



ProFoUnd: Program-analysis–based Form Understanding

(joint work with M. Benedikt,
T. Furche, A. Savvides)

PIERRE SENELLART





The Deep Web

Definition (Deep Web, Hidden Web, Invisible Web)

All the content on the Web that is not directly accessible through **hyperlinks**. In particular: HTML forms, Web services.



Size estimate: 500 times more content than on the **surface Web!**
[BrightPlanet, 2001]. Hundreds of thousands of deep Web databases
[Chang et al., 2004]



Example

- *Yellow Pages* and other directories;
- Library catalogs;
- Weather services;
- US Census Bureau data;
- etc.

Discovering Knowledge from the Deep Web

[Varde et al., 2009]

- Content of the deep Web hidden to classical Web search engines (they just follow links)
- But very valuable and high quality!
- Even services allowing access through the surface Web (e.g., e-commerce) have more semantics when accessed from the deep Web
- How to **benefit** from this information?

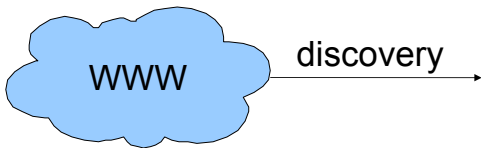
Focus here: Automatic, unsupervised, methods



Notes on the Extensional Approach

- Main issues:
 - Discovering services
 - Choosing appropriate data to submit forms
 - Use of data found in result pages to bootstrap the siphoning process
 - Ensure good coverage of the database
- Approach **favored by Google**, used in production [Madhavan et al., 2006]
- Not always feasible (huge load on Web servers)

Intensional Approach



Google Scholar **Advanced Scholar Search** [Advanced Search Tips](#) | [About Google Scholar](#)

Find articles with all of the words 10 results | Search Scholar

with the **exact phrase**

with **at least one** of the words

without the words

where my words occur anywhere in the article

Author Return articles written by
e.g., "P.J. Hayes" or McCarthy

Publication Return articles published in
e.g., J Biol Chem or Nature

Date Return articles published between -
e.g., 1996

probing

Google Scholar data Search [Advanced Scholar Search](#) [Database Information](#) [Help](#)

Scholar All articles **Recent Articles** Results 1 - 19 of about 91,669,000 for data [Database](#) (0.18 seconds)

1. Fisher D. The use of multiple measurements in economic problems. [View Article](#)

2. Psychol. 400. Sereno, SA. Grant, W. H. 1986. [View Article](#)

... Calhoun A, Pines D, Conroy E, Carter T, Higgins D. Between-group analysis of increasing data. [View Article](#)

3. Gao L. 2005. [View Article](#)

The protein kinase encoded by the *AI1* proto-oncogene is a target of the PDGF-activated... [View Article](#)

TF PRINZEL, SUNG-LI YANG, TO CHAN H, BATA A, KALLAUSSKAS, DR MORRISON, DR KAPLAN, PHILIPSONIS, Cell. [View Article](#)

4. Gao L. 2005. [View Article](#)

RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J recombination. [View Article](#)

FD. [View Article](#)

5. Gao L. 2005. [View Article](#)

Both genetic and biochemical data point toward a physiological role for the complex as the above begin-opening activity in "E2" recombination. [View Article](#)

6. Gao L. 2005. [View Article](#)

Random data analysis and measurement procedures. [View Article](#)

7. Gao L. 2005. [View Article](#)

8. Gao L. 2005. [View Article](#)

9. Gao L. 2005. [View Article](#)

10. Gao L. 2005. [View Article](#)

11. Gao L. 2005. [View Article](#)

12. Gao L. 2005. [View Article](#)

13. Gao L. 2005. [View Article](#)

14. Gao L. 2005. [View Article](#)

15. Gao L. 2005. [View Article](#)

16. Gao L. 2005. [View Article](#)

17. Gao L. 2005. [View Article](#)

18. Gao L. 2005. [View Article](#)

19. Gao L. 2005. [View Article](#)

Form wrapped as a Web service

analyzing

query





Notes on the Intensional Approach

- More **ambitious** [Chang et al., 2005, Senellart et al., 2008]
- Main issues:
 - Discovering services
 - Understanding the structure and semantics of a form
 - Understanding the structure and semantics of result pages
 - Semantic analysis of the service as a whole
- No significant load imposed on Web servers



Outline

Introduction

Wrapping Web Forms

Form Analysis

Information Extraction from Deep Web Pages

ProFoUnd

Conclusions



Outline

Introduction

Wrapping Web Forms

Form Analysis

Information Extraction from Deep Web Pages

ProFoUnd

Conclusions



Forms

Analyzing the **structure** of HTML forms.

