POLITECNICO
MILANO 1863

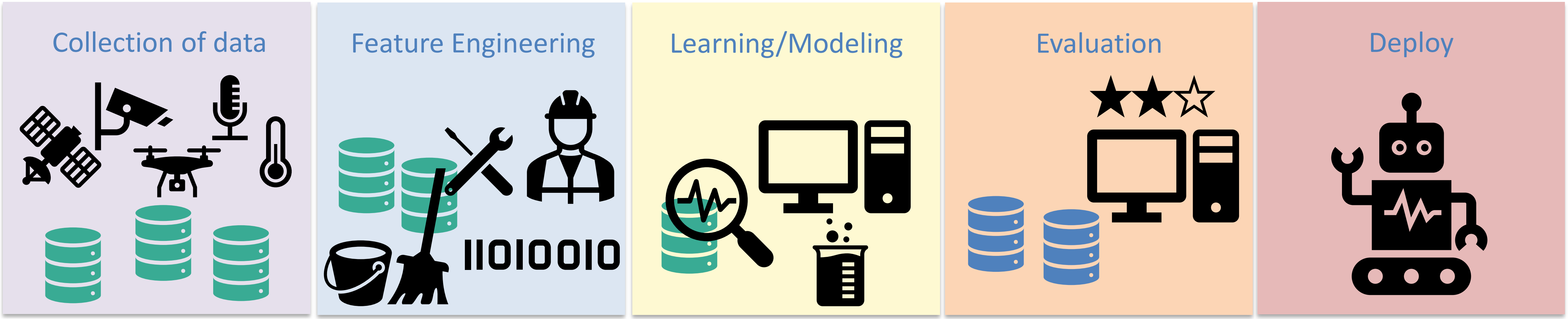
Data Analysis for Smart Agriculture

- Data Analysis Introduction -

Prof. Matteo Matteucci – matteo.matteucci@polimi.it

Data Analysis: the process

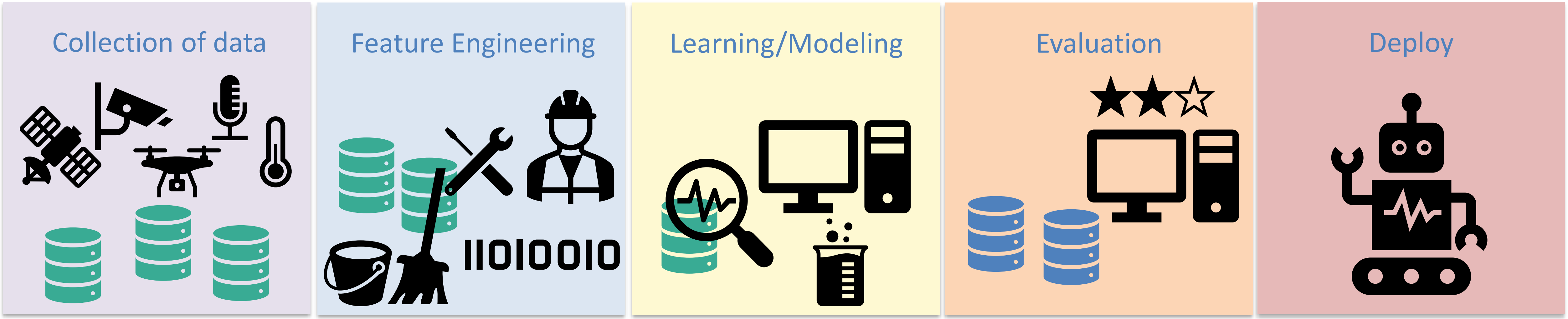
Data Analysis: the process



Feature Engineering

- Data Cleaning/Pre-Processing: Are there errors or inconsistencies in the data we need to eliminate?
- Feature Extraction: Need to elaborate existing variables to create new ones?
- Feature Selection: Which data we actually need to answer the posed question?

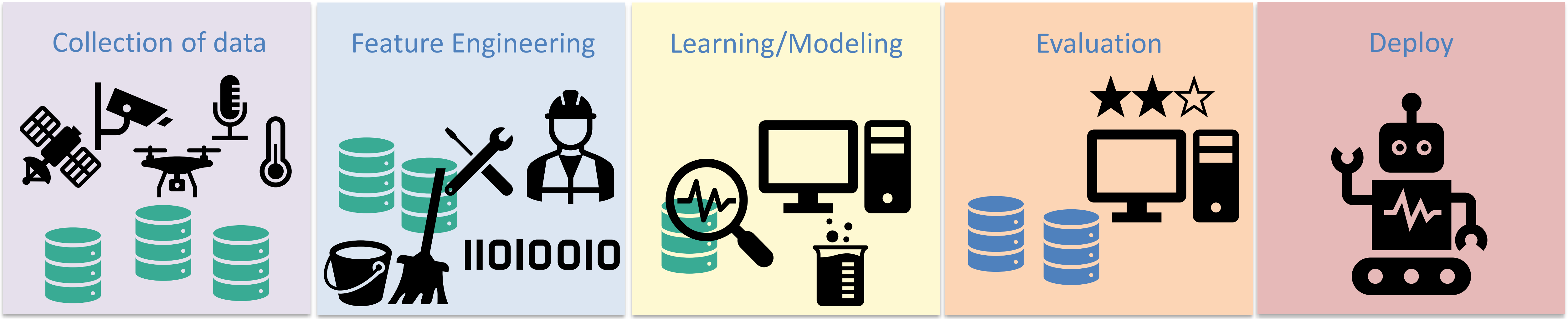
Data Analysis: the process



Learning/Modeling

- Select the learning task: *Classification*? *Regression*? *Clustering*? etc.
- Select the algorithm and perform learning/modeling

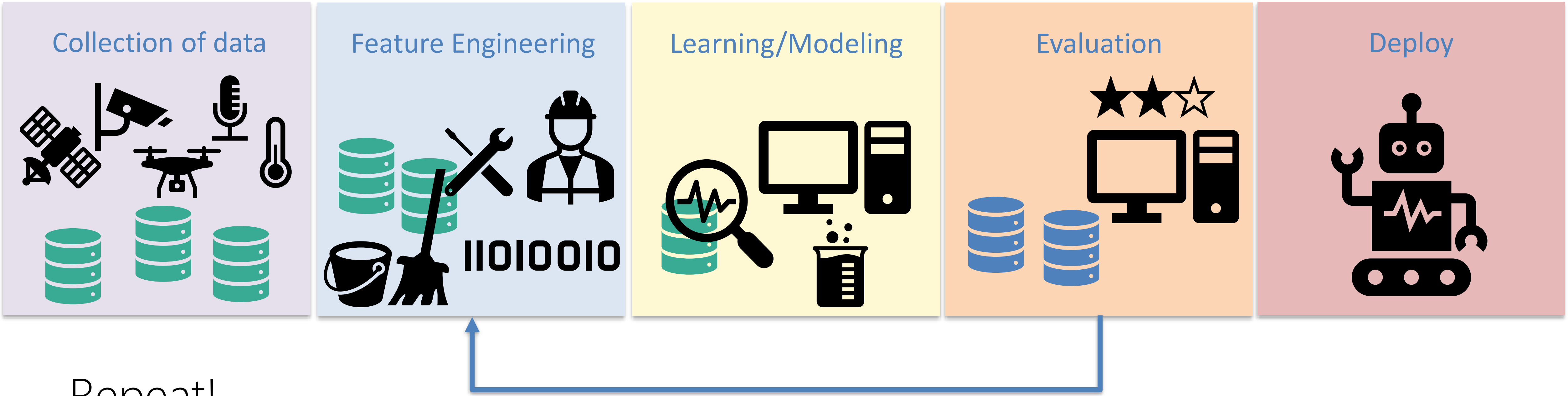
Data Analysis: the process



Evaluation

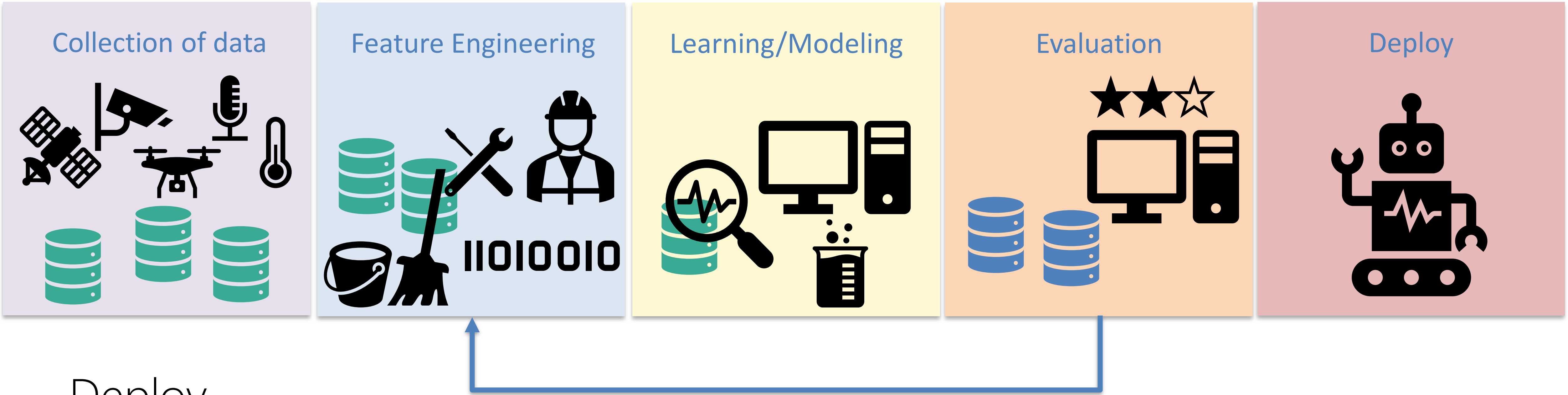
- Assess the performance of the learned model

Data Analysis: the process



Repeat!

