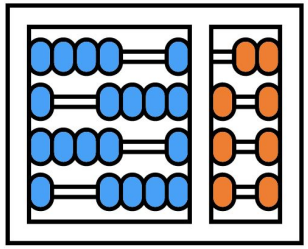


Instituto de Computação

UNIVERSIDADE ESTADUAL DE CAMPINAS

Capacitação profissional em tecnologias de Inteligência Artificial

Machine Learning Overview

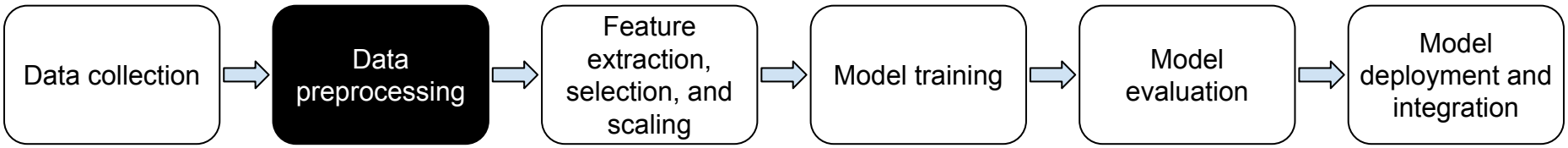
Prof. Edson Borin

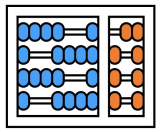
https://www.ic.unicamp.br/~edson

Institute of Computing - UNICAMP

ML Process

Data preprocessing

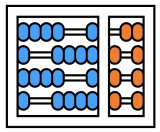




ML Process - Data preprocessing

Data plays an important role in machine learning

- **Bad data leads to bad models**!

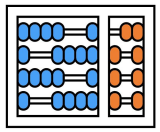


ML Process - Data preprocessing

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

Potential data problems

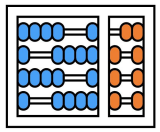
- Invalid duplicate items;
- Incorrect format;
- Attribute dependencies;
- Missing values;
- Missing value
- Invalid value
- *etc.*



ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	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	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	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	M	1	26	Ytali	Rome

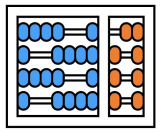


ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F	1	15	Iceland	
3	333		000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	**555**	Alex	15/03/2000	A	1	23	Germany	Berlin
6	**555**	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	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	M	1	26	Ytali	Rome

Invalid duplicate item



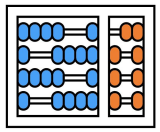
ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F	1	15	Iceland	
3	333		000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	**555**	Alex	15/03/2000	A	1	23	Germany	Berlin
6	**555**	Peter	**1983-12-01**	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/18	F	0	0	Portugal	Lisbon
9	999	Anne		F	0	5	Switzerland	Geneva
10	101010	Paul		M	1	26	Ytali	Rome

Invalid duplicate item

Incorrect format



ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F			Iceland	
3	333		2000	F			Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	**555**	Alex	15/03/2000	**A**	1	23	Germany	Berlin
6	**555**	Peter	**1983-12-01**	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/48	F	0	0	Portugal	Lisbon
9	999	Anne		F	0	5	Switzerland	Geneva
10	101010	Paul		M	1	26	Ytali	Rome

Invalid duplicate item

Invalid value

Incorrect format

ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F			Iceland	
3	333		2000	F			Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	**555**	Alex	15/03/2000	**A**	1	23	Germany	Berlin
6	**555**	Peter	**1983-12-01**	M			Italy	Rome
7	777	Calvin	05/05/1995	M			Italy	Italy
8	888	Roxane	03/08/18	F	0	0	Portugal	Lisbon
9	999	Anne		F	**0**	**5**	Switzerland	Geneva
10	101010	Paul		M	1	26	Ytali	Rome

Invalid duplicate item

Invalid value

Attribute dependency

Incorrect format

ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F			Iceland	
3	333		2000	F			Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	**555**	Alex	15/03/2000	**A**	1	23	Germany	Berlin
6	**555**	Peter	**1983-12-01**	M			Italy	Rome
7	777	Calvin	05/05/1995	M			Italy	Italy
8	888	Roxane	03/08/48	F	0	0	Portugal	Lisbon
9	999	Anne		F	**0**	**5**	Switzerland	Geneva
10	101010	Paul		M	1	26	Ytali	Rome

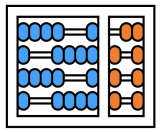
Missing value

Invalid duplicate item

Invalid value

Attribute dependency

Incorrect format



ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F			Iceland	
3	333		2000	F			Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	
5	**555**	Alex	15/03/2000	**A**	1	23	Germany	
6	**555**	Peter	**1983-12-01**	M			Italy	
7	777	Calvin	05/05/1995	M			Italy	**Italy**
8	888	Roxane	03/08/18	F	0	0	Portugal	Lisbon
9	999	Anne		F	**0**	**5**	Switzerland	Geneva
10	101010	Paul		M	1	26	Ytali	Rome

Missing value

Invalid duplicate item

Invalid value

Value that should be in another column

Attribute dependency

Incorrect format

ML Process - Data preprocessing

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

#	Id	Name	Birthday	Sex	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222		978	F			Iceland	
3	333		2000	F			Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	
5	**555**	Alex	15/03/2000	**A**	1	23	Germany	
6	**555**	Peter	**1983-12-01**	M			Italy	
7	777	Calvin	05/05/1995	M			Italy	**Italy**
8	888	Roxane	03/08/18	F	0	0	Portugal	
9	999	Anne		F	**0**	**5**	Switzerland	
10	101010	Paul		M	1	26	**Ytali**	Rome

