



Institute of Information Systems
Database and Artificial Intelligence Group

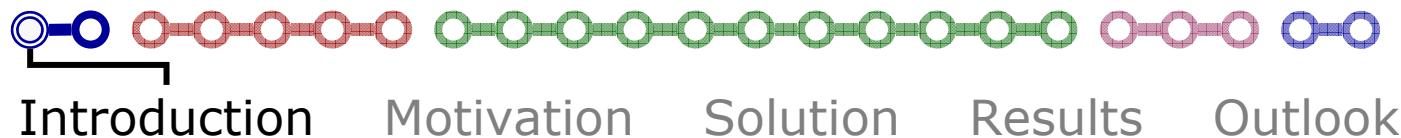
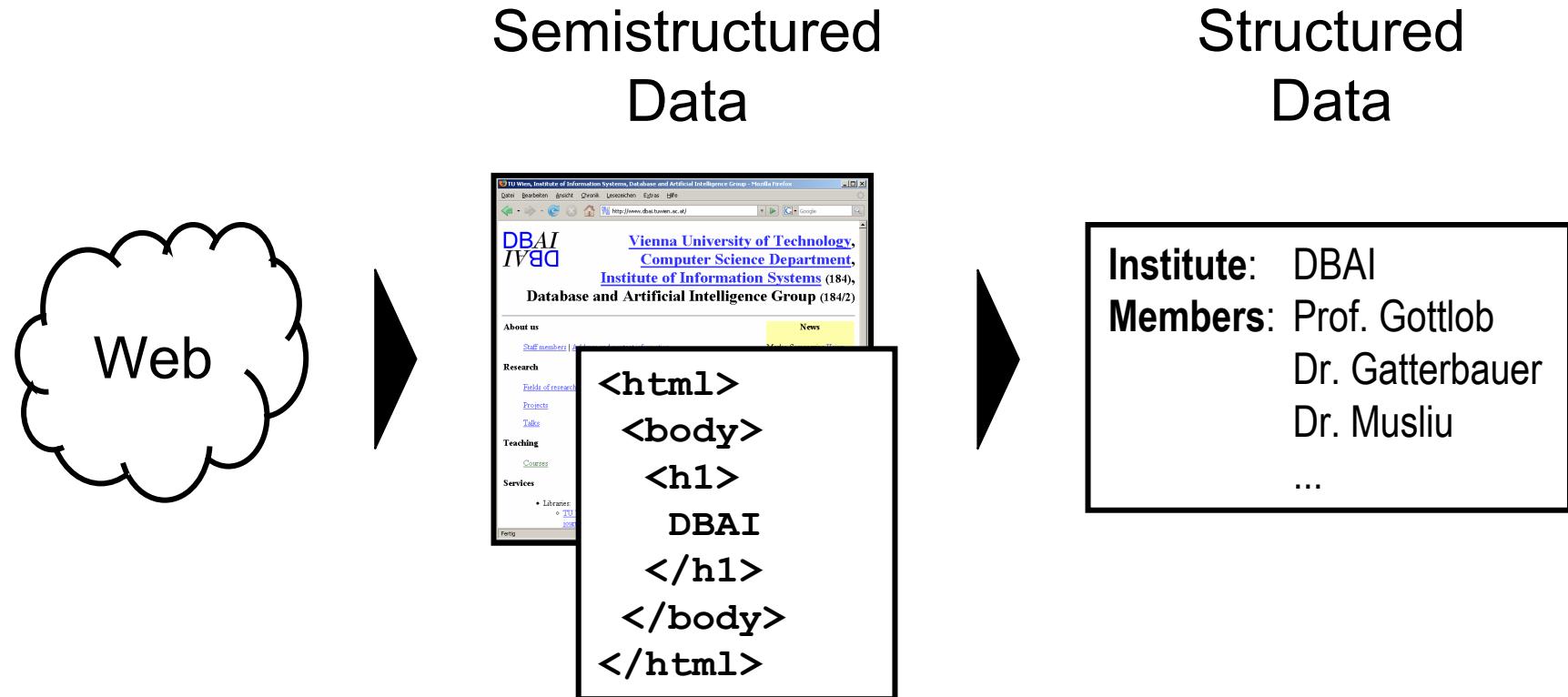
Functional Semantic Analysis of Web Pages on the Visual Layer

Presentation of the Master's Thesis
by Bernhard Pollak - 9326613

Supervision
Prof. Georg Gottlob
Dr. Wolfgang Gatterbauer

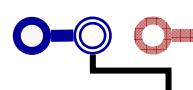
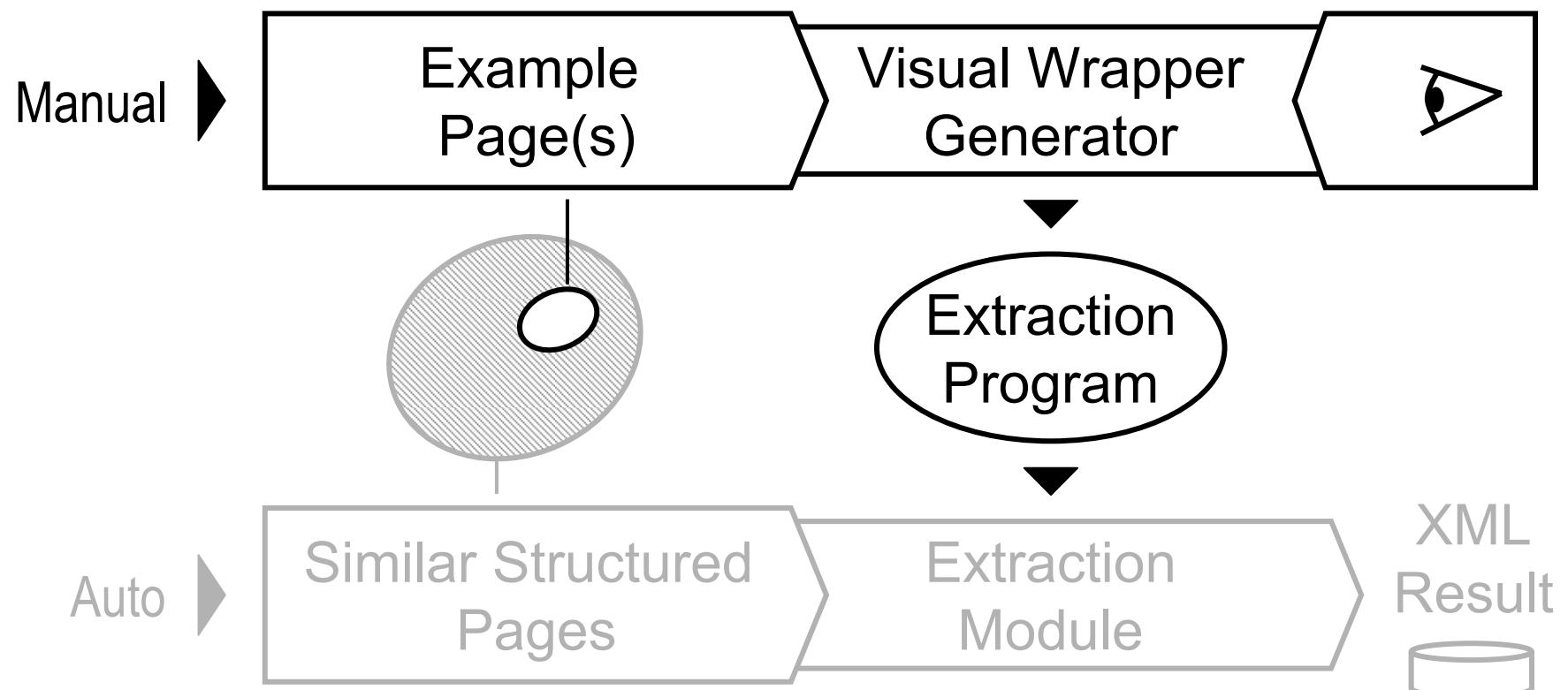
Date: 2008-01-22

