Domain specific induction for data wrangling automation

Lidia Contreras-Ochando
Cèsar Ferri
José Hernández-Orallo
Susumu Katayama
Fernando Martínez-Plumed
María José Ramírez-Quintana

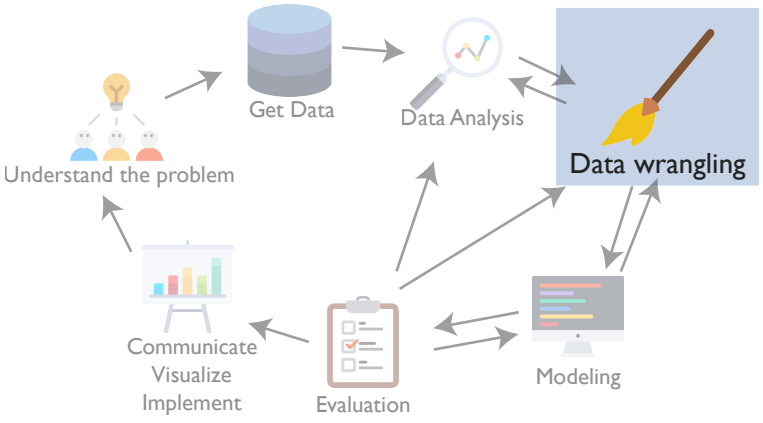
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ
dmip
宮崎大学 University of Miyazaki

What is data wrangling?

Row No.	Starting Station	Starting Time	Ending Time	Ending Station
1	001	03/10/2016 00:18:36	03/10/2016 00:32:12	004
2	001	03/10/2016 00:25:45	03/10/2016 00:38:07	006
...	...	...	...	...
69852	001	6-10-16 20:35	6-10-16 14:39	007

- Different formats and domains
- To be normalised

Data Science process:



Data wrangling

This step includes:

- It is the most tedious, boring and repetitive step
- It takes up to **80%** of the project time

Automating data wrangling process is essential to reduce time and cost

Approach:

Inductive Programming

The program receives:
- Some examples
- Background Knowledge

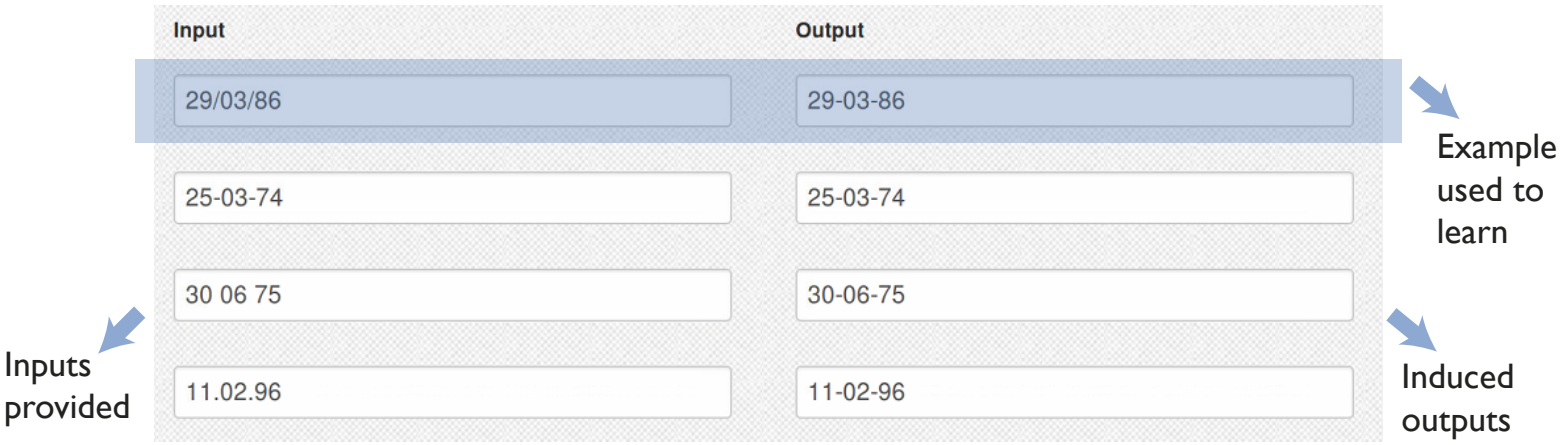
→ The result is a hypothesis on how to obtain new examples by using the knowledge.

\+

Domain specific Background Knowledge

Dates | Emails | Names | Words

Demo app:



First results:

The DSI approach outperforms the results obtained by other data wrangling tools

Thanks! =)

Questions?

Poster & Demo
liconoc@upv.es
@liconoc