R for Data Science

Introduction to Data Analytics

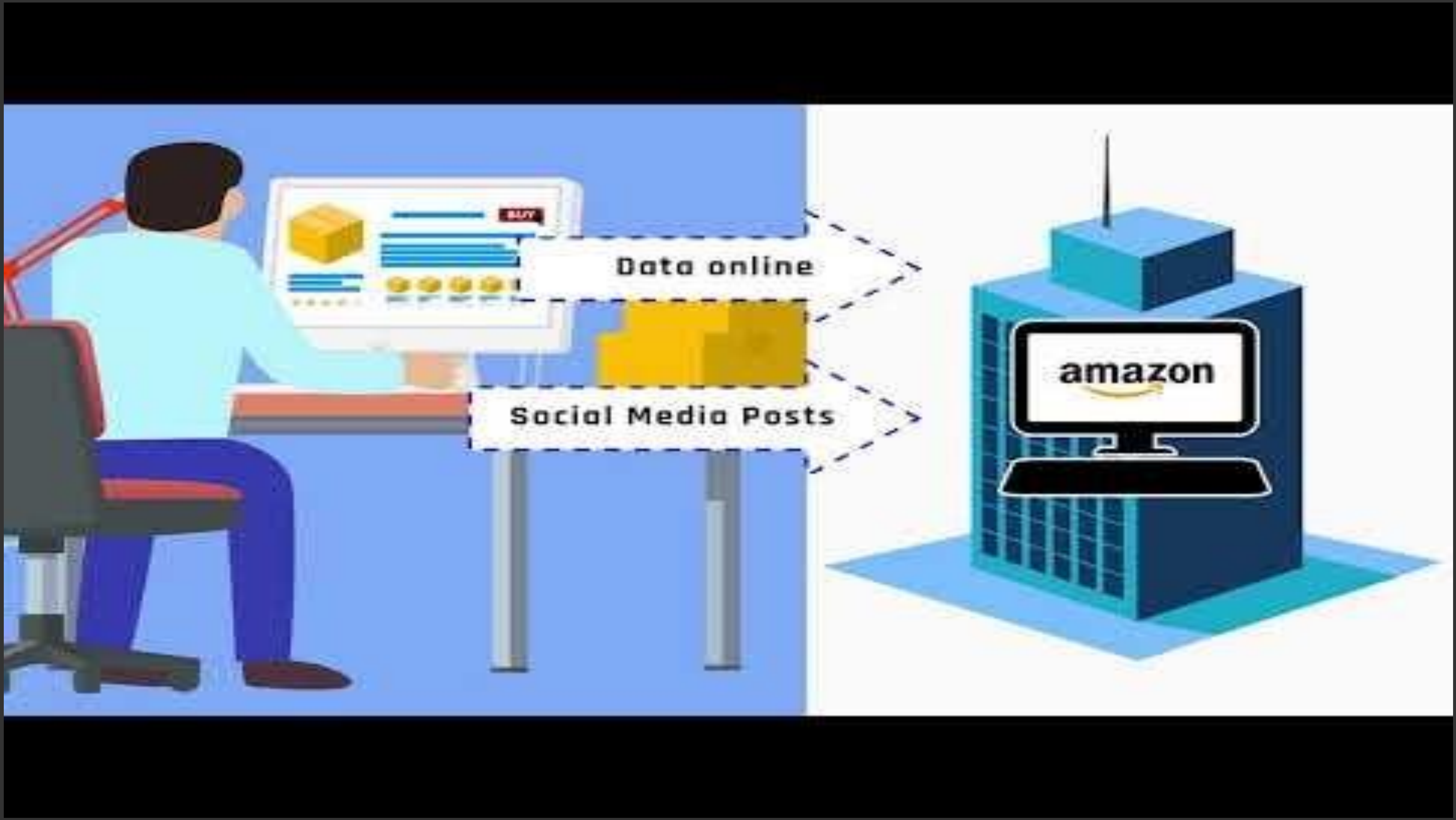
Agenda

- Motivation
- Introduction to machine learning
- Linear regression model
  - Example case – TransportEY
    - Simplest regression model
    - Interpreting regression output
  - Regression fundamentals*
  - Assumptions*
- Case study demonstration
- Regression assignment
- Concluding remarks

*Optional

Motivation

Economics of technology[1]



[1] https://www.youtube.com/watch?v=5G0PbwtiMJk

Main strengths of computers and humans

- Computers excel at
  - Remembering facts precisely without error or difficulty recalling
  - Performing repeated operations/actions without getting distracted
  - Numeric calculations

- Humans excel at
  - Critical thinking
  - Strategical thinking
  - Creativity
  - Communication
  - Using biases and context

What is analytics?

Analytics is about discovering patterns in the data. The goal is to summarize large amounts of data so it can be better understood (and applied).

Goals of Analytics	Business Objectives
Data Understanding **Visualization** **Summarizing**	• Identify key trends and understand historical business performance • Compare performance across the business, through time and against external metrics • Identify opportunities through of business under/over performance
Model Design **Variable Selection** **Dimension Reduction**	• Identify performance drivers • Decompose the impact multiple changing factors • Model how operational performance against peers • Model sensitivity across variables (e.g. Price sensitive across Brand, Channel, Customer type).
Prediction **Forecasting** **Classification**	• Classify and group customers for targeted marketing efforts • Predict how customers will behave to new or updated offerings • Forecast financials for business planning

The goal of analytics is to make better decisions by leveraging data

What is *Advanced* Analytics?

Advanced Analytics has the same objectives of all analytics – to make better decisions by understanding the data/business, designing more accurate models and forecasts

Advanced Analytics is statistics for large datasets

- Large in both the number of rows (size n) and in the number of variables (dimension p)

With large datasets you cannot:

- Plot every variable to look for interactions or transformations
- Test each individual variable for significance (T-test)
- Choose the best model amongst a set of candidate models (F-test)

A different approach is needed when n and p get really big. Lots of techniques exists for various situations (Text Mining, Neural Networks, Deep Learning, Etc.)