Web Information Extraction

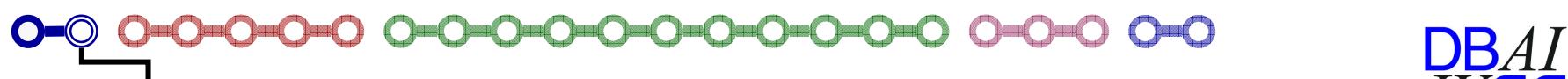


Wrapper

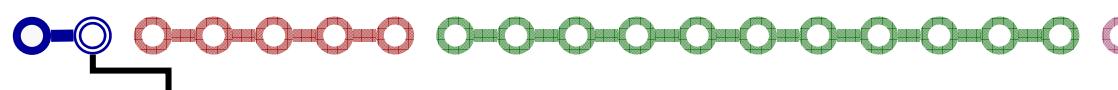
Lixto Concept ¹



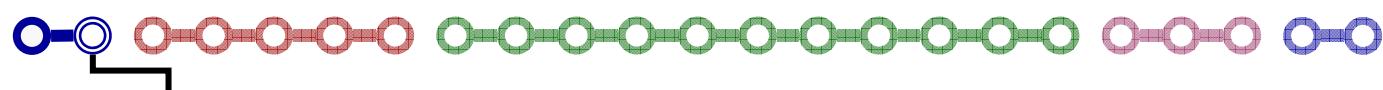
Introduction



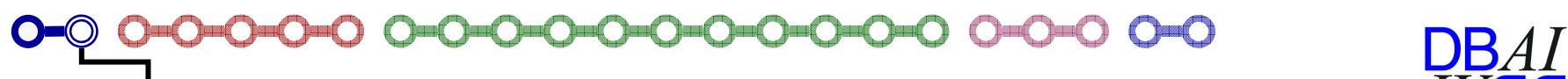
Motivation



Solution



Results

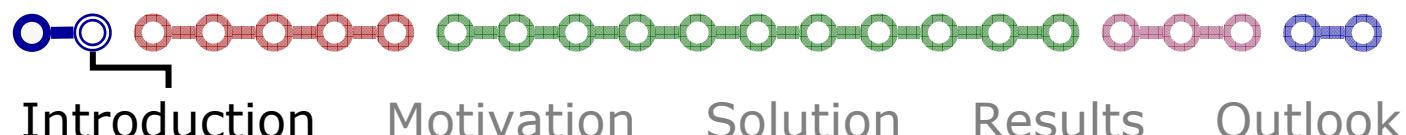
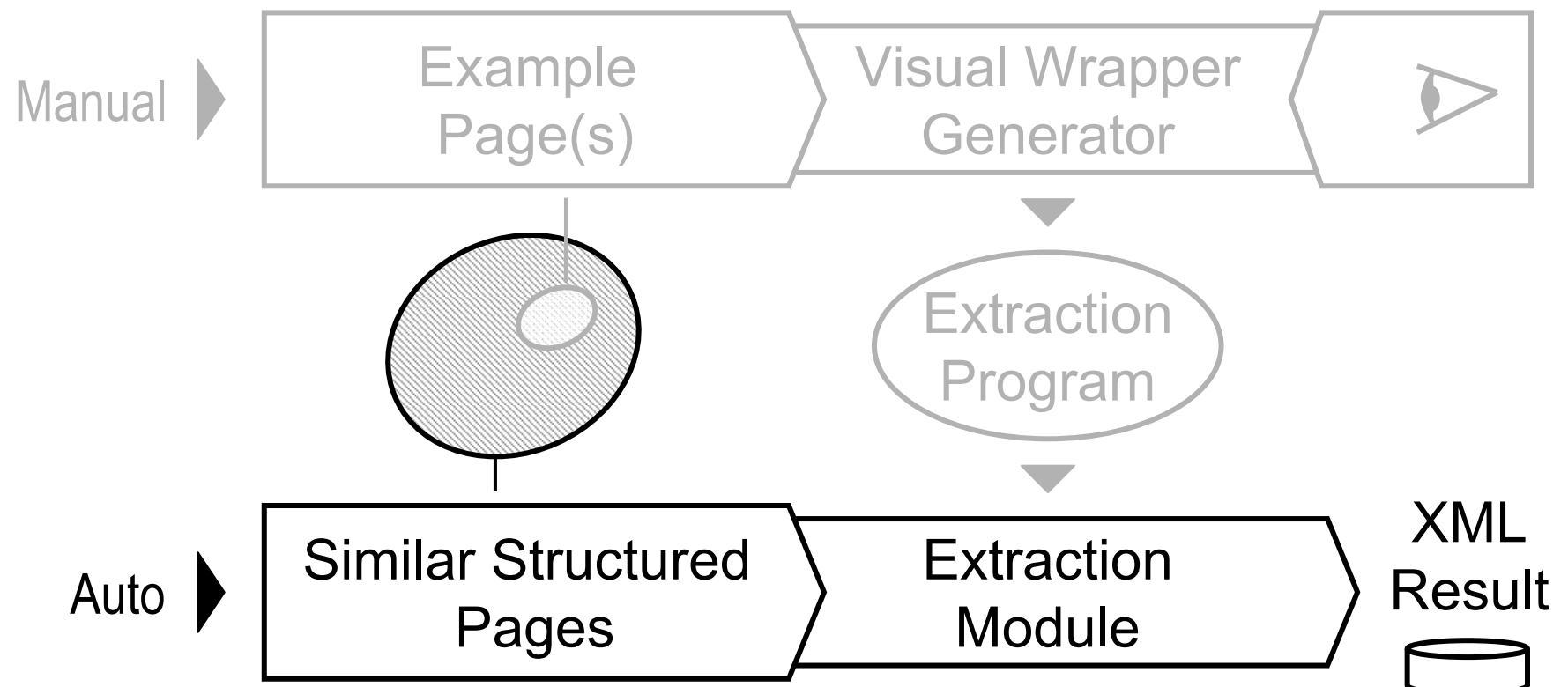


Outlook

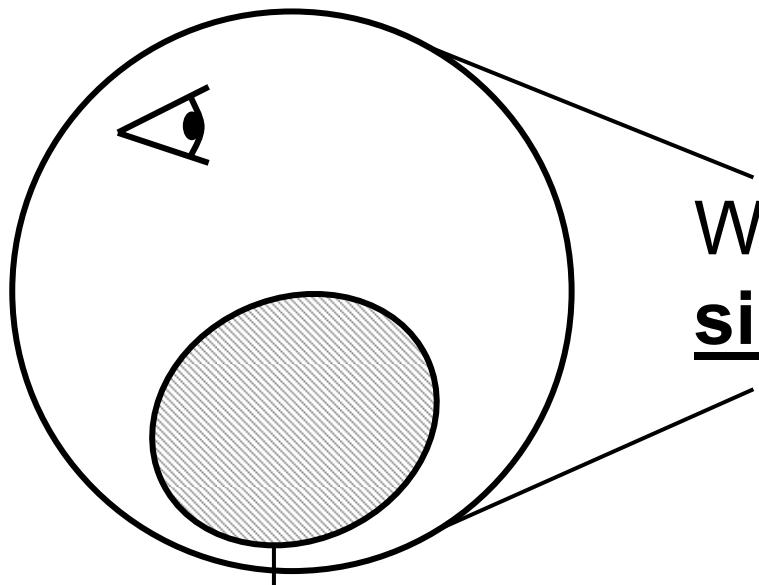
DBAI
DBAI

Wrapper

Lixto Concept ¹



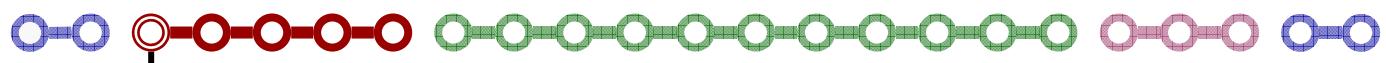
The Problem



What about (visual)
similar pages ?

