Technology and Application of Big Data

Qing LIAO(廖清)
School of Computer Science and Technology
HIT

Course Details

- Instructor:
  - Qing LIAO, liaoqing@hit.edu.cn
  - Rm. 303B, Building C
  - Office hours: by appointment
- Course web site:
  - liaoqing.me
- Reference books/materials:
  - Big data courses from University of California
  - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
  - Papers
- Grading Scheme:
  - Paper Report 30%
  - Final Exam 70%

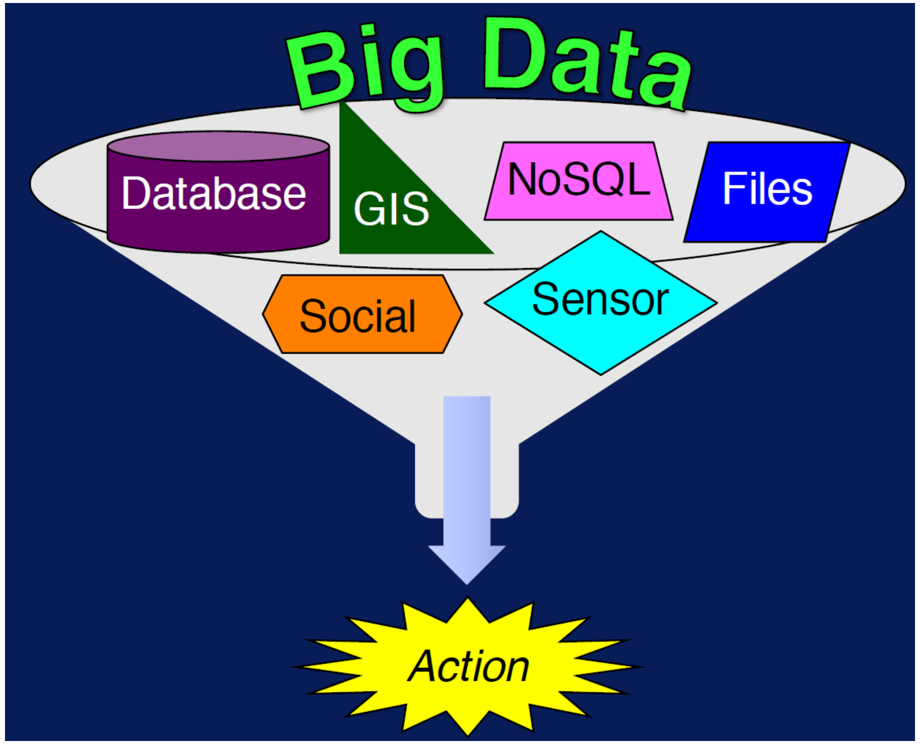
What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data
- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning would help
    - But not necessary