You need experience to make the black box work

- Always be aware of **overfitting & false discovery**
- Important to understand how your **assumptions** impact the output

Why data analytics is no longer something for in the future

- Increased availability of hardware and software
  - Unlimited supply of cheap computation power
  - Hardware infrastructure mature enough
  - Power and user-friendliness of software
- Improved availability and quality of data
  - Advanced systems have collected useful and clean data
  - Cost of data storage has decreased substantially
- Sufficient knowledge
  - ‘Scale-up phase’
  - Online resources and documentation are accessible and of high quality
  - Growing portion of the workforce possess the required skills

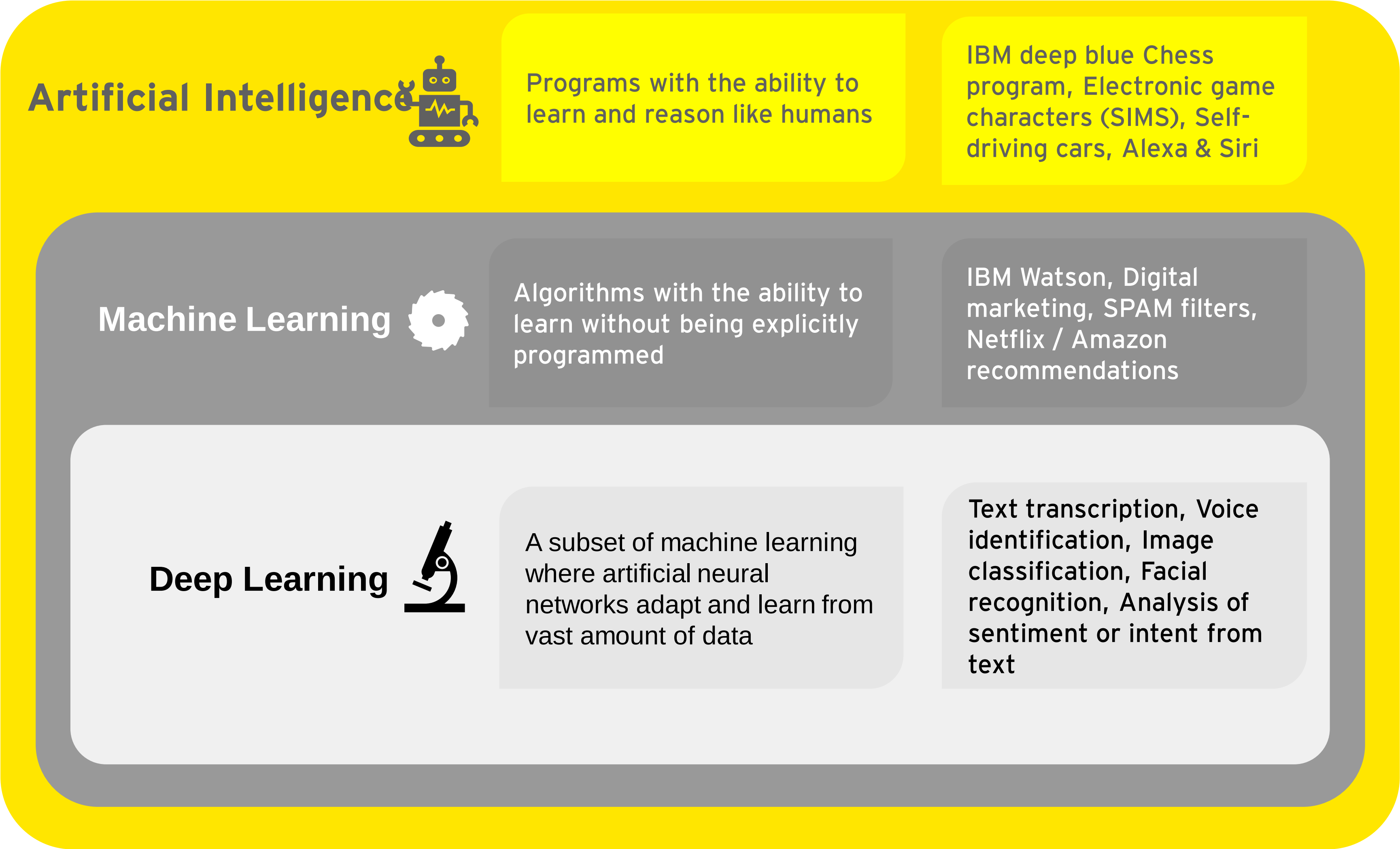
Important side notes

- Data analytics is merely a tool, not a solution to everything
  - Complement rather than a substitute
    - Human interaction, for example, will never be fully replaced
  - Critical thinking and taking responsibility more important than ever before

- Need for clear guidelines and procedures
  - The biggest challenge is choosing when it is appropriate to use data analytics
    - Ethics
    - Quality
    - Efficiency
  - Policies should be established for
    - When to use data analytics
    - When to outsource
    - When not to use data analytics

Introduction to machine learning

Artificial intelligence, machine learning & deep learning



Artificial Intelligence

Programs with the ability to learn and reason like humans

IBM deep blue Chess program, Electronic game characters (SIMS), Self-driving cars, Alexa & Siri

Machine Learning

Algorithms with the ability to learn without being explicitly programmed

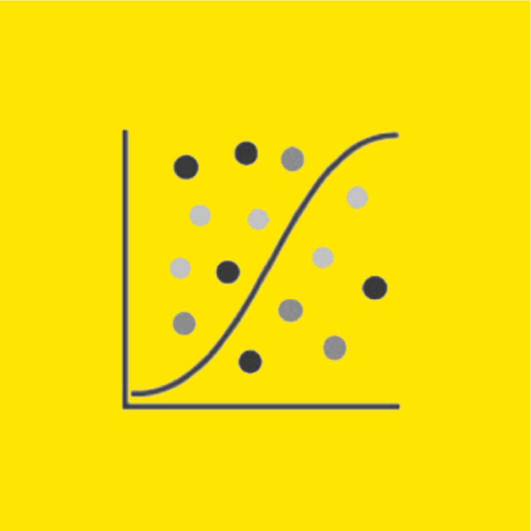
IBM Watson, Digital marketing, SPAM filters, Netflix / Amazon recommendations

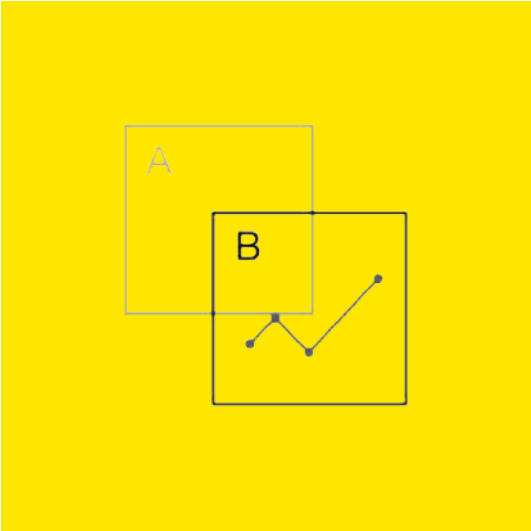
Deep Learning

A subset of machine learning where artificial neural networks adapt and learn from vast amount of data

Text transcription, Voice identification, Image classification, Facial recognition, Analysis of sentiment or intent from text

Types of data analytics





Descriptive

- Describe **what happened**
- Employed heavily across all industries and in scientific research

Predictive

- Anticipate **what will happen** (inherently probabilistic)
- Employed in data-driven organizations as a key source of insights

Prescriptive

- Provide recommendations on **what to do** to achieve the goals
- Employed heavily by leading data and internet companies

