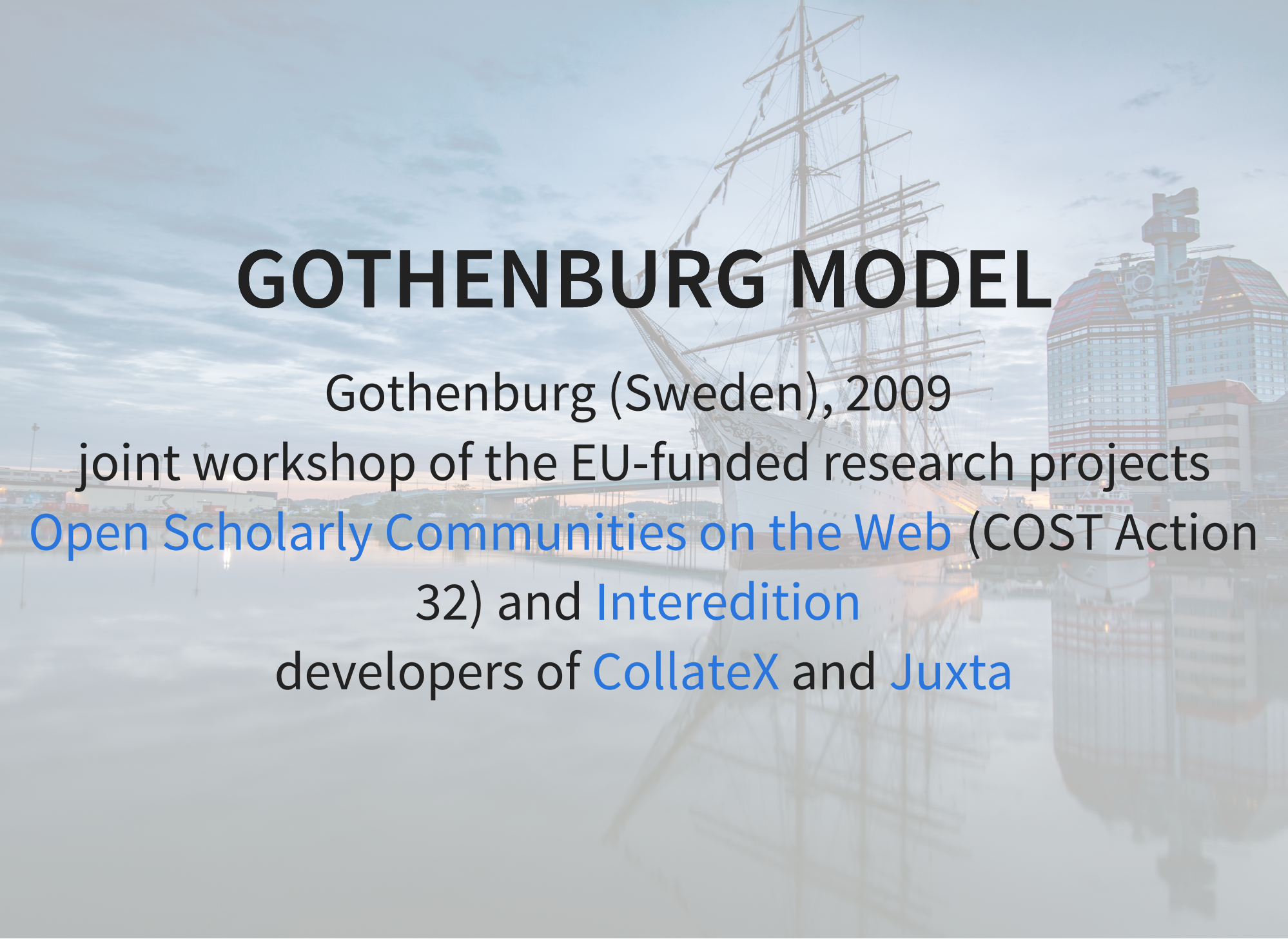
THE GOTHENBURG MODEL AND COLLATEX

Programme doctoral en études numériques.
Formation Textes et éditions numériques, 25 et 26 avril 2019, Université de Lausanne

Elena Spadini (Université de Lausanne)