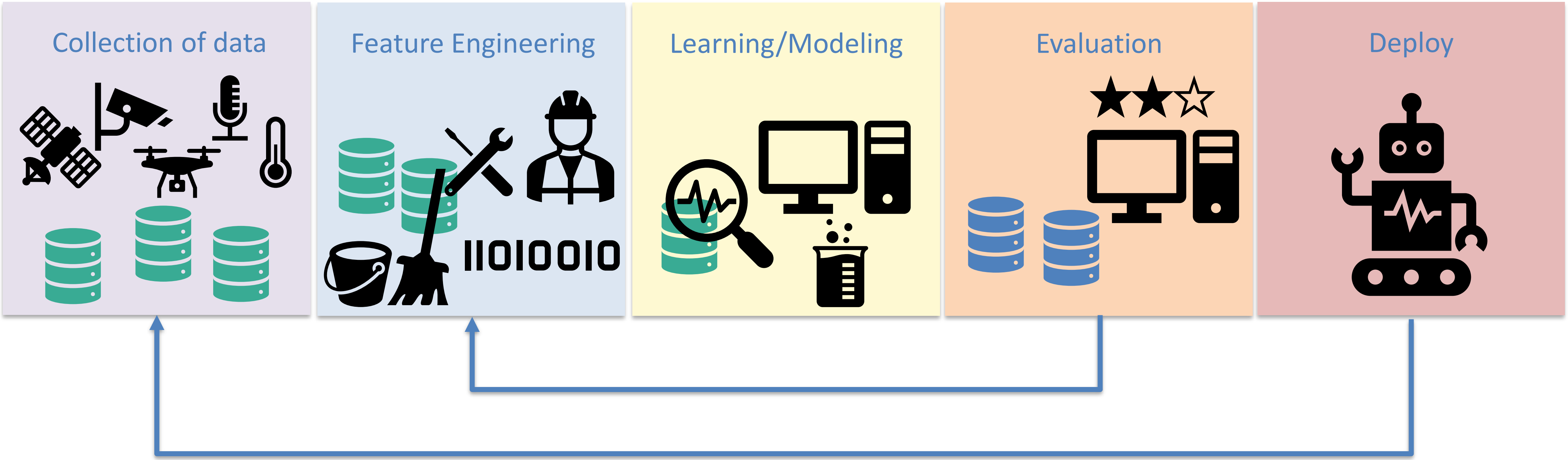
Data Analysis: the process



Deploy

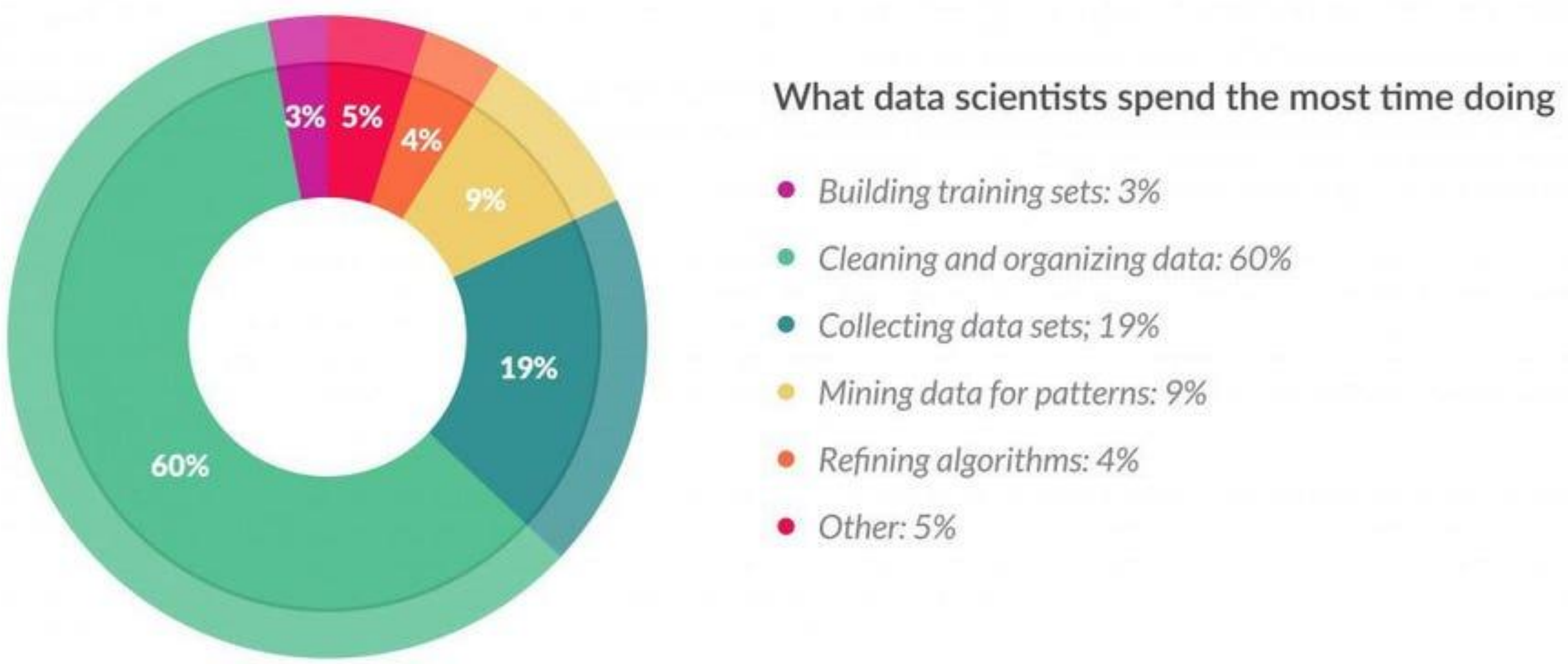
- The learned model is ready to be used in a real application. Until..

Data Analysis: the process



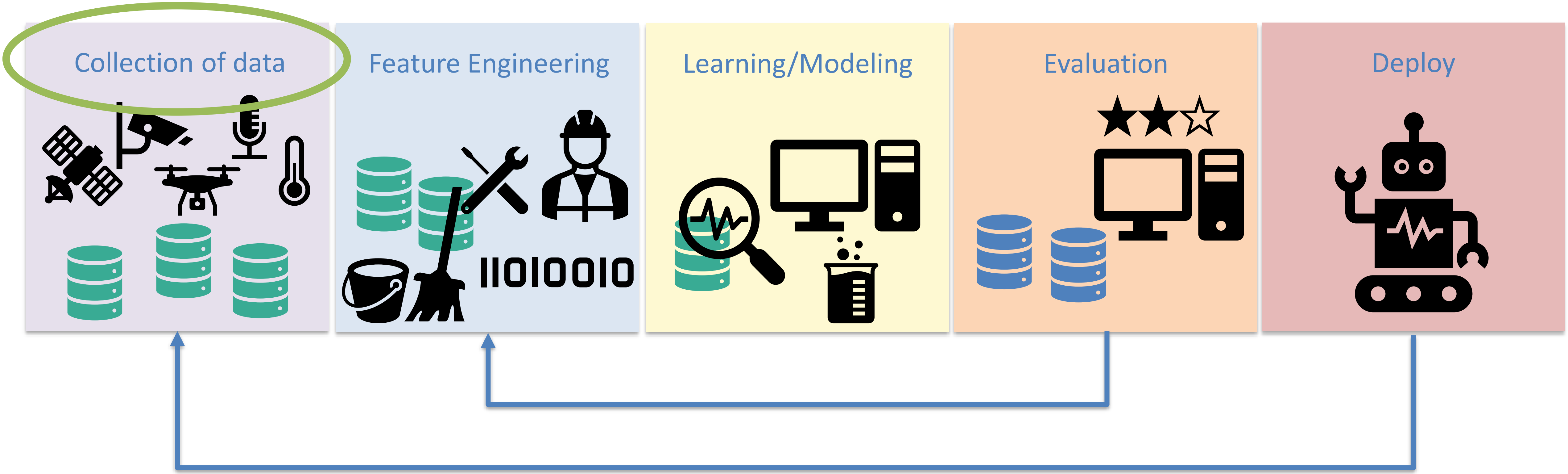
… until new data is available …

Data Analysis is all about …



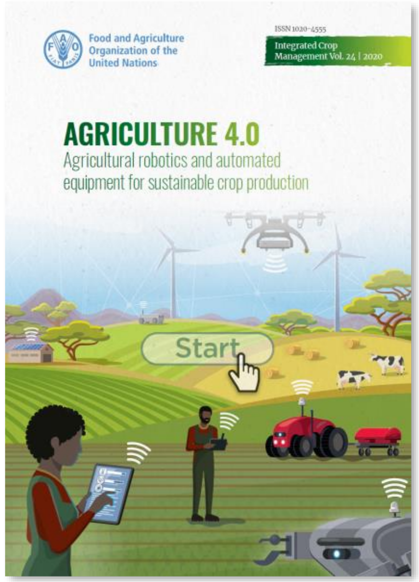
FigureEight (CrowdFlower) Data Science Report

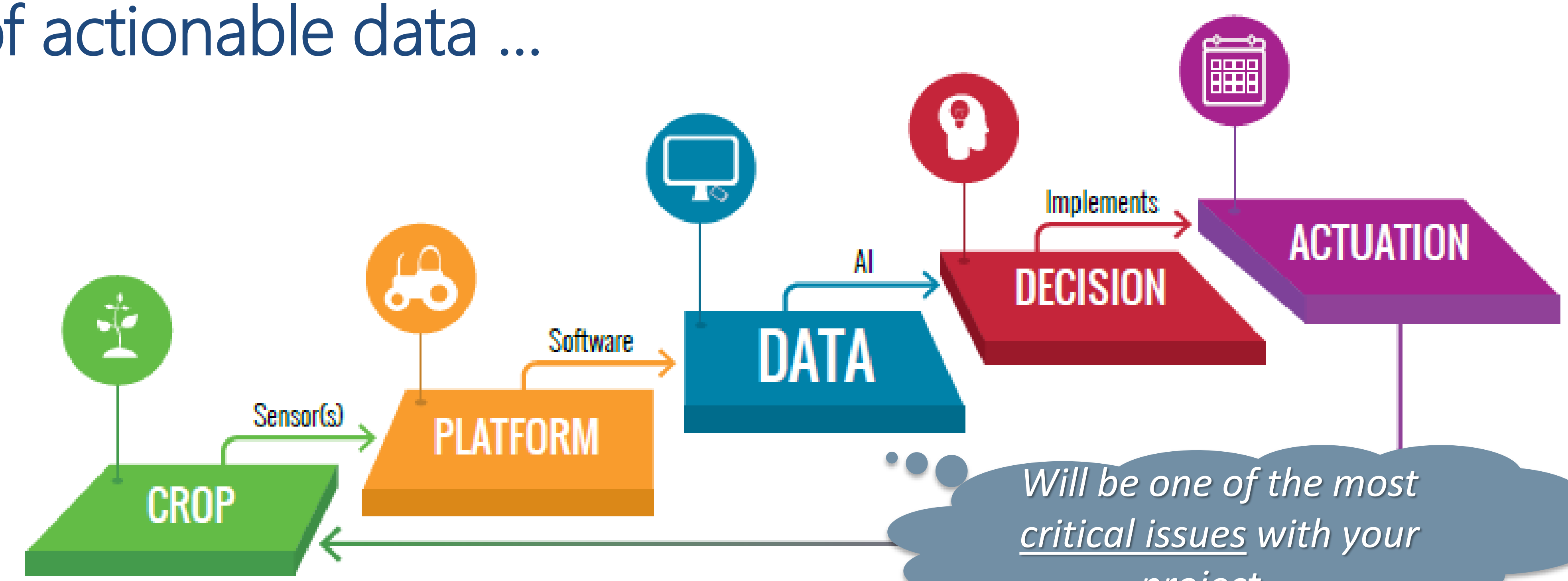
Data Analysis: the process



… until new data is available …

The value of actionable data …





Will be one of the most <u>critical issues</u> with your project …

days *minutes*

Data Analysis: the process

Feature Engineering

- Data Cleaning/Pre-Processing: Are there errors or inconsistencies in the data we need to eliminate?
- Feature Extraction: Need to elaborate existing variables to create new ones?
- Feature Selection: Which data we actually need to answer the posed question?

Features and Instances

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Gentoo	Biscoe	45.8	14.2	219.0	4700.0	Female
1	Gentoo	Biscoe	50.8	15.7	226.0	5200.0	Male
2	Chinstrap	Dream	46.9	16.6	192.0	2700.0	Female
3	Adelie	Torgersen	41.4	18.5	202.0	3875.0	Male
4	Adelie	Torgersen	34.6	21.1	198.0	4400.0	Male

Features and Instances …

Instances (or Examples) Features (or Attributes, or Variables)

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Gentoo	Biscoe	45.8	14.2	219.0	4700.0	Female
1	Gentoo	Biscoe	50.8	15.7	226.0	5200.0	Male
2	Chinstrap	Dream	46.9	16.6	192.0	2700.0	Female
3	Adelie	Torgersen	41.4	18.5	202.0	3875.0	Male
4	Adelie	Torgersen	34.6	21.1	198.0	4400.0	Male

Instances

- The atomic elements of information from a dataset
- Also known as examples, records, or prototypes,

Features

- Measures aspects of an instance
- Also known as attributes or variables
- Each instance is composed of a certain number of features

..and Concepts

Instances (or Examples)

Features (or Attributes, or Variables)

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Gentoo	Biscoe	45.8	14.2	219.0	4700.0	Female
1	Gentoo	Biscoe	50.8	15.7	226.0	5200.0	Male
2	Chinstrap	Dream	46.9	16.6	192.0	2700.0	Female
3	Adelie	Torgersen	41.4	18.5	202.0	3875.0	Male
4	Adelie	Torgersen	34.6	21.1	198.0	4400.0	Male

Concept

Concept

- Special content inside the data
- The things to be learned

Features Types (1/3)

Categorical Features

- Values are distinct symbols from a predefined set
- Typically used as labels or names in a non-numerical format

	species	island	sex
0	Gentoo	Biscoe	Female
1	Gentoo	Biscoe	Male
2	Chinstrap	Dream	Female
3	Adelie	Torgersen	Male
		Torgersen	Male

Nominal (Categorical) Features

- No relation is implied among nominal values
- No ordering, nor distance measure
- Only equality tests can be performed

Ordinal (Categorical) Features

- Impose order on values
- No distance between values defined
- Addition and subtraction don't make sense
- Distinction between nominal and ordinal not always clear

e.g.,
- Education level
  'high-school' < 'BS' < 'MS'
- Satisfaction rating
  'dislike' < 'neutral' < 'like'

Features Types (2/3)

Numerical Features

- Not only ordered but measured in fixed and equal units
- Sometimes they are divided into "discrete" (e.g., number of enrolled students) and "continuous" (e.g., height, weight)

Interval (Numerical) Features

- Difference of two values makes sense
- Zero point is not defined
- No concept of ratio between measurements

e.g.,
- Temperature in degrees
- Year
- …

Ratio (Numerical) Features

- Zero point is "naturally" defined
- Depending on the scientific knowledge
- Ratio between measurements makes sense

e.g.,
- Height
- Weight
- Salary
- …

Features Types (3/3)

What Features Types in Practice?