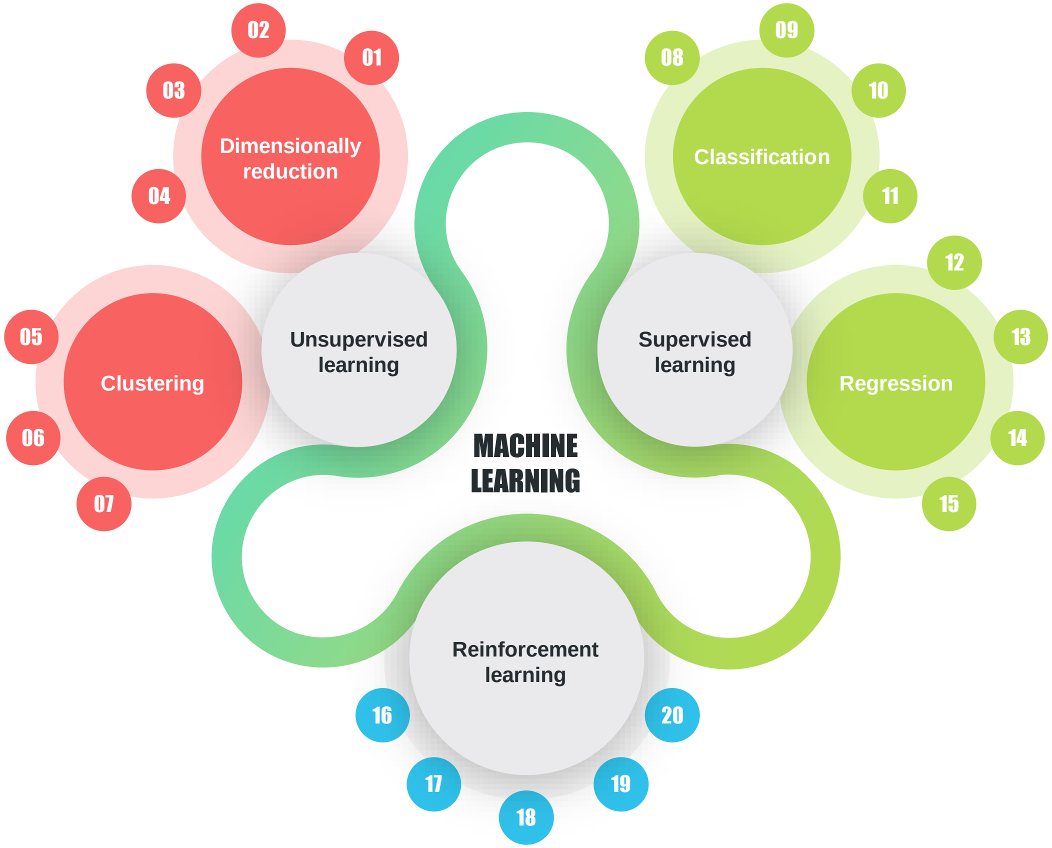
What is machine learning?

✓ A branch of **artificial intelligence**, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.

✓ As intelligence requires knowledge, it is necessary for the computers to acquire knowledge (They do so by using historical data)

01. Feature Elicitation
02. Structure Discovery
03. Meaningful compression
04. Big data Visualisation

05. Recommended Systems
06. Targeted Marketing
07. Customer Segmentation

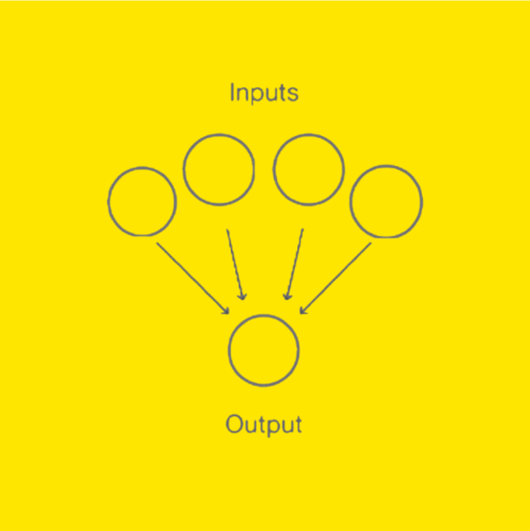


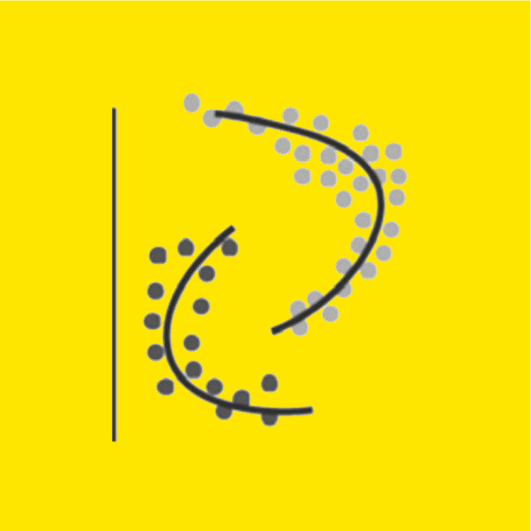
08. Fraud Detection
09. Image Classification
10. Customer Retention
11. Credit Risk, Payments & Business Failure

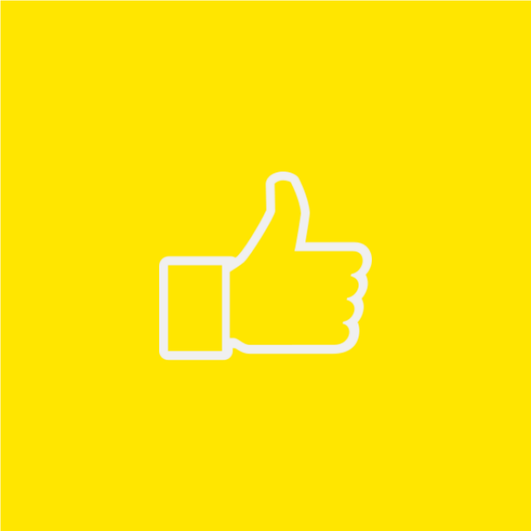
12. Forecasting
13. Predictions
14. Process Optimization
15. New Insights

16. Real-Time Decisions
17. Game AI
18. Learning Tasks
19. Skill Acquisition
20. Robot Navigation

Major types of machine learning







Supervised learning

- An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output

Unsupervised learning

- An algorithm explores input data without being given an explicit output variable

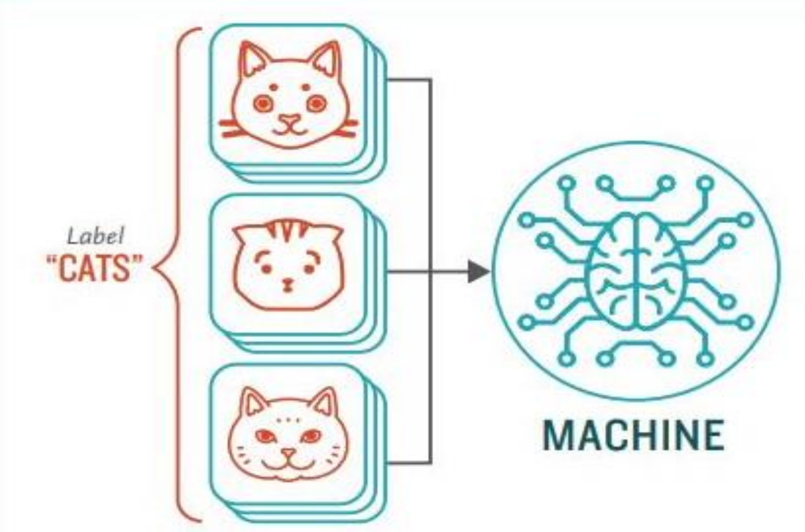
Reinforcement learning

- An algorithm learns to perform a task simply by trying to maximize rewards it receives for its actions