Previous Section

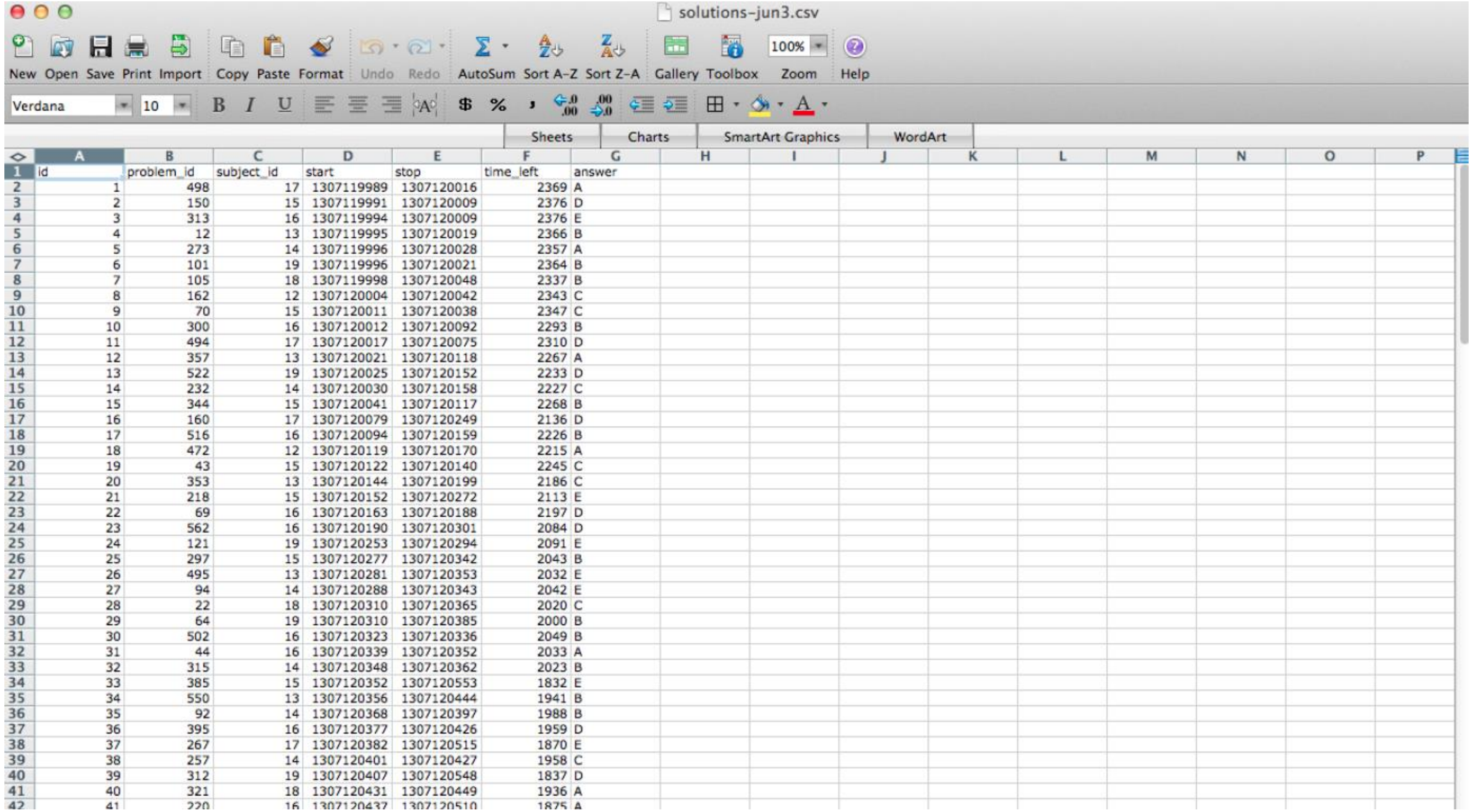
- How to Get Value from Big Data

Acquire | Prepare | Analyze | Report | ACT



Acquire Data - Information Extraction

What you wish data looked like



id	problem_id	subject_id	start	stop	time_left	answer
1	498	17	1307119989	1307120016	2369	A
2	150	15	1307119991	1307120009	2376	D
3	313	16	1307119994	1307120009	2376	E
4	12	13	1307119995	1307120019	2366	B
5	273	14	1307119996	1307120028	2357	A
6	101	19	1307119996	1307120021	2364	B
7	105	18	1307119998	1307120048	2337	B
8	162	12	1307120004	1307120042	2343	C
9	70	15	1307120011	1307120038	2347	C
10	300	16	1307120012	1307120092	2293	B
11	494	17	1307120017	1307120075	2310	D
12	357	13	1307120021	1307120118	2267	A
13	522	19	1307120025	1307120152	2233	D
14	232	14	1307120030	1307120158	2227	C
15	344	15	1307120041	1307120117	2268	B
16	160	17	1307120079	1307120249	2136	D
17	516	16	1307120094	1307120159	2226	B
18	472	12	1307120119	1307120170	2215	A
19	43	15	1307120122	1307120140	2245	C
20	353	13	1307120144	1307120199	2186	C
21	218	15	1307120152	1307120272	2113	E
22	69	16	1307120163	1307120188	2197	D
23	562	16	1307120190	1307120301	2084	D
24	121	19	1307120253	1307120294	2091	E
25	297	15	1307120277	1307120342	2043	B
26	495	13	1307120281	1307120353	2032	E
27	94	14	1307120288	1307120343	2042	E
28	22	18	1307120310	1307120365	2020	C
29	64	19	1307120310	1307120385	2000	B
30	502	16	1307120323	1307120336	2049	B
31	44	16	1307120339	1307120352	2033	A
32	315	14	1307120348	1307120362	2023	B
33	385	15	1307120352	1307120553	1832	E
34	550	13	1307120356	1307120444	1941	B
35	92	14	1307120368	1307120397	1988	B
36	395	16	1307120377	1307120426	1959	D
37	267	17	1307120382	1307120515	1870	E
38	257	14	1307120401	1307120427	1958	C
39	312	19	1307120407	1307120548	1837	D
40	321	18	1307120431	1307120449	1936	A
41	220	16	1307120437	1307120510	1875	A

