Exploiting Structured Data in Textual Content from the Web: Methods, Techniques and Applications

Altigran Soares da Silva (alti)

alti@icomp.ufam.edu.br

Instituto de Computação

Universidade Federal do Amazonas

UFAM

iComp Instituto de Computação

SBBD2012 SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS

Acknowledgments

- Joint Work with people from the BDRI Group at UFAM and InWeb at UFMG
- Industrial cooperation

neemu

zunnit

UOL

- Support

inweb

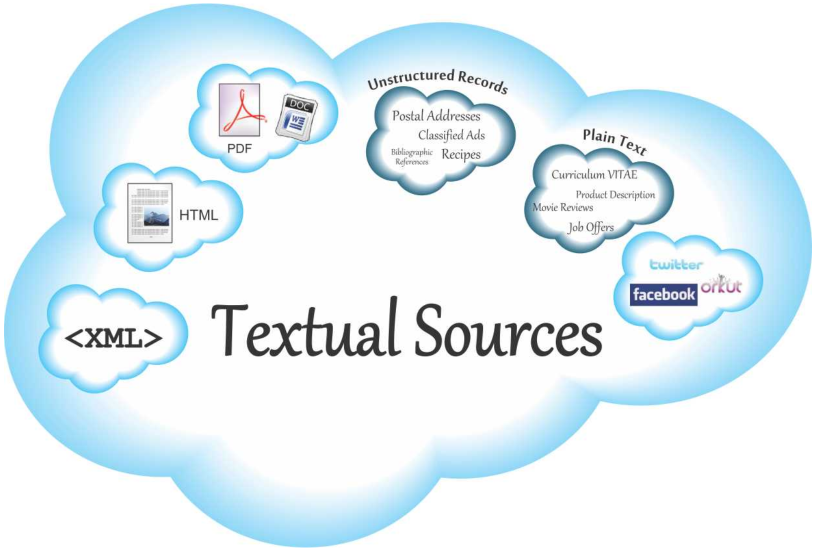
CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico
60 ANOS

FAPEAM

Where is the data?

- Data of interest is no longer only in databases
  - They are, though, available in on-line sources
  - In particular: textual sources
    - Social networks, Wikis, Blogs, Web of Data, RSS, e-mail, …
- Search engines are effective and popular tools
- Consensus:
  - its possible to better exploit them



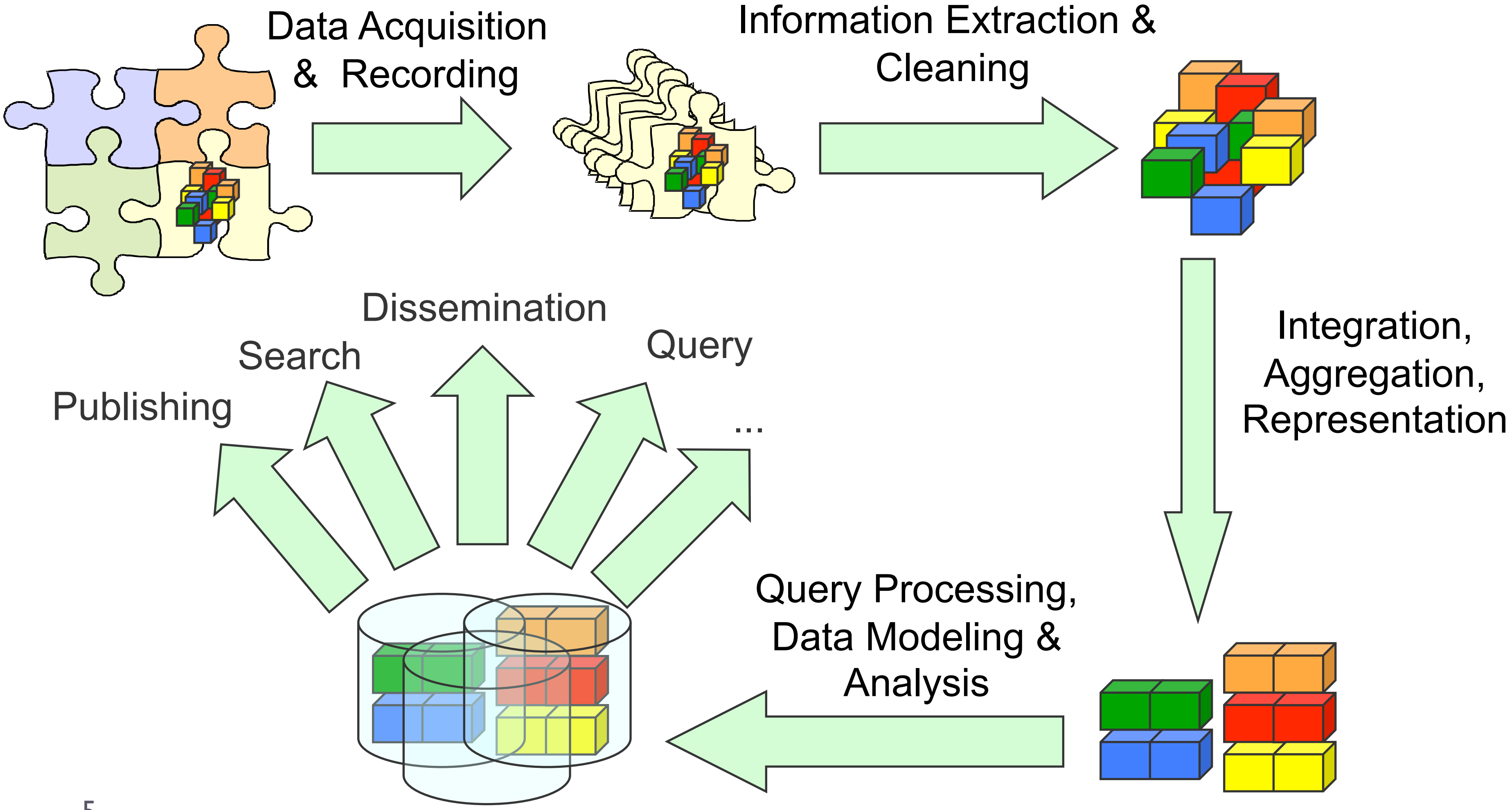
How to deal with it?

- Textual Sources
  - The structure is only implicit
  - Meta-data is a luxury
  - Constraints are a utopia
- We do need semantics!
- Multiple proposals to increase the expressive power
  - Syntactically: e.g., XML technology, RDF, etc.
  - Semantically: e.g., Semantic Web, Linked Data, etc.
- Challenge: adoption of standards
  - Governance is needed, and it is good!!
  - But, the web was born messy and its is likely to remain like that

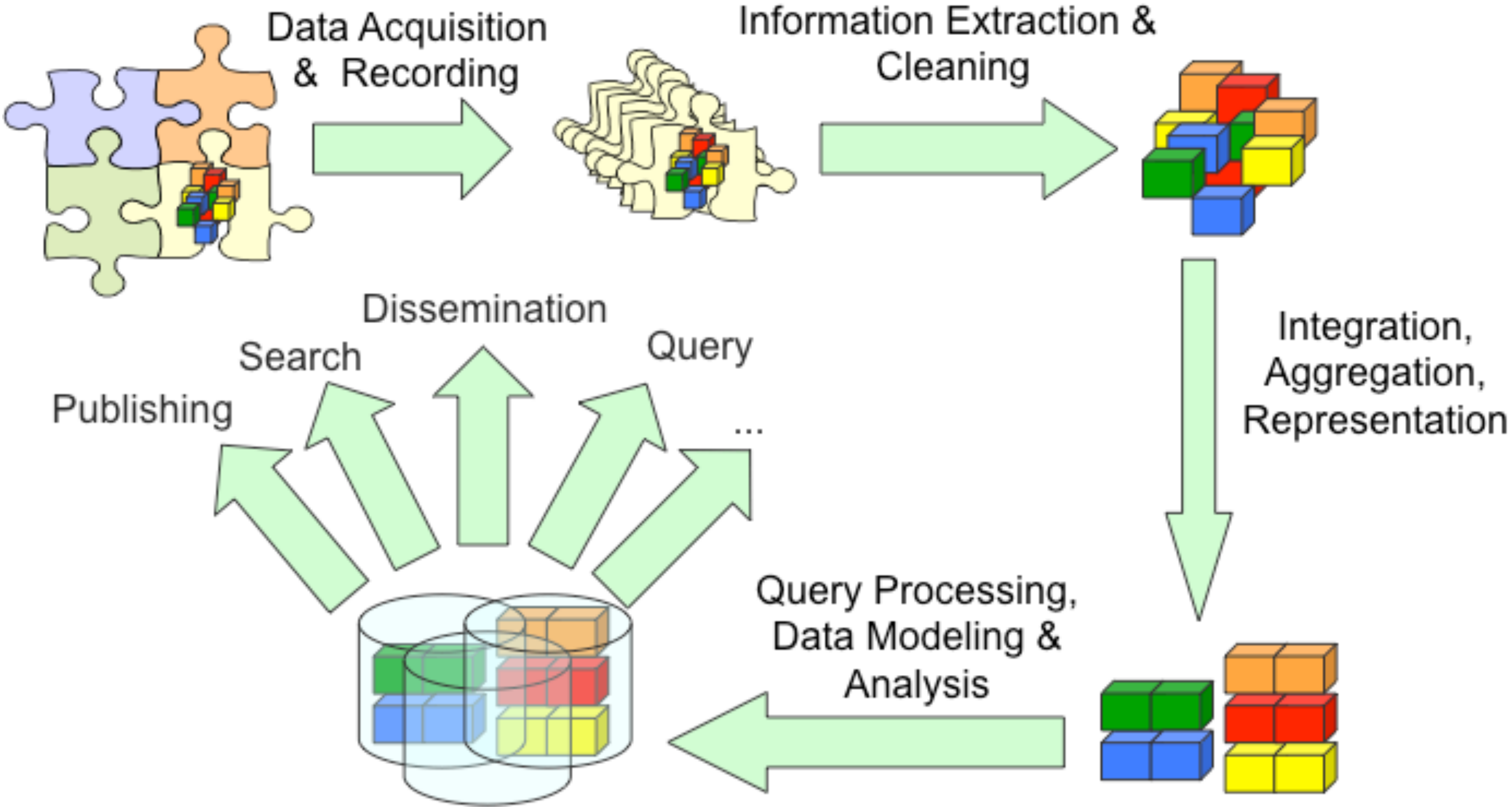
Any alternative ?

- Possible alternative perspective:
  - Methods & Techniques for “automatically” gathering, extracting , enriching and exploiting data available in textual Web sources
- By no means new!
  - It has been out there for more than a decade!
- New impulse: Industrial needs
  - Advances in Data Management, Information Retrieval, Machine Learning, Data Mining, Artificial Intelligence, …
- Research on this subject is immediately applicable
  - Motivates a continuous feedback between industry and academia

Many Problems …



It is Big Data !


Data Acquisition & Recording
Information Extraction & Cleaning
Integration, Aggregation, Representation
Query Processing, Data Modeling & Analysis
Publishing
Search
Dissemination
Query
...

The Big Data Analysis Pipeline

H.V. Jagadish – ACM SIGMOD Blog - 05/06/2012

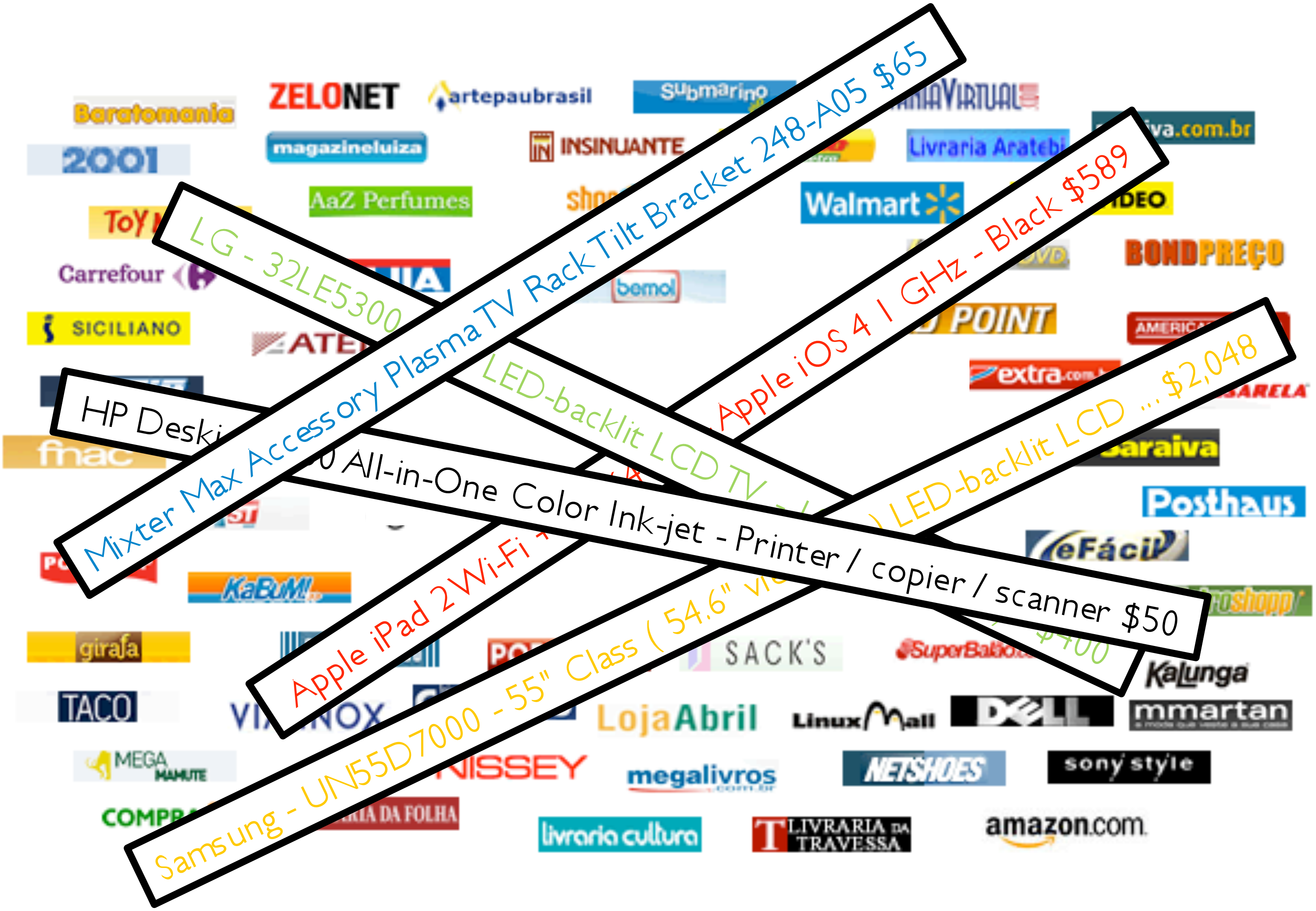
Challenges & Opportunities w/ Big Data – Online report

e-Shopping Aggregation

- e-Shopping Aggregators receive and/or crawl hundreds of thousands unstructured product offers from thousands of stores
- Available as ordinary unstructured textual descriptions
- Different “styles” depending on the source and on the type of product

Apple iPad 2 Wi-Fi + 3G 64 GB - Apple iOS 4 1 GHz - Black $589
LG - 32LE5300 - 32" LED-backlit LCD TV - 1080p (FullHD) - $400
Samsung - UN55D7000 - 55" Class (54.6" viewable) LED-backlit LCD ... $2,048
Mixter Max Accessory Plasma TV Rack Tilt Bracket 248-A05 $65
HP Deskjet 3050 All-in-One Color Ink-jet - Printer / copier / scanner $50

e-Shopping Aggregation

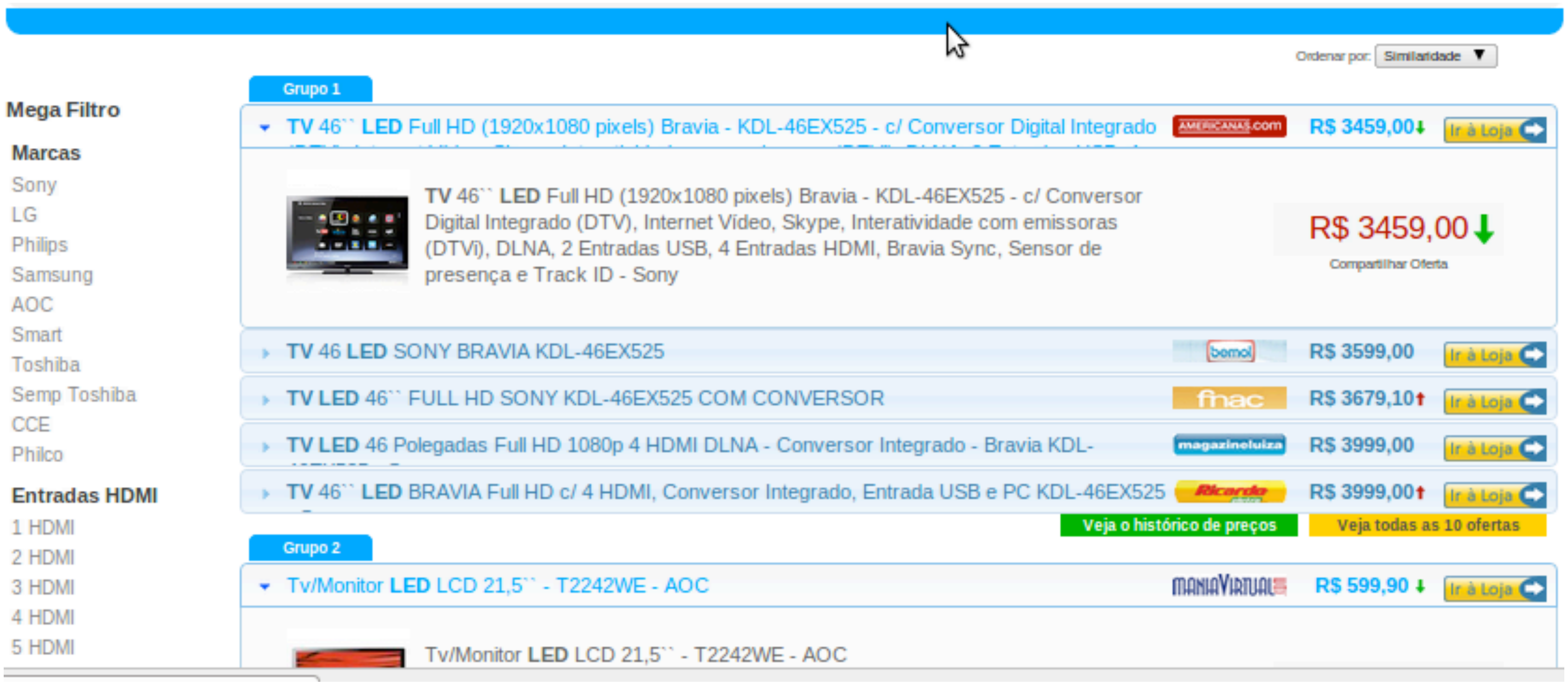


e-Shopping Aggregation

- Main Tasks/Services
  - Crawl product offers over the Web
  - Product aggregation: cluster offers of a same product
  - Categorization: put offers in the right category
  - Structured search: e.g., search by brand
  - Product comparison: e.g., give me the cheapest 3D 40” TV
- Easier if data in offers is correctly segmented and labeled

Type	*Brand*	*Size*	*Screen Type*	*Price*
TV	Samsung	55“	LED-backlit	$2,048

Live showcase: neemu.com by neemu



Ordenar por: Similaridade ▼

Mega Filtro

Marcas

- Sony
- LG
- Philips
- Samsung
- AOC
- Smart
- Toshiba
- Semp Toshiba
- CCE
- Philco

Entradas HDMI

- 1 HDMI
- 2 HDMI
- 3 HDMI
- 4 HDMI
- 5 HDMI

Grupo 1

Produto	Loja	Preço	
TV 46`` **LED** Full HD (1920x1080 pixels) Bravia - KDL-46EX525 - c/ Conversor Digital Integrado	AMERICANAS.com	R$ 3459,00↓	Ir à Loja
TV 46 **LED** SONY BRAVIA KDL-46EX525	bemol	R$ 3599,00	Ir à Loja
TV LED 46`` FULL HD SONY KDL-46EX525 COM CONVERSOR	fnac	R$ 3679,10↑	Ir à Loja
TV LED 46 Polegadas Full HD 1080p 4 HDMI DLNA - Conversor Integrado - Bravia KDL-	magazineluiza	R$ 3999,00	Ir à Loja
TV 46`` **LED** BRAVIA Full HD c/ 4 HDMI, Conversor Integrado, Entrada USB e PC KDL-46EX525	Ricardo Eletro	R$ 3999,00↑	Ir à Loja

TV 46`` **LED** Full HD (1920x1080 pixels) Bravia - KDL-46EX525 - c/ Conversor Digital Integrado (DTV), Internet Vídeo, Skype, Interatividade com emissoras (DTVi), DLNA, 2 Entradas USB, 4 Entradas HDMI, Bravia Sync, Sensor de presença e Track ID - Sony

R$ 3459,00↓

Compartilhar Oferta

Veja o histórico de preços

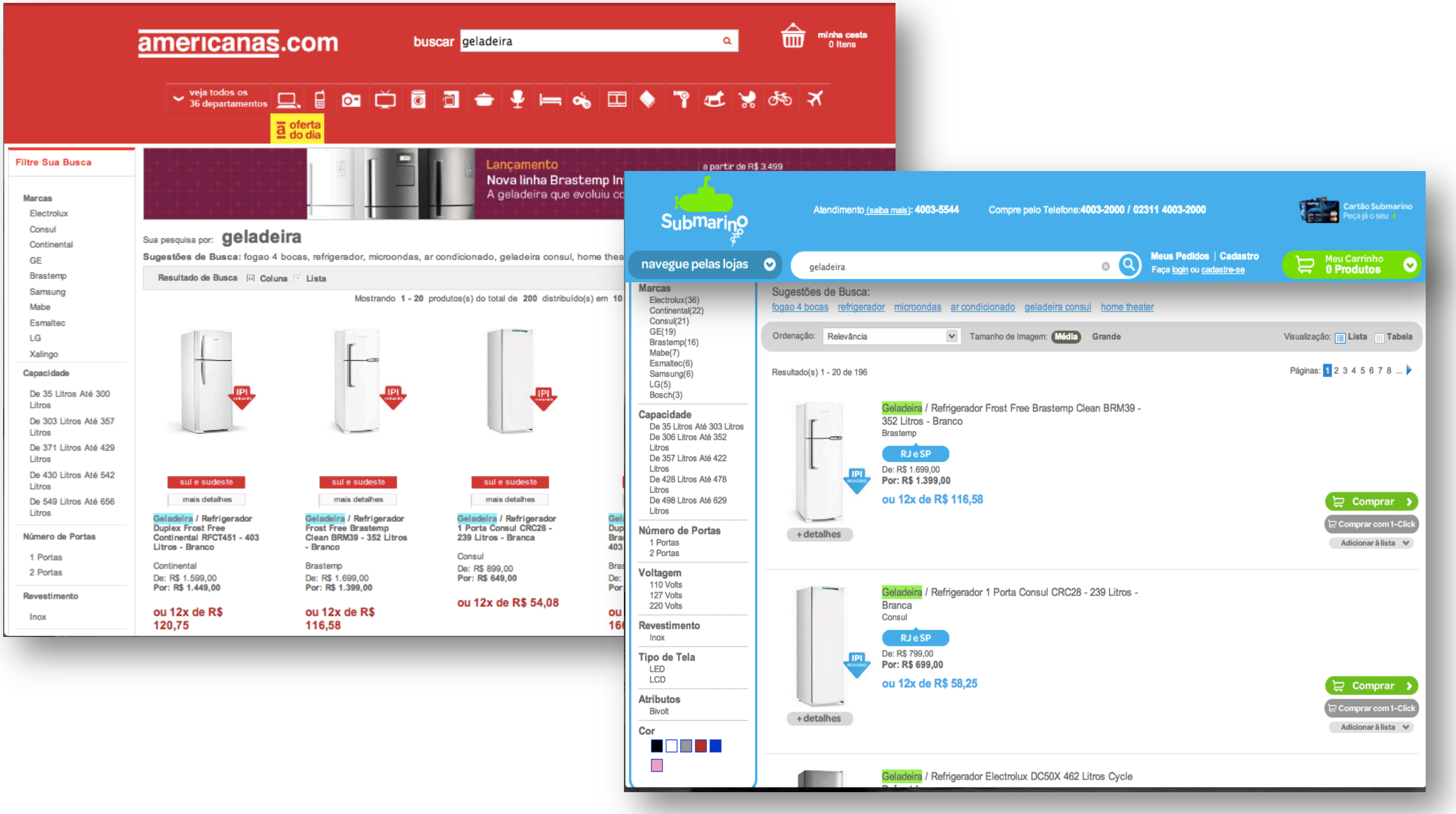
Veja todas as 10 ofertas

Grupo 2

Produto	Loja	Preço	
Tv/Monitor **LED** LCD 21,5`` - T2242WE - AOC	MANIAVIRTUAL	R$ 599,90↓	Ir à Loja

Tv/Monitor **LED** LCD 21,5`` - T2242WE - AOC

Also powered by neemu



americanas.com

buscar geladeira

minha cesta
0 itens

veja todos os
36 departamentos

oferta do dia

Filtre Sua Busca

Marcas
- Electrolux
- Consul
- Continental
- GE
- Brastemp
- Samsung
- Mabe
- Esmaltec
- LG
- Xalingo

Capacidade
- De 35 Litros Até 300 Litros
- De 303 Litros Até 357 Litros
- De 371 Litros Até 429 Litros
- De 430 Litros Até 542 Litros
- De 549 Litros Até 656 Litros

Número de Portas
- 1 Portas
- 2 Portas

Revestimento
- Inox

Lançamento
Nova linha Brastemp In
A geladeira que evoluiu co
a partir de R$ 3.499