Authors	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	
Title	<input type="text"/>	Year	<input type="text"/>	Page	<input type="text"/>
Conference	<input type="text"/>	ID	<input type="text"/>		
Journal	<input type="text"/>	Volume	<input type="text"/>	Number	<input type="text"/>
<input type="button" value="Search"/>	<input type="button" value="Reset"/>	Maximum of <input type="text" value="100"/> matches			

Goal

Associating to each form field the appropriate **domain concept**.



1st Step: Structural Analysis

1. Build a **context** for each field:
 - label tag;
 - id and name attributes;
 - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



1st Step: Structural Analysis

1. Build a **context** for each field:
 - label tag;
 - id and name attributes;
 - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



1st Step: Structural Analysis

1. Build a **context** for each field:
 - label tag;
 - id and name attributes;
 - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



1st Step: Structural Analysis

1. Build a **context** for each field:
 - label tag;
 - id and name attributes;
 - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



2nd Step: Confirm Annotations w/ Probing

For each field annotated with a concept c :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



2nd Step: Confirm Annotations w/ Probing

For each field annotated with a concept c :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



2nd Step: Confirm Annotations w/ Probing

For each field annotated with a concept c :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



2nd Step: Confirm Annotations w/ Probing

For each field annotated with a concept c :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of c (chosen representatively of the frequency distribution of c).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



How well does this work?

- **Good results** in practice [Senellart et al., 2008]

	Initial annot.		Confirmed annot.	
	$p(\%)$	$r(\%)$	$p(\%)$	$r(\%)$
Average	49	73	82	73

- Probing raises precision **without hurting recall**
- Clustering according to **DOM paths**: much better than previous approaches
- But some critical assumptions:
 - It is possible to query a field with a **subword**
 - All form fields are **independent**
 - No field is **required**



How well does this work?

- **Good results** in practice [Senellart et al., 2008]

	Initial annot.		Confirmed annot.	
	$p(\%)$	$r(\%)$	$p(\%)$	$r(\%)$
Average	49	73	82	73

- Probing raises precision **without hurting recall**
- Clustering according to **DOM paths**: much better than previous approaches
- But some critical assumptions:
 - It is possible to query a field with a **subword**
 - All form fields are **independent**
 - No field is **required**



Outline

Introduction

Wrapping Web Forms

Form Analysis

Information Extraction from Deep Web Pages

ProFoUnd

Conclusions

Result Pages

Pages resulting from a given form submission:

- share the **same structure**;
- set of **records** with fields;
- **unknown** presentation!

Find: remi gilleron Documents
Citations

Searching for PHRASE remi gilleron.
Restrict to: [Leader](#) [Title](#) Order by: [Expected citations](#) [Title](#) [Usage](#) [Data](#) Try: [Google](#) [CiteSeer](#) [Google Books](#) [Yahoo!](#) [MSN](#) [EML](#) [DBLP](#)
7 documents found. Order: number of citations.

