Università degli Studi di Padova

Machine Learning 2024/2025

AMCO
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

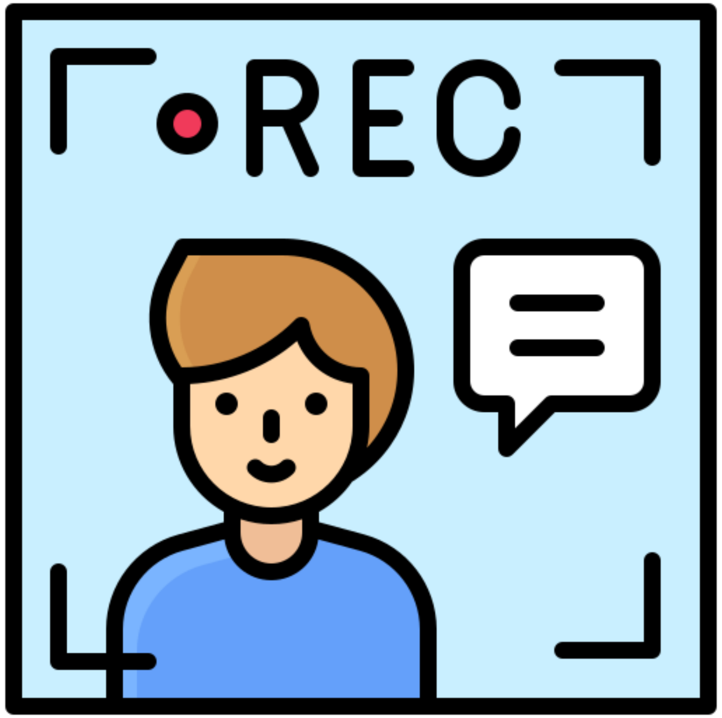
Lecture #02 Introduction & Basic Statistics

Gian Antonio Susto

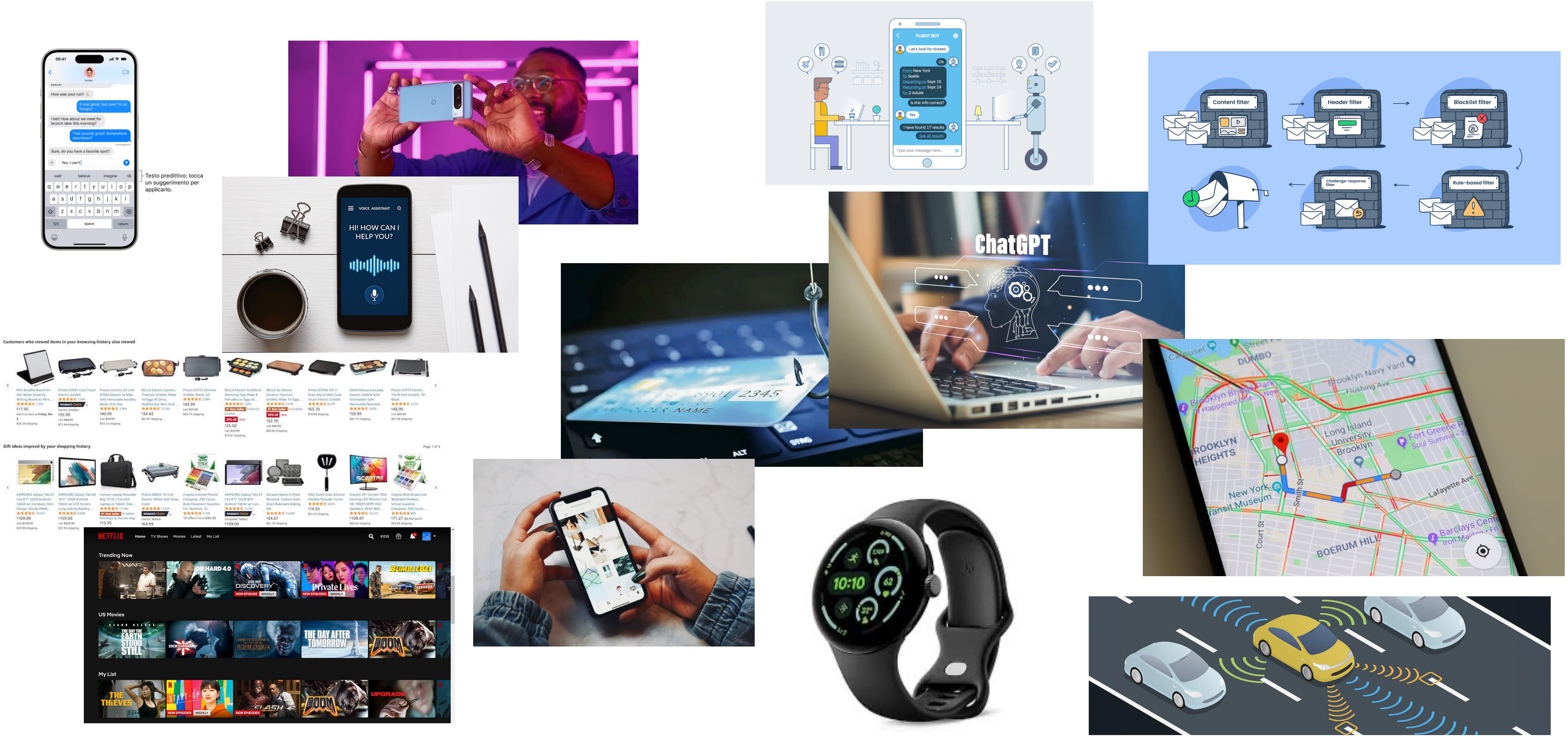
Before starting

IPSE DIXIT:
‘Lecture and laboratories recordings will be made available shortly after the lecture.’

... from this lecture! As I forgot to record the first one...



Many technologies, let’s put some order



Many technologies, let's put some order

Different tasks (objectives), different models, different data type... and different stages of development!

Many technologies, let's put some order

Different tasks (objectives), different models, different data type... and different stages of development!

A task refers to the specific problem that the model is designed to solve. It defines the objective that the model aims to achieve based on input data

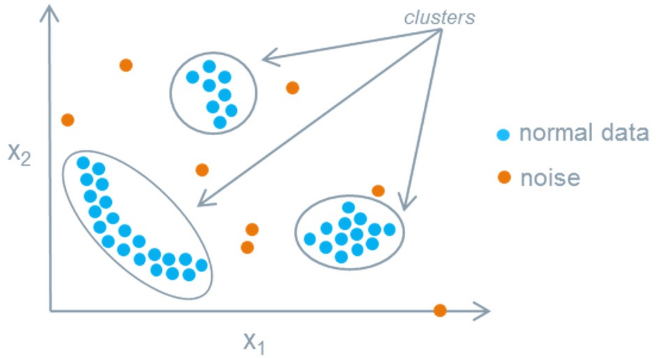
Supervised Learning

Setup: Observation of the environment

Data: (x,y)

Task: learn a map from inputs x to outputs y

Unsupervised Learning



Setup: Observation of the environment

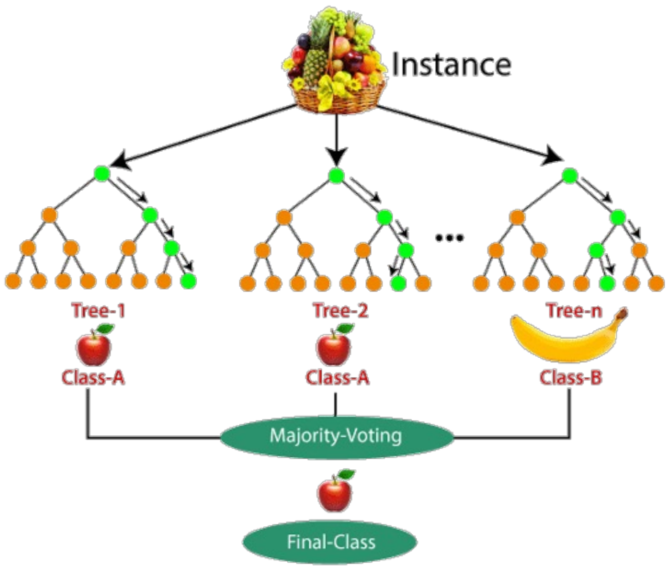
Data: x (no labels)

