What is a text within the Digital Humanities, or some of them at least?

Manfred Thaller, Universität zu Köln

Digital Humanities 2012, July 20th 2012

Information I

Shannon

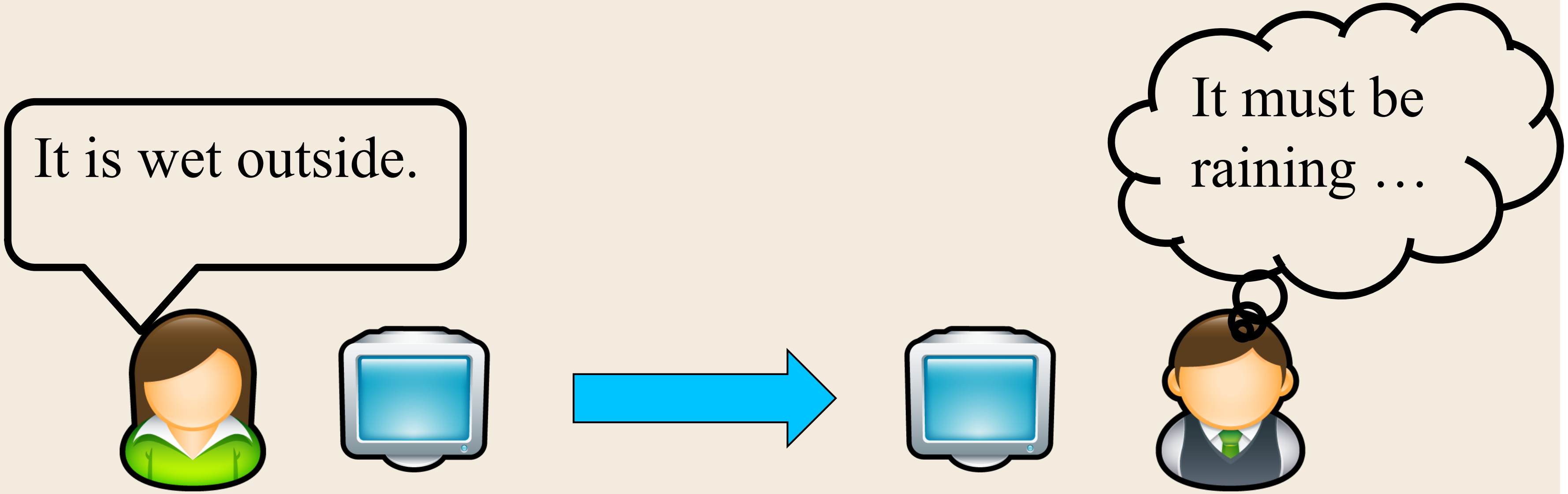
Claude Shannon: "A Mathematical Theory of Communication", Bell System Technical Journal, 1948.

Shannon

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

(Shannon, 1948, 379)

Shannon



It is wet outside.

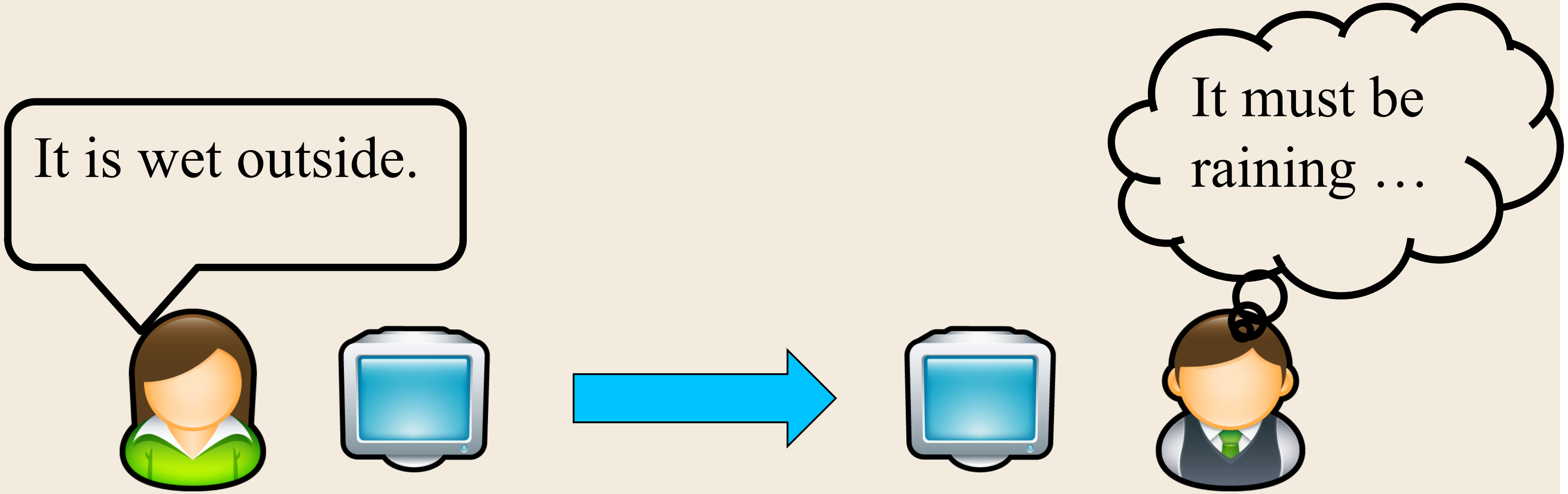
It must be raining …

Shannon

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.

(Shannon, 1948, 379)

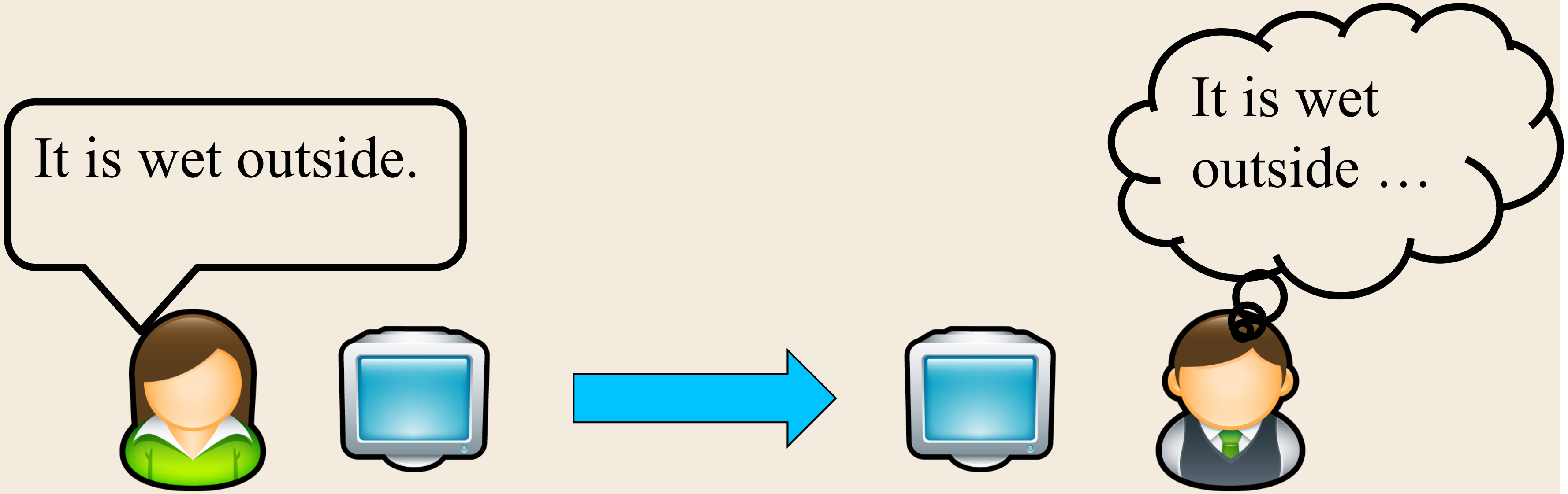
Shannon



It is wet outside.