[PAC Learning under Helpful Distributions - Denis, Gilleron \(1997\)](#) [Correct]
[110 citations](#)
Helpful Distributions y Francois Denis, Remi Gilleron LFL, LRA 369 CNRS, Université de Lille 1 59655
1 59655 Villeneuve d'Ascq FRANCE e-mail: denis.gilleron@lfl.fr Abstract A PAC model under helpful
on Algorithmic Learning Theory ALT'97 (Denis and Gilleron, 1997)Introduction it seems that many

Sort by	Sort by	Sort by
Relevance	Title	Year
● 81%	Grindhouse Director Screenwriter Producer	2007
● N/A	Death Proof Director	2007
● 59%	Hostel Executive Producer	2006
● N/A	Reservoir Dogs/Bad Lieutenant Director	2006
● N/A	Inglorious Bastards Director	2006
● 97%	Double Dare Featured	2005
● 78%	Sin City Additional Directing	2005
● 29%	The Muppets: Wizard of Oz Star	2005
● 0%	Dalry Calhoun Executive Producer	2005
● 85%	Kill Bill Vol. 2 Director Screenwriter	2004
● 100%	2 Channel: A Magnificent Obsession Featured	2004
● 85%	Kill Bill Vol. 1 Director Screenwriter Producer	2003

Goal

Building **wrappers** for a given kind of result pages, in a fully automatic, **unsupervised**, way.

Simplification: restriction to a domain of interest, with some **domain knowledge**.



Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for **all:xml**

1. cs.LO/0601085 [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Riccardo Pucella**, **Vicky Weissman**

Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004

Subj-class: Logic in Computer Science: Cryptography and Security

ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :

Title: VOFfilter, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Chen-zhou Cui** (1), **Markus Dotensky** (2), **Peter Quinn** (2), **Yong-heng Zhao** (1), **Francoise Genova** (3) ((1)NAO China, (2) ESO, (3) CDS)

Comments: Accepted for publication in ChJA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :

Title: Matching Subsequences in Trees

Authors: **Philip Bille**, **Inge Li Goertz**

Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: **Thierry Despeyroux** (INRIA Rocquencourt / INRIA Sophia Antipolis)

Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouarain** (INRIA Rocquencourt), **Cosmin Creangescu** (INRIA Rocquencourt), **François Dang Ngoc** (INRIA Rocquencourt, PRISM - UVSQ),

Nicolas Dieu (INRIA Rocquencourt), **Philippe Pucheral** (INRIA Rocquencourt, PRISM - UVSQ)

Subj-class: Cryptography and Security: Databases

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.



Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for **all:xml**

1. cs.LO/0601085 [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Riccardo Pucella**, **Vicky Weissman**

Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004

Subj-class: Logic in Computer Science: Cryptography and Security

ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :

Title: VOFfilter, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Chen-zhou Cui** (1), **Markus Dotensky** (2), **Peter Quinn** (2), **Yong-heng Zhao** (1), **Francoise Genova** (3) ((1)NAO China, (2) ESO, (3) CDS)

Comments: Accepted for publication in ChJA (9 pages, 2 figures, 185KB)

3. cs.DS/0512061 [abs, ps, pdf, other] :

Title: Matching Subsequences in Trees

Authors: **Philip Bille**, **Inge Li Goertz**

Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: **Thierry Despeyroux** (INRIA Rocquencourt / INRIA Sophia Antipolis)

Subj-class: Information Retrieval

5. cs.CR/0510013 [abs, pdf] :

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouarain** (INRIA Rocquencourt), **Cosmin Creangescu** (INRIA Rocquencourt), **François Dang Ngoc** (INRIA Rocquencourt, PRISM - UVSQ),

Nicolas Dieu (INRIA Rocquencourt), **Philippe Pucheral** (INRIA Rocquencourt, PRISM - UVSQ)

Subj-class: Cryptography and Security: Databases

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.



Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for **all:xml**

1. **cs.LO/0601085** [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Riccardo Pucella**, **Vicky Weissman**

Comments: 32 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security) 2006

Subj-class: **Logic in Computer Science**; **Cryptography and Security**

ACM-class: H.2.7; K.4.4

2. **astro-ph/0512493** [abs, pdf] :

Title: VOFiler, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Chen-zhou Cui** (1), **Markus Dolansky** (2), **Peter Quinn** (2), **Yong-heng Zhao** (1), **Francoise Genov** (3) ((1)NAO China, (2) **ESO**, (3) CDS)

Comments: Accepted for publication in CHAA (9 pages, 2 figures, 185KB)

3. **cs.DS/0512061** [abs, ps, pdf, other] :