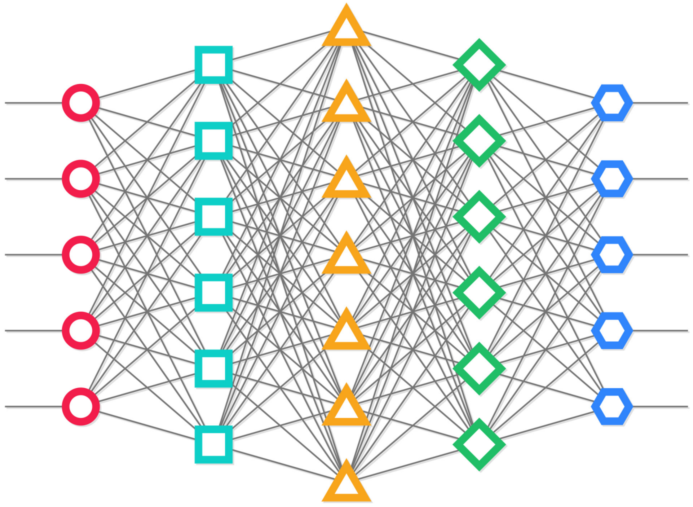
Task: learn patterns in input data

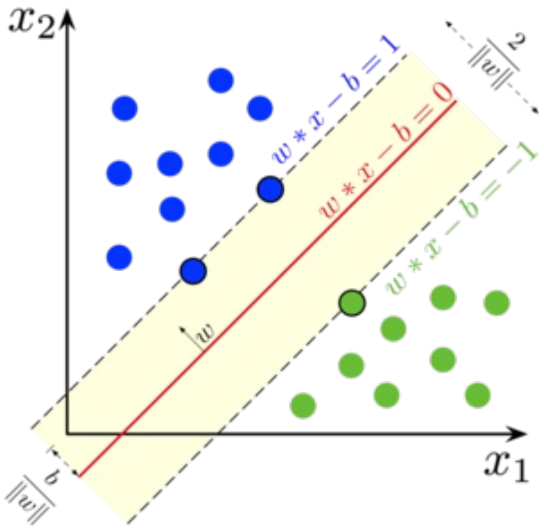
Many technologies, let's put some order

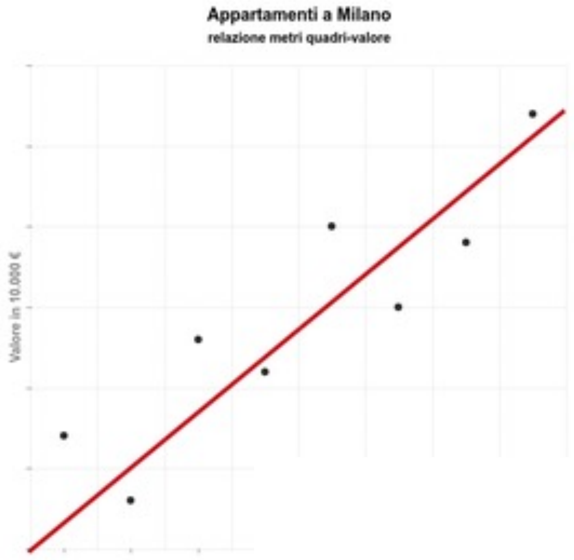
Different tasks (objectives), different models, different data type... and different stages of development!

A model is a mathematical representation of patterns and relationships within data. It is trained using an algorithm to make predictions or decisions based on input data



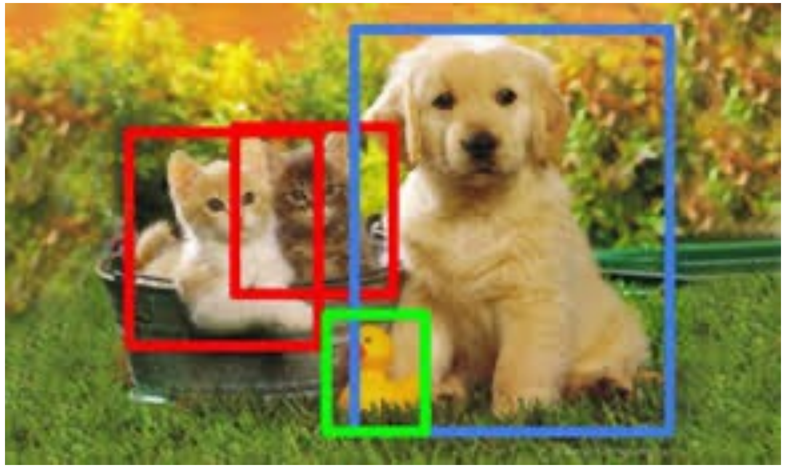


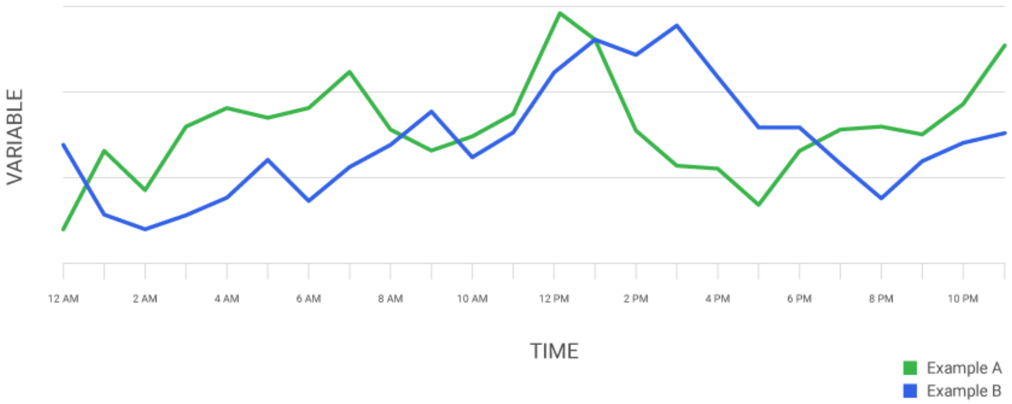




Many technologies, let's put some order

Different tasks (objectives), different models, different data type... and different stages of development!





39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Tabular Data (the 'design matrix') - *x*

n observations: the number of times the phenomenon we need to 'describe' is available in our data through historical examples

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Tabular Data (the ‘design matrix’) - *x*

What will be better? High values of *p* and/or *n*?

p attributes (variables, features) potentially related to the phenomenon under examination

n observations: the number of times the phenomenon we need to 'describe' is available in our data through historical examples

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Tabular Data (the 'design matrix') - *x*

Not always so easy to define!

p attributes (variables, features) potentially related to the phenomenon under examination

n observations: the number of times the phenomenon we need to 'describe' is available in our data through historical examples

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Tabular Data (the ‘design matrix’) - *x*

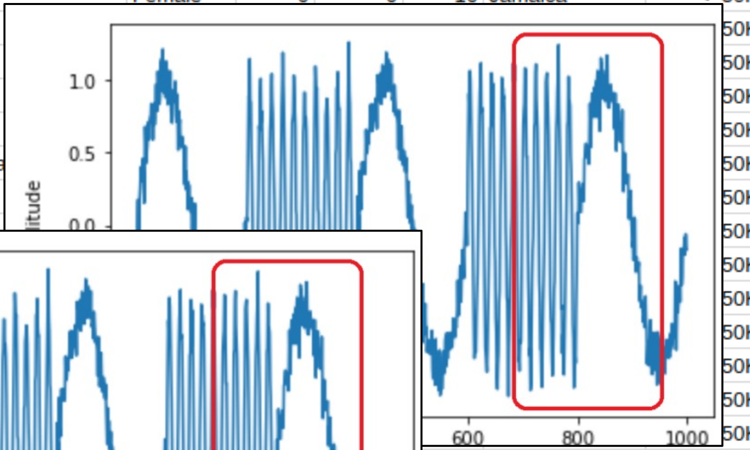
Not always so easy to define!

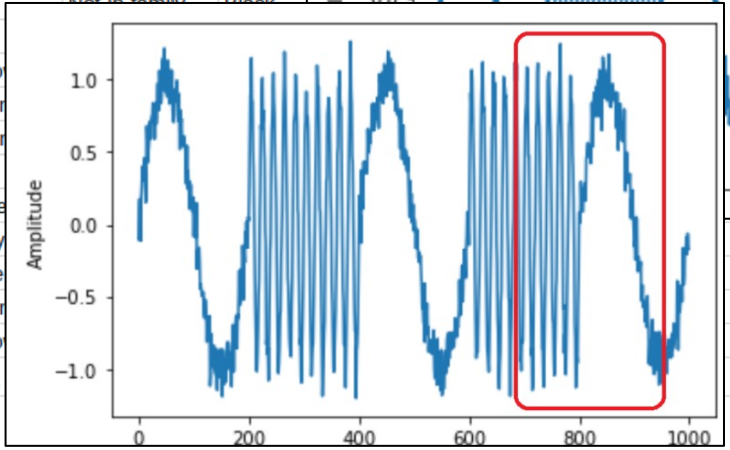
p attributes (variables, features) potentially related to the phenomenon under examination

n observations: the number of times the phenomenon we need to 'describe' is available in our data through historical examples

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White						50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White						50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White						50K
37	Private					Exec-managerial	Husband	Black						50K
30	State-gov					Prof-specialty	Husband	Asian-Pa						50K
23	Private					Adm-clerical	Own-child	White						50K
32	Private					Sales								50K
40	Private					Craft-repair								50K
34	Private					Transport-mo								50K
25	Self-emp-not-inc					Farming-fishi								50K
32	Private					Machine-op-i								50K
38	Private					Sales								50K
43	Self-emp-not-inc					Exec-manage								50K
40	Private					Prof-specialty						60	United-States	>50K
54	Private					Other-service						20	United-States	<=50K
35	Federal-gov					Farming-fishi						40	United-States	<=50K
43	Private					Transport-mo						40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support						40	United-States	<=50K

For example, a system described by (multiple) time-serie data: maybe we need to extract quantities from each time window





Tabular Data (the ‘design matrix’) - *x*

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

At same point, this matrix (or part of this... more on this later) will be fed to a machine learning model!

To do so, data has:

- to be ‘consistent’;
- all variables should ‘be treated’ equally (unless we have a priori knowledge, all variables can contribute to understand/describe the phenomena);
- we should make life easy for a model and do not provide redundant/useless information

Tabular Data (the ‘design matrix’) - *x*

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

At same point, this matrix (or part of this... more on this later) will be fed to a machine learning model!

Statistical moments/quantities can be of help!

To do so, data has:

- to be ‘consistent’;
- all variables should ‘be treated’ equally (unless we have a priori knowledge, all variables can contribute to understand/describe the phenomena);
- we should make life easy for a model and do not provide redundant/useless information

Many technologies, let’s put some order

Different tasks (objectives), different models, different data type… and different stages of development!