How supervised learning works

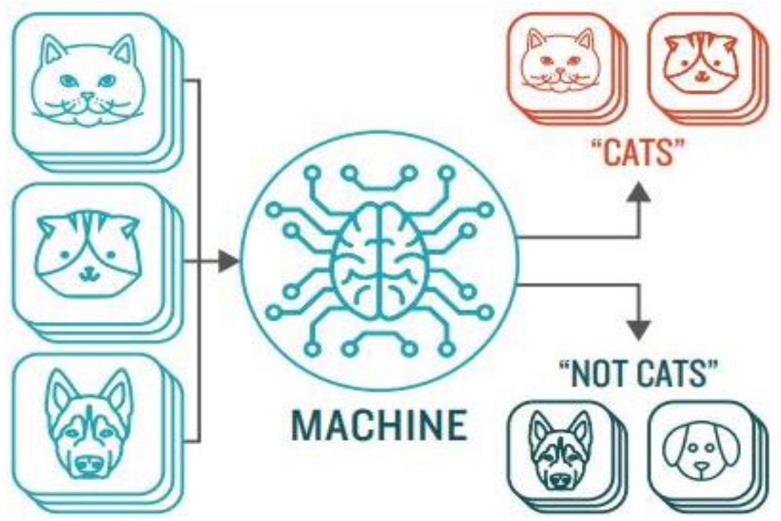
STEP 1

Provide the machine learning algorithm categorized or "labeled" input & output data to learn from

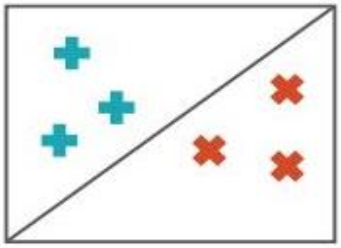


STEP 2

Feed the machine new, un-labeled information to see if it tags new data correctly. If not, continue refining the algorithm.

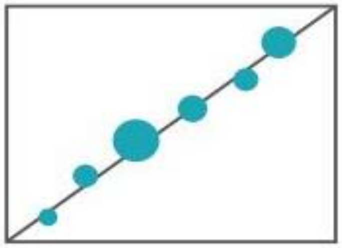


TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLASSIFICATION

Sorting items into categories



REGRESSION

Identifying real values (dollars, weight, etc.)

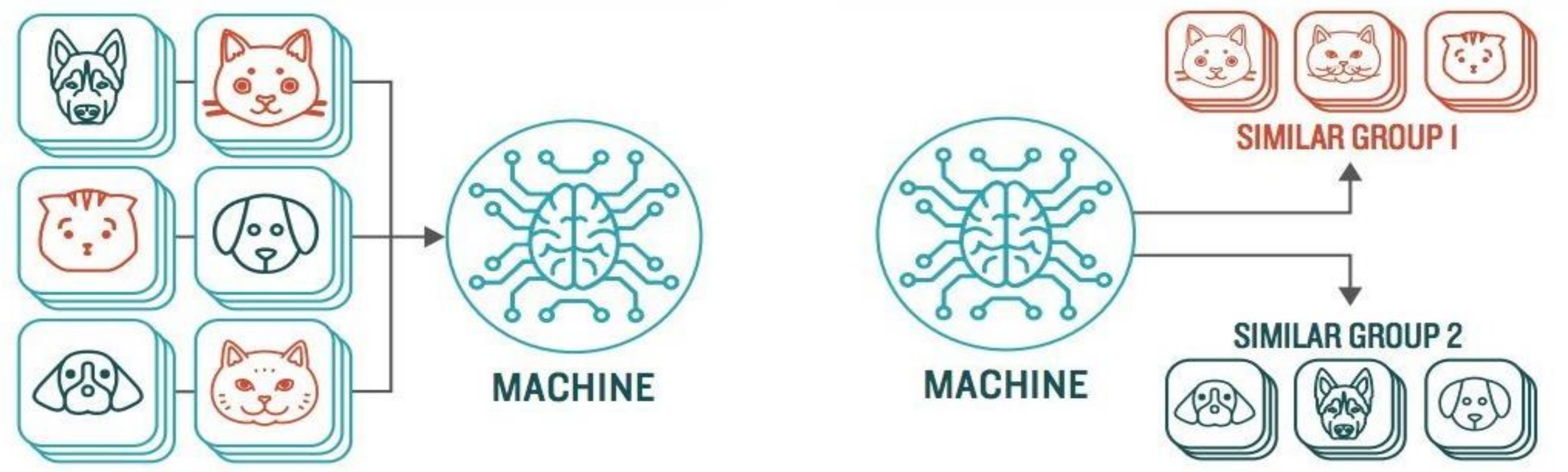
How unsupervised learning works

STEP 1

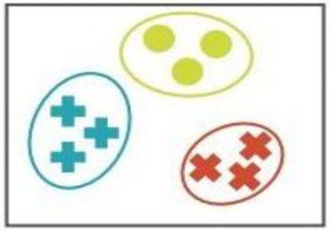
Provide the machine learning algorithm un-categorized, unlabeled data to see what pattern it finds

STEP 2

Observe and learn from the patterns the machine identifies.



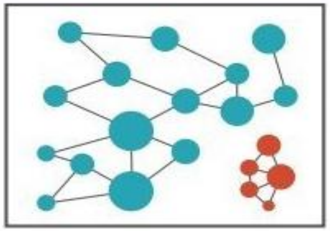
TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?