Missing value

Invalid duplicate item

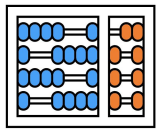
Invalid value

Value that should be in another column

Attribute dependency

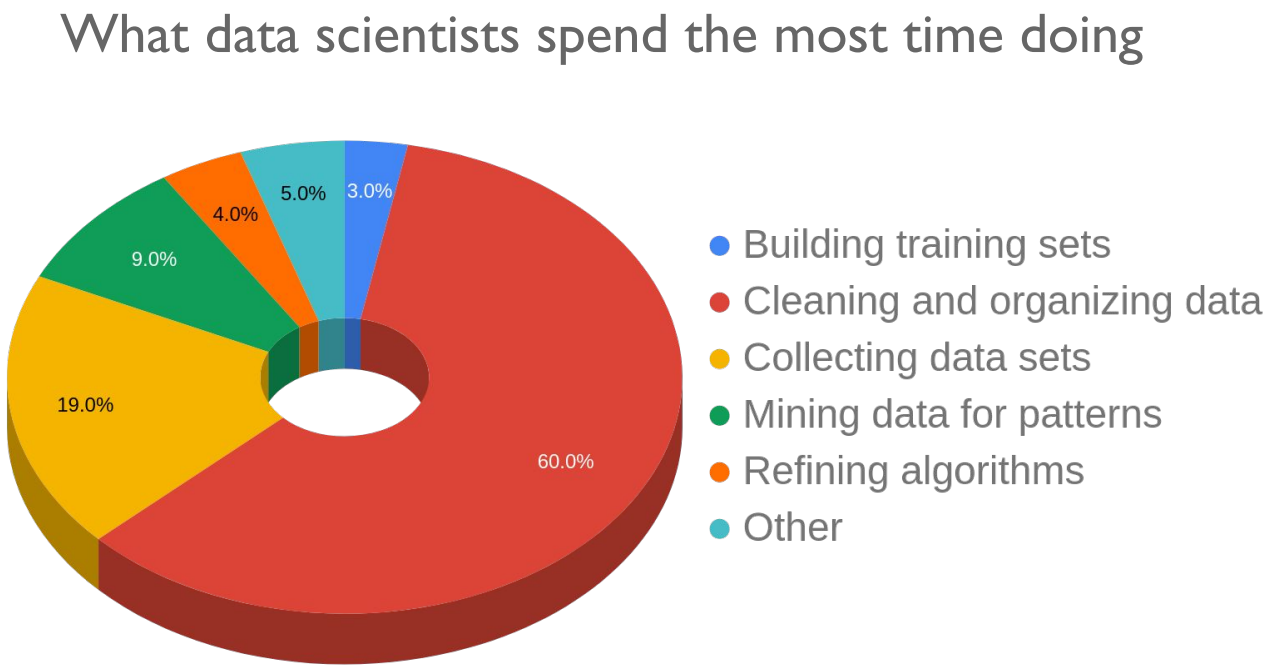
Incorrect format

Misspelling



ML Process - Data preprocessing

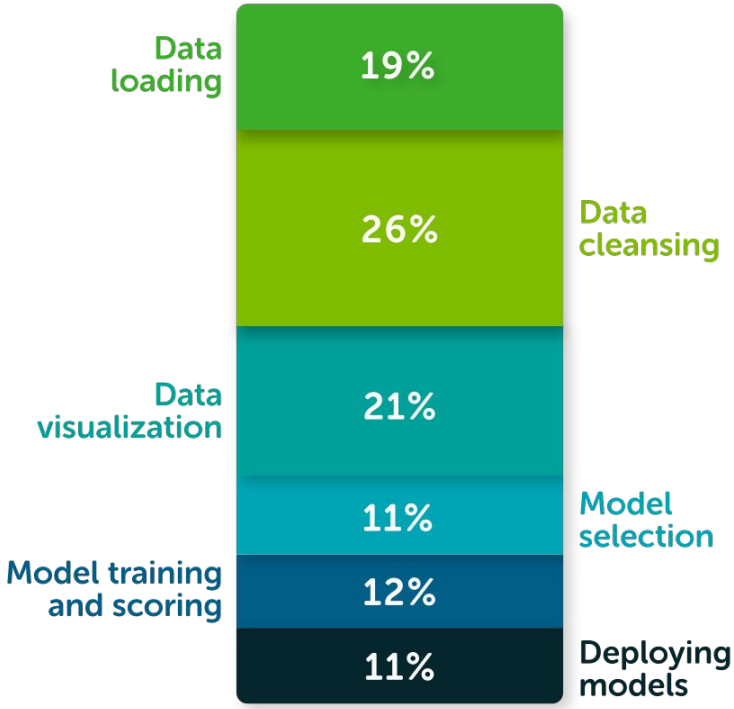
Data cleaning: modify data to ensure it is accurate and correct



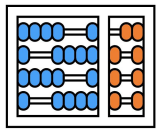
Source: 2016 Data Science Report - CrowdFlower

ML Process - Data preprocessing

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



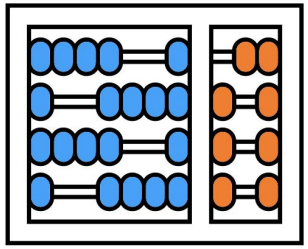
Source: The State of Data Science 2020 - Moving from hype toward maturity
https://www.anaconda.com/state-of-data-science-2020



ML Process - Data preprocessing

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