Problem
- Definition
- Expected Impact
- Evaluation metric

Collection
- Conversion
- Parsing
- Aggregation
- Alignment

Cleaning
- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

Modelling
- Feature Extraction
- Building
- Evaluation/ Comparison

Roll-out
- On-line implementation
- Business outcome
- Improvement

A Machine Learning pipeline

Problem
- Definition
- Expected Impact
- Evaluation metric

Collection
- Conversion
- Parsing
- Aggregation
- Alignment

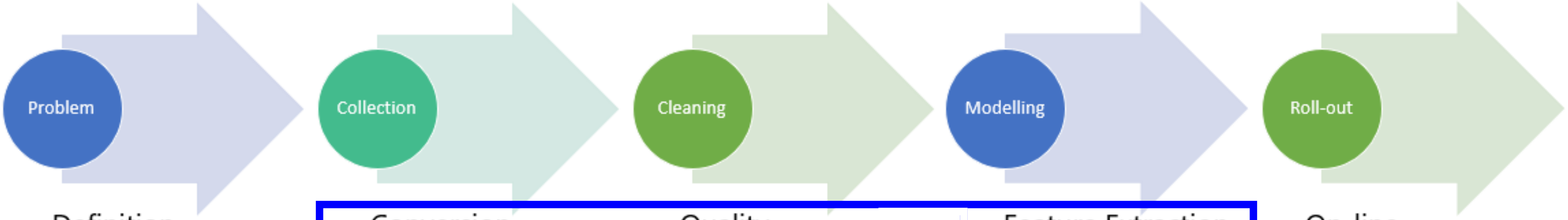
Cleaning
- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

Modelling
- Feature Extraction
- Building
- Evaluation/ Comparison

Roll-out
- On-line implementation
- Business outcome
- Improvement

Main focus of the course (from week 3)

A Machine Learning pipeline



- Definition
- Expected Impact
- Evaluation metric

- Conversion
- Parsing
- Aggregation
- Alignment

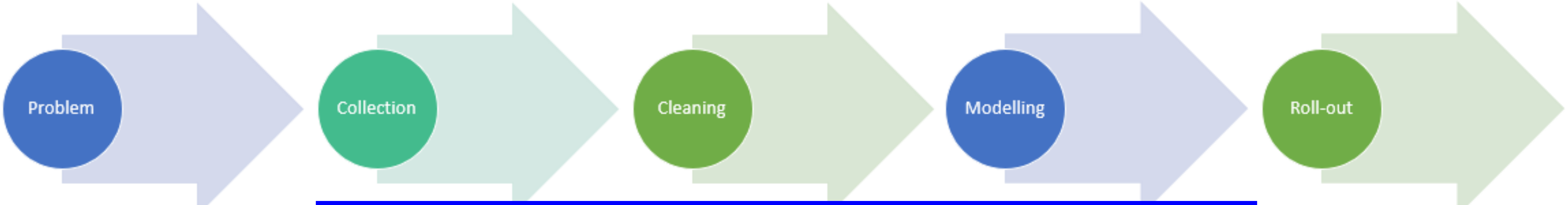
- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

- Feature Extraction
- Building
- Evaluation/ Comparison

- On-line implementation
- Business outcome
- Improvement

Preprocessing

A Machine Learning pipeline



Problem
- Definition
- Expected Impact
- Evaluation metric

Collection
- Conversion
- Parsing
- Aggregation
- Alignment

Cleaning
- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

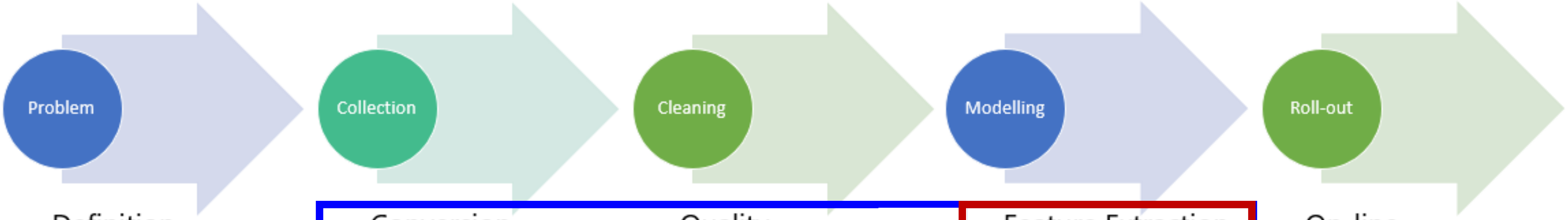
Modelling
- Feature Extraction
- Building
- Evaluation/ Comparison

Roll-out
- On-line implementation
- Business outcome
- Improvement

Preprocessing

We will have a dedicated lecture

A Machine Learning pipeline



- Definition
- Expected Impact
- Evaluation metric

- Conversion
- Parsing
- Aggregation
- Alignment

- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

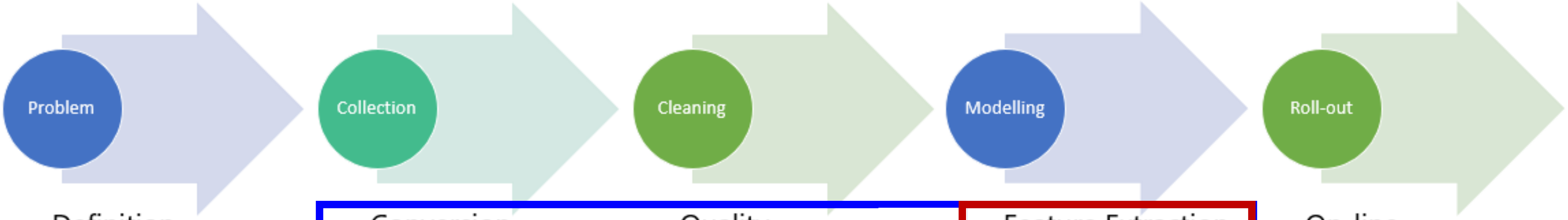
Preprocessing

- Feature Extraction
- Building
- Evaluation/ Comparison

Many approaches, but also statistical ones!

- On-line implementation
- Business outcome
- Improvement

A Machine Learning pipeline



- Definition
- Expected Impact
- Evaluation metric

- Conversion
- Parsing
- Aggregation
- Alignment

- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

Preprocessing

- Feature Extraction
- Building
- Evaluation/ Comparison

- On-line implementation
- Business outcome
- Improvement

Many approaches, but also statistical ones!

Statistical tools, not only for preprocessing!

A Machine Learning pipeline

- Definition
- Expected Impact
- Evaluation metric

- Conversion
- Parsing
- Aggregation
- Alignment

- Quality
- Reconciliation
- Missing data handling
- Denoising
- Outlier detection

We will have a dedicated lecture

- Feature Extraction
- Building
- Evaluation/ Comparison

Many approaches, but also statistical ones!

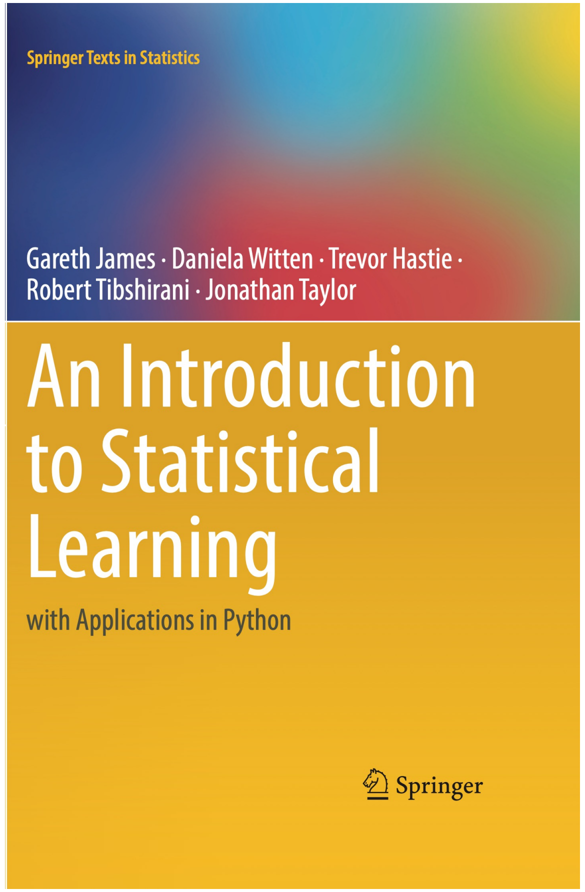
Not relevant only for pre-

- On-line implementation
- Business outcome
- Improvement

Why use Statistics in Machine Learning?

Why Use Statistics in ML?

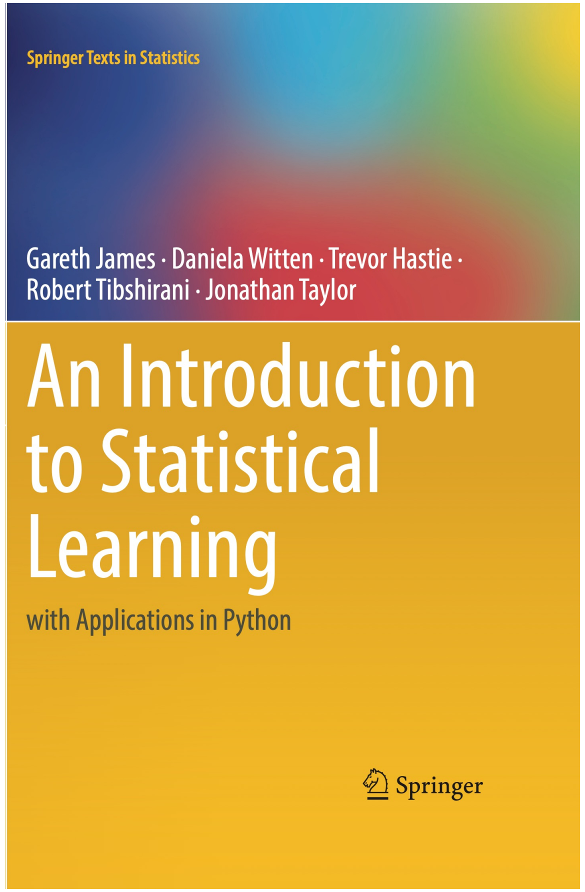
1. [Preprocessing] Data Understanding – Descriptive statistics (mean, variance, distributions) help explore and clean data, identifying patterns and outliers.
2. [Preprocessing] Feature Engineering – Techniques like correlation analysis, PCA, and scaling rely on statistical principles.
3. [Building] Probability & Uncertainty – ML often deals with probabilistic models (e.g., Naïve Bayes) and uncertainty estimation.
4. [Building] Generalization & Inference – Concepts like overfitting, hypothesis testing, and bias-variance tradeoff come from statistics.
5. [Evaluation] Model Evaluation – Metrics like MSE, MAE, accuracy, precision, and recall are rooted in statistical concepts.