Acquire Data - Information Extraction

What dose data really looked like

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```
```

Acquire Data - Information Extraction

- Information Extraction (IE)

  - Transforming unstructured textual information into structured information via extraction rules.

- A set of extraction rules suitable to extract information from a Web site is called a Wrapper

  - Wrapper induction (supervised learning)
  - Automatic extraction (unsupervised learning)

Acquire Data - Information Extraction

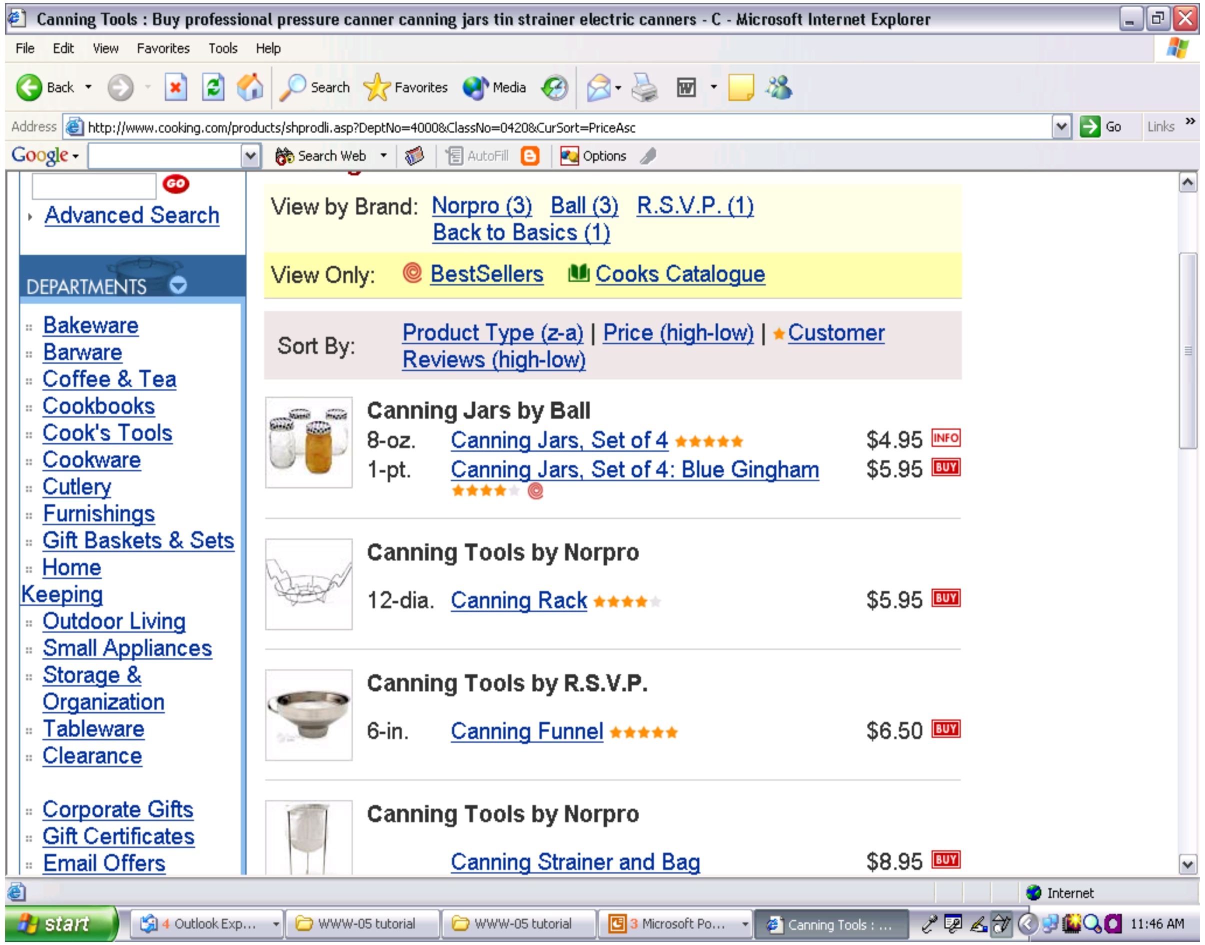
- Two types of data rich pages

➢ List pages
  - Each such page contains one or more lists of data records.
  - Each list in a specific region in the page

➢ Detail pages
  - Each such page focuses on a single object.
  - But can have a lot of related and unrelated information

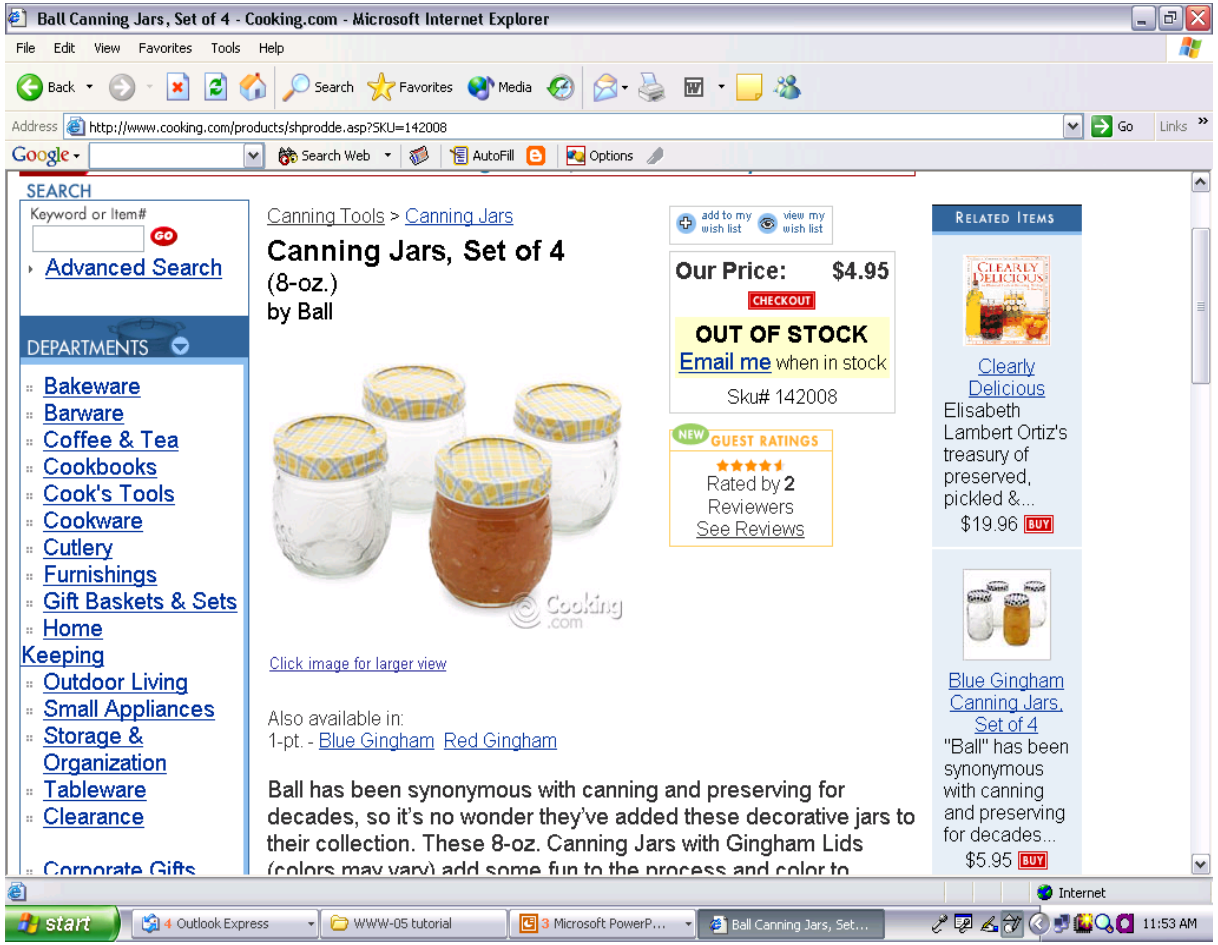
Acquire Data - Information Extraction

- ➢List pages



Acquire Data - Information Extraction

- ➢Detail pages



Acquire Data - Information Extraction

- Extraction results

(a). An example page segment

image 1	Cabinet Organizers by Copco	9-in.	Round Turntable: White	*****	$4.95
image 1	Cabinet Organizers by Copco	12-in.	Round Turntable: White	*****	$7.95
image 2	Cabinet Organizers	14.75x9	Cabinet Organizer (Non-skid): White	*****	$7.95
image 3	Cabinet Organizers	22x6	Cookware Lid Rack	****	$19.95

(b). Extraction results

Acquire Data - Information Extraction

- Wrapper induction (supervised learning)
  - Using machine learning to generate extraction rules.