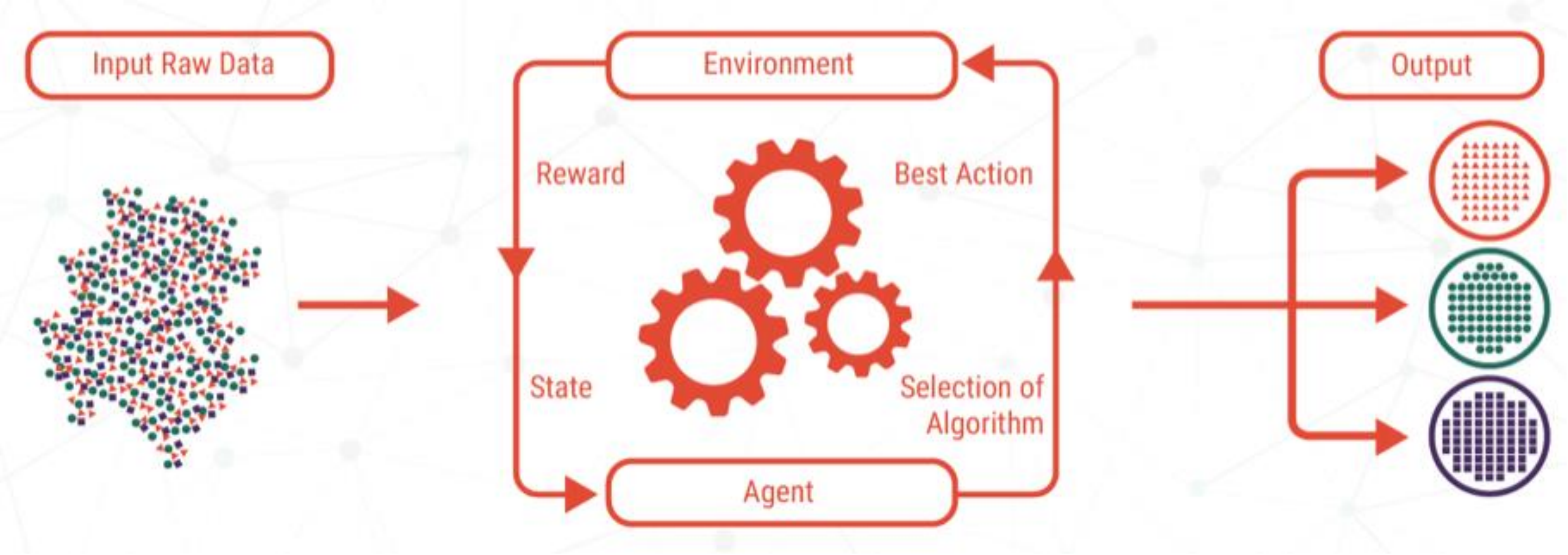
ANOMALY DETECTION

Identifying abnormalities in data

For Example: Is a hacker intruding in our network?

How reinforcement learning works

REINFORCEMENT LEARNING



- An approach to AI which relies on reward-based learning
- Learning from positive and negative reinforcement on decisions/actions taken in specific states/observations
- The machine learns how to act in a certain environments to maximize rewards

Linear regression

- Highly interpretable, standard method for modeling the past relationship between independent input variables and dependent output variables to help predict future values of the output variables
  - Sample business use cases:
    - Understand product-sales drivers such as competition prices, distribution, advertisement, etc.
    - Forecast revenue streams based on previous sales and characteristics of the market and competition
    - Test the results of different pricing strategies in order to recommend a pricing policy

Example case[1]

[1] https://colab.research.google.com/drive/1YuDlidH5w63IrHxmeESo9KXGavPnOKUM

Goal of regression analysis

- Examine the relationship between two or more variables of interest
  - Determine the influence of one or more variables on another variable (ceteris paribus)
    - Derive the size of the effect
    - Identify the statistical properties of the estimated effects

Example case – TransportEY (1/8)

- Case introduction
  - Looking at the success of Amazon and its expansions into the transport industry, the EY management has decided to pilot a new service line: TransportEY.
  - TransportEY is a package delivery service that aims to compete with Amazon. Due to the potential threat the company could form to Amazon, Amazon has shown interest in purchasing TransportEY and hired us to perform diligence for the potential transaction.

Example case – TransportEY (2/8)

- Problem definition
  - Due to a lack of structure at the launch of the company, the price of transporting a package was determined per order by people from the sales department. One of its salesmen claims that the price charged consists of a constant base tariff for shipping and a variable fee based on the weight of the package and that not all salesmen charge the exact same prices
  - We want to test the claim that the price is based on a constant fee and a variable fee based on the weight of a package and aim to estimate the pricing function used by the sales department

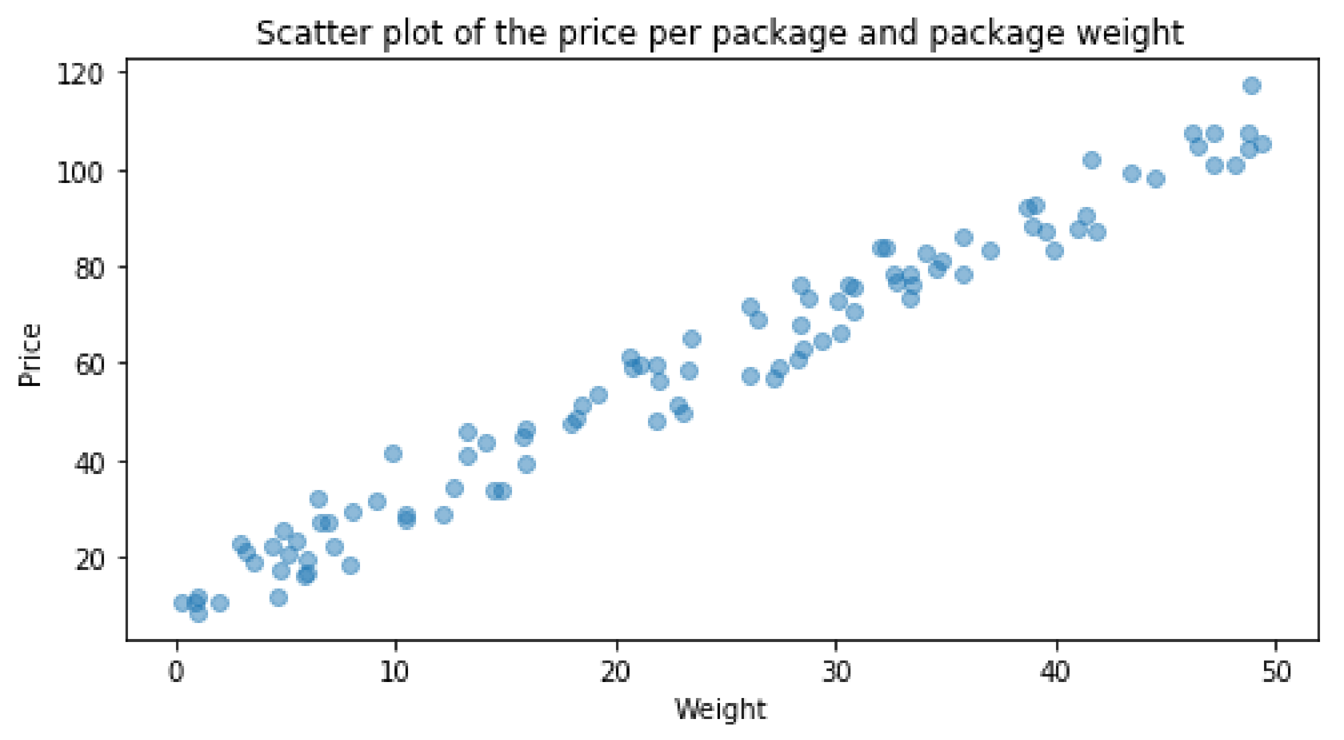
Example case – TransportEY (3/8)