Sua pesquisa por: **geladeira**

Sugestões de Busca: fogao 4 bocas, refrigerador, microondas, ar condicionado, geladeira consul, home thea

Resultado de Busca | Coluna | Lista

Mostrando **1 - 20** produtos(s) do total de **200** distribuído(s) em **10**

sul e sudeste
mais detalhes
Geladeira / Refrigerador Duplex Frost Free Continental RFCT451 - 403 Litros - Branco
Continental
De: R$ 1.599,00
Por: R$ 1.449,00
ou 12x de R$ 120,75

sul e sudeste
mais detalhes
Geladeira / Refrigerador Frost Free Brastemp Clean BRM39 - 352 Litros - Branco
Brastemp
De: R$ 1.699,00
Por: R$ 1.399,00
ou 12x de R$ 116,58

sul e sudeste
mais detalhes
Geladeira / Refrigerador 1 Porta Consul CRC28 - 239 Litros - Branca
Consul
De: R$ 899,00
Por: R$ 649,00
ou 12x de R$ 54,08

Submarino

Atendimento (saiba mais): **4003-5544** Compre pelo Telefone:**4003-2000 / 02311 4003-2000**

Cartão Submarino
Peça já o seu

navegue pelas lojas

geladeira

Meus Pedidos | Cadastro
Faça login ou cadastre-se

Meu Carrinho
0 Produtos

Marcas
- Electrolux(36)
- Continental(22)
- Consul(21)
- GE(19)
- Brastemp(16)
- Mabe(7)
- Esmaltec(6)
- Samsung(6)
- LG(5)
- Bosch(3)

Capacidade
- De 35 Litros Até 303 Litros
- De 306 Litros Até 352 Litros
- De 357 Litros Até 422 Litros
- De 428 Litros Até 478 Litros
- De 498 Litros Até 629 Litros

Número de Portas
- 1 Portas
- 2 Portas

Voltagem
- 110 Volts
- 127 Volts
- 220 Volts

Revestimento
- Inox

Tipo de Tela
- LED
- LCD

Atributos
- Bivolt

Cor
■ □ ■ ■ ■ ■

Sugestões de Busca:
fogao 4 bocas | refrigerador | microondas | ar condicionado | geladeira consul | home theater

Ordenação: Relevância | Tamanho de Imagem: **Média** **Grande** | Visualização: **Lista** **Tabela**

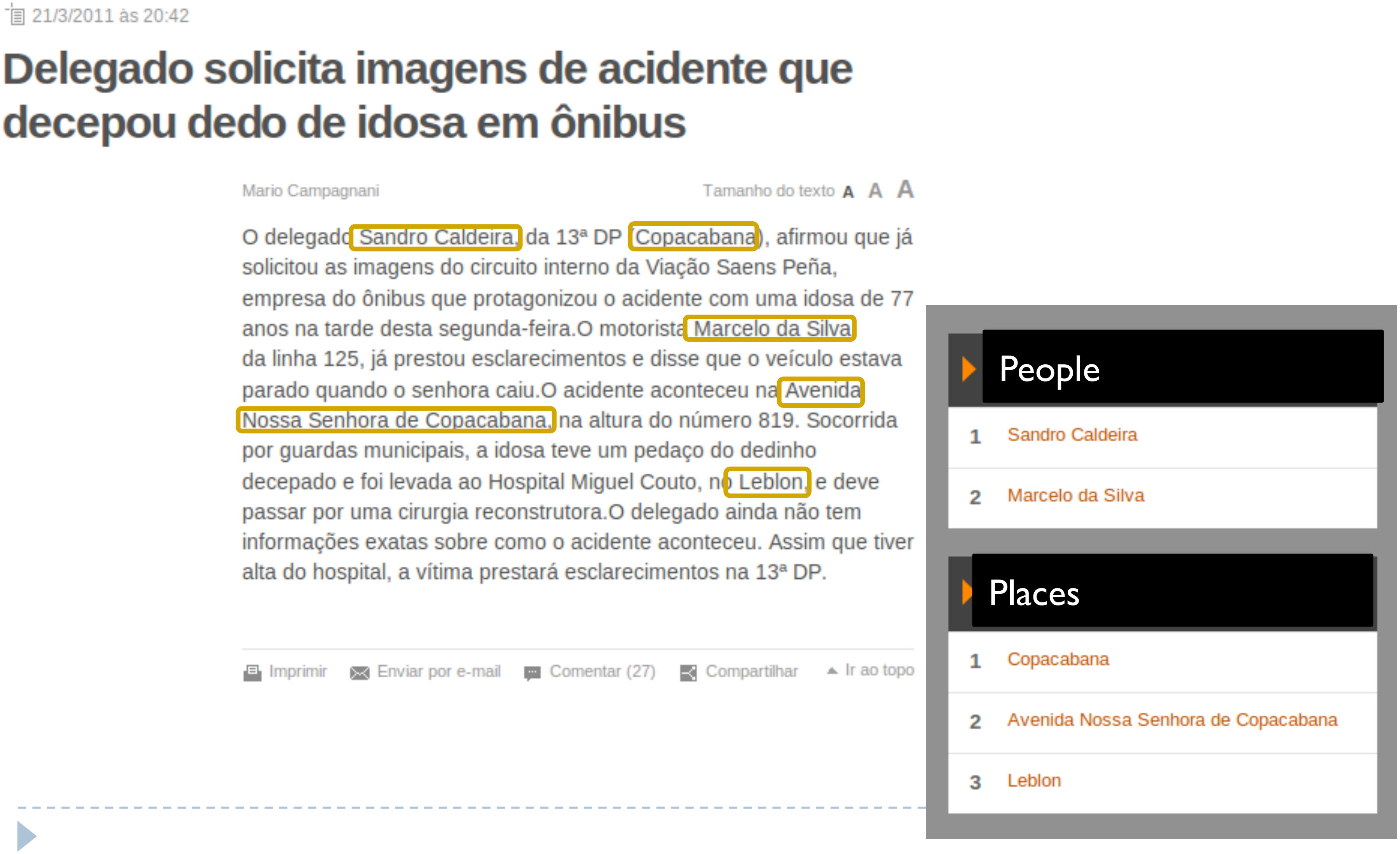
Resultado(s) 1 - 20 de 196 | Páginas: 1 2 3 4 5 6 7 8 ...

Geladeira / Refrigerador Frost Free Brastemp Clean BRM39 - 352 Litros - Branco
Brastemp
RJ e SP
De: R$ 1.699,00
Por: R$ 1.399,00
ou 12x de R$ 116,58
+detalhes
Comprar
Comprar com 1-Click
Adicionar à lista

Geladeira / Refrigerador 1 Porta Consul CRC28 - 239 Litros - Branca
Consul
RJ e SP
De: R$ 799,00
Por: R$ 699,00
ou 12x de R$ 58,25
+detalhes
Comprar
Comprar com 1-Click
Adicionar à lista

Geladeira / Refrigerador Electrolux DC50X 462 Litros Cycle

Entity recognition by zunnit



21/3/2011 às 20:42

Delegado solicita imagens de acidente que decepou dedo de idosa em ônibus

Mario Campagnani

O delegado Sandro Caldeira, da 13ª DP (Copacabana), afirmou que já solicitou as imagens do circuito interno da Viação Saens Peña, empresa do ônibus que protagonizou o acidente com uma idosa de 77 anos na tarde desta segunda-feira.O motorista Marcelo da Silva da linha 125, já prestou esclarecimentos e disse que o veículo estava parado quando o senhora caiu.O acidente aconteceu na Avenida Nossa Senhora de Copacabana, na altura do número 819. Socorrida por guardas municipais, a idosa teve um pedaço do dedinho decepado e foi levada ao Hospital Miguel Couto, no Leblon, e deve passar por uma cirurgia reconstrutora.O delegado ainda não tem informações exatas sobre como o acidente aconteceu. Assim que tiver alta do hospital, a vítima prestará esclarecimentos na 13ª DP.

People

1 Sandro Caldeira

2 Marcelo da Silva

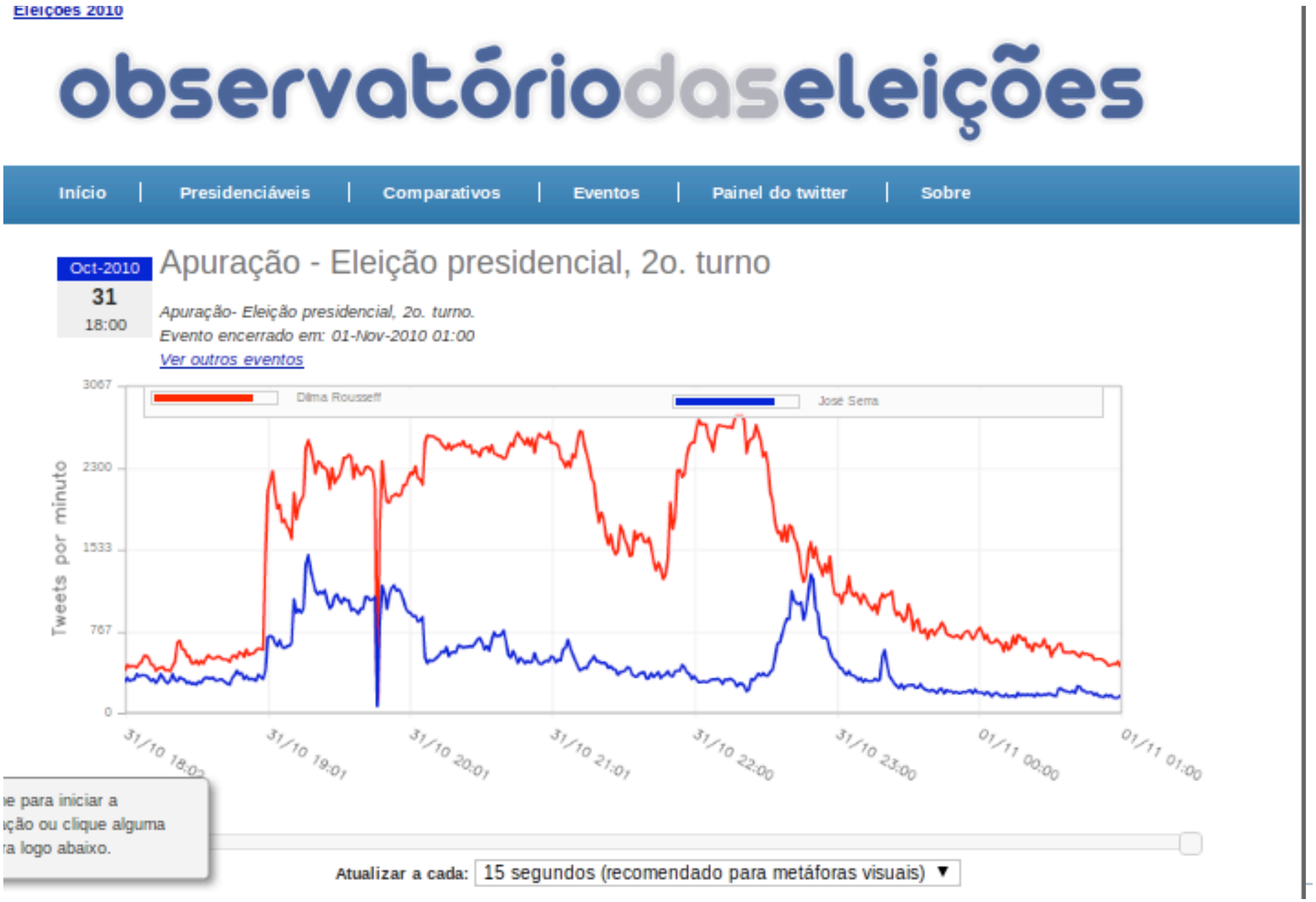
Places

1 Copacabana

2 Avenida Nossa Senhora de Copacabana

3 Leblon

Entity Disambiguation at inweb



Eleições 2010
observatóriodaseleições
Início | Presidenciáveis | Comparativos | Eventos | Painel do twitter | Sobre
Oct-2010
31
18:00

Apuração - Eleição presidencial, 2o. turno

Apuração- Eleição presidencial, 2o. turno.

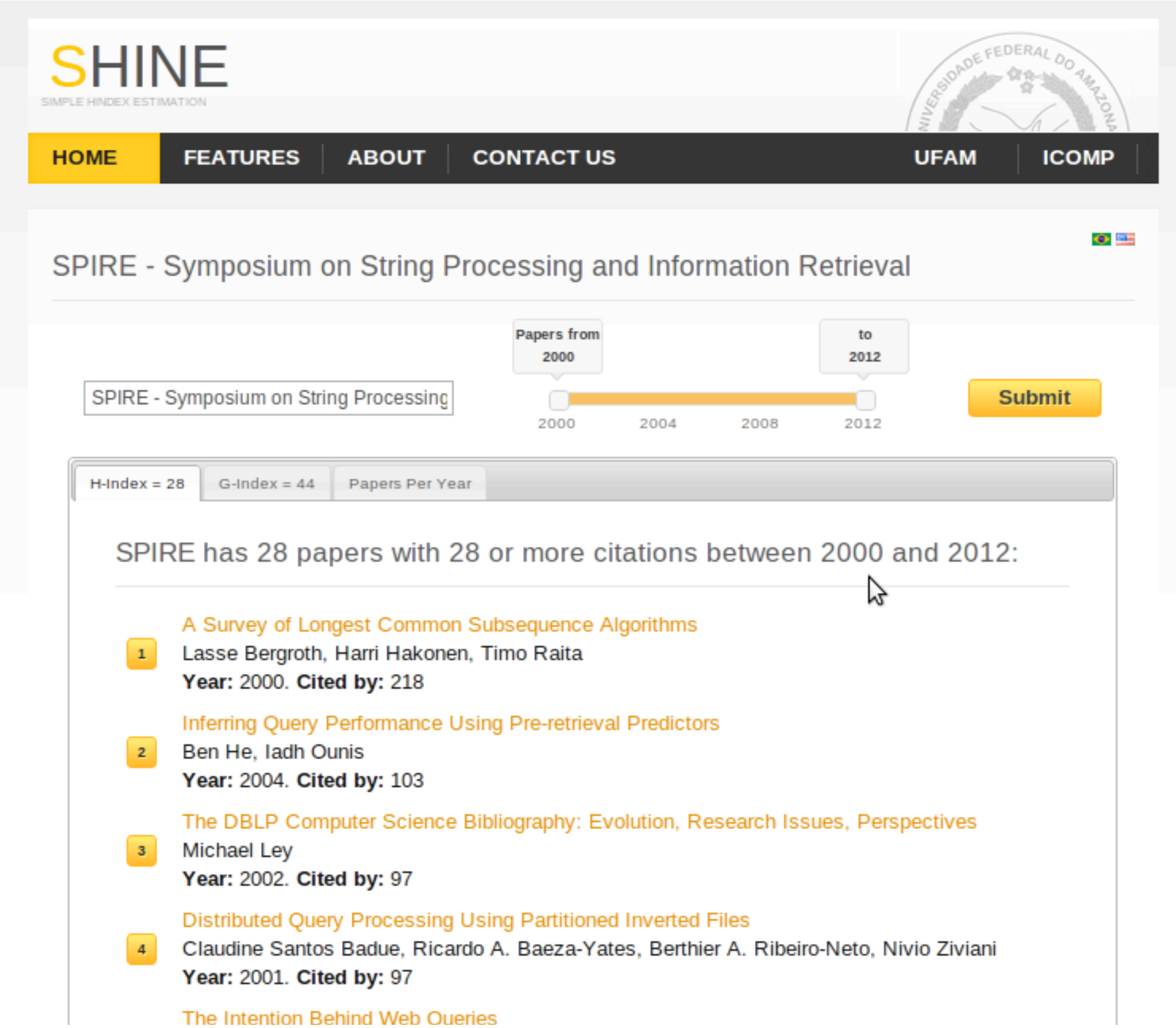
Evento encerrado em: 01-Nov-2010 01:00

Ver outros eventos

ie para iniciar a
ição ou clique alguma
ra logo abaixo.

Atualizar a cada: 15 segundos (recomendado para metáforas visuais) ▼

Management of Bib. References in SHINE



SPIRE - Symposium on String Processing and Information Retrieval

SPIRE - Symposium on String Processing

Papers from 2000 to 2012

2000 2004 2008 2012

H-Index = 28 | G-Index = 44 | Papers Per Year

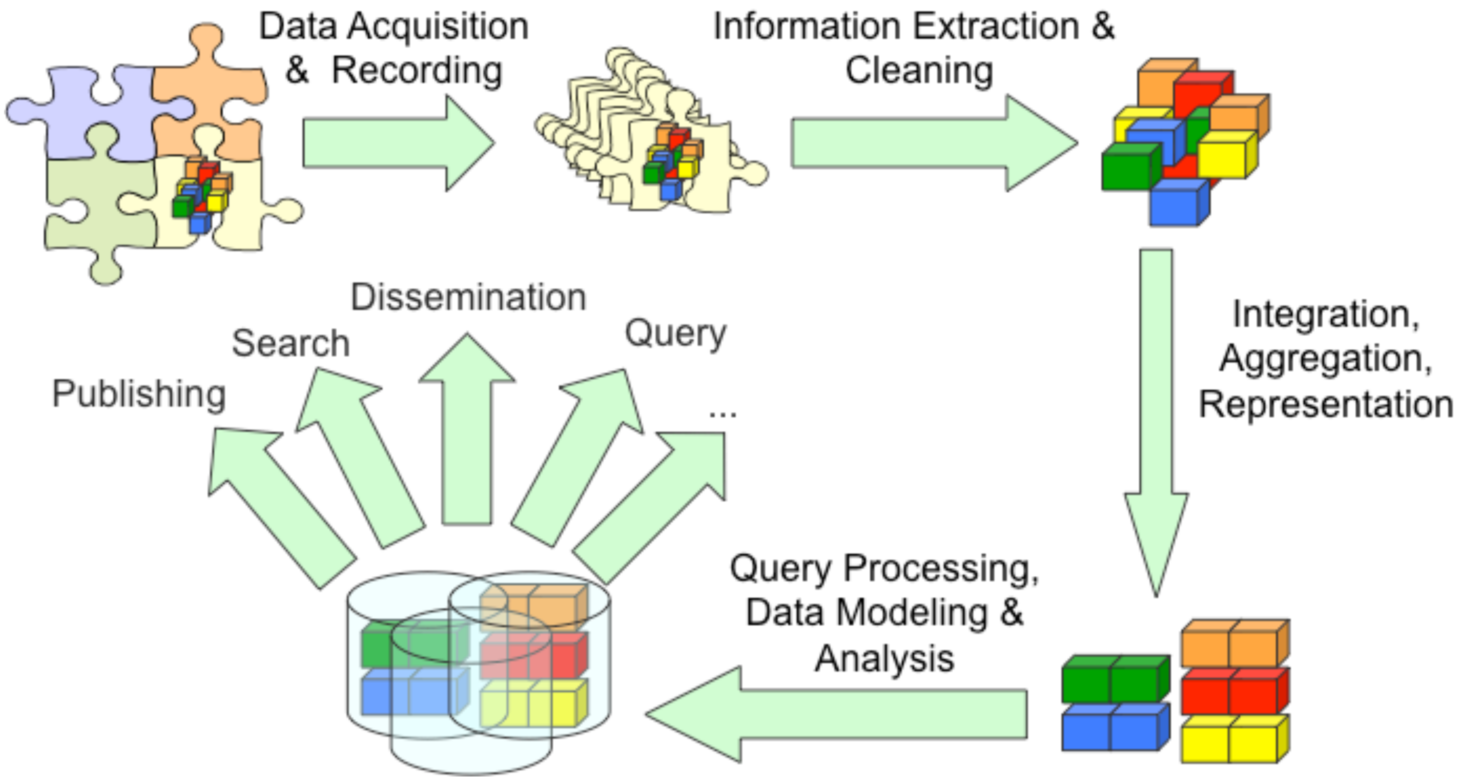
SPIRE has 28 papers with 28 or more citations between 2000 and 2012:

1. A Survey of Longest Common Subsequence Algorithms
   Lasse Bergroth, Harri Hakonen, Timo Raita
   Year: 2000. **Cited by:** 218
2. Inferring Query Performance Using Pre-retrieval Predictors
   Ben He, Iadh Ounis
   Year: 2004. **Cited by:** 103
3. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives
   Michael Ley
   Year: 2002. **Cited by:** 97
4. Distributed Query Processing Using Partitioned Inverted Files
   Claudine Santos Badue, Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto, Nivio Ziviani
   Year: 2001. **Cited by:** 97

The Intention Behind Web Queries

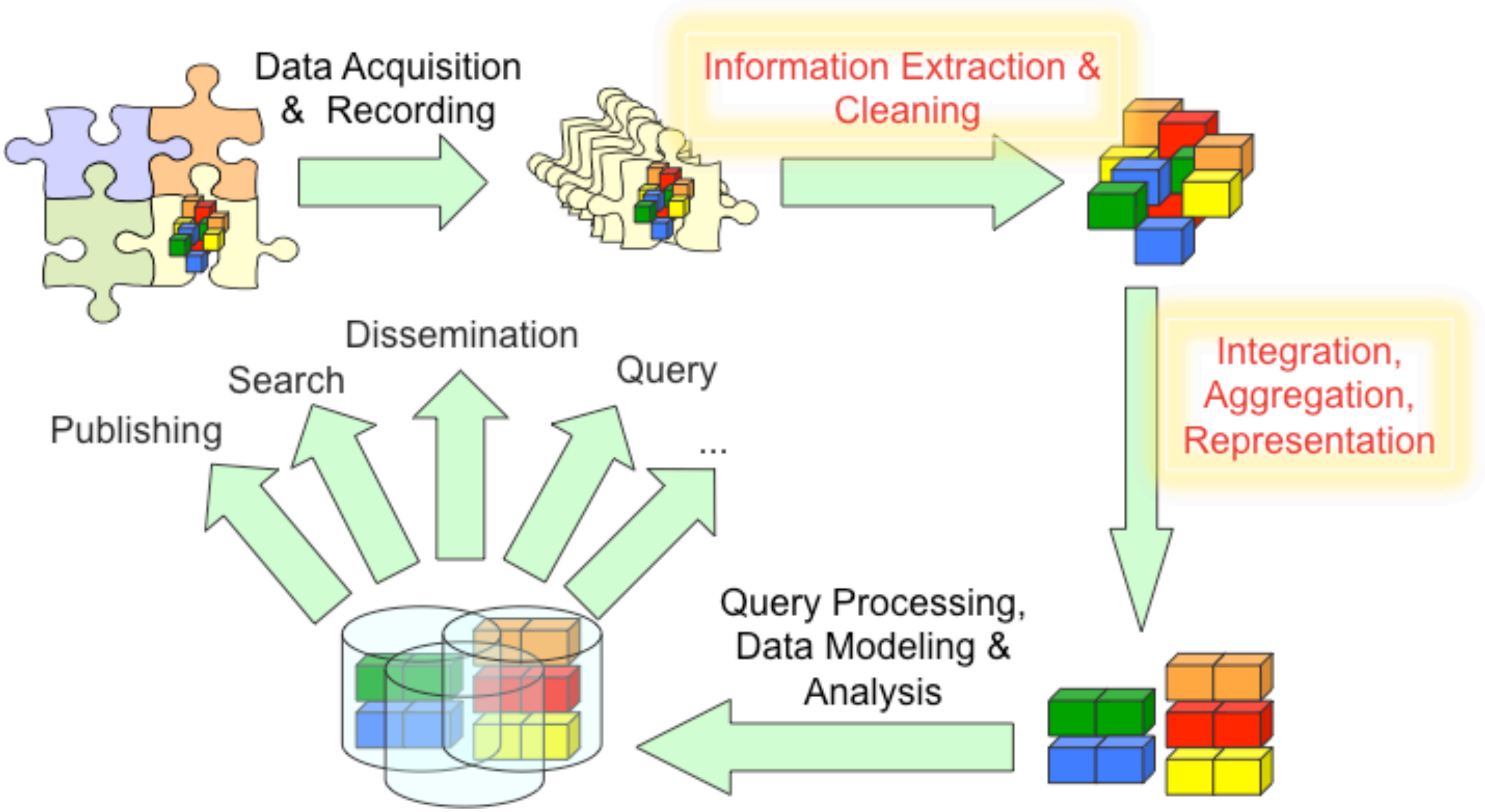
Structured Data in Textual Content

- We have studied, developed, published and applied methods and techniques for all of these problems



Structured Data in Textual Content

- In this talk, focus on 3 specific results for two problems

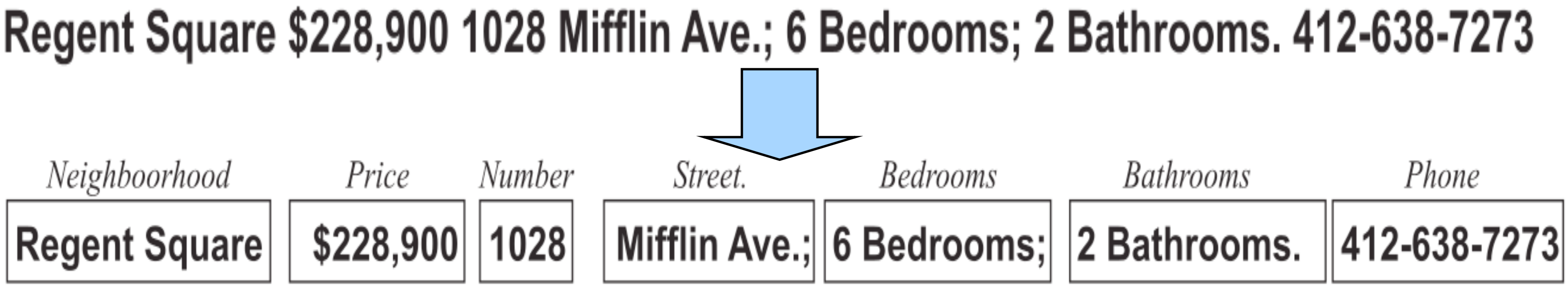


In this talk

- Information Extraction
  - ONDUX [SIGMOD'10] and JUDIE [SIGMOD'11]
- Filling of Web Forms
  - IForm [VLDB'11]
- Complex Schema Matching
  - EvoMatch [IS'13]