  - The user marks the target items in a few training pages.
  - The system learns extraction rules from these pages.
  - The rules are applied to extract items from other pages.
- Many wrapper induction systems, e.g.,
  - WIEN (Kushmerick et al, IJCAI-97),
  - Softmealy (Hsu and Dung, 1998),
  - Stalker (Muslea et al. Agents-99),
  - BWI (Freitag and Kushmerick, AAAI-00),
  - WL2 (Cohen et al. WWW-02).

We will only focus on Stalker, which also has a commercial version, Fetch

Acquire Data - Information Extraction

- Stalker: A hierarchical wrapper induction system
- Hierarchical wrapper learning
  - Extraction is isolated at different levels of hierarchy
  - This is suitable for nested data records (embedded list)
- Each item is extracted independent of others.

- Each target item is extracted using two rules
  - A start rule for detecting the beginning of the target item.
  - A end rule for detecting the ending of the target item.

Acquire Data - Information Extraction

- An example nested tuple type

- *name* (of type *string*),
- *image* (of type *image-file*), and
- *differentSizes* (a *set* type), consists of a set of tuples with the attributes:
  - *size* (of type *string*), and
  - *price* (of type *string*).

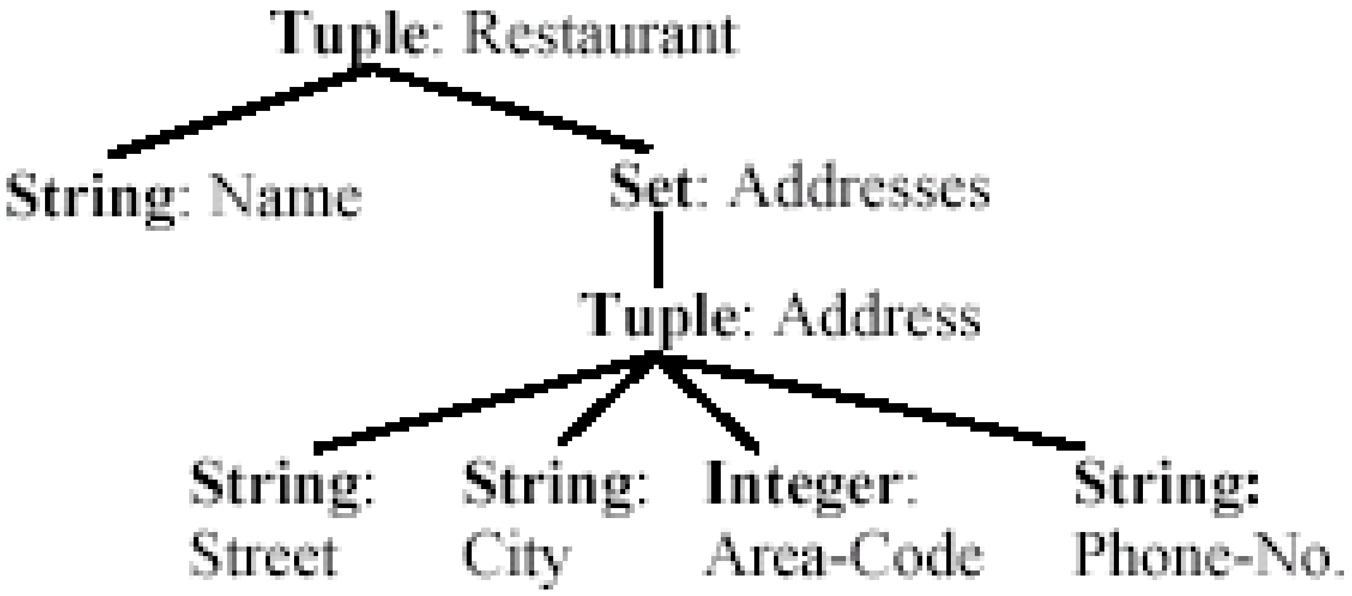
```
tuple  product  ( name:             string;
                  image:            image-file;
                  differentSizes:   set  ( size:    string;
                                           price:   string; ))
```

Acquire Data - Information Extraction

➢Hierarchical representation: tree