Similar Structured
Pages

Means *similar structured*
with regards to HTML



Introduction

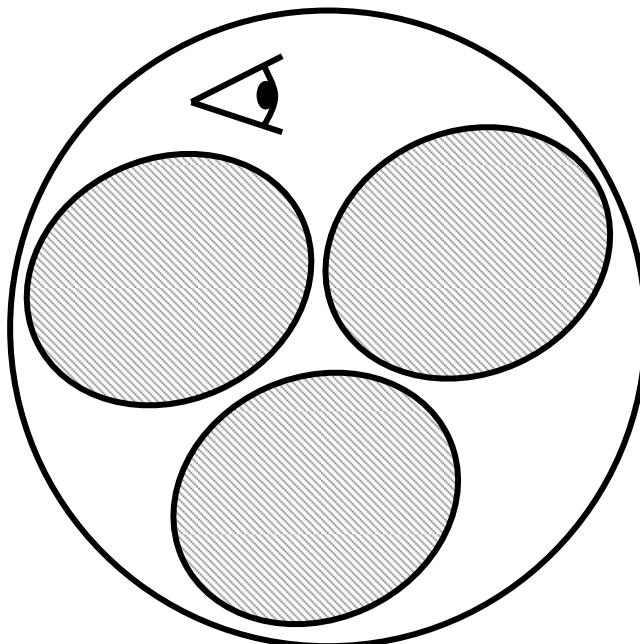
Motivation

Solution

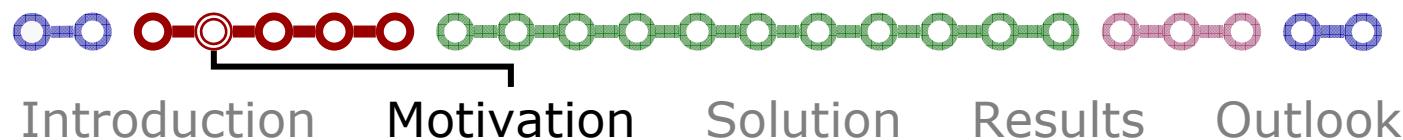
Results

Outlook

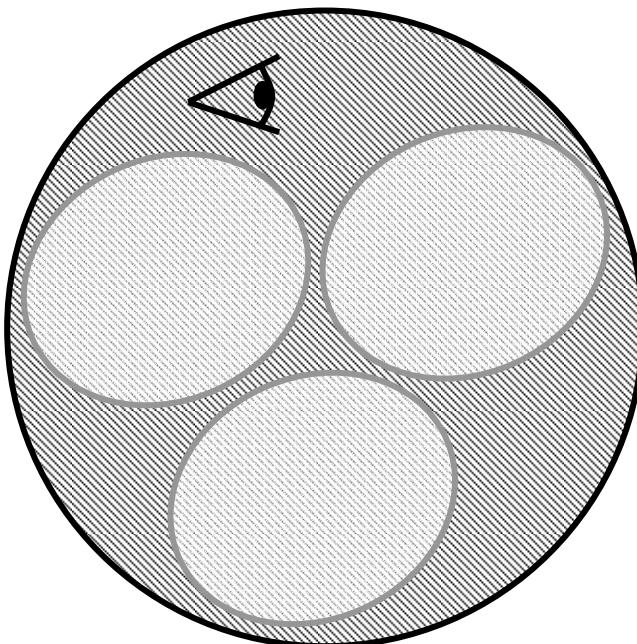
Multiple Wrappers ?



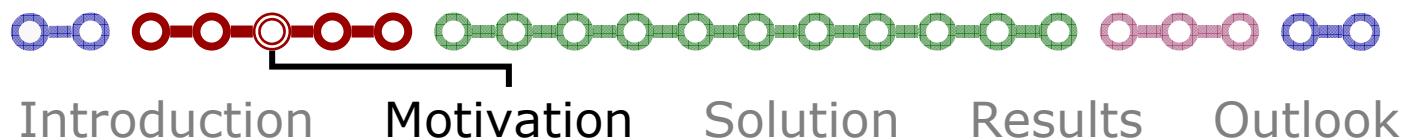
Needs multiple manual wrapper definitions and higher maintaining efforts



Visual Approach ?

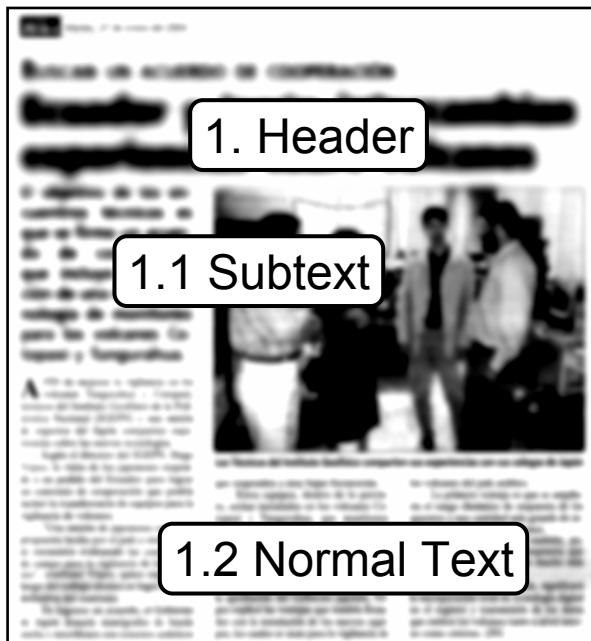


Try to use general visual rules for reducing special wrapper dependence

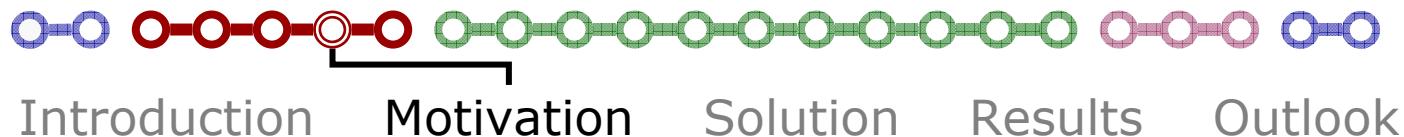


What could be deduced ?