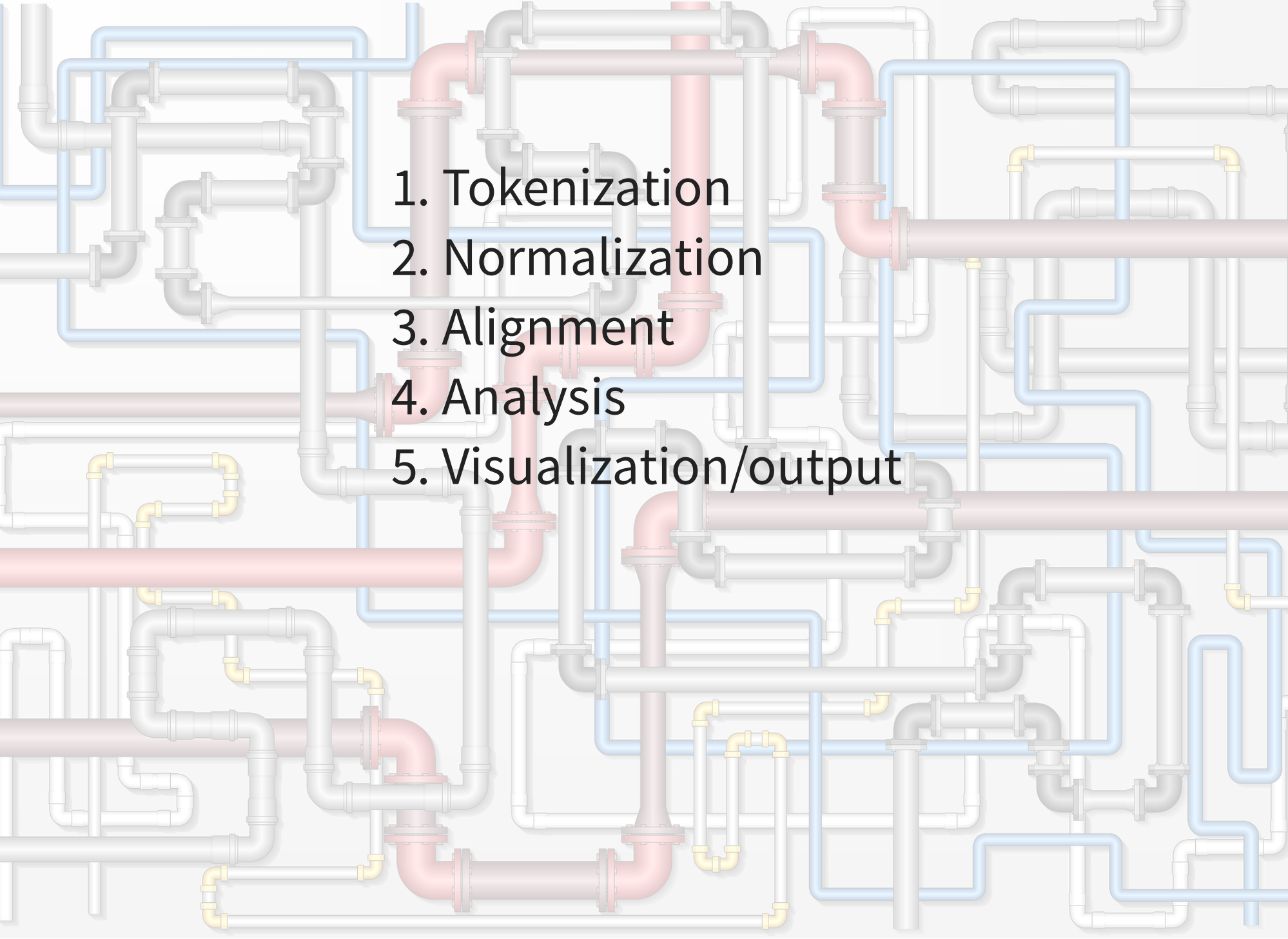
GOTHENBURG MODEL

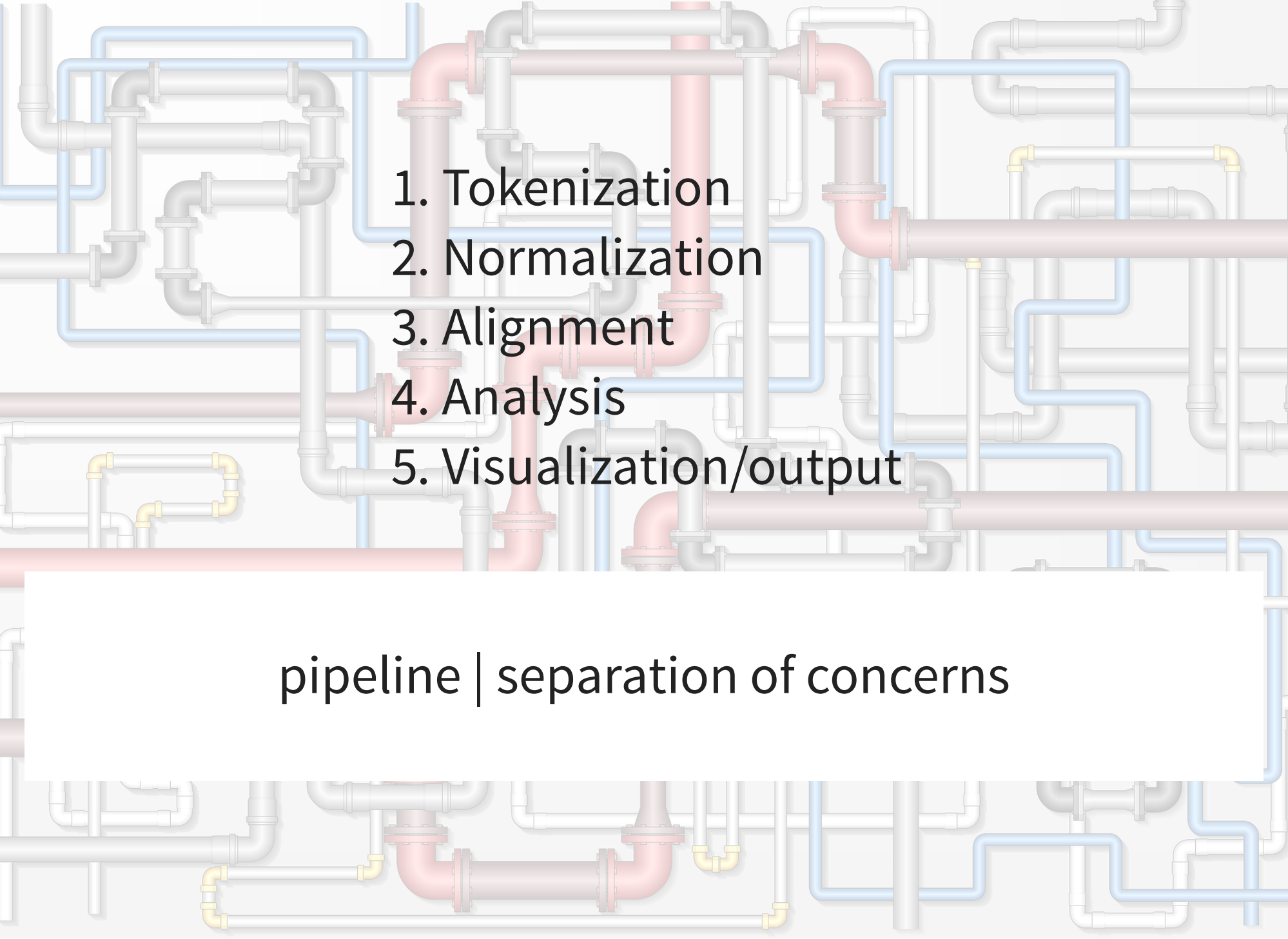
Gothenburg (Sweden), 2009

joint workshop of the EU-funded research projects Open Scholarly Communities on the Web (COST Action 32) and Interedition

developers of CollateX and Juxta

1. Tokenization
2. Normalization
3. Alignment
4. Analysis
5. Visualization/output


1. Tokenization
2. Normalization
3. Alignment
4. Analysis
5. Visualization/output


1. Tokenization
2. Normalization
3. Alignment
4. Analysis
5. Visualization/output
pipeline | separation of concerns

Step 0. Transcription or digitizazion+OCR/HWR

1. TOKENIZATION

division of a continuous text into units to be aligned, called **tokens**

1. TOKENIZATION

division of a continuous text into units to be aligned, called **tokens**

Exemple: *Peter's cat.*

1. TOKENIZATION

division of a continuous text into units to be aligned, called **tokens**

any level of granularity: syllables, words, lines, phrases, verses, paragraphs, text nodes ...