It must be raining …

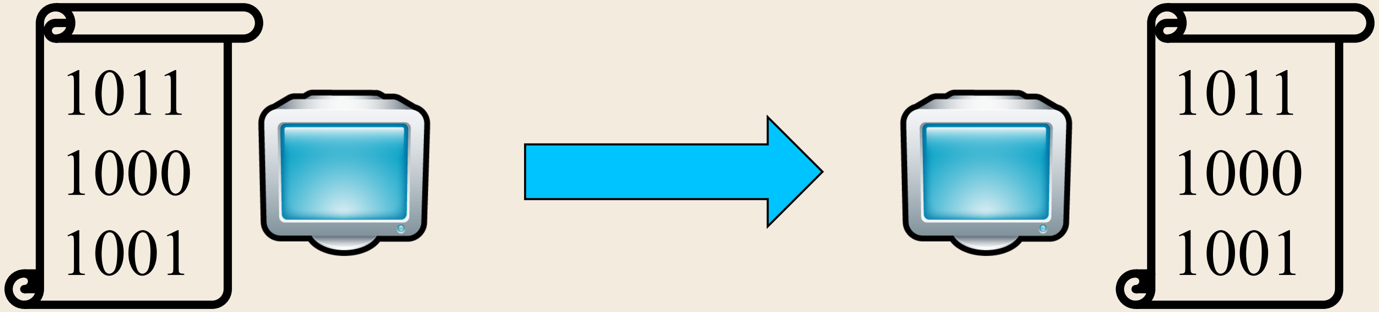
Shannon



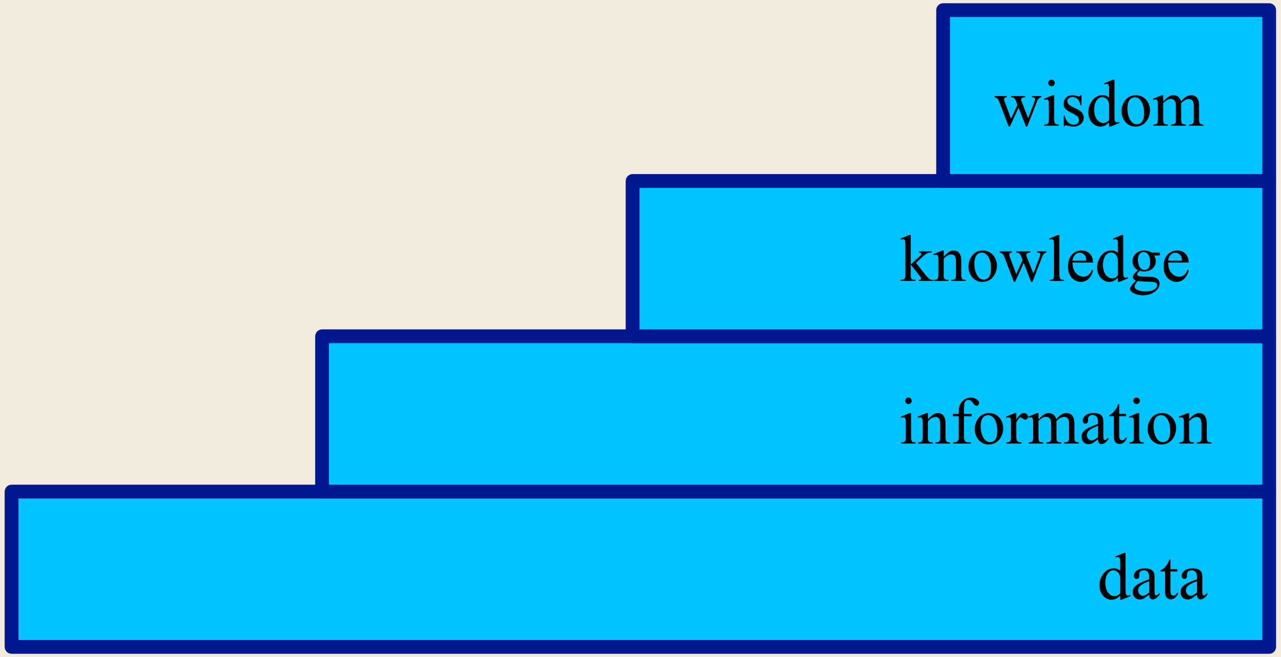
It is wet outside.

It is wet outside …

Shannon



„Ladder of Knowledge“



wisdom

knowledge

information

data

Information

Data

Data → Information

Data are stored. E.g.: *22°C*.

Information are data interpreted within a context:
"*In this lecture hall the temperature is 22°C*".

This context is fixed and identical for all recipients of information.

Information ➔ Knowledge

Knowledge is the result of a more complex process.

E.g. the decision, derived from the room temperature of 22 ° centigrade, to get out of your jacket; or not.

This context is different between recipients of information.

So …

Data	Information
22 ° C	22 ° C in lecture hall M
22	22 °
‘00000000 00010110’	22 [NOT ASCII { 0, 22 }]

Langefors

Langefors “Infological Equation”: original

$$\mathbf{I = i\,(D, S, t)}$$

I ::= *Information*

i() ::= *interpretative process*

D ::= *Data*

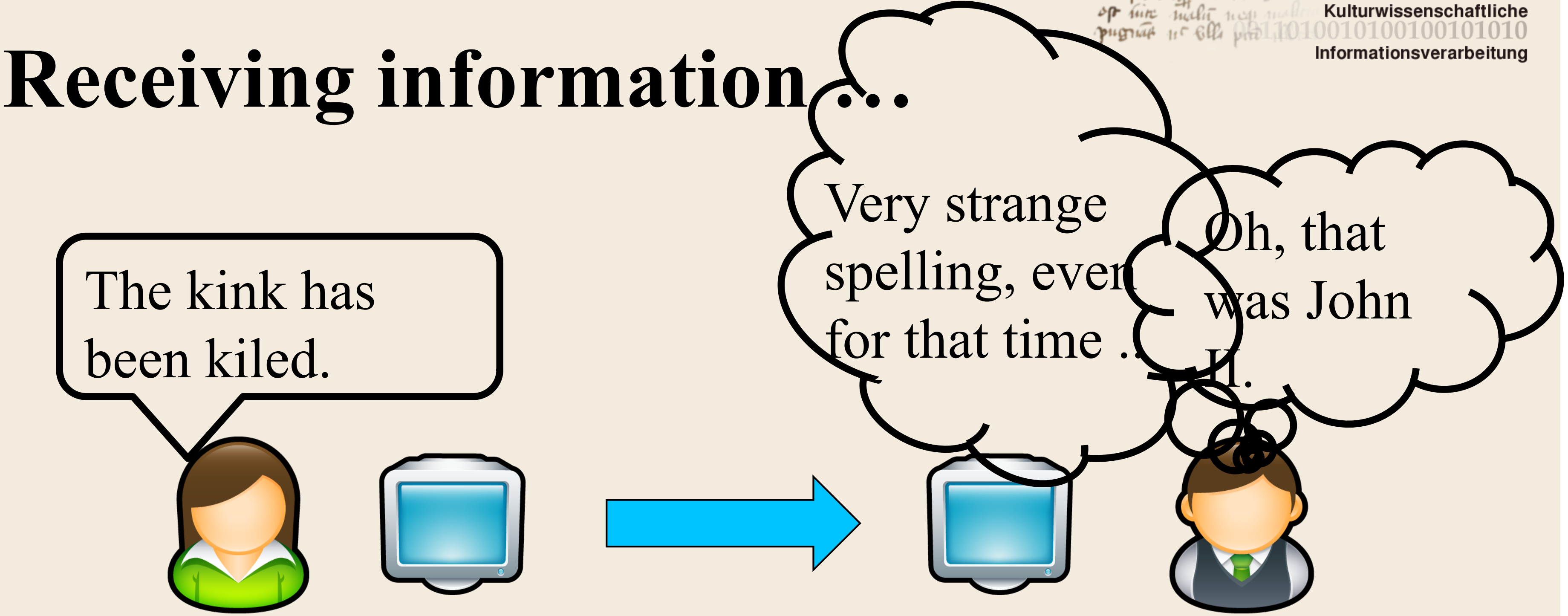
S ::= *Previous knowledge*

t ::= time

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

Information II

Receiving information …

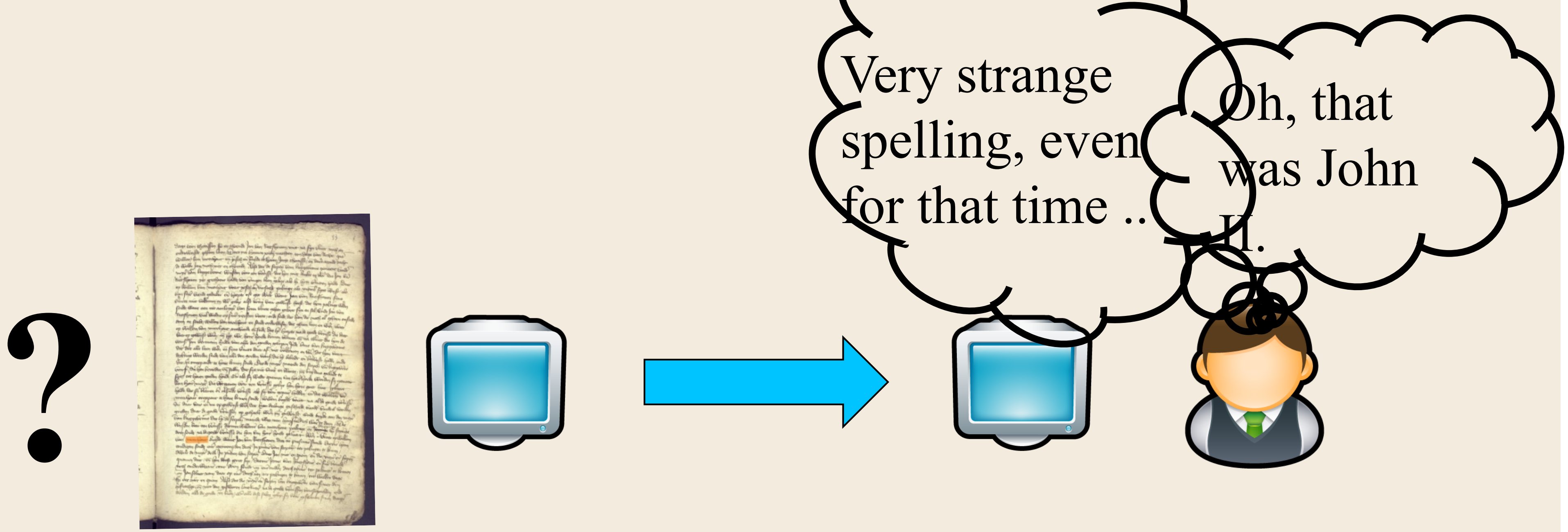


The kink has been kiled.

Very strange spelling, even for that time …

Oh, that was John II.

Receiving information …



?

Very strange spelling, even for that time ..

Oh, that was John II.

Notice: We can not consult the sender any more ….