Restaurant Name: **Good Noodles**

- 205 Willow, *Glen*, Phone 1-*773*-366-1987
- 25 Oak, *Forest*, Phone (800) 234-7903
- 324 Halsted St., *Chicago*, Phone 1-*800*-996-5023
- 700 Lake St., *Oak Park*, Phone: (708) 798-0008



To extract each target item (a node), the wrapper needs a rule that extracts the item from its parent.

Acquire Data - Information Extraction

- Extraction using two rules
  - a start rule and a end rule.
- The start rule identifies the beginning of the node and the end rule identifies the end of the node.
- For a list node, list iteration rules are needed to break the list into individual data records (tuple instances).

Acquire Data - Information Extraction

- The extraction rules are based on the idea of landmarks.
  - Each landmark is a sequence of consecutive tokens.
- Landmarks are used to locate the beginning and the end of a target item.
- Rules use landmarks

Restaurant Name: **Good Noodles**

- 205 Willow, *Glen*, Phone 1-*773*-366-1987
- 25 Oak, *Forest*, Phone (800) 234-7903
- 324 Halsted St., *Chicago*, Phone 1-*800*-996-5023
- 700 Lake St., *Oak Park*, Phone: (708) 798-0008

Acquire Data - Information Extraction

- An example
  - Let us try to extract the restaurant name “Good Noodles”. Rule R1 can to identify the beginning :

    R1: *SkipTo*() // start rule
  - This rule means that the system should start from the beginning of the page and skip all the tokens until it sees the first tag. is a landmark.
  - Similarly, to identify the end of the restaurant name, we use:

    R2: *SkipTo*() // end rule