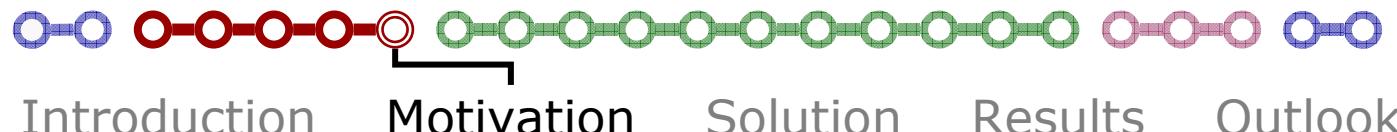
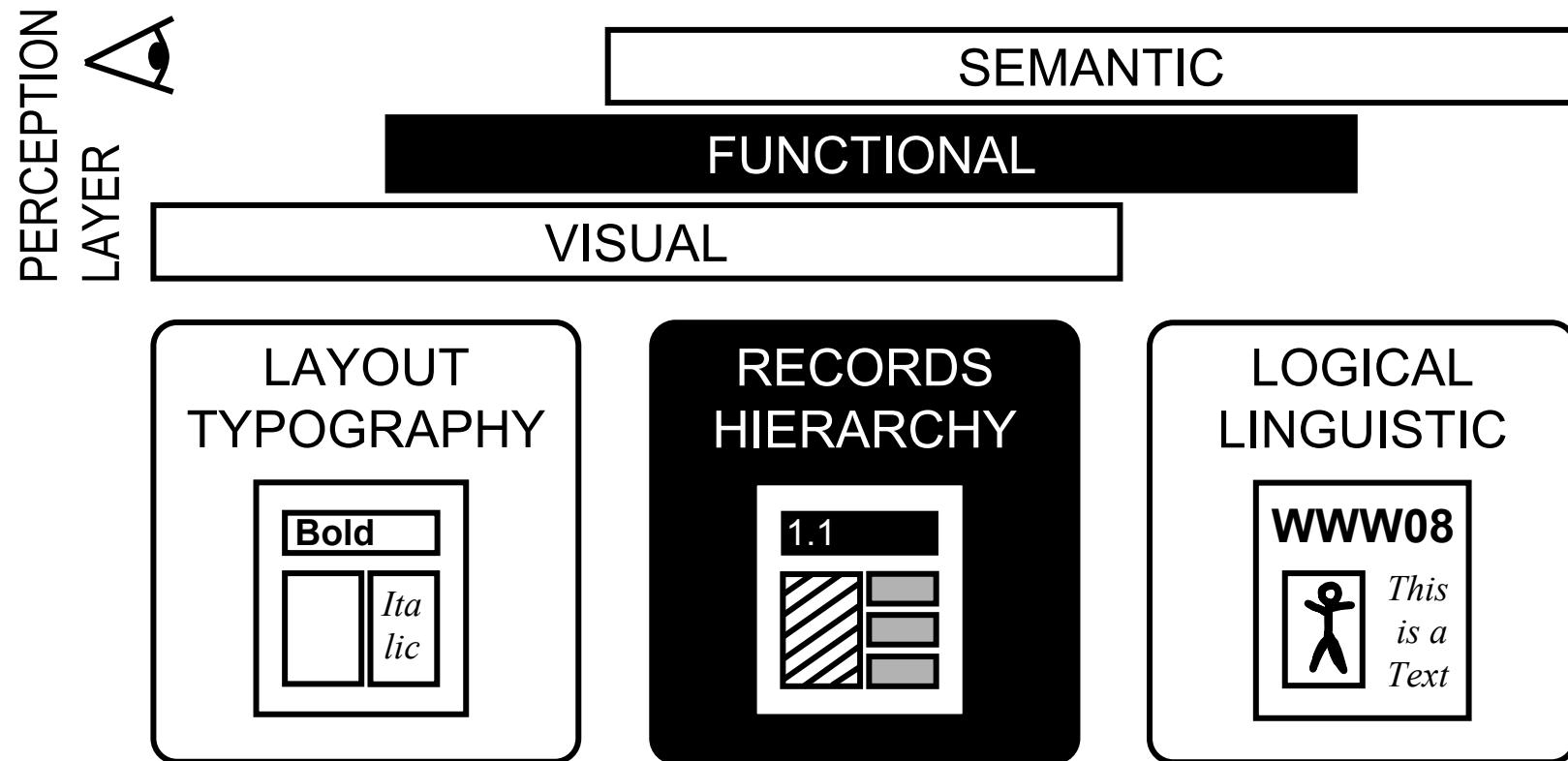
Newspaper²



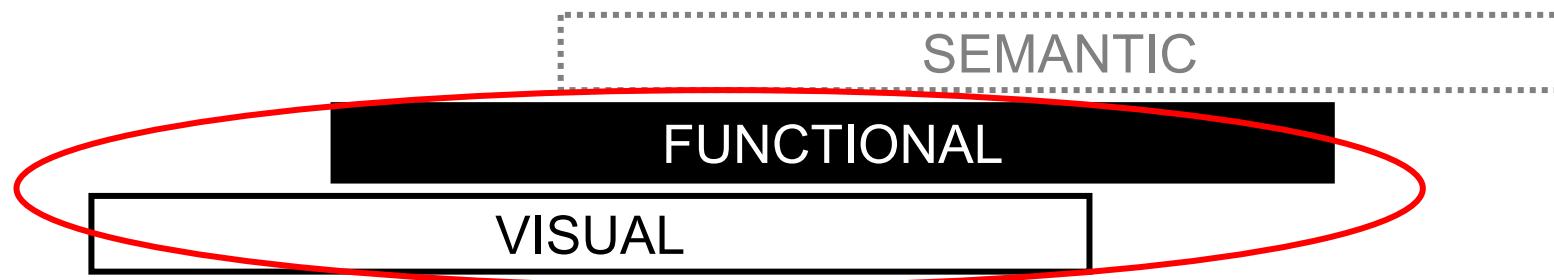
Semantic is present
even without knowing
the content



Functional Semantics³



The Solution



The **REDEVILA** approach

REcord
DEtection on the
Visual
LAyer



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



Introduction

Motivation

Solution

Results

Outlook

X-Tagging

Box Identification

1

Without X-Tagging

John is
running wrapping
errors

With X-Tagging

John is
running

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

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



Introduction

Motivation

Solution

Results

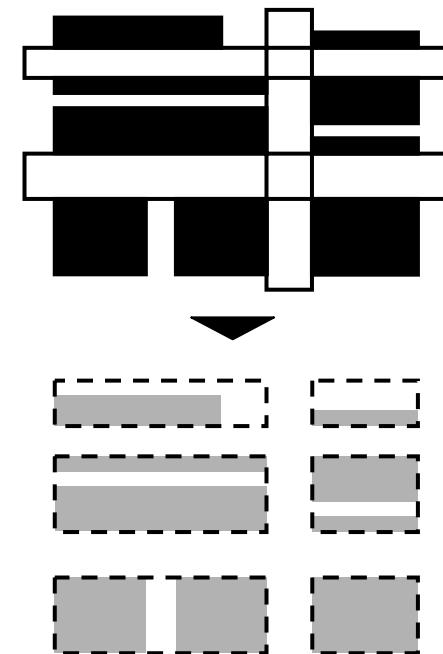
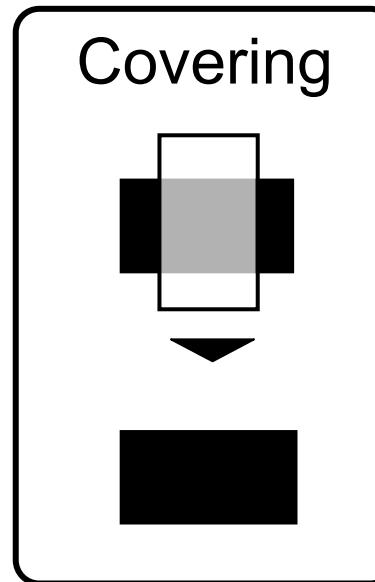
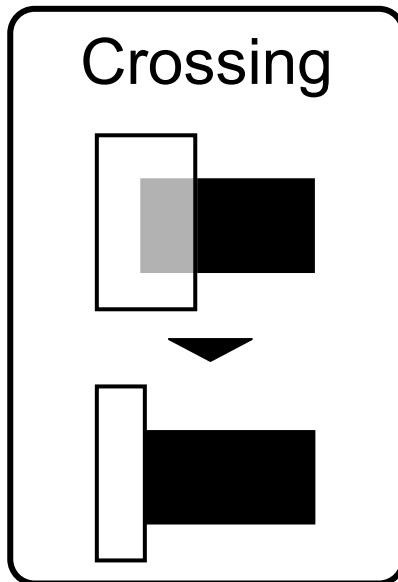
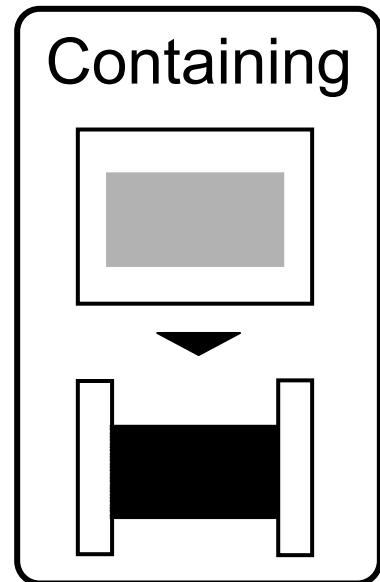
Outlook

DBAI

VIPS Algorithm⁴

Basic Operations

Inversion



Introduction

Motivation

Solution

Results

Outlook

Segmentation Example 5

B1

Flanking Menus

With the popularity of three column layouts, this layout is bound to be useful to many. You may have seen this technique used at [dynamic ribbon device](#). In fact, this "flanking menus" technique was devised by BlueRobot for that site. Surprisingly, the technique has caused quite a bit of talk. The concept is simple: a content box with large margins is flanked by two additional (menu) boxes.

