DBAI
Institute of Information Systems
Database and Artificial Intelligence Group

Functional Semantic Analysis of Web Pages on the Visual Layer

Presentation of the Master‘s Thesis

by Bernhard Pollak - 9326613

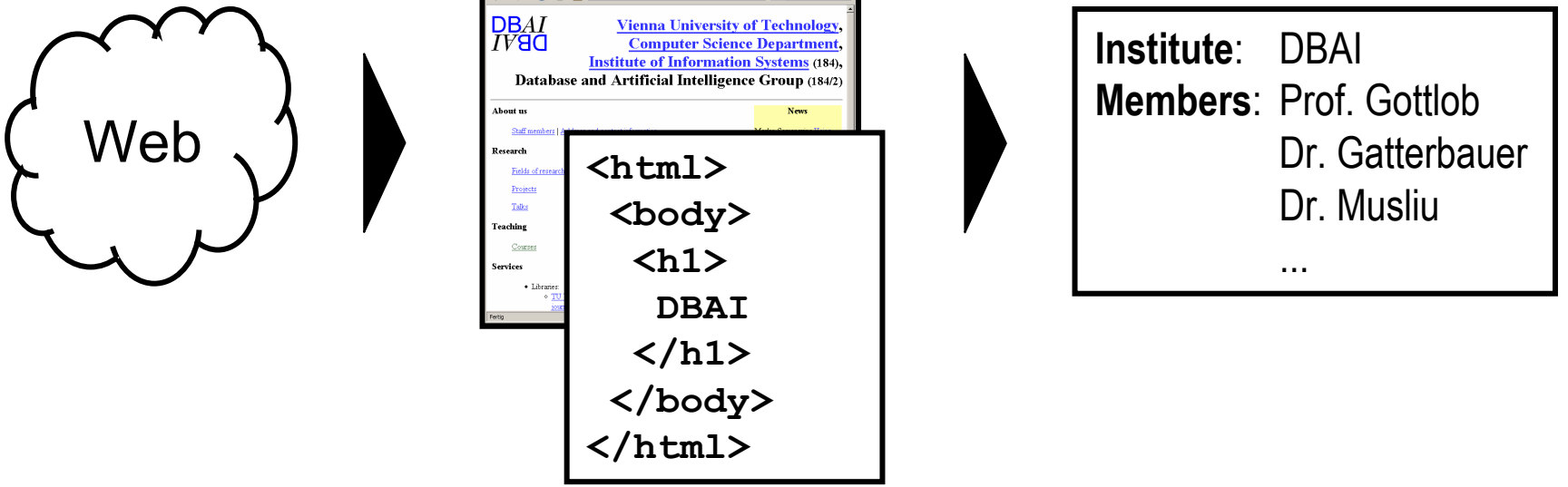
Supervision
Prof. Georg Gottlob
Dr. Wolfgang Gatterbauer