- Hypothesis:
  - Price consists of a constant base tariff and a variable fee based on the weight of a package

- Additional goal:
  - Estimate the pricing function used by the sales department

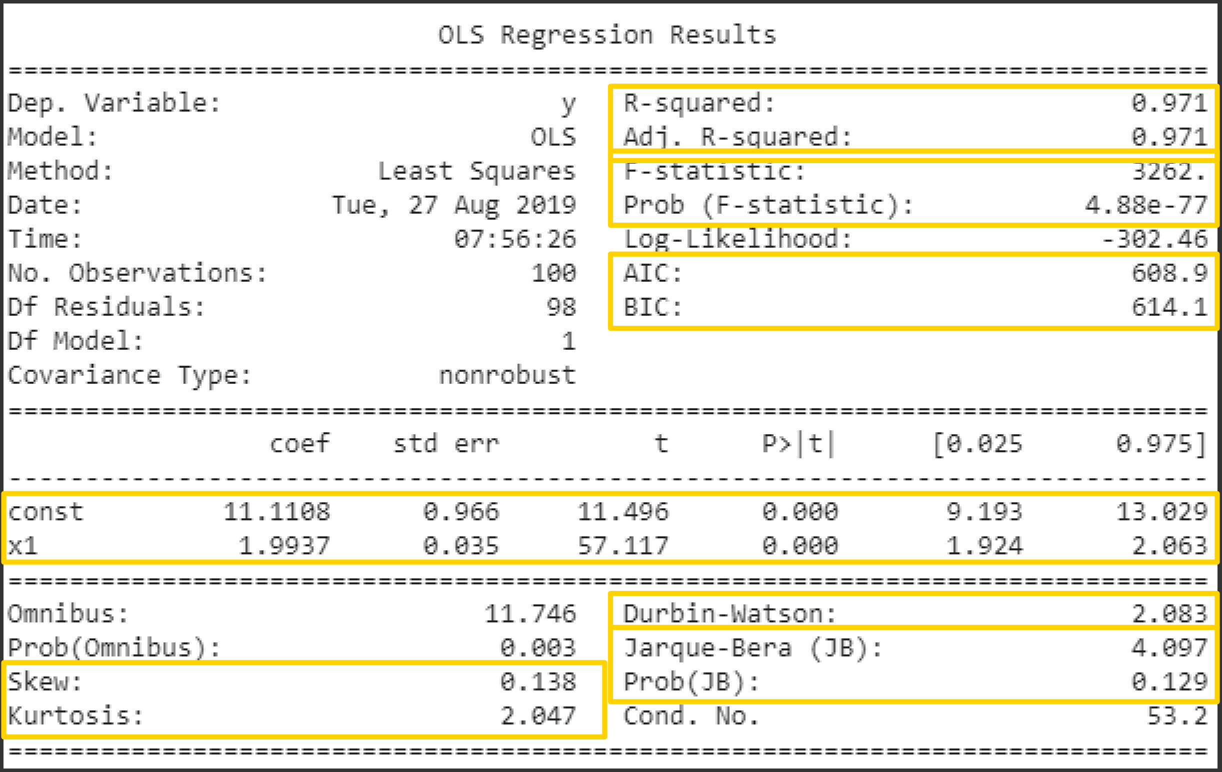
Example case – TransportEY (4/8)

- The sales data provided to us is shown in the scatter plot below



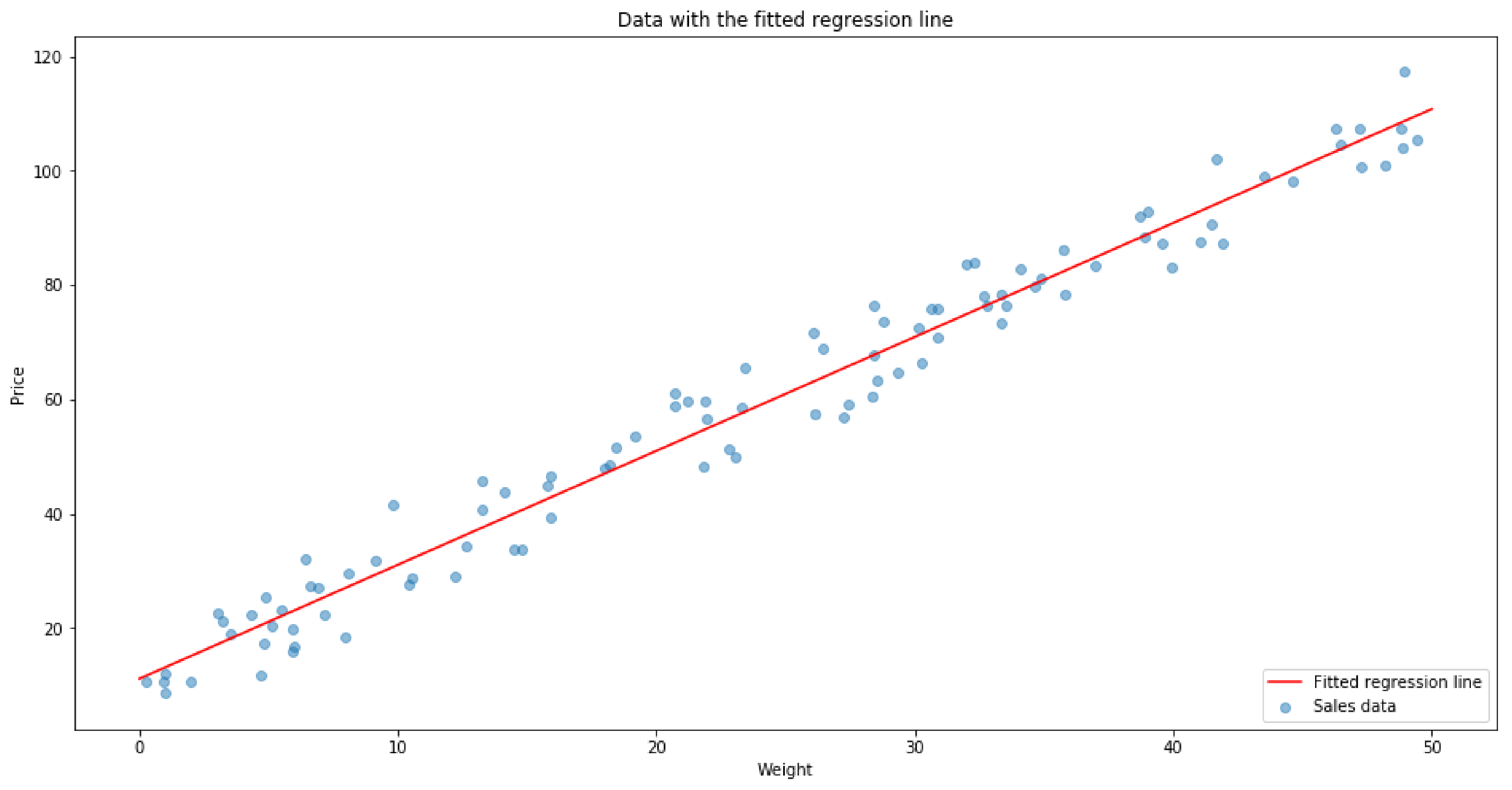
Example case – TransportEY (5/8)