Exemple: *Peter's cat.*

TOKENIZATION ISSUES

- White spaces and punctuation are not enough
- multiple tokenizations are possible (vd. punctuation)
- markup
- language specific issues (contractions, superscription)

COLLATEX DEFAULT TOKENIZATION

- It divides the text into tokens at white space
- Punctuation is tokenized separately from alphanumeric characters

Example: Peter's cat.

Peter ' s cat .

2. NORMALIZATION

from handwriting or print into a digital representation (if not already in transcription)

prepare the texts for the collation in order to obtain "good" results (for example, upper/lower-case, formal vs substantive)

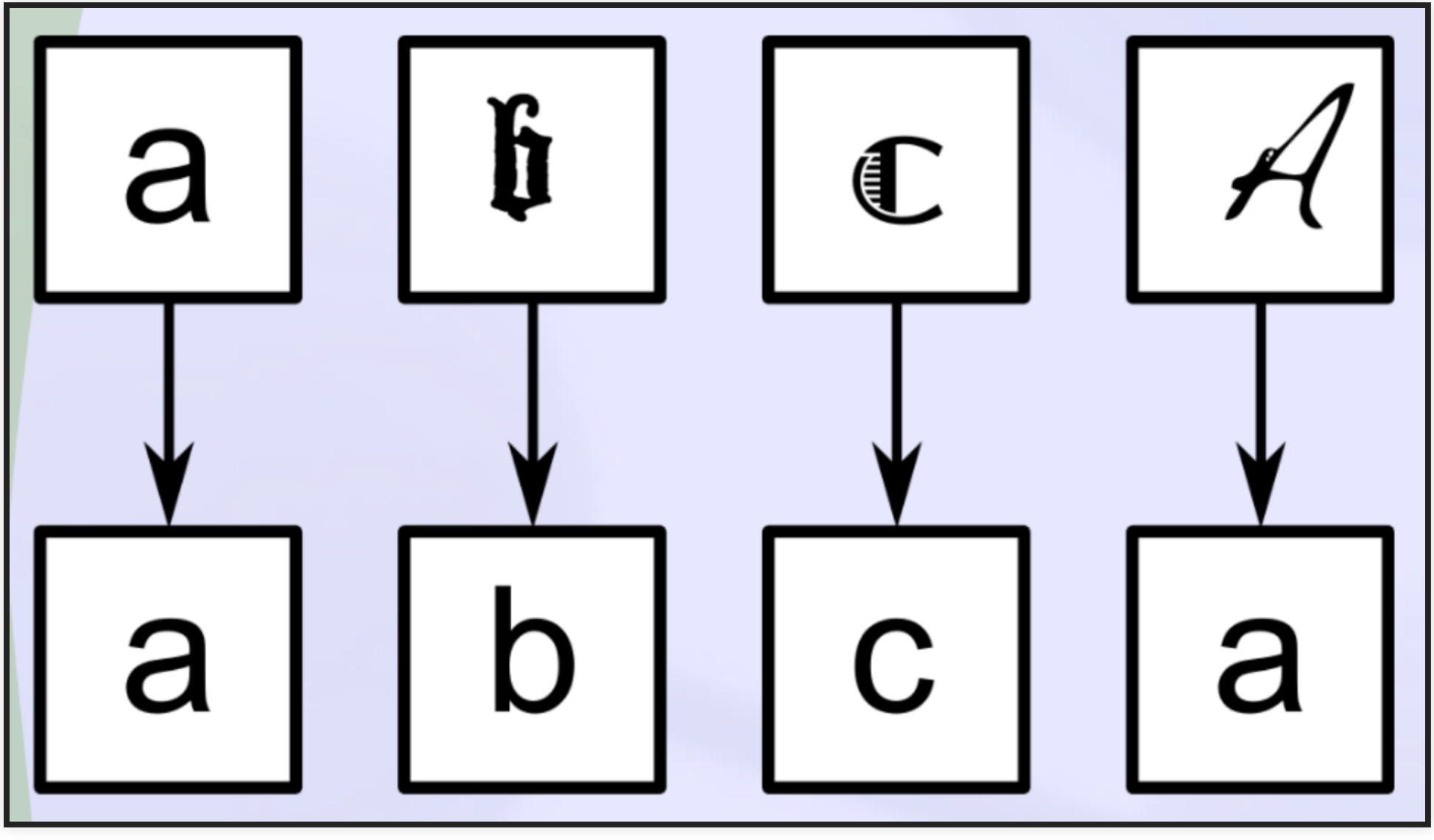
TYPES OF NORMALIZATION

- upper vs lower cases
- 3 vs three
- letterforms (graphemes)
- ... and much more

NORMALIZATION IN COLLATEX

By default, it removes trailing white space at the ends of tokens.