Date: 2008-01-22

Web Information Extraction

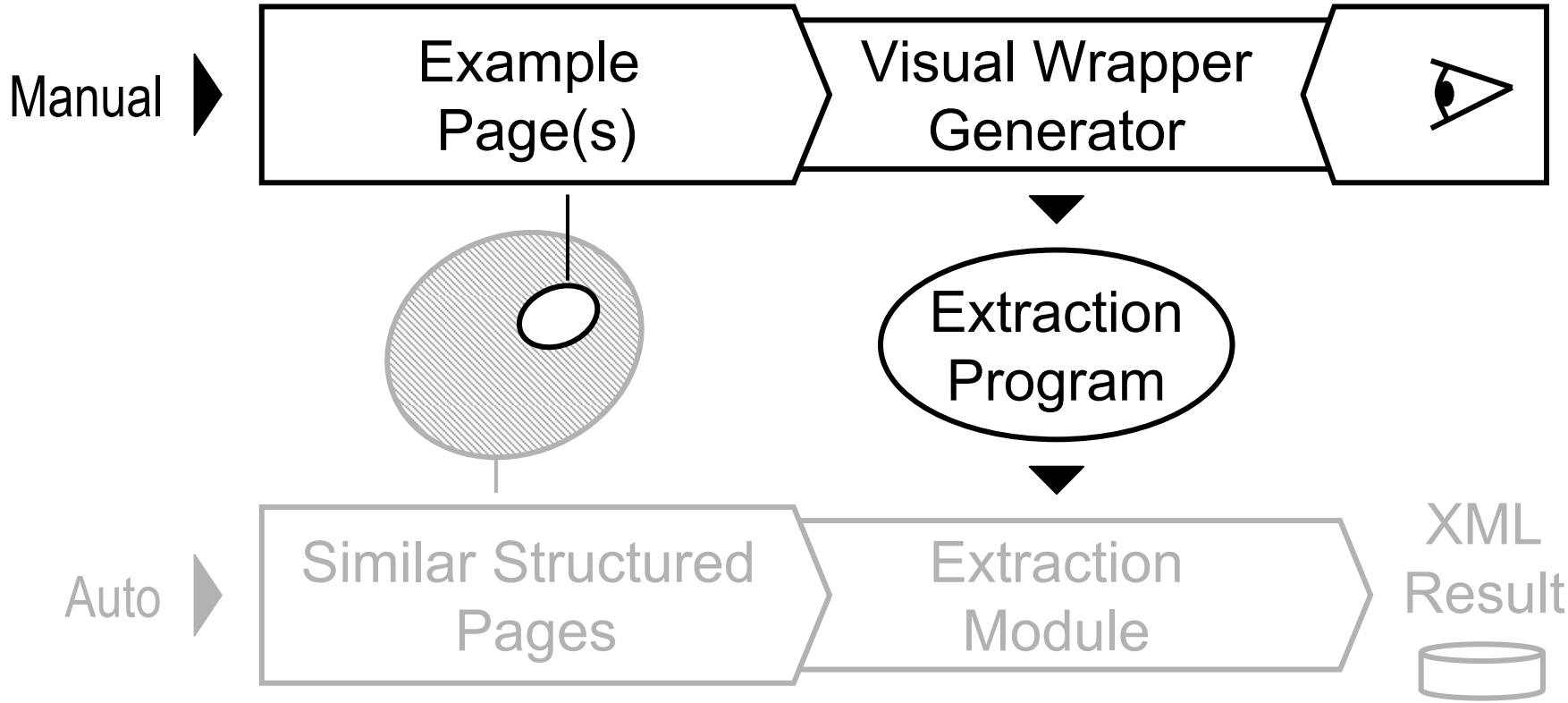
Semistructured Data

Structured Data



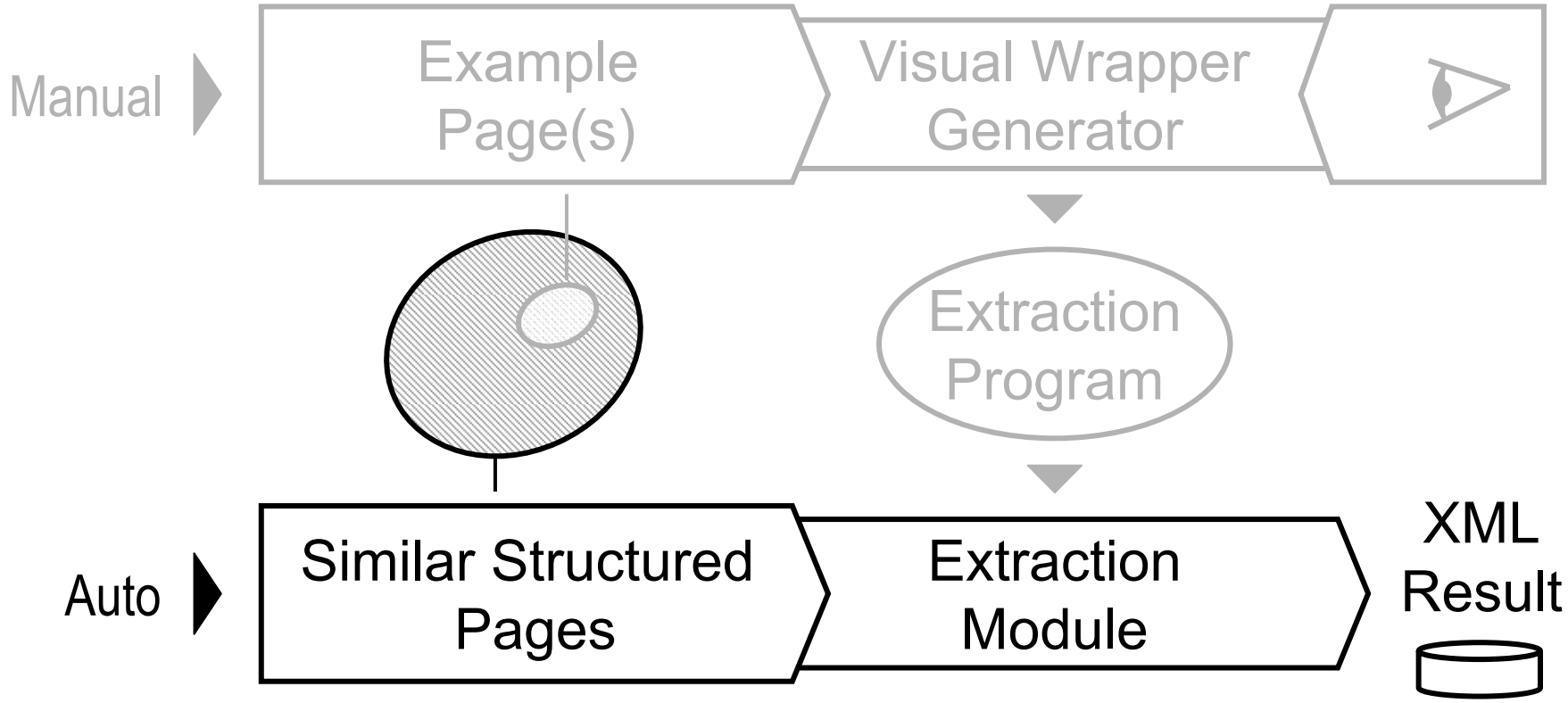
Wrapper

Lixto Concept [1]



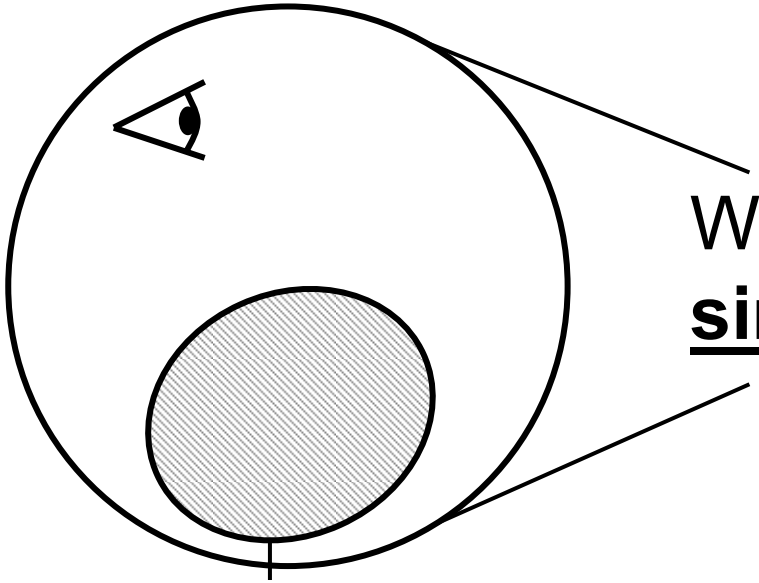
Wrapper

Lixto Concept [1]



DBAI

The Problem

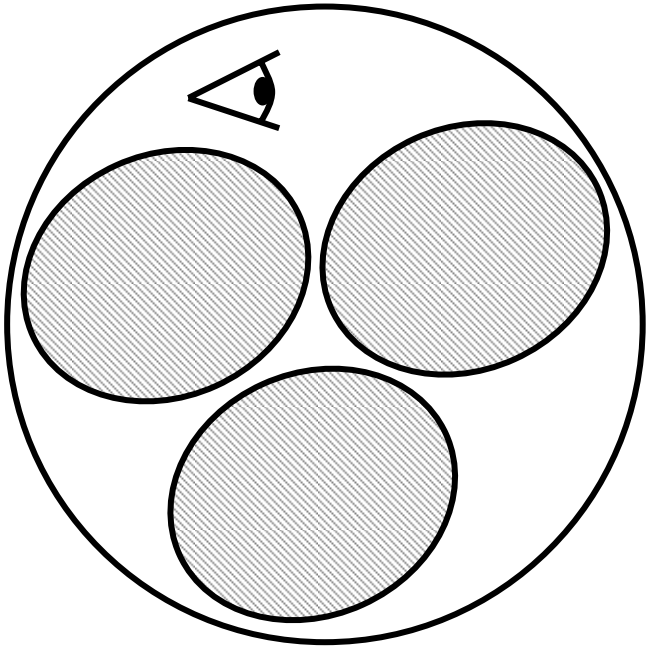


What about (visual) **similar** pages ?

Similar Structured Pages

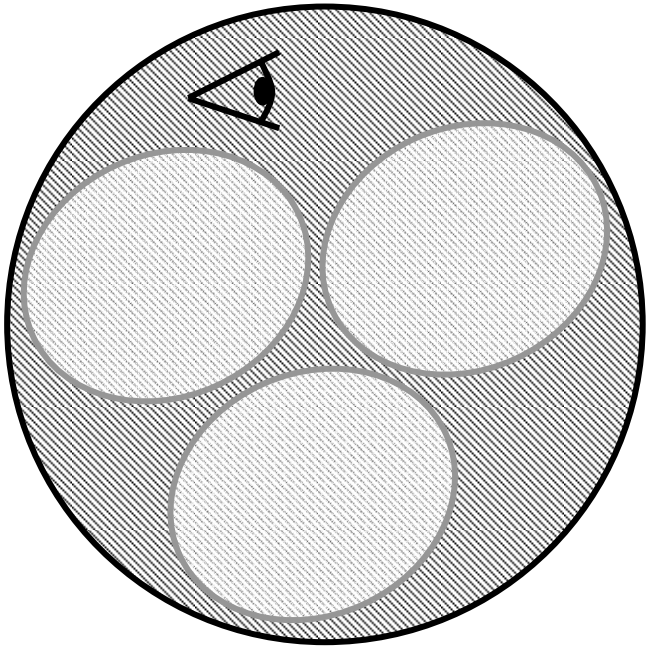
Means *similar structured* with regards to HTML

Multiple Wrappers ?



Needs **multiple** manual wrapper definitions and higher maintaining efforts

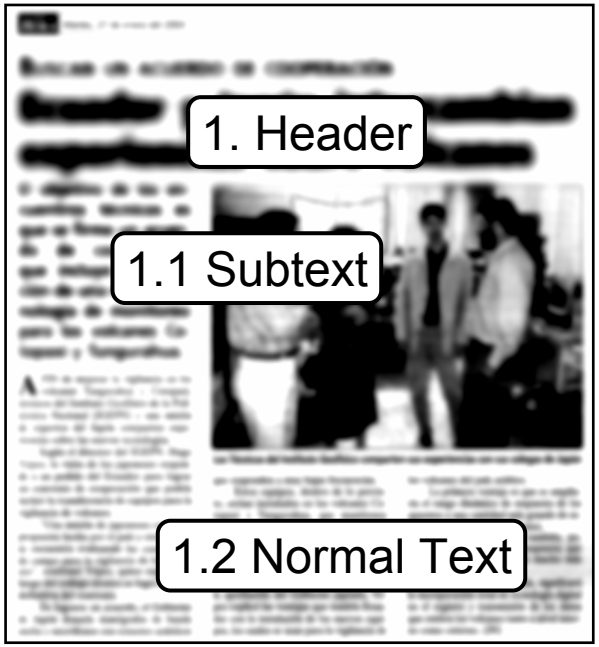
Visual Approach ?