```
1:  <p> Restaurant Name: <b>Good Noodles</b><br><br>
2:  <li> 205 Willow, <i>Glen</i>, Phone 1-<i>773</i>-366-1987</li>
3:  <li> 25 Oak, <i>Forest</i>, Phone (800) 234-7903 </li>
4:  <li> 324 Halsted St., <i>Chicago</i>, Phone 1-<i>800</i>-996-5023 </li>
5:  <li> 700 Lake St., <i>Oak Park</i>, Phone: (708) 798-0008 </li>  </p>
```

Acquire Data - Information Extraction

➢Rules are not unique

- Note that a rule may not be unique. For example, we can also use the following rules to identify the beginning of the name:

  R3: *SkiptTo*(Name *_Punctuation_* *_HtmlTag_*)

- **R3** means that we skip everything till the word “Name” followed by a punctuation symbol and then a HTML tag. In this case, “Name *_Punctuation_* *_HtmlTag_*” together is a landmark.
  - *_Punctuation_* and *_HtmlTag_* are **wildcards**.

```
1:  <p> Restaurant Name: <b>Good Noodles</b><br><br>
2:  <li> 205 Willow, <i>Glen</i>, Phone 1-<i>773</i>-366-1987</li>
3:  <li> 25 Oak, <i>Forest</i>, Phone (800) 234-7903 </li>
4:  <li> 324 Halsted St., <i>Chicago</i>, Phone 1-<i>800</i>-996-5023 </li>
5:  <li> 700 Lake St., <i>Oak Park</i>, Phone: (708) 798-0008 </li>  </p>
```

Acquire Data - Information Extraction

- Wrapper maintenance
  - Wrapper verification: If the site changes, does the wrapper know the change?
  - Wrapper repair: If the change is correctly detected, how to automatically repair the wrapper?
  - One way to deal with both problems is to learn the characteristic patterns of the target items.
  - These patterns are then used to monitor the extraction to check whether the extracted items are correct.

Acquire Data - Information Extraction

- Wrapper maintenance
  - Re-labeling: If they are incorrect, the same patterns can be used to locate the correct items assuming that the page changes are minor formatting changes.
  - Re-learning: re-learning produces a new wrapper.
  - Difficult problems: These two tasks are extremely difficult because it often needs contextual and semantic information to detect changes and to find the new locations of the target items.
  - Wrapper maintenance is still an active research area.

Acquire Data - Information Extraction

- Active learning
  - help identify informative unlabeled examples in learning automatically.

1. Randomly select a small subset L of unlabeled examples from U.
2. Manually label the examples in L, and $U = U - L$.
3. Learn a wrapper W based on the labeled set L.
4. Apply W to U to find a set of informative examples L.
5. Stop if $L = \emptyset$, otherwise go to step 2.

Acquire Data - Information Extraction

- Wrapper induction (supervised) has two main shortcomings:
  - It is unsuitable for a large number of sites due to the manual labeling effort.