Discrete Features

- Values in a finite or countable set
- Examples: counts, set of words in a collection of documents
- Often represented as integer variables
- Note: Binary features are a special case of discrete features

Continuous Features

- Values are real numbers
- Examples: temperature, height, or weight
- Usually represented as floating-point numbers

Features Types: Exercise!!!

	pickup	dropoff	passengers	distance	fare	tip	tolls	total	color	payment	pickup_zone	dropoff_zone	pickup_borough	dropoff_borough
0	2019-03-23 20:21:09	2019-03-23 20:27:24	1	1.60	7.0	2.15	0.0	12.95	yellow	credit card	Lenox Hill West	UN/Turtle Bay South	Manhattan	Manhattan
1	2019-03-04 16:11:55	2019-03-04 16:19:00	1	0.79	5.0	0.00	0.0	9.30	yellow	cash	Upper West Side South	Upper West Side South	Manhattan	Manhattan
2	2019-03-27 17:53:01	2019-03-27 18:00:25	1	1.37	7.5	2.36	0.0	14.16	yellow	credit card	Alphabet City	West Village	Manhattan	Manhattan
3	2019-03-10 01:23:59	2019-03-10 01:49:51	1	7.70	27.0	6.15	0.0	36.95	yellow	credit card	Hudson Sq	Yorkville West	Manhattan	Manhattan
4	2019-03-30 13:27:42	2019-03-30 13:37:14	3	2.16	9.0	1.10	0.0	13.40	yellow	credit card	Midtown East	Yorkville West	Manhattan	Manhattan
5	2019-03-11 10:37:23	2019-03-11 10:47:31	1	0.49	7.5	2.16	0.0	12.96	yellow	credit card	Times Sq/Theatre District	Midtown East	Manhattan	Manhattan
6	2019-03-26 21:07:31	2019-03-26 21:17:29	1	3.65	13.0	2.00	0.0	18.80	yellow	credit card	Battery Park City	Two Bridges/Seward Park	Manhattan	Manhattan
7	2019-03-22 12:47:13	2019-03-22 12:58:17	0	1.40	8.5	0.00	0.0	11.80	yellow	NaN	Murray Hill	Flatiron	Manhattan	Manhattan
8	2019-03-23 11:48:50	2019-03-23 12:06:14	1	3.63	15.0	1.00	0.0	19.30	yellow	credit card	East Harlem South	Midtown Center	Manhattan	Manhattan
9	2019-03-08 16:18:37	2019-03-08 16:26:57	1	1.52	8.0	1.00	0.0	13.30	yellow	credit card	Lincoln Square East	Central Park	Manhattan	Manhattan
10	2019-03-16 10:02:25	2019-03-16 10:22:29	1	3.90	17.0	0.00	0.0	17.80	yellow	cash	LaGuardia Airport	Astoria	Queens	Queens
11	2019-03-20 19:39:42	2019-03-20 19:45:36	1	1.53	6.5	2.16	0.0	12.96	yellow	credit card	Upper West Side South	Manhattan Valley	Manhattan	Manhattan
12	2019-03-18 21:27:14	2019-03-18 21:34:16	1	1.05	6.5	1.00	0.0	11.30	yellow	credit card	Murray Hill	Midtown Center	Manhattan	Manhattan
13	2019-03-19 07:55:25	2019-03-19 08:09:17	1	1.75	10.5	0.00	0.0	13.80	yellow	cash	Lincoln Square West	Times Sq/Theatre District	Manhattan	Manhattan
14	2019-03-27 12:13:34	20[illegible]27 12:25:48	0	2.90	11.5	0.00	0.0	14.80	yellow	cash	Financial District North	Two Bridges/Seward Park	Manhattan	Manhattan
15	2019-03-1[illegible]	[illegible]8:13:57	3	2.09	13.5	0.00	0.0	16.80	yellow	cash	Upper West Side North	Clinton East	Manhattan	Manhattan
16	2019-0[illegible]	[illegible]-03-15 12:54:28	1	2.12	13.0	0.00	0.0	16.30	yellow	cash	East Chelsea	Meatpacking/West Village West	Manhattan	Manhattan
[illegible]	2019[illegible]	2019-03-23 21:02:07	1	2.60	10.5	2.00	0.0	16.30	yellow	credit card	Midtown Center	East Harlem South	Manhattan	Manhattan
[illegible]8	[illegible]-27 0[illegible]	[illegible]19-03-27 06:38:10	1	2.18	9.5	1.92	0.0	14.72	yellow	credit card	Gramercy	Midtown Center	Manhattan	Manhattan
[illegible]	201[illegible]22:04:25	20[illegible]03-25 22:11:30	6	1.08	6.5	1.08	0.0	11.38	yellow	credit card	East Chelsea	East Chelsea	Manhattan	Manhattan

Data Pre-Processing and Cleaning

Typically, raw data is not ready for being used by the data analysis algorithms but it must undergo several transformations

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Data Pre-Processing and Cleaning

Typically, raw data is not ready for being used by the data analysis algorithms but it must undergo several transformations

- Categorical to numerical encoding (mandatory)

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Data Pre-Processing and Cleaning

Typically, raw data is not ready for being used by the data analysis algorithms but it must undergo several transformations

- Categorical to numerical encoding (mandatory)
- Data Cleaning: Missing values, duplicates, outliers

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Data Pre-Processing and Cleaning

Typically, raw data is not ready for being used by the data analysis algorithms but it must undergo several transformations

- Categorical to numerical encoding (mandatory)
- Data Cleaning: Missing values, duplicates, outliers
- Numerical to numerical transformations: Normalization

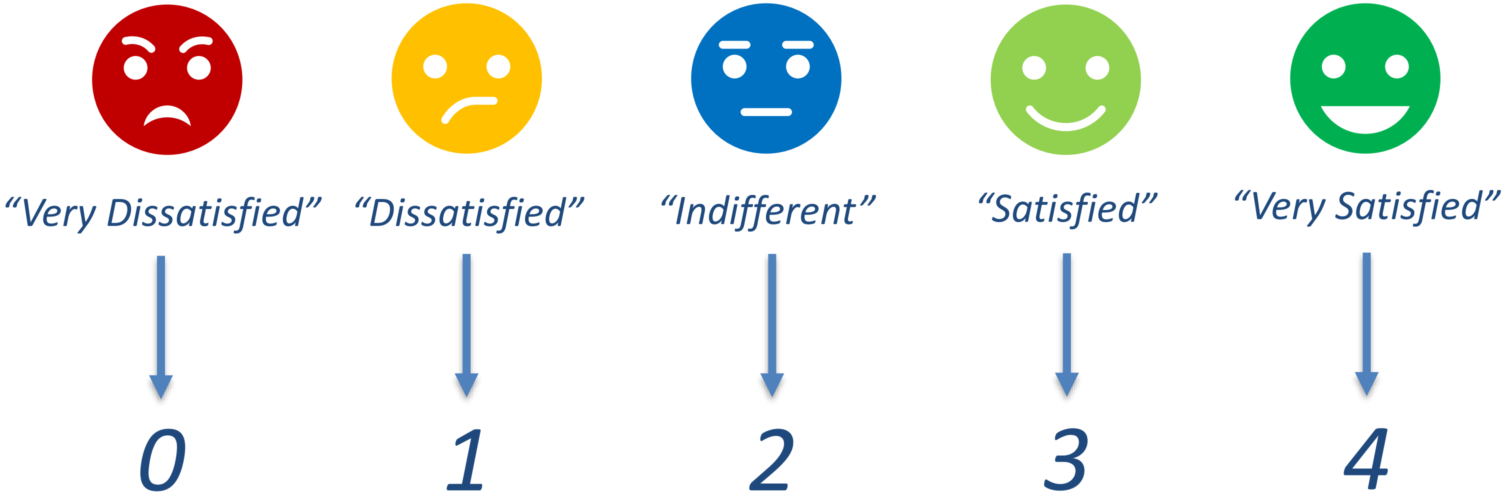
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Categorical to Numerical Encoding (1/2)

Ordinal Features

- Encoding should preserve the information regarding the order

<u>Integer/Label Encoding</u>: Assign an integer to each value, preserving the ordering (e.g., customer satisfaction)



Categorical to Numerical Encoding (2/2)

Nominal Features

- Assigning integer values is no longer suitable for nominal categorical variables
- It induces an order that does not exists

One-Hot Encoding: e.g., “island” feature the penguins dataset: S = {“Biscoe”, “Dream”, “Torgersen”}

	0	*1*	*2*
“Biscoe” →	**1**	**0**	**0**
“Dream” →	**0**	**1**	**0**
“Torgersen” →	**0**	**0**	**1**

Increased number of features, from 1 to |S|=3

Data Cleaning: Missing Values (1/2)

Reasons for missing values

- Faulty equipment, incorrect measurements, missing cells in manual data entry, censored/anonymous data
- Very frequent in questionnaires for medical scenarios
- Censored/anonymous data

Frequently represented by

- Out-of-range values
- NaN
- Special values (e.g., -1)
- …

Data Cleaning: Missing Values (2/2)

Types of missing values

- Missing completely at random (MCAR): when the distribution of missing values does not depend on either the observed data or the unobserved data (e.g., random sampling from a population)
- Missing at random (MAR): when the distribution of missing values depends on the observed data, but not on the unobserved one (e.g., sampling from a population with a probability which depends on some known property)
- Missing not at random (MNAR): when the distribution of missing values depends on the unobserved data.

MAR / MNAR are difficult to identify, domain knowledge often required

Data Cleaning: Dealing with Missing Values (1/2)

Discarding examples with missing values (a.k.a., “list-wise deletion”)

- Easy to implement, but works when few examples have missing values, otherwise data are heavily reduced
- Excluded examples could be informative
- Deployed model cannot deal with missing values

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Dream	43.2	18.5	192.0	4100.0	Male
1	Adelie	Dream	36.0	17.1	187.0	3700.0	Female
2	Gentoo	Biscoe	44.5	14.7	214.0	4850.0	Female
3	Adelie	Biscoe	41.3	21.1	195.0	4400.0	Male
4	Chinstrap	Dream	51.4	19.0	201.0	3950.0	Male
~~5~~	~~Gentoo~~	~~Biscoe~~	~~47.3~~	~~13.8~~	~~216.0~~	~~4725.0~~	~~NaN~~
6	Adelie	Biscoe	41.1	19.1	188.0	4100.0	Male

Data Cleaning: Dealing with Missing Values (2/2)

Imputation Methods assign new values based on the dataset

Continuous Features

- Mean value imputation
- Mode value imputation (most frequent value)
- Replace with a constant value (e.g., mean) and add a new categorical feature as missing values indicator (1 = value is missing, 0 = value is not missing)
- Regression model

Categorical Features

- Mode value imputation (most frequent value)
- Insert additional “Unknown” category

Data Cleaning: Inaccurate Values

Data has not been collected for machine learning

- Errors and omissions that don't affect original purpose of data (e.g., age of customer)
- Typographical errors in nominal attributes, thus values need to be checked for consistency
- Typographical and measurement errors in numeric attributes, thus outliers need to be identified

We should know our data..
Statistics and visualization are powerful tools!

Errors may be deliberate (e.g., wrong zip codes or phone numbers)

Data Exploration

Preliminary exploration of data during which

- Statistics are to summarize properties of the data
- Visualization tools are used to convert data into a visual format

In order to

- Tune the pre-processing tools
- Identify general patterns and trends which may help in selecting the right learning algorithm
- Detect outliers

Data Exploration: Summary Statistics (1/2)

Frequency and Mode

- Frequency of a feature value: Percentage of time the value occurs in the dataset. E.g., frequency of ‘Dream’ value in the penguins dataset is
- The mode is the most frequent feature value
- Frequency and mode are typically used with categorical features

Mean and Median

- $\mathrm{mean}(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$
- $median(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$
- Both measure of the location of the data; median is more robust to outlier

Data Exploration: Summary Statistics (2/2)

Percentiles

- Given an ordinal or continuous feature $\boldsymbol{x}$ and a number p, the p-th percentile is a value $\boldsymbol{x}$_p of $\boldsymbol{x}$_such that p% of the observed values of $\boldsymbol{x}$ are less than $\boldsymbol{x}$_p
- E.g, 25-th percentile is the value $\boldsymbol{x}$_25 such that 25% of all values of $\boldsymbol{x}$ are less than $\boldsymbol{x}$_25

Range and Variance (both are measures of spread)

- $\mathbf{range}(x) = \mathbf{max}(x) - \mathbf{min}(x)$
- $\mathbf{variance}(x) = \sigma_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$

Data Exploration: Data Visualization (1/4)

Exploit humas ability of capturing patterns from visual information …
there are several visual tools that can be exploited

- Bar plots
- Histograms
- Scatter plots
- Heatmaps
- …

(attend the practical lectures and you'll see)

Data Exploration: Data Visualization (2/4)

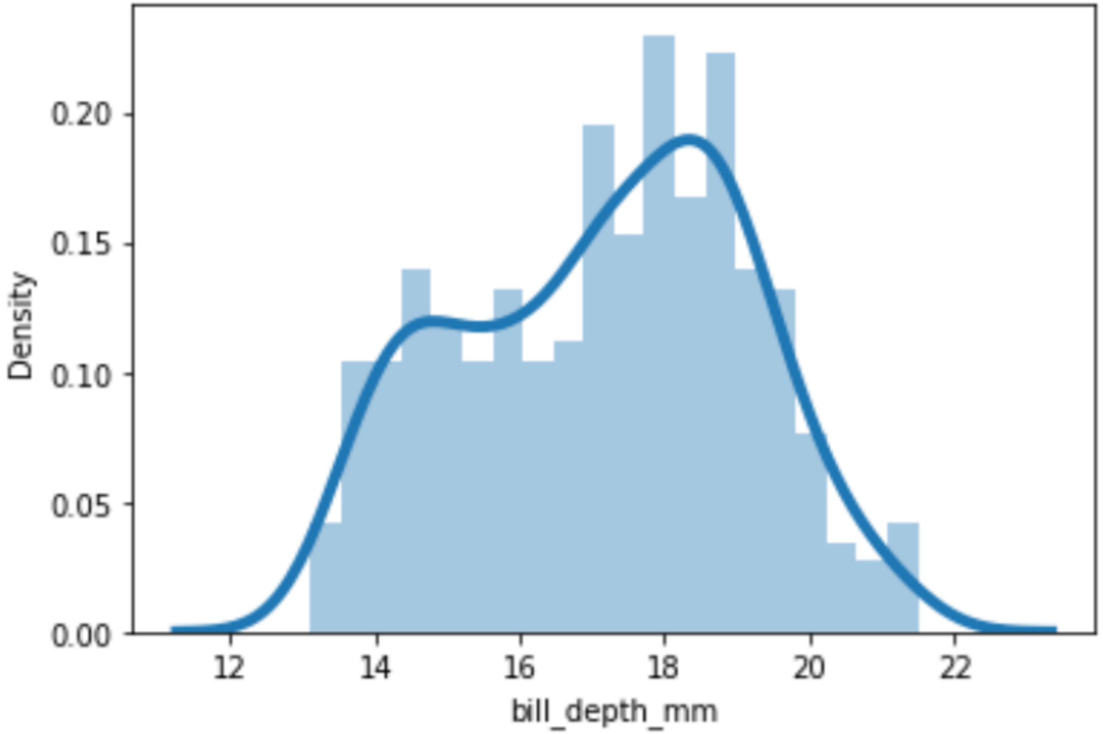
Bar plots

- They use horizontal or vertical bars to compare categories.
- One axis shows the compared categories, the other axis represents a discrete value