Langefors

Langefors “Infological Equation”: original

$$\mathbf{I = i\ (D, S, t)}$$

I ::= *Information*

i() ::= *interpretative process*

D ::= *Data*

S ::= *Previous knowledge*

t ::= time

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

Langefors

Langefors "Infological Equation": generalization 1

$$\mathbf{I_2 = i\,(I_1, S_2, t)}$$

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

I ::= *Information*

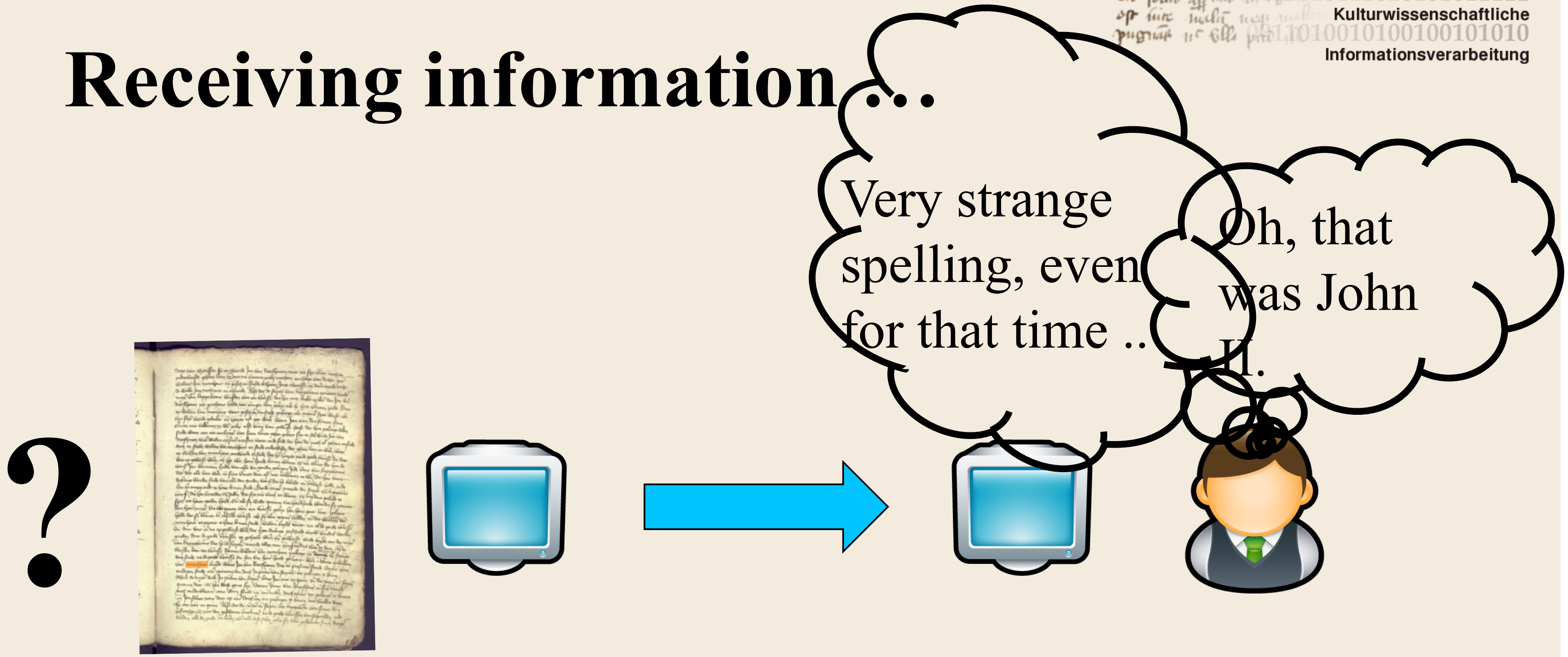
i() ::= *interpretative process*

D ::= *Data*

S ::= *Previous knowledge*

t ::= time

Receiving information …

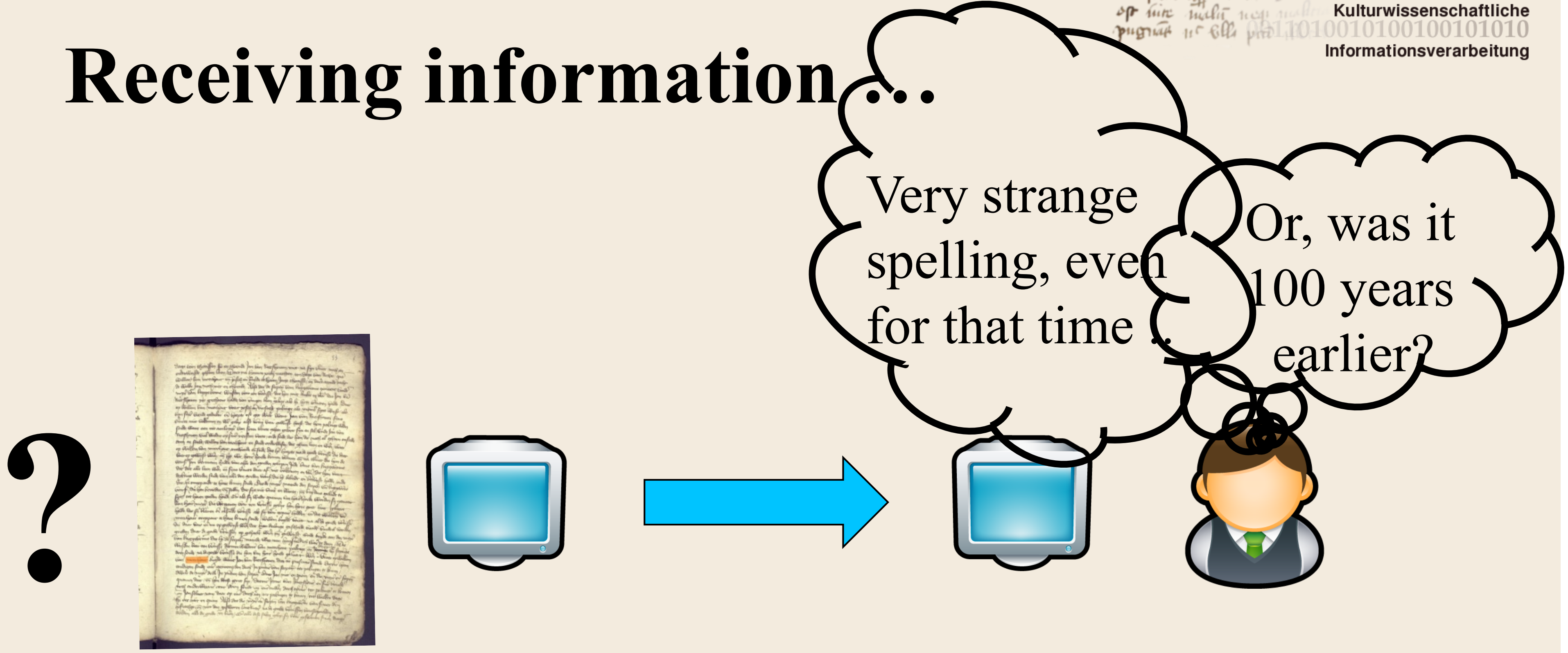


Very strange spelling, even for that time ..

Oh, that was John II.

Notice: We can not consult the sender any more ….

Receiving information …



Very strange spelling, even for that time …

Or, was it 100 years earlier?

Notice: We can not consult the sender any more ….

Langefors

Langefors “Infological Equation”: generalization 2

$$\mathbf{I_x = i\,(I_{x-1}, S_x, t)}$$

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

I ::= *Information*

i() ::= *interpretative process*

D ::= *Data*

S ::= *Previous knowledge*

t ::= time

Langefors

Langefors “Infological Equation”: generalization 3

$$S_x = s\,(I_{x-1}, t)$$

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

I ::= *Information*

i() ::= *interpretative process*

D ::= *Data*

S ::= *Previous knowledge, s = knowledge generating process*

t ::= time

Langefors

Langefors “Infological Equation”: generalization 4

$$\mathbf{I_x = i\ (I_{x-\alpha}, S_{x-\beta}, t)}$$

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

I ::= *Information*

i() ::= *interpretative process*

D ::= *Data*

S ::= *Previous knowledge*

t ::= time

Langefors

Langefors “Infological Equation”: generalization 5

$$\mathbf{I_x = i\,(I_{x-\alpha}, s(I_{x-\beta}, t), t)}$$

Börje Langefors, *Essays on Infology,* Studentliteratur: Lund, 1995

I ::= *Information*

i() ::= *interpretative process*

D ::= *Data*

S ::= *Previous knowledge*

t ::= time

Remember …

Data	Information
22 ° C	22 ° C in lecture hall M
22	22 °
‘00000000 00010110’	22 [NOT ASCII { 0, 22 }]

Changeable datatypes

```
int myVariable;
char myVariable;
temperature myVariable;

obj myVariable;
myVariable.useAsInt();
myVariable.useAsChar();
myVariable.addInterpretation(temperature,Centigrade);
```

Langefors

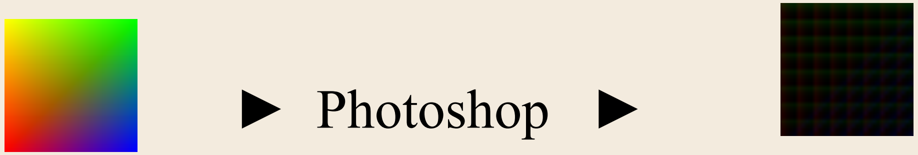
Notes:

(1) If this is so, the assumption of Comp. Sci., that information is represented by structures on which algorithms operate, can be replaced by a more general understanding, according to which information is a state of a set of perpetually active algorithms.

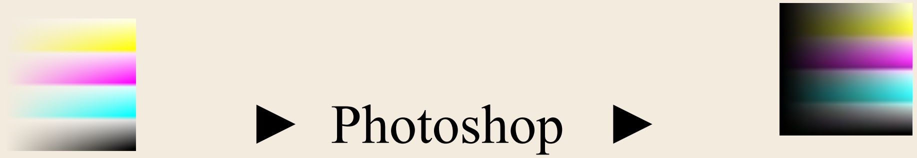
(2) Has that any practical meaning?