An important benefit of this technique is the order of elements in the HTML source. Here, the order is essentially content, menu one, menu two. For old browsers, text-only browsers, screen-readers, and many alternative devices, this means that the content is displayed before the menus. And, after all, most users visit a page for its content.

Known Issues

This layout fails in IE4.5/Mac. That browser has poor support for CSS absolute positioning, yet it recognizes and executes the CSS @import statement used to hide CSS from broken browsers. Currently, there is no known solution.

[Return to the Layout Reservoir](#) :: [View the CSS](#)



Introduction

Motivation

Solution

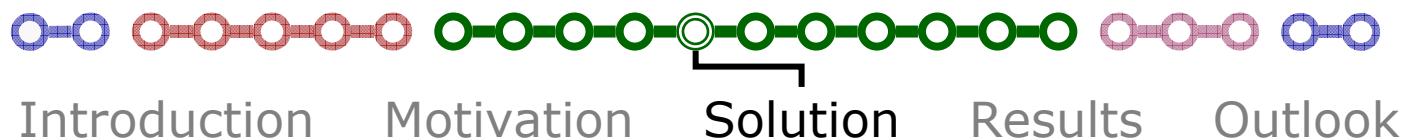
Results

Outlook

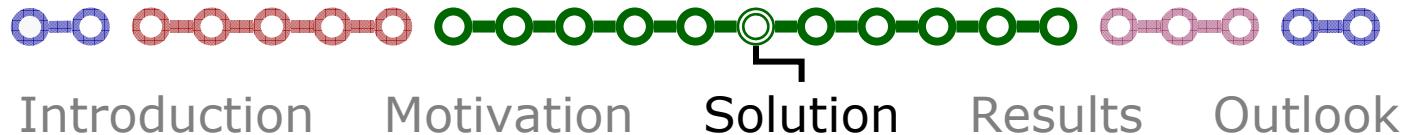
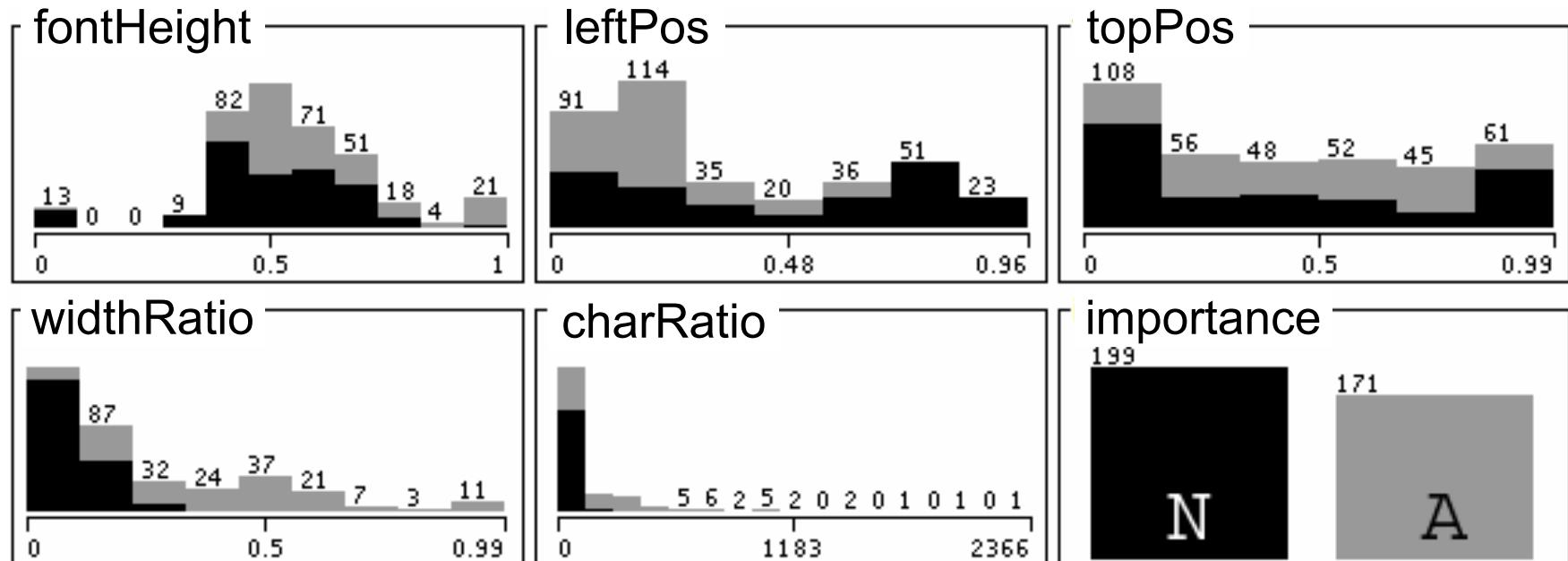
WEKA Toolkit⁶

Important vs. Noisy segments

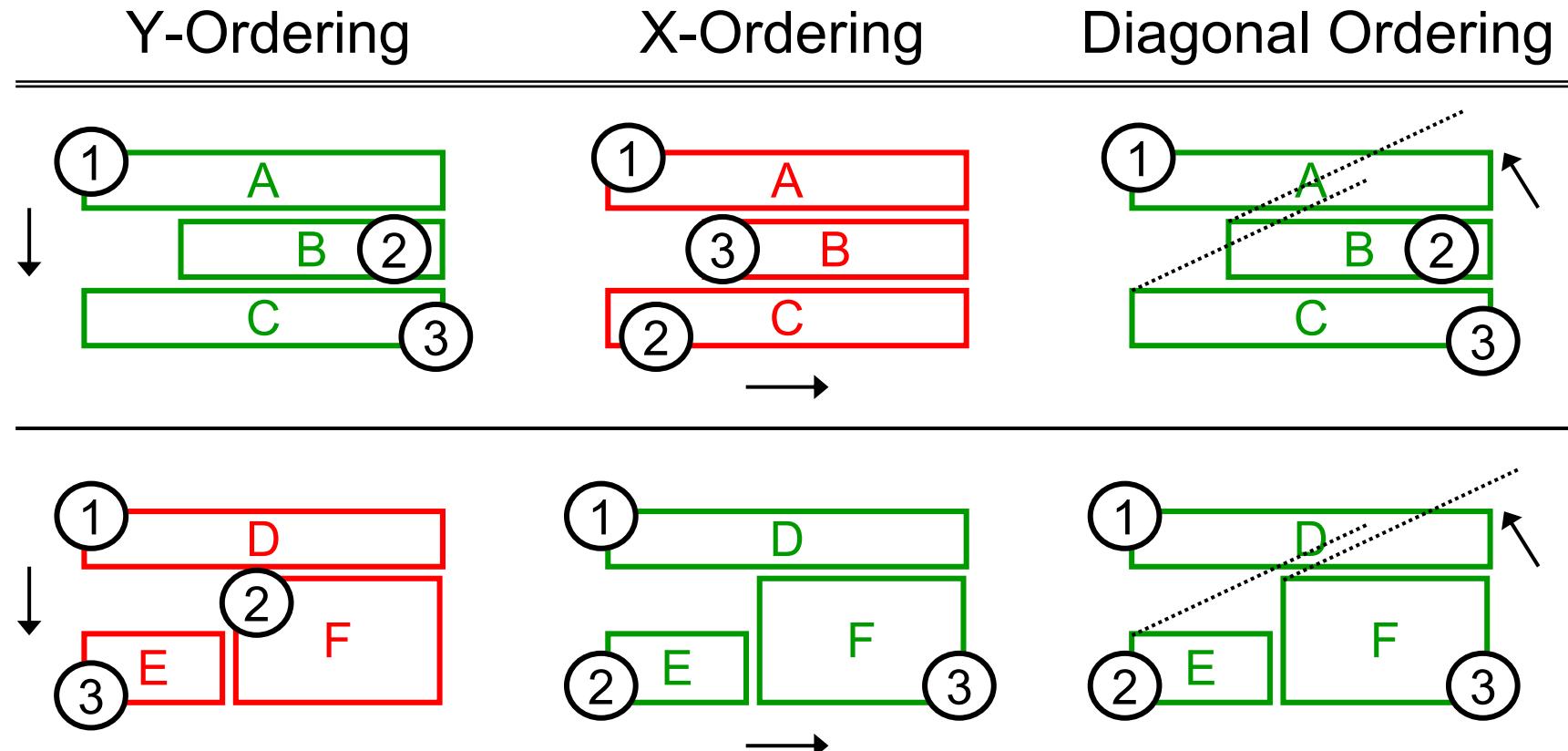
- 370 segments from web pages
- WEKA machine learning toolkit
- Feature reduction
- PART algorithm
 - C4.5 decision tree algorithm