Why use Statistics in Machine Learning?

Why Use Statistics in ML?

1. [Preprocessing] Data Understanding – Descriptive statistics (mean, variance, distributions) help explore and clean data, identifying patterns and outliers.
2. [Preprocessing] Feature Engineering – Techniques like correlation analysis, PCA, and scaling rely on statistical principles.
3. [Building] Probability & Uncertainty – ML often deals with probabilistic models (e.g., Naïve Bayes) and uncertainty estimation.
4. [Building] Generalization & Inference – Concepts like overfitting, hypothesis testing, and bias-variance tradeoff come from statistics.
5. [Evaluation] Model Evaluation – Metrics like MSE, MAE, accuracy, precision, and recall are rooted in statistical concepts.



Statistical moments and other statistical quantities

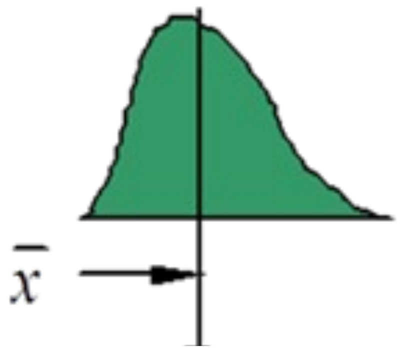
Statistical moments in Machine Learning

Moments of a Random Variable X

1.First Moment (Mean) – Central Tendency

$E[X] = \mu$

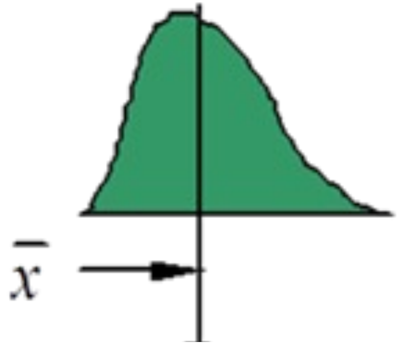
The expected value of X, representing the average outcome.



Statistical moments in Machine Learning

Moments of a Random Variable X

1.First Moment (Mean) – Central Tendency
E[X] = μ
The expected value of X, representing the average outcome.



Difference between Statistical Moments (theoretical) and Statistical Moments Computed with Sampling (with available datA)

$$\mu_k = E[X^k] = \int_{-\infty}^{\infty} x^k f(x)dx \quad \text{(for continuous distributions)}$$

$$\mu_k = \sum_x x^k P(X = x) \quad \text{(for discrete distributions)}$$

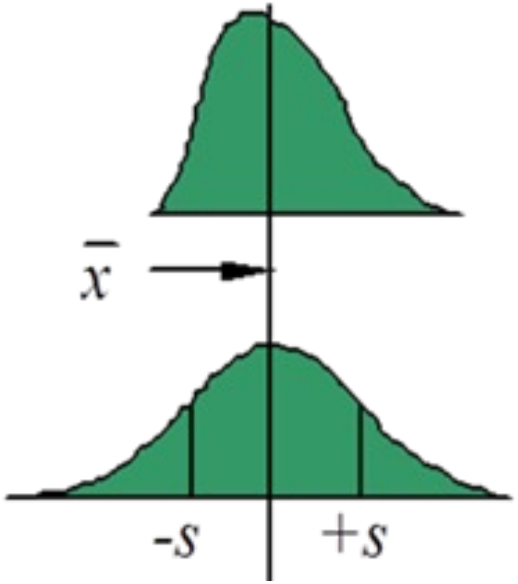
$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

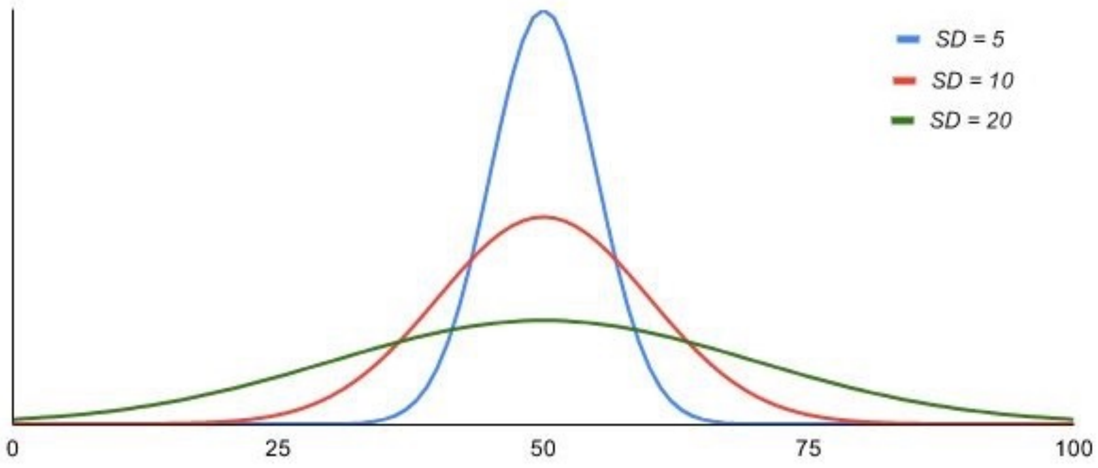
Statistical moments in Machine Learning

Moments of a Random Variable X

1.First Moment (Mean) – Central Tendency
$E[X] = \mu$
The expected value of X, representing the average outcome.

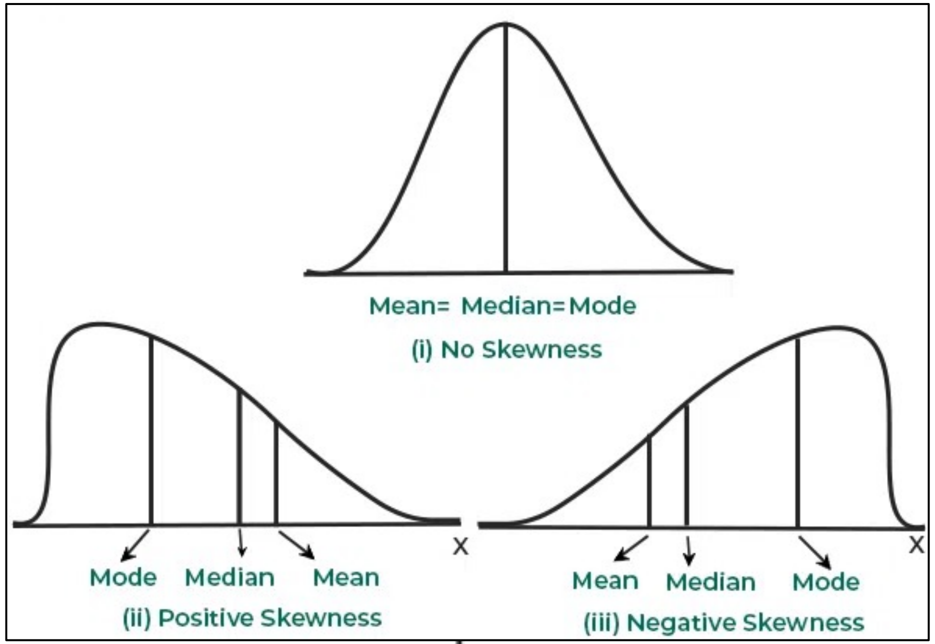
2.Second Moment (Variance) – Spread of Data
$Var(X) = E[(X - \mu)^2] = \sigma^2$
Measures how far values of X deviate from the mean.





Distributions with the same mean, but different variance

Statistical moments in Machine Learning

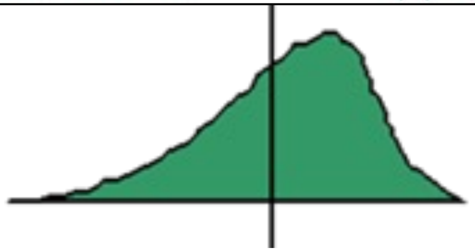


Moments of a Random Variable X

1.First Moment (Mean) – Central Tendency
$E[X] = \mu$
The expected value of X, representing the average outcome.

2.Second Moment (Variance) – Spread of Data
$Var(X) = E[(X - \mu)^2] = \sigma^2$
Measures how far values of X deviate from the mean.

3.Third Moment (Skewness) – Asymmetry of Distribution
$Skew(X) = E[(X - \mu)^3] / \sigma^3$
Indicates whether the distribution leans right (negative skew) or left (positive skew).