A practical interlude

Planets: the problem

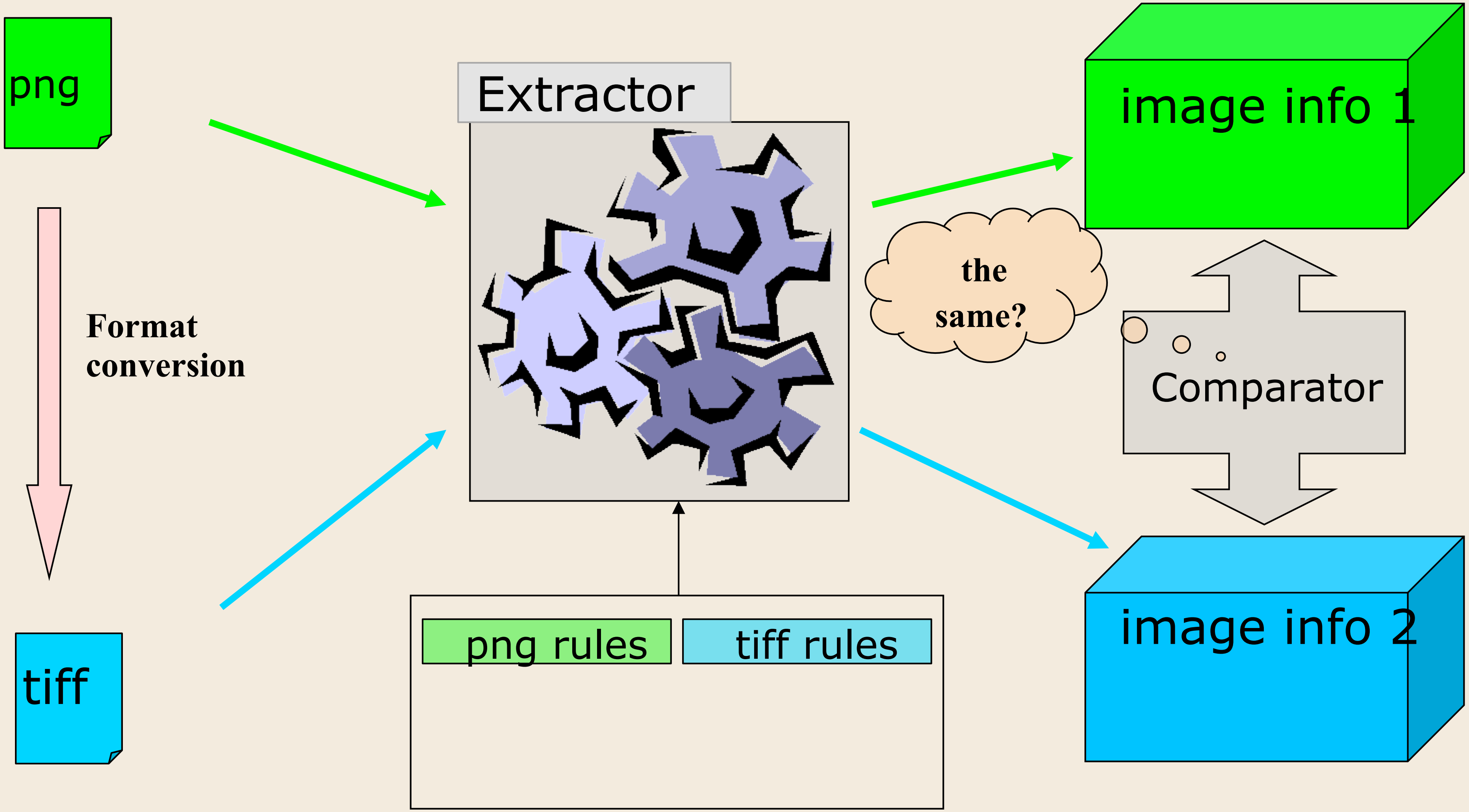


► Photoshop ►

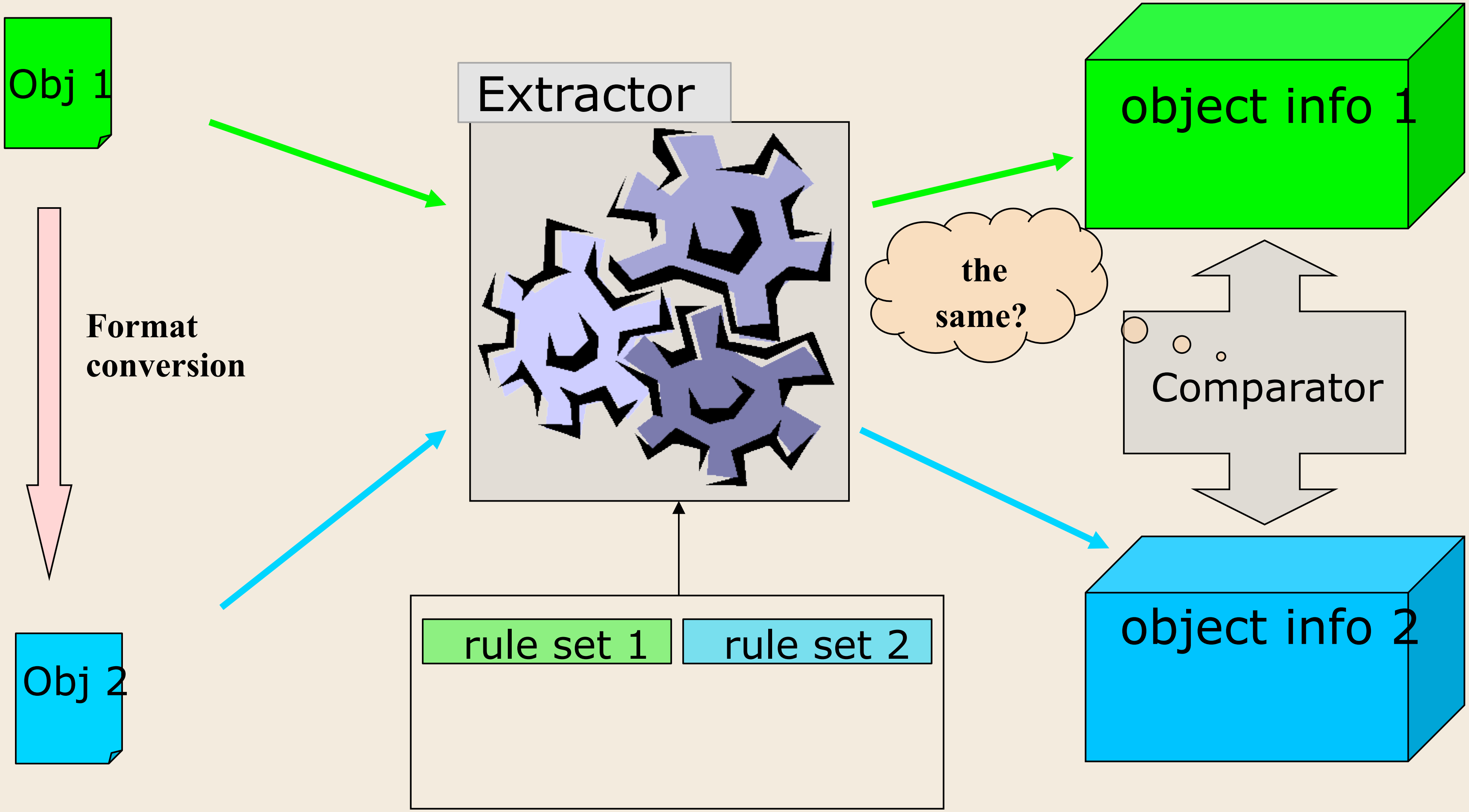


► Photoshop ►

Planets: the vision 1



Planets: the vision 2



Obj 1

Format
conversion

Obj 2

Extractor

rule set 1

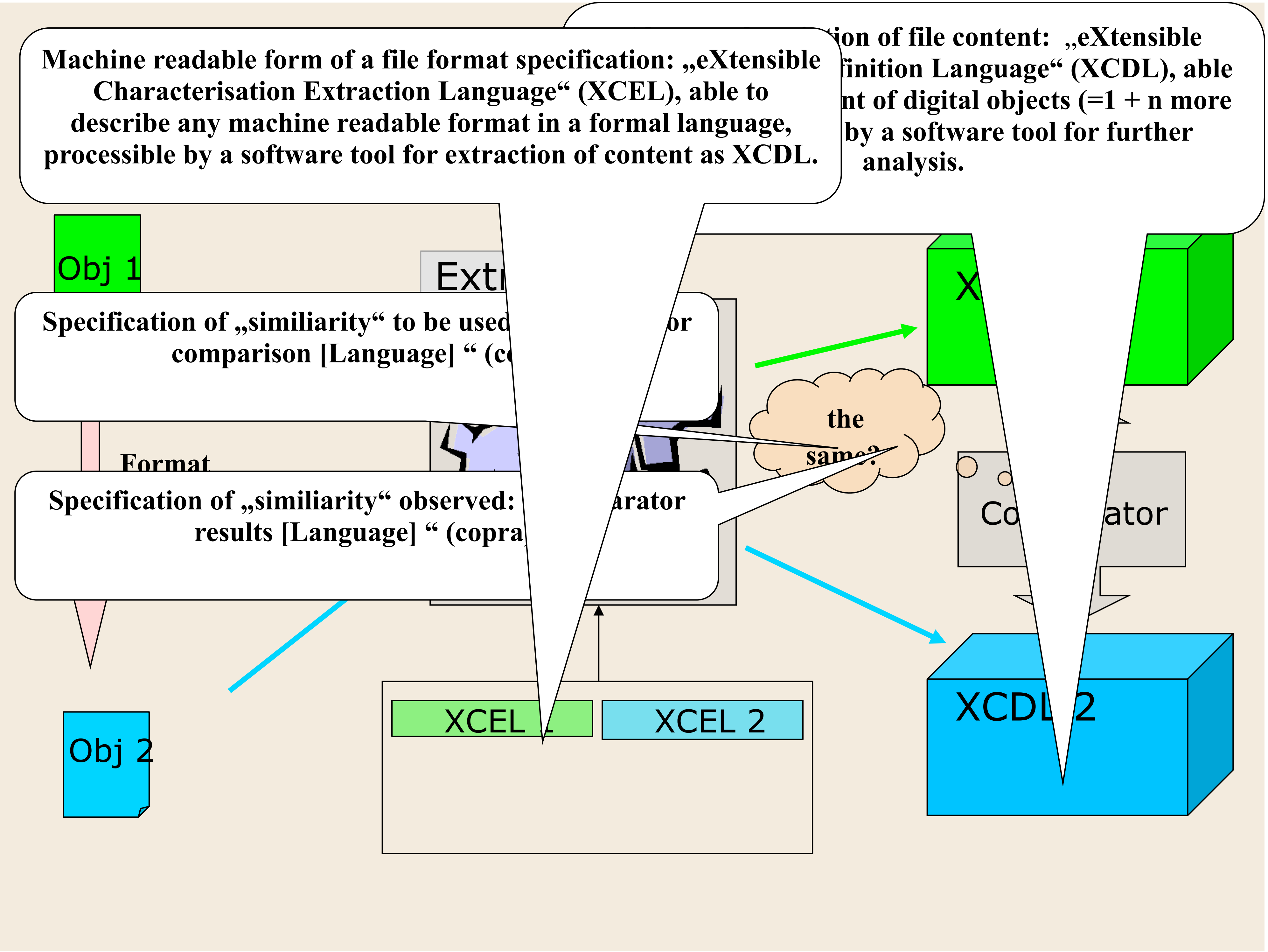
rule set 2

object info 1

the
same?