Try to use general visual rules for **reducing** special wrapper dependence

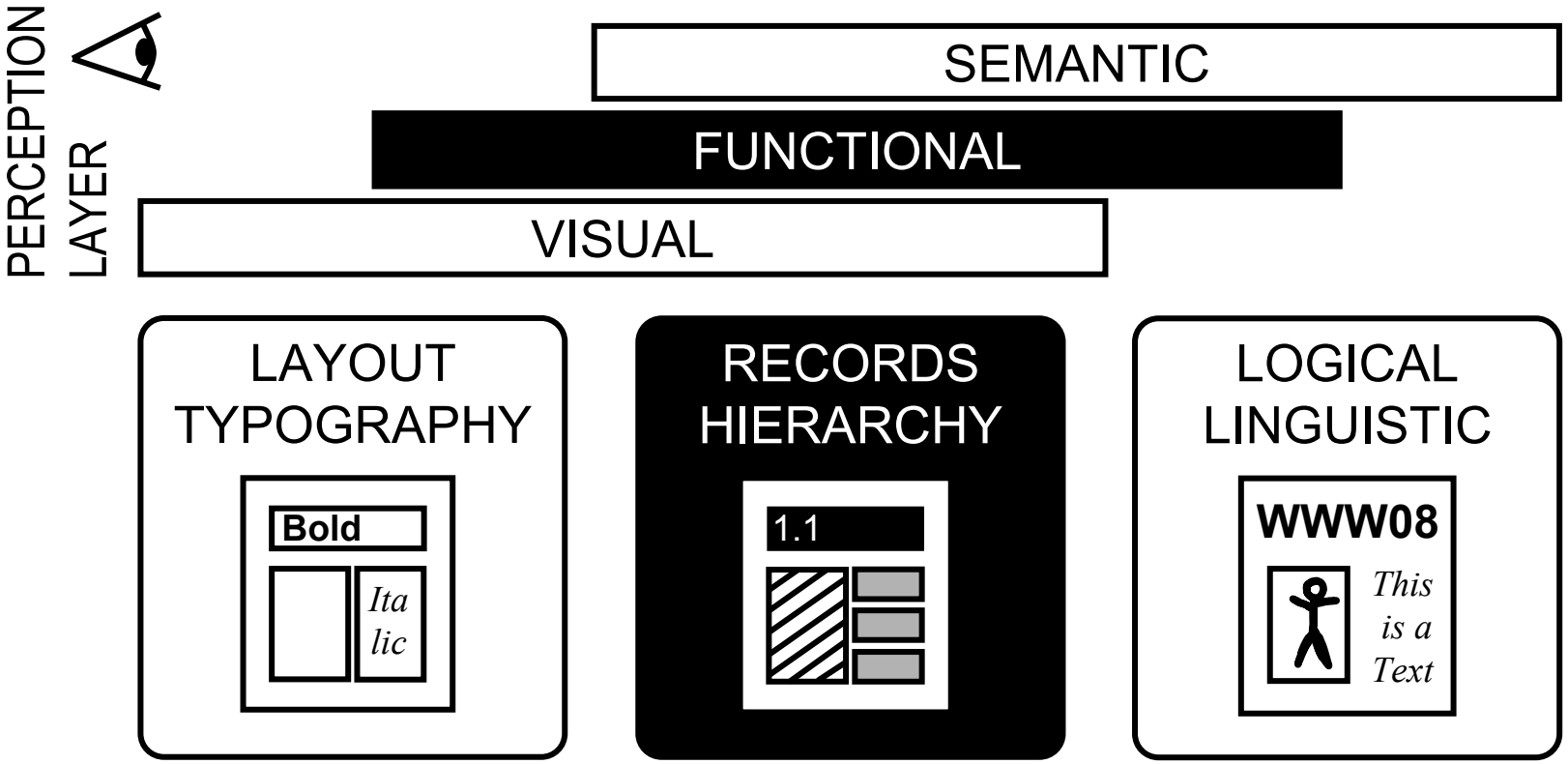
What could be deduced ?

Newspaper [2]

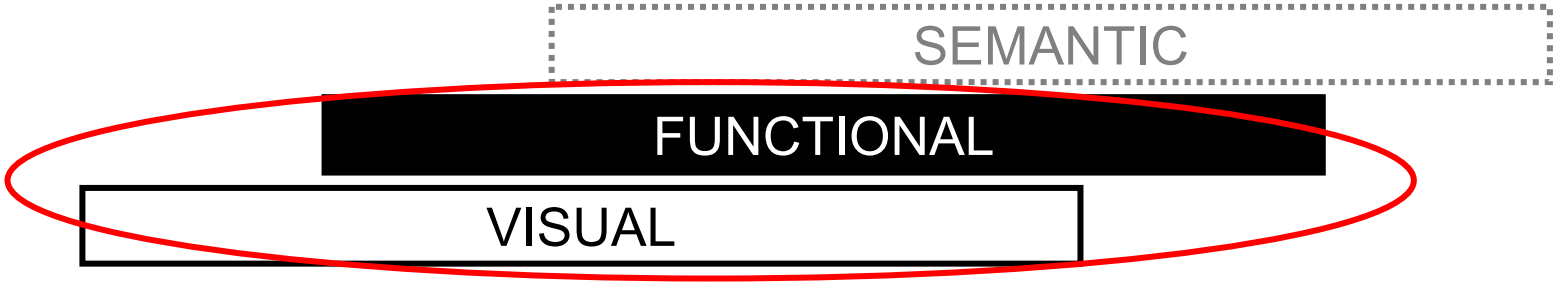


Semantic is present even without knowing the content

Functional Semantics [3]