Data Exploration: Data Visualization (3/4)

Histograms

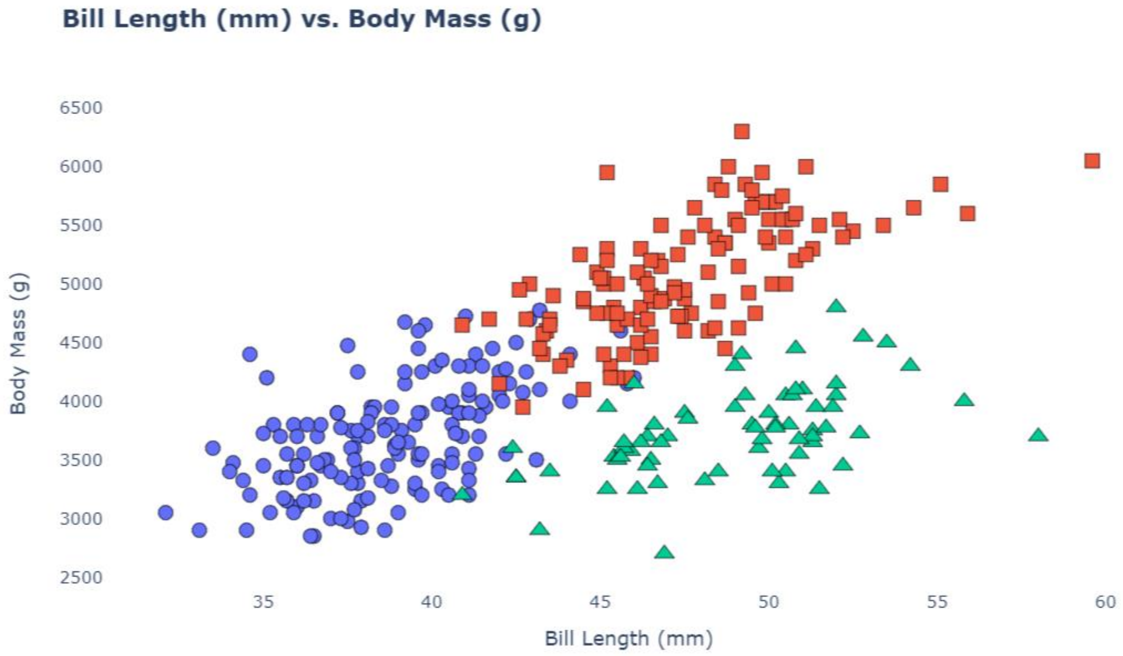
- They estimate the probability distribution of a continuous variables
- They are representations of tabulated frequencies depicted as adjacent rectangles, erected over discrete intervals (bins).
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins



Data Exploration: Data Visualization (4/4)

Scatter plots

- Used to compare two (or more) attributes
- Attributes values used to determine the position of the point
- Two-dimensional scatter plots most common, but 3D plots also used
- Often additional attributes can be displayed by using size, shape, and color of the markers
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes



Data Cleaning: Outliers (1/4)

Outliers are data objects that do not comply with the general behavior or model of the data, that is, values that appear as anomalous

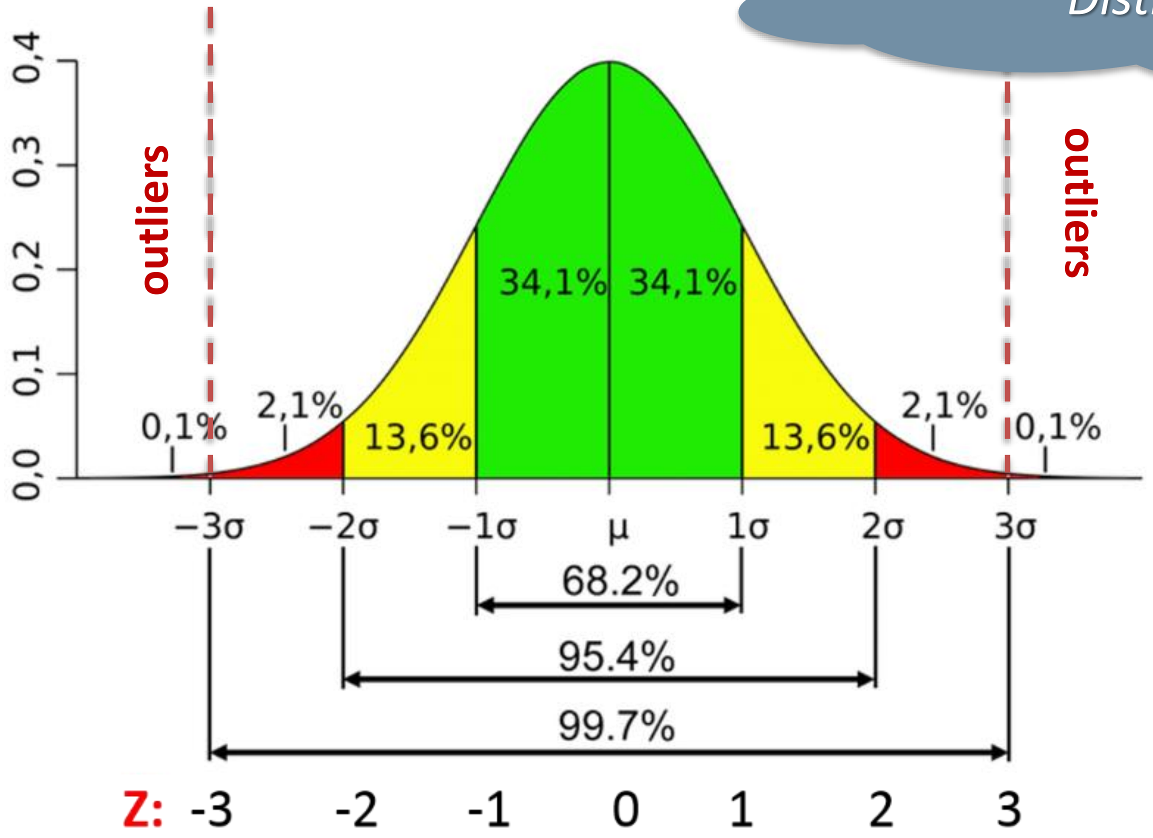
Outliers may be detected using

- Manual inspection and knowledge of reasonable values
- Statistical tests that assume a distribution or probability model for the data
- Distance measures where objects that are a substantial distance from any other cluster are considered outliers

Data Cleaning: Outliers (2/4)

Z-score : number of standard deviations of an observation w.r.t. to mean

Works with Normal Distributions …



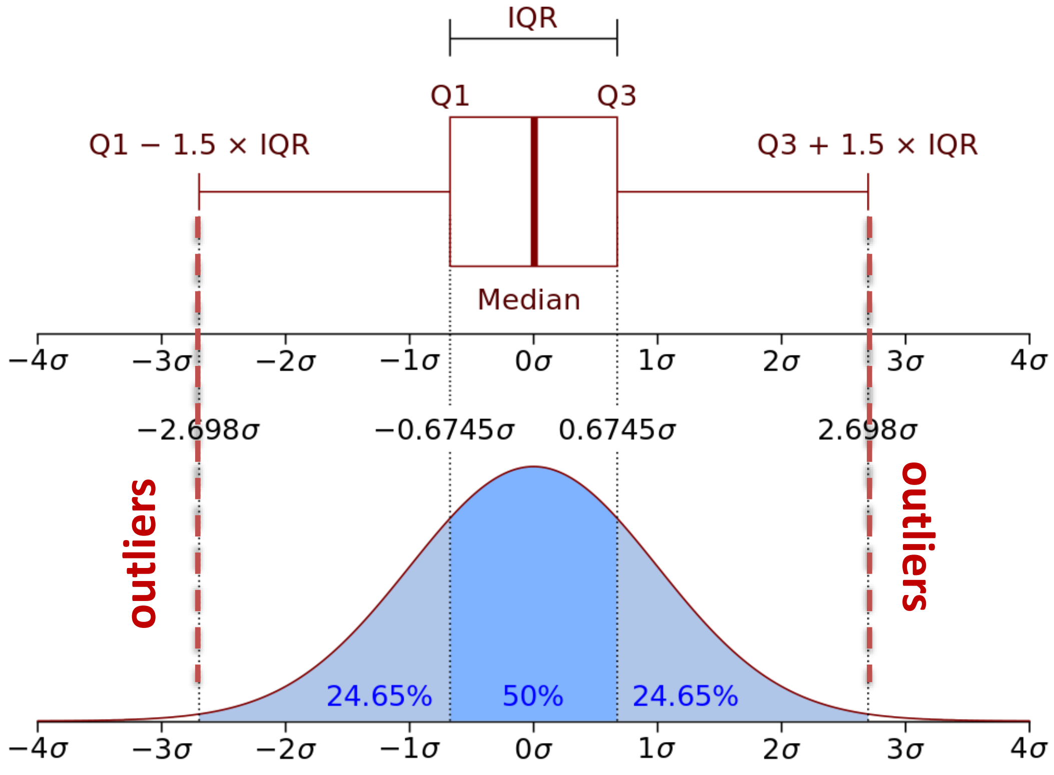
Adapted from source

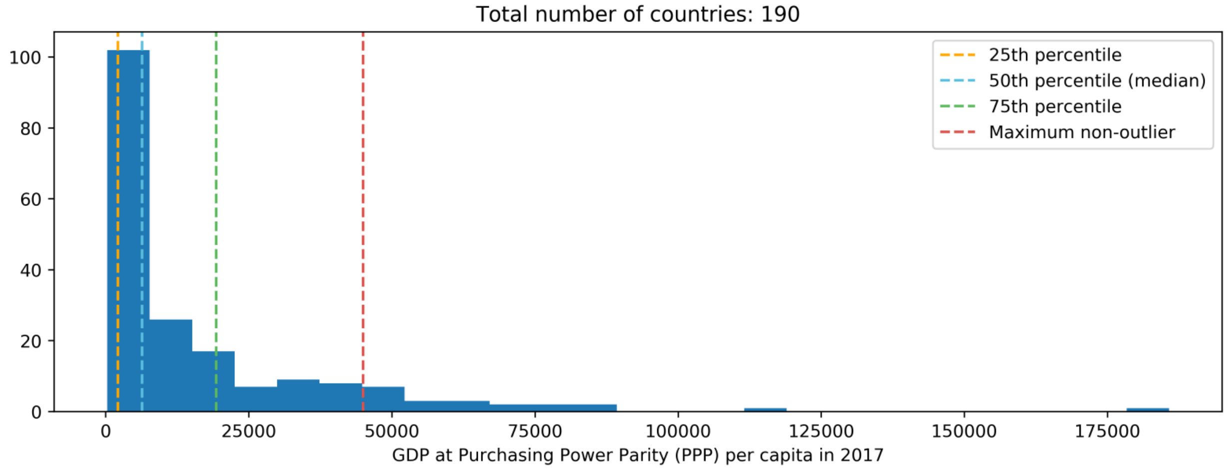
Data Cleaning: Outliers (3/4)

Percentiles work with any distribution …

Interquartile Range (IQR): spread difference between the first and third quartiles of data, i.e., the 25-th and 75-th percentiles

Works with Skewed Distributions …





Adapted from source1 and source2

Data Cleaning: Outliers (4/4)

Outliers are typically filtered out by eliminating containing data points

Trimming

- Eliminate the outlier data values

Imputation

- Typically, by using boundary values (clapping); e.g, observations > 99-th percentile = 99-th percentile
- Symmetric clapping: Winsorizing; e.g., a 10% Winsorizing, consider the 5th and 95th percentiles and set the values below the 5th percentile to the 5th percentile itself and values above the 95th percentile to the 95th percentile itself

Data Pre-Processing: Normalization

Features having different scales can cause unwanted effect during fit

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Data Pre-Processing: Normalization