Title: **Matching Subsequences in Trees**

Authors: **Philip Bille**, **Inge LI Goertz**

Subj-class: **Data Structures and Algorithms**

4. **cs.IR/0510025** [abs, ps, pdf, other] :

Title: **Practical Semantic Analysis of Web Sites and Documents**

Authors: **Etienne Despres** (**IRISA**, **IRISA**), **Yves-Michel Buge** (**IRISA**, **IRISA**)

Subj-class: **Information Retrieval**

5. **cs.CR/0510013** [abs, pdf] :

Title: **Safe Data Sharing and Data Dissemination on Smart Devices**

Authors: **Luc Bouganim** (**IRISA**, **IRISA**), **Cosmin Cremareanu** (**IRISA**, **IRISA**), **François Dang Ngoc** (**IRISA**, **IRISA**), **PRISM - UVSQ**,

Nicolas Tiedj (**IRISA**, **IRISA**), **Philippe Buchera** (**IRISA**, **IRISA**), **PRISM - UVSQ**

Subj-class: **Cryptography and Security**; **Databases**

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.



Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for **all:xml**

1. **cs.LO/0601085** [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Riccardo Pucella**, **Vicky Weissman**

Comments: 32 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security) 2006

Subj-class: **Logic in Computer Science**; **Cryptography and Security**

ACM-class: H.2.7; K.4.4

2. **astro-ph/0512493** [abs, pdf] :

Title: VOFitter, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Chen-zhou Cui** (1), **Markus Dolansky** (2), **Peter Quinn** (2), **Yong-heng Zhao** (1), **Francoise Genov** (3) ((1)NAO China, (2) **ESO**, (3) CDS)

Comments: Accepted for publication in CHAA (9 pages, 2 figures, 185KB)

3. **cs.DS/0512061** [abs, ps, pdf, other] :

Title: Matching Subsequences in Trees

Authors: **Philippe Billè**, **Inge LI Goertz**

Subj-class: **Data Structures and Algorithms**

4. **cs.IR/0510025** [abs, ps, pdf, other] :

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: **Thierry Desjardins** (**IRISA**, **IRISA**), **Yves-Alexandre Grandjean** (**IRISA**, **IRISA**)

Subj-class: **Information Retrieval**

5. **cs.CR/0510013** [abs, pdf] :

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouganim** (**IRISA**, **IRISA**), **Cosmin Cremene** (**IRISA**, **IRISA**), **François Dang Ngoc** (**IRISA**, **IRISA**), **PRISM - UVSQ**,

Nicolas Tiedj (**IRISA**, **IRISA**), **Philippe Buchera** (**IRISA**, **IRISA**), **PRISM - UVSQ**

Subj-class: **Cryptography and Security**; **Databases**

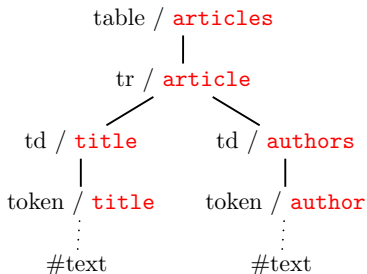
Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.

Both **incomplete** and **imprecise**!

Unsupervised Wrapper Induction

- Use the pre-annotation as the input of a structural supervised machine learning process.
- Purpose: remove outliers, generalize incomplete annotations.





How well does this work?

- Good, but not great, results [Senellart et al., 2008]

	Title		Author		Date	
	F_g	F_x	F_g	F_x	F_g	F_x
Average	44	63	64	70	85	76

- F_g : F -measure (%) of the annotation by the gazetteer.
 - F_x : F -measure (%) of the annotation by the induced wrapper.
- **Main issue:** the machine learning assumes that the initial annotation is really the reference one



Outline

Introduction

Wrapping Web Forms

ProFoUnd

JavaScript and the Deep Web

Form Understanding through JavaScript Analysis

Conclusions



Outline

Introduction

Wrapping Web Forms

ProFoUnd

JavaScript and the Deep Web

Form Understanding through JavaScript Analysis

Conclusions



Better Form Analysis

What

Where

eg. Restaurants
Hairdressers
Telstra
Apple Stores




Better Form Analysis


What

Find

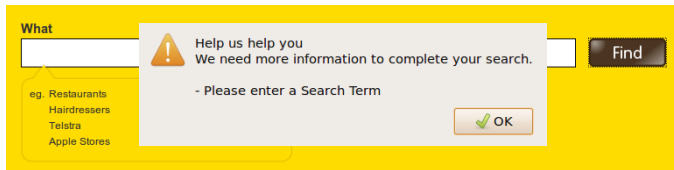
eg. Restaurants
Hairdressers
Telstra
Apple Stores

 Help us help you
We need more information to complete your search.

