ProFoUnd:
Program-analysis-based Form Understanding

Andreas Savvides
(work done with M. Benedikt,
T.Furche and P. Senellart)
Looking deep under the surface

• Surface Web:
  – Bread & butter of traditional search engines
  – All about hyperlinks

• Deep Web:
  – Search engines struggle at surfacing hidden content
  – Plethora of valuable information behind Web services and HTML forms
  – Valuable data for information extraction systems, e.g. DIADEM
Dealing with a deep Web search interface

1. Find relevant websites

2. Identify a search interface

3. Deduce input fields of the search interface and corresponding meta-data

4. Query a hidden database
JavaScript and the Deep Web

- Client-side **integrity constraints** enforced using JavaScript

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

The first system to provide **integrity** constraint identification for deep Web search interfaces based on JavaScript analysis.
Evaluation & Moving Forward

• 70 randomly selected real-estate websites
  – 100% precision and 63% recall

• Where did we fall short?
  – eval, obfuscation, system limitations

• What about…
  – Runtime execution?
  – Ajax?
  – Server-side constraints?
Demonstration?

Come find us at our demo booth all day long on Friday for a showcase of the system or simply to learn more!