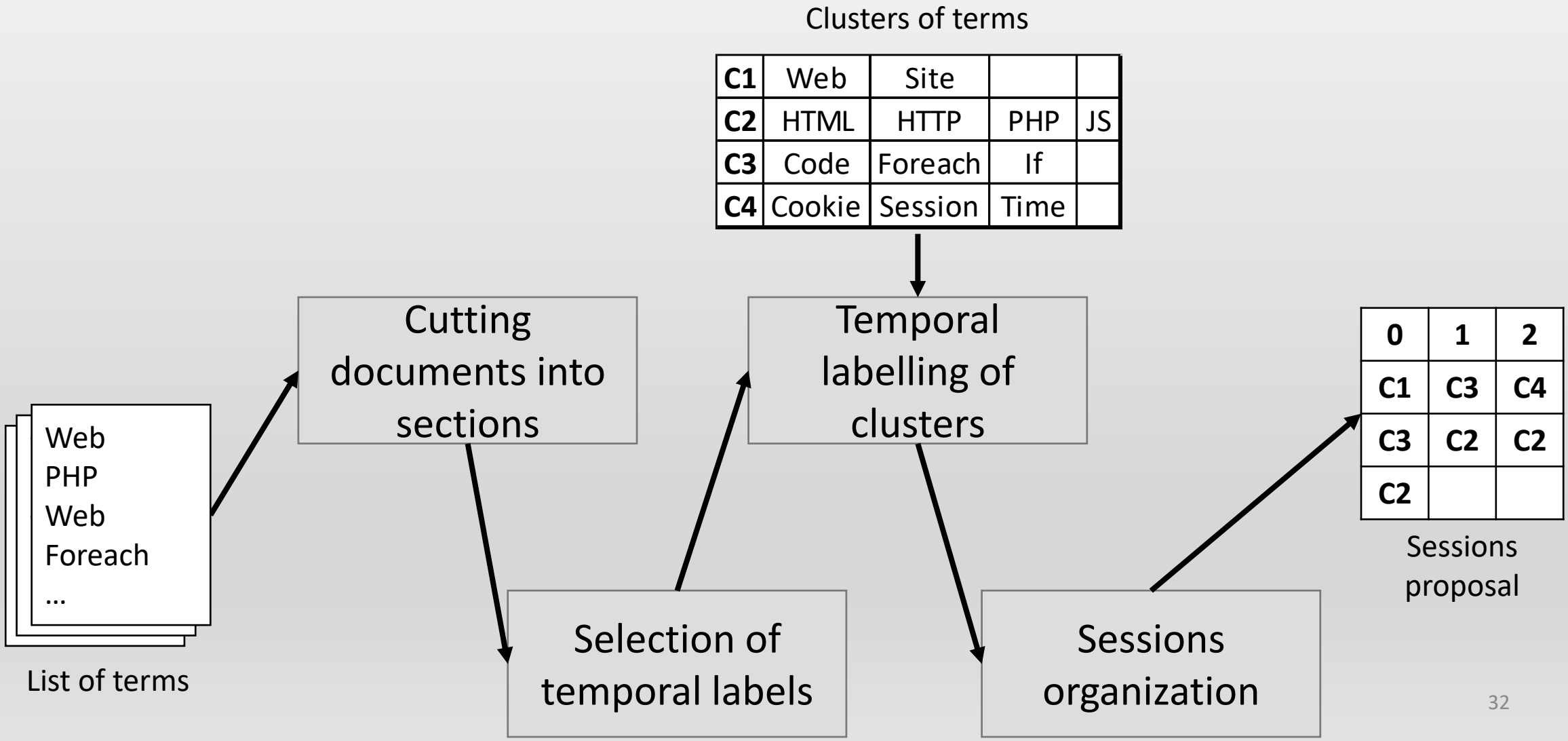
Structural Analysis

Temporal Analysis

0	1	2
Web, PHP, HTML	PHP, Foreach, If	Cookies, Session, Time
PHP, Foreach, If	Web, HTTP, HTML, JS	Web, HTTP, HTML, JS
Web, HTTP, HTML, JS		

Sessions proposal

Future Work: Temporal analysis



Thanks for your attention

Knowledge:

[1] Ikujiro Nonaka and Hirotaka Takeuchi. The knowledge-creating company. Harvard business review, 85(7/8) :162, 2007.

[2] Klaus North and Gita Kumta. Knowledge management : Value creation through organizational learning. Springer, 2018.

[3] Jawad Syed, Peter A Murray, Donald Hislop, and Yusra Mouzughi. The Palgrave handbook of knowledge management. Springer, 2018.

Data science and Natural Language Processing:

[4] Ali Jaffal. Aide à l'utilisation et à l'exploitation de l'Analyse de Concepts Formels pour des non-spécialistes de l'analyse des données. PhD thesis, Université Panthéon - Sorbonne - Paris I, 2019.

[5] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In New methods in language processing, page 154, 1994.

[6] Roberto Navigli and Simone Paolo Ponzetto. Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193 :217–250, 2012.

[7] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation : a unified approach. Transactions of the Association for Computational Linguistics, 2 :231–244, 2014.

[8] Rudolf Wille. Restructuring lattice theory : An approach based on hierarchies of concepts. In Ivan Rival, editor, Ordered Sets, volume 83 of NATO Advanced Study Institutes Series, pages 445–470. Springer Netherlands, 1982.

[9] Lior Rokach and Oded Maimon. Clustering methods. In Data mining and knowledge discovery handbook, pages 321–352. Springer, 2005.