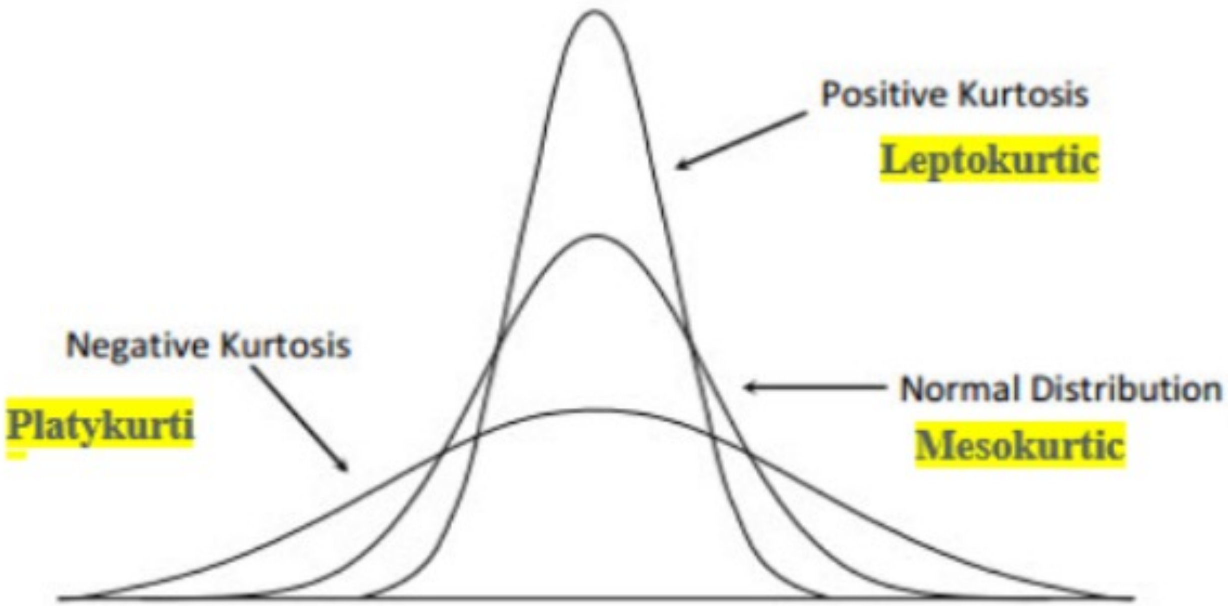


Statistical moments in Machine Learning

Normal Kurtosis (K = 3) - Mesokurtic
•The distribution has the same shape as the normal distribution.

High Kurtosis (K > 3) - Leptokurtic
•Heavier tails than the normal distribution, meaning more extreme values (outliers).
•The distribution is more "peaked" in the center and has longer tails.

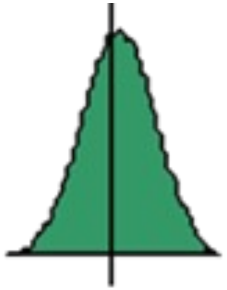
Low Kurtosis (K < 3, K_excess < 0) - Platykurtic
•Lighter tails than the normal distribution, meaning fewer extreme values.
•The distribution is "flatter" in the center with shorter tails.



4.Fourth Moment (Kurtosis) – Tailedness of Distribution

$Kurt(X) = E[(X - \mu)^4] / \sigma^4$

Measures how heavy or light the tails of the distribution are compared to a normal distribution.



Statistical moments in Machine Learning

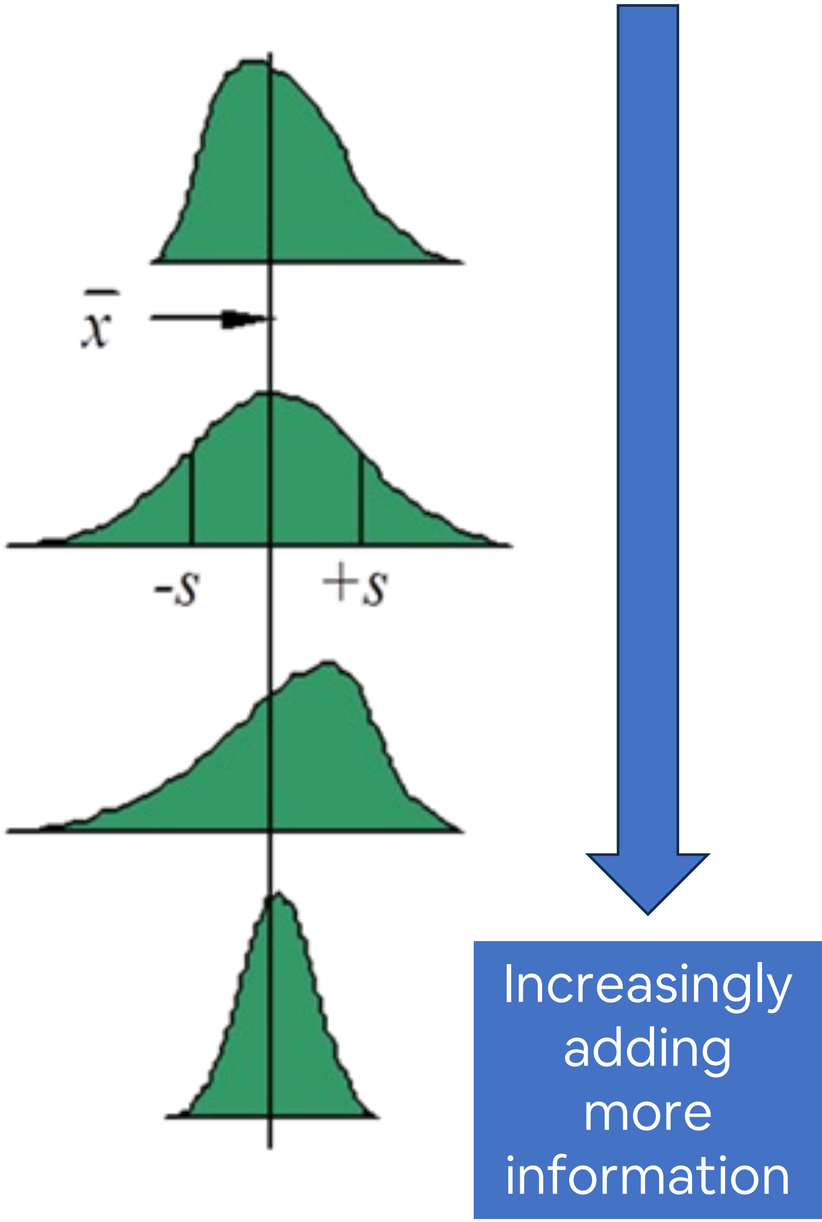
Moments of a Random Variable X

1.First Moment (Mean) – Central Tendency
$E[X] = \mu$
The expected value of X, representing the average outcome.

2.Second Moment (Variance) – Spread of Data
$Var(X) = E[(X - \mu)^2] = \sigma^2$
Measures how far values of X deviate from the mean.

3.Third Moment (Skewness) – Asymmetry of Distribution
$Skew(X) = E[(X - \mu)^3] / \sigma^3$
Indicates whether the distribution leans right (negative skew) or left (positive skew).

4.Fourth Moment (Kurtosis) – Tailedness of Distribution
$Kurt(X) = E[(X - \mu)^4] / \sigma^4$
Measures how heavy or light the tails of the distribution are compared to a normal distribution.



A numerical example

A *n = 10* dataset

X=[10,15,20,25,30,35,40,45,50,55]

Let's compute the mean:

$$\mu = \frac{\sum X_i}{n} = \frac{10+15+20+25+30+35+40+45+50+55}{10} = \frac{325}{10} = 32.5$$

A numerical example

Let’s compute the variance:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$$

A numerical example

Let's compute the variance:

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{n}$$

$$(10 - 32.5)^2 = 552.25$$

$$(15 - 32.5)^2 = 306.25$$

$$(20 - 32.5)^2 = 156.25$$

$$(25 - 32.5)^2 = 56.25$$

$$(30 - 32.5)^2 = 6.25$$

$$(35 - 32.5)^2 = 6.25$$

$$(40 - 32.5)^2 = 56.25$$

$$(45 - 32.5)^2 = 156.25$$

$$(50 - 32.5)^2 = 306.25$$

$$(55 - 32.5)^2 = 552.25$$

A numerical example

Let's compute the variance:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$$

$$(10 - 32.5)^2 = 552.25$$

$$(15 - 32.5)^2 = 306.25$$

$$(20 - 32.5)^2 = 156.25$$

$$(25 - 32.5)^2 = 56.25$$

$$(30 - 32.5)^2 = 6.25$$

$$(35 - 32.5)^2 = 6.25$$

$$(40 - 32.5)^2 = 56.25$$

$$(45 - 32.5)^2 = 156.25$$

$$(50 - 32.5)^2 = 306.25$$

$$(55 - 32.5)^2 = 552.25$$

$$552.25 + 306.25 + 156.25 + 56.25 + 6.25 + 6.25 + 56.25 + 156.25 + 306.25 + 552.25 = 2062.5$$

$$\sigma^2 = \frac{2062.5}{10} = 206.25$$

A numerical example

Let's compute the skewness:

$$\text{Skew}(X) = \frac{\sum(X_i - \mu)^3}{n \cdot \sigma^3}$$

$$(10 - 32.5)^3 = -12903.125$$
$$(15 - 32.5)^3 = -5359.375$$
$$(20 - 32.5)^3 = -1953.125$$
$$(25 - 32.5)^3 = -421.875$$
$$(30 - 32.5)^3 = -15.625$$
$$(35 - 32.5)^3 = 15.625$$
$$(40 - 32.5)^3 = 421.875$$
$$(45 - 32.5)^3 = 1953.125$$
$$(50 - 32.5)^3 = 5359.375$$
$$(55 - 32.5)^3 = 12903.125$$

$$-12903.125 + (-5359.375) + (-1953.125) + (-421.875) + (-15.625) + 15.625 + 421.875 + 1953.125 + 5359.375 + 12903.125 = 0$$

$$\text{Skew}(X) = \frac{0}{10 \cdot (206.25)^{3/2}} = 0$$

A numerical example

Let's compute the skewness:

$$\text{Skew}(X) = \frac{\sum (X_i - \mu)^3}{n \cdot \sigma^3}$$

$$(10 - 32.5)^3 = -12903.125$$
$$(15 - 32.5)^3 = -5359.375$$
$$(20 - 32.5)^3 = -1953.125$$
$$(25 - 32.5)^3 = -421.875$$
$$(30 - 32.5)^3 = -15.625$$
$$(35 - 32.5)^3 = 15.625$$
$$(40 - 32.5)^3 = 421.875$$
$$(45 - 32.5)^3 = 1953.125$$
$$(50 - 32.5)^3 = 5359.375$$
$$(55 - 32.5)^3 = 12903.125$$

We have a simmetric distribution!