  - Wrapper maintenance is very costly. The Web is a dynamic environment. Sites change constantly. Since rules learnt by wrapper induction systems mainly use formatting tags, if a site changes its formatting templates, existing extraction rules for the site become invalid.

Acquire Data - Information Extraction

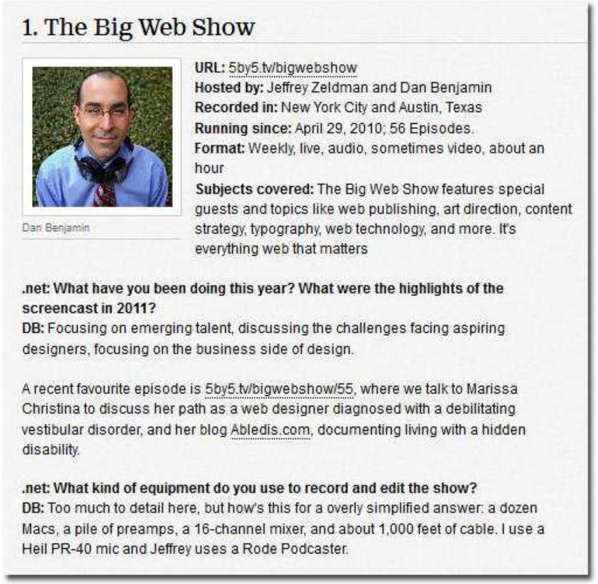
- Unsupervised learning is possible
  - Due to these problems, automatic (or unsupervised) extraction has been studied.
  - Automatic extraction is possible because data records (tuple instances) in a Web site are usually encoded using a very small number of fixed templates.
  - It is possible to find these templates by mining repeated patterns.

Acquire Data - Information Extraction

- ➢Automatic Extraction (unsupervised learning)

“Automatic Extraction of Top-k Lists from the Web”-ICDE 2013

(b)



(c)

(a)

(d)

(e)

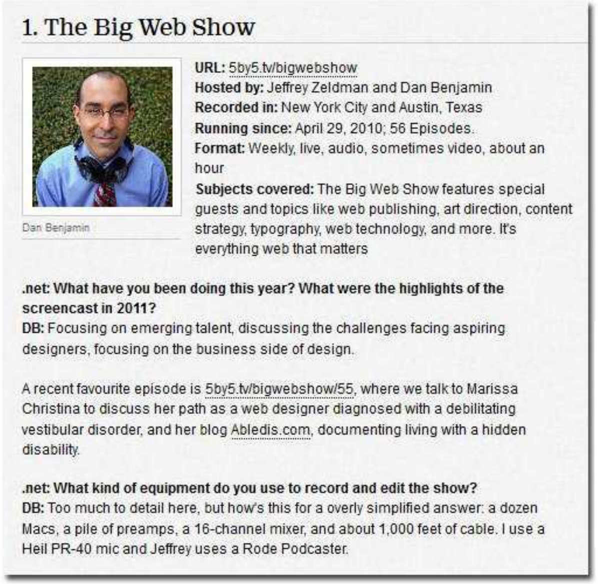
Acquire Data - Information Extraction

➢Automatic Extraction (unsupervised learning)

(b)

"Automatic Extraction of Top-k Lists from the Web"-ICDE 2013

Motivation: *Compared to other structured information on the web (including web tables), information in top-k lists is larger and richer, of higher quality, and generally more interesting.*


1. The Big Web Show

Dan Benjamin

URL: 5by5.tv/bigwebshow
Hosted by: Jeffrey Zeldman and Dan Benjamin
Recorded in: New York City and Austin, Texas
Running since: April 29, 2010; 56 Episodes.
Format: Weekly, live, audio, sometimes video, about an hour
Subjects covered: The Big Web Show features special guests and topics like web publishing, art direction, content strategy, typography, web technology, and more. It's everything web that matters

.net: What have you been doing this year? What were the highlights of the screencast in 2011?
DB: Focusing on emerging talent, discussing the challenges facing aspiring designers, focusing on the business side of design.

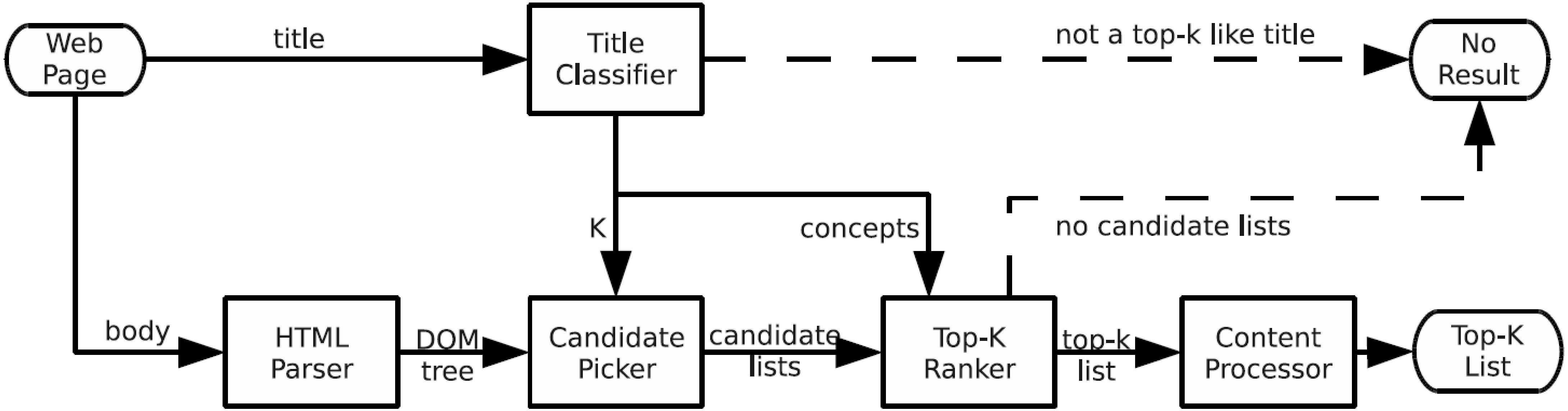
A recent favourite episode is 5by5.tv/bigwebshow/55, where we talk to Marissa Christina to discuss her path as a web designer diagnosed with a debilitating vestibular disorder, and her blog Abledis.com, documenting living with a hidden disability.

.net: What kind of equipment do you use to record and edit the show?
DB: Too much to detail here, but how's this for a overly simplified answer: a dozen Macs, a pile of preamps, a 16-channel mixer, and about 1,000 feet of cable. I use a Heil PR-40 mic and Jeffrey uses a Rode Podcaster.

Index	Name	Image	Url	Hosted by	Recorded in	Running since	Format	...
1	The Big Web Show	[image]	[link]	Zeldman et al.	NYC & Austin, TX	April 29, 2010	Weekly, live...	...
2	Boagworld	[image]	[link]	Boag et al.	a barn in Hampshire	August 2005	Weekly, audio...	...
3	Creative Coding	[image]	[link]	Lee-Delisle et al.	Brighton, Truro...	January 2011	Every two...	...
...	...	...	...	...	...	...	...	...