For each token, there is a **t** and a **n** property.



3. ALIGNMENT

Alignement is performed on **n** (shadow copy), but **t** is available all the time

ALIGNMENT COMPLEXITIES

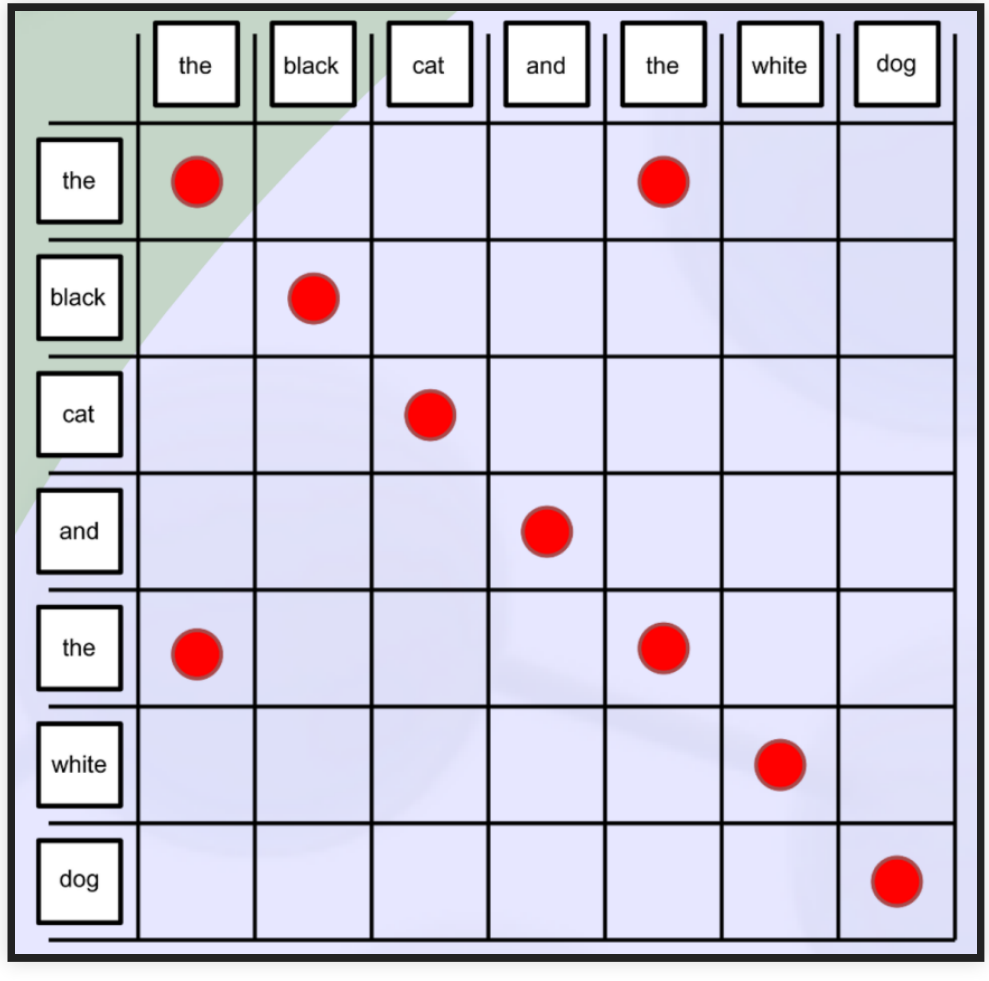
Computational complexities

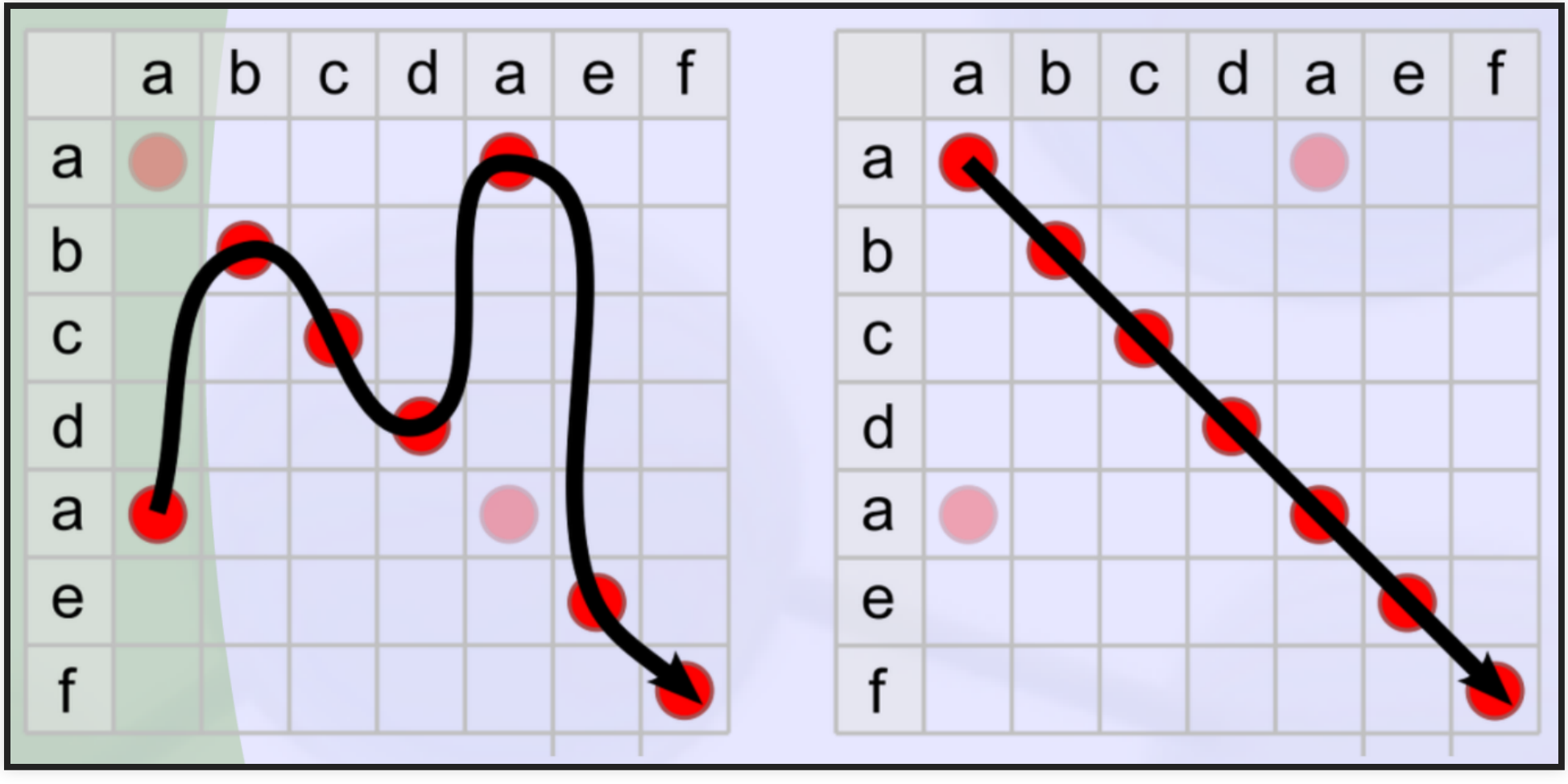
↓

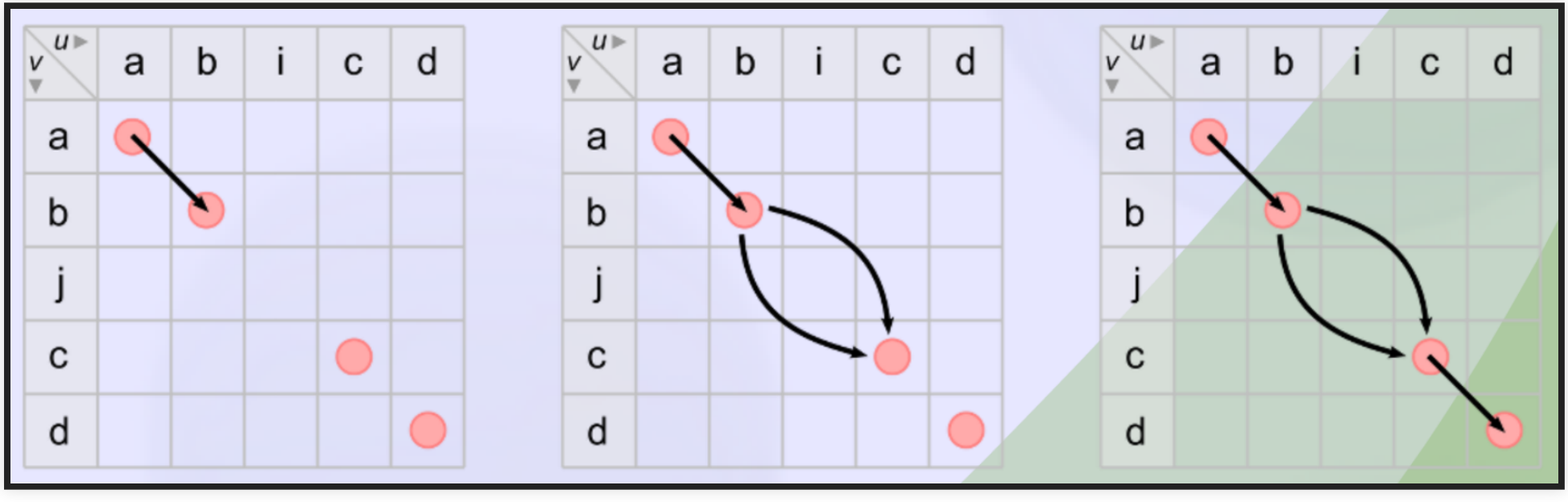
heuristics

ALIGNMENT COMPLEXITIES

Repetition




a b c d a e f
a
b
c
d
a
e
f
a b c d a e f
a
b