- The regression output is shown below

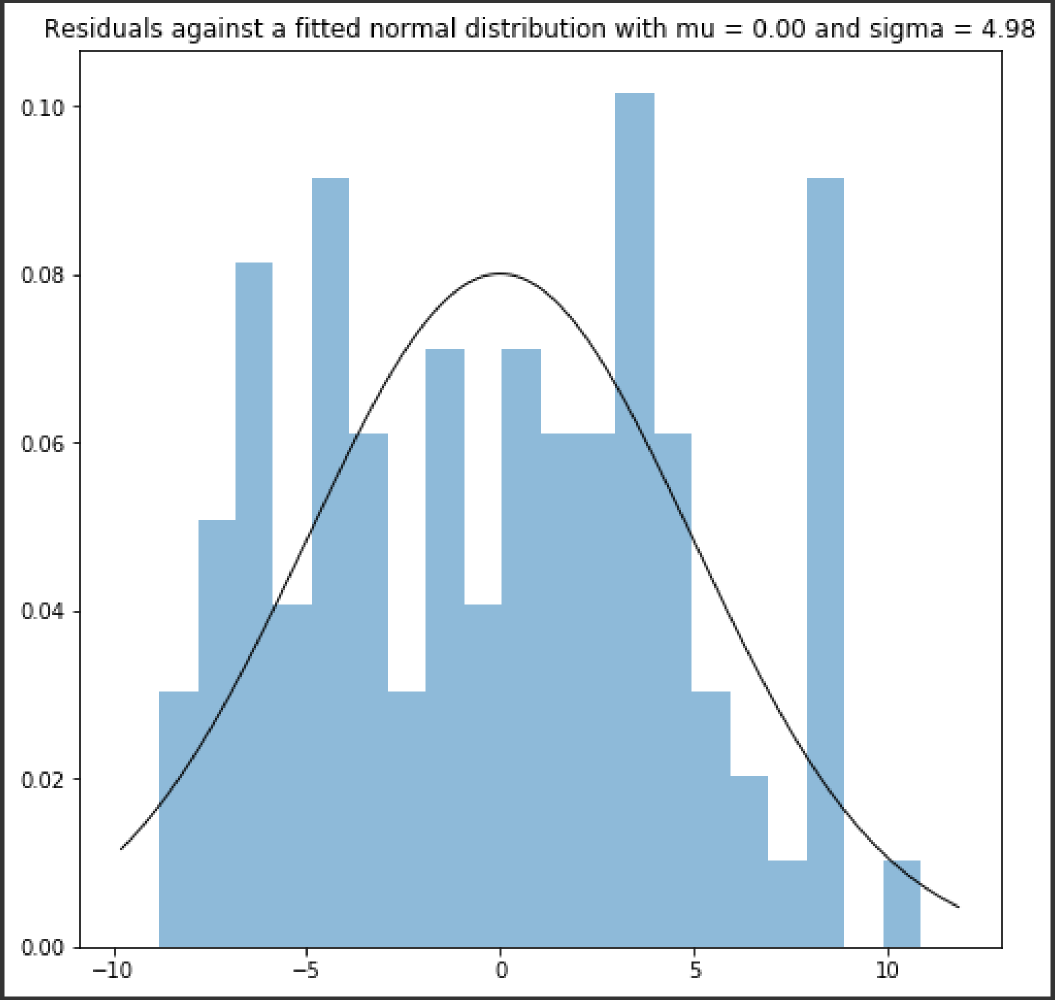


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.971
Model:                            OLS   Adj. R-squared:                  0.971
Method:                 Least Squares   F-statistic:                     3262.
Date:                Tue, 27 Aug 2019   Prob (F-statistic):           4.88e-77
Time:                        07:56:26   Log-Likelihood:                -302.46
No. Observations:                 100   AIC:                             608.9
Df Residuals:                      98   BIC:                             614.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.1108      0.966     11.496      0.000       9.193      13.029
x1             1.9937      0.035     57.117      0.000       1.924       2.063
==============================================================================
Omnibus:                       11.746   Durbin-Watson:                   2.083
Prob(Omnibus):                  0.003   Jarque-Bera (JB):                4.097
Skew:                           0.138   Prob(JB):                        0.129
Kurtosis:                       2.047   Cond. No.                         53.2
==============================================================================
```

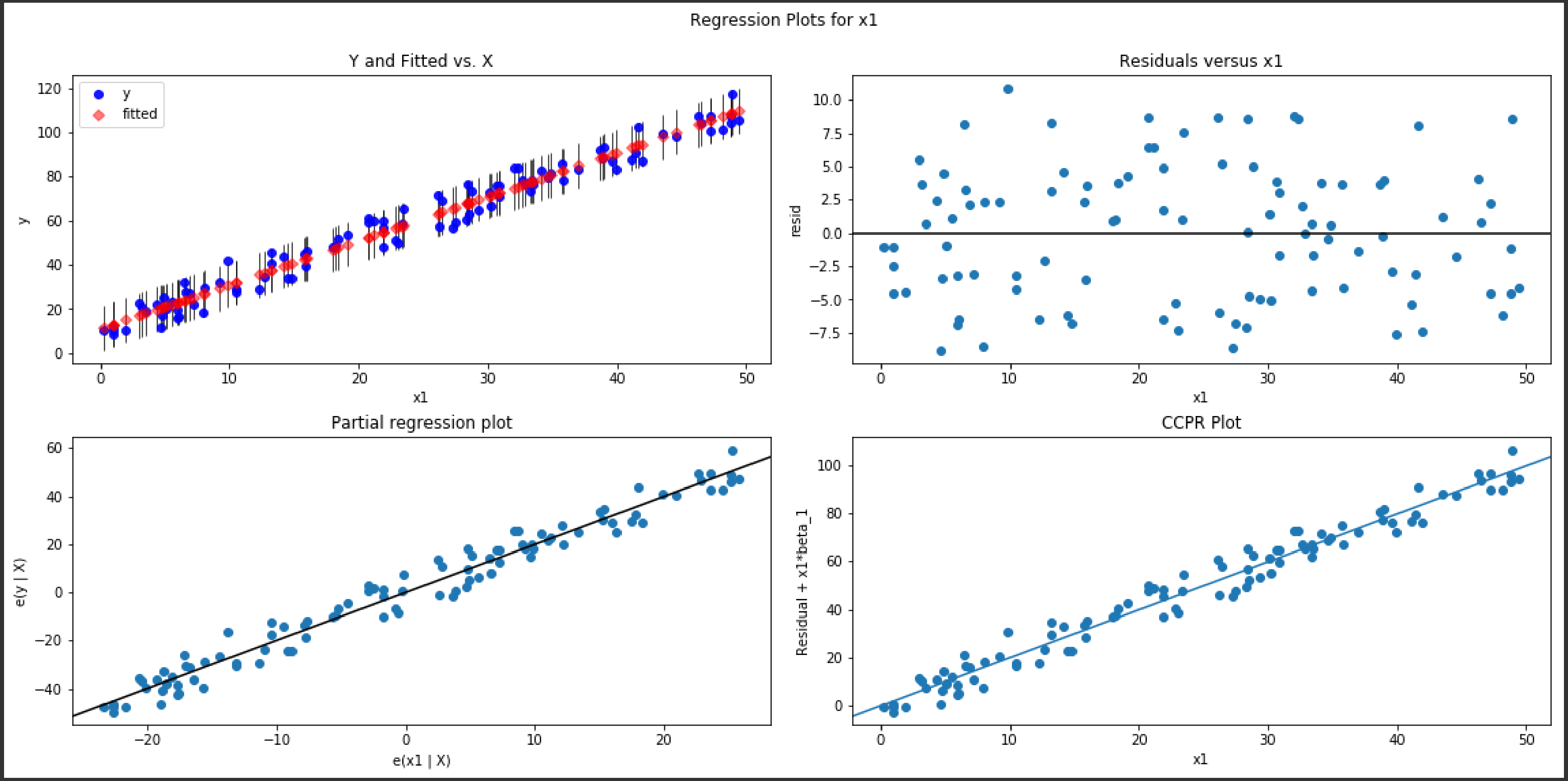
Example case – TransportEY (6/8)



Example case – TransportEY (7/8)



Example case – TransportEY (8/8)

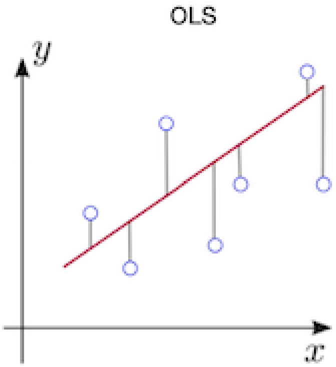


Regression fundamentals

Ordinary least squares (OLS)

- Starting point
  - Set of points in a scatter diagram
    - $(x_i, y_i),\ i = 1, 2, \ldots, N$
  - Goal
    - Find the line that gives the best fit to these points
      - $y = a + bx$

- Terminology
  - $y \equiv$ dependent variable
  - $x \equiv$ explanatory variable
  - Measure the deviations e_i of the observations from the line vertically, that is, $e_i = y_i - (a + bx_i)$



Criterion functions and deriving the optimization problem

- How do we determine which line fits the data best?
  - Criterion function $S(a, b)$
    - Special cases
      - $S_{ABS}(a, b) = \sum_{i=1}^{N} |e_i|$
      - $S_{OLS}(a, b) = \sum_{i=1}^{N} e_i^2$
      - $S_{REG}(a, b; \lambda) = \sum_{i=1}^{N} R(e_i) + \lambda L(a, b)$