$$-12903.125 + (-5359.375) + (-1953.125) + (-421.875) + (-15.625) + 15.625 + 421.875 + 1953.125 + 5359.375 + 12903.125 = 0$$

$$\text{Skew}(X) = \frac{0}{10 \cdot (206.25)^{3/2}} = 0$$

A numerical example

Let's compute the kurtosis:

$$\text{Kurt}(X) = \frac{\sum (X_i - \mu)^4}{n \cdot \sigma^4}$$

$$(10 - 32.5)^4 = 3013025.5625$$
$$(15 - 32.5)^4 = 938906.25$$
$$(20 - 32.5)^4 = 244140.625$$
$$(25 - 32.5)^4 = 31640.625$$
$$(30 - 32.5)^4 = 39.0625$$
$$(35 - 32.5)^4 = 39.0625$$
$$(40 - 32.5)^4 = 31640.625$$
$$(45 - 32.5)^4 = 244140.625$$
$$(50 - 32.5)^4 = 938906.25$$
$$(55 - 32.5)^4 = 3013025.5625$$

The kurtosis is less than 3, so the distribution is platykurtic (fewer "tails" compared to the normal distribution).

$$3013025.56 + 938906.25 + 244140.62 + 31640.62 + 39.06 + 39.06 + 31640.62 + 244140.62 + 938906.25 + 3013025.56 = 8282825$$

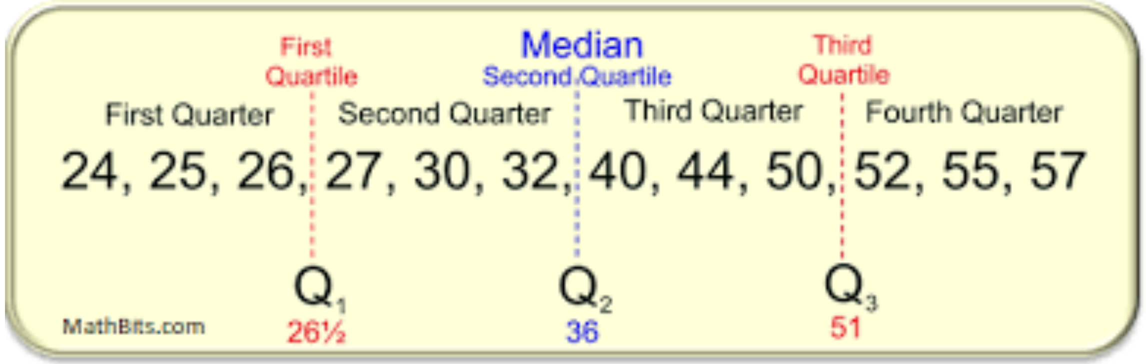
$$\text{Kurt}(X) = \frac{8282825}{10 \cdot (206.25)^2} = \frac{8282825}{42514.0625} = 1.8$$

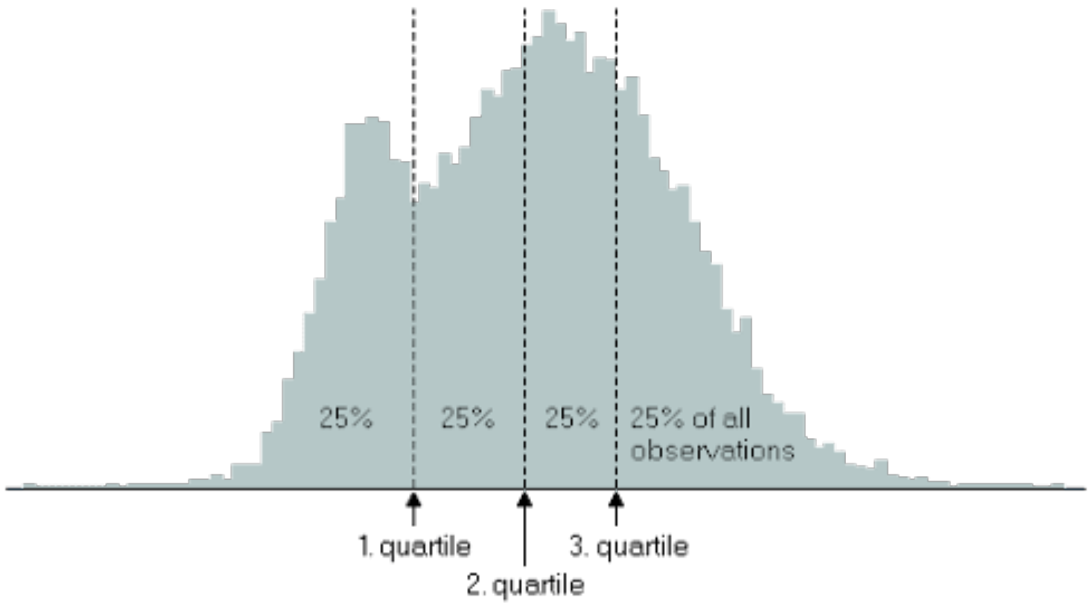
Quartiles

Quartiles divide a dataset into **four equal parts**, helping to understand the distribution and spread of the data.

Quartile Definitions:
- *Q1* (First Quartile, 25%) – The value below which 25% of the data falls.
- *Q2* (Second Quartile, 50%) – The median, the value that splits the data into two equal halves.
- *Q3* (Third Quartile, 75%) – The value below which 75% of the data falls.
- Interquartile Range (*IQR*) – The range between Q1 and *Q3*, measuring the spread of the middle 50% of the data:

$$IQR = Q3 - Q1$$





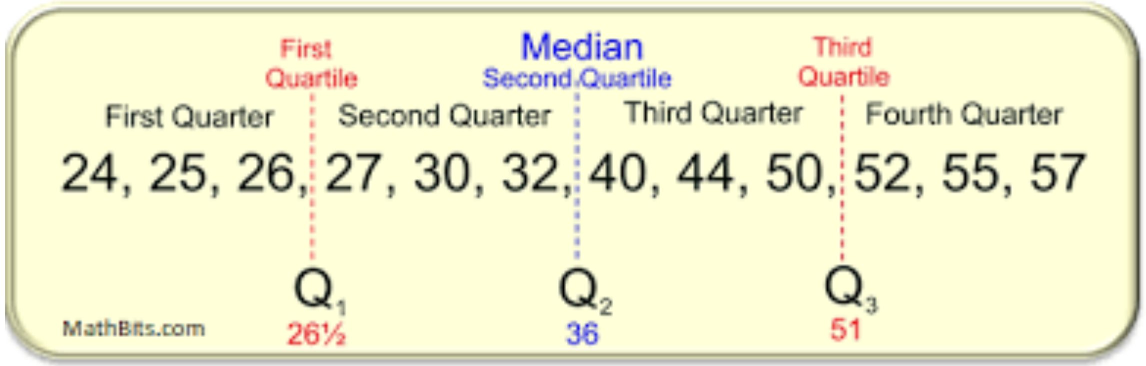
Quartiles

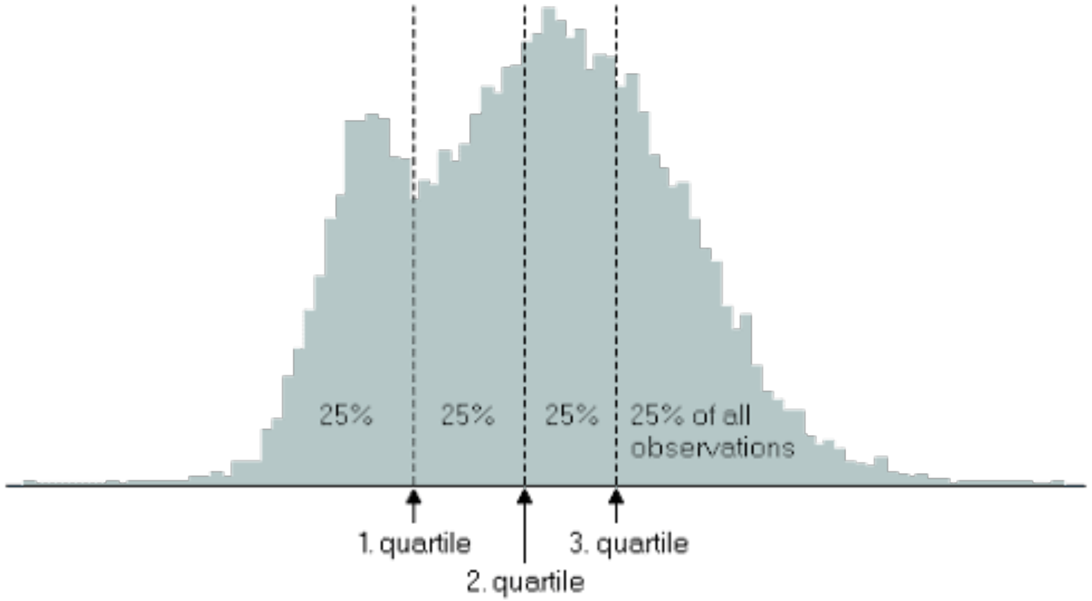
Quartiles divide a dataset into **four equal parts**, helping to understand the distribution and spread of the data.