Features having different scales can cause unwanted effect during fit

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Min-Max Normalization: scales the values in the [0, 1] range

$$x_i' = \frac{x_i - \min_i x_i}{\max_i x_i - \min_i x_i}$$

Data Pre-Processing: Normalization

Features having different scales can cause unwanted effect during fit

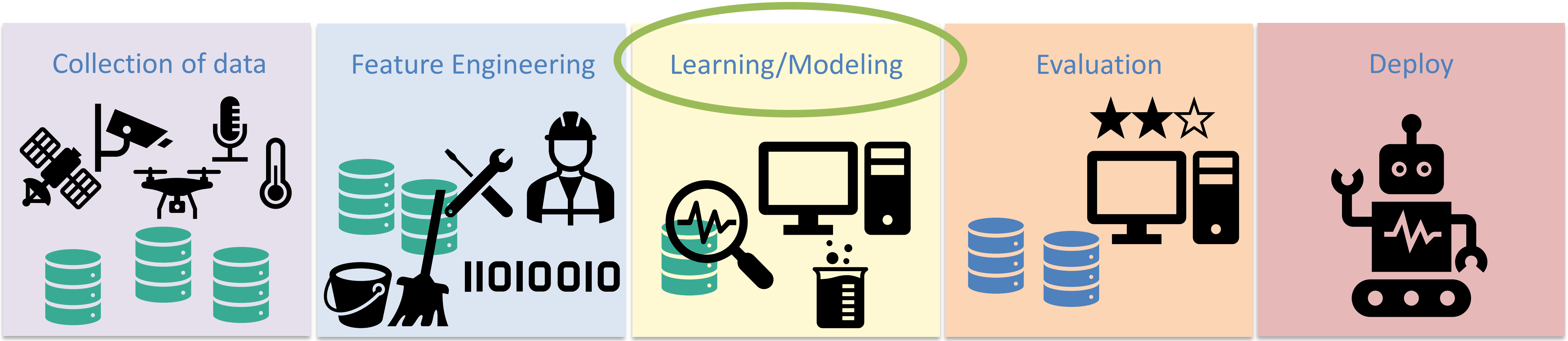
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

Standard Score Normalization (a.k.a. standardization): forces features to have mean of 0 and standard deviation of 1

$$x_i' = \frac{x_i - \mu}{\sigma}$$

If data was normally distributed, most of it (68%) will lie in the range [-1, 1]

Data Analysis: the process



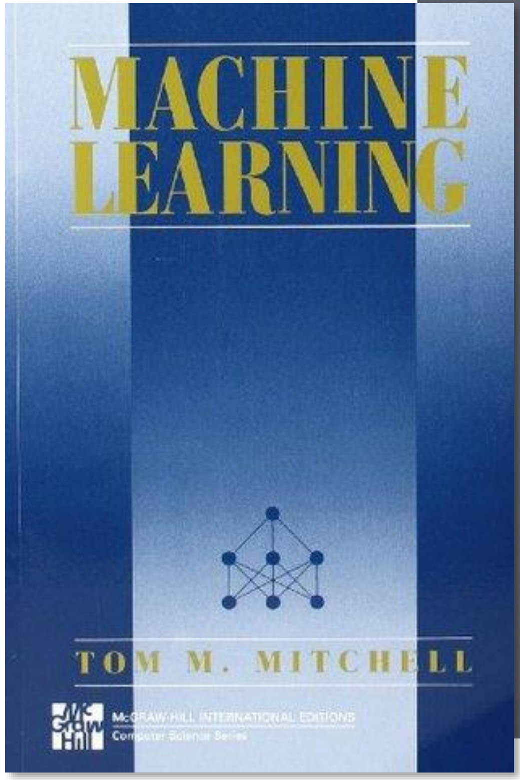
Learning/Modeling

- Select the learning task: *Classification*? *Regression*? *Clustering*? etc.
- Select the algorithm and perform learning/modeling

Machine Learning

A formal definition

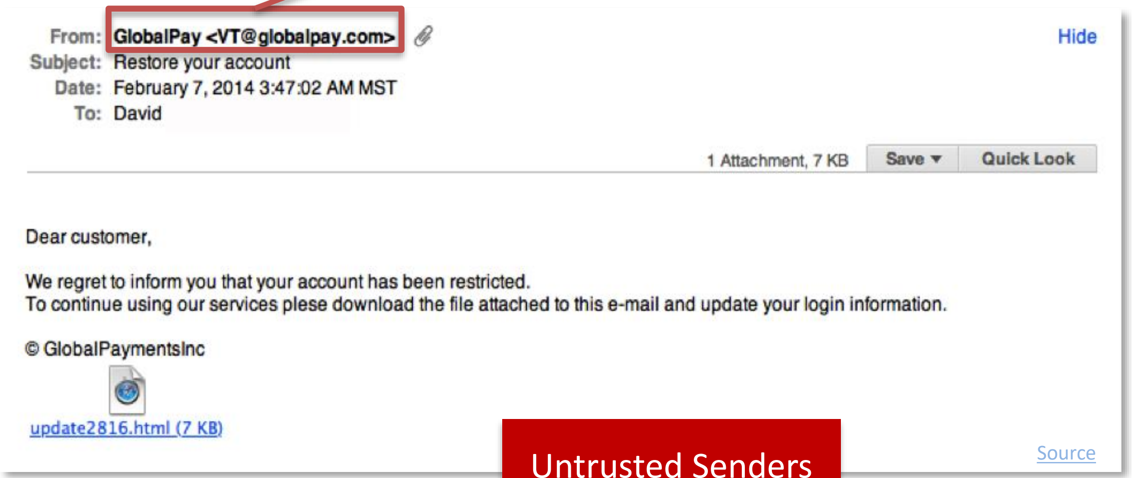
"A computer program is said to learn from experience E with respect to some class of task T and a performance measure P, if its performance at tasks in T, as measured by P, improves because of experience E."



Spam filtering example

Untrusted Senders

February 7, 2014


From: GlobalPay <VT@globalpay.com>
Subject: Restore your account
Date: February 7, 2014 3:47:02 AM MST
To: David

Hide

1 Attachment, 7 KB Save ▾ Quick Look

Dear customer,

We regret to inform you that your account has been restricted.
To continue using our services plese download the file attached to this e-mail and update your login information.

© GlobalPaymentsInc

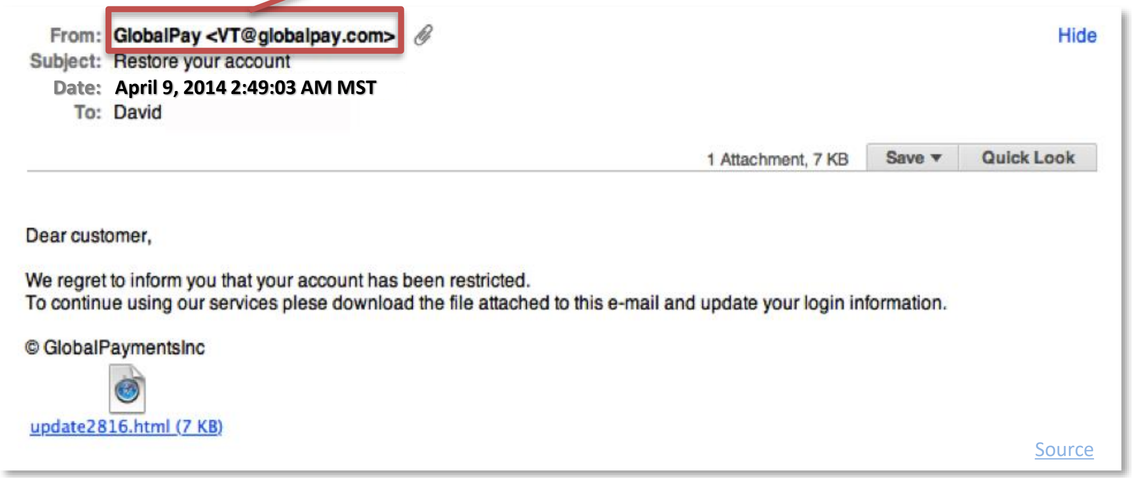
update2816.html (7 KB)

Source

Untrusted Senders

?

April 9, 2014


From: GlobalPay <VT@globalpay.com>
Subject: Restore your account
Date: April 9, 2014 2:49:03 AM MST
To: David

Hide

1 Attachment, 7 KB Save ▾ Quick Look

Dear customer,

We regret to inform you that your account has been restricted.
To continue using our services plese download the file attached to this e-mail and update your login information.

© GlobalPaymentsInc

update2816.html (7 KB)

Source

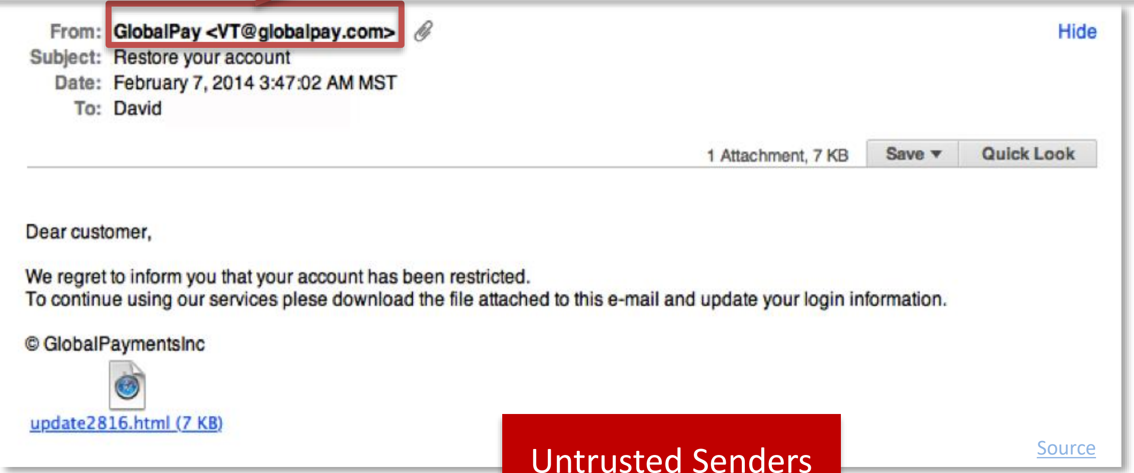
Spam filtering example

Task T

Untrusted Senders

Experience E

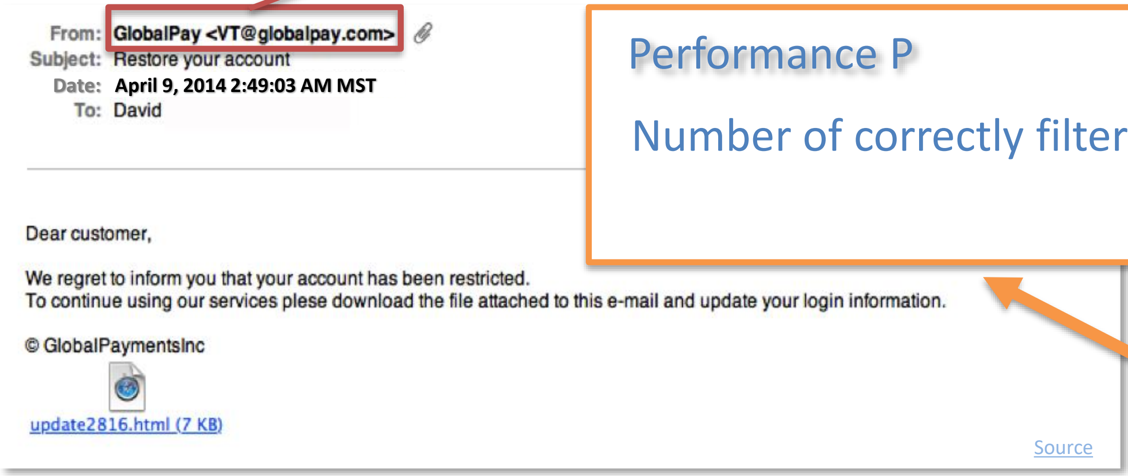
February 7, 2014


From: GlobalPay <VT@globalpay.com>
Subject: Restore your account
Date: February 7, 2014 3:47:02 AM MST
To: David
Hide
1 Attachment, 7 KB Save ▾ Quick Look
Dear customer,
We regret to inform you that your account has been restricted.
To continue using our services plese download the file attached to this e-mail and update your login information.
© GlobalPaymentsInc
update2816.html (7 KB)
Source

Untrusted Senders

?

April 9, 2014


From: GlobalPay <VT@globalpay.com>
Subject: Restore your account
Date: April 9, 2014 2:49:03 AM MST
To: David
Dear customer,
We regret to inform you that your account has been restricted.
To continue using our services plese download the file attached to this e-mail and update your login information.
© GlobalPaymentsInc
update2816.html (7 KB)
Source

Performance P

Number of correctly filtered e-mails

Are we really learning?

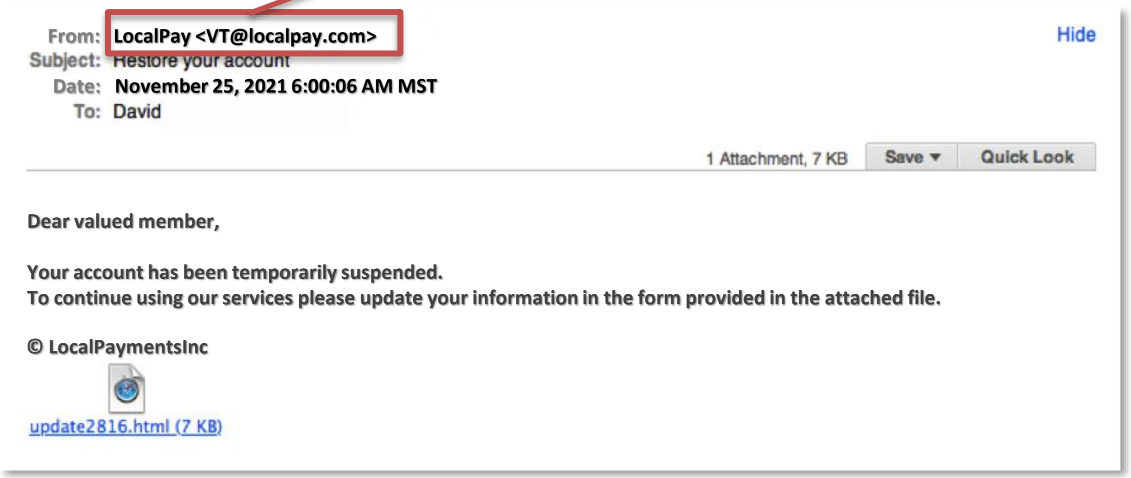
Spam filtering example

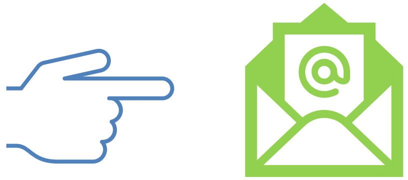
?