Comparator

object info 2



ion of file content: „eXtensible finition Language“ (XCDL), able nt of digital objects (=1 + n more by a software tool for further analysis.

Machine readable form of a file format specification: „eXtensible Characterisation Extraction Language“ (XCEL), able to describe any machine readable format in a formal language, processible by a software tool for extraction of content as XCDL.

Specification of „similiarity“ to be used or comparison [Language] “ (c

Specification of „similiarity“ observed: arator results [Language] “ (copra

Text in XCDL

This is a text

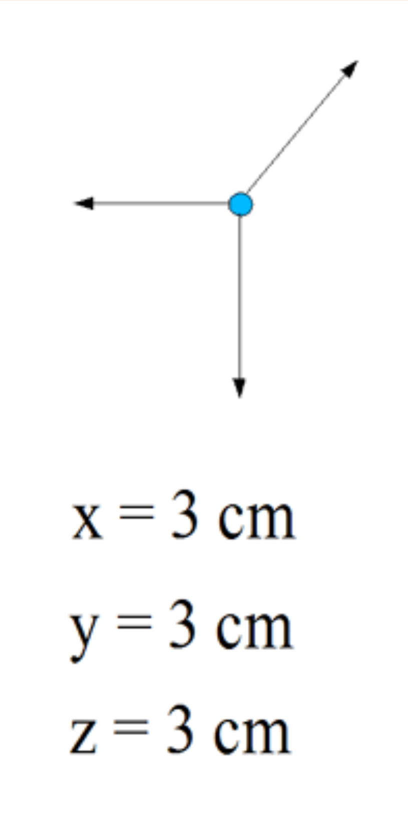
```
<refData id="1">54 68 69 73 20 69 73 20 61 20 74 65 78 74</refData>
…
<property>
<name>fontsize</name>
<rawVal>
<val>48</val>
<type>unsignedInt8</type>
</rawVal>
<dataRef> <!-- property refers to discrete part of reference data-->
<ref id="1" start="0" end="3"/>
<ref id="1" start=“10" end="12"/>
</dataRef>
</property>
```

Image in XCDL

```
<refData id="1">7A 11 9B 77 34 89 72 11 29 F4 DA 9C B2 23 56 93 86 83 82 65 …</refData>
…
<property>
<name>title</name>
<rawVal>
<val>Ebstorf Mappa Mundi</val>
<type>ASCII</type>
</rawVal>
<dataRef>
<ref id="1" start="0" end="13455"/>
</dataRef>
</property>
```

Generalizing the practical solution

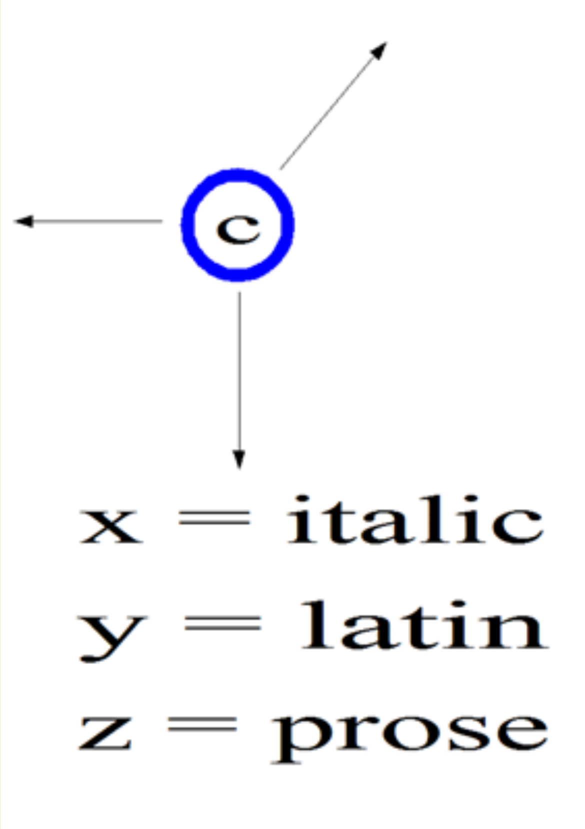
Dimensions: geometry



Allows to make statements about the proximity of two objects on the "y" axis.

Irrespective of the "shape" of the object.

Dimensions: textual / conceptual



x = italic
y = latin
z = prose

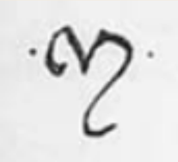
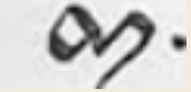
Allows to make statements about the proximity of two objects on the "y" axis.

Irrespective of the “object" that is at the abstract position.

Dimensions: metrics

Dimensions are by definition orthogonal.

Dimensions can have any sort of metric:

- Rational: { - ∞ … + ∞ }
- Integer range: { 0 … 100 }
- Nominal: { medieval, early modern, modern }
- Image: {  ,  }
- …

Four texts …

(1) <person><surname><bold>Biggin</bold></surname></person>

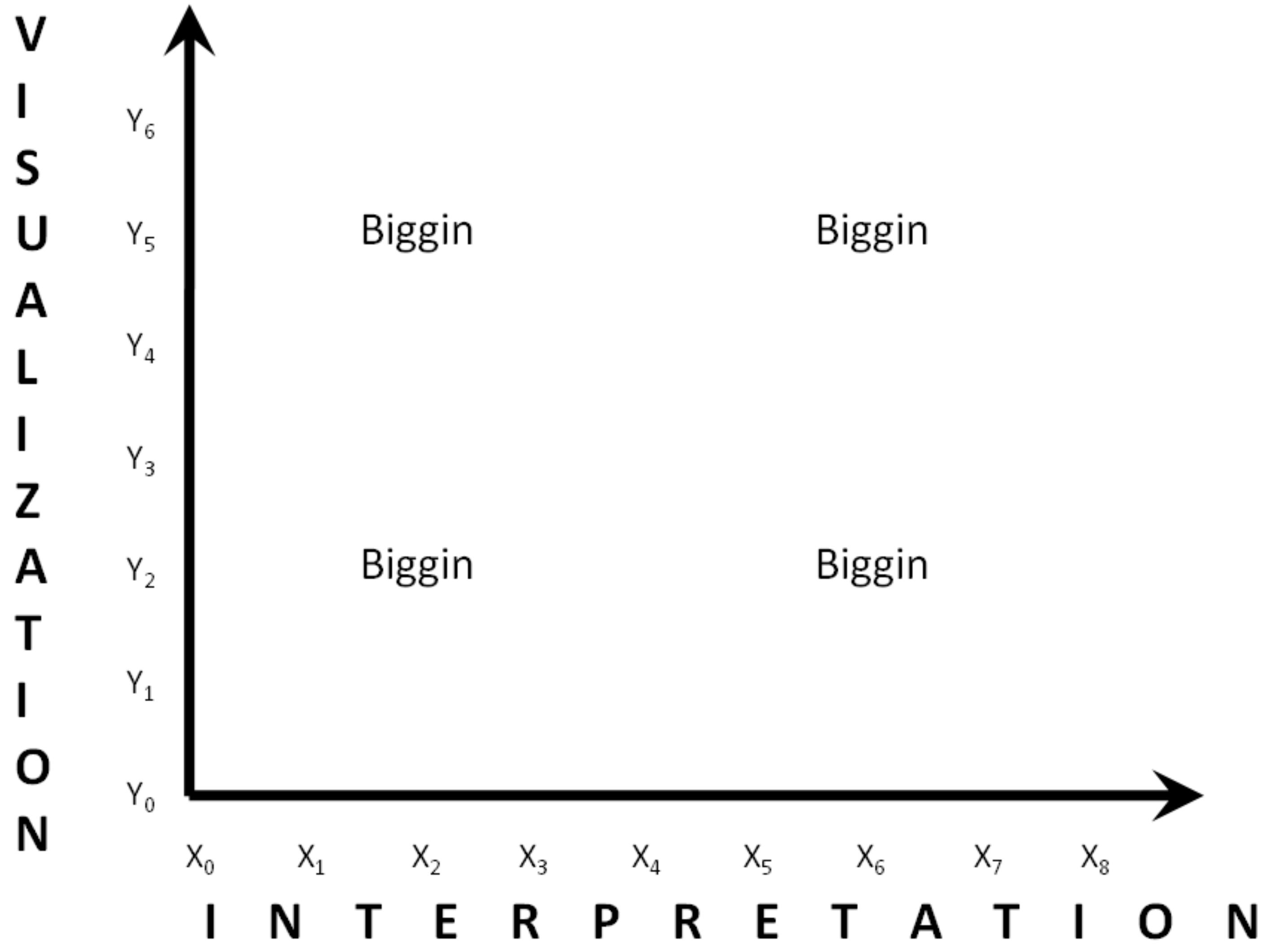
(2) <person><surname><italics>Biggin</italics></surname></person>

(3) <airfield><name><bold>Biggin</bold></name></airfield>

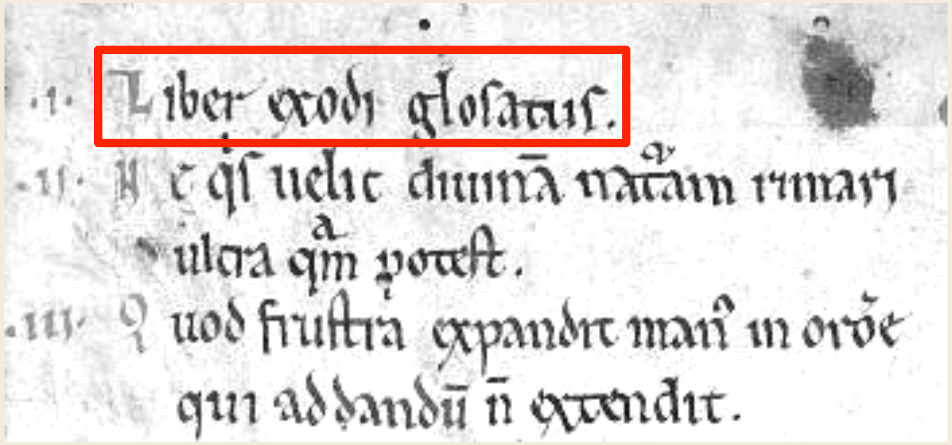
(4) <airfield><name><italics>Biggin</italics></name></airfield>

Which of the chunks are more similar to each other: (1) and (2) or (1) and (3)?