10	Unmatched Style	[image]	[link]	Crawford et al.	Columbia, SC	2009	Weekly, pre-recorded...	...

Acquire Data - Information Extraction

- Automatic Extraction (unsupervised learning)



System Overview

Acquire Data - Information Extraction

- Automatic Extraction (unsupervised learning)
  - Title Classifier

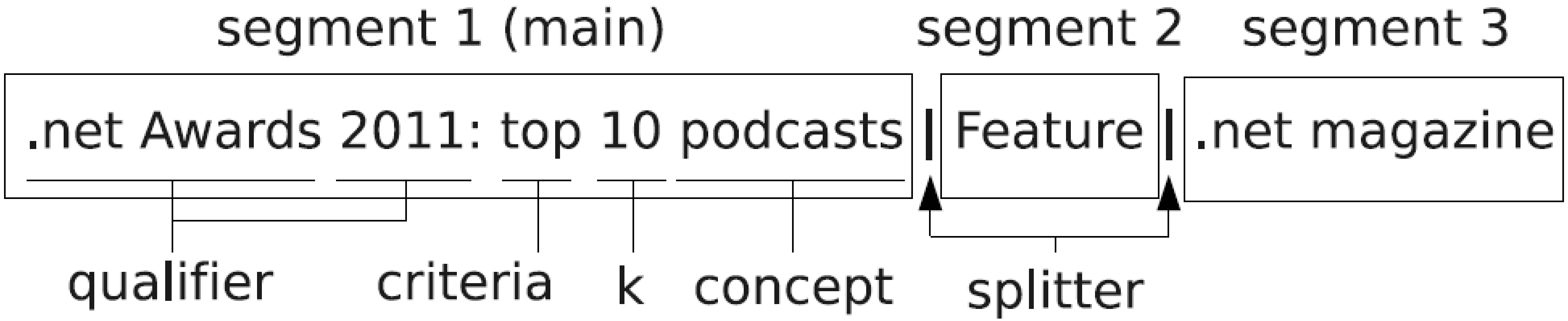
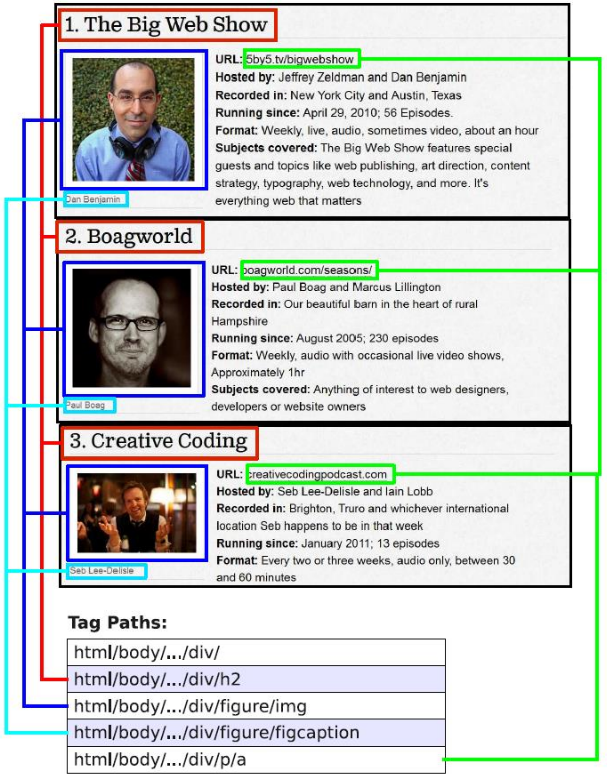
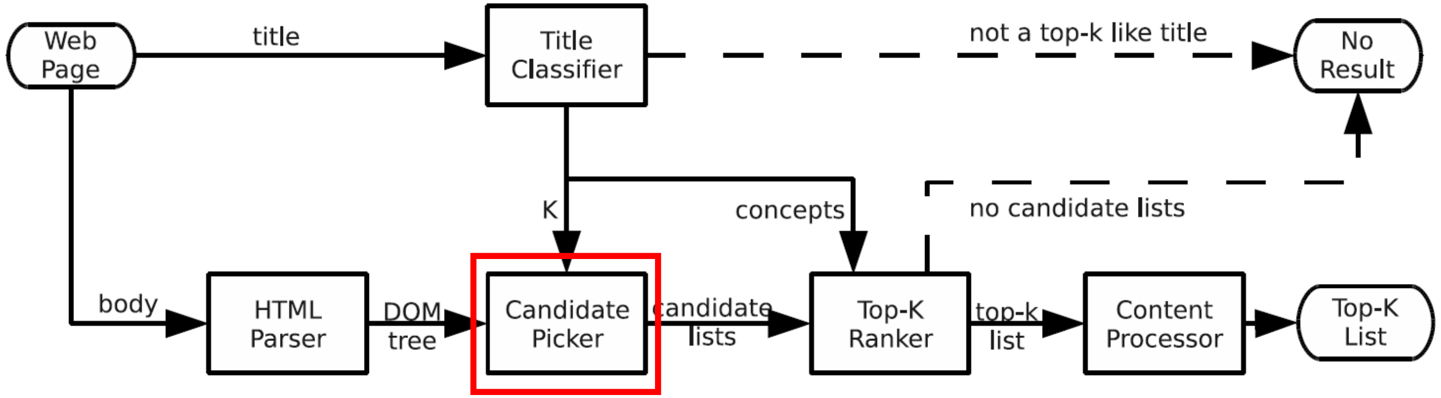


Fig. 3. A Sample Top-K Title

Acquire Data - Information Extraction

- Automatic Extraction (unsupervised learning)
  - Candidate Picker



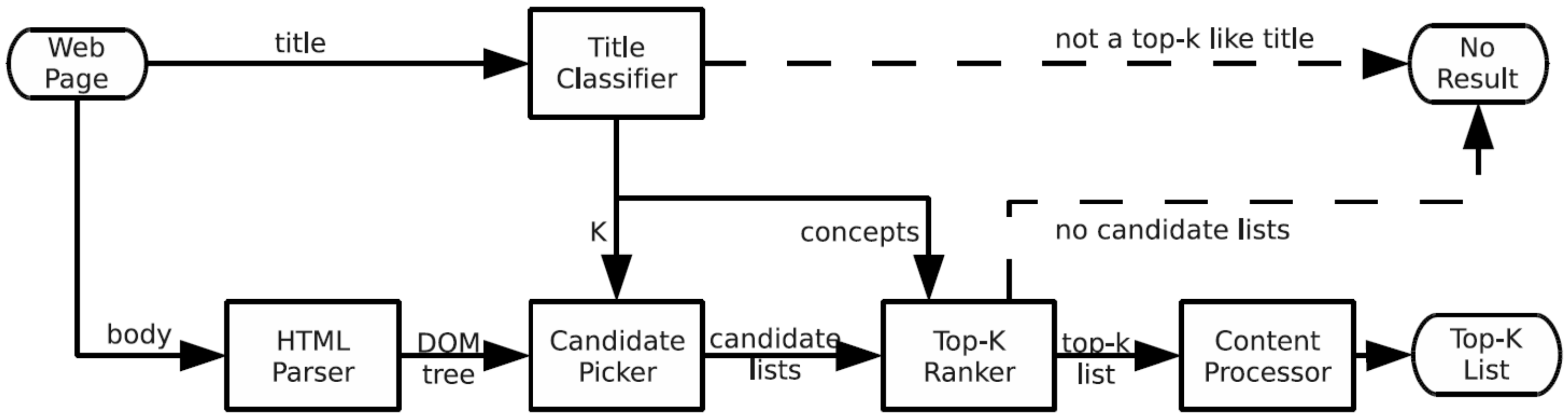


Acquire Data - Information Extraction

- Automatic Extraction (unsupervised learning)
  - Top-K Ranker： Researchers assume that one or more items from the main list should be instances of that central concept from the title.
  - For example, if the title contains the concept “scientist”, then the items of the main list should be instances of the “scientist” concept.
  - Based on “Probase”
  - Calculate the P-Score of each candidate

Acquire Data - Information Extraction

- Automatic Extraction (unsupervised learning)
  - Content Processor
    - Infer the structure of text nodes
    - Conceptualize the list attributes
    - Detect when and where



Acquire Data - Information Extraction

Wrapper induction

- Advantages:
  - Only the target data are extracted as the user can label only data items that he/she is interested in.
  - Due to manual labeling, there is no integration issue for data extracted from multiple sites as the problem is solved by the user.
- Disadvantages:
  - It is not scalable to a large number of sites due to significant manual efforts. Even finding the pages to label is non-trivial.
  - Wrapper maintenance (verification and repair) is very costly if the sites change frequently.

Acquire Data - Information Extraction

Automatic extraction

- Advantages:
  - It is scalable to a huge number of sites due to the automatic process.
  - There is little maintenance cost.
- Disadvantages:
  - It may extract a large amount of unwanted data because the system does not know what is interesting to the user. Domain heuristics or manual filtering may be needed to remove unwanted data.
  - Extracted data from multiple sites need integration, i.e., their schemas need to be matched.

Acquire Data - Information Extraction

- In terms of extraction accuracy, it is reasonable to assume that wrapper induction is more accurate than automatic extraction. However, there is no reported comparison.
- Applications
  - Wrapper induction should be used in applications in which the number of sites to be extracted and the number of templates in these sites are not large.
  - Automatic extraction is more suitable for large scale extraction tasks which do not require accurate labeling or integration.
- **Still an active research area**.