The Solution

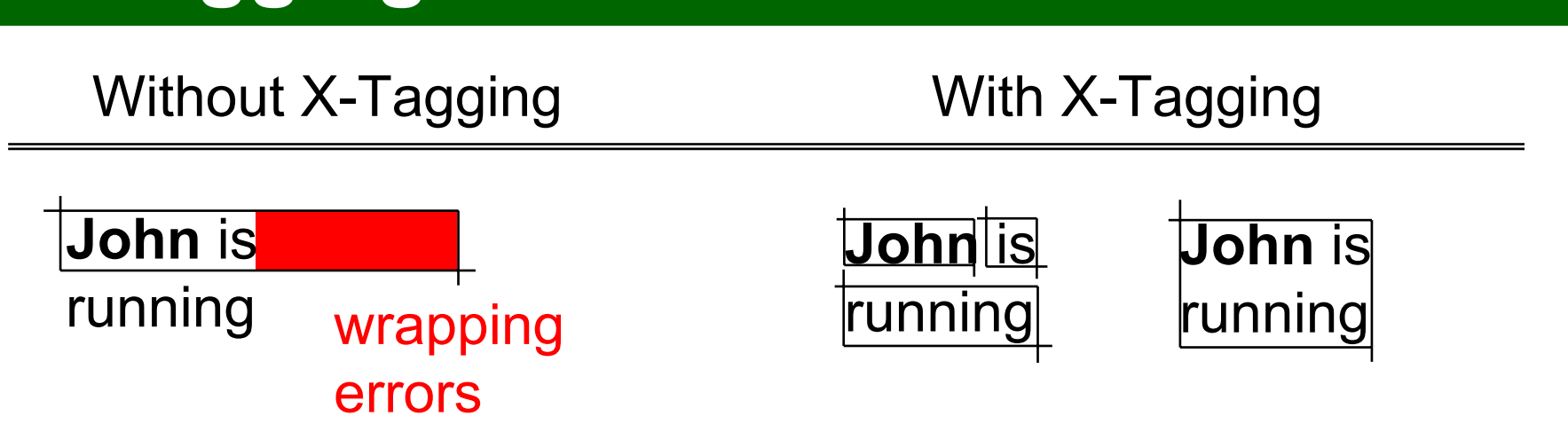


The **REDEVILA** approach

RE cord
DE tection on the
VI sual
LA yer

1. Box Identification
2. Segmentation
3. Classification
4. Ordering
5. Hierarchy

X-Tagging



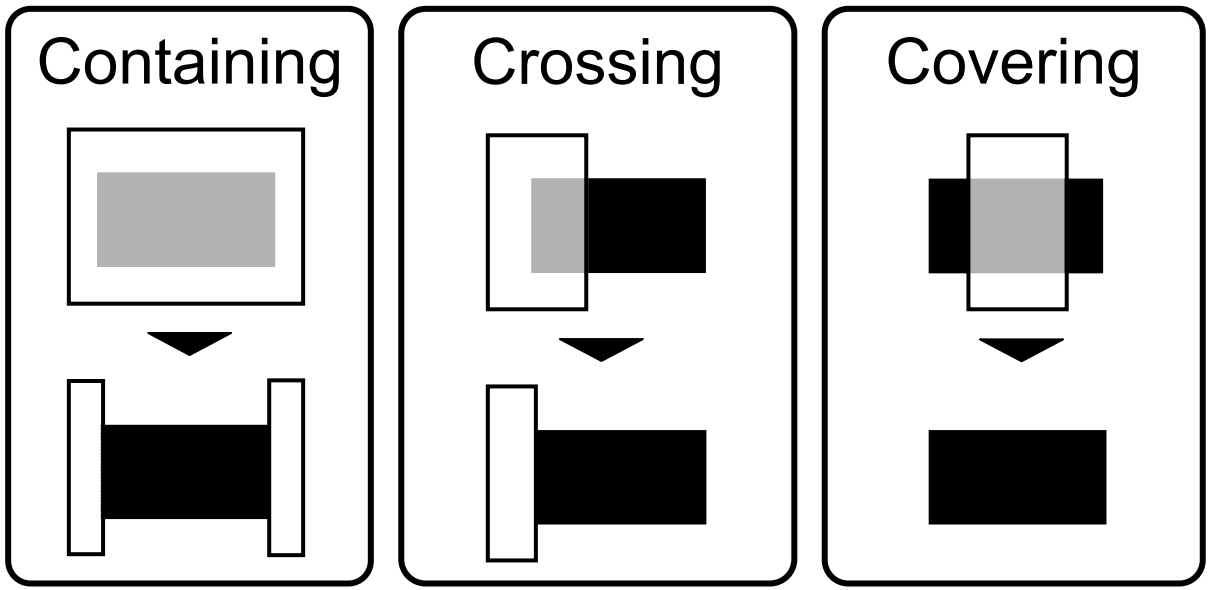

```
<html>
  <body>
    <b>John</b> is
    running
  </body>
</html>
```

```
<html>
  <body>
    <b><x>John</x></b><x>is<x>
    <x>text</x>
  </body>
</html>
```

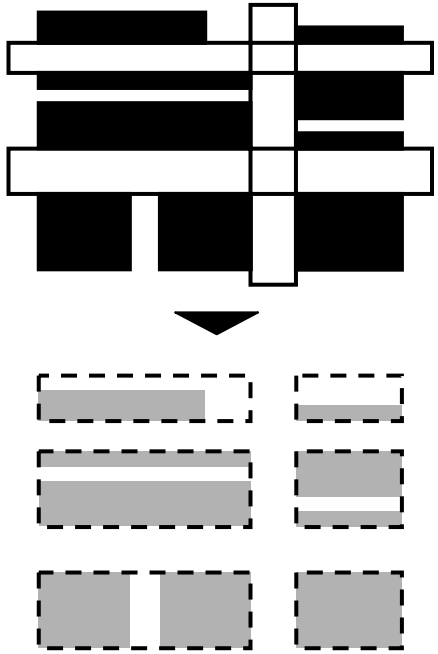

DBAI

VIPS Algorithm [4]

Basic Operations

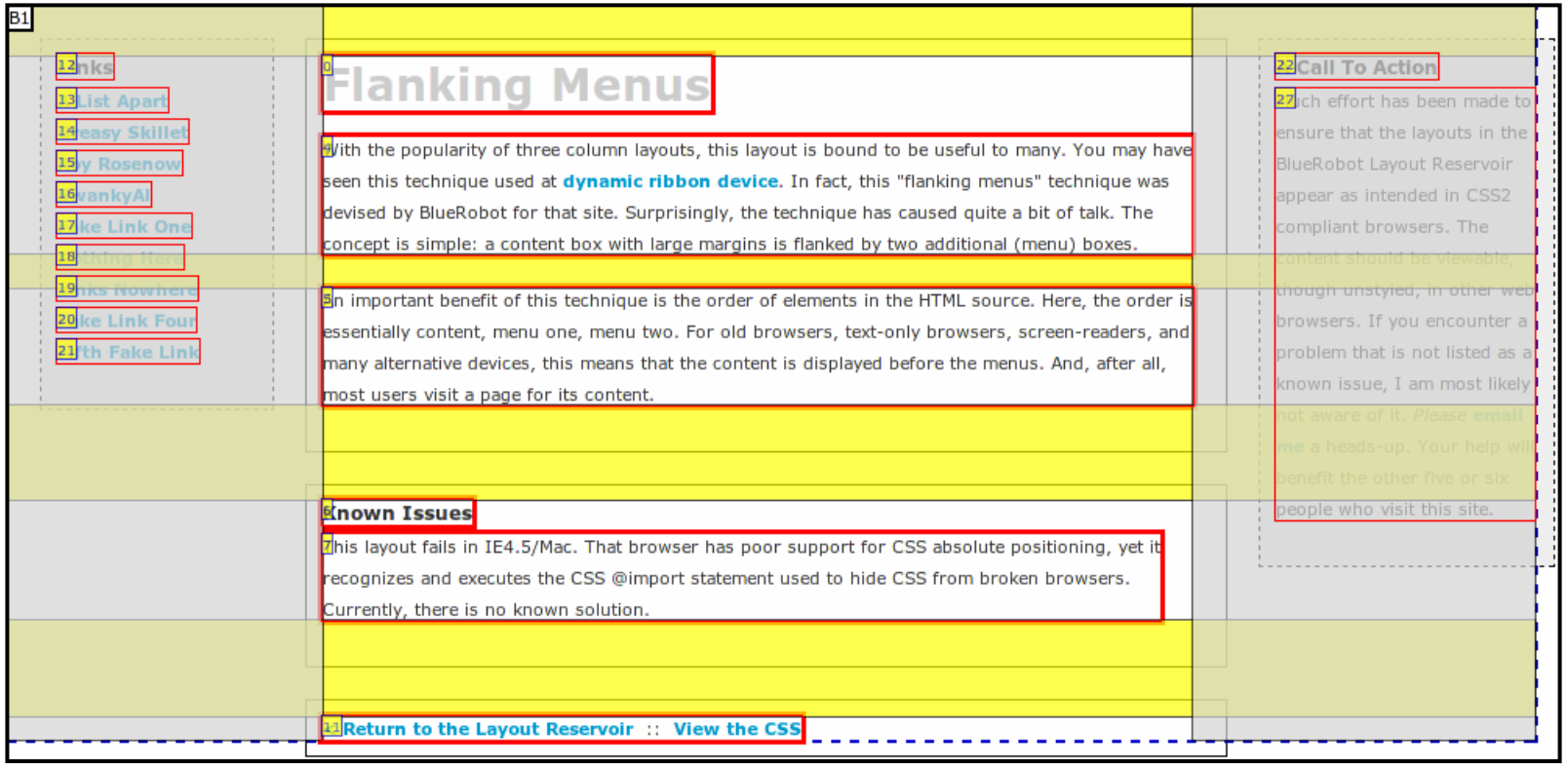


Invertion



DBAI

Segmentation Example [5]