… in a coordinate space.



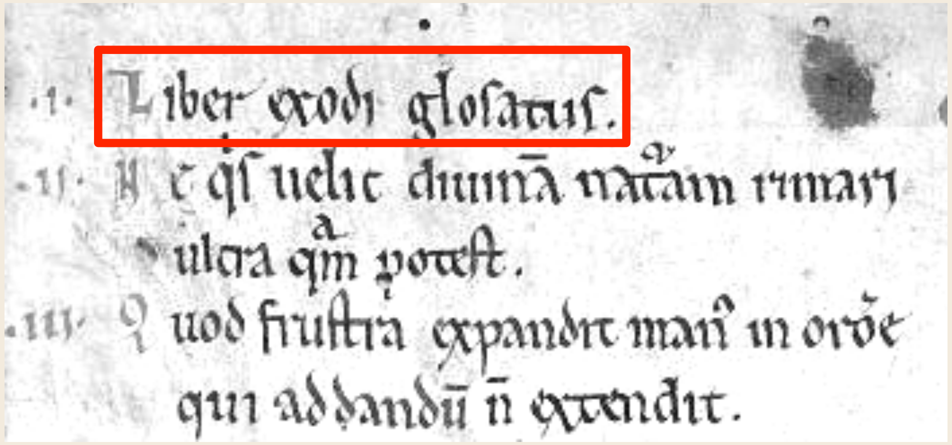
An image in a textual coordinate space



·i· Liber exodi glosatus.
·ii· Hec q̄s uelit diuinā naťam rimari
ulcra q̄m potest.
·iii· Quod frustra expandit manū in orōe
qui addandū ñ extendit.

Liber exodi glosatus

An text in an image coordinate space

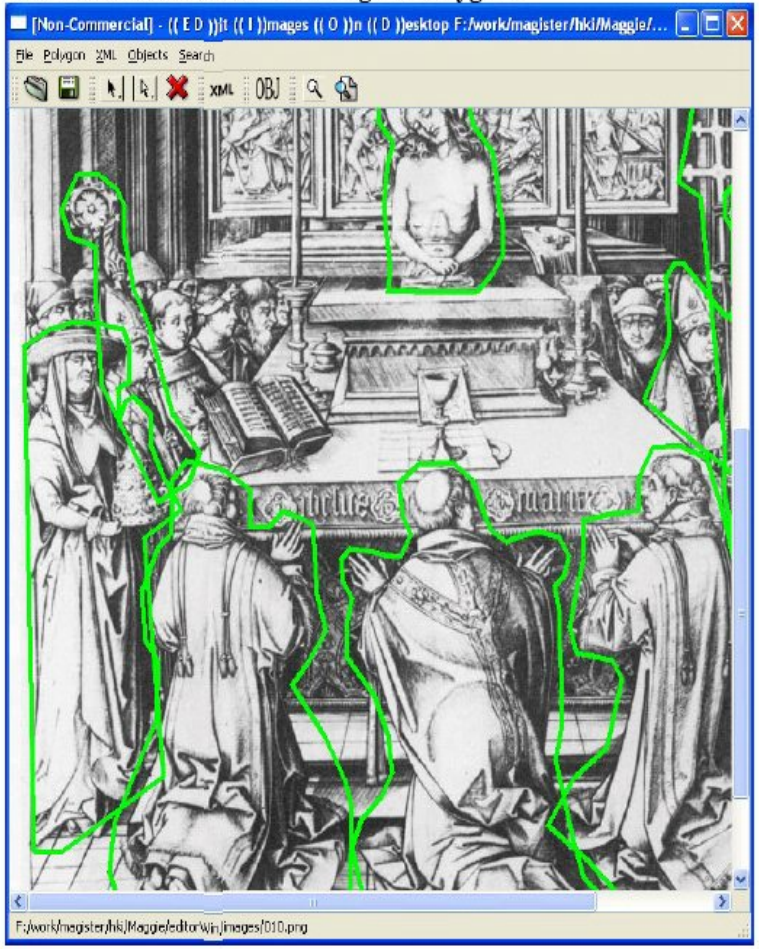


·i· Liber exodi glosatus.
·ii· N e q̄s uelit diuinā naturā rimari
ultra q̄m potest.
·iii· Q uod frustra expandit man̄ in orđe
qui addandū n̄ extendit.

Liber exodi glosatus

An image in a semantic coordinate space

B.1 Bildbetrachter mit Darstellung der Polygone


[Non-Commercial] - ((E D))it ((I))mages ((O))n ((D))esktop F:/work/magister/hki/Maggie/...
File
Polygon
XML
Objects
Search
XML
OBJ

F:/work/magister/hki/Maggie/editor/win/images/010.png

Bishop

Cardinal

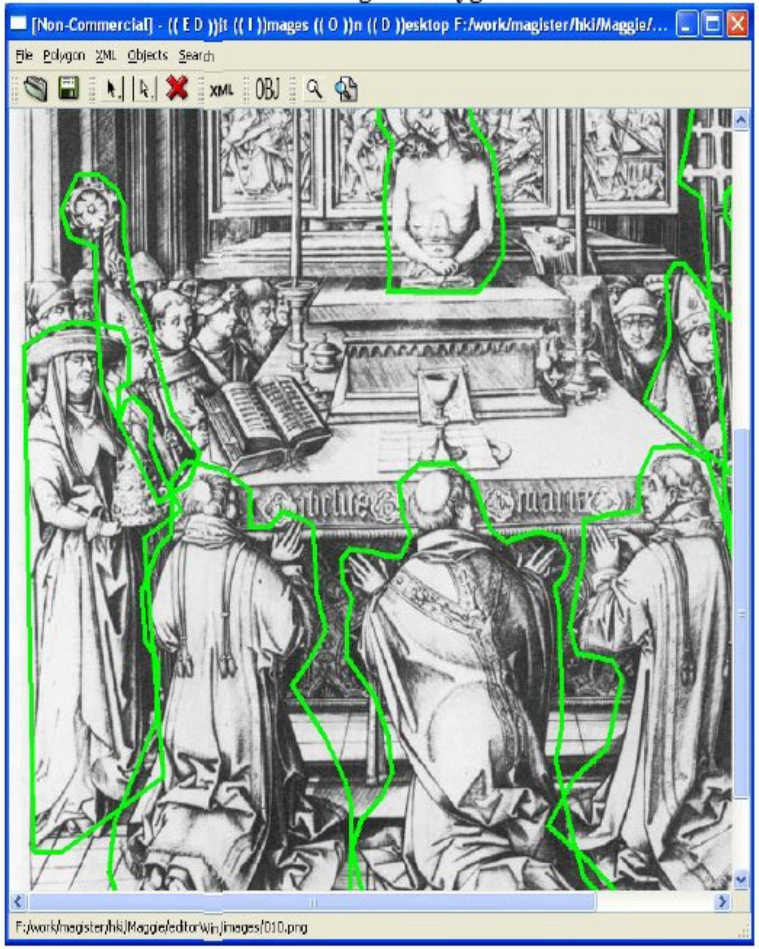
Monk

Priest

Monk

Semantics in an image coordinate space

B.1 Bildbetrachter mit Darstellung der Polygone


[Non-Commercial] - ((E D))it ((I))mages ((O))n ((D))esktop F:/work/magister/hki/Maggie/...
File Polygon XML Objects Search
XML
OBJ
F:/work/magister/hki/Maggie/editorWin/images/010.png