  - Objective
    - Choose (a, b) such that $S(a, b)$ is minimized, that is,

$$(a, b) = \operatorname{argmin}_{a,b} S(a, b)$$

Derivation of OLS estimators

- The criterion function for OLS is

$$S_{OLS}(a,b) = \sum_{i=1}^{N} e_i^2$$

- The optimization problem can be solved algebraically
  - OLS estimators in scalar notation
    - $a = \bar{y} - b\bar{x} = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - b\sum_{i=1}^{N} x_i\right),$
    - $b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2},$
  - In matrix notation

$$\vec{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$$

OLS assumptions

OLS assumptions on model and model parameters

A1: ***Linear model***
The data on $y_1, y_2, \dots, y_N$ have been generated by

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

for $\mathrm{i} = 1, 2, \dots, N$.

A2: ***Fixed regressors***
The N observations on the explanatory variable $x_1, x_2, \dots, x_N$ are *fixed numbers* and they satisfy $\sum_{i=1}^{N}(x_i - \bar{x})^2 > 0$.

A3: ***Constant parameters***
The parameters α, β and σ are *fixed unknown numbers* with $\sigma > 0$.

OLS assumptions on the disturbances

A4: ***Random mean zero disturbances***
The N disturbances $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ are *random variables with zero mean*, that is,

$$\mathbf{E}[\varepsilon_i] = 0,$$

for $\mathrm{i} = 1, 2, \dots, N$.

A5: ***Homoskedasticity***
The variances of the N disturbances $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ *exist* and are *all equal to*

$$\mathbf{E}\left[\varepsilon_i^2\right] = \sigma^2,$$

for $\mathrm{i} = 1, 2, \dots, N$.

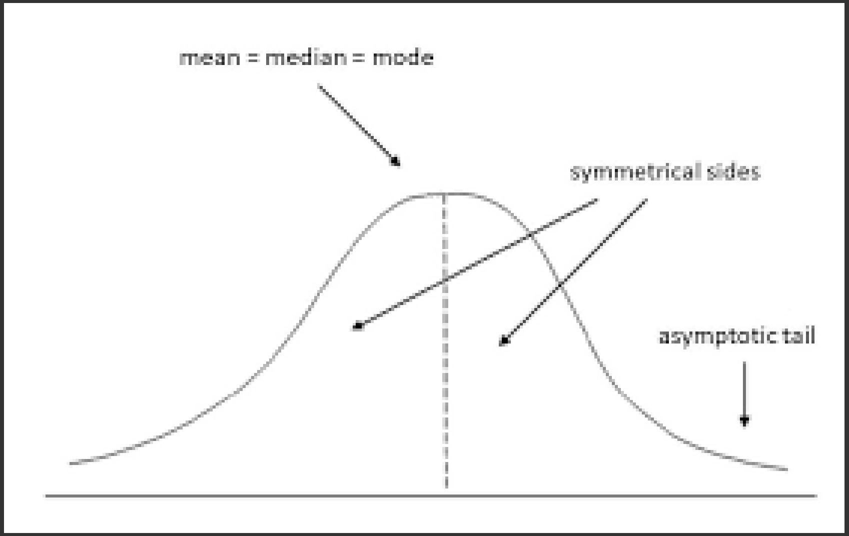
A6: ***Uncorrelated disturbances***
All pairs of disturbances $(\varepsilon_i, \varepsilon_j)$ are uncorrelated,

$$\mathbf{E}\left[\varepsilon_i \varepsilon_j\right] = 0,$$

for $\mathrm{i}, \mathrm{j} = 1, 2, \dots, N$ with $\mathrm{i} \neq \mathrm{j}$.

OLS assumptions on the probability distribution

A7: *Normality*
The N disturbances $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ are jointly normally distributed.



Regression assignment[1]

[1] https://colab.research.google.com/drive/1lmfsq0iYCtpQq3p5a6RbYKIZpNz-sjpv

Concluding remarks