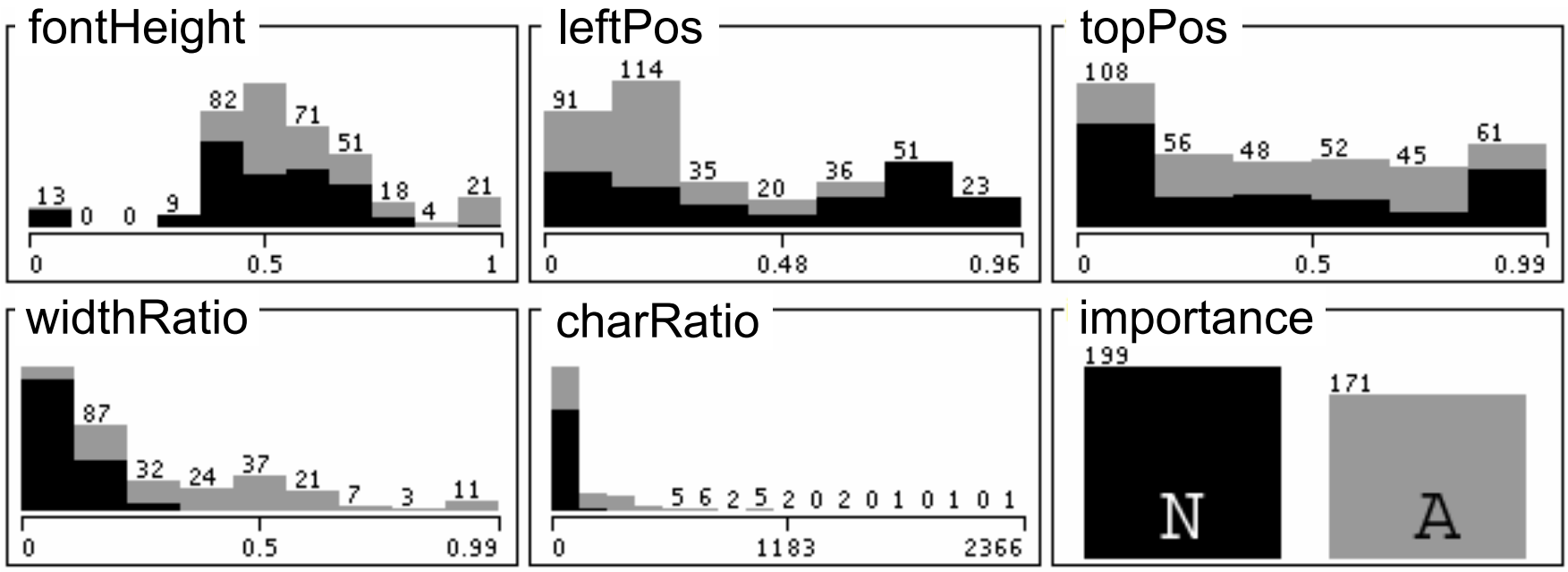


WEKA Toolkit [6]

Important vs. Noisy segments

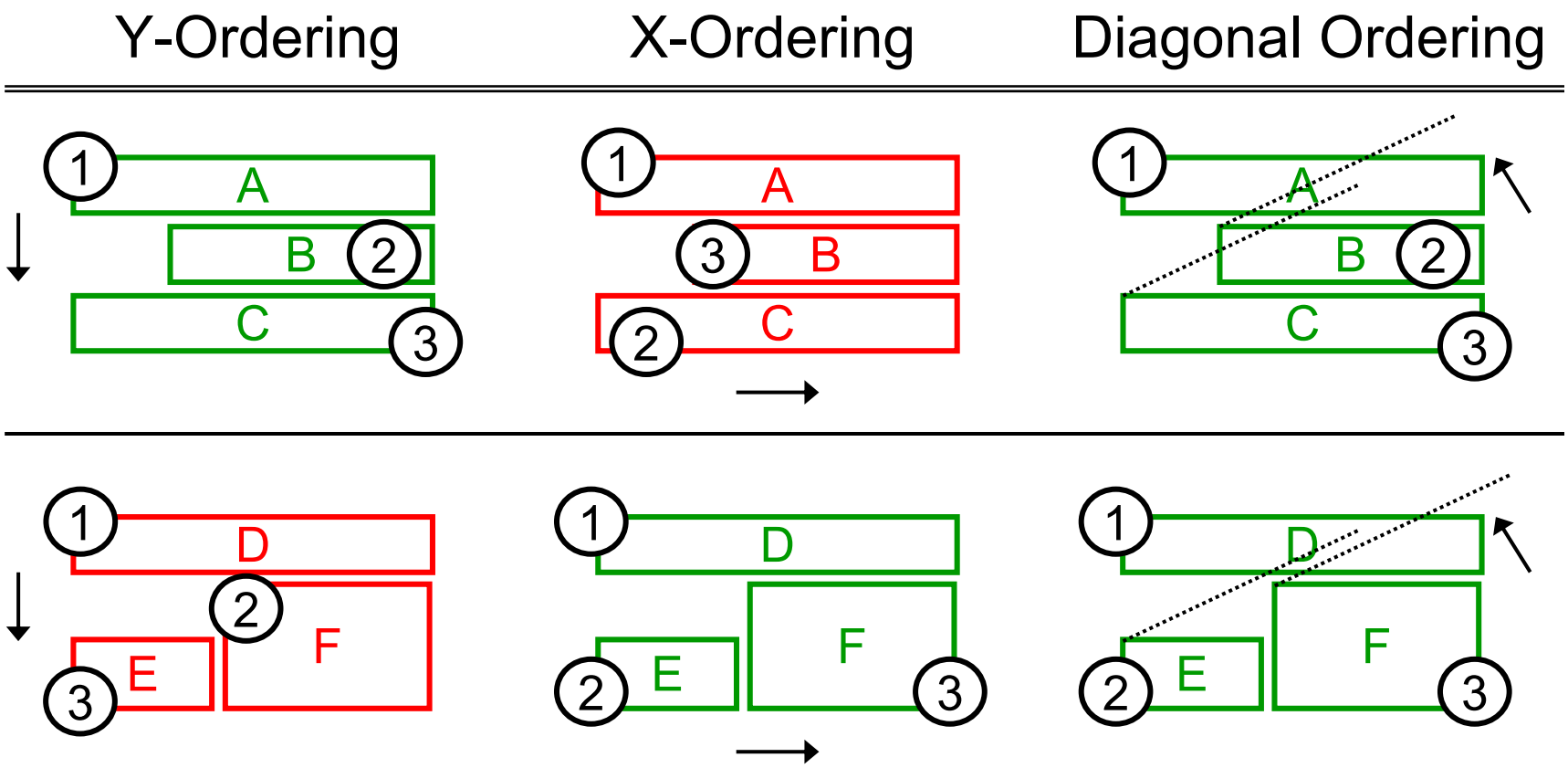
- 370 segments from web pages
- WEKA machine learning toolkit
- Feature reduction
- PART algorithm
  - C4.5 decision tree algorithm

Final Feature Set



DBAI

Diagonal Ordering



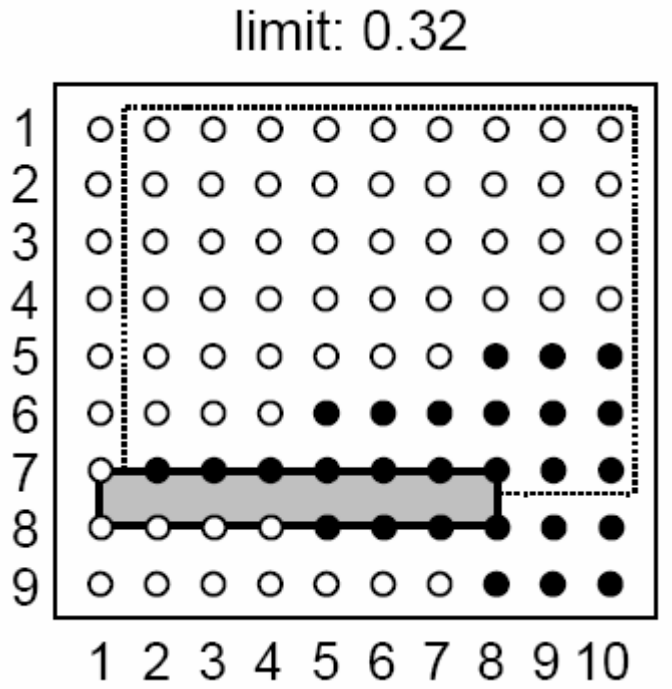
Diagonal Ordering Limit

Limit for the arctan between the two box corners:

$$limit = \frac{1}{2 + \frac{2\,b_{max}}{w_{max}} \sqrt{\frac{b_{max}}{w_{max}}}}$$