Bishop

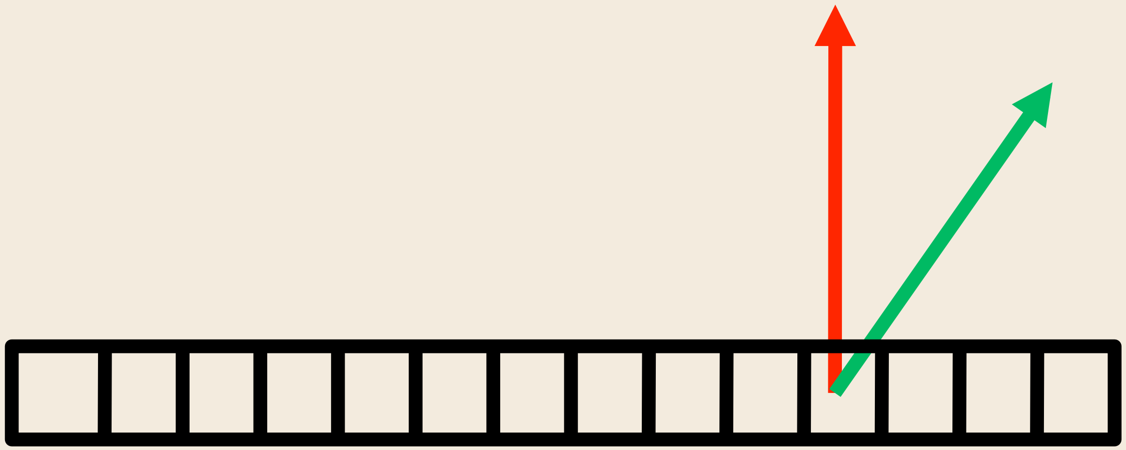
Cardinal

Monk

Priest

Monk

Generalization 1

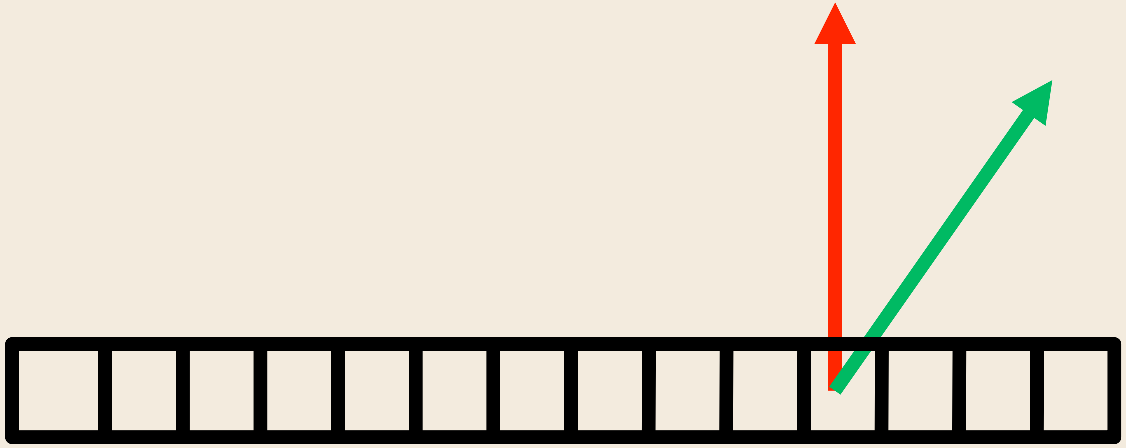


Biggin

Visualization {bold, italic}

Interpretation {surname, topographic name}

Generalization 2

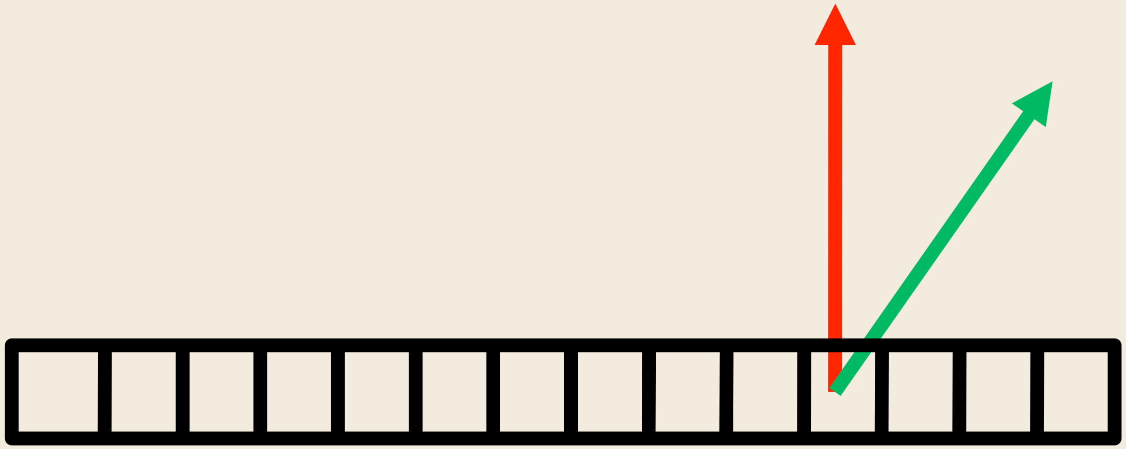


Series of atomic content tokens

Conceptual dimension 1

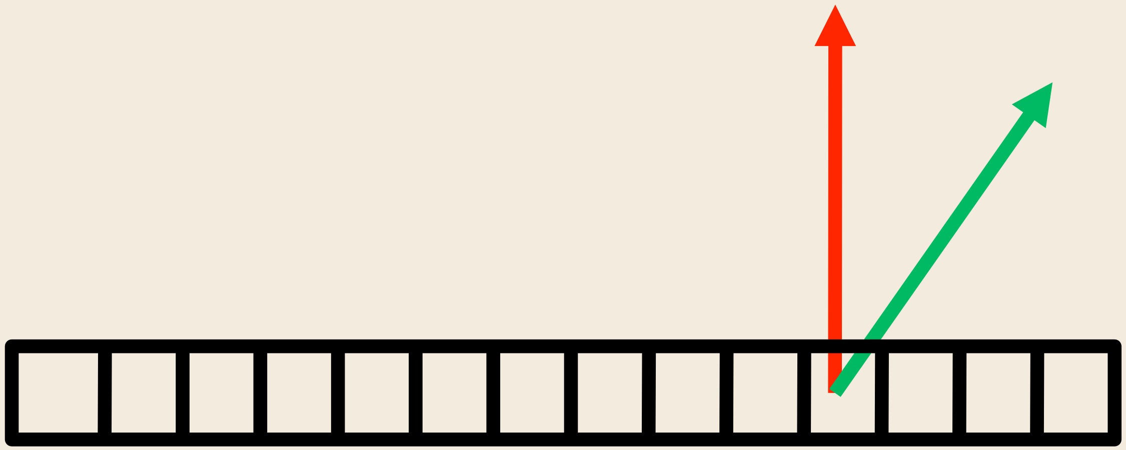
Conceptual dimension 2

Generalization 3



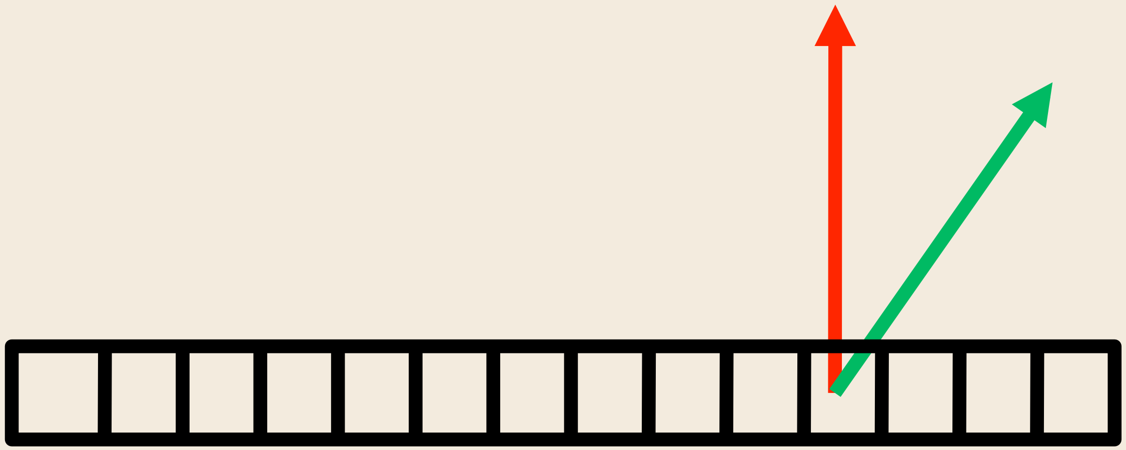
$\{ T, C_1, C_2\}$

Generalization 4



$\{ T, \{ C_1, C_2, \ldots, C_n \} \}$

Generalization 5

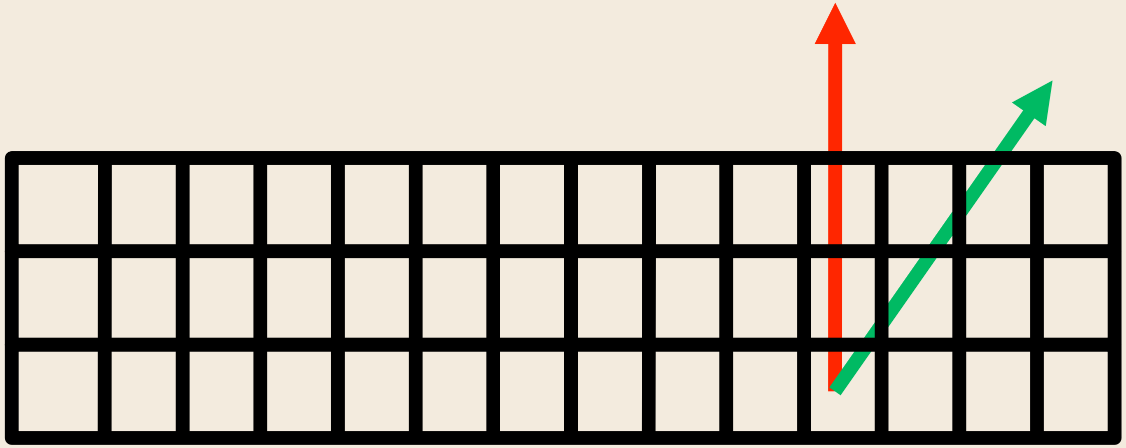


$\{ T, C_n \}$

(1) Texts are sequences of content carrying atomic tokens.

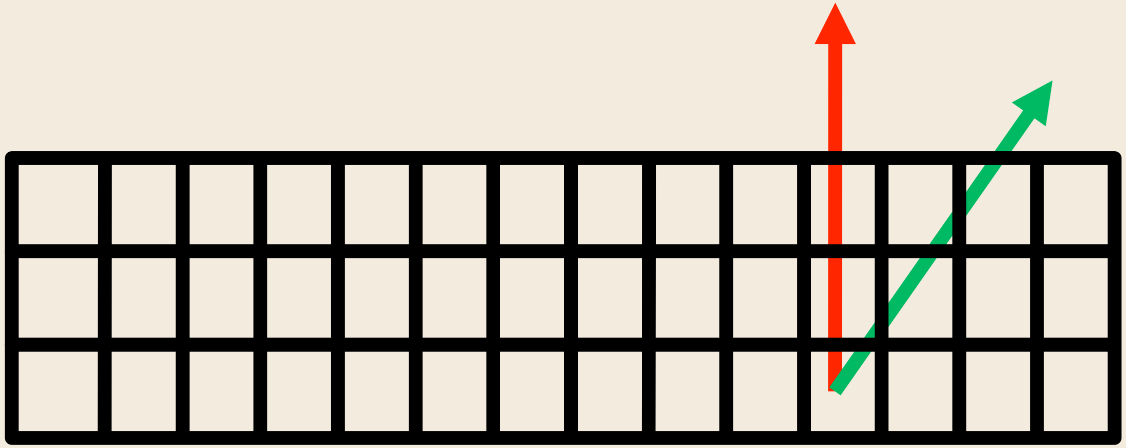
(2) Each of these tokens has a position in an *n*-dimensional conceptual universe.