c
d
a
e
f


u
v
a
b
i
c
d
a
b
j
c
d

ALIGNMENT COMPLEXITIES

Order effects

Pairwise vs multiple (progressive) alignment

PAIRWISE ALIGNMENT

EXAMPLE

A: Dalla collina si vede una grande casa rossa.
B: Dal belvedere si vede una grande casa azzurra.
C: Dalla collina si vede una piccola casa rossa.
D: Dal belvedere si vedono tante case.

STEP 1: PAIRWISE ALIGNMENT USING 'A' AS THE BASE MANUSCRIPT.

A	Dalla	collina	si	vede	una	grande	casa	rossa
B	Dal	belvedere	si	vede	una	grande	casa	azzurra

Dalla] dal B
collina] belvedere B
rossa] azzurra B

A	Dalla	collina	si	vede	una	grande	casa	rossa
C	Dalla	collina	si	vede	una	piccola	casa	rossa

grande] piccola C

A	Dalla	collina	si	vede	una	grande	casa	rossa
D	Dal	belvedere	si	vedono	tante		case	

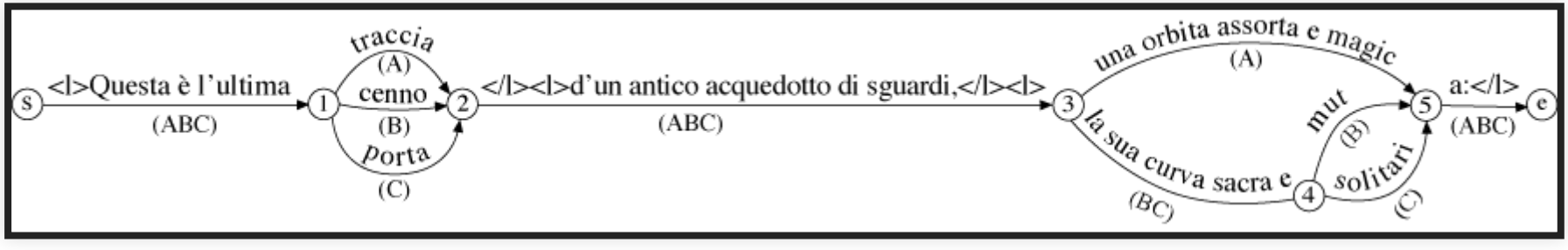
Dalla] dal D
collina] belvedere D
vede] vedono D
una] tante D
grande] om. D