b_{max} = maximum width of the two boxes

w_{max} = maximum width of parent structure



Hierarchy Detection

- Monohierarchical structures
- Multitopological Grid
- Hierarchy model: *b.x.x*

  b = record start flag *{true, false}*, *x* = hierarchy depth

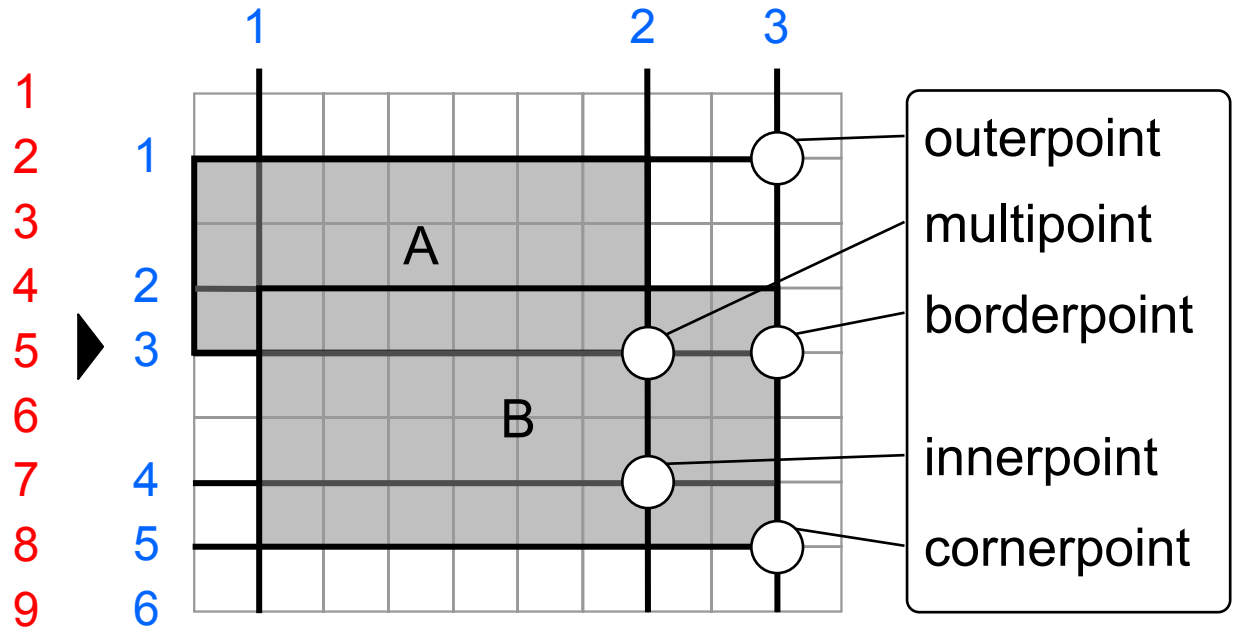
- Record start

$$fontLimit = 11 + (0.25 \cdot maxFontHeight)$$

DBAI

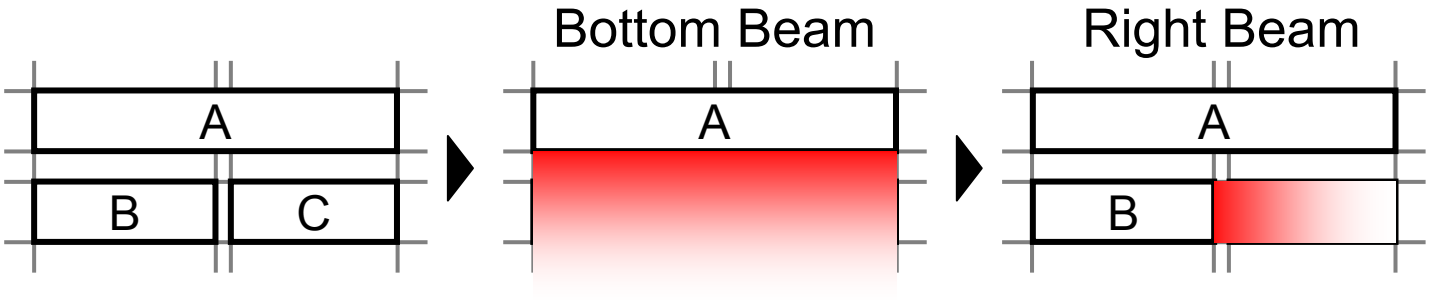
Multitopological Grid Concept

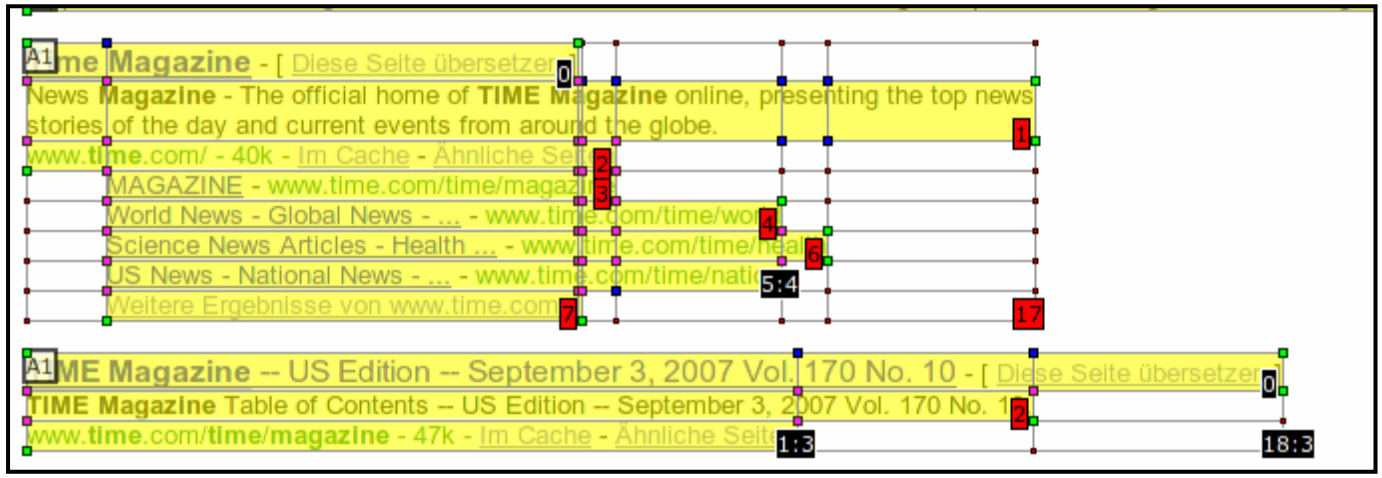
Minimal Grid: Screen Coordinates ▶ Logical Coordinates



DBAI

Multitopological Grid Example [7]