Quartile Definitions:
- *Q1* (First Quartile, 25%) – The value below which 25% of the data falls.
- *Q2* (Second Quartile, 50%) – The median, the value that splits the data into two equal halves.
- *Q3* (Third Quartile, 75%) – The value below which 75% of the data falls.
- Interquartile Range (*IQR*) – The range between Q1 and *Q3*, measuring the spread of the middle 50% of the data:

$$IQR = Q3 - Q1$$

The median is used many times instead of the mean, as it is a robust quantity w.r.t. 'outliers' (strange data)





A numerical example (odd numbers)

Same dataset previously seen (*n = 11*).

$$X = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$Q2 =$$

$$Q3 =$$

$$Q1 =$$

$$IQR =$$

A numerical example (odd numbers)

Same dataset previously seen (*n = 11*).

$$X = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$Q2 = 30$$

$$Q3 =$$

$$Q1 =$$

$$IQR =$$

A numerical example (odd numbers)

Same dataset previously seen (*n = 11*).

$$X = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$Q2 = 30$$

$$Q3 = 45$$

$$Q1 = 15$$

$$IQR = Q3 - Q1 = 45 - 15 = 30$$

A numerical example (even numbers)

Same dataset previously seen (*n = 10*).

$$X = [10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$Q1 =$$

$$Q2 =$$

$$Q3 =$$

$$IQR =$$

A numerical example (even numbers)

Same dataset previously seen (*n = 10*).

$$X = [10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$Q1 = \frac{20 + 25}{2} = 22.5$$

$$Q2 = \frac{30 + 35}{2} = 32.5$$

$$Q3 = \frac{45 + 50}{2} = 47.5$$

$$IQR = Q3 - Q1 = 47.5 - 22.5 = 25$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%]$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:
- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%]$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[9.09\%, 18.18\%, 27.27\%, 36.36\%, 45.45\%, 54.55\%, 63.64\%, 72.73\%, 81.82\%, 90.91\%, 100\%]$$

Common convention (for the theoric part of the exam): no interpolation!

Procedure:

- Order the array
- Compute the exact percentage
- Always take the closest number which percentage is AT LEAST 25%/50%/75% for Q1/Q2/Q3 respectively

$$[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$$

$$[9.09\%, 18.18\%, 27.27\%, 36.36\%, 45.45\%, 54.55\%, 63.64\%, 72.73\%, 81.82\%, 90.91\%, 100\%]$$

Mode

The value that appears the most on a dataset. It is an important quantity when dealing with categhorical data.

Example:

$$X = [1, 2, 2, 3, 4, 7, 9, 10, 10, 10, 12]$$

$$\text{Mean} = \frac{\sum X_i}{n} = \frac{1+2+2+3+4+7+9+10+10+10+12}{11} = \frac{70}{11} \approx 6.36$$

$$\text{Median} = 7$$

$$\text{Mode} = 10$$

Statistical moments (and quantities) in Machine Learning

Statistical moments and quartiles provide a structured way to summarize and understand data distributions, which is essential for building, evaluating, and interpreting machine learning models.

Moments help in other pre-processing steps, such as feature engineering or missing data handling.

Not only that: they can be useful in easing the ‘training’ procedure*

**What ‘training’ means it will be clearer when we talk about modelling, but it means tuning/finding the ‘right’ parameters in a given model*

‘Understand’ a dataset: Iris dataset

-‘Iris Classification’ dataset, Ronald Fisher (1936)
- Available on UCI ML Repository/Kaggle/… everywhere!
- L = 3 classes problem: classify Setosa, Versicolour and Virginica iris from data containing sepal and petal width and length – n = 150 samples, p = 4 variables

Wikipedia

Iris dataset

https://www.kaggle.com/datasets/uciml/iris

# SepalLengthCm Length of the sepal (in cm)	# SepalWidthCm Width of the sepal (in cm)	# PetalLengthCm Length of the petal (in cm)	# PetalWidthCm Width of the petal (in cm)	A Species Species name
 4.3 7.9	 2 4.4	 1 6.9	 0.1 2.5	**3** unique values
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa

Iris dataset

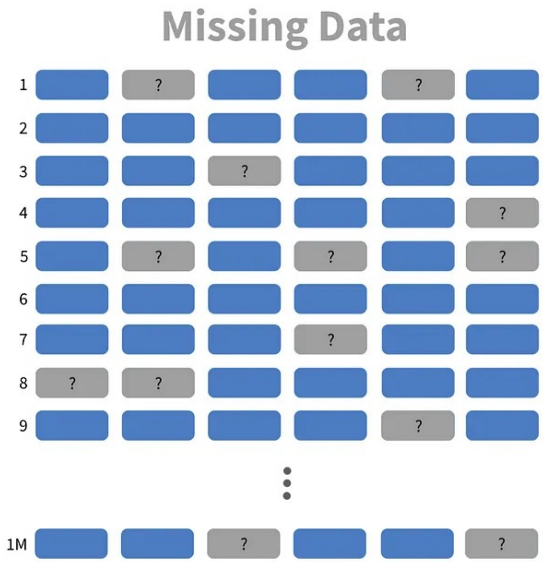
https://www.kaggle.com/datasets/uciml/iris

	Statistic	Value
0	Mean (First Moment)	3.758000
1	Variance (Second Moment)	3.095503
2	Skewness (Third Moment)	-0.274884
3	Kurtosis (Fourth Moment)	-1.402103
4	Q1 (First Quartile, 25%)	1.600000

# SepalLengthCm	# SepalWidthCm	# PetalLengthCm	# PetalWidthCm	Species
Length of the sepal (in cm)	Width of the sepal (in cm)	Length of the petal (in cm)	Width of the petal (in cm)	Species name
	4.4	 1 – 6.9	 0.1 – 2.5	3 unique values
		1.4	0.2	Iris-setosa
		1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa

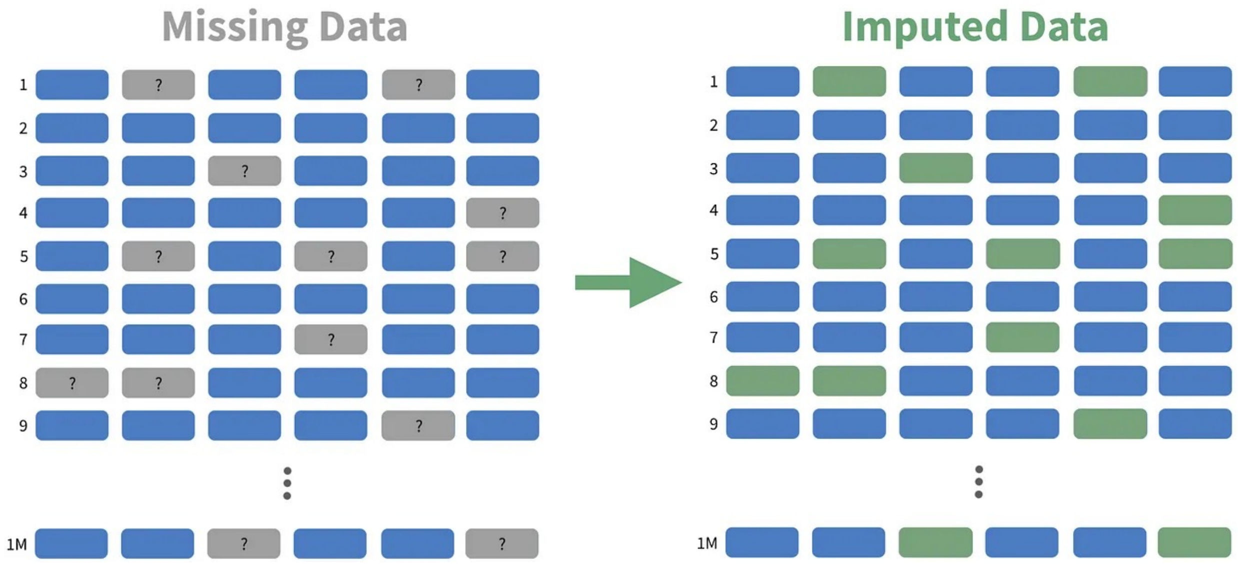
‘Correct’ a dataset

- Missing data are common in practical problems!
- Several models do not work with missing information
- Typically, we prefer not to throw away sample, instead we prefer to ‘impute’ data



‘Correct’ a dataset

- Missing data are common in practical problems!
- Several models do not work with missing information
- Typically, we prefer not to throw away sample, instead we prefer to ‘impute’ data
-Mean and median of a variable are typical choices