IETS

- Information extraction by text segmentation (IETS)
  - Extracting semi-structured data records by identifying attribute in continuous text
  - bibliographic citations, product descriptions, classified ads, etc



Regent Square $228,900 1028 Mifflin Ave.; 6 Bedrooms; 2 Bathrooms. 412-638-7273

Neighboorhood	*Price*	*Number*	*Street.*	*Bedrooms*	*Bathrooms*	*Phone*
Regent Square	$228,900	1028	Mifflin Ave.;	6 Bedrooms;	2 Bathrooms.	412-638-7273

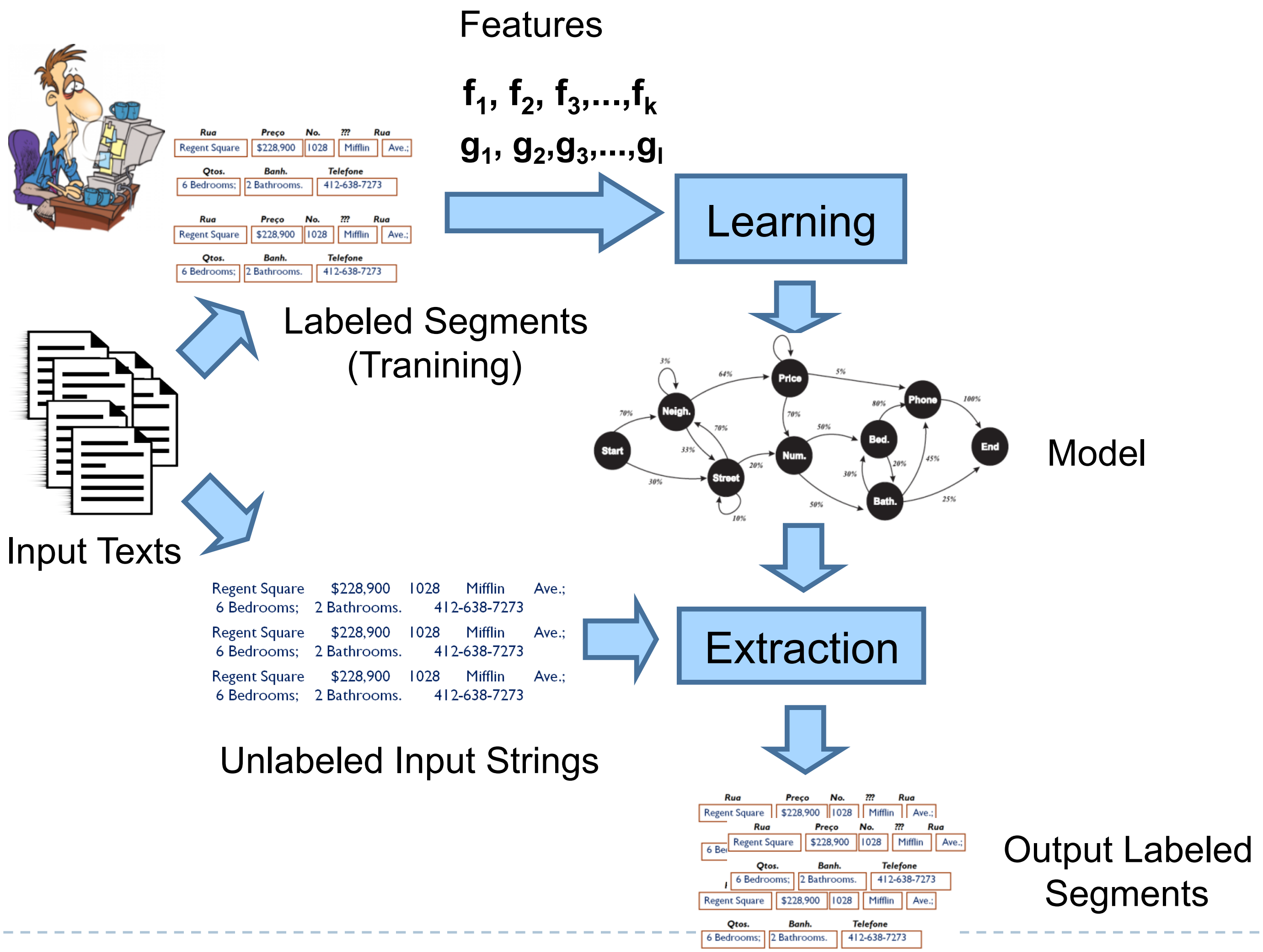
  - Ungrammatical text – not suitable for NLP methods

Supervised Methods

- Current IETS methods use probabilistic frameworks such as HMM or CRF
- Learn a model for extracting data related to a domain
- Supervised IETS methods
  - Require training data from each source

```
<Neighboorhood>Regent Square </Neighboorhood> <Price> $228,900 </Price>
<No>1028 </No><Street>Mifflin Ave, </Street> <Bed>6 Bedrooms </Bed> <Bath> 2
Bathrooms </Bath> <Phone>412-638-7273 </Phone>
```

Supervised IETS



Features

$f_1, f_2, f_3,\ldots,f_k$

$g_1, g_2,g_3,\ldots,g_l$

Learning

Rua Preço No. ??? Rua
Regent Square $228,900 1028 Mifflin Ave.;
Qtos. Banh. Telefone
6 Bedrooms; 2 Bathrooms. 412-638-7273

Rua Preço No. ??? Rua
Regent Square $228,900 1028 Mifflin Ave.;
Qtos. Banh. Telefone
6 Bedrooms; 2 Bathrooms. 412-638-7273

Labeled Segments
(Tranining)

Input Texts

Model

Extraction

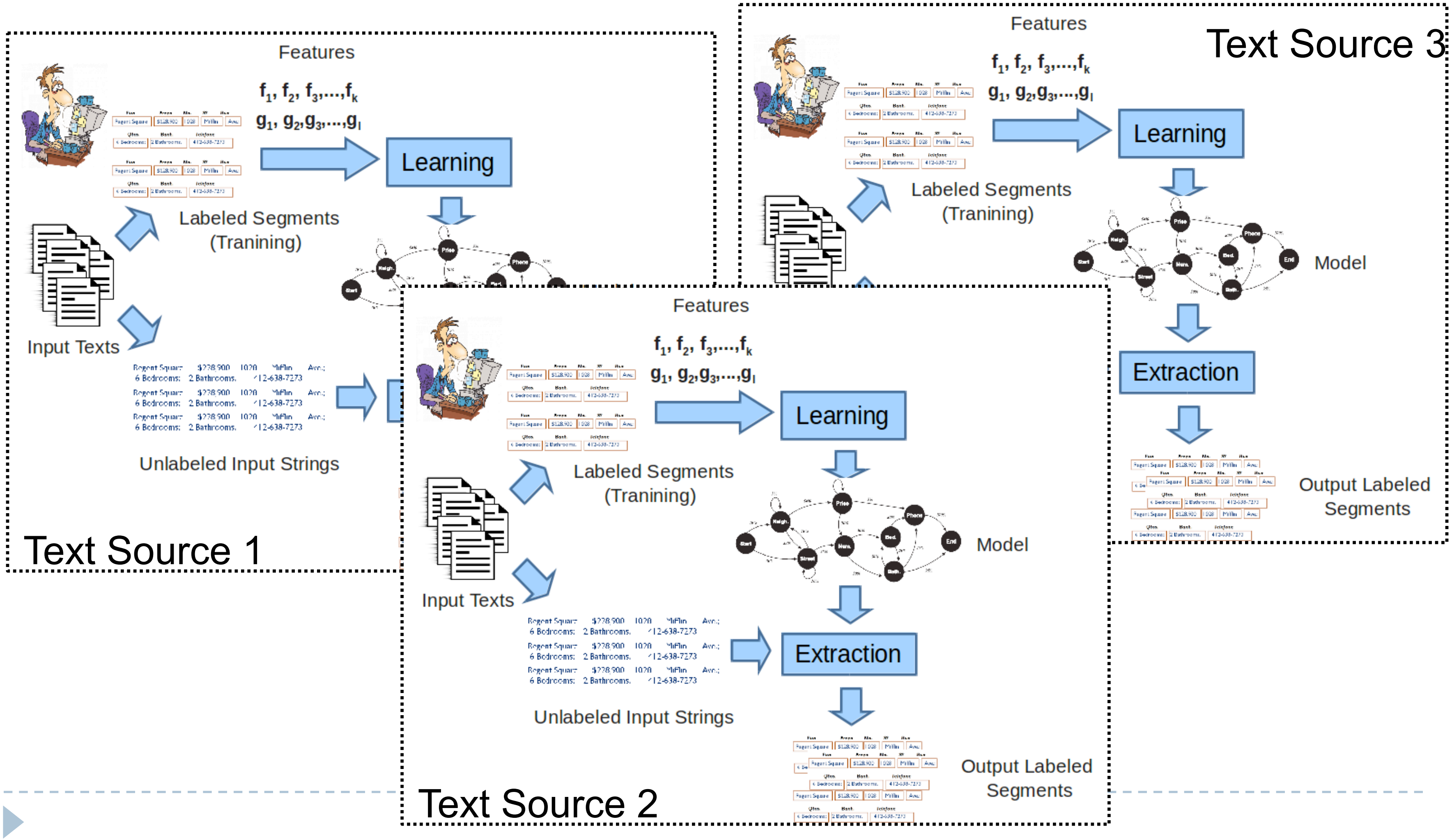
Regent Square $228,900 1028 Mifflin Ave.;
6 Bedrooms; 2 Bathrooms. 412-638-7273
Regent Square $228,900 1028 Mifflin Ave.;
6 Bedrooms; 2 Bathrooms. 412-638-7273
Regent Square $228,900 1028 Mifflin Ave.;
6 Bedrooms; 2 Bathrooms. 412-638-7273

Unlabeled Input Strings

Rua Preço No. ??? Rua
Regent Square $228,900 1028 Mifflin Ave.;
Qtos. Banh. Telefone
6 Bedrooms; 2 Bathrooms. 412-638-7273

Output Labeled
Segments

Supervised IETS



Text Source 1

Features

$f_1, f_2, f_3, \dots, f_k$

$g_1, g_2, g_3, \dots, g_l$

Learning

Labeled Segments (Tranining)

Input Texts

Unlabeled Input Strings

Text Source 3

Features

$f_1, f_2, f_3, \dots, f_k$

$g_1, g_2, g_3, \dots, g_l$

Learning

Labeled Segments (Tranining)

Model

Extraction

Output Labeled Segments

Text Source 2

Features

$f_1, f_2, f_3, \dots, f_k$

$g_1, g_2, g_3, \dots, g_l$

Learning

Labeled Segments (Tranining)

Model

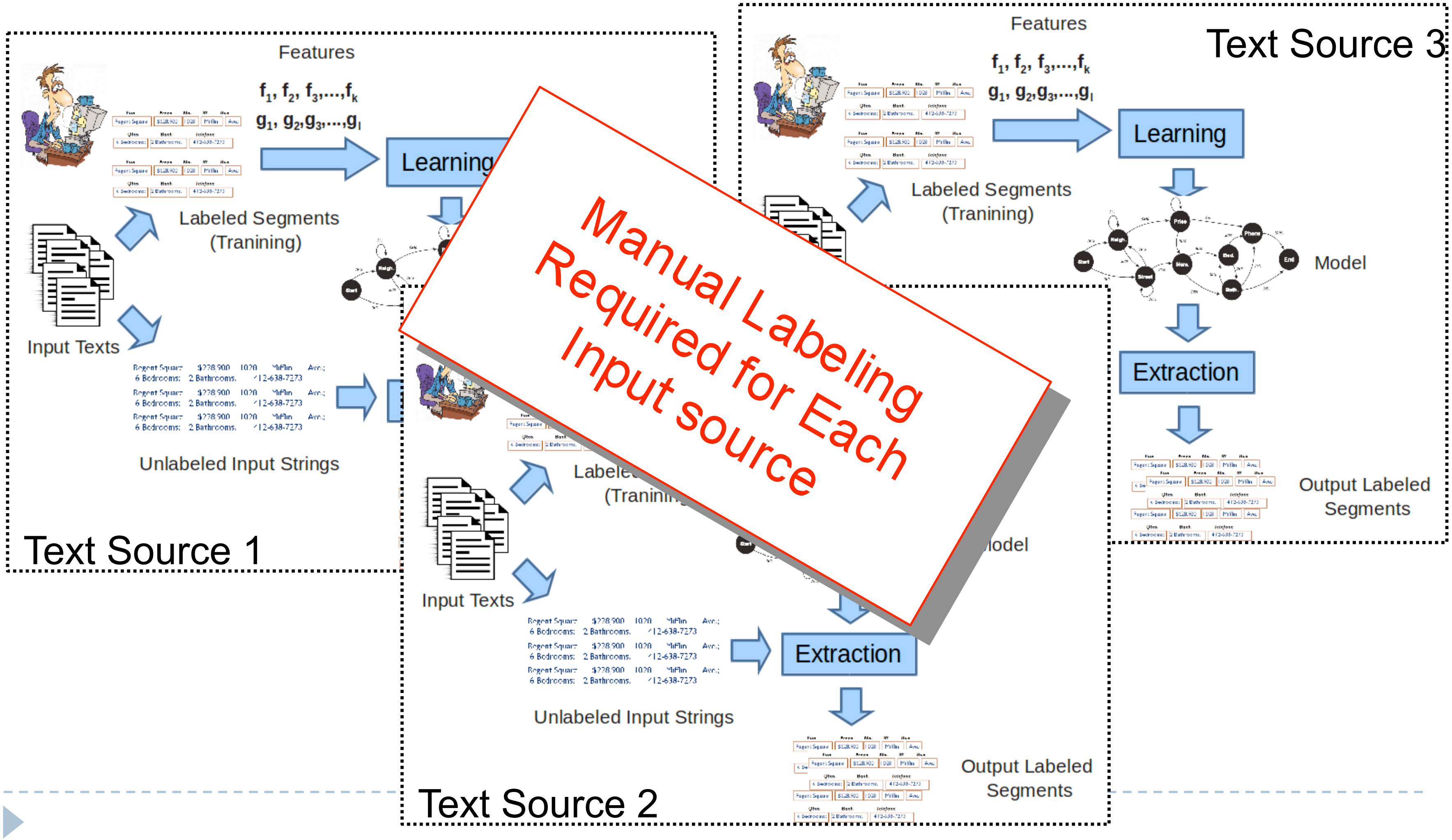
Input Texts

Extraction

Unlabeled Input Strings

Output Labeled Segments

Supervised IETS

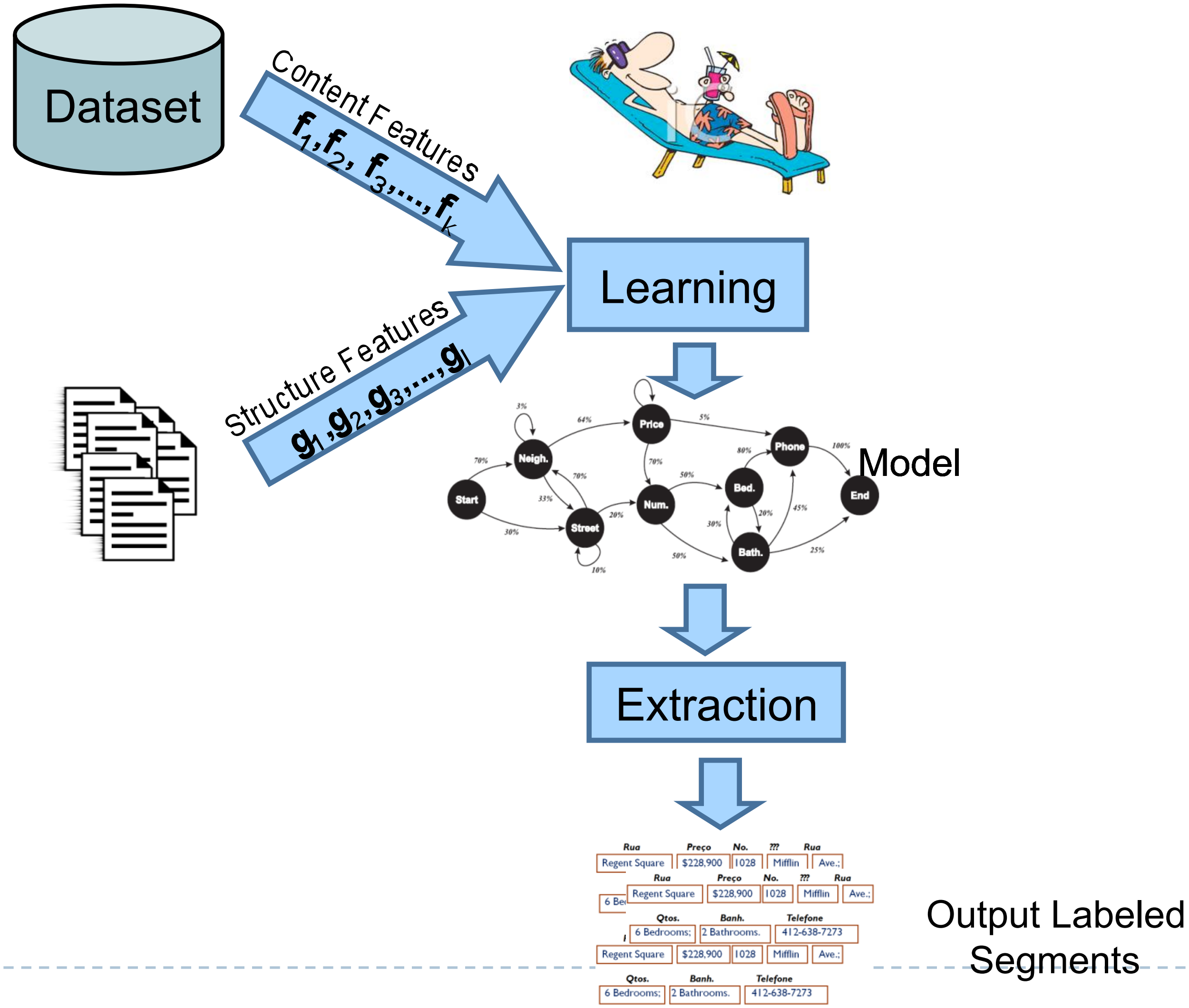


Manual Labeling Required for Each Input source

Unsupervised IETS methods

- Learn from datasets
  - Dictionaries, knowledge bases, references tables, etc.
- No need for manual training for each input
- Source Independent
- IETS methods
  - Unsup. CRF (Zhao et al. @SIAM ICDM'08)
  - ONDUX (Cortez et al. @SIGMOD'10)
  - JUDIE (Cortez et al. @SIGMOD'11)

Unsupervised IETS
ONDUX & JUDIE



Dataset

Content Features $f_1, f_2, f_3, \ldots, f_k$

Structure Features $g_1, g_2, g_3, \ldots, g_l$

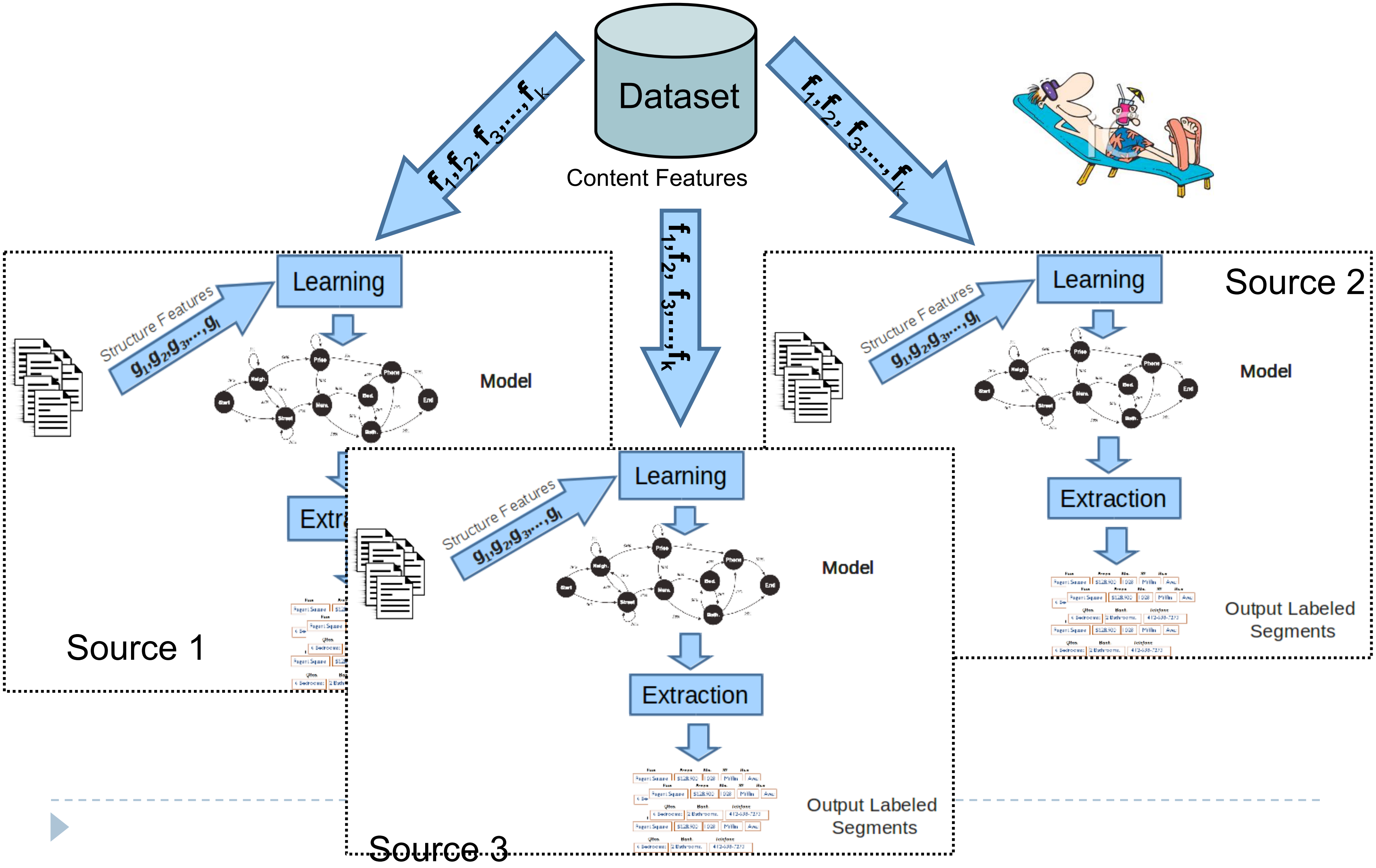
Learning

Model

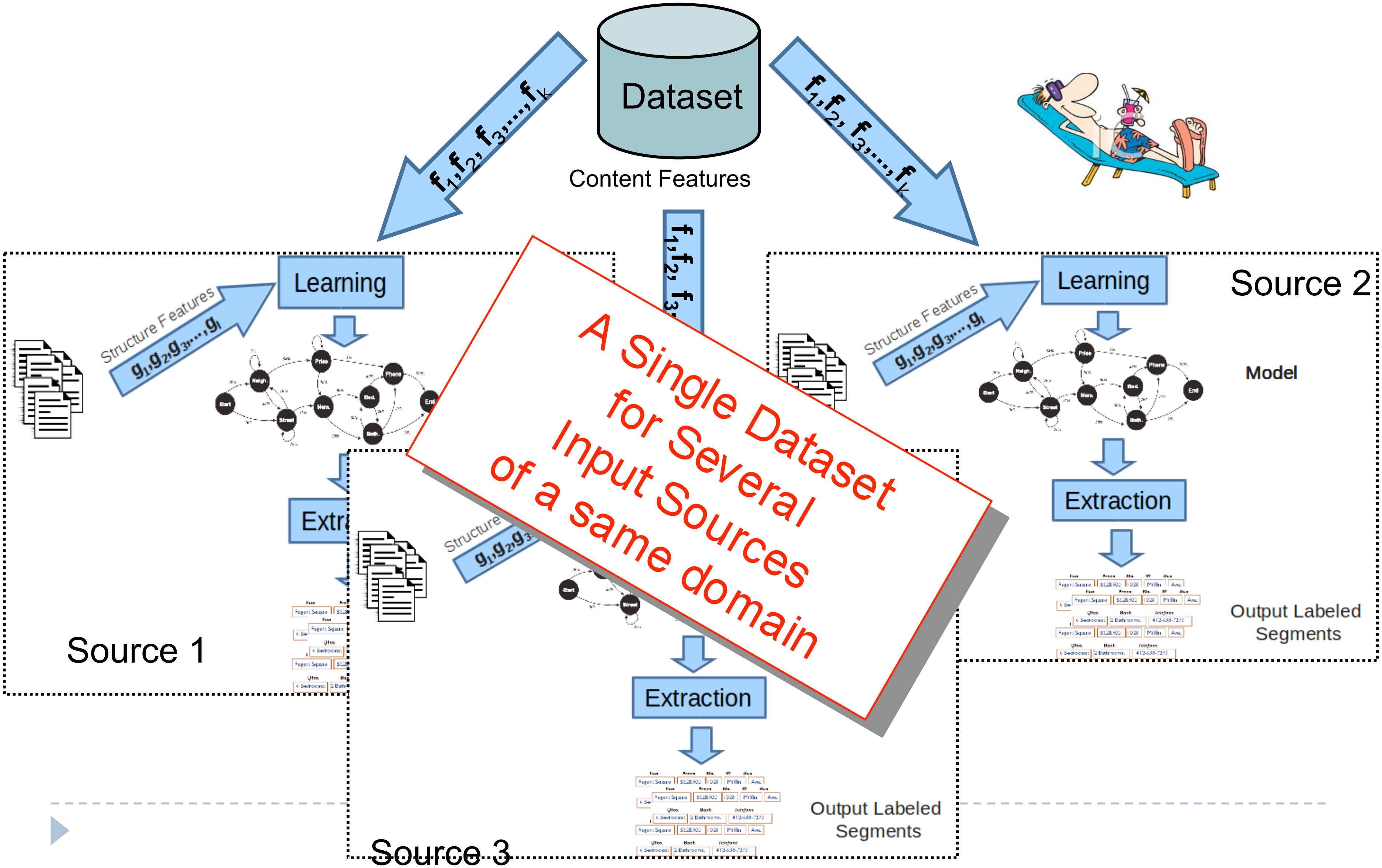
Extraction

Output Labeled Segments

Unsupervised IETS - ONDUX & JUDIE



Unsupervised IETS - ONDUX & JUDIE



A Single Dataset for Several Input Sources of a same domain

Features

- IETS methods rely on two types of features:
- Content (or state) features:
  - Related to the contents of the tokens/strings
- Structure (or transition) features:
  - Related to the location of tokens/strings in a sequence

Content Features we use

- Vocabulary:
  - Similarity betweew strings in the input and values of an attribute from the KB
- Value Range:
  - How close a numeric string in the input is from the mean value of a set of numeric values of an attribute in the KB
- Format:
  - Common style often used to represent values of some attributes
  - URLs, e-mails, telephone numbers, etc

Structure Features we use

- Features
  - Positioning:
    - position of the values of a given attribute within the input
  - Sequencing:
    - relative order of attribute values within the input
- Assumption:
  - Some regularity in the appearance of attribute values within the input texts
  - Does not necessarily mean assuming a fixed order of appearance

Content x Structure Features

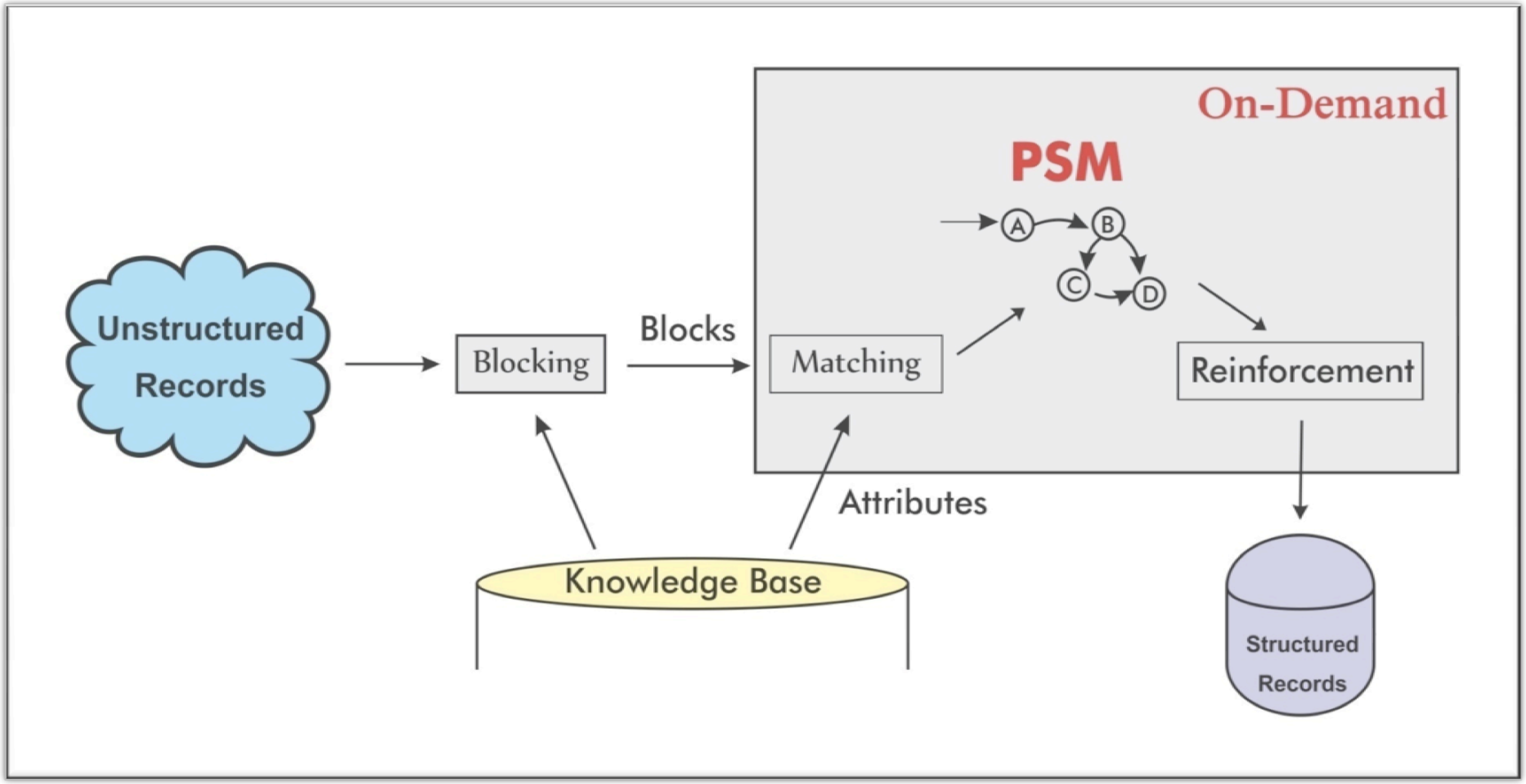
- Content Features
  - Domain-dependent but input-independent
  - For a given attribute A, can be computed from a any representative set of values in domain of A
    - e.g., from a previous existing dataset
- Structure Features
  - Dependent of the placement of attributes values on the input
  - Thus, they are input-dependent

Unsupervised IETS methods

Method	Content Features	Structure Features
Mansuri@ICDE'06	Dictionaries	Seed instances
Agichtein@SIGKDD'04	Reference Tables	Sample, assumed to have a fixed order
Zhao@SICDM'08	Reference Tables	Sample, assumed to have a fixed order
Cortez@JASIST'09	Bibliographic Files	Heuristics for the bibliographic domain
Cortez@SIGMOD'10	Knowledge Bases	Automatically Induced
Cortez@SIGMOD'11	Knowledge Bases	Automatically Induced – multiple records

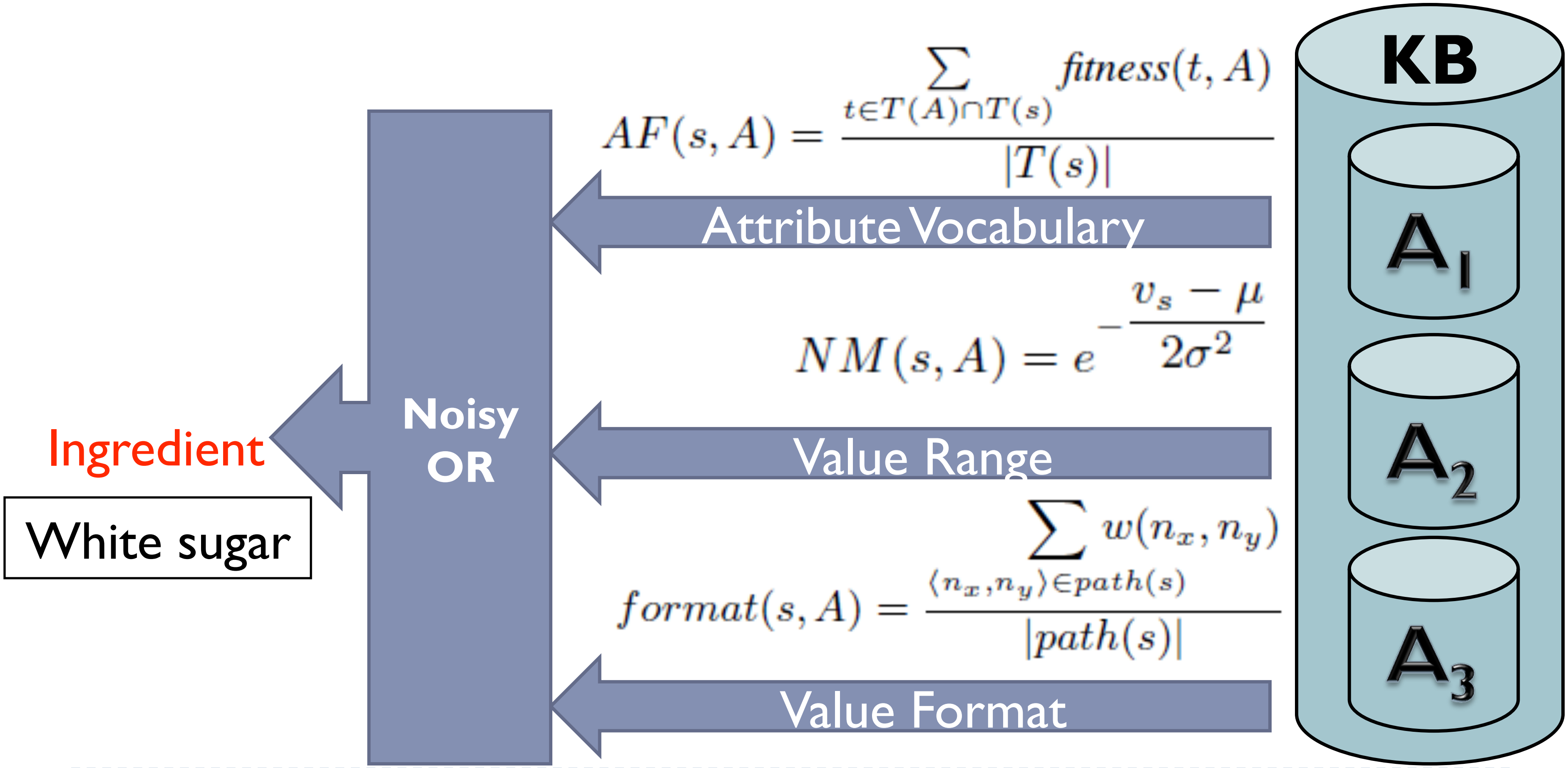
ONDUX

- General View



Features – Content Related

- Features Considered:

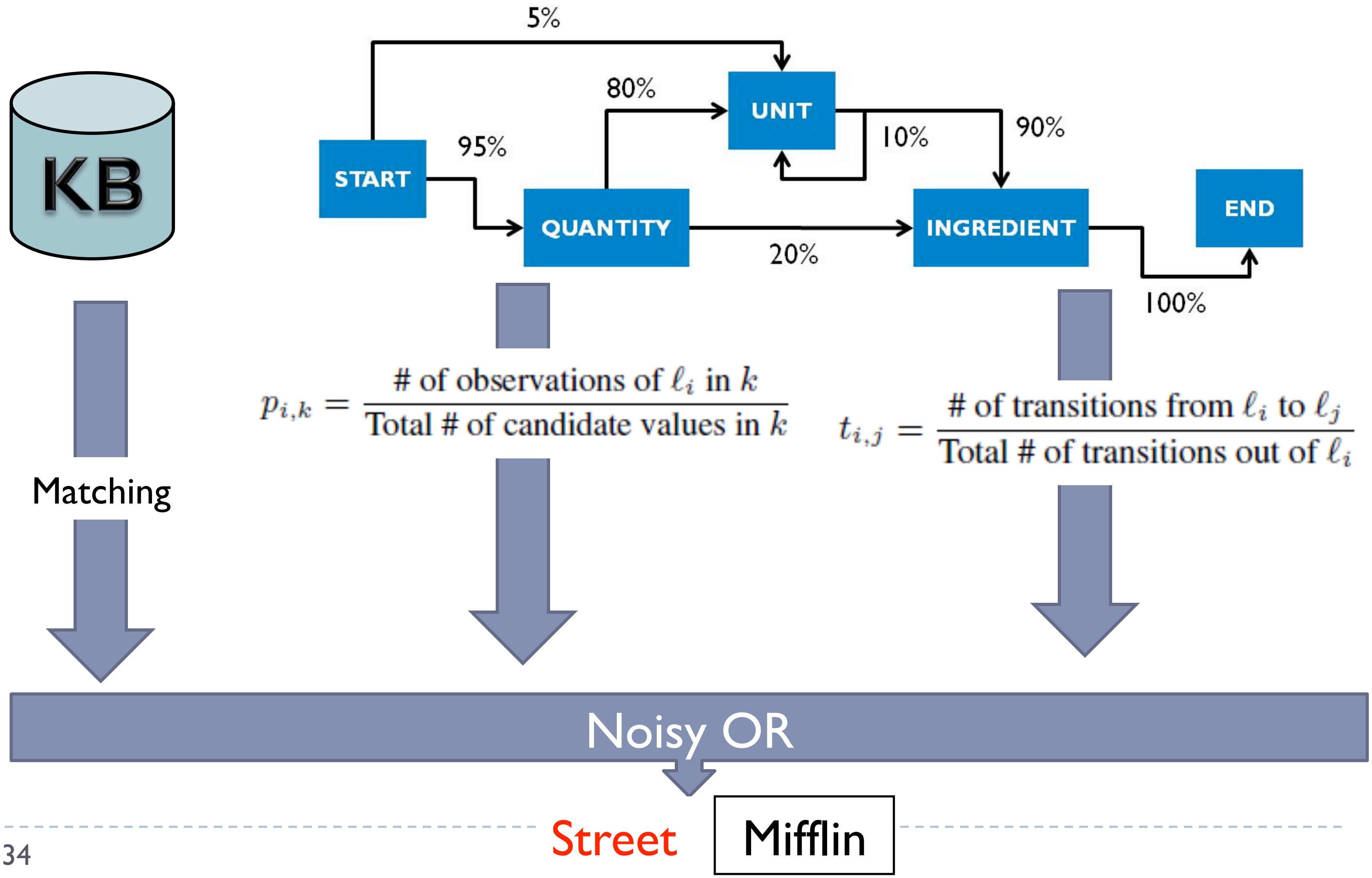


$$AF(s,A) = \frac{\sum_{t \in T(A) \cap T(s)} fitness(t,A)}{|T(s)|}$$

$$NM(s,A) = e^{-\frac{v_s - \mu}{2\sigma^2}}$$

$$format(s,A) = \frac{\sum_{\langle n_x, n_y \rangle \in path(s)} w(n_x, n_y)}{|path(s)|}$$

Adding Structure Related Features



$$p_{i,k} = \frac{\text{\# of observations of } \ell_i \text{ in } k}{\text{Total \# of candidate values in } k}$$

$$t_{i,j} = \frac{\text{\# of transitions from } \ell_i \text{ to } \ell_j}{\text{Total \# of transitions out of } \ell_i}$$

Matching

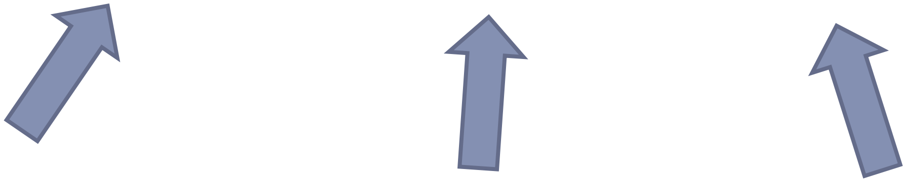
Noisy OR

Street Mifflin

ONDUX

- Reinforcement
  - Once the PSM is built, we combine the matching, positioning and sequencing evidences using the Bayesian operator *OR*.

$$FS(B,a_i)=1-((1-M(B,a_i))\times(1-t_{j,i})\times(1-p_{i,k}))$$

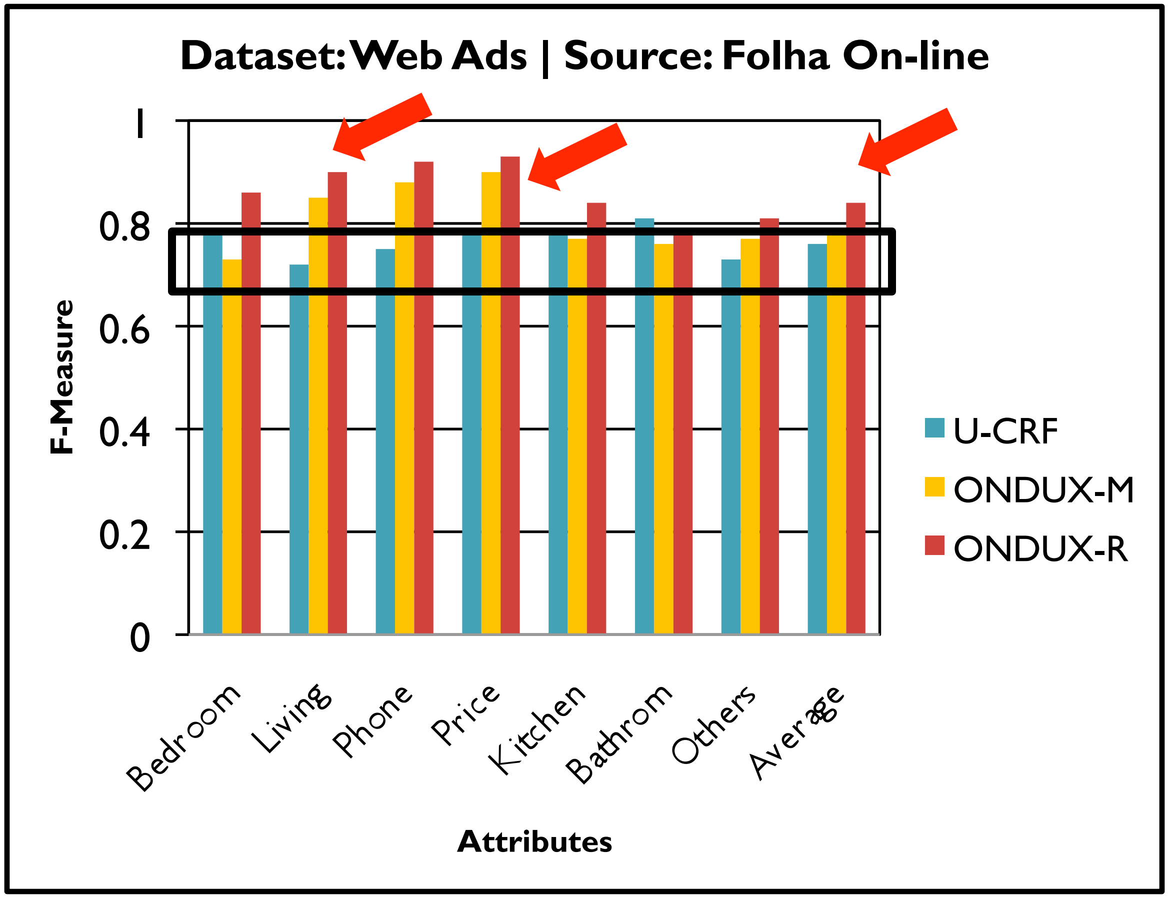


Matching Result

Sequence

Positioning

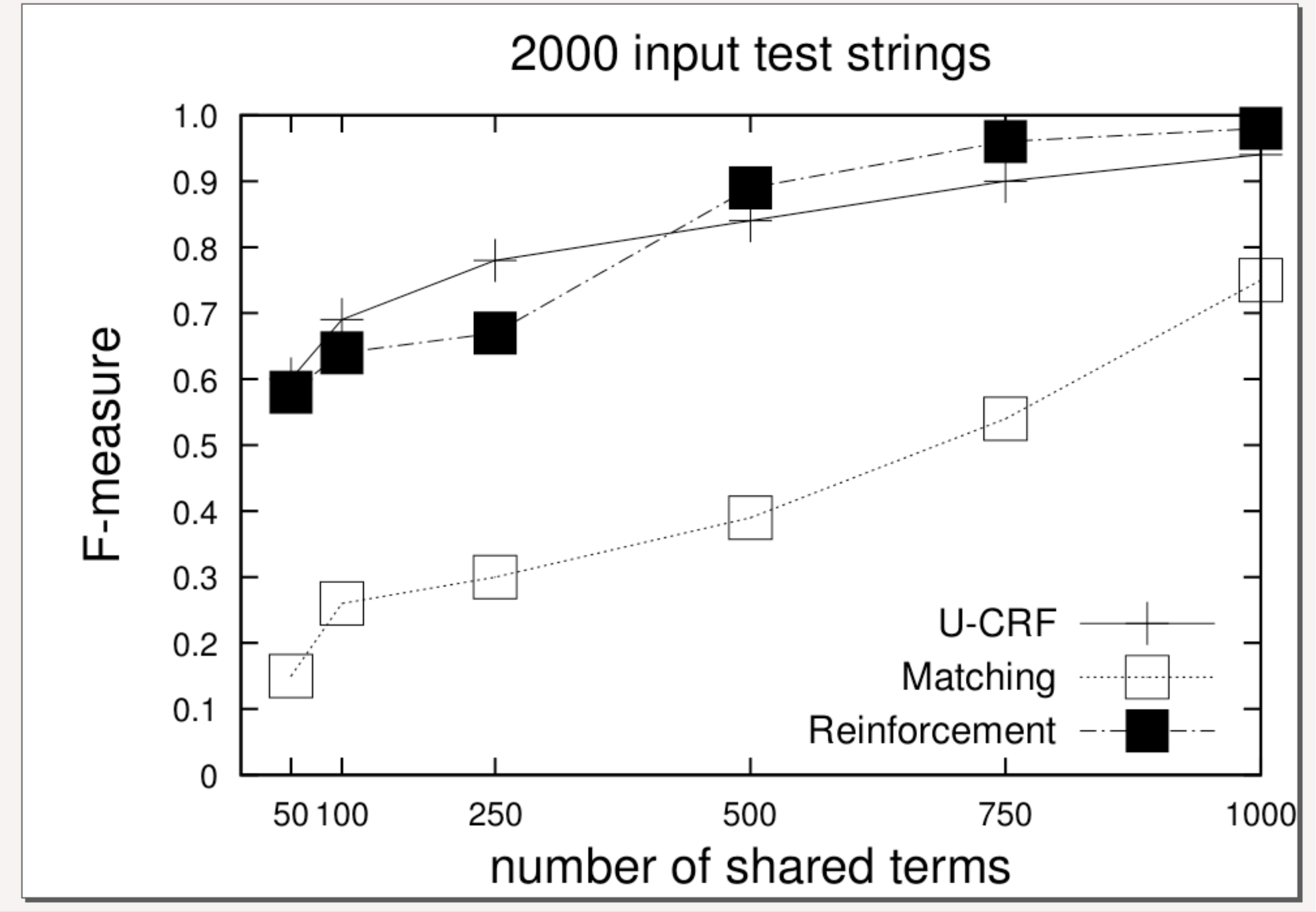
Experimental Results



U-CRF presented a poor performance (very heterogeneous dataset)

Due to the Matching Phase and the PSM that is learned *On-Demand*, ONDUX achieve very high quality results

Reinforcement



JUDIE

Chocolate Cake Recipe

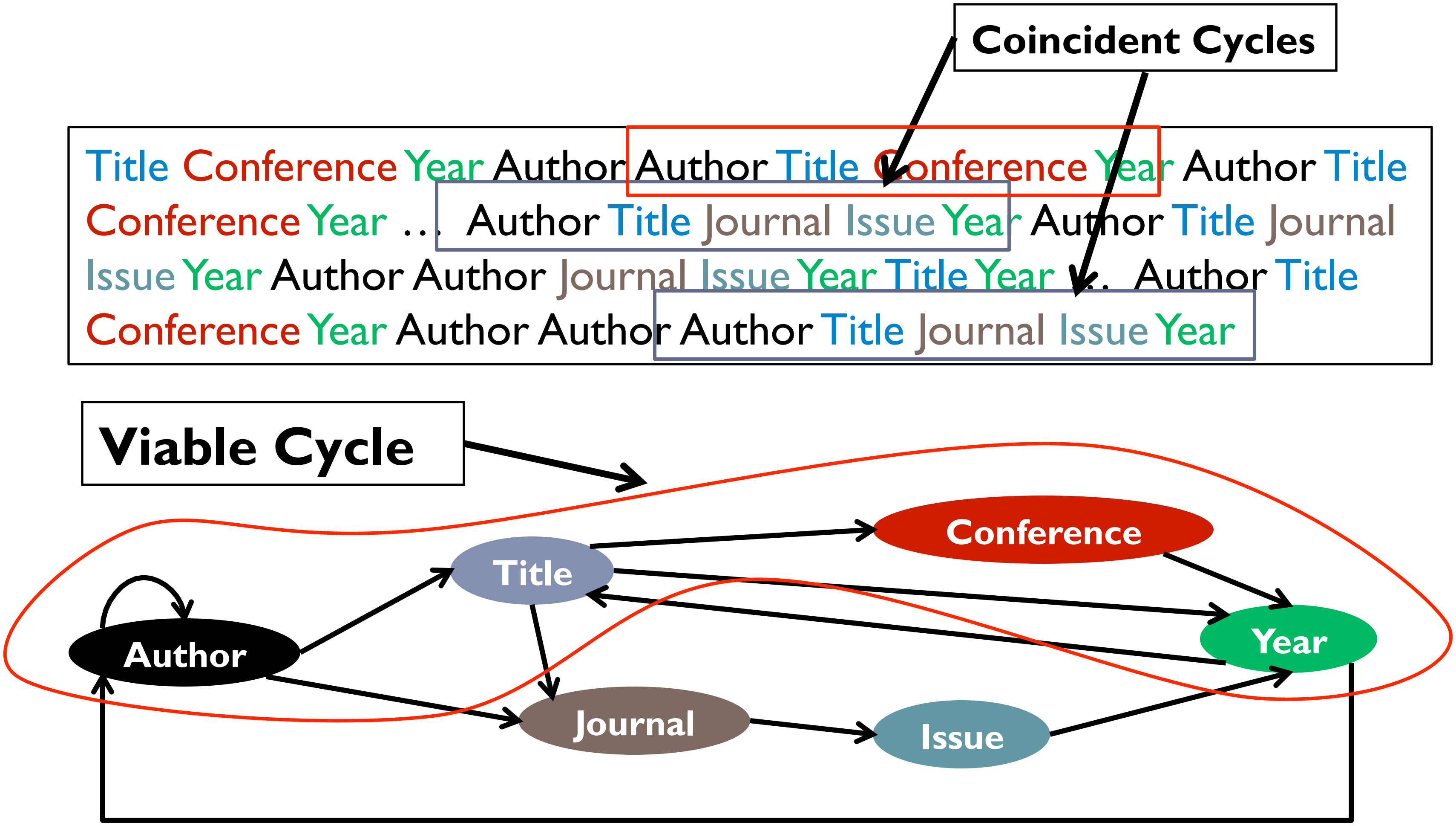
1/2 cup butter 2 eggs 4 cups white sugar ground cinnamon 2 tablespoons dark rum 6 chopped pecans 1/2 cup milk 1 1/2 cups applesauce 2 cups all-purpose flour 1/4 cup cocoa powder 2 teaspoons baking soda 1/8 teaspoon salt 1 cup raisins 1/4 cup dark rum

Quantity	Unit	Ingredient
1/2	cup	butter
2		eggs
4	cups	white sugar
		ground cinnamon
2	tablespoons	dark rum
6		chopped pecans

JUDIE

- Joint Unsupervised Structure Discovery and Information Extraction
  - Detects the structure of each individual record being extracted without any user intervention
  - Looks for frequent patterns of label repetitions or **cycles**
- Integrates this algorithm in the IE process
  - Accomplished by successive refinement steps that alternate information extraction and structure discovery.

The SD Algorithm



Coincident Cycles

Title Conference Year Author Author Title Conference Year Author Title Conference Year … Author Title Journal Issue Year Author Title Journal Issue Year Author Author Journal Issue Year Title Year …. Author Title Conference Year Author Author Author Title Journal Issue Year

Viable Cycle

Comparison with baselines – Attribute Level

Attribute	JUDIE	ONDUX	U-CRF
Author	0.88	0.922	0.87
Title	0.70	0.79	0.69
Booktitle	0.86	0.89	0.56
Journal	0.84	0.90	0.55
Volume	0.90	0.96	0.43
Pages	0.86	0.84	0.50
Date	0.87	0.89	0.49
Average	**0.86**	**0.88**	**0.58**

CORA

Attribute	JUDIE	ONDUX	U-CRF
Bedroom	0.82	0.86	0.79
Living	0.89	0.90	0.72
Phone	0.87	0.92	0.75
Price	0.92	0.93	0.78
Kitchen	0.83	0.84	0.78
Bathroom	0.77	0.79	0.81
Others	0.73	0.79	0.71
Average	**0.84**	**0.85**	**0.76**

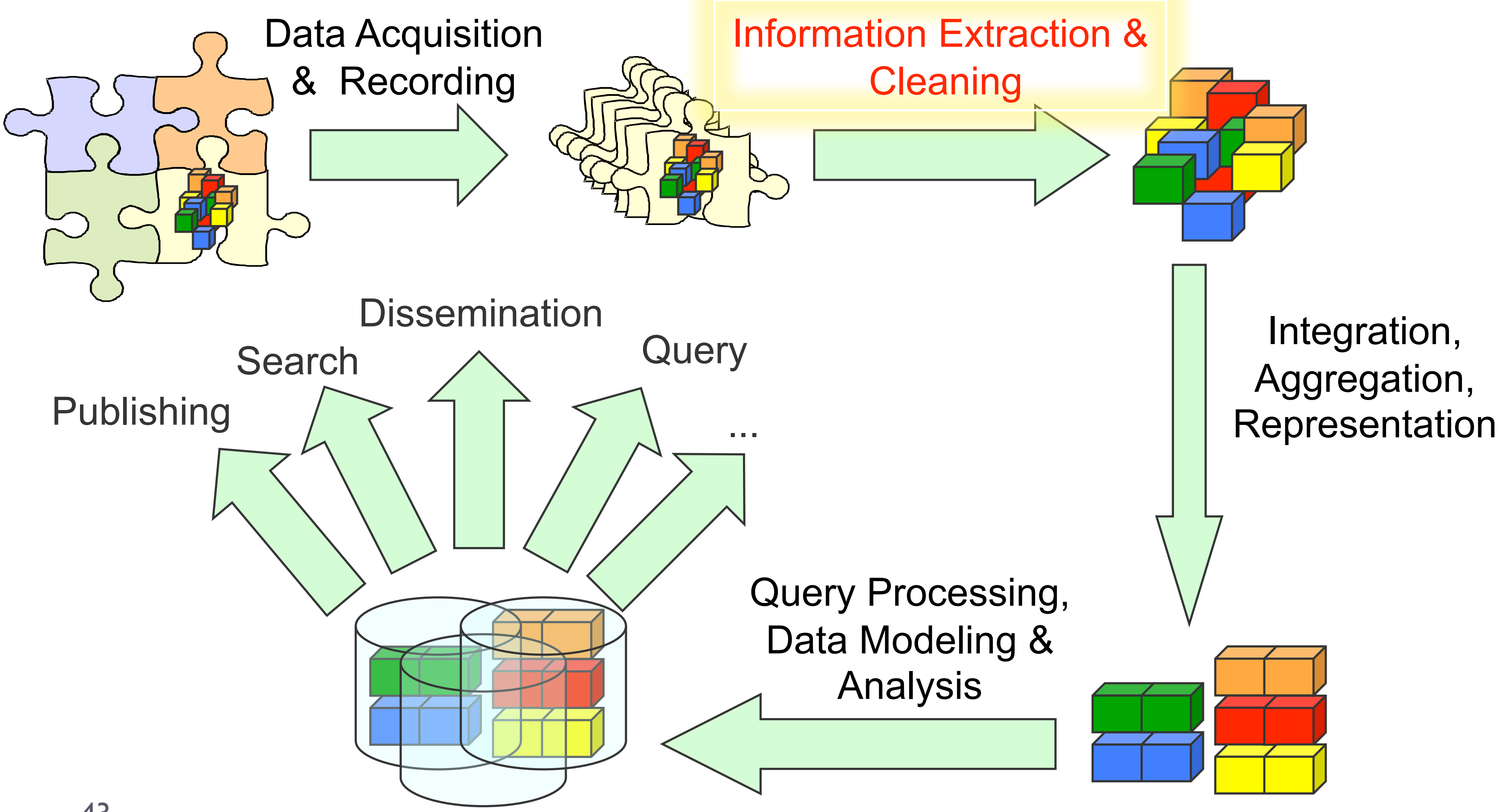
Web Ads

- Results very close to ONDUX and even better than U-CRF
- Recall: JUDIE faces a harder task.

More details ….

- Cortez, Silva, Gonçalves & Moura. *ONDUX: on-demand unsupervised learning for information extraction.* SIGMOD 2010
- Cortez, Oliveira, Silva, Moura & Laender: *Joint unsupervised structure discovery and information extraction.* SIGMOD 2011

One more …

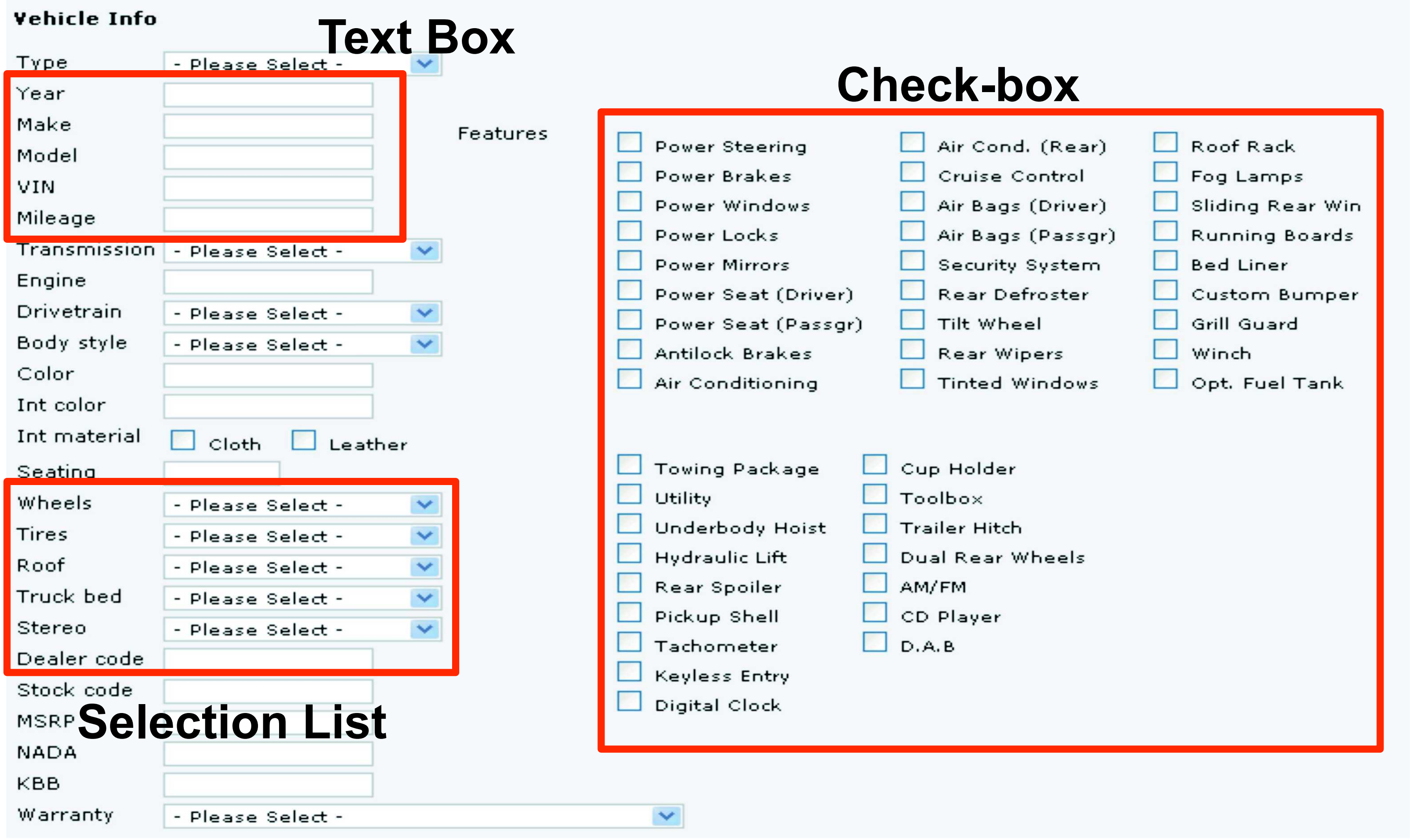

Data Acquisition
& Recording
Information Extraction &
Cleaning
Integration,
Aggregation,
Representation
Query Processing,
Data Modeling &
Analysis
Publishing
Search
Dissemination
Query
…

The Form Filling Problem

- Goal:
  - To automatically fill out the fields of a given **form-based** interface with **values extracted** from a **data-rich free text document**.
    1. Extracting values from the input text;
    2. Filling out the fields of the target form using them.

Example

- Form-based interface



Text Box

Check-box

Selection List

Vehicle Info

Type	- Please Select -
Year	
Make	
Model	
VIN	
Mileage	
Transmission	- Please Select -
Engine	
Drivetrain	- Please Select -
Body style	- Please Select -
Color	
Int color	
Int material	□ Cloth □ Leather
Seating	
Wheels	- Please Select -
Tires	- Please Select -
Roof	- Please Select -
Truck bed	- Please Select -
Stereo	- Please Select -
Dealer code	
Stock code	
MSRP	
NADA	
KBB	
Warranty	- Please Select -

Features

□ Power Steering
□ Power Brakes
□ Power Windows
□ Power Locks
□ Power Mirrors
□ Power Seat (Driver)
□ Power Seat (Passgr)
□ Antilock Brakes
□ Air Conditioning

□ Air Cond. (Rear)
□ Cruise Control
□ Air Bags (Driver)
□ Air Bags (Passgr)
□ Security System
□ Rear Defroster
□ Tilt Wheel
□ Rear Wipers
□ Tinted Windows

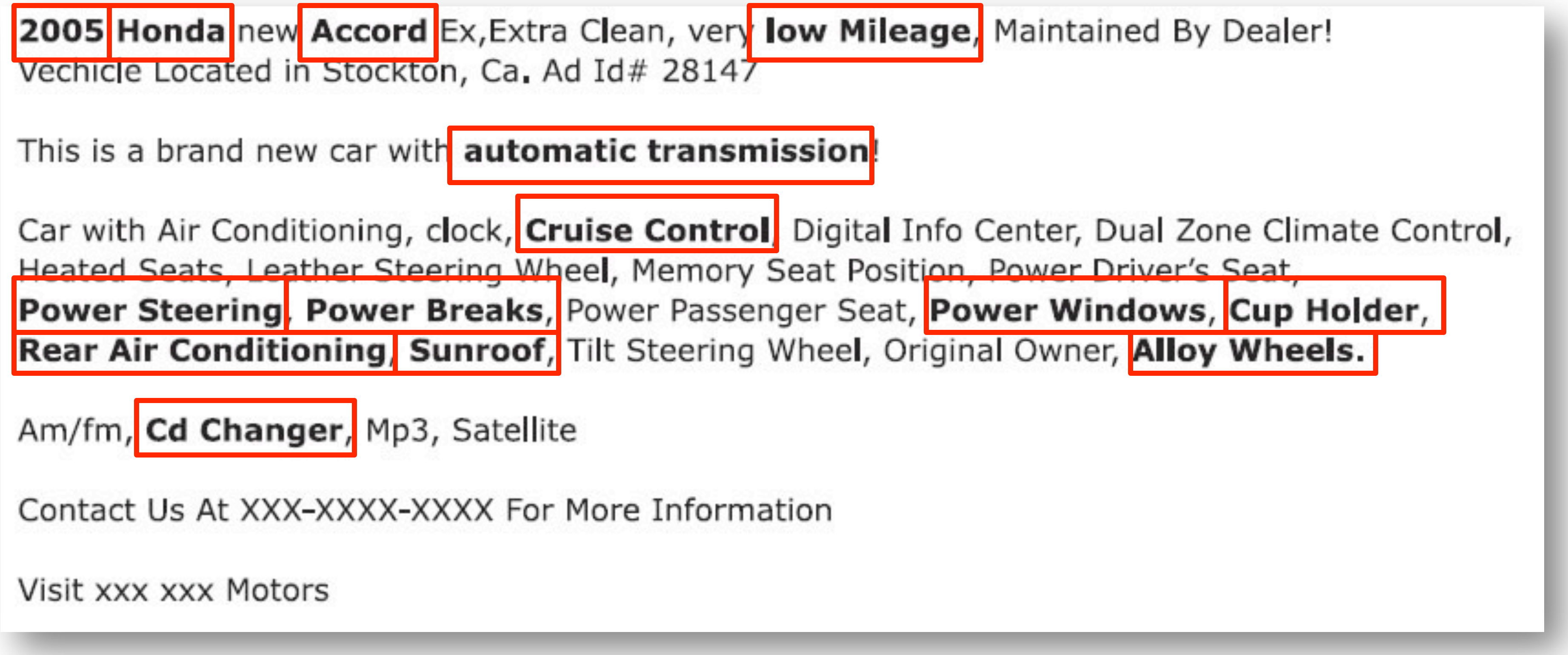
□ Roof Rack
□ Fog Lamps
□ Sliding Rear Win
□ Running Boards
□ Bed Liner
□ Custom Bumper
□ Grill Guard
□ Winch
□ Opt. Fuel Tank

□ Towing Package
□ Utility
□ Underbody Hoist
□ Hydraulic Lift
□ Rear Spoiler
□ Pickup Shell
□ Tachometer
□ Keyless Entry
□ Digital Clock

□ Cup Holder
□ Toolbox
□ Trailer Hitch
□ Dual Rear Wheels
□ AM/FM
□ CD Player
□ D.A.B

Example

- Data-rich free text document



2005 **Honda** new **Accord** Ex,Extra Clean, very **low Mileage**, Maintained By Dealer! Vechicle Located in Stockton, Ca. Ad Id# 28147

This is a brand new car with **automatic transmission**!

Car with Air Conditioning, clock, **Cruise Control**, Digital Info Center, Dual Zone Climate Control, Heated Seats, Leather Steering Wheel, Memory Seat Position, Power Driver's Seat, **Power Steering**, **Power Breaks**, Power Passenger Seat, **Power Windows**, **Cup Holder**, **Rear Air Conditioning**, **Sunroof**, Tilt Steering Wheel, Original Owner, **Alloy Wheels.**

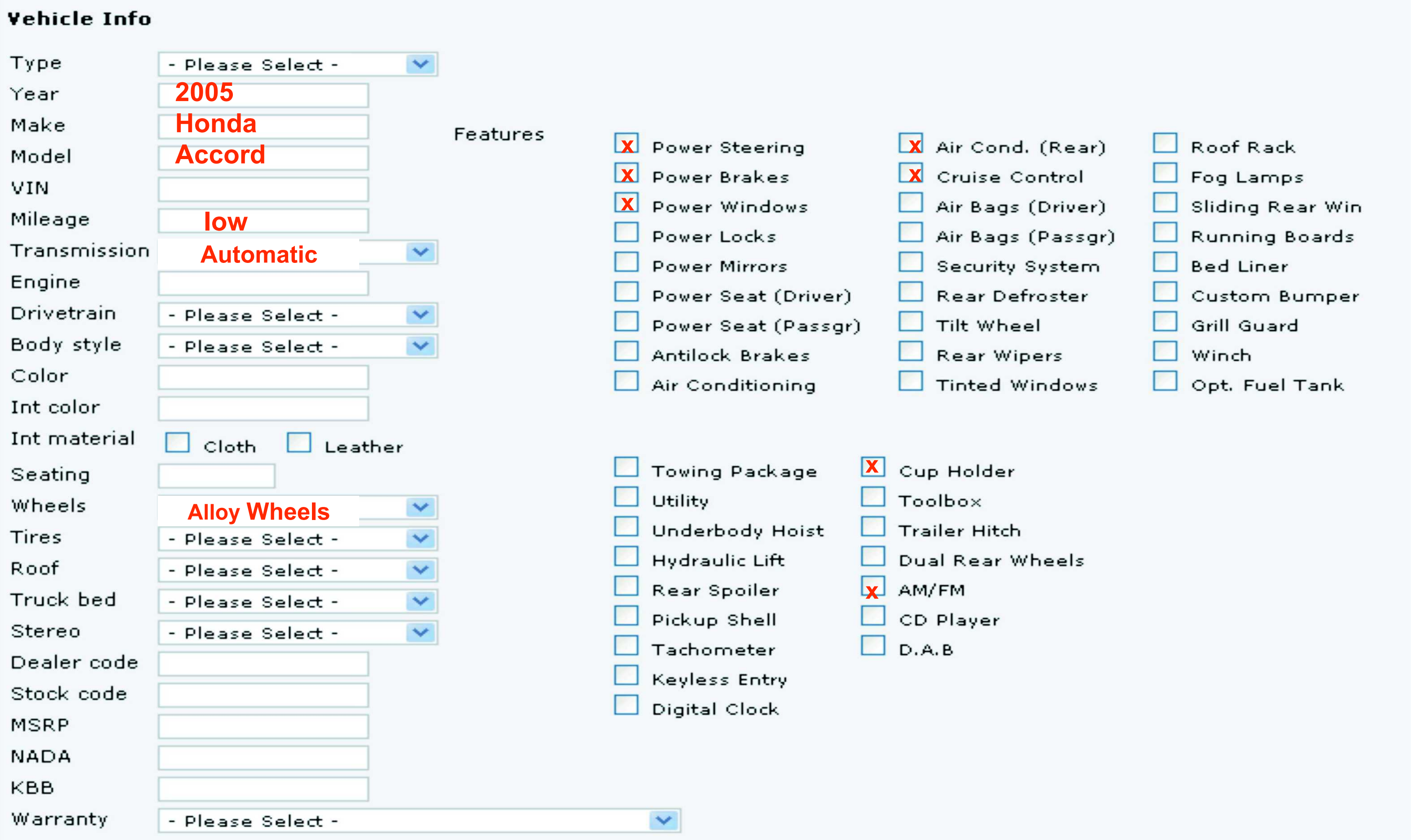
Am/fm, **Cd Changer**, Mp3, Satellite

Contact Us At XXX-XXXX-XXXX For More Information

Visit xxx xxx Motors

Example

- Form Filling



Vehicle Info

Type: - Please Select -
Year: 2005
Make: Honda
Model: Accord
VIN:
Mileage: low
Transmission: Automatic
Engine:
Drivetrain: - Please Select -
Body style: - Please Select -
Color:
Int color:
Int material: ☐ Cloth ☐ Leather
Seating:
Wheels: Alloy Wheels
Tires: - Please Select -
Roof: - Please Select -
Truck bed: - Please Select -
Stereo: - Please Select -
Dealer code:
Stock code:
MSRP:
NADA:
KBB:
Warranty: - Please Select -

Features

☒ Power Steering
☒ Power Brakes
☒ Power Windows
☐ Power Locks
☐ Power Mirrors
☐ Power Seat (Driver)
☐ Power Seat (Passgr)
☐ Antilock Brakes
☐ Air Conditioning

☒ Air Cond. (Rear)
☒ Cruise Control
☐ Air Bags (Driver)
☐ Air Bags (Passgr)
☐ Security System
☐ Rear Defroster
☐ Tilt Wheel
☐ Rear Wipers
☐ Tinted Windows

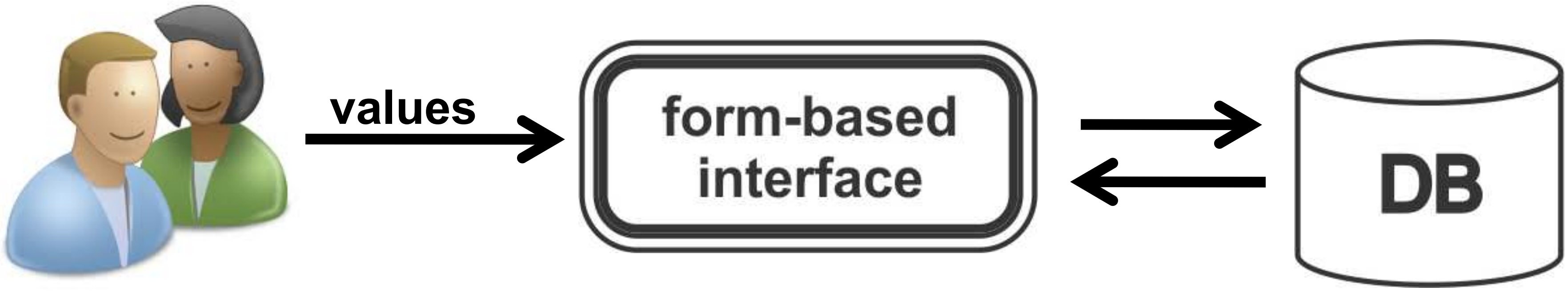
☐ Roof Rack
☐ Fog Lamps
☐ Sliding Rear Win
☐ Running Boards
☐ Bed Liner
☐ Custom Bumper
☐ Grill Guard
☐ Winch
☐ Opt. Fuel Tank

☐ Towing Package
☐ Utility
☐ Underbody Hoist
☐ Hydraulic Lift
☐ Rear Spoiler
☐ Pickup Shell
☐ Tachometer
☐ Keyless Entry
☐ Digital Clock

☒ Cup Holder
☐ Toolbox
☐ Trailer Hitch
☐ Dual Rear Wheels
☒ AM/FM
☐ CD Player
☐ D.A.B

Common usage of Web Forms

- A user manually fills each form field
  - Text-box, selection list, check-box and radio button
- Tedious, error prone and repetitive process



values

form-based interface

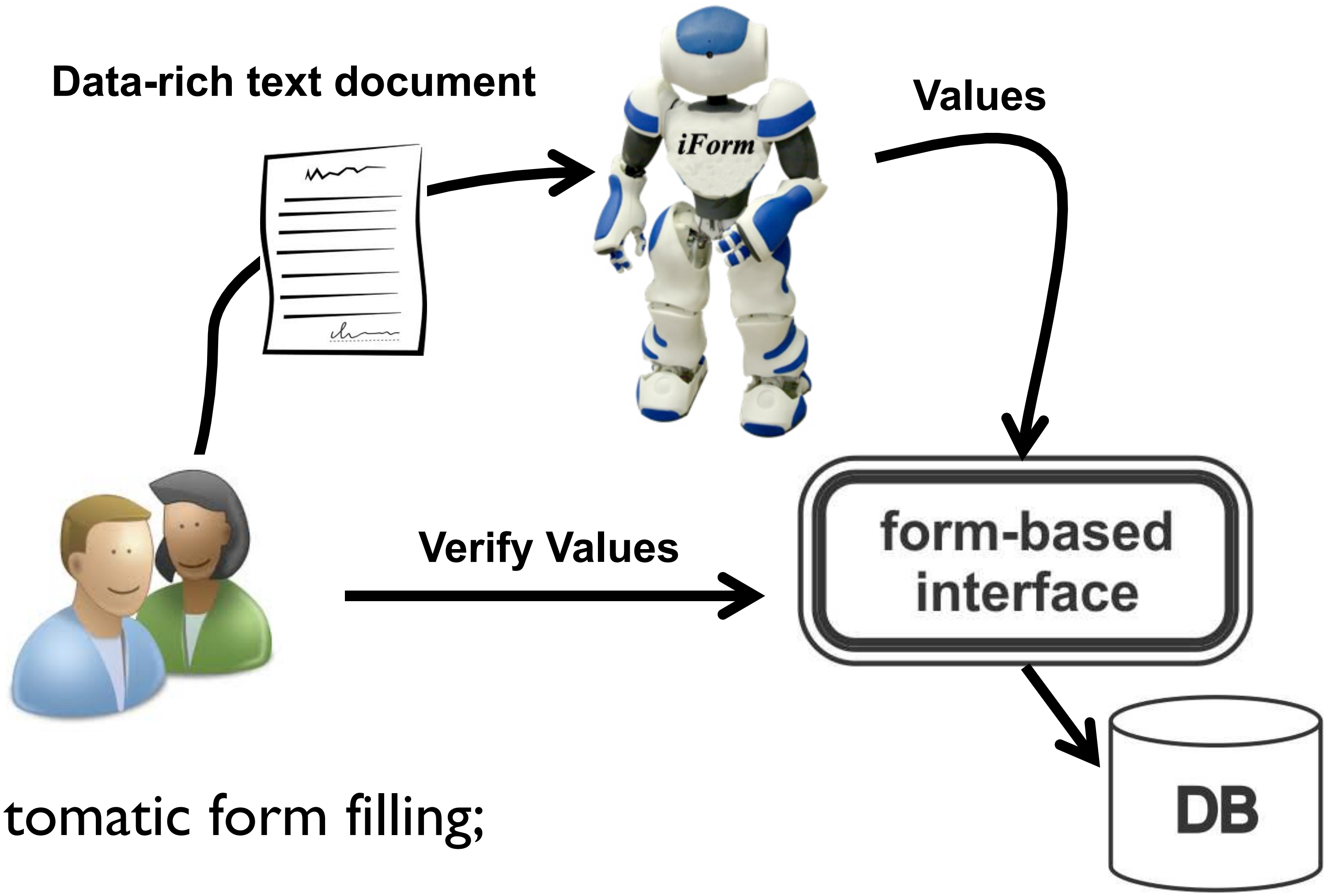
DB

Our Aproach

- IForm: Information Extraction + Form Filling
- A Probabilistic Approach for Automatically Filling Form-Based Web Interfaces
  - Appeared in PVLBD 2010 / VLDB 2011
  - With Guilherme Toda, Eli Cortez and Edleno Moura

iForm

- Information Extraction + Form Filling



- Automatic form filling;

iForm

- A **probabilistic** approach for **automatically filling** form-based interface

- Relies on a model that estimates the probability of each field in the form given the input text based on the **values previously used** for filling the form.

- Exploits features related to the **content** and **style**, which are combined through a **Bayesian framework**
  - tokens (words) composing each segment
  - wording style of each segment

Related Work – Information Extraction

- CRF (*Conditional Random Fields*): state-of-the-art information extraction approach

- Lafferty, J. et al [ICML,2001]
- Peng and McCallum [IPM, 2006]
- Mansuri and Sarawagi [ICDE, 2006]
- Kristjansson et al [IAAA, 2004]

- Usually requires training instances manually labeled
- Extracts all segments in a input text
  - Iform extracts only relevant segments

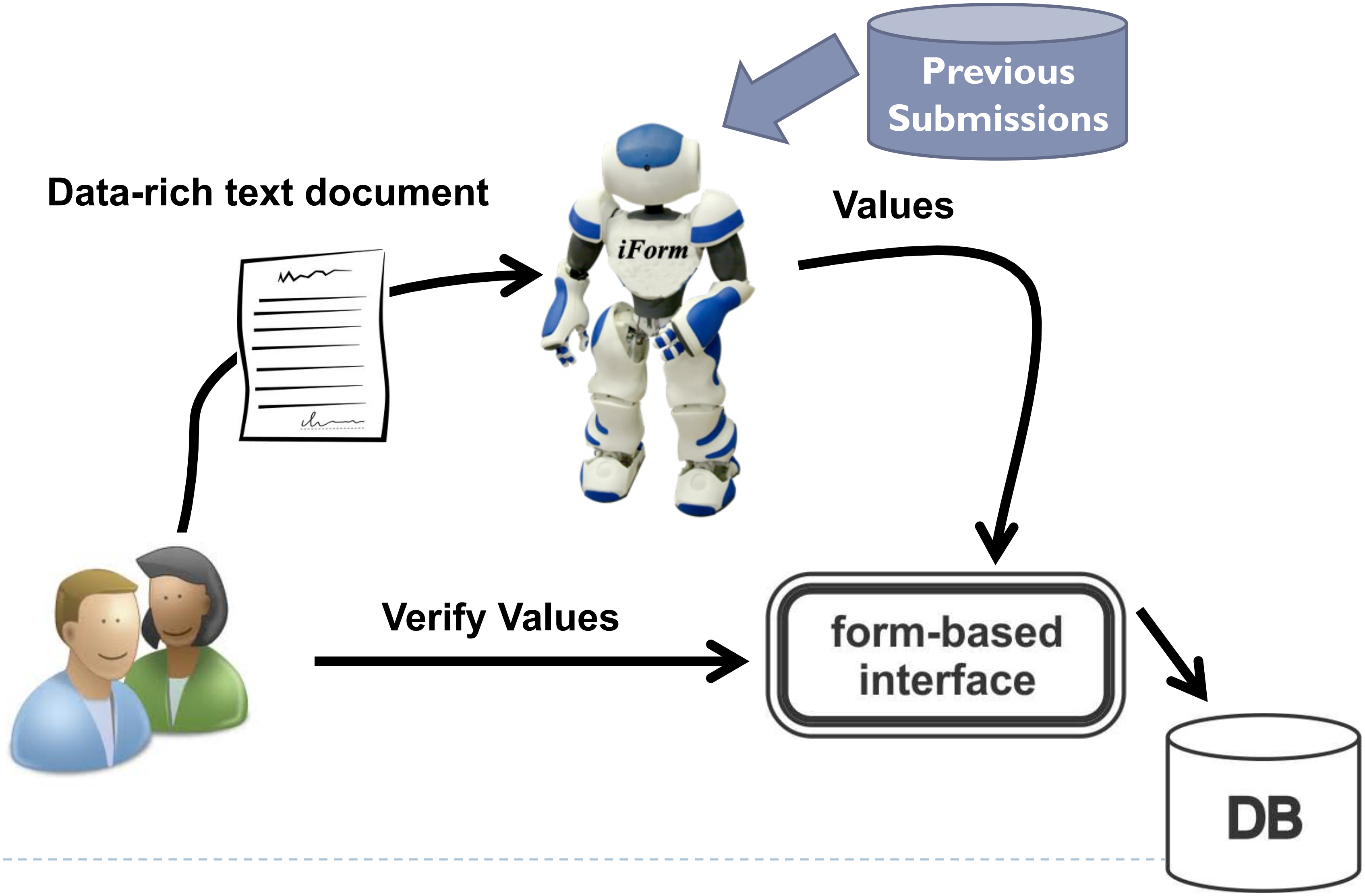
Related Work – Form Filling

- Chen et al. [ICDE, 2010]
  - USHER, a system used to automatically **adapt the form design** according to user experience.

- M. Al-Muhammed e Embley D. [ICDE,2007]
  - An approach that relies on a **manually built** ontology to guide the user in the form filling process.

- iCRF - Kristjansson et al [IAAA, 2004] - Baseline
  - CRF approach for the task of automatically filling web forms.
  - Relies on content and positioning features extracted from training instances
  - Model requires **training instances** to be **manually labeled**.

iForm - Overview

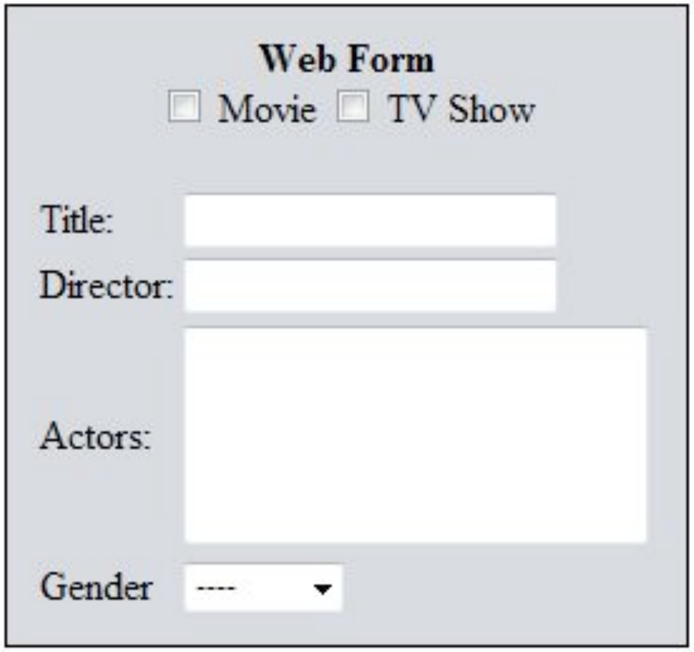


iForm - Scenario

Shutter Island is a 2010 American psychological thriller film directed by Martin Scorsese. The film is based on Dennis Lehane's 2003 novel of the same name . Starring Leonardo DiCaprio, Mark Ruffalo and Ben Kingsley.

Movie Review - Data-rich text

Web Form



Web Form

□ Movie □ TV Show

Title:

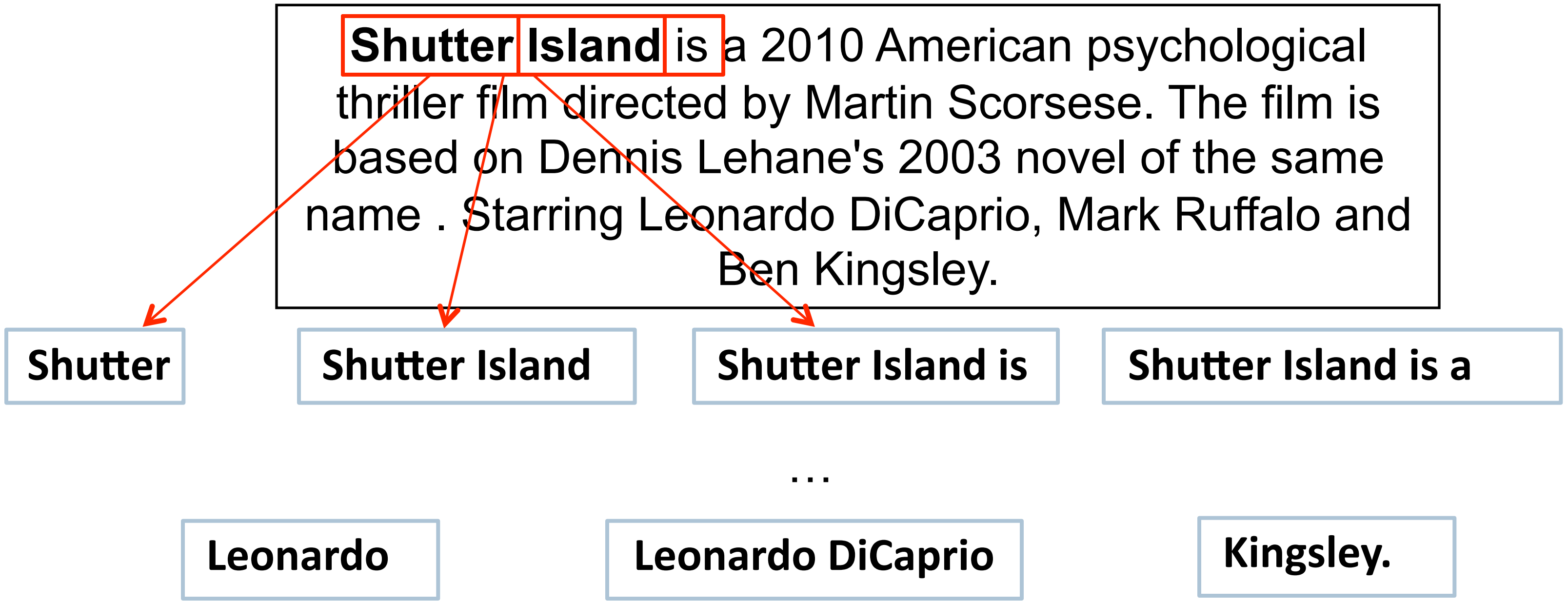
Director:

Actors:

Gender ---- ▾

iForm – Selecting plausible segments

- *What is the* probability of a form field given each *text segment?*



Shutter Island is a 2010 American psychological thriller film directed by Martin Scorsese. The film is based on Dennis Lehane's 2003 novel of the same name . Starring Leonardo DiCaprio, Mark Ruffalo and Ben Kingsley.

Shutter

Shutter Island

Shutter Island is

Shutter Island is a

…

Leonardo

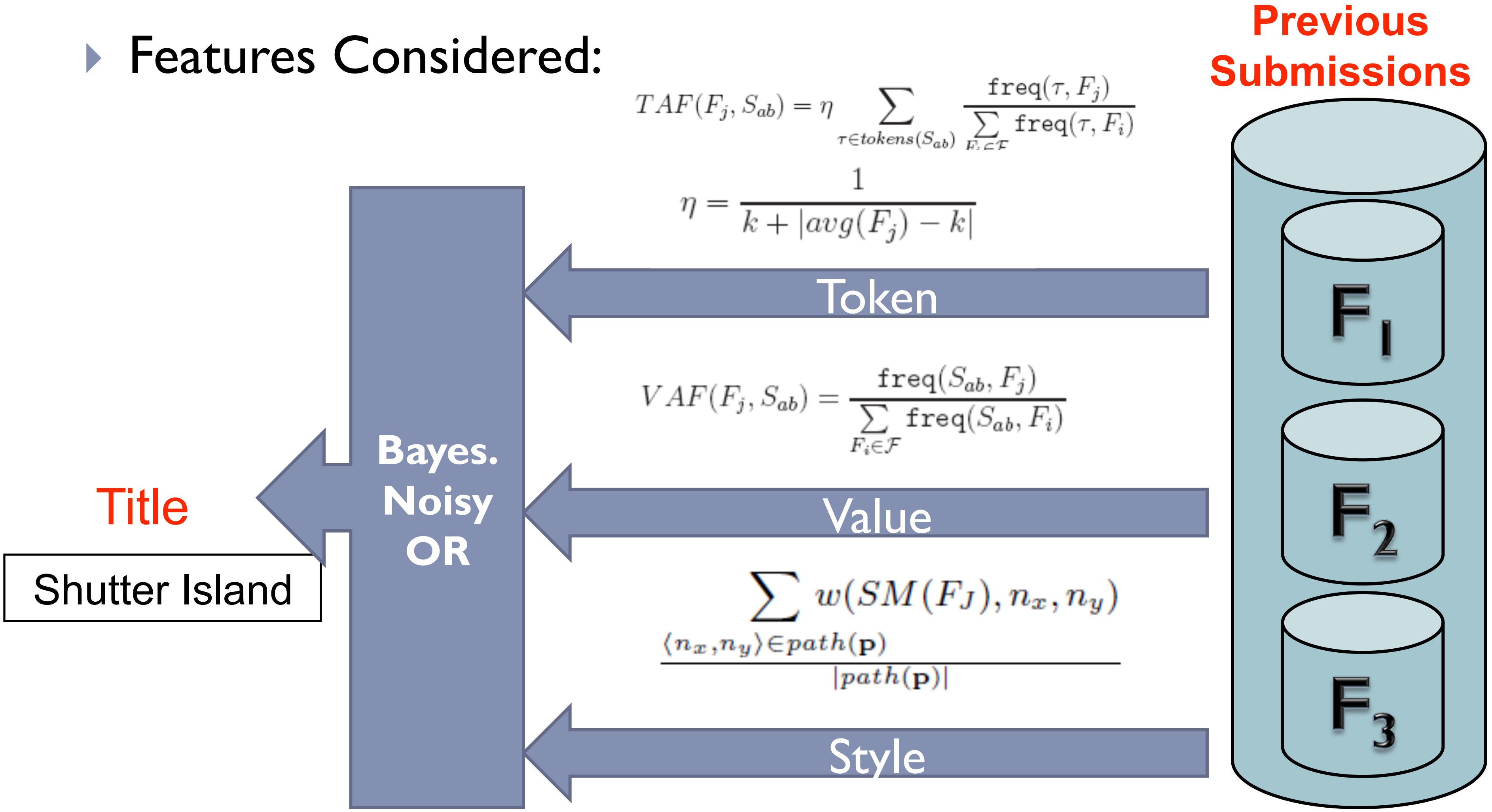
Leonardo DiCaprio

Kingsley.

Redundant computation of several probabilities can be avoided by using dynamic programming.

iForm - Features

- Features Considered:



Previous Submissions

$$TAF(F_j, S_{ab}) = \eta \sum_{\tau \in tokens(S_{ab})} \frac{\mathtt{freq}(\tau, F_j)}{\sum_{F_i \in \mathcal{F}} \mathtt{freq}(\tau, F_i)}$$

$$\eta = \frac{1}{k + |avg(F_j) - k|}$$

Token

$$VAF(F_j, S_{ab}) = \frac{\mathtt{freq}(S_{ab}, F_j)}{\sum_{F_i \in \mathcal{F}} \mathtt{freq}(S_{ab}, F_i)}$$

Value

$$\frac{\sum_{\langle n_x, n_y \rangle \in path(\mathbf{p})} w(SM(F_J), n_x, n_y)}{|path(\mathbf{p})|}$$

Style

Bayes. Noisy OR

Title

Shutter Island

F_1

F_2

F_3

iForm – Token Similarity

- Likelihood of each **token** present in the segment occurring in each field

$$TAF(F_j, S_{ab}) = \eta \sum_{\tau \in tokens(S_{ab})} \frac{\mathtt{freq}(\tau, F_j)}{\sum_{F_i \in \mathcal{F}} \mathtt{freq}(\tau, F_i)}$$

$$\eta = \frac{1}{k + |avg(F_j) - k|}$$

Average number of words of each field

Shutter **Island**

Previous Submissions

Actors	Title	Director	Genre
Joshua Jackson	Shutter	Masayuki …	Terror
Mark Man	Shutter Bug	Paul J.	Animation
Mark Rufallo	…	…	…
Leonardo DiCaprio	The Departed	Martin …	Thriller
Ewan Mcgregor,	The Island	Michael B.	Action
Marlon Brando	The Island of Dr. ..	John Frank	Terror

iForm – Value Similarity

- Likelihood of the **value** present in the segment occurring in each field

Mark Ruffalo

$$VAF(F_j, S_{ab}) = \frac{\mathtt{freq}(S_{ab}, F_j)}{\sum_{F_i \in \mathcal{F}} \mathtt{freq}(S_{ab}, F_i)}$$

Previous Submissions

Actors
Seth Rogen
Ben Affleck
Jim Carrey,
Zooey Deschanel
Ethan Hawke
Mark Ruffalo

Title	Director	Genre
Kung Fu Panda	Mark Osborne	Animation
Daredevil	Mark S. Johson	Action
…	…	…
Yes Man	Peyton Reed	Comedy
What Doesn't	Brian Goodma	Action
Zodiac	David Fincher	Thriller

iForm – Style Similarity

- Given a text segment, we encode it according to a taxonomy of symbols.

Ben Kingsley

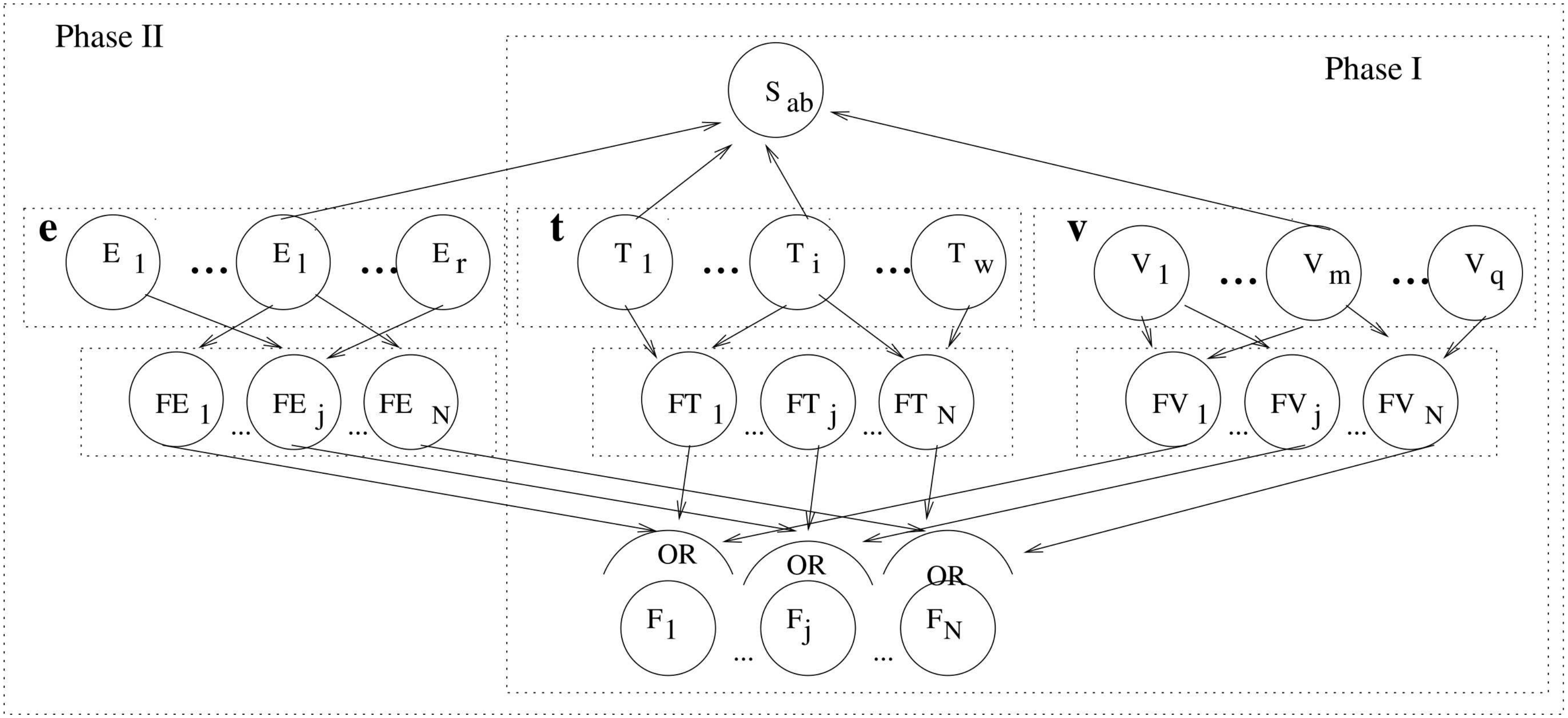
[A-Z][a-z]+   [A-Z][a-z]+

- Verifies the likelihood of the sequence following the same **wording style** of the known values for each field

$$\frac{\sum_{\langle n_x, n_y \rangle \in path(\mathbf{p})} w(SM(F_J), n_x, n_y)}{|path(\mathbf{p})|}$$

iForm – Combining all probabilities

- iForm models the computation of the **probability of a field given a segment** using a **Bayesian network**.



iForm – Mapping Segments to Fields

- Given the set of text segments such that theirs probability $P(f_j \mid S_{ab})$ is above a threshold ε
  - iForm aims at finding a ***mapping*** between candidate values and form fields with a **maximum aggregate probability**
    - Select non-overlaping segments.
- Accomplished by means of a two-phase procedure

iForm – Mapping Segments to Fields

- In the first phase, we begin by computing the candidate values for each field based only on content-based features (token + value).
  - The initial mapping is composed by the set of all candidate values C_j for all fields and contains **segment-field** pairs.
- Goal: To find a subset of segment-field pairs $\langle S_{ab}, F_j \rangle$ in the mapping whose probabilities are maximum.
  - iForm relies on a simple greedy heuristic to find an approximate solution.

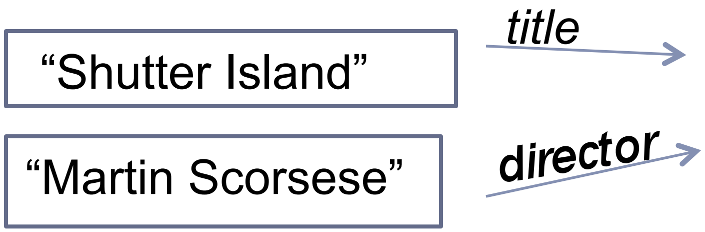
iForm – Mapping Segments to Fields

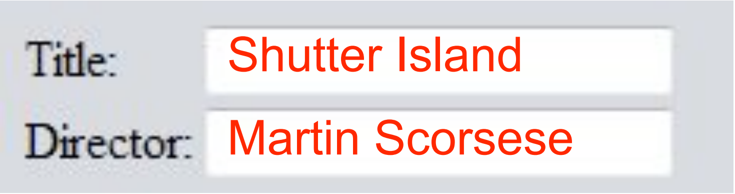
- Extracts the pair $\langle S_{ab}, F_j \rangle$ with the **highest probability** from the initial mapping and verifies if the current field was already filled with a text segment.

- To deal with fields that were not mapped to a segment, we use the probabilities derived from the **style-related features,** in the second phase**.**
  - We adopt the two phase mapping after verifying through experiments that the style-related feature is less precise than the other two features adopted.

iForm – Filling Form-based interfaces

- Uses the final mapping to fill out the form fields
  - Text Boxes: Mapped text segments as a field values.

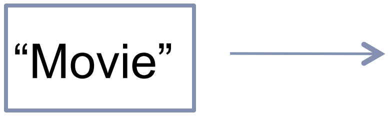


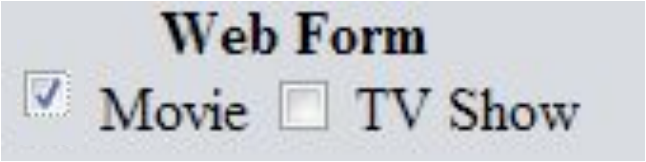


Title: Shutter Island

Director: Martin Scorsese

  - Check boxes: *Set true for mapped fields.*





Web Form

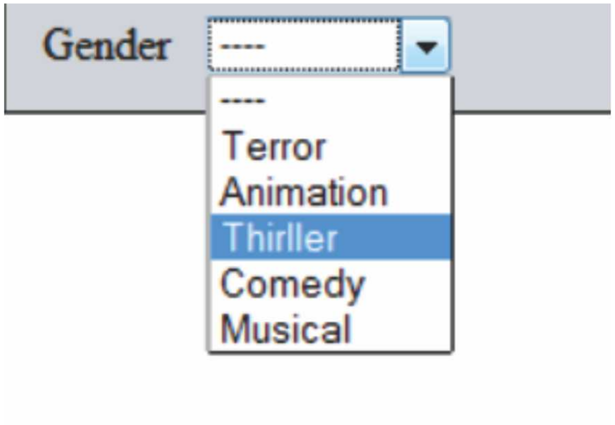
☑ Movie □ TV Show

iForm – Filling Form-based interfaces

- Selection list
  - iForm aims at finding an item such that its similarity with the extracted value is maximum – "softTF-IDF"

$$soft(A,B)=\frac{\sum_{(a,b)\in close(\theta,A,B)} w(a,A)\cdot w(b,B)\cdot s(a,b)}{\sqrt{\sum_{a\in A} w(a,A)^2}\cdot\sqrt{\sum_{b\in B} w(b,B)^2}}$$

"psychological thriller" →

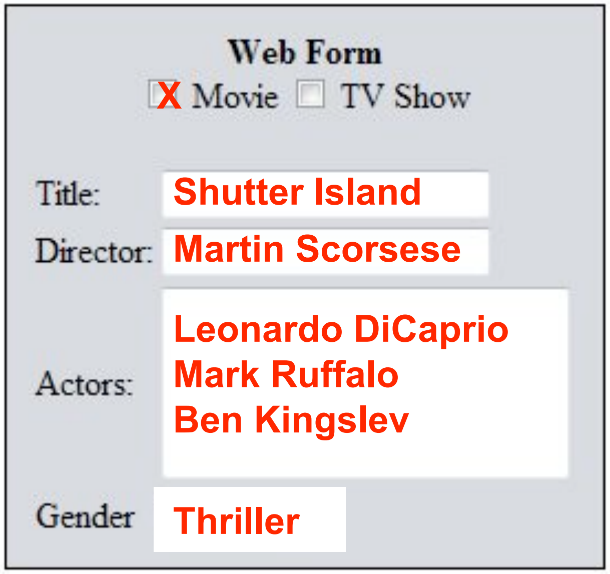


iForm - Overview

Shutter Island is a 2010 American psychological thriller film directed by Martin Scorsese. The film is based on Dennis Lehane's 2003 novel of the same name . Starring Leonardo DiCaprio, Mark Ruffalo and Ben Kingsley.

Previous Submissions

Web Form



Web Form

☒ Movie □ TV Show

Title:	Shutter Island
Director:	Martin Scorsese
Actors:	Leonardo DiCaprio Mark Ruffalo Ben Kingslev
Gender	Thriller

Evaluation – Multi-typed web forms

Movies

Type of Field	# Fields	P	R	F
Text Box	4	0.74	0.69	0.71
Submission-Level		**0.73**	**0.67**	**0.69**

iForm achieved high quality results in all datasets

Cars

Type of Field	# Fields	P	R	F
Text Box	5	0.78	0.73	0.76
Check Box	30	0.79	0.79	0.79
Average		0.79	0.78	0.79
Submission-Level		**0.77**	**0.73**	**0.75**

The quality of iForm was almost the same for the text box and the check box fields.

Evaluation – Multi-typed web forms

Cellphones

Type of Field	# Fields	P	R	F
Text Box	2	0.89	0.69	0.78
Check Box	35	0.94	0.94	0.94
Average		0.94	0.93	0.93
Submission-Level		**0.96**	**0.94**	**0.95**

Filling quality above 0.90. In fact, more than 90% of each submission was correctly entered in the web form interface.

Books 1

Type of Field	# Fields	P	R	F
Text Box	4	0.88	0.67	0.76
Drop Down	1	0.96	0.96	0.96
Average		0.90	0.73	0.80
Submission-Level		**0.89**	**0.67**	**0.76**

Precision levels are above 0.8 in all cases, and submission-level f-measure results for this dataset is above 0.7.

Evaluation – Comparison with iCRF

Jobs

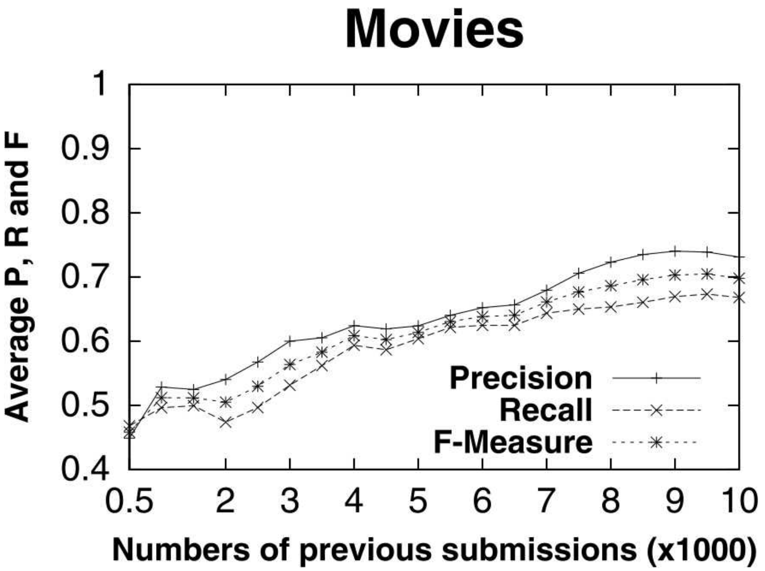
Field	iForm	iCRF
Application	**0.82**	0.37
Area	0.18	0.23
City	**0.70**	0.65
Company	**0.41**	0.17
Country	0.77	0.87
Desired Degree	**0.57**	0.37
Language	**0.84**	0.69
Platform	**0.47**	0.38
Recruiter	**0.44**	0.22
Req. Degree	0.31	0.59
Salary	0.22	0.25
State	**0.85**	0.81
Title	**0.72**	0.49

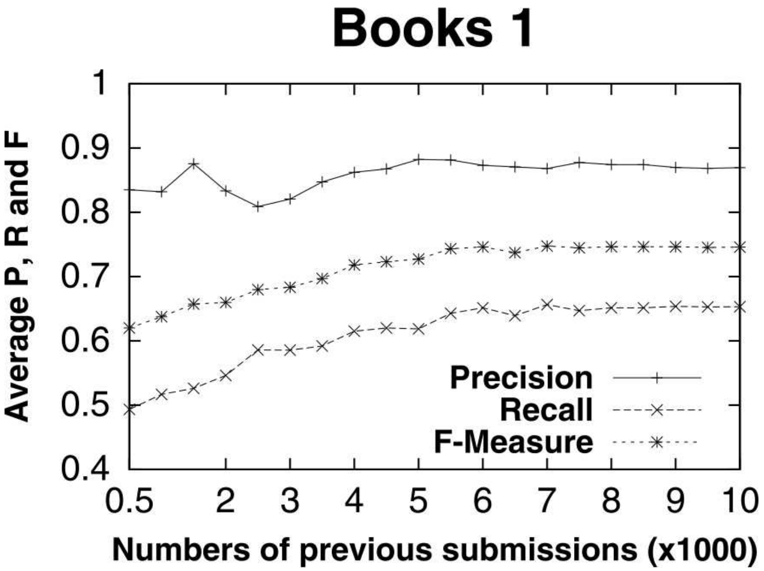
iForm had superior F-measure levels in nine fields.

The lower quality obtained by iCRF is explained by the fact that segments to be extracted from typical free text inputs, such as jobs postings, may not appear in a regular context.

iForm was designed to conveniently exploit these field-related features from previous submissions

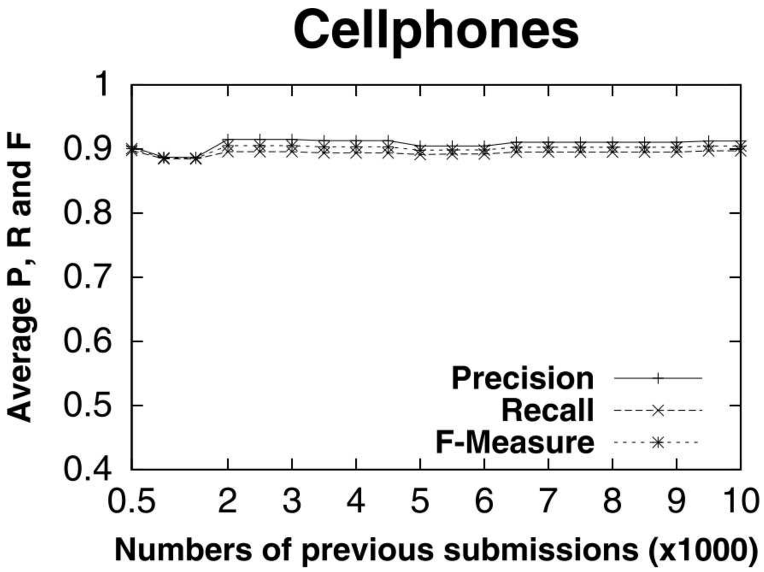
Previous Submissions Impact

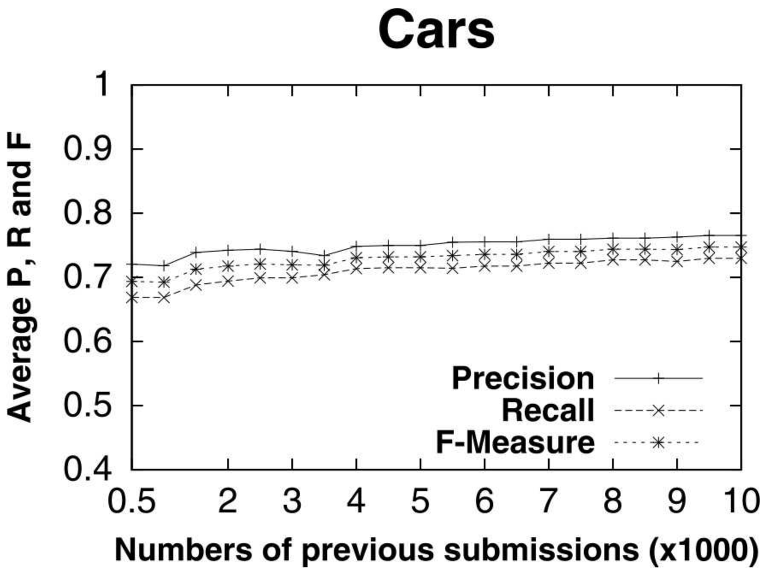




For the Movies and Books 1 datasets, the quality achieved by iForm increases proportionally with the number of previous submissions

Previous Submissions Impact



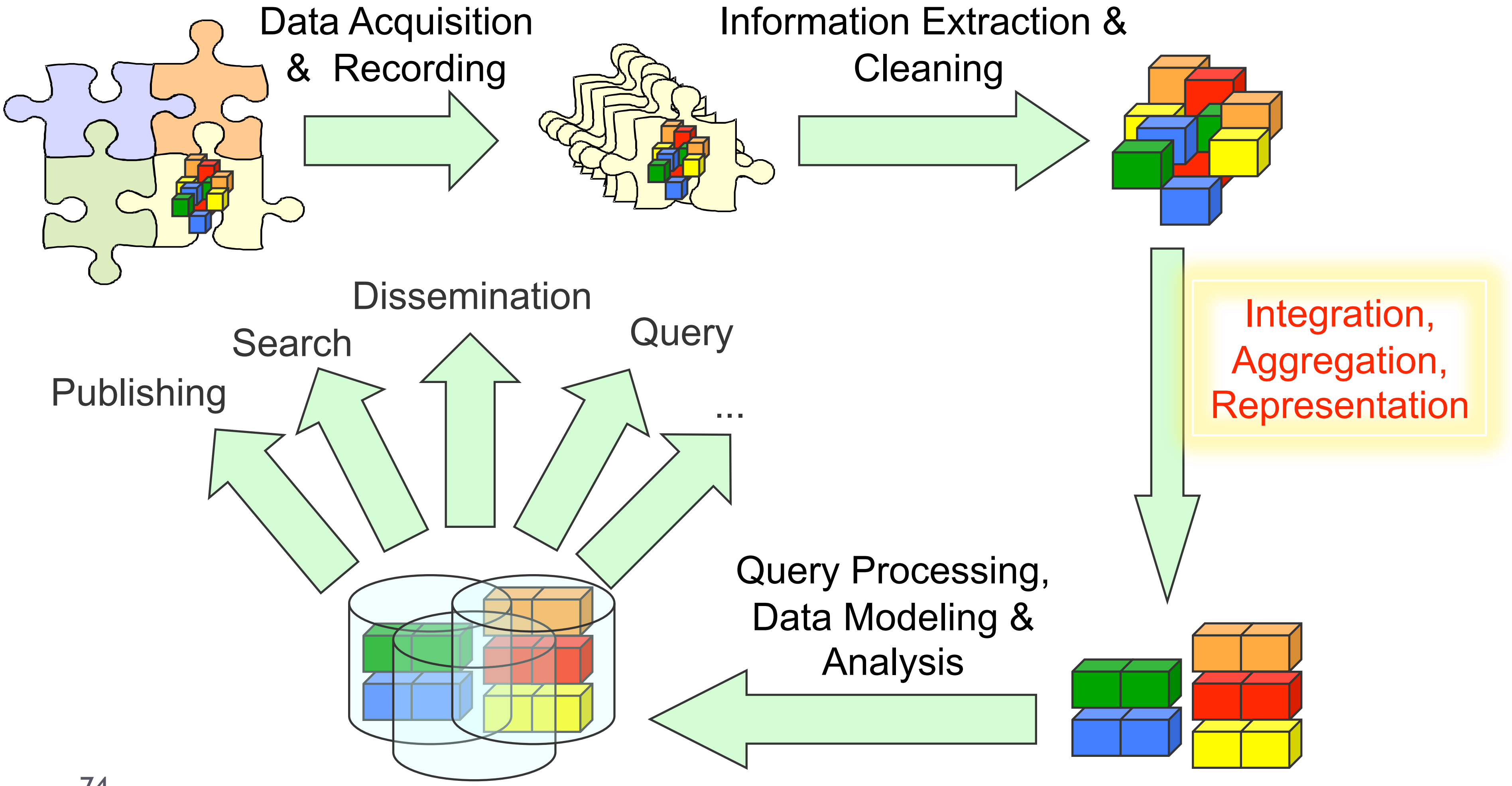


Notice that F-measure values stabilize at around 3000 previous submissions and remain the same until 10000. Besides, even starting with a small number of submissions, iForm is able to help decrease the human effort in the form filling task.

Conclusions

- A **probabilistic** approach for **automatically filling** form-based interface
- Relies on a model that estimates the probability of each field in the form given the input text based on the values **previously used for filling the form.**
- Achieved good results in comparison with iCRF
  - Our experiments demonstrate that our approach is able to properly deal with different types of input fields, such as text boxes, pull-down lists and check boxes
- More in
  - Toda, Cortez, Silva & Moura: *A Probabilistic Approach for Automatically Filling* Form-Based *Web Interfaces*. VLDB 2011

The last one …



Data Acquisition & Recording

Information Extraction & Cleaning

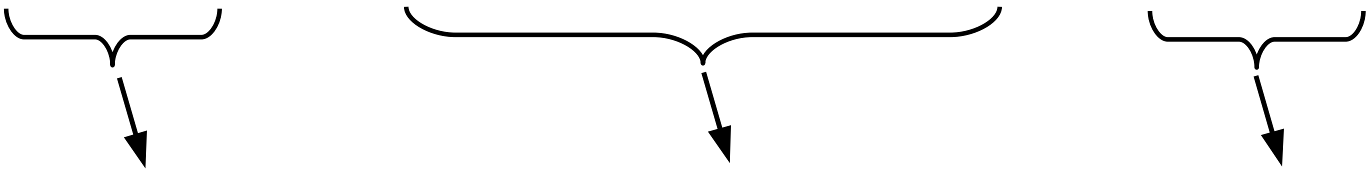
Integration, Aggregation, Representation

Query Processing, Data Modeling & Analysis

Complex Schema Matching

- A group of elements from a given schema match a group of elements from another schema.

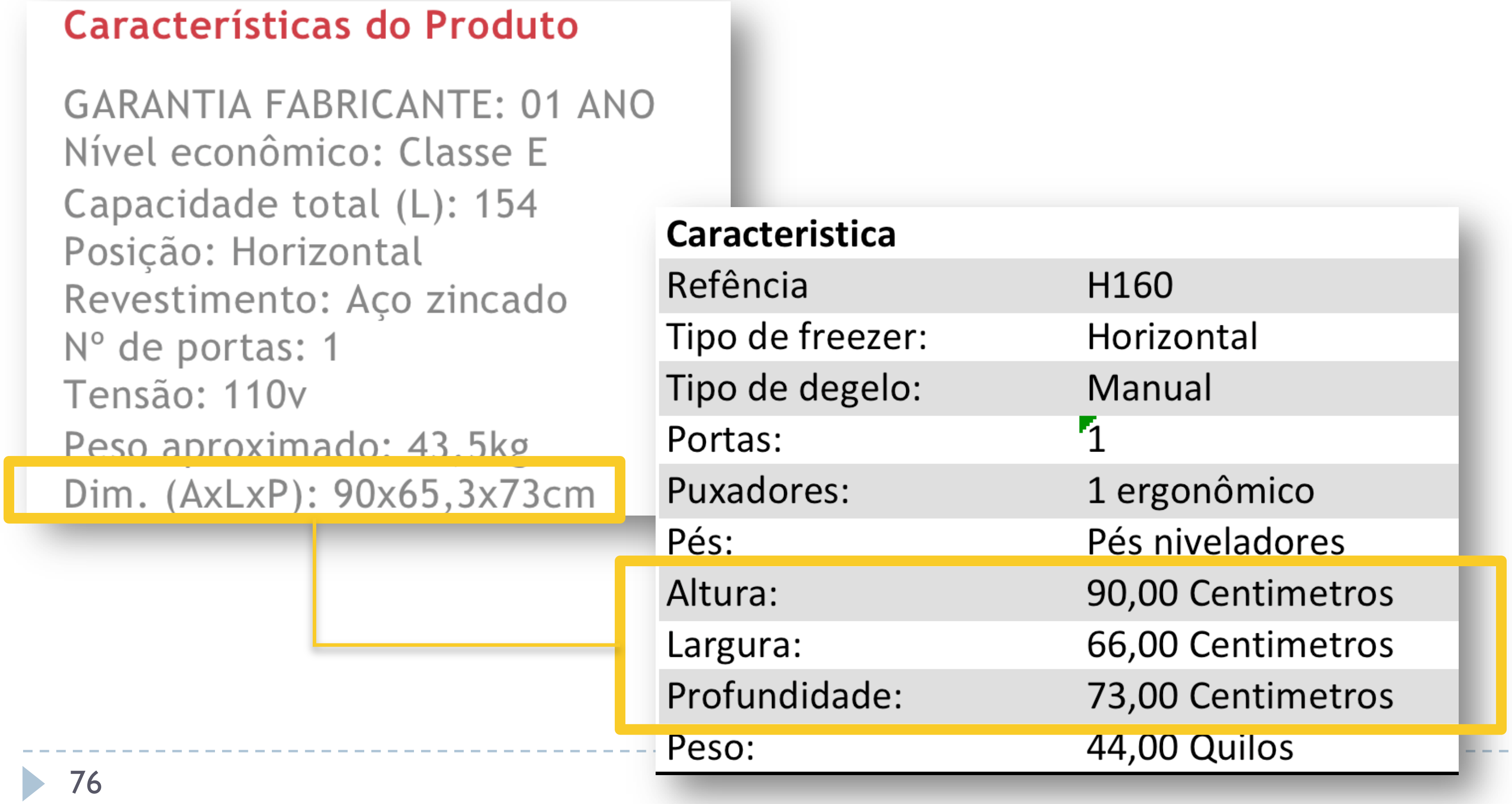
given name	surname	street number	address 1	address 2	suburb
rose	leslie	26	coranderrk street	rowethorpe	hill end
katheri	hand	18	derrington crescent	homewood	kingsthorpe
mary	white	23	prescott street		bonbeach



full name	age	address	area
leslei rose	43	coranderrk 26, rowethorpe	hi end
katherine hand	33	derrington crescent 18 , homewood	kingsthorpe
mary wite	39	prescott str	bonbeach

Complex Schema Matching

- A group of elements from a given schema matches a group of elements from another schema.



Características do Produto

GARANTIA FABRICANTE: 01 ANO
Nível econômico: Classe E
Capacidade total (L): 154
Posição: Horizontal
Revestimento: Aço zincado
Nº de portas: 1
Tensão: 110v
Peso aproximado: 43,5kg
Dim. (AxLxP): 90x65,3x73cm

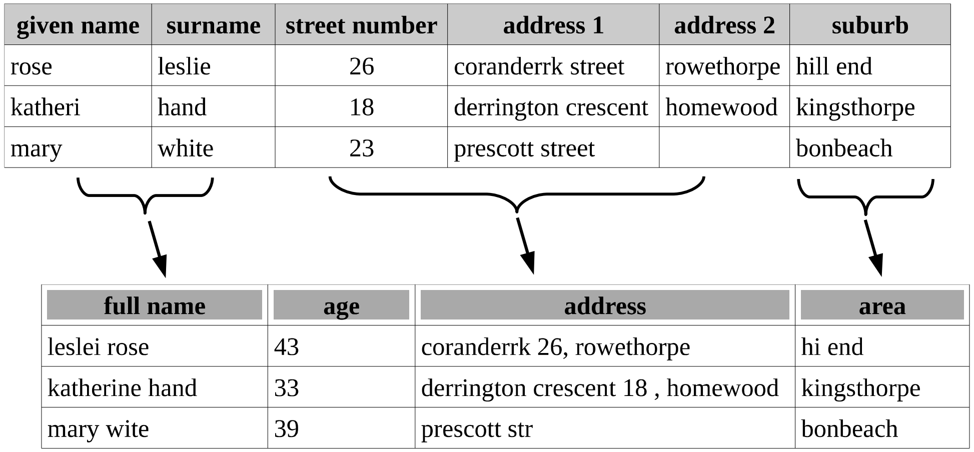
Caracteristica	
Refência	H160
Tipo de freezer:	Horizontal
Tipo de degelo:	Manual
Portas:	1
Puxadores:	1 ergonômico
Pés:	Pés niveladores
Altura:	90,00 Centimetros
Largura:	66,00 Centimetros
Profundidade:	73,00 Centimetros
Peso:	44,00 Quilos

Our approach

- An Evolutionary Approach to Complex Schema Matching
  - Just accepted to Information Systems to appear in 2013
  - With Moises Carvalho, Alberto Laender & Marcos Gonçalves
- Given two input schema, use an evolutionary process to generate *Schema Matching Solutions* for them
- Start from an initial set of possible spurious/meaningless schema matching solution
- Hopefully reach a final meaningful schema matching solution
- Use a *fitness* function to evaluate and refine the solutions been generated

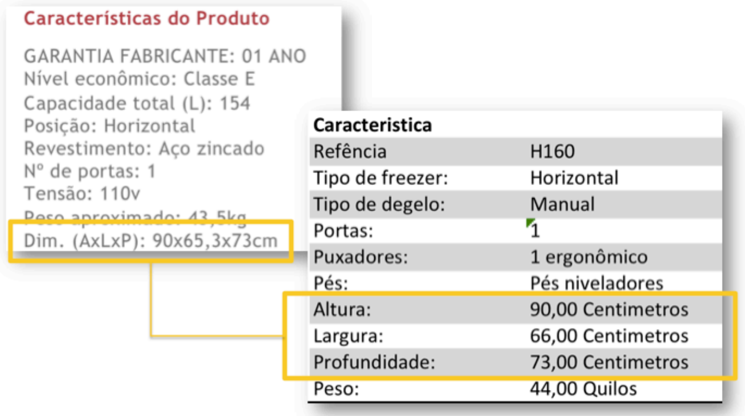
Requirements and Assumptions

- Schemata are known, but we can’t rely on attribute names
  - Different labels, noisy label extraction
- Instances are known, we rely on them
  - Assumed to be abundant



given name	surname	street number	address 1	address 2	suburb
rose	leslie	26	coranderrk street	rowethorpe	hill end
katheri	hand	18	derrington crescent	homewood	kingsthorpe
mary	white	23	prescott street		bonbeach

full name	age	address	area
leslei rose	43	coranderrk 26, rowethorpe	hi end
katherine hand	33	derrington crescent 18 , homewood	kingsthorpe
mary wite	39	prescott str	bonbeach

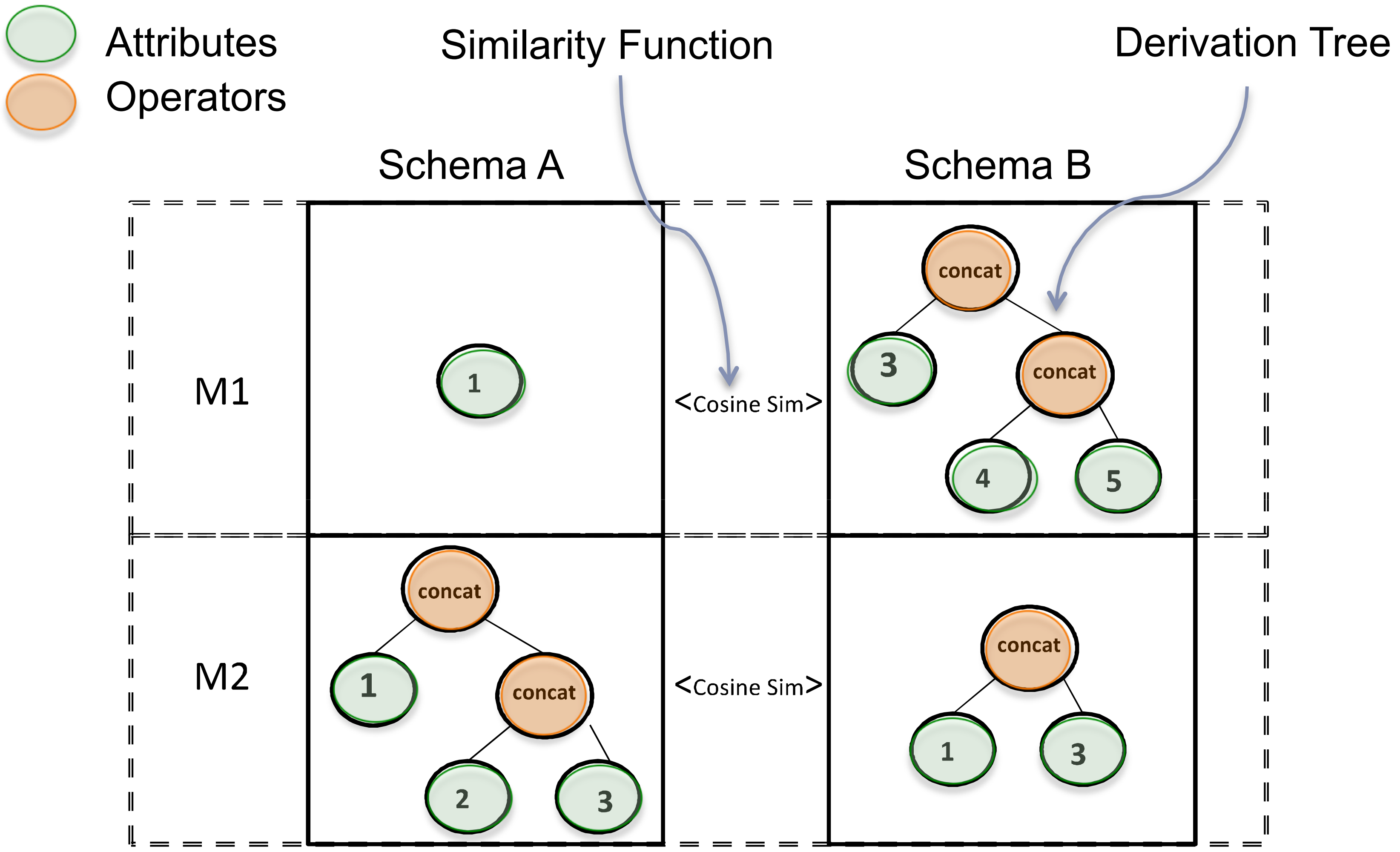


Características do Produto

GARANTIA FABRICANTE: 01 ANO
Nível econômico: Classe E
Capacidade total (L): 154
Posição: Horizontal
Revestimento: Aço zincado
Nº de portas: 1
Tensão: 110v
Peso aproximado: 43,5kg
Dim. (AxLxP): 90x65,3x73cm

Caracteristica	
Refência	H160
Tipo de freezer:	Horizontal
Tipo de degelo:	Manual
Portas:	1
Puxadores:	1 ergonômico
Pés:	Pés niveladores
Altura:	90,00 Centimetros
Largura:	66,00 Centimetros
Profundidade:	73,00 Centimetros
Peso:	44,00 Quilos

Schema Matching Solutions (SMS)



Attributes

Operators

Similarity Function

Derivation Tree

Schema A

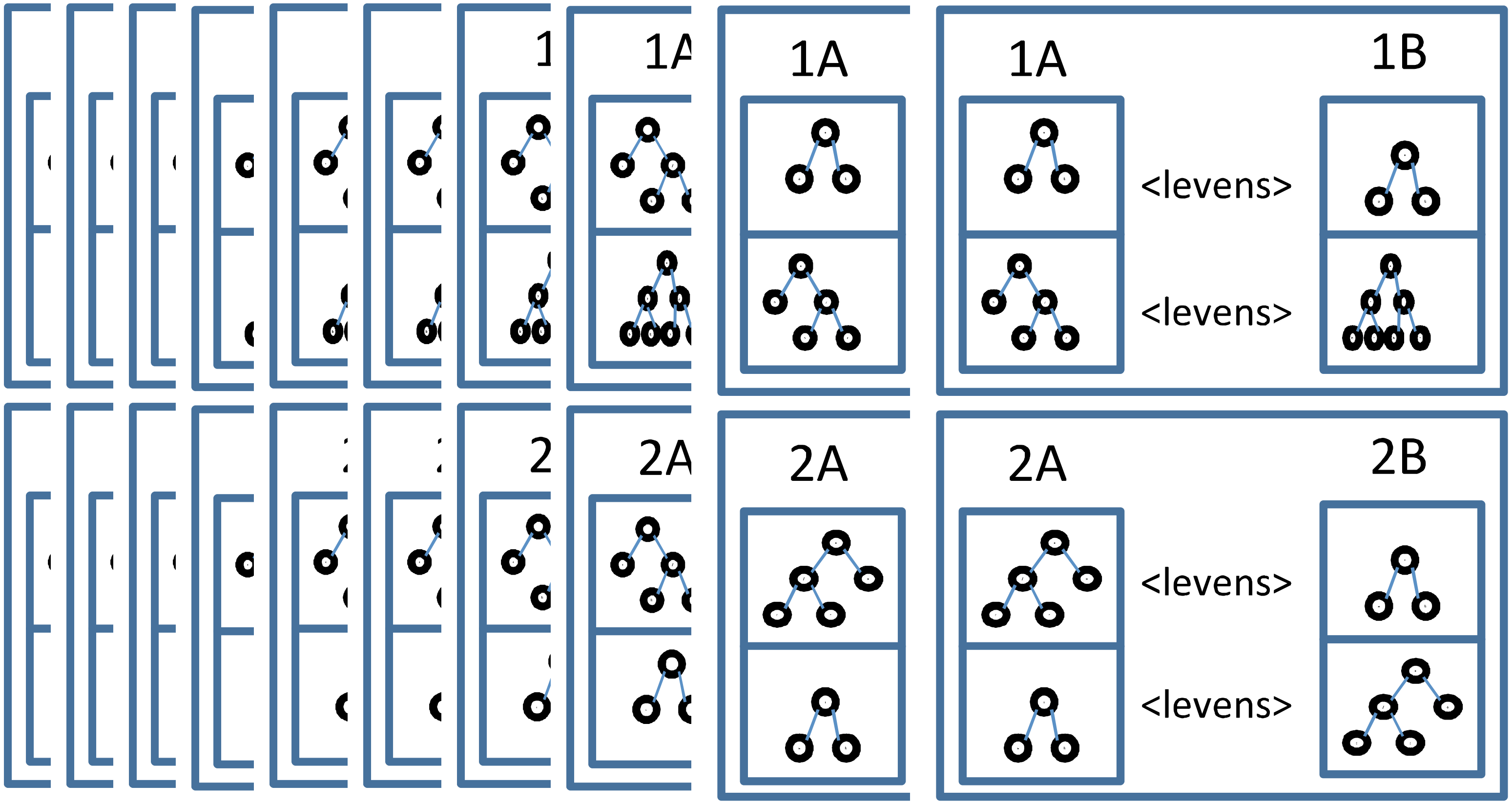
Schema B

M1

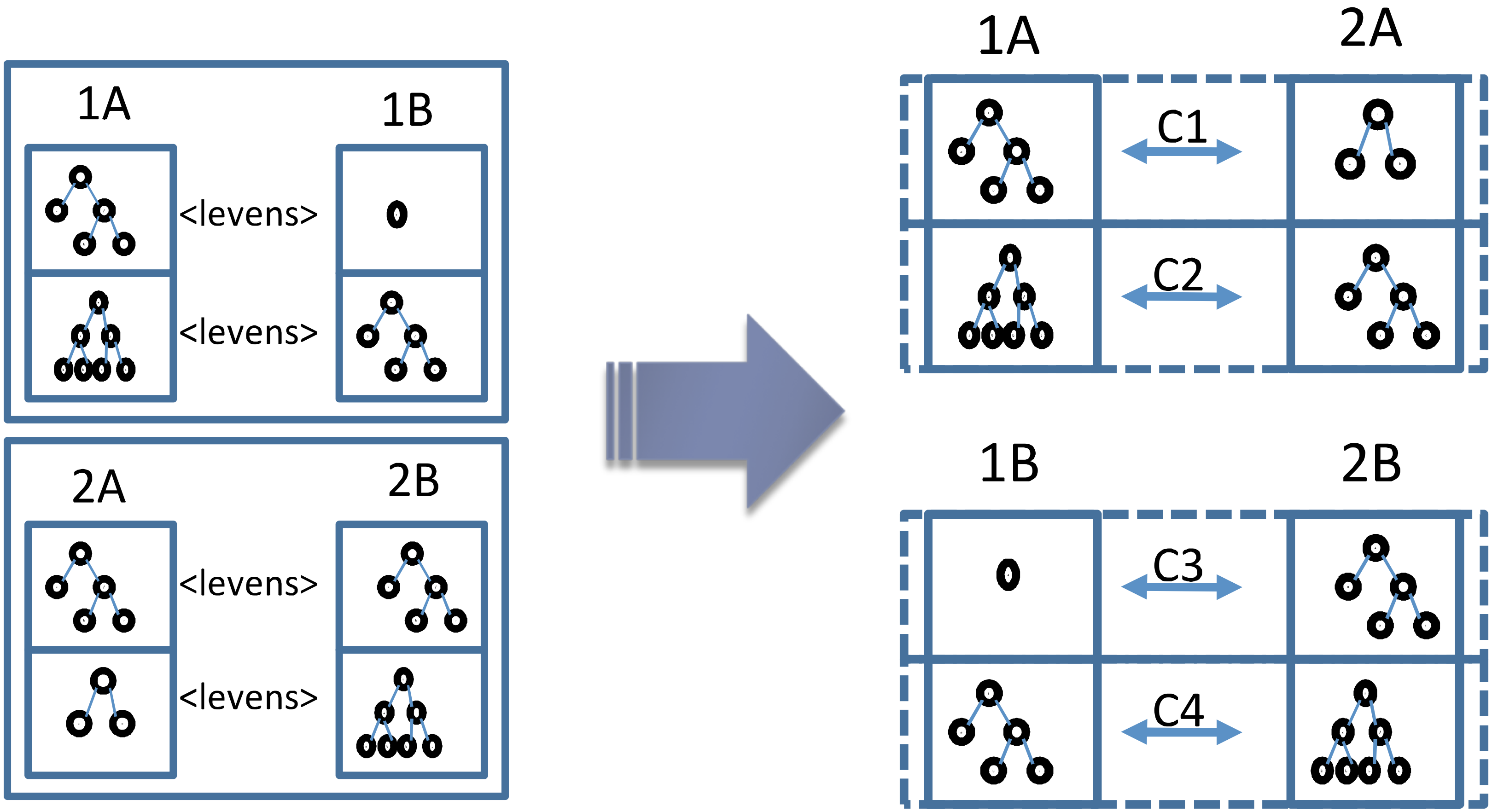
M2

SMS Evolution

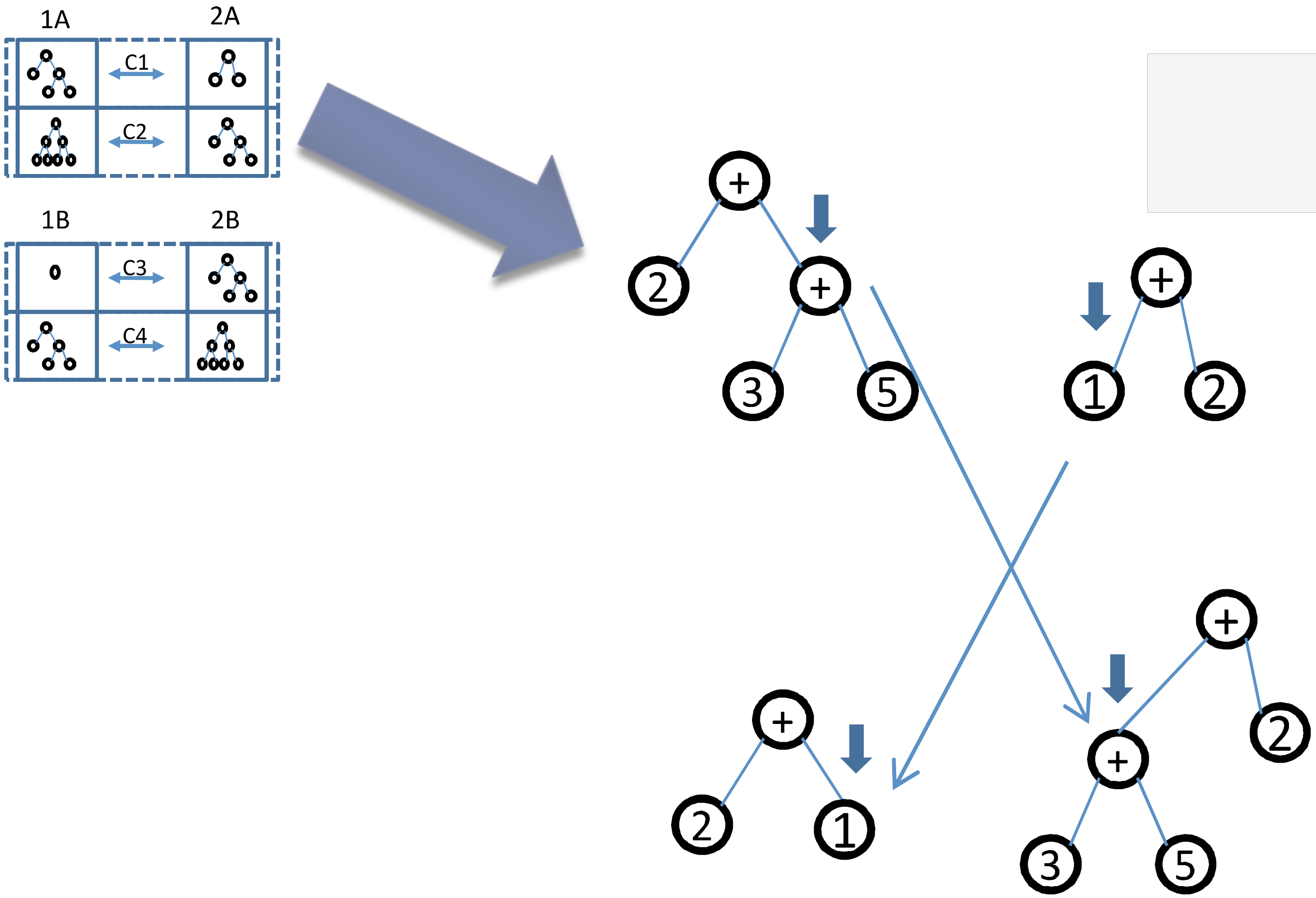
K evolutionary steps



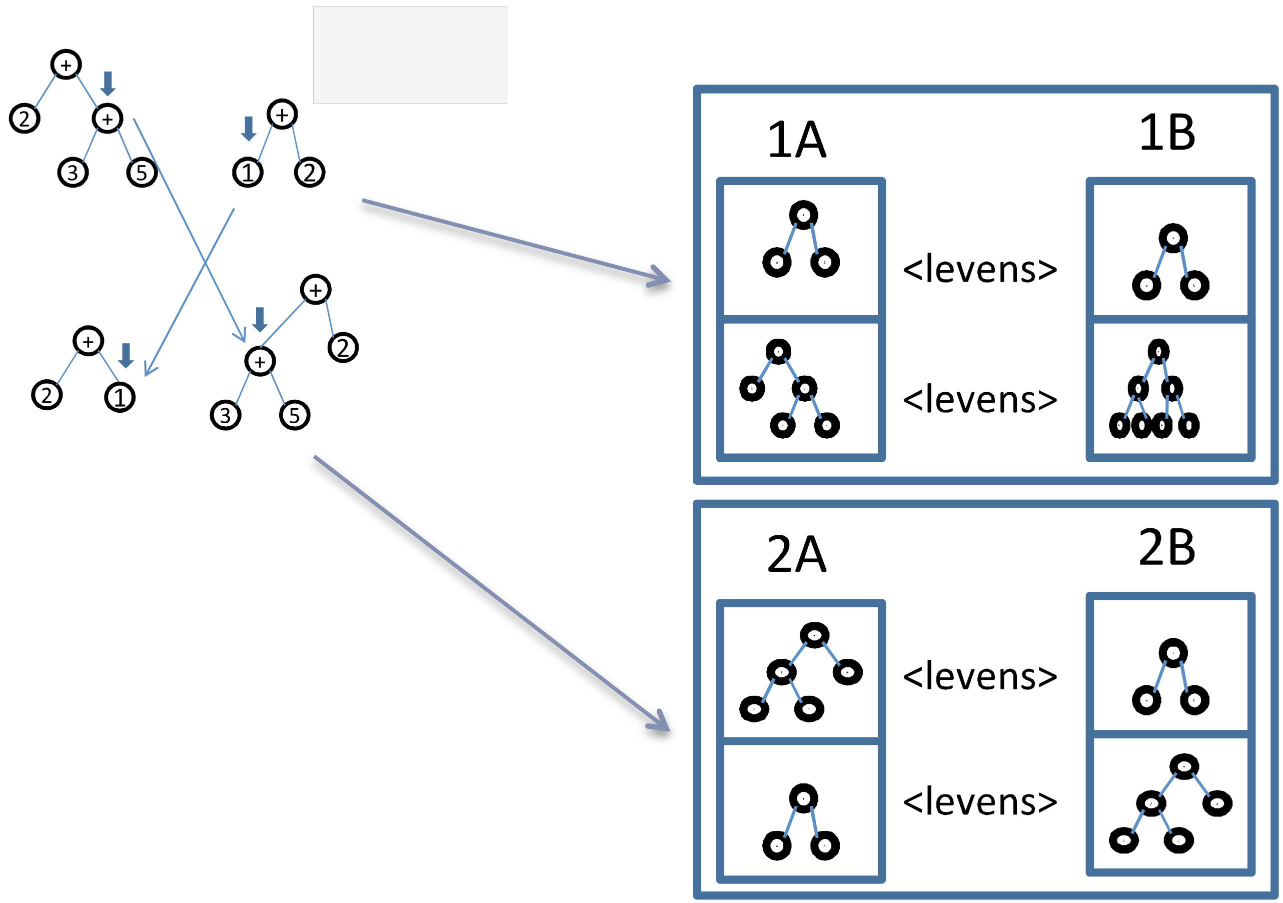
SMS Evolution: A Single Step



SMS Evolution: Crossover



SMS Evolution: New Solution



SMS Evolution – Details

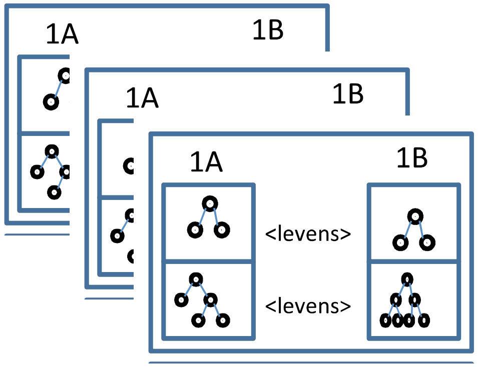
- Setup
  - Similarity Functions (e.g., Jaro, Consine, Prob. Density, etc.)
  - Data types with operators
    - STRING: concatenation, insertion, substitution, etc.
    - DATE: sum, sub, conversion (e.g., year to days), etc
    - NUMBER: sum, mult, etc.
- Next Generation
  - k individuals with the fitness value above a threshold ε is selected for mutation and crossover

SMS Evolution – Details

- Fitness: which solutions are good?
- General idea
  - Given a SMS, evaluate its matches
  - In good matches, similarity functions must give high values

$$f(S) = \frac{\sum_{i=1}^{n} eval(m_i)}{n}$$

$$m_i =$$



SMS Evolution – Details

- Two different entities
- Entity-oriented Strategy:
  - Assumes a non-negligible overlap between the instances
  - First, use similarity functions to look for similar entities
  - Then, verify if the match can detect these entities
- Value-oriented Strategy
  - Assumes an empty or negligible overlap between the instances
  - First, use similarity functions to look for similar attributes
  - Then verify if the match can detect these entities

SMS Evolution – Details

- Constraints
  - For a given match, all attributes, operations, similarity functions should be of same data type
  - The set of possible similarity functions can be select by a specialists
- These are practical constraints
  - The evolutionary process could be carried out without them
  - But using them we narrow the solution space and save some time

Experiments - Datasets

Characteristic	Synthetic 1, 2, 3	Real State	Inventory
Total of Elements in File A	12	32	44
Total of Elements in File B	7	19	38
Total of 1-1 Matches	7	7	27
String Matches	3	6	11
Numerical Matches	4	1	16
Total of Complex Matches	2	12	11
String Matches	2	5	4
Numerical Matches	0	7	7

Characteristic	Real Estate	Car Dealers	Restaurants
Total of Elements in Table A	7	28	6
Total of Elements in Table B	6	8	9
Total of 1-1 Matches	6	5	2
String Matches	3	5	2
Numerical Matches	3	1	0

Experiments - Results

Overlap

Partial (ST1)

Matches	Accuracy
All 1-1 Matches	57%
String 1-1 Matches	100%
Numeric 1-1 Matches	24%
All Complex Matches	75%
String Complex Matches	75%

Full (ST2)

Matches	Accuracy
All 1-1 Matches	100%
Strings 1-1 Matches	100%
Numeric 1-1 Matches	100%
All Complex Matches	100%
String Complex Matches	100%

Non-Overlap

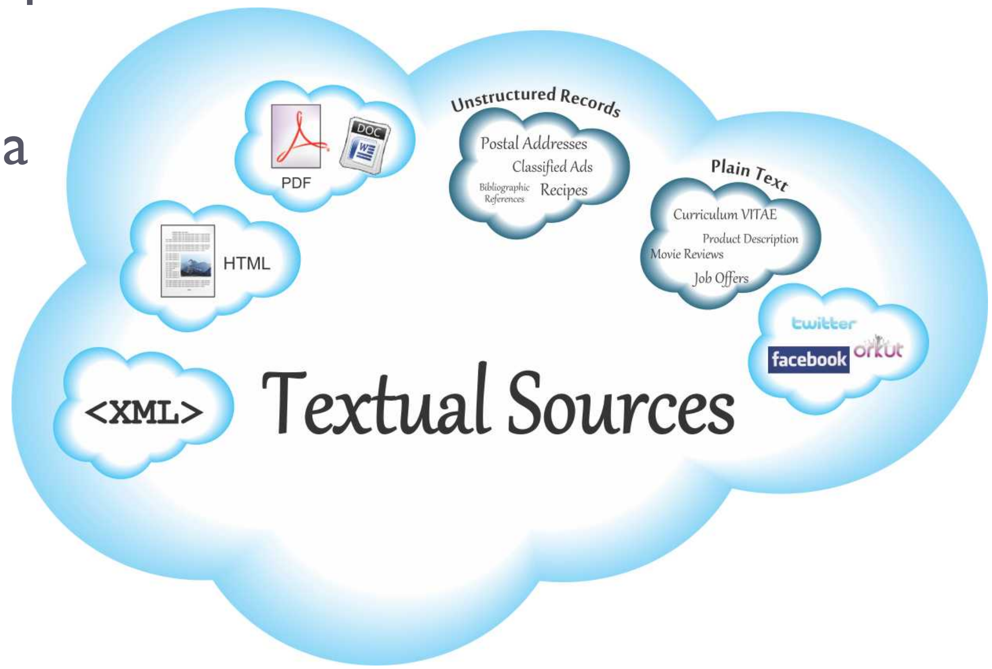
Matches	Accuracy
RS All 1-1 Matches	85%
RS String 1-1 Matches	100%
RS Numeric 1-1 Matches	0%
RS All Complex Matches	25%
RS String Complex Matches	60%
RS Numeric Complex Matches	0%
INV All 1-1 Matches	40%
INV String 1-1 Matches	100%
INV Numeric 1-1 Matches	0%
INV All Complex Matches	20%
INV String Complex Matches	56%
INV Numeric Complex Matches	0%
ST3 All 1-1 Matches	42%
ST3 String 1-1 Matches	100%
ST3 Numeric 1-1 Matchings	0%
ST3 All Complex Matches	100%
ST3 String Complex Matches	100%

Experiments – Examples of Matches

- Inventory dataset:
  - ship-address = (ship-address + ship-postal-code) +
    (ship-city + ship-country)
- Real State dataset:
  - house-address = (house-street + house-city) + house-zip-code
- Synthetic 3 dataset:
  - fullname = forename + surname

Conclusions and Remarks

- Data of interest is no longer in databases, although they are in on-line sources
- In particular: Textual Sources
  - The structure is only implicit
  - Meta-data is a luxury
  - Constraints are a utopia



Other areas can help a lot

- Information Retrieval
  - IR models, text indexing, relevance metrics, language models, etc.
- Data/Text Mining
  - Rule Mining, Learning, Categorization, Graph Models
- Artificial Intelligence
  - Ontologies, Automated Reasoning
- ….

An expanded set of CS foundations is helpful!

- Computer Science Theory for the Information Age
  - Upcoming book by John Hopcroft and Ravindran Kannan
- From the TOC
  - High-Dimensional Space
  - Random Graphs
  - Singular Value Decomposition (SVD)
  - Markov Chains
  - Learning and VC-dimension
  - Algorithms for Massive Data Problems
  - Clustering
  - Graphical Models and Belief Propagation

This is the theory for the next 30 years !!

Many other approaches

- Named Entity Recognition (NER)
  - E.g. Sarawagi@FTD'08, Ratinov@CoNLL'09
- Open Information Extraction
  - Unsupervised NER over massive text collections, e.g., the Web
  - Oren Etzioni (e.g., EMNLP-CoNLL'12, WWW'08, IJICAI'07)
- Hidden Web
  - Juliana Freire (e.g., WWW'07, ICDE'07, WebD'10)
- Web Tables
  - Alon Halevy, Mike Cafarela (e.g., PVLDB'08, CIDR'07)
- NoDB – Scientific Data!
  - Anastacia Ailamaki (e.g., SIGMOD'12)