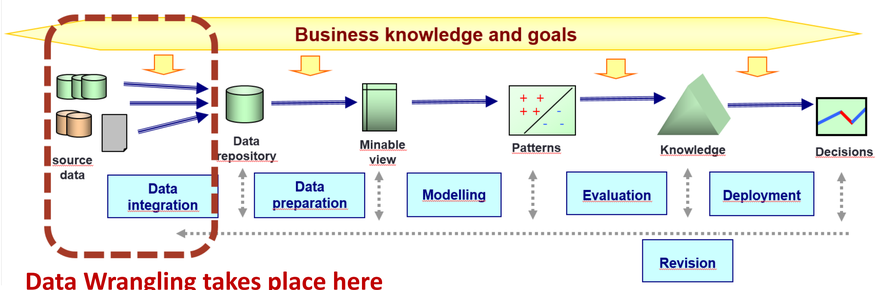
General-Purpose Inductive Programming for Data Wrangling Automation

Lidia Contreras-Ochando, Fernando Martínez-Plumed, Cèsar Ferri, José Hernández-Orallo and María José Ramírez-Quintana
Universitat Politècnica de València (UPV), Spain
{liconoc, fmartinez, cferri, jorallo, mramirez}@dsic.upv.es

DATA WRANGLING IN DATA SCIENCE



Data Wrangling takes place here

- **50%-80% of effort.**
- It depends on **previous knowledge**.

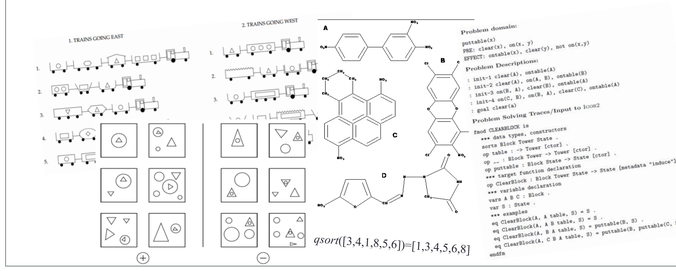
(Semi-)Automating Data Wrangling would have a very significant impact

INDUCTIVE PROGRAMMING (IP)

- Machines program themselves (programming by example, program synthesis, ILP, ...).

Input:
- Data **D** (usually small sets of data)
- Background Knowledge **BK** (facts, functions, constraints, etc.)

Output:
- Hypothesis **h** (a program, possibly rich and complex)



$qsort([],[]) \leftarrow$
$qsort([X|T],S) \leftarrow part(X,T,L1,L2), qsort(L1,S1), qsort(L2,S2), app(S1,[X|S2],S)$

prod(s(X0), X1) = sum(prod(X0, X1), X1)
prod(0, X0) = 0

delOdds(L,R) ← L=[], R=[]
delOdds(L,R) ← L=[HL|TL], odd(HL), delOdds(TL,TR), R=TR
delOdds(L,R) ← L=[HL|TL], ¬odd(HL), delOdds(TL,TR), R=[HL|TR]

$B_m \forall x.((group\_likes\ x) = \text{if } ((\wedge_{\text{s}} \blacklozenge B_s\ likes\ B_s\ likes)\ x) \text{ then } \top \text{ else if } ((\wedge_{\text{s}} \blacklozenge B_s\ likes\ B_s\ likes)\ top)\ x) \text{ then } \top \text{ else } \bot).$

1.0: likes(X,Y):- friendof(X,Y).
0.8: likes(X,Y):- friendof(X,Z), likes(Z,Y).
0.5: friendof(john,mary).
0.5: friendof(mary,pedro).
0.5: friendof(mary,tom).
0.5: friendof(pedro,tom).

```
% Bush syntactic model:
12 :: imperative_sentence(verb_phrase(verb,noun_phrase(noun)))
13 :: affirmative_sentence(noun_phrase(noun),verb_phrase(verb))
 3 :: imperative_sentence(verb_phrase(verb,noun_phrase(noun)),
      prop_pred(prep,noun_phrase(det,noun))))
 3 :: imperative_sentence(verb_phrase(verb,noun_phrase(noun)),
      prop_pred(prep,noun_phrase(noun))))
 4 :: affirmative_sentence(connector,
      affirmative_sentence(noun_phrase(noun),
      verb_phrase(verb(aux,verb,verb),noun_phrase(noun))))
 4 :: affirmative_sentence(noun_phrase(noun),
      affirmative_sentence(noun_phrase(noun),
      verb_phrase(verb(aux,aux,verb)))
 4 :: imperative_sentence(verb_phrase(verb))
 7 :: affirmative_sentence(noun_phrase(noun),
      affirmative_sentence(noun_phrase(noun),
      verb_phrase(verb,noun_phrase(det,noun)))
...
```

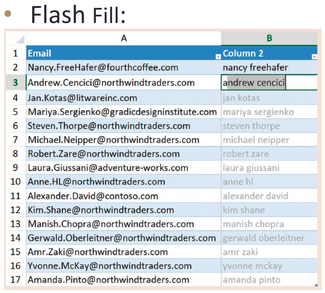
IP FOR DATA WRANGLING: GPDL + BK!