Untrusted Senders

NO

November 25, 2021





Spam filtering example

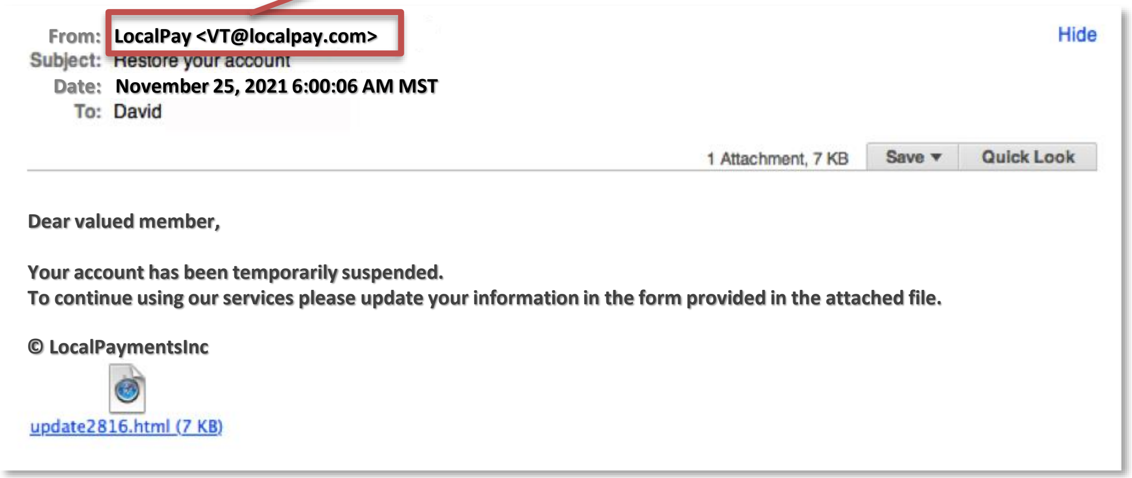
Learning is not «memorizing»...

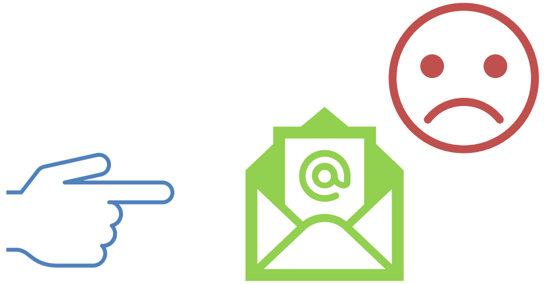
November 25, 2021

?

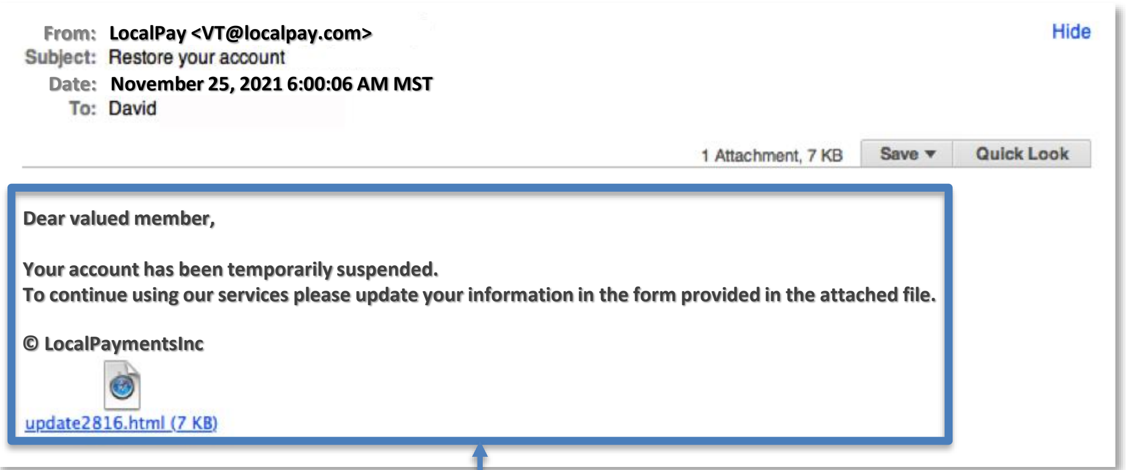
Untrusted Senders

NO





Spam filtering example



From: LocalPay <VT@localpay.com>
Subject: Restore your account
Date: November 25, 2021 6:00:06 AM MST
To: David

Hide

1 Attachment, 7 KB Save ▾ Quick Look

Dear valued member,

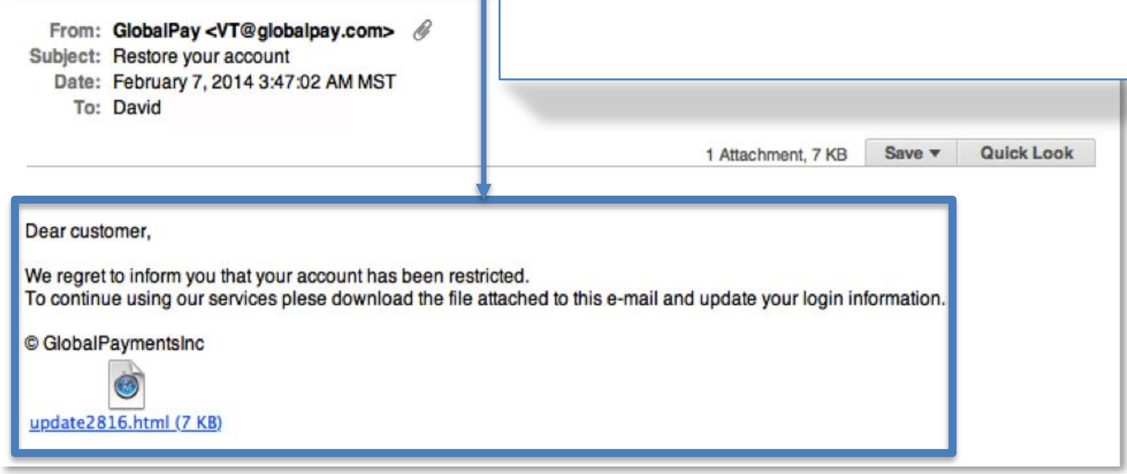
Your account has been temporarily suspended.
To continue using our services please update your information in the form provided in the attached file.

© LocalPaymentsInc

update2816.html (7 KB)

What if we look to the content?

Learning is about «finding patterns in data»…



From: GlobalPay <VT@globalpay.com>
Subject: Restore your account
Date: February 7, 2014 3:47:02 AM MST
To: David

1 Attachment, 7 KB Save ▾ Quick Look

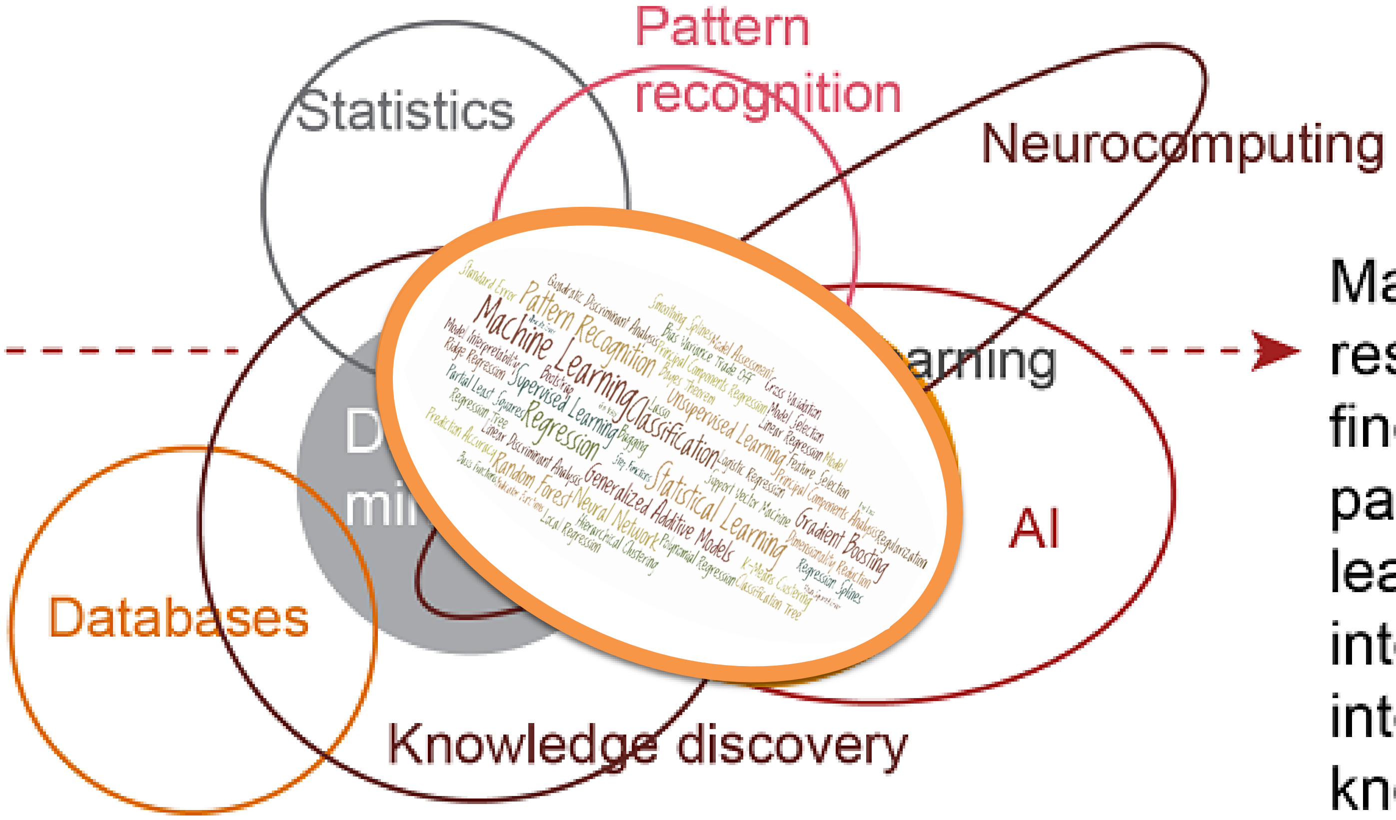
Dear customer,

We regret to inform you that your account has been restricted.
To continue using our services plese download the file attached to this e-mail and update your login information.

© GlobalPaymentsInc

update2816.html (7 KB)

Machine Learning



Machine learning is a category of research and algorithms focused on finding patterns in data and using those patterns to make predictions. Machine learning falls within the artificial intelligence (AI) umbrella, which in turn intersects with the broader field of knowledge discovery and data mining.