STEP 2: THE RESULTS OF THE PAIRWISE ALIGNMENT ARE MERGED

Dalla] dal B, D
collina] belvedere B, D
vede] vedono D
una] tante D
grande] piccola C, om. D
casa] case D
rossa] azzurra B, om. D

ALIGNMENT COMPLEXITIES

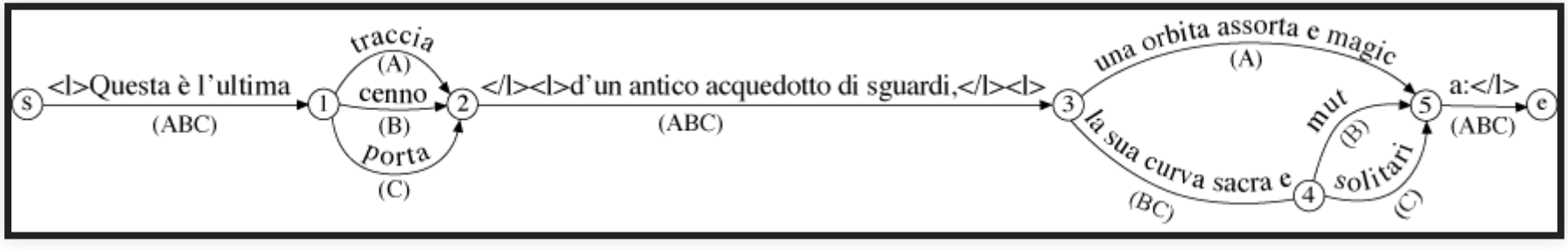
Order effects: pairwise vs multiple (progressive) alignment

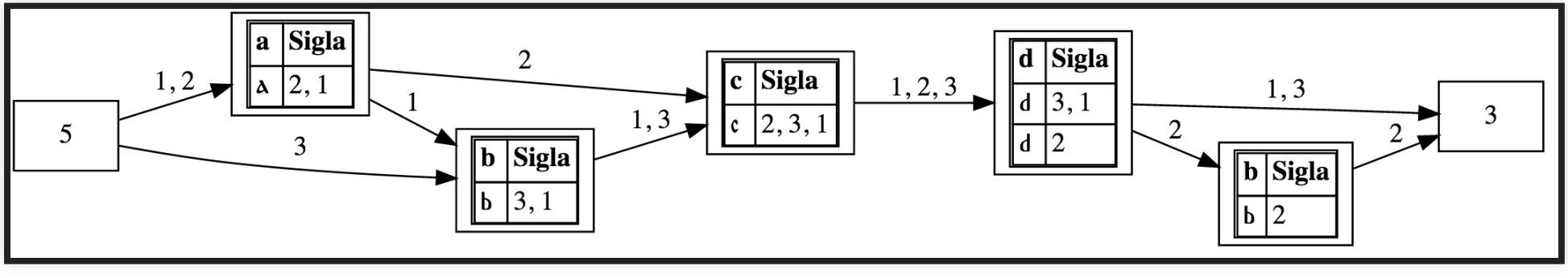


Progressively compare and merge the result of the comparison into the graph

ALIGNMENT COMPLEXITIES

Order effects: pairwise vs multiple (progressive) alignment





Progressively compare and merge the result of the comparison into the graph

ALIGNMENT COMPLEXITIES

exact vs near (fuzzy) matching

The big gray koala
The grey koala

4. ANALYSIS

computational analysis of the output of the alignment operation, for

- improving it (loop)
- enrich it

ANALYSIS IN COLLATEX

Near-matching

5. OUTPUT AND VISUALIZATION

- Catview
- Traviz
- Apparatus vs. Graph – an Interface as Scholarly Argument

COLLATEX OUTPUTS

output	segmentation	near_match	layout	indent
table	yes	yes	yes	no
html	yes	yes	yes	no
html2	yes	yes	no	no
svg_simple	yes	yes	no	no
svg	yes	yes	no	no
xml	yes	yes	no	no
tei	yes	yes	no	yes

CollateX in action

These slides reuse materials from

- http://collatex.net/
- https://github.com/DiXiT-eu/collatex-tutorial
- https://github.com/interedition/collatex

Elena Spadini, The Gothenburg Model and CollateX

Questions?