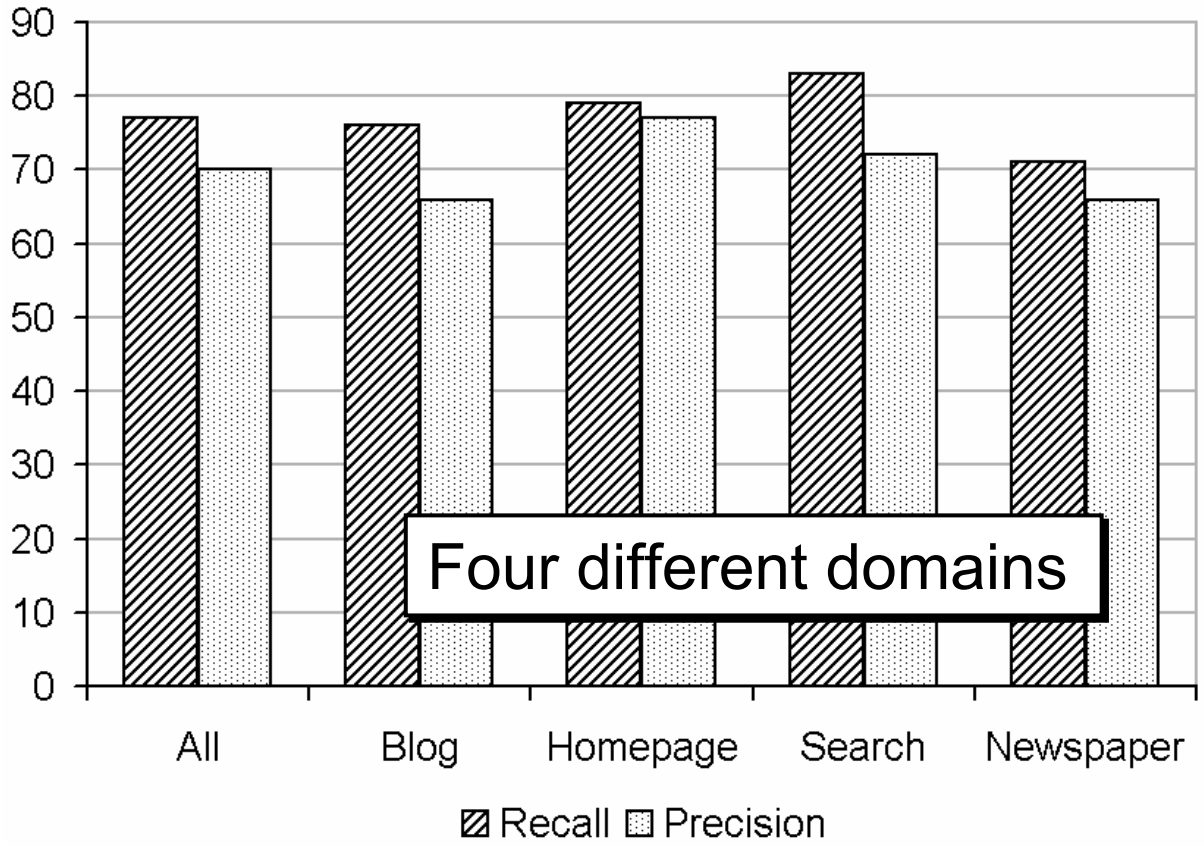
DBAI

Experimental Results

Web Pages: 85
Record Count: 1086

Correct: 836
False Positives: 351
False Negatives: 241

Recall: 77%
Precision: 70%
F-Measure: 73%



Semantic (Domain) Dependence [8]

Webpage

REDEVILA Result

DBAI

REDEVILA Example [9]

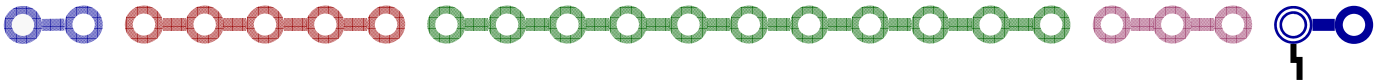
Webpage

REDEVILA Result

DBAI

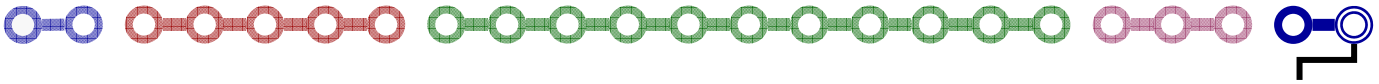
Conclusion

- Domain independence not satisfying
- Definition of distance difficult
- Would make current wrapper approaches more robust
- Potential for single record detection
- Clear separation between "tag" and "visual" approaches



Future Work

- Introducing domain dependence
- Automatic rule generation for the MT Grid
- Considering colored headers
- Considering the layout (column) structure
- Integration with tag information
- Integration of table models with substructured lists



Thank you for your attention

References

1. R. Baumgartner, S. Flesca, and G. Gottlob. Visual web information extraction with lixto. In VLDB ’01: Proceedings of the 27th International Conference on Very Large Data Bases, pages 119–128, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
2. http://www.bosai.go.jp/e/international
3. D. S. Doermann, A. Rosenfeld, and E. Rivlin. The function of documents. In ICDAR ’97: Proceedings of the 4th International Conference on Document Analysis and Recognition, pages 1077–1081, Washington, DC, USA, 1997. IEEE Computer Society.
4. D. Cai, S. Yu, J. Wen, and W. Ma. Extracting content structure for web pages based on visual representation. In Proc. 5th Asian-PacificWeb Conference (Web Technologies and Applications), pages 406–417. Springer, April 2003.
5. http://bluerobot.com/web/layouts/layout3.html
6. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.Morgan Kaufmann, San Francisco, 2nd edition, 2005.
7. http://www.google.at
8. http://the1review.com
9. http://www.google.com