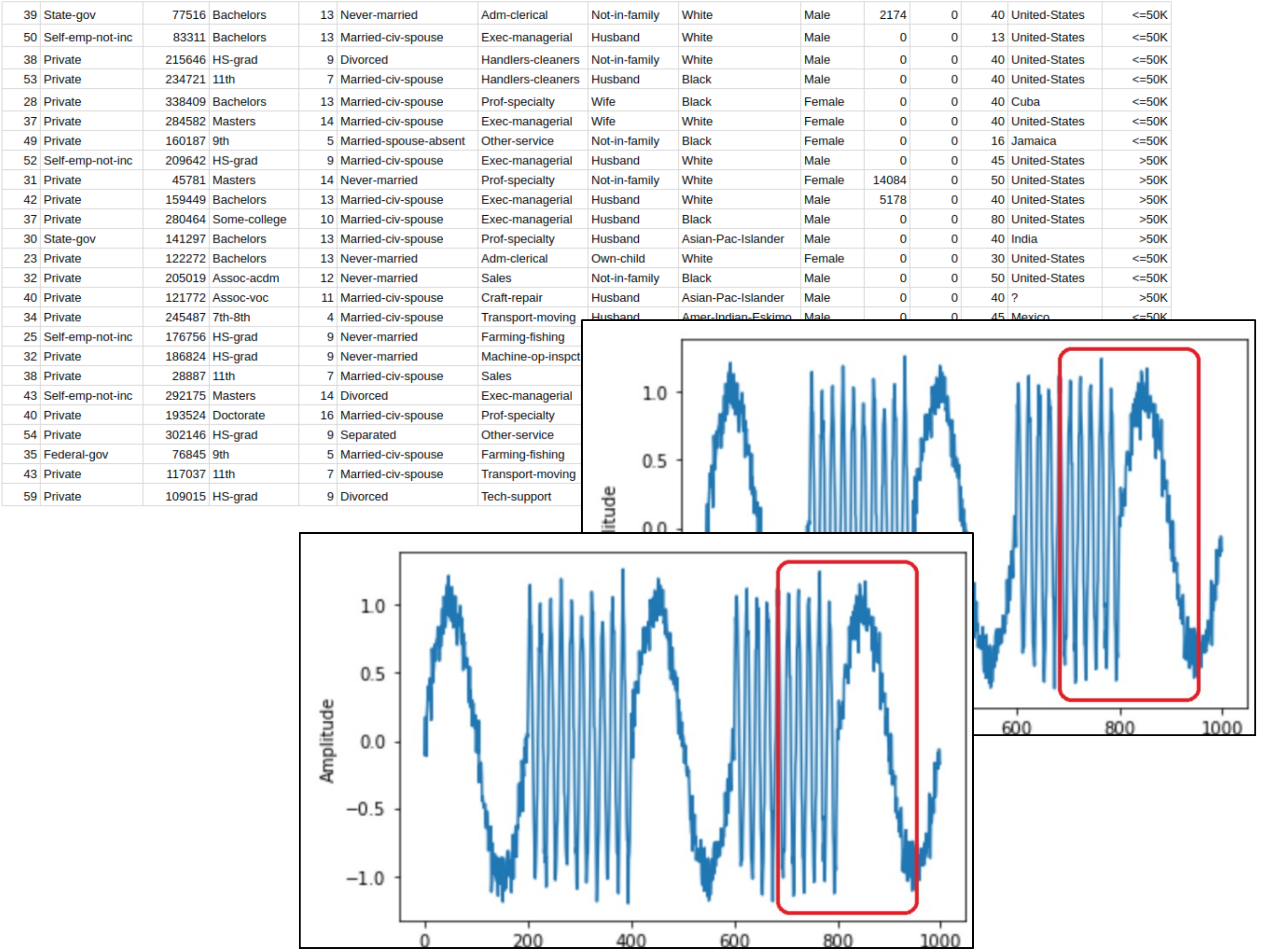


Enhance a dataset: feature engineering

- As said, data do not always present themselves in an easy tabular form

- We may use statistical moments/quantities (but also rule-based) for feature engineering

- Feature engineering is the process of creating, selecting and transforming ‘features’ (variables)



39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing								
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct								
38	Private	28887	11th	7	Married-civ-spouse	Sales								
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial								
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty								
54	Private	302146	HS-grad	9	Separated	Other-service								
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing								
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving								
59	Private	109015	HS-grad	9	Divorced	Tech-support								

Reduce a dataset!

- If not informative for the task, variables should be removed for efficiency and for better ‘engineering’ of a productive solutions
- This is typically not known a priori, and it should be done after/during modelling

	Feature_1	Feature_2	Constant_Var
1	54.96714153011233	23.636499344142013	100
2	48.61735698828815	23.668090197068675	100
3	56.47688538100692	26.084844859190753	100
4	65.23029856408026	30.495128632644757	100
5	47.658466252766644	28.638900372842315	100
6	47.6586304305082	25.824582803960837	100
7	65.79212815507391	32.23705789444759	100
8	57.67434729152909	22.789877213040835	100
9	45.30525614065048	25.842892970704362	100
10	55.42560043585965	27.327236865873836	100

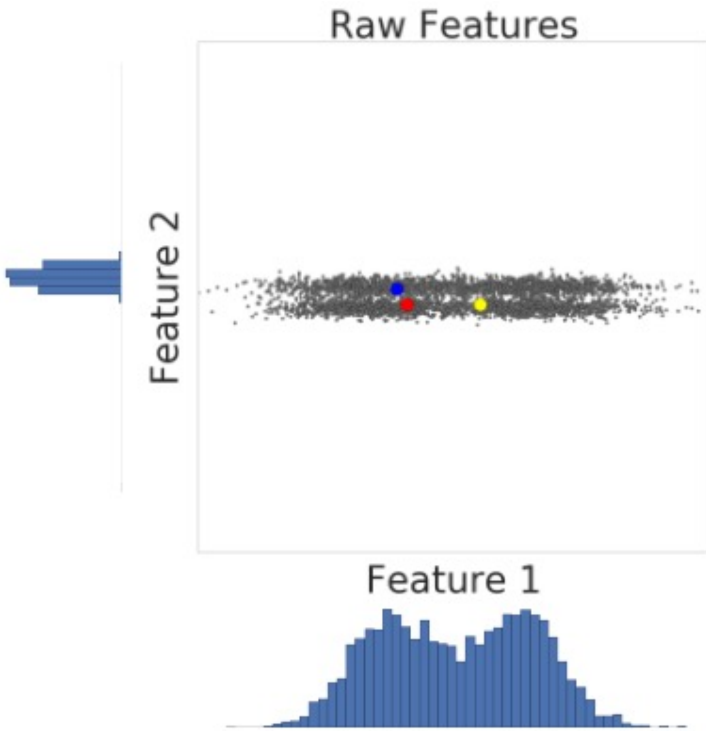
Reduce a dataset!

- If not informative for the task, variables should be removed for efficiency and for better ‘engineering’ of a productive solutions
- This is typically not known a priori, and it should be done after/during modelling
- However, if a variable is constant (variance = 0), we should get rid of it!

	Feature_1	Feature_2	Constant_Var
1	54.96714153011233	23.636499344142013	100
2	48.61735698828815	23.668090197068675	[illegible]
3	56.47688538100692	26.084844859190753	[illegible]
4	65.23029856408026	30.495128632644757	[illegible]
5	47.658466252766644	28.638900372842315	100
6	47.6586304305082	25.824582803960837	[illegible]
7	65.79212815507391	32.23705789444759	[illegible]
8	57.67434729152909	22.789877213040835	[illegible]
9	45.30525614065048	25.842892970704362	100
10	55.42560043585965	27.327236865873836	100

Optimize a dataset for modelling: data normalization

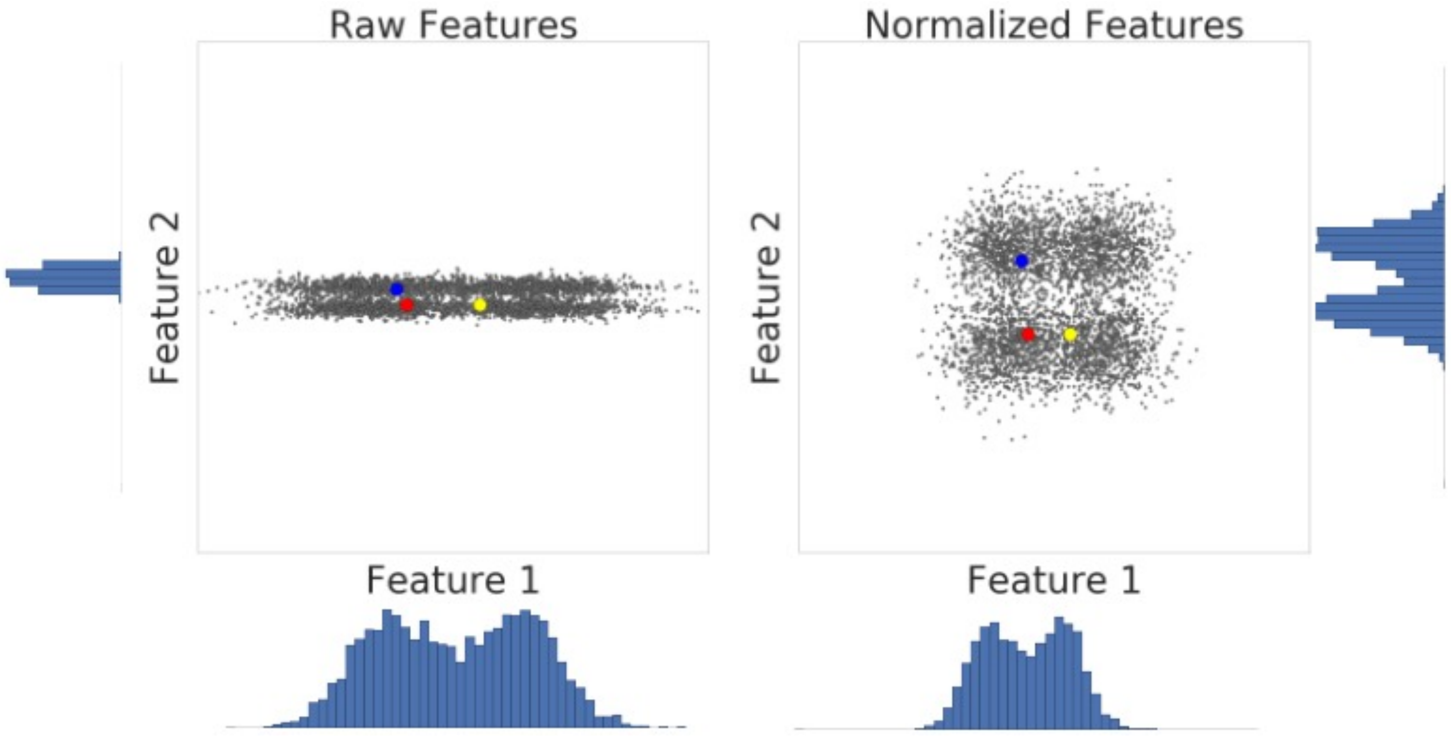
- A-priori, variables can be equally important in a ML task
- However, variable have different range values and one can ‘dominate’ the others
- Normalization can be of great help and it can speed up processing



Optimize a dataset for modelling: data normalization

- Z-score normalization (standardization) transforms each data (variables) to have a mean of 0 and a standard deviation of 1

$$Z = \frac{X - \mu}{\sigma}$$



Optimize a dataset for modelling: data normalization

X	Y
10	200
12	220
14	250
16	260
18	280

	X	Y
0	-1.264911	-1.315071
1	-0.632456	-0.688847
2	0.000000	0.250490
3	0.632456	0.563602
4	1.264911	1.189826

Optimize a dataset for modelling: data normalization

Pay attention: data normalization is a task that can make you save a lot of time during building of a model, but it is typically a forgotten step

Thank you!

Gian Antonio Susto