

# Instituto de Computação UNIVERSIDADE ESTADUAL DE CAMPINAS



# Capacitação profissional em tecnologias de Inteligência Artificial

#### **Machine Learning Overview**

**Prof. Edson Borin** 

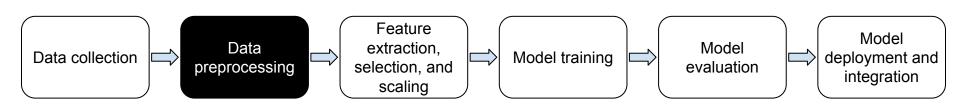
https://www.ic.unicamp.br/~edson
Institute of Computing - UNICAMP



#### ML Process



### Data preprocessing







Data plays an important role in machine learning

Bad data leads to bad models!





**Data cleaning:** modify data to ensure it is accurate and correct

#### Potential data problems

- Invalid duplicate items;
- Incorrect format;
- Attribute dependencies;
- Missing values;
- Missing value
- Invalid value
- etc.





#	ld	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	М	0	0	Ireland	Dublin
2	222	Mary	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	М	0	0	France	Paris
5	555	Alex	15/03/2000	Α	1	23	Germany	Berlin
6	555	Peter	1983-12-01	М	1	10	Italy	Rome
7	777	Calvin	05/05/1995	М	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	М	1	26	Ytali	Rome





#	ld	Name	Birthday		Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/	1990	М	0	0	Ireland	Dublin
2	222	Inval	id	978	F	1	15	Iceland	
3	333	luplicate	item	2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/	1997	М	0	0	France	Paris
5	555	Alex	15/03/2	15/03/2000		1	23	Germany	Berlin
6	555	Peter	1983-1	2-01	М	1	10	Italy	Rome
7	777	Calvin	05/05/	1995	М	0	0	Italy	Italy
8	888	Roxane	03/08/	1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/	05/09/1992		0	5	Switzerland	Geneva
10	101010	Paul	14/11/	1992	М	1	26	Ytali	Rome





#	ld	Name	Birthday		Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/	1990	М	0	0	Ireland	Dublin
2	222	Inval	id	978	F	1	15	Iceland	
3	333	duplicate	eitem	000	F	0	0	Spain	Madrid
4	444	Mark	01/11/	01/11/1997		0	0	France	Paris
5	555	Alex	15/03/2	15/03/2000		1	23	Germany	Berlin
6	555	Peter	1983-1	2-01	М	1	10	Italy	Rome
7	777	Calvin	05/05/	1/95	М	0	0	Italy	Italy
8	888	Roxane	03/08	18	F	0	0	Portugal	Lisbon
9	999	Anne	ncorrect		F	0	5	Switzerland	Geneva
10	101010	Paul	TOTM	format		1	26	Ytali	Rome





#	ld	Name	Birthday		Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/	1990	М	0	0	Ireland	Dublin
2	222	Inval	id	978	F			Iceland	
3	333	duplicate item 200		2000	F		alid lue	Spain	Madrid
4	444	Mark	01/11/1997		М			France	Paris
5	555	Alex	15/03/2000		Α	1	23	Germany	Berlin
6	555	Peter	1983-1	2-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/	1/95	М	0	0	Italy	Italy
8	888	Roxane	U3/U8	18	F	0	0	Portugal	Lisbon
9	999	Anne	Incorrect format		F	0	5	Switzerland	Geneva
10	101010	Paul			М	1	26	Ytali	Rome





#	ld	Name	Birth	day	Sex	IsTeacher #Students		Country	City
1	111	John	31/12/	1990	М	0	0	Ireland	Dublin
2	222	Inval	id	978	F	la.	ralid	Iceland	
3	333	duplicate	e item 2000		F		alid lue	Spain	Madrid
4	444	Mark	01/11/	1997	М			France	Paris
5	555	Alex	15/03/2000		Α	1 23		Germany	Berlin
6	555	Peter	1983-1	2-01	M	At	Attribute		Rome
7	777	Calvin	05/05/	1/95	М	dep	endency	Italy	Italy
8	888	Roxane	U3/U8	18	F	0	0	Portugal	Lisbon
9	999	Anne	Incorrect format		F	0	5	Switzerland	Geneva
10	101010	Paul			М	1	26	Ytali	Rome