Generalization 6



$\{ X, Y, C_n \}$

Generalization 7



$\{ T_1, T_2, C_n \}$

(1) Images are planes of content carrying atomic tokens.

(2) Each of these tokens has a position in an *n*-dimensional conceptual universe.

Generalization 8

$$\mathbf{I \ ::= \{ \{ T_1, T_2, \ldots\ T_m\}, C_n \ \}}$$

(1) Information objects are *m*-dimensional arrangements of content carrying atomic tokens.

(2) Each of these tokens has a position in an *n*-dimensional conceptual universe.

Generalization 9

$$\mathbf{I ::= \{T_m,\ C_n\ \}}$$

(1) Information objects are *m*-dimensional arrangements of content carrying atomic tokens.

(2) Each of these tokens has a position in an *n*-dimensional conceptual universe.

(3) All of this, of course, is recursive …

Another practical interlude

Virtual Research Environments

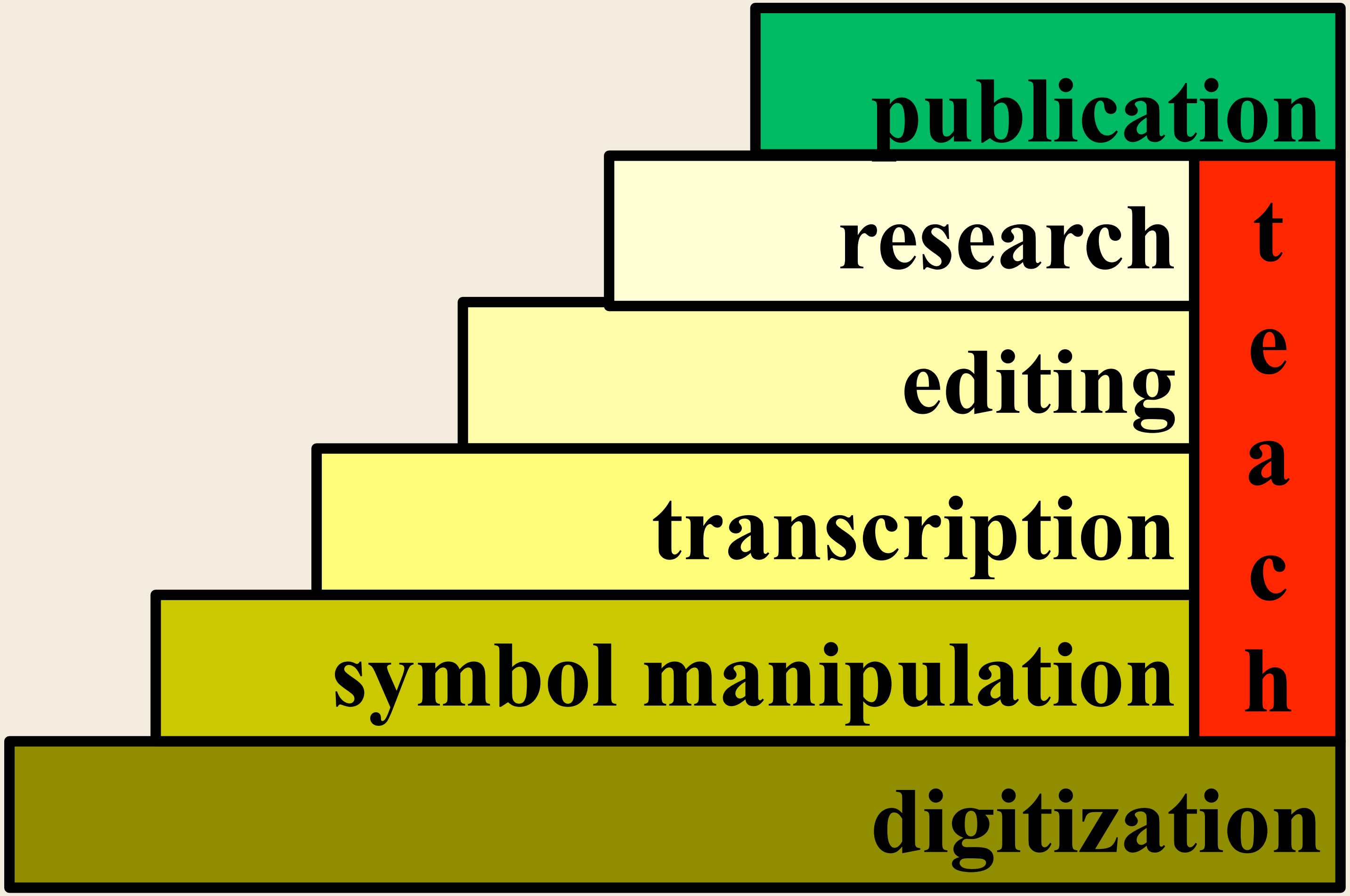
http://www.monasterium.net/

➔

Virtuelles deutsches Urkundennetz

(Virtual network of German charters)

A model of historical research



publication

research

editing

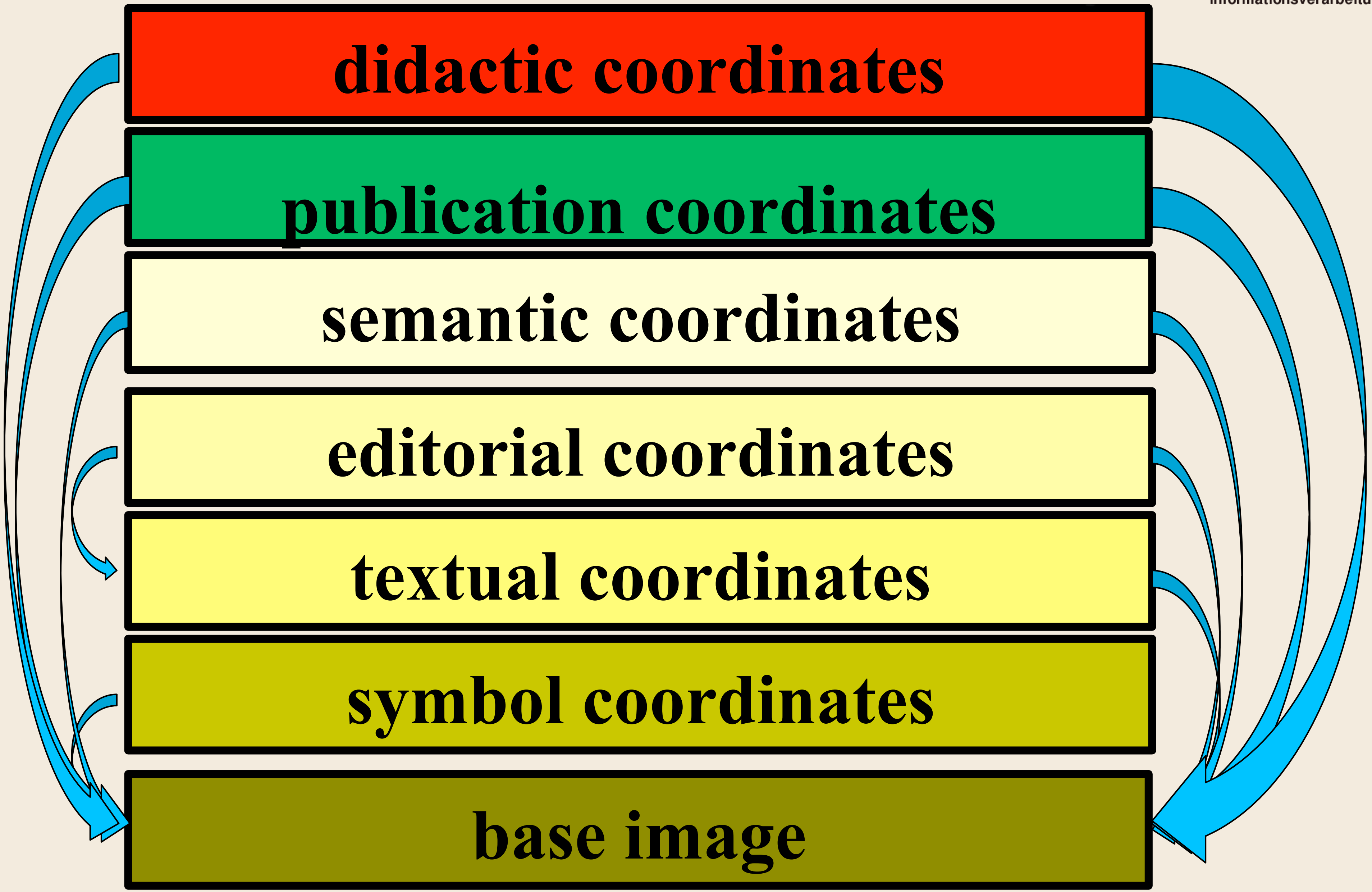
transcription

symbol manipulation

digitization

t e a c h

A model of historical research



didactic coordinates

publication coordinates

semantic coordinates

editorial coordinates

textual coordinates

symbol coordinates

base image

Conclusion

Summary

(1) All texts, for which we cannot consult the producer, should be understood as a sequence of tokens, where we should keep the representation of the tokens and the representation of our interpretation thereof completely separate.

(2) Such representations can be grounded in information theory.

(3) These representations are useful as blueprints for software on highly divergent levels of abstraction.

Thank you!

manfred.thaller@uni-koeln.de