Source: SAS, 2014 and PwC, 2016 *and Matteucci, 2017*

Machine Learning

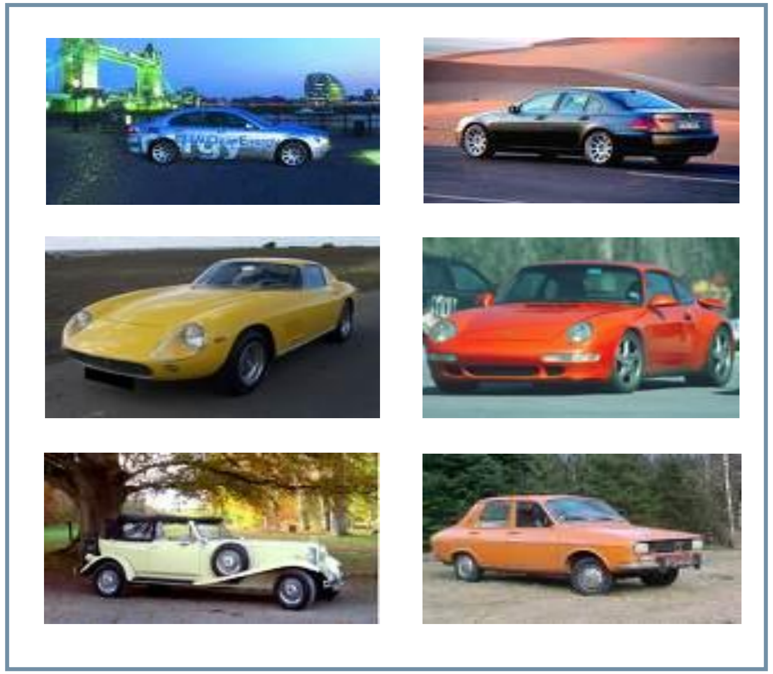
Machine Learning Paradigms

Imagine you have a certain experience D, i.e., data, and let's name it

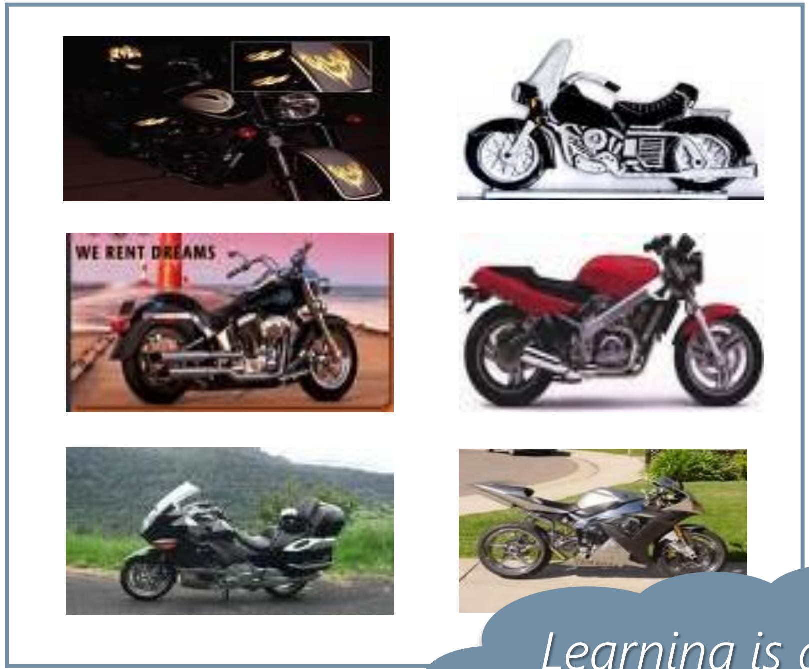
$$D = x_1, x_2, x_3, \dots, x_N$$

- ***Supervised learning***: given the desired outputs $t_1, t_2, t_3, \dots, t_N$ learn to produce the correct output given a new set of input
- ***Unsupervised learning***: exploit regularities in D to build a representation to be used for reasoning or prediction
- ***Reinforcement learning***: producing actions $a_1, a_2, a_3, \dots, a_N$ which affect the environment, and receiving rewards $r_1, r_2, r_3, \dots, r_N$ learn to act in order to maximize rewards in the long term

Supervised learning: Classification



Cars



Motorcy

Learning is about modeling …

Hand-crafted Features

Learned Classifier

Motorcycle

Supervised learning: Regression


12000 $


15000 $

6000 $


2000 $


8000 $


22000 $


4000 $


28000 $


6000 $


35000 $


→ Hand-crafted Features → Learned Regressor → **3800 $**

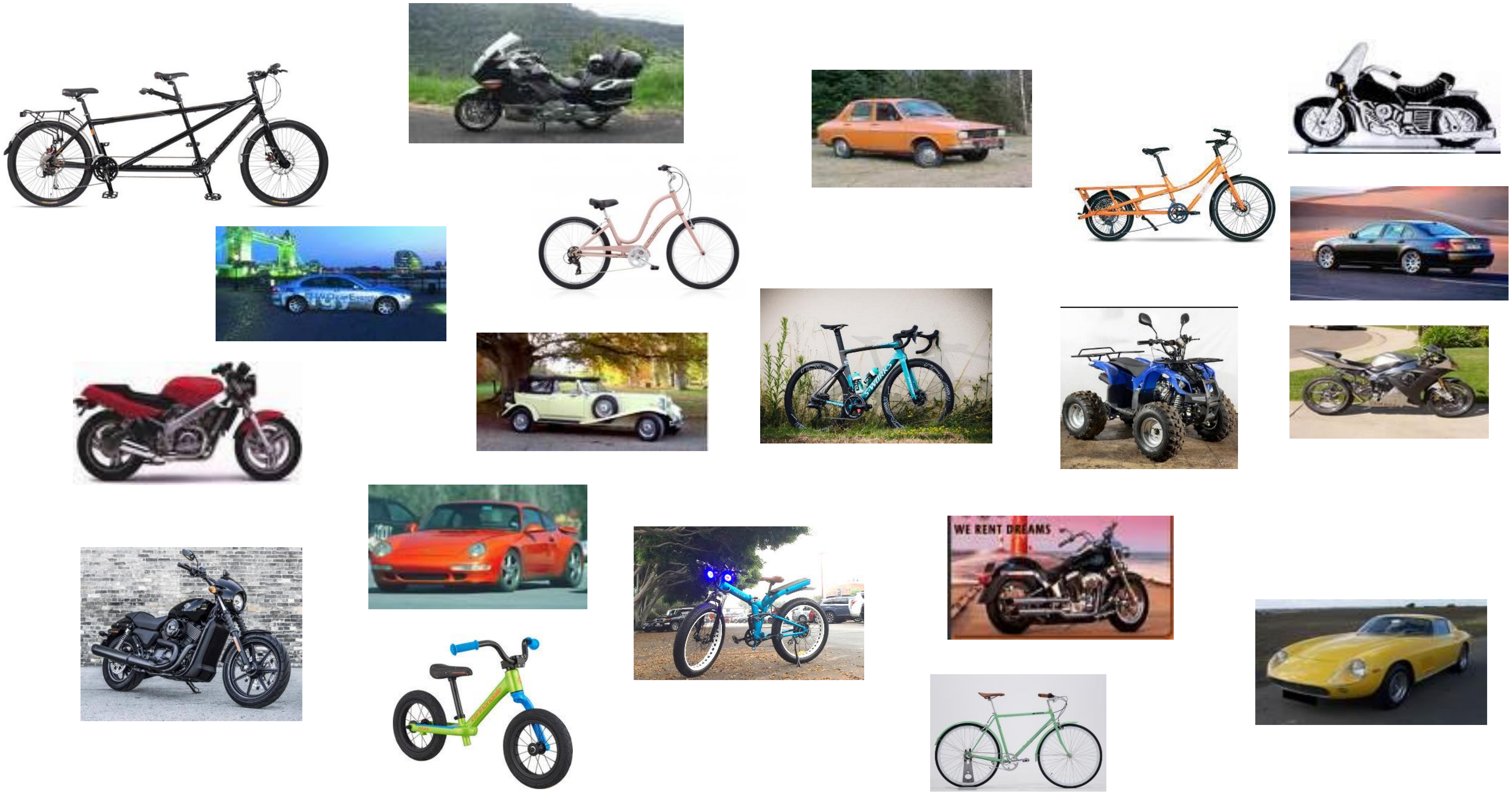
Machine Learning Paradigms

Imagine you have a certain experience D, i.e., data, and let's name it

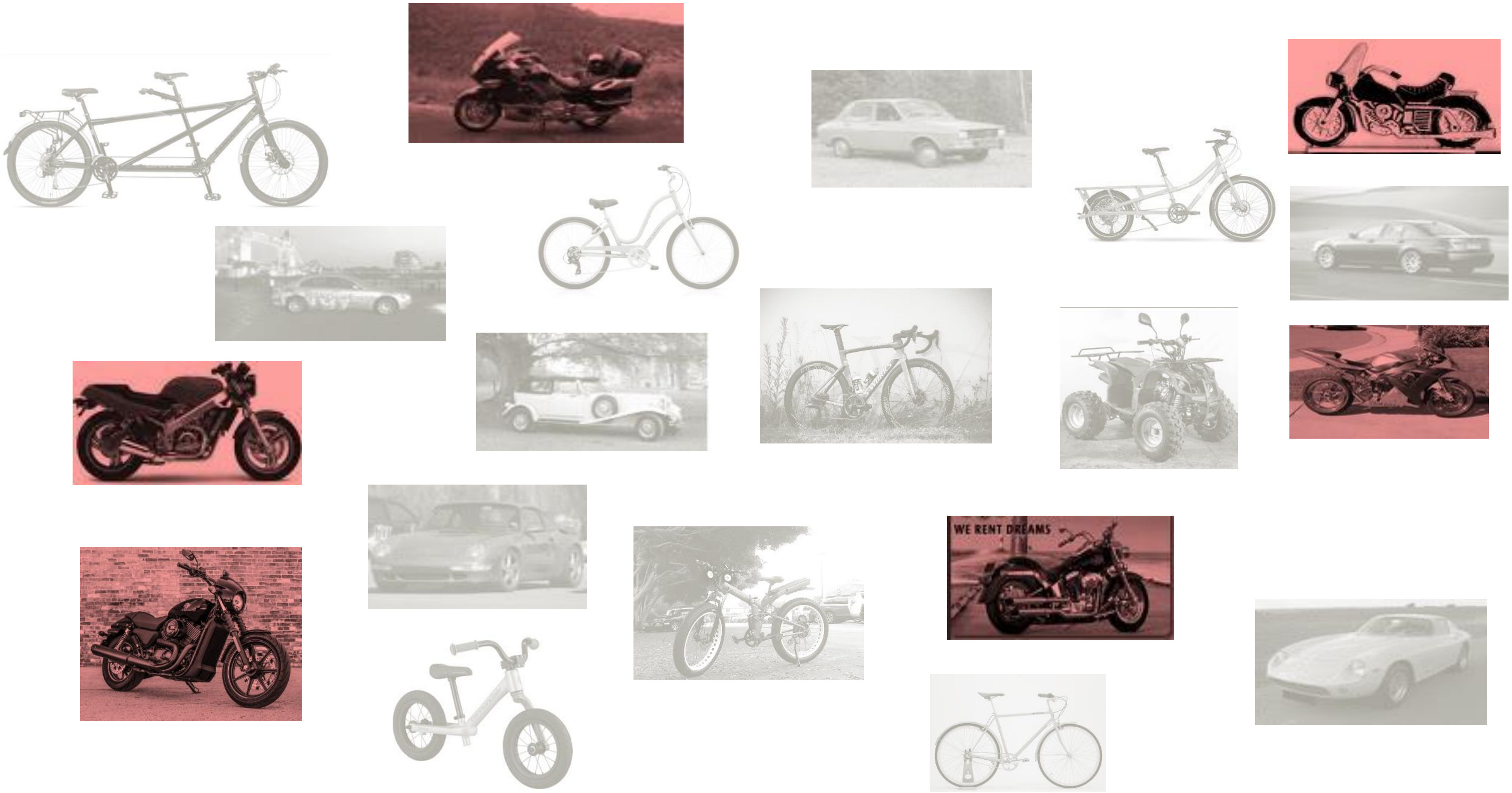
$$D = x_1, x_2, x_3, \ldots, x_N$$

- ***Supervised learning***: given the desired outputs $t_1, t_2, t_3, \ldots, t_N$ learn to produce the correct output given a new set of input
- ***Unsupervised learning***: exploit regularities in D to build a representation to be used for reasoning or prediction
- ***Reinforcement learning***: producing actions $a_1, a_2, a_3, \ldots, a_N$ which affect the environment, and receiving rewards $r_1, r_2, r_3, \ldots, r_N$ learn to act in order to maximize rewards in the long term