Missing value

Data cleaning: modify data to ensure it is accurate

and correct

#	ld	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	М	0	0	Ireland	Dublin
2	222	Inval	Invalid 978		lan.	ralid	Iceland	
3	333	duplicate	e item j <sub>000</sub>	F		alid lue	Spain	Madrid
4	444	Mark	01/11/1997	М	Value		France	Paris
5	555	Alex	15/03/2000	Α	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	At	tribute	Italy	Rome
7	777	Calvin	05/05/1/95	М	dep	endency	Italy	Italy
8	888	Roxane	U3/U8 18	F	0	0	Portugal	Lisbon
9	999	Anne	Incorrect	F	0	5	Switzerland	Geneva
10	101010	Paul	format	М	1	26	Ytali	Rome





Missing value

Data cleaning: modify data to ensure it is accurate

and correct

#	ld	Name	Birth	day	Sex	IsTeacher	#Students	Country	City	
1	111	John	31/12/	1990	М	0	0	Ireland	Dublin	
2	222	Inval			F	la.	Invalid value			
3	333	duplicate			F				Madrid	
4	444	Mark	01/11/	1997	М			France	Value that should be in another column	
5	555	Alex	15/03/2	2000	Α			Germany		
6	555	Peter	1983-1	2-01	M	At	tribute	Italy	e	
7	777	Calvin	05/05/	1/95	М	dep	endency	Italy	Italy	
8	888	Roxane	03/08	18	F	0	0	Portugal	Lisbon	
9	999	Anne	Incorre		F	0	5	Switzerland	Geneva	
10	101010	Paul	forma	al	М	1	26	Ytali	Rome	





Missing value

Data cleaning: modify data to ensure it is accurate

and correct

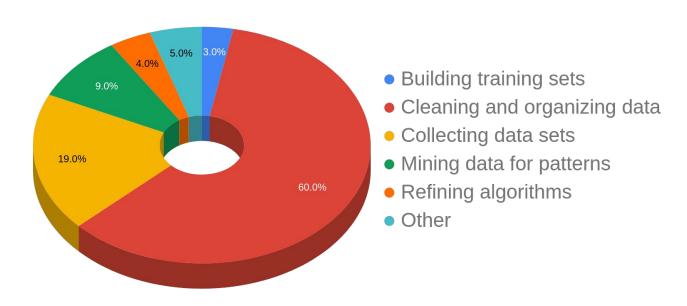
#	ld	Name	Birth	day	Sex	IsTeache	#Students	Country	City	
1	111	John	31/12/	1990	М	0	0	Ireland	Dublin	
2	222	Invali	id	978	F		volid	Iceland		
3	333	duplicate	item	item <sub>000</sub>			Invalid value		Madrid	
4	444	Mark	01/11/	1997	М	1 23		France	Value that should be in another column	
5	555	Alex	15/03/2	2000	Α			Germany		
6	555	Peter	1983-1	2-01	M	A	\ttribute	Italy	e	
7	777	Calvin	05/05/	1/95	М	dep	pendency	Italy	Italy	
8	888	Roxane	U3/U8	18	F	0	0	Portugal	Misspelli	na
9	999	Anne	Incorre		F	0	5	Switzerland	•	9
10	101010	Paul	forma	11	М	1	26	Ytali	Rome	





### **Data cleaning**: modify data to ensure it is accurate and correct

What data scientists spend the most time doing

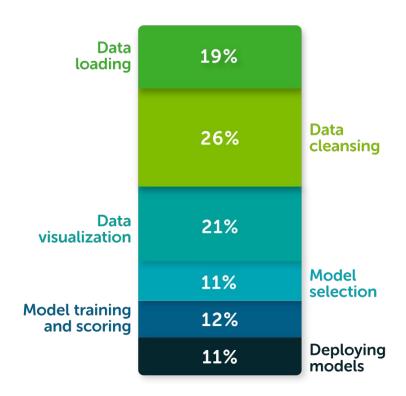


Source: 2016 Data Science Report - CrowdFlower





### **Data cleaning:** modify data to ensure it is accurate and correct



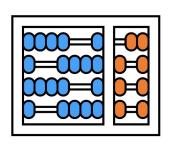
Source: The State of Data Science 2020 - Moving from hype toward maturity https://www.anaconda.com/state-of-data-science-2020





**Data conversion**: change data into a representation form suitable for the ML model

- Text and categorical attributes may need to be encoded as numbers
- Text may need to be encoded as a word vector (e.g., word2vec model, BERT model)
- Value data may be converted to category to simplify problem and/or improve learning
- Etc.



# Instituto de Computação UNIVERSIDADE ESTADUAL DE CAMPINAS



# Capacitação profissional em tecnologias de Inteligência Artificial

#### **Machine Learning Overview**

**Prof. Edson Borin** 

https://www.ic.unicamp.br/~edson
Institute of Computing - UNICAMP