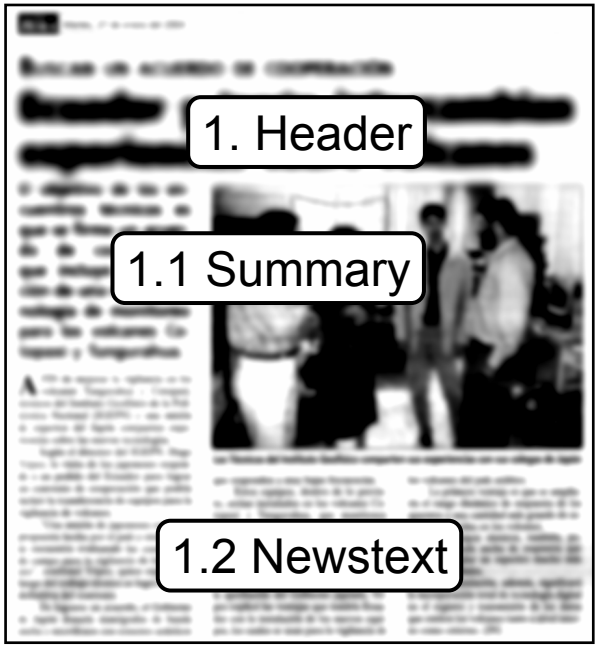
DBAI

BACKUP

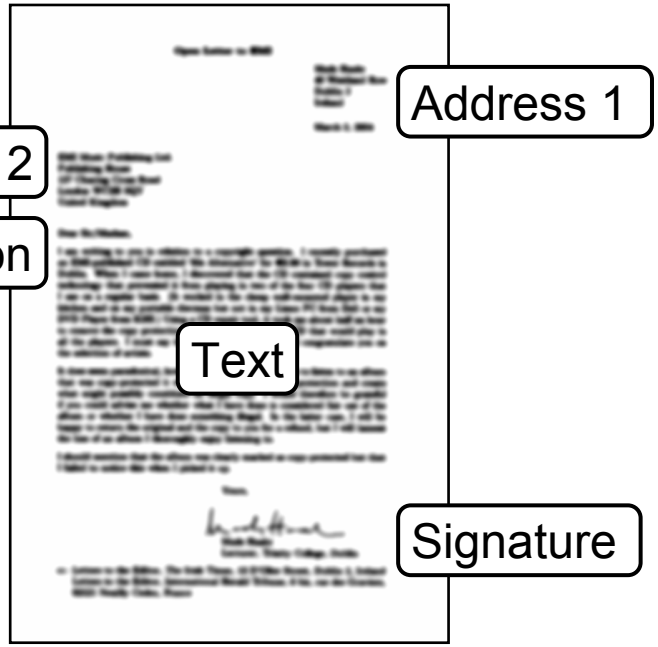
BACKUP

Domain Dependent Functional Semantics

NEWSPAPER

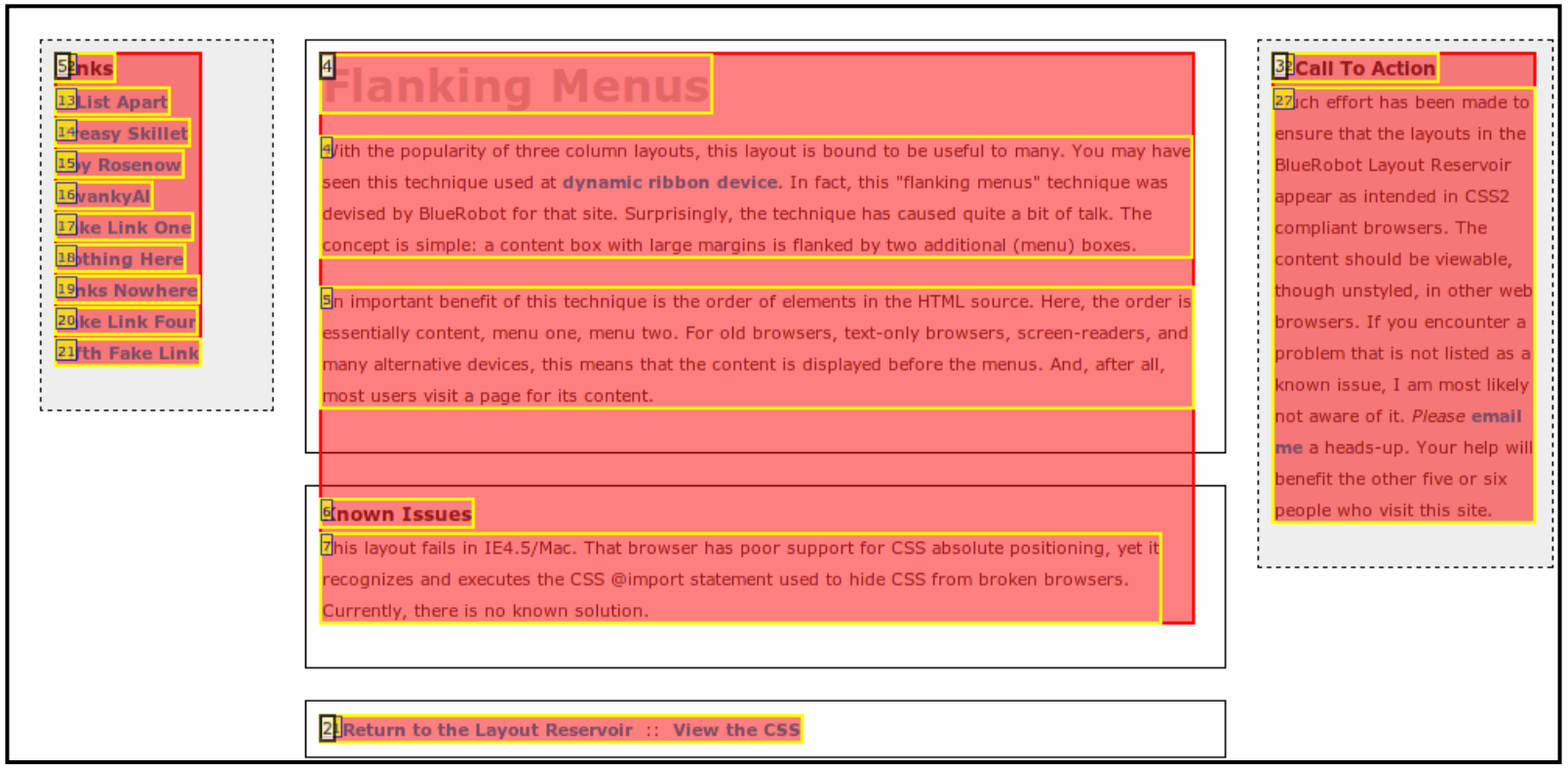


LETTER



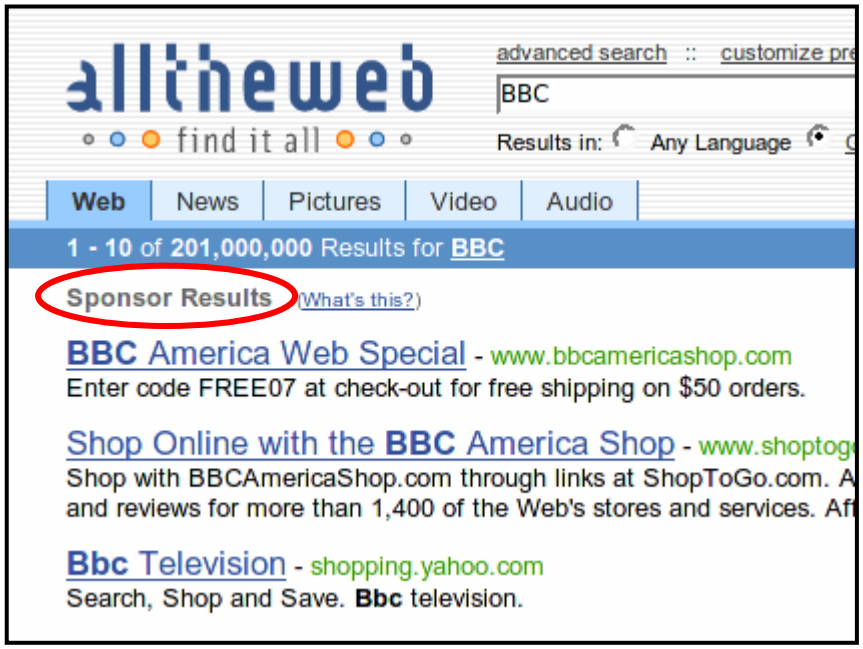
DBAI

Segmentation Example II

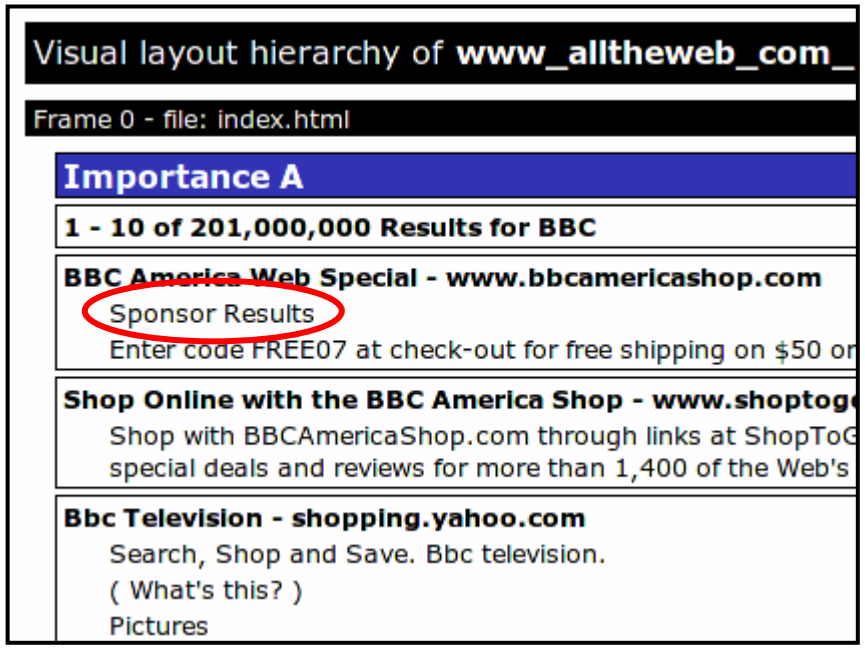


Problem: "Small Line Above" Rule

Webpage



REDEVILA Result

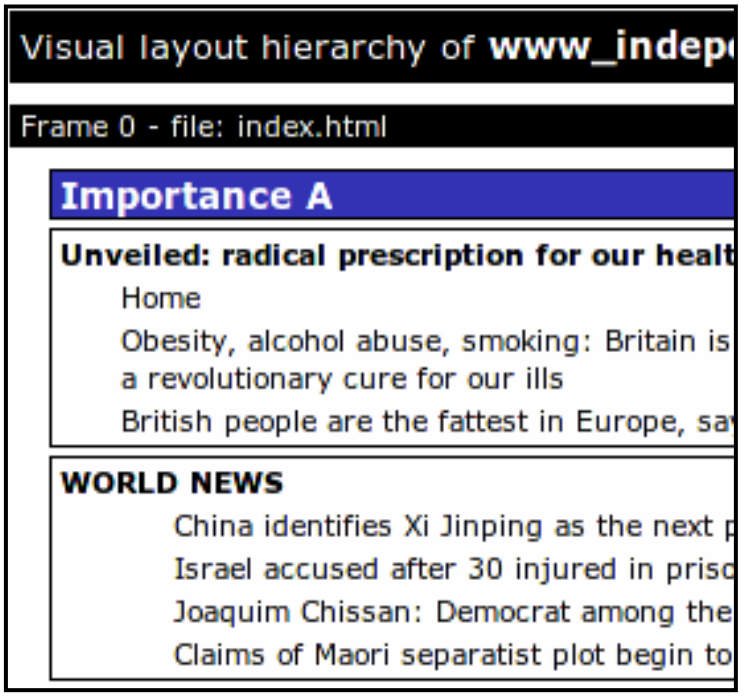


DBAI

REDEVILA Example I

Webpage

REDEVILA Result



DBAI