Unsupervised learning: Clustering

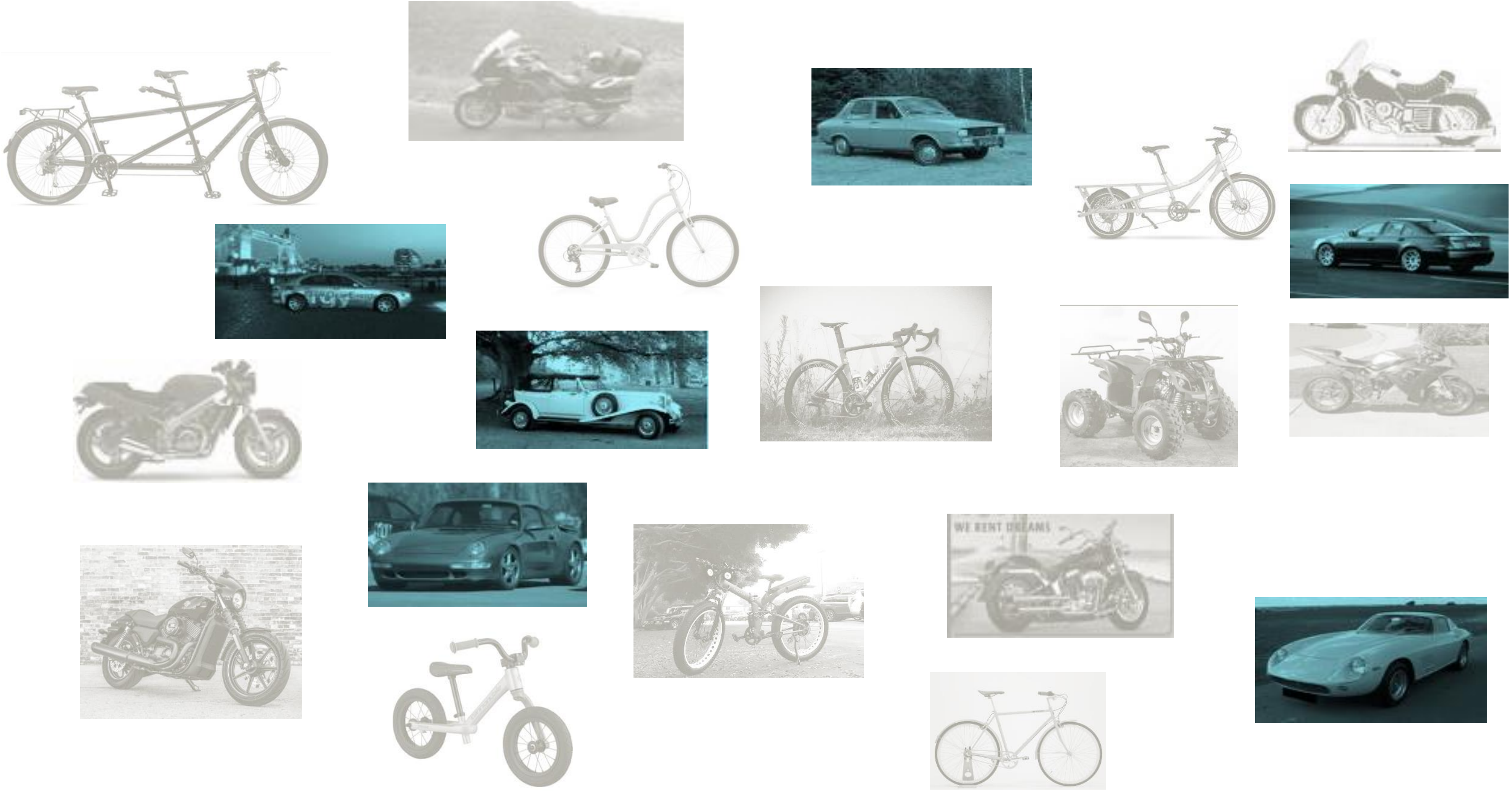


Unsupervised learning: Clustering

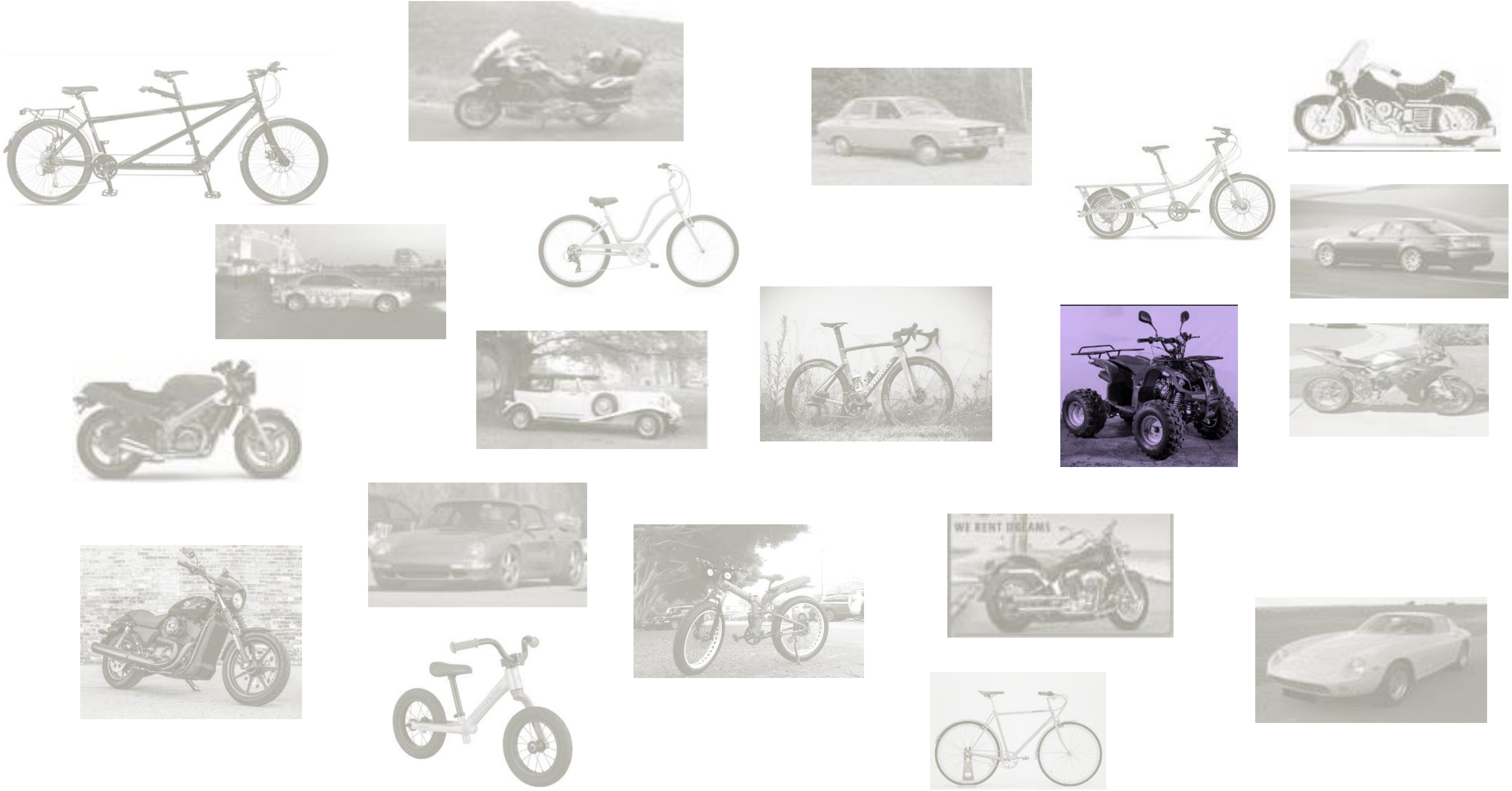


Unsupervised learning: Clustering

Unsupervised learning: Clustering



Unsupervised learning: Clustering



Unsupervised learning: Clustering











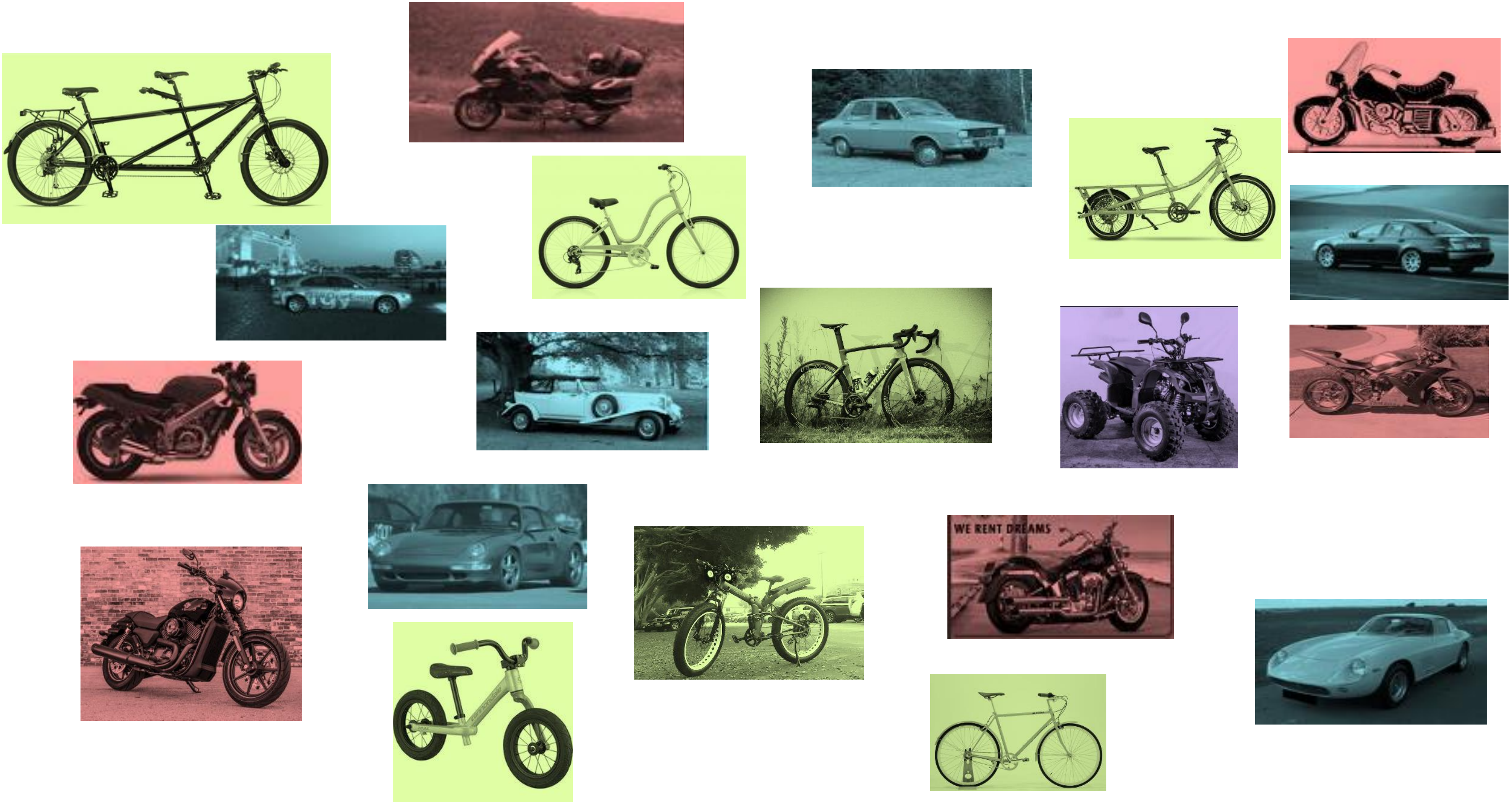








Unsupervised learning: Clustering



Clustering vs. Classification

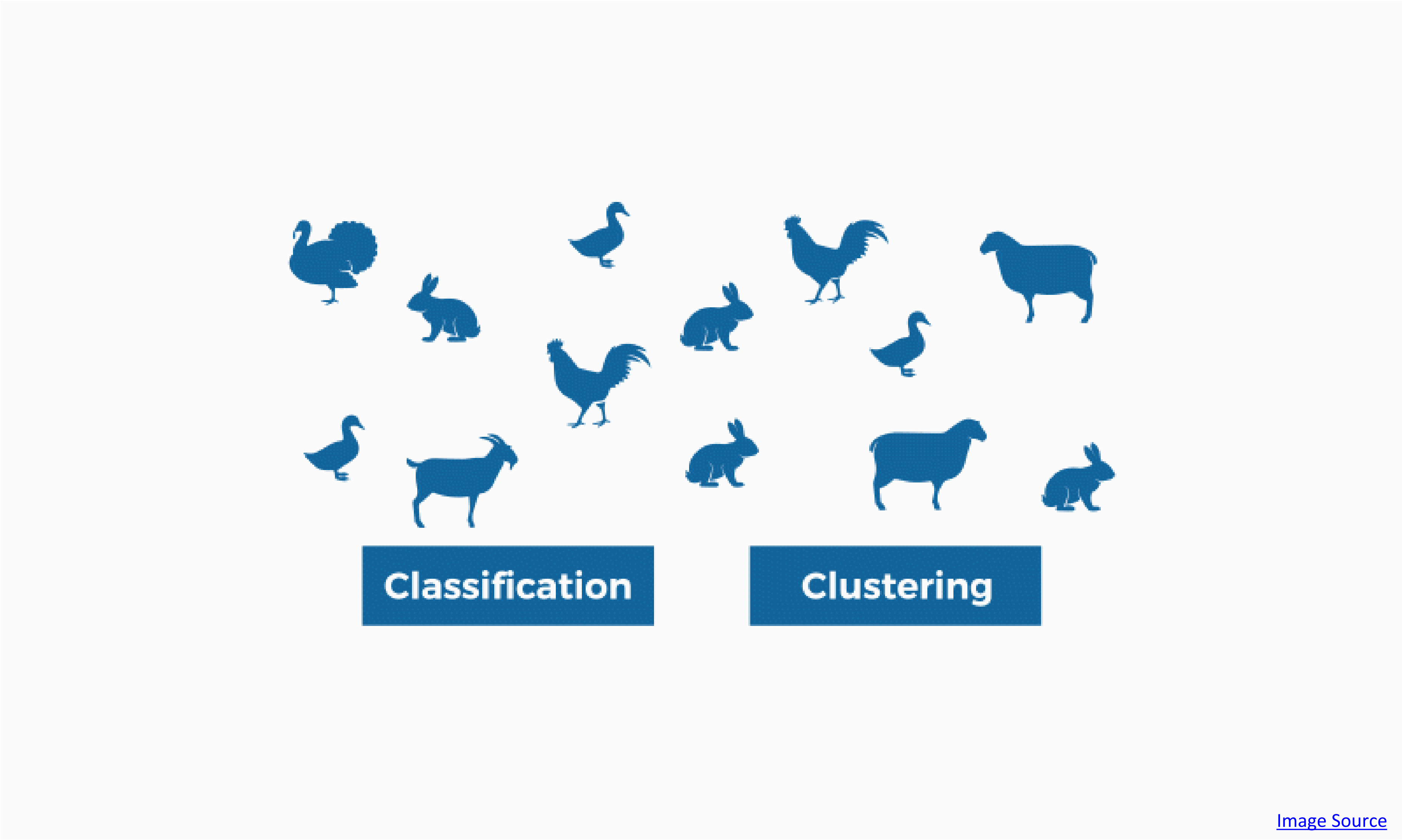


Image Source

Supervised/Unsupervised Learning: Reasoning Time ☺

Let's try to identify together some examples of learning tasks in agriculture

<u>Classification</u>	<u>Regression</u>	<u>Clustering</u>