- Please enter a Search Term



Better Form Analysis



The screenshot shows a search interface on a yellow background. At the top left, the word "What" is displayed above a search input field. Below the input field, there is a list of suggestions: "eg. Restaurants", "Hairdressers", "Telstra", and "Apple Stores". To the right of the input field is a "Find" button. A modal alert box is centered over the form, containing a warning icon, the text "Help us help you We need more information to complete your search.", and a sub-message "- Please enter a Search Term". An "OK" button with a green checkmark is located at the bottom right of the alert box.

```
// Do not submit unless form is valid
$j("#searchForm").submit(function(event) {
    $j("#searchFormLocationClue").val($j("#searchFormLocationClue").val().trim());
    if ($j("#searchFormBusinessClue").val().isEmpty()) {
        alert('Help us help you\nWe need more information to
            complete your search.\n\n- Please enter a Search Term');
        return false;
    } else {
        return true;
    }
});
```



JavaScript: the Data Language of the Web

- Lots of JavaScript code on the Web (source is always available!)
- Lots of information can be gained by **static analysis** of this code:
 - **Required** fields
 - **Dependencies** between fields (if x is filled in, so should be y ; the value of x should be less than that of y ; etc.)
 - **Datatype** of each fields (regular expressions, numeric types, dates, etc.)
- Is this feasible in practice?



Outline

Introduction

Wrapping Web Forms

ProFoUnd

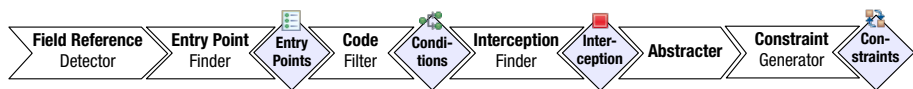
JavaScript and the Deep Web

Form Understanding through JavaScript Analysis

Conclusions



ProFoUnd architecture



- **Entry points** are HTML event attributes, setting of event handlers in code, etc. (event: *click* on a submit button, *submit* on a form)
- **Conditions** are (in)equality tests on form field values (possibly aliased)
- **Interceptions** are interruptions of the form submission process (error messages, simple return `false`; in event handler, etc.)



Abstracting the code

- Rice's theorem: no hope in a **sound** and **complete** constraint finder
- But that's ok! **Anything** that we can learn is more than what we have at the moment.
- **Coarse abstraction** of the JS code:
 - Only conditions on the code flow from entry points to interceptions are considered.
 - We consider only a simple subset of the JS language; anything beyond that is ignored.
 - Side-effects are mostly ignored
- As a consequence: no guarantee of either soundness or completeness \Rightarrow **only experimental guarantees**



Engineering issues to deal with

- Extracting a Web form model: **DIADEM**'s tools <http://www.diadem-project.info/>
- Parsing JavaScript: **Mozilla Rhino** (but see later)
- JavaScript frameworks: **ad-hoc support** for most popular ones (jQuery, Prototype, ASP.NET generated code, YUI, Dojo, MooTools)
- Evaluating JavaScript code (e.g., to determine what a jQuery selector (`$("#form#lookup .product")`) returns): **Mozilla JS engine**
- Abstraction, alias references, etc.: ProFoUnd core, **developed from scratch**

ProFoUnd interface [Benedikt et al., 2012]

The screenshot displays the ProFoUnd interface with three numbered callouts:

- 1**: Points to the browser window showing a web page with a "Keyword:" field and a "Price (USD)" section. The "Minimum:" field contains "0" (highlighted in green) and the "Maximum:" field contains "10000" (highlighted in blue). A "Submit Query" button is visible below.
- 2**: Points to the "Constraints" panel on the left, which lists several constraints. The selected constraint is "CONSTRAINT: LE" with conditions "on <INPUT id=min>" and "on <INPUT id=max>".
- 3**: Points to the "Javascript / Event handler" panel at the bottom, which contains the following code:

```
if (min > max)
{
  $(".label_price").css("color", "red");
  return false;
}
```

1. Web page view, with fields highlighted
2. Constraints found:
 $min < max, max \neq 0,$
 $product \neq "$
3. JS fragment for the highlighted constraint



Preliminary evaluation

- 70 real-estate websites containing search forms
- 30 out of 70 use client-side validation, with a total of 35 constraints
- **100% precision**: all identified constraints are correct
- **63% recall**: 22 out of 35 JS-enforced constraints were found
- Why did we miss some?
 - Use of complex JavaScript features, such as `eval`
 - Code obfuscation by introducing extra layers of computation
 - Limitations of the abstracter – work in progress!



Outline

Introduction

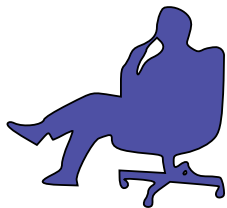
Wrapping Web Forms

ProFoUnd

Conclusions



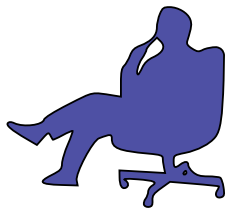
- Exploiting data from the deep Web in an automatic manner:
non-trivial, largely open problem
- Classical techniques exploit both **domain knowledge** and the structure of **forms and result pages**
- Possible to get very precise information about the behavior of Web forms by **static analysis of client-side code**



- Use a **real JS parser** (Rhino has lots of limitations); trying with SpiderMonkey, Mozilla's JS engine
- **Large-scale evaluation**, application to deep Web crawling
- **Type inference** for form fields: regular expressions, simple datatypes
- Combining with **dynamic analysis**
- **Type inference** for AJAX applications: static analysis of AJAX calls to determine input and output types (possibly JSON or XML types)

PhD Opportunity

PhD scholarship on this topic at U. Oxford, looking for excellent candidates!



- Use a **real JS parser** (Rhino has lots of limitations); trying with SpiderMonkey, Mozilla's JS engine
- **Large-scale evaluation**, application to deep Web crawling
- **Type inference** for form fields: regular expressions, simple datatypes
- Combining with **dynamic analysis**
- **Type inference** for AJAX applications: static analysis of AJAX calls to determine input and output types (possibly JSON or XML types)

PhD Opportunity

PhD scholarship on this topic at U. Oxford, looking for excellent candidates!



Merci.

Michael Benedikt, Tim Furche, Andreas Savvides, and Pierre Senellart. ProFoUnd: Program-analysis-based form understanding. In *Proc. WWW*, Lyon, France, April 2012. Demonstration.

BrightPlanet. The deep Web: Surfacing hidden value. White Paper, July 2001.

Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang. Structured databases on the Web: Observations and implications. *SIGMOD Record*, 33(3):61–70, September 2004.

Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the Web. In *Proc. CIDR*, Asilomar, USA, January 2005.

Jayant Madhavan, Alon Y. Halevy, Shirley Cohen, Xin Dong, Shawn R. Jeffery, David Ko, and Cong Yu. Structured data meets the Web: A few observations. *IEEE Data Engineering Bulletin*, 29 (4):19–26, December 2006.

Pierre Senellart, Avin Mittal, Daniel Muschick, Rémi Gilleron, and Marc Tommasi. Automatic wrapper induction from hidden-Web sources with domain knowledge. In *Proc. WIDM*, pages 9–16, Napa, USA, October 2008.

Aparna Varde, Fabian M. Suchanek, Richi Nayak, and Pierre Senellart. Knowledge discovery over the deep Web, semantic Web and XML. In *Proc. DASFAA*, pages 784–788, Brisbane, Australia, April 2009. Tutorial.