- **Re-use** any IP sysem based on a **general-purpose declarative language** (GPDL):
  - GOLEM, Progol, (F)FOIL, ADATE, DIALOGS, FLIP, IGOR I/II, Aleph, MagicHaskeller, Metagol, gErl, ...
  - Users may edit the solutions written in languages such as Haskell , Erlang or Prolog.
- User can choose or modify the libraries and **background knowledge** (BK)

IP FOR DATA MANIPULATION: THE DSL APPROACH

- Data manipulation automation using DSLs
  - Define IP systems over Domain Specific Languages (DSLs) fitting the domain.
  - **Expressive enough** to cover the problems in the domain.
  - **Restrictive enough** to enable efficient search.
- It has been a success! (Gulwani 2011-2016)
  - FlashFill, FlashExtractText, FlashRelate, FlashNormalize, BlinkFill, ...
- Limitations:
  - Systems (IP engines) **must be redesigned** for each domain...
  - **Lack of flexibility**: customisation, including background knowledge, ...

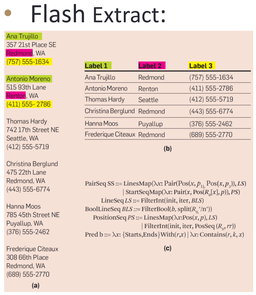
- Flash Fill:



	A	B
1	Email	Column 2
2	Nancy.FreeHafer@fourthcoffee.com	nancy freehafer
3	Andrew.Cencici@northwindtraders.com	andrew cencici
4	Jan.Kotas@litwareinc.com	jan kotas
5	Mariya.Sergienko@gradicdesigninstitute.com	mariya sergienko
6	Steven.Thorpe@northwindtraders.com	steven thorpe
7	Michael.Neipper@northwindtraders.com	michael neipper
8	Robert.Zare@northwindtraders.com	robert zare
9	Laura.Giussani@adventure-works.com	laura giussani
10	Anne.HL@northwindtraders.com	anne hl
11	Alexander.David@contoso.com	alexander david
12	Kim.Shane@northwindtraders.com	kim shane
13	Manish.Chopra@northwindtraders.com	manish chopra
14	Gerwald.Oberleitner@northwindtraders.com	gerwald oberleitner
15	Amr.Zaki@northwindtraders.com	amr zaki
16	Yvonne.McKay@northwindtraders.com	yvonne mckay
17	Amanda.Pinto@northwindtraders.com	amanda pinto

(a)

- Flash Extract:



Ana Trujillo
357 21st Place SE
Redmond, WA
(757) 555-1634

Antonio Moreno
515 93th Lane
Renton, WA
(411) 555-2786

Thomas Hardy
742 17th Street NE
Seattle, WA
(412) 555-5719

Christina Berglund
475 22th Lane
Redmond, WA
(443) 555-6774

Hanna Moos
785 45th Street NE
Puyallup, WA
(378) 555-2462

Frederique Citeaux
308 68th Place
Redmond, WA
(888) 555-2770

Label 1	Label 2	Label 3
Ana Trujillo	Redmond	(757) 555-1634
Antonio Moreno	Renton	(411) 555-2786
Thomas Hardy	Seattle	(412) 555-5719
Christina Berglund	Redmond	(443) 555-6774
Hanna Moos	Puyallup	(378) 555-2462
Frederique Citeaux	Redmond	(888) 555-2770

(b)

(c)

EXAMPLE

- Attribute transformation performed with:
  - **MagicHaskeller** (Katayama 2004-today): http://nautilus.cs.miyazaki-u.ac.jp/~skata/MagicHaskeller.html
  - **Metagol** (Muggleton et al. 2014-today) .https://github.com/metagol/metagol
  - **gErl** (Martinez-Plumed et al. 2012-today). https://github.com/nandomp/gErl

	Gender	GenderOK	Birthdate	BirthdateOk	Postcode	PostcodeOK	Score	ScoreOk	Km	metresOK	Weight	WeightOK
#1	Male	M	3 1 1971	1971 1 3	46 025	46025	5.5, 4.6, 5.8	5.3	5	5000	f "CAMP DRY DBL NDL 3.6 OZ"	"3.6 OZ"
#2	Female	F	4 5 1993	1993 5 4	46225	46225	3.5	3.5	3	3000	f "DRY NDL 0.23 KG"	"0.23 Kg"
#3	...	...	...	...	...	...	...	...	...	...	...	...

STATUS

- **Analyse common transformations** in data wrangling (from tutorials/books/users/systems and DSL-based wrangling tools).
  - *Feature transformations.*
  - *Row transformations.*
  - *Table transformations.*
  - *Integration from several tables.*
  - *Other kinds of formatting.*
- **Define a library** (BK) that ***can*** solve many of these common problems.
  - Trade-off between efficiency and power (syntactic and semantic domain).
- BK selection by **interaction** (asking the user)
  - Do you think this is a date? An address? Use the appropriate BK for the task.
- With MagicHaskeller:
  - Identify the function library (LibTH.hs) to make it easily editable :
  - Modifiable by the user and coding new functions.
- With Metagol:
  - Try to resolve some type problems
  - Two "libraries" to be selected for Metagol: Metarules and proper BK.
- With gErl:
  - Needs learning operators, data representation and examples navigation.
  - Identify BIFs o new functions to be added in the background knowledge.

Download the paper:

https://goo.gl/YvC75L