# .Web Scraping. Collecting data and working with them... ☺


Carina F. Dorneles

carina.dorneles@ufsc.br

PPGCC
Programa de Pós-Graduação em Ciência da Computação

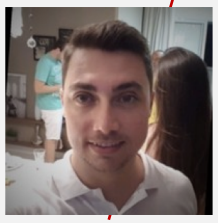
UFSC

# Collecting data .....☺

# Collecting data

.**Crawlers** *Undergraduate Students*

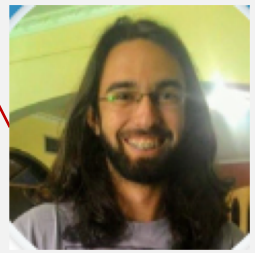
- **Web Forms**
- **Web Tables**
- **Research Questionnaires**
- **Medical papers and package insert**
- **Researcher's publication**
- **Metadata**
- **Citation**
- **Q&As**



eonardo Bres dos Santos

Arthur Machado Branco

Larissa Taw

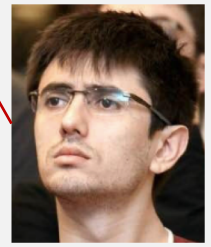


Luiz Philipi Machado da Silva

Eduardo Picinin

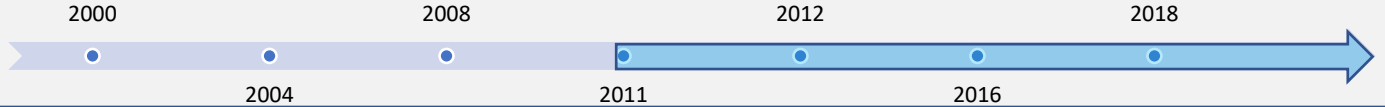
Marcelo Scheidt

Gilney Nathanael Mathias



Lucas Viana Knochenhauer

*Working since*




**... and working with them .....**

Data Extraction
Data Similarity
Ranking

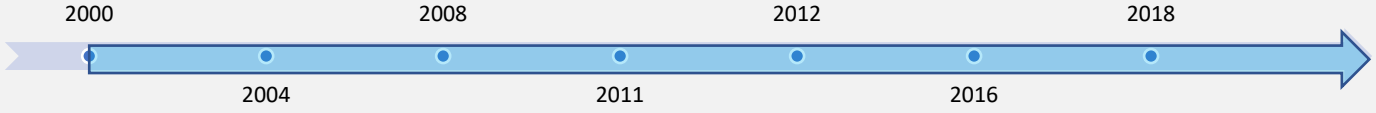
- **Data extraction**
  - **Segmentation**
  - **Noise removal**
  - **Record extraction**
  - **Entity extraction**



Roberto Panerai Velloso
Master Student (2012-2014)
PhD Student (2015 - now)

Edimar Manica (UFRGS)
Master Student (2011-2013)
PhD Student (2013 - 2017)

*Working since*



# Data Similarity

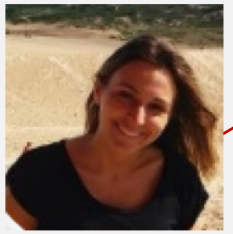
- Metrics
- Clustering
- Classification

Karine B. de Oliveira
Master Student (2011-2013)

Richard Henrique de Souza
PhD Student (2015 - now)

Larissa Lautert
Master Student (2011-2013)

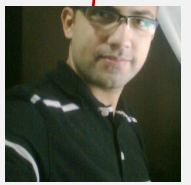
Rodrigo Gonçalves
PhD Student (2016 - now)



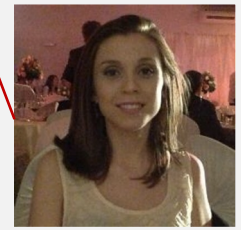
Gabriela B. Colonetti
Undergrad. Student - 2016



Jaime Mendes da Silva
Undergrad. Student - 2016

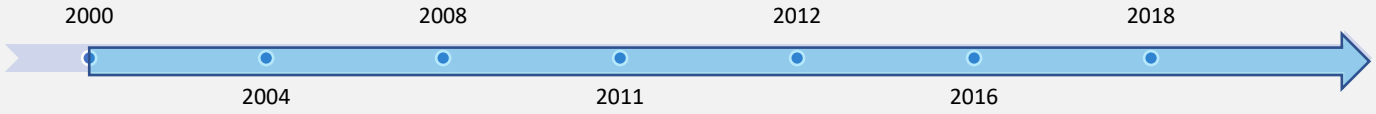


Gleidson C. da Silva
Master Student (2012-2014)

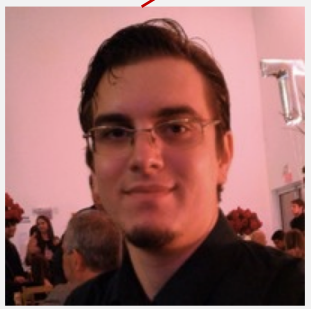


Renata Tomaz Siega
Undergrad. Student - 2016

*Working since*



- **Ranking**
  - **Weighting adjustment**
  - **Features definition**



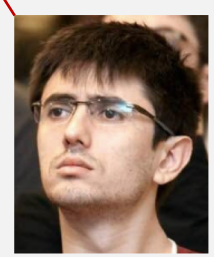
Felipe Born de Jesus
Master Student (2015-2017 )



José Henrique Calenzo
Master Student (2014-2016)

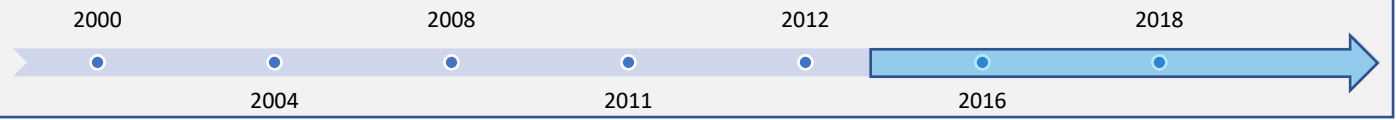


Leandro Amâncio
Master Student (2015-2017)



Lucas Viana Knochenhauer
Master Student (2016 - now )

*Working since*



Web

Crawler

Web Subset

Data Extraction

Data Similarity

Ranking