Final Feature Set



Diagonal Ordering



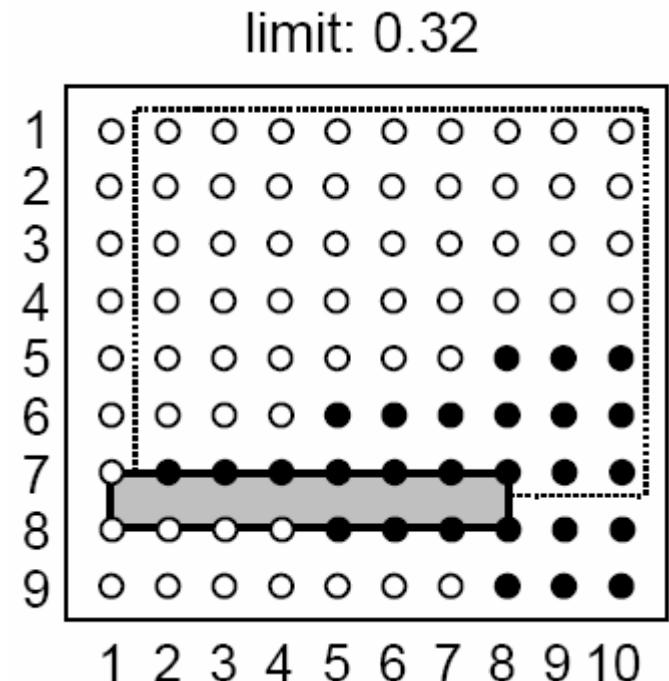
Diagonal Ordering Limit

Limit for the arctan between the two box corners:

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

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

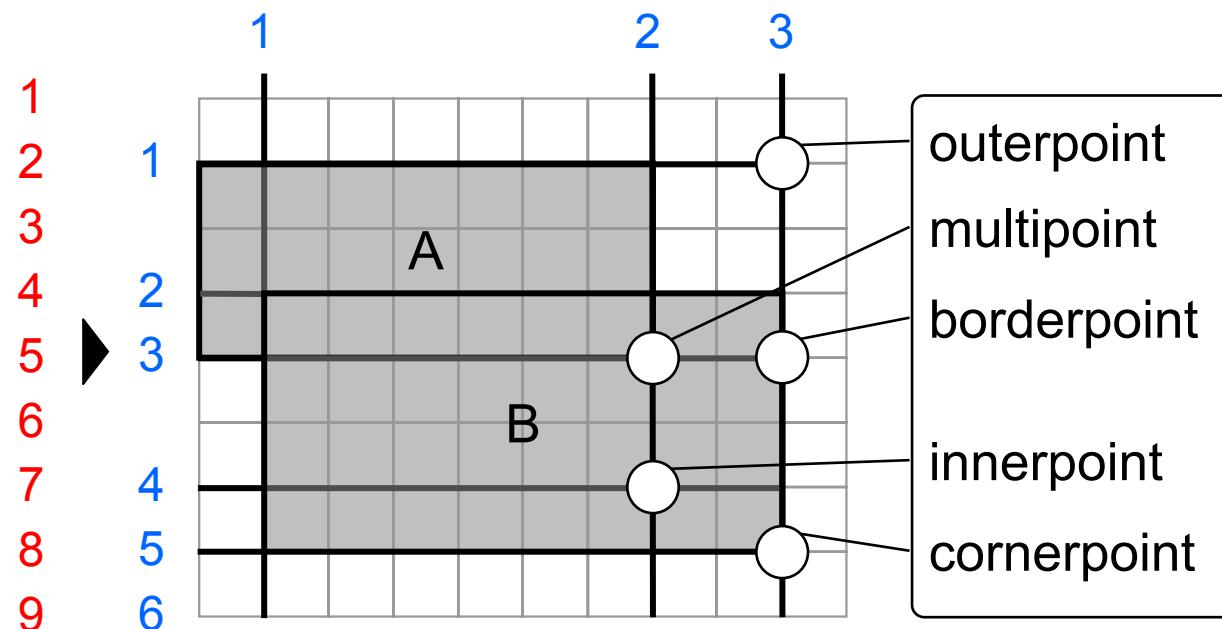
- Record start

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

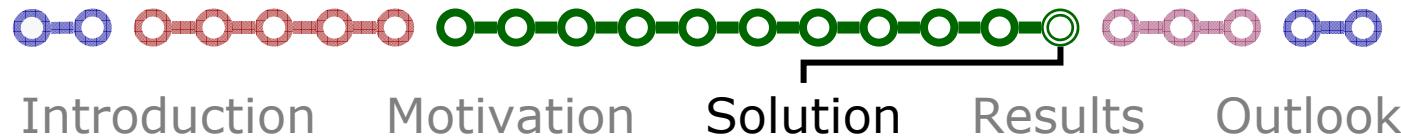
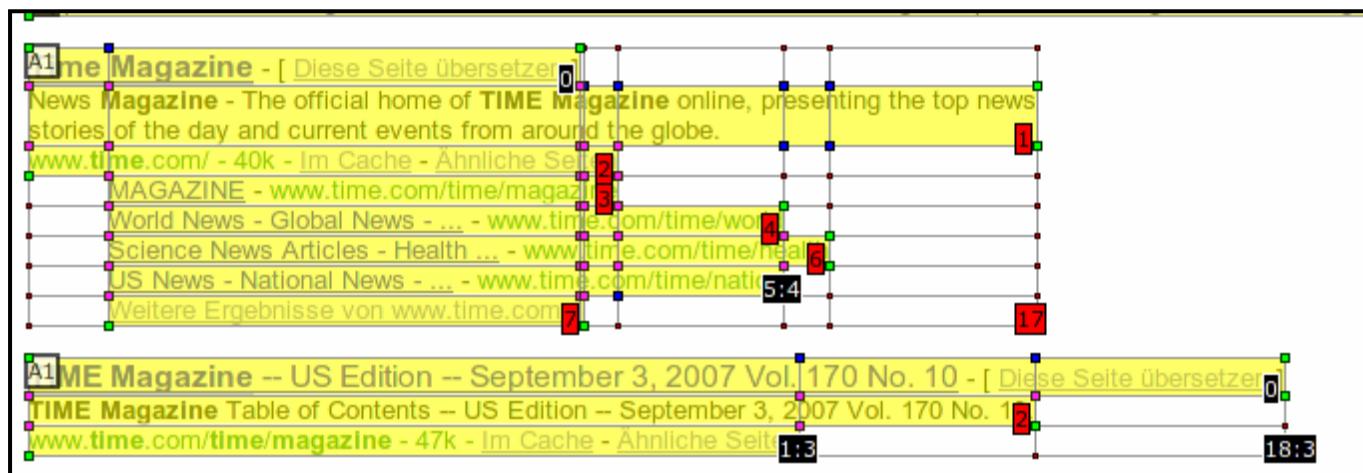
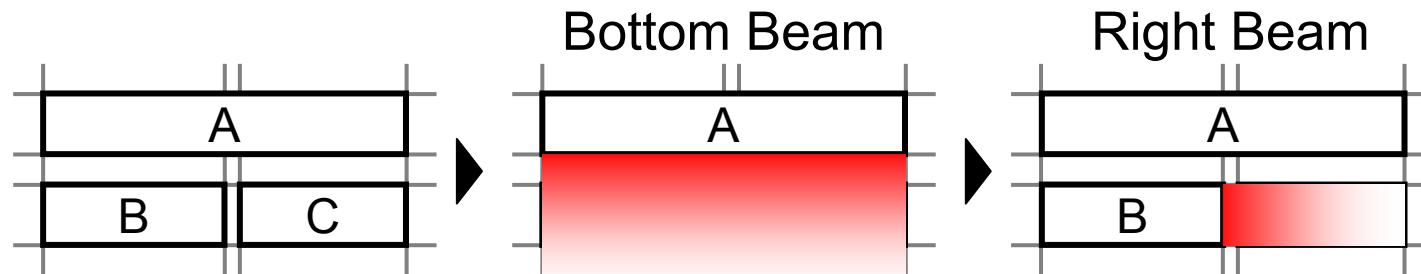


Multitopological Grid Concept

Minimal Grid: Screen Coordinates → Logical Coordinates



Multitopological Grid Example ⁷



Experimental Results

Web Pages: 85

Record Count: 1086

Correct: 836

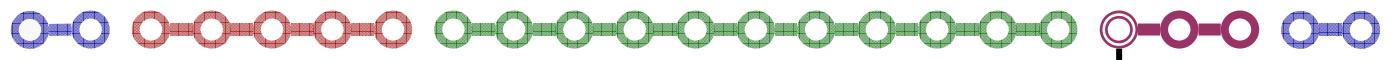
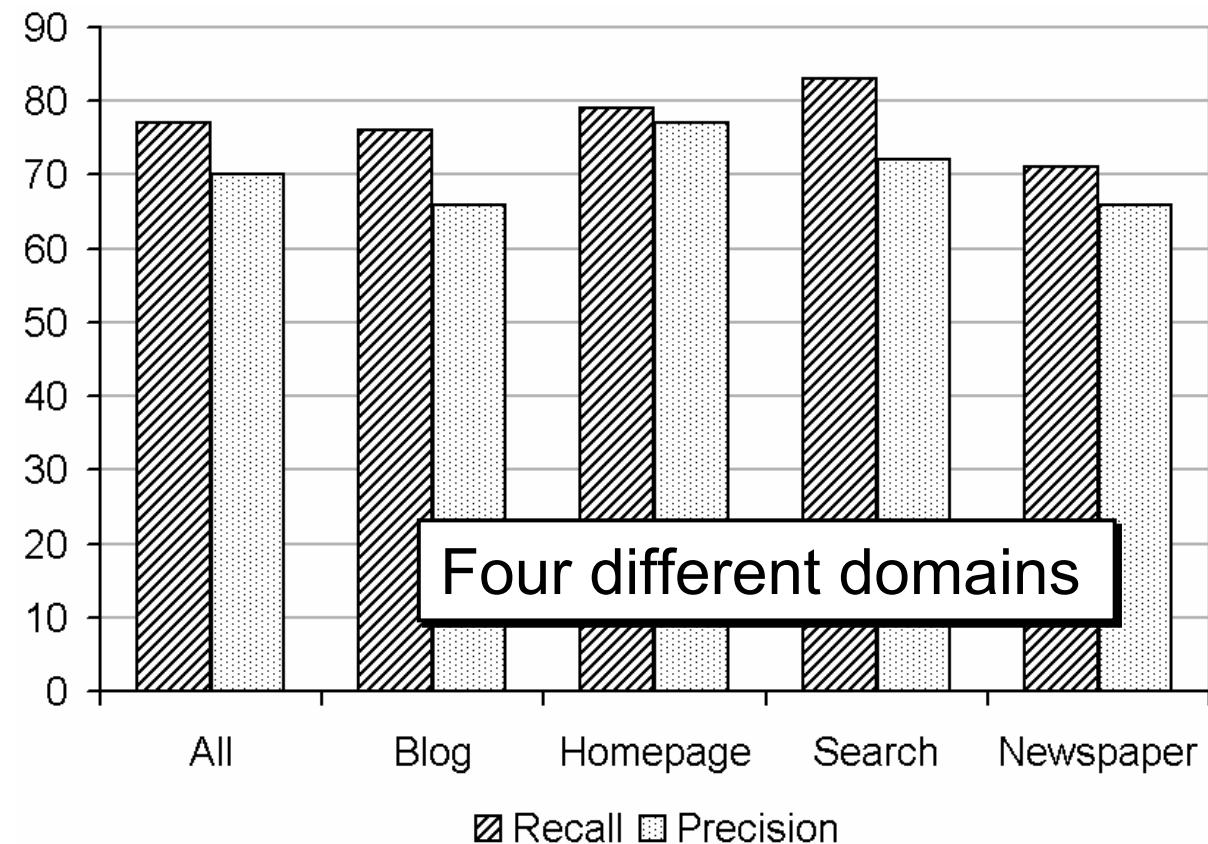
False Positives: 351

False Negatives: 241

Recall: 77%

Precision: 70%

F-Measure: 73%



Introduction

Motivation

Solution

Results

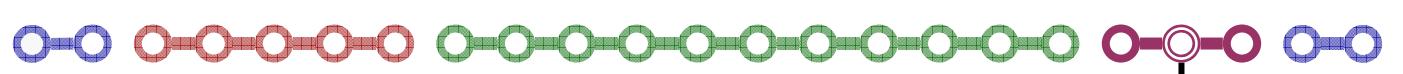
Outlook

Semantic (Domain) Dependence ⁸

Webpage



REDEVILA Result



Introduction

Motivation

Solution

Results

Outlook

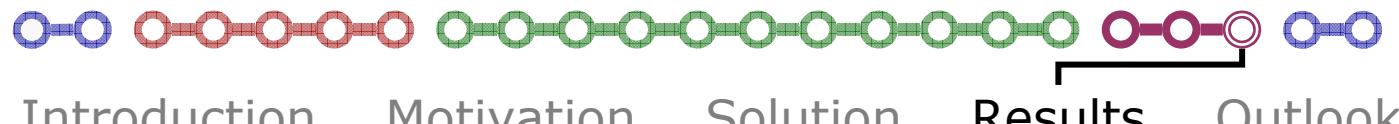
REDEVILA Example ⁹

Webpage

The screenshot shows a Google search results page. At the top is the Google logo and a search bar containing "new york times". Below the search bar are two tabs: "Web" and "News", with "Web" being selected. The first result is a sponsored link for "The New York Times Online" with the URL "www.nytimes.com". The second result is a regular search result for "The New York Times - Breaking News, World News & Multi". This result includes links for "Today's Paper", "World", "Sports", "Opinion", and "More results from nytimes.com".

REDEVILA Result

The screenshot shows the REDEVILA search results for "new york times". The first result is "The New York Times Online" with the URL "www.ny times .com". It is described as "Continuous coverage of news from arou". Below it is a "Sponsored Link". The second result is "The New York Times - Breaking News, World News & Multi" with the URL "www.ny times .com/". It is described as "Online edition of the newspaper's news and commentary. [R". Below it are several sub-links: "Similar pages", "Today's Paper - www.nytimes.com/pages/todayspaper/index.html", "World - www.nytimes.com/pages/world/", "Sports - www.nytimes.com/pages/sports/", "Opinion - select.nytimes.com/", and "More results from nytimes.com ». The third result is "Today's Paper - New York Times" with the URL "www.ny times .com/pages/todayspaper/index.html". It is described as "Use the Today's Paper page to see all the headlines from the". Below it is a "Similar pages" link.



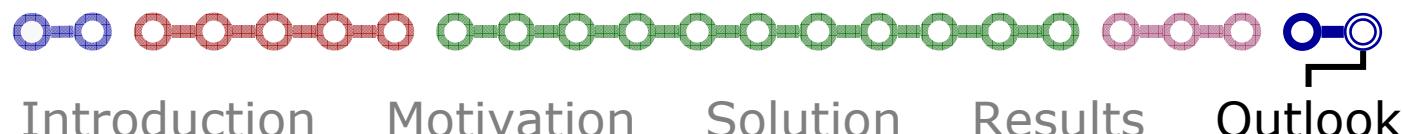
Conclusion

- 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

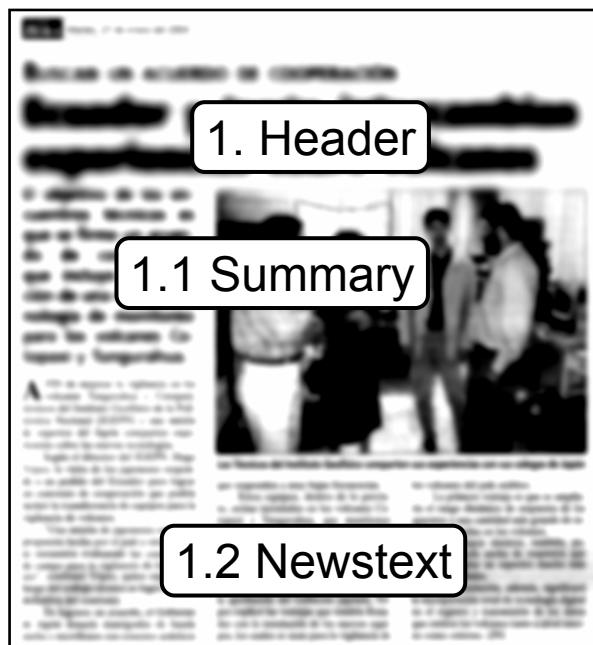
1. R. Baumgartner, S. Flesca, and G. Gottlob. Visual web information extraction with lixto. In VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases, pages 119–128, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
2. <http://www.bosai.go.jp/e/international>
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>

BACKUP

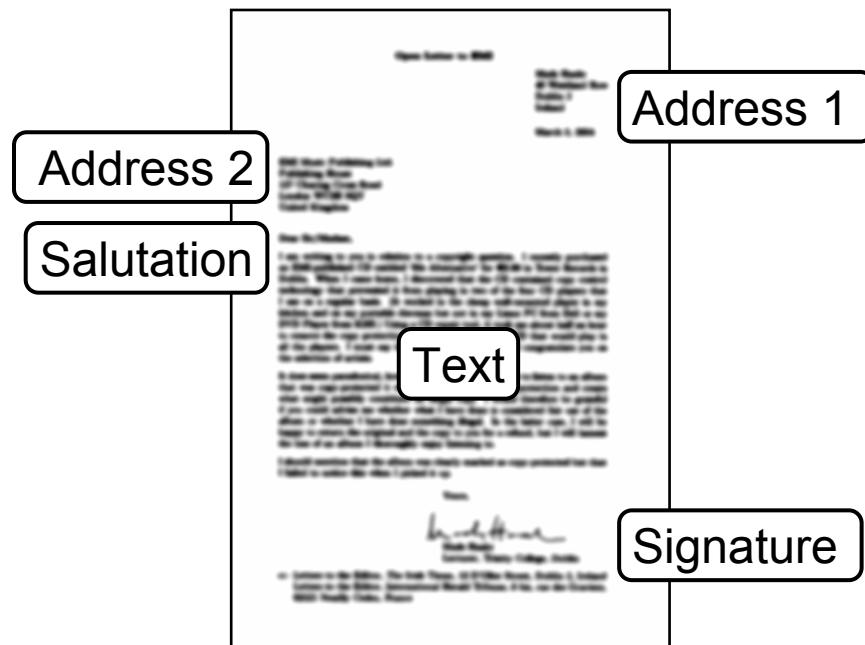
BACKUP

Domain Dependent Functional Semantics

NEWSPAPER



LETTER



Segmentation Example II

5) [Links](#)
13) [List Apart](#)
14) [easy Skillet](#)
15) [My Rosenow](#)
16) [vankyaAI](#)
17) [Fake Link One](#)
18) [Nothing Here](#)
19) [Links Nowhere](#)
20) [Fake Link Four](#)
21) [Fake Link](#)

4) Flanking Menus

With the popularity of three column layouts, this layout is bound to be useful to many. You may have seen this technique used at [dynamic ribbon device](#). In fact, this "flanking menus" technique was devised by BlueRobot for that site. Surprisingly, the technique has caused quite a bit of talk. The concept is simple: a content box with large margins is flanked by two additional (menu) boxes.

An important benefit of this technique is the order of elements in the HTML source. Here, the order is essentially content, menu one, menu two. For old browsers, text-only browsers, screen-readers, and many alternative devices, this means that the content is displayed before the menus. And, after all, most users visit a page for its content.

Known Issues

This layout fails in IE4.5/Mac. That browser has poor support for CSS absolute positioning, yet it recognizes and executes the CSS @import statement used to hide CSS from broken browsers. Currently, there is no known solution.

2) [Return to the Layout Reservoir](#) ::: [View the CSS](#)

Problem: "Small Line Above" Rule

Webpage

The screenshot shows the Alltheweb search results for 'BBC'. At the top, there's a search bar with 'advanced search :: customize pre' and a dropdown for 'BBC'. Below the search bar is a navigation menu with 'Web', 'News', 'Pictures', 'Video', and 'Audio' buttons. A blue banner displays '1 - 10 of 201,000,000 Results for BBC'. Underneath, a link 'Sponsor Results' is circled in red. Below it, there are two sponsored links: 'BBC America Web Special' and 'Shop Online with the BBC America Shop'. At the bottom, there's another link 'Bbc Television'.

REDEVILA Result

The screenshot shows the REDEVILA visual layout hierarchy for the website 'www.alltheweb.com'. It starts with 'Frame 0 - file: index.html' and 'Importance A'. Below that, it lists '1 - 10 of 201,000,000 Results for BBC'. The first result is 'BBC America Web Special - www.bbcamericanashop.com', which also has a 'Sponsor Results' link circled in red. Other results include 'Shop Online with the BBC America Shop - www.shoptogo.com' and 'Bbc Television - shopping.yahoo.com'.

REDEVILA Example I

Webpage

[Home](#)

Unveiled: radical prescription for our health crisis

Obesity, alcohol abuse, smoking: Britain is among the most unhealthy countries in Europe. Now a pioneering NHS adviser is proposing a revolutionary cure for our ills

- British people are the fattest in Europe, says Government report

WORLD NEWS

- China identifies Xi Jinping as the next party leader

UK NEWS

- UK population 2031 **NEW**
- Brown assai betrayal ove
- Russian shor art and polit
- Plan to cull d dismay by a
- The engineer mistaken for
- Rise in crime offenders'

REDEVILA Result

Visual layout hierarchy of [www_indep](#)

Frame 0 - file: index.html

Importance A

Unveiled: radical prescription for our health crisis

Home
Obesity, alcohol abuse, smoking: Britain is among the most unhealthy countries in Europe. Now a pioneering NHS adviser is proposing a revolutionary cure for our ills
British people are the fattest in Europe, says Government report

WORLD NEWS

China identifies Xi Jinping as the next party leader
Israel accused after 30 injured in prison riot
Joaquim Chissan: Democrat among the claims
Claims of